UNIVERZA V MARIBORU
FAKULTETA ZA LOGISTIKO

Učbenik

STATISTIKA IN UVOD V REGRESIJSKE MODELE V MATLABU PRI OPTIMIZACIJI LOGISTIČNIH PROCESOV

Dejan Dragan

Celje, April 2014

Naslov:

Statistika in uvod v regresijske modele v Matlabu pri Optimizaciji logističnih procesov

Izdajatelj:

Univerza v Mariboru, Fakulteta za logistiko

Avtor:

doc. dr. Dejan Dragan

Oblikovanje naslovnice:

doc. dr. Dejan Dragan

Število izvodov: 20

KAZALO

Predgovor

To delo je zasnovano kot gradivo, ki obravnava snov na področju univariantne statistike in osnovnih statističnih linearnih regresijskih modelov. Učbenik predstavlja sestavni del obravnave prenovljenih vsebin obveznega predmeta "Optimizacija logističnih procesov", ki se predava v 1. letniku na podiplomskem študiju Fakultete za logistiko. Poleg tega je snov tega učbenika koristen uvod v teorijo regresijskih modelov, ki se bo prav tako pojavljala v okviru predmeta "Optimizacija logističnih procesov", ponekod pa tudi v kontekstu prenovljenega obveznega predmeta "Stohastični procesi v logistiki" v 2. letniku podiplomskega študija.

Delo na začetku predstavi splošen uvod v statistiko in analizo podatkov, ter njuno vlogo pri analizi časovnih vrst kot pomembni kategoriji podatkov na številnih področjih znanosti, tudi v okviru logistike in operacijskih raziskav. Nato se preletijo kratke teoretične osnove teorije verjetnosti, potrebne za razumevanje nadaljnje snovi. Sledijo osnove opisne statistike, pregled lastnosti posebnih verjetnostnih porazdelitev, ter porazdelitve vzorčnih statistik. V sklepnem delu učbenika pa sledijo klasična področja univariantne statistike, kot so npr. preverjanje hipotez, intervali zaupanja, ter korelacija in regresija.

Kar se tiče regresije, je slednji namenjeno precej obsežno zadnje poglavje "Korelacija in regresija". To poglavje predstavlja uvod v osnove regresijskih modelov s poudarkom na enostavnih linearnih regresijskih modelih v obliki regresijske premice. Slednja je vključena v statistične mehanizme za preučevanje linearne povezanosti med določeno odvisno in določeno neodvisno spremenljivko. Pri tem pa so obravnavane tudi vzorčne porazdelitve, statistike, hipoteze in intervali zaupanja, ki se tičejo ocenjenih regresijskih parametrov, ocenjenega matematičnega upanja odvisne spremenljivke, ter napovedi odvisne spremenljivke. Kot je razvidno iz gradiva, se da že z uporabo enostavnih linearnih regresijskih modelov rešiti marsikateri regresijski problem v okviru optimizacije in upravljanja logističnih procesov ter operacijskih raziskav. Še bolj pomembna pa so teoretična ozadja enostavnih linearnih regresijskih modelov za razumevanje obravnave bolj kompleksne regresijske teorije, ki vključuje npr. multiplo in logistično regresijo, regresijske modele pri obravnavi časovnih vrst, obravnavo regresije in regresijskih problemov v okviru ekonometrije, in podobno.

Ker se učbenik marsikje posveti razlagi dokaj zahtevnih statističnih konceptov, se pričakuje vsaj osnovno predznanje študentov s področja visokošolske matematike, vključno s poznavanjem funkcij več spremenljivk in znanjem integralnega oz. diferencialnega računa. V gradivu je teoretična obravnava v veliki meri podkrepljena tudi z doslednimi in natančnimi izpeljavami, s konkretnimi primeri in nazornimi postopki reševanja. Verjamemo, da to opravičuje relativno velik obseg tega učbenika, saj bralec lahko brez večjih težav ujame ritem razlage in slednjemu sledi do konca izračunov.

Poleg tega so v učbeniku v večini primerov dodani tudi računalniški programi in izpisi ustreznih rezultatov v komandnem oknu, ki uporabljajo programsko orodje Matlab in ponekod tudi kličejo funkcije njegovega "Statistics Toolbox-a". Tudi obsežnost nekaterih izmed teh programov in izpisov njihovih rezultatov seveda vpliva na celotno obsežnost tega gradiva. Prisotna sta dva poglavitna razloga za podrobno obravnavo primerov v Matlabu. Prvi se skriva v namenu, da bralec spozna izreden pomen pomoči sodobnih programskih orodij pri reševanju problemov s področja statistike, analize podatkov in statističnega modeliranja. Drugi razlog pa je želja, da bi študentje na praktičnih primerih kar se da dobro utrdili svoje znanje pri uporabi Matlaba, ki se zaradi svoje specifičnosti vse bolj uveljavlja v reševanju različnih problemov tako s področja naravoslovnih, tehniških, ekonomskih, pa tudi družboslovnih ved. Glede na takšno koncentracijo rešenih primerov v Matlabu verjamemo, da gre za eno izmed redkih pedagoških gradiv, ki bi se tako podrobno lotilo ilustracije uporabe tega orodja.

To gradivo seveda ni v celoti originalno, pač pa se opira na številne učbenike in druga gradiva. Večina slednjih je navedenih v seznamu literature, zato na tem mestu naštejmo le tista gradiva, na katera smo se najbolj opirali pri tvorbi tega dela: Dragan: Stohastični procesi v logistiki in Upravljanje logističnih sistemov, Jesenko: Statistika v organizaciji in managementu, Brvar: Statistika, Artenjak: Poslovna statistika, Hsu: Schaum's Outline of Probability, Random Variables, and Random Processes, Bernstein, Schaum's Outline of Theory and Problems of Elements of Statistics II, Kutner, Applied Linear Statistical Models, Montgomery, Applied Statistics and Probability for Engineers, Walpole, Probability & Statistics for Engineers & Scientists, Jurišić, Verjetnostni račun in statistika, Turk, Verjetnostni račun in statistika, Žibert, Verjetnost in statistika v tehniki in naravoslovju, Košmelj K., Uporabna statistika, Košmelj B., Statistično sklepanje, Ljubič, Predvidevanje in napovedovanje v oskrbovalni verigi, ter deli prof. Usenika.

Delo je zasnovano na takšen način, da v smislu interdisciplinarnosti pokriva različna področja znanosti s problematiko statistike, analize podatkov in statističnega modeliranja, čeprav je težišče usmerjeno predvsem k reševanju praktičnih primerov iz logistike, organizacije dela, operacijskih raziskav in ekonometrije. Zaradi svoje raznolikosti vsekakor verjamemo, da bi bilo gradivo primerno tudi za raziskovalce in študente drugih fakultet, raziskovalnih institutov in podobnih organizacij, še posebej tistih z naravoslovno-tehniškim karakterjem.

Poudarimo še, da je to delo šele prva verzija gradiva, zato ni izključena možnost določenih tiskarskih in podobnih napak. Za morebitne napake se bralcu že vnaprej opravičujemo in bomo hvaležni za vsak kritičen komentar.

Avtor

1 UVOD

1.1. Statistika in analiza podatkov

Statistika je področje znanosti, ki obravnava zbiranje, organiziranje, analizo, interpretacijo in ustrezno predstavitev podatkov. Ukvarja se z vsemi vidiki podatkov, vključno s planiranjem zbiranja podatkov v smislu načrtovanja eksperimentov.

V vsakdanjem življenju večkrat govorimo o statistiki. Kdo še ni slišal za statistiko življenskih stroškov, statistiko cen ali za statistiko prebivalstva. Včasih si s tem pojmom predstavljamo na primer povprečne plače v državi ali povprečne življenske stroške. V vsakem primeru pa je pojem statistike povezan z velikimi količinami podatkov, ki jih urejamo v tabele ali pa iz njih sestavljamo kakšne grafikone in računamo najrazličnejše količine kot pokazatelje določenih zakonitosti pojavov [Jesenko].

Nekateri obravnavajo statistiko kot samostojno področje znanosti, medtem ko je za druge statistika posebna veja matematike, ki se ukvarja z zbiranjem in interpretacijo podatkov. Verjetno imajo prvi bolj pravilen pogled zaradi njenih empiričnih korenin, iz katerih je statistika izšla, in zaradi njene predvsem aplikativne orientiranosti.

Tako lahko pod pojmom statistika razumemo različne stvari, od zbiranja podatkov, do ustanov, ki se ukvarjajo z zbiranjem in obdelavo podatkov. Za nas je statistika predvsem znanost, ki se ukvarja z zakonitostmi množičnih pojavov. Obravnava vprašanja, ki izvirajo iz izkušenj, za matematično orodje pa uporablja verjetnostni račun. V statistiki je množičen pojav vsak takšen pojav, ki se v prostoru in času pojavlja v velikem številu. Množični pojavi so npr. lastnosti in pojavi v množici ljudi, večkratne meritve, delovne operacije, ki se ponavljajo, nesreče na delovnih mestih, serije izdelkov, itn [Šrekl].

Definicijo statistike bi lahko opredelili na naslednji način [Jesenko]:

Statistika je znanost, ki razvija in uporablja čim bolj učinkovite metode zbiranja, urejanja in interpretiranja numeričnih podatkov, pri čemer pa možnosti napak zaključkov in ocen temeljijo na verjetnostnem računu.

Z večpomensko besedo "statistika" običajno mislimo na "podatke", v pedagoškem in znanstvenoraziskovalnem procesu pa jo pojmujemo kot znanstveno disciplino, ki se ukvarja z zbiranjem podatkov, z razvojem metod za obdelavo podatkov, z analizo, in s predstavitvijo izidov statistične analize [Artenjak].

Pojem statistika ima naslednje pomene [Brvar]:

- Statistika je znanost, ki razvija metode o zbiranju statističnih podatkov, njhovi analizi in predstavitvi.
- Statistika je zbiranje statističnih podatkov, njihova obdelava in objavljanje.
- Statistika so sistematično zbrani podatki o enotah populacije.
- Statistika je organizacija, ki se ukvarja z zbiranjem podatkov, njihovo obdelavo in objavljanjem (statistična organizacija).

Sama beseda izvira iz latinske besede "ratio status" in italijanske soznačnice "ragione di stato" - državni interes ter iz izpeljanke statista - oseba, ki je spretna v izvajanju državnih zadev [Artenjak].

Predmet statističnega opazovanja so s statistično raziskavo opredeljeni pojavi, predpogoj za smotrnost teh raziskav pa zagotavljata množičnost pojavljanja v času in prostoru ter variiranje vrednosti spremenljivk. Statistika namreč stremi za tem, da iz množice podatkov na osnovi različnih statističnih pristopov z izločitvijo nebistvenega ugotavlja pomembne kvantitativne in kvalitativne značilnosti opazovanih pojavov, ki jih v statistiki imenujejo statistični parametri [Artenjak].

Različna okolja nenehno ustvarjajo obilico podatkov, ki jih seveda ne bi mogli predelati, če ne bi poznali različnih načinov njihovega prikazovanja in zgoščevanja. S pomočjo statističnega urejanja in obdelave ter metod statistične analize lahko ponavadi nepregledno množico različnih podatkov, z nikakršno ali pa zelo skromno informativno vrednostjo, predelamo v manjše število kar najbolj koristnih informacij. Zato je poglavitni cilj statistike preoblikovanje podatkov opazovanih pojavov v statistične parametre in modele, ki jih lahko koristno uporabimo pri takšnem ali drugačnem odločanju [Artenjak].

Za uresničitev tovrstnih ciljev pa se pred statistiko postavljajo naslednje naloge [Artenjak]:

- zbiranje, urejanje in prikazovanje podatkov,
- prikaz možnosti in omejitev uporabe statističnih metod,
- statistična analiza in vrednotenje njenih izidov, ter
- predstavitev izidov statistične analize.

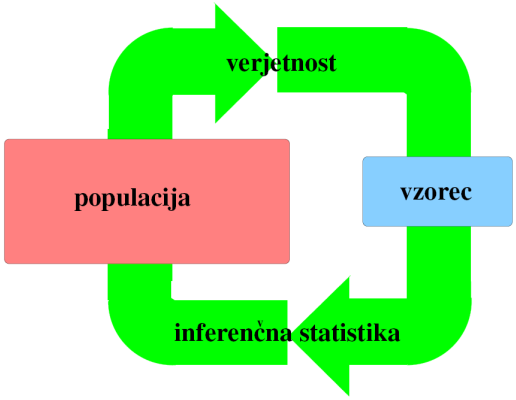
Področja uporabe statistike so številna. Tako lahko npr. govorimo o statistiki prebivalstva, statistiki kmetijstva, statistiki industrije in podobno. V ekonomiji so predmet statističnega proučevanja socialno-ekonomski pojavi na mikro in makroravni. Statistika se pojavlja tudi v številnih drugih disciplinah, vključujoč naravoslovne in tehniške znanosti, socialne znanosti, ter poslovne znanosti.

Statistiko kot znanstveno-analitično metodologijo delimo na opisno (deskriptivno) in analitično (matematično, induktivno) statistiko. Opisna statistika se ukvarja z organiziranjem, povzemanjem in opisovanjem zbirk podatkov, pri čemer statistični zaključki veljajo izrecno za uporabljeno empirično gradivo [Artenjak].

Po drugi strani pa analitična statistika jemlje vzorce podatkov (delno opazovanje - vzorec), pri čemer se izvede statistična analiza. Izide slednje pa nato po določenem statističnem pravilu posplošimo na celotno statistično množico. Pri tem gre za takoimenovano statistično sklepanje oz. inferenčnost o populaciji, ki temelji na teoriji verjetnosti [Jurišič, Artenjak].

Povezava med statistiko in verjetnostjo

Statistika in verjetnost sta tesno povezani. Razlika je v tem, da teorija verjetnosti izhaja iz danih parametrov znane populacije z namenom dedukcije verjetnosti, ki se nanašajo na posamezne vzorce. Po drugi strani pa statistična inferenca deluje v obratni smeri, torej gre za induktivno sklepanje o parametrih celotne ali večje populacije na osnovi danih vzorcev. Torej nam verjetnost pomaga oceniti, kakšen bo vzorec, ki ga bomo izbrali iz dane in dobro poznane populacije, medtem ko nam inferenčna statistika pomaga delati zaključke o celotni populaciji na osnovi danih vzorcev (Slika 1) [Jurišič].



Slika 1: Povezava med statistiko in verjetnostjo [Jurišić]

Osnovni statistični pojmi

Kakršenkoli je že problem, ki ga proučujemo s statističnimi sredstvi, je vselej povezan s proučevanjem kakšnega pojava. Proučujemo pa lahko le učinke tega pojava na okolje, na katerega pojav deluje, in na podlagi teh učinkov sklepamo o lastnostih pojava. Okolje, na katerega pojav učinkuje, je običajno sestavljeno iz množice elementov, ki zaradi učinkov pojava dobijo določene lastnosti [Jesenko].

Vzemimo na primer čas proizvodnje določenega izdelka, ki je rezultat tehnologije, organizacije in še drugih faktorjev. Iz izkušenj pa nam je poznano, da nikoli ne moremo doseči enakih proizvodnih časov za vse izdelke določene vrste, pač pa so ti bolj ali manj različni med seboj zaradi različnih naključnih ali individualnih vplivov. Tako lahko rečemo, da je proizvodni čas izdelka rezultat določljivih (deterministčnih) in naključnih učinkov [Jesenko]. Določljivi učinki so rezultat delovanja pojava na elemente, s katerimi pojav opazujemo, naključni učinki pa nastajajo zato, ker elementi na učinke pojava različno reagirajo [Jesenko].

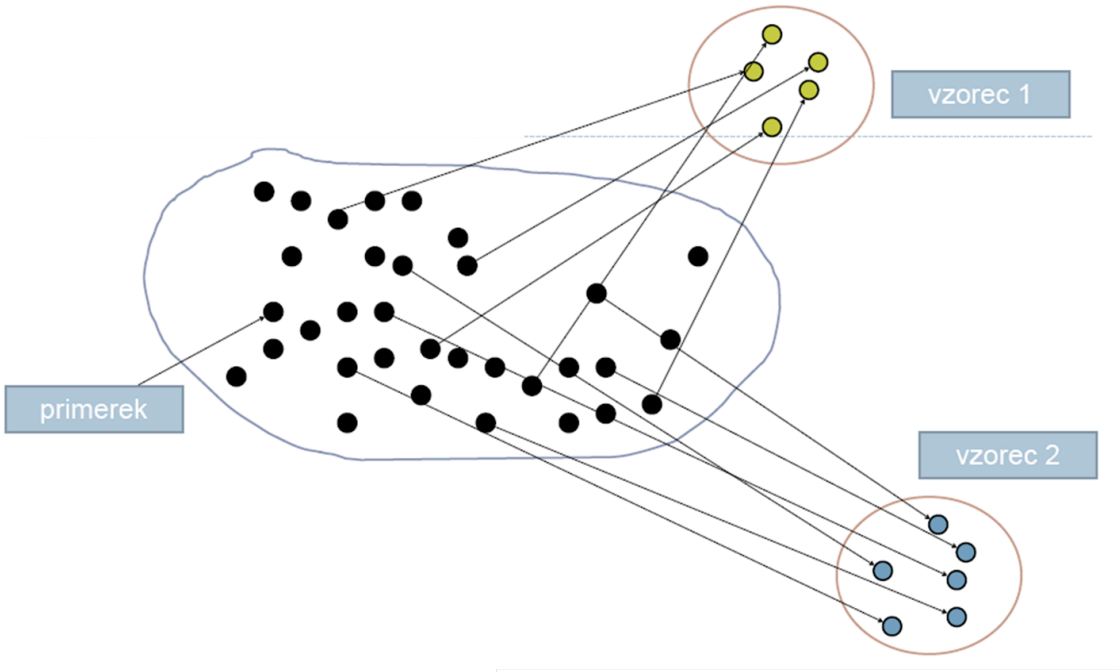
Pojave takšne vrste imenujemo **množični pojavi**. Množici vseh elementov, na katere pojav učinkuje, pravimo **populacija**, samim elementom pa pravimo **enote**. Množične pojave potemtakem vselej proučujemo tako, da proučujemo njihove učinke na ustrezno populacijo. Zato pogosto sam pojav kar poistovetimo s populacijo [Jesenko].

Množični pojav, ki je opredeljen krajevno, časovno in stvarno, imenujemo tudi **statistična množica** [Bastič]. Posameznim pojavom, ki izpolnjujejo opredelitvene pogoje, pa

pravimo **statistične enote**. Slednje imajo najrazličnejše lastnosti, proučevane značilnosti oz. lastnosti statističnih enot, ki jih analiziramo, pa imenujemo **statistične spremenljivke** [Bastič, Artenjak]. Statistična spremenljivka je torej lastnost statistične enote, ki lahko zavzame katerokoli vrednost iz določenega nabora možnih vrednosti [Brvar].

Za konkretno enoto ugotavljamo vrednost statistične spremenljivke, za konkretno statistično množico pa njene značilnosti, ki jih imenujemo **statistični parametri** [Artenjak]. Gre za kazalce, ki jih dobimo kot rezultat statističnega analiziranja populacije, ki izražajo njene lastnosti, določene z verjetnostno porazdelitvijo naključne spremenljivke v opazovani populaciji [Brvar].

Populacije so ponavadi sestavljene iz velikega števila enot (primerkov), pogosto jih imajo celo neskončno mnogo. Ker je določen pojav neracionalno opazovati na celotni populaciji, ga preučujemo samo na primerno izbranem delu populacije, ki mu pravimo **statistični vzorec** [Jesenko]. Če je vzorec izbran tako, da ima vsaka enota populacije enako verjetnost za izbiro, takšnemu vzorcu pravimo **enostavni naključni vzorec** ali kar **naključni vzorec** [Jesenko]. Slika 1a ilustrira populacijo in primere vzorcev z določenim naborom enot (primerkov) [Žibert].



Slika 1a: Ilustracija populacije in primerov vzorcev z določenim naborom enot (primerkov) [Žibert]

Gotovo z večanjem vzorca dobimo zanesljivejše informacije o pojavu, čeprav se pri tem lahko povečuje možnost napak pri opazovanju, merjenju, ali kasnejši obdelavi. Za vzorec pravimo, da je **reprezentativen**, če nam da podobne rezultate, kot bi jih dobili s preučevanjem celotne populacije [Jesenko]. Vsekakor velja, da vzorec dejansko analiziramo, medtem ko populacijo lahko le ocenjujemo.

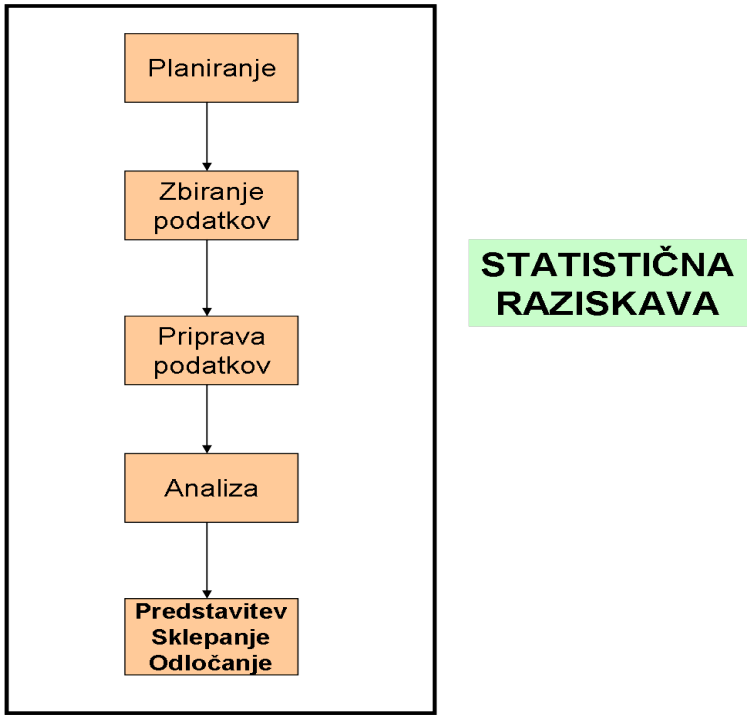
Konkretne vrednosti spremenljivk določenega pojava izražamo z **indikatorji (kazalci)** [Brvar]. Indikator je torej konkretno zaznana vrednost ali odsev spremenljivke pri opazovanju določenega konkretnega pojava.

Spremenljvike so lahko **številske ali numerične**, in **opisne ali atributne** [Brvar]. Vrednosti numerične spremenljivke izražamo s kvantitativnimi podatki - števili, na primer: starost, temperatura, tlak, itn. Vrednosti opisne spremenljivke pa izražamo s kvalitativnimi podatki, na primer: vremenske okoliščine, vidljivost, itn.

Statistika preučuje predvsem naključne spremenljivke, ki lahko zavzamejo vrednosti v kateremkoli intervalu z znano verjetnostjo [Brvar]. V primeru diskretnih naključnih spremenljivk ima spremenljivka končno število vrednosti, v primeru zveznih naključnih spremenljivk pa lahko zavzame na določenem intervalu katerokoli vrednost.

Potek statistične raziskave

Kot za vsako raziskovalno področje, velja tudi za statistične raziskave, da potekajo v več zaporednih, sicer funkcijsko različnih, vendar logično povezanih etapah [Artenjak] (glej sliko 2).



Slika 2: Potek statistične raziskave [Artenjak]

V fazi planiranja natančno opredelimo cilje raziskave, jo opredelimo s stvarnega, prostorskega in časovnega vidika, izberemo ustrezne statistične metode, ter razrešimo organizacijske probleme [Jesenko].

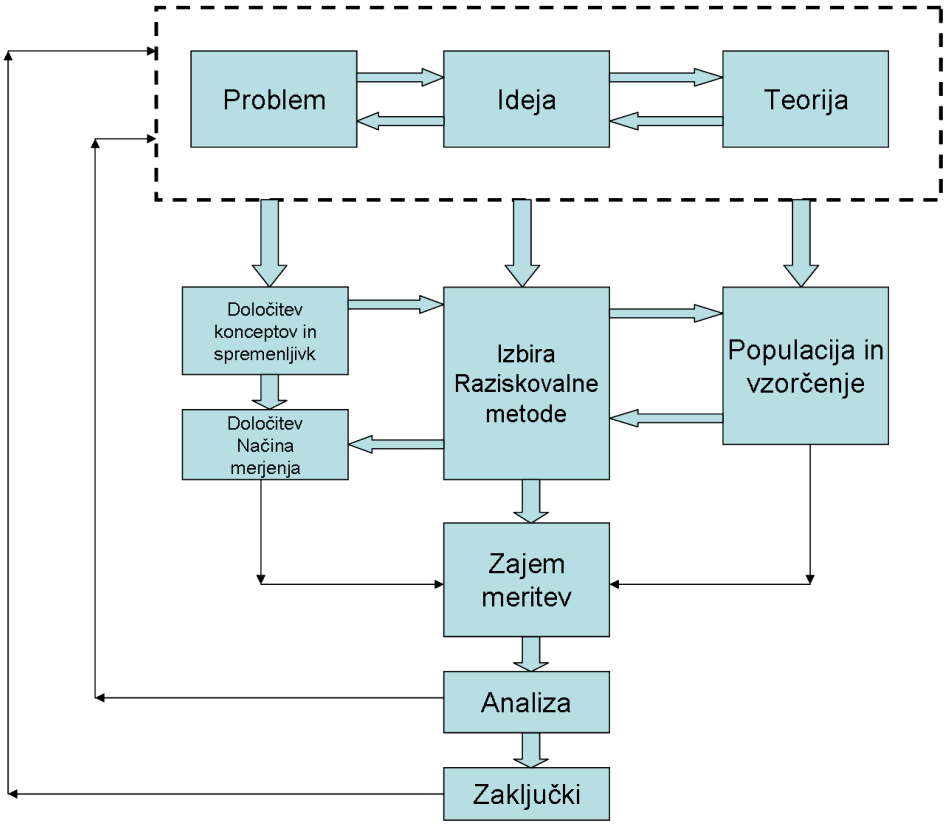
Pri zbiranju podatkov je od njihove kakovosti in količine v veliki meri odvisna tudi uporabnost izidov raziskave. Zato mora biti raziskava v tej etapi dobro zastavljena z vidika opredelitve načina zbiranja podatkov in vidika izvajalcev, oboje pa je seveda odvisno od razpoložljivega časa in denarnih sredstev [Jesenko].

V fazi priprave podatkov le-te urejamo, zgoščujemo in odkrivamo morebitne napake. Prikažemo jih v preglednicah in grafično tako, da so podatki že primerni za uporabo določene statistične metode [Jesenko].

V fazi analize podatkov le-te analiziramo na osnovi statističnih metod, pri čemer velja, da je statistična analiza najpomembnejši del vsake statistične raziskave [Jesenko].

V končni fazi poskušamo izide statistične raziskave smiselno obrazložiti, nakazati njeno koristnost in morda navesti razloge za smotrnost nadaljnjih raziskav, ter v določenih primerih tudi izvesti morebitne ukrepe in odločitve na osnovi raziskave [Jesenko].

Še bolj natančen vpogled v potek statistične raziskave prikazuje slika 3. Na njej je prikazan osnovni potek raziskovalnega procesa od zaznave problema naprej, vključujoč ves uporabljen statistični aparat [Brvar, Babbie].



Slika 3: Potek statističnega raziskovalnega procesa [Brvar, Babbie]

Napake v statističnih raziskavah

Uporabnost izidov statistične analize podatkov je odvisna tudi od kakovosti zbranih podatkov in od natančnosti računskega postopka pri uporabi določene statistične metode. Napake, ki se lahko pojavijo pri izvajanju statistične raziskave, delimo na naslednje skupine [Artenjak]:

- Sistematične napake (zaradi napak merilnega instrumenta),
- Merilne napake pri odčitavanju merilnega instrumenta,
- Napake pri zaokroževanju, ter
- Razvrstitvene napake.

Sistematične napake so posledica napake merilnega instrumenta. Tudi, če je instrument natančen, lahko pride do napak pri odčitavanju njegovih meritev. Izmerjene rezultate večkrat tudi zaokrožimo, pri čemer pri do zaokrožitvenih napak. Razvrstitvene napake pa so napake zaradi razvrščanja podatkov in se v glavnem pojavijo pri ročni obdelavi podatkov, pri računalniški obdelavi pa zaradi napak pri vnosu podatkov [Artenjak].

Statistični model

Statistični model je matematična formulacija povezav med statističnimi spremenljivkami v obliki matematičnih izrazov in enačb. Tovrsten model pojasnjuje, kako so ena ali več naključnih spremenljivk povezane z eno ali več drugimi spremenljivkami.

Inferenčna statistika

Inferenčna statistika uporablja vzorce v opazovanih podatkih, da bi izvršila določene zaključke o opazovani populaciji, vključujoč fenomene negotovosti. Inferenčna statistika lahko izvaja naslednje naloge:

- Odgovori z DA/NE na vprašanja, ki se tičejo podatkov (testiranje hipotez),
- Ocenjevanje numeričnih karakteristik podatkov,
- Opisovanje povezav med podatki (korelacija),

- Modeliranje odnosov med podatki (npr. z regresijsko analizo),
- Napovedovanje, predikcija in ocenjevanje nemerljivih vrednosti v populaciji,
- Napovedovanje, ekstrapolacija in interpolacija časovnih vrst ali prostorskih podatkov,
- Podatkovno rudarjenje, itn.

Korelacija

Koncept korelacije med spremenljivkami zna biti zelo zavajajoč. Statistična analiza določenega niza podatkov pogosto razkrije, da določeni dve spremenljivki simultano skupaj variirata po enakih zakonitostih, da izgleda, kot da bi bili povezani. Tedaj pravimo, da sta spremenljivki korelirani med seboj. Vendar se po drugi strani lahko zgodi, da sta v direktni kavzalni (vzročno-posledični) povezavi ali pa ne. Namreč, njuna korelacija je lahko povzročena tudi s strani neke tretje spremenljivke, zato na osnovi korelacije ne moremo takoj sklepati tudi o direktni medsebojni kavzalnosti dveh spremenljivk.

Negotovost

Negotovost v spremenljivkah opazovanega pojava je preučevana s pomočjo teorije verjetnosti. Slednja je uporabljena na področju matematične statistike (oz. statistične teorije) za preučevanje vzorčnih porazdelitev vzorčnih statistik, in gledano bolj splošno, s pomočjo teorije verjetnosti preučujemo tudi lastnosti statističnih procedur in postopkov.

Eksperimentalne in opazovalne študije

Eden glavnih ciljev statistične raziskave je preučevanje kavzalnosti spremenljivk opazovanega pojava in še posebej izvajanje zaključkov o vplivu sprememb neodvisnih spremenljivk na odvisne spremenljivke. Obstajata dva glavna tipa kavzalnih statističnih študij:

- Eksperimentalne študije, in
- Opazovalne študije.

Eksperimentalna študija izvaja meritve na opazovanem sistemu, manipulira z njim, nato pa ponovno izvaja dodatne meritve s pomočjo enake procedure, ter ugotavlja, če je ponovna manipulacija nad sistemom spremenila vrednosti meritev.

Za razliko od eksperimentalne študije, opazovalna študija ne vpleta eksperimentalnih manipulacij, pač pa le zajame podatke in preučuje korelacije med neodvisnimi in odvisnimi spremenljivkami.

Specializirane statistične discipline

Statistične tehnike se uporabljajo na širokem spektru znanstvenih področij. Naštejmo nekatera glavna področja uporabne statistike:

- Zavarovalniška statistika,
- Informacijska ekonomika in ekonometrija,
- Energijska statistika,
- Inženirska statistika,
- Geografska statistika,
- Statistika v psihologiji,
- Zanesljivost naprav,
- Socialna statistika, itn.

Statistične metode, analize in modeli

Obstaja zelo širok spekter statističnih metod, analiz in modelov. Naštejmo nekatere glavne izmed njih:

- Univariantne statistične metode,
- Multivariantne statistične metode,
- Statistično razvrščanje in klasifikacija,
- Analiza variance,
- Korelacijska analiza,
- Statistični testi,

- Regresijska analiza in izgradnja regresijskih modelov,
- Faktorska analiza,
- Analiza in modeliranje časovnih vrst,
- Analiza strukturnih podatkov,
- Modeliranje strukturnih enačb,
- Tehnike ponovnega vzorčenja (resampling), itn.

Silovit razmah moči računalnikov v zadnjih letih je občutno povečal uporabnost metod in tehnik v statističnih znanostih. V starejši zgodovini so bili statistični modeli skoraj vselej iz družine linearnih modelov, vendar pa so močni računalniki, podkrepljeni z učinkovitimi numeričnimi algoritmi, izredno vplivali tudi na razvoj nelinearnih modelov (npr. nevronskih mrež, itn), generaliziranih linearnih modelov, ter večnivojskih modelov v novejšem času.

Prav tako so močnejši računalniki tudi pripomogli k večji popularnosti računalniško intenzivnih metod na osnovi ponovnega vzorčenja (npr. Bootstrap metoda), kot tudi k večji uporabnosti Bayesovih modelov in tehnik iz družine "Markov Chain Monte Carlo".

1.2. Vloga statistike pri analizi in napovedovanju časovnih vrst

Statistika ima pomembno vlogo na številnih področjih znanosti, tudi pri obravnavi podatkov v obliki časovnih vrst v okviru logistike in operacijskih raziskav. Zato bomo v tem poglavju na kratko sklenili splošen uvod v statistiko s predstavitvijo problematike analize in napovedovanja časovnih vrst, ter kakšno vlogo pri tem igra inferenčna statistika.

Predvidevanje in napovedovanje je staro kot človeštvo. Od nekdaj človek želi vedeti, kaj se bo zgodilo v prihodnosti in da bi to izvedel, se zateka k "strokovnjakom", vedeževalcem, ki znajo včasih na osnovi analize trenutnih in preteklih dogodkov, največkrat pa kar počez, bolj ali manj verjetno napovedati prihodnost [Ljubič].

Tudi v poslovnem okolju želi človek videti v prihodnost in napovedati obnašanje poslovnih procesov, ki ustvarjajo novo vrednost. Takšni procesi se npr. tičejo povpraševanja oziroma prodaje, nakupa in porabe materialov, proizvodnega ciklusa in dobavnih časov, razpoložjivosti delovnih sredstev, kakovosti proizvodov in še marsičesa. Vendar pa predvidevanje in napovedovanje ni zgolj ugibanje, pač pa organizirana, s podatki in matematičnimi ter statističnimi metodami podprta dejavnost [Ljubič].

Predvidevanje je aktivnost ocenjevanja bodočih dogodkov oziroma dejavnosti. Razume se kot ocenjevanje zunanjih objektivnih razmer za poslovanje v določenem prihodnjem časovnem obdobju, torej ocenjevanje možnih razvojev zunanjih neodvisnih spremenljivk oz. razvoja nepredvidenih rpihodnjih dogodkov. Predstavlja obsežno področje metod, ki se ukvarjajo z ocenjevanjem možnih bodočih dogodkov [Ljubič].

Napovedovanje je poskus, pri katerem želimo vnaprej napovedati najbolj verjeten izid določene naključne spremenljivke. Je posebna kategorija predvidevanja, kamor spadajo matematične in statistične metode za izdelavo napovedi. Predvidevanje in napovedovanje sta torej projekciji pričakovanega obnašanja poslovnega procesa v prihodnosti, ki jo podajajo pogoji v okolju. Planiranje pa se opredeljuje kot določanje nabora akcij, ki naj se sprovedejo, da bi se doseglo (ali preseglo) napovedane vrednosti poslovnih procesov oziroma dogodkov [Ljubič].

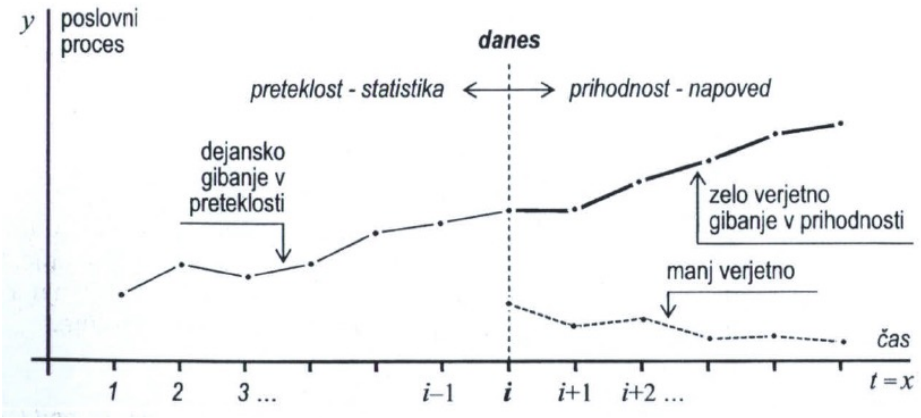
Ker ima mnogo poslovnih procesov stohastični značaj, se napovedi dogajanj v prihodnosti podrejajo stohastičnim zakonitostim in seveda niso popolnoma zanesljive. Zato se napovedovanje včasih - sicer zmotno - imenuje tudi stohastično planiranje. Vendar pa je negotovost dejstvo, ki je ves čas prisotno pri vsaki obliki planiranja in zaradi tega se ne sme vnaprej odklanjati napovedovanja kot orodja za planiranje [Ljubič].

Slika 5 prikazuje tabelo razlik med napovedovanjem in planiranjem [Ljubič].

	Napovedovanje	*Planiranje*
Kdo izvaja ?	Analitik, analitska skupina ali služba	Praviloma management na različnih ravneh
Na čem sloni ?	V veliki meri na statističnih in matematičnih metodah ali proceduralno točno določenih subjektivnih metodah	Pretežno na objektivnih deterministični metodah, pa tudi na subjektivnih metodah
Zaporedje ?	Napoved je običajno prva	Plan sledi napovedi
Nagrajevanje ?	Analitik za točnost napovedi, prepoznavanje tveganj in priložnosti ter njihovo analitično ovrednotenje	Poslovna funkcija za doseganje/preseganje planskih ciljev
Subjektivnost v procesu ?	V analitičnem delu ni prisotna, je pa prisotna pri oceni tveganj in priložnosti. Vplivi so ocenjeni analitično npr. s stohastičnimi modeli	Prisotna ves čas, pogosto v obliki "mehkih" ciljev

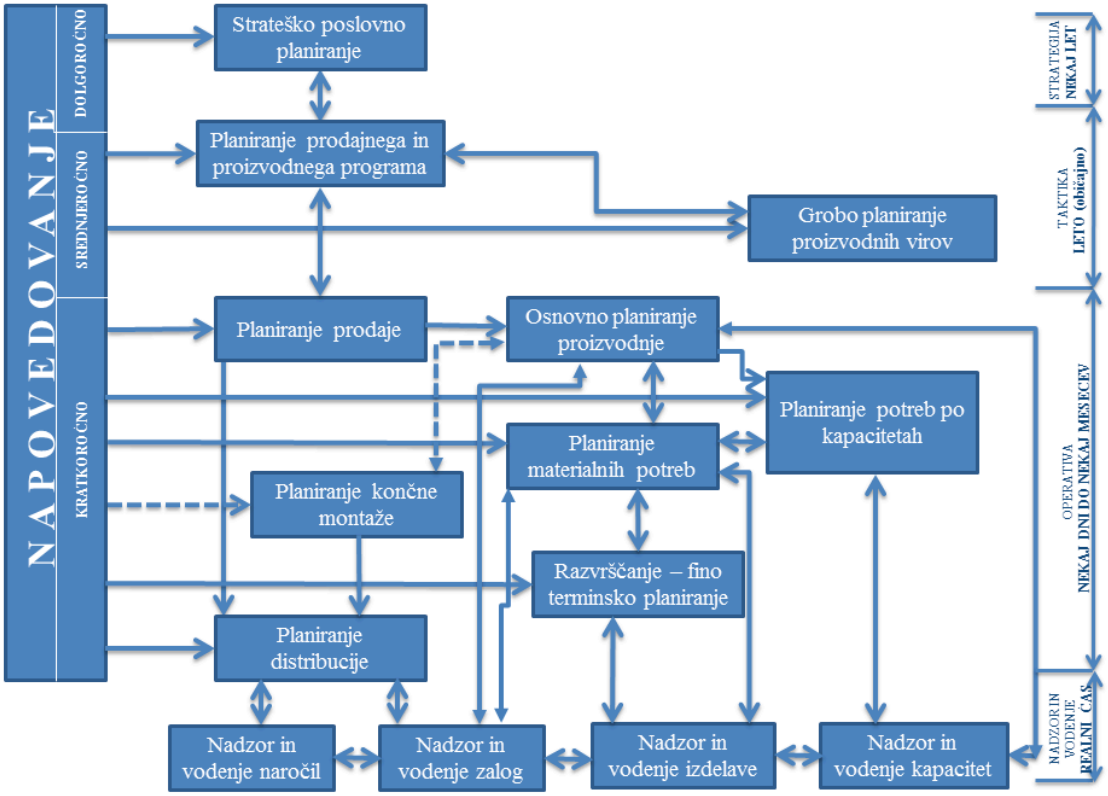
Slika 5: Razlike med napovedovanjem in planiranjem [Ljubič]

Osnova za napovedovanje je načelo vztrajnosti poslovnih procesov. To je podobno načelu fizikalne vztrajnosti, kjer poslovni proces želi vztrajati v smeri, v kateri se giblje toliko časa, dokler ga ne zmoti nek zunanji dogodek, na katerega ni mogoče več vplivati. lz dogajanj v preteklosti se torej sme preko trenutnih dogajanj sklepati na dogajanja v prihodnosti (slika 6) [Ljubič].



Slika 6: Načelo vztrajnosti poslovnih procesov [Ljubič]

Slika 7 prikazuje vlogo napovedovanja v sistemu planiranja v proizvodnem okolju [Ljubič].



Slika 7: Vloga napovedovanja v sistemu planiranja v proizvodnem okolju [Ljubič]

Napovedovanje povpraševanja in logističnih zahtev

Pri upravljanju logističnih sistemov potrebujemo napovedovanje zato, da lahko predvidimo stopnjo prihodnjih ekonomskih aktivnosti, potrebnih za sledenje preskrbe želenim zahtevam [Dragan 1].

Logistične zahteve, ki jih ima smisel napovedovati, so [Dragan 1]:

- Povpraševanje in zahteve strank, naročnikov, kupcev (da lahko planiramo produkcijo, oz. naročila izdelkov, oz. distribucijske aktivnosti),
- Napovedovanje gibanja bodočih cen,
- Napovedovanje bodočih stroškov delovne sile,
- Napovedovanje prodaje določenih artiklov v prihodnosti, da vemo planirati naročila blaga in stroške skladiščenja, itn.

Pri tem morata biti izpolnjena naslednja ključna pogoja za potrebe napovedovanja [Dragan 1]:

- Vzorec obnašanja spremenljivke, ki jo napovedujemo, se v bližnji prihodnosti ne sme preveč spremeniti, ter

- Spremenljivka, ki jo napovedujemo, mora vsaj delno zaviseti bodisi od svojih preteklih vrednosti ali/in preteklih vrednosti drugih spremenljivk.

Napovedovanje je zlasti pomembno pri optimalnem planiranju in upravljanju zalog. Ločimo več tipov napovedovanj, kot npr. [Dragan 1]:

- Napovedovanje na dolgi rok (1 do 5 let): Uporabimo ga tedaj, ko se moramo odločiti, če bi dali nov artikel na trg oz. umaknili star artikel, ter takrat, ko načrtujemo novo logistično mrežo.

- Napovedovanje na srednji rok (nekaj mesecev do 1 leta): Uporabimo ga tedaj, ko izvajamo taktične logistične odločitve, kot npr. pri postavitvi letnega plana proizvodnje in distribucije, pri načrtovanju zalog za nekaj mesecev naprej, ter pri. prerazporeditvi (alokaciji) blaga v skladišču za nekaj mesecev naprej.

- Napovedovanje na kratek rok (nekaj dni do nekaj tednov): Namenjeno za razvrščanje virov in blaga z namenom doseganja kratkoročnih in srednjeročnih proizvodnih in distribucijskih načrtov (postavitev nekaj tedenskega plana proizvodnje, distribucije in skladiščenja blaga).

Agregacija napovedi

Agregacija napovedi določa, kako podrobna je napoved. Sega od najpodrobnejših napovedi za posamezne proizvode po lokacijah ali proizvode preko bolj zgoščenih napovedi za družine proizvodov in proizvodne programe. do agregiranih napovedi za strateške poslovne enote in cel poslovni sistem [Ljubič].

Glede na vsebino se ločujejo [Ljubič]:

- **Ekonomske napovedi**, ki napovedujejo poslovni ciklus s predvidevanjem npr. stopnje inflacije, finančnih virov, investicij in drugih gospodarskih indikatorjev za daljše obdobje,

- **Tehnološke napovedi**, ki napovedujejo tehnološke dejavnike: razvoj novih proizvodov, proizvodnih procesov in zmogljivosti v srednjeročnem obdobju, ter

- **Napovedi povpraševanja** - prodaje pa so kratkoročna projekcija potreb tržišča po proizvodih oziroma storitvah. Le-te služijo kot vhod za operativno planiranje prodaje, proizvodnje in financ.

Značaj procesov, ki jih napovedujemo

Poslovni procesi oziroma pojavi, ki se napovedujejo, so [Ljubič]:

- **endogeni**, če so odvisni izključno od časa in jih je zato mogoče dokaj zanesljivo napovedati, ter

- **eksogeni**, na katere poleg časa vpliva še mnogo drugih več ali manj nepoznanih dejavnikov, predvsem ekološkega in sociološkega značaja (ki so načeloma manj predvidljivi), zatorej jih ni mogoče zanesljivo napovedati.

Če se opazuje srednjeročno in kratkoročno gibanje nekega pojava, se ločuje [Ljubič]:

- **stacionarne pojave**, ki so dokaj stabilni (se v kratkem časovnem obdobju le malo spreminjajo), nanje ima prevladujoč vpliv čas, vplivi drugih faktorjev pa so zanemarljivi. Ti pojavi so torej endogenega značaja,

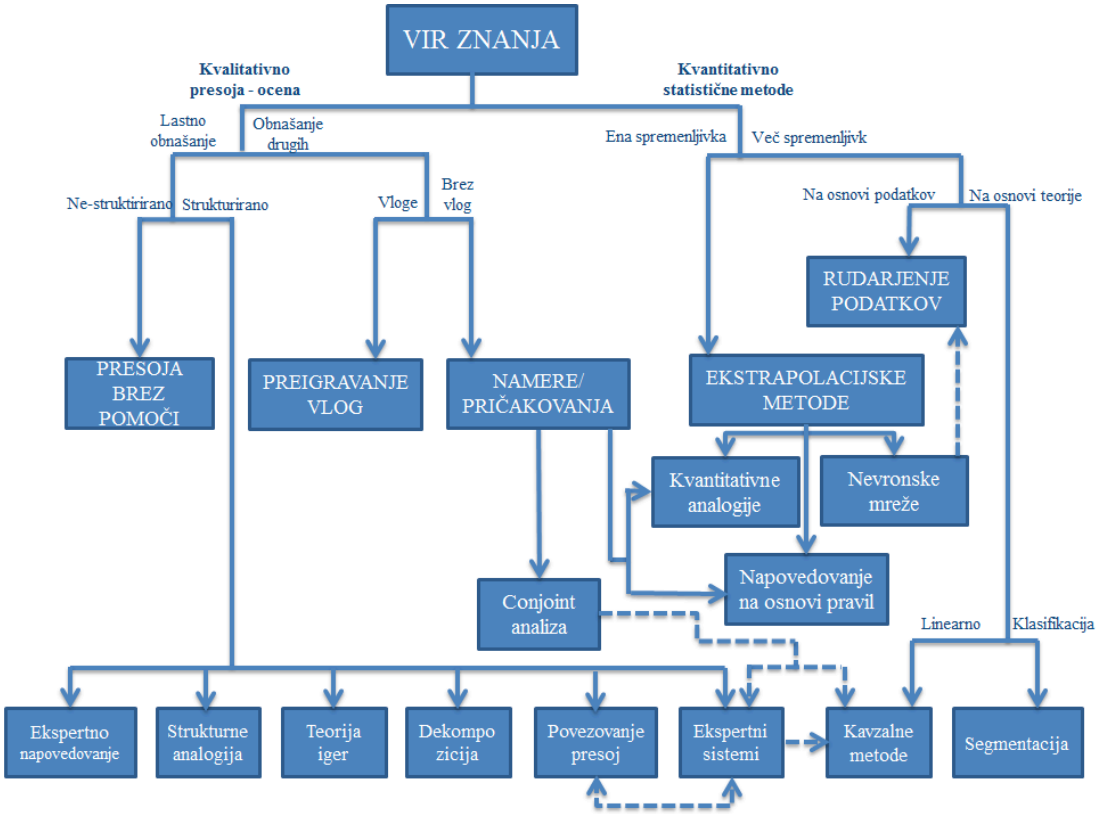
- **nestacionarne pojave**, ki se spreminjajo razmeroma hitro, nanje vpliva tako čas kot vrsta drugih faktorjev, vendar je mogoče vpliv le-teh ugotoviti in opredeliti. Ti pojavi so torej po svojem značaju eksogeni.

Metode za napovedovanje

Slika 8 prikazuje drevo metod napovedovanja [Ljubič]. **Kvalitativne metode** temeljijo na subjektivni presoji - oceni, **kvantitativne metode** pa zahtevajo relevantne numerične podatke o tem, kako in pod kakšnimi pogoji se je opazovani pojav obnašal v preteklosti.

Kvalitativne metode se delijo na tiste, ki predvidevajo lastno obnašanje nekoga in na tiste, kjer (nevtralni) strokovnjaki predvidevajo, kako se bodo obnašali drugi. Prve so uporabne, kadar odgovori temeljijo na znanju tistega, ki napoveduje, druge pa, če obstoji znanje o pričakovanem obnašanju drugih ljudi ali organizacij [Ljubič].

Nestrukturirane metode obravnavajo informacije neformalno, brez nekih posebnih pravil. Strukturirane metode analizirajo informacije s formalnimi postopki, pravila analize so opredeljena vnaprej in se uporabljajo dosledno [Ljubič].



Slika 8: Drevo metod napovedovanja [Ljubič]

Metode ekspertnega napovedovanja uporabljajo znanje strokovnjakov (vsaj dveh ali več), ki napovedujejo v osebnem stiku po strukturirani poti. Metod je več glede na delovno

okolje oz. razpoložjiv čas, omejitve, porazdelitev znanja, dostopnost strokovnjakov in njihove motivacije, zahteve po zaupnosti, itn [Ljubič].

Pri struktumih analogijah strokovnjaki iščejo ciljnemu pojavu podobne, analogne pojave in ugotavljajo podobnosti in razlike med njimi. Nato se primerja vsak rezultat analognih pojavov z možnim rezultatom ciljnega pojava. Rezultat (ali odločitev) najviše ocenjenega analognega pojava se uporabi kot napoved [Ljubič].

Teorija iger skuša razložiti, modelirati in predvideti obnašanje v socialnem okolju. Za to išče pravila obnahnja v dolohnih situacijah tako, da bi vsi udeleženci imeli čim večjo korist. Je uporabna predvsem za naknadno (posteriorno) analizo, manj za samo napovedovanje [Ljubič].

Analiza namer z anketiranjem (Conjoint analiza) omogoča preverjanje, kako različne situacije vplivajo na namere ljudi. Analitik na primer ljudem pokaže različne oblike nekega proizvoda in jih sprašuje, kaj bi kupili. Nato pa se za kvantificiranje razmerja namer ljudi in značilnosti proizvoda uporabljajo statistične analize, podobne regresiji (npr., kako bi uvedba ali opustitev neke značilnosti proizvoda vplivala na prodajo) [Ljubič].

Rudarjenje podatkov (Data Mining) je proces analize podatkov iz različnih vidikov, njihove kategorizacije in povzemanja uporabne informacije. Omogoča ugotavljanje razmerij med dejavniki, ki vplivajo na pojav in s tem tudi na napovedovanje [Ljubič].

Ekstrapolacijske metode predpostavljajo, da je gibanje nekega pojava endogeno in so podatki o njem zapisani kot **časovna vrsta**. Napoved vrednosti pojava v prihodnosti ugotovijo s statistično analizo podatkov o vrednosti obravnavanega pojava v preteklosti, nakar se izvede ekstrapolacija v prihodnost. Kadar je vzorec gibanja podatkov v časovni vrsti znan, kadar se ve, da v njej obstojijo komponente osnovne vrednosti, odstopanja (belega šuma), trenda, periodičnih ter sezonskih nihanj in vzorec velja daljše časovno obdobje, potem se uporabljajo metode za časovne vrste **s fiksnim vzorcem** (fiksne formule). Mednje sodijo različne **metode povprečij in eksponentnega glajenja**. Če pa vzorec gibanja ni znan, metode napovedovanja za časovne vrste **z odprtim vzorcem** najprej analizirajo časovno vrsto in ugotovijo vzorec gibanja, nato pa zanjo zgradijo

specifičen model oz. formulo za ekstrapolacijo. Sem spadata npr. **Fourierova analiza** in **Box-Jenkinsova metoda napovedovanja** [Ljubič].

Za napovedovanje je mogoče uporabiti tudi **nevronske mreže** (Neural networks), ki obravnavajo informacije podobno, kot človeški možgani. Po svoji logiki so dokaj blizu rudarjenju podatkov [Ljubič].

Napovedovanje na osnovi pravil kombinira poznavanje pojava in statistične tehnike in uporablja ekspertni sistem, da ekstrapolira časovne vrste. Največ značilnosti časovne vrste se določa avtomatsko, nekatere dejavnike pa ugotavljajo strokovnjaki. Posebej primerno je za ugotavljanje vzrokov za trend gibanja pojavov [Ljubič].

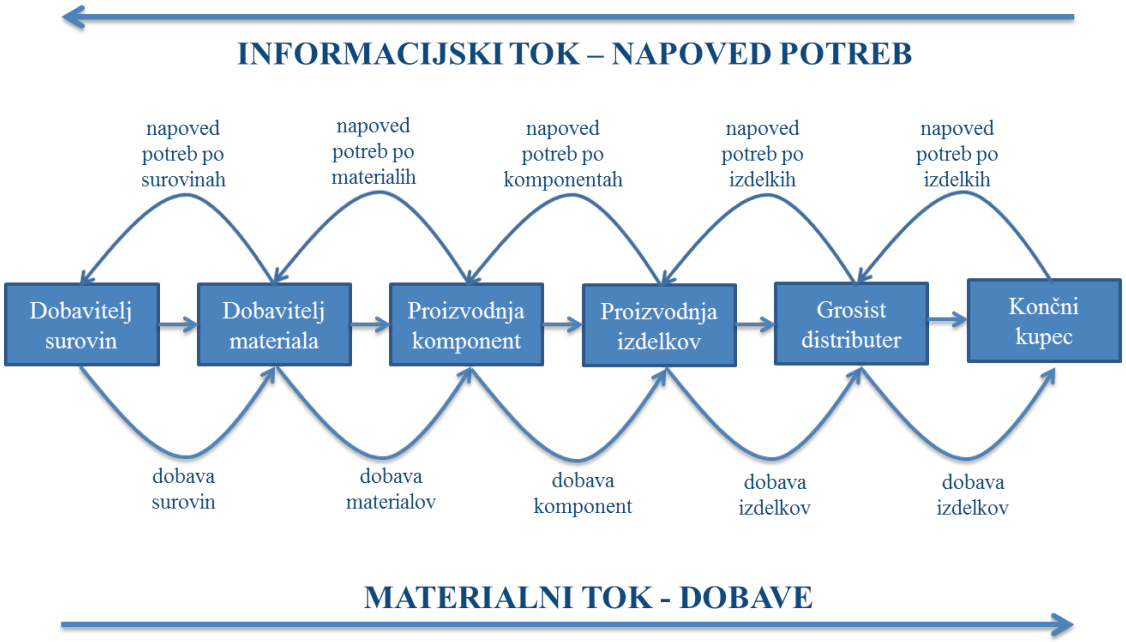
Linearni modeli na osnovi teorije obravnavajo probleme, ki imajo linearne parametre, katerih vpliv je enakovreden. Če pa so problemi sestavljeni iz večih podproblemov, ki na spremembe reagirajo različno, se jih rešuje z modeli klasifikacije. Le-ti proučujejo vsak podproblem posebej in jih nato povežejo navzkriž, da dobijo napoved [Ljubič].

Kavzalne (vzročno-posledične) oziroma **korelacijske metode** trdijo, da je gibanje nekega pojava v korelaciji z mnogimi faktorji, izmed katerih je eden lahko tudi čas. Imajo torej eksogen značaj. S primernimi statističnimi metodami je mogoče to korelacijo ugotoviti in ovrednotiti, to je določiti enačbo funkcije, ki idealizirano dovolj dobro predstavlja pojav. Ob predpostavki, da korelacijska razmerja veljajo dalj časa, se tako lahko napove vrednost opazovanega pojava za vse možne vrednosti faktorjev, s katerimi le-ta korelira. V to skupino sodi predvsem **regresijska analiza** [Ljubič].

Segmentacija razgrajuje celoten problem v drevesno strukturo podproblemov in išče informacije o relacijah med podproblemi. V to skupino spadajo na primer **input-output analiza**, **analiza skupin** (Cluster Analysis) in **analiza sistemske dinamike**, uporabne predvsem za dolgoročne ekonomske napovedi [Ljubič].

Napovedovanje v oskrbni verigi

Informacijski in materialni tok v oskrbovalni verigi prikazuje slika 9 [Ljubič].



Slika 9: Informacijski in materialni tok v oskrbovalni verigi [Ljubič]

lnformacijski tok napovedovanja v oskrbovalni verigi teče po verigi nazaj. Povpraševanje na trgu zaznava le detajlist oziroma končni kupec, ki tako lahko napove potrebe po proizvodih. Napoved posreduje grosistu - distributerju. Ta združi in morebiti prilagodi napovedi detajlistov in združeno napoved preda proizvajalcu (končnih) izdelkov. Le-ta spet združi prispele napovedi, jih morebiti prilagodi in pretvori v osnovni plan proizvodnje. lz plana proizvodnje ugotovi potrebe po komponentah in materialih ter jih kot napovedi preda proizvajalcem komponent oziroma dobaviteljem materiala in surovin. Materialni tok pri tem teče po verigi naprej, od dobaviteljev surovin in materialov preko proizvajalcev komponent in končnih izdelkov do distributerjev in detajlistov, ki izdelke prodajo končnemu kupcu [Ljubič].

Hitrost in zanesljivost tako informacijskega kot materialnega toka sta pri tem zelo pomembni. Prepočasen in prekinjen informacijski tok poleg ostalega povzroči tudi takoimenovan učinek biča (bullwhip effect) in destabilizira oskrbno verigo. Prepočasen materialni tok pa rezultira zlasti v neracionalnem poslovanju [Ljubič].

Planiranje je procesna funkcija, ki jo mora izvajati vsaka poslovna funkcija v takoimenovani verigi vrednosti - poslovnih sistemih tako fizične proizvodnje kot storitev, torej se v prenesenem pomenu tiče vsakogar v podjetju. Predvidevanje in napovedovanje kot izhodišče za planiranje imata tako velik pomen v vseh poslovnih funkcijah [Ljubič].

Ključna vprašanja v sistemu napovedovanja so, kdo sploh potrebuje napovedi? Načeloma vse poslovne funkcije, zlasti pa **marketing, prodaja, proizvodnja, logistika in finance**. Pri tem so pomembne predvsem **napovedi povpraševanja oziroma prodaje** in odtod izvedeni plani prodaje, marketinga, proizvodnje in materialnih potreb, ter finančni plani. Pomembno je tudi vprašanje, kaj naj se napoveduje? V proizvodnem okolju izdelave na zalogo se načeloma napoveduje potrebe po (končnih) izdelkih in iz njih izvede napovedi potreb po materialu. V okolju sestavljanja/montaže po naročilu, kjer potrebe po (končnih) izdelkih niso znane, se napoveduje potrebe po standardnih komponentah in prav tako se iz le-teh izvede napovedi potreb po materialu. Tudi v okoljih izdelave po naročilu, oziroma razvoja in izdelave po naročilu, potrebe po izdelkih niso znane vnaprej, zato se lahko napoveduje le potrebe po materialu [Ljubič].

Proces napovedovanja

Proces napovedovanja (slika 10) povezuje udeležence s tehnologijo in viri. Začne z izborom metode za napovedovanje, ter z ugotavljanjem, ali so na razpolago ustrezni statistični podatki o pojavu, ki se ga želi napovedovati. Če podatkov ni ali če so nezanesljivi, se bodo uporabile kvalitativne metode izkustvenega ocenjevanja [Ljubič].

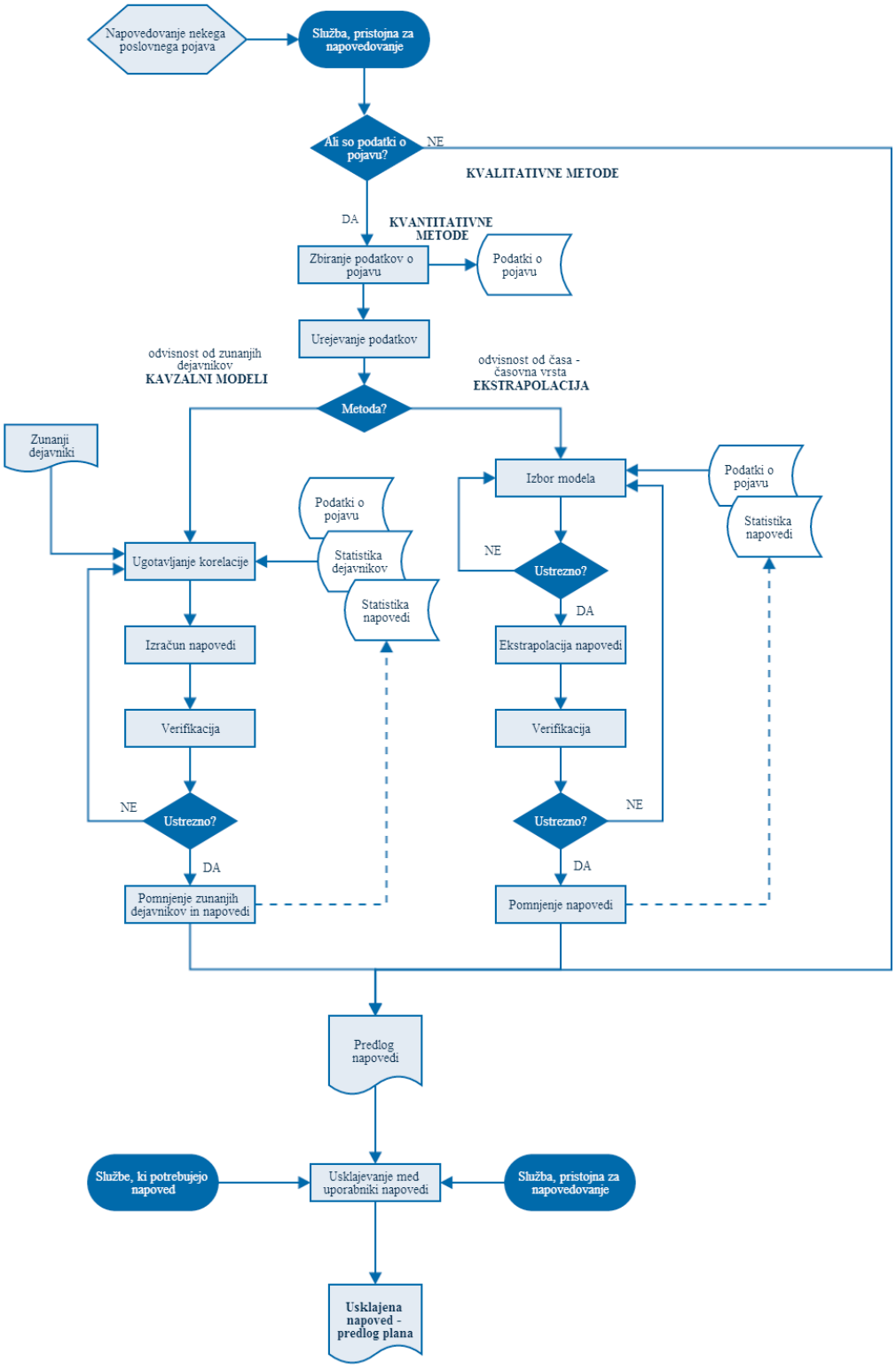
Kadar pa so razpoložljivi statistični podatki o pojavu, ki naj se ga napoveduje, se uporabljajo kvantitativne metode napovedovanja. Podatke se najprej preveri, analizira in uredi. ***Sledi odločitev, katera skupina metod se bo uporabila: ali ekstrapolacija, kadar je jasno, da je gibanje pojava dokaj stacionarno in odvisno zgolj od časa, ali kavzalne metode - korelacija, če je gibanje nestacionarno in poleg časa odvisno tudi od drugih zunanjih faktorjev*** [Ljubič].

Ekstrapolacija se začne z izborom modela izračuna, za katerega se ocenjuje, da bi lahko bil najustreznejši. Nadaljuje se s samim izračunom napovedi in verifikacijo rezultatov. Če

je evidentno, da rezultati niso realni, se izbere drug model izračuna in se račun ponovi. Ko so rezultati ustrezni, se shrani izhodiščne podatke in napoved [Ljubič].

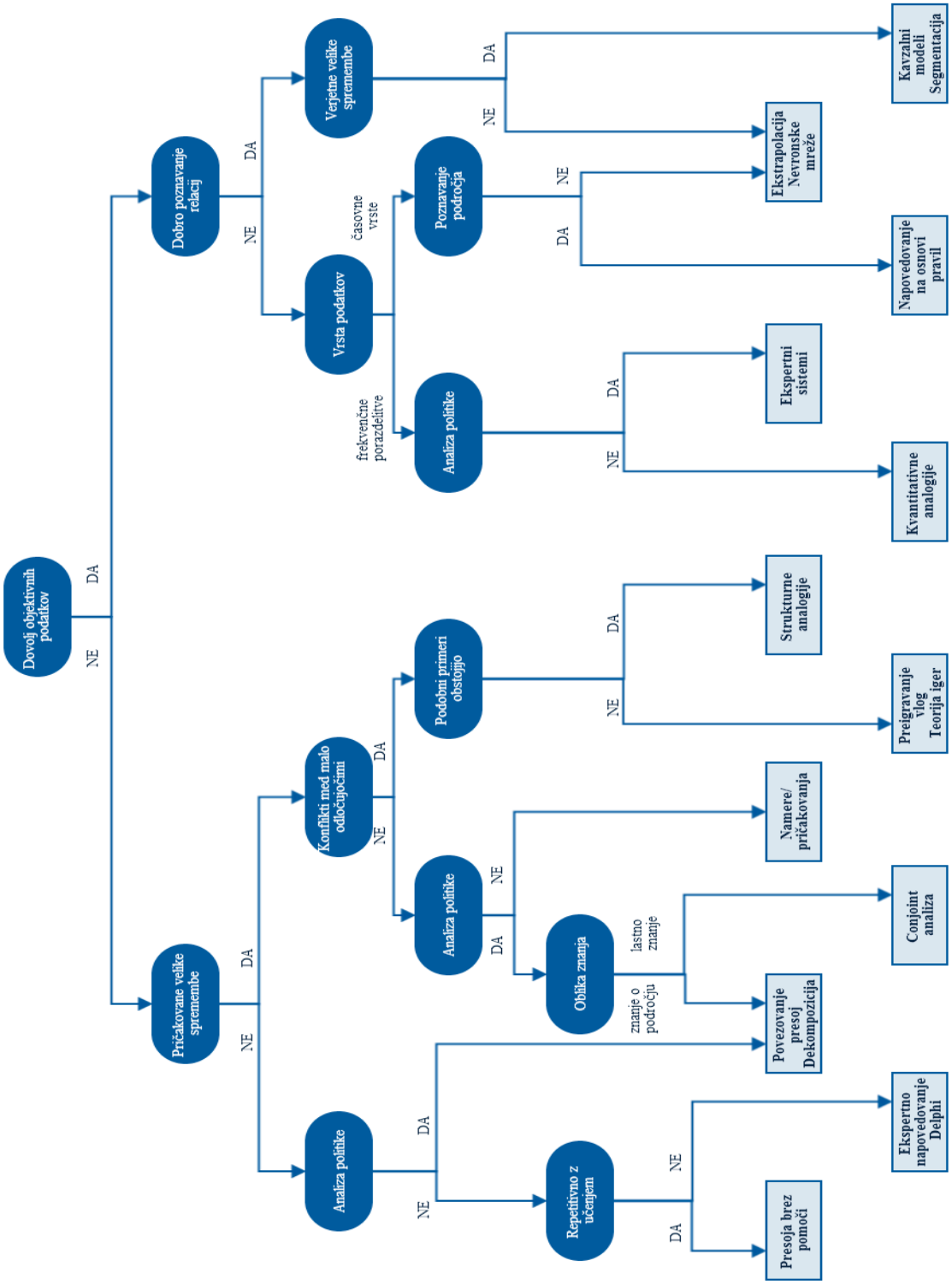
Če pa se bodo uporabile korelacijske metode, je treba najprej ugotoviti, s katerimi (zunanjimi) dejavniki pojav korelira, pri čemer morajo seveda biti na razpolago podatki o zunanjih dejavnikih. V nadaljevanju se odloči za nek model izračuna in se z izbranim modelom izračuna napoved. lzračunana napoved se preveri vizualno in/ali analitično. Tudi v tem primeru velja, da se takrat, kadar rezultati že na prvi pogled niso primerni, izbere nek drug model izračuna, s katerim se račun ponovi. To je pogosto potrebno ponoviti večkrat. Ko je napoved ustrezna, se jo shrani, prav tako se shranijo izhodiščni podatki o pojavu in o zunanjih faktorjih, ki na pojav vplivajo [Ljubič].

Predlog napovedi, dobljen po katerikoli poti, morajo uporabniki napovedi uskladiti in morebiti popraviti. To se izvede najpogosteje na usklajevalnih sestankih, kjer si uporabniki izmenjajo argumente za spremembo napovedi in določijo dokončno napoved. Le-ta pa je izhodišče za vse planske akcije vseh zainteresiranih poslovnih funkcij [Ljubič].



Slika 10: Proces napovedovanja [Ljubič, Armstrong]

Slika 11 prikazuje proces izbora ustrezne metode napovedovanja. Več podrobnosti o načinu izbiranja ustreznih metod si bralec pogleda v literaturi [Ljubič, Armstrong].



Slika 11: Proces izbora ustrezne metode napovedovanja [Ljubič, Armstrong]

Časovne vrste

Časovne vrste so sekvence (nizi) podatkov, izmerjenih v določenih časovnih trenutkih. Tipični primeri časovnih vrst so vrednost delnic na borzi, letni volumen pretoka na reki Nil, itn. Uporabljane so v statistiki, signalnem procesiranju, razpoznavanju vzorcev, ekonometriji, matematičnih financah, napovedovanju vremena ali potresov, sistemih za vodenje, astronomiji, komunikacijah, itn.
Analiza časovnih vrst združuje metode za analiziranje podatkov časovnih vrst z namenom, da bi se iz njih pridobilo kar največ koristnih informacij in karakteristik oz. statističnih značilnosti. **Napovedovanje časovnih vrst** pa je uporaba določenih modelov, da bi se tvorile napovedi prihodnjih vrednosti časovnih vrst na osnovi njihovih preteklih vrednosti.

Potrebno je ločiti regresijsko analizo od analize časovnih vrst. Pri regresijski analizi gre namreč zgolj za testiranje teorij, pri katerih naj bi trenutne vrednosti ene ali več neodvisnih časovnih vrst vplivale na trenutno vrednost neke druge časovne vrste.

Stohastični modeli časovnih vrst v splošnem upoštevajo dejstvo, da so meritve, ki so časovno bliže skupaj, bolj povezane med seboj, kot pa meritve, ki so časovno daleč vsakasebi.

Metode za analizo časovnih vrst lahko delimo na dve skupini:

- Metode v frekvenčnem prostoru, ter
- Metode v časovnem prostoru.

Metode v frekvenčnem prostoru vključujejo npr. **spektralno ali valčno analizo** (Wavelet Analysis), metode v časovnem prostoru pa npr. **avtokorelacijsko in križnokorelacijsko analizo**.

Metode za analizo časovnih vrst lahko delimo tudi na **parametrične in neparametrične metode**. Parametrične metode predpostavljajo, da ima obravnavan **stacionarni stohastični proces** določeno strukturo, ki je lahko opisana z majhnim številom parametrov (npr. pri **avtoregresivnem modelu** ali **modelu s premikajočim se**

povprečjem), ki jih je potrebno oceniti. V nasprotju s tem se pri neparametričnih metodah ocenjuje **kovarianca ali spekter procesa**, ne da bi se predpostavila določena struktura zanj.

Nenazadnje lahko metode za analizo časovnih vrst delimo tudi na linearne ali nelinearne, ter univariantne ali multivariantne.

Vloga statistike pri analizi in napovedovanju časovnih vrst

Inferenčna statistika igra pomembno vlogo pri statistični analizi časovnih vrst, katere glavni namen je uporabiti različne statistične teste za potrebe kvantitativnega opisa vseh naključnih procesov, ki tvorijo določeno sekvenco opazovanj.

Tako lahko vpeljemo teste za testiranje normalnosti časovne vrste, homogenosti, stacionarnosti in invertibilnosti, periodičnosti in sezonskosti, testiranje prisotnosti trenda, testiranje kvalitete in stabilnosti modela časovne vrste, testiranje signifikantnosti njegovih parametrov, in podobno.

Dejstvo je, da preprosto ne moremo izvesti kvalitetne konstrukcije modelov časovnih vrst, če pri tem ne upoštevamo dosledno vseh potrebnih statističnih testov, ki nam povedo več o naravi opazovanih časovnih vrst. Več o različnih oblikah uporabe statističnih testov pri načrtovanju modelov časovnih vrst si lahko bralec pogleda v ustrezni literaturi.

2 KRATEK PREGLED TEORIJE VERJETNOSTI

Matematična teorija verjetnosti nam daje osnovna orodja za konstrukcijo in analizo matematičnih modelov, ki opisujejo naključne fenomene. Pri študiju tovrstnih fenomenov pa se srečujemo z eksperimenti, pri katerih izida ne moremo napovedati vnaprej [Soong]. Verjetnostni račun se torej ukvarja z zakonitostmi naključnih izidov pri ponovitvah poskusov, ki potekajo pod enakimi ali vsaj zelo podobnimi pogoji.

Verjetnostni račun se v zadnjih desetletjih vedno krepkeje uveljavlja ne le kot samostojna panoga teoretične matematike, pač pa tudi kot zelo uspešen sklop raziskovalnih metod na številnih drugih znanstvenih področjih, kot npr. pri fiziki, astronomiji, biologiji, ekonomiji, psihologiji, inženirskih znanostih, itn. Uporabnost verjetnostnega izračuna izvira zlasti iz njegove povezave s statistiko, katere teoretične osnove temeljijo na zakonih verjetnostnega računa [Vadnal].

Koncept naključnosti so poznali že Egipčani in Grki, pri čemer so izide pojasnjevali z voljo bogov. Leta 1662 je plemič Chevalier de Mere zastavil matematiku Pascalu vprašanje, zakaj določene stave prinašajo dobiček, druge pa ne. Pascal se je o tem začel dopisovati s Fermatom in iz tega so nastali začetki verjetnostnega računa. Istega leta je Anglež John Graunt sestavil na osnovi podatkov prve zavarovalniške tabele. Teorijo verjetnosti kot uporabno vedo pa je utrdil Bernoulli v začetku 18 stoletja. Leta 1865 Mendel uporabi verjetnostno analizo pri razlagi dednosti, konec 19. stoletja pa teorija verjetnosti prodre v fiziko (začetek statistične fizike). V 20. stoletju se teorija verjetnosti razširi praktično na vsa področja znanosti in tehnike [Dragan 2].

V naravi se srečujemo z dvema tipoma dogodkov [Dragan 2]:

- deterministični in
- naključni.

Razlika je seveda v tem, da prve lahko točno predvidimo, drugim pa ne moremo vnaprej napovedati izida. Poglejmo dva primera determinističnih dogodkov:

- Primer 1: Če vodo ohladimo na -5 stopinj, se bo čez nekaj časa ustvaril led.

- Primer 2: Če kroglo spustimo z višine 1m od tal, bo ta padla na tla v času 0.45s.

Oba dogodka sta deterministična. To pomeni, da pri ponovitvi poskusa natančno vemo, kaj se bo zgodilo. Poglejmo si še tri primere naključnih dogodkov:

- Primer 3: Kolikšna bo vrednost delnice Mercatorja čez 1 mesec?
- Primer 4: Kdo bo letošnji zmagovalec lige prvakov?
- Primer 5: Vplačam srečko za loterijo. Ali bom zadel?

V teh primerih gre za naključne dogodke, kjer je njihov izid nemogoče povsem zanesljivo napovedati. Kljub temu pa življenje terja od nas odločitve, četudi ne moremo vedno jasno opredeliti njihovih posledic. Na primer se moramo odločiti:

- Ali naj kupimo delnice Mercatorja ali Krke?
- Ali je bolje naložiti denar v nepremičnino?
- Direktorja logističnega podjetja zanima povračilna doba investicije, da se lažje odloči, ali iti v investicijo ali ne.

Čeprav pride intuicija vedno prav (seveda tistim, ki jo imajo), je za učinkovito odločanje zelo koristno, če premoremo kvantitativno oceno možnosti (verjetnosti) za posamezne opcije. Orodja za to pa najdemo v teoriji verjetnosti [Dragan 2].

V svetu kompleksnih dinamičnih procesov je negotovost osrednja lastnost. Za sistematično računanje z negotovostjo pa potrebujemo [Dragan 2]:

- konsistenten in
- logičen

sistem razmišljanja. Tega pa vsebuje teorija verjetnosti, ki nam s pomočjo matematičnega aparata pomaga ocenjevati stopnjo negotovosti v naključnih procesih.

Pojem verjetnosti poznamo že iz vsakdanjega življenja [Vadnal]. Pogosto pravimo o kakšnem dogodku, da je zelo verjeten, da je malo verjeten, da je neverjeten itn. Npr. pri metu kocke pravimo, da je verjetno, da vržemo šest pik. Pri loteriji je zelo malo verjetno,

da zadanemo glavni dobitek, bolj verjetno je, da zadanemo kakšen manjši dobitek, najverjetneje pa je, da ostanemo praznik rok. Vsebina takšnih izjav je prej ko slej nedoločena in marsikdaj odvisna tudi od človekovega razpoloženja.

S takšnimi in podobnimi izjavami se ukvarja verjetnostni račun [Vadnal]. Pri njem srečujemo pojme, ki ji deloma poznamo iz vsakdanjega življenja, vendar pa ti običajno vsebujejo premalo opredeljeno vsebino v primerjavi z ustreznimi pojmi verjetnostnega računa.

Verjetnostni račun proučuje naključne pojave. To so pojavi, ki ob ponovljenih poskusih (pri enakih pogojih) rezultirajo v različnih izidih. Vendar, če število ponovitev postane veliko, se pokažejo določene konsistentne lastnosti.

Kot vemo iz osnov verjetnostnega računa, so temeljni pojmi, ki se pojavijo pri verjetnostnem računu, naslednji [Dragan 2, Vadnal]:

- poskus,
- dogodek, ter
- verjetnost dogodka.

Od bralca tega dela se pričakuje, da bo seznanjen z osnovami verjetnostnega računa, kot npr., z lastnostmi verjetnosti, pogojno verjetnostjo, zakonom totalnih verjetnosti in Bayesovim pravilom (glej npr. vire [Dragan 2, Benjamin, Bertsekas, Grinstead, Usenik 2, Vadnal]). Zato bomo v nadaljevanju preleteli zgolj nekatere osnovne pojme teorije verjetnosti, potrebne pri razumevanju kasnejše snovi.

2.1 Diskretne naključne spremenljivke

Pri nekaterih poskusih so lahko izidi tudi števila. Npr., pri kockanju je izid število pik, torej neko naravno število med 1 in 6. Pri *n*-kratni ponovitvi poskusa je izid frekvenca danega dogodka A, torej katerokoli celo število od 0 do n. Pri streljanju v tarčo pa imamo lahko za izid razdaljo med lego zadetka in sredino tarče, se pravi kakšno nenegativno realno število, seveda ne neomejeno veliko.

Na takšne poskuse lahko gledamo, kot da jim je prirejena neka količina, ki ima lahko različne vrednosti. Katero vrednost pa ta količina zavzame pri dani ponovitvi poskusa, pa je odvisno od naključja (slučaja). Zato imenujemo takšne količine naključne (slučajne) spremenljivke [Jamnik 1].

Pri slučajnih spremenljivkah moramo poznati dvoje [Usenik 2]:

- zalogo vrednosti,
- predpis, ki določa verjetnosti, da naključna spremenljivka zavzame določeno vrednost (porazdelitven zakon).

Slučajna spremenljivka je torej količina, ki ima svojo vrednost odvisno od slučaja in je natanko določena s svojo zalogo vrednosti ter s svojim porazdelitvenim zakonom [Usenik 2].

Slučajne spremenljivke običajno označujemo z *X, Y, Z*, itn, ali pa tudi $X_1, X_2, X_3, \ldots$, pripadajoče vrednosti slučajnih spremenljivk pa označujemo z malimi črkami *x, y, z*, itn, oz. $x_1, x_2, x_3, \ldots$.

Dogodek, da slučajna spremenljivka zavzame določeno vrednost iz svoje zaloge, označujemo z $(X = x)$, $(x_1 < X < x_2)$, $(X \geq x)$, itn [Usenik 2].

Slučajne spremenljivke v splošnem delimo na:

- diskretne slučajne spremenljivke, ter
- zvezne slučajne spremenljivke.

Diskretne slučajne spremenljivke so tiste spremenljivke, katerih zaloga vrednosti je neko končno ali neskončno zaporedje. Pri zveznih slučajnih spremenljivkah pa zalogo vrednosti ne moremo numerirati z naravnimi števili, pač pa je zaloga vrednosti določen končen ali neskončen interval realnih števil [Usenik 2].

Poglejmo si nekaj pojmov [Jamnik 1]:

- $(X = x)$ je dogodek, da zavzame naključna spremenljivka *X* vrednost *x*,
- $P(x) = P(X = x)$ pa pomeni verjetnost tega dogodka,
- $(X > x)$ je dogodek, da zavzame naključna spremenljivka *X* vrednost, ki je večja od *x*,
- $P(X > x)$ pa je verjetnost tega dogodka. Itn...

Funkcijo $P(x) = P(X = x)$ običajno imenujemo tudi funkcija porazdelitve verjetnosti [Dragan 2]. V terminologiji v različni literaturi se srečamo tudi z izrazom verjetnostna funkcija, s katero običajno podajamo splošno obliko porazdelitvenega zakona diskretnih slučajnih spremenljivk [Usenik 2]. V povezavi s tem si poglejmo naslednjo definicijo [Usenik 2]:

Verjetnostna funkcija p_k *diskretne slučajne spremenljivke X je funkcija, ki ima pri vsakem mogočem k svojo vrednost enako verjetnosti dogodka* $(X = x_k)$*, torej velja:* $p_k = P(X = x_k) = P(x_k)$*. Pri tem k preteče vse tiste cele vrednosti, za katere spada* x_k *v zalogo vrednosti diskretne slučajne spremenljivke X.*

Zaradi preglednosti običajno zapišemo pri diskretnih naključnih spremenljivkah zalogo vrednosti in porazdelitveni zakon v obliki takoimenovane verjetnostne sheme. Npr., za naključno spremenljivko *X* zapišemo shemo takole:

$$X: \begin{pmatrix} x_1 & x_2 & x_3 \ldots & x_n \\ p_1 & p_2 & p_3 \ldots & p_n \end{pmatrix} \tag{2.1}$$

Seveda pri tem velja, da je [Jamnik 1]:

$$p_1 + p_2 + \ldots p_n = 1 \tag{2.2}$$

saj sestavljajo $x_1, x_2, x_3, \ldots x_n$ popoln sistem vrednosti.

Poglejmo si dva primera, odkoder bo popolnoma jasna razlika med diskretnimi in zveznimi naključnimi spremenljivkami.

Primer 2.1.:

Enkrat vržemo dva kovanca. Naj diskretna spremenljivka X označuje število padlih grbov. Poiščite verjetnostno shemo.

X je diskretna naključna diskretna spremenljivka, ki lahko zavzame vrednosti 0,1,2 z verjetnostmi:

$$P[X=0] = P[(\check{S},\check{S})] = \frac{1}{4}$$
$$P[X=1] = P[(\check{S},G),(G,\check{S})] = \frac{2}{4}$$
$$P[X=2] = P[(G,G)] = \frac{1}{4}$$

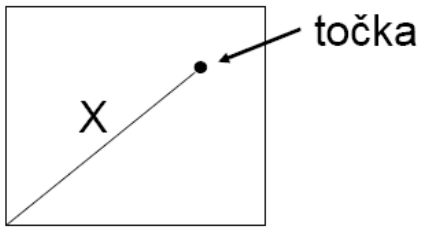
kjer $\check{S}$ pomeni padlo številko, G pa padli grb.

Verjetnostna shema torej je:

$$X: \begin{pmatrix} 0 & 1 & 2 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{pmatrix}$$

Primer 2.2.:

Znotraj kvadrata stranice 1 naključno izberemo zočko. Oddaljenost točke od levega spodnjega roba je naključna spremenljivka X. Ker X lahko zavzame katerokoli realno vrednost med 0 in $\sqrt{2}$, je X zvezna naključna spremenljivka (glej sliko 12).



Slika 12: Ilustracija primera zvezne naključne spremenljivke X

2.1.1 Primeri diskretnih porazdelitev

Porazdelitveni zakon diskretne naključne spremenljivke se imenuje na kratko diskretna porazdelitev [Jamnik 1]. Naštejmo nekatere važnejše diskretne porazdelitve [Jamnik 1]:

- Binomska porazdelitev,
- Enakomerna porazdelitev, ter
- Poissonova porazdelitev.

Diskretna naključna spremenljivka je porazdeljena **enakomerno**, če sestavljajo njeno zalogo vrednosti števila $x_1, x_2, x_3, ...x_n$ (vsa različna med seboj) in za vsak k od 1 do n velja [Jamnik 1]:

$$P(X = x_k) = \frac{1}{n} \tag{2.3}$$

Zgled za enakomerno diskretno porazdelitev je npr. število pik pri kockanju [Jamnik 1].

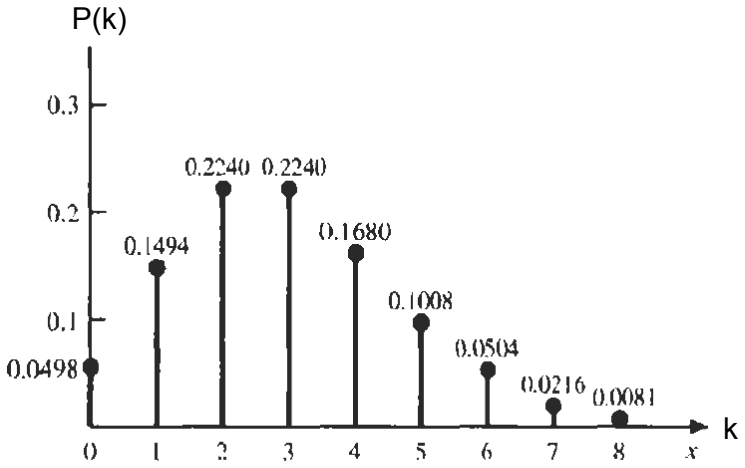
Za **binomsko porazdelitev** je značilno, da je porazdeljena naključna spremenljivka X po takoimenovanem binomskem zakonu, pri čemer ima zalogo vrednosti $\{0,1,2,...,n\}$. Po binomskem zakonu je porazdeljena frekvenca dogodka A v n ponovitvah poskusa, v katerem ima A verjetnost p. Verjetnostna funkcija se glasi ($0 < p < 1$):

$$P(X = k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}, \quad k = 0,1,2,...,n \tag{2.4}$$

Naključna spremenljivka, porazdeljena po **Poissonovem zakonu**, ima zalogo vrednosti $\{0,1,2,...\}$ in verjetnostno funkcijo [Jamnik 1]:

$$p_k = P(X = k) = \frac{\lambda^k}{k!} \cdot e^{-\lambda}, \quad k = 0,1,2,... \tag{2.5}$$

Primer Poissonove porazdelitve je za parameter $\lambda = 3$ ilustriran na sliki 13.



Slika 13: Primer Poissonove porazdelitve za $\lambda = 3$

Seveda poleg naštetih diskretnih porazdelitev obstajajo še tudi druge porazdelitve, kot npr.:

- Pascalova porazdelitev,
- Geometrijska porazdelitev,
- Hipergeometrijska porazdelitev, itn.

Preden gremo na natančnejšo obravnavo binomske porazdelitve in z njo tesno povezanega Bernoullijevega poskusa, pa si v nadaljevanju oglejmo še nekaj lastnosti takoimenovane kumulativne porazdelitve.

2.2 Kumulativna porazdelitev verjetnosti

Funkcija kumulativne porazdelitve verjetnosti *F(x)* naključne spremenljivke *X* je definirana na naslednji način [Dragan 2]:

$$F(x) = P[X \le x], \quad -\infty < x < \infty \tag{2.6}$$

Funkcija (2.6) je zanimiva zato, ker se da tudi iz nje razbrati veliko informacije o naključnem eksperimentu. Poglejmo si nekaj njenih lastnosti [Hsu]:

$$
\begin{aligned}
&1.\quad 0 \le F(x) \le 1 \\
&2.\quad F(x_1) \le F(x_2), \qquad \check{c}e \;\; x_1 \le x_2 \\
&3.\quad F(-\infty) = 0 \\
&4.\quad F(\infty) = 1
\end{aligned}
\tag{2.7}
$$

Velja tudi naslednje [Hsu]:

$$
\begin{aligned}
&F(b) = P(X \le b) \\
&F(a) = P(X \le a) \\
&F(b) - F(a) = P(X \le b) - P(X \le a) = \\
&= \underbrace{P(X \le a)}_{P(X \le b)} + P(a \le X \le b) - P(X \le a) = P(a \le X \le b)
\end{aligned}
\tag{2.8}
$$

torej :

$$F(b) - F(a) = P(a \le X \le b)$$

Prav tako še velja:

$$
\begin{aligned}
&F(a) = P(X \le a) \\
&P(X \le a) + P(X > a) = 1 \\
&P(X > a) = 1 - P(X \le a) \\
&P(X > a) = 1 - F(a)
\end{aligned}
\tag{2.9}
$$

Zveza med kumulativno porazdelitvijo verjetnosti in verjetnostno funkcijo je pri diskretnih naključnih spremenljivkah naslednja:

$$F(x) = P(X \le x) = \sum_{x_k \le x} P(x_k) \tag{2.10}$$

Princip kumulativne porazdelitve verjetnosti osvetlimo na naslednjem primeru [Dragan 2]:

Primer 2.3.: *Enkrat vržemo dve kocki. Pri tem naključna spremenljivka X predstavlja vsoto padlih številk. Prostor vseh možnih dogodkov in pripadajoče vrednosti X so ilustrirane na sliki 14. Izračunajte verjetnosti za posamezne izide spremenljivke X, ter narišite funkciji porazdelitve verjetnosti in kumulativne porazdelitve verjetnosti.*

(1,1) 2	(1,2) 3	(1,3) 4	(1,4) 5	(1,5) 6	(1,6) 7
(2,1) 3	(2,2) 4	(2,3) 5	(2,4) 6	(2,5) 7	(2,6) 8
(3,1) 4	(3,2) 5	(3,3) 6	(3,4) 7	(3,5) 8	(3,6) 9
(4,1) 5	(4,2) 6	(4,3) 7	(4,4) 8	(4,5) 9	(4,6) 10
(5,1) 6	(5,2) 7	(5,3) 8	(5,4) 9	(5,5) 10	(5,6) 11
(6,1) 7	(6,2) 8	(6,3) 9	(6,4) 10	(6,5) 11	(6,6) 12

Slika 14: Prostor vseh možnih dogodkov in pripadajoče vrednosti naključne spremenljivke X pri enkratnem metu dveh kock

Najprej izračunajmo verjetnosti za posamezne izide spremenljivke *X*:

$$
\begin{aligned}
&P(1) = P[X=1] = 0 \\
&P(2) = P[X=2] = P\left[\{(1,1)\}\right] = \frac{1}{36} \\
&P(3) = P[X=3] = P\left[\{(1,2),(2,1)\}\right] = \frac{2}{36} \\
&P(4) = P[X=4] = P\left[\{(1,3),(2,2),(3,1)\}\right] = \frac{3}{36} \\
&P(5) = \frac{4}{36}, P(6) = \frac{5}{36}, P(7) = \frac{6}{36}, P(8) = \frac{5}{36}, P(9) = \frac{4}{36} \\
&P(10) = \frac{3}{36}, P(11) = \frac{2}{36}, P(12) = \frac{1}{36} \\
&P(i) = P[X=i] = 0, \qquad i = 13,14,...
\end{aligned}
\tag{2.11}
$$

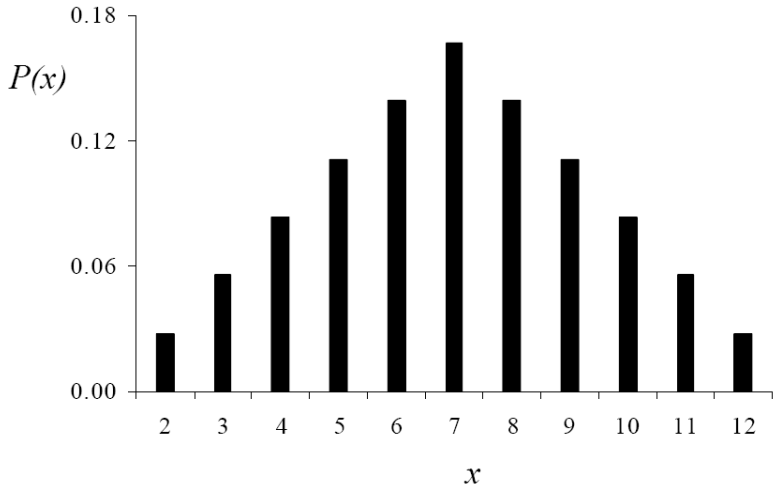
Nato na osnovi izračunanih izrazov v (2.11) narišemo funkcijo porazdelitve verjetnosti, kar prikazuje slika 15. Vrednosti funkcije kumulativne porazdelitve verjetnosti pa izračunamo s pomočjo izraza (2.10). Tako dobimo:

$$F(x)=\begin{cases} 0 & x<2 \\ 0+\frac{1}{36} & 2\le x<3 \\ 0+\frac{1}{36}+\frac{2}{36} & 3\le x<4 \\ 0+\frac{1}{36}+\frac{2}{36}+\frac{3}{36} & 4\le x<5 \\ \quad\quad ... & \end{cases} \tag{2.12}$$

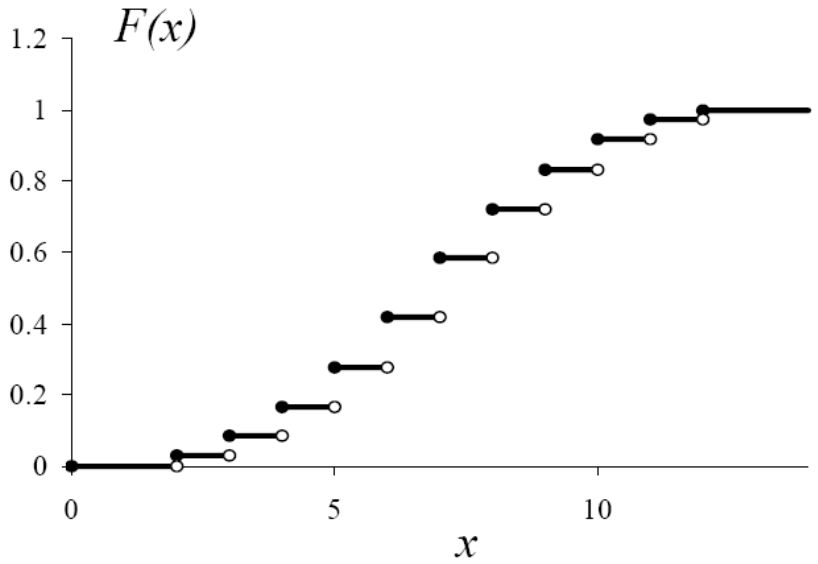
oziroma, če izračunamo vse vrednosti, dobimo:

$$F(x)=\begin{cases} 0 & x<2 \\ \frac{1}{36} & 2\le x<3 \\ \frac{3}{36} & 3\le x<4 \\ \frac{6}{36} & 4\le x<5 \\ \frac{10}{36} & 5\le x<6 \\ \frac{15}{36} & 6\le x<7 \\ \frac{21}{36} & 7\le x<8 \\ \frac{26}{36} & 8\le x<9 \\ \frac{30}{36} & 9\le x<10 \\ \frac{33}{36} & 10\le x<11 \\ \frac{35}{36} & 11\le x<12 \\ 1 & 12\le x \end{cases} \tag{2.13}$$

Nazadnje na osnovi izračunanega izraza (2.13) narišemo še funkcijo kumulativne porazdelitve verjetnosti, kar prikazuje slika 16.



Slika 15: Funkcija porazdelitve verjetnosti pri enkratnem metu dveh kock



Slika 16: Funkcija kumulativne porazdelitve verjetnosti pri enkratnem metu dveh kock

2.3 Bernoullijeva porazdelitev

Predpostavimo, da imamo Bernoullijev poskus, pri katerem sta možna dva izida (S - success oz. uspeh, ter F - failure oz. neuspeh):

- ugoden izid (S), in
- neugoden izid (F).

Tipičen primer takšnega poskusa je proizvodnja, kjer je na koncu izdelek dober ali slab (izmet). Predpostavimo tudi, da je verjetnost za ugoden izid p, za neugoden izid pa $q = 1 - p$. Potem lahko definiramo takoimenovano Bernoullijevo naključno spremenljivko X, s katero opišemo izid Bernoullijevega poskusa [Dragan 2]:

$$X = \begin{cases} 0 & \text{če je izid F (neuspešen)} \\ 1 & \text{če je izid S (uspešen)} \end{cases} \tag{2.14}$$

Porazdelitev verjetnosti za takšno spremenljivko pa je definirana na naslednji način [Hsu, Dragan 2] (Bernoullijeva porazdelitev):

$$P(k) = P(X = k) = p^k \cdot \underbrace{(1-p)}_{q}^{1-k}, \qquad k = 0,1 \tag{2.15}$$

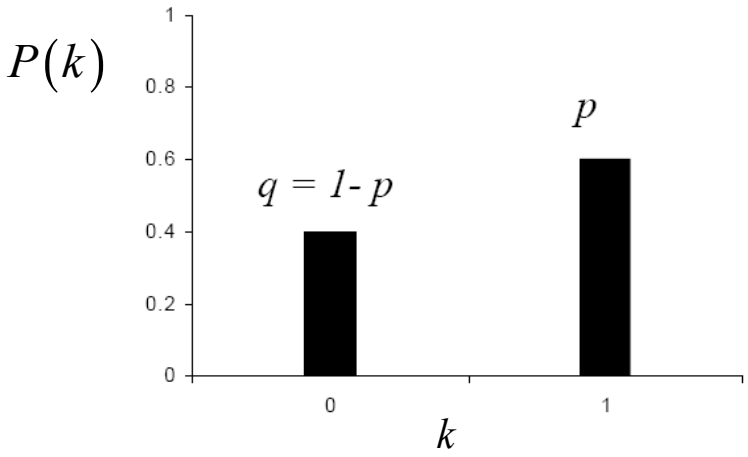
Izraz (2.15) lahko zapišemo tudi drugače:

$$\begin{aligned} P(0) &= P(X = 0) = p^0 \cdot \underbrace{(1-p)}_{q}^{1-0} = q \\ P(1) &= P(X = 1) = p^1 \cdot \underbrace{(1-p)}_{q}^{1-1} = p \end{aligned} \tag{2.16}$$

Torej velja:

$$P(k) = P[X = k] = \begin{cases} q & k = 0 \\ p & k = 1 \end{cases} \tag{2.17}$$

Ilustracijo Bernoullijeve porazdelitve prikazuje slika 17.



Slika 17: Ilustracija Bernoullijev porazdelitve

2.4 Binomska porazdelitev

Dva poskusa imenujemo med seboj neodvisna, če je vsak dogodek iz enega poskusa neodvisen od kateregakoli dogodka v drugem poskusu [Jamnik 1]. Tudi za več poskusov velja podobno. Če so dogodki teh poskusov med seboj v celoti neodvisni, potem so tudi poskusi med seboj neodvisni. V zaporedju neodvisnih poskusov ni potrebno, da bi bili poskusi med seboj enaki. Preprosteje pa je, če so. Zaporedje enakih neodvisnih poskusov lahko imenujemo tudi ponavljanje istega poskusa. Tako zaporedje je npr. metanje kocke, serijska proizvodnja, itn [Jamnik 1].

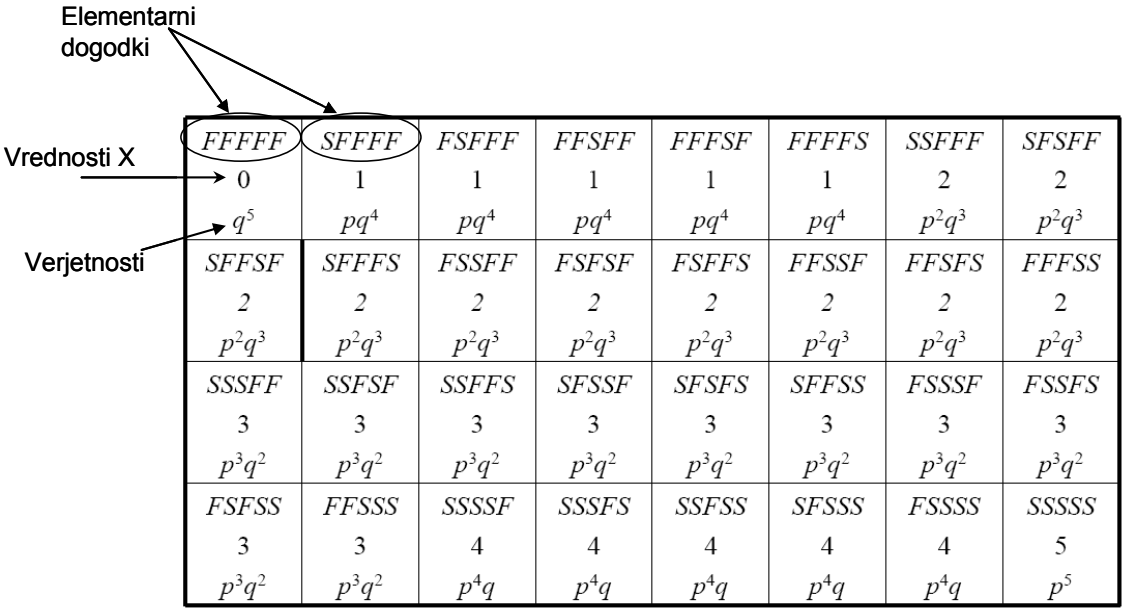
Med zaporedji enakih neodvisnih poskusov so še posebej zanimiva zaporedja Bernoullijevih poskusov, torej takšnih poskusov, kjer sta možna le dva izida (npr. met kovanca).

V nadaljevanju predpostavimo, da imamo opravka z Bernoullijevim poskusom, ki ga ponovimo *n*-krat, pri čemer je seveda vsaka ponovitev poskusa neodvisna od prejšnjih poskusov.

Naj bo naključna spremenljivka X število ugodnih izidov v seriji *n*-tih Bernoullijevih poskusov. Potem lahko ugotovimo, da so možne vrednosti X enake:

$$X \in \{0,1,2,...,n\} \tag{2.18}$$

Če je npr. $n = 5$, potem lahko prostor elementarnih dogodkov, pripadajoče vrednosti naključne spremenljivke X in pripadajočo verjetnost elementarnih dogodkov podamo s sliko 18 (npr. imamo opravka s petkratnim metanjem kovanca, S = uspeh, F = neuspeh) [Dragan 2].



FFFFF	*SFFFF*	*FSFFF*	*FFSFF*	*FFFSF*	*FFFFS*	*SSFFF*	*SFSFF*
0	1	1	1	1	1	2	2
q^5	pq^4	pq^4	pq^4	pq^4	pq^4	p^2q^3	p^2q^3
SFFSF	*SFFFS*	*FSSFF*	*FSFSF*	*FSFFS*	*FFSSF*	*FFSFS*	*FFFSS*
2	2	2	2	2	2	2	2
p^2q^3	p^2q^3	p^2q^3	p^2q^3	p^2q^3	p^2q^3	p^2q^3	p^2q^3
SSSFF	*SSFSF*	*SSFFS*	*SFSSF*	*SFSFS*	*SFFSS*	*FSSSF*	*FSSFS*
3	3	3	3	3	3	3	3
p^3q^2	p^3q^2	p^3q^2	p^3q^2	p^3q^2	p^3q^2	p^3q^2	p^3q^2
FSFSS	*FFSSS*	*SSSSF*	*SSSFS*	*SSFSS*	*SFSSS*	*FSSSS*	*SSSSS*
3	3	4	4	4	4	4	5
p^3q^2	p^3q^2	p^4q	p^4q	p^4q	p^4q	p^4q	p^5

Slika 18: Ilustracija prostora elementarnih dogodkov, pripadajoče vrednosti naključne spremenljivke X in pripadajoče verjetnosti elementarnih dogodkov pri petkratni ponovitvi Bernoullijevega poskusa (S = uspel poskus z verjetnostjo p, F = neuspel poskus z verjetnostjo q)

Iz slike 18 je razvidno, da X vsakič zavzame takšno vrednost, kolikor je bilo uspešnih poskusov S v posameznem elementarnem dogodku (petkratna ponovitev poskusa). Posamezne verjetnosti petkratne ponovitve poskusa so pa enake produktu verjetnosti posameznega poskusa (ki, kot vemo, zavzamejo p v primeru uspeha S in q v primeru neuspeha F). Tako je npr. $P(FFFFF) = q \cdot q \cdot q \cdot q \cdot q = q^5$, podobno je $P(SFFFF) = p \cdot q \cdot q \cdot q \cdot q = p \cdot q^4$, itn.

Na osnovi slike 18 lahko izračunamo verjetnosti za nastop posameznih vrednosti naključne spremenljivke X. Tako dobimo:

$$
\begin{aligned}
&P(0) = P(X=0) = P(FFFFF) = q^5 \\
&P(1) = P(X=1) = \\
&= P(SFFFF) + P(FSFFF) + P(FFSFF) + P(FFFSF) + P(FFFFS) = \\
&= p\cdot q^4 + p\cdot q^4 + p\cdot q^4 + p\cdot q^4 + p\cdot q^4 = 5\cdot p\cdot q^4 \\
&P(2) = P(X=2) = \ldots = 10\cdot p^2\cdot q^3 \\
&P(3) = P(X=3) = \ldots = 10\cdot p^3\cdot q^2 \\
&P(4) = P(X=4) = \ldots = 5\cdot p^4\cdot q \\
&P(5) = P(X=5) = P(SSSSS) = p^5
\end{aligned}
\tag{2.19}
$$

Izračune v izrazu (2.19) lahko tudi bolj pregledno zapišemo v obliki verjetnostne sheme, prikazane na sliki 19.

x	0	1	2	3	4	5
$P(x) = P[X = x]$	q^5	$5pq^4$	$10p^2q^3$	$10p^3q^2$	$5p^4q$	p^5

Slika 19: Verjetnostna shema za naključno spremenljivko X pri petkratni ponovitvi Bernoullijevega poskusa

Na podoben način, kot smo sklepali pri n = 5, torej petkratni ponovitvi Bernoullijevega poskusa, lahko sklepamo tudi za poljuben n. Tedaj je izid zaporedja Bernoullijevih poskusov neko poljubno zaporedje uspehov S in neuspehov F, ki ima dolžino n:

$$
SSFSFFSFFF\ldots FSSSFFSFSFFS \tag{2.20}
$$

Denimo je v tem zaporedju k uspehov in n-k neuspehov. Verjetnost, da je v zaporedju k uspehov, je:

$$
P(k\ uspehov\ S) = \underbrace{p\cdot p\cdot p\cdot p\cdot \ldots \cdot p}_{k-krat} = p^k \tag{2.21}
$$

Verjetnost, da je v zaporedju *n-k* neuspehov, je:

$$P\left(n-k\ neuspehov\ F\right)=\underbrace{q\cdot q\cdot q\cdot q\cdot \ldots \cdot q}_{(n-k)-krat}=q^{n-k} \tag{2.22}$$

Verjetnost, da se v seriji n poskusov zgodi *k* uspešnih poskusov in *n-k* neuspešnih poskusov, torej je:

$$P\left(k\ uspehov\ S\ in\ n-k\ neuspehov\ F\right)=\binom{n}{k}\cdot \underbrace{P\left(k\ uspehov\ S\right)}_{p^k}\cdot \underbrace{P\left(n-k\ neuspehov\ F\right)}_{q^{n-k}} \tag{2.23}$$

kjer smo z $\binom{n}{k}$ še upoštevali število vseh možnih zaporedij, kjer se zgodi *k* uspehov in *n-k* neuspehov [Jamnik 1, Dragan 2].

Na osnovi zgornjih izrazov tako pridemo do izraza za Binomsko diskretno porazdelitev naključne spremenljivke *X*, ki jo izrazimo z naslednjo verjetnostno funkcijo (seveda je *p + q = 1*) [Jamnik 1, Dragan 2]:

$$P\left(k\right)=P\left(X=k\right)=\binom{n}{k}\cdot p^k\cdot \underbrace{\left(1-p\right)}_{q}^{n-k},\quad k=0,1,2,\ldots,n \tag{2.24}$$

Princip binomske porazdelitve bomo osvetlili na naslednjem primeru:

Kovanec mečemo 7-krat, t.j. n = 7. Naj bo X naključna spremenljivka, ki pove, kolikokrat je padlo pismo P. Poiščite binomsko porazdelitev naključne spremenljivke X!

Ker sta pri vsakem poskusu možna le dva izida P (pismo) ali Š (številka), ima *X* očitno binomsko porazdelitev, kjer sedemkrat ponovimo Bernoullijev poskus. Seveda velja:

$$\begin{aligned} p&=P\left(P\right)=\frac{1}{2}\\ q&=P\left(\check{S}\right)=1-p=1-\frac{1}{2}=\frac{1}{2}\end{aligned} \tag{2.25}$$

Verjetnosti za nastop posameznih vrednosti naključne spremenljivke X izračunamo s pomočjo izraza (2.24):

$$P(k)=P(X=k)=\binom{7}{k}\cdot p^{k}\cdot(1-p)^{7-k},\quad k=0,1,2,...,7$$

$$P(k)=P(X=k)=\binom{7}{k}\cdot\left(\frac{1}{2}\right)^{k}\cdot\left(\frac{1}{2}\right)^{7-k},\quad k=0,1,2,...,7 \qquad (2.26)$$

$$P(k)=P(X=k)=\binom{7}{k}\cdot\left(\frac{1}{2}\right)^{7},\quad k=0,1,2,...,7$$

1 2 3

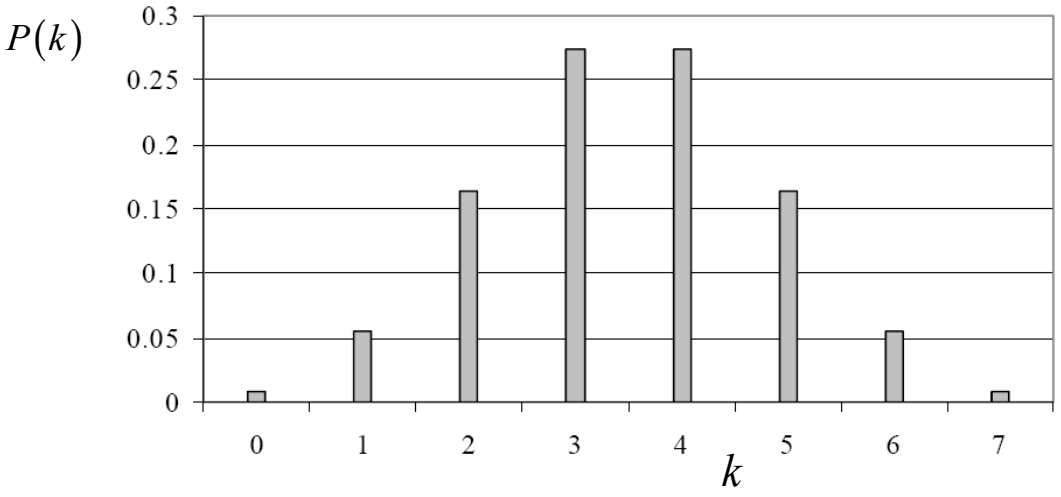
q

Če bi v izrazu (2.26) izračunali verjetnosti za vse $k=0,1,...,7$, bi dobili verjetnostno shemo, kot jo prikazuje slika 20.

k	0	1	2	3	4	5	6	7
$P(k)$	$^{1}/_{128}$	$^{7}/_{128}$	$^{21}/_{128}$	$^{35}/_{128}$	$^{35}/_{128}$	$^{21}/_{128}$	$^{7}/_{128}$	$^{1}/_{128}$

Slika 20: Verjetnostna shema za naključno spremenljivko X pri sedemkratnem metu kovanca

Binomsko porazdelitev naključne spremenljivke *X* bi lahko na osnovi slike 20 predstavili tudi v obliki grafa, ki ga prikazuje slika 21.



Slika 21: Binomska porazdelitev naključne spremenljivke X pri sedemkratnem metu kovanca

2.5 Zvezne naključne spremenljivke

Za zgled tovrstnih spremenljivk uporabimo temperaturo v sobi. Ker je le-ta katerokoli realno število, recimo v intervalu med -10 stopinj in +40 stopinj, je teh števil neskončno mnogo. Kolikšna je torej verjetnost, da bo temperatura $X = 12.5$ stopinj? Kot se izkaže, je tovrstno vprašanje neprimerno zastavljeno. Zakaj? Zato, ker je elementarnih dogodkov za določeno temperaturo neskončno mnogo, saj je neskončno število možnih izidov poskusa merjenja temperature. Če bi uporabili isto logiko kot pri diskretnih naključnih spremenljivkah, bi dobili, da je verjetnost:

$$P\left(X = 12.5^{o}\right) = \frac{1}{\textit{število vseh možnih izidov merjenja temperature}} = \frac{1}{\infty} = 0,$$

torej bi dobili popolnoma nesmiseln rezultat. Bolj smiselno bi se bilo vprašati, kakšna je verjetnost, da izmerjena temperatura zavzame vrednost npr. v intervalu med 12 in 13 stopinj, torej $P\left[12 < X \le 13\right] = ?$

Če imamo opravka z zveznimi naključnimi spremenljivkami, potem njihova zaloga vrednosti vsebuje (končen ali neskončen) interval realnih števil [Hsu]. Tedaj je kumulativna funkcija porazdelitve verjetnosti zvezna funkcija, za katero obstaja tudi odvod $\frac{dF(x)}{dx}$ [Hsu]. Potem lahko vpeljemo tudi pojem zvezne funkcije porazdelitve **gostote** verjetnosti, ki se jo izračuna na naslednji način (za neko zvezno naključno spremenljivko *X*) [Hsu]:

$$f(x) = \lim_{\Delta x \to 0} \frac{F\left(x + \Delta x\right) - F(x)}{\Delta x} = \frac{dF(x)}{dx} \tag{2.27}$$

Iz izraza (2.27) pa lahko izrazimo tudi:

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f\left(t\right) dt \tag{2.28}$$

torej povezavo med kumulativno funkcijo verjetnosti in funkcijo porazdelitve gostote verjetnosti.

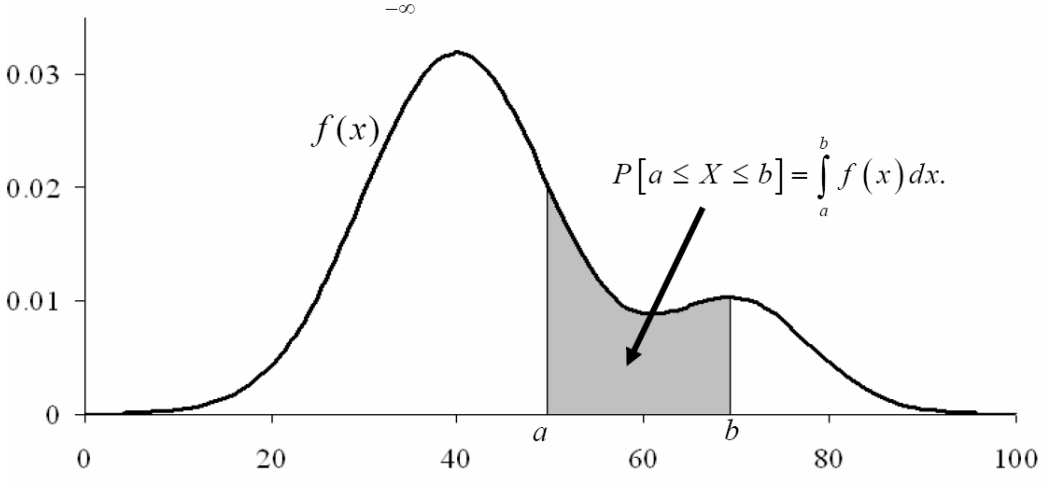
Za zvezno spremenljivko X veljajo naslednje lastnosti [Usenik 2]:

$$
\begin{aligned}
&1.\ Ker\ je\ F(x)\ monotono\ naraščajoča\ funkcija,\ sledi:\\
&f(x)=\frac{dF}{dx}\geq 0\\
&2.\ Ker\ je\ F(\infty)=1,\ sledi:\\
&F(\infty)=P(X\leq\infty)=\int_{-\infty}^{\infty} f(t)\,dt=1\\
&3.\ Ker\ je\ F(b)-F(a)=P(a\leq X\leq b),\ sledi:\\
&\int_{-\infty}^{b} f(t)\,dt-\int_{-\infty}^{a} f(t)\,dt=\int_{-\infty}^{a} f(t)\,dt+\int_{a}^{b} f(t)\,dt-\int_{-\infty}^{a} f(t)\,dt=\int_{a}^{b} f(t)\,dt=P(a\leq X\leq b)\\
&torej:\\
&P(a\leq X\leq b)=\int_{a}^{b} f(t)\,dt
\end{aligned}
\tag{2.29}
$$

Primer zvezne funkcije porazdelitve gostote verjetnosti $f(x)$ in verjetnosti $P(a\leq X\leq b)$, da se zvezna naključna spremenljivka X nahaja na intervalu $[a,b]$, prikazuje slika 22 [Dragan 2]. Kot je razvidno iz slike 22, je $P(a\leq X\leq b)$ enaka ploščini pod funkcijo $f(x)$ na intervalu $[a,b]$, ki jo dobimo z integracijo $f(x)$ na tem intervalu.

funkcija porazdelitve gostote verjetnosti, *f*(*x*)

$$\int_{-\infty}^{\infty} f(x)\,dx = 1.$$

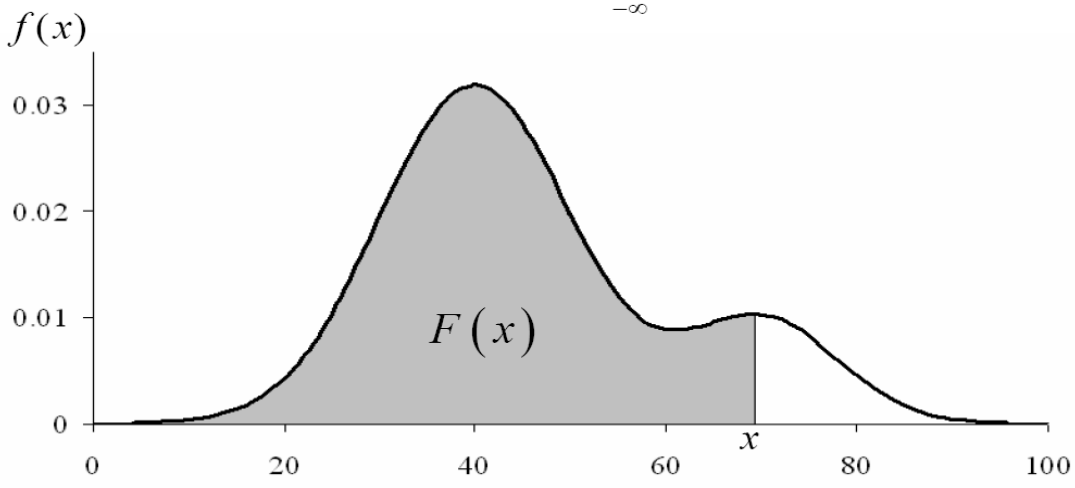


Slika 22: Primer zvezne funkcije porazdelitve gostote verjetnosti $f(x)$ *in verjetnosti* $P(a \le X \le b)$*, da se zvezna naključna spremenljivka X nahaja na intervalu* $[a,b]$.

Primer zvezne kumulativne funkcije porazdelitve verjetnosti $F(x)$ pa prikazuje slika 23 [Dragan 2]. Kot je razvidno iz slike 23, je $F(x)$ enaka ploščini pod funkcijo $f(x)$ na intervalu $[-\infty, x]$, ki jo dobimo z integracijo $f(x)$ na tem intervalu.

Funkcija kumulativne porazdelitve verjetnosti, *F*(*x*)

$$F(x) = P[X \le x] = \int_{-\infty}^{x} f(t)\,dt.$$



Slika 23: Primer zvezne kumulativne funkcije porazdelitve verjetnosti $F(x)$

Porazdelitveni zakon zvezne naključne spremenljivke se imenuje na kratko zvezna porazdelitev [Jamnik 1]. Naštejmo nekatere važnejše zvezne porazdelitve [Jamnik 1]:

- Uniformna (enakomerna zvezna) porazdelitev,
- Normalna (Gaussova) porazdelitev,
- Eksponentna porazdelitev,
- Porazdelitev hi kvadrat,
- Studentova porazdelitev, itn.

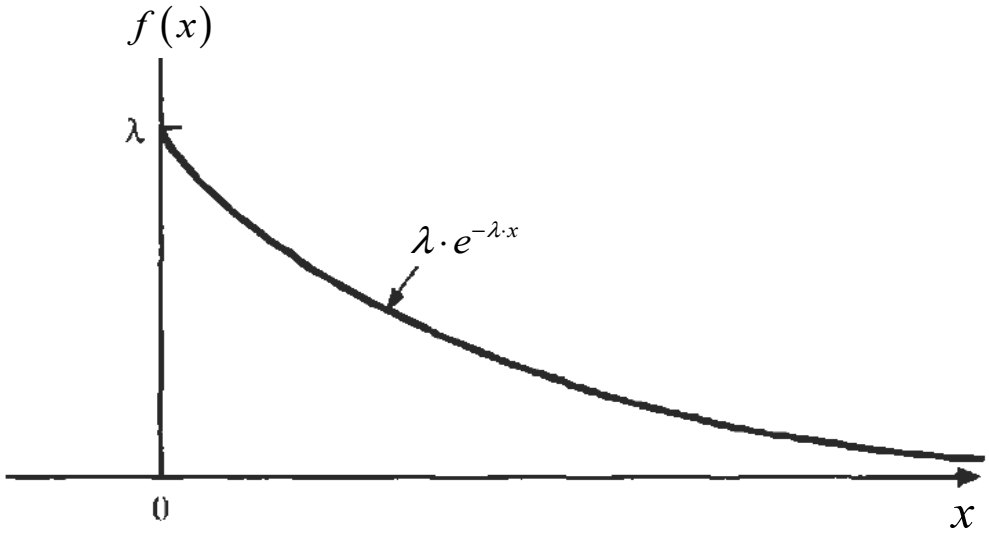
V nadaljevanju si bomo na kratko ogledali uniformno, normalno in eksponentno porazdelitev.

2.6 Eksponentna porazdelitev

Naključna spremenljivka se imenuje eksponentna s parametrom λ, če zanjo velja naslednja porazdelitev gostote verjetnosti [Hsu]:

$$f(x)=\begin{cases}\lambda\cdot e^{-\lambda\cdot x}, & x\geq 0\\ 0, & x<0\end{cases} \tag{2.30}$$

Primer eksponentne porazdelitve je prikazan na sliki 24.



Slika 24: Primer eksponentne porazdelitve

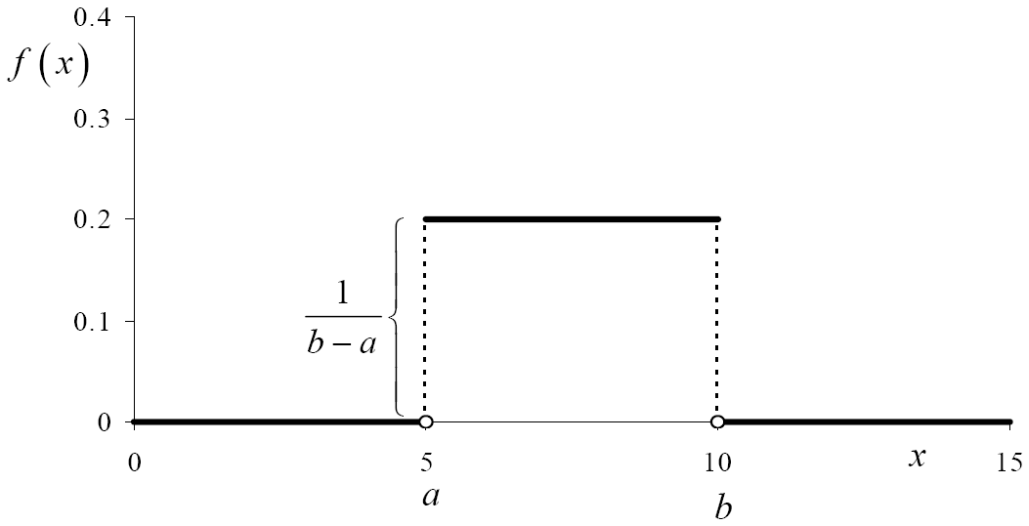
Najbolj zanimiva lastnost eksponentne porazdelitve je takoimenovana "memoryless" (Markovska) lastnost [Hsu]. To npr. pomeni, da če je življenjska doba nekega izdelka (komponente) eksponentno porazdeljena, potem je izdelek, ki je že bil v uporabi nekaj časa (ur), prav tako dober kot nov izdelek z obzirom na preostalo količino časa trajanja izdelka (do okvare). Torej izdelek "pozabi", koliko časa je že obratoval [Hsu].

2.7 Uniformna porazdelitev

Naključna spremenljivka X ima Uniformno porazdelitev na intervalu $[a,b]$, če ima naslednjo funkcijo porazdelitve gostote verjetnosti:

$$f(x)=\begin{cases}\dfrac{1}{b-a} & a\le x\le b\\ 0 & \text{sicer}\end{cases} \tag{2.31}$$

Primer uniformne porazdelitve naključne spremenljivke X je prikazan na sliki 25.



Slika 25: Primer uniformne porazdelitve

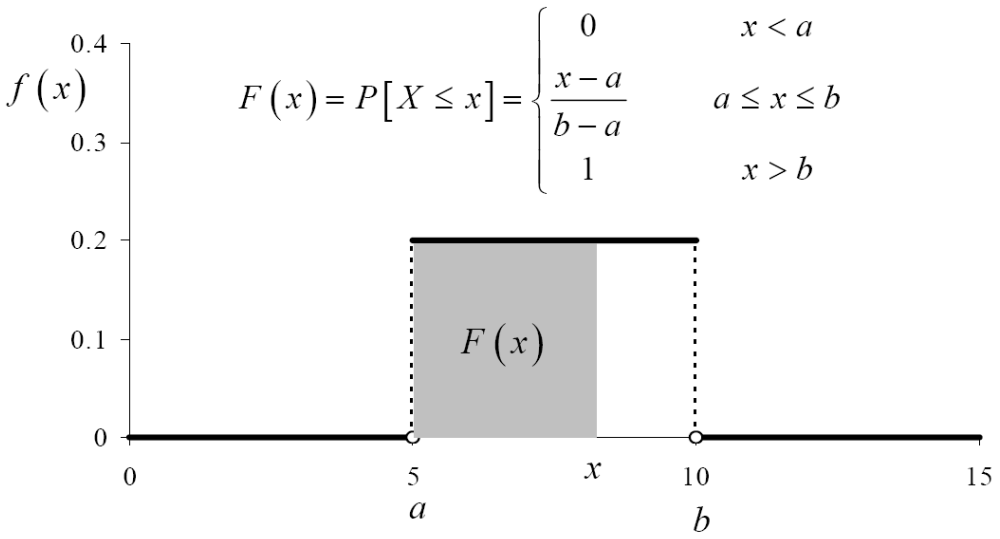
Na osnovi izraza (2.28) lahko izračunamo tudi kumulativno funkcijo porazdelitve verjetnosti $F(x)$. Na intervalu $x \in (-\infty, a)$ je njena vrednost enaka:

$$F(x) = P(X \le x) = \int_{-\infty}^{x} 0 \cdot dt = 0, \qquad -\infty < x < a \tag{2.32}$$

Na intervalu $x \in [a, b]$ je njena vrednost enaka (glej sliko 26):

$$F(x) = P(X \le x) = \int_{a}^{x} \frac{1}{b-a} \cdot dt = \frac{1}{b-a}(x-a), \qquad a \le x \le b \tag{2.33}$$

$$F(x) = \frac{x-a}{b-a}, \qquad a \le x \le b$$



Slika 26: Slika za pomoč izračuna $F(x)$ na intervalu $x \in [a, b]$

Na intervalu $x > b$ pa je njena vrednost enaka:

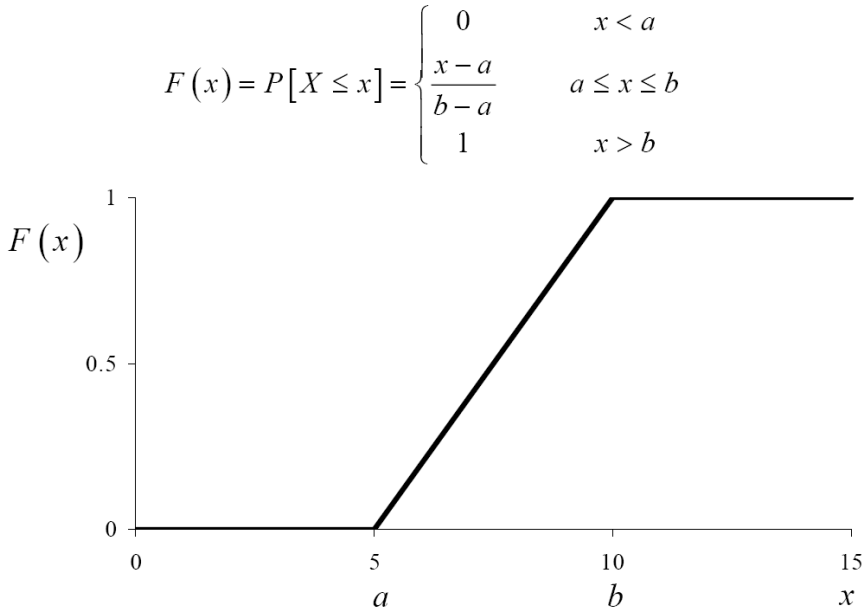
$$F(x) = P(X \le x) = \int_{a}^{b} \frac{1}{b-a} \cdot dt = \frac{1}{b-a}(b-a), \qquad x > b \tag{2.34}$$

$$F(x) = 1, \qquad x > b$$

Če izraze (2.32), (2.33) in (2.34) združimo, torej dobimo za $F(x)$ naslednji izraz:

$$F(x) = P[X \le x] = \begin{cases} 0 & x < a \\ \dfrac{x-a}{b-a} & a \le x \le b \\ 1 & x > b \end{cases} \tag{2.35}$$

Če funkcijo v izrazu (2.35) narišemo, dobimo primer poteka, kot ga prikazuje slika 27.



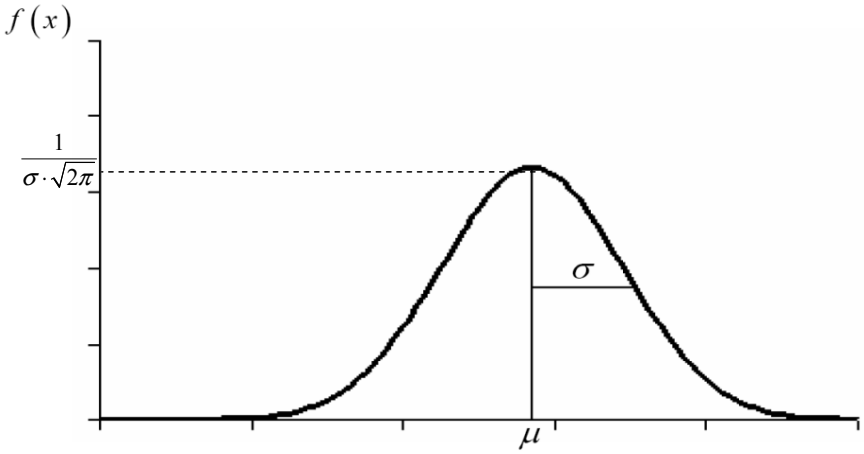
Slika 27: Kumulativna funkcija porazdelitve verjetnosti $F(x)$ pri uniformni porazdelitvi zvezne naključne spremenljivke X

2.8 Normalna porazdelitev

Zvezna naključna spremenljivka *X* ima normalno porazdelitev s srednjo vrednostjo μ in standardno deviacijo σ, če zanjo velja naslednja funkcija porazdelitve gostote verjetnosti [Jamnik 1, Usenik 2]:

$$f(x) = \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{2.36}$$

Primer normalne porazdelitve spremenljivke *X* prikazuje slika 28.



Slika 28: Primer normalne porazdelitve zvezne naključne spremenljivke X

Pri normalni porazdelitvi je μ lahko katerokoli realno število, σ pa je poljubno pozitivno število. Kot se izkaže, parameter μ določa lego krivulje, parameter σ pa njeno obliko [Jamnik 1]. Velja tudi, da tovrstna funkcija doseže maksimum v točki $\left(\mu, \frac{1}{\sigma \cdot \sqrt{2\pi}} \right)$, kar lahko preprosto pokažemo z iskanjem ekstrema te funkcije. V ta namen funkcijo (2.36) najprej odvajamo in dobljeni odvod enačimo z 0, pri čemer dobimo:

$$\begin{aligned} &\frac{df(x)}{dx} = \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}} \cdot \left[-\frac{1}{\sigma^2}(x-\mu) \right] = 0 \\ &\textit{Odtod pa sledi}: \\ &x - \mu = 0 \\ &\textit{oziroma}: \\ &x^* = \mu \end{aligned} \tag{2.37}$$

Če dobljeni izraz $x^* = \mu$ vstavimo v funkcijo (2.36), dobimo:

$$f(x^*) = \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot e^{-\frac{(x^*-\mu)^2}{2\sigma^2}} = \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot e^{-\frac{(\mu-\mu)^2}{2\sigma^2}} = \frac{1}{\sigma \cdot \sqrt{2\pi}} \tag{2.38}$$

Torej res velja za ekstremno točko:

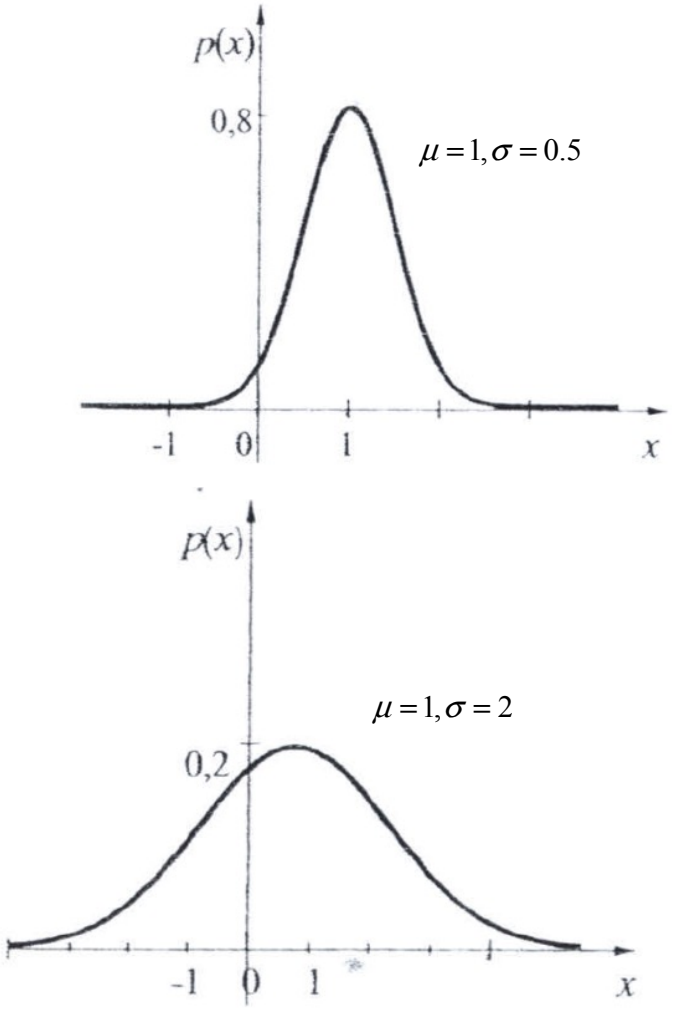
$$(x^*, f(x^*)) = \left(\mu, \frac{1}{\sigma \cdot \sqrt{2\pi}} \right) \tag{2.39}$$

Najpreprostejšo normalno porazdelitev dobimo tedaj, ko velja: $\mu = 0, \sigma = 1$. Tedaj izraz (2.36) preide v obliko:

$$f(x) = \frac{1}{1 \cdot \sqrt{2\pi}} \cdot e^{-\frac{(x-0)^2}{2 \cdot 1^2}} = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{x^2}{2}} \tag{2.40}$$

ki ji pravimo standardizirana normalna porazdelitev [Jamnik 1].

Slika 29 prikazuje primer normalne porazdelitve pri dveh različnih vrednostih parametrov μ in σ [Jamnik 1, Usenik 2]. Iz slike 29 je razvidno, da čim manjši kot je parameter σ, tem bolj izrazito je teme krivulje oziroma je krivulja tem bolj stisnjena okrog temena [Jamnik 1].



Slika 29: Primer normalne porazdelitve pri dveh različnih vrednostih parametra σ [Jamnik 1, Usenik 2]

2.9 Matematično upanje (pričakovanje)

Porazdelitve naključnih spremenljivk so pogostokrat preširok pojem za konkretno uporabo, zato želimo iz porazdelitvenega zakona najti le nekatere osnovne značilnosti, ki jih poskušamo oceniti z določenim številom. Tovrstne značilnosti pa tudi imenujemo z imenom *številske karakteristike* [Usenik 2].

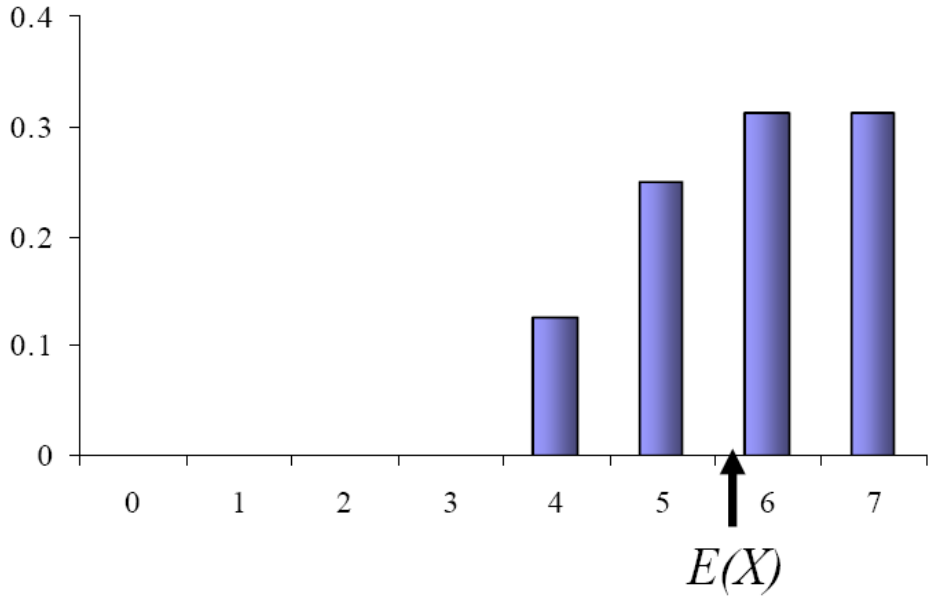
Slednje predstavljajo določene sumarne podatke o naključni spremenljivki, recimo z navedbo, okrog katere povprečne vrednosti je porazdeljena zaloga vrednosti, kako so možne vrednosti naključne spremenljivke razpršene okrog povprečja, in podobno [Jamnik 1].

Najbolj značilne številske karakteristike določene naključne spremenljivke X so [Usenik 2]:

- matematično upanje (pričakovanje) $E(X)=\mu_X$,
- varianca ali disperzija $D(X)=VAR(X)=\sigma_X^2$, ter
- standardna deviacija $\sqrt{D(X)}=\sqrt{VAR(X)}=\sigma_X$

V tem poglavju se bomo omejili na obravnavo matematičnega upanja, kasneje pa bomo obdelali tudi drugi dve številski karakteristiki. Matematičnemu upanju $E(X)$ velikokrat pravimo tudi povprečna vrednost naključne spremenljivke X. Predstavlja namreč število, pri katerem se običajno ustali povprečje realizacij naključne spremenljivke X v primeru velikega števila realizacij [Jamnik 1]. Zgodovinsko gledano pa ime matematično upanje izvira iz loterijskih iger, s katerim so poimenovali upanje na dobitek [Jamnik 1].

Matematično upanje $E(X)$ si lahko interpretiramo tudi kot center gravitacije porazdelitve verjetnosti naključne spremenljivke X, kar je ilustrirano na primeru na sliki 30.



Slika 30: Interpretacija matematičnega upanja $E(X)$ kot centra gravitacije porazdelitve verjetnosti naključne spremenljivke X

V primeru, da imamo opravka z diskretno naključno spremenljivko *X*, ki ima določeno porazdelitev verjetnosti *p(x)* oz. ima verjetnostno shemo:

$$X: \begin{pmatrix} x_1 & x_2 & x_3 \ldots & x_n \\ p_1 & p_2 & p_3 \ldots & p_n \end{pmatrix} \tag{2.41}$$

je matematično upanje $E(X)$ definirano kot [Jamnik 1]:

$$E(X) = \sum_x x \cdot p(x) = \sum_{i=1}^{n} x_i \cdot p(x_i) = \sum_{i=1}^{n} x_i \cdot p_i \tag{2.42}$$

Če ima naključna spremenljivka *X* neskončno zalogo vrednosti, imamo [Jamnik 1, Usenik 2]:

$$E(X) = \sum_x x \cdot p(x) = \sum_{i=1}^{\infty} x_i \cdot p(x_i) = \sum_{i=1}^{\infty} x_i \cdot p_i \tag{2.43}$$

pod pogojem, da je vrsta konvergentna in velja:

$$\sum_{i=1}^{\infty} |x_i| \cdot p_i < \infty \tag{2.44}$$

Če ta pogoj ne velja, potem za dotično naključno spremenljivko matematično upanje ne obstaja [Usenik 2].

Poglejmo si naslednji primer [Usenik 2]:

Primer 2.4.: *Pri streljanju v tarčo zadene strelec v center tarče z verjetnostjo 0.1, dva cm od centra tarče z verjetnostjo 0.3, štiri cm od centra z verjetnostjo 0.3, šest cm od centra z verjetnostjo 0.2 in osem cm od centra z verjetnostjo 0.1. Izračunajte matematično upanje naključne spremenljivke X, ki jo uporabimo v problemu.*

Gotovo za naključno spremenljivko X velja naslednja verjetnostna shema:

$$X: \begin{pmatrix} x_1 & x_2 & x_3 & x_4 & x_5 \\ p_1 & p_2 & p_3 & p_4 & p_5 \end{pmatrix} = \begin{pmatrix} 0 & 2 & 4 & 6 & 8 \\ 0.1 & 0.3 & 0.3 & 0.2 & 0.1 \end{pmatrix} \tag{2.45}$$

S pomočjo izraza (2.42) izračunamo matematično upanje:

$$E(X) = \sum_{i=1}^{5} x_i \cdot p_i = x_1 \cdot p_1 + x_2 \cdot p_2 + x_3 \cdot p_3 + x_4 \cdot p_4 + x_5 \cdot p_5 = \tag{2.46}$$
$$= 0 \cdot 0.1 + 2 \cdot 0.3 + 4 \cdot 0.3 + 6 \cdot 0.2 + 8 \cdot 0.1 = 3.8$$

Dobljeni rezultat pomeni, da bo strelec v povprečju (zato ime povprečna vrednost) pričakoval zadetek, ki bo 3.8 cm oddaljen od centra tarče. Poleg tega tudi velja naslednje. Če bi bili posamezni rezultati pri streljanju ovrednoteni z denarnim dobitkom, bi povprečna vrednost strelcu kazala, koliko je vredno njegovo upanje na dobitek, ki bi ga osvojil z enim strelom [Usenik 2].

V primeru, da imamo opravka z zvezno naključno spremenljivko X, ki ima določeno porazdelitev gostote verjetnosti *f(x)*, je matematično upanje $E(X)$ definirano kot [Jamnik 1]:

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x)\, dx \tag{2.47}$$

pri čemer matematično upanje obstaja le tedaj, ko velja [Jamnik 1]:

$$\int_{-\infty}^{\infty} |x| \cdot f(x)\,dx < \infty \tag{2.48}$$

Poglejmo si naslednji primer [Jamnik 1]:

Primer 2.5.: *Izračunajte matematično upanje naključne spremenljivke X, če zanjo velja normalna porazdelitev.*

Če vstavimo (2.36) v izraz (2.47), dobimo:

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x)\,dx = \int_{-\infty}^{\infty} x \cdot \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx = \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot \int_{-\infty}^{\infty} x \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx \tag{2.49}$$

Vpeljimo substitucijo:

$$\begin{aligned} x &= \mu + z \cdot \sigma \\ z &= \frac{x-\mu}{\sigma} \\ dz &= \frac{1}{\sigma}\,dx \end{aligned} \tag{2.50}$$

Z njo izraz (2.49) preide v obliko:

$$\begin{aligned} E(X) &= \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot \int_{-\infty}^{\infty} (\mu + z \cdot \sigma) \cdot e^{-\frac{z^2}{2}} \cdot \sigma \cdot dz = \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{\infty} (\mu + z \cdot \sigma) \cdot e^{-\frac{z^2}{2}} dz = \\ &= \frac{1}{\sqrt{2\pi}} \left(\int_{-\infty}^{\infty} \mu \cdot e^{-\frac{z^2}{2}} dz + \int_{-\infty}^{\infty} z \cdot \sigma \cdot e^{-\frac{z^2}{2}} dz \right) = \\ &= \frac{\mu}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}} dz + \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z \cdot e^{-\frac{z^2}{2}} dz \end{aligned} \tag{2.51}$$

Dokazati se da, da velja naslednje:

$$\begin{aligned} \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}} dz &= \sqrt{2\pi} \\ \int_{-\infty}^{\infty} z e^{-\frac{z^2}{2}} dz &= 0 \end{aligned} \tag{2.52}$$

s čimer izraz (2.51) preide v obliko:

$$E(X) = \frac{\mu}{\sqrt{2\pi}} \cdot \sqrt{2\pi} + \frac{\sigma}{\sqrt{2\pi}} \cdot 0$$
$$E(X) = \mu \tag{2.53}$$

Poglejmo si še naslednji primer:

Primer 2.6.: *Izračunajte matematično upanje naključne spremenljivke X, če zanjo velja uniformna porazdelitev.*

Če vstavimo (2.31) v izraz (2.47), dobimo:

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x)\,dx = \int_{a}^{b} x \cdot \frac{1}{b-a}\,dx = \frac{1}{b-a}\int_{a}^{b} x \cdot dx =$$
$$= \frac{1}{b-a}\left(\frac{x^2}{2}\right)_a^b = \frac{1}{2(b-a)}\left(b^2 - a^2\right) = \frac{a+b}{2} \tag{2.54}$$

2.10 Varianca in standardna deviacija

2.10.1 Varianca

Disperzija ali varianca naključne spremenljivke X je definirana kot povprečna vrednost kvadriranih odklonov naključne spremenljivke od njene povprečne vrednosti. Z njo torej merimo razpršenost naključne spremenljivke okoli njenega povprečja [Usenik 2].

Varianca je v splošnem definirana na naslednji način [Usenik 2]:

$$VAR(X) = D(X) = E\left[\left\{X - E(X)\right\}^2\right] \tag{2.55}$$

Pri diskretni naključni spremenljivki izraz (2.55) preide v obliko:

$$VAR(X) = D(X) = \sum_i \left\{x_i - E(X)\right\}^2 \cdot p_i \tag{2.56}$$

pri zvezni naključni spremenljivki pa preide v obliko:

$$VAR(X) = D(X) = \int_{-\infty}^{\infty} \{x - E(X)\}^2 \cdot f(x)\, dx \tag{2.57}$$

Poskušajmo izraz (2.56) še nekoliko razviti:

$$\begin{aligned} VAR(X) &= D(X) = \sum_i \{x_i^2 - 2 \cdot x_i \cdot E(X) + E^2(X)\} \cdot p_i = \\ &= \sum_i x_i^2 \cdot p_i - 2 \cdot E(X) \cdot \sum_i x_i \cdot p_i + E^2(X) \underbrace{\sum_i p_i}_{1} = \\ &= E(X^2) - 2 \cdot E(X) \cdot E(X) + E^2(X) \end{aligned} \tag{2.58}$$

torej velja:

$$VAR(X) = D(X) = E(X^2) - E^2(X)$$

Do končnega rezultata v izrazu (2.58) bi prišli tudi, če bi še nekoliko razvili izraz (2.57) za zvezno naključno spremenljivko. Torej vedno velja [Usenik 2]:

$$VAR(X) = D(X) = E(X^2) - E^2(X) \tag{2.59}$$

Poglejmo si naslednji primer [Usenik 2]:

Primer 2.7.: *Za diskretno naključno spremenljivko, ki ima podano naslednjo verjetnostno shemo:*

$$X : \begin{pmatrix} x_1 & x_2 & x_3 & x_4 & x_5 \\ p_1 & p_2 & p_3 & p_4 & p_5 \end{pmatrix} = \begin{pmatrix} -2 & -1 & 0 & 1 & 3 \\ 0.1 & \lambda & 0.4 & 0.2 & 0.1 \end{pmatrix} \tag{2.60}$$

določite:

a) število λ*,*

b) matematično upanje, ter

c) varianco.

Ker mora veljati:

$$p_1 + p_2 + p_3 + p_4 + p_5 = 1 \tag{2.61}$$

sledi:

$$\begin{aligned} &0.1 + \lambda + 0.4 + 0.2 + 0.1 = 0.8 + \lambda = 1 \\ &\lambda = 0.2 \end{aligned} \tag{2.62}$$

Matematično upanje izračunamo na osnovi izraza (2.42):

$$\begin{aligned} &E(X) = \sum_{i=1}^{5} x_i \cdot p_i = x_1 \cdot p_1 + x_2 \cdot p_2 + x_3 \cdot p_3 + x_4 \cdot p_4 + x_5 \cdot p_5 = \\ &= -2 \cdot 0.1 + (-1) \cdot 0.2 + 0 \cdot 0.4 + 1 \cdot 0.2 + 3 \cdot 0.1 = 0.1 \end{aligned} \tag{2.63}$$

Varianco izračunamo na osnovi izraza (2.59). V ta namen pa moramo prej izračunati še $E\left(X^2\right)$:

$$\begin{aligned} &E\left(X^2\right) = \sum_{i=1}^{5} x_i^{\,2} \cdot p_i = x_1^{\,2} \cdot p_1 + x_2^{\,2} \cdot p_2 + x_3^{\,2} \cdot p_3 + x_4^{\,2} \cdot p_4 + x_5^{\,2} \cdot p_5 = \\ &= (-2)^2 \cdot 0.1 + (-1)^2 \cdot 0.2 + 0^2 \cdot 0.4 + 1^2 \cdot 0.2 + 3^2 \cdot 0.1 = 0.4 + 0.2 + 0.2 + 0.9 = 1.7 \end{aligned} \tag{2.64}$$

Varianca torej je:

$$VAR(X) = D(X) = E\left(X^2\right) - E^2(X) = 1.7 - (0.1)^2 = 1.69 \tag{2.65}$$

Poglejmo si še en primer:

Primer 2.8.: *Za diskretno naključno spremenljivko s Poissonovo porazdelitvijo izračunajte varianco!*

Varianco izračunamo na osnovi izraza (2.59). V ta namen pa moramo prej izračunati tako $E(X)$, kot tudi $E\left(X^2\right)$. Izračunajmo najprej $E(X)$. Na osnovi izrazov (2.5) in (2.43) lahko zapišemo:

$$E(X)=\sum_k k\cdot p(k)=\sum_{k=1}^{\infty} k\cdot p(k)=$$
$$=\sum_{k=1}^{\infty} k\cdot p_k=\sum_{k=1}^{\infty} k\cdot\frac{\lambda^k}{k!}\cdot e^{-\lambda}=\frac{\lambda}{\lambda}\cdot\sum_{k=1}^{\infty} k\cdot\frac{\lambda^k}{k!}\cdot e^{-\lambda}=\lambda\cdot e^{-\lambda}\cdot\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!} \tag{2.66}$$
vpeljimo $n=k-1,$ *sledi*:
$$E(X)=\lambda\cdot e^{-\lambda}\cdot\underbrace{\sum_{n=0}^{\infty}\frac{\lambda^n}{(n)!}}_{e^{\lambda}}=\lambda\cdot e^{-\lambda}\cdot e^{\lambda}=\lambda$$

Izračunajmo še $E\left(X^2\right)$. V ta namen moramo najprej izračunati $E\left(X(X-1)\right)$ [Hsu]:

$$E\left(X(X-1)\right)=\sum_k k(k-1)\cdot p(k)=$$
$$=\sum_{k=1}^{\infty} k(k-1)\cdot p(k)=\sum_{k=1}^{\infty} k(k-1)\cdot p_k$$
$$=\sum_{k=1}^{\infty} k(k-1)\cdot\frac{\lambda^k}{k!}\cdot e^{-\lambda}=\underbrace{1(1-1)\cdot\frac{\lambda^1}{1!}\cdot e^{-\lambda}}_{0}+\sum_{k=2}^{\infty} k(k-1)\cdot\frac{\lambda^k}{k!}\cdot e^{-\lambda}= \tag{2.67}$$
$$=\frac{\lambda^2}{\lambda^2}\sum_{k=2}^{\infty}\frac{\lambda^k}{(k-2)!}\cdot e^{-\lambda}=\lambda^2\cdot e^{-\lambda}\sum_{k=2}^{\infty}\frac{\lambda^{k-2}}{(k-2)!}$$
vpeljimo $n=k-2,$ *sledi*:
$$E\left(X(X-1)\right)=\lambda^2\cdot e^{-\lambda}\cdot\underbrace{\sum_{n=0}^{\infty}\frac{\lambda^n}{(n)!}}_{e^{\lambda}}=\lambda^2\cdot e^{-\lambda}\cdot e^{\lambda}=\lambda^2$$

Po drugi strani velja:

$$E\left(X(X-1)\right)=E\left(X^2-X\right)=E\left(X^2\right)-E(X)=E\left(X^2\right)-\lambda \tag{2.68}$$

Če izenačimo izraza (2.67) in (2.68), dobimo:

$$E\left(X^2\right)-\lambda=\lambda^2 \tag{2.69}$$
$$E\left(X^2\right)=\lambda^2+\lambda$$

Varianca torej je:

$$VAR(X) = D(X) = E\left(X^2\right) - E^2(X) = \lambda^2 + \lambda - \lambda^2 = \lambda \tag{2.70}$$

Poglejmo si še en primer:

Primer 2.9.: *Za zvezno naključno spremenljivko z uniformno porazdelitvijo izračunajte varianco!*

Varianco izračunamo na osnovi izraza (2.59). Za matematično upanje na osnovi izraza (2.54) vemo, da je enako $\frac{a+b}{2}$. Torej moramo izračunati še $E\left(X^2\right)$:

$$\begin{aligned} E\left(X^2\right) &= \int_{-\infty}^{\infty} x^2 \cdot f(x)\,dx = \int_a^b x^2 \cdot \frac{1}{b-a}\,dx = \frac{1}{b-a}\int_a^b x^2 \cdot dx = \\ &= \frac{1}{b-a}\left(\frac{x^3}{3}\right)_a^b = \frac{1}{3(b-a)}\left(b^3 - a^3\right) = \frac{\left(a^2 + a\cdot b + b^2\right)(b-a)}{3(b-a)} = \\ &= \frac{\left(a^2 + a\cdot b + b^2\right)}{3} \end{aligned} \tag{2.71}$$

Varianca torej je:

$$\begin{aligned} VAR(X) &= D(X) = E\left(X^2\right) - E^2(X) = \frac{\left(a^2 + a\cdot b + b^2\right)}{3} - \left(\frac{a+b}{2}\right)^2 = \\ &= \frac{\left(a^2 + a\cdot b + b^2\right)}{3} - \frac{a^2 + 2\cdot a\cdot b + b^2}{4} = \frac{4\left(a^2 + a\cdot b + b^2\right) - 3\left(a^2 + 2\cdot a\cdot b + b^2\right)}{12} = \\ &= \frac{\left(4a^2 + 4a\cdot b + 4b^2\right) - \left(3a^2 + 6\cdot a\cdot b + 3b^2\right)}{12} = \\ &= \frac{\left(a^2 - 2a\cdot b + b^2\right)}{12} = \frac{(a-b)^2}{12} \end{aligned} \tag{2.72}$$

2.10.2 Standardna deviacija

Ker je disperzija številska karakteristika, ki jo računamo iz kvadriranih odklonov od povprečja, lahko dobijo veliki odkloni prevelik vpliv [Usenik 2]. Da bi ta vpliv vsaj delno razvrednotili, vzamemo za dodatno mero razpršenosti naključne spremenljivke le pozitivni kvadratni koren variance in dobimo takoimenovano **standardno deviacijo**:

$$\sqrt{D(X)} = \sqrt{VAR(X)} = \sigma_X \tag{2.73}$$

2.11 Številske karakteristike za različne porazdelitve in pregled osnovnih lastnosti

V tem poglavju bomo najprej podali pregledno tabelo številskih karakteristik (matematičnega upanja in variance) za nekatere poglavitne porazdelitve diskretnih oz. zveznih naključnih spremenljivk. Nekatere smo tudi izpeljali v prejšnjih poglavjih, izpeljave ostalih pa si bralec lahko pogleda v literaturi [Hsu, Jamnik 1]. Slika 31 prikazuje tabelo številskih karakteristik različnih porazdelitev [Hsu, Jamnik 1].

PORAZDELITEV	MATEMATIČNO UPANJE	VARIANCA
BINOMSKA	$n \cdot p$	$n \cdot p \cdot q$
POISSONOVA	λ	λ
NORMALNA	μ	σ^2
UNIFORMNA ZVEZNA	$\frac{a+b}{2}$	$\frac{(a-b)^2}{12}$
EKSPONENTNA	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

Slika 31: Tabela številskih karakteristik poglavitnih porazdelitev diskretnih oz. zveznih naključnih spremenljivk

V nadaljevanju si še na kratko poglejmo nekaj osnovnih lastnosti matematičnega upanja in variance [Jamnik 1 , Usenik 2]. Gre namreč za karakteristiki, ki ju v verjetnostnem računu veliko uporabljamo, še zlasti pa imata pomembno vlogo v statistiki [Usenik 2].

Podali bomo le osnovne lastnosti, dokaze zanje pa si bralec lahko pogleda v literaturi [Jamnik 1, Usenik 2].

***Izrek 1**: Če za naključni spremenljivki X in Y obstajata njuni matematični upanji $E(X)$ in $E(Y)$, potem obstaja tudi matematično upanje njune vsote in velja:*

$$E(X+Y)=E(X)+E(Y) \tag{2.74}$$

Ta izrek je mogoče posplošiti tudi za *n* naključnih spremenljivk. Tedaj velja:

$$E(X_1+X_2+...+X_n)=E(X_1)+E(X_2)+...+E(X_n) \tag{2.75}$$

***Izrek 2**: Če za naključno spremenljivko X obstaja matematično upanje $E(X)$ in je a poljubno realno število, potem obstaja tudi matematično upanje naključne spremenljivke a.X in velja:*

$$E(a\cdot X)=a\cdot E(X) \tag{2.76}$$

Podobno velja tudi naslednji izrek.

***Izrek 3**: Če za naključni spremenljivki X in Y obstajata njuni matematični upanji $E(X)$ in $E(Y)$ in sta λ in μ poljubni realni števili, potem velja naslednji izraz:*

$$E(\lambda\cdot X+\mu\cdot Y)=\lambda\cdot E(X)+\mu\cdot E(Y) \tag{2.77}$$

Ta izrek je mogoče posplošiti tudi za *n* naključnih spremenljivk. Tedaj velja:

$$E(\lambda_1X_1+\lambda_2X_2+...+\lambda_nX_n)=\lambda_1\cdot E(X_1)+\lambda_2\cdot E(X_2)+...+\lambda_n\cdot E(X_n) \tag{2.78}$$

***Izrek 4**: Če sta naključni spremenljivki X in Y med seboj neodvisni in obstajata njuni matematični upanji $E(X)$ in $E(Y)$, potem velja naslednji izraz (nekoreliranost!):*

$$E(X \cdot Y) = E(X) \cdot E(Y) \tag{2.79}$$

Če pa sta med seboj odvisni (korelirani), pa izraz (2.79) seveda ne velja:

$$E(X \cdot Y) \neq E(X) \cdot E(Y) \tag{2.80}$$

Izrek (2.79) je mogoče posplošiti tudi za *n* med seboj neodvisnih naključnih spremenljivk. Tedaj velja:

$$E(X_1 \cdot X_2 \cdot \ldots \cdot X_n) = E(X_1) \cdot E(X_2) \cdot \ldots \cdot E(X_n) \tag{2.81}$$

Tudi za varianco veljajo določene lastnosti.

***Izrek 5**: Če za naključno spremenljivko X obstaja disperzija $D(X)$ in je a poljubno realno število, potem obstaja naslednji izraz:*

$$D(a \cdot X) = a^2 \cdot D(X) \tag{2.82}$$

***Izrek 6**: Če sta naključni spremenljivki X in Y med seboj nekorelirani in obstajata njuni disperziji $D(X)$ in $D(Y)$, potem velja naslednji izraz:*

$$D(X + Y) = D(X) + D(Y) \tag{2.83}$$

Ta izrek je mogoče posplošiti tudi za *n* med seboj paroma nekoreliranih naključnih spremenljivk. Tedaj velja:

$$D(X_1 + X_2 + ... + X_n) = D(X_1) + D(X_2) + ... + D(X_n) \tag{2.84}$$

Seveda se v primeru, ko so naključne spremenljivke korelirane med seboj, nekateri zgoraj podani izrazi nekoliko spremenijo, saj je potrebno upoštevati tudi takoimenovane kovariance med naključnimi spremenljivkami. Več o tem si bralec lahko pogleda v literaturi [Hsu, Jamnik 1, Usenik 2].

2.12 Pričakovanje funkcij naključnih spremenljivk

Denimo imamo dano diskretno naključno spremenljivko *X*, ki ima funkcijo porazdelitve verjetnosti *p(x)*. Dano imamo tudi funkcijo te naključne spremenljivke *g(X)*. Potem je pričakovana vrednost funkcije *g(X)*, to je $E\left[g(X)\right]$, enaka:

$$E\left[g(X)\right] = \sum_x g(x)\,p(x) = \sum_i g(x_i)\,p(x_i) \tag{2.85}$$

V primeru zveznih naključnih spremenljivk velja podobno sklepanje.

Denimo imamo dano zvezno naključno spremenljivko *X*, ki ima funkcijo porazdelitve gostote verjetnosti *f(x)*. Dano imamo tudi funkcijo te naključne spremenljivke *g(X)*. Potem je pričakovana vrednost funkcije *g(X)*, to je $E\left[g(X)\right]$, enaka:

$$E\left[g(X)\right] = \int_{-\infty}^{\infty} g(x) \cdot f(x)\,dx \tag{2.86}$$

Poglejmo si naslednji primer:

Primer 2.10.: *Dano imamo uniformno porazdeljeno zvezno naključno spremenljivko X, ki predstavlja dolžino stranice kvadrata in je izbrana naključno med vrednostima 0 in b. Dano imamo tudi funkcijo naključne spremenljivke* $A = g(X) = X^2$, *ki predstavlja površino kvadrata. Poiščite* $E\left[g(X)\right]$!

Ker je X uniformno porazdeljena na intervalu *(a,b) = (0,b),* velja:

$$f(x)=\begin{cases}\frac{1}{b-0} & 0\le x\le b\\ 0 & x<0, x>b\end{cases}=\begin{cases}\frac{1}{b} & 0\le x\le b\\ 0 & x<0, x>b\end{cases} \tag{2.87}$$

Na osnovi izraza (2.86) lahko izračunamo $E\left[g(X)\right]$:

$$\begin{aligned} E\left[g(X)\right] &= \int_{-\infty}^{\infty} g(x)\cdot f(x)\,dx \\ E\left[X^2\right] &= \int_{-\infty}^{\infty} x^2\cdot f(x)\,dx \\ E\left[X^2\right] &= \int_0^b x^2\cdot\frac{1}{b}\,dx=\frac{1}{b}\left(\frac{x^3}{3}\right)_0^b=\frac{b^2}{3} \end{aligned} \tag{2.88}$$

Ker je maksimalna površina kvadrata enaka b^2, je torej $E\left[g(X)\right]$ enaka eni tretjini te vrednosti.

2.13 Transformacijska metoda

V tem poglavju si bomo pogledali transformacijsko metodo za izračun porazdelitve gostote verjetnosti funkcij naključnih spremenljivk.

Teorem:

Naj bo X naključna spremenljivka s funkcijo porazdelitve gostote verjetnosti f(x) in funkcijo naključne spremenljivke U = h(X). Če predpostavimo, da je h(x) bodisi striktno monotono naraščajoča bodisi padajoča funkcija, potem je funkcija porazdelitve gostote verjetnosti za U, to je g(U), definirana z naslednjim izrazom:

$$g(u)=f\left(h^{-1}(u)\right)\cdot\left|\frac{dh^{-1}(u)}{du}\right|=f(x)\cdot\left|\frac{dx(u)}{du}\right| \tag{2.89}$$

Poglejmo si dokaz tega teorema.

Najprej izrazimo spremenljivko *x* kot inverzno funkcijo spremenljivke *u*:

$$U=h(X)\quad\Rightarrow\quad u=h(x)\quad\Rightarrow\quad x=h^{-1}(u) \tag{2.90}$$

Za kumulativno funkcijo porazdelitve verjetnosti naključne spremenljivke U gotovo velja:

$$G(u) = P[U \le u] = P\left[h(X) \le u\right] \tag{2.91}$$

Pokazati se da, da izraz (2.91) preide v obliko:

$$G(u) = \begin{cases} P\left[X \le h^{-1}(u)\right] & h \text{ striktno monotno narašča} \\ P\left[X \ge h^{-1}(u)\right] & h \text{ striktno monotno pada} \end{cases} \tag{2.92}$$

oziroma:

$$G(u) = \begin{cases} F\left(h^{-1}(u)\right) & h \text{ striktno monotno narašča} \\ 1 - F\left(h^{-1}(u)\right) & h \text{ striktno monotno pada} \end{cases} \tag{2.93}$$

kjer je F kumulativna funkcija porazdelitve verjetnosti spremenljivke X.

Relacija med funkcijo porazdelitve gostote verjetnosti naključne spremenljivke U in kumulativno funkcijo te spremenljivke je seveda naslednja:

$$g(u) = \frac{dG(u)}{du} \tag{2.94}$$

pri čemer dobimo:

$$\begin{aligned} g(u) &= \frac{d}{du}\left[\begin{cases} F\left(h^{-1}(u)\right) & h \text{ striktno monotno narašča} \\ 1 - F\left(h^{-1}(u)\right) & h \text{ striktno monotno pada} \end{cases}\right] = \\ &= \begin{cases} \dfrac{dF\left(h^{-1}(u)\right)}{dh^{-1}} \cdot \dfrac{dh^{-1}(u)}{du} & h \text{ striktno monotno narašča} \\ -\dfrac{dF\left(h^{-1}(u)\right)}{dh^{-1}} \cdot \dfrac{dh^{-1}(u)}{du} & h \text{ striktno monotno pada} \end{cases} \end{aligned} \tag{2.95}$$

Seveda velja tudi:

$$f\left[h^{-1}(u)\right] = \frac{dF\left(h^{-1}(u)\right)}{dh^{-1}} = \frac{dF(x)}{dx} \tag{2.96}$$

Zato izraz (2.95) preide v obliko:

$$g(u)=\begin{cases} f\left[h^{-1}(u)\right]\cdot\dfrac{dh^{-1}(u)}{du} & h \text{ striktno monotno narašča} \\ -f\left[h^{-1}(u)\right]\cdot\dfrac{dh^{-1}(u)}{du} & h \text{ striktno monotno pada} \end{cases} \tag{2.97}$$

$$g(u)=f\left[h^{-1}(u)\right]\cdot\left|\frac{dh^{-1}(u)}{du}\right|=f(x)\cdot\left|\frac{dx}{du}\right|$$

Poglejmo si naslednji primer :

Primer 2.11.: *Denimo ima zvezna naključna spremenljivka X normalno porazdelitev s srednjo vrednostjo* μ *in varianco* σ^2. *Poiščite porazdelitev gostote verjetnosti funkcije te naključne spremenljivke* $U=h(X)=e^X$.

Najprej zapišimo inverzno funkcijo, to je:

$$U=h(X)=e^X \quad \Rightarrow \quad u=h(x)=e^x \quad \Rightarrow \quad x=h^{-1}(u)=\ln(u) \tag{2.98}$$

Nato tvorimo odvod v izrazu (2.89), pri čemer dobimo:

$$\frac{dh^{-1}(u)}{du}=\frac{dx(u)}{du}=\frac{d}{du}\left(\ln(u)\right)=\frac{1}{u} \tag{2.99}$$

Na osnovi (2.36) lahko zapišemo:

$$f\left(h^{-1}(u)\right)=f\left(\ln(u)\right)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(\ln(u)-\mu)^2}{2\sigma^2}} \tag{2.100}$$

Nazadnje lahko na osnovi izrazov (2.89), (2.99) in (2.100) zapišemo naslednjo porazdelitev gostote verjetnosti za spremenljivko *U*:

$$g(u)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(\ln(u)-\mu)^2}{2\sigma^2}}\cdot\left|\frac{1}{u}\right|=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(\ln(u)-\mu)^2}{2\sigma^2}}\cdot\frac{1}{u},\quad u>0 \tag{2.101}$$

Porazdelitev, ki smo jo dobili, imenujemo **log-normalna porazdelitev**.

2.14 Združeno porazdeljene naključne spremenljivke

Pogosto imamo opravka z dvema ali več naključnimi spremenljivkami (*X*, *Y*, itn), definiranimi pri istem naključnem eksperimentu [Dragan 2]. Tedaj pravimo tudi, da imamo opravka z multivariantnimi porazdelitvami teh spremenljivk.

Če imamo npr. opravka z dvema naključnima spremenljivkama *X* in *Y*, potem lahko njuno funkcijo združene porazdelitve verjetnosti zapišemo na naslednji način [Dragan 2]:

$$p(x,y)=P(X=x,Y=y) \tag{2.102}$$

Izraz (2.102) lahko pri diskretnih (bivariantnih) spremenljivkah *X* in *Y* zapišemo tudi nekoliko drugače [Hsu]:

$$p(x_i,y_j)=P(X=x_i,Y=y_j) \tag{2.103}$$

pri čemer imata spremenljivki zalogo vrednosti (x_i,y_j) za določen niz celih števil *i* in *j*.

Če imamo opravka s spremenljivkami $X_1,X_2,...,X_k$, to je s *k* diskretnimi naključnimi spremenljivkami, potem lahko funkcijo združene porazdelitve verjetnosti zanje napišemo na naslednji način [Hsu]:

$$p(x_1,x_2,...,x_k)=P(X_1=x_1,X_2=x_2,...,X_k=x_k) \tag{2.104}$$

Za funkcijo (2.104) veljajo naslednje lastnosti [Hsu]:

$$\begin{aligned}
&1. \quad 0 \le p(x_1,\text{K}\ ,x_k) \le 1 \\
&2. \quad \sum_{x_1}\text{K}\sum_{x_k} p(x_1,\text{K}\ ,x_k) = 1 \\
&3. \quad P\left[(X_1,\text{K}\ ,X_k)\in A\right] = \sum_{(X_1,\text{K}\ ,X_k)\in R_A} \ldots \sum p(x_1,\text{K}\ ,x_k), \\
&\textit{pri čemer } (x_1,\text{K}\ ,x_k)\in A
\end{aligned} \tag{2.105}$$

Tretja lastnost pomeni, da lahko verjetnost nekega *k*-dimenzionalnega dogodka *A* najdemo tako, da izraz (2.104) sumiramo preko vseh točk v *k*-dimenzionalnem prostoru dimenzije R_A, ki pripada dogodku *A* [Hsu].

Če pa imamo opravka s spremenljivkami $X_1, X_2, ..., X_n$, to je z *n* zveznimi naključnimi spremenljivkami, potem lahko funkcijo združene porazdelitve gostote verjetnosti zanje napišemo na naslednji način [Hsu]:

$$f(x_1, x_2, ..., x_n) \tag{2.106}$$

Za funkcijo (2.106) veljajo naslednje lastnosti [Hsu]:

$$\begin{aligned}
&1. \quad f(x_1,\text{K}\ ,x_n) \ge 0 \\
&2. \quad \int_{-\infty}^{\infty}\text{K}\int_{-\infty}^{\infty} f(x_1,\text{K}\ ,x_n)\,dx_1\cdot\text{K}\cdot dx_n = 1 \\
&3. \quad P\left[(X_1,\text{K}\ ,X_n)\in A\right] = \int_{(X_1,\text{K}\ ,X_n)\in R_A} \text{K}\int f(x_1,\text{K}\ ,x_n)\,dx_1\cdot\text{K}\cdot dx_n \\
&\textit{pri čemer } (x_1,\text{K}\ ,x_n)\in A
\end{aligned} \tag{2.107}$$

Tretja lastnost ima podoben pomen kot v izrazu (2.105) pri diskretnih naključnih spremenljivkah.

Poglejmo si naslednji primer:

Zapišitev bivariantno normalno porazdelitev dveh naključnih spremenljivk X in Y!

Podobno kot v izrazu (2.36) za univariantno normalno porazdelitev ene naključne spremenljivke *X*, lahko v tem primeru zapišemo funkcijo združene porazdelitve gostote verjetnosti na naslednji način [Hsu]:

$$f(x,y)=\frac{1}{2\cdot\pi\cdot\sigma_x\cdot\sigma_y\cdot\sqrt{1-\rho^2}}\cdot e^{-\frac{1}{2\left(1-\rho^2\right)}\cdot\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2-2\cdot\rho\cdot\left(\frac{x-\mu_x}{\sigma_x}\right)\cdot\left(\frac{y-\mu_y}{\sigma_y}\right)+\left(\frac{y-\mu_y}{\sigma_y}\right)^2\right]} \qquad (2.108)$$

kjer sta $\sigma_x^{\,2},\mu_x$ varianca oz srednja vrednost naključne spremenljivke *X*, $\sigma_y^{\,2},\mu_y$ sta varianca oz srednja vrednost naključne spremenljivke *Y*, medtem ko je ρ takoimenovan **korelacijski koeficient** spremenljivk *X* in *Y* [Hsu].

Poglejmo si še en primer:

Zapišite k-variantno normalno porazdelitev k naključnih spremenljivk $X_1,X_2,...,X_k$!

Če označimo z **x** vektor, ki pripada naključnim spremenljivkam $X_1,X_2,...,X_k$:

$$\mathbf{x}=\begin{bmatrix}x_1\\x_2\\\mathrm{M}\\x_k\end{bmatrix} \qquad (2.109)$$

potem lahko zapišemo funkcijo združene porazdelitve gostote verjetnosti na naslednji način [Hsu]:

$$f(x_1,\mathrm{K}\ ,x_k)=f(\mathbf{x})=\frac{1}{(2\cdot\pi)^{k/2}\left|\det\mathbf{K}\right|^{1/2}}e^{-\frac{1}{2}\cdot(\mathbf{x}-\boldsymbol{\mu})^T\cdot\mathbf{K}^{-1}\cdot(\mathbf{x}-\boldsymbol{\mu})} \qquad (2.110)$$

pri čemer je **μ** vektor srednjih vrednosti naključnih spremenljivk, **K** pa kovariančna matrika. Slednja imata obliko [Hsu]:

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \text{M} \\ \mu_k \end{bmatrix} = \begin{bmatrix} E(X_1) \\ E(X_2) \\ \text{M} \\ E(X_k) \end{bmatrix} \tag{2.111}$$

$$\mathbf{K} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \text{L} & \sigma_{1k} \\ \sigma_{21} & \sigma_{22} & \text{L} & \sigma_{2k} \\ \text{M} & \text{M} & \text{O} & \text{M} \\ \sigma_{k1} & \sigma_{k2} & \text{L} & \sigma_{kk} \end{bmatrix}, \qquad \sigma_{ij} = COV(X_i, X_j) \tag{2.112}$$

Več podrobnosti o strukturi pravkar izpeljanih izrazov si lahko bralec pogleda v literaturi [Hsu].

2.15 Mejne porazdelitve

Poglejmo si naslednjo definicijo:

Če imamo opravka s spremenljivkami $X_1, X_2, ..., X_q, X_{q+1}, ...X_k$, *to je s k diskretnimi naključnimi spremenljivkami, ki imajo funkcijo združene porazdelitve verjetnosti* $p(x_1, x_2, ..., x_q, x_{q+1}, ...x_k)$, *potem je funkcija* ***mejne*** *združene porazdelitve verjetnosti spremenljivk* $X_1, X_2, ..., X_q$ *enaka [Hsu]:*

$$p_{x_1 \ldots x_q}(x_1, \ldots, x_q) = \sum_{x_{q+1}} \text{K} \sum_{x_n} p_{x_1 \ldots x_n}(x_1, \ldots, x_n) \tag{2.113}$$

kar lahko enostavneje zapišemo kot:

$$p_{12\text{K}\,q}(x_1, \ldots, x_q) = \sum_{x_{q+1}} \text{K} \sum_{x_n} p(x_1, \text{K}\,, x_n) \tag{2.114}$$

Podobno velja za zvezne naključne spremenljivke, kjer imamo definicijo:

Če imamo opravka s spremenljivkami $X_1, X_2, ..., X_q, X_{q+1}, ...X_k$, *to je s k zveznimi naključnimi spremenljivkami, ki imajo funkcijo združene porazdelitve gostote verjetnosti* $f\left(x_1, x_2, ..., x_q, x_{q+1}, ...x_k\right)$, *potem je funkcija* ***mejne*** *združene porazdelitve gostote verjetnosti spremenljivk* $X_1, X_2, ..., X_q$ *enaka [Hsu]:*

$$f_{12K\,q}\left(x_1, \ldots, x_q\right) = \int_{-\infty}^{\infty} \mathrm{K} \int_{-\infty}^{\infty} f\left(x_1, \mathrm{K}\ , x_n\right) \cdot dx_{q+1} \cdot \mathrm{K}\ \cdot dx_n \tag{2.115}$$

Poglejmo si naslednji primer:

Primer 2.12.: *Naj bodo spremenljivke X, Y, Z tri združeno porazdeljene naključne spremenljivke z naslednjo funkcijo združene porazdelitve gostote verjetnosti:*

$$f\left(x, y, z\right) = \begin{cases} K \cdot \left(x^2 + y \cdot z\right) & 0 \le x \le 1,\ 0 \le y \le 1,\ 0 \le z \le 1 \\ 0 & \text{sicer} \end{cases} \tag{2.116}$$

a) Poiščite vrednost za K.
b) Ugotovite mejno porazdelitev za spremenljivko X.
c) Ugotovite združeno mejno porazdelitev za par spremenljivk: X, Y.

a) Najprej bomo izračunali vrednost za parameter *K*. Na osnovi 2. lastnosti izraza (2.107) velja:

$$\begin{gathered} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f\left(x, y, z\right) dxdydz = 1 \\ \int_0^1 \int_0^1 \int_0^1 K\left(x^2 + yz\right) dxdydz = 1 \end{gathered} \tag{2.117}$$

Odtod sledi:

$$K\int_0^1\int_0^1\left[\frac{x^3}{3}+xyz\right]_{x=0}^{x=1}dydz = K\int_0^1\int_0^1\left(\frac{1}{3}+yz\right)dydz$$

$$K\int_0^1\left[\frac{1}{3}y+z\frac{y^2}{2}\right]_{y=0}^{y=1}dz = K\int_0^1\left(\frac{1}{3}+z\frac{1}{2}\right)dz \qquad (2.118)$$

$$K\left[\frac{z}{3}+\frac{z^2}{4}\right]_0^1 = K\left(\frac{1}{3}+\frac{1}{4}\right) = K\frac{7}{12} = 1$$

$$K = \frac{12}{7}$$

b) Na osnovi izraza (2.115) lahko zapišemo:

$$f_1(x) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y,z)\,dydz \qquad (2.119)$$

oziroma:

$$f_1(x) = \frac{12}{7}\cdot\int_0^1\int_0^1\left(x^2+yz\right)dydz \qquad (2.120)$$

Izraz (2.120) bomo izpeljali do konca:

$$f_1(x) = \frac{12}{7}\cdot\int_0^1\left[x^2y+\frac{y^2}{2}z\right]_{y=0}^{y=1}dz = \frac{12}{7}\cdot\int_0^1\left(x^2+\frac{1}{2}z\right)dz$$

$$f_1(x) = \frac{12}{7}\cdot\left[x^2z+\frac{z^2}{4}\right]_0^1 = \frac{12}{7}\cdot\left(x^2+\frac{1}{4}\right) \qquad \text{pri } 0\le x\le 1 \qquad (2.121)$$

Na podoben način bi lahko izračunali tudi mejno porazdelitev $f_2(y)$ za spremenljivko *Y* oz. $f_3(z)$ za spremenljivko *Z*.

c) Zopet izhajamo iz izraza (2.115) in tako lahko zapišemo:

$$f_{12}(x,y)=\int_{-\infty}^{\infty} f(x,y,z)\,dz=\frac{12}{7}\int_{0}^{1}\left(x^2+yz\right)dz$$

$$f_{12}(x,y)=\frac{12}{7}\left[x^2z+y\frac{z^2}{2}\right]_{z=0}^{z=1} \qquad (2.122)$$

$$f_{12}(x,y)=\frac{12}{7}\left(x^2+\frac{1}{2}y\right) \qquad \text{pri } 0\le x\le 1,\quad 0\le y\le 1$$

Na podoben način bi lahko npr. izračunali tudi združeno mejno porazdelitev $f_{13}(x,z)$ za spremenljivki X in Z, ter $f_{23}(y,z)$ za spremenljivki Y in Z.

2.16 Pogojne porazdelitve

Poglejmo si naslednjo definicijo:

Če imamo opravka s spremenljivkami $X_1, X_2, ..., X_q, X_{q+1}, ...X_k$, *to je s k diskretnimi naključnimi spremenljivkami, ki imajo funkcijo združene porazdelitve verjetnosti* $p(x_1, x_2, ..., x_q, x_{q+1}, ...x_k)$ *, potem je funkcija* ***pogojne*** *združene porazdelitve verjetnosti spremenljivk* $X_{q+1}, X_{q+2}, ..., X_k$ *pri danih* $X_1 = x_1, X_2 = x_2, ..., X_q = x_q$ *enaka [Hsu]:*

$$p_{q+1\mathrm{K}\,k|1\ldots q}\left(x_{q+1},\ldots,x_k\,\middle|\,x_1,\ldots,x_q\right)=\frac{p\left(x_1,\mathrm{K}\ ,x_k\right)}{p_{1\mathrm{K}\,q}\left(x_1,\ldots,x_q\right)} \qquad (2.123)$$

pri čemer je $p_{12\mathrm{K}\,q}(x_1,\ldots,x_q)$ funkcija mejne združene porazdelitve verjetnosti spremenljivk $X_1, X_2, ..., X_q$, podana z izrazom (2.114).

Podobno velja za zvezne naključne spremenljivke, kjer imamo definicijo:

Če imamo opravka s spremenljivkami $X_1, X_2, ..., X_q, X_{q+1}, ...X_k$, *to je s k zveznimi naključnimi spremenljivkami, ki imajo funkcijo združene porazdelitve gostote verjetnosti* $f\left(x_1, x_2, ..., x_q, x_{q+1}, ...x_k\right)$, *potem je funkcija* ***pogojne*** *združene porazdelitve gostote verjetnosti spremenljivk* $X_{q+1}, X_{q+2}, ..., X_k$ *pri danih* $X_1 = x_1, X_2 = x_2, ..., X_q = x_q$ *enaka [Hsu]:*

$$f_{q+1\mathrm{K}\ k|1\ldots q}\left(x_{q+1}, \ldots, x_k \middle| x_1, \ldots, x_q\right) = \frac{f\left(x_1, \mathrm{K}\ , x_k\right)}{f_{1\mathrm{K}\ q}\left(x_1, \ldots, x_q\right)} \tag{2.124}$$

pri čemer je $f_{12\mathrm{K}\ q}\left(x_1, \ldots, x_q\right)$ funkcija mejne združene porazdelitve gostote verjetnosti spremenljivk $X_1, X_2, ..., X_q$, podana z izrazom (2.115).

Poglejmo si naslednji primer:

Primer 2.13.: *Naj bodo spremenljivke X, Y, Z tri združeno porazdeljene naključne spremenljivke z naslednjo funkcijo združene porazdelitve gostote verjetnosti:*

$$f\left(x, y, z\right) = \begin{cases} \frac{12}{7} \cdot \left(x^2 + y \cdot z\right) & 0 \le x \le 1,\ 0 \le y \le 1,\ 0 \le z \le 1 \\ 0 & \text{sicer} \end{cases} \tag{2.125}$$

a) Poiščite pogojno porazdelitev za spremenljivko Z pri danih spremenljivkah X=x in Y=y.
b) Poiščite pogojno porazdelitev za spremenljivki Y, Z pri dani spremenljivki X=x.

a) Na osnovi prejšnjega primera vemo, da velja izraz (2.122) za združeno mejno porazdelitev para spremenljivk *X* in *Y*. Tako lahko na osnovi izraza (2.124) zapišemo:

$$f_{3|12}\left(z|x,y\right)=\frac{f\left(x,y,z\right)}{f_{12}\left(x,y\right)}=\frac{\frac{12}{7}\left(x^2+yz\right)}{\frac{12}{7}\left(x^2+\frac{1}{2}y\right)}=\frac{x^2+yz}{x^2+\frac{1}{2}y}\qquad \text{pri}\ \ 0\le z\le 1 \tag{2.126}$$

Seveda tudi velja: $0\le x\le 1,\quad 0\le y\le 1.$

b) Na osnovi prejšnjega primera vemo, da velja izraz (2.121) za mejno porazdelitev spremenljivke *X*. Tako lahko na osnovi izraza (2.124) zapišemo:

$$f_{23|1}\left(y,z|x\right)=\frac{f\left(x,y,z\right)}{f_1\left(x\right)}=\frac{\frac{12}{7}\left(x^2+yz\right)}{\frac{12}{7}\left(x^2+\frac{1}{4}\right)}=\frac{x^2+yz}{x^2+\frac{1}{4}}\quad \text{pri}\ \ 0\le y\le 1,\ 0\le z\le 1 \tag{2.127}$$

Seveda tudi velja: $0\le x\le 1.$

2.17 Neodvisnost vektorjev naključnih spremenljivk

Poglejmo si naslednjo definicijo:

Če imamo opravka s spremenljivkami $X_1,X_2,...,X_q,X_{q+1},...X_k$, *to je s k zveznimi naključnimi spremenljivkami, ki imajo funkcijo združene porazdelitve gostote verjetnosti* $f\left(x_1,x_2,...,x_q,x_{q+1},...x_k\right)$, *potem so spremenljivke* $X_1,X_2,...,X_q$ ***<u>neodvisne</u>*** *od spremenljivk* $X_{q+1},X_{q+2},...,X_k$, *če velja [Hsu]:*

$$f\left(x_1,\text{K},x_k\right)=f_{\text{lK}\,q}\left(x_1,\ldots,x_q\right)\cdot f_{q+\text{lK}\,k}\left(x_{q+1},\ldots,x_k\right) \tag{2.128}$$

Podobna definicija velja tudi za diskretne naključne spremenljivke.

Poglejmo si še eno definicijo:

Če imamo opravka s spremenljivkami $X_1, X_2, ..., X_k$, *to je s k zveznimi naključnimi spremenljivkami, ki imajo funkcijo združene porazdelitve gostote verjetnosti* $f(x_1, x_2, ..., x_k)$, *potem so spremenljivke* $X_1, X_2, ..., X_k$ ***vzajemno neodvisne*** , *če velja [Hsu]:*

$$f(x_1, \text{K}\ , x_k) = f_1(x_1) \cdot f_2(x_2) \cdot \ldots \cdot f_k(x_k) \tag{2.129}$$

Podobna definicija velja tudi za diskretne naključne spremenljivke.

Poglejmo si naslednji primer:

Primer 2.14.: *Za naključni spremenljivki X in Y imamo dano naslednjo združeno porazdelitev gostote verjetnosti*:

$$f(x,y) = \begin{cases} \frac{1}{8} \cdot (x+y), & 0 < x < 2, \quad 0 < y < 2 \\ 0 & \textit{sicer} \end{cases} \tag{2.130}$$

Ali sta spremenljivki X in Y medsebojno neodvisni?

Za medsebojno neodvisnost mora veljati:

$$f(x,y) = f(x) \cdot f(y) \tag{2.131}$$

torej moramo najprej poiskati mejni porazdelitvi $f(x)$ in $f(y)$. Najprej izračunamo mejno porazdelitev za spremenljivko *X*:

$$\begin{aligned} f(x) &= \int_{-\infty}^{\infty} f(x,y)\,dy = \int_0^2 \frac{1}{8} \cdot (x+y)\,dy = \frac{1}{8} \cdot \left(xy + \frac{y^2}{2} \right)_{y=0}^{y=2} = \\ &= \frac{1}{8} \cdot \left(x \cdot 2 + \frac{4}{2} \right) = \frac{1}{4} \cdot (x+1), \quad \textit{pri} \quad 0 < x < 2 \end{aligned} \tag{2.132}$$

Potem izračunamo še mejno porazdelitev za spremenljivko *Y*:

$$f(y)=\int_{-\infty}^{\infty} f(x,y)\,dx=\int_{0}^{2}\frac{1}{8}\cdot(x+y)\,dx=\frac{1}{8}\cdot\left(\frac{x^2}{2}+xy\right)_{x=0}^{x=2}= \\ =\frac{1}{8}\cdot\left(\frac{4}{2}+2\cdot y\right)=\frac{1}{4}\cdot(y+1),\quad pri\quad 0<y<2 \tag{2.133}$$

Očitno velja:

$$f(x,y)\neq f(x)\cdot f(y) \tag{2.134}$$

in zato spremenljivki *X* in *Y* nista medsebojno neodvisni (sta odvisni).

2.18 Pričakovanje za porazdelitve več spremenljivk

Poglejmo si naslednjo definicijo:

Če imamo opravka s spremenljivkami $X_1, X_2, ..., X_n$, *to je z n zveznimi naključnimi spremenljivkami, ki imajo funkcijo združene porazdelitve gostote verjetnosti* $f(x_1, x_2, ..., x_n)$, *potem za matematično upanje funkcije* $g(X_1, X_2, ..., X_n)$ *teh spremenljivk velja [Hsu]:*

$$E\left[g(X_1,\text{K},X_n)\right]=\int_{-\infty}^{\infty}\text{K}\int_{-\infty}^{\infty} g(x_1,\text{K},x_n)\cdot f(x_1,\text{K},x_n)\,dx_1\cdot\text{K}\cdot dx_n \tag{2.135}$$

Poglejmo si naslednji primer:

Primer 2.15.: *Naj bodo X, Y, Z tri združeno porazdeljene naključne spremenljivke s funkcijo združene porazdelitve gostote verjetnosti:*

$$f(x,y,z)=\begin{cases}\frac{12}{7}\left(x^2+yz\right) & 0\le x\le 1,\ 0\le y\le 1,\ 0\le z\le 1\\ 0 & \text{sicer}\end{cases} \tag{2.136}$$

Poiščite matematično upanje $E\left(g(X,Y,Z)\right)=E(X\cdot Y\cdot Z)$*!*

Na osnovi izraza (2.135) lahko zapišemo:

$$\begin{aligned} E\left[g\left(X,Y,Z\right)\right] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g\left(x,y,z\right)\cdot f\left(x,y,z\right)dxdydz \\ E\left[XYZ\right] &= \int_{0}^{1}\int_{0}^{1}\int_{0}^{1} xyz\frac{12}{7}\left(x^2+yz\right)dxdydz \end{aligned} \tag{2.137}$$

Odtod sledi:

$$E\left[XYZ\right] = \frac{12}{7}\int_{0}^{1}\int_{0}^{1}\int_{0}^{1}\left(x^3yz+xy^2z^2\right)dxdydz \tag{2.138}$$

Izraz (2.138) bomo poskušali rešiti s trojno integracijo, kar nam bo dalo:

$$\begin{aligned} E\left[XYZ\right] &= \frac{12}{7}\int_{0}^{1}\int_{0}^{1}\left[\frac{x^4}{4}yz+\frac{x^2}{2}y^2z^2\right]_{x=0}^{x=1}dydz = \frac{12}{7}\int_{0}^{1}\int_{0}^{1}\left[\frac{1}{4}yz+\frac{1}{2}y^2z^2\right]dydz = \\ &= \frac{3}{7}\int_{0}^{1}\int_{0}^{1}\left(yz+2y^2z^2\right)dydz = \frac{3}{7}\int_{0}^{1}\left[\frac{y^2}{2}z+2\frac{y^3}{3}z^2\right]_{y=0}^{y=1}dz = \frac{3}{7}\int_{0}^{1}\left(\frac{1}{2}z+\frac{2}{3}z^2\right)dz = \\ &= \frac{3}{7}\left[\frac{z^2}{4}+\frac{2z^3}{9}\right]_0^1 = \frac{3}{7}\left(\frac{1}{4}+\frac{2}{9}\right) = \frac{3}{7}\left(\frac{17}{36}\right) = \frac{17}{84} \end{aligned} \tag{2.139}$$

V nadaljevanju si bomo pogledali še nekatera pravila za računanje pričakovanih vrednosti [Dragan 2].

Denimo imamo dano funkcijo združene porazdelitve gostote verjetnosti $f\left(x_1,x_2,...,x_n\right)$. Zanima nas, kako izračunati $E\left[g\left(X_1,\mathrm{K}\ ,X_n\right)\right]$ v primeru, če je $g\left(X_1,\mathrm{K}\ ,X_n\right)=X_1$, torej želimo izračunati $E\left[X_1\right]$. Na osnovi izraza (2.135) lahko zapišemo:

$$\begin{aligned} E\left[g\left(X_1,\mathrm{K}\ ,X_n\right)\right] &= \int_{-\infty}^{\infty}\mathrm{K}\int_{-\infty}^{\infty} g\left(x_1,\mathrm{K}\ ,x_n\right)\cdot f\left(x_1,\mathrm{K}\ ,x_n\right)dx_1\cdot\mathrm{K}\cdot dx_n \\ E\left[X_1\right] &= \int_{-\infty}^{\infty}\mathrm{K}\int_{-\infty}^{\infty} x_1\cdot f\left(x_1,\mathrm{K}\ ,x_n\right)dx_1\mathrm{K}\ dx_n \end{aligned} \tag{2.140}$$

Pokazati se da, da je ta izraz enak izrazu:

$$E[X_1] = \int_{-\infty}^{\infty} x_1 \cdot f_1(x_1)\, dx_1 \tag{2.141}$$

Torej lahko izračunamo $E[X_1]$ bodisi s pomočjo združene porazdelitve $f(x_1, x_2, ..., x_n)$, bodisi s pomočjo mejne porazdelitve za X_1, to je s pomočjo $f_1(x_1)$.

Poglejmo si dokaz, da je izraz (2.140) res enak izrazu (2.141):

$$\begin{aligned}
&\int_{-\infty}^{\infty} \mathrm{K} \int_{-\infty}^{\infty} x_1 \cdot f(x_1, \mathrm{K}, x_n)\, dx_1 \mathrm{K}\, dx_n = \\
&= \int_{-\infty}^{\infty} x_1 \left[\int_{-\infty}^{\infty} \mathrm{K} \int_{-\infty}^{\infty} \underbrace{f(x_1, \mathrm{K}, x_n)\, dx_n}_{f(x_1, \mathrm{K}, x_{n-1})} \mathrm{K}\, dx_2 \right] dx_1 = \\
&= \int_{-\infty}^{\infty} x_1 \left[\int_{-\infty}^{\infty} \mathrm{K} \int_{-\infty}^{\infty} \underbrace{f(x_1, \mathrm{K}, x_{n-1})\, dx_{n-1}}_{f(x_1, \mathrm{K}, x_{n-2})} \mathrm{K}\, dx_2 \right] dx_1 = \ldots = \\
&= \int_{-\infty}^{\infty} x_1 \left[\underbrace{\int_{-\infty}^{\infty} f(x_1, x_2) dx_2}_{f_1(x_1)} \right] dx_1 = \int_{-\infty}^{\infty} x_1 \cdot f_1(x_1)\, dx_1
\end{aligned} \tag{2.142}$$

Na podoben način lahko storimo posplošitev za *n* naključnih spremenljivk. To pomeni, da lahko izračunamo matematično upanje *i*-te naključne spremenljivke $E(X_i)$ bodisi iz združene porazdelitve za spremenljivke X_1, K, X_n, bodisi iz mejne porazdelitve za spremenljivko X_i:

$$E[X_i] = \int_{-\infty}^{\infty} \mathrm{K} \int_{-\infty}^{\infty} x_i \cdot f(x_1, \mathrm{K}, x_n)\, dx_1 \mathrm{K}\, dx_n = \int_{-\infty}^{\infty} x_i \cdot f_i(x_i)\, dx_i \tag{2.143}$$

To zakonitost bomo, kot bo razvidno, koristno uporabili kasneje v nadaljevanju.

Kot smo videli že v poglavju 2.11, v izrazu (2.78), za matematično upanje linearne kombinacije *n* naključnih spremenljivk velja lastnost linearnosti, ki pravi:

$$E\left[a_1X_1+\mathrm{L}\ +a_nX_n\right]=a_1E\left[X_1\right]+\mathrm{L}\ +a_nE\left[X_n\right] \tag{2.144}$$

Podajmo dokaz, da res velja izraz (2.144):

$$E\left[a_1X_1+\mathrm{L}\ +a_nX_n\right]=\int_{-\infty}^{\infty}\mathrm{K}\int_{-\infty}^{\infty}\left(a_1x_1+\mathrm{K}\ +a_nx_n\right)\cdot f\left(x_1,\mathrm{K}\ ,x_n\right)dx_1\mathrm{K}\ dx_n= \tag{2.145}$$

$$=a_1\int_{-\infty}^{\infty}\mathrm{K}\int_{-\infty}^{\infty}x_1\cdot f\left(x_1,\mathrm{K}\ ,x_n\right)dx_1\mathrm{K}\ dx_n+\ldots+a_n\int_{-\infty}^{\infty}\mathrm{K}\int_{-\infty}^{\infty}x_n\cdot f\left(x_1,\mathrm{K}\ ,x_n\right)dx_1\mathrm{K}\ dx_n$$

Če upoštevamo še izraz (2.143), sledi:

$$E\left[a_1X_1+\mathrm{L}\ +a_nX_n\right]=a_1\int_{-\infty}^{\infty}x_1f_1\left(x_1\right)dx_1+\ldots+a_n\int_{-\infty}^{\infty}x_nf_n\left(x_n\right)dx_n= \tag{2.146}$$

$$=a_1E\left[X_1\right]+\mathrm{L}\ +a_nE\left[X_n\right]$$

in tako smo dokazali, da velja linearnost.

V nadaljevanju si poglejmo še lastnost multiplikativnosti:

Če imamo opravka s spremenljivkami $X_1,X_2,\ldots,X_q$, *to je s q naključnimi spremenljivkami, ki so neodvisne od spremenljivk* $X_{q+1},X_{q+2},\ldots,X_k$, *potem velja naslednji izraz:*

$$\begin{aligned}&E\left[g_{1\ldots k}\left(X_1,\mathrm{K}\ ,X_q,X_{q+1},\mathrm{K}\ ,X_k\right)\right]=\\&=E\left[g_{1\ldots q}\left(X_1,\mathrm{K}\ ,X_q\right)\cdot g_{q+1\ldots k}\left(X_{q+1},\mathrm{K}\ ,X_k\right)\right]=\\&=E\left[g_{1\ldots q}\left(X_1,\mathrm{K}\ ,X_q\right)\right]\cdot E\left[g_{q+1\ldots k}\left(X_{q+1},\mathrm{K}\ ,X_k\right)\right]\end{aligned} \tag{2.147}$$

V preprostem primeru za *k = 2*, ko imamo torej le dve neodvisni naključni spremenljivki X_1 in X_2, izraz (2.147) preide v obliko:

$$E\left[g_{12}(X_1,X_2)\right]=E\left[g_1(X_1)\cdot g_2(X_2)\right]=E\left[g_1(X_1)\right]\cdot E\left[g_2(X_2)\right] \tag{2.148}$$

Poglejmo si naslednji primer:

Primer 2.16.: *Denimo velja:*

$$\begin{aligned} &g_1(X_1)=X_1 \\ &g_2(X_2)=X_2 \\ &X_1 \ in \ \ X_2 \ \ sta \ \ neodvisna. \end{aligned} \tag{2.149}$$

Pokažite, da velja izraz (2.148).

Napišemo lahko:

$$\begin{aligned} &E\left[g_{12}(X_1,X_2)\right]=E\left[g_1(X_1)\cdot g_2(X_2)\right]=E\left[X_1\cdot X_2\right]= \\ &=E\left[g_1(X_1)\right]\cdot E\left[g_2(X_2)\right]=E\left[X_1\right]\cdot E\left[X_2\right] \end{aligned} \tag{2.150}$$

Podajmo dokaz za veljavnost izraza (2.150):

$$\begin{aligned} &E\left[g_{12}(X_1,X_2)\right]=E\left[X_1\cdot X_2\right]=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}x_1\cdot x_2\cdot \underbrace{f_{12}(x_1,x_2)}_{f_1(x_1)\cdot f_2(x_2)}dx_1dx_2= \\ &=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}x_1\cdot x_2\cdot f_1(x_1)\cdot f_2(x_2)\,dx_1dx_2= \\ &=\int_{-\infty}^{\infty}x_1f_1(x_1)\,dx_1\cdot\int_{-\infty}^{\infty}x_2f_2(x_2)\,dx_2=E\left[X_1\right]E\left[X_2\right] \end{aligned} \tag{2.151}$$

2.19 Maksimalna podobnost

Začnimo razlago tovrstne problematike v obliki naslednjega problema [Dragan 2]:

Imamo na razpolagi podatke o nekem (naključnem) procesu. Na primer, potek temperature v sobi, ali potek borznega indeksa. Predpostavimo, da poznamo model oz. mehanizem, ki je "proizvedel" podatke, ne poznamo pa njegovih parametrov.

Metoda **maksimalne (največje) podobnosti (verjetja)** (ang. **Maximum Likelihood**) nam pomaga pri reševanju tovrstnih problemov.

Problematiko bomo dodatno osvetlili in razložili na naslednjih dveh primerih [Dragan 2]:

1. primer (primer 2.17.):

Nekdo opravi N zaporednih metov kovanca. Predpostavimo, da je padlo n pisem in N-n številk. Kolikšna je verjetnost q, da bo pri novem metu kovanca padlo pismo?

2. primer (primer 2.18.):

Objekt neznane mase m tehtamo N-krat. Zaradi omejene natančnosti tehtnice, kakor tudi zaradi naključnih vplivov pri izvedbi poskusa, dobimo pri merjenju vsakič nekoliko drugačen rezultat. Zanima nas optimalna ocena mase m in ocena variance naključnih vplivov.

1. Najprej poskusimo rešiti 1. primer:

Predpostavimo, da je verjetnost, da pri metu kovanca pade pismo, enaka *q*. Torej je verjetnost, da pade številka, enaka *1-q*. Predpostavimo, da so meti med seboj statistično neodvisni. Označimo z y_k izid meta z zaporedno številko *k*. Verjetnost za realizacijo vseh *N* metov je potem enaka:

$$p(y_1, y_2, ..., y_N) = p(y_1) \cdot p(y_2) \cdot ... \cdot p(y_N) \qquad (2.152)$$

Od *N*-tih izidov pade *n* pisem. Torej se bo v produktu na desni strani izraza (2.152) *n*-krat pojavil faktor *q*. Namreč, za verjetnost za dogodek "pade pismo" velja:

$$p(y_i = 'pade\ \ pismo') = q \qquad (2.153)$$

Podobno sklepamo, da bo od preostalih izidov padlo *N-n* številk. Torej se bo v produktu na desni strani izraza (2.152) *N-n* - krat pojavil faktor *1-q*. Namreč, za verjetnost za dogodek "pade številka" velja:

$$p(y_i = \text{' pade } \check{s}tevilka'}) = 1 - q \tag{2.154}$$

Zato izraz (2.152) preide v naslednjo obliko:

$$\begin{aligned} p(y_1, y_2, ..., y_N) &= \underbrace{q \cdot q \cdot \ldots \cdot q}_{n-krat} \cdot \underbrace{(1-q) \cdot (1-q) \cdot \ldots \cdot (1-q)}_{N-n\ -krat} \\ p(y_1, y_2, ..., y_N) &= q^n \cdot (1-q)^{N-n} = L(q) \end{aligned} \tag{2.155}$$

Torej je združena gostota verjetnosti za realizacijo dogodka $\{y_1, y_2, ..., y_N\} = \{y_1 \cap y_2 \cap ... \cap y_N\}$ neka funkcija $L(q)$ z argumentom q. Funkcija L se imenuje funkcija največje podobnosti. Optimalna vrednost argumenta q pa je tista vrednost q^*, pri kateri $L(q^*)$ doseže maksimalno vrednost. Zato se postopek imenuje metoda maksimalne podobnosti (ali maksimalnega verjetja) [Dragan 2].

Za nastop ekstrema (maksimuma) funkcije v izrazu (2.155) se potreben pogoj glasi:

$$\frac{dL(q)}{dq} = 0 \tag{2.156}$$

Izračunajmo ta odvod:

$$\begin{aligned} &\frac{dL(q)}{dq} = \frac{d}{dq}\left[q^n (1-q)^{N-n}\right] = \frac{d}{dq}\left[q^n\right] \cdot (1-q)^{N-n} + q^n \cdot \frac{d}{dq}\left[(1-q)^{N-n}\right] = \\ &= n \cdot q^{n-1} \cdot (1-q)^{N-n} + q^n \cdot (N-n) \cdot (1-q)^{N-n-1} \cdot (-1) = \\ &= n \cdot q^{n-1} \cdot (1-q)^{N-n} - q^n \cdot (N-n) \cdot (1-q)^{N-n-1} \end{aligned} \tag{2.157}$$

Izenačimo izraz (2.157) z 0 in dobimo:

$$\begin{aligned} & n \cdot q^{n-1} \cdot (1-q)^{N-n} - q^n \cdot (N-n) \cdot (1-q)^{N-n-1} = 0 \\ & (1-q)^{N-n-1} \cdot \left(n \cdot q^{n-1} \cdot (1-q) - q^n \cdot (N-n) \right) = 0 \\ & n \cdot q^{n-1} \cdot (1-q) - q^n \cdot (N-n) = 0 \\ & n \cdot (1-q) - q \cdot (N-n) = 0 \\ & n - nq - Nq + nq = 0 \\ & n - Nq = 0 \end{aligned} \tag{2.158}$$

Torej velja, da mora biti ocena verjetnosti za faktor *q* enaka:

$$\hat{q} = \frac{n}{N} \tag{2.159}$$

Seveda mora v primeru, če je kovanec "pošten" (kar ni nujno res), veljati:

$$\lim_{N \to \infty} \frac{n}{N} = \frac{1}{2} \tag{2.160}$$

2. Poskušajmo rešiti še 2. primer:

Primer tehtanja je morda eden najenostavnejših primerov, s katerim ponazorimo temeljni problem, kako iz podatkov priti do parametrov modela dogodkov. Na tehtanje lahko gledamo kot na naključni dogodek. Izid *k*-tega merjenja lahko zapišemo na naslednji način:

$$y_k = m + e_k \tag{2.161}$$

pri čemer je *m* dejanska, vendar neznana masa predmeta, e_k pa naključna vrednost, ki se tekom merjenja prišteje k vrednosti *m*. Na ta način s spremenljivko e_k ponazorimo naključni izid eksperimenta. Predpostavimo lahko, da za spremenljivko e_k velja normalna porazdelitev s srednjo vrednostjo 0 in varianco σ^2. Potem na osnovi izraza (2.36) za porazdelitev gostote verjetnosti spremenljivke e_k velja naslednji izraz:

$$p(e_k) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{e_k^2}{2\sigma^2}} \tag{2.162}$$

ki z upoštevanjem izraza (2.161) preide v obliko:

$$p(y_k - m) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_k - m)^2}{2\sigma^2}} = p(y_k), \qquad k = 1,2,...,N \tag{2.163}$$

Pod predpostavko, da je vsaka meritev statistično neodvisna od drugih meritev, je funkcija združene gostote verjetnosti za nastop dogodka $\{y_1, y_2, ..., y_N\}$ enaka:

$$p(y_1, y_2, ..., y_N) = p(y_1) \cdot p(y_2) \cdot ... \cdot p(y_N) \tag{2.164}$$

Če upoštevamo izraz (2.163), izraz (2.164) preide v obliko:

$$L(m,\sigma) = p(y_1, y_2, ..., y_N) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_1 - m)^2}{2\sigma^2}} \cdot \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_2 - m)^2}{2\sigma^2}} \cdot ... \cdot \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_N - m)^2}{2\sigma^2}} \tag{2.165}$$
$$= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^N \cdot e^{-\frac{(y_1 - m)^2}{2\sigma^2} - \frac{(y_2 - m)^2}{2\sigma^2} - ... - \frac{(y_N - m)^2}{2\sigma^2}} = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^N \cdot e^{-\frac{(y_1 - m)^2 + (y_2 - m)^2 + ... + (y_N - m)^2}{2\sigma^2}}$$

Za nastop ekstrema (maksimuma) funkcije v izrazu (2.165) se potrebna pogoja glasita:

$$\frac{\partial L(m,\sigma)}{\partial m} = 0$$
$$\frac{\partial L(m,\sigma)}{\partial \sigma} = 0 \tag{2.166}$$

Če izračunamo prvi parcialni odvod, dobimo:

$$\begin{aligned}
&\frac{\partial}{\partial m}L(m,\sigma)=\frac{\partial}{\partial m}\left(\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^N\cdot e^{-\frac{(y_1-m)^2+(y_2-m)^2+\ldots+(y_N-m)^2}{2\sigma^2}}\right)=\\
&=\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^N\cdot e^{-\frac{(y_1-m)^2+(y_2-m)^2+\ldots+(y_N-m)^2}{2\sigma^2}}\cdot\frac{\partial}{\partial m}\left(-\frac{(y_1-m)^2+(y_2-m)^2+\ldots+(y_N-m)^2}{2\sigma^2}\right)=\\
&=\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^N\cdot e^{-\frac{(y_1-m)^2+(y_2-m)^2+\ldots+(y_N-m)^2}{2\sigma^2}}\cdot\left(\frac{2(y_1-m)+2(y_2-m)+\ldots+2(y_N-m)}{2\sigma^2}\right)=\\
&=\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^N\cdot e^{-\frac{(y_1-m)^2+(y_2-m)^2+\ldots+(y_N-m)^2}{2\sigma^2}}\cdot\left(\frac{(y_1-m)+(y_2-m)+\ldots+(y_N-m)}{\sigma^2}\right)
\end{aligned}\tag{2.167}$$

Če ta izraz enačimo z 0, dobimo:

$$\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^N\cdot e^{-\frac{(y_1-m)^2+(y_2-m)^2+\ldots+(y_N-m)^2}{2\sigma^2}}\cdot\left(\frac{(y_1-m)+(y_2-m)+\ldots+(y_N-m)}{\sigma^2}\right)=0 \tag{2.168}$$

$$(y_1-m)+(y_2-m)+\ldots+(y_N-m)=0$$

$$y_1+y_2+\ldots+y_N-m\cdot N=0$$

Torej velja, da mora biti ocena za maso *m* enaka:

$$\hat{m}=\frac{\sum_{i=1}^{N}y_i}{N} \tag{2.169}$$

Na podoben način bi nato izračunali še $\frac{\partial L(m,\sigma)}{\partial\sigma}=0$. Po daljši izpeljavi bi dobili naslednji rezultat [Dragan 2]:

$$\hat{\sigma}^2=\frac{\sum_{i=1}^{N}(y_i-\hat{m})^2}{N} \tag{2.170}$$

pri čemer bi $\hat{m}$ seveda izračunali s pomočjo izraza (2.169).

2.20 Momenti naključnih spremenljivk

Predmet tega poglavja je skupina številskih karakteristik, imenovanih momenti. Mednje spadata tudi matematično upanje in varianca.

Definicija [Jamnik 3]:

Naj bo r poljubno nenegativno celo število. Če obstaja veličina:

$$\mu_r = E\left[\left(X - E(X)\right)^r\right] = E\left[\left(X - \mu\right)^r\right] \tag{2.171}$$

jo imenujemo ***r-ti centralni moment*** *(****centralni moment reda r****)naključne spremenljivke X glede na vrednost* μ *(matematično upanje). Pri naključnih spremenljivkah, ki nimajo določene srednje vrednosti, centralni moment ni določljiv.*

Pri diskretnih naključnih spremenljivkah je centralni moment enak:

$$\mu_r = \sum_{i=0}^{m} \left(x_i - \mu\right)^r \cdot p\left(x_i\right) \tag{2.172}$$

pri zveznih pa je enak:

$$\mu_r = \int_{-\infty}^{\infty} \left(x - \mu\right)^r \cdot f\left(x\right) dx \tag{2.173}$$

Včasih je bolj ugodno, da uporabljamo moment glede na izhodišče (vrednost 0), ki mu pravimo tudi **začetni moment (reda *r*)**, ne pa glede na matematično upanje.

Začetni moment reda *r* je definiran za diskretne naključne spremenljivke na naslednji način:

$$\mu_r' = E\left(X^r\right) = \sum_{i=0}^{m} \left(x_i\right)^r \cdot p\left(x_i\right) \tag{2.174}$$

za zvezne spremenljivke pa:

$$\mu_r' = E\left(X^r\right) = \int_{-\infty}^{\infty} (x)^r \cdot f(x)\,dx \tag{2.175}$$

Povezava med začetnim in centralnim momentom je naslednja [Jamnik 3, Jesenko]:

$$\mu_r = \sum_{k=0}^{r} (-1)^{r-k} \binom{r}{k} (\mu)^{r-k} \mu_k' \tag{2.176}$$

torej je centralni momentt linearna kombinacija začetnih momentov. Izraz (2.176) je uporaben, saj je marsikje začetne momente laže računati kot pa centralne momente.

Glede na izraz (2.171) za prve tri centralne momente velja:

$$\begin{aligned}
&\mu_0 = E\left[\left(X - E(X)\right)^0\right] = E\left[\left(X-\mu\right)^0\right] = E[1] = 1 \\
&\mu_1 = E\left[\left(X - E(X)\right)^1\right] = E\left[\left(X-\mu\right)^1\right] = E[X] - E(\mu) = E[X] - E\left(E[X]\right) = 0 \\
&\mu_2 = E\left[\left(X - E(X)\right)^2\right] = E\left[\left(X-\mu\right)^2\right] = E\left[(X)^2 - 2X\mu + \mu^2\right] = \\
&= E\left(X^2\right) - 2\mu E(X) + E\left[\mu^2\right] = E\left(X^2\right) - 2E^2(X) + \mu^2 E[1] = \\
&= E\left(X^2\right) - E^2(X) = VAR(X)
\end{aligned} \tag{2.177}$$

Za prve tri začetne momente pa velja:

$$\begin{aligned}
&\mu_0' = E\left(X^0\right) = E(1) = 1 \\
&\mu_1' = E\left(X^1\right) = E(X) = \mu \\
&\mu_2' = E\left(X^2\right)
\end{aligned} \tag{2.178}$$

V nadaljevanju izrazimo drugi in tretji centralni moment z začetnimi momenti na osnovi izraza (2.176):

$$\begin{aligned}
&\mu_2 = \sum_{k=0}^{2} (-1)^{2-k} \binom{2}{k} (\mu)^{2-k} \mu_k' = \\
&= (-1)^{2-0} \binom{2}{0} (\mu)^{2-0} \mu_0' + (-1)^{2-1} \binom{2}{1} (\mu)^{2-1} \mu_1' + (-1)^{2-2} \binom{2}{2} (\mu)^{2-2} \mu_2' = \\
&= \mu^2 - 2\mu \cdot \mu_1' + \mu_2' = \mu^2 - 2\mu^2 + \mu_2' = \mu_2' - \mu^2 = \mu_2' - \mu_1'^2 = \\
&= E\left(X^2\right) - E^2(X) = VAR(X)
\end{aligned} \tag{2.179}$$

$$\begin{aligned}
&\mu_3 = \sum_{k=0}^{3} (-1)^{3-k} \binom{3}{k} (\mu)^{3-k} \mu_k' = \\
&= (-1)^{3-0} \binom{3}{0} (\mu)^{3-0} \mu_0' + (-1)^{3-1} \binom{3}{1} (\mu)^{3-1} \mu_1' + \\
&+ (-1)^{3-2} \binom{3}{2} (\mu)^{3-2} \mu_2' + (-1)^{3-3} \binom{3}{3} (\mu)^{3-3} \mu_3' = \\
&= -(\mu)^3 + 3(\mu)^2 \mu_1' - 3\mu \cdot \mu_2' + \mu_3' = \\
&= \mu_3' - 3\mu \cdot \mu_2' + 2\mu^3
\end{aligned} \tag{2.180}$$

Na podoben način bi dobili tudi četrti centralni moment, ki ima obliko [Jesenko]:

$$\mu_4 = \mu_4' - 4\mu_3'\mu + 6\mu^2 \cdot \mu_2' - 3\mu^4 \tag{2.180A}$$

Kot se izkaže, imajo takoimenovane nesimetrične porazdelitve lihe centralne momente [Jamnik 3]. Zato je priporočljivo v meri za nesimetričnost vzeti kakšen moment lihega reda, najbolje kar najnižjega, torej μ_3 (saj je $\mu_1 = 0$). Da pa bomo izločili še razpršenost porazdelitve, za mero asimetrije vzamemo [Jamnik 3, Jesenko]:

$$g_1 = \frac{\mu_3}{\left(\sqrt{VAR(X)}\right)^3} \tag{2.181}$$

Četrti centralni moment μ_4 pa običajno uporabljamo za merjenje koničavosti ali sploščenosti porazdelitve realizacij naključne spremenljivke glede na matematično upanje [Jesenko]. Pri tem upeljemo koeficient sploščenosti [Jesenko]:

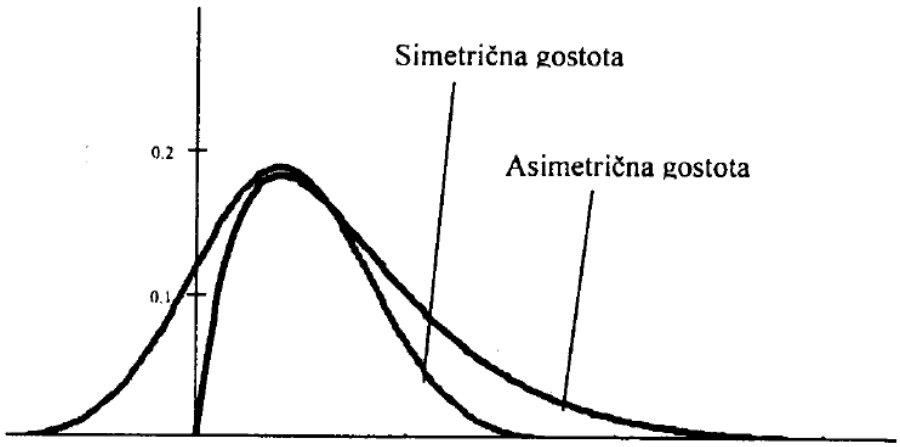
$$g_2 = \frac{\mu_4}{\left(\sqrt{VAR(X)}\right)^4} \tag{2.181A}$$

Vrednost tega koeficienta je vedno pozitivna. Večja ko je njegova vrednost, bolj so porazdelitve koničaste glede na matematično upanje, in obratno, manjše ko so njegove vrednosti, bolj so porazdelitve sploščene glede na matematično upanje.

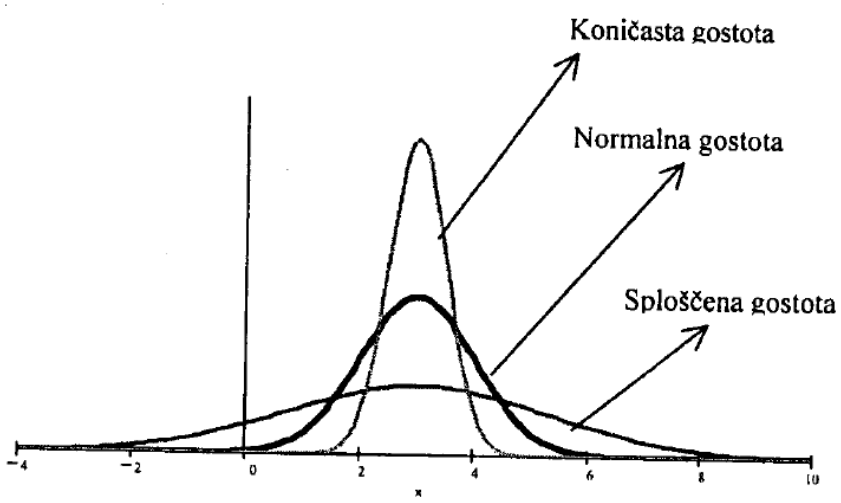
Velikokrat koeficient sploščenosti vpeljemo na takšen način, da primerjamo dano porazdelitev z normalno porazdelitvijo, pri čemer tvorimo koeficient:

$$\gamma = g_2 - 3 = \frac{\mu_4}{\left(\sqrt{VAR(X)}\right)^4} - 3 \qquad (2.181B)$$

Primer simetrične in nesimetrične porazdelitve gostote verjetnosti prikazuje slika 32, primer koničavosti in sploščenosti pa slika 33 [Jesenko].



Slika 32: Primer simetrične in nesimetrične porazdelitve gostote verjetnosti [Jesenko]



Slika 33: Primer koničavosti in sploščenosti porazdelitve gostote verjetnosti [Jesenko]

Primer 2.19.:

Izračunajte asimetrijo eksponentne porazdelitve: $f(x)=e^{-x}$!

Kot se izkaže, za začetne momente eksponentne porazdelitve velja:

$$\mu_r' = E\left(X^r\right) = \frac{r!}{\lambda^r} \tag{2.182}$$

Torej imamo:

$$\mu_2' = E\left(X^2\right) = \frac{2!}{\lambda^2} \tag{2.183}$$
$$\mu_3' = E\left(X^3\right) = \frac{3!}{\lambda^3}$$

Na osnovi slike 31 vemo, da velja:

$$E(X) = \mu = \frac{1}{\lambda} \tag{2.184}$$
$$VAR(X) = \frac{1}{\lambda^2}$$

Za asimetrijo torej velja:

$$g_1 = \frac{\mu_3' - 3\mu \cdot \mu_2' + 2\mu^3}{\left(\sqrt{VAR(X)}\right)^3} = \frac{\frac{3!}{\lambda^3} - 3\frac{1}{\lambda} \cdot \frac{2!}{\lambda^2} + 2\left(\frac{1}{\lambda}\right)^3}{\left(\sqrt{\frac{1}{\lambda^2}}\right)^3} =$$
$$= \frac{\frac{6}{\lambda^3} - \frac{6}{\lambda^3} + \frac{2}{\lambda^3}}{\frac{1}{\lambda^3}} = 2 \qquad (2.185)$$

Torej ima eksponentna porazdelitev ne glede na vrednost parametra λ vedno asimetrijo enako 2.

2.21 Rodovne funkcije momentov

Pri obravnavi naključnih spremenljivk se pokažejo za zelo koristne nekatere transformacije njihovih porazdelitvenih zakonov, kot npr. transformacija v **rodovne funkcije momentov (funkcije generiranja momentov)**.

Rodovna funkcija momentov je za diskretne naključne spremenljivke definirana na naslednji način [Hsu, Jesenko]:

$$M(t) = E\left(e^{t \cdot X}\right) = \sum_{i=0}^{m} e^{t \cdot x_i} \cdot p(x_i) \qquad (2.186)$$

in za zvezne spremenljivke:

$$M(t) = E\left(e^{t \cdot X}\right) = \int_{-\infty}^{\infty} e^{t \cdot x} \cdot f(x)\,dx \qquad (2.187)$$

kjer je *t* realna spremenljivka. Opozoriti moramo, da rodovna funkcija momentov ne obstaja za vse tipe naključnih spremenljivk.

Povezava med začetnim momentom reda *r* in rodovno funkcijo momentov je naslednja [Hsu, Jesenko]:

$$\mu_r' = E\left(X^r\right) = \left.\frac{d^r M(t)}{dt^r}\right|_{t=0} \tag{2.188}$$

Poglejmo si dokaz za izraz (2.188). V ta namen najprej razvijmo funkcijo $e^{t\cdot X}$ v Maclaurinovo vrsto:

$$e^{t\cdot X} = 1 + t\cdot X + \frac{1}{2!}(t\cdot X)^2 + \ldots + \frac{1}{k!}(t\cdot X)^k + \ldots \tag{2.189}$$

Rodovna funkcija momentov dobi obliko:

$$\begin{aligned} M(t) &= E\left(e^{t\cdot X}\right) = E\left(1 + t\cdot X + \frac{1}{2!}(t\cdot X)^2 + \ldots + \frac{1}{k!}(t\cdot X)^k + \ldots\right) = \\ &= 1 + t\cdot E(X) + \frac{t^2}{2!}E\left(X^2\right) + \ldots + \frac{t^k}{k!}E\left(X^k\right) + \frac{t^{k+1}}{(k+1)!}E\left(X^{k+1}\right) + \ldots \end{aligned} \tag{2.190}$$

Tvorimo prvi odvod te funkcije:

$$\begin{aligned} \frac{dM(t)}{dt} &= \frac{d}{dt}\left[1 + t\cdot E(X) + \frac{t^2}{2!}E\left(X^2\right) + \ldots + \frac{t^k}{k!}E\left(X^k\right) + \frac{t^{k+1}}{(k+1)!}E\left(X^{k+1}\right) + \ldots\right] = \\ &= E(X) + \frac{2t^1}{2!}E\left(X^2\right) + \ldots + \frac{k\cdot t^{k-1}}{k!}E\left(X^k\right) + \frac{(k+1)\cdot t^k}{(k+1)!}E\left(X^{k+1}\right) + \ldots \end{aligned} \tag{2.191}$$

Tvorimo drugi odvod te funkcije:

$$\begin{aligned} \frac{d^2M(t)}{dt^2} &= \frac{d}{dt}\left[E(X) + \frac{2t^1}{2!}E\left(X^2\right) + \ldots + \frac{k\cdot t^{k-1}}{k!}E\left(X^k\right) + \frac{(k+1)\cdot t^k}{(k+1)!}E\left(X^{k+1}\right)\ldots\right] = \\ &= \frac{2}{2!}E\left(X^2\right) + \ldots + \frac{k(k-1)\cdot t^{k-2}}{k!}E\left(X^k\right) + \frac{(k+1)\cdot k\cdot t^{k-1}}{(k+1)!}E\left(X^{k+1}\right)\ldots \end{aligned} \tag{2.192}$$

Postopek odvajanja nadaljujemo. Pri *k-tem* odvodu dobimo:

$$\begin{aligned}\frac{d^k M(t)}{dt^k} &= \frac{k(k-1)\cdot(k-2)\cdot\ldots\cdot 1\cdot t^{k-k}}{k!}E\left(X^k\right)+\\ &+\frac{(k+1)(k)\cdot(k-1)\cdot\ldots\cdot 2\cdot t^{k-k+1}}{(k+1)!}E\left(X^{k+1}\right)+\ldots=\\ &=E\left(X^k\right)+t\cdot E\left(X^{k+1}\right)+t^2\cdot E\left(X^{k+2}\right)+\ldots\end{aligned} \tag{2.193}$$

Če v izraz (2.193) vstavimo $t = 0$, dobimo:

$$\frac{d^k M(0)}{dt^k} = E\left(X^k\right) \tag{2.194}$$

Tako smo dokazali, da velja izraz (2.188).

Primer 2.20.:

Poiščite rodovno funkcijo momentov in začetne momente za naključno spremenljivko s porazdelitvijo gostote verjetnosti:

$$f(x)=\begin{cases}e^{-x}, & x\geq 0\\ 0, & x<0\end{cases} \tag{2.195}$$

Rodovna funkcija momentov je:

$$M(t)=E\left(e^{t\cdot X}\right)=\int_0^\infty e^{t\cdot x}\cdot e^{-x}dx=\int_0^\infty e^{-x(1-t)}dx=\left.\frac{e^{-x(1-t)}}{(t-1)}\right|_0^\infty=-\frac{1}{(t-1)}=\frac{1}{1-t} \tag{2.196}$$

Za izračun začetnih momentov je najprej potrebno izračunati zaporedne odvode rodovne funkcije momentov:

$$\begin{aligned}
&\frac{dM(t)}{dt} = \frac{d}{dt}\left(\frac{1}{1-t}\right) = \frac{-1\cdot(-1)}{(1-t)^2} = \frac{1}{(1-t)^2} \\
&\frac{d^2M(t)}{dt^2} = \frac{d}{dt}\left[\frac{1}{(1-t)^2}\right] = \frac{-1\cdot 2\cdot(1-t)(-1)}{(1-t)^4} = \frac{2!}{(1-t)^3} \\
&\frac{d^3M(t)}{dt^3} = \frac{3!}{(1-t)^4} \\
&\ldots \\
&\frac{d^rM(t)}{dt^r} = \frac{r!}{(1-t)^{r+1}} \\
&\ldots
\end{aligned} \tag{2.197}$$

Začetni momenti so:

$$\mu_r' = E\left(X^r\right) = \left.\frac{d^rM(t)}{dt^r}\right|_{t=0} = \left.\frac{r!}{(1-t)^{r+1}}\right|_{t=0} = r! \tag{2.198}$$

Primer 2.21.:

Poiščite rodovno funkcijo momentov, začetne in centralne momente, ter koeficient sploščenosti za naključno spremenljivko s porazdelitvijo gostote verjetnosti:

$$p(x) = \frac{1}{8}\binom{3}{x}, \quad x = 0,1,2,3 \tag{2.199}$$

Rodovna funkcija momentov je:

$$M(t)=E\left(e^{t\cdot X}\right)=\sum_{i=0}^{3}e^{t\cdot x_i}\cdot\frac{1}{8}\binom{3}{x_i}=\frac{1}{8}\left[e^{t\cdot x_0}\cdot\binom{3}{x_0}+e^{t\cdot x_1}\cdot\binom{3}{x_1}+e^{t\cdot x_2}\cdot\binom{3}{x_2}+e^{t\cdot x_3}\cdot\binom{3}{x_3}\right]=$$
$$=\frac{1}{8}\left[e^{t\cdot 0}\cdot\binom{3}{0}+e^{t\cdot 1}\cdot\binom{3}{1}+e^{t\cdot 2}\cdot\binom{3}{2}+e^{t\cdot 3}\cdot\binom{3}{3}\right]=\frac{1}{8}\left[1+e^{t}\cdot 3+e^{t\cdot 2}\cdot 3+e^{t\cdot 3}\right]= \tag{2.200}$$
$$=\frac{1}{8}\left[1+3e^{t}+3e^{2t}+e^{3t}\right]$$

Za izračun začetnih momentov je najprej potrebno izračunati zaporedne odvode rodovne funkcije momentov:

$$\begin{aligned}
\frac{dM(t)}{dt}&=\frac{d}{dt}\left(\frac{1}{8}\left[1+3e^{t}+3e^{2t}+e^{3t}\right]\right)=\frac{1}{8}\left(3e^{t}+6e^{2t}+3e^{3t}\right)\\
\frac{d^2M(t)}{dt^2}&=\frac{d}{dt}\left[\frac{1}{8}\left(3e^{t}+6e^{2t}+3e^{3t}\right)\right]=\frac{1}{8}\left(3e^{t}+12e^{2t}+9e^{3t}\right)\\
\frac{d^3M(t)}{dt^3}&=\frac{d}{dt}\left[\frac{1}{8}\left(3e^{t}+12e^{2t}+9e^{3t}\right)\right]=\frac{1}{8}\left(3e^{t}+24e^{2t}+27e^{3t}\right)\\
\frac{d^4M(t)}{dt^4}&=\frac{d}{dt}\left[\frac{1}{8}\left(3e^{t}+24e^{2t}+27e^{3t}\right)\right]=\frac{1}{8}\left(3e^{t}+48e^{2t}+81e^{3t}\right)
\end{aligned} \tag{2.201}$$

Začetni momenti so:

$$\mu_r'=E\left(X^r\right)=\left.\frac{d^rM(t)}{dt^r}\right|_{t=0},\ r=1,2,3,4$$
$$\begin{aligned}
\mu_1'&=\frac{dM(0)}{dt}=\frac{1}{8}\left(3e^{0}+6e^{0}+3e^{0}\right)=\frac{12}{8}=\frac{3}{2}\\
\mu_2'&=\frac{d^2M(0)}{dt^2}=\frac{1}{8}\left(3e^{0}+12e^{0}+9e^{0}\right)=\frac{24}{8}=3\\
\mu_3'&=\frac{d^3M(0)}{dt^3}=\frac{1}{8}\left(3e^{0}+24e^{0}+27e^{0}\right)=\frac{54}{8}=\frac{27}{4}\\
\mu_4'&=\frac{d^4M(0)}{dt^4}=\frac{1}{8}\left(3e^{0}+48e^{0}+81e^{0}\right)=\frac{132}{8}=\frac{33}{2}
\end{aligned} \tag{2.202}$$

Na osnovi izrazov (2.179), (2.180) in (2.180A) izračunamo centralne momente:

$$\begin{aligned}
&\mu_2 = \mu_2' - \mu_1'^2 = 3 - \left(\frac{3}{2}\right)^2 = 3 - \left(\frac{9}{4}\right) = \frac{3}{4} \\
&\mu_3 = \mu_3' - 3\mu \cdot \mu_2' + 2\mu^3 = \mu_3' - 3\mu_1' \cdot \mu_2' + 2\mu_1'^3 = \frac{27}{4} - 3 \cdot \frac{3}{2} \cdot 3 + 2\left(\frac{3}{2}\right)^3 = \\
&= \frac{27}{4} - \frac{27}{2} + 2\frac{27}{8} = \frac{27}{4} - \frac{54}{4} + \frac{27}{4} = 0 \\
&\mu_4 = \mu_4' - 4\mu_3'\mu + 6\mu^2 \cdot \mu_2' - 3\mu^4 = \frac{33}{2} - 4\frac{27}{4} \cdot \frac{3}{2} + 6\left(\frac{3}{2}\right)^2 \cdot 3 - 3\left(\frac{3}{2}\right)^4 = \\
&= \frac{33}{2} - \frac{81}{2} + 18 \cdot \frac{9}{4} - \frac{243}{16} = -24 + \frac{162}{4} - \frac{243}{16} = -\frac{96}{4} + \frac{162}{4} - \frac{243}{16} = \frac{66}{4} - \frac{243}{16} = \\
&= \frac{264}{16} - \frac{243}{16} = \frac{21}{16}
\end{aligned} \tag{2.203}$$

pri čemer smo upoštevali $\mu = \mu_1'$. Ker je tretji centralni moment enak 0, sklepamo, da je porazdelitev simetrična.

Izračunajmo še koeficient sploščenosti na osnovi izraza (2.181B):

$$\begin{aligned}
&\gamma = g_2 - 3 = \frac{\mu_4}{\left(\sqrt{VAR(X)}\right)^4} - 3 = \frac{\mu_4}{\left(\sqrt{\mu_2}\right)^4} - 3 = \frac{\frac{21}{16}}{\left(\sqrt{\frac{3}{4}}\right)^4} - 3 = \\
&= \frac{\frac{21}{16}}{\left(\frac{3}{4}\right)^2} - 3 = \frac{\frac{21}{16}}{\left(\frac{9}{16}\right)} - 3 = \frac{21}{9} - \frac{27}{9} = -\frac{2}{3}
\end{aligned} \tag{2.204}$$

Ker je koeficient sploščenosti negativen, sklepamo, da je dana porazdelitev v primerjavi z normalno sploščena.

Primer 2.22.:

Za Bernoullijevo porazdelitev poiščite rodovno funkcijo momentov, začetne momente, srednjo vrednost in varianco.

Kot vemo na osnovi izraza (2.17), velja:

$$P(k) = P[X = k] = \begin{cases} 1-p & k = 0 \\ p & k = 1 \end{cases} \tag{2.205}$$

Rodovna funkcija momentov je:

$$M(t) = E\left(e^{t \cdot X}\right) = \sum_{k=0}^{1} e^{t \cdot k} \cdot P(k) = e^{t \cdot 0} \cdot P(0) + e^{t \cdot 1} \cdot P(1) = \\ = 1 - p + e^{t} \cdot p \tag{2.206}$$

Za izračun začetnih momentov je najprej potrebno izračunati zaporedne odvode rodovne funkcije momentov:

$$\frac{dM(t)}{dt} = \frac{d}{dt}\left(1 - p + e^{t} \cdot p\right) = e^{t} \cdot p \\ \frac{d^2 M(t)}{dt^2} = \frac{d}{dt}\left[e^{t} \cdot p\right] = e^{t} \cdot p \tag{2.207}$$

Začetni momenti so:

$$\mu_r' = E\left(X^r\right) = \left.\frac{d^r M(t)}{dt^r}\right|_{t=0}, \; r = 1,2 \\ \mu_1' = \frac{dM(0)}{dt} = \left(e^{0} \cdot p\right) = p \\ \mu_2' = \frac{d^2 M(0)}{dt^2} = \left(e^{0} \cdot p\right) = p \tag{2.208}$$

Srednja vrednost je:

$$\mu = \mu_1' = E(X) = p \tag{2.209}$$

Varianca pa je:

$$VAR(X) = \mu_2 = \mu_2' - \mu_1'^2 = p - p^2 = p(1-p) \tag{2.210}$$

Primer 2.23.:

Za Poissonovo porazdelitev poiščite rodovno funkcijo momentov, prva dva začetna momenta, srednjo vrednost in varianco.

Rodovna funkcija momentov je:

$$M(t) = E\left(e^{t\cdot X}\right) = \sum_{k=0}^{\infty} e^{t\cdot k}\cdot\frac{\lambda^k}{k!}\cdot e^{-\lambda} = e^{-\lambda}\cdot\sum_{k=0}^{\infty}\cdot\frac{\left(\lambda\cdot e^t\right)^k}{k!} = e^{-\lambda}\cdot e^{\lambda\cdot e^t} = e^{\lambda\cdot\left(e^t-1\right)} \tag{2.211}$$

Za izračun prvih dveh začetnih momentov je najprej potrebno izračunati prva odvoda rodovne funkcije momentov:

$$\begin{aligned}
&\frac{dM(t)}{dt} = \frac{d}{dt}\left(e^{\lambda\cdot\left(e^t-1\right)}\right) = e^{\lambda\cdot\left(e^t-1\right)}\cdot\lambda\cdot e^t = e^{\lambda\cdot\left(e^t-1\right)+t}\cdot\lambda \\
&\frac{d^2M(t)}{dt^2} = \frac{d}{dt}\left[e^{\lambda\cdot\left(e^t-1\right)+t}\cdot\lambda\right] = e^{\lambda\cdot\left(e^t-1\right)+t}\cdot\lambda\cdot\left(\lambda\cdot e^t+1\right) = \\
&= \lambda^2 e^{\lambda\cdot\left(e^t-1\right)+2t} + e^{\lambda\cdot\left(e^t-1\right)+t}\cdot\lambda = \\
&= \left(\lambda\cdot e^t\right)^2 e^{\lambda\cdot\left(e^t-1\right)} + \lambda\cdot e^t\cdot e^{\lambda\cdot\left(e^t-1\right)}
\end{aligned} \tag{2.212}$$

Začetna momenta sta:

$$\begin{aligned}
&\mu_r' = E\left(X^r\right) = \left.\frac{d^r M(t)}{dt^r}\right|_{t=0}, \; r = 1,2 \\
&\mu_1' = \frac{dM(0)}{dt} = \left(e^{\lambda\cdot\left(e^0-1\right)+0}\cdot\lambda\right) = \lambda \\
&\mu_2' = \frac{d^2M(0)}{dt^2} = \left(\left(\lambda\cdot e^0\right)^2 e^{\lambda\cdot\left(e^0-1\right)} + \lambda\cdot e^0\cdot e^{\lambda\cdot\left(e^0-1\right)}\right) = \lambda^2+\lambda
\end{aligned} \tag{2.213}$$

Srednja vrednost je:

$$\mu = \mu_1' = E(X) = \lambda \tag{2.214}$$

Varianca pa je:

$$VAR(X) = \mu_2 = \mu_2' - \mu_1'^2 = \lambda^2+\lambda-\lambda^2 = \lambda \tag{2.215}$$

Primer 2.24.:

Za eksponentno porazdelitev poiščite rodovno funkcijo momentov, prva dva začetna momenta, srednjo vrednost in varianco.

Rodovna funkcija momentov je (glej izraz (2.30)):

$$\begin{aligned}
&M(t) = E\left(e^{t\cdot X}\right) = \int_{-\infty}^{\infty} e^{t\cdot x}\cdot f(x)\,dx = \int_0^{\infty} e^{t\cdot x}\cdot\lambda\cdot e^{-\lambda\cdot x}dx = \\
&= \lambda\int_0^{\infty} e^{-x(\lambda-t)}dx = \frac{\lambda}{t-\lambda}e^{-x(\lambda-t)}\Big|_0^{\infty} = \frac{\lambda}{\lambda-t}
\end{aligned} \tag{2.216}$$

Za izračun prvih dveh začetnih momentov je najprej potrebno izračunati prva odvoda rodovne funkcije momentov:

$$\frac{dM(t)}{dt}=\frac{d}{dt}\left(\frac{\lambda}{\lambda-t}\right)=\frac{-\lambda(-1)}{(\lambda-t)^2}=\frac{\lambda}{(\lambda-t)^2}$$
$$\frac{d^2M(t)}{dt^2}=\frac{d}{dt}\left[\frac{\lambda}{(\lambda-t)^2}\right]=\frac{-\lambda\cdot 2\cdot(\lambda-t)(-1)}{(\lambda-t)^4}=\frac{2\lambda}{(\lambda-t)^3} \qquad (2.217)$$

Začetna momenta sta:

$$\mu_r'=E\left(X^r\right)=\left.\frac{d^rM(t)}{dt^r}\right|_{t=0},\ r=1,2$$
$$\mu_1'=\frac{dM(0)}{dt}=\left(\frac{\lambda}{(\lambda-0)^2}\right)=\frac{1}{\lambda} \qquad (2.218)$$
$$\mu_2'=\frac{d^2M(0)}{dt^2}=\left(\frac{2\lambda}{(\lambda-0)^3}\right)=\frac{2}{\lambda^2}$$

Srednja vrednost je:

$$\mu=\mu_1'=E(X)=\frac{1}{\lambda} \qquad (2.219)$$

Varianca pa je:

$$VAR(X)=\mu_2=\mu_2'-\mu_1'^2=\frac{2}{\lambda^2}-\left(\frac{1}{\lambda}\right)^2=\frac{1}{\lambda^2} \qquad (2.220)$$

Primer 2.25.:

Za normalno porazdelitev N(0,1) poiščite rodovno funkcijo momentov, prva dva začetna momenta, srednjo vrednost in varianco.

Rodovna funkcija momentov je (glej izraz (2.36)):

$$M(t)=E\left(e^{t\cdot X}\right)=\int_{-\infty}^{\infty}e^{t\cdot x}\cdot f(x)dx=\int_{-\infty}^{\infty}e^{t\cdot x}\cdot\frac{1}{1\cdot\sqrt{2\pi}}\cdot e^{-\frac{(x-0)^2}{2\cdot 1^2}}dx= \tag{2.221}$$
$$=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}e^{t\cdot x}\cdot e^{-\frac{(x)^2}{2}}dx=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}e^{-\frac{(x)^2}{2}+t\cdot x}dx$$

Eksponent eksponentne funkcije lahko zapišemo kot:

$$-\frac{(x)^2}{2}+t\cdot x=-\frac{(x-t)^2}{2}+\frac{t^2}{2} \tag{2.222}$$

Sledi:

$$M(t)=E\left(e^{t\cdot X}\right)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}e^{-\frac{(x-t)^2}{2}+\frac{t^2}{2}}dx= \tag{2.223}$$
$$=\frac{1}{\sqrt{2\pi}}e^{\frac{t^2}{2}}\int_{-\infty}^{\infty}e^{-\frac{(x-t)^2}{2}}dx$$

Dokazati se da, da velja naslednje:

$$\int_{-\infty}^{\infty}e^{-\frac{(x-t)^2}{2}}dx=\sqrt{2\pi} \tag{2.224}$$

Sledi:

$$M(t)=E\left(e^{t\cdot X}\right)=\frac{1}{\sqrt{2\pi}}e^{\frac{t^2}{2}}\sqrt{2\pi}=e^{\frac{t^2}{2}} \tag{2.225}$$

Za izračun prvih dveh začetnih momentov je najprej potrebno izračunati prva odvoda rodovne funkcije momentov:

$$\frac{dM(t)}{dt} = \frac{d}{dt}\left(e^{\frac{t^2}{2}} \right) = e^{\frac{t^2}{2}} \cdot t$$
$$\frac{d^2M(t)}{dt^2} = \frac{d}{dt}\left[e^{\frac{t^2}{2}} \cdot t \right] = e^{\frac{t^2}{2}} + t \cdot e^{\frac{t^2}{2}} \cdot t = e^{\frac{t^2}{2}} \left(1 + t^2 \right) \tag{2.226}$$

Začetna momenta sta:

$$\mu_r' = E\left(X^r \right) = \left. \frac{d^r M(t)}{dt^r} \right|_{t=0}, r = 1,2$$
$$\mu_1' = \frac{dM(0)}{dt} = \left(e^{\frac{0^2}{2}} \cdot 0 \right) = 0 \tag{2.227}$$
$$\mu_2' = \frac{d^2M(0)}{dt^2} = \left(e^{\frac{0^2}{2}} \left(1 + 0^2 \right) \right) = 1$$

Srednja vrednost je:

$$\mu = \mu_1' = E(X) = 0 \tag{2.228}$$

Varianca pa je:

$$VAR(X) = \mu_2 = \mu_2' - \mu_1'^2 = 1 - (0)^2 = 1 \tag{2.229}$$

2.22 Produktni moment, kovarianca in korelacijski koeficient

Začeni produktni moment redov *r* in *s* dveh naključnih spremenljivk X in Y z dvodimenzionalno porazdelitvijo verjetnosti *p(x,y)* pri diskretnih spremenljivkah oz. dvodimenzionalno porazdelitvijo gostote verjetnosti *f(x,y)* pri zveznih spremenljivkah je matematično upanje produkta naključne spremenljivke $X^r \cdot Y^s$. Tovrsten začetni moment je enak [Hsu, Jesenko]:

$$\mu_{r,s}{}' = E\left(X^r \cdot Y^s\right) = \begin{cases} \sum_{y_j}\sum_{x_i} (x_i)^r \left(y_j\right)^s \cdot p\left(x_i, y_j\right) .....pri\ diskretnih\ spremenljivkah \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x)^r (y)^s \cdot f(x,y)\,dx\,dy .....pri\ zveznih\ spremenljivkah \end{cases} \tag{2.230}$$

<u>Centralni produktni moment redov *r* in *s*</u> dveh naključnih spremenljivk X in Y z dvodimenzionalno porazdelitvijo verjetnosti *p(x,y)* pri diskretnih spremenljivkah oz. dvodimenzionalno porazdelitvijo gostote verjetnosti *f(x,y)* pri zveznih spremenljivkah je matematično upanje produkta naključne spremenljivke $\left(X-\mu_X\right)^r \cdot \left(Y-\mu_Y\right)^s$. Tovrsten centralni moment je enak [Jesenko]:

$$\mu_{r,s} = E\left(\left(X-\mu_X\right)^r \cdot \left(Y-\mu_Y\right)^s\right) =$$
$$= \begin{cases} \sum_{y_j}\sum_{x_i} \left(x_i-\mu_X\right)^r \left(y_j-\mu_Y\right)^s \cdot p\left(x_i, y_j\right) .....pri\ diskretnih\ spremenljivkah \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left(x-\mu_X\right)^r \left(y-\mu_Y\right)^s \cdot f(x,y)\,dx\,dy .....pri\ zveznih\ spremenljivkah \end{cases} \tag{2.231}$$

Začetni produktni moment redov *r=1* in *s=1* je enak:

$$\mu_{1,1}{}' = E\left(X \cdot Y\right) = \begin{cases} \sum_{y_j}\sum_{x_i} (x_i)\left(y_j\right) \cdot p\left(x_i, y_j\right) .....pri\ diskretnih\ spremenljivkah \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x)(y) \cdot f(x,y)\,dx\,dy .....pri\ zveznih\ spremenljivkah \end{cases} \tag{2.232}$$

Centralni produktni moment redov *r=1* in *s=1* imenujemo tudi kovarianca naključnih spremenljivk X in Y in je enak [Hsu, Jesenko]:

$$\mu_{1,1} = \sigma_{XY} = COV(X,Y) = E\left((X-\mu_X)\cdot(Y-\mu_Y)\right) = \\ = \begin{cases} \sum_{y_j}\sum_{x_i}(x_i-\mu_X)(y_j-\mu_Y)\cdot p(x_i,y_j).....pri\ diskretnih\ spremenljivkah \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_X)(y-\mu_Y)\cdot f(x,y)\,dx\,dy.....pri\ zveznih\ spremenljivkah \end{cases} \tag{2.233}$$

Izraz (2.233) lahko zapišemo še nekoliko drugače:

$$\mu_{1,1} = COV(X,Y) = E\left((X-\mu_X)\cdot(Y-\mu_Y)\right) = E\left(XY - X\cdot\mu_Y - \mu_X\cdot Y + \mu_X\mu_Y\right) = \\ = E(X\cdot Y) - \underbrace{E(X)}_{\mu_X}\cdot\mu_Y - \underbrace{E(Y)}_{\mu_Y}\cdot\mu_X + \mu_X\mu_Y = E(X\cdot Y) - \mu_X\mu_Y = \\ = E(X\cdot Y) - E(X)E(Y) = \mu_{1,1}' - \mu_X\mu_Y \tag{2.234}$$

Če sta *X* in *Y* neodvisni (nekorelirani), potem sledi:

$$E(X\cdot Y) = E(X)E(Y) \\ \mu_{1,1} = COV(X,Y) = E(X\cdot Y) - E(X)E(Y) = 0 \tag{2.235}$$

Kadar obstaja velika verjetnost, da pri velikih (majhnih) vrednostih naključne spremenljivke *X* dobimo velike (majhne) vrednosti naključne spremenljivke *Y*, je kovarianca pozitivna, sicer pa je negativna. Očitno torej kovarianca meri **povezanost** med spremenljivkama *X* in *Y* [Jesenko].

Opozorimo še, da v primeru, če sta spremenljivki X in Y neodvisni, se da dokazati, da sta gotovo nekorelirani. Obratno pa ni nujno res, torej, če vemo, da sta spremenljivki nekorelirani, ni nujno, da sta tudi neodvisni med seboj [Hsu].

Poglejmo si še, kaj je to korelacijski koeficient [Hsu]:

$$\rho(X,Y)=\frac{COV(X,Y)}{\sqrt{VAR(X)\cdot VAR(Y)}}=\frac{E(XY)-E(X)E(Y)}{\sqrt{\left\{E\left(X^2\right)-E^2(X)\right\}\left\{E\left(Y^2\right)-E^2(Y)\right\}}} \tag{2.236}$$

oz.

$$\rho_{XY}=\frac{\sigma_{XY}}{\sigma_X\cdot\sigma_Y}$$

za katerega velja:

$$\begin{aligned}&\left|\rho(X,Y)\right|\le 1\\&-1\le\rho(X,Y)\le 1\end{aligned} \tag{2.237}$$

V primeru, da imamo opravka z *n* naključnimi spremenljivkami $X_1, X_2, ..., X_n$, potem smo v izrazu (2.112.) videli, kakšen izgled ima kovariančna matrika. Kovariance parov spremenljivk X_i in X_j imajo obliko:

$$\sigma_{ij}=COV\left(X_i,X_j\right)=E\left(\left(X_i-\mu_i\right)\cdot\left(X_j-\mu_j\right)\right) \tag{2.238}$$

pri čemer je varianca posameznih naključnih spremenljivk X_i enaka:

$$\sigma_i^2=VAR\left(X_i\right)=E\left[\left\{\underbrace{X_i-E\left(X_i\right)}_{\mu_i}\right\}^2\right] \tag{2.239}$$

Korelacijski koeficient ima obliko:

$$\rho_{ij}=\rho\left(X_i,X_j\right)=\frac{COV\left(X_i,X_j\right)}{\sqrt{VAR\left(X_i\right)\cdot VAR\left(X_j\right)}}=\frac{\sigma_{ij}}{\sigma_i\cdot\sigma_j} \tag{2.240}$$

Denimo je naključna spremenljivka Y linearna kombinacija naključnih spremenljivk $X_1, X_2, ..., X_n$ in velja:

$$Y = \sum_{i=1}^{n} a_i \cdot X_i \tag{2.241}$$

Če so spremenljivke X_i neodvisne med seboj, potem velja [Hsu]:

$$\sigma_{ij} = COV\left(X_i, X_j\right) = \begin{cases} \sigma_i^2, & i = j \\ 0, & i \neq j \end{cases} \tag{2.242}$$

in kovariančna matrika zavzame vrednost:

$$\mathbf{K} = \begin{bmatrix} \sigma_1^2 & 0 & \mathrm{L} & 0 \\ 0 & \sigma_2^2 & \mathrm{L} & 0 \\ \mathrm{M} & \mathrm{M} & \mathrm{O} & \mathrm{M} \\ 0 & 0 & \mathrm{L} & \sigma_n^2 \end{bmatrix} \tag{2.243}$$

Dokazati se da, da v primeru neodvisnih naključnih spremenljivk $X_1, X_2, ..., X_n$ varianca spremenljivke Y zavzame vrednost [Hsu]:

$$VAR(Y) = VAR\left(\sum_{i=1}^{n} a_i \cdot X_i\right) = \sum_{i=1}^{n} a_i^2 \cdot VAR\left(X_i\right) \tag{2.244}$$

Primer 2.26.:

Dani imamo naključni spremenljivki X in Y, katerih porazdelitvi gostote verjetnosti sta enaki:

$$\begin{aligned} f_X(x) &= \frac{1}{2} \cdot x, \quad 0 \le x \le 2 \\ f_Y(y) &= \frac{1}{2} \cdot (2 - y), \quad 0 \le y \le 2 \end{aligned} \tag{2.245}$$

in je ugotovljeno, da je $E\left(X \cdot Y\right) = 1$. *Poiščite korelacijski koeficient!*

Najprej bomo poiskali matematični upanji $E(X), E(Y)$. V ta namen tvorimo:

$$E(X)=\int_{-\infty}^{\infty} x\cdot f_X(x)\,dx=\int_0^2 x\cdot\frac{x}{2}\,dx=\left(\frac{x^3}{6}\right)_0^2=\frac{4}{3}$$
$$E(Y)=\int_{-\infty}^{\infty} y\cdot f_Y(y)\,dy=\int_0^2 y\cdot\left(1-\frac{y}{2}\right)dy=\left(\frac{y^2}{2}-\frac{y^3}{6}\right)_0^2=\left(\frac{2^2}{2}-\frac{2^3}{6}\right)= \tag{2.246}$$
$$=\left(2-\frac{4}{3}\right)=\frac{2}{3}$$

Če hočemo izračunati $\rho(X,Y)$, moramo prej izračunati še $E(X^2), E(Y^2)$. Sledi:

$$E(X^2)=\int_{-\infty}^{\infty} x^2\cdot f_X(x)\,dx=\int_0^2 x^2\cdot\frac{x}{2}\,dx=\left(\frac{x^4}{8}\right)_0^2=2$$
$$E(Y^2)=\int_{-\infty}^{\infty} y^2\cdot f_Y(y)\,dy=\int_0^2 y^2\cdot\left(1-\frac{y}{2}\right)dy= \tag{2.247}$$
$$=\left(\frac{y^3}{3}-\frac{y^4}{8}\right)_0^2=\left(\frac{2^3}{3}-\frac{2^4}{8}\right)=\left(\frac{8}{3}-2\right)=\frac{2}{3}$$

Za korelacijski koeficient $\rho(X,Y)$ velja:

$$\rho(X,Y)=\frac{E(XY)-E(X)E(Y)}{\sqrt{\left\{E(X^2)-E^2(X)\right\}\left\{E(Y^2)-E^2(Y)\right\}}}=\frac{1-\frac{4}{3}\cdot\frac{2}{3}}{\sqrt{\left\{2-\left(\frac{4}{3}\right)^2\right\}\left\{\frac{2}{3}-\left(\frac{2}{3}\right)^2\right\}}} \tag{2.248}$$

in dobimo:

$$\rho(X,Y)=\frac{\frac{1}{9}}{\sqrt{\left\{2-\left(\frac{16}{9}\right)\right\}\left\{\frac{2}{3}-\left(\frac{4}{9}\right)\right\}}}=\frac{\frac{1}{9}}{\sqrt{\left\{\frac{2}{9}\right\}\left\{\frac{2}{9}\right\}}}=\frac{\frac{1}{9}}{\frac{2}{9}}=\frac{1}{2} \tag{2.249}$$

2.23 Zakon velikih števil in centralni limitni izrek

Šibek zakon velikih števil

Denimo imamo niz neodvisnih, identično porazdeljenih naključnih spremenljivk $X_1, X_2, ..., X_n$, pri čemer ima vsaka srednjo vrednost $E(X_i) = \mu$. Denimo tudi velja:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{X_1 + ... + X_n}{n} \tag{2.250}$$

Potem za vsak $\varepsilon > 0$ sledi:

$$\lim_{n\to\infty} P\left(\left|\bar{X}_n - \mu\right| > \varepsilon\right) = 0 \tag{2.251}$$

Izraz (2.251) je znan pod imenom **šibek zakon velikih števil** [Hsu], pri čemer je $\bar{X}_n$ vzorčna srednja vrednost.

Močen zakon velikih števil

Denimo imamo niz neodvisnih, identično porazdeljenih naključnih spremenljivk $X_1, X_2, ..., X_n$, pri čemer ima vsaka srednjo vrednost $E(X_i) = \mu$. Potem za vsak $\varepsilon > 0$ sledi:

$$P\left(\lim_{n\to\infty}\left|\bar{X}_n - \mu\right| > \varepsilon\right) = 0 \tag{2.252}$$

Izraz (2.252) je znan pod imenom **močen zakon velikih števil** [Hsu], pri čemer je $\bar{X}_n$ vzorčna srednja vrednost. Ta zakon nam pove, da v limiti $\bar{X}_n$ konvergira k vrednosti μ.

Centralni limitni teorem

Ta teorem je eden najpomembnejših rezultatov v teoriji verjetnosti. Obstaja več verzij tega teorema, pri čemer si bomo ogledali le najpreprostejšo verzijo.

Denimo imamo niz neodvisnih, identično porazdeljenih naključnih spremenljivk $X_1, X_2, ..., X_n$, pri čemer ima vsaka srednjo vrednost μ in varianco σ^2. Vpeljimo novo spremenljivko:

$$Z_n = \frac{X_1 + ... + X_n - n \cdot \mu}{\sigma\sqrt{n}} = \frac{\sum_{i=1}^{n} X_i - n \cdot \mu}{\sigma\sqrt{n}} = \\ = \frac{n\left(\frac{\sum_{i=1}^{n} X_i}{n} - \mu\right)}{\sigma\sqrt{n}} = \frac{n\left(\bar{X}_n - \mu\right)}{\sigma\sqrt{n}} = \frac{\left(\bar{X}_n - \mu\right)}{\frac{\sigma}{\sqrt{n}}} \tag{2.253}$$

Kot se izkaže, velja naslednje [Hsu]:

$$\lim_{n\to\infty} Z_n = \lim_{n\to\infty} \frac{\left(\bar{X}_n - \mu\right)}{\frac{\sigma}{\sqrt{n}}} = \lim_{n\to\infty} \frac{\left(\frac{\sum_{i=1}^{n} X_i}{n} - \mu\right)}{\frac{\sigma}{\sqrt{n}}} = N(0,1) \tag{2.254}$$

ali:

$$\lim_{n\to\infty} F_{Z_n}(z) = \lim_{n\to\infty} P(Z_n \le z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{t^2}{2}} dt = \Phi(z) \tag{2.255}$$

kjer je *N(0,1)* **standardna normalna porazdelitev,** $\Phi(z)$ pa je njena kumulativna verjetnostna porazdelitev, katere vrednosti se da prebrati iz ustrezne tabele.

Torej **centralni limitni teorem** govori o tem, da je pri dovolj velikem *n* porazdelitev vsote $X_1 + ... + X_n$ približno enaka normalni porazdelitvi, ne glede na to, kakšno porazdelitev imajo posamezne spremenljivke X_i [Hsu].

Primer 2.27.:

Dokažite šibek zakon velikih števil!

Pri šibkem zakonu velikih števil imamo:

$$\lim_{n\to\infty} P\left(\left|\bar{X}_n - \mu\right| > \varepsilon\right) = 0 \tag{2.256}$$

pri čemer je $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{X_1 + \ldots + X_n}{n}$, $E\left(X_i\right) = \mu$ in $VAR\left(X_i\right) = \sigma^2$.

Na osnovi izraza (2.146) velja tudi:

$$E\left(\bar{X}_n\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = E\left(\frac{X_1 + \ldots + X_n}{n}\right) = \frac{E\left(X_1\right) + \ldots + E\left(X_n\right)}{n} = \tag{2.257}$$
$$= \frac{\mu + \mu + \ldots + \mu}{n} = \frac{n\mu}{n} = \mu$$

na osnovi izraza (2.244) pa velja:

$$VAR\left(\bar{X}_n\right) = VAR\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = VAR\left(\sum_{i=1}^{n} \frac{X_i}{n}\right) = VAR\left(\frac{X_1}{n} + \ldots + \frac{X_n}{n}\right) = \tag{2.258}$$
$$= \frac{1}{n^2} VAR\left(X_1\right) + \ldots + \frac{1}{n^2} VAR\left(X_n\right) = \frac{1}{n^2}\left(\sigma^2 + \ldots + \sigma^2\right) = \frac{n \cdot \sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Pri dokazu si bomo pomagali s takoimenovanim **Chebyshevim zakonom**, ki se v splošni obliki glasi [Hsu]:

$$P\left(\left|X - \mu_X\right| \geq a\right) \leq \frac{VAR\left(X\right)}{a^2} = \frac{\sigma_X^2}{a^2} \tag{2.259}$$

oz. v našem primeru:

$$P\left(\left|\bar{X}_n - \mu\right| \geq \varepsilon\right) \leq \frac{VAR\left(\bar{X}_n\right)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \tag{2.260}$$

Ker v limiti velja:

$$\lim_{n\to\infty} \frac{\sigma^2}{n\varepsilon^2} = 0 \tag{2.261}$$

sledi:

$$\lim_{n\to\infty} P\left(\left|\bar{X}_n - \mu\right| > \varepsilon\right) = 0 \tag{2.262}$$

in tako je zakon dokazan.

Primer 2.28.:

Denimo imamo naključno spremenljivko X s porazdelitvijo gostote verjetnosti $f(x)$*, prav tako imamo set neodvisnih naključnih spremenljivk* $X_1, X_2, ..., X_n$*, pri čemer ima vsaka spremenljivka gostoto verjetnosti* $f(x)$*. Potem se set* $X_1, X_2, ..., X_n$ *imenuje naključni vzorec velikosti n spremenljivke X, ki ima srednjo vrednost* μ *in varianco* σ^2*. Koliko vzorcev naključne spremenljivke X je potrebno vzeti, da bo verjetnost, da vzorčna srednja vrednost* $\bar{X}_n$ *ne bo odstopala od dejanske srednje vrednosti* μ *za več kot* $\varepsilon = \frac{\sigma}{10}$*, enaka vsaj 95%?*

Izhajamo iz izraza (2.260):

$$P\left(\left|\bar{X}_n - \mu\right| \geq \varepsilon\right) \leq \frac{VAR\left(\bar{X}_n\right)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \tag{2.263}$$

oziroma:

$$P\left(\left|\bar{X}_n-\mu\right|\geq\frac{\sigma}{10}\right)\leq\frac{\sigma^2}{n\left(\frac{\sigma}{10}\right)^2}$$
$$1-P\left(\left|\bar{X}_n-\mu\right|<\frac{\sigma}{10}\right)\leq\frac{\sigma^2}{n\left(\frac{\sigma}{10}\right)^2}=\frac{100}{n} \tag{2.264}$$

Tako dobimo:

$$1-\frac{100}{n}\leq P\left(\left|\bar{X}_n-\mu\right|<\frac{\sigma}{10}\right) \tag{2.265}$$

Hkrati glede na tekst naloge velja tudi:

$$P\left(\left|\bar{X}_n-\mu\right|<\frac{\sigma}{10}\right)\geq 0.95 \tag{2.266}$$

Izenačimo oba pogoja in dobimo:

$$\begin{aligned}&1-\frac{100}{n}\geq\frac{95}{100}\\&100-\frac{10000}{n}\geq 95\\&100\cdot n-10000\geq 95\cdot n\\&5\cdot n\geq 10000\\&n\geq 2000\end{aligned} \tag{2.267}$$

Torej moramo vzeti vsaj 2000 vzorcev, da bo verjetnost, da vzorčna srednja vrednost $\bar{X}_n$ ne bo odstopala od dejanske srednje vrednosti μ za več kot $\varepsilon=\frac{\sigma}{10}$, enaka vsaj 95%.

3 UVOD V STATISTIKO

Statistika se je kot znanost, ki razvija metode o zbiranju podatkov, njihovi analizi in predstavitvi, začela razvijati šele proti koncu 19. stoletja, z deli K. Pearsona, R. Fisherja, J. Neymana in E. Pearsona. Dela teh statistikov so bila namenjena obravnavi zelo različne problematike in so vodila tudi do oblikovanja nekaterih posebnih statističnih disciplin. Ti statistiki so pokazali, da je mogoče, če so izpolnjene določene predpostavke, z opazovanjem dela enot pojava ocenjevati lastnosti celotnega pojava Pri tem je mogoče tudi ugotoviti kakovost ocenjevanja in verjetnost nepravilnih sklepov [Košmelj B.].

V delih K. Pearsona so bile začrtane osnove za ocenjevanje lastnosti populacije, če so vzorci dovolj veliki. R. Fisher je preučeval predvsem problematiko vzorcev z majhnim številom enot in prvi uvedel **statistično preizkušanje domnev**. Pri tem se je omejil na upoštevanje ene domneve, ki jo je imenoval **ničelna domneva**. V tridesetih letih 20. stoletja sta J. Neyman in E. Pearson ugotovila, da je smiselno nasproti ničelni domnevi postaviti še alternativno domnevo. Pri tem sta upoštevala napaki 1. in 2. vrste ter opredelila postopek, kakršen je znan kot statistično preizkušanje domnev še danes [Košmelj B.].

Statistično sklepanje je statistični pristop, ki temelji na opazovanju dela enot, izvedenem ob določenih predpostavkah, ter vključuje sklepanje o lastnostih pojava z navajanjem verjetnosti za pravilnost sklepov [Košmelj B.].

Statistično sklepanje se je razvilo v dveh smereh [Košmelj B.]:

- v ocenjevanje parametrov, ter
- v preizkušanje domnev.

Statistično sklepanje je bistveni sestavni del mnogih statističnih disciplin, kot so npr. [Košmelj B.]:

- načrtovanje poskusov,
- kontrola kakovosti,
- analiza časovnih vrst,
- multivariantna analiza,

- neparametrične metode,
- teorija odločitev, itn.

Bistveno za kakovost statističnih podatkov, dobljenih z vzorčenjem, je navajanje **standardnih napak** in **stopenj tveganja** (verjetnosti za nepravilne sklepe). Trditve na osnovi vzorca namreč niso postavljene s popolno gotovostjo, pač pa le z neko verjetnostjo. Postopki statističnega sklepanja so sicer naravnani tako, da je velika verjetnost, da je sklep, dobljen z vzorčnimi podatki, pravilen, vendar je tveganje, da je napačen, vseeno prisotno [Košmelj B.].

Poudariti še velja, da pri opazovanju pojavov ni mogoče vedno izpolniti predpostavke, na katerih temelji statistično sklepanje. Zato so bili pri opazovanju in analizi pojavov razviti tudi drugi pristopi, kot npr. **analiza podatkov**, ki se prav tako pogosto uporablja v statistični analizi. Pomemben del analize podatkov se imenuje **odkrivalna analiza podatkov** (Exploratory Data Analysis), ki jo je uvedel Tukey po letu 1970, ki se nanaša na metode proučevanja, prikazovanja in povzemanja bistva statističnih podatkov, brez statističnega sklepanja ali modeliranja. Takšne metode npr. so [Košmelj B.]:

- opisna statistika,
- grafični prikaz kvantilov,
- grafična analiza osamelcev,
- multivariantna metoda razvrščanja v skupine,
- korenspodenčna analiza, itn.

3.1 Prvine statistike

Skupnost enot, ki sestavljajo pojav, ki ga preučujemo s statističnimi opazovanji, imenujemo statistična populacija (populacija). Če označimo posamezne enote z $e_1, e_2, ..., e_N$, potem zapišemo populacijo z N enotami na naslednji način:

$$P(e_1, e_2, ..., e_N) \tag{3.1}$$

Denimo imamo opravka z neko fakulteto, kjer želijo preučiti povezanost med socialno-ekonomskimi dejavniki in uspehom študentov. Zato so za študente zbrali podatke o kraju bivanja, prejemkih in uspehu v 1. letniku. Vsekakor je v tem primeru proučevana enota študent, populacija pa vsi študentje fakultete.

S statističnimi opazovanji želimo spoznati lastnosti populacije, ki jih opisujemo s statističnimi parametri, kot npr. z aritmetično sredino, standardno deviacijo, regresijskim koeficientom, itn. Parametri so npr. delež študentov iz določenega kraja bivanja, povprečni prejemki študenta, ali delež študentov, ki so v roku opravili vse obveznosti 1. letnika.

Da bi lahko v skladu s cilji statističnega raziskovanja izračunali vrednosti ustreznih parametrov, moramo seveda opazovati posamezne enote. Če opazujemo vse enote populacije (*N* študentov) in zanje izračunamo vrednost ustreznega parametra, govorimo o **pravi vrednosti parametra**. Če pa se pri opazovanju omejimo le na nekatere enote (vzorec z *n* enotami), dobimo le **(vzorčno) oceno parametra** [Košmelj B.].

Enote e_i se po vrednostih opazovane lastnosti ponavadi razlikujejo. Če označimo lastnosti enot s spremenljivkami $Y_1, Y_2, ..., Y_k$ (v našem primeru npr. kraj bivanja, prejemki in uspeh v 1. letniku) in vrednosti slednjih z $y_1, y_2, ..., y_k$, potem lahko **pravo vrednost** parametra za spremenljivko Y zapišemo kot:

$$\theta_Y = f\left(y_1, y_2, ..., y_N\right) \tag{3.2}$$

njegovo **vzorčno oceno** pa kot:

$$\hat{\theta}_Y = f\left(y_1, y_2, ..., y_n\right) \tag{3.3}$$

Če nas npr. zanima parameter aritmetična sredina, potem lahko za pravo vrednost zapišemo:

$$\theta_Y = \mu_Y = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{3.4}$$

za vzorčno oceno pa:

$$\hat{\theta}_Y = \hat{\mu}_Y = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{3.5}$$

Spremenljivke opisujejo različne lastnosti enot. Njihove vrednosti so lahko izražene številsko (numerično) ali opisno (atributivno). Kot se izkaže, z vidika izračunavanja parametrov ta delitev pogosto ne daje zadostne osnove za izbiro ustreznega parametra oz. ne omogoča ustrezne razlage dobljenih rezultatov. Zato je z vidika izračunavanja parametrov pomembnejša **razvrstitev spremenljivk glede na mersko lestvico**.

Slika 34 prikazuje možne merske lestvice opredelitev različnosti spremenljivk, nekaj primerov spremenljivk in možne parametre glede na mersko lestvico [Košmelj B.].

Merska lestvica	Opredelitev različnosti med dvema vrednostma spremenljivke	Nekaj primerov spremenljivk	Možni parametri
Imenska	z neenakostjo ali enakostjo $y_A \neq y_B$ ali $y_A = y_B$	spol, rojstni kraj, hišna številka, serija proizvoda	število enot, modus, delež enot, koeficient kontingence
Urejenostna	z urejanjem po velikosti $y_A \geq y_B$ ali $y_A \leq y_B$	kvalifikacija, ocena izgleda proizvoda	kvantili, koeficient korelacije ranga
Razmična	z razliko $y_A - y_B = d$	temperatura na Celzijevi skali, koledarski čas	aritmetična sredina, standardni odklon, koeficient korelacije
Razmernostna	z razmerjem $y_A / y_B = k$	dohodek, starost, višina računa	geometrijska sredina, harmonična sredina

Slika 34: Možne merske lestvice opredelitev različnosti spremenljivk, nekaj primerov spremenljivk in možni parametri glede na mersko lestvico [Košmelj B.].

Pri statističnem opazovanju predpostavimo, da zajamemo v opazovanje vse enote populacije in želimo dobiti o njej kar najboljšo sliko. To nam včasih uspe v večji, včasih pa v manjši meri, pri čemer pride tudi do določenih napak.

Denimo ugotavljamo vrednost kakšne spremenljivke, npr. Y pri enoti e_i. Potem je $\hat{y}_j$ opazovana vrednost te spremenljivke, y_j pa dejanska vrednost te spremenljivke. Napako definiramo na naslednji način [Košmelj B.]:

$$d_j = \hat{y}_j - y_j, \quad j = 1, 2, ..., N \tag{3.6}$$

Če je $d_j = 0$, potem je opazovana vrednost točna.

Če nato iz opazovanih vrednosti izračunavamo vrednosti parametrov, se v njih odraža kakovost opazovanih vrednosti. Razlika med pravo in ocenjeno vrednostjo parametra je **merilo pristranskosti**, definirano na naslednji način [Košmelj B.]:

$$\Delta\theta_Y = \hat{\theta}_Y - \theta_Y \tag{3.7}$$

Če je $\Delta\theta_Y = 0$, potem je vrednost parametra točna (nepristranska).

Pristranskost parametra je možno izraziti tudi relativno [Košmelj B.]:

$$\Delta\theta_Y(\%) = \frac{\hat{\theta}_Y - \theta_Y}{\theta_Y} \cdot 100 = \frac{\Delta\theta_Y}{\theta_Y} \cdot 100 \tag{3.8}$$

Napake, pri katerih se njihov skupen učinek kompenzira, so naključne napake, ki ne povzročijo pristranskosti parametrov. Takšen primer je npr. parameter aritmetična sredina. Pri nekaterih drugih parametrih, npr. pri varianci, pa se skupen učinek naključnih napak ne izravna, zato pride do pristranskosti ocene. Ker je zaradi tega netočnost ocenjenega parametra lahko relativno velika, so tedaj tovrstne napake gotovo silno nezaželene [Košmelj B.].

3.2 Vzorčna opazovanja

Vzorčna opazovanja so v nekaterih primerih smiselna, v drugih pa celo edina možna. Pri tem je izbira enot eno najpomembnejših vprašanj, ki se tukaj porajajo. Odpirata se dve možnosti: bodisi upoštevamo naključno izbiro enot (naključni vzorec), ali pa izberemo vzorce na kakšen drug, nenaključen način (nenaključni vzorec) [Košmelj B.].

Pri naključnih vzorcih je enotam vzorčenja (posamezne enote, ali skupine enot) zagotovljena enaka možnost, da so izbrane v vzorec, pri čemer je znana tudi verjetnost za izbiro enote v vzorec. Pri nenaključnih vzorcih pa slednja ni znana, zato tudi ni mogoče izračunati ustreznih kazalcev za ocenjevanje kakovosti vzorčnih ocen [Košmelj B.].

Statistična teorija daje prednost vzorčenju, pri katerem je vzorec naključen, saj je s tem zagotovljena večja objektivnost pri izbiti enot v opazovanje in možnost ugotavljanja kazalcev za ocenjevanje kakovosti vzorčnih ocen [Košmelj B.]

Postopek, v katerem se zberejo enote vzorčenja in ocenjujejo parametri pri naključnih vzorcih, imenujemo verjetnostno vzorčenje ali na kratko vzorčenje, statistična disciplina, ki se s tem ukvarja, pa teorija vzorčenja [Košmelj B.].

3.3 Vzorčenje in kakovost vzorčnih ocen

Glede na to, da je ob enakih pogojih mogoče iz dane populacije z *N* enotami izbrati veliko število vzorcev brez ponavljanja s po *n* enot in da ob danem opazovanju izberemo le en vzorec izmed vseh možnih, se odpira vprašanje, kakšne so vzorčne ocene parametra pri posameznih vzorcih. ***Kot se izkaže, se vzorčne ocene parametra praviloma razlikujejo med seboj, ko izberemo različne vzorce*** [Košmelj B.].

Primer 3.1.:

Denimo imamo štiri enote e_1, e_2, e_3, e_4 v populaciji (*N=4*). Pare dveh enot v vzorcu lahko izberemo na $\binom{N}{n} = \binom{4}{2} = 6$ načinov, torej lahko dobimo šest možnih vzorčnih ocen parametra (glej sliko 35).

Zaporedna številka vzorca	Enote v vzorcu	Vzorčne ocene parametra θ
1	e_1, e_2	$\hat{\theta}_1$
2	e_1, e_3	$\hat{\theta}_2$
3	e_1, e_4	$\hat{\theta}_3$
4	e_2, e_3	$\hat{\theta}_4$
5	e_2, e_4	$\hat{\theta}_5$
6	e_3, e_4	$\hat{\theta}_6$

Slika 35: Šest možnih vzorčnih ocen parametra

Torej s posameznim vzorcem dobimo vzorčno oceno parametra, ki je odvisna od vzorčnih vrednosti opazovane spremenljivke, torej vrednosti, ki jih dobimo v vzorec pri izbranih enotah. Seveda so vzorčne ocene parametra odvisne od vzorca do vzorca. ***Spremeljivko, katere vrednosti so vzorčne ocene*** $\hat{\theta}_i, \quad i = 1, 2, ..., \binom{N}{n}$ ***in so funkcija v vzorec zajetih vrednosti, imenujemo cenilka, pri kateri imamo opravka s porazdelitvijo vseh vzorčnih ocen parametra [Košmelj B.]. Ta porazdelitev ima osrednje mesto pri vzorčenju, saj so iz nje izpeljane mere za presojo kvalitete ocenjevanja parametrov na osnovi vzorcev*** [Košmelj B.].

Porazdelitev vzorčnih ocen bi lahko poimenovali tudi porazdelitev cenilke. Tipične mere za opis slednje so npr. aritmetična sredina, varianca in standardna deviacija.

Pričakovana vrednost vzorčnih ocen

Iz vzorčnih ocen $\hat{\theta}_i, \quad i = 1, 2, ..., \binom{N}{n}$ lahko izračunamo aritmetično sredino na naslednji način [Košmelj B.]:

$$\mu_{\hat{\theta}} = E\left(\hat{\theta}\right) = \frac{1}{\binom{N}{n}} \sum_{i=1}^{\binom{N}{n}} \hat{\theta}_i \tag{3.9}$$

Aritmetično sredino vzorčnih ocen parametra je mogoče izračunati tudi kot tehtano sredino, kjer so ponderji verjetnosti. Ker ima lahko več vzorcev enako oceno parametra, predpostavimo, da je npr. *s* različnih vzorčnih ocen. Potem lahko zapišemo [Košmelj B.]:

$$E\left(\hat{\theta}\right)=\sum_{i=1}^{s}\hat{\theta}_i\cdot P\left(\hat{\theta}_i\right) \tag{3.10}$$

kjer je $P\left(\hat{\theta}_i\right)$ verjetnost za posamezno oceno, ter je $s\le\binom{N}{n}$. Seveda velja tudi: $\sum_{i=1}^{s}P\left(\hat{\theta}_i\right)=1.$

Primer 3.2.:

Dano imamo populacijo, ki šteje *N=5* enot u_1,u_2,u_3,u_4,u_5 z vrednostmi spremenljivke *Y*:

$$\begin{aligned} y_1&=4\\ y_2&=6\\ y_3&=9\\ y_4&=8\\ y_5&=3 \end{aligned} \tag{3.11}$$

Ocenjujemo vrednost parametra μ_Y, ki predstavlja srednjo vrednost spremenljivke *Y*. Prava vrednost tega parametra je:

$$\mu_Y=\frac{1}{N}\sum_{i=1}^{N}y_i=\frac{1}{5}\left(y_1+y_2+y_3+y_4+y_5\right)=\frac{1}{5}\left(4+6+9+8+3\right)=6 \tag{3.12}$$

Denimo jemljemo vzorce po tri enote (brez ponavljanja). Vseh možnih načinov izbire treh vzorcev iz populacije petih vzorcev je $\binom{N}{n}=\binom{5}{3}=10$. Denimo je $s=\binom{N}{n}=10$ in smo uspeli izbrati vse možne vzorce (glej sliko 36 [Košmelj B.]).

Zaporedna številka vzorca v	Enote v vzorcu	Vrednosti y_{vi} spremenljivke Y	$\hat{\mu}_Y$
1	u_1, u_2, u_3	4, 6, 9	$6\frac{1}{3}$
2	u_1, u_2, u_4	4, 6, 8	6
3	u_1, u_2, u_5	4, 6, 3	$4\frac{1}{3}$
4	u_1, u_3, u_4	4, 9, 8	7
5	u_1, u_3, u_5	4, 9, 3	$5\frac{1}{3}$
6	u_1, u_4, u_5	4, 8, 3	5
7	u_2, u_3, u_4	6, 9, 8	$7\frac{2}{3}$
8	u_2, u_3, u_5	6, 9, 3	6
9	u_2, u_4, u_5	6, 8, 3	$5\frac{2}{3}$
10	u_3, u_4, u_5	9, 8, 3	$6\frac{2}{3}$

Slika 36: Izbrani vzorci, vrednosti spremenljivke Y pri njih, ter vzorčna ocena [Košmelj B.]

Vzorčne ocene tvorimo na naslednji način:

$$\begin{aligned}
\hat{\mu}_{Y1} &= \frac{1}{3}(4+6+9) = \frac{19}{3} \\
\hat{\mu}_{Y2} &= \frac{1}{3}(4+6+8) = \frac{18}{3} = 6 \\
\hat{\mu}_{Y3} &= \frac{1}{3}(4+6+3) = \frac{13}{3} \\
\hat{\mu}_{Y4} &= \frac{1}{3}(4+9+8) = \frac{21}{3} = 7 \\
\hat{\mu}_{Y5} &= \frac{1}{3}(4+9+3) = \frac{16}{3} \\
\hat{\mu}_{Y6} &= \frac{1}{3}(4+8+3) = \frac{15}{3} = 5 \\
\hat{\mu}_{Y7} &= \frac{1}{3}(6+9+8) = \frac{23}{3} \\
\hat{\mu}_{Y8} &= \frac{1}{3}(6+9+3) = \frac{18}{3} = 6 \\
\hat{\mu}_{Y9} &= \frac{1}{3}(6+8+3) = \frac{17}{3} \\
\hat{\mu}_{Y10} &= \frac{1}{3}(9+8+3) = \frac{20}{3}
\end{aligned} \tag{3.13}$$

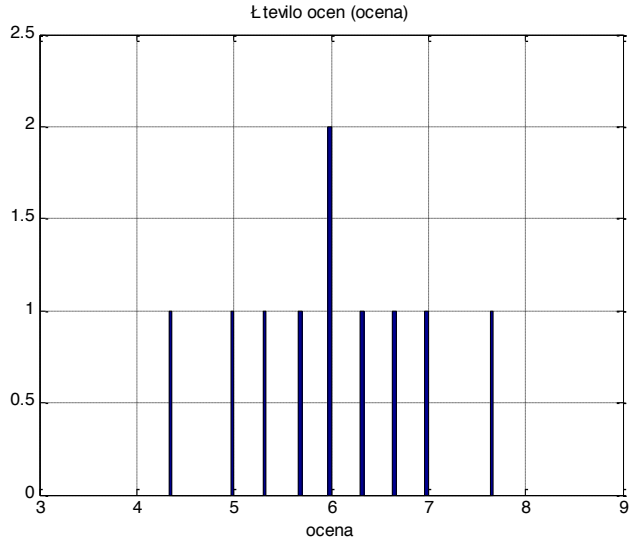
Na osnovi izraza (3.9) lahko izračunamo aritmetično sredino (pričakovano vrednost) vzorčnih ocen na prvi način:

$$\mu_{\hat{\theta}} = E\left(\hat{\theta}\right) = \frac{1}{\binom{N}{n}} \sum_{i=1}^{\binom{N}{n}} \hat{\theta}_i$$
$$\mu_{\hat{\mu}_Y} = E\left(\hat{\mu}_Y\right) = \frac{1}{10} \sum_{i=1}^{10} \hat{\mu}_{Yi} = \frac{1}{10}\left(\frac{19+18+13+21+16+15+23+18+17+20}{3}\right) = 6 \tag{3.14}$$

na osnovi izraza (3.10) pa še na drugi način:

$$E\left(\hat{\theta}\right) = \sum_{i=1}^{s} \hat{\theta}_i \cdot P\left(\hat{\theta}_i\right)$$
$$E\left(\hat{\mu}_Y\right) = \sum_{i=1}^{10} \hat{\mu}_{Yi} \cdot P\left(\hat{\mu}_{Yi}\right) = \hat{\mu}_{Y1} \cdot P\left(\hat{\mu}_{Y1}\right) + \hat{\mu}_{Y2} \cdot P\left(\hat{\mu}_{Y2}\right) + \ldots + \hat{\mu}_{Y10} \cdot P\left(\hat{\mu}_{Y10}\right) =$$
$$= \frac{19}{3} \cdot \frac{1}{10} + \frac{18}{3} \cdot \frac{1}{10} + \ldots + \frac{20}{3} \cdot \frac{1}{10} = \tag{3.15}$$
$$= \frac{1}{10}\left(\frac{19+18+13+21+16+15+23+18+17+20}{3}\right) = 6$$

Narišimo še porazdelitev vzorčnih ocen parametra aritmetična sredina, ki jo prikazuje slika 37.



Slika 37: Porazdelitev vzorčnih ocen parametra aritmetična sredina

Programski ukazi v Matlabu za izris slike 37 so bili naslednji:

```
% slika1.m

clc
clear
close all

y = [19/3 6 13/3 7 16/3 5 23/3 6 17/3 20/3];

hist(y,100)

axis([3 9 0 2.5])
grid

title('Število ocen (ocena)')
xlabel('ocena')
```

Kot lahko vidimo, so štiri vzorčne ocene $\hat{\mu}_{Yi}$ večje od prave vrednosti parametra $\mu_Y = 6$, štiri pa manjše. Dve vzorčni oceni sta enaki pravi vrednosti parametra, vse druge ocene pa se pojavljajo le enkrat.

Varianca in standardna deviacija vzorčnih ocen

Razlike med vzorčnimi ocenami parametra $\hat{\theta}_i$ merimo z varianco vzorčnih ocen (varianca cenilke). Izračunamo jo na naslednji način [Košmelj B.]:

$$VAR\left(\hat{\theta}\right) = \frac{1}{\binom{N}{n}} \sum_{i=1}^{\binom{N}{n}} \left(\hat{\theta}_i - \mu_{\hat{\theta}_i}\right)^2 \tag{3.16}$$

Varianca vzorčnih ocen je kazalec za kvaliteto vzorčnih ocen in eden temeljnih kazalcev v teoriji vzorčenja. Tudi varianco vzorčnih ocen lahko izrazimo kot pričakovano vrednost vsote kvadratov odklonov posameznih ocen od pričakovane vrednosti teh ocen, to je [Košmelj B.]:

$$VAR\left(\hat{\theta}\right) = E\left[\left\{\hat{\theta} - E\left(\hat{\theta}\right)\right\}^2\right] = \sum_{i=1}^{s} \left\{\hat{\theta}_i - E\left(\hat{\theta}\right)\right\}^2 \cdot P\left(\hat{\theta}_i\right) \tag{3.17}$$

Če se povrnemo k prejšnjemu primeru, dobimo varianco vzorčnih ocen na prvi način na osnovi izraza (3.16):

$$
\begin{aligned}
VAR\left(\hat{\theta}\right) &= \frac{1}{\binom{N}{n}} \sum_{i=1}^{\binom{N}{n}} \left(\hat{\theta}_i - \mu_{\hat{\theta}_i}\right)^2 \\
VAR\left(\hat{\mu}_Y\right) &= \frac{1}{10} \sum_{i=1}^{10} \left(\hat{\mu}_{Yi} - \mu_{\hat{\mu}_Y}\right)^2 = \frac{1}{10} \sum_{i=1}^{10} \left(\hat{\mu}_{Yi} - 6\right)^2 = \\
&= \frac{1}{10}\left[\left(\frac{19}{3} - 6\right)^2 + \left(\frac{18}{3} - 6\right)^2 + \ldots + \left(\frac{20}{3} - 6\right)^2\right] = 0.87
\end{aligned}
\tag{3.18}
$$

Programski ukazi v Matlabu za ta izračun so bili naslednji:

```
% primer1.m

clc
clear
close all

y = [19/3 6 13/3 7 16/3 5 23/3 6 17/3 20/3]

ysr = mean(y)
N = length(y)

var = 0;

for i = 1:10
    var = var + (y(i)-ysr)^2;
end

var = var/N
```

Če se povrnemo k prejšnjemu primeru, dobimo varianco vzorčnih ocen na drugi način na osnovi izraza (3.17):

$$
\begin{aligned}
VAR\left(\hat{\theta}\right) &= E\left[\left\{\hat{\theta} - E\left(\hat{\theta}\right)\right\}^2\right] = \sum_{i=1}^{s} \left\{\hat{\theta}_i - E\left(\hat{\theta}\right)\right\}^2 \cdot P\left(\hat{\theta}_i\right) \\
VAR\left(\hat{\mu}_Y\right) &= E\left[\left\{\hat{\mu}_Y - E\left(\hat{\mu}_Y\right)\right\}^2\right] = \sum_{i=1}^{10} \left\{\hat{\mu}_{Yi} - E\left(\hat{\mu}_Y\right)\right\}^2 \cdot P\left(\hat{\mu}_{Yi}\right) = \\
&= \sum_{i=1}^{10} \left\{\hat{\mu}_{Yi} - 6\right\}^2 \cdot \frac{1}{10} = \frac{1}{10}\left[\left(\frac{19}{3} - 6\right)^2 + \left(\frac{18}{3} - 6\right)^2 + \ldots + \left(\frac{20}{3} - 6\right)^2\right] = 0.87
\end{aligned}
\tag{3.19}
$$

1 2 3

Standardno deviacijo vzorčnih ocen dobimo na naslednji način:

$$\sigma_{\hat{\theta}} = STD\left(\hat{\theta}\right) = \sqrt{VAR\left(\hat{\theta}\right)} \tag{3.20}$$

oz. v našem primeru je enaka:

$$\sigma_{\hat{\mu}_Y} = STD\left(\hat{\mu}_Y\right) = \sqrt{VAR\left(\hat{\mu}_Y\right)} = \sqrt{0.87} = 0.93 \tag{3.21}$$

Natančnost vzorčne ocene

Natančnost vzorčne ocene je tem večja, čim manjša je varianca vzorčnih ocen, kar pomeni, da se vzorčne ocene parametra $\hat{\theta}_i$ čim manj razlikujejo med seboj in so posledično čim bliže njihovi pričakovani vrednosti $E\left(\hat{\theta}_Y\right)$. Pri tem pa se ne oziramo na dejstvo, a se vzorčne ocene porazdeljujejo okoli prave vrednosti parametra θ ali ne [Košmelj B.].

Pristranskost vzorčne ocene

Pristranskost lahko nastane zaradi več razlogov: zaradi pristranskosti cenilke, zaradi pristranskosti v vzorčnem postopku (neustrezna izbira enot v vzorec), zaradi pristranskosti v opazovanju, itn. Pristranskost se nanaša na razliko $E\left(\hat{\theta}_Y\right) - \theta_Y$, pri čemer za nepristransko cenilko velja: $E\left(\hat{\theta}_Y\right) = \theta_Y$. Ker prave vrednosti parametra θ_Y običajno ne poznamo, tudi ne moremo izračunati učinka pristranskosti. Kljub temu pa lahko na pristranskost vplivamo z ustreznim načrtovanjem vzorca, pri čemer poskušamo odpraviti vse dejavnike, ki povzročajo pristranskost [Košmelj B.].

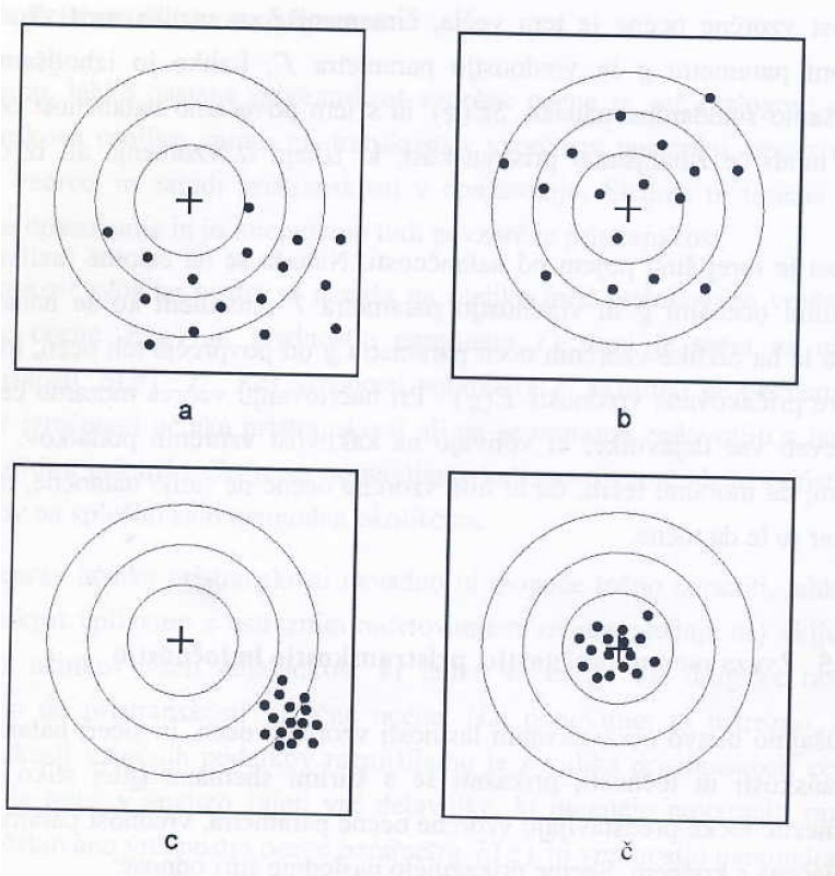
Točnost vzorčne ocene

Točnost vzorčne ocene je tem večja, čim manjše so razlike med vzorčnimi ocenami $\hat{\theta}_{Yi}$ in pravo vrednostjo parametra θ_Y. Točnost lahko izboljšamo, če povečamo natančnost

opazovanja in/ali zmanjšamo pristranskost, ki izhaja iz vzorčenja ali drugih virov [Košmelj B.].

Slika 38 prikazuje odnose med natančnostjo, pristranskostjo in točnostjo [Košmelj B.]. Pri tem velja naslednje:

a) majhna natančnost, velika pristranskost, majhna točnost,
b) majhna natančnost, majhna pristranskost, majhna točnost,
c) velika natančnost, velika pristranskost, majhna točnost,
č) velika natančnost, majhna pristranskost, velika točnost.

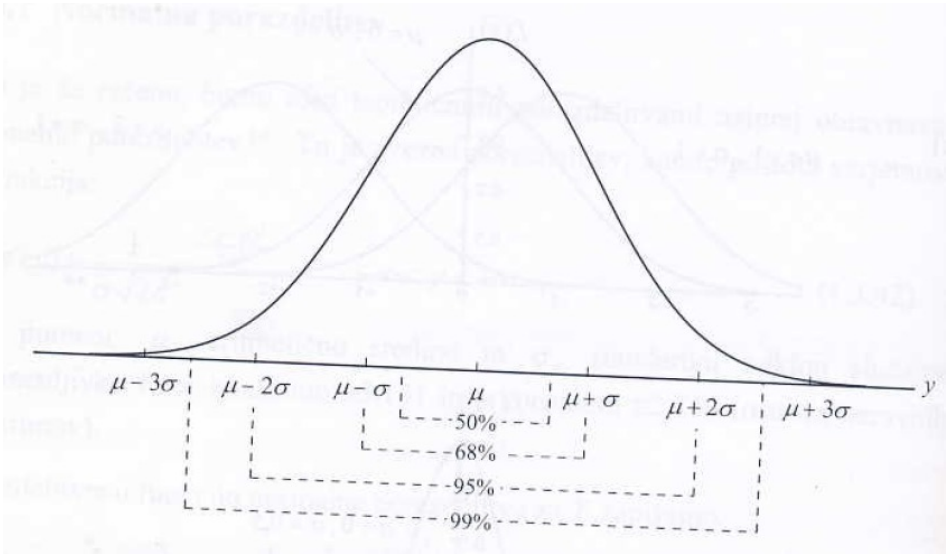


Slika 38: Odnosi med natančnostjo, pristranskostjo in točnostjo [Košmelj B.]

3.4 Pomen normalne porazdelitve pri vzorčenju

Normalno porazdelitev smo nekoliko spoznali že v poglavju 2.8. Ima naslednje lastnosti [Košmelj B.]:

- Je unimodalna in simetrična,
- Aritmetična sredina je enaka mediani in modusu,
- Asimptotično se približuje abscisni osi, ploščina pod krivuljo je enaka 1, pri čemer je 50% ploščine levo in 50% ploščine desno od aritmetične sredine.
- Na intervalu $[\mu_Y - 0.67\sigma_Y, \mu_Y + 0.67\sigma_Y]$ je približno 50% ploščine,
- Na intervalu $[\mu_Y - \sigma_Y, \mu_Y + \sigma_Y]$ je približno 68% ploščine,
- Na intervalu $[\mu_Y - 2\sigma_Y, \mu_Y + 2\sigma_Y]$ je približno 95% ploščine,
- Na intervalu $[\mu_Y - 2.58\sigma_Y, \mu_Y + 2.58\sigma_Y]$ je približno 99% ploščine (glej sliko 39 [Košmelj B.]).



Slika 39: Ilustracija različnih vrednosti ploščine pod krivuljo pri različnih intervalih za normalno porazdelitev [Košmelj B.]

Standardizirana normalna porazdelitev

To je normalna porazdelitev za takoimenovano standardizirano spremenljivko *Z*. Njeno vrednost za posamezno enoto e_i izračunamo na naslednji način:

$$z_i = \frac{y_i - \mu_Y}{\sigma_Y} \tag{3.22}$$

pri čemer ima spremenljivka *Y* normalno porazdelitev:

$$f\left(y_i\right)=\frac{1}{\sigma_Y\cdot\sqrt{2\pi}}\cdot e^{-\frac{(y_i-\mu_Y)^2}{2\sigma_Y{}^2}} \tag{3.23}$$

Porazdelitev gostote verjetnosti za spremenljivko *Z* dobimo na naslednji način:

$$\begin{aligned}
&F_Z(z_i)=P(Z\le z_i)=P\left(\frac{Y-\mu_Y}{\sigma_Y}\le z_i\right)=P\left(Y\le z_i\cdot\sigma_Y+\mu_Y\right)=\\
&=\int_{-\infty}^{z_i\cdot\sigma_Y+\mu_Y}\frac{1}{\sigma_Y\cdot\sqrt{2\pi}}\cdot e^{-\frac{(y_i-\mu_Y)^2}{2\sigma_Y{}^2}}dy_i\\
&t_i=\frac{y_i-\mu_Y}{\sigma_Y},\quad dy_i=\sigma_Y\cdot dt_i\\
&Sledi:\\
&F_Z(z_i)=\int_{-\infty}^{\frac{z_i\cdot\sigma_Y+\mu_Y-\mu_Y}{\sigma_Y}}\frac{1}{\sigma_Y\cdot\sqrt{2\pi}}\cdot e^{-\frac{(t_i)^2}{2}}\sigma_Y\cdot dt_i\\
&F_Z(z_i)=\int_{-\infty}^{z_i}\frac{1}{\sqrt{2\pi}}\cdot e^{-\frac{(t_i)^2}{2}}dt_i\\
&f_Z\left(z_i\right)=\frac{dF_Z(z_i)}{dz_i}=\frac{1}{\sqrt{2\pi}}\cdot e^{-\frac{(z_i)^2}{2}}
\end{aligned} \tag{3.24}$$

Aritmetična sredina in varianca za spremenljivko *Z* sta enaki:

$$\begin{aligned}
&\mu_Z=E\left(Z\right)=E\left(\frac{Y-\mu_Y}{\sigma_Y}\right)=\frac{1}{\sigma_Y}E\left(Y-\mu_Y\right)=\frac{1}{\sigma_Y}\left(E\left(Y\right)-\mu_Y\right)=0\\
&\sigma_Z^2=E\left[\left(Z-\mu_Z\right)^2\right]=E\left[\left(Z-0\right)^2\right]=E\left(\left\{\frac{Y-\mu_Y}{\sigma_Y}\right\}^2\right)=\frac{1}{\sigma_Y{}^2}\underbrace{E\left[\left(Y-\mu_Y\right)^2\right]}_{\sigma_Y{}^2}=1
\end{aligned} \tag{3.25}$$

Vrednost z_i pove, za koliko standardnih odklonov se vrednost y_i razlikuje od svojega povprečja (aritmetične sredine) in tudi, v kateri smeri. Tako vrednost $z_i=+2$ pove, da je vrednost y_i večja od svojega povprečja za dva standardna odklona. Vrednost $z_i=-1.5$ pa npr. pove, da je vrednost y_i manjša od svojega povprečja za -1.5 standardnega odklona.

Tako je mogoče po vrednosti z_i sklepati na mesto enote e_i v populaciji ali vzorcu, pa tudi primerjati posamezne vrednosti med različnimi populacijami [Košmelj B.].

Za standardizirano normalno porazdelitev so vrednosti tabelirane, kar prikazujeta sliki 40 in 41. Na sliki 40 je podana tabela za $-3.49 \le z \le -0.09$, na sliki 41 pa tabela za $0 \le z \le 3.49$ [Košmelj B.]. V obeh tabelah so pri različnih *z-jih* podane vrednosti kumulativne funkcije $F(z)$, ki izražajo delež ploščine za standardizirano normalno porazdelitev na intervalu $-\infty < t \le z$. Zaradi simetričnosti standardizirane normalne porazdelitve tudi velja relacija:

$$F(-Z) = 1 - F(Z) \tag{3.26}$$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
-1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
-1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
-1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
-1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
-0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
-0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
-0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
-0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
-0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
-0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
-0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
-0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
-0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
-0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

Slika 40: Tabelirane vrednosti kumulativne funkcije standardizirane normalne porazdelitve pri $-3.49 \le z \le -0.09$ *[Košmelj B.]*

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Slika 41: Tabelirane vrednosti kumulativne funkcije standardizirane normalne porazdelitve pri $0 \le z \le 3.49$ *[Košmelj B.]*

V tabeli na sliki 42 je prikazanih nekaj najpomembnejših vrednosti kumulativne funkcije za standardizirano normalno porazdelitev (glej tudi sliki 40 in 41) [Košmelj B.].

z_i	$F(z_i)$	Odstotek ploščine pod standardizirano normalno porazdelitvijo med -∞ in z_i
-3,29	0,0005	0,05
-3,09	0,0010	0,1
-2,58	0,0049	0,5
-2,33	0,0102	1,0
-1,96	0,0250	2,5
-1,64	0,0505	5,0
0,00	0,5000	50,0
1,64	0,9495	95,0
1,96	0,9750	97,5
2,33	0,9898	99,0
2,58	0,9951	99,5
3,09	0,9990	99,9
3,29	0,9995	99,95

Slika 42: Nekaj najpomembnejših vrednosti kumulativne funkcije za standardizirano normalno porazdelitev [Košmelj B.]

Primer 3.3.:

Utemeljite, da je na intervalu $[y_1, y_2] = [\mu_Y - 2\sigma_Y, \mu_Y + 2\sigma_Y]$ *približno 95% ploščine pri normalni porazdelitvi!*

Postavimo:

$$\begin{aligned} y_1 &= \mu_Y + z_1\sigma_Y = \mu_Y - 2\sigma_Y \quad \Rightarrow \quad z_1 = -2 \\ y_2 &= \mu_Y + z_2\sigma_Y = \mu_Y + 2\sigma_Y \quad \Rightarrow \quad z_2 = 2 \end{aligned} \tag{3.27}$$

Nato gremo gledati tabeli na slikah 40 in 41 in odčitamo:

$$\begin{aligned} z_1 = -2 \quad &\Rightarrow \quad F(z_1) = 0.0228 \\ z_2 = 2 \quad &\Rightarrow \quad F(z_2) = 0.9772 \end{aligned} \tag{3.28}$$

Sledi:

$$F(z_2) - F(z_1) = 0.9772 - 0.0228 = 0.9544 \tag{3.29}$$

Torej je na intervalu $[y_1, y_2] = [\mu_Y - 2\sigma_Y, \mu_Y + 2\sigma_Y]$ dejansko približno 95% ploščine pri normalni porazdelitvi!

Normalna porazdelitev je pomembna tako s stališča porazdelitve opazovanih vrednosti, verjetnostnih porazdelitev, kot tudi porazdelitev vzorčnih ocen. Pri proučevanju pojavov, ko opazujemo lastnosti enot, lahko dobimo pri razvrščanju slednjih po vrednosti opazovane spremenljivke porazdelitve, ki so normalne ali vsaj približno normalne. To še posebej velja na področju naravoslovja in tehnologije, manj pa pri proučevanju socialno-ekonomskih pojavov [Košmelj B.]. Tudi napake pri ponavljajočih se meritvah dane lastnosti, npr. dolžine, višine, mase, itn, so velikokrat normalno porazdeljene [Košmelj B.].

V nadaljevanju si poglejmo še nekaj primerov.

Primer 3.4.:

Iz izkušenj vemo, da imajo proizvodi, proizvedeni na stroju A, povprečno dolžino 30 cm in standardni odklon 0.2 cm, pri čemer je dolžina proizvodov normalno porazdeljena. Odgovorite:

a) Kolikšen je delež proizvodov, ki so krajši od 30.3 cm?

b) Kolikšen je delež proizvodov, ki so dolgi med 30 in 30.5 cm?

c) Kolikšen je odstotek proizvodov, ki so krajši od 29.8 cm?

d) Kolikšen je odstotek proizvodov, ki so dolgi med 29.7 in 30.2 cm?

e) Kolikšno je število proizvodov z dolžino med 30.2 cm in 30.4 cm, če je proizvedenih 5000 proizvodov?

V splošnem za standardizirano spremenljivko velja:

$$z_i = \frac{y_i - \mu_Y}{\sigma_Y} = \frac{y_i - 30}{0.2} \tag{3.30}$$

a) Najprej izračunamo z_i:

$$z_i = \frac{30.3 - 30}{0.2} = \frac{3}{2} = 1.5 \tag{3.31}$$

Nato gremo gledati tabelo na sliki 40 in odčitamo:

$$z_i = 1.5 \;\Rightarrow\; F(z_i) = 0.9332 \;\Rightarrow\; Delež(Y < 30.3) = 0.9332 \tag{3.32}$$

Torej je delež proizvodov, ki so krajši od 30.3 cm, enak 0.9332. Do približno enakega rezultata bi prišli tudi z numerično integracijo za integral $F_Z(z_i) = \int_{-\infty}^{z_i} \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{(t_i)^2}{2}} dt_i$.

Ustrezen program v Matlabu bi bil:

```
% numericna integracija (num_int.m):

clear,
clc

ch = input('Zelis numer. integ. stand. normal.
porazd.?');
if ch == 1
   f = 'exp(-x^2/2)/sqrt(2*pi)'
else
   f = input('Vnesi funkcijo, npr. x/2','s')
end

a = input('Vnesi spodnjo mejo')
b = input('Vnesi zgornjo mejo')
dx = input('Vnesi korak')

if length(dx) == 0
   dx = 0.05
end

x1 = [a:dx:b];

I = 0;

N = round((b-a)/dx)

for i = 1:N
    x = x1(i);
    I = I + dx*eval(f);
end

disp('Rezultat numericne integracije je:')
I
```

Izpis komandnega okna bi bil:

```
Zelis numer. integ. stand. normal. porazd.?1

f =
exp(-x^2/2)/sqrt(2*pi)

Vnesi spodnjo mejo-100
a =
 -100

Vnesi zgornjo mejo1.5
b =
   1.5000

Vnesi korak
dx =
   0.0500

N =
     2030

Rezultat numericne integracije je:
I =
   0.9299
```

b) Najprej izračunamo z_i:

$$z_i = \frac{30.5 - 30}{0.2} = \frac{5}{2} = 2.5 \tag{3.33}$$

Nato gremo gledati tabelo na sliki 40 in odčitamo:

$$z_i = 2.5 \;\Rightarrow\; F(z_i) = 0.9938 \;\Rightarrow\; Delež(30 < Y < 30.5) = 0.9938 - 0.5 = 0.4938 \tag{3.34}$$

Torej je delež proizvodov, ki so dolgi med 30 in 30.5 cm, enak 0.4938. Do približno enakega rezultata za $F(z_i)$ bi prišli tudi z numerično integracijo z Matlabom, pri čemer bi dobili rezultat 0.9933 pri koraku 0.05.

c) Najprej izračunamo z_i:

$$z_i = \frac{29.8 - 30}{0.2} = \frac{-2}{2} = -1 \tag{3.35}$$

Nato gremo gledati tabelo na sliki 39 in odčitamo:

$$z_i = -1 \;\Rightarrow\; F(z_i) = 0.1587 \;\Rightarrow\; Odstotek(Y < 29.8) = 15.87\% \tag{3.36}$$

Torej je odstotek proizvodov, ki so krajši od 29.8 cm, enak 15.87%. Do približno enakega rezultata za $F(z_i)$ bi prišli tudi z numerično integracijo z Matlabom, pri čemer bi dobili rezultat 0.1574 pri koraku 0.01.

d) Najprej izračunamo z_{i1}:

$$z_{i1} = \frac{29.7 - 30}{0.2} = -\frac{3}{2} = -1.5 \tag{3.37}$$

Nato gremo gledati tabelo na sliki 39 in odčitamo:

$$z_{i1} = -1.5 \;\Rightarrow\; F(z_{i1}) = 0.0668 \;\Rightarrow\; Delež(Y < 29.7) = 0.0668 \tag{3.38}$$

Nato izračunamo z_{i2}:

$$z_{i2} = \frac{30.2 - 30}{0.2} = \frac{2}{2} = 1 \tag{3.39}$$

Nato gremo gledati tabelo na sliki 40 in odčitamo:

$$z_{i2} = 1 \;\Rightarrow\; F(z_{i2}) = 0.8413 \;\Rightarrow\; Delež(Y < 30.2) = 0.8413 \tag{3.40}$$

Sledi:

$$\begin{aligned} &Delež(29.7 < Y < 30.2) = Delež(Y < 30.2) - Delež(Y < 29.7) = \\ &= 0.8413 - 0.0668 = 0.7745 \end{aligned} \tag{3.41}$$

Torej je proizvodov z dolžino med 29.7 cm in 30.2 cm približno 77.45%. Do približno enakega rezultata za $F(z_{i1})$ bi prišli tudi z numerično integracijo z Matlabom, pri čemer bi dobili rezultat 0.0662 pri koraku 0.01. Za $F(z_{i2})$ pa bi dobili rezultat 0.8401 pri koraku 0.01.

e) Najprej izračunamo z_{i1} in odčitamo $F(z_{i1})$:

$$z_{i1} = \frac{30.2 - 30}{0.2} = 1 \;\Rightarrow\; F(z_{i1}) = 0.8413 \;\Rightarrow\; Delež(Y < 30.2) = 0.8413 \tag{3.42}$$

Nato izračunamo z_{i2} in odčitamo $F(z_{i2})$:

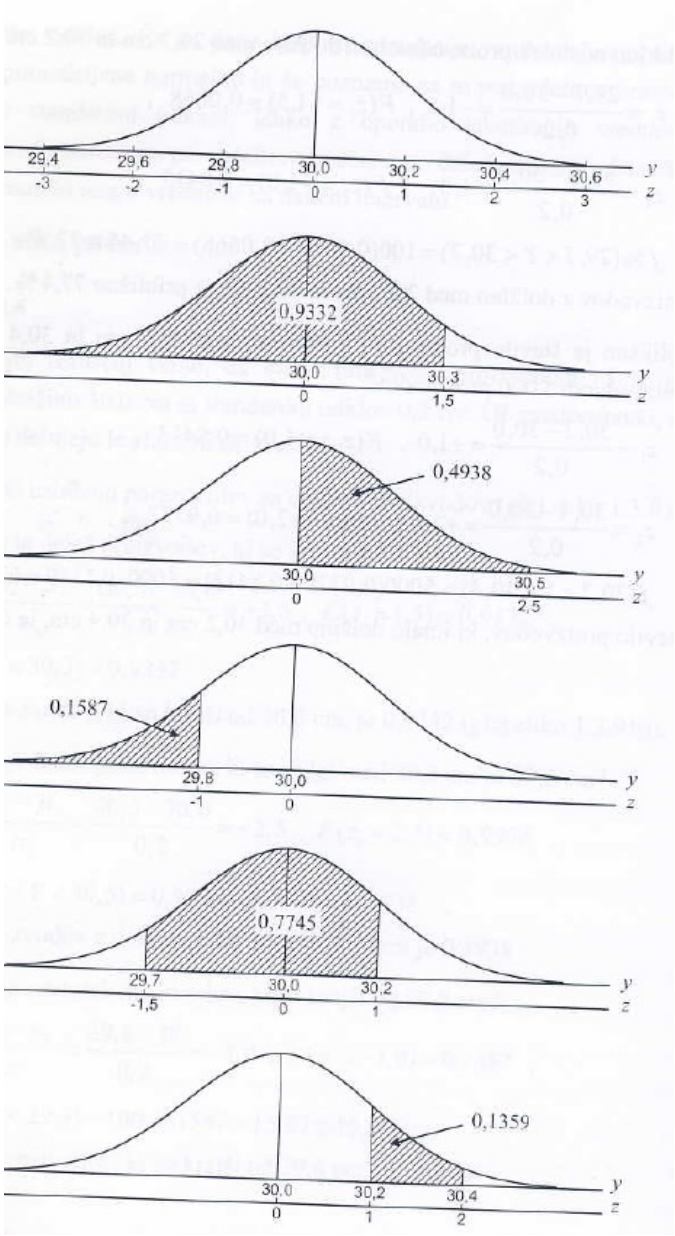
$$z_{i2} = \frac{30.4 - 30}{0.2} = 2 \;\Rightarrow\; F(z_{i2}) = 0.9772 \;\Rightarrow\; Delež(Y < 30.4) = 0.9772 \tag{3.43}$$

Sledi:

$$\begin{aligned} &Delež\left(30.2 < Y < 30.4\right) = Delež\left(Y < 30.4\right) - Delež\left(Y < 30.2\right) = \\ &= 0.9772 - 0.8413 = 0.1359 \end{aligned} \tag{3.44}$$

$$Delež\left(30.2 < Y < 30.4\right)_{5000} = 5000 \cdot 0.1359 = 679.5$$

Torej je izmed 5000 proizvodov delež proizvodov, ki imajo dolžino med 30.2 in 30.4 cm, enak cca 680 proizvodov. Vsi rezultati tega primera so ustrezno ilustrirani na sliki 43 [Košmelj B.].



Slika 43: Ilustracija rezultatov za primer normalne porazdelitve dolžin proizvodov [Košmelj B.]

Stopnja tveganja

Denimo imamo nek dogodek A. Verjetnost, da dogodek nastopi, je enaka $P(A)$, da ne nastopi, pa:

$$\alpha = P(\bar{A}) = 1 - P(A) \tag{3.45}$$

pri čemer α imenujemo **tveganje za nenastop dogodka** [Košmelj B.]. V statistični analizi, ki jo izvajamo na vzorčnih podatkih, postavljamo vse trditve ob neki stopnji tveganja, ki pa ne sme biti visoka. Običajno mora veljati:

$$\alpha \leq 0.05 \tag{3.46}$$

Če določamo stopnjo tveganja vnaprej, potem upoštevamo poleg že omenjene največje stopnje tveganja v višini 0.05 najpogosteje še stopnjo tveganja v višini 0.01 in 0.001 [Košmelj B.].

V splošnem za stopnjo tveganja pri izračunani spodnji meji opazovane spremenljivke velja [Košmelj B.]:

$$\alpha = P(Y \leq y_S) \tag{3.47}$$

Pri normalni porazdelitvi bi imeli:

$$\alpha = P(Y \leq y_S) = P(Z\sigma_Y + \mu_Y \leq y_S) = P\left(Z \leq \frac{y_S - \mu_Y}{\sigma_Y}\right) = P(Z \leq z_s) = F(z_s) \tag{3.48}$$

V splošnem za stopnjo tveganja pri izračunani zgornji meji opazovane spremenljivke velja [Košmelj B.]:

$$\alpha = P(Y \geq y_Z) \tag{3.49}$$

Pri normalni porazdelitvi bi imeli:

$$\alpha = P\left(Y \geq y_Z\right) = P\left(Z\sigma_Y + \mu_Y \geq y_Z\right) = P\left(Z \geq \frac{y_Z - \mu_Y}{\sigma_Y}\right) = \\ = P\left(Z \geq z_Z\right) = 1 - F\left(z_Z\right) \tag{3.50}$$

Vrnimo se k prejšnjemu primeru, kjer imajo proizvodi, proizvedeni na stroju A, povprečno dolžino 30 cm in standardni odklon 0.2 cm, pri čemer je dolžina proizvodov normalno porazdeljena.

Izračunajte tisto dolžino proizvoda, pri kateri je verjetnost, da ima naključno izbrani proizvod večjo dolžino od izračunane spodnje meje, enaka 0.95.

Imamo:

$$P\left(Y \geq y_S\right) = P\left(Z\sigma_Y + \mu_Y \geq y_S\right) = P\left(Z \geq \frac{y_S - \mu_Y}{\sigma_Y}\right) = P\left(Z \geq z_s\right) = 1 - F\left(z_s\right) = 0.95 \tag{3.51}$$
$$F\left(z_s\right) = \alpha = 1 - 0.95 = 0.05$$

Odčitamo iz tabele na sliki 39 tisti z_S, pri katerem je $F\left(z_S\right) = 0.05$:

$$z_s = -1.64 \tag{3.52}$$

Spodnjo mejo pri stopnji tveganja $\alpha = 0.05$ izračunamo takole:

$$y_S = z_s\sigma_Y + \mu_Y = -1.64 \cdot 0.2 + 30 = 29.67 \tag{3.53}$$

Izračunajte tisto dolžino proizvoda, pri kateri je verjetnost, da ima naključno izbrani proizvod krajšo dolžino od izračunane zgornje meje, enaka 0.95.

Imamo:

$$P(Y \le y_Z) = P(Z\sigma_Y + \mu_Y \le y_Z) = P\left(Z \le \frac{y_Z - \mu_Y}{\sigma_Y}\right) = P(Z \le z_Z) = F(z_Z) = 0.95 \quad (3.54)$$
$$\alpha = 1 - F(z_Z) = 1 - 0.95 = 0.05$$

Odčitamo iz tabele na sliki 40 tisti z_Z, pri katerem je $F(z_Z) = 0.95$:

$$z_Z = 1.64 \quad (3.55)$$

Zgornjo mejo pri stopnji tveganja $\alpha = 0.05$ izračunamo takole:

$$y_Z = z_Z\sigma_Y + \mu_Y = 1.64 \cdot 0.2 + 30 = 30.33 \quad (3.56)$$

Do sedaj smo se seznanili s stopnjo tveganja pri dveh **enostranskih trditvah**, ko smo opazovali interval glede na spodnjo mejo oz. interval glede na zgornjo mejo. Lahko pa obravnavamo tudi stopnjo tveganja pri dvostranski trditvi, pri čemer velja [Košmelj B.]:

$$\alpha_{skupen} = 2\alpha = P(Y \le y_S) + P(Y \ge y_Z) \quad (3.57)$$

V primeru normalne porazdelitve bi ob upoštevanju izrazov (3.48) in (3.50) imeli:

$$\alpha + \alpha = F(z_s) + 1 - F(z_Z) = F(z_s) + 1 - \left(F(z_s) + P(z_s \le Z \le z_Z)\right) = 1 - P(z_s \le Z \le z_Z) \quad (3.58)$$
$$\alpha = \frac{1 - P(z_s \le Z \le z_Z)}{2}$$

Izračunajte tisto dolžino proizvoda, pri kateri je verjetnost, da ima naključno izbrani proizvod dolžino med izračunanima spodnjo in zgornjo mejo, enaka 0.95.

Imamo:

$$\begin{aligned} &P(z_s \le Z \le z_Z) = 0.95 \\ &\alpha = \frac{1 - P(z_s \le Z \le z_Z)}{2} = \frac{1 - 0.95}{2} = 0.025 \end{aligned} \tag{3.59}$$

Na osnovi izraza (3.48) dobimo:

$$F(z_s) = \alpha = 0.025 \tag{3.60}$$

Odčitamo iz tabele na sliki 39 tisti z_S, pri katerem je $F(z_S) = 0.025$:

$$z_s = -1.96 \tag{3.61}$$

Spodnjo mejo pri stopnji tveganja $\alpha = 0.025$ izračunamo takole:

$$y_S = z_s \sigma_Y + \mu_Y = -1.96 \cdot 0.2 + 30 = 29.608 \approx 29.61 \tag{3.62}$$

Na osnovi izraza (3.50) dobimo:

$$F(z_Z) = 1 - \alpha = 1 - 0.025 = 0.975 \tag{3.63}$$

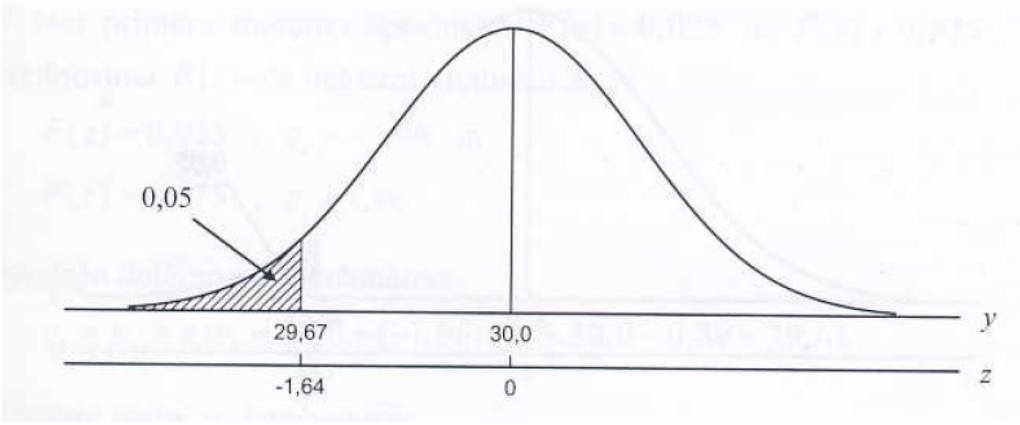
Odčitamo iz tabele na sliki 40 tisti z_Z, pri katerem je $F(z_Z) = 0.975$:

$$z_Z = 1.96 \tag{3.64}$$

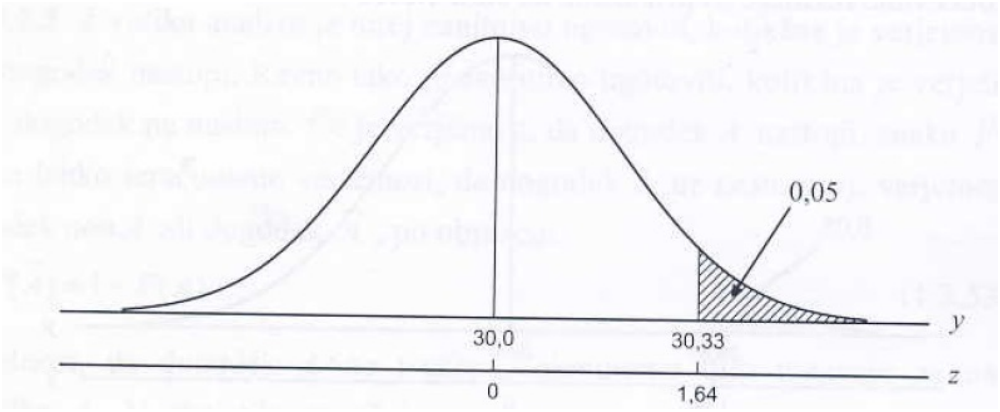
Zgornjo mejo pri stopnji tveganja $\alpha = 0.025$ izračunamo takole:

$$y_Z = z_Z\sigma_Y + \mu_Y = 1.96 \cdot 0.2 + 30 = 30.39 \tag{3.65}$$

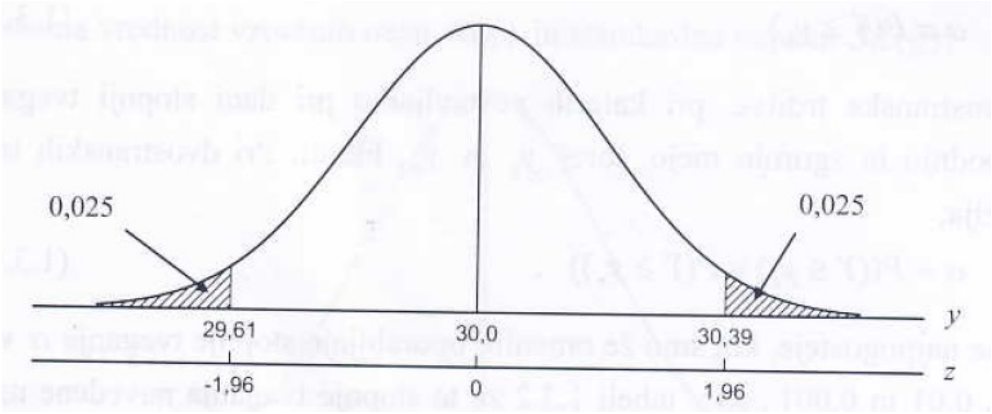
Slike 44, 45 in 46 ilustrirajo določanje spodnje meje in zgornje meje pri enostranski trditvi, ter obeh meja pri dvostranski trditvi [Košmelj B.].



Slika 44: Določanje spodnje meje pri $\alpha = 0.05$ [Košmelj B.]



Slika 45: Določanje zgornje meje pri $\alpha = 0.05$ [Košmelj B.]



Slika 46: Določanje spodnje in zgornje meje pri $\alpha_{skupen} = 0.05$ [Košmelj B.]

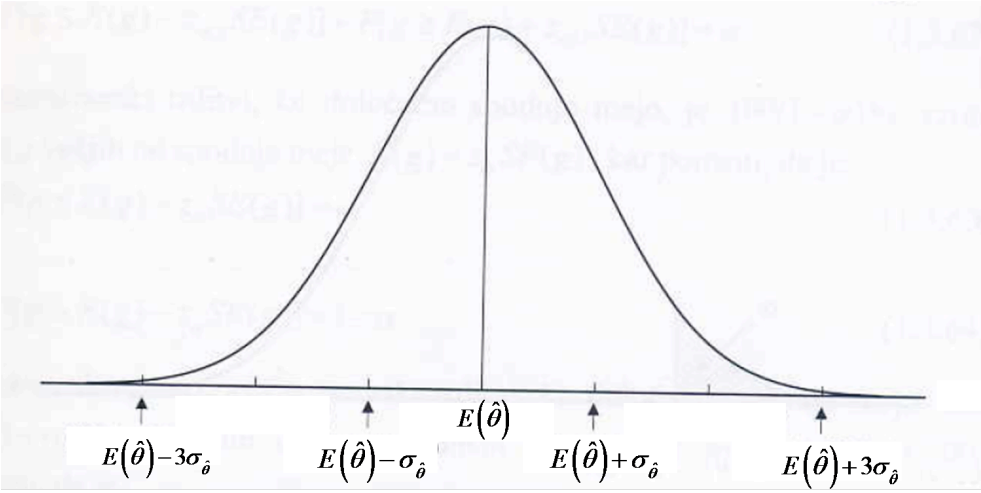
Tabela na sliki 47 prikazuje vrednosti kumulativne funkcije in standardizirane spremenljivke pri najbolj tipičnih stopnjah tveganja za normalno porazdelitev [Košmelj B.].

Trditev	α	$F(z)$	z
Enostranska	0,05	0,05 ali 0,95	-1,64 ali 1,64
	0,01	0,01 ali 0,99	-2,33 ali 2,33
	0,001	0,001 ali 0,999	-3,09 ali 3,09
Dvostranska	0,05	0,025 in 0,975	-1,96 in 1,96
	0,01	0,005 in 0,995	-2,58 in 2,58
	0,001	0,0005 in 0,9995	-3,29 in 3,29

Slika 47: Vrednosti kumulativne funkcije in standardizirane spremenljivke pri najbolj tipičnih stopnjah tveganja za normalno porazdelitev [Košmelj B.]

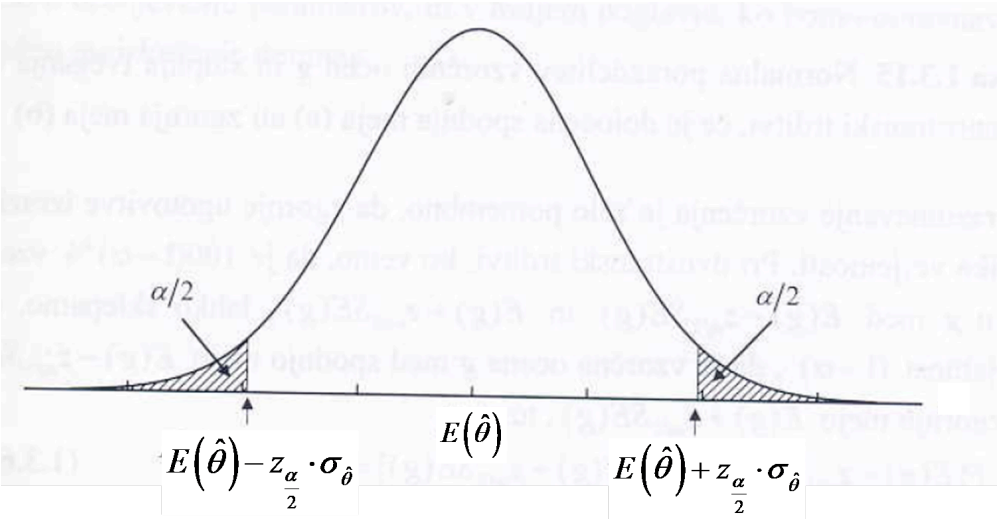
Normalna porazdelitev kot porazdelitev vzorčnih ocen

Kot se izkaže, se vzorčne ocene dokaj pogosto porazdeljujejo po normalni porazdelitvi. Če je vzorec velik, npr. večji od 100 enot ($n > 100$), je porazdelitev vzorčnih ocen parametra $\hat{\theta}_i$ (npr. aritmetične sredine) približno normalna (glej sliko 48) [Košmelj B.].



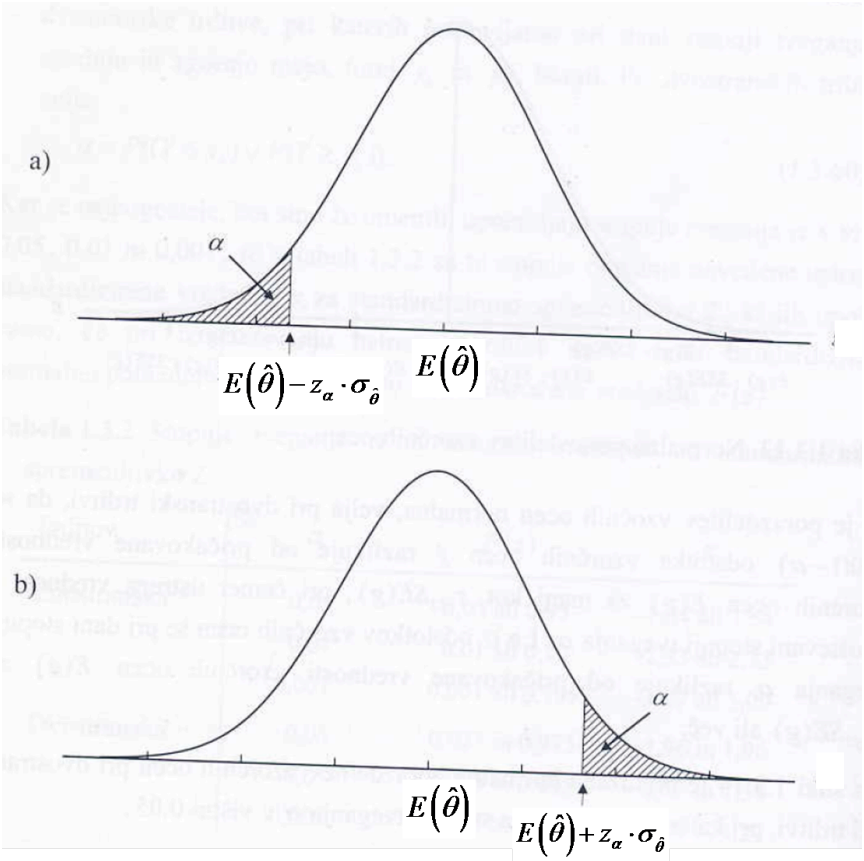
Slika 48: Normalna porazdelitev vzorčnih ocen parametra [Košmelj B.]

Na sliki 49 je prikazana normalna porazdelitev vzorčnih ocen pri dvostranski trditvi, pri kateri je upoštevana stopnja tveganja v višini α [Košmelj B.].



Slika 49: Normalna porazdelitev vzorčnih ocen pri dvostranski trditvi, pri kateri je upoštevana stopnja tveganja v višini α [Košmelj B.].

Na sliki 50 pa je prikazana normalna porazdelitev vzorčnih ocen pri enostranskih trditvah, pri katerih je upoštevana stopnja tveganja v višini α [Košmelj B.].



Slika 50: Normalna porazdelitev vzorčnih ocen pri enostranskih trditvah, pri katerih je upoštevana stopnja tveganja v višini α [Košmelj B.].

Pri dvostranski trditvi velja naslednji izraz [Košmelj B.]:

$$P\left[E\left(\hat{\theta}\right)-z_{\frac{\alpha}{2}}\cdot\sigma_{\hat{\theta}}<\hat{\theta}<E\left(\hat{\theta}\right)+z_{\frac{\alpha}{2}}\cdot\sigma_{\hat{\theta}}\right]=1-\alpha \tag{3.66}$$

Po drugi strani velja:

$$\begin{aligned}&P\left[E\left(\hat{\theta}\right)-z_{\frac{\alpha}{2}}\cdot\sigma_{\hat{\theta}}<\hat{\theta}<E\left(\hat{\theta}\right)+z_{\frac{\alpha}{2}}\cdot\sigma_{\hat{\theta}}\right]=\\&=1-\left[P\left[E\left(\hat{\theta}\right)+z_{\frac{\alpha}{2}}\cdot\sigma_{\hat{\theta}}<\hat{\theta}\right]+P\left[E\left(\hat{\theta}\right)-z_{\frac{\alpha}{2}}\cdot\sigma_{\hat{\theta}}>\hat{\theta}\right]\right]\end{aligned} \tag{3.67}$$

Torej očitno sledi:

$$P\left[E\left(\hat{\theta}\right)+z_{\frac{\alpha}{2}}\cdot\sigma_{\hat{\theta}}<\hat{\theta}\right]+P\left[E\left(\hat{\theta}\right)-z_{\frac{\alpha}{2}}\cdot\sigma_{\hat{\theta}}>\hat{\theta}\right]=\alpha \tag{3.68}$$

Pri enostranski trditvi za spodnjo mejo velja naslednji izraz [Košmelj B.]:

$$P\left[E\left(\hat{\theta}\right)-z_{\alpha}\cdot\sigma_{\hat{\theta}}\geq\hat{\theta}\right]=\alpha \tag{3.69}$$

Po drugi strani velja:

$$P\left[E\left(\hat{\theta}\right)-z_{\alpha}\cdot\sigma_{\hat{\theta}}\geq\hat{\theta}\right]=1-P\left[E\left(\hat{\theta}\right)-z_{\alpha}\cdot\sigma_{\hat{\theta}}<\hat{\theta}\right] \tag{3.70}$$

Torej očitno sledi:

$$P\left[E\left(\hat{\theta}\right)-z_{\alpha}\cdot\sigma_{\hat{\theta}}<\hat{\theta}\right]=1-\alpha \tag{3.71}$$

Pri enostranski trditvi za zgornjo mejo velja naslednji izraz [Košmelj B.]:

$$P\left[E\left(\hat{\theta}\right)+z_{\alpha}\cdot\sigma_{\hat{\theta}}\geq\hat{\theta}\right]=1-\alpha \tag{3.72}$$

Po drugi strani velja:

$$P\left[E\left(\hat{\theta}\right)+z_{\alpha}\cdot\sigma_{\hat{\theta}}\geq\hat{\theta}\right]=1-P\left[E\left(\hat{\theta}\right)+z_{\alpha}\cdot\sigma_{\hat{\theta}}<\hat{\theta}\right] \tag{3.73}$$

Torej očitno sledi:

$$P\left[E\left(\hat{\theta}\right)+z_{\alpha}\cdot\sigma_{\hat{\theta}}<\hat{\theta}\right]=\alpha \tag{3.74}$$

4 OPISNA STATISTIKA

Opisna statistika je skupina statističnih metod, ki se ukvarjajo s povzemanjem pridobljenih podatkov. Te metode iščejo opisne (meta) podatke o populaciji in njenih sestavnih delih, da bi ustvarile pregledni opis.

Temelje opisne statistike sestavlja grafični opis, katerega osnova je predstavitev s pomočjo grafov, tabelarni opis, ki ustvarja pregled s pomočjo tabel in statistični povzetki, ki na osnovi nekaterih računov predstavijo bistvene podatke (npr. srednje vrednosti).

Naloga opisne statistike je mnogokrat prikaz vsaj dveh lastnosti statistične populacije oziroma njenih gradnikov, statističnih enot. Prva izmed teh so skupne lastnosti statističnih enot, tj., kako se različne enote med seboj povezujejo oziroma kakšne skupne lastnosti ustvarjajo. Druga izmed dveh lastnosti pa je raznolikost statističnih enot, opisna statistika torej predstavi spremenljivost (variabilnost). Temeljni in največkrat uporabni podatki opisne statistike za prikaz skupnih lastnosti so aritmetična sredina, mediana in modus določene populacije, pa tudi nekatere druge vrednosti, kot kumulativne frekvence in kvantili. Spremenljivost statističnih enot v populaciji pa je moč predstaviti z varianco, standardnim odklonom in razponom vrednosti.

Opisna statistika je prvi korak analize podatkov, ko so bili ti zbrani in urejeni. Sledi ji nadaljnja obdelava in izvajanje sklepov, če je za njih moč izpeljati dovolj podatkov.

4.1 Urejanje in prikazovanje podatkov

Urejen in uporaben ter prijazen prikaz podatkov je prvi korak pri spoznavanju podatkov. Velikokrat je od tega odvisen uspeh nadaljnjih korakov statističnih raziskav [Brvar]. Čeprav gre za preprosta znanja pri uporabi opisne statistike, slednjo uporabniki pogostokrat podcenijo in si zato "pridelajo" nepotrebne probleme pri nadaljnji analitični obravnavi [Brvar].

Urejanje podatkov pomeni zapisati le-te na način, ki olajša prepoznavanje značilnosti opazovane množice, hkrati pa tudi že omogoča obdelavo podatkov z uporabo izbrane statistične metode. najenostavnejši način urejanja je prikazovanje podatkov s **statističnimi vrstami**. Prikaz podatkov s **strukturami** in **porazdelitvami** pa vsebinsko pomeni zgoščevanje podatkov, kar sicer zmanjšuje informativno vrednost podatkov, vendar pa nudi večje možnosti proučevanja kvantitativnih značilnosti opazovane statistične množice [Artenjak].

Statistične vrste

Statistična vrsta predstavlja zapis vrednosti spremenljivke na takšen način, kot smo jih pridobivali tekom raziskave. Običajno so tako zbrane vrednosti **neurejene**. Statistična vrsta pa je **urejena**, če so vrednosti razvrščene po določenem kriteriju. Ločimo [Artenjak]:

- časovne vrste,
- krajevne vrste in
- stvarne vrste.

Slika 51 prikazuje primere neurejene oz. urejene vrste, ter časovne, krajevne ali stvarne statistične vrste [Artenjak].

Primer 1.1 Neurejena in urejena statistična vrsta

A ***Ocene študenta so***: 10, 6, 6, 8, 9, 7
B ***Ocene študenta so***: 10, 9, 8, 7, 6, 6 ali 6, 6, 7, 8, 9, 10

Prikaza vrednosti spremenljivke pod B se imenujeta *ranžirna vrsta.*

Primer 1.2 Časovna statistična vrsta

Vrednost prodaje od leta 1997 do leta 2002:

Leto	*1997*	*1998*	*1999*	*2000*	*2001*	*2002*
Vrednost prodaje v mio SIT	*567*	*890*	*804*	*911*	*942*	*922*

Primer 1.3 Krajevna statistična vrsta

Vrednost prodaje v letu 2001 po trgih:

Trg	*A*	*B*	*C*	*D*	*E*
Vrednost prodaje v mio SIT	*241*	*361*	*125*	*54*	*161*

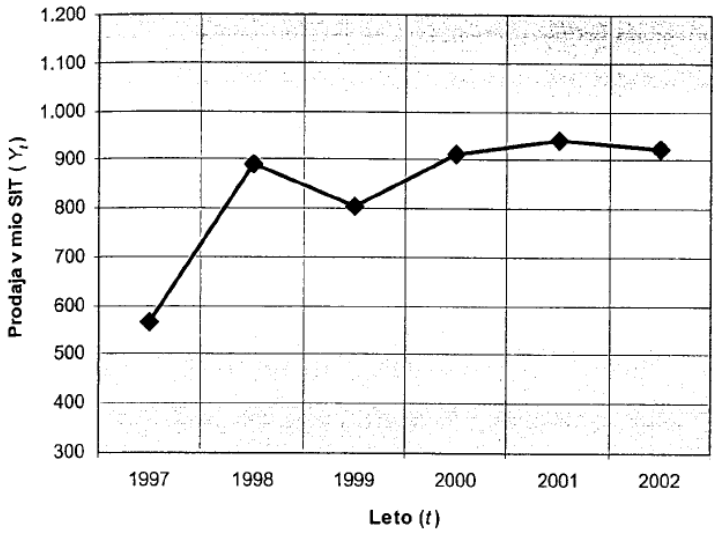
Primer 1.4 Stvarna statistična vrsta

Ocena študenta iz posameznih predmetov:

Predmet	*Računovodstvo*	*Statistika*	*Pravo*
Ocena	*7*	*8*	*10*

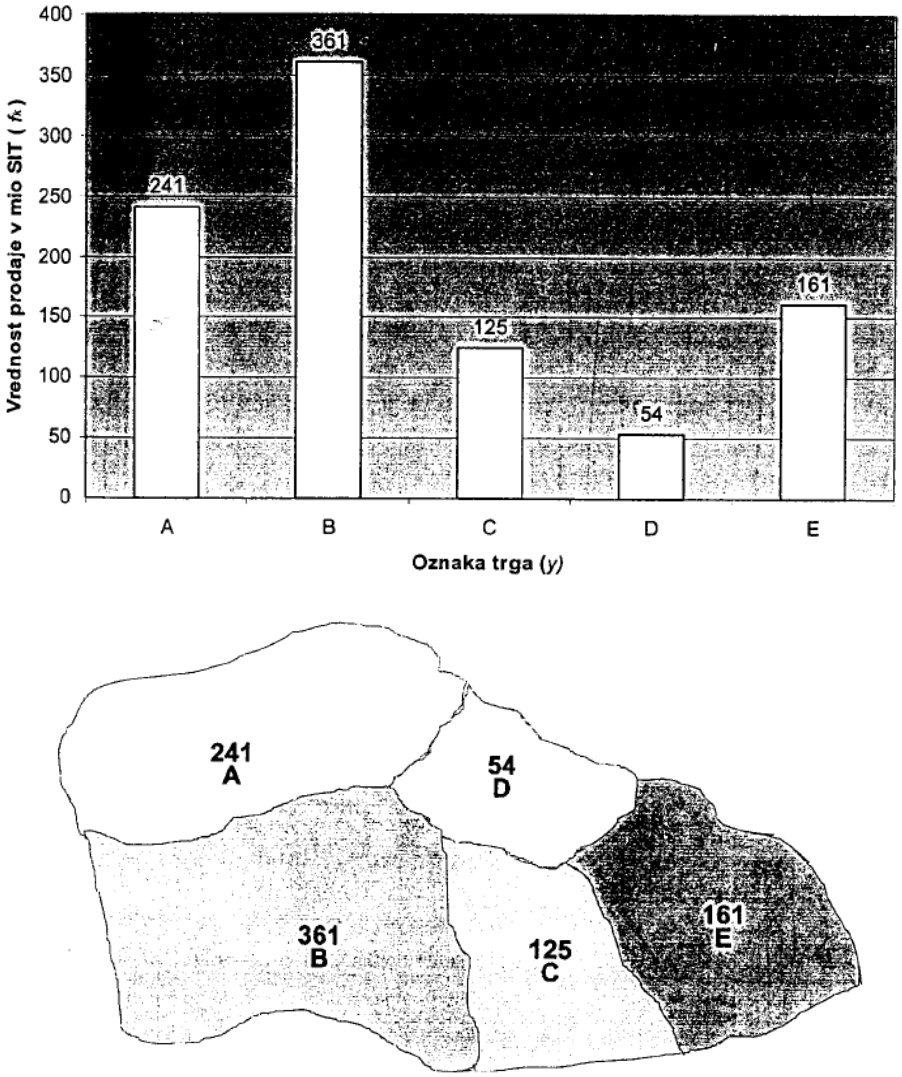
Slika 51: Primeri neurejene oz. urejene vrste, ter časovne, krajevne ali stvarne statistične vrste [Artenjak].

Statistične vrste pogosto prikazujemo z grafikoni, ker na enostaven in viden način posredujejo osnovne informacije o opazovanem pojavu. Časovno vrsto lahko prikažemo s stolpi ali pa z **linijskim grafikonom**, kot npr. na sliki 52. Slednji je bolj prikladen, saj je pri njem bolj razvidna tendenca spreminjanja vrednosti spremenljivke [Artenjak].



Slika 52: Primer linijskega grafikona za časovno vrsto [Artenjak]

Krajevne vrste običajno prikazujemo s stolpi ali **kartogrami** (glej sliko 52), stvarne vrste pa najpogosteje s stolpi [Artenjak].



Slika 52: Primer prikaza krajevne vrste s stolpi in s kartogrami [Artenjak]

Strukture

Skupine enot izbrane statistične množice, ki jim pripada ista vrednost opisne ali številske spremenljivke, imenujemo strukture. Število, ki pokaže pogostost pojavljanja vrednosti spremenljivke, je **frekvenca**. Strukture prikazujemo v **enostavnih**, **sestavljenih** in **kombiniranih** tabelah (preglednicah) [Artenjak].

Enostavna preglednica prikazuje strukturo le ene statistične množice po eni spremenljivki. Sestavljena preglednica predstavlja prikaz struktur več množic po isti spremenljivki. Kombinacijska preglednica pa npr. prikazuje strukturo ene same množice, ki jo

proučujemo po dveh spremenljivkah hkrati [Artenjak]. Več o strukturah podatkov lahko bralec zasledi v literaturi [Artenjak].

4.1.1 Frekvenčne porazdelitve

Statistične vrste, ki prikazujejo sestavo populacije po skupinah vrednosti številske spremenljivke, to je po razredih, imenujemo **frekvenčne porazdelitve**. Slednje prikazujejo porazdelitev vrednosti številske spremenjivke med enotami populacije in s tem variiranje vrednosti spremenljivke. Tako dobimo s prikazom frekvenčnih porazdelitev nazorno sliko o variiranju in zgoščevanju opazovanega pojava [Pfajfar].

Frekvenca razreda je število enot proučevane populacije, ki imajo vrednost spremenljivke v mejah razreda. Pri tem se pojavi vprašanje, koliko razredov oblikovati za spremenljivko, da bo čim nazorneje prikazana zakonitost njenega pojavljanja v proučevani populaciji. Če bo razredov preveč, prikaz ne bo nazoren in ne bo prišla do izraza osnovna zakonitost pojavljanja spremenljivke. Če pa je razredov premalo, pa se bo preveč zabrisala variabilnost vrednosti spremenljivke. trdnega pravila za določitev števila razredov žal ni. ***Izkustveno pravilo pravi, daj naj bi imeli od 8 do 16 razredov oz. od najmanj 5 do največ 20 razredov***. Za manše populacije naj bi imeli manj razredov in za večje več razredov. Poznamo pa tudi takimenovano **Sturgesovo pravilo**, ki pravi [Pfajfar]:

$$K \approx 1 + 3.3 \cdot \log N \qquad (4.1)$$

Primer 4.1.:

Imamo porazdelitev zaposlenih v RS septembra 2008 po velikosti bruto plače. Uradna statistika je pripravila frekvenčne porazdelitve za gospodarstvo kot celoto, po sektorjih dejavnosti in celo po upravnih enotah. Pri tem je za porazdelitve uporabila 19 razredov (K=19). Bilo je potrebno razvrstiti N = 678529 zaposlednih. Koliko razredov bi dalo Sturgesovo pravilo [Pfajfar]?

$$K \approx 1+3.3\cdot \log N \approx 1+3.3\cdot \log 678529 = 20.244 \qquad (4.2)$$

Za prikaz celotne populacije je izbrano število (4.2), to je 20 razredov, primerno, manj pa seveda za delne populacije. V vsakem primeru *K* izberemo tako, da sorazmerno narašča z *N*, pri čemer je potrebno preizkusiti več variant, paziti, da je razvidna osnovna zakonitost in variabilnost spremenljivke, ter se odločiti za najboljšo varianto [Pfajfar].

Primer 4.2.:

Denimo imamo opravka z diskretno populacijo, ki lahko zavzame vrednosti $x_1, x_2, ..., x_n, \quad n \le N$. Zanima nas število otrok na družino v neki krajevni skupnosti (lastnost pojava za populacijo), ki ima 100 družin (populacija). Iz matičnega urada so bili posredovani podatki, kot jih prikazuje slika 53 [Jesenko].

1,0,4,1,1,0,3,2,1,2,2,0,1,3,0,1,1,2,5,0,2,2,1,0,3,
2,2,1,2,0,1,4,1,1,0,3,2,2,2,1,0,3,1,2,2,1,4,0,1,1,
0,2,2,3,1,2,2,0,5,1,1,2,4,2,2,1,0,1,1,2,2,3,0,1,2,
2,2,1,0,0,2,2,1,4,2,0,1,1,0,2,2,3,1,0,3,1,1,2,2,0

Slika 53: Podatki o številu otrok v posamezni družini za 100 družin [Jesenko]

Iz slike 53 vidimo, da lahko opazovana spremenljivka zavzame 6 različnih vrednosti $0,1,2,3,4,5$, število vseh enot je pa *N=100*. Ker so podatki v tabeli na sliki 53 zelo nepregledni, jih preoblikujemo tako, da za vsako vrednost zapišemo, kolikokrat nastopi.

Številu enakih vrednosti x_i pravimo **frekvenca** f_i te vrednosti. Frekvence so seveda močno odvisne od števila vseh preučevanih enot *N*, medtem ko **relativne frekvence** niso odvisne od števila preučevanih enot. Relativna frekvenca je definirana na naslednji način [Jesenko]:

$$p_i = \frac{f_i}{N} \qquad (4.3)$$

Tabelo, v kateri prikazujemo vrednosti spremenljivke in pripadajoče frekvence, imenujemo **frekvenčna porazdelitev** (glej sliko 54) [Jesenko].

Vrednosti znaka X	Frekvenca
x_1	f_1
x_2	f_2
●	●
●	●
●	●
x_n	f_n

Slika 54: Tabela frekvenčne porazdelitve [Jesenko]

Za frekvenčno porazdelitev velja [Jesenko]:

$$\sum_{i=1}^{n} f_i = N \tag{4.4}$$

Tabelo, v kateri prikazujemo vrednosti spremenljivke in pripadajoče relativne frekvence, imenujemo **relativna frekvenčna porazdelitev** (glej sliko 55) [Jesenko].

Vrednosti znaka X	Relativna frekvenca
x_1	p_1
x_2	p_2
●	●
●	●
●	●
x_n	p_n

Slika 55: Tabela relativne frekvenčne porazdelitve [Jesenko]

Za relativno frekvenčno porazdelitev velja [Jesenko]:

$$\sum_{i=1}^{n} p_i = \sum_{i=1}^{n} \frac{f_i}{N} = \frac{1}{N} \sum_{i=1}^{n} f_i = \frac{N}{N} = 1 \tag{4.5}$$

V primeru števila otrok na družino v neki krajevni skupnosti na osnovi tabele na sliki 53 sestavimo tabelo frekvenčne porazdelitve, kot jo prikazuje slika 56 [Jesenko].

Število otrok	Frekvenca
0	20
1	31
2	33
3	9
4	5
5	2

Slika 56: Tabela frekvenčne porazdelitve za primer števila otrok na družino v neki krajevni skupnosti [Jesenko]

Na osnovi izraza (4.3) dobimo relativne frekvence: $\frac{20}{100}, \frac{31}{100}, \frac{33}{100}, \frac{9}{100}, \frac{5}{100}, \frac{2}{100}$, ki nam dajo tabelo relativne frekvenčne porazdelitve, prikazano na sliki 57 [Jesenko].

Število otrok	Relativna frekvenca
0	0,20
1	0,31
2	0,33
3	0,09
4	0,05
5	0,02

Slika 57: Tabela relativne frekvenčne porazdelitve za primer števila otrok na družino v neki krajevni skupnosti [Jesenko]

Razrede vpeljemo tedaj, če imamo diskretno populacijo, vendar bi tabeli frekvenčne in relativne frekvečne porazdelitve še vedno dali dokaj nepregledne rezultate [Nemec]. Vpeljemo pa jih tudi v primeru, ko imamo zvezno populacijo, kjer so vse vrednosti, ki jih opazovana spremenljivka lahko zavzame, različne med seboj. Sicer se v tem primeru posamezne vrednosti tudi večkrat ponovijo, vendar je to posledica zaokroževanja podatkov [Jesenko].

Razred je enolično določen s sredino razreda in širino razreda, ali pa s spodnjo in zgornjo mejo razreda [Jesenko]. Meje razredov ponavadi zberemo tako, da na njih ni nobene

vrednosti. Zaradi lažjega računanja jih je koristno izbirati tako, da so sredine razredov čim bolj enostavna števila, razredi pa enako široki [Jesenko].

Denimo imamo *n* razredov. Potem je frekvenca razreda število podatkov v razredu. Če vsi podatki leže na nekem intervalu $[a,b]$, potem je **širina razreda, če so razredi enako široki**, enaka [Jesenko]:

$$\Delta x = \frac{b-a}{n} \tag{4.6}$$

Meje razredov označimo z [Jesenko]:

$$a = r_0 < r_1 < r_2 < \ldots < r_i < \ldots < r_n = b \tag{4.7}$$

Tabelo, s katero zapišemo razrede in pripadajoče frekvence $f_1, f_2, \ldots, f_n$, imenujemo **razredna frekvenčna porazdelitev** [Jesenko] (glej sliko 58).

Razred	Frekvenca
$[r_0, r_1)$	f_1
$[r_1, r_2)$	f_2
.	.
.	.
.	.
.	.
$[r_{n-1}, r_n)$	f_n

Slika 58: Razredna frekvenčna porazdelitev [Jesenko]

Število vseh proučevanih enot dobimo pri razredni frekvenčni porazdelitvi na naslednji način [Jesenko]:

$$N = \sum_{i=1}^{n} f_i \tag{4.8}$$

relativno frekvenco razreda pa z izrazom [Jesenko]:

$$p_i = \frac{f_i}{N} \tag{4.9}$$

Tabelo, s katero zapišemo razrede in pripadajoče relativne frekvence $p_1, p_2, ..., p_n$, imenujemo **razredna relativna frekvenčna porazdelitev** [Jesenko] (glej sliko 59).

Razred	Relativna frekvenca
$[r_0, r_1)$	p_1
$[r_1, r_2)$	p_2
.	.
.	.
.	.
.	.
$[r_{n-1}, r_n)$	p_n

Slika 59: Razredna relativna frekvenčna porazdelitev [Jesenko]

Primer 4.3.:

Tabela na sliki 60 prikazuje življenjsko dobo 80 TV sprejemnikov določenega tipa v letih. Dolžina vseh razredov naj bo 1. Podatke zapišite v obliki razredne frekvenčne in razredne relativne frekvenčne porazdelitve [Jesenko]!

0,02;1,95;8,16;0,09;2,78;3,17;8,56;1,78;6,89;5,12;0,87;2,08;1,12;9,09;8,43;7,08
2,60;6,91;0,23;8,24;6,05;5,15;4,28;5,92;9,05;1,16;7,09;7,78;2,08;3,97;3,15;4,45
5,03;9,15;7,24;6,12;4,55;3,05;4,76;5,18;5,65;4,18;5,74;6,34;7,25;4,16;5,22;5,98
2,18;8,16;7,26;5,76;4,62;6,45;5,43;5,18;6,05;5,04;4,88;3,15;5,32;4,65;5,54;3,12
6,23;4,39;5,28;6,37;7,23;4,16;3,23;5,58;7,03;5,91;6,08;4,75;5,16;4,62;5,69;6,15

Slika 60: Življenjska doba 80 TV sprejemnikov določenega tipa v letih [Jesenko]

Razredna frekvenčna porazdelitev ima obliko, prikazano na sliki 61 [Jesenko].

Razred	Frekvenca
0 - 1	4
1 – 2	4
2 – 3	5
3 – 4	7
4 – 5	13
5 – 6	20
6 – 7	11
7 – 8	8
8 – 9	5
9 - 10	3

Slika 61: Razredna frekvenčna porazdelitev za življenjsko dobo televizorjev [Jesenko]

Na osnovi izraza (4.9) dobimo razredne relativne frekvence: $\frac{4}{80},\frac{4}{80},\frac{5}{80},\frac{7}{80},\frac{13}{80},\frac{20}{80},...,\frac{3}{80}$, ki nam dajo tabelo razredne relativne frekvenčne porazdelitve, prikazano na sliki 62 [Jesenko].

Razred	Relativna frekvenca
0 - 1	0,05
1 – 2	0,05
2 – 3	0,06
3 – 4	0,09
4 – 5	0,16
5 – 6	0,25
6 – 7	0,14
7 – 8	0,10
8 – 9	0,06
9 - 10	0,04

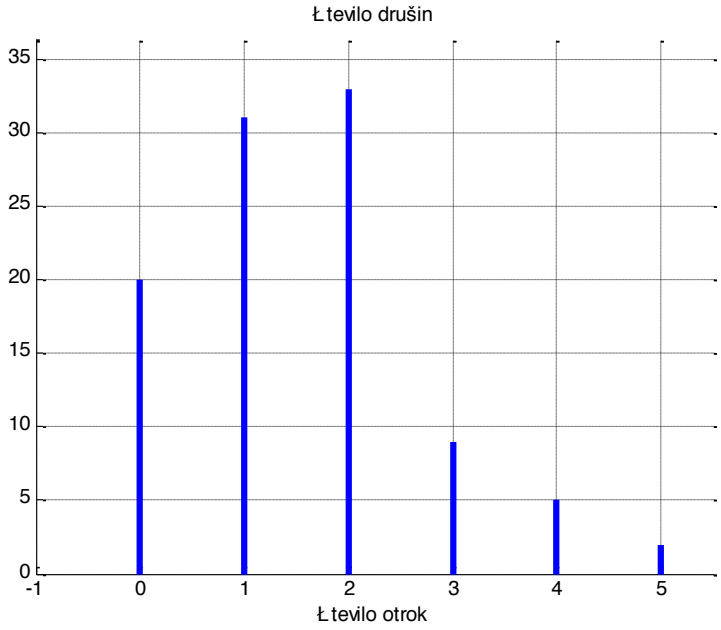
Slika 61: Razredna relativna frekvenčna porazdelitev za življenjsko dobo televizorjev [Jesenko]

4.1.2 Grafično prikazovanje frekvenčnih porazdelitev

Zaradi nazornejšega prikaza frekvenčnih porazdelitev je včasih ugodneje le-te prikazati v grafični obliki. Ločimo npr. med [Jesenko]:

- paličnimi diagrami,
- frekvenčnimi poligoni, in
- histogrami.

Na osnovi tabele na sliki 56 lahko narišemo palični diagram, ki ga prikazuje slika 62.



Slika 62: Palični diagram za primer števila otrok

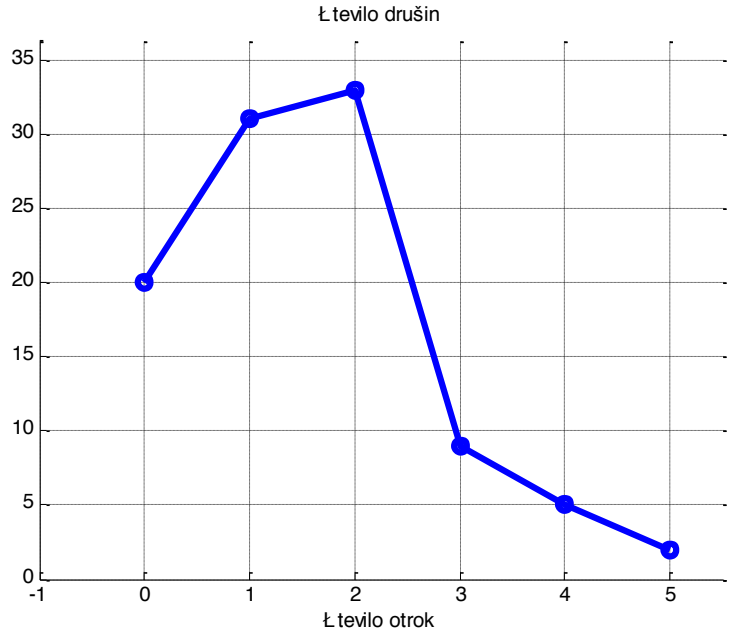
Pri generaciji slike 62 je bil uporabljen naslednji program v Matlabu:

```
% palicni.m

clc
clear
close all

x = input('Vnesi abscisni vektor')
y = input('Vnesi oordinatni vektor')
xstr = input('Vnesi naziv x osi','s')
ystr = input('Vnesi naziv y osi','s')

figure
hold on

for i=1:length(x)

plot(x(i)*ones(10,1)',linspace(0,y(i),10),'LineWidth',3)
end

grid

if min(x) == 0
   xx = min(x) - (x(2)-x(1));
else
   xx = 0.9*min(x)
end

axis([xx max(x)*1.1 0 max(y)*1.1])

title(ystr)
xlabel(xstr)
```

Izgled komandnega okna v Matlabu je naslednji:

```
Vnesi abscisni vektor[0 1 2 3 4 5]
x =
    0    1    2    3    4    5

Vnesi oordinatni vektor[20 31 33 9 5 2]
y =
   20   31   33    9    5    2

Vnesi naziv x osiŠtevilo otrok
xstr =
Število otrok

Vnesi naziv y osiŠtevilo družin
ystr =
Število družin
```

Na osnovi tabele na sliki 56 lahko narišemo tudi frekvenčni poligon, ki ga prikazuje slika 63.



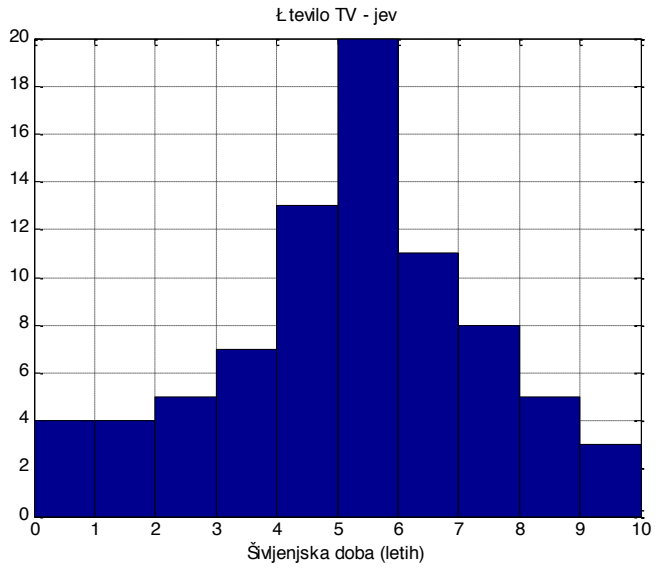
Slika 63: Frekvenčni poligon za primer števila otrok

Pri generaciji slike 63 je bil uporabljen naslednji program v Matlabu:

```
% frek_polig.m

clc
clear
close all

x = input('Vnesi abscisni vektor')
y = input('Vnesi oordinatni vektor')
xstr = input('Vnesi naziv x osi','s')
ystr = input('Vnesi naziv y osi','s')

figure
hold on

plot(x,y,'LineWidth',3)
hold on
plot(x,y,'o','LineWidth',3)

grid

if min(x) == 0
   xx = min(x) - (x(2)-x(1));
else
   xx = 0.9*min(x)
end

axis([xx max(x)*1.1 0 max(y)*1.1])

title(ystr)
xlabel(xstr)
```

Izgled komandnega okna je enak kot v primeru paličnega diagrama.

Na osnovi tabele na sliki 61 pa lahko narišemo histogram, ki ga prikazuje slika 64.



Slika 64: Histogram za primer življenjske dobe televizorjev

Pri generaciji slike 64 je bil uporabljen naslednji program v Matlabu:

```
% hist_uni.m

clc
clear
close all

x = input('Vnesi vektor razredov')
dx = input('Vnesi širino razreda')
f = input('Vnesi vektor frekvenc')

xstr = input('Vnesi naziv x osi','s')
ystr = input('Vnesi naziv y osi','s')

bar(x+dx/2,f,1)

grid

title(ystr)
xlabel(xstr)
```

Izgled komandnega okna v Matlabu je naslednji:

```
Vnesi vektor razredov0:9
x =
   0   1   2   3   4   5   6   7   8   9

Vnesi širino razreda1
dx =
   1

Vnesi vektor frekvenc[4 4 5 7 13 20 11 8 5 3]
f =
   4   4   5   7   13   20   11   8   5   3

Vnesi naziv x osiŽivljenjska doba (letih)
xstr =
Življenjska doba (letih)

Vnesi naziv y osiŠtevilo TV - jev
ystr =
Število TV - jev
```

Primer 4.4.:

Denimo smo dobili podatke o številu zaposlenih v različnih podjetjih. Tabela na sliki 65 prikazuje razredno frekvenčno porazdelitev [Nemec].

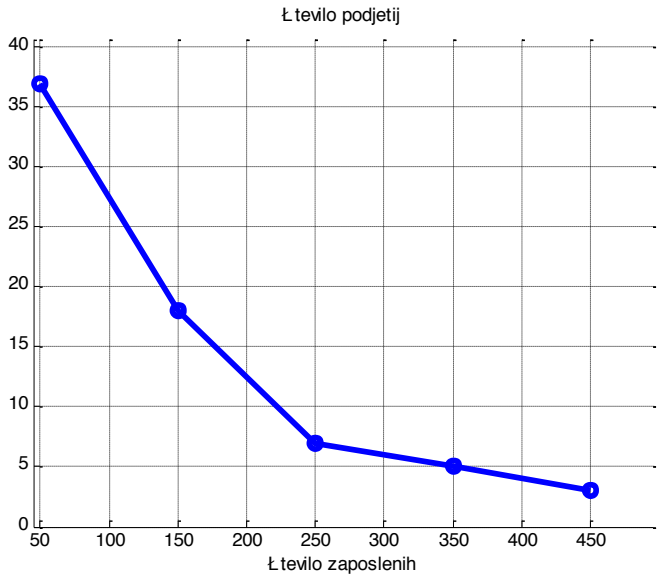
Število zaposlenih	**Število podjetij**	$x_{k\min}$	$x_{k\max}$	x_k
Do 100	37	0	100	50
Nad 100 do 200	18	100	200	150
Nad 200 do 3 00	7	200	300	250
Nad 300 do 4 00	5	300	400	350
Nad 400 do 5 00	3	400	500	450
Skupaj	70			

Slika 65: Razredna frekvenčna porazdelitev o številu zaposlenih v različnih podjetjih [Nemec]

Kot je razvidno iz slike 65, so označene tudi meje in sredine razredov. Če želimo razredno frekvenčno porazdelitev na sliki 65 predstaviti s frekvenčnim poligonom, moramo upoštevati sredine razredov, ki jih dobimo na naslednji način [Nemec]:

$$x_k = \frac{x_{k\min} + x_{k\max}}{2} \tag{4.10}$$

Razredni frekvenčni poligon je prikazan na sliki 66.

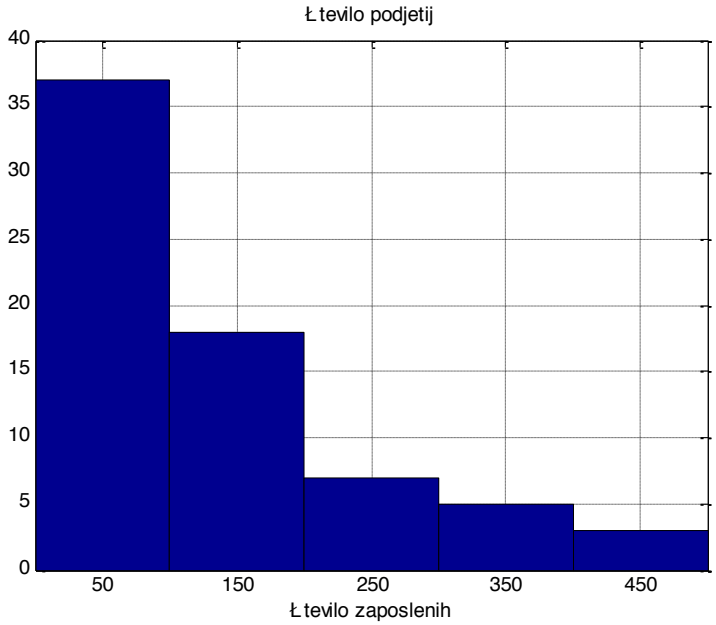


Slika 66: Razredni frekvenčni poligon za primer števila zaposlenih v podjetjih

Pri generaciji slike 66 smo uporabili enak program **frek_polig.m** kot v primeru števila otrok pri različnih družinah. Izgled komandnega okna v Matlabu je tokrat naslednji:

```
Vnesi abscisni vektor[50 150 250 350 450]
x =
   50   150   250   350   450

Vnesi oordinatni vektor[37 18 7 5 3]
y =
   37    18     7     5     3

Vnesi naziv x osiŠtevilo zaposlenih
xstr =
Število zaposlenih

Vnesi naziv y osiŠtevilo podjetij
ystr =
Število podjetij
```

Histogram za ta primer je prikazan na sliki 67.



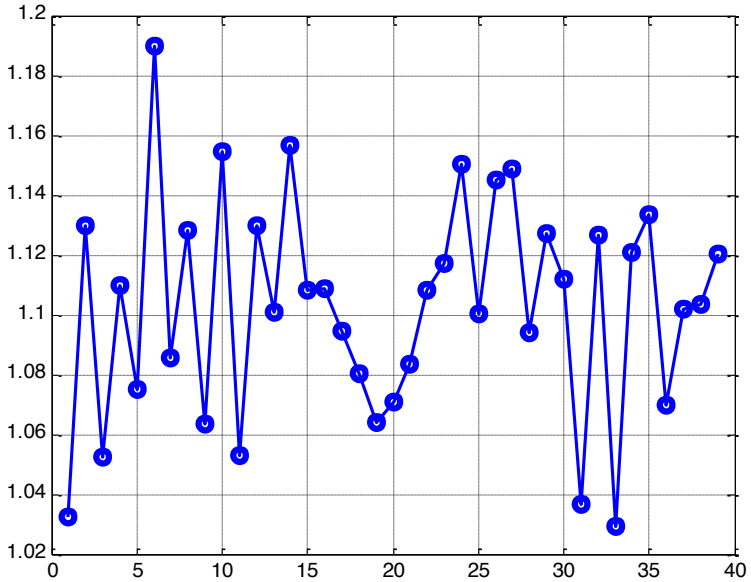
Slika 67: Histogram za primer števila zaposlenih v podjetjih

Pri generaciji slike 67 smo uporabili enak program **hist_uni.m** kot v primeru televizorjev. Izgled komandnega okna v Matlabu je tokrat naslednji:

```
Vnesi vektor razredov0:100:400
x =
    0   100   200   300   400

Vnesi širino razreda100
dx =
   100

Vnesi vektor frekvenc[37 18 7 5 3]
f =
   37    18     7     5     3

Vnesi naziv x osiŠtevilo zaposlenih
xstr =
Število zaposlenih

Vnesi naziv y osiŠtevilo podjetij
ystr =
Število podjetij
```

Primer 4.5.:

Dano imamo časovno vrsto, shranjeno v datoteki **yvalues.txt**, ki jo prikazuje slika 68 [Žibert].

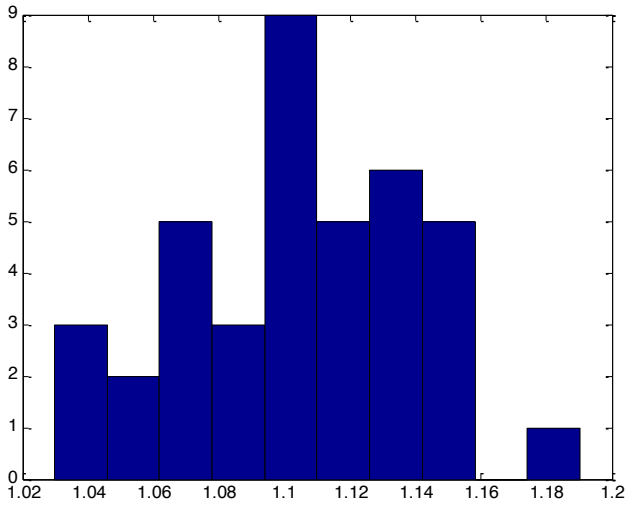


*Slika 68: Časovna vrsta, shranjena v datoteki **yvalues.txt** [Žibert]*

Program v Matlabu, ki je narisal sliko 68, je naslednji:

```
% primer časovne vrste v datoteki yvalues.txt
%
% plot_yvalues.m

clear
clc
close all

data = importdata('yvalues.txt');
y=data.data;

plot(y,'LineWidth',1.5)
hold on
plot(y,'o','LineWidth',3)

grid
```

Histogram za to časovno vrsto narišemo v Matlabu z ukazom hist(y), pri čemer dobimo sliko 69.



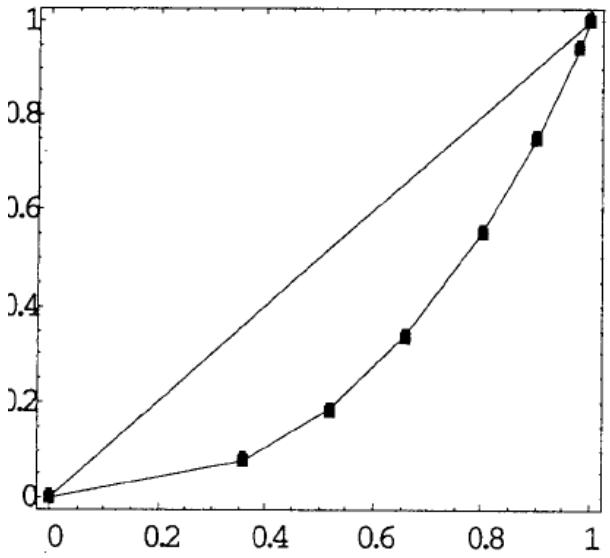
*Slika 69: Histogram za časovno vrsto, shranjeno v datoteki **yvalues.txt***

Koncentracija pojava

Porazdelitev agregata (skupne vrednosti spremenljivke za opazovane enote) med *n* enot statistične množice je lahko zelo različna. Če npr. opazujemo porazdelitev narodnega bogatstva med državami, lahko ugotovimo, da ima 30% razvitih držav v rokah kar 70% svetovnega bogatstva. Takšen pojav imenujemo **koncentracija** [Nemec]. V splošnem lahko za koncentracijo določenega pojava ugotovimo, da leži med dvema skrajnostima:

- popolna razpršenost (agregat je enakomerno porazdeljen med enote) in
- popolna koncentracija (agregat je skoncentriran v eni sami enoti).

Koncentracijo običajno analiziramo s takoimenovanim **Lorenzovim grafikonom**. Pri tem dobimo konveksen lik, katerega površina je sorazmerna koncentraciji pojava. Grafikon v primeru popolne razpršenosti leži na diagonali kvadrata s stranico 100%. Pri popolni koncentraciji pa zavzame poligon celotno površino pod diagonalo [Nemec]. Primer Lorenzovega grafikona je prikazan na sliki 70 [Jesenko].



Slika 70: Primer Lorenzovega grafikona [Jesenko]

Več o koncentraciji pojavov si lahko bralec pogleda v literaturi [Jesenko, Pfajfar, Nemec].

4.2 Kvantili in rangi

Za posamezno statistično enoto ali skupino statističnih enot je zelo pomembno, da poznamo njen položaj med ostalimi enotami preučevane statistične množice glede na opazovano značilnost. Kvantili so teoretično pomembni zaradi opredeljevanja značilnosti teoretičnih porazdelitev, praktično pa zaradi določanja položaja enote v statistični množici [Artenjak].

4.2.1 Rang

Naj bo $x_1 \leq x_2 \leq ... \leq x_N$ urejena vrsta vrednosti statistične spremenljivke. **Mesto *k*, ki ga ima neka vrednost x_k v urejeni vrsti, imenujemo rang R_{x_k}** [Jesenko]. Urejeni množici podatkov, bodisi v naraščajočem ali padajočem vrstnem redu, pa pravimo **ranžirna vrsta** [Jesenko].

Primer 4.6.:

Deset smučarjev je peljalo smuk po neki progi in so dosegli čase vožnje v sekundah, kot jih prikazuje slika 71 [Jesenko].

Smučar	Čas vožnje
S1	52,3
S2	51,8
S3	53,1
S4	52,6
S5	54,3
S6	51,6
S7	52,9
S8	53,7
S9	52,5
S10	53,1

Slika 71: Časi vožnje 10 smučarjev [Jesenko]

V nadaljevanju tvorimo ranžirno vrsto po naraščajočem vrstnem redu časa vožnje, kateri priredimo range. Tako dobimo rezultat, ki ga prikazuje slika 72 [Jesenko].

Smučar	Čas vožnje	Rang
S1	52,3	3
S2	51,8	2
S3	53,1	8
S4	52,6	5
S5	54,3	10
S6	51,6	1
S7	52,9	6
S8	53,7	9
S9	52,5	4
S10	53,1	7

Slika 72: Dodelitev ranga [Jesenko]

Za izračun rangov uporabimo naslednji program v Matlabu:

```
% rang.m

clear
close all
clc

x=[52.3 51.8 53.1 52.6 54.3 51.6 52.9 53.7 52.5
53.1]

disp('nesortirana vrsta in indeks smucarja:')

[x' (1:10)']


[xsort,ind_smuc] = sort(x);


disp('ranzirna vrsta    rang
indeks_smucarja')

R=[xsort' (1:10)' ind_smuc']

disp('indeks_smucarja     nesortirana vrsta
rang')

F = sortrows(R,3);

[F(:,3) x' F(:,2)]
```

Izgled komandnega okna je naslednji:

```
x =

  Columns 1 through 8

   52.3000   51.8000   53.1000   52.6000   54.3000   51.6000   52.9000   53.7000

  Columns 9 through 10

   52.5000   53.1000

nesortirana vrsta in indeks smucarja:

ans =

   52.3000    1.0000
   51.8000    2.0000
   53.1000    3.0000
   52.6000    4.0000
   54.3000    5.0000
   51.6000    6.0000
   52.9000    7.0000
   53.7000    8.0000
   52.5000    9.0000
   53.1000   10.0000

ranzirna vrsta    rang    indeks_smucarja

R =

   51.6000    1.0000    6.0000
   51.8000    2.0000    2.0000
   52.3000    3.0000    1.0000
   52.5000    4.0000    9.0000
   52.6000    5.0000    4.0000
   52.9000    6.0000    7.0000
```

```
   53.1000    7.0000    3.0000
   53.1000    8.0000   10.0000
   53.7000    9.0000    8.0000
   54.3000   10.0000    5.0000
indeks_smucarja     nesortirana vrsta    rang
ans =
    1.0000   52.3000    3.0000
    2.0000   51.8000    2.0000
    3.0000   53.1000    7.0000
    4.0000   52.6000    5.0000
    5.0000   54.3000   10.0000
    6.0000   51.6000    1.0000
    7.0000   52.9000    6.0000
    8.0000   53.7000    9.0000
    9.0000   52.5000    4.0000
   10.0000   53.1000    8.0000
```

V obravnavanem primeru ima smučar S_8 rang $R_{S_8} = 9$, kar pove veliko več kot pa sam čas vožnje, ki ga je dosegel. Vemo namreč, da je bil le en smučar počasnejši od njega, osem pa hitrejših, kar sam čas vožnje ne pove.

4.2.2 Določanje absolutnega in kvantilnega ranga iz frekvenčne porazdelitve

Pomanjkljivost ranga je v tem. da nam daje zadovoljivo informacijo o položaju vrednosti statistične spremenljivke v ranžirni vrsti le, če poznamo dolžino vrste (število vseh vrednosti v njej). Primerneje je, da mesto neke vrednosti v ranžirni vrsti izrazimo z njenim relativnim mestom in v ta namen vpeljemo **kvantilni rang** [Jesenko].

Kvantilni rang je enak (glej sliko 73) [Jesenko]:

$$P_{x_k} = \frac{R_{x_k} - 0.5}{N} \tag{4.11}$$

kjer je *N* dolžina ranžirne vrste.

Slika 73: Ilustracija kvantilnega ranga [Jesenko]

Primer 4.7.:

Kakšen je kvantilni rang vrednosti statistične spmremenjivke, ki ima v ranžirni vrsti mesto 52, če je vrsta dolga $N = 114$*?*

Na osnovi izraza (4.10) izračunamo:

$$P_{x_{52}} = \frac{R_{x_{52}} - 0.5}{114} = \frac{52 - 0.5}{114} = 0.45 \tag{4.11}$$

Kvantilni rang 0.45 pove, da je 45% vrednosti v ranžirni vrsti pred vrednostjo z rangom 52, kar pri naraščajočem vrstnem redu pomeni, da je 45% vrednosti manjših ali enakih vrednosti x_{52}.

Kadar so vrednosti podane v obliki frekvenčne porazdelitve, izračunamo **absolutni rang** za vrednost $r_{k-1} \le x \le r_k$ na naslednji način [Jesenko]:

$$R_{x_k} = \sum_{i=1}^{k-1} f_i + \frac{x_k - r_{k-1}}{r_k - r_{k-1}} \cdot f_k \tag{4.12}$$

kjer je:

$$F_{k-1} = \sum_{i=1}^{k-1} f_i \tag{4.13}$$

kumulativna frekvenca *k-1*. razreda.

Primer 4.8.:

Na vzorcu 200 krav molznic so proučevali količino namolženega mleka na dan. Podatki v litrih so podani na sliki 74 [Jesenko]. Izračunajte kvantilni rang količine 16.3 litrov namolženega mleka na dan.

Količina	Frekvenca
5-7	2
7-9	8
9-11	14
11-13	22
13-15	28
15-17	38
17-19	44
19-21	24
21-23	14
23-25	6

Slika 74: Razredna frekvenčna porazdelitev količine namolženega mleka krav molznic na dan [Jesenko]

Vrednost 16.3 se nahaja v 6. razredu: $16.3 \in [15,17]$. Izračunajmo najprej izraz $\sum_{i=1}^{k-1} f_i$:

$$\sum_{i=1}^{k-1} f_i = \sum_{i=1}^{5} f_i = f_1 + f_2 ... + f_5 = 2+8+14+22+28 = 74 \tag{4.14}$$

Sledi na osnovi izraza (4.12) za absolutni rang:

$$R_{16.3} = 74 + \frac{16.3-15}{17-15} \cdot 38 = 74 + \frac{1.3}{2} \cdot 38 = 98.7 \tag{4.15}$$

Kvantilni rang pa je:

$$P_{16.3} = \frac{R_{16.3} - 0.5}{200} = \frac{98.7 - 0.5}{200} = 0.491 \tag{4.16}$$

Torej je 49.1% namolžene količine mleka do vrednosti 16.3 litra na dan.

4.2.3 Kvantili

Kvantilni rang izračuna za vsako vrednost x_k delež ranžirne vrste oz. frekvenčne porazdelitve (urejene vrste), v katerem so vse vrednosti do x_{k-1} manjše od x_k. Tiste vrednosti x_k, ki delijo urejeno vrsto na *k* delov z enakim deležem enot, imenujemo **kvantile** [Brvar].

Kvantile delimo na [Brvar]:

- **Mediano**, ki urejeno vrsto razdeli na dva dela,
- **Kvartile**, ki zavzamejo tri vrednosti spremenljivke, s katero so opazovane enote v urejeni vrsti razdeljene na štiri enake dele,
- **Kvintile**, ki zavzamejo štiri vrednosti spremenljivke, s katero so opazovane enote v urejeni vrsti razdeljene na pet enakih delov,
- **Sekstile**, ki zavzamejo pet vrednosti opazovane spremenljivke in je z njimi urejena vrsta razdeljena na šest enakih delov,
- **Oktile**, ki zavzamejo sedem vrednosti opazovane spremenljivke in je z njimi urejena vrsta razdeljena na osem enakih delov,
- **Decile**, ki zavzamejo devet vrednosti opazovane spremenljivke in je z njimi urejena vrsta razdeljena na deset enakih delov,
- **Centile**, ki zavzamejo 99 vrednosti opazovane spremenljivke in je z njimi urejena vrsta razdeljena na 100 enakih delov.

Prvi kvartil Q_1 je tista vrednost opazovane spremenljivke, pri kateri ima 25% enot manjšo vrednost, 75% enot pa večjo vrednost. Drugi kvartil Q_2 ali mediana je tista vrednost opazovane spremenljivke, pri kateri ima 50% enot manjšo vrednost, 50% enot pa večjo vrednost. Tretji kvartil Q_3 je tista vrednost opazovane spremenljivke, pri kateri ima 75% enot manjšo vrednost, 25% enot pa večjo vrednost [Brvar]. Podobna logika velja za ostale tipe kvantilov.

Katerikoli kvantil izračunamo s preureditvijo izraza (4.12), pri čemer upoštevamo tudi izraz (4.11) z zanemaritvijo 0.5:

$$R_{x_k} = P_{x_k} \cdot N = \underbrace{\sum_{i=1}^{k-1} f_i}_{F_{k-1}} + \frac{\%_k - r_{k-1}}{\Delta\%_k} \cdot f_k$$

$$P_{x_k} \cdot N - F_{k-1} = \frac{\%_k - r_{k-1}}{\Delta\%_k} \cdot f_k \qquad (4.17)$$

$$\left(P_{x_k} \cdot N - F_{k-1}\right)\frac{(\Delta\%_k)}{f_k} = \%_k - r_{k-1}$$

$$\%_k = r_{k-1} + \frac{(\Delta\%_k)\cdot\left(P_{x_k} \cdot N - F_{k-1}\right)}{f_k} = kvantil$$

Primer 4.9.:

Izračunajmo prvi, drugi in tretji kvantil za frekvenčno porazdelitev ovadenih oseb v Sloveniji v letu 2006 po starosti (glej sliko 75) [Brvar].

Razredi starosti	Frekvenca (f)	Kumulativna frekvenca (F)	Relativna kumulativna frekvenca (H)
14–17	1550	1550	0,09
18–20	1790	3340	0,19
21–30	5279	8619	0,49
31–40	3731	12350	0,70
41–50	3087	15437	0,87
51–70[46]	2312	17749	1,00
Skupaj	17749		

Slika 75: Frekvenčna porazdelitev ovadenih oseb v Sloveniji v letu 2006 po starosti [Brvar]

Kvantilni rang prvega kvartila Q_1 znaša 0.25. To pomeni, da je 25% vrednosti manjših od vrednosti Q_1, ki jo iščemo. Če je celotno število oseb 17749, potem je 25% te vrednosti enako 4437.25. Za določitev začetka kvantilnega razreda opazujemo kumulativne frekvence. Kumulativna frekvenca za razred 18-20 je 3340, za razred 21-30 pa 8619. Ker se vrednost 4437.25 nahaja med vrednostima 3340 in 8619, sklepamo, da je kvantilni razred tisti, kjer ima kumulativna frekvenca vrednost 8619, začetek tega razreda pa je 21.

Sledi:

$$\%_k = r_{k-1} + \frac{(\Delta\%_k)\cdot\left(P_{x_k}\cdot N - F_{k-1}\right)}{f_k} = kvantil$$
$$Q_1 = 21 + \frac{(10)\cdot(0.25\cdot 17749 - 3340)}{5279} = 23.08 \qquad (4.18)$$

Prvi kvartil znaša 23.08 let, kar pomeni, da je 25% ovadenih oseb v RS v letu 2006 mlajših od 23.08 let.

Podobno izračunamo drugi kvartil ali mediano, ki leži v razredu 31-40 let:

$$Q_2 = 31 + \frac{(10)\cdot(0.5\cdot 17749 - 8619)}{3731} = 31.68 \qquad (4.19)$$

Torej je polovica ovadenih oseb mlajših od 31.68 let, polovica pa starejša od 31.68 let.

Tretji kvartil leži v razredu 41-50, saj je $0.75\cdot 17749 = 13311.75$. Njegova vrednost je:

$$Q_3 = 41 + \frac{(10)\cdot(0.75\cdot 17749 - 12350)}{3087} = 44.11 \qquad (4.20)$$

Torej je 75% povzročiteljev mlajših od 44.11 let, 25% pa starejših od 44.11 let.

4.3 Srednje vrednosti

Pri opisovanju značilnosti numeričnih spremenljivk ne navajamo vseh vrednosti spremenljivk, pač pa določimo le nekaj vrednosti, ki morajo čimbolj stvarno predstavljati vse opazovane vrednosti. Takšne vrednosti so srednje vrednosti. Razlikujemo med [Artenjak]:

- Aritmetično sredino,
- Mediano,

- Geometrično sredino,
- Harmonično sredino, in
- Modusom.

Aritmetično sredino izračunamo iz vseh vrednosti. Modus je gostiščnica, okoli katere je največja zgostitev vrednosti. Mediana je središčnica, ki predstavlja sredino ranžirne vrste. Geometrijska sredina je uporabna za izračun različnih statističnih parametrov pri časovnih vrstah. Harmonična sredina se uporablja dokaj redko, vendar je včasih edina srednja vrednost, s katero lahko dobimo vsebinsko pravilne rešitve [Artenjak].

4.3.1 Aritmetična sredina

Vrednost spremenljivke za *i-to* enoto je izid vsote splošnih ($\bar{y}$) in individualnih vplivov (e_i):

$$y_i = \bar{y} + e_i \tag{4.21}$$

Domnevamo lahko, da se individualni vplivi v medsebojni vsoti ozničijo:

$$\sum_{i=1}^{N} e_i = \sum_{i=1}^{N} (y_i - \bar{y}) = 0 \tag{4.22}$$

Odtod dobimo aritmetično sredino [Artenjak]:

$$\sum_{i=1}^{N} (y_i - \bar{y}) = \sum_{i=1}^{N} (y_i) - \sum_{i=1}^{N} (\bar{y}) = \sum_{i=1}^{N} (y_i) - \bar{y} \cdot N = 0 \tag{4.23}$$

$$\bar{y} = \frac{1}{N} \sum_{i=1}^{N} (y_i)$$

Primer 4.10.:

Desetkrat smo opazovali porabo avtomobila, izraženo v litrih, na 100 km, ter dobili naslednje podatke: 8.9, 8.4, 10.1, 8.7, 7.8, 10.3, 9.5, 9.4, 8.8, 9.0. Kakšna je povprečna poraba goriva na 100 prevoženih kilometrov?

$$\bar{y} = \frac{1}{10}\sum_{i=1}^{10}(y_i) = \frac{1}{10}(8.9 + 8.4 + ... + 9.0) = 9.09 \tag{4.24}$$

Matlab bi ta rezultat izračunal na naslednji način:

```
y=[8.9, 8.4, 10.1, 8.7, 7.8, 10.3, 9.5, 9.4, 8.8, 9.0]
y_sr = mean(y)
```

lahko pa tudi z ukazom:

```
y_sr=sum(y)/length(y)
```

Aritmetična sredina za linearno funkcijo $z(x, y) = a + bx + cy$ je enaka [Artenjak]:

$$\bar{z} = a + b\bar{x} + c\bar{y} \tag{4.25}$$

Dokaz lahko bralec zasledi v literaturi [Artenjak].

Denimo imamo opravka z vsoto kvadratov odklonov individualnih vrednosti y_i od neke konstante A. Slednja je najmanjša tedaj, če velja: $A^* = \bar{y}$.

Dokaz:

Imamo:

$$S = \sum_{i=1}^{N}(y_i - A)^2 \tag{4.25}$$

Odvajajmo po A in enačimo z 0:

$$\begin{aligned}
&\frac{dS}{dA} = \frac{d}{dA}\left[\sum_{i=1}^{N}(y_i - A)^2\right] = \sum_{i=1}^{N}\frac{d}{dA}\left[(y_i - A)^2\right] = \\
&= \sum_{i=1}^{N}\left[2(y_i - A)(-1)\right] = -2\cdot\sum_{i=1}^{N}(y_i - A) = 0 \\
&\sum_{i=1}^{N}(y_i) - N\cdot A = 0 \\
&A^* = \frac{1}{N}\sum_{i=1}^{N}(y_i) = \overline{y}
\end{aligned} \tag{4.26}$$

Skupno aritmetično sredino množice izračunamo iz aritmetičnih sredin delnih množic s pomočjo izraza [Artenjak]:

$$\overline{y} = \frac{N_1\overline{y}_1 + N_2\overline{y}_2 + \ldots + N_r\overline{y}_r}{N_1 + N_2 + \ldots + N_r} = \frac{\sum_{k=1}^{r} N_k\overline{y}_k}{\sum_{k=1}^{r} N_k} = \frac{\sum_{k=1}^{r} N_k\overline{y}_k}{N} \tag{4.27}$$

Primer 4.11.:

Prometni policisti so na nekem odseku tri dni merili hitrost motornih vozil in izmerili naslednje povprečne hitrosti:

1.dan: 68.5 km/h.....230 vozil
2.dan: 66.2 km/h.....320 vozil
3.dan: 64.0 km/h.....188 vozil

Kolikšna je povprečna hitrost vozil na dotičnem odseku?
Imamo:

$$\begin{aligned}
N_1 &= 230, \overline{y}_1 = 68.5 \\
N_2 &= 320, \overline{y}_2 = 66.2 \\
N_3 &= 188, \overline{y}_3 = 64.0
\end{aligned} \tag{4.28}$$

Povprečna hitrost vozil na dotičnem odseku je:

$$\bar{y} = \frac{N_1\bar{y}_1 + N_2\bar{y}_2 + N_3\bar{y}_3}{N_1 + N_2 + N_3} = \frac{230 \cdot 68.5 + 320 \cdot 66.2 + 188 \cdot 64}{230 + 320 + 188} = 66.35\ km/h \tag{4.29}$$

Velikokrat se zgodi, da so podatki, ki popisujejo populacijo, podani v obliki frekvenčne porazdelitve. **Aritmetično sredino za diskretno frekvenčno porazdelitev** izračunamo z izrazom [Jesenko]:

$$\bar{x} = \frac{\sum_{i=1}^{n} f_i \cdot x_i}{N} \tag{4.30}$$

Aritmetično sredino za zvezno frekvenčno porazdelitev pa računamo z izrazom [Jesenko]:

$$\bar{x} = \frac{\sum_{i=1}^{n} f_i \cdot x_{si}}{N} \tag{4.31}$$

kjer je x_{si} sredina *i-tega* razreda.

Primer 4.12.:

Na vzorcu 200 krav molznic so dobili frekvenčno porazdelitev za količino namolženega mleka, ki smo jo prikazali že na sliki 74. Izračunajte aritmetično sredino [Jesenko]!

Na osnovi izraza (4.31) zapišemo:

$$\begin{aligned} \bar{x} &= \frac{\sum_{i=1}^{n} f_i \cdot x_{si}}{N} = \frac{\sum_{i=1}^{10} f_i \cdot x_{si}}{200} = \frac{1}{200}\left(f_1 \cdot x_{s1} + \ldots + f_{10} \cdot x_{s10}\right) = \\ &= \frac{1}{200}\left(2 \cdot 6 + 8 \cdot 8 + 14 \cdot 10 + 22 \cdot 12 + 28 \cdot 14 + 38 \cdot 16 + 44 \cdot 18 + 24 \cdot 20 + 14 \cdot 22 + 6 \cdot 24\right) = \\ &= 16.02 \end{aligned} \tag{4.32}$$

Aritmetična sredina je torej 16.02 litra mleka na dan in predstavlja količino namolženega mleka na kravo, če ne bi bilo individualnih učinkov. Če le-teh ne bi bilo, bi pri vseh kravah namolzli enako količino mleka na dan.

4.3.2 Mediana

Naj bodo $x_1 \le x_2 \le ... \le x_N$ vrenosti, urejene po velikosti. Potem mediano določimo na naslednji način [Jesenko]:

$$Me = \begin{cases} x_{\frac{N+1}{2}}, & N\ lih \\ \dfrac{x_{\frac{N+2}{2}} + x_{\frac{N}{2}}}{2}, & N\ sod \end{cases} \tag{4.33}$$

V primerjavi z aritmetično sredino je mediana stabilnejša, ker je manj občutljiva na spremembe podatkov.

Primer 4.13.:

Imamo 40 podatkov sortiranih v urejeno vrsto po velikosti, kot jih prikazuje slika 76. Izračunajte mediano [Brvar]!

29, 33, 33, 34, 39, 42, 43, 44, 44, 45, 49, 49, 54, 55, 56, 57, 59, 60, 61, 61, 66, 66, 67, 67, 67, 72, 72, 73, 74, 77, 82, 83, 86, 87, 88, 88, 89, 91, 95, 95.

Slika 76: 40 podatkov urejenih po velikosti [Brvar]

Ker je *N = 40 = sodo število*, bomo uporabili izraz:

$$Me = \frac{x_{\frac{N+2}{2}} + x_{\frac{N}{2}}}{2} = \frac{x_{\frac{40+2}{2}} + x_{\frac{40}{2}}}{2} = \frac{x_{21} + x_{20}}{2} = \frac{66+61}{2} = 63.5 \tag{4.34}$$

Če imamo dane podatke, urejene v tabeli frekvenčne porazdelitve, si pomagamo z izrazom (4.17):

$$\%_k = r_{k-1} + \frac{(\Delta\%_k)\cdot\left(P_{x_k}\cdot N - F_{k-1}\right)}{f_k} = r_{k-1} + \frac{(\Delta\%_k)\cdot\left(R_{x_k} - F_{k-1}\right)}{f_k} = r_{k-1} + \frac{(\Delta\%_k)\cdot\left(\frac{N}{2} - F_{k-1}\right)}{f_k} \tag{4.35}$$

Sledi:

$$Me = x_{spk} + \frac{(\Delta\%_k)\cdot\left(\frac{N}{2} - F_{k-1}\right)}{f_k}$$

kjer je $x_{spk} = r_{k-1}$ spodnja meja medianinega razreda, f_k njegova frekvenca in $\Delta\%_k$ njegova širina. Rang R_{x_k} predstavlja mesto mediane v urejeni vrsti, ki je $\frac{N}{2}$.

Primer 4.14.:

Tabela na sliki 77 prikazuje frekvenčno porazdelitev starosti povzročiteljev hujših prometnih nesreč v RS v letu 2006. Izračunajte mediano [Brvar]!

Razredi starosti	Frekvenca (f_i)	Kumulativna frekvenca (F_i)	Relativna kumulativna frekvenca (H_i)
1–10	9	9	0,01
10–20	119	128	0,12
20–30	320	448	0,43
30–40	209	657	0,63
40–50	148	805	0,78
50–60	127	932	0,90
60–70	46	978	0,94
70–80	47	1025	0,99
80–90	11	1036	1,00
Skupaj	1036		

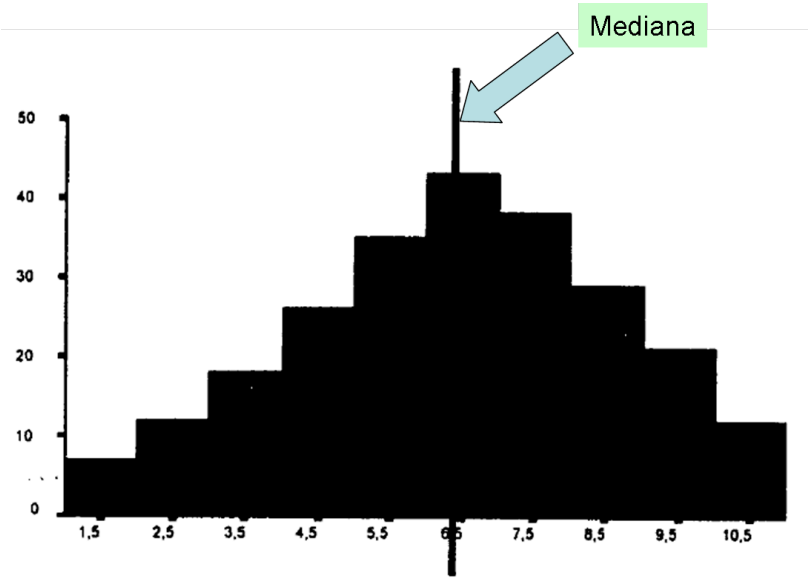
Slika 77: Frekvenčna porazdelitev starosti povzročiteljev prometnih nesreč v RS v letu 2006 [Brvar]

Ker je $\frac{N}{2} = 518$, hitro ugotovimo, da se mediana nahaja v razredu 30-40. Tako dobimo:

$$Me = x_{spk} + \frac{(\Delta\%_k)\cdot\left(\frac{N}{2} - F_{k-1}\right)}{f_k} = 30 + \frac{(10)\cdot(518-448)}{209} = 33.35 \tag{4.36}$$

Torej starost 33.3 let deli populacijo povzročiteljev hujših prometnih nesreč v RS v letu 2006 na dva enaka dela: 50% povzročiteljev je mlajših od 33.3 let, 50% pa starejših.

Grafično mediana pomeni tisto točko na abscisni osi, ki razdeli ploščino histograma na dva enaka dela (glej sliko 78) [Jesenko].



Slika 78: Grafični pomen mediane na histogramu [Jesenko]

V nadaljevanju se vrnimo na primer časovne vrste, shranjene v datoteki **yvalues.txt**, ki jo prikazuje slika 68. Slabost aritmetične sredine je v tem, da je zelo občutljiva na vrednosti, ki zelo odstopajo (ang. outliers). Več kot je takšnih vrednosti, slabša je ocena. Tu se mediana gotovo mnogo bolje obnese.

V Matlabu bi aritmetično sredino izračunali na naslednji način:

```
% primer aritmeticne sredine časovne vrste v datoteki yvalues.txt
%
% aritsre_yvalues.m

clear
clc
close all

data = importdata('yvalues.txt');
y=data.data;

N = length(y)

disp('Aritmeticna sredina na 1. nacin:')

arit_sred1 = mean(y)
```

```
disp('Aritmeticna sredina na 2. nacin:')

arit_sred2 = sum(y)/N
```

Izgled komandnega okna bi bil:

```
N =

   39

Aritmeticna sredina na 1. nacin:

arit_sred1 =

   1.1035

Aritmeticna sredina na 2. nacin:

arit_sred2 =

   1.1035
```

V Matlabu bi mediano izračunali na naslednji način:

```
% primer mediane časovne vrste v datoteki yvalues.txt
%
% med_yvalues.m

clear
clc
close all

data = importdata('yvalues.txt');
y=data.data;

y = sort(y);
yold = y;

while 1 == 1

    ch = input('Želiš sodo (1) ali liho (0) število podatkov');
    if ch == 1
       y = y(1:length(y)-1);
    end

    N = length(y)

    disp('mediana na 1. način:')
    median(y)

    disp('mediana na 2. način:')

    if abs(N/2-round(N/2))>0   % lih
       med = y((N+1)/2)
    else                       % sod
       med = (y(N/2)+y((N+2)/2))/2
    end

    ch = input('Želiš končati 1-Da,0-Ne')
    if ch == 1
       break
    end

    y = yold;

end
```

Izgled komandnega okna bi bil:

```
Želiš sodo (1) ali liho (0) število podatkov0
N =
    39
mediana na 1. način:
ans =
    1.1088
mediana na 2. način:
med =
    1.1088
Želiš končati 1-Da,0-Ne0
ch =
     0
Želiš sodo (1) ali liho (0) število podatkov1
N =
    38
mediana na 1. način:
ans =
    1.1064
mediana na 2. način:
med =
    1.1064
Želiš končati 1-Da,0-Ne
```

4.3.3 Geometrična sredina

Včasih se vrednosti statisitčne spremenljivke izražajo z odstotki, kot npr. dvig plač, odstotek osipa študentov, itn. Tedaj aritmetična sredina nima posebnega pomena, ker se skupna sprememba ne popisuje z vsoto, pač pa produktom [Jesenko].

Geometrijsko sredino po potrebi uporabljamo tudi v primeru analize časovnih vrst, ko želimo npr. spoznati povprečni verižni indeks, povprečni koeficient dinamike ali povprečno stopnjo rasti, da bi lahko analizirali preteklo obdobje in predvideli prihodnji razvoj vrednosti spremennljivke [Artenjak].

Geometrijska srednja vrednost je definirana na naslednji način [Jesenko, Brvar]:

$$\bar{x}_G = \sqrt[N]{x_1 \cdot x_2 \cdot \ldots \cdot x_N} \tag{4.37}$$

Primer 4.15.:

Pet let zaporedoma smo zasledovali rast prodaje v nekem trgovskem podjetju. Podatki so prikazani na sliki 79. Izračunajte povprečni dvig prodaje [Jesenko]!

Leto	1.leto	2.leto	3.leto	4.leto	5.leto
Dvig prodaje	6%	12%	7%	11%	4%

Slika 79: Rast prodaje v nekem trgovskem podjetju [Jesenko]

Uporabimo izraz (4.37):

$$\bar{x}_G = \sqrt[5]{x_1 \cdot x_2 \cdot \ldots \cdot x_5} = \sqrt[5]{1.06 \cdot 1.12 \cdot 1.07 \cdot 1.11 \cdot 1.04} = 1.0796 \tag{4.38}$$

Povprečen dvig prodaje je torej cca. 8%.

Primer 4.16.:

Štejemo insekte na rastlinah v 5 gredicah. V prvi smo prešteli 10 insektov, v drugi 1 insekt, v tretji 1000 insektov, v četrti 1 insekt, ter v peti 10 insektov. Izračunajte povprečje insektov v gredicah [Žibert].

Če bi šli računati aritmetično sredino, bi dobili:

$$\frac{10+1+1000+1+10}{5} = 204.4$$

Takšen rezultat ni zadovoljiv, saj ne prikazuje "povprečnega" dogajanja v gredicah, ker se insekti lahko izredno hitro razmnožujejo.

Bolj pravilen je izračun z uporabo geometrične sredine, ki da rezultat:

$$\bar{x}_G = \sqrt[5]{x_1 \cdot x_2 \cdot \ldots \cdot x_5} = \sqrt[5]{10 \cdot 1 \cdot 1000 \cdot 1.10} = 10 \tag{4.39}$$

V Matlabu bi ta rezultat izračunali na naslednji način:

```
insekti=[10 1 1000 1 10];
gs=geomean(insekti)
gs =
   10.0000
```

4.3.4 Harmonična sredina

Uporablja se za izračun povprečij statističnih koeficientov, npr. povprečne hitrosti [Bajt]. Izračunamo jo na naslednji način [Jesenko, Brvar]:

$$\bar{x}_H = \frac{N}{\sum_{i=1}^{N}\left(\frac{1}{x_i}\right)} \tag{4.40}$$

Primer 4.17.:

Za tri delavce, ki porabijo 3, 4, in 5 minut za izdelavo 1 kosa enakega izdelka, ugotovite povprečno porabo časa za izdelavo 1 kosa izdelka!

Na osnovi izraza (4.40) dobimo:

$$\bar{t}_H = \frac{N}{\sum_{i=1}^{N}\left(\frac{1}{t_i}\right)} = \frac{3}{\sum_{i=1}^{3}\left(\frac{1}{t_i}\right)} = \frac{3}{\frac{1}{t_1}+\frac{1}{t_2}+\frac{1}{t_3}} = \frac{3}{\frac{1}{3}+\frac{1}{4}+\frac{1}{5}} = 3.83\ \text{min}\,ut \tag{4.41}$$

Kot se izkaže, bi z uporabo aritmetične sredine dobili napačen rezultat (glej [Artenjak]).

Primer 4.18.:

Kraja A in B sta oddaljena 120 km. Voznik je od kraja A do kraja B vozil s hitrostjo 120 km/h, v obratni smeri pa 80 km/h. Izračunajte povprečno hitrost vožnje v obe smeri [Brvar]!

Na osnovi izraza (4.40) dobimo:

$$\bar{v}_H = \frac{N}{\sum_{i=1}^{N}\left(\frac{1}{v_i}\right)} = \frac{2}{\sum_{i=1}^{2}\left(\frac{1}{v_i}\right)} = \frac{2}{\frac{1}{v_1}+\frac{1}{v_2}} = \frac{2}{\frac{1}{120}+\frac{1}{80}} = 96\ km/h \tag{4.42}$$

Če bi uporabili aritmetično sredino, bi dobili rezultat 100 km/h, kar pa ni pravilno (glej [Brvar]).

V Matlabu bi rezultat (4.42) izračunali na naslednji način:

```
v=[120 80];
hs = harmmean(v)
hs =
   96.0000
```

4.3.5 Modus

Modus ali modalna vrednost je tista vrednost statistične spremenljivke v množici podatkov, ki se največkrat ponovi. Pri diskretni frekvenčni porazdelitvi je modus enak tisti vrednosti, ki ima največjo frekvenco. Pri zveznih frekvenčnih porazdelitvah z enako širino razredov Δx pa je modus tista vrednost, ki pripada razredu z največjo frekvenco (modalni razred) [Jesenko]. Predpostavimo, da je to *k-ti* razred $\left[r_{k-1}, r_k\right]$ s širino Δx in frekvenco f_k. Potem je modus enak [Jesenko]:

$$Mo = r_{k-1} + \frac{(\Delta x)\cdot(f_k - f_{k-1})}{(f_k - f_{k-1}) + (f_k - f_{k+1})} \tag{4.43}$$

Primer 4.19.:

Na vzorcu 200 krav molznic so proučevali količino namolženega mleka na dan. Podatki v litrih so podani na sliki 80 [Jesenko]. Izračunajte modus.

Količina	Frekvenca
5-7	2
7-9	8
9-11	14
11-13	22
13-15	28
15-17	38
17-19	44
19-21	24
21-23	14
23-25	6

Slika 80: Razredna frekvenčna porazdelitev količine namolženega mleka krav molznic na dan [Jesenko]

Iz slike 80 vidimo, da je modalni razred 17-19, saj ima največjo frekvenco, enako 44. Modus je enak:

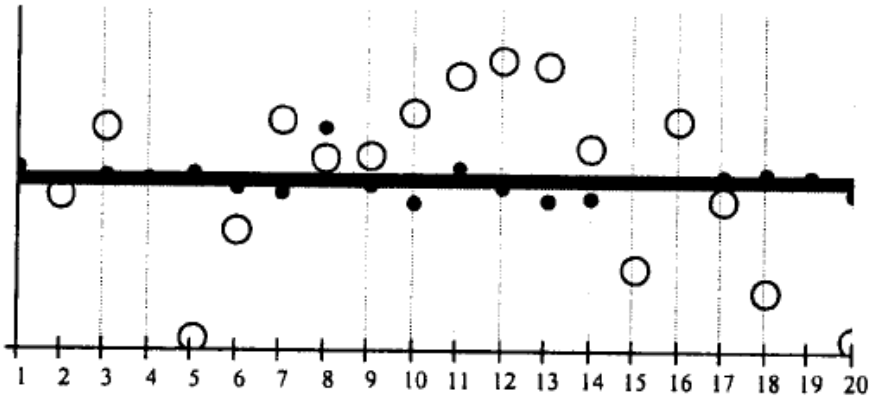
$$Mo = 17 + \frac{(2)\cdot(44-38)}{(44-38)+(44-24)} = 17.46 \tag{4.44}$$

Dobljeni rezultat pomeni, da pri največjem številu krav iz opazovane populacije v povprečju namolzejo 17.46 litra mleka na dan.

4.4 Variabilnost, asimetrija in sploščenost

Mere srednje vrednosti, ki smo jih vpeljali, podajajo le informacije o srednji vrednosti. Ničeasr pa ne povedo o tem, koliko so posamezne vrednosti spremenljivke oddaljene od srednje vrednosti [Jesenko].

Na sliki 81 po prikazane vrednosti statistične sprmenljivke. V smeri abscisne osi imamo zaporedno številko posamezne vrednosti, v smeri oordinatne osi pa vrednost spremenljivke. Vzporednica z abscisno osjo prikazuje aritmetično sredino [Jesenko].



Slika 81: Razpršenost podatkov okoli aritmetične sredine [Jesenko]

Kot je razvidno iz slike 81, sta prikazani dve množici podatkov (ena z velikimi krogci, druga z črnimi pikami), ki imata isto aritmetično sredino, a povsem drugačno porazdelitev okoli te sredine.

Mere, s katerimi merimo velikosti odmikov posameznih vrednosti od srednjih vrednosti, se imenujejo mere **variabilnosti** ali **razpršenosti** podatkov [Jesenko].

4.4.1 Variabilnost

Kakovost informacij, ki jih dobimo z izračunom mer variabilnosti, je različna. Na splošno pa velja, da je informacija tem boljša, čimveč podatkov smo uporabili pri njenem izračunu. Na splošno poznamo **absolutne in relativne mere variabilnosti**. Med absolutne mere spadajo variacijski razmik, kvartilni razmik in decilni razmik, povprečni absolutni odklon od aritmetične sredine in mediane ter **praktično in teoretično najpomembnejša mera variabilnosti - varianca** [Artenjak].

Variacijski razmik R_V je interval, v katerem leže vse vrednosti spremenljivke. Določen je na naslednji način [Jesenko]:

$$R_V = x_{\max} - x_{\min} \tag{4.45}$$

Primer 4.20.:

V 13 krajih so zabeležili temperaturo, kar prikazuje slika 82. Izračunajte variacijski razmik [Artenjak].

R_i:	1	2	3	4	5	6	7	8	9	10	11	12	13
y_i:	-5	-3	-1	-1	0	0	2	4	6	7	7	9	10

Slika 82: Zabeležena temperatura v 13 krajih [Artenjak]

Variacijski razmik je enak:

$$R_V = x_{\max} - x_{\min} = 10 - (-5) = 15 \text{ stopinj } C \tag{4.46}$$

Variacijski razmik ima to slabo lastnost, da je odvisen le od ekstremnih vrednosti spremenljivke. Lahko pa se zgodi, da najmanjša in največja vrednost zelo odskočita od ostalih vrednosti, kar povzroči velik variacijski razmik, kljub temu, da se vse ostale vrednosti zelo malo razlikujejo in je variabilnost podatkov v resnici majhna. To pomanjkljivost odpravi **kvartilni razmik**. Definiramo ga kot razliko med prvim in tretjim kvartilom [Jesenko]:

$$R_Q = Q_3 - Q_1 \tag{4.47}$$

Ta mera izloči ekstremne vrednosti in je zato boljša od variacijskega razmika, ima pa to pomanjkljivost, da upošteva le polovico vseh opazovanih vrednosti [Jesenko].

Podobno bi lahko vpeljali tudi **decilni razmik**, ki predstavlja razliko med devetim in prvim decilom [Artenjak, Bajt]:

$$R_D = D_9 - D_1 \tag{4.48}$$

Decilni razmik je uporabljen na 80% vrednosti spremenljivke, odrezanih pa je 10% vrednosti pod prvim decilom in 10% vrednosti nad devetim decilom.

V primeru temperature 13 krajev (glej sliko 82) imamo opravka s **problemom določitve kvantila iz ranžirne vrste**. Za ta namen obstaja poseben postopek, ki se glasi [Artenjak]:

1. Denimo je kvantilni rang P_i znan.

2. Rang R_i izračunamo na osnovi izraza:

$$R_i = N \cdot P_i + 0.5 \tag{4.49}$$

3. Za neenačbo $R_0 \le R_i \le R_l$, kamor pade R_i, odčitamo R_0, R_l, nato pa priredimo in odčitamo vrednosti y_0, y_l istoležnih členov, pri čemer velja neenačba $y_0 \le y_i \le y_l$.

4. Vrednost kvantila y_i izračunamo z izrazom:

$$y_i = y_0 + \frac{R_i - R_0}{R_l - R_0}\left(y_l - y_0\right) \tag{4.50}$$

Najprej izračunajmo kvartilni razmik za primer temperature krajev. Pri izračunu Q_3 imamo:

$$\begin{aligned} &P_i = 0.75 \\ &R_i = 0.75 \cdot 13 + 0.5 = 10.25 \\ &R_0 = 10,\ R_l = 11 \\ &y_0 = 7,\ y_l = 7 \\ &Q_3 = y_i = y_0 + \frac{R_i - R_0}{R_l - R_0}\left(y_l - y_0\right) = 7 + \frac{10.25 - 10}{11 - 10}\left(7 - 7\right) = 7 \end{aligned} \tag{4.51}$$

Pri izračunu Q_1 imamo:

$$\begin{aligned} &P_i = 0.25 \\ &R_i = 0.25 \cdot 13 + 0.5 = 3.75 \\ &R_0 = 3,\ R_l = 4 \\ &y_0 = -1,\ y_l = -1 \\ &Q_1 = y_i = y_0 + \frac{R_i - R_0}{R_l - R_0}\left(y_l - y_0\right) = -1 + \frac{3.75-3}{4-3}\left(-1+1\right) = -1 \end{aligned} \tag{4.52}$$

Kvartilni razmik torej je:

$$R_Q = Q_3 - Q_1 = 7 - \left(-1\right) = 8\ \textit{stopinj}\ C \tag{4.53}$$

Nato izračunajmo decilni razmik za primer temperature krajev. Pri izračunu D_9 imamo:

$$\begin{aligned} &P_i = 0.90 \\ &R_i = 0.90 \cdot 13 + 0.5 = 12.2 \\ &R_0 = 12,\ R_l = 13 \\ &y_0 = 9,\ y_l = 10 \\ &D_9 = y_i = y_0 + \frac{R_i - R_0}{R_l - R_0}\left(y_l - y_0\right) = 9 + \frac{12.2-12}{13-12}\left(10-9\right) = 9.2 \end{aligned} \tag{4.54}$$

Pri izračunu D_1 imamo:

$$\begin{aligned} &P_i = 0.1 \\ &R_i = 0.1 \cdot 13 + 0.5 = 1.8 \\ &R_0 = 1,\ R_l = 2 \\ &y_0 = -5,\ y_l = -3 \\ &D_1 = y_i = y_0 + \frac{R_i - R_0}{R_l - R_0}\left(y_l - y_0\right) = -5 + \frac{1.8-1}{2-1}\left(-3+5\right) = -5 + 1.6 = -3.4 \end{aligned} \tag{4.55}$$

Decilni razmik torej je:

$$R_D = D_9 - D_1 = 9.2 - (-3.4) = 12.6 \ \textit{stopinj C} \tag{4.56}$$

Sklep kvartilnega razmika: V 50% (ker gledamo le 50% podatkov) srednje toplih krajev je bila največja temperaturna razlika v povprečju 8 stopinj C [Artenjak].

Sklep decilnega razmika: V 80% (ker gledamo 80% podatkov) srednje toplih krajev pa je bila največja temperaturna razlika v povprečju že 12.6 stopinj C [Artenjak].

Povprečni absolutni odklon

Za mero variabilnosti lahko vzamemo tudi veličino, ki je določena kot povprečni absolutni odmik vrednosti spremenljivke od srednje vrednosti (aritmetične srednie ali mediane).

Povprečni absolutni odmik vrednosti spremenljivke od aritmetične sredine je enak [Jesenko]:

$$d = \frac{1}{N}\sum_{i=1}^{N}|y_i - \bar{y}| \tag{4.57}$$

Povprečni absolutni odmik vrednosti spremenljivke od mediane je enak [Jesenko]:

$$d = \frac{1}{N}\sum_{i=1}^{N}|y_i - Me| \tag{4.58}$$

Primer 4.21.:

V 13 krajih so zabeležili temperaturo, kar prikazuje slika 83 [Artenjak]. Izračunajte povprečni absolutni odklon od aritmetične sredine in od mediane.

R_i:	1	2	3	4	5	6	7	8	9	10	11	12	13
y_i:	-5	-3	-1	-1	0	0	2	4	6	7	7	9	10

Slika 83: Zabeležena temperatura v 13 krajih [Artenjak]

Arimetična sredina je enaka:

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i = \frac{1}{13}(-5-3-1-1+0+0+2+4+6+7+7+9+10) = \frac{35}{13} = 2.6923 \tag{4.59}$$

Povprečni absolutni odmik vrednosti spremenljivke od aritmetične sredine je enak:

$$d = \frac{1}{13}\sum_{i=1}^{13}|y_i - 2.7| = \frac{1}{13}(|-5-2.7|+|-3-2.7|+\ldots+|10-2.7|) = 4.1302 \tag{4.60}$$

Za izračun lahko uporabimo naslednji program v Matlabu:

```
% abs_odkl.m

clear
clc
close all

y=input('Vnesi podatke y')
y = sort(y);  % zaradi mediane

N = length(y)

ch = input('Izračun glede na aritmetično sredino (1) ali mediano (2)');

d = 0;

if ch ==1
  disp('aritmetična sredina je:')
  ysr = mean(y)
else
  disp('mediana je:')
  if abs(N/2-round(N/2))>0   % lih
     ysr = y((N+1)/2)
  else                       % sod
     ysr = (y(N/2)+y((N+2)/2))/2
  end
end
```

```
for i = 1:N
    d = d + abs(y(i)-ysr);
end

disp('absolutni odklon je:')
d/N
```

Izgled komandnega okna je naslednji:

```
Vnesi podatke y[-5 -3 -1 -1 0 0 2 4 6 7 7 9 10]

y =

   -5   -3   -1   -1    0    0    2    4    6    7    7    9   10

N =

   13

Izračun glede na aritmetično sredino (1) ali mediano (2)1

aritmetična sredina je:

ysr =

   2.6923

absolutni odklon je:

ans =

   4.1302
```

Povprečni absolutni odmik vrednosti spremenljivke od mediane je enak:

$$Me = \begin{cases} x_{\frac{N+1}{2}}, & N\ lih \\ \dfrac{x_{\frac{N+2}{2}} + x_{\frac{N}{2}}}{2}, & N\ sod \end{cases} = x_{\frac{N+1}{2}} = x_{\frac{13+1}{2}} = x_7 = 2 \tag{4.61}$$

Povprečni absolutni odmik vrednosti spremenljivke od mediane je enak:

$$d = \frac{1}{13}\sum_{i=1}^{13}|y_i - 2| = \frac{1}{13}\left(|-5-2| + |-3-2| + \ldots + |10-2|\right) = 4.0769 \tag{4.62}$$

Izgled komandnega okna v Matlabu je tokrat naslednji:

```
Vnesi podatke y[-5 -3 -1 -1 0 0 2 4 6 7 7 9 10]
y =
    -5    -3    -1    -1     0     0     2     4     6     7     7     9    10
N =
    13
Izračun glede na aritmetično sredino (1) ali mediano (2)2
mediana je:
ysr =
     2
absolutni odklon je:
ans =
    4.0769
```

Kadar so podatki podani v obliki diskretne frekvenčne porazdelitve, povprečni absolutni odmik od aritmetične sredine oz. mediane izračunamo na naslednji način [Jesenko, Artenjak]:

$$d = \frac{1}{N}\sum_{i=1}^{n}\left|y_i - \bar{y}\right| \cdot f_i \tag{4.63}$$

in:

$$d = \frac{1}{N}\sum_{i=1}^{n}\left|y_i - Me\right| \cdot f_i \tag{4.64}$$

pri čemer aritmetično sredino izračunamo z izrazom (4.30).

Kadar so podatki podani v obliki zvezne frekvenčne porazdelitve, povprečni absolutni odmik od aritmetične sredine oz. mediane izračunamo na naslednji način [Jesenko, Artenjak]:

$$d = \frac{1}{N}\sum_{i=1}^{n}\left|y_{si} - \bar{y}\right| \cdot f_i \tag{4.65}$$

in:

$$d = \frac{1}{N}\sum_{i=1}^{n} |y_{si} - Me| \cdot f_i \tag{4.66}$$

Primer 4.22.:

Tabela na sliki 84 prikazuje proizvodne čase (v minutah) določene vrste izdelka za izbrani vzorec. Izračunajte povprečni absolutni odklon od mediane [Jesenko].

Čas izdelave	Frekvenca
22-26	22
26-30	28
30-34	36
34-38	41
38-42	45
42-46	38
46-50	32
50-54	19
54-58	12
58-62	4

Slika 84: Proizvodni časi (v minutah) določene vrste izdelka za izbrani vzorec [Jesenko]

Na osnovi izraza (4.35) najprej izračunamo mediano. Kot se izkaže, je $\frac{N}{2} = \frac{\sum frekvenc}{2} = \frac{22+28+36+41+...}{2} = \frac{277}{2} = 138.5$. Vsota frekvenc do vključno razreda 34-38 je 127, do vključno razreda 38-42 pa 172. Zato hitro ugotovimo, da je medianin razred 38-42. Mediana torej je:

$$Me = x_{spk} + \frac{(\Delta\%) \cdot \left(\frac{N}{2} - F_{k-1}\right)}{f_k} = 38 + \frac{(4)\cdot(138.5-127)}{45} = 39.02 \tag{4.67}$$

Povprečni absolutni odmik od mediane je enak:

$$d = \frac{1}{N}\sum_{i=1}^{n}|y_{si} - Me|\cdot f_i = \frac{1}{277}\sum_{i=1}^{10}|y_{si} - 39.02|\cdot f_i =$$
$$= \frac{1}{277}\left(|y_{s1} - 39.02|\cdot f_1 + \ldots + |y_{s10} - 39.02|\cdot f_{10}\right) = \qquad (4.68)$$
$$= \frac{1}{277}\left(|24 - 39.02|\cdot 22 + \ldots + |60 - 39.02|\cdot 4\right) = 7.4749$$

Torej je povprečni absolutni odklon od mediane proizvodnih časov izdelka enak 7.4749 minut.

Za izračun rezultata (4.68) lahko uporabimo naslednji program v Matlabu:

```
% abs_odkl1.m

clear
clc
close all

s = input('Vnesi vektor sredin razredov')
f = input('Vnesi frekvence')
med = input('Vnesi mediano')

n = length(f)
N = sum(f)

d = 0;

for i=1:n
    d = d + f(i)*abs(s(i)-med);
end

disp('absolutni odklon je:')
d/N
```

Izpis komandnega okna pa je naslednji:

```
Vnesi vektor sredin razredov[24 28 32 36 40 44 48 52 56 60]
s =
   24   28   32   36   40   44   48   52   56   60
Vnesi frekvence[22 28 36 41 45 38 32 19 12 4]
f =
   22   28   36   41   45   38   32   19   12    4
Vnesi mediano39.02
med =
   39.0200
n =
    10
N =
   277
```

```
absolutni odklon je:
ans =
    7.4749
```

Varianca iz podatkov

Varianca je napogosteje uporabljena mera, s katero merimo variabilnost podatkov. Definirana je na naslednji način [Jesenko, Artenjak]:

$$\sigma^2_{pristr} = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2 =$$
$$= \frac{1}{N}\sum_{i=1}^{N}\left(x_i^{\,2} - 2x_i \cdot \bar{x} + \bar{x}^2\right) = \frac{1}{N}\left[\sum_{i=1}^{N}\left(x_i^{\,2}\right) - 2\cdot\bar{x}\sum_{i=1}^{N}\left(x_i\right) + \bar{x}^2 \cdot N\right] = \tag{4.69}$$
$$= \frac{1}{N}\sum_{i=1}^{N}\left(x_i^{\,2}\right) - \bar{x}^2$$

Kot se izkaže, je ocena variance v izrazu (4.69) pristranska. Nepristranska vrednost pa je:

$$\sigma^2 = \sigma^2_{nepristr} = \frac{N}{N-1}\sigma^2_{pristr} = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2 \tag{4.70}$$

Primer 4.23.:

Tabela na sliki 85 prikažuje težo učencev 8. razreda osnovne šole. Določite varianco [Jesenko]!

i	x_i kg	i	x_i kg	i	x_i kg	i	x_i kg
1	42,3	11	45,7	21	51,4	31	55,4
2	42,6	12	45,7	22	51,5	32	55,9
3	43,1	13	45,9	23	51,8	33	56,1
4	43,5	14	50,1	24	52,4	34	56,5
5	43,8	15	50,4	25	52,6	35	56,8
6	44,2	16	50,5	26	52,9	36	58,3
7	44,4	17	50,7	27	53,2	37	59,7
8	44,9	18	50,8	28	53,9	38	62,4
9	45,0	19	51,0	29	54,2	39	66,5
10	45,3	20	51,3	30	54,8	40	71,2

Slika 85: Teža učencev 8. razreda osnovne šole [Jesenko]

Najprej izračunajmo aritmetično sredino:

$$\bar{x} = \frac{1}{40}\sum_{i=1}^{40}(x_i) = \frac{1}{40}(42.3 + ... + 71.2) = 51.4675 \tag{4.71}$$

Če nato izračunamo pristransko varianco, dobimo:

$$\sigma^2_{pristr} = \frac{1}{N}\sum_{i=1}^{N}(x_i^2) - \bar{x}^2 = \frac{1}{40}\sum_{i=1}^{40}(x_i^2) - 51.4675^2 = \tag{4.72}$$

$$= \frac{1}{40}(42.3^2 + ... + 71.2^2) - 51.4675^2 = 41.8737$$

Program **var_pod.m** v Matlabu za ta izračun je bil naslednji:

```
% var_pod.m
%
% program izračuna pristransko varianco.
%
% mozen je primer iz jesenko knjige, ali poljuben primer
%
% pristranska varianca se izracuna iz razvite koncne predelane formule.
%

clear
clc
close all

ch = input('podatki iz jesenko primer 4.9.2 1-da,0-ne');
if ch == 1
   y = [42.3 42.6 43.1 43.5 43.8 44.2 44.4 44.9 45.0 45.3 45.7 45.7 45.9...
        50.1 50.4 50.5 50.7 50.8 51.0 51.3 51.4 51.5 51.8 52.4 52.6 52.9...
        53.2 53.9 54.2 54.8 55.4 55.9 56.1 56.5 56.8 58.3 59.7 62.4 66.5...
        71.2]
else
   y=input('Vnesi podatke y')
end
```

```
N = length(y)

d = 0;
for i=1:length(y)
    d = d + y(i);
end

ysr = d/N

m2 = 0;

for i=1:N
    m2 = m2 + y(i)^2;
end

m2 = m2/N

var = m2 - ysr^2
```

Izgled komandnega okna je naslednji:

```
podatki iz jesenko primer 4.9.2 1-da,0-ne1
y =
  Columns 1 through 8
   42.3000   42.6000   43.1000   43.5000   43.8000   44.2000   44.4000   44.9000
  Columns 9 through 16
   45.0000   45.3000   45.7000   45.7000   45.9000   50.1000   50.4000   50.5000
  Columns 17 through 24
   50.7000   50.8000   51.0000   51.3000   51.4000   51.5000   51.8000   52.4000
  Columns 25 through 32
   52.6000   52.9000   53.2000   53.9000   54.2000   54.8000   55.4000   55.9000
  Columns 33 through 40
   56.1000   56.5000   56.8000   58.3000   59.7000   62.4000   66.5000   71.2000


N =
    40
ysr =
   51.4675
m2 =
  2.6908e+003
var =
   41.8737
```

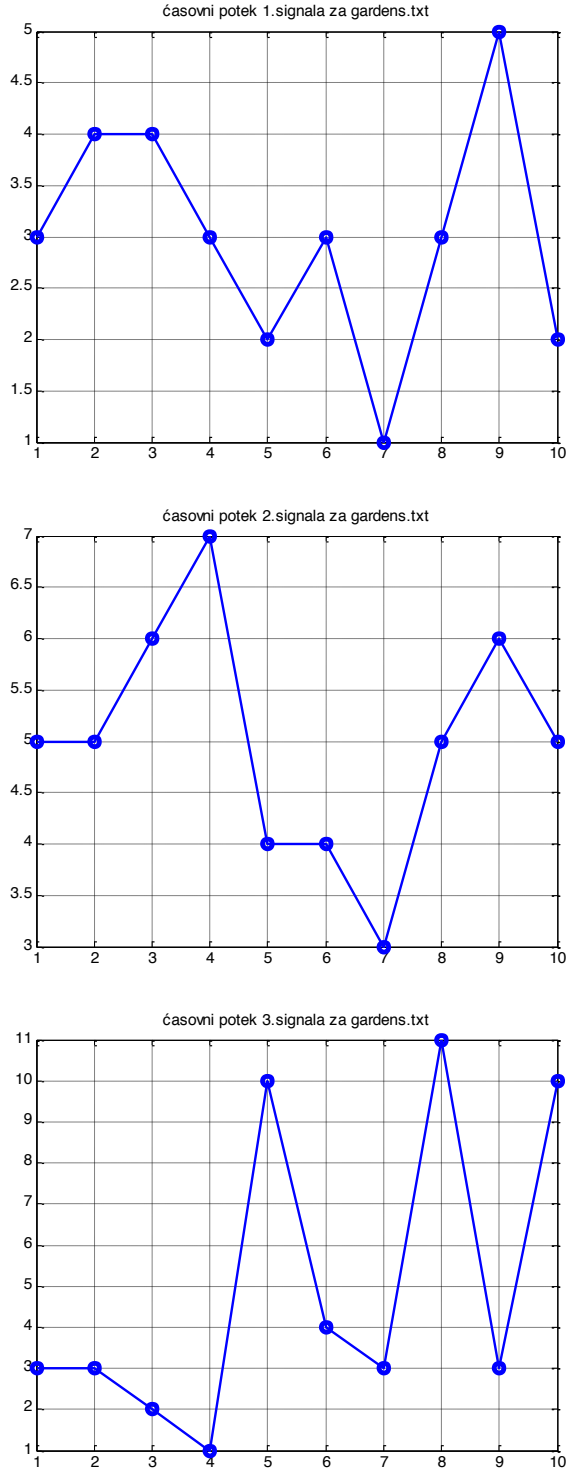
Nepristransko varianco dobimo na naslednji način:

$$\sigma^2 = \sigma^2_{nepristr} = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2 = \frac{N}{N-1}\sigma^2_{pristr} = \tag{4.73}$$

$$= \frac{40}{39}\cdot 41.8737 = 42.9473$$

Primer 4.24.:

*Dane imamo tri časovne vrste, shranjene v datoteki **gardens.txt**, ki jih prikazuje slika 86. Zanje izračunajte variance [Žibert]!*



*Slika 86: Tri časovne vrste, shranjene v datoteki **gardens.txt** [Žibert]*

Za izračun varianc smo uporabili program **var_pod1.m**, ki ima naslednjo obliko:

```
% var_pod1.m
%
% program izračuna pristransko in nepristransko varianco.
%
% mozen je primer iz jesenko knjige, ali zibert knjige, ali poljuben primer
%
% pristranska varianca se izracuna iz zacetne formule.
% nepristranska varianca se izracuna s korekturo pristranske, ali pa s standardnim
% matlab ukazom.

clear
clc
close all

ch = input('podatki iz jesenko primer 4.9.2 (1), iz zibert (2), drugi (3)');
if ch == 1
   y = [42.3 42.6 43.1 43.5 43.8 44.2 44.4 44.9 45.0 45.3 45.7 45.7 45.9...
        50.1 50.4 50.5 50.7 50.8 51.0 51.3 51.4 51.5 51.8 52.4 52.6 52.9...
        53.2 53.9 54.2 54.8 55.4 55.9 56.1 56.5 56.8 58.3 59.7 62.4 66.5...
        71.2]
elseif ch == 2
   garden=importdata('gardens.txt');
   ch = input('izracun var1(1) var2(2) var3(3)')
   y=garden.data;
   y = y(:,ch);
   plot(y,'LineWidth',1.5)
   hold on
   plot(y,'o','LineWidth',3)
   title(['časovni potek ' num2str(ch) '.signala za gardens.txt'])
   grid
else
   y=input('Vnesi podatke y')
end

N = length(y)

d = 0;
for i=1:length(y)
    d = d + y(i);
end

ysr = d/N

var = 0;

for i=1:N
    var = var + (y(i)-ysr)^2;
end

disp('Pristranska varianca in standard. deviacija:')

var = var/N
sqrt(var)

disp('Nepristranska varianca in deviacija (1. način):')

var1 = var * (N/(N-1))
sqrt(var1)

disp('Nepristranska varianca in deviacija (2. način):')

VAR(y)
STD(y)
```

Izgled komandnega okna pri izračunu variance prve časovne vrste je naslednji:

```
podatki iz jesenko primer 4.9.2 (1), iz zibert (2), drugi (3)2
izracun var1(1) var2(2) var3(3)1
ch =
     1
N =
    10
ysr =
     3
Pristranska varianca in standard. deviacija:
var =
    1.2000
ans =
    1.0954


Nepristranska varianca in deviacija (1. način):
var1 =
    1.3333
ans =
    1.1547
Nepristranska varianca in deviacija (2. način):
ans =
    1.3333
ans =
    1.1547
```

Izgled komandnega okna pri izračunu variance druge časovne vrste je naslednji:

```
podatki iz jesenko primer 4.9.2 (1), iz zibert (2), drugi (3)2
izracun var1(1) var2(2) var3(3)2
ch =
     2
N =
    10
ysr =
     5
Pristranska varianca in standard. deviacija:
var =
    1.2000
ans =
    1.0954
```

```
Nepristranska varianca in deviacija (1. način):
var1 =
    1.3333
ans =
    1.1547
Nepristranska varianca in deviacija (2. način):
ans =
    1.3333
ans =
    1.1547
```

Izgled komandnega okna pri izračunu variance tretje časovne vrste je naslednji:

```
podatki iz jesenko primer 4.9.2 (1), iz zibert (2), drugi (3)2
izracun var1(1) var2(2) var3(3)3
ch =
     3
N =
    10
ysr =
     5
Pristranska varianca in standard. deviacija:
var =
   12.8000
ans =
    3.5777

Nepristranska varianca in deviacija (1. način):
var1 =
   14.2222
ans =
    3.7712
Nepristranska varianca in deviacija (2. način):
ans =
   14.2222
ans =
    3.7712
```

Pri tem primeru so bili podatki naslednji:

```
>> garden=importdata('gardens.txt')
garden =
    data: [10x3 double]
    textdata: {'gardenA'  'gardenB'  'gardenC'}
    colheaders: {'gardenA'  'gardenB'  'gardenC'}
>> header = garden.textdata
header =
  'gardenA'    'gardenB'    'gardenC'
>> data = garden.data
data =
    3    5    3
    4    5    3
    4    6    2
    3    7    1
    2    4   10
    3    4    4
    1    3    3
    3    5   11
    5    6    3
    2    5   10
```

Dobili smo torej naslednje rezultate za dotične časovne vrste:

1. časovna vrsta: $\bar{x} = 3, \quad \sigma^2 = 1.3333$
2. časovna vrsta: $\bar{x} = 5, \quad \sigma^2 = 1.3333$
3. časovna vrsta: $\bar{x} = 5, \quad \sigma^2 = 14.2222$

Torej sta varianci 1. in 2. časovne vrste enaki. Prav tako sta enaki srednji vrednosti 2. in 3. časovne vrste. Tu se lahko poraja vrsta zanimivih vprašanj, kot npr., kakšna je narava vzorcev z enakimi povprečji oz. variancami, na katera lahko odgovorimo z nadaljnjimi statističnimi testi. Vsekakor je varianca npr. zelo primerna za merjenje zaupanja v ocene parametrov in testiranje hipotez [Žibert].

Varianca pri diskretni frekvenčni porazdelitvi

Definirana je na naslednji način [Jesenko, Artenjak]:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n} f_i \cdot (x_i - \bar{x})^2 =$$

$$= \frac{1}{N}\sum_{i=1}^{n} f_i \cdot (x_i^2 - 2x_i \cdot \bar{x} + \bar{x}^2) = \frac{1}{N}\left[\sum_{i=1}^{n} f_i \cdot (x_i^2) - 2 \cdot \bar{x}\underbrace{\sum_{i=1}^{n}(f_i \cdot x_i)}_{\bar{x} \cdot N} + \bar{x}^2 \cdot \underbrace{\sum_{i=1}^{n}(f_i)}_{N}\right] = \tag{4.74}$$

$$= \frac{1}{N}\sum_{i=1}^{n} f_i \cdot (x_i^2) - \bar{x}^2$$

Varianca pri zvezni frekvenčni porazdelitvi

Definirana je na naslednji način [Jesenko, Artenjak]:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n} f_i \cdot (x_{si} - \bar{x})^2 =$$

$$= \frac{1}{N}\sum_{i=1}^{n} f_i \cdot (x_{si}^2 - 2x_{si} \cdot \bar{x} + \bar{x}^2) = \frac{1}{N}\left[\sum_{i=1}^{n} f_i \cdot (x_{si}^2) - 2 \cdot \bar{x}\underbrace{\sum_{i=1}^{n}(f_i \cdot x_{si})}_{\bar{x} \cdot N} + \bar{x}^2 \cdot \underbrace{\sum_{i=1}^{n}(f_i)}_{N}\right] = \tag{4.75}$$

$$= \frac{1}{N}\sum_{i=1}^{n} f_i \cdot (x_{si}^2) - \bar{x}^2$$

V primeru, ko so podatki podani v obliki zvezne frekvenčne porazdelitve, pride do določene napake zaradi računanja s sredinami razredov, ki se povečuje z večanjem širine razreda. To napako pa lahko zmanjšamo s takoimenovanim Sheppardovim popravkom [Jesenko]:

$$\hat{\sigma}^2 = \sigma^2 - \frac{(\Delta x)^2}{12} = \frac{1}{N}\sum_{i=1}^{n} f_i \cdot (x_{si} - \bar{x})^2 - \frac{(\Delta x)^2}{12}$$
$$= \frac{1}{N}\sum_{i=1}^{n} f_i \cdot (x_{si}{}^2) - \bar{x}^2 - \frac{(\Delta x)^2}{12} \qquad (4.76)$$

Standardna deviacija in variacijski koeficient

Standardna deviacija je definirana kot koren variance: $\sigma = \sqrt{\sigma^2}$, variacijski koeficient pa kot: $KV = \frac{\sigma}{\bar{x}}$. Deviacijo je zaželjeno vpeljati, saj je varianca izražena v kvadratnih merskih enotah, o čemer nimamo najboljše predstave. Variacijski koeficient pa vpeljemo tedaj, ko bi želeli iz mere za variabilnost izločiti vpliv merske enote [Jesenko].

Primer 4.25.:

Dane imamo podatke (proizvodne čase nekega izdelka v minutah) v tabeli na sliki 87. Izračunajte varianco, deviacijo in variacijski koeficient [Jesenko]!

Razred	Frekvenca f_i	Sredina razredov x_{si}	$f_i.x_{si}$	$f_i.(x_{si})^2$
22-26	22	24	528	12672
26-30	28	28	784	21952
30-34	36	32	1152	36864
34-38	41	36	1476	53136
38-42	45	40	1800	72000
42-46	38	44	1672	73568
46-50	32	48	1536	73728
50-54	19	52	988	51376
54-58	12	56	672	37632
58-62	4	60	240	14400
vsota	**N = 277**		**10848**	**447328**

Slika 87: Podatki za proizvodne čase nekega izdelka v minutah v obliki zvezne frekvenčne porazdelitve

Najprej izračunamo aritmetično sredino:

$$\begin{aligned}\bar{x} &= \frac{1}{N}\sum_{i=1}^{n} f_i \cdot (x_{si}) = \frac{1}{277}\sum_{i=1}^{10} f_i \cdot (x_{si}) = \frac{1}{277}(f_1 \cdot x_{s1} + ... + f_{10} \cdot x_{s10}) = \\ &= \frac{1}{277}(22 \cdot 24 + ... + 4 \cdot 60) = 39.1625\end{aligned} \tag{4.77}$$

Varianca v prvem približku je enaka:

$$\begin{aligned}\sigma'^2 &= \frac{1}{277}\sum_{i=1}^{10} f_i \cdot (x_{si}{}^2) - 39.1625^2 = \\ &= \frac{1}{277} \cdot (22 \cdot 24^2 + ... + 4 \cdot 60^2) - 39.1625^2 = \\ &= \frac{1}{277} \cdot 447328 - 39.1625^2 = 81.2047\end{aligned} \tag{4.78}$$

Korigirana varianca s Shepperdovim popravkom je enaka:

$$\hat{\sigma}^2 = \sigma'^2 - \frac{(\Delta x)^2}{12} = 81.2047 - \frac{(4)^2}{12} = 79.8713 \tag{4.79}$$

Varianca proizvodnih časov je torej 81.2047 $\min^2$, korigirana varianca pa 79.8713 $\min^2$.

Deviacija je enaka:

$$\hat{\sigma} = \sqrt{79.8713} = 8.9371 \tag{4.80}$$

Variacijski koeficient pa je enak:

$$KV = \frac{\hat{\sigma}}{\bar{x}} = \frac{8.9371}{39.1625} = 0.2282 \tag{4.81}$$

kar pomeni 22.82%.

Za izračune je bil uporabljen naslednji program v Matlabu:

```
% var_pod2.m
%
%

clear
clc
close all

ch = input('podatki iz jesenko primer 4.9.6 (1), iz artenjak knjige (2), drugo(3)');
if ch == 1
   f = [22 28 36 41 45 38 32 19 12 4]
   s = [24 28 32 36 40 44 48 52 56 60]
elseif ch == 2
   f = [11 51 86 64 33 5]
   s = [2.5 7.5 12.5 17.5 22.5 27.5]
else
   f=input('Vnesi vektor frekvenc f')
   s=input('Vnesi vektor sredin razredov s')
end

dx = s(2) - s(1)
N = sum(f)
n = length(f)

d = 0;
for i=1:n
    d = d + f(i)*s(i);
end

xsr = d/N

m2 = 0;

for i=1:n
    m2 = m2 + f(i)*s(i)^2;
end

m2 = m2/N

disp('Prvi približek variance:')

var = m2 - xsr^2

disp('Korigirana varianca s Shepperdom:')

vark = var - dx^2/12

disp('Standardna deviacija je:')

stdk = sqrt(vark)

disp('Variacijski koeficient je:')

kv = stdk/xsr
```

Izgled komandnega okna je naslednji:

```
podatki iz jesenko primer 4.9.6 (1), iz artenjak knjige (2), drugo(3)1
f =
   22   28   36   41   45   38   32   19   12    4
s =
   24   28   32   36   40   44   48   52   56   60
```

```
dx =
     4
N =
   277
n =
    10
xsr =
   39.1625
m2 =
  1.6149e+003
Prvi približek variance:
var =
   81.2047
Korigirana varianca s Shepperdom:
vark =
   79.8713
Standardna deviacija je:
stdk =
    8.9371
Variacijski koeficient je:
kv =
    0.2282
```

Primer 4.26.:

Dane imamo podatke (odstotki pokvarjenega sadja v 250 zabojih) v tabeli na sliki 88. Izračunajte varianco, deviacijo in variacijski koeficient [Artenjak]!

Odstotek pokvarjenega sadja *(y)*	f_k	y_k	$f_k y_k$	$f_k y_k^2$
manj kot 5	11	2,5	27,5	68,75
od 5 do manj kot 10	51	7,5	382,5	2.868,75
od 10 do manj kot 15	86	12,5	1.075	13.437,50
od 15 do manj kot 20	64	17,5	1.120	19.600
od 20 do manj kot 25	33	22,5	742,5	16.706,25
od 25 do manj kot 30	5	27,5	137,5	3.781,25
Skupaj *(N)*	**250**		**3.485,0**	**56.462,50**

Slika 88: Podatki za odstotke pokvarjenega sadja v 250 zabojih v obliki zvezne frekvenčne porazdelitve [Artenjak]

Najprej izračunamo aritmetično sredino:

$$\begin{aligned}\bar{x} &= \frac{1}{N}\sum_{i=1}^{n} f_i \cdot (x_{si}) = \frac{1}{250}\sum_{i=1}^{6} f_i \cdot (x_{si}) = \frac{1}{250}\left(f_1 \cdot x_{s1} + \ldots + f_6 \cdot x_{s6}\right) = \\ &= \frac{1}{250}\left(11 \cdot 2.5 + \ldots + 5 \cdot 27.5\right) = 13.9400\end{aligned} \tag{4.82}$$

Varianca v prvem približku je enaka:

$$\begin{aligned}\vartheta\!\!\!/ &= \frac{1}{250}\sum_{i=1}^{6} f_i \cdot \left(x_{si}{}^2\right) - 13.94^2 = \\ &= \frac{1}{250} \cdot \left(11 \cdot 2.5^2 + \ldots + 5 \cdot 27.5^2\right) - 13.94^2 = \\ &= \frac{1}{250} \cdot 56463 - 13.94^2 = 31.5264\end{aligned} \tag{4.83}$$

Korigirana varianca s Shepperdovim popravkom je enaka:

$$\hat{\sigma}^2 = \vartheta\!\!\!/ - \frac{(\Delta x)^2}{12} = 31.5264 - \frac{(5)^2}{12} = 29.4431 \tag{4.84}$$

Deviacija je enaka:

$$\hat{\sigma} = \sqrt{29.4431} = 5.4261 \tag{4.85}$$

Variacijski koeficient pa je enak:

$$KV = \frac{\hat{\sigma}}{\bar{x}} = \frac{5.4261}{13.9400} = 0.3893 \tag{4.86}$$

kar pomeni 38.93% variacije pokvarjenega sadja v 250 zabojih sadja.

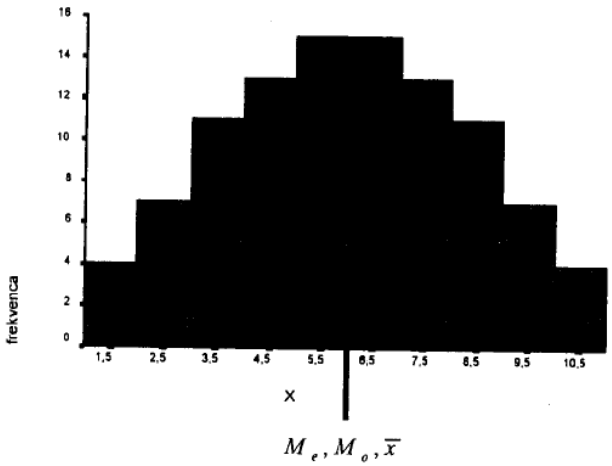
Tudi tokrat smo za izračune uporabili program v Matlabu **var_pod2.m**. Izpis komandnega okna je tokrat naslednji:

```
podatki iz jesenko primer 4.9.6  (1), iz artenjak knjige (2), drugo(3)2
f =
    11    51    86    64    33     5
s =
    2.5000    7.5000   12.5000   17.5000   22.5000   27.5000
dx =
     5
N =
   250
n =
     6
xsr =
   13.9400
m2 =
  225.8500
Prvi približek variance:
var =
   31.5264
Korigirana varianca s Shepperdom:
vark =
   29.4431
Standardna deviacija je:
stdk =
    5.4261
Variacijski koeficient je:
kv =
    0.3893
```

4.4.2 Asimetrija

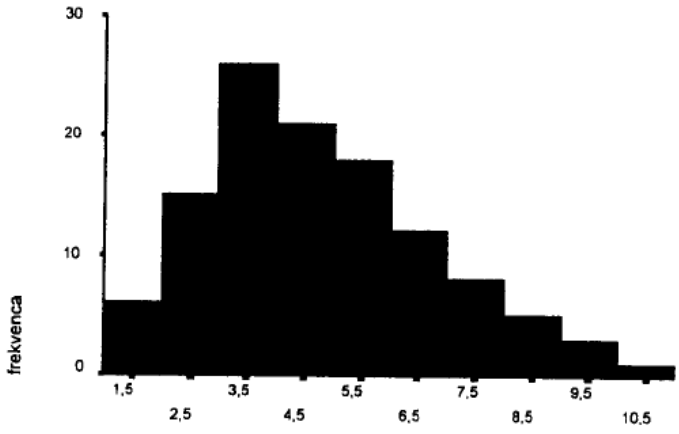
Za frekvenčne porazdelitve je značilno, da so lahko **simetrične, asimetrične v levo in asimetrične v desno**.

Množica podatkov z enim samim modusom je simetrična, kadar se medsebojno ujemajo aritmetična sredina, mediana in modus in velja: $\overline{x} = Me = Mo$ (glej sliko 89) [Jesenko].



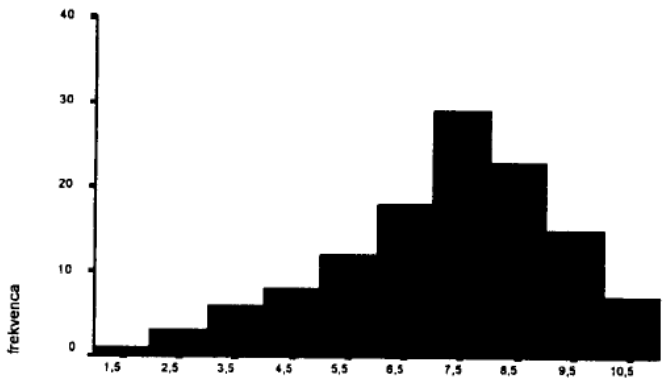
Slika 89: Simetrična frekvenčna porazdelitev [Jesenko]

Množica podatkov z enim samim modusom je asimetrična v desno, kadar velja: $\overline{x} > Me > Mo$ (glej sliko 90) [Jesenko].



Slika 90: Asimetrična frekvenčna porazdelitev v desno [Jesenko]

Množica podatkov z enim samim modusom je asimetrična v levo, kadar velja: $\overline{x} < Me < Mo$ (glej sliko 91) [Jesenko].



Slika 91: Asimetrična frekvenčna porazdelitev v levo [Jesenko]

Analitična kazalca za določanje smeri in moči asimetrije, ki dasta natančnejšo informacijo o asimetriji frekvenčne porazdelitve, sta **koeficient asimetrije na osnovi mediane in koeficient asimetrije na osnovi modusa** [Tominc].

Koeficient asimetrije na osnovi mediane je [Tominc, Artenjak]:

$$KA_{Me} = \frac{3(\bar{y} - Me)}{\sigma} \tag{4.87}$$

Koeficient asimetrije na osnovi modusa je [Tominc, Artenjak]:

$$KA_{Mo} = \frac{(\bar{y} - Mo)}{\sigma} \tag{4.88}$$

Če sta $KA_{Me}, KA_{Mo} < 0$, je porazdelitev asimetrična v levo, sicer pa je asimetrična v desno ($KA_{Me}, KA_{Mo} > 0$), ali simetrična ($KA_{Me}, KA_{Mo} = 0$).

Primer 4.27.:

V primeru proučevanja 250 zabojev po odstotku pokvarjenega sadja v zaboju (glej sliko 88) so bile izračunane naslednje vrednosti parametrov:

$$\begin{aligned} \bar{x} &= 13.9400 \\ \hat{\sigma} &= \sqrt{29.4431} = 5.4261 \end{aligned} \tag{4.89}$$

Izračunajte koeficient asimetrije na osnovi modusa in mediane!

Najprej moramo izračunati še modus in mediano. Za izračun mediane prvo ugotovimo potek kumulativne frekvence. To lahko naredimo z naslednjima ukazoma v Matlabu:

```
x=[11 51 86 64 33 5]
x =
   11   51   86   64   33    5
>> kf = cumsum(x)
kf =
   11   62  148  212  245  250
```

Ker je $\frac{N}{2}=125$, hitro ugotovimo, da se mediana nahaja v 3. razredu, to je od 10 do manj kot 15. Tako dobimo:

$$Me = x_{spk} + \frac{(\Delta\%_k)\cdot\left(\frac{N}{2}-F_{k-1}\right)}{f_k} = 10 + \frac{(5)\cdot(125-62)}{86} = 13.66 \tag{4.90}$$

Izračunajmo še modus. Modalni razred je 3. razred, to je od 10 do manj kot 15, saj tam leži največja frekvenca 86. Sledi:

$$Mo = r_{k-1} + \frac{(\Delta x)\cdot(f_k - f_{k-1})}{(f_k - f_{k-1})+(f_k - f_{k+1})} = \tag{4.91}$$
$$= 10 + \frac{(5)\cdot(86-51)}{(86-51)+(86-64)} = 13.07$$

Koeficient asimetrije na osnovi mediane je:

$$KA_{Me} = \frac{3(\bar{x}-Me)}{\hat{\sigma}} = \frac{3(13.94-13.66)}{5.4261} = 0.1548 \tag{4.92}$$

Koeficient asimetrije na osnovi modusa je:

$$KA_{Mo} = \frac{(\bar{x} - Mo)}{\hat{\sigma}} = \frac{(13.94 - 13.07)}{5.4261} = 0.1603 \tag{4.93}$$

Torej je porazdelitev rahlo asimetrična v desno. Oba koeficienta se sicer rahlo razlikujeta, kar je razumljivo, vendar sta po predznaku praviloma vselej skladna [Artenjak].

Za podatke $x_1, x_2, ..., x_N$, ki niso podani v obliki frekvenčne porazdelitve, se občasno vzame tudi naslednji koeficient asimetričnosti [Jesenko]:

$$g_1 = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^3}{\left(\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}\right)^3} \tag{4.94}$$

katerega podobnost z izrazom (2.181) je očitna. Če je $g_1 = 0$, je množica podatkov simetrična. Bolj, ko je g_1 negativen, bolj je množica podatkov asimetrična v levo, in obratno, bolj ko je pozitiven, bolj je množica podatkov asimetrična v desno.

Pri diskretni frekvenčni porazdelitvi je koeficient asimetričnosti enak [Jesenko]:

$$g_1 = \frac{\frac{1}{N}\sum_{i=1}^{n} f_i \cdot (x_i - \bar{x})^3}{\left(\sqrt{\frac{1}{N}\sum_{i=1}^{n} f_i \cdot (x_i - \bar{x})^2}\right)^3} \tag{4.95}$$

Pri zvezni frekvenčni porazdelitvi je koeficient asimetričnosti enak [Jesenko]:

$$g_1 = \frac{\frac{1}{N}\sum_{i=1}^{n} f_i \cdot (x_{si} - \bar{x})^3}{\left(\sqrt{\frac{1}{N}\sum_{i=1}^{n} f_i \cdot (x_{si} - \bar{x})^2}\right)^3} \tag{4.96}$$

Primer 4.28.:

Tabela na sliki 92 prikazuje proizvodne čase (v minutah) določene vrste izdelka za izbrani vzorec. Izračunajte koeficient asimetričnosti [Jesenko].

Čas izdelave	Frekvenca
22-26	22
26-30	28
30-34	36
34-38	41
38-42	45
42-46	38
46-50	32
50-54	19
54-58	12
58-62	4

Slika 92: Proizvodni časi (v minutah) določene vrste izdelka za izbrani vzorec [Jesenko]

Aritmetično sredino smo izračunali že prej (glej izraz (4.77)) in znaša: $\bar{x} = 39.1625 \approx 39.2$.

Slika 93 prikazuje delne izračune.

Razred	Frekvenca f_i	Sredina razreda x_{si}	$f_i.(x_{si}-\bar{x})^2$	$f_i.(x_{si}-\bar{x})^3$
22-26	22	24	5082,88	-77259,776
26-30	28	28	3512,32	-39337,984
30-34	36	32	1866,24	-13436,928
34-38	41	36	419,84	-1343,488
38-42	45	40	28,80	23,040
42-46	38	44	875,52	4202,496
46-50	32	48	2478,08	21807,104
50-54	19	52	3112,96	39845,888
54-58	12	56	3386,88	56899,584
58-62	4	60	1730,56	35995,648
vsota	**N = 277**		**22494,08**	**27395,581**

Slika 93: Delni izračuni za primer proizvodnih časov [Jesenko]

Če vstavimo te rezultate v izraz (4.96), dobimo:

$$g_1 = \frac{\frac{1}{277}27395.581}{\left(\sqrt{\frac{1}{277}22494.08}\right)^3} = \frac{98.9010}{731.7836} = 0.1352 \qquad (4.97)$$

Vidimo torej, da je obravnavana množica podatkov rahlo asimetrična v desno. Izračune smo opravili z naslednjim programom v Matlabu:

```
% asim_pod.m
%
%

clear
clc
close all

ch = input('podatki iz jesenko primer (1), drugo(2)');
if ch == 1
   f = [22 28 36 41 45 38 32 19 12 4]
   s = [24 28 32 36 40 44 48 52 56 60]
else
   f=input('Vnesi vektor frekvenc f')
   s=input('Vnesi vektor sredin razredov s')
end

N = sum(f)
n = length(f)

d = 0;
for i=1:n
    d = d + f(i)*s(i);
end

xsr = d/N

if ch == 1
   ch1 = input('zelis zaokroziti xsr 1-da,0-ne');
   if ch1 == 1
      xsr = 39.2
   end
end


for i=1:n
    p2(i) = f(i)*(s(i)-xsr)^2;
    p3(i) = f(i)*(s(i)-xsr)^3;
end

disp('delni izračuni:')
p2'
p3'

p2s = sum(p2)
p3s = sum(p3)

stevec = p3s/N
imenov = (p2s/N)^1.5
g1=stevec/imenov;

disp('koeficient asimetrije je:')
g1
```

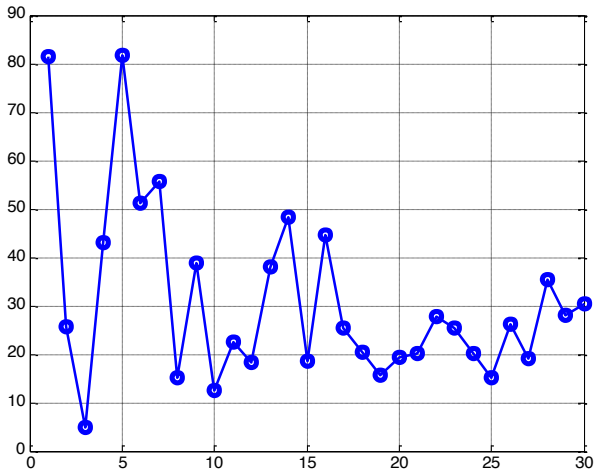
Izgled komandnega okna je naslednji:

```
podatki iz jesenko primer (1), drugo(2)1

f =

   22   28   36   41   45   38   32   19   12    4

s =

   24   28   32   36   40   44   48   52   56   60
```

```
N =

   277

n =

    10

xsr =

   39.1625

zelis zaokroziti xsr 1-da,0-ne1
xsr =

   39.2000

delni izračuni:

ans =

  1.0e+003 *

    5.0829
    3.5123
    1.8662
    0.4198
    0.0288
    0.8755
    2.4781
    3.1130
    3.3869
    1.7306

ans =

  1.0e+004 *

   -7.7260
   -3.9338
   -1.3437
   -0.1343
    0.0023
    0.4202
    2.1807
    3.9846
    5.6900
    3.5996

p2s =

  2.2494e+004

p3s =

  2.7396e+004

stevec =

   98.9010

imenov =

  731.7836

koeficient asimetrije je:

g1 =

    0.1352
```

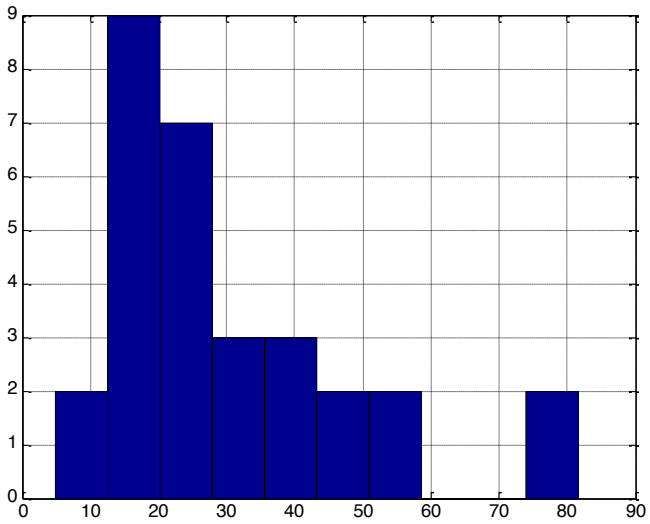
Primer 4.29.:

Dano imamo časovno vrsto iz datoteke **skewdata.txt**, ki jo prikazuje slika 94 [Žibert].



*Slika 94: Časovno vrsta iz datoteke **skewdata.txt** [Žibert]*

Njen histogram ima obliko, ki jo prikazuje slika 95.



*Slika 95: Histogram časovne vrste iz datoteke **skewdata.txt***

Kot se izkaže, koeficient asimetrije (4.94) zavzame naslednjo vrednost:

$$g_1 = \frac{\frac{1}{30}\sum_{i=1}^{30}(x_i - \bar{x})^3}{\left(\sqrt{\frac{1}{30}\sum_{i=1}^{30}(x_i - \bar{x})^2}\right)^3} = 1.3877 \tag{4.98}$$

Za izračun smo uporabili naslednji program v Matlabu:

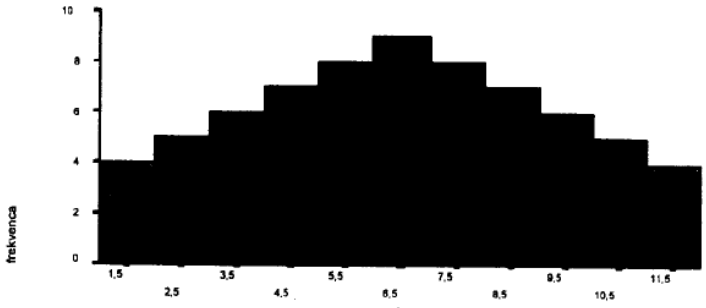
```
% asim_pod1.m

clear
clc
close all

data = importdata('skewdata.txt');
data = data.data;
x = data;

plot(x,'LineWidth',1.5)
hold on
plot(x,'o','LineWidth',3)
grid
figure
hist(x)
grid

n = length(x)

d = 0;
for i=1:n
    d = d + x(i);
end

xsr = d/n

for i=1:n
    p2(i) = (x(i)-xsr)^2;
    p3(i) = (x(i)-xsr)^3;
end

p2s = sum(p2);
p3s = sum(p3);

stevec = p3s/n;
imenov = (p2s/n)^1.5;
g1=stevec/imenov;

disp('koeficient asimetrije - 1. nacin:')
g1

disp('koeficient asimetrije - 2. nacin (s skewness.m):')
g1a = skewness(data)
```

Izpis komandnega okna je naslednji:

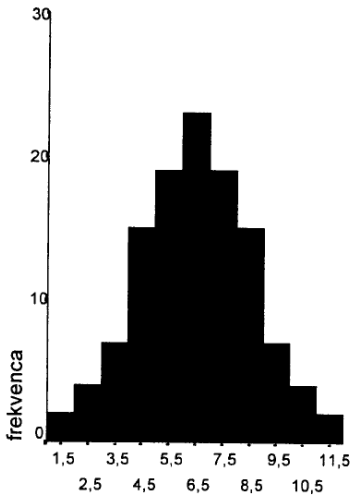
```
n =
    30
xsr =
   30.9687
koeficient asimetrije - 1. nacin:
g1 =
    1.3877
koeficient asimetrije - 2. nacin (s skewness.m):
g1a =
    1.3877
```

4.4.3 Sploščenost

Porazdelitev podatkov je lahko bolj sploščena (slika 96) ali pa bolj koničasta (slika 97) [Jesenko].



Slika 96: Sploščena porazdelitev podatkov [Jesenko]



Slika 97: Koničasta porazdelitev podatkov [Jesenko]

Porazdelitve z isto srednjo vrednostjo, isto mero variabilnosti in stopnjo asimetrije niso nujno med sebolj enake, pač pa so lahko ene bolj, druge pa manj sploščene. Sploščenost normalne porazdelitve imamo za idealno, zato jo primerjamo s sploščenostjo drugih porazdelitev [Artenjak].

Za podatke $x_1, x_2, ..., x_N$, ki niso podani v obliki frekvenčne porazdelitve, je koeficient sploščenosti enak [Jesenko]:

$$g_2 = \frac{\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^4}{\left(\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2\right)^2} \tag{4.99}$$

pri čemer je sorodnost z izrazom (2.181A) očitna. Velikokrat od vrednosti izraza (4.99) odštejemo še vrednost 3, kar nam omogoči primerjavo z normalno porazdelitvijo (glej tudi izraz (2.181B)). Tako dobimo **Fisherjev koeficient sploščenosti** [Jesenko]:

$$\gamma = g_2 - 3 \tag{4.100}$$

Kadar je le-ta pozitiven, je množica podatkov v primerjavi z normalno porazdelitvijo koničasta, če pa je negativen, je množica podatkov v primerjavi z normalno porazdelitvijo sploščena [Jesenko].

Pri diskretni frekvenčni porazdelitvi je koeficient sploščenosti enak [Jesenko]:

$$g_2 = \frac{\frac{1}{N}\sum_{i=1}^{n} f_i \cdot \left(x_i - \bar{x}\right)^4}{\left(\frac{1}{N}\sum_{i=1}^{n} f_i \cdot \left(x_i - \bar{x}\right)^2\right)^2} \tag{4.101}$$

Pri zvezni frekvenčni porazdelitvi je koeficient sploščenosti enak [Jesenko]:

$$g_2 = \frac{\frac{1}{N}\sum_{i=1}^{n} f_i \cdot \left(x_{si} - \overline{x}\right)^4}{\left(\frac{1}{N}\sum_{i=1}^{n} f_i \cdot \left(x_{si} - \overline{x}\right)^2\right)^2} \tag{4.102}$$

Primer 4.30.:

Tabela na sliki 98 prikazuje proizvodne čase (v minutah) določene vrste izdelka za izbrani vzorec. Izračunajte koeficient sploščenosti [Jesenko].

Čas izdelave	Frekvenca
22-26	22
26-30	28
30-34	36
34-38	41
38-42	45
42-46	38
46-50	32
50-54	19
54-58	12
58-62	4

Slika 98: Proizvodni časi (v minutah) določene vrste izdelka za izbrani vzorec [Jesenko]

Aritmetično sredino smo izračunali že prej (glej izraz (4.77)) in znaša: $\overline{x} = 39.1625 \approx 39.2$.

Slika 99 prikazuje delne izračune [Jesenko].

Razred	Frekvenca f_i	Sredina razreda x_{si}	$f_i \cdot \left(x_{si} - \overline{x}\right)^2$	$f_i \cdot \left(x_{si} - \overline{x}\right)^4$
22-26	22	24	5082,88	1174348,5952
26-30	28	28	3512,32	440585,4208
30-34	36	32	1866,24	96745,8816
34-38	41	36	419,84	4299,1616
38-42	45	40	28,80	18,4320
42-46	38	44	875,52	20171,9808
46-50	32	48	2478,08	191902,5152
50-54	19	52	3112,96	510027,3664
54-58	12	56	3386,88	955913,0112
58-62	4	60	1730,56	748709,4784
vsota	**N = 277**		**22494,08**	**4142721,8432**

Slika 99: Delni izračuni za proizvodne čase (v minutah) določene vrste izdelka za izbrani vzorec [Jesenko]

Če vstavimo te rezultate v izraz (4.99), dobimo:

$$g_2 = \frac{\frac{1}{277}\sum_{i=1}^{277}(x_i - \bar{x})^4}{\left(\frac{1}{277}\sum_{i=1}^{277}(x_i - \bar{x})^2\right)^2} = \frac{\frac{1}{277}\left(4.1427 \cdot 10^6\right)}{\left(\frac{1}{277} \cdot 22494\right)^2} = 2.2679 \tag{4.103}$$

Za Fisherjev koeficient sploščenosti potem sledi:

$$\gamma = g_2 - 3 = 2.2679 - 3 = -0.7321 \tag{4.104}$$

Vidimo torej, da je obravnavana množica podatkov v primerjavi z normalno porazdelitvijo sploščena. Izračune smo opravili z naslednjim programom v Matlabu:

```
% splos_pod.m
%
clear
clc
close all

ch = input('podatki iz jesenko primer (1), drugo(2)');
if ch == 1
   f = [22 28 36 41 45 38 32 19 12 4]
   s = [24 28 32 36 40 44 48 52 56 60]
else
   f=input('Vnesi vektor frekvenc f')
   s=input('Vnesi vektor sredin razredov s')
end

N = sum(f)
n = length(f)

d = 0;
for i=1:n
    d = d + f(i)*s(i);
end

xsr = d/N

if ch == 1
   ch1 = input('zelis zaokroziti xsr 1-da,0-ne');
   if ch1 == 1
      xsr = 39.2
```

```
   end
end
for i=1:n
    p2(i) = f(i)*(s(i)-xsr)^2;
    p4(i) = f(i)*(s(i)-xsr)^4;
end

disp('delni izračuni:')
p2'
p4'

p2s = sum(p2)
p4s = sum(p4)

stevec = p4s/N
imenov = (p2s/N)^2
g2=stevec/imenov;

disp('koeficient sploscenosti je:')
g2
```

Izgled komandnega okna je naslednji:

```
podatki iz jesenko primer (1), drugo(2)1
f =
    22    28    36    41    45    38    32    19    12     4
s =
    24    28    32    36    40    44    48    52    56    60
N =
   277
n =
    10
xsr =
   39.1625
zelis zaokroziti xsr 1-da,0-ne1
xsr =
   39.2000
delni izračuni:
ans =
  1.0e+003 *
    5.0829
    3.5123
    1.8662
    0.4198
    0.0288
    0.8755
    2.4781
    3.1130
    3.3869
```

```
    1.7306

ans =
  1.0e+006 *
    1.1743
    0.4406
    0.0967
    0.0043
    0.0000
    0.0202
    0.1919
    0.5100
    0.9559
    0.7487
p2s =
  2.2494e+004
p4s =
  4.1427e+006
stevec =
  1.4956e+004
imenov =
  6.5944e+003
koeficient sploscenosti je:
g2 =
    2.2679
```

Primer 4.31.:

*Dano imamo časovno vrsto iz datoteke **skewdata.txt**, ki smo jo prikazali na sliki 94 [Žibert]. Izračunajte sploščenost!*

Kot se izkaže, koeficient sploščenosti (4.99) zavzame naslednjo vrednost:

$$g_2 = \frac{\frac{1}{30}\sum_{i=1}^{30}\left(x_i - \bar{x}\right)^4}{\left(\frac{1}{30}\sum_{i=1}^{30}\left(x_i - \bar{x}\right)^2\right)^2} = 4.5993 \tag{4.105}$$

Za Fisherjev koeficient sploščenosti potem sledi:

$$\gamma = g_2 - 3 = 4.5993 - 3 = 1.5993 \tag{4.106}$$

Za izračun smo uporabili naslednji program v Matlabu:

```
% splos_pod1.m
clear
clc
close all

data = importdata('skewdata.txt');
data = data.data;
x = data;

plot(x,'LineWidth',1.5)
hold on
plot(x,'o','LineWidth',3)
grid
figure
hist(x)
grid

n = length(x)

d = 0;
for i=1:n
   d = d + x(i);
end

xsr = d/n

for i=1:n
   p2(i) = (x(i)-xsr)^2;
   p4(i) = (x(i)-xsr)^4;
end

p2s = sum(p2);
p4s = sum(p4);

stevec = p4s/n;
imenov = (p2s/n)^2;

disp('Koeficient sploscenosti:')

g2=stevec/imenov

g2fish = g2 - 3;

disp('Fisherjev koeficient sploscenosti - 1. nacin:')
g2fish
```

```
disp('Fisherjev koeficient sploscenosti - 2. nacin (s kurtosis.m in odstejes 3):')
g2fisha = kurtosis(data) – 3
```

Izpis komandnega okna je naslednji:

```
n =
    30
xsr =
   30.9687
Koeficient sploscenosti:
g2 =
    4.5993
Fisherjev koeficient sploscenosti - 1. nacin:
g2fish =
    1.5993
Fisherjev koeficient sploscenosti - 2. nacin (s kurtosis.m in odstejes 3):
g2fisha =
    1.5993
```

5 POSEBNE VERJETNOSTNE PORAZDELITVE

V tem poglavju bomo opisali nekatere pomembne diskretne in zvezne naključne spremeljivke, ki jih v statistiki pogosto uporabljamo. Tako bomo obravnavali diskretne spremenljivke: binomsko naključno spremenljivko, negativno binomsko naključno spremenljivko, hipergeometrično ter Poissonovo naključno spremenljivko. Poleg tega pa bomo obravnavali tudi zvezne spremenljivke: standardno normalno, gama in beta naključno spremenljivko.

5.1 Diskretne porazdelitve

5.1.1 Binomska porazdelitev

Binomske naključne spremenljivke smo se dotaknili že pri obravnavi teorije verjetnosti v poglavju 2.4, kjer smo prišli do izraza (2.24). Zapišimo ta izraz še enkrat:

$$P(x) = P(X = x) = \binom{n}{x} \cdot p^x \cdot \underbrace{(1-p)}_{q}^{n-x}, \quad x = 0,1,2,...,n \tag{5.1}$$

Na osnovi izraza (2.186) lahko izračunamo rodovno funkcijo momentov:

$$M(t) = E\left(e^{t \cdot X}\right) = \sum_{x=0}^{n} e^{t \cdot x} \cdot p(x) = \sum_{x=0}^{n} e^{t \cdot x} \cdot \binom{n}{x} \cdot p^x \cdot q^{n-x} = \sum_{x=0}^{n} \binom{n}{x} \cdot \left(e^t \cdot p\right)^x \cdot q^{n-x} \tag{5.2}$$

Kot vemo, se binomski izrek glasi:

$$(a+b)^n = \sum_{x=0}^{n} \binom{n}{x} \cdot b^x \cdot a^{n-x} \tag{5.3}$$

Tako dobimo:

$$M(t) = \left(e^t \cdot p + q\right)^n \tag{5.4}$$

Če želimo poiskati matematično upanje binomske naključne spremenljivke, lahko uporabimo izraz (2.178), od koder se vidi, da velja:

$$\mu_1' = E\left(X^1\right) = E\left(X\right) = \mu \tag{5.5}$$

Če upoštevamo izraz (2.188), dobimo:

$$\begin{aligned} \mu_r' &= E\left(X^r\right) = \left.\frac{d^r M\left(t\right)}{dt^r}\right|_{t=0} \\ \mu_1' &= E\left(X^1\right) = \left.\frac{d^1 M\left(t\right)}{dt^1}\right|_{t=0} = \frac{dM\left(0\right)}{dt} \end{aligned} \tag{5.6}$$

Tvorimo $\frac{dM\left(t\right)}{dt}$:

$$\frac{dM\left(t\right)}{dt} = \frac{d}{dt}\left(e^t \cdot p + q\right)^n = n\left(e^t \cdot p + q\right)^{n-1} \cdot e^t \cdot p \tag{5.7}$$

Sledi za matematično upanje:

$$\begin{aligned} &\mu_1' = E\left(X^1\right) = \frac{dM\left(0\right)}{dt} = n\left(e^0 \cdot p + q\right)^{n-1} \cdot e^0 \cdot p = \\ &= n\underbrace{\left(p + q\right)^{n-1}}_{1} \cdot p = n \cdot p \end{aligned} \tag{5.8}$$

Če želimo izračunati varianco, upoštevamo izraza (2.177) in (2.178). Varianca je enaka:

$$\mu_2 = E\left(X^2\right) - E^2\left(X\right) = \mu_2' - \left(\mu_1'\right)^2 \tag{5.9}$$

μ_2' izračunamo na naslednji način:

$$\mu_r' = E\left(X^r\right) = \left.\frac{d^r M(t)}{dt^r}\right|_{t=0} \qquad (5.10)$$
$$\mu_2' = E\left(X^2\right) = \left.\frac{d^2 M(t)}{dt^2}\right|_{t=0} = \frac{d^2 M(0)}{dt^2}$$

Sledi:

$$\begin{aligned}
\mu_2' &= \frac{d}{dt}\left(n\left(e^t \cdot p + q\right)^{n-1} \cdot e^t \cdot p\right)_{t=0} = \\
&= n \cdot p \cdot \frac{d}{dt}\left(\left(e^t \cdot p + q\right)^{n-1} \cdot e^t\right)_{t=0} = \\
&= n \cdot p \cdot \left((n-1)\left(e^t \cdot p + q\right)^{n-2} \cdot e^t \cdot p \cdot e^t + \left(e^t \cdot p + q\right)^{n-1} \cdot e^t\right)_{t=0} = \\
&= \left(n \cdot p^2 \cdot e^{2t}(n-1)\left(e^t \cdot p + q\right)^{n-2} + n \cdot p \cdot \left(e^t \cdot p + q\right)^{n-1} \cdot e^t\right)_{t=0} = \\
&= \left(n \cdot p^2 \cdot e^0 (n-1)\left(e^0 \cdot p + q\right)^{n-2} + n \cdot p \cdot \left(e^0 \cdot p + q\right)^{n-1} \cdot e^0\right) = \\
&= \left(n \cdot p^2 (n-1)(p+q)^{n-2} + n \cdot p \cdot (p+q)^{n-1}\right) = \\
&= \left(n \cdot p^2 (n-1)(1)^{n-2} + n \cdot p \cdot (1)^{n-1}\right) = \\
&= \left(n \cdot p^2 (n-1) + n \cdot p\right) = n^2 \cdot p^2 - n \cdot p^2 + n \cdot p
\end{aligned} \qquad (5.11)$$

Torej je varianca enaka:

$$\begin{aligned}
\mu_2 &= \mu_2' - \left(\mu_1'\right)^2 = n^2 \cdot p^2 - n \cdot p^2 + n \cdot p - (n \cdot p)^2 = \\
&= n \cdot p - n \cdot p^2 = n \cdot p(1-p) = n \cdot p \cdot q
\end{aligned} \qquad (5.12)$$

Na podoben način bi lahko tudi pokazali, da za koeficient asimetričnosti velja [Jesenko, Krishnamoorthy]:

$$g_1 = \frac{q-p}{\sqrt{n \cdot p \cdot q}} = \frac{1-p-p}{\sqrt{n \cdot p \cdot (1-p)}} = \frac{1-2p}{\sqrt{n \cdot p \cdot (1-p)}} \qquad (5.13)$$

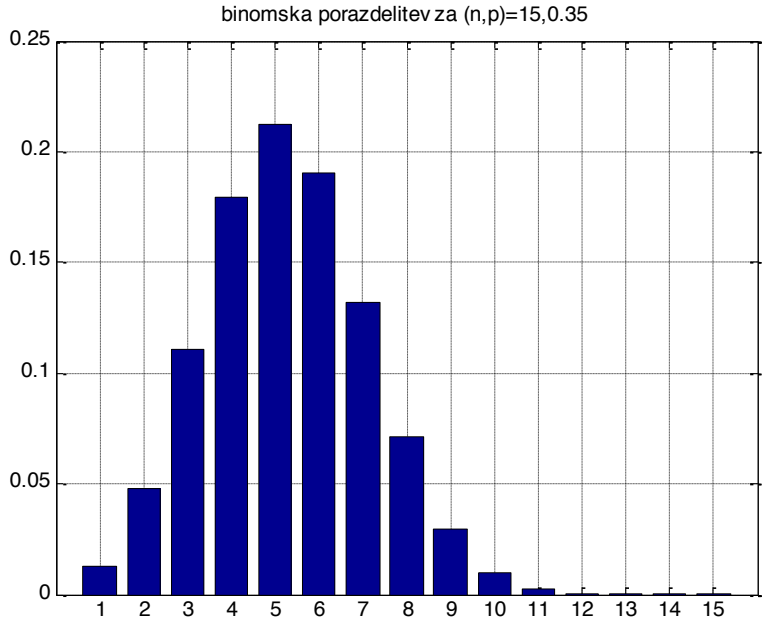
za koeficient sploščenosti pa velja [Jesenko, Krishnamoorthy]:

$$g_2 = 3 + \frac{1 - 6p \cdot q}{n \cdot p \cdot q} = 3 - \frac{6}{n} + \frac{1}{n \cdot p \cdot (1-p)} \tag{5.14}$$

oz. za Fisherjev koeficient sploščenosti [Jesenko]:

$$\gamma = g_2 - 3 = \frac{1 - 6p \cdot q}{n \cdot p \cdot q} = -\frac{6}{n} + \frac{1}{n \cdot p \cdot (1-p)} \tag{5.15}$$

Porazdelitveni zakon binomske naključne spremenljivke lahko seveda ponazorimo tudi grafično. Primer porazdelitve pri $n = 15$ in $p = 0.35$ prikazuje slika 100.



Slika 100: Primer binomske porazdelitve pri n= 15 in p = 0.35

Za izris slike 100 smo uporabili naslednji program v Matlabu:

```
% binomska porazdelitev: binomska1.m

clear
clc
close all

n = input('Vnesi n')
p = input('Vnesi p')
q = 1 - p
x = 1:1:n

warning off
```

```
for i=1:n
    P(i) = nchoosek(n,i)*p^i*q^(n-i);
end

disp('Binomska porazdelitev je:')
P

if n/20 < 1
   dx = n/20;
else
   dx = 0.8;
end

bar(x,P,dx)

grid

astr = ['binomska porazdelitev za (n,p)=' num2str(n) ',' num2str(p) ];
title(astr)

warning on
```

Izgled komandnega okna je naslednji:

```
Vnesi n15
n =
   15

Vnesi p0.35
p =
   0.3500

q =
   0.6500

x =

   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15

Binomska porazdelitev je:
P =

 Columns 1 through 11

   0.0126   0.0476   0.1110   0.1792   0.2123   0.1906   0.1319   0.0710   0.0298   0.0096   0.0024

 Columns 12 through 15

   0.0004   0.0001   0.0000   0.0000
```

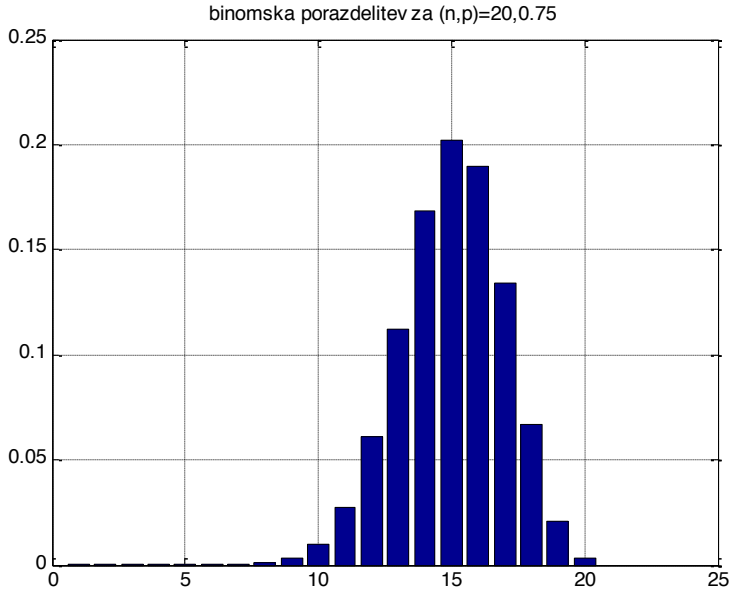
Primer 5.1.:

V sanatoriju so ugotovili, da ozdravi 75% bolnikov, zbolelih za neko boleznijo, če jih zdravijo z določeno terapijo. Trenutno se po dotični terapiji zdravi 20 bolnikov. Kakšna je verjetnost, da jih bo ozdravelo natanko 16? Kakšna je verjetnost, da jih bo ozdravelo največ osem (torej 8 ali manj)? Izračunajte tudi matematično upanje, varianco in deviacijo! [Jesenko]

Iz naloge lahko razberemo naslednje:

$n = 20,$
$p = 0.75$
$q = 1 - p = 0.25$

Z uporabo programa **binomska1.m** lahko izrišemo porazdelitev, kar prikazuje slika 101.



Slika 101: Primer binomske porazdelitve pri n= 20 in p = 0.75

Izgled komandnega okna je naslednji:

```
Vnesi n20
n =
    20

Vnesi p0.75
p =
    0.7500

q =
    0.2500

x =
  Columns 1 through 19

     1     2     3     4     5     6     7     8     9    10    11    12    13    14    15    16    17    18    19
  Column 20
    20

Binomska porazdelitev je:
P =
  Columns 1 through 11

    0.0000    0.0000    0.0000    0.0000    0.0000    0.0000    0.0002    0.0008    0.0030    0.0099    0.0271

  Columns 12 through 20

    0.0609    0.1124    0.1686    0.2023    0.1897    0.1339    0.0669    0.0211    0.0032
```

Odtod lahko takoj vidimo, da je $P(X=16)=P(16)=\binom{20}{16}\cdot 0.75^{16}\cdot 0.25^{20-16}=0.1897$. Do tega rezultata bi prišli tudi, če bi klicali program **binomska2.m**:

```
% binomska porazdelitev: binomska2.m

clear
clc
close all

n = input('Vnesi n')
p = input('Vnesi p')
q = 1 - p
x = input('Vnesi x')

disp('Verjetnost je')

P = nchoosek(n,x)*p^x*q^(n-x)
```

Izgled komandnega okna bi bil naslednji:

```
Vnesi n20
n =
   20

Vnesi p0.75
p =
   0.7500

q =
   0.2500

Vnesi x16
x =
   16

Verjetnost je
P =
   0.1897
```

Torej je 18.97% verjetnosti, da bo ozdravelo natanko 16 bolnikov.

Verjetnost, da jih bo ozdravelo največ osem (torej 8 ali manj), dobimo na naslednji način:

$$P(X\le 8)=P(0)+P(1)+\ldots+P(8)=$$
$$=\binom{20}{0}\cdot 0.75^{0}\cdot 0.25^{20-0}+\binom{20}{1}\cdot 0.75^{1}\cdot 0.25^{20-1}+\ldots+\binom{20}{8}\cdot 0.75^{8}\cdot 0.25^{20-8}= \qquad (5.16)$$
$$=9.35\cdot 10^{-4}$$

Matematično upanje izračunamo na osnovi izraza (5.8):

$$\mu_1' = n \cdot p = 20 \cdot 0.75 = 15 \tag{5.17}$$

Varianco izračunamo na osnovi izraza (5.12):

$$\mu_2 = n \cdot p \cdot q = 20 \cdot 0.75 \cdot 0.25 = 3.75 \tag{5.18}$$

Standardni odklon pa je:

$$\sigma = \sqrt{n \cdot p \cdot q} = \sqrt{20 \cdot 0.75 \cdot 0.25} = \sqrt{3.75} = 1.9365 \tag{5.19}$$

Za zgornje izračune smo uporabili naslednji program v Matlabu:

```
% binomska porazdelitev: binomska3.m

clear
clc
close all

n = input('Vnesi n')
p = input('Vnesi p')
q = 1 - p
xmax = input('Vnesi xmax')

x = 1:1:xmax

P = 0;

for i=1:1:xmax
    P = P + nchoosek(n,i)*p^i*q^(n-i);
end

disp('Vsota verjetnosti je:')

P

disp('Matematicno upanje je:')

MU = n*p

disp('Varianca je:')

VAR = n*p*q

disp('Deviacija je:')

STD = sqrt(VAR)
```

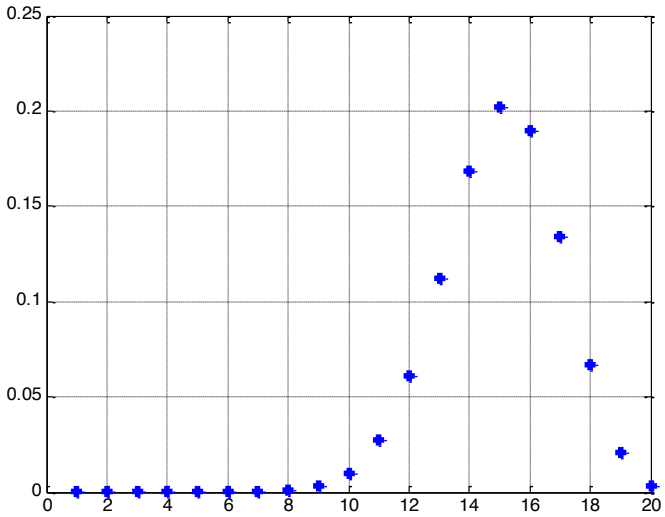
Izgled komandnega okna je naslednji:

```
Vnesi n20
n =
   20

Vnesi p0.75
p =
   0.7500

q =
   0.2500

Vnesi xmax8
xmax =
    8

x =
    1    2    3    4    5    6    7    8

Vsota verjetnosti je:
P =
  9.3539e-004

Matematicno upanje je:
MU =
   15

Varianca je:
VAR =
   3.7500

Deviacija je:
STD =
   1.9365
```

Torej v povprečju ozdravi 15 bolnikov izmed 20 zdravljenih. Število ozdravelih bolnikov pa v povprečju odstopa od 15 pacientov za 1.93 pacienta.

Poglejmo si še, kako bi narisali porazdelitev verjetnosti za ta primer s standardnimi Matlab ukazi (slika 102):

```
>> x = 1:20;
>> y = binopdf(x,20,0.75)

y =

  Columns 1 through 12

   0.0000    0.0000    0.0000    0.0000    0.0000    0.0000    0.0002    0.0008    0.0030    0.0099    0.0271    0.0609

  Columns 13 through 20

   0.1124    0.1686    0.2023    0.1897    0.1339    0.0669    0.0211    0.0032

>> plot(x,y,'+','LineWidth',3)
>> grid
```



Slika 102: Primer binomske porazdelitve pri n= 20 in p = 0.75 (standardni Matlab ukazi)

5.1.2 Geometrijska porazdelitev

Denimo, da opazujemo zaporedne Bernoullijeve poskuse, ki so med seboj neodvisni. Ponavljamo jih, dokler se ne zgodi prvi uspeh, ki ima verjetnost *p*. Definiramo naključno spremenljivko *X*, ki meri število poskusov, dokler se ne zgodi uspešen dogodek [Elezović].

Velja naslednje:

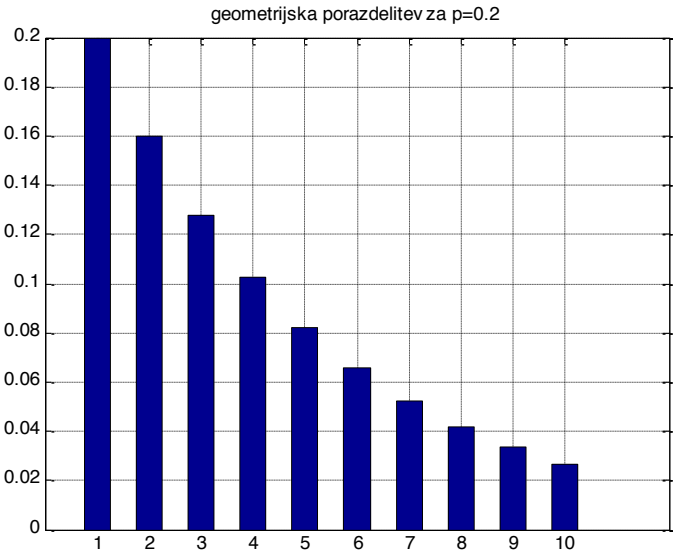
$$P(X=x)=P(X)=\underbrace{(1-p)\cdot(1-p)\cdot\ldots\cdot(1-p)}_{x-1\ neuspehov}\cdot\underbrace{p}_{uspeh}=\underbrace{(1-p)^{x-1}}_{q}\cdot p \qquad (5.20)$$

$x = 1,2,3,...$

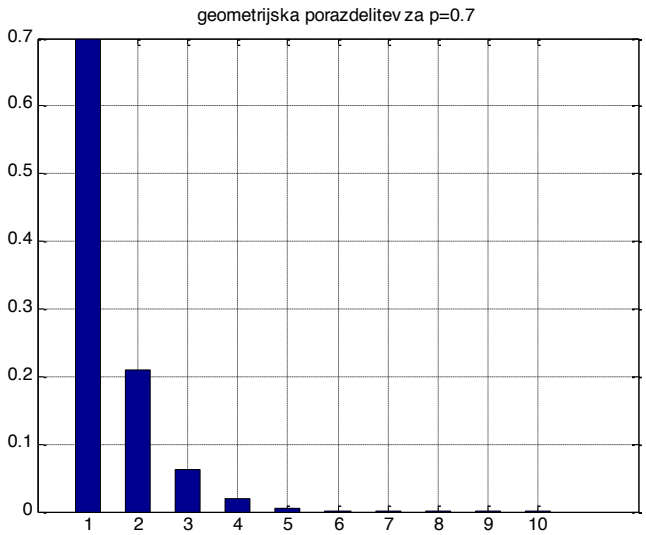
Matematično upanje izračunamo na naslednji način [Turk]:

$$\begin{aligned}
&E(X)=\sum_{x=1}^{\infty}x\cdot P(x)=\sum_{x=1}^{\infty}x\cdot(1-p)^{x-1}\cdot p=p\sum_{x=1}^{\infty}x\cdot(1-p)^{x-1}=\\
&=p\sum_{x=1}^{\infty}\frac{d}{dp}\left[-(1-p)^{x}\right]=-p\frac{d}{dp}\left(\sum_{x=1}^{\infty}(1-p)^{x}\right)=-p\frac{d}{dp}\left(\sum_{x=0}^{\infty}(1-p)^{x}-1\right)=\\
&=-p\frac{d}{dp}\left(\sum_{x=0}^{\infty}(1-p)^{x}\right)=-p\frac{d}{dp}\left(\frac{1}{1-(1-p)}\right)=-p\frac{d}{dp}\left(\frac{1}{p}\right)=-p\left(-\frac{1}{p^{2}}\right)=\\
&=\frac{1}{p}
\end{aligned} \qquad (5.21)$$

Sliki 103 in 104 prikazujeta porazdelitev verjetnosti za geometrijsko spremenljivko pri dveh različnih vrednostih *p*, pri čemer je prikazanih 10 vrednosti ($x \leq 10$).



Slika 103: Geometrijska porazdelitev pri p = 0.2



Slika 104: Geometrijska porazdelitev pri p = 0.7

Pri tem smo uporabili naslednji program v Matlabu:

```
% geometrijska porazdelitev: geomet1.m

clear
clc
close all

p = input('Vnesi p')
n = input('Za kako velik n naj pokazem plot?')
q = 1 - p
x = 1:1:n

warning off

for i=1:n
    P(i) = p*q^(i-1);
end

disp('Geometrijska porazdelitev je:')
P

if n/20 < 1
   dx = n/20;
else
   dx = 0.8;
end

bar(x,P,dx)

grid

astr = ['geometrijska porazdelitev za p=' num2str(p)];
title(astr)

warning on

disp('Geometrijska porazdelitev s standardnimi matlab ukazi:')

P1 = geopdf(x-1,p)

ch = input('Zelis dodatne izracune 1-DA,0-NE');
if ch == 0
   break
end

xmax = input('Vnesi xmax')

x = 1:1:xmax

P = 0;

for i=1:1:xmax
    P = P + p*q^(i-1);
end

disp('Vsota verjetnosti je:')

P

disp('Matematicno upanje je:')

MU = 1/p
```

Izpis komandnega okna pri $p = 0.2$ je naslednji:

```
Vnesi p0.2
p =
    0.2000
Za kako velik n naj pokazem plot?10
n =
    10
q =
    0.8000
x =

     1     2     3     4     5     6     7     8     9    10
Geometrijska porazdelitev je:
P =

    0.2000    0.1600    0.1280    0.1024    0.0819    0.0655    0.0524    0.0419    0.0336    0.0268

Geometrijska porazdelitev s standardnimi matlab ukazi:
P1 =
    0.2000    0.1600    0.1280    0.1024    0.0819    0.0655    0.0524    0.0419    0.0336    0.0268

Zelis dodatne izracune 1-DA,0-NE0
```

Izpis komandnega okna pri $p = 0.7$ je naslednji:

```
Vnesi p0.7
p =
    0.7000
Za kako velik n naj pokazem plot?10
n =
    10
q =
    0.3000
x =
     1     2     3     4     5     6     7     8     9    10
Geometrijska porazdelitev je:
P =
    0.7000    0.2100    0.0630    0.0189    0.0057    0.0017    0.0005    0.0002    0.0000    0.0000

Geometrijska porazdelitev s standardnimi matlab ukazi:
P1 =

    0.7000    0.2100    0.0630    0.0189    0.0057    0.0017    0.0005    0.0002    0.0000    0.0000

Zelis dodatne izracune 1-DA,0-NE0
```

Primer 5.2.:

Mečemo kocko, pri čemer število metanj kocke do uspešnega dogodka, ko pade šestica, meri naključna spremenljivka z geometrijsko porazdelitvijo. Kolikšna je verjetnost, da se bo šestica pojavila pri šestih metih?

V splošnem velja:

$$P(X=x)=P(X)=\underbrace{(1-p)\cdot(1-p)\cdot\ldots\cdot(1-p)}_{x-1\ neuspehov}\cdot\underbrace{p}_{uspeh}=\underbrace{\left(1-\frac{1}{6}\right)^{x-1}}_{q}\cdot\frac{1}{6}= \tag{5.22}$$

$$=\left(\frac{5}{6}\right)^{x-1}\cdot\frac{1}{6},\qquad x=1,2,3,\ldots$$

Da se bo šestica pojavila pri šestih metih, je verjetnost enaka:

$$P(X\le 6)=$$

$$=\left(\frac{5}{6}\right)^{1-1}\cdot\frac{1}{6}+\left(\frac{5}{6}\right)^{2-1}\cdot\frac{1}{6}+\left(\frac{5}{6}\right)^{3-1}\cdot\frac{1}{6}+\left(\frac{5}{6}\right)^{4-1}\cdot\frac{1}{6}+\left(\frac{5}{6}\right)^{5-1}\cdot\frac{1}{6}+\left(\frac{5}{6}\right)^{6-1}\cdot\frac{1}{6}= \tag{5.23}$$

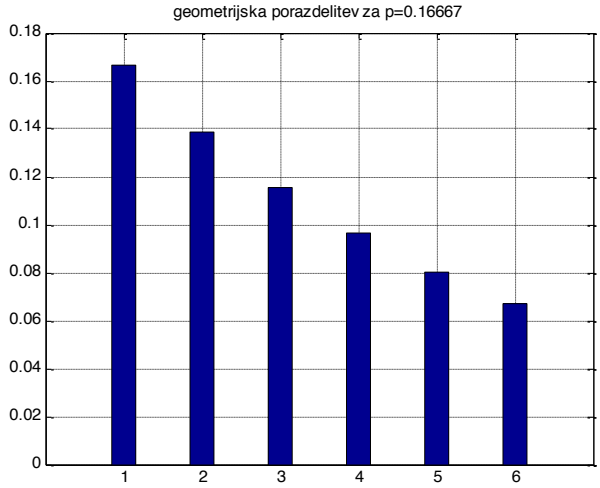
$$=\frac{1}{6}\left[1+\frac{5}{6}+\left(\frac{5}{6}\right)^{2}+\left(\frac{5}{6}\right)^{3}+\left(\frac{5}{6}\right)^{4}+\left(\frac{5}{6}\right)^{5}\right]=0.6651$$

Za ta izračun lahko uporabimo tudi program **geomet1.m.** Izpis komandnega okna je naslednji:

```
Vnesi p1/6
p =
    0.1667
Za kako velik n naj pokazem plot?6
n =
     6
q =
    0.8333
x =
     1     2     3     4     5     6
Geometrijska porazdelitev je:
P =
    0.1667    0.1389    0.1157    0.0965    0.0804    0.0670

Geometrijska porazdelitev s standardnimi matlab ukazi:
P1 =
    0.1667    0.1389    0.1157    0.0965    0.0804    0.0670

Zelis dodatne izracune 1-DA,0-NE1
Vnesi xmax6
xmax =
     6
x =
     1     2     3     4     5     6

Vsota verjetnosti je:
P =
    0.6651

Matematicno upanje je:
MU =
     6
```

Kot vidimo, je program izračunal tudi matematično upanje, ki je: $E(X)=\frac{1}{p}=\frac{1}{1/6}=6$. Torej je pričakovano število ponavljanja meta kocke, da bi padla šestica, enako 6.

Slika 105 prikazuje porazdelitev verjetnosti za geometrijsko spremenljivko pri vrednosti $p=\frac{1}{6}$, pri čemer je prikazanih 6 vrednosti ($x \le 6$).



Slika 105: Geometrijska porazdelitev pri p = 1/6, prikazanih 6 vrednosti

Seveda verjetnost, da še ni dosežen ugoden dogodek, z naraščanjem števila neuspelih poskusov pada. To z drugimi besedami pomeni, da z naraščanjem števila neuspelih poskusov raste verjetnost, da se bo le zgodil uspešen dogodek.

Na primer, da se bo šestica pojavila pri osmih metih, je verjetnost enaka:

$$\begin{aligned} & P(X \le 8) = \\ & = \left(\frac{5}{6}\right)^{1-1} \cdot \frac{1}{6} + \left(\frac{5}{6}\right)^{2-1} \cdot \frac{1}{6} + \left(\frac{5}{6}\right)^{3-1} \cdot \frac{1}{6} + \\ & + \left(\frac{5}{6}\right)^{4-1} \cdot \frac{1}{6} + \left(\frac{5}{6}\right)^{5-1} \cdot \frac{1}{6} + \left(\frac{5}{6}\right)^{6-1} \cdot \frac{1}{6} + + \left(\frac{5}{6}\right)^{7-1} \cdot \frac{1}{6} + \left(\frac{5}{6}\right)^{8-1} \cdot \frac{1}{6} = \\ & = \frac{1}{6}\left[1 + \frac{5}{6} + \left(\frac{5}{6}\right)^{2} + \left(\frac{5}{6}\right)^{3} + \left(\frac{5}{6}\right)^{4} + \left(\frac{5}{6}\right)^{5} + \left(\frac{5}{6}\right)^{6} + \left(\frac{5}{6}\right)^{7}\right] = 0.7674 \end{aligned} \tag{5.24}$$

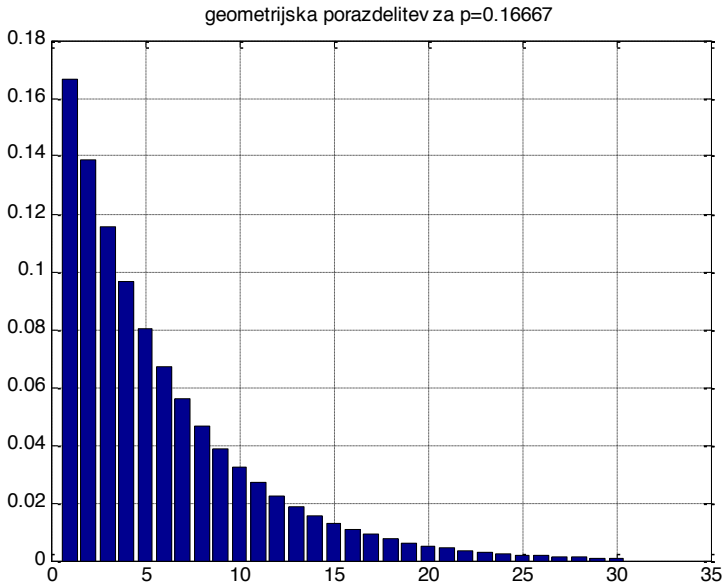
torej se verjetnost poveča že na 76.74%.

Na primer, da se bo šestica pojavila pri petnajstih metih, pa je verjetnost enaka:

$$\begin{aligned} &P(X \le 15) = \\ &= \left(\frac{5}{6}\right)^{1-1} \cdot \frac{1}{6} + \left(\frac{5}{6}\right)^{2-1} \cdot \frac{1}{6} + \left(\frac{5}{6}\right)^{3-1} \cdot \frac{1}{6} + \ldots + \left(\frac{5}{6}\right)^{15-1} \cdot \frac{1}{6} = \\ &= \frac{1}{6}\left[1 + \frac{5}{6} + \left(\frac{5}{6}\right)^{2} + \ldots + \left(\frac{5}{6}\right)^{14}\right] = 0.9351 \end{aligned} \tag{5.25}$$

torej se verjetnost poveča že na 93.51%.

Slika 106 prikazuje porazdelitev verjetnosti za geometrijsko spremenljivko pri vrednosti $p = \frac{1}{6}$, pri čemer je prikazanih 30 vrednosti ($x \le 30$).



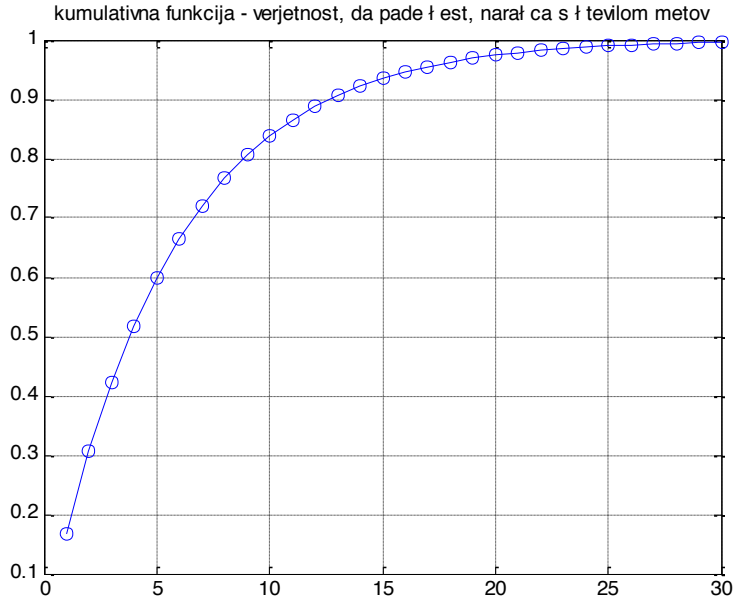
Slika 106: Geometrijska porazdelitev pri p = 1/6, prikazanih 30 vrednosti

Kot se izkaže, je verjetnost, da se bo šestica pojavila pri 30 metih, enaka:

$$\begin{aligned} &P(X \le 30) = \\ &= \left(\frac{5}{6}\right)^{1-1} \cdot \frac{1}{6} + \left(\frac{5}{6}\right)^{2-1} \cdot \frac{1}{6} + \left(\frac{5}{6}\right)^{3-1} \cdot \frac{1}{6} + \ldots + \left(\frac{5}{6}\right)^{30-1} \cdot \frac{1}{6} = \\ &= \frac{1}{6}\left[1 + \frac{5}{6} + \left(\frac{5}{6}\right)^{2} + \ldots + \left(\frac{5}{6}\right)^{29}\right] = 0.9958 \end{aligned} \tag{5.26}$$

Verjetnosti, da se bo šestica pojavila pri *x* metih, lahko dobimo z naslednjimi ukazi (kumulativna funkcija!, glej sliko 107):

```
>> x=1:30
x =
  Columns 1 through 20
    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16   17   18   19   20
  Columns 21 through 30
   21   22   23   24   25   26   27   28   29   30

>> F = geocdf(x-1,1/6)
F =

  Columns 1 through 12
   0.1667   0.3056   0.4213   0.5177   0.5981   0.6651   0.7209   0.7674   0.8062   0.8385   0.8654   0.8878

  Columns 13 through 24
   0.9065   0.9221   0.9351   0.9459   0.9549   0.9624   0.9687   0.9739   0.9783   0.9819   0.9849   0.9874

  Columns 25 through 30
   0.9895   0.9913   0.9927   0.9939   0.9949   0.9958

>> plot(F)
>> hold on
>> plot(F,'o')
>> grid
>> title('kumulativna funkcija - verjetnost, da pade šest, narašca s številom metov')
```



Slika 107: Kumulativna funkcija: Verjetnost, da se bo šestica pojavila pri x metih, narašča s številom metov

Lahko pa pokličemo tudi program:

```
% geom_kumf.m

clear
clc
close all

x=1:30
```

```
F = geocdf(x-1,1/6)

plot(F)
hold on
plot(F,'o')
grid
title('kumulativna funkcija - verjetnost, da pade šest, narašca s številom metov')
```

5.1.3 Negativna binomska (Pascalova) porazdelitev

Tudi pri tej porazdelitvi opazujemo zaporedne Bernoullijeve poskuse. Želimo, da se uspeh ponovi *r-krat*. Najprej ponavljamo poskuse, dokler se ne zgodi prvi uspeh, ki ima verjetnost *p*. Nato nadaljujemo poskuse, dokler se ne zgodi drugi uspeh, ki ima verjetnost *p*. Itn. To dogajanje opazujemo, dokler se ne zgodi *r-ti* uspeh. Naključna spremenljivka *X*, ki meri število poskusov, dokler se ne zgodi uspešen dogodek *r-tič*, je porazdeljena po Pascalovi oz. negativni binomski porazdelitvi [Turk].

Pri izvajanju poskusov lahko zapišemo naslednjo delno verjetnost:

$$P_d = \underbrace{(1-p)\cdot(1-p)\cdot\ldots\cdot(1-p)}_{n_1-1\ neuspehov}\cdot\underbrace{p}_{1.uspeh}\cdot\underbrace{(1-p)\cdot(1-p)\cdot\ldots\cdot(1-p)}_{n_2-1\ neuspehov}\cdot\underbrace{p}_{2.uspeh}\cdot\ldots$$
$$\cdot\underbrace{(1-p)\cdot(1-p)\cdot\ldots\cdot(1-p)}_{n_r-1\ neuspehov}\cdot\underbrace{p}_{r.uspeh} = \tag{5.27}$$
$$= (1-p)^{n_1-1+n_2-1+\ldots+n_r-1}\cdot p^{1+1+\ldots+1} = (1-p)^{(n_1+n_2+\ldots+n_r)-r}\cdot p^r$$

Število vseh poskusov do vključno *r.tega* uspeha je enako:

$$x = n_1 - 1 + 1 + n_2 - 1 + 1 + n_r - 1 + 1 = n_1 + n_2 + \ldots + n_r \tag{5.28}$$

Sledi:

$$P_d = (1-p)^{x-r}\cdot p^r \tag{5.29}$$

Seveda velja, da je $x = r, r+1, r+2, \ldots$, saj se mora zgoditi vsaj *r* poskusov ali več, da lahko računamo na *r* uspehov. Da bi lahko nastavili verjetnostno funkcijo za Pascalovo porazdelitev, moramo seveda s členom $\binom{x-1}{r-1}$ še upoštevati število vseh možnih zaporedij, podobno, kot smo to imeli pri binomski porazdelitvi. Tako dobimo [Jesenko, Turk]:

$$P(X=x) = \binom{x-1}{r-1} \cdot P_d = \binom{x-1}{r-1} \cdot (1-p)^{x-r} \cdot p^r, \qquad (5.30)$$
$$x = r, r+1, r+2, \ldots$$
$$q = 1-p$$

Če je *r = 1*, sledi:

$$P(X=x) = \binom{x-1}{1-1} \cdot (1-p)^{x-1} \cdot p^1, \qquad (5.31)$$
$$x = 1, 1+1, 1+2, \ldots$$
$$q = 1-p$$

in dobimo:

$$P(X=x) = 1 \cdot (1-p)^{x-1} \cdot p, \qquad (5.32)$$
$$x = 1, 2, 3, \ldots$$
$$q = 1-p$$

Torej očitno Pascalova porazdelitev preide v geometrijsko pri *r = 1*.

V nadaljevanju izračunajmo matematično upanje za Pascalovo porazdelitev. V ta namen tvorimo:

$$\begin{aligned}
&\mu = E(X) = \sum_{x=r}^{\infty} x \cdot P(x) = \sum_{x=r}^{\infty} x \cdot \binom{x-1}{r-1} \cdot (1-p)^{x-r} \cdot p^r = \\
&= \sum_{x=r}^{\infty} x \cdot \frac{(x-1)!}{(r-1)! \cdot (x-1-r+1)!} \cdot (1-p)^{x-r} \cdot p^r = \\
&= \frac{p \cdot r}{p \cdot r} \sum_{x=r}^{\infty} x \cdot \frac{(x-1)!}{(r-1)! \cdot (x-r)!} \cdot (q)^{x-r} \cdot p^r = \\
&= \frac{r}{p} \sum_{x=r}^{\infty} x \cdot \frac{(x-1)!}{r(r-1)! \cdot (x-r)!} \cdot (q)^{x-r} \cdot p^{r+1} = \\
&= \frac{r}{p} \sum_{x=r}^{\infty} \frac{(x)!}{(r)! \cdot (x-r)!} \cdot (q)^{x-r} \cdot p^{r+1} = \\
&= \frac{r}{p} \sum_{x=r}^{\infty} \binom{x}{r} \cdot (q)^{x-r} \cdot p^{r+1}
\end{aligned} \tag{5.33}$$

Vpeljemo novi spremenljivki $x = y-1, r = k-1$ in dobimo:

$$\begin{aligned}
&\mu = \frac{r}{p} \sum_{y=k-1+1}^{\infty} \binom{y-1}{k-1} \cdot (q)^{y-1-k+1} \cdot p^{k-1+1} = \\
&= \frac{r}{p} \sum_{y=k}^{\infty} \binom{y-1}{k-1} \cdot (q)^{y-k} \cdot p^k = \\
&= \frac{r}{p} \underbrace{\left[P(y=k) + P(y=k+1) + \ldots \right]}_{1} = \frac{r}{p}
\end{aligned} \tag{5.34}$$

Ta rezultat bi lahko dobili tudi z drugačnim razmišljanjem. Na osnovi izraza (5.28) lahko zapišemo:

$$\begin{aligned}
&x = n_1 + n_2 + \ldots + n_r \\
&E(X) = E(N_1 + N_2 + \ldots + N_r) = \\
&= E(N_1) + E(N_2) + \ldots + E(N_r) = \\
&= \frac{1}{p} + \frac{1}{p} + \ldots + \frac{1}{p} = \frac{r}{p}
\end{aligned} \tag{5.35}$$

Torej, ker je Pascalova naključna spremenljivka X vsota geometrijskih naključnih spremenljivk N_i, lahko uporabimo izraz (5.21) pri izračunu matematičnega upanja.

V nadaljevanju izračunajmo drugi moment za Pascalovo porazdelitev. V ta namen tvorimo:

$$\begin{aligned}
&E\left(X^2\right)=\sum_{x=r}^{\infty}x^2\cdot P(x)=\sum_{x=r}^{\infty}x^2\cdot\binom{x-1}{r-1}\cdot(1-p)^{x-r}\cdot p^r=\\
&=\sum_{x=r}^{\infty}x^2\cdot\frac{(x-1)!}{(r-1)!\cdot(x-1-r+1)!}\cdot(1-p)^{x-r}\cdot p^r=\\
&=\frac{p\cdot r}{p\cdot r}\sum_{x=r}^{\infty}x\cdot\frac{(x)!}{(r-1)!\cdot(x-r)!}\cdot(q)^{x-r}\cdot p^r=\\
&=\frac{r}{p}\sum_{x=r}^{\infty}x\cdot\frac{(x)!}{r(r-1)!\cdot(x-r)!}\cdot(q)^{x-r}\cdot p^{r+1}=\\
&=\frac{r}{p}\sum_{x=r}^{\infty}x\cdot\frac{(x)!}{(r)!\cdot(x-r)!}\cdot(q)^{x-r}\cdot p^{r+1}=\\
&=\frac{r}{p}\sum_{x=r}^{\infty}x\binom{x}{r}\cdot(q)^{x-r}\cdot p^{r+1}
\end{aligned}\tag{5.33}$$

Vpeljemo novi spremenljivki $x=y-1, r=k-1$ in dobimo:

$$\begin{aligned}
&E\left(X^2\right)=\frac{r}{p}\sum_{y=k-1+1}^{\infty}(y-1)\binom{y-1}{k-1}\cdot(q)^{y-1-k+1}\cdot p^{k-1+1}=\\
&=\frac{r}{p}\sum_{y=k}^{\infty}y\binom{y-1}{k-1}\cdot(q)^{y-k}\cdot p^k-\frac{r}{p}\sum_{y=k}^{\infty}\binom{y-1}{k-1}\cdot(q)^{y-k}\cdot p^k=\\
&=\frac{r}{p}\left(\sum_{y=k}^{\infty}y\binom{y-1}{k-1}\cdot(q)^{y-k}\cdot p^k-\sum_{y=k}^{\infty}\binom{y-1}{k-1}\cdot(q)^{y-k}\cdot p^k\right)=\\
&=\frac{r}{p}\left(E(Y)-1\right)=\frac{r}{p}\left(\frac{k}{p}-1\right)=\frac{r}{p}\left(\frac{r+1}{p}-1\right)=\frac{r(r+1)}{p^2}-\frac{r}{p}
\end{aligned}\tag{5.34}$$

Varianca je enaka:

$$\begin{aligned}
&VAR(X)=E\left(X^2\right)-E^2(X)=\frac{r(r+1)}{p^2}-\frac{r}{p}-\left(\frac{r}{p}\right)^2=\\
&=\frac{r}{p^2}-\frac{r}{p}=\frac{r-r\cdot p}{p^2}=\frac{r\cdot q}{p^2}
\end{aligned}\tag{5.35}$$

Tudi ta rezultat bi lahko dobili z drugačnim razmišljanjem. Na osnovi izraza (5.28) lahko zapišemo:

$$\begin{aligned} &x = n_1 + n_2 + \ldots + n_r \\ &VAR(X) = VAR(N_1 + N_2 + \ldots + N_r) = \\ &= VAR(N_1) + VAR(N_2) + \ldots + VAR(N_r) \end{aligned} \tag{5.36}$$

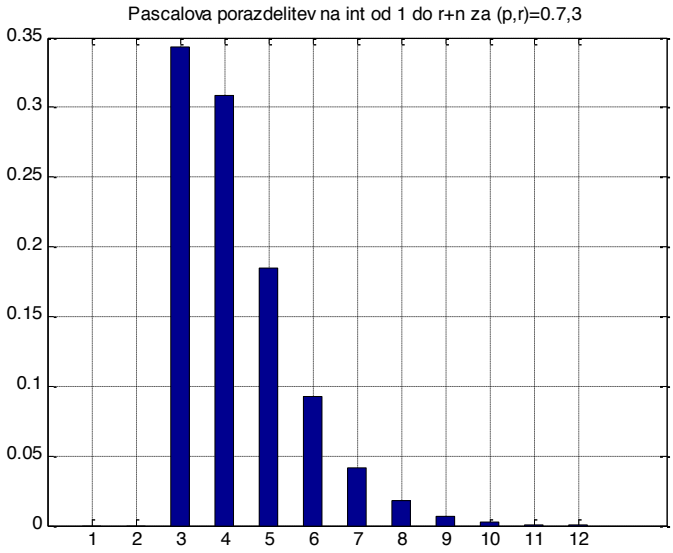
Torej, ker je Pascalova naključna spremenljivka X vsota geometrijskih naključnih spremenljivk N_i, lahko uporabimo izraz (5.21) pri izračunu variance. Kot se izkaže, za varianco geometrijske naključne spremenljivke velja [Turk, Krishnamoorthy]:

$$VAR(N) = \frac{1-p}{p^2} \tag{5.37}$$

Če to upoštevamo, dobimo:

$$VAR(X) = \frac{1-p}{p^2} + \frac{1-p}{p^2} + \ldots + \frac{1-p}{p^2} = r\left(\frac{1-p}{p^2}\right) = \frac{r \cdot q}{p^2} \tag{5.38}$$

V nadaljevanju si poglejmo, kakšna so Pascalove porazdelitve za štiri različne vrednosti *r* in *p*. Prikazujejo jih slike 108-111.



Slika 108: Pascalova porazdelitev za p = 0.7, r = 3, r+n = 12

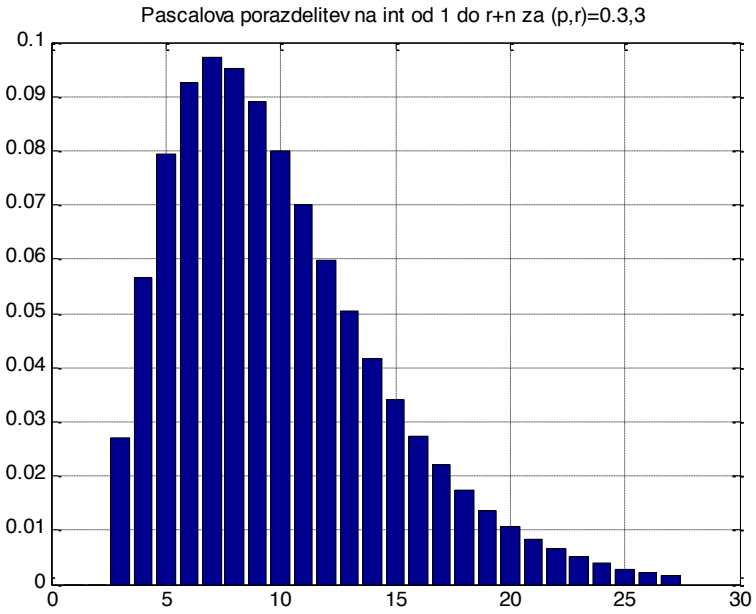
Pri izrisu slike 108 smo si pomagali z naslednjim programom v Matlabu:

```
% pascalova porazdelitev: pascal.m

clear
clc
close all

p = input('Vnesi p')
n = input('Za kako velik n naj pokazem plot?')
r = input('Vnesi r')
q = 1 - p
x = r:1:r+n

warning off

P(1:r-1)=0;

for i=x
    P(i) = nchoosek(i-1,r-1)*p^r*q^(i-r);
end

disp('Pascalova porazdelitev na int. od 1 do r+n je:')
P

if n/20 < 1
   dx = n/20;
else
   dx = 0.8;
end

bar(1:r+n,P,dx)

grid

astr = ['Pascalova porazdelitev na int od 1 do r+n za (p,r)=' num2str(p) ',' num2str(r)];
title(astr)

warning on

disp('Pascalova porazdelitev na int. od 1 do r+n s standardnimi matlab ukazi:')

P1(1:r-1)=0;
K1 = nbinpdf(0:n,r,p);
P1 =[P1 K1]

ch = input('Zelis dodatne izracune 1-DA,0-NE');
if ch == 0
   break
end

disp('Matematicno upanje je:')

MU = r/p

%-------------------------------------------------------------------------
% dodatni izracuni:
%-------------------------------------------------------------------------

ch = input('Zelis izracun za eno verjetnost 1-Da,0-Ne');
if ch == 1
   i = input('i=?')
   Pi = nchoosek(i-1,r-1)*p^r*q^(i-r)
   break
end


xmax = input('Vnesi xmax, do koder naj se tvori vsota verjetnosti')

x = r:1:xmax

clear P
P = 0;
```

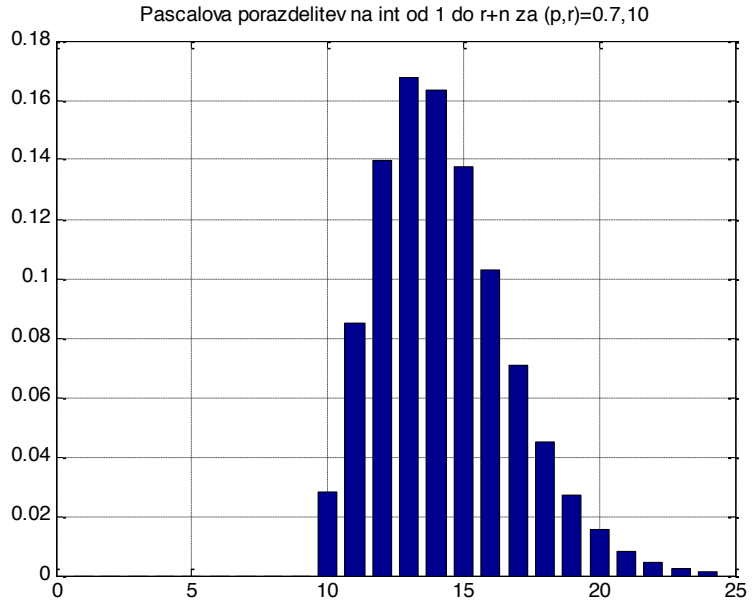
```
for i=x
    P = P + nchoosek(i-1,r-1)*p^r*q^(i-r);
end

disp('Vsota verjetnosti na int. od r do xmax je:')

P

disp('Vsota verjetnosti na int. od r do xmax z matlab ukazom nbincdf je:')

P1z = nbincdf(0:xmax,r,p);
P1 = P1z(xmax-r+1)

disp('Vrednosti kumulativne funkcije na int. od 1 do xmax+r:')

P2z(1:r-1)=0;
P2z = [P2z P1z]

% Izris kumulativne funkcije:
figure
plot(P2z)
hold on
plot(P2z,'o')
grid
title('kumulativna funkcija na int. od 1 do xmax+r')
```

Izgled komandnega okna za sliko 108 je naslednji:

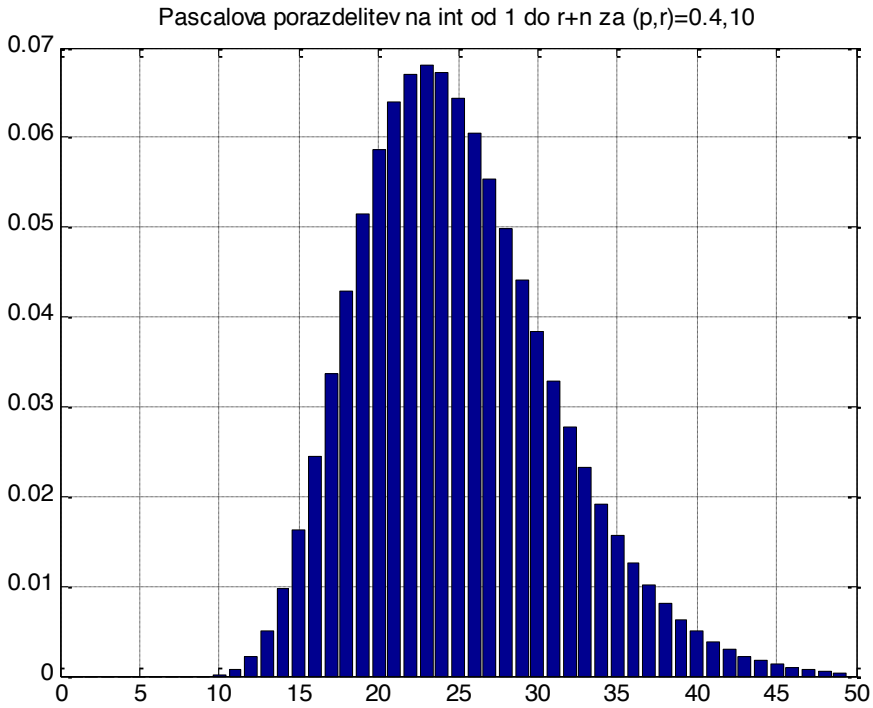
```
Vnesi p0.7
p =
    0.7000
Za kako velik n naj pokazem plot?9
n =
     9
Vnesi r3
r =
     3
q =
    0.3000
x =
     3     4     5     6     7     8     9    10    11    12

Pascalova porazdelitev na int. od 1 do r+n je:
P =
  Columns 1 through 11

0         0    0.3430    0.3087    0.1852    0.0926    0.0417    0.0175    0.0070
0.0027    0.0010

  Column 12

    0.0004

Pascalova porazdelitev na int. od 1 do r+n s standardnimi matlab ukazi:
P1 =
  Columns 1 through 11

0         0    0.3430    0.3087    0.1852    0.0926    0.0417    0.0175    0.0070
0.0027    0.0010

  Column 12

    0.0004

Zelis dodatne izracune 1-DA,0-NE0
```



Slika 109: Pascalova porazdelitev za p = 0.3, r = 3, r+n = 27



Slika 110: Pascalova porazdelitev za p = 0.7, r = 10, r+n = 24



Slika 111: Pascalova porazdelitev za p = 0.4, r = 10, r+n = 49

Primer 5.3..

Iz preteklih opazovanj vemo, da 40% študentov, ki pridejo na izpit, le-tega tudi opravi. Profesor se je nekega dne odločil, da bo izpraševal tako dolgo, dokler ne bodo izpita opravili 4 študentje, Kakšna je verjetnost, da bo izprašal natanko 7 študentov? Koliko študentov pa mora profesor v povprečju izprašati, da bodo štirje opravili izpit?

Na osnovi podatkov naloge lahko postavimo:

$p = 0.4$
$q = 1 - p = 0.6$
$r = 4$
$P(X = 7) = ?$
$E(X) = ?$

Dobimo:

$$P(X=7)=\binom{x-1}{4-1}\cdot(1-p)^{x-4}\cdot p^4=\binom{x-1}{3}\cdot(0.6)^{x-4}\cdot 0.4^4=$$
$$=\binom{7-1}{3}\cdot(0.6)^{7-4}\cdot 0.0256=\binom{6}{3}\cdot(0.6)^3\cdot 0.0256= \tag{5.39}$$
$$=\binom{6}{3}\cdot 0.216\cdot 0.0256=20\cdot 0.216\cdot 0.0256=0.1106$$

Torej je verjetnost, da bo moral profesor vprašati sedem študentov, da bodo štirje opravili izpit, enaka 0.1106.

Za matematično upanje velja:

$$E(X)=\frac{r}{p}=\frac{4}{4/10}=10 \tag{5.40}$$

Deset študentov mora profesor v povprečju izprašati, da bodo štirje opravili izpit.

Izgled komandnega okna za ta izračun je naslednji:

```
Vnesi p0.4
p =
   0.4000
Za kako velik n naj pokazem plot?3
n =
     3
Vnesi r4
r =
     4
q =
   0.6000
x =
    4    5    6    7

Pascalova porazdelitev na int. od 1 do r+n je:
P =
        0         0         0    0.0256    0.0614    0.0922    0.1106

Pascalova porazdelitev na int. od 1 do r+n s standardnimi matlab ukazi:
P1 =

        0         0         0    0.0256    0.0614    0.0922    0.1106

Zelis dodatne izracune 1-DA,0-NE1
Matematicno upanje je:
MU =
    10

Zelis izracun za eno verjetnost 1-Da,0-Ne1
i=?7
i =
     7
Pi =
   0.1106
```

Primer 5.4.:

Izračunajte pričakovano število metov kocke, dokler se šestica ne bo pojavila trikrat. Izračunajte tudi verjetnost, da so potrebni največ štirje meti, da bi padla šestica trikrat.

Na osnovi podatkov naloge lahko postavimo:

$$p = \frac{1}{6}$$
$$q = 1 - p = \frac{5}{6}$$
$$r = 3$$
$$P(X \le 4) = ?$$
$$E(X) = ?$$

Dobimo:

$$\begin{aligned} & P(X \le 4) = P(X = 3) + P(X = 4) = \\ & = \binom{3-1}{3-1} \cdot \left(\frac{5}{6}\right)^{3-3} \cdot \left(\frac{1}{6}\right)^{3} + \binom{4-1}{3-1} \cdot \left(\frac{5}{6}\right)^{4-3} \cdot \left(\frac{1}{6}\right)^{3} = \\ & = \left(\frac{1}{6}\right)^{3} + \binom{3}{2} \cdot \left(\frac{5}{6}\right)^{1} \cdot \left(\frac{1}{6}\right)^{3} = \left(\frac{1}{6}\right)^{3} \left(1 + 3 \cdot \left(\frac{5}{6}\right)\right) = \\ & = \left(\frac{1}{6}\right)^{3} \left(\frac{21}{6}\right) = \frac{21}{6^4} = 0.0162 \end{aligned} \tag{5.41}$$

Torej je verjetnost, da so potrebni največ štirje meti, da bi padla šestica trikrat, enaka 0.0162.

Za matematično upanje velja:

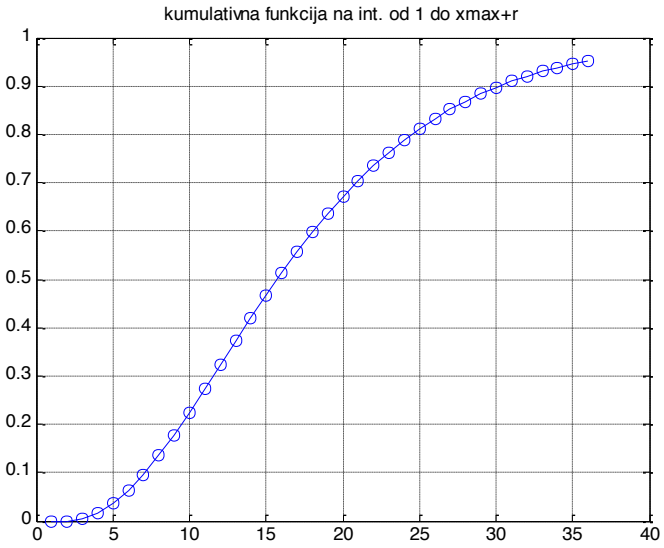
$$E(X) = \frac{r}{p} = \frac{3}{1/6} = 18 \tag{5.42}$$

Pričakovano število metov kocke, dokler se šestica ne bo pojavila trikrat, je enako 18.

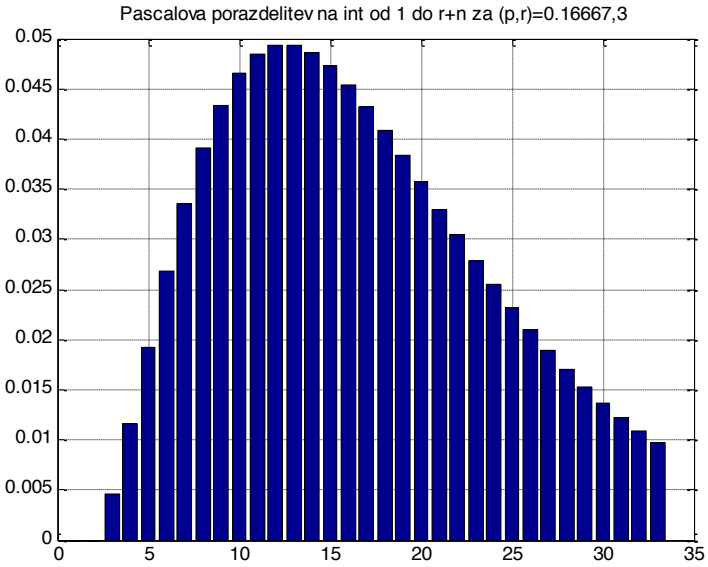
Izgled komandnega okna je naslednji:

```
Vnesi p1/6
p =
    0.1667
Za kako velik n naj pokazem plot?1
n =
     1
Vnesi r3
r =
     3
q =
    0.8333
x =
     3     4

Pascalova porazdelitev na int. od 1 do r+n je:
P =
         0         0    0.0046    0.0116

Pascalova porazdelitev na int. od 1 do r+n s standardnimi matlab ukazi:
P1 =

         0         0    0.0046    0.0116

Zelis dodatne izracune 1-DA,0-NE1
Matematicno upanje je:
MU =
    18

Zelis izracun za eno verjetnost 1-Da,0-Ne0
Vnesi xmax, do koder naj se tvori vsota verjetnosti4
xmax =
     4

x =
     3     4

Vsota verjetnosti na int. od r do xmax je:
P =
    0.0162

Vsota verjetnosti na int. od r do xmax z matlab ukazom nbincdf je:
P1 =
    0.0162

Vrednosti kumulativne funkcije na int. od 1 do xmax+r:
P2z =
         0         0    0.0046    0.0162    0.0355    0.0623    0.0958
```

Podobno kot pri geometrijski porazdelitvi tudi tukaj velja, da z naraščanjem števila neuspelih poskusov raste verjetnost, da se bo le zgodilo predpisano število uspešnih dogodkov. Tudi tu lahko ta fenomen merimo s pomočjo kumulativne funkcije. Slika 112 prikazuje kumulativno funkcijo, če kocko vržemo 36 krat, pri čemer ima verjetnostna funkcija obliko prikazano na sliki 113.



Slika 112: Kumulativna funkcija, če kocko vržemo 36 krat



Slika 113: Pascalova porazdelitev pri p = 1/6, r = 3, r+n = 33

Kot je razvidno iz slike 113, najprej verjetnost, da se bo pri x-tem poskusu realiziralo predpisano število uspešnih dogodkov, raste. Ko pa poskuse izvedemo že precejkrat, začne ta verjetnost padati. Pri zelo velikem številu poskusov namreč verjetnost, da se pri x-tem poskusu zgodi (doseže) predpisano število uspešnih dogodkov, praktično pade na 0, saj se je to skoraj gotovo zgodilo že prej.

Poglejmo si še komandno okno, ki je generiralo sliki 113 in 114:

```
Vnesi p1/6
p =
    0.1667
Za kako velik n naj pokazem plot?30
n =
    30
Vnesi r3
r =
     3
q =
    0.8333
x =
  Columns 1 through 19
     3     4     5     6     7     8     9    10    11    12    13    14    15    16    17    18    19    20    21
  Columns 20 through 31
    22    23    24    25    26    27    28    29    30    31    32    33

Pascalova porazdelitev na int. od 1 do r+n je:
P =
  Columns 1 through 11
         0         0    0.0046    0.0116    0.0193    0.0268    0.0335    0.0391    0.0434    0.0465    0.0485
  Columns 12 through 22
    0.0493    0.0493    0.0486    0.0473    0.0454    0.0433    0.0409    0.0383    0.0357    0.0330    0.0304
  Columns 23 through 33
    0.0279    0.0255    0.0231    0.0210    0.0189    0.0170    0.0153    0.0137    0.0122    0.0109    0.0097

Pascalova porazdelitev na int. od 1 do r+n s standardnimi matlab ukazi:
P1 =
  Columns 1 through 11
         0         0    0.0046    0.0116    0.0193    0.0268    0.0335    0.0391    0.0434    0.0465    0.0485
  Columns 12 through 22
    0.0493    0.0493    0.0486    0.0473    0.0454    0.0433    0.0409    0.0383    0.0357    0.0330    0.0304
  Columns 23 through 33
    0.0279    0.0255    0.0231    0.0210    0.0189    0.0170    0.0153    0.0137    0.0122    0.0109    0.0097

Zelis dodatne izracune 1-DA,0-NE1
Matematicno upanje je:
MU =
    18
Zelis izracun za eno verjetnost 1-Da,0-Ne0
Vnesi xmax, do koder naj se tvori vsota verjetnosti33
xmax =
    33
x =
  Columns 1 through 18
     3     4     5     6     7     8     9    10    11    12    13    14    15    16    17    18    19    20
  Columns 19 through 31
    21    22    23    24    25    26    27    28    29    30    31    32    33
```

```
Vsota verjetnosti na int. od r do xmax je:
P =
   0.9300
Vsota verjetnosti na int. od r do xmax z matlab ukazom nbincdf je:
P1 =
   0.9300

Vrednosti kumulativne funkcije na int. od 1 do xmax+r:
P2z =
 Columns 1 through 11
      0       0   0.0046   0.0162   0.0355   0.0623   0.0958   0.1348   0.1783   0.2248   0.2732
 Columns 12 through 22
   0.3226   0.3719   0.4205   0.4678   0.5132   0.5565   0.5973   0.6357   0.6713   0.7044   0.7348
 Columns 23 through 33
   0.7627   0.7882   0.8113   0.8323   0.8512   0.8682   0.8835   0.8972   0.9094   0.9203   0.9300
 Columns 34 through 36
   0.9386   0.9462   0.9529
```

5.1.4 Hipergeometrična porazdelitev

Denimo imamo N elementov, med katerimi jih ima M določeno lastnost, N - M pa te lastnosti nima. Na slepo izberemo brez vračanja n elementov izmed N elementov. Da bo imelo dano lastnost natanko x elementov izmed n izbranih na slepo, obstaja točno $\binom{M}{x}$ načinov. Da ne bo imelo dano lastnost natanko n - x elementov izmed n izbranih na slepo, pa obstaja točno $\binom{N-M}{n-x}$ načinov. Da bo imelo dano lastnost natanko x elementov izmed n izbranih na slepo, in hkrati ne bo imelo dano lastnost natanko n - x elementov izmed n izbranih na slepo, obstaja točno $\binom{M}{x}\binom{N-M}{n-x}$ načinov. Vseh možnih načinov izbora n elementov izmed N elementov pa je $\binom{N}{n}$. Verjetnost, da bo imelo dano lastnost natanko x elementov izmed n izbranih na slepo, in hkrati ne bo imelo dano lastnost natanko n - x elementov izmed n izbranih na slepo, je enaka (Hipergeometrična porazdelitev $Hip(N,M,n)$) [Jesenko, Turk, Jurišić]:

$$P(X=x)=\frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}},\quad x=0,1,2,...,n \qquad (5.42)$$
$$x\le M,\quad n-x\le N-M$$

V nadaljevanju izračunajmo matematično upanje za hipergeometrično porazdelitev. V ta namen tvorimo (upoštevamo, da je prvi člen v vsakem primeru 0 in se vrsta začne z 1):

$$\mu=E(X)=\sum_{x=1}^{n}x\cdot P(x)=\sum_{x=1}^{n}x\cdot\frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}=$$
$$=\sum_{x=1}^{n}x\cdot\frac{\frac{M!}{x!\cdot(M-x)!}\binom{N-M}{n-x}}{\binom{N}{n}}=\frac{M}{M}\sum_{x=1}^{n}\frac{\frac{M!}{(x-1)!\cdot(M-x)!}\binom{N-M}{n-x}}{\binom{N}{n}}=$$
$$=M\sum_{x=1}^{n}\frac{\frac{(M-1)!}{(x-1)!\cdot(M-x)!}\binom{N-M}{n-x}}{\binom{N}{n}}=M\sum_{x=1}^{n}\frac{\binom{M-1}{x-1}\binom{N-M}{n-x}}{\binom{N}{n}}= \qquad (5.43)$$
$$=M\sum_{x=1}^{n}\frac{\binom{M-1}{x-1}\binom{N-M}{n-x}}{\frac{N!}{n!\cdot(N-n)!}\cdot\frac{Nn}{Nn}}=M\sum_{x=1}^{n}\frac{\binom{M-1}{x-1}\binom{N-M}{n-x}}{\frac{(N-1)!}{(n-1)!\cdot(N-n)!}\cdot\frac{N}{n}}=$$
$$=\frac{n\cdot M}{N}\sum_{x=1}^{n}\frac{\binom{M-1}{x-1}\binom{N-M}{n-x}}{\binom{N-1}{n-1}}$$

Vpeljemo novi spremenljivki $y = x-1, m = n-1$ in dobimo:

$$\mu = \frac{n\cdot M}{N}\sum_{y=o}^{n-1}\frac{\binom{M-1}{y}\binom{N-M}{m+1-y-1}}{\binom{N-1}{m}} =$$

$$= \frac{n\cdot M}{N}\sum_{y=o}^{m}\frac{\binom{M-1}{y}\binom{N-M}{m-y}}{\binom{N-1}{m}}, \qquad (5.44)$$

$y+1 \le M \quad oz.\ y \le M-1 \quad in$

$m+1-y-1 \le N-M \quad oz.\ m-y \le (N-1)-(M-1)$

Sledi :

$$\mu = \frac{n\cdot M}{N}\underbrace{\left[P(y=0)+P(y=1)+\ldots+P(y=m)\right]}_{1} = \frac{n\cdot M}{N}$$

Na podoben način (in s podobnimi prijemi kot pri Pascalovi porazdelitvi) bi dobili tudi varianco, ki se glasi [Jesenko, Krishnamoorthy]:

$$\sigma^2 = VAR(X) = \frac{n\cdot M(N-M)(N-n)}{N^2(N-1)} =$$

$$= \frac{n\cdot M}{N}\cdot\frac{(N-M)(N-n)}{N(N-1)} = \frac{n\cdot M}{N}\cdot\frac{N\left(1-\frac{M}{N}\right)(N-n)}{N(N-1)} = \qquad (5.45)$$

$$= n\cdot\left(\frac{M}{N}\right)\cdot\left(1-\frac{M}{N}\right)\cdot\frac{(N-n)}{(N-1)}$$

Primer 5.5.:

Varnostni inšpektor se je odločil, da bo v nekem podjetju, ki ima 30 strojev, na slepo izbral 5 strojev in jih temeljito pogledal. V podjetju pa je, kot se izkaže, 8 strojev takšnih, ki ne ustrezajo varnostnim normam. Inšpektor je sklenil, da bo pregledal vse stroje, če bo vsaj eden od 5 izbranih neustrezen. Kakšna je verjetnost, da s pregledom ne bo nadaljeval?Koliko varnostno neustreznih strojev lahko pričakujemo med 5 pregledanimi?

Na osnovi podatkov naloge lahko zapišemo:

$N = 30$

$M = 8$

$n = 5$

$P(vseh\ \ pet\ ustreznih) = P(X = 0) = ?$

$E(X) = ?$

Dobimo:

$$P(X=0) = \frac{\binom{8}{0}\binom{30-8}{5-0}}{\binom{30}{5}} = \frac{\binom{8}{0}\binom{22}{5}}{\binom{30}{5}} = 0.1848 \qquad (5.46)$$

ter:

$$E(X) = \frac{n \cdot M}{N} = \frac{5 \cdot 8}{30} = \frac{4}{3} = 1.33 \qquad (5.46)$$

Torej je 18.48% verjetnosti, da s pregledom ne bo nadaljeval. Med pregledanimi stroji pa bo v povprečju 1.33 varnostno neustreznih.

Primer 5.6.:

Proizvajalec ima na zalogi 300 rezervnih delov, od katerih jih je 100 od lokalnega dobavitelja. Če so štirje deli izbrani naključno, kakšna je verjetnost, da so vsi od lokalnega dobavitelja? Kakšna je verjetnost, da sta vsaj dva dela od lokalnega dobavitelja? Kakšna je verjetnost, da je vsaj en del od lokalnega dobavitelja [Montgomery 1]? Izrišite tudi Hipergeometrično porazdelitev tega primera.

Na osnovi podatkov naloge lahko zapišemo:

$N = 300$
$M = 100$
$n = 4$
$P(X = 4) = ?$
$P(X \geq 2) = ?$
$P(X \geq 1) = ?$
$E(X) = ?$
$VAR(X) = ?$

Dobimo:

$$P(X = 4) = \frac{\binom{100}{4}\binom{300-100}{4-4}}{\binom{300}{4}} = \frac{\binom{100}{4}\binom{200}{0}}{\binom{300}{4}} = 0.0119 \tag{5.47}$$

ter

$$\begin{aligned} P(X \geq 2) &= P(X = 2) + P(X = 3) + P(X = 4) = \\ &= \frac{\binom{100}{2}\binom{300-100}{4-2}}{\binom{300}{4}} + \frac{\binom{100}{3}\binom{300-100}{4-3}}{\binom{300}{4}} + \frac{\binom{100}{4}\binom{300-100}{4-4}}{\binom{300}{4}} = \\ &= \frac{\binom{100}{2}\binom{200}{2}}{\binom{300}{4}} + \frac{\binom{100}{3}\binom{200}{1}}{\binom{300}{4}} + \frac{\binom{100}{4}\binom{200}{0}}{\binom{300}{4}} = 0.2978 + 0.0978 + 0.0119 = \\ &= 0.4075 \end{aligned} \tag{5.48}$$

in

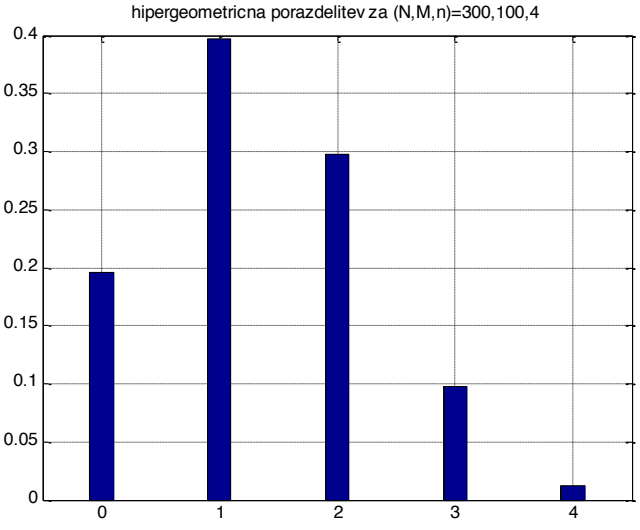
$$\begin{aligned} P(X \geq 1) &= P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) = \\ &= 1 - P(X = 0) = 1 - \frac{\binom{100}{0}\binom{300-100}{4-0}}{\binom{300}{4}} = 1 - \frac{\binom{100}{0}\binom{200}{4}}{\binom{300}{4}} = \\ &= 1 - 0.1955 = 0.8045 \end{aligned} \tag{5.49}$$

Matematično upanje in varianca sta enaka:

$$E(X)=\frac{n\cdot M}{N}=\frac{4\cdot 100}{300}=\frac{4}{3}=1.33 \tag{5.50}$$

$$\begin{aligned}\sigma^2=VAR(X)&=n\cdot\left(\frac{M}{N}\right)\cdot\left(1-\frac{M}{N}\right)\cdot\frac{(N-n)}{(N-1)}=\\&=4\cdot\left(\frac{100}{300}\right)\cdot\left(1-\frac{100}{300}\right)\cdot\frac{(300-4)}{(300-1)}=4\cdot\frac{1}{3}\cdot\frac{2}{3}\cdot\frac{296}{299}=0.88\end{aligned} \tag{5.51}$$

Izris hipergeometrične porazdelitve za ta primer prikazuje slika 114.



Slika 114: Izris hipergeometrične porazdelitve za primer rezervnih delov

Pri izrisu slike 114 smo si pomagali z naslednjim programom v Matlabu:

```
% hipergeometrična porazdelitev: hipergeo.m

clear
clc
close all

N = input('Vnesi N')
M = input('Vnesi M')
n = input('Vnesi n')

if n == 0
   disp('n mora biti vecji od 0')
   break
end

ch = input('izracun samo za en x(1) ali vec x (2)')
if ch == 1
   x = input('x=')
   if x>n
      disp('x ne sme biti vecji od n')
      break
   end
else
   x = 0:1:n
end

warning off

if length(x) > 1
   for i=x
       P(i+1) = nchoosek(M,i)*nchoosek(N-M,n-i)/nchoosek(N,n);
   end
else
   P = nchoosek(M,x)*nchoosek(N-M,n-x)/nchoosek(N,n);
end

disp('Hipergeometrična porazdelitev je:')
P

disp('Hipergeometricna porazdelitev s standardnim matlab ukazom je:')
P1 = hygepdf(x,N,M,n)


if ch == 1
   break
end

if n/20 < 1
   dx = n/20;
else
   dx = 0.8;
end

bar(x,P,dx)

grid

astr = ['hipergeometricna porazdelitev za (N,M,n)=' num2str(N) ',' num2str(M) ','
num2str(n) ];
title(astr)

warning on
```

Pri tem je bil izpis komandnega okna naslednji:

```
Vnesi N300
N =
  300
Vnesi M100
M =
  100
Vnesi n4
n =
    4
izracun samo za en x(1) ali vec x (2)2
ch =
    2
x =
    0    1    2    3    4

Hipergeometrična porazdelitev je:
P =
   0.1955    0.3970    0.2978    0.0978    0.0119

Hipergeometricna porazdelitev s standardnim matlab ukazom je:
P1 =
   0.1955    0.3970    0.2978    0.0978    0.0119
```

Do rezultata (5.48) bi prišli v Matlabu tudi s pomočjo kumulativne funkcije:

```
>> P = 1 - hygecdf([1],300,100,4)
P =
   0.4074
```

Do rezultata (5.49) bi prišli v Matlabu tudi s pomočjo kumulativne funkcije:

```
>> P = 1 - hygecdf([0],300,100,4)
P =
   0.8045
```

Kadar velja, da so *N, M* in *N – M* veliki v primerjavi z *n*, potem Hipergeometrična porazdelitev gravitira k binomski in velja [Čuljak]:

$$Hip\left(N,M,n\right) \approx Bin\left(n, p = \frac{M}{N}\right) \tag{5.52}$$

Primer 5.7.:

Primerjajte hipergeometrično in binomsko porazdelitev za [Montgomery 1]:

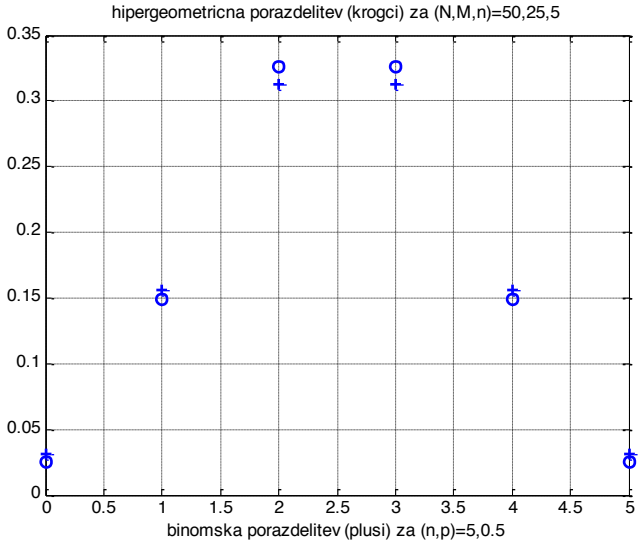
$N = 50$

$M = 25$

$n = 5$

$p \approx \frac{M}{N} = 0.5$

Slika 115 prikazuje primerjavo obeh porazdelitev.



Slika 115: Primerjava hipergeometrične in binomske porazdelitve za N = 50, M = 25, n = 5

Za izris slike 115 smo si pomagali z naslednjim programom v Matlabu:

```
% hipergeometrična porazdelitev: hipergeo1.m

clear
clc
close all

warning off

N = input('Vnesi N')
M = input('Vnesi M')
n = input('Vnesi n')

p = M/N
q=1-p

if n == 0
   disp('n mora biti vecji od 0')
   break
end

x = 0:1:n
for i=x
    P(i+1) = nchoosek(M,i)*nchoosek(N-M,n-i)/nchoosek(N,n);
    P1(i+1) = nchoosek(n,i)*p^i*q^(n-i);
```

```
end


disp('Hipergeometrična porazdelitev je:')
P

disp('Binomska porazdelitev je:')
P1

plot(x,P,'o','LineWidth',2)
hold on
plot(x,P1,'+','LineWidth',2)

grid

astr = ['hipergeometricna porazdelitev (krogci) za (N,M,n)=' num2str(N) ',' num2str(M) ','
num2str(n) ];
title(astr)

astr = ['binomska porazdelitev (plusi) za (n,p)=' num2str(n) ',' num2str(p)];
xlabel(astr)

warning on
```

Komandno okno ima naslednji izgled:

```
Vnesi N50
N =
   50
Vnesi M25
M =
   25
Vnesi n5
n =
    5
p =
   0.5000
q =
   0.5000

x =

    0    1    2    3    4    5

Hipergeometrična porazdelitev je:

P =

   0.0251   0.1493   0.3257   0.3257   0.1493   0.0251

Binomska porazdelitev je:

P1 =

   0.0313   0.1563   0.3125   0.3125   0.1563   0.0313
```

Kot vidimo, se v tem primeru hipergeometrična in binomska porazdelitev približno ujemata.

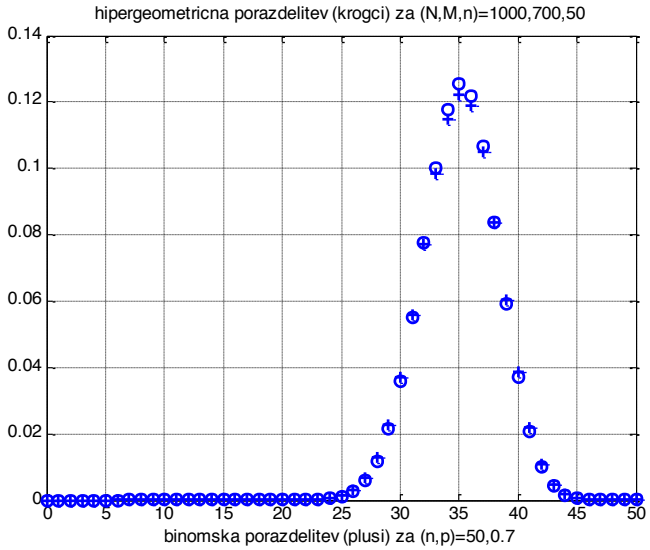
Primer 5.8.:

Spisek računov strank pri velikem podjetju vsebuje 1000 strank. Izmed teh jih je 700 naročilo vsaj enega od proizvodov podjetja v zadnjih 3 mesecih. Da bi se opravila evalvacija načrtovanja novega proizvoda, je 50 strank naključno izbranih iz spiska. Kakšna je verjetnost, da je več kot 45 vzorčenih strank naročilo nov izdelek v zadnjih treh mesecih [Montgomery 1]. Primerjajte binomsko in hipergeometrično porazdelitev. Poiščite tudi matematično upanje.

Na osnovi podatkov naloge velja:

$N = 1000$
$M = 700$
$n = 50$
$p \approx \frac{M}{N} = 0.7$
$P(X > 45) = ?$
$E(X) = ?$

Slika 116 prikazuje primerjavo obeh porazdelitev. Pri tem smo ponovno uporabili program **hipergeo1.m**.



Slika 116: Primerjava hipergeometrične in binomske porazdelitve za N = 1000, M = 700, n = 50

$$E(X) = \frac{n \cdot M}{N} = \frac{50 \cdot 700}{1000} = 35 \tag{5.53}$$

Torej se v vzorcu 50 strank v povprečju pričakuje, da jih je 35 kupilo nov izdelek v zadnjih 3 mesecih. Kot vidimo, se vrh obeh porazdelitev na sliki 116 ujema s pričakovano vrednostjo. Iz slike 116 vidimo tudi, da se obe porazdelitvi izredno dobro ujemata.

Verjetnost $P(X > 45)$ lahko izračunamo na dva načina, preko hipergeometrične ali preko binomske porazdelitve. Če jo računamo preko binomske porazdelitve, velja [Montgomery 1]:

$$P(X > 45) \approx \sum_{x=46}^{50} \binom{50}{x} \cdot 0.7^x \cdot 0.3^{50-x} = 0.00017 \tag{5.54}$$

Torej je izredno majhna verjetnost, da je več kot 45 vzorčenih strank naročilo nov izdelek v zadnjih treh mesecih.

5.1.5 Poissonova porazdelitev

Poissonove porazdelitve smo se nekoliko dotaknili že v poglavju 2.1.1 (glej izraz (2.5)), v poglavju 2.10.1 (glej izraze (2.66) do (2.70)), kjer smo izračunali matematično upanje $E(X) = \lambda$, drugi moment $E(X^2) = \lambda^2 + \lambda$, ter varianco $VAR(X) = \lambda$, ter v poglavju 2.21 (glej izraze (2.211) do (2.215)), kjer smo tudi izračunali rodovno funkcijo momentov $M(t) = e^{\lambda \cdot (e^t - 1)}$.

V nadaljevanju bomo poskušali izpeljati Poissonovo porazdelitev. Pri tem si pomagamo z razmišljanjem iz prometa [Jurišić]. Denimo imamo strokovnjake za promet, ki opazujejo neko križišče. Radi bi napovedali, kakšna je verjetnost, da v danem časovnem intervalu skozi križišče zapelje npr. 100 avtomobilov. Pri tem npr. definirajo naključno spremenljivko X = *število avtov, ki pridejo v križišče na uro*, ter v modelu privzamejo še 2 predpostavki: vsaka ura je enaka kot vsaka druga ura (čeprav to ne drži, saj je npr.

gostota prometa v konici večja kot ponoči), ter, da so dogodki prihodov avtov v posameznih časovnih intervalih med seboj neodvisni.

Za začetek štejejo promet in tako dobijo oceno za pričakovano število avtov na časovno enoto: $E(X)=\lambda$, npr. 9 avtov na uro. Nato poskusijo v modelu uporabiti binomsko porazdelitev. To je povsem upravičeno, saj je Poissonova porazdelitev po pomenu enaka binomski, saj tudi predstavlja število uspehov [Turk]. Za binomsko porazdelitev velja (glej izraz (5.8)): $E(X)=n\cdot p$, pri čemer je n število ponovitev poskusa, en poskus predstavlja opazovanje v določenem časovnem intervalu Δt, p pa je verjetnost, da je v dani časovni enoti Δt prišel mimo vsaj en avto (uspešen dogodek). Če izenačimo obe pričakovani vrednosti, dobimo:

$$\begin{aligned} E(X) &= \lambda = n\cdot p \\ p &= \frac{\lambda}{n} \end{aligned} \qquad (5.55)$$

Če je recimo opazovani časovni interval 1 ura (60 minut) in časovna enota $\Delta t = 1\ \text{min}$, potem je $n = 60$ in sledi za binomsko porazdelitev:

$$P(X=k)=P_{60}(k)=\binom{60}{k}\cdot\left(\frac{\lambda}{60}\right)^{k}\cdot\left(1-\frac{\lambda}{60}\right)^{60-k} \qquad (5.56)$$

Vendar pa lahko v tem primeru naletijo na problem, da gre v časovni enoti $\Delta t = 1\ \text{min}$ v križišče več kot en avto. Potem je potrebno časovno enoto zmanjšati npr. na sekunde: $\Delta t = 60\ sek$. Če opazovani časovni interval še vedno 1 ura (3600 sek), potem je $n = 3600$ in sledi za binomsko porazdelitev:

$$P(X=k)=P_{3600}(k)=\binom{3600}{k}\cdot\left(\frac{\lambda}{3600}\right)^{k}\cdot\left(1-\frac{\lambda}{3600}\right)^{3600-k} \qquad (5.57)$$

Če tudi to ni dovolj, vzamejo še manjšo časovno enoto. Kaj pa se zgodi, če bi šel *n* proti neskončno?

Potem bi imeli [Jurišić, Turk, Jesenko]:

$$\begin{aligned}
&\lim_{n\to\infty} P_n(k) = \lim_{n\to\infty}\left[\binom{n}{k}\cdot\left(\frac{\lambda}{n}\right)^k\cdot\left(1-\frac{\lambda}{n}\right)^{n-k}\right] = \lim_{n\to\infty}\left[\frac{n!}{(n-k)!\,k!}\cdot\left(\frac{\lambda}{n}\right)^k\cdot\left(1-\frac{\lambda}{n}\right)^{n-k}\right] = \\
&= \frac{\lambda^k}{k!}\lim_{n\to\infty}\left[\frac{n!}{(n-k)!}\cdot\frac{1}{n^k}\cdot\left(1-\frac{\lambda}{n}\right)^{n-k}\right] = \\
&= \frac{\lambda^k}{k!}\lim_{n\to\infty}\left[\frac{1\cdot 2\cdot\ldots\cdot n}{1\cdot 2\cdot\ldots(n-k)}\cdot\frac{1}{n^k}\cdot\left(1-\frac{\lambda}{n}\right)^{n-k}\right] = \\
&= \frac{\lambda^k}{k!}\lim_{n\to\infty}\left[\frac{(n-k+1)\cdot\ldots\cdot n}{n^k}\cdot\left(1-\frac{\lambda}{n}\right)^{n-k}\right] = \\
&= \frac{\lambda^k}{k!}\lim_{n\to\infty}\left[\frac{(n-k+1)}{n}\cdot\frac{(n-k+2)}{n}\cdot\ldots\cdot\frac{n}{n}\cdot\left(1-\frac{\lambda}{n}\right)^{n-k}\right] = \\
&= \frac{\lambda^k}{k!}\lim_{n\to\infty}\left[\frac{n\left(1-\frac{k-1}{n}\right)}{n}\cdot\frac{n\left(1-\frac{k-2}{n}\right)}{n}\cdot\ldots\cdot 1\cdot\left(1-\frac{\lambda}{n}\right)^{n-k}\right] = \\
&= \frac{\lambda^k}{k!}\lim_{n\to\infty}\left[\left(1-\frac{k-1}{n}\right)\cdot\left(1-\frac{k-2}{n}\right)\cdot\ldots\cdot 1\cdot\left(1-\frac{\lambda}{n}\right)^{n-k}\right] = \\
&= \frac{\lambda^k}{k!}\lim_{n\to\infty}\left[\left(1-\frac{\lambda}{n}\right)^{n-k}\right]
\end{aligned} \tag{5.58}$$

Sledi:

$$\begin{aligned}
&\lim_{n\to\infty} P_n(k) = \frac{\lambda^k}{k!}\lim_{n\to\infty}\left[\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-k}\right] = \\
&= \frac{\lambda^k}{k!}\lim_{n\to\infty}\left[\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-k}\right]
\end{aligned} \tag{5.58}$$

Dokazati se da, da velja [Jurišić]:

$$\lim_{n\to\infty}\left[\left(1-\frac{\lambda}{n}\right)^{-k}\right]=1 \tag{5.59}$$

Ker je po definiciji število *e* enako [Turk]:

$$e=\lim_{m\to\infty}\left[\left(1+\frac{1}{m}\right)^{m}\right] \tag{5.60}$$

sledi (vpeljemo novo spremenljivko $m=-\frac{n}{\lambda}$):

$$\begin{aligned}&\lim_{n\to\infty}P_n(k)=\frac{\lambda^k}{k!}\lim_{n\to\infty}\left[\left(1-\frac{\lambda}{n}\right)^{n}\right]=\frac{\lambda^k}{k!}\lim_{m\to\infty}\left[\left(1-\frac{\lambda}{(-m\cdot\lambda)}\right)^{-m\cdot\lambda}\right]=\\&=\frac{\lambda^k}{k!}\lim_{m\to\infty}\left[\left(1+\frac{1}{m}\right)^{m}\right]^{-\lambda}=\frac{\lambda^k}{k!}[e]^{-\lambda}\end{aligned} \tag{5.61}$$

Tako smo dokazali, da v limiti binomska porazdelitev gravitira k Poissonovi porazdelitvi, ki ima obliko [Jesenko]:

$$P(X=x)=P(x)=\frac{\lambda^x}{x!}e^{-\lambda},\qquad x=0,1,2,3,... \tag{5.62}$$

Poissonov obrazec govori o naslednjem izreku, ki velja za velike *n* in majhne verjetnosti *p* ($n\geq 20,\ p\leq 0.05$) in povezuje Poissonovo in binomsko porazdelitev [Jurišić]:

$$P_n(k)\approx\frac{(n\cdot p)^k}{k!}e^{-n\cdot p},\qquad k=0,1,2,3,... \tag{5.63}$$

Poissonova porazdelitev izraža verjetnost števila dogodkov, ki se zgodijo v danem časovnem (krajevnem, prostorninskem, ...) intervalu, če vemo, da se ti dogodki pojavijo s

poznano povprečno frekvenco in neodvisno od časa, ko se je zgodil zadnji dogodek [Jurišić]. Zapišemo torej lahko [Jurišič]:

$$P\left(Število\ dogodkov = k\right) = \frac{\lambda^k}{k!} e^{-\lambda}, \qquad k = 0,1,2,3,... \tag{5.64}$$

kjer je λ dani parameter in predstavlja pričakovano pogostost nekega dogodka. Poissonova porazdelitev je še posebej pomembna v naslednjih primerih [Jurišić]

- V analizi prometa,
- V teoriji množične strežbe,
- Analiza števila dostopov do omrežnega strežnika na minuto,
- Analiza števila klicev na bazni postaji na minuto,
- Analiza števila mutacij RNK v danem intervalu po določeni količini prejete radiacije, itn.

Pri primerjavi z binomsko porazdelitvijo opozorimo še, da ima Poissonova naključna spmremenljivka neomejeno zalogo vrednosti, pri Bernoullijevi porazdelitvi pa število uspehov seveda ne more preseči števila Bernoullijevih poskusov *n*.

Naštejmo še nekaj primerov naključnih spremenljivk, porazdeljenih po Poissonovi porazdelitvi [Turk]:

- Na osnovi opazovanj v pretekosti, lahko ocenimo pričakovano število poplav λ v določenem časovnem obdobju (npr. 10 let),

- Denimo vemo, da se na nekem cestnem odseku v nekem časovnem obdobju v povprečju zgodi λ nesreč. Število nesreč, ki se zgode na tem odseku in obdobju, je naključna spremenljivka, porazdeljena po Poissonu.

- Število pojavov kritične obremenitve na konstrukcijo je Poissonovo porazdeljena naključna spremenljivka. Parameter porazdelitve λ predstavlja povprečno število pojava kritične obremenitve v nekem časovnem obdobju.

- Poissonovo porazdeljeno spremenljivko lahko uporabimo tudi za primere, kjer opazujemo pojav na nekem krajevnem območju. Npr. obravnavamo porazdelitev sil na neki konstrukciji. Potem je število delujočih sil na konstrukcijo Poissonovo porazdeljena naključna spremenljivka.

Parameter Poissonove porazdelitve lahko v primeru časdovno nespremenljivih razmer definiramo tudi kot linearno funkcijo časa [Turk, Dragan 2]: $\lambda = \vartheta \cdot t$. Potem dobimo:

$$P\left(Y\left(t\right)=y\right)=\frac{\left(\vartheta\cdot t\right)^{y}}{y!}e^{-\vartheta\cdot t},\qquad y=0,1,2,3,... \tag{5.65}$$

V takšnih primerih so naključne spremenljivke funkcija časa in jim pravimo tudi **stohastični (naključni) procesi**. Pri tem pravkar opisanemu procesu pravimo **Poissonov proces** [Turk, Dragan 2]. Več o tej problematiki si lahko bralec ogleda v delu [Dragan 2].

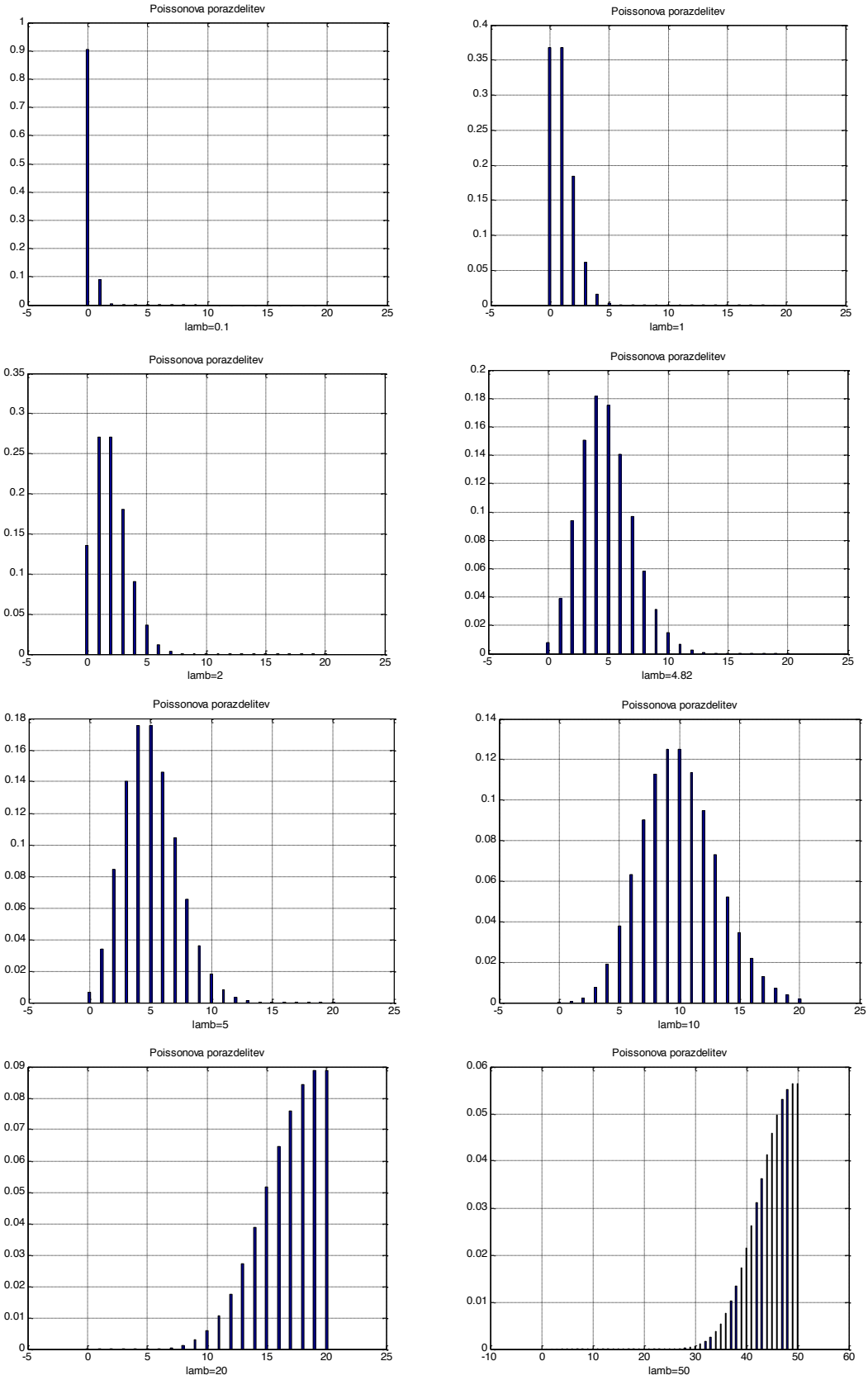
V nadaljevanju si še poglejmo, kakšna sta koeficienta asimetrije in sploščenosti za Poissonovo porazdelitev. Dokazati se da, da dobimo naslednja rezultata [Jesenko]:

$$\begin{aligned} g_1 &= \frac{1}{\sqrt{\lambda}} \\ g_2 &= 3+\frac{1}{\lambda} \\ \gamma &= g_2 - 3 = \frac{1}{\lambda} \end{aligned} \tag{5.66}$$

Izgled Poissonove porazdelitve za:

$$\lambda = 0.1,\ \ \lambda = 1,\ \ \lambda = 2,\ \ \lambda = 4.82,$$
$$\lambda = 5,\ \ \lambda = 10,\ \ \lambda = 20,\ \ \lambda = 50$$

prikazuje slika 117.



Slika 117: Izgled Poissonove porazdelitve za: $\lambda = 0.1,\ \lambda = 1,\ \lambda = 2,\ \lambda = 4.82,$ $\lambda = 5,\ \lambda = 10,\ \lambda = 20,\ \lambda = 50$

Pri izrisu slike 117 smo si pomagali z naslednjim programom v Matlabu:

```
% poisson.m
%

clc
clear
close all

while 1==1

    x = input('x=');
    dx = input('dx=');

    if (length(x) == 0)||(length(dx) == 0)
       dx = 1;
       x = 0:dx:15;
    end

    lamb = input('lamb = ')     % lamb = ?
    y = poisspdf(x,lamb);
    bar(x,y,0.2)
    title('Poissonova porazdelitev')
    xlabel(['lamb=' num2str(lamb)])

    grid

    ch = input('Zelis izhod 1(DA)/0(NE)')
    if ch == 1
       return
    end

    figure


end
```

Primer 5.9.:

Opravka imamo z dogodkom, ki se v povprečju zgodi enkrat v 50 letih, kar pomeni, da je verjetnost, da se takšen dogodek zgodi v posameznem letu, enaka $p \approx \frac{\lambda}{n} = \frac{1}{50}$. *Primeri takšnih dogodkov so: 50-letni snežni metež, 50-letno neurje, 50-leten potres, itn. Določite verjetnost, da se v 50 letih zgodi vsaj en takšen dogodek. Določite tudi verjetnosti, da se zgodijo eden, dva ali trije takšni dogodki v 50 letih. Rešite nalogo s pomočjo binomske in s pomočjo Poissonove porazdelitve [Turk].*

Ker je $n = 50, p = \frac{1}{50}$, je n velik in p majhen, zato velja približna enakost obeh porazdelitev. Najprej rešimo s pomočjo binomske porazdelitve.

$$P(X>0)=1-P(X=0)=1-\binom{50}{0}\left(\frac{1}{50}\right)^{0}\left(1-\frac{1}{50}\right)^{50-0}= \\ =1-\left(\frac{49}{50}\right)^{50}=0.6358 \tag{5.67}$$

Torej je verjetnost, da se v 50 letih zgodi vsaj en takšen dogodek, enaka 0.6358.

Verjetnost, da se zgodi en takšen dogodek, je enaka:

$$P(X=1)=\binom{50}{1}\left(\frac{1}{50}\right)^{1}\left(1-\frac{1}{50}\right)^{50-1}=50\left(\frac{1}{50}\right)\left(\frac{49}{50}\right)^{49}=\left(\frac{49}{50}\right)^{49}=0.3716 \tag{5.68}$$

Verjetnost, da se zgodita dva takšna dogodka, je enaka:

$$P(X=2)=\binom{50}{2}\left(\frac{1}{50}\right)^{2}\left(1-\frac{1}{50}\right)^{50-2}=1225\left(\frac{1}{2500}\right)\left(\frac{49}{50}\right)^{48}=0,1858 \tag{5.69}$$

Verjetnost, da se zgodijo trije takšni dogodki, je enaka:

$$P(X=3)=\binom{50}{3}\left(\frac{1}{50}\right)^{3}\left(1-\frac{1}{50}\right)^{50-3}=19600\left(\frac{1}{50^{3}}\right)\left(\frac{49}{50}\right)^{47}=0,0607 \tag{5.70}$$

Sedaj rešimo še s pomočjo Poissonove porazdelitve.

$$P(X>0)=1-P(X=0)=1-\frac{\lambda^{0}}{0!}e^{-\lambda}=1-\frac{(n\cdot p)^{0}}{0!}e^{-(n\cdot p)}= \\ =1-\frac{\left(50\cdot\frac{1}{50}\right)^{0}}{0!}e^{-\left(50\cdot\frac{1}{50}\right)}=1-e^{-1}=0.6321 \tag{5.71}$$

Verjetnost, da se zgodi en takšen dogodek, je enaka:

$$P(X=1)=\frac{\lambda^1}{1!}e^{-\lambda}=\frac{1^1}{1!}e^{-1}=0.3679 \tag{5.72}$$

Verjetnost, da se zgodita dva takšna dogodka, je enaka:

$$P(X=2)=\frac{\lambda^2}{2!}e^{-\lambda}=\frac{1^2}{2!}e^{-1}=0.1839 \tag{5.73}$$

Verjetnost, da se zgodijo trije takšni dogodki, je enaka:

$$P(X=3)=\frac{\lambda^3}{3!}e^{-\lambda}=\frac{1^3}{3!}e^{-1}=0.0613 \tag{5.74}$$

Vidimo lahko, da so razlike med izračunanimi verjetnostmi obeh porazdelitev, binomske in Poissonove, zelo majhne.

Primer 5.10.:

Dokažite, da za Poissonovo porazdelitev velja enakost:

$$P(X=x)=\frac{\lambda}{x}\cdot P(X=x-1) \tag{5.75}$$

Tvorimo kvocient:

$$\frac{P(X=x)}{P(X=x-1)}=\frac{\frac{\lambda^x}{x!}e^{-\lambda}}{\frac{\lambda^{x-1}}{(x-1)!}e^{-\lambda}}=\frac{\frac{\lambda^x}{x!}}{\frac{\lambda^{x-1}}{(x-1)!}}=\lambda\frac{(x-1)!}{x!}=\lambda\cdot\frac{1}{x} \tag{5.76}$$

Relacijo (5.76) lahko uporabimo za hitrejši izračun zaporednih členov Poissonove porazdelitve.

Primer 5.11.:

Napišite program v Matlabu, ki bo izračunal prvih 16 verjetnosti za binomsko in Poissonovo porazdelitev za $n=150, p=0.05$ *(*$\lambda = n \cdot p = 150 \cdot \frac{5}{100} = 7.5$*). Podajte prikaz komandnega okna.*

Program je:

```
% poisson1.m
%

warning off

clc
clear
close all

n = 150
p = 0.05
q=1-p
lamb = n*p

for i=0:1:15
    po(i+1) = lamb^i*exp(-lamb)/factorial(i);
    bi(i+1) = nchoosek(n,i)*p^i*q^(n-i);
end

disp('prvih 16 verjetnosti za Poissona:')
po

disp('prvih 16 verjetnosti za binomsko:')
bi

warning on
```

Izpis komandnega okna je:

```
n =
   150
p =
    0.0500
q =
    0.9500
lamb =
    7.5000
```

```
prvih 16 verjetnosti za Poissona:
po =
  Columns 1 through 11
    0.0006    0.0041    0.0156    0.0389    0.0729    0.1094    0.1367    0.1465    0.1373    0.1144    0.0858
  Columns 12 through 16
    0.0585    0.0366    0.0211    0.0113    0.0057

prvih 16 verjetnosti za binomsko:
bi =
  Columns 1 through 11
    0.0005    0.0036    0.0141    0.0366    0.0708    0.1088    0.1384    0.1499    0.1410    0.1171    0.0869
  Columns 12 through 16
    0.0582    0.0355    0.0198    0.0102    0.0049
```

Primer 5.12.:

Število novorojenčkov, rojenih v enem tednu v neki porodnišnici, je Poissonova naključna spremenljivka (glej sliko 118). Zapišite njen porazdelitveni zakon in izračunajte verjetnost $P(X \geq 2)$*, da število novorojenčkov, rojenih v enem tednu, ne bo manj kot dva (torej dva ali več).*

Št. novorojenčkov x_i	0	1	2	3	4	5	6	7	8	9	10
Št. tednov f_i	2	5	7	9	15	8	6	4	2	1	1

Slika 118: Frekvenčna porazdelitev po tednih za rojstva novorojenčkov v porodnišnici [Jesenko]

Iz danih podatkov lahko izračunamo aritmetično sredino, ki podaja oceno za povprečno realizacijo naključne spremenljivke (matematično upanje):

$$\bar{x} = E(x) = \lambda = \frac{\sum_{i=0}^{10} f_i \cdot x_i}{N} = \frac{f_1 \cdot x_1 + \ldots + f_{10} \cdot x_{10}}{N} = \frac{2 \cdot 0 + \ldots + 1 \cdot 10}{2 + 5 + \ldots + 1} = \tag{5.77}$$

$$= \frac{245}{60} = 4.0833$$

Poissonov porazdelitveni zakon bo torej imel obliko:

$$P(X=x)=\frac{\lambda^x}{x!}e^{-\lambda}=\frac{4.0833^x}{x!}e^{-4.0833}, \qquad x=0,1,2,... \tag{5.78}$$

Verjetnost $P(X\geq 2)$, da število novorojenčkov, rojenih v enem tednu, ne bo manj kot dva, je enaka:

$$\begin{aligned} &P(X\geq 2)=1-P(X=0)-P(X=1)= \\ &=1-\left(\frac{4.0833^0}{0!}e^{-4.0833}+\frac{4.0833^1}{1!}e^{-4.0833}\right)= \\ &=1-(0.0169+0.0688)=1-0.0857=0.9143 \end{aligned} \tag{5.79}$$

Pri izračunih smo uporabili naslednji program v Matlabu:

```
% poisson2.m

clear
clc
close all

x = 0:10

f = [2 5 7 9 15 8 6 4 2 1 1]

N = sum(f)

st = x*f'

lamb = st/N

p0 = lamb^0*exp(-lamb)/factorial(0)

p1 = lamb^1*exp(-lamb)/factorial(1)

p_vsaj_2 = 1 - p0 - p1
```

komandno okno pa ima izgled:

```
x =
   0   1   2   3   4   5   6   7   8   9  10
f =
   2   5   7   9  15   8   6   4   2   1   1
```

```
N =
    60
st =
   245
lamb =
    4.0833
p0 =
    0.0169
p1 =
    0.0688
p_vsaj_2 =
    0.9143
```

5.1.6 Multidimenzionalna binomska porazdelitev

Multidimenzionalna binomska porazdelitev je neposredna posplošitev binomske. Pri binomski porazdelitvi smo predpostavljali, da sta pri poskusu možna samo dva izida. Kaj pa, če je možnih več izidov? Takšen primer je na primer anketa, kjer sprašujemo, če ljudje podpirajo vlado, so proti njej, ali pa so neopredeljeni. V tem primeru ima vsak poskus očitno tri možne izide.

Naključne spremenljivke $X_1,...,X_k$ določajo multidimenzionalno binomsko naključno spremenljivko tedaj, ko velja [Jesenko]:

$$\begin{aligned} &P\left(X_1 = x_1,...,X_k = x_k\right) = \frac{n!}{x_1!\cdot...\cdot x_k!}\cdot p_1^{x_1}\cdot...\cdot p_k^{x_k}, \\ &x_i = 0,1,...,k, \quad i = 1,2,...,k, \\ &n = \sum_{i=1}^{k} x_i, \qquad \sum_{i=1}^{k} p_i = 1 \end{aligned} \tag{5.80}$$

Primer 5.13.:

Pri proizvodnji nekega izdelka je ugotovljeno, da je 80% izdelkov brezhibnih, 15% jih ima določene manjše napake, 5% pa jih je neuporabnih. Kontrolor je na slepo izbral 10 izdelkov. Kakšna je verjetnost, da bodo štirje brezhibni, 5 jih bo imelo manjše napake, en bo pa neuporaben?

Na osnovi podatkov naloge lahko zapišemo:

$n = 10$
$p_1 = 0.8$
$p_2 = 0.15$
$p_3 = 0.05$
$P(X_1 = 4, X_2 = 5, X_3 = 1) = ?$

Imamo:

$$P(X_1 = 4, X_2 = 5, X_3 = 1) = = \frac{10!}{4! \cdot 5! \cdot 1!} \cdot 0.8^4 \cdot 0.15^5 \cdot 0.05^1 = \tag{5.81}$$
$$= 0.002$$

Pri izračunu smo si pomagali z naslednjim programom v Matlabu:

```
% multibin.m

clear
clc
close all

disp('Vektor vrednosti:')

x = [4 5 1]

disp('Vektor verjetnosti:')

p = [0.8 0.15 0.05]

n = sum(x)

stev = factorial(n)*p(1)^x(1)*p(2)^x(2)*p(3)^x(3)
imen = factorial(x(1))*factorial(x(2))*factorial(x(3))

Verj = stev/imen
```

Izgled komandnega okna je naslednji:

```
Vektor vrednosti:
x =
    4    5    1
Vektor verjetnosti:
p =
    0.8000    0.1500    0.0500
n =
    10
stev =
    5.6435
imen =
        2880
Verj =
    0.0020
```

5.1.7 Multidimenzionalna hipergeometrična porazdelitev

Pri enodimenzionalni hipergeometrčni naključni spremenljivki smo imeli opravka z množico N elementov, v kateri jih en del imel določeno lastnost, preostali pa te lastnosti niso imeli. Problem sedaj posplošimo tako, da ima od vseh N elementov M_1 elementov lastnost L_1, M_2 elementov lastnost L_2, itn,..., M_k elementov pa lastnost L_k.

Naključne spremenljivke $X_1,...,X_k$ določajo multidimenzionalno hipergeometrično naključno spremenljivko tedaj, ko velja [Jesenko]:

$$P\left(X_1 = x_1,...,X_k = x_k\right) = \frac{\binom{M_1}{x_1}\cdot\binom{M_2}{x_2}\cdot\ldots\cdot\binom{M_k}{x_k}}{\binom{N}{n}}, \tag{5.82}$$

$$x_i = 0,1,...,n, \quad x_i \le M_i, \quad i = 1,2,...,k,$$

$$n = \sum_{i=1}^{k} x_i, \qquad \sum_{i=1}^{k} M_i = N$$

Primer 5.14.:

V skupini 20 študentov jih je šest iz 1. letnika, trije iz 2. letnika, sedem iz 3. letnika, ter štirje iz 4. letnika. Na slepo želimo za neko aktivnost izbrati 12 študentov. Kakšna je verjetnost, da jih bomo iz vsakega letnika izbrali enako število, torej tri? [Jesenko]

Na osnovi podatkov naloge lahko zapišemo:

$N = 20$
$M_1 = 6$
$M_2 = 3$
$M_3 = 7$
$M_4 = 4$
$n = 12$
$x_1 = x_2 = x_3 = x_4 = 3$
$P\left(X_1 = 3, X_2 = 3, X_3 = 3, X_4 = 3\right) = ?$

Dobimo:

$$P(X_1=3,X_2=3,X_3=3,X_4=3)=\frac{\binom{M_1}{x_1}\cdot\binom{M_2}{x_2}\cdot\binom{M_3}{x_3}\cdot\binom{M_4}{x_4}}{\binom{N}{n}}= \tag{5.83}$$

$$=\frac{\binom{6}{3}\cdot\binom{3}{3}\cdot\binom{7}{3}\cdot\binom{4}{3}}{\binom{20}{12}}=0.0222$$

Pri izračunu smo si pomagali z naslednjim programom v Matlabu:

```
% multihip.m

clear
clc
close all

disp('Vektor vrednosti naključnih spremenljivk:')

x = [3 3 3 3]

disp('Vektor stevila elementov mnozic z doloceno lastnostjo:')

M = [6 3 7 4]

N = sum(M)
n = sum(x)

disp('Verjetnost je:')

stev = nchoosek(M(1),x(1))* nchoosek(M(2),x(2))*nchoosek(M(3),x(3))*nchoosek(M(4),x(4))
imen = nchoosek(N,n)

Verj = stev/imen
```

Izgled komandnega okna je naslednji:

```
Vektor vrednosti naključnih spremenljivk:
x =
   3   3   3   3
Vektor stevila elementov mnozic z doloceno lastnostjo:
M =
   6   3   7   4
N =
   20
n =
   12


Verjetnost je:
```

```
stev =
        2800
imen =
      125970
Verj =
    0.0222
```

5.2 Zvezne porazdelitve

V tem poglavju bomo podrobneje spoznali nekatere zvezne naključne spremenljivke, ki jih v statistiki pogosto uporabljamo.

5.2.1 Uniformna porazdelitev

Uniformno zvezno naključno spremenljivko smo že spoznali v poglavju 2.7, izračunali njeno matematično upanje $E(X)=\frac{a+b}{2}$ v poglavju 2.9 (glej izraz (2.54)), ter izračunali njeno varianco $VAR(X)=\frac{(a-b)^2}{12}$ v poglavju 2.10 (glej izraz (2.72)).

Rodovno funkcijo momentov izračunamo na naslednji način [Krishnamoorthy]:

$$M(t)=E\left(e^{tX}\right)=\int_a^b e^{t\cdot x}\cdot\frac{1}{b-a}dx=\frac{1}{b-a}\cdot\frac{1}{t}\cdot\left(e^{t\cdot x}\right)_a^b=\frac{1}{b-a}\cdot\frac{1}{t}\cdot\left(e^{t\cdot b}-e^{t\cdot a}\right) \qquad (5.84)$$

Primer 5.15.:

Nek stroj nenehno polni buteljke, katerih volumen je lahko katerakoli vrednost med 0.7 in 0.755 dcl. Predpostavimo, da so te vrednosti uniformno porazdeljene. Izračunajte

verjetnost, da je volumen naključno izbrane buteljke večji od 0.75 dcl. Izračunajte tudi pričakovano vrednost naključne spremenljivke.

Porazdelitvena funkcija je enaka:

$$f(x)=\begin{cases}\dfrac{1}{b-a} & a\le x\le b\\ 0 & \text{sicer}\end{cases}=\begin{cases}\dfrac{1}{0.755-0.7} & 0.7\le x\le 0.755\\ 0 & \text{sicer}\end{cases}=$$
$$=\begin{cases}\dfrac{1}{0.055} & 0.7\le x\le 0.755\\ 0 & \text{sicer}\end{cases} \tag{5.85}$$

Verjetnost, da je volumen naključno izbrane buteljke večji od 0.75 dcl, je enaka:

$$P(X>0.75)=\int_{0.75}^{0.755}\frac{1}{0.055}\,dx=\frac{1}{0.055}(0.755-0.75)=\frac{0.005}{0.055}=0.0909 \tag{5.86}$$

Pričakovana vrednost naključne spremenljivke je:

$$E(X)=\frac{a+b}{2}=\frac{0.7+0.755}{2}=0.7275 \tag{5.87}$$

5.2.2 Normalna in standardna normalna porazdelitev

Normalno naključno spremenljivko smo že spoznali v poglavju 2.8, ter izračunali njeno matematično upanje $E(X)=\mu$ v poglavju 2.9 (glej izraz (2.53)). Rodovno funkcijo momentov izračunamo na naslednji način:

$$M(t)=E\left(e^{t\cdot X}\right)=\int_{-\infty}^{\infty}e^{t\cdot x}\cdot f(x)\,dx=\int_{-\infty}^{\infty}e^{t\cdot x}\cdot\frac{1}{\sigma\cdot\sqrt{2\pi}}\cdot e^{-\frac{(x-\mu)^2}{2\cdot\sigma^2}}dx=$$
$$=\frac{1}{\sigma\cdot\sqrt{2\pi}}\int_{-\infty}^{\infty}e^{t\cdot x-\frac{(x-\mu)^2}{2\cdot\sigma^2}}dx=\frac{1}{\sigma\cdot\sqrt{2\pi}}\int_{-\infty}^{\infty}e^{\frac{2\cdot\sigma^2\cdot t\cdot x-(x-\mu)^2}{2\cdot\sigma^2}}dx= \tag{5.88}$$
$$=\frac{1}{\sigma\cdot\sqrt{2\pi}}\int_{-\infty}^{\infty}e^{-\frac{-2\cdot\sigma^2\cdot t\cdot x+(x-\mu)^2}{2\cdot\sigma^2}}dx$$

Pokazati se da, da velja naslednji izraz [Jesenko]:

$$-2\cdot\sigma^2\cdot t\cdot x+(x-\mu)^2=\left[x-\left(\mu+t\cdot\sigma^2\right)\right]^2-2\cdot\sigma^2\cdot t\cdot\mu-t^2\sigma^4 \tag{5.89}$$

Sledi:

$$M(t)=\frac{1}{\sigma\cdot\sqrt{2\pi}}\int_{-\infty}^{\infty}e^{-\frac{\left[x-\left(\mu+t\cdot\sigma^2\right)\right]^2-2\cdot\sigma^2\cdot t\cdot\mu-t^2\sigma^4}{2\cdot\sigma^2}}dx=$$
$$=\frac{1}{\sigma\cdot\sqrt{2\pi}}e^{-\frac{-2\cdot\sigma^2\cdot t\cdot\mu-t^2\sigma^4}{2\cdot\sigma^2}}\underbrace{\int_{-\infty}^{\infty}e^{-\frac{\left[x-\left(\mu+t\cdot\sigma^2\right)\right]^2}{2\cdot\sigma^2}}dx}_{\sigma\cdot\sqrt{2\pi}}= \tag{5.90}$$
$$=e^{t\cdot\mu+\frac{t^2\sigma^2}{2}}$$

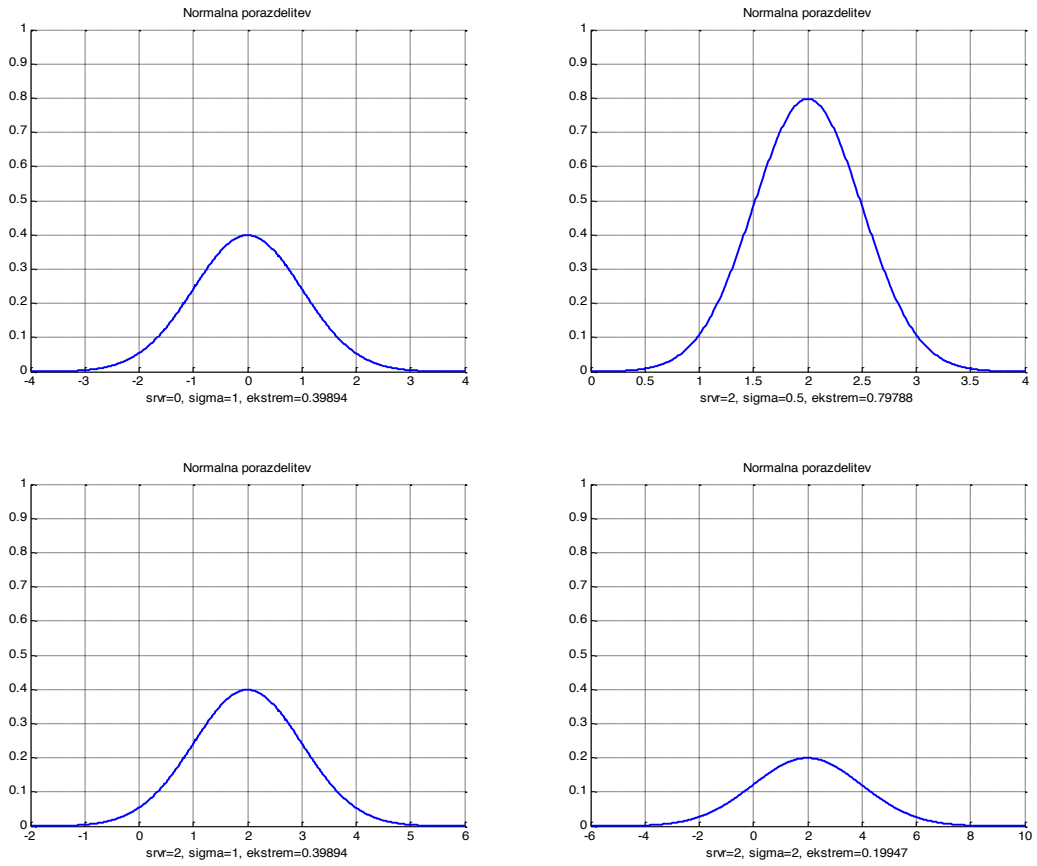
Seveda v primeru standardne normalne porazdelitve $N(0,1)$ dobimo rodovno funkcijo momentov $M(t)=e^{\frac{t^2}{2}}$, kar smo pokazali že v poglavju 2.21 (glej izraz (2.225)).

Na osnovi rodovne funkcije momentov (5.90) bi sedaj lahko pokazali tudi, da je varianca pri normalni porazdelitvi enaka $VAR(X)=\sigma^2$, podobno, kot smo to v poglavju 2.21 naredili za standardno normalno porazdelitev $N(0,1)$ (glej izraz (2.229)).

V poglavju 2.8 smo tudi pokazali, da normalna porazdelitev doseže maksimum v točki $\left(\mu, \frac{1}{\sigma \cdot \sqrt{2\pi}}\right)$. Naslednji program v Matlabu prikazuje, kako lahko narišemo normalno porazdelitev:

```
% normal.m
%

clc
clear
close all

while 1==1

    srvr = input('srvr = ')
    sigma = input('sigma=')
    disp('ekstrem je pri:')
    ekstrem=1/sigma/sqrt(2*pi)

    xsp=srvr-4*sigma
    xzg=srvr+4*sigma

    x =xsp:0.01:xzg;
    y = normpdf(x,srvr,sigma);
    hold on

    plot(x,y,'LineWidth',2)
    title('Normalna porazdelitev')
    xlabel(['srvr=' num2str(srvr) ', sigma=' num2str(sigma) ', ekstrem='
    num2str(ekstrem)])
    d = axis;
    axis([d(1) d(2) d(3) 1])
    grid

    ch = input('Zelis izhod 1(DA)/0(NE)')
    if ch == 1
       return
    end

    figure


end
```

Slika 118 prikazuje normalno porazdelitev pri različnih vrednostih parametrov (μ,σ).



Slika 118: Normalna porazdelitev pri različnih vrednostih parametrov (μ,σ).

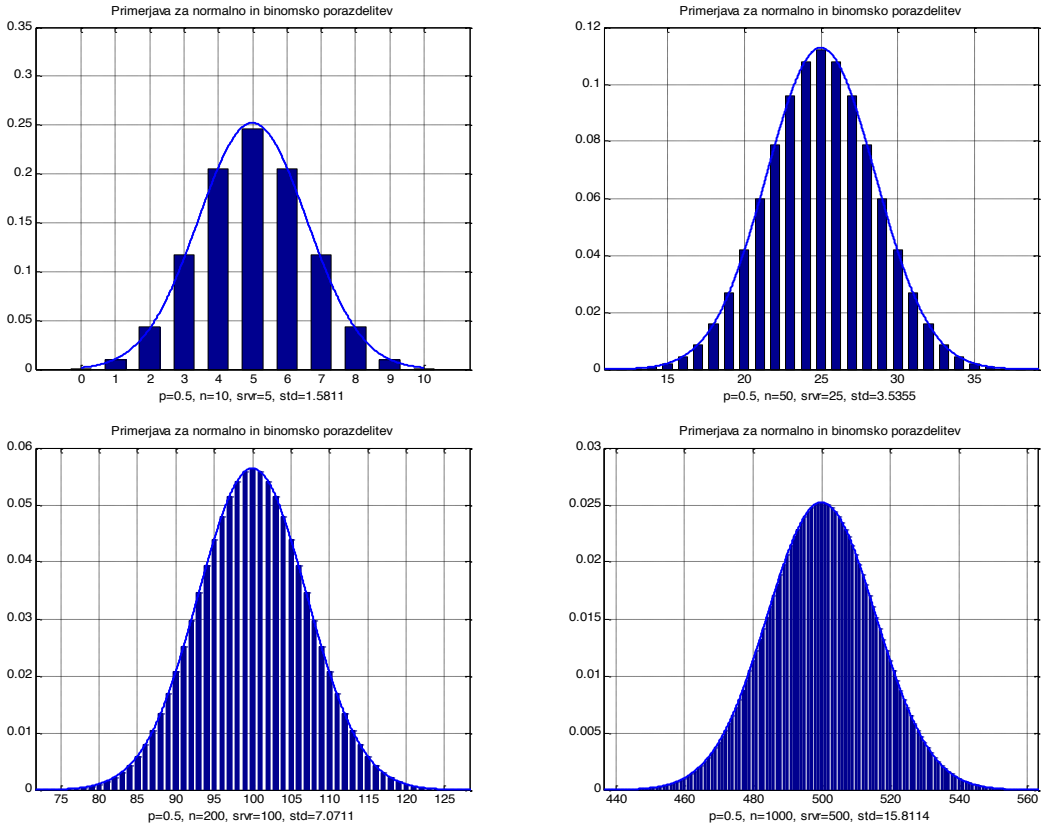
Povezava med binomsko in normalno porazdelitvijo

Povezava izhaja iz Laplaceovega točkovnega obrazca in se glasi [Jurišić]:

$$P(X=x,n,p,q)\approx\frac{1}{\sqrt{2\pi}}\cdot\frac{1}{\sqrt{n\cdot p\cdot q}}\cdot e^{-\frac{(x-n\cdot p)^2}{2\cdot n\cdot p\cdot q}}=N\left(n\cdot p,\sqrt{n\cdot p\cdot q}\right)=N(\mu,\sigma) \tag{5.91}$$

Ta povezava velja pri velikih *n* in ko je *p* blizu 0.5, izpeljemo pa jo s pomočjo zakona velikih števil oz. centralnega limitnega teorema [Hsu, Jurišić].

Slika 119 prikazuje normalno in binomsko porazdelitev pri različnih vrednostih parametrov *n* in *p*.



Slika 119: Normalna in binomska porazdelitev pri različnih vrednostih parametrov n in p

Pri izrisu slike 119 smo si pomagali z naslednjim programom v Matlabu:

```
% normal1.m

clc
clear
close all

p=input('p=')
q=1-p
n=input('n=')
x1=0:1:n;
x2=0:0.01:n;

srvr=n*p
sigma=sqrt(n*p*q)

B = binopdf(x1,n,p);
N = normpdf(x2,srvr,sigma);

bar(x1,B,0.6);
hold on
plot(x2,N,'LineWidth',1.5)

xsp=srvr-4*sigma
xzg=srvr+4*sigma

d=axis
axis([xsp xzg d(3) d(4)])

title('Primerjava za normalno in binomsko porazdelitev')

xlabel(['p=' num2str(p) ', n=' num2str(n) ', srvr=' num2str(srvr) ', std='
num2str(sigma)])

grid
```

Težave pri računanju verjetnosti in prehod na standardno normalno porazdelitev

Kumulativno funkcijo za normalno porazdelitev lahko izrazimo na naslednji način:

$$F(X)=P(X\le x)=\int_{-\infty}^{x}\frac{1}{\sigma\cdot\sqrt{2\pi}}\cdot e^{-\frac{(t-\mu)^2}{2\sigma^2}}dt \tag{5.92}$$

Normalna naključna spremenljivka je ena najpomembnejših naključnih spremenljivk v statistiki in je v različnih problemih pogostokrat uporabljena. Žal pa se pri tovrstni uporabi srečamo z računanjem integrala, ki ga analitično neposredno ne moremo izračunati. Problem je rešen tako, da vpeljemo standardno normalno porazdelitev [Jesenko]. To smo nekoliko že spoznali v poglavju 3.4, kjer smo pokazali, da moramo vpeljati naslednjo transformacijo nad normalno naključno spremenljivko:

$$z=\frac{x-\mu}{\sigma} \tag{5.93}$$

pri čemer dobimo porazdelitev gostote verjetnosti:

$$f(z)=\frac{1}{\sqrt{2\pi}}\cdot e^{-\frac{(z)^2}{2}} \tag{5.94}$$

kar smo pokazali v izrazu (3.24).

V nadaljevanju pokažimo, da velja enakost: $P(x_1\le X\le x_2)=P(z_1\le Z\le z_2)$. Pri tem upoštevamo relacije:

$$\begin{aligned} &z=\frac{x-\mu}{\sigma},\quad dz=\frac{dx}{\sigma}\\ &z_1=\frac{x_1-\mu}{\sigma},z_2=\frac{x_2-\mu}{\sigma} \end{aligned} \tag{5.95}$$

Sledi:

$$\begin{aligned} &P(x_1\le X\le x_2)=\int_{x_1}^{x_2}\frac{1}{\sigma\cdot\sqrt{2\pi}}\cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx=\int_{\frac{x_1-\mu}{\sigma}}^{\frac{x_2-\mu}{\sigma}}\frac{1}{\sigma\cdot\sqrt{2\pi}}\cdot e^{-\frac{(z)^2}{2}}\sigma dz=\\ &=\frac{1}{\sqrt{2\pi}}\int_{z_1}^{z_2}e^{-\frac{(z)^2}{2}}dz=P(z_1\le Z\le z_2) \end{aligned} \tag{5.96}$$

Funkcija porazdelitve gostote verjetnosti za standardno normalno naključno spremenljivko ima obliko:

$$F(z)=P(Z\le z)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z}e^{-\frac{(t)^2}{2}}dt \tag{5.97}$$

Vrednosti funkcije (5.97) so zapisane na slikah 40 in 41 v poglavju 3.4.

Primer 5.16.:

Poiščite verjetnosti:

$$F(z=1.82)=P(Z\le 1.82)=?$$
$$F(z=-0.89)=P(Z\le -0.89)=?$$
$$P(1.23\le Z\le 1.65)=P(Z\le 1.65)-P(Z\le 1.23)=F(z=1.65)-F(z=1.23)=?$$
$$P(-0.28\le Z\le 1.14)=P(Z\le 1.14)-P(Z\le -0.28)=F(z=1.14)-F(z=-0.28)=?$$

Da bi izračunali te rezultate, gremo lahko pogledat v tabeli na slikah 40 oz. 41 in dobimo:

$$F(z=1.82)=P(Z\le 1.82)=0.9656$$
$$F(z=-0.89)=P(Z\le -0.89)=0.1867$$
$$P(1.23\le Z\le 1.65)=P(Z\le 1.65)-P(Z\le 1.23)=F(z=1.65)-F(z=1.23)=$$
$$=0.9505-0.8907=0.0598$$
$$P(-0.28\le Z\le 1.14)=P(Z\le 1.14)-P(Z\le -0.28)=F(z=1.14)-F(z=-0.28)=$$
$$=0.8729-0.3897=0.4832$$

Na tem mestu je vredno poudariti, da v današnjem času sodobnih programskih orodij tabele hitro izgubljajo na pomenu. Tako bomo v nadaljevanju pokazali, kako se da dotične rezultate izračunati z Matlabom.

Uporabili bomo program **cum_fun.m**, ki ima naslednjo obliko:

```
% kumulativna funkcija (cum_fun.m):
%
% klic je npr. F=cum_fun(-5,10,'',2,[3 3])
% ekvivalent bi bil F = normcdf(10,3,3)-normcdf(-5,3,3)
%

function F = cum_sum(xsp,xzg,f,flag,par);

clc
close all
dx = 0.01;
```

```
if length(f) == 0
   if exist('flag','var') == 0
         disp('nisi vnesel flaga')
         F = 0
         return
   end
   if flag == 1
       srvr = 0;
       sigma = 1;
       f = 'exp(-i^2/2)/sqrt(2*pi)'
   elseif flag == 2
      if exist('par','var') == 0
         disp('nisi vnesel parametrov srvr in sigma')
         F = 0
         return
      else
         srvr = par(1);
         sigma = par(2);
      end
      f = 'exp(-(i-srvr)^2/2/sigma^2)/sqrt(2*pi)/sigma'
   end
end


x = [xsp:dx:xzg];

I = 0;

for i = x
    I = I + dx*eval(f);
end

disp('Verjetnost na izbranem intervalu je:')
F=I

% Za normalno porazdelitev se nariše tudi graf, po potrebi normaliziran:
% (funkcija normaldistribution pobrana iz file exchange)

if (length(f)>0)&&(flag==2)
    ch = input('Zelis normaliziran graf 0-ne,1-da');
    G1=normaldistribution(xzg,srvr,sigma,ch,1);
    G2=normaldistribution1(xsp,srvr,sigma,ch,1);
    disp('Verjetnost na izbranem intervalu z normaldistribution.m je:')
    G=G1-G2
    ylabel(['Verjetnost na izbranem intervalu ' num2str(xsp) ':' num2str(xzg) ' je:'
             num2str(G)])
    if ch == 1
       disp('POZOR: GRAF JE NORMALIZIRAN!!!!!!!!!!!!!!!')
    end
    disp('             ')
    disp('Vnesi procente na graf!!!!!!!!!!!!!!!')
    gtext(num2str(G),'FontSize',24)
    disp('             ')
    ch1 = input('zelis vnesti spodnjo mejo na graf 1-da, 0-ne')
    if ch1 == 1
       disp('             ')
       disp('Vnesi spodnjo mejo na graf!!!!!!!!!!!!!!!')
       gtext(num2str(xsp),'FontSize',15)
    end
    disp('             ')
    disp('Vnesi zgornjo mejo na graf!!!!!!!!!!!!!!!')
    gtext(num2str(xzg),'FontSize',15)
end

disp('          ')
disp('Verjetnost na izbranem intervalu s standardnimi matlab ukazi je:')

F1 = normcdf(xzg,srvr,sigma)-normcdf(xsp,srvr,sigma)
```

Program je sicer prirejen zlasti za normalno porazdelitev, vendar bi se z njim dalo (z manjšimi modifikacijami) računati verjetnosti tudi pri drugih porazdelitvah. Za izris grafa program kliče tudi funkcijo **normaldistribution.m**, pridobljeno na MathWorks File Exchange, ki pa je bila še nekoliko prilagojena za naše potrebe.

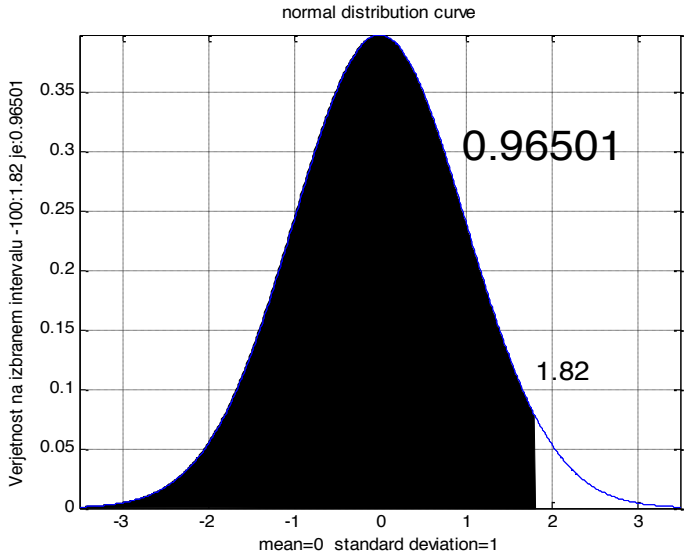
Če bi želeli najprej izračunati verjetnost $F(z=1.82)=P(Z\leq 1.82)=?$, bi bil klic funkcije naslednji (z argumentom **[0 1]** poudarimo, da gre za standardno porazdelitev):

```
cum_fun(-100,1.82,'',2,[0 1])
```

Komandno okno bi imelo naslednji izgled:

```
f =
exp(-(i-srvr)^2/2/sigma^2)/sqrt(2*pi)/sigma

Verjetnost na izbranem intervalu je:
F =
   0.9660

Zelis normaliziran graf 0-ne,1-da0

Verjetnost na izbranem intervalu z normaldistribution.m je:
G =
   0.9650

Vnesi procente na graf!!!!!!!!!!!!!!!
zelis vnesti spodnjo mejo na graf 1-da, 0-ne0
ch1 =
    0
Vnesi zgornjo mejo na graf!!!!!!!!!!!!!!!

Verjetnost na izbranem intervalu s standardnimi matlab ukazi je:
F1 =
   0.9656
```

Do rahlih razlik pri rezultatu pride, ker se za izračune uporabijo malce drugačne numerične tehnike. Poleg dobljenega rezultata program nariše tudi graf, ki ga prikazuje slika 120.



Slika 120: Ploščina pod normalno porazdelitveno funkcijo, ki predstavlja verjetnost

$$F(z=1.82)=P(Z\le 1.82)=0.9656$$

Na podoben način bi program uporabili tudi za ostale izračune, to je:

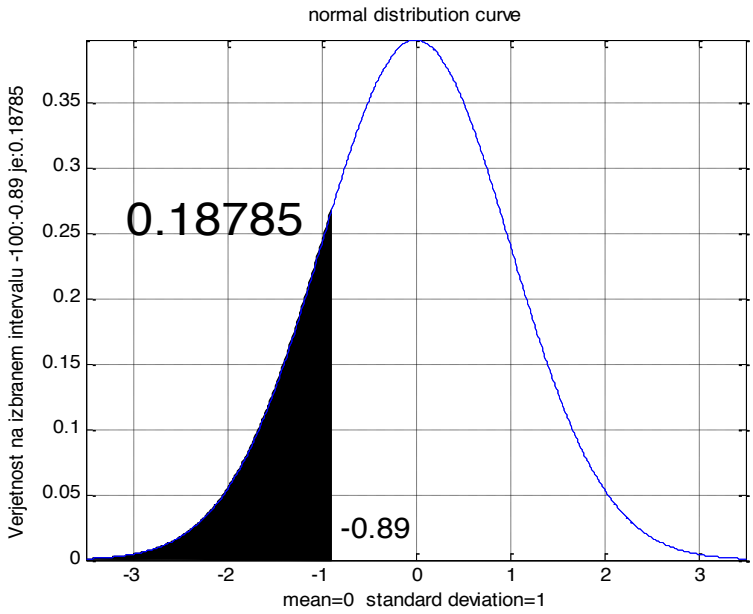
$$F(z=-0.89)=P(Z\le -0.89)=0.1867$$
$$P(1.23\le Z\le 1.65)=P(Z\le 1.65)-P(Z\le 1.23)=F(z=1.65)-F(z=1.23)=$$
$$=0.9505-0.8907=0.0598$$
$$P(-0.28\le Z\le 1.14)=P(Z\le 1.14)-P(Z\le -0.28)=F(z=1.14)-F(z=-0.28)=$$
$$=0.8729-0.3897=0.4832$$

Slike 121, 122 in 123 prikazujejo ploščine pod normalno porazdelitveno funkcijo, ki predstavljajo izračunane verjetnosti. Pri tem smo izvedli naslednje klice funkcije cum_fun.m:

- **cum_fun(-100,-0.89,'',2,[0 1])**

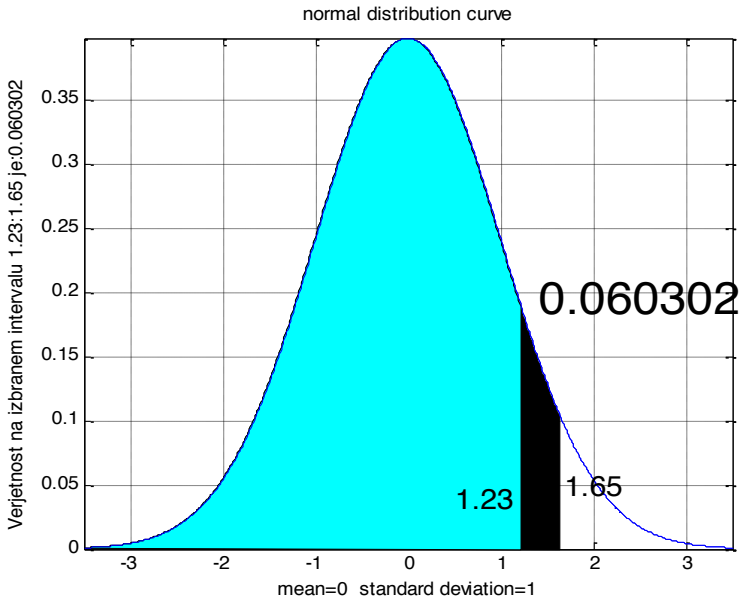
- **cum_fun(1.23,1.65,'',2,[0 1])**

- **cum_fun(-0.28,1.14,'',2,[0 1])**



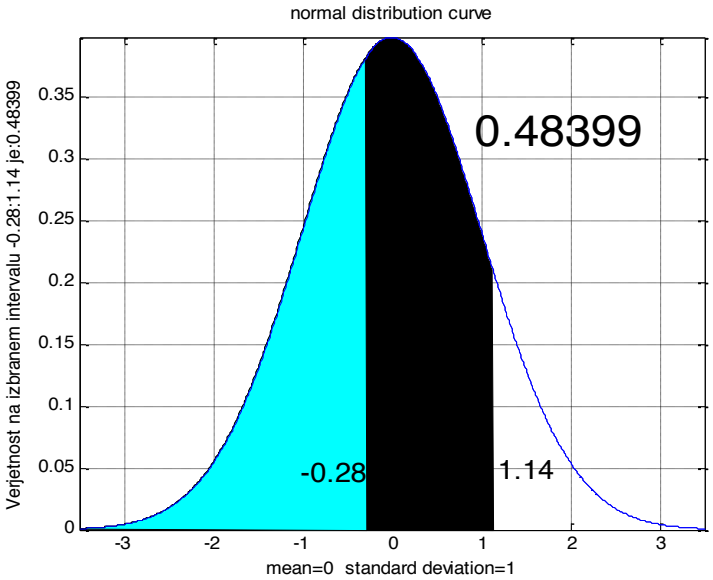
Slika 121: Ploščina pod normalno porazdelitveno funkcijo, ki predstavlja verjetnost

$$F(z=-0.89)=P(Z\le -0.89)=0.18785$$



Slika 122: Ploščina pod normalno porazdelitveno funkcijo, ki predstavlja verjetnost

$$P(1.23\le Z\le 1.65)=P(Z\le 1.65)-P(Z\le 1.23)=F(z=1.65)-F(z=1.23)=0.06$$



Slika 123: Ploščina pod normalno porazdelitveno funkcijo, ki predstavlja verjetnost

$$P(-0.28 \le Z \le 1.14) = P(Z \le 1.14) - P(Z \le -0.28) = F(z = 1.14) - F(z = -0.28) = 0.483$$

Primer 5.17.:

Proizvodni čas izdelka je normalna naključna spremenljivka z matematičnim upanjem 35 minut in standardno deviacijo 1.3 minute. Kakšna je verjetnost za to, da se nek izdelek ne bo izdeloval več kot 33 minut? [Jesenko]

Zapišimo podatke naloge:

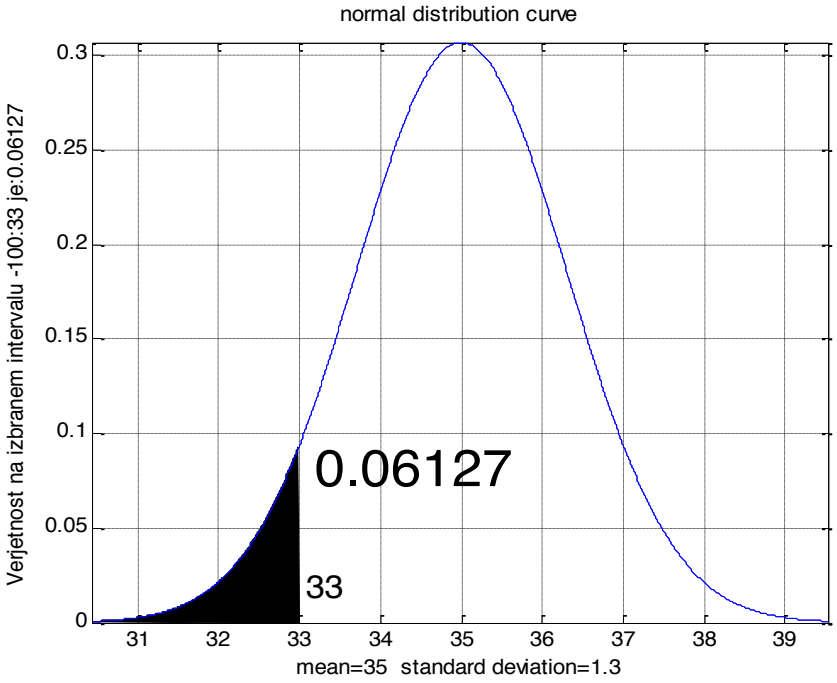
$X \in N(\mu, \sigma) = N(35, 1.3)$
$E(X) = \mu = 35$
$STD(X) = \sigma = 1.3$
$P(X \le 33) = ?$

Če bi želeli uporabiti program **cum_fun.m** brez transformacije v standardno naključno spremenljivko, bi izvedli ukaz:

```
cum_fun(-100,33,'',2,[35 1.3])
```

Komandno okno bi imelo izgled (glej sliko 124):

```
f =

exp(-(i-srvr)^2/2/sigma^2)/sqrt(2*pi)/sigma

Verjetnost na izbranem intervalu je:

F =

    0.0624

Zelis normaliziran graf 0-ne,1-da0

Verjetnost na izbranem intervalu z normaldistribution.m je:

G =

    0.0613

Vnesi procente na graf!!!!!!!!!!!!!!!

zelis vnesti spodnjo mejo na graf 1-da, 0-ne0

ch1 =

     0

Vnesi zgornjo mejo na graf!!!!!!!!!!!!!!!

Verjetnost na izbranem intervalu s standardnimi matlab ukazi je:

F1 =

    0.0620
```



Slika 124: Ploščina pod normalno porazdelitveno funkcijo, ki predstavlja verjetnost

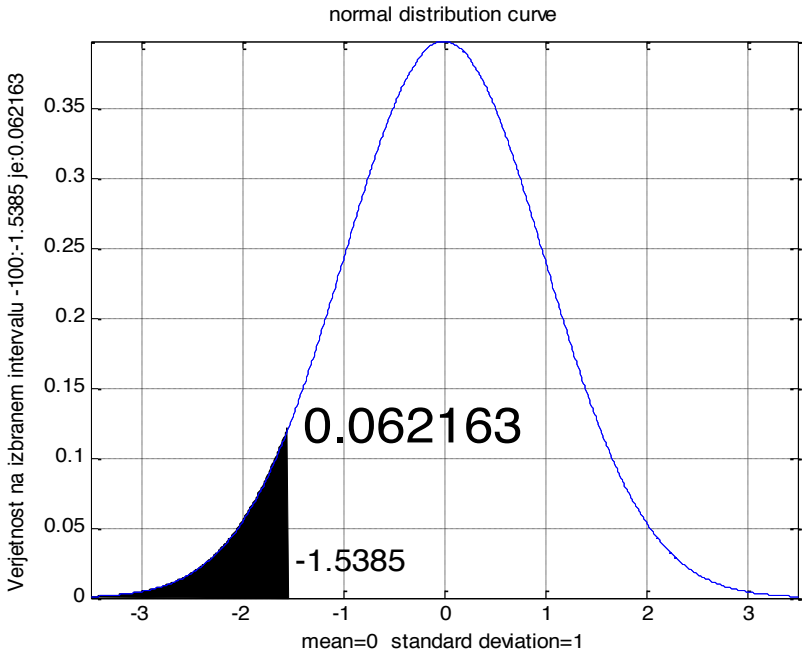
$$P(X \le 33) = 0.06127$$

Če bi pa želeli uporabiti program **cum_fun.m** z transformacijo v standardno naključno spremenljivko, bi izvedli ukaz ($z = \frac{x-35}{1.3} = \frac{33-35}{1.3} = -1.5385$):

```
cum_fun(-100,-1.5385,'',2,[0 1])
```

Komandno okno bi imelo izgled (glej sliko 125):

```
f =
exp(-(i-srvr)^2/2/sigma^2)/sqrt(2*pi)/sigma

Verjetnost na izbranem intervalu je:
F =
   0.0624

Zelis normaliziran graf 0-ne,1-da0

Verjetnost na izbranem intervalu z normaldistribution.m je:
G =
   0.0622

Vnesi procente na graf!!!!!!!!!!!!!!!
zelis vnesti spodnjo mejo na graf 1-da, 0-ne0
ch1 =
    0

Vnesi zgornjo mejo na graf!!!!!!!!!!!!!!!

Verjetnost na izbranem intervalu s standardnimi matlab ukazi je:
F1 =
   0.0620
```



Slika 125: Ploščina pod normalno porazdelitveno funkcijo, ki predstavlja verjetnost

$$P(Z \le -1.5385) = 0.0621$$

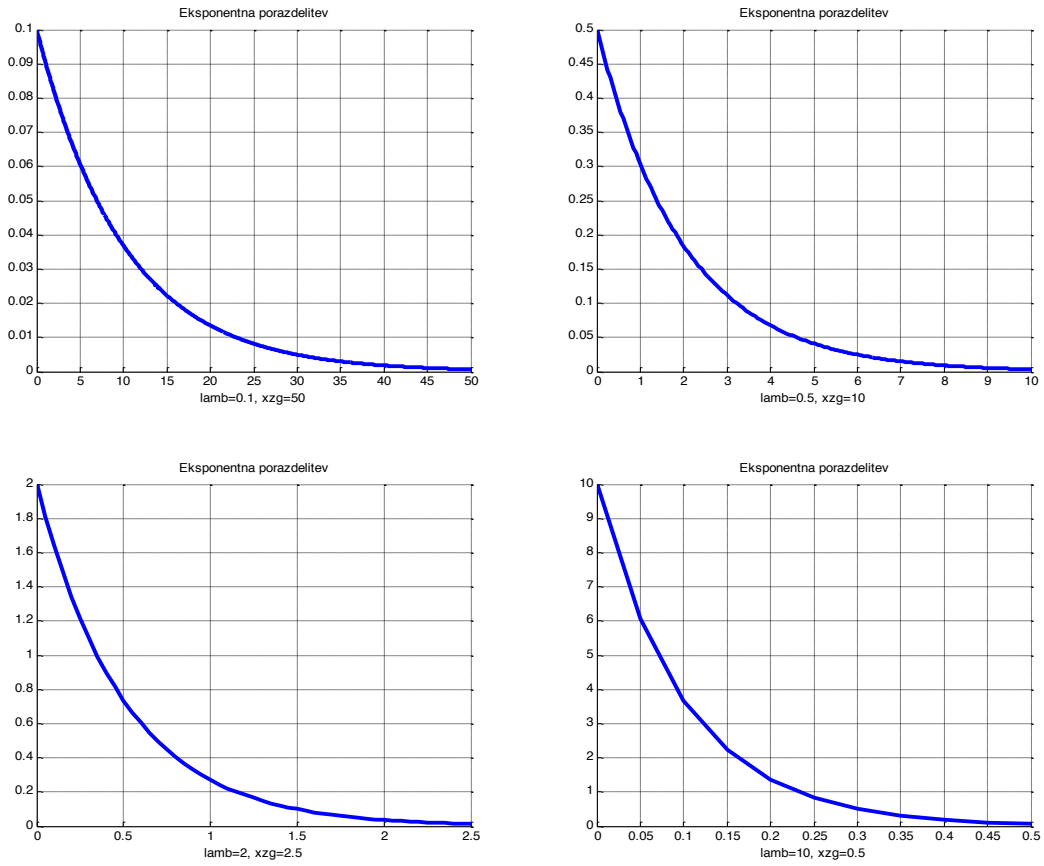
5.2.3 Eksponentna porazdelitev

Eksponentno zvezno naključno spremenljivko smo že spoznali v poglavju 2.6, izračunali njeno matematično upanje $E(X) = \frac{1}{\lambda}$ v poglavju 2.21 (glej izraz (2.219)), ter izračunali njeno varianco $VAR(X) = \frac{1}{\lambda^2}$ v poglavju 2.21 (glej izraz (2.220)). Prav tako smo v poglavju 2.21 izračunali rodovno funkcijo momentov $M(t) = \frac{\lambda}{\lambda - t}$ (glej izraz 2.216).

Kumulativno funkcijo izračunamo na naslednji način:

$$F(x) = P(X \le x) = \int_0^x \lambda \cdot e^{-\lambda \cdot t} dt = \lambda \cdot \frac{1}{(-\lambda)} \left(e^{-\lambda \cdot t} \right)_0^x = \tag{5.98}$$
$$= -\left(e^{-\lambda \cdot x} - 1 \right) = 1 - e^{-\lambda \cdot x}, \quad x \ge 0$$

Slika 126 prikazuje eksponentno porazdelitev pri različnih vrednostih parametra λ.



Slika 126: Eksponentna porazdelitev pri različnih vrednostih parametra λ.

Za izris slike 126 smo si pomagali z naslednjim programom v Matlabu:

```
% eksp.m

clc
clear
close all

while 1==1

    lamb = input('lamb = ')

    xsp=0
    xzg=5/lamb

    x =xsp:0.05:xzg;
    y = exppdf(x,1/lamb);
    hold on

    plot(x,y,'LineWidth',3)
    title('Eksponentna porazdelitev')
    xlabel(['lamb=' num2str(lamb) ', xzg=' num2str(xzg)])
    grid

    ch = input('Zelis izhod 1(DA)/0(NE)')
    if ch == 1
       return
    end

    figure


end
```

Primer 5.18.:

Čas trajanja žarnic je eksponentna naključna spremenljivka. V množici žarnic vzamemo vzorec in ugotovimo, da 5% žarnic sveti do 100 ur. Določite parameter λ. Kolikšna je verjetnost, da bo neka žarnica svetila več kot 200 ur?

Zapišimo podatke naloge:

$$X \in EXP(\lambda)$$
$$P(X \le 100) = 0.05 = F(100)$$
$$\lambda = ?$$
$$P(X \ge 200) = ?$$

Zapišemo:

$$\begin{aligned} &F(x) = 1 - e^{-\lambda \cdot x}, \quad x \ge 0 \\ &F(100) = 1 - e^{-\lambda \cdot 100} \\ &e^{-\lambda \cdot 100} = 1 - 0.05 = 0.95 \\ &\ln e^{-\lambda \cdot 100} = \ln 0.95 \\ &-\lambda \cdot 100 = \ln 0.95 \\ &\lambda = -\frac{\ln 0.95}{100} = 5.12 \cdot 10^{-4} \end{aligned} \tag{5.99}$$

Sledi:

$$\begin{aligned} &P(X \ge 200) = 1 - P(X < 200) = \\ &= 1 - F(200) = 1 - \left(1 - e^{-\lambda \cdot 200}\right) = \\ &= e^{-\lambda \cdot 200} = e^{-5.12 \cdot 10^{-4} \cdot 200} = e^{-0.1026} = 0.9025 \end{aligned} \tag{5.100}$$

Drugi način:

$$\begin{aligned} &P(X \ge 200) = \int_{200}^{\infty} \lambda \cdot e^{-\lambda \cdot t} dt = \lambda \cdot \frac{1}{(-\lambda)} \left(e^{-\lambda \cdot t}\right)_{200}^{\infty} = \\ &= -\left(0 - e^{-\lambda \cdot 200}\right) = e^{-\lambda \cdot 200} = 0.9025 \end{aligned} \tag{5.101}$$

5.2.4 Gama porazdelitev

Naključna spremenljivka se imenuje **Gama naključna spremenljivka** s parametroma $\left(a>0, \lambda=\frac{1}{b}>0\right)$, če ima njena porazdelitev gostote verjetnosti naslednjo obliko:

$$\begin{aligned} f(x) &= \frac{\lambda \cdot e^{-\lambda \cdot x} \cdot (\lambda \cdot x)^{a-1}}{\Gamma(a)} = \frac{\lambda \cdot e^{-\lambda \cdot x} \cdot (\lambda)^{a-1} (x)^{a-1}}{\Gamma(a)} = \\ &= \frac{\frac{1}{b} \cdot e^{-\frac{1}{b} \cdot x} \cdot \left(\frac{1}{b}\right)^{a-1} (x)^{a-1}}{\Gamma(a)} = \frac{e^{-\frac{1}{b} \cdot x} \cdot (x)^{a-1}}{b^a \cdot \Gamma(a)} \qquad , x>0 \\ f(x) &= 0, \qquad x<0 \end{aligned} \tag{5.102}$$

kjer je $\Gamma(a)$ takoimenovana **gama funkcija**, ki ima obliko [Jesenko, Hsu]:

$$\Gamma(a) = \int_0^\infty x^{a-1} \cdot e^{-x} dx, \qquad a>0 \tag{5.103}$$

S pomočjo Per Partes integracije se da pokazati (glej [Hsu]), da za gama funkcijo velja naslednja lastnost [Jesenko, Hsu]:

$$\begin{aligned} &\Gamma(a) = (a-1) \cdot \Gamma(a-1), \qquad a>0 \\ &oz. \\ &\Gamma(a+1) = (a) \cdot \Gamma(a) \end{aligned} \tag{5.104}$$

Za gama funkcijo velja tudi ta lastnost:

$$\Gamma(1) = \int_0^\infty x^{1-1} \cdot e^{-x} dx = \int_0^\infty e^{-x} dx = -\left(e^{-x}\right)_0^\infty = 1 \tag{5.105}$$

V matematiki je gama funkcija posplošitev funkcije faktoriela, definirana pa je za realna in kompleksna števila, ni pa definirana za negativna cela števila in ničlo.

Če je število *a* naravno število ($a \in N$), potem lahko gama funkcijo zapišemo v naslednji obliki:

$$\begin{aligned}
&\Gamma(a+1) = (a)\cdot\Gamma(a) = \\
&= (a)\cdot\underbrace{(a-1)\cdot\Gamma(a-1)}_{\Gamma(a)} = \\
&= (a)\cdot(a-1)\underbrace{(a-2)\cdot\Gamma(a-2)}_{\Gamma(a-1)} = \ldots = \\
&= (a)\cdot(a-1)\cdot(a-2)\cdot\ldots\cdot 2\cdot\underbrace{\Gamma(a-(a-1))}_{\Gamma(1)} = \\
&= (a)\cdot(a-1)\cdot(a-2)\cdot\ldots\cdot 2\cdot\underbrace{\Gamma(1)}_{1} = a!
\end{aligned} \tag{5.106}$$

oz.

$$\Gamma(a) = (a-1)! \tag{5.107}$$

Podobno, kot smo pri obravnavi Bernoullijevih poskusov vpeljali Pascalovo porazdelitev, ki predstavlja število uspehov do *k-tega* uspeha, si lahko predstavljamo Gama naključno spremenljivko kot čas do *k-tega* uspeha [Turk]. Podobno kot Poissonova in eksponentna porazdelitev opisujeta Poissonov stohastični proces (glej Dragan, Stohastični procesi v logistiki), je z njim povezana tudi Gama naključna spremenljivka. Imenujejo jo tudi Erlangova ali Pearsonova porazdelitev tretjega tipa [Turk]. Gama naključno spremenljivko lahko obravnavamo kot vsoto eksponentno porazdeljenih naključnih spremenljivk z istim parametrom λ, pri čemer njeno porazdelitev gostote verjetnosti lahko izpeljemo preko konvolucijskih integralov (glej [Turk]). Velja tudi, da Gama porazdelitev preide v **Erlangovo porazdelitev**, če je število *a* naravno število in velja: $\Gamma(a) = (a-1)!$. Slednja se glasi [Montgomery 1, Turk]:

$$\begin{aligned}
&f(x) = \frac{\lambda\cdot e^{-\lambda\cdot x}\cdot(\lambda\cdot x)^{a-1}}{(a-1)!} = \frac{\lambda\cdot e^{-\lambda\cdot x}\cdot(\lambda)^{a-1}(x)^{a-1}}{(a-1)!} = \\
&= \frac{e^{-\lambda\cdot x}\cdot(\lambda)^{a}(x)^{a-1}}{(a-1)!}, \quad x>0 \quad in \quad a = 1,2,3,\ldots
\end{aligned} \tag{5.108}$$

Za gama funkcijo velja še naslednja lastnost [Hsu]:

$$\Gamma\left(\frac{1}{2}\right)=\int_0^\infty x^{\frac{1}{2}-1}\cdot e^{-x}dx=\int_0^\infty x^{-\frac{1}{2}}\cdot e^{-x}dx$$

$$\textit{nova spremenljivka}:\quad y=x^{\frac{1}{2}},dy=\frac{1}{2}x^{-\frac{1}{2}}dx,x=y^2$$

$$\Gamma\left(\frac{1}{2}\right)=\int_0^\infty 2\cdot e^{-y^2}dy=2\int_0^\infty e^{-y^2}dy=\int_{-\infty}^\infty e^{-y^2}dy \qquad (5.109)$$

$$\textit{nova spremenljivka}:\quad y=\frac{z}{\sqrt{2}},dy=\frac{dz}{\sqrt{2}}$$

$$\Gamma\left(\frac{1}{2}\right)=\int_{-\infty}^\infty e^{-\frac{z^2}{2}}\frac{dz}{\sqrt{2}}=\frac{1}{\sqrt{2}}\underbrace{\int_{-\infty}^\infty e^{-\frac{z^2}{2}}dz}_{\sqrt{2\pi}}=\sqrt{\pi}$$

Rodovna funkcija momentov je enaka:

$$M(t)=E\left(e^{t\cdot X}\right)=\int_{-\infty}^{\infty}e^{t\cdot x}\cdot f(x)dx=\int_0^\infty e^{t\cdot x}\cdot\frac{e^{-\frac{1}{b}\cdot x}\cdot(x)^{a-1}}{b^a\cdot\Gamma(a)}dx=$$

$$=\frac{1}{b^a\cdot\Gamma(a)}\int_0^\infty e^{t\cdot x}\cdot e^{-\frac{1}{b}\cdot x}\cdot(x)^{a-1}dx=\frac{1}{b^a\cdot\Gamma(a)}\int_0^\infty e^{-\left(\frac{1}{b}\cdot x-t\cdot x\right)}\cdot(x)^{a-1}dx=$$

$$=\frac{1}{b^a\cdot\Gamma(a)}\int_0^\infty e^{-\left(\frac{1}{b}-t\right)\cdot x}\cdot(x)^{a-1}dx$$

$$\textit{nova spremenljivka}:\ u=\left(\frac{1}{b}-t\right)\cdot x,du=\left(\frac{1}{b}-t\right)\cdot dx \qquad (5.110)$$

$$M(t)=\frac{1}{b^a\cdot\Gamma(a)}\int_0^\infty e^{-u}\cdot\left(\frac{u}{\left(\frac{1}{b}-t\right)}\right)^{a-1}\cdot\frac{du}{\left(\frac{1}{b}-t\right)}=\frac{1}{b^a\cdot\Gamma(a)}\int_0^\infty e^{-u}\cdot\frac{u^{a-1}}{\left(\frac{1}{b}-t\right)^a}du=$$

$$=\frac{1}{b^a\cdot\Gamma(a)\left(\frac{1}{b}-t\right)^a}\int_0^\infty e^{-u}\cdot u^{a-1}du=\frac{1}{b^a\cdot\Gamma(a)\left(\frac{1}{b}-t\right)^a}\Gamma(a)=\frac{1}{b^a\cdot\left(\frac{1}{b}-t\right)^a}$$

Tako dobimo:

$$M(t)=\frac{1}{b^a\cdot\left(\frac{1}{b}-t\right)^a}=\frac{1}{b^a\cdot\left(\frac{1-b\cdot t}{b}\right)^a}=\frac{1}{(1-b\cdot t)^a}=(1-b\cdot t)^{-a} \tag{5.111}$$

Matematično upanje in varianco bi dobili preko odvajanja rodovne funkcije momentov. Kot se izkaže, bi bil rezultat naslednji [Jesenko, Krishnamoorthy]:

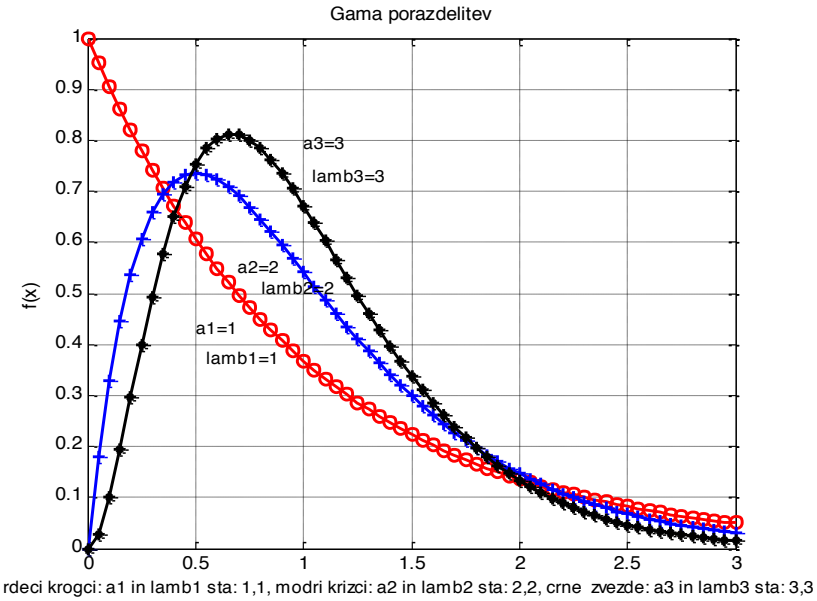
$$\begin{aligned} E(X)&=a\cdot b=a\cdot\frac{1}{\lambda}\\ VAR(x)&=a\cdot b^2=a\cdot\frac{1}{\lambda^2}\end{aligned} \tag{5.112}$$

Izračunajmo še kumulativno funkcijo [Martinez]:

$$\begin{aligned} &F(x)=P(X\le x)=\int_0^x\frac{e^{-\frac{1}{b}\cdot t}\cdot(t)^{a-1}}{b^a\cdot\Gamma(a)}dt=\frac{1}{b^a\cdot\Gamma(a)}\int_0^x e^{-\frac{1}{b}\cdot t}\cdot(t)^{a-1}dt\\ &nova\ \ spremenljivka:\ \ u=\frac{t}{b},du=\frac{dt}{b}\\ &F(x)=\frac{1}{b^a\cdot\Gamma(a)}\int_0^{\frac{x}{b}}e^{-u}\cdot(b\cdot u)^{a-1}\cdot b\cdot du=\frac{b^a}{b^a\cdot\Gamma(a)}\int_0^{\frac{x}{b}}e^{-u}\cdot(u)^{a-1}\cdot du=\\ &=\frac{1}{\Gamma(a)}\int_0^{\frac{x}{b}}e^{-u}\cdot(u)^{a-1}\cdot du=\frac{1}{\Gamma(a)}\int_0^{\lambda\cdot x}e^{-u}\cdot(u)^{a-1}\cdot du,\qquad x>0\end{aligned} \tag{5.113}$$

V nadaljevanju lahko narišemo nekaj primerov Gama porazdelitve (glej sliko 127), pri čemer vzamemo [Martinez]:

$$\begin{aligned} &a=1,\lambda=1,b=\frac{1}{\lambda}=1\\ &a=2,\lambda=2,b=\frac{1}{\lambda}=\frac{1}{2}\\ &a=3,\lambda=3,b=\frac{1}{\lambda}=\frac{1}{3}\end{aligned} \tag{5.114}$$



Slika 127: Trije primeri Gama porazdelitve

Pri izrisu slike 127 smo si pomagali z naslednjim programom v Matlabu:

```
% gama.m
%

clc
clear
close all


x = 0:0.05:3;

% Narisali bomo poteke gama porazdelitve pri treh lambdah in treh a-jih:

a1 = 1
a2 = 2
a3 = 3

lamb1 = 1
lamb2 = 2
lamb3 = 3

y1 = gampdf(x,a1,1/lamb1);
y2 = gampdf(x,a2,1/lamb2);
y3 = gampdf(x,a3,1/lamb3);


% Narisemo porazdelitve:

plot(x,y1,'r','LineWidth',2)
hold on
plot(x,y1,'ro','LineWidth',2)

plot(x,y2,'b','LineWidth',2)
plot(x,y2,'b+','LineWidth',2)

plot(x,y3,'k','LineWidth',2)
plot(x,y3,'k*','LineWidth',2)

title('Gama porazdelitev')
xlabel(['rdeci krogci: a1 in lamb1 sta: ' num2str(a1) ',' num2str(lamb1)...
        ', modri krizci: a2 in lamb2 sta: ' num2str(a2) ',' num2str(lamb2)...
        ', crne  zvezde: a3 in lamb3 sta: ' num2str(a3) ',' num2str(lamb3)],'FontSize',9)
```

```
ylabel('f(x)')
grid

disp('Vnesi a1 na graf!!!!!!!!!!!!!!!')
gtext(['a1=' num2str(a1)],'FontSize',9)
disp('Vnesi lamb1 na graf!!!!!!!!!!!!!!!')
gtext(['lamb1=' num2str(lamb1)],'FontSize',9)


disp('Vnesi a2 na graf!!!!!!!!!!!!!!!')
gtext(['a2=' num2str(a2)],'FontSize',9)
disp('Vnesi lamb2 na graf!!!!!!!!!!!!!!!')
gtext(['lamb2=' num2str(lamb2)],'FontSize',9)

disp('Vnesi a3 na graf!!!!!!!!!!!!!!!')
gtext(['a3=' num2str(a3)],'FontSize',9)
disp('Vnesi lamb3 na graf!!!!!!!!!!!!!!!')
gtext(['lamb3=' num2str(lamb3)],'FontSize',9)
```

Povezava med Gama in eksponentno naključno spremenljivko

Če v izrazu (5.102) postavimo $a = 1$, dobimo:

$$f(x)=\frac{e^{-\frac{1}{b}\cdot x}\cdot(x)^{1-1}}{b^1\cdot\Gamma(1)}=\frac{e^{-\frac{1}{b}\cdot x}}{b}=\frac{1}{b}\cdot e^{-\frac{1}{b}\cdot x}=\lambda\cdot e^{-\lambda\cdot x},\qquad x>0 \tag{5.115}$$

in smo očitno prešli iz gama spremenljivke na eksponentno naključno spremenljivko.

Primer 5.19.:

Potrošnja materiala v nekem proizvodnem procesu je naključni proces. V povprečju se vsak dan porabi 20 komadov. Vsak mesec se nabavi 640 komadov potrošnega materiala, ki predstavlja vrednost parametra a. Naj bo X gama naključna spremenljivka, ki predstavlja čas (v dnevih), v katerem se porabi celotna zaloga. Kolikšna je verjetnost, da zmanjka potrošnega materiala? Koliko mora biti mesečna nabava a^*, da bi bila verjetnost, da zmanjka tekom meseca zaloge, enaka 0.01?

Na osnovi podatkov naloge lahko zapišemo:

$X\in\Gamma(a,\lambda)$
$a=640$
$\lambda=20$
$P(X<30)=?$
$a^*=?$

Imamo:

$$F(x) = P(X \le x) = \frac{1}{\Gamma(a)} \int_0^{\frac{x}{b}} e^{-u} \cdot (u)^{a-1} \cdot du = \frac{1}{\Gamma(a)} \int_0^{\lambda \cdot x} e^{-u} \cdot (u)^{a-1} \cdot du$$
$$F(30) = P(X \le 30) = \frac{1}{\Gamma(640)} \int_0^{20 \cdot x} e^{-u} \cdot (u)^{640-1} \cdot du = \frac{1}{\Gamma(640)} \int_0^{20 \cdot 30} e^{-u} \cdot (u)^{639} \cdot du = \tag{5.116}$$
$$= \frac{1}{\Gamma(640)} \int_0^{600} e^{-u} \cdot (u)^{639} \cdot du$$

Za izračun izraza (5.116) si bomo pomagali z naslednjim programom:

```
% gama kumulativna funkcija (gama1.m):
%
%

clear
clc
close all
dx = 0.01;
xsp = input('xsp=')
xzg = input('xzg=')
a = input('a=')
lamb = input('lamb=')

disp('Verjetnost na izbranem intervalu s standard matlab ukazom je:')

F1 = gamcdf(xzg,a,1/lamb)
```

Komandno okno ima izgled:

```
xsp=-100
xsp =
  -100

xzg=30
xzg =
    30

a=640
a =
   640

lamb=20
lamb =
    20

Verjetnost na izbranem intervalu s standard matlab ukazom je:
F1 =
    0.0546
```

Dobimo torej rezultat:

$$F(30) = P(X \le 30) = 0.0546 \tag{5.117}$$

Torej je verjetnost, da zmanjka potrošnega materiala, enaka 0.0546.

Izračunajmo še, koliko mora biti mesečna nabava a^*, da bi bila verjetnost, da zmanjka tekom meseca zaloge, enaka 0.01:

$$F(30) = P(X \le 30) = \frac{1}{\Gamma\left(a^*\right)} \int_0^{20 \cdot x} e^{-u} \cdot \left(u\right)^{a^*-1} \cdot du = 0.01 \qquad (5.118)$$

$$\Rightarrow \quad a^* = ?$$

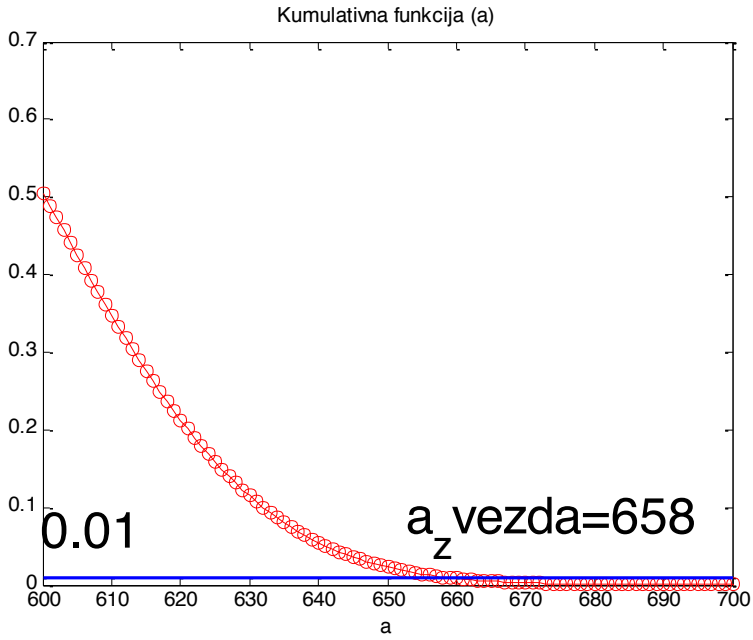
Pri izračunu si pomagamo z naslednjim programom v Matlabu (glej sliko 128):

```
% gama kumulativna funkcija (gama2.m):
%
%
clear
clc
close all
dx = 0.01;
xsp = input('xsp=')
xzg = input('xzg=')
lamb = input('lamb=')
disp('- Izracun s standard matlab ukazi, gremo preko mnozice moznih a-jev od 600 do 700')
disp('      ')
a = [600:1:700];
F1 = [];

for i=a
   F1 = [F1; gamcdf(xzg,i,1/lamb)];
end

disp('Vrednost a-ja, da bo verjetnost P<30 cimbolj enaka 0.01, je enaka:')
aopt=599 + find(abs(F1-0.01)<3e-4)

disp('Pri njem je verjetnost P<30 enaka:')
P=F1(aopt-599)

plot(a,F1,'r','LineWidth',1)
hold on
plot(a,F1,'ro','LineWidth',1)
title('Kumulativna funkcija (a)')
xlabel('a')
x1=600;
x2=700;
x=linspace(x1,x2);
y=0.01*ones(100,1);
plot(x,y,'LineWidth',2)
```

```
disp('- vnesi na graf a_zvezda in verjetnost 0.01!!!')
gtext(['a_zvezda=' num2str(aopt)],'FontSize',24)
gtext(['0.01'],'FontSize',24)
```

Izgled komandnega okna je naslednji:

```
xsp=-100
xsp =
  -100
xzg=30
xzg =
    30
lamb=20
lamb =
    20
- Izracun s standard matlab ukazi, gremo preko mnozice moznih a-jev od 600 do 700

Vrednost a-ja, da bo verjetnost P<30 cimbolj enaka 0.01, je enaka:
aopt =
   658

Pri njem je verjetnost P<30 enaka:
P =
    0.0102
- vnesi na graf a_zvezda in verjetnost 0.01!!!
```



Slika 128: Iskanje tiste vrednosti a^**, pri kateri bo verjetnost* $F(30) = P(X \le 30)$ *enaka 0.01. Izkaže se, da je to vrednost 658, kjer* $F(30) = P(X \le 30)$ *doseže vrednost 0.0102.*

Torej mora biti mesečna nabava $a^* \approx 658$ kosov, da bi bila verjetnost, da zmanjka tekom meseca zaloge, enaka $F(30) = P(X \leq 30) = 0.01$. Če primerjamo s prejšnjo situacijo, mora torej biti mesečna nabava večja za 18 kosov, da verjetnost zmanjkanja zalog zmanjšamo iz 0.0546 na 0.0102.

5.2.5 Hi kvadrat porazdelitev

Za statistiko še posebej pomembna je takoimenovana Hi kvadrat porazdelitev χ^2, ki jo dobimo iz Gama porazdelitve, če postavimo: $a = \frac{n}{2}, b = \frac{1}{\lambda} = 2$. Potem dobimo [Jesenko, Krishnamoorthy]:

$$f(x) = \frac{e^{-\frac{1}{2}\cdot x} \cdot (x)^{\frac{n}{2}-1}}{2^{\frac{n}{2}} \cdot \Gamma\left(\frac{n}{2}\right)} = \frac{e^{-\frac{1}{2}\cdot x} \cdot (x)^{\frac{n-2}{2}}}{2^{\frac{n}{2}} \cdot \Gamma\left(\frac{n}{2}\right)}, \quad x > 0 \tag{5.119}$$

Rodovna funkcija momentov je enaka:

$$M(t) = \frac{1}{2^{\frac{n}{2}} \cdot \left(\frac{1}{2} - t\right)^{\frac{n}{2}}} = \frac{1}{2^{\frac{n}{2}} \cdot \left(\frac{1-2\cdot t}{2}\right)^{\frac{n}{2}}} = \frac{1}{(1-2\cdot t)^{\frac{n}{2}}} = (1-2\cdot t)^{-\frac{n}{2}} \tag{5.120}$$

Matematično upanje in varianca sta enaka:

$$E(X) = a \cdot b = \frac{n}{2} \cdot 2 = n$$
$$VAR(x) = a \cdot b^2 = \frac{n}{2} \cdot 2^2 = 2 \cdot n \tag{5.121}$$

Število *n* = *df* se imenuje število prostostnih stopenj ali kar stopnje prostosti [Jesenko].

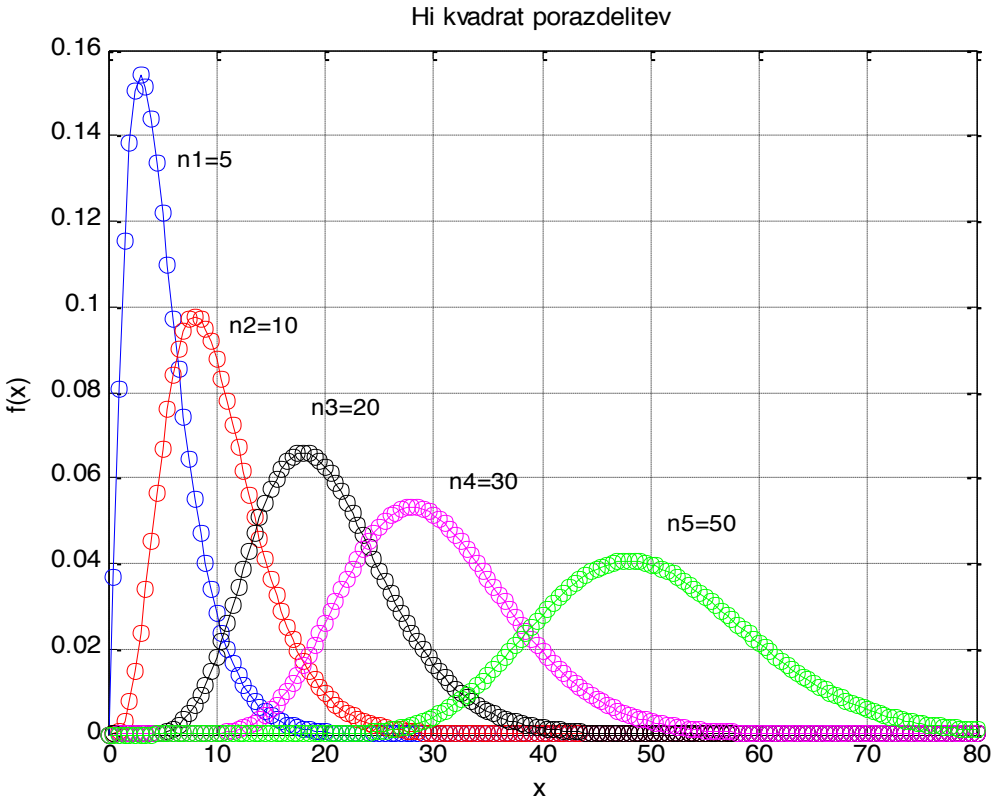
Hi kvadrat porazdelitev se uporablja za določanje porazdelitev vzorčnih varianc in je pomembna pri takoimenovanih "goodness of fit" testih [Martinez].

Denimo imamo set neodvisnih naključnih spremenljivk $X_1, X_2, ..., X_n$, ki so vse **normalno** porazdeljene. Potem velja, da ima naključna spremenljivka X:

$$X = X_1^2 + X_2^2 + ... + X_n^2 \tag{5.122}$$

hi kvadrat porazdelitev [Krishnamoorthy].

V nadaljevanju lahko narišemo nekaj primerov Hi kvadrat porazdelitve (glej sliko 129), pri čemer vzamemo [Krishnamoorthy]: $n = 5, 10, 20, 30, 50$.



Slika 129: Nekaj primerov Hi kvadrat porazdelitve

Pri izrisu slike 129 smo si pomagali z naslednjim programom v Matlabu:

```
% chikvad.m
%

clc
clear
close all


x = 0:0.5:80;

% Narisali bomo poteke hi kvadrat porazdelitve pri:

n1 = 5
n2 = 10
n3 = 20
n4 = 30
n5 = 50

y1 = chi2pdf(x,n1);
y2 = chi2pdf(x,n2);
y3 = chi2pdf(x,n3);
y4 = chi2pdf(x,n4);
y5 = chi2pdf(x,n5);

% Narisemo porazdelitve:

plot(x,y1,'b','LineWidth',1)
hold on
plot(x,y1,'bo','LineWidth',1)

plot(x,y2,'r','LineWidth',1)
plot(x,y2,'ro','LineWidth',1)

plot(x,y3,'k','LineWidth',1)
plot(x,y3,'ko','LineWidth',1)

plot(x,y4,'m','LineWidth',1)
plot(x,y4,'mo','LineWidth',1)

plot(x,y5,'g','LineWidth',1)
plot(x,y5,'go','LineWidth',1)


title('Hi kvadrat porazdelitev')
xlabel('x')
ylabel('f(x)')
grid

disp('Vnesi n1 na graf (modra)!!!!!!!!!!!!!!!')
gtext(['n1=' num2str(n1)],'FontSize',9)

disp('Vnesi n2 na graf (rdeca)!!!!!!!!!!!!!!!')
gtext(['n2=' num2str(n2)],'FontSize',9)

disp('Vnesi n3 na graf (crna)!!!!!!!!!!!!!!!')
gtext(['n3=' num2str(n3)],'FontSize',9)

disp('Vnesi n4 na graf (magenta)!!!!!!!!!!!!!!!')
gtext(['n4=' num2str(n4)],'FontSize',9)

disp('Vnesi n5 na graf (zelena)!!!!!!!!!!!!!!!')
gtext(['n5=' num2str(n5)],'FontSize',9)
```

Primer 5.20.:

Pri vrednosti prostostne stopnje $n = df = 13$ *izračunajte verjetnost* $P(X \le 12.3)$.

Za izračun si bomo pomagali z naslednjim programom v Matlabu:

```
% chi kvadrat kumulativna funkcija (chikvad1.m):
%
%

clear
clc
close all
dx = 0.1;
xsp = input('xsp=')
xzg = input('xzg=')
n = input('n=')

disp('Verjetnost na izbranem intervalu s standard matlab ukazom je:')

F1 = chi2cdf(xzg,n)
```

Izpis komandnega okna je naslednji:

```
xsp=-100
xsp =
 -100

xzg=12.3
xzg =
  12.3000

n=13
n =
   13

Verjetnost na izbranem intervalu s standard matlab ukazom je:
F1 =
   0.4968
```

Torej je verjetnost $P(X \le 12.3)$ enaka 0.4968.

Primer 5.21.:

Dan imamo vzorec, ki je porazdeljen s hi kvadrat naključno spremenljivko $X \in \chi^2(x,n) = \chi^2(x,df) = \chi^2(x,20)$. *Od katere vrednosti* x_p *naprej bo verjetnost* $P(X > x_p) = 1 - P(X \le x_p) = 0.1$?

Vemo, da je $n = df = 20$ in $P(X \le x_p) = 1 - 0.1 = 0.9$. Pri izračunu si pomagamo z naslednjim programom v Matlabu:

```
% chi kvadrat kumulativna funkcija (chikvad2.m):
%
%

clear
clc
close all

n = input('prostostna stopnja n=')
F = input('kumulativna verjetnost F=')

disp('iskani x s standard matlab ukazom je enak:')

x_isk = chi2inv(F,n)
```

Izpis komandnega okna je naslednji:

```
prostostna stopnja n=20
n =
   20
kumulativna verjetnost F=0.9
F =
   0.9000
iskani x s standard matlab ukazom je enak:
x_isk =
  28.4120
```

Torej bo od vrednosti $x_p = 28.4120$ naprej verjetnost $P\left(X > x_p\right) = 1 - P\left(X \le x_p\right) = 0.1.$

Preizkus: Pokličemo program **chikvad1.m** in preverimo, če je prav. Izpis komandnega okna je naslednji:

```
xsp=-100
xsp =
  -100

xzg=28.41
xzg =
  28.4100

n=20
n =
   20

Verjetnost na izbranem intervalu s standard matlab ukazom je:
F1 =
   0.9000
```

Torej se očitno nismo zmotili pri izračunu, saj pride $P\left(X \le x_p\right) = P\left(X \le 28.41\right) = 0.9$

Primer 5.22.:

Denimo, da vemo, da je $P(X \leq 6) = 0.8$. *Izračunajte vrednost prostostne stopnje.*

Pri izračunu si pomagamo z naslednjim programom v Matlabu:

```
% chi kvadrat kumulativna funkcija (chikvad3.m):
%
%

clear
clc
close all

xsp = -100;
xzg = input('xzg=')
F = input('F=')

n = 1:0.1:100;

for i = 1:length(n)
   F1(i) = chi2cdf(xzg,n(i));
end

disp('Vrednost prostostne stopnje je enaka:')

nopt_t = n(find(abs(F1-F)<0.7*1e-2))
```

Izpis komandnega okna je naslednji:

```
xzg=6
xzg =
    6

F=0.8
F =
   0.8000

Vrednost prostostne stopnje je enaka:
nopt_t =
    4
```

Torej je vrednost prostostne stopnje enaka: $n = df = 4$.

Preizkus: Če pokličemo program **chikvad2.m**, dobimo:

```
prostostna stopnja n=4
n =
    4

kumulativna verjetnost F=0.8
F =
   0.8000

iskani x s standard matlab ukazom je enak:
x_isk =
   5.9886
```

Če pa pokličemo program **chikvad1.m**, dobimo:

```
xsp=-100
xsp =
  -100

xzg=6
xzg =
     6

n=4
n =
     4

Verjetnost na izbranem intervalu s standard matlab ukazom je:
F1 =
    0.8009
```

5.2.6 Beta porazdelitev

Naključna spremenljivka se imenuje **Beta naključna spremenljivka** s parametroma $(a>0, b>0)$, če ima njena porazdelitev gostote verjetnosti naslednjo obliko [Jesenko, Krishnamoorthy]:

$$f(x)=\frac{\Gamma(a+b)}{\Gamma(a)\cdot\Gamma(b)}(x)^{a-1}(1-x)^{b-1}, \qquad 0<x<1 \tag{5.123}$$

Beta naključna spremenljivka temelji na takoimenovani **Beta funkciji**, za katero velja [Jesenko, Krishnamoorthy]:

$$B(a,b)=\int_0^1 x^{a-1}\cdot(1-x)^{b-1}\,dx=\frac{\Gamma(a)\cdot\Gamma(b)}{\Gamma(a+b)} \tag{5.124}$$

Torej je porazdelitev gostote verjetnosti enaka:

$$f(x)=\frac{1}{B(a,b)}(x)^{a-1}(1-x)^{b-1}, \qquad 0<x<1 \tag{5.125}$$

Matematično upanje je enako [Jesenko, Krishnamoorthy]:

$$\begin{aligned} E(x) &= \int_{-\infty}^{\infty} x \cdot f(x)\,dx = \int_{0}^{1} x \cdot \frac{\Gamma(a+b)}{\Gamma(a)\cdot\Gamma(b)} (x)^{a-1}(1-x)^{b-1}\,dx = \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\cdot\Gamma(b)} \int_{0}^{1} (x)^{a}(1-x)^{b-1}\,dx = \frac{\Gamma(a+b)}{\Gamma(a)\cdot\Gamma(b)} B(a+1,b) = \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\cdot\Gamma(b)} \cdot \frac{\Gamma(a+1)\cdot\Gamma(b)}{\Gamma(a+1+b)} = \frac{\Gamma(a+b)}{\Gamma(a)} \cdot \frac{\Gamma(a+1)}{\Gamma(a+1+b)} = \\ &= \frac{\Gamma(a+b)}{\Gamma(a)} \cdot \frac{a\cdot\Gamma(a)}{(a+b)\Gamma(a+b)} = \frac{a}{(a+b)} \end{aligned} \tag{5.126}$$

Za izračun variance moramo najprej izračunati drugi moment:

$$\begin{aligned} E(x^2) &= \int_{-\infty}^{\infty} x^2 \cdot f(x)\,dx = \int_{0}^{1} x^2 \cdot \frac{\Gamma(a+b)}{\Gamma(a)\cdot\Gamma(b)} (x)^{a-1}(1-x)^{b-1}\,dx = \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\cdot\Gamma(b)} \int_{0}^{1} (x)^{a+1}(1-x)^{b-1}\,dx = \frac{\Gamma(a+b)}{\Gamma(a)\cdot\Gamma(b)} B(a+2,b) = \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\cdot\Gamma(b)} \cdot \frac{\Gamma(a+2)\cdot\Gamma(b)}{\Gamma(a+2+b)} = \frac{\Gamma(a+b)}{\Gamma(a)} \cdot \frac{\Gamma(a+2)}{\Gamma(a+2+b)} = \\ &= \frac{\Gamma(a+b)}{\Gamma(a)} \cdot \frac{(a+1)\cdot\Gamma(a+1)}{(a+1+b)\Gamma(a+1+b)} = \frac{\Gamma(a+b)}{\Gamma(a)} \cdot \frac{(a+1)\cdot a\cdot\Gamma(a)}{(a+1+b)(a+b)\Gamma(a+b)} = \\ &= \frac{(a+1)\cdot a}{(a+1+b)(a+b)} \end{aligned} \tag{5.127}$$

Varianca je torej enaka:

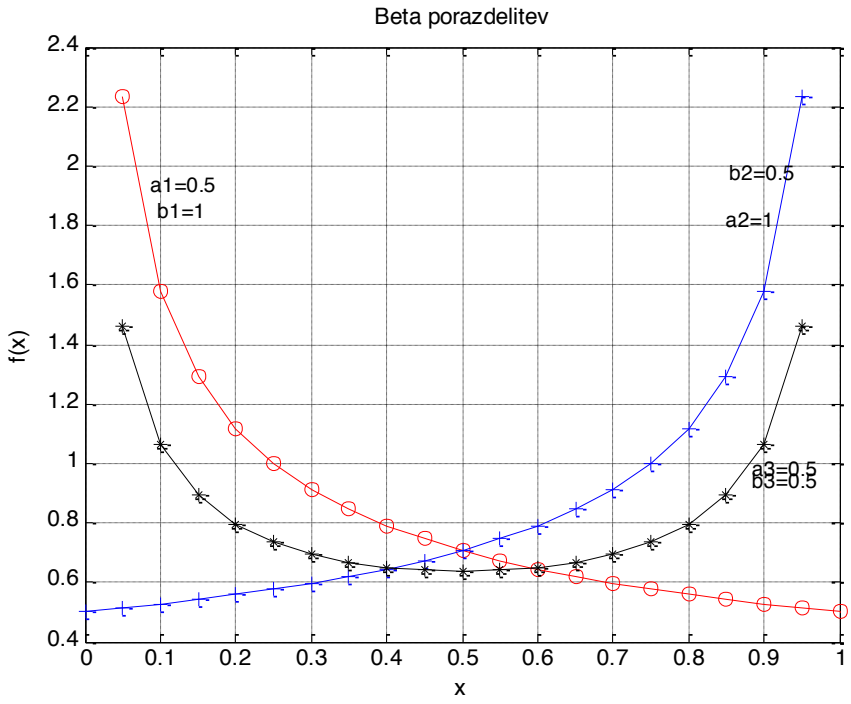
$$\begin{aligned} VAR(X) &= E(X^2) - E^2(X) = \frac{(a+1)\cdot a}{(a+1+b)(a+b)} - \left(\frac{a}{(a+b)}\right)^2 = \\ &= \frac{(a+1)\cdot a}{(a+1+b)(a+b)} - \frac{a^2}{(a+b)^2} \end{aligned} \tag{5.128}$$

Po izpeljavi dobimo [Jesenko, Krishnamoorthy]:

$$VAR(X) = \frac{b \cdot a}{(a+1+b)(a+b)^2} \tag{5.129}$$

V nadaljevanju lahko narišemo nekaj primerov Beta porazdelitve (glej sliko 130), pri čemer vzamemo [Krishnamoorthy]:

$$\begin{aligned} a &= 0.5, b = 1 \\ a &= 1, b = 0.5 \\ a &= 0.5, b = 0.5 \end{aligned} \tag{5.130}$$



Slika 130: Nekaj primerov Beta porazdelitve

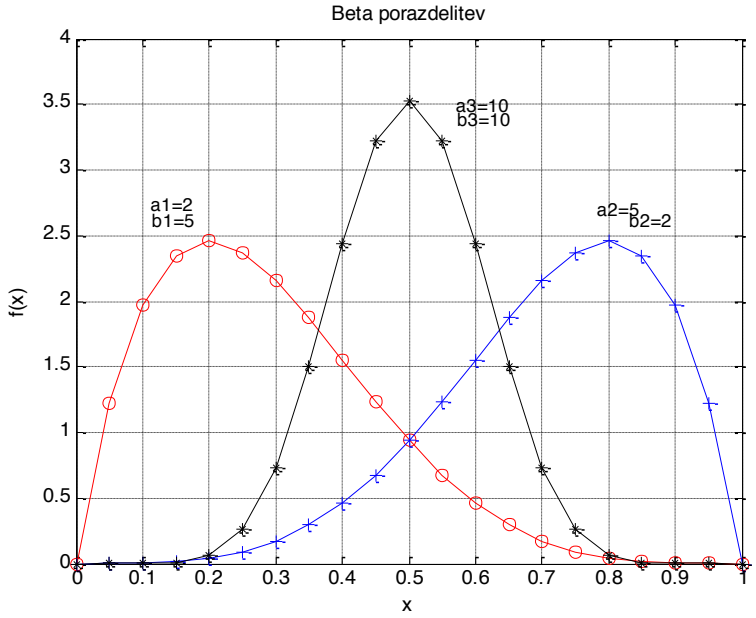
Pri izrisu slike 130 smo si pomagali z naslednjim programom v Matlabu:

```
% beta1.m
%

clc
clear
close all
```

```
x = 0:0.05:1;

% Narisali bomo poteke beta porazdelitve pri treh b-jih in treh a-jih:

a1 = input('a1=')
a2 = input('a2=')
a3 = input('a3=')

b1 = input('b1=')
b2 = input('b2=')
b3 = input('b3=')

y1 = betapdf(x,a1,b1);
y2 = betapdf(x,a2,b2);
y3 = betapdf(x,a3,b3);



% Narisemo porazdelitve:

plot(x,y1,'r','LineWidth',1)
hold on
plot(x,y1,'ro','LineWidth',1)

plot(x,y2,'b','LineWidth',1)
plot(x,y2,'b+','LineWidth',1)

plot(x,y3,'k','LineWidth',1)
plot(x,y3,'k*','LineWidth',1)



title('Beta porazdelitev')
xlabel('x')
ylabel('f(x)')
grid

disp('Vnesi a1 na graf!!!!!!!!!!!!!!!!')
gtext(['a1=' num2str(a1)],'FontSize',9)
disp('Vnesi b1 na graf!!!!!!!!!!!!!!!!')
gtext(['b1=' num2str(b1)],'FontSize',9)



disp('Vnesi a2 na graf!!!!!!!!!!!!!!!!')
gtext(['a2=' num2str(a2)],'FontSize',9)
disp('Vnesi b2 na graf!!!!!!!!!!!!!!!!')
gtext(['b2=' num2str(b2)],'FontSize',9)

disp('Vnesi a3 na graf!!!!!!!!!!!!!!!!')
gtext(['a3=' num2str(a3)],'FontSize',9)
disp('Vnesi b3 na graf!!!!!!!!!!!!!!!!')
gtext(['b3=' num2str(b3)],'FontSize',9)
```

V nadaljevanju narišimo še nekaj primerov Beta porazdelitve (glej sliko 131), pri čemer vzamemo [Krishnamoorthy]:

$$\begin{aligned} a &= 2, b = 5 \\ a &= 5, b = 2 \\ a &= 10, b = 10 \end{aligned} \tag{5.131}$$



Slika 131: Še nekaj primerov Beta porazdelitve

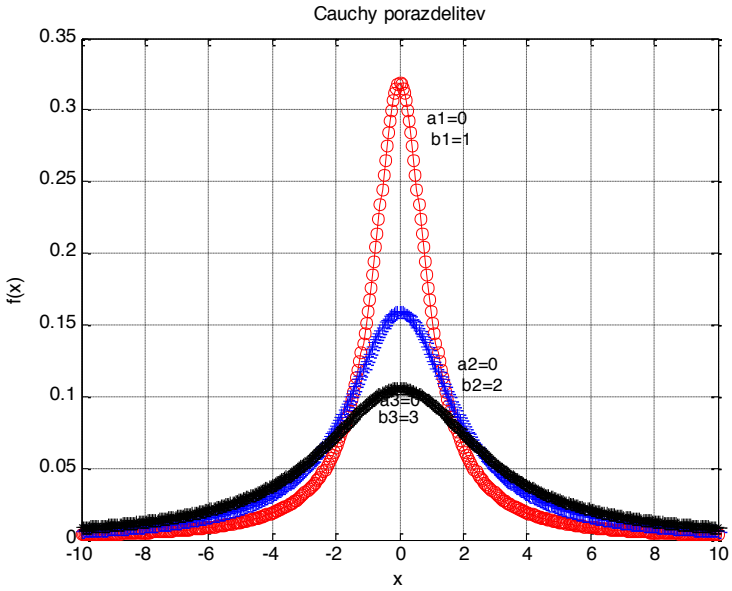
5.2.7 Cauchyjeva porazdelitev

Naključna spremenljivka se imenuje **Cauchyjeva naključna spremenljivka** s parametroma $(a, b>0)$, če ima njena porazdelitev gostote verjetnosti naslednjo obliko:

$$f(x)=\frac{\frac{b}{\pi}}{(x-a)^2+b^2}, \qquad -\infty<x<\infty$$
$$f(x)=\frac{\frac{b}{\pi}}{(x-a)^2+b^2}=\frac{1}{b^2}\cdot\frac{\frac{b}{\pi}}{\left(\frac{x-a}{b}\right)^2+1}=\frac{\frac{1}{b\cdot\pi}}{\left(\frac{x-a}{b}\right)^2+1} \tag{5.132}$$

Pri tem je *a* lokacijski parameter, ki določa vrh porazdelitve, *b* je pa skalirni parameter. Kot se izkaže, srednja vrednost in momenti sploh ne obstajajo, mediana in modus pa sta enaka parametru *a*.

V nadaljevanju lahko narišemo nekaj primerov Cauchyjeve porazdelitve (glej sliko 132), pri čemer vzamemo *a = 0*, ter *b = 1, 2, 3* [Krishnamoorthy].



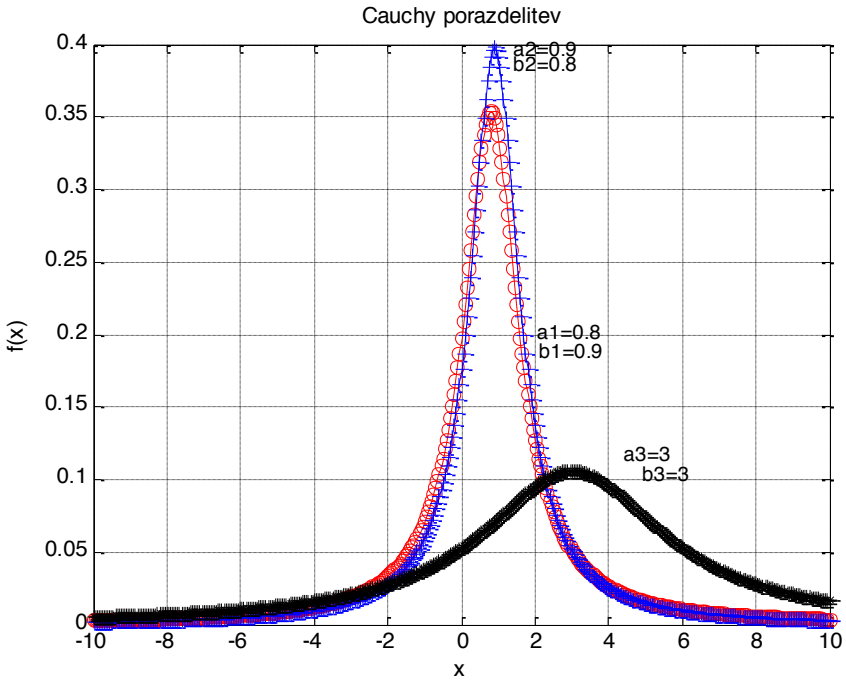
Slika 132: Trije primeri Cauchyjeve porazdelitve

V nadaljevanju narišimo še en primer Cauchyjeve porazdelitve (glej sliko 133), pri čemer vzamemo:

$a = 0.8, b = 0.9$

$a = 0.9, b = 0.8$

$a = 3, b = 3$



Slika 133:Še trije primeri Cauchyjeve porazdelitve

Pri izrisu slik 132 in 133 smo si pomagali z naslednjim programom v Matlabu:

```
% cauchy.m
%

clc
clear
close all



x = -10:0.05:10;

% Narisali bomo poteke cauchy porazdelitve pri treh b-jih in treh a-jih:

a1 = input('a1=')
a2 = input('a2=')
a3 = input('a3=')

b1 = input('b1=')
b2 = input('b2=')
```

```
b3 = input('b3=')



for i=1:length(x)
   y1(i)=b1/((x(i)-a1)^2+b1^2)/pi;
   y2(i)=b2/((x(i)-a2)^2+b2^2)/pi;
   y3(i)=b3/((x(i)-a3)^2+b3^2)/pi;
end



% Narisemo porazdelitve:

plot(x,y1,'r','LineWidth',1)
hold on
plot(x,y1,'ro','LineWidth',1)

plot(x,y2,'b','LineWidth',1)
plot(x,y2,'b+','LineWidth',1)

plot(x,y3,'k','LineWidth',1)
plot(x,y3,'k*','LineWidth',1)



title('Cauchy porazdelitev')
xlabel('x')
ylabel('f(x)')
grid

disp('Vnesi a1 na graf!!!!!!!!!!!!!!!!')
gtext(['a1=' num2str(a1)],'FontSize',9)
disp('Vnesi b1 na graf!!!!!!!!!!!!!!!!')
gtext(['b1=' num2str(b1)],'FontSize',9)



disp('Vnesi a2 na graf!!!!!!!!!!!!!!!!')
gtext(['a2=' num2str(a2)],'FontSize',9)
disp('Vnesi b2 na graf!!!!!!!!!!!!!!!!')
gtext(['b2=' num2str(b2)],'FontSize',9)

disp('Vnesi a3 na graf!!!!!!!!!!!!!!!!')
gtext(['a3=' num2str(a3)],'FontSize',9)
disp('Vnesi b3 na graf!!!!!!!!!!!!!!!!')
gtext(['b3=' num2str(b3)],'FontSize',9)
```

Izračunajmo še kumulativno funkcijo za Cauchyjevo porazdelitev [Krishnamoorthy]:

$$F(x)=P(X\le x)=\int_{-\infty}^{x}\frac{\frac{b}{\pi}}{(t-a)^2+b^2}dt=\frac{b}{\pi}\int_{-\infty}^{x}\frac{1}{(t-a)^2+b^2}dt=$$

$$\frac{1}{b\cdot\pi}\int_{-\infty}^{x}\frac{1}{\left(\frac{t-a}{b}\right)^2+1}dt \qquad (5.133)$$

$$\text{nova spremenljivka } u=\frac{t-a}{b},\ du=\frac{dt}{b}$$

$$F(x)=\frac{1}{b\cdot\pi}\int_{-\infty}^{\frac{x-a}{b}}\frac{1}{(u)^2+1}\cdot b\cdot du=\frac{1}{\pi}\left[\arctan(u)\right]_{-\infty}^{\frac{x-a}{b}}=$$

$$=\frac{1}{\pi}\left(\arctan\left(\frac{x-a}{b}\right)+\frac{\pi}{2}\right)=\frac{1}{\pi}\cdot\arctan\left(\frac{x-a}{b}\right)+\frac{1}{2}$$

Primer 5.23.:

Dana imamo parametra a = 1, b = 2. Izračunajte $F(1.2)=P(X\le 1.2)$.

Imamo:

$$F(x)=P(X\le x)=\frac{1}{\pi}\cdot\arctan\left(\frac{x-a}{b}\right)+\frac{1}{2}$$

$$F(1.2)=P(X\le 1.2)=\frac{1}{\pi}\cdot\arctan\left(\frac{1.2-1}{2}\right)+\frac{1}{2}=\frac{1}{\pi}\cdot\arctan\left(\frac{0.2}{2}\right)+\frac{1}{2}= \qquad (5.134)$$

$$=\frac{1}{\pi}\cdot\arctan\left(\frac{1}{10}\right)+\frac{1}{2}=0.5317$$

Pri izračunu si lahko pomagamo z naslednjim programom v Matlabu:

```
% cauchy2.m
clear
clc
close all
a=input('a=')
b=input('b=')
```

```
xzg = input('xzg=')

F=0.5+atan((xzg-a)/b)/pi;
disp('Verjetnost je:')
F
```

Izpis komandnega okna je naslednji:

```
a=1
a =
     1
b=2
b =
     2
xzg=1.2
xzg =
    1.2000

Verjetnost je:
F =
    0.5317
```

6 PORAZDELITVE VZORČNIH STATISTIK

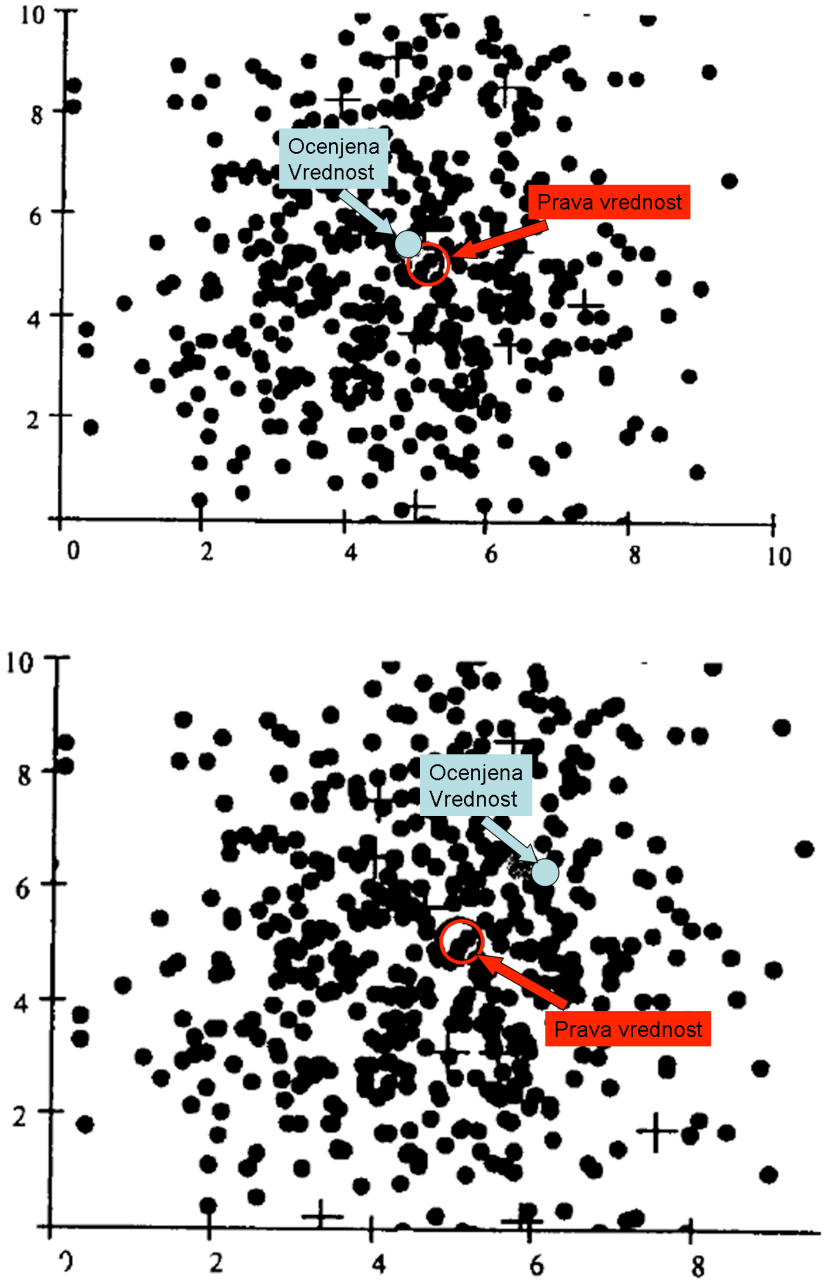
6.1 Naključni vzorci

Statistično sklepanje se v glavnem nanaša na zaključke, ki izhajajo iz naključnih izidov skrbno načrtovanih poskusov. Naključni izidi so ponavadi podmnožice oz. vzorci meritev ali opazovanj iz večje množice, imenovane populacija [Jesenko].

Ni si težko predstavljati, da vsak vzorec ne vodi k pravilnim zaključkom, ki se nanašajo na celotno populacijo, iz katere vzorci izhajajo. Kljub temu pa večinokrat zaključki, narejeni s pomočjo statističnih metod, vodijo k pravilnim odločitvam [Jesenko].

Slika 134 prikazuje računalniško simulacijo, kjer smo izbrali naključni vzorec iz populacije in izračunali določen parameter. Pri tem so elementi poluplacije prikazani z malimi pikami, elementi vzorca pa s križci. Izbrali smo dva različna, enako velika vzorca. Kot je očitno iz slike 134, v prvem primeru ocenjena vrednost parametra leži dokaj blizu pravi vrednosti, v drugem primeru pa je ocenjena vrednost občutno oddaljena od prave vrednosti. Problem torej je, kako se sploh lahko zanesemo na ocene, kako jih lahko izboljšamo, itn [Jesenko].

V praksi so običajno populacije, iz katerih izbiramo vzorce, končne, vendar kljub temu tako velike, da jih lahko obravnavamo kot neskončne.



Slika 134: Ocenjena in prava vrednost parametra populacije pri dveh različnih vzorcih enake velikosti [Jesenko].

Poglejmo si naslednjo definicijo [Jesenko]:

Če so $X_1,...,X_n$ **neodvisne** in **enako porazdeljene** naključne spremenljivke, potem sestavljajo **naključni vzorec** iz neskončne populacije z dano skupno porazdelitvijo. Če so $x_1,...,x_n$ realizacije naključnih spremenljivk $X_1,...,X_n$, ter je $p(x_1,...,x_n)$ njihov multidimenzionalni porazdelitveni zakon, slednjega lahko zapišemo kot [Jesenko]:

$$p(x_1,...,x_n) = p(x_1)\cdot p(x_2)\cdot ...\cdot p(x_n) \tag{6.1}$$

Poznamo dve zelo znameniti statistiki, aritmetično sredino vzorca in varianco vzorca, ki se glasita [Jesenko, Košmelj K., Elezović, Ross]:

$$\begin{aligned} \bar{X} &= \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}(X_1 + ... + X_n) \\ VAR(\bar{X}) &= S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \end{aligned} \tag{6.2}$$

Ko vzorec dejansko izberemo in pridemo do realizacije naključnih spremenljivk $X_1,...,X_n$, potem lahko izračunamo nepristranski vrednosti obeh statistik na naslednji način [Jesenko]:

$$\begin{aligned} \bar{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + ... + x_n) \\ VAR(\bar{x}) &= s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \end{aligned} \tag{6.3}$$

V poglavju 4.4.1 (glej izraza (4.69) in (4.70)) smo omenili, da deljenje z *n* daje pristransko varianco, zato moramo deliti z *n - 1*.

6.2 Porazdelitev aritmetične sredine vzorcev

Statistike so naključne spremenljivke, zato se njihove realizacije spreminjajo od vzorca do vzorca. Pri statističnem sklepanju so zelo pomembni porazdelitveni zakoni teh naključnih spremenljivk, pravimo pa jim **vzorčne porazdelitve** [Jesenko].

Denimo sta $\mu = E(X)$ in $\sigma^2 = VAR(X)$ aritmetična sredina in varianca neke populacije. Če je $X_1,...,X_n$ nek vzorec iz te populacije, potem velja:

$$\begin{aligned}
&E(\bar{X}) = \mu_{\bar{X}} = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}E(X_1 + ... + X_n) = \frac{1}{n}E(X_1) + ... + \frac{1}{n}E(X_n) = \\
&= \frac{1}{n}\mu + ... + \frac{1}{n}\mu = \frac{1}{n}\cdot n \cdot \mu = \mu \\
&VAR(\bar{X}) = \sigma^2_{\bar{X}} = VAR\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}VAR(X_1 + ... + X_n) = \\
&= \frac{1}{n^2}VAR(X_1) + ... + \frac{1}{n^2}VAR(X_n) = \\
&= \frac{1}{n^2}\sigma^2 + ... + \frac{1}{n^2}\sigma^2 = \frac{1}{n^2}\cdot n \cdot \sigma^2 = \frac{\sigma^2}{n} \\
&\sigma_{\bar{X}} = \sqrt{VAR(\bar{X})} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}
\end{aligned} \tag{6.4}$$

Kot vidimo, je $E(\bar{X}) = \mu$ in je zato ta ocena nepristranska. Standardno deviacijo $\sigma_{\bar{X}}$ imenujemo **standardna ocena napake aritmetične sredine**. Kot vidimo, se le-ta zmanjšuje, ko večamo velikost vzorca, pri čemer se aritmetična sredina vzorca vse bolj bliža aritmetični sredini populacije.

Če bi bilo matematično upanje populacije znano, potem bi lahko varianco izračunali na naslednji način [Elezović]:

$$VAR(\bar{X}) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 \tag{6.5}$$

Pokažimo, da je tovrstna varianca nepristranska [Elezović]:

$$
\begin{aligned}
&E\left(VAR\left(\bar{X}\right)\right)=E\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\mu\right)^2\right)=\frac{1}{n}\sum_{i=1}^{n}E\left(X_i-\mu\right)^2=\\
&=\frac{1}{n}\sum_{i=1}^{n}E\left(X_i^2-2X_i\cdot\mu+\mu^2\right)=\\
&=\frac{1}{n}\sum_{i=1}^{n}\left[E\left(X_i^2\right)-2\mu\cdot E\left(X_i\right)+\mu^2\right]=\frac{1}{n}\sum_{i=1}^{n}\left[E\left(X_i^2\right)-\mu^2\right]=\\
&=\frac{1}{n}\sum_{i=1}^{n}\left[VAR\left(X_i\right)+E^2\left(X_i\right)-\mu^2\right]=\frac{1}{n}\sum_{i=1}^{n}\left[VAR\left(X_i\right)+\mu^2-\mu^2\right]=\\
&=\frac{1}{n}\sum_{i=1}^{n}\left[VAR\left(X_i\right)\right]=\frac{1}{n}\left(\sigma^2+\ldots+\sigma^2\right)=\sigma^2
\end{aligned}
\tag{6.6}
$$

Kot se izkaže, disperija nepristranske statistike $VAR\left(\bar{X}\right)=\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\mu\right)^2$ gre proti 0, ko velikost vzorca narašča, zato je ta statistika veljavna [Elezović].

Pokažimo še, zakaj je statistika $VAR\left(\bar{X}\right)=\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2$ pristranska. Tvorimo matematično upanje te statistike:

$$
\begin{aligned}
&E\left(VAR\left(\bar{X}\right)\right)=E\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2\right)=\frac{1}{n}\sum_{i=1}^{n}E\left[\left(X_i-\bar{X}\right)^2\right]=\\
&=\frac{1}{n}\sum_{i=1}^{n}\left(\left[E\left(X_i-\bar{X}\right)\right]^2+VAR\left(X_i-\bar{X}\right)\right)=\\
&=\frac{1}{n}\sum_{i=1}^{n}\left(\left[E\left(X_i\right)-\bar{X}\right]^2+VAR\left(X_i-\bar{X}\right)\right)=\\
&=\frac{1}{n}\sum_{i=1}^{n}\left(\left[\mu-\mu\right]^2+VAR\left(X_i-\bar{X}\right)\right)=\frac{1}{n}\sum_{i=1}^{n}VAR\left(X_i-\bar{X}\right)=\\
&=\frac{1}{n}\sum_{i=1}^{n}VAR\left(X_i-\frac{1}{n}\sum_{j=1}^{n}X_j\right)=\frac{1}{n}\sum_{i=1}^{n}VAR\left(X_i-\frac{1}{n}X_i-\frac{1}{n}\sum_{\substack{j=1\\j\neq i}}^{n}X_j\right)=\\
&=\frac{1}{n}\sum_{i=1}^{n}VAR\left(\frac{n-1}{n}X_i-\frac{1}{n}\sum_{\substack{j=1\\j\neq i}}^{n}X_j\right)=\frac{\left(n-1\right)^2}{n^3}\sum_{i=1}^{n}VAR\left(X_i\right)+\frac{1}{n^3}\sum_{i=1}^{n}VAR\left(\sum_{\substack{j=1\\j\neq i}}^{n}X_j\right)=\\
&=\frac{\left(n-1\right)^2}{n^3}n\cdot\sigma^2+\frac{1}{n^3}\sum_{i=1}^{n}\left(n-1\right)\cdot\sigma^2=\frac{\left(n-1\right)^2}{n^2}\sigma^2+\frac{1}{n^3}\left(n-1\right)\cdot n\cdot\sigma^2
\end{aligned}
\tag{6.7}
$$

Dobimo:

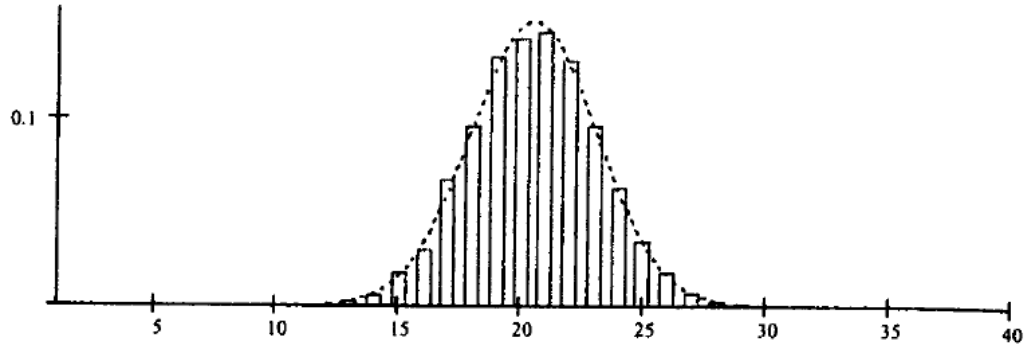
$$\begin{aligned} E\left(VAR\left(\bar{X}\right)\right) &= \frac{(n-1)^2}{n^2}\sigma^2 + \frac{1}{n^3}(n-1)\cdot n\cdot\sigma^2 = \\ &= \frac{\sigma^2}{n^2}\left((n-1)^2 + (n-1)\right) = \frac{\sigma^2}{n^2}(n-1)\left((n-1)+1\right) = \\ &= \frac{\sigma^2}{n^2}(n-1)\cdot n = \frac{n-1}{n}\cdot\sigma^2 \end{aligned} \tag{6.8}$$

Torej očitno dobimo pristransko oceno. Da bi dobili nepristransko oceno variance, bi morali statistiko $VAR\left(\bar{X}\right) = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$ množiti z $\frac{n}{n-1}$. Tako bi dobili nepristransko oceno variance, podane v izrazu (5.137). Izkaže se tudi, da je nepristranska statistika $VAR\left(\bar{x}\right) = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$ veljavna, saj se da dokazati, da gre disperija te statistike proti 0, ko velikost vzorca narašča [Elezović].

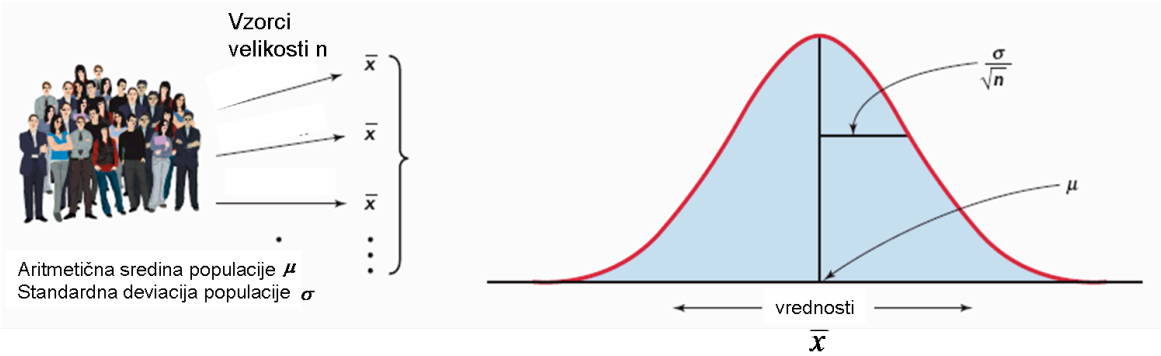
Poudarimo še, da izraz (6.4) velja ne glede na to, kakšen tip porazdelitve ima populacija. Torej ima porazdelitev aritmetične sredine vzorcev srednjo vrednost enako srednji vrednosti populacije, njegova varianca pa je enaka varianci populacije, deljeni s številom elementov v vzorcu (velikostjo vzorca).

Pomen centralnega limitnega izreka

Denimo izvedemo simulacijo, kjer ponavljamo vzorčenje in pri tem računamo aritmetično srdino vsakokratnega vzorca. Seveda vsakič dobimo drugačen rezultat. Nato interval, na katerem se nahajajo dobljene aritmetične sredine, razdelimo na več razredov in narišemo relativni frekvenčni histogram. Rezultat tovrstne simulacije prikazujeta sliki 135 in 136 [Jesenko, Moore].



Slika 135: Ilustracija centralnega limitnega izreka



Slika 136: Še ena ilustracija centralnega limitnega izreka

Iz simulacije je razvidno, da se aritmetična sredina vzorcev porazdeljuje normalno, torej je normalna naključna spremenljivka. ***To velja ne glede na tip porazdelitve populacije, če je velikost vzorca dovolj velika.***

Iz teh dejstev izhaja centralni limitni izrek, ki smo ga na kratko spoznali že v poglavju 2.23 (glej izraze (2.253) do (2.255)). **Centralni limitni izrek** pravi naslednje [Jesenko]:

Če je $X_1,...,X_n$ naključni vzorec iz neskončne **poljubno porazdeljene** populacije z aritmetično sredino μ in standardno deviacijo σ, potem naključna spremenljivka $Z=\dfrac{\left(\bar{X}-\mu\right)}{\dfrac{\sigma}{\sqrt{n}}}$ za **dovolj velik vzorec** ($n\geq 30$) teži k standardizirani normalni naključni spremenljivki: $\lim\limits_{n\to\infty} Z=N(0,1)$. Naključna spremenljivka $\bar{X}$ pa teži k normalni porazdelitvi: $\bar{X}\in N\left(\mu,\dfrac{\sigma}{\sqrt{n}}\right)=N\left(E\left(\bar{X}\right),\sqrt{VAR\left(\bar{X}\right)}\right)$. Če pa imamo normalno porazdeljeno populacijo in velja: $X\in N(\mu,\sigma)$, potem je naključna spremenljivka $Z=\dfrac{\left(\bar{X}-\mu\right)}{\dfrac{\sigma}{\sqrt{n}}}$ enaka standardizirani normalni naključni spremenljivki: $Z=N(0,1)$, **ne glede na velikost vzorca *n***.

Primer 6.1.:

Avtomat za točenje vode je nastavljen tako, da v povprečju natoči 150 ml vode s standardnim odklonom 9 ml. Kakšna je verjetnost, da bo v naključnem vzorcu velikosti 36 ***povprečno*** *natočena količina vode vsaj 148 ml?*

V tem primeru sta varianca in deviacija populacije znani in enaki $\mu = 150,\ \sigma = 9$. Iz podatkov tudi razvidimo, da se zanimamo za srednjo vrednost vzorca: $\bar{x} = 148$. Ker je vzorec dovolj velik ($n = 36$), nas porazdelitev populacije pravzprav sploh ne zanima, saj vemo, da naključna spremenljivka $\bar{X}$ gotovo teži k normalni porazdelitvi. Uporabili bomo centralni limitni izrek in v ta namen izračunali:

$$z = \frac{(\bar{x} - \mu)}{\frac{\sigma}{\sqrt{n}}} = \frac{(148 - 150)}{\frac{9}{\sqrt{36}}} = -\frac{2 \cdot 6}{9} = -\frac{4}{3} \tag{6.9}$$

Sledi:

$$P(\bar{X} \geq \bar{x}) = P(\bar{X} \geq 148) = P\left(Z \geq -\frac{4}{3}\right) = 1 - P\left(Z < -\frac{4}{3}\right) = 0.9032 \tag{6.10}$$

Za izračun smo uporabili naslednje ukaze v Matlabu:

```
x= -100:0.05:-4/3+0.05;
F=normcdf(x,0,1);
F1=1-F(length(F))
```

Porazdelitev aritmetične sredine vzorcev iz končne populacije

Če je $\bar{X}$ aritmetična sredina naključnega vzorca velikosti *n*, izbranega iz končne populacije velikosti *N* z aritmetično sredino μ in deviacijo σ, potem velja naslednje [Jesenko]:

$$E\left(\bar{X}\right)=\mu$$
$$VAR\left(\bar{X}\right)=\frac{\sigma^2}{n}\cdot\frac{N-n}{N-1} \tag{6.11}$$

Kot lahko vidimo, se varianca aritmetične sredine vzorca, izbranega iz končne populacije, razlikuje od variance aritmetične sredine vzorca, izbranega iz neskončne populacije, za korekcijski faktor $\frac{N-n}{N-1}=\frac{1-\frac{n}{N}}{1-\frac{1}{N}}$. Očitno smemo za veliko populacijo ($N>>n$) končno populacijo obravnavati kot neskončno. V praksi to napravimo tedaj, ko velja: $\frac{n}{N}\leq 0.05$.

Problemi z neznanim standardnim odklonom populacije

Centralni limitni izrek pride zelo prav, ko je varianca populacije poznana. Žal pa v praksi večinokrat temu ni tako in variance ne poznamo. Zato moramo neznano varianco σ^2 nadomestiti z varianco vzorca $VAR\left(\bar{x}\right)=s^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2$, ki jo lahko izračunamo. Tako pridemo do takoimenovane *t* naključne spremenljivke, izpeljane na osnovi *z* naključne spremenljivke, ki se pojavi v centralnem limitnem izreku, ter χ^2 naključne spremenljivke, predstavljene v poglavju 5.2.5. Kot bomo videli, se pri kombiniranju teh dveh spremenljivk neznana varianca populacije pokrajša, v *t* porazdelitvi pa se namesto nje pojavi poznana spremenljivka *S*, ki pripada vzorcem.

Še preden pa gremo na obravnavo t naključne spremenljivke, si poglejmo še nekaj lastnosti glede statistik, ki so χ^2 naključne spremenljivke.

6.3 Hi-kvadrat statistike

Velja naslednji **izrek** [Jesenko]:

Če je Z standardizirana normalna naključna spremenljivka ($Z \in N(0,1)$*), potem je* Z^2 *enaka* χ^2 *naključni spremenljivki z eno prostostno stopnjo (n=1). Torej velja:* $Z^2 \in \chi^2(n) = \chi^2(1)$.

Da bi to dokazali, bomo uporabili transformacijsko metodo, predstavljeno v poglavju 2.13. Če je $Z \in N(0,1)$, potem velja:

$$f(z) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{(z)^2}{2}} \tag{6.12}$$

Najprej zapišimo inverzno (transformacijsko) funkcijo, to je:

$$U = h(Z) = Z^2 \quad \Rightarrow \quad u = h(z) = z^2 \quad \Rightarrow \quad z = h^{-1}(u) = \sqrt{u} \tag{6.13}$$

Velja:

$$f\left(h^{-1}(u)\right) = f\left(\sqrt{u}\right) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{\left(\sqrt{u}\right)^2}{2}} = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{u}{2}} \tag{6.14}$$

Nato tvorimo naslednji odvod:

$$\frac{dh^{-1}(u)}{du} = \frac{dz(u)}{du} = \frac{d}{du}\left(\sqrt{u}\right) = \frac{1}{2\sqrt{u}} \tag{6.15}$$

Ker je spremenljivka u samo pozitivna in je kot transformacijska funkcija simetrična, lahko zapišemo:

$$g(u)=f\left(h^{-1}(u)\right)\cdot 2\cdot\frac{dh^{-1}(u)}{du}= \\ =\frac{1}{\sqrt{2\pi}}\cdot e^{-\frac{u}{2}}\cdot 2\cdot\frac{1}{2\sqrt{u}}=\frac{1}{\sqrt{2\pi u}}\cdot e^{-\frac{u}{2}}=\frac{1}{\sqrt{\pi}\cdot 2^{\frac{1}{2}}}u^{-\frac{1}{2}}\cdot e^{-\frac{u}{2}} \tag{6.16}$$

Po drugi strani velja, da če bi vstavili v funkcijo v izrazu (5.119) vrednost *n = 1*, bi dobili:

$$f(u)=\frac{e^{-\frac{1}{2}\cdot u}\cdot(u)^{\frac{1-2}{2}}}{2^{\frac{1}{2}}\cdot\Gamma\left(\frac{1}{2}\right)}=\frac{e^{-\frac{1}{2}\cdot u}\cdot(u)^{\frac{-1}{2}}}{2^{\frac{1}{2}}\cdot\Gamma\left(\frac{1}{2}\right)}=\frac{e^{-\frac{1}{2}\cdot u}\cdot(u)^{\frac{-1}{2}}}{2^{\frac{1}{2}}\cdot\sqrt{\pi}} \tag{6.17}$$

kjer smo na osnovi izraza (5.109) upoštevali lastnost: $\Gamma\left(\frac{1}{2}\right)=\sqrt{\pi}$. Vidimo, da sta izraza (6.16) in (6.17) enaka, s tem pa smo dokazali izrek.

Velja tudi naslednji **izrek** [Jesenko]:

Če so $Z_i,\ i=1,...,n$ *neodvisne standardizirane normalne naključne spremenljivke (*$Z_i\in N(0,1)$*), potem je spremenljivka* $\eta=\sum_{i=1}^{n}Z_i^2$ *enaka* χ^2 *naključni spremenljivki z n prostostnimi stopnjami. Torej velja:* $\eta=\sum_{i=1}^{n}Z_i^2\in\chi^2(n)$.

Dokaz:

Ker za vsako izmed naključnih spremenljivk $Z_i,\ i=1,...,n$ velja: $Z_i^2\in\chi^2(1)$, za kvadrirane naključne spremenljivke sledi naslednja rodovna funkcija momentov na osnovi izraza (5.120):

$$M_{Z_i^2}(t)=(1-2\cdot t)^{-\frac{n}{2}}=(1-2\cdot t)^{-\frac{1}{2}} \tag{6.18}$$

Za spremenljivko $\eta = \sum_{i=1}^{n} Z_i^2$ pa je rodovna funkcija momentov enaka produktu rodovnih funkcij momentov posameznih kvadriranih naključnih spremenljivk Z_i^2:

$$M_\eta(t) = (1-2\cdot t)^{-\frac{1}{2}} \cdot (1-2\cdot t)^{-\frac{1}{2}} \cdot \ldots \cdot (1-2\cdot t)^{-\frac{1}{2}} = \\ = \left((1-2\cdot t)^{-\frac{1}{2}} \right)^n = (1-2\cdot t)^{-\frac{n}{2}} \tag{6.19}$$

Vidimo, da je ta izraz enak izrazu (5.120), s tem pa smo dokazali izrek.

Velja še en **izrek** [Jesenko]:

Če so $\xi_i,\ i = 1,\ldots,k$ *neodvisne* χ^2 *naključne spremenljivke z* $n_1,\ldots,n_k$ *prostostnimi stopnjami, potem je spremenljivka* $\eta = \sum_{i=1}^{k} \xi_i$ *enaka* χ^2 *naključni spremenljivki z* $n_1 + \ldots + n_k$ *prostostnimi stopnjami. Torej velja:* $\eta = \sum_{i=1}^{n} \xi_i \in \chi^2(n_1 + \ldots + n_k)$.

Prav tako velja **izrek** [Jesenko]:

Denimo sta $\bar{X}$ *in* S^2 *aritmetična sredina in varianca naključnega vzorca velikosti n, izbranega iz normalne populacije z aritmetično sredino* μ *in standardnim odklonom* σ. *Potem je naključna spremenljivka* $\frac{(n-1)\cdot S^2}{\sigma^2}$ *enaka* χ^2 *naključni spremenljivki z n - 1 prostostnimi stopnjami.*

Dokaz:

Pokazati se da, da velja [Jesenko]:

$$\sum_{i=1}^{n} (X_i - \mu)^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2 + n\cdot(\bar{X} - \mu)^2 \tag{6.20}$$

Delimo ta izraz z σ^2 in dobimo:

$$\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2=\frac{1}{\sigma^2}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2+n\cdot\left(\frac{\bar{X}-\mu}{\sigma}\right)^2= \tag{6.21}$$
$$=\frac{1}{\sigma^2}\cdot S^2\cdot(n-1)+\left(\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}\right)^2$$

kjer smo upoštevali tudi izraz (6.2). Sedaj velja naslednje [Jesenko]:

$$\begin{aligned}&1.\quad \sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2\in\chi^2(n)\\ &2.\quad \left(\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}\right)^2\in\chi^2(1)\\ &3.\quad \frac{1}{\sigma^2}\cdot S^2\cdot(n-1)\in\chi^2(n-1)\text{.....(}\textit{se da dokazati}\text{)}\end{aligned} \tag{6.22}$$

Prva dva izraza v izrazu (6.22) sledita iz prejšnjih izrekov. Da velja tudi tretji izraz, če veljata prva dva izraza, pa se da tudi dokazati. Dokaz lahko bralec zasledi v [Jesenko]. S tem je ta izrek dokazan.

Zaradi velike uporabnosti naključne spremenljivke χ^2 so vrednosti $\chi^2(\alpha,n)$ tabelirane za različne stopnje prostosti in verjetnosti α, pri čemer velja [Jesenko]:

$$\begin{aligned}&P\left(\chi^2\geq\chi_{zg}{}^2(\alpha,n)\right)=1-P\left(\chi^2<\chi_{zg}{}^2(\alpha,n)\right)=\alpha\\ &P\left(\chi^2<\chi_{zg}{}^2(\alpha,n)\right)=1-\alpha\end{aligned} \tag{6.23}$$

Iz izraza (6.23) lahko določimo zgornjo mejo oz. kritično vrednost. Tabele so običajno zgrajene za $\alpha=0.995,0.975,0.95,0.05,0.025,0.005$ ter za stopnje prostosti $n=1,2,...,30$. **Kot se izkaže, lahko za večje število prostostnih stopenj spremenljivko χ^2 nadomestimo z normalno naključno spremenljivko**.

Primer 6.2.:

Predpostavimo, da je debelina izdelane pločevine naključna spremenljivka. Menimo, da je proizvodni proces pod nadzorom, če variabilnost debeline pločevine, ki je podana z znanim standardnim odklonom, ni večja od $\sigma = 0.35$ *stotink cm. Za nadzor proizvodnega procesa občasno izbiramo naključne vzorce velikosti* $n = 25$. *Proces uide nadzoru, če velja:*

$$\begin{aligned} &\frac{(n-1)\cdot S^2}{\sigma^2} \geq \chi_{zg}^{\ 2}\left(\alpha,(n-1)\right) \\ &\frac{24\cdot S^2}{0.35^2} \geq \chi_{zg}^{\ 2}\left(0.01,24\right) \\ &195.9184\cdot S^2 \geq \chi_{zg}^{\ 2}\left(0.01,24\right) \end{aligned} \tag{6.24}$$

pri čemer smo vzeli $\alpha = 0.01$. *Kaj lahko sklepamo o proizvodnem procesu, če je standardni odklon izbranega vzorca enak* $s = 0.41$ *stotink cm?*

Vrednost $\chi_{zg}^{\ 2}\left(0.01,24\right) = 42.979$ dobimo z naslednjim ukazom v Matlabu:

```
>> x=chi2inv(1-0.01,24)
x =
   42.9798
```

pri čemer velja:

$$\begin{aligned} &P\left(\chi^2 < \chi_{zg}^{\ 2}\left(\alpha,n\right)\right) = 1-\alpha \\ &P\left(\chi^2 < \chi_{zg}^{\ 2}\left(0.01,24\right)\right) = 1-0.01 = 0.99 \\ &P\left(\chi^2 < 42.979\right) = 1-0.01 = 0.99 \end{aligned} \tag{6.25}$$

Preverimo v Matlabu, če to res drži:

```
>> F=chi2cdf(-100:0.1:42.979,24);
>> F1=F(length(F))
F1 =
   0.9898
```

Vrednost $\frac{(n-1)\cdot s^2}{\sigma^2}$ je enaka:

$$\frac{(n-1)\cdot s^2}{\sigma^2} = \frac{24}{0.35^2}\cdot s^2 = 195.9184\cdot s^2 = \\ = 195.9184\cdot 0.41^2 = 32.9339 \tag{6.26}$$

Ker je $32.9339 < 42.797$, sklepamo, da proces ni ušel nadzoru.

6.4 Studentova t naključna spremenljivka

Centralni limitni izrek pride zelo prav, ko je varianca populacije poznana. Žal pa v praksi večinokrat temu ni tako in variance ne poznamo. Zato moramo neznano varianco σ^2 nadomestiti z varianco vzorca $VAR(\bar{x}) = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$, ki jo lahko izračunamo. Iz centralnega limitnega izreka vemo, da če je $X_1,...,X_n$ naključni vzorec iz neskončne **poljubno porazdeljene** populacije z aritmetično sredino μ in standardno deviacijo σ, potem naključna spremenljivka $Z = \frac{(\bar{X} - \mu)}{\frac{\sigma}{\sqrt{n}}}$ za **dovolj velik vzorec** ($n \geq 30$) teži k standardizirani normalni naključni spremenljivki: $\lim_{n\to\infty} Z = N(0,1)$. Sedaj pa nastane vprašanje, kako se porazdeli naključna spremenljivka $\frac{(\bar{X} - \mu)}{\frac{S}{\sqrt{n}}}$, ki jo moramo vpeljati, če variance σ^2 ne poznamo.

Izrek:

*Naj bosta χ^2 z n prostostnimi stopnjami in **standardizirana normalna porazdelitev** Z neodvisni naključni spremenljivki. Potem je naključna spremenljivka, definirana z izrazom:*

$$T = \frac{Z}{\sqrt{\frac{\chi^2}{n}}} \tag{6.27}$$

***Studentova t naključna spremenljivka** z n prostostnimi stopnjami, ki ima porazdelitveni zakon enak [Jesenko]:*

$$f(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi \cdot n} \cdot \Gamma\left(\frac{n}{2}\right)} \cdot \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}}, \quad -\infty < t < \infty \tag{6.28}$$

Dokaz:

Ker sta $x \in \chi^2$ in $z \in Z$ neodvisni naključni spremenljivki, bo njuna združena porazdelitev gostote verjetnosti enaka:

$$f(x,z) = f(x) \cdot f(z) = \frac{e^{-\frac{1}{2} \cdot x} \cdot (x)^{\frac{n-2}{2}}}{2^{\frac{n}{2}} \cdot \Gamma\left(\frac{n}{2}\right)} \cdot \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{(z)^2}{2}}, \qquad x > 0, \quad -\infty < z < \infty \tag{6.29}$$

Pokazati se da, da velja naslednja relacija za neko tranformirano gostoto verjetnosti [Jesenko]:

$$g(x,t) = f\left(x, z(x,t)\right) \cdot \frac{\partial}{\partial t}\left[z(x,t)\right] \tag{6.30}$$

če vpeljemo transformacijo: $z(x,t) = t \cdot \sqrt{\frac{x}{n}}$.

Sledi:

$$g(x,t) = \frac{e^{-\frac{1}{2} \cdot x} \cdot (x)^{\frac{n-2}{2}}}{2^{\frac{n}{2}} \cdot \Gamma\left(\frac{n}{2}\right)} \cdot \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{\left(t \cdot \sqrt{\frac{x}{n}}\right)^2}{2}} \cdot \frac{\partial}{\partial t}\left[t \cdot \sqrt{\frac{x}{n}}\right] =$$

$$= \frac{e^{-\frac{1}{2} \cdot x} \cdot (x)^{\frac{n-2}{2}}}{2^{\frac{n}{2}} \cdot \Gamma\left(\frac{n}{2}\right)} \cdot \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{(t)^2 \cdot x}{2 \cdot n}} \cdot \sqrt{\frac{x}{n}} = \frac{1}{2^{\frac{n}{2}} \cdot \Gamma\left(\frac{n}{2}\right)} \cdot \frac{1}{\sqrt{2\pi n}} \sqrt{x}\,(x)^{\frac{n}{2}-1} e^{-\frac{1}{2} \cdot x} e^{-\frac{(t)^2 \cdot x}{2 \cdot n}} = \tag{6.31}$$

$$= \frac{1}{2^{\frac{n}{2}} \cdot \Gamma\left(\frac{n}{2}\right)} \cdot \frac{1}{\sqrt{2\pi n}} (x)^{\frac{1}{2}} (x)^{\frac{n}{2}-1} e^{-\frac{1}{2} \cdot x\left(1+\frac{t^2}{n}\right)} =$$

$$= \frac{1}{2^{\frac{n}{2}} \cdot \Gamma\left(\frac{n}{2}\right)} \cdot \frac{1}{\sqrt{2\pi n}} (x)^{\frac{n-1}{2}} e^{-\frac{1}{2} \cdot x\left(1+\frac{t^2}{n}\right)}, \qquad x > 0, -\infty < t < \infty$$

V nadaljevanju integrirajmo obe strani enačbe na naslednji način:

$$f(t)=\int_0^{\infty} g(x,t)\,dx=\frac{1}{2^{\frac{n}{2}}\cdot\Gamma\left(\frac{n}{2}\right)}\cdot\frac{1}{\sqrt{2\pi n}}\int_0^{\infty}(x)^{\frac{n-1}{2}}\,e^{-\frac{1}{2}\cdot x\left(1+\frac{t^2}{n}\right)}\,dx$$

$$nova\ spremenljivka:\quad u=\frac{x}{2}\cdot\left(1+\frac{t^2}{n}\right),\quad x=\frac{2u}{\left(1+\frac{t^2}{n}\right)},\qquad dx=\frac{2\cdot du}{\left(1+\frac{t^2}{n}\right)}$$

$$f(t)=\frac{1}{2^{\frac{n}{2}}\cdot\Gamma\left(\frac{n}{2}\right)}\cdot\frac{1}{\sqrt{2\pi n}}\int_0^{\infty}\left(\frac{2u}{\left(1+\frac{t^2}{n}\right)}\right)^{\frac{n-1}{2}}\cdot e^{-u}\cdot\frac{2\cdot du}{\left(1+\frac{t^2}{n}\right)}=$$

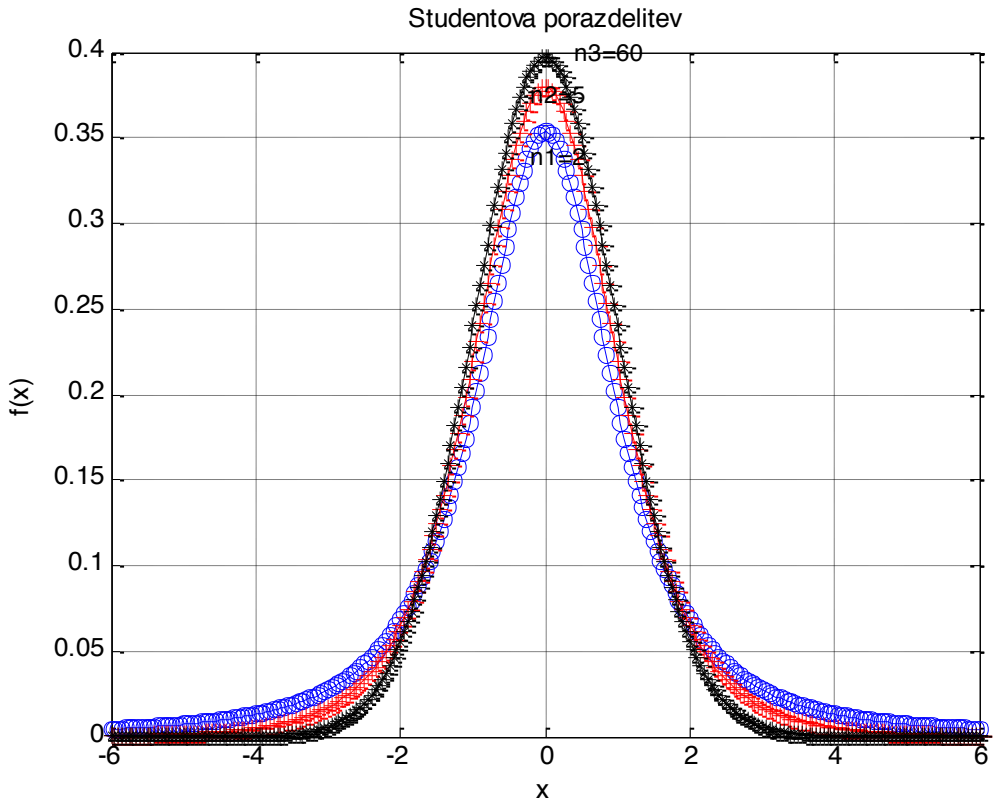
$$=\frac{1}{2^{\frac{n}{2}}\cdot\Gamma\left(\frac{n}{2}\right)}\cdot\frac{1}{\sqrt{2\pi n}}\cdot\frac{2^{\frac{n+1}{2}}}{\left(1+\frac{t^2}{n}\right)^{\frac{n+1}{2}}}\int_0^{\infty}(u)^{\frac{n-1}{2}}\cdot e^{-u}\cdot du=\qquad (6.32)$$

$$=\frac{1}{\Gamma\left(\frac{n}{2}\right)}\cdot\frac{1}{\sqrt{2\pi n}}\cdot\frac{2^{\frac{1}{2}}}{\left(1+\frac{t^2}{n}\right)^{\frac{n+1}{2}}}\cdot\Gamma\left(\frac{n+1}{2}\right)=$$

$$=\frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n}\cdot\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}$$

Tako smo dokazali izrek.

V nadaljevanju lahko narišemo nekaj primerov Studentove porazdelitve (glej sliko 137), pri čemer vzamemo [Krishnamoorthy]: *n = 2,5,60*.



Slika 137: Trije primeri Studentove porazdelitve

Pri izrisu slike 137 smo si pomagali z naslednjim programom v Matlabu:

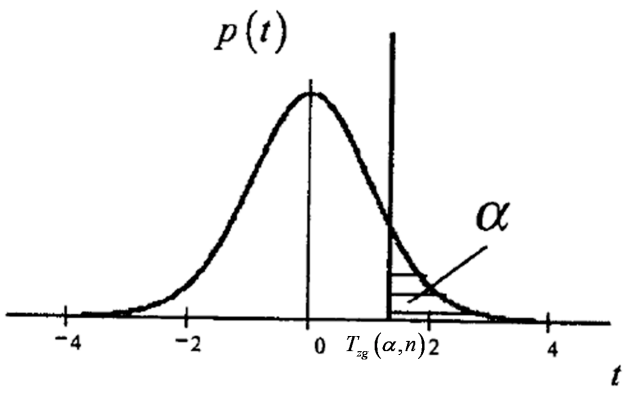
```
% student.m
%
clc
clear
close all

x = -6:0.05:6;

% Narisali bomo poteke studentove porazdelitve pri:
n1 = 2
n2 = 5
n3 = 60

y1 = tpdf(x,n1);
y2 = tpdf(x,n2);
y3 = tpdf(x,n3);

% Narisemo porazdelitve:

plot(x,y1,'b','LineWidth',1)
hold on
plot(x,y1,'bo','LineWidth',1)
```

```
plot(x,y2,'r','LineWidth',1)
plot(x,y2,'r+','LineWidth',1)

plot(x,y3,'k','LineWidth',1)
plot(x,y3,'k*','LineWidth',1)

title('Studentova porazdelitev')
xlabel('x')
ylabel('f(x)')
grid

disp('Vnesi n1 na graf (modra)!!!!!!!!!!!!!!!!')
gtext(['n1=' num2str(n1)],'FontSize',9)

disp('Vnesi n2 na graf (rdeca)!!!!!!!!!!!!!!!!')
gtext(['n2=' num2str(n2)],'FontSize',9)

disp('Vnesi n3 na graf (crna)!!!!!!!!!!!!!!!!')
gtext(['n3=' num2str(n3)],'FontSize',9)
```

Zaradi velike uporabnosti naključne spremenljivke T so vrednosti $T(\alpha,n)$ tabelirane za različne stopnje prostosti in verjetnosti α, pri čemer velja [Jesenko]:

$$P\left(T \geq T_{zg}(\alpha,n)\right) = 1 - P\left(T < T_{zg}(\alpha,n)\right) = \alpha \qquad (6.33)$$
$$P\left(T < T_{zg}(\alpha,n)\right) = 1 - \alpha$$

Iz izraza (6.33) lahko določimo zgornjo mejo oz. kritično vrednost (glej sliko 138). Tabele so običajno zgrajene za različne verjetnosti α ter za stopnje prostosti $n \leq 30$. **Za večje število prostostnih stopenj pa namesto spremenljivke T uporabimo standardizirano normalno naključno spremenljivko**.



Slika 138: Ilustracija zgornje meje oz. kritične vrednosti pri Studentovi t porazdelitvi

Veliko uporabnih lastnosti *t* naključne spremenljivke sloni na naslednjem **izreku** [Jesenko]:

Denimo sta $\bar{X}$ *in* S^2 *aritmetična sredina in varianca naključnega vzorca velikosti n, izbranega iz* ***normalne populacije*** *z aritmetično sredino* μ *in standardnim odklonom* σ. *Potem je naključna spremenljivka* $T = \frac{\left(\bar{X} - \mu\right)}{\frac{S}{\sqrt{n}}}$ *enaka t naključni spremenljivki z n - 1 prostostnimi stopnjami.*

Dokaz:

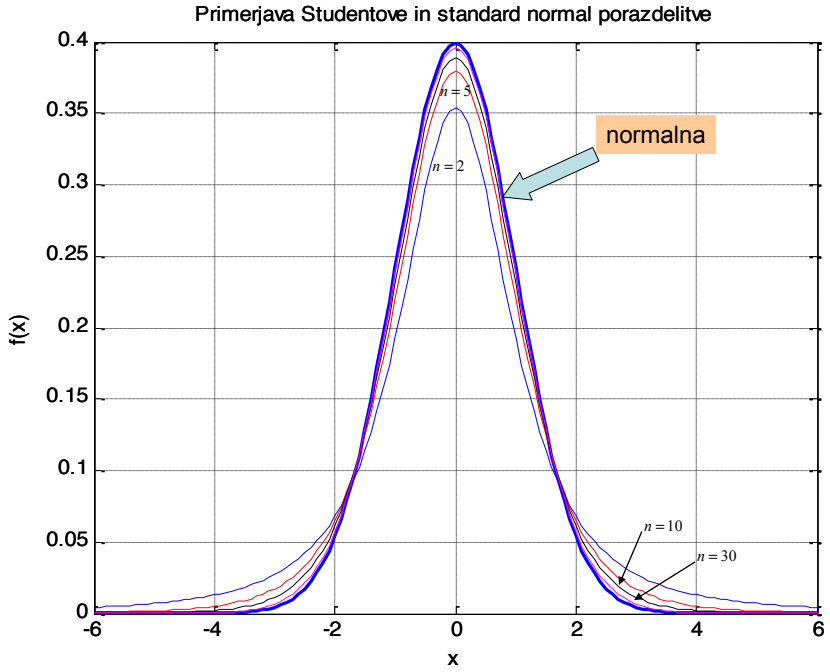
Že prej smo videli iz izraza (6.22), da velja: $\frac{1}{\sigma^2} \cdot S^2 \cdot (n-1) \in \chi^2 (n-1)$. Če je $Z = \frac{\left(\bar{X} - \mu\right)}{\frac{\sigma}{\sqrt{n}}}$ standardizirana **normalna** spremenljivka, potem za izraz (6.27) sledi (pri *n-1* prostostni stopnji):

$$T = \frac{Z}{\sqrt{\frac{\chi^2}{n-1}}} = \frac{\frac{\left(\bar{X} - \mu\right)}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{\frac{1}{\sigma^2} \cdot S^2 \cdot (n-1)}{n-1}}} = \frac{\frac{\left(\bar{X} - \mu\right)}{\frac{\sigma}{\sqrt{n}}}}{\frac{S}{\sigma}} \cdot \sqrt{n} = \frac{\left(\bar{X} - \mu\right)}{S} \cdot \sqrt{n} \qquad (6.34)$$

Torej se spremenljivka $\frac{\left(\bar{X} - \mu\right)}{\frac{S}{\sqrt{n}}}$ porazdeli po *t* naključni porazdelitvi. Kot vidimo, smo se z vpeljavo spremenljivke *T* znebili neznane variance σ^2, saj se je pokrajšala v izrazu (6.34). Namesto nje sedaj lahko operiramo z varianco vzorca S^2, ki pa jo lahko izračunamo.

Primerjava standardne normalne in Studentove porazdelitve

Slika 139 prikazuje primerjavo standardne normalne in Studentove porazdelitve.



Slika 139: Primerjava standardne normalne in Studentove porazdelitve (n = 2, 5, 10, 30)

Primerjava pokaže podobnost obeh porazdelitev, le da ima Studentova porazdelitev več verjetnosti koncentrirane v obeh repih in manj v sredini. Iz slike 139 je tudi razvidno, kako se z večanjem parametra *n* Studentova porazdelitev vedno bolj bliža standardni normalni porazdelitvi in jo pri *n = 30* skoraj že popolnoma ujame.

Pri izrisu slike 139 smo si pomagali z naslednjim programom v Matlabu:

```
% student1.m
%

clc
clear
close all

x = -6:0.1:6;

% Potek standard. normalne porazdelitve:

y = normpdf(x,0,1);

% Narisali bomo poteke studentove porazdelitve pri:

n1 = 2
n2 = 5
```

```
n3 = 10
n4 = 30

y1 = tpdf(x,n1);
y2 = tpdf(x,n2);
y3 = tpdf(x,n3);
y4 = tpdf(x,n4);

% Narisemo porazdelitve:

plot(x,y,'LineWidth',2)
hold on

plot(x,y1,'b','LineWidth',1)
plot(x,y2,'r','LineWidth',1)
plot(x,y3,'k','LineWidth',1)
plot(x,y4,'m','LineWidth',1)

title('Primerjava Studentove in standard normal porazdelitve')
xlabel('x')
ylabel('f(x)')
grid
```

6.5 Fisherjeva F statistika

Naključna spremenljivka, ki igra pomembno vlogo v povezavi z vzorčenjem iz normalne populacije, je *F* naključna spremenljivka.

Izrek:

Naj bosta ξ in η neodvisni χ^2 naključni spremenljivki z n_1 in n_2 prostostnimi stopnjami. Naključna spremenljivka:

$$F = \frac{\frac{\xi}{n_1}}{\frac{\eta}{n_2}} \tag{6.35}$$

se imenuje *F* naključna spremenljivka s porazdelitvijo gostote verjetnosti [Jesenko]:

$$f(x) = \frac{\Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)} \cdot \left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}} \cdot x^{\frac{n_1}{2}-1} \cdot \left(1+\frac{n_1}{n_2}\cdot x\right)^{-\frac{1}{2}(n_1+n_2)} \tag{6.36}$$

za $x > 0$.

Dokaz:

Ker sta ξ in η neodvisni χ^2 naključni spremenljivki, bo njuna združena porazdelitev gostote verjetnosti enaka ($u \in \xi, v \in \eta$):

$$\begin{aligned} f(u,v) = f(u)\cdot f(v) &= \frac{e^{-\frac{1}{2}\cdot u}\cdot (u)^{\frac{n_1-2}{2}}}{2^{\frac{n_1}{2}}\cdot\Gamma\left(\frac{n_1}{2}\right)} \cdot \frac{e^{-\frac{1}{2}\cdot v}\cdot (v)^{\frac{n_2-2}{2}}}{2^{\frac{n_2}{2}}\cdot\Gamma\left(\frac{n_2}{2}\right)} = \\ &= \frac{e^{-\frac{1}{2}\cdot(u+v)}\cdot (u)^{\frac{n_1-2}{2}}}{2^{\frac{n_1+n_2}{2}}\cdot\Gamma\left(\frac{n_1}{2}\right)} \cdot \frac{(v)^{\frac{n_2-2}{2}}}{\Gamma\left(\frac{n_2}{2}\right)} \end{aligned} \tag{6.37}$$

Tranformirana gostota verjetnosti je [Jesenko]:

$$g(x,v) = f\left(x(u,v),v\right)\cdot\frac{\partial}{\partial x}\left[u(x,v)\right] \tag{6.38}$$

pri čemer vpeljemo transformacijo: $x(u,v) = \frac{n_2}{n_1}\cdot\frac{u}{v}$.

Sledi:

$$\begin{aligned} g(x,v) &= \frac{e^{-\frac{1}{2}\left(\frac{n_1}{n_2}\cdot v\cdot x+v\right)}\cdot\left(\frac{n_1}{n_2}\cdot v\cdot x\right)^{\frac{n_1-2}{2}}}{2^{\frac{n_1+n_2}{2}}\cdot\Gamma\left(\frac{n_1}{2}\right)}\cdot\frac{(v)^{\frac{n_2-2}{2}}}{\Gamma\left(\frac{n_2}{2}\right)}\cdot\frac{\partial}{\partial x}\left[\frac{n_1}{n_2}\cdot v\cdot x\right] = \\ &= \frac{e^{-\frac{v}{2}\left(\frac{n_1}{n_2}\cdot x+1\right)}\cdot\left(\frac{n_1}{n_2}\cdot v\cdot x\right)^{\frac{n_1}{2}-1}}{2^{\frac{n_1+n_2}{2}}\cdot\Gamma\left(\frac{n_1}{2}\right)}\cdot\frac{(v)^{\frac{n_2-2}{2}}}{\Gamma\left(\frac{n_2}{2}\right)}\cdot\frac{n_1}{n_2}\cdot v = \\ &= \frac{e^{-\frac{v}{2}\left(\frac{n_1}{n_2}\cdot x+1\right)}\cdot\left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}}(v)^{\frac{n_1+n_2}{2}-1}(x)^{\frac{n_1}{2}-1}}{2^{\frac{n_1+n_2}{2}}\cdot\Gamma\left(\frac{n_1}{2}\right)}\cdot\frac{1}{\Gamma\left(\frac{n_2}{2}\right)} \end{aligned} \tag{6.39}$$

V nadaljevanju integrirajmo obe strani enačbe na naslednji način:

$$f(x) = \int_0^{\infty} g(x,v)\,dv = \int_0^{\infty}\frac{e^{-\frac{v}{2}\left(\frac{n_1}{n_2}\cdot x+1\right)}\cdot\left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}}(v)^{\frac{n_1+n_2}{2}-1}(x)^{\frac{n_1}{2}-1}}{2^{\frac{n_1+n_2}{2}}\cdot\Gamma\left(\frac{n_1}{2}\right)}\cdot\frac{1}{\Gamma\left(\frac{n_2}{2}\right)}\cdot dv \tag{6.40}$$

Dobimo:

$$f(x)=\frac{\left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}}(x)^{\frac{n_1}{2}-1}}{2^{\frac{n_1+n_2}{2}}\cdot\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)}\int_0^{\infty}e^{-\frac{v}{2}\cdot\left(\frac{n_1}{n_2}\cdot x+1\right)}\cdot(v)^{\frac{n_1+n_2}{2}-1}\cdot dv$$

$$\textit{nova spremenljivka}:\ s=\frac{v}{2}\cdot\left(\frac{n_1}{n_2}\cdot x+1\right),\quad v=\frac{2s}{\left(\frac{n_1}{n_2}\cdot x+1\right)},\ dv=\frac{2\cdot ds}{\left(\frac{n_1}{n_2}\cdot x+1\right)} \tag{6.41}$$

$$f(x)=\frac{\left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}}(x)^{\frac{n_1}{2}-1}}{2^{\frac{n_1+n_2}{2}}\cdot\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)}\int_0^{\infty}e^{-s}\cdot\left(\frac{2s}{\left(\frac{n_1}{n_2}\cdot x+1\right)}\right)^{\frac{n_1+n_2}{2}-1}\cdot\frac{2\cdot ds}{\left(\frac{n_1}{n_2}\cdot x+1\right)}$$

Sledi:

$$f(x)=\frac{\left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}}(x)^{\frac{n_1}{2}-1}}{2^{\frac{n_1+n_2}{2}}\cdot\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)}\cdot\frac{(2)^{\frac{n_1+n_2}{2}}}{\left(\frac{n_1}{n_2}\cdot x+1\right)^{\frac{n_1+n_2}{2}}}\int_0^{\infty}e^{-s}\cdot(s)^{\frac{n_1+n_2}{2}-1}\cdot ds=$$

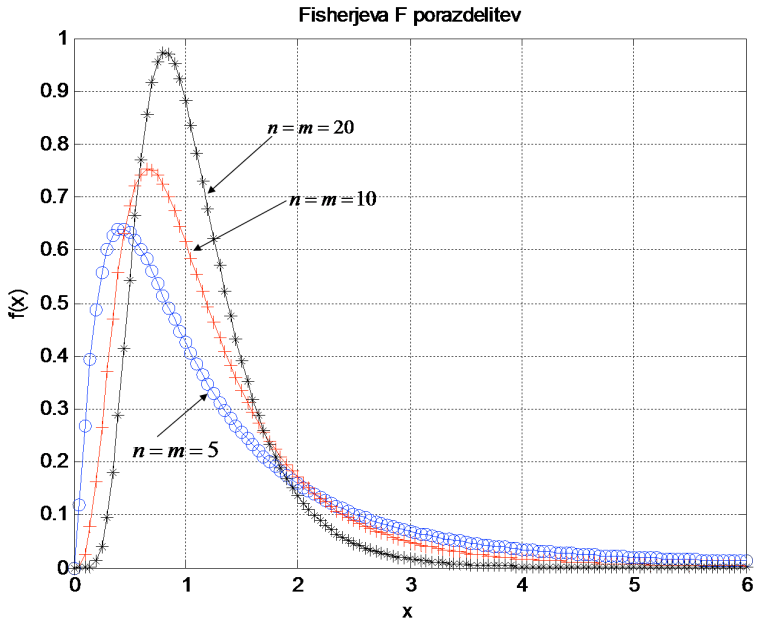
$$=\frac{\left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}}(x)^{\frac{n_1}{2}-1}}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)}\cdot\frac{1}{\left(\frac{n_1}{n_2}\cdot x+1\right)^{\frac{n_1+n_2}{2}}}\cdot\Gamma\left(\frac{n_1+n_2}{2}\right)= \tag{6.42}$$

$$=\frac{\Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)}\cdot\left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}}\cdot x^{\frac{n_1}{2}-1}\cdot\left(1+\frac{n_1}{n_2}\cdot x\right)^{-\frac{1}{2}(n_1+n_2)}$$

Tako smo dobili rezultat, ki smo ga želeli dokazati.

Slika 140 prikazuje primere Fisherjeve F porazdelitve za parametre [Krishnamoorthy]:

$$n=m=5$$
$$n=m=10$$
$$n=m=20$$



Slika 140: Primeri Fisherjeve F porazdelitve

Za izris slike 140 smo uporabili naslednji program v Matlabu:

```
% fisher.m
%

clc
clear
close all


x = 0:0.05:6;

% Narisali bomo poteke fisher porazdelitve pri:

n1 = 5
m1 = 5
n2 = 10
m2 = 10
n3 = 20
m3 = 20

y1 = fpdf(x,n1,m1);
y2 = fpdf(x,n2,m2);
y3 = fpdf(x,n3,m3);

% Narisemo porazdelitve:

plot(x,y1,'b','LineWidth',1)
hold on
plot(x,y1,'bo','LineWidth',1)

plot(x,y2,'r','LineWidth',1)
plot(x,y2,'r+','LineWidth',1)

plot(x,y3,'k','LineWidth',1)
plot(x,y3,'k*','LineWidth',1)

title('Fisherjeva F porazdelitev')
xlabel('x')
ylabel('f(x)')
grid
```

Uporabnost F naključne spremenljivke nastopa predvsem pri problemih, povezanih s primerjanjem varianc σ_1^2 in σ_2^2 dveh normalnih populacij. Sklepi v zvezi s tem vprašanjem so naslonjeni na neodvisne vzorce velikosti n_1 in n_2 iz dveh populacij [Jesenko].

Glede na to, da velja (glej (6.22)) [Jesenko]:

$$\chi_1^{\,2} = \frac{1}{\sigma_1^{\,2}} \cdot S_1^{\,2} \cdot (n_1 - 1)$$
$$\chi_2^{\,2} = \frac{1}{\sigma_2^{\,2}} \cdot S_2^{\,2} \cdot (n_2 - 1) \qquad (6.43)$$

z $n_1 - 1$ oz. $n_2 - 1$ prostostnimi stopnjami, sledi naslednji **izrek** [Jesenko]:

Če sta $S_1^{\,2}$ in $S_2^{\,2}$ varianci dveh neodvisnih naključnih vzorcev velikosti n_1 in n_2 iz dveh normalnih populacij z variancama σ_1^2 in σ_2^2, potem je [Jesenko]:

$$F = \frac{\dfrac{S_1^2}{\sigma_1^2}}{\dfrac{S_2^2}{\sigma_2^2}} = \frac{\sigma_2^2}{\sigma_1^2} \cdot \frac{S_1^2}{S_2^2} \qquad (6.44)$$

F naključna spremenljivka z $n_1 - 1$ in $n_2 - 1$ prostostnimi stopnjami.

7 STATISTIČNO OCENJEVANJE PARAMETROV

7.1 Uvod

Vzorčenja in kakovosti vzorčnih ocen smo se nekoliko dotaknili že v poglavju 3.3. Naloge statističnega sklepanja delimo na [Jesenko, Košmelj B.]:

- probleme ocenjevanja in
- probleme testiranja hipotez.

Po svoji vsebini spadata obe nalogi med probleme odločanja. ***Osnovna razlika je v tem, da moramo pri ocenjevanju določiti vrednost enega ali več parametrov na zvezni množici alternativ, pri testiranju hipotez pa moramo odločati o sprejetju ali zavrnitvi vrednosti enega ali več parametrov*** [Jesenko].

Kadar uporabimo neko statistiko, da z njo ocenimo kakšen parameter populacije, temu postopku pravimo **točkasto ocenjevanje**, sami vrednosti **točkaste statistike (cenilke)** pa rečemo **točkasta ocena parametra**. Tako je na primer S^2 cenilka variance σ^2, medtem ko je s^2 točkasta ocena tega parametra [Jesenko].

Cenilke so naključne spremenljivke, zato je ena glavnih nalog določanje njihovih verjetnostnih porazdelitev. Pri tovrstnih ocenah se namreč vedno pojavlja vprašanje, kako blizu so točkaste ocene pravim vrednostim parametrov populacije, to je, kako dobre so [Jesenko].

Lastnosti, ki jih imajo cenilke, so [Jesenko, Košmelj B.]:

- nepristranskost,
- konsistentnost,
- zadostnost in
- krepkost.

V nadaljevanju bomo v splošnem z Θ označevali cenilko pravega parametra θ, njegovo oceno pa z $\hat{\theta}$.

7.2 Nepristranske cenilke

Poglejmo si naslednjo definicijo [Jesenko, Košmelj B.]:

Statistika Θ *je nepristranska cenilka parametra* θ*, ko velja:*

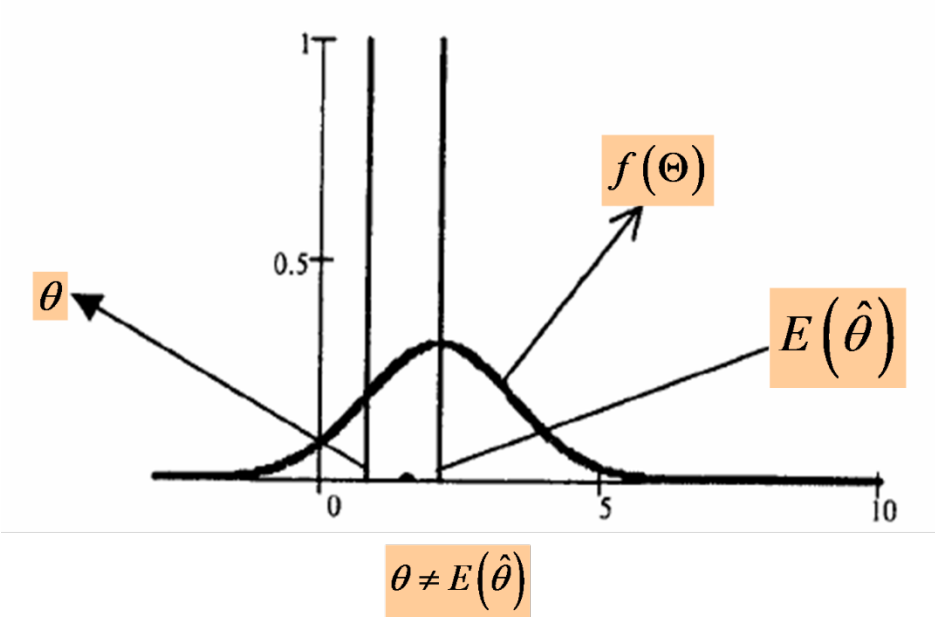
$$E\left(\hat{\theta}\right)=\theta \tag{7.1}$$

Tako npr. velja:

$$E\left(S^2\right)=\sigma^2 \tag{7.2}$$

kar smo pokazali v poglavju 6.2 (glej izraza (6.7) in (6.8), ko smo pristranski varianci dodali korekcijski faktor $\frac{n}{n-1}$, da je postala nepristranska).

Slika 141 prikazuje ilustracijo pristranske cenilke Θ.



Slika 141: Ilustracija pristranske cenilke Θ

7.3 Najučinkovitejše cenilke

Kadar moramo izbirati med več nepristranskimi cenilkami za določen parameter, običajno izberemo tisto, ki ima najmanjšo varianco (**najbolj učinkovita nepristranska cenilka**). Pri tovrstnem ugotavljanju se naslonimo na Cramer-Raovo neenačbo [Jesenko]:

$$VAR(\Theta) \geq \frac{1}{n \cdot E\left[\left(\frac{\partial \ln f(x)}{\partial \hat{\theta}}\right)^2\right]} \tag{7.3}$$

kjer je $f(x)$ porazdelitev gostote verjetnosti populacije.

Izrek:

Če je Θ nepristranska cenilka parametra θ in velja [Jesenko]:

$$VAR(\Theta) = \frac{1}{n \cdot E\left[\left(\frac{\partial \ln f(x)}{\partial \hat{\theta}}\right)^2\right]} \tag{7.4}$$

potem je Θ najbolj učinkovita nepristranska cenilka parametra populacije. Veličina $n \cdot E\left[\left(\frac{\partial \ln f(x)}{\partial \hat{\theta}}\right)^2\right]$ je informacija o parametru θ, ki jo določa vzorec. Večja ko je, manjša je varianca cenilke.

7.4 Dosledne cenilke

V izrazu 6.4 smo imeli:

$$\begin{aligned} E(\bar{X}) &= \mu \\ VAR(\bar{X}) &= \sigma_{\bar{X}}^2 = \frac{\sigma^2}{n} \end{aligned} \tag{7.5}$$

Kot je razvidno, je varianca tem manjša, čim večja je velikost vzorca *n*. Torej se z večanjem velikosti vzorca ocena približuje pravi vrednosti parametra populacije.

Definicija:

Statistika Θ *je* ***dosledna (konsistentna)*** *cenilka parametra* θ *natanko tedaj, ko za vsako realno število c > 0 velja (verjetnostna konvergenca) [Jesenko]:*

$$\lim_{n\to\infty} P\left(\left|\Theta-\theta\right|<c\right)=1 \tag{7.6}$$

Ta definicija nam pove, da je napaka ocene, izračunane iz cenilke pri zadosti velikem vzorcu, zanemarljiva. Ocena se praktično ujema s pravo vrednostjo parametra populacije, iz katere je vzorec izbran.

Izrek:

Če je Θ nepristranska cenilka parametra θ in velja:

$$\lim_{n\to\infty} VAR\left(\Theta\right)=0 \tag{7.7}$$

potem je Θ dosledna cenilka parametra θ [Jesenko].

7.5 Zadostne cenilke

Cenilka Θ je zadostna, če uporabi vse informacije vzorca, ki so relevantne za oceno parametra θ. Drugače rečeno, vse kar izvemo o parametru θ iz posameznih elementov vzorca, lahko izvemo tudi iz cenilke Θ [Jesenko].

7.6 Metoda momentov

Metoda momentov je ena od metod za določanje cenilk parametrov. Moment reda *k* vzorca velikosti *n* je [Jesenko]:

$$m_k' = \frac{\sum_{i=1}^{n}\left(x_i\right)^k}{n} \tag{7.8}$$

Po tej metodi ocenjujemo neznane parametre, ki nastopajo v porazdelitvenih zakonih populacij, na takšen način, da rešimo *r* enačb [Jesenko]:

$$m_k' = \mu_k', \quad k = 1,2,...,r \tag{7.9}$$

kolikor je parametrov, ki jih moramo oceniti. V teh enačbah pomeni μ_k' začetni moment reda *k*, izračunan iz porazdelitvenega zakona.

Primer 7.1.:

Dano imamo Rayleighevo naključno spremenljivko, ki ima porazdelitev gostote verjetnosti [Jesenko, Krishnamoorthy]:

$$f(x) = 2 \cdot a \cdot x \cdot e^{-a \cdot x}, \quad x > 0 \tag{7.10}$$

Ker v izrazu (7.10) nastopa le en parameter ($r = 1$), velja (imamo le eno enačbo):

$$m_1' = \mu_1', \quad k = 1 \tag{7.11}$$

Izraz (7.8) preide v obliko:

$$m_1' = \frac{\sum_{i=1}^{n} (x_i)^1}{n} = \frac{\sum_{i=1}^{n} x_i}{n} = \overline{x} \tag{7.12}$$

Matematično upanje Rayleigheve naključne spremenljivke je enako [Jesenko, Krishnamoorthy]:

$$E(X) = \mu_1' = \mu = \frac{1}{2}\sqrt{\frac{\pi}{a}} \tag{7.13}$$

Izenačimo in dobimo oceno parametra *a* iz podatkov vzorca:

$$\begin{aligned} m_1' &= \mu_1' \\ \bar{x} &= \frac{1}{2}\sqrt{\frac{\pi}{\hat{a}}} \\ \hat{a} &= \frac{\pi}{4\cdot\bar{x}^2} = \frac{\pi}{4}\cdot\frac{1}{\left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2} \end{aligned} \tag{7.14}$$

Cenilka je v tem primeru enaka:

$$\hat{A} = \frac{\pi}{4\cdot\bar{x}^2} \tag{7.15}$$

Primer 7.2.:

Dano imamo Gama naključno spremenljivko, ki ima porazdelitev gostote verjetnosti [Jesenko, Krishnamoorthy]:

$$f(x) = \frac{e^{-\frac{1}{b}\cdot x}\cdot(x)^{a-1}}{b^a\cdot\Gamma(a)} \quad ,x>0 \tag{7.16}$$

Kot vemo na osnovi izraza (5.112), velja:

$$\begin{aligned} E(X) &= a\cdot b = a\cdot\frac{1}{\lambda} = \mu_1' \\ VAR(x) &= a\cdot b^2 = a\cdot\frac{1}{\lambda^2} \end{aligned} \tag{7.17}$$

Začetni moment drugega reda je enak:

$$E(X^2) = VAR(X) + E^2(X) = a\cdot b^2 + (a\cdot b)^2 = a\cdot b^2(1+a) = \mu_2' \tag{7.18}$$

Izraz (7.8) preide v obliko:

$$
\begin{aligned}
m_k' &= \frac{\sum_{i=1}^{n} (x_i)^k}{n}, \quad k = 1,2, \quad \text{ker } r = 2 \\
m_1' &= \frac{\sum_{i=1}^{n} (x_i)^1}{n}, \\
m_2' &= \frac{\sum_{i=1}^{n} (x_i)^2}{n}
\end{aligned}
\tag{7.19}
$$

Sistem enačb (7.9) preide v obliko:

$$
\begin{aligned}
m_k' &= \mu_k', \quad k = 1,2 \\
m_1' &= \mu_1' \\
m_2' &= \mu_2'
\end{aligned}
\tag{7.20}
$$

Dobimo:

$$
\begin{aligned}
m_1' &= \frac{\sum_{i=1}^{n} (x_i)^1}{n} = a \cdot b \\
m_2' &= \frac{\sum_{i=1}^{n} (x_i)^2}{n} = a \cdot b^2 (1 + a)
\end{aligned}
\tag{7.21}
$$

Rešimo ta sistem enačb:

$$
\begin{aligned}
& a = \frac{m_1'}{b} \\
& m_2' = \frac{m_1'}{b} \cdot b^2 \left(1 + \frac{m_1'}{b} \right) = m_1' \cdot b \left(1 + \frac{m_1'}{b} \right) = m_1' \cdot b \left(\frac{b + m_1'}{b} \right) = \\
& = m_1' \cdot \left(b + m_1' \right) \\
& \textit{Sledi}: \\
& b = \frac{m_2'}{m_1'} - m_1' = \frac{m_2' - m_1'^2}{m_1'} \\
& a = \frac{m_1'}{\frac{m_2' - m_1'^2}{m_1'}} = \frac{m_1'^2}{m_2' - m_1'^2}
\end{aligned}
\tag{7.21}
$$

Oceni parametrov *a* in *b* iz podatkov vzorca torej sta:

$$
\begin{aligned}
& \hat{b} = \frac{m_2' - m_1'^2}{m_1'} \\
& \hat{a} = \frac{m_1'^2}{m_2' - m_1'^2}, \\
& \textit{kjer sta}: \\
& m_1' = \frac{\sum_{i=1}^{n} (x_i)^1}{n}, \\
& m_2' = \frac{\sum_{i=1}^{n} (x_i)^2}{n}
\end{aligned}
\tag{7.22}
$$

V nekaterih pogledih ta metoda deluje slabše kot metoda največje podobnosti, saj ima slednja višjo verjetnost, da je ocena blizu prave vrednosti. Vendar pa se v nekaterih primerih, kot npr. v pravkar prikazanem primeru za Gama porazdelitev, odreže metoda momentov veliko bolje, saj je z metodo največje podobnosti izredno težko ali celo nemogoče analitično poiskati rezultate. Po drugi strani pa je lahko analitično reševanje z metodo momentov za takšne primere dokaj preprosto. Ocene z metodo momentov se tudi

lahko uporabijo kot začetne aproksimacije rešitev metode največje podobnosti, pri čemer nato izvedemo nadaljnje izboljšave s pomočjo Newton-Raphsonove metode. Metoda momentov se lahko včasih slabše odreže tudi v primeru male velikosti vzorca, ko ocene lahko celo padejo izven definicijskega prostora parametrov. To se pri metodi največje podobnosti ne more nikoli zgoditi.

7.7 Metoda največje podobnosti

Metodo največje podobnosti smo nekoliko spoznali že v poglavju 2.19. Ta metoda daje **ocene parametrov največje verjetnosti** in se imenuje tudi **metoda največje verjetnosti**. Bistvo te metode je v tem, da izberemo oceno neznanega parametra tako, da ima porazdelitveni zakon oz. porazdelitev verjetnosti (porazdelitev gostote verjetnosti) za izbrani vzorec največjo vrednost [Jesenko].

Definicija:

Če so $x_1,...,x_n$ med seboj neodvisne vrednosti naključnega vzorca, izbranega iz populacije s parametrom Θ, je funkcija največje verjetnosti vzorca (oz. njen naravni logaritem) enaka [Jesenko]:

$$\begin{aligned} L\left(\hat{\theta}\right) &= f\left(x_1,...,x_n,\hat{\theta}\right) = f\left(x_1,\hat{\theta}\right)\cdot f\left(x_2,\hat{\theta}\right)\cdot ....\cdot f\left(x_n,\hat{\theta}\right) \\ \ln L\left(\hat{\theta}\right) &= \ln f\left(x_1,...,x_n,\hat{\theta}\right) = \ln f\left(x_1,\hat{\theta}\right)+...+\ln f\left(x_n,\hat{\theta}\right) \end{aligned} \tag{7.23}$$

Optimalno oceno s to metodo poiščemo na naslednji način:

$$\frac{d}{d\hat{\theta}}\ln L\left(\hat{\theta}\right)=0 \quad \rightarrow \quad \hat{\theta}_{opt} \tag{7.24}$$

torej da poiščemo ekstrem (maksimum) te funkcije (enega parametra).

V primeru, če imamo več (*r*) parametrov, velja:

(7.25)

$$L\left(\hat{\theta}_i\right) = f\left(x_1,...,x_n,\hat{\theta}_i\right) = f\left(x_1,\hat{\theta}_i\right)\cdot f\left(x_2,\hat{\theta}_i\right)\cdot...\cdot f\left(x_n,\hat{\theta}_i\right)$$
$$\ln L\left(\hat{\theta}_i\right) = \ln f\left(x_1,...,x_n,\hat{\theta}_i\right) = \ln f\left(x_1,\hat{\theta}_i\right)+...+\ln f\left(x_n,\hat{\theta}_i\right)$$
$$\hat{\theta}_i = \left\{\hat{\theta}_1,...,\hat{\theta}_r\right\}$$

Optimalne ocene s to metodo v tem primeru poiščemo na naslednji način:

$$\frac{\partial}{\partial\hat{\theta}_j}\ln L\left(\hat{\theta}_i\right)=0 \quad , \quad j=1,...,r \qquad \rightarrow \quad \hat{\theta}_{iopt} = \left\{\hat{\theta}_{1\,opt},...,\hat{\theta}_{r\,opt}\right\} \tag{7.26}$$

torej da rešimo sistem *r* enačb, ki jih da parcialno odvajanje po parametrih in enačenje rezultatov odvajanja z 0. Seveda s tem tudi poiščemo ekstrem (maksimum) funkcije več parametrov.

Primer 7.3.:

Dano imamo Paretovo naključno spremenljivko, ki ima porazdelitev gostote verjetnosti [Jesenko, Krishnamoorthy]:

$$f\left(x\right) = \frac{a}{x^{a+1}} \qquad , x>1, a>0 \tag{7.27}$$

Poiščite oceno parametra a s pomočjo metode največje podobnosti. Poiščite nato oceno tudi z metodo momentov.

Funkcija največje verjetnosti vzorca je enaka:

$$\begin{aligned} &L\left(\hat{\theta}\right) = f\left(x_1,...,x_n,\hat{\theta}\right) = f\left(x_1,\hat{\theta}\right)\cdot f\left(x_2,\hat{\theta}\right)\cdot...\cdot f\left(x_n,\hat{\theta}\right) \\ &L\left(\hat{a}\right) = f\left(x_1,...,x_n,\hat{a}\right) = f\left(x_1,\hat{a}\right)\cdot f\left(x_2,\hat{a}\right)\cdot...\cdot f\left(x_n,\hat{a}\right) = \\ &= \frac{\hat{a}}{x_1^{\hat{a}+1}}\cdot\frac{\hat{a}}{x_2^{\hat{a}+1}}\cdot...\cdot\frac{\hat{a}}{x_n^{\hat{a}+1}} = \frac{\hat{a}^n}{\left(x_1\cdot x_2\cdot...\cdot x_n\right)^{\hat{a}+1}} \end{aligned} \tag{7.28}$$

Njen naravni logaritem je enak:

(7.29)

$$\ln L(\hat{a}) = \ln\left(\frac{\hat{a}^n}{(x_1 \cdot x_2 \cdot \ldots \cdot x_n)^{\hat{a}+1}}\right) = \ln \hat{a}^n - \ln(x_1 \cdot x_2 \cdot \ldots \cdot x_n)^{\hat{a}+1} =$$
$$= n \cdot \ln \hat{a} - (\hat{a}+1) \cdot \ln(x_1 \cdot x_2 \cdot \ldots \cdot x_n)$$

Odvajajmo in enačimo z 0:

$$\begin{aligned} &\frac{d}{d\hat{a}} \ln L(\hat{a}) = \frac{d}{d\hat{a}}\left(n \cdot \ln \hat{a} - (\hat{a}+1) \cdot \ln(x_1 \cdot x_2 \cdot \ldots \cdot x_n)\right) = \\ &= \frac{n}{\hat{a}} - \ln(x_1 \cdot x_2 \cdot \ldots \cdot x_n) = 0 \\ &Sledi: \\ &\hat{a}_{opt} = \frac{n}{\ln(x_1 \cdot x_2 \cdot \ldots \cdot x_n)} = \frac{n}{\ln(x_1) + \ldots + \ln(x_n)} \end{aligned} \tag{7.29}$$

Tako dobimo oceno največje verjetnosti parametra *a*. Pripadajoča cenilka največje verjetnosti pa je:

$$\hat{A} = \frac{n}{\ln(X_1) + \ldots + \ln(X_n)} \tag{7.30}$$

Pri čemer je *X* Paretova naključna spremenljivka.

Poiščimo sedaj oceno še z metodo momentov. Matematično upanje Paretove naključne spremenljivke je enako:

$$\begin{aligned} &E(X) = \int_{-\infty}^{\infty} x \cdot f(x) dx = \int_{1}^{\infty} x \cdot \frac{a}{x^{a+1}} dx = a \int_{1}^{\infty} x^{-(a)} dx = \\ &= a \cdot \frac{1}{-(a)+1}\left(x^{-(a)+1}\right)_1^{\infty} = \frac{-a}{1-a} = \frac{a}{a-1} \end{aligned} \tag{7.31}$$

Dobimo:

$$
\begin{aligned}
m_1' &= \frac{\sum_{i=1}^{n}(x_i)^1}{n} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x} \\
E(X) &= \mu_1' = \frac{a}{a-1} \\
m_1' &= \mu_1' \\
\bar{x} &= \frac{a}{a-1} \\
(a-1)\bar{x} &= a \\
a\cdot\bar{x} - \bar{x} &= a \\
a(\bar{x}-1) &= \bar{x} \\
\hat{a} &= \frac{\bar{x}}{(\bar{x}-1)} = \frac{\frac{\sum_{i=1}^{n} x_i}{n}}{\left(\frac{\sum_{i=1}^{n} x_i}{n} - 1\right)}
\end{aligned}
\qquad (7.32)
$$

Nadaljevanje primera 7.3.

Dano imamo populacijo s 1000 enotami, ki ima naslednjo Paretovo porazdelitev:

$$
\begin{aligned}
f(x) &= \frac{a}{x^{a+1}} \quad , x>1, a>0 \\
f(x) &= \frac{2}{x^{2+1}} = \frac{2}{x^3} \quad , x>1 \\
a &= 2
\end{aligned}
\qquad (7.33)
$$

Torej je prava vrednost parametra enaka 2. Srednja vrednost je enaka [Jesenko, Krishnamoorthy]:

$$E(X)=\mu=\frac{a}{a-1}=\frac{2}{2-1}=2 \tag{7.34}$$

a) Generirajte M = 1000 vzorcev te populacije s pomočjo Matlaba. Narišite normiran histogram in teoretično porazdelitev za populacijo. Ocenite vrednost parametra populacije z metodo največje podobnosti in metodo momentov ter izračunajte srednjo vrednost, če bi vzeli pri izračunu vse enote populacije.

b) Pri različnih velikostih vzorca iz populacije $(50 \le N \le 500)$ *izračunajte oceno parametra z obema metodama in narišite potek* $\hat{a}(N)$, $N = 50, 52, 54, ..., 500$. *Za obe metodi narišite tudi histograma porazdelitve te ocene glede na* $50 \le N \le 500$ *ter izračunajte aritmetično sredino* $E(\hat{a}(N))$ *in standardni odklon* $STD(\hat{a}(N))$.

c) Pri fiksni velikost vzorca $N = 50$ *naključno potegnite vzorec iz populacije stokrat* $(K = 100)$, *izračunajte oceno parametra z obema metodama in narišite potek* $\hat{a}(i, N) = \hat{a}(i, 50)$, $i = 1, 2, ..., 100$, *kjer je i indeks (zaporedna številka) naključnega vzorca. Za obe metodi narišite tudi histograma porazdelitve te ocene glede na* $i = 1, 2, ..., 100$ *ter izračunajte aritmetično sredino* $E(\hat{a}(i, 50))$ *in standardni odklon* $STD(\hat{a}(i, 50))$.

Za izračune v točki b) smo uporabili naslednji program v Matlabu:

```
% ocenj_par.m
% ocenjevanje parametra Pareto porazdelitve z max likeli in z metodo
% momentov (VZORCE RAZLICNIH VELIKOSTI VLECEMO IZ POPULACIJE)

clc
clear
close all

% Inicializacija nakljucnega generatorja:

rand('seed',0)

a = input('Vnesi pravi parameter a')
M = input('Vnesi velikost populacije M');
dN = input('korak narascanja velikosti vzorca')
Nmin = input('Vnesi min velikost vzorca Nmin');
Nmax = input('Vnesi max velikost vzorca Nmax');

if length(a) == 0
   a = 2
end
if length(M) == 0
   M = 1000
end
```

```
if length(dN) == 0
   dN = 2
end
if length(Nmin) == 0
   Nmin = 50
end
if length(Nmax) == 0
   Nmax = 500
end

N = Nmin:dN:Nmax;


% Generiranje populacije Pareto porazdelitve:

disp('- Generiram populacijo Pareto porazdelitve')
disp('        ')

x = randraw('pareto', [1 a], [1 M]);   % dobljena funkcija iz file exchange

% Normiran histogram in teoreticna porazdelitev za populacijo (funkcija norm_hist na osnovi Martinez,
str.103):

disp('srednja vrednost populacije:')

srvr = norm_hist(x,'z(k) = par/(y(k)^(par+1));',0.2/a,1,10/a+1,M/(2*a),'Pareto',a,1)


% Ocenjevanje z obema metodama za celo populacijo:

disp('ocenjevanje z metodo max. likelihood (cela populacija):')

im = 0;

for i=1:M
    im = im + log(x(i));
end

a1 = M/im

disp('ocenjevanje z metodo momentov (cela populacija):')

a2 = mean(x)/(mean(x)-1)


% Generiranje vzorcev Pareto porazdelitve in ocenjevanje parametra - a_oc(N):
% (velikost vzorca N se veca)

disp('- Generiram vzorce Pareto porazdelitve, velikost N jim narasca')
disp('        ')

a1 = []; a2 = [];

for j=N

    x1 = ran_poteg(x,j);   % nakljucno vlecemo j enot iz populacije - vzorec je velikosti j, ki se veca

    im = 0;

    for i=1:j
        im = im + log(x1(i));
    end

    a1 = [a1 j/im];                      % ocena z likelihood

    a2 = [a2 mean(x1)/(mean(x1)-1)];     % ocena z metodo momentov

end

% Izris parametra a_oc(N), ocenjenega z likelihood (MLE) - vzorec se veca na x osi:

figure
plot(N,a1)
hold on
plot(N,a1,'o')
plot(N,a*ones(length(N),1)','r','LineWidth',3)
title('Ocena parametra a z max. likelihood: a_oc(N) ')
xlabel(['Velikost vzorca N,   pravi parameter a je: ' num2str(a)])
grid

% Izris parametra a_oc(N), ocenjenega z metodo momentov (MM) - vzorec se veca na x osi:

figure
plot(N,a2)
hold on
plot(N,a2,'o')
plot(N,a*ones(length(N),1)','r','LineWidth',3)
title('Ocena parametra a z metodo momentov: a_oc(N) ')
```

```
xlabel(['Velikost vzorca N,   pravi parameter a je: ' num2str(a)])
grid


disp('srednja vrednost za oceno a1 pri spremenljivi velikosti vzorca N (MLE):')
mean(a1)

disp('deviacija za oceno a1 pri spremenljivi velikosti vzorca N (MLE):')
std(a1)


disp('srednja vrednost za oceno a2 pri spremenljivi velikosti vzorca N (MM):')
mean(a2)

disp('deviacija za oceno a2 pri spremenljivi velikosti vzorca N (MM):')
std(a2)

% Izris histograma za a1(N):

figure
histfit(a1,30)
grid
title('histogram ocene parametra a z max. likelihood')
xlabel(['a_oc(N),   pravi parameter a je: ' num2str(a)])
ylabel('Stevilo ocen(a_oc)')

% Izris histograma za a2(N):

figure
histfit(a2,30)
grid
title('histogram ocene parametra a z metodo momentov')
xlabel(['a_oc(N),   pravi parameter a je: ' num2str(a)])
ylabel('Stevilo ocen(a_oc)')
```

Kot vidimo, program kliče tudi funkcijo **randraw.m**, ki je bila pridobljena na Matlab File Exchange spletni strani. Prav tako program kliče funkcijo **norm_hist.m**, ki ima naslednjo obliko [Martinez]:

```
% Normiran histogram in teoreticna porazdelitev za vhodni signal x:

function srvr = norm_hist(x,f,dy,ymin,ymax,nbins,fstring,par,axst)

M = length(x);

y = ymin:dy:ymax;             % za teoreticno porazdelitev

for k=1:length(y)             % generiramo teoreticno porazdelitev f
    eval(f)
end

[B,h] = hist(x,nbins);        % generiramo histogram vhodnega signala

B = B/(h(2)-h(1))/M;          % normiramo histogram vhodnega signala (ga uglasimo s f)

bar(h,B,1,'r')
hold on
plot(y,z,'LineWidth',2)
plot(y,z,'o')
hold off
xlabel('x')
ylabel('f(x)')
grid
title(['Normiran histogram populacije s ' fstring ' porazdelitvijo in teoreticna porazdelitev'])
d = axis;
axis([axst ymax d(3) d(4)])

srvr = mean(x);

return
```

Program kliče tudi funkcijo **ran_poteg.m**, ki ima obliko:

```
function y = ran_poteg(x,N)

z = [];
t = 1:length(x);
```

```
for i=1:N
    while 1==1
          k = randi([min(t) max(t)],1);
          if length(intersect(t,k)) > 0
              break
          end
    end
    z = [z;k];
    t = setdiff(t,k);
end

y = x(z);

return
```

Za izračune v točki c) smo uporabili naslednji program v Matlabu:

```
% ocenj_par1.m
% ocenjevanje parametra Pareto porazdelitve z max likeli in z metodo
% momentov (VZOREC ISTE VELIKOSTI VECKRAT VLECEMO IZ POPULACIJE)

clc
clear
close all

% Inicializacija nakljucnega generatorja:

rand('seed',0)

a = input('Vnesi pravi parameter a')
M = input('Vnesi velikost populacije M');
K = input('Vnesi stevilo potegov vzorca K')
N = input('Vnesi velikost vzorca N');


if length(a) == 0
   a = 2
end
if length(M) == 0
   M = 1000
end
if length(K) == 0
   K = 100
end
if length(N) == 0
   N = 50
end


% Generiranje populacije Pareto porazdelitve:

disp('- Generiram populacijo Pareto porazdelitve')
disp('        ')

x = randraw('pareto', [1 a], [1 M]);    % dobljena funkcija iz file exchange


% Normiran histogram in teoreticna porazdelitev za populacijo (funkcija norm_hist na osnovi Martinez,
str.103):

disp('srednja vrednost populacije:')

srvr = norm_hist(x,'z(k) = par/(y(k)^(par+1));',0.2/a,1,10/a+1,M/(2*a),'Pareto',a,1)


% Ocenjevanje z obema metodama za celo populacijo:

disp('ocenjevanje z metodo max. likelihood (cela populacija):')

im = 0;

for i=1:M
    im = im + log(x(i));
end

a1 = M/im

disp('ocenjevanje z metodo momentov (cela populacija):')

a2 = mean(x)/(mean(x)-1)


% Generiranje vzorcev Pareto porazdelitve in ocenjevanje parametra - a_oc(ind.vzorca velikosti N):
% (Velikost vzorca N skoz enaka, a veckrat vlecemo vzorec iz populacije)
```

```
disp('- Generiram K vzorcev velkosti N Pareto porazdelitve')
disp('         ')

a1 = []; a2 = [];









for j=1:K                    % K krat vlecemo vzorec

    x1  =  ran_poteg(x,N);        % vzorec vedno velikosti N (nakljucno potegnemo vsakic N enot iz
populacije)

    im = 0;

    for i=1:N
        im = im + log(x1(i));
    end

    a1 = [a1 N/im];                      % ocena z likelihood

    a2 = [a2 mean(x1)/(mean(x1)-1)];     % ocena z metodo momentov

end

% Izris parametra a_oc(indeks_vzorca), ocenjenega z likelihood (MLE):
% (na x osi je indeks vzorca velikosti N, vseh skupaj jih je K)

figure
plot(1:K,a1)
hold on
plot(1:K,a1,'o')
plot(1:K,a*ones(K,1)','r','LineWidth',3)
title('Ocena parametra a z max. likelihood: a_oc(ind_vzorca)')
xlabel(['Indeks vzorca velikosti N,   pravi parameter a je: ' num2str(a)])
grid
d=axis;
axis([d(1) d(2) 0.7*d(3) 1.3*d(4)])

% Izris parametra a_oc(N), ocenjenega z metodo momentov (MM) - vec vzorcev K, vsi velikosti N:

figure
plot(1:K,a2)
hold on
plot(1:K,a2,'o')
plot(1:K,a*ones(K,1)','r','LineWidth',3)
title('Ocena parametra a z metodo momentov: a_oc(ind_vzorca)')
xlabel(['Indeks vzorca velikosti N,   pravi parameter a je: ' num2str(a)])
grid
d=axis;
axis([d(1) d(2) 0.7*d(3) 1.3*d(4)])

disp('srednja vrednost za oceno a1 pri veckrat vlecenem vzorcu velikosti N (MLE):')
mean(a1)

disp('deviacija za oceno a1 pri veckrat vlecenem vzorcu velikosti N (MLE):')
std(a1)


disp('srednja vrednost za oceno a2 pri veckrat vlecenem vzorcu velikosti N (MM):')
mean(a2)

disp('deviacija za oceno a2 pri veckrat vlecenem vzorcu velikosti N (MM):')
std(a2)

% Izris histograma za a1(ind_vzorca velikosti N):

figure
histfit(a1,25)
grid
title('histogram ocene parametra a z max. likelihood')
xlabel(['a_oc(indeks vzorca velikosti N),   pravi parameter a je: ' num2str(a)])
ylabel('Stevilo ocen(a_oc)')

% Izris histograma za a2(ind_vzorca velikosti N):

figure
histfit(a2,25)
grid
title('histogram ocene parametra a z metodo momentov')
```

```
xlabel(['a_oc(indeks vzorca velikosti N),   pravi parameter a je: ' num2str(a)])
ylabel('Stevilo ocen(a_oc)')
```

Kot vidimo iz obeh programov, smo v njih opravili tudi izračune za točko a). Če poženemo program **ocenj_par.m**, dobimo naslednji izpis v komandnem oknu:

```
Vnesi pravi parameter a
a =
    []
Vnesi velikost populacije M
korak narascanja velikosti vzorca
dN =
    []
Vnesi min velikost vzorca Nmin
Vnesi max velikost vzorca Nmax

a =
     2

M =
        1000

dN =
     2

Nmin =
    50

Nmax =
   500

- Generiram populacijo Pareto porazdelitve

srednja vrednost populacije:
srvr =
    1.9896

ocenjevanje z metodo max. likelihood (cela populacija):
a1 =
    2.0154

ocenjevanje z metodo momentov (cela populacija):
a2 =
    2.0105

- Generiram vzorce Pareto porazdelitve, velikost N jim narasca

srednja vrednost za oceno a1 pri spremenljivi velikosti vzorca N (MLE):
ans =
    2.0138

deviacija za oceno a1 pri spremenljivi velikosti vzorca N (MLE):
ans =
    0.1439

srednja vrednost za oceno a2 pri spremenljivi velikosti vzorca N (MM):
ans =
    2.0228

deviacija za oceno a2 pri spremenljivi velikosti vzorca N (MM):
ans =
    0.1694
```

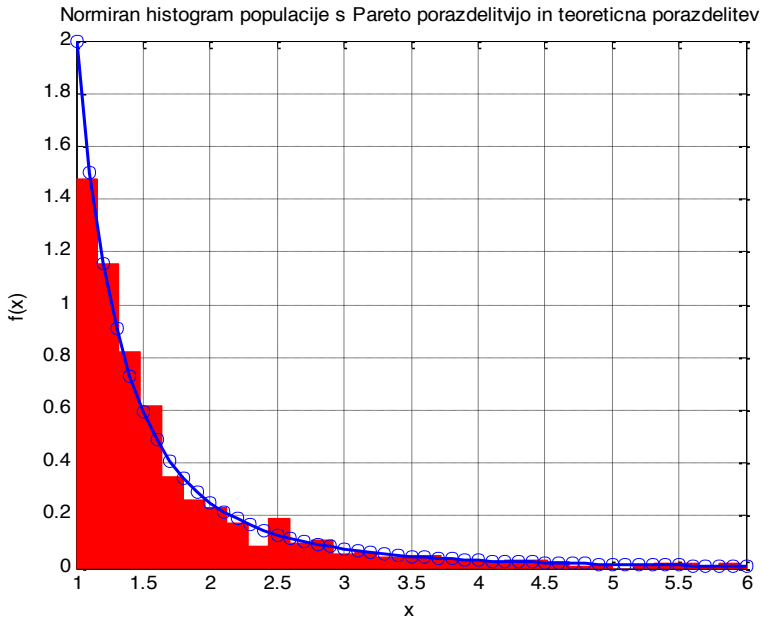
Dobimo torej naslednje rezultate, če pri računanju upoštevamo celo populacijo (MLE-metoda največje podobnosti, MM - metoda momentov):

$$\hat{\mu} = \frac{1}{M}\sum_{i=1}^{M} x_i = 1.9896$$

$$\hat{a}_{MLE} = \frac{M}{\ln(x_1) + ... + \ln(x_M)} = 2.0154$$

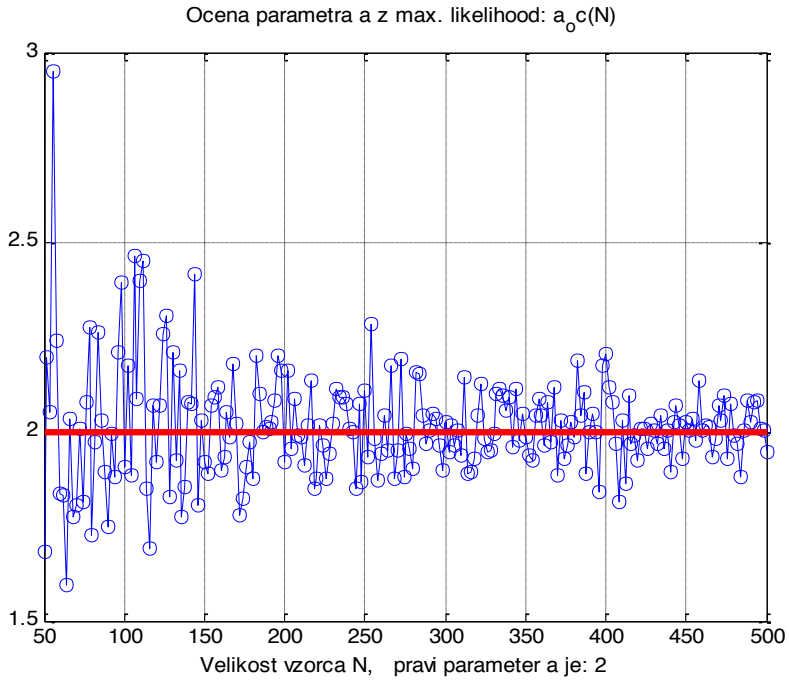
$$\hat{a}_{MM} = \frac{\overline{x}}{(\overline{x} - 1)} = \frac{\frac{\sum_{i=1}^{M} x_i}{M}}{\left(\frac{\sum_{i=1}^{M} x_i}{M} - 1\right)} = 2.0105$$

Če bi pri izračunu $\hat{\mu}$ upoštevali $\frac{\hat{a}_{MLE}}{\hat{a}_{MLE} - 1}$, bi dobili za malenkost drugačen rezultat 1.9848. Ča pa bi upoštevali $\frac{\hat{a}_{MM}}{\hat{a}_{MM} - 1}$, bi pa dobili enak rezultat 1.9896. Slika 142 prikazuje normiran histogram in teoretično porazdelitev za populacijo.



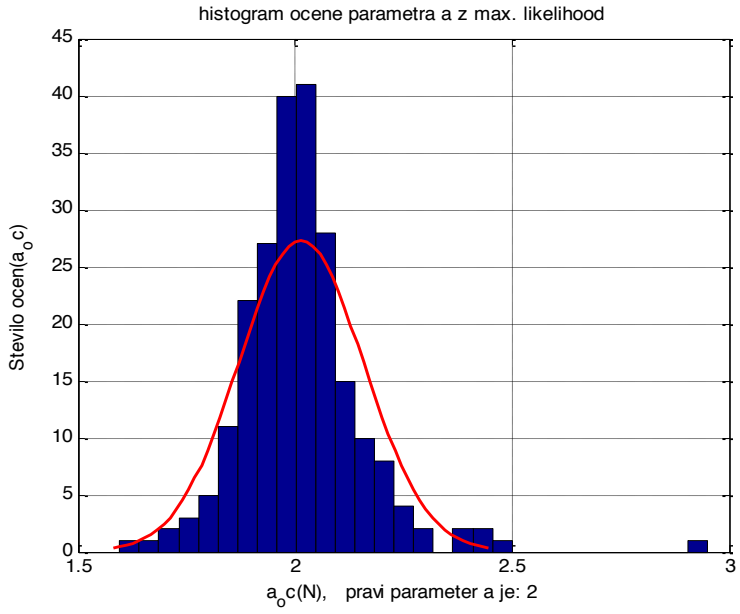
Slika 142: Normiran histogram in teoretično porazdelitev za Paretovo populacijo.

Slika 143 prikazuje potek parametra $\hat{a}_{MLE}(N)$, $N = 50,52,54,...,500$, ocenjenega z metodo največje podobnosti.



Slika 143: Potek parametra $\hat{a}_{MLE}(N)$, $N = 50,52,54,...,500$, ocenjenega z metodo največje podobnosti.

Kot lahko vidimo, z naraščanjem velikosti vzorca N varianca ocenjenega parametra pada, ocenjeni rezultati pa so vse bliže pravi vrednosti parametra. Histogram porazdelitve ocenjenega parametra $\hat{a}_{MLE}(N)$ glede na spreminjajočo se velikost vzorca ($50 \le N \le 500$) prikazuje slika 144.

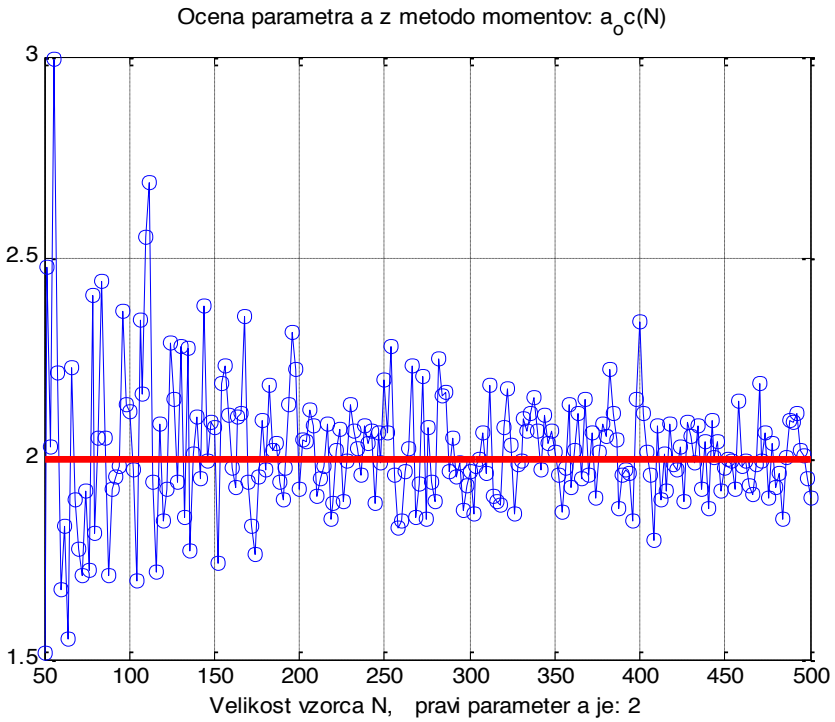


Slika 144: Histogram porazdelitve ocenjenega parametra $\hat{a}_{MLE}(N)$ glede na spreminjajočo se velikost vzorca ($50 \le N \le 500$) - metoda največje podobnosti.

Aritmetična sredina in standardni odklon pri metodi največje podobnosti sta:

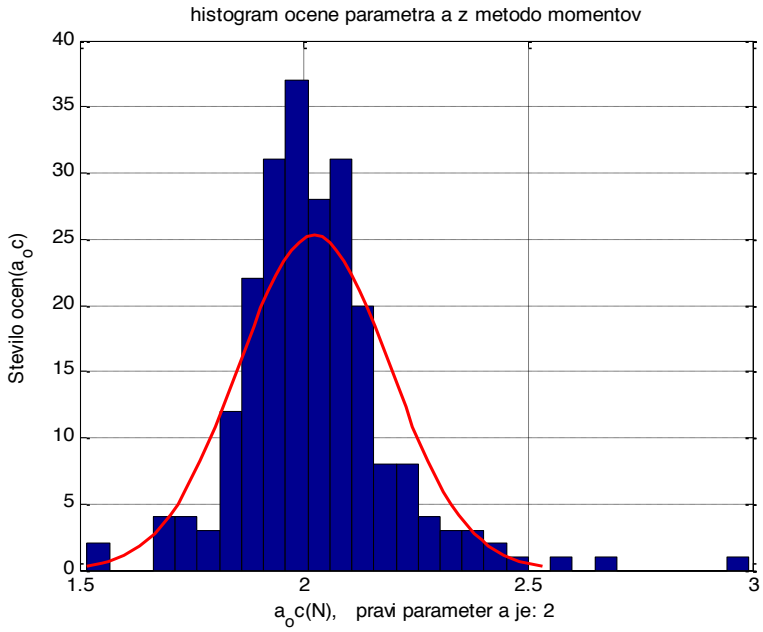
$$E\left(\hat{a}_{MLE}(N)\right) = 2.0138$$
$$STD\left(\hat{a}_{MLE}(N)\right) = 0.1439$$

Slika 145 prikazuje potek parametra $\hat{a}_{MM}(N)$, $N = 50, 52, 54, ..., 500$, ocenjenega z metodo momentov.



Slika 145: Potek parametra $\hat{a}_{MM}(N)$, $N = 50, 52, 54, ..., 500$, *ocenjenega z metodo momentov.*

Tudi tu z naraščanjem velikosti vzorca N varianca ocenjenega parametra pada, ocenjeni rezultati pa so vse bliže pravi vrednosti parametra. Histogram porazdelitve ocenjenega parametra $\hat{a}_{MM}(N)$ glede na spreminjajočo se velikost vzorca ($50 \le N \le 500$) prikazuje slika 146.



Slika 146: Histogram porazdelitve ocenjenega parametra $\hat{a}_{MM}(N)$ glede na spreminjajočo se velikost vzorca ($50 \le N \le 500$) - metoda momentov.

Aritmetična sredina in standardni odklon pri metodi momentov sta:

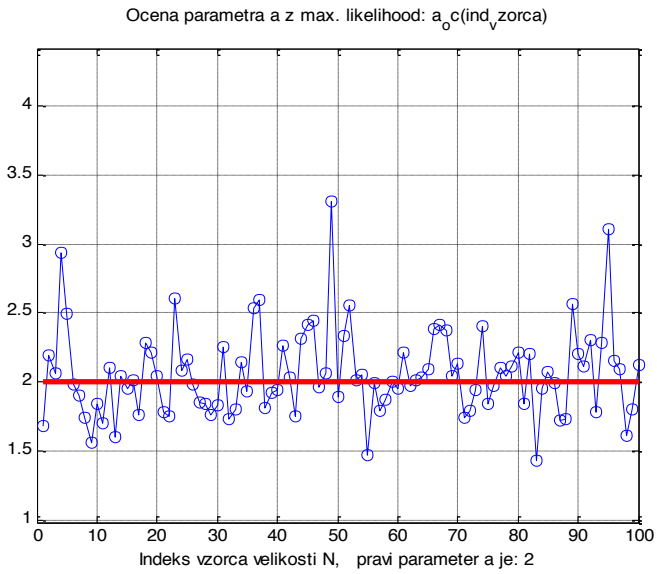
$$E\left(\hat{a}_{MM}(N)\right) = 2.0228$$
$$STD\left(\hat{a}_{MM}(N)\right) = 0.1694$$

Če poženemo program **ocenj_par1.m**, dobimo naslednji izpis v komandnem oknu:

```
Vnesi pravi parameter a
a =
    []
Vnesi velikost populacije M
Vnesi stevilo potegov vzorca K
K =
    []
Vnesi velikost vzorca N
a =
    2
M =
      1000
K =
  100
N =
   50
- Generiram populacijo Pareto porazdelitve

srednja vrednost populacije:
srvr =
   1.9896
ocenjevanje z metodo max. likelihood (cela populacija):
a1 =
   2.0154
ocenjevanje z metodo momentov (cela populacija):
a2 =
   2.0105

- Generiram K vzorcev velkosti N Pareto porazdelitve

srednja vrednost za oceno a1 pri veckrat vlecenem vzorcu velikosti N (MLE):
ans =
   2.0570
deviacija za oceno a1 pri veckrat vlecenem vzorcu velikosti N (MLE):
ans =
```

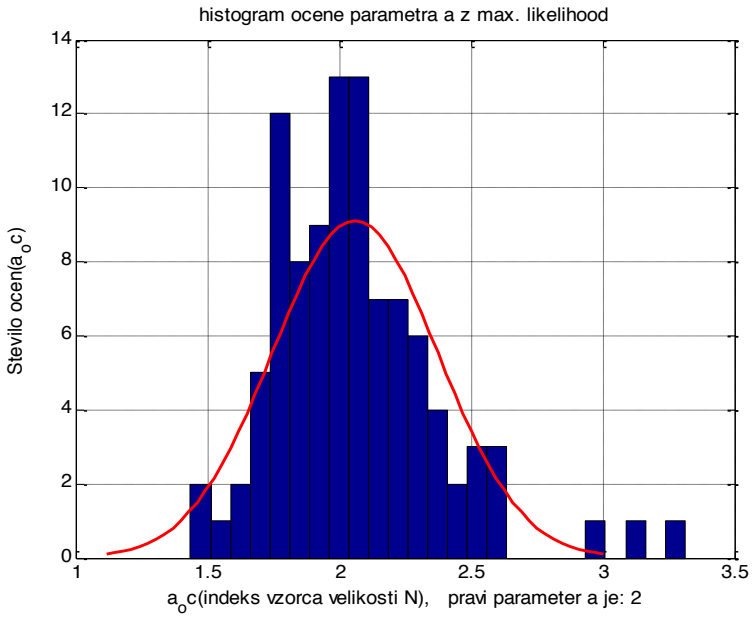
```
    0.3154

srednja vrednost za oceno a2 pri veckrat vlecenem vzorcu velikosti N (MM):
ans =
    2.1064
deviacija za oceno a2 pri veckrat vlecenem vzorcu velikosti N (MM):
ans =
    0.3569
```

Slika 147 prikazuje potek parametra $\hat{a}_{MLE}(i,N)=\hat{a}_{MLE}(i,50),\ i=1,2,...,100$, ocenjenega z metodo največje podobnosti.



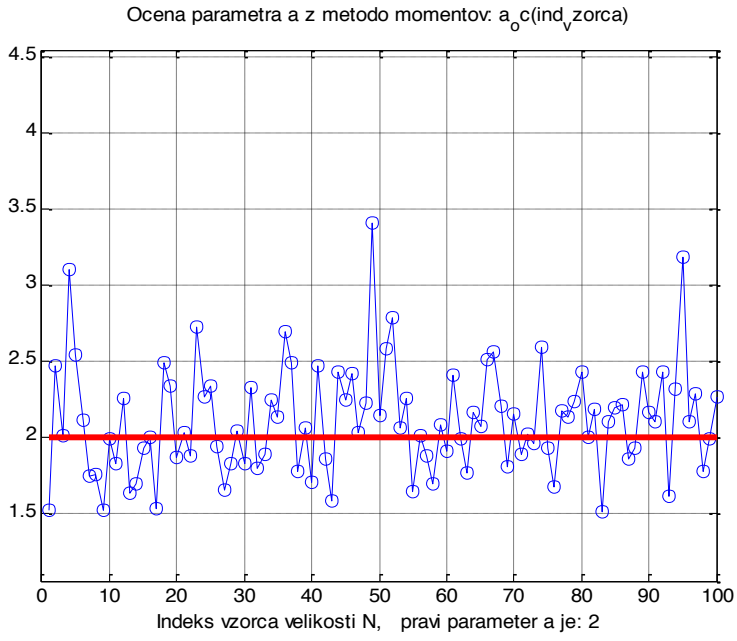
Slika 147: Potek parametra $\hat{a}_{MLE}(i,N)=\hat{a}_{MLE}(i,50),\ i=1,2,...,100$, ocenjenega z metodo največje podobnosti.

Histogram porazdelitve ocene $\hat{a}_{MLE}(i,50)$ glede na indeks vzorca $i=1,2,...,100$ prikazuje slika 148.



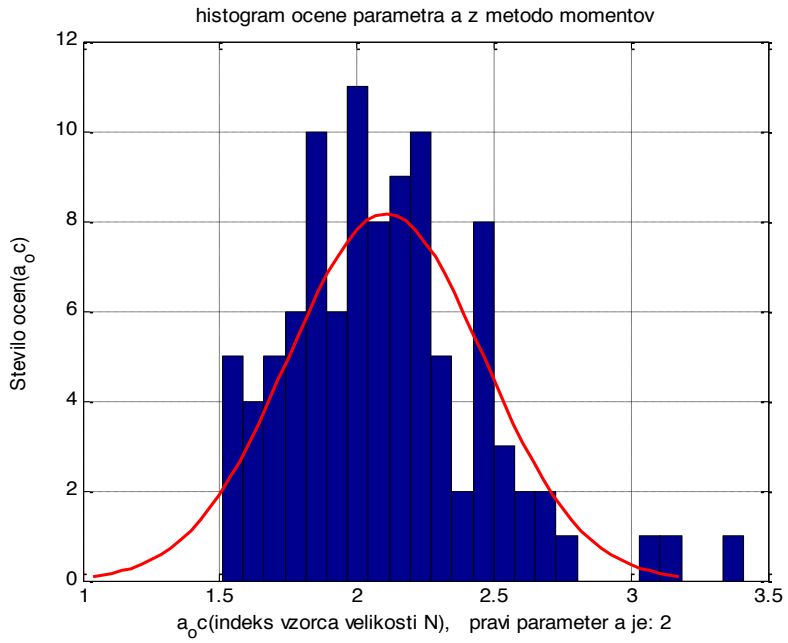
Slika 148: Histogram porazdelitve ocene $\hat{a}_{MLE}(i,50)$ *glede na indeks vzorca* $i=1,2,...,100$

Slika 149 prikazuje potek parametra $\hat{a}_{MM}(i,N)=\hat{a}_{MM}(i,50),\ i=1,2,...,100$, ocenjenega z metodo momentov.



Slika 149: Potek parametra $\hat{a}_{MM}(i,N)=\hat{a}_{MM}(i,50),\ i=1,2,...,100$*, ocenjenega z metodo momentov.*

Histogram porazdelitve ocene $\hat{a}_{MM}(i,50)$ glede na indeks vzorca $i=1,2,...,100$ prikazuje slika 150.



Slika 150: Histogram porazdelitve ocene $\hat{a}_{MM}(i,50)$ glede na indeks vzorca $i=1,2,...,100$

Aritmetična sredina in standardni odklon pri metodi največje podobnosti sta:

$$E\left(\hat{a}_{MLE}(i,50)\right)=2.057$$
$$STD\left(\hat{a}_{MLE}(i,50)\right)=0.3154$$

pri metodi momentov pa sta:

$$E\left(\hat{a}_{MM}(i,50)\right)=2.1064$$
$$STD\left(\hat{a}_{MM}(i,50)\right)=0.3569$$

7.8 Intervali zaupanja

O stopnji tveganja in intervalih zaupanja smo nekoliko spregovorili že v poglavju 3.3. Čeprav je točkasto ocenjevanje najbolj običajen način določanja ocen parametrov populacije, pa ostaja odprtih pri tem vrsto vprašanj, kot npr., koliko informacije o parametru vsebuje točkasta ocena, kako velika je napaka ocene, itn [Jesenko].

Intervalska ocena parametra θ je nek interval $\left[\hat{\theta}_1, \hat{\theta}_2\right]$, kjer sta $\hat{\theta}_1$ in $\hat{\theta}_2$ realizaciji ustreznih statistik Θ_1 in Θ_2, pri čemer velja [Jesenko]:

$$P\left(\hat{\theta}_1 \le \theta \le \hat{\theta}_2\right) = 1-\alpha \tag{7.35}$$

kjer je $1-\alpha$ vnaprej izbrana verjetnost (**stopnja zaupanja**). Pri njej ima interval $\left[\hat{\theta}_1, \hat{\theta}_2\right]$ $(1-\alpha)\cdot 100\%$ **interval zaupanja** za parameter θ, pri čemer sta $\hat{\theta}_1$ in $\hat{\theta}_2$ **meji zaupanja**. Verjetnost, da bo parameter θ ležal znotraj intervala $\left[\hat{\theta}_1, \hat{\theta}_2\right]$, je enaka $1-\alpha$, da ne bo ležal znotraj intervala $\left[\hat{\theta}_1, \hat{\theta}_2\right]$, pa je enaka α. Tej verjetnosti pravimo **stopnja tveganja**.

7.9 Ocenjevanje aritmetične sredine

V tem poglavju bomo obravnavali ocenjevanje aritmetične sredine populacije, pri čemer bomo upoštevali tudi intervale zaupanja.

Da bi prikazali, kako lahko ocenimo velikost napake pri točkasti oceni, vzemimo, da je aritmetična sredina vzorca točkasta ocena za aritmetično sredino **normalne** populacije pri znani varianci σ^2. Na osnovi izraza (6.4) vemo, da velja:

$$\begin{aligned} E\left(\bar{X}\right) &= \mu_{\bar{X}} = \mu \\ VAR\left(\bar{X}\right) &= \sigma^2_{\bar{X}} = \frac{\sigma^2}{n} \end{aligned} \tag{7.36}$$

Zapišemo lahko [Jesenko]:

$$P\left(z_{1-\frac{\alpha}{2}} \le Z \le z_{\frac{\alpha}{2}} \right) = 1-\alpha \tag{7.37}$$

pri čemer je $Z = \dfrac{\left(\bar{X}-\mu\right)}{\dfrac{\sigma}{\sqrt{n}}}$. Za $z_{1-\frac{\alpha}{2}}$ in $z_{\frac{\alpha}{2}}$ izberemo tisto posebno vrednost naključne spremenljivke *Z*, da velja [Jesenko]:

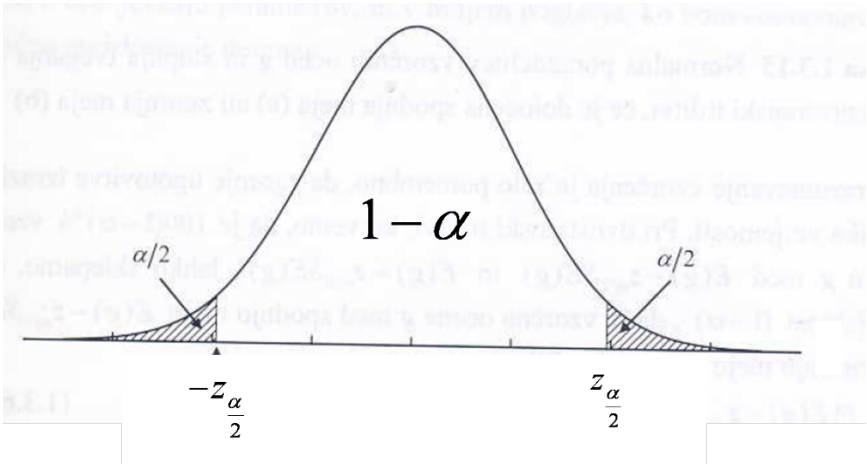
$$\begin{aligned} &\int_{-\infty}^{z_{1-\frac{\alpha}{2}}} \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{(t)^2}{2}} dt = \frac{\alpha}{2} \quad in \\ &\int_{z_{\frac{\alpha}{2}}}^{\infty} \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{(t)^2}{2}} dt = \frac{\alpha}{2} \end{aligned} \tag{7.38}$$

Ker je porazdelitev gostote verjetnosti standardne normalne spremenljivke simetrična glede na oordinatno os, gotovo velja:

$$z_{1-\frac{\alpha}{2}} = -z_{\frac{\alpha}{2}} \tag{7.39}$$

Tako dobimo (glej sliko 151):

$$P\left(-z_{\frac{\alpha}{2}} \le Z \le z_{\frac{\alpha}{2}} \right) = 1-\alpha \tag{7.40}$$



Slika 151: Verjetnost za interval zaupanja in verjetnost za interval tveganja

Sledi:

$$P\left(-z_{\frac{\alpha}{2}} \le \frac{\left(\bar{X}-\mu\right)}{\frac{\sigma}{\sqrt{n}}} \le z_{\frac{\alpha}{2}}\right)=1-\alpha$$

$$P\left(-z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \le \left(\bar{X}-\mu\right) \le z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right)=1-\alpha \qquad (7.41)$$

ali

$$P\left(\left|\bar{X}-\mu\right| \le z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right)=1-\alpha$$

Odtod sledi **izrek**:

Če je aritmetična sredina $\bar{X}$ *naključnega vzorca velikosti n iz normalne populacije z znano varianco* σ^2 *uporabljena kot cenilka aritmetične sredine populacije* μ*, je* $1-\alpha$ *verjetnost, da bo napaka točkaste ocene manjša ali enaka* $z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$ *[Jesenko].*

Primer 7.4.:

Stroj polni vreče z apnom. Teža napolnjenih vreč je normalna naključna spremenljivka s standardnim odklonom $\sigma = 0.7\ kg$*. Kaj lahko z verjetnostjo 0.99 sklepamo na osnovi vzorca 50 vreč o največji napaki točkaste ocene za povprečno težo vreč, napolnjenih z apnom [Jesenko]?*

Na osnovi podatkov naloge lahko zapišemo:

$$\begin{aligned}
&\alpha = 0.99 - 1 = 0.01 \\
&\frac{\alpha}{2} = \frac{0.01}{2} = 0.005 \\
&P\left(\left| \bar{X} - \mu \right| \le z_{0.005} \cdot \frac{0.7}{\sqrt{50}} \right) = 0.99 \\
&P\left(\left| \bar{X} - \mu \right| \le z_{0.005} \cdot 0.099 \right) = 0.99 \\
&e_{\max} = z_{0.005} \cdot 0.099
\end{aligned} \tag{7.42}$$

kjer je $e_{\max}$ maksimalna možna napaka točkaste ocene za povprečno težo. Vrednost $z_{0.005}$ dobimo z naslednjim ukazom v Matlabu:

```
>> z=abs(norminv(0.005,0,1))
z =
   2.5758
```

Sledi: $e_{\max} = z_{0.005} \cdot 0.099 = 2.5758 \cdot 0.099 = 0.255$.

Torej lahko na osnovi izbranega vzorca z 99% verjetnostjo trdimo, da napaka točkaste ocene za povprečno težo ne bo presegla 0.255 kg.

Izraz (7.41) lahko zapišemo tudi v naslednji obliki:

$$\begin{aligned}
&P\left(-z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \le \left(\bar{X} - \mu \right) \le z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \right) = 1 - \alpha \\
&P\left(-z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} - \bar{X} \le \left(-\mu \right) \le z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} - \bar{X} \right) = 1 - \alpha \\
&P\left(z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} + \bar{X} \ge \left(\mu \right) \ge -z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} + \bar{X} \right) = 1 - \alpha \\
&P\left(\bar{X} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \le \left(\mu \right) \le \bar{X} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \right) = 1 - \alpha \\
&P\left(\bar{X} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \right) = 1 - \alpha
\end{aligned} \tag{7.43}$$

Na osnovi tega izraza sledi naslednji **izrek**:

Če je $\bar{X}$ *vrednost aritmetične sredine naključnega vzorca velikosti n iz* ***normalne populacije*** *z* ***znano*** *varianco* σ^2*, potem je:*

$$\left[\bar{X} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right] \tag{7.44}$$

$(1-\alpha)\cdot 100\%$ **interval zaupanja za aritmetično sredino populacije**, ki ji vzorec pripada.

Primer 7.5.:

Kot se izkaže, je povprečna koncentracija cinka, izračunana na osnovi vzorca meritev iz 36 različnih lokacij na reki, enaka 2.6 grama na mililiter. Standardni odklon ***normalne*** *populacije je znan in je enak 0.3 grama na mililiter. Poiščite 95% in 99% interval zaupanja za aritmetično sredino koncentracije cinka v reki [Walpole].*

Primer spada v kategorijo: normalna populacija, znana varianca, velik vzorec (n>30).

Na osnovi podatkov naloge lahko zapišemo:

$$\begin{aligned}
&\bar{X} = 2.6 \\
&\sigma = 0.3 \\
&n = 36 \\
&\alpha_1 = 1 - 0.95 = 0.05 \quad \Rightarrow \quad \frac{\alpha_1}{2} = 0.025 \\
&\alpha_2 = 1 - 0.99 = 0.01 \quad \Rightarrow \quad \frac{\alpha_2}{2} = 0.005 \\
&I_1 = \left[\bar{X} - z_{\frac{\alpha_1}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\frac{\alpha_1}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right] = ? \\
&I_2 = \left[\bar{X} - z_{\frac{\alpha_2}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\frac{\alpha_2}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right] = ?
\end{aligned} \tag{7.45}$$

Vrednosti $z_{0.005}$ in $z_{0.025}$ dobimo z ukazoma v Matlabu:

```
>> z1=abs(norminv(0.005,0,1))
```

```
z1 =

    2.5758
>> z2=abs(norminv(0.025,0,1))

z2 =

    1.9600
```

Sledi:

$$I_1 = \left[2.6 - 1.96 \cdot \frac{0.3}{\sqrt{36}}, 2.6 + 1.96 \cdot \frac{0.3}{\sqrt{36}} \right] = [2.502, 2.698] \qquad (7.46)$$

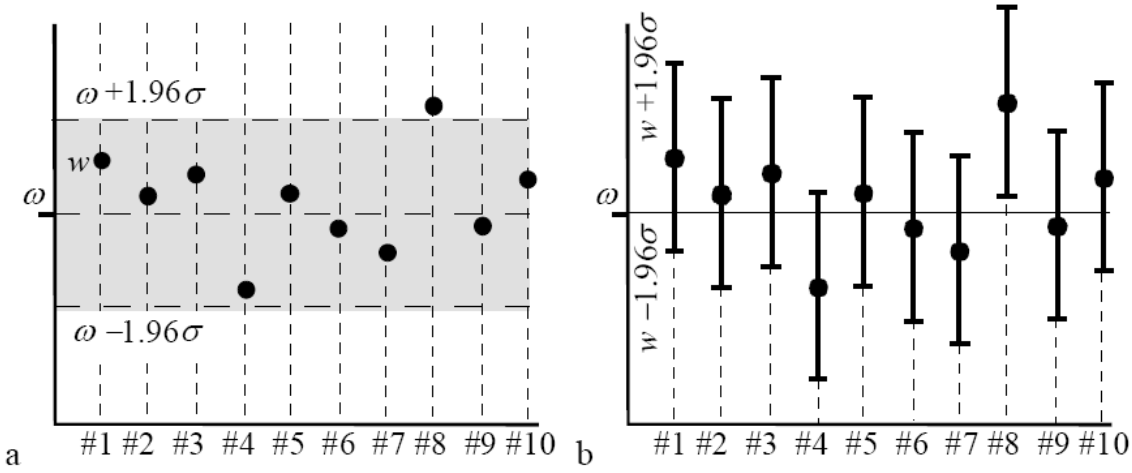
$$I_2 = \left[2.6 - 2.57 \cdot \frac{0.3}{\sqrt{36}}, 2.6 + 2.57 \cdot \frac{0.3}{\sqrt{36}} \right] = [2.4715, 2.7285]$$

Torej lahko s 95% verjetnostjo trdimo, da se povprečna koncentracija cinka nahaja znotraj intervala med 2.502 in 2.698, ter z 99% verjetnostjo trdimo, da se povprečna koncentracija cinka nahaja znotraj intervala med 2.4715 in 2.7285.

Primer 7.6.:

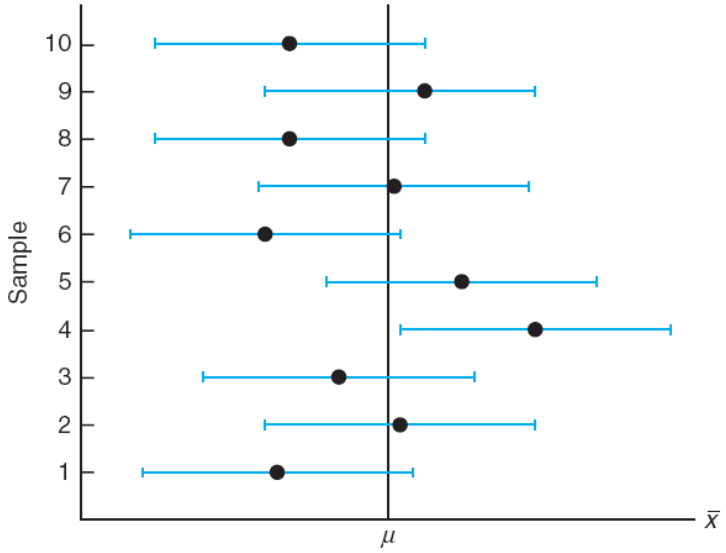
Na dva načina ilustrirajte interpretacijo intervala zaupanja!

Denimo imamo populacijo, katere aritmetična sredina je enaka $\mu = \omega$, spremenljivka $w = \hat{\omega}$ pa predstavlja oceno aritmetične sredine populacije. Potem si lahko interpretiramo npr. 95% interval zaupanja na način, kot ga prikazuje slika 152 [De Sa].



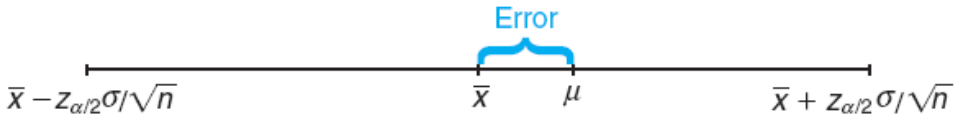
Slika 152: Dve interpretaciji 95% intervala zaupanja [De Sa]

Pri prvi interpretaciji (a) na sliki 152 lahko trdimo, da bo pri daljšem ocenjevanju 95% ocen padlo v šrafiran interval, pri drugi interpretaciji (b) na sliki 152 pa lahko trdimo, da bo pri daljšem ocenjevanju 95% intervalov vsebovalo neznan parameter – aritmetično sredino populacije. Podoben pomen kot slika 152 b ima tudi slika 153 [Walpole].



Slika 153: Interpretacija intervala zaupanja v oceno aritmetične sredine populacije pri različnih vzorčnih ocenah (sample - vzorec) [Walpole]

Slika 154 pa elegantno ilustrira napako v ocenjevanju aritmetične sredine populacije [Walpole].



Slika 154: Ilustracija napake (error) v intervalnem ocenjevanju aritmetične sredine populacije [Walpole]

Primer 7.7.:

*Dano imamo **normalno** populacijo s poznanim standardnim odklonom* $\sigma = 1.5$. *Iz populacije je vzet vzorec velikosti 9 enot, ki ima aritmetično sredino* $\bar{x} = 14.3$. *Poiščite 95% interval zaupanja za aritmetično sredino populacije [Bernstein].*

Primer spada v kategorijo: normalna populacija, znana varianca, mali vzorec (n < 30).

Kljub temu, da je majhen vzorec, vseeno lahko uporabimo interval zaupanja (7.44), saj gre za normalno populacijo. Dobimo:

$$\alpha = 1 - 0.95 = 0.05 \quad \Rightarrow \quad \frac{\alpha}{2} = 0.025 \quad \Rightarrow \quad z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$$

$$\left[\bar{X} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right] = \left[14.3 - 1.96 \cdot \frac{1.5}{\sqrt{9}}, 14.3 + 1.96 \cdot \frac{1.5}{\sqrt{9}}\right] = \tag{7.47}$$

$$= [14.3 - 1.96 \cdot 0.5, 14.3 + 1.96 \cdot 0.5] = [14.3 - 0.98, 14.3 + 0.98] =$$

$$= [13.32, 15.28]$$

Določitev velikosti vzorca

V izrazu (7.42) smo se seznanili z $e_{\max}$, ki je maksimalna možna napaka točkaste ocene za aritmetično sredino populacije. Zanjo velja:

$$e_{\max} = z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \tag{7.48}$$

Odtod pa lahko določimo velikost vzorca pri **poznani varianci** [Bernstein]:

$$e_{\max}{}^2 = z_{\frac{\alpha}{2}}{}^2 \cdot \frac{\sigma^2}{n}$$
$$n = \left(\frac{z_{\frac{\alpha}{2}} \cdot \sigma}{e_{\max}} \right)^2 \tag{7.49}$$

Primer 7.8.:

*Načrtujemo eksperiment, ki vključuje naključni vzorec **normalne** populacije s poznanim standardnim odklonom* $\sigma = 1.5$. *Pri tem moramo vzeti takšen interval, ki bo dal 95% zaupanje v oceno aritmetične sredine populacije, pri čemer maksimalna možna napaka točkaste ocene ne sme preseči vrednost* $e_{\max} = 1$. *Kako velik naj bo vzorec? [Bernstein]*

Imamo:

$$\alpha = 1 - 0.95 = 0.05 \quad \Rightarrow \quad \frac{\alpha}{2} = 0.025 \quad \Rightarrow \quad z_{\frac{\alpha}{2}} = z_{0.025} = 1.96 \tag{7.50}$$
$$\sigma = 1.5$$
$$e_{\max} = 1$$

Sledi:

$$n = \left(\frac{z_{\frac{\alpha}{2}} \cdot \sigma}{e_{\max}} \right)^2 = \left(\frac{1.96 \cdot 1.5}{1} \right)^2 = 8.6436 \approx 9 \tag{7.51}$$

Torej mora vzorec vsebovati 9 enot populacije.

Primer 7.9.:

Dano imamo normalno populacijo $N\left(\mu,\sigma^2\right)=N\left(\mu,4\right)$*, iz katere je izvlečen vzorec s frekvenčno porazdelitvijo:*

x_j	0	1	2	3	4
n_j	1	4	6	12	2

Ocenite parameter μ *in izračunajte 90% interval zaupanja v vzorčno oceno parametra [Elezović].*

Ocena aritmetične sredine je enaka:

$$\overline{x}=\frac{0\cdot 1+1\cdot 4+2\cdot 6+3\cdot 12+4\cdot 2}{1+4+6+12+2}=\frac{60}{25}=2.4 \tag{7.52}$$

Imamo še:

$$\begin{aligned}&\alpha=1-0.90=0.10 \quad \Rightarrow \quad \frac{\alpha}{2}=0.05 \quad \Rightarrow \quad z_{\frac{\alpha}{2}}=z_{0.05}=1.6449\\&\sigma=2\\&n=25\end{aligned} \tag{7.53}$$

Za izračun $z_{\frac{\alpha}{2}}=z_{0.05}=1.6449$ smo poklicali naslednji ukaz v Matlabu:

```
>> z=abs(norminv(0.05,0,1))

z =

   1.6449
```

Dobimo:

$$\begin{aligned}&\left[\overline{X}-z_{\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}},\overline{X}+z_{\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}}\right]=\\&=\left[2.4-1.6449\cdot\frac{2}{\sqrt{25}},2.4-1.6449\cdot\frac{2}{\sqrt{25}}\right]=\left[1.742,3.058\right]\end{aligned} \tag{7.54}$$

Primer 7.10.:

Prometni inženirji ugotavljajo, če je potrebno razširiti določen odsek ceste. V ta namen zmerijo z radarjem 85 vozil iz izračunajo povprečno hitrost 66.3 km/h. Iz prejšnjih študij vedo, da je $\sigma = 8.3$ *km/h. Kakšen je* ***<u>približen</u>*** *95% interval zaupanja v oceno aritmetične sredine populacije? [Bernstein]*

Primer spada v kategorijo: neznana (poljubna) populacija, znana varianca, velik vzorec (n > 30).

Kljub temu, da je porazdelitev populacije poljubna, lahko vseeno uporabimo centralni limitni teorem, saj je velikost vzorca velika (n > 30). Torej lahko predpostavimo, da se vzorčna porazdelitev za $\bar{X}$ porazdeljuje približno normalno.

Torej imamo:

$$\left[\bar{X} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right] = \\ = \left[66.3 - 1.96 \cdot \frac{8.3}{\sqrt{85}}, 66.3 + 1.96 \cdot \frac{8.3}{\sqrt{85}}\right] = \left[64.5355, 68.0645\right] \tag{7.55}$$

<u>Primer 7.11.:</u>

Veriga trgovin s 24 urnim servisom išče lokacijo za novo trgovino. V ta namen 60 zaporednih noči merijo pretok prometa (število vozil, ki gredo mimo določene lokacije) v intervalu od 22. ure do 5. ure. Kot se izkaže, na osnovi meritev izračunajo $\bar{x} = 238.2, s = 31.32$. *Ker je promet gostejši med vikendi, porazdelitev ni normalna. Določite 95% interval zaupanja za aritmetično sredino populacije [Bernstein].*

Primer spada v kategorijo: neznana (poljubna) populacija, neznana varianca, velik vzorec (n > 30).

Kljub temu, da je porazdelitev populacije poljubna, lahko vseeno uporabimo centralni limitni teorem, saj je velikost vzorca velika (n > 30). Torej lahko predpostavimo, da se vzorčna porazdelitev za $\bar{X}$ porazdeljuje približno normalno.

Dobimo:

$$\left[\bar{X} - z_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}, \bar{X} + z_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}\right] =$$
$$= \left[238.2 - 1.96 \cdot \frac{31.32}{\sqrt{60}}, 238.2 + 1.96 \cdot \frac{31.32}{\sqrt{60}}\right] = \tag{7.56}$$
$$= [238.2 - 7.9251, 238.2 + 7.9251] = [230.2749, 246.1251]$$

Vpeljava *t* porazdelitve

V primeru neznane variance in normalne populacije vpeljemo *t* porazdelitev, ne glede na to, a je velik ali mali vzorec. Tako dobimo **točen** izračun za interval zaupanja. Sicer bi lahko ostali pri *z* porazdelitvi, vendar bi potem dobili zgolj **približen** interval zaupanja. Možni izračuni intervala zaupanja so prikazani v tabeli na sliki 155 [Bernstein].

Table 14.3

Population	Sample size	Confidence interval	Exact or approximate	Defining section
Known σ				
Normal	$n \geq 30$	$\bar{x} \pm z_{\alpha/2}\sigma_{\bar{x}}$	Exact	14.9
Normal	$n < 30$	$\bar{x} \pm z_{\alpha/2}\sigma_{\bar{x}}$	Exact	14.9
Not normal	$n \geq 30$	$\bar{x} \pm z_{\alpha/2}\sigma_{\bar{x}}$	Approximate	14.13
Not normal	$n < 30$	?	?	?
Unknown σ				
Normal	$n \geq 30$	$\bar{x} \pm t_{\alpha/2,\nu}s_{\bar{x}}$	Exact for exact t	14.23
		or		
		$\bar{x} \pm z_{\alpha/2}s_{\bar{x}}$	Approximate	14.23
Normal	$n < 30$	$\bar{x} \pm t_{\alpha/2,\nu}s_{\bar{x}}$	Exact	14.21
Not normal	$n \geq 30$	$\bar{x} \pm z_{\alpha/2}s_{\bar{x}}$	Approximate	14.24
Not normal	$n < 30$	?	?	?

Slika 155: Možni izračuni intervalov zaupanja (Population - populacija, Sample size - velikost vzorca, Exact - točen, Approximate - približen, known - poznana, unknown - neznana, normal - normalna, not normal - ni normalna (poljubna)) [Bernstein]

V primeru majhnega vzorca, ki pripada poljubni porazdelitvi, ni nobenega standarda glede izračuna intervalov zaupanja. Možna rešitev bi bila v uporabi testov brez porazdelitev ali v neparameterskih testih [Kmenta].

Interval zaupanja za *t* statistiko izpeljemo podobno kot pri *z* statistiki (glej izraz 7.43):

$$
\begin{aligned}
&P\left(-t_{\frac{\alpha}{2},n-1}\cdot\frac{s}{\sqrt{n}}\le\left(\bar{X}-\mu\right)\le t_{\frac{\alpha}{2},n-1}\cdot\frac{s}{\sqrt{n}}\right)=1-\alpha\\
&P\left(-t_{\frac{\alpha}{2},n-1}\cdot\frac{s}{\sqrt{n}}-\bar{X}\le\left(-\mu\right)\le t_{\frac{\alpha}{2},n-1}\cdot\frac{s}{\sqrt{n}}-\bar{X}\right)=1-\alpha\\
&P\left(t_{\frac{\alpha}{2},n-1}\cdot\frac{s}{\sqrt{n}}+\bar{X}\ge\left(\mu\right)\ge-t_{\frac{\alpha}{2},n-1}\cdot\frac{s}{\sqrt{n}}+\bar{X}\right)=1-\alpha\\
&P\left(\bar{X}-t_{\frac{\alpha}{2},n-1}\cdot\frac{s}{\sqrt{n}}\le\left(\mu\right)\le\bar{X}+t_{\frac{\alpha}{2},n-1}\cdot\frac{s}{\sqrt{n}}\right)=1-\alpha\\
&P\left(\bar{X}-t_{\frac{\alpha}{2},n-1}\cdot\frac{s}{\sqrt{n}}\le\mu\le\bar{X}+t_{\frac{\alpha}{2},n-1}\cdot\frac{s}{\sqrt{n}}\right)=1-\alpha
\end{aligned}
\qquad (7.57)
$$

Primer 7.12.:

Mizarski mojster je želel določiti povprečni čas sušenja novo prepleskanih smrekovih letev. Izbral je 15 letev, jih prepleskal in dobil povprečni čas sušenja 123.5 min in standardni odklon 11.2 minute. določite 95% interval zaupanja za povprečni čas sušenja letev. Predpostavimo normalno populacijo, kjer je mali vzorec (n = 15).

Imamo:

$$
\begin{aligned}
&\alpha=1-0.95=0.05 \quad\Rightarrow\quad \frac{\alpha}{2}=0.025 \quad\Rightarrow\quad t_{\frac{\alpha}{2},n-1}=t_{0.025,14}=2.1448\\
&\sigma=11.2\\
&n=15\\
&s=123.5
\end{aligned}
\qquad (7.58)
$$

Pri tem smo $t_{\frac{\alpha}{2},n-1}=t_{0.025,14}=2.1448$ dobili z naslednjim ukazom v Matlabu:

```
>> t = abs(tinv(0.025,15-1))

t =

   2.1448
```

Sledi:

$$\begin{aligned}&\left[\bar{X}-t_{\frac{\alpha}{2},n-1}\cdot\frac{s}{\sqrt{n}},\bar{X}+t_{\frac{\alpha}{2},n-1}\cdot\frac{s}{\sqrt{n}}\right]=\\&=\left[123.5-2.1448\cdot\frac{11.2}{\sqrt{15}},123.5+2.1448\cdot\frac{11.2}{\sqrt{15}}\right]=\\&=\left[117.2976,129.7024\right]\end{aligned} \tag{7.59}$$

Torej smemo s 95% zaupanjem trditi, da je povprečni čas sušenja smrekovih letev med 117.29 minute in 129.70 minute.

Primer 7.13.:

Vzemimo, da je spremenljivka X število ur branja dnevnih časopisov na teden, ki se porazdeljuje normalno. Na osnovi podatkov 7 naključno izbranih oseb (mali vzorec) ocenimo interval zaupanja za aritmetično sredino pri 10% tveganju ($\alpha=0.1$). Podatki so [Jurišić]:

1. oseba: 5 ur

2. oseba: 7 ur

3. oseba: 9 ur

4. oseba: 7 ur

5. oseba: 6 ur

6. oseba: 10 ur

7. oseba: 5 ur

Izračunajmo:

$$\bar{X}=\frac{\sum_{i=1}^{7}x_i}{n}=\frac{5+7+9+7+6+10+5}{7}=\frac{49}{7}=7 \tag{7.60}$$

in:

$$
\begin{aligned}
s^2 &= \frac{\sum_{i=1}^{7}\left(x_i-\bar{X}\right)^2}{n-1} = \\
&= \frac{(5-7)^2+(7-7)^2+(9-7)^2+(7-7)^2+(6-7)^2+(10-7)^2+(5-7)^2}{7-1} = \\
&= \frac{4+0+4+0+1+9+4}{6} = \frac{22}{6} = 3.6667
\end{aligned}
\tag{7.61}
$$

Kritično vrednost $t_{\frac{\alpha}{2},n-1} = t_{0.05,6} = 1.9432$ dobimo z naslednjim ukazom v Matlabu:

```
>> t = abs(tinv(0.05,7-1))

t =

   1.9432
```

Sledi:

$$
\begin{aligned}
&\left[\bar{X}-t_{\frac{\alpha}{2},n-1}\cdot\frac{s}{\sqrt{n}},\bar{X}+t_{\frac{\alpha}{2},n-1}\cdot\frac{s}{\sqrt{n}}\right] = \\
&= \left[7-1.9432\cdot\frac{\sqrt{3.6667}}{\sqrt{7}},7+1.9432\cdot\frac{\sqrt{3.6667}}{\sqrt{7}}\right] = \\
&= \left[5.5936,8.4064\right]
\end{aligned}
\tag{7.62}
$$

7.10 Ocenjevanje razlike aritmetičnih sredin

Denimo imamo 2 neskončni populaciji. Ena naj ima aritmetično sredino in varianco $\mu_1 = E(X_1), \sigma_1^2 = VAR(X_1),$ druga pa aritmetično sredino in varianco $\mu_2 = E(X_2), \sigma_2^2 = VAR(X_2)$ (glej sliko 156). Če iz prve populacije izberemo naključni

vzorec $X_{11}, X_{12}, ..., X_{1n_1}$ velikosti n_1, iz druge populacije pa izberemo naključni vzorec $X_{21}, X_{22}, ..., X_{2n_2}$ velikosti n_2, potem velja podobno kot pri izrazu (6.4) [Jesenko]:

$$\begin{aligned} &E\left(\bar{X}\right) = E\left(\bar{X}_1 - \bar{X}_2\right) = E\left(\bar{X}_1\right) - E\left(\bar{X}_2\right) = \\ &= E\left(\frac{1}{n}\sum_{i=1}^{n_1} X_{1i}\right) - E\left(\frac{1}{n}\sum_{i=1}^{n_2} X_{2i}\right) = \mu_1 - \mu_2 = \mu \\ &VAR\left(\bar{X}\right) = VAR\left(\bar{X}_1 - \bar{X}_2\right) = VAR\left(\bar{X}_1\right) + VAR\left(-\bar{X}_2\right) = \\ &= VAR\left(\bar{X}_1\right) + VAR\left(\bar{X}_2\right) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \end{aligned} \tag{7.63}$$

Izrek:

Če sta populaciji, iz katerih izbiramo vzorca, normalni, potem je naključna spremenljivka $\bar{X} = \bar{X}_1 - \bar{X}_2$ *tudi normalna naključna spremenljivka z matematičnim upanjem* $\mu = \mu_1 - \mu_2$ *in varianco* $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$ *[Jesenko].*

Odtod sledi, da je [Jesenko]:

$$\begin{aligned} Z &= \frac{\left(\bar{X} - \mu\right)}{\sqrt{VAR\left(\bar{X}\right)}} = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \\ &= \frac{\left(\bar{X}_1 - \mu_1\right) - \left(\bar{X}_2 - \mu_2\right)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \end{aligned} \tag{7.64}$$

standardizirana naključna spremenljivka in pri dani verjetnosti α vedno lahko dobimo takšni vrednosti $-z_{\frac{\alpha}{2}}$ in $z_{\frac{\alpha}{2}}$, da velja:

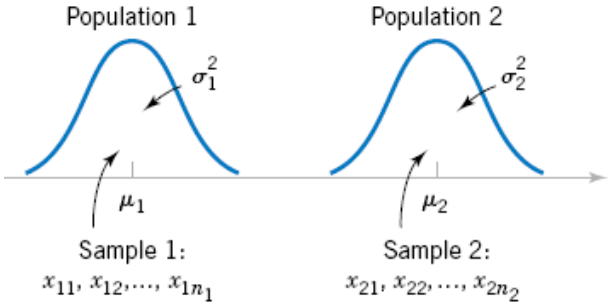
$$P\left(-z_{\frac{\alpha}{2}} \le Z \le z_{\frac{\alpha}{2}}\right) = 1 - \alpha \tag{7.65}$$

Sledi:

$$P\left(-z_{\frac{\alpha}{2}} \le \frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \le z_{\frac{\alpha}{2}}\right)=1-\alpha \qquad (7.66)$$

Po podobni izpeljavi kot v izrazu (7.43) dobimo:

$$P\left((\bar{X}_1-\bar{X}_2)-z_{\frac{\alpha}{2}}\cdot\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \le \mu_1-\mu_2 \le (\bar{X}_1-\bar{X}_2)+z_{\frac{\alpha}{2}}\cdot\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right)=1-\alpha \qquad (7.67)$$



Slika 156: Dve neodvisni populaciji (population) (sample - vzorec) [Montgomery 1]

Na osnovi tega izraza sledi naslednji **izrek**:

Če sta $\bar{X}_1$ *in* $\bar{X}_2$ *vrednosti aritmetičnih sredin neodvisnih naključnih vzorcev velikosti* velikosti n_1 in n_2 *iz dveh* ***normalnih populacij*** *z* ***znanima*** *variancama* σ_1^2 *in* σ_2^2*, potem je:*

$$\left[(\bar{X}_1-\bar{X}_2)-z_{\frac{\alpha}{2}}\cdot\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}, (\bar{X}_1-\bar{X}_2)+z_{-\frac{\alpha}{2}}\cdot\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right] \qquad (7.68)$$

$(1-\alpha)\cdot 100\%$ **interval zaupanja za razliko aritmetičnih sredin populacij** $\mu=\mu_1-\mu_2$, katerima vzorca pripadata.

Če sta velikosti obeh vzorcev $n_1 \ge 30$ ***in*** $n_2 \ge 30$***, velja centralni limitni teorem in lahko zahtevo o normalnosti populacij opustimo*** [Jesenko].

Primer 7.14.:

Določite 95% interval zaupanja za razliko med povprečnima časoma učenja določene snovi študentov in študentk. Na osnovi naključnih vzorcev 40 študentov in 50 študentk so ugotovili, da je bil povprečni čas učenja študentov 12.5 ure in povprečni čas učenja študentk 11.9 ure. Standardna odklona obeh populacij sta znana: $\sigma_1 = 1.5$ *ure in* $\sigma_2 = 2.1$ *ure.*

Na osnovi podatkov naloge lahko zapišemo:

$$\begin{aligned}
&\bar{X}_1 = 12.5 \\
&\bar{X}_2 = 11.9 \\
&\sigma_1 = 1.5 \\
&\sigma_2 = 2.1 \\
&n_1 = 40 \\
&n_2 = 50 \\
&\alpha = 1 - 0.95 = 0.05 \quad \Rightarrow \quad \frac{\alpha}{2} = 0.025 \\
&\left[\left(\bar{X}_1 - \bar{X}_2 \right) - z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \left(\bar{X}_1 - \bar{X}_2 \right) + z_{-\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right] = ?
\end{aligned} \tag{7.69}$$

Dobimo:

$$\begin{aligned}
&\left[(12.5 - 11.9) - 1.96 \cdot \sqrt{\frac{1.5^2}{40} + \frac{2.1^2}{50}}, (12.5 - 11.9) + 1.96 \cdot \sqrt{\frac{1.5^2}{40} + \frac{2.1^2}{50}} \right] = \\
&= [-0.1449, 1.3449]
\end{aligned} \tag{7.70}$$

<u>Če sta vzorca večja od 30, varianci (poljubnih!) populacij pa neznani, ju lahko nadomestimo z vzorčnima variancama in postopamo na enak način kot pri znanih variancah [Jesenko].</u>

<u>Primer 7.15.:</u>

Iz dveh populacij z neznanima variancama sta vzeta naključna vzorca velikosti $n_1 = 45$ *in* $n_2 = 40$. *Za prvi vzorec je ugotovljeno:* $\bar{x}_1 = 2.16, s_1 = 0.358$, *za drugi vzorec pa:*

$\bar{x}_2 = 1.98, s_2 = 0.352$. *Kakšen je približen 95% interval zaupanja za razliko* $\mu = \mu_1 - \mu_2$ *[Bernstein]?*

Imamo:

$$\left[\left(\bar{X}_1 - \bar{X}_2\right) - z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \left(\bar{X}_1 - \bar{X}_2\right) + z_{-\frac{\alpha}{2}} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right] =$$
$$= \left[\left(2.16 - 1.98\right) - 1.96 \cdot \sqrt{\frac{0.358^2}{45} + \frac{0.352^2}{40}}, \left(2.16 - 1.98\right) + 1.96 \cdot \sqrt{\frac{0.358^2}{45} + \frac{0.352^2}{40}}\right] = \quad (7.71)$$
$$= \left[0.18 - 0.1511, 0.18 + 0.1511\right] = \left[0.0289, 0.3311\right]$$

<u>Za majhne vzorce in neznane variance opisani postopki žal več ne veljajo, tudi pri normalnih populacijah ne</u> [Bernstein].

Če sta varianci σ_1^2 in σ_2^2 enaki (σ^2), potem izraz (7.64) preide v obliko:

$$Z = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}} = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\sigma \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad (7.72)$$

Neznano varianco σ^2 nadomestimo z naslednjo uteženo varianco na osnovi obeh vzorcev [Jesenko, Montgomery 1, Bernstein]:

$$\sigma^2 \approx S_p^2 = \frac{(n_1-1)\cdot S_1^{\,2} + (n_2-1)\cdot S_2^{\,2}}{n_1+n_2-2} \tag{7.73}$$

Tako dobimo:

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p \cdot \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1-1)\cdot S_1^{\,2} + (n_2-1)\cdot S_2^{\,2}}{n_1+n_2-2}} \cdot \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \tag{7.74}$$

Dokazati se da [Jesenko], da se spremenljivka v izrazu (7.74) porazdeljuje v skladu s *t* porazdelitvijo z n_1+n_2-2 prostostnimi stopnjami, tako lahko zapišemo:

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p \cdot \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1-1)\cdot S_1^{\,2} + (n_2-1)\cdot S_2^{\,2}}{n_1+n_2-2}} \cdot \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \tag{7.75}$$

Dokazati se tudi da, da tako pridemo do prave in ne le približne rešitve pri ocenjevanju [Bernstein]. Podobno kot v izrazu (7.57) lahko zapišemo:

$$P\left((\bar{X}_1 - \bar{X}_2) - t_{\frac{\alpha}{2},n_1+n_2-2}\cdot S_p \cdot \sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \le \mu \le (\bar{X}_1 - \bar{X}_2) + t_{\frac{\alpha}{2},n_1+n_2-2}\cdot S_p \cdot \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right) = 1-\alpha \tag{7.76}$$

$$\mu = \mu_1 - \mu_2$$

Interval zaupanja za razliko aritmetičnih sredin $\mu = \mu_1 - \mu_2$ pa je enak:

$$\left[(\bar{X}_1 - \bar{X}_2) - t_{\frac{\alpha}{2},n_1+n_2-2}\cdot S_p \cdot \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}, (\bar{X}_1 - \bar{X}_2) + t_{\frac{\alpha}{2},n_1+n_2-2}\cdot S_p \cdot \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right] \tag{7.77}$$

Primer 7.16.:

Primerjali so količino alkohola v dveh vrstah viskija. Deset steklenic viskija vrste A je v povprečju vsebovalo 39.8% alkohola s standardnim odklonom 0.9%, medtem ko je osem steklenic viskija vrste B v povprečju vsebovalo 41.2% alkohola s standardnim odklonom 1.1%. Predpostavimo, da gre za dva neodvisna naključna vzorca iz normalnih populacij z enako varianco. Določite 95% interval zaupanja za razliko aritmetičnih sredin vsebnosti alkohola v obeh vrstah viskija [Jesenko].

Na osnovi podatkov naloge lahko zapišemo:

$$\begin{aligned}
&\bar{X}_1 = 39.8 \\
&\bar{X}_2 = 41.2 \\
&s_1 = 0.9 \\
&s_2 = 1.1 \\
&n_1 = 10 \\
&n_2 = 8 \\
&\sigma_1 = \sigma_2 = \sigma \\
&\alpha = 1 - 0.95 = 0.05 \quad \Rightarrow \quad \frac{\alpha}{2} = 0.025 \\
&\left[\left(\bar{X}_1 - \bar{X}_2 \right) - t_{\frac{\alpha}{2}, n_1+n_2-2} \cdot S_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \left(\bar{X}_1 - \bar{X}_2 \right) + t_{\frac{\alpha}{2}, n_1+n_2-2} \cdot S_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right] = ?
\end{aligned} \tag{7.78}$$

Dobimo:

$$\begin{aligned}
&\left[\left(\bar{X}_1 - \bar{X}_2 \right) - t_{\frac{\alpha}{2}, n_1+n_2-2} \cdot S_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \left(\bar{X}_1 - \bar{X}_2 \right) + t_{\frac{\alpha}{2}, n_1+n_2-2} \cdot S_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right] = \\
&= \left[\left(39.8 - 41.2 \right) - 2.12 \cdot S_p \cdot \sqrt{\frac{1}{10} + \frac{1}{8}}, \left(39.8 - 41.2 \right) + 2.12 \cdot S_p \cdot \sqrt{\frac{1}{10} + \frac{1}{8}} \right] = \\
&= \left[-1.4 - 1.0056 \cdot S_p, -1.4 + 1.0056 \cdot S_p \right]
\end{aligned} \tag{7.79}$$

pri čemer smo $t_{\frac{\alpha}{2}, n_1+n_2-2}$ dobili z naslednjim ukazom v Matlabu:

```
>> t=abs(tinv(0.025,10+8-2))

t =

   2.1199
```

Izračunati moramo še S_p:

$$S_p = \sqrt{\frac{(n_1-1)\cdot S_1^{\,2} + (n_2-1)\cdot S_2^{\,2}}{n_1+n_2-2}} = \sqrt{\frac{(10-1)\cdot 0.9^2 + (8-1)\cdot 1.1^2}{10+8-2}} = \tag{7.80}$$

$$= \sqrt{\frac{9\cdot 0.9^2 + 7\cdot 1.1^2}{16}} = 0.25\cdot\sqrt{9\cdot 0.9^2 + 7\cdot 1.1^2} = 0.9925$$

Sledi:

$$\begin{aligned} &\left[-1.4-1.0056\cdot 0.9925, -1.4+1.0056\cdot 0.9925\right] = \\ &= \left[-1.4-0.998, -1.4+0.998\right] = \left[-2.398, -0.402\right] \end{aligned} \tag{7.81}$$

Če pa varianci $\sigma_1^{\,2}$ in $\sigma_2^{\,2}$ nista enaki, je postopek reševanja **nekoliko drugačen od pravkar prikazanega**. Podrobnosti o tem si bralec lahko pogleda v virih [Jurišić, Montgomery 1].

7.11 Ocenjevanje deležev

V mnogih problemih moramo oceniti delež, verjetnost, odstotek ali stopnjo, kot recimo delež slabih izdelkov v veliki pošiljki, odstotek osnovnošolcev, ki imajo IQ nad 115, stopnjo umrljivosti zaradi neke bolezni, itn. V večini teh primerov predpostavljamo, da je populacija, iz katere izbiramo vzorec, binomska. Naloga pa je, kako oceniti verjetnost *p* in intervale zaupanja v binomski populaciji [Jesenko]. Kot vemo, za binomsko populacijo velja (glej izraze (5.8), (5.12) in (5.91)):

(7.82)

$$\mu = E(X) = n \cdot p$$
$$\sigma^2 = VAR(X) = n \cdot p \cdot q$$
$$P(X = x, n, p, q) \approx \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\sqrt{n \cdot p \cdot q}} \cdot e^{-\frac{(x-n \cdot p)^2}{2 \cdot n \cdot p \cdot q}} = N\left(n \cdot p, \sqrt{n \cdot p \cdot q}\right) =$$
$$= N\left(n \cdot p, \sqrt{n \cdot p \cdot (1-p)}\right)$$

Tako je naključna spremenljivka:

$$Z = \frac{X - n \cdot p}{\sqrt{n \cdot p \cdot (1-p)}} \tag{7.82}$$

približno standardizirana normalna naključna spremenljivka [Jesenko]. V poglavju 2.19 smo videli, da je točkasta ocena parametra z metodo največje podobnosti enaka (glej (2.160)):

$$\hat{P} = \frac{X}{n} \tag{7.83}$$

kjer je X binomska naključna spremenljivka za število uspehov v vzorcu, n pa je velikost vzorca [Bernstein, Montgomery 1]. Če v izraz:

$$P\left(-z_{\frac{\alpha}{2}} \le Z \le z_{\frac{\alpha}{2}}\right) = 1 - \alpha \tag{7.84}$$

vstavimo izraz (7.82), dobimo:

$$P\left(-z_{\frac{\alpha}{2}} \le \frac{X - n \cdot p}{\sqrt{n \cdot p \cdot (1-p)}} \le z_{\frac{\alpha}{2}}\right) = 1 - \alpha \tag{7.85}$$

oziroma:

$$P\left(-z_{\frac{\alpha}{2}} \le \frac{\frac{X}{n}-p}{\sqrt{\frac{p\cdot(1-p)}{n}}} \le z_{\frac{\alpha}{2}}\right)=1-\alpha \qquad (7.86)$$

$$P\left(-z_{\frac{\alpha}{2}} \le \frac{\hat{P}-p}{\sqrt{\frac{p\cdot(1-p)}{n}}} \le z_{\frac{\alpha}{2}}\right)=1-\alpha$$

Po daljši izpeljavi in rešitvi neenačb dobimo naslednji izraz [Jesenko, Turk]:

$$P\left(\begin{array}{l}\frac{n\cdot\hat{P}+\frac{1}{2}\cdot z_{\frac{\alpha}{2}}^{\,2}}{n+z_{\frac{\alpha}{2}}^{\,2}}-\frac{z_{\frac{\alpha}{2}}}{n+z_{\frac{\alpha}{2}}^{\,2}}\sqrt{\frac{n\cdot\hat{P}\left(n-n\cdot\hat{P}\right)}{n}+\frac{1}{4}\cdot z_{\frac{\alpha}{2}}^{\,2}} \le p \le \\ \le \frac{n\cdot\hat{P}+\frac{1}{2}\cdot z_{\frac{\alpha}{2}}^{\,2}}{n+z_{\frac{\alpha}{2}}^{\,2}}+\frac{z_{\frac{\alpha}{2}}}{n+z_{\frac{\alpha}{2}}^{\,2}}\sqrt{\frac{n\cdot\hat{P}\left(n-n\cdot\hat{P}\right)}{n}+\frac{1}{4}z_{\frac{\alpha}{2}}^{\,2}}\end{array}\right)=1-\alpha \qquad (7.87)$$

Preoblikujmo še nekoliko spodnjo mejo:

$$\frac{n\cdot\hat{P}+\frac{1}{2}\cdot z_{\frac{\alpha}{2}}^{\,2}}{n+z_{\frac{\alpha}{2}}^{\,2}}-\frac{z_{\frac{\alpha}{2}}}{n+z_{\frac{\alpha}{2}}^{\,2}}\sqrt{\frac{n\cdot\hat{P}\left(n-n\cdot\hat{P}\right)}{n}+\frac{1}{4}\cdot z_{\frac{\alpha}{2}}^{\,2}}= \tag{7.88}$$

$$=\frac{\hat{P}+\frac{1}{2n}\cdot z_{\frac{\alpha}{2}}^{\,2}}{1+\frac{z_{\frac{\alpha}{2}}^{\,2}}{n}}-\frac{z_{\frac{\alpha}{2}}}{1+\frac{z_{\frac{\alpha}{2}}^{\,2}}{n}}\sqrt{\frac{\hat{P}\left(n-n\cdot\hat{P}\right)}{n^2}+\frac{1}{4\cdot n^2}\cdot z_{\frac{\alpha}{2}}^{\,2}}=$$

$$=\frac{1}{1+\frac{z_{\frac{\alpha}{2}}^{\,2}}{n}}\left(\hat{P}+\frac{1}{2n}\cdot z_{\frac{\alpha}{2}}^{\,2}-z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{P}\left(1-\hat{P}\right)}{n}+\frac{1}{4\cdot n^2}\cdot z_{\frac{\alpha}{2}}^{\,2}}\right)$$

Tako dobimo:

$$P\left(\begin{array}{l}\frac{1}{1+\frac{z_{\frac{\alpha}{2}}^{\,2}}{n}}\left(\hat{P}+\frac{1}{2n}\cdot z_{\frac{\alpha}{2}}^{\,2}-z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{P}\left(1-\hat{P}\right)}{n}+\frac{1}{4\cdot n^2}\cdot z_{\frac{\alpha}{2}}^{\,2}}\right)\le p\le \\ \le\frac{1}{1+\frac{z_{\frac{\alpha}{2}}^{\,2}}{n}}\left(\hat{P}+\frac{1}{2n}\cdot z_{\frac{\alpha}{2}}^{\,2}+z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{P}\left(1-\hat{P}\right)}{n}+\frac{1}{4\cdot n^2}\cdot z_{\frac{\alpha}{2}}^{\,2}}\right)\end{array}\right)=1-\alpha \tag{7.89}$$

oziroma:

$$P\left(\begin{array}{l}\frac{\hat{P}+\frac{1}{2n}\cdot z_{\frac{\alpha}{2}}^{\,2}-z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{P}\left(1-\hat{P}\right)}{n}+\frac{1}{4\cdot n^2}\cdot z_{\frac{\alpha}{2}}^{\,2}}}{1+\frac{z_{\frac{\alpha}{2}}^{\,2}}{n}}\le p\le \\ \le\frac{\hat{P}+\frac{1}{2n}\cdot z_{\frac{\alpha}{2}}^{\,2}+z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{P}\left(1-\hat{P}\right)}{n}+\frac{1}{4\cdot n^2}\cdot z_{\frac{\alpha}{2}}^{\,2}}}{1+\frac{z_{\frac{\alpha}{2}}^{\,2}}{n}}\end{array}\right)=1-\alpha \tag{7.90}$$

Primer 7.17.:

V daljšem obdobju opazujemo košarkarja, ki meče na koš. V 695 metih je zadel 413 krat. določite 95% interval zaupanja za verjetnost, da bo pri metu na koš košarkar koš zadel.

Imamo:

$$\hat{P} = \frac{413}{695} = 0.5942$$
$$z_{\frac{\alpha}{2}} = z_{\frac{0.05}{2}} = 1.96$$
$$n = 695$$

Sledi:

$$P\left(\frac{0.5942 + \frac{1}{2 \cdot 695} \cdot 1.96^2 - 1.96 \cdot \sqrt{\frac{0.5942(1-0.5942)}{695} + \frac{1}{4 \cdot 695^2} \cdot 1.96^2}}{1 + \frac{1.96^2}{695}} \le p \le \right.$$
$$\left. \le \frac{0.5942 + \frac{1}{2 \cdot 695} \cdot 1.96^2 + 1.96 \cdot \sqrt{\frac{0.5942(1-0.5942)}{695} + \frac{1}{4 \cdot 695^2} \cdot 1.96^2}}{1 + \frac{1.96^2}{695}} \right) = 1 - 0. \quad (7.91)$$

Po izračunu dobimo naslednji rezultat:

$$p \in [0.5573, 0.6301] \quad (7.92)$$

Torej s 95% verjetnostjo trdimo, da je verjetnost, da košarkar zadane koš, na intervalu med 0.5573 in 0.6301.

Pri izračunu smo si pomagali z naslednjim programom v Matlabu:

```
% delez_oc.m

P = 0.5942
z = 1.96
n = 695

koren = sqrt(P*(1-P)/n + z^2/4/n^2);
st1 = P + z^2/2/n - z*koren;
st2 = P + z^2/2/n + z*koren;
im= 1 + z^2/n;

SP = st1/im
ZG = st2/im
```

Izračune lahko poenostavimo, če bi ob predpostavki, da je *n* dovolj velik, v izrazu (7.86) nadomestili verjetnost *p* pod korenom kar z njeno oceno $\hat{P}$ [Turk, Jesenko, Montgomery 1]:

$$P\left(-z_{\frac{\alpha}{2}} \le \frac{\hat{P}-p}{\sqrt{\frac{\hat{P}\cdot\left(1-\hat{P}\right)}{n}}} \le z_{\frac{\alpha}{2}}\right)=1-\alpha \tag{7.93}$$

sledi :

$$P\left(\hat{P}-z_{\frac{\alpha}{2}}\cdot\sqrt{\frac{\hat{P}\cdot\left(1-\hat{P}\right)}{n}} \le p \le \hat{P}+z_{\frac{\alpha}{2}}\cdot\sqrt{\frac{\hat{P}\cdot\left(1-\hat{P}\right)}{n}}\right)=1-\alpha$$

Do tega rezultata bi lahko prišli tudi iz izraza (7.90), če bi določene člene zaradi velikega *n* zanemarili. Vrnimo se k primeru s košarkarjem in izračunajmo interval zaupanja:

$$P\left(0.5942-1.96\cdot\sqrt{\frac{0.5942\cdot\left(1-0.5942\right)}{695}} \le p \le 0.5942+1.96\cdot\sqrt{\frac{0.5942\cdot\left(1-0.5942\right)}{695}}\right) = \tag{7.94}$$

$$=1-0.05$$

$$P\left(0.5942-0.0365 \le p \le 0.5942+0.0365\right)=0.95$$

Dobimo:

$$p\in\left[0.5577,0.6307\right] \tag{7.95}$$

Vidimo, da se ta rezultat ne razlikuje bistveno od prejšnjega.

Ocenjevanje razlike deležev

Včasih nas zanima ocena razlike parametrov p_1 in p_2 dveh binomskih naključnih spremenljivk. Oceno poskušamo določiti na osnovi vzorcev velikosti n_1 in n_2. Takšen

primer je npr. razlika med deležema volilcev in volilk, ki bodo volili določeno stranko na volitvah [Jesenko].

Naj bosta X_1 in X_2 binomski naključni spremenljivki, katerih realizaciji predstavljata število nastopov nekega dogodka v n_1 oz. n_2. Označimo z $\hat{P}_1 = \frac{X_1}{n_1}$ in $\hat{P}_2 = \frac{X_2}{n_2}$ oceni, dobljeni na osnovi vzorcev. Razlika $\hat{P}_1 - \hat{P}_2$ je cenilka, iz katere izračunamo oceno za razliko verjetnosti $p_1 - p_2$. Velja naslednje:

$$
\begin{aligned}
&\mu_1 = E(X_1) = n_1 \cdot p_1 \\
&\sigma_1^2 = VAR(X_1) = n_1 \cdot p_1 \cdot q_1 \\
&\mu_2 = E(X_2) = n_2 \cdot p_2 \\
&\sigma_2^2 = VAR(X_2) = n_2 \cdot p_2 \cdot q_2 \\
&E(X_1 - X_2) = \mu_1 - \mu_2 = n_1 \cdot p_1 - n_2 \cdot p_2 \\
&E(\hat{P}_1 - \hat{P}_2) = E\left(\frac{X_1}{n_1} - \frac{X_2}{n_2}\right) = \frac{n_1 \cdot p_1}{n_1} - \frac{n_2 \cdot p_2}{n_2} = p_1 - p_2 \\
&VAR(X_1 - X_2) = VAR(X_1) + VAR(X_2) = n_1 \cdot p_1 \cdot q_1 + n_2 \cdot p_2 \cdot q_2 \\
&VAR(\hat{P}_1 - \hat{P}_2) = VAR\left(\frac{X_1}{n_1} - \frac{X_2}{n_2}\right) = \frac{1}{n_1^2} VAR(X_1) + \frac{1}{n_2^2} VAR(X_2) = \\
&= \frac{1}{n_1^2} n_1 \cdot p_1 \cdot q_1 + \frac{1}{n_2^2} n_2 \cdot p_2 \cdot q_2 = \frac{p_1 \cdot q_1}{n_1} + \frac{p_2 \cdot q_2}{n_2} = \frac{p_1 \cdot (1 - p_1)}{n_1} + \frac{p_2 \cdot (1 - p_2)}{n_2}
\end{aligned}
\tag{7.96}
$$

Za velike vzorce lahko binomski naključni spremenljivki in njuno razliko nadomestimo z normalno naključno spremenljivko, tako da je [Jesenko]:

$$
Z = \frac{(\hat{P}_1 - \hat{P}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1 \cdot (1 - p_1)}{n_1} + \frac{p_2 \cdot (1 - p_2)}{n_2}}} \tag{7.97}
$$

standardizirana naključna spremenljivka.

Potem lahko zapišemo:

$$P\left(-z_{\frac{\alpha}{2}} \le \frac{\left(\hat{P}_1-\hat{P}_2\right)-\left(p_1-p_2\right)}{\sqrt{\frac{p_1\cdot\left(1-p_1\right)}{n_1}+\frac{p_2\cdot\left(1-p_2\right)}{n_2}}} \le z_{\frac{\alpha}{2}}\right)=1-\alpha \qquad (7.98)$$

sledi :

$$P\left(\begin{array}{l}\left(\hat{P}_1-\hat{P}_2\right)-z_{\frac{\alpha}{2}}\cdot\sqrt{\frac{p_1\cdot\left(1-p_1\right)}{n_1}+\frac{p_2\cdot\left(1-p_2\right)}{n_2}} \le p_1-p_2 \le \\ \le\left(\hat{P}_1-\hat{P}_2\right)+z_{\frac{\alpha}{2}}\cdot\sqrt{\frac{p_1\cdot\left(1-p_1\right)}{n_1}+\frac{p_2\cdot\left(1-p_2\right)}{n_2}}\end{array}\right)=1-\alpha$$

Podobno kot prej lahko izračune poenostavimo, če bi ob predpostavki, da je *n* dovolj velik, v izrazu (7.98) nadomestili obe verjetnosti pod korenom kar z njunima ocenama [Turk, Jesenko, Montgomery 1]. Tako bi dobili:

$$P\left(\begin{array}{l}\left(\hat{P}_1-\hat{P}_2\right)-z_{\frac{\alpha}{2}}\cdot\sqrt{\frac{\hat{P}_1\cdot\left(1-\hat{P}_1\right)}{n_1}+\frac{\hat{P}_2\cdot\left(1-\hat{P}_2\right)}{n_2}} \le p_1-p_2 \le \\ \le\left(\hat{P}_1-\hat{P}_2\right)+z_{\frac{\alpha}{2}}\cdot\sqrt{\frac{\hat{P}_1\cdot\left(1-\hat{P}_1\right)}{n_1}+\frac{\hat{P}_2\cdot\left(1-\hat{P}_2\right)}{n_2}}\end{array}\right)=1-\alpha \qquad (7.99)$$

Primer 7.18.:

Z naključnim vzorcem 320 volilcev ugotovimo, da podpira določeno stranko na volitvah 252 volilcev. Z naključnim vzorcem 185 volilk pa ugotovimo, da podpira isto stranko na volitvah 102 volilk. Določite 99% interval zaupanja za razliko dejanskih deležev volilcev in volilk, ki podpirajo dotično stranko [Jesenko].

Imamo:

$$\hat{P}_1 = \frac{252}{320} = 0.7875$$
$$\hat{P}_2 = \frac{102}{185} = 0.5514$$
$$z_{\frac{\alpha}{2}} = z_{\frac{0.01}{2}} = 2.58 \quad (z\ matlabom)$$
$$n_1 = 320$$
$$n_2 = 185$$

kjer smo uporabili ukaz v Matlabu:

```
>> z=abs(norminv(0.01/2,0,1))

z =

   2.5758
```

Sledi:

$$\left(\hat{P}_1 - \hat{P}_2\right) - z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{P}_1 \cdot \left(1-\hat{P}_1\right)}{n_1} + \frac{\hat{P}_2 \cdot \left(1-\hat{P}_2\right)}{n_2}} = \tag{7.100}$$
$$= \left(0.7875 - 0.5514\right) - 2.58 \cdot \sqrt{\frac{0.7875 \cdot \left(1-0.7875\right)}{320} + \frac{0.5514 \cdot \left(1-0.5514\right)}{185}} =$$
$$= 0.2361 - 0.1113 = 0.1248$$

Podobno dobimo:

$$\left(\hat{P}_1 - \hat{P}_2\right) + z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{P}_1 \cdot \left(1-\hat{P}_1\right)}{n_1} + \frac{\hat{P}_2 \cdot \left(1-\hat{P}_2\right)}{n_2}} = 0.2361 + 0.1113 = 0.3491 \tag{7.101}$$

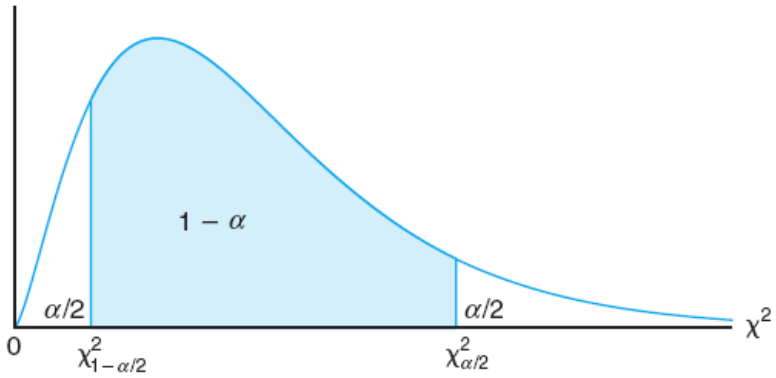
Torej je interval zaupanja enak:

$$p_1 - p_2 \in \left[0.1248, 0.3491\right] \tag{7.102}$$

7.12 Ocenjevanje variance

Dan imamo naključni vzorec velikosti n iz normalne populacije. Zanima nas $(1-\alpha)\cdot 100\%$ interval zaupanja za varianco σ^2. Iz poglavja 6.3 vemo (glej izraz (6.22)), da velja: $\frac{1}{\sigma^2}\cdot S^2\cdot(n-1)\in \chi^2(n-1)$. Torej lahko zapišemo (glej sliko 157) [Jesenko, Jurišić]:

$$
\begin{aligned}
&P\left(\chi^2_{1-\frac{\alpha}{2}}(n-1)\le\frac{1}{\sigma^2}\cdot S^2\cdot(n-1)\le\chi^2_{\frac{\alpha}{2}}(n-1)\right)=1-\alpha\\
&P\left(\chi^2_{n-1,1-\frac{\alpha}{2}}\cdot\frac{1}{S^2\cdot(n-1)}\le\frac{1}{\sigma^2}\le\chi^2_{n-1,\frac{\alpha}{2}}\cdot\frac{1}{S^2\cdot(n-1)}\right)=1-\alpha\\
&P\left(\frac{S^2\cdot(n-1)}{\chi^2_{n-1,1-\frac{\alpha}{2}}}\ge\sigma^2\ge\frac{S^2\cdot(n-1)}{\chi^2_{n-1,\frac{\alpha}{2}}}\right)=1-\alpha\\
&P\left(\frac{S^2\cdot(n-1)}{\chi^2_{n-1,\frac{\alpha}{2}}}\le\sigma^2\le\frac{S^2\cdot(n-1)}{\chi^2_{n-1,1-\frac{\alpha}{2}}}\right)=1-\alpha
\end{aligned}
\qquad (7.103)
$$



Slika 157: Ilustracija izraza $P\left(\chi^2_{1-\frac{\alpha}{2}}(n-1) \le \underbrace{\frac{1}{\sigma^2}\cdot S^2\cdot(n-1)}_{\chi^2(n-1)} \le \chi^2_{\frac{\alpha}{2}}(n-1)\right)=1-\alpha$

[Walpole]

Primer 7.19.:

Na 21 naključno izbranih izdelkih normalne populacije smo ugotovili, da je bil standardni odklon časa izdelave 5.8 min. Določite 99% interval zaupanja za σ^2. *ki pomeni resnično variabilnost proizvodnega časa za izbrani izdelek [Jesenko].*

Imamo:

$$\begin{aligned} &n = 21 \\ &S = 5.8 \\ &\alpha = 0.01 \\ &\chi^2_{n-1,\frac{\alpha}{2}} = \chi^2_{21-1,\frac{0.01}{2}} = 39.9968 \\ &\chi^2_{n-1,1-\frac{\alpha}{2}} = \chi^2_{21-1,1-\frac{0.01}{2}} = 7.4338 \end{aligned} \tag{7.104}$$

kjer smo kritični vrednosti Hi kvadrat statistike dobili z naslednjima ukazoma v Matlabu (funkcija chi2inv deluje po obratni logiki):

```
>> hi1=abs(chi2inv(1-0.01/2,21-1))
hi1 =
   39.9968
>> hi2=abs(chi2inv(0.01/2,21-1))
hi2 =
    7.4338
```

Tako dobimo:

(7.105)

$$\sigma^2 \in \left(\frac{S^2 \cdot (n-1)}{\chi^2_{n-1,\frac{\alpha}{2}}}, \frac{S^2 \cdot (n-1)}{\chi^2_{n-1,1-\frac{\alpha}{2}}} \right)$$
$$\sigma^2 \in \left(\frac{5.8^2 \cdot (21-1)}{39.9968}, \frac{5.8^2 \cdot (21-1)}{7.4338} \right)$$
$$\sigma^2 \in (16.8213, 90.5055)$$

Primer 7.20.:

V poglavju 7.9 smo imeli primer, da je spremenljivka X število ur branja dnevnih časopisov na teden, ki se porazdeljuje normalno. Izračunali smo $s^2 = 3.6667$ *(glej izraz (7.61)). Izračunajte interval zaupanja za varianco pri 10% tveganju [Jurišić].*

Imamo:

$$\begin{aligned} & n = 7 \\ & S^2 = 3.6667 \\ & \alpha = 0.1 \\ & \chi^2_{n-1,\frac{\alpha}{2}} = \chi^2_{7-1,\frac{0.1}{2}} = 12.5916 \\ & \chi^2_{n-1,1-\frac{\alpha}{2}} = \chi^2_{7-1,1-\frac{0.1}{2}} = 1.6354 \end{aligned} \qquad (7.106)$$

kjer smo kritični vrednosti Hi kvadrat statistike dobili z naslednjima ukazoma v Matlabu:

```
>> hi1=abs(chi2inv(1-0.1/2,7-1))
hi1 =
   12.5916
>> hi2=abs(chi2inv(0.1/2,7-1))
hi2 =
    1.6354
```

Tako dobimo:

$$\sigma^2 \in \left(\frac{S^2 \cdot (n-1)}{\chi^2_{n-1,\frac{\alpha}{2}}}, \frac{S^2 \cdot (n-1)}{\chi^2_{n-1,1-\frac{\alpha}{2}}} \right)$$
$$\sigma^2 \in \left(\frac{3.6667^2 \cdot (7-1)}{12.5916}, \frac{3.6667^2 \cdot (7-1)}{1.6354} \right) \qquad (7.107)$$
$$\sigma^2 \in (1.7472, 13.4525)$$

7.13 Ocenjevanje kvocienta varianc

V poglavju 6.5. (glej izraz (6.44)) smo dejali: Če sta $S_1{}^2$ in $S_2{}^2$ varianci dveh neodvisnih naključnih vzorcev velikosti n_1 in n_2 iz dveh normalnih populacij z variancama σ_1^2 in σ_2^2, potem je [Jesenko]:

$$F = \frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}} = \frac{\sigma_2^2}{\sigma_1^2} \cdot \frac{S_1^2}{S_2^2} \qquad (7.108)$$

F naključna spremenljivka z $n_1 - 1$ in $n_2 - 1$ prostostnimi stopnjami. Zapišemo lahko naslednji izraz [Jesenko]:

(7.109)

$$P\left(f_{1-\frac{\alpha}{2}}\left(n_1-1,n_2-1\right) \le \frac{\sigma_2^2}{\sigma_1^2}\cdot\frac{S_1^2}{S_2^2} \le f_{\frac{\alpha}{2}}\left(n_1-1,n_2-1\right) \right)=1-\alpha$$

$$P\left(f_{1-\frac{\alpha}{2}}\left(n_1-1,n_2-1\right)\frac{S_2^2}{S_1^2} \le \frac{\sigma_2^2}{\sigma_1^2} \le f_{\frac{\alpha}{2}}\left(n_1-1,n_2-1\right)\frac{S_2^2}{S_1^2} \right)=1-\alpha$$

$$P\left(\frac{1}{f_{1-\frac{\alpha}{2}}\left(n_1-1,n_2-1\right)\frac{S_2^2}{S_1^2}} \ge \frac{\sigma_1^2}{\sigma_2^2} \ge \frac{1}{f_{\frac{\alpha}{2}}\left(n_1-1,n_2-1\right)\frac{S_2^2}{S_1^2}} \right)=1-\alpha$$

$$P\left(\frac{\frac{S_1^2}{S_2^2}}{f_{\frac{\alpha}{2}}\left(n_1-1,n_2-1\right)} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{\frac{S_1^2}{S_2^2}}{f_{1-\frac{\alpha}{2}}\left(n_1-1,n_2-1\right)} \right)=1-\alpha$$

kjer sta $f_{1-\frac{\alpha}{2}}\left(n_1-1,n_2-1\right), f_{\frac{\alpha}{2}}\left(n_1-1,n_2-1\right)$ vrednosti *F* naključne spremenljivke.

Dokazati se da velja naslednje [Jesenko, Bernstein]:

$$\frac{1}{f_{1-\frac{\alpha}{2}}\left(n_1-1,n_2-1\right)} = f_{\frac{\alpha}{2}}\left(n_2-1,n_1-1\right) \tag{7.110}$$

torej se pri **<u>recipročnem pravilu stopnji obrneta</u>**. Tako dobimo:

$$P\left(\frac{\frac{S_1^2}{S_2^2}}{f_{\frac{\alpha}{2}}\left(n_1-1,n_2-1\right)} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{S_1^2}{S_2^2}\cdot f_{\frac{\alpha}{2}}\left(n_2-1,n_1-1\right) \right)=1-\alpha \tag{7.111}$$

<u>Primer 7.21.:</u>

Imamo podoben primer kot je bil primer, ko so primerjali količino alkohola v dveh vrstah viskija (glej izraz (7.78)). Za podatke tega primera poiščite 98% interval zaupanja za kvocient varianc [Jesenko].

Imamo:

$$
\begin{aligned}
&s_1 = 0.9 \\
&s_2 = 1.1 \\
&n_1 = 10 \\
&n_2 = 8 \\
&\alpha = 1 - 0.98 = 0.02 \\
&f_{\frac{\alpha}{2}}\left(n_1 - 1, n_2 - 1\right) = f_{\frac{0.02}{2}}\left(10 - 1, 8 - 1\right) = f_{0.01}\left(9, 7\right) = 6.7188 \\
&f_{\frac{\alpha}{2}}\left(n_2 - 1, n_1 - 1\right) = f_{\frac{0.02}{2}}\left(8 - 1, 10 - 1\right) = f_{0.01}\left(7, 9\right) = 5.6129
\end{aligned}
\tag{7.112}
$$

pri čemer smo $f_{\frac{\alpha}{2}}\left(n_1 - 1, n_2 - 1\right)$ in $f_{\frac{\alpha}{2}}\left(n_2 - 1, n_1 - 1\right)$ dobili z naslednjima ukazoma v Matlabu:

```
>> f1 = finv(1-0.02/2,10-1,8-1)
f1 =
   6.7188

>> f2 = finv(1-0.02/2,8-1,10-1)
f2 =
   5.6129
```

Sledi:

$$
\begin{aligned}
&P\left(\frac{\frac{0.9^2}{1.1^2}}{6.7188} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{0.9^2}{1.1^2} \cdot 5.6129 \right) = 1 - 0.02 \\
&P\left(0.0996 \le \frac{\sigma_1^2}{\sigma_2^2} \le 3.7574 \right) = 0.98 \\
&\frac{\sigma_1^2}{\sigma_2^2} \in \left[0.0996, 3.7574 \right]
\end{aligned}
\tag{7.113}
$$

8 PREVERJANJE HIPOTEZ

8.1 Uvod

Pogosto želimo preverjati predpostavke o vrednostih določenih parametrov populacije na osnovi podatkov, dobljenih na vzorcu. Npr., tehnolog želi v tovarni avtomobilskih gum vedeti na osnovi podatkov vzorca, če določena vrsta gum zdrži vsaj 40000 km. Ali, vinogradnik želi na osnovi preizkusa ugotoviti, če določena vrsta gnojila vpliva na večji donos grozdja. Ali, proizvajalec pralnih praškov želi izvedeti na osnovi vzorca, če je z uporabo novega praška možno odstraniti 90% vseh madežev. Ali, sociolog bi se želel na osnovi vzorca odločiti, če sta krvna skupina in barva oči neodvisni spremenljivki. Obstaja še veliko podobnih problemov, vsi pa spadajo v **področje statističnega testiranja hipotez** [Jesenko].

V vseh navedenih primerih gre za odločitve o tem, ali imajo parametri populacije neko določeno vrednost. Včasih se te odločitve nanašajo tudi na tip ali naravo populacije. Npr. tehnolog v tovarni avtomobilskih gum bi se lahko spraševal tudi to, če je naključna spremenljivka število prevoženih km enaka Weibullovi naključni spremenljivki [Jesenko].

Metodologiji **ocenjevanja parametrov** in **testiranja hipotez** obe uporabljata empirične vzorce za sklepanje o neznanih značilnostih populacije, vendar med njima obstaja **opazna razlika**. Pri problemu ocenjevanja parametrov je informacija vzorca uporabljena za to, da se oceni vrednost nekega parametra θ, pri čemer se njegovo oceno umesti v nek interval zaupanja. Pri problemu testiranja hipotez pa se vzpostavi statistična hipoteza, ki daje tekmovalne (alternativne) predpostavke glede parametra θ, in je informacija vzorca uporabljena za **odločanje** z določeno stopnjo zaupanja, katera od tekmovalnih predpostavk naj bo sprejeta kot pravilna [Bernstein].

Z drugimi besedami, pri testiranju hipotez gre za odločanje, ali sprejmemo ali zavrnemo izjavo (predpostavko) o določenem parametru. Ta predpostavka se imenuje **hipoteza**, **odločitvena procedura** glede hipotez pa se imenuje **testiranje hipotez** [Montgomery 1].

Resnica o veljavnosti statistične hipoteze ni nikoli znana z absolutno gotovostjo, razen če ne bi raziskali cele populacije. Ker je to nepraktično, namesto tega vzamemo vzorec in se na njegovih podatkih odločimo, če drži določena hipoteza o populaciji. Če ta hipoteza na vzorcu ne drži, potem jo zavrnemo [Walpole].

Poglejmo si nekaj **definicij**:

- **Raziskovalna domneva** (hipoteza) je še nedokazana trditev, ki jo želimo potrditi ali zavrniti z raziskovalnim delom [Košmelj K.].

- **Statistična domneva** je še nedokazana trditev o lastnosti naključne spremenljivke. Opredelimo dve statistični domnevi: **ničelno domnevo** H_0 in **alternativno domnevo** H_1 [Košmelj K.].

- **Statistična domneva** je vsaka domneva o neznani porazdelitvi vrednosti slučajne spremenljivke *X*. Ta je lahko **parametrična domneva**, to je domneva o vrednosti nekega parametra porazdelitve, in **neparametrična domneva**, to je domneva o neki neparametrični lastnosti (tip porazdelitve, neodvisnost ...) porazdelitve slučajne

spremenljivke *X*. Če domneva natančno določa predpostavljeno porazdelitev, pravimo, da je to **enostavna domneva**, v nasprotnem primeru, ko so nekateri parametri porazdelitve nedoločeni, pravimo, da je **domneva sestavljena** [Turk].

- **Statistična domneva** (ali hipoteza) je vsaka domneva o porazdelitvi slučajne spremenljivke *X* na populaciji. Če poznamo vrsto (obliko) porazdelitve in postavljamo/raziskujemo domnevo o parametru, govorimo o **parametrični domnevi**. Če pa je vprašljiva tudi sama vrsta porazdelitve, je **domneva neparametrična**. Domneva je **enostavna**, če natančno določa porazdelitev (njeno vrsto in točno vrednost parametra), sicer je sestavljena [Jurišić].

- **Statistična domneva** je trditev ali predpostavka o porazdelitvi ene ali več naključnih spremenljivk. Če v celoti določa porazdelitev, je hipoteza **enostavna**, sicer je pa **sestavljena**. Enostavna hipoteza mora potemtakem določati ne le obliko porazdelitve naključne spremenljivke, pač pa tudi vrednosti vseh parametrov na njej. Če npr. predpostavimo, da je vzorec izbran iz binomske populacije z vrednostjo parametra *p = 90%,* gre za enostavno domnevo. Če pa nas npr. zanima, če je parameter $\theta \geq 40000\ km$, gre za sestavljeno domnevo, saj parameter θ nima neke določene vrednosti [Jesenko].

Statistična domneva je lahko pravilna ali napačna. Želimo seveda sprejeti pravilno domnevo in zavrniti napačno. Težava je v tem, da o pravilnosti/napačnosti domneve ne moremo biti gotovi, če jo ne preverimo na celotni populaciji. Ponavadi se odločamo le na podlagi vzorca. Če vzorčni podatki preveč odstopajo od domneve, rečemo, da niso skladni z domnevo, oziroma, da so razlike značilne, in domnevo zavrnemo. Če pa podatki domnevo podpirajo, jo ne zavrnemo - včasih jo celo sprejmemo. To ne pomeni, da je domneva pravilna, temveč, da ni zadostnega razloga za zavrnitev [Jurišić].

Primer:

Naj bo $X \in N(\mu,\sigma)$. Če poznamo σ, je domneva $H:\mu = 0$ enostavna. Če pa parametra σ ne poznamo, je sestavljena. Primer sestavljene hipoteze je tudi $H:\mu > 0$.

Pojem ničelne in nasprotne (alternativne) hipoteze

Postavitev hipoteze, **da "ni razlike"**, nas pripelje do pojma **ničelne hipoteze** H_0. Gre za trditev o lastnosti populacije, za katero predpostavljamo (verjamemo), da drži. Hkrati je to tudi trditev, ki jo test poskuša ovreči [Jurišić]. V matematičnem smislu ima vedno enačaj.

Alternativne hipoteze $H_i,\ \ i=1,...,k$ so trditve, ki so nasprotne ničelni hipotezi in jih poskušamo s testiranjem dokazati [Jurišić].

Primeri hipotez [Turk]

- Parametrični domnevi dvostranskega testa:
  $H_0 : \mu = 0$
  $H_1 : \mu \neq 0$
- Parametrični domnevi enostranskega testa:
  $H_0 : \sigma = 16$
  $H_1 : \sigma > 16$
- Neparametrični domnevi:
  H_0 : *Porazdelitev je norma*lna
  H_1 : *Porazdelitev ni norma*lna

Testiranje hipotez

Testiranje hipotez je uporaba natanko določenih postopkov za odločitev o tem, ali sprejeti ničelno hipotezo ali pa jo zavrniti v korist nasprotne hipoteze. Vzemimo primer:

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta = \theta_1$$

Najprej moramo opraviti preizkus, da pridobimo podatke naključnega vzorca in iz njih izračunamo vrednost izbrane statistike (**statistika testa hipoteze** Θ), ter se nato glede na njeno vrednost odločimo o sprejetju ali zavrnitvi ničelne hipoteze. Postopek testiranja je

torej zgrajen na delitvi vseh možnih vrednosti statistike testa hipoteze na dve območji [Jesenko]:

- Območje sprejemanja ničelne hipoteze H_0,
- Območje zavračanja ničelne hipoteze H_0.

Postopek testiranja hipotez poteka v naslednjih fazah [Turk]:

1. Postavimo ničelno in alternativno hipotezo,

2. Izberemo statistiko, ki ustreza ničelni domnevi in določimo njeno porazdelitev,

3. Izberemo stopnjo tveganja α. Na osnovi stopnje tveganja in porazdelitve statistike določimo kritično območje.

4. Na vzorčnih podatkih izračunamo vrednost statistike.

5. Izvedemo sklep:

- **Če vrednost statistike pade v kritično območje (območje zavrnitve ničelne domneve), ničelno domnevo zavrnemo in sprejmemo alternativno domnevo ob stopnji tveganja α.**
- **Če vrednost statistike ne pade v kritično območje, ničelne hipoteze ne moremo zavrniti ob stopnji tveganja α.**

Možne napake pri testiranju hipotez

Opisani postopek pa lahko pripelje do dveh vrst napak [Jesenko]:

- Prava vrednost je: $\theta = \theta_0$, mi pa na osnovi testa sklepamo, da je $\theta = \theta_1$ (napaka 1. vrste)
- Prava vrednost je: $\theta = \theta_1$, mi pa na osnovi testa sklepamo, da je $\theta = \theta_0$ (napaka 2. vrste)

Primer:

Imamo naslednjo ničelno in nasprotno hipotezo [Nemec]:

$$H_0 : M_H = M$$
$$H_1 : M_H \neq M$$

Pomen obeh napak pri testiranju je prikazan na sliki 158.

	Rezultat statističnega preizkusa	
Dejansko:	**Pravilna hipoteza H_0**	**Pravilna hipoteza H_1**
$M_H = M$	Pravilen zaključek	Napaka 1. vrste
$M_H \neq M$	Napaka 2. vrste	Pravilen zaključek

Slika 158: Pomen obeh napak pri testiranju [Nemec]

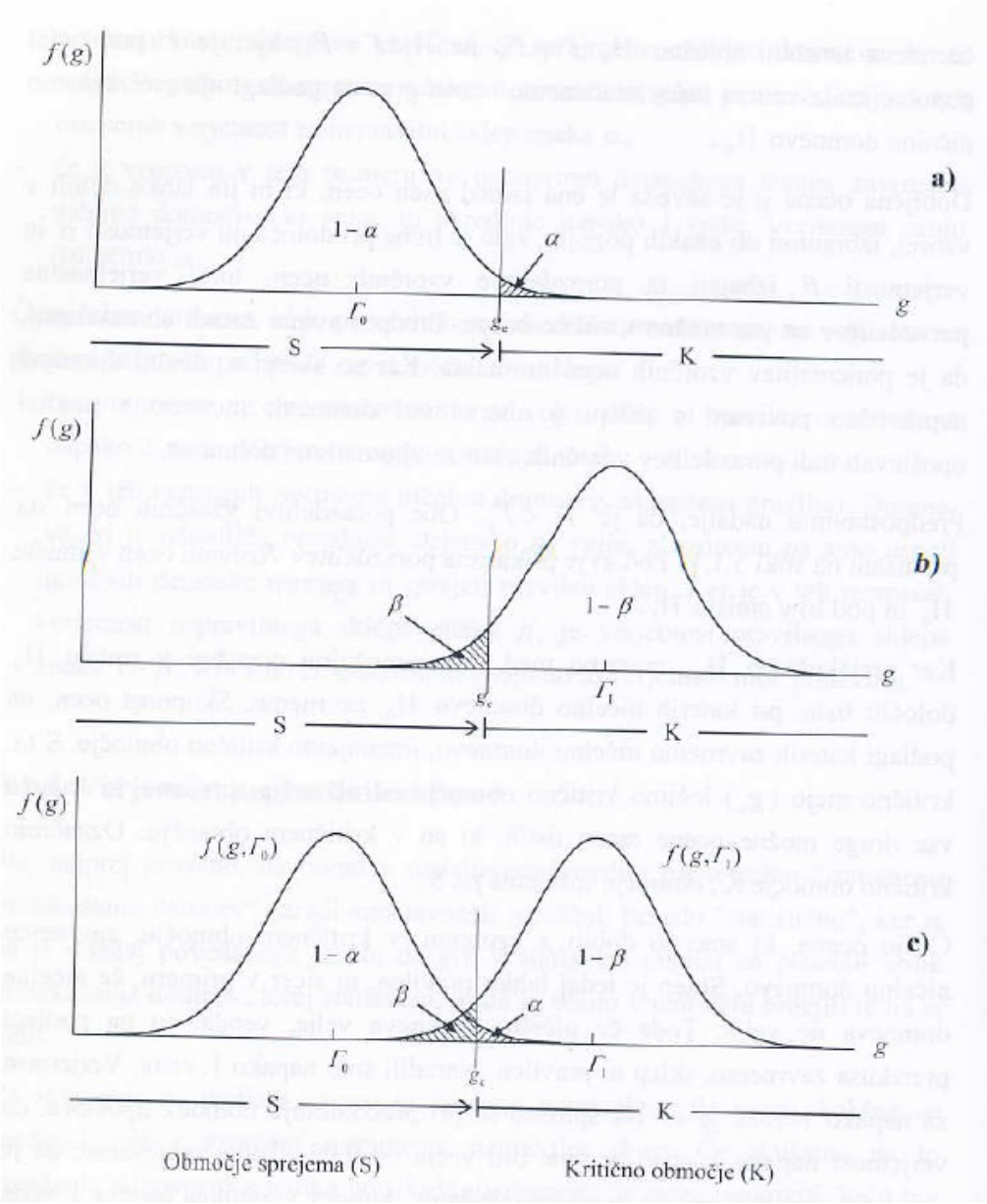
Verjetnosti napak testiranja

- Napaka 1. vrste je zavrnitev ničelne hipoteze, ko je ta pravilna. Verjetnost, da naredimo napako 1. vrste, označimo s simbolom α in ji pravimo stopnja tveganja, $(1-\alpha)$ pa je stopnja zaupanja [Jurišić, Jesenko].

- Napaka 2. vrste je sprejetje ničelne hipoteze, ko je ta napačna. Verjetnost, da naredimo napako 2. vrste, označimo s simbolom β [Jurišić, Jesenko] (glej sliki 159 in 159a [Košmelj B.]).

Sklep na podlagi vzorca	**Dejansko stanje**	
	H_0 velja	**H_0 ne velja**
H_0 velja	Sklep je pravilen (verjetnost sklepa je 1-α)	Sklep ni pravilen – napaka 2. vrste (verjetnost sklepa je β)
H_0 ne velja	Sklep ni pravilen – napaka 1. vrste (verjetnost sklepa je α)	Sklep je pravilen (verjetnost sklepa je 1-β, imenovana moč preizkusa)

Slika 159: Pravilni in nepravilni sklepi pri preizkušanju domnev [Košmelj B.]



Slika 159a: Porazdelitvi vzorčnih ocen v smislu ničelne in alternativne domneve ter verjetnost napake 1. in 2. vrste za nesestavljeno domnevo: $\begin{matrix} H_0 : \Gamma = \Gamma_0 \\ H_1 : \Gamma = \Gamma_1 \end{matrix}$ *(*g_c *je kritična meja) [Košmelj B.]*

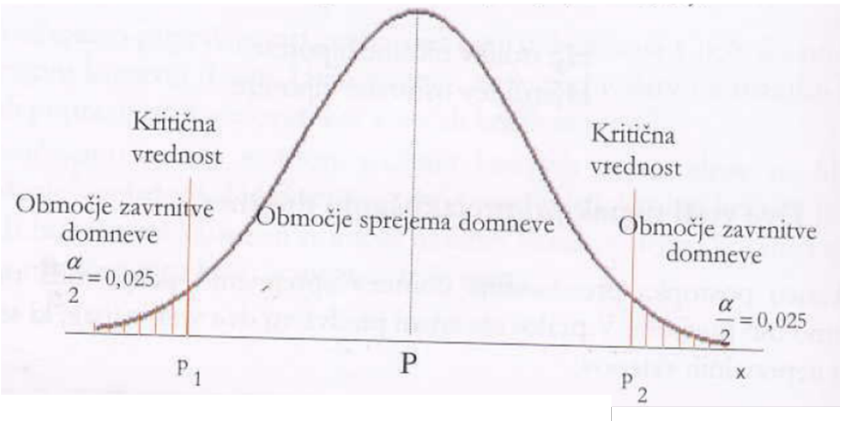
<u>Območje sprejemanja in zavračanja domnev</u>

Običajno imenujemo **območje zavračanja ničelne hipoteze kritično območje testa**, verjetnost α, da vrednost statistike testa hipoteze pade v kritično območje, pa imenujemo stopnja pomembnosti (značilnosti) testa [Jesenko].

Postopki testov, s katerimi preizkušamo domneve, so takšni, da izključujejo možnost, da bi storili napako 2. vrste. Dopuščajo pa zavrnitev ničelne hipoteze s stopnjo tveganja α. Vsi testi temeljijo na primerjavi med dejansko izračunanimi vrednostmi in kritičnimi vrednostmi, ki razmejujejo območje sprejema ali zavrnitve ničelne hipoteze pri določeni stopnji tveganja. **Kritična vrednost je torej vrednost statistike, ki ustreza dani stopnji značilnosti in je določena z njeno porazdelitvijo [Brvar]**.

Tako se srečamo s pojmom **takoimenovane *P*-vrednosti**, ki predstavlja (bistveno) stopnjo značilnosti (signifikantnosti) testa in je največja vrednost parametra α, ki smo jo pripravljeni sprejeti glede na dan vzorec (zgornja meja za napako 1. vrste) [Jurišić].

Razlike med dejansko izračunanimi in kritičnimi vrednostmi so značilne ali pa niso značilne. Značilnost (stopnjo pomembnosti testa) izrazimo z najmanjšo stopnjo tveganja α, s katero še lahko zavrnemo ničelno hipotezo. **Običajno je stopnja tveganja** $\alpha = 0.05$ **najvišja stopnja, za katero še pravimo, da je razlika značilna** (glej sliko 160) [Brvar].



Slika 160: Območje sprejemanja in zavračanja domnev [Brvar]

Vrste testov

Ničelna hipoteza vedno navaja, da ima proučevani parameter določeno vrednost. Alternativna
hipoteza pa trdi, da je vrednost tega parametra večja, manjša ali preprosto ni enaka tej vrednosti, ki je zapisana v ničelni hipotezi. **Tako poznamo eno in dvorepe statistične teste** (angleško one-tailed or two-tailed test). **Enorepi test** uporabimo v primeru, če alternativna hipoteza trdi, da je vrednost večja ali manjša od vrednosti, zapisane v ničelni hipotezi. Kadar pa v alternativni hipotezi postavimo trditev, da je vrednost parametra preprosto različna (večja ali manjša) od vrednosti, postavljene v ničelni hipotezi, uporabimo **dvorepi test** [Tominc, Kramberger].

Primer 8.1.:

Podjetje za proizvodnjo pralnih praškov je želelo preizkusiti nov pralni prašek. Postavilo je ničelno hipotezo:

$$H_0 : p = 85\% \tag{8.1}$$

ki pomeni, da novi pralni prašek očisti 85% vseh madežev, ter nasprotno hipotezo:

$$H_1 : p = 60\% \tag{8.2}$$

ki pomeni procent očiščenih madežev s prejšnjimi praški. Statistika testa hipoteze naj bo T in predstavlja število očiščenih madežev v 25 preizkusih, ki so jih opravili. Očitno je ta statistika binomska naključna spremenljivka. Ničelno hipotezo bomo sprejeli, če je bilo število število očiščenih madežev večje od 17 ($T = x > 17$), sicer jo bomo zavrnili. Poiščite napaki α, β [Jesenko].

Območje sprejemanja ničelne hipoteze je:

$$x = 18, 19, 20, ..., 25 \tag{8.3}$$

Kritično območje pa je:

$$x = 0,1,2,...,17 \tag{8.4}$$

Verjetnost α za nastop napake 1. vrste je verjetnost, da ničelno hipotezo zavržemo (očiščenih madežev v vzorcu največ 17), kljub temu, da je pravilna (novi pralni prašek je učinkovit).

Izračunamo jo tako:

$$\alpha = P(T \leq 17, p = 0.85) = 0.0255 \tag{8.5}$$

pri čemer smo si pomagali z Matlabom:

```
>> b=binocdf(0:17,25,0.85);
>> b=b(length(b))
b =
   0.0255
```

Verjetnost β za nastop napake 2. vrste je verjetnost, da ničelno hipotezo sprejmemo (očiščenih madežev v vzorcu več kot 17), kljub temu, da ni pravilna (novi pralni prašek ni učinkovit). Izračunamo jo tako:

$$\beta = P(T > 17, p = 0.60) = 0.1536 \tag{8.6}$$

pri čemer smo si pomagali z Matlabom:

```
>> b=binocdf(0:17,25,0.60);
>> b=1-b(length(b))
b =
   0.1536
```

Dober test je takšen, pri katerem sta obe napaki kar se da majhni, kar daje dobre možnosti za pravilno odločitev. Kot vidimo iz primera, je napaka druge vrste precej velika, vendar jo lahko zmanjšamo, če spremenimo kritično območje [Jesenko]. Če bi npr. vzeli območje sprejemanja ničelne hipoteze (OSNH) in kritično območje (KO):

$$KO = (T \le 18) \qquad (8.7)$$
$$OSNH = (T > 18)$$

bi dobili:

$$\alpha = P(T \le 18, p = 0.85) = 0.0695 \qquad (8.8)$$

in:

$$\beta = P(T > 18, p = 0.60) = 0.0736 \qquad (8.9)$$

kjer smo si spet pomagali z Matlabom:

```
>> b=binocdf(0:18,25,0.85);
>> b=b(length(b))
b =
   0.0695

>> b=binocdf(0:18,25,0.60);
>> b=1-b(length(b))
b =
   0.0736
```

Kot vidimo, smo zmanjšali napako β, vendar smo povečali napako α. Edini način, da bi občutno zmanjšali obe napaki, je povečanje velikosti vzorca [Jesenko].

V praksi imamo redko opravka z enostavnimi hipotezami. Prejšnji primer s praški bi lahko postavili tudi tako:

$$H_0 : p \ge 0.85 \qquad (8.10)$$
$$H_1 : p < 0.85$$

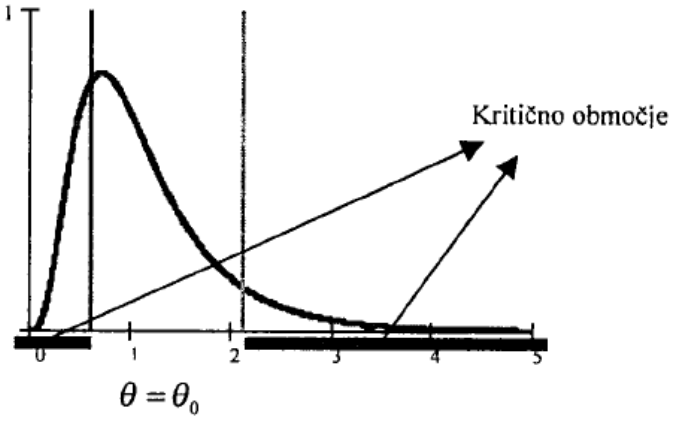
Določitev kritičnega območja je v tem primeru bolj zapletena naloga. Namreč, pri enostavnih hipotezah sta α, β konstanti, pri sestavljenih pa postaneta funkcija parametra θ:

$$\alpha = \alpha(\theta) \qquad (8.11)$$
$$\beta = \beta(\theta)$$

Denimo, da želimo testirati naslednji hipotezi:

$$H_0 : \theta = \theta_0 \qquad (8.12)$$
$$H_1 : \theta \neq \theta_0$$

V tem primeru je smiselno, da sprejmemo ničelno hipotezo, če je točkasta ocena $\hat{\theta}$ parametra θ blizu θ_0, ter jo zavrnemo, če je veliko večja ali veliko manjša od θ_0. Tako je razumljivo, da naj bo v tem primeru kritično območje sestavljeno iz obeh delov porazdelitve gostote verjetnosti statistike testa Θ. Takšen test imenujemo **dvostranski test**, kot smo že omenili (slika 161) [Jesenko].



Slika 161: Dvostranski test hipoteze [Jesenko]

Sedaj pa vzemimo, da želimo testirati naslednji hipotezi:

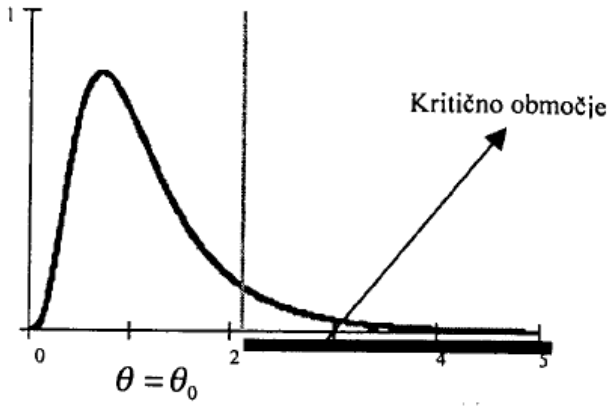
$$H_0 : \theta = \theta_0 \qquad (8.13)$$
$$H_1 : \theta < \theta_0$$

V tem primeru je smiselno, da sprejmemo ničelno hipotezo, če je točkasta ocena $\hat{\theta}$ parametra θ blizu θ_0, ter jo zavrnemo, če je veliko manjša od θ_0. V tem primeru je razumljivo, da naj bo kritično območje sestavljeno le iz spodnjega dela porazdelitve gostote verjetnosti statistike testa Θ. Takšen test imenujemo **enostranski test (od spodaj)** [Jesenko].

Sedaj pa vzemimo, da želimo testirati naslednji hipotezi:

$$H_0 : \theta = \theta_0 \qquad (8.14)$$
$$H_1 : \theta > \theta_0$$

V tem primeru je smiselno, da sprejmemo ničelno hipotezo, če je točkasta ocena $\hat{\theta}$ parametra θ blizu θ_0, ter jo zavrnemo, če je veliko večja od θ_0. V tem primeru je razumljivo, da naj bo kritično območje sestavljeno le iz zgornjega dela porazdelitve gostote verjetnosti statistike testa Θ. Takšen test imenujemo **enostranski test (od zgoraj)** (slika 162) [Jesenko].



Slika 162: Enostranski test hipoteze (od zgoraj) [Jesenko]

Natančnejša opredelitev postopka preizkušanja domnev

Postopek preizkušanja domnev smo v grobem že omenili. Po Ferligoju [Ferligoj] je izveden v naslednjih 5 korakih [Brvar]:

1. Na podlagi opredelitve statističnega problema postavimo ničelno in alternativno domnevo.

2. Izberemo čimbolj nepristransko cenilko za parameter, ki ga ocenjujemo, in porazdelitev, v kateri ta cenilka nastopa.

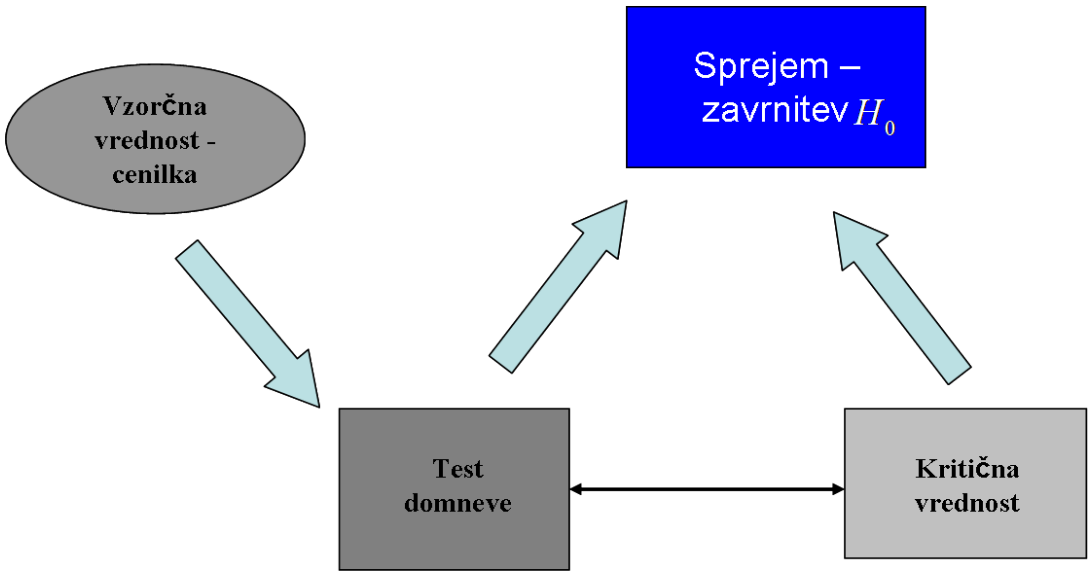
3. Izberemo stopnjo značilnosti α. Na njeni podlagi in podlagi porazdelitve statistike določimo kritično območje.

4. Na vzorčnih podatkih izračunamo vrednost statistike v kritičnem območju porazdelitve ali zunaj nje.

5. Podamo sklep:

- Če izračunana vrednost statistike pade v kritično območje, ničelno domnevo zavrnemo in sprejmemo alternativno ob izbranem α.
- Če pa pade izven kritičnega območja, ničelne domneve ne moremo zavreči in sprejmemo sklep, da vzorčni podatki kažejo na neznačilne razlike med populacijskim parametrom in vzorčno oceno.

Postopek sklepanja končne ocene prikazuje slika 163 [Brvar].



Slika 163: Postopek sklepanja končne ocene [Brvar]

Računalniški programi običajno za vsak rezultat (vrednost statistike) izračunajo najmanjšo verjetnost, pri kateri ničelno hipotezo še lahko zavrnemo. Ta verjetnost se imenuje, kot smo že omenili, *P*-vrednost. Z vpeljavo te vrednosti pa se nekoliko spremenijo tudi koraki v postopku testiranja domnev [Brvar]:

1. Na podlagi opredelitve statističnega problema postavimo ničelno in alternativno domnevo.

2. Izberemo čimbolj nepristransko cenilko za parameter, ki ga ocenjujemo, in porazdelitev, v kateri ta cenilka nastopa.

3. Izberemo stopnjo značilnosti α.

4. Na vzorčnih podatkih izračunamo vrednost statistike in pripadajočo *P*-vrednost.

5. Podamo sklep, to je, preverimo, ali je P-vrednost manjša ali enaka α, in glede na to domnevo zavrnemo, ali pa jo sprejmemo (ali pa se vzdržimo presoje).

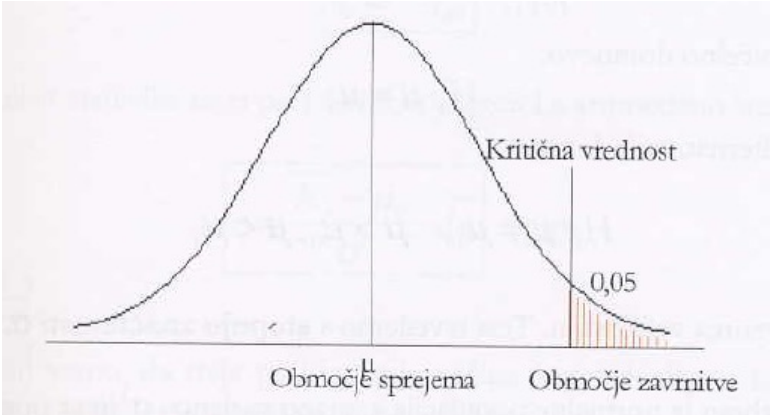
Primer 8.2.:

Imamo naslednji hipotezi:

H_0 : *parameter testa* = *določena vrednost*
H_1 : *parameter testa* > *določena vrednost*

Ilustrirajte test aritmetične sredine!

Ilustracija je prikazana na sliki 164 [Brvar].



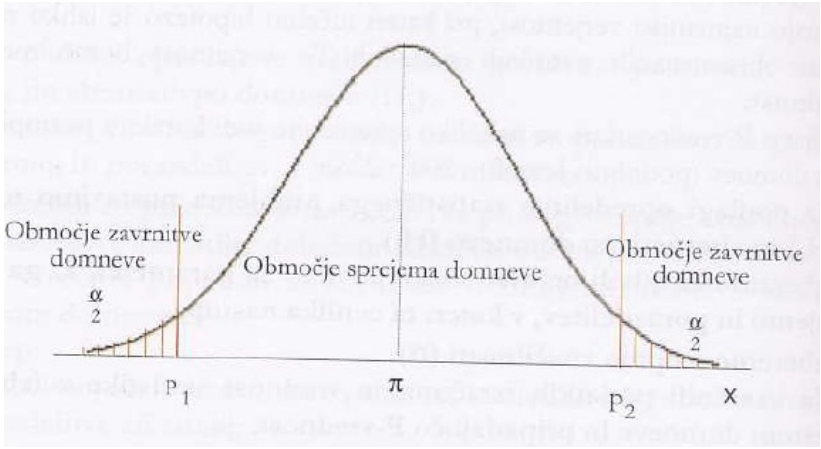
Slika 164: Enostranski test aritmetične sredine [Brvar]

Imamo naslednji hipotezi:

H_0 : *parameter testa = določena vrednost*
H_1 : *parameter testa ≠ določena vrednost*

Ilustrirajte test aritmetične sredine!

Ilustracija je prikazana na sliki 165 [Brvar].



Slika 165: Dvostranski test aritmetične sredine [Brvar]

Primer 8.3.:

Postavimo domnevo o vrednosti populacijskega deleža [Brvar]:

$$H : \hat{P} = 0.33 \tag{8.15}$$

Sestavljamo naključne vzorce velikosti n = 700 in na vsakem vzorcu ugotovimo vzorčni delež $\hat{P}$. *Ob predpostavki, da je domneva pravilna, vemo, da se vzorčni deleži porazdelujejo približno normalno (glej izraz (7.93)) in velja:*

$$Z = \frac{\hat{P} - p}{\sqrt{\frac{\hat{P} \cdot \left(1 - \hat{P}\right)}{n}}} \tag{8.16}$$

Okoli vrednosti $\hat{P}$ *naredimo območje sprejemanja domneve in izven tega območja kritično območje zavračanja domneve. Denimo, da je območje zavračanja določeno s 5%*

vzorcev ($\alpha = 0.05$), ki imajo ekstremne vrednosti deležev: 2.5% na levi in 2.5% na desni strani porazdelitve. Določite obe vrednosti, ki definirata območje zavračanja hipoteze!

Na osnovi izraza (7.93) lahko zapišemo:

$$P\left(\hat{P} - z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{P}\cdot\left(1-\hat{P}\right)}{n}} \le p \le \hat{P} + z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{P}\cdot\left(1-\hat{P}\right)}{n}}\right) = 1-\alpha \tag{8.17}$$

$$P\left(0.33 - 1.96 \cdot \sqrt{\frac{0.33\cdot(1-0.33)}{700}} \le p \le 0.33 + 1.96 \cdot \sqrt{\frac{0.33\cdot(1-0.33)}{700}}\right) = 0.95$$

$$P\left(0.33 - 0.0348 \le p \le 0.33 + 0.0348\right) = 0.95$$

$$P\left(0.2952 \le p \le 0.3648\right) = 0.95$$

Torej sta pod vrednostjo $p_1 = 0.2952$ in nad vrednostjo $p_2 = 0.3648$ območji zavračanja hipoteze.

V nadaljevanju si bomo pogledali nekaj testov hipotez, ki jih največkrat uporabljamo. Pri vseh bomo predpostavili, da so vzorci izbrani iz normalne populacije, ali pa so tako veliki, da lahko vzamemo normalno populacijo kot približek.

8.2 Test aritmetične sredine

Predpostavimo, da želimo testirati ničelno hipotezo:

$$H_0 : \mu = \mu_0 \tag{8.18}$$

pri eni izmed nasprotnih hipotez:

$$H_1 : \mu \neq \mu_0 \qquad (8.19)$$
$$H_1 : \mu > \mu_0$$
$$H_1 : \mu < \mu_0$$

na osnovi vzorca velikosti n, izbranega iz normalne populacije z znano varianco σ^2 in pri napaki 1. vrste α.

Statistika testa hipoteze je v tem primeru, v skladu s centralnim limitnim teoremom, enaka (glej poglavje 7.9):

$$Z = \frac{\left(\bar{X} - \mu_0\right)}{\frac{\sigma}{\sqrt{n}}} \qquad (8.20)$$

Kritično območje je glede na nasprotne hipoteze določeno na naslednji način:

$$-z_{\frac{\alpha}{2}} \leq Z \leq z_{\frac{\alpha}{2}}$$
$$ali \qquad (8.21)$$
$$Z \geq z_{\alpha}$$
$$ali$$
$$Z \leq -z_{\alpha}$$

Primer 8.4.:

Predpostavimo, da je standardni odklon 10 dag težkih zavitkov kave (normalna populacija), polnjenih v neki pražarni, enak 0.32 dag. Da bi ugotovili, če je polnjenje nekega določenega dne pod nadzorom, to je, da je prava povprečna teža zavitkov $\mu = 10\,dag$*, smo izbrali vzorec 36 zavitkov in izračunali, da je bila povprečna teža 10.11 dag. Glede na to, da bi v primeru, ko je* $\mu > 10\,dag$*, pražarna imela izgubo, če je* $\mu < 10\,dag$*, bi pa kupci imeli izgubo, testirajte hipotezo* $\mu = 10\,dag$ *pri nasprotni hipotezi* $\mu \neq 10\,dag$ *za stopnjo pomembnosti 0.01.[Jesenko]*

Imamo:

$$\begin{aligned} &H_0 : \mu = \mu_0 = 10 \\ &H_1 : \mu \neq \mu_0 = 10 \\ &\sigma = 0.32 \\ &n = 36 \\ &\bar{x} = 10.11 \\ &\alpha = 0.01 \\ &z_{\frac{\alpha}{2}} = z_{\frac{0.01}{2}} = z_{0.005} = 2.57 \end{aligned} \tag{8.22}$$

Vrednost statistike je enaka:

$$z = \frac{(\bar{x} - \mu_0)}{\frac{\sigma}{\sqrt{n}}} = \frac{(10.11 - 10)}{\frac{0.32}{\sqrt{36}}} = \frac{(0.11) \cdot 6}{0.32} = 2.0625 \tag{8.23}$$

Ker je neenakost:

$$\begin{aligned} &-z_{\frac{\alpha}{2}} \leq z \leq z_{\frac{\alpha}{2}} \\ &-2.57 \leq 2.0625 \leq 2.57 \end{aligned} \tag{8.24}$$

izpolnjena in smo v območju zaupanja, ničelne hipoteze ne moremo zavreči. Torej lahko rečemo, da je prava povprečna teža zavitkov $\mu = 10\, dag$, proces polnjenja vrečk pa dotičnega dne pod nadzorom.

Ko imamo opravka z velikimi vzorci iz populacij, ki niso normalno porazdeljene, imajo pa znano varianco, lahko na osnovi centralnega limitnega izreka še vedno uporabimo isto statistiko za test hipoteze kot v primeru normalne populacije. Če je varianca populacije neznana, jo pa nadomestimo z varianco vzorca [Jesenko].

Primer 8.5.:

V proizvodnji avtomobilskih žarnic so na osnovi vzorca 100 žarnic neznane populacije ugotovili, da je bila povprečna življenjska doba žarnic 742 ur s standardnim odklonom 48 ur. Testirajte ničelno hipotezo $\mu = 750\, ur$ *pri nasprotni hipotezi* $\mu < 750\, ur$ *pri stopnji pomembnosti 0.05.[Jesenko]*

Imamo:

$$
\begin{aligned}
&H_0 : \mu = \mu_0 = 750 \\
&H_1 : \mu < \mu_0 = 750 \\
&\sigma = 48 \\
&n = 100 \\
&\bar{x} = 742 \\
&\alpha = 0.05 \\
&z_\alpha = z_{0.05} = 1.6449 \quad \left(\text{z=norminv(1-0.05,0,1)}\right)
\end{aligned}
\tag{8.25}
$$

Vrednost statistike je enaka:

$$
z = \frac{\left(\bar{x} - \mu_0\right)}{\frac{\sigma}{\sqrt{n}}} = \frac{\left(742 - 750\right)}{\frac{48}{\sqrt{100}}} = \frac{\left(-8\right)\cdot 10}{48} = -1.6667 \tag{8.26}
$$

Ker je neenakost:

$$
\begin{aligned}
&z \le -z_\alpha \\
&-1.6667 \le -1.6449
\end{aligned}
\tag{8.27}
$$

izpolnjena, nismo v območju zaupanja (saj je to večje od -1.6449), pač pa v kritičnem območju, zato moramo ničelno hipotezo zavreči. Torej sprejmemo nasprotno hipotezo: $\mu < 750$ *ur*.

Kadar je velikost vzorca majhna in je varianca normalne populacije neznana, ne moremo več uporabiti *z* statistike, pač pa preiti na *t* statistiko (glej poglavje 7.9):

$$
T = \frac{\left(\bar{X} - \mu_0\right)}{\frac{S}{\sqrt{n}}} \tag{8.28}
$$

z *n - 1* prostostnimi stopnjami.

Kritično območje je glede na nasprotne hipoteze določeno na naslednji način:

$$
\begin{aligned}
&-t_{\frac{\alpha}{2},n-1} \le T \le t_{\frac{\alpha}{2},n-1} \\
&ali \\
&T \ge t_{\alpha,n-1} \\
&ali \\
&T \le -t_{\alpha,n-1}
\end{aligned}
\tag{8.29}
$$

Primer 8.6.:

Specifikacija za določeno vrsto žice določa, da se v povprečju žica pretrga, če jo obremenimo z 250 kg. Iz različno izbranih kolutov žice smo izbrali 7 koncev žice (normalna populacija) in videli, da so se pretrgali pri obremenitvah 271.6 kg, 231.8 kg, 248.5 kg, 261.9 kg, 238.8 kg, 255.6 kg in 252.1 kg. Testirajte ničelno hipotezo $\mu = 250$ *kg pri nasprotni hipotezi* $\mu < 250$ *kg za stopnjo pomembnosti 0.05 [Jesenko].*

Imamo:

$$
\begin{aligned}
&H_0 : \mu = \mu_0 = 250 \\
&H_1 : \mu < \mu_0 = 250 \\
&n = 7 \\
&\alpha = 0.05 \\
&t_\alpha = t_{0.05} = -1.9432 \quad \text{(t=tinv(0.05,7-1))}
\end{aligned}
\tag{8.30}
$$

Izračunajmo aritmetično sredino vzorca:

$$
\bar{x} = \frac{271.6+231.8+248.5+261.9+238.8+255.6+252.1}{7} = 251.4714 \tag{8.31}
$$

in varianco oz. deviacijo:

$$
s^2 = \frac{(271.6-251.4714)^2 + (231.8-251.4714)^2 + ... + (252.1-251.4714)^2}{6} = 181.2857 \tag{8.32}
$$

$$
s = 13.4642
$$

Ta rezultat lahko dobimo z naslednjimi ukazi v Matlabu:

```
>> x=[271.6 231.8 248.5 261.9 238.8 255.6 252.1];
>> var=sum((x-mean(x)).^2)/6
var =
  181.2857
>> std=sqrt(var)
std =
   13.4642
```

Seveda bi pa lahko uporabili tudi ukaza **var(x)** oz**. std(x).**

Vrednost statistike je enaka:

$$t = \frac{(\bar{x} - \mu_0)}{\frac{s}{\sqrt{n}}} = \frac{(\bar{x} - \mu_0)}{\frac{s}{\sqrt{n}}} = \frac{(251.4714 - 250)}{\frac{13.4642}{\sqrt{7}}} = \frac{(1.4714)}{13.4642}\sqrt{7} = 0.2891 \qquad (8.33)$$

Ker neenakost:

$$\begin{aligned} & t \le -t_\alpha \\ & 0.2891 \le -1.9432 \end{aligned} \qquad (8.34)$$

ni izpolnjena, saj je $0.2891 > -1.9432$, smo v območju zaupanja (saj je to večje od -1.9432), torej lahko ničelno hipotezo sprejmemo oz. je ne moremo zavrniti. Seveda smo pri tem sklepu izpostavljeni napaki 2. vrste.

8.3 Test razlike aritmetičnih sredin

Velikokrat nas zanimajo hipoteze, povezane z razlikami med aritmetičnima sredinama dveh populacij, kot npr., če moški izvršijo neko delo hitreje kot ženske, če so povprečni tedenski izdatki v enem mestu večji kot v drugem, če je uspešnost dopoldanske izmene dela v podjetju večja od popoldanske, itn [Jesenko].

Denimo imamo 2 normalni populaciji. Ena naj ima aritmetično sredino in varianco $\mu_1 = E(X_1), \sigma_1^2 = VAR(X_1),$ druga pa aritmetično sredino in varianco

$\mu_2 = E(X_2), \sigma_2^2 = VAR(X_2)$. Iz prve populacije izberemo naključni vzorec velikosti n_1, iz druge populacije pa izberemo naključni vzorec velikosti n_2.

Predpostavimo, da želimo testirati ničelno hipotezo:

$$H_0 : \mu_1 - \mu_2 = \delta \tag{8.35}$$

pri eni izmed nasprotnih hipotez:

$$\begin{aligned} H_1 &: \mu_1 - \mu_2 \neq \delta \\ H_1 &: \mu_1 - \mu_2 > \delta \\ H_1 &: \mu_1 - \mu_2 < \delta \end{aligned} \tag{8.36}$$

Statistika testa hipoteze je v tem primeru enaka (glej izraz (7.64)):

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(\bar{X}_1 - \bar{X}_2) - (\delta)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{8.37}$$

Kritično območje je glede na nasprotne hipoteze določeno na naslednji način:

$$\begin{aligned} & -z_{\frac{\alpha}{2}} \leq Z \leq z_{\frac{\alpha}{2}} \\ & ali \\ & Z \geq z_\alpha \\ & ali \\ & Z \leq -z_\alpha \end{aligned} \tag{8.38}$$

kjer je:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\delta)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{8.39}$$

vrednost standardizirane normalne naključne spremenljivke, $\overline{x}_1, \overline{x}_2$ pa sta aritmetični sredini vzorcev.

Kadar sta varianci populacij σ_1^2 in σ_2^2 neznani, ju lahko pri velikih vzorcih nadomestimo z variancama vzorcev s_1^2 in s_2^2 , statistika testa hipotez pa ostane zaradi centralnega limitnega izreka enaka [Jesenko].

Primer 8.7.:

Ugotoviti so želeli, če zrak v mestu A v povprečju vsebuje za 2 g/m3 ogljikovega dioksida več kot zrak v mestu B. V določenem obdobju so v mestu A na naključnem vzorcu velikosti 55 merili prisotnost CO2 v zraku. Ugotovili so, da je bila povprečna vrednost CO2 enaka 28.2 g/m3, standardni odklon pa 2.3 g/m3. Prav tako so v tem obdobju v mestu B na naključnem vzorcu velikosti 45 merili prisotnost CO2 v zraku. Ugotovili so, da je bila povprečna vrednost CO2 enaka 27.1 g/m3, standardni odklon pa 2.8 g/m3. Testirajte ničelno hipotezo $H_0 : \mu_1 - \mu_2 = 2\ g/m^3$ *pri nasprotni hipotezi* $H_1 : \mu_1 - \mu_2 \neq 2\ g/m^3$*, če je* $\alpha = 0.05$ *[Jesenko].*

Imamo:

$$
\begin{aligned}
&H_0 : \mu_1 - \mu_2 = \delta = 2\ g/m^3 \\
&H_1 : \mu_1 - \mu_2 = \delta \neq 2\ g/m^3 \\
&n_1 = 55 \\
&n_2 = 45 \\
&\overline{x}_1 = 28.2\ g/m^3 \\
&s_1 = 2.3\ g/m^3 \\
&\overline{x}_2 = 27.1\ g/m^3 \\
&s_2 = 2.8\ g/m^3 \\
&\alpha = 0.05 \\
&\left| z_{\frac{\alpha}{2}} \right| = \left| z_{\frac{0.05}{2}} \right| = 1.96 \quad \text{(z=abs(norminv(0.05/2,0,1))}
\end{aligned}
\tag{8.40}
$$

Sledi:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\delta)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(28.2 - 27.1) - (2)}{\sqrt{\frac{2.3^2}{55} + \frac{2.8^2}{45}}} = \frac{-0.9}{0.52} = -1.7308 \tag{8.41}$$

Ker je neenakost:

$$\begin{aligned} & -z_{\frac{\alpha}{2}} \le z \le z_{\frac{\alpha}{2}} \\ & -1.96 \le -1.7308 \le 1.96 \end{aligned} \tag{8.42}$$

izpolnjena in smo v območju zaupanja, ničelne hipoteze ne moremo zavreči. Torej lahko rečemo, da je razlika med vsebnostima CO2 v obeh opazovanih mestih: 28.2 - 27.1 = 1.1 g/m3 statistično pomembno različna od 2 g/m3 [Jesenko].

Kadar sta velikosti vzorcev majhni, varianci pa neznani, pravkar opisanega testa ne moremo uporabiti [Jesenko]. Statistika testa hipoteze je v tem primeru enaka (glej izraz (7.75)):

$$\begin{aligned} T &= \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1 - 1) \cdot S_1^{\,2} + (n_2 - 1) \cdot S_2^{\,2}}{n_1 + n_2 - 2}} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \\ &= \frac{(\bar{X}_1 - \bar{X}_2) - (\delta)}{\sqrt{\frac{(n_1 - 1) \cdot S_1^{\,2} + (n_2 - 1) \cdot S_2^{\,2}}{n_1 + n_2 - 2}} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \end{aligned} \tag{8.43}$$

in imamo *t* porazdelitev z $n_1 + n_2 - 2$ prostostnimi stopnjami.

Kritično območje je glede na nasprotne hipoteze določeno na naslednji način:

$$\begin{aligned} & -t_{\frac{\alpha}{2}, n_1 + n_2 - 2} \le T \le t_{\frac{\alpha}{2}, n_1 + n_2 - 2} \\ & \textit{ali} \\ & T \ge t_{\alpha, n_1 + n_2 - 2} \\ & \textit{ali} \\ & T \le -t_{\alpha, n_1 + n_2 - 2} \end{aligned} \tag{8.44}$$

Primer 8.8.:

Mizarski mojster je želel ugotoviti, če lahko z enakim lakom dveh različnih proizvajalcev polakira enako površino desk. Izbral je po štiri enako velike embalaže laka pri obeh proizvajalcih. Z vsako embalažo laka prvega proizvajalca je v povprečju polakiral 342 dm2 desk, standardni odklon pa je bil 24 dm2. Z vsako embalažo laka drugega proizvajalca pa je v povprečju polakiral 361 dm2 desk, standardni odklon pa je bil 26 dm2. Predpostavimo, da sta obe populaciji porabe laka normalni z enako varianco. Testirajte ničelno hipotezo $H_0 : \mu_1 - \mu_2 = 0\ dm^2$ *pri nasprotni hipotezi* $H_1 : \mu_1 - \mu_2 < 0\ dm^2$*, če je* $\alpha = 0.05$ *[Jesenko].*

Imamo:

$$
\begin{aligned}
&H_0 : \mu_1 - \mu_2 = 0\ dm^2 \\
&H_1 : \mu_1 - \mu_2 < 0\ dm^2 \\
&n_1 = 4 \\
&n_2 = 4 \\
&\bar{x}_1 = 342\, dm^2 \\
&s_1 = 24\, dm^2 \\
&\bar{x}_2 = 361\, dm^2 \\
&s_2 = 26\, dm^2 \\
&\alpha = 0.05 \\
&-t_{\alpha, n_1+n_2-2} = -t_{0.05,4+4-2} = -1.9432 \quad \text{(t=tinv(0.05,4+4-2))}
\end{aligned}
\tag{8.45}
$$

Sledi:

(8.46)

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\delta)}{\sqrt{\frac{(n_1 - 1)\cdot s_1^{\,2} + (n_2 - 1)\cdot s_2^{\,2}}{n_1 + n_2 - 2}} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} =$$

$$= \frac{(342 - 361) - (0)}{\sqrt{\frac{(4-1)\cdot 24^2 + (4-1)\cdot 26^2}{4+4-2}} \cdot \sqrt{\frac{1}{4} + \frac{1}{4}}} =$$

$$= \frac{-19}{\sqrt{\frac{3\cdot 24^2 + 3\cdot 26^2}{6}} \cdot \sqrt{\frac{1}{2}}} = \frac{-19}{\sqrt{\frac{\left(24^2 + 26^2\right)}{4}}} =$$

$$= \frac{-38}{\sqrt{\left(24^2 + 26^2\right)}} = -1.0739$$

Ker neenakost:

$$t \le -t_{\alpha, n_1 + n_2 - 2} \qquad (8.47)$$
$$-1.0739 \le -1.9432$$

ni izpolnjena in velja $-1.0739 > -1.9432$, smo v območju zaupanja, torej ničelne hipoteze ne moremo zavreči. Torej lahko sklepamo, da mizarski mojster z lakoma obeh proizvajalcev v povprečju polakira enako površino desk.

8.4 Test variance

Primerov, kjer lahko uporabimo test variance, je veliko. Npr. vzemimo proizvodno podjetje, ki je dobilo zelo stroge zahteve glede kakovosti izdelkov in zato želi testirati variabilnost kakovosti. Podoben je primer, ko želi farmacevt preveriti, če je variabilnost nekega zdravila v predpisanih mejah. Včasih pa test variance izvedemo, preden izvedemo test kakšnega drugega parametra. Na primer, pri testu enakosti aritmetičnih sredin včasih zahtevamo enakost varianc populacij, iz katerih izbiramo vzorce [Jesenko].

Dan imamo naključni vzorec velikosti *n* iz normalne populacije. Testirati želimo ničelno hipotezo:

$$H_0 : \sigma^2 = \sigma_0^{\,2} \qquad (8.48)$$

pri eni izmed nasprotnih hipotez:

$$\begin{aligned} H_1 &: \sigma^2 \neq {\sigma_0}^2 \\ H_1 &: \sigma^2 > {\sigma_0}^2 \\ H_1 &: \sigma^2 < {\sigma_0}^2 \end{aligned} \tag{8.49}$$

Iz poglavja 6.3 (glej izraz (6.22)) in poglavja 7.12 (glej izraz (7.103)) vemo, da velja:

$$\chi^2(n-1) = \frac{1}{\sigma^2} \cdot S^2 \cdot (n-1) \tag{8.50}$$

kar je statistika testa hipoteze. Kritično območje je glede na nasprotne hipoteze določeno na naslednji način:

$$\begin{aligned} &\left(\chi^2(n-1) \leq {\chi^2}_{1-\frac{\alpha}{2}}(n-1) \right) \vee \left(\chi^2(n-1) \geq {\chi^2}_{\frac{\alpha}{2}}(n-1) \right) \\ &\textit{ali} \\ &\chi^2(n-1) \leq {\chi^2}_{1-\alpha}(n-1) \\ &\textit{ali} \\ &\chi^2(n-1) \geq {\chi^2}_{\alpha}(n-1) \end{aligned} \tag{8.51}$$

Primer 8.9.:

V nekem izdelku, ki ga izdelujejo v podjetju, je debelina določenega sestavnega elementa njegova kritična komponenta. Na naključnem vzorcu 16 sestavnih delov smo ugotovili, da je bila varianca debeline $s^2 = 0.72$ *pri izbrani merski enoti. Proizvodni proces je pod nadzorom, če variabilnost debeline kritičnih sestavnih delov, izražena z varianco, ni večja od 0.43. Predpostavimo, da je debelina teh elementov normalna naključna spremenljivka. Testirajte ničelno hipotezo* $\sigma^2 = 0.43$, *pri nasprotni hipotezi* $\sigma^2 > 0.43$, *če je* $\alpha = 0.05$ *[Jesenko].*

Imamo:

$$\begin{aligned} &H_0 : \sigma^2 = \sigma_0{}^2 = 0.43 \\ &H_0 : \sigma^2 > \sigma_0{}^2 = 0.43 \\ &n = 16 \\ &s^2 = 0.72 \\ &\alpha = 0.05 \\ &\chi^2{}_{\alpha}(n-1) = \chi^2{}_{0.05}(16-1) = 24.9958 \qquad (\text{hi=chi2inv(1-0.05,16-1)}) \end{aligned} \tag{8.52}$$

Sledi:

$$\begin{aligned} &\chi^2(n-1) = \frac{1}{\sigma^2} \cdot s^2 \cdot (n-1) = \frac{1}{0.43} \cdot 0.72 \cdot (16-1) = \\ &= 25.1163 \end{aligned} \tag{8.53}$$

Ker je neenakost:

$$\begin{aligned} &\chi^2(n-1) \geq \chi^2{}_{\alpha}(n-1) \\ &25.1163 \geq 24.9958 \end{aligned} \tag{8.54}$$

izpolnjena, smo v kritičnem območju, zato moramo ničelno hipotezo zavrniti. Torej velja obratna hipoteza, da je varianca večja od 0.43, zato sklepamo, da proizvodni proces ni pod nadzorom. Kot se izkaže, pa pri $\alpha = 0.01$ ničelne hipoteze ne bi zavrnili, iz česar se vidi, kako pomemben je izbor parametra α.

Test enakosti oz. kvocienta varianc

Denimo imamo $S_1{}^2$ in $S_2{}^2$, ki sta varianci dveh neodvisnih naključnih vzorcev velikosti n_1 in n_2 iz dveh normalnih populacij z variancama σ_1^2 in σ_2^2. Potem je (glej (7.108) :

$$F = \frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}} = \frac{\sigma_2^2}{\sigma_1^2} \cdot \frac{S_1^2}{S_2^2} \tag{8.55}$$

F naključna spremenljivka z $n_1 - 1$ in $n_2 - 1$ prostostnimi stopnjami. Testirati želimo ničelno hipotezo:

$$H_0 : \sigma_1^2 = \sigma_2^2 \tag{8.56}$$

pri eni izmed nasprotnih hipotez:

$$\begin{aligned} H_1 &: \sigma_1^2 \neq \sigma_2^2 \\ H_1 &: \sigma_1^2 > \sigma_2^2 \\ H_1 &: \sigma_1^2 < \sigma_2^2 \end{aligned} \tag{8.57}$$

Pri ničelni hipotezi je: $F = \frac{\sigma_2^2}{\sigma_1^2} \cdot \frac{S_1^2}{S_2^2} = \frac{S_1^2}{S_2^2}$. Kritično območje je glede na nasprotne hipoteze določeno na naslednji način:

$$\begin{aligned} &\left(\frac{s_1^2}{s_2^2} \geq f_{\frac{\alpha}{2}}\left(n_1 - 1, n_2 - 1\right) \right) \vee \left(\frac{s_1^2}{s_2^2} \leq f_{1-\frac{\alpha}{2}}\left(n_1 - 1, n_2 - 1\right) = \frac{1}{f_{\frac{\alpha}{2}}\left(n_2 - 1, n_1 - 1\right)} \right) \\ &\left(\frac{s_1^2}{s_2^2} \geq f_{\frac{\alpha}{2}}\left(n_1 - 1, n_2 - 1\right) \right) \vee \left(\frac{s_2^2}{s_1^2} \geq f_{\frac{\alpha}{2}}\left(n_2 - 1, n_1 - 1\right) \right) \text{.....(dvorepi)} \\ &\textit{ali} \\ &\frac{s_1^2}{s_2^2} \leq f_{1-\alpha}\left(n_1 - 1, n_2 - 1\right) = \frac{1}{f_{\alpha}\left(n_2 - 1, n_1 - 1\right)} \\ &\frac{s_2^2}{s_1^2} \geq f_{\alpha}\left(n_2 - 1, n_1 - 1\right) \text{.....(levi enorepi)} \\ &\textit{ali} \\ &\frac{s_1^2}{s_2^2} \geq f_{\alpha}\left(n_1 - 1, n_2 - 1\right) \text{.....(desni enorepi)} \end{aligned} \tag{8.58}$$

Pri dvorepem testu še velja, da se levi pogoj testira pri $s_1^2 > s_2^2$, desni pa pri $s_2^2 > s_1^2$, pri čemer se testira le eden ali drug pogoj, odvisno pač od tega, kakšno je razmerje varianc [Jesenko].

Primer 8.10.:

Primerjali so variabilnost nateznih trdnosti dveh vrst žice. Za prvo vrsto žice so iz meritev na naključnem vzorcu velikosti 17 izračunali standardni odklon 5.6 kp/mm2. Za drugo vrsto žice so iz meritev na naključnem vzorcu velikosti 12 izračunali standardni odklon 3.1 kp/mm2. Predpostavimo, da predstavljajo meritve dva neodvisna naključna vzorca iz dveh normalnih populacij. Testirajte ničelno hipotezo $H_0 : \sigma_1^2 = \sigma_2^2$ *pri nasprotni hipotezi* $H_1 : \sigma_1^2 \neq \sigma_2^2$*, pri čemer je* $\alpha = 0.02$ *[Jesenko].*

Imamo:

$$
\begin{aligned}
&s_1 = 5.6 \\
&s_2 = 3.1 \\
&n_1 = 17 \\
&n_2 = 12 \\
&\alpha = 0.02 \\
&s_1^2 > s_2^2, \quad \textit{zato gledamo kritično vrednost}: \\
&f_{\frac{\alpha}{2}}(n_1 - 1, n_2 - 1) = f_{\frac{0.02}{2}}(17-1, 12-1) = f_{0.01}(16,11) = 4.2134
\end{aligned}
\tag{8.59}
$$

kjer smo si pomagali z Matlabom:

```
>> f1 = finv(1-0.02/2,17-1,12-1)
f1 =
   4.2134
```

Izračunajmo sedaj razmerje:

$$
\frac{s_1^2}{s_2^2} = \frac{5.6^2}{3.1^2} = 3.2633 \tag{8.60}
$$

Ker neenakost:

$$\frac{s_1^{\,2}}{s_2^{\,2}} \geq f_{\frac{\alpha}{2}}\left(n_1 - 1, n_2 - 1\right) \qquad (8.61)$$
$$3.2633 \geq 4.2124$$

ni izpolnjena, nismo v kritičnem območju (smo v območju zaupanja), zato ničelne hipoteze ne moremo zavreči. To pomeni, da lahko sklepamo o enakosti varianc, torej, da sta variabilnosti natezne trdnosti obeh vrst žice enaki.

8.5 Test deleža

Veliko je primerov, kjer predpostavljamo, da je vzorec izbran iz binomske populacije, za katero testiramo ničelno hipotezo:

$$H_0 : p = p_0 \qquad (8.62)$$

pri eni izmed nasprotnih hipotez:

$$\begin{aligned} H_1 &: p \neq p_0 \\ H_1 &: p > p_0 \\ H_1 &: p < p_0 \end{aligned} \qquad (8.63)$$

kjer je p parameter binomske porazdelitve in pomeni verjetnost nastopa nekega naključnega dogodka pri izvedbi določenega poskusa. Binomska naključna spremenljivka pomeni število nastopov uspešnega dogodka pri n ponovitvah poskusa.

Majhna velikost vzorca

Pri stopnji pomembnosti α je kritično območje glede na nasprotne hipoteze pri levorepem testu določeno na naslednji način [Walpole]:

$$P = P\left(X \le x^* \middle| p = p_0\right) \le \alpha \qquad (8.64)$$
$$0 \le x \le x^* \le n$$

kjer je x^* število uspehov v vzorcu, pogojeno s problemom. Če je neenakost izpolnjena, pademo v kritično območje. Pri stopnji pomembnosti α je kritično območje glede na nasprotne hipoteze pri desnorepem testu določeno na naslednji način [Walpole]:

$$\begin{aligned} &P\left(X \le x^* \middle| p = p_0\right) \ge 1-\alpha \\ &oz. \\ &1-P\left(X \ge x^* \middle| p = p_0\right) \ge 1-\alpha \\ &oz. \\ &P\left(X \ge x^* \middle| p = p_0\right) \le \alpha \\ &0 \le x^* \le x \le n \end{aligned} \qquad (8.65)$$

Če je neenakost izpolnjena, pademo v kritično območje. Pri obojestranskem testu pa kritično območje določimo takole [Walpole]:

$$P = \begin{cases} 2P\left(X \le x^* \middle| p = p_0\right), & če \quad x^* < n \cdot p_o \\ 2P\left(X \ge x^* \middle| p = p_0\right), & če \quad x^* > n \cdot p_o \end{cases} \qquad (8.66)$$
$$P \le \alpha ?$$

Če je neenakost izpolnjena, pademo v kritično območje.
Verjetnosti, ki se pojavijo v izrazih (8.64) oz. (8.65), računamo na naslednji način:

$$\begin{aligned} P\left(X \le x^* \middle| p = p_0\right) &= \sum_{x=0}^{x^*} \binom{n}{x} \cdot p_0^{\,x} \cdot \underbrace{\left(1-p_0\right)}_{q_0}^{n-x}, \quad x^* \le n \\ P\left(X \ge x^* \middle| p = p_0\right) &= \sum_{x=x^*}^{n} \binom{n}{x} \cdot p_0^{\,x} \cdot \underbrace{\left(1-p_0\right)}_{q_0}^{n-x} \end{aligned} \qquad (8.67)$$

Podobno seveda velja tudi za verjetnosti pri dvorepem testu.

Primer 8.11.:

Strokovnjak trdi, da so toplotne črpalke instalirane v 70% vseh domov v nekem mestu. Preverite to trditev, če je naključni obhod 15 domov pokazal, da ima le 8 domov instaliranih toplotne črpalke. Pri tem je $\alpha = 0.1$ *[Walpole].*

Imamo:

$$\begin{aligned} &H_0 : p = p_0 = 0.7 \\ &H_1 : p \neq p_0 = 0.7 \\ &n = 15 \\ &\alpha = 0.1 \\ &x^* = 8 \\ &n \cdot p_0 = 15 \cdot 0.7 = 10.5 \\ &x^* < n \cdot p_0 \end{aligned} \tag{8.68}$$

Sledi:

$$\begin{aligned} P &= 2P\left(X \le x^* \middle| p = p_0\right) = 2 \cdot \sum_{x=0}^{x^*} \binom{n}{x} \cdot \underbrace{p_0^{\,x} \cdot \left(1 - p_0\right)^{n-x}}_{q_0} = \\ &= 2 \cdot \sum_{x=0}^{8} \binom{15}{x} \cdot 0.7^x \cdot 0.3^{15-x} = 0.2623 \end{aligned} \tag{8.69}$$

kjer smo si pomagali z Matlabom:

```
>> P=2*binocdf(0:8,15,0.7);
>> P=P(length(P))
P =
   0.2623
```

Ker je $P = 0.2623 > \alpha = 0.1$, nismo v kritičnem območju, zato ničelne hipoteze ne moremo zavreči. Torej naredimo sklep, da ni zadovoljivega razloga, da bi dvomili, da so toplotne črpalke instalirane v 70% vseh domov v nekem mestu.

Primer 8.12.:

Kriminalisti so na vzorcu 20 naključno izbranih ropov želeli preizkusiti novo metodo za odkrivanje storilcev. Po tej metodi so razjasnili 8 ropov. Testirajte ničelno hipotezo, da je uspešnost te metode 60% pri nasprotni hipotezi, da je uspešnost manjša od 60%. Stopnja pomembnosti naj bo 0.05 [Jesenko].

Imamo:

$$\begin{aligned} &H_0 : p = p_0 = 0.6 \\ &H_1 : p < p_0 = 0.6 \\ &n = 20 \\ &\alpha = 0.05 \\ &x^* = 8 \end{aligned} \tag{8.70}$$

Sledi:

$$\begin{aligned} P\left(X \le x^* \middle| p = p_0\right) &= \sum_{x=0}^{x^*} \binom{n}{x} \cdot p_0^{\,x} \cdot \underbrace{(1-p_0)}_{q_0}^{n-x} = \\ &= \sum_{x=0}^{8} \binom{20}{x} \cdot 0.6^x \cdot 0.4^{20-x} = 0.0565 \end{aligned} \tag{8.71}$$

kjer smo si pomagali z Matlabom:

```
>> P=binocdf(0:8,20,0.6);
>> P=P(length(P))
P =
   0.0565
```

Ker je $P = 0.0565 > \alpha = 0.05$, nismo v kritičnem območju, zato ničelne hipoteze ne moremo zavreči. Torej lahko naredimo sklep, da je uspešnost metode za odkrivanje storilcev enaka 60%.

Velika velikost vzorca

Uporabimo testno statistiko (glej izraz (7.82)) [Montgomery1, Walpole]:

$$Z = \frac{X^* - n \cdot p_0}{\sqrt{n \cdot p_0 \cdot (1 - p_0)}} = \frac{\frac{X^*}{n} - p_0}{\sqrt{\frac{p_0 \cdot (1 - p_0)}{n}}} = \frac{\hat{P} - p_0}{\sqrt{\frac{p_0 \cdot (1 - p_0)}{n}}} \tag{8.72}$$

ki je približno standardizirana normalna naključna spremenljivka.

Kritično območje je glede na nasprotne hipoteze določeno na naslednji način:

$$\begin{aligned} & -z_{\frac{\alpha}{2}} \le Z \le z_{\frac{\alpha}{2}} \\ & ali \\ & Z \ge z_{\alpha} \\ & ali \\ & Z \le -z_{\alpha} \end{aligned} \tag{8.73}$$

Primer 8.13.:

Proizvajalec polprevodnikov proizvaja krmilnike za avtomobilske motorje. Naročnik zahteva, da izmet ne sme preseči 5%. Proizvajalec zajame naključni vzorec 200 naprav in ugotovi, da so 4 pokvarjene. A proizvajalec izpolnjuje zahteve naročnika, če je stopnja pomembnosti 0.05 [Montgomery 1]?

Imamo:

$$
\begin{aligned}
&H_0 : p = p_0 = 0.05 \\
&H_1 : p < p_0 = 0.05 \\
&n = 200 \\
&\alpha = 0.05 \\
&x^* = 4 \\
&z_\alpha = z_{0.05} = -1.6449 \qquad \text{(z=norminv(0.05,0,1))}
\end{aligned}
\tag{8.74}
$$

Sledi:

$$
z = \frac{x^* - n \cdot p_0}{\sqrt{n \cdot p_0 \cdot (1 - p_0)}} = \frac{4 - 200 \cdot 0.05}{\sqrt{200 \cdot 0.05 \cdot (1 - 0.05)}} = \frac{-6}{3.0822} = -1.9467 \tag{8.75}
$$

Ker je neenakost:

$$
\begin{aligned}
&z \le -z_\alpha \\
&-1.9467 \le -1.6449
\end{aligned}
\tag{8.76}
$$

izpolnjena, smo v kritičnem območju. Zato ničelno hipotezo zavržemo in velja nasprotna hipoteza, torej lahko rečemo, da je defektnega izmeta manj kot 5% in proizvajalec izpolnjuje zahteve naročnika.

Izračune bi lahko opravili tudi z naslednjim programom v Matlabu:

```
% test deleza delez_test.m (veliki vzorci)

clear
clc
close all

n = input('n=')
p0 = input('p0=')
xzv = input('xzv=')
alfa = input('alfa=')

ch = input('levi rep(1) desni rep (2) dvorepni (3)');
```

```
z = (xzv-n*p0)/sqrt(n*p0*(1-p0))

if ch == 1
    zkrit=norminv(alfa,0,1)
    if z < zkrit
       disp('Padli smo v kriticno obmocje - zavrzi nicelno hipotezo')
    else
       disp('Padli smo v obmocje zaupanja - obdrzi nicelno hipotezo')
    end
elseif ch == 2
    zkrit=norminv(1-alfa,0,1)
    if z > zkrit
       disp('Padli smo v kriticno obmocje - zavrzi nicelno hipotezo')
    else
       disp('Padli smo v obmocje zaupanja - obdrzi nicelno hipotezo')
    end
else
    zkrit=norminv([alfa/2 1-alfa/2],0,1)
    if (z < zkrit(1))||(z>zkrit(2))
       disp('Padli smo v kriticno obmocje - zavrzi nicelno hipotezo')
    else
       disp('Padli smo v obmocje zaupanja - obdrzi nicelno hipotezo')
    end
end
```

Izpis komandnega okna je naslednji:

```
n=200
n =
  200
p0=0.05
p0 =
   0.0500
xzv=4
xzv =
    4
alfa=0.05
alfa =
   0.0500

levi rep(1) desni rep (2) dvorepni (3)1

z =
  -1.9467

zkrit =
  -1.6449

Padli smo v kriticno obmocje - zavrzi nicelno hipotezo
```

Primer 8.14.:

Vzorec velikosti 100 je vzet na osnovi Bernoullijevih poskusov iz velike binomske populacije, ki ima neznan delež uspešnih elementov. Pri tem dobimo na osnovi vzorca: $\hat{P} = 0.49$. *Naredite desnorepi test, kjer ničelna hipoteza pomeni* $p_0 = 0.40$ *pri* $\alpha = 0.05$ *[Bernstein].*

Imamo:

$$
\begin{aligned}
&H_0 : p = p_0 = 0.40 \\
&H_1 : p > p_0 = 0.40 \\
&\hat{P} = 0.49 \\
&n = 100 \\
&\alpha = 0.05 \\
&x^* = \hat{P} \cdot n = 49 \\
&z_{1-\alpha} = z_{0.95} = 1.6449 \qquad \text{(z=norminv(0.95,0,1))}
\end{aligned}
\tag{8.77}
$$

Sledi:

$$
z = \frac{x^* - n \cdot p_0}{\sqrt{n \cdot p_0 \cdot (1 - p_0)}} = \frac{49 - 100 \cdot 0.40}{\sqrt{100 \cdot 0.40 \cdot (1 - 0.40)}} = \frac{9}{4.8990} = 1.8371 \tag{8.78}
$$

Ker je neenakost:

$$
\begin{aligned}
&z \geq z_\alpha \\
&1.8371 \geq 1.6449
\end{aligned}
\tag{8.79}
$$

izpolnjena, smo v kritičnem območju. Zato ničelno hipotezo zavržemo in velja nasprotna hipoteza, torej lahko rečemo, da je $p > p_0 = 0.40$.

Ponovno bi si lahko pomagali s programom **delez_test.m**, pri čemer bi dobili naslednji izpis komandnega okna:

```
n=100
n =
   100
p0=0.40
p0 =
```

```
    0.4000
xzv=49
xzv =
    49
alfa=0.05
alfa =
    0.0500

levi rep(1) desni rep (2) dvorepni (3)2

z =
    1.8371

zkrit =
    1.6449

Padli smo v kriticno obmocje - zavrzi nicelno hipotezo
```

Testiranje hipotez o razlikah med več deleži

V nekaterih problemih aplikativnega raziskovanja moramo na osnovi vzorcev odločati o tem, ali so razlike med več deleži pomembne, ali pa jih lahko obravnavamo kot naključne. Tako nas npr. zanima, če so deleži volilcev v 4 mestih, ki podpirajo isto stranko, enaki ali ne. Denimo imamo vrednosti $x_1, x_2, ..., x_k$ k neodvisnih binomskih naključnih spremenljivk $X_1, X_2, ..., X_k$ s parametri $n_1, n_2, ..., n_k$ in $p_1, p_2, ..., p_k$. Če so števila $n_1, n_2, ..., n_k$ zadosti velika, lahko uporabimo standardizirane normalne naključne spremenljivke [Jesenko]:

$$Z_i = \frac{X_i - n_i \cdot p_i}{\sqrt{n_i \cdot p_i \cdot (1 - p_i)}}, \quad i = 1, 2, ..., k \tag{8.80}$$

V poglavju 5.2.5 smo videli, da ima naključna spremenljivka (glej izraz (5.122)):

$$Z = Z_1^{\,2} + Z_2^{\,2} + ... + Z_k^{\,2} = \sum_{i=1}^{k} Z_i^{\,2} \tag{8.81}$$

hi kvadrat porazdelitev. Torej spadata:

$$\chi^2 = \sum_{i=1}^{k} \left(\frac{X_i - n_i \cdot p_i}{\sqrt{n_i \cdot p_i \cdot (1 - p_i)}} \right)^2 = \sum_{i=1}^{k} \frac{(X_i - n_i \cdot p_i)^2}{n_i \cdot p_i \cdot (1 - p_i)}, \quad i = 1,2,...,k \tag{8.82}$$

in njena realizacija :

$$\chi^2 = \sum_{i=1}^{k} \left(\frac{x_i - n_i \cdot p_i}{\sqrt{n_i \cdot p_i \cdot (1 - p_i)}} \right)^2 = \sum_{i=1}^{k} \frac{(x_i - n_i \cdot p_i)^2}{n_i \cdot p_i \cdot (1 - p_i)}, \quad i = 1,2,...,k$$

v razred hi kvadrat naključnih spremenljivk s *k* prostostnimi stopnjami. Če postavimo:

$$\begin{aligned} &H_0 : p_1 = p_2 = ... = p_k = p_o \\ &H_1 : vsaj\ en\ \ p_i\ \ različen\ od\ \ ostalih \end{aligned} \tag{8.83}$$

potem kritično območje določimo na naslednji način [Jesenko]:

$$\chi^2 = \sum_{i=1}^{k} \frac{(x_i - n_i \cdot p_0)^2}{n_i \cdot p_0 \cdot (1 - p_0)} \geq \chi^2(\alpha, k) \tag{8.84}$$

Če p_o ni znan, vzamemo namesto njega [Jesenko]:

$$\hat{p} = \frac{x_1 + ... + x_k}{n_1 + ... + n_k} \tag{8.85}$$

in kritično območje določimo na naslednji način [Jesenko]:

$$\chi^2 = \sum_{i=1}^{k} \frac{(x_i - n_i \cdot \hat{p})^2}{n_i \cdot \hat{p} \cdot (1 - \hat{p})} \geq \chi^2(\alpha, k-1) \tag{8.86}$$

Uredimo podatke vseh *k* vzorcev v tabelo na sliki 166.

	A	$\overline{A}$
Vzorec 1	x_1	$n_1 - x_1$
Vzorec 2	x_2	$n_2 - x_2$
	.....	.....
	.....	.....
Vzorec k	x_k	$n_k - x_k$

Slika 166: Ureditev podatkov k vzorcev v tabelo [Jesenko]

Kot je razvidno iz tabele na sliki 166, so v 1. stolpu podana števila nastopov uspešnih dogodkov, v 2. stolpu pa števila nastopov neuspešnih dogodkov. Za 1. in 2. stolp vpeljimo pojem empiričnih frekvenc na naslednji način [Jesenko]:

$$\begin{aligned} f_{i1} &= x_i, \quad i = 1,....,k \\ f_{i2} &= n_i - x_i, \quad i = 1,....,k \end{aligned} \tag{8.87}$$

Če velja ničelna hipoteza, bi bile teoretične frekvence za 1. in 2. stolp na sliki 166 naslednje [Jesenko]:

$$\begin{aligned} e_{i1} &= n_i \cdot p_o, \quad i = 1,....,k \\ e_{i2} &= n_i - n_i \cdot p_o = n_i \cdot (1 - p_o), \quad i = 1,....,k \end{aligned} \tag{8.88}$$

Če pa p_o ni znan, vzamemo namesto njega oceno $\hat{p}$ in dobimo:

$$\begin{aligned} e_{i1} &= n_i \cdot \hat{p}, \quad i = 1,....,k \\ e_{i2} &= n_i - n_i \cdot \hat{p} = n_i \cdot (1 - \hat{p}), \quad i = 1,....,k \end{aligned} \tag{8.89}$$

Preuredimo števec v izrazu (8.86):

$$\begin{aligned} &(x_i - n_i \cdot \hat{p})^2 = (x_i + n_i - n_i - n_i \cdot \hat{p})^2 = (x_i - n_i + n_i \cdot (1 - \hat{p}))^2 = \\ &= (-f_{i2} + e_{i2})^2 = (f_{i2} - e_{i2})^2 \end{aligned} \tag{8.90}$$

Dokazati se da, da velja tudi enakost [Jesenko]:

$$\left(f_{i2}-e_{i2}\right)^2=\left(f_{i1}-e_{i1}\right)^2 \tag{8.91}$$

Ulomek v izrazu (8.86) lahko zapišemo na naslednji način:

$$\begin{aligned}
&\frac{\left(x_i-n_i\cdot\hat{p}\right)^2}{n_i\cdot\hat{p}\cdot\left(1-\hat{p}\right)}=\frac{\left(x_i-n_i\cdot\hat{p}\right)^2}{n_i}\cdot\frac{1}{\hat{p}\cdot\left(1-\hat{p}\right)}=\frac{\left(x_i-n_i\cdot\hat{p}\right)^2}{n_i}\cdot\left(\frac{1}{\hat{p}}+\frac{1}{\left(1-\hat{p}\right)}\right)=\\
&=\frac{\left(x_i-n_i\cdot\hat{p}\right)^2}{n_i}\cdot\left(\frac{1}{\hat{p}}\right)+\frac{\left(x_i-n_i\cdot\hat{p}\right)^2}{n_i}\cdot\left(\frac{1}{\left(1-\hat{p}\right)}\right)=\\
&=\frac{\left(x_i-n_i\cdot\hat{p}\right)^2}{e_{i1}}+\frac{\left(x_i-n_i\cdot\hat{p}\right)^2}{e_{i2}}=\frac{\left(f_{i2}-e_{i2}\right)^2}{e_{i1}}+\frac{\left(f_{i2}-e_{i2}\right)^2}{e_{i2}}=\\
&=\frac{\left(f_{i1}-e_{i1}\right)^2}{e_{i1}}+\frac{\left(f_{i2}-e_{i2}\right)^2}{e_{i2}}
\end{aligned} \tag{8.92}$$

kjer smo upoštevali tudi enakost (8.91). Vstavimo izraz (8.92) v izraz (8.86) in dobimo:

$$\begin{aligned}
&\chi^2=\sum_{i=1}^{k}\left(\frac{\left(f_{i1}-e_{i1}\right)^2}{e_{i1}}+\frac{\left(f_{i2}-e_{i2}\right)^2}{e_{i2}}\right)\geq\chi^2\left(\alpha,k-1\right)\\
&\chi^2=\sum_{i=1}^{k}\sum_{j=1}^{2}\frac{\left(f_{ij}-e_{ij}\right)^2}{e_{ij}}\geq\chi^2\left(\alpha,k-1\right)
\end{aligned} \tag{8.93}$$

Če je neenakost v izrazu (8.93) izpolnjena, smo v kritičnem območju in moramo ničelno hipotezo zavrniti ter sprejeti nasprotno hipotezo, torej, da je vsaj en delež različen od ostalih. Če pa neenakost ni izpolnjena (smo v območju zaupanja), pa ničelno hipotezo sprejmemo in sklepamo, da so vsi deleži enaki.

Primer 8.15.:

V predvolilnem obdobju je želela neka stranka ugotoviti, če je delež tistih, ki jo bodo volili, enak v 5 regijah. V ta namen so izbrali različno velike vzorce po posameznih regijah in na podlagi ankete, ki so jo izvedli, dobili rezultate, prikazane na sliki 167. Ali je delež volilcev, ki bodo to stranko volili, v vseh 5 regijah enak ($\alpha = 0.05$)? [Jesenko]

	Podpirajo stranko	Število anketiranih
Regija 1	123	250
Regija 2	92	180
Regija 3	163	300
Regija 4	112	200
Regija 5	76	150

Slika 167: Število volilcev, ki podpirajo stranko v posameznih regijah.

Imamo:

$$\begin{aligned} &H_0 : p_1 = p_2 = ... = p_5 \\ &H_1 : vsaj\ en\ p_i\ različen\ od\ ostalih \\ &\chi^2(\alpha, k-1) = \chi^2(0.05, 4) = 9.4877 \qquad \text{(chi2inv(1-0.05,4))} \end{aligned} \tag{8.94}$$

Izračunajmo empirične frekvence:

$$\begin{aligned} &f_{i1} = x_i, \quad i = 1,....,k \\ &f_{11} = x_1 = 123 \qquad n_1 = 250 \\ &f_{21} = x_2 = 92 \qquad n_2 = 180 \\ &f_{31} = x_3 = 163 \qquad n_3 = 300 \\ &f_{41} = x_4 = 112 \qquad n_4 = 200 \\ &f_{51} = x_5 = 76 \qquad n_5 = 150 \\ &f_{i2} = n_i - x_i, \quad i = 1,....,k \\ &f_{12} = n_1 - x_1 = 127 \\ &f_{22} = n_2 - x_2 = 88 \\ &f_{32} = n_3 - x_3 = 137 \\ &f_{42} = n_4 - x_4 = 88 \\ &f_{52} = n_5 - x_5 = 74 \end{aligned} \tag{8.95}$$

Ocena $\hat{p}$ je enaka:

$$\hat{p} = \frac{x_1 + ... + x_5}{n_1 + ... + n_5} = \frac{123+92+163+112+76}{250+180+300+200+150} = \frac{566}{1080} = 0.5241 \tag{8.96}$$

Sledi:

$$\begin{aligned}
&e_{i1} = n_i \cdot \hat{p} = n_i \cdot 0.5241, \quad i = 1,....,k \\
&e_{11} = n_1 \cdot \hat{p} = 250 \cdot 0.5241 = 131.025 \approx 131 \\
&e_{21} = n_2 \cdot \hat{p} = 180 \cdot 0.5241 = 94.339 \approx 94 \\
&e_{31} = n_3 \cdot \hat{p} = 300 \cdot 0.5241 = 157.23 \approx 157 \\
&e_{41} = n_4 \cdot \hat{p} = 200 \cdot 0.5241 = 104.82 \approx 105 \\
&e_{51} = n_5 \cdot \hat{p} = 150 \cdot 0.5241 = 78.615 \approx 79 \\
&e_{i2} = n_i - n_i \cdot \hat{p} = n_i \cdot (1 - \hat{p}) = n_i \cdot 0.4759, \quad i = 1,....,k \\
&e_{12} = n_1 \cdot 0.4759 = 250 \cdot 0.4759 = 118.975 \approx 119 \\
&e_{22} = n_2 \cdot 0.4759 = 180 \cdot 0.4759 = 85.662 \approx 86 \\
&e_{32} = n_3 \cdot 0.4759 = 300 \cdot 0.4759 = 142.77 \approx 143 \\
&e_{42} = n_4 \cdot 0.4759 = 200 \cdot 0.4759 = 95.18 \approx 95 \\
&e_{52} = n_5 \cdot 0.4759 = 150 \cdot 0.4759 = 71.385 \approx 71
\end{aligned} \tag{8.97}$$

Dobimo torej:

$$\begin{aligned}
&\chi^2 = \sum_{i=1}^{5}\sum_{j=1}^{2} \frac{(f_{ij} - e_{ij})^2}{e_{ij}} = \sum_{i=1}^{5}\left(\frac{(f_{i1} - e_{i1})^2}{e_{i1}} + \frac{(f_{i2} - e_{i2})^2}{e_{i2}} \right) = \\
&= \frac{(f_{11} - e_{11})^2}{e_{11}} + \frac{(f_{12} - e_{12})^2}{e_{12}} + \frac{(f_{21} - e_{21})^2}{e_{21}} + \frac{(f_{22} - e_{22})^2}{e_{22}} + \frac{(f_{31} - e_{31})^2}{e_{31}} + \frac{(f_{32} - e_{32})^2}{e_{32}} + \\
&+ \frac{(f_{41} - e_{41})^2}{e_{41}} + \frac{(f_{42} - e_{42})^2}{e_{42}} + \frac{(f_{51} - e_{51})^2}{e_{51}} + \frac{(f_{52} - e_{52})^2}{e_{52}} = \\
&= \frac{(123-131)^2}{131} + \frac{(127-119)^2}{119} + \frac{(92-94)^2}{94} + \frac{(88-86)^2}{86} + \frac{(163-157)^2}{157} + \frac{(137-143)^2}{143} \\
&+ \frac{(112-105)^2}{105} + \frac{(88-95)^2}{95} + \frac{(76-79)^2}{79} + \frac{(74-71)^2}{71}
\end{aligned} \tag{8.98}$$

Sledi:

$$\chi^2 = \frac{64}{131} + \frac{64}{119} + \frac{4}{94} + \frac{4}{86} + \frac{36}{157} + \frac{36}{143} + \frac{49}{105} + \frac{49}{95} + \frac{9}{79} + \frac{9}{71} = \\ = 2.8196 \tag{8.99}$$

Torej neenakost:

$$\chi^2 \geq \chi^2(\alpha, k-1) \\ 2.8196 \geq 9.4877 \tag{8.100}$$

ni izpolnjena (smo v območju zaupanja), zato ničelno hipotezo sprejmemo in sklepamo, da so deleži volilcev, ki bodo volili dotično stranko v vseh 5 regijah, enaki.

Pravkar prikazane izračune bi lahko izvedli tudi z naslednjim programom v Matlabu:

```
% vec_delezev.m (tabela k X 2)

clear
clc
close all

alfa = input('alfa=')
k = input('st_vrst k=')

fprvi = input('Vnesi prvi stolp')
fdrugi = input('Vnesi drugi stolp')

poc = sum(fprvi)/sum(fdrugi)

fij = [fprvi fdrugi-fprvi]

eprvi = fdrugi*poc
edrugi = fdrugi*(1-poc)

eij = [eprvi edrugi]

hi2krit = chi2inv(1-alfa,k-1)

hi2 = 0;

for i = 1:k
    for j = 1:2
    hi2 = hi2 + (fij(i,j)-eij(i,j))^2/eij(i,j);
    end
end

disp('hi2 je enak:')
hi2

if hi2 > hi2krit
   disp('Padli smo v kriticno obmocje - zavrzi nicelno hipotezo')
else
   disp('Padli smo v obmocje zaupanja - sprejmi nicelno hipotezo')
end
```

Izpis komandnega okna je naslednji:

```
alfa=0.05
alfa =
    0.0500
st_vrst k=5
k =
     5
Vnesi prvi stolp[123 92 163 112 76]'
fprvi =
   123
    92
   163
   112
    76
Vnesi drugi stolp[250 180 300 200 150]'
fdrugi =
   250
   180
   300
   200
   150
poc =
    0.5241
fij =
   123   127
    92    88
   163   137
   112    88
    76    74
eprvi =
  131.0185
   94.3333
  157.2222
  104.8148
   78.6111
edrugi =
  118.9815
   85.6667
  142.7778
   95.1852
   71.3889
eij =
  131.0185  118.9815
   94.3333   85.6667
  157.2222  142.7778
  104.8148   95.1852
   78.6111   71.3889
```

```
hi2krit =
   9.4877
hi2 je enak:
hi2 =
   2.8157
Padli smo v obmocje zaupanja - sprejmi nicelno hipotezo
```

8.6 Kontingenčne tabele

Velikokrat nas zanima, če sta določeni **opisni** spremenljivki *X* in *Y* v populaciji med seboj neodvisni, ali pa sta povezani. Populacija naj bo glede na spremenljivko *X* razdeljena na razrede $x_1, x_2, ..., x_k$, glede na spremenljivko *Y* pa razdeljena na razrede $y_1, y_2, ..., y_r$. Obe spremenljivki skupaj razčlenita populacijo na $k \cdot r$ razredov (x_i, y_j). Izberimo iz dotične populacije vzorec velikosti *n* in tudi njega klasificirajmo glede na spremenljivki *X* in *Y*. Označimo z f_{ij} frekvenco razreda (x_i, y_j) v vzorcu. Frekvence vseh razredov lahko zapišemo v takoimenovano **kontingenčno tabelo**, prikazano na sliki 168 [Košmelj K.].

	Y						
X	y_1	y_2	⋯	y_j	⋯	y_r	**Skupaj**
x_1	f_{11}	f_{12}		f_{1j}		f_{1r}	$f_{1\cdot}$
x_2	f_{21}	f_{22}		f_{2j}		f_{2r}	$f_{2\cdot}$
⋯							⋯
x_i	f_{i1}	f_{i2}		f_{ij}		f_{ir}	$f_{i\cdot}$
⋯							⋯
x_k	f_{k1}	f_{k2}		f_{kj}		f_{kr}	$f_{r\cdot}$
Skupaj	$f_{\cdot 1}$	$f_{\cdot 2}$	⋯	$f_{\cdot j}$	⋯	$f_{\cdot r}$	n

Slika 168: Kontingenčna tabela [Košmelj K.]

Pri tem velja:

$$f_{i\bullet} = \sum_{j=1}^{r} f_{ij}, \quad i = 1, ..., k \tag{8.101}$$

in:

$$f_{\bullet j} = \sum_{i=1}^{k} f_{ij}, \quad j = 1,...,r \tag{8.102}$$

Vsota vseh frekvenc je enaka številu obravnavanih enot v vzorcu [Jesenko, Košmelj K.]:

$$n = \sum_{i=1}^{k} \sum_{j=1}^{r} f_{ij} \tag{8.103}$$

Označimo s p_{ij} verjetnost, da bo neka enota pripadala *i-ti* vrsti in *j-tem* stolpu, s $p_{i\bullet}$ verjetnost, da bo neka enota pripadala *i-ti* vrsti, in s $p_{\bullet j}$ verjetnost, da bo neka enota pripadala *j-tem* stolpu. Ničelna hipoteza je [Jesenko, Košmelj K.]:

$$\begin{aligned} &H_0 : p_{ij} = p_{i\bullet} \cdot p_{\bullet j}, \;\; i = 1,...,k, \;\; j = 1,...,r \\ &(\textit{kar velja, če sta } X \textit{ in } Y \textit{ neodvisni}) \end{aligned} \tag{8.104}$$

nasprotna hipoteza pa je, da vsaj za en par indeksov *i* in *j* velja:

$$\begin{aligned} &H_1 : p_{ij} \neq p_{i\bullet} \cdot p_{\bullet j}, \;\; \textit{vsaj za en par } (i, j) \\ &(\textit{kar velja, če sta } X \textit{ in } Y \textit{ odvisni}) \end{aligned} \tag{8.105}$$

Definirajmo ocenjene verjetnosti [Jesenko]:

$$\begin{aligned} \hat{p}_{i\bullet} &= \frac{f_{i\bullet}}{n} \\ \hat{p}_{\bullet j} &= \frac{f_{\bullet j}}{n} \end{aligned} \tag{8.106}$$

Pri veljavnosti ničelne hipoteze bo za teoretične frekvence veljalo [Jesenko]:

$$\begin{aligned} &e_{ij} = \hat{p}_{i\bullet} \cdot \hat{p}_{\bullet j} \cdot n = \frac{f_{i\bullet}}{n} \cdot \frac{f_{\bullet j}}{n} \cdot n = \frac{f_{i\bullet} \cdot f_{\bullet j}}{n}, \\ &i = 1,2,...,k, \quad j = 1,2,...,r \end{aligned} \tag{8.107}$$

Kot se izkaže, se lahko o veljavnosti ničelne hipoteze lahko odločimo na osnovi naslednje testne (**Pearsonove** χ^2) statistike [Jesenko, Košmelj K.]:

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{r} \frac{\left(f_{ij} - e_{ij}\right)^2}{e_{ij}} \tag{8.108}$$

Če njena vrednost preseže določeno kritično vrednost, zavržemo ničelno hipotezo. Poglejmo si, kako dobimo stopnjo prostosti *SP* za kritično vrednost [Jesenko, Košmelj K.]:

$$\begin{aligned} &SP = \text{število neodvisnih podatkov} - \text{število parametrov} \\ &SP = \left(k \cdot r - 1\right) - \left[\left(k-1\right) + \left(r-1\right)\right] = \\ &= k \cdot r - 1 - k + 1 - r + 1 = k \cdot r - k - r + 1 = \\ &= k\left(r-1\right) - \left(r-1\right) = \left(r-1\right)\left(k-1\right) \end{aligned} \tag{8.109}$$

Veljavnost ničelne hipoteze torej testiramo z neenakostjo:

$$\chi^2 \geq \chi^2\left(\alpha, \left(k-1\right)\cdot\left(r-1\right)\right) \tag{8.110}$$

Statistika testa hipoteze, ki jo uporabimo tukaj, je le približno χ^2 s $\left(k-1\right)\cdot\left(r-1\right)$ stopnjami prostosti in zato ta test običajno uporabljamo le, če nobena od teoretičnih frekvenc ni manjša od 5. Če pa ta pogoj ni izpolnjen, združujemo vrstice ali stolpe, da dobimo večje frekvence, kar potem zmanjša število prostostnih stopenj [Jesenko].

Primer 8.16.:

Zanima nas povezava med inteligenčnim kvocientom (IQ) ljudi, ki so zaključili obsežen izobraževalen program v podjetju, in njihovo delovno uspešnostjo. Rezultate poizkusa prikazuje tabela na sliki 169. [Jesenko]. Testirajte ničelno hipotezo, da delovna uspešnost ni odvisna od velikosti IQ udeleženca izobraževalnega programa ($\alpha = 0.01$*).*

		Delovna uspešnost			
		Slaba	**Primerna**	**Dobra**	
IQ	**podpovprečen**	67	64	25	156
	povprečen	42	76	56	174
	nadpovprečen	10	23	37	70
		119	163	118	400

Slika 169: Kontingenčna tabela za primer ugotavljanja povezanosti med delovno uspešnostjo in IQ-jem. [Jesenko]

Imamo:

$$
\begin{aligned}
&k = 3 \\
&r = 3 \\
&\alpha = 0.01 \\
&f_{ij} \text{ odčitamo iz tabele} \\
&f_{i\bullet} = \sum_{j=1}^{3} f_{ij}, \quad i = 1,2,3 \\
&f_{\bullet j} = \sum_{i=1}^{3} f_{ij}, \quad j = 1,2,3 \\
&n = \sum_{i=1}^{3}\sum_{j=1}^{3} f_{ij} \\
&e_{ij} = \frac{f_{i\bullet} \cdot f_{\bullet j}}{n}, \quad i = 1,2,3, \quad j = 1,2,3 \\
&\chi^2 = \sum_{i=1}^{3}\sum_{j=1}^{3} \frac{\left(f_{ij} - e_{ij}\right)^2}{e_{ij}} \\
&\chi^2\left(0.01, 2\cdot 2\right) = \chi^2\left(0.01, 4\right) = 13.2767
\end{aligned}
\tag{8.111}
$$

Hipotezi sta:

$$
\begin{aligned}
&H_0 \ldots\ldots\ldots\ldots \textit{delovna neuspešnost neodvisna od IQ} \\
&H_1 \ldots\ldots\ldots\ldots \textit{delovna neuspešnost odvisna od IQ}
\end{aligned}
\tag{8.112}
$$

Za izračune uporabimo naslednji program v Matlabu:

```
% konting.m (kontingencna tabela)

clear
clc
close all

alfa = input('alfa=')
k = input('st_vrst k=')
r = input('st_stolpov r=')

F = input('Vnesi matriko empiricnih frekvenc')

for i = 1:k
   fipika(i) = sum(F(i,:));
end

fipika

for j = 1:r
   fjpika(j) = sum(F(:,j));
end

fjpika

n = sum(fipika)

hi2krit = chi2inv(1-alfa,(k-1)*(r-1))

hi2 = 0;

eij = fipika'*fjpika/n

for i = 1:k
   for j = 1:r
   hi2 = hi2 + (F(i,j)-eij(i,j))^2/eij(i,j);
   end
end

disp('hi2 je enak:')
hi2

if hi2 > hi2krit
  disp('Padli smo v kriticno obmocje - zavrzi nicelno hipotezo')
else
  disp('Padli smo v obmocje zaupanja - sprejmi nicelno hipotezo')
end
```

Izpis komandnega okna je naslednji:

```
alfa=0.01
alfa =
   0.0100
st_vrst k=3
k =
    3
st_stolpov r=3
r =
    3

Vnesi matriko empiricnih frekvenc[67 64 25;42 76 56;10 23 37]
F =
   67   64   25
   42   76   56
   10   23   37

fipika =
```

```
   156   174    70

fjpika =
   119   163   118

n =
   400

hi2krit =
   13.2767

eij =
   46.4100   63.5700   46.0200
   51.7650   70.9050   51.3300
   20.8250   28.5250   20.6500

hi2 je enak:
hi2 =
   41.0143

Padli smo v kriticno obmocje - zavrzi nicelno hipotezo
```

Dobimo torej rezultate:

$$
\begin{aligned}
&f_{i\bullet} = \begin{bmatrix} 156 & 174 & 70 \end{bmatrix}^T \\
&f_{\bullet j} = \begin{bmatrix} 119 & 163 & 118 \end{bmatrix} \\
&n = 400 \\
&e_{ij} = \begin{bmatrix} 46.41 & 63.57 & 46.02 \\ 51.76 & 70.905 & 51.33 \\ 20.825 & 28.525 & 20.65 \end{bmatrix} \\
&\chi^2 = 41.0143 \\
&\chi^2\left(0.01, 4\right) = \chi^2_{krit} = 13.2767
\end{aligned}
\tag{8.113}
$$

Ker je $\chi^2 = 41.0143 > \chi^2_{krit} = 13.2767$, moramo ničelno hipotezo zavreči. Torej velja nasprotna hipoteza, da je delovna uspešnost odvisna od IQ-ja udeleženca izobraževalnega programa, oz. sta obe opisni spremenljivki med seboj povezani.

8.7 Prilagoditveni test

Procedure testiranja hipotez, ki smo jih do sedaj spoznali, so bile načrtovane za probleme, kjer je bila verjetnostna porazdelitev populacije poznana. Ali je predpostavka o privzeti porazdelitvi sprejemljiva, pa se sploh nismo spraševali. Obstajajo torej tudi drugačne hipoteze, kjer verjetnostnih porazdelitev populacij ne poznamo [Montgomery 1]. Prilagoditveni test (Goodness of Fit Test) uporabljamo, ko želimo ugotoviti, če vzorec določene velikosti pripada populaciji, ki ima predpisan porazdelitveni zakon. Torej nas zanima, kako na osnovi zbranih podatkov preverjamo domnevo o verjetnostni porazdelitvi [Košmelj K.].

Predpostavimo najprej, da ima populacija diskreten porazdelitven zakon $p(x_i) = p_i$, kjer so x_i vrednosti statistične spremenljivke populacije. Iz slednje izberemo vzorec velikosti N in dobljene vrednosti zapišemo v obliki empirične frekvenčne porazdelitve (glej sliko 170).

Vrednost	x_1	x_2	…	x_i	…	x_n
Frekvenca	f_1	f_2	…	f_i	…	f_n

Slika 170: Frekvenčna porazdelitev empiričnih frekvenc

Vsota vseh empiričnih frekvenc iz slike 170 je:

$$N = \sum_{i=1}^{n} f_i \tag{8.114}$$

Vsaki vrednosti x_i priredimo teoretično frekvenco na naslednji način [Jesenko]:

$$e_i = N \cdot p(x_i),\ i = 1,...,n \tag{8.115}$$

Nobena od teoretičnih frekvenc ne sme biti manjša od 5, sicer test ni dovolj natančen. Če pa ta pogoj ni izpolnjen, združimo dve ali več vrednosti v eno. Ničelna hipoteza je, da ima

populacija, iz katere vzorec izhaja, porazdelitven zakon $p(x_i) = p_i$, torej da velja [Košmelj K.]:

$$H_0 : X \in \begin{bmatrix} x_1 & x_2 & \dots & x_n \\ p(x_1) & p(x_2) & \dots & p(x_n) \end{bmatrix} \tag{8.116}$$

Alternativna hipoteza pa je:

$$H_1 : X \notin \begin{bmatrix} x_1 & x_2 & \dots & x_n \\ p(x_1) & p(x_2) & \dots & p(x_n) \end{bmatrix} \tag{8.117}$$

torej da dejanska porazdelitev populacije ni enaka predpostavljeni teoretični porazdelitvi. Kot se izkaže, se lahko o veljavnosti ničelne hipoteze lahko odločimo na osnovi naslednje testne (**Pearsonove** χ^2) statistike [Jesenko, Košmelj K.]:

$$\chi^2 = \sum_{i=1}^{n} \frac{(f_i - e_i)^2}{e_i} \tag{8.118}$$

Ideja statističnega preizkusa je naslednja. Če se dejanske empirične in pričakovane teoretične frekvence med seboj dovolj dobro ujemajo, ničelno hipotezo sprejmemo, sicer pa jo zavržemo [Košmelj K.]. Mera ujemanja, kot je razvidno, temelji na kvadratu razlike $(f_i - e_i)^2$. Bolj ko se dejanske empirične in pričakovane teoretične frekvence med seboj razlikujejo, večja je vrednost χ^2 in bolj je verjetno, da bo prekoračila kritično vrednost. Če je ničelna hipoteza pravilna, je Pearsonova χ^2 statistika porazdeljena približno po χ^2 porazdelitvi, kjer je število stopenj naslednje:

$$\begin{aligned} &SP = število\ neodvisnih\ \ podatkov - število\ \ parametrov \\ &SP = (n-1) - t \end{aligned} \tag{8.119}$$

kjer število parametrov zavisi od predpostavljene teoretične porazdelitve. Veljavnost ničelne hipoteze torej testiramo z neenakostjo:

$$\chi^2 \geq \chi^2_{krit} = \chi^2(\alpha, n-1-t) \tag{8.120}$$

Če vrednost χ^2 preseže kritično vrednost, ničelno hipotezo zavržemo, sicer pa jo sprejmemo.

Praktični primeri tega testa, ki nas npr. zanimajo:

- Ali je število tipkarskih napak na eni tipkani strani porazdeljeno po Poissonovi porazdelitvi,
- Ali je rojevanje mladičev v nekem gnezdu porazdeljeno po Poissonovi porazdelitvi,
- Ali je inteligenčni kvocient ljudi porazdeljen po normalni porazdelitvi,
- Ali je masa nekega izdelka v proizvodnji porazdeljena po normalni porazdelitvi, itn.

Primer 8.17.:

Imamo frekvenčno porazdelitev števila napak na eni strani, ki ga prikazuje slika 171. Ugotovite, če je število napak na eni strani Poissonova naključna spremenljivka [Jesenko]. Stopnja pomembnosti je 0.05.

Število napak (x_i)	Število strani (f_i)
0	15
1	46
2	89
3	98
4	71
5	38
6	21
7	9
8	3
9	1

Slika 171: Frekvenčna porazdelitev števila napak na eni strani [Jesenko]

Imamo:

$$
\begin{aligned}
&n = 10 \\
&\alpha = 0.05 \\
&t = 1 \\
&f_i \text{ odčitamo iz tabele} \\
&N = \sum_{i=1}^{10} f_i \\
&e_i = N \cdot p\left(x_i\right),\ i = 1,...,10 \\
&p\left(x_i\right) = \frac{\hat{\lambda}^{x_i} \cdot e^{-\hat{\lambda}}}{x_i!},\ i = 1,...,10 \\
&\chi^2 = \sum_{i=1}^{10} \frac{\left(f_i - e_i\right)^2}{e_i} \\
&\chi^2_{krit} = \chi^2\left(\alpha, n-1-t\right) = \chi^2\left(0.05, n-1-1\right) = \chi^2\left(0.05, 8\right)
\end{aligned}
\tag{8.121}
$$

Hipotezi sta:

$$
\begin{aligned}
H_0 &: X \in Poisson\left(\hat{\lambda}\right) \\
H_1 &: X \notin Poisson\left(\hat{\lambda}\right)
\end{aligned}
\tag{8.122}
$$

Parameter $\hat{\lambda}$ izračunamo iz podatkov na naslednji način [Jesenko]:

$$
\hat{\lambda} = \frac{\sum_{i=1}^{n} f_i \cdot x_i}{N} = \frac{\sum_{i=1}^{n} f_i \cdot x_i}{\sum_{i=1}^{n} f_i} = \frac{\sum_{i=1}^{10} f_i \cdot x_i}{\sum_{i=1}^{10} f_i} = ... = 3.1049
\tag{8.123}
$$

pri čemer je:

$N = ... = 391$

Teoretične frekvence izračunamo z izrazom: $e_i = N \cdot p\left(x_i\right),\ i = 1,...,10,$ kjer je $p\left(x_i\right) = \frac{\hat{\lambda}^{x_i} \cdot e^{-\hat{\lambda}}}{x_i!},\ i = 1,...,10$. Zapisane so v tabeli na sliki 172 [Jesenko].

$p(x_i)$	e_i
0,04483	17,52885
0,13919	54,42462
0,21609	84,4904
0,22364	87,4436
0,17359	67,87502
0,1078	42,14848
0,05578	21,81085
0,02474	9,67423
0,0096	3,75464
0,00331	1,29529

Slika 172: Teoretične verjetnosti $p(x_i)$ in teoretične frekvence e_i [Jesenko]

Ker sta zadnji dve teoretični frekvenci manjši od 5, bi ju lahko združili. Seveda bi morali potem združiti tudi empirični frekvenci. Na ta način bi sicer izgubili eno prostostno stopnjo, vendar pa bi bil test bolj zanesljiv [Jesenko].

Pri izračunih dobimo še:

$$\chi^2 = \sum_{i=1}^{10} \frac{(f_i - e_i)^2}{e_i} = \ldots = 4.0323 \qquad (8.124)$$
$$\chi^2_{krit} = \chi^2(0.05, 8) = 15.5073$$

Za izračune smo uporabili naslednji program v Matlabu:

```
% goodnes.m (ugotavljanje tipa diskretne porazdelitve)

clear
clc
close all
ch = 0;
ch1=0;

alfa = input('alfa=')

fij = input('Vnesi stolp empiricnih frekvenc')
xi = input('Vnesi vektor vrednosti xi')

N = sum(fij)
n = length(xi)

ch = input('poisson(1), enakomerna(2), poljubna(3)');

if ch == 1
  f = 'par^i*exp(-par)/factorial(i)'
elseif ch == 2
  n1 = input('stevilo enot v vzorcu')
  pteo = 1/n
  f = 'n1*pteo'
else
  f = input('vnesi vektor teoreticnih verjetnosti')
end

if ch == 1
   t = input('stevilo parametrov t =')
else
   t = 0;
end

if ch == 1        % poisson
  par = input('par= (enter za izracun iz podatkov)')
```

```
  if length(par) == 0
    ch1 = input('iz podatkov-1, iz frekv.porazdelitve-2');
    if ch1 == 1
      pod = input('vnesi vektor podatkov pod=')
      par = sum(pod)/n        % ocenj. lambda z max.likeli pri danih stevilskih podatkih
      st_raz=input('st_raz=') % stevilo razredov
      clear xi
      xi=0:1:st_raz-1
    else
      par = fij*xi'/N       % ocenj.lambda iz podatkov pri dani empir.frekv.porazd.
    end
  end
end

if ch1 == 1
  for i=xi
    f1(i+1)=eval(f);
  end
end

eij = [];

for i=xi
  if ch == 2         % enakomerna
    f1=eval(f);
    eij = [eij f1];
  elseif ch == 1      % poissonova
    if ch1 == 1      % iz stevilskih podatkov - razbijemo na razrede
      if (i == 0)||(i==1)
        eij = [eij f1(i+1)];         % primer nesrec - glej ross, str. 495
      elseif i == 2
        eij = [eij f1(i+1)+f1(i+2)];
      elseif i == 3
        eij = [eij f1(i+2)];
      else
        eij = [eij 1-sum(eij)];
      end
    else         % iz frekvencne porazdelitve
      f1=eval(f);
      eij = [eij N*f1];
    end
  else           % poljubna
    eij = [eij N*f(i)];
  end
end

if ch1 == 1    % poissonova iz stevilskih podatkov
  eij = N*eij
  disp('teoreticne verjetnosti so:')
  eij/N
else
  eij
end
```

```
hi2krit = chi2inv(1-alfa,length(fij)-t-1)

hi2 = 0;

for i = 1:length(fij)
   hi2 = hi2 + (fij(i)-eij(i))^2/eij(i);
end

disp('hi2 je enak:')
hi2

if hi2 > hi2krit
  disp('Padli smo v kriticno obmocje - zavrzi nicelno hipotezo')
else
  disp('Padli smo v obmocje zaupanja - sprejmi nicelno hipotezo')
end
```

Izpis komandnega okna je naslednji:

```
alfa=0.05
alfa =
   0.0500

Vnesi stolp empiricnih frekvenc[15 46 89 98 71 38 21 9 3 1]
fij =
   15   46   89   98   71   38   21    9    3    1

Vnesi vektor vrednosti xi0:9
xi =
    0    1    2    3    4    5    6    7    8    9
N =
  391
n =
   10

poisson(1), enakomerna(2), poljubna(3)1
f =
par^i*exp(-par)/factorial(i)

stevilo parametrov t =1
t =
    1

par= (enter za izracun iz podatkov)
par =
    []
iz podatkov-1, iz frekv.porazdelitve-22
par =
   3.1049
```

```
eij =
  17.5289  54.4246  84.4904  87.4436  67.8750  42.1485  21.8108   9.6742   3.7546   1.2953

hi2krit =
  15.5073

hi2 je enak:
hi2 =
   4.0323

Padli smo v obmocje zaupanja - sprejmi nicelno hipotezo
>>
```

Torej, ker je $\chi^2 = 4.0323 < \chi^2_{krit} = 15.5073$, smo v območju zaupanja in ničelno hipotezo sprejmemo. Zaključimo lahko, da je število napak, ki jih je napravila tipkarica na eni strani, Poissonova naključna spremenljivka.

Primer 8.18.:

Zanima nas, če so nesreče v nekem industrijskem obratu porazdeljene po Poissonovi naključni spremenljivki. Denimo, da imamo tedensko število nesreč preko 30 tednov, kot ga prikazuje slika 173 (spremenljivka $y_i, i = 1,...,30$) [Ross 1]. Stopnja pomembnosti je 0.05.

8	0	0	1	3	4	0	2	12	5
1	8	0	2	0	1	9	3	4	5
3	3	4	7	4	0	1	2	1	2

Slika 173: Tedensko število nesreč preko 30 tednov [Ross 1]

Testirajte hipotezo, da so nesreče porazdeljene po Poissonovi naključni spremenljivki [Ross 1]!

Hipotezi sta:

$$H_0 : Y \in Poisson\left(\hat{\lambda}\right)$$
$$H_1 : Y \notin Poisson\left(\hat{\lambda}\right) \qquad (8.125)$$

Dane vrednosti bomo opazovali v nekem številu razredov, denimo jih vzamemo pet. Potem lahko rečemo, da je izid števila nesreč v določenem dnevu v 1. razredu, če ni nič

nesreč, v 2. razredu, če je ena nesreča, v 3. razredu, če sta dve ali tri nesreče, v 4. razredu, če so 4 nesreče, in v 5. razredu, če je več kot 4 nesreč. Torej, če je porazdelitev res Poissonova s srednjo vrednostjo λ, potem velja [Ross 1]:

$$\begin{aligned} p_1 &= P(Y=0) = e^{-\lambda} \\ p_2 &= P(Y=1) = \frac{\lambda^1 \cdot e^{-\lambda}}{1!} = \lambda \cdot e^{-\lambda} \\ p_3 &= P(Y=2) + P(Y=3) = \frac{\lambda^2 \cdot e^{-\lambda}}{2!} + \frac{\lambda^3 \cdot e^{-\lambda}}{3!} \\ p_4 &= P(Y=4) = \frac{\lambda^4 \cdot e^{-\lambda}}{4!} \\ p_5 &= 1 - (p_1 + p_2 + p_3 + p_4) \end{aligned} \tag{8.126}$$

Ker bomo tudi dane podatke združevali v razrede, dobimo na osnovi slike 173 naslednjo empirično frekvenčno razredno porazdelitev:

$$[x_i \quad f_i] = \begin{bmatrix} 1 & 6\ ničel \\ 2 & 5\ enk \\ 3 & 8\ dvojk\ ali\ trojk \\ 4 & 4\ štirke \\ 5 & 7\ jih\ je\ več\ kot\ štirke \end{bmatrix} \tag{8.127}$$

Imamo še:

$$n_{y_i} = 30$$
$$\alpha = 0.05$$
$$t = 1$$
$$f_i \text{ odčitamo iz tabele}$$
$$N = \sum_{i=1}^{5} f_i = 6+5+8+4+7 = 30$$
$$e_i = N \cdot \begin{cases} \hat{p}(x_1) = \hat{p}(0) \\ \hat{p}(x_2) = \hat{p}(1) \\ \hat{p}(x_3) + \hat{p}(x_4) = \hat{p}(2) + \hat{p}(3) \\ \hat{p}(x_5) = \hat{p}(4) \\ 1 - \sum_{i=0}^{4} \hat{p}(i) \end{cases} = N \cdot \begin{cases} \hat{p}_1 \\ \hat{p}_2 \\ \hat{p}_3 \\ \hat{p}_4 \\ \hat{p}_5 \end{cases} \tag{8.128}$$
$$\hat{p}(x_i) = \frac{\hat{\lambda}^{x_i} \cdot e^{-\hat{\lambda}}}{x_i!} = \frac{\hat{\lambda}^{j} \cdot e^{-\hat{\lambda}}}{j!} = \hat{p}(j),\ i = 1,...,5,\ j = 0,...,4$$
$$\chi^2 = \sum_{i=1}^{5} \frac{(f_i - e_i)^2}{e_i}$$
$$\chi^2_{krit} = \chi^2(\alpha, n-1-t) = \chi^2(0.05, 5-1-1) = \chi^2(0.05, 3)$$

Izračunati moramo tudi parameter $\hat{\lambda}$ iz podatkov v tabeli na sliki 173, kar storimo na osnovi ocene z metodo največjega verjetja [Ross 1]:

$$\hat{\lambda} = \frac{\sum_{i=1}^{n_{y_i}} y_i}{n_{y_i}} = \frac{95}{30} = 3.1667 \tag{8.129}$$

Ocenjene teoretične verjetnosti so enake:

$$\begin{aligned} \hat{p}_1 &= e^{-\hat{\lambda}} = 0.0421 \\ \hat{p}_2 &= \hat{\lambda} \cdot e^{-\hat{\lambda}} = 0.1335 \\ \hat{p}_3 &= \frac{\hat{\lambda}^2 \cdot e^{-\hat{\lambda}}}{2!} + \frac{\hat{\lambda}^3 \cdot e^{-\hat{\lambda}}}{3!} = 0.4343 \\ \hat{p}_4 &= \frac{\hat{\lambda}^4 \cdot e^{-\hat{\lambda}}}{4!} = 0.1766 \\ \hat{p}_5 &= 1 - (\hat{p}_1 + \hat{p}_2 + \hat{p}_3 + \hat{p}_4) = 0.2135 \end{aligned} \tag{8.130}$$

Ocenjene teoretične frekvence so enake:

$$e_i = N \cdot \hat{p}_i, i = 1,2,3,4,5$$

$$\begin{aligned} e_1 &= 30 \cdot 0.0421 = 1.2643 \\ e_2 &= 30 \cdot 0.1335 = 4.0037 \\ e_3 &= 30 \cdot 0.4343 = 13.0304 \\ e_4 &= 30 \cdot 0.1766 = 5.2973 \\ e_5 &= 30 \cdot 0.2135 = 6.4043 \end{aligned} \tag{8.131}$$

Dobimo:

$$\begin{aligned} \chi^2 &= \sum_{i=1}^{5} \frac{(f_i - e_i)^2}{e_i} = \ldots = 20.3013 \\ \chi^2_{krit} &= \chi^2(0.05,3) = 7.8147 \end{aligned} \tag{8.132}$$

Ker pademo v kritično območje, moramo ničelno hipotezo zavrniti, torej ne moremo predpostaviti, da so nesreče porazdeljene po Poissonovi naključni spremenljivki. Razlog je v tem, da je bilo preveč tednov, ko se ni zgodilo nič nesreč, da bi lahko ničelna hipoteza držala [Ross 1].

Tudi pri izračunih za ta primer smo si pomagali s programom **goodnes.m**. Izpis komandnega okna je bil naslednji:

```
alfa=0.05
alfa =
   0.0500

Vnesi stolp empiricnih frekvenc[6 5 8 4 7]
fij =
    6    5    8    4    7

Vnesi vektor vrednosti xi0:29
xi =
  Columns 1 through 17
    0    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16
  Columns 18 through 30
   17   18   19   20   21   22   23   24   25   26   27   28   29

N =
   30
n =
   30
```

```
poisson(1), enakomerna(2), poljubna(3)1
f =
par^i*exp(-par)/factorial(i)

stevilo parametrov t =1
t =
     1
par= (enter za izracun iz podatkov)
par =
     []
iz podatkov-1, iz frekv.porazdelitve-21

vnesi vektor podatkov pod=[8 0 0 1 3 4 0 2 12 5 1 8 0 2 0 1 9 3 4 5 3 3 4 7 4 0 1 2 1 2]
pod =
  Columns 1 through 16
    8    0    0    1    3    4    0    2   12    5    1    8    0    2    0    1
  Columns 17 through 30
    9    3    4    5    3    3    4    7    4    0    1    2    1    2

par =
    3.1667

st_raz=5
st_raz =
     5

xi =
     0     1     2     3     4

eij =
    1.2643    4.0037   13.0304    5.2973    6.4043
teoreticne verjetnosti so:
ans =
    0.0421    0.1335    0.4343    0.1766    0.2135

hi2krit =
    7.8147

hi2 je enak:
hi2 =
   20.3013

Padli smo v kriticno obmocje - zavrzi nicelno hipotezo
```

Poglejmo si, kaj bi se spremenilo, če bi izraz (8.126) imel obliko:

$$
\begin{aligned}
p_1 &= P(Y=0) = e^{-\lambda} \\
p_2 &= P(Y=1) = \frac{\lambda^1 \cdot e^{-\lambda}}{1!} = \lambda \cdot e^{-\lambda} \\
p_3 &= P(Y=2) + P(Y=3) = \frac{\lambda^2 \cdot e^{-\lambda}}{2!} + \frac{\lambda^3 \cdot e^{-\lambda}}{3!} \\
p_4 &= P(Y=4) = \frac{\lambda^4 \cdot e^{-\lambda}}{4!} + \frac{\lambda^5 \cdot e^{-\lambda}}{5!} \\
p_5 &= 1 - (p_1 + p_2 + p_3 + p_4)
\end{aligned}
\tag{8.133}
$$

torej, da bi v 4. razredu obravnavali 4 ali 5 nesreč. Tokrat bi na osnovi slike 173 dobili naslednjo empirično frekvenčno razredno porazdelitev:

$$
[x_i \quad f_i] = \begin{bmatrix} 1 & 6\ ničel \\ 2 & 5\ enk \\ 3 & 8\ dvojk\ ali\ trojk \\ 4 & 6\ štirk\ ali\ petk \\ 5 & 5\ jih\ je\ več\ kot\ petk \end{bmatrix}
\tag{8.134}
$$

Ocenjene teoretične verjetnosti bi bile enake:

$$
\begin{aligned}
\hat{p}_1 &= e^{-\hat{\lambda}} = 0.0421 \\
\hat{p}_2 &= \hat{\lambda} \cdot e^{-\hat{\lambda}} = 0.1335 \\
\hat{p}_3 &= \frac{\hat{\lambda}^2 \cdot e^{-\hat{\lambda}}}{2!} + \frac{\hat{\lambda}^3 \cdot e^{-\hat{\lambda}}}{3!} = 0.4343 \\
\hat{p}_4 &= \frac{\hat{\lambda}^4 \cdot e^{-\hat{\lambda}}}{4!} + \frac{\hat{\lambda}^5 \cdot e^{-\hat{\lambda}}}{5!} = 0.288 \\
\hat{p}_5 &= 1 - (\hat{p}_1 + \hat{p}_2 + \hat{p}_3 + \hat{p}_4) = 0.1016
\end{aligned}
\tag{8.135}
$$

Ocenjene teoretične frekvence so enake:

$$
e_i = N \cdot \hat{p}_i, i = 1,2,3,4,5
$$

$$
\begin{aligned}
e_1 &= 30 \cdot 0.0421 = 1.2643 \\
e_2 &= 30 \cdot 0.1335 = 4.0037 \\
e_3 &= 30 \cdot 0.4343 = 13.0304 \\
e_4 &= 30 \cdot 0.2884 = 8.6522 \\
e_5 &= 30 \cdot 0.1016 = 3.0493
\end{aligned}
\tag{8.136}
$$

Dobimo:

$$\chi^2 = \sum_{i=1}^{5} \frac{\left(f_i - e_i\right)^2}{e_i} = \ldots = 21.989$$
$$\chi^2_{krit} = \chi^2\left(0.05, 3\right) = 7.8147 \qquad (8.137)$$

Tudi tokrat velja, da moramo, ker pademo v kritično območje, ničelno hipotezo zavrniti, torej ne moremo predpostaviti, da so nesreče porazdeljene po Poissonovi naključni spremenljivki.

Za potrebe izračunov smo napisali rahlo modificiran program v Matlabu:

```
% goodnes1.m (ross, str.495)
clear
clc
close all

alfa = input('alfa=')
fij = input('Vnesi stolp empiricnih frekvenc')

N = sum(fij)
n = 30
f = 'par^i*exp(-par)/factorial(i)'
t = 1;
pod = [8 0 0 1 3 4 0 2 12 5 1 8 0 2 0 1 9 3 4 5 3 3 4 7 4 0 1 2 1 2]

par = sum(pod)/n
st_raz=5

xi=0:1:st_raz
for i=xi
   f1(i+1)=eval(f);
end
```

```
eij = [];
for i=xi
   if i == st_raz
      break
   end
   if (i == 0)||(i==1)
      eij = [eij f1(i+1)];          % primer nesrec - glej ross, str. 495
   elseif i == 2
      eij = [eij f1(i+1)+f1(i+2)];
   elseif i == 3
      eij = [eij f1(i+2)+f1(i+3)];
   else
      eij = [eij 1-sum(eij)];
   end
end

eij = N*eij

disp('teoreticne verjetnosti so:')
eij/N

hi2krit = chi2inv(1-alfa,length(fij)-t-1)




hi2 = 0;
for i = 1:length(fij)
   hi2 = hi2 + (fij(i)-eij(i))^2/eij(i);
end

disp('hi2 je enak:')
hi2

if hi2 > hi2krit
  disp('Padli smo v kriticno obmocje - zavrzi nicelno hipotezo')
else
  disp('Padli smo v obmocje zaupanja - sprejmi nicelno hipotezo')
end
```

Izpis komandnega okna pa je naslednji:

```
alfa=0.05
alfa =
   0.0500

Vnesi stolp empiricnih frekvenc[6 5 8 6 5]
fij =
    6    5    8    6    5
```

```
N =
   30
n =
   30
f =
par^i*exp(-par)/factorial(i)

pod =
  Columns 1 through 17
    8    0    0    1    3    4    0    2   12    5    1    8    0    2    0    1    9
  Columns 18 through 30
    3    4    5    3    3    4    7    4    0    1    2    1    2

par =
   3.1667

st_raz =
    5

xi =
    0    1    2    3    4    5

eij =
   1.2643    4.0037   13.0304    8.6522    3.0493



teoreticne verjetnosti so:
ans =
   0.0421    0.1335    0.4343    0.2884    0.1016

hi2krit =
   7.8147

hi2 je enak:
hi2 =
  21.9890

Padli smo v kriticno obmocje - zavrzi nicelno hipotezo
```

Primer 8.19.:

V določenem časovnem obdobju je bilo 500 nesreč pri delu, od tega 130 ob ponedeljkih, 90 ob torkih, 100 ob sredah, 90 ob četrtkih in 90 ob petkih. Na podlagi teh podatkov bomo preverili domnevo, da je porazdelitev nesreč pri delu po dnevih v tednu enakomerna ($\alpha = 0.05$) [Košmelj K.].

Hipotezi sta:

$$\begin{aligned} H_0 &: X \in Enakomerna\left(\hat{p}\right) \\ H_1 &: X \notin Enakomerna\left(\hat{p}\right) \end{aligned} \tag{8.138}$$

Če je porazdelitev res enakomerna, potem velja:

$$\begin{aligned} p_1 &= \frac{1}{5} = p_{pon} \\ p_2 &= \frac{1}{5} = p_{tor} \\ p_3 &= \frac{1}{5} = p_{sre} \\ p_4 &= \frac{1}{5} = p_{čet} \\ p_5 &= \frac{1}{5} = p_{pet} \\ p &= p_i = \frac{1}{5}, i = 1,2,3,4,5 \end{aligned} \tag{8.139}$$

Empirično frekvenčno porazdelitev zapišemo na naslednji način:

$$\begin{bmatrix} x_i & f_i \end{bmatrix} = \begin{bmatrix} 1 & 130 \\ 2 & 90 \\ 3 & 100 \\ 4 & 90 \\ 5 & 90 \end{bmatrix} \begin{matrix} pon \\ tor \\ sre \\ čet \\ pet \end{matrix} \tag{8.140}$$

Imamo še:

$$
\begin{aligned}
&n = 5 \\
&\alpha = 0.05 \\
&t = 0 \\
&f_i \text{ odčitamo iz izraza } (8.140) \\
&N = \sum_{i=1}^{5} f_i = 500 \\
&p(x_i) = \hat{p}_i = \frac{1}{5},\ i = 1,...,5 \\
&e_i = N \cdot p(x_i),\ i = 1,...,5 \\
&e_i = 500 \cdot \frac{1}{5},\ i = 1,...,5 \\
&e_i = 100,\ i = 1,...,5 \\
&\chi^2 = \sum_{i=1}^{5} \frac{(f_i - e_i)^2}{e_i} \\
&\chi^2_{krit} = \chi^2(\alpha, n-1-t) = \chi^2(0.05, n-1-0) = \chi^2(0.05, 4) = 9.4877
\end{aligned}
\tag{8.141}
$$

Dobimo:

$$
\begin{aligned}
&\chi^2 = \sum_{i=1}^{5} \frac{(f_i - e_i)^2}{e_i} = \frac{(f_1 - e_1)^2}{e_1} + \frac{(f_2 - e_2)^2}{e_2} + \frac{(f_3 - e_3)^2}{e_3} + \frac{(f_4 - e_4)^2}{e_4} + \frac{(f_5 - e_5)^2}{e_5} = \\
&= \frac{(130-100)^2}{100} + \frac{(90-100)^2}{100} + \frac{(100-100)^2}{100} + \frac{(90-100)^2}{100} + \frac{(90-100)^2}{100} = \\
&= 9+1+0+1+1 = 12
\end{aligned}
\tag{8.142}
$$

Ker je $\chi^2 > \chi_{krit}{}^2$, zavrnemo ničelno hipotezo. Torej trdimo, da število nesreč ni enakomerno porazdeljeno po dnevih v tednu, glede na dani vzorec. Največja odstopanja so očitno v ponedeljek, torej podatki nakazujejo, da je ob ponedeljkih več nesreč, kot bi pričakovali pri enakomerni porazdelitvi.

Verjetnost *p* zaupanja v dobljeni rezultat glede na dani vzorec dobimo na naslednji način:

$$
\begin{aligned}
&p \approx P(\chi^2 > 12) = 1 - P(\chi^2 \le 12) = 1 - 0.9826 = 0.0174 \\
&(\text{1-chi2cdf(12,4)})
\end{aligned}
\tag{8.143}
$$

Pri izračunih si ponovno lahko pomagamo s programom **goodnes.m**. Izpis komandnega okna je naslednji:

```
alfa=0.05
alfa =
    0.0500

Vnesi stolp empiricnih frekvenc[130 90 100 90 90]
fij =
   130    90   100    90    90

Vnesi vektor vrednosti xi1:5
xi =
     1     2     3     4     5

N =
   500
n =
     5

poisson(1), enakomerna(2), poljubna(3)2
stevilo enot v vzorcu500
n1 =
   500

pteo =
    0.2000

f =
n1*pteo

eij =
   100   100   100   100   100

hi2krit =
    9.4877


hi2 je enak:
hi2 =
    12

Padli smo v kriticno obmocje - zavrzi nicelno hipotezo
```

Primer 8.20.:

200 pacientom, ki imajo želodčnega raka, je vzeta kri. Izkaže se, da jih ima 92 kri tipa A, 20 jih ima tip B, 4 imajo tip AB, in 84 jih ima tip O. So ti podatki dovolj signifikantni, da sprejmemo hipotezo da imajo bolniki enako porazdelitev tipov krvi kot celotna populacija (41% jih ima tip A, 9% ima tip B, 4% imajo tip AB, in 46% ima tip O) [Ross 2].

Hipotezi sta:

$$\begin{aligned} H_0 &: Velja \qquad p_1 = 0.41, p_2 = 0.09, p_3 = 0.04, p_4 = 0.46 \\ H_1 &: Ne\ velja \quad p_1 = 0.41, p_2 = 0.09, p_3 = 0.04, p_4 = 0.46 \end{aligned} \tag{8.144}$$

Empirično frekvenčno porazdelitev zapišemo na naslednji način:

$$\begin{bmatrix} x_i & f_i \end{bmatrix} = \begin{bmatrix} 1 & 92 \\ 2 & 20 \\ 3 & 4 \\ 4 & 84 \\ & \end{bmatrix} \begin{matrix} tip\ A \\ tip\ B \\ tip\ AB \\ tip\ O \\ \end{matrix} \tag{8.145}$$

Imamo še:

$$\begin{aligned} & n = 4 \\ & \alpha = 0.05 \\ & t = 0 \\ & f_i\ odčitamo\ iz\ izraza\ (8.145) \\ & N = \sum_{i=1}^{4} f_i = 200 \\ & p_1 = 0.41, p_2 = 0.09, p_3 = 0.04, p_4 = 0.46 \\ & e_i = N \cdot p_i,\ i = 1,...,4 \\ & e_i = 200 \cdot \begin{bmatrix} 0.41 & 0.09 & 0.04 & 0.46 \end{bmatrix} \\ & e_i = \begin{bmatrix} 82 & 18 & 8 & 92 \end{bmatrix} \\ & \chi^2 = \sum_{i=1}^{4} \frac{(f_i - e_i)^2}{e_i} \\ & \chi^2_{krit} = \chi^2(\alpha, n-1-t) = \chi^2(0.05, n-1-0) = \chi^2(0.05, 3) = 7.8147 \end{aligned} \tag{8.146}$$

Dobimo:

$$\begin{aligned} \chi^2 &= \sum_{i=1}^{4} \frac{(f_i - e_i)^2}{e_i} = \frac{(f_1 - e_1)^2}{e_1} + \frac{(f_2 - e_2)^2}{e_2} + \frac{(f_3 - e_3)^2}{e_3} + \frac{(f_4 - e_4)^2}{e_4} = \\ &= \frac{(92-82)^2}{82} + \frac{(20-18)^2}{18} + \frac{(4-8)^2}{8} + \frac{(84-92)^2}{92} = \\ &= \frac{100}{82} + \frac{4}{18} + \frac{16}{8} + \frac{64}{92} = 4.1374 \end{aligned} \tag{8.147}$$

Ker je $\chi^2 < \chi_{krit}{}^2$, sprejmemo ničelno hipotezo. Torej trdimo, da imajo bolniki enako porazdelitev tipov krvi kot celotna populacija. Tudi v tem primeru si lahko pomagamo s programom **goodnes.m**. Izpis komandnega okna je naslednji:

```
alfa=0.05
alfa =
   0.0500

Vnesi stolp empiricnih frekvenc[92 20 4 84]
fij =
   92   20    4   84

Vnesi vektor vrednosti xi1:4
xi =
    1    2    3    4

N =
  200
n =
    4

poisson(1), enakomerna(2), poljubna(3)3
vnesi vektor teoreticnih verjetnosti[0.41 0.09 0.04 0.46]
f =
   0.4100    0.0900    0.0400    0.4600

eij =
   82   18    8   92

hi2krit =
   7.8147

hi2 je enak:
hi2 =
   4.1374

Padli smo v obmocje zaupanja - sprejmi nicelno hipotezo
```

Primer 8.21.:

Imamo binomsko populacijo, za katero predpostavljamo, da je dejanski parameter enak $p = \frac{1}{3}$. *Iz populacije vzamemo vzorec velikosti 26306, pri čemer so podatki v obliki*

frekvenčne porazdelitve podani na sliki 174. Preverite ničelno hipotezo, da je $p = \frac{1}{3}$ *(*$\alpha = 0.05$*) [Vidakovic].*

x_i	f_i
0	185
1	1149
2	3265
3	5475
4	6114
5	5194
6	3067
7	1331
8	403
9	105
10	14
11	4
12	0

Slika 174: Podatki v obliki frekvenčne porazdelitve [Vidakovic] (i=1,…,13)

Hipotezi sta:

$$H_0 : p = \frac{1}{3} \qquad (8.148)$$
$$H_1 : p \neq \frac{1}{3}$$

Imamo še:

$$
\begin{aligned}
&n = 13 \\
&\alpha = 0.05 \\
&t = 0 \\
&f_i \text{ odčitamo iz tabele} \\
&N = \sum_{i=1}^{13} f_i = 26306 \\
&p_i = p(x_i) = \binom{n-1}{x_i} \cdot p^{x_i} \cdot (1-p)^{n-1-x_i}, x_i = 0,1,...,n-1, i = 1,2,...,13 \\
&p_i = p(x_i) = \binom{12}{x_i} \cdot \left(\frac{1}{3}\right)^{x_i} \cdot \left(1-\frac{1}{3}\right)^{12-x_i}, x_i = 0,1,...,12, i = 1,2,...,13 \\
&e_i = N \cdot p_i,\ i = 1,...,13
\end{aligned}
\tag{8.149}
$$

$$
\chi^2 = \sum_{i=1}^{13} \frac{(f_i - e_i)^2}{e_i}
$$

$$
\chi^2_{krit} = \chi^2(\alpha, n-1-t) = \chi^2(0.05, 13-1-0) = \chi^2(0.05, 12) = 21.0261
$$

Za izračune uporabimo naslednji program v Matlabu:

```
% goodnes2.m (vidakovic, str.510)

clear
clc
close all

alfa = input('alfa=')

fij = input('Vnesi stolp empiricnih frekvenc')

N = sum(fij)
n = 13
t = 0

disp('teoreticne verjetnosti so:')
f = binopdf(0:12,12,1/3)

disp('teroreticne frekvence so:')
eij = N*f

hi2krit = chi2inv(1-alfa,n-t-1)

hi2 = 0;
for i = 1:length(fij)
   hi2 = hi2 + (fij(i)-eij(i))^2/eij(i);
```

```
end

disp('hi2 je enak:')
hi2

if hi2 > hi2krit
  disp('Padli smo v kriticno obmocje - zavrzi nicelno hipotezo')
else
  disp('Padli smo v obmocje zaupanja - sprejmi nicelno hipotezo')
end

disp('verjetnost p iz podatkov je:')
pval= 1-chi2cdf(hi2,n-t-1)
```

Izpis komandnega okna je naslednji:

```
alfa=0.05
alfa =
   0.0500

Vnesi stolp empiricnih frekvenc[185 1149 3265 5475 6114 5194 3067 1331 403 105 14 4 0]
fij =
  Columns 1 through 8
       185       1149       3265       5475       6114       5194       3067       1331
  Columns 9 through 13
       403        105         14          4          0

N =
      26306
n =
    13
t =
     0

teoreticne verjetnosti so:
f =
  Columns 1 through 10
    0.0077    0.0462    0.1272    0.2120    0.2384    0.1908    0.1113    0.0477    0.0149    0.0033
  Columns 11 through 13
    0.0005    0.0000    0.0000

teroreticne frekvence so:
eij =
  1.0e+003 *
  Columns 1 through 10
    0.2027    1.2165    3.3454    5.5756    6.2726    5.0180    2.9272    1.2545    0.3920    0.0871
  Columns 11 through 13
    0.0131    0.0012    0.0000
```

```
hi2krit =
   21.0261

hi2 je enak:
hi2 =
   41.3122

Padli smo v kriticno obmocje - zavrzi nicelno hipotezo

verjetnost p iz podatkov je:
pval =
  4.3449e-005
```

Verjetnost *p* zaupanja v dobljeni rezultat glede na dani vzorec torej je:

$$p \approx P\left(\chi^2 > 41.3122\right) = 1 - P\left(\chi^2 \le 41.3122\right) = 4.34 \cdot 10^{-5} \tag{8.150}$$

Dobimo še:

$$\chi^2 = 41.3122 > \chi_{krit}{}^2 = 21.0261 \tag{8.151}$$

Ker smo padli v kritično območje, moramo ničelno hipotezo zavreči, torej dan vzorec ne potrdi hipoteze, da je $p = \frac{1}{3}$ od binomske populacije.

Prilagoditveni test pri zveznih naključnih spremenljivkah

Denimo, da imamo opravka z zveznim porazdelitvenim zakonom populacije $p(x)$. Podatke vzorca velikosti *N*, izbranega iz te populacije moramo zapisati v obliki tabele, prikazane na sliki 175. Teoretične frekvence $e_i, i = 1,...,n$ dobimo tako, da ploščine pod krivuljo $p(x)$ nad vsakim razredom pomnožimo z *N*. Potem je postopek testiranja enak kot v primeru diskretno porazdeljene populacije [Jesenko]. Izdelavo razredov lahko naredimo na več načinov. Nekateri statistiki priporočajo uporabo kvantilov (npr. kvartile), včasih pa so meje vsebinsko določene [Košmelj K.].

Razred	$r_1 - r_2$	$r_2 - r_3$	...	$r_{i-1} - r_i$	...	$r_{n-1} - r_n$
Frekvenca	f_1	f_2	...	f_i	...	f_n

Slika 175: Generacija frekvenčne porazdelitve pri zvezni populaciji

Primer 8.22.:

Zanima nas, ali lahko za maso določenega izdelka privzamemo normalno porazdelitev. V vzorcu je bilo N=300 izdelkov, njihove mase (v gramih g) pa so dane na sliki 176 [Košmelj K.].

498, 494, 501, 506, 506, 509, 489, 499, 505, 495, 497, 492. 491, 495, 496, 489, 497, 498, 501,
498, 498, 498, 507, 500, 499, 497, 510, 504, 512, 497, 508. 492, 503, 505, 510, 500, 497, 503,

498, 504, 493, 496, 492, 498, 500, 500. 498, 511, 491, 496, 487, 507, 494, 497, 504, 502, 504,
503, 493, 494, 503, 502, 495, 499, 501, 503, 501, 495, 509, 502, 500, 504, 504, 497, 495, 506,
494, 492, 504, 503, 511, 507, 507, 501, 500, 502, 500, 495, 491, 504, 502, 503, 501, 495, 506,
498, 496, 496, 498, 498, 497, 504, 503, 497, 507, 491, 503, 499, 500, 497, 497, 504, 504, 502,
503, 508, 502, 503, 509, 498, 505, 501, 506, 499, 496, 505, 497, 503, 503, 498, 498, 493, 510,
497, 500, 499, 514, 506, 504, 507, 501, 503, 499, 494, 506, 499, 493, 504, 504, 502, 502, 497,
504, 503, 505, 486, 502, 507, 491, 500, 507, 508, 501, 501, 509, 500, 497, 505, 500, 505, 502,
507, 499, 495, 492, 487, 501, 501, 500, 501, 501, 507, 500, 494, 493, 501, 505, 492, 497, 503,
499, 498, 499, 498, 496, 503, 500, 506, 501, 500, 501, 501, 500, 494, 501, 501, 504, 496, 503,
501, 501, 504, 509, 493, 494, 503, 506, 496, 489, 496, 501, 502, 495, 496, 502, 501, 497, 501,
495, 496, 505, 508, 497, 499, 501, 509, 503, 508, 501, 508, 497, 495, 506, 505, 491, 506, 499,
494, 491, 496, 502, 505, 500, 500, 501, 498, 505, 498, 512, 504, 503, 495, 504, 507, 496, 498,
493, 507, 505, 501, 489, 503, 503, 506, 495, 506, 497, 496, 491, 507, 495, 508, 494, 498, 493,
508, 500, 502, 505, 510, 500, 495, 497, 504, 504, 503, 493, 493, 499, 497.

Slika 176: Mase izdelkov [Košmelj K.]

Testirajte hipotezo, da so mase izdelkov (spremenljivka y_i) porazdeljene po normalni naključni spremenljivki! ($\alpha = 0.05$)

Hipotezi sta:

$$H_0 : Y \in N(\hat{\mu}, \hat{\sigma})$$
$$H_0 : Y \notin N(\hat{\mu}, \hat{\sigma}) \qquad (8.152)$$

Na osnovi podatkov vzorca bomo skušali ugotoviti, ali ima populacija, kateri vzorec pripada, normalno porazdelitev. Ker obeh parametrov ne poznamo, jih bomo skušali

oceniti iz danih podatkov. Tako lahko izračunamo s pomočjo Matlaba (program bo podan kasneje):

$$\begin{aligned}\hat{\mu} &= 500.2767\\ \hat{\sigma} &= 5.1739\\ y_{\min} &= 486\\ y_{\max} &= 514\end{aligned} \tag{8.153}$$

Dane vrednosti bomo opazovali v nekem številu razredov, denimo jih vzamemo devet (podatke razvrstimo v razrede širine 4g). Potem dobimo empirično frekvenčno porazdelitev, prikazano na sliki 177.

Masa (g)	Dejanska frekvenca
do 488	3
nad 488 do 492	18
nad 492 do 496	50
nad 496 do 500	77
nad 500 do 504	89
nad 504 do 508	48
nad 508 do 512	14
nad 512 do 516	1
nad 516	0
Skupaj	300

Slika 177: Empirična frekvenčna porazdelitev $f_i, i = 1,...,9$ za dano maso izdelkov iz vzorca [Košmelj K.]

Nato definiramo sredine razredov na naslednji način:

$$x_i = [486 \quad 490 \quad 494 \quad 498 \quad 502 \quad 506 \quad 510 \quad 514 \quad 518] \tag{8.154}$$

V njih izračunamo vrednost funkcije:

$$\begin{aligned} f(x_i) &= \frac{1}{\hat{\sigma}\cdot\sqrt{2\pi}}\cdot e^{-\frac{(x_i-\hat{\mu})^2}{2\hat{\sigma}^2}}\\ f'(x_i) &= \frac{f(x_i)}{\sum_{i=1}^{9} f(x_i)}\end{aligned} \tag{8.155}$$

Teoretične frekvence določimo z izrazom:

$$e_i = \left(\sum_{i=1}^{9} f_i \right) \cdot f'(x_i) = N \cdot f'(x_i) = 300 \cdot f'(x_i) \tag{8.156}$$

Dobimo:

$$e_i = \begin{bmatrix} 2.06 & 12.88 & 44.36 & 84.05 & 87.6 & 50.21 & 15.83 & 2.74 & 0.262 \end{bmatrix} \tag{8.157}$$

Tvorimo vrednost statistike:

$$\chi^2 = \sum_{i=1}^{9} \frac{(f_i - e_i)^2}{e_i} = \ldots = 5.4833 \tag{8.158}$$

in njeno kritično vrednost:

$$\chi^2_{krit} = \chi^2(\alpha, n-1-t) = \chi^2(0.05, 9-1-2) = \chi^2(0.05, 6) = 12.5916 \tag{8.159}$$

Ker je $\chi^2 < \chi^2_{krit}$, pademo v območje zaupanja in sprejmemo ničelno hipotezo, torej, da na osnovi danega vzorca lahko sklepamo, da je njegova populacija normalna. Verjetnost *p* zaupanja v dobljeni rezultat glede na dani vzorec je enaka:

$$\begin{aligned} &p \approx P(\chi^2 > 5.4833) = 1 - P(\chi^2 \le 5.4833) = 0.4835 \\ &(\text{1-chi2cdf(5.4833,6)}) \end{aligned} \tag{8.160}$$

Prikazali smo postopek, kot ga predlaga Trauth [Trauth]. Morda je težava prikazanih izračunov v tem, da so ponekod teoretične frekvence premajhne. Zato bomo v nadaljevanju razrede s premajhnimi teoretičnimi frekvencami združili. Potem dobimo frekvenčno porazdelitev, prikazano na sliki 178.

Masa (g)	Dejanske frekvence	Pričakovane Frekvence	Prispevek k χ^2-statistiki
do 492	21	16,45	1,26
nad 492 do 496	50	44,82	0,60
nad 496 do 500	77	82,33	0,35
nad 500 do 504	89	85,63	0,13
nad 504 do 508	48	50,44	0,12
nad 508	15	20,33	1,40
Skupaj	300	300,00	3,85

Slika 178: Empirična frekvenčna porazdelitev $f_i, i = 1,...,6$ *(dejanske frekvence) za dano maso izdelkov iz vzorca. Zraven so dodane tudi teoretične frekvence* $e_i, i = 1,...,6$ *(pričakovane frekvence), ki jih bomo izračunali. [Košmelj K.]*

Tokrat bomo teoretične frekvence izračunali s pomočjo kumulativne funkcije za normalno porazdelitev. Izračuni deležev kumulativne funkcije glede na razrede so naslednji:

$$\begin{aligned} F_1 &= F(492) \\ F_2 &= F(496) - F(492) \\ F_3 &= F(500) - F(496) \\ F_4 &= F(504) - F(500) \\ F_5 &= F(508) - F(504) \\ F_6 &= 1 - F(508) \end{aligned} \tag{8.160}$$

Teoretične frekvence pa določimo z izrazom:

$$e_i = N \cdot F_i = 300 \cdot F_i, i = 1,...,6 \tag{8.161}$$

in so prikazane v 2. stolpu tabele na sliki 178 (stolp pričakovane frekvence).

Tvorimo vrednost statistike:

$$\chi^2 = \sum_{i=1}^{6} \frac{(f_i - e_i)^2}{e_i} = ... = 3.8476 \tag{8.162}$$

in njeno kritično vrednost:

$$\chi^2_{krit} = \chi^2(\alpha, n-1-t) = \chi^2(0.05, 6-1-2) = \chi^2(0.05, 3) = 7.8147 \qquad (8.163)$$

Ker je $\chi^2 < \chi^2_{krit}$, pademo v območje zaupanja in sprejmemo ničelno hipotezo, torej, da na osnovi danega vzorca lahko sklepamo, da je njegova populacija normalna. Verjetnost *p* zaupanja v dobljeni rezultat glede na dani vzorec je enaka:

$$p \approx P(\chi^2 > 3.8476) = 1 - P(\chi^2 \le 3.8476) = 0.2784 \qquad (8.164)$$
$$(1\text{-chi2cdf}(3.8476,3))$$

V nadaljevanju prikažimo program v Matlabu, ki je opravil vse potrebne izračune:

```
% goodnes3.m (KOSMELJ, str.182) - za normalno porazdelitev

clear
clc
close all

% nalozimo podatke:

pod =  [498,    494,    501,    506,    506,    509,    489,    499,    505,    495,
497,    492,    491,    495,    496,    489,    497,    498,    501,...
498,    498,    498,    507,    500,    499,    497,    510,    504,    512,    497,
508,    492,    503,    505,    510,    500,    497,    503,    498,...
504,    493,    496,    492,    498,    500,    500,    498,    511,    491,    496,
487,    507,    494,    497,    504,    502,    504,...
503,    493,    494,    503,    502,    495,    499,    501,    503,    501,    495,
509,    502,    500,    504,    504,    497,    495,    506,...
494,    492,    504,    503,    511,    507,    507,    501,    500,    502,    500,
495,    491,    504,    502,    503,    501,    495,    506,...
498,    496,    496,    498,    498,    497,    504,    503,    497,    507,    491,
503,    499,    500,    497,    497,    504,    504,    502,...
503,    508,    502,    503,    509,    498,    505,    501,    506,    499,    496,
505,    497,    503,    503,    498,    498,    493,    510,...
497,    500,    499,    514,    506,    504,    507,    501,    503,    499,    494,
506,    499,    493,    504,    504,    502,    502,    497,...
504,    503,    505,    486,    502,    507,    491,    500,    507,    508,    501,
501,    509,    500,    497,    505,    500,    505,    502,...
507,    499,    495,    492,    487,    501,    501,    500,    501,    501,    507,
500,    494,    493,    501,    505,    492,    497,    503,...
499,    498,    499,    498,    496,    503,    500,    506,    501,    500,    501,
501,    500,    494,    501,    501,    504,    496,    503,...
501,    501,    504,    509,    493,    494,    503,    506,    496,    489,    496,
501,    502,    495,    496,    502,    501,    497,    501,...
495,    496,    505,    508,    497,    499,    501,    509,    503,    508,    501,
508,    497,    495,    506,    505,    491,    506,    499,...
494,    491,    496,    502,    505,    500,    500,    501,    498,    505,    498,
512,    504,    503,    495,    504,    507,    496,    498,...
493,    507,    505,    501,    489,    503,    503,    506,    495,    506,    497,
496,    491,    507,    495,    508,    494,    498,    493,...
508,    500,    502,    505,    510,    500,    495,    497,    504,    504,    503,
493,    493,    499,    497]


alfa = input('alfa=')

minpod = min(pod)
maxpod = max(pod)
```

```
srvr = mean(pod)    % ocena 1. parametra iz podatkov
stdev = std(pod)    % ocena 2. parametra iz podatkov

% dva parametra ocenjujemo iz podatkov (ker normalna porazdelitev)
t = 2

N = length(pod)      % 300 podatkov

% definiramo sredine razredov glede na podatke:

x=[486 490 494 498 502 506 510 514 518]

n = length(x)

% generiramo empiricne frekvence:

f_obs=hist(pod,x)

% 1. nacin izracuna teoreticnih frekvenc, kot predlaga trauth str 70:
% (tu v bistvu normalno porazdelitev gledamo le v tockah)

f_exp=normpdf(x,mean(pod),std(pod));    % hkrati ocenimo parametra
f_exp=f_exp/sum(f_exp);
f_exp=sum(f_obs)*f_exp                          % uglasimo velikostni razred s tistim od
empiricnih frekvenc

hi2 = sum((f_obs-f_exp).^2./f_exp)      % vrednost cenilke

hi2krit = chi2inv(1-alfa,n-t-1)         % kriticna vrednost cenilke


if hi2 > hi2krit
   disp('Padli smo v kriticno obmocje - zavrzi nicelno hipotezo')
else
   disp('Padli smo v obmocje zaupanja - sprejmi nicelno hipotezo')
end

%-------------------------------------------------------------------------

% Ker so bile nekatere frekvence manjse od 5, bomo raje zdruzevali razrede v manj razredov

% Nov, zmanjsan nabor razredov:

disp('------------------------------------------------------------------------')
disp('Nov izracun pri manjsem stevilu razredov')
disp('------------------------------------------------------------------------')

x = [min(pod) 492 496 500 504 508 max(pod)]

% Poiscemo empiricne frekvence:

for i = 1:length(x)-1
    if i == 1
       fij(i) = length(find((pod>=x(i))&(pod<x(i+1)+1)));
    else
       fij(i) = length(find((pod>x(i))&(pod<x(i+1)+1)));
    end
end
fij

x = [492 496 500 504 508];

n = length(x)+1

% Sedaj gremo racunat teoreticne frekvence. Vendar jih tokrat tvorimo kot
% razlike kumulativnih verjetnosti normalne porazdelitve glede na dane meje
% razredov:

f1=normcdf(-100:1:x(1),srvr,stdev);
f1=f1(length(f1));
eijz = f1*N;

for i = 1:length(x)-1
    f1=normcdf(-100:1:x(i),srvr,stdev);
    f1=f1(length(f1));
    f2=normcdf(-100:1:x(i+1),srvr,stdev);
```

```
    f2=f2(length(f2));
    eij(i) = (f2-f1)*N ;
end


f1=normcdf(-100:1:x(length(x)),srvr,stdev);
f1=f1(length(f1));
f2=1;
eijk = (f2-f1)*N;

% To so zracunane teoreticne frekvence:

eij = [eijz eij eijk]


hi2 = sum((fij-eij).^2./eij)            % vrednost cenilke

hi2krit = chi2inv(1-alfa,n-t-1)         % kriticna vrednost cenilke


if hi2 > hi2krit
   disp('Padli smo v kriticno obmocje - zavrzi nicelno hipotezo')
else
   disp('Padli smo v obmocje zaupanja - sprejmi nicelno hipotezo')
end

disp('          ')
disp('verjetnost p, da nicelna hipoteza drzi, je:')
p = 1-chi2cdf(hi2,n-t-1)


disp('------------------------------------------------------------------------------------
---')
disp('Nov izracun pri manjsem stevilu razredov - uporaba funkcije chi2gof (default
alfa=0.05)')
disp('------------------------------------------------------------------------------------
---')

[h,p,st]                                                                          =
chi2gof(pod,'cdf',@(z)normcdf(z,mean(pod),std(pod)),'nparams',2,'edges',[min(pod)  492  496
500  504  508 max(pod)])

disp('- ce nicelna hipoteza drzi, je h=0, sicer je 1. p je verjetnost, da nicelna hipoteza
drzi')
disp('- st.df je st. stopenj, st.O so empiricne, st.E pa teoreticne frekvence')

hi2krit = chi2inv(1-alfa,st.df)         % kriticna vrednost cenilke

if st.chi2stat > hi2krit
   disp('Padli smo v kriticno obmocje - zavrzi nicelno hipotezo')
else
   disp('Padli smo v obmocje zaupanja - sprejmi nicelno hipotezo')
end
```

Izpis komandnega okna je naslednji:

```
pod =
  Columns 1 through 13
   498   494   501   506   506   509   489   499   505   495   497   492   491
  Columns 14 through 26
   495   496   489   497   498   501   498   498   498   507   500   499   497
  Columns 27 through 39
   510   504   512   497   508   492   503   505   510   500   497   503   498
  Columns 40 through 52
   504   493   496   492   498   500   500   498   511   491   496   487   507
  Columns 53 through 65
```

```
  494   497   504   502   504   503   493   494   503   502   495   499   501
 Columns 66 through 78
  503   501   495   509   502   500   504   504   497   495   506   494   492
 Columns 79 through 91
  504   503   511   507   507   501   500   502   500   495   491   504   502
 Columns 92 through 104
  503   501   495   506   498   496   496   498   498   497   504   503   497
 Columns 105 through 117
  507   491   503   499   500   497   497   504   504   502   503   508   502
 Columns 118 through 130
  503   509   498   505   501   506   499   496   505   497   503   503   498
 Columns 131 through 143
  498   493   510   497   500   499   514   506   504   507   501   503   499
 Columns 144 through 156
  494   506   499   493   504   504   502   502   497   504   503   505   486
 Columns 157 through 169
  502   507   491   500   507   508   501   501   509   500   497   505   500
 Columns 170 through 182
  505   502   507   499   495   492   487   501   501   500   501   501   507
 Columns 183 through 195
  500   494   493   501   505   492   497   503   499   498   499   498   496
 Columns 196 through 208
  503   500   506   501   500   501   501   500   494   501   501   504   496
 Columns 209 through 221
  503   501   501   504   509   493   494   503   506   496   489   496   501
 Columns 222 through 234
  502   495   496   502   501   497   501   495   496   505   508   497   499
 Columns 235 through 247
  501   509   503   508   501   508   497   495   506   505   491   506   499
 Columns 248 through 260
  494   491   496   502   505   500   500   501   498   505   498   512   504
 Columns 261 through 273
  503   495   504   507   496   498   493   507   505   501   489   503   503
 Columns 274 through 286
  506   495   506   497   496   491   507   495   508   494   498   493   508
 Columns 287 through 299
  500   502   505   510   500   495   497   504   504   503   493   493   499
 Column 300
  497

alfa=0.05
alfa =
    0.0500

minpod =
   486
maxpod =
   514
srvr =
  500.2767
```

```
stdev =
    5.1739
t =
     2
N =
   300
x =
   486   490   494   498   502   506   510   514   518
n =
     9

f_obs =
     3    18    50    77    89    48    14     1     0
f_exp =
  Columns 1 through 8
    2.0568   12.8790   44.3599   84.0470   87.5944   50.2173   15.8363    2.7471
  Column 9
    0.2621

hi2 =
    5.4833
hi2krit =
   12.5916

Padli smo v obmocje zaupanja - sprejmi nicelno hipotezo

---------------------------------------------------------------------
Nov izracun pri manjsem stevilu razredov
---------------------------------------------------------------------
x =
   486   492   496   500   504   508   514
fij =
    21    50    77    89    48    15

n =
     6

eij =
   16.4500   44.8210   82.3322   85.6344   50.4368   20.3256

hi2 =
    3.8476
hi2krit =
    7.8147

Padli smo v obmocje zaupanja - sprejmi nicelno hipotezo

verjetnost p, da nicelna hipoteza drzi, je:
p =
    0.2784
```

```
-------------------------------------------------------------------------------
Nov izracun pri manjsem stevilu razredov - uporaba funkcije chi2gof (default alfa=0.05)
-------------------------------------------------------------------------------
h =
     0
p =
    0.1943
st =
        chi2stat: 4.7098
         df: 3
         edges: [486 492 496 500 504 508 514]
          O: [15 41 71 89 61 23]
          E: [16.4500 44.8210 82.3322 85.6344 50.4368 20.3256]

- ce nicelna hipoteza drzi, je h=0, sicer je 1. p je verjetnost, da nicelna hipoteza drzi
- st.df je st. stopenj, st.O so empiricne, st.E pa teoreticne frekvence

hi2krit =
    7.8147

Padli smo v obmocje zaupanja - sprejmi nicelno hipotezo
```

Kot je razvidno iz programa **goodnes3.m**, smo na koncu za izračune pri manjšem številu razredov uporabili še standardno Matlabovo funkcijo **chi2gof**. Iz izpisa komandnega okna je razvidno, da dobimo podobne rezultate. Kot je razvidno, pridejo teoretične frekvence enake kot v našem delu programa, empirične frekvence pa pridejo rahlo drugačne. Zato pride tudi vrednost cenilke nekoliko drugačna ($\chi^2 = 4.7098$), prav tako verjetnost zaupanja v dobljeni rezultat ($p = 0.1943$).

8.8 Primeri uporabe Matlaba pri ocenjevanju parametrov in preverjanju hipotez

Primer 8.23.:

Dano imamo eksponentno naključno spremenljivko s parametrom λ. Pokažite na primeru, kako iz podatkov vzorca eksponentne populacije ocenimo vrednost parametra s pomočjo metode momentov [Vidakovic].

Za metodo momentov velja (glej izraza (7.8) in (7.9)):

$$m_k' = \frac{\sum_{i=1}^{n} (x_i)^k}{n}$$
$$m_k' = \mu_k', \quad k = 1, 2, ..., r \tag{8.165}$$

Ker pri eksponentni funkciji nastopa le en parameter ($r = 1$), velja (imamo le eno enačbo):

$$m_1' = \mu_1', \quad k = 1 \tag{8.166}$$

Dobimo:

$$m_1' = \frac{\sum_{i=1}^{n} (x_i)^1}{n} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x} \tag{8.167}$$

Matematično upanje eksponentne naključne spremenljivke je enako $E(X)=\frac{1}{\lambda}$ (glej poglavje 5.2.3), torej dobimo:

$$\frac{\sum_{i=1}^{n} x_i}{n} = E(X) = \frac{1}{\lambda}$$
$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i} \qquad (8.168)$$

Denimo imamo eksponentno porazdelitev populacije s parametrom $\lambda = 3$. Program v Matlabu, kjer le-to simuliramo z vzorcem velikosti 10^6, ter ocenjujemo parameter z metodo momentov, bi bil:

```
% pogl8_8_1.m
% simuliran vzorec populacije:
X = exprnd(1/3, 10e6, 1);
disp('ocenjena vrednost parametra z metodo momentov je:')
lamb_oc = 1/mean(X)
```

Izpis komandnega okna:

```
>> pogl8_8_1
ocenjena vrednost parametra z metodo momentov je:

lamb_oc =

   3.0001
```

Primer 8.24.:

Pokažite uporabo funkcije mle (maximum likelihood estimation) za simuliran vzorec beta populacije (glej poglavje 5.2.6), kjer želimo oceniti parametra a in b z metodo največjega verjetja [Vidakovic]. Prava parametra sta: $a = 2, b = 3$.

Program v Matlabu je:

```
% pogl8_8_2.m

% simuliran vzorec beta populacije:

X = betarnd(2,3,[1 1000]);

disp('ocenjena vrednost parametrov a in b z metodo mle je:')

thetahat = mle(X,'distribution', 'beta');

aoc = thetahat(1)
boc = thetahat(2)
```

Izpis komandnega okna:

```
>> pogl8_8_2
ocenjena vrednost parametrov a in b z metodo mle je:

aoc =

   1.9952

boc =

   3.0069
```

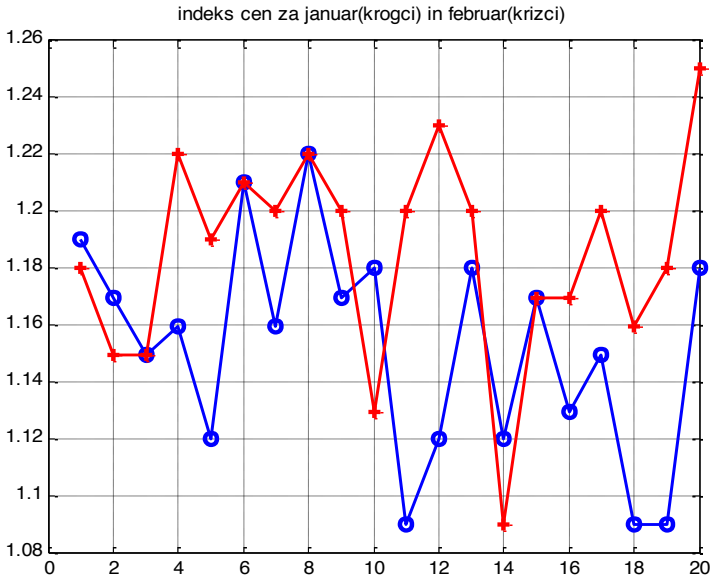
Primer 8.25.:

Dani imamo časovni vrsti indeksa cen za mesec januar in februar, prikazani na sliki 179. Velikosti obeh vzorcev sta 20. Populacija je normalna z znano standardno deviacijo $\sigma = s_0 = 0.04$ *in neznanim povprečjem* μ. *Preverite ničelno hipotezo za obe časovni vrsti (vzorca), da je dejansko povprečje populacije enako predpostavljenemu:* $\mu_0 = 1.16$. *[Žibert]* $(\alpha = 0.05)$.

Hipotezi sta:

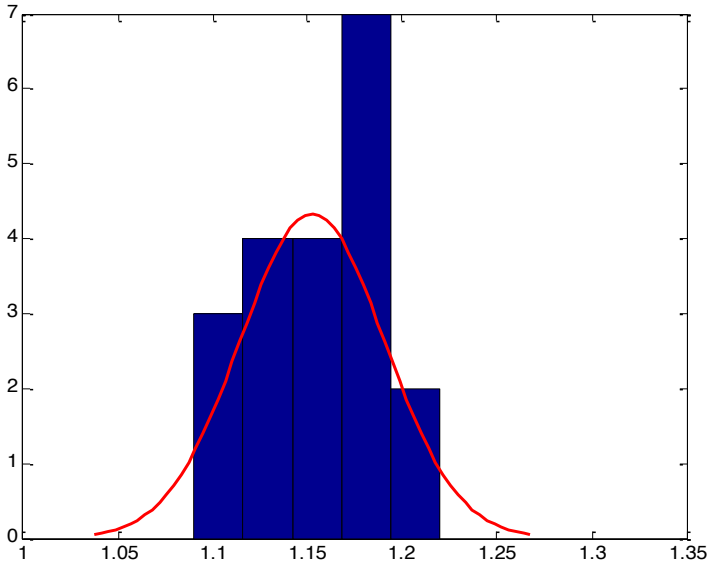
$$\left.\begin{aligned} H_0 &: \mu_1 = \mu_0 = 1.16 \\ H_1 &: \mu_1 \neq \mu_0 = 1.16 \end{aligned}\right\} za\ 1.časovno\ vrsto$$

$$\left.\begin{aligned} H_0 &: \mu_2 = \mu_0 = 1.16 \\ H_1 &: \mu_2 \neq \mu_0 = 1.16 \end{aligned}\right\} za\ 2.časovno\ vrsto \qquad (8.169)$$



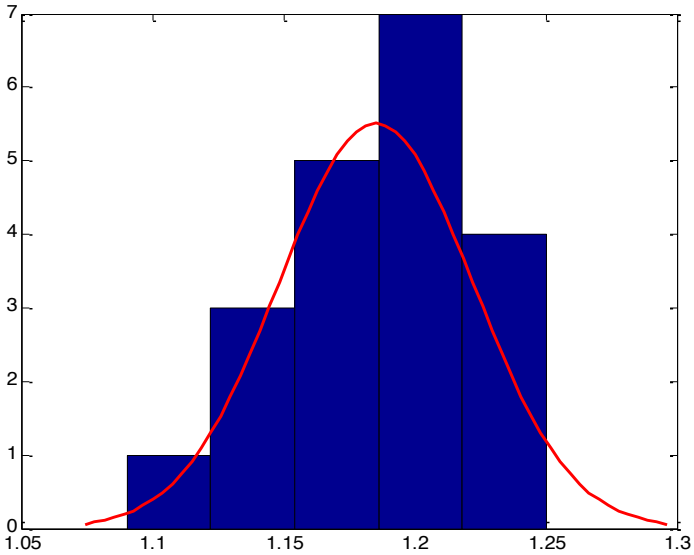
Slika 179: Časovni vrsti indeksa cen za mesec januar in februar

Histogram z uporabo funkcije **histfit** za 1. časovno vrsto je prikazan na sliki 180.



*Slika 180: Histogram z uporabo funkcije **histfit** za 1. časovno vrsto*

Histogram z uporabo funkcije **histfit** za 2. časovno vrsto je prikazan na sliki 181.



*Slika 181: Histogram z uporabo funkcije **histfit** za 2. časovno vrsto*

Iz slik 180 in 181 vidimo, da ima pri 1. časovni vrsti krivulja ujemanja vrh blizu predpostavljene vrednosti $\mu_0 = 1.16$, pri 2. časovni vrsti pa ne. Že iz tega sklepamo, da bo test ničelno hipotezo pri 1. časovni vrsti verjetno potrdil, pri drugi časovni vrsti pa verjetno zavrnil.

Kljub temu, da gre za mali vzorec pri obeh časovnih vrstah, vseeno lahko na osnovi izraza (7.41) zapišemo:

$$\begin{aligned} P\left(-z_{\frac{\alpha}{2}} \le \frac{\left(\bar{X}_1 - \mu_1\right)}{\frac{\sigma}{\sqrt{n}}} \le z_{\frac{\alpha}{2}}\right) &= 1-\alpha \ldots\ldots\ldots\ldots 1.\ \textit{časovna vrsta} \\ P\left(-z_{\frac{\alpha}{2}} \le \frac{\left(\bar{X}_2 - \mu_2\right)}{\frac{\sigma}{\sqrt{n}}} \le z_{\frac{\alpha}{2}}\right) &= 1-\alpha \ldots\ldots\ldots\ldots 2.\ \textit{časovna vrsta} \end{aligned} \qquad (8.170)$$

saj gre za vzorca iz normalne populacije. Pri tem velja, da sta cenilki (testni statistiki) za 1. oz. 2. časovno vrsto naslednji:

$$Z_1 = \frac{\left(\bar{X}_1 - \mu_1\right)}{\frac{\sigma}{\sqrt{n}}} \text{.....1. časovna vrsta}$$
$$Z_2 = \frac{\left(\bar{X}_2 - \mu_2\right)}{\frac{\sigma}{\sqrt{n}}} \text{.....2. časovna vrsta} \tag{8.171}$$

Seveda pri ničelni hipotezi v izrazih (8.170) in (8.171) upoštevamo: $\mu_1 = \mu_2 = \mu_0 = 1.16$.

Intervalni oceni na osnovi izraza (7.44) sta:

$$I_{januar} = \left[\bar{X}_1 - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \bar{X}_1 + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right]$$
$$I_{februar} = \left[\bar{X}_2 - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \bar{X}_2 + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right] \tag{8.172}$$

Dobimo:

$$\bar{x}_1 = 1.1525$$
$$\bar{x}_2 = 1.1850$$
$$z_1 = \frac{\left(\bar{x}_1 - \mu_0\right)}{\frac{\sigma}{\sqrt{n}}} = \frac{\left(1.1525 - 1.16\right)}{\frac{0.04}{\sqrt{20}}} = -0.8385$$
$$z_2 = \frac{\left(\bar{x}_2 - \mu_0\right)}{\frac{\sigma}{\sqrt{n}}} = \frac{\left(1.1850 - 1.16\right)}{\frac{0.04}{\sqrt{20}}} = 2.7951 \tag{8.173}$$

Kritična vrednost je enaka:

$$z_{\frac{\alpha}{2}} = z_{\frac{0.05}{2}} = 1.96 \tag{8.174}$$

Območje zaupanja je enako:

$$\left(-z_{\frac{\alpha}{2}}, z_{\frac{\alpha}{2}}\right) = \left(-1.96, 1.96\right) \tag{8.175}$$

Ker velja:

$$\begin{aligned} z_1 &= -0.8385 \in \left(-z_{\frac{\alpha}{2}}, z_{\frac{\alpha}{2}} \right) = (-1.96, 1.96) \\ z_2 &= 2.7951 \notin \left(-z_{\frac{\alpha}{2}}, z_{\frac{\alpha}{2}} \right) = (-1.96, 1.96) \end{aligned} \tag{8.176}$$

test ničelno hipotezo pri 1. časovni vrsti potrdi, pri drugi časovni vrsti pa jo zavrne.

Verjetnost zaupanja v rezultat (p-vrednost oz p-value) pri obeh časovnih vrstah je enaka:

$$\begin{aligned} p_1 &= P\left(Z_1 \in \left(-z_{\frac{\alpha}{2}}, z_{\frac{\alpha}{2}} \right) \middle| H_0 : \mu_1 = \mu_0 = 1.16 \right) = 0.4017 > \alpha = 0.05 \\ p_2 &= P\left(Z_2 \in \left(-z_{\frac{\alpha}{2}}, z_{\frac{\alpha}{2}} \right) \middle| H_0 : \mu_2 = \mu_0 = 1.16 \right) = 0.0052 < \alpha = 0.05 \end{aligned} \tag{8.177}$$

in pomeni verjetnost, da dobimo vrednost testne statistike kot ekstrem glede na dejansko dobljeno vrednost, pri pogoju, da je ničelna hipoteza res prava [Vidakovic]. Če je p vrednost velika (večja od α), to pomeni dodatno potrditev, da lahko ničelno hipotezo sprejmemo, in obratno. Matlab ukaza za p-vrednost sta bila:

```
p1 = 2*(1-normcdf(abs(z1))) % p-vrednost za obojestranski
test
p2 = 2*(1-normcdf(abs(z2))) % p-vrednost za obojestranski
test
```

Intervalni oceni prideta:

(8.178)

$$I_{januar} = \left[\bar{X}_1 - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \bar{X}_1 + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \right] =$$

$$= \left[1.1525 - 1.96 \cdot \frac{0.04}{\sqrt{20}}, 1.1525 + 1.96 \cdot \frac{0.04}{\sqrt{20}} \right] = [1.1350, 1.1700]$$

$$I_{februar} = \left[\bar{X}_2 - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \bar{X}_2 + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \right] =$$

$$= \left[1.1850 - 1.96 \cdot \frac{0.04}{\sqrt{20}}, 1.1850 + 1.96 \cdot \frac{0.04}{\sqrt{20}} \right] = [1.1675, 1.2025]$$

Kot vidimo, predpostavljena vrednost za povprečje $\mu_0 = 1.16$ pri 1. časovni vrsti pade znotraj intervalne ocene, pri 2. časovni vrsti pa pade izven. Tudi po tem lahko sklepamo, da test ničelno hipotezo pri 1. časovni vrsti potrdi, pri drugi časovni vrsti pa jo zavrne.

Vse izračune smo dobili z naslednjim programom v Matlabu:

```
% pog18_8_3.m

clear
clc
close all

disp('znana deviacija:')
s0 = 0.04  % znan standardni odklon populacije

disp('predpostavljena aritmeticna sredina:')
m0 = 1.16  % predpostavljeno povprečje populacije

% podamo obe casovni vrsti in histograma:

cene_jan= [1.19 1.17 1.15 1.16 1.12 1.21 1.16 1.22 1.17 1.18 1.09 1.12 1.18 1.12 1.17 1.13
1.15 1.09 1.09 1.18]
cene_feb= [1.18 1.15 1.15 1.22 1.19 1.21 1.20 1.22 1.20 1.13 1.20 1.23 1.20 1.09 1.17 1.17
1.20 1.16 1.18 1.25]

plot(cene_jan,'LineWidth',1.5)
hold on
plot(cene_jan,'o','LineWidth',2.5)

plot(cene_feb,'r','LineWidth',1.5)
hold on
plot(cene_feb,'r+','LineWidth',2.5)

grid
title('indeks cen za januar(krogci) in februar(krizci)')

figure
histfit(cene_jan)
figure
histfit(cene_feb)

n = 20

% naredimo oceno povprecja za obe casovni vrsti

m1 = mean(cene_jan)
m2 = mean(cene_feb)

% Tvorimo z vrednosti za obe casovni vrsti:

z1 = (m1 - m0) / (s0/sqrt(n)) % testna statistika za m1
z2 = (m2 - m0) / (s0/sqrt(n)) % testna statistika za m2
```

```
meja = norminv([0.025 0.975],0,1) %leva in desna meja porazdelitve za alfa=0.05

if (z1<meja(1))||(z1>meja(2))
    disp('zavrni nicelno hipotezo za 1. casovno vrsto')
else
    disp('sprejmi nicelno hipotezo za 1. casovno vrsto')
end

if (z2<meja(1))||(z2>meja(2))
    disp('zavrni nicelno hipotezo za 2. casovno vrsto')
else
    disp('sprejmi nicelno hipotezo za 2. casovno vrsto')
end



disp('verjetnost zaupanja v rezultat pri 1. in 2. casovni vrsti:')

p1 = 2*(1-normcdf(abs(z1))) % p-vrednost za obojestranski test
p2 = 2*(1-normcdf(abs(z2))) % p-vrednost za obojestranski test

disp('interval zaupanja v srednjo vrednost za 1. casovno vrsto')
I1 = m1 - meja(2)*s0/sqrt(n);
I2 = m1 + meja(2)*s0/sqrt(n);
I = [I1,I2]


disp('interval zaupanja v srednjo vrednost za 2. casovno vrsto')
I1 = m2 - meja(2)*s0/sqrt(n);
I2 = m2 + meja(2)*s0/sqrt(n);
I = [I1,I2]
```

Izpis komandnega okna je naslednji:

```
znana deviacija:
s0 =
   0.0400


predpostavljena aritmeticna sredina:
m0 =
   1.1600


cene_jan =
  Columns 1 through 8
    1.1900    1.1700    1.1500    1.1600    1.1200    1.2100    1.1600    1.2200
  Columns 9 through 16
    1.1700    1.1800    1.0900    1.1200    1.1800    1.1200    1.1700    1.1300
  Columns 17 through 20
    1.1500    1.0900    1.0900    1.1800


cene_feb =
  Columns 1 through 8
    1.1800    1.1500    1.1500    1.2200    1.1900    1.2100    1.2000    1.2200
  Columns 9 through 16
    1.2000    1.1300    1.2000    1.2300    1.2000    1.0900    1.1700    1.1700
  Columns 17 through 20
    1.2000    1.1600    1.1800    1.2500


n =
```

```
    20

m1 =
    1.1525
m2 =
    1.1850

z1 =
   -0.8385
z2 =
    2.7951
meja =
   -1.9600    1.9600

sprejmi nicelno hipotezo za 1. casovno vrsto
zavrni nicelno hipotezo za 2. casovno vrsto

verjetnost zaupanja v rezultat pri 1. in 2. casovni vrsti:
p1 =
    0.4017
p2 =
    0.0052

interval zaupanja v srednjo vrednost za 1. casovno vrsto
I =
    1.1350    1.1700
interval zaupanja v srednjo vrednost za 2. casovno vrsto
I =
    1.1675    1.2025
```

Izračune bi lahko opravili tudi s standardno funkcijo v Matlabu **ztest**. Klic funkcije in izpis v komandnem oknu za 1. časovno vrsto bi bila naslednja:

```
[h,p,ci]=ztest(cene_jan,1.16,0.04,0.05,'both')

h =
     0

p =
    0.4017

ci =
    1.1350    1.1700
```

Klic funkcije in izpis v komandnem oknu za 2. časovno vrsto pa bi bila naslednja:

```
>> [h,p,ci]=ztest(cene_feb,1.16,0.04,0.05,'both')


h =
     1

p =
    0.0052

ci =
    1.1675    1.2025
```

Kot lahko vidimo, dobimo popolnoma enake rezultate.

Primer 8.26.:

Imamo popolnoma enak primer prejšnjemu, le da tokrat standardni odklon populacije ni znan. Preverite ničelno hipotezo za obe časovni vrsti (vzorca), da je dejansko povprečje populacije enako predpostavljenemu: $\mu_0 = 1.16$. *[Žibert](*$\alpha = 0.05$*).*

Kadar je velikost vzorca majhna in je varianca normalne populacije neznana, ne moremo več uporabiti *z* statistike, pač pa preiti na *t* statistiko. Testni statistiki za obe časovni vrsti sta tokrat enaki (velikost vzorca je pri obeh *n = 20*) (glej izraz (8.28)):

$$T_1 = \frac{\left(\bar{X}_1 - \mu_0\right)}{\frac{s_1}{\sqrt{n}}}$$
$$T_2 = \frac{\left(\bar{X}_2 - \mu_0\right)}{\frac{s_2}{\sqrt{n}}} \qquad (8.179)$$

z *n - 1* prostostnimi stopnjami.

Kritično območje je za obe časovni vrsti glede na nasprotno hipotezo določeno na naslednji način (glej izraz (8.29)):

(8.180)

$$-t_{\frac{\alpha}{2},n-1} \le T_1 \le t_{\frac{\alpha}{2},n-1}$$
$$-t_{\frac{\alpha}{2},n-1} \le T_2 \le t_{\frac{\alpha}{2},n-1}$$

Intervalni oceni na osnovi izraza (7.57) sta:

$$I_{januar} = \left[\bar{X}_1 - t_{\frac{\alpha}{2},n-1} \cdot \frac{s_1}{\sqrt{n}}, \bar{X}_1 + t_{\frac{\alpha}{2},n-1} \cdot \frac{s_1}{\sqrt{n}} \right]$$
$$I_{februar} = \left[\bar{X}_2 - t_{\frac{\alpha}{2},n-1} \cdot \frac{s_2}{\sqrt{n}}, \bar{X}_2 + t_{\frac{\alpha}{2},n-1} \cdot \frac{s_2}{\sqrt{n}} \right] \quad (8.181)$$

Dobimo:

$$\bar{x}_1 = 1.1525$$
$$\bar{x}_2 = 1.1850$$
$$s_1 = 0.0384$$
$$s_2 = 0.0371 \quad (8.182)$$
$$t_1 = \frac{(\bar{x}_1 - \mu_0)}{\frac{s_1}{\sqrt{n}}} = \frac{(1.1525 - 1.16)}{\frac{0.0384}{\sqrt{20}}} = -0.8741$$
$$t_2 = \frac{(\bar{x}_2 - \mu_0)}{\frac{s_2}{\sqrt{n}}} = \frac{(1.1850 - 1.16)}{\frac{0.0371}{\sqrt{20}}} = 3.0166$$

Kritična vrednost je enaka:

$$t_{\frac{\alpha}{2},n-1} = t_{\frac{0.05}{2},19} = 2.093 \quad (8.183)$$

Območje zaupanja je enako:

$$\left(-t_{\frac{\alpha}{2},n-1}, t_{\frac{\alpha}{2},n-1} \right) = (-2.093, 2.093) \quad (8.184)$$

Ker velja:

$$t_1 = -0.8741 \in \left(-t_{\frac{\alpha}{2},n-1}, t_{\frac{\alpha}{2},n-1} \right) = (-2.093, 2.093) \tag{8.185}$$
$$t_2 = 3.0166 \notin \left(-t_{\frac{\alpha}{2},n-1}, t_{\frac{\alpha}{2},n-1} \right) = (-2.093, 2.093)$$

test ničelno hipotezo pri 1. časovni vrsti potrdi, pri drugi časovni vrsti pa jo zavrne.

Verjetnost zaupanja v rezultat (p-vrednost oz p-value) pri obeh časovnih vrstah je enaka:

$$p_1 = P\left(T_1 \in \left(-t_{\frac{\alpha}{2},n-1}, t_{\frac{\alpha}{2},n-1} \right) \middle| H_0 : \mu_1 = \mu_0 = 1.16 \right) = 0.393 > \alpha = 0.05 \tag{8.186}$$
$$p_2 = P\left(T_2 \in \left(-t_{\frac{\alpha}{2},n-1}, t_{\frac{\alpha}{2},n-1} \right) \middle| H_0 : \mu_2 = \mu_0 = 1.16 \right) = 0.0071 < \alpha = 0.05$$

Matlab ukaza za p-vrednost sta bila:

```
p1 = tcdf(t1,n-1);
p1 = 2*min(p1,1-p1)  % p-vrednost za obojestranski test

p2 = tcdf(t2,n-1);
p2 = 2*min(p2,1-p2)  % p-vrednost za obojestranski test
```

Intervalni oceni prideta:

(8.187)

$$I_{januar} = \left[\bar{X}_1 - t_{\frac{\alpha}{2},n-1} \cdot \frac{s_1}{\sqrt{n}}, \bar{X}_1 + t_{\frac{\alpha}{2},n-1} \cdot \frac{s_1}{\sqrt{n}} \right] =$$

$$= \left[1.1525 - 2.093 \cdot \frac{0.0384}{\sqrt{20}}, 1.1525 + 2.093 \cdot \frac{0.0384}{\sqrt{20}} \right] = [1.1345, 1.1705]$$

$$I_{februar} = \left[\bar{X}_2 - t_{\frac{\alpha}{2},n-1} \cdot \frac{s_2}{\sqrt{n}}, \bar{X}_2 + t_{\frac{\alpha}{2},n-1} \cdot \frac{s_2}{\sqrt{n}} \right] =$$

$$= \left[1.1850 - 2.093 \cdot \frac{0.0371}{\sqrt{20}}, 1.1850 + 2.093 \cdot \frac{0.0371}{\sqrt{20}} \right] = [1.1677, 1.2023]$$

Kot vidimo, predpostavljena vrednost za povprečje $\mu_0 = 1.16$ pri 1. časovni vrsti pade znotraj intervalne ocene, pri 2. časovni vrsti pa pade izven. Tudi po tem lahko sklepamo, da test ničelno hipotezo pri 1. časovni vrsti potrdi, pri drugi časovni vrsti pa jo zavrne.

Vse izračune smo dobili z naslednjim programom v Matlabu:

```
% pogl8_8_4.m

clear
clc
close all

disp('predpostavljena aritmeticna sredina:')
m0 = 1.16  % predpostavljeno povprečje populacije

% podamo obe casovni vrsti:

cene_jan= [1.19 1.17 1.15 1.16 1.12 1.21 1.16 1.22 1.17 1.18 1.09 1.12 1.18 1.12 1.17 1.13
1.15 1.09 1.09 1.18]
cene_feb= [1.18 1.15 1.15 1.22 1.19 1.21 1.20 1.22 1.20 1.13 1.20 1.23 1.20 1.09 1.17 1.17
1.20 1.16 1.18 1.25]

n = 20

% naredimo oceno povprecja in oceno odklona za obe casovni vrsti

m1 = mean(cene_jan)
m2 = mean(cene_feb)
s1 = std(cene_jan)
s2 = std(cene_feb)

% Tvorimo t vrednosti za obe casovni vrsti:

t1 = (m1 - m0) / (s1/sqrt(n)) % testna statistika za m1
t2 = (m2 - m0) / (s2/sqrt(n)) % testna statistika za m2

meja = tinv([0.025 0.975],n-1) %leva in desna meja porazdelitve za alfa=0.05

if (t1<meja(1))||(t1>meja(2))
```

```
    disp('zavrni nicelno hipotezo za 1. casovno vrsto')
else
    disp('sprejmi nicelno hipotezo za 1. casovno vrsto')
end

if (t2<meja(1))||(t2>meja(2))
    disp('zavrni nicelno hipotezo za 2. casovno vrsto')
else
    disp('sprejmi nicelno hipotezo za 2. casovno vrsto')
end

disp('verjetnost zaupanja v rezultat pri 1. in 2. casovni vrsti:')

p1 = tcdf(t1,n-1);
p1 = 2*min(p1,1-p1)  % p-vrednost za obojestranski test

p2 = tcdf(t2,n-1);
p2 = 2*min(p2,1-p2)  % p-vrednost za obojestranski test

disp('interval zaupanja v srednjo vrednost za 1. casovno vrsto')
I1 = m1 - meja(2)*s1/sqrt(n);
I2 = m1 + meja(2)*s1/sqrt(n);
I = [I1,I2]


disp('interval zaupanja v srednjo vrednost za 2. casovno vrsto')
I1 = m2 - meja(2)*s2/sqrt(n);
I2 = m2 + meja(2)*s2/sqrt(n);
I = [I1,I2]
```

Izpis komandnega okna je naslednji:

```
predpostavljena aritmeticna sredina:
m0 =
   1.1600


cene_jan =
  Columns 1 through 8
    1.1900    1.1700    1.1500    1.1600    1.1200    1.2100    1.1600    1.2200
  Columns 9 through 16
    1.1700    1.1800    1.0900    1.1200    1.1800    1.1200    1.1700    1.1300
  Columns 17 through 20
    1.1500    1.0900    1.0900    1.1800


cene_feb =
  Columns 1 through 8
    1.1800    1.1500    1.1500    1.2200    1.1900    1.2100    1.2000    1.2200
  Columns 9 through 16
    1.2000    1.1300    1.2000    1.2300    1.2000    1.0900    1.1700    1.1700
  Columns 17 through 20
    1.2000    1.1600    1.1800    1.2500


n =
   20


m1 =
```

```
    1.1525
m2 =
    1.1850
s1 =
    0.0384
s2 =
    0.0371

t1 =
   -0.8741
t2 =
    3.0166

meja =
   -2.0930    2.0930

sprejmi nicelno hipotezo za 1. casovno vrsto
zavrni nicelno hipotezo za 2. casovno vrsto

verjetnost zaupanja v rezultat pri 1. in 2. casovni vrsti:
p1 =
    0.3930
p2 =
    0.0071




interval zaupanja v srednjo vrednost za 1. casovno vrsto
I =
    1.1345    1.1705

interval zaupanja v srednjo vrednost za 2. casovno vrsto
I =
    1.1677    1.2023
```

Izračune bi lahko opravili tudi s standardno funkcijo v Matlabu **ttest**. Klic funkcije in izpis v komandnem oknu za 1. časovno vrsto bi bila naslednja:

```
[h,p,ci]=ttest(cene_jan,1.16,0.05,'both')


h =
     0

p =
    0.3930

ci =
```

```
    1.1345    1.1705
```

Klic funkcije in izpis v komandnem oknu za 2. časovno vrsto pa bi bila naslednja:

```
>> [h,p,ci]=ttest(cene_feb,1.16,0.05,'both')


h =
     1

p =
    0.0071

ci =
    1.1677    1.2023
```

Kot lahko vidimo, dobimo popolnoma enake rezultate.

Primer 8.27.:

Dani imamo časovni vrsti indeksa cen za mesec januar in februar, prikazani na sliki 179. Velikosti obeh vzorcev sta 20. Populacija je normalna s predpostavljeno standardno deviacijo $\sigma = \sigma_0 = s_0 = 0.04$. *Preverite ničelno hipotezo za obe časovni vrsti (vzorca), da je dejanska standardna deviacija enaka predpostavljeni. [Žibert]* ($\alpha = 0.05$).

Hipotezi sta:

$$\left.\begin{aligned} H_0 &: \sigma_1^2 = \sigma_0^2 = 0.04^2 \\ H_1 &: \sigma_1^2 \neq \sigma_0^2 = 0.04^2 \end{aligned}\right\} za\ 1.časovno\ vrsto$$

$$\left.\begin{aligned} H_0 &: \sigma_2^2 = \sigma_0^2 = 0.04^2 \\ H_1 &: \sigma_2^2 \neq \sigma_0^2 = 0.04^2 \end{aligned}\right\} za\ 2.časovno\ vrsto \qquad (8.188)$$

Testni statistiki za obe časovni vrsti sta tokrat enaki (velikost vzorca je pri obeh $n = 20$) (glej izraz (8.50)):

$$\begin{aligned}\chi_1^2(n-1) &= \frac{1}{\sigma_1^2}\cdot S_1^2\cdot(n-1) = \frac{1}{\sigma_0^2}\cdot S_1^2\cdot(n-1)\\ \chi_2^2(n-1) &= \frac{1}{\sigma_2^2}\cdot S_2^2\cdot(n-1) = \frac{1}{\sigma_0^2}\cdot S_2^2\cdot(n-1)\end{aligned} \tag{8.189}$$

Kritično območje je za obe časovni vrsti glede na nasprotno hipotezo določeno na naslednji način (glej izraz (8.51)):

$$\begin{aligned}&\left(\chi_1^2(n-1)\le\chi^2_{1-\frac{\alpha}{2}}(n-1)\right)\vee\left(\chi_1^2(n-1)\ge\chi^2_{\frac{\alpha}{2}}(n-1)\right)\\ &\left(\chi_2^2(n-1)\le\chi^2_{1-\frac{\alpha}{2}}(n-1)\right)\vee\left(\chi_2^2(n-1)\ge\chi^2_{\frac{\alpha}{2}}(n-1)\right)\end{aligned} \tag{8.190}$$

Intervalni oceni na osnovi izraza (7.105) sta:

$$\begin{aligned}I_{januar} &= \left(\frac{S_1^2\cdot(n-1)}{\chi^2_{n-1,\frac{\alpha}{2}}},\frac{S_1^2\cdot(n-1)}{\chi^2_{n-1,1-\frac{\alpha}{2}}}\right)\\ I_{februar} &= \left(\frac{S_2^2\cdot(n-1)}{\chi^2_{n-1,\frac{\alpha}{2}}},\frac{S_2^2\cdot(n-1)}{\chi^2_{n-1,1-\frac{\alpha}{2}}}\right)\end{aligned} \tag{8.191}$$

Dobimo:

(8.192)

$$s_1 = 0.0384$$
$$s_2 = 0.0371$$
$$\chi_1^2(n-1) = \frac{1}{\sigma_0^2} \cdot s_1^2 \cdot (n-1) = \frac{1}{0.04^2} \cdot 0.0384^2 \cdot (20-1) = 17.4844$$
$$\chi_2^2(n-1) = \frac{1}{\sigma_0^2} \cdot s_2^2 \cdot (n-1) = \frac{1}{0.04^2} \cdot 0.0371^2 \cdot (20-1) = 16.3125$$

Območje zaupanja OZ pride:

$$OZ = [8.9065, 32.8523] \tag{8.193}$$

Ker pri obeh časovnih vrstah velja:

$$\chi_1^2(n-1) \in OZ = [8.9065, 32.8523] \tag{8.194}$$
$$\chi_2^2(n-1) \in OZ = [8.9065, 32.8523]$$

test ničelno hipotezo potrdi za obe časovni vrsti.

Verjetnost zaupanja v rezultat (p-vrednost oz p-value) pri obeh časovnih vrstah je enaka:

$$p_1 = 0.8858 > \alpha = 0.05 \tag{8.195}$$
$$p_2 = 0.7273 > \alpha = 0.05$$

Matlab ukaza za p-vrednost sta bila:

```
p1 = chi2cdf(hi1,n-1);
p1 = 2*min(p1,1-p1)  % p-vrednost za obojestranski test

p2 = chi2cdf(hi2,n-1);
p2 = 2*min(p2,1-p2)  % p-vrednost za obojestranski test
```

Intervalni oceni prideta:

$$I_{januar} = \left(\frac{s_1^2 \cdot (n-1)}{\chi^2_{n-1,\frac{\alpha}{2}}}, \frac{s_1^2 \cdot (n-1)}{\chi^2_{n-1,1-\frac{\alpha}{2}}} \right) = \left(\frac{0.0384^2 \cdot (20-1)}{32.8523}, \frac{0.0384^2 \cdot (20-1)}{8.9065} \right) = \quad (8.196)$$

$$= (0.0009, 0.0031)$$

$$I_{februar} = \left(\frac{s_2^2 \cdot (n-1)}{\chi^2_{n-1,\frac{\alpha}{2}}}, \frac{s_2^2 \cdot (n-1)}{\chi^2_{n-1,1-\frac{\alpha}{2}}} \right) = \left(\frac{0.0371^2 \cdot (20-1)}{32.8523}, \frac{0.0371^2 \cdot (20-1)}{8.9065} \right) =$$

$$= (0.0008, 0.0029)$$

Ker je predpostavljena varianca enaka: $\sigma_0^2 = s_0^2 = 0.04^2 = 0.0016$, očitno pri obeh časovnih vrstah pade znotraj intervalne ocene. Tudi po tem lahko sklepamo, da test ničelno hipotezo potrdi pri obeh časovnih vrstah.

Vse izračune smo dobili z naslednjim programom v Matlabu:

```
% pog18_8_5.m

clear
clc
close all

disp('predpostavljen standardni odklon:')
s0 = 0.04

% podamo obe casovni vrsti:

cene_jan= [1.19 1.17 1.15 1.16 1.12 1.21 1.16 1.22 1.17 1.18 1.09 1.12 1.18 1.12 1.17 1.13
1.15 1.09 1.09 1.18]
cene_feb= [1.18 1.15 1.15 1.22 1.19 1.21 1.20 1.22 1.20 1.13 1.20 1.23 1.20 1.09 1.17 1.17
1.20 1.16 1.18 1.25]

n = 20

% naredimo oceno odklona za obe casovni vrsti

s1 = std(cene_jan)
s2 = std(cene_feb)

% Tvorimo hi vrednosti za obe casovni vrsti:

hi1 = (n-1)*s1^2 / s0^2 % testna stat. za s1
hi2 = (n-1)*s2^2 / s0^2 % testna stat. za s2

meja = chi2inv([0.025 0.975],n-1) %leva in desna meja porazdelitve za alfa=0.05

if (hi1<meja(1))||(hi1>meja(2))
    disp('zavrni nicelno hipotezo za 1. casovno vrsto')
else
    disp('sprejmi nicelno hipotezo za 1. casovno vrsto')
end

if (hi2<meja(1))||(hi2>meja(2))
    disp('zavrni nicelno hipotezo za 2. casovno vrsto')
else
    disp('sprejmi nicelno hipotezo za 2. casovno vrsto')
end

disp('verjetnost zaupanja v rezultat pri 1. in 2. casovni vrsti:')
```

```
p1 = chi2cdf(hi1,n-1);
p1 = 2*min(p1,1-p1)  % p-vrednost za obojestranski test

p2 = chi2cdf(hi2,n-1);
p2 = 2*min(p2,1-p2)  % p-vrednost za obojestranski test

disp('interval zaupanja v varianco za 1. casovno vrsto')
I1 = s1^2*(n-1)/meja(2);
I2 = s1^2*(n-1)/meja(1);
I = [I1,I2]


disp('interval zaupanja v varianco za 2. casovno vrsto')
I1 = s2^2*(n-1)/meja(2);
I2 = s2^2*(n-1)/meja(1);
I = [I1,I2]

disp('predpostavljena varianca je:')
s0^2
```

Izpis komandnega okna je naslednji:

```
predpostavljen standardni odklon:
s0 =
   0.0400


cene_jan =
  Columns 1 through 8
    1.1900    1.1700    1.1500    1.1600    1.1200    1.2100    1.1600    1.2200
  Columns 9 through 16
    1.1700    1.1800    1.0900    1.1200    1.1800    1.1200    1.1700    1.1300
  Columns 17 through 20
    1.1500    1.0900    1.0900    1.1800


cene_feb =
  Columns 1 through 8
    1.1800    1.1500    1.1500    1.2200    1.1900    1.2100    1.2000    1.2200
  Columns 9 through 16
    1.2000    1.1300    1.2000    1.2300    1.2000    1.0900    1.1700    1.1700
  Columns 17 through 20
    1.2000    1.1600    1.1800    1.2500


n =
   20


s1 =
   0.0384
s2 =
   0.0371


hi1 =
   17.4844
hi2 =
   16.3125
```

```
meja =
   8.9065   32.8523

sprejmi nicelno hipotezo za 1. casovno vrsto
sprejmi nicelno hipotezo za 2. casovno vrsto

verjetnost zaupanja v rezultat pri 1. in 2. casovni vrsti:
p1 =
   0.8858
p2 =
   0.7273

interval zaupanja v varianco za 1. casovno vrsto
I =
   0.0009    0.0031
interval zaupanja v varianco za 2. casovno vrsto
I =
   0.0008    0.0029

predpostavljena varianca je:
ans =
   0.0016
```

Izračune bi lahko opravili tudi s standardno funkcijo v Matlabu **vartest**. Klic funkcije in izpis v komandnem oknu za 1. časovno vrsto bi bila naslednja:

```
>> [h,p,ci]=vartest(cene_jan,0.04^2,0.05,'both')


h =
    0
p =
   0.8858
ci =
   0.0009    0.0031
```

Klic funkcije in izpis v komandnem oknu za 2. časovno vrsto pa bi bila naslednja:

```
>> [h,p,ci]=vartest(cene_feb,0.04^2,0.05,'both')


h =
    0
p =
   0.7273
ci =
   0.0008    0.0029
```

Kot lahko vidimo, dobimo popolnoma enake rezultate.

Primer 8.28.:

Imamo primer ugotavljanja koncentracij arzena v pitni vodi v okrožju Phoenix in ruralnih predelih Arizone, ZDA. Izvedene so bile meritve v mestnih predelih okrožja Phoenixa in na podeželju Arizone. Meritve prikazuje slika 182. Testirajte ničelno hipotezo, da sta koncentraciji v mestnih predelih in na podeželju v povprečju enaki [Žibert] ($\alpha = 0.05$).

Metro Phoenix ($\bar{x}_1 = 12.5, s_1 = 7.63$)	Rural Arizona ($\bar{x}_2 = 27.5, s_2 = 15.3$)
Phoenix, 3	Rimrock, 48
Chandler, 7	Goodyear, 44
Gilbert, 25	New River, 40
Glendale, 10	Apachie Junction, 38
Mesa, 15	Buckeye, 33
Paradise Valley, 6	Nogales, 21
Peoria, 12	Black Canyon City, 20
Scottsdale, 25	Sedona, 12
Tempe, 15	Payson, 1
Sun City, 7	Casa Grande, 18

Slika 182: Meritve koncentracij v mestnih predelih (metro Phoenix) in na podeželju (Rural Arizona) (rural -podeželje) [Žibert]

Če predvidevamo, da sta neznani varianci med seboj enaki, lahko pri majhnih vzorcih uporabimo izraze (7.72) do (7.77). Če pa upravičeno sklepamo, da neznani varianci med seboj nista enaki, potem se izkaže, **da ne obstaja neka točna t-statistika za testiranje** $H_0 : \mu_1 - \mu_2 = \Delta_0$ [Montgomery 1]. Kljub temu pa se izkaže, da je statistika [Montgomery 1]:

(8.197)

$$T_0^* = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - \left(\Delta_0\right)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

porazdeljena **<u>približno</u>** v skladu s t statistiko s stopnjo prostosti [Montgomery 1]:

$$\nu = \frac{\left(\frac{S_1^{\,2}}{n_1} + \frac{S_2^{\,2}}{n_2}\right)^2}{\frac{\left(\frac{S_1^{\,2}}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^{\,2}}{n_2}\right)^2}{n_2 - 1}} \tag{8.198}$$

Postavimo hipotezi:

$$\begin{aligned} H_0 &: \mu_1 - \mu_2 = \Delta_0 = 0 \\ H_1 &: \mu_1 - \mu_2 = \Delta_0 \neq 0 \end{aligned} \tag{8.199}$$

Izraz (8.197) preide pri ničelni hipotezi v obliko:

$$T_0^* = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - (0)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - (0)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{\left(\bar{X}_1 - \bar{X}_2\right)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \tag{8.200}$$

kar je naša testna statistika. Pri konkretno zajetem vzorcu iz mestnih in ruralnih predelov zavzame vrednost:

$$t_0^* = \frac{\left(\bar{x}_1 - \bar{x}_2\right)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(12.5 - 27.5)}{\sqrt{\frac{7.63^2}{10} + \frac{15.3^2}{10}}} = \frac{-15}{5.4065} = -2.7669 \tag{8.201}$$

Stopnja prostosti zavzame vrednost:

$$\nu = \frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1-1}+\frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2-1}} = \frac{\left(\frac{7.63^2}{10}+\frac{15.3^2}{10}\right)^2}{\frac{\left(\frac{7.63^2}{10}\right)^2}{9}+\frac{\left(\frac{15.3^2}{10}\right)^2}{9}} = 9\cdot\frac{\left(7.63^2+15.3^2\right)^2}{\left(7.63^2\right)^2+\left(15.3^2\right)^2} = 13.1956 \quad (8.202)$$

$$\nu \approx 13$$

Območje zaupanja OZ dobimo z ukazom:

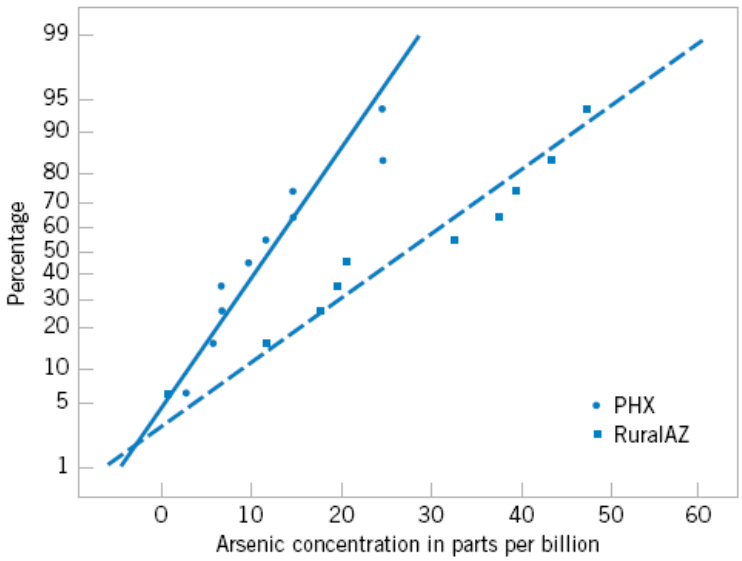
```
>> OZ = tinv([0.025, 0.975], 13)

OZ =
   -2.1604    2.1604
```

Ker velja:

$$t_0^* = -2.7669 \notin OZ = \left[-2.1604, 2.1604\right] \quad (8.203)$$

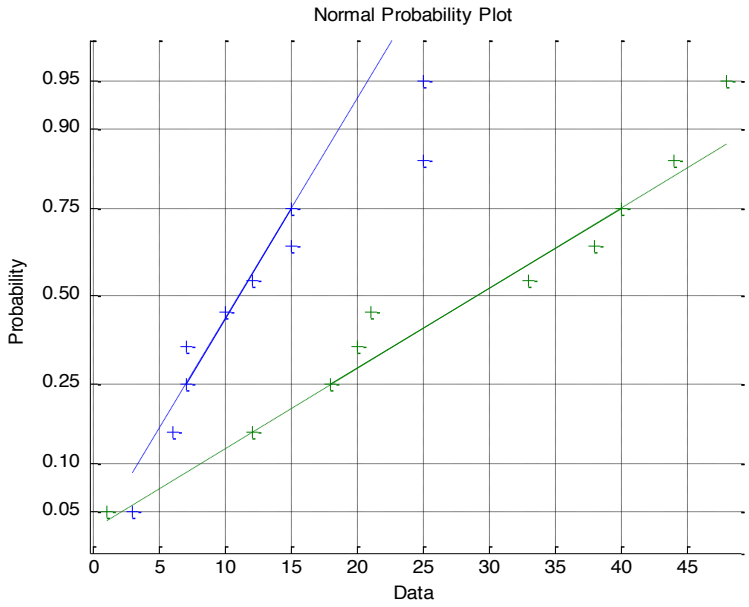
ničelno hipotezo zavržemo. Torej sklepamo, da v povprečju koncentraciji arzena v mestnih in urbanih predelih dejansko nista enaki, kot smo tudi predvidevali.

Slika 183 prikazuje takoimenovan ***Normal probability plot*** za obe koncentraciji. Iz tovrstnega grafa se ugotavlja, če je nek vzorec normalno porazdeljen. Če je, se podatki približno ujemajo s premico, sicer pa ne. Iz slike 183 je tudi razvidno, da se koncentraciji v lastnostih precej razlikujeta med seboj.



Slika 183: ***Normal probability plot*** *za obe koncentraciji (PHX - mestni predeli, Rural AZ - ruralni predeli) (x - koncentracija arzenika v delcih na milijardo, y - procenti)[Montgomery 1].*

V Matlabu bi imela slika 183 izgled, kot ga prikazuje slika 184.



Slika 184: ***Normal probability plot*** *za obe koncentraciji v Matlabu.*

Za izračune lahko uporabimo naslednji program v Matlabu:

```
% pogl8_8_6.m

clear
clc
close all

% podatki:
```

```
metro_phx = [3 7 25 10 15 6 12 25 15 7];
rural_phx = [48 44 40 38 33 21 20 12 1 18];


% normal probability plot:

normplot([metro_phx(:), rural_phx(:)])


n1 = length(metro_phx); % stevilo meritev metro
n2 = length(rural_phx); % stevilo meritev rural

m1 = mean(metro_phx) % izracunano povprecje: metro
m2 = mean(rural_phx) % izracunano povprecje: rural

s1 = std(metro_phx) % izracunan std: metro
s2 = std(rural_phx) % izracunan std: rural



t0 = (m1 - m2) / sqrt(s1^2/n1 + s2^2/n2) % testna statistika

nu = (s1^2/n1 + s2^2/n2)^2 / ((s1^2/n1)^2/(n1-1) + (s2^2/n2)^2/(n2-1)) % stevilo
prostostnih stopenj
nu = round(nu)

OZ = tinv([0.025, 0.975], nu) % meje za zavrnitev H0 pri alfa = 0.05


if (t0<OZ(1))||(t0>OZ(2))
    disp('zavrni nicelno hipotezo')
else
    disp('sprejmi nicelno hipotezo')
end
```

Izpis komandnega okna je naslednji:

```
m1 =
  12.5000


m2 =
  27.5000


s1 =
   7.6340
s2 =
  15.3496


t0 =
  -2.7669


nu =
  13.1956
nu =
   13


OZ =
  -2.1604   2.1604
```

zavrni nicelno hipotezo

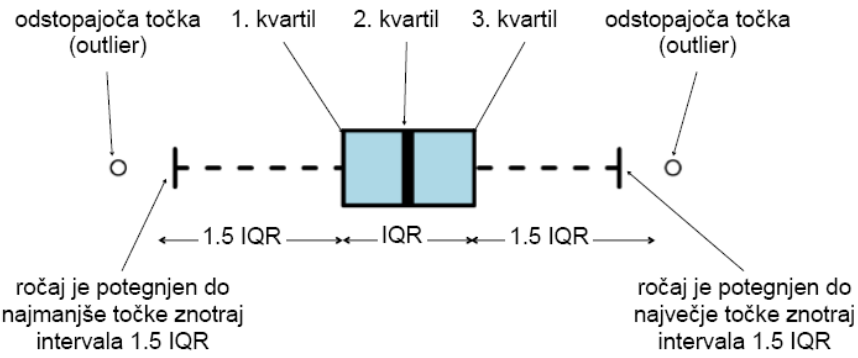
Izračune bi lahko opravili tudi s standardno funkcijo v Matlabu **ttest2**. Klic funkcije in izpis v komandnem oknu bi bila naslednja:

```
>> [h, p, ci, stat] = ttest2(metro_phx, rural_phx, 0.05, 'both', 'unequal')


h =
    1
p =
   0.0158
ci =
  -26.6941   -3.3059

stat =
      tstat: -2.7669
      df: 13.1956
      sd: [7.6340 15.3496]
```

Kot lahko vidimo, dobimo iste rezultate. Ker je *h=1*, se zavrne ničelna hipoteza. Vidimo tudi, da funkcija ttest2 da intervalno oceno (ci) za neznano razliko $\mu = \mu_1 - \mu_2$. Slednje sicer v tem gradivu nismo izpeljali, jo pa lahko bralec najde v ustrezni literaturi.

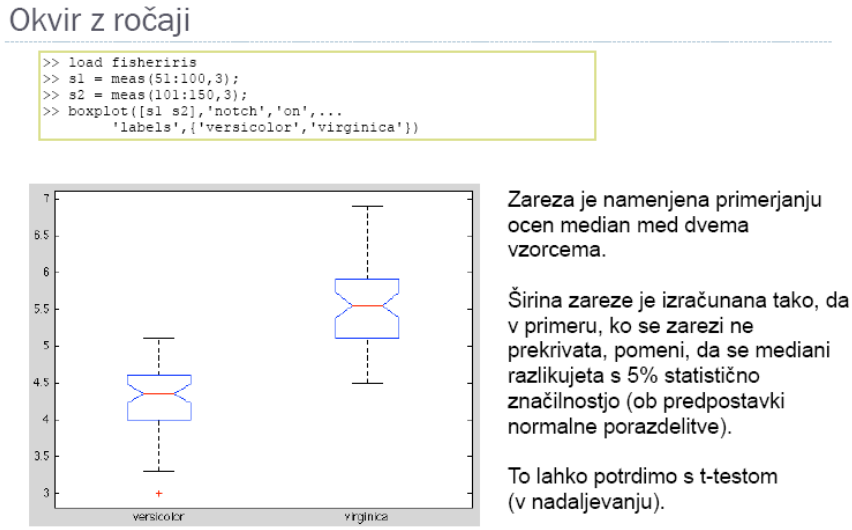
Okvir z ročaji

Okvir z ročaji (**boxplot**) ima v splošnem izgled, kot ga prikazuje slika 185 [Žibert]. Gre za postopek v opisni statistiki, kjer prikazujemo numerične podatke s stališča njihovih kvartilov. Tovrstni okvirji merijo diference med vzorci (ali populacijami) brez predpostavke o njihovi porazdelitvi (**neparametrični** pristop). Govorijo tudi o stopnji variabilnosti oz. disperzije in sploščenosti podatkov, in identificirajo osamelce. Meja (zareza) znotraj ročaja pomeni drugi kvartil (mediano).



Slika 185: Okvir z ročaji (boxplot) [Žibert]

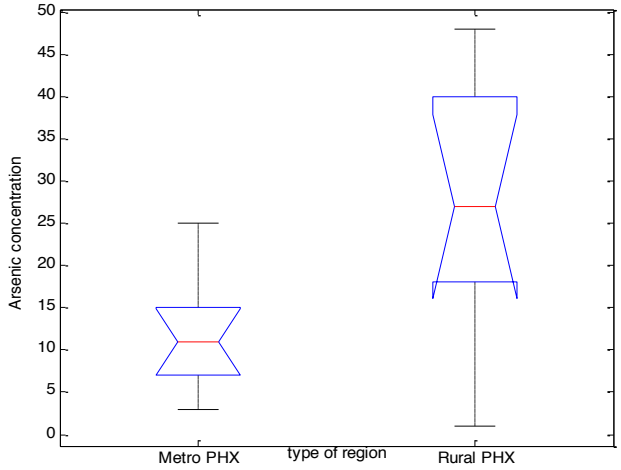
Odstopajoče točke (osamelci oz. outliers) so vsi podatki, ki padejo izven področja, ki ga določa debelina ročajev. Širina zareze je izračunana tako (glej sliko 186), da v primeru, ko se zarezi ne prekrivata, to pomeni, da se mediani med seboj razlikujeta s 5% statistično značilnostjo (pri normalni porazdelitvi) [Žibert].



Slika 186: Primer primerjave median z okvirjema z ročaji [Žibert]

V našem primeru koncentracij arzena v mestnih oz. ruralnih predelov bi uporabili naslednje ukaze v Matlabu za primerjavo obeh okvirjev z ročaji (glej sliko 187):

```
% pogl_8_8_6a.m

metro_phx = [3 7 25 10 15 6 12 25 15 7];
rural_phx = [48 44 40 38 33 21 20 12 1 18];

boxplot([metro_phx(:), rural_phx(:)],1)
set(gca,'XTick',[1 2])
set(gca,'XtickLabel',{'Metro PHX','Rural PHX'})
xlabel('type of region')
ylabel('Arsenic concentration')
```

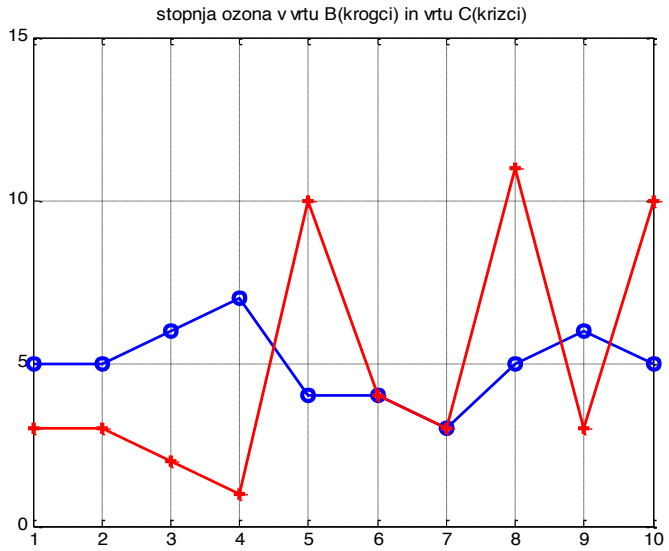


Slika 187: Primerjava median z okvirjema z ročaji v primeru koncentracije arzena v mestnih oz. ruralnih predelov Arizone [Žibert]

Tudi slika 187 nam nakazuje, da se koncentraciji v lastnostih precej razlikujeta med seboj.

Primer 8.29.:

Imamo primer, ko želimo primerjati koncentracijo ozona v vrtovih B in C. Gre za test enakosti oz. kvocienta varianc pri normalni porazdelitvi. Časovni vrsti obeh koncentracij prikazuje slika 188. Testirajte ničelno hipotezo: $H_0 : \sigma_1^2 = \sigma_2^2$ *ob nasprotni hipotezi* $H_1 : \sigma_1^2 \neq \sigma_2^2$ *[Žibert] (*$\alpha = 0.05$*).*



Slika 188: Časovni vrsti koncentracij ozona v vrtovih B in C

Imamo varianci S_1^2 in S_2^2, ki sta varianci obeh časovnih vrst velikosti n_1 in n_2 dveh neodvisnih populacij (za vrta B in C) z neznanima variancama σ_1^2 in σ_2^2 . Potem je testna statistika (glej (8.55) :

$$F_0 = \frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}} = \frac{\sigma_2^2}{\sigma_1^2} \cdot \frac{S_1^2}{S_2^2} \tag{8.204}$$

F naključna spremenljivka z $n_1 - 1$ in $n_2 - 1$ prostostnimi stopnjami. Testirati želimo hipotezi:

$$\begin{aligned} H_0 &: \sigma_1^2 = \sigma_2^2 \\ H_1 &: \sigma_1^2 \neq \sigma_2^2 \end{aligned} \tag{8.205}$$

Pri ničelni hipotezi je: $F_0 = \frac{\sigma_2^2}{\sigma_1^2} \cdot \frac{S_1^2}{S_2^2} = \frac{S_1^2}{S_2^2}$. Kritično območje je glede na nasprotne hipoteze določeno na naslednji način (glej (8.58)):

$$\left(\frac{s_1^2}{s_2^2} \geq f_{\frac{\alpha}{2}}\left(n_1 - 1, n_2 - 1\right) \right) \vee \left(\frac{s_1^2}{s_2^2} \leq f_{1-\frac{\alpha}{2}}\left(n_1 - 1, n_2 - 1\right) = \frac{1}{f_{\frac{\alpha}{2}}\left(n_2 - 1, n_1 - 1\right)} \right)$$

$$\left(\frac{s_1^2}{s_2^2} \geq f_{\frac{\alpha}{2}}\left(n_1 - 1, n_2 - 1\right) \right) \vee \left(\frac{s_2^2}{s_1^2} \geq f_{\frac{\alpha}{2}}\left(n_2 - 1, n_1 - 1\right) \right) \text{.....(dvorepi)}$$

ali (8.206)

$$\frac{s_1^2}{s_2^2} \leq f_{1-\alpha}\left(n_1 - 1, n_2 - 1\right) = \frac{1}{f_{\alpha}\left(n_2 - 1, n_1 - 1\right)}$$

$$\frac{s_2^2}{s_1^2} \geq f_{\alpha}\left(n_2 - 1, n_1 - 1\right) \text{.....(levi enorepi)}$$

ali

$$\frac{s_1^2}{s_2^2} \geq f_{\alpha}\left(n_1 - 1, n_2 - 1\right) \text{.....(desni enorepi)}$$

pri čemer seveda gledamo izraz za dvorepi test.

Intervalna ocena je določena z izrazom (glej (7.109)):

$$P\left(\frac{\frac{S_1^2}{S_2^2}}{f_{\frac{\alpha}{2}}\left(n_1-1,n_2-1\right)}\le\frac{\sigma_1^2}{\sigma_2^2}\le\frac{\frac{S_1^2}{S_2^2}}{f_{1-\frac{\alpha}{2}}\left(n_1-1,n_2-1\right)}\right)=1-\alpha \tag{8.207}$$

Vrednost testne statistike je enaka:

$$\begin{aligned} s_1^2 &= 1.3333 \\ s_2^2 &= 14.2222 \\ F_0 &= \frac{s_1^2}{s_2^2}=\frac{1.3333}{14.2222}=0.0938 \end{aligned} \tag{8.208}$$

Območje zaupanja OZ za testno statistiko je enako:

$$OZ=\left[0.2484,4.0260\right] \tag{8.209}$$

Ker velja:

$$F_0=0.0938\notin OZ=\left[0.2484,4.0260\right] \tag{8.210}$$

ničelno hipotezo zavržemo. Torej sklepamo, da varianci koncentracij ozona v vrtovih B in C nista enaki.

Verjetnost zaupanja v rezultat (p-vrednost oz p-value) je enaka:

$$p=0.0016<\alpha=0.05 \tag{8.211}$$

Matlab ukaz za p-vrednost je bil:

```
p = fcdf(F0,n1-1,n2-1); % p-vrednost za obojestranski test
p = 2*min(p,1-p)
```

Intervalna ocena pride:

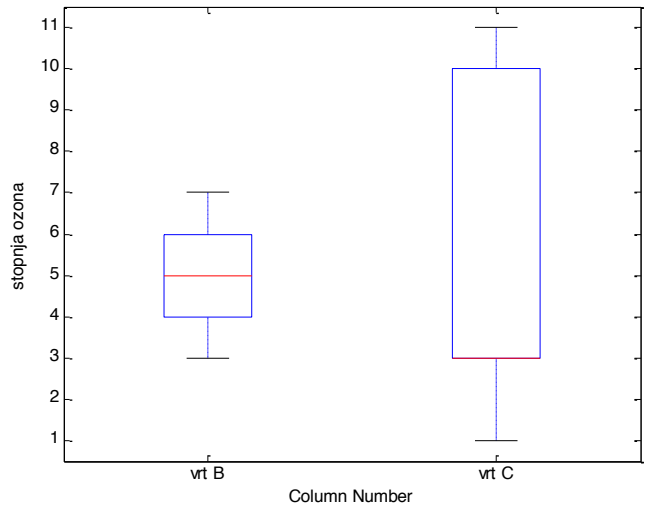
$$I = \left(\frac{\frac{s_1^2}{s_2^2}}{f_{\frac{\alpha}{2}}(n_1-1, n_2-1)}, \frac{\frac{s_1^2}{s_2^2}}{f_{1-\frac{\alpha}{2}}(n_1-1, n_2-1)} \right) \tag{8.212}$$

$$I = \left(\frac{\frac{1.3333}{14.2222}}{4.026}, \frac{\frac{1.3333}{14.2222}}{0.2484} \right) = \left(\frac{0.0938}{4.026}, \frac{0.0938}{0.2484} \right) = (0.0233, 0.3774)$$

$$\frac{\sigma_1^2}{\sigma_2^2} \in I = (0.0233, 0.3774)$$

in govori o tem, znotraj katerega intervala naj bi se nahajal kvocient neznanih varianc $\frac{\sigma_1^2}{\sigma_2^2}$.

Primerjavo median z okvirjema z ročaji za obe časovni vrsti prikazuje slika 189.



Slika 189: Primerjava median z okvirjema z ročaji za obe časovni vrsti

Tudi slika 189 nam nakazuje, da se varianci koncentracij ozona za obe časovni vrsti obeh vrtov v lastnostih precej razlikujeta med seboj.

Za izračune lahko uporabimo naslednji program v Matlabu:

```
% pogl_8_8_7.m

clear
clc
close all

data = importdata('f.test.data.txt')

X = data.data;

pod1 = X(:,1);
pod2 = X(:,2);

% Izris obeh casovnih brst (stopnji ozona v vrtih B in C):

plot(pod1,'LineWidth',1.5)
hold on
plot(pod1,'o','LineWidth',2.5)

plot(pod2,'r','LineWidth',1.5)
hold on
plot(pod2,'r+','LineWidth',2.5)

grid
title('stopnja ozona v vrtu B(krogci) in vrtu C(krizci)')
d = axis;
axis([d(1) d(2) 0 15])


% Izris okvirjev z rocaji za obe casovni vrsti:

figure
boxplot(X)
set(gca,'XTick',[1 2])
set(gca,'XtickLabel',{'vrt B','vrt C'})
ylabel('stopnja ozona')

% Ocena varianc za obe casovni vrsti

disp('Ocena varianc za obe casovni vrsti:')

V1 = var(X(:,1)) % ocenjena varianca vrta B
V2 = var(X(:,2)) % ocenjena varianca vrta C

n1 = length(X(:,1)); % stevilo meritev vrt B
n2 = length(X(:,2)); % stevilo meritev vrt C

% Test hipoteze o enakosti varianc obeh casovnih vrst

disp('Vrednost testne statistike:')

F0 = V1 / V2 % testna statistika

alf = 0.05; % stopnja zavrnitve

disp('Območje zaupanja za testno statistiko')

meje = finv([alf/2, 1-alf/2], n1-1, n2-1) %mejne vrednosti zavrnitve H0

if (F0<meje(1))||(F0>meje(2))
    disp('zavrni nicelno hipotezo')
else
    disp('sprejmi nicelno hipotezo')
```

```
end

p = fcdf(F0,n1-1,n2-1); % p-vrednost za obojestranski test
p = 2*min(p,1-p)

% Intervalna ocena je:

disp('Intervalna ocena je:')

I = [V1/V2/meje(2) V1/V2/meje(1)]
```

Izpis komandnega okna je naslednji:

```
data =
        data: [10x2 double]
        textdata: {'gardenB'  'gardenC'}
        colheaders: {'gardenB'  'gardenC'}

Ocena varianc za obe casovni vrsti:
V1 =
   1.3333
V2 =
  14.2222

Vrednost testne statistike:
F0 =
   0.0938

Območje zaupanja za testno statistiko
meje =
   0.2484    4.0260

zavrni nicelno hipotezo

p =
   0.0016

Intervalna ocena je:
I =
   0.0233    0.3774
```

Izračune bi lahko opravili tudi s standardno funkcijo v Matlabu **vartest2**. Klic funkcije in izpis v komandnem oknu bi bila naslednja:

```
>> [h,p,ci, stat] = vartest2(X(:,1), X(:,2), 0.05, 'both')


h =
    1
```

```
p =
    0.0016

ci =
    0.0233
    0.3774

stat =
    fstat: 0.0938
    df1: 9
    df2: 9
```

Kot lahko vidimo, dobimo iste rezultate. Ker je $h=1$, se zavrne ničelna hipoteza.

9 KORELACIJA IN REGRESIJA

9.1 Uvod

Pri statističnih proučevanjih nas večkrat zanimajo zveze med pojavi, ki jih proučujemo. Te zveze nam omogočajo izraziti en pojav ali več pojavov z ostalimi. Za primer vzemimo, da nas zanima povezanost vlaganj v reklamo z doseženo prodajo. Poznavanje te zveze bi nam omogočilo predvideti prodajo pri izbranem vložku v reklamo [Jesenko]. Podobno bi nas lahko zanimalo, kako število dni, v katerih smo bili na dieti, vpliva na izgubo telesne teže. Na osnovi te zveze bi lahko predvideli, za koliko kilogramov bomo shujšali po določenem številu dni diete. Idealno bi bilo, če bi medsebojne odvisnosti pojavov poznali povsem natančno, kajti to bi nam omogočalo natanko predvideti, kako stanje enega pojava vpliva na določeno stanje drugega pojava, kar pa je le redko možno [Jesenko]. Zato se največkrat zadovoljimo, če nam uspe **predvideti le povprečne vrednosti stanj pojavov**. Ne morerno na primer predvideti povsem natančno, koliko bo nek študent po 8 letih po diplomiranju na neki fakulteti zaslužil, pač pa lahko na osnovi primerno izbranih podatkov predvidimo, koliko bo po 8 letih povprečno zaslužil diplomant dotične fakultete [Jesenko].

V statistiki se pri preučevanju množičnih pojavov običajno srečujemo z dvema vrstama spremenljivk. Na eni strani imamo spremenljivke, ki jih lahko opazujemo in merimo z merilnimi inštrumenti ali štejemo (število požarov na časovno enoto, izpostavljenost, zaprašenost, količina nevarne snovi v prostoru, število zaposlenih, delovni čas, osvetljenost, itn), imamo pa tudi spremenljivke, katerih vrednosti npr. dobimo s pomočjo vprašalnikov (znanje, pričakovana reakcija na določen dogodek, odnos do posameznih pojavov, izobrazba, kvalifikacija, itn.). Spremenljivke lahko merimo s poljubnimi fizikalnimi ali drugačnimi enotami (točkovanje). Pri tem v splošnem ločimo tri tipe spremenljivk:

- **urejenostne (ordinalne) spremenljivke** (vrednosti omogočajo kvečjemu ureditev enot po velikosti, npr., ocena čistoče, ocena vzdrževanja naprav),
- **imenske (nominalne) spremenljivke** (vrednosti omogočajo le razlikovanje z enakostjo ali neenakostjo med seboj, npr. vrsta dejavnosti),

- **razmernostne spremenljivke** (vrednosti omogočajo tudi primerjavo razmerij med vrednostima dvojic).

Kadar govorimo o enem samem pojavu, ki je odvisen od ene same spremenljivke, je pojav opisan, če lahko najdemo povezavo med opazovano spremenljivko in pojavom. Takšne spremenljivke imenujemo **merjene spremenljivke, indikatorji ali kazalci**. Stanje, ki vpliva na te kazalce, imenujemo **latentna spremenljivka**. Vsako stanje z neznanimi odvisnostmi vpliva na svoje indikatorje. Npr., na požarno varnost v objektu vplivajo vgrajeni varovalni sistemi, poučenost o nevarnostih, usposobljenost za reakcije ob nevarnostih, itn. Vsakega od teh indikatorjev lahko na delovnem mestu izmerimo, vendar nam te količine še nič ne povedo o stopnji varnosti v objektu. Presoja stopnje požarne varnosti prešteva zgolj prekoračitve sprejemljivih vrednosti indikatorjev in ne upošteva moči vpliva posameznega indikatorja. Spremenljivke, ki spremljajo analizo varnosti, sestavljajo sistem, ki ga je potrebno obravnavati kot celoto in ne zgolj kot singularnosti posameznih spremenljivk.

Z **regresijsko in korelacijsko analizo** ugotavljamo medsebojno odvisnost med dvema ali več skupinami spremenljivk [Bastič]. S **korelacijsko analizo** ugotavljamo **jakost odvisnosti**, **z regresijsko analizo** pa je mogoče odvisnost med odvisno in eno (ali več) neodvisnimi spremenljivkami **izraziti v obliki regresijske enačbe** [Bastič]. Korelacijska analiza (kakor tudi noben drugi matematični postopek) pa **ne omogoča ugotavljanja vzročnosti**. Le–to je mogoče ugotavljati na osnovi poznavanja pojavov oz. študija relevantne teorije [Bastič].

Študij odvisnosti med eno odvisno in eno neodvisno spremenljivko (**enostavna regresija**) je najenostavneje pričeti s prikazom dvojic vrednosti obeh spremenljivk v **razsevnem grafikonu** (angl. **scatter diagram**). Ta omogoča ugotoviti obliko, smer in jakost odvisnosti [Bastič]. Oblika je lahko linearna ali krivuljčna, smer je lahko pozitivna (z naraščanjem vrednosti neodvisne spremenljivke naraščajo tudi vrednosti odvisne) ali negativna, glede na jakost pa je lahko bolj ali manj močna [Bastič].

0 regresiji torej govorimo, kadar sta dva ali več pojavov (veličin) v medsebojni odvisnosti. Regresija je enostavna, kadar nastopata v medsebojni odvisnosti samo dva pojava (veličini), kadar pa nastopa v medsebojni odvisnosti več pojavov, govorimo o **večkratni ali multipli regresiji** [Jesenko]. Tako bi na primer za količino kmetijskega

pridelka neke kulture ugotovili, da je le-ta lahko odvisna od kakovosti zemlje, porabljene količine gnojila, vremenskih pogojev, kakovosti semena in drugih dejavnikov. Vpliv navedenih dejavnikov na pridelano količino ni nujno istosmeren niti enako intenziven, zato je koristno, da z regresijsko
analizo ugotovimo značilnosti delovanja raznih dejavnikov, da bi v kombinaciji z uporabo drugih metod kolikostne ekonomske analize lahko dosegli ekonomsko gledano najboljši učinek [Artenjak].

Naloga regresije je, poiskati takšno funkcijo $y = f(x)$, ki najbolje podaja medsebojno odvisnost pojavov. Odvisnost je enostranska $X \to Y$, kadar je veličina X vzrok, veličina Y pa posledica [Jesenko]. Vzemimo za primer čas, potreben za to, da preberemo neko število strani v knjigi. Očitno je vzrok število strani v knjigi, posledica pa čas branja. Pri iskanju funkcije $y = f(x)$, bi bil vzrok neodvisna, posledica pa odvisna spremenljivka [Jesenko].

Odvisnost je dvostranska $X \leftrightarrow Y$, kadar ni možno določiti, kaj je vzrok in kaj posledica. Za primer vzemimo medsebojno odvisnost velikosti bratov in sester. Vzemimo, da smo z opazovanji ugotovili, da med tema pojavoma obstaja neka določena odvisnost v smislu, da večji kot je brat, večja je tudi sestra in obratno. V tem primeru pa ni možno določiti vzroka in posledice. Regresijska analiza se prav zato ponavadi omejuje le na enostranske odvisne pojave [Jesenko]. Osnovni problem, ki ga pri enostavni regresiji rešujemo, je, kako pri dani realizaciji pojava $X = x$ najti funkcijo, ki bi določala pripadajočo realizacijo pojava $Y = y$. Formalno gledano, sta količini X in Y naključni spremenljivki, zato njunih vrednosti ne moremo vnaprej natanko predvideti. Vrednosti spremenljivke Y torej ne moremo vnaprej natanko predvideti, določimo pa lahko njeno matematično upanje. Dejanske vrednosti pa "nihajo" okrog matematičnega upanja v skladu s porazdelitvenim zakonom naključne spremenljivke Y [Jesenko].

V splošnem lahko zapišemo:

$$\begin{aligned} y(x) &= f(x) + \varepsilon \\ Y(X) &= f(X) + \varepsilon \end{aligned} \qquad (9.1)$$

kjer predstavlja ε naključne vplive. Pri regresijski analizi ponavadi predpostavljamo, da je ε normalna naključna spremenljivka z matematičnim upanjem 0 in nekim standardnim odklonom, torej [Jesenko]:

$$\begin{aligned} &\varepsilon \in N(0,\sigma) \\ &E(\varepsilon) = 0 \\ &STD(\varepsilon) = \sigma \end{aligned} \tag{9.2}$$

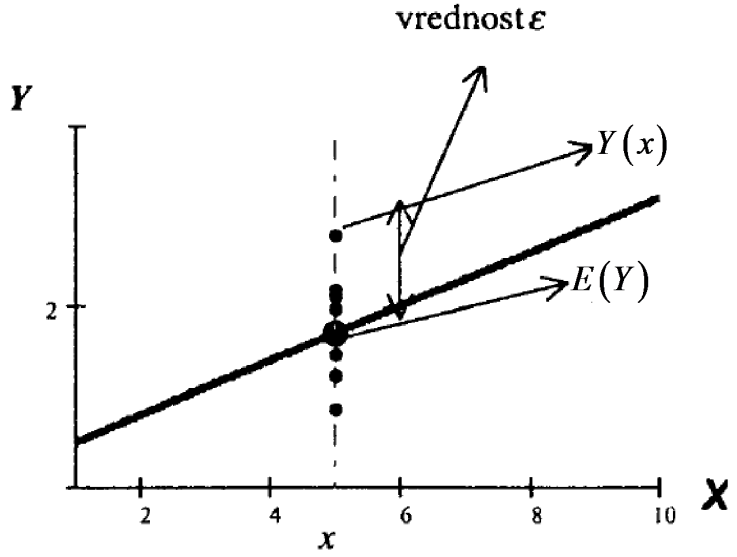
Sledi:

$$E(Y) = E\left(f(X)\right) + E(\varepsilon) = E\left(f(X)\right) + 0 = E\left(f(X)\right) = f(X) \tag{9.3}$$

Če vstavimo ta izraz v izraz (9.1), dobimo:

$$\begin{aligned} &Y(x) = f(X) + \varepsilon \\ &\varepsilon = Y(x) - f(X) = Y(x) - E(Y) \end{aligned} \tag{9.4}$$

Realizacije naključne spremenljivke ε torej lahko obravnavamo kot odmike realiziranih vrednosti naključne spremenljivke (pojava) Y od njenega matematičnega upanja $E(Y)$ [Jesenko]. Naključno spremenljivko ε imenujemo tudi **napaka modela**, samemu modelu (9.1) pa pravimo **regresijski model** [Jesenko] (glej sliko 190).



Slika 190: Regresijski model [Jesenko]

9.2 Korelacijska in regresijska povezanost

Če je spremenljivka *y* funkcijsko povezana s *k* neodvisnimi spremenljivkami, to na splošno zapišemo v obliki (**korelacijska odvisnost**) [Artenjak]:

$$y = f\left(x_1, x_2, ..., x_k\right) + \varepsilon \tag{9.5}$$

Če imamo linearno odvisnost ene odvisne spremenljivke od več neodvisnih spremenljivk, sledi:

$$y = a_0 + a_1 \cdot x_1 + a_2 \cdot x_2 + ... + a_k \cdot x_k + \varepsilon \tag{9.6}$$

Tu je z vidika ocenjevanja **regresijskih koeficientov** $a_j, j = 1, ..., k$, pomembna delitev spremenljivk na **regresande** (**odvisne spremenljivke, pojasnjene spremenljivke, endogene spremenljivke**) in **regresorje** (**pojasnjevalne spremenljivke, neodvisne spremenljivke, eksogene spremenljivke**), ki so opredeljeni s koeficienti pri posameznih neodvisnih spremenljivkah ali kombinacijah neodvisnih spremenljivk [Artenjak]. Zato je

število regresorjev lahko manjše, enako ali večje od števila neodvisnih spremenljivk. Velikokrat se zgodi, da so neodvisne spremenljivke obenem tudi regresorji.

Če dobimo z merjenjem *N* meritev odvisne in neodvisnih spremenljivk, sledi:

$$y(i) = a_0 + a_1 \cdot x_1(i) + a_2 \cdot x_2(i) + ... + a_k \cdot x_k(i) + \varepsilon(i), \quad i = 1, 2, ..., N \tag{9.7}$$

V izrazu (9.7) je število regresorjev enako številu neodvisnih spremenljivk (*k*) in tudi regresorji so enaki neodvisnim spremenljivkam [Artenjak]. a_0 je konstanta, regresijski koeficienti $a_j, j = 1, ..., k$ pa pokažejo, za koliko enot se v povprečju spremeni vrednost odvisne spremenljivke, če se j-ta neodvisna spremenljivka spremeni za enoto, vrednosti ostalih spremenljivk pa ostanejo nespremenjene. Zapis (9.7) prikazuje populacijsko regresijsko odvisnost, za razliko od funkcijske odvisnosti, kjer ni člena napak (odklonov) $\varepsilon(i)$.

Če bi npr. postavili, da je izpitna ocena odvisna izključno od časa študija, potem bi študentje z enako dolžino časovne priprave dobili enako oceno *(funkcijska odvisnost)*. Dejansko pa bi ocena teh študentov lahko bila različna zaradi delovanja drugih dejavnikov, to je na primer različnih vedenjskih značilnosti posameznih študentov, predhodne šolske izobrazbe itn *(korelacijska odvisnost)* [Artenjak].

Izraz (9.7) lahko zapišemo tudi v vektorsko-matrični obliki. V ta namen tvorimo enačbe:

$$\begin{aligned}
&y(1) = a_0 + a_1 \cdot x_1(1) + a_2 \cdot x_2(1) + ... + a_k \cdot x_k(1) + \varepsilon(1) \\
&y(2) = a_0 + a_1 \cdot x_1(2) + a_2 \cdot x_2(2) + ... + a_k \cdot x_k(2) + \varepsilon(2) \\
&... \\
&y(N) = a_0 + a_1 \cdot x_1(N) + a_2 \cdot x_2(N) + ... + a_k \cdot x_k(N) + \varepsilon(N)
\end{aligned} \tag{9.8}$$

Sledi:

$$\underbrace{\begin{bmatrix} y(1) \\ y(2) \\ \dots \\ y(N) \end{bmatrix}}_{\mathbf{y}} = \underbrace{\begin{bmatrix} 1 & x_1(1) & x_2(1) & \dots & x_k(1) \\ 1 & x_1(2) & x_2(2) & \dots & x_k(2) \\ 1 & \dots & \dots & \dots & \dots \\ 1 & x_1(N) & x_2(N) & \dots & x_k(N) \end{bmatrix}}_{\boldsymbol{\Psi}} \cdot \underbrace{\begin{bmatrix} a_0 \\ a_1 \\ \dots \\ a_k \end{bmatrix}}_{\boldsymbol{\Theta}} + \underbrace{\begin{bmatrix} \varepsilon(1) \\ \varepsilon(2) \\ \dots \\ \varepsilon(N) \end{bmatrix}}_{\boldsymbol{\varepsilon}} \tag{9.9}$$

Vrste korelacij

Korelacijo med dvema pojavoma lahko razlikujemo po [Nemec]:

- smiselnosti,
- smeri,
- obliki.

Ko opazujemo korelacijo po smiselnosti, ugotavljamo medsebojen vpliv posameznih pojavov. Ob tem ločimo naslednje primere [Nemec]:

Vzročna (enosmerna) korelacija: O vzročni korelaciji govorimo, ko je en pojav vzrok, drugi pa posledica. Velikost posledičnega pojava je odvisna od velikosti vzročnega pojava. Tako npr. lahko ugotovimo, da je število škodljivcev, ki prežive zimo, odvisno od temperaturnih razmer, in ne obratno. Nizke zimske temperature vplivajo na število prezimljenih škodljivcev.

Čista (dvosmerna) korelacija: S čisto korelacijo imamo opravka tedaj, ko ni vnaprej določeno, kateri izmed obeh pojavov je vzrok in kateri posledica. Čisto korelacijo srečamo npr. tam, kjer imamo opravka s porabo substitucijskih dobrin. Če opazujemo prodajo surovega masla in margarine, lahko ugotovimo, da se ob večji prodaji surovega masla zmanjšuje prodaja margarine. Trdimo pa lahko tudi obratno, ob večji prodaji margarine se zmanjšuje prodaja surovega masla. Katerega izmed obeh pojavov bomo obravnavali kot vzrok in katerega kot posledico, je odvisno od namena in naročnika raziskave [Nemec]. V primeru, da bo svojo prodajo analiziral proizvajalec surovega masla, bo opazoval, kako je odvisna prodaja surovega masla od prodaje margarine. V tem primeru bo neodvisna spremenljivka prodaja

margarine. Če pa bo naročnik raziskave proizvajalec margarine, bo opazoval, kako je odvisna prodaja margarine od prodaje surovega masla. Zato bo v tem primeru neodvisna spremenljivka prodaja surovega masla [Nemec].

Avtokorelacija: O avtokorelaciji govorimo, če **vpliva pojav sam nase**. Avtokorelacijo srečamo npr. pri časovnih vrstah. Tu je velikost pojava v nekem časovnem obdobju odvisna tudi od velikosti pojava v predhodnem obdobju. Tako je npr. velikost osnovne črede goveje živine v tekočem letu odvisna od velikosti osnovne črede v preteklem obdobju, saj je povečanje črede predvsem odvisno od reprodukcijskih sposobnosti živali v čredi [Nemec].

Ko opazujemo korelacijo **po smeri**, ugotavljamo spreminjanje velikosti odvisnega pojava, če se spreminja velikost neodvisnega pojava. Ob tem ima korelacija pozitivno smer, če se ob povečanju vrednosti neodvisnega pojava poveča tudi vrednost odvisnega pojava. Npr., ob večjem številu sončnih dni je stopnja sladkorja v grozdju višja. Pri negativni smeri se ob povečanju vrednosti neodvisnega pojava vrednost odvisnega pojava zmanjša. Npr., ob povečanju prodaje margarine se zmanjša prodaja surovega masla [Nemec].
Pri analizi korelacije moramo ugotoviti **smiselnost povezave, njeno obliko in jakost z namenom, da lahko pri znanih vrednostih neodvisne spremenljivke poiščemo oceno za vrednost odvisne spremenljivke**. [Nemec]

Statistično modeliranje in regresija

Statistično modeliranje se uporablja za določanje modelov iz vzorcev. Statističen model je običajno matematična funkcija (v primeru parametričnega modela), ki ji na podlagi vzorca določimo parametre, tako da se kar najbolje ujema s podatki v vzorcu [Žibert].

Statističen model uporabljamo [Žibert]:

- za bolj zgoščeno opisovanje podatkov iz vzorca,
- za napovedovanje dogodkov,
- za razvrščanje novih primerkov v različne populacije (razrede).

Zahteve pri statističnih modelih so [Žibert]:

- da se čim bolj natančno ujemajo s podatki, s katerih so ocenjeni,
- da imajo lastnost posploševanja,
- da so čim manj kompleksni.

Regresijska analiza je postopek določitve modelov, da se ujemajo s podatki [Žibert]:

- postopek modeliranja (regresije) je odvisen od modela,
- če je model parametričen, potem z regresijsko analizo ocenjujemo parametre modela,
- če je model linearen, potem s postopki linearne algebre lahko določamo parametre modela z minimizacijo napake ujemanja modela s podatki,
- če je model nelinearen in parametričen, uporabimo drugačne postopke iskanja optimalnih parametrov modela (nelinearna optimizacija).
- če je model neparametričen (npr. regresijsko drevo), uporabimo drugačne postopke.

Pri regresiji določamo odvisnost med odzivno spremenljivko Y in opisno spremenljivko X [Žibert]:

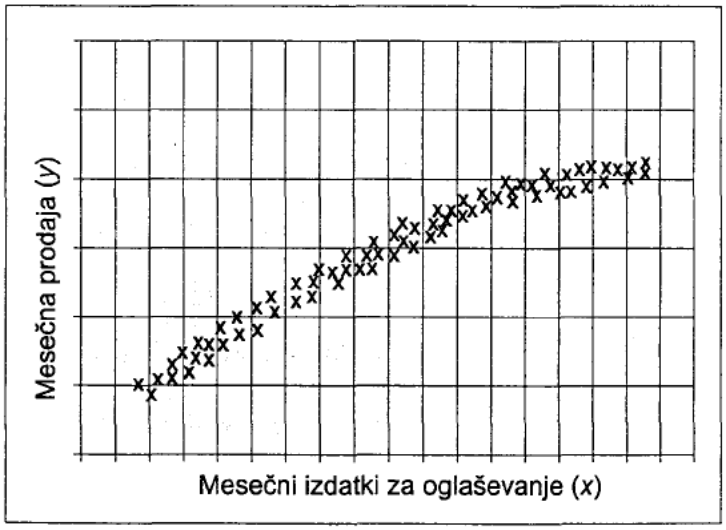
- Y in X sta zvezni spremenljivki (če je X diskretna (kategorijska) in Y zvezna spremenljivka, lahko npr. izvedemo analizo variance),
- če je X vektor, imamo **multiplo regresijo**,
- če je Y vektor, imamo **multivariatno regresijo**.

9.3 Povezanost med številskima spremenljivkama

V analiziranju povezanosti med dvema številskima spremenljivkama y in x je primerno, da množici vrednosti spremenljivk *prikažemo* grafično *s točkami* v *razsevnem diagramu,* ki je lahko zelo koristen pripomoček za odločanje o tem, kako naj analitično nadaljujemo, če je to sploh smotrno. Majhno število točk v razsevnem diagramu pri tem sicer ne bo v veliko pomoč pri statističnem sklepanju, vendar pa imamo v večini primerov veliko

število opazovanj, zato nam razsevni diagram lahko prikaže zakonitost povezanosti med spremenljivkama [Artenjak].

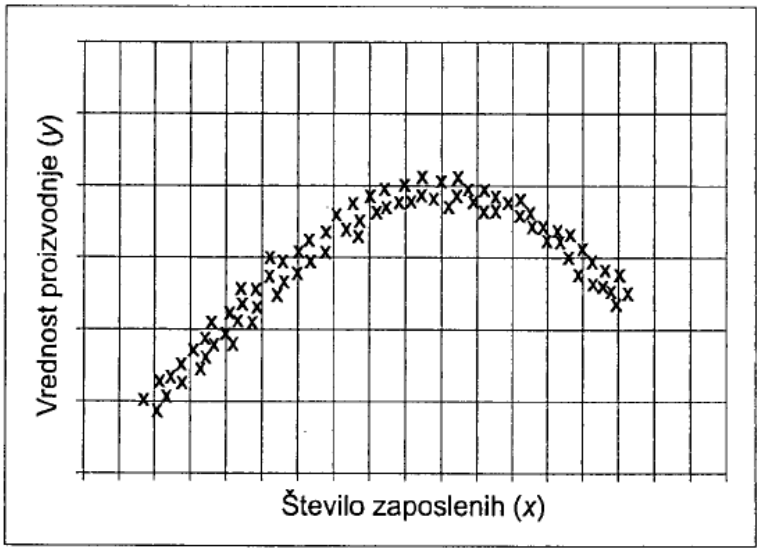
Slika 191 nam prikazuje razsevni diagram mesečne prodaje v odvisnosti od mesečnih izdatkov za oglaševanje [Artenjak].



Slika 191: Razsevni diagram mesečne prodaje v odvisnosti od mesečnih izdatkov za oglaševanje [Artenjak].

S slike 191 lahko razberemo, da se s povečanjem izdatkov za televizijsko oglaševanje povečuje na splošno tudi prodaja. Ta povezanost ni funkcijska, niti ne pove, kaj je vzrok in kaj je posledica. Velja pa ugotovitev, da je v mesecih z visoko prodajo običajno tudi visok izdatek za televizijsko reklarniranje prodaje. Šele v povezavi z ekonomsko teorijo (teorija marketinga) na podjetniški ravni velja spoznanje, da je raven prodaje posledica ravni izdatkov za televizijsko oglaševanje (vzrok) [Artenjak].

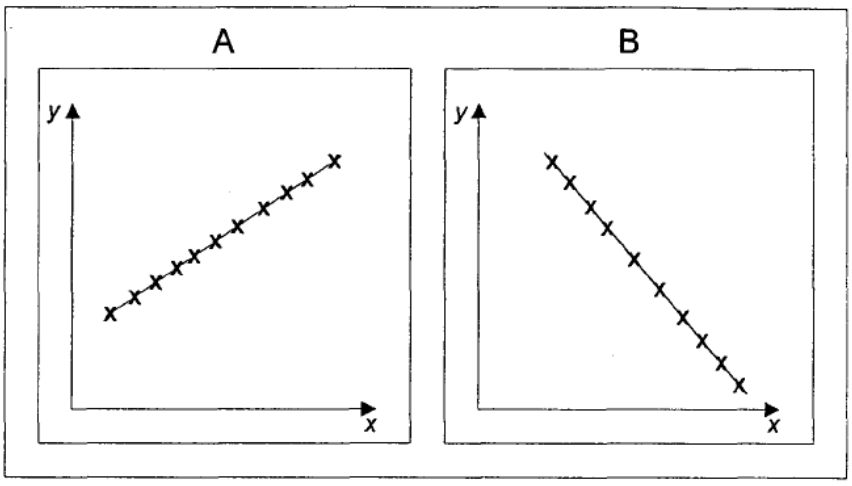
Slika 192 nam prikazuje razsevni diagram vrednosti proizvodnje v odvisnosti od števila zaposlenih [Artenjak].



Slika 192: Razsevni diagram vrednosti proizvodnje v odvisnosti od števila zaposlenih [Artenjak].

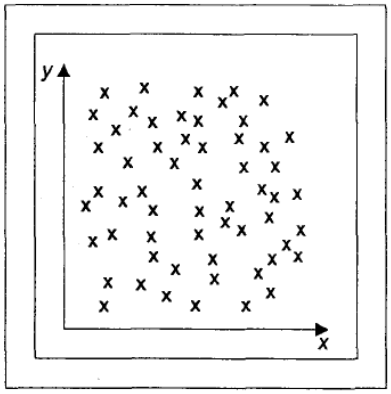
Na sliki 192 vidimo, kakšna je povezanost med številom zaposlenih in proizvedeno količino nekega izdelka v podjetju s fiksnim kapitalskim vložkom. Tu je oblika povezanosti bolj prepoznavna kot v prejšnjem primeru, ker so točke zgoščene v ožjem pasu, kot je to na sliki 191. Vzorec vrisanih točk odslikava prisotnost **zakona padajočega donosa**.[Artenjak]

Da bi lažje presojali o značilnostih povezanosti, si oglejmo še dva skrajna primera na sliki 193. Na sliki A in sliki B sta prikazani dve po smeri sicer različni, po obliki pa popolni povezavi, ker se vse točke nahajajo na premici. S povečevanjem vrednosti ene brez izjeme raste (A) oziroma pada (B) vrednost druge spremenljivke. To povečevanje oziroma zmanjševanje je linearno, zato tej vrsti povezanosti pravimo funkcijska in ne korelacijska povezanost [Artenjak].



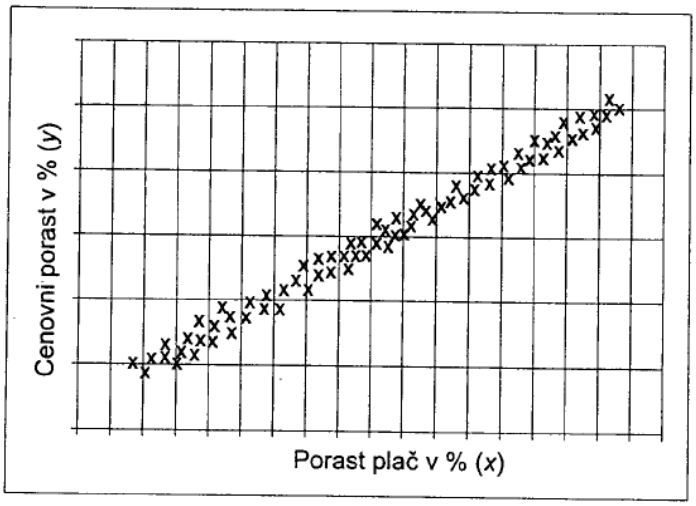
Slika 193: Dve po smeri sicer različni, po obliki pa popolni povezavi [Artenjak].

Kot nasprotje popolni povezavi na slikah A in B na sliki 193 lahko prikažemo primer, ko za spremenljivki *y* in *x* pravimo, da sploh nista povezani (glej sliko 194).



Slika 194: Popolna nepovezanost dveh spremenljivk [Artenjak].

Tudi če na osnovi razsevnega diagrama ugotavljamo, da je prisotna močna linearna povezanost med spremenljivkama, kot to vidimo na sliki 195, to še ne pomeni, da smemo v našem primeru reči, da je rast plač vzrok za rast cen, to je inflacije. Lahko bi tudi trdili, da je rast cen vzrok za rast plač. V vsakem konkretnem primeru posebej je torej potrebno na podlagi izkustva in teoretičnih spoznanj določiti, ali sta spremenljivki po vsebini enosmemo ali tudi dvosmerno povezani [Artenjak].



Slika 195: Primer močne linearne povezanosti med spremenljivkama [Artenjak].

Če povzamerno, pri proučevanju povezanosti dveh spremenljivk govorimo o dvorazsežnostnih porazdelitvah, ki jih prikazujemo na tri načine [Artenjak]:

- v preglednici z vrsto parov vrednosti spremenljivk opazovanih enot, ali v korelacijski preglednici,
- s točkami v razsevnem diagramu,
- v funkcijski obliki z regresijsko krivuljo.

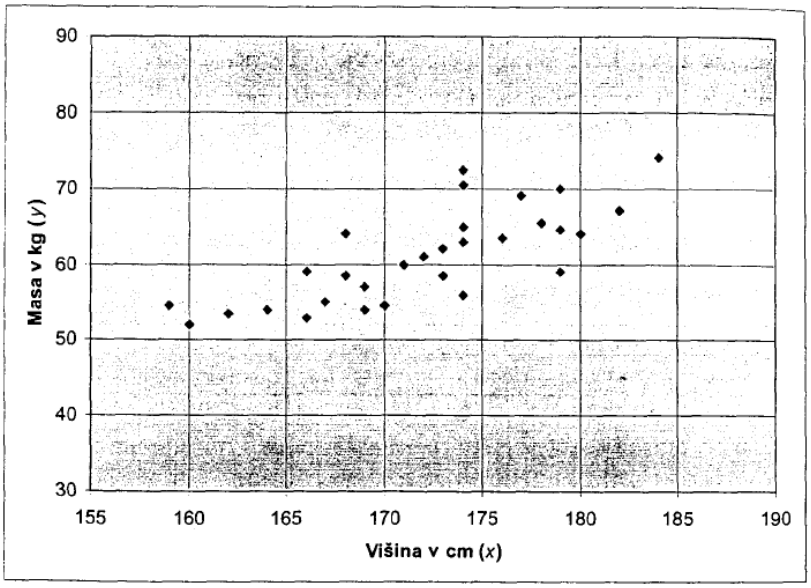
Primer 9.1.:

Imamo 30 študentov prvega letnika, za katere smo opazovali telesno višino in telesno maso. Rezultati meritev so prikazani na sliki 196. Narišite razsevni diagram! [Artenjak].

Zaporedna številka	Telesna masa v kg	Telesna višina v cm	Zaporedna številka	Telesna masa v kg	Telesna višina v cm
1	54,5	170	16	63,5	176
2	59	166	17	52	160
3	64	168	18	54	164
4	60	171	19	55	167
5	61	172	20	56	174
6	62	173	21	70	179
7	72,5	174	22	64	180
8	54	169	23	64,5	179
9	65	174	24	69	177
10	67	182	25	53,5	162
11	65,5	178	26	58,5	173
12	54,5	159	27	74	184
13	53	166	28	57	169
14	58,5	168	29	59	179
15	63	174	30	70,5	174

Slika 196: Tabela rezultatov meritev telesnih višin in mas za 30 študentov 1. letnika [Artenjak]

Razsevni diagram za tabelo na sliki 196 je prikazan na sliki 197. Glede na razpršenost točk ugotavljamo, da je oblika povezanosti lineama, smer je pozitivna, jakost pa srednje močna, ker so točke blizu namišljene regresijske funkcije. V tem primeru to pomeni, da so težji študentje v povprečju višji ali pa tudi, da so višji študentje v povprecju težji, saj za spremenljivki velja, da sta v dvosmemi odvisnosti.

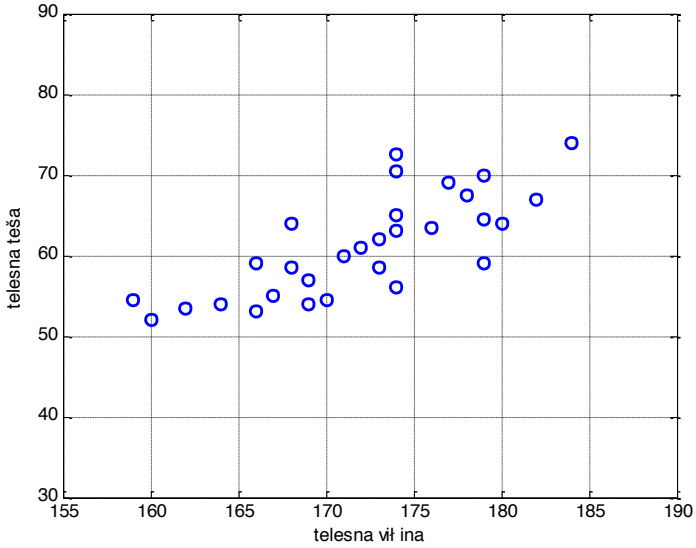


Slika 197: Razsevni diagram glede na tabelo rezultatov meritev telesnih višin in mas za 30 študentov 1. letnika [Artenjak]

Do slike 197 bi lahko prišli tudi z naslednjimi ukazi v Matlabu:

```
% razsevni.m

clear
clc
close all

%podatki:

x = [170  166  168  171  172  173  174  169  174  182  178  159  166  168  ...
     174  176  160  164  167  174  179  180  179  177  162  173  184  169  179  174 ];
y = [54.5 59.0 64.0 60.0 61.0 62.0 72.5 54.0 65.0 67.0 67.5 54.5 53.0 58.5 ...
     63.0 63.5 52.0 54.0 55.0 56.0 70.0 64.0 64.5 69.0 53.5 58.5 74.0 57.0 59.0 70.5];

plot(x,y,'o','LineWidth',2)
grid
xlabel('telesna višina')
ylabel('telesna teža')
d = axis;
axis([d(1) 190 30 90])
```

Dobili bi sliko 198. Uporabili pa bi lahko tudi standarden ukaz **scatter(…).**



Slika 198: Razsevni diagram glede na tabelo rezultatov meritev telesnih višin in mas za 30 študentov 1. letnika – z Matlabom

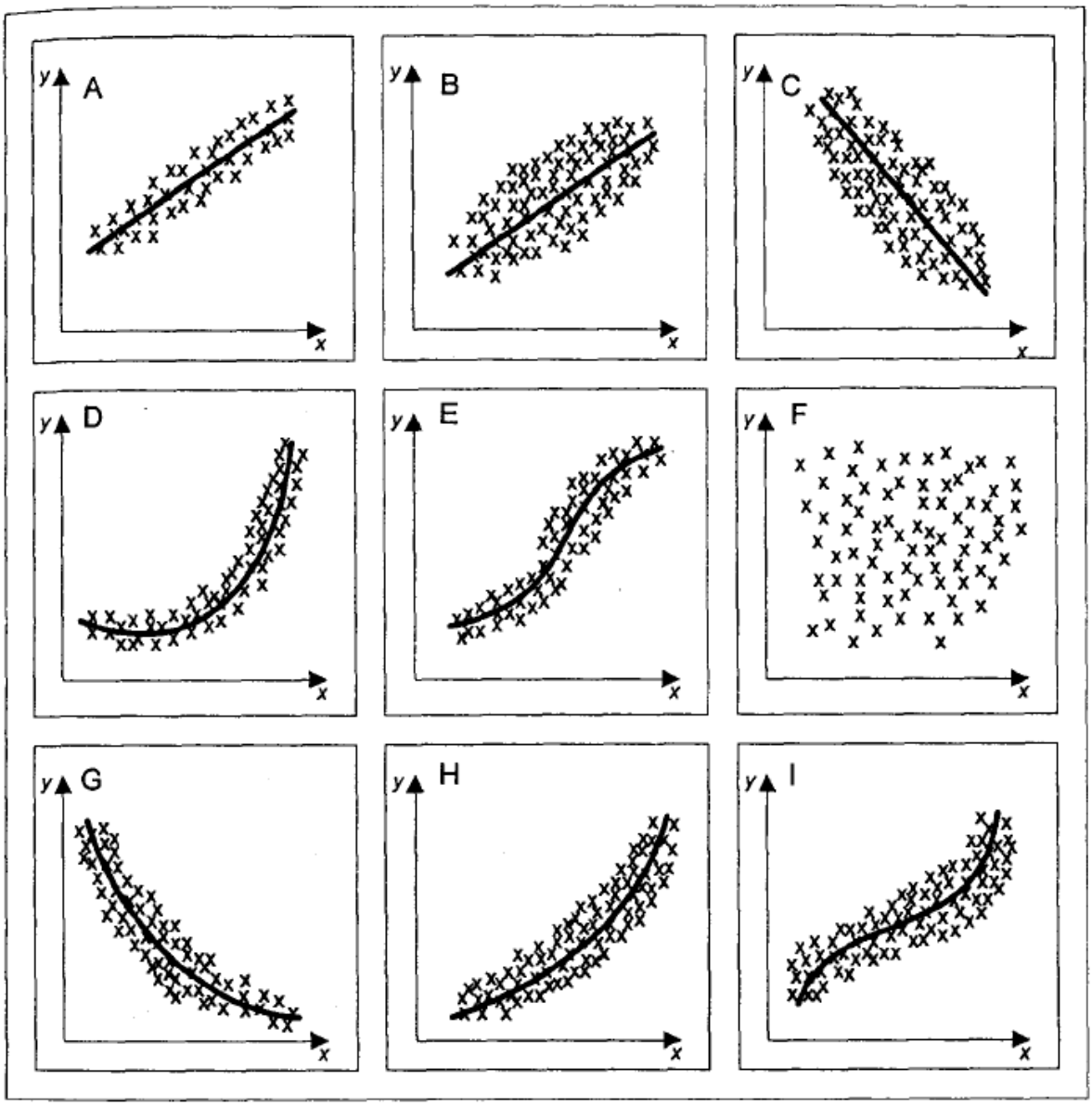
Grafično proučevanje povezanosti med dvema spremenljivkama, kot smo to doslej pokazali, je dokaj preprosto, dobimo pa uporabne podatke glede oblike, smeri in jakosti povezave. Na slikah razsevnih diagramov na sliki 199 si lahko pogledamo nekatere najbolj značilne primere te povezanosti [Artenjak].

Če v razsevne diagrame na sliki 199 (korelacijske grafikone) vrišemo ali pa si samo zamislimo krivuljo, ki se točkam čim bolje prilega, lahko ugotovimo naslednje [Artenjak]:

- V razsevnih diagramih A, B in C se točkam najbolje prilega premica, v D, G, H in I parabola, ter v E 1ogistična krivulja, medtem ko v primeru F spremenljivki nista povezani.

- V primerih pod A, D, E, G in I sta spremenljivki močno povezani, ker so točke razmeščene blizu zamišljene krivulje. V primeru pod H je povezanost srednje močna in v primerih pod B in C je povezanost med spremenljivkama šibka.

- Smer povezanosti je pozitivna v primerih pod A, B, E in H ter negativna v primerih pod C in G.



Slika 199: Nekateri najbolj značilni primeri povezanosti med dvema spremenljivkama [Artenjak]

Pri analitičnem določanju enačbe regresijske krivulje:

$$\hat{y} = f\left(x_i, \hat{a}_0, \hat{a}_1, \hat{a}_2, ...\right) \tag{9.10}$$

je potrebno ne glede na tip funkcije določiti (oceniti) vrednosti parametrov $\hat{a}_j, j = 1,...,k$ tako, da se krivulja regresijske funkcije čim bolj prilega stvarnim podatkom. **Kot merilo boljše ali slabše prilagojenosti krivulje vzamemo vsoto kvadratov odklonov** stvarnih

vrednosti y od vrednosti $\hat{y}$ na krivulji regresijske funkcije. Če vrednosti parametrov izbrane regresijske funkcije določimo tako, da je vsota kvadratov odklonov stvarnih od teoretičnih vrednosti najmanjša, torej da velja [Artenjak]:

$$S = \sum_{i=1}^{N} \left(y(i) - \hat{y}(i) \right)^2 = \sum_{i=1}^{N} \left(e(i) \right)^2 \Rightarrow \min \tag{9.11}$$

kjer je $e(i)$ napaka modela, potem govorimo o **metodi najmanjših kvadratov** [Artenjak].

9.4 Linearna povezanost med številskima spremenljivkama

Dosedanja razglabljanja o korelaciji in regresiji veljajo na splošno ne glede na tip regresijske funkcije. Vendar je ugotavljanje jakosti ter analitično določanje enačbe regresijske funkcije najenostavnejše, če predpostavljamo, da je povezanost med dvema spremenljivkama linearna. Linearne odvisnosti so v praksi zelo pomembne in jih najpogosteje uporabljamo [Artenjak]. Obenem pa je dobro teoretično razumevanje korelacijske in linearne povezanosti med spremenljivkama pomembno tudi za pravilno uporabo in vrednotenje izidov nelinearne in multiple linearne regresije [Artenjak].

Korelacijska in regresijska analiza je sistematično proučevanje odvisnosti med opazovanima pojavoma in jo zelo zgoščeno sestavljajo tile zaporedni koraki [Artenjak]:

1. Opredelitev odvisne in neodvisne spremenljivke, merskih enot ter vrednosti spremenljivk.

2. Prikaz parov vrednosti spremenljivk v razsevnem diagramu ter opredelitev osnovnih značilnosti povezanosti med spremenljivkama (jakost, oblika, smer).

3. Izračun kazalcev korelacijske povezanosti (korelacijski in determinacijski koeficient).

4. Izračun koeficientov enačbe regresijske funkcije.

5. Ocenjevanje vrednosti odvisne spremenljivke na osnovi enačbe regresijske funkcije.

V nadaljevanju si poglejmo najpreprostejši primer linearne regresije, ko vzamemo za regresijsko funkcijo kar regresijsko premico (model 1. reda).

9.4.1 Regresijska premica

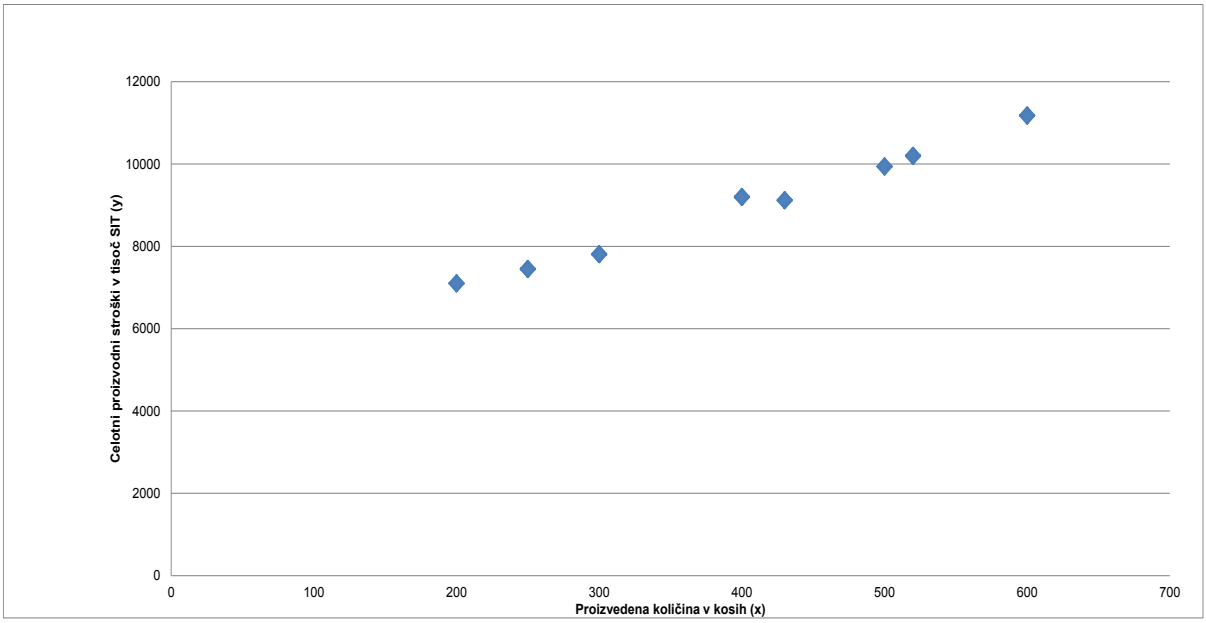
Denimo imamo primer analize odvisnosti celotnih proizvodnih stroškov *y* od proizvedene količine *x*, kar je prikazano na sliki 200 [Artenjak].

Preglednica 3.2 Proizvodnja in celotni proizvodni stroški izdelka v osmih mesecih

Mesec	I.	II.	III.	IV.	V.	VI.	VII.	VIII.
Proizvodnja v kosih	200	250	400	300	500	600	520	430
Celotni proizvodni stroški v tisoč SIT	7.100	7.450	9.200	7.810	9.940	11.18 0	10.200	9.120

Slika 200: Odvisnost celotnih proizvodnih stroškov y od proizvedene količine x [Artenjak]

Razsevni diagram ima obliko, prikazano na sliki 201 [Artenjak].



Slika 201: Razsevni diagram za odvisnost celotnih proizvodnih stroškov y od proizvedene količine x [Artenjak]

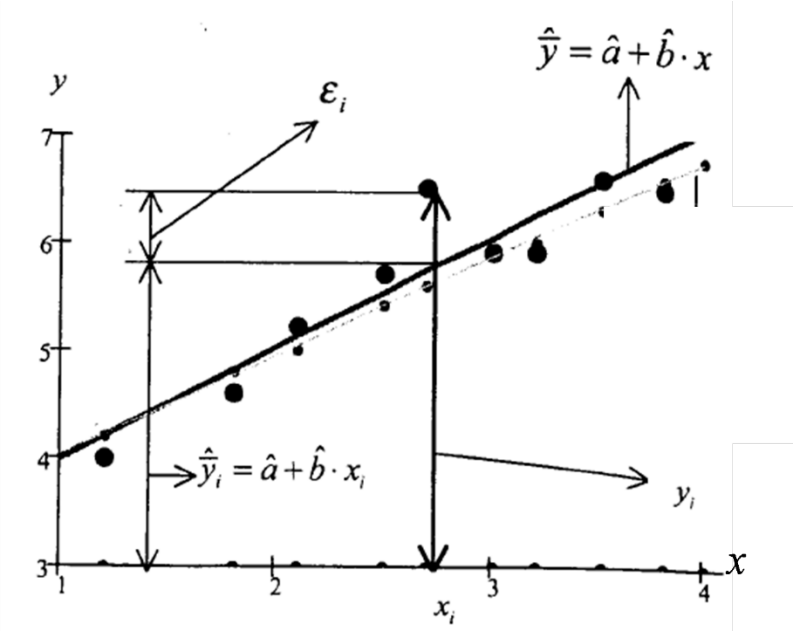
Tudi brez poznavanja teorije stroškov ni težko ugotoviti, da so celotni proizvodni stroški odvisna spremenljivka y in mesečno proizvedena količina izdelka neodvisna spremenljivka (x). Na razsevnem diagramu na sliki 201 je razvidno, da imamo opraviti z močno, pozitivno in lineamo korelacijsko povezanostjo med celotnimi proizvodnimi stroški in proizvedeno količino [Artenjak].

Na osnovi izraza (9.1) lahko zapišemo:

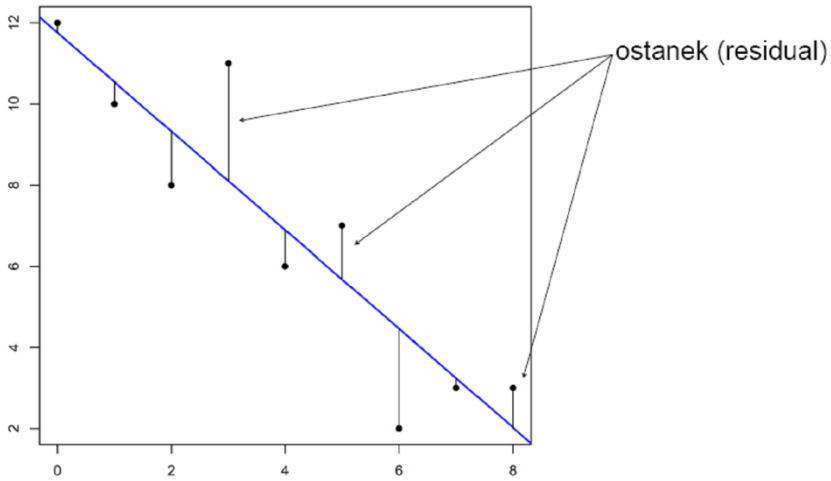
$$\begin{aligned} y(x) &= f(x) + \varepsilon \\ y(x) &= a + b \cdot x + \varepsilon \\ E(y) &= \bar{y} = E(a + b \cdot x + \varepsilon) = a + b \cdot E(x) = a + b \cdot x \end{aligned} \tag{9.12}$$

Sedaj se pojavi vprašanje, kako oceniti parametra a in b. Vzemimo, da imamo naključni vzorec $x_1, ..., x_n$, ki določa n stanj pojava X, ter pripadajoči naključni vzorec $y_1, ..., y_n$, ki določa n stanj pojava Y. Na ta vzorca lahko gledamo tudi kot na množico urejenih parov $\{(x_1, y_1), ..., (x_n, y_n)\}$, ki predstavljajo množico točk v ravnini. Kako sedaj poiskati premico, ki se tem točkam najbolje prilega. To nalogo največkrat rešujemo z že omenjeno metodo najmanjših kvadratov. Z njo poiščemo oceni parametrov $\hat{a}$ in $\hat{b}$, ki pripadata

točkasti oceni matematičnega upanja odvisne spremenljivke $\hat{\bar{y}} = \hat{a} + \hat{b} \cdot x$ (glej sliki 202 in 203) [Jesenko].



Slika 202: Metoda najmanjših kvadratov [Jesenko]



Slika 203: Metoda najmanjših kvadratov [Žibert]

Na osnovi slike 203 lahko zapišemo:

$$
\begin{aligned}
&y = \hat{\bar{y}} + \varepsilon = \hat{a} + \hat{b} \cdot x + \varepsilon \\
&oz. \\
&y(i) = \hat{\bar{y}}(i) + \varepsilon(i) = \hat{a} + \hat{b} \cdot x(i) + \varepsilon(i) \\
&in \\
&\varepsilon(i) = y(i) - \hat{\bar{y}}(i) = y(i) - \hat{a} - \hat{b} \cdot x(i) \\
&i = 1,...,n
\end{aligned}
\tag{9.13}
$$

Če opazujemo sliki 202 in 203, gotovo velja, da se bo premica tem bolj prilegala podatkom, čim manjši bodo residuali (napake oz. pogreški modela), točneje, čim manjša bo vsota kvadratov pogreškov modela. Tako lahko zapišemo pri danem naključnem vzorcu velikosti *n*:

$$
S = \sum_{i=1}^{n} \varepsilon_i^{\,2} = \sum_{i=1}^{n} \varepsilon(i)^2 \rightarrow \min \tag{9.14}
$$

oz.:

$$
S\left(\hat{a},\hat{b}\right) = \sum_{i=1}^{n} \varepsilon(i)^2 = \sum_{i=1}^{n} \left(y(i) - \hat{\bar{y}}(i)\right)^2 = \sum_{i=1}^{n} \left[y(i) - \hat{a} - \hat{b} \cdot x(i)\right]^2 \rightarrow \min \tag{9.15}
$$

Kot vidimo, je kriterijska funkcija $S\left(\hat{a},\hat{b}\right)$ odvisna od dveh ocenjenih parametrov ter izmerjenih podatkov odvisne in neodvisne spremenljivke. Sedaj moramo izvesti naslednjo matematično operacijo:

$$
\min_{\hat{a},\hat{b}} S\left(\hat{a},\hat{b}\right) \rightarrow \hat{a}^*, \hat{b}^* \; optima \ln a \;\; parametra \rightarrow S\left(\hat{a}^*, \hat{b}^*\right) = \min_{\hat{a},\hat{b}} S\left(\hat{a},\hat{b}\right) \tag{9.16}
$$

torej moramo očitno poiskati ekstrem funkcije $S\left(\hat{a},\hat{b}\right)$. Ker gre za funkcijo dveh spremenljivk (parametrov), moramo poiskati parcialna odvoda funkcije po obeh parametrih ter dobljena rezultata enačiti z 0. Tako bomo dobili sistem dveh enačb z dvema neznankama, ki ga bomo morali rešiti. Končni rezultat bosta optimalna ocenjena parametra.

Izvedemo torej:

$$\begin{aligned}\frac{\partial S\left(\hat{a},\hat{b}\right)}{\partial \hat{a}} &= 0 \\ \frac{\partial S\left(\hat{a},\hat{b}\right)}{\partial \hat{b}} &= 0\end{aligned} \tag{9.17}$$

Dobimo:

$$\frac{\partial S\left(\hat{a},\hat{b}\right)}{\partial \hat{a}} = \frac{\partial S}{\partial \hat{a}}\left(\sum_{i=1}^{n}\left[y(i)-\hat{a}-\hat{b}\cdot x(i)\right]^2\right) = 2\cdot\sum_{i=1}^{n}\left[y(i)-\hat{a}-\hat{b}\cdot x(i)\right](-1) = 0$$

sledi :

$$\begin{aligned}&\sum_{i=1}^{n}\left[y(i)-\hat{a}-\hat{b}\cdot x(i)\right] = 0 \\ &\sum_{i=1}^{n}\left[y(i)\right]-\hat{a}\cdot n-\hat{b}\cdot\sum_{i=1}^{n}\left[x(i)\right] = 0 \\ &\sum_{i=1}^{n}\left[y(i)\right] = \hat{a}\cdot n+\hat{b}\cdot\sum_{i=1}^{n}\left[x(i)\right]\end{aligned} \tag{9.18}$$

in:

$$\frac{\partial S\left(\hat{a},\hat{b}\right)}{\partial \hat{b}} = \frac{\partial S}{\partial \hat{b}}\left(\sum_{i=1}^{n}\left[y(i)-\hat{a}-\hat{b}\cdot x(i)\right]^2\right) = 2\cdot\sum_{i=1}^{n}\left[y(i)-\hat{a}-\hat{b}\cdot x(i)\right]\left(-x(i)\right) = 0$$

sledi :

$$\begin{aligned}&\sum_{i=1}^{n}\left[x(i)\cdot y(i)-\hat{a}\cdot x(i)-\hat{b}\cdot x^2(i)\right] = 0 \\ &\sum_{i=1}^{n}\left[x(i)\cdot y(i)\right]-\hat{a}\cdot\sum_{i=1}^{n}\left[x(i)\right]-\hat{b}\cdot\sum_{i=1}^{n}\left[x^2(i)\right] = 0 \\ &\sum_{i=1}^{n}\left[x(i)\cdot y(i)\right] = \hat{a}\cdot\sum_{i=1}^{n}\left[x(i)\right]+\hat{b}\cdot\sum_{i=1}^{n}\left[x^2(i)\right]\end{aligned} \tag{9.19}$$

Ko rešimo sistem enačb, po daljši izpeljavi dobimo [Jesenko, Artenjak]:

$$\hat{a} = \frac{\sum_{i=1}^{n}\left[x^2(i)\right]\cdot\sum_{i=1}^{n}\left[y(i)\right]-\sum_{i=1}^{n}\left[x(i)\right]\cdot\sum_{i=1}^{n}\left[x(i)\cdot y(i)\right]}{n\cdot\sum_{i=1}^{n}\left[x^2(i)\right]-\left(\sum_{i=1}^{n}\left[x(i)\right]\right)^2}$$
$$\hat{b} = \frac{n\cdot\sum_{i=1}^{n}\left[x(i)\cdot y(i)\right]-\sum_{i=1}^{n}\left[x(i)\right]\cdot\sum_{i=1}^{n}\left[y(i)\right]}{n\cdot\sum_{i=1}^{n}\left[x^2(i)\right]-\left(\sum_{i=1}^{n}\left[x(i)\right]\right)^2} \qquad (9.20)$$

Če števec in imenovalec v izrazu za $\hat{b}$ delimo z n^2, dobimo:

$$\hat{b} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left[x(i)\cdot y(i)\right]-\frac{1}{n^2}\sum_{i=1}^{n}\left[x(i)\right]\cdot\sum_{i=1}^{n}\left[y(i)\right]}{\frac{1}{n}\cdot\sum_{i=1}^{n}\left[x^2(i)\right]-\frac{1}{n^2}\left(\sum_{i=1}^{n}\left[x(i)\right]\right)^2} =$$
$$= \frac{\frac{1}{n}\sum_{i=1}^{n}\left[x(i)\cdot y(i)\right]-\left(\frac{1}{n}\sum_{i=1}^{n}\left[x(i)\right]\right)\cdot\left(\frac{1}{n}\sum_{i=1}^{n}\left[y(i)\right]\right)}{\frac{1}{n}\cdot\sum_{i=1}^{n}\left[x^2(i)\right]-\left(\frac{1}{n}\sum_{i=1}^{n}\left[x(i)\right]\right)^2} \qquad (9.21)$$

Vpeljimo novi oznaki:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n}\left[x(i)\right]$$
$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n}\left[y(i)\right] \qquad (9.22)$$

Sledi:

(9.23)

$$\hat{b} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left[x(i)\cdot y(i)\right] - \bar{x}\cdot\bar{y}}{\frac{1}{n}\cdot\sum_{i=1}^{n}\left[x^2(i)\right] - \bar{x}^2}$$

Izraz v števcu je kovarianca med spremenljivkama x in y, izraz v imenovalcu pa varianca spremenljivke x [Artenjak] (glej tudi poglavje 2.22). Velja namreč (če nepristransko kovarianco aproksimiramo s pristransko):

$$\begin{aligned} C_{xy} &= \frac{1}{n-1}\sum_{i=1}^{n}\left[\left(x(i)-\bar{x}\right)\cdot\left(y(i)-\bar{y}\right)\right] \approx \\ &\approx \frac{1}{n}\sum_{i=1}^{n}\left[\left(x(i)\cdot y(i) - \bar{x}\cdot y(i) - x(i)\cdot\bar{y} + \bar{x}\cdot\bar{y}\right)\right] = \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[\left(x(i)\cdot y(i)\right)\right] - 2\cdot\bar{x}\cdot\bar{y} + \bar{x}\cdot\bar{y} = \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[\left(x(i)\cdot y(i)\right)\right] - \bar{x}\cdot\bar{y} \end{aligned} \tag{9.24}$$

in (če nepristransko varianco aproksimiramo s pristransko):

$$\begin{aligned} C_{xx} = \sigma_x^2 &= \frac{1}{n-1}\sum_{i=1}^{n}\left[\left(x(i)-\bar{x}\right)\cdot\left(x(i)-\bar{x}\right)\right] \approx \\ &\approx \frac{1}{n}\sum_{i=1}^{n}\left[\left(x(i)^2 - \bar{x}\cdot x(i) - x(i)\cdot\bar{x} + \bar{x}^2\right)\right] = \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[x(i)^2\right] - 2\cdot\bar{x}^2 + \bar{x}^2 = \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[x(i)^2\right] - \bar{x}^2 \end{aligned} \tag{9.25}$$

Dobimo torej:

$$\hat{b} = \frac{C_{xy}}{C_{xx}} = \frac{C_{xy}}{\sigma_x^2} \tag{9.26}$$

Če pa števec in imenovalec izraza (9.21) množimo z *n* (kar je velikokrat praksa), dobimo:

$$\hat{b} = \frac{\sum_{i=1}^{n}\left[x(i)\cdot y(i)\right] - \frac{1}{n}\sum_{i=1}^{n}\left[x(i)\right]\cdot\sum_{i=1}^{n}\left[y(i)\right]}{\sum_{i=1}^{n}\left[x^2(i)\right] - \frac{1}{n}\left(\sum_{i=1}^{n}\left[x(i)\right]\right)^2} = \tag{9.27}$$
$$= \frac{S_{xy}}{S_{xx}}$$

Predelajmo sedaj izraz (9.18) :

$$\begin{aligned}
&\sum_{i=1}^{n}\left[y(i)\right] = \hat{a}\cdot n + \hat{b}\cdot\sum_{i=1}^{n}\left[x(i)\right] \\
&\frac{1}{n}\sum_{i=1}^{n}\left[y(i)\right] = \hat{a} + \hat{b}\cdot\frac{1}{n}\cdot\sum_{i=1}^{n}\left[x(i)\right] \\
&\hat{a} = \left(\frac{1}{n}\sum_{i=1}^{n}\left[y(i)\right]\right) - \hat{b}\cdot\left(\frac{1}{n}\cdot\sum_{i=1}^{n}\left[x(i)\right]\right) \\
&\hat{a} = \bar{y} - \hat{b}\cdot\bar{x}
\end{aligned} \tag{9.28}$$

Na osnovi izrazov (9.27) in (9.28) lahko izračunamo oceni obeh parametrov. Ko sta parametra $\hat{a},\hat{b}$ enkrat ocenjena, ju lahko uporabimo pri izbrani vrednosti $X = x(i) = x_i$ v točkasti oceni za $E\left(Y\middle|x_i\right)$ [Jesenko]:

$$\hat{\bar{y}}(i) = \hat{a} + \hat{b}\cdot x(i), \quad i = 1,...,n \tag{9.29}$$

Točke $\left(x(i), \hat{\bar{y}}(i)\right)$ za vse $i = 1,...,n$ ležijo na ocenjeni regresijski premici.

Primer 9.2.:

Tabela na sliki 204 podaja vzorec osmih ocenjenih vrednosti stanovanj, kot so jih ocenili cenilci, ter dejansko doseženo prodajno ceno. Vrednosti so podane v milijonih SIT [Jesenko]. Izračunajte in narišite ocenjeno regresijsko premico s katero bo mogoče napovedati povprečno prodajno ceno stanovanja pri znani uradno ocenjeni vrednosti.

Ocenjena vrednost (x_i)	Prodajna cena (y_i)
125	132
83	88
182	177
135	138
147	146
112	121
211	203
76	87

Slika 204: Vzorec osmih ocenjenih vrednosti stanovanj, kot so jih ocenili cenilci, ter dejanska dosežena prodajna cena [Jesenko]

Določimo najprej parameter $\hat{b}$. V ta namen izračunamo vrednosti, ki jih prikazuje tabela na sliki 205.

$x_i \cdot y_i$	x_i^2	y_i^2
16500	15625	17424
7304	6889	7744
32214	33124	31329
18630	18225	19044
21462	21609	21316
13552	12544	14641
42833	44521	41209
6612	5776	7569

Slika 205: Nekateri vmesni izračuni [Jesenko]

Nato izračunajmo vse vsote:

$$\begin{aligned}
&\sum_{i=1}^{8} x(i) = 125 + 83 + \ldots + 76 = 1071 \\
&\sum_{i=1}^{8} y(i) = 132 + 88 + \ldots + 87 = 1092 \\
&\sum_{i=1}^{8} x(i) \cdot y(i) = 125 \cdot 132 + 83 \cdot 88 + \ldots + 76 \cdot 87 = \\
&16500 + 7304 + \ldots + 6612 = 159107 \\
&\sum_{i=1}^{8} x(i)^2 = 125^2 + 83^2 + \ldots + 76^2 = \\
&= 15625 + 6889 + \ldots + 5776 = 158313 \\
&\sum_{i=1}^{8} y(i)^2 = 132^2 + 88^2 + \ldots + 87^2 = \\
&= 17424 + 7744 + \ldots + 7569 = 160276
\end{aligned} \tag{9.30}$$

Izračunajmo sedaj:

$$\begin{aligned}
S_{xy} &= \sum_{i=1}^{8} x(i) \cdot y(i) - \frac{1}{8}\left(\sum_{i=1}^{8} x(i)\right)\left(\sum_{i=1}^{8} y(i)\right) = 159107 - \frac{1}{8} \cdot 1071 \cdot 1092 = 12916 \\
S_{xx} &= \sum_{i=1}^{8} x(i)^2 - \frac{1}{8}\left(\sum_{i=1}^{8} x(i)\right)^2 = 158313 - \frac{1}{8} \cdot 1071^2 = 14933
\end{aligned} \tag{9.31}$$

Sledi:

$$\hat{b} = \frac{S_{xy}}{S_{xx}} = \frac{12916}{14933} = 0.8649 \tag{9.32}$$

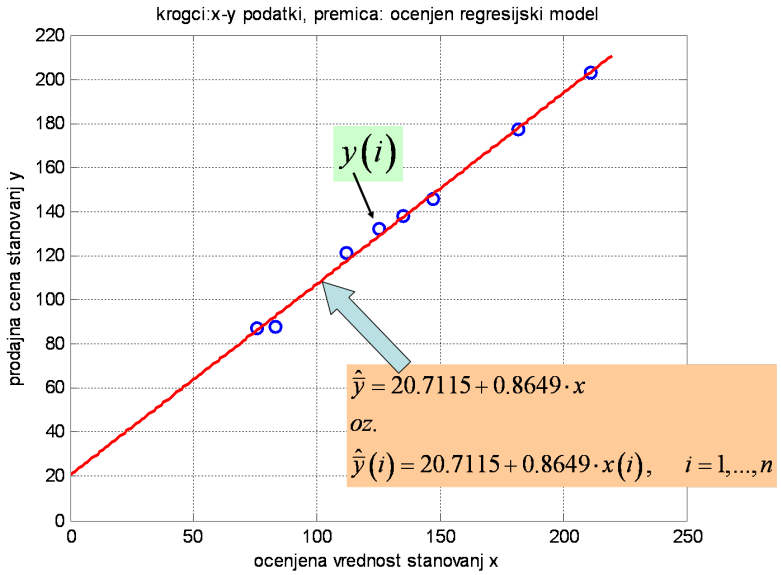
ter:

$$\begin{aligned}
&\hat{a} = \bar{y} - \hat{b} \cdot \bar{x} = \frac{1}{n}\sum_{i=1}^{n} y(i) - \hat{b} \cdot \frac{1}{n}\sum_{i=1}^{n} x(i) = \\
&= \frac{1}{8}\sum_{i=1}^{8} y(i) - 0.8649 \cdot \frac{1}{8}\sum_{i=1}^{8} x(i) = \\
&= \frac{1}{8} \cdot 1092 - 0.8649 \cdot \frac{1}{8} \cdot 1071 = 20.7115
\end{aligned} \tag{9.33}$$

Ocenjena regresijska premica ima torej obliko:

$$\hat{y} = 20.7115 + 0.8649 \cdot x$$
$$oz.$$
$$\hat{\bar{y}}(i) = 20.7115 + 0.8649 \cdot x(i), \quad i = 1,...,n \tag{9.34}$$

Podatke iz tabele na sliki 204 in ocenjeno regresijsko premico prikazuje slika 206.



Slika 206: Podatki (krogci) in ocenjena regresijska premica

Za izris slike 206 in vse izračune smo si pomagali z naslednjim programom v Matlabu:

```
% mnk.m

clear
clc
close all

x = [125 83 182 135 147 112 211 76]
y = [132 88 177 138 146 121 203 87]

n = length(x)

disp('vsota x je:')
sx = sum(x)

disp('vsota y je:')
sy = sum(y)
```

```
disp('vsota x^2 je:')
sx2 = x*x'

disp('vsota y^2 je:')
sy2 = y*y'

disp('vsota x.y je:')
sxy = x*y'

disp('Sxy je:')
Sxy = sxy - sx*sy/n

disp('Sxx je:')
Sxx = sx2 - sx^2/n

disp('boc=')
boc = Sxy/Sxx

disp('aoc=')
aoc = sy/n - boc*sx/n

plot(x,y,'o','LineWidth',2)
grid
xlabel('ocenjena vrednost stanovanj x')
ylabel('prodajna cena stanovanj y')
d = axis
axis([0 250 0 d(4)])

hold on
x = 0:1:220;
yoc = aoc + boc*x;
plot(x,yoc,'r','LineWidth',1.5)
title('krogci:x-y podatki, premica: ocenjen regresijski model')
```

Izpis komandnega okna je naslednji:

```
x =
  125   83  182  135  147  112  211   76

y =
  132   88  177  138  146  121  203   87

n =
   8

vsota x je:
sx =
    1071

vsota y je:
```

```
sy =
      1092

vsota x^2 je:
sx2 =
    158313

vsota y^2 je:
sy2 =
    160276

vsota x.y je:
sxy =
    159107

Sxy je:
Sxy =
  1.2916e+004

Sxx je:
Sxx =
  1.4933e+004

boc=
boc =
   0.8649

aoc=
aoc =
  20.7110
```

Primer 9.3.:

Dane imamo podatke za tanin x in rast y, kot jih prikazuje slika 207 [Žibert]. Izračunajte in narišite ocenjeno regresijsko premico.

tanin	rast
0	12
1	10
2	8
3	11
4	6
5	7
6	2
7	3
8	3

Slika 207: Tanin (x) in rast (y) [Žibert]

Dobimo:

$$\begin{aligned} S_{xy} &= \sum_{i=1}^{9} x(i)\cdot y(i) - \frac{1}{9}\left(\sum_{i=1}^{9} x(i)\right)\left(\sum_{i=1}^{9} y(i)\right) = \ldots = -73 \\ S_{xx} &= \sum_{i=1}^{9} x(i)^2 - \frac{1}{9}\left(\sum_{i=1}^{9} x(i)\right)^2 = \ldots = 60 \end{aligned} \tag{9.35}$$

Sledi:

$$\hat{b} = \frac{S_{xy}}{S_{xx}} = \frac{-73}{60} = -1.2167 \tag{9.36}$$

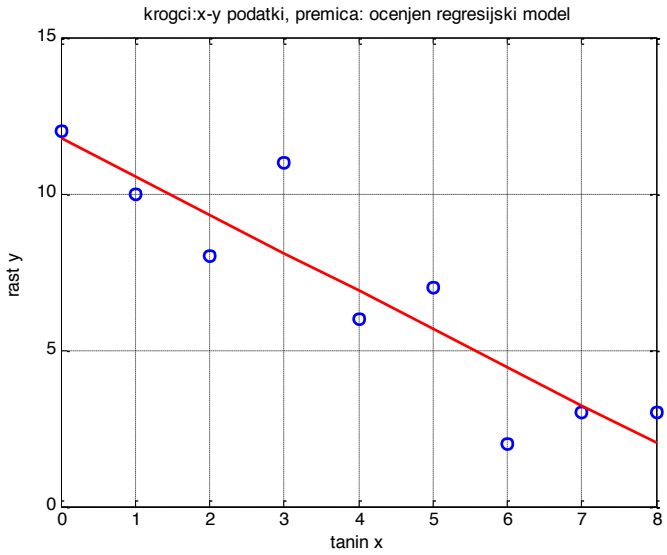
ter:

$$\begin{aligned} \hat{a} &= \bar{y} - \hat{b}\cdot\bar{x} = \frac{1}{n}\sum_{i=1}^{n} y(i) - \hat{b}\cdot\frac{1}{n}\sum_{i=1}^{n} x(i) = \\ &= \frac{1}{9}\sum_{i=1}^{9} y(i) + 1.2167\cdot\frac{1}{9}\sum_{i=1}^{9} x(i) = \\ &= \ldots = 11.7557 \end{aligned} \tag{9.37}$$

Ocenjena regresijska premica ima torej obliko:

$$\begin{aligned} &\hat{\bar{y}} = 11.7556 - 1.2167\cdot x \\ &oz. \\ &\hat{\bar{y}}(i) = 11.7556 - 1.2167\cdot x(i), \quad i = 1,\ldots,n \end{aligned} \tag{9.38}$$

Podatke iz tabele na sliki 207 in ocenjeno regresijsko premico prikazuje slika 208.



Slika 208: Podatki (krogci) iz tabele na sliki 207 in ocenjena regresijska premica

Za izris slike 208 in vse izračune smo si pomagali z naslednjim programom v Matlabu:

```
% mnk2.m

clear
clc
close all

tanin = importdata('tannin.txt');
data = tanin.data;

% podatki:
x = data(:,2); %tanin
y = data(:,1); %rast

disp('Sxx je:')
Sxx=sum(x.^2)-sum(x)^2/length(x)

disp('Sxy je:')
Sxy=sum(x.*y)-sum(x)*sum(y)/length(x)

% ocenjena parametra:

boc = Sxy/Sxx

aoc = mean(y) - boc*mean(x)

plot(x,y,'o','LineWidth',2)
grid
xlabel('tanin x')
ylabel('rast y')
d = axis
axis([0 d(2) 0 15])

hold on
x = 0:1:8;
yoc = aoc + boc*x;
plot(x,yoc,'r','LineWidth',1.5)
title('krogci:x-y podatki, premica: ocenjen regresijski model')
```

Izpis komandnega okna pa je naslednji:

```
Sxx je:
Sxx =
```

```
    60

Sxy je:
Sxy =
   -73

boc =
   -1.2167

aoc =
   11.7556
```

Pri izračunih in izrisu slike 208 bi si lahko pomagali tudi s standardnim ukazom v Matlabu: **regress(…)** Uporabili bi ga na naslednji način:

```
% mnk3.m

clear
clc
close all

tanin = importdata('tannin.txt');
data = tanin.data;

% podatki:
x = data(:,2); %tanin
y = data(:,1); %rast

X = [ones(length(x),1), x];
par = regress(y,X);

aoc = par(1)
boc = par(2)

hold on;
plot(x,y, 'o')
plot(x, aoc + boc*x, 'r-')
title('Linearna regresija')
xlabel('tanin'); ylabel('rast')
hold off;
```

Praktični in teoretični pomen linearne regresijske premice

Linearna regresijska funkcija z ocenjenimi koeficienti je pomembna zaradi teh teoretičnih značilnosti [Artenjak]:

1 . Seštevek odstopanj stvarnih vrednosti odvisne spremenljivke y od regresijskih vrednosti $\hat{\bar{y}}$ je enak 0.

2. Seštevek kvadratov teh odstopanj je minimalen. To izvira iz metode ocenjevanja.

3. Seštevek zmnožkov vrednosti odvisne spremenljivke $\hat{\bar{y}}$ in odklonov ε je enak 0.

4. Seštevek zmnožkov vrednosti neodvisne spremenljivke x in odklonov ε je prav tako enak 0.

5. Iz prve značilnosti sledi, da je aritmetična sredina stvarnih vrednosti enaka aritmetični sredini regresijskih (teoretičnih) vrednosti.

Stopnjo prilagojenosti izbrane regresijske funkcije stvarnim podatkom ugotavljamo tudi na osnovi naslednjega kazalca relativnih odklonov [Artenjak]:

$$\varepsilon(i)(\%) = \frac{y(i) - \hat{\bar{y}}(i)}{y(i)} \cdot 100, \quad i = 1,...,n \tag{9.39}$$

ki kaže relativne odklone spremenljivke ε.

V praktične namene pa je linearna regresijska funkcija pomembna iz teh razlogov [Artenjak]:

1. Enačba regresijske premice prikazuje povezanost med spremenljivkama tako, da lahko za poljubno vrednost neodvisne spremenljivke iz njenega definicijskega območja ugotovimo vrednost odvisne spremenljivke, pri tem pa seveda ne upoštevamo naključnih in drugih vplivov.

2. Koeficient *b* ima dvojen pomen:

a) predznak pokaže smer povezanosti med spremenljivkama, zato se tudi imenuje smerni koeficient;

b) vrednost pa pove, za koliko enot se v povprečju spremeni vrednost odvisne spremenljivke, če se neodvisna spremenljivka spremeni za enoto.

3. Koeficient *a* matematično pomeni odsek na oordinatni osi, nima pa vselej tudi ekonomske ali kakšne druge fizikalne razlage.

Primer 9.4.:

Ponovno vzemimo primer proizvodnje in celotnih proizvodnih stroškov izdelka v 8 mesecih (glej sliki 200 in 201). Izračunajte in narišite ocenjeno regresijsko premico [Artenjak].

Dobimo:

$$\begin{aligned}
&\sum_{i=1}^{8} x(i) = 3200, \quad \bar{x} = \frac{1}{8} \cdot 3200 = 400 \\
&\sum_{i=1}^{8} y(i) = 72000, \quad \bar{y} = \frac{1}{8} \cdot 72000 = 9000 \\
&\sum_{i=1}^{8} x(i) \cdot y(i) = 30.209.100 \\
&\sum_{i=1}^{8} x(i)^2 = 1.417.800 \\
&\sum_{i=1}^{8} y(i)^2 = 662.559.000
\end{aligned} \tag{9.40}$$

Dobimo tudi:

$$\begin{aligned}
S_{xy} &= \sum_{i=1}^{8} x(i) \cdot y(i) - \frac{1}{8}\left(\sum_{i=1}^{8} x(i)\right)\left(\sum_{i=1}^{8} y(i)\right) = 1.409.100 \\
S_{xx} &= \sum_{i=1}^{8} x(i)^2 - \frac{1}{8}\left(\sum_{i=1}^{8} x(i)\right)^2 = 137800
\end{aligned} \tag{9.41}$$

Sledi:

$$\hat{b} = \frac{S_{xy}}{S_{xx}} = \frac{1.409.100}{137800} = 10.2257 \tag{9.42}$$

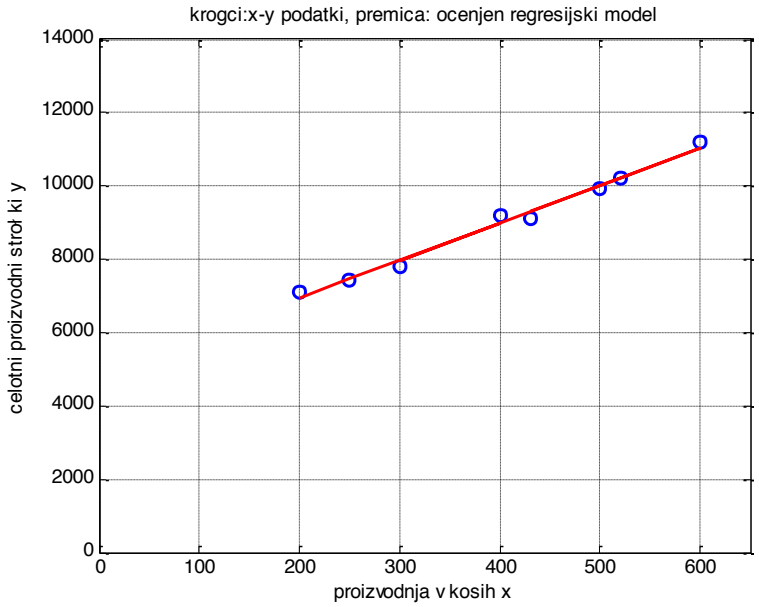
ter:

$$\begin{aligned} \hat{a} &= \bar{y} - \hat{b} \cdot \bar{x} = \frac{1}{n}\sum_{i=1}^{n} y(i) - \hat{b} \cdot \frac{1}{n}\sum_{i=1}^{n} x(i) = \\ &= \frac{1}{8}\sum_{i=1}^{8} y(i) - 10.2257 \cdot \frac{1}{8}\sum_{i=1}^{8} x(i) = \\ &= \frac{1}{8} \cdot 72000 - 10.2257 \cdot \frac{1}{8} \cdot 3200 = \\ &= 9000 - 10.2257 \cdot 400 = 4909.7 \end{aligned} \tag{9.43}$$

Ocenjena regresijska premica ima torej obliko:

$$\begin{aligned} &\hat{\bar{y}} = 4909.7 + 10.2257 \cdot x \\ &oz. \\ &\hat{\bar{y}}(i) = 4909.7 + 10.2257 \cdot x(i), \quad i = 1,...,n \end{aligned} \tag{9.44}$$

Podatke in ocenjeno regresijsko premico prikazuje slika 209.



Slika 209: Podatki in ocenjena regresijska premica

Ocenjena regresijska premica ima pri minimalnem in maksimalnem številu proizvedenih kosov naslednji vrednosti:

$$\begin{aligned}
&\hat{\bar{y}}_{\min} = \hat{\bar{y}}(1) = 4909.7 + 10.2257 \cdot x(1) = \\
&= 4909.7 + 10.2257 \cdot 200 = 6954.8 \\
&\hat{\bar{y}}_{\max} = \hat{\bar{y}}(6) = 4909.7 + 10.2257 \cdot x(6) = \\
&= 4909.7 + 10.2257 \cdot 600 = 11045
\end{aligned} \qquad (9.45)$$

Kot je razvidno iz tabele na sliki 200, so namreč najmanj proizvedli v 1. mesecu (200 kosov pri stroških 7100 tisoč SIT), največ pa v 6. mesecu (600 kosov pri stroških 11180 tisoč SIT).

Na podlagi regresijske enačbe (3.18) v tabeli na sliki 210 prikazujemo vrednosti odvisne spremenljivke, regresijske vrednosti in odklone. Primer potrjuje prej navedena teoretična spoznanja [Artenjak].

Mesec i	**Celotni stvarni proizvodni stroški** y_i	**Celotni teoretični proizvodni stroški** $\hat{y}_i$	**Odstopanja stroškov** $y_i - \hat{y}_i$	**Relativna odstopanja stroškov** $e\%_i$
1	7.100	6.955	+145	+2,04
2	7.450	7.466,25	-16,25	-0,21
3	9.200	9.000	+200	+2,17
4	7.810	7.977,5	-167,5	-2,14
5	9.940	10.022,5	-82,5	-0,83
6	11.180	11.045	+135	+1,21
7	10.200	10.227,0	-27,0	-0,27
8	9.120	9.306,75	-186,75	+2,05
Skupaj	**72.000**	**72.000**	**0**	

Slika 210: Mesečni stvarni in z regresijsko premico izračunani celotni proizvodni stroški ter mesečna absolutna in relativna odstopanja celotnih proizvodnih stroškov [Artenjak]. ($\hat{y}$ pomeni oceno matematičnega upanja odvisne spremenljivke za dani x).

Regresijska funkcija (9.44) ima naslednji pomen [Artenjak]:

a) Za vsako proizvodnjo od 0 do maksimalne proizvodnje, **denimo 1.000 kosov izdelka**, lahko ocenimo celotne proizvodne stroške. Odstopanje dejanskih stroškov od ocenjenih bi bilo na osnovi regresijske enačbe (9.44) zanemarljivo, ker smo v tabeli na sliki 210 ugotovili, da so mesečna odstopanja celotnih stvarnih proizvodnih stroškov od teoretičnih (regresijskih) stroškov v razmiku od -2,14 % do +2,17 %.

b) Za vsak dodatno proizvedeni kos izdelka se celotni proizvodni stroški povečajo v povprečju za 10.225 SIT (zaradi parametra $\hat{b} = 10.225\ SIT$).

c) Tudi če ne proizvajamo, imamo stroške. To so stalni stroški, ki v našem primeru v povprečju znašajo cca. 4.910 SIT (zaradi parametra $\hat{a} = 4909.7\ SIT$).

Pri uporabi enačbe regresijske krivulje za ocenjevanje (predvidevanje) lahko glede na vrednost neodvisne spremenljivke ločimo dva primera ocenjevanja:

1. Ocenjevanje na podlagi **interpolacije**, ko za vrednost neodvisne spremenljivke *x* izberemo vrednosti iz razmika med $x_{\min}$ in $x_{\max}$. V našem primeru bi bile to vrednosti od 200 do 600 kosov

2. Ocenjevanje na podlagi **ekstrapolacije**, ko za vrednosti neodvisne spremenljivke *x* uporabimo vrednosti, ki so izven razmika med $x_{\min}$ in $x_{\max}$. V našem primeru so to vrednosti spremenljivke od 0 do 199 kosov in od 601 do 1.000 kosov izdelka. Ekstrapolacijo sicer pogosto uporabljamo za predvidevanje (napovedovanje) v primeru časovnih vrst.

Za zgled uporabe enačbe regresijske premice ocenimo celotne proizvodne stroške pri proizvodnji 400 kosov izdelka.

Ker je:

$$\begin{aligned} & i = 3, x(3) = 400\ kosov \\ & y(3) = 9200\ SIT \\ & \hat{\tilde{y}}(3) = 4909.7 + 10.2257 \cdot x(3) = \\ & = 4909.7 + 10.2257 \cdot 400 = 9000 \\ & \varepsilon(3)(\%) = \frac{y(3) - \hat{\tilde{y}}(3)}{y(3)} \cdot 100\% = \frac{200}{9200} \cdot 100\% = 2.17\% \end{aligned} \qquad (9.46)$$

to pomeni, da bi bili na osnovi proizvodnje 400 kosov izdelka celotni ocenjeni proizvodni stroški enaki 9 mio SIT. Zaradi drugih in naključnih vplivov je ta ocena celotnih proizvodnih stroškov praviloma različna od dejanskih stroškov, o čemer se lahko prepričamo o dejanskih stroških za takšno količinsko proizvodnjo, ki jih navajamo v tabeli na sliki 210 za mesec marec (*i = 3*). Tedaj so stvarni stroški enaki 9200 SIT, torej so glede na ocenjene podcenjeni za 200 SIT (2,17 %).

Za vse izračune smo uporabili naslednji program v Matlabu:

```
% mnk4.m

clear
clc
close all


% podatki:
x = [200 250 400 300 500 600 520 430]
y = [7100 7450 9200 7810 9940 11180 10200 9120]

n = length(x)

disp('vsota x je:')
sx = sum(x)

disp('vsota y je:')
sy = sum(y)

disp('vsota x^2 je:')
sx2 = x*x'

disp('vsota y^2 je:')
sy2 = y*y'

disp('vsota x.y je:')
sxy = x*y'

disp('Sxy je:')
Sxy = sxy - sx*sy/n

disp('Sxx je:')
Sxx = sx2 - sx^2/n

disp('boc=')
boc = Sxy/Sxx

disp('aoc=')
aoc = sy/n - boc*sx/n

plot(x,y,'o','LineWidth',2)
grid
xlabel('proizvodnja v kosih x')
ylabel('celotni proizvodni stroški y')
d = axis
axis([0 650 0 14000])

hold on
yoc = aoc + boc*x;
plot(x,yoc,'r','LineWidth',1.5)
title('krogci:x-y podatki, premica: ocenjen regresijski model')

disp('Regresijska vrednost pri min stevilu kosov')
yoc_min = aoc + boc*min(x)

disp('Regresijska vrednost stroskov pri max stevilu kosov')
yoc_max = aoc + boc*max(x)

% Izracunajmo se relativne pogreske:

for i = 1:n
    e(i) = (y(i)-yoc(i))*100/y(i);
end

disp('y/1000      y_ocenj/1000  (y-y_ocenj)     e_relat(%) ')
[y'/1000 yoc'/1000 (y-yoc)' e']
```

Komandno okno ima naslednji izgled:

```
x =
   200   250   400   300   500   600   520   430

y =
  Columns 1 through 6
      7100       7450       9200       7810       9940      11180
  Columns 7 through 8
     10200       9120
n =
     8

vsota x je:
sx =
       3200
vsota y je:
sy =
      72000
vsota x^2 je:
sx2 =
     1417800
vsota y^2 je:
sy2 =
   662559000
vsota x.y je:
sxy =
    30209100

Sxy je:
Sxy =
     1409100

Sxx je:
Sxx =
      137800

boc=
boc =
   10.2257
aoc=
aoc =
  4.9097e+003
```

```
Regresijska vrednost pri min stevilu kosov
yoc_min =
  6.9549e+003

Regresijska vrednost stroskov pri max stevilu kosov
yoc_max =
  1.1045e+004

y/1000     y_ocenj/1000  (y-y_ocenj)    e_relat(%)
ans =
    7.1000    6.9549    145.1379     2.0442
    7.4500    7.4661    -16.1466    -0.2167
    9.2000    9.0000    200.0000    2.1739
    7.8100    7.9774    -167.4311   -2.1438
    9.9400   10.0226   -82.5689    -0.8307
   11.1800   11.0451  134.8621    1.2063
   10.2000   10.2271  -27.0827    -0.2655
    9.1200    9.3068    -186.7707  -2.0479
```

9.4.2 Regresijski model kot polinomska funkcija časa

V praksi se velikokrat zgodi, da je neodvisna spremenljivka čas, recimo pri napovedovanju časovnih vrst povpraševanja v prihodnosti. Zato bomo v tem poglavju pokazali, kako razviti regresijski model v tovrstnih primerih [Dragan 1].

Denimo imamo opravka z določenimi meritvami povpraševanja $d(t),\ \ t=1,2,...,N$, za katere želimo načrtati regresijski model. Model bo imel nalogo, da aproksimira obnašanje meritev povpraševanja, ter poskuša projecirati trend njihovega gibanja v prihodnostni dogodkovni horizont. Pri tem se bomo omejili le na uporabo linearnih modelov (linearnih v parametrih).

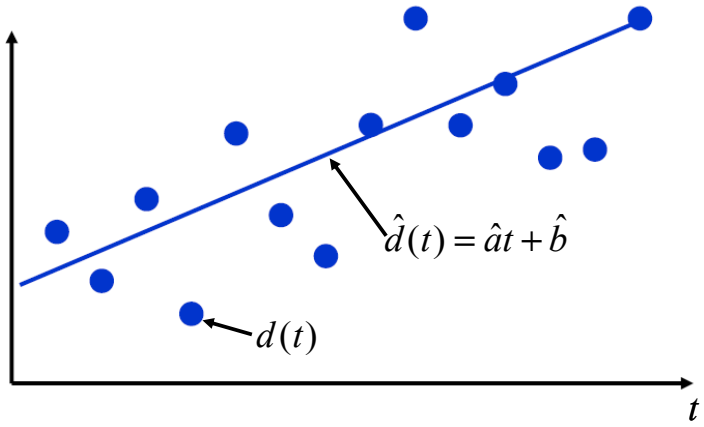
Če se odločimo za aproksimacijo oz. posnemanje obnašanja meritev povpraševanja z modelom kot polinomsko funkcijo časa, seveda lahko izberemo različne stopnje oz. rede polinoma. Običajno izberemo takšen polinom, ki bo še dovolj dobro ujel vzorec obnašanja meritev, ne da bi imel pri tem po nepotrebnem prevelik red.

Če želimo posnemati obnašanje meritev s premico (polinomom 1. reda), bo imel linearni model (1. reda) naslednjo strukturo:

$$\hat{d}(t) = \hat{a}t + \hat{b} \quad \Longrightarrow \quad \text{Regresija v odvisnosti od časa} \tag{9.47}$$

kjer oznaka »^« pomeni izhod modela (aproksimacijo meritve povpraševanja), pri čemer sta $\hat{a}, \hat{b}$ neznana parametra, ki ju želimo oceniti na osnovi danih podatkov (meritev).

Posnemanje meritev povpraševanja z modelom 1. reda (9.47) si lahko ponazorimo s sliko 211. Na tej sliki nam pike predstavljajo izmerjene meritve povpraševanja *d(t),* premica pa nam predstavlja potek modela, ki opisuje linearno naraščajoč trend teh meritev (ko sta seveda parametra $\hat{a}, \hat{b}$ že ocenjena iz podatkov).



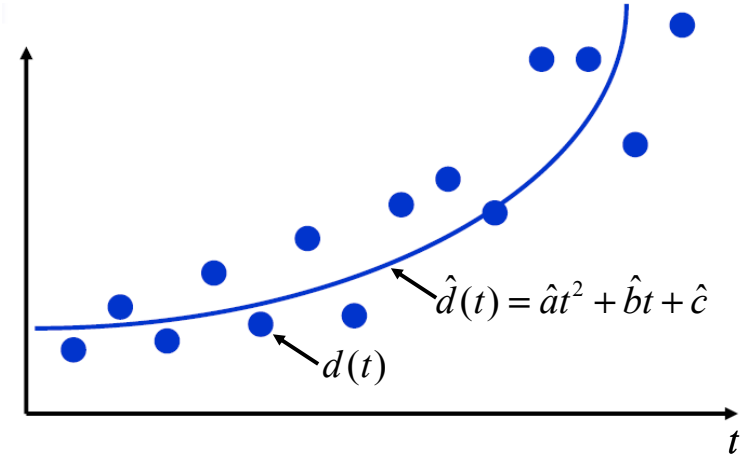
Slika 211: Posnemanje meritve povpraševanja z modelom 1. reda

Če pa npr. želimo posnemati obnašanje meritev s parabolo (polinomom 2. reda), bo imel linearni model (2. reda) naslednjo strukturo:

$$\hat{d}(t) = \hat{a}t^2 + \hat{b}t + \hat{c} \tag{9.48}$$

pri čemer so $\hat{a}, \hat{b}, \hat{c}$ neznani parametri, ki jih želimo oceniti na osnovi danih podatkov (meritev). Posnemanje meritev povpraševanja z modelom 2. reda (9.48) si lahko ponazorimo s sliko 212. Na tej sliki nam pike predstavljajo izmerjene meritve

povpraševanja d(t), parabola pa nam predstavlja potek modela, ki opisuje parabolično naraščajoč trend teh meritev (ko so seveda parametri $\hat{a},\hat{b},\hat{c}$ že ocenjeni iz podatkov).
Podobna logika kot za model 1. in 2. reda, velja tudi za modele višjih redov. Če meritve povpraševanja nakazujejo trend višjega reda, potem poskušamo uporabiti takšno stopnjo oz. red modela, ki bo dovolj dobro opisal tovrsten trend meritev.



Slika 212: Posnemanje meritve povpraševanja z modelom 2. reda

V nadaljevanju si poglejmo skalarno izpeljavo ocenjenih parametrov ($\hat{a},\hat{b}$), če imamo linearni model 1. reda (9.47). Seveda izpeljava poteka podobno, kot smo že pokazali prej za regresijsko funkcijo splošne neodvisne spremenljivke.

Skalarna izpeljava ocenjenih parametrov za linearni model 1. reda

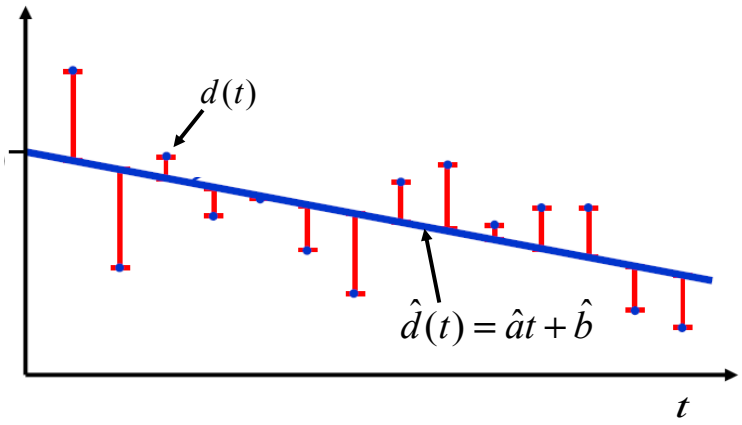
Najprej definirajmo pogrešek modela, ki je razlika med meritvami povpraševanja *d(t)* in napovedmi modela $\hat{d}(t) = \hat{a}t + \hat{b}$ v posameznih časovnih trenutkih $t = 1,2,...,N$:

$$e(t) = d(t) - \hat{d}(t), \quad t = 1,2,...,N \tag{9.49}$$

Definirajmo tudi kriterijsko funkcijo kot vsoto kvadratov pogreškov modela v posameznih časovnih trenutkih $t = 1,2,...,N$, pri čemer upoštevamo tudi relacijo iz izraza (9.47):

$$J(\hat{a},\hat{b}) = \frac{1}{N}.\sum_{t=1:N} e^2(t) = \frac{1}{N}.\sum_{t=1:N}\left[d(t)-\hat{d}(t)\right]^2 = \frac{1}{N}.\sum_{t=1:N}\left[d(t)-\left(\hat{a}t+\hat{b}\right)\right]^2 \qquad (9.50)$$

Slika 213 prikazuje pogreške modela (navpične črte) v posameznih časovnih trenutkih $t = 1,2,...,N$, ki predstavljajo razlike med meritvami povpraševanja *d(t)* in napovedmi modela $\hat{d}(t) = \hat{a}t + \hat{b}$. Če bomo uspeli minimizirati kriterijsko funkcijo (9.50), to je voto kvadratov pogreškov, bomo gotovo dosegli, da se bo premica $\hat{d}(t) = \hat{a}t + \hat{b}$ najtesneje (najbolje) prilegala meritvam *d(t),* model pa bo najbolje opisoval obnašanje oz. potek njihovega linearnega trenda.



Slika 213: Ilustracija pogreškov modela

Iz izraza (9.50) je razvidno, da je kriterijska funkcija *J* odvisna od dveh parametrov $\hat{a},\hat{b}$. Problem minimizacije te kriterijske funkcije se torej preslika na iskanje takšnih parametrov $\hat{a},\hat{b}$, pri katerih bo funkcija *J* minimalna. Očitno gre za problem iskanja ekstrema (minimuma) funkcije *J*, zato moramo poiskati njena parcialna odvoda po parametrih $\hat{a},\hat{b}$, ter ju izenačiti z 0. Matematično to zapišemo na naslednji način:

$$\frac{\partial}{\partial \hat{a}} J(\hat{a},\hat{b}) = 0 \qquad \frac{\partial}{\partial \hat{b}} J(\hat{a},\hat{b}) = 0 \tag{9.51}$$

Po obeh parcialnih odvajanjih izraza in krajši izpeljavi dobimo naslednji sistem 2 enačb z dvema neznankama $\hat{a},\hat{b}$ [Dragan 1]:

$$\begin{aligned} \frac{1}{N}.\sum_{t=1:N} t.d(t) &= \frac{1}{N}.\hat{a}.\sum_{t=1:N} t^2 + \frac{1}{N}.\hat{b}.\sum_{t=1:N} t, \\ \frac{1}{N}.\sum_{t=1:N} d(t) &= \frac{1}{N}.\hat{a}.\sum_{t=1:N} t + \hat{b} \end{aligned} \tag{9.52}$$

Če rešimo ta sistem enačb, dobimo naslednja rezultata za ocenjena parametra $\hat{a},\hat{b}$ [Dragan 1]:

$$\hat{a} = \frac{\frac{1}{N}\sum_{t=1:N} t.d(t) - \bar{t}.\bar{d}}{\frac{1}{N}\sum_{t=1:N} t^2 - \bar{t}^2}, \quad kjer\ \ je: \quad \bar{t} = \frac{1}{N}.\sum_{t=1:N} t; \quad \bar{d} = \frac{1}{N}.\sum_{t=1:N} d(t); \tag{9.53}$$

$$\hat{b} = \bar{d} - \hat{a}.\bar{t} \tag{9.54}$$

Kot vidimo, dobimo zelo podobne rezultate kot pri izpeljavi regresijske funkcije splošne neodvisne spremenljivke, le da smo obrnili parametra $\hat{a},\hat{b}$. Z opazovanjem rezultatov (9.53) in (9.54) lahko vidimo, da ocenjena parametra zavisita le od meritev povpraševanja

d(t) in časov $t = 1,2,...,N$, torej ju dejansko lahko ocenimo le na osnovi danih podatkov (meritev).

Iz pravkar opisanega postopka za ocenjevanje parametrov $\hat{a},\hat{b}$ je razvidno, da se že pri izpeljavi modela 1. reda soočimo z dokaj zapletenimi računskimi operacijami. Seveda kompleksnost tovrstnih izpeljav pri modelih višjih redov strmo narašča, če uporabimo takšen postopek.

Zato je bolj priporočljivo, da se poslužimo vektorsko matrične izpeljave ocenjenih parametrov. Sicer bomo le-to v nadaljevanju spet ponazorili le za model 1. reda, vendar pa velja poudariti, da je dobljeni rezultat, vektor ocenjenih parametrov, splošen in velja za poljuben red modela, če seveda prav sestavimo matriko $\boldsymbol{\Psi}$, ki se v rezultatu pojavi.

Vektorsko matrična izpeljava ocenjenih parametrov za linearni model 1. reda

Če meritve povpraševanja predstavimo v vektorski obliki, dobimo naslednji vektor danih meritev povpraševanja:

$$\underbrace{\begin{bmatrix} d(1) \\ ... \\ d(N) \end{bmatrix}}_{\mathbf{d}} \rightarrow meritve \tag{9.55}$$

Za posamezne časovne trenutke $t = 1,2,...,N$ tvorimo nato naslednji sistem enačb, ki ga dobimo, če čase $t = 1,2,...,N$ vstavljamo v model 1. reda (9.47):

$$\hat{d}(t)=\hat{a}t+\hat{b}\rightarrow\begin{cases}\hat{d}(1)=\hat{a}.1+\hat{b}\\ \ldots\\ \hat{d}(N)=\hat{a}.N+\hat{b}\end{cases}\rightarrow\underbrace{\begin{bmatrix}\hat{d}(1)\\ \ldots\\ \hat{d}(N)\end{bmatrix}}_{\hat{\mathbf{d}}}=\underbrace{\begin{bmatrix}1 & 1\\ 2 & 1\\ \ldots & \ldots\\ N & 1\end{bmatrix}}_{\boldsymbol{\Psi}}.\underbrace{\begin{bmatrix}\hat{a}\\ \hat{b}\end{bmatrix}}_{\hat{\boldsymbol{\Theta}}} \tag{9.56}$$

pri čemer smo dobljeni sistem enačb (9.56) zapisali v bolj kompaktni vektorsko matrični obliki $\hat{\mathbf{d}}=\boldsymbol{\Psi}.\hat{\boldsymbol{\Theta}}$**.** Pri tem je $\hat{\mathbf{d}}$ vektor napovedi (ocen) modela za posamezne časovne trenutke $t=1,2,...,N$, $\boldsymbol{\Psi}$ je takoimenovana regresorska (informacijska) matrika, $\hat{\boldsymbol{\Theta}}$ pa je vektor ocenjenih parametrov $\hat{a},\hat{b}$. Ker je v sistemu enačb (9.56) več enačb (*N*) kot neznank (le dve), gre za primer **predoločenega sistema enačb**.

Na osnovi izrazov (9.55) in (9.56) lahko tvorimo tudi vektor pogreškov kot razliko vektorjev meritev povpraševanja in njihovih ocen, ki jih daje model:

$$\mathbf{e}=\mathbf{d}-\hat{\mathbf{d}}=\mathbf{d}-\boldsymbol{\Psi}.\hat{\boldsymbol{\Theta}} \tag{9.57}$$

Definirajmo sedaj kriterijsko funkcijo kot vsoto kvadratov pogreškov modela v posameznih časovnih trenutkih $t=1,2,...,N$, podobno, kot smo to storili pri skalarni izpeljavi. Za razliko od slednje pa tokrat vsoto kvadratov pogreškov raje izrazimo s produktom $\mathbf{e}^{\mathbf{T}}.\mathbf{e}$, pri čemer upoštevamo tudi izraz (9.57). Tako dobimo naslednjo skalarno funkcijo *J* vektorske spremenljivke $\hat{\boldsymbol{\Theta}}$**:**

$$J(\hat{\boldsymbol{\Theta}})=\frac{1}{N}.\sum_{t=1:N}e^2(t)=\frac{1}{N}.\mathbf{e}^{\mathbf{T}}\mathbf{e}=\frac{1}{N}.\left(\mathbf{d}-\boldsymbol{\Psi}.\hat{\boldsymbol{\Theta}}\right)^{\mathbf{T}}\left(\mathbf{d}-\boldsymbol{\Psi}.\hat{\boldsymbol{\Theta}}\right) \tag{9.58}$$

Dobljeni izraz še malce preuredimo z medsebojnim množenjem matrik oz. vektorjev, pri čemer dobimo:

$$J(\hat{\boldsymbol{\Theta}}) = \frac{1}{N}.\left(\mathbf{d}^{\mathbf{T}}\mathbf{d} - \mathbf{d}^{\mathbf{T}}\boldsymbol{\Psi}.\hat{\boldsymbol{\Theta}} - \hat{\boldsymbol{\Theta}}^{\mathbf{T}}\boldsymbol{\Psi}^{\mathbf{T}}\mathbf{d} + \hat{\boldsymbol{\Theta}}^{\mathbf{T}}\boldsymbol{\Psi}^{\mathbf{T}}\boldsymbol{\Psi}.\hat{\boldsymbol{\Theta}}\right) \tag{9.59}$$

Iz izraza (9.59) je razvidno, da je kriterijska funkcija *J* odvisna od vektorja ocenjenih parametrov $\hat{\boldsymbol{\Theta}}$. Problem minimizacije te kriterijske funkcije se torej preslika na iskanje takšnega vektorja $\hat{\boldsymbol{\Theta}}$, pri katerem bo funkcija *J* minimalna. Očitno gre za problem iskanja ekstrema (minimuma) funkcije *J*, zato moramo poiskati njen parcialni odvod po vektorju $\hat{\boldsymbol{\Theta}}$, ter dobljeni rezultat izenačiti z 0. Matematično to zapišemo na naslednji način:

$$\begin{aligned}
&\frac{\partial}{\partial\hat{\boldsymbol{\Theta}}}J(\hat{\boldsymbol{\Theta}}) = \frac{1}{N}.\left(0 - \left(\mathbf{d}^{\mathbf{T}}\boldsymbol{\Psi}\right)^{T} - \boldsymbol{\Psi}^{\mathbf{T}}\mathbf{d} + 2\boldsymbol{\Psi}^{\mathbf{T}}\boldsymbol{\Psi}.\hat{\boldsymbol{\Theta}}\right) = \\
&= \frac{1}{N}.\left(-2\boldsymbol{\Psi}^{\mathbf{T}}\mathbf{d} + 2\boldsymbol{\Psi}^{\mathbf{T}}\boldsymbol{\Psi}.\hat{\boldsymbol{\Theta}}\right) = 0 \rightarrow \\
&\rightarrow \boldsymbol{\Psi}^{\mathbf{T}}\mathbf{d} = \boldsymbol{\Psi}^{\mathbf{T}}\boldsymbol{\Psi}.\hat{\boldsymbol{\Theta}}
\end{aligned} \tag{9.60}$$

pri čemer smo uporabili tudi pravila za odvajanje skalarne funkcije po vektorski spremenljivki [Matko].

Če izraz (9.60) še malce preoblikujemo, dobimo naslednji rezultat za vektor ocenjenih parametrov [Matko, Dragan 1]:

$$\hat{\boldsymbol{\Theta}} = \left(\boldsymbol{\Psi}^{\mathbf{T}}\boldsymbol{\Psi}\right)^{-1}\boldsymbol{\Psi}^{\mathbf{T}}\mathbf{d} = \begin{bmatrix}\hat{a}\\ \hat{b}\end{bmatrix} \tag{9.61}$$

Vektor ocenjenih parametrov zavisi le od meritev povpraševanja *d(t)* in časov $t = 1,2,...,N$ v regresorski matriki, torej ga dejansko lahko ocenimo le na osnovi danih podatkov (meritev).

Seveda nam rezultat (9.61) da popolnoma enaka ocenjena parametra $\hat{a},\hat{b}$, kot bi ju dobili pri skalarni izpeljavi. Je pa rezultat (9.61) bolj splošen, saj velja za poljuben red modela, če seveda prav sestavimo matriko $\boldsymbol{\Psi}$, ki se v rezultatu pojavi. Po drugi strani pa rezultata (9.53) in (9.54) veljata le za model 1. reda.

Tako npr. izraz (9.61) velja tudi za model 2. reda s strukturo:

$$\hat{d}(t) = \hat{a}.t^2 + \hat{b}\cdot t + \hat{c}, \quad t = 1,2,...,N \tag{9.62}$$

le da ima v tem primeru sistem enačb (9.56) malce drugačno obliko, in sicer:

$$\underbrace{\begin{bmatrix} \hat{d}(1) \\ \hat{d}(2) \\ \dots \\ \dots \\ \hat{d}(N) \end{bmatrix}}_{\hat{\mathbf{d}}} = \underbrace{\begin{bmatrix} 1^2 & 1 & 1 \\ 2^2 & 2 & 1 \\ 3^2 & 3 & 1 \\ \dots & \dots & 1 \\ N^2 & N & 1 \end{bmatrix}}_{\boldsymbol{\Psi}} . \underbrace{\begin{bmatrix} \hat{a} \\ \hat{b} \\ \hat{c} \end{bmatrix}}_{\hat{\boldsymbol{\Theta}}} \tag{9.63}$$

torej je regresorska matrika $\boldsymbol{\Psi}$ malce drugačna od primera s 1. redom modela, vektor ocenjenih parametrov $\hat{\boldsymbol{\Theta}}$ pa tokrat vsebuje tri ocenjene parametre. Rezultat zanj je torej v tem primeru naslednji:

$$\hat{\boldsymbol{\Theta}} = \left(\boldsymbol{\Psi}^{\mathrm{T}}\boldsymbol{\Psi}\right)^{-1}\boldsymbol{\Psi}^{\mathrm{T}}\mathbf{d} = \begin{bmatrix} \hat{a} \\ \hat{b} \\ \hat{c} \end{bmatrix} \tag{9.64}$$

kjer je struktura matrike $\boldsymbol{\Psi}$ določena v izrazu (9.63). Seveda pa ima vektor meritev povpraševanja enako obliko, kot jo je imel v izrazu (9.55).

Poudarimo še, da so rezultati vektorsko matrične izpeljave še posebej primerni za izračun ocenjenih parametrov, ko tovrstna izračunavanja opravimo z računalnikom oz. računalniškim programom. Pri "ročnem računanju" pa izpeljava ocenjenih parametrov na takšen način načeloma ni tako enostavna, saj imamo lahko kar nekaj dela z invertiranjem matrike.

Ko so ocenjeni parametri modela enkrat izračunani, lahko na osnovi izraza (9.47) izračunamo ocene modela, to je aproksimacije meritev povpraševanja, za pretekli časovni horizont $t = 1, 2, ..., N$:

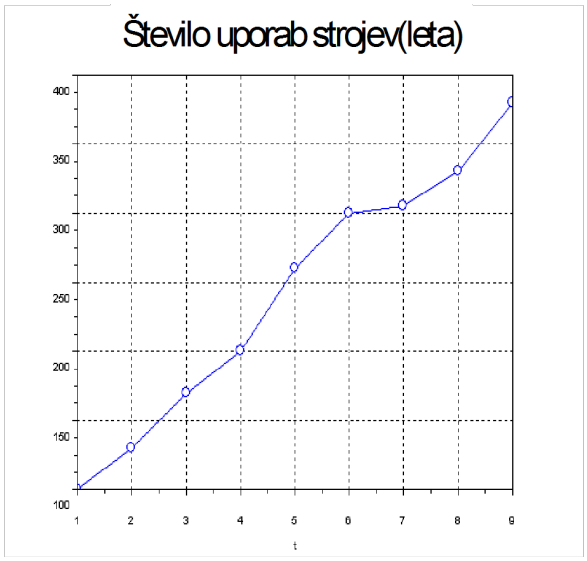
$$\begin{aligned} \hat{d}(1) &= \hat{a}.1 + \hat{b} \\ &... \\ \hat{d}(N) &= \hat{a}.N + \hat{b} \end{aligned} \tag{9.65}$$

Dobljeni regresijski model pa seveda lahko uporabimo tudi za napovedovanje trenda časovnih vrst povpraševanja v prihodnosti.

Primer 9.5.:

Banka želi predvideti število uporab strojev za štetje denarja (v milijonih) v odvisnosti od časa (let). Ta odvisnost je za prvih 9 let prikazana v tabeli na sliki 214. Kolikšno bo predvideno število uporab strojev v 10., 11. in 12. letu, če za oceno povpraševanja uporabimo premico? [Dragan 1]

Leto t	Število uporab strojev d(t)
1	100
2	130
3	170
4	200
5	260
6	300
7	305
8	330
9	380



Slika 214.: Časovni potek števila uporab strojev (v milijonih) v posameznih letih in tabelarični prikaz numeričnih vrednosti. t = čas v letih, d(t) = Število uporab strojev (čas v letih)

Ocenjena parametra $\hat{a},\hat{b}$ dobimo na osnovi izraza:

$$\hat{a}=\frac{\frac{1}{N}\sum_{t=1:N} t.d(t)-\overline{t}.\overline{d}}{\frac{1}{N}\sum_{t=1:N} t^2-\overline{t}^2} \qquad (9.66)$$

$$\hat{b}=\overline{d}-\hat{a}.\overline{t}$$

Najprej se lotimo izračuna za ocenjeni parameter $\hat{a}$. V ta namen najprej izračunamo povprečni vrednosti časa in meritev povpraševanja, ki se se glasita (glej podatke na sliki 214, kjer imamo danih $N = 9$ meritev):

$$\begin{aligned}\overline{t} &= \frac{1}{N}.\sum_{t=1:N} t = \frac{1}{9}(1+2+...+9)=5\\ \overline{d} &= \frac{1}{N}.\sum_{t=1:N} d(t) = \frac{1}{9}(100+130+170+...+380)=241.66667\end{aligned} \qquad (9.67)$$

Ostale delne izraze za parameter $\hat{a}$ izračunamo na naslednji način (glej podatke na sliki 214):

$$\frac{1}{N}\sum_{t=1:N} t.d(t)=\frac{1}{9}(1.100+2.130+3.170+...+9.380)=1440.5556$$

$$\frac{1}{N}\sum_{t=1:N} t^2=\frac{1}{9}(1^2+2^2+...+9^2)=31.666667 \qquad (9.68)$$

$$\overline{t}.\overline{d}=5\cdot 241.66667=1208.3333$$

$$\overline{t}^2=5^2=25$$

Tako dobimo naslednji vrednosti za ocenjena parametra modela $\hat{a},\hat{b}$:

$$\hat{a} = \frac{\frac{1}{N}\sum_{t=1:N} t.d(t) - \bar{t}\,.\bar{d}}{\frac{1}{N}\sum_{t=1:N} t^2 - \bar{t}^{\,2}} = \frac{1440.556 - 1208.3333}{31.666667 - 25} = 34.833333 \tag{9.69}$$

$$\hat{b} = \bar{d} - \hat{a}.\bar{t} = 241.66667 - 34.833333 \cdot 5 = 67.5$$

Enačba modela se glasi:

$$\begin{aligned} &\hat{d}(t) = \hat{a}t + \hat{b}, \quad t = 1,...,9,... \\ &\hat{d}(t) = 34.833333 \cdot t + 67.5, \quad t = 1,...,9,... \end{aligned} \tag{9.70}$$

Napovedi modela za gibanje trenda gibanja meritev povpraševanja v prihodnjem časovnem horizontu ($t > N$) se potemtakem glasijo:

$$\begin{aligned} &\hat{d}(10) = 34.833333 \cdot 10 + 67.5 = 415.83333 \\ &\hat{d}(11) = 34.833333 \cdot 11 + 67.5 = 450.66667 \\ &\hat{d}(12) = 34.833333 \cdot 12 + 67.5 = 485.5 \end{aligned} \tag{9.71}$$

Izračunajmo ocenjena parametra še na vektorsko matrični način. Če meritve povpraševanja predstavimo v vektorski obliki, dobimo naslednji vektor danih meritev povpraševanja:

$$\underbrace{\begin{bmatrix} d(1) \\ \dots \\ d(9) \end{bmatrix}}_{\mathbf{d}} \rightarrow meritve \qquad 1\ 2\ 3$$

$$\underbrace{\begin{bmatrix} 100 \\ \dots \\ 380 \end{bmatrix}}_{\mathbf{d}} \rightarrow meritve \tag{9.72}$$

Za posamezne časovne trenutke $t = 1, 2, ..., 9$ tvorimo nato naslednji sistem enačb, ki ga dobimo, če čase $t = 1, 2, ..., 9$ vstavljamo v model 1. reda (9.47):

$$\hat{d}(t) = \hat{a}t + \hat{b} \rightarrow \begin{cases} \hat{d}(1) = \hat{a} \cdot 1 + \hat{b} \\ \dots \\ \hat{d}(9) = \hat{a} \cdot 9 + \hat{b} \end{cases} \rightarrow \underbrace{\begin{bmatrix} \hat{d}(1) \\ \dots \\ \hat{d}(9) \end{bmatrix}}_{\hat{\mathbf{d}}} = \underbrace{\begin{bmatrix} 1 & 1 \\ 2 & 1 \\ \dots & \dots \\ 9 & 1 \end{bmatrix}}_{\boldsymbol{\Psi}} \cdot \underbrace{\begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix}}_{\hat{\boldsymbol{\Theta}}} \tag{9.73}$$

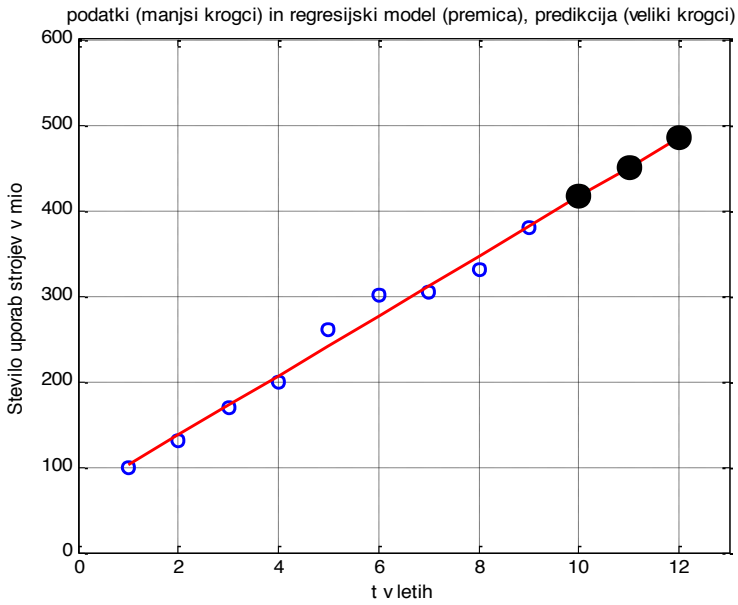
Vektor ocenjenih parametrov je enak:

$$\hat{\boldsymbol{\Theta}} = \left(\boldsymbol{\Psi}^{\mathbf{T}}\boldsymbol{\Psi}\right)^{-1}\boldsymbol{\Psi}^{\mathbf{T}}\mathbf{d} = \begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix}$$

$$\hat{\boldsymbol{\Theta}} = \begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix} = \left(\begin{bmatrix} 1 & 1 \\ 2 & 1 \\ \dots & \dots \\ 9 & 1 \end{bmatrix}^T \cdot \begin{bmatrix} 1 & 1 \\ 2 & 1 \\ \dots & \dots \\ 9 & 1 \end{bmatrix} \right)^{-1} \cdot \begin{bmatrix} 1 & 1 \\ 2 & 1 \\ \dots & \dots \\ 9 & 1 \end{bmatrix}^T \cdot \begin{bmatrix} 100 \\ \dots \\ 380 \end{bmatrix} = \begin{bmatrix} 34.833 \\ 67.5 \end{bmatrix} \tag{9.74}$$

Seveda dobimo enak rezultat kot pri skalarnem računanju.

Slika 215 prikazuje podatke, ocenjeno regresijsko premico ter napovedi uporab strojev za 10., 11. in 12. leto.



Slika 215: Podatki za prvih 9 let, ocenjena regresijska premica ter napovedi uporab strojev za 10., 11. in 12. leto.

Pri izrisu slike 215 in vseh izračunih smo si pomagali z naslednjim programom v Matlabu:

```
% regresija.m

clear
clc
close all

% podatki:

t = 1:1:9
d = [100 130 170 200 260 300 305 330 380]

N = length(d);

%------------------------------------------------------------------------
%  1. Skalarni izracun
%------------------------------------------------------------------------

disp('Skalarni izračun:')

t_s = sum(t)/N;
d_s = sum(d)/N;

disp('t_s=')
t_s
disp('d_s=')
d_s

mean_t_d = t*d'/N;
mean_t_t = t*t'/N;

disp('mean_t_d=')
mean_t_d
disp('mean_t_t=')
mean_t_t

clen1 = mean_t_d
```

```
clen2 = t_s*d_s
clen3 = mean_t_t
clen4 = t_s*t_s

% ocenjena parametra:

a = (clen1 - clen2)/(clen3 - clen4)
b = d_s - a*t_s

plot(t,d,'o','LineWidth',2)
hold on
grid
xlabel('t v letih')
ylabel('Stevilo uporab strojev v mio')
title('podatki (manjsi krogci) in regresijski model (premica), predikcija (veliki
krogci)')

t1 = 1:1:12;

d_o = a*t1+b;    % model

plot(t1,d_o,'r','LineWidth',1.5)
plot(t1(10:12),d_o(10:12),'ko','LineWidth',7)

axis([0 13 0 600])



disp('d_oc(10)     d_oc(11)       d_oc(12)')
[d_o(10) d_o(11) d_o(12)]


%------------------------------------------------------------------------
%  2. Vektorski izracun izracun
%------------------------------------------------------------------------

disp('Vektorski izracun')

psi = [t' ones(N,1)]

theta = inv(psi'*psi)*psi'*d';

a = theta(1)
b = theta(2)
```

Komandno okno ima naslednji izgled:

```
t =
   1   2   3   4   5   6   7   8   9
d =
  100  130  170  200  260  300  305  330  380


Skalarni izračun:
t_s=
t_s =
   5
d_s=
d_s =
  241.6667
mean_t_d=
mean_t_d =
  1.4406e+003
```

```
mean_t_t=
mean_t_t =
   31.6667

clen1 =
  1.4406e+003
clen2 =
  1.2083e+003
clen3 =
   31.6667
clen4 =
    25

a =
   34.8333
b =
   67.5000

d_oc(10)     d_oc(11)       d_oc(12)
ans =
  415.8333  450.6667  485.5000

Vektorski izracun

psi =
     1     1
     2     1
     3     1
     4     1
     5     1
     6     1
     7     1
     8     1
     9     1

a =
   34.8333
b =
   67.5000
```

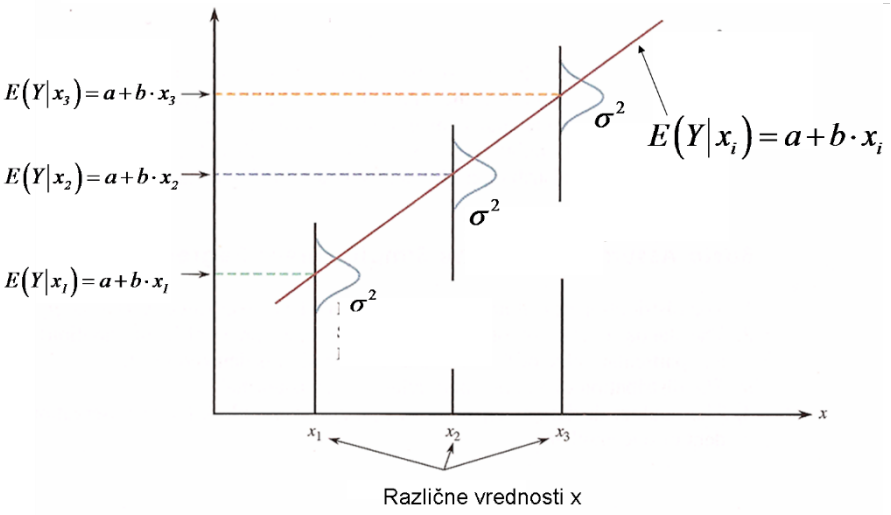
9.4.3 Statistike, hipoteze in intervali zaupanja pri linearni regresijski premici

Pri določanju regresijskega modela smo vrednosti neodvisne spremenljivke $x_i, i=1,...,n$ obravnavali kot konstante, pripadajoče vrednosti odvisne spremenljivke $y_i, i=1,...,n$ pa kot realizacije naključne spremenljivke $Y|x_i$. Ker predpostavljamo, da je napaka modela normalna naključna spremenljivka $\varepsilon \in N(0,\sigma)$ z normalnim porazdelitvenim zakonom, sledi (glej sliko 216):

$$
\begin{aligned}
&p(\varepsilon) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\varepsilon^2}{2\sigma^2}} \\
&p\left(y-(a+b\cdot x_i)\right) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\left[y-(a+b\cdot x_i)\right]^2}{2\sigma^2}} = p\left(y|x_i\right) \\
&in: \\
&E\left(Y|x_i\right) = a+b\cdot x_i = f\left(x_i\right) \\
&STD\left(Y|x_i\right) = \sigma \\
&Y|x_i \in N\left(E\left(Y|x_i\right), STD\left(Y|x_i\right)\right) = N\left(a+b\cdot x_i, \sigma\right)
\end{aligned}
\qquad (9.75)
$$

kjer so parametri a,b,σ enaki za vsak indeks *i*.

$$Y|x_i \in N\left(E\left(Y|x_i\right), STD\left(Y|x_i\right)\right) = N\left(a+b\cdot x_i, \sigma\right)$$



Slika 216: Pomen spreminjajoče se srednje vrednosti in standardnega odklona odvisne spremenljivke

Naloga regresijske analize se v glavnem nanaša na določanje ocen teh parametrov, testiranje hipotez o njih in na napovedi, določene z ocenjeno regresijsko funkcijo.

Poiščimo ocene $\hat{a},\hat{b},\hat{\sigma}=s$ za parametre a,b,σ z metodo največje verjetnosti. Funkcija največje verjetnosti je [Jesenko]:

$$L(\hat{a},\hat{b},s)=\frac{1}{\sqrt{2\pi}s}e^{-\frac{\left[y_1-\left(\hat{a}+\hat{b}\cdot x_1\right)\right]^2}{2\cdot s^2}}\cdot\frac{1}{\sqrt{2\pi}s}e^{-\frac{\left[y_2-\left(\hat{a}+\hat{b}\cdot x_2\right)\right]^2}{2\cdot s^2}}\cdot\ldots\cdot\frac{1}{\sqrt{2\pi}s}e^{-\frac{\left[y_n-\left(\hat{a}+\hat{b}\cdot x_n\right)\right]^2}{2\cdot s^2}} \tag{9.76}$$

$$=\left(\frac{1}{\sqrt{2\pi}s}\right)^n\cdot e^{-\frac{\left[y_1-\left(\hat{a}+\hat{b}\cdot x_1\right)\right]^2}{2\cdot s^2}-\frac{\left[y_2-\left(\hat{a}+\hat{b}\cdot x_2\right)\right]^2}{2\cdot s^2}-\ldots-\frac{\left[y_n-\left(\hat{a}+\hat{b}\cdot x_n\right)\right]^2}{2\cdot s^2}}=\left(\frac{1}{\sqrt{2\pi}s}\right)^n\cdot e^{-\frac{1}{2\cdot s^2}\cdot\sum_{i=1}^{n}\left[y_i-\left(\hat{a}+\hat{b}\cdot x_i\right)\right]^2}$$

Če jo logaritmiramo, dobimo:

$$\begin{aligned}
&ln\,L(\hat{a},\hat{b},s)=ln\left(\left(\frac{1}{\sqrt{2\pi}s}\right)^n\cdot e^{-\frac{1}{2\cdot s^2}\cdot\sum_{i=1}^{n}\left[y_i-\left(\hat{a}+\hat{b}\cdot x_i\right)\right]^2}\right)=\\
&=ln\left(\left(\frac{1}{\sqrt{2\pi}s}\right)^n\right)+ln\left(e^{-\frac{1}{2\cdot s^2}\cdot\sum_{i=1}^{n}\left[y_i-\left(\hat{a}+\hat{b}\cdot x_i\right)\right]^2}\right)=\\
&=n\cdot ln\left(\left(\sqrt{2\pi}s\right)^{-1}\right)-\frac{1}{2\cdot s^2}\cdot\sum_{i=1}^{n}\left[y_i-\left(\hat{a}+\hat{b}\cdot x_i\right)\right]^2=\\
&=-n\cdot\left[ln\left(2\pi\right)^{\frac{1}{2}}+ln\left(s\right)\right]-\frac{1}{2\cdot s^2}\cdot\sum_{i=1}^{n}\left[y_i-\left(\hat{a}+\hat{b}\cdot x_i\right)\right]^2=\\
&=-n\cdot ln\left(s\right)-\frac{n}{2}\cdot ln\left(2\pi\right)-\frac{1}{2\cdot s^2}\cdot\sum_{i=1}^{n}\left[y_i-\left(\hat{a}+\hat{b}\cdot x_i\right)\right]^2
\end{aligned} \tag{9.77}$$

Parcialno sedaj odvajajmo logaritmsko funkcijo največjega verjetja po $\hat{a},\hat{b},\hat{\sigma}=s$ in enačimo z nič. Dobimo:

$$\frac{\partial}{\partial \hat{a}} ln\, L(\hat{a},\hat{b},s) = \frac{\partial}{\partial \hat{a}}\left(-n\cdot ln(s) - \frac{n}{2}\cdot ln(2\pi) - \frac{1}{2\cdot s^2}\cdot \sum_{i=1}^{n}\left[y_i - \left(\hat{a}+\hat{b}\cdot x_i\right)\right]^2 \right) =$$
$$= \frac{2}{2\cdot s^2}\sum_{i=1}^{n}\left[y_i - \left(\hat{a}+\hat{b}\cdot x_i\right)\right] = 0 \rightarrow \ \frac{1}{s^2}\sum_{i=1}^{n}\left[y_i - \left(\hat{a}+\hat{b}\cdot x_i\right)\right] = 0$$

$$\frac{\partial}{\partial \hat{b}} ln\, L(\hat{a},\hat{b},s) = \frac{\partial}{\partial \hat{b}}\left(-n\cdot ln(s) - \frac{n}{2}\cdot ln(2\pi) - \frac{1}{2\cdot s^2}\cdot \sum_{i=1}^{n}\left[y_i - \left(\hat{a}+\hat{b}\cdot x_i\right)\right]^2 \right) =$$
$$= \frac{2}{2\cdot s^2}\sum_{i=1}^{n}\left[\left\{ y_i - \left(\hat{a}+\hat{b}\cdot x_i\right)\right\}\cdot x_i \right] = 0 \rightarrow \ \frac{1}{s^2}\sum_{i=1}^{n}\left[\left\{ y_i - \left(\hat{a}+\hat{b}\cdot x_i\right)\right\}\cdot x_i \right] = 0 \qquad (9.78)$$

$$\frac{\partial}{\partial s} ln\, L(\hat{a},\hat{b},s) = \frac{\partial}{\partial s}\left(-n\cdot ln(s) - \frac{n}{2}\cdot ln(2\pi) - \frac{1}{2\cdot s^2}\cdot \sum_{i=1}^{n}\left[y_i - \left(\hat{a}+\hat{b}\cdot x_i\right)\right]^2 \right) =$$
$$= -\frac{n}{s} - \frac{1}{2}\cdot \sum_{i=1}^{n}\left[y_i - \left(\hat{a}+\hat{b}\cdot x_i\right)\right]^2 \left(\frac{-2}{s^3} \right) = -\frac{n}{s} + \frac{1}{s^3}\cdot \sum_{i=1}^{n}\left[y_i - \left(\hat{a}+\hat{b}\cdot x_i\right)\right]^2 = 0$$

Če prvi dve enačbi množimo z vrednostjo s^2, dobimo za ocenjena parametra $\hat{a},\hat{b}$ enaka rezultata (9.27) in (9.28), kot pri metodi najmanjših kvadratov, če bi izhajali iz izrazov (9.18) in (9.19). Rešimo še tretjo enačbo, da dobimo (pri ocenjenih $\hat{a},\hat{b}$) še oceno največje verjetnosti za parameter σ, to je $\hat{\sigma} = s$:

$$\begin{aligned}
&-\frac{n}{s} + \frac{1}{s^3}\cdot \sum_{i=1}^{n}\left[y_i - \left(\hat{a}+\hat{b}\cdot x_i\right)\right]^2 = 0 \\
&-n + \frac{1}{s^2}\cdot \sum_{i=1}^{n}\left[y_i - \left(\hat{a}+\hat{b}\cdot x_i\right)\right]^2 = 0 \\
&\frac{1}{s^2}\cdot \sum_{i=1}^{n}\left[y_i - \left(\hat{a}+\hat{b}\cdot x_i\right)\right]^2 = n \\
&s^2 = \frac{1}{n}\cdot \sum_{i=1}^{n}\left[y_i - \left(\hat{a}+\hat{b}\cdot x_i\right)\right]^2 \\
&s = \sqrt{\frac{1}{n}\cdot \sum_{i=1}^{n}\left[y_i - \left(\hat{a}+\hat{b}\cdot x_i\right)\right]^2}
\end{aligned} \qquad (9.79)$$

Poskušajmo dobljeni izraz izraziti z veličinami $\hat{b}, S_{xx}, S_{xy}, S_{yy}$. Veličini S_{xx}, S_{xy} smo že omenili v prejšnjih izpeljavah. V splošnem lahko veličine S_{xx}, S_{xy}, S_{yy} izrazimo tudi na naslednji način [Jesenko]:

$$
\begin{aligned}
S_{xx} &= \sum_{i=1}^{n} \left[x_i - \bar{x} \right]^2 \\
S_{yy} &= \sum_{i=1}^{n} \left[y_i - \bar{y} \right]^2 \\
S_{xy} &= \sum_{i=1}^{n} \left[x_i - \bar{x} \right] \cdot \left[y_i - \bar{y} \right]
\end{aligned}
\tag{9.80}
$$

Nadalje lahko zapišemo:

$$
\begin{aligned}
&\left(\hat{a} + \hat{b} \cdot x_i \right) = \bar{y} - \hat{b} \cdot \bar{x} + \hat{b} \cdot x_i = \bar{y} + \hat{b} \cdot \left(x_i - \bar{x} \right) \\
&\hat{a} \\
&y_i - \left(\hat{a} + \hat{b} \cdot x_i \right) = y_i - \left(\bar{y} + \hat{b} \cdot \left(x_i - \bar{x} \right) \right) = \left(y_i - \bar{y} \right) - \hat{b} \cdot \left(x_i - \bar{x} \right) \\
&\left(y_i - \left(\hat{a} + \hat{b} \cdot x_i \right) \right)^2 = \left(\left(y_i - \bar{y} \right) - \hat{b} \cdot \left(x_i - \bar{x} \right) \right)^2 = \\
&= \left(y_i - \bar{y} \right)^2 - 2 \cdot \hat{b} \cdot \left(x_i - \bar{x} \right) \cdot \left(y_i - \bar{y} \right) + \hat{b}^2 \cdot \left(x_i - \bar{x} \right)^2 \\
&\sum_{i=1}^{n} \left[y_i - \left(\hat{a} + \hat{b} \cdot x_i \right) \right]^2 = \sum_{i=1}^{n} \left[\left(y_i - \bar{y} \right)^2 - 2 \cdot \hat{b} \cdot \left(x_i - \bar{x} \right) \cdot \left(y_i - \bar{y} \right) + \hat{b}^2 \cdot \left(x_i - \bar{x} \right)^2 \right] = \\
&= \sum_{i=1}^{n} \left[\left(y_i - \bar{y} \right)^2 \right] - 2 \cdot \hat{b} \cdot \sum_{i=1}^{n} \left[\left(x_i - \bar{x} \right) \cdot \left(y_i - \bar{y} \right) \right] + \hat{b}^2 \cdot \sum_{i=1}^{n} \left[\left(x_i - \bar{x} \right)^2 \right] = \\
&= S_{yy} - 2 \cdot \hat{b} \cdot S_{xy} + \hat{b}^2 \cdot S_{xx} = S_{yy} - 2 \cdot \hat{b} \cdot S_{xy} + \left(\frac{S_{xy}}{S_{xx}} \right)^2 \cdot S_{xx} = \\
&= S_{yy} - 2 \cdot \hat{b} \cdot S_{xy} + \frac{S_{xy} \cdot S_{xy}}{S_{xx}} = S_{yy} - 2 \cdot \hat{b} \cdot S_{xy} + \hat{b} \cdot S_{xy} = S_{yy} - \hat{b} \cdot S_{xy}
\end{aligned}
\tag{9.81}
$$

Dobimo torej:

$$
s = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} \left[y_i - \left(\hat{a} + \hat{b} \cdot x_i \right) \right]^2} = \sqrt{\frac{1}{n} \cdot \left(S_{yy} - \hat{b} \cdot S_{xy} \right)}
\tag{9.82}
$$

S pomočjo ocen največje verjetnosti parametrov $\hat{a}, \hat{b}, \hat{\sigma} = s$ bomo v nadaljevanju določili njihove intervalske ocene in jih uporabili tudi pri testiranju hipotez, povezanih z njimi. Točkaste ocene, ki smo jih poiskali, so zgrajene na podatkih naključnih vzorcev, kar pomeni, da so realizacije nekih statistik [Jesenko]. Označimo z $\hat{B}$ statistiko, katere realizacija je ocena

regresijskega koeficienta b , ki smo jo označili z $\hat{b}$. Statistika je enaka [Jesenko]:

$$\begin{aligned}
&\hat{B}=\frac{S_{xY}}{S_{xx}}=\frac{\sum_{i=1}^{n}\left[x_i-\bar{x}\right]\cdot\left[Y|x_i-\bar{Y}\right]}{S_{xx}}=\frac{\sum_{i=1}^{n}\left[x_i-\bar{x}\right]\cdot\left[Y|x_i-E\left(Y|x_i\right)\right]}{S_{xx}}= \\
&=\frac{\sum_{i=1}^{n}\left[x_i-\bar{x}\right]\cdot Y|x_i}{S_{xx}}-\frac{\sum_{i=1}^{n}\left[x_i-\bar{x}\right]\cdot E\left(Y|x_i\right)}{S_{xx}}= \\
&=\frac{\sum_{i=1}^{n}\left[x_i-\bar{x}\right]\cdot Y|x_i}{S_{xx}}-\frac{E\left(Y|x_i\right)}{S_{xx}}\cdot\underbrace{\sum_{i=1}^{n}\left[x_i-\bar{x}\right]}_{0}= \\
&=\frac{\sum_{i=1}^{n}\left[x_i-\bar{x}\right]\cdot Y|x_i}{S_{xx}}=\sum_{i=1}^{n}\left(\frac{\left[x_i-\bar{x}\right]}{S_{xx}}\right)\cdot Y|x_i
\end{aligned} \tag{9.83}$$

Dobimo torej linearno kombinacijo naključnih spremenljivk $Y|x_i, i=1,...,n$. Ker so le-te normalne naključne spremenljivke (glej izraz (9.75)), je tudi njihova linearna kombinacija, to je spremenljivka $\hat{B}$, normalna naključna spremenljivka.

Poglejmo, kakšno je njeno matematično upanje [Jesenko]:

$$\begin{aligned}
&E\left(\hat{B}\right)=E\left(\sum_{i=1}^{n}\left(\frac{\left[x_i-\bar{x}\right]}{S_{xx}}\right)\cdot Y|x_i\right)=\sum_{i=1}^{n}\left(\frac{\left[x_i-\bar{x}\right]}{S_{xx}}\right)\cdot E\left(Y|x_i\right)= \\
&=\sum_{i=1}^{n}\left(\frac{\left[x_i-\bar{x}\right]}{S_{xx}}\right)\cdot\left(a+b\cdot x_i\right)
\end{aligned} \tag{9.84}$$

kjer smo upoštevali izraz (9.12). Sledi:

$$
\begin{aligned}
E\left(\hat{B}\right) &= \sum_{i=1}^{n}\left(\frac{[x_i-\bar{x}]}{S_{xx}}\right)\cdot(a)+\sum_{i=1}^{n}\left(\frac{[x_i-\bar{x}]}{S_{xx}}\right)\cdot(b\cdot x_i)= \\
&= \frac{a}{S_{xx}}\underbrace{\sum_{i=1}^{n}[x_i-\bar{x}]}_{0}+\frac{b}{S_{xx}}\left(\sum_{i=1}^{n}x_i^2-\bar{x}\cdot\sum_{i=1}^{n}(x_i)\right)= \\
&= \frac{b}{S_{xx}}\left(\sum_{i=1}^{n}x_i^2-\frac{\sum_{i=1}^{n}(x_i)}{n}\cdot\sum_{i=1}^{n}(x_i)\right)=\frac{b}{S_{xx}}\left(\sum_{i=1}^{n}x_i^2-\frac{1}{n}\cdot\left(\sum_{i=1}^{n}(x_i)\right)^2\right)= \\
&= \frac{b}{S_{xx}}\cdot S_{xx}=b
\end{aligned}
\tag{9.85}
$$

Varianca spremenljivke $\hat{B}$ pa je:

$$
\begin{aligned}
VAR\left(\hat{B}\right) &= VAR\left(\sum_{i=1}^{n}\left(\frac{[x_i-\bar{x}]}{S_{xx}}\right)\cdot Y|x_i\right)=\sum_{i=1}^{n}\left(\frac{[x_i-\bar{x}]}{S_{xx}}\right)^2\cdot VAR\left(Y|x_i\right)= \\
&= \sum_{i=1}^{n}\left(\frac{[x_i-\bar{x}]}{S_{xx}}\right)^2\cdot\sigma^2=\frac{\sigma^2}{S^2{}_{xx}}\cdot\sum_{i=1}^{n}\left([x_i-\bar{x}]\right)^2=\frac{\sigma^2}{S^2{}_{xx}}\cdot S_{xx}=\frac{\sigma^2}{S_{xx}}
\end{aligned}
\tag{9.86}
$$

Za normalno naključno spremenljivko $\hat{B}$ torej velja:

$$
\hat{B}\in N\left(E\left(\hat{B}\right),STD\left(\hat{B}\right)\right)=N\left(b,\sqrt{\frac{\sigma^2}{S_{xx}}}\right) \tag{9.87}
$$

Standardizirana normalna naključna spremenljivka zanjo je:

$$
Z=\frac{\hat{b}-b}{\sigma_{\hat{b}}}=\frac{\hat{b}-b}{\sqrt{\frac{\sigma^2}{S_{xx}}}}=\frac{\hat{b}-b}{\sigma}\cdot\sqrt{S_{xx}}\in N(0,1) \tag{9.88}
$$

Ker je σ neznana standardna deviacija, moramo uvesti t statistiko na naslednji način [Jesenko]:

$$t = \frac{\hat{b} - b}{\frac{s_\varepsilon}{\sqrt{S_{xx}}}} = \frac{\hat{b} - b}{s_\varepsilon}\sqrt{S_{xx}} \in t(n-2) \tag{9.89}$$

ki ima *n - 2* stopnji prostosti. Pri tem je s_ε **<u>standardna ocena napake modela</u>** in je enaka [Jesenko]:

$$s_\varepsilon = \sqrt{\frac{1}{n-2}\cdot\sum_{i=1}^{n}[\varepsilon_i]^2} = \sqrt{\frac{1}{n-2}\cdot\sum_{i=1}^{n}\left[y_i - \left(\hat{a} + \hat{b}\cdot x_i\right)\right]^2} = \tag{9.90}$$
$$= \sqrt{\frac{1}{n-2}\cdot\left(S_{yy} - \hat{b}\cdot S_{xy}\right)}$$

t statistika, na kateri temelji test hipoteze o regresijskem parametru *b*, je torej enaka:

$$t = \frac{\hat{b} - b}{\sqrt{\frac{1}{n-2}\cdot\left(S_{yy} - \hat{b}\cdot S_{xy}\right)}}\sqrt{S_{xx}} \in t(n-2) \tag{9.91}$$

Relacija med oceno največje verjetnosti za parameter σ, to je $\hat{\sigma} = s$, ter standardno oceno napake modela s_ε je pa enaka:

$$s_\varepsilon = \sqrt{\frac{n}{n-2}}\cdot s \tag{9.92}$$

Kot se izkaže, je ${s_\varepsilon}^2$ nepristranska ocena za σ^2. Velja namreč, da je spremenljivka $\frac{n\cdot s^2}{\sigma^2}$ porazdeljena po $\chi^2(n-2)$ porazdelitvi, pri čemer velja: $E\left(\frac{n\cdot s^2}{\sigma^2}\right) = E\left(\chi^2(n-2)\right) = n-2$ [Jesenko].

Sledi:

$$E\left(s_\varepsilon{}^2\right)=E\left(\frac{n}{n-2}\cdot s^2\right)=\frac{\sigma^2}{\sigma^2}\cdot E\left(\frac{n}{n-2}\cdot s^2\right)= \\ =\frac{\sigma^2}{n-2}\cdot E\left(\frac{n}{\sigma^2}\cdot s^2\right)=\frac{\sigma^2}{n-2}\cdot\left(n-2\right)=\sigma^2 \tag{9.93}$$

torej je $s_\varepsilon{}^2$ res nepristranska ocena za σ^2.

V nadaljevanju si poglejmo še testno statistiko za ocenjen parameter $\hat{a}$. Na osnovi izraza (9.28) lahko zapišemo:

$$\hat{A}=\overline{y}-\hat{B}\cdot\overline{x} \tag{9.94}$$

Če upoštevamo izraz (9.83), dobimo:

$$\hat{A}=\overline{y}-\left(\sum_{i=1}^{n}\left(\frac{\left[x_i-\overline{x}\right]}{S_{xx}}\right)\cdot Y\middle|x_i\right)\cdot\overline{x} \tag{9.95}$$

Pri nekem izbranem vzorcu sledi:

$$\hat{a}=\overline{y}-\left(\sum_{i=1}^{n}\left(\frac{\left[x_i-\overline{x}\right]}{S_{xx}}\right)\cdot y_i\right)\cdot\overline{x} \tag{9.96}$$

Dobimo:

$$\begin{aligned}
&\hat{a} = \bar{y} - \left(\sum_{i=1}^{n} \left(\frac{[x_i - \bar{x}]}{S_{xx}} \right) \cdot y_i \right) \cdot \bar{x} = \\
&= \sum_{i=1}^{n} \frac{y_i}{n} - \sum_{i=1}^{n} \left(\frac{[x_i - \bar{x}]}{S_{xx}} \right) \cdot \bar{x} \cdot y_i = \\
&= \sum_{i=1}^{n} \left(\frac{y_i}{n} - \frac{[x_i - \bar{x}]}{S_{xx}} \cdot \bar{x} \cdot y_i \right) = \sum_{i=1}^{n} \left(\frac{y_i \cdot S_{xx} - n \cdot [x_i - \bar{x}] \cdot \bar{x} \cdot y_i}{n \cdot S_{xx}} \right) = \\
&= \sum_{i=1}^{n} \left(\frac{y_i \cdot S_{xx} - n \cdot x_i \cdot \bar{x} \cdot y_i + n \cdot \bar{x}^2 \cdot y_i}{n \cdot S_{xx}} \right) = \\
&= \sum_{i=1}^{n} \left(\frac{S_{xx} - n \cdot x_i \cdot \bar{x} + n \cdot \bar{x}^2}{n \cdot S_{xx}} \right) \cdot y_i
\end{aligned} \tag{9.97}$$

Pripadajoča statistika bo:

$$\hat{A} = \sum_{i=1}^{n} \left(\frac{S_{xx} - n \cdot x_i \cdot \bar{x} + n \cdot \bar{x}^2}{n \cdot S_{xx}} \right) \cdot Y \big| x_i \tag{9.98}$$

kar je normalna naključna spremenljivka, saj predstavlja linearno kombinacijo normalnih naključnih spremenljivk $Y \big| x_i$. Njeno matematično upanje je:

$$E\left(\hat{A}\right) = E\left(\bar{y} - \hat{B} \cdot \bar{x}\right) = \bar{y} - E\left(\hat{B}\right) \cdot \bar{x} = \bar{y} - b \cdot \bar{x} = a \tag{9.99}$$

Varianca pa bo enaka:

$$\begin{aligned}
&VAR\left(\hat{A}\right) = VAR\left(\bar{y} - \hat{B} \cdot \bar{x}\right) = VAR\left(\bar{y}\right) + VAR\left(\hat{B} \cdot \bar{x}\right) = \\
&= VAR\left(\bar{y}\right) + \bar{x}^2 \cdot VAR\left(\hat{B}\right) = VAR\left(\bar{y}\right) + \bar{x}^2 \cdot \frac{\sigma^2}{S_{xx}} = \\
&= VAR\left(\frac{1}{n} \sum_{i=1}^{n} y_i \right) + \bar{x}^2 \cdot \frac{\sigma^2}{S_{xx}} = \frac{1}{n^2} \cdot VAR\left(\sum_{i=1}^{n} y_i \right) + \bar{x}^2 \cdot \frac{\sigma^2}{S_{xx}} = \\
&= \frac{1}{n^2} \cdot \sum_{i=1}^{n} VAR\left(y_i\right) + \bar{x}^2 \cdot \frac{\sigma^2}{S_{xx}} = \frac{1}{n^2} \cdot \left(\sigma^2 + \sigma^2 + \ldots + \sigma^2 \right) + \bar{x}^2 \cdot \frac{\sigma^2}{S_{xx}} = \\
&= \frac{n \cdot \sigma^2}{n^2} \cdot + \bar{x}^2 \cdot \frac{\sigma^2}{S_{xx}} = \frac{\sigma^2}{n} \cdot + \bar{x}^2 \cdot \frac{\sigma^2}{S_{xx}} = \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right) \cdot \sigma^2 = \left(\frac{S_{xx} + n \cdot \bar{x}^2}{n \cdot S_{xx}} \right) \cdot \sigma^2
\end{aligned} \tag{9.100}$$

Za normalno naključno spremenljivko $\hat{A}$ torej velja:

$$\hat{A} \in N\left(E\left(\hat{A}\right), STD\left(\hat{A}\right)\right) = N\left(a, \sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right) \cdot \sigma^2}\right) \tag{9.101}$$

Standardizirana normalna naključna spremenljivka zanjo je:

$$Z = \frac{\hat{a} - a}{\sigma_{\hat{a}}} = \frac{\hat{a} - a}{\sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right) \cdot \sigma^2}} = \frac{\hat{a} - a}{\sqrt{\left(\frac{S_{xx} + n \cdot \bar{x}^2}{n \cdot S_{xx}}\right) \cdot \sigma^2}} = \frac{\hat{a} - a}{\sigma} \cdot \sqrt{\frac{n \cdot S_{xx}}{S_{xx} + n \cdot \bar{x}^2}} \in N(0,1) \tag{9.102}$$

Ker je σ neznana standardna deviacija, moramo uvesti t statistiko na naslednji način [Jesenko]:

$$\begin{aligned} t &= \frac{\hat{a} - a}{\sqrt{\left(\frac{S_{xx} + n \cdot \bar{x}^2}{n \cdot S_{xx}}\right) \cdot s_\varepsilon{}^2}} = \frac{\hat{a} - a}{s_\varepsilon} \cdot \sqrt{\frac{n \cdot S_{xx}}{S_{xx} + n \cdot \bar{x}^2}} = \\ &= \frac{\hat{a} - a}{\sqrt{\frac{1}{n-2} \cdot \left(S_{yy} - \hat{b} \cdot S_{xy}\right)}} \cdot \sqrt{\frac{n \cdot S_{xx}}{S_{xx} + n \cdot \bar{x}^2}} \in t(n-2) \end{aligned} \tag{9.103}$$

Na njej temelji test hipoteze o regresijskem parametru *a*.

Primer 9.6.:

Tabela na sliki 217 podaja vzorec osmih ocenjenih vrednosti stanovanj, kot so jih ocenili cenilci, ter dejansko doseženo prodajno ceno. Vrednosti so podane v milijonih SIT [Jesenko]. Kot se izkaže, sta ocenjena parametra regresijske premice enaka: $\hat{b} = 0.8649, \hat{a} = 20.7115$ *(glej izraza (9.32) in (9.33)). Preizkusite ničelno hipotezo* $b = 1$ *pri nasprotni hipotezi* $b < 1$*, če vzamemo* $\alpha = 0.01$ *[Jesenko]. Preizkusite tudi ničelno hipotezo* $a = 20$ *pri nasprotni hipotezi* $a > 20$*, če vzamemo* $\alpha = 0.01$*.*

Ocenjena vrednost (x_i)	Prodajna cena (y_i)
125	132
83	88
182	177
135	138
147	146
112	121
211	203
76	87

Slika 217: Vzorec osmih ocenjenih vrednosti stanovanj, kot so jih ocenili cenilci, ter dejanska dosežena prodajna cena [Jesenko]

Imamo:

$$\begin{aligned}
&H_0 : b = 1 \\
&H_1 : b < 1 \\
&H_0 : a = 20 \\
&H_1 : a > 20 \\
&\alpha = 0.01 \\
&S_{xx} = 14933 \quad (glej\ \ izraz\ (9.31)) \\
&S_{xy} = 12916 \quad (glej\ \ izraz\ (9.31)) \\
&S_{yy} = 11218 \quad (se\ izkaže\ [Jesenko]) \\
&\hat{b} = 0.8649 \\
&\hat{a} = 20.7115
\end{aligned} \tag{9.103}$$

Najprej izračunajmo:

$$\begin{aligned}
s_\varepsilon &= \sqrt{\frac{1}{n-2}\cdot\left(S_{yy} - \hat{b}\cdot S_{xy}\right)} = \sqrt{\frac{1}{8-2}\cdot\left(11218 - 0.8649\cdot 12916\right)} = \\
&= \sqrt{\frac{1}{6}\cdot\left(11218 - 0.8649\cdot 12916\right)} = 2.8088
\end{aligned} \tag{9.104}$$

Nato izračunajmo:

$$t = \frac{\hat{b} - b}{\frac{s_\varepsilon}{\sqrt{S_{xx}}}} = \frac{0.8649 - 1}{2.8088}\sqrt{14933} = -5.8776 \tag{9.105}$$

Kritična vrednost t statistike je:

$$t_{krit} = t(\alpha, n-2) = t(0.01, 6) = -3.1427 \tag{9.106}$$

Ker je $t < t_{krit}$, smo padli v kritično območje, zato moramo zavrniti ničelno hipotezo, da je $b = 1$. Lahko pa sprejmemo nasprotno hipotezo, da je $b < 1$.

Izračunajmo še:

$$t = \frac{\hat{a} - a}{s_\varepsilon} \cdot \sqrt{\frac{n \cdot S_{xx}}{S_{xx} + n \cdot \bar{x}^2}} = \frac{20.711 - 20}{2.8088} \cdot \sqrt{\frac{8 \cdot 14933}{14933 + 8 \cdot \left(\frac{1071}{8}\right)^2}} =$$
$$= \frac{0.711}{2.8088} \cdot \sqrt{\frac{8 \cdot 14933}{14933 + \left(\frac{1071^2}{8}\right)}} = 0.2199 \tag{9.107}$$

Kritična vrednost t statistike je:

$$t_{krit} = t(1-\alpha, n-2) = t(0.99, 6) = 3.1427 \tag{9.108}$$

Ker je $t < t_{krit}$, smo padli v območje zaupanja, zato lahko sprejmemo ničelno hipotezo, to je, da velja: $a = 20$.

Pri izračunih smo si pomagali z naslednjim programom v Matlabu:

```
% batest.m

clear
clc
close all
```

```
x = [125 83 182 135 147 112 211 76]
y = [132 88 177 138 146 121 203 87]

n = length(x)

disp('vsota x je:')
sx = sum(x)

disp('vsota y je:')
sy = sum(y)

disp('vsota x^2 je:')
sx2 = x*x'

disp('vsota y^2 je:')
sy2 = y*y'

disp('vsota x.y je:')
sxy = x*y'

disp('Sxy je:')
Sxy = sxy - sx*sy/n

disp('Sxx je:')
Sxx = sx2 - sx^2/n

disp('Syy je:')
Syy = sy2 - sy^2/n

disp('boc=')
boc = Sxy/Sxx

disp('aoc=')
aoc = sy/n - boc*sx/n

% test za parameter boc:

alfa = 0.01
b = 1
db = boc - b

se=sqrt((Syy-boc*Sxy)/(n-2)) % standardna ocena napake modela

% vrednost testne t statistike:

t=db*sqrt(Sxx)/se

% kriticna vrednost t statistike:

tkrit = -tinv(1-alfa,n-2)
```

```
if t<tkrit
    disp('zavrni nicelno hipotezo')
else
    disp('sprejmi nicelno hipotezo')
end

% test za parameter aoc:

a = 20
da = aoc - a

% vrednost testne t statistike:

t=da*sqrt(n*Sxx/(Sxx+n*(sx/n)^2))/se

tkrit = tinv(1-alfa,n-2)

if t>tkrit
    disp('zavrni nicelno hipotezo')
else
    disp('sprejmi nicelno hipotezo')
end
```

Izpis komandnega okna je naslednji:

```
x =
  125   83  182  135  147  112  211   76
y =
  132   88  177  138  146  121  203   87
n =
   8

vsota x je:
sx =
     1071
vsota y je:
sy =
     1092
vsota x^2 je:
sx2 =
    158313
vsota y^2 je:
sy2 =
    160276
vsota x.y je:
sxy =
    159107
```

```
Sxy je:
Sxy =
  1.2916e+004
Sxx je:
Sxx =
  1.4933e+004
Syy je:
Syy =
      11218

boc=
boc =
    0.8649
aoc=
aoc =
   20.7110

alfa =
    0.0100
b =
     1
db =
   -0.1351

se =
    2.8088

t =
   -5.8776
tkrit =
   -3.1427
zavrni nicelno hipotezo

a =
    20
da =
    0.7110

t =
    0.2199
tkrit =
    3.1427
sprejmi nicelno hipotezo
```

Intervalne ocene – intervali zaupanja za parametra *a* in *b*

Interval zaupanja za regresijski parameter *b* dobimo na naslednji način:

$$
\begin{aligned}
&P\left(-t_{\frac{\alpha}{2},n-2} \le t \le t_{\frac{\alpha}{2},n-2}\right)=1-\alpha \\
&P\left(-t_{\frac{\alpha}{2},n-2} \le \frac{\hat{b}-b}{\frac{s_\varepsilon}{\sqrt{S_{xx}}}} \le t_{\frac{\alpha}{2},n-2}\right)=1-\alpha \\
&P\left(-t_{\frac{\alpha}{2},n-2} \cdot \frac{s_\varepsilon}{\sqrt{S_{xx}}} \le \hat{b}-b \le t_{\frac{\alpha}{2},n-2} \cdot \frac{s_\varepsilon}{\sqrt{S_{xx}}}\right)=1-\alpha \\
&P\left(\hat{b}-t_{\frac{\alpha}{2},n-2} \cdot \frac{s_\varepsilon}{\sqrt{S_{xx}}} \le b \le \hat{b}+t_{\frac{\alpha}{2},n-2} \cdot \frac{s_\varepsilon}{\sqrt{S_{xx}}}\right)=1-\alpha \\
&b \in \left(\hat{b}-t_{\frac{\alpha}{2},n-2} \cdot \frac{s_\varepsilon}{\sqrt{S_{xx}}}, \hat{b}+t_{\frac{\alpha}{2},n-2} \cdot \frac{s_\varepsilon}{\sqrt{S_{xx}}}\right)
\end{aligned}
\qquad (9.109)
$$

Interval zaupanja za regresijski parameter *a* dobimo na naslednji način:

$$
\begin{aligned}
&P\left(-t_{\frac{\alpha}{2},n-2} \le t \le t_{\frac{\alpha}{2},n-2}\right)=1-\alpha \\
&P\left(-t_{\frac{\alpha}{2},n-2} \le \frac{\hat{a}-a}{\sqrt{\left(\frac{S_{xx}+n\cdot\overline{x}^2}{n\cdot S_{xx}}\right)\cdot {s_\varepsilon}^2}} \le t_{\frac{\alpha}{2},n-2}\right)=1-\alpha \\
&P\left(-t_{\frac{\alpha}{2},n-2}\cdot\sqrt{\left(\frac{S_{xx}+n\cdot\overline{x}^2}{n\cdot S_{xx}}\right)\cdot {s_\varepsilon}^2} \le \hat{a}-a \le t_{\frac{\alpha}{2},n-2}\cdot\sqrt{\left(\frac{S_{xx}+n\cdot\overline{x}^2}{n\cdot S_{xx}}\right)\cdot {s_\varepsilon}^2}\right)=1-\alpha \\
&P\left(-t_{\frac{\alpha}{2},n-2}\cdot\sqrt{\left(\frac{1}{n}+\frac{\overline{x}^2}{S_{xx}}\right)}\cdot s_\varepsilon \le \hat{a}-a \le t_{\frac{\alpha}{2},n-2}\cdot\sqrt{\left(\frac{1}{n}+\frac{\overline{x}^2}{S_{xx}}\right)}\cdot s_\varepsilon\right)=1-\alpha \\
&a\in\left(\hat{a}-t_{\frac{\alpha}{2},n-2}\cdot\sqrt{\left(\frac{1}{n}+\frac{\overline{x}^2}{S_{xx}}\right)}\cdot s_\varepsilon,\hat{a}+t_{\frac{\alpha}{2},n-2}\cdot\sqrt{\left(\frac{1}{n}+\frac{\overline{x}^2}{S_{xx}}\right)}\cdot s_\varepsilon\right)
\end{aligned}
\tag{9.110}
$$

Vzorčna porazdelitev za oceno matematičnega upanja odvisne spremenljivke

Imamo enačbo regresijske premice:

$$\hat{\hat{y}}=\hat{a}+\hat{b}\cdot x \tag{9.111}$$

Spremenljivka $\hat{\hat{y}}$ je normalno porazdeljena, saj je linearna kombinacija parametrov $\hat{a},\hat{b}$, ki sta normalno porazdeljena. Ugotovimo zanjo matematično upanje in varianco.

Matematično upanje je enako:

$$E\left(\hat{\hat{y}}\right)=E\left(\hat{a}+\hat{b}\cdot x\right)=E\left(\hat{a}\right)+E\left(\hat{b}\right)\cdot x=a+b\cdot x \tag{9.112}$$

Varianca je enaka:

$$VAR\left(\hat{\bar{y}}\right)=VAR\left(\underbrace{\bar{y}-\hat{b}\cdot\bar{x}}_{\hat{a}}+\hat{b}\cdot x\right)=VAR\left(\bar{y}+\hat{b}\cdot\left(x-\bar{x}\right)\right)=VAR\left(\bar{y}\right)+VAR\left(\hat{b}\cdot\left(x-\bar{x}\right)\right) \tag{9.113}$$

$$=VAR\left(\bar{y}\right)+VAR\left(\hat{b}\right)\cdot\left(x-\bar{x}\right)^2=VAR\left(\frac{1}{n}\sum_{i=1}^{n}y_i\right)+\frac{\sigma^2}{S_{xx}}\cdot\left(x-\bar{x}\right)^2=$$

$$=\frac{1}{n^2}\cdot\sum_{i=1}^{n}VAR\left(y_i\right)+\frac{\sigma^2}{S_{xx}}\cdot\left(x-\bar{x}\right)^2=\frac{1}{n^2}\cdot\left(\sigma^2+\sigma^2+\ldots+\sigma^2\right)+\frac{\sigma^2}{S_{xx}}\cdot\left(x-\bar{x}\right)^2=$$

$$=\frac{n\cdot\sigma^2}{n^2}+\frac{\sigma^2}{S_{xx}}\cdot\left(x-\bar{x}\right)^2=\sigma^2\left(\frac{1}{n}+\frac{\left(x-\bar{x}\right)^2}{S_{xx}}\right)$$

Za normalno naključno spremenljivko $\hat{\bar{y}}$ torej velja:

$$\hat{\bar{Y}}\in N\left(E\left(\hat{\bar{Y}}\right),STD\left(\hat{\bar{Y}}\right)\right)=N\left(a+b\cdot x,\sqrt{\sigma^2\left(\frac{1}{n}+\frac{\left(x-\bar{x}\right)^2}{S_{xx}}\right)}\right) \tag{9.114}$$

Standardizirana normalna naključna spremenljivka zanjo je:

$$Z=\frac{\hat{\bar{Y}}-\bar{Y}}{\sigma_{\hat{\bar{Y}}}}=\frac{\hat{\bar{Y}}-\bar{Y}}{\sqrt{\sigma^2\left(\frac{1}{n}+\frac{\left(x-\bar{x}\right)^2}{S_{xx}}\right)}}\in N\left(0,1\right) \tag{9.115}$$

Ker je σ neznana standardna deviacija, moramo uvesti t statistiko na naslednji način [Kutner]:

$$t=\frac{\hat{\bar{Y}}-\bar{Y}}{\sqrt{s_e{}^2\left(\frac{1}{n}+\frac{\left(x-\bar{x}\right)^2}{S_{xx}}\right)}}\in t\left(n-2\right) \tag{9.116}$$

Intervalna ocena – interval zaupanja za oceno matematičnega upanja odvisne spremenljivke

Interval zaupanja za normalno naključno spremenljivko $\hat{\bar{y}}$ dobimo na naslednji način:

$$
\begin{aligned}
&P\left(-t_{\frac{\alpha}{2},n-2} \le t \le t_{\frac{\alpha}{2},n-2}\right)=1-\alpha \\
&P\left(-t_{\frac{\alpha}{2},n-2} \le \frac{\hat{\bar{Y}}-\bar{Y}}{\sqrt{s_e{}^2\left(\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}\right)}} \le t_{\frac{\alpha}{2},n-2}\right)=1-\alpha \\
&P\left(-t_{\frac{\alpha}{2},n-2}\cdot\sqrt{s_e{}^2\left(\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}\right)} \le \hat{\bar{Y}}-\bar{Y} \le t_{\frac{\alpha}{2},n-2}\cdot\sqrt{s_e{}^2\left(\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}\right)}\right)=1-\alpha \\
&\bar{Y}\in\left(\hat{\bar{Y}}-t_{\frac{\alpha}{2},n-2}\cdot s_e\cdot\sqrt{\left(\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}\right)},\hat{\bar{Y}}+t_{\frac{\alpha}{2},n-2}\cdot s_e\cdot\sqrt{\left(\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}\right)}\right)
\end{aligned}
\tag{9.117}
$$

Pri dani vrednosti neodvisne spremenljivke $X=x_i$ torej z verjetnostjo $(1-\alpha)$ pričakujemo, da bo matematično upanje odvisne spremenljivke padlo v naslednji interval:

$$
\bar{Y}\Big|x_i\in\left(\hat{\bar{Y}}\Big|x_i-t_{\frac{\alpha}{2},n-2}\cdot s_e\cdot\sqrt{\left(\frac{1}{n}+\frac{(x_i-\bar{x})^2}{S_{xx}}\right)},\hat{\bar{Y}}\Big|x_i+t_{\frac{\alpha}{2},n-2}\cdot s_e\cdot\sqrt{\left(\frac{1}{n}+\frac{(x_i-\bar{x})^2}{S_{xx}}\right)}\right)
$$

oz.

$$
\bar{Y}\Big|x_i\in\left(\begin{array}{l}\left(\hat{a}+\hat{b}\cdot x_i\right)-t_{\frac{\alpha}{2},n-2}\cdot s_e\cdot\sqrt{\left(\frac{1}{n}+\frac{(x_i-\bar{x})^2}{S_{xx}}\right)}, \\ \left(\hat{a}+\hat{b}\cdot x_i\right)+t_{\frac{\alpha}{2},n-2}\cdot s_e\cdot\sqrt{\left(\frac{1}{n}+\frac{(x_i-\bar{x})^2}{S_{xx}}\right)}\end{array}\right),\quad i=1,2,...,n
\tag{9.118}
$$

Predikcijski interval zaupanja

Imamo:

$$\hat{\bar{y}}(i) = \hat{a} + \hat{b} \cdot x(i),\ i = 1,...,n \tag{9.119}$$

pri čemer smo ocenili parametra na intervalu $i = 1,...,n$. Nato pa nas zanima enokoračna napoved za novo vrednost odvisne spremenljivke y_{nov}, pri čemer se predpostavi, da je nova vrednost neodvisne spremenljivke x_{nov} znana vnaprej (npr. naročilo v proizvodnji za izdelavo določenega števila kosov za naslednji mesec je znano vnaprej, vodstvo pa zanima, koliko delovnih ur bo potrebnih za realizacijo naročila). Zanima nas, kakšen bi bil ocenjeni predikcijski interval zaupanja za pričakovano vrednost odvisne spremenljivke y_{nov}.

Pri dani vrednosti neodvisne spremenljivke x_{nov} je vrednost ocenjene regresijske premice naslednja:

$$\hat{\bar{y}}_{nov} = \hat{a} + \hat{b} \cdot x_{nov} \tag{9.120}$$

Poleg tega velja:

$$y_{nov} = a + b \cdot x_{nov} + \varepsilon_{nov} \tag{9.121}$$

Seveda je Y_{nov} normalna naključna spremenljivka s porazdelitvenim zakonom $N(a + b \cdot x_{nov}, \sigma)$, pri čemer je $\varepsilon_{nov} \in N(0, \sigma)$. Tvorimo naslednjo razliko:

$$y_{nov} - \hat{\bar{y}}_{nov} = (a + b \cdot x_{nov} + \varepsilon_{nov}) - (\hat{a} + \hat{b} \cdot x_{nov}) \tag{9.122}$$

Ta razlika je linearna kombinacija normalnih naključnih spremenljivk, zato ima normalno porazdelitev. Njeno matematično upanje je:

$$E\left(y_{nov}-\hat{\bar{y}}_{nov}\right)=E\left(\left(a+b\cdot x_{nov}+\varepsilon_{nov}\right)-\left(\hat{a}+\hat{b}\cdot x_{nov}\right)\right)=0 \tag{9.123}$$

Varianca je enaka:

$$\begin{aligned}&VAR\left(y_{nov}-\hat{\bar{y}}_{nov}\right)=VAR\left(a+b\cdot x_{nov}+\varepsilon_{nov}\right)+VAR\left(\hat{\bar{y}}_{nov}\right)=\\&=VAR\left(\varepsilon_{nov}\right)+VAR\left(\hat{\bar{y}}_{nov}\right)=\sigma^2+\sigma^2\left(\frac{1}{n}+\frac{\left(x_{nov}-\bar{x}\right)^2}{S_{xx}}\right)=\\&=\sigma^2\left(1+\frac{1}{n}+\frac{\left(x_{nov}-\bar{x}\right)^2}{S_{xx}}\right)\end{aligned} \tag{9.124}$$

pri čemer smo upoštevali izraz (9.113), kjer smo *x* nadomestili z x_{nov}. Za normalno naključno spremenljivko $y_{nov}-\hat{\bar{y}}_{nov}$ torej velja:

$$y_{nov}-\hat{\bar{y}}_{nov}\in N\left(E\left(y_{nov}-\hat{\bar{y}}_{nov}\right),STD\left(y_{nov}-\hat{\bar{y}}_{nov}\right)\right)=N\left(0,\sqrt{\sigma^2\left(1+\frac{1}{n}+\frac{\left(x_{nov}-\bar{x}\right)^2}{S_{xx}}\right)}\right) \tag{9.125}$$

Standardizirana normalna naključna spremenljivka zanjo je:

$$Z=\frac{y_{nov}-\hat{\bar{y}}_{nov}}{\sigma_{y_{nov}-\hat{\bar{y}}_{nov}}}=\frac{y_{nov}-\hat{\bar{y}}_{nov}}{\sqrt{\sigma^2\left(1+\frac{1}{n}+\frac{\left(x_{nov}-\bar{x}\right)^2}{S_{xx}}\right)}}\in N(0,1) \tag{9.126}$$

Ker je σ neznana standardna deviacija, moramo uvesti t statistiko na naslednji način:

$$t=\frac{y_{nov}-\hat{\bar{y}}_{nov}}{\sqrt{s_e{}^2\left(1+\frac{1}{n}+\frac{\left(x_{nov}-\bar{x}\right)^2}{S_{xx}}\right)}}\in t(n-2) \tag{9.127}$$

Interval zaupanja za normalno naključno spremenljivko $y_{nov} - \hat{\bar{y}}_{nov}$ dobimo na naslednji način:

$$P\left(-t_{\frac{\alpha}{2},n-2} \le t \le t_{\frac{\alpha}{2},n-2}\right) = 1-\alpha$$

$$P\left(-t_{\frac{\alpha}{2},n-2} \le \frac{y_{nov} - \hat{\bar{y}}_{nov}}{\sqrt{s_e{}^2\left(1+\frac{1}{n}+\frac{\left(x_{nov}-\bar{x}\right)^2}{S_{xx}}\right)}} \le t_{\frac{\alpha}{2},n-2}\right) = 1-\alpha$$

$$P\left(-t_{\frac{\alpha}{2},n-2} \cdot \sqrt{s_e{}^2\left(1+\frac{1}{n}+\frac{\left(x_{nov}-\bar{x}\right)^2}{S_{xx}}\right)} \le y_{nov} - \hat{\bar{y}}_{nov} \le t_{\frac{\alpha}{2},n-2} \cdot \sqrt{s_e{}^2\left(1+\frac{1}{n}+\frac{\left(x_{nov}-\bar{x}\right)^2}{S_{xx}}\right)}\right) = 1- \quad (9.128)$$

$$y_{nov} \in \left(\hat{\bar{y}}_{nov} - t_{\frac{\alpha}{2},n-2} \cdot s_e \cdot \sqrt{\left(1+\frac{1}{n}+\frac{\left(x_{nov}-\bar{x}\right)^2}{S_{xx}}\right)}, \hat{\bar{y}}_{nov} + t_{\frac{\alpha}{2},n-2} \cdot s_e \cdot \sqrt{\left(1+\frac{1}{n}+\frac{\left(x_{nov}-\bar{x}\right)^2}{S_{xx}}\right)}\right)$$

Predikcijski interval zaupanja pri večkratnem opazovanju odvisne spremenljivke

Denimo nas zanima enokoračna napoved za novo srednjo vrednost odvisne spremenljivke $\bar{y}_{nov}$, pri čemer je neodvisna spremenljivke x_{nov} znana vnaprej, ter bomo izvedli pri njeni enaki vrednosti *m* novih opazovanj neodvisne spremenljivke (npr. naročilo v proizvodnji za izdelavo določenega (enakega) števila kosov za naslednje *m = 3* mesece je znano vnaprej, vodstvo pa zanima, koliko delovnih ur bo v povprečju potrebnih za realizacijo tega naročila v naslednjih treh mesecih).

Pri dani (enaki) vrednosti neodvisne spremenljivke x_{nov} je srednja vrednost ocenjene regresijske premice naslednja:

$$\hat{\bar{y}}_{nov} = \hat{a} + \hat{b} \cdot x_{nov} \tag{9.129}$$

Poleg tega velja za srednjo vrednost naslednjih *m* opazovanj odvisne spremenljivke:

$$\begin{aligned}
&\bar{y}_{nov} = \frac{1}{m} \cdot \sum_{i=1}^{m} \left(y_{nov}(i) \right) = \frac{1}{m} \cdot \sum_{i=1}^{m} \left(a + b \cdot x_{nov} + \varepsilon_{nov}(i) \right) = \\
&= a + b \cdot x_{nov} + \frac{1}{m} \cdot \sum_{i=1}^{m} \left(\varepsilon_{nov}(i) \right) \\
&E\left(\bar{y}_{nov} \right) = a + b \cdot x_{nov} + E\left(\frac{1}{m} \cdot \sum_{i=1}^{m} \left(\varepsilon_{nov}(i) \right) \right) = \\
&= a + b \cdot x_{nov} + \frac{1}{m} \cdot \sum_{i=1}^{m} \underbrace{E\left(\varepsilon_{nov}(i) \right)}_{0} = a + b \cdot x_{nov} \\
&VAR\left(\bar{y}_{nov} \right) = VAR\left(a + b \cdot x_{nov} + \frac{1}{m} \cdot \sum_{i=1}^{m} \left(\varepsilon_{nov}(i) \right) \right) = \\
&= \frac{1}{m^2} \cdot \sum_{i=1}^{m} VAR\left(\varepsilon_{nov}(i) \right) = \frac{m \cdot \sigma^2}{m^2} = \frac{\sigma^2}{m}
\end{aligned} \tag{9.130}$$

Seveda je $\bar{y}_{nov}$ normalna naključna spremenljivka s porazdelitvenim zakonom $N\left(a + b \cdot x_{nov}, \sqrt{\frac{\sigma^2}{m}} \right)$, pri čemer je $\varepsilon_{nov}(i) \in N(0, \sigma), i = 1, ..., m$. Tvorimo naslednjo razliko:

$$\bar{y}_{nov} - \hat{\bar{y}}_{nov} = a + b \cdot x_{nov} + \frac{1}{m} \cdot \sum_{i=1}^{m} \left(\varepsilon_{nov}(i) \right) - \left(\hat{a} + \hat{b} \cdot x_{nov} \right) \tag{9.131}$$

Ta razlika je linearna kombinacija normalnih naključnih spremenljivk, zato ima normalno porazdelitev. Njeno matematično upanje je:

(9.132)

$$E\left(\bar{y}_{nov}-\hat{\bar{y}}_{nov}\right)=E\left(a+b\cdot x_{nov}+\frac{1}{m}\cdot\sum_{i=1}^{m}\left(\varepsilon_{nov}(i)\right)-\left(\hat{a}+\hat{b}\cdot x_{nov}\right)\right)=$$
$$=E\left(\frac{1}{m}\cdot\sum_{i=1}^{m}\left(\varepsilon_{nov}(i)\right)\right)=0$$

Varianca je enaka:

$$VAR\left(\bar{y}_{nov}-\hat{\bar{y}}_{nov}\right)=VAR\left(\bar{y}_{nov}\right)+VAR\left(\hat{\bar{y}}_{nov}\right)=$$
$$=\frac{\sigma^2}{m}+\sigma^2\left(\frac{1}{n}+\frac{\left(x_{nov}-\bar{x}\right)^2}{S_{xx}}\right)= \qquad (9.133)$$
$$=\sigma^2\left(\frac{1}{m}+\frac{1}{n}+\frac{\left(x_{nov}-\bar{x}\right)^2}{S_{xx}}\right)$$

pri čemer smo upoštevali izraz (9.113), kjer smo x nadomestili z x_{nov}. Za normalno naključno spremenljivko $\bar{y}_{nov}-\hat{\bar{y}}_{nov}$ torej velja:

$$\bar{y}_{nov}-\hat{\bar{y}}_{nov}\in N\left(E\left(\bar{y}_{nov}-\hat{\bar{y}}_{nov}\right),STD\left(\bar{y}_{nov}-\hat{\bar{y}}_{nov}\right)\right)=N\left(0,\sqrt{\sigma^2\left(\frac{1}{m}+\frac{1}{n}+\frac{\left(x_{nov}-\bar{x}\right)^2}{S_{xx}}\right)}\right) \qquad (9.134)$$

Standardizirana normalna naključna spremenljivka zanjo je:

$$Z=\frac{\bar{y}_{nov}-\hat{\bar{y}}_{nov}}{\sigma_{\bar{y}_{nov}-\hat{\bar{y}}_{nov}}}=\frac{\bar{y}_{nov}-\hat{\bar{y}}_{nov}}{\sqrt{\sigma^2\left(\frac{1}{m}+\frac{1}{n}+\frac{\left(x_{nov}-\bar{x}\right)^2}{S_{xx}}\right)}}\in N(0,1) \qquad (9.135)$$

Ker je σ neznana standardna deviacija, moramo uvesti t statistiko na naslednji način:

$$t = \frac{\bar{y}_{nov} - \hat{\bar{y}}_{nov}}{\sqrt{s_e^{\,2}\left(\frac{1}{m} + \frac{1}{n} + \frac{\left(x_{nov} - \bar{x}\right)^2}{S_{xx}}\right)}} \in t\left(n-2\right) \tag{9.136}$$

Interval zaupanja za normalno naključno spremenljivko $\bar{y}_{nov} - \hat{\bar{y}}_{nov}$ dobimo na naslednji način:

$$\begin{gathered}
P\left(-t_{\frac{\alpha}{2},n-2} \le t \le t_{\frac{\alpha}{2},n-2}\right) = 1-\alpha \\
P\left(-t_{\frac{\alpha}{2},n-2} \le \frac{\bar{y}_{nov} - \hat{\bar{y}}_{nov}}{\sqrt{s_e^{\,2}\left(\frac{1}{m} + \frac{1}{n} + \frac{\left(x_{nov} - \bar{x}\right)^2}{S_{xx}}\right)}} \le t_{\frac{\alpha}{2},n-2}\right) = 1-\alpha \\
P\left(-t_{\frac{\alpha}{2},n-2} \cdot \sqrt{s_e^{\,2}\left(\frac{1}{m} + \frac{1}{n} + \frac{\left(x_{nov} - \bar{x}\right)^2}{S_{xx}}\right)} \le \bar{y}_{nov} - \hat{\bar{y}}_{nov} \le t_{\frac{\alpha}{2},n-2} \cdot \sqrt{s_e^{\,2}\left(\frac{1}{m} + \frac{1}{n} + \frac{\left(x_{nov} - \bar{x}\right)^2}{S_{xx}}\right)}\right) = 1- \\
\bar{y}_{nov} \in \left(\hat{\bar{y}}_{nov} - t_{\frac{\alpha}{2},n-2} \cdot s_e \cdot \sqrt{\left(\frac{1}{m} + \frac{1}{n} + \frac{\left(x_{nov} - \bar{x}\right)^2}{S_{xx}}\right)}, \hat{\bar{y}}_{nov} + t_{\frac{\alpha}{2},n-2} \cdot s_e \cdot \sqrt{\left(\frac{1}{m} + \frac{1}{n} + \frac{\left(x_{nov} - \bar{x}\right)^2}{S_{xx}}\right)}\right)
\end{gathered} \tag{9.137}$$

Primer 9.7.:

Tabela na sliki 217 podaja vzorec osmih ocenjenih vrednosti stanovanj, kot so jih ocenili cenilci, ter dejansko doseženo prodajno ceno. Vrednosti so podane v milijonih SIT [Jesenko]. Kot se izkaže, sta ocenjena parametra regresijske premice enaka:

$\hat{b} = 0.8649, \hat{a} = 20.7115$ *(glej izraza (9.32) in (9.33)). Izračunajte interval zaupanja za oba ocenjena parametra. Izračunajte tudi interval zaupanja za prodajno ceno stanovanja, če je njegova ocenjena vrednost enaka* $x_{nov} = 175$ *milijonov SIT (ta vrednost ni bila uporabljena pri ocenjevanju parametrov). Narišite spodnjo in zgornjo mejo zaupanja vrednosti ocenjene regresijske premice pri* $x_i, i = 1,...,8$, *ki so bile uporabljene pri ocenjevanju obeh parametrov.(*$\alpha = 0.05$*). Kaj bi se zgodilo, če bi uvedli razširjen interval zaupanja, ki bi upošteval tudi vrednost* $x_{nov} = 175$ *milijonov SIT ?*

Imamo:

$$\begin{aligned}
&\alpha = 0.05 \\
&S_{xx} = 14933 \quad (glej\ \ izraz\ (9.31)) \\
&S_{xy} = 12916 \quad (glej\ \ izraz\ (9.31)) \\
&S_{yy} = 11218 \quad (se\ izkaže\ [Jesenko]) \\
&\hat{b} = 0.8649 \\
&\hat{a} = 20.7115 \\
&s_e = 2.8088 \quad (glej\ \ izraz\ (9.104)) \\
&\overline{x} = \frac{1071}{8} = 133.875 \quad (glej\ \ izraz\ (9.30)) \\
&n = 8
\end{aligned} \tag{9.138}$$

Kritična vrednost t statistike je:

$$t_{krit} = t\left(\frac{\alpha}{2}, n-2\right) = t\left(0.025, 6\right) = 2.4469 \tag{9.139}$$

Interval zaupanja za regresijski parameter *b* je:

(9.140)

$$b \in \left(\hat{b} - t_{\frac{\alpha}{2},n-2} \cdot \frac{s_\varepsilon}{\sqrt{S_{xx}}}, \hat{b} + t_{\frac{\alpha}{2},n-2} \cdot \frac{s_\varepsilon}{\sqrt{S_{xx}}} \right)$$
$$b \in (0.8087, 0.9211)$$

Interval zaupanja za regresijski parameter *a* je:

$$a \in \left(\hat{a} - t_{\frac{\alpha}{2},n-2} \cdot \sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)} \cdot s_\varepsilon, \hat{a} + t_{\frac{\alpha}{2},n-2} \cdot \sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)} \cdot s_\varepsilon \right) \tag{9.141}$$
$$a \in (12.7992, 28.6228)$$

Ocenjena prodajna cena stanovanja, če je njegova ocenjena vrednost enaka $x_{nov} = 175$ *milijonov SIT*, je enaka:

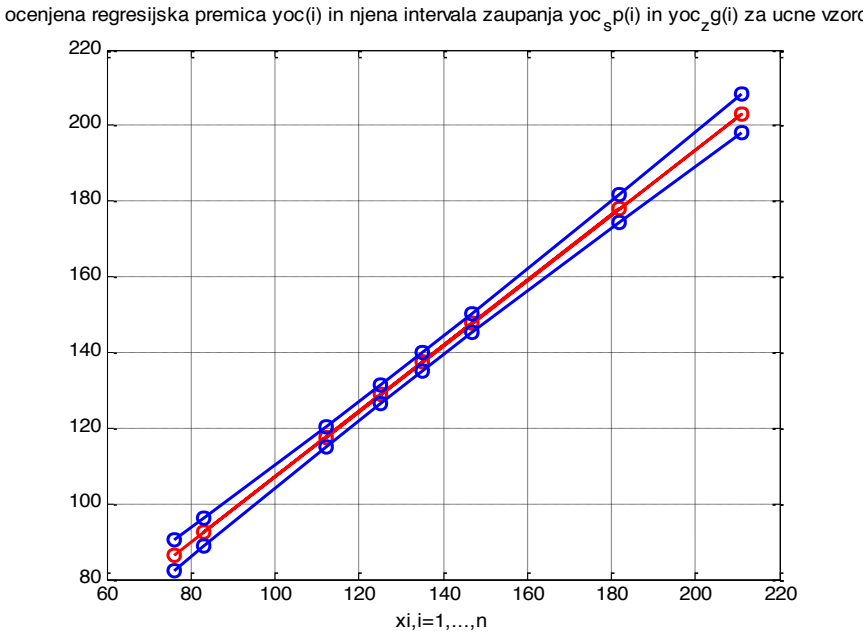
$$\hat{\bar{y}}_{nov} = \hat{a} + \hat{b} \cdot x_{nov} = 172.0692 \tag{9.142}$$

Njen interval zaupanja pa je na osnovi izraza (9.128) enak:

$$y_{nov} \in \left(\hat{\bar{y}}_{nov} - t_{\frac{\alpha}{2},n-2} \cdot s_e \cdot \sqrt{\left(1 + \frac{1}{n} + \frac{(x_{nov} - \bar{x})^2}{S_{xx}} \right)}, \hat{\bar{y}}_{nov} + t_{\frac{\alpha}{2},n-2} \cdot s_e \cdot \sqrt{\left(1 + \frac{1}{n} + \frac{(x_{nov} - \bar{x})^2}{S_{xx}} \right)} \right) \tag{9.143}$$
$$y_{nov} \in (164.4213, 179.717)$$

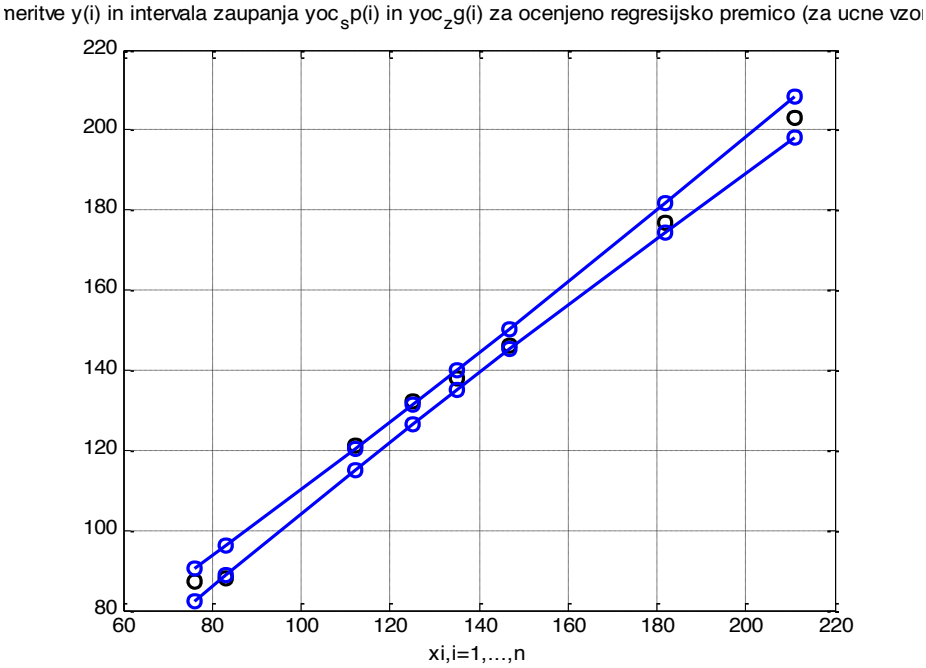
Torej s 95% verjetnostjo napovedujemo, da bo prodajna cena stanovanja dosegla vrednost med 164.4212 in 179.717 milijona SIT, če bo njegova ocenjena vrednost enaka 175 mio SIT.

Slika 218 prikazuje spodnjo in zgornjo mejo zaupanja vrednosti ocenjene regresijske premice pri $x_i, i = 1,...,8$, ki so bile uporabljene pri ocenjevanju obeh parametrov. Na njej je narisana tudi ocenjena regresijska premica. Pri tem je bil za izračun mej uporabljen izraz (9.118).



Slika 218: Ocenjena regresijska premica, ter njeni spodnja in zgornja meja zaupanja pri vrednostih $x_i, i = 1,...,8$*, ki so bile uporabljene pri ocenjevanju obeh parametrov.*

Slika 219 prikazuje primerjavo po velikosti sortiranih vrednosti odvisne spremenljivke, ki so bile uporabljene pri ocenjevanju parametrov, ter intervala zaupanja ocenjene regresijske premice.



Slika 219: Primerjava vrednosti odvisne spremenljivke, ki so bile uporabljene pri ocenjevanju parametrov, ter intervala zaupanja ocenjene regresijske premice.

Številčna primerjava teh vrednosti je (izpis v komandnem oknu Matlaba):

```
sortirane vrednosti spodnje meje, meritev, in zgornje meje zaupanja so:

ans =

  Columns 1 through 7

   82.3817   88.7441  114.8566  126.3433  135.0423  145.3123  174.4861
   87.0000   88.0000  121.0000  132.0000  138.0000  146.0000  177.0000
   90.5057   96.2519  120.3038  131.3046  139.9037  150.3914  181.7609

  Column 8

  198.2338
  203.0000
  208.1776
```

Kot lahko vidimo, pade pet vrednosti meritev odvisne spremenljivke znotraj obeh meja, tri vrednosti pa padejo za malenkost izven.

V splošnem velja, da se mora pri neki novi vrednosti neodvisne spremenljivke $x_{nov} \in x_{novj}, j = 1,2,...,$ ki ni bila uporabljena pri ocenjevanju, interval zaupanja ocenjenih vrednosti odvisne spremenljivke povečati. Razlog je v tem, da bo nova meritev odvisne spremenljivke zaradi nove vrednosti $\varepsilon_{nov} \in N(0,\sigma)$ povečala vpliv skupne variance naključnih vplivov za σ^2, ki ga moramo dodatno zajeti v intervalu zaupanja. Tako lahko slednjega zapišemo na osnovi izraza (9.143) v naslednji obliki [Jesenko]:

$$y_{nov} \in \left(\hat{\bar{y}}_{nov} - t_{\frac{\alpha}{2},n-2} \cdot s_e \cdot \sqrt{\left(1 + \frac{1}{n} + \frac{(x_{nov} - \bar{x})^2}{S_{xx}}\right)}, \hat{\bar{y}}_{nov} + t_{\frac{\alpha}{2},n-2} \cdot s_e \cdot \sqrt{\left(1 + \frac{1}{n} + \frac{(x_{nov} - \bar{x})^2}{S_{xx}}\right)} \right) \quad (9.144)$$

$$x_{nov} \in (x_{nov1}, x_{nov2}, ...)$$

$$\hat{\bar{y}}_{nov} = \hat{a} + \hat{b} \cdot x_{nov}$$

Če želimo narisati na skupnem grafu dane meritve $x_i, y_i, i = 1,...,n$, novo vrednost neodvisne spremenljivke iz množice $x_{nov} \in x_{novj}, j = 1,2,...,$ ocenjeno regresijsko premico, ter osnovni in razširjen interval zaupanja, lahko zapišemo (glej sliko 220):

$$x' = [x_i, x_{nov}], \quad i = 1,...,n$$

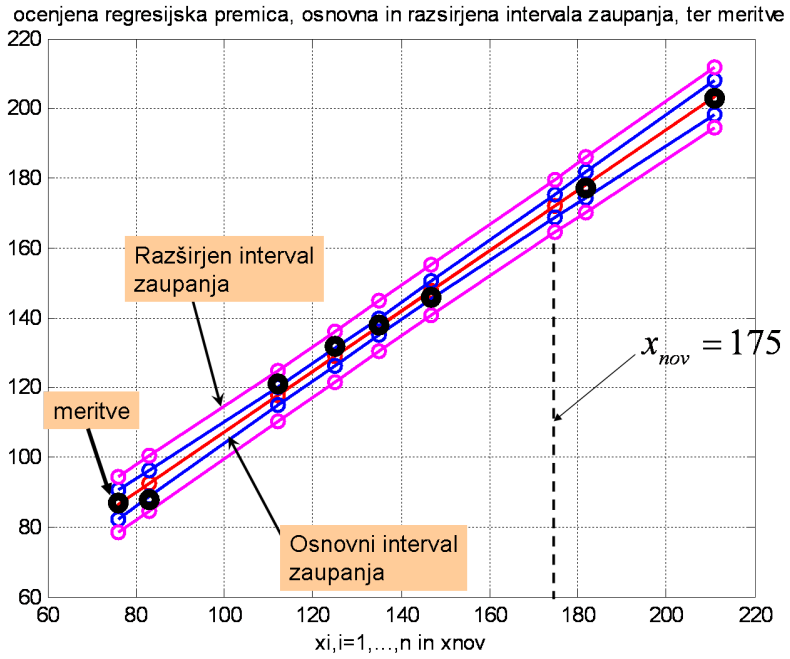
$$\hat{\hat{y}}_{nov} = \hat{a} + \hat{b} \cdot x'$$

$$y_{osnovni} \in \left(\hat{\hat{y}}_{nov} - t_{\frac{\alpha}{2},n-2} \cdot s_e \cdot \sqrt{\left(\frac{1}{n} + \frac{(x' - \bar{x})^2}{S_{xx}} \right)}, \hat{\hat{y}}_{nov} + t_{\frac{\alpha}{2},n-2} \cdot s_e \cdot \sqrt{\left(\frac{1}{n} + \frac{(x' - \bar{x})^2}{S_{xx}} \right)} \right) \tag{9.145}$$

$$y_{razširjen} \in \left(\hat{\hat{y}}_{nov} - t_{\frac{\alpha}{2},n-2} \cdot s_e \cdot \sqrt{\left(1 + \frac{1}{n} + \frac{(x' - \bar{x})^2}{S_{xx}} \right)}, \hat{\hat{y}}_{nov} + t_{\frac{\alpha}{2},n-2} \cdot s_e \cdot \sqrt{\left(1 + \frac{1}{n} + \frac{(x' - \bar{x})^2}{S_{xx}} \right)} \right)$$

$$x_{nov} \in (x_{nov1}, x_{nov2}, ...)$$

Pri tem načeloma velja, da bi lahko pri dani vrednosti neodvisne spremenljivke $X = x_i, i = 1,...,n$ z verjetnostjo $(1-\alpha)$ pričakovali, da bo matematično upanje odvisne spremenljivke padlo v osnovni interval, za vrednost $x_{nov} \in x_{novj}, j = 1,2,...$ pa ne, saj je nismo uporabili pri ocenjevanju. Za razliko od osnovnega intervala pa bi moral razširjen interval pokriti tudi vrednost $x_{nov} \in x_{novj}, j = 1,2,...$ z verjetnostjo $(1-\alpha)$.



Slika 220: Meritve, ocenjena regresijska premica, ter osnovni in razširjen interval zaupanja

Pri izračunih in izrisu slik smo si pomagali z naslednjim programom v Matlabu:

```
% intzaup.m

clear
clc
close all

pr = input('primer 1(1) ali primer 2(2)');
if pr == 1
   x = [125 83 182 135 147 112 211 76]
   y = [132 88 177 138 146 121 203 87]
   alfa = 0.05
   xnov = 175
else
   n = 25
   Sxy = 70690
   Sxx = 19800
   Syy = 307203
   xsr = 70
   ysr = 312.28
   sx = n*xsr
   sy = n*ysr
   xtst1 = 65
   xtst2 = 100
   xnov = 100
   alfa = 0.05
   alfa1 = 0.10
   xnov1 = [100 100 100]
end

if pr == 1

   n = length(x)

   disp('vsota x je:')
   sx = sum(x)

   disp('vsota y je:')
   sy = sum(y)

   disp('vsota x^2 je:')
   sx2 = x*x'

   disp('vsota y^2 je:')
   sy2 = y*y'

   disp('vsota x.y je:')
   sxy = x*y'

   disp('Sxy je:')
   Sxy = sxy - sx*sy/n

   disp('Sxx je:')
   Sxx = sx2 - sx^2/n

   disp('Syy je:')
   Syy = sy2 - sy^2/n

end

disp('boc=')
boc = Sxy/Sxx
```

```
disp('aoc=')
aoc = sy/n - boc*sx/n


se=sqrt((Syy-boc*Sxy)/(n-2)) % standardna ocena napake modela

% kriticna vrednost t statistike:

tkrit = abs(tinv(alfa/2,n-2))

disp('interval zaupanja za parameter b:')

Ib = [boc-tkrit*se/sqrt(Sxx),boc+tkrit*se/sqrt(Sxx)]

disp('interval zaupanja za parameter a:')

if pr == 1

   Ia = [aoc-tkrit*se*sqrt(1/n+(sx/n)^2/Sxx),aoc+tkrit*se*sqrt(1/n+(sx/n)^2/Sxx)]

else

   tkrit = abs(tinv(alfa1/2,n-2))
   Ia = [aoc-tkrit*se*sqrt(1/n+(sx/n)^2/Sxx),aoc+tkrit*se*sqrt(1/n+(sx/n)^2/Sxx)]

end

% generirajmo interval zaupanja za vse xi,i=1,...,n, pri katerih sta bila
% ocenjena a in b:

if pr == 1

    for i=1:length(x)
        yoc(i)= aoc+boc*x(i);
        yoc_sp(i) = aoc+boc*x(i)-tkrit*se*sqrt(1/n+(x(i)-(sx/n))^2/Sxx);
        yoc_zg(i) = aoc+boc*x(i)+tkrit*se*sqrt(1/n+(x(i)-(sx/n))^2/Sxx);
    end

    plot(x,yoc,'r','LineWidth',2)
    hold on
    plot(x,yoc,'ro','LineWidth',2)
    plot(sort(x),sort(yoc_sp),'b','LineWidth',2)
    plot(x,yoc_sp,'bo','LineWidth',2)
    plot(sort(x),sort(yoc_zg),'b','LineWidth',2)
    plot(x,yoc_zg,'bo','LineWidth',2)
    grid
    title('ocenjena regresijska premica yoc(i) in njena intervala zaupanja yoc_sp(i) in ...yoc_zg(i) za ucne vzorce')
    xlabel('xi,i=1,...,n')
    comp_mer_iz % primerjamo meritve z osnovnima intervaloma zaupanja pri učnih vzorcih
    compare1    % primerjamo meritve z osnovnim in razširjenim intervalom zaupanja (ko dodamo ...xnov)

end

disp('interval zaupanja za ocenjeno regresijsko premico pri xnov (ni del ucne mnozice):')

yoc_nov = aoc + boc*xnov

Iyoc_nov=[yoc_nov-tkrit*se*sqrt(1+1/n+(xnov-(sx/n))^2/Sxx),...
          yoc_nov+tkrit*se*sqrt(1+1/n+(xnov-(sx/n))^2/Sxx)]

if pr == 1
   return
end

disp('interval zaupanja za ocenjeno regresijsko premico pri xtst1 (je en od ucne mnozice xi-jev):')

yoc_tst1 = aoc + boc*xtst1
```

```
Iyoc_tst1=[yoc_tst1-tkrit*se*sqrt(1/n+(xtst1-(sx/n))^2/Sxx),...
          yoc_tst1+tkrit*se*sqrt(1/n+(xtst1-(sx/n))^2/Sxx)]

disp('interval zaupanja za ocenjeno regresijsko premico pri xtst2 (je en od ucne mnozice xi-jev):')

yoc_tst2 = aoc + boc*xtst2

Iyoc_tst2=[yoc_tst2-tkrit*se*sqrt(1/n+(xtst2-(sx/n))^2/Sxx),...
          yoc_tst2+tkrit*se*sqrt(1/n+(xtst2-(sx/n))^2/Sxx)]

disp('interval zaupanja za ocenjeno regresijsko premico pri m prihodnjih xnov1 (niso del ucne mnozice):')

m = length(xnov1)

yoc_nov1 = aoc + boc*xnov1(1)

Iyoc_nov1=[yoc_nov1-tkrit*se*sqrt(1/m+1/n+(xnov1(1)-(sx/n))^2/Sxx),...
          yoc_nov1+tkrit*se*sqrt(1/m+1/n+(xnov1(1)-(sx/n))^2/Sxx)]
```

Program **comp_mer_iz.m** ima obliko:

```
% comp_mer_iz.m

% primerjajmo ucne meritve in osnovna intervala zaupanja pri ocenjeni regresij.
% premici (za ucne vzorce):

figure
plot(sort(x),sort(y),'ko','LineWidth',1.5)
hold on
plot(sort(x),sort(yoc_sp),'b','LineWidth',2)
plot(sort(x),sort(yoc_sp),'bo','LineWidth',2)
plot(sort(x),sort(yoc_zg),'b','LineWidth',2)
plot(sort(x),sort(yoc_zg),'bo','LineWidth',2)
grid
title('meritve y(i) in intervala zaupanja yoc_sp(i) in yoc_zg(i) za ocenjeno regresijsko
premico (za ucne vzorce)')
xlabel('xi,i=1,...,n')

disp('sortirane vrednosti spodnje meje, meritev, in zgornje meje zaupanja pri ucnih
vzorcih so:')
[sort(yoc_sp);sort(y);sort(yoc_zg)]
```

Program **compare1.m** ima obliko:

```
% compare1.m
% Narisimo meritve, ocenjeno regresijsko premico, osnovni  interval zaupanja (na podlagi
ucnih
% vzorcev) in razsirjen interval zaupanja, ce dodamo novo vrednost neodvisne
% spremenljivke, ki ni bila uporabljena pri ocenjevanju.

xstar = x
x = [x xnov]   % neodvisna spremenljivka pri ocenjevanju in nova vrednost, ki je ni bilo
pri ocenjevanju

for i=1:length(x)
    yoc(i)= aoc+boc*x(i);
    % osnovni interval zaupanja:
    yoc_sp(i) = aoc+boc*x(i)-tkrit*se*sqrt(1/n+(x(i)-(sx/n))^2/Sxx);
    yoc_zg(i) = aoc+boc*x(i)+tkrit*se*sqrt(1/n+(x(i)-(sx/n))^2/Sxx);
    % razsirjen interval zaupanja:
    yoc_sp1(i) = aoc+boc*x(i)-tkrit*se*sqrt(1+1/n+(x(i)-(sx/n))^2/Sxx);
    yoc_zg1(i) = aoc+boc*x(i)+tkrit*se*sqrt(1+1/n+(x(i)-(sx/n))^2/Sxx);
end

% ocenjena regresijska premica:

figure
plot(x,yoc,'r','LineWidth',2)
hold on
plot(x,yoc,'ro','LineWidth',2)

% osnovni interval zaupanja:

plot(sort(x),sort(yoc_sp),'b','LineWidth',2)
plot(x,yoc_sp,'bo','LineWidth',2)
plot(sort(x),sort(yoc_zg),'b','LineWidth',2)
plot(x,yoc_zg,'bo','LineWidth',2)

% razsirjen interval zaupanja:

plot(sort(x),sort(yoc_sp1),'m','LineWidth',2)
plot(x,yoc_sp1,'mo','LineWidth',2)
plot(sort(x),sort(yoc_zg1),'m','LineWidth',2)
plot(x,yoc_zg1,'mo','LineWidth',2)

grid
title('ocenjena regresijska premica, osnovna in razsirjena intervala zaupanja, ter
meritve')
xlabel('xi,i=1,...,n in xnov')

% meritve:
```

```
plot(sort(xstar),sort(y),'ko','LineWidth',4)
```

Komandno okno ima izgled:

```
primer 1(1) ali primer 2(2)1

x =
  125   83  182  135  147  112  211   76
y =
  132   88  177  138  146  121  203   87

alfa =
   0.0500

xnov =
  175

n =
    8

vsota x je:
sx =
      1071
vsota y je:
sy =
      1092
vsota x^2 je:
sx2 =
    158313
vsota y^2 je:
sy2 =
    160276
vsota x.y je:
sxy =
    159107

Sxy je:
Sxy =
 1.2916e+004
Sxx je:
Sxx =
 1.4933e+004
Syy je:
Syy =
     11218

boc=
boc =
```

```
    0.8649
aoc=
aoc =
   20.7110
se =
    2.8088

tkrit =
    2.4469

interval zaupanja za parameter b:
Ib =
    0.8087    0.9211

interval zaupanja za parameter a:
Ia =
   12.7992   28.6228

sortirane vrednosti spodnje meje, meritev, in zgornje meje zaupanja pri ucnih vzorcih so:
ans =
   82.3817   88.7441  114.8566  126.3433  135.0423  145.3123  174.4861  198.2338
   87.0000   88.0000  121.0000  132.0000  138.0000  146.0000  177.0000  203.0000
   90.5057   96.2519  120.3038  131.3046  139.9037  150.3914  181.7609  208.1776

xstar =
   125    83   182   135   147   112   211    76
x =
   125    83   182   135   147   112   211    76   175

interval zaupanja za ocenjeno regresijsko premico pri xnov (ni del ucne mnozice):

yoc_nov =
  172.0692

Iyoc_nov =
  164.4213  179.7170
```

Primer 9.8.:

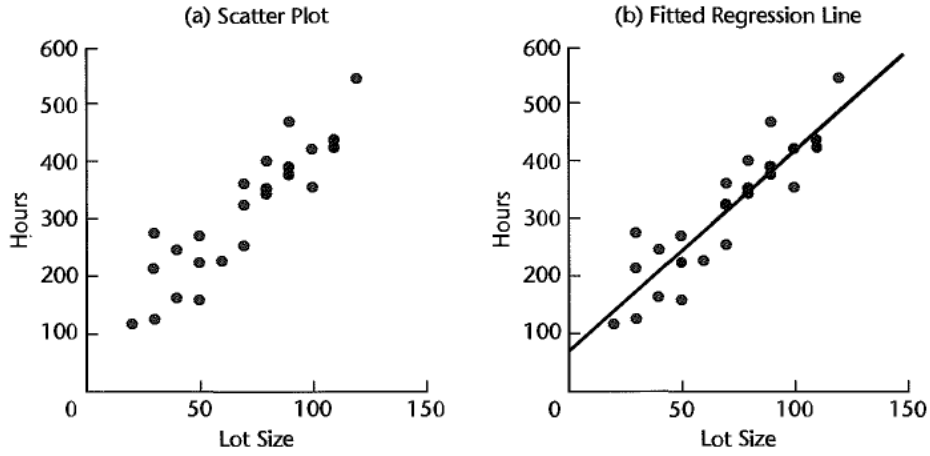
Neko podjetje proizvaja hladilnike in rezervne dele zanje [Kutner]. Glede na to, da je pri proizvodnji enega dela prihajalo do precejšnjih odstopanj v velikosti obsega njegove produkcije, je želelo podjetje določiti optimalno velikost obsega proizvodnje. Pri tem je bila relacija med velikostjo obsega in številom porabljenih delovnih ur ključnega pomena. Izvedene so bile meritve v 25 proizvodnih ciklih, kjer se je opazovala velikost obsega $x_i, i = 1,...,25$ *in število delovnih ur* $y_i, i = 1,...,25$ *(glej tabelo na sliki 221) [Kutner].*

TABLE 1.1 Data on Lot Size and Work Hours and Needed Calculations for Least Squares Estimates—Toluca Company Example.

Run i	(1) Lot Size X_i	(2) Work Hours Y_i	(3) $X_i - \bar{X}$	(4) $Y_i - \bar{Y}$	(5) $(X_i - \bar{X})(Y_i - \bar{Y})$	(6) $(X_i - \bar{X})^2$	(7) $(Y_i - \bar{Y})^2$
1	80	399	10	86.72	867.2	100	7,520.4
2	30	121	−40	−191.28	7,651.2	1,600	36,588.0
3	50	221	−20	−91.28	1,825.6	400	8,332.0
...	...	...	...	...	...	...	...
23	40	244	−30	−68.28	2,048.4	900	4,662.2
24	80	342	10	29.72	297.2	100	883.3
25	70	323	0	10.72	0.0	0	114.9
Total	1,750	7,807	0	0	70,690	19,800	307,203
Mean	70.0	312.28					

Slika 221: Velikost obsega produkcije (lot size) $x_i, i = 1,...,25$ *in število delovnih ur (work hours)* $y_i, i = 1,...,25$ *[Kutner]*

Podatki so prikazani na sliki 222, kjer je prikazana tudi regresija s premico [Kutner].



Slika 222: Podatki (lot size – velikost obsega proizvodnje, hours – delovne ure) in regresija s premico (fitted regression line) [Kutner]

Izračunajte vrednost ocenjenih parametrov. Ocenite mejo zaupanja za parameter b pri $\alpha_1 = 0.05$. *Ocenite mejo zaupanja za parameter a pri* $\alpha_2 = 0.1$. *Izračunajte tudi interval zaupanja za* ***srednjo vrednost*** *števila delovnih ur (*$\alpha_2 = 0.1$*), če sta bili naročili* $x_{tst1} = 65\,kosov$ *ali* $x_{tst2} = 100\,kosov$ *uporabljeni pri ocenjevanju parametrov. Izračunajte tudi interval zaupanja za* ***vrednost*** *števila delovnih ur (*$\alpha_2 = 0.1$*) za naslednji mesec, če vemo vnaprej za naročilo* $x_{nov} = 100\,kosov$ *za naslednji mesec (ta vrednost ni bila uporabljena pri ocenjevanju parametrov). Izračunajte tudi interval zaupanja za* ***srednjo vrednost*** *števila delovnih ur za naslednje 3 mesece (*$\alpha_2 = 0.1$*), če je vnaprej znano (vedno enako) naročilo* $x_{nov1} = 100\,kosov$ *za naslednje 3 mesece (ta vrednost ni bila uporabljena pri ocenjevanju parametrov).*

Imamo (nekatere veličine so podane v [Kutner]):

$$\begin{aligned}
&\alpha_1 = 0.05 \\
&\alpha_2 = 0.1 \\
&S_{xx} = 19800 \quad (Kutner) \\
&S_{xy} = 70690 \quad (Kutner) \\
&S_{yy} = 307203 \quad (Kutner) \\
&\bar{x} = 70 \quad (Kutner) \\
&\bar{y} = 312.28 \quad (Kutner) \\
&n = 25
\end{aligned} \tag{9.146}$$

Ocenjena parametra na osnovi (9.27) in (9.28) prideta:

$$\begin{aligned}
\hat{b} &= \frac{S_{xy}}{S_{xx}} = 3.5702 \\
\hat{a} &= \bar{y} - \hat{b} \cdot \bar{x} = 62.3659
\end{aligned} \tag{9.147}$$

Vrednost s_e je enaka:

$$s_\varepsilon = \sqrt{\frac{1}{n-2} \cdot \left(S_{yy} - \hat{b} \cdot S_{xy} \right)} = 48.8233 \tag{9.148}$$

Kritična vrednost t statistike za parameter *b* je:

$$t_{krit1} = t\left(\frac{\alpha_1}{2}, n-2 \right) = t\left(0.025, 23\right) = 2.0687 \tag{9.149}$$

Interval zaupanja za regresijski parameter *b* je:

$$b \in \left(\hat{b} - t_{\frac{\alpha_1}{2}, n-2} \cdot \frac{s_\varepsilon}{\sqrt{S_{xx}}}, \hat{b} + t_{\frac{\alpha_1}{2}, n-2} \cdot \frac{s_\varepsilon}{\sqrt{S_{xx}}} \right) \tag{9.150}$$

$$b \in \left(2.8524, 4.2880\right)$$

Kritična vrednost t statistike za parameter *a* je:

$$t_{krit2} = t\left(\frac{\alpha_2}{2}, n-2 \right) = t\left(0.05, 23\right) = 1.7139 \tag{9.151}$$

Interval zaupanja za regresijski parameter *a* je:

$$a \in \left(\hat{a} - t_{\frac{\alpha_2}{2}, n-2} \cdot \sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)} \cdot s_\varepsilon, \hat{a} + t_{\frac{\alpha_2}{2}, n-2} \cdot \sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)} \cdot s_\varepsilon \right) \tag{9.152}$$

$$a \in \left(17.5011, 107.2306\right)$$

Ocenjena vrednost odvisne spremenljivke in interval zaupanja za **srednjo vrednost** števila delovnih ur, če je bilo naročilo $x_{tst1} = 65\,kosov$ uporabljeno pri ocenjevanju, sta:

$$\hat{\bar{y}}_{tst1} = \hat{a} + \hat{b} \cdot x_{tst1} = 294.429\,ur \tag{9.153}$$

$$\bar{y}_{tst1} \in \left(\hat{\bar{y}}_{tst1} - t_{\frac{\alpha_2}{2},n-2} \cdot s_e \cdot \sqrt{\left(\frac{1}{n} + \frac{\left(x_{tst1} - \bar{x} \right)^2}{S_{xx}} \right)}, \hat{\bar{y}}_{tst1} + t_{\frac{\alpha_2}{2},n-2} \cdot s_e \cdot \sqrt{\left(\frac{1}{n} + \frac{\left(x_{tst1} - \bar{x} \right)^2}{S_{xx}} \right)} \right) \tag{9.154}$$

$$\bar{y}_{tst1} \in (277.4313, 311.4267)$$

Torej lahko zaključimo z 90% verjetnostjo, da bo srednje število delovnih ur, potrebno za izdelavo 65 kosov, nekje med 277.4 in 311.4 ure.

Ocenjena vrednost odvisne spremenljivke in interval zaupanja za **srednjo vrednost** števila delovnih ur, če je bilo naročilo $x_{tst2} = 100\,kosov$ uporabljeno pri ocenjevanju, sta:

$$\hat{\bar{y}}_{tst2} = \hat{a} + \hat{b} \cdot x_{tst2} = 419.3861\,ur \tag{9.155}$$

$$\bar{y}_{tst2} \in \left(\hat{\bar{y}}_{tst2} - t_{\frac{\alpha_2}{2},n-2} \cdot s_e \cdot \sqrt{\left(\frac{1}{n} + \frac{\left(x_{tst2} - \bar{x} \right)^2}{S_{xx}} \right)}, \hat{\bar{y}}_{tst2} + t_{\frac{\alpha_2}{2},n-2} \cdot s_e \cdot \sqrt{\left(\frac{1}{n} + \frac{\left(x_{tst2} - \bar{x} \right)^2}{S_{xx}} \right)} \right) \tag{9.156}$$

$$\bar{y}_{tst2} \in (394.9251, 443.847)$$

Torej lahko zaključimo z 90% verjetnostjo, da bo srednje število delovnih ur, potrebno za izdelavo 100 kosov, nekje med 394.9251 in 443.847 ure.

Ocenjena vrednost odvisne spremenljivke in interval zaupanja za **vrednost** števila delovnih ur za naslednji mesec, če vemo vnaprej za naročilo $x_{nov} = 100\,kosov$ za naslednji mesec (ta vrednost ni bila uporabljena pri ocenjevanju parametrov), sta enaka:

$$\hat{\bar{y}}_{nov} = \hat{a} + \hat{b} \cdot x_{nov} = 419.3861\, ur \tag{9.157}$$

$$y_{nov} \in \left(\hat{\bar{y}}_{nov} - t_{\frac{\alpha_2}{2},n-2} \cdot s_e \cdot \sqrt{\left(1 + \frac{1}{n} + \frac{\left(x_{nov} - \bar{x} \right)^2}{S_{xx}} \right)}, \hat{\bar{y}}_{nov} + t_{\frac{\alpha_2}{2},n-2} \cdot s_e \cdot \sqrt{\left(1 + \frac{1}{n} + \frac{\left(x_{nov} - \bar{x} \right)^2}{S_{xx}} \right)} \right) \tag{9.158}$$

$$y_{nov} \in \left(332.2072, 506.5649 \right)$$

Torej lahko zaključimo z 90% verjetnostjo, da bo napoved za število delovnih ur, potrebnih za izdelavo 100 kosov v naslednjem mesecu, nekje med 332.2072 in 506.5649 ure. Kot vidimo, je tukaj interval zaupanja za **vrednost** odvisne spremenljivke (v naslednjem mesecu) širši kot v izrazu (9.156) za njeno **srednjo vrednost** (v splošnem), saj bo nova meritev odvisne spremenljivke prinesla s seboj nov naključni vpliv, ki ga je potrebno upoštevati.

Ocenjena vrednost odvisne spremenljivke in interval zaupanja za **srednjo vrednost** števila delovnih ur za naslednje 3 mesece, če vemo vnaprej za naročilo $x_{nov1} = 100\, kosov$ za naslednje 3 mesece (ta vrednost ni bila uporabljena pri ocenjevanju parametrov), sta enaka:

$$\hat{\bar{y}}_{nov1} = \hat{a} + \hat{b} \cdot x_{nov1} = 419.3861\, ur \tag{9.159}$$

$$\bar{y}_{nov1} \in \left(\hat{\bar{y}}_{nov1} - t_{\frac{\alpha_2}{2},n-2} \cdot s_e \cdot \sqrt{\left(\frac{1}{m} + \frac{1}{n} + \frac{\left(x_{nov1} - \bar{x} \right)^2}{S_{xx}} \right)}, \hat{\bar{y}}_{nov1} + t_{\frac{\alpha_2}{2},n-2} \cdot s_e \cdot \sqrt{\left(\frac{1}{m} + \frac{1}{n} + \frac{\left(x_{nov1} - \bar{x} \right)^2}{S_{xx}} \right)} \right) \tag{9.160}$$

$$\bar{y}_{nov1} \in \left(365.2356, 473.5365 \right)$$

Tukaj je interval zaupanja nekoliko ožji kot v izrazu (9.158), saj pri napovedi **srednje vrednosti** za naslednje 3 mesece potrebujemo ožji interval zaupanja (manj tvegamo, da napoved ne bi bila točna), kot pa pri napovedi **ene same vrednosti** za naslednji mesec

[Kutner]. Še vedno pa je ta interval šiši kot pri izrazu (9.156), saj bodo nove 3 meritve odvisne spremenljivke prinesle s seboj nove 3 naključne vplive, ki jih je potrebno upoštevati.

Tudi tukaj za vse izračune uporabimo program **intzaup.m**. Izgled komandnega okna je naslednji:

```
primer 1(1) ali primer 2(2)2

n =
    25
Sxy =
       70690
Sxx =
       19800
Syy =
      307203
xsr =
    70
ysr =
  312.2800
sx =
        1750
sy =
  7.8070e+003
xtst1 =
    65
xtst2 =
   100
xnov =
   100
alfa =
    0.0500
alfa1 =
    0.1000
xnov1 =
   100   100   100

boc=
boc =
    3.5702
aoc=
aoc =
   62.3659

se =
   48.8233
```

```
tkrit =
   2.0687



interval zaupanja za parameter b:
Ib =
   2.8524    4.2880

interval zaupanja za parameter a:

tkrit =
   1.7139

Ia =
  17.5011  107.2306

interval zaupanja za ocenjeno regresijsko premico pri xnov (ni del ucne mnozice):
yoc_nov =
  419.3861
Iyoc_nov =
  332.2072  506.5649

interval zaupanja za ocenjeno regresijsko premico pri xtst1 (je en od ucne mnozice xi-jev):
yoc_tst1 =
  294.4290
Iyoc_tst1 =
  277.4315  311.4264

interval zaupanja za ocenjeno regresijsko premico pri xtst2 (je en od ucne mnozice xi-jev):
yoc_tst2 =
  419.3861
Iyoc_tst2 =
  394.9251  443.8470

interval zaupanja za ocenjeno regresijsko premico pri m prihodnjih xnov1 (niso del ucne mnozice):
m =
    3
yoc_nov1 =
  419.3861
Iyoc_nov1 =
  365.2356  473.5365
```

Primer 9.9.:

Vodstvo podjetja za izdelavo gradbenih materialov ima občutek, da je povpraševanje po pošiljkah mavčnih plošč (plasterboard shipments) povezano s številom gradbenih

dovoljenj (construction permits) v določeni občini v preteklem četrtletju. Podatki so prikazani na sliki 223 [Monks].

Table 9-7

Construction Permits (X)	Plasterboard Shipments (Y)
15	6
9	4
40	16
20	6
25	13
25	9
15	10
35	16

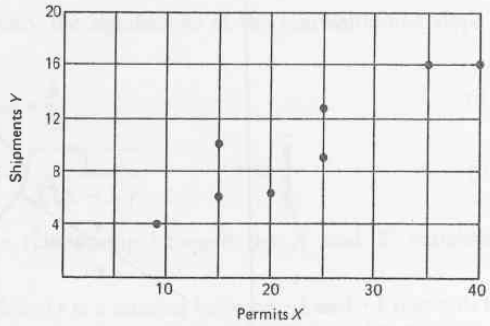


Fig. 9-5

Slika 223: Povpraševanje po pošiljkah mavčnih plošč (plasterboard shipments) v odvisnosti od števila gradbenih dovoljenj (construction permits) [Monks].

Ocenite parametra a in b, če za regresijo uporabimo premico. Določite točkasto oceno pošiljk mavčnih plošč, če je število gradbenih dovoljenj v preteklem četrtletju enako 30. Ocenite in narišite 95% predikcijski interval zaupanja za določeno število pošiljk, če je bilo število gradbenih dovoljenj v preteklem četrtletju enako 30 ($x = 30$) ali večje kot 30 ($x > 30$).

Ocenjena parametra prideta:

$$\hat{b} = \frac{S_{xy}}{S_{xx}} = \ldots = 0.3953$$
$$\hat{a} = \overline{y} - \hat{b} \cdot \overline{x} = \ldots = 0.907 \qquad (9.161)$$

Kritična vrednost t statistike je:

$$t_{krit} = t\left(\frac{\alpha}{2}, n-2\right) = t\left(0.025, 6\right) = 2.4469 \qquad (9.162)$$

Vrednost s_e je enaka:

$$s_\varepsilon = \sqrt{\frac{1}{n-2}\cdot\left(S_{yy} - \hat{b}\cdot S_{xy}\right)} = \ldots = 2.1994 \tag{9.163}$$

Ocenjena **<u>srednja vrednost</u>** odvisne spremenljivke pri $x = 30$ pride:

$$\hat{\bar{y}}_{x=30} = \hat{\bar{y}}(30) = \hat{a} + \hat{b}\cdot 30 = 12.7674 \tag{9.164}$$

Njen interval zaupanja za **<u>vrednost</u>** odvisne spremenljivke je enak:

$$y_{x=30} \in \left(\hat{\bar{y}}_{x=30} - t_{\frac{\alpha}{2},n-2}\cdot s_e \cdot \sqrt{\left(1+\frac{1}{n}+\frac{(30-\bar{x})^2}{S_{xx}}\right)}, \hat{\bar{y}}_{x=30} + t_{\frac{\alpha}{2},n-2}\cdot s_e \cdot \sqrt{\left(1+\frac{1}{n}+\frac{(30-\bar{x})^2}{S_{xx}}\right)} \right) \tag{9.165}$$

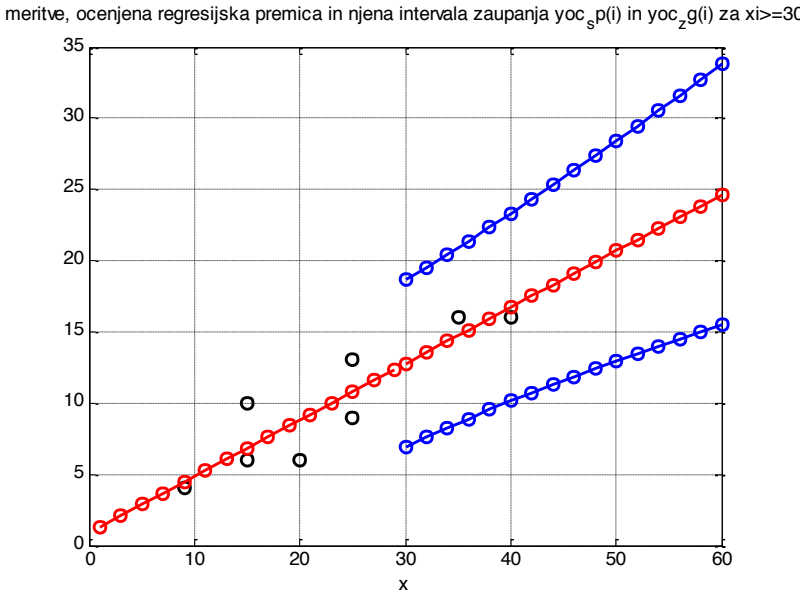
$$y_{nov} \in (6.9009, 18.634)$$

95% predikcijski interval zaupanja za določeno število pošiljk, če je bilo število gradbenih dovoljenj v preteklem četrtletju enako 30 ($x = 30$) ali večje kot 30 ($x > 30$), dobimo na osnovi naslednjega izraza (glej sliki 224 in 225):

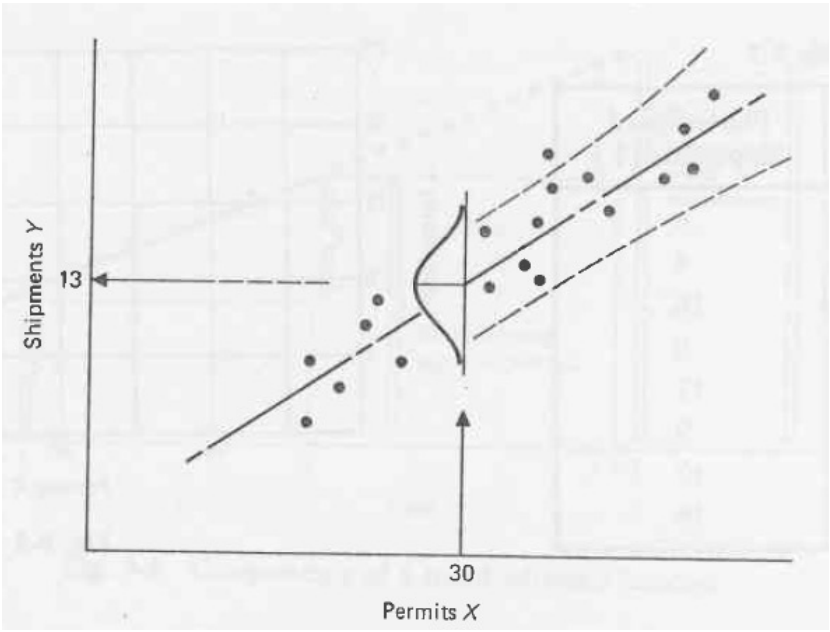
$$y(x) \in \left(\hat{\bar{y}}(x) - t_{\frac{\alpha}{2},n-2}\cdot s_e \cdot \sqrt{\left(1+\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}\right)}, \hat{\bar{y}}(x) + t_{\frac{\alpha}{2},n-2}\cdot s_e \cdot \sqrt{\left(1+\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}\right)} \right) \tag{9.166}$$

$$x \in (30, 31, \ldots)$$

$$\hat{\bar{y}}(x) = \hat{a} + \hat{b}\cdot x$$



Slika 224: Meritve, ocenjena regresijska premica in interval zaupanja za $x \geq 30$ *(Izris v Matlabu)*



Slika 224: Meritve, ocenjena regresijska premica in interval zaupanja za $x \geq 30$ *(Izris v [Monks])*

Uporabili smo naslednji program vMatlabu:

```
% intzaup1.m

clear
clc
```

```
close all

x = [15 9 40 20 25 25 15 35]
y = [6 4 16 6 13 9 10 16]

alfa = 0.05
xnov = 30
xnov1 = [30:2:60];

n = length(x)

disp('vsota x je:')
sx = sum(x)

disp('vsota y je:')
sy = sum(y)


disp('vsota x^2 je:')
sx2 = x*x'

disp('vsota y^2 je:')
sy2 = y*y'

disp('vsota x.y je:')
sxy = x*y'

disp('Sxy je:')
Sxy = sxy - sx*sy/n

disp('Sxx je:')
Sxx = sx2 - sx^2/n

disp('Syy je:')
Syy = sy2 - sy^2/n

disp('boc=')
boc = Sxy/Sxx

disp('aoc=')
aoc = sy/n - boc*sx/n

disp('Vrednost odvisne spremenljivke pri x=30')
ynov = aoc + boc*xnov

se=sqrt((Syy-boc*Sxy)/(n-2)) % standardna ocena napake modela (dobljena glede na ucne vzorce)

% kriticna vrednost t statistike:

tkrit = abs(tinv(alfa/2,n-2))

% generirajmo interval zaupanja za vse xi>=30:

for i=1:length(xnov1)
    yoc(i)= aoc+boc*xnov1(i);
    yoc_sp(i) = yoc(i)-tkrit*se*sqrt(1+1/n+(xnov1(i)-(sx/n))^2/Sxx);
    yoc_zg(i) = yoc(i)+tkrit*se*sqrt(1+1/n+(xnov1(i)-(sx/n))^2/Sxx);
end
```

```
disp('interval zaupanja za ynov pri x=30')
I = [yoc_sp(1) yoc_zg(1)]

plot(x,y,'ko','LineWidth',2)
hold on

plot(xnov1,yoc,'r','LineWidth',2)

plot(xnov1,yoc,'ro','LineWidth',2)
plot(sort(xnov1),sort(yoc_sp),'b','LineWidth',2)
plot(xnov1,yoc_sp,'bo','LineWidth',2)
plot(sort(xnov1),sort(yoc_zg),'b','LineWidth',2)
plot(xnov1,yoc_zg,'bo','LineWidth',2)
grid
title('meritve, ocenjena regresijska premica in njena intervala zaupanja yoc_sp(i) in yoc_zg(i) za
xi>=30')
xlabel('x')

xm = 1:2:30;
ym = aoc + boc*xm;
plot(xm,ym,'r','LineWidth',2)
plot(xm,ym,'ro','LineWidth',2)
```

Izgled komandnega okna je naslednji:

```
x =
   15    9   40   20   25   25   15   35
y =
    6    4   16    6   13    9   10   16

alfa =
    0.0500
xnov =
    30
n =
     8

vsota x je:
sx =
   184
vsota y je:
sy =
    80
vsota x^2 je:
sx2 =
      5006
vsota y^2 je:
sy2 =
   950
vsota x.y je:
sxy =
      2146

Sxy je:
Sxy =
```

```
   306
Sxx je:
Sxx =
   774
Syy je:
Syy =
   150

boc=
boc =
    0.3953
aoc=
aoc =
    0.9070

Vrednost odvisne spremenljivke pri x=30
ynov =
   12.7674

se =
    2.1994

tkrit =
    2.4469

interval zaupanja za ynov pri x=30
I =
    6.9009   18.6340
```

9.4.4 Determinacijski koeficient - koeficient določenosti

Varianco $VAR(Y)$ količine Y imenujemo **<u>skupna ali začetna varianca</u>** [Jesenko]. Njena točkasta ocena, izračunana pri izbranem vzorcu $\{y_1,...,y_n\}$, je enaka [Jesenko]:

$$s_Y{}^2 = \frac{1}{n-1}\cdot\sum_{i=1}^{n}(y_i-\bar{y})^2 = \frac{S_{yy}}{n-1} \tag{9.167}$$

Pri izbranem vzorcu so realizacije napake regresijskega modela enake $\varepsilon_i = y_i - \hat{a} - \hat{b}\cdot x_i, i = 1,...,n$. Varianco napake regresijskega modela ocenimo z izrazom:

$$s_e{}^2 = \frac{1}{n-2}\cdot\sum_{i=1}^{n}\left[y_i-\left(\hat{a}+\hat{b}\cdot x_i\right)\right]^2 = \frac{1}{n-2}\cdot\left(S_{yy}-\hat{b}\cdot S_{xy}\right) \tag{9.168}$$

To je najboljša ocena, kajti pripadajoča cenilka ima od vseh cenilk variance napake modela najmanjšo varianco. Varianci napake modela pravimo tudi **nepojasnena varianca** in geometrijsko pomeni povprečni kvadratni odmik točk $\{x_i, y_i\}$ od ocenjene regresijske premice $\hat{\hat{y}}_i = \hat{a} + \hat{b} \cdot x_i, i = 1,...,n$. Razliki med začetno in nepojasneno varianco pa pravimo **pojasnena varianca** in ima obliko [Jesenko]:

$$s_{XY}^2 = s_Y{}^2 - s_e{}^2 = \frac{S_{yy}}{n-1} - \frac{1}{n-2} \cdot \left(S_{yy} - \hat{b} \cdot S_{xy} \right) \tag{9.169}$$

Primer 9.10.:

Tabela na sliki 217 podaja vzorec osmih ocenjenih vrednosti stanovanj, kot so jih ocenili cenilci, ter dejansko doseženo prodajno ceno. Vrednosti so podane v milijonih SIT [Jesenko]. Kot se izkaže, sta ocenjena parametra regresijske premice enaka: $\hat{b} = 0.8649, \hat{a} = 20.7115$ *(glej izraza (9.32) in (9.33)). Izračunajte začetno, pojasneno in nepojasneno varianco.*

Imamo:

$$\begin{aligned} &S_{xx} = 14933 \quad (\textit{glej izraz } (9.31)) \\ &S_{xy} = 12916 \quad (\textit{glej izraz } (9.31)) \\ &S_{yy} = 11218 \quad (\textit{se izkaže } [\textit{Jesenko}]) \\ &\hat{b} = 0.8649 \\ &\hat{a} = 20.7115 \\ &s_e = 2.8088 \quad (\textit{glej izraz } (9.104)) \end{aligned} \tag{9.170}$$

Začetna varianca je:

$$s_Y{}^2 = \frac{S_{yy}}{n-1} = \frac{11218}{7} = 1602.6 \tag{9.171}$$

Nepojasnena varianca je:

$$s_e^{\,2} = 2.8088^2 = 7.8894 \tag{9.172}$$

Pojasnena varianca pa je enaka:

$$s_{XY}^2 = s_Y^{\,2} - s_e^{\,2} = 1602.6 - 7.8894 = 1594.7 \tag{9.173}$$

Ugotovimo torej, da je nepojasnena varianca izredno majhna v primerjavi z začetno varianco, kar se vidi tudi na sliki 206, saj ležijo točke $\{x_i, y_i\}$ skoraj na regresijski premici.

Determinacijski koeficient - koeficient določenosti

Smiselno je, da zasnujemo neko mero „kakovosti" ocenjene regresijske premice prav na primerjavi nepojasnene in skupne variance [Jesenko].

Definicija:

*Koeficient **določenosti (determinacijski koeficient)**, ki meri linearno povezavo med vzrokom X in posledico Y, je enak [Jesenko]:*

$$D = r^2 = \frac{s_{XY}^2}{s_Y^{\,2}} = \frac{s_Y^{\,2} - s_e^{\,2}}{s_Y^{\,2}} = 1 - \frac{s_e^{\,2}}{s_Y^{\,2}} \tag{9.174}$$

Skrajni primeri so:

1. $s_Y = s_e \quad \rightarrow \quad D = 0 \quad (s_{XY} = 0)$
2. $s_e = 0 \quad \rightarrow \quad D = 1 \quad (s_{XY} = s_Y)$
3. $0 < D < 1$

V 1. primeru med količino X in količino Y ni nobene linearne odvisnosti. V 2. primeru med količino X in količino Y obstaja popolna matematična povezava v obliki linearne funkcije (napaka modela je 0). V 3. primeru med X in Y obstaja **verjetna** linearna povezava.

Koeficient določenosti nam torej pove, kako dobro podaja ocenjena regresijska premica odvisnost med pojavoma X in Y. Bližje, ko je 1, bolj sta pojava linearno odvisna. Bliže, ko je koeficient določenosti vrednosti 0, slabše nam ocenjena regresijska premica popisuje linearno povezanost med proučevanima pojavoma [Jesenko]. Pri tem količini $U = 1 - D = 1 - r^2 = \frac{s_e^{\,2}}{s_Y^{\,2}}$ pravimo **koeficient nedoločenosti** [Jesenko].

Primer 9.11.:

Za podatke prejšnjega primera izračunajte koeficient določenosti!

$$D = r^2 = \frac{s_{XY}^2}{s_Y^{\,2}} = \frac{s_Y^{\,2} - s_e^{\,2}}{s_Y^{\,2}} = 1 - \frac{s_e^{\,2}}{s_Y^{\,2}} = 1 - \frac{7.8894}{1602.6} = 0.9951 \tag{9.175}$$

Na osnovi tega rezultata lahko zaključimo, da je med doseženo prodajno ceno stanovanja in njegovo ocenjeno vrednostjo 99.5% linearna odvisnost.

Izpeljava determinacijskega koeficienta [Artenjak]

Zapišimo naslednji izraz:

$$y_i-\bar{y}=y_i-\bar{y}+\hat{\bar{y}}_i-\hat{\bar{y}}_i=\left(\hat{\bar{y}}_i-\bar{y}\right)+\left(y_i-\hat{\bar{y}}_i\right)$$

kvadrirajmo :

$$\left(y_i-\bar{y}\right)^2=\left(\hat{\bar{y}}_i-\bar{y}\right)^2+2\cdot\left(\hat{\bar{y}}_i-\bar{y}\right)\left(y_i-\hat{\bar{y}}_i\right)+\left(y_i-\hat{\bar{y}}_i\right)^2$$

$$\left(y_i-\bar{y}\right)^2=\left(\hat{\bar{y}}_i-\bar{y}\right)^2+2\cdot\left(\hat{\bar{y}}_i-\bar{y}\right)(\varepsilon_i)+\left(y_i-\hat{\bar{y}}_i\right)^2$$

sumirajmo : (9.176)

$$\sum\left(y_i-\bar{y}\right)^2=\sum\left(\hat{\bar{y}}_i-\bar{y}\right)^2+2\cdot\sum\left(\hat{\bar{y}}_i-\bar{y}\right)(\varepsilon_i)+\sum\left(y_i-\hat{\bar{y}}_i\right)^2$$

$$\sum\left(y_i-\bar{y}\right)^2=\sum\left(\hat{\bar{y}}_i-\bar{y}\right)^2+2\cdot\sum\left(\hat{a}+\hat{b}\cdot x_i-\bar{y}\right)(\varepsilon_i)+\sum\left(y_i-\hat{\bar{y}}_i\right)^2$$

$$\sum\left(y_i-\bar{y}\right)^2=$$

$$=\sum\left(\hat{\bar{y}}_i-\bar{y}\right)^2+2\cdot\hat{a}\sum(\varepsilon_i)+2\cdot\hat{b}\cdot\sum(x_i)(\varepsilon_i)-2\cdot\bar{y}\sum(\varepsilon_i)+\sum\left(y_i-\hat{\bar{y}}_i\right)^2$$

Predpostavimo lahko, da približno velja pri dovolj veliki velikosti vzorca:

$$\begin{aligned}\sum(x_i)(\varepsilon_i)&\approx 0\\ \sum(\varepsilon_i)&\approx 0\end{aligned} \qquad (9.177)$$

Sledi:

$$\sum\left(y_i-\bar{y}\right)^2\approx\sum\left(\hat{\bar{y}}_i-\bar{y}\right)^2+\sum\left(y_i-\hat{\bar{y}}_i\right)^2$$

in :

$$\frac{\sum\left(y_i-\bar{y}\right)^2}{\sum\left(y_i-\bar{y}\right)^2}=1\approx\frac{\sum\left(\hat{\bar{y}}_i-\bar{y}\right)^2+\sum\left(y_i-\hat{\bar{y}}_i\right)^2}{\sum\left(y_i-\bar{y}\right)^2}=\frac{\sum\left(\hat{\bar{y}}_i-\bar{y}\right)^2}{\sum\left(y_i-\bar{y}\right)^2}+\frac{\sum\left(y_i-\hat{\bar{y}}_i\right)^2}{\sum\left(y_i-\bar{y}\right)^2}$$

torej : (9.178)

$$\frac{\frac{1}{n-1}\sum\left(\hat{\bar{y}}_i-\bar{y}\right)^2}{\frac{1}{n-1}S_{YY}}+\frac{\frac{1}{n-1}\sum(\varepsilon_i)^2}{\frac{1}{n-1}S_{YY}}\approx 1$$

$$\frac{\frac{1}{n-1}\sum\left(\hat{\bar{y}}_i-\bar{y}\right)^2}{s_Y^2}+\frac{\frac{1}{n-1}\sum(\varepsilon_i)^2}{s_Y^2}\approx 1$$

$$\frac{\frac{1}{n-1}\sum\left(\hat{\bar{y}}_i-\bar{y}\right)^2}{s_Y^2}\approx 1-\frac{\frac{1}{n-1}\sum(\varepsilon_i)^2}{s_Y^2}\approx 1-\frac{\frac{1}{n-2}\sum(\varepsilon_i)^2}{s_Y^2}=1-\frac{s_e^2}{s_Y^2}=r^2$$

Povezava med determinacijskim in korelacijskim koeficientom pri regresijski premici

Korelacijski koeficient bomo podrobneje obravnavali v naslednjem poglavju, zato na tem mestu samo podajmo dokaz, da je **kvadrat korelacijskega koeficienta za regresijsko premico približno enak determinacijskem koeficientu**.

Kot se izkaže, je na osnovi izračuna z metodo največjega verjetja korelacijski koeficient v splošnem enak [Jesenko]:

$$\rho = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}} \quad \rightarrow \quad \rho^2 = \frac{S^2{}_{xy}}{S_{xx} \cdot S_{yy}} \tag{9.179}$$

Po drugi strani je determinacijski koeficient pri premici enak:

$$D = r^2 = \frac{s_{XY}^2}{s_Y{}^2} = \frac{s_Y{}^2 - s_e{}^2}{s_Y{}^2} = \frac{\frac{S_{yy}}{n-1} - \frac{1}{n-2}\left(S_{yy} - \hat{b} \cdot S_{xy}\right)}{\frac{S_{yy}}{n-1}} =$$

$$= \frac{\frac{S_{yy}}{n-1} - \frac{1}{n-2}\left(S_{yy}\right) + \frac{1}{n-2}\left(\hat{b} \cdot S_{xy}\right)}{\frac{S_{yy}}{n-1}} = \frac{\frac{S_{yy}}{n-1} - \frac{1}{n-2}\left(S_{yy}\right) + \frac{1}{n-2}\left(\frac{S_{xy}}{S_{xx}} \cdot S_{xy}\right)}{\frac{S_{yy}}{n-1}} \approx \tag{9.180}$$

$$\approx \frac{\frac{1}{n-2}\left(\frac{S_{xy}}{S_{xx}} \cdot S_{xy}\right)}{\frac{S_{yy}}{n-1}} \approx \frac{\left(\frac{S_{xy}}{S_{xx}} \cdot S_{xy}\right)}{S_{yy}} = \frac{S_{xy}{}^2}{S_{xx} \cdot S_{yy}}$$

Če primerjamo izraza (9.179) in (9.180), vidimo, da sta približno enaka. Torej pri regresijski premici približno velja:

$$D = r^2 \approx \rho^2 \tag{9.181}$$

in je kvadrat korelacijskega koeficienta približno enak determinacijskem koeficientu.

Pokažimo še, kako iz izraza $\frac{\frac{1}{n-1}\sum\left(\hat{\bar{y}}_i-\bar{y}\right)^2}{s_Y^2}$ v (9.178) pridemo do izraza (9.179).

Napišemo lahko (če imamo regresijsko premico, seveda):

$$\frac{\frac{1}{n-1}\sum\left(\hat{\bar{y}}_i-\bar{y}\right)^2}{s_Y^2}=\frac{\frac{1}{n-1}\sum\left(\hat{a}+\hat{b}\cdot x_i-\bar{y}\right)^2}{s_Y^2}=\frac{\frac{1}{n-1}\sum\left(\bar{y}-\hat{b}\cdot\bar{x}+\hat{b}\cdot x_i-\bar{y}\right)^2}{s_Y^2}=$$

$$=\frac{\frac{1}{n-1}\sum\left(\hat{b}\cdot\left(x_i-\bar{x}\right)\right)^2}{s_Y^2}=\frac{\frac{1}{n-1}\cdot\hat{b}^2\cdot S_{xx}}{s_Y^2}=\frac{\frac{1}{n-1}\cdot\left(\frac{S_{xy}}{S_{xx}}\right)^2\cdot S_{xx}}{s_Y^2}=\frac{\frac{1}{n-1}\cdot\left(\frac{S_{xy}}{S_{xx}}\right)^2\cdot S_{xx}}{\frac{1}{n-1}\cdot S_{yy}}= \quad (9.182)$$

$$=\frac{\left(\frac{S_{xy}}{S_{xx}}\right)^2\cdot S_{xx}}{S_{yy}}=\frac{\left(\frac{S_{xy}{}^2}{S_{xx}}\right)}{S_{yy}}=\frac{S_{xy}{}^2}{S_{xx}\cdot S_{yy}}=\rho^2$$

Pri regresijski premici torej za determinacijski koeficient velja:

$$\begin{aligned} D=r^2&=\frac{s_{XY}^2}{s_Y{}^2}=\frac{s_Y{}^2-s_e{}^2}{s_Y{}^2}=1-\frac{s_e{}^2}{s_Y{}^2} \quad (1.način)\\ D\approx\rho^2&=\frac{S^2{}_{xy}}{S_{xx}\cdot S_{yy}}=\frac{\frac{1}{n-1}\sum\left(\hat{\bar{y}}_i-\bar{y}\right)^2}{s_Y^2} \quad (2.način)\end{aligned} \qquad (9.183)$$

9.4.5 Korelacijski koeficient in normalna korelacijska analiza

Vzemimo, da proučujemo dvostransko odvisna pojava $X\leftrightarrow Y$. Zanima nas, kako močno sta linearno odvisna. Predpostavimo, da sta populaciji, s katerima proučujemo pojava, normalni:

$$\begin{aligned} X&\in N\left(\mu_X,\sigma_X\right)\\ Y&\in N\left(\mu_Y,\sigma_Y\right)\end{aligned} \qquad (9.184)$$

Iz obeh populacij zberemo vzorca $\{x_1,...,x_n\}$ in $\{y_1,...,y_n\}$, ki nam določata v ravnini množico točk $\{x_i,y_i\}$. V izrazu (9.24) smo vpeljali kovarianco, ki nam meri odvisnost dveh naključnih spremenljivk. Slednja ima to pomanjkljivost, da je njena velikost odvisna od izbire merskih enot, s katerimi izražamo podatke. Tej pomanjkljivosti se lahko izognemo tako, da kovarianco normiramo, pri čemer dobimo **Pearsonov koeficient korelacije** ρ:

$$\rho=\frac{COV(X,Y)}{\sigma_X\cdot\sigma_Y}=\frac{C_{xy}}{\sigma_X\cdot\sigma_Y}\approx\frac{\frac{1}{n}\sum_{i=1}^{n}\left[\left(x(i)-\bar{x}\right)\cdot\left(y(i)-\bar{y}\right)\right]}{s_X\cdot s_Y}=$$
$$=\frac{\frac{1}{n}\sum_{i=1}^{n}\left[\left(x(i)-\bar{x}\right)\cdot\left(y(i)-\bar{y}\right)\right]}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x(i)-\bar{x}\right)^2}\cdot\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y(i)-\bar{y}\right)^2}}=\frac{\frac{1}{n}S_{XY}}{\frac{1}{n}\sqrt{\sum_{i=1}^{n}\left(x(i)-\bar{x}\right)^2\cdot\sum_{i=1}^{n}\left(y(i)-\bar{y}\right)^2}}= \tag{9.185}$$
$$=\frac{S_{XY}}{\sqrt{S_{XX}\cdot S_{YY}}}$$

Pogostokrat se za korelacijski koeficient uporablja tudi naslednji zapis:

$$\rho \approx \frac{\frac{1}{n}\sum_{i=1}^{n}\left[\left(x(i)-\bar{x}\right)\cdot\left(y(i)-\bar{y}\right)\right]}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x(i)-\bar{x}\right)^2}\cdot\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y(i)-\bar{y}\right)^2}} = \frac{\sum_{i=1}^{n}\left[\left(x(i)-\bar{x}\right)\cdot\left(y(i)-\bar{y}\right)\right]}{\sqrt{\sum_{i=1}^{n}\left(x(i)-\bar{x}\right)^2}\cdot\sqrt{\sum_{i=1}^{n}\left(y(i)-\bar{y}\right)^2}} =$$

$$= \frac{\sum_{i=1}^{n}\left[\left(x(i)\cdot y(i)-\bar{x}\cdot y(i)-x(i)\cdot\bar{y}+\bar{x}\cdot\bar{y}\right)\right]}{\sqrt{\sum_{i=1}^{n}\left(x(i)-\bar{x}\right)^2\cdot\sum_{i=1}^{n}\left(y(i)-\bar{y}\right)^2}} = \frac{\sum_{i=1}^{n}\left[x(i)\cdot y(i)\right]-2\cdot\bar{x}\cdot\bar{y}\cdot n+\bar{x}\cdot\bar{y}\cdot n}{\sqrt{\sum_{i=1}^{n}\left(x(i)-\bar{x}\right)^2\cdot\sum_{i=1}^{n}\left(y(i)-\bar{y}\right)^2}} =$$

$$= \frac{\sum_{i=1}^{n}\left[x(i)\cdot y(i)\right]-\bar{x}\cdot\bar{y}\cdot n}{\sqrt{\left(\sum_{i=1}^{n}x(i)^2-\bar{x}^2\cdot n\right)\left(\sum_{i=1}^{n}y(i)^2-\bar{y}^2\cdot n\right)}} = \frac{\sum_{i=1}^{n}\left[x(i)\cdot y(i)\right]-\frac{1}{n^2}\sum_{i=1}^{n}x(i)\cdot\sum_{i=1}^{n}y(i)\cdot n}{\sqrt{\left(\sum_{i=1}^{n}x(i)^2-\bar{x}^2\cdot n\right)\left(\sum_{i=1}^{n}y(i)^2-\bar{y}^2\cdot n\right)}} = \quad (9.186)$$

$$= \frac{\frac{1}{n}\left(n\sum_{i=1}^{n}\left[x(i)\cdot y(i)\right]-\sum_{i=1}^{n}x(i)\cdot\sum_{i=1}^{n}y(i)\right)}{\sqrt{\left(\sum_{i=1}^{n}x(i)^2-\frac{1}{n^2}\left(\sum_{i=1}^{n}x(i)\right)^2\cdot n\right)\left(\sum_{i=1}^{n}y(i)^2-\frac{1}{n^2}\left(\sum_{i=1}^{n}y(i)\right)^2\cdot n\right)}} =$$

$$= \frac{n\sum_{i=1}^{n}\left[x(i)\cdot y(i)\right]-\sum_{i=1}^{n}x(i)\cdot\sum_{i=1}^{n}y(i)}{\sqrt{\left(n\sum_{i=1}^{n}x(i)^2-\left(\sum_{i=1}^{n}x(i)\right)^2\right)\left(n\sum_{i=1}^{n}y(i)^2-\left(\sum_{i=1}^{n}y(i)\right)^2\right)}}$$

Vrednost korelacijskega koeficienta leži med -1 in +l. Kadar sta naključni spremenljivki neodvisni, vemo, da je njuna kovarianca 0 in zato je tudi $\rho = 0$, kar pomeni, da med njima ni nobene medsebojne odvisnosti. Bolj ko se pomika proti +1, bolj sta naključni spremenljivki odvisni v pozitivnem smislu (večja ko je realizacija ene, večja je tudi realizacija druge spremenljivke). Kadar pa se ρ pomika proti -1, sta naključni spremenljivki odvisni v negativnem smislu (s povečevanjem realizacije ene se zmanjšuje realizacija druge naključne spremenljivke) [Jesenko].

Če želimo priti do rezultata (9.185) oz. (9.186), to storimo tako, da za **dvorazsežni** porazdelitven zakon vpeljemo funkcijo največjega verjetja ter poščemo njen maksimum [Jesenko]. Dvorazsežni **normalni** porazdelitven zakon smo predstavili v izrazu (2.108) in ima obliko:

$$f(x,y)=\frac{1}{2\cdot\pi\cdot\sigma_x\cdot\sigma_y\cdot\sqrt{1-\rho^2}}\cdot e^{-\frac{1}{2\left(1-\rho^2\right)}\cdot\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2-2\cdot\rho\cdot\left(\frac{x-\mu_x}{\sigma_x}\right)\cdot\left(\frac{y-\mu_y}{\sigma_y}\right)+\left(\frac{y-\mu_y}{\sigma_y}\right)^2\right]} \qquad (9.187)$$

Pri *n* vzorcih dobi obliko:

$$f(x_i,y_i)=\frac{1}{2\cdot\pi\cdot s_x\cdot s_y\cdot\sqrt{1-\rho^2}}\cdot e^{-\frac{1}{2\left(1-\rho^2\right)}\cdot\left[\left(\frac{x_i-\bar{x}}{s_x}\right)^2-2\cdot\rho\cdot\left(\frac{x_i-\bar{x}}{s_x}\right)\cdot\left(\frac{y_i-\bar{y}}{s_y}\right)+\left(\frac{y_i-\bar{y}}{s_y}\right)^2\right]},i=1,\ldots, \qquad (9.188)$$

Funkcija največjega verjetja je enaka:

$$\begin{aligned}
&L=f(x_1,y_1)\cdot f(x_2,y_2)\cdot\ldots\cdot f(x_n,y_n)=\\
&=\frac{1}{2\cdot\pi\cdot s_x\cdot s_y\cdot\sqrt{1-\rho^2}}\cdot e^{-\frac{1}{2\left(1-\rho^2\right)}\cdot\left[\left(\frac{x_1-\bar{x}}{s_x}\right)^2-2\cdot\rho\cdot\left(\frac{x_1-\bar{x}}{s_x}\right)\cdot\left(\frac{y_1-\bar{y}}{s_y}\right)+\left(\frac{y_1-\bar{y}}{s_y}\right)^2\right]}\cdot\\
&\cdot\frac{1}{2\cdot\pi\cdot s_x\cdot s_y\cdot\sqrt{1-\rho^2}}\cdot e^{-\frac{1}{2\left(1-\rho^2\right)}\cdot\left[\left(\frac{x_2-\bar{x}}{s_x}\right)^2-2\cdot\rho\cdot\left(\frac{x_2-\bar{x}}{s_x}\right)\cdot\left(\frac{y_2-\bar{y}}{s_y}\right)+\left(\frac{y_2-\bar{y}}{s_y}\right)^2\right]}\cdot\\
&\ldots\\
&\cdot\frac{1}{2\cdot\pi\cdot s_x\cdot s_y\cdot\sqrt{1-\rho^2}}\cdot e^{-\frac{1}{2\left(1-\rho^2\right)}\cdot\left[\left(\frac{x_n-\bar{x}}{s_x}\right)^2-2\cdot\rho\cdot\left(\frac{x_n-\bar{x}}{s_x}\right)\cdot\left(\frac{y_n-\bar{y}}{s_y}\right)+\left(\frac{y_n-\bar{y}}{s_y}\right)^2\right]}=\\
&=\frac{1}{\left(2\cdot\pi\right)^n\cdot {s_x}^n\cdot {s_y}^n\cdot\left(1-\rho^2\right)^{\frac{n}{2}}}\cdot e^{-\frac{1}{2\left(1-\rho^2\right)}\cdot\sum_{i=1}^{n}\left[\left(\frac{x_i-\bar{x}}{s_x}\right)^2-2\cdot\rho\cdot\left(\frac{x_i-\bar{x}}{s_x}\right)\cdot\left(\frac{y_i-\bar{y}}{s_y}\right)+\left(\frac{y_i-\bar{y}}{s_y}\right)^2\right]}
\end{aligned} \qquad (9.189)$$

Dobimo torej funkcijo: $L=L(\bar{x},\bar{y},s_X,s_Y,\rho)$, ki jo moramo parcialno odvajati po vseh petih parametrih, parcialne odvode enačiti z 0 in rešiti sistem petih enačb s petimi neznankami.

Po daljši izpeljavi dobimo naslednje rezultate [Jesenko]:

$$\begin{aligned}
\bar{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\
\bar{y} &= \frac{1}{n}\sum_{i=1}^{n} y_i \\
s_X &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x(i)-\bar{x}\right)^2} \\
s_Y &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y(i)-\bar{y}\right)^2} \\
\rho &= \frac{\frac{1}{n}\sum_{i=1}^{n}\left[\left(x(i)-\bar{x}\right)\cdot\left(y(i)-\bar{y}\right)\right]}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x(i)-\bar{x}\right)^2}\cdot\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y(i)-\bar{y}\right)^2}}
\end{aligned} \tag{9.190}$$

Primer 9.12.:

Z naključnim vzorcem velikosti n = 15 želimo ugotoviti, če obstaja pri ženskah določene starosti linearna odvisnost med srčnim utripom X in zgornjim krvnim tlakom Y. Izračunajte Pearsonov koeficient korelacije, še predpostavljamo, da sta tako srčni utrip kot krvni tlak normalno porazdeljena. Podatki so prikazani na sliki 225 [Jesenko].

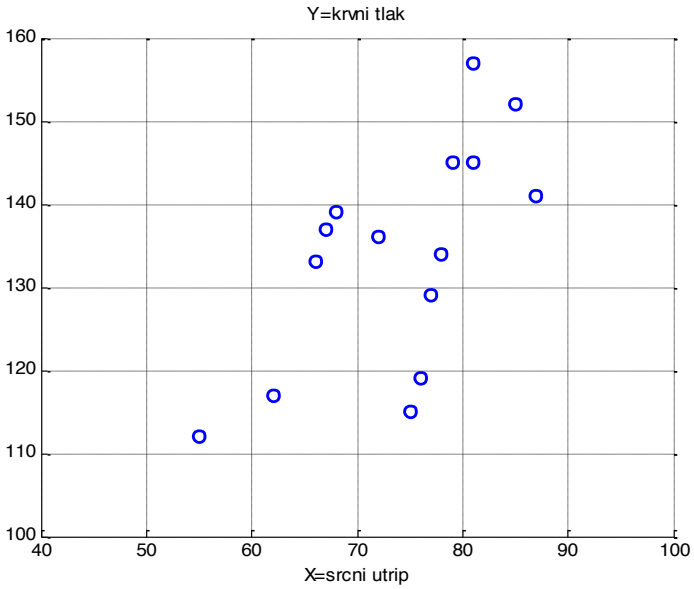
Utrip (x_i)	Krvni tlak (y_i)
77	129
68	139
55	112
76	119
79	145
66	133
85	152
62	117
81	145
72	136
75	115
67	137
78	134
87	141
81	157

Slika 225: Podatki za srčni utrip in krvni tlak [Jesenko]

Dobimo:

$$\begin{aligned} S_{xx} &= 1100.9 \\ S_{xy} &= 1093.1 \\ S_{yy} &= 2566.9 \\ \rho &= \frac{S_{XY}}{\sqrt{S_{XX} \cdot S_{YY}}} = 0.6502 \end{aligned} \tag{9.191}$$

Razsevni diagram prikazuje slika 226.



Slika 226: Razsevni diagram

Iz razsevnega diagrama razberemo, da obstaja določena povezanost med srčnim utripom in krvnim tlakom, saj točke v ravnini niso popolnoma naključno porazdeljene. Vsekakor velja, da bliže ko so točke premici, večja je linearna povezanost med pojavoma, in obratno. Determinacijski koeficient pride enak:

$$D = \rho^2 = \frac{S_{XY}{}^2}{S_{XX} \cdot S_{YY}} = 0.6502^2 = 0.4228 \qquad (9.192)$$

Torej lahko sklepamo, da je pri ženskah določene starosti 42.2% linearna odvisnost med srčnim utripom in zgornjim krvnim tlakom [Jesenko].

Pomagali smo si z naslednjim programom v Matlabu:

```
% koefkor.m

clear
clc
close all


x = [77 68 55 76 79 66 85 62 81 72 75 67 78 87 81]
y = [129 139 112 119 145 133 152 117 145 136 115 137 134 141 157]

n = length(x)

sx = sum(x);
sy = sum(y);
```

```
sx2 = x*x';
sy2 = y*y';
sxy = x*y';
Sxy = sxy - sx*sy/n
Sxx = sx2 - sx^2/n
Syy = sy2 - sy^2/n

disp('Koeficient korelacije je enak:')

ro = Sxy/sqrt(Sxx*Syy)

scatter(x,y,'LineWidth',2)
title('Y=krvni tlak')
xlabel('X=srcni utrip')
grid
axis([40 100 100 160])

disp('Determinacijski koeficient je enak:')

D = ro^2
```

Izpis komandnega okna pa je naslednji:

```
x =
  Columns 1 through 13
    77   68   55   76   79   66   85   62   81   72   75   67   78
  Columns 14 through 15
    87   81
y =
  Columns 1 through 13
   129  139  112  119  145  133  152  117  145  136  115  137  134
  Columns 14 through 15
   141  157
n =
    15


Sxy =
  1.0931e+003
Sxx =
  1.1009e+003
Syy =
  2.5669e+003


Koeficient korelacije je enak:
ro =
    0.6502


Determinacijski koeficient je enak:
D =
    0.4228
```

9.4.6 Interval zaupanja za korelacijski koeficient

Cenilka *R* koeficienta korelacije za naključni vzorec iz dvorazsežne normalne populacije je precej zapletena. Zato se običajno pri določanju intervala zaupanja za parameter ρ in tudi pri testiranju hipoteze o ρ naslonimo na naslednjo statistiko [Jesenko]:

$$\hat{R} = \frac{1}{2} \cdot \ln \frac{1+R}{1-R} \tag{9.193}$$

Za to statistiko **približno** velja [Jesenko]:

$$\hat{R} \in N\left(E\left(\hat{R}\right), VAR\left(\hat{R}\right)\right) = N\left(\frac{1}{2} \cdot \ln \frac{1+\rho}{1-\rho}, \frac{1}{n-3}\right) \tag{9.194}$$

Odtod sledi, da velja:

$$\begin{aligned} Z &= \frac{\hat{R} - \hat{\rho}}{STD\left(\hat{R}\right)} = \frac{\frac{1}{2} \cdot \ln \frac{1+R}{1-R} - \frac{1}{2} \cdot \ln \frac{1+\rho}{1-\rho}}{\sqrt{\frac{1}{n-3}}} = \\ &= \frac{\sqrt{n-3}}{2} \cdot \left(\ln \frac{1+R}{1-R} - \ln \frac{1+\rho}{1-\rho} \right) = \frac{\sqrt{n-3}}{2} \cdot \left(\ln \frac{\frac{1+R}{1-R}}{\frac{1+\rho}{1-\rho}} \right) = \\ &= \frac{\sqrt{n-3}}{2} \cdot \ln \left(\frac{(1+R) \cdot (1-\rho)}{(1-R) \cdot (1+\rho)} \right) \in N(0,1) \end{aligned} \tag{9.195}$$

Statistiko Z uporabimo za test ničelne hipoteze pri ustrezni nasprotni hipotezi:

$$\begin{aligned} H_0 &: \rho = \rho_0 \\ H_1 &: \rho \neq \rho_0 \\ H_1 &: \rho < \rho_0 \\ H_1 &: \rho > \rho_0 \end{aligned} \tag{9.196}$$

Na osnovi te statistike lahko tudi zgradimo ustrezen interval zaupanja za parameter ρ:

$$P\left(-z_{\frac{\alpha}{2}} \le Z \le z_{\frac{\alpha}{2}}\right) = 1-\alpha \qquad (9.197)$$
$$P\left(-z_{\frac{\alpha}{2}} \le \frac{\sqrt{n-3}}{2}\cdot \ln\left(\frac{(1+R)\cdot(1-\rho)}{(1-R)\cdot(1+\rho)}\right) \le z_{\frac{\alpha}{2}}\right) = 1-\alpha$$

Po daljši izpeljavi dobimo naslednji interval zaupanja za parameter ρ [Jesenko]:

$$\rho \in \left[\frac{(1+R)-(1-R)\cdot e^{\frac{2\cdot z_{\alpha/2}}{\sqrt{n-3}}}}{(1+R)+(1-R)\cdot e^{\frac{2\cdot z_{\alpha/2}}{\sqrt{n-3}}}}, \frac{(1+R)-(1-R)\cdot e^{-\frac{2\cdot z_{\alpha/2}}{\sqrt{n-3}}}}{(1+R)+(1-R)\cdot e^{-\frac{2\cdot z_{\alpha/2}}{\sqrt{n-3}}}}\right] \qquad (9.198)$$

Primer 9.13.:

Z naključnim vzorcem velikosti n = 15 želimo ugotoviti, če obstaja pri ženskah določene starosti linearna odvisnost med srčnim utripom X in zgornjim krvnim tlakom Y. Predpostavljamo, da sta tako srčni utrip kot krvni tlak normalno porazdeljena. Podatki so prikazani na sliki 225 [Jesenko]. Izračunajte interval zaupanja za korelacijski koeficient. Napravite test hipoteze $H_0 : \rho = 0.70$ *pri nasprotni hipotezi* $H_1 : \rho < 0.70$ *(*$\alpha = 0.05$*).*

Upoštevamo rezultate od prej in tekst naloge:

$$\begin{aligned} S_{xx} &= 1100.9 \\ S_{xy} &= 1093.1 \\ S_{yy} &= 2566.9 \\ r &= \frac{S_{XY}}{\sqrt{S_{XX}\cdot S_{YY}}} = 0.6502 \\ n &= 15 \\ D = r^2 &= \frac{{S_{XY}}^2}{S_{XX}\cdot S_{YY}} = 0.6502^2 = 0.4228 \end{aligned} \qquad (9.199)$$

$$H_0 : \rho = r_0 = 0.70$$
$$H_1 : \rho < r_0 = 0.70$$

Kritična vrednost Z statistike za interval zaupanja je enaka:

$$z_{krit} = \left| z_{\frac{\alpha}{2}} \right| = \left| z_{\frac{0.05}{2}} \right| = 1.96 \tag{9.200}$$

Interval zaupanja za parameter ρ je enak:

$$\rho \in \left[\frac{(1+r)-(1-r)\cdot e^{\frac{2\cdot z_{\alpha/2}}{\sqrt{n-3}}}}{(1+r)+(1-r)\cdot e^{\frac{2\cdot z_{\alpha/2}}{\sqrt{n-3}}}}, \frac{(1+r)-(1-r)\cdot e^{-\frac{2\cdot z_{\alpha/2}}{\sqrt{n-3}}}}{(1+r)+(1-r)\cdot e^{-\frac{2\cdot z_{\alpha/2}}{\sqrt{n-3}}}} \right]$$
$$\rho \in \left[\frac{(1+0.6502)-(1-0.6502)\cdot e^{\frac{2\cdot 1.96}{\sqrt{12}}}}{(1+0.6502)+(1-0.6502)\cdot e^{\frac{2\cdot 1.96}{\sqrt{12}}}}, \frac{(1+0.6502)-(1-0.6502)\cdot e^{-\frac{2\cdot 1.96}{\sqrt{12}}}}{(1+0.6502)+(1-0.6502)\cdot e^{-\frac{2\cdot 1.96}{\sqrt{12}}}} \right] \tag{9.201}$$
$$\rho \in [0.2069, 0.872]$$

Kritična vrednost Z statistike za test hipotez je enaka:

$$z_{krit} = z_{\alpha} = z_{0.05} = -1.6449 \tag{9.202}$$

Vrednost Z statistike pri ničelni hipotezi je enaka:

$$\begin{aligned} z &= \frac{\sqrt{n-3}}{2} \cdot \ln\left(\frac{(1+r)\cdot(1-r_0)}{(1-r)\cdot(1+r_0)} \right) = \\ &= \frac{\sqrt{12}}{2} \cdot \ln\left(\frac{(1+0.6502)\cdot(1-0.7)}{(1-0.6502)\cdot(1+0.7)} \right) = -0.3174 \end{aligned} \tag{9.203}$$

Ker je $z > z_{krit}$, smo v območju zaupanja hipoteze, zato ničelne hipoteze ne moremo zavrniti. Torej lahko sklepamo, da je $\rho = 0.70$, kar pomeni, da je koeficient korelacije med srčnim utripom in zgornjim krvnim tlakom pri ženskah določene starosti enak 0.70.

Pomagali smo si z naslednjim programom v Matlabu:

```
% koefkor1.m

clear
clc
close all

x = [77 68 55 76 79 66 85 62 81 72 75 67 78 87 81]
y = [129 139 112 119 145 133 152 117 145 136 115 137 134 141 157]

r0 = 0.70
alfa = 0.05
n = length(x)

sx = sum(x);
sy = sum(y);
sx2 = x*x';
sy2 = y*y';
sxy = x*y';
Sxy = sxy - sx*sy/n
Sxx = sx2 - sx^2/n
Syy = sy2 - sy^2/n

disp('Koeficient korelacije je enak (točkasta ocena):')
r = Sxy/sqrt(Sxx*Syy)

disp('kritična vrednost Z statistike za interval zaupanja je:')
zkrit = abs(norminv(alfa/2,0,1))

disp('interval zaupanja za korelacijski koeficient je:')
Isp = (1 + r - (1 - r)*exp(2*zkrit/sqrt(n-3))) / (1 + r + (1 - r)*exp(2*zkrit/sqrt(n-3)))
Izg = (1 + r - (1 - r)*exp(-2*zkrit/sqrt(n-3))) / (1 + r + (1 - r)*exp(-2*zkrit/sqrt(n-
3)))
I = [Isp Izg]

disp('kritična vrednost Z statistike za test hipotez je:')
zkrit = norminv(alfa,0,1)

disp('vrednost z statistike pri ničelni hipotezi:')
```

```
z = (sqrt(n-3)/2)*log( (1+r)*(1-r0)/(1-r)/(1+r0) )


if z < zkrit
   disp('Zavrni ničelno hipotezo')
else
   disp('Sprejmi ničelno hipotezo')
end
```

Izpis komandnega okna je naslednji:

```
x =

  77   68   55   76   79   66   85   62   81   72   75   67   78   87   81

y =

  129  139  112  119  145  133  152  117  145  136  115  137  134  141  157

r0 =

   0.7000

alfa =

   0.0500

n =

   15

Sxy =

  1.0931e+003

Sxx =

  1.1009e+003

Syy =

  2.5669e+003

Koeficient korelacije je enak (točkasta ocena):

r =

   0.6502

kritična vrednost Z statistike za interval zaupanja je:

zkrit =

   1.9600

interval zaupanja za korelacijski koeficient je:

Isp =

   0.2069

Izg =

   0.8720

I =
```

```
    0.2069    0.8720

kritična vrednost Z statistike za test hipotez je:

zkrit =

   -1.6449

vrednost z statistike pri ničelni hipotezi:

z =

   -0.3174


Sprejmi ničelno hipotezo
```

9.4.7 Še nekaj primerov linearne regresije

Napovedovanje s pomočjo linearne regresije je pomembno tudi na področju operacijskih raziskav. Slika 227 prikazuje nekaj tipičnih primerov [Winston].

Odvisna spremenljivka	**Neodvisna spremenljivka**
Prodaja nekega proizvoda	Cena proizvoda
Prodaja avtomobilov	Obrestna mera
Skupni proizvodni stroški	Število proizvedenih enot

Slika 227: Nekaj tipičnih primerov, kjer bi želeli uvesti napovedovanje

V nadaljevanju si bomo pogledali nekaj takšnih primerov.

Primer 9.14.:

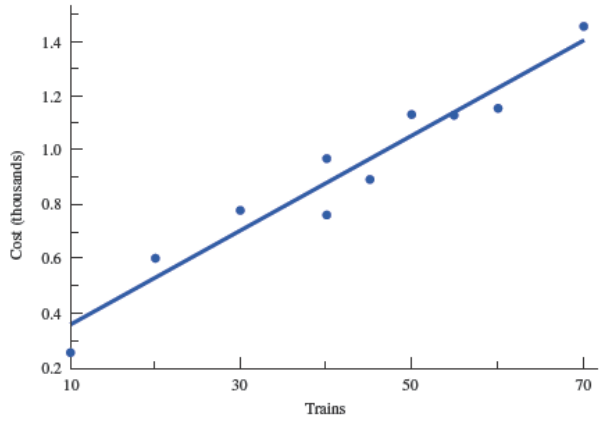
Tovarna igrač izdeluje vlake. Tabela na sliki 228 prikazuje stroške proizvodnje vlakov (cost of producing trains) v odvisnosti od števila izdelanih vlakov (trains produced) v odbobju zadnjih 10 tednov [Winston].

TABLE 9
Weekly Cost Data on Trains

Week	Trains Produced	Cost of Producing Trains
1	10	$257.40
2	20	$601.60
3	30	$782.00
4	40	$765.40
5	45	$895.50
6	50	$1,133.00
7	60	$1,152.80
8	55	$1,132.70
9	70	$1,459.20
10	40	$970.10

Slika 228: Stroški proizvodnje vlakov (cost of producing trains) v odvisnosti od števila izdelanih vlakov (trains produced) v odbobju zadnjih 10 tednov [Winston].

Če narišemo razsevni diagram in ocenjeno regresijsko premico, dobimo sliko 229.



Slika 229: Razsevni diagram in ocenjena regresijska premica [Winston]

Ocenite parametra regresijske premice. Določite koeficient določenosti in koeficient korelacije. Izračunajte še vse ostale relevantne statistične značilke! Komentirajte fenomen homoskedastičnosti, heteroskedastičnosti in avtokorelacije! $(\alpha = 0.05)$. *Kolišna je napoved za stroške, če bi se proizvedlo 75 vlakov?*

Imamo:

$$\begin{aligned} \alpha &= 0.05 \\ n &= 10 \\ x_{nov} &= 75 \end{aligned} \tag{9.204}$$

Dobimo:

$$\begin{aligned} S_{xx} &= 3010 \\ S_{xy} &= 53757 \\ S_{yy} &= 1.0218 \cdot 10^6 \\ \overline{x} &= 42 \\ \overline{y} &= 914.97 \end{aligned} \tag{9.205}$$

Ocenjena parametra sta:

$$\begin{aligned} \hat{b} &= \frac{S_{xy}}{S_{xx}} = \ldots = 17.8593 \\ \hat{a} &= \overline{y} - \hat{b} \cdot \overline{x} = \ldots = 164.8779 \end{aligned} \tag{9.206}$$

Kritična vrednost t statistike je:

$$t_{krit} = t\left(\frac{\alpha}{2}, n-2\right) = t\left(0.025, 8\right) = 2.3060 \tag{9.207}$$

Vrednost s_e je enaka:

$$s_\varepsilon = \sqrt{\frac{1}{n-2} \cdot \left(S_{yy} - \hat{b} \cdot S_{xy}\right)} = \ldots = 87.8246 \tag{9.208}$$

Interval zaupanja za regresijski parameter *b* je:

$$\begin{aligned} b &\in \left(\hat{b} - t_{\frac{\alpha}{2}, n-2} \cdot \frac{s_\varepsilon}{\sqrt{S_{xx}}}, \hat{b} + t_{\frac{\alpha}{2}, n-2} \cdot \frac{s_\varepsilon}{\sqrt{S_{xx}}} \right) \\ b &\in \left(14.1679, 21.5508\right) \end{aligned} \tag{9.209}$$

Interval zaupanja za regresijski parameter *a* je:

$$a \in \left(\hat{a} - t_{\frac{\alpha}{2},n-2} \cdot \sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)} \cdot s_\varepsilon, \hat{a} + t_{\frac{\alpha}{2},n-2} \cdot \sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)} \cdot s_\varepsilon \right) \tag{9.210}$$
$$a \in (-2.8685, 332.6243)$$

Ocenjena vrednost odvisne spremenljivke in **<u>približno 95%</u>** interval zaupanja za **<u>vrednost</u>** stroškov, če bi izdelali $x_{nov} = 75\, vlakov$ (ta vrednost ni bila uporabljena pri ocenjevanju parametrov), sta enaka:

$$\hat{\bar{y}}_{nov} = \hat{a} + \hat{b} \cdot x_{nov} = 1504.3 \tag{9.211}$$

$$y_{nov} \in \left(\hat{\bar{y}}_{nov} - t_{\frac{\alpha}{2},n-2} \cdot s_e, \hat{\bar{y}}_{nov} + t_{\frac{\alpha}{2},n-2} \cdot s_e \right) \tag{9.212}$$
$$y_{nov} \in (1301.8, 1706.9)$$

Kot vidimo, smo pri izračunu (9.212) vzeli poenostavljen izračun, kar je sicer **<u>pogosta praksa na področju operacijskih raziskav</u>**. Če pa smo bolj dosledni, bi morali vzeti:

$$y_{nov} \in \left(\hat{\bar{y}}_{nov} - t_{\frac{\alpha}{2},n-2} \cdot s_e \cdot \sqrt{\left(1 + \frac{1}{n} + \frac{\left(x_{nov} - \bar{x} \right)^2}{S_{xx}} \right)}, \hat{\bar{y}}_{nov} + t_{\frac{\alpha}{2},n-2} \cdot s_e \cdot \sqrt{\left(1 + \frac{1}{n} + \frac{\left(x_{nov} - \bar{x} \right)^2}{S_{xx}} \right)} \right) \tag{9.213}$$
$$y_{nov} \in (1259.5, 1749.2)$$

V tem primeru dobimo seveda širši interval zaupanja, kar je bolj pravilno, verjetnost, da je $y_{nov} \in (1259.5, 1749.2)$, pa je **<u>res enaka točno 95%.</u>**

Začetna varianca je enaka:

$$s_Y{}^2 = \frac{S_{yy}}{n-1} = 1.1353 \cdot 10^5 \tag{9.214}$$

Nepojasnena varianca je enaka:

$$s_e^{\,2} = 87.8246^2 = 7713.2 \tag{9.215}$$

Pojasnena varianca je enaka:

$$s_{XY}^2 = s_Y^{\,2} - s_e^{\,2} = 1.1353 \cdot 10^5 - 7713.2 = 1.0582 \cdot 10^5 \tag{9.216}$$

Korelacijski koeficient je enak:

$$r = \frac{S_{XY}}{\sqrt{S_{XX} \cdot S_{YY}}} = 0.9693 \tag{9.217}$$

Determinacijski koeficient, izračunan na 1. način, je enak:

$$D = 1 - \frac{s_e^{\,2}}{s_Y^{\,2}} = 0.9321 \tag{9.218}$$

Determinacijski koeficient, izračunan na 2. način, je enak:

$$D = r^2 = \frac{S_{xy}^{\,2}}{S_{xx} \cdot S_{yy}} = 0.9396 \tag{9.219}$$

Faktor **SST (sum of squares total)** je enak:

$$SST = S_{yy} = 1.0218 \cdot 10^6 \tag{9.220}$$

Faktor **SSE (sum of squares error)** je enak:

$$SSE = (n-2) \cdot s_e^2 = 61705 \tag{9.221}$$

Faktor **SSR (sum of squares regression)** je enak:

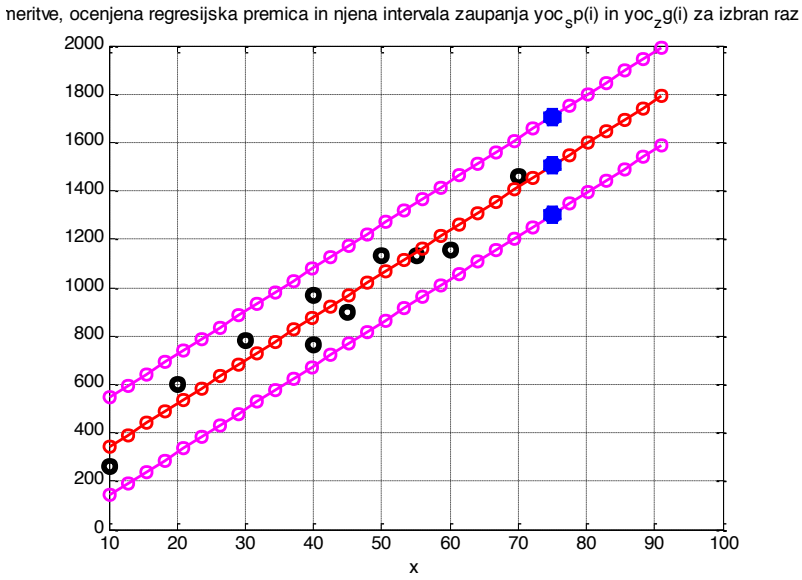
$$SSR = SST - SSE = 9.6006 \cdot 10^5 \tag{9.222}$$

Determinacijski koeficient, izračunan na 3. način, je enak:

$$D = \frac{SSR}{SST} = 0.9396 \tag{9.223}$$

Vidimo, da ta način da enak rezultat kot pri 2. načinu.

Slika 230 prikazuje meritve, ocenjeno regresijsko premico, ter **fiksni** interval zaupanja za napoved odvisne spremenljivke pri vrednosti neodvisne spremenljivke, ki ni bila upoštevana pri ocenjevanju parametrov. Z zvezdicami je označena tudi točkasta in intervalna ocena za napoved, če bi neodvisna spremeljivka zavzela vrednost 75 vlakov.

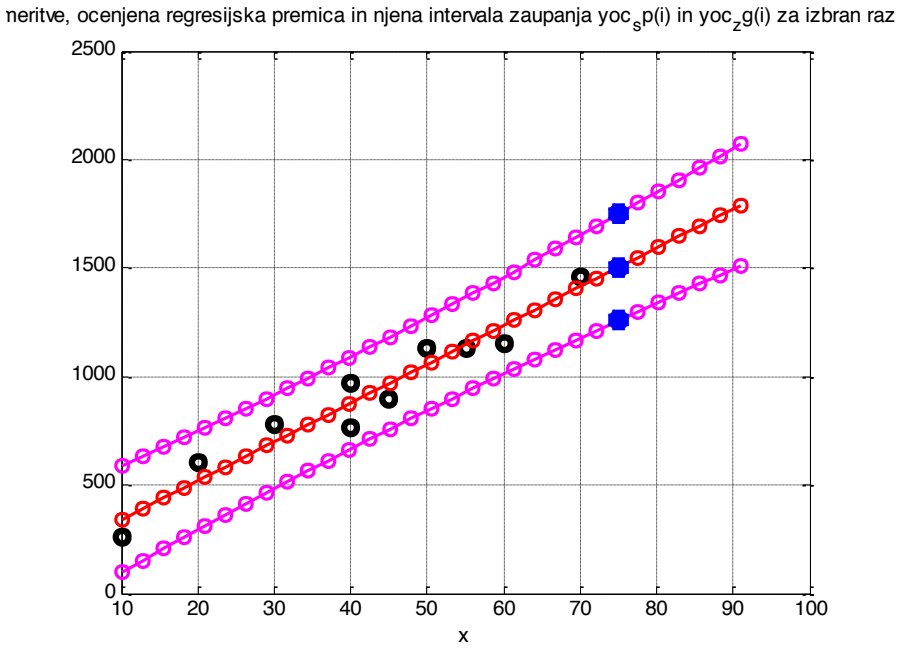


*Slika 230: Meritve, ocenjena regresijska premica, ter **fiksni** interval zaupanja za napoved odvisne spremenljivke pri vrednosti neodvisne spremenljivke, ki ni bila upoštevana pri ocenjevanju parametrov. (x=število izdelanih vlakov, y=stroški proizvodnje)*

Fiksni interval zaupanja za napoved odvisne spremenljivke pri vrednosti neodvisne spremenljivke, ki ni bila upoštevana pri ocenjevanju parametrov, smo tvorili na naslednji način:

$$
\begin{aligned}
& x' = \left[x_{novj} \right], \quad j = 1,\ldots,m \\
& \hat{\bar{y}}_{nov} = \hat{a} + \hat{b} \cdot x' \\
& y_{nov} \in \left(\hat{\bar{y}}_{nov} - t_{\frac{\alpha}{2},n-2} \cdot s_e , \hat{\bar{y}}_{nov} + t_{\frac{\alpha}{2},n-2} \cdot s_e \right) \\
& x_{nov} \in \left(x_{nov1}, x_{nov2}, \ldots \right)
\end{aligned}
\qquad (9.224)
$$

Slika 231 prikazuje meritve, ocenjeno regresijsko premico, ter **spremenljiv** interval zaupanja za napoved odvisne spremenljivke pri vrednosti neodvisne spremenljivke, ki ni bila upoštevana pri ocenjevanju parametrov. Z zvezdicami je označena tudi točkasta in intervalna ocena za napoved, če bi neodvisna spremeljivka zavzela vrednost 75 vlakov.



*Slika 231: Meritve, ocenjena regresijska premica, ter **spremenljiv** interval zaupanja za napoved odvisne spremenljivke pri vrednosti neodvisne spremenljivke, ki ni bila upoštevana pri ocenjevanju parametrov. (x=število izdelanih vlakov, y=stroški proizvodnje)*

Spremenljiv interval zaupanja za napoved odvisne spremenljivke pri vrednosti neodvisne spremenljivke, ki ni bila upoštevana pri ocenjevanju parametrov, smo tvorili na naslednji način:

$$
\begin{aligned}
&x' = \left[x_{novj} \right], \quad j = 1,\ldots,m \\
&\hat{\bar{y}}_{nov} = \hat{a} + \hat{b} \cdot x' \\
&y_{nov} \in \left(\hat{\bar{y}}_{nov} - t_{\frac{\alpha}{2},n-2} \cdot s_e \cdot \sqrt{\left(1 + \frac{1}{n} + \frac{(x' - \bar{x})^2}{S_{xx}} \right)}, \hat{\bar{y}}_{nov} + t_{\frac{\alpha}{2},n-2} \cdot s_e \cdot \sqrt{\left(1 + \frac{1}{n} + \frac{(x' - \bar{x})^2}{S_{xx}} \right)} \right) \\
&x_{nov} \in \left(x_{nov1}, x_{nov2}, \ldots \right)
\end{aligned}
\qquad (9.225)
$$

V tem primeru dobimo seveda širši interval zaupanja, kot pri fiksnem intervalu, kar je bolj pravilno, verjetnost, da y_{nov} pade v meje zaupanja, pa je **res enaka točno 95%.** Pri fiksnem intervalu pa je bila ta verjetnost **le približno 95%**.

Uporabili smo naslednji program v Matlabu (**intzaup2.m**):

```
% intzaup2.m
% default je example 9.5 na strani 170 Schaum OM knjige
clear
clc
close all

mode = input('mode za sort ascend-1,descend-2');
if mode == 1
   str = 'ascend';
else
   str = 'descend';
end
x = input('x=')
y = input('y=')
if length(x) == 0
    x = [15 9 40 20 25 25 15 35]
end
if length(y) == 0
    y = [6 4 16 6 13 9 10 16]
end

alfa = input('alfa=')
if length(alfa)==0
   alfa = 0.05
end
xnov = input('xnov=')
if length(xnov)==0
   xnov = 30
end

ch = input('Zelis poenostavljen interval zaupanja: 1-da,0-ne')
n = length(x)
sx = sum(x);
sy = sum(y);
sx2 = x*x';
sy2 = y*y';
sxy = x*y';
Sxy = sxy - sx*sy/n
Sxx = sx2 - sx^2/n
Syy = sy2 - sy^2/n
ysr = sy/n
xsr = sx/n

disp('boc=')
boc = Sxy/Sxx
disp('aoc=')
aoc = sy/n - boc*sx/n

disp('standardna ocena napake modela:')
se=sqrt((Syy-boc*Sxy)/(n-2)) % standardna ocena napake modela (dobljena glede na ucne vzorce)
```

```
% kriticna vrednost t statistike za intervale zaupanja za a, b in meje od yoc:
disp('kriticna vrednost t statistike:')
tkrit = abs(tinv(alfa/2,n-2))

disp('intervala zaupanja za parametra b in a:')
Ib = [boc-tkrit*se/sqrt(Sxx),boc+tkrit*se/sqrt(Sxx)]
Ia = [aoc-tkrit*se*sqrt(1/n+(sx/n)^2/Sxx),aoc+tkrit*se*sqrt(1/n+(sx/n)^2/Sxx)]

disp('Napoved pri x=xnov')
ynov = aoc + boc*xnov

disp('interval zaupanja za ynov pri x=xnov')

if ch == 0
   yoc_sp_nov = ynov-tkrit*se*sqrt(1+1/n+(xnov-(sx/n))^2/Sxx);
   yoc_zg_nov = ynov+tkrit*se*sqrt(1+1/n+(xnov-(sx/n))^2/Sxx);
else
   yoc_sp_nov = ynov-tkrit*se;
   yoc_zg_nov = ynov+tkrit*se;
end
I = [yoc_sp_nov yoc_zg_nov]

disp('skupna varianca:')
VARsk = Syy/(n-1)
disp('nepojasnena varianca:')
VARe = se^2
disp('pojasnena varianca:')
VARxy = VARsk - VARe

disp('determinacijski koeficient (1. način):')
D = 1 - VARe/VARsk

disp('korelacijski koeficient:')
ro = Sxy/sqrt(Sxx*Syy)

disp('determinacijski koeficient (2. način):')
D = ro^2

xmin = input('vnesi xmin, kjer naj bo interval zaupanja za meritve, ki niso bile pri ocenjevanju:')
xmax = input('vnesi xmax, kjer naj bo interval zaupanja za meritve, ki niso bile pri ocenjevanju:')
dx = input('vnesi korak dx:')

if length(xmin)==0
   xmin = min(x)
end
if length(xmax)==0
   xmax = max(x)
end
if length(dx)==0
   dx = (xmax-xmin)/30
end

x1 = [xmin:dx:xmax];
```

```
% generirajmo interval zaupanja za vse x1:


if ch == 0
   for i=1:length(x1)
       yoc(i)= aoc+boc*x1(i);
       yoc_sp(i) = yoc(i)-tkrit*se*sqrt(1+1/n+(x1(i)-(sx/n))^2/Sxx);
       yoc_zg(i) = yoc(i)+tkrit*se*sqrt(1+1/n+(x1(i)-(sx/n))^2/Sxx);
   end
else
  for i=1:length(x1)
       yoc(i)= aoc+boc*x1(i);
       yoc_sp(i) = yoc(i)-tkrit*se;
       yoc_zg(i) = yoc(i)+tkrit*se;
  end
end


plot(x,y,'ko','LineWidth',3)
hold on
plot(x1,yoc,'r','LineWidth',2)
plot(x1,yoc,'ro','LineWidth',2)


plot(sort(x1),sort(yoc_sp,str),'m','LineWidth',2)
plot(x1,yoc_sp,'mo','LineWidth',2)
plot(sort(x1),sort(yoc_zg,str),'m','LineWidth',2)
plot(x1,yoc_zg,'mo','LineWidth',2)
grid
title('meritve, ocenjena regresijska premica in njena intervala zaupanja yoc_sp(i) in yoc_zg(i) za izbran razpon')
xlabel('x')


plot(xnov,ynov,'b*','LineWidth',10)
plot(xnov,yoc_sp_nov,'b*','LineWidth',10)
plot(xnov,yoc_zg_nov,'b*','LineWidth',10)


SSE = (n-2)*VARe
SST = Syy
SSR = SST - SSE
disp('determinacijski koeficient (3. način):')
SSR/SST


disp('vhod     prava vrednost y (meritve)    ocenjena vrednost y      pogresek')
for i = 1:length(x)
    yocn(i)= aoc + boc*x(i);
    e(i) = y(i) - yocn(i);
end
[x' y' yocn' e']
figure
plot(x,e,'o','LineWidth',4)
hold on
grid
title('pogresek modela e(x) pri meritvah, uporabljenih pri ocenjevanju')
xlabel('razsevni diagram, x os')
d = axis;
axis([d(1) d(2) d(3)-1.5*abs(min(e)) d(4)+1.5*abs(max(e))])


disp('vsota pogreskov modela pri meritvah, uporabljenih pri ocenjevanju:')
sum(e)
disp('vsota x*e:')
x*e'
disp('vsota yoc*e')
yocn*e'
```

```
figure
plot(x,y,'o','LineWidth',4)
hold on
plot(x,yocn,'r+','LineWidth',8)
grid
title('izhod modela y in njegova ocena yoc pri meritvah, uporabljenih pri ocenjevanju')
xlabel('razsevni diagram, x os')
```

Primer izpisa komandnega okna je naslednji (npr. pri spremenljivem intervalu zaupanja):

```
x=[10 20 30 40 45 50 60 55 70 40]
x =
   10   20   30   40   45   50   60   55   70   40
y=[257.4 601.6 782 765.4 895.5 1133 1152.8 1132.7 1459.2 970.1]
y =
  1.0e+003 *
   0.2574   0.6016   0.7820   0.7654   0.8955   1.1330   1.1528   1.1327   1.4592   0.9701

alfa=0.05
alfa =
   0.0500

xnov=75
xnov =
   75

Zelis poenostavljen interval zaupanja: 1-da,0-ne0
ch =
    0
n =
   10

Sxy =
  5.3757e+004
Sxx =
      3010
Syy =
  1.0218e+006
ysr =
  914.9700
xsr =
    42
boc=
boc =
   17.8593
aoc=
aoc =
  164.8779

standardna ocena napake modela:
se =
   87.8246
```

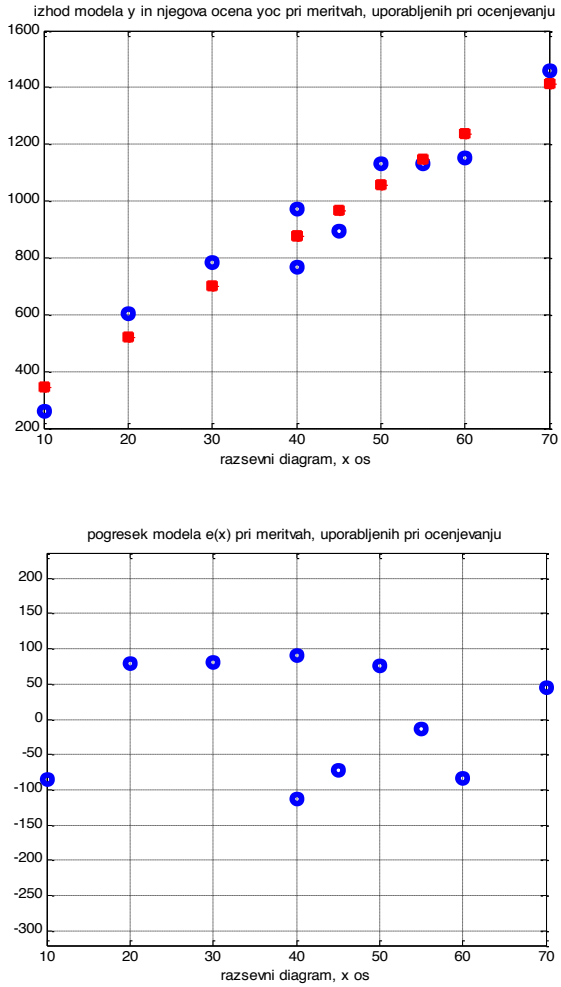
```
kriticna vrednost t statistike:
tkrit =
    2.3060

intervala zaupanja za parametra b in a:
Ib =
   14.1679   21.5508
Ia =
   -2.8685  332.6243

Napoved pri x=xnov
ynov =
   1.5043e+003
interval zaupanja za ynov pri x=xnov
I =
  1.0e+003 *
    1.2595    1.7492

skupna varianca:
VARsk =
  1.1353e+005
nepojasnena varianca:
VARe =
  7.7132e+003

pojasnena varianca:
VARxy =
  1.0582e+005
determinacijski koeficient (1. način):
D =
    0.9321

korelacijski koeficient:
ro =
    0.9693

determinacijski koeficient (2. način):
D =
    0.9396

vnesi xmin, kjer naj bo interval zaupanja za meritve, ki niso bile pri ocenjevanju:
xmin =
     []
vnesi xmax, kjer naj bo interval zaupanja za meritve, ki niso bile pri ocenjevanju:max(x)*1.3
xmax =
    91
vnesi korak dx:
dx =
     []
xmin =
    10
dx =
    2.7000
```

```
SSE =
  6.1705e+004
SST =
  1.0218e+006
SSR =
  9.6006e+005

determinacijski koeficient (3. način):
ans =
    0.9396

vhod     prava vrednost y (meritve)    ocenjena vrednost y      pogresek
ans =
  1.0e+003 *
    0.0100    0.2574    0.3435   -0.0861
    0.0200    0.6016    0.5221    0.0795
    0.0300    0.7820    0.7007    0.0813
    0.0400    0.7654    0.8793   -0.1139
    0.0450    0.8955    0.9685   -0.0730
    0.0500    1.1330    1.0578    0.0752
    0.0600    1.1528    1.2364   -0.0836
    0.0550    1.1327    1.1471   -0.0144
    0.0700    1.4592    1.4150    0.0442
    0.0400    0.9701    0.8793    0.0908

vsota pogreskov modela pri meritvah, uporabljenih pri ocenjevanju:
ans =
  1.2506e-012

vsota x*e:
ans =
  2.3647e-011
vsota yoc*e
ans =
  6.2573e-010
```

Kot vidimo iz programa intzaup2.m in izpisa komandnega okna, smo izračunali tudi:

$$\begin{aligned}\sum_{i=1}^{n}\varepsilon(i)&\approx 0\\ \sum_{i=1}^{n}\varepsilon(i)\cdot x(i)&\approx 0\\ \sum_{i=1}^{n}\varepsilon(i)\cdot \hat{\bar{y}}(i)&\approx 0\end{aligned} \tag{9.226}$$

Slika 232 prikazuje razsevni diagram izhoda modela y in njegove ocene $\hat{\bar{y}}$, ter pogreška modela glede na neodvisno spremenljivko za podatke pri ocenjevanju.



Slika 232: Razsevni diagram za $y(x), \hat{y}(x)$ in $\varepsilon(x)$ *(podatki pri ocenjevanju).*

Pogreške modela v numerični obliki prikazuje slika 233 [Winston].

Computations of Errors

x_i	y_i	$\hat{y}_i$	e_i
10	257.4	343.5	−86.1
20	601.6	522.1	79.5
30	782.0	700.7	81.3
40	765.4	879.3	−113.9
45	895.5	968.5	−73.0
50	1,133.0	1,057.8	75.2
60	1,152.8	1,236.4	−83.6
55	1,132.7	1,147.1	−14.4
70	1,459.2	1,415	44.2
40	970.1	879.3	90.8

Slika 233: Pogreški modela v numerični obliki [Winston]

S podobnim sklepanjem kot pri izrazih (9.212) in (9.213) lahko seveda zapišemo tudi naslednje [Winston]:

$$P\left(\hat{\bar{y}}_{nov} - s_e \le y_{nov} \le \hat{\bar{y}}_{nov} + s_e\right) \approx 0.68$$
$$P\left(\hat{\bar{y}}_{nov} - 2\cdot s_e \le y_{nov} \le \hat{\bar{y}}_{nov} + 2\cdot s_e\right) \approx 0.95$$
ter bolj točno:
$$P\left(\hat{\bar{y}}_{nov} - s_e \cdot \sqrt{\left(1+\frac{1}{n}+\frac{\left(x_{nov}-\bar{x}\right)^2}{S_{xx}}\right)} \le y_{nov} \le \hat{\bar{y}}_{nov} + s_e \cdot \sqrt{\left(1+\frac{1}{n}+\frac{\left(x_{nov}-\bar{x}\right)^2}{S_{xx}}\right)}\right) \approx 0.68 \quad (9.227)$$
$$P\left(\hat{\bar{y}}_{nov} - 2\cdot s_e \cdot \sqrt{\left(1+\frac{1}{n}+\frac{\left(x_{nov}-\bar{x}\right)^2}{S_{xx}}\right)} \le y_{nov} \le \hat{\bar{y}}_{nov} + 2\cdot s_e \cdot \sqrt{\left(1+\frac{1}{n}+\frac{\left(x_{nov}-\bar{x}\right)^2}{S_{xx}}\right)}\right) \approx 0.95$$

V nadaljevanju moramo na osnovi teksta obravnavanega primera komentirati tudi fenomen homoskedastičnosti, heteroskedastičnosti in avtokorelacije.

Homoskedastičnost

Homoskedastičnost pomeni, da varianca pogreška modela ne zavisi od vrednosti neodvisne spremenljivke (glej sliko 234) [Winston].

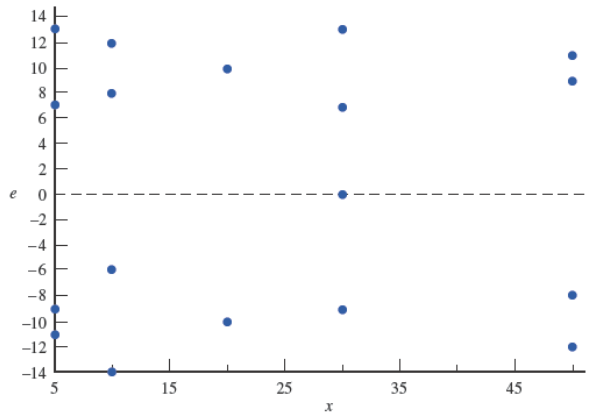
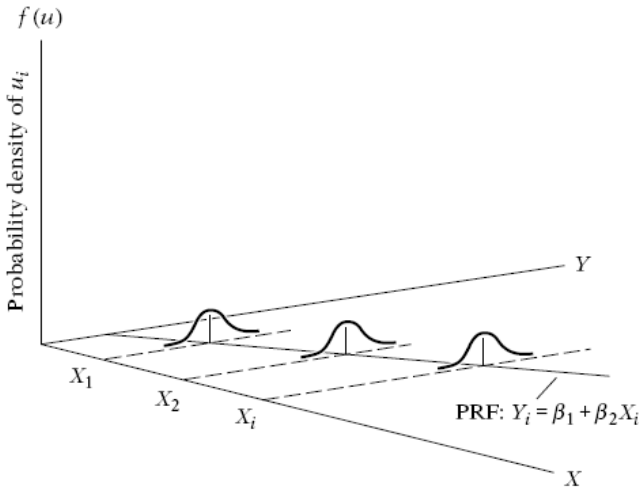


FIGURE 10
Homoscedasticity

Slika 234: Primer homoskedastičnosti (pogrešek e nima tendence spreminjanja velikosti glede na neodvisno spremenljivko x) [Winston]

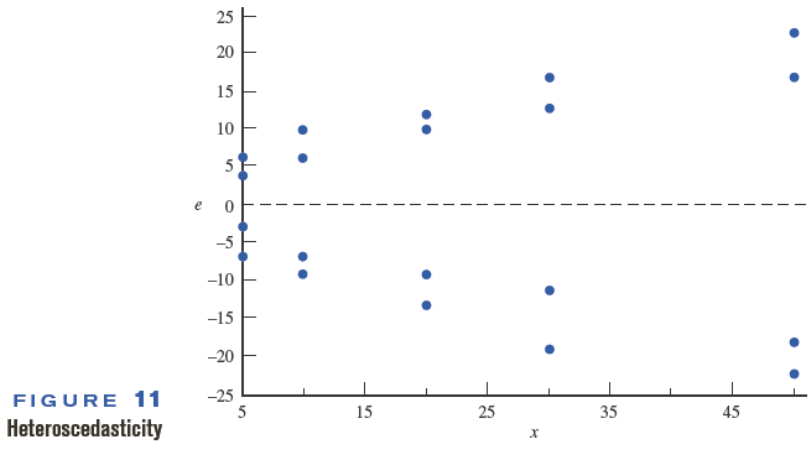
Morda še nazorneje je homoskedastičnost prikazana na sliki 235 [Gujarati].



Slika 235: Primer homoskedastičnosti (pogrešek u = e nima tendence spreminjanja velikosti glede na neodvisno spremenljivko x) [Gujarati]

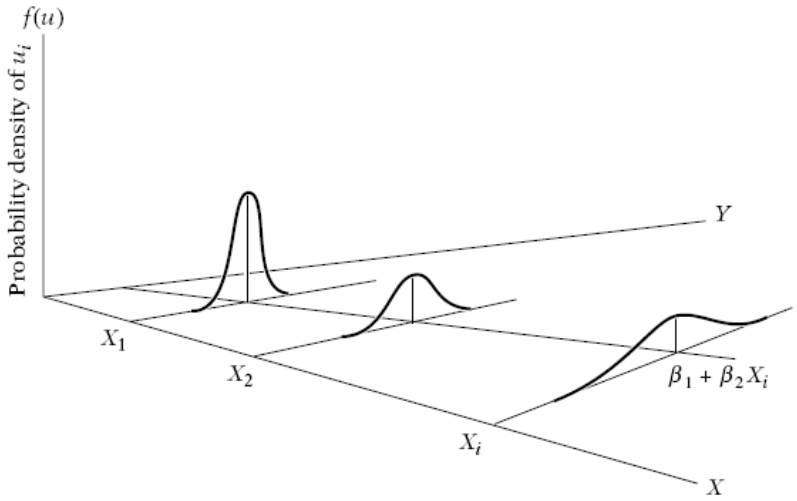
Heteroskedastičnost

Heteroskedastičnost pomeni, da varianca pogreška modela zavisi od vrednosti neodvisne spremenljivke (glej sliko 236) [Winston].



Slika 236: Primer heteroskedastičnosti (pogrešek e ima tendenco spreminjanja velikosti glede na neodvisno spremenljivko x) [Winston]

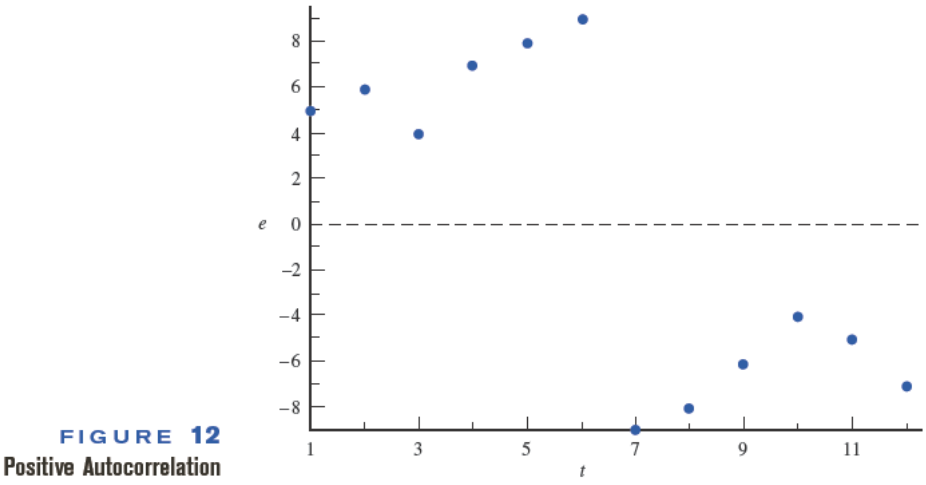
Morda še nazorneje je heteroskedastičnost prikazana na sliki 237 [Gujarati].



Slika 237: Primer heteroskedastičnosti (pogrešek u = e ima tendenco spreminjanja velikosti glede na neodvisno spremenljivko x) [Gujarati]

Avtokorelacija

Do nje prihaja pogosto tedaj, ko so podatki zbrani preko časa (kot v našem primeru). Pomeni pa, da pogreški niso neodvisni med seboj [Winston]. **Pozitivno avtokorelacijo** imamo tedaj, ko si npr. predznaki pogreškov sledijo v naslednjem vrstnem redu: + + + + + - - - - - -. Torej pozitivnemu pogrešku sledi pozitiven pogrešek, negativnemu pogrešku pa negativen pogrešek [Winston]. Torej imajo pogreški med seboj pozitivno linearno povezavo. Primer pozitivne avtokorelacije je prikazan na sliki 238 [Winston].



Slika 238: Primer pozitivne avtokorelacije [Winston]

Pri **negativni avtokorelaciji** pa si npr. predznaki pogreškov sledijo v naslednjem vrstnem redu: + - + - + - + - + - + -. To pomeni, da vsakemu pozitivnemu pogrešku sledi negativen pogrešek, in obratno [Winston]. Primer negativne avtokorelacije je prikazan na sliki 239 [Winston]. Torej imajo zaporedni pogreški med seboj negativno linearno povezavo in niso neodvisni.

Če pogreški modela ne kažejo niti lastnosti pozitivne, niti negativne avtokorelacije, so neodvisni med seboj, avtokorelacije pa ni.

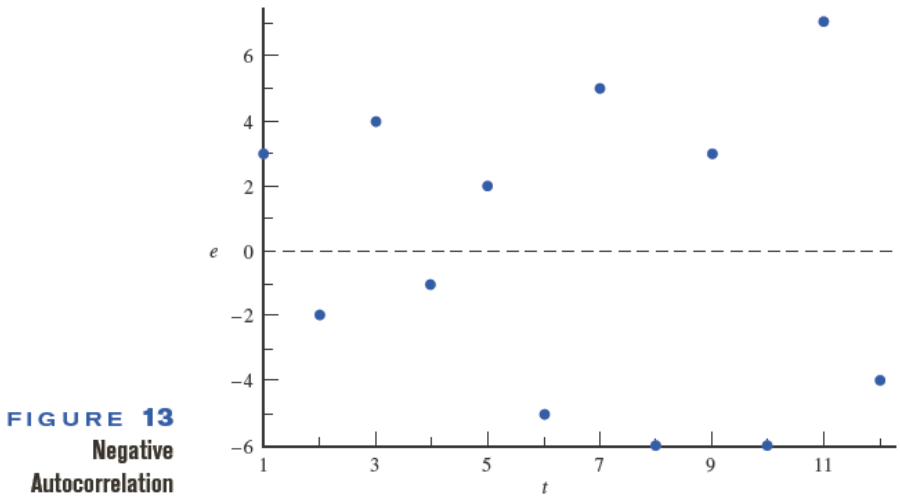


FIGURE 13
Negative
Autocorrelation

Slika 239: Primer negativne avtokorelacije [Winston]

Poglejmo si na sliki 240, kakšne rezultate bi za naš primer dal program v Excelu [Winston].

	A	B	C	D	E	F	G	H
1	trains	cost	yihatr	e				
2	10	257.4	9.32060032	248.0794	COST	REGRESSION		
3	20	601.6	18.6412006	582.958799	EXAMPLE			
4	30	782	27.961801	754.038199				
5	40	765.4	37.2824013	728.117599				
6	45	895.5	41.9427014	853.557299				
7	50	1133	46.6030016	1086.397				
8	60	1152.8	55.9236019	1096.8764				
9	55	1132.7	51.2633018	1081.4367				
10	70	1459.2	65.2442022	1393.9558				
11	40	970.1	37.2824013	932.817599				
12								
13								
14								
15		SUMMARY OUTPUT						
16								
17		*Regression Statistics*						
18		Multiple R	0.9693343					
19		R Square	0.9396089					
20		Adjusted R Squ	0.93206					
21		Standard Error	87.824643					
22		Observations	10					
23								
24		ANOVA						
25			*df*	*SS*	*MS*	*F*	*Significance F*	
26		Regression	1	960057.16	960057.16	124.46988 87	3.72837E-06	
27		Residual	8	61705.344	7713.168			
28		Total	9	1021762.5				
29								
30			*Coefficients*	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
31		Intercept	164.87791	72.743329	2.2665708	0.053174264	-2.8686199	332.62443
32		trains	17.859336	1.6007855	11.156607	3.72837E-06	14.167915 14	21.550756
33								
34								
35								
36								

Slika 240: Rezultati v Excelu [Winston]

S pozornim opazovanjem slike 240 bi ugotovili, da dobimo enake rezultate kot pri naših izračunih v Matlabu. Podrobnosti glede izračunov v Excelu si lahko bralec pogleda v [Winston].

Poglejmo si še, kako bi rezultate dobili s pomočjo standardne Matlabove funkcije **regress(...)**:

```
% regres_mnk.m

clear
clc
close all

x=[10 20 30 40 45 50 60 55 70 40]'
y=[257.4 601.6 782 765.4  895.5 1133 1152.8 1132.7 1459.2 970.1]'
X = [ones(size(x)) x]
alfa = 0.05

[b,bint,r,rint,stats] = regress(y,X,alfa);

disp('ocenjena parametra sta:')
b
```

```
disp('intervala zaupanja za ocenjena parametra:')
bint

disp('determinacijski koeficient:')
stats(1)

disp('nepojasnena varianca:')
stats(4)

disp('pogresek je:')
r

disp('interval zaupanja za pogresek')
rint

disp('x    spodnja meja pogreska    pogresek     zgornja meja pogreska:')
[x rint(:,1) r rint(:,2)]

disp('Standardna ocena napake modela je:')
se = sqrt(stats(4))

SSE = (length(x)-2)*stats(4)
```

Izpis v komandnem oknu bi imel obliko:

```
x =
   10
   20
   30
   40
   45
   50
   60
   55
   70
   40
y =
  1.0e+003 *
   0.2574
   0.6016
   0.7820
   0.7654
   0.8955
   1.1330
   1.1528
   1.1327
   1.4592
   0.9701


X =
    1   10
    1   20
    1   30
    1   40
    1   45
    1   50
    1   60
```

```
     1    55
     1    70
     1    40

alfa =
    0.0500

ocenjena parametra sta:
b =
  164.8779
   17.8593

intervala zaupanja za ocenjena parametra:
bint =
   -2.8685  332.6243
   14.1679   21.5508

determinacijski koeficient:
ans =
    0.9396

nepojasnena varianca:
ans =
  7.7132e+003

pogresek je:
r =
  -86.0713
   79.5354
   81.3420
 -113.8513
  -73.0480
   75.1553
  -83.6380
  -14.4414
   44.1686
   90.8487

interval zaupanja za pogresek
rint =
 -229.6438   57.5013
  -93.2213   252.2920
 -105.5241   268.2082
 -293.5140   65.8114
 -267.9688   121.8728
 -116.9393   267.2499
 -262.0422   94.7661
 -212.9295   184.0468
 -124.6403   212.9775
  -98.5073   280.2046

x    spodnja meja pogreska    pogresek    zgornja meja pogreska:
ans =
   10.0000  -229.6438   -86.0713    57.5013
```

```
   20.0000  -93.2213   79.5354   252.2920
   30.0000 -105.5241   81.3420   268.2082
   40.0000 -293.5140 -113.8513   65.8114
   45.0000 -267.9688  -73.0480   121.8728
   50.0000 -116.9393   75.1553   267.2499
   60.0000 -262.0422  -83.6380   94.7661
   55.0000 -212.9295  -14.4414   184.0468
   70.0000 -124.6403   44.1686   212.9775
   40.0000  -98.5073   90.8487   280.2046

Standardna ocena napake modela je:

se =
   87.8246

SSE =
   6.1705e+004
```

Primer 9.15.:

Podatki na sliki 241 prikazujejo povezavo med odvisno spremenljivko y (prihodki od prodaje - Sales revenue v mio dolarjev) in neodvisnima spremenljivkama (številom prodajnih referentov - Number of sales representatives x_1 in ceno proizvoda - Product price x_2) [Bronson].

Year	Sales Revenue ($ millions)	Number of Sales Representatives	Product Price ($)
1	1.2	25	0.95
2	1.5	25	0.93
3	2.0	25	0.92
4	3.5	26	0.90
5	4.1	28	0.87
6	5.6	28	0.85

Slika 241: Podatki naloge [Bronson]

a.) Napovejte prihodek prodaje za naslednje, 7. leto, če povečamo število referentov na $x_{nov} = 30$. *b.) Napovejte prihodek prodaje za naslednje, 7. leto, če spustimo prodajno ceno na* $x_{nov} = 0.82$ *dolarja. Primerjajte rezultate obeh regresij tako, da primerjate oba determinacijska koeficienta.*

Uporabimo program **intzaup2.m**. za regresijo pri a.). Dobimo naslednji izpis v komandnem oknu (pri fiksnem intervalu zaupanja za oceno napovedi):

```
x=[25 25 25 26 28 28]
x =
   25   25   25   26   28   28
y=[1.2 1.5 2 3.5 4.1 5.6]
y =
   1.2000   1.5000   2.0000   3.5000   4.1000   5.6000

alfa=
alfa =
    []
alfa =
   0.0500

xnov=30
xnov =
   30

Zelis poenostavljen interval zaupanja: 1-da,0-ne1
ch =
    1
n =
    6

Sxy =
   11.7167
Sxx =
   10.8333
Syy =
   14.7083
ysr =
    2.9833
xsr =
   26.1667

boc=
boc =
    1.0815
aoc=
aoc =
  -25.3169

standardna ocena napake modela:
se =
    0.7135

kriticna vrednost t statistike:
tkrit =
```

```
    2.7764

intervala zaupanja za parametra b in a:
Ib =
    0.4797    1.6834
Ia =
  -41.0865   -9.5473

Napoved pri x=xnov
ynov =
    7.1292
interval zaupanja za ynov pri x=xnov
I =
    5.1482    9.1102

skupna varianca:
VARsk =
    2.9417
nepojasnena varianca:
VARe =
    0.5091
pojasnena varianca:
VARxy =
    2.4326

determinacijski koeficient (1. način):
D =
    0.8269

korelacijski koeficient:
ro =
    0.9282
determinacijski koeficient (2. način):
D =
    0.8616

vnesi xmin, kjer naj bo interval zaupanja za meritve, ki niso bile pri ocenjevanju:min(x)*0.7
xmin =
   17.5000

vnesi xmax, kjer naj bo interval zaupanja za meritve, ki niso bile pri ocenjevanju:max(x)*1.3
xmax =
   36.4000
vnesi korak dx:
dx =
     []
dx =
    0.6300

SSE =
    2.0363
SST =
   14.7083
SSR =
```

```
   12.6720


determinacijski koeficient (3. način):
ans =
    0.8616


vhod     prava vrednost y (meritve)    ocenjena vrednost y      pogresek
ans =
   25.0000    1.2000    1.7215   -0.5215
   25.0000    1.5000    1.7215   -0.2215
   25.0000    2.0000    1.7215    0.2785
   26.0000    3.5000    2.8031    0.6969
   28.0000    4.1000    4.9662   -0.8662
   28.0000    5.6000    4.9662    0.6338


vsota pogreskov modela pri meritvah, uporabljenih pri ocenjevanju:
ans =
 -2.8866e-015
vsota x*e:
ans =
 -3.7659e-013
vsota yoc*e
ans =
 -3.3484e-013
```

Dobimo torej rezultate:

$$S_{xx} = 10.8333 \qquad S_{xy} = 11.7167 \qquad S_{yy} = 14.7083$$

$$\overline{x} = 26.1667 \qquad \overline{y} = 2.9833$$

$$\hat{b} = \frac{S_{xy}}{S_{xx}} = \ldots = 1.0815, \qquad \hat{a} = \overline{y} - \hat{b} \cdot \overline{x} = \ldots = \text{-}25.3169$$

$$s_\varepsilon = \sqrt{\frac{1}{n-2} \cdot \left(S_{yy} - \hat{b} \cdot S_{xy} \right)} = 0.7135$$

$$b \in \left(\hat{b} - t_{\frac{\alpha}{2},n-2} \cdot \frac{s_\varepsilon}{\sqrt{S_{xx}}}, \hat{b} + t_{\frac{\alpha}{2},n-2} \cdot \frac{s_\varepsilon}{\sqrt{S_{xx}}} \right) = \left(0.4797, 1.6834\right)$$

$$a \in \left(\hat{a} - t_{\frac{\alpha}{2},n-2} \cdot \sqrt{\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}} \right)} \cdot \ s_\varepsilon, \hat{a} + t_{\frac{\alpha}{2},n-2} \cdot \sqrt{\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}} \right)} \cdot \ s_\varepsilon \right) = \left(\text{-}41.0865, \text{-}9.5473\right)$$

$$\hat{\overline{y}}_{nov} = \hat{a} + \hat{b} \cdot x_{nov} = 7.1292$$

$$y_{nov} \in \left(\hat{\overline{y}}_{nov} - t_{\frac{\alpha}{2},n-2} \cdot s_e, \hat{\overline{y}}_{nov} + t_{\frac{\alpha}{2},n-2} \cdot s_e \right) = \left(5.1482, 9.1102\right)$$

$$y_{nov} \in \left(\hat{\overline{y}}_{nov} - t_{\frac{\alpha}{2},n-2} \cdot s_e \cdot \sqrt{\left(1 + \frac{1}{n} + \frac{\left(x_{nov} - \overline{x} \right)^2}{S_{xx}} \right)}, \hat{\overline{y}}_{nov} + t_{\frac{\alpha}{2},n-2} \cdot s_e \cdot \sqrt{\left(1 + \frac{1}{n} + \frac{\left(x_{nov} - \overline{x} \right)^2}{S_{xx}} \right)} \right) \quad (9.228)$$

$$y_{nov} \in \left(3.9826, 10.2759\right)$$

$$s_Y{}^2 = \frac{S_{yy}}{n-1} = 2.9417$$

$$s_e{}^2 = 0.7135^2 = 0.5091$$

$$s_{XY}^2 = s_Y{}^2 - s_e{}^2 = 2.4326$$

$$r = \frac{S_{XY}}{\sqrt{S_{XX} \cdot S_{YY}}} = 0.9282$$

$$D = 1 - \frac{s_e{}^2}{s_Y{}^2} = 0.8269$$

$$D = r^2 = \frac{S_{xy}{}^2}{S_{xx} \cdot S_{yy}} = 0.8616$$

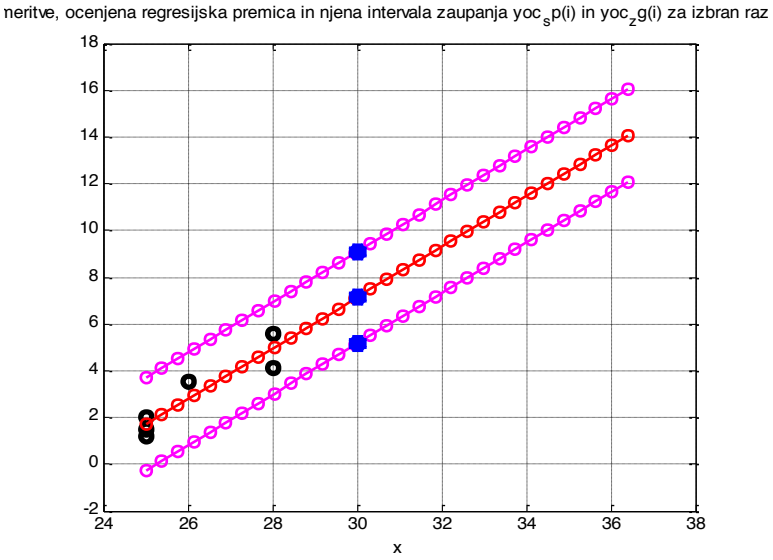
$$SST = S_{yy} = 14.7083$$

$$SSE = \left(n-2\right) \cdot s_e^2 = 2.0363$$

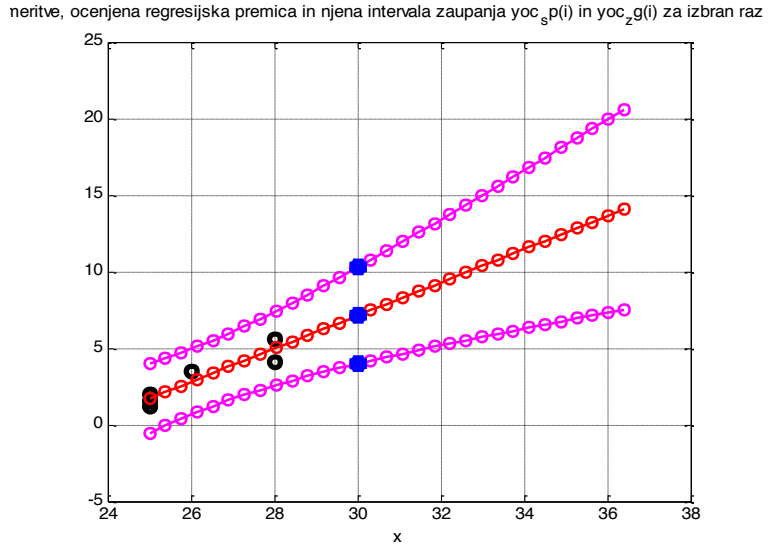
$$SSR = SST - SSE = 12.6720$$

$$D = \frac{SSR}{SST} = 0.8616$$

Sliki 242 in 243 prikazujeta meritve, ocenjeno regresijsko premico, ter **fiksni** oz. **spremenljiv** interval zaupanja za napoved odvisne spremenljivke pri vrednosti neodvisne spremenljivke, ki ni bila upoštevana pri ocenjevanju parametrov. Z zvezdicami je označena tudi točkasta in intervalna ocena za napoved, če bi neodvisna spremeljivka zavzela vrednost 30 referentov.

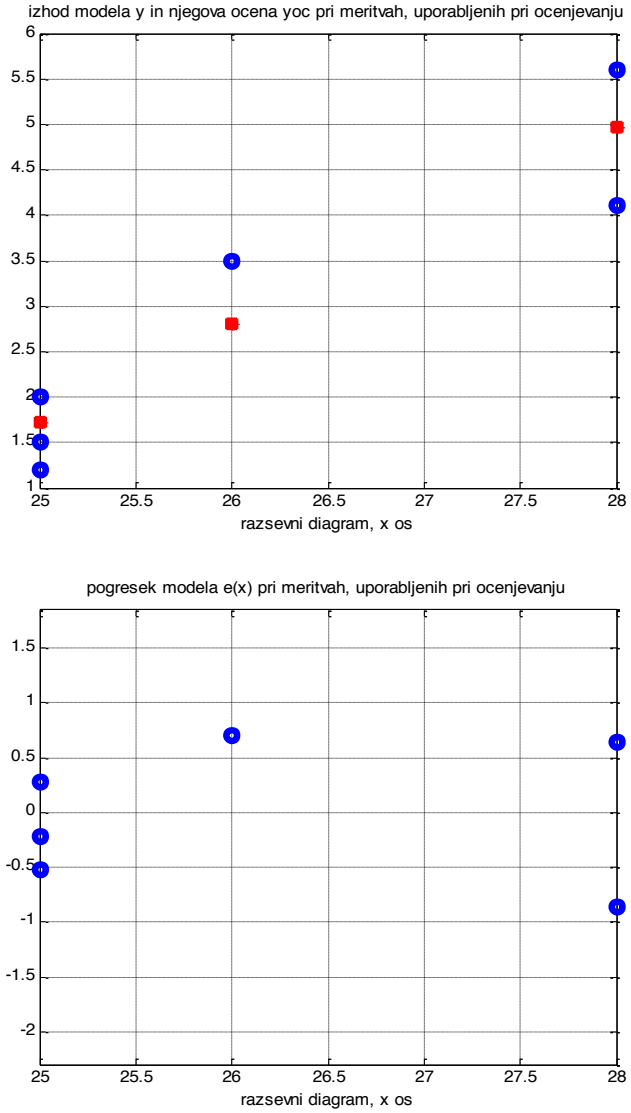


*Slika 242: Meritve, ocenjena regresijska premica, ter **fiksni** interval zaupanja za napoved odvisne spremenljivke pri vrednosti neodvisne spremenljivke, ki ni bila upoštevana pri ocenjevanju parametrov. (x=število referentov, y=prihodki prodaje)*



*Slika 243: Meritve, ocenjena regresijska premica, ter **spremenljiv** interval zaupanja za napoved odvisne spremenljivke pri vrednosti neodvisne spremenljivke, ki ni bila upoštevana pri ocenjevanju parametrov. (x=število referentov, y=prihodki prodaje)*

Slika 244 prikazuje razsevni diagram izhoda modela y in njegove ocene $\hat{y}$, ter pogreška modela glede na neodvisno spremenljivko za podatke pri ocenjevanju.



Slika 244: Razsevni diagram za $y(x), \hat{y}(x)$ in $\varepsilon(x)$ (podatki pri ocenjevanju).

Uporabimo program **intzaup2.m**. še za regresijo pri b.). Dobimo naslednji izpis v komandnem oknu (pri fiksnem intervalu zaupanja za oceno napovedi):

```
mode za sort ascend-1,descend-22
x=[0.95 0.93 0.92 0.9 0.87 0.85]
x =
    0.9500    0.9300    0.9200    0.9000    0.8700    0.8500
y=[1.2 1.5 2 3.5 4.1 5.6]
y =
    1.2000    1.5000    2.0000    3.5000    4.1000    5.6000

alfa=
alfa =
     []
alfa =
    0.0500

xnov=0.82
xnov =
    0.8200

Zelis poenostavljen interval zaupanja: 1-da,0-ne1
ch =
     1
n =
     6
Sxy =
   -0.3177
Sxx =
    0.0071
Syy =
   14.7083
ysr =
    2.9833
xsr =
    0.9033

boc=
boc =
  -44.5327
aoc=
aoc =
   43.2112

standardna ocena napake modela:
se =
    0.3748
kriticna vrednost t statistike:
tkrit =
```

```
    2.7764

intervala zaupanja za parametra b in a:
Ib =
  -56.8522  -32.2132
Ia =
   32.0745   54.3480

Napoved pri x=xnov
ynov =
    6.6944

interval zaupanja za ynov pri x=xnov
I =
    5.6539    7.7349

skupna varianca:
VARsk =
    2.9417
nepojasnena varianca:
VARe =
    0.1404
pojasnena varianca:
VARxy =
    2.8012

determinacijski koeficient (1. način):
D =
    0.9523

korelacijski koeficient:
ro =
   -0.9807

determinacijski koeficient (2. način):
D =
    0.9618

vnesi xmin, kjer naj bo interval zaupanja za meritve, ki niso bile pri ocenjevanju:min(x)*0.7
xmin =
    0.5950
vnesi xmax, kjer naj bo interval zaupanja za meritve, ki niso bile pri ocenjevanju:
xmax =
     []
vnesi korak dx:
dx =
     []
xmax =
    0.9500
dx =
    0.0118

SSE =
    0.5618
```

```
SST =
   14.7083
SSR =
   14.1466

determinacijski koeficient (3. način):
ans =
    0.9618

vhod     prava vrednost y (meritve)    ocenjena vrednost y      pogresek
ans =
    0.9500    1.2000    0.9051    0.2949
    0.9300    1.5000    1.7958   -0.2958
    0.9200    2.0000    2.2411   -0.2411
    0.9000    3.5000    3.1318    0.3682
    0.8700    4.1000    4.4678   -0.3678
    0.8500    5.6000    5.3584    0.2416

vsota pogreskov modela pri meritvah, uporabljenih pri ocenjevanju:
ans =
  5.7510e-014
vsota x*e:
ans =
 -1.8374e-014
vsota yoc*e
ans =
  3.3042e-012
```

Sedaj pa torej dobimo rezultate:

$S_{xx} = 0.0071 \qquad S_{xy} = \text{-}0.3177 \qquad S_{yy} = 14.7083$

$\bar{x} = 0.9033 \qquad \bar{y} = 2.9833$

$\hat{b} = \frac{S_{xy}}{S_{xx}} = ... = \text{-}44.5327, \qquad \hat{a} = \bar{y} - \hat{b} \cdot \bar{x} = ... = 43.2112$

$s_\varepsilon = \sqrt{\frac{1}{n-2} \cdot \left(S_{yy} - \hat{b} \cdot S_{xy} \right)} = 0.3748$

$b \in \left(\hat{b} - t_{\frac{\alpha}{2},n-2} \cdot \frac{s_\varepsilon}{\sqrt{S_{xx}}}, \hat{b} + t_{\frac{\alpha}{2},n-2} \cdot \frac{s_\varepsilon}{\sqrt{S_{xx}}} \right) = \left(\text{-}56.8522, \text{-}32.2132 \right)$

$a \in \left(\hat{a} - t_{\frac{\alpha}{2},n-2} \cdot \sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)} \cdot s_\varepsilon, \hat{a} + t_{\frac{\alpha}{2},n-2} \cdot \sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)} \cdot s_\varepsilon \right) = \left(32.0745, 54.3480 \right)$

$\hat{\bar{y}}_{nov} = \hat{a} + \hat{b} \cdot x_{nov} = 6.6944$

$y_{nov} \in \left(\hat{\bar{y}}_{nov} - t_{\frac{\alpha}{2},n-2} \cdot s_e, \hat{\bar{y}}_{nov} + t_{\frac{\alpha}{2},n-2} \cdot s_e \right) = \left(5.6539, 7.7349 \right)$

$$y_{nov} \in \left(\hat{\bar{y}}_{nov} - t_{\frac{\alpha}{2},n-2} \cdot s_e \cdot \sqrt{\left(1 + \frac{1}{n} + \frac{\left(x_{nov} - \bar{x} \right)^2}{S_{xx}} \right)}, \hat{\bar{y}}_{nov} + t_{\frac{\alpha}{2},n-2} \cdot s_e \cdot \sqrt{\left(1 + \frac{1}{n} + \frac{\left(x_{nov} - \bar{x} \right)^2}{S_{xx}} \right)} \right) \quad (9.229)$$

$y_{nov} \in \left(5.1722, 8.2166 \right)$

$s_Y{}^2 = \frac{S_{yy}}{n-1} = 2.9417$

$s_e{}^2 = 0.3748^2 = 0.1404$

$s_{XY}^2 = s_Y{}^2 - s_e{}^2 = 2.8012$

$r = \frac{S_{XY}}{\sqrt{S_{XX} \cdot S_{YY}}} = \text{-}0.9807$

$D = 1 - \frac{s_e{}^2}{s_Y{}^2} = 0.9523$

$D = r^2 = \frac{S_{xy}{}^2}{S_{xx} \cdot S_{yy}} = 0.9618$

$SST = S_{yy} = 14.7083$

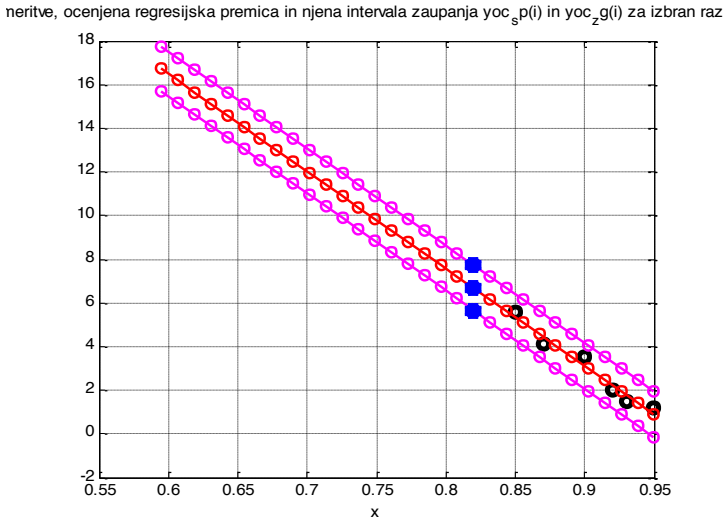
$SSE = (n-2) \cdot s_e^2 = 0.5618$

$SSR = SST - SSE = 14.1466$

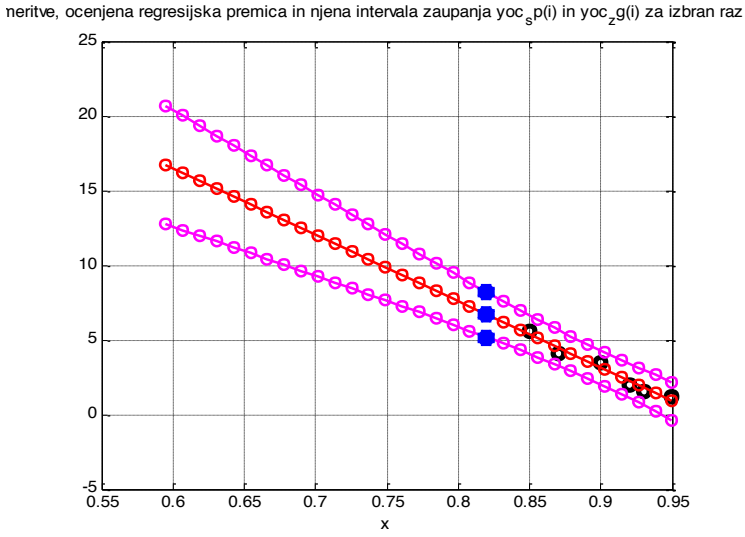
$D = \frac{SSR}{SST} = 0.9618$

Sliki 245 in 246 prikazujeta meritve, ocenjeno regresijsko premico, ter **fiksni** oz. **spremenljiv** interval zaupanja za napoved odvisne spremenljivke pri vrednosti neodvisne

spremenljivke, ki ni bila upoštevana pri ocenjevanju parametrov. Z zvezdicami je označena tudi točkasta in intervalna ocena za napoved, če bi neodvisna spremenljivka zavzela vrednost cene proizvoda 0.82.

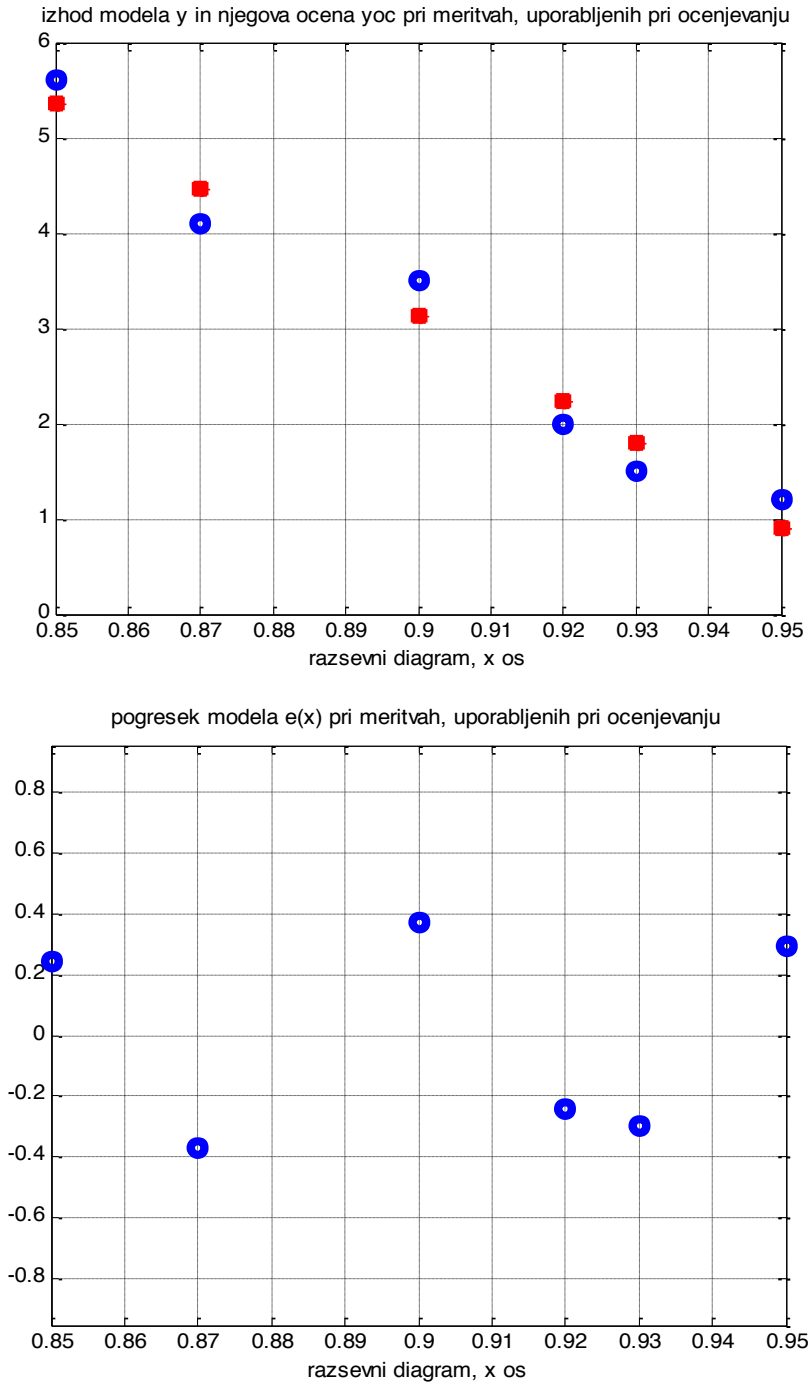


*Slika 245: Meritve, ocenjena regresijska premica, ter **fiksni** interval zaupanja za napoved odvisne spremenljivke pri vrednosti neodvisne spremenljivke, ki ni bila upoštevana pri ocenjevanju parametrov. (x=cena proizvoda, y=prihodki prodaje)*



*Slika 246: Meritve, ocenjena regresijska premica, ter **spremenljiv** interval zaupanja za napoved odvisne spremenljivke pri vrednosti neodvisne spremenljivke, ki ni bila upoštevana pri ocenjevanju parametrov. (x=cena proizvoda, y=prihodki prodaje)*

Slika 247 prikazuje razsevni diagram izhoda modela y in njegove ocene $\hat{\bar{y}}$, ter pogreška modela glede na neodvisno spremenljivko za podatke pri ocenjevanju.



Slika 247: Razsevni diagram za $y(x), \hat{\bar{y}}(x)$ in $\varepsilon(x)$ *(podatki pri ocenjevanju).*

Če primerjamo determinacijski koeficient v izrazih (9.228) in (9.229), vidimo, da je pri modelu za drugo regresijo determinacijski koeficient (0.9618) večji od tistega pri prvi

regresiji (0.8616). Torej je model pri drugi regresiji bolj zanesljiv in uporaben kot pa tisti pri prvi regresiji, saj ima zaradi močnejše linearne povezave med neodvisno in odvisno spremenljivko večjo izrazno moč v obliki regresijske premice [Bronson].

Primer 9.16.:

Podatki na sliki 248 prikazujejo celotne prihodke industrijske prodaje x in prihodke letne prodaje y podjetja ABC, ki proizvaja obleke za otroke [Bronson].

Year	Industry Sales (x) ($ millions)	ABC's Sales (y) ($ millions)
1	1103	105
2	1250	117
3	1097	110
4	955	101
5	945	97
6	903	92
7	1025	104
8	1170	116

Slika 248: celotni prihodki industrijske prodaje x in prihodki letne prodaje y podjetja ABC, ki proizvaja obleke za otroke [Bronson]

Če je ocena celotnih prihodkov industrijske prodaje za naslednje (9.) leto $x_{nov} = 1300$ *mio dolarjev, kakšna je potem napoved prihodkov letne prodaje* $\hat{\bar{y}}_{nov} \pm 95\%$ *interval zaupanja podjetja ABC? Izračunajte korelacijski koeficient in ga razložite. Koliko variabilnosti spremenljivke y je razloženo s spremenljivko x?*

Uporabimo program **intzaup2.m**. za regresijo. Dobimo naslednji izpis v komandnem oknu (pri fiksnem intervalu zaupanja za oceno napovedi):

```
mode za sort ascend-1,descend-21
x=[1103 1250 1097 955 945 903 1025 1170]
x =
  Columns 1 through 7
      1103       1250       1097        955        945        903       1025
  Column 8
      1170
y=[105 117 110 101 97 92 104 116]
y =
   105   117   110   101    97    92   104   116

alfa=
alfa =
     []
alfa =
    0.0500

xnov=1300
xnov =
      1300

Zelis poenostavljen interval zaupanja: 1-da,0-ne1
ch =
     1
n =
     8

Sxy =
       7099
Sxx =
      101414
Syy =
   539.5000
ysr =
   105.2500
xsr =
       1056

boc=
boc =
    0.0700
aoc=
aoc =
   31.3298

standardna ocena napake modela:
se =
    2.6636

kriticna vrednost t statistike:
tkrit =
    2.4469
```

```
intervala zaupanja za parametra b in a:
Ib =
    0.0495    0.0905
Ia =
    9.5949   53.0647

Napoved pri x=xnov
ynov =
  122.3300
interval zaupanja za ynov pri x=xnov
I =
  115.8125  128.8476

skupna varianca:
VARsk =
   77.0714
nepojasnena varianca:
VARe =
    7.0948
pojasnena varianca:
VARxy =
   69.9767

determinacijski koeficient (1. način):
D =
    0.9079

korelacijski koeficient:
ro =
    0.9597

determinacijski koeficient (2. način):
D =
    0.9211

vnesi xmin, kjer naj bo interval zaupanja za meritve, ki niso bile pri ocenjevanju:
xmin =
     []
vnesi xmax, kjer naj bo interval zaupanja za meritve, ki niso bile pri ocenjevanju:1350
xmax =
        1350
vnesi korak dx:
dx =
     []
xmin =
   903
dx =
   14.9000

SSE =
   42.5686
SST =
  539.5000
SSR =
```

```
  496.9314
determinacijski koeficient (3. način):
ans =
    0.9211


vhod     prava vrednost y (meritve)    ocenjena vrednost y      pogresek
ans =
  1.0e+003 *
    1.1030    0.1050    0.1085   -0.0035
    1.2500    0.1170    0.1188   -0.0018
    1.0970    0.1100    0.1081    0.0019
    0.9550    0.1010    0.0982    0.0028
    0.9450    0.0970    0.0975   -0.0005
    0.9030    0.0920    0.0945   -0.0025
    1.0250    0.1040    0.1031    0.0009
    1.1700    0.1160    0.1132    0.0028


vsota pogreskov modela pri meritvah, uporabljenih pri ocenjevanju:
ans =
  5.6843e-014
vsota x*e:
ans =
  5.9117e-011
vsota yoc*e
ans =
  5.9401e-012
```

Dobimo torej rezultate:

$$S_{xx} = 101414 \qquad S_{xy} = 7099 \qquad S_{yy} = 539.5000$$

$$\overline{x} = 1056 \qquad \overline{y} = 105.2500$$

$$\hat{b} = \frac{S_{xy}}{S_{xx}} = \ldots = 0.0700, \quad \hat{a} = \overline{y} - \hat{b} \cdot \overline{x} = \ldots = 31.3298$$

$$s_\varepsilon = \sqrt{\frac{1}{n-2} \cdot \left(S_{yy} - \hat{b} \cdot S_{xy} \right)} = 2.6636$$

$$b \in \left(\hat{b} - t_{\frac{\alpha}{2},n-2} \cdot \frac{s_\varepsilon}{\sqrt{S_{xx}}}, \hat{b} + t_{\frac{\alpha}{2},n-2} \cdot \frac{s_\varepsilon}{\sqrt{S_{xx}}} \right) = \left(0.0495, 0.0905 \right)$$

$$a \in \left(\hat{a} - t_{\frac{\alpha}{2},n-2} \cdot \sqrt{\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}} \right)} \cdot s_\varepsilon, \hat{a} + t_{\frac{\alpha}{2},n-2} \cdot \sqrt{\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}} \right)} \cdot s_\varepsilon \right) = \left(9.5949, 53.0647 \right)$$

$$\hat{\overline{y}}_{nov} = \hat{a} + \hat{b} \cdot x_{nov} = 122.3300$$

$$y_{nov} \in \left(\hat{\overline{y}}_{nov} - t_{\frac{\alpha}{2},n-2} \cdot s_e, \hat{\overline{y}}_{nov} + t_{\frac{\alpha}{2},n-2} \cdot s_e \right) = \left(115.8125, 128.8476 \right)$$

$$y_{nov} \in \left(\hat{\overline{y}}_{nov} - t_{\frac{\alpha}{2},n-2} \cdot s_e \cdot \sqrt{\left(1 + \frac{1}{n} + \frac{\left(x_{nov} - \overline{x} \right)^2}{S_{xx}} \right)}, \hat{\overline{y}}_{nov} + t_{\frac{\alpha}{2},n-2} \cdot s_e \cdot \sqrt{\left(1 + \frac{1}{n} + \frac{\left(x_{nov} - \overline{x} \right)^2}{S_{xx}} \right)} \right) \quad (9.230)$$

$$y_{nov} \in \left(113.8021, 130.8580 \right)$$

$$s_Y{}^2 = \frac{S_{yy}}{n-1} = 77.0714$$

$$s_e{}^2 = 2.6636^2 = 7.0948$$

$$s_{XY}^2 = s_Y{}^2 - s_e{}^2 = 69.9767$$

$$r = \frac{S_{XY}}{\sqrt{S_{XX} \cdot S_{YY}}} = 0.9597$$

$$D = 1 - \frac{s_e{}^2}{s_Y{}^2} = 0.9079$$

$$D = r^2 = \frac{S_{xy}{}^2}{S_{xx} \cdot S_{yy}} = 0.9211$$

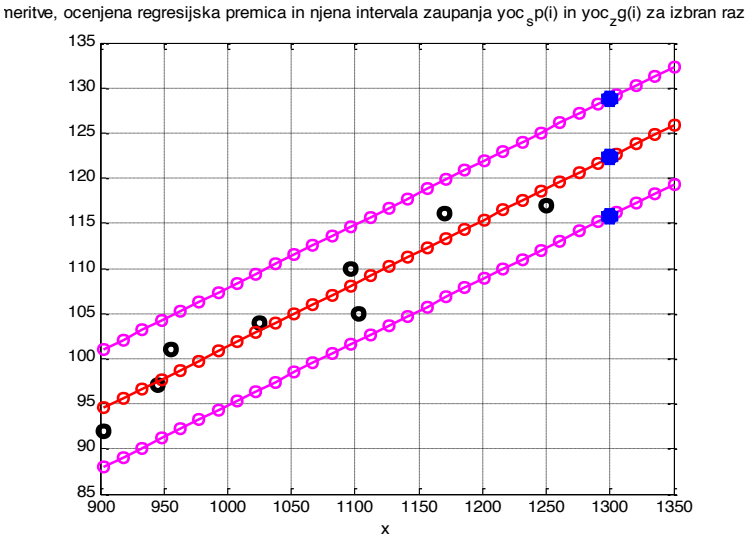
$$SST = S_{yy} = 539.5000$$

$$SSE = \left(n-2 \right) \cdot s_e^2 = 42.5686$$

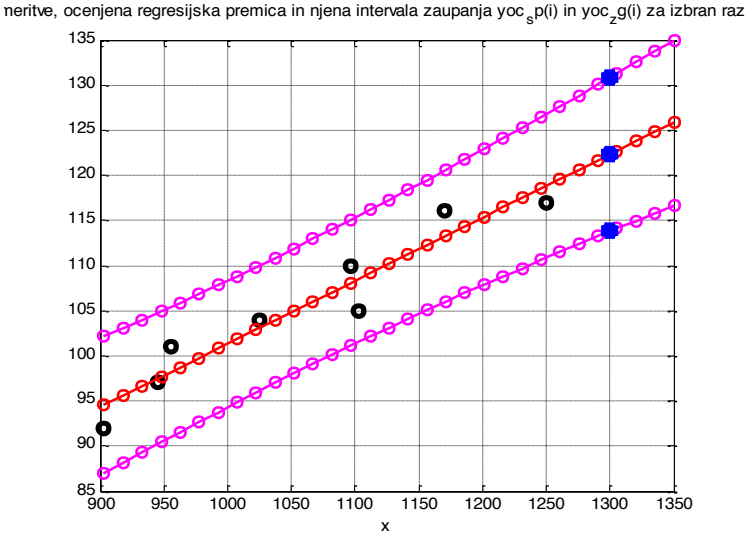
$$SSR = SST - SSE = 496.9314$$

$$D = \frac{SSR}{SST} = 0.9211$$

Sliki 249 in 250 prikazujeta meritve, ocenjeno regresijsko premico, ter **fiksni** oz. **spremenljiv** interval zaupanja za napoved odvisne spremenljivke pri vrednosti neodvisne spremenljivke, ki ni bila upoštevana pri ocenjevanju parametrov. Z zvezdicami je označena tudi točkasta in intervalna ocena za napoved $\hat{\bar{y}}_{nov}$.

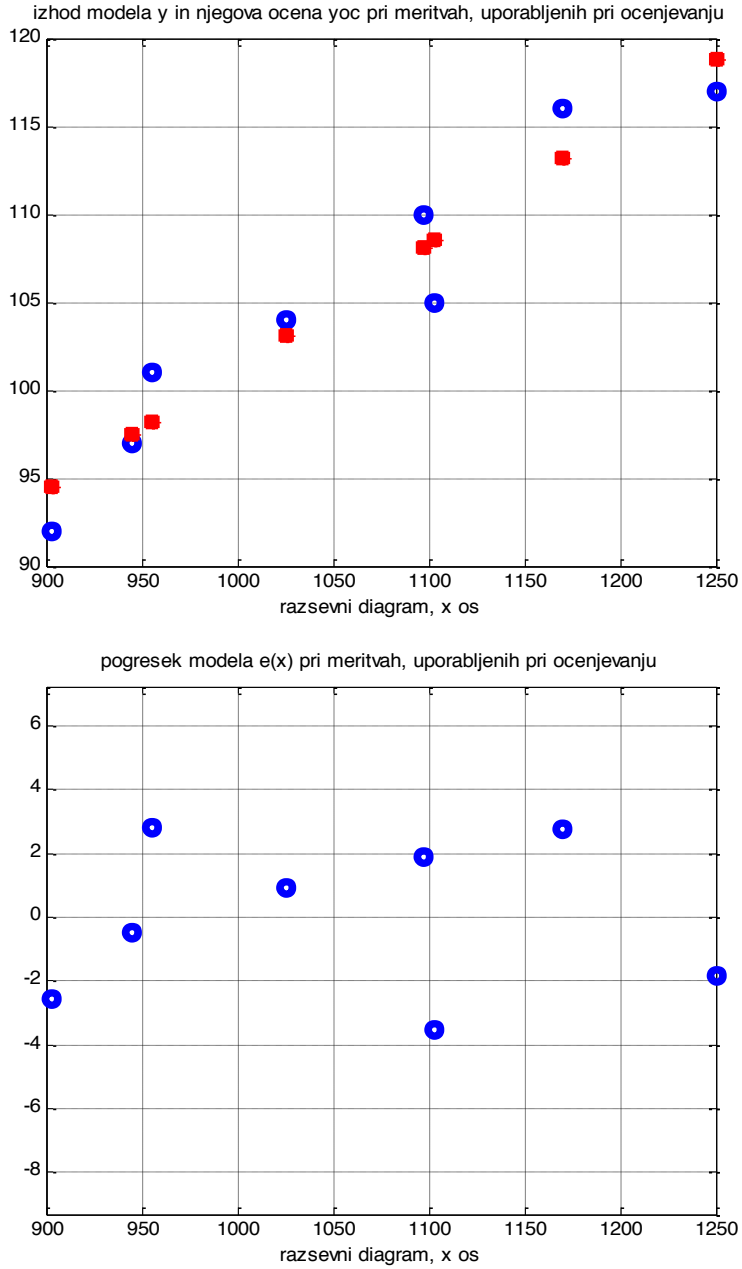


*Slika 249: Meritve, ocenjena regresijska premica, ter **fiksni** interval zaupanja za napoved odvisne spremenljivke pri vrednosti neodvisne spremenljivke, ki ni bila upoštevana pri ocenjevanju parametrov.*



*Slika 250: Meritve, ocenjena regresijska premica, ter **spremenljiv** interval zaupanja za napoved odvisne spremenljivke pri vrednosti neodvisne spremenljivke, ki ni bila upoštevana pri ocenjevanju parametrov.*

Slika 251 prikazuje razsevni diagram izhoda modela y in njegove ocene $\hat{y}$, ter pogreška modela glede na neodvisno spremenljivko za podatke pri ocenjevanju.



Slika 251: Razsevni diagram za $y(x), \hat{y}(x)$ in $\varepsilon(x)$ *(podatki pri ocenjevanju).*

Korelacijski koeficient z vrednostjo 0.9597 nakazuje zelo močno pozitivno linearno povezanost med industrijsko prodajo in prodajo ABC podjetja. Determinacijski koeficient z vrednostjo 0.9211 nakazuje, da je 92.1% variabilnosti prodaje ABC podjetja pojasneno z industrijsko prodajo.

Primer 9.17.:

Podatki na sliki 252 prikazujejo povezavo med odvisno spremenljivko y (prihodki od prodaje - Sales revenue v mio dolarjev) in neodvisnima spremenljivkama (številom prodajnih referentov - Number of sales representatives x_1 in ceno proizvoda - Product price x_2) [Bronson].

Year	Sales Revenue ($ millions)	Number of Sales Representatives	Product Price ($)
1	1.2	25	0.95
2	1.5	25	0.93
3	2.0	25	0.92
4	3.5	26	0.90
5	4.1	28	0.87
6	5.6	28	0.85

Slika 252: Podatki naloge [Bronson]

Napovejte vrednost prihodkov prodaje za naslednje (7.) leto, če tokrat uporabimo regresijo na osnovi časovnih vrst s časom t = x kot neodvisno spremenljivko $(x_{nov} = t_{nov} = 7)$.

Uporabimo program **intzaup2.m**. za regresijo. Dobimo naslednji izpis v komandnem oknu (pri fiksnem intervalu zaupanja za oceno napovedi):

```
mode za sort ascend-1,descend-21
x=1:1:6
x =
    1    2    3    4    5    6
y=[1.2 1.5 2 3.5 4.1 5.6]
y =
   1.2000    1.5000    2.0000    3.5000    4.1000    5.6000

alfa=
alfa =
    []
alfa =
   0.0500

xnov=7
```

```
xnov =
     7
Zelis poenostavljen interval zaupanja: 1-da,0-ne1
n =
     6


Sxy =
   15.6500
Sxx =
   17.5000
Syy =
   14.7083
ysr =
    2.9833
xsr =
    3.5000


boc=
boc =
    0.8943
aoc=
aoc =
   -0.1467
standardna ocena napake modela:
se =
    0.4221


kriticna vrednost t statistike:
tkrit =
    2.7764
intervala zaupanja za parametra b in a:
Ib =
    0.6141    1.1744
Ia =
   -1.2377    0.9444


Napoved pri x=xnov
ynov =
    6.1133
interval zaupanja za ynov pri x=xnov
I =
    4.9413    7.2853


skupna varianca:
VARsk =
    2.9417
nepojasnena varianca:
VARe =
    0.1782
pojasnena varianca:
VARxy =
    2.7635


determinacijski koeficient (1. način):
```

```
D =
    0.9394
korelacijski koeficient:
ro =
    0.9755

determinacijski koeficient (2. način):
D =
    0.9515

vnesi xmin, kjer naj bo interval zaupanja za meritve, ki niso bile pri ocenjevanju:
xmin =
     []
vnesi xmax, kjer naj bo interval zaupanja za meritve, ki niso bile pri ocenjevanju:9
xmax =
     9
vnesi korak dx:
dx =
     []
xmin =
     1
dx =
    0.2667

SSE =
    0.7128
SST =
   14.7083
SSR =
   13.9956

determinacijski koeficient (3. način):
ans =
    0.9515

vhod     prava vrednost y (meritve)    ocenjena vrednost y      pogresek
ans =
    1.0000    1.2000    0.7476    0.4524
    2.0000    1.5000    1.6419   -0.1419
    3.0000    2.0000    2.5362   -0.5362
    4.0000    3.5000    3.4305    0.0695
    5.0000    4.1000    4.3248   -0.2248
    6.0000    5.6000    5.2190    0.3810

vsota pogreskov modela pri meritvah, uporabljenih pri ocenjevanju:
ans =
 -1.4433e-015
vsota x*e:
ans =
 -6.6613e-015
vsota yoc*e
ans =
 -5.5511e-015
```

Dobimo torej rezultate:

$$S_{xx} = 17.5000 \qquad S_{xy} = 15.6500 \qquad S_{yy} = 14.7083$$

$$\overline{x} = 3.5000 \qquad \overline{y} = 2.9833$$

$$\hat{b} = \frac{S_{xy}}{S_{xx}} = \ldots = 0.8943, \qquad \hat{a} = \overline{y} - \hat{b} \cdot \overline{x} = \ldots = -0.1467$$

$$s_\varepsilon = \sqrt{\frac{1}{n-2} \cdot \left(S_{yy} - \hat{b} \cdot S_{xy} \right)} = 0.4221$$

$$b \in \left(\hat{b} - t_{\frac{\alpha}{2},n-2} \cdot \frac{s_\varepsilon}{\sqrt{S_{xx}}}, \hat{b} + t_{\frac{\alpha}{2},n-2} \cdot \frac{s_\varepsilon}{\sqrt{S_{xx}}} \right) = (0.6141, 1.1744)$$

$$a \in \left(\hat{a} - t_{\frac{\alpha}{2},n-2} \cdot \sqrt{\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}} \right)} \cdot s_\varepsilon, \hat{a} + t_{\frac{\alpha}{2},n-2} \cdot \sqrt{\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}} \right)} \cdot s_\varepsilon \right) = (-1.2377, 0.9444)$$

$$\hat{\hat{y}}_{nov} = \hat{a} + \hat{b} \cdot x_{nov} = 6.1133$$

$$y_{nov} \in \left(\hat{\hat{y}}_{nov} - t_{\frac{\alpha}{2},n-2} \cdot s_e, \hat{\hat{y}}_{nov} + t_{\frac{\alpha}{2},n-2} \cdot s_e \right) = (4.9413, 7.2853)$$

$$y_{nov} \in \left(\hat{\hat{y}}_{nov} - t_{\frac{\alpha}{2},n-2} \cdot s_e \cdot \sqrt{\left(1 + \frac{1}{n} + \frac{\left(x_{nov} - \overline{x} \right)^2}{S_{xx}} \right)}, \hat{\hat{y}}_{nov} + t_{\frac{\alpha}{2},n-2} \cdot s_e \cdot \sqrt{\left(1 + \frac{1}{n} + \frac{\left(x_{nov} - \overline{x} \right)^2}{S_{xx}} \right)} \right) \quad (9.231)$$

$$y_{nov} \in (4.5121, 7.7146)$$

$$s_Y{}^2 = \frac{S_{yy}}{n-1} = 2.9417$$

$$s_e{}^2 = 0.4221^2 = 0.1782$$

$$s_{XY}^2 = s_Y{}^2 - s_e{}^2 = 2.7635$$

$$r = \frac{S_{XY}}{\sqrt{S_{XX} \cdot S_{YY}}} = 0.9755$$

$$D = 1 - \frac{s_e{}^2}{s_Y{}^2} = 0.9394$$

$$D = r^2 = \frac{S_{xy}{}^2}{S_{xx} \cdot S_{yy}} = 0.9515$$

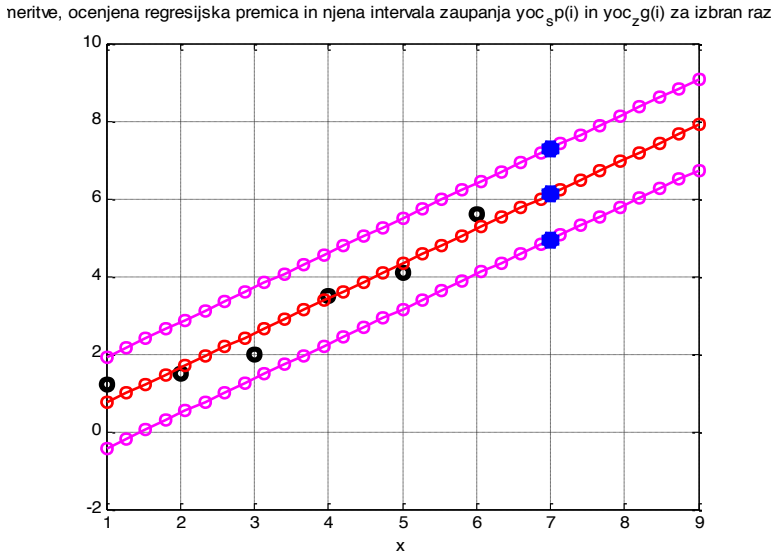
$$SST = S_{yy} = 14.7083$$

$$SSE = (n-2) \cdot s_e^2 = 0.7128$$

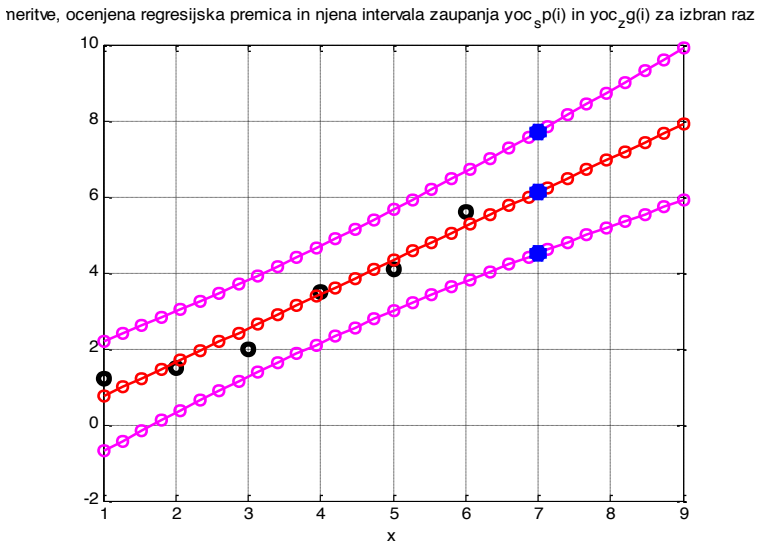
$$SSR = SST - SSE = 13.9956$$

$$D = \frac{SSR}{SST} = 0.9515$$

Sliki 253 in 254 prikazujeta meritve, ocenjeno regresijsko premico, ter **fiksni** oz. **spremenljiv** interval zaupanja za napoved odvisne spremenljivke pri vrednosti neodvisne spremenljivke, ki ni bila upoštevana pri ocenjevanju parametrov. Z zvezdicami je označena tudi točkasta in intervalna ocena za napoved $\hat{\bar{y}}_{nov}$.

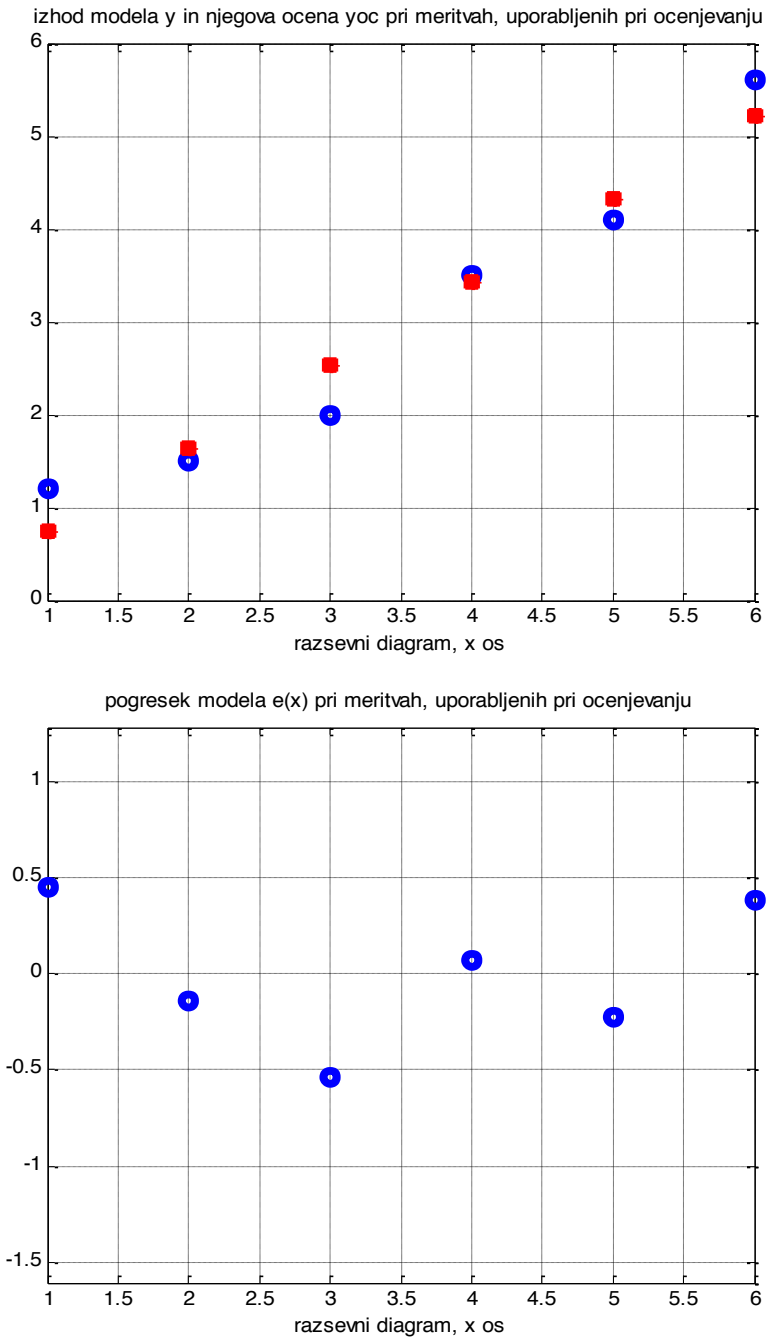


*Slika 253: Meritve, ocenjena regresijska premica, ter **fiksni** interval zaupanja za napoved odvisne spremenljivke pri vrednosti neodvisne spremenljivke, ki ni bila upoštevana pri ocenjevanju parametrov. (x=čas t, y(t)=prihodki prodaje)*



*Slika 254: Meritve, ocenjena regresijska premica, ter **spremenljiv** interval zaupanja za napoved odvisne spremenljivke pri vrednosti neodvisne spremenljivke, ki ni bila upoštevana pri ocenjevanju parametrov. (x=čas t, y(t)=prihodki prodaje)*

Slika 255 prikazuje razsevni diagram izhoda modela y in njegove ocene $\hat{y}$, ter pogreška modela glede na neodvisno spremenljivko za podatke pri ocenjevanju.



Slika 255: Razsevni diagram za $y(x), \hat{y}(x)$ in $\varepsilon(x)$ *(podatki pri ocenjevanju).*

Primer 9.18.:

Podatki na sliki 256 predstavljajo število zaposlenih v računalniškem sektorju (v tisočih) glede na čas v letih [Bronson].

Year	1	2	3	4	5	6	7	8	9	10	11	12	13
Employ	271	304	337	365	416	476	542	589	631	676	740	775	792

Slika 256: Število zaposlenih v računalniškem sektorju (v tisočih) glede na čas v letih [Bronson]

Poiščite regresijsko premico za regresijo na osnovi časovnih vrst. Poiščite napoved za število zaposlenih $\hat{\bar{y}}_{nov} \pm 95\%$ *interval zaupanja v naslednjem (14.) letu (*$x_{nov} = 14$*).*

Uporabimo program **intzaup2.m**. za regresijo. Dobimo naslednji izpis v komandnem oknu (pri fiksnem intervalu zaupanja za oceno napovedi):

```
mode za sort ascend-1,descend-21
x=1:1:13
x =
    1   2   3   4   5   6   7   8   9   10   11   12   13
y=|271 304 337 365 416 476 542 589 631 676 740 775 792|
y =
  271   304   337   365   416   476   542   589   631   676   740   775   792

alfa=
alfa =
    ||
alfa =
   0.0500

xnov=14
xnov =
   14

Zelis poenostavljen interval zaupanja: 1-da,0-ne1
ch =
    1

n =
```

```
   13


Sxy =
        8569
Sxx =
   182
Syy =
  4.0641e+005
ysr =
  531.8462
xsr =
     7


boc=
boc =
   47.0824
aoc=
aoc =
  202.2692


standardna ocena napake modela:
se =
   16.4053


kriticna vrednost t statistike:
tkrit =
    2.2010


intervala zaupanja za parametra b in a:
Ib =
   44.4059   49.7589
Ia =
  181.0253  223.5132


Napoved pri x=xnov
ynov =
  861.4231


interval zaupanja za ynov pri x=xnov
I =
  825.3154  897.5308


skupna varianca:
VARsk =
  3.3867e+004
nepojasnena varianca:
VARe =
  269.1324
pojasnena varianca:
VARxy =
  3.3598e+004


determinacijski koeficient (1. način):
D =
```

```
    0.9921


korelacijski koeficient:
ro =
    0.9964
determinacijski koeficient (2. način):
D =
    0.9927


vnesi xmin, kjer naj bo interval zaupanja za meritve, ki niso bile pri ocenjevanju:
xmin =
     []
vnesi xmax, kjer naj bo interval zaupanja za meritve, ki niso bile pri ocenjevanju:16
xmax =
    16
vnesi korak dx:
dx =
     []
xmin =
     1
dx =
    0.5000
SSE =
  2.9605e+003
SST =
  4.0641e+005
SSR =
  4.0345e+005
determinacijski koeficient (3. način):
ans =
    0.9927
vhod     prava vrednost y (meritve)    ocenjena vrednost y      pogresek
ans =
    1.0000  271.0000   249.3516     21.6484
    2.0000  304.0000   296.4341      7.5659
    3.0000  337.0000   343.5165     -6.5165
    4.0000  365.0000   390.5989    -25.5989
    5.0000  416.0000   437.6813    -21.6813
    6.0000  476.0000   484.7637     -8.7637
    7.0000  542.0000   531.8462     10.1538
    8.0000  589.0000   578.9286     10.0714
    9.0000  631.0000   626.0110      4.9890
   10.0000  676.0000  673.0934      2.9066
   11.0000  740.0000  720.1758     19.8242
   12.0000  775.0000  767.2582      7.7418
   13.0000  792.0000  814.3407    -22.3407


vsota pogreskov modela pri meritvah, uporabljenih pri ocenjevanju:
ans =
  5.6843e-013
vsota x*e:
ans =
  3.8085e-012
vsota yoc*e
```

```
ans =
  2.9740e-010
```

Dobimo torej rezultate:

$S_{xx} = 182 \qquad S_{xy} = 8569 \qquad S_{yy} = 4.0641\text{e+}005$

$\bar{x} = 7 \qquad \bar{y} = 531.8462$

$\hat{b} = \frac{S_{xy}}{S_{xx}} = \ldots = 47.0824, \qquad \hat{a} = \bar{y} - \hat{b} \cdot \bar{x} = \ldots = 202.2692$

$s_\varepsilon = \sqrt{\frac{1}{n-2} \cdot \left(S_{yy} - \hat{b} \cdot S_{xy} \right)} = 16.4053$

$b \in \left(\hat{b} - t_{\frac{\alpha}{2},n-2} \cdot \frac{s_\varepsilon}{\sqrt{S_{xx}}}, \hat{b} + t_{\frac{\alpha}{2},n-2} \cdot \frac{s_\varepsilon}{\sqrt{S_{xx}}} \right) = \left(44.4059, 49.7589 \right)$

$a \in \left(\hat{a} - t_{\frac{\alpha}{2},n-2} \cdot \sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)} \cdot s_\varepsilon, \hat{a} + t_{\frac{\alpha}{2},n-2} \cdot \sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)} \cdot s_\varepsilon \right) = \left(181.0253, 223.5132 \right)$

$\hat{\bar{y}}_{nov} = \hat{a} + \hat{b} \cdot x_{nov} = 861.4231$

$y_{nov} \in \left(\hat{\bar{y}}_{nov} - t_{\frac{\alpha}{2},n-2} \cdot s_e, \hat{\bar{y}}_{nov} + t_{\frac{\alpha}{2},n-2} \cdot s_e \right) = \left(825.3154, 897.5308 \right)$

$$y_{nov} \in \left(\hat{\bar{y}}_{nov} - t_{\frac{\alpha}{2},n-2} \cdot s_e \cdot \sqrt{\left(1 + \frac{1}{n} + \frac{\left(x_{nov} - \bar{x} \right)^2}{S_{xx}} \right)}, \hat{\bar{y}}_{nov} + t_{\frac{\alpha}{2},n-2} \cdot s_e \cdot \sqrt{\left(1 + \frac{1}{n} + \frac{\left(x_{nov} - \bar{x} \right)^2}{S_{xx}} \right)} \right) \quad (9.232)$$

$y_{nov} \in \left(819.5295, 903.3167 \right)$

$s_Y{}^2 = \frac{S_{yy}}{n-1} = 3.3867\text{e+}004$

$s_e{}^2 = 16.4053^2 = 269.1324$

$s_{XY}^2 = s_Y{}^2 - s_e{}^2 = 3.3598\text{e+}004$

$r = \frac{S_{XY}}{\sqrt{S_{XX} \cdot S_{YY}}} = 0.9964$

$D = 1 - \frac{s_e{}^2}{s_Y{}^2} = 0.9921$

$D = r^2 = \frac{S_{xy}{}^2}{S_{xx} \cdot S_{yy}} = 0.9927$

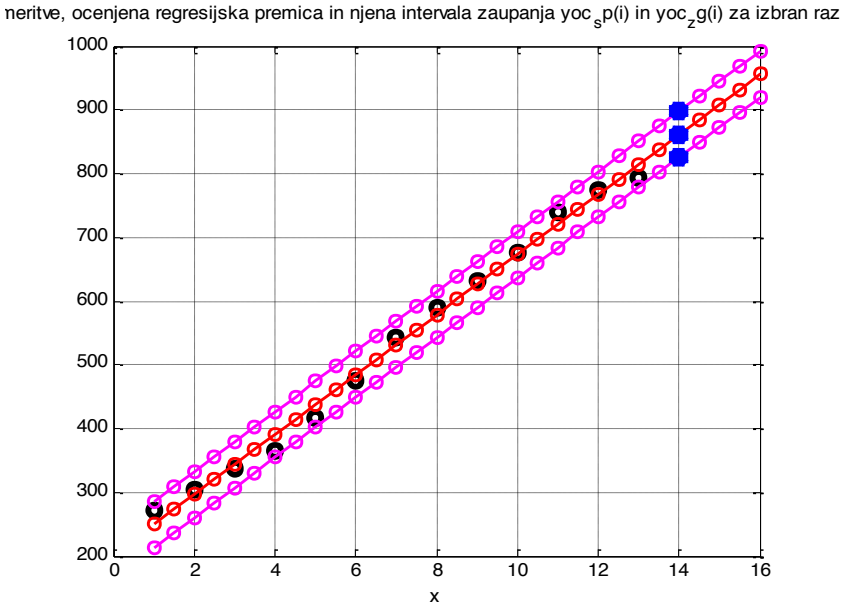
$SST = S_{yy} = 4.0641\text{e+}005$

$SSE = \left(n-2 \right) \cdot s_e^2 = 2.9605\text{e+}003$

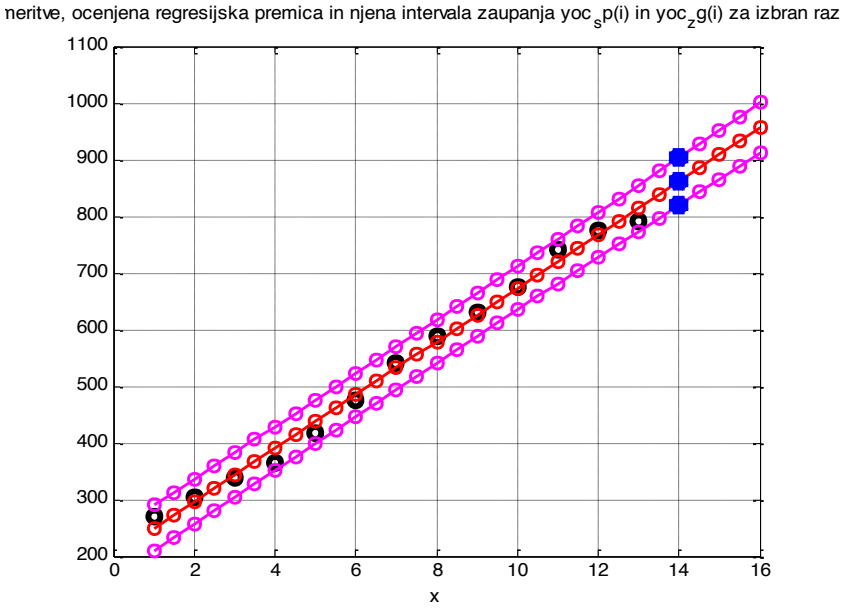
$SSR = SST - SSE = 4.0345\text{e+}005$

$D = \frac{SSR}{SST} = 0.9927$

Sliki 257 in 258 prikazujeta meritve, ocenjeno regresijsko premico, ter **fiksni** oz. **spremenljiv** interval zaupanja za napoved odvisne spremenljivke pri vrednosti neodvisne spremenljivke, ki ni bila upoštevana pri ocenjevanju parametrov. Z zvezdicami je označena tudi točkasta in intervalna ocena za napoved za 14. leto.

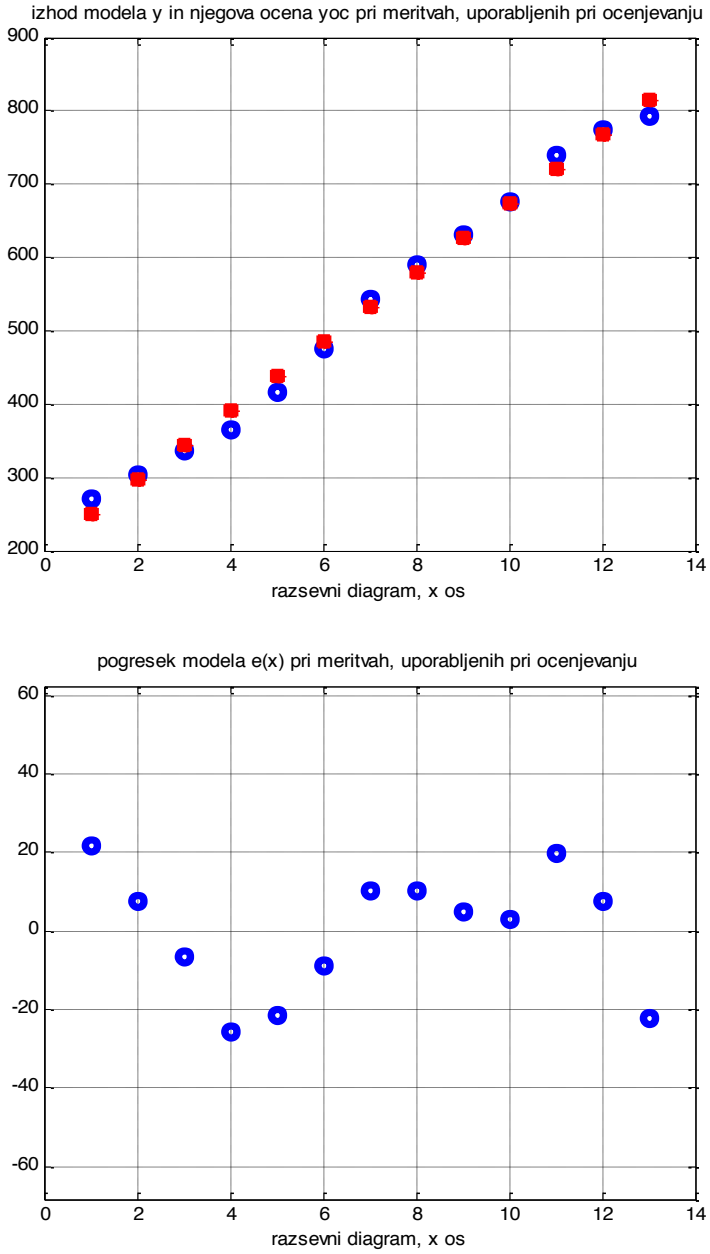


*Slika 257: Meritve, ocenjena regresijska premica, ter **fiksni** interval zaupanja za napoved odvisne spremenljivke pri vrednosti neodvisne spremenljivke, ki ni bila upoštevana pri ocenjevanju parametrov. (x=čas t, y(t)=število zaposlenih v tisočih)*



*Slika 258: Meritve, ocenjena regresijska premica, ter **spremenljiv** interval zaupanja za napoved odvisne spremenljivke pri vrednosti neodvisne spremenljivke, ki ni bila upoštevana pri ocenjevanju parametrov. (x=čas t, y(t)=število zaposlenih v tisočih)*

Slika 259 prikazuje razsevni diagram izhoda modela y in njegove ocene $\hat{y}$, ter pogreška modela glede na neodvisno spremenljivko za podatke pri ocenjevanju.



Slika 259: Razsevni diagram za $y(x), \hat{y}(x)$ in $\varepsilon(x)$ (podatki pri ocenjevanju).

Kot vidimo iz primerjave slik 257 in 258, razlike med fiksnim in spremenljivim intervalom zaupanja za oceno napovedi odvisne spremenljivke praktično skoraj ni. Razlog

je v tem, da sta korelacijski in determinacijski koeficient tako blizu 1, da praktično velja popolna linearna povezanost med neodvisno in odvisno spremenljivko.

Primer 9.19.:

Na sliki 260 so prikazani podatki časovne vrste povpraševanja po nekem proizvodu v 24 mesecih [Taha]. Poiščite napoved povpraševanja za $x_{nov} = 25$. *mesec.*

Mo, x_i	Demand, y_i	Mo, x_i	Demand, y_i
1	46	13	54
2	56	14	42
3	54	15	64
4	43	16	60
5	57	17	70
6	56	18	66
7	67	19	57
8	62	20	55
9	50	21	52
10	56	22	62
11	47	23	70
12	56	24	72

Slika 260: Podatki časovne vrste povpraševanja po nekem proizvodu v 24 mesecih [Taha].

Uporabimo program **intzaup2.m**. za regresijo. Dobimo naslednji izpis v komandnem oknu (pri fiksnem intervalu zaupanja za oceno napovedi):

```
mode za sort ascend-1,descend-21
x=1:1:24
```

```
x =
  Columns 1 through 14
    1    2    3    4    5    6    7    8    9   10   11   12   13   14
  Columns 15 through 24
   15   16   17   18   19   20   21   22   23   24
y=[46 56 54 43 57 56 67 62 50 56 47 56 54 42 64 60 70 66 57 55 52 62 70 72]



y =
  Columns 1 through 14
   46   56   54   43   57   56   67   62   50   56   47   56   54   42
  Columns 15 through 24
   64   60   70   66   57   55   52   62   70   72

alfa=
alfa =
     []
alfa =
    0.0500

xnov=25
xnov =
    25

Zelis poenostavljen interval zaupanja: 1-da,0-ne1
ch =
     1
n =
    24

Sxy =
   667
Sxx =
        1150
Syy =
  1.5925e+003
ysr =
   57.2500
xsr =
   12.5000

boc=
boc =
    0.5800
aoc=
aoc =
    50
standardna ocena napake modela:
se =
    7.4028

kriticna vrednost t statistike:
tkrit =
    2.0739
```

```
intervala zaupanja za parametra b in a:
Ib =
    0.1273    1.0327
Ia =
   43.5312   56.4688

Napoved pri x=xnov
ynov =
   64.5000
interval zaupanja za ynov pri x=xnov
I =
   49.1475   79.8525

skupna varianca:
VARsk =
   69.2391
nepojasnena varianca:
VARe =
   54.8018
pojasnena varianca:
VARxy =
   14.4373

determinacijski koeficient (1. način):
D =
    0.2085

korelacijski koeficient:
ro =
    0.4929

determinacijski koeficient (2. način):
D =
    0.2429

vnesi xmin, kjer naj bo interval zaupanja za meritve, ki niso bile pri ocenjevanju:
xmin =
     []
vnesi xmax, kjer naj bo interval zaupanja za meritve, ki niso bile pri ocenjevanju:29
xmax =
    29

vnesi korak dx:
dx =
     []
xmin =
     1
dx =
    0.9333

SSE =
  1.2056e+003
SST =
```

```
  1.5925e+003
SSR =
  386.8600

determinacijski koeficient (3. način):
ans =
    0.2429

vhod     prava vrednost y (meritve)    ocenjena vrednost y      pogresek
ans =
    1.0000   46.0000   50.5800   -4.5800
    2.0000   56.0000   51.1600    4.8400
    3.0000   54.0000   51.7400    2.2600
    4.0000   43.0000   52.3200   -9.3200
    5.0000   57.0000   52.9000    4.1000
    6.0000   56.0000   53.4800    2.5200
    7.0000   67.0000   54.0600   12.9400
    8.0000   62.0000   54.6400    7.3600
    9.0000   50.0000   55.2200   -5.2200
   10.0000   56.0000   55.8000    0.2000
   11.0000   47.0000   56.3800   -9.3800
   12.0000   56.0000   56.9600   -0.9600
   13.0000   54.0000   57.5400   -3.5400
   14.0000   42.0000   58.1200  -16.1200
   15.0000   64.0000   58.7000    5.3000
   16.0000   60.0000   59.2800    0.7200
   17.0000   70.0000   59.8600   10.1400
   18.0000   66.0000   60.4400    5.5600
   19.0000   57.0000   61.0200   -4.0200
   20.0000   55.0000   61.6000   -6.6000
   21.0000   52.0000   62.1800  -10.1800
   22.0000   62.0000   62.7600   -0.7600
   23.0000   70.0000   63.3400    6.6600
   24.0000   72.0000   63.9200    8.0800

vsota pogreskov modela pri meritvah, uporabljenih pri ocenjevanju:
ans =
  7.1054e-015
vsota x*e:
ans =
     0
vsota yoc*e
ans =
  6.8212e-013
```

Dobimo torej rezultate:

$S_{xx} = 1150 \qquad S_{xy} = 667 \qquad S_{yy} = 1.5925\text{e+}003$

$\overline{x} = 12.5 \qquad \overline{y} = 57.25$

$$\hat{b} = \frac{S_{xy}}{S_{xx}} = \ldots = 0.58, \qquad \hat{a} = \overline{y} - \hat{b} \cdot \overline{x} = \ldots = 50$$

$$s_\varepsilon = \sqrt{\frac{1}{n-2} \cdot \left(S_{yy} - \hat{b} \cdot S_{xy} \right)} = 7.4028$$

$$b \in \left(\hat{b} - t_{\frac{\alpha}{2},n-2} \cdot \frac{s_\varepsilon}{\sqrt{S_{xx}}}, \hat{b} + t_{\frac{\alpha}{2},n-2} \cdot \frac{s_\varepsilon}{\sqrt{S_{xx}}} \right) = (0.1273, 1.0327)$$

$$a \in \left(\hat{a} - t_{\frac{\alpha}{2},n-2} \cdot \sqrt{\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}} \right)} \cdot s_\varepsilon, \hat{a} + t_{\frac{\alpha}{2},n-2} \cdot \sqrt{\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}} \right)} \cdot s_\varepsilon \right) = (43.5312, 56.4688)$$

$$\hat{\hat{y}}_{nov} = \hat{a} + \hat{b} \cdot x_{nov} = 64.5$$

$$y_{nov} \in \left(\hat{\hat{y}}_{nov} - t_{\frac{\alpha}{2},n-2} \cdot s_e, \hat{\hat{y}}_{nov} + t_{\frac{\alpha}{2},n-2} \cdot s_e \right) = (49.1475, 79.8525)$$

$$y_{nov} \in \left(\hat{\hat{y}}_{nov} - t_{\frac{\alpha}{2},n-2} \cdot s_e \cdot \sqrt{\left(1 + \frac{1}{n} + \frac{\left(x_{nov} - \overline{x} \right)^2}{S_{xx}} \right)}, \hat{\hat{y}}_{nov} + t_{\frac{\alpha}{2},n-2} \cdot s_e \cdot \sqrt{\left(1 + \frac{1}{n} + \frac{\left(x_{nov} - \overline{x} \right)^2}{S_{xx}} \right)} \right) \quad (9.233)$$

$$y_{nov} \in (47.8403, 81.1597)$$

$$s_Y{}^2 = \frac{S_{yy}}{n-1} = 69.2391$$

$$s_e{}^2 = 7.4028^2 = 54.8018$$

$$s_{XY}^2 = s_Y{}^2 - s_e{}^2 = 14.4373$$

$$r = \frac{S_{XY}}{\sqrt{S_{XX} \cdot S_{YY}}} = 0.4929$$

$$D = 1 - \frac{s_e{}^2}{s_Y{}^2} = 0.2085$$

$$D = r^2 = \frac{S_{xy}{}^2}{S_{xx} \cdot S_{yy}} = 0.2429$$

$$SST = S_{yy} = 1.5925\text{e+}003$$

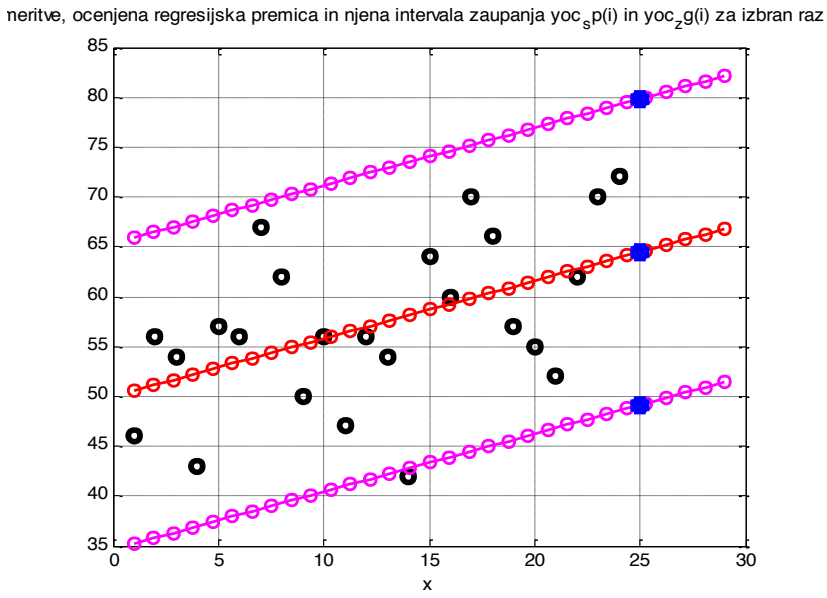
$$SSE = (n-2) \cdot s_e^2 = 1.2056\text{e+}003$$

$$SSR = SST - SSE = 386.8600$$

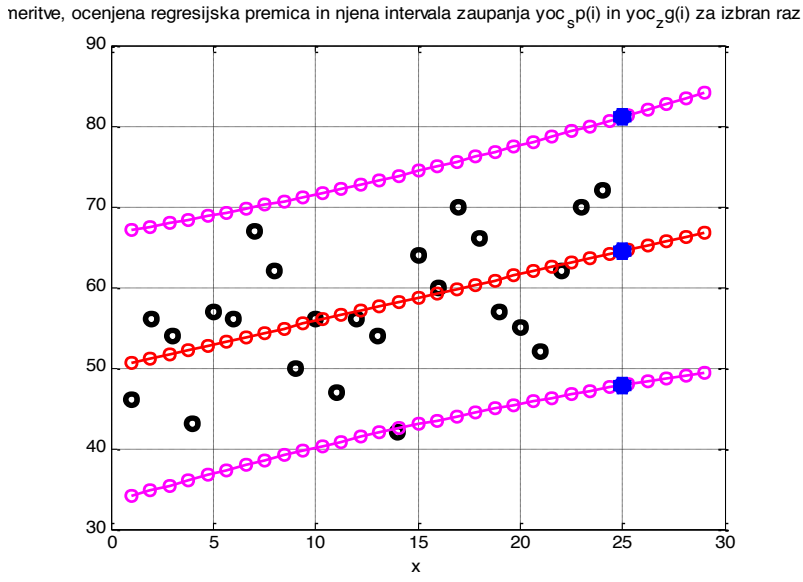
$$D = \frac{SSR}{SST} = 0.2429$$

Sliki 261 in 262 prikazujeta meritve, ocenjeno regresijsko premico, ter **fiksni** oz. **spremenljiv** interval zaupanja za napoved odvisne spremenljivke pri vrednosti neodvisne

spremenljivke, ki ni bila upoštevana pri ocenjevanju parametrov. Z zvezdicami je označena tudi točkasta in intervalna ocena za napoved za 25. mesec.

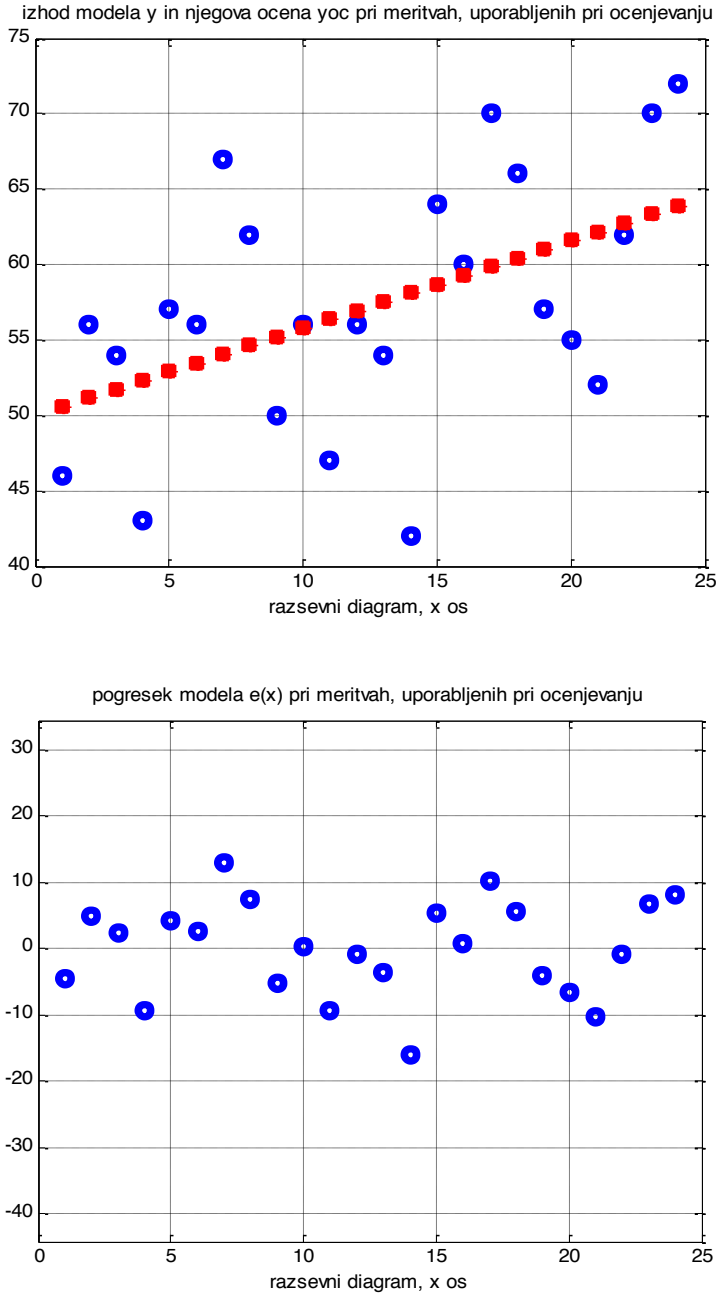


*Slika 261: Meritve, ocenjena regresijska premica, ter **fiksni** interval zaupanja za napoved odvisne spremenljivke pri vrednosti neodvisne spremenljivke, ki ni bila upoštevana pri ocenjevanju parametrov. (x=čas t v mesecih, y(t)=povpraševanje)*



*Slika 262: Meritve, ocenjena regresijska premica, ter **spremenljiv** interval zaupanja za napoved odvisne spremenljivke pri vrednosti neodvisne spremenljivke, ki ni bila upoštevana pri ocenjevanju parametrov. (x=čas t v mesecih, y(t)=povpraševanje)*

Slika 263 prikazuje razsevni diagram izhoda modela y in njegove ocene $\hat{y}$, ter pogreška modela glede na neodvisno spremenljivko za podatke pri ocenjevanju.



Slika 263: Razsevni diagram za $y(x), \hat{y}(x)$ in $\varepsilon(x)$ *(podatki pri ocenjevanju).*

Kot vidimo iz primera, je korelacijski koeficient relativno nizek. To nazazuje možnost, da dobljeni linearni model v obliki premice morda ni najbolj primeren za regresijo [Taha].

9.4.8 Testiranje kakovosti regresijske premice

V praksi je uporaba analize determinacijskega koeficienta zelo priljubljena, predvsem zaradi enostavne uporabe. Vendar se je potrebno zavedati, da je ta koeficient odvisen od velikosti vzorca in tipa aplikacije. Sprašujemo se, katera vrednost detrminacijskega koeficienta je še primerna za uporabo. Npr., nek kemik bo zadovoljen z regresijsko premico, če bo detrminacijski koeficient vsaj 0.99, po drugi strani pa bo nek psiholog zadovoljen že s koeficientom višjim od 0.7. Detrminacijski koeficient zna biti nevaren, če primerjamo med seboj primernost več različnih modelov. Namreč, če dodamo v model dodatni regresor, se zaradi tega lahko pomanjša SSE in zato poveča $D = \frac{SSR}{SST} = \frac{SSR}{SSR + SSE}$ (Glej izraze (9.221) do (9.223)). Pri tem pa smo z predeterminiranostjo modela umetno povečali koeficient D [Walpole].

Velikokrat problem analiziranja kvalitete ocenjene regresijske premice rešujemo z **analizo variance (ANOVA)**. To je procedura, kjer totalno variabilnost odvisne spremenljivke razbijemo na ustrezne komponente, ki so nato obravnavane na sistematičen način [Walpole].

Vsoto kvadratov pogreškov SSE lahko izrazimo na naslednji način:

$$
\begin{aligned}
&SSE = (n-2)\cdot s_e{}^2 = \sum_{i=1}^{n} \varepsilon^2(i) = \sum_{i=1}^{n} \left(y(i) - \hat{\bar{y}}(i)\right)^2 = \sum_{i=1}^{n} \left(y(i) - \hat{a} - \hat{b}\cdot x(i)\right)^2 = \\
&= \sum_{i=1}^{n} \left(y(i) - \left(\bar{y} - \hat{b}\cdot\bar{x}\right) - \hat{b}\cdot x(i)\right)^2 = \sum_{i=1}^{n} \left(\left(y(i) - \bar{y}\right) - \hat{b}\cdot\left(x(i) - \bar{x}\right)\right)^2 = \\
&= \sum_{i=1}^{n} \left(\left(y(i) - \bar{y}\right)\right)^2 - 2\cdot\hat{b}\cdot\sum_{i=1}^{n} \left(\left(y(i) - \bar{y}\right)\cdot\left(x(i) - \bar{x}\right)\right) + \hat{b}^2\cdot\sum_{i=1}^{n} \left(\left(x(i) - \bar{x}\right)\right)^2 = \\
&= S_{yy} - 2\cdot\hat{b}\cdot S_{xy} + \hat{b}^2\cdot S_{xx} = S_{yy} - 2\cdot\left(\frac{S_{xy}}{S_{xx}}\right)\cdot S_{xy} + \left(\frac{S_{xy}}{S_{xx}}\right)^2\cdot S_{xx} = \\
&= S_{yy} - \left(\frac{S_{xy}}{S_{xx}}\right)\cdot S_{xy} = S_{yy} - \hat{b}\cdot S_{xy}
\end{aligned}
\qquad (9.234)
$$

sledi :

$$S_{yy} = SSE + \hat{b}\cdot S_{xy}$$

Po drugi strani velja (glej (9.220) in (9.222)):

$$SST = S_{yy} = SSR + SSE \tag{9.235}$$

Torej je očitno:

$$SSR = \hat{b} \cdot S_{xy} \tag{9.236}$$

Tvorimo še:

$$\begin{aligned}
&\sum_{i=1}^{n}\left(\hat{\bar{y}}(i)-\bar{y}\right)^2 = \sum_{i=1}^{n}\left(\hat{a}+\hat{b}\cdot x(i)-\bar{y}\right)^2 = \sum_{i=1}^{n}\left(\left(\bar{y}-\hat{b}\cdot\bar{x}\right)+\hat{b}\cdot x(i)-\bar{y}\right)^2 = \\
&= \sum_{i=1}^{n}\left(\hat{b}\cdot\left(x(i)-\bar{x}\right)\right)^2 = \hat{b}^2\sum_{i=1}^{n}\left(\left(x(i)-\bar{x}\right)\right)^2 = \hat{b}^2\cdot S_{xx} = \left(\frac{S_{xy}}{S_{xx}}\right)^2\cdot S_{xx} = \\
&= \left(\frac{S_{xy}}{S_{xx}}\right)\cdot S_{xy} = \hat{b}\cdot S_{xy}
\end{aligned} \tag{9.237}$$

Torej veljajo naslednje relacije:

$$\begin{aligned}
&SSR = \hat{b}\cdot S_{xy} = \sum_{i=1}^{n}\left(\hat{\bar{y}}(i)-\bar{y}\right)^2 \\
&SSE = (n-2)\cdot s_e{}^2 = \sum_{i=1}^{n}\varepsilon^2(i) = \sum_{i=1}^{n}\left(y(i)-\hat{\bar{y}}(i)\right)^2 \\
&SST = S_{yy} = \sum_{i=1}^{n}\left(y(i)-\bar{y}(i)\right)^2 \\
&SST = SSR + SSE \\
&\underbrace{\sum_{i=1}^{n}\left(y(i)-\bar{y}(i)\right)^2}_{SST} = \underbrace{\sum_{i=1}^{n}\left(\hat{\bar{y}}(i)-\bar{y}\right)^2}_{SSR} + \underbrace{\sum_{i=1}^{n}\left(y(i)-\hat{\bar{y}}(i)\right)^2}_{SSE}
\end{aligned} \tag{9.238}$$

S temi izpeljavami smo tudi potrdili točnost izraza (9.178).

Faktor SSR se imenuje tudi regresijska vsota kvadratov. Ker je odvisen od S_{xy}, v sebi preko parametra $\hat{b}$ skriva vpliv regresijske premice in preko nje neodvisne spremenljivke na variacijo vrednosti odvisne spremenljivke. Po drugi strani pa SSE vključuje naključne vplive, ki niso povezani z vrednostmi neodvisne spremenljivke [Walpole]. Gotovo velja, da če je vrednost parametra $\hat{b}$ blizu 0, je vpliv regresijske premice in preko nje neodvisne spremenljivke na variacijo vrednosti odvisne spremenljivke majhen, regresijska premica pa ni najbolj primeren model. Torej lahko testiramo hipotezi:

$$\begin{aligned} H_0 &: b = 0 \\ H_1 &: b \neq 0 \end{aligned} \tag{9.239}$$

Pokazati se da, da sta pri predpostavki ničelne hipoteze $\frac{SSR}{\sigma^2}$ in $\frac{SSE}{\sigma^2}$ vrednosti neodvisnih hi kvadrat spremenljivk s stopnjama prostosti 1 oz. *n - 2* [Walpole]. Prav tako $\frac{SST}{\sigma^2}$ predstavlja vrednost hi kvadrat spremenljivke s stopnjo prostosti *n - 1* [Walpole]. Torej lahko zgornji hipotezi testiramo z naslednjo F statistiko [Walpole]:

$$F = \frac{\frac{SSR}{1}}{\frac{SSE}{n-2}} = \frac{SSR}{\frac{(n-2)\cdot s_e^{\,2}}{n-2}} = \frac{SSR}{s_e^{\,2}} \tag{9.240}$$

Ničelno hipotezo zavrnemo v primeru, če velja [Walpole]:

$$F = \frac{SSR}{s_e^{\,2}} > F_{krit} = F(\alpha, 1, n-2) \tag{9.241}$$

V literaturi se velikokrat pojavijo tudi naslednje oznake [Montgomery 1]:

$$\begin{aligned} MSE &= s_e^{\,2} = \frac{SSE}{n-2} \\ MSR &= \frac{SSR}{1} \\ F &= \frac{SSR}{s_e^{\,2}} = \frac{MSR}{MSE} \\ \hat{b} &= \hat{\beta}_1 \end{aligned} \tag{9.242}$$

Testna procedura za ANOVO za testiranje signifikantnosti regresije je običajno izvedena v obliki tabele, kot jo prikazuje slika 264 [Montgomery 1].

Table 11-3 Analysis of Variance for Testing Significance of Regression

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Regression	$SS_R = \hat{\beta}_1 S_{xy}$	1	MS_R	MS_R/MS_E
Error	$SS_E = SS_T - \hat{\beta}_1 S_{xy}$	$n - 2$	MS_E	
Total	SS_T	$n - 1$		

Slika 264: Tabela za testno proceduro za ANOVO za testiranje signifikantnosti regresije (source of variation - vir variacije, sum of squares - vsota kvadratov, degrees of freedom - prostostne stopnje, mean square - srednja vsota kvadratov, error - napaka) [Montgomery 1]

Tovrstne testne procedure nam običajno dajo tudi p vrednost v naslednji obliki:

$$p = P\left(F < F_{krit}\right) \tag{9.243}$$

ki je seveda vezana na ničelno oz. nasprotno hipotezo v izrazu (9.239). Sklep bi torej bil, da če nam uspe zavrniti ničelno hipotezo, potem je regresija signifikantna, regresijska premica pa predstavlja ustrezen model.

Drugi testi

Za preverjanje ustreznosti in kvalitete regresijske premice je priporočljivo uporabiti še nekatere druge teste. Tako lahko izvedemo na slednje teste [Žibert, Walpole, Vidakovic, Montgomery 1, Gujarati]:

1. Preverjanje normalnosti porazdelitve standardiziranih (normiranih) residualov (pogreškov) s pomočjo takoimenovanega **q-q diagrama** oz. **normal probability** diagrama (glej sliko 265 [Žibert]),

2. Preverjanje normalnosti porazdelitve standardiziranih (normiranih) residualov (pogreškov) s pomočjo opazovanja histograma (glej sliko 266 [Vidakovic]),

3. **Jarque-Bera** test normalnosti pogreška [Gujarati].

Kot vidimo, so vsi trije testi vezani na ugotavljanje, če je pogrešek modela res normalno porazdeljen, saj na tem sloni vsa izpeljana teorija za regresijsko premico kot model. Pri q-q diagramu na horizontalno os nanašamo pogreške, na vertikalno os pa pričakovane vrednosti pogreškov, če bi bili dejansko normalno porazdeljeni. Torej, če je pogrešek res normalno porazdeljen, bi moral biti q-q diagram približno enak premici [Gujarati]. Kar se tiče histograma pogreškov, na dotični diagram narišemo tudi teoretično porazdelitev, ki se najbolj ujema s histogramom (funkcija **histfit** v Matlabu). Če je slednja podobna normalni porazdelitvi, sklepamo, da je histogram pogreška približno normalno porazdeljen [Gujarati].

Pri Jarque-Bera testu normalnosti pogreška vpeljemo naslednjo testno statistiko [Gujarati]:

$$JB = n \cdot \left[\frac{S^2}{6} + \frac{(K-3)^2}{24} \right] \tag{9.244}$$

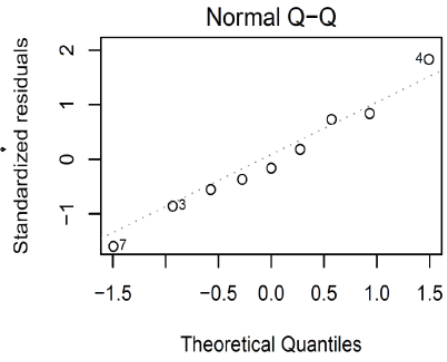
pri čemer je S stopnja asimetrije, K pa stopnja sploščenosti pogreška. Če je pogrešek res približno normalno porazdeljen, velja [Gujarati]:

$$\begin{aligned} &S \approx 0 \\ &K \approx 3 \\ &JB = n \cdot \left[\frac{S^2}{6} + \frac{(K-3)^2}{24} \right] \approx 0 \end{aligned} \tag{9.245}$$

Pri ničelni hipotezi, da so residuali normalno porazdeljeni, sta Jarque in Bera pokazala, da JB statistika v asimptotičnem smislu (pri dovolj veliki velikosti vzorca) sledi hi kvadrat porazdelitvi. Če je vrednost JB statistike zelo različna od 0 in p vrednost zadovoljivo majhna, zavržemo ničelno hipotezo (torej pogrešek ni normalno porazdeljen). Če pa je vrednost JB statistike blizu 0 in p vrednost razumljivo velika, pa sprejmemo ničelno hipotezo (torej pogrešek je normalno porazdeljen) [Gujarati]. Če je velikost vzorca premajhna, si lahko pomagamo tako, da jo umetno povečamo s pomočjo Monte Carlo simulacij. Takšen mehanizem ima vgrajen tudi funkcija **jbtest** v Matlabu.

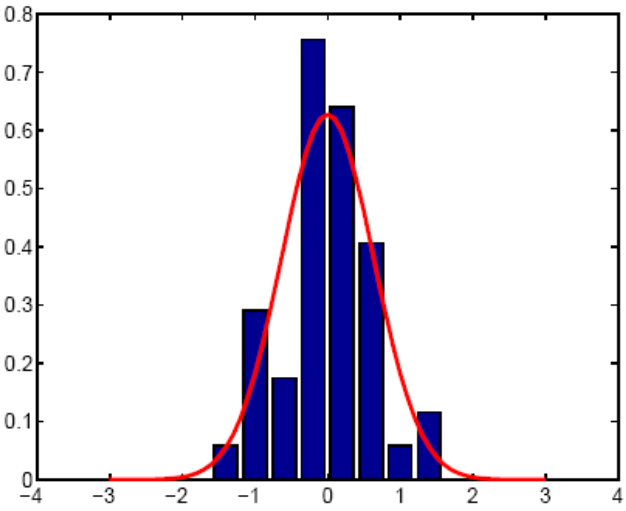
- Preverjanje normalnosti porazdelitve residualov:

$d_i = e_i/\sqrt{\hat{\sigma}^2}$ →



označene so točke, ki najbolj izstopajo od premice

Slika 265: Preverjanje normalnosti porazdelitve standardiziranih (normiranih) residualov (pogreškov) s pomočjo q-q diagrama [Žibert]. Pri tem kot standardizirane residuale običajno vzamemo: $e_{i\,\mathrm{stand}} = \frac{e_i}{s_e}, i = 1,\ldots,n.$



Slika 266: Preverjanje normalnosti porazdelitve standardiziranih (normiranih) residualov (pogreškov) s pomočjo opazovanja normiranega ali nenormiranega histograma [Vidakovic]. Na abscisi imamo pogrešek, na oordinati pa histogram in ocenjeno teoretično porazdelitev pogreška.

Detekcija točk z največjim vplivom na ocenjene parametre in kvaliteto modela

Določene točke (**outliers**) lahko zelo odstopajo od povprečja ostalih točk in s tem kvarijo kvaliteto ocenjevanja parametrov in s tem modela pri linearni regresiji z metodo najmanjših kvadratov. V ta namen lahko vpeljemo takoimenovano Cookovo razdaljo pri linearni regresiji, ki ima obliko [Žibert]:

$$D_i = \frac{\sum_{j=1}^{n}\left(\hat{\bar{y}}(j)-\hat{\bar{y}}(j,i)\right)^2}{p\cdot\frac{\sum_{j=1}^{n}\left(y(j)-\hat{\bar{y}}(j)\right)^2}{n}} \qquad (9.246)$$

kjer je p število parametrov modela, $\hat{\bar{y}}(j,i)$ pa ocena z modelom brez *i-te* meritve [Žibert]. Pri tem velja, da so točke s Cookovo razdaljo večjo od 1 problematične in jih je potrebno podrobneje analizirati ali pa izključiti iz regresije.

V nadaljevanju bomo prikazali program **regres_mnk1.m**, ki ilustrira delovanje opisanih testov kvalitete modela na podatkih za 4 primere. Primeri so izbrani iz virov [Walpole], [Montgomery 1], [Vidakovic] in [Gujarati]. Iz programov bo razvidna tudi uporaba standardnih Matlabovih funkcij **regress** in **regstats**. Na koncu programa sta dodani tudi funkciji **archtest** in **lbqtest**, ki sta sicer značilni za analizo časovnih vrst. Prva testira, če v pogrešku modela ni arch (avtoregresivnega pogojno heteroskedastičnega) efekta, druga pa uprizori takoimenovani Ljung-Boxov test adekvatnosti modela. Podrobnejšo razlago obeh tovrstnih testov je mogoče zaslediti v ustrezni literaturi za analizo časovnih vrst. Program regres_mnk1.m ima naslednji izgled:

```
% regres_mnk1.m

clear
clc
close all

ch = input('izberi primere 1, 2, 3, 4');

if ch == 1  % walpole, str. 299

   x= [3 7 11 15 18 27 29 30 30 31 31 32 33 33 34 36 36 36 37 38 39 39 39 40 41 42 42 43 44 45 46 47
50]';
   y= [5 11 21 16 16 28 27 25 35 30 40 32 34 32 34 37 38 34 36 38 37 36 45 39 41 40 44 37 44 46 46 49
51]';
   xstr='redukcija trdne snovi (%)'
   ystr='redukcija kemijske potrebe po kisiku (%)'

elseif ch == 2 % montgomery, str. 373

   x=[0.99 1.02 1.15 1.29 1.46 1.36 0.87 1.23 1.55 1.4 1.19 1.15 0.98 1.01 1.11 1.2 1.26 1.32 1.43
0.95]';
   y=[90.01 89.05 91.43 93.74 96.73 94.45 87.59 91.77 99.42 93.65 93.54 92.52 90.56 89.54 89.85 90.39
93.25 93.41 94.98 87.33]';
   xstr='ogljikovodik (%)'
   ystr = 'čistost (%)'

elseif ch == 3  % vidakovic, str. 601

   x=[-8.1 -16.1 -0.9 -7.8 -29.0 -19.2 -18.9 -10.6 -2.8...
      -25.0 -3.1 -7.8 -13.9 -4.5 -11.6 -2.1 -2.0 -9.0 -11.2 -0.2...
      -6.1 -1 -3.6 -8.2 -0.5 -2.0 -1.6 -11.9 -0.7 -1.2 -14.3 -0.8...
      -16.8 -5.1 -9.5 -17.0 -3.3 -0.7 -3.3 -13.6 -1.9 -10.0 -13.5]';
   y=[ 4.8 4.1 5.2 5.5 5 3.4 3.4 4.9 5.6 3.7 3.9 ...
       4.5 4.8 4.9 3.0 4.6 4.8 5.5 4.5 5.3 4.7 6.6 5.1 3.9 ...
       5.7 5.1 5.2 3.7 4.9 4.8 4.4 5.2 5.1 4.6 3.9 5.1 5.1 ...
       6.0 4.9 4.1 4.6 4.9 5.1]';
   xstr = 'deficit'
   ystr = 'log C peptid'

else

   z = importdata('data_3.2.shd');  % gujarati, str. 149
   x = z(:,2);
   y = z(:,1);
   ystr = 'izdatki za hrano (v tisočih rupijah)'
   xstr = 'celotni izdatki (v tisočih rupijah)'

end
```

```
n = length(x)

sx = sum(x);
sy = sum(y);
sx2 = x'*x;
sy2 = y'*y;
sxy = x'*y;
Sxy = sxy - sx*sy/n
Sxx = sx2 - sx^2/n
Syy = sy2 - sy^2/n
ysr = sy/n
xsr = sx/n

alfa = input('alfa=')
if length(alfa)==0
   alfa = 0.05
end

nrep = input('stevilo monte carlo replikacij za jb test')
if length(nrep)==0
   nrep = []
end



disp('boc=')
boc = Sxy/Sxx

disp('aoc=')
aoc = sy/n - boc*sx/n

disp('standardna ocena napake modela:')
se=sqrt((Syy-boc*Sxy)/(n-2)) % standardna ocena napake modela (dobljena glede na ucne vzorce)


SSR = boc * Sxy
SSE = (n-2)*se^2
SST = SSR + SSE   % mora priti enak kot Syy


nreg = 1;  % prostost. stopnja regresije


F = SSR/se^2  % testna statistika za nicelno hipotezo b=0 napram nasprotni, da b <> 0

Fkrit = finv(1-alfa,nreg,n-2)  % kriticna vrednost

if F > Fkrit
   disp('Zavrni nicelno hipotezo, da je b = 0, torej je model premice signifikanten')
else
   disp('Sprejmi nicelno hipotezo, da je b = 0, torej model premice ni signifikanten')
end


disp('p vrednost je:')

p = fcdf(F,nreg,n-2);
p = min(p,1-p)


disp('korelacijski koeficient:')

ro = Sxy/sqrt(Sxx*Syy)

disp('determinacijski koeficient:')

D = ro^2

% Tvorjenje vrednosti modela - premice:

yoc = aoc + boc*x;

% Narisemo meritve in regresijsko premico

plot(x,y,'ko','LineWidth',3)
hold on

plot(x,yoc,'r','LineWidth',2)
plot(x,yoc,'ro','LineWidth',2)

grid
title('razsevni diagram za izhod modela y in njegova ocena yoc pri meritvah')
xlabel(xstr)
ylabel(ystr)

% Tvorimo in narisemo pogreske modela
```

```
for i = 1:length(x)
    e(i) = y(i) - yoc(i);
end

figure
plot(x,e,'o','LineWidth',4)
hold on
grid
title('pogresek modela e(x) pri meritvah, uporabljenih pri ocenjevanju')
xlabel('razsevni diagram, x os')
d = axis;
axis([d(1) d(2) d(3)-1.5*abs(min(e)) d(4)+1.5*abs(max(e))])


disp('%------------------------------------------------------------------------')
disp('% Izracun s standarnim regress ukazom:')
disp('%------------------------------------------------------------------------')

n = length(x)

X = [ones(size(x)) x];

[b,bint,r,rint,stats] = regress(y,X,alfa);

disp('ocenjena parametra sta:')
aoc = b(1)
boc = b(2)

disp('intervala zaupanja za ocenjena parametra:')
Iaoc = bint(1,:)
Iboc = bint(2,:)

disp('determinacijski koeficient:')
D = stats(1)

disp('x    spodnja meja pogreska    pogresek     zgornja meja pogreska:')
[x rint(:,1) r rint(:,2)]

disp('Standardna ocena napake modela je:')
se = sqrt(stats(4))

SSE = (n-2)*se^2

Sxy = x'*y - sum(x)*sum(y)/n

SSR = boc*Sxy
SST = SSE + SSR

disp('F vrednost je (1. nacin - iz regress):')
F = stats(2)

disp('F vrednost je (2. nacin - rocno):')
F = SSR/se^2

disp('p vrednost je - iz regress:')
p = stats(3)

disp('p vrednost je - na 2. nacin:')
p = fcdf(F,nreg,n-2);
p = min(p,1-p)

Fkrit = finv(1-alfa,nreg,n-2)

if F > Fkrit
   disp('Zavrni nicelno hipotezo, da je b = 0, torej je model premice signifikanten')
else
   disp('Sprejmi nicelno hipotezo, da je b = 0, torej model premice ni signifikanten')
end


%------------------------------------------------------------------------
% drugi testi
%------------------------------------------------------------------------

stats = regstats(y,x,'linear',{'cookd','standres'});

% izris q-q plota za standardizirane residuale - pogreske (preverjamo, ce je pogresek normalno
porazdeljen)

figure
qqplot(stats.standres)
grid
xlabel('Standardni normalni kvantili - qq plot za test normalnosti pogreska')

disp('primerjava standardiziranih residualov in normiranih pogreskov:')

[stats.standres (e/se)']
```

```
% izris cookovih razdalj

figure
stem(stats.cookd)
grid
title('cookove razdalje')

% izris histograma normiranih pogreskov

figure
histfit(e/se,7)
title('histfit od pogreska')
xlabel('pogresek e')


% Jarque-Bera test normalnosti pogreska (glej Gujarati, str. 148) - pozor
% (zanesljiv je le za dovolj velik vzorec)

s = skewness(e)    % mora biti priblizno 0
k = kurtosis(e)    % mora biti priblizno 3
JB = n*(s^2/6+(k-3)^2/24) % mora biti priblizno 0

[h,p]=jbtest(e,alfa,nrep)  % ce vrne 0 in je p dovolj velik, velja nicelna hipoteza, da je e normalen

if (h == 0)&&(p>0.05)
   disp('skoraj gotovo je e normalen')
else
   disp('skoraj gotovo e ni normalen')
end

% generiranje avtokorelacije pogreska:

figure
autocorr(e,[],1)
title('avtokorelacija pogreska modela')

disp('srednja vrednost pogreska')
mean(e)

disp('stand. deviacija pogreska:')
std(e)

% test arch efekta v pogresku za arch redov 1 do 10. Ce so vsi h-ji enaki
% 0 in p-ji dovolj veliki, ni nikjer arch efekta. 3. stolp je vrednost statistike, 4. stolp pa kriticna
vrednost

disp('arch test za pogreseek:')
disp('H,P,Stat,CV')

[H,P,Stat,CV] = archtest(e,[1:1:n-2]',0.10);
[H,P,Stat,CV]

% ljung box test za pogresek glede lack of fit oz. adekvatnosti modela. Če
% so vsi h-ji enaki 0, in p-ji dovolj veliki, je model adekvaten.

disp('ljung-box test za pogresek:')
disp('H,P,Qstat,CV')

[H,P,Qstat,CV] = lbqtest(e, [1:1:n-1]', 0.10);
[H,P,Qstat,CV]
```

Kot vidimo v programu, smo dodali tudi funkcijo **autocorr**, ki izriše avtokorelacijo pogreška modela. Prav tako smo v programu primerjali standardizirane residuale, pridobljene s funkcijo regstats, ter normirane pogreške $e_{istand} = \frac{e_i}{s_e}, i = 1,...,n$. Do rahlih odstopanj med tema spremenljivkama pride zato, ker funkcija regstats pri izračunih vplete še takoimenovano **QR dekompozicijo**.

V nadaljevanju si poglejmo, kako bi ta program izvedel teste za primere, izbrane iz virov [Walpole], [Montgomery 1], [Vidakovic] in [Gujarati].

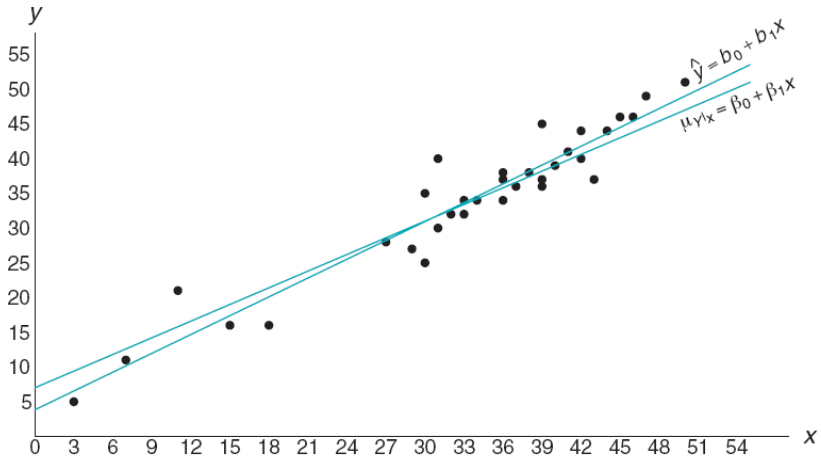
Primer 9.20.:

Odpad v usnjarski industriji je kemijsko zelo kompleksen in so zanj značilne velika kemijska potreba po kisiku, nestanovitne trdnine, in druge polucijske snovi. Na sliki 267 so prikazani podatki za 33 vzorcev za redukcijo trdnin (solids reduction) v % kot neodvisno spremenljivko x, ter redukcijo kisika (oxygen demand reduction) v % kot odvisno spremenljivko y. Na sliki 268 pa je za podatke prikazan razsevni diagram z ocenjeno regresijsko premico [Walpole].

Table 7.1: Measures of Reduction in Solids and Oxygen Demand

Solids Reduction, x (%)	Oxygen Demand Reduction, y (%)	Solids Reduction, x (%)	Oxygen Demand Reduction, y (%)
3	5	36	34
7	11	37	36
11	21	38	38
15	16	39	37
18	16	39	36
27	28	39	45
29	27	40	39
30	25	41	41
30	35	42	40
31	30	42	44
31	40	43	37
32	32	44	44
33	34	45	46
33	32	46	46
34	34	47	49
36	37	50	51
36	38		

Slika 267: Podatki za 33 vzorcev za redukcijo trdnin (solids reduction) v % kot neodvisno spremenljivko x, ter redukcijo kisika (oxygen demand reduction) v % kot odvisno spremenljivko y. [Walpole]



Slika 268: Razsevni diagram z ocenjeno regresijsko premico [Walpole]

Uporabimo program **regres_mnk1.m** za izračune. Izpis rezultatov v komandnem oknu je naslednji:

```
izberi primere 1, 2, 3, 41
xstr =
redukcija trdne snovi (%)
ystr =
redukcija kemijske potrebe po kisiku (%)
n =
   33

Sxy =
  3.7521e+003
Sxx =
  4.1522e+003
Syy =
  3.7139e+003
ysr =
   34.0606
xsr =
   33.4545

alfa=
alfa =
alfa =
    0.0500

stevilo monte carlo replikacij za jb test100
nrep =
   100

boc=
boc =
    0.9036
aoc=
aoc =
    3.8296

standardna ocena napake modela:
se =
    3.2295
SSR =
  3.3906e+003
SSE =
  323.3273
SST =
  3.7139e+003

F =
  325.0795
Fkrit =
```

```
    4.1596

Zavrni nicelno hipotezo, da je b = 0, torej je model premice signifikanten
p vrednost je:
p =
     0

korelacijski koeficient:
ro =
    0.9555
determinacijski koeficient:
D =
    0.9129

%------------------------------------------------------------------------
% Izracun s standarnim regress ukazom:
%------------------------------------------------------------------------
n =
    33
ocenjena parametra sta:
aoc =
    3.8296
boc =
    0.9036

intervala zaupanja za ocenjena parametra:
Iaoc =
    0.2229    7.4364
Iboc =
    0.8014    1.0059
determinacijski koeficient:
D =
    0.9129

x    spodnja meja pogreska    pogresek    zgornja meja pogreska:
ans =
    3.0000   -7.2964   -1.5406    4.2152
    7.0000   -5.1399    0.8449    6.8296
   11.0000    1.6823    7.2303   12.7782
   15.0000   -7.6715   -1.3843    4.9030
   18.0000  -10.3055   -4.0952    2.1151
   27.0000   -6.7866   -0.2280    6.3306
   29.0000   -9.5145   -3.0353    3.4439
   30.0000  -12.1400   -5.9389    0.2621
   30.0000   -2.3465    4.0611   10.4686
   31.0000   -8.3952   -1.8426    4.7100
   31.0000    2.3110    8.1574   14.0038
   32.0000   -7.3320   -0.7462    5.8395
   33.0000   -6.2417    0.3501    6.9420
   33.0000   -8.2143   -1.6499    4.9146
   34.0000   -7.1434   -0.5535    6.0364
   36.0000   -5.9445    0.6392    7.2229
   36.0000   -4.9205    1.6392    8.1989
   36.0000   -8.8899   -2.3608    4.1683
```

```
   37.0000   -7.8306   -1.2644    5.3017
   38.0000   -6.7442   -0.1681    6.4080
   39.0000   -8.5944   -2.0717    4.4509
   39.0000   -9.5395   -3.0717    3.3960
   39.0000   -0.2578    5.9283   12.1143
   40.0000   -7.5235   -0.9754    5.5727
   41.0000   -6.4254    0.1210    6.6674
   42.0000   -8.2821   -1.7826    4.7168
   42.0000   -4.2635    2.2174    8.6982
   43.0000  -11.8511   -5.6863    0.4785
   44.0000   -6.0898    0.4101    6.9099
   45.0000   -4.9525    1.5064    7.9654
   46.0000   -5.8565    0.6028    7.0621
   47.0000   -3.6633    2.6991    9.0616
   50.0000   -4.3338    1.9882    8.3103

Standardna ocena napake modela je:
se =
    3.2295
SSE =
  323.3273
Sxy =
  3.7521e+003
SSR =
  3.3906e+003

SST =
  3.7139e+003

F vrednost je (1. nacin - iz regress):
F =
  325.0795
F vrednost je (2. nacin - rocno):
F =
  325.0795

p vrednost je - iz regress:
p =
     0
p vrednost je - na 2. nacin:
p =
     0

Fkrit =
    4.1596

Zavrni nicelno hipotezo, da je b = 0, torej je model premice signifikanten

primerjava standardiziranih residualov in normiranih pogreskov:
ans =
   -0.5522   -0.4770
    0.2923    0.2616
    2.4308    2.2388
   -0.4549   -0.4286
```

```
   -1.3277   -1.2680
   -0.0721   -0.0706
   -0.9568   -0.9399
   -1.8702   -1.8389
    1.2789    1.2575
   -0.5798   -0.5705
    2.5670    2.5259
   -0.2347   -0.2311
    0.1101    0.1084
   -0.5188   -0.5109
   -0.1741   -0.1714
    0.2012    0.1979
    0.5159    0.5076
   -0.7429   -0.7310
   -0.3982   -0.3915
   -0.0530   -0.0520
   -0.6539   -0.6415
   -0.9696   -0.9511
    1.8713    1.8356
   -0.3083   -0.3020
    0.0383    0.0375
   -0.5657   -0.5520
    0.7036    0.6866
   -1.8086   -1.7607
    0.1308    0.1270
    0.4817    0.4665
    0.1934    0.1866
    0.8687    0.8358
    0.6476    0.6156

s =
    0.6682
k =
    3.6651
JB =
    3.0641
h =
     0
p =
    0.0800

skoraj gotovo je e normalen

srednja vrednost pogreska
ans =
 -3.4989e-015
stand. deviacija pogreska:
ans =
    3.1787

arch test za pogreseek:
H,P,Stat,CV
ans =
         0    0.4344    0.6109    2.7055
```

```
     0   0.6746   0.7873    4.6052
     0   0.9039   0.5672    6.2514
     0   0.9697   0.5379    7.7794
     0   0.9678   0.9334    9.2364
     0   0.9697   1.3354   10.6446
     0   0.9549   2.0872   12.0170
     0   0.9588   2.5619   13.3616
     0   0.9470   3.3857   14.6837
     0   0.9621   3.6416   15.9872
     0   0.7390   7.7099   17.2750
     0   0.4721  11.6751   18.5493
     0   0.4058  13.5562   19.8119
     0   0.3156  15.9636   21.0641
     0   0.3630  16.2906   22.3071
     0   0.3856  17.0000   23.5418
     0   0.5238  16.0000   24.7690
     0   0.6620  15.0000   25.9894
     0   0.7837  14.0000   27.2036
     0   0.8774  13.0000   28.4120
     0   0.9396  12.0000   29.6151
     0   0.9747  11.0000   30.8133
     0   0.9913  10.0000   32.0069
     0   0.9976   9.0000   33.1962
     0   0.9995   8.0000   34.3816
     0   0.9999   7.0000   35.5632
     0   1.0000   6.0000   36.7412
     0   1.0000   5.0000   37.9159
     0   1.0000   4.0000   39.0875
     0   1.0000   3.0000   40.2560
     0   1.0000   2.0000   41.4217


ljung-box test za pogresek:
H,P,Qstat,CV
ans =
     0   0.2610   1.2634    2.7055
     0   0.5247   1.2898    4.6052
     0   0.6107   1.8194    6.2514
     0   0.6833   2.2863    7.7794
     0   0.2302   6.8736    9.2364
     0   0.2651   7.6468   10.6446
     0   0.3638   7.6574   12.0170
     0   0.3683   8.6989   13.3616
     0   0.4610   8.7469   14.6837
     0   0.5554   8.7560   15.9872
     0   0.5446   9.8424   17.2750
     0   0.5934  10.2572   18.5493
     0   0.6700  10.2910   19.8119
     0   0.4663  13.7787   21.0641
     0   0.3659  16.2465   22.3071
     0   0.4295  16.3396   23.5418
     0   0.4209  17.5013   24.7690
     0   0.4869  17.5314   25.9894
     0   0.4247  19.5077   27.2036
     0   0.2341  24.1931   28.4120
```

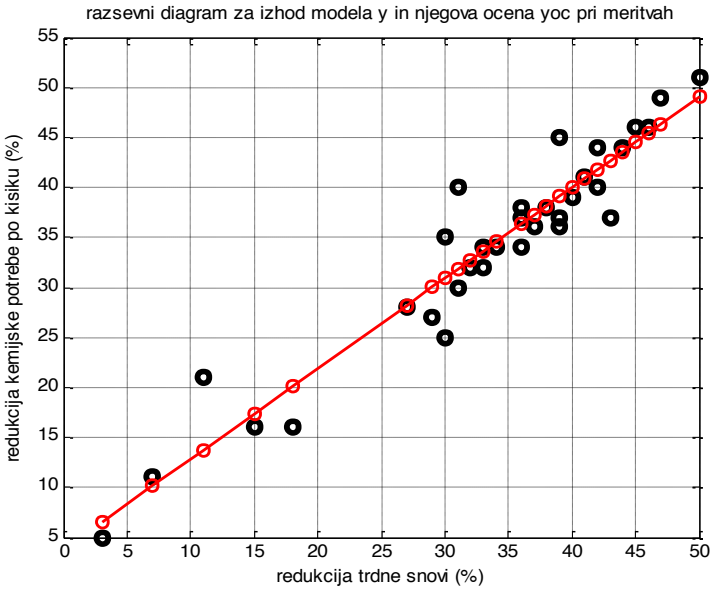
```
    0    0.1893   26.4656   29.6151
    0    0.2274   26.5848   30.8133
    0    0.2709   26.6553   32.0069
    0    0.3144   26.7881   33.1962
    0    0.1486   32.3338   34.3816
    0    0.1691   32.7614   35.5632
    0    0.2014   32.8693   36.7412
    0    0.2367   32.9739   37.9159
    0    0.2545   33.5914   39.0875
    0    0.2612   34.5023   40.2560
    0    0.3025   34.5362   41.4217
    0    0.3430   34.6399   42.5847
```

Dobimo torej:

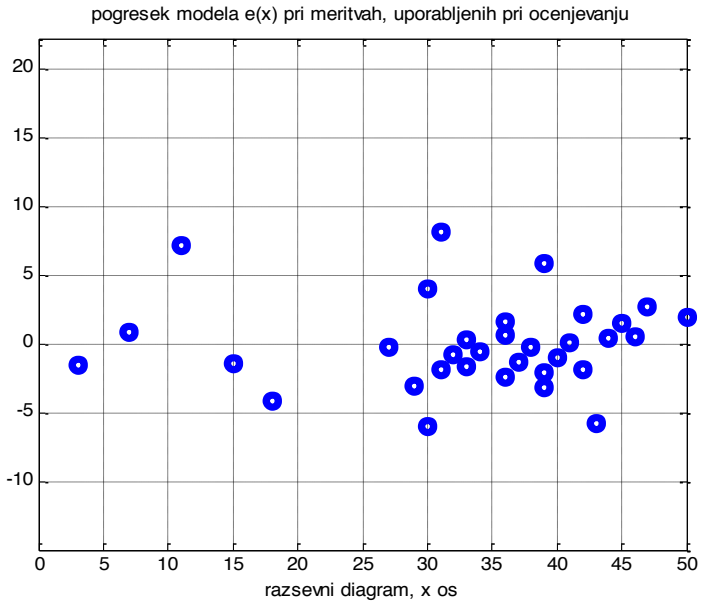
$$
\begin{aligned}
&n = 33 \\
&S_{xy} = 3752.1,\ S_{xx} = 4152.2,\ \ S_{yy} = 3713.9 \\
&\overline{y} = 34.0606,\ \overline{x} = 33.4545 \\
&\alpha = 0.05 \\
&n_{rep} = 100\ (za\ Monte\ Carlo\ \ pri\ JB\ \ testu) \\
&\hat{b} = 0.9036,\quad \hat{a} = 3.8296 \\
&s_e = 3.2295 \\
&SSR = 3390.6,\quad SSE = 323.3273,\quad SST = 3713.9 \\
&F = 325.0795 > F_{krit} = 4.1596\ \ (p = 0) \\
&mod\ el\ \ premice\ \ je\ signifikanten \\
&r = 0.9555,\quad\ D = r^2 = 0.9129 \\
&I_a = [0.2229, 7.4364] \\
&I_b = [0.8014, 1.0059] \\
&S = 0.6682,\quad K = 3.6651,\quad JB = 3.0641 \\
&za\ \ JB:\ \ h = 0,\ \ p = 0.08\ \rightarrow skoraj\ \ gotovo\ \ \ je\ \ e\ \ normalen
\end{aligned}
\tag{9.247}
$$

Tudi test za arch in Ljung-Box test potrdita, da je model ustrezen, saj dasta v 1. stolpu same ničle, v drugem stolpu pa so p vrednosti dovolj signifikantne.

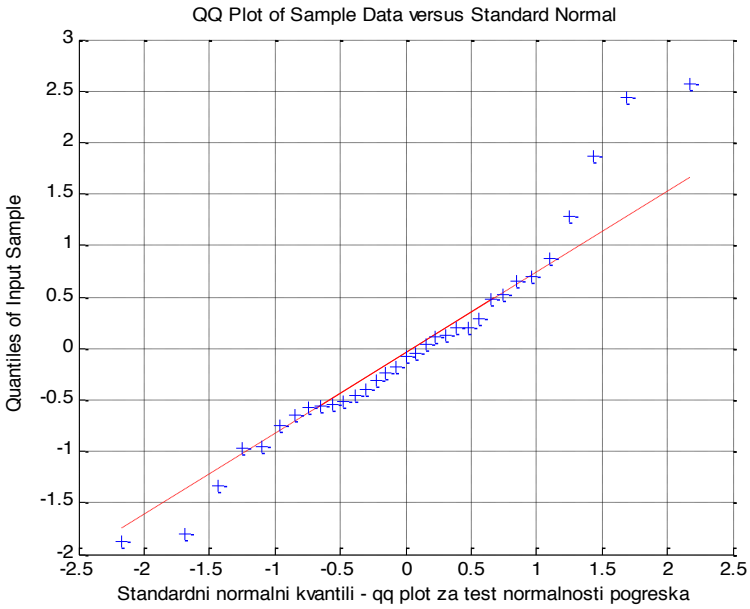
Slike 269 do 274 prikazujejo: podatke in ocenjeno regresijsko premico, razsevni diagram pogreška modela, q-q diagram za pogrešek, Cookove razdalje za pogrešek, histogram pogreška in avtokorelacijo pogreška. Tudi na osnovi teh slik sklepamo, da je uporabljeni model ustrezen.



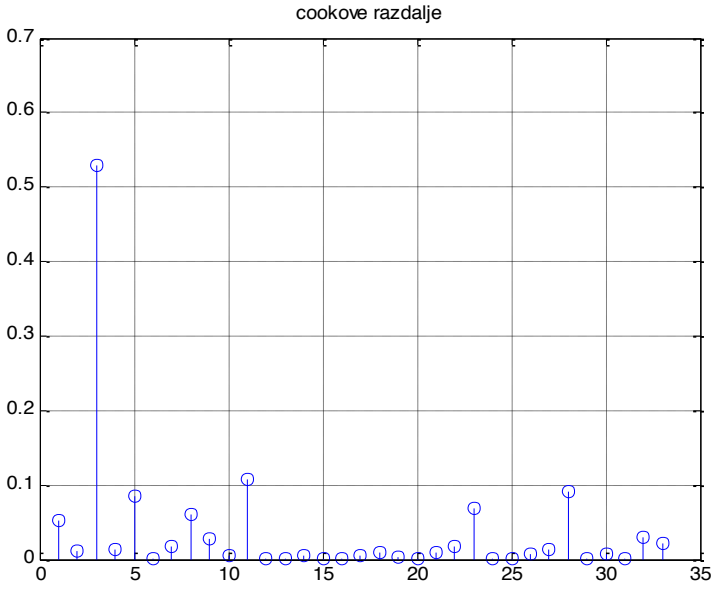
Slika 269: Podatki in ocenjena regresijska premica



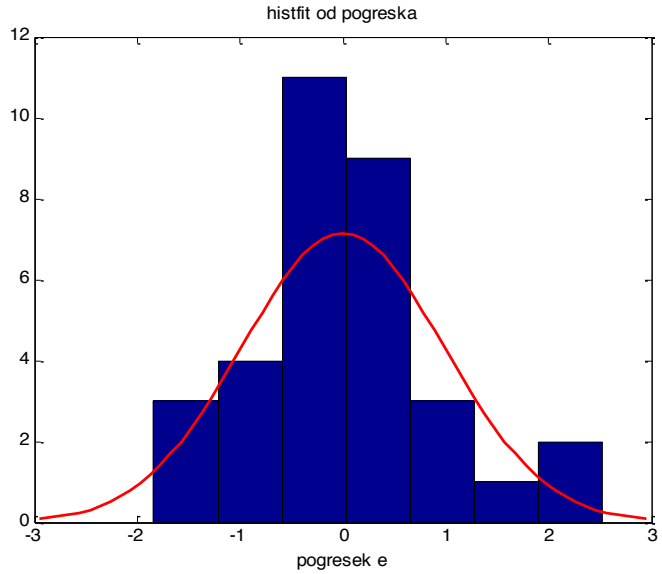
Slika 270: Razsevni diagram pogreška



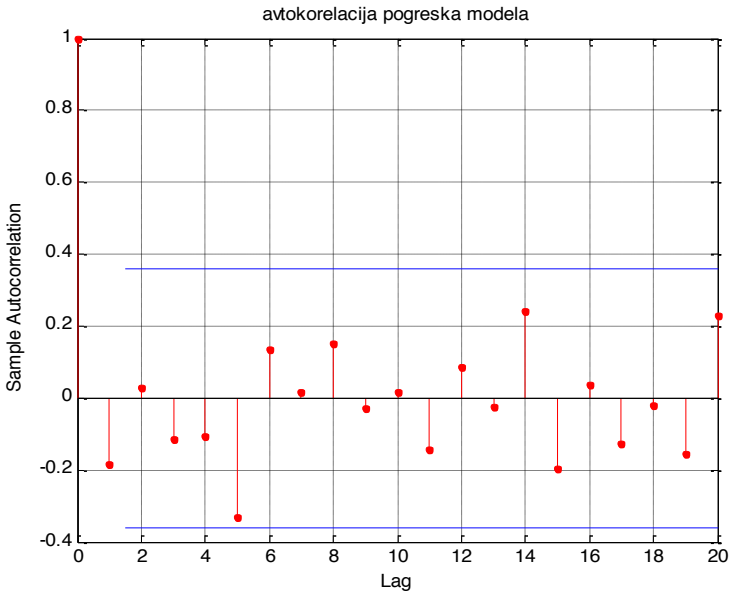
Slika 271: q-q diagram pogreška



Slika 272: Cookove razdalje za pogrešek



Slika 273: Histogram pogreška in ocenjena teoretična porazdelitev



Slika 274: Avtokorelacija pogreška

Primer 9.21.:

Dane imamo podatke na sliki 275, ki prikazujejo odvisnost čistosti kisika (purity of oxygen) y v % v kemijski distilacijski tovarni v odvisnosti od nivoja ogljikovodika (hydrocarbon level) v % (neodvisna spremenljivka x) [Montgomery 1].

Table 11-1 Oxygen and Hydrocarbon Levels

Observation Number	Hydrocarbon Level x(%)	Purity y(%)
1	0.99	90.01
2	1.02	89.05
3	1.15	91.43
4	1.29	93.74
5	1.46	96.73
6	1.36	94.45
7	0.87	87.59
8	1.23	91.77
9	1.55	99.42
10	1.40	93.65
11	1.19	93.54
12	1.15	92.52
13	0.98	90.56
14	1.01	89.54
15	1.11	89.85
16	1.20	90.39
17	1.26	93.25
18	1.32	93.41
19	1.43	94.98
20	0.95	87.33

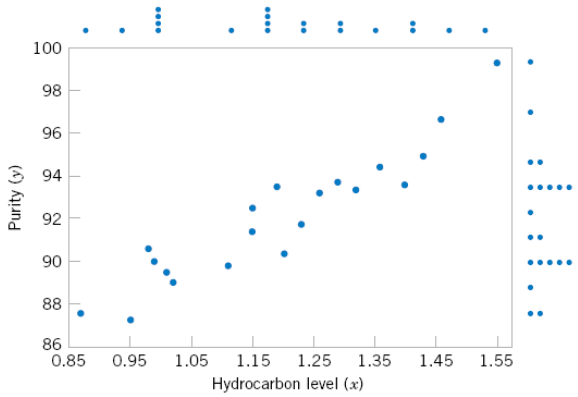


Figure 11-1 Scatter diagram of oxygen purity versus hydrocarbon level from Table 11-1.

Slika 275: Odvisnost čistosti kisika (purity of oxygen) y v % v kemijski distilacijski tovarni v odvisnosti od nivoja ogljikovodika (hydrocarbon level) v % (neodvisna spremenljivka x) [Montgomery 1].

Uporabimo program **regres_mnk1.m** za izračune. Izpis rezultatov v komandnem oknu je naslednji:

```
izberi primere 1, 2, 3, 42
xstr =
ogljikovodik (%)
ystr =
čistost (%)

n =
    20

Sxy =
   10.1774
Sxx =
    0.6809
```

```
Syy =
  173.3769
ysr =
   92.1605
xsr =
    1.1960

alfa=
alfa =
alfa =
    0.0500

stevilo monte carlo replikacij za jb test100
nrep =
   100

boc=
boc =
   14.9475
aoc=
aoc =
   74.2833

standardna ocena napake modela:
se =
    1.0865

SSR =
  152.1271
SSE =
   21.2498
SST =
  173.3769

F =
  128.8617
Fkrit =
    4.4139

Zavrni nicelno hipotezo, da je b = 0, torej je model premice signifikanten
p vrednost je:
p =
  1.2273e-009

korelacijski koeficient:
ro =
    0.9367
determinacijski koeficient:
D =
    0.8774
%-------------------------------------------------------------------------
% Izracun s standarnim regress ukazom:
%-------------------------------------------------------------------------
n =
```

```
    20


ocenjena parametra sta:
aoc =
   74.2833
boc =
   14.9475


intervala zaupanja za ocenjena parametra:
Iaoc =
   70.9356   77.6311
Iboc =
   12.1811   17.7139


determinacijski koeficient:
D =
    0.8774


x    spodnja meja pogreska    pogresek     zgornja meja pogreska:
ans =
    0.9900   -1.2332    0.9287    3.0905
    1.0200   -2.7003   -0.4797    1.7408
    1.1500   -2.3285   -0.0429    2.2426
    1.2900   -2.0976    0.1744    2.4464
    1.4600   -1.5157    0.6234    2.7625
    1.3600   -2.4017   -0.1619    2.0779
    0.8700   -1.7848    0.3024    2.3896
    1.2300   -3.1398   -0.8987    1.3423
    1.5500    0.1736    1.9681    3.7626
    1.4000   -3.6268   -1.5598    0.5072
    1.1900   -0.6943    1.4692    3.6327
    1.1500   -1.1754    1.0471    3.2696
    0.9800   -0.4151    1.6282    3.6715
    1.0100   -2.0661    0.1597    2.3856
    1.1100   -3.2406   -1.0250    1.1905
    1.2000   -3.9211   -1.8303    0.2605
    1.2600   -2.1483    0.1329    2.4140
    1.3200   -2.8450   -0.6040    1.6370
    1.4300   -2.8411   -0.6782    1.4847
    0.9500   -3.2524   -1.1534    0.9455


Standardna ocena napake modela je:
se =
    1.0865
SSE =
   21.2498
Sxy =
   10.1774
SSR =
  152.1271
SST =
  173.3769


F vrednost je (1. nacin - iz regress):
```

```
F =
  128.8617
F vrednost je (2. nacin - rocno):
F =
  128.8617

p vrednost je - iz regress:
p =
  1.2273e-009
p vrednost je - na 2. nacin:
p =
  1.2273e-009

Fkrit =
    4.4139

Zavrni nicelno hipotezo, da je b = 0, torej je model premice signifikanten

primerjava standardiziranih residualov in normiranih pogreskov:
ans =
    0.9072    0.8547
   -0.4643   -0.4415
   -0.0406   -0.0395
    0.1659    0.1605
    0.6232    0.5737
   -0.1561   -0.1490
    0.3123    0.2783
   -0.8494   -0.8271
    2.0697    1.8114
   -1.5227   -1.4356
    1.3873    1.3522
    0.9904    0.9637
    1.5961    1.4985
    0.1550    0.1470
   -0.9735   -0.9434
   -1.7283   -1.6845
    0.1259    0.1223
   -0.5772   -0.5559
   -0.6694   -0.6242
   -1.1440   -1.0616

s =
    0.1566
k =
    2.2085
JB =
    0.6038
h =
     0
p =
    0.6360
skoraj gotovo je e normalen

srednja vrednost pogreska
```

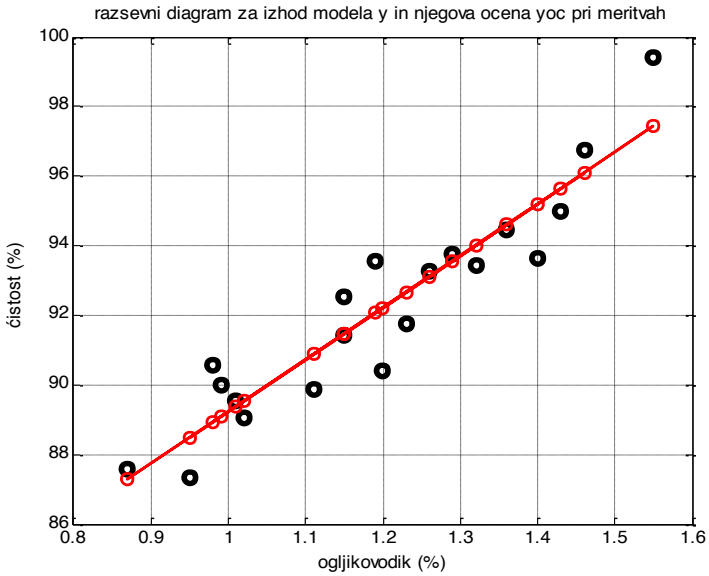
```
ans =

  1.2790e-014

stand. deviacija pogreska:

ans =

    1.0575


arch test za pogreseek:

H,P,Stat,CV

ans =

     0    0.3011    1.0695    2.7055
     0    0.6078    0.9958    4.6052
     0    0.8394    0.8420    6.2514
     0    0.9695    0.5400    7.7794
     0    0.9420    1.2300    9.2364
     0    0.9633    1.4418   10.6446
     0    0.9527    2.1236   12.0170
     0    0.4927    7.4138   13.3616
     0    0.2776   10.9728   14.6837
     0    0.4405   10.0000   15.9872
     0    0.6219    9.0000   17.2750
     0    0.7851    8.0000   18.5493
     0    0.9022    7.0000   19.8119
     0    0.9665    6.0000   21.0641
     0    0.9921    5.0000   22.3071
     0    0.9989    4.0000   23.5418
     0    0.9999    3.0000   24.7690
     0    1.0000    2.0000   25.9894


ljung-box test za pogresek:

H,P,Qstat,CV

ans =

     0    0.6850    0.1646    2.7055
     0    0.3960    1.8529    4.6052
     0    0.3242    3.4734    6.2514
     0    0.3530    4.4132    7.7794
     0    0.4247    4.9285    9.2364
     0    0.5502    4.9505   10.6446
     0    0.2502    9.0349   12.0170
     0    0.2415   10.3466   13.3616
     0    0.2207   11.8700   14.6837
     0    0.2786   12.0966   15.9872
     0    0.2887   13.0709   17.2750
     0    0.3239   13.6457   18.5493
     0    0.3991   13.6479   19.8119
     0    0.4741   13.6767   21.0641
     0    0.4537   14.9687   22.3071
     0    0.5259   14.9829   23.5418
     0    0.5959   14.9941   24.7690
     0    0.6622   14.9970   25.9894
     0    0.6496   16.1150   27.2036
```

Dobimo torej:

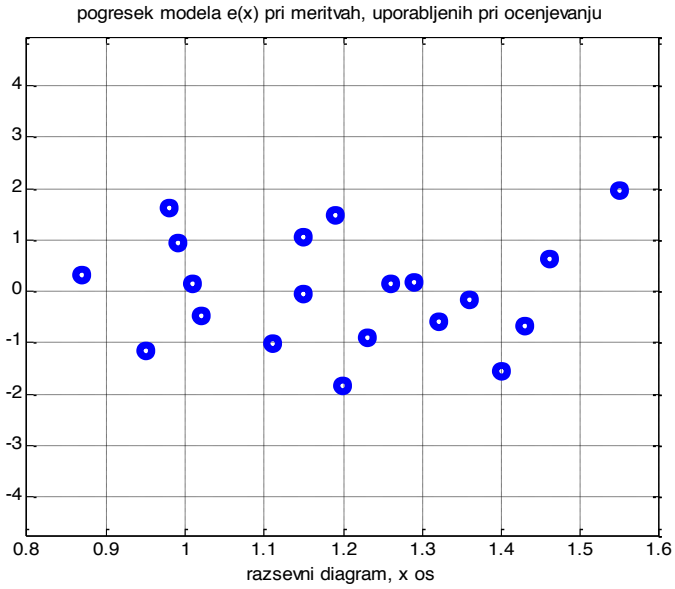
$$
\begin{aligned}
&n = 20 \\
&S_{xy} = 10.1774,\ S_{xx} = 0.6809,\ \ S_{yy} = 173.3769 \\
&\bar{y} = 92.1605,\ \bar{x} = 1.196 \\
&\alpha = 0.05 \\
&n_{rep} = 100\ (za\ Monte\ Carlo\ \ pri\ JB\ \ testu) \\
&\hat{b} = 14.9475,\quad \hat{a} = 74.2833 \\
&s_e = 1.0865 \\
&SSR = 152.1271,\quad SSE = 21.2498,\quad SST = 173.3769 \\
&F = 128.8617 > F_{krit} = 4.4139\ \ \left(p = 1.2273e\text{-}009\right) \\
&mod\,el\ \ premice\ \ je\ signifikanten \\
&r = 0.9397,\qquad D = r^2 = 0.8774 \\
&I_a = \left[70.9356, 77.6311\right] \\
&I_b = \left[12.1811, 17.7139\right] \\
&S = 0.1566,\quad K = 2.2085,\quad JB = 0.6038 \\
&za\ \ JB:\ \ h = 0,\ \ p = 0.636\ \rightarrow skoraj\ \ gotovo\ \ \ je\ \ e\ \ normalen
\end{aligned}
\tag{9.248}
$$

Tudi test za arch in Ljung-Box test potrdita, da je model ustrezen, saj dasta v 1. stolpu same ničle, v drugem stolpu pa so p vrednosti dovolj signifikantne.

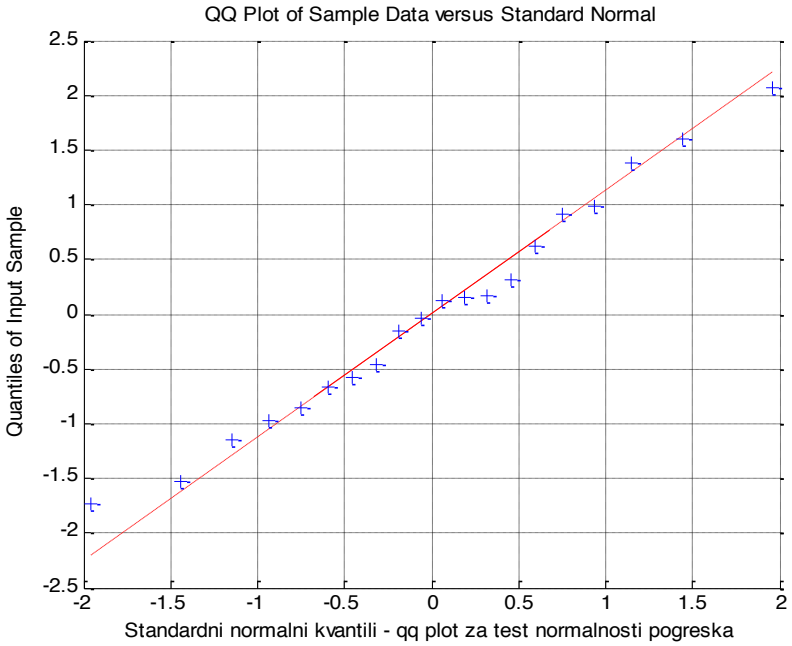
Slike 276 do 281 prikazujejo: podatke in ocenjeno regresijsko premico, razsevni diagram pogreška modela, q-q diagram za pogrešek, Cookove razdalje za pogrešek, histogram pogreška in avtokorelacijo pogreška. Tudi na osnovi teh slik sklepamo, da je uporabljeni model ustrezen.



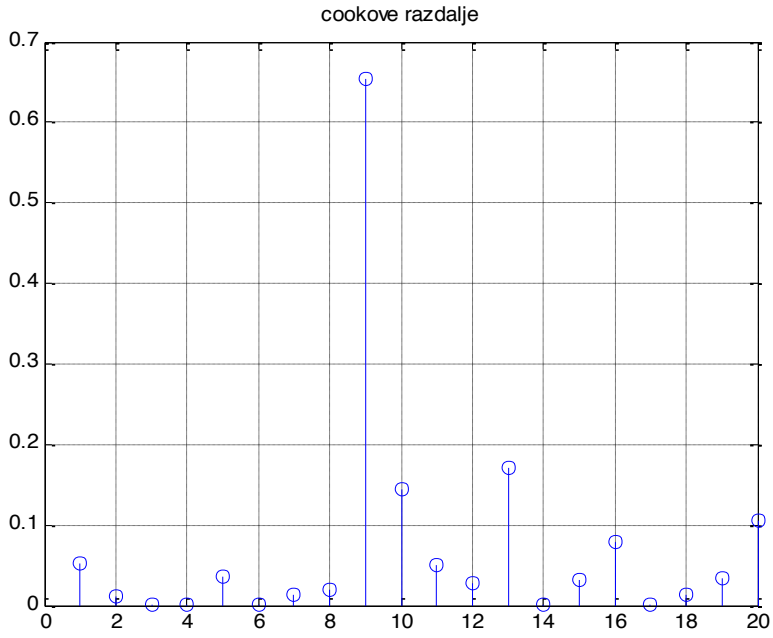
Slika 276: Podatki in ocenjena regresijska premica



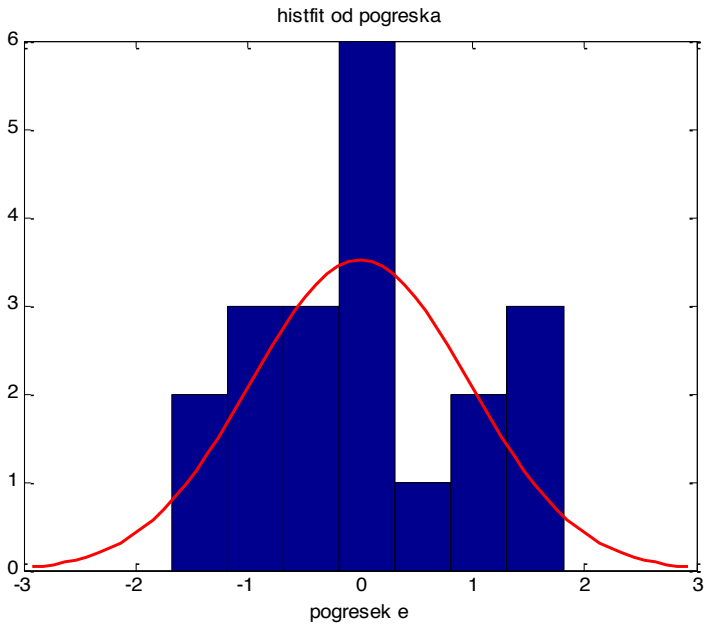
Slika 277: Razsevni diagram pogreška



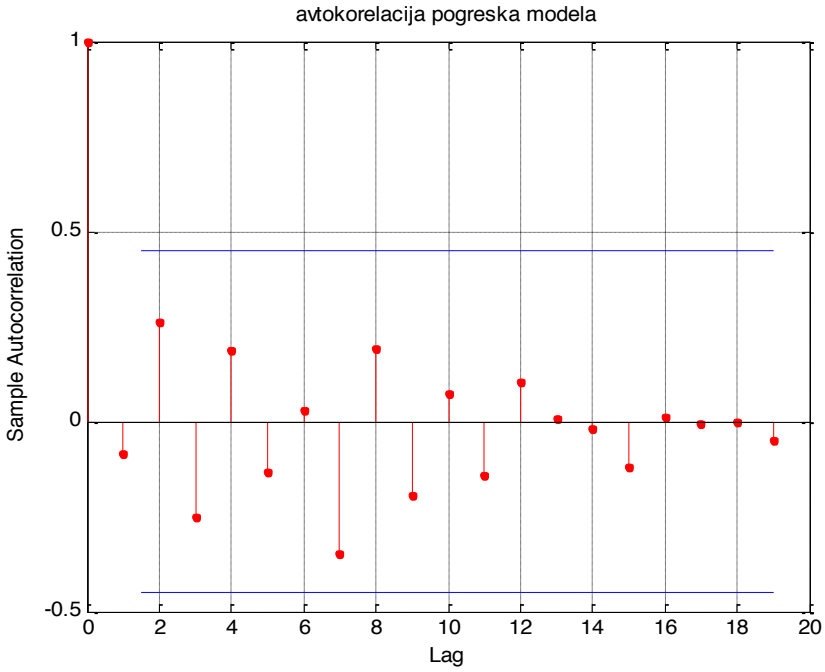
Slika 278: q-q diagram pogreška



Slika 279: Cookove razdalje za pogrešek



Slika 280: Histogram pogreška in ocenjena teoretična porazdelitev



Slika 281: Avtokorelacija pogreška

Primer 9.22.:

Sladkorna bolezen je stanje, označeno s hiperglikemijo, ki izhaja iz nezmožnosti telesa uporabiti glukoze v krvi za energijo. Pri sladkorni bolezni tipa 1, v trebušni slinavki ne bo več insulina in zato glukoza v krvi ne more vstopiti v celice, ki se uporabljajo za proizvodnjo energije. Cilj je bil raziskati odvisnost višine seruma C peptida na različnih drugih dejavnikih, da bi razumeli vzorce izločanja insulina. Podatki na sliki 282 v okviru tovrstnih raziskav prikazujejo odvisnost log C peptida (y) od deficita (x) kot mere kislosti. [Vidakovic]

Deficit (x)	−8.1	−16.1	−0.9	−7.8	−29.0	−19.2	−18.9	−10.6	−2.8	−25.0	−3.1
Log C-peptide (y)	4.8	4.1	5.2	5.5	5	3.4	3.4	4.9	5.6	3.7	3.9
Deficit (x)	−7.8	−13.9	−4.5	−11.6	−2.1	−2.0	−9.0	−11.2	−0.2	−6.1	−1
Log C-peptide (y)	4.5	4.8	4.9	3.0	4.6	4.8	5.5	4.5	5.3	4.7	6.6
Deficit (x)	−3.6	−8.2	−0.5	−2.0	−1.6	−11.9	−0.7	−1.2	−14.3	−0.8	−16.8
Log C-peptide (y)	5.1	3.9	5.7	5.1	5.2	3.7	4.9	4.8	4.4	5.2	5.1
Deficit (x)	−5.1	−9.5	−17.0	−3.3	−0.7	−3.3	−13.6	−1.9	−10.0	−13.5	
Log C-peptide (y)	4.6	3.9	5.1	5.1	6.0	4.9	4.1	4.6	4.9	5.1	

Slika 282: Odvisnost log C peptida (y) od deficita (x) kot mere kislosti [Vidakovic]

Uporabimo program **regres_mnk1.m** za izračune. Izpis rezultatov v komandnem oknu je naslednji:

```
izberi primere 1, 2, 3, 43
xstr =
deficit
ystr =
log C peptid

n =
    43

Sxy =
  105.3477
Sxx =
  2.1310e+003
Syy =
   21.8070
ysr =
    4.7465
xsr =
   -8.1488
```

```
alfa=
alfa =
alfa =
    0.0500

stevilo monte carlo replikacij za jb test100
nrep =
   100

boc=
boc =
    0.0494
aoc=
aoc =
    5.1494

standardna ocena napake modela:
se =
    0.6363
SSR =
    5.2079
SSE =
   16.5990
SST =
   21.8070

F =
   12.8637
Fkrit =
    4.0785

Zavrni nicelno hipotezo, da je b = 0, torej je model premice signifikanten
p vrednost je:
p =
  8.8412e-004

korelacijski koeficient:
ro =
    0.4887
determinacijski koeficient:
D =
    0.2388

%------------------------------------------------------------------------
% Izracun s standarnim regress ukazom:
%------------------------------------------------------------------------

n =
    43

ocenjena parametra sta:
aoc =
    5.1494
boc =
```

```
    0.0494

intervala zaupanja za ocenjena parametra:
Iaoc =
    4.8496    5.4491
Iboc =
    0.0216    0.0773

determinacijski koeficient:
D =
    0.2388

x    spodnja meja pogreska    pogresek    zgornja meja pogreska:
ans =
   -8.1000   -1.2346    0.0511    1.3367
  -16.1000   -1.5169   -0.2534    1.0100
   -0.9000   -1.1739    0.0951    1.3642
   -7.8000   -0.5278    0.7362    2.0003
  -29.0000    0.2167    1.2843    2.3518
  -19.2000   -2.0212   -0.8002    0.4208
  -18.9000   -2.0372   -0.8150    0.4071
  -10.6000   -1.0062    0.2747    1.5556
   -2.8000   -0.6739    0.5891    1.8520
  -25.0000   -1.4063   -0.2135    0.9794
   -3.1000   -2.3251   -1.0961    0.1329
   -7.8000   -1.5467   -0.2638    1.0192
  -13.9000   -0.9331    0.3378    1.6087
   -4.5000   -1.3085   -0.0269    1.2547
  -11.6000   -2.7551   -1.5759   -0.3967
   -2.1000   -1.7120   -0.4455    0.8209
   -2.0000   -1.5220   -0.2505    1.0210
   -9.0000   -0.4646    0.7956    2.0557
  -11.2000   -1.3782   -0.0957    1.1868
   -0.2000   -1.1045    0.1605    1.4256
   -6.1000   -1.4314   -0.1478    1.1358
   -1.0000    0.3240    1.5001    2.6761
   -3.6000   -1.1501    0.1286    1.4073
   -8.2000   -2.1012   -0.8440    0.4132
   -0.5000   -0.6788    0.5754    1.8295
   -2.0000   -1.2244    0.0495    1.3234
   -1.6000   -1.1420    0.1297    1.4015
  -11.9000   -2.1126   -0.8611    0.3905
   -0.7000   -1.4814   -0.2147    1.0519
   -1.2000   -1.5574   -0.2900    0.9773
  -14.3000   -1.3164   -0.0424    1.2315
   -0.8000   -1.1784    0.0902    1.3588
  -16.8000   -0.4564    0.7812    2.0187
   -5.1000   -1.5766   -0.2972    0.9821
   -9.5000   -2.0405   -0.7797    0.4811
  -17.0000   -0.4447    0.7911    2.0268
   -3.3000   -1.1642    0.1138    1.3917
   -0.7000   -0.3513    0.8853    2.1218
   -3.3000   -1.3644   -0.0862    1.1919
  -13.6000   -1.6479   -0.3770    0.8938
```

```
   -1.9000   -1.7207   -0.4554    0.8099
  -10.0000   -1.0373    0.2450    1.5273
  -13.5000   -0.6435    0.6180    1.8796

Standardna ocena napake modela je:
se =
    0.6363
SSE =
   16.5990
Sxy =
  105.3477
SSR =
    5.2079
SST =
   21.8070

F vrednost je (1. nacin - iz regress):
F =
   12.8637
F vrednost je (2. nacin - rocno):
F =
   12.8637

p vrednost je - iz regress:
p =
  8.8412e-004
p vrednost je - na 2. nacin:
p =
  8.8412e-004

Fkrit =
    4.0785

Zavrni nicelno hipotezo, da je b = 0, torej je model premice signifikanten

primerjava standardiziranih residualov in normiranih pogreskov:
ans =
    0.0812    0.0803
   -0.4093   -0.3983
    0.1532    0.1495
    1.1708    1.1571
    2.2961    2.0184
   -1.3115   -1.2576
   -1.3336   -1.2809
    0.4374    0.4317
    0.9433    0.9258
   -0.3653   -0.3355
   -1.7538   -1.7227
   -0.4194   -0.4145
    0.5415    0.5309
   -0.0429   -0.0423
   -2.5132   -2.4767
   -0.7148   -0.7002
   -0.4020   -0.3937
```

```
    1.2654    1.2503
   -0.1525   -0.1504
    0.2592    0.2523
   -0.2353   -0.2323
    2.4153    2.3576
    0.2055    0.2021
   -1.3421   -1.3264
    0.9281    0.9043
    0.0795    0.0778
    0.2085    0.2039
   -1.3740   -1.3533
   -0.3461   -0.3375
   -0.4667   -0.4558
   -0.0681   -0.0667
    0.1453    0.1418
    1.2652    1.2277
   -0.4737   -0.4671
   -1.2405   -1.2254
    1.2823    1.2432
    0.1820    0.1788
    1.4269    1.3913
   -0.1379   -0.1355
   -0.6039   -0.5926
   -0.7311   -0.7158
    0.3899    0.3851
    0.9896    0.9713


s =
    0.0406
k =
    3.1651
JB =
    0.0606
h =
     0
p =
    0.9650
skoraj gotovo je e normalen
srednja vrednost pogreska
ans =
  5.6802e-016
stand. deviacija pogreska:
ans =
    0.6287


arch test za pogreseek:
H,P,Stat,CV
ans =
         0    0.2870    1.1334    2.7055
         0    0.4610    1.5486    4.6052
         0    0.4338    2.7381    6.2514
         0    0.6112    2.6889    7.7794
         0    0.5081    4.2926    9.2364
         0    0.5785    4.7325   10.6446
```

```
     0   0.4237   7.0495   12.0170
     0   0.5504   6.8731   13.3616
     0   0.6708   6.6761   14.6837
     0   0.7157   7.1032   15.9872
     0   0.6039   9.1953   17.2750
     0   0.7197   8.8029   18.5493
     0   0.7287   9.5698   19.8119
     0   0.7487  10.1823   21.0641
     0   0.6037  12.9819   22.3071
     0   0.6835  12.8529   23.5418
     0   0.6878  13.7055   24.7690
     0   0.7252  14.0589   25.9894
     0   0.7752  14.1452   27.2036
     0   0.6489  17.0624   28.4120
     0   0.3995  22.0000   29.6151
     0   0.5207  21.0000   30.8133
     0   0.6419  20.0000   32.0069
     0   0.7520  19.0000   33.1962
     0   0.8424  18.0000   34.3816
     0   0.9091  17.0000   35.5632
     0   0.9529  16.0000   36.7412
     0   0.9784  15.0000   37.9159
     0   0.9914  14.0000   39.0875
     0   0.9970  13.0000   40.2560
     0   0.9991  12.0000   41.4217
     0   0.9998  11.0000   42.5847
     0   1.0000  10.0000   43.7452
     0   1.0000   9.0000   44.9032
     0   1.0000   8.0000   46.0588
     0   1.0000   7.0000   47.2122
     0   1.0000   6.0000   48.3634
     0   1.0000   5.0000   49.5126
     0   1.0000   4.0000   50.6598
     0   1.0000   3.0000   51.8051
     0   1.0000   2.0000   52.9485

ljung-box test za pogresek:
H,P,Qstat,CV
ans =
     0   0.7593   0.0939   2.7055
  1.0000   0.0738   5.2128   4.6052
  1.0000   0.0636   7.2763   6.2514
  1.0000   0.0128  12.6993   7.7794
  1.0000   0.0116  14.7309   9.2364
  1.0000   0.0079  17.4177  10.6446
  1.0000   0.0037  21.0224  12.0170
  1.0000   0.0065  21.2558  13.3616
  1.0000   0.0059  23.1436  14.6837
  1.0000   0.0048  25.2740  15.9872
  1.0000   0.0070  25.7975  17.2750
  1.0000   0.0077  27.0042  18.5493
  1.0000   0.0040  30.4662  19.8119
  1.0000   0.0060  30.7489  21.0641
  1.0000   0.0083  31.1938  22.3071
```

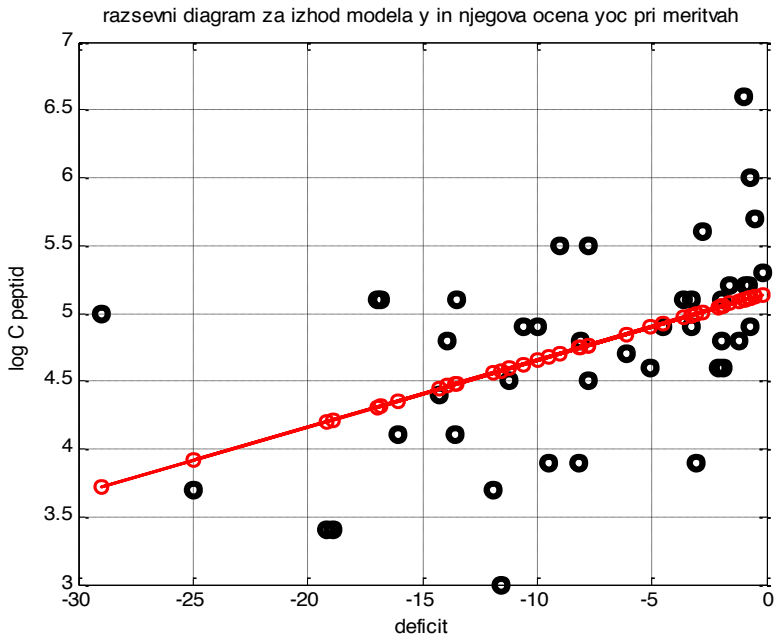
```
1.0000    0.0127   31.2086   23.5418
1.0000    0.0107   33.1774   24.7690
1.0000    0.0137   33.7124   25.9894
1.0000    0.0149   34.7651   27.2036
1.0000    0.0174   35.5510   28.4120
1.0000    0.0244   35.5778   29.6151
1.0000    0.0323   35.7438   30.8133
1.0000    0.0348   36.7170   32.0069
1.0000    0.0417   37.2048   33.1962
1.0000    0.0551   37.2142   34.3816
1.0000    0.0602   38.0305   35.5632
1.0000    0.0574   39.4716   36.7412
1.0000    0.0648   40.0977   37.9159
1.0000    0.0687   41.0089   39.0875
1.0000    0.0785   41.5282   40.2560
1.0000    0.0844   42.3327   41.4217
1.0000    0.0938   42.9335   42.5847
1.0000    0.0785   45.0645   43.7452
1.0000    0.0731   46.6286   44.9032
1.0000    0.0902   46.6407   46.0588
1.0000    0.0622   49.8516   47.2122
1.0000    0.0734   50.1258   48.3634
1.0000    0.0696   51.5926   49.5126
1.0000    0.0772   52.1721   50.6598
1.0000    0.0941   52.1737   51.8051
     0    0.1120   52.2467   52.9485
     0    0.1334   52.2537   54.0902
```

Dobimo torej:

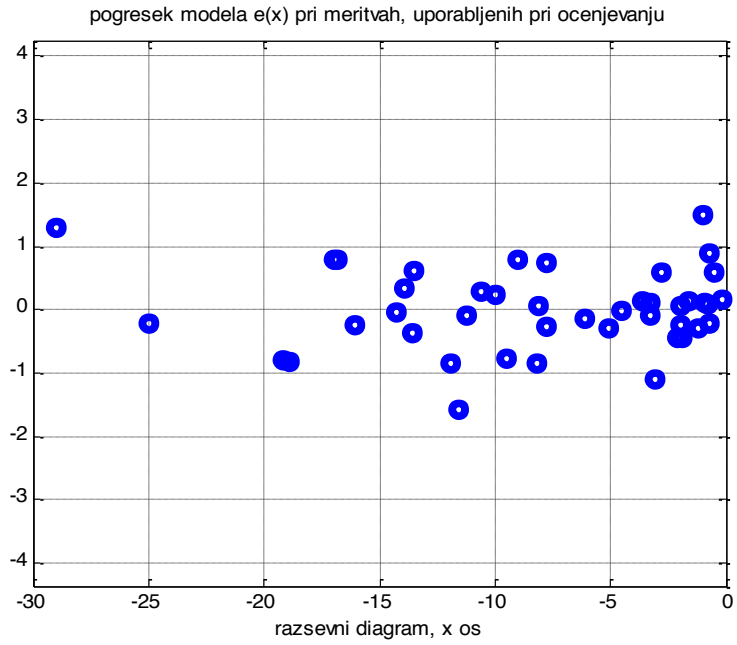
$$
\begin{aligned}
&n = 43 \\
&S_{xy} = 105.3477,\ S_{xx} = 2.1310e{+}003,\ S_{yy} = 21.8070 \\
&\bar{y} = 4.7465,\ \bar{x} = -8.1488 \\
&\alpha = 0.05 \\
&n_{rep} = 100\ (za\ Monte\ Carlo\ pri\ JB\ testu) \\
&\hat{b} = 0.0494,\quad \hat{a} = 5.1494 \\
&s_e = 0.6363 \\
&SSR = 5.2079,\quad SSE = 16.5990,\quad SST = 21.8070 \\
&F = 12.8637 > F_{krit} = 4.0785\ \left(p = 8.8412e{-}004\right) \\
&mod\ el\ \ premice\ \ je\ signifikanten \\
&r = 0.4887,\quad D = r^2 = 0.2388 \\
&I_a = \left[4.8496, 5.4491\right] \\
&I_b = \left[0.0216, 0.0773\right] \\
&S = 0.0406,\quad K = 3.1651,\quad JB = 0.0606 \\
&za\ \ JB:\ \ h = 0,\ \ p = 0.9650 \rightarrow skoraj\ \ gotovo\ \ \ je\ \ e\ \ normalen
\end{aligned}
\tag{9.249}
$$

Kot vidimo, sta korelacijski in determinacijski koeficient relativno nizka, vendar kljub temu F test nakaže, da je model premice signifikanten. Tudi test za arch potrdi, da je model ustrezen, saj da v 1. stolpu same ničle, v drugem stolpu pa so p vrednosti dovolj signifikantne. Pri Ljung Box testu se sicer v 1. stolpu pojavijo pretežno enke in so p vrednosti dokaj nizke, vendar je ta, kot smo že dejali, bolj merodajen za testiranje modelov časovnih vrst.

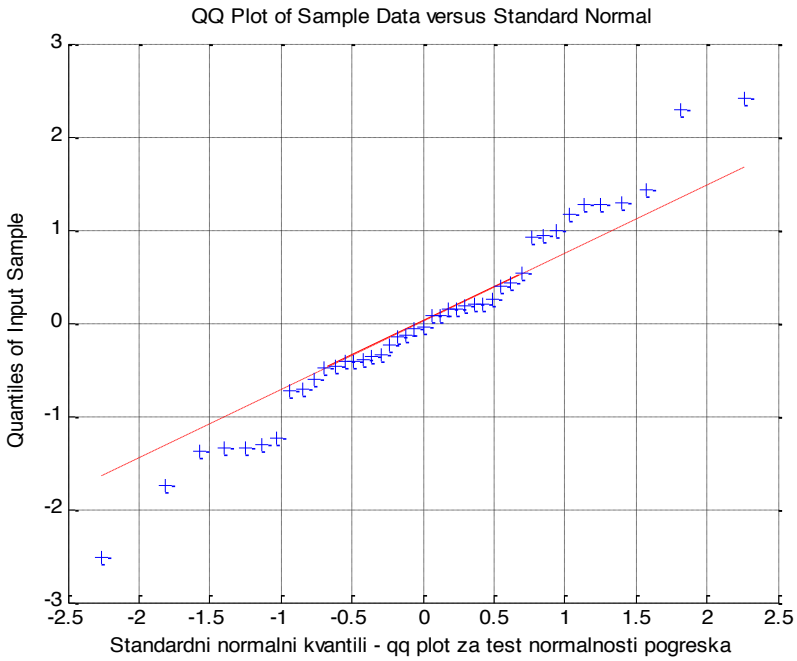
Slike 283 do 288 prikazujejo: podatke in ocenjeno regresijsko premico, razsevni diagram pogreška modela, q-q diagram za pogrešek, Cookove razdalje za pogrešek, histogram pogreška in avtokorelacijo pogreška. Tudi na osnovi teh slik sklepamo, da je uporabljeni model ustrezen.



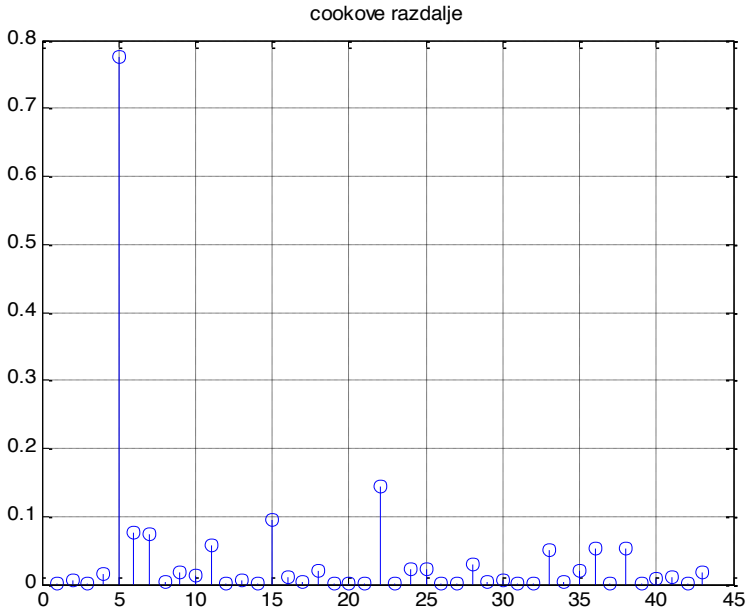
Slika 283: Podatki in ocenjena regresijska premica



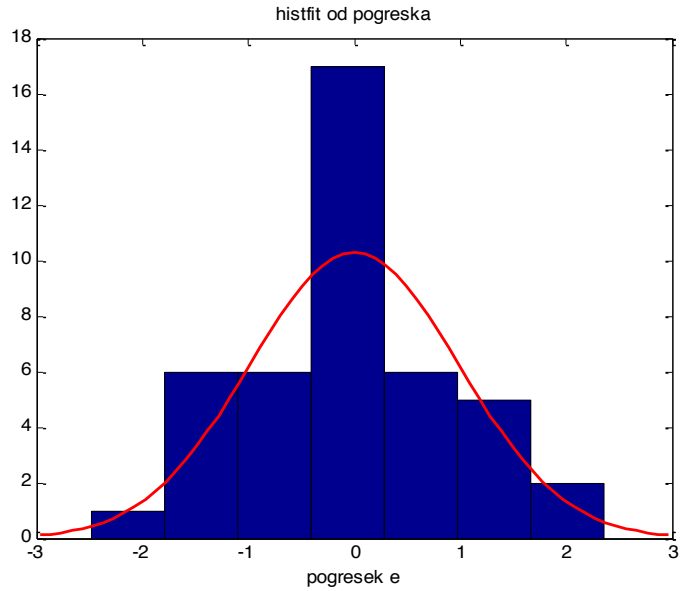
Slika 284: Razsevni diagram pogreška



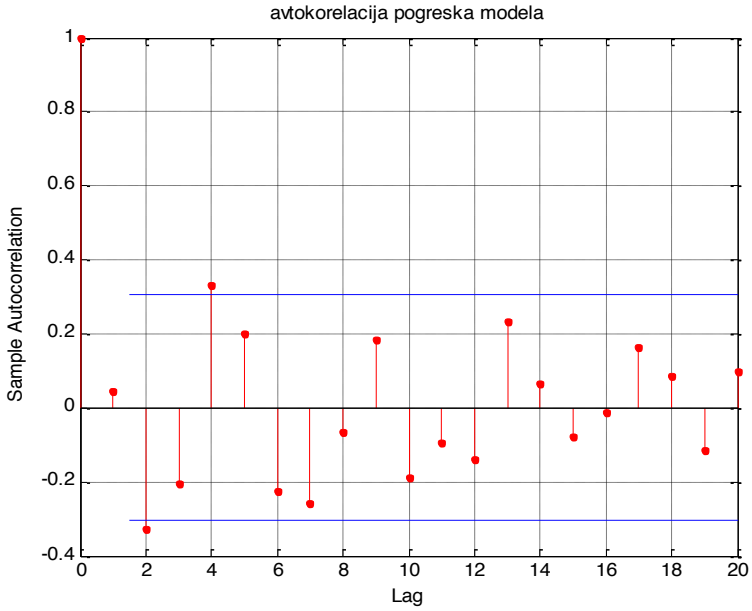
Slika 285: q-q diagram pogreška



Slika 286: Cookove razdalje za pogrešek



Slika 287: Histogram pogreška in ocenjena teoretična porazdelitev



Slika 288: Avtokorelacija pogreška

Primer 9.23.:

Slika 289 prikazuje porabo denarja (v rupijah) za 55 indijskih ruralnih gospodinjstev za hrano (food expenditure) in za celotne stroške (total expenditure) na mesec. Zanima nas, koliko informacije o stroških za hrano (odvisna spremenljivka y) je vsebovano v celotnih stroških (neodvisna spremenljivka x) [Gujarati].

TABLE 2.8 FOOD AND TOTAL EXPENDITURE (RUPEES)

Observation	Food expenditure	Total expenditure	Observation	Food expenditure	Total expenditure
1	217.0000	382.0000	29	390.0000	655.0000
2	196.0000	388.0000	30	385.0000	662.0000
3	303.0000	391.0000	31	470.0000	663.0000
4	270.0000	415.0000	32	322.0000	677.0000
5	325.0000	456.0000	33	540.0000	680.0000
6	260.0000	460.0000	34	433.0000	690.0000
7	300.0000	472.0000	35	295.0000	695.0000
8	325.0000	478.0000	36	340.0000	695.0000
9	336.0000	494.0000	37	500.0000	695.0000
10	345.0000	516.0000	38	450.0000	720.0000
11	325.0000	525.0000	39	415.0000	721.0000
12	362.0000	554.0000	40	540.0000	730.0000
13	315.0000	575.0000	41	360.0000	731.0000
14	355.0000	579.0000	42	450.0000	733.0000
15	325.0000	585.0000	43	395.0000	745.0000
16	370.0000	586.0000	44	430.0000	751.0000
17	390.0000	590.0000	45	332.0000	752.0000
18	420.0000	608.0000	46	397.0000	752.0000
19	410.0000	610.0000	47	446.0000	769.0000
20	383.0000	616.0000	48	480.0000	773.0000
21	315.0000	618.0000	49	352.0000	773.0000
22	267.0000	623.0000	50	410.0000	775.0000
23	420.0000	627.0000	51	380.0000	785.0000
24	300.0000	630.0000	52	610.0000	788.0000
25	410.0000	635.0000	53	530.0000	790.0000
26	220.0000	640.0000	54	360.0000	795.0000
27	403.0000	648.0000	55	305.0000	801.0000
28	350.0000	650.0000			

Source: Chandan Mukherjee, Howard White, and Marc Wuyts, *Econometrics and Data Analysis for Developing Countries,* Routledge, New York, 1998, p. 457.

Slika 289: Poraba denarja (v rupijah) za 55 indijskih ruralnih gospodinjstev za hrano (food expenditure) - spremenljivka y in za celotne stroške (total expenditure) - spremenljivka x na mesec. [Gujarati]

Uporabimo program **regres_mnk1.m** za izračune. Izpis rezultatov v komandnem oknu je naslednji:

```
izberi primere 1, 2, 3, 44
ystr =
izdatki za hrano (v tisočih rupijah)
xstr =
celotni izdatki (v tisočih rupijah)
n =
    55

Sxy =
  3.1827e+005
Sxx =
  7.2862e+005
Syy =
  3.7592e+005
ysr =
  373.3455
xsr =
  639.0364

alfa=
alfa =
alfa =
    0.0500

stevilo monte carlo replikacij za jb test100
nrep =
   100

boc=
boc =
    0.4368
aoc=
aoc =
   94.2088

standardna ocena napake modela:
se =
   66.8557
SSR =
  1.3902e+005
SSE =
  2.3689e+005
SST =
  3.7592e+005

F =
   31.1035
Fkrit =
    4.0230

Zavrni nicelno hipotezo, da je b = 0, torej je model premice signifikanten
p vrednost je:
p =
```

```
  8.4513e-007

korelacijski koeficient:
ro =
    0.6081
determinacijski koeficient:
D =
    0.3698

%------------------------------------------------------------------------
% Izracun s standarnim regress ukazom:
%------------------------------------------------------------------------

n =
    55
ocenjena parametra sta:
aoc =
   94.2088
boc =
    0.4368

intervala zaupanja za ocenjena parametra:
Iaoc =
   -7.7961  196.2137
Iboc =
    0.2797    0.5939

determinacijski koeficient:
D =
    0.3698

x    spodnja meja pogreska    pogresek     zgornja meja pogreska:
ans =
  382.0000 -171.2787  -44.0697   83.1393
  388.0000 -194.3972  -67.6906   59.0161
  391.0000  -89.8095   37.9990  165.8075
  415.0000 -134.8265   -5.4844  123.8577
  456.0000  -99.0620   31.6064  162.2748
  460.0000 -165.8787  -35.1408   95.5971
  472.0000 -131.8831   -0.3825  131.1180
  478.0000 -109.5499   21.9966  153.5432
  494.0000 -105.9498   26.0077  157.9652
  516.0000 -107.1296   25.3979  157.9254
  525.0000 -131.4504    1.4666  134.3837
  554.0000 -107.4706   25.7992  159.0689
  575.0000 -163.8642  -30.3738  103.1166
  579.0000 -125.9073    7.8790  141.6652
  585.0000 -158.4334  -24.7419  108.9496
  586.0000 -113.9437   19.8213  153.5863
  590.0000  -95.4234   38.0741  171.5716
  608.0000  -72.7904   60.2115  193.2134
  610.0000  -84.0213   49.3379  182.6971
  616.0000 -114.2635   19.7170  153.6976
  618.0000 -182.5588  -49.1566   84.2456
```

```
  623.0000 -230.5819  -99.3406   31.9006
  627.0000  -81.4373   51.9121  185.2616
  630.0000 -202.1370  -69.3983   63.3404
  635.0000  -95.2970   38.4177  172.1324
  640.0000 -280.9078 -153.7664  -26.6249
  648.0000 -108.2046   25.7392  159.6829
  650.0000 -162.0372  -28.1345  105.7683
  655.0000 -124.4101    9.6815  143.7731
  662.0000 -132.4685    1.6238  135.7161
  663.0000  -45.7413   86.1870  218.1153
  677.0000 -200.5970  -67.9283   64.7404
  680.0000   21.3254  148.7613  276.1972
  690.0000  -96.1010   37.3932  170.8874
  695.0000 -233.5502 -102.7909   27.9684
  695.0000 -190.6707  -57.7909   75.0890
  695.0000  -28.5854  102.2091  233.0037
  720.0000  -91.7428   41.2889  174.3207
  721.0000 -127.6491    5.8521  139.3534
  730.0000   -1.6864  126.9208  255.5280
  731.0000 -186.0298  -53.5160   78.9978
  733.0000  -97.3333   35.6104  168.5541
  745.0000 -157.5405  -24.6313  108.2779
  751.0000 -125.1967    7.7478  140.6924
  752.0000 -221.2146  -90.6890   39.8367
  752.0000 -158.4375  -25.6890  107.0596
  769.0000 -116.5905   15.8853  148.3611
  773.0000  -83.6327   48.1381  179.9088
  773.0000 -210.4352  -79.8619   50.7113
  775.0000 -154.9824  -22.7356  109.5113
  785.0000 -188.2754  -57.1036   74.0681
  788.0000   48.4670  171.5859  294.7049
  790.0000  -38.8421   90.7123  220.2667
  795.0000 -211.3518  -81.4717   48.4084
  801.0000 -264.9402 -139.0926  -13.2450

Standardna ocena napake modela je:
se =
   66.8557
SSE =
  2.3689e+005
Sxy =
  3.1827e+005
SSR =
  1.3902e+005
SST =
  3.7592e+005

F vrednost je (1. nacin - iz regress):
F =
   31.1035
F vrednost je (2. nacin - rocno):
F =
   31.1035
```

```
p vrednost je - iz regress:
p =
  8.4513e-007
p vrednost je - na 2. nacin:
p =
  8.4513e-007


Fkrit =
    4.0230


Zavrni nicelno hipotezo, da je b = 0, torej je model premice signifikanten


primerjava standardiziranih residualov in normiranih pogreskov:
ans =
   -0.6983   -0.6592
   -1.0700   -1.0125
    0.6000    0.5684
   -0.0859   -0.0820
    0.4887    0.4728
   -0.5428   -0.5256
   -0.0059   -0.0057
    0.3382    0.3290
    0.3985    0.3890
    0.3875    0.3799
    0.0223    0.0219
    0.3914    0.3859
   -0.4598   -0.4543
    0.1192    0.1179
   -0.3743   -0.3701
    0.2998    0.2965
    0.5757    0.5695
    0.9095    0.9006
    0.7452    0.7380
    0.2977    0.2949
   -0.7423   -0.7353
   -1.4999   -1.4859
    0.7837    0.7765
   -1.0477   -1.0380
    0.5799    0.5746
   -2.3212   -2.3000
    0.3886    0.3850
   -0.4247   -0.4208
    0.1462    0.1448
    0.0245    0.0243
    1.3016    1.2891
   -1.0264   -1.0160
    2.2483    2.2251
    0.5655    0.5593
   -1.5551   -1.5375
   -0.8743   -0.8644
    1.5463    1.5288
    0.6261    0.6176
    0.0888    0.0875
    1.9271    1.8984
```

```
   -0.8127   -0.8005
    0.5409    0.5326
   -0.3748   -0.3684
    0.1180    0.1159
   -1.3814   -1.3565
   -0.3913   -0.3842
    0.2427    0.2376
    0.7360    0.7200
   -1.2210   -1.1945
   -0.3477   -0.3401
   -0.8751   -0.8541
    2.6313    2.5665
    1.3917    1.3568
   -1.2513   -1.2186
   -2.1393   -2.0805


s =
    0.1198
k =
    3.2345
JB =
    0.2576
h =
     0
p =
    0.8770
skoraj gotovo je e normalen

srednja vrednost pogreska
ans =
 -5.0642e-014
stand. deviacija pogreska:
ans =
   66.2338

arch test za pogreseek:
H,P,Stat,CV
ans =
         0    0.7873    0.0728    2.7055
         0    0.5732    1.1129    4.6052
         0    0.3798    3.0778    6.2514
         0    0.4456    3.7173    7.7794
         0    0.6056    3.6183    9.2364
         0    0.5311    5.0998   10.6446
         0    0.1888    9.9962   12.0170
         0    0.2542   10.1574   13.3616
         0    0.3529    9.9696   14.6837
         0    0.4474    9.9220   15.9872
         0    0.5576    9.6997   17.2750
         0    0.6049   10.1258   18.5493
         0    0.7085    9.8206   19.8119
         0    0.7917    9.5869   21.0641
         0    0.6551   12.3145   22.3071
         0    0.6567   13.2185   23.5418
```

```
     0   0.7151   13.3112   24.7690
     0   0.7927   12.9814   25.9894
     0   0.6459   16.1695   27.2036
     0   0.6902   16.4203   28.4120
     0   0.6055   18.6827   29.6151
     0   0.6919   18.2358   30.8133
     0   0.7688   17.7905   32.0069
     0   0.7620   18.8118   33.1962
     0   0.7156   20.5832   34.3816
     0   0.5831   23.8757   35.5632
     0   0.4110   28.0000   36.7412
     0   0.5182   27.0000   37.9159
     0   0.6255   26.0000   39.0875
     0   0.7250   25.0000   40.2560
     0   0.8105   24.0000   41.4217
     0   0.8783   23.0000   42.5847
     0   0.9276   22.0000   43.7452
     0   0.9604   21.0000   44.9032
     0   0.9802   20.0000   46.0588
     0   0.9911   19.0000   47.2122
     0   0.9964   18.0000   48.3634
     0   0.9987   17.0000   49.5126
     0   0.9996   16.0000   50.6598
     0   0.9999   15.0000   51.8051
     0   1.0000   14.0000   52.9485
     0   1.0000   13.0000   54.0902
     0   1.0000   12.0000   55.2302
     0   1.0000   11.0000   56.3685
     0   1.0000   10.0000   57.5053
     0   1.0000    9.0000   58.6405
     0   1.0000    8.0000   59.7743
     0   1.0000    7.0000   60.9066
     0   1.0000    6.0000   62.0375
     0   1.0000    5.0000   63.1671
     0   1.0000    4.0000   64.2954
     0   1.0000    3.0000   65.4224
     0   1.0000    2.0000   66.5482


ljung-box test za pogresek:
H,P,Qstat,CV
ans =
     0   0.5094    0.4352    2.7055
     0   0.6684    0.8058    4.6052
     0   0.1848    4.8291    6.2514
     0   0.1713    6.3989    7.7794
     0   0.2203    7.0048    9.2364
     0   0.2194    8.2636   10.6446
     0   0.1588   10.5648   12.0170
     0   0.1993   11.0427   13.3616
     0   0.2594   11.2432   14.6837
     0   0.2897   11.9304   15.9872
     0   0.3486   12.2024   17.2750
     0   0.2919   14.1390   18.5493
     0   0.3062   15.0192   19.8119
```

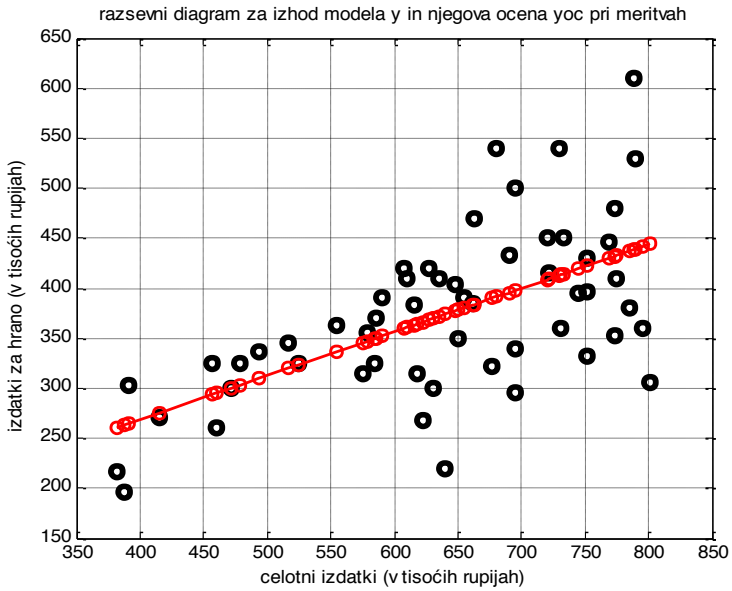
```
     0    0.3768   15.0202   21.0641
     0    0.3680   16.2137   22.3071
     0    0.2312   19.7609   23.5418
     0    0.2223   21.0897   24.7690
1.0000    0.0895   26.4673   25.9894
1.0000    0.0215   33.4240   27.2036
1.0000    0.0227   34.5362   28.4120
1.0000    0.0313   34.5864   29.6151
1.0000    0.0426   34.6035   30.8133
1.0000    0.0446   35.6659   32.0069
1.0000    0.0590   35.6749   33.1962
1.0000    0.0704   36.0766   34.3816
1.0000    0.0600   38.0428   35.5632
1.0000    0.0685   38.6281   36.7412
1.0000    0.0790   39.1190   37.9159
1.0000    0.0799   40.2521   39.0875
1.0000    0.0998   40.2667   40.2560
1.0000    0.0920   41.8747   41.4217
     0    0.1059   42.2694   42.5847
     0    0.1039   43.5325   43.7452
1.0000    0.0911   45.4253   44.9032
     0    0.1110   45.4565   46.0588
     0    0.1229   45.9952   47.2122
     0    0.1356   46.5232   48.3634
     0    0.1375   47.5639   49.5126
     0    0.1618   47.6259   50.6598
     0    0.1902   47.6261   51.8051
     0    0.2211   47.6265   52.9485
     0    0.2544   47.6288   54.0902
     0    0.2713   48.1791   55.2302
     0    0.2956   48.5233   56.3685
     0    0.3329   48.5233   57.5053
     0    0.2821   51.0429   58.6405
     0    0.3169   51.0692   59.7743
     0    0.3368   51.5462   60.9066
     0    0.3168   53.1691   62.0375
     0    0.3174   54.1989   63.1671
     0    0.2389   57.7899   64.2954
     0    0.2621   58.0543   65.4224
     0    0.1683   62.7790   66.5482
     0    0.1476   64.8780   67.6728
```

Dobimo torej:

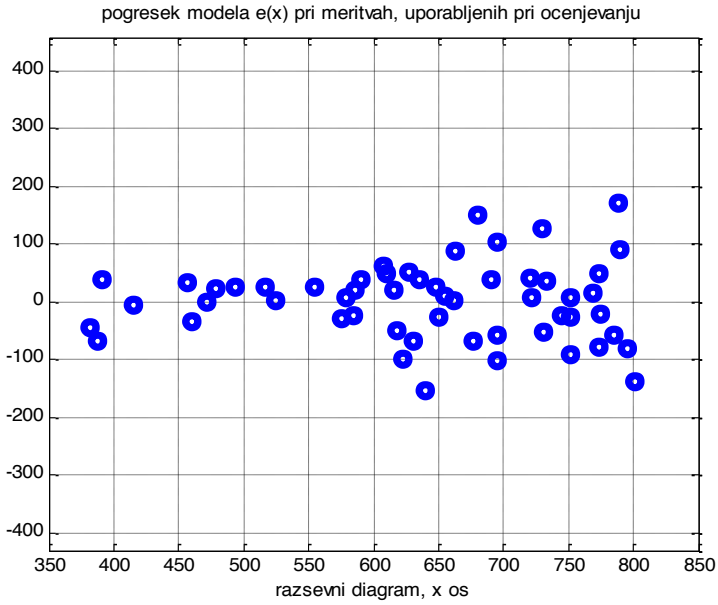
$$
\begin{aligned}
&n = 55 \\
&S_{xy} = 3.1827e{+}005,\ S_{xx} = 7.2862e{+}005,\ \ S_{yy} = 3.7592e{+}005 \\
&\bar{y} = 373.3455,\ \bar{x} = 639.0364 \\
&\alpha = 0.05 \\
&n_{rep} = 100\ (za\ Monte\ Carlo\ \ pri\ JB\ \ testu) \\
&\hat{b} = 0.4368,\quad \hat{a} = 94.2088 \\
&s_e = 66.8557 \\
&SSR = 1.3902e{+}005,\quad SSE = 2.3689e{+}005,\quad SST = 3.7592e{+}005 \\
&F = 31.1035 > F_{krit} = 4.0230\ \ \left(p = 8.4513e\text{-}007\right) \\
&mod\ el\ \ premice\ \ je\ signifikanten \\
&r = 0.6081,\qquad D = r^2 = 0.3698 \\
&I_a = \left[-7.7961, 196.2137\right] \\
&I_b = \left[0.2797, 0.5939\right] \\
&S = 0.1198,\quad K = 3.2345,\quad JB = 0.2576 \\
&za\ \ JB:\ \ h = 0,\ \ p = 0.8770\ \rightarrow skoraj\ \ gotovo\ \ \ je\ \ e\ \ normalen
\end{aligned}
\tag{9.250}
$$

Kot nam nakazuje determinacijski koeficient, je le cca. 37% variabilnosti v porabi hrane pojasneno v celotni porabi gospodinjstev. Glede na to, da gre za takoimenovane **cross-sectional podatke**, to ni nič nenavadnega, saj je razlog ponavadi v raznolikosti enot v vzorcu [Gujarati]. F test nakaže, da je model premice signifikanten. Tudi test za arch potrdi, da je model ustrezen, saj da v 1. stolpu same ničle, v drugem stolpu pa so p vrednosti dovolj signifikantne. Pri Ljung Box testu se sicer v 1. stolpu pojavijo nekatere enke in so p vrednosti pri njih dokaj nizke, vendar je ta, kot smo že dejali, bolj merodajen za testiranje modelov časovnih vrst, kot pa cross-sectional modelov.

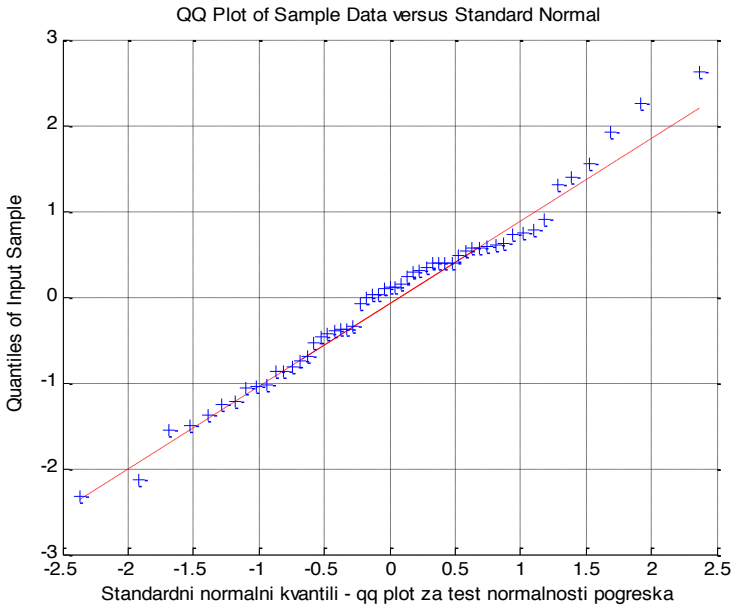
Slike 290 do 295 prikazujejo: podatke in ocenjeno regresijsko premico, razsevni diagram pogreška modela, q-q diagram za pogrešek, Cookove razdalje za pogrešek, histogram pogreška in avtokorelacijo pogreška. Tudi na osnovi teh slik sklepamo, da je uporabljeni model ustrezen.



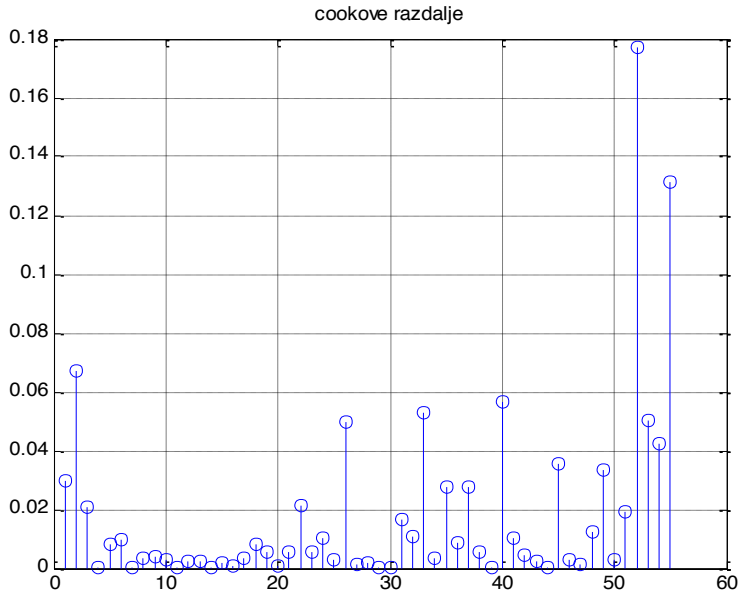
Slika 290: Podatki in ocenjena regresijska premica



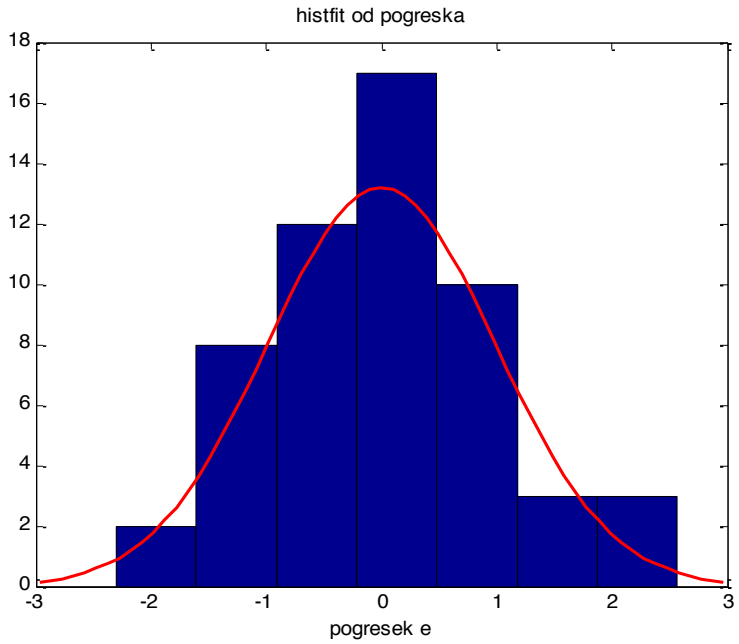
Slika 291: Razsevni diagram pogreška



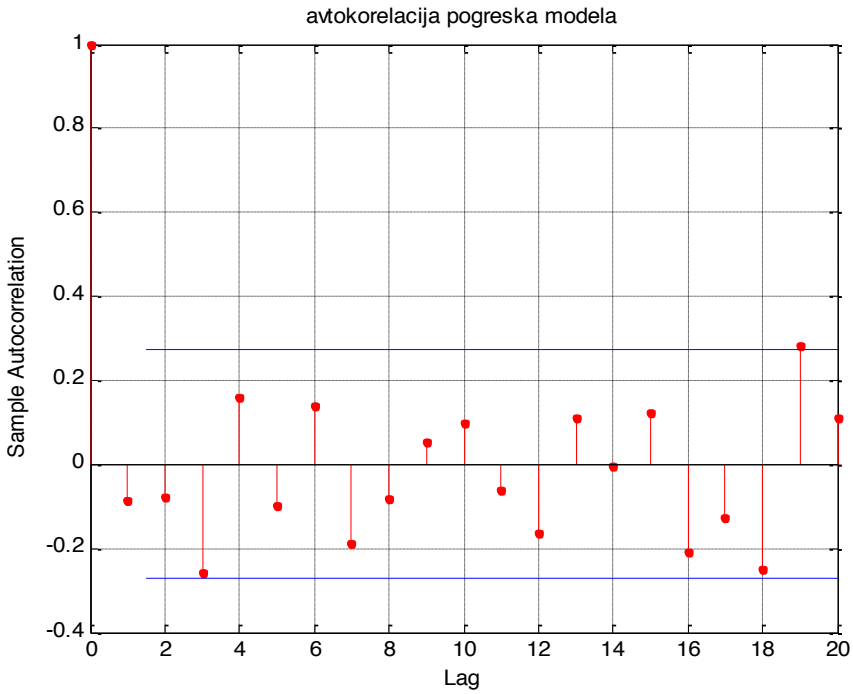
Slika 292: q-q diagram pogreška



Slika 293: Cookove razdalje za pogrešek



Slika 294: Histogram pogreška in ocenjena teoretična porazdelitev



Slika 295: Avtokorelacija pogreška

9.4.9 Sklep enostavnih regresijskih modelov v obliki regresijske premice

V dokaj obsežnem poglavju 9.4 smo natančno predstavili pomen enostavnih regresijskih modelov v obliki regresijske premice. Pojasnili smo tudi vsa teoretična statistična ozadja, ki veljajo za tovrstne modele, ter razložili temeljne principe. Pomen uporabe teh modelov smo ilustrirali na številnih primerih, tako splošnih, kot tudi bolj specializiranih s področja logistike, organizacije dela in operacijskih raziskav. Prav tako smo pri večini primerov ponazorili, kako se da raznovrstne probleme učinkovito reševati s pomočjo programskega orodja Matlab. Še posebej pa je razumevanje principov uporabe enostavnih regresijskih modelov v obliki regresijske premice pomembno za kasnejšo nadgradnjo snovi iz teorije regresijskih modelov, kjer nastopijo bolj kompleksni modeli in primeri regresij, kot npr.: modeli multiple regresije, modeli logistične regresije, različne vrste regresijskih modelov pri obravnavi časovnih vrst, različni modeli in z njimi povezano reševanje regresijskih problemov v okviru ekonometrije, in podobno. Srčno upamo, da smo s tem učbenikom bralcu podali zadovoljivo osnovo iz statistike in regresijskih modelov ne le v okviru predmeta Optimizacija logističnih procesov, pač pa tudi pri pridobivanju kasnejših znanj tako pri nadaljnjem študiju, kot tudi kasneje v praksi.

LITERATURA

1. Artenjak J., Poslovna statistika, Ekonomsko-poslovna fakulteta, 2003.

2. Armstrong J.S.: Principles of Forecasting: A Handbook for Researchers and Practitioners, Springer, 2001.

3. Babbie F.: The Practice of Social Research, 10th Edition, Thomson Learning, 2004.

4. Bajt A., Štiblar F.: Statistika za družboslovce, GV Založba, Ljubljana, 2002.

5. Bastič M.: Metode raziskovanja, Ekonomsko-Poslovna fakulteta, Maribor, 2006.

6. Benjamin J.R., Cornell, C.A.: Probability, Statistics, and Decision for Civil Engineers, McGraw-Hill, 1970.

7. Bernstein, S., Bernstein, R.: Schaum's outline of theory and problems of elements of statistics II: Inferential statistics, New York: McGraw-Hill, 1999.

8. Bertsekas D.P., Tsitsiklis, J.N.: Introduction to Probability, Athena Scientific, Belmont USA, 2002.

9. Blejec M., Lovrenčič M., Perman M., Štraus M.: Statistika, Visoka šola za podjetništvo, Piran, 2003.

10. Bronson R., Naadimuthu G.: Schaum's Outline of Operations Research, McGraw-Hill, 2nd ed., 1997.

11. Brvar, B.: Statistika, Fakulteta za varnostne vede, Ljubljana, 2007.

12. Čuljak V.: Vjerovatnost i Statistika, Građevinski fakultet, Sveučilište u Zagrebu, 2011.

13. De Sa J.P.M.: Applied Statistics using SPSS, Statistica, Matlab and R, 2nd ed., Springer, 2007.

14. Dragan D.: Upravljanje logističnih sistemov, Učbenik, Fakulteta za logistiko, Univerza v Mariboru, 2009 ([Dragan 1]).

15. Dragan D.: Stohastični procesi v logistiki, Učbenik, Fakulteta za logistiko, Univerza v Mariboru, 2013 ([Dragan 2]).

16. Elezović N.: Vjerovatnost i Statistika: Diskretna Vjerovatnost, Zagreb, 2007.

17. Ferligoj A., Manfreda K.L, Žiberna A.: Osnove statistike na prosojnicah, Fakulteta za družbene vede, Ljubljana, 2011.

18. Grinstead C.M., Snell J.L.: Introduction to Probability, American Mathematical Society, 2 Revised edition, 1997.

19. Gujarati D., Porter D.: Basic Econometrics, 5th ed., McGraw-Hill, 2008.
Hines W.W., Montgomery D.C.: Probability and Statistics in Engineering and Management Science, John Wiley&Sons, 1990.

20. Hsu H.: Schaum's Outline of Probability, Random Variables, and Random Processes, McGraw-Hill, 1997

21. Jamnik R.: Matematika, Društvo matematikov, fizikov in astronomov, Ljubljana, 1985 ([Jamnik 1]).

22. Jamnik, R.: Uvod v matematično statistiko, DMFA Slovenije, 1976 ([Jamnik 2]).

23. Jamnik R.: Verjetnostni račun in statistika, DMFA Slovenije, 1986 ([Jamnik 3]).

24. Jesenko, J.: Statistika v organizaciji in managementu, Fakulteta za organizacijske vede, Univerza v Mariboru, Založba Moderna organizacija, 2001.

25. Jurišić A., Verjetnostni račun in statistika, Fakulteta za računalništvo in informatiko, 2012.

26. Kmenta J.: Počela Ekonometrije, Mate, 1997.

27. Košmelj K., Uporabna statistika, druga dopolnjena izdaja, Biotehniška fakulteta, Ljubljana, 2007.

28. Košmelj B., Rovan J.: Statistično sklepanje, Ljubljana, 2007.

29. Kottegoda N.T., Rosso R.: Statistics, Probability and Reliability for Civil and Environmental Engineering, McGraw-Hill, 1997.

30. Krishnamoorthy K.: Handbook of Statistical Distributions with Applications, Chapman and Hall, 2006.

31. Kutner M. H., Neter J., Nachtsheim C. J., Li W.: Applied Linear Regression Models, McGraw-Hill College, 2004.

32. Ljubič T.: Predvidevanje in napovedovanje v oskrbovalni verigi, Založba Moderna Organizacija, 2008.

33. Pfajfar L.: Osnovna statistika za ekonomske in poslovne vede, Ekonomska fakulteta, Ljubljana, 2011.

34. Martinez W.L., Martinez A.R.: Computational Statistics Handbook with Matlab, 2nd ed., Chapman and Hall, 2007.

35. Matko D.: Identifikacije, Fakulteta za elektrotehniko, Ljubljana, 1992.

36. Monks J. G.: Theory And Problems Of Operations Managment, Schaum's Outline Series In Business, McGraw-Hill, 1985.

37. Montgomery D.C., Runger G.C.: Applied Statistics and Probability for Engineers, Wiley, 4th Ed., 2006 ([Montgomery 1]).

38. Montgomery D.C., Runger G.C., Hubele N.F.: Engineering Statistics, Wiley, 2011 ([Montgomery 2]).

39. Moore D. S.: The Basic Practice of Statistics, 2nd ed., W. H. Freeman, 1999.

40. Nemec J.: Statistika, Fakulteta za kmetijstvo in biosistemske vede, Maribor, 2009.

41. Ross S.M.: Introduction to Probability and Statistics for Engineers and Scientists, 4th ed., Academic Press, 2009 ([Ross 1]).

42. Ross S.M.: Introductory Statistics, 3th ed., Academic Press, 2010 ([Ross 2]).

43. Soong T.T.: Probability and Statistics for Engineers, Wiley, 2004.

44. Spiegel M.R.: Theory and Problems of Probability and Statistics, McGraw-Hill, 1997.

45. Šrekl J.: Statistika, Fakulteta za kemijo in kemijsko tehnologijo, Ljubljana, 2009.

46. Taha H.A.: Operations Research: An Introduction, 6th ed., Prentice Hall, 1996.

47. Tominc P.: Statistične metode: Uporaba v prometu, Fakulteta za gradbeništvo, 2000.

48. Tominc P., Kramberger T.: Statistične metode v logistiki, Fakulteta za logistiko, 2007.

49. Trauth M.: Matlab Recipes for Earth Sciences, 3rd ed., Springer, 2010.

50. Turk G., Verjetnostni račun in statistika, Skripta, Ljubljana, 2001.

51. Usenik J.: Matematične metode v prometu, Univerza v Ljubljani, Fakulteta za pomorstvo in promet, Portorož, 1998 ([Usenik 1]).

52. Usenik J.: Matematične metode z zbirko nalog: Verjetnostni račun, Univerza v Ljubljani, Fakulteta za pomorstvo in promet, Portorož, 1995 ([Usenik 2]).

53. Vadnal A.: Elementarni uvod v verjetnostni račun, Državna založba Slovenije, Ljubljana, 1988.

54. Vidakovic B.: Statistics for Bioengineering Sciences: With Matlab and WinBugs Support, Springer, 2011.

55. Walpole R.E., Myers R.H., Myers S.L., Ye K.: Probability & Statistics for Engineers & Scientists, Pearson Higher Education, 2007.

56. Winston, W. L.: Operations Research, Applications and Algorithms, Duxbury Press, International Thomson Publishing, 1994.

57. Žibert J., Verjetnost in statistika v tehniki in naravoslovju, Skripta, Fakulteta za matematiko, naravoslovje in informacijske tehnologije, Univerza na Primorskem, 2012.