**PRIRODNO–MATEMATIČKI FAKULTET**
**UNIVERZITETA U NIŠU**

**Prof. dr BILJANA POPOVIĆ**

# MATEMATIČKA STATISTIKA I STATISTIČKO MODELOVANJE

NIŠ, 2003

**Autor:**
Dr Biljana Popović, redovni profesor Prirodno–matematičkog fakulteta u Nišu

**MATEMATIČKA STATISTIKA I STATISTIČKO MODELOVANJE**

**Recenzenti:**
Dr Zagorka Lozanov–Crvenković, redovni profesor Prirodno–matematičkog fakulteta u Novom Sadu
Dr Mila Stojaković, redovni profesor Fakulteta tehničkih nauka u Novom Sadu

**Tehničko uredjivanje:**
Autor

Odlukom Naučno–nastavnog veća Prirodno–matematičkog fakulteta u Nišu, broj 373/1–01 od 10. jula 2002. godine, rukopis je odobren za štampu kao univerzitetski udžbenik.

**Izdavač:**
Prirodno–matematički fakultet, Niš

**YU ISBN** 86–83481–13–1

**Štampa:** "Sven", Niš

**Tiraž:** 100 primeraka

# Sadržaj

# Predgovor

Ova knjiga je rezultat nastavnog rada autora u periodu dužem od jedne decenije. Autor je nastavnik, najpre na predmetu Uvod u matematičku statistiku, a od 1994/95. školske godine na predmetu Matematička statistika i statističko modelovanje na Odseku za matematiku, nekada Filozofskog, a sada Prirodno-matematičkog fakulteta u Nišu. Dakle, knjiga nastaje prvenstveno kao udžbenik za studente matematike i to smera Diplomirani matematičar za računarstvo i informatiku Prirodno-matematičkog fakulteta, za istoimeni predmet. U tom pogledu ona u potpunosti pokriva teorijske sadržaje predmeta Matematička statistika i statističko modelovanje predvidjene programom. U toku školske 2001/2002. godine, sadražaj ove knjige je svojim najvećim delom bio dostupan studentima na sajtu Fakulteta u vidu autorizovanih predavanja čime je prošao i proveru čitljivosti i prihvatljivosti za ciljnu grupu čitalaca kojima je namenjen.

Za korišćenje udžbenika u celosti neophodno je opšte matematičko znanje do nivoa teorije mera i integrala, medjutim, činjenica je da se njegov veći deo može čitati i koristiti već sa stečenim znanjem klasičnog kursa Matematičke analize II. Posebno matematičko znanje kojim se podrazumeva da raspolažu studenti, korisnici udžbenika, je standardni kurs savtemene Teorije verovatnoća.

Udžbenik će, takodje, u dobroj meri moći da zadovolji kurseve Matematička statistika i Statističko modelovanje na smeru Diplomirani matematičar za matematiku ekonomije Odseka za matematiku Prirodno-matematičkog fakulteta u Nišu, i to u delu opštih znanja iz ovih oblasti.

Osim namene da se koristi kao udžbenik na osnovnim studijama matematike, udžbenik bi bio i polazište specijalističkih, magistarskih i doktroskih studija iz oblasti Matematičke statistike i primena za studente koji na svojim osnovnim studijama nisu bili u prilici da ovladaju znanjima iz ove oblasti u dovoljnoj meri za potrebe postdiplomske nastave.

Autor se zahvaljuje recenzentima, a posebno prof. dr Zagorki Lozanov–Crvenković na veoma korisnim sugestijama.

Niš, oktobra 2003. godine Autor

# Glava 1

# Uvod

Matematička statistika je primenjena matematička disciplina srodna teoriji verovatnoće. Bazira se na pitanjima i metodama teorije verovatnoće, ali rešava svoje specifične (probleme) zadatke svojim metodama. (Svaka matematička teorija se razvija u okviru nekog modela koji opisuje odredjeni krug realnih pojava čijim se proučavanjem i bavi data teorija.)

U teoriji verovatnoće se polazi od pretpostavke da je poznat prostor verovatnoća $(\Omega, \mathcal{F}, P)$, gde je $\Omega$ skup svih elementarnih ishoda, $\mathcal{F}$ je $\sigma$-algebra na skupu $\Omega$ a $P$ je verovatnoća.

Verovatnoća $P$, u praktičnim problemima koje treba rešavati, nije u potpunosti poznata. U većini slučajeva se pretpostavlja da $P \in \mathcal{P}$, gde je $\mathcal{P} = \{P\}$ familija verovatnoća. Takvi praktični problemi nazivaju se statističkim modelima.

Dakle, za razliku od modela teorije verovatnoća, statistički model je $(\Omega, \mathcal{F}, \mathcal{P})$.

**Primer 1.** (Šema Bernulija.) Obavlja se $n$ nezavisnih opita u kojima se realizuje 0 ili 1 sa verovatnoćama redom $1 - p = q$ i $p$, $0 \leq p \leq 1$. Ishod ovog eksperimenta je

$$\Omega = \{\omega : \omega = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n), \varepsilon_i = 0, 1\}.$$

Pri tome je verovatnoća pojedinog elementarnog ishoda

$$P(\omega) = p^{\sum \varepsilon_i} q^{n - \sum \varepsilon_i}.$$

Ako verovatnoća $p$ nije prethodno poznata, označićemo je sa $\theta$ i tu oznaku ćemo nadalje koristiti za svaki nepoznati parametar. U tom slučaju jedina informacija koju imamo o parametru ovog primera je da je $\theta \in \Theta = [0, 1]$. Tačnije, imamo jedino informaciju da raspodela verovatnoća kojom ovaj eksperiment opisujemo pripada familiji $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$, gde je $P_\theta = \theta^{\sum \varepsilon_i}(1 - \theta)^{n - \sum \varepsilon_i}$. $\triangle$

U prethodnom primeru je definisan jedan statistički model, dakle model koji u sebi sadrži neku vrstu neodredjenosti. Zadatak matematičke statistike je da se korišćenjem informacije dobijene posmatranjem ishoda eksperimenta, dakle statističkih podataka, smanji ta neodredjenost, odnosno da se, što je moguće tačnije, izvrši izbor $P \in \mathcal{P}$.

Matematička statistika je nauka o statističkom zaključivanju. Statističko zaključivanje podrazumeva rešavanje zadataka obrnutih od onih koje rešava teorija verovatnoće: ona

utvrdjuje strukturu statističkih modela prema rezultatima sprovedenih posmatranja, dakle, odredjuje prostor verovatnoća na osnovu eksperimenta. Pri tome posmatranja ne mogu biti proizvoljna. Naime, ona moraju biti ekvivalentna statističkom eksperimentu:

- može se ponavljati proizvoljan broj puta pod istim uslovima,
- unapred je definisano šta se registruje u eksperimentu pri čemu su poznati svi mogući ishodi i
- ishod pojedinačnog eksperimenta nije unapred poznat.

Za prve svesne pokušaje definisanja i primene statističkog zaključivanja uzimaju se popisi stanovništva koje su sprovodili vladari još nekoliko vekova pre naše ere radi utvrdjivanja broja vojnih podanika ili poreskih obveznika. Zasnivanje statistike kao nauke vezuje se za pojavu škole "političkih aritmetičara" u Engleskoj u $XVII$ veku. Po nekima, delo "Natural and Political Observations upon the Bills of Mortality", koje je napisao Dž. Grant (J. Graunt) i objavio 1622. godine, označava početak statistike kao nauke. Dugo vremena je statistika smatrana naučnim metodom za proučavanje društvenih nauka. Medjutim, matematičari koji su neminovno bili uključeni u konstituisanje, formalno definisanje, i postali odgovorni za razvoj statističkog metoda zaključivanja, odgovorni su i za početak primene statistike u prirodnim naukama. Tu ideju medju prvima je prihvatio engleski biolog Galton (Sir Francis Galton, 1822-1911), koji je primenio statistički metod u istraživanjima u biologiji. Teorijski doprinos razvoju matematičke statistike dao je medju prvima švajcarski matematičar Jakob Bernuli (Jacob Bernoulli, 1654-1705) definišući i obrazlažući zakon velikih brojeva u svom delu "Ars conjectandi". Krupan korak u tom pravcu dao je i francuski astronom i matematičar Laplas (Pierre Simon, Marquis de Laplace, 1749-1827). Poznato je njegovo delo "Théorie analytique de probabilités". Buran razvoj matematičke statistike kao teorijske discipline u $XX$ veku omogućen je, pre svega, razvojem teorije verovatnoća u ovom periodu.

# Glava 2

# Teorija uzoraka

## 2.1 Osnovni pojmovi

Statistički eksperiment se izvodi nad elementima nekog skupa na kojima se posmatra jedno ili više zajedničkih svojstava.

Definicija 1. *Populacija* ili *generalni skup* ili *osnovni skup* je skup elemenata čija se zajednička svojstva izučavaju statističkim metodima. Populacija se simbolički beleži sa $\Omega$, a njen element sa $\omega$.

Definicija 2. *Obeležje* je zajedničko svojstvo elemenata jedne populacije (koje se ispituje). Obeležje može biti kvantitativno (numeričko) ili kvalitativno (atributivno).

Pri izvodjenju statističkog eksperimenta polazi se od pretpostavke da se tom prilikom realizuju neki slučajni dogadjaji. Dakle, pretpostavlja se da se ishod eksperimenta može prikazati slučajnom veličinom $X$. Ukoliko je eksperiment ponavljan $n$ puta, ishod se predstavlja slučajnim vektorom $\mathbf{X} = (X_1, X_2, \ldots, X_n)$. Pri proučavanju ovog slučajnog vektora poželjno je poznavati njegovu raspodelu. S tim u vezi reći ćemo da treba odrediti gustinu raspodele obeležja, a nadalje ćemo to pojasniti. Ovde će se koristiti termin gustina raspodele u uopštenom značenju, tj. vezivaće se i za slučajne promenljive diskretnog tipa.

**Primer 2.** Za slučajnu promenljivu sa binomnom raspodelom $\mathcal{B}(1, p)$, kazaćemo da ima gustinu raspodele

$$f(x) = \begin{cases} p^x(1-p)^{1-x}, & x = 0, 1 \\ 0, & x \neq 0, 1 \end{cases} . \quad \triangle$$

Neka je $Y$ slučajna promenljiva definisana kao funkcija slučajnih promenljivih

$$X_1, X_2, \ldots, X_n,$$

tj. neka je $Y = u(X_1, X_2, \ldots, X_n)$. Odredjivanje gustine raspodele ove slučajne promenljive na osnovu poznavanja zajedničke gustine raspodele vektora slučajnih promenljivih $\mathbf{X} = (X_1, X_2, \ldots, X_n)$, u oznaci $f(x_1, x_2, \ldots, x_n)$, $(x_1, x_2, \ldots, x_n) \in R^n$, je jedan od zadataka matematičke statistike. Sam slučajni vektor $\mathbf{X}$ i funkcije od njegovih komponenata su okosnica matematičke statistike.

DEFINICIJA 3. *Uzorak* je deo populacije na kome se ispituje posmatrano obeležje. Broj elemenata u uzorku se naziva *obim uzorka*.

Na uzorku se sprovodi statistički eksperiment. Ishod tog eksperimenta će biti vektor $\mathbf{X}$, koji je po svojim karakteristikama slučajna promenljiva.Vektor $\mathbf{X}$ još zovemo *slučajnim uzorkom* za razliku od njegove *realizovane vrednosti* po obavljenom eksperimentu.

DEFINICIJA 4. Vektor $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ koji predstavlja realizaciju vektora $\mathbf{X}$ po obavljenom eksperimentu zovemo *realizovani uzorak*.

**U daljem tekstu će se pod uzorkom podrazumevati slučajni uzorak.**

Detaljnije o uzorku i načinima za izbor uzoraka biće reči nadalje. Populacija ima nešto širi smisao od izvesnog dogadjaja u teoriji verovatnoće, dok je obeležje nešto širi pojam od pojma slučajne promenljive. Naime, izvesan dogadjaj je skup svih mogućih elementarnih ishoda jednog eksperimenta, pri čemu se podrazumevaju različiti ishodi. Populacija je, medjutim, skup svih elemenata na kojima se posmatra neko svojstvo (skup ljudi, skup sijalica, deo tla, itd.). Definišimo funkciju iz skupa $\Omega$, populacije, u skup koji čine kategorije jednog svojstva. Preciznije, na skupu $\Omega$ se definiše relacija ekvivalencije: "dva elementa populacije su u relaciji ako su im jednake vrednosti obeležja koje se na elementima populacije posmatra". Tom relacijom se vrši razbijanje skupa $\Omega$ na klase ekvivalencije, odnosno, definiše se faktor skup. Klase ekvivalencije su kategorije, te se najpre definiše preslikavanje populacije na faktor skup jednom funkcijom tako što se svakom elementu populacije pridružuje njegova klasa ekvivalencije. Iz poslednjeg skupa je moguće definisati novu funkciju sa vrednostima u skupu realnih brojeva, $R$, koja je, zapravo, slučajna promenljiva. Kompozicija ovih funkcija je *obeležje*. U tom smislu se može govoriti o raspodeli obeležja posredstvom raspodele ovako definisane slučajne promenljive, te će se i obeležje, kao i slučajna promenljiva, označavati velikim slovom latinice sa kraja abecede, $X$,$Y$,$Z$,.... U vezi sa uopštenjem pojma gustine raspodele smatraće se da svako obeležje ima svoju gustinu raspodele.

**Primer 3.** Za populaciju ćemo uzeti studente Prirodno-matematičkog fakulteta u Nišu. Neka je obeležje koje posmatramo na toj populaciji "obrazovni profil". U ovom momentu ćemo posmatrati samo osnovni profil, tj. matematika, fizika, hemija, biologija, geografija. Ovih 5 kategorija bi činile razbijanje skupa $\Omega$. Dakle, studenti istog odseka – obrazovnog profila bi činili jednu klasu ekvivalencije. Nadalje bismo svakom odseku pridružili broj (kod), recimo neka su to prirodni brojevi od 1 do 5. Time bi bila definisana slučajna promenljiva. $\triangle$

Sa gledišta matematičke statistike dato obeležje $X$ je potpuno odredjeno ako je odredjena njegova raspodela, $P\{X \in S\}$, gde je $S \in \mathcal{B}_1$, a $(R, \mathcal{B}_1, P)$ fazni prostor. To je istovremeno i jedan od glavnih problema kojima se bavi matematička statistika: odredjivanje raspodele obeležja. Pri tome je moguće da unapred nije poznata familija dopustivih raspodela ili da je ona poznata, a da iz nje treba napraviti pravi izbor ocenom vrednosti nepoznatih parametara koji u raspodeli figurišu. Dakle, osnovni problem statističkog zaključivanja je da na osnovu statističkog eksperimenta nešto zaključi o raspodeli obeležja.

Zaključivanje o raspodeli obeležja vrši se na osnovu izabranog uzorka. Otuda je važno da izabrani uzorak bude reprezentativan, tj. da bude takav da se sa dovoljnom tačnošću

zaključak o raspodeli posmatranog obeležja dobijenoj na uzorku može da ekstrapoluje na čitavu populaciju.

## 2.2 Pojam slučajnog broja

Za izbor reprezentativnog uzorka preporučuje se slučajni izbor, tj. izbor elemenata populacije u uzorak na slučajan način. Da bi se realizovao slučajni izbor često se koristi tablica slučajnih brojeva.

Razmotrimo dekadni brojni sistem. Za zapisivanje nekog realnog broja u dekadnom brojnom sistemu koristi se dest cifara: 0,1, 2, 3, 4, 5, 6, 7, 8, 9. Ako pretpostavimo da vršimo eksperiment u kome je jednako verovatan izbor bilo koje od navedenih deset cifara, svaka cifra će biti izabrana sa verovatnoćom 0,1. Slučajna promenljiva kojom se opisuje ovaj eksperiment ima diskretnu uniformnu raspodelu. Ponavljanjem eksperimenta proizvoljan broj puta (pod istim uslovima, pri čemu su poznati svi mogući ishodi eksperimenta, ali ni u jednom pojedinom eksperimentu nije unapred poznat ishod – statistički eksperiment) dobio bi se niz slučajnih brojeva ili, preciznije, slučajnih cifara. Potreba za ovakvim nizom i formiranjem čitave tablice slučajnih brojeva (Tablica 6) biće jasnija u narednim poglavljima. Tablica 6 je samo deo tablice od 1000000 slučajnih cifara sačinjene 1955. godine u SAD od strane korporacije pod nazivom "Rand Corporation". Tehnika kojom je dobijena tablica koristi ideju ruleta. Naime, pomenuta tablica dobijena je pomoću ruleta sa deset polja od kojih je svako polje odgovaralo po jednoj dekadnoj cifri (pri čemu je elektronika i mehanika sistema morala da zadovolji posebno visoke zahteve). Otuda se statističke tehnike koje koriste slučajne brojeve zovu metod Monte Karlo, prema gradu poznatom po kockarnicama.

Prema tablici slučajnih brojeva dekadnog brojnog sistema moguće je napraviti i tablice slučajnih brojeva drugih brojnih sistema, na pr. binarnog brojnog sistema identifikujući, recimo, sve parne cifre sa 0, a neparne sa 1.

Postoje i neki drugi fizički sistemi koji su se koristili kao generatori slučajnih brojeva. Jedan primer je emisija čestica radioaktivnog izvora zračenja, pri čemu se beleži broj čestica u jedinici vremena registrovanih na barijeri.

Kako se koristi tablica slučajnih brojeva?

**Primer 4.** Ako bi nam iz bilo kog razloga bilo potrebno da imamo 15 dvocifrenih brojeva ne većih od 63, koji su uniformno raspodeljeni, tj. ako je u pitanju slučajni eksperiment sa raspodelom $P\{X = n\} = \frac{1}{90}$ gde je $n$ dvocifren broj, trebalo bi iz tablice po nekoj strategiji (ili redom) čitati grupe od po dve cifre izostavljajući one grupe koje počinju nulom sve dok ne izaberemo 15 dvocifrenih brojeva ne većih od 63. Na putu do tog cilja ignorisali bismo sve grupe cifara koje bi protumačili kao dvocifren broj veći od 63 na koje bismo u tablici naišli. Na primer, čitajmo grupe od po dve cifre iz prvog i drugog reda Tablice 6. Dobijamo redom

$$51, 77, 27, 46, 40, 42, 33, 12, 90, 44, 46, 62, 12, 40, 33, 23, 49$$

i odbacujemo brojeve 77 i 90. Ako bi za eksperiment bilo prihvatljivo da se brojevi ponavljaju, u ovom trenutku bismo završili čitanje. Medjutim, ako se brojevi ne smeju

ponavljati, ignorisali bismo $46, 12, 40$ i 33 kada se drugi put jave u pročitanom nizu i pročitali bismo još naredne brojeve

$$49, 18, 35, 87, 06, 56, 82, 19$$

i odbacili 87, 06 i 82. $\triangle$

O primeni tablice slučajnih brojeva biće nadalje još reči.

Naglasimo da je sa pojavom računara pomenuta tablica izgubila na značaju, ali ne i metod. Naime, tablica nije pogodna za korišćenje pri obradi podataka na računaru, jer pre svega usporava rad paralelnim radom, a drugo nije pogodno ni da se tablica unese u memoriju računara jer bi zauzela, odnosno, blokirala veliki deo memorije za aktivno korišćenje. S toga se prilikom rada na računaru koriste tzv. pseudoslučajni brojevi. Pseudoslučajni brojevi "dosta dobro" sa statističke tačke gledišta aproksimiraju tablicu slučajnih brojeva, tj. pomenutu uniformnu raspodelu, a generišu se pomoću algoritama programiranih na računaru. Jedan od najčešće korišćenih algoritama je linearni kongruentni metod kod koga se niz brojeva $x_0, x_1, x_2, \ldots$ dobija preko formule

$$x_{n+1} = (ax_n + c) mod M \quad .$$

Čitava teorija je usmerena na to da se konstante $x_0, a, c$ i $M$ odaberu tako da se dobije što duži niz brojeva. Prema definiciji je jasno da je dužina niza različitih brojeva najviše $M$. Najčešće je $M = 2^k, \quad k \geq 1$ ($k$ se po pravilu uzima kao vrlo veliki broj).

Kvalitet koji treba da zadovolji algoritam da bi dobijeni niz "dovoljno dobro" aproksimirao niz slučajnih brojeva je predmet iz domena testiranja statističkih hipoteza.

Treba naglasiti da su neki iracionalni brojevi, odnosno njihove značajne cifre, kao što su broj $\pi$ i $\sqrt{7}$, izvanredni prirodni generatori niza slučajnih cifara.

## 2.3 Slučajni izbori bez i sa vraćanjem

Bez obzira da li je populacija konačna (obima $N$, $N < \infty$) ili beskonačna, moguće je iz nje na različite načine izabrati uzorke istog obima $n$ (za konačnu populaciju $n < N$) za različite prirodne brojeve $n$. Dakle, može se govoriti o kolekciji $\mathcal{S}$ svih uzoraka iste populacije $\Omega$, $\mathcal{S} = \{s\}$, gde je sa $s$ označen proizvoljan uzorak posmatrane populacije, $s \subset \Omega$, dok će obim uzorka $s$ biti obeležen sa $n(s)$. Ako u uzorku $s$ ima istih elemenata, sa $\nu(s)$ možemo označiti broj različitih elemenata u uzorku $s$.

Definicija 5. Broj različitih elemenata u uzorku je *efektivni obim uzorka.*

U vezi sa efektivnim obimom uzorka za konačnu populaciju uvodi se pojam stope izbora:

Definicija 6. *Stopa izbora uzorka* ili *frakcija uzorka* je funkcija od uzorka $s$ definisana kao količnik efektivnog obima uzorka i obima populacije,

$$f(s) = \frac{\nu(s)}{N} \quad .$$

Ako uzorak $s$ redukujemo samo na različite elemente, dobićemo uzorak $\tilde{s}$ čiji će obim biti $\nu(s)$, tj. $\nu(\tilde{s}) = \nu(s) = n(\tilde{s})$. Ako medju svim elementima skupa $\mathcal{S}$ izvršimo ovakvu redukciju, dobićemo skup $\tilde{\mathcal{S}} = \{\tilde{s}\}$.

Zaključivanje na osnovu uzorka po pravilu zavisi od načina izbora elemenata populacije u uzorak. Način izbora uzorka zove se *plan uzorka* ili *strategija izbora*. Formalna definicija plana je sledeća:

Definicija 7. *Plan uzorka* je zakon raspodele neke slučajne promenljive $S$ definisane na skupu $\mathcal{S}$, tj. $\{P(S = s), s \in \mathcal{S}\}$.

Nadalje će biti korišćena oznaka $P(S = s) = p(s)$.

U tom smislu se uopštava pojam slučajnog uzorka o kome je već bilo reči, podrazumevajući da je uzorak slučajan i kada je dobijen na osnovu poznatog plana, tj. na osnovu zadate raspodele verovatnoća.

Za $\omega \in \Omega$ definisaćemo indikator, u smislu da li uočeni element populacije pripada izabranom uzorku $s$, na sledeći način:

$$I_s(\omega) = \begin{cases} 1, & \omega \in s \\ 0, & \omega \notin s \end{cases} .$$

Pri tome je

$$P(I_s = 1) = \sum_{s \ni \omega} p(s) \quad .$$

Kraće ćemo označiti

$$P(I_s = 1) = \pi \quad .$$

Za prebrojivu, a naročito za konačnu populaciju, po potrebi se definiše bijektivna funkcija na skup prirodnih brojeva, tj. na prvih $N$ prirodnih brojeva, za konačnu populaciju, čime se svaki element populacije identifikuje sa svojim "mestom" u populaciji. Tada je moguće definisati *indikator uključenja $i$–tog elementa populacije u uzorak* kao slučajnu promenljivu

$$I_s(\omega_i) = \begin{cases} 1, & \omega_i \in s \\ 0, & \omega_i \notin s \end{cases}$$

sa raspodelom

$$P(I_s(\omega_i) = 1) = \sum_{s \ni \omega_i} p(s) \quad .$$

Koristićemo oznaku

$$P(I_s(\omega_i) = 1) = \pi_i \quad .$$

Pomenuta bijekcija se primenjuje kada je bitan redosled izbora elemenata u uzorak.

Neka je iz populacije $\Omega$ na kojoj posmatramo obeležje $X$ uzet slučajni uzorak obima $n$, $\mathbf{X} = (X_1, X_2, \ldots, X_n)$. Kao što je već rečeno, $X_i$ je vrednost obeležja na $i$–tom elementu uzorka, odnosno, $(X_1, X_2, \ldots, X_n)$ je niz slučajnih promenljivih. Sa gledišta teorije verovatnoće, najjednostavniji je *prost* slučajni uzorak kod koga se pretpostavlja da su slučajne promenljive $X_1, X_2, \ldots, X_n$ nezavisne i da svaka ima istu raspodelu kao obeležje $X$. U terminima planova uzoraka to znači da su sve verovatnoće $p(s)$ pri $n(s) = n$ medju sobom jednake.

Najopštija podela planova slučajnih uzoraka je na uzorke sa vraćanjem (ponavljanjem) i uzorke bez vraćanja (ponavljanja). Uzorak sa vraćanjem pretpostavlja strategiju izbora kod koje se jedan isti element populacije može više puta javiti u uzorku, odnosno biti izabran. To bi se moglo dogoditi ukoliko se posle izbora elemnta u uzorak i registrovanja vrednosti obeležja na njemu, on ponovo vraća u populaciju. Otuda i naziv ove strategije. Kod uzorka bez vraćanja takva mogućnost ne postoji, odnosno po izboru elementa u uzorak jednom, on se više ne vraća u populaciju. Kod beskonačne populacije se ove dve strategije u praksi ne razlikuju, jer je mala verovatnoća ponovnog izbora istog elementa populacije u uzorak. Ukoliko je populacija konačna, ali daleko većeg obima nego što je obim uzorka, verovatnoće svih uzoraka, $p(s)$, konstantnog obima su približno jednake, pa se uzorak izabran po bilo kojoj od navedenih strategija može smatrati prostim sa verovatnoćom 1. (Da bi ova tvrdnja opstala, moralo bi se pristupiti opširnom dokazivanju, što ovde neće biti sprovedeno.) Situacija u kojoj se izbor sa vraćanjem i izbor bez vraćanja bitno razlikuju je izbor iz konačne populacije iz koje se uzima uzorak čiji obim nije zanemarljivo mali u odnosu na obim populacije. Nadalje će biti više reči o tome.

* * *

Za izbor uzorka iz uredjene populacije može se koristiti tablica slučajnih brojeva.

Ukoliko nam je potreban uzorak obima 20 iz populacije obima 1000, koja je uredjena, čitali bismo grupe od po tri cifre zajedno iz tablice slučajnih brojeva. Dobijene brojeve bismo tumačili kao redne brojeve elemenata populacije. Ukoliko bi medju pročitanim grupama bila grupa 000, to bismo protumačili kao da je reč o poslednjem elementu populacije. Broj grupa koje bismo pročitali bi zavisio od strategije izbora, a ne samo od obima uzorka. Za uzorak sa vraćanjem pročitali bismo tačno 20 grupa. Za uzorak bez vraćanja bismo morali da izostavimo svaku ponovljenu grupu i da nastavimo čitanje dok ne pročitamo 20 različitih grupa cifara, tj. rednih brojeva.

**Primer 5.** Neka je data populacija od 100 elemenata. Koristeći tablicu slučajnih brojeva modelirati realizovani uzorak bez vraćanja od 20 elemenata iz ove populacije.

Ova populacija je očigledno uredjena. Iz tablice slučajnih brojeva čitaćemo redne brojeve elemenata populacije koje ćemo uzeti u uzorak. Ako se odlučimo za petnaesti red Tablice 6 i čitemo po dve cifre dobijamo:

$$85, 65, 93, 60, 81, 50, 88, 41, 40, 70, 74, 95,$$

sad možemo da nastavimo sa čitanjem u šesnaestom redu, pri čemu možemo da "pročitamo" i 0 na kraju dela tablice slučajnih brojeva koja se nalazi na kraju petnaestog reda Tablice 6 ili da je izostavimo. Recimo da je "pročitamo", dobijamo

$$05, 51, 89, 00, 56, 52, 53, 11.$$

Broj 74 se javlja dva puta i njega izostavljamo na mestu kada se drugi put javi (izmedju grupa "00" i "56"), jer smo već prethodno uzeli 74–ti element populacije u uzorak, a uzorak je bez vraćanja. Jasno da 05 nalaže da uzmemo 5–ti element populacije u uzorak, a 00 da uzmemo 100–ti. $\triangle$

**Primer 6.** Planira se sondiranje terena u 8 tačaka radi ispitivanja sastava tla. Preciznost merenja je do $0,10m$, a površina ispitivanog terena je $68a$(ari).

U mapu te lokacije treba uneti Dekartov koordinatni sistem tako da ucrtane ose budu tangente ispitivane parcele i utvrditi dimenzije minimalnog pravougaonika koji u potpunosti pokriva ispitivani teren sa dvema stranicama na osama, a zatim mesta za sondiranje odrediti uz pomoć tablice slučajnih brojeva.

Neka je pravougaonik dimenzije $100 \times 80m$. Čitaćemo iz Tablice 6 iz prvog i drugog reda uporedo grupe od po prve 4 cifre iz svake grupe kolona radi formiranja uredjenih parova koordinata i množiti dobijene brojeve sa $10^{-2}$:

$$x: \ 51,77 \quad 74,64 \quad 42,33 \quad 29,04 \quad 46,62 \quad 45,93 \quad 60,17 \quad 52,07 \quad 25,42\ldots$$

$$y: \ 24,03 \quad 23,49 \quad 83,58 \quad 06,56 \quad 21,96 \quad 30,58 \quad 02,13 \quad 75,79 \quad 45,40\ldots$$

Uredjeni par $(42,33; 83,58)$ se odbacuje jer izlazi iz područja definisanog pravougaonika. Takodje će biti tačaka koje treba odbaciti jer ne pripadaju definisanom području koje se ispituje. $\triangle$

### 2.3.1 Uzorak bez vraćanja iz konačne populacije

Plan uzorka obima $n$ bez vraćanja definisan je na kolekciji $\tilde{\mathcal{S}}$ uzoraka bez ponavljanja elemenata:

$$p(s) = \begin{cases} \frac{1}{\binom{N}{n}}, & n(s) = \nu(s) = n \\ 0 \quad , & \text{u ostalim slučajevima} \end{cases} \quad , \quad s \in \tilde{\mathcal{S}}$$

ukoliko je populacija konačna, $N \geq n$. To otuda što je svaki uzorak $\tilde{s}$ kombinacija bez ponavljanja $n$–te klase od $N$ elemenata, a u slučaju da je izbor "fer", svaki takav uzorak biće izabran sa verovatnoćom $\frac{1}{\binom{N}{n}}$. Ovakvih uzoraka koji sadrže fiksirani element $\omega$ ima tačno $\binom{N-1}{n-1}$ pa je verovatnoća

$$\pi = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N}$$

za neuredjenu populaciju. Ukoliko je pak populacija uredjena, plan uzorka bez vraćanja biće

$$p(s) = \begin{cases} \frac{1}{n!\binom{N}{n}}, & n(s) = \nu(s) = n \\ 0 \quad , & \text{u ostalim slučajevima} \end{cases} \quad , \quad s \in \tilde{\mathcal{S}} \quad ,$$

jer se radi o varijacijama bez ponavljanja, a verovatnoća $\pi_i$ je

$$\pi_i = \sum_{s \ni \omega_i} \frac{1}{n!\binom{N}{n}} = \frac{n}{N} \quad .$$

Dakle, verovatnoća izbora proizvoljnog ali fiksiranog elementa populacije u uzorak po strategiji izbora bez vraćanja, za zadati obim uzorka $(n)$ je konstantna i iznosi $\frac{n}{N}$ bez obzira da li je populacija uredjena ili ne.

Generalno gledano, slučajni izbor bez vraćanja dobija se bilo izvlačenjem svih $n$ elemenata iz populacije odjednom, bilo izvlačenjem jednog po jednog elementa ne vraćajući ga više u populaciju.

Izbor bez vraćanja ima konstantnu stopu izbora $f = n/N$, jer je $\nu(s) = n(s) = n$.

### 2.3.2 Uzorak sa vraćanjem iz konačne populacije

Plan izbora sa vraćanjem zasniva se na činjenici da je u svakom izvlačenju verovatnoća izbora pojedinog elementa populacije u uzorak ista i jednaka $1/N$. Prema tome

$$p(s) = \begin{cases} \frac{1}{N^n}, & n(s) = n \\ 0 \quad , & \text{inače} \end{cases} \quad , \quad s \in \mathcal{S} \quad ,$$

bez obzira da li je populacija uredjena ili ne, a verovatnoća da $i$–ti član populacije bude uključen u uzorak je

$$\pi_i = 1 - \left(\frac{N-1}{N}\right)^n \quad , i = 1, 2, ..., N \quad ,$$

dakle ista kao i kada se radi o verovatnoći $\pi$ za element $\omega$ neuredjene populacije.

Ovakav uzorak nema konstantnu stopu izbora, jer $\nu(s)$ varira za isti obim uzorka $n$.

## 2.4 Neki specijalni planovi uzoraka

### 2.4.1 Stratifikovani uzorak

U mnogim realnim situacijama prirodno je podeliti populaciju na podgrupe koje treba proučavati.

**Primer 7.** Treba izvršiti anketno istraživanje u preduzećima državnog i privatnog sektora. Preduzeća su elementi populacije iz koje treba uzeti uzorak. Medjutim, neka preduzeća su vrlo velika i zapošljavaju više hiljada radnika, dok su druga mala i zapošljavaju svega nekoliko lica. Bilo koja ocena na osnovu direktnog slučajnog uzorka izabranog iz celine skupa preduzeća neće biti realna. Postupak kojim se postiže značajno poboljšanje preciznosti zaključivanja na osnovu slučajnog uzorka jeste stratifikacija. Tako se preduzeća mogu podeliti prema broju radnika na velika, srednja i mala. $\triangle$

Definicija 8. *Stratifikacija (raslojavanje)* podrazumeva podelu populacije na delove - *stratume (slojeve)*, disjunktne podskupove čija unija obuhvata celu populaciju, sa zahtevom postizanja što veće homogenosti unutar stratuma (sloja).

Homogenost se ostvaruje prema nekom zajedničkom svojstvu elemenata populacije, na pr. starosna dob, pol, tip preduzeća i sl.

Kao što je rečeno, stratumi su medju sobom disjunktni, a svi zajedno obuhvataju celu populaciju. Dakle, čine jedno razbijanje populacije potpunim sistemom dogadjaja.

Kako je osnovni cilj matematičke statistike ocenjivanje raspodele obeležja posmatranog na populaciji, cilj stratifikacije je da se postigne veća tačnost ocene, ekonomičnost ili

jednostavnost ispitivanja i slično. U nekim situacijama je ispitivanje i jedino moguće sprovesti po stratumima.

Tehnika stratifikacije podrazumeva rešavanje odredjenih zadataka, odnosno pronalaženje odgovora na sledeća pitanja:

- Kako formirati stratume i koliko njih ?
- Kako izabrati (raspodeliti, alocirati) ukupan uzorak uočavajući pojedine stratume, tj. alocirati uzorak po stratumima ?
- Kako sprovesti statističko zaključivanje na osnovu dobijenog stratifikovanog uzorka ?

Na ova pitanja se može i treba vratiti kasnije, pošto se obrade tačkaste i intervalne ocene parametara, dok ćemo se ovde još malo zadržati na definisanju stratifikovanog uzorka.

Neka je populacija obima $N$ podeljena na $L$ disjunktnih strataума obima $N_l$, gde je $l = 1, 2, \ldots, L$, pri čemu je $N_1 + N_2 + \ldots + N_L = N$. Pretpostavimo da su obimi strataума poznate veličine. Udeo $l$-tog strataума u uzorku može se meriti veličinom $w_l = \frac{N_l}{N}$. Očigledno je $w_1 + w_2 + \ldots + w_L = 1$. Dakle, $w_l$ bi se mogla protumačiti kao klasična definicija verovatnoće da se pri slučajnom izboru elemenata populacije, izabere element $l$-tog strataума.

Stratumi se u istoj populaciji mogu odrediti na različite načine.

**Primer 8.** Radi ispitivanja uspeha na studijama na Prirodno-matematičkom fakultetu treba izvršiti stratifikaciju svih upisanih studenata u prvu godinu studija.

Već pri samom upisu studenti su podeljeni po odsecima koje možemo prihvatiti kao stratume (slojeve). Dakle, studenti jednog odseka bi činili jedan stratum. U tom slučaju bilo bi onoliko strataума koliko ima odseka na Prirodno-matematičkom fakultetu.

Medjutim, s obzirom na cilj istraživanja, intuitivno bi bilo prihvatljivije definisati stratume prema postignutom uspehu u srednjoj školi. Dakle, ako prihvatimo četiri uobičajene kategorije uspeha: odličan, vrlo dobar, dobar i dovoljan, populaciju studenata upisanih u prvu godinu studija posmatranog fakulteta podelili bi na četiri strataума.

Konačno za koji način podele na stratume bismo se opredelili zavisilo bi i od odgovora koji se istraživanjem traži, tj. da li je akcenat, recimo, na profesionalnoj orijentaciji srednjoškolaca (prva podela) ili na validnosti ocenjivanja u srednjim škoalama (druga podela). $\triangle$

Statistički kriterijum za "bolju" stratifikaciju spada u domen testiranja statističkih hipoteza.

Iz stratifikovane populacije se uzima uzorak $s$ čiji je obim $n(s)$, a koga čine podskupovi (poduzorci) $s_1, s_2, \ldots, s_L$ pri čemu je $s_i$, $i = 1, 2, \ldots, L$ deo uzorka $s$ koji je uzet iz $i$-tog strataума. Dakle, $s_i \cap s_j = \emptyset$, $\sum_{i=1}^{L} s_i = s$, pa važi

$$n(s_1) + n(s_2) + \cdots + n(s_L) = n(s).$$

Po pravilu se uvodi pretpostavka da su izvlačenja iz različitih strataума nezavisna. Stopa izbora $i$-tog strataума je $f_i = \frac{n(s_i)}{N_i}$. Ukoliko je $f_i = c$ za svako $i = 1, 2, \ldots, L$, reč

je o proporcionalnoj raspodeli obima uzorka po stratumima. Svakako najjednostavniji metod razmeštaja uzorka po stratumima je izbor jednakog broja elemenata iz svakog sloja, tj. ako važi $n(s_i) = \frac{n}{L}$ za svako $i = 1, 2, \ldots, L$. Ovakav razmeštaj (alokacija) uzorka po slojevima se uglavnom primenjuje kada su slojevi približno istog obima. U protivnom se, po pravilu, koristi alokacija sa konstantnom frakcijom.

Izvlačenja iz stratuma se mogu vršiti takodje sa i bez vraćanja pri čemu dobijamo stratifikovani slučajni uzorak sa vraćanjem ili bez vraćanja.

Ukoliko se na $i$-tom stratumu vrednost posmatranog obeležja na populaciji $X$, označi sa $X^{(i)}$, slučajni uzorak će biti vektor

$$X = (X_1^{(1)}, X_2^{(1)}, \ldots, X_{n(s_1)}^{(1)}, X_1^{(2)}, X_2^{(2)}, \ldots, X_{n(s_2)}^{(2)}, \ldots, X_1^{(L)}, X_2^{(L)}, \ldots, X_{n(s_L)}^{(L)})$$

gde se uočavaju podvektori koji odgovaraju pojedinim stratumima.

### 2.4.2 Grupni uzorak

Grupni uzorak takodje podrazumeva prethodnu podelu cele populacije na disjunktne delove. Kod grupnog uzorka se, medjutim, ne pretpostavlja podela prema zajedničkom svojstvu u ciju postizanja homogenosti grupe. Naprotiv, princip podele na grupe je praktične prirode i može biti po teritorijalnom principu ili nekom sličnom.

Neka je razmatrana populacija podeljena po nekom principu na više disjunktnih grupa. Za stratifikovani uzorak je potrebno iz svake grupe izabrati odredjeni broj elemenata populacije. Nasuprot tome, za grupni uzorak treba izabrati odredjeni broj grupa na slučajan način i uzeti sve elemente iz izabranih grupa u uzorak. Grupni uzorak se još zove uzorak skupina ili uzorak serija (serija u proizvodnji nekog artikla npr.). Osnovna jedinica izbora ovog tipa uzorka je grupa (za razliku od slučajnog uzorka i stratifikovanog uzorka kod kojih je osnovni element izbora bio element populacije).

**Primer 9.** Na teritoriji Srbije treba sprovesti anketu o gledanosti informativnog TV programa.

Nepostojanje upotrebljivih spiskova stanovnika Srbije u momentu sprovodjenja ankete samo je jedan od razloga koji onemogućava izbor prostog slučajnog uzorka ili stratifikovanog uzorka. Drugi bi razlog mogao biti ekonomski aspekt istraživanja, jer bi bilo neekonomično da personal koji anketira stanovništvo obilazi sve delove teritorije Srbije. Zbog svega toga je opravdano izvršiti grupisanje stanovništva po teritorijalnim jedinicama, recimo opštinama, pa anketirati sve stanovnike slučajno izabranih opstina. $\triangle$

Grupni uzorak podrazumeva da u definisanim grupama ima konačno mnogo elemenata populacije $N_1, N_2, \ldots, N_i, \ldots$ pa se takav uzorak može predstaviti slučajnim vektorom

$$X = (X_1^{(k_1)}, X_2^{(k_1)}, \ldots, X_{N_{k_1}}^{(k_1)}, X_1^{(k_2)}, X_2^{(k_2)}, \ldots, X_{N_{k_2}}^{(k_2)}, \ldots, X_1^{(k_l)}, X_2^{(k_l)}, \ldots, X_{N_{k_l}}^{(k_l)}),$$

gde je $k_i$ oznaka grupe sa ukupno $N_{k_i}$ elemenata populacije u sebi.

Izbor sa i bez vraćanja kod grupnog uzorka odnosio bi se na ponovni, ili ne, izbor istih grupa u uzorak.

### 2.4.3 Višeetapni uzorak

Grupni uzorak je po strukturi jednostavan, ali kada je obim grupa veliki može biti nepraktičan, ili davati manju tačnost. Povezivanje metoda grupnog i stratifikovanog uzorka daje nam ideju izbora uzorka u dve ili više etapa. Naime, u prvoj etapi od svih (disjunktnih) grupa na koje je populacija podeljena biramo na slučajan način odredjeni broj grupa, a zatim u drugoj etapi iz svake grupe izabrane u prvoj etapi biramo takodje na slučajan način odredjeni broj elemenata. Ovakav uzorak zove se dvoetapni uzorak.

**Primer 10.** Ako u prethodnom primeru ne vršimo anketiranje svih stanovnika odabranih opština, već odabranog dela stanovništva iz svake odabrane opštine (grupe) dobićemo dvoetapni uzorak. Prvu etapu čini izbor grupa, tj. opština iz kojih će se u drugoj etapi birati odredjeni elementi – stanovnici koji im pripadaju. (Ukoliko bi se u prvoj etapi izabrale sve postojeće grupe na koje je populacija podeljena, dvoetapni uzorak bi se sveo na stratifikovani.) △

**Primer 11.** Ako u odabranim opštinama u prvoj etapi prethodnog primera uočimo mesne zajednice, pa izaberemo na slučajan način odredjen broj mesnih zajednica iz odabranih grupa za dalju analizu, a zatim u trećoj etapi odaberemo na slučajan način po odredjeni broj stanovnika u uzorak za anketiranje dobićemo takodje troetapni uzorak.

Ukoliko bi grupe prve etape bile podeljene na disjunktne podgrupe, pa iz svake od grupa izaberemo u drugoj etapi podgrupe iz kojih ćemo tek u trećoj etapi birati elemente u uzorak, formirali bismo troetapni uzorak. △

Po istom principu može se formirati bilo koji višeetapni uzorak sa unapred definisanim konačnim brojem etapa.

Višeetapni izbor se još zove i *klasterizacija*.

### 2.4.4 Sistematski uzorak

Veoma pogodan metod izbora uzorka iz konačne uredjene populacije sastoji se u sledećem:

Neka je $N = nk$, gde je $n$ zadati obim uzorka i $k$ takodje prirodan broj. Uzima se slučajni broj izmedju 1 i $k$ (iz tablice slučajnih brojeva), pretpostavimo da je to broj $i$. Tada se uzorak formira od elemenata populacije čiji su redni brojevi

$$i, i+k, i+2k, \ldots, i+(n-1)k \quad ,$$

tj. uzorak sadrži prvi slučajno izabrani element populacije i svaki sledeći $k$–ti po redu brojeći od tog prvog.

Pogodnost ovog uzorka sastoji se u tome što prvi korak (izbor prvog elementa uzorka) odredjuje uzorak u celini. Medjutim, ova strategija izbora je primenjiva samo ukoliko je redosled elemenata populacije u uredjenju koje je na populaciji definisano **slučajan**. O ovome će još biti reči.

Uočimo da je za datu populaciju procedura sistematskog uzorka ustvari izbor jedne od $k$ grupa (na koje je podeljena cela populacija) sa verovatnoćom $\frac{1}{k}$. U ovom slučaju grupe odredjuju skupovi indekasa:

$$\{1, k+1, 2k+1, \ldots, (n-1)k+1\}, \ldots, \{i, k+i, 2k+i, \ldots, (n-1)k+i\}, \ldots, \{k, 2k, 3k, \ldots, nk\}.$$

Verovatnoća da ovakvim podskupom $i$–ti element populacije bude izabran u uzorak je $\pi_i = \frac{1}{k}$.

U slučaju kada je $N = nk+c$, $c < k$, $c \in N$, neke grupe sadrže $n$ elemenata populacije, a druge $n+1$ elemenata, tj. veličine grupa nisu iste od grupe do grupe, a verovatnoća izbora elemenata populacije u uzorak je takodje $\pi_i = \frac{1}{k}$ bilo da je uzorak obima $n$ ili $n+1$. (Naime, ako je slučajno izabrani redni broj prvog elementa koji treba uzeti u uzorak iz uredjene populacije $i$ takav da je $i < c < k$, opisanim pravilom dobiće se uzorak obima $n+1$.)

Sistematski uzorak se još zove *periodični* ili *mehanički* uzorak.

Primena tablice slučajnih brojeva nije od suštinskog značaja za izbor sistematskog uzorka. Otuda se često prvi element periodičnog uzorka bira kao središnji u prvom intervalu izbora.

Sistematski uzorak ima odredjene prednosti nad slučajnim uzorkom, jer je pravilo izbora sasvim prosto, ne zahteva tablice slučajnih brojeva, pa ni potpunu numeraciju populacije, a izvodi se znatno brže.

**Primer 12.** Treba proceniti učestanost javljanja alergijskog bronhitisa medju pacijentima jedne zdravstvene ustanove.

S obzirom da pacijente reprezentuju njihovi zdravstveni kartoni koji se nalaze u kartoteci zdravstvene ustanove, iz kartoteke treba uzeti kartone pacijenata koji će činiti uzorak uz pomoć lenjira definišući dužinsko rastojanje izmedju dva izabrana kartona. Pri tome su mala odstupanja od zadate dužine bez značaja, kao i to da li su kartoni uredno poredjani po brojevima.

Obratimo pažnju na činjenicu da je redosled pristizanja pacijenata u zdravstvenu ustanovu, čime je utvrdjen redosled otvorenih kartona, slučajan. Zbog toga redni broj kartona čini slučajno uredjenje u populaciji. $\triangle$

Sistematski uzorak je intuitivno prihvatljiv – "ravnomerno" je rasporedjen po populaciji, ne dopušta slučajna grupisanja ili "propuštanja" nekih delova populacije, što se kod slučajnog izbora može desiti.

Sistematski uzorak se može uporediti sa stratifikovanim uzorkom kod koga stratumi predstavljaju elemente na intervalu dužine $k$, pri čemu se iz svakog od njih bira po jedan element.

Sistematski uzorak se može koristiti i u kombinaciji sa ostalim metodima izbora uzorka. Na primer, kod stratifikovanog uzorka se elementi unutar stratuma mogu birati periodično. Kod grupnog uzorka se grupe mogu birati periodično. Kod višeetapnog se mogu kombinovati sistematski i slučajni izbor na više načina– u svakoj od etapa izbor može biti periodičan ili slučajan.

Medjitim, ako pretpostavimo da su elementi populacije slučajno rasporedjeni u niz, ili da je obeležje koje se posmatra nezavisno od rasporeda elemenata populacije, sistematski uzorak postaje samo jedan vid slučajnog uzorka bez vraćanja. Iako se ova logika veoma često koristi treba biti oprezan. Na primer, kada se prate sezonske pojave, tj. obeležja koja imaju sezonska kolebanja (kao što je temperatura vazduha, broj turista i sl.), može se desiti da se sezonska kolebanja u vrednosti obeležja poklope sa periodom izbora i daju pogrešnu sliku o obeležju. Zbog toga se o ovome mora voditi računa pri donošenju odluke o sistematskom izboru.

* * *

Konstatujmo da, svaka od strategija izbora ima za posledicu odredjenu tačnost u ocenjivanju nepoznatih parametara obeležja, kao i testiranju odgovarajućih hipoteza.

## 2.5 Empirijska funkcija raspodele

Vratimo se slučajnom uzorku uopšte i razmotrimo još neke važne pojmove vezane za uzorak.

Okosnica naučne oblasti koju zovemo matematičkom statistikom ili, jednostavno, statistikom, je funkcija od uzorka opisana sledećom definicijom:

DEFINICIJA 9. *Statistika* je funkcija od uzorka čiji analitički izraz ne zavisi od nepoznatih parametara obeležja, tj. funkcija od uzorka i poznatih konstanata.

Primeri nekih statistika su:

$$T_n = \sum_{i=1}^{n} X_i \quad - \text{total uzorka}$$

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \quad - \text{sredina uzorka}$$

$$\overline{S}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \quad - \text{disperzija uzorka}$$

$$\overline{S}_n = \sqrt{\overline{S}_n^2} \quad - \text{uzoračka standardna devijacija}$$

$$\tilde{S}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \quad - \text{popravljena disperzija uzorka}$$

$$R = X_{max} - X_{min} \quad - \text{raspon uzorka}\,.$$

Za dva obeležja $X$ i $Y$ i uzorak $((X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n))$ iz populacije na kojoj se posmatra dvodimenziono obeležje $(X, Y)$ može se definisati statistika

$$R_{XY} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)(Y_i - \overline{Y}_n)}{\overline{S}_X \overline{S}_Y} \quad - \text{uzoračka koeficijent korelacije}\,,$$

gde su sa $\overline{S}_X$ i $\overline{S}_Y$ označene uzoračke standardne devijacije za obeležja $X$ i $Y$ redom.

Posebno mesto medju statistikama imaju tzv. statistike poretka. Ove se statistike definišu posredstvom varijacionog niza:

DEFINICIJA 10. *Varijacioni niz* čine elementi uzorka poredjani u neopadajućem poretku.

Za uzorak $(X_1, X_2, \ldots, X_n)$ varijacioni niz čini niz slučajnih promenljivih sačinjen od elemenata ovog uzorka u oznaci $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ za koji važi

$$X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)} \quad .$$

Za realizovane vrednosti varijacionog niza koristi se isti termin *varijacioni niz*, bez opasnosti od zabune, a označavaju se malim slovima:

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)} \quad .$$

Definicija 11. *Statistika poretka reda k* uzorka obima $n$, $1 \leq k \leq n$, je $k$–ti element varijacionog niza posmatranog uzorka, dakle slučajna promenljiva $X_{(k)}$.

Neka je uzorak $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ prost slučajni uzorak iz populacije sa obeležjem $X$ čija je funkcija raspodele $F$. **U definisanju funkcije raspodele biće sve vreme korišćena neprekidnost s desna.** Za svako $x \in R$ definisaćemo slučajnu veličinu $\mu_n(x)$ kao broj elemenata uzorka $\mathbf{X}$ koji su manji ili jednaki $x$, tj.

Definicija 12.

$$\mu_n(x) = card\{j | X_j \leq x, j = 1, 2, \ldots, n\} \quad , \quad x \in R \quad .$$

Nadalje se može definisati slučajna promenljiva $S_n(x)$ koja daje vrednosti slučajne promenljive $\mu_n(x)$ u relativnom odnosu prema obimu uzorka:

Definicija 13. *Empirijska funkcija raspodele* uzorka $\mathbf{X}$ je statistika

$$S_n(x) \stackrel{\text{def}}{=} \frac{\mu_n(x)}{n} \quad , \quad x \in R \quad .$$

Slučajna promenljiva $S_n(x)$ je statistika čiji je kodomen skup

$$\{0, 1/n, 2/n, \ldots, (n-1)/n, 1\}$$

ili njegov pravi podskup sa verovatnoćama

$$P\{S_n(x) = k/n\} = P\{\mu_n(x) = k\} = \binom{n}{k} (F(x))^k (1 - F(x))^{n-k}, \quad k = 0, 1, \ldots, n.$$

Ovo otuda što, prema definiciji, slučajna promenljiva $\mu_n(x)$ ima binomnu raspodelu, $\mathcal{B}(n, p)$ sa $p = P\{X \leq x\} = F(x)$, $x \in R$. Statistiku $S_n(x)$ možemo posmatrati i kao aritmetičku sredinu indikatora

$$I_{A_i} = \begin{cases} 1, & \omega \in A_i \\ 0, & \omega \notin A_i \end{cases} \quad ,$$

$A_i = \{\omega | X_i(\omega) \leq x\}$, a s obzirom da je $E(I_{A_i}) = F(x)$ za fiksirano $x \in R$, važi teorema:

**Teorema 2.5.1** *Za fiksirano $x \in R$, $S_n(x) \longrightarrow F(x)$, $n \to \infty$ skoro izvesno, tj.*

$$P\{S_n(x) \to F(x)\,,\, n \to \infty\} = 1.$$

**Dokaz.**Tvrdjenje sledi na osnovu Borelovog zakona velikih brojeva.□

Za realizovani uzorak $(x_1, x_2, \ldots, x_n)$, $S_n(x)$, $x \in R$, je monotono neopadajuća funkcija sa mogućim skokovima u tačkama varijacionog niza $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$:

$$S_n(x) = \frac{k}{n}, \quad x \in [x_{(k)}, x_{(k+1)}), k = 0, 1, \ldots, n \quad .$$

Pri tome su uvedene oznake $x_{(0)} = -\infty$, i u tom slučaju je i leva granica intervala otvorena, i $x_{(n+1)} = +\infty$. Ukoliko su svi elementi u realizovanom uzorku različiti, skokovi su veličine $1/n$.

Konvergencija o kojoj je bilo reči u prethodnoj teoremi, ostvaruje se i uniformno po $x \in R$. O tome govori tzv. centralna teorema matematičke statistike. Jedan od njenih oblika je sledeći.

**Teorema 2.5.2 (Glivenko-Kanteli)** *Neka je $F$ funkcija raspodele obeležja $X$ i $S_n(x)$, $x \in R$, empirijska funkcija raspodele uzorka obima $n$ iz populacije sa obeležjem $X$. Tada važi*

$$P\{\sup_{x \in R} |S_n(x) - F(x)| \to 0\,,\, n \to \infty\} = 1. \tag{2.1}$$

**Dokaz.** Neka je $F$ proizvoljna funkcija raspodele obeležja diskretnog ili apsolutno neprekidnog tipa i neka je $\varepsilon$ proizvoljan realan broj za koji važi $0 < \varepsilon < 1$. Za tako izabrano $\varepsilon$ i zadatu funkciju $F$, moguće je izabrati konačan broj tačaka

$$z_0, z_1, \ldots, z_N \in \overline{R} \qquad (\overline{R} = R \cup \{-\infty, +\infty\})$$

takvih da je

$$-\infty = z_0 < z_1 < \ldots < z_{N-1} < z_N = +\infty$$
$$F(z_k - 0) - F(z_{k-1}) \leq \varepsilon, \quad k = 1, \ldots, N.$$

Na primer, moguće je izabrati skup $\{z_j\}$ tako da on sadrži sve tačke prekida (ako ima tačaka prekida) funkcije $F$ u kojima je skok funkcije $F$ veći od $\varepsilon/2$. Tada za proizvoljno $z \in [z_{k-1}, z_k)$ važi

$$S_n(z) - F(z) \leq S_n(z_k - 0) - F(z_{k-1}) \leq S_n(z_k - 0) - F(z_k - 0) + \varepsilon.$$

Slično i

$$S_n(z) - F(z) \geq S_n(z_{k-1}) - F(z_k - 0) \geq S_n(z_{k-1}) - F(z_{k-1}) - \varepsilon.$$

Definišimo sledeće skupove

$$\overline{B}_k = \{\omega | (S_n(z_k - 0))(\omega) \longrightarrow F(z_k - 0),\ \ n \to \infty\}$$

$$B_k = \{\omega | (S_n(z_k))(\omega) \longrightarrow F(z_k),\ \ n \to \infty\}$$

$$B = \bigcap_{k=0}^{N} B_k \overline{B}_k.$$

Tada, prema prethodnoj teoremi, dogadjaji $B_k$ i $\overline{B}_k$ se realizuju skoro izvesno, tj.

$$P(B_k) = P(\overline{B}_k) = 1.$$

Otuda je

$$P(B) = 1.$$

Ovo s toga što se za svako $\omega \in B$ može naći uzorak dovoljno velikog obima $n(\omega)$ takav da kadgod je $n \geq n(\omega)$ tada je $B_0\overline{B}_0 \subset B_1\overline{B}_1 \subset \ldots \subset B_N\overline{B}_N$.

Dakle, za dovoljno veliko $n \geq n(\omega)$, biće

$$|S_n(z_k - 0) - F(z_k - 0)| < \varepsilon, \quad k = 0, 1, \ldots, N$$

i

$$|S_n(z_k) - F(z_k)| < \varepsilon, \quad k = 0, 1, \ldots, N,$$

za svako $k$, te je

$$\sup_{z \in R} |S_n(z) - F(z)| \leq 2\varepsilon.$$

Time je teorema dokazana.□

Sledeće dve teoreme govore o raspodeli važnih statistika baziranih na empirijskoj funkciji raspodele. Ovde ćemo ih navesti bez dokaza.

**Teorema 2.5.3 (Kolmogorova)** *Ako je funkcija $F$ neprekidna, tada za proizvoljno fiksirano $t > 0$ statistika $D_n = \sup_{x \in R} |S_n(x) - F(x)|$ ima raspodelu za koju važi*

$$\lim_{n \to \infty} P\{\sqrt{n} D_n \leq t\} = K(t) = \sum_{j=-\infty}^{+\infty} (-1)^j e^{-2j^2t^2}.$$

**Teorema 2.5.4 (Smirnova)** *Neka su $S_{1n_1}$ i $S_{2n_2}$ dve empirijske funkcije raspodele sačinjene na osnovu dva nezavisna uzorka obima $n_1$ i $n_2$ iz iste populacije sa obeležjem $X$ i*

$$D_{n_1 n_2} = \sup_{x \in R} |S_{1n_1}(x) - S_{2n_2}(x)|.$$

*Tada, ako je teorijska funkcija raspodele $F$ neprekidna, za proizvoljno fiksirano $t > 0$,*

$$\lim_{n_1, n_2 \to \infty} P\left\{ \sqrt{\frac{n_1 n_2}{n_1 + n_2}} D_{n_1 n_2} \leq t \right\} = K(t).$$

## 2.6 Sredjivanje i prikazivanje realizovanih uzoraka

Eksperimentalni podaci se radi statističke obrade predstavljaju na dva osnovna načina: *tabl̃ično* i *grafički*. Tablični metod daje podatke sredjene u obliku tabele. Grafički metod te tabele ilustruje prigodnim skicama, kartama, grafikonima...

### 2.6.1 Tablični metod prikaza podataka – organizovanje baza podataka

**Tablice kvantitativnih obeležja**

Niz dobijenih podataka poredjanih u rastućem poretku (rangiranih) daje varijacioni niz obeležja. On pruža polaznu osnovu za dalja razmatranja u vezi sa raspodelom.

**Primer 13.** U 20 odeljenja nižih razreda osnovne škole registrovan je broj učenika sa natprosečnim sposobnostima: 5, 6, 8, 10, 9, 8, 4, 7, 7, 3, 6, 4, 8, 7, 6, 6, 5, 3, 6, 6. Varijacioni niz uzorka je: 3, 3, 4, 4, 5, 5, 6, 6, 6, 6, 6, 6, 7, 7, 7, 8, 8, 8, 9, 10. Za odredjivanje **raspodele obeležja** koristi se sledeća tabela:

| broj učenika sa natprosečnim sposobnostima | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|
| $f = k$ | 2 | 2 | 2 | 6 | 3 | 3 | 1 | 1 |
| $f^* = k/n$ | 0,1 | 0,1 | 0,1 | 0,3 | 0,15 | 0,15 | 0,05 | 0,05 |
| $f^*_{\%} = k/n$ [%] | 10 | 10 | 10 | 30 | 15 | 15 | 5 | 5 |
| $\Sigma f = n_x$ | 2 | 4 | 6 | 12 | 15 | 18 | 19 | 20 |
| $\Sigma f^* = n_x/n$ | 0,1 | 0,2 | 0,3 | 0,6 | 0,75 | 0,9 | 0,95 | 1 |
| $\Sigma f^*_{\%} = n_x/n$ [%] | 10 | 20 | 30 | 60 | 75 | 90 | 95 | 100 |

U tabeli su korišćene oznake: $k$–broj odeljenja sa posmatranim brojem natprosečnih učenika, $f$–apsolutna učestanost, $f^*$–relativna učestanost, $f^*_{\%}$–procentualna učestanost, $n_x$–broj odeljenja sa ne više od $x$ natprosečnih učenika, $\Sigma f$–zbirna učestanost (kumulativna), $\Sigma f^*$–zbirna relativna učestanost i $\Sigma f^*_{\%}$–zbirna procentualna učestanost. $\triangle$

Primetimo da je broj $\frac{n_x}{n}$, zapravo, realizovana vrednost empirijske funkcije raspodele za zadato $x \in R$.

Kod obeležja apsolutno neprekidnog tipa ili diskretnih obeležja sa velikim brojem različitih vrednosti, podaci se sredjuju po unapred odabranim intervalima. Broj i raspored tih intervala zavisi od broja podataka i samog obeležja.

**Primer 14.** Beležene su minimalne jačine struje koje predstavljaju prag osetljivosti jednog mišića 60 posmatranih pacijenata i dobijeni su sledeći rezultati:

| Red.br. | Jačina (mA) | Red.br. | Jačina (mA) | Red.br. | Jačina (mA) |
|---|---|---|---|---|---|
| 1 | 7,80 | 21 | 6,23 | 41 | 6,36 |
| 2 | 9,28 | 22 | 7,27 | 42 | 5,98 |
| 3 | 8,70 | 23 | 6,98 | 43 | 5,16 |
| 4 | 5,30 | 24 | 4,84 | 44 | 11,40 |
| 5 | 5,63 | 25 | 10,53 | 45 | 8,59 |
| 6 | 6,54 | 26 | 8,00 | 46 | 8,12 |
| 7 | 7,80 | 27 | 7,28 | 47 | 10,30 |
| 8 | 7,73 | 28 | 9,16 | 48 | 10,80 |
| 9 | 6,76 | 29 | 5,02 | 49 | 11,87 |
| 10 | 12,06 | 30 | 8,08 | 50 | 11,62 |
| 11 | 16,44 | 31 | 3,95 | 51 | 11,34 |
| 12 | 9,55 | 32 | 6,77 | 52 | 9,50 |
| 13 | 3,71 | 33 | 5,24 | 53 | 6,43 |
| 14 | 8,97 | 34 | 6,32 | 54 | 6,21 |
| 15 | 7,38 | 35 | 9,64 | 55 | 10,42 |
| 16 | 5,02 | 36 | 10,97 | 56 | 7,71 |
| 17 | 5,18 | 37 | 8,79 | 57 | 6,55 |
| 18 | 7,51 | 38 | 7,93 | 58 | 7,33 |
| 19 | 4,92 | 39 | 7,91 | 59 | 7,25 |
| 20 | 4,82 | 40 | 14,52 | 60 | 4,92 |

Strogog pravila za izbor broja i dužine intervala nema, ali je moguće upravljanje po formuli koja preporučuje $k$ intervala, gde je

$$k \geq 1 + 3,322 \log_{10} n = 1 + \log_2 n$$

za obim uzorka n. Medjutim, ne preporučuje se više od $5 \cdot \log_{10} n$ intervala, tj.

$$k \leq 5 \cdot \log_{10} n.$$

Za razmatrani primer, obim uzorka je $n = 60$, pa je donja granica broja intervala jednaka $k = 1 + \log_2 60 \approx 7$, a gornja granica broja intervala je $k \leq 5 \cdot \log_{10} 60 \approx 9$. Znači, može se uzeti 7, 8 ili 9 intervala.

Broj intervala $k$ se može odrediti i na jedan od sledećih načina: $k \approx \sqrt{n}$, $k \approx 2\sqrt[3]{n}$ ili $k \approx 5 \log_{10} n$.

Bez obzira na način odredjivanja broja intervala, dužine intervala se odredjuju na sledeći način. Odredjuju se najmanja $x_{min}$ i najveća $x_{max}$ vrednost u realizovanom uzorku, a zatim se dužina intervala računa po formuli:

$$h = \frac{x_{max} - x_{min}}{k},$$

pri čemu se vodi računa da su granice intervala jednostavne za rad (celi brojevi, brojevi deljivi sa 5 i sl.).

Uzmimo da je broj intervala 8. Zatim ćemo odrediti najmanju i najveću vrednost uzorka. One su redom 3,71 i 16,44. Tada je dužina intervala jednaka

$$h = \frac{16,44 - 3,71}{8} = 1,591.$$

Kako dobijeni broj nije pogodan za rad, to se može uzeti drugi pogodniji broj, recimo 2. U donjoj tabeli odredjeni su intervali, apsolutne, relativne, zbirne i zbirne relativne učestanosti minimalnih struja razmatranog niza:

| interval | sredina intervala | $f$ | $f^* = \frac{f}{n}$ | $f^*_{\%}$ | $\Sigma f$ | $\Sigma f^*$ | $\Sigma f^*_{\%}$ |
|---|---|---|---|---|---|---|---|
| [2,4) | 3 | 2 | 0,03 | 3 | 2 | 0,03 | 3 |
| [4,6) | 5 | 12 | 0,20 | 20 | 14 | 0,23 | 23 |
| [6,8) | 7 | 22 | 0,36 | 36 | 36 | 0,59 | 59 |
| [8,10) | 9 | 12 | 0,20 | 12 | 48 | 0,79 | 79 |
| [10,12) | 11 | 9 | 0,15 | 15 | 57 | 0,94 | 94 |
| [12,14) | 13 | 1 | 0,02 | 2 | 58 | 0,96 | 96 |
| [14,16) | 15 | 1 | 0,02 | 2 | 59 | 0,98 | 98 |
| [16,18] | 17 | 1 | 0,02 | 2 | 60 | 1,00 | 100 |

.

△

**Primer 15.** 50 studenata je polagalo ispit iz statistike i dobijeni su sledeći rezultati po broju osvojenih poena od mogućih 100: 17, 73, 85, 43, 36, 21, 0, 35, 50, 32, 75, 21, 78, 41, 92, 70, 80, 84, 55, 42, 79, 45, 98, 62, 46, 45, 79, 2, 17, 19, 49, 42, 32, 6, 8, 39, 4, 28, 48, 86, 26, 60, 92, 15, 85, 26, 14, 69, 55, 94. Podaci se mogu srediti na sledeći način:

Ukupno ima $n = 50$ podataka. Najmanji dozvoljeni broj intervala je $k = 1 + 3.322 \log_{10} 50 = 6.64 \approx 7$, kada se broj zaokruži. Najveći dozvoljeni broj intervala je $5 \cdot \log_{10} 50 = 8.49 \approx 8$. Tako se može raditi sa 7 ili 8 intervala.

Neka je broj intervala $k = 7$. Najmanja i najveća vrednost uzorka su $x_{min} = 0$ i $x_{max} = 98$, tako da je dužina intervala

$$h = \frac{98 - 0}{7} = 14.$$

Sada se podaci grupišu po intervalima: $[0, 14)$, $[14, 28)$, $[28, 42)$, $[42, 56)$, $[56, 70)$, $[70, 84)$ i $[84, 98]$, i dobija se sledeća tabela:

| Broj bodova | $[0, 14)$ | $[14, 28)$ | $[28, 42)$ | $[42, 56)$ | $[56, 70)$ | $[70, 84)$ | $[84, 98]$ |
|---|---|---|---|---|---|---|---|
| Broj studenata ($k$) | 5 | 9 | 7 | 11 | 3 | 7 | 8 |
| $f^* = k/n$ | 0,1 | 0,18 | 0,14 | 0,22 | 0,06 | 0,14 | 0,16 |
| $f^*_{\%} = k/n$ [%] | 10 | 18 | 14 | 22 | 6 | 14 | 16 |
| $\Sigma f = n_x$ | 5 | 14 | 21 | 32 | 35 | 42 | 50 |
| $\Sigma f^* = n_x/n$ | 0,1 | 0,28 | 0,42 | 0,64 | 0,7 | 0,84 | 1 |
| $\Sigma f^*_{\%} = n_x/n$ [%] | 10 | 28 | 42 | 64 | 70 | 84 | 100 |

.

△

Intervali **ne moraju** biti jednakih dužina, što preporučuje sâmo konkretno obeležje.

Najčešće, sredine intervala reprezentuju realizovane vrednosti obeležja kod izračunavanja realizovanih vrednosti statistika (o čemu će još biti reči). Kao sredina intervala $[a, b)$, ali takodje i intervala $[a, b]$ koristi se broj $(b + a)/2$. Ovo otuda što je kod obeležja apsolutno neprekidnog tipa verovatnoća realizacije pojedine tačke sa realne prave jednaka 0.

### Tablice kvalitativnih obeležja

U slučaju kvalitativnog obeležja može se takodje sačiniti tabela.

**Primer 16.** Testom za proveru motornih sposobnosti je meren nivo sposobnosti učenika jednog odeljenja i dobijeni rezultati su svrstani u tri kategorije: nizak (n), srednji (s) i visok (v) nivo sposobnosti. U odeljenju je registrovan sledeći niz podataka: n, n, s, v, s, s, s, n, v, v, s, s, s, n, v, v, v, s, v, n, n, s, v, s. Na osnovu niza realizacija dobijena je tabela

| nivo motornih sposobnosti | n | s | v | $\Sigma$ |
|---|---|---|---|---|
| $f$ broj učenika | 6 | 10 | 8 | 24 |
| $f^*$ relativna učestanost | 0,25 | 0,42 | 0,33 | 1 |

.

△

**Tablice za dvodimenzionalno obeležje**

Ukoliko se posmatraju dva obeležja $X$ i $Y$ istovremeno (koja su moguće zavisna, tj. dvodimenzionalno obeležje) tabela je oblika:

| $X\backslash Y$ | | | |
|---|---|---|---|
| | | | |
| | | | |
| | | | |

.

Dobijena tabela se naziva i tabela kontingencije. Sam postupak formiranja tabele je jednostavan. Ukoliko drugačije nije naglašeno, postupak je sledeći: odredjuju se intervali za svako obeležje posebno a zatim se realizovani uzorak grupiše po dobijenim intervalima.

Naravno, ukoliko se radi o diskretnom obeležju kao nekoj od komponenata ili obema komponentama posmatranog dvodimenzionalnog obeležja, utvrdjuju se apsolutne (relativne, procentualne) učestanosti odgovarajućih parova u realizovanom uzorku i unose u tabelu.

**Primer 17.** 38 osoba konkuriše za jednu vrstu posla. Poslodavca zanima njihova stručna i intelektualna sposobnost. Zbog toga ove osobe rade testove stručnosti (TS) i inteligencije (TI). Dobijeni su sledeći rezultati:

| Redni broj kandidata | TS | TI | Redni broj kandidata | TS | TI |
|---|---|---|---|---|---|
| 1 | 70 | 112 | 20 | 55 | 120 |
| 2 | 75 | 121 | 21 | 60 | 100 |
| 3 | 80 | 100 | 22 | 58 | 102 |
| 4 | 85 | 102 | 23 | 60 | 104 |
| 5 | 75 | 120 | 24 | 74 | 97 |
| 6 | 48 | 98 | 25 | 48 | 94 |
| 7 | 52 | 111 | 26 | 53 | 89 |
| 8 | 50 | 120 | 27 | 58 | 129 |
| 9 | 51 | 105 | 28 | 79 | 116 |
| 10 | 55 | 110 | 29 | 82 | 145 |
| 11 | 46 | 134 | 30 | 84 | 130 |
| 12 | 87 | 100 | 31 | 81 | 115 |
| 13 | 72 | 99 | 32 | 52 | 120 |
| 14 | 70 | 91 | 33 | 55 | 109 |
| 15 | 63 | 101 | 34 | 68 | 110 |
| 16 | 56 | 104 | 35 | 73 | 112 |
| 17 | 60 | 115 | 36 | 46 | 121 |
| 18 | 72 | 116 | 37 | 52 | 130 |
| 19 | 78 | 119 | 38 | 80 | 90 |

.

Posmatraju se dva obeležja: stručnost i inteligencija, posredstvom testova kao mernih instrumenata za posmatrana obeležja. Broj intervala za oba obeležja može biti 6, 7 ili 8. Neka se za svako od obeležja podaci grupišu u po 6 intervala. Sada se odredjuju dužine intervala za svako obeležje posebno. Za prvo obeležje (rezultati testa stručnosti) dužine intervala su $h_1 = (87 - 46)/6 = 6,83 \approx 7$, tako da se dobijaju intervali: $[46, 53)$, $[53, 60)$,

$[60, 67)$, $[67, 74)$, $[74, 81)$ i $[81, 87]$. Za drugo obeležje (rezultati testa inteligencije) dužine intervala su $h_2 = (145 - 89)/6 = 9,33$ te neka je $h_2 = 10$ radi lakšeg računanja. Za drugi test se dobijaju sledeći intervali: $[89, 99)$, $[99, 109)$, $[109, 119)$, $[119, 129)$, $[129, 139)$ i $[139, 145]$.

Sada se vrši prebrojavanje podataka po intervalima iz sledeće tabele:

| TS\TI | $[89, 99)$ | $[99, 109)$ | ... | $[129, 139)$ | $[139, 145]$ |
|---|---|---|---|---|---|
| $[46, 53)$ | | | ... | | |
| $[53, 60)$ | | | ... | | |
| $[60, 67)$ | | | ... | | |
| $[67, 74)$ | | | ... | | |
| $[74, 81)$ | | | ... | | |
| $[81, 87]$ | | | ... | | |

.

Odgovarajuće apsolutne učestanosti su:

| TS\TI | $[89, 99)$ | $[99, 109)$ | ... | $[129, 139)$ | $[139, 145]$ |
|---|---|---|---|---|---|
| $[46, 53)$ | 2 | 1 | ... | 2 | 0 |
| $[53, 60)$ | 1 | 2 | ... | 1 | 0 |
| $[60, 67)$ | 0 | 3 | ... | 0 | 0 |
| $[67, 74)$ | 1 | 1 | ... | 0 | 0 |
| $[74, 81)$ | 2 | 1 | ... | 0 | 0 |
| $[81, 87]$ | 0 | 2 | ... | 1 | 1 |

.

$\triangle$

### 2.6.2 Grafički metodi prikaza podataka

Raspodela obeležja grafički se prikazuje preko (običnih) učestanosti ili preko zbirnih učestanosti (naročito zbirnih relativnih učestanosti, tj. empirijske funkcije raspodele).

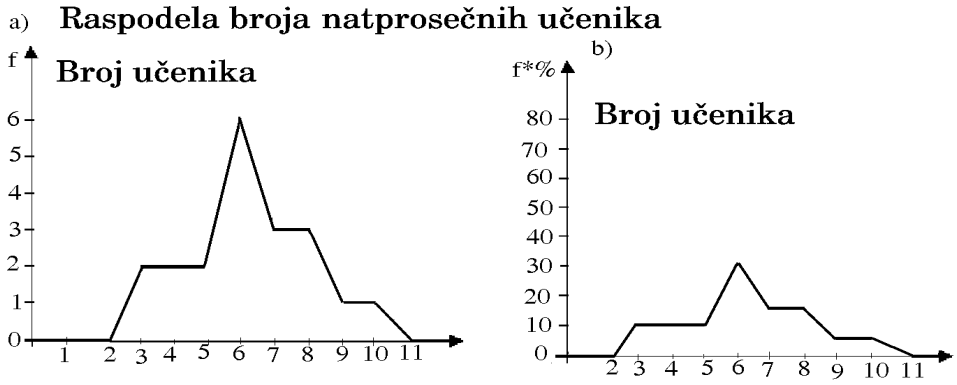
Grafički metodi prikaza podataka su najčešće: poligon, histogram, kumulativna kriva, razni dijagrami i slično.

#### A. Grafici kvantitativnih obeležja

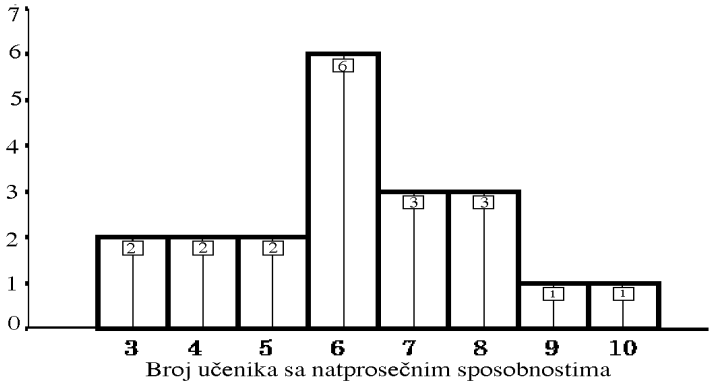
Na slikama od 2.1 do 2.3 prikazani su podaci koji se odnose na primer 13. Figure na slikama 2.1 a), b) i 2.4 b) su poligoni, na slikama 2.2, 2.3 b) su trakasti dijagrami, a na slici 2.4 a) je histogram.

1. **Histogram**

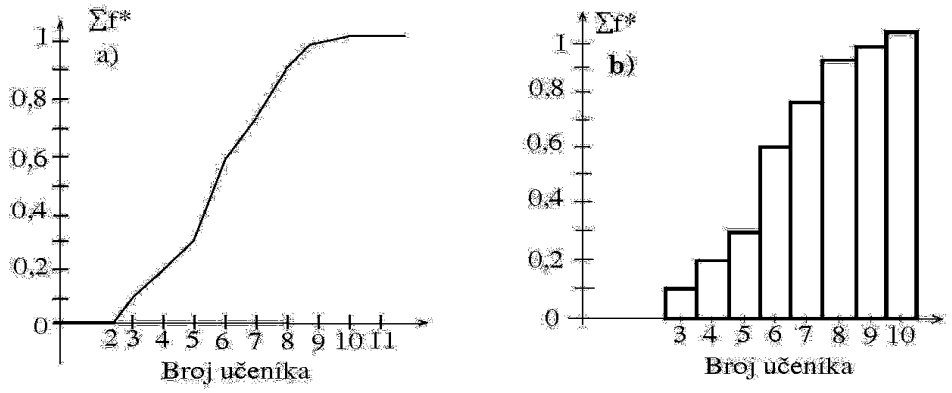
   Histogram se može primenjivati samo za grafičko prikazivanje realizovanog uzorka iz populacije sa obeležjem $X$ apsolutno neprekidnog tipa. Za uzorak koji je u tom slučaju intervalno sredjen, podaci se prikazuju na sledeći način. Oblast vrednosti posmatranog obeležja je razbijena na intervale dužine $h$. Ovi intervali se prikazuju na apscisnoj osi koordinatnog sistema pripremljenog za grafičko predstavljanje realizovanog uzorka $\mathbf{x}$ obima $n$. Nad svakim od intervala se konstruiše pravougaonik čija je visina $\nu/(nh)$, odnosno površina $\nu/n$, gde je $\nu$ broj elemenata realizovanog uzorka koji pripadaju uočenom intervalu. Figura koja predstavlja uniju upravo konstruisanih pravougaonika zove se histogram relativnih učestanosti.



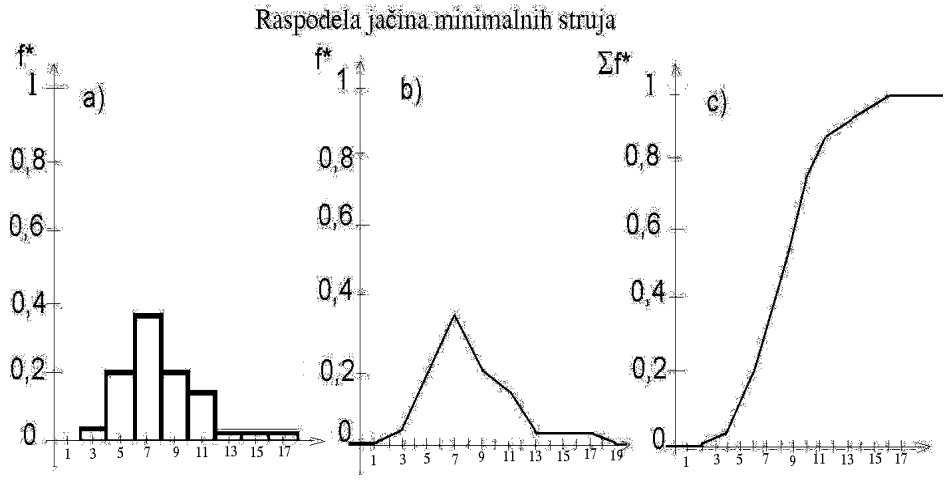

Slika 2.1: Poligoni: a) apsolutnih učestanosti; b) relativnih učestanosti u %.



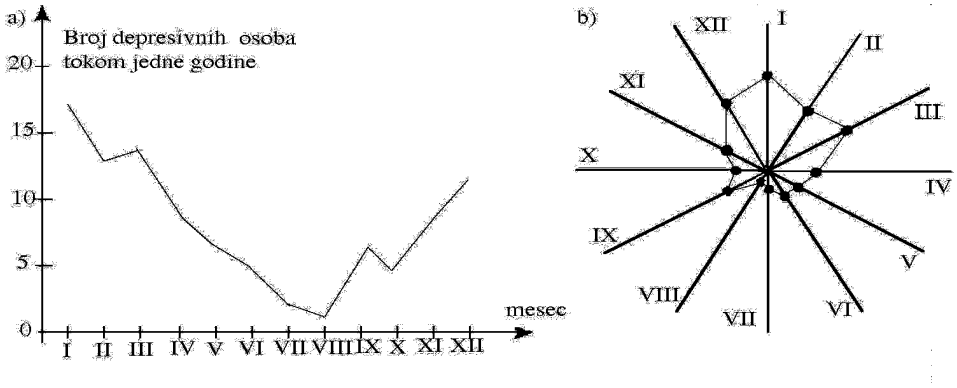

Slika 2.2: Trakasti dijagram apsolutnih učestanosti iz primera 13



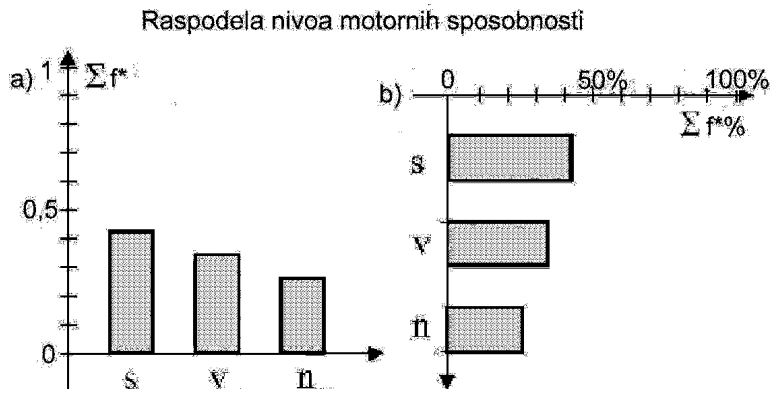

Slika 2.3: a) Kumulativna kriva relativnih učestanosti (ogiva relativnih učestanosti); b) trakasti dijagram zbirnih relativnih učestanosti iz primera 13



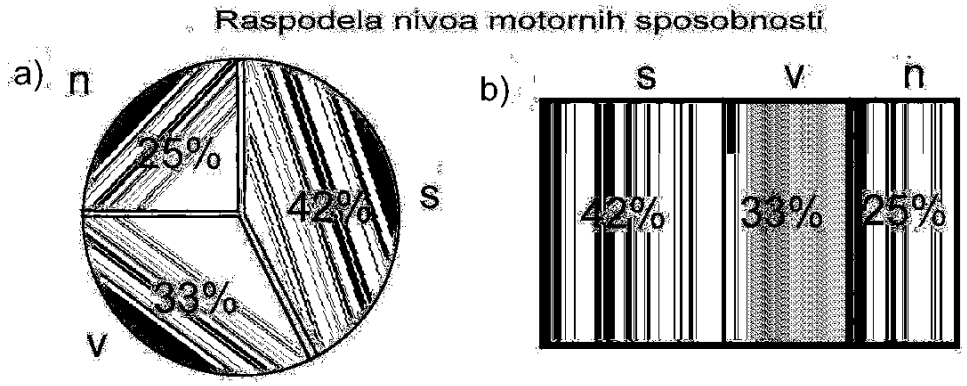

Slika 2.4: a) Histogram relativnih učestanosti; b) poligon relativnih učestanosti; c) kumulativna kriva iz primera 14



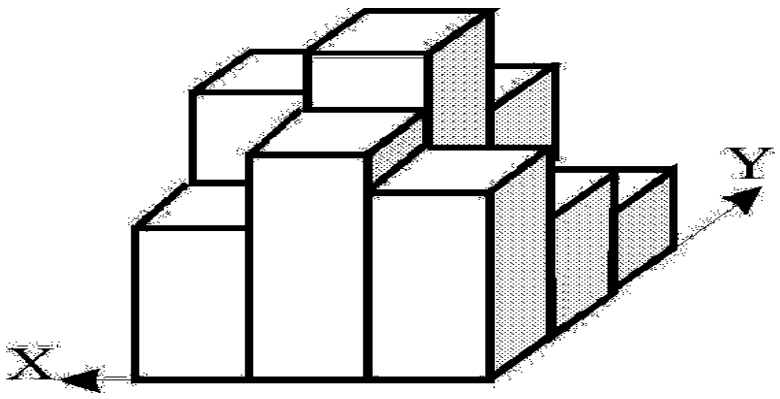

Slika 2.5: Ilustracija pojave koja ima ciklični karakter: a) linearni dijagram; b) zvezdasti dijagram.




Slika 2.6: a) Vertikalni i b) horizontalni trakasti dijagram za primer16




Slika 2.7: Podela a) kruga ("pita") i b) pravougaonika za prikazivanje učestanosti iz primera 16




Slika 2.8: Stereogram

Za slučajni uzorak $\mathbf{X}$ obima $n$, količnik $\nu/n$ je slučajna promenljiva. Ako se ima u vidu da je $n$ proizvoljan prirodan broj, može se govoriti o nizu slučajnih promenljivih za koji važi Bernulijev zakon velikih brojeva,

$$P\left\{\left|\frac{\nu}{n}-p\right|>\varepsilon\right\}\longrightarrow 0,\quad n\to\infty,$$

gde je $p$ verovatnoća da obeležje $X$ ima vrednost u odgovarajućem intervalu. Ako je dužina intervala $h$ dovoljno mala, a gustina $f$ obeležja neprekidna, tada je ta verovatnoća približno jednaka $f(z)h$, gde je $z$ sredina odgovarajućeg intervala. To znači da je pri velikom obimu uzorka i maloj dužini intervala, visine konstruisanih pravougaonika moguće posmatrati kao približne vrednosti gustine raspodele koje odgovaraju sredinama intervala, odnosno, gornja granica histograma se može posmatrati kao statistički analogon gustine raspodele posmatranog obeležja.

Važno je, medjutim, naglasiti da je histogram primenjiv samo u početnoj fazi istraživanja. Ovo otuda što se ne smeju zanemariti njegovi nedostaci, a to su neodredjenost u načinu formiranja intervala i gubitak informacija pri grupisanju podataka, jer se koristi samo broj koji pokazuje koliko se je elemenata realizovanog uzorka našlo u odredjenom intervalu, a ne i sami elementi uzorka.

Analogno opisanom postupku može se konstruisati i histogram zbirnih (kumulativnih) relativnih učestanosti, čija bi se gornja granica mogla posmatrati kao statistički analogon funkcije raspodele posmatranog obeležja.

Veoma često primenjuje se i histogram odgovarajućih procentualnih učestanosti.

2. **Poligon**

Iz histograma relativnih i histograma zbirnih relativnih učestanosti dobijaju se poligon relativnih i poligon zbirnih relativnih učestanosti. Oba se konstruišu tako što se sredine gornjih stranica susednih pravougaonika odgovarajućeg histograma spoje dužima, čime uočene sredine postaju temena poligonalnih linija. Poligon relativnih učestanosti "počinje" temenom na apscisnoj osi iz sredine intervala koji neposredno prethodi prvom intervalu u kome je učestanost različita od nule, i "završava" se takodje na istoj osi temenom koje je sredina neposredno susednog intervala poslednjem intervalu u kome je relativna učestanost takodje različita od nule. Poligon zbirnih relativnih učestanosti "počinje" kao i prethodni, a "završava" nad intervalom u kome je zbirna učestanost jednaka jedinici. Oba su poligona otvorena i logički se nastavljaju horizontalnim polupravim na oba kraja.

Poligon zbirnih relativnih učestanosti se zove i **kumulativna kriva** ili **ogiva**. Može se vršiti izgladjivanje kumulativne krive pri čemu se umesto poligonalne linije dobija glatka kriva koja prolazi kroz temena ove poligonalne linije.

3. **Dijagram**

Dijagrami se koriste za grafičko predstavljanje realizovanih vrednosti obeležja diskretnog tipa kvantitativnih ili kvalitativnih. Mogu biti linijski i površinski. U linijske dijagrame spadaju i svi poligoni učestanosti, ali i više od toga. Na pr. zvezdasti dijagram (videti sliku 2.5 b)). Površinski dijagram može biti trakasti

(slike 2.2, 2.3 i 2.6) ili kružni (slika 2.7 a)) ili kvadratni, odnosno, pravougaoni (slika 2.7 b)) i slično. Površinski dijagrami se rade po principu delova površi srazmernih odgovarajućim učestanostima.

## 2.7 Modeliranje raspodela metodom Monte Karlo

Opšti princip statističkog modelovanja (simulacije) bi se sastojao u sledećem:

*Treba odrediti približnu vrednost neke realne veličine a. U tom cilju bira se slučajna veličina X sa raspodelom takvom da je $E(X) = a$. Na osnovu realizovanog uzorka*

$$\mathbf{x} = (x_1, x_2, \ldots, x_n)$$

*iz populacije sa obeležjem X, odredjuje se približna vrednost veličine a kao ocena matematičkog očekivanja $E(X)$,*

$$\hat{a} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad .$$

Više reči o samoj oceni biće u narednom poglavlju. Ovde treba prokomentarisati sa kojom tačnošću se vrši ovakvo ocenjivanje. Naime, prema centralnoj graničnoj teoremi,

$$P\{|\overline{X}_n - a| \leq \varepsilon\} \approx 2\Phi(\varepsilon\frac{\sqrt{n}}{\overline{s_n}}),$$

gde je

$$\overline{s_n} = \sqrt{\overline{s_n}^2},$$

odnosno

$$\overline{s_n}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x_n})^2.$$

Tačnost ove ocene je reda $1/\sqrt{n}$, što se najčešće ne smatra velikom tačnošću.

Nadalje ćemo se baviti praktičnim rešavanjem problema statističkog modelovanja same raspodele obeležja $X$ koje je u definiciji problema istaknuto.To podrazumeva modeliranje realizovanog uzorka iz populacije sa ovim obeležjem.

### 2.7.1 Modeliranje diskretne raspodele sa konačno mnogo vrednosti

Kada je obeležje $X$ sa diskretnom raspodelom sa konačno mnogo vrednosti, zadatak se sastoji u tome da je potrebno simulirati raspodelu slučajne promenljive

$$X: \quad \begin{pmatrix} x_1 & x_2 & \dots & x_k \\ p_1 & p_2 & \cdots & p_k \end{pmatrix}, \quad \sum_{i=1}^{k} p_i = 1.$$

S obzirom da je $p_i \in [0,1]$ za $i = 1, 2, \dots, k$, podelimo segment $[0,1]$ na $k$ podintervala

$$\Delta_1 = [0, p_1), \quad \Delta_2 = [p_1, p_1 + p_2), \dots, \Delta_k = [p_1 + p_2 + \dots + p_{k-1}, 1].$$

Iza toga se vrši izbor $n$ slučajnih brojeva izmedju 0 i 1 sa željenim brojem značajnih cifara iz tablice slučajnih brojeva ili na neki drugi način. Time je definisana nova slučajna promenljiva $Z$ kao slučajno izabrani broj. Recimo da je takvim izborom dobijen niz brojeva $z_1, z_2, \dots, z_n$. Nadalje, za svaki od dobijenih brojeva $z$ utvrdjujemo kom intervalu $\Delta_j$, $j = 1, 2, \dots, k$, pripada. Neka je $z \in \Delta_l$, $(l = 1, 2, \dots, k)$. Tada prihvatimo da je realizovan dogadjaj $\{X = x_l\}$ i tako redom. Dakle,

$$P\{X = x_l\} = P\{Z \in \Delta_l\} = d(\Delta_l) = p_l.$$

**Primer 18.** Neka slučajna promenljiva $X$ ima raspodelu sledećeg oblika:

$$\begin{pmatrix} -1 & 0 & 1 \\ 0.1 & 0.3 & 0.6 \end{pmatrix}$$

Modelirati deset vrednosti ove slučajne promenljive, koristeći tablicu slučajnih brojeva.

Čitamo brojeve iz Tablice 6 slučajnih brojeva iz trećeg reda. Segment $[0,1]$ delimo na podintervale $[0, 0.1), [0.1, 0.4), [0.4, 1]$. Sada uzimamo po jednu cifru iz tablice slučajnih brojeva i množimo sa $10^{-1}$. Tako se od niza cifara

$$4, 5, 9, 3, 9, 6, 0, 1, 7, 3$$

dobijaju brojevi

$$0.4,\ 0.5,\ 0.9,\ 0.3,\ 0.9,\ 0.6,\ 0.0,\ 0.1,\ 0.7,\ 0.3\,.$$

Broj 0.4 pripada intervalu $[0.4, 1]$, te prihvatamo da se realizovala vrednost 1 slučajne promenljive $X$ (jedna modelirana vrednost). Zatim isti princip zaključivanja primenjujemo i na ostale izabrane brojeve, i konačno dobijamo sledeći niz modeliranih vrednosti:

$$1, 1, 1, 0, 1, 1, -1, 0, 1, 0\,.$$

△

## 2.7.2 Modeliranje raspodela apsolutno neprekidnog tipa

1. **Modeliranje uniformne raspodele $\mathcal{U}[a, b]$**

   Kako sama tablica slučajnih brojeva odslikava uniformnu raspodelu, označimo sa $\eta$ izbor $k$–tocifrenih prirodnih brojeva iz Tablice 6. Dakle, pročitane grupe od po $k$ cifara smatramo $k$–tocifrenim brojevima $\eta$ i zatim izvršimo množenja sa $10^{-k}$ sa

ciljem da modeliramo vrednosti slučajne promenljive $\xi : \mathcal{U}[0,1]$, $\xi = \eta \cdot 10^{-k}$. Veza izmedju slučajnih promenljivih $X : \mathcal{U}[a,b]$ i $\xi : \mathcal{U}[0,1]$ je

$$X = a + (b-a)\xi.$$

Poslednjom transformacijom se upravo izvrši modelovanje slučajne promenljive $X$, odnosno njenih realizovanih vrednosti.

2. **Opšti slučaj**

Koristićemo slučajni izbor broja iz intervala $[0,1]$, tj. slučajnu promenljivu $\xi : \mathcal{U}[0,1]$. Posredstvom te slučajne promenljive modeliraćemo vrednosti svake druge slučajne promenljive apsolutno neprekidnog tipa.

**Teorema 2.7.1** *Neka je slučajna promenljiva $X$ apsolutno neprekidnog tipa sa funkcijom raspodele $F$. Rešenje slučajne jednačine*

$$F(X) = \xi \tag{2.2}$$

*po nepoznatoj $X$, pri čemu je $\xi : \mathcal{U}[0,1]$, je slučajna promenljiva čija je funkcija raspodele baš $F$.*

**Dokaz.** Kako je slučajna promenljiva $X$ po pretpostavci apsolutno neprekidnog tipa, onda postoji interval $(a,b) \subset R$ na kome je funkcija $F$ monotono rastuća za $x \in (a,b)$ pri čemu $a$ može biti i $-\infty$, a $b$ može biti $+\infty$. Neka je najpre $F$ monotono rastuća za svako $x \in R$. Sledi da za svako $y \in (0,1)$ postoji tačno jedan $x \in R$ takav da je $F(x) = y$, tj. na intervalu $(a,b)$ funkcija $F$ je bijektivna funkcija, te ima inverznu. Dakle, jednačina (2.2) ima jedinstveno rešenje na tom intervalu. Otuda

$$P\{X \le x\} = P\{F^{-1}(\xi) \le x\} = P\{\xi \le F(x)\} = F_\xi(F(x)) = F(x)$$

$$x \in R\,.$$

Ili postoji interval $[a,b) \subset R$ tako da važi

$$F(x) = \begin{cases} 0, & x < a \\ g(x), & a \le x < b \\ 1, & x \ge b \end{cases}$$

pri čemu postoji $g^{-1}$ na $[a,b)$ pa je za

$$\begin{array}{ll} x < a & P\{X \le x\} = F(x) = 0 \\ a \le x < b & P\{X \le x\} = P\{g^{-1}(\xi) \le x\} = P\{\xi \le g(x)\} = g(x) \\ x \ge b & P\{X \le x\} = F(x) = 1 \end{array}$$

□

**Primer 19 (Eksponencijalna raspodela).** Neka slučajna promenljiva $X$ ima gustinu raspodele

$$f(x) = \begin{cases} 0, & x < 0 \\ \lambda e^{-\lambda x}, & x \geq 0 \end{cases} \quad , \quad x \in R \quad , \quad \lambda > 0,$$

odnosno funkciju raspodele

$$F(x) = \begin{cases} 0, & x < 0 \\ 1 - e^{-\lambda x}, & x \geq 0 \end{cases} \quad , \quad x \in R \quad , \quad \lambda > 0.$$

Tada se rešavanjem jednačine (2.2) po $X$ za $x \geq 0$ dobija

$$X = -\frac{1}{\lambda} \ln(1 - \xi) \, . \triangle$$

**Primer 20 (Normalna raspodela).** Neka slučajna promenljiva $X : \mathcal{N}(m \, , \, \sigma^2)$. Njene vrednosti ćemo simulirati pomoću standardizovane slučajne promenljive

$$X^* = \frac{X - m}{\sigma} \, : \mathcal{N}(0 \, , \, 1)$$

Posmatraćemo dva slučaja:

- (i)$x^* < 0$
  Rešavamo jednačinu (2.2) za

  $$F(x^*) = 0,5 - \Phi(-x^*) \, .$$

  Drugim rečima iz tablice za normalnu normiranu raspodelu čitamo vrednost $-x^*$ za koju važi

  $$\Phi(-x^*) = 0,5 - \xi \, ,$$

  gde je $\xi$ realizovana vrednost slučajne promenljive $\xi$ i dobijamo vrednost za $x$ kao $x = \sigma x^* + m$.
- (ii)$x^* \geq 0$
  Rešavamo jednačinu (2.2) za

  $$F(x^*) = 0,5 + \Phi(x^*) \, .$$

  Drugim rečima iz tablice za normalnu normiranu raspodelu čitamo vrednost $x^*$ za koju važi

  $$\Phi(x^*) = \xi - 0,5$$

  i zatim dolazimo do rešenja kao u slučaju $(i)$

Do brojeva $\xi$ dolazimo iz tablice slučajnih brojeva, ili nekim generatorom slučajnih brojeva. Ukoliko dobijemo $\xi \in [0 \, , \, 1/2)$ primenićemo rešenje $(i)$, a ukoliko dobijemo $\xi \in [1/2 \, , \, 1]$ primenićemo $(ii)$.$\triangle$

# Glava 3

# Ocenjivanje parametara

Jedan od najčešće rešavanih problema u matematičkoj statistici je ocenjivanje nepoznatog parametra, ili više njih, raspodele obeležja na osnovu uzorka. Pri tome se primenjuju dva tipa ocena: tačkaste i intervalne ocene.

Neka je $X$ obeležje koje se posmatra na populaciji. Problem izbora raspodele obeležja $X$ iz familije dopustivih raspodela $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ se može definisati i kao izbor iz familije funkcija raspodele $\{F(x;\theta), \theta \in \Theta\}$ ili gustina raspodele $\{f(x;\theta), \theta \in \Theta\}$. Dakle, svakim od pomenuta tri skupa se na odredjeni način definiše familija dopustivih raspodela za obeležje $X$.

## 3.1 Tačkasto ocenjivanje

Već smo istakli da je osnovni zadatak matematičke statistike da na osnovu eksperimenta odredi raspodelu posmatranog obeležja na populaciji. Dakle, ako na osnovu nekih prethodnih istraživanja ili na neki drugi način dodjemo do familije dopustivih raspodela za posmatrano obeležje $X$, problem odredjivanja konkretne raspodele se svodi na odredjivanje tačne vrednosti parametra $\theta$. Medjutim, statističkim postupcima samo možemo ocenjivati nepoznati parametar na osnovu uzorka i to sa odredjenom tačnošću.

Tačkasto ocenjivanje je jedan od načina za ocenjivanje prave vrednosti nepoznatog parametra i njime ćemo se najpre baviti.

Metod tačkastog ocenjivanja sastoji se u sledećem:

Treba definisati statistiku $Y = u(X_1, X_2, \ldots, X_n)$ tako da za realizovani uzorak $(x_1, x_2, \ldots, x_n)$, broj $y = u(x_1, x_2, \ldots, x_n)$ bude "dobra" ocena za $\theta$.

Obično, ali ne uvek, skup vrednosti ocene $Y$ se poklapa sa $\Theta$. Valjanost ocene utvrdjuje se na osnovu odredjenih kriterijuma o kojima će nadalje biti reči.

Definicija 14. Statistika $Y = u(X_1, X_2, \ldots, X_n)$ na osnovu uzorka $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ iz populacije sa obeležjem $X$, čija raspodela pripada familiji dopustivih raspodela $\{f(x;\theta), \theta \in \Theta\}$, je *nepristrasna* ili *centrirana ocena parametra* $\theta$, ako je njeno matematičko očekivanje jednako vrednosti parametra $\theta$, tj.

$$E(Y) = \theta.$$

**Primer 21.** ( Ocena za matematičko očekivanje)
Neka je dat slučajni uzorak

$$\mathbf{X} = (X_1, X_2, \ldots, X_n)$$

iz populacije sa obeležjem $X$ čija raspodela pripada familiji dopustivih raspodela

$$\{f(x;\theta), \theta \in \Theta\} \quad \text{za koju je} \quad E(X) = \theta\,.$$

Primer takve familije je

$$\left\{\mathcal{N}(m, \sigma^2), m \in R, \sigma^2 \in R^+\right\}.$$

Posmatrajmo statistiku

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Dokažimo da je statistika $\overline{X}_n$ nepristrasna ocena parametra $\theta$.
Na osnovu osobina matematičkog očekivanja sledi

$$E(\overline{X}_n) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}E\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} EX_i = \frac{1}{n}\sum_{i=1}^{n} EX = \frac{1}{n}n\theta = \theta.$$

Primetimo da je sredina uzorka nepristrasna ocena matematičkog očekivanja proizvoljne raspodele koja ima matematičko očekivanje. $\triangle$

Pristrasne ocene se opisuju pomoću pomeraja $b(\theta) = E(Y) - \theta$. Očigledno, nepristrasne ocene su one za koje je $b(\theta) = 0$.

**Primer 22.** (Ocena za disperziju)
Pod uslovom da obeležje $X$ ima disperziju i da je $\theta = D(X)$ ispitajmo da li je uzoračka disperzija $\overline{S}_n^2$, dobijena na osnovu prostog slučajnog uzorka, nepristrasna ocena za $\theta$.
Kako je uzorak prost,
$E(\overline{S}_n^2) = \frac{1}{n}\sum_{i=1}^{n} E(X_i - \overline{X}_n)^2 = \frac{1}{n}\sum_{i=1}^{n} E(X_i^2 - 2X_i\overline{X}_n + \overline{X}_n^2) =$
$= \frac{1}{n}\sum_{i=1}^{n}(E(X_i^2) - 2\frac{1}{n}\sum_{j=1}^{n} E(X_iX_j) + E(\overline{X}_n^2)) =$
$= \frac{1}{n}\sum_{i=1}^{n}(E(X^2) - 2\frac{n-1}{n}(EX)^2 - 2\frac{1}{n}E(X^2) + E(\frac{1}{n^2}\sum_{i,j=1}^{n} X_iX_j)) =$
$= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{n-2}{n}E(X^2) - 2\frac{n-1}{n}(EX)^2 + \frac{1}{n^2}nE(X^2) + \frac{n^2-n}{n^2}(EX)^2\right) =$
$= \frac{1}{n}n\frac{n-1}{n}(EX^2 - (EX)^2) = \frac{n-1}{n}D(X)$ . $\triangle$

Definicija 15. Statistika $Y = u(X_1, X_2, \ldots, X_n)$ na osnovu uzorka $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ iz populacije sa obeležjem $X$, čija raspodela pripada familiji dopustivih raspodela $\{f(x;\theta), \theta \in \Theta\}$, je *asimptotski nepristrasna* ili *asimptotski centrirana ocena* parametra $\theta$, ako

$$E(Y) \to \theta \quad , \quad n \to \infty \quad .$$

**Primer 23.** Kako je

$$E(\overline{S}_n^2) = \frac{n-1}{n} D(X)\,,$$

zaključujemo da je disperzija uzorka asimptotski nepristrasna ocena disperzije obeležja (ukoliko je uzorak prost).

Asimptotski nepristrasna ocena se može popraviti do nepristrasne ocene. Tako, kada je reč o disperziji uzorka, na osnovu nje se može definisati popravljena disperzija uzorka:

$$\widetilde{S}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{n}{n-1}\overline{S}_n^2.$$

U uzorcima većeg obima **pristrasnost** ocene $\overline{S}_n^2$ je zanemarljiva (na primer, ako se radi sa tačnošću $10^{-2}$ pomenuta pristrasnost se ne oseća za uzorak čiji je obim veći od 200). Medjutim, kada je obim uzorka mali, kao ocena za disperziju obeležja $X$ uzima se **popravljena disperzija** uzorka.△

Jedna od mera bliskosti ocene i prave vrednosti parametra je srednjekvadratno odstupanje statistike (ocene) $Y$ od prave vrednosti parametra:

$$E(Y-\theta)^2 = E(Y - E(Y) + E(Y) - \theta)^2 = D(Y) + b^2(\theta)$$

Dakle, u slučaju da je ocena $Y$ nepristrasna, srednjekvadratno odstupanje te ocene od prave vrednosti parametra je upravo disperzija statistike $Y$.

Definicija 16. Statistika $Y = u(X_1, X_2, \ldots, X_n)$ je *najbolja ocena parametra* $\theta$ ako je nepristrasna i ako je disperzija $D(Y)$ manja ili jednaka od disperzije bilo koje druge nepristrasne ocene za $\theta$.

I pored pogodnosti koje pruža klasa nepristrasnih statistika, ovakvu klasu ne treba idealizovati. Nekada je zahtev da je ocena nepristrasna prestrog, a nekada nepristrasna ocena i ne postoji.

Nije uvek neophodno oceniti sam parametar $\theta$ raspodele $f(x;\theta)$, već treba oceniti neku funkciju od parametra $\tau(\theta)$, $\theta \in \Theta$. Tada se od ocene $T = T(X_1, X_2, \ldots, X_n)$ za funkciju $\tau$ traži da bude nepristrasna, tj. da je $E(T) = \tau(\theta)$, što nije trivijalna posledica postojanja nepristrasne ocene za $\theta$.

**Primer 24.** Neka obeležje $X$ ima Puasonovu raspodelu $\mathcal{P}(\theta)$. Oceniti funkciju $\tau(\theta) = 1/\theta$ na osnovu uzorka obima $n = 1$. Lako je uočiti da je sredina uzorka nepristrasna ocena za $\theta$. Medjutim, ako tražimo nepristrasnu ocenu funkcije $\tau$, $T = T(X_1)$, ona mora da zadovolji uslov $E(T) = 1/\theta$. Dakle,

$$E(T) = \sum_{x=0}^{\infty} T(x) e^{-\theta}\frac{\theta^x}{x!} = \frac{1}{\theta} \quad ,$$

odnosno, važila bi jednakost

$$\sum_{x=0}^{\infty} T(x)\frac{\theta^{x+1}}{x!} = e^{\theta}$$

Primenjujući Tejlorov razvoj funkcije $e^\theta$, dobija se

$$\theta \sum_{x=0}^{\infty} T(x)\frac{\theta^x}{x!} = \sum_{t=0}^{\infty} \frac{\theta^t}{t!}$$

što znači da funkcija $T$ zavisi od $\theta$. Medjutim, ovo protivureči zahtevu da $T$ bude statistika. Dakle, u ovom slučaju ne postoji nepristrasan ocena funkcije $\tau$ na osnovu zadatog uzorka. $\triangle$

Druga moguća mera odstupanja ocene od prave vrednosti parametra je *apsolutno odstupanje*, $|Y-\theta|$, koje predstavlja jednu slučajnu promenljivu, te se kao mera valjanosti ocene može uzeti verovatnoća postizanja gornje granice apsolutne greške,

$$P\{|Y-\theta| < \varepsilon\} = p.$$

Tada se za unapred zadato dovoljno malo $\varepsilon > 0$, odredjuje $p$ ili obrnuto, na osnovu poznavanja broja $p$ odredjuje se $\varepsilon$. Dakle, ako je poznata raspodela verovatnoća za $Y$, problem je, sa teorijske tačke gledišta, rešiv. Medjutim, i kada nam ova raspodela nije poznata, moguće je pod odredjenim uslovima odrediti $\varepsilon$ za zadato $p$. Ako je, recimo, $\varepsilon$ oblika

$$\varepsilon = k\sqrt{D(Y)}, \quad \text{za neko} \quad k \geq 1$$

i $Y$ nepristrasnu ocenu parametra $\theta$, prema Čebiševljevoj nejednakosti je

$$1 - P\left\{|Y-\theta| < k\sqrt{D(Y)}\right\} = P\left\{|Y-\theta| \geq k\sqrt{D(Y)}\right\} \leq \frac{D(Y)}{k^2 D(Y)} = \frac{1}{k^2} \quad ,$$

tj.

$$p \geq 1 - \frac{1}{k^2} \quad .$$

Tako, za $k = 2$ je $p \geq 0,75$. Slično se može odrediti $k$ za zadato $p$. Ovaj se rezultat može koristiti i za odredjivanje obima uzorka ukoliko se zadaju vrednosti za $\varepsilon$ i $p$, o čemu će nadalje biti reči.

U vezi sa apsolutnim odstupanjem može se iskazati sledeći kriterijum valjanosti tačkastih ocena.

Definicija 17. Statistika $Y = u(X_1, X_2, \ldots, X_n)$ je *postojana ocena* za $\theta$ ako ona konvergira u verovatnoći ka $\theta$, tj.

$$Y \xrightarrow{P} \theta, \quad n \to \infty.$$

Ako je ova konvergencija skoro izvesno (skoro sigurno),

$$Y \xrightarrow{s.i.} \theta, \quad n \to \infty,$$

onda je ocena *strogo postojana*.

## 3.2 Odredjivanje obima uzorka

Povećanje obima uzorka je postupak na koji se u statistici po pravilu računa. Medjutim, praktično, kadgod obim uzorka nije slučajan, statistički postupak se sprovodi nad tačno odredjenim, konačnim obimom uzorka $n$. Otuda da bi primena pojedinog statističkog postupka dala zadovoljavajuće rezultate, neophodno je prilikom planiranja eksperimenta predvideti i obim uzorka, $n$.

Planiranje eksperimenta je po svojoj suštini postupak kojim se planira dobijanje odredjene količine informacija. "Proizvod" koji se eksperimentom stvara jeste informacija. Cilj je dobiti što veću količinu informacija praveći pri tome što manje troškove, dakle dobiti što kvalitetniji proizvod po što nižoj ceni. Plan eksperimenta direktno utiče na količinu dobijenih informacija pri svakom merenju. U tom svetlu biće govora o izboru obima uzorka $n$. Jedan od prvih koraka koje istraživač mora da načini u svom istraživanju je odredjivanje obima uzorka. Obim uzorka direktno utiče na preciznost ocene, odnosno na valjanost zaključaka dobijenih primrnom statističkih metoda. Preciznost ocene se može iskazati veličinom greške dobijene ocene. Pri tome primenjeni metod za procenu valjanosti ocene, odnosno veličinu greške, jeste u vezi sa obimom uzorka.

Postupak za odredjivanje broja $n$ zavisi od parametra koji se ocenjuje, statistike kojom se ocenjuje i od toga da li su drugi relevantni parametri obeležja poznati ili se takodje procenjuju na osnovu uzorka (ukoliko je reč o ocenjivanju parametara). Imajući sve to u vidu, definiše se maksimalna veličina greške sa kojom treba raditi u ocenjivanju.

Pomenućemo samo neke kriterijume koji se koriste za odredjivanje obima uzorka u zavisnosti od zadate tačnosti ocenjivanja:

- srednjekvadratno odstupanje, $E(\widehat{\theta}-\theta)^2$, i standardno odstupanje, $\sqrt{E(\widehat{\theta}-\theta)^2}$
- apsolutna greška, manja od $\varepsilon$, ocene, verovatnoće (poverenja) $1-\alpha$, tj. $P\{|\widehat{\theta}-\theta|<\varepsilon\}=1-\alpha$
- koeficijent varijacije ocene, $C_v(\widehat{\theta})=\frac{\sqrt{E(\widehat{\theta}-\theta)^2}}{E(\widehat{\theta})}$
- relativna greška ocene, $\delta=\frac{|\widehat{\theta}-\theta|}{\theta}$

Obim uzorka se planira tako da se postigne zadata tačnost ocene.

**Primer 25.** Iz uzorka $(X_1, X_2, \ldots, X_n)$ treba postići tačnost ocene nepoznatog parametra $\theta$ sa apsolutnom greškom ne većom od $\varepsilon$ poverenja $1-\alpha$.

Dakle,

$$P\{|\theta-\widehat{\theta}|<\varepsilon\}=1-\alpha$$
$$P\{-\varepsilon<\theta-\widehat{\theta}<\varepsilon\}=1-\alpha$$
$$P\{\widehat{\theta}-\varepsilon<\theta<\widehat{\theta}+\varepsilon\}=1-\alpha\,.$$

Kako je

$$\widehat{\theta}=\widehat{\theta}(X_1,\ldots,X_n)=\widehat{\theta}(n)\,,$$

to će se izborom $\varepsilon$ odrediti najmanji $n$ za koji se postiže nivo poverenja $1-\alpha$.△

Očigledno, da bismo odredili n, treba da znamo raspodelu statistike $\widehat{\theta}$. Ovaj problem se rešava konkretno, mada najčešće primenom centralne granične teoreme. Pri tome može da nastupi novi problem zbog nepoznavanja matematičkog očekivanja ili disperzije obeležja. Tada se ovi parametri moraju oceniti na osnovu prethodnog uzorka manjeg obima, ili na osnovu nepotpune informacije o ovim karakteristikama obeležja, kao što je, recimo informacija $D(X) \leq \sigma_0^2$.

**Primer 26.** Neka se ocenjuje matematičko očekivanje, $m$, obeležja X beskonačne populacije. Nepristrasna ocena ovog parametra je $\overline{X}_n$(sredina uzorka). Dakle,

$$P\{|m - \overline{X}_n| < \varepsilon\} = 1 - \alpha \quad , \quad m = EX = E\overline{X}_n$$

$$P\left\{\frac{|m - \overline{X}_n|}{\frac{\sigma}{\sqrt{n}}} < \frac{\varepsilon\sqrt{n}}{\sigma}\right\} = 1 - \alpha \quad , \quad \sigma = \sqrt{D(X)}\,.$$

Prema centralnoj graničnoj teoremi slučajna promenljiva $\frac{\overline{X}_n - m}{\sigma}\sqrt{n}$ ima $\mathcal{N}(0,1)$ raspodelu kada $n \to \infty$, pa se u gornjoj jednakosti verovatnoća može odrediti iz

$$2\Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right) = 1 - \alpha\,,$$

odakle sledi da je

$$\frac{\varepsilon\sqrt{n}}{\sigma} = z_{\frac{1-\alpha}{2}}\,, \text{odnosno}\,, n \geq \left(\frac{\sigma z_{\frac{1-\alpha}{2}}}{\varepsilon}\right)^2$$

garantuje traženu tačnost.

Za zadato $\alpha$ vrednost $z_{\frac{1-\alpha}{2}}$ se može pročitati iz tablice, te se i odredjuje na prikazani način. Ukoliko je $\sigma$ nepoznato, ali se zna da je $\sigma \leq \sigma_0$, ova se informacija može upotrebiti tako što će se na osnovu

$$n \geq \left(\frac{\sigma z_{\frac{1-\alpha}{2}}}{\varepsilon}\right)^2$$

izabrati prirodan broj $n_0$ za 1 veći od najvećeg celog dela izraza na desnoj strani, tj.

$$n_0 = \left[\left(\frac{\sigma_0 z_{\frac{1-\alpha}{2}}}{\varepsilon}\right)^2\right] + 1\,. \tag{3.1}$$

Isti način zaključivanja bi se primenio i kod odredjivanja obima uzorka sa vraćanjem iz konačne populacije. Razume se, kada je $n \ll N$. $\triangle$

**Primer 27.** Ukoliko je u pitanju uzorak bez vraćanja iz konačne populacije biće

$$D(\overline{X}_n) = \frac{D(X)}{n}\left(1 - \frac{n}{N}\right)$$

(videti Glavu 11), pa za ocenjivanje matematičkog očekivanja po istom kriterijumu kao u prethodnom primeru imamo

$$P\left\{\frac{|m - \overline{X}_n|}{\sqrt{\frac{\sigma^2}{n}\left(1 - \frac{n}{N}\right)}} < \frac{\varepsilon}{\sqrt{\frac{\sigma^2}{n}\left(1 - \frac{n}{N}\right)}}\right\} = 1 - \alpha$$

$$\frac{\varepsilon\sqrt{nN}}{\sigma\sqrt{N-n}} = z_{\frac{1-\alpha}{2}}\,,$$

odnosno,

$$n \geq \frac{1}{\left(\frac{\varepsilon}{\sigma z_{\frac{1-\alpha}{2}}}\right)^2 + \frac{1}{N}}\,.$$

Ponovo, ako ne poznajemo $\sigma$ već znamo samo da je $\sigma \leq \sigma_0$, imaćemo da obim uzorka za zahtevanu tačnost ocene treba da zadovolji uslov

$$n \geq \frac{1}{\left(\frac{\varepsilon}{\sigma_0 z_{\frac{1-\alpha}{2}}}\right)^2 + \frac{1}{N}}\,.$$

Označimo sa $n_1$ najmanji prirodni broj koji zadovoljava poslednju nejednakost. Dakle,

$$n_1 = \left[\frac{1}{\left(\frac{\varepsilon}{\sigma_0 z_{\frac{1-\alpha}{2}}}\right)^2 + \frac{1}{N}}\right] + 1\,. \tag{3.2}$$

Ukoliko uporedimo obim uzorka $n_1$ sa obimom $n_0$ iz prethodnog primera možemo konstatovati da se kod uzorka bez vraćanja zahteva manji obim uzorka za postizanje iste tačnosti kod tačkastog ocenjivanja očekivane vrednosti obeležja, nego kod uzorka sa vraćanjem. Zaista, iz (3.1)

$$n_0 - 1 + t = \left(\frac{\sigma_0 z_{\frac{1-\alpha}{2}}}{\varepsilon}\right)^2\,,\ \text{gde je}\,, 0 \leq t < 1\,,$$

tj.

$$\left(\frac{\varepsilon}{\sigma_0 z_{\frac{1-\alpha}{2}}}\right)^2 = \frac{1}{n_0 - 1 + t}\,,$$

pa se zamenom u (3.2) dobija relacija

$$n_1 = \left[\frac{N(n_0 - 1 + t)}{N + n_0 - 1 + t}\right] + 1\,.$$

Kako je

$$\frac{1}{n_0 - 1 + t} > \frac{1}{n_0}$$

to je

$$\frac{1}{n_0 - 1 + t} + \frac{1}{N} > \frac{1}{n_0}\,,$$

odnosno,

$$\left[\frac{1}{\frac{1}{n_0 - 1 + t} + \frac{1}{N}}\right] + 1 \leq n_0\,.$$

Drugim rečima,

$$n_1 \leq n_0\,,$$

što je i trebalo dokazati. $\triangle$

## 3.3 Dovoljne statistike

Još jedan od načina da se govori o valjanosti tačkaste ocene za nepoznati parametar raspodele obeležja je kriterijum dovoljnosti.

Neka je data statistika $Y_1 = u_1(X_1, \ldots, X_n)$ za ocenu nepoznatog parametra $\theta \in \Theta \subset R$. Posmatrajmo istovremeno još $(n-1)$-nu statistiku istog uzorka:

$$\begin{aligned} Y_2 &= u_2(X_1, \ldots, X_n) \\ &\vdots \\ Y_n &= u_n(X_1, \ldots, X_n) \end{aligned}$$

i to takvih da je transformacija

$$\begin{aligned} y_1 &= u_1(x_1, \ldots, x_n) \\ y_2 &= u_2(x_1, \ldots, x_n) \\ &\vdots \\ y_n &= u_n(x_1, \ldots, x_n) \end{aligned} \tag{3.3}$$

"$1-1$". Zajednička gustina raspodele vektora statistika $(Y_1, Y_2, \ldots, Y_n)$ je:

$$g(y_1, y_2, \ldots, y_n; \theta) = |J| f(w_1(y_1, \ldots, y_n), \cdots, w_n(y_1, \ldots, y_n); \theta)\,, \tag{3.4}$$

gde je sa $f$ označena zajednička gustina vektora $\mathbf{X}$, za obeležje apsolutno neprekidnog tipa i bez faktora $|J|$, odnosno Jakobijana, za obeležje diskretnog tipa. Sa $\mathbf{w} = (w_i, i = 1, 2, \ldots, n)$ je označena inverzna transformacija za transformaciju $\mathbf{u} = (u_i, i = 1, 2, \ldots, n)$.

Ukoliko transformacija (3.3) nije "$1-1$" u celom $R^n$, onda je desna strana relacije (3.4) suma od, recimo, $k$ izraza tog oblika. Broj sabiraka, $k$, odgovara ukupnom broju oblasti na koje je izvršeno razbijanje prostora $R^n$ i to takvih da je u svakoj od njih transformacija (3.3) "$1-1$".

Ukoliko je uzorak prost, zajednička gustina vektora $\mathbf{X}$ bi bila

$$f(x_1, x_2, \ldots, x_n; \theta) = f(x_1; \theta) f(x_2; \theta) \cdots f(x_n; \theta)$$

što bi proizvelo i adekvatne promene u relaciji (3.4).

Uslovna gustina za $(Y_2, Y_3, \ldots, Y_n)$ pod uslovom $Y_1 = y_1$ data je sa

$$h(y_2, \ldots, y_n | y_1; \theta) = \frac{g(y_1, y_2, \ldots, y_n; \theta)}{g_1(y_1; \theta)}, \quad \text{za} \quad g_1 > 0\,,$$

gde je $g_1(y_1; \theta)$ marginalna gustina za $Y_1$. U opštem slučaju $h(y_2, \ldots, y_n | y_1; \theta)$ zavisi od $\theta$. Medjutim, posebno važni slučajevi su oni u kojima gustina $h$ ne zavisi od parametra $\theta$.

Definicija 18. Neka je $n$ fiksiran prirodni broj i $(X_1, X_2, \ldots, X_n)$ uzorak obima $n$ iz populacije sa obeležjem čija raspodela pripada familiji dopustivih raspodela $\{f(x; \theta), \theta \in \Theta\}$. Statistika $Y_1 = u_1(X_1, \ldots, X_n)$ je *dovoljna statistika* za $\theta$ ako i samo ako za bilo koje druge statistike $Y_2 = u_2(X_1, \ldots, X_n), \ldots, Y_n = u_n(X_1, \ldots, X_n)$, za koje Jakobijan pripadajuće transformacije nije nula, uslovna gustina raspodele $h(y_2, \ldots, y_n | y_1)$ slučajnih promenljivih $Y_2, \ldots, Y_n$ pod uslovom $Y_1 = y_1$, ne zavisi od parametra $\theta$ za bilo koju fiksiranu vrednost $y_1$.

Pri tome se podrazumeva ne samo da analitički izraz uslovne gustine ne zavisi od $\theta$, već i njena oblast definisanosti takodje. Upozoravamo na nedozvoljenu zavisnost sledećim primerom.

**Primer 28.** Funkcija

$$f(x) = \begin{cases} \frac{1}{2}, & \theta - 1 < x < \theta + 1, \\ 0, & \text{van} \end{cases}$$

gde $\theta \in (-\infty, +\infty)$, zavisi od $\theta$. $\triangle$

**Primer 29.** Neka je $(X_1, X_2)$ prost slučajni uzorak obima 2 iz raspodele $\{\mathcal{B}(2,\theta),\ 0 < \theta < 1\}$. Ispitajmo da li je $Y_1 = X_1 + X_2$ dovoljna statistika. U tu svrhu posmatramo

$$\begin{aligned} f(x_1;\theta)f(x_2;\theta) &= \begin{cases} \binom{2}{x_1}\theta^{x_1}(1-\theta)^{2-x_1}\binom{2}{x_2}\theta^{x_2}(1-\theta)^{2-x_2} & , \quad (x_1,x_2) \in \{(0,0),\ldots,(2,2)\} \\ 0 & , \quad \text{inače} \end{cases} \\ &= \begin{cases} \frac{2!2!\theta^{x_1+x_2}(1-\theta)^{4-(x_1+x_2)}}{x_1!(2-x_1)!x_2!(2-x_2)!} & , \quad (x_1,x_2) \in \{(0,0),\ldots,(2,2)\} \\ 0 & , \quad \text{inače} \quad . \end{cases} \end{aligned}$$

Odavde dobijamo da je:

$$\begin{aligned} g_1(y_1) &= \begin{cases} \binom{4}{y_1}\theta^{y_1}(1-\theta)^{4-y_1} & , \quad y_1 \in \{0,1,2,3,4\} \\ 0 & , \quad \text{inače} \end{cases} \\ &= \begin{cases} \frac{4!\theta^{y_1}(1-\theta)^{4-y_1}}{y_1!(4-y_1)!} & , \quad y_1 \in \{0,1,2,3,4\} \\ 0 & , \quad \text{inače} \quad . \end{cases} \end{aligned}$$

Uzmimo drugu statistiku $Y_2 = X_2$ sa raspodelom $\mathcal{B}(2,\theta)$, pa će zajednička gustina statistika $Y_1$ i $Y_2$ biti:

$$g(y_1,y_2;\theta) = \begin{cases} \frac{4\theta^{y_1}(1-\theta)^{4-y_1}}{(y_1-y_2)!(2-y_1+y_2)!(2-y_2)!y_2!} & , \quad (y_1,y_2) \in \{(0,0),\ldots,(4,2)\} \\ 0 & , \quad \text{inače} \quad . \end{cases}$$

Uslovna gustina će biti:

$$\begin{aligned} h(y_2|y_1;\theta) &= \begin{cases} \frac{\frac{4\theta^{y_1}(1-\theta)^{4-y_1}}{(y_1-y_2)!(2-y_1+y_2)!(2-y_2)!y_2!}}{\frac{4!\theta^{y_1}(1-\theta)^{4-y_1}}{y_1!(4-y_1)!}} & , \quad (y_1,y_2) \in \{(0,0),\ldots,(4,2)\} \\ 0 & , \quad \text{inače} \end{cases} \\ &= \begin{cases} \frac{y_1!(4-y_1)!}{3!(y_1-y_2)!(2-y_1+y_2)!(2-y_2)!y_2!} & , \quad (y_1,y_2) \in \{(0,0),\ldots,(4,2)\} \\ 0 & , \quad \text{inače} \quad . \end{cases} \end{aligned}$$

Dakle, uslovna gustina za $Y_2$ pod uslovom $Y_1 = y_1$ ne zavisi od $\theta$, čime smo dokazali da je statistika $Y_1$ dovoljna. $\triangle$

### 3.3.1 Kriterijumi egzistencije dovoljne statistike

Provera dovoljnosti neke statisrtike direktno po definiciji se retko koristi zbog složenog eksplicitnog odredjivanja gustina raspodele koje u definiciji učestvuju. Jednostavniji način za proveru dovoljnosti daju naredne teoreme.

**Teorema 3.3.1** *(Fišer-Nejmanov kriterijum)*
*Neka je $(X_1, X_2, \ldots, X_n)$ slučajni uzorak iz populacije sa obeležjem $X$ sa gustinom raspodele koja pripada familiji dopustivih raspodela $\{f(x;\theta),\ \theta \in \Theta\}$. Neka je $Y_1 = u_1(X_1, \ldots, X_n)$ statistika čija je gustina raspodele $g_1(y_1;\theta)$. Tada je $Y_1$ dovoljna statistika za $\theta$ ako i samo ako*

$$f(x_1, x_2, \ldots x_n;\theta) = g_1(u_1(x_1, x_2, \ldots, x_n);\theta)H(x_1, x_2, \ldots, x_n) \tag{3.5}$$

*gde za svaku vrednost funkcije $u_1$, funkcija $H$ ne zavisi od $\theta$.*

**Dokaz.** Dokaz ćemo izvesti samo za obeležje apsolutno neprekidnog tipa.

Najpre uvedimo sledeću bijektivnu transformaciju $\mathbf{u} : R^n \to R^n$ i njoj inverznu $\mathbf{w} : R^n \to R^n$, $\mathbf{u} = (u_1, \ldots, u_n)$, $\mathbf{w} = (w_1, \ldots, w_n)$:

$$\begin{array}{rcll} y_1 &=& u_1(x_1, \ldots, x_n) & x_1 = w_1(y_1, \ldots, y_n) \\ y_2 &=& u_2(x_1, \ldots, x_n) & x_2 = w_2(y_1, \ldots, y_n) \\ & \vdots & & \\ y_n &=& u_n(x_1, \ldots, x_n) & x_n = w_n(y_1, \ldots, y_n) \end{array} \tag{3.6}$$

takve da je Jakobijan

$$J = \left| \frac{\partial x_i}{\partial y_j} \right|_{i,j=1,\ldots,n} \neq 0.$$

Pretpostavimo da važi uslov (3.5) i dokažimo da je $Y_1$ dovoljna statistika.

Odgovarajuća zajednička gustina raspodele statistika $Y_1, Y_2, \ldots, Y_n$ biće:

$$\begin{aligned} g(y_1, y_2, \ldots, y_n;\theta) &= f(w_1(y_1, \ldots, y_n), \ldots, w_n(y_1, \ldots, y_n);\theta)|J| \\ &= g_1(y_1;\theta)H(w_1(y_1, \ldots, y_n), \ldots, w_n(y_1, \ldots, y_n))|J|\,. \end{aligned}$$

S druge strane, kako je uslovna gustina data sa:

$$\begin{aligned} h(y_2, \ldots, y_n|y_1) &= \frac{g(y_1, \ldots, y_n;\theta)}{g_1(y_1;\theta)} = \\ &= \frac{g_1(y_1;\theta)H(w_1(y_1, \ldots, y_n), \ldots, w_n(y_1, \ldots, y_n))|J|}{g_1(y_1;\theta)} = \\ &= H(w_1(y_1, \ldots, y_n), \ldots, w_n(y_1, \ldots, y_n))|J| \end{aligned}$$

i očigledno ne zavisi od parametra $\theta$, sledi da je $Y_1$ dovoljna statistika.

Dokažimo i drugi smer iskaza teoreme.

Pretpostavimo sada da je $Y_1$ dovoljna statistika i dokažimo da važi uslov (3.5). U tom slučaju uslovna gustina ne zavisi od parametra $\theta$:

$$h(y_2, \ldots, y_n|y_1) = \frac{g(y_1, \ldots, y_n; \theta)}{g_1(y_1; \theta)},$$

tj.

$$g(y_1, \ldots, y_n; \theta) = g_1(y_1; \theta) h(y_2, \ldots, y_n|y_1).$$

Koristeći inverznu transformaciju $\mathbf{w}$ čiji je Jakobijan $J^* = \left|\frac{\partial y_i}{\partial x_j}\right|_{i,j=1,\ldots,n} \neq 0$, dobijamo da je zajednička gustina vektora $\mathbf{X}$

$$f(x_1, \ldots, x_n; \theta) = g(u_1(x_1, \ldots, x_n), \ldots, u_n(x_1, \ldots, x_n); \theta)|J^*| =$$
$$= g_1(u_1(x_1, \ldots, x_n; \theta)) h(u_2(x_1, \ldots, x_n), \ldots, u_n(x_1, \ldots, x_n)|u_1(x_1, \ldots, x_n))|J^*|.$$

Očigledno za $H(x_1, \ldots, x_n)$ se može uzeti $h(u_2(x_1, \ldots, x_n), \ldots, u_n(x_1, \ldots, x_n)|u_1(x_1, \ldots, x_n))|J^*|$, te je i drugi smer iskaza teoreme dokazan.

Za diskretan slučaj dokaz se razlikuje samo u tome što izostaje Jakobijan. □

Dakle, Fišer-Nejmanov kriterijum nam pokazuje da činjenica da je neka statistika dovoljna, nije uslovljena izborom preostalih $n-1$ statistika (pod uslovom da je Jakobijan korišćene transformacije različit od nule).

**Primer 30.** Neka su $Y_1, \ldots, Y_n$ statistike poretka na osnovu prostog slučajnog uzorka $\mathbf{X} = (X_1, \ldots, X_n)$, $Y_i = X_{(i)}, i = 1, \ldots, n$ iz raspodele čija je gustina

$$f(x; \theta) = \begin{cases} e^{-(x-\theta)} & , \quad \theta < x < \infty \quad , -\infty < \theta < \infty \\ 0 & , \quad \text{inače}. \end{cases}$$

Gustina raspodele statistike $Y_1$ je

$$g_1(y_1; \theta) = \begin{cases} ne^{-n(y_1-\theta)} & , \quad \theta < y_1 < \infty \\ 0 & , \quad \text{inače}. \end{cases}$$

Zajednička gustina raspodele vektora $\mathbf{X}$ je

$$e^{-(x_1-\theta)} e^{-(x_2-\theta)} \cdots e^{-(x_n-\theta)} = g_1(\min_i x_i; \theta) \left\{ \frac{exp(-x_1 - \cdots - x_n)}{n \cdot exp(-n \cdot \min_i x_i)} \right\}.$$

Dakle, koristeći se Fišer–Nejmanovim kriterijumom, zaključujemo da je statistika poretka reda jedan dovoljna statistika za ocenu parametra $\theta$ posmatrane raspodele. △

Fišer-Nejmanov kriterijum zahteva poznavanje gustine raspodele za $Y_1$, što može da oteža ili čak onemogući njegovu praktičnu primenu. Naredni kriterijum izbegava neophodnost poznavanja ove gustine.

**Teorema 3.3.2** *(Teorema faktorizacije)*

*Neka je $(X_1, X_2, \ldots, X_n)$ uzorak iz populacije sa obeležjem čija gustina raspodele pripada familiji dopustivih raspodela $\{f(x; \theta), \theta \in \Theta\}$. Statistika $Y_1 = u_1(X_1, \ldots, X_n)$ je dovoljna statistika za $\theta$ ako i samo ako postoje dve nenegativne funkcije $k$ i $K$, takve da je*

$$f(x_1, x_2, \cdots, x_n; \theta) = k(u_1(x_1, x_2, \ldots, x_n); \theta) K(x_1, x_2, \ldots, x_n), \tag{3.7}$$

*gde za svaku konkretnu vrednost $u_1(x_1, \ldots, x_n)$ funkcija $K(x_1, \ldots, x_n)$ ne zavisi od $\theta$.*

**Dokaz.** Dokaz izvodimo samo za apsolutno neprekidni slučaj, dok je za diskretni analogan.

Pretpostavimo da važi faktorizacija (3.7). Definišimo istu transformaciju kao u dokazu Fišer-Nejmanovog kriterijuma. Zajednička gustina raspodele statistika $Y_1, \ldots, Y_n$ je

$$\begin{aligned} g(y_1, \ldots, y_n; \theta) &= f(w_1(y_1, \ldots, y_n), \ldots, w_n(y_1, \ldots, y_n); \theta)|J| \\ &= k(y_1; \theta)K(w_1(y_1, \ldots, y_n), \ldots, w_n(y_1, \ldots, y_n))|J| \,. \end{aligned}$$

Sada treba dokazati da uslovna gustina ne zavisi od parametra $\theta$ ili da važi teorema Fišer-Nejmana. U tu svrhu odredimo marginalnu gustinu za $Y_1$.

$$\begin{aligned} g_1(y_1; \theta) &= \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} g(y_1, y_2, \ldots, y_n; \theta) dy_2 \cdots dy_n = \\ &= \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} k(y_1; \theta)K(w_1(y_1, \ldots, y_n), \ldots, w_n(y_1, \ldots, y_n))|J| dy_2 \ldots dy_n = \\ &= k(y_1; \theta)m(y_1) \end{aligned}$$

gde je

$$m(y_1) = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} K(w_1(y_1, \ldots, y_n), \ldots, w_n(y_1, \ldots, y_n))|J| dy_2 \ldots dy_n$$

Za $m(y_1) = 0$ biće $g_1(y_1; \theta) = 0$. Medjutim, za $m(y_1) \neq 0$ ,

$$f(x_1, \ldots, x_n; \theta) = g_1(u_1(x_1, \ldots, x_n); \theta) \frac{K(x_1, \ldots, x_n)}{m(u_1(x_1, \ldots, x_n))} \,.$$

Prema Fišer-Nejmanovom kriterijumu, statistika $Y_1$ je dovoljna statistika za parametar $\theta$.

Obrnuto, pretpostavimo da je $Y_1$ dovoljna statistika. Tada se za $k$ iz faktorizacije (3.7) može uzeti funkcija $g_1(u_1(x_1, \ldots, x_n); \theta)$, a za $K$ funkcija $h(u_2(x_1, \ldots, x_n), \ldots,$ $u_n(x_1, \ldots, x_n)|u_1(x_1, \ldots, x_n))$ iz definicije dovoljne statistike, čime se dokazuje da faktorizacija važi. □

**Teorema 3.3.3** *Ako je $Y_1 = u_1(X_1, \ldots, X_n)$ dovoljna statistika za parametar $\theta$ raspodele obeležja $X$ iz koje je uzet uzorak, tada proizvoljna bijektivna funkcija $Z = u(Y_1)$ koja ne zavisi od $\theta$, tj. $Z = u(u_1(X_1, \ldots, X_n)) = v(X_1, \ldots, X_n)$, je takodje dovoljna statistika za $\theta$.*

**Dokaz.** Prema teoremi faktorizacije je

$$f(x_1, \ldots, x_n; \theta) = k(u_1(x_1, \ldots, x_n); \theta)K(x_1, \ldots, x_n).$$

S obzirom na bijekciju $z = u(y_1)$, sledi da postoji inverzna funkcija $w$ takva da je $y_1 = w(z)$. S druge strane je $y_1 = u_1(x_1, \ldots, x_n)$, te je

$$u_1(x_1, \ldots, x_n) = w(z) = w(u(u_1(x_1, \ldots, x_n))) = w(v(x_1, \ldots, x_n))$$

koja ne zavisi od $\theta$, te je

$$\begin{aligned} f(x_1, \ldots, x_n; \theta) = k(w(v(x_1, \ldots, x_n)); \theta)K(x_1, \ldots, x_n) = k(w(z); \theta)K(x_1, \ldots, x_n) = \\ = k_1(z; \theta)K(x_1, \ldots, x_n) \,, \text{ gde je } k_1 = k \circ w. \end{aligned}$$

Kako je $k_1$ funkcija (od uzorka) samo od $z$ i od $\theta$, a $K$ ne zavisi od $\theta$, to je prema teoremi faktorizacije, $Z$ takodje dovoljna statistika za parametar $\theta$.□

### 3.3.2 Najbolja ocena za parametar

Dovoljne statistike igraju važnu ulogu u odredjivanju dobre ocene za parametar. Ako je $\widehat{\theta}$ nepristrasna ocena za $\theta$, a $Y$ dovoljna statistika za isti parametar, tada je moguće naći neku funkciju od $Y$ koja će takodje biti nepristrasna ocena za $\theta$, a neće imati veću disperziju od $\widehat{\theta}$. Teorijsku podlogu za taj zaključak daje sledeća teorema koju navodimo bez dokaza.

**Teorema 3.3.4** *(Rao-Blekvelova teorema) Neka su $X$ i $Y$ slučajne promenljive takve da $Y$ ima očekivanje $\theta$ i konačnu disperziju $D(Y)$. Neka je $E(Y|X=x)=\varphi(x)$. Tada je $E(\varphi(X))=\theta$ i $D(\varphi(X))\leq D(Y)$.*

**Posledica 1** *Neka je $Y_1=u_1(X_1,\ldots,X_n)$ dovoljna statistika za parametar $\theta$. Neka je $Y_2=u_2(X_1,\ldots,X_n)$ neka druga statistika (koja nije funkcija od uzorka samo posredstvom $Y_1$) koja je nepristrasna ocena parametra $\theta$. Tada $E(Y_2|Y_1=y_1)=\varphi(y_1)$ definiše statistiku $\varphi(Y_1)$. Ova statistika, koja je funkcija od dovoljne statistike za $\theta$, je nepristrasna ocena za $\theta$ i njena disperzija je manja od disperzije za $Y_2$.*

**Dokaz.** Pošto je $Y_1$ dovoljna statistika za $\theta$, uslovna gustina raspodele za $Y_2$ pod uslovom $Y_1=y_1$ ne zavisi od $\theta$, tako da je $E(Y_2|Y_1=y_1)=\varphi(y_1)$ funkcija samo od $y_1$ (a ne i od $\theta$). Dakle, $\varphi(Y_1)$ je statistika. Prema Rao-Blekvelovoj teoremi,

$$E(\varphi(Y_1))=E(Y_2)=\theta\,,$$

jer je $Y_2$ nepristrasna ocena parametra $\theta$ i

$$D(\varphi(Y_1))\leq D(Y_2)\,.$$

No, kako $Y_2$ nije funkcija od uzorka samo preko $Y_1$, to je

$$D(\varphi(Y_1))<D(Y_2)\,.\square$$

Ova posledica nam ukazuje na to da u traganju za najboljom ocenom parametra $\theta$, pažnju možemo da ograničimo na dovoljnu statistiku ako ona postoji, jer polazeći od nje dolazimo do nepristrasne ocene za parametar čija je disperzija manja od disperzije bilo koje druge nepristrasne ocene.

### 3.3.3 Kompletnost

Definicija kompletnosti pripada teoriji mera. Mi ćemo je ovde koristiti da iskažemo jedno važno svojstvo dovoljnih statistika.

Definicija 19. Neka je dato obeležje $X$ čija raspodela pripada familiji dopustivih raspodela

$$\{f(x;\theta),\theta\in\Theta\}. \tag{3.8}$$

Izaberimo proizvoljnu Borelovu funkciju $u$ gde je $u(X)$ slučajna promenljiva koja ne zavisi od parametra $\theta$ i takvu da postoji $E(u(X))$. Ako je $E(u(X))=0$ za svako $\theta\in\Theta$, samo pod uslovom da je $u(x)=0$ u svakoj tački $x\in R$ u kojoj je bar jedan element familije (3.8) strogo pozitivan (može biti jednaka nuli samo na skupu mere nula), tada je familija (3.8) *kompletna*.

Mi ćemo ovde koristiti (zbog jednostavnosti dokazivanja) samo neprekidne funkcije $u$, a ne Borelove uopšte, kako je iskazano definicijom.

**Primer 31.** Neka je familija dopustivih raspodela $\{\mathcal{U}(0;\theta), \theta > 0\}$, tj.

$$f(x;\theta) = \begin{cases} \frac{1}{\theta}, & 0 < x < \theta \\ 0, & \text{inače} \end{cases} .$$

Dokažimo da je familija $\{f(x;\theta), \theta > 0\}$ kompletna.

Podjimo od pretpostavke da je $E(u(X)) = 0$, gde je $u$ neprekidna funkcija. Utvrdimo kada je ova pretpostavka moguća. Dakle,

$$E(u(X)) = \int_0^\theta u(x)\frac{1}{\theta}dx = \frac{1}{\theta}\int_0^\theta u(x)dx = 0\,,$$

a s druge strane poslednja jednakost je tačna ako i samo ako je

$$\int_0^\theta u(x)dx = 0\,.$$

Ako je $\theta > 0$ i nadjemo izvod po gornjoj granici, tj. po $\theta$, (što je čitaocu poznato kao izvod parametarskog integrala) dobićemo:

$$J'_\theta = \int_0^\theta u'_\theta(x)dx + \theta' \cdot u(\theta) - 0 = 0 \quad .$$

Iz poslednje jednakosti dobijamo da je $u(\theta) = 0$ za svako $\theta > 0$, a odavde sledi da je i $u(x) = 0$, za svako $x > 0$, pa je familija kompletna. $\triangle$

**Primer 32.** Neka je u pitanju diskretna raspodela i neka je obeležje $X : \mathcal{B}(1;\theta)$. Dokažimo da je familija $\{f(x;\theta), 0 \le \theta < 1\}$ kompletna. Dakle,

$$f(x;\theta) = \begin{cases} \theta^x(1-\theta)^{1-x}, & x = 0, 1 \\ 0, & \text{inače} \end{cases} .$$

Podjimo, kao i u predhodnom primeru, od pretpostavke da je $E(u(X)) = 0$ i da je funkcija $u$ neprekidna. Tada je:

$$0 = u(0)\theta^0(1-\theta)^1 + u(1)\theta^1(1-\theta)^0 = u(0) - \theta u(0) + u(1)\theta = (u(1) - u(0))\theta + u(0)$$

Linearna funkcija je identički jednaka nuli ako i samo ako su joj koeficijenti jednaki nuli, tj.

$$u(0) = 0$$
$$u(1) - u(0) = 0\,,$$

a odavde dobijamo da je

$$u(0) = u(1) = 0\,,$$

čime smo dokazali kompletnost familije. $\triangle$

### 3.3.4 Jedinstvenost najbolje statistike za parametar

Videli smo da, ako je $Y_1 = u_1(X_1, \ldots, X_n)$ bila dovoljna statistika za parametar $\theta$ i ako je postojala bilo koja nepristrasna statistika $Y_2$ za parametar $\theta$, koja nije bila funkcija od uzorka samo preko $Y_1$, tada je postojala bar još jedna funkcija od $Y_1$ različita od $Y_1$, koja je bila nepristrasna statistika za $\theta$. Tako se naše traganje za boljom statistikom (po kriterijumu srednje kvadratnog odstupanja) za $\theta$ može da suzi samo na funkcije od $Y_1$.

**Teorema 3.3.5** *Neka je $Y_1 = u_1(X_1, \ldots, X_n)$ dovoljna statistika za parametar $\theta$ raspodele obeležja $X$ posmatranog na populaciji iz koje je uzet uzorak, i neka je familija gustina raspodele statistike $Y_1$*

$$\{g_1(y_1; \theta), \theta \in \Theta\} \tag{3.9}$$

*kompletna. Ako postoji neprekidna funkcija $\varphi$ od $Y_1$ (definisana u posledici 1) koja je nepristrasna ocena parametra $\theta$, (tj. postoji statistika $\varphi(Y_1)$ takva da je $E(\varphi(Y_1)) = \theta$) tada je funkcija $\varphi$ skoro sigurno jedinstvena najbolja statistika za parametar $\theta$.*

**Dokaz.** Pretpostavimo suprotno tj. da sem funkcije $\varphi(Y_1)$ postoji još neka funkcija $\psi(Y_1)$, i neka su obe neprekidne funkcije koje ne zavise od parametra $\theta$, a nepristrasne su ocene parametra $\theta$. Tada je:

$$E(\varphi(Y_1) - \psi(Y_1)) = E(\varphi(Y_1)) - E(\psi(Y_1)) = \theta - \theta = 0.$$

Kako je (3.9) kompletna familija, to znači da je funkcija $\varphi - \psi = 0$ za sve vrednosti argumenta za koje je bar jedan element familije (3.9) strogo pozitivan. Dakle,

$$\varphi(Y_1) - \psi(Y_1) = 0$$

skoro sigurno, tj.

$$\varphi(Y_1) = \psi(Y_1)$$

skoro sigurno. Da je funkcija $\varphi$ i najbolja statistika zaključuje se prema teoremi Rao-Blekvela iz činjenice da je $D(\varphi(Y_1)) \leq D(Y_2)$ za bilo koju drugu nepristrasnu statistiku $Y_2$ istog uzorka. □

Ova teorema je samo specijalan slučaj opštije teoreme Lemana i Šefea.

Kao što je u definiciji kompletne familije naglašeno da se ona može iskazati i za funkcije koje nisu neprekidne (u tom slučaju se dokaz izvodi aparatom teorije mera), tako se i teorema o jedinstvenosti najbolje statistike za parametar koja se odnosi na kompletnu familiju može iskazati i bez pretpostavke o neprekidnosti, čime se u okviru ovog kursa nećemo baviti.

Iskaz da je $Y_1$ dovoljna statistika za parametar $\theta$, $\theta \in \Theta$, za koju je familija

$$\{g_1(y_1; \theta), \theta \in \Theta\}$$

funkcija gustina raspodele kompletna, zamenjuje se, radi jednostavnijeg izražavanja, iskazom *kompletna dovoljna statistika za $\theta$.*

**Primer 33.** Neka je data Puasonova raspodela $\{\mathcal{P}(\theta),\, 0 < \theta < \infty\}$ sa gustinom

$$f(x;\theta) = \begin{cases} \frac{\theta^x e^{-\theta}}{x!}, & x \in \{0,1,2,\ldots\} \\ 0, & \text{inače} \end{cases}$$

i prost slučajni uzorak $\mathbf{X}$ obima $n$ iz te raspodele. Dokazati da je $Y_1 = \sum_{i=1}^n X_i$ kompletna dovoljna statistika za parametar $\theta$.

Dovoljnost statistike sledi na osnovu

$$\begin{aligned} \prod_{i=1}^n f(x_i;\theta) &= \prod_{i=1}^n \frac{\theta^{x_i} e^{-\theta}}{x_i!} = \\ &= \exp\left(\ln\theta \cdot \sum_{i=1}^n x_i - n\theta\right) \cdot \frac{1}{\prod_{i=1}^n x_i!} \end{aligned}$$

i teoreme faktorizacije.

Kompletnost familije $\{g_1(y_1;\theta)\,,\, 0 < \theta < \infty\}$ sledi iz činjenice da je familija Puasonovih raspodela kompletna, tj. činjenice da $Y_1$ ima $\mathcal{P}(n\theta)$ raspodelu (što je poznato iz Teorije verovatnoće). Prema tome $Y_1$ ima gustinu

$$g_1(y_1;\theta) = \begin{cases} \frac{(n\theta)^{y_1} e^{-n\theta}}{y_1!}, & y_1 \in \{0,1,2,\ldots\} \\ 0, & \text{inače} \end{cases}$$

Uzmimo proizvoljnu neprekidnu funkciju $u$ za koju je $E(u(Y_1)) = 0$ i dokažimo da je to moguće samo ako je $u(y_1) = 0$ u tačkama $y_1 = 0, 1, 2, \ldots$.

Dakle, za svako $\theta > 0$

$$\begin{aligned} 0 = E(u(Y_1)) &= \sum_{y_1=0}^{\infty} u(y_1) \frac{(n\theta)^{y_1} e^{-n\theta}}{y_1!} = \\ &= e^{-n\theta}\left(u(0) + u(1)\frac{n\theta}{1!} + u(2)\frac{(n\theta)^2}{2!} + \cdots\right). \end{aligned}$$

Pošto $e^{-n\theta}$ nije jednako nuli, pokazali smo da je

$$0 = u(0) + nu(1)\theta + \left(\frac{n^2 u(2)}{2}\right)\theta^2 + \cdots.$$

Medjutim, ako takav beskonačni (stepeni) red konvergira ka nuli za svako $\theta > 0$, tada svaki koeficijenat uz odgovarajući stepen promenljive mora biti 0. To jest

$$u(0) = 0,\; nu(1) = 0,\; \frac{n^2 u(2)}{2} = 0, \ldots$$

i otuda

$$0 = u(0) = u(1) = u(2) = \cdots$$

što je i trebalo dokazati.

Dakle, $E(Y_1) = n\theta$, pa je tražena funkcija $\varphi$ koja daje jedinstvenu najbolju statistiku za parametar $\theta$

$$\varphi(Y_1) = \frac{Y_1}{n} = \overline{X}_n \quad .\triangle$$

Navedimo još jednu, iz operativnih razloga važnu teoremu:

**Teorema 3.3.6** *Neka je $(X_1, \ldots, X_n)$ uzorak iz raspodele $\{f(x;\theta), \theta \in \Theta\}$ gde je $\Theta$ interval. Neka je $Y_1 = u_1(X_1, \ldots, X_n)$ dovoljna statistika za $\theta$ sa kompletnom familijom gustina raspodele $\{g_1(y_1;\theta), \theta \in \Theta\}$. Neka je $Z = u(X_1, \ldots, X_n)$ bilo koja druga statistika (koja nije samo funkcija od $Y_1$). Ako raspodela za $Z$ ne zavisi od $\theta$, tada je $Z$ nezavisna slučajna promenljiva u odnosu na slučajnu promenljivu $Y_1$.*

**Dokaz.** Dokaz ćemo izvesti samo u specijalnom slučaju i to kada je obeležje X apsolutno neprekidnog tipa i kada je uslovna gustina raspodele statistike Z pod uslovom $Y_1 = y_1$ neprekidna funkcija od $y_1$.

Neka su redom:

$g_2(z)$ –marginalna gustina za Z,
$h(z|Y_1 = y_1)$ –uslovna gustina za Z pri uslovu $Y_1 = y_1$,
$g(y_1, z;\theta)$ –zajednička gustina za $(Y_1, Z)$.

Tada je integral:

$$\int_{-\infty}^{+\infty} h(z|y_1)g_1(y_1;\theta)dy_1 = \int_{-\infty}^{+\infty} \frac{g(y_1,z;\theta)}{g_1(y_1;\theta)} g_1(y_1;\theta)dy_1 = g_2(z)\,.$$

Odatle sledi da je

$$\int_{-\infty}^{+\infty} (g_2(z) - h(z|y_1))g_1(y_1;\theta)dy_1 = 0\,.$$

Kako je $Y_1$ dovoljna statistika za parametar $\theta$, to $h(z|y_1)$ ne zavisi od $\theta$ i, kako je familija kompletna i funkcija $g_2(z)$ konstanta po $y_1$ i ne zavisi od $\theta$, to je i razlika $g_2(z) - h(z|y_1)$ neprekidna funkcija od $y_1$ i ne zavisi od $\theta$. Dakle, biće $g_2(z) - h(z|y_1) = 0$. Sledi $g_2(z) = h(z|y_1)$. Odnosno, dobili smo da je:

$$h(z|y_1) = \frac{g(y_1,z;\theta)}{g_1(y_1;\theta)}$$

$$g(y_1,z;\theta) = g_1(y_1;\theta)h(z|y_1)$$

$$g(y_1,z;\theta) = g_1(y_1;\theta)g_2(z)\,.$$

Dakle, zajednička gustina je proizvod marginalnih gustina, a samim tim su $Y_1$ i Z nezavisne slučajne promenljive. □

**Primer 34.** Sada smo u mogućnosti da dodokažemo da su sredina i disperzija prostog slučajnog uzorka iz normalne raspodele $\{\mathcal{N}(m,\sigma^2), m \in R, \sigma^2 \in R^+\}$ nezavisne slučajne promenljive. Ranije smo se uverili da je sredina uzorka $\overline{X}_n$ za svako dato $\sigma^2 > 0$ kompletna dovoljna statistika za $m$ gde je $-\infty < m < \infty$. Da bismo dokazali pomenutu nezavisnost dovoljno je pokazati (prema Teoremi 3.3.6) da raspodela za $\overline{S}_n^2$ ne zavisi od $m$. U tu svrhu koristićemo funkciju generatrise momenata slučajne promenljive $\overline{S}_n^2$:

$$\begin{aligned} M(t) &= E(e^{t\overline{S}_n^2}) = \\ &= \int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} \exp\left(\frac{t}{n}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2\right)\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-\sum_{i=1}^{n}\frac{(x_i-m)^2}{2\sigma^2}\right)dx_1 \ldots dx_n\,. \end{aligned}$$

Funkcija $M(t)$ je definisana za $t < \frac{n}{2\sigma^2}$.

Napisani integral je komplikovan za izračunavanje i nećemo ga izračunavati, a činjenicu da $M(t)$ ne zavisi od $m$ dobićemo na sledeći način. Uvodimo $1-1$ transformaciju

$$\omega_1 = x_1 - m \quad , \quad \omega_2 = x_2 - m \quad , \ldots , \quad \omega_n = x_n - m \, .$$

Očigledno, njen Jakobijan je jednak 1. Koristićemo ovu transformaciju za uvodjenje smene u integral kojim je definisana posmatrana funkcija generatrisa momenata, pa su sume

$$\sum_{i=1}^{n}(x_i - m)^2 = \sum_{i=1}^{n}\omega_i^2 \quad , \quad \sum_{i=1}^{n}(x_i - \overline{x}_n)^2 = \sum_{i=1}^{n}(\omega_i - \overline{\omega})^2 \, .$$

Dakle,

$$M(t) = \int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} \exp\left(\frac{t}{n}\sum_{i=1}^{n}(\omega_i - \overline{\omega})^2\right)\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\omega_i^2\right) d\omega_1 \ldots d\omega_n$$

što, očigledno, ne zavisi od $m$. $\triangle$

U slučaju regularne eksponencijalne klase, o čemu će nadalje biti reči, teorema je tipa *ako i samo ako*. Teorema važi i u slučaju višedimenzionog parametra.

Uopštenje je moguće vršiti i u pravcu višedimenzionog obeležja sa višedimenzionim parametrom.

### 3.3.5 Dovoljna statistika za višedimenzioni parametar

Pojam dovoljnosti statistike se proširuje i na slučaj višedimenzionog nepoznatog parametra. Proširuju se i teoreme o kojima smo govorili u jednodimenzionom slučaju. Na primeru dvodimenzionog parametra obeležja apsolutno neprekidnog tipa to bi izgledalo ovako:

Definicija 20. Neka je data familija gustina raspodela $\{f(x;\theta), \theta = (\theta_1, \theta_2) \in \mathbf{\Theta} \subset R^2\}$, dopustiva za obeležje $X$ populacije iz koje je uzet uzorak obima $n$. Na osnovu uzorka je definisano $n$ statistika: $Y_1 = u_1(X_1, \ldots, X_n), \ldots, Y_n = u_n(X_1, \ldots, X_n)$, čiji je Jakobijan različit od nule, i sa zajedničkom gustinom raspodele $g(y_1, \ldots, y_n; \theta) = g(y_1, \ldots, y_n; \theta_1, \theta_2)$. Neka je $g_{12}(y_1, y_2; \theta_1, \theta_2)$ zajednička gustina raspodele statistika $Y_1$ i $Y_2$. Tada će dvodimenziona statistika $(Y_1, Y_2)$ biti *dovoljna statistika za dvodimenzioni parametar* $(\theta_1, \theta_2)$ ako i samo ako uslovna gustina raspodele data sa:

$$h(y_3, \ldots, y_n | y_1, y_2) = \frac{g(y_1, \ldots, y_n; \theta_1, \theta_2)}{g_{12}(y_1, y_2; \theta_1, \theta_2)}$$

ne zavisi bar od jednog od parametara $\theta_1$ ili $\theta_2$.

Fišer-Nejmanov kriterijum ima tada sledeću formulaciju:

Potreban i dovoljan uslov da dvodimenziona statistika $(Y_1, Y_2)$ iz prethodne definicije bude dovoljna statistika dvodimenzionog nepoznatog parametra $(\theta_1, \theta_2)$ je

$$f(x_1, \ldots, x_n; \theta_1, \theta_2) = g_{12}(u_1(x_1, \ldots, x_n), u_2(x_1, \ldots, x_n); \theta_1, \theta_2) H(x_1, \ldots, x_n)$$

gde za sve moguće vrednosti $u_1$ i $u_2$ funkcija $H(x_1,\dots,x_n)$ ne zavisi od jedne ili obe komponente nepoznatog parametra.

Potreban i dovoljan uslov da $(Y_1,Y_2)$ bude dovoljna statistika prema teoremi faktorizacije bi izgledao ovako:

$$f(x_1,\dots,x_n;\theta_1,\theta_2)=k(u_1(x_1,\dots,x_n),u_2(x_1,\dots,x_n);\theta_1,\theta_2)K(x_1,\dots,x_n)$$

gde za sve vrednosti iz kodomena funkcija $u_1$ i $u_2$ funkcija $K(x_1,\dots,x_n)$ ne zavisi bar od jedne od komponenata nepoznatog parametra.

**Primer 35.** Neka je dat prost uzorak $\mathbf{X}=(X_1,\dots,X_n)$ iz populacije sa normalnom raspodelom obeležja, tj. sa familijom dopustivih raspodela

$$\{\mathcal{N}(\theta_1,\theta_2)\,,\ -\infty<\theta_1<+\infty\,,\ 0<\theta_2<\infty\}\,.$$

Gustina raspodele obeležja $X$ je

$$f(x;\theta_1,\theta_2)=\exp\left(-\frac{1}{2\theta_2}x^2+\frac{\theta_1}{\theta_2}x-\frac{\theta_1^2}{2\theta_2}-\ln\sqrt{2\pi\theta_2}\right)\,.$$

Posmatrajmo statistike

$$Y_1=\sum_{i=1}^n X_i\quad\text{i}\quad Y_2=\sum_{i=1}^n X_i^2\,.$$

Lako je dokazati da je $(Y_1,Y_2)$ dvodimenziona dovoljna kompletna statistika za parametar $(\theta_1,\theta_2)$. Štaviše, ako definišemo preslikavanje $1-1$ na sledeći način:

$$Z_1=\frac{Y_1}{n}=\overline{X}_n\quad,\quad Z_2=\frac{Y_2-\frac{Y_1^2}{n}}{n-1}=\frac{\sum_{i=1}^n(X_i-\overline{X}_n)^2}{n-1}=\tilde{S}_n^2\,,$$

onda će $(Z_1,Z_2)$ biti dvodimenziona kompletna dovoljna statistika za $(\theta_1,\theta_2)$ i jedina takva da je nepristrasna, tj.

$$E(Z_1)=\theta_1\quad,\quad E(Z_2)=\theta_2\,.$$

Detaljnije ćemo se ovde pozabaviti samo dovoljnošću statistike $(Y_1,Y_2)$. Zaista, zajednička gustina ovog prostog uzorka je

$$\prod_{i=1}^n f(x_i;\theta_1,\theta_2)=\prod_{i=1}^n\frac{1}{(2\pi\theta_2)^{\frac{1}{2}}}e^{-\frac{(x_i-\theta_1)^2}{2\theta_2}}=$$

$$=(2\pi)^{-\frac{n}{2}}\theta_2^{-\frac{n}{2}}e^{-\frac{1}{2\theta_2}(\sum_{i=1}^n x_i^2-2\theta_1\sum_{i=1}^n x_i+n\theta_1^2)}$$

gde se može uzeti

$$k(y_1,y_2;\theta_1,\theta_2)=e^{-\frac{1}{2\theta_2}(y_2-2\theta_1y_1+n\theta_1^2)}$$

i

$$K=(2\pi\theta_2)^{-\frac{n}{2}}$$

pa prema teoremi faktorizacije sledi tvrdjenje.

$\triangle$

## 3.4 Regularna familija gustina raspodele

U daljem izlaganju u vezi sa ocenjivanjem parametara biće od interesa da zajedničku gustinu raspodele slučajnog uzorka posmatramo kao funkciju od nepozatog parametra, jednodimenzionog ili višedimenzionog, zavisno od definicije problema koji ćemo rešavati.

DEFINICIJA 21. Neka je dat uzorak $\mathbf{X} = (X_1, \ldots, X_n)$ iz populacije sa obeležjem $X$ čija raspodela pripada familiji dopustivih raspodela $\{f(x;\theta),\ \theta \in \mathbf{\Theta}\}$. *Funkcija verodostojnosti* za parametar $\theta$, u oznaci $L(\theta)$, je zajednička gustina raspodele vektora $\mathbf{X}$ posmatrana kao funkcija od $\theta$

$$L(\theta) = L(\theta; x_1, \ldots, x_n) = f(x_1, \ldots, x_n; \theta) \quad , \quad (x_1, \ldots, x_n) \in R^n \, .$$

Uvedimo sada definiciju regularne familije gustina raspodele za jednodimenzioni parametar:

DEFINICIJA 22. Neka je data funkcija $u : R^n \longrightarrow R$ i familija dopustivih raspodela $\{f(x;\theta), \theta \in \Theta\}$ za obeležje $X$ neke populacije iz koje se uzima uzorak obima $n$. Kazaćemo da je *familija dopustivih raspodela regularna* ako za funkciju $u$ i familiju dopustivih raspodela važe sledeći *uslovi regularnosti*:

- $(R_1)$ $\Theta$ je interval (ili ceo skup $R$)
- $(R_2)$ Skup $K = \{\mathbf{x} = (x_1, \ldots, x_n) : L(\theta, \mathbf{x}) > 0\} \subset R^n$ ne zavisi od $\theta$
- $(R_3)$ Funkcija $L(\theta)$ je diferencijabilna po $\theta$ i važi[1]

$$\frac{\partial}{\partial\theta} \int_{\mathbf{R}^n} L(\theta, \mathbf{x}) d\mathbf{x} = \int_{\mathbf{R}^n} \frac{\partial L(\theta, \mathbf{x})}{\partial\theta} d\mathbf{x}$$

$$\frac{\partial}{\partial\theta} \int_{\mathbf{R}^n} u(\mathbf{x}) L(\theta, \mathbf{x}) d\mathbf{x} = \int_{\mathbf{R}^n} u(\mathbf{x}) \frac{\partial L(\theta, \mathbf{x})}{\partial\theta} d\mathbf{x}$$

i analogno za diskretan slučaj

$$\frac{\partial}{\partial\theta} \sum_{\mathbf{x}\in\mathbf{K}} L(\theta, \mathbf{x}) = \sum_{\mathbf{x}\in\mathbf{K}} \frac{\partial L(\theta, \mathbf{x})}{\partial\theta}$$

$$\frac{\partial}{\partial\theta} \sum_{\mathbf{x}\in\mathbf{K}} u(\mathbf{x}) L(\theta, \mathbf{x}) = \sum_{\mathbf{x}\in\mathbf{K}} u(\mathbf{x}) \frac{\partial L(\theta, \mathbf{x})}{\partial\theta} \, .$$

**Teorema 3.4.1** *( Rao-Kramerova donja granica)*

*Neka je dat uzorak $\mathbf{X} = (X_1, \ldots, X_n)$ iz populacije sa obeležjem $X$ čija raspodela pripada familiji dopustivih raspodela $\{f(x;\theta), \theta \in \Theta\}, \Theta \subset R$, koja je regularna. Neka je $Y$ statistika posmatranog uzorka sa konačnim drugim momentom i takva da je nepristrasna ocena diferencijabilne funkcije $\varphi(\theta)$ parametra $\theta \in \Theta$, $\varphi : \Theta \longrightarrow R$, posmatrane familije. Tada važi*

$$D(Y) \geq \frac{(\varphi'(\theta))^2}{E\left(\frac{\partial \ln L(\theta)}{\partial\theta}\right)^2}$$

*za svako $\theta \in \Theta$.*

[1] U daljem tekstu biće korišćena oznaka $\int_{\mathbf{R}^n} f(\mathbf{x}) d\mathbf{x} = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} f(x_1, \ldots, x_n) dx_1 \ldots dx_n$

**Dokaz.**Dokaz ćemo navesti za apsolutno neprekidan slučaj, a diskretan slučaj se razmatra analogno.

Kako je funkcija verodostojnosti funkcija gustine raspodele vektora $\mathbf{X}$, to je

$$\int_{\mathbf{R}^n} L(\theta;\mathbf{x})d\mathbf{x} = 1\,. \tag{3.10}$$

Neka je

$$Y = u(\mathbf{X})\,.$$

Tada, zbog apsolutne neprekidnosti i nepristrasnosti statistike $Y$ imamo

$$E(Y) = \int_{\mathbf{R}^n} u(\mathbf{x})L(\theta;\mathbf{x})d\mathbf{x} = \varphi(\theta) \quad . \tag{3.11}$$

Diferenciranjem po $\theta$ jednakosti (3.10) i (3.11) i na osnovu osobine regularnosti, dobijamo:

$$\int_{\mathbf{R}^n} \frac{\partial L(\theta;\mathbf{x})}{\partial\theta}d\mathbf{x} = \frac{\partial}{\partial\theta}\int_{\mathbf{R}^n} L(\theta;\mathbf{x})d\mathbf{x} = 0\,, \tag{3.12}$$

kao i

$$\frac{\partial}{\partial\theta}\int_{\mathbf{R}^n} u(\mathbf{x})L(\theta,\mathbf{x})d\mathbf{x} = \frac{\partial\varphi(\theta)}{\partial\theta}\,,$$

tj.

$$\int_{\mathbf{R}^n} u(\mathbf{x})\frac{\partial L(\theta;\mathbf{x})}{\partial\theta}d\mathbf{x} = \varphi'(\theta)\,. \tag{3.13}$$

Iz (3.12) dobijamo

$$0 = \int_{\mathbf{R}^n} \frac{\partial L(\theta;\mathbf{x})}{\partial\theta}d\mathbf{x} = \int_{\mathbf{R}^n} \left(\frac{1}{L(\theta;\mathbf{x})}\frac{\partial L(\theta;\mathbf{x})}{\partial\theta}\right) L(\theta;\mathbf{x})d\mathbf{x} = \int_{\mathbf{R}^n} \frac{\partial \ln L(\theta;\mathbf{x})}{\partial\theta}L(\theta;\mathbf{x})d\mathbf{x} =$$
$$= E\left(\frac{\partial \ln L(\theta;\mathbf{X})}{\partial\theta}\right)\,,$$

odakle

$$0 = 0\cdot\varphi(\theta) = \int_{\mathbf{R}^n} \varphi(\theta)\frac{\partial \ln L(\theta;\mathbf{x})}{\partial\theta}L(\theta;\mathbf{x})d\mathbf{x}\,. \tag{3.14}$$

Iz (3.13) dobijamo

$$\varphi'(\theta) = \int_{\mathbf{R}^n} u(\mathbf{x})\frac{\partial L(\theta;\mathbf{x})}{\partial\theta}d\mathbf{x} = \int_{\mathbf{R}^n} u(\mathbf{x})\frac{\partial \ln L(\theta;\mathbf{x})}{\partial\theta}L(\theta;\mathbf{x})d\mathbf{x}\,. \tag{3.15}$$

Oduzimanjem (3.14) od (3.15) dobijamo:

$$\varphi'(\theta) - 0 = \int_{\mathbf{R}^n} (u(\mathbf{x}) - \varphi(\theta))\frac{\partial \ln L(\theta;\mathbf{x})}{\partial\theta}L(\theta;\mathbf{x})d\mathbf{x}\,,$$

tj.

$$\varphi'(\theta) = \int_{\mathbf{R}^n} (u(\mathbf{x}) - \varphi(\theta))\frac{\partial \ln L(\theta;\mathbf{x})}{\partial\theta}L(\theta;\mathbf{x})d\mathbf{x}\,.$$

Koristeći nejednakost Koši- Švarc-Bunjakovskog dobijamo:

$$\begin{aligned}
|\varphi'(\theta)|^2 &= \left|\int_{\mathbf{R}^n}(u(\mathbf{x})-\varphi(\theta))\frac{\partial \ln L(\theta;\mathbf{x})}{\partial\theta}L(\theta;\mathbf{x})d\mathbf{x}\right|^2 \\
&\leq \int_{\mathbf{R}^n}(u(\mathbf{x})-\varphi(\theta))^2 L(\theta;\mathbf{x})d\mathbf{x}\int_{\mathbf{R}^n}\left(\frac{\partial \ln L(\theta;\mathbf{x})}{\partial\theta}\right)^2 L(\theta;\mathbf{x})d\mathbf{x} = \\
&= D(Y)\cdot D\left(\frac{\partial \ln L(\theta;\mathbf{X})}{\partial\theta}\right) = D(Y)E\left[\left(\frac{\partial \ln L(\theta;\mathbf{X})}{\partial\theta}\right)^2\right],
\end{aligned}$$

odnosno

$$D(Y) \geq \frac{|\varphi'(\theta)|^2}{E\left(\frac{\partial \ln L(\theta;X_1,\ldots,X_n)}{\partial\theta}\right)^2} \quad ,$$

što je i trebalo dokazati. □

Primetimo izraz koji igra važnu ulogu u definisanju Rao–Kramerove donje granice:

$$I(\theta) = E\left(\frac{\partial \ln L(\theta;X_1,\ldots,X_n)}{\partial\theta}\right)^2$$

i koji je u statističkoj literaturi poznat pod nazivom *Fišerova količina informacija*.

Rao–Kramerova nejednakost nam omogućava da kod regularnih familija dopustivih raspodela odredimo donju granicu disperzija svih nepristrasnih statistika za ocenu nepoznatog parametra raspodele. Detaljnije ćemo se baviti samo jednodimenzionim slučajem.

Imajući u vidu definiciju najbolje statistike za parametar, zaključujemo da je u regularnom slučaju to ona statistika čija disperzija dostiže Rao-Kramerovu donju granicu. U tom slučaju definišemo najefikasniju statistiku za parametar.

Definicija 23. *Efikasnost nepristrasne statistike* u regularnom slučaju tačkastog ocenjivanja parametra raspodele je količnik Rao–Kramerove donje granice i disperzije statistike koja se koristi kao ocena za posmatrani parametar.

Definicija 24. *Najefikasnija statistika* za parametar regularne familije je nepristrasna statistika čija je efikasnost jednaka jedinici.

**Primer 36** Statistika $\overline{X}_n$ je najefikasnija statistika za parametar $\theta$ Puasonove raspodele $\mathcal{P}(\theta)$ na osnovu prostog slučajnog uzorka. Dokazati! △

**Primer 37.** Neka je dat prost slučajni uzorak obima $n$, $n > 1$, iz populacije sa obeležjem $X$ čija raspodela pripada familiji normalnih raspodela $\{\mathcal{N}(m,\theta);\theta > 0\}$. Ranije smo dokazali da je statistika $\widetilde{S}_n^2$ nepristrasna ocena parametra $\theta$. Lako se može proveriti da je Rao-Kramerova donja granica za parametar $\theta$ normalne raspodele ($m$ je poznati parametar) jednaka $\frac{2\theta^2}{n}$, a njena disperzija $\frac{2\theta^2}{n-1}$, pa je njena efikasnost $\frac{n-1}{n}$. Zaključujemo da je ispitivana statistika samo asimptotski najefikasnija.△

Navešćemo bez dokaza još neke činjenice u vezi sa Rao-Kramerovom donjom granicom.

Ako za neku nepristrasnu ocenu utvrdimo da joj disperzija dostiže Rao-Kramerovu donju granicu, tada je moguće dokazati da je ta statistika takodje i dovoljna statistika za posmatrani parametar. Drugim rečima, klasa najefikasnijih statistika je potklasa klase dovoljnih statistika.

Najzad, konstatujmo da postoji odgovarajuća nejednakost i za neregularne gustine raspodela, ali se time ovde nećemo baviti.

### 3.4.1 Eksponencijalna klasa funkcija gustina raspodele

DEFINICIJA 25. Neka je $\{f(x;\theta), \theta \in \Theta\}$ familija dopustivih gustina raspodele apsolutno neprekidnog tipa takvih da je $\Theta = \{\theta | \gamma < \theta < \delta\}$, $\gamma, \delta \in \bar{R}$, $\gamma < \delta$, tj. interval, i neka je:

$$f(x;\theta) = \begin{cases} \exp\left(p(\theta)K(x) + S(x) + q(\theta)\right), & a < x < b \\ 0, \quad \text{inače} \quad , & a, b \in R \end{cases} \tag{3.16}$$

pri čemu je $p(\theta)$ netrivijalna neprekidna funkcija parametra $\theta$, a funkcije $S(x)$ i $K'(x)$ su neprekidne po $x$, pri čemu $K'(x)$ nije identički jednaka 0. Tada se familija raspodela definisana gustinom (3.16) zove *eksponencijalna klasa gustina raspodele* u neprekidnom slučaju.

Definicija se može iskazati i za raspodele diskretnog tipa, s tim što tada umesto intervala $(a, b)$ imamo diskretan skup tačaka. U tom slučaju bismo govorili o regularnoj eksponencijalnoj klasi gustina raspodele diskretnog tipa.

Kada interval $(a, b)$, odnosno njegove granice, ne zavise od $\theta$, može se dokazati da je eksponencijalna klasa regularna, tj. da imamo *regularni slučaj eksponencijalne klase.*

**Primer 38.** Neka je data familija $\{\mathcal{N}(0, \theta), \theta > 0\}$. Dakle,

$$f(x;\theta) = \frac{1}{\sqrt{2\pi\theta}} e^{-\frac{x^2}{2\theta}},$$

odnosno,

$$f(x;\theta) = e^{-\ln\sqrt{2\pi\theta}} e^{-\frac{x^2}{2\theta}} = \exp\left(-\frac{1}{2\theta}x^2 - \ln\sqrt{2\pi\theta}\right)$$

je gustina raspodele obeležja $X$ sa normalnom raspodelom. Ovde su očigledno funkcije:

$$p(\theta) = -\frac{1}{2\theta} \quad , \quad K(x) = x^2 \quad , \quad S(x) = 0 \quad , \quad q(\theta) = -\ln\sqrt{2\pi\theta} \quad , \quad x \in R\,.$$

Vidimo da je $p$ netrivijalna, $S$ i $K'$ su neprekidne, pa su uslovi eksponencijalnosti zadovoljeni.△

**Primer 39.** Neka je data raspodela diskretnog tipa $\{\mathcal{P}(\theta), 0 < \theta < \infty\}$ gde je

$$f(x;\theta) = \begin{cases} \frac{\theta^x e^{-\theta}}{x!}, & x = 0, 1, 2, \ldots \\ 0, \quad \text{inače} \quad . \end{cases}$$

Odavde se dobija

$$\frac{\theta^x e^{-\theta}}{x!} = \exp(x\ln\theta - \ln(x!) - \theta)$$

pri čemu je:

$$q(\theta) = -\theta, \quad p(\theta) = \ln\theta, \quad K(x) = x, \quad S(x) = -\ln(x!)\,,$$

pa bi Puasonova raspodela bila primer eksponencijalne klase u diskretnom slučaju. $\triangle$

Ako uzmemo prost slučajni uzorak $\mathbf{X} = (X_1, \ldots, X_n)$ iz populacije čija raspodela pripada klasi eksponencijalnih funkcija gustina raspodele, onda bi ovaj vektor imao gustinu raspodele datu sa

$$f(x_1;\theta)\cdots f(x_n;\theta) = \begin{cases} \exp(p(\theta)\sum_{i=1}^n K(x_i) + nq(\theta))\exp(\sum_{i=1}^n S(x_i)), & a < x_i < b, \\ & i = 1, \ldots, n \\ 0, & \text{inače} \quad . \end{cases}$$

U vezi sa prostim slučajnim uzorkom navodimo bez dokaza sledeću teoremu.

**Teorema 3.4.2** *Neka je za obeležje $X$ $\{f(x;\theta), \gamma < \theta < \delta\}$, $\gamma, \delta \in R$, familija dopustivih gustina raspodele iz eksponencijalne klase i neka imamo prost slučajni uzorak $\mathbf{X} = (X_1, \ldots, X_n)$ konstantnog obima $n$ iz populacije sa obeležjem $X$. Statistika*

$$Y_1 = \sum_{i=1}^n K(X_i)$$

*je dovoljna statistika za $\theta$, a familija $\{g_1(y_1;\theta), \gamma < \theta < \delta\}$ gustina raspodele za $Y_1$ je kompletna, tj. $Y_1$ je kompletna dovoljna statistika za $\theta$.*

Navedimo još jedno važno svojstvo regularnog slučaja eksponencijalne klase, a to je da Teorema 3.3.6 daje potrebne i dovoljne uslove ukoliko familija dopustivih raspodela pripada ovoj klasi.

Eksponencijalna klasa gustina raspodele se definiše i u slučaju višedimenzionog parametra.

Definicija 26. Neka je $\boldsymbol{\Theta} = \{\boldsymbol{\theta} | \boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)'\}$ parametarski prostor. Familija dopustivih raspodela $\{f(x;\boldsymbol{\theta}); \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ obeležja $X$ posmatranog na populaciji je eksponencijalna, ako ima gustinu

$$f(x;\boldsymbol{\theta}) = C(\boldsymbol{\theta}) \exp\left(\sum_{j=1}^k Q_j(\boldsymbol{\theta}) T_j(x)\right) U(x)\,,$$

gde su $C$ i $Q_j$ merljive funkcije na parametarskom prostoru $\boldsymbol{\Theta}$, a $T_j$ i $U$ su merljive funkcije po $x$ i još važi da je $C(\boldsymbol{\theta}) > 0$ i $U(x) \geq 0$ i to za svako $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ i svako $x$ iz skupa vrednosti za obeležje $X$.

# 3.5 Metodi tačkastog ocenjivanja parametara

## 3.5.1 Metod maksimalne verodostojnosti

Metod maksimalne verodostojnosti je opšti metod za ocenjivanje nepoznatih parametara raspodele i može se primenjivati (sa više ili manje uspeha) kod regularnih i neregularnih familija, zatim kod proizvoljnih uzoraka, prostih ili ne itd.

Metod maksimalne verodostojnosti se primenjuje, kako za jednodimenzioni, tako i za višedimenzioni parametar.

Neka je data familija dopustivih gustina raspodele $\{f(x;\boldsymbol{\theta}),\, \boldsymbol{\theta} \in \Theta\}$ obeležja X i uzorak $(X_1,\ldots,X_n)$ sa zajedničkom gustinom raspodele $f(x_1,\ldots,x_n;\boldsymbol{\theta})$ iz populacije sa obeležjem $X$. Kako smo već istakli, zajedničku gustinu možemo da posmatramo kao funkciju parametra $\boldsymbol{\theta}$ i u tom slučaju se ova funkcija zove funkcija verodostojnosti i najčešće se označava sa $L(\boldsymbol{\theta};x_1,\ldots,x_n) = f(x_1,\ldots,x_n;\boldsymbol{\theta})$.

Definicija 27. *Ocenom maksimalne verodostojnosti* parametra $\theta \in \Theta \subset R$ zvaćemo onu statistiku $Y = u(X_1,\ldots,X_n)$ koja daje maksimum funkcuje $L$ po $\theta \in \Theta$ za svako $\mathbf{x} = (x_1,\ldots,x_n)$ iz uzoračkog prostora $\mathcal{X} \subset R^n$ realizovanih uzoraka fiksnog obima $n$. Oznaka za ocenu maksimalne verodostojnosti biće $\widehat{\theta} = u(X_1,\ldots,X_n)$, tj. za proizvoljni $\mathbf{x} \in \mathcal{X}$, $\widehat{\theta} = u(x_1,\ldots,x_n)$ je ona funkcija za koju je

$$L(\widehat{\theta};x_1,\ldots,x_n) = \max_{\theta\in\Theta} L(\theta;x_1,\ldots,x_n).$$

Ako je $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_r)$ višedimenzioni parametar, ocena maksimalne verodostojnosti će, jasno, biti $r$–dimenziona statistika $\widehat{\boldsymbol{\theta}} = (\widehat{\theta_1},\ldots,\widehat{\theta_r})$ koja maksimalizuje funkciju verodostojnosti $L(\boldsymbol{\theta})$

U slučaju da je uzorak prost, zajednička gustina je jednaka proizvodu marginalnih gustina, pa je:

$$L(\boldsymbol{\theta};x_1,\ldots,x_n) = f(x_1;\boldsymbol{\theta})f(x_2;\boldsymbol{\theta})\ldots f(x_n;\boldsymbol{\theta})\,.$$

**Primer 40.** Neka je dat prost slučajni uzorak $(X_1,\ldots,X_n)$ iz populacije sa obeležjem $X$ čija raspodela pripada familiji dopustivih raspodela

$$\{\mathcal{N}(\theta,1), \theta \in R\}.$$

Ocena maksimalne verodostojnosti će se dobiti na sledeći način:

$$\begin{aligned} L(\theta;x_1,\ldots,x_n) &= f(x_1;\theta)f(x_2;\theta)\ldots f(x_n;\theta) = \\ &= \frac{1}{\sqrt{2\pi}}e^{-\frac{(x_1-\theta)^2}{2}}\frac{1}{\sqrt{2\pi}}e^{-\frac{(x_2-\theta)^2}{2}}\cdots\frac{1}{\sqrt{2\pi}}e^{-\frac{(x_n-\theta)^2}{2}} = \\ &= (2\pi)^{-\frac{n}{2}}e^{-\frac{1}{2}\sum_{i=1}^n(x_i-\theta)^2}\,. \end{aligned}$$

Logaritmovanjem leve i desne strane, dobićemo:

$$\ln L(\theta) = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\sum_{i=1}^n(x_i-\theta)^2.$$

Da bismo našli ocenu maksimalne verodostojnosti parametra $\theta$ treba naći maksimum funkcije $L$, i to ćemo učiniti posredno odredjujući maksimum funkcije $\ln L$. Dakle, iz

$$\frac{\partial \ln L(\theta)}{\partial \theta} = -\frac{1}{2}\sum_{i=1}^{n} 2(x_i - \theta)(-1) = 0$$

sledi

$$\sum_{i=1}^{n} x_i - n\theta = 0\,,$$

odnosno,

$$\theta = \frac{1}{n}\sum_{i=1}^{n} x_i\,,$$

pa je tražena statistika

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i\,,$$

tj.

$$\widehat{\theta} = \overline{X_n}\,.$$

Uočimo da je ova ocena nepristrasna.△

**Primer 41.** Neka je dat prost slučajni uzorak $(X_1, \ldots, X_n)$ iz populacije sa obeležjem $X$ čija raspodela pripada familiji dopustivih raspodela

$$\{\mathcal{U}(0, \theta]; 0 < \theta < 1\}.$$

Gustina je

$$f(x;\theta) = \begin{cases} \frac{1}{\theta}, & 0 < x \leq \theta \\ 0, & \text{inače} \quad . \end{cases}$$

Funkcija verodostojnosti je

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta) = \frac{1}{\theta^n}, \quad 0 < x_i \leq \theta, \quad i = 1, \ldots, n.$$

Ova funkcija će imati maksimum kada je $\theta$ najmanje moguće, pa ćemo za ocenu maksimalne verodostojnosti uzeti statistiku

$$\widehat{\theta} = \max_{i\in\{1,\ldots,n\}} X_i \;\; = \;\; X_{(n)}\,.$$

Dakle, ocena maksimalne verodostojnosti je statistika poretka reda $n$. Ova ocena nije nepristrasna, ali jeste asimptotski nepristrasna. △

**Teorema 3.5.1** *Ako postoji jedinstvena dovoljna statistika $Y = u(X_1, \ldots, X_n)$ za parametar $\theta$ na osnovu uzorka $\mathbf{X} = (X_1, \ldots, X_n)$ iz populacije sa obeležjem $X$ čija raspodela pripada familiji dopustivih raspodela koja zavisi od $\theta$ i ako, takodje, postoji ocena maksimalne verodostojnosti $\widehat{\theta}$ za parametar $\theta$, tada je $\widehat{\theta}$ funkcija od dovoljne statistike $Y$, tj. $\widehat{\theta} = \upsilon(Y)$.*

**Dokaz.** Neka je $g(y;\theta)$ gustina raspodele statistike $Y$. Tada je funkcija verodostojnosti (prema Fišer-Nejmanovom kriterijumu):

$$L(\theta; x_1, \ldots, x_n) \;=\; g(u(x_1, \ldots, x_n); \theta) H(x_1, \ldots, x_n),$$

pa će maksimum funkcije verodostojnosti biti funkcija od uzorka samo preko funkcije $u$, što znači da je $\widehat{\theta}$ funkcija od dovoljne statistike. □

Pod odredjenim uslovima ocena maksimalne verodostojnosti je strogo postojana ocena za parametar $\theta$, a u regularnom slučaju je asimptotski normalna i asimptotski najefikasnija.

### 3.5.2 Metod momenata

Još jedan metod tačkastog ocenjivanja parametara uveo je Pirson. To je tzv. metod momenata. Da bismo izložili ovaj metod uvešćemo još neke definicije teorije uzoraka, kao i pretpostavku da posmatrano obeležje $X$ ima momente do nekog reda $r$.

Definicija 28. Za uzorak $(X_1, \ldots, X_n)$ obima $n$ iz populacije sa obeležjem $X$, *uzorački moment reda $k$* je statistika

$$A_k = \frac{1}{n} \sum_{i=1}^{n} X_i^k \quad , \quad k = 1, 2, \ldots, r\,,$$

a *uzorački centralni moment reda $k$* je statistika

$$M_k = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X}_n)^k \quad , \quad k = 1, 2, \ldots, r\,.$$

Specijalni slučajevi ovako definisanih momenata su sredina i disperzija uzorka, pominjani već ranije.

Da bi se istakla zavisnost uzoračkih momenata od obima uzorka $n$, koriste se oznake $A_{nk}$ i $M_{nk}$ za običan i centralni uzorački moment reda $k$.

**Teorema 3.5.2** *Uzorački moment reda $k$ prostog slučajnog uzorka je nepristrasna i postojana ocena teorijskog momenta reda $k$ datog obeležja.*

**Dokaz.** Označimo sa $\alpha_k = EX^k$ $(k \le r)$ teorijski moment reda $k$ obeležja $X$ populacije iz koje je uzet uzorak $(X_1, \ldots, X_n)$. Tada je:

$$E(A_{nk}) = E\left(\frac{1}{n} \sum_{i=1}^{n} X_i^k\right) = \frac{1}{n} \sum_{i=1}^{n} E(X_i^k) = \frac{1}{n} n\alpha_k = \alpha_k,$$

što dokazuje da je uzorački moment nepristrasna ocena. Dokažimo postojanost ove ocene koristeći nejednakost Čebiševa:

$$P\{|A_{nk} - \alpha_k| \ge \varepsilon\} \le \frac{E(A_{nk} - \alpha_k)^2}{\varepsilon^2} = \frac{D(A_{nk})}{\varepsilon^2}\,,$$

gde je

$$D(A_{nk}) = E(A_{nk}^2) - (EA_{nk})^2 = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i^k\right)^2 - \alpha_k^2 .$$

Kako je

$$\begin{aligned} E\left(\frac{1}{n}\sum_{i=1}^{n} X_i^k\right)^2 &= E\left(\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} X_i^k X_j^k\right) = \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} E(X_i^k X_j^k) = \\ &= \frac{1}{n^2} nEX^{2k} + \frac{1}{n^2}(n^2-n)EX^k EX^k = \\ &= \frac{1}{n}\alpha_{2k} + \frac{n-1}{n}\alpha_k^2 \quad , \end{aligned}$$

to je

$$D(A_{nk}) = \frac{1}{n}\alpha_{2k} + \frac{n-1}{n}\alpha_k^2 - \alpha_k^2 = \frac{\alpha_{2k} - \alpha_k^2}{n} ,$$

odnosno,

$$P\{|A_{nk} - \alpha_k| \geq \varepsilon\} \leq \frac{\alpha_{2k} - \alpha_k^2}{n\varepsilon^2} \quad \longrightarrow 0 \quad , \quad n \to \infty . \square$$

Analogno tvrdjenje ovome važi i za centralne momente $M_{nk}$, kao i za bilo koje druge uzoračke karakteristike koje su definisane kao neprekidne funkcije od konačnog broja statistika $A_{nk}$. Ovo je tačno kao posledica sledeće teoreme teorije verovatnoće koju navodimo bez dokaza:

**Teorema 3.5.3** *Neka niz $r$–dimenzionih slučajnih promenljivih čiji je opšti član $(\xi_1(n), \ldots, \xi_r(n))$, konvergira u verovatnoći ka konstanti $(c_1, \ldots, c_r)$. Tada za neprekidnu funkciju $\varphi : R^r \longrightarrow R$, važi*

$$\zeta(n) = \varphi(\xi_1(n), \ldots, \xi_r(n)) \xrightarrow{P} \varphi(c_1, \ldots, c_r) , \quad n \longrightarrow \infty \quad .$$

**Teorema 3.5.4** *Uzoračkи moment $A_{nk}$ prostog slučajnog uzorka je asimptotski normalan, tj. ima približno $\mathcal{N}(\alpha_k, \frac{\alpha_{2k}-\alpha_k^2}{n})$ raspodelu kada obim uzorka neograničeno raste.*

**Dokaz.** Polazimo od prostog slučajnog uzorka$(X_1, \ldots, X_n)$ iz populacije sa obeležjem $X$ proizvoljne raspodele koja ima momente $E(X^k) = \alpha_k$ i $D(X^k) = \alpha_{2k} - \alpha_k^2$. Tada je, kao u Teoremi 3.5.2,

$$A_{nk} = \frac{1}{n}\sum_{i=1}^{n} X_i^k \quad , \quad D(A_{nk}) = \frac{\alpha_{2k} - \alpha_k^2}{n} .$$

Koristeći centralnu graničnu teoremu zaključujemo da niz slučajnih promenljivih definisanih opštim članom

$$\eta(n) = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i^k - \alpha_k}{\sqrt{\frac{\alpha_{2k}-\alpha_k^2}{n}}} = \frac{A_{nk} - \alpha_k}{\sqrt{\alpha_{2k}-\alpha_k^2}}\sqrt{n}$$

konvergira u raspodeli ka slučajnoj promenljivoj sa $\mathcal{N}(0,1)$ raspodelom. Izražavajući $A_{nk}$ kao funkciju od $\eta(n)$, dobijamo

$$A_{nk} = \eta(n)\sqrt{\frac{\alpha_{2k}-\alpha_k^2}{n}} + \alpha_k\,.$$

Kako je $A_{nk}$ linearna funkcija od $\eta(n)$ i $\eta(n)$ ima asimptotski normalnu raspodelu, to i $A_{nk}$ ima, za dovoljno veliki obim uzorka, približno normalnu raspodelu $\mathcal{N}(\alpha_k, \frac{\alpha_{2k}-\alpha_k^2}{n})$. □

Ova teorema omogućava da se za veliki obim uzorka oceni greška ocenjivanja pri oceni teorijskog momenta odgovarajućim uzoračkim momentom:

$$P\{|A_{nk} - \alpha_k| < t\} = 2\Phi\left(t\sqrt{\frac{n}{\alpha_{2k}-\alpha_k^2}}\right)\,.$$

**Postupak metoda momenata**

Neka je data familija dopustivih raspodela $\{f(x;\boldsymbol{\theta}), \boldsymbol{\theta} = (\theta_1,\ldots,\theta_r) \in \boldsymbol{\Theta}\}$ obeležja $X$ i uzorak $(X_1,\ldots,X_n)$ iz populacije sa ovim obeležjem. Treba oceniti $\theta_i$, $i = 1,\ldots,r$.

Postojanje momenata $\alpha_k$, $k = 1,\ldots,r$, gde je $\alpha_k = \alpha_k(\boldsymbol{\theta})$, je samo potreban uslov za primenu metoda momenata. Sam postupak se sastoji u sledećem.

Neka su $a_k$ realizovane vrednosti uzoračkih momenata $A_{nk}$. Izjednačavaćemo teorijske momente sa realizovanim vrednostima uzoračkih momenata, tj.

$$\alpha_k(\theta_1,\ldots,\theta_r) = \alpha_k(\boldsymbol{\theta}) = a_k\,, \quad k = 1,\ldots,r\,.$$

Na taj način se dobija sistem od $r$ jednačina po nepoznatim parametrima $\theta_1,\ldots,\theta_r$. Ako je ovaj sistem rešiv, dobićemo tačkaste ocene parametara. Dovoljno je da korespondencija bude obostrano jednoznačna izmedju parametarskih promenljivih $\theta_1,\ldots,\theta_r$ i vrednosti statistika, $a_1,\ldots,a_r$, tj. da postoje takve funkcije $\varphi_i$ pomoću kojih možemo da nadjemo jedinstveno rešenje sistema u obliku:

$$\theta_i = \varphi_i(a_1,\ldots,a_r)\,, \quad i = 1,\ldots,r\,.$$

Zamenom realizovanih vrednosti statistika na mestu argumenata funkcija $\varphi_i$ samim statistikama, dobićemo statistike, koje ćemo zvati statistikama (ocenama) metoda momenata:

$$\widetilde{\theta}_i = \varphi_i(A_{n1},\ldots,A_{nr})\,.$$

Ako su $\varphi_i$ neprekidne funkcije tada će ocene parametara $\theta_i$ dobijene metodom momenata biti postojane.

**Primer 42.** (Gama raspodela)

Neka obeležje X ima gustinu raspodele

$$\{f(x;\boldsymbol{\theta}), \boldsymbol{\theta} = (\theta_1, \theta_2) \in \Theta = (0, +\infty) \times (0, +\infty)\},$$

gde je

$$f(x;\boldsymbol{\theta}) = \begin{cases} \frac{x^{\theta_2 - 1} e^{-\frac{x}{\theta_1}}}{\theta_1^{\theta_2} \Gamma(\theta_2)}, & x > 0 \\ 0, & x \leq 0 \end{cases} .$$

Momenti gama raspodele su

$$\alpha_k = \int_0^{+\infty} \frac{x^{\theta_2 + k - 1} e^{-\frac{x}{\theta_1}}}{\theta_1^{\theta_2} \Gamma(\theta_2)} dx = \theta_1^k \theta_2 (\theta_2 + 1) \cdots (\theta_2 + k - 1), \; k = 1, 2, \ldots \quad .$$

Za $k = 1, 2$ dobija se sistem koji treba rešiti po $\theta_1$ i $\theta_2$:

$$\alpha_1 = \theta_1 \theta_2 \quad , \quad \alpha_2 = \theta_1^2 \theta_2 (\theta_2 + 1).$$

Rešenje sistema je:

$$\theta_1 = \frac{\alpha_2 - \alpha_1^2}{\alpha_1} \quad , \quad \theta_2 = \frac{\alpha_1^2}{\alpha_2 - \alpha_1^2},$$

odakle slede ocene metodom momenata

$$\widetilde{\theta}_1(X) = \frac{A_{n2} - A_{n1}^2}{A_{n1}}, \quad \text{tj.} \quad \widetilde{\theta}_1(X) = \frac{\overline{S}_n^2}{\overline{X}_n}$$

i

$$\widetilde{\theta}_2(X) = \frac{A_{n1}^2}{A_{n2} - A_{n1}^2}, \quad \text{odnodsno}, \quad \widetilde{\theta}_2(X) = \frac{\overline{X}_n^2}{\overline{S}_n^2} . \triangle$$

## 3.6 Statistike poretka

S obzirom na to da su statistike poretka osnov za dobijanje ocena izvesnog broja parametara raspodele obeležja, zadržaćemo se malo detaljnije na problemu odredjivanja raspodela statistika poretka, kao i na nekim konkretnim primenama.

Neka je $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ prost slučajni uzorak iz populacije sa obeležjem $X$ čija je funkcija raspodele $F$. Kao što smo rekli, niz $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$, gde je $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$ je niz statistika poretka datog uzorka.

Označimo sa $F_k$ funkciju raspodele statistike poretka reda $k$, $k = 1, 2, \ldots, n$. Tada je za svako $x \in R$

$$\begin{aligned} F_n(x) &= P\{X_{(n)} \leq x\} = P(\cap_{i=1}^n \{X_{(i)} \leq x\}) = P(\cap_{i=1}^n \{X_i \leq x\}) = \\ &= \prod_{i=1}^n P\{X_i \leq x\} = \prod_{i=1}^n F(x) = (F(x))^n . \end{aligned}$$

Takodje,

$$F_1(x) = P\{X_{(1)} \leq x\} = 1 - P\{X_{(1)} > x\} = 1 - \prod_{i=1}^n (1 - F(x)) = 1 - (1 - F(x))^n$$

ili uopšte:

$$\begin{aligned} F_k(x) = P\{X_{(k)} \le x\} &= P\{\text{bar } k \text{ elemenata u uzorku je manje ili jednako } x\} = \\ &= \sum_{i=k}^{n} \binom{n}{i} (F(x))^i (1-F(x))^{n-i}. \end{aligned}$$

Na osnovu ovako dobijenih funkcija raspodele relativno je jednostavno odrediti gustine raspodela. Mi ćemo se zadržati samo na gustinama za obeležje apsolutno neprekidnog tipa.

Zbog jednostavnosti, preći ćemo na oznake $Y_k \equiv X_{(k)}$.

**Teorema 3.6.1** *Neka je $Y_1 \le Y_2 \le \ldots \le Y_n$ varijacioni niz prostog uzorka $(X_1, \ldots, X_n)$ obeležja $X$ čija je gustina raspodele $f(x)$, neprekidna i strogo pozitivna za $x \in (a,b)$. Tada je zajednička gustina raspodele vektora $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$:*

$$g(y_1, \ldots, y_n) = \begin{cases} n! f(y_1) f(y_2) \ldots f(y_n), & a < y_1 < y_2 < \ldots < y_n < b \\ 0, & \textit{inače} \end{cases} .$$

Dokaz nećemo izvoditi u opštem slučaju, već samo ilustraciju dokaza na primeru obima uzorka $n = 3$.

Dakle, neka je dat prost slučajni uzorak $(X_1, X_2, X_3)$. Uredivši ga dobijamo $(X_{(1)}, X_{(2)}, X_{(3)})$, odnosno statistike $Y_1 \le Y_2 \le Y_3$ gde je:

$$Y_1 = X_{(1)}\,, \quad Y_2 = X_{(2)}\,, \quad Y_3 = X_{(3)}\,.$$

Kako je

$$P\{a < X_1 = X_2 < b, a < X_3 < b\} = \int_a^b \int_a^b \int_{x_2}^{x_2} f(x_1) f(x_2) f(x_3) dx_1 dx_2 dx_3 = 0\,,$$

posmatraćemo samo slučaj stroge nejednakosti.

Zajednička gustina raspodele za vektor $(X_1, X_2, X_3)$ je:

$$f(x_1, x_2, x_3) = \begin{cases} f(x_1) f(x_2) f(x_3), & a < x_i < b,\ i = 1,2,3 \\ 0, & \text{inače}\,. \end{cases}$$

Posmatrajmo sledeće skupove:

$$A_1 = \{(x_1, x_2, x_3) : a < x_1 < x_2 < x_3 < b\}$$

$$A_2 = \{(x_1, x_2, x_3) : a < x_1 < x_3 < x_2 < b\}$$

$$A_3 = \{(x_1, x_2, x_3) : a < x_3 < x_1 < x_2 < b\}$$

$$A_4 = \{(x_1, x_2, x_3) : a < x_3 < x_2 < x_1 < b\}$$

$$A_5 = \{(x_1, x_2, x_3) : a < x_2 < x_1 < x_3 < b\}$$

$$A_6 = \{(x_1, x_2, x_3) : a < x_2 < x_3 < x_1 < b\}.$$

Neka je $y_1 = \min\{x_1, x_2, x_3\}$, $y_3 = \max\{x_1, x_2, x_3\}$. Svaka od oblasti $A_i, i = 1, \ldots, 6$ se može obostrano jednoznačno da preslika u oblast $B = \{(y_1, y_2, y_3) : a < y_1 < y_2 < y_3 < b\}$ i za svako takvo preslikavanje odredimo Jakobijan:

$$J_k = \left| \frac{\partial x_i}{\partial y_j} \right|_{i,j=1,2,3} \quad , \quad k = 1, 2, \ldots, 6 \, .$$

Apsolutna vrednost svakog od ovih Jakobijana iznosi 1 tj.

$$J_1 = \begin{vmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{vmatrix} = 1, \ldots \qquad J_5 = \begin{vmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{vmatrix} = -1, \ldots$$

Tako je zajednička gustina raspodele statistika $Y_1, Y_2, Y_3$ jednaka:

$$\begin{aligned} g(y_1, y_2, y_3) &= \begin{cases} |J_1| f(y_1) f(y_2) f(y_3) + \cdots + |J_6| f(y_1) f(y_2) f(y_3), & a < y_1 < y_2 < y_3 < b \\ 0, & \text{u ostalim slučajevima} \, , \end{cases} \\ &= \begin{cases} (3!) f(y_1) f(y_2) f(y_3), & a < y_1 < y_2 < y_3 < b \\ 0, & \text{u ostalim slučajevima} \end{cases} \end{aligned}$$

što je i trebalo dokazati.

Zaključujemo da su statistike poretka, za razliku od slučajnih promenljivih iz prostog slučajnog uzorka iz koga su nastale, stohastički zavisne.

Nadjimo sada marginalne gustine statistika poretka takodje u slučaju obeležja $X$ apsolutno neprekidnog tipa. Neka je gustina raspodele obeležja $X$ označena sa $f$ i neka je ona strogo pozitivna na intervalu $(a, b)$, za neke $a, b \in \overline{R}$ i $a < b$, tj. $f(x) > 0$ za $x \in (a, b)$. Tada je

$$F(x) = \begin{cases} 0, & x \le a \\ \int_a^x f(w) dw, & a < x < b \\ 1, & x \ge b \quad . \end{cases}$$

Očigledno za $a < x < b$ je:

$$1 - F(x) = F(b) - F(x) = \int_x^b f(w) dw.$$

Dakle,

$$g_n(y_n) = \int_a^{y_n} dy_{n-1} \int_a^{y_{n-1}} dy_{n-2} \ldots \int_a^{y_2} (n!) f(y_1) f(y_2) \ldots f(y_n) dy_1 \, , \; a < y_n < b \, .$$

Kako je

$$\int_a^{y_2} f(y_1) dy_1 = F(y_2) \, ,$$

dobijamo

$$\int_a^{y_3} F(y_2) f(y_2) dy_2 = \frac{(F(y_3))^2}{2} \, ,$$

jer je $F(a) = 0$. Tada je:

$$g_n(y_n) = n! \frac{(F(y_n))^{n-1}}{(n-1)!} f(y_n) \, ,$$

odnosno

$$g_n(y_n) = \begin{cases} n(F(y_n))^{n-1} f(y_n), & a < y_n < b \\ 0, & \text{inače} \end{cases}$$

i slično

$$g_1(y_1) = \begin{cases} n(1 - F(y_1))^{n-1} f(y_1), & a < y_1 < b \\ 0, & \text{inače}\,. \end{cases}$$

Konačno, za proizvoljno $y_j, j = 1, \ldots, n$ uočimo da je

$$\int_a^x (F(w))^{\alpha-1} f(w)dw = \frac{(F(x))^\alpha}{\alpha} \quad , \quad \alpha > 0$$

i

$$\int_y^b (1 - F(w))^{\beta-1} f(w)dw = \frac{(1 - F(y))^\beta}{\beta} \quad , \quad \beta > 0$$

pa se dobija

$$g_j(y_j) = \begin{cases} \frac{n!}{(n-j)!(j-1)!}(1 - F(y_j))^{n-j}(F(y_j))^{j-1} f(y_j), & a < y_j < b \\ 0, & \text{inače}\,. \end{cases}$$

Marginalnu gustinu vektora $(Y_i, Y_j)$, $Y_i \leq Y_j$, za bilo koje $1 \leq i < j \leq n$ izračunavamo na sledeći način:

$$g_{ij}(y_i, y_j) = \int_a^{y_i} dy_{i-1} \int_a^{y_{i-1}} dy_{i-2} \ldots \int_a^{y_2} dy_1 \int_{y_i}^{y_j} dy_{i+1} \int_{y_{i+1}}^{y_j} dy_{i+2} \ldots$$
$$\ldots \int_{y_{j-2}}^{y_j} dy_{j-1} \int_{y_j}^{b} dy_{j+1} \int_{y_{j+1}}^{b} dy_{j+2} \ldots \int_{y_{n-1}}^{b} (n!) f(y_1) \ldots f(y_n) dy_n\,.$$

Kako je za $\gamma > 0$

$$\int_x^y (F(y) - F(w))^{\gamma-1} f(w)dw = -\frac{(F(y) - F(w))^\gamma}{\gamma}\Big|_x^y = \frac{(F(y) - F(x))^\gamma}{\gamma}\,,$$

to je konačno marginalna gustina vektora $(Y_i, Y_j)$ data sa

$$g_{ij}(y_i, y_j) =$$

$$= \begin{cases} \frac{n!}{(i-1)!(j-i-1)!(n-j)!}(F(y_i))^{i-1}(F(y_j) - F(y_i))^{j-i-1}(1 - F(y_j))^{n-j} f(y_i) f(y_j), \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad a < y_i < y_j < b \\ 0, \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{inače}\,. \end{cases}$$

Statistike poretka su osnov za ocenjivanje nekih važnih parametara obeležja, kao što je, recimo, medijana ili uopšte kvantil reda $p$.

### 3.6.1 Primena statistika poretka u odredjivanju tačkastih ocena za kvantile

Navedimo, najpre, definiciju kvantila. U literaturi se, uglavnom, mogu naći sledeće definicije ovog pojma:

DEFINICIJA 29. *Kvantil reda p* , u oznaci $M_p$, $0 < p < 1$, slučajne promenljive $X$ sa funkcijom raspodele $F$, je

$$M_p = \sup\{x \in R \,|\, F(x) < p\}\,.$$

DEFINICIJA 30. *Kvantil reda p*, u oznaci $M_p$, $0 < p < 1$, slučajne promenljive $X$ sa funkcijom raspodele $F$, je bilo koje rešenje jednačine $F(x) = p \quad , \quad 0 < p < 1$ , po $x \in R$ ili rešenje sistema nejednačina

$$P\{X < x\} \le p \le P\{X \le x\} \qquad , \qquad 0 < p < 1\,,$$

po $x$, ukoliko rešenje pomenute jednačine ne postoji.

Ako je obeležje X apsolutno neprekidnog tipa sa strogo rastućom funkcijom raspodele $F$ za svako $x \in R$ za koje je $0 < F(x) < 1$, obe definicije kvantila reda $p$ daju jedinstvenu vrednost koja se, praktično, dobija rešavanjem jednačine:

$$F(x) = p \quad , \quad 0 < p < 1\,,$$

po $x$.

Kod obeležja diskretnog tipa situacija je složenija utoliko što se prvom od navedenih definicija dolazi do jedinstvene brojne vrednosti koja se naziva kvantilom, dok se drugom definicijom u nekim slučajevima dobija jedinstvena vrednost kao kvantil, a u nekim se dolazi do poluotvorenog intervala čije sve tačke zadovoljavaju definiciju. Dakle, u tom slučaju, kvantil nije jedinstvena brojčana vrednost, nego ceo interval.[2]

Posebno često korišćeni kvantil je kvantil reda $p = 0,5$, koji se zbog svoje važnosti i posebno imenuje. Njegovo ime je *medijana*.

**Primer 43.** Neka je data slučajna promenljiva $X$ sa raspodelom

$$X : \begin{pmatrix} 1 & 2 & 3 \\ 1/4 & 1/4 & 1/2 \end{pmatrix}.$$

Kvantil reda $1/3$ je $M_{1/3} = 2$, jer je $P\{X < 2\} = 1/4 < 1/3$, a $P\{X \le 2\} = 1/2 > 1/3$. Medijana je, medjutim po drugoj definiciji, ceo interval $[2, 3)$. $\triangle$

Kada se red kvantila izražava u procentima, $100p\%$, umesto termina kvantil koristi se termin *percentil*. Za neke kvantile se koriste posebni nazivi vezano za red kvantila. Tako, za $p \in \{0,25; 0,50; 0,75\}$ koristi se naziv *kvartil*, za $p \in \{0,1; 0,2; 0,3; 0,4; 0,5; 0,6; 0,7; 0,8; 0,9\}$ se koristi naziv *decil* i slično.

Pristupimo sada ocenjivanju kvantila na osnovu uzorka. Statistike koje se koriste kao ocene kvantila jednim imenom se zovu *uzorački kvantili*.

U slučaju diskretnog obeležja ili uzorka malog obima $n$, kvantil reda $p$ se ocenjuje odgovarajućim linearnim kombinacijama statistika poretka i to, ako je

[2]U praktičnim primenama se, u takvim slučajevima, može koristiti proizvoljna tačka pomenutog intervala. Najčešće se koristi sredina intervala ili donja granica.

- $(n-1)p$ prirodan broj, kvantil $M_p$ se ocenjuje statistikom

$$\widehat{M}_p = X_{(1+(n-1)p)}\,,$$

- $(n-1)p$ nije prirodan broj, kvantil $M_p$ se ocenjuje statistikom

$$\begin{aligned}\widehat{M}_p &= (k-(n-1)p)X_{(k)} + (1+(n-1)p-k)X_{(k+1)} = \\ &= X_{(k)} + (1+(n-1)p-k)(X_{(k+1)} - X_{(k)})\,,\end{aligned}$$

  gde je $k$ prirodan broj takav da je $k < 1+(n-1)p < k+1$.

Izloženi postupak, zapravo, identifikuje član varijacionog niza (ako takav postoji) koji deli varijacioni niz u odnosu $100p\%$ prema $100(1-p)\%$ po analogiji sa teorijskim značenjem kvantila slučajne promenljive.

O oceni kvantila obeležja apsolutno neprekidnog tipa biće reči u okviru narednog potpoglavlja.

## 3.7 O još nekim primerima tačkastih ocena

Ne ulazeći u to kojim metodom su ocene dobijene, navešćemo neke konkretne primere tačkastih ocena parametara obeležja o kojima do sada nije bilo reči.

Realizovani uzorci iz populacije sa obeležjem apsolutno neprekidnog tipa se, po pravilu, sredjuju intervalno. Već je rečeno da se kod intervalnog sredjivanja uzorka gubi izvesna količina informacija koju uzorak u sebi nosi. U ocenama koje treba definisati za intervalno sredjen uzorak se ne koriste sami elementi uzorka, već reprezentanti intervala, najčešće sredine intervala, kao objekti nad kojima se definišu statistike (dok je uzorak slučajan i intervali su slučajni, pa i njihovi reprezentanti). Tako, na primer, sredina uzorka se odredjuje kao aritmetička sredina srednjeg intervala po kojima se uzorak sredjuje. No, nije uvek opravdano koristiti sredine intervala, kao što će se videti iz nekih niže navedenih primera.

### Ocena kvantila obeležja apsolutno neprekidnog tipa

Uzorački kvantil uzorka obima $n$ za obeležje apsolutno neprekidnog tipa je statistika koja se bazira na granicama intrvala i apsolutnim učestanostima elemenata uzorka u intervalima po kojima se realizovani uzorak sredjuje.

*Uzorački kvantil reda p* je statistika

$$M_p = a_p + h_p \frac{np - \Sigma_p}{N_p}\,,$$

gde je $a_p$ – donja granica kvantilnog intervala, $h_p$ – dužina kvantilnog intervala, $N_p$ – apsolutna učestanost u kvantilnom intervalu, a $\Sigma_p$ – zbirna učestanost do granice $a_p$, tj. do kvantilnog intervala.

*Kvantilni interval* je interval u kome je zbirna apsolutna učestanost prvi put veća ili jednaka $np$.

### Ocene moda

Još jedan od centara grupisanja vrednosti slučajne promenljive je mod slučajne promenljive. Iz praktičnih razloga navodimo ovde, ne samo njegovu ocenu, odnosno statistiku kojom se ocenjuje, već i samu definiciju.

Definicija 31. *Mod* slučajne promenljive je ona vrednost slučajne promenljive za koju gustina raspodele dostiže lokalni maksimum.

Iz definicije sledi da slučajna promenljiva može da ima više modova, pa se prema broju modova koriste nazivi: *unimodalna, bimodalna, trimodalna* i *multimodalna* slučajna promenljiva.

Za mod slučajne promenljive $X$ se najčešće koristi oznaka $M_o(X)$.

Što se tiče ocena moda, razlikuju se postupci ocenjivanja za diskretna i obeležja apsolutno neprekidnog tipa.

Kod diskretnog obeležja, ocena moda (modova) je element varijacionog niza koji ima najveću apsolutnu učestanost u svojoj okolini.

Kod obeležja apsolutno neprekidnog tipa, tj. kod uzorka koji se sredjuje intervalno, najčešće su u upotrebi sledeće statistike:

- $M_o = a_\mu + h_\mu \frac{N_0 - N_1}{(N_0 - N_1) + (N_0 - N_2)}$
- $M_o = a_\mu + h_\mu \frac{N_2}{N_1 + N_2}$

gde je $a_\mu$ – leva granica modalnog intervala, $h_\mu$ – dužina modalnog intervala, $N_0$ – apsolutna učestanost u modalnom intervalu, $N_1$ – apsolutna učestanost u intervalu ispred modalnog i $N_2$ – apsolutna učestanost u intervalu iza modalnog. Modalni interval je interval u kome je apsolutna učestanost elemenata realizovanog uzorka najveća u odnosu na susedne intervale – prethodni i naredni. Ukoliko je prvi interval modalni, onda je $N_1 = 0$, a ukoliko je poslednji interval modalni, tada je $N_2 = 0$.

Primetimo da je u poslednja dva slučaja korišćena ista oznaka za parametar raspodele i njegovu ocenu. To u statističkim razmatranjima nije izuzetak i može se tolerisati kadgod nema opasnosti od zabune.

### Ocene koeficijenta korelacije

Brojne su statistike kojima se ocenjuje koeficijent korelacije obeležja. Ovde ćemo se, medjutim, zadržati samo na jednom od njih, dok se više podataka o ocenjivanju koeficijenta korelacije može naći u knjizi: B. Popović, M. Ristić: "Statistika u psihologiji", Mrlješ, Beograd, 2000.

Statistika koju navodimo je u literaturi poznata pod nazivom *uzoračka* ili *Pirsonov koeficijent korelacije*.

Za ispitivanje linearne povezanosti dva obeležja $X$ i $Y$ na elementima iste populacije uzima se uzorak oblika

$$((X_1, Y_1), \ldots, (X_n, Y_n)), \tag{3.17}$$

pri čemu je obim uzorka $n$ konstantan. Par slučajnih promenljivih $(X_i, Y_i)$ predstavlja vrednost dvodimenzionog obeležja na $i$–tom elementu uzorka.

Sama statistika je već pominjana u prethodnoj glavi ove knjige i definisana je sa

$$R_{X,Y} = \frac{\frac{1}{n}\Sigma_{i=1}^{n}(X_i - \overline{X}_n)(Y_i - \overline{Y}_n)}{\overline{S}_X \overline{S}_Y},$$

gde su $\overline{S}_X$ i $\overline{S}_Y$ uzoračke standardne devijacije za obeležja $X$ i $Y$ redom na osnovu uzorka (3.17).

## 3.8 Oblasti poverenja

Za razliku od tačkastog ocenjivanja parametara, kod kojeg se ocenom smatra statistika, tj. njena realizovana vrednost, oblasti poverenja imaju tu ulogu da se proceni skup, odnosno podskup, odgovarajućeg realnog prostora unutar koga se može smatrati da će se "naći" prava vrednost parametra. U najjednostavnijem slučaju, tj. kod jednodimenzionog parametra, najjednostavnija oblast poverenja biće interval, deo realne prave.

### 3.8.1 Intervali poverenja

Razjasnićemo pitanje intervala poverenja na jednom primeru.

**Primer 44.** Neka je $(X_1, X_2, \ldots, X_n)$ prost slučajni uzorak iz populacije sa obeležjem $X$ čija raspodela pripada familiji dopustivih raspodela $\{\mathcal{N}(m, 4), -\infty < m < \infty\}$. Kao što je poznato, ako je $m$ prava vrednost parametra, slučajna promenljiva

$$Z_0 = \frac{\overline{X}_n - m}{2}\sqrt{n} \quad \text{ima} \quad \mathcal{N}(0,1) \quad \text{raspodelu.}$$

U tom slučaju je za zadato $\alpha$, $0 < \alpha < 1$, odnosno $\gamma = 1 - \alpha$, moguće uvek odrediti broj $z_{\frac{1-\alpha}{2}}$, tako da je

$$P\{|Z_0| \leq z_{\frac{1-\alpha}{2}}\} = 1 - \alpha.$$

Dakle, biće:

$$P\left\{\left|\frac{\overline{X}_n - m}{2}\sqrt{n}\right| \leq z_{\frac{1-\alpha}{2}}\right\} = 1 - \alpha$$

$$P\left\{\left|\frac{\overline{X}_n - m}{2}\sqrt{n}\right| \leq z_{\frac{1-\alpha}{2}}\right\} = P\{\overline{X}_n - z_{\frac{1-\alpha}{2}}\frac{2}{\sqrt{n}} \leq m \leq \overline{X}_n + z_{\frac{1-\alpha}{2}}\frac{2}{\sqrt{n}}\} = 1 - \alpha.$$

Odavde se dobija interval kao skup mogućih vrednosti za parametar $m$:

$$\left[\overline{X}_n - z_{\frac{1-\alpha}{2}}\frac{2}{\sqrt{n}} \quad , \quad \overline{X}_n + z_{\frac{1-\alpha}{2}}\frac{2}{\sqrt{n}}\right].$$

Ovaj interval ćemo zvati interval poverenja za nepoznato matematičko očekivanje normalne raspodele nivoa poverenja $\gamma = 1 - \alpha$.$\triangle$

Posmatrajmo interval dobijen u prethodnom primeru:

$$\left[\overline{X}_n - z_{\frac{1-\alpha}{2}}\frac{2}{\sqrt{n}} \quad , \quad \overline{X}_n + z_{\frac{1-\alpha}{2}}\frac{2}{\sqrt{n}}\right].$$

Njegove granice su slučajne veličine.

DEFINICIJA 32. Ako je bar jedna granica intervala slučajna promenljiva, interval se naziva *slučajni interval*.

Iz definicije slučajnog intervala je jasno da je i njegova dužina $d$ slučajna veličina, pa se može govoriti o očekivanoj dužini, $E(d)$, slučajnog intervala.

Dakle, u gornjem primeru dobijen je jedan slučajni interval, a njegova očekivana dužina je: $E(d) = \frac{4z_{\frac{1-\alpha}{2}}}{\sqrt{n}}$.

U ovom primeru je dužina intervala neslučajna. Medjutim, u sledećem primeru je ona slučajna.

**Primer 45.** Neka je $X : \mathcal{N}(0, \sigma^2), \sigma^2 > 0$. Kolika je verovatnoća da slučajni interval $(|X|, |10X|)$ sadrži tačku $\sigma$? Koja je očekivana dužina ovog intervala?

Odgovor je jednostavan s obzirom na činjenicu da $\frac{X}{\sigma}$ ima $\mathcal{N}(0,1)$ raspodelu i niz dogadjaja jednakih verovatnoća (ekvivalentni dogadjaji):

$$P\{|X| < \sigma < 10|X|\} = P\left\{\frac{\sigma}{10} < |X| < \sigma\right\} = P\left\{\frac{\sigma}{10} < X < \sigma \vee -\sigma < X < -\frac{\sigma}{10}\right\}.$$

Dakle, verovatnoća da slučajni interval sadrži tačku $\sigma$ jednaka je:

$$P\left\{\frac{\sigma}{10} < |X| < \sigma\right\} = P\left\{\frac{\sigma}{10} < X < \sigma\right\} + P\left\{-\sigma < X < -\frac{\sigma}{10}\right\} =$$

$$= 2P\left\{\frac{\sigma}{10} < X < \sigma\right\} = 2P\left\{\frac{1}{10} < \frac{X}{\sigma} < 1\right\} = 2\Phi(1) - 2\Phi(0,1) = 0,60\,.$$

Dužina posmatranog slučajnog intervala je $10|X| - |X| = 9|X|$. Kako je

$$E|X| = 2\int_0^\infty \frac{x}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} dx = \sigma\sqrt{\frac{2}{\pi}}\,,$$

Očekivana dužina ovog intervala je $9E|X| \sim 7,2\sigma$. $\triangle$

**Primer 46.** Pokazali smo da sredina uzorka obima $n$, $\overline{X}_n$, iz populacije sa obeležjem $X$ čija raspodela pripada familiji dopustivih raspodela $\{\mathcal{N}(m, \sigma^2),\ -\infty < m < +\infty,\ \sigma^2 > 0\}$ ima $\mathcal{N}(m, \frac{\sigma^2}{n})$ raspodelu. Kolika je verovatnoća da slučajni interval

$$\left(\overline{X}_n - 2\frac{\sigma}{\sqrt{n}} \quad , \quad \overline{X}_n + 2\frac{\sigma}{\sqrt{n}}\right)$$

sadrži tačku $m$? Kolika je očekivana dužina ovog slučajnog intervala?

Kao i u prethodnom primeru, koristićemo ekvivalentne dogadjaje:

$$P\left\{\overline{X}_n - 2\frac{\sigma}{\sqrt{n}} < m < \overline{X}_n + 2\frac{\sigma}{\sqrt{n}}\right\} = P\left\{-2\frac{\sigma}{\sqrt{n}} < \overline{X}_n - m < 2\frac{\sigma}{\sqrt{n}}\right\} =$$
$$= P\left\{-2 < \frac{\overline{X}_n - m}{\frac{\sigma}{\sqrt{n}}} < 2\right\}.$$

Kako je raspodela slučajne promenljive

$$\frac{\overline{X}_n - m}{\frac{\sigma}{\sqrt{n}}} : \mathcal{N}(0,1)\,,$$

to je tražena verovatnoća $2\Phi(2) = 0,954$ , a očekivana dužina intervala

$$E(d) = 4\frac{\sigma}{\sqrt{n}}\,. \triangle$$

**Primer 47.** Neka je $(X_1, X_2, X_3, X_4)$ prost slučajni uzorak iz populacije sa obeležjem $X$, čija je gustina raspodele

$$f(x) = \begin{cases} \frac{1}{\theta}, & 0 < x < \theta,\ \theta > 0 \\ 0, & \text{inače} \end{cases}$$

tj. obeležje $X$ ima raspodelu koja pripada familiji dopustivih raspodela $\{\mathcal{U}(0,\theta), \theta > 0\}$.

Odredićemo 95%–tni interval poverenja za $\theta$ koristeći statistiku poretka $Y_4$ ovog uzorka. S obzirom da se radi o uniformnoj raspodeli,

$$F(x) = \begin{cases} 0, & x \le 0 \\ \frac{x}{\theta}, & 0 < x < \theta \\ 1, & x \ge \theta \end{cases},$$

odnosno, gustina raspodele statistike poretka maksimalnog reda je

$$g_4(y_4) = \begin{cases} 4(F(y_4))^3 f(y_4), & 0 < y_4 < \theta \\ 0, & \text{inače} \end{cases}$$
$$= \begin{cases} \frac{4}{\theta^4} y_4^3, & 0 < y_4 < \theta \\ 0, & \text{inače}\,. \end{cases}$$

Ako izaberemo realne brojeve $0 < c_1 < c_2 \le 1$ takve da je:

$$0,95 = P\{c_1\theta < Y_4 < c_2\theta\} = \frac{4}{\theta^4}\int_{c_1\theta}^{c_2\theta} y_4^3 dy_4 = \tag{3.18}$$

$$= (c_2 - c_1)(c_2 + c_1)(c_2^2 + c_1^2)$$

dobićemo

$$0,95 = P\left\{\frac{Y_4}{c_2} < \theta < \frac{Y_4}{c_1}\right\}.$$

Dakle, za realizovanu vrednost za $Y_4$, $y_4$, interval

$$\left(\frac{y_4}{c_2}\,,\,\frac{y_4}{c_1}\right)$$

je jedan 95%–tni interval poverenja za $\theta$. Konstante $c_1$ i $c_2$ koje zadovoljavaju uslov (3.18) su, na primer,

$$c_1 = \sqrt[4]{0,05} \quad , \quad c_2 = 1,$$

gde smo izabrali $c_2 = 1$, a onda izračunali $c_1$.$\triangle$

**Primer 48.** Neka je $(X_1, \ldots, X_{10})$ prost uzorak iz populacije sa obeležjem $X$ koje ima raspodelu $\mathcal{N}(m, \sigma^2)$, gde je $m$ poznata karakteristika. Neka je

$$Y = \sum_{i=1}^{10}(X_i - m)^2.$$

Kolika je verovatnoća da slučajni interval $(\frac{Y}{20,5}\,,\,\frac{Y}{3,25})$ sadrži pravu vrednost parametra $\sigma^2$?

S obzirom na činjenicu o raspodeli slučajne promenljive

$$\frac{Y}{\sigma^2} : \chi^2_{10}$$

i činjenice da je

$$P\left\{\frac{Y}{20,5} < \sigma^2 < \frac{Y}{3,25}\right\} = P\left\{3,25 < \frac{Y}{\sigma^2} < 20,5\right\}$$

može se odrediti tražena verovatnoća i ona iznosi 0,95. Dužina posmatranog slučajnog intervala je $d = Y\left(\frac{1}{3,25} - \frac{1}{20,5}\right) = 0,26Y$. S druge strane, kako je

$$E\left(\frac{Y}{\sigma^2}\right) = 10 \quad \text{sledi da je} \quad E(Y) = 10\sigma^2\,,$$

pa je očekivana dužina intervala $2,6\sigma^2$. $\triangle$

Opšti problem koji je rešavan kroz sve prethodne primere je sledeći. Neka je $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ uzorak iz populacije sa obeležjem $X$ čija raspodela pripada familiji dopustivih raspodela $\{f(x;\theta),\, \theta \in \Theta\}$. Treba odrediti interval poverenja za nepoznati parametar $\theta$ na osnovu zadatog uzorka $\mathbf{X}$. Problem odredjivanja intervalne ocene parametra $\theta$ se svodi na to da se odrede dve statistike $Y_1 = u_1(X_1, \ldots, X_n)$ i $Y_2 = u_2(X_1, \ldots, X_n)$ takve da je

$$P\{Y_1 \le Y_2\} = 1$$

i

$$P\{Y_1 \le \theta \le Y_2\} = \gamma\,,$$

gde je $\gamma$ zadata verovatnoća koju zovemo nivo poverenja. Precizno,

DEFINICIJA 33. Neka su $Y_1 = u_1(X_1, \ldots, X_n)$ i $Y_2 = u_2(X_1, \ldots, X_n)$ dve statistike na osnovu istog uzorka $\mathbf{X} = (X_1, \ldots, X_n)$ iz populacije sa obeležjem $X$ čija je familija dopustivih raspodela $\{f(x;\theta),\ \theta \in \Theta\}$, takve da pod uslovom da je $\theta$ prava vrednost parametra (što je nadalje naglašeno u indeksu mere $P$), važi da je:

$$P_\theta\{Y_1 \leq Y_2\} = 1$$

i

$$P_\theta\{Y_1 \leq \theta \leq Y_2\} \geq \gamma, \quad 0 \leq \gamma \leq 1.$$

Tada se interval $[Y_1, Y_2]$ zove *dvostrani interval poverenja* za nepoznati parametar $\theta$ sa *nivoom poverenja* $\gamma$ . Ukoliko je jedna od granica neslučajna veličina, interval će biti *jednostrani interval poverenja*. Oba jednim imenom zovemo *interval poverenja*.

Prirodno je tražiti da interval poverenja bude što uži u tom smislu da očekivanje dužine intervala poverenja, $E(Y_2 - Y_1)$, bude što manje. S duge strane, nivo poverenja treba da bude što veći, i obično se uzima da je $\gamma = 0,95$ ili $\gamma = 0,99$. (Uobičajeno je da se nivo poverenja izražava u procentima kao $100\gamma\%$, što je u prethodnim primerima već korišćeno.) Ova dva zahteva su uglavnom oprečna i ne može im se istovremeno udovoljiti, a izlaz se donekle nalazi u povećanju obima uzorka, mada se to ne može uzeti kao pravilo.

Interval poverenja je po definiciji slučajni interval. Medjutim, kada se eksperiment obavi i na osnovu realizovanog uzorka dobiju realizovane vrednosti statistika koje su granice intervala, dobija se realizovani interval poverenja koji je interval na realnoj pravoj. I za ovako dobijeni interval koristi se naziv *interval poverenja (nivoa poverenja $\gamma$)* bez opasnosti od zabune. Važno je, medjutim, pravilno razumevanje nivoa poverenja.

Osvrnimo se na tumačenje verovatnoće $\gamma$.

Pogrešno je tumačiti da sa verovatnoćom $\gamma$ realizovani interval poverenja sadrži parametar $\theta$, već je verovatnoća $\gamma$ samo verovatnoća da slučajni interval $[Y_1\,,\, Y_2]$ prekrije nepoznatu pravu vrednost parametra $\theta$. Verovatnoću $\gamma$ možemo interpretirati i ovako: zamislimo da smo "uzeli" 100 realizovanih uzoraka istog obima $n$ i dobili nizove brojeva

$$(x_{11}, \ldots, x_{1n}), (x_{21}, \ldots, x_{2n}), \ldots, (x_{100;1}, \ldots, x_{100;n})\,,$$

a zatim na osnovu njih izračunali intervale poverenja

$$[y_{11}\,,\, y_{12}]\,, \quad [y_{21}\,,\, y_{22}]\,, \ \ldots\,, \quad [y_{100;1}\,,\, y_{100;n}]\,.$$

Tada na te intervale možemo gledati kao na realizacije slučajnog intervala $[Y_1, Y_2]$. Kako je $P\{Y_1 \leq \theta \leq Y_2\} = \gamma$ i tumačeći verovatnoću kao graničnu vrednost relativnih učestanosti, možemo reći da približno $100\gamma\%$ realizovanih intervala prekriva nepoznati parametar $\theta$, a ostalih $100(1-\gamma)\%$ realizovanih intervala ga ne prekriva.

Predjimo sada na opšti postupak dobijanja intervala poverenja za proizvoljni nepoznati jednodimenzioni parametar raspodele.

Dakle, neka je obeležje $X$ sa raspodelom koja pripada familiji dopustivih raspodela $\{f(x;\theta), \theta \in \Theta\}$. Pretpostavićemo još da je $\Theta$ neki interval u $R$ (inače intervalno ocenjivanje ne bi imalo nikavog smisla). Uočimo, takodje da je uzorački prostor $\mathcal{X} \subset R^n$. Izaberimo realan broj $\gamma$, $0 \leq \gamma \leq 1$. Polazeći od uzorka $\mathbf{X} = (X_1, X_2, \ldots, X_n)$, tražićemo funkcije $u_1(\mathbf{x})$ i $u_2(\mathbf{x})$, takve da je $u_1 \leq u_2$ za svako $\mathbf{x} \in \mathcal{X}$ i da je za svako $\theta \in \Theta$:

1) skup $\{\mathbf{x} : \theta \in [u_1(\mathbf{x}), u_2(\mathbf{x})]\}$ merljiv,
2) $P_\theta\{\omega : \theta \in [u_1(\mathbf{X}), u_2(\mathbf{X})]\} \geq \gamma$.

DEFINICIJA 34. Pretpostavimo da je $\{f(x;\theta), \theta \in \Theta\}$ familija dopustivih raspodela obeležja $X$, $\mathbf{X} = (X_1, \ldots, X_n)$ uzorak obima $n$ iz populacije sa obeležjem $X$ i $T : \mathcal{X} \times \Theta \longrightarrow R$ funkcija koja zadovoljava sledeća dva uslova:
- Raspodela verovatnoća slučajne veličine $T(X_1, \ldots, X_n; \theta)$ **ne zavisi** od parametra $\theta$.
- Za svako $\mathbf{x} = (x_1, \ldots, x_n) \in \mathcal{X}$ funkcija $T(x_1, \ldots, x_n; \theta)$ je **neprekidna** i **strogo monotona** funkcija argumenta $\theta$. Pri tome karakter monotonosti ne sme da zavisi od $\mathbf{x}$.
Tada se slučajna veličina $T(X_1, \ldots, X_n; \theta)$ zove *centralna statistika* za parametar $\theta$.

Dokazaćemo sledeću teoremu:

**Teorema 3.8.1** *Neka postoji centralna statistika $T : \mathcal{X} \times \Theta \longrightarrow R$ koja je za svako fiksirano $\theta \in \Theta$ merljiva funkcija na skupu $\mathcal{X}$. Označimo sa $A$ kodomen funkcije $T$ tj. $A = T(\mathcal{X} \times \Theta) \subset R$ i još pretpostavimo da je za svako $a \in A$ i za svako $\mathbf{x} \in \mathcal{X}$ rešiva po $\theta$ jednačina $a = T(\mathbf{x}, \theta)$. Kada je $\theta$ fiksirano, koristićemo oznaku $T(\mathbf{x}, \theta) \equiv T_\theta(\mathbf{x})$. Tada je funkcija $T_\theta(\mathbf{X})$ slučajna promenljiva čija raspodela ne zavisi od $\theta$ i na osnovu nje je uvek moguće naći interval poverenja za parametar $\theta$.*

**Dokaz.** Pod uslovima navedenim u teoremi, za svako fiksirano $\gamma \in [0, 1]$ moguće je odrediti dva realna broja $t_1(\gamma)$ i $t_2(\gamma)$ takva da za svako $\theta \in \Theta$ važi da je:

$$P_\theta\{t_1(\gamma) \leq T_\theta(\mathbf{X}) \leq t_2(\gamma)\} \geq \gamma\,,$$

jer je u pitanju strogo monotona funkcija po $\theta$ pa sigurno možemo naći $t_1(\gamma)$ i $t_2(\gamma)$, s tim što to ne moraju da budu jedinstveni realni brojevi.

Pošto je funkcija $T$ monotona po $\theta$ za fiksirano $\mathbf{x}$, rešićemo jednačine po parametru $\theta$ i dobićemo jedinstvena rešenja koja su granice intervala.

Znači, imamo sledeći sistem jednačina i ograničenja

$$t_1(\gamma) = T(\mathbf{x}, \theta) \tag{3.19}$$

$$t_2(\gamma) = T(\mathbf{x}, \theta) \tag{3.20}$$

$$t_1(\gamma) \leq t_2(\gamma)\,.$$

Označimo rešenja ovog sistema sa $T_{1\theta}(\mathbf{x}) = u_1(\mathbf{x})$ i $T_{2\theta}(\mathbf{x}) = u_2(\mathbf{x})$. Dakle, rešavanjem ovog sistema po $\theta$ dobijamo granice intervala, i ako umesto fiksiranog uzmemo slučajni vektor $\mathbf{X}$, dobićemo slučajni interval i to:

$$[T_{1\theta}(\mathbf{X}), T_{2\theta}(\mathbf{X})]\,.\square$$

Pokazaćemo kako se primenom centralne statistike u slučaju raspodele apsolutno neprekidnog tipa može efektivno odrediti interval poverenja za nepoznati parametar $\theta$.

Neka je $g(t)$ gustina raspodele slučajne promenljive $T = T(X_1, \ldots, X_n; \theta)$, koja ne zavisi od parametra $\theta$. Neka je $\gamma \in (0, 1)$ zadati nivo poverenja. Kada je $T$ apsolutno neprekidnog tipa, iz uslova:

$$\gamma = P_\theta\{t_1 \leq T(X_1, \ldots, X_n; \theta) \leq t_2\} = \int_{t_1}^{t_2} g(t)dt$$

odredjujemo konstante $t_1$ i $t_2$. Te konstante nisu jednoznačno odredjene. Fiksirajmo dve konstante za koje je

$$\gamma = \int_{t_1}^{t_2} g(t)dt.$$

Za tako fiksirane konstante $t_1$ i $t_2$ i fiksirano $(x_1, \ldots, x_n)$ odredimo rešenja jednačina: $T(x_1, \ldots, x_n; \theta) = t_1$ i $T(x_1, \ldots, x_n; \theta) = t_2$ po $\theta$ koja su jedinstvena zbog monotonosti funkcije T. Označimo rešenja sa $y_1 = u_1(x_1, \ldots, x_n)$, $y_2 = u_2(x_1, \ldots, x_n)$ i odgovarajuće statistike su $Y_1 = u_1(X_1, \ldots, X_n)$, $Y_2 = u_2(X_1, \ldots, X_n)$. Tada će za njih važiti

$$P_\theta\{Y_1 < \theta < Y_2\} = \gamma.$$

Primetimo da uključivanje granica intervala u sam interval nema značaja kod obeležja apsolutno neprekidnog tipa.

Za diskretan slučaj postupak je analogan, medjutim, rešenje se dobija iz nejednačine:

$$P_\theta\{c_1 \leq T(X_1, \ldots, X_n; \theta) \leq c_2\} \geq \gamma.$$

Ako pretpostavimo da je $T$ monotono opadajuća funkcija (bez smanjenja opštosti) po $\theta$ i označimo sa $T_1(\mathbf{x}, \gamma)$ rešenje jednačine (3.20) , a sa $T_2(\mathbf{x}, \gamma)$ rešenje jednačine (3.19), očigledno $T_2(\mathbf{x}, \gamma) \geq T_1(\mathbf{x}, \gamma)$.

Ako za neko $\theta$ pri fiksiranom $\mathbf{x}$ važi nejednakost:

$$t_1(\gamma) \leq T(\mathbf{x}, \gamma) \leq t_2(\gamma) \tag{3.21}$$

tada je i

$$T_1(\mathbf{x}, \gamma) \leq \theta \leq T_2(\mathbf{x}, \gamma)\,. \tag{3.22}$$

Obrnuto, svako $\mathbf{x}$ koje zadovoljava relaciju (3.22), zadovoljava takodje i relaciju (3.21), pa je

$$\{\mathbf{x} : t_1(\gamma) \leq T(\mathbf{x}, \theta) \leq t_2(\gamma)\} = \{\mathbf{x} : \theta \in [T_1(\mathbf{x}, \gamma), T_2(\mathbf{x}, \gamma)]\}.$$

Tada je $[T_1(\mathbf{x}, \gamma), T_2(\mathbf{x}, \gamma)]$ tj. $[T_1(\mathbf{X}, \gamma), T_2(\mathbf{X}, \gamma)]$ interval poverenja sa nivoom poverenja $\gamma$.

Zadržaćemo se sada na specijalnim intervalima poverenja od kojih smo neke već razmatrali kroz prethodne primere.

**Ocena za matematičko očekivanje** $m$ kod normalne raspodele $\mathcal{N}(m, \sigma^2)$ kada je $\sigma^2$ **poznato**:

Interval poverenja za parametar $m$ imali smo u prvom primeru i on iznosi:

$$I_m = \left[ \overline{X}_n - z_{\frac{1-\alpha}{2}} \frac{\sigma}{\sqrt{n}} \quad , \quad \overline{X}_n + z_{\frac{1-\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right].$$

**Ocena za matematičko očekivanje** $m$ kod normalne raspodele $\mathcal{N}(m, \sigma^2)$ kada $\sigma^2$ **nije poznato**:

Najpre treba oceniti $\sigma^2$ pa koristimo sledeću centralnu statistiku koja ima $\chi^2$ raspodelu sa $(n-1)$ stepeni slobode:

$$\frac{n\overline{S}_n{}^2}{\sigma^2} = \chi^2_{n-1} \quad , \quad \overline{S}_n{}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \quad .$$

Definišimo slučajnu promenljivu:

$$\frac{Z^*}{\sqrt{\frac{\chi^2_{n-1}}{n-1}}} = t_{n-1}\,, \quad \text{gde je} \quad Z^* = \frac{\overline{X}_n - m}{\sqrt{\frac{\sigma^2}{n}}}$$

sa normalnom normiranom raspodelom, pa statistika $t_{n-1}$ ima studentovu raspodelu sa $(n-1)$ stepeni slobode, jer su $Z^*$ i $\overline{S}_n{}^2$ nezavisne slučajne promenljive. Otuda

$$\frac{\overline{X}_n - m}{\sqrt{\overline{S}^2_n}}\sqrt{n-1} = t_{n-1}.$$

Ovo će biti tražena centralna statistika za parametar $m$ na osnovu koje ćemo odrediti interval poverenja:

$$I_m = \left[ \overline{X}_n - t_{n-1;\frac{1-\alpha}{2}} \sqrt{\frac{\overline{S}^2_n}{n-1}} \quad , \quad \overline{X}_n + t_{n-1;\frac{1-\alpha}{2}} \sqrt{\frac{\overline{S}^2_n}{n-1}} \right]$$

sa nivoom poverenja $1-\alpha$, gde je broj $t_{n-1;\frac{1-\alpha}{2}}$ odredjen iz uslova

$$P\left\{ |t_{n-1}| \le t_{n-1;\frac{1-\alpha}{2}} \right\} = 1 - \alpha\,.$$

**Interval poverenja za razliku matematičkih očekivanja** $m_1 - m_2$ dva nezavisna obeležja sa normalnim raspodelama i jednakim disperzijama kada je $\sigma^2$ **nepoznato**[3]:

Neka su data dva nezavisna obeležja sa normalnom raspodelom i jednakim disperzijama $X : \mathcal{N}(m_1, \sigma^2)$ i $Y : \mathcal{N}(m_2, \sigma^2)$. Njihova razlika, kao što je poznato, ima takodje normalnu raspodelu. Koristićemo ovo svojstvo i na osnovu dva nezavisna uzorka, po jedan iz svake od ovih raspodela, ocenićemo razliku njihovih očekivanja.

Neka su pomenuti uzorci redom $(X_1, X_2, \ldots, X_{n_1})$ i $(Y_1, Y_2, \ldots, Y_{n_2})$. Uočimo statistiku

$$\overline{X}_{n1} - \overline{Y}_{n2} : \mathcal{N}\left( m_1 - m_2\,,\, \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} \right).$$

[3]Za slučaj poznate disperzije procedura će biti analogna, ali sa odgovarajućim statistikama i njeno sprovodjenje se ostavlja čitaocu.

Posmatrajmo sledeće slučajne promenljive:

$$\frac{\overline{X}_{n1}-\overline{Y}_{n2}-(m_1-m_2)}{\sqrt{\frac{\sigma^2}{n_1}+\frac{\sigma^2}{n_2}}}:\mathcal{N}(0,1)\,,$$

$$\frac{n_1\overline{S}_{n1}{}^2}{\sigma^2}=\chi^2_{n_1-1}\,,$$

$$\frac{n_2\overline{S}_{n2}{}^2}{\sigma^2}=\chi^2_{n_2-1}$$

i

$$\frac{n_1\overline{S}_{n1}{}^2}{\sigma^2}+\frac{n_2\overline{S}_{n2}{}^2}{\sigma^2}=\chi^2_{n_1+n_2-2}\,.$$

Odavde ćemo imati statistiku koja ima studentovu raspodelu sa $n_1+n_2-2$ stepena slobode:

$$\frac{\frac{\overline{X}_{n1}-\overline{Y}_{n2}-(m_1-m_2)}{\sqrt{\frac{\sigma^2}{n_1}+\frac{\sigma^2}{n_2}}}}{\sqrt{\frac{\frac{n_1\overline{S}_{n1}{}^2}{\sigma^2}+\frac{n_2\overline{S}_{n2}{}^2}{\sigma^2}}{n_1+n_2-2}}}=t_{n_1+n_2-2}\,.$$

Na taj način smo izabrali centralnu statistiku za parametar $m_1-m_2$ na osnovu koje ćemo odrediti traženi interval poverenja

$$I_{m_1-m_2}=\left[\overline{X}_{n1}-\overline{Y}_{n2}-t_{n_1+n_2-2;\frac{1-\alpha}{2}}\sqrt{\frac{n_1\overline{S}^2_{n1}+n_2\overline{S}^2_{n2}}{n_1n_2(n_1+n_2-2)}(n_1+n_2)}\quad,\right.$$

$$\left.\overline{X}_{n1}-\overline{Y}_{n2}+t_{n_1+n_2-2;\frac{1-\alpha}{2}}\sqrt{\frac{n_1\overline{S}^2_{n1}+n_2\overline{S}^2_{n2}}{n_1n_2(n_1+n_2-2)}(n_1+n_2)}\right].$$

**Ocena disperzije** $\sigma^2$ kod normalne raspodele kada je $m$ **poznato**:

Neka je data familija dopustivih raspodela $\{\mathcal{N}(m,\sigma^2),\sigma^2>0\}$, gde je očekivanje $m$ poznato. Uočimo slučajnu promenljivu sa $\chi^2$ raspodelom i $n$ stepeni slobode, koja će imati ulogu centralne statistike

$$Y=\sum_{i=1}^{n}\frac{(X_i-m)^2}{\sigma^2}\,.$$

Dalje, ako je $\gamma=1-\alpha$ nivo poverenja, treba odrediti realne brojeve $a$ i $b$, $a<b$, takve da je:

$$P\{a\le Y\le b\}=\gamma=1-\alpha\,.$$

U prethodnim procedurama smo koristili simetričnost raspodele centralne statistike za odredjivanje granica intervala i brojeve $a$ i $b$ smo odredjivali na jedinstven način. S obzirom da $\chi^2$–raspodela ima gustinu koja je asimetrična, pitanje odredjivanja ovih brojeva je interesantno utoliko što se poštuje isti princip u razmišljanju kao i prethodno, te se oni odredjuju tako da važi

$$P\{\chi^2_n<a\}=P\{\chi^2_n>b\}=\frac{\alpha}{2}$$

(mada ne neophodno). Otuda će nadalje biti korišćene oznake za odgovarajuće kvantile $\chi^2$–raspodele.

$$P\left\{\chi^2_{n;\alpha/2} < \sum_{i=1}^{n} \frac{(X_i - m)^2}{\sigma^2} < \chi^2_{n;1-\alpha/2}\right\} = 1 - \alpha\,,$$

tj.

$$P\left\{\frac{\sum_{i=1}^n (X_i - m)^2}{\chi^2_{n;1-\alpha/2}} < \sigma^2 < \frac{\sum_{i=1}^n (X_i - m)^2}{\chi^2_{n;\alpha/2}}\right\} = 1 - \alpha\,.$$

Odavde je, očigledno, interval poverenja za disperziju $\sigma^2$ kada je očekivanje poznato :

$$I_{\sigma^2} = \left[\frac{\sum_{i=1}^n (X_i - m)^2}{\chi^2_{n;1-\alpha/2}} \quad , \quad \frac{\sum_{i=1}^n (X_i - m)^2}{\chi^2_{n;\alpha/2}}\right].$$

Pomoću ovog intervala se može odrediti i odgovarajući interval poverenja za standardnu devijaciju koristeći definiciju standardne devijacije:

$$I_{\sigma} = \left[\sqrt{\frac{\sum_{i=1}^n (X_i - m)^2}{\chi^2_{n;1-\alpha/2}}} \quad , \quad \sqrt{\frac{\sum_{i=1}^n (X_i - m)^2}{\chi^2_{n;\alpha/2}}}\right].$$

**Ocena disperzije** $\sigma^2$ kod normalne raspodele kada je $m$ **nepoznato**:

Polazi se od iste familije $\{\mathcal{N}(m, \sigma^2), \sigma^2 > 0\}$ i koristi statistika koja ima $\chi^2$ raspodelu sa $(n-1)$ stepeni slobode:

$$\frac{n\overline{S}_n{}^2}{\sigma^2} = \chi^2_{n-1}.$$

Ideja je ista kao i u predhodnom slučaju,

$$P\left\{a \le \frac{n\overline{S}_n{}^2}{\sigma^2} \le b\right\} = \gamma = 1 - \alpha,$$

pa kad se obavi neophodno izračunavanje, dobijamo sledeće granice slučajnog intervala:

$$I_{\sigma^2} = \left[\frac{n\overline{S}_n{}^2}{b}\,,\, \frac{n\overline{S}_n{}^2}{a}\right].$$

Pri tome se brojevi $a$ i $b$ odredjuju kao i kod prethodne ocene, tj. iz uslova[4]

$$P\{\chi^2_{n-1} \le a\} = \frac{\alpha}{2} \quad , \quad P\{\chi^2_{n-1} \le b\} = 1 - \frac{\alpha}{2}\,.$$

**Interval poverenja za količnik disperzija** dva nezavisna obeležja sa normalnim raspodelama:

Neka su data dva uzorka $(X_1, \ldots, X_{n_1})$ i $(Y_1, \ldots, Y_{n_2})$ za obeležja redom $X : \{\mathcal{N}(m_X, \sigma^2_X)\,,\, \sigma^2_X > 0\}$ i $Y : \{\mathcal{N}(m_Y, \sigma^2_Y), \sigma^2_Y > 0\}$ pri čemu očekivanja nisu poznata. Ukoliko se

[4]koji nije obavezan, ali je uobičajen

disperzije ovih obeležja uporedjuju medju sobom kroz njihov količnik, ima smisla odredjivanje intervala poverenja za količnik

$$\frac{\sigma_Y^2}{\sigma_X^2}\,.$$

Ocena se dobija na osnovu saznanja da slučajne promenljive imaju raspodele kako sledi

$$\frac{n_1\overline{S}_X^2}{\sigma_X^2} = \chi^2_{n_1-1} \qquad , \qquad \frac{n_2\overline{S}_Y^2}{\sigma_Y^2} = \chi^2_{n_2-1}$$

$$\frac{\frac{n_1\overline{S}_X^2}{\sigma_X^2(n_1-1)}}{\frac{n_2\overline{S}_Y^2}{\sigma_Y^2(n_2-1)}} = F_{n_1-1;n_2-1}\,.$$

Sam interval poverenja će biti:

$$I_{\frac{\sigma_Y^2}{\sigma_X^2}} = \left[ a\frac{\frac{n_2\overline{S}_Y^2}{n_2-1}}{\frac{n_1\overline{S}_X^2}{n_1-1}} \quad , \quad b\frac{\frac{n_2\overline{S}_Y^2}{n_2-1}}{\frac{n_1\overline{S}_X^2}{n_1-1}} \right],$$

za $a$ i $b$ koji zadovoljavaju uslov

$$P\{a \le F_{n_1-1;n_2-1} \le b\} = 1-\alpha$$

i, po pravilu, uslov

$$P\{F_{n_1-1;n_2-1} < a\} = P\{F_{n_1-1;n_2-1} > b\} = \frac{\alpha}{2}\,.$$

*

Svi navedeni intervali poverenja su tzv. dvostrani, a zadržimo se još malo na jednostranim intervalima. Pogotovu kod disperzije i standardne devijacije su u upotrebi, tzv. jednostrani intervali poverenja kod kojih je samo jedna granica slučajna. Ovakav interval se koristi u prilici kada je jedna granica od većeg interesa za istraživanje u kome se primenjuje intervalno ocenjivanje kao statistička procedura. Izložićemo ovaj način razmišljanja na primeru disperzije.

Parametarski prostor za disperziju je po pravilu $\Theta = [0\,, +\infty)$. Jednostrani gornji interval poverenja (gornja granica mu je slučajna) bi se dobio na osnovu razmišljanja

$$P\{0 \le \sigma^2 \le \varphi(X_1, X_2, \ldots, X_n)\} = 1-\alpha\,,$$

a jednostrani donji (donja granica mu je slučajna)

$$P\{\varphi(X_1, X_2, \ldots, X_n) \le \sigma^2 < +\infty\} = 1-\alpha\,.$$

### 3.8.2 Neparametarski intervali poverenja za kvantile

Postoje intervali poverenja čije granice nisu u vezi sa raspodelom obeležja čiji se parametri ocenjuju, odnosno ne zavisi od te raspodele. U tom slučaju se za interval poverenja kaže da je neparametarski. Primer neparametarskih intervala poverenja su intervali poverenja za kvantile.

Intervalna ocena kvantila se, kao što je bio slučaj i sa tačkastom, bazira na statistikama poretka. Ona se odredjuje iz uslova

$$P\{X_{(r)} \leq M_p \leq X_{(s)}\} \geq \gamma \quad , \quad 1 \leq r < s \leq n\,.$$

Pokazaćemo, da ako $X$ ima raspodelu apsolutno neprekidnog tipa, onda slučajni interval $[X_{(r)}, X_{(s)}], r < s$ prekriva pravu vrednost kvantila $M_p$ sa verovatnoćom koja zavisi od $r$, $s$ i $p$, ali ne i od raspodele obeležja $X$. Otuda će on predstavljati neparametarski interval (za razliku od prethodnih primera intervala poverenja koji su zavisili od raspodele obeležja $X$ koja je bila normalna, i sve korišćene statistike su bile zasnovane na toj pretpostavci).

Zaista,

$$\begin{aligned}\{\omega : X_{(r)}(\omega) \leq M_p\} = \\ = \{\omega : X_{(r)}(\omega) \leq M_p \wedge X_{(s)}(\omega) \geq M_p\} \cup \{\omega : X_{(r)}(\omega) \leq M_p \wedge X_{(s)} < M_p\}\,,\end{aligned}$$

te kako je

$$\{\omega : X_{(s)}(\omega) < M_p\} \subset \{\omega : X_{(r)}(\omega) \leq M_p\}\,,$$

sledi da je

$$\begin{aligned}P\{X_{(r)} \leq M_p\} &= P\{X_{(r)} \leq M_p \wedge X_{(s)} \geq M_p\} + P\{X_{(r)} \leq M_p \wedge X_{(s)} < M_p\} = \\ &= P\{X_{(r)} \leq M_p \leq X_{(s)}\} + P\{X_{(s)} < M_p\}\,,\end{aligned}$$

a odavde je (u apsolutno neprekidnom slučaju)

$$P\{X_{(r)} \leq M_p \leq X_{(s)}\} = P\{X_{(r)} \leq M_p\} - P\{X_{(s)} < M_p\} = \sum_{i=r}^{s-1} \binom{n}{i} p^i (1-p)^{n-i},$$

što je i trebalo dokazati.

Ovde je još važno konstatovati da $s$ i $r$ nisu na jedinstven način odredjeni, ali uvodjenjem nekog dodatnog kriterijuma kao na primer kod intervala poverenja čija je centralna statistika imala $\chi^2$–raspodelu, može se prevazići neodredjenost.

**Primer 49.** Neka je

3,1; 3,1; 4,5; 4,6; 4,6; 4,9; 4,9; 5,2; 5,2; 5,3; 5,3; 5,8; 6,0; 6,0; 6,2; 6,7; 7,1; 7,5; 7,9; 8,0; 8,2; 8,4; 8,5; 9,3; 9,7; 9,8; 9,9; 10,0; 10,0; 10,1; 10,4; 10,5; 11,6; 11,6; 11,7; 11,8; 12,4; 12,6; 12,9; 13,4

varijacioni niz realizovanog uzorka obima 40. Odrediti interval poverenja nivoa poverenja 0,90 za kvantil reda $p = 0,25$ (prvi kvartil) na osnovu ovog uzorka.

U varijacionom nizu realizovanog uzorka traže se brojevi $x_{(r)}$ i $x_{(s)}$ koji bi bili granice intervala poverenja za $M_{0,25}$ sa nivoom poverenja $0,9$. Dakle, $n = 40, p = 0,25$ daju

$$\sum_{k=r}^{s-1} \left( \begin{array}{c} 40 \\ k \end{array} \right) \cdot 0,25^k \cdot 0,75^{40-k} = P\{x_{(r)} \leq X \leq x_{(s)}\} = P\{r \leq Y \leq s\} =$$

$$= P\left\{\frac{r-40\cdot 0,25}{\sqrt{40\cdot 0,25\cdot 0,75}} \le \frac{Y-10}{\sqrt{7,5}} \le \frac{s-10}{\sqrt{7,5}}\right\} = P\{c_1 \le Y^* \le c_2\} = 0,90$$

tj. aproksimacijom normalnom raspodelom: $Y^* \sim \mathcal{N}(0,1)$, treba da je

$$\frac{1}{\sqrt{2\pi}}\int_{c_1}^{c_2} e^{-\frac{y^2}{2}}dy = 0,90.$$

Ako se izabere $r = 6$ i $s = 15$ dobija se

$$P\{-1,46 \le Y^* \le 1,83\} = 0,42786 + 0,46562 = 0,89348 \approx 0,9.$$

Dakle, traženi interval ima granice $x_{(6)} = 4,9$ i $x_{(15)} = 6,2$, tj. 90%-tni interval poverenja za nepoznati kvartil reda $0,25$ je:

$$I_{M_{0,25}} = [4,9\,;\,6,2].$$

Jasno da ovaj izbor nije jedinstven, već je samo jedan od mogućih izbora. △

### 3.8.3 Višedimenzione oblasti poverenja

U mnogim praktičnim primerima jednodimenzionog parametra, moguće je konstruisati intervale poverenja, ali je to uvek povezano sa pretpostavkom o monotonosti za centralnu statistiku, kao i pretpostavkom da je $\Theta$ interval. Uslov monotonosti, medjutim, nije uvek moguće ostvariti, a ako $\Theta$ ne bi bio interval, interval poverenja ne bi imao nikakvog smisla.

Oblasti poverenja rešavaju opštiji problem.

Ponovo posmatramo merljiv uzorački prostor $\mathcal{X}$ za populaciju sa obeležjem $X$ čija raspodela pripada familiji dopustivih raspodela $\mathcal{F}_\theta = \{f(x,\theta), \theta \in \mathbf{\Theta}\}$, pri čemu je $\mathbf{\Theta}$ višedimenzioni merljiv skup, tj. podskup odgovarajućeg višedimenzionog realnog prostora, a $\mathbf{X} = (X_1, \ldots, X_n)$ slučajni uzorak.

Neka je $K \subset \mathcal{X} \times \mathbf{\Theta}$, $K_\theta = \{\mathbf{x} : (\mathbf{x},\theta) \in K\}$, $K(\mathbf{x}) = \{\theta : (\mathbf{x},\theta) \in K\}$. Neka je $K_\theta$ merljiv podskup od $\mathcal{X}$ za svako $\theta \in \mathbf{\Theta}$. Takodje pretpostavimo da je za svako $\mathbf{x} \in \mathcal{X}$ $K(\mathbf{x}) \neq \emptyset$, tačnije, da je $\{\mathbf{X} : K(\mathbf{X}) = \emptyset\} \subset L$ za koji važi da za svako $\theta \in \mathbf{\Theta} : P_\theta(L) = 0$. Neka je $\beta$ realan broj, $0 < \beta < 1$, takav da je

$$\beta \le \inf_{\theta\in\mathbf{\Theta}} P_\theta(K_\theta(\mathbf{X}))$$

gde je

$$K_\theta(\mathbf{X}) = \{\omega : \mathbf{X}(\omega) \in K_\theta\}.$$

Tada je

$$P_\theta\{\mathbf{X} \in K_\theta\} \ge \beta,$$

a oblast $K_\theta(\mathbf{X})$ ćemo zvati oblast poverenja za parametar $\theta$ sa nivoom poverenja $\beta$.

**Primer 50.** Za oblast poverenja dvodimenzionog parametra $(m, \sigma^2)$ normalne raspodele nije moguće uzeti pravougaonik

$$\left[\overline{X}_n - \frac{\overline{S}_n}{\sqrt{n-1}}t_{n-1,\beta}, \overline{X}_n + \frac{\overline{S}_n}{\sqrt{n-1}}t_{n-1,\beta}\right] \times \left[\frac{n\overline{S}_n{}^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}, \frac{n\overline{S}_n{}^2}{\chi^2_{n-1,\frac{\alpha}{2}}}\right] \quad ,$$

jer slučajne veličine kojima je pravougaonik definisan $\frac{\overline{X}_n - m}{\overline{S}_n}$ i $\overline{S}_n{}^2$ nisu nezavisne.

Oblast poverenja moguće je bazirati na dvodimenzionoj statistici $(\overline{X}_n, \overline{S}_n{}^2)$ čije su komponente nezavisne.

Uočimo skup:

$$K_\theta = \left\{ \mathbf{x} : \sqrt{\frac{n}{\sigma^2}} |\overline{x}_n - m| \leq t = z_{\frac{1-\alpha_1}{2}} \wedge t_1(\beta_2) = \chi^2_{n-1,\frac{\alpha_2}{2}} \leq \frac{n\overline{s}_n{}^2}{\sigma^2} \leq t_2(\beta_2) = \chi^2_{n-1,1-\frac{\alpha_2}{2}} \right\}.$$

Neka je $\beta = \beta_1\beta_2$ i $\Phi(z_{\frac{1-\alpha_1}{2}}) = \frac{1-\alpha_1}{2}$, gde je $\beta_1 = 1-\alpha_1$ i $\beta_2 = 1-\alpha_2$. Na osnovu nezavisnosti $\overline{X}_n$ i $\overline{S}_n{}^2$ sledi da je:

$$P_\theta\{\mathbf{X} \in K_\theta\} = P_\theta \left\{ \frac{|\overline{X}_n - m|}{\sqrt{\frac{\sigma^2}{n}}} \leq z_{\frac{1-\alpha_1}{2}} \right\} P_\theta \left\{ t_1(\beta_2) \leq \frac{n\overline{S}_n{}^2}{\sigma^2} \leq t_2(\beta_2) \right\} = \beta_1\beta_2 = \beta.$$

Oblast poverenja će biti oblika

$$K_\theta = \{\mathbf{x} : \mathbf{x} = (x_1, x_2)\},$$

ograničena parabolom $x_2 = \frac{(\overline{x}_n - x_1)^2 n}{z_{\frac{1-\alpha_1}{2}}}$, i dvema paralelnim pravama

$$x_2 = \frac{n\overline{s}_n{}^2}{t_2(\beta_2)} \quad , \quad x_2 = \frac{n\overline{s}_n{}^2}{t_1(\beta_2)}.$$

Preciznije

$$K_\theta = \left\{ \mathbf{x} : \mathbf{x} = (x_1, x_2) \wedge x_2 \geq \frac{(\overline{x}_n - x_1)^2 n}{z_{\frac{1-\alpha_1}{2}}} \wedge x_2 \geq \frac{n\overline{s}_n^2}{t_2(\beta_2)} \wedge x_2 \leq \frac{n\overline{s}_n^2}{t_1(\beta_2)} \right\}. \triangle$$

# Glava 4

# Testiranje statističkih hipoteza

Testiranje statističkih hipoteza predstavlja vid statističkog zaključivanja koji se primenjuje u situacijama u kojima se unapred pretpostavlja postojanje odredjene veze medju izučavanim pojavama ili kada se razmatra raspodela obeležja kojom se posmatrana pojava karakteriše. U psihologiji se recimo povezuju: emocije i izraz lica, uticaj prve impresije i tumačenje kasnijih podataka, u kontroli kvaliteta: povezuje se proizvodna smena sa brojem defektnih proizvoda i još mnogi drugi primeri se mogu dati u tom smislu. Sa druge strane, može da se iznese bilo kakva pretpostavka o obeležju koje karakteriše izučavanu pojavu, kao, na primer: visina stanovništva jedne regije (ili u celini) prati normalnu raspodelu i slično. Pretpostavke mogu da se odnose i samo na pojedine karakteristike obeležja kao što je očekivanje, medijana i slično. Svaka od pretpostavki može da bude tačna ili pogrešna.

## 4.1 Osnovni pojmovi

Definicija 35. Tvrdjenje o posmatranim pojavama i procesima na jednoj ili više populacija, koje može da se iskaže kao tvrdjenje o raspodeli jednog ili više obeležja je *statistička hipoteza*.

Drugim rečima, svaka pretpostavka da obeležje $\mathbf{X}$ ima raspodelu koja pripada nekom skupu dopustivih raspodela naziva se statistička hipoteza.

Definicija 36. *Testiranje statističke hipoteze* je postupak provere hipoteze.

Cilj testiranja hipoteze je njeno prihvatanje ili odbacivanje pri čemu se verifikacija obavlja nekim unapred utvrdjenim statističkim metodom.

Hipoteza se testira na osnovu uzorka i donosi se odluka o njenom prihvatanju ili odbacivanju, pri čemu se govori i o verovatnoći pogrešnog zaključka.

**Primer 51.** Ispituju se motorne sposobnosti radnika zaposlenih u dva pogona jedne fabrike. Pri tome treba utvrditi da li postoji razlika u motornoj snazi šake radnika koji pripadaju dvema definisanim grupama. S tim u vezi prirodno je postaviti dve hipoteze:

- Postoje razlike u motornoj snazi šaka radnika u prvom i drugom pogonu, i

- Ne postoje razlike u motornoj snazi šaka radnika u prvom i drugom pogonu. $\triangle$

U ovom slučaju postupak statističkog zaključivanja podrazumeva da se jedna od postavljenih hipoteza uzima za polaznu ili tzv. **nultu hipotezu** i označava se sa $H_0$. Druga hipoteza naziva se u tom slučaju **alternativna hipoteza** i označava sa $H_1$, ili, redje, $H_a$. Koja hipoteza od postavljenih se uzima kao nulta, zavisi od samog problema. Opšti je princip da se za nultu hipotezu po pravilu uzima ona koja se lakše verifikuje, tj. za koju je lakše utvrditi verovatnoće izvodjenja pogrešnih zaključaka.

Hipoteza može biti prosta ili složena. Hipoteza je **prosta** ako u potpunosti odredjuje raspodelu obeležja kojom se bavi, u protivnom je **složena**.

Navodimo nekoliko primera najrasprostranjenijih matematičkih formulacija statističkih hipoteza.

- Hipoteza o obliku raspodele posmatranog obeležja $X$

$$H_0 : F_X(x) = F_0(x)\,, \quad x \in R\,,$$

gde je $F_0$ potpuno odredjena zadata raspodela ili

$$H_0 : F_X \in \mathcal{F}\,,$$

gde je $\mathcal{F}$ zadata familija funkcija raspodele.

- Hipoteza o homogenosti (istovrsnosti) kojom se proverava jednakost raspodela, recimo $k$, slučajnih vektora istih dimenzija

$$H_0 : F_1(\mathbf{x}) = \ldots = F_k(\mathbf{x})\,, \quad \mathbf{x} = (x_1, \ldots, x_n) \in R^n\,,$$

gde je $F_i$ funkcija raspodele vektora $(X_{i1}, \ldots, X_{in}), i = 1, \ldots, k$. Ovakva hipoteza se primenjuje, na primer, u situacijama kada se za više uzoraka istog obima proverava da li su uzeti iz iste populacije.

- Hipoteza o nezavisnosti dva obeležja $X$ i $Y$

$$H_0 : F(x, y) = F_X(x)F_Y(y)\,, \quad (x, y) \in R^2\,,$$

gde je $F(x, y)$ funkcija raspodele slučajnog vektora $(X, Y)$. Ovakva hipoteza može biti iskazana i za više od dva obeležja.

- Hipoteza o slučajnosti se odnosi na višedimenziono obeležje ili bilo koji vektor slučajnih promenljivih

$$\mathbf{X} = (X_1, \ldots, X_n)$$

i njome se testira da li su komponente $X_i$ nezavisne i jendako raspodeljene, tj. da li je moguće razmatrati $\mathbf{X}$ kao prost slučajni uzorak iz raspodele neke slučajne veličine $\xi$

$$H_0 : F_{\mathbf{X}}(\mathbf{x}) = F_\xi(x_1) \ldots F_\xi(x_n)\,, \quad \mathbf{x} = (x_1, \ldots, x_n) \in R^n$$

gde je $F_\xi$ funkcija raspodele slučajne promenljive $\xi$.

Pravilo za testiranje hipoteze $H_0$ je **statistički test**. Pravilo testiranja najčešće koristi neku statistiku. Statistika čijim se posredstvom vrši testiranje zove se **test statistika**.

Statistički testovi kod kojih raspodela test statistike bitno zavisi od raspodele posmatranog obeležja su **parametarski testovi**, a testovi kod kojih raspodela test statistike ne zavisi od raspodele posmatranog obeležja su **neparametarski testovi**.

Parametarski testovi najčešće služe za proveru hipoteze o parametrima posmatranih raspodela, a neparametarski za proveru oblika raspodele, zavisnosti obeležja (dva ili više), slučajnosti niza dogadjaja i slično. Nadalje ćemo se baviti nekim konkretnim testovima iz svake od ovih grupa.

Svaki realizovani uzorak obima $n$, $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ definiše jednu tačku $n$–dimenzionog realnog euklidskog prostora, $(x_1, x_2, \ldots, x_n) \in R^n$. Pri testiranju statističkih hipoteza po pravilu se definiše skup $C \subset R^n$ koji služi kao kriterijum za odbacivanje, odnosno prihvatanje nulte hipoteze i to: ako $(x_1, x_2, \ldots, x_n) \in C$, tada se hipoteza $H_0$ odbacuje, a ako $(x_1, x_2, \ldots, x_n) \in C^c$, tada nema razloga da se $H_0$ odbaci.

Skup svih tačaka $C \subset R^n$ za koje se $H_0$ odbacuje je **kritična oblast testa**. Kritična oblast se najčešće iskazuje preko kritične vrednosti test statistike. Tako, ako je test statistika jednodimenziona funkcija $n$–dimenzionog argumenta, posredstvom test statistike se $n$–dimenziona oblast prostora $R^n$ prevodi u jednodimenzionu, tj. $C \subset R$. Pri tome se, bez opasnosti od zabune, koristi ista oznaka $C$.

Odlukom o prihvatanju ili odbacivanju nulte hipoteze moguće je načiniti dve vrste grešaka. Moguće je da je nulta hipoteza odbačena, a da je ona faktički tačna. Takvim zaključivanjem čini se **greška prve vrste**. Verovatnoća da se učini greška prve vrste se najčešće označava sa $\alpha$. Situacija pri kojoj se čini ova greška, u postupku u kome se primenjuje kritična oblast, nastaje kada realizovani uzorak pripadne kritičnoj oblasti, iako je $H_0$ tačna, pa se $H_0$ odbaci. Prema tome, $\alpha$ može da se izrazi kao:

$$\alpha = P_{H_0}\{(X_1, X_2, \ldots, X_n) \in C\},$$

kako se najčešće označava činjenica da je $\alpha$ uslovna verovatnoća

$$\alpha = P\{(X_1, X_2, \ldots, X_n) \in C | H_0\}.$$

Tada se još kaže da je $C$ kritična oblast veličine $\alpha$. U praksi se veličina kritične oblasti iskazuje u procentima, a najčešće korišćene vrednosti za $\alpha$ su 1% i 5%, tj. $\alpha = 0,01$ i $\alpha = 0,05$.

Verovatnoća $\alpha$ se zove **prag značajnosti** ili **nivo značajnosti** testa.

**Greška druge vrste** čini se kada se nulta hipoteza prihvati, a zapravo nije tačna. To se dešava ako realizovani uzorak ne pripadne kritičnoj oblasti, a nulta hipoteza faktički nije tačna. Verovatnoća da se načini greška druge vrste najčešće se označava sa $\beta$ i može da se izrazi kao:

$$\beta = P_{H_1}\{(X_1, X_2, \ldots, X_n) \in C^c\},$$

odnosno

$$\beta = P\{(X_1, X_2, \ldots, X_n) \in C^c | H_1\}.$$

Dakle, greška druge vrste čini se kada je faktički tačna alternativna hipoteza $H_1$, a prihvati se hipoteza $H_0$.

Šematski se verovatnoće pravilnih i pogrešnih odluka prikazuju na sledeći način:

| stvarna situacija → odluka ↓ | $H_0$ | $H_1$ |
|---|---|---|
| $H_0$ | $1-\alpha$ | $\beta$ |
| $H_1$ | $\alpha$ | $1-\beta$ |

.

Prirodno je da težimo da nadjemo test, tj. kritičnu oblast $C$, takvu da verovatnoće grešaka, $\alpha$ i $\beta$, budu što manje. Takav zahtev u opštem slučaju je protivurečan, jer najčešće smanjivanje $\alpha$ dovodi do povećanja $\beta$, i obratno. Otuda se u statistici postupa tako što se jedna od dveju verovatnoća $\alpha$ ili $\beta$ fiksira (najčešće $\alpha$), a onda se druga odredi da bude najmanja moguća u zadatim uslovima testiranja. Kaže se da se odredjuje najbolja kritična oblast veličine $\alpha$, kada se za fiksirano $\alpha$ odredjuje kritična oblast za koju je $\beta$ najmanje medju svim kritičnim oblastima veličine $\alpha$. U tom smislu skup $C$ mora da zadovolji odredjene kriterijume optimalnosti o kojima će nadalje biti još reči.

Pomenimo ovde i termin **značajnost testa** koji je, po pravilu, u vezi sa jednodimenzionom kritičnom oblašću vezanom za odredjenu test statistiku. Naime, pod značajnošću testa podrazumeva se veličina kritične oblasti čija je granica realizovana vrednost test statistike u konkretno rešavanom problemu testiranja pojedine statističke hipoteze. U tom smislu se u komunikacijama, odnosno razmeni informacija o nekom eksperimentu, koristi izraz "značajnost veća (recimo) od 5%" ili "značajnost manja od 5%".

Posmatrajmo nadalje obeležje $X$ čija raspodela pripada familiji dopustivih raspodela $\{F(x;\theta), \theta \in \Theta\}$. Statističkom hipotezom se često definiše podskup ovog skupa raspodela kao skup dopustivih raspodela. Formalno, neka je $\Lambda \subset \Theta$. Tada se nulta i alternativna hipoteza mogu iskazati na sledeći način

$$H_0 : \theta \in \Lambda\,, \quad H_1 : \theta \in \Delta \qquad \text{za neki skup} \quad \Delta \subseteq \Lambda^c = \Theta \setminus \Lambda\,. \tag{4.1}$$

Ako je $\Lambda$ jednočlan skup, nulta hipoteza je prosta, u protivnom ona je složena. Na isti način se može govoriti i o alternativnoj hipotezi.

O verovatnoći odbacivanja nulte hipoteze može da se govori i u terminima funkcije moći testa.

Definicija 37. *Funkcija moći* statističkog testa za testiranje nulte hipoteze $H_0$ protiv alternativne $H_1$ je funkcija nulte hipoteze za sve raspodele iz familije dopustivih raspodela, koja daje verovatnoću da uzorak pripadne kritičnoj oblasti testa, tj. verovatnoću odbacivanja nulte hipoteze.

Ukoliko se hipoteze iskazuju u terminima izbora raspodele iz familije dopustivih raspodela izborom vrednosti parametra kao što je navedeno u (4.1), funkcija moći se može posmatrati kao funkcija od $\theta$

$$M(\theta) = P_\theta\{(X_1, \ldots, X_n) \in C\} \quad , \quad \theta \in \Theta,$$

tj.

$$M(\theta) = P\{(X_1, \ldots, X_n) \in C|\theta\}.$$

Vrednost funkcije moći za pojedinu vrednost parametra $\theta$ (ili u opštem slučaju za prostu nultu hipotezu) zove se **moć testa**.

Nasuprot funkciji moći je funkcija koju zovemo operativnom karakteristikom testa, koja daje verovatnoću prihvatanja nulte hipoteze, tj. verovatnoću da uzorak ne pripadne kritičnoj oblasti.

DEFINICIJA 38. *Operativna karakteristika testa* je funkcija

$$N(\theta) = 1 - M(\theta) \quad , \quad \theta \in \Theta,$$

odnosno

$$N(\theta) = P_\theta\{(X_1, \ldots, X_n) \in C^c\} = P\{(X_1, \ldots, X_n) \in C^c|\theta\}.$$

**Primer 52.** Neka je $\Theta = \{\theta_0, \theta_1\}$ dvočlani skup. Neka su nulta i alternativna hipoteza redom $H_0 : \theta = \theta_0$ i $H_1 : \theta = \theta_1$. Tada je moć testa za vrednost argumenta $\theta = \theta_0$ prag značajnosti testa,

$$M(\theta_0) = P_{\theta_0}\{(X_1, \ldots, X_n) \in C\} = \alpha,$$

a za vrednost argumenta $\theta = \theta_1$ operativna karakteristika testa će biti verovatnoća greške druge vrste

$$N(\theta_1) = 1 - M(\theta_1) = P_{\theta_1}\{(X_1, \ldots, X_n) \in C^c\} = \beta \,.\triangle$$

Efektivni postupak za odredjivanje najbolje kritične oblasti zadate veličine $\alpha$ za testiranje nulte proste protiv alternativne proste hipoteze daje sledeća teorema poznata kao teorema Nejman-Pirsona.

**Teorema 4.1.1** *(Nejman-Pirson)*
*Neka je $(X_1, \ldots, X_n)$ uzorak iz populacije sa obeležjem $X$ čija raspodela pripada familiji dopustivih raspodela $\{f(x;\theta), \theta \in \Theta\}$, $\Theta = \{\theta_0, \theta_1\}$, $\theta_0 \neq \theta_1$ i neka je $L(\theta; x_1, \ldots, x_n)$ funkcija verodostojnosti posmatranog uzorka. Neka je izabran realan broj $k > 0$ i neka je skup $C \subset \mathbf{R}^n$ takav da važi:*

- *(i)*
$$\frac{L(\theta_0; x_1, \ldots, x_n)}{L(\theta_1; x_1, \ldots, x_n)} \leq k \quad \textit{za svako} \quad (x_1, \ldots, x_n) \in C$$
- *(ii)*
$$\frac{L(\theta_0; x_1, \ldots, x_n)}{L(\theta_1; x_1, \ldots, x_n)} \geq k \quad \textit{za svako} \quad (x_1, \ldots, x_n) \in C^c$$
- *(iii)*
$$\alpha = P_{\theta_0}\{(X_1, \ldots, X_n) \in C\}.$$

*Tada je $C$ najbolja kritična oblast veličine $\alpha$ za testiranje nulte proste hipoteze $H_0(\theta = \theta_0)$ protiv alternativne, takodje proste hipoteze, $H_1(\theta = \theta_1)$.*

**Dokaz.** Dokaz navodimo samo za apsolutno neprekidno obeležje, a u diskretnom slučaju dokaz je analogan.

Ako je $C$, definisana uslovima teoreme, jedina kritična oblast veličine $\alpha$, dokaz je završen. Ako postoji još neka kritična oblast veličine $\alpha$ označimo je sa $A$,

$$A \subset \mathbf{R}^n \quad \text{takva da je} \quad P_{\theta_0}\{(X_1, \ldots, X_n) \in A\} = \alpha\,,$$

treba pokazati da je verovatnoća greške druge vrste manja za oblast $C$ nego za oblast $A$, tj. da važi:

$$P_{\theta_1}\{(X_1, \ldots, X_n) \in C^c\} \leq P_{\theta_1}\{(X_1, \ldots, X_n) \in A^c\}$$

što je ekvivalentno sa

$$P_{\theta_1}\{(X_1, \ldots, X_n) \in C\} \geq P_{\theta_1}\{(X_1, \ldots, X_n) \in A\}\,. \tag{4.2}$$

Označimo sa:

$$\int\int \cdots \int_B L(\theta_i; x_1, \ldots, x_n) dx_1 \ldots dx_n = \int_B L(\theta_i) \quad , \quad i = 0, 1$$

za proizvoljnu obalast $B \subseteq R^n$. Treba, dakle, pokazati da je:

$$\int_C L(\theta_1) - \int_A L(\theta_1) \geq 0.$$

Kako je $C = (C \cap A) \cup (C \cap A^c)$ i $A = (A \cap C) \cup (A \cap C^c)$, dobijamo

$$\int_C L(\theta_1) - \int_A L(\theta_1) = \int_{C \cap A^c} L(\theta_1) - \int_{A \cap C^c} L(\theta_1).$$

Kako je s druge strane $L(\theta_1; x_1, \ldots, x_n) \geq \frac{1}{k} L(\theta_0; x_1, \ldots, x_n)$ za svako $(x_1, \ldots, x_n) \in C$, ista nejednakost važi i za svako $(x_1, \ldots, x_n) \in C \cap A^c$, pa je

$$\int_{C \cap A^c} L(\theta_1) \geq \frac{1}{k} \int_{C \cap A^c} L(\theta_0).$$

Slično je $L(\theta_1; x_1, \ldots, x_n) \leq \frac{1}{k} L(\theta_0; x_1, \ldots, x_n)$ za svako $(x_1, \ldots, x_n) \in A \cap C^c$ pa je

$$\int_{A \cap C^c} L(\theta_1) \leq \frac{1}{k} \int_{A \cap C^c} L(\theta_0).$$

Tako dobijamo da je

$$\int_{C \cap A^c} L(\theta_1) - \int_{A \cap C^c} L(\theta_1) \geq \frac{1}{k} \left[ \int_{C \cap A^c} L(\theta_0) - \int_{A \cap C^c} L(\theta_0) \right].$$

Medjutim, s druge strane je

$$\begin{aligned} \int_{C \cap A^c} L(\theta_0) - \int_{A \cap C^c} L(\theta_0) &= \int_{C \cap A^c} L(\theta_0) + \int_{C \cap A} L(\theta_0) - \int_{C \cap A} L(\theta_0) - \int_{A \cap C^c} L(\theta_0) = \\ &= \int_C L(\theta_0) - \int_A L(\theta_0) = 0\,. \end{aligned}$$

Dakle, sledi da je

$$\int_C L(\theta_1) - \int_A L(\theta_1) \geq \frac{1}{k}\left[\int_{C\cap A^c} L(\theta_0) - \int_{A\cap C^c} L(\theta_0)\right] = 0\,,$$

što je i trebalo dokazati. □

Primetimo sledeće. Kako su $\theta_0$ i $\theta_1$ poznati brojevi, to je $\frac{L(\theta_0;X_1,\ldots,X_n)}{L(\theta_1;X_1,\ldots,X_n)}$ jedna statistika čiju raspodelu možemo da odredimo, a kako je $C$ skup svih mogućih vrednosti $(x_1,\ldots,x_n)$ uzorka $(X_1,\ldots,X_n)$ za koje je

$$\frac{L(\theta_0;x_1,\ldots,x_n)}{L(\theta_1;x_1,\ldots,x_n)} \leq k\,,$$

veličinu kritične oblasti $\alpha$ možemo odrediti iz

$$\alpha = P_{\theta_0}\{(X_1,\ldots,X_n)\in C\} = P_{\theta_0}\left\{\frac{L(\theta_0;X_1,\ldots,X_n)}{L(\theta_1;X_1,\ldots,X_n)} \leq k\right\}\,,$$

kao što iz iste jednakosti za zadato $\alpha$ možemo odrediti $k$. U apsolutno neprekidnom slučaju je to moguće odrediti na jedinstven način, dok će kod diskretnog obeležja, moguće, biti potrebna dodatna informacija.

Primetimo, takodje, da u dokazu teoreme nije korišćena činjenica da su hipoteze bile definisane izborom parametra raspodele. Osim toga, istaknimo da teorema važi i za uzorak koji nije prost, što takodje sledi iz dokaza. Navedimo jedan primer odredjivanja najbolje kritične oblasti Nejman-Pirsonovom teoremom kod koga su dve proste hipoteze drugačije definisane.

**Primer 53.** Neka je dat prost slučajni uzorak $\mathbf{X} = (X_1,\ldots,X_n)$ iz populacije sa obeležjem $X$ čija raspodela pripada skupu dopustivih raspodela

$$\{f_1(x), f_2(x)\}\,,$$

$$f_1(x) = \begin{cases} \frac{e^{-1}}{x!}, & x = 0,1,2,\ldots \\ 0, & \text{inače} \end{cases}$$

$$f_2(x) = \begin{cases} (\frac{1}{2})^{x+1}, & x = 0,1,2,\ldots \\ 0, & \text{inače} \end{cases}\,.$$

Postavimo dve proste hipoteze

$$H_0 : f = f_1\,, \quad H_1 : f = f_2\,.$$

Količnik odgovarajućih funkcija verodostojnosti će tada biti

$$\frac{g(x_1,\ldots,x_n)}{h(x_1,\ldots,x_n)} = \frac{\frac{e^{-n}}{x_1!\ldots x_n!}}{(\frac{1}{2})^n(\frac{1}{2})^{x_1+\ldots+x_n}} = \frac{(2e^{-1})^n 2^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n (x_i!)}\,.$$

Za realan broj $k > 0$ razmotrimo nejednakost

$$\frac{g(x_1,\ldots,x_n)}{h(x_1,\ldots,x_n)} \leq k,$$

tj.

$$\frac{(2e^{-1})^n 2^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n (x_i!)} \leq k\,.$$

Iz nje sledi

$$n\ln 2 - n + (\sum_{i=1}^n x_i)\ln 2 - \sum_{i=1}^n \ln(x_i!) \leq \ln k$$

$$(\sum_{i=1}^n x_i)\ln 2 - \sum_{i=1}^n \ln(x_i!) \leq \ln k + n - n\ln 2$$

$$\sum_{i=1}^n \ln\left(\frac{2^{x_i}}{x_i!}\right) \leq k_1 \quad .$$

Dakle, najbolja kritična oblast veličine $\alpha$ za definisano testiranje je oblika

$$C = \left\{(x_1, x_2, \ldots, x_n) : \sum_{i=1}^n \ln\left(\frac{2^{x_i}}{x_i!}\right) \leq c\right\}\,. \triangle$$

Još jedan aspekt ove teoreme zavredjuje posebnu pažnju. Radi se o broju parametara koji se javlja u raspodeli posmatranog obeležja, tj. o dimenziji parametra. Pažljivom analizom dokaza vidi se da dimenzija parametra nije od značaja za dokaz, niti se iskaz teoreme vezuje za dimenziju parametra. Dakle, gustina raspodele obeležja koje je predmet testiranja može da zavisi od parametra proizvoljne dimenzije, odnosno od proizvoljnog konačnog broja parametara. Ono što je bitno, to je da su obe, i nulta i alternativna hipoteza, proste, tj. da u potpunosti odredjuju raspodelu.

### 4.1.1 Uniformno najmoćniji testovi

Nadalje ćemo razmatrati mogućnosti za testiranje nulte proste protiv alternativne složene hipoteze. Razmotrimo pažljivo sledeći primer.

**Primer 54.** Neka obeležje $X$ ima gustinu raspodele koja pripada familiji $\{\mathcal{N}(\theta, 1), \theta \in \{\theta_0, \theta_1\}, \theta_0 < \theta_1\}$. Primenom Nejman-Pirsonove teoreme odredimo najbolju kritičnu oblast za testiranje proste nulte hipoteze $H_0 : \theta = \theta_0$ protiv proste alternativne $H_1 : \theta = \theta_1$ na osnovu prostog slučajnog uzorka $\mathbf{X} = (X_1, \ldots, X_n)$.

Funkcija verodostojnosti je data sa:

$$L(\theta_j; x_1, \ldots, x_n) = \frac{1}{(\sqrt{2\pi})^n} \exp\left\{-\sum_{i=1}^n \frac{(x_i - \theta_j)^2}{2}\right\}, \quad j = 0, 1\,.$$

Da bismo našli kritičnu oblast $C$ po teoremi Nejman-Pirsona, posmatramo sledeću nejednakost:

$$\frac{L(\theta_0)}{L(\theta_1)} = \frac{\frac{1}{(\sqrt{2\pi})^n} \exp\left\{-\sum_{i=1}^n \frac{(x_i-\theta_0)^2}{2}\right\}}{\frac{1}{(\sqrt{2\pi})^n} \exp\left\{-\sum_{i=1}^n \frac{(x_i-\theta_1)^2}{2}\right\}} = \exp\left\{-\frac{1}{2}\sum_{i=1}^n [(x_i-\theta_0)^2 - (x_i-\theta_1)^2]\right\} \leq k.$$

Iz ove nejednakosti sledi da je

$$-\frac{1}{2}\sum_{i=1}^{n}[(x_i-\theta_0)^2-(x_i-\theta_1)^2]\le \ln k\,.$$

Posle kraćeg računanja imaćemo niz nejednakosti

$$(\theta_0-\theta_1)\sum_{i=1}^{n}(\theta_0+\theta_1-2x_i)\ge -2\ln k\,,$$

$$\sum_{i=1}^{n}(\theta_0+\theta_1)-2\sum_{i=1}^{n}x_i\le\frac{2\ln k}{\theta_1-\theta_0}\,,$$

$$\frac{1}{n}\sum_{i=1}^{n}x_i\ge\frac{\ln k}{(\theta_0-\theta_1)n}+\frac{\theta_0+\theta_1}{2}\,,$$

$$\frac{1}{n}\sum_{i=1}^{n}x_i\ge c\,,\quad c=\text{const.}\,.\,\triangle$$

Obratimo pažnju da je za oblik kritične oblasti u navedenom primeru od značaja bila činjenica da je $\theta_0<\theta_1$. Medjutim, još važnije je uočiti da bi kritična oblast zadržala ovaj oblik (ne i vrednost konstante $c$) i za svaki drugi broj $\theta_1\in(\theta_0,+\infty)$.

Ograničenje $\theta_0<\theta_1$ ima za posledicu da smo bili u mogućnosti da na jedinstven način odredimo najbolju kritičnu oblast. S tim u vezi možemo problem testiranja da postavimo na sledeći način.

**Primer 55.** Neka obeležje $X$ ima gustinu raspodele koja pripada familiji $\{\mathcal{N}(\theta,1),\theta\in R\}$. Testirati hipotezu $H_0:\theta=\theta_0$ protiv alternativne $H_1:\theta>\theta_0$ sa pragom značajnosti $\alpha$ na osnovu prostog slučajnog uzorka $\mathbf{X}=(X_1,\ldots,X_n)$.

Iz prethodnog primera sledi da će za svako fiksirano $\theta_1\in(\theta_0,+\infty)$ kritična oblast veličine $\alpha$ biti odredjena kao

$$C=[c,+\infty)\,,$$

gde je $c$ odredjeno po kriterijumu

$$\alpha=P\left\{\sum_{i=1}^{n}X_i\ge c\right\}\,.$$

Preciznije,

$$C=\left\{(x_1,\ldots,x_n):\sum_{i=1}^{n}x_i\ge\frac{n}{2}(\theta_1+\theta_0)-\frac{\ln k}{\theta_1-\theta_0}\right\}.$$

Konstatujemo da će prema Nejman-Pirsonovoj teoremi oblast $C$ biti najbolja kritična oblast za testiranje nulte proste protiv svake alternativne proste hipoteze sadržane u alternativnoj složenoj hipotezi. $\triangle$

Definicija 39. Kritična oblast $C$ je *uniformno najmoćnija oblast* veličine $\alpha$ za testiranje proste hipoteze $H_0$ protiv alternativne složene hipoteze $H_1$ ako je skup $C$ najbolja kritična oblast veličine $\alpha$ za testiranje $H_0$ protiv svake proste hipoteze sadržane u $H_1$. Test definisan ovom kritičnom oblašću, zove se *uniformno najmoćniji test* sa pragom značajnosti $\alpha$ za testiranje proste hipoteze $H_0$ protiv alternativne složene $H_1$.

Uniformno najmoćniji test ne mora uvek da postoji, medjutim, kada postoji, Nejman-Pirsonova teorema daje tehniku za njegovo nalaženje.

**Primer 56.** Posmatrajmo obeležje $X$ i familiju njegovih dopustivih raspodela definisanu u prethodnom primeru.

Možemo konstatovati da ako definišemo nultu i alternativnu hipotezu na sledeći način

$$H_0 : \theta = \theta_0 \quad , \quad H_1 : \theta \neq \theta_0$$

neće postojati uniformno najmoćnija oblast, pa ni uniformno najmoćniji test. Zaista, za $\theta_1 < \theta_0$ kritična oblast će biti odredjena sa

$$\sum_{i=1}^{n} x_i \leq \frac{n}{2}(\theta_1 + \theta_0) - \frac{\ln k}{\theta_1 - \theta_0} \, . \triangle$$

### 4.1.2 Test količnika verodostojnosti

Postavlja se pitanje možemo li testirati nultu složenu hipotezu protiv alternativne takodje složene. U tu svrhu koristimo intuitivni test, test količnika verodostojnosti, koji koristi ideju teoreme Nejman-Pirsona.

**Primer 57.** Neka je dato obeležje $X : \{\mathcal{N}(\theta_1, \theta_2), -\infty < \theta_1 < +\infty, \theta_2 > 0\}$ i treba testirati nultu složenu hipotezu $H_0 : (\theta_1 = 0, \theta_2 > 0)$ protiv alternativne složene hipoteze $H_1 : (\theta_1 \neq 0, \theta_2 > 0)$ na osnovu prostog slučajnog uzorka $\mathbf{X} = (X_1, \ldots, X_n)$.

Posmatrajmo parametarske prostore definisane na sledeći način:

$$A_0 = \{(\theta_1, \theta_2), \theta_1 = 0, \theta_2 > 0\}$$

$$\Theta = \{(\theta_1, \theta_2), -\infty < \theta_1 < +\infty, \theta_2 > 0\} \, .$$

Sada se postavljene hipoteze mogu iskazati u terminima ovih prostora

$$H_0 : (\theta_1, \theta_2) \in A_0 \quad \text{i} \quad H_1 : (\theta_1, \theta_2) \in A_0^c \, ,$$

gde je $A_0^c = \Theta \setminus A_0$.

Podjimo od funkcije verodostojnosti za postavljene hipoteze, koja je sada funkcija dve promenljive, odnosno od dvodimenzionog parametra, razlikujući joj vrednost na skupu $\Theta$ i skupu $A_0$:

$$L(\theta_1, \theta_2; x_1, \ldots, x_n) = \frac{1}{(2\pi\theta_2)^{\frac{n}{2}}} \exp\left\{-\frac{1}{2\theta_2}\sum_{i=1}^{n}(x_i - \theta_1)^2\right\} = L(\Theta)$$

$$L(\theta_1, \theta_2; x_1, \ldots, x_n) = \frac{1}{(2\pi\theta_2)^{\frac{n}{2}}} \exp\left\{-\frac{1}{2\theta_2}\sum_{i=1}^{n} x_i^2\right\} = L(A_0) \, .$$

Uočimo količnik $\frac{L(A_0)}{L(\Theta)}$. Kako analitički izrazi za $L(A_0)$ i $L(\Theta)$ zavise od nepoznatih parametara, to se jednostavnom zamenom realizovanog uzorka slučajnim neće dobiti

statistike. Medjutim, baš zbog pomenute zavisnosti ima smisla odredjivati maksimum funkcija $L(A_0)$ i $L(\Theta)$ po $\theta = (\theta_1, \theta_2)$ na $A_0$ i na $\Theta$, redom, za svaki od realizovanih uzoraka $\mathbf{x} = (x_1, \ldots, x_n) \in \mathcal{X}$, pa se zatim posmatra sledeći količnik

$$\frac{\max_{(\theta_1,\theta_2)\in A_0} L(A_0)}{\max_{(\theta_1,\theta_2)\in\Theta} L(\Theta)} = \frac{L(\hat{A}_0)}{L(\hat{\Theta})} = \lambda = \lambda(x_1, \ldots, x_n)\,,$$

čiji analitički izraz neće zavisiti od nepoznatih parametara.

Sprovedeni postupak, zapravo, znači zamenu nepoznatih parametara njihovim ocenama maksimalne verodostojnosti.

Dobijeni količnik se zove **količnik verodostojnosti**. Kako je $A_0 \subset \Theta$, onda je količnik manji od jedinice, pa je $0 \leq \lambda \leq 1$. Za nultu hipotezu su problematične male vrednosti ovog količnika, odnosno, funkcije $\lambda(x_1, \ldots, x_n)$, $(x_1, \ldots, x_n) \in \mathcal{X}$, jer ukazuju na to da je verodostojnost $H_0$ mala u poredjenju sa $H_1$, tj. da podaci favorizuju $H_1$ u odnosu na $H_0$.

Neka je $\lambda_0$ pozitivan pravi razlomak. Princip testiranja količnikom verodostojnosti nalaže da se hipoteza $H_0 : (\theta_1, \theta_2) \in A_0$ odbaci ako i samo ako je

$$\lambda(x_1, \ldots, x_n) = \lambda \leq \lambda_0$$

za neko $0 < \lambda_0 < 1$. Funkcija $\lambda(X_1, \ldots, X_n)$ je slučajna promenljiva, tj. statistika, pa je prag značajnosti ovog testa dat sa

$$\alpha = P_{H_0}\{\lambda(X_1, \ldots, X_n) \leq \lambda_0\}.$$

Ostaje još da primer dovršimo efektivnim odredjivanjem maksimuma funkcija

$$\ln L(A_0) = -\frac{n}{2}\ln(2\pi\theta_2) - \frac{1}{2\theta_2}\sum_{i=1}^{n} x_i^2$$

i

$$\ln L(\Theta) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\theta_2 - \frac{1}{2\theta_2}\sum_{i=1}^{n}(x_i - \theta_1)^2.$$

Uobičajenim postupkom odredjivanja parcijalnih izvoda po $\theta_1$ i $\theta_2$ sledi:

$$\frac{\partial \ln L(A_0)}{\partial\theta_2} = -\frac{n}{2}\frac{1}{\theta_2} + \frac{1}{2\theta_2^2}\sum_{i=1}^{n} x_i^2 = 0\,,$$

$$\theta_2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2\,.$$

Ovu vrednost uvrstimo u izraz za $L(A_0)$ da bismo dobili $L(\hat{A}_0)$:

$$L(\hat{A}_0) = \left(\frac{2\pi}{n}\sum_{i=1}^{n} x_i^2\right)^{-\frac{n}{2}} e^{-\frac{n}{2}} = \left(\frac{ne^{-1}}{2\pi\sum_{i=1}^{n} x_i^2}\right)^{\frac{n}{2}}.$$

Ovo isto uradimo i za $L(\hat{\Theta})$, dakle,

$$\frac{\partial \ln L(\Theta)}{\partial\theta_1} = \frac{\sum_{i=1}^{n}(x_i - \theta_1)}{\theta_2}$$

$$\frac{\partial \ln L(\Theta)}{\partial \theta_2} = \frac{\sum_{i=1}^{n}(x_i-\theta_1)^2}{2\theta_2^2} - \frac{n}{2\theta_2}\,.$$

Izjednačavanjem ovih izvoda sa nulom, dobićemo vrednosti za $\theta_1$ i $\theta_2$ koje iznose redom:

$$\theta_1 = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}_n$$

$$\theta_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x}_n)^2 = \overline{s}_n{}^2.$$

Konačno se dobija da je:

$$\lambda = \frac{1}{\left(1+\frac{n\overline{x}_n{}^2}{\sum_{i=1}^{n}(x_i-\overline{x}_n)^2}\right)^{\frac{n}{2}}}\,, \quad \lambda \le \lambda_0$$

odnosno, dalje

$$\frac{n\overline{x}_n{}^2}{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x}_n)^2} \ge (n-1)(\lambda_0^{-\frac{2}{n}}-1)\,,$$

tj.

$$\frac{\sqrt{n}|\overline{x}_n|}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x}_n)^2}} \ge \sqrt{(n-1)(\lambda_0^{-\frac{2}{n}}-1)} = c\,.$$

U ovom konkretnom primeru dobili smo statistiku

$$\frac{|\overline{X}_n|}{\sqrt{\widetilde{S}_n^2}}\sqrt{n}$$

koja ima studentovu raspodelu sa $n-1$ stepenom slobode i koja omogućava da $n$-dimenzionu kritičnu oblast testa prevedemo na jednodimenzionu, a koja će biti unija intervala, $(-\infty,-c]\cup[c,+\infty)$. Dakle, kad izračunamo $c$ tražimo $\lambda_0$ i vraćamo se na kritičnu oblast $C=[0,\lambda_0]$. $\triangle$

Test količnika verodostojnosti je intuitivni test i ne postoji strogi razlog za njegovu primenu u smislu da je uniformno najmoćniji. Medjutim, za većinu praktičnih problema opisanim postupkom količnika verodostojnosti se dobija najmoćniji mogući test za zadate uslove. Nažalost, za statistiku $\lambda(X_1,\ldots,X_n)$ ovog testa se uvek ne nalazi raspodela medju poznatim raspodelama. Dva najčešća uslova koja su potrebna da bi se odredila raspodela test statistike do koje se dolazi metodom količnika verodostojnosti su:

- regularnost (možda i u nekoj oslabljenoj formi, ali se uglavnom ovi uslovi baziraju na diferencijabilnosti) i
- da oblast u kojoj je funkcija verodostojnosti strogo pozitivna **ne zavisi** od nepoznatih vrednosti parametara.

Razmotrimo činjenicu da smo do rezultata dobijenog u prethodnom primeru mogli doći i drugačijim razmišljanjem.

**Primer 58.** Neka raspodela obeležja X pripada familiji dopustivih raspodela

$$\{\mathcal{N}(m,\sigma^2),\, m\in(-\infty,+\infty),\, \sigma^2\in(0,\infty)\}.$$

Želimo da na osnovu prostog uzorka obima $n$ testiramo nultu hipotezu $H_0 : m = m_0$ protiv alternativne $H_1 : m \neq m_0$. U vezi sa tačkastim ocenama parametara, naveli smo da je statistika $(Z_1, Z_2)$ dvodimenziona kompletna dovoljna statistika za dvodimenzioni parametar $(m,\sigma^2)$ gde je $Z_1 = \overline{X}_n$ i $Z_2 = \frac{n}{n-1}\overline{S}_n^2$. Dakle, naš test se mora da bazira na ovim statistikama, pa će količnik funkcija verodostojnosti biti:

$$\frac{L(m_0;x_1,\ldots,x_n)}{L(m_1;x_1,\ldots,x_n)} = \varphi(z_1,z_2) = \varphi(z_1(x_1,\ldots,x_n), z_2(x_1,\ldots,x_n))$$

za neko $m_1 \neq m_0$.

Koristeći se ovom činjenicom, pribećićemo postupku koji je jednostavniji od direktnog izračunavanja.

Dokazali smo ranije da su $\overline{X}_n$ i $\overline{S}_n^2$ nezavisne slučajne promenljive pa su takve i

$$\frac{\overline{X}_n - m_0}{\frac{\sigma}{\sqrt{n}}} \quad , \quad \frac{n\overline{S}_n^2}{\sigma^2}\,.$$

Za poslednje znamo da imaju sledeće raspodele:

$$\frac{\overline{X}_n - m_0}{\frac{\sigma}{\sqrt{n}}} : \mathcal{N}(0,1) \quad , \quad \frac{n\overline{S}_n^2}{\sigma^2} : \chi^2_{n-1}.$$

Njihov količnik ima Studentovu raspodelu sa $n-1$ stepeni slobode:

$$t_{n-1} = \frac{\frac{\overline{X}_n - m_0}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{\frac{n\overline{S}_n^2}{\sigma^2}}{n-1}}} = \frac{\overline{X}_n - m_0}{\overline{S}_n}\sqrt{n-1} \quad , \quad \overline{S}_n = \sqrt{\overline{S}_n^2}\,.$$

S obzirom na simetričnost studentove raspodele oko svoje očekivane vrednosti, dakle oko nule, i činjenice da ako je $m_0$ tačna vrednost parametra $m$,

$$E_{m_0}\left(\frac{\overline{X}_n - m_0}{\overline{S}_n}\sqrt{n-1}\right) = E(t_{n-1}) = 0 \quad \text{i} \quad H_1(m \neq m_0)\,,$$

kritičnu oblast $C$ izabraćemo na sledeći način:

$$C = \left\{(x_1,\ldots,x_n) : \frac{|\overline{x}_n - m_0|}{\overline{s}_n}\sqrt{n-1} \geq k\right\}, \quad k > 0\,.$$

Dakle, za datu vrednost $\alpha$ broj $k$ odredjujemo iz uslova:

$$P_{m_0}\left\{\frac{|\overline{X}_n - m_0|}{\overline{S}_n}\sqrt{n-1} \geq k\right\} = P\{|t_{n-1}| \geq k\} = \alpha\,,$$

gde $k$ odredjujemo iz tablice za Studentovu raspodelu. $\triangle$

Načinimo sada generalizaciju.

Neka je $(X_1, \ldots, X_n)$ uzorak obima $n$ iz populacije sa obeležjem $X$ čija raspodela pripada familiji dopustivih raspodela $\{f(x; \theta_1, \theta_2, \ldots, \theta_m), (\theta_1, \theta_2, \ldots, \theta_m) \in \Theta\}$. Zadržaćemo se na parametarskom prostoru $\Theta = \{(\theta_1, \theta_2, \ldots, \theta_m)\}$. Neka je $A_0 \subset \Theta$. Želimo da testiramo (prostu ili složenu ) hipotezu $H_0 : (\theta_1, \theta_2, \ldots, \theta_m) \in A_0$, $dim A_0 = k < m$, protiv svih alternativnih. Definišimo funkcije verodostojnosti posmatranog uzorka

$$L(A_0) = \varphi(x_1, \ldots, x_n; \theta_1, \theta_2, \ldots, \theta_m) \quad , \quad (\theta_1, \theta_2, \ldots, \theta_m) \in A_0$$

i

$$L(\Theta) = \varphi(x_1, \ldots, x_n; \theta_1, \theta_2, \ldots, \theta_m) \quad , \quad (\theta_1, \theta_2, \ldots, \theta_m) \in \Theta \, .$$

Neka su, kao i ranije, $L(\widehat{A}_0)$ i $L(\widehat{\Theta})$ maksimumi gornjih funkcija po $(\theta_1, \theta_2, \ldots, \theta_m) \in A_0$ i $(\theta_1, \theta_2, \ldots, \theta_m) \in \Theta$ redom (za koje pretpostavljamo da postoje). Količnik

$$\frac{L(\widehat{A}_0)}{L(\widehat{\Theta})} = \lambda = \lambda(x_1, \ldots, x_n)$$

se zove **količnik verodostojnosti**. Neka je $\lambda_0$ pozitivan realan broj manji od jedinice. Princip testiranja količnikom verodostojnosti nalaže da se hipoteza $H_0 : (\theta_1, \theta_2, \ldots, \theta_m) \in A_0$ odbaci ako i samo ako

$$\lambda(x_1, \ldots, x_n) = \lambda \leq \lambda_0 \, .$$

Funkcija $\lambda(X_1, \ldots, X_n)$ je slučajna promenljiva, pa je prag značajnosti ovog testa dat sa

$$\alpha = P_{H_0}\{\lambda(X_1, \ldots, X_n) \leq \lambda_0\} \, .$$

Sledeći primer ilustruje generalizaciju.

**Primer 59.** Neka su obeležja X i Y nezavisna sa raspodelama $\mathcal{N}(\theta_1, \theta_3)$ i $\mathcal{N}(\theta_2, \theta_3)$ redom, gde su $\theta_1, \theta_2, \theta_3$ nepoznati parametri definisani parametarskim prostorom

$$\Theta = \{(\theta_1, \theta_2, \theta_3); -\infty < \theta_1 < \infty, -\infty < \theta_2 < \infty, 0 < \theta_3 < \infty\}.$$

Neka su $(X_1, \ldots, X_n)$ i $(Y_1, \ldots, Y_m)$ nezavisni prosti uzorci iz ovih raspodela. Neka je $A_0 = \{(\theta_1, \theta_2, \theta_3); -\infty < \theta_1 = \theta_2 < \infty, 0 < \theta_3 < \infty\}$. Testira se hipoteza $H_0 : (\theta_1, \theta_2, \theta_3) \in A_0$ protiv svih alternativnih.

Funkcija verodostojnosti se formira iz prostog uzorka obima $n+m > 2$, $(X_1, \ldots, X_n, Y_1, \ldots, Y_m)$ pa je,

$$L(\Theta) = \left(\frac{1}{2\pi\theta_3}\right)^{\frac{n+m}{2}} \exp\left\{-\frac{\sum_{i=1}^{n}(x_i - \theta_1)^2 + \sum_{j=1}^{m}(y_j - \theta_2)^2}{2\theta_3}\right\}$$

i

$$L(A_0) = L(\Theta)\Big|_{\theta_1 = \theta_2} \, .$$

Rešenje se dobija uobičajenim odredjivanjem maksimuma iz jednačina dobijenih pomoću parcijalnih izvoda

$$\frac{\partial \ln L(A_0)}{\partial \theta_1} \, , \, \frac{\partial \ln L(A_0)}{\partial \theta_3} \, , \, \frac{\partial \ln L(\Theta)}{\partial \theta_1} \, , \, \frac{\partial \ln L(\Theta)}{\partial \theta_2} \, , \, \frac{\partial \ln L(\Theta)}{\partial \theta_3}$$

i njihovim izjednačavanjem sa nulom. Dakle,

$$L(\widehat{A_0}) = (2\pi e u_2)^{-\frac{n+m}{2}},$$

$$u_2 = \frac{\sum_{i=1}^{n}(x_i - u_1)^2 + \sum_{j=1}^{m}(y_j - u_1)^2}{n+m},$$

$$u_1 = \frac{\sum_{i=1}^{n} x_i + \sum_{j=1}^{m} y_j}{n+m}$$

i

$$L(\widehat{\Theta}) = (2\pi e v_3)^{-\frac{n+m}{2}},$$

$$v_3 = \frac{\sum_{i=1}^{n}(x_i - v_1)^2 + \sum_{j=1}^{m}(y_j - v_2)^2}{n+m},$$

$$v_2 = \frac{\sum_{j=1}^{m} y_j}{m},$$

$$v_1 = \frac{\sum_{i=1}^{m} x_i}{n}.$$

Otuda,

$$\lambda(x_1, \ldots, x_n, y_1, \ldots, y_m) = \left(\frac{v_3}{u_2}\right)^{\frac{n+m}{2}} \leq \lambda_0$$

vodi do test statistike

$$T = \frac{\overline{X}_n - \overline{Y}_m}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 + \sum_{j=1}^{m}(Y_j - \overline{Y}_m)^2}} \sqrt{\frac{nm(n+m-2)}{n+m}},$$

koja ima studentovu raspodelu sa $n+m-2$ stepena slobode. $\triangle$

U vezi sa generalizacijom ovog postupka navedimo bez dokaza sledeću teoremu.

**Teorema 4.1.2** *Neka uzorak $(X_1, \ldots, X_n)$ ima zajedničku gustinu raspodele, odnosno funkciju verodostojnosti $L(\theta)$, gde je $\theta \in \Theta$ višedimenzioni parametar. Neka je $m$ dimenzija parametarskog prostora $\Theta$, a $k$ dimenzija parametra definisanog hipotezom $H_0 : \theta \in A_0$, tj. dimenzija prostora $A_0$. Tada za veliko $n$, statistika*

$$-2 \ln \lambda(X_1, \ldots, X_n) = -2 \ln \frac{L(\widehat{A}_0)}{L(\widehat{\Theta})}$$

*ima približno $\chi^2$ raspodelu sa $m-k$ stepeni slobode.*

Ova teorema omogućava da odredimo granicu kritične oblasti za veliki obim uzorka, bez obzira na raspodelu posmatranog obeležja. Medjutim, koji je obim uzorka, $n$, dovoljno veliki nije moguće odrediti u opštem slučaju, već će brzina konvergencije zavisiti od raspodele obeležja koje se posmatra.

## 4.2 Parametarski testovi

U ovom odeljku navešćemo samo nekoliko važnijih testova za testiranje parametarskih hipoteza. Svi testovi ovog poglavlja su testovi količnika verodostojnosti.

Već smo istakli da ne postoji "univerzalni" obim uzorka koji će garantovati valjanost statističkih zaključaka sa zadatom tačnošću. Kada je reč o testiranju parametarskim testovima, u tom smislu je posebno zanimljiv test za nepoznato matematičko očekivanje obeležja, koji za dovoljno veliki obim uzorka, može da se tretira kao neparametarski u smislu gore navedene definicije. Naime, ako se raspolaže uzorkom dovoljno velikog obima, test statistika će imati asimptotski normalnu normiranu raspodelu, bez obzira na raspodelu obeležja o čijem se očekivanju radi.

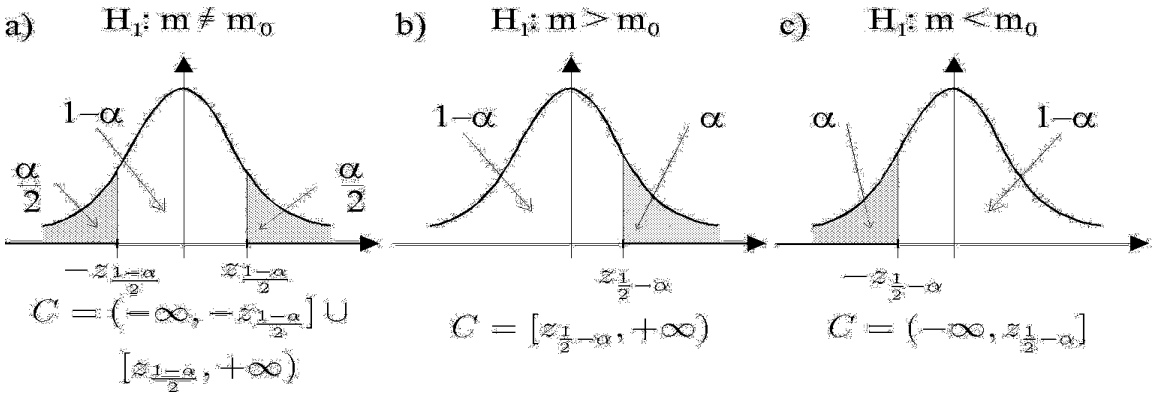
Formulacija "veliki uzorak" u smislu testova ovog odeljka je uzorak čiji je obim $n \geq 30$.

### 4.2.1 Test za srednju vrednost obeležja za velike uzorke

Kod testa za srednju vrednost, $m$, testira se nulta hipoteza $H_0(m = m_0)$, protiv alternativne hipoteze koja može da bude trojaka: $H_1(m \neq m_0)$, $H_1(m > m_0)$ ili $H_1(m < m_0)$, na osnovu prostog uzorka $\mathbf{X} = (X_1, \ldots, X_n)$. Testiranje srednje vrednosti se bazira na sredini uzorka, $\overline{X}_n$. U slučaju da disperzija obeležja $X$ čija se srednja vrednost ocenjuje, nije poznata, koristi se statistika

$$Z_0 = \frac{\overline{X}_n - m_0}{\widetilde{S}_n} \cdot \sqrt{n} \tag{4.3}$$

koja ima približno normalnu raspodelu $\mathcal{N}(0,1)$ za **veliki obim uzorka**, bez obzira na raspodelu obeležja $X$. Jasno da je $\widetilde{S}_n$ ocena nepoznate standardne devijacije obeležja $X$, te je $\widetilde{S}_n/\sqrt{n}$ ocena parametra $\sqrt{D(\overline{X}_n)}$.




Slika 4.1: Testiranje srednje vrednosti ($m$) obeležja za velike uzorke: oblasti prihvatanja nulte hipoteze $H_0(m = m_0)$ i kritične oblasti za različite alternativne hipoteze $H_1$. Kritična oblast je deo apscisne ose "ispod" šrafiranog dela.

Kritična oblast veličine $\alpha$ za testiranje $H_0$ protiv $H_1(m \neq m_0)$, odredjuje se iz uslova

$$P_{H_0}\{|Z_0| \geq c\} = \alpha.$$

S obzirom na raspodelu statistike $Z_0$, vrednost $c$ se najčešće zapisuje kao $c = z_{\frac{1-\alpha}{2}}$ (slika 4.1). Ili uopšte, u odnosu na sve alternativne hipoteze kritična oblast veličine $\alpha$ odredjuje se prema tabeli:

| $H_0$ | $H_1$ | $H_0$ se odbacuje ako se za realizovani uzorak dobije |
|---|---|---|
| $m = m_0$ | $m \neq m_0$ | $\left\| \frac{\overline{x}_n - m_0}{\widetilde{s}_n/\sqrt{n}} \right\| \geq z_{0,5-\alpha/2}$ |
| $m = m_0$ | $m > m_0$ | $\frac{\overline{x}_n - m_0}{\widetilde{s}_n/\sqrt{n}} \geq z_{0,5-\alpha}$ |
| $m = m_0$ | $m < m_0$ | $\frac{\overline{x}_n - m_0}{\widetilde{s}_n/\sqrt{n}} \leq -z_{0,5-\alpha}$ |

.

Medjutim, ako je obim uzorka mali, test statistika (4.3) nema normalnu raspodelu čak ni kod normalne raspodele obeležja $X$. Kada uzorak ima $n$–dimenzionu normalnu raspodelu, a obim uzorka je mali, (4.3) ima Studentovu raspodelu, o čemu će nadalje biti reči.

Ukoliko je disperzija posmatranog obeležja (odnekud) poznata, koristiće se $\sigma/\sqrt{n}$, $\sigma = \sqrt{D(X)}$, za standardizaciju statistike $\overline{X}_n$, tj. test statistika će biti

$$Z_0 = \frac{\overline{X}_n - m_0}{\sigma}\sqrt{n}$$

i njena raspodela će biti takodje normalna normirana za veliki obim uzorka, bez obzira na raspodelu posmatranog obeležja.

### 4.2.2 Parametarska testiranja kod normalne raspodele

Narednih šest grupa testova odnose se na normalnu raspodelu, tj. na obeležje $X$ čija raspodela pripada familiji dopustivih raspodela $\{\mathcal{N}(m, \sigma^2),\ m \in R,\ \sigma^2 \in R^+\}$.

1. Testira se hipoteza o nepoznatom matematičkom očekivanju obeležja $X$, $H_0(m = m_0)$. Pri tome se razlikuju dva slučaja:

   (a) $\sigma^2$ poznato, ili $\sigma^2$ nepoznato a obim uzorka veliki. Test statistika za slučaj poznate disperzije je

   $$Z_0 = \frac{\overline{X}_n - m_0}{\sigma}\sqrt{n}\,.$$

   U ovom slučaju $Z_0$ ima tačnu raspodelu $\mathcal{N}(0,1)$, dok za nepoznato $\sigma^2$

   $$Z_0 = \frac{\overline{X}_n - m_0}{\overline{S}_n}\sqrt{n-1}\,,$$

   ima samo asimptotski raspodelu $\mathcal{N}(0,1)$. Kritične oblasti koje odgovaraju pojedinim alternativnim hipotezama za realizovanu vrednost $z_0$ statistike $Z_0$ prikazane su u tabeli:

| $H_0$ | $H_1$ | $C$ |
|---|---|---|
| $m = m_0$ | $m \neq m_0$ | $\lvert z_0\rvert \geq z_{0,5-\alpha/2}$ |
| $m = m_0$ | $m > m_0$ | $z_0 \geq z_{0,5-\alpha}$ |
| $m = m_0$ | $m < m_0$ | $z_0 \leq -z_{0,5-\alpha}$ |

.

Kritične oblasti su prikazane na slici 4.1.

**Primer 60.** Za sledeće rezultate:

| Br. poena | Br. studenata |
|---|---|
| [50,60) | 4 |
| [60,70) | 17 |
| [70,80) | 24 |
| [80,90) | 10 |
| [90,100] | 5 |

koji su dobijeni testiranjem praga osetljivosti, testirati hipotezu da je srednja vrednost jednaka 75 za prag značajnosti $\alpha = 0,01$, ako je disperzija poznata i iznosi 100.

Testira se hipoteza $H_0(m = 75)$ protiv alternativne $H_1(m \neq 75)$. Disperzija je poznata i iznosi 100, prag značajnosti je $\alpha = 0,01$, dok je uzoračka sredina jednaka

$$\overline{x}_{60} = \frac{1}{60}(4 \cdot 55 + 17 \cdot 65 + 24 \cdot 75 + 10 \cdot 85 + 5 \cdot 95) = 74,17.$$

Tako se dobija da je realizovana vrednost test statistike jednaka

$$\frac{74,17 - 75}{10}\sqrt{60} = -0,64.$$

Kako je $z_{0,495} = 2,575$, to je kritična oblast

$$C = (-\infty; -2,575] \cup [2,575; +\infty).$$

Kako $-0,64 \notin C$, to se hipoteza $H_0$ prihvata. (Značajnost ovog testa je 0,5222, dakle veća od 0,01.) $\triangle$

(b) $\sigma^2$ nepoznato i obim uzorka mali. Test statistika

$$t_0 = \frac{\overline{X}_n - m_0}{\widetilde{S}_n}\sqrt{n}$$

ima Studentovu raspodelu sa $n - 1$ stepeni slobode. Tabela odgovarajućih kritičnih oblasti je:

| $H_0$ | $H_1$ | $C$ |
|---|---|---|
| $m = m_0$ | $m \neq m_0$ | $\lvert t_0\rvert \geq t_{n-1;\frac{1-\alpha}{2}}$ |
| $m = m_0$ | $m > m_0$ | $t_0 \geq t_{n-1;0,5-\alpha}$ |
| $m = m_0$ | $m < m_0$ | $t_0 \leq -t_{n-1;0,5-\alpha}$ |

,

gde se konstante $t_{n-1,\frac{1-\alpha}{2}}$ i $t_{n-1,\frac{1}{2}-\alpha}$, tj. granice kritične oblasti, čitaju iz tablice Studentove raspodele. Grafički prikaz bi bio analogan onome sa slike 4.1

2. Testira se hipoteza o jednakosti srednjih vrednosti dvaju nezavisnih obeležja $X$ i $Y$ sa pretpostavljenim raspodelama:

$$X \ : \ \mathcal{N}(m_X, \sigma_X^2), \quad Y \ : \ \mathcal{N}(m_Y, \sigma_Y^2)$$

na osnovu prostih nezavisnih uzoraka $\mathbf{X} = (X_1, \ldots, X_{n_X})$ i $\mathbf{Y} = (Y_1, \ldots, Y_{n_Y})$ obima $n_X$ i $n_Y$ redom,

$$H_0(m_X = m_Y), \quad \text{odnosno,} \quad H_0(m_X - m_Y = 0).$$

Razmatraju se dva slučaja:

(a) $\sigma_X^2$ i $\sigma_Y^2$ poznate, ili $\sigma_X^2$ i $\sigma_Y^2$ nepoznate i obimi uzoraka veliki.
Za testiranje navedene nulte hipoteze koristi se test statistika

$$Z_0 = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}}, \text{ odnosno }, Z_0 = \frac{\overline{X} - \overline{Y}}{\sqrt{\widehat{D}(\overline{X} - \overline{Y})}},$$

gde je

$$\widehat{D}(\overline{X} - \overline{Y}) = \frac{\overline{S}_{n_X}^2}{n_X} + \frac{\overline{S}_{n_Y}^2}{n_Y} \quad \text{i} \quad \overline{X} \equiv \overline{X}_{n_X} \quad , \quad \overline{Y} \equiv \overline{Y}_{n_Y}$$

od kojih se prva koristi za poznate disperzije i ima tačno normalnu normiranu raspodelu, a druga ima približno $\mathcal{N}(0,1)$ raspodelu ako je hipoteza $H_0$ tačna. Kritične oblasti veličine $\alpha$ za odgovarajuće alternativne hipoteze $H_1$ date su tabelom:

| $H_0$ | $H_1$ | $C$ |
|---|---|---|
| $m_X = m_Y$ | $m_X \neq m_Y$ | $\lvert z_0 \rvert \geq z_{0,5-\alpha/2}$ |
| $m_X = m_Y$ | $m_X > m_Y$ | $z_0 \geq z_{0,5-\alpha}$ |
| $m_X = m_Y$ | $m_X < m_Y$ | $z_0 \leq -z_{0,5-\alpha}$ |

.

(b) $\sigma_X^2$ i $\sigma_Y^2$ nepoznate i $\sigma_X^2 = \sigma_Y^2$. Koristi se test statistika

$$t_0 = \frac{(\overline{X} - \overline{Y})\sqrt{\frac{n_X \cdot n_Y}{n_X + n_Y}(n_X + n_Y - 2)}}{\sqrt{(n_X - 1)\widetilde{S}_X^2 + (n_Y - 1)\widetilde{S}_Y^2}}, \tag{4.4}$$

koja, pod navedenim uslovima, ima približno $t_{n_X+n_Y-2}$ raspodelu. Tabela odgovarajućih kritičnih oblasti izgleda ovako:

| $H_0$ | $H_1$ | $C$ |
|---|---|---|
| $m_X = m_Y$ | $m_X \neq m_Y$ | $\lvert t_0 \rvert \geq t_{n_X+n_Y-2;0,5-\alpha/2}$ |
| $m_X = m_Y$ | $m_X > m_Y$ | $t_0 \geq t_{n_X+n_Y-2;0,5-\alpha}$ |
| $m_X = m_Y$ | $m_X < m_Y$ | $t_0 \leq -t_{n_X+n_Y-2;0,5-\alpha}$ |

.

Testiranje u okviru ove tačke se može vršiti i u slučaju nepoznatih, a različitih disperzija $\sigma_X^2$ i $\sigma_Y^2$. Test statistika je ponovo oblika (4.4), ali se granice kritičnih oblasti odredjuju tzv. aproksimacijom Kohrena (Cochren) o kojoj ovde neće biti reči.

**Primer 61.** Posmatrane su dve grupe radnika jedne fabrike i meren je njihov koeficijent inteligencije. Za prvu grupu od 16 radnika dobijeno je da je $\overline{x}_{16} = 114$ i $\overline{s}_X = 82$. Za drugu grupu od 14 radnika dobijeno je da je $\overline{y}_{14} = 121$ i $\overline{s}_Y = 60$. Da li postoje bitne razlike izmedju srednjih vrednosti koeficijenata inteligencije ovih dveju grupa radnika ako je prag značajnosti $\alpha = 0,05$ ?

Testira se hipoteza $H_0(m_1 = m_2)$ protiv hipoteze $H_1(m_1 \neq m_2)$, pri čemu su disperzije nepoznate. Test statistika ima realizovanu vrednost $-2,258$. Kako je $t_{28;0,475} = 2,048$, to je kritična oblast $C = (-\infty; -2,048] \cup [2,048; +\infty)$. Kako $-2,258 \in C$, to se hipoteza $H_0$ odbacuje, tj. zaključuje se da postoje bitne razlike izmedju koeficijenata inteligencije posmatranih dveju grupa radnika. (Značajnost ovog testa je 0,032 što je manje od 0,05.) $\triangle$

3. Testira se hipoteza o jednakosti srednjih vrednosti dvaju obeležja $X$ i $Y$ posmatranih istovremeno na istoj populaciji sa pretpostavljenim raspodelama:

$$X \ : \ \mathcal{N}(m_X, \sigma_X^2), \quad Y \ : \ \mathcal{N}(m_Y, \sigma_Y^2)$$

na osnovu dvodimenzionog uzorka obima $n$. Najčešće se radi o tome da se, zapravo, na istim jedinkama (ljudima, životinjama) utvrdjuju vrednosti jednog ispitivanog obeležja pri postojanju različitih uslova izvodjenja eksperimenta, pa se rezultati jednog merenja označe sa $X^{(1)}$, a drugog sa $X^{(2)}$. "Rezultat" oba izvršena merenja na uzorku obima $n$ je slučajni vektor:

$$\left((X_1^{(1)}, X_1^{(2)}), (X_2^{(1)}, X_2^{(2)}), \ldots, (X_n^{(1)}, X_n^{(2)})\right)$$

Testira se hipoteza:

$$H_0(m_1 = m_2), \text{ odnosno, } \quad H_0(m_1 - m_2 = 0)$$

i kaže se da se radi o testu za matematičko očekivanje kod sparenih uzoraka. Test statistika koja se pri tome koristi

$$t_0 = \frac{\overline{D}_n}{\sqrt{\frac{\sum_{i=1}^n (D_i - \overline{D}_n)^2}{n-1}}},$$

gde je

$$\overline{D}_n = \frac{1}{n}\sum_{i=1}^n D_i\,, \ D_i = X_i^{(1)} - X_i^{(2)}\,,$$

ima Studentovu raspodelu sa $n-1$ stepeni slobode. Kritične oblasti se odredjuju kao kod testa za matematičko očekivanje normalne raspodele sa nepoznatom disperzijom i malim obimom uzorka.

**Primer 62.** Neka je na grupi od 10 ljudi meren broj pozitivnih reakcija pod dejstvom dva stresora i to najpre fizičke prirode, pri čemu je stres izazivan električnim šokom, a zatim psihološke prirode - glasna muzika. U tabeli je dat broj pozitivnih reakcija:

| Stresor \ Osoba | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Elektrošok | 6 | 8 | 4 | 8 | 6 | 4 | 5 | 5 | 6 | 7 |
| Glasna muzika | 8 | 9 | 9 | 12 | 9 | 7 | 9 | 9 | 8 | 11 |
| $d_i$ | $-2$ | $-1$ | $-5$ | $-4$ | $-3$ | $-3$ | $-4$ | $-4$ | $-2$ | $-4$ |

.

Testirati hipotezu o jednakosti matematičkih očekivanja za $\alpha = 0,05$.

Dobija se da je $\overline{d}_{10} = -3,2$ i $\sum(d_i - \overline{d}_{10})^2 = 13,6$, tako da test statistika ima realizovanu vrednost

$$t_0 = \frac{-3,2}{\sqrt{13,6}} \cdot \sqrt{9} = -2,603.$$

Kritična oblast je

$$C = (-\infty; -2,262] \cup [2,262; +\infty)$$

i kako $-2,603 \in C$, to se hipoteza $H_0$ odbacuje, tj. zaključuje se da ima razlike u očekivanom broju pozitivnih reakcija pod dejstvom fizičkog i psihološkog stresora kod posmatrane grupe ispitanika. (Značajnost ovog testa je 0,029, dakle manja od 0,05.) $\triangle$

4. Test koji se odnosi na testiranje disperzije obeležja sa normalnom raspodelom ima nultu hipotezu:

$$H_0(\sigma^2 = \sigma_0^2),$$

gde je $\sigma_0^2$ fiksiran pozitivan realan broj. U tom slučaju, test statistika je

$$\chi_0^2 = \frac{(n-1)\tilde{S}_n^2}{\sigma_0^2}$$

koja ima približno $\chi^2$ raspodelu sa $n-1$ stepeni slobode: $\chi^2_{n-1}$.

Kritične oblasti veličine $\alpha$, za različite alternativne hipoteze, date su u tabeli:

| $H_0$ | $H_1$ | $C$ |
|---|---|---|
| $\sigma^2 = \sigma_0^2$ | $\sigma^2 \neq \sigma_0^2$ | $\chi_0^2 \leq \chi^2_{n-1;\frac{\alpha}{2}} \vee \chi_0^2 \geq \chi^2_{n-1;1-\frac{\alpha}{2}}$ |
| $\sigma^2 = \sigma_0^2$ | $\sigma^2 > \sigma_0^2$ | $\chi_0^2 \geq \chi^2_{n-1;1-\alpha}$ |
| $\sigma^2 = \sigma_0^2$ | $\sigma^2 < \sigma_0^2$ | $\chi_0^2 \leq \chi^2_{n-1;\alpha}$ |

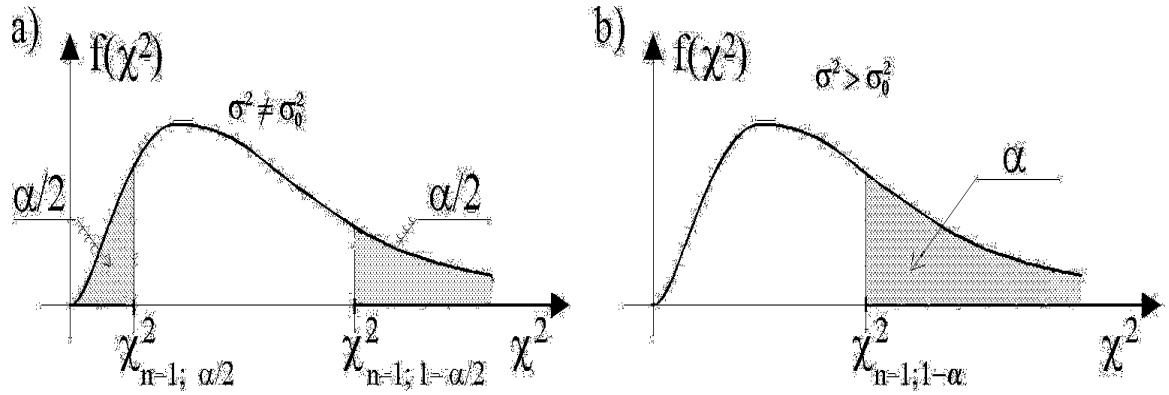
.

Deo apscisne ose ispod šrafirane površine je kritična oblast veličine $\alpha$ i na slici 4.2 a) odgovara oblasti odbacivanja razmatrane nulte hipoteze protiv alternativne $H_1(\sigma^2 \neq \sigma_0^2)$ i na odgovarajući način 4.2 b).

Navedena statistika se koristi za slučaj nepoznatog matematičkog očekivanja. Za slučaj poznatog $m$, koristi se test statistika

$$\chi_0^2 = \frac{\sum_{i=1}^n (X_i - m)^2}{\sigma_0^2}$$

koja ima $\chi^2$ raspodelu sa $n$ stepeni slobode, pa se u tom smislu i kritične oblasti razlikuju od gore navedenih, tj. razlikuju se samo u broju stepeni slobode.




Slika 4.2: Testiranje disperzije ($\sigma^2$) obeležja za velike uzorke: kritične oblasti za $H_0(\sigma^2 = \sigma_0^2)$ protiv alternativnih hipoteza a) $H_1(\sigma^2 \neq \sigma_0^2)$, i b) $H_1(\sigma^2 > \sigma_0^2)$.

**Primer 63.** Merenjem koeficijenta inteligencije 50 učenika dobijeno je da je disperzija jednaka $\overline{s}_{50}^2 = 2,45$. Testirati hipotezu da je standardno odstupanje veće od 2 za prag značajnosti $\alpha = 0,05$.

Testira se hipoteza $H_0(\sigma^2 = 4)$ protiv hipoteze $H_1(\sigma^2 > 4)$. Kako je $\chi^2_{49;0,95} = 67,5$, to je kritična oblast $C = [67,5; +\infty)$. Test statistika ima realizovanu vrednost $(49 \cdot 2,45)/4 = 30,01$ i ona ne pripada oblasti $C$, tako da se hipoteza $H_0$ prihvata. (Značajnost ovog testa je 0,99, dakle, veća od 0,05.) $\triangle$

5. Često se ukazuje potreba za uporedjivanjem dva obeležja po njihovim disperzijama. Koriste se dva nezavisna uzorka $(X_1, \ldots, X_{n_X})$ i $(Y_1, \ldots, Y_{n_Y})$ obeležja $X$ i $Y$ čije su raspodele redom $\mathcal{N}(m_X, \sigma_X^2)$ i $\mathcal{N}(m_Y, \sigma_Y^2)$. Testira se nulta hipoteza

$$H_0(\sigma_X^2 = \sigma_Y^2)$$

protiv odgovarajuće složene hipoteze, kao i u prethodnim testovima. U slučaju da su $m_X$ i $m_Y$ poznate veličine, koristi se test statistika

$$F_0 = \frac{n_X \sum\limits_{i=1}^{n_Y} (Y_i - m_Y)^2}{n_Y \sum\limits_{i=1}^{n_X} (X_i - m_X)^2}$$

koja ima Fišerovu raspodelu $F_{n_Y, n_X}$. Kada $m_X$ i $m_Y$ nisu poznate, koristi se

$$F_0 = \frac{\widetilde{S}_{n_Y}^2}{\widetilde{S}_{n_X}^2}$$

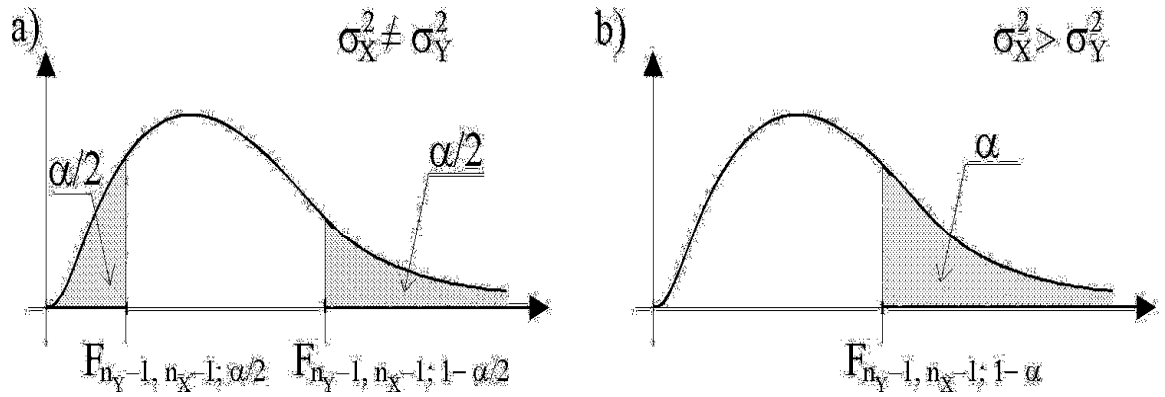
koja ima približno Fišerovu raspodelu $F_{n_Y-1, n_X-1}$. Tabela odgovarajućih kritičnih oblasti u poslednjem slučaju je:

| $H_1$ | $C$ |
|---|---|
| $\sigma_X^2 \neq \sigma_Y^2$ | $F_0 \leq F_{n_Y-1,n_X-1;\frac{\alpha}{2}} \vee F_0 \geq F_{n_Y-1,n_X-1;1-\frac{\alpha}{2}}$ |
| $\sigma_X^2 > \sigma_Y^2$ | $F_0 \geq F_{n_Y-1,n_X-1;1-\alpha}$ |
| $\sigma_X^2 < \sigma_Y^2$ | $F_0 \leq F_{n_Y-1,n_X-1;\alpha}$ |

.

dok za slučaj poznatih $m_X$ i $m_Y$ treba na odgovarajući način prilagoditi broj stepeni slobode statistika u tabeli.

Kritične oblasti za odgovarajuće složene alternativne hipoteze prikazane su na slici 4.3.




Slika 4.3: Kritične oblasti kod primene Fišerove raspodele za testiranje nulte protiv alternativnih hipoteza a) $H_1(\sigma_X^2 \neq \sigma_Y^2)$, i b) $H_1(\sigma_X^2 > \sigma_Y^2)$.

Ovaj način testiranja koristi se i na istoj populaciji za dva nezavisna uzorka ako treba da se utvrdi da li je došlo do promene raspodele obeležja na populaciji pod dejstvom nekog procesa, recimo pre i posle rada i slično.

6. Testiranje koeficijenta korelacije

Za slučajni vektor $(X, Y)$ čija raspodela pripada familiji dvodimenzionih normalnih raspodela $\{\mathcal{N}(m_X, m_Y, \sigma_X^2, \sigma_Y^2, \rho), m_X \in R, m_Y \in R, \sigma_X^2 \in R^+, \sigma_Y^2 \in R^+, \rho \in [0,1]\}$ testira se nulta hipoteza

$$H_0 : \rho = \rho_0, \quad -1 < \rho_0 < 1$$

protiv odgovarajućih alternativnih. Razlikuju se dva slučaja: $\rho_0 = 0$ i $\rho_0 \neq 0$. Ovo je posledica različitih test statistika impliciranih pomenutim vrednostima koeficijenta korelacije. Dakle, za dva obeležja testira se postojanje linearne veze medju njima na sledeći način:

(a) Testiranje nulte hipoteze

$$H_0 : \rho = 0$$

na osnovu uzoračkog koeficijenta korelacije i uzorka obima $n$.

Pod pretpostavkom da je nulta hipoteza tačna, test statistika

$$t_0 = \frac{R_{XY}\sqrt{n-2}}{\sqrt{1-R_{XY}^2}}$$

ima Studentovu raspodelu sa $n-2$ stepena slobode. $R_{XY}$ je, kao i do sada, uzoračka koeficijent korelacije. Otuda je tabela odgovarajućih kritičnih oblasti za prag značajnosti $\alpha$:

| $H_0$ | $H_1$ | $C$ |
|---|---|---|
| $\rho = 0$ | $\rho \neq 0$ | $\lvert t_0\rvert \geq t_{n-2;0,5-\alpha/2}$ |
| $\rho = 0$ | $0 < \rho < 1$ | $t_0 \geq t_{n-2;0,5-\alpha}$ |
| $\rho = 0$ | $-1 < \rho < 0$ | $t_0 \leq -t_{n-2;0,5-\alpha}$ |

.

Treba se podsetiti ranije iznete činjenice da ukoliko vektor $(X, Y)$ ima dvodimenzionalnu normalnu raspodelu, saznanje o tome da je $H_0 : \rho = 0$ tačna, znači ne samo nekorelisanost, već i nezavisnost obeležja $X$ i $Y$.

**Primer 64.** Grupa od 16 studenata pokazala je na ispitu iz matematike sledeći uspeh:

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Pismeni | 90 | 90 | 80 | 90 | 92 | 88 | 90 | 63 |
| Usmeni | 84 | 84 | 82 | 94 | 90 | 85 | 89 | 62 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Pismeni | 70 | 54 | 78 | 86 | 99 | 84 | 56 | 85 |
| Usmeni | 65 | 52 | 72 | 90 | 98 | 89 | 58 | 85 |

.

Da li je na 5% pragu značajnosti koeficijent korelacije blizak nuli ?

Testira se hipoteza $H_0(\rho = 0)$ protiv hipoteze $H_1(\rho \neq 0)$. Odgovarajuće uzoračke sredine su $\overline{x}_{16} = 80,94$ i $\overline{y}_{16} = 79,94$, dok su uzoračke disperzije $\overline{s}_X = 166,9$ i $\overline{s}_Y = 177,66$, respektivno. Uzoračka koeficijent korelacije ima vrednost $r_{XY} = 0,964$. Prema tome, vrednost test statistike je

$$t_0 = \frac{0,964\sqrt{14}}{\sqrt{1-0,964^2}} = 13,565\,.$$

Kritična oblast je $C = (-\infty; -2,145] \cup [2,145; +\infty)$ i kako 13,565 pripada oblasti $C$, to se hipoteza $H_0$ odbacuje. $\triangle$

(b) Testiranje nulte hipoteze

$$H_0 : \rho = \rho_0\,,$$

gde je $\rho_0 \neq 0$, tj. $\rho_0 \in (-1, 0) \cup (0, 1)$, takodje na bazi uzorka obima $n$.

Koristi se test statistika

$$Z = \frac{1}{2}\ln\left(\frac{1+R_{XY}}{1-R_{XY}}\right)$$

koja, pod pretpostavkom da je nulta hipoteza tačna, ima približno normalnu raspodelu

$$Z : \mathcal{N}\left(\frac{1}{2}\ln\left(\frac{1+\rho_0}{1-\rho_0}\right) + \frac{\rho_0}{2(n-1)}\,,\ \frac{1}{n-3}\right).$$

Standardizovanjem ovakve slučajne promenljive omogućeno je korišćenje tablice za normalnu normiranu raspodelu i tabela odgovarajućih kritičnih oblasti veličine $\alpha$ kao u prethodnom slučaju, gde je $z_0$ realizovana vrednost statistike

$$Z_0 = \frac{Z - \frac{1}{2}\ln(\frac{1+\rho_0}{1-\rho_0}) - \frac{\rho_0}{2(n-1)}}{1/\sqrt{n-3}}\,.$$

### 4.2.3 Testiranje parametra binomne raspodele

Ako treba testirati verovatnoću $p$ realizacije nekog dogadjaja $A$ preko uzorka obima $n$, zapravo se vrši testiranje hipoteze o parametru binomne raspodele, tj. testiranje nulte hipoteze $H_0(p = p_0)$ protiv svih alternativnih, pod pretpostavkom da je uzorak uzet iz populacije sa obeležjem $S_n : \mathcal{B}(n, p)$. Za mali obim uzorka, kritične oblasti se odredjuju direktno iz definicije binomne raspodele. Medjutim, za veliki obim uzorka, što ovde podrazumeva $n > 50$ i $np_0 > 10$, koristi se normalna aproksimacija binomne raspodele i statistika

$$Z_0 = \frac{S_n - np_0}{\sqrt{np_0(1-p_0)}}$$

koja ima približno $\mathcal{N}(0, 1)$ raspodelu. Dakle:

| $H_0$ | $H_1$ | $C$ |
|---|---|---|
| $p = p_0$ | $p \neq p_0$ | $\lvert z_0\rvert \geq z_{0,5-\alpha/2}$ |
| $p = p_0$ | $p > p_0$ | $z_0 \geq z_{0,5-\alpha}$ |
| $p = p_0$ | $p < p_0$ | $z_0 \leq -z_{0,5-\alpha}$ |

.

**Primer 65.** Anketom se ispituju šanse jednog kandidata na izborima. Medju 100 slučajno izabranih glasača 55 njih se izjasnilo da bi glasalo za tog kandidata. Neka je $p$ verovatnoća da je slučajno izabrani anketirani simpatizer posmatranog kandidata. Testirati hipotezu da će posmatrani kandidat dobiti 50% glasova celokupnog biračkog tela protiv svih alternativnih. Dakle, testira se nulta hipoteza $H_0(p = 0,5)$ i neka je $\alpha = 0,01$, protiv: a) $H_1(p \neq 0,5)$, b) $H_1(p > 0,5)$ i c) $H_1(p < 0,5)$.

Realizovana vrednost test statistike je

$$z_0 = \frac{55 - 100 \cdot 0,5}{\sqrt{100 \cdot 0,5 \cdot 0,5}} = 1.$$

(a) Dobija se da je $z_{0,495} = 2,575$, tako da je kritična oblast $C = (-\infty; -2,575] \cup [2,575; +\infty)$. Kako $1 \notin C$, to se hipoteza $H_0$ prihvata. (Značajnost ovog testa je 0,317, dakle, veća od 0,01) To istovremeno znači da nema smisla dalje vršiti testiranja protiv preostale dve alternativne hipoteze. $\triangle$

## 4.3 Neparametarski testovi

Dve su vrste problema koji se najčešće rešavaju neparametarskim testovima:

1. problem jednog uzorka – ispituju se:
   (a) parametri raspodele (pre svega kvantili),
   (b) slučajnost uzorka, i
   (c) saglasnost uzorka sa pretpostavljenom raspodelom.
2. problem dva uzorka – ispituju se:
   (a) zavisnost dva obeležja,
   (b) uporedjuju se raspodele dva obeležja, i sl.

### 4.3.1 Test Kolmogorov – Smirnova

Test Kolmogorov–Smirnova se koristi samo kod obeležja apsolutno neprekidnog tipa, tj. kod takvih obeležja kod kojih je funkcija raspodele $F$ neprekidna. Testira se nulta hipoteza

$$H_0 : F(x) = F_0(x)\,, \quad x \in R\,,$$

gde je $F_0$ neka odredjena, takodje neprekidna, funkcija raspodele. Test Kolmogorov–Smirnova direktno primenjuje centralnu teoremu matematičke statistike, te na osnovu uzorka obima $n$, $(X_1, \ldots, X_n)$, odredjuje empirijsku funkciju raspodele $S_n(x)$, $x \in R$, i definiše statistiku

$$D_n = \sup_{-\infty<x<+\infty} |S_n(x) - F_0(x)|.$$

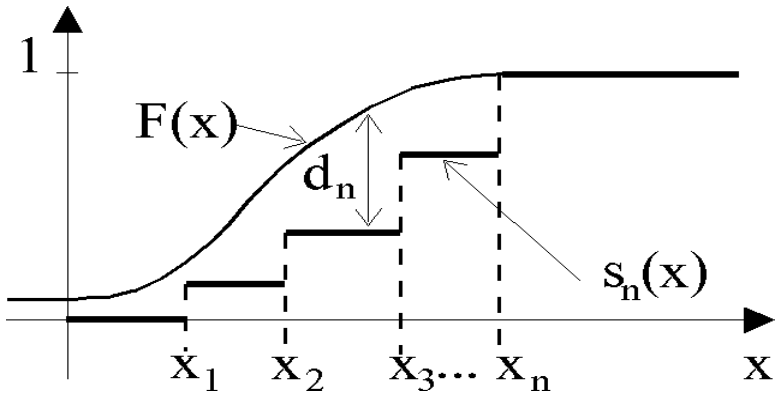
Pod pretpostavkom da je nulta hipoteza tačna, statistika $D_n$ ima raspodelu Kolmogorova. Na osnovu realizovanog uzorka $(x_1, \ldots, x_n)$ treba odrediti realizovanu vrednost statistike $D_n$:

$$d_n = \sup_{-\infty<x<+\infty} |s_n(x) - F_0(x)|.$$

Realizovana vrednost empirijske funkcije raspodele je stepenasta funkcija $s_n(x)$, $x \in R$, i ima konačan broj "stepenika" (najviše $n+1$, u slučaju da su svi elementi realizovanog uzorka različiti). Uočimo još jednom da su $s_n$ i $F$ monotono neopadajuće funkcije. Odredjivanje supremuma, u praksi, se, po pravilu, svodi na odredjivanje maksimuma apsolutnih razlika $|s_n(x-0) - F_0(x-0)|$ (zbog neprekidnosti s desna) na segmentima definisanim uzorkom (slika 4.4).

Granica kritične oblasti za zadati prag značajnosti $\alpha$: $d_{n,1-\alpha}$ čita se iz tablice Kolmogorova. Kritična oblast odredjuje se na sledeći način:

| $H_0$ | $H_1$ | $C$ |
|---|---|---|
| $F = F_0$ | $F \neq F_0$ | $d_n \geq d_{n,1-\alpha}$ |




Slika 4.4: Odredjivanje vrednosti $d_n$ kod testa Kolmogorov–Smirnova.

**Primer 66.** Sledeća tabela prikazuje rezultate testa inteligencije 53 dečaka:

| IQ | Br. dece |
|---|---|
| [65,75) | 1 |
| [75,85) | 2 |
| [85,95) | 10 |
| [95,105) | 12 |
| [105,115) | 14 |
| [115,125) | 11 |
| [125,135] | 3 |

.

Ispitati saglasnost ovih podataka sa normalnim zakonom raspodele koristeći test Kolmogorov–Smirnova sa 1% pragom značajnosti.

Neka je obeležje $X$ koeficijent inteligencije deteta. Testira se hipoteza

$$H_0(\text{podaci su saglasni sa } \mathcal{N}(m, \sigma^2) \text{ raspodelom}).$$

Parametri raspodele $m$ i $\sigma^2$ su nepoznati. Parametar $m$ se ocenjuje uzoračkom sredinom a $\sigma^2$ uzoračkom disperzijom. Tako se dobija $\hat{m} = 105,28$ i $\hat{\sigma}^2 = 184,047$. Znači, testira se hipoteza $H_0($ podaci su saglasni sa $\mathcal{N}(105,28; 184,047)$ raspodelom). Dalji postupak je sadržan u sledećoj tabeli:

| $x_i + 0$ | $n_i$ | $\sum_i$ | $s_{53}(x_i - 0)$ | $t_i$ | $F_0(x_i) = 0,5 \pm \Phi(\lvert t_i \rvert)$ | $\lvert s_{53}(x_i - 0) - F_0(x_i) \rvert$ |
|---|---|---|---|---|---|---|
| 75 | 1 | 1 | 0,019 | -2,232 | 0,0129 | 0,0061 |
| 85 | 2 | 3 | 0,057 | -1,495 | 0,0681 | 0,0111 |
| 95 | 10 | 13 | 0,245 | -0,758 | 0,2236 | 0,0214 |
| 105 | 12 | 25 | 0,472 | -0,021 | 0,4920 | 0,0200 |
| 115 | 14 | 39 | 0,736 | 0,716 | 0,7642 | 0,0282 |
| 125 | 11 | 50 | 0,943 | 1,454 | 0,9265 | 0,0165 |
| 135 | 3 | 53 | 1 | 2,191 | 0,9857 | 0,0143 |

.

Ovde je $n_i$ apsolutna učestanost $i$-tog intervala, $\sum_i$ zbirna učestanost do tog intervala (uključujući i taj interval), a $t_i = \frac{(x_i - 0) - \overline{x}_{53}}{\sqrt{s^2_{53}}}$. Primećuje se da je maksimalna razlika $|s_{53}(x-0) - F_0(x)| \equiv |s_{53}(x-0) - F_0(x-0)|$ jednaka $0,0282$, tako da je $d_{53} = 0,0282$. Kako je $1 - \alpha = 0,99$, to iz tablice Kolmogorova sledi da je $d_{53;0,99} = 0,23$, tako da je

kritična oblast $C = [0,23; +\infty)$. Kako $0,0280 \notin C$, to se hipoteza $H_0$ prihvata, odnosno nema razloga da se odbaci. Dakle, na osnovu rezultata testa, može se tvrditi da je IQ normalno raspodeljen na populaciji dečaka ispitivanog uzrasta. $\triangle$

Test Kolmogorov–Smirnova koristi se i za testiranje jednakosti raspodela dvaju obeležja $X$ i $Y$ apsolutno neprekidnog tipa na osnovu nezavisnih uzoraka $(X_1, X_2, \ldots, X_{n_1})$ i $(Y_1, Y_2, \ldots, Y_{n_2})$. Koristi se statistika

$$D_{\frac{n_1 n_2}{n_1+n_2}} = \sup_{-\infty<x<+\infty} |S_X(x) - S_Y(x)|$$

za testiranje nulte hipoteze

$$H_0 : F_X = F_Y$$

protiv alternativne

$$H_1 : F_X \neq F_Y,$$

gde su $S_X(x)$ i $S_Y(x)$, $x \in R$, odgovarajuće empirijske funkcije raspodele obeležja $X$ i $Y$ na osnovu posmatranih uzoraka. I ovde se u praksi supremum zamenjuje maksimumom. Odgovarajuća kritična oblast veličine $\alpha$ je:

| $H_0$ | $H_1$ | $C$ |
|---|---|---|
| $F_X = F_Y$ | $F_X \neq F_Y$ | $d_{\frac{n_1 n_2}{n_1+n_2}} \geq d_{\frac{n_1 n_2}{n_1+n_2},1-\alpha}$ |

.

**Primer 67.** Slučajno izabrani dečaci iz dve škole podvrgnuti su testu agresivnosti. Dobijeni su sledeći rezultati:

| Broj poena na testu | [75,85) | [85,95) | [95,105) | [105,115) | [115,125) | [125,135] |
|---|---|---|---|---|---|---|
| Br. dečaka I škole | 3 | 10 | 12 | 14 | 11 | 3 |
| Br. dečaka II škole | 0 | 2 | 13 | 30 | 5 | 1 |

.

Testirati hipotezu da su uzorci iz populacije sa istom raspodelom obeležja sa pragom značajnosti 0,05.

Testira se hipoteza $H_0(F_X = F_Y)$ protiv hipoteze $H_1(F_X \neq F_Y)$. Postupak računanja dat je u tabeli:

| $x_i + 0$ | $n_i$ | $\sum_i$ | $s_X(x_i - 0)$ | $m_i$ | $\sum_i$ | $s_Y(x_i - 0)$ | $\lvert s_X(x_i - 0) - s_Y(x_i - 0)\rvert$ |
|---|---|---|---|---|---|---|---|
| 85 | 3 | 3 | 0,0566 | 0 | 0 | 0 | 0,0566 |
| 95 | 10 | 13 | 0,2453 | 2 | 2 | 0,0392 | 0,2061 |
| 105 | 12 | 25 | 0,4717 | 13 | 15 | 0,2941 | 0,1776 |
| 115 | 14 | 39 | 0,7358 | 30 | 45 | 0,8824 | 0,1466 |
| 125 | 11 | 50 | 0,9434 | 5 | 50 | 0,9804 | 0,0370 |
| 135 | 3 | 53 | 1 | 1 | 51 | 1 | 0 |

.

Kako je $(53 \cdot 51)/(53 + 51) = 25,99$ a to približno jednako 26, to je $d_{26} = 0,2061$. S druge strane, kritična oblast je oblika $C = [d_{26;0,95}; +\infty) = [0,264; +\infty)$. Vrednost 0,2061 ne pripada kritičnoj oblasti $C$, što znači da se hipoteza $H_0$ prihvata, tj. uzorci su sa istom raspodelom obeležja (za dati prag značajnosti). Odnosno, može se smatrati da u stepenu agresivnosti kod dečaka dveju ispitivanih škola nema razlike. $\triangle$

### 4.3.2 Pirsonov $\chi^2$ test

Jedan od najčešće primenjivanih neparametarskih testova je Pirsonov ili $\chi^2$ test.

Test nosi naziv po svom autoru Karlu Pirsonu, koji ga je definisao i uveo u statističku praksu 1900. godine. Njegov alternativni naziv, $\chi^2$ test, potiče od raspodele test statistike kojom se koristi. Ovde ćemo samo približiti ideju o raspodeli test statistike, a nećemo se baviti dokazom upravo navedene tvrdnje.

Podjimo od slučajne promenljive u oznaci $X_1$ sa binomnom raspodelom $\mathcal{B}(n, p_1)$. Standardizovana slučajna promenljiva $X_1^* = \frac{X_1 - np_1}{\sqrt{np_1(1-p_1)}}$ ima asimptotski (prema Muavr–Laplasovoj teoremi) normalnu normiranu raspodelu. Otuda, kada $n \to \infty$, slučajna promenljiva $Q_1 = (X_1^*)^2$ ima asimptotski $\chi^2$–raspodelu sa 1 stepenom slobode, $\chi_1^2$. Uvodeći novu slučajnu promenljivu $X_2 = n - X_1$ i parametar $p_2 = 1 - p_1$, slučajna promenljiva $Q_1$ se može da predstavi kao

$$Q_1 = \frac{(X_1 - np_1)^2}{np_1(1-p_1)} = \frac{(X_1 - np_1)^2}{np_1} + \frac{(X_2 - np_2)^2}{np_2}\,.$$

Dakle, zbir ovako definisanih slučajnih promenljivih ima asimptotski $\chi_1^2$–raspodelu kada se $n$ uvećava (recimo za $n \geq 50$).

Posmatrajmo sada slučajni vektor $(X_1, X_2, \ldots, X_{k-1})$ dimenzije $k-1$ sa multinomnom raspodelom $\mathcal{M}(n, p_1, p_2, \ldots, p_{k-1})$. Definišimo slučajnu promenljivu $X_k = n - (X_1 + \cdots + X_{k-1})$ i parametar $p_k = 1 - (p_1 + \cdots + p_{k-1})$. Može se pokazati da slučajna promenljiva

$$Q_{k-1} = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i}\,, \quad np_i \geq 5 \text{ za svako } i = 1, \ldots, k$$

konvergira u raspodeli ka slučajnoj promenljivoj sa $\chi_{k-1}^2$ raspodelom. U literaturi se može da nadje upozorenje da je ova aproksimacija dobra tek kada je $n$ dovoljno veliko tako da svako $np_i$, $i = 1, \ldots, k$ bude najmanje 5.

Slučajna promenljiva $Q_{k-1}$ služi kao osnov za definiciju Pirsonovog testa.

Pretpostavimo da se uzorački prostor nekog eksperimenta razbija na konačan broj medjusobno disjunktnih skupova $A_1, \ldots, A_k$. Neka je $P(A_i) = p_i$, $i = 1, \ldots, k$, gde je $p_k = 1 - (p_1 + \cdots + p_{k-1})$, što znači da je $p_i$ verovatnoća da je ishod ovog slučajnog eksperimenta u skupu $A_i$. Pretpostavljamo da se slučajni eksperiment ponavlja $n$ nezavisnih puta pod istim uslovima, pa ćemo slučajnom promenljivom $X_i$ da označimo koliko je puta ishod eksperimenta pripao skupu $A_i$. Drugim rečima, $X_1, \ldots, X_k$, $X_k = n - (X_1 + \cdots + X_{k-1})$ su apsolutne učestanosti sa kojima ishod eksperimenta pripada respektivno skupovima $A_1, \ldots, A_k$. Tada je zajednička gustina raspodele za $X_1, \ldots, X_{k-1}$ multinomna sa parametrima $n, p_1, \ldots, p_{k-1}$. Nadalje se testiranje odnosi na nultu hipotezu o pomenutoj multinomnoj raspodeli:

$$H_0 : p_1 = p_{01}, p_2 = p_{02}, \ldots, p_{k-1} = p_{0,k-1}\,,$$

protiv svih alternativnih, gde su $p_{01}, \ldots, p_{0,k-1}$ brojevi, $0 < p_{0i} < 1$, $i = 1, 2, \ldots, k-1$ i $p_{01} + p_{02} + \cdots + p_{0,k-1} < 1$.

Jasno je sada u kakvoj je vezi ovaj test sa slučajnom promenljivom $Q_{k-1}$. Ako je $H_0$ tačna hipoteza, slučajna promenljiva

$$Q_{k-1} = \sum_{i=1}^{k} \frac{(X_i - np_{0i})^2}{np_{0i}}$$

ima približno $\chi^2_{k-1}$–raspodelu.

Uočimo sledeće. Kada je $H_0$ tačna, onda je $np_{0i}$ očekivanje od $X_i$. Otuda je i intuitivno jasno da eksperimentalna vrednost slučajne promenljive $Q_{k-1}$ ne treba da je velika ako je $H_0$ tačna. Na osnovu ovoga, za unapred zadati prag značajnosti $\alpha$ odredjujemo granicu kritične oblasti $c$ kao

$$P\{Q_{k-1} \geq c\} = \alpha\,.$$

S obzirom na raspodelu kojoj teži, umesto oznake $Q_{k-1}$, ili neke druge, za ovu slučajnu promenljivu se koristi oznaka baš $\chi^2_{k-1}$.

Neki najkarakterističniji primeri primene $\chi^2$ testa izneti su u narednim odeljcima.

### A. Ispitivanje saglasnosti uzorka sa pretpostavljenom raspodelom

Kao i kod testa Kolmogorov–Smirnova, testira se hipoteza da je nepoznata raspodela $F$ posmatranog obeležja $X$ jednaka zadatoj – poznatoj raspodeli $F_0$, tj.

$$H_0 : F = F_0,$$

protiv alternativne, da su raspodele različite. Bitna razlika u odnosu na test Kolmogorov-Smirnova je u tome što se $\chi^2$ test ne ograničava samo na raspodele apsolutno neprekidnog tipa, već se može primeniti na bilo koju raspodelu $F_0$.

Postupak testiranja sprovodi se tako što se oblast vrednosti za $X$ deli na odredjen broj $(k)$ disjunktnih grupa (skupova), tj. realna prava se podeli na $k$ disjunktnih intervala $S_1$, $S_2$, ..., $S_k$, čija je unija skup $R$:

$$S_1, S_2, \ldots, S_k \subset R,\ \cup_{i=1}^{k} S_i = R,\ S_i \cap S_j = \emptyset \text{ za } i \neq j.$$

Sledeći korak je izračunavanje teorijskih verovatnoća $P\{X \in S_i\} = p_{0i}$, $i = 1, \ldots, k-1$ i $p_{0k} = 1 - (p_{01} + \cdots + p_{0,k-1})$, uz pretpostavku da je nulta hipoteza tačna. Zatim se sračunavaju teorijske apsolutne učestanosti u intervalima $S_i$ zadatog uzorka obima $n$:

$$\hat{n}_i = np_{0i}\,.$$

Drugim rečima, $\hat{n}_i$ je očekivani broj elemenata uzorka u svakom intervalu $S_i$ (otuda i oznaka $e_i$ koja se koristi u literaturi) pod uslovom da je nulta hipoteza tačna. On se uporedjuje sa realizovanim u eksperimentu brojem elemenata uzorka u tom intervalu, tj. sa $n_i$ (u literaturi se koristi i oznaka $o_i$). Ako je $H_0$ tačna, odstupanja $\hat{n}_i$ od $n_i$ ne bi smela da budu velika. Za meru odstupanja u svakoj grupi uzima se relativno odstupanje

$$\frac{(n_i - \hat{n}_i)^2}{\hat{n}_i}\,,$$

pa se kao test statistika koristi

$$\chi_0^2 = \sum_{i=1}^{k} \frac{(n_i - \hat{n}_i)^2}{\hat{n}_i} = \sum_{i=1}^{k} \frac{n_i^2}{\hat{n}_i} - n\,.$$

Ona za veliki obim uzorka $n$, pod pretpostavkom da je nulta hipoteza tačna i svi parametri raspodele $F_0$ u nultoj hipotezi poznati, ima približno $\chi^2$ raspodelu sa $k-1$ stepeni slobode.

Kritična oblast veličine $\alpha$ dobija se iz uslova

$$\alpha = P_{H_0}\{\chi_0^2 \geq c\},$$

jer realizacija dogadjaja $\{\chi_0^2 \geq c\}$ signalizira veliko odstupanje $F$ od $F_0$.

Dakle, naredna tabela prikazuju kritičnu oblast veličine $\alpha$ za razmatrani test.

| $H_0$ | $H_1$ | $C$ |
|---|---|---|
| $F = F_0$ | $F \neq F_0$ | $\chi_0^2 \geq \chi^2_{k-1,1-\alpha}$ |

.

Ova kritična oblast se odredjuje analogno sa onom prikazanom na slici 4.2 *b*).

Pri razbijanju skupa $R$ na intervale ne postoji strogo pravilo o broju intervala. U praksi se rukovodi logikom sredjivanja realizovanog uzorka, medjutim, ne treba birati nijedan interval $S_i$ u kome bi se dobilo $\hat{n}_i < 5$, odnosno takav interval treba priključiti prethodnom ili narednom intervalu.

Ukoliko neki od parametara raspodele $F_0$ treba oceniti na osnovu uzorka (parametar nije unapred poznat) da bi se odredile verovatnoće $p_{0i}$, onda za svaki procenjeni parametar treba smanjiti broj stepeni slobode za 1. Dakle, ako se ocenjuje ukupno $l$ parametara, broj stepeni slobode statistike $\chi_0^2$ je: $k - l - 1$.

**Primer 68.** Testiranjem 200 ispitanika dobijen je realizovani uzorak koji je, posle intervalnog sredjivanja, kako sledi:

| Broj bodova | [10,12) | [12,14) | [14,16) | [16,18) | [18,20) | [20,22] | $n$ |
|---|---|---|---|---|---|---|---|
| Broj ispitanika | 10 | 26 | 56 | 64 | 30 | 14 | 200 |

.

Sa pragom značajnosti $\alpha = 0,01$ testirati hipotezu da je raspodela rezultata testova normalna.

Parametre normalne raspodele $m$ i $\sigma^2$ treba oceniti, na osnovu uzorka, sredinom uzorka i disperzijom uzorka redom:

$$\bar{x}_{200} = 16,2 \quad \text{i} \quad \bar{s}^2_{200} = 6,08 \quad \Rightarrow \bar{s}_{200} = 2,47.$$

Skup realnih brojeva se deli na intervale: $S_1 = (-\infty, 12)$, $S_2 = [12, 14)$, $S_3 = [14, 16)$, $S_4 = [16, 18)$, $S_5 = [18, 20)$ i $S_6 = [20, +\infty)$. Ukoliko je nulta hipoteza tačna, tada je

$$\begin{aligned}
p_{01} &= P_{H_0}\{X \in (-\infty, 12)\} = \\
&= P_{H_0}\{-\infty < X^* < \frac{12 - 16,2}{2,47}\} = 0,0446, \\
p_{02} &= P_{H_0}\{X \in [12, 14)\} = 0,1421\,, \\
p_{03} &= 0,2814, \quad p_{04} = 0,2992, \quad p_{05} = 0,1709\,, \\
p_{06} &= 1 - p_{01} - p_{02} - p_{03} - p_{04} - p_{05} = 0,0618.
\end{aligned}$$

Dakle, $\hat{n}_1 = 200 \cdot 0,0446 = 8,92$, $\hat{n}_2 = 200 \cdot 0,1421 = 28,42$, $\hat{n}_3 = 56,28$, $\hat{n}_4 = 59,84$, $\hat{n}_5 = 34,18$ i $\hat{n}_6 = 12,36$, te je $\chi^2_{6-2-1} = 1,3562$.

Kako je kritična oblast definisanog testa $[11,3; +\infty)$, nema razloga da se nulta hipoteza odbaci; dakle, sa pragom značajnosti 0,01 nema značajne razlike raspodele posmatranog obeležja od normalne raspodele. (Značajnost ovog testa je 0,716). $\triangle$

**Primer 69.** Testira se hipoteza o normalnoj raspodeli zarada radnika jednog preduzeća prema slučajnom uzorku od 70 radnika iz dva pogona ($1-50$ prvi pogon, $51-70$ drugi pogon): 970, 650, 890, 1230, 680, 1010, 740, 480, 690, 820, 990, 860, 1040, 820, 1100, 540, 730, 670, 880, 530, 680, 790, 780, 850, 900, 700, 770, 890, 930, 1000, 1180, 1010, 850, 830, 940, 980, 740, 1110, 810, 840, 620, 790, 480, 990, 1060, 800, 700, 590, 920, 810, 1040, 710, 1100, 1070, 1010, 830, 1020, 760, 780, 1140, 700, 750, 900, 780, 970, 960, 710, 660, 410, 560.

| interval | $n_i$ | *) | $\hat{p}_{0i}$ | $\hat{n}_i = 70 \cdot \hat{p}_{0i}$ |
|---|---|---|---|---|
| $(-\infty;$ 600) | 7 | $(-\infty;$ -1,42) | 0,0778 | 5,4 |
| [600; 700) | 7 | [-1,42; -0,84) | 0,1226 | 8,6 |
| [700; 800) | 16 | [-0,84; -0,26) | 0,1970 | 13,8 |
| [800; 900) | 14 | [-0,26; 0,32) | 0,2281 | 16,0 |
| [900; 1000) | 11 | [0,32; 0,90) | 0,1904 | 13,3 |
| [1000; 1100) | 9 | [0,90; 1,48) | 0,1147 | 8,0 |
| [1100; $+\infty$) | 6 | [1,48; $+\infty$) | 0,0694 | 4,9 |
| $\Sigma$ | 70 | / | 1 | 70 |

*)–centrirani i normirani intervali

Ocenjuju se parametri $m$ i $\sigma^2$ normalne raspodele

$$\hat{m} = \overline{x}_{70} = \frac{1}{70}(550 \cdot 7 + 650 \cdot 7 + \ldots + 1150 \cdot 6) = 844,29$$

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \sqrt{\overline{s}_n^2} =$$

$$\sqrt{\frac{1}{70}(550^2 \cdot 7 + 650^2 \cdot 7 + \ldots + 1150^2 \cdot 6) - 844,29^2} = 172,28$$

$$\Rightarrow \hat{p}_{0i} : z_d = \frac{x_d - \hat{m}}{\hat{\sigma}};\ z_g = \frac{x_g - \hat{m}}{\hat{\sigma}},\quad \hat{p}_{0i} = \Phi(z_g) - \Phi(z_d)$$

$$\chi^2_0 = \frac{(7-5,4)^2}{5,4} + \frac{(7-8,6)^2}{8,6} + \ldots + \frac{(6-4,9)^2}{4,9} = 2,14.$$

Ima $k = 7$ intervala i $l = 2$ ocenjena parametra, znači da je broj stepeni slobode $k - l - 1 = 7 - 2 - 1 = 4$.

Osim uobičajenog načina zaključivanja o prihvatanju ili odbacivanju nulte hipoteze koji je korišćen i u prethodnom primeru, ovde će biti izložen još jedan, karakterističan za $\chi^2$ test:

S obzirom da je očekivanje slučajne promenljive $\chi^2_\nu$ jednako $\nu$,

$$E(\chi^2_\nu) = \nu,$$

to je bez obzira na prag značajnosti $\alpha$ moguć zaključak da odstupanja u uzorku od teorijski pretpostavljene raspodele nisu značajna ako je $\chi_0^2 < k - l - 1$, tj. manje od očekivane vrednosti test statistike i da u tom slučaju hipotezu $H_0$ treba prihvatiti.

Kako je u ovom primeru $\chi_0^2 = 2,14 < 4$, hipoteza $H_0$ se prihvata. $\triangle$

**Primer 70.** Izdvojena je grupa talentovanih učenika i beležen je njihov koeficijent inteligencije. Dobijeni su sledeći rezultati:

| IQ | $[120, 124)$ | $[124, 130)$ | $[130, 136)$ | $[136, 140]$ |
|---|---|---|---|---|
| Br. učenika | 10 | 50 | 35 | 5 |

.

Ispitati saglasnost ovih podataka sa $\chi^2$ raspodelom za $\alpha = 0,01$.

Broj stepeni slobode pretpostavljene $\chi^2$ raspodele treba oceniti, recimo sredinom uzorka jer je $E(\chi_\nu^2) = \nu$, pa se dobija $\hat{\nu} = \overline{x}_{100} = 129,15 \approx 129$. Testira se hipoteza $H_0(X$ ima $\chi_{129}^2$ raspodelu). Koristimo 4 intervala: $S_1 = (-\infty, 124)$, $S_2 = [124, 130)$, $S_3 = [130, 136)$ i $S_4 = [136, +\infty)$. Broj nepoznatih, tj. ocenjenih parametara je $l = 1$. Verovatnoće su:

$$p_{01} = P_{H_0}\{-\infty < X < 124\} = 0,392,$$

$$p_{02} = P_{H_0}\{124 \leq X < 130\} = 0,149,$$

$$p_{03} = P_{H_0}\{130 \leq X < 136\} = 0,139,$$

$$p_{04} = 1 - (p_{01} + p_{02} + p_{03}) = 0,32.$$

Broj stepeni slobode $\chi_0^2$ statistike je $4 - 1 - 1 = 2$, tako da je $\chi_2^2 = 159,25$. Kritična oblast je $C = \left[\chi_{2;0,99}^2, +\infty\right) = [9,21; +\infty)$. Kako $159,25 \in C$, to se hipoteza $H_0$ odbacuje, tj. zaključuje se da ovi podaci nisu saglasni sa $\chi_{129}^2$ raspodelom. $\triangle$

U praksi se za testiranje saglasnosti sa zadatom raspodelom često koristi formulacija ispitivanje saglasnosti očekivanih, $e_i$ i opserviranih vrednosti, $o_i$, pogotovu kod raspodela diskretnog tipa. U slučaju obeležja diskretnog tipa primenjuje se opšti princip koji je gore izložen na veoma jednostavan način. Ovaj slučaj se posebno ističe i zbog toga što obrazlaže kako se testira saglasnost raspodele kvalitativnog obeležja sa unapred pretpostavljenom diskretnom raspodelom.

**Primer 71.** Proizvodjač iznosi na tržište šest različitih žvakaćih guma u tipiziranom pakovanju. Na osnovu uzorka obima 60 u kome je registrovano 13, 18, 11, 8, 5, 5 prodatih komada po tipovima žvakaćih guma redom, testirati hipotezu da je verovatnoća prodaje za svaki tip žvakaće gume ista. Testiranje izvršiti sa pragom značajnosti $\alpha = 0.05$.

Neka je $A_i$, $i = 1, 2, \ldots, 6$ tip žvakaće gume. Tada je

$$H_0 : P(A_i) = p_{0i} = \frac{1}{6}, \quad i = 1, \ldots, 6\,,$$

pa je

$$e_i = np_{0i} = 60\frac{1}{6} = 10, \quad i = 1, \ldots, 6\,,$$

jer treba, zapravo, testirati saglasnost uzorka sa diskretnom uniformnom raspodelom.

Ako je $X_i$ učestanost sa kojom je dogadjaj $A_i$ ishod eksperimenta, tada je

$$\chi^2_5 = \sum_{i=1}^{5} \frac{(o_i - e_i)^2}{e_i} = \frac{(13-10)^2}{10} + \frac{(18-10)^2}{10} + \frac{(11-10)^2}{10} + \frac{(8-10)^2}{10} + \frac{(5-10)^2}{10} + \frac{(5-10)^2}{10} = 12,8$$

realizovana vrednost test statistike koja ima $\chi^2$–raspodelu sa 5 stepeni slobode. Granica kritične oblasti zadovoljava uslov

$$P(\chi^2_5 \geq 11,1) = 0,05\,,$$

odnosno kritična oblast je $C = [11,1;\, +\infty)$. Kako je $12,8 > 11,1$ hipotezu $H_0$ odbacujemo sa 5%–nim pragom značajnosti. $\triangle$

### B. Testiranje jednakosti dve multinomne raspodele

Posmatrajmo dve nezavisne multinomne raspodele sa parametrima $n_j, p_{1j}, p_{2j}, \ldots, p_{kj}$, $j = 1, 2$ respektivno. Neka $X_{ij}$, $i = 1, 2, \ldots, k$, $j = 1, 2$ predstavljaju odgovarajuće učestanosti. Ako su $n_1$ i $n_2$ veliki, slučajna promenljiva

$$\sum_{j=1}^{2}\sum_{i=1}^{k} \frac{(X_{ij} - n_j p_{ij})^2}{n_j p_{ij}} = \sum_{i=1}^{k} \frac{(X_{i1} - n_1 p_{i1})^2}{n_1 p_{i1}} + \sum_{i=1}^{k} \frac{(X_{i2} - n_2 p_{i2})^2}{n_2 p_{i2}}$$

je zbir dve stohastički nezavisne slučajne promenljive od kojih svaka ima raspodelu $\chi^2_{k-1}$. To znači da je navedena slučajna promenljiva sa raspodelom $\chi^2_{2k-2}$.
Testiramo hipotezu

$$H_0 : p_{11} = p_{12},\, p_{21} = p_{22}, \ldots,\, p_{k1} = p_{k2}$$

gde su svi $p_{i1} = p_{i2}$, $i = 1, 2, \ldots, k$ nepoznati. Stoga su nam potrebne tačkaste ocene ovih parametara. Statistika maksimalne verodostojnosti za ove parametre u slučaju kada je $p_{i1} = p_{i2}$ je

$$\widehat{p}_{ij} = \frac{X_{i1} + X_{i2}}{n_1 + n_2}\,, \quad i = 1, 2, \ldots, k\,, \quad j = 1, 2\,.$$

Primetimo da nam je potrebna samo $k - 1$ ocena, jer ćemo ocenu za $p_{k1} = p_{k2}$ imati samim tim što smo našli ocene za prvih $k - 1$ verovatnoća. Dakle, slučajna promenljiva

$$\sum_{j=1}^{2}\sum_{i=1}^{k} \frac{(X_{ij} - n_j \frac{X_{i1}+X_{i2}}{n_1+n_2})^2}{n_j \frac{X_{i1}+X_{i2}}{n_1+n_2}}$$

ima približno $\chi^2$ raspodelu sa $2k - 2 - (k - 1) = k - 1$ stepeni slobode.

**Primer 72.** Testirati hipotezu o jednakoj zastupljenosti četiri tipa ličnosti u populaciji stanovništva dva grada na osnovu nezavisnih uzoraka obima 100 sa 5%–nim pragom značajnosti:

| Tip ličnosti | kolerik | sangvinik | melanholik | flegmatik |
|---|---|---|---|---|
| Ostvarene učestanosti u gradu I | 30 | 25 | 23 | 22 |
| Ostvarene učestanosti u gradu II | 25 | 27 | 23 | 25 |

.

Ocenimo najpre, na osnovu datog uzorka, nepoznate verovatnoće:

$$\widehat{p}_{11} = \widehat{p}_{12} = \frac{30+25}{100+100} = 0,275 \quad , \quad \widehat{p}_{21} = \widehat{p}_{22} = \frac{25+27}{200} = 0,26 \quad ,$$

$$\widehat{p}_{31} = \widehat{p}_{32} = \frac{23+23}{200} = 0,23 \quad \text{i} \quad \widehat{p}_{41} = \widehat{p}_{42} = \frac{22+25}{200} = 0,235\,.$$

S obzirom da je $n_1 = n_2 = 100$, to je

$$n_j\widehat{p}_{1j} = 27,5\,, \quad n_j\widehat{p}_{2j} = 26\,, \quad n_j\widehat{p}_{3j} = 23\,, \quad n_j\widehat{p}_{4j} = 23,5\,, \quad j = 1,2\,.$$

Test statistika za proveru postavljene nulte hipoteze

$$\chi_0^2 = \sum_{j=1}^{2}\sum_{i=1}^{k} \frac{(X_{ij} - n_j\widehat{p}_{ij})^2}{n_j\widehat{p}_{ij}}$$

će imati $\chi^2$–raspodelu sa 4-1=3 stepena slobode, te je njena realizovana vrednost

$$\chi_3^2 = \frac{(30-27,5)^2}{27,5} + \frac{(25-26)^2}{26} + \frac{(23-23)^2}{23} + \frac{(22-23,5)^2}{23,5} +$$

$$+\frac{(25-27,5)^2}{27,5} + \frac{(27-26)^2}{26} + \frac{(23-23)^2}{23} + \frac{(25-23,5)^2}{23,5} = 0,72\,,$$

a kritična oblast je $[7,81; +\infty)$. Zaključak je da se sa 5%–nim pragom značajnosti može smatrati da je približno podjednak broj svakog od tipova ličnosti u oba grada, jer $0,72 \notin C$. $\triangle$

### C. Ispitivanje nezavisnosti $\chi^2$ testom (tabele kontingencije)

$\chi^2$ testom često se ispituje i nezavisnost dva obeležja iste populacije. Dakle, za dva obeležja $X$ i $Y$ jedne populacije, testira se hipoteza

$$H_0 : X \text{ i } Y \text{ su nezavisna obeležja}$$

protiv alternativne da nisu nezavisna.

Testiranje se obavlja tako što se na uzorku obima $n$ iz posmatrane populacije registruju "vrednosti" oba obeležja (obeležja ne moraju biti numerička, kvantitativna, već mogu biti i kvalitativna). Pri tome se formira tabela (tabela kontingencije, slučajnosti) koja u polju $(i,j)$ ima podatak $n_{ij}$ o broju elemenata u uzorku, kod kojih obeležje $X$ ima vrednost $x_i$, a obeležje $Y$ vrednost $y_j$:

| $X\downarrow / Y\rightarrow$ | $y_1$ | $y_2$ | $\ldots$ | $y_r$ | $n_{i\bullet}$ |
|---|---|---|---|---|---|
| $x_1$ | $n_{11}$ | $n_{12}$ | $\ldots$ | $n_{1r}$ | $n_{1\bullet}=\sum_j n_{1j}$ |
| $x_2$ | $n_{21}$ | $n_{22}$ | $\ldots$ | $n_{2r}$ | $n_{2\bullet}=\sum_j n_{2j}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $x_k$ | $n_{k1}$ | $n_{k2}$ | $\ldots$ | $n_{kr}$ | $n_{k\bullet}=\sum_j n_{kj}$ |
| $n_{\bullet j}$ | $n_{\bullet 1}=$ $\sum_i n_{i1}$ | $n_{\bullet 2}=$ $\sum_i n_{i2}$ | $\ldots$ | $n_{\bullet r}=$ $\sum_i n_{ir}$ | $n$ |

.

U tabeli su korišćene sledeće oznake:

- $n_{ij}$ je broj parova za koje je $X=x_i$ i $Y=y_j$;
- $n_{i\bullet}$ je broj parova za koje je $X=x_i$;
- $n_{\bullet j}$ je broj parova za koje je $Y=y_j$;

Prema tome, nulta hipoteza: $X$ i $Y$ su nezavisna obeležja, može da se iskaže kao

$$H_0 : \forall(i,j)\ p_{ij}=p_{i\bullet}p_{\bullet j},$$

a alternativna

$$H_1 : \exists(i,j)\ p_{ij}\neq p_{i\bullet}p_{\bullet j},$$

gde je $p_{ij}=P\{X=x_i\wedge Y=y_j\}$, $p_{i\bullet}=P\{X=x_i\}$, $p_{\bullet j}=P\{Y=y_j\}$.

Verovatnoće $p_{i\bullet}$ i $p_{\bullet j}$ se ocenjuju na osnovu relativnih učestanosti

$$\widehat{p}_{i\bullet}=\frac{n_{i\bullet}}{n},\ \widehat{p}_{\bullet j}=\frac{n_{\bullet j}}{n}.$$

Ako je hipoteza $H_0$ tačna, nepoznata verovatnoća $p_{ij}$ ocenjuje se na sledeći način

$$\widehat{p}_{ij}=\widehat{p}_{i\bullet}\widehat{p}_{\bullet j}=(\frac{n_{i\bullet}}{n})(\frac{n_{\bullet j}}{n})=\frac{n_{i\bullet}n_{\bullet j}}{n^2}.$$

Očekivani broj parova za koje je $X=x_i$ i $Y=y_j$ ocenjuje se sa

$$\widehat{n}_{ij}=n\widehat{p}_{ij}=\frac{n_{i\bullet}n_{\bullet j}}{n}.$$

Odstupanje od hipoteze o nezavisnosti za "ćeliju" $(i,j)$ je

$$\frac{(n_{ij}-\widehat{n}_{ij})^2}{\widehat{n}_{ij}}\,.$$

Otuda je ukupno odstupanje za sve ćelije

$$\chi_0^2=\sum_{i=1}^{k}\sum_{j=1}^{r}\frac{(n_{ij}-\widehat{n}_{ij})^2}{\widehat{n}_{ij}}\,.$$

Ako je hipoteza $H_0$ tačna, $\chi_0^2$ ima $\chi^2$ raspodelu sa $(k-1)\cdot(r-1)$ stepeni slobode. Kritična oblast testa je:

| $H_0$ | $H_1$ | $C$ |
|---|---|---|
| $X$ i $Y$ su nezavisna obeležja | $X$ i $Y$ nisu nezavisna obeležja | $\chi_0^2 \geq \chi^2_{(k-1)\cdot(r-1);1-\alpha}$ |

Dakle, ako je za realizovani uzorak $\chi_0^2 < \chi^2_{(k-1)\cdot(r-1);1-\alpha}$, $H_0$ se prihvata. Medjutim, i ovde se može uporedjivati realizovana vrednost statistike i očekivanje slučajne promenljive sa odgovarajućom $\chi^2$ raspodelom ($H_0$ se prihvata ako je $\chi_0^2 < (k-1)(r-1)$ ).

**Primer 73.** Ispituje se nezavisnost visine i težine stanovnika jedne regije. Svi rezultati merenja su svrstani u sledeće kategorije: visoki i gojazni, visoki i negojazni, niski i gojazni i niski i negojazni i prikazani tabelom:

| $Y$ (visina) $\rightarrow$ <br> $X$ (težina) $\downarrow$ | visoki | niski | $n_{i\bullet}$ | $\widehat{p}_{i\bullet}$ |
|---|---|---|---|---|
| gojazni | 14 | 36 | 50 | 0,161 |
| negojazni | 59 | 201 | 260 | 0,839 |
| $n_{\bullet j}$ | 73 | 237 | 310 | / |
| $\widehat{p}_{\bullet j}$ | 0,236 | 0,765 | / | 1 |

To daje:

$$\widehat{p}_{11} = \frac{50 \cdot 73}{310^2} \;\Rightarrow\; \widehat{n}_{11} = 310 \cdot \widehat{p}_{11} = 11,8$$
$$\widehat{p}_{12} = \frac{50 \cdot 237}{310^2} \;\Rightarrow\; \widehat{n}_{12} = 310 \cdot \widehat{p}_{12} = 38,2$$
$$\widehat{p}_{21} = \frac{260 \cdot 73}{310^2} \;\Rightarrow\; \widehat{n}_{21} = 310 \cdot \widehat{p}_{21} = 61,2$$
$$\widehat{p}_{22} = \frac{260 \cdot 237}{310^2} \;\Rightarrow\; \widehat{n}_{22} = 310 \cdot \widehat{p}_{22} = 198,8\,,$$

pa je

$$\chi_0^2 = \frac{(14-11,8)^2}{11,8} + \ldots + \frac{(201-198,8)^2}{198,8} = 0,64.$$

Broj stepeni slobode je $(2-1)(2-1) = 1$ pa je kritična oblast $[3,84;+\infty)$. Dakle, $H_0$ se prihvata, tj. može se smatrati da su za tu populaciju visina i težina nezavisna obeležja. $\triangle$

Uočimo da je u prethodnom primeru broj stepeni slobode primenjene test statistike sa $\chi^2$–raspodelom bio 1. Po pravilu se u takvim slučajevima primene Pirsonovog testa, iz razloga u koje u okviru ovog kursa nećemo ulaziti, vrši tzv. Jejcova (Yates' correction) korekcija ($\chi^2_{\text{korigovano}} < \chi_0^2$):

$$\chi^2_{\text{korigovano}} = \sum_{i=1}^{k} \frac{(|o_i - e_i| - 0,5)^2}{e_i},$$

gde je $o_i$ opservirana apsolutna učestanost, a $e_i$ očekivana apsolutna učestanost.

Dakle, u poslednjem primeru bi trebalo izvršiti korekciju.

Korigovana vrednost iz primera je: $\chi^2_{\text{korigovano}} = 0,38$, a za prag značajnosti $\alpha = 0,05$ odgovarajuća kritična oblast je $C = [3,84; +\infty)$. Kako vrednost 0,38 ne pripada skupu $C$ prihvata se hipoteza da su obeležja nezavisna, tj. zaključak se ne menja u odnosu na onaj koji je načinjen bez korekcije.

Uočimo da smo korigovanjem došli do manje vrednosti za test statistiku na osnovu istog uzorka. Ovo je zaključak koji važi uvek pri primeni Jejcove korekcije. Otuda, ako je realizovana vrednost test statistike manja od leve granice kritične oblasti, korekciju ne treba ni vršiti, jer neće uticati na zaključak o prihvatanju nulte hipoteze.

**Primer 74.** Ispitano je 500 osoba kojima je postavljeno pitanje da li bi na predstojećim izborima svoje poverenje poklonili kandidatu koji bi bio muškarac ili radije kandidatu koji bi bio ženskog pola i dobijeni su sledeći rezultati:

| Birači \ Kandidat | Muškarac | Žena |
|---|---|---|
| Muškarci | 180 | 90 |
| Žene | 130 | 100 |

Ispitati da li izbor kandidata zavisi od pola birača za $\alpha = 0,05$.

Testira se hipoteza $H_0$(izbor kandidata ne zavisi od pola birača). Realizovana vrednost test statistike je $\chi^2_0 = 5,43$, a broj stepeni slobode joj je 1. Kritična oblast je, kao i u prethodnom primeru, $C = [3,84\,;+\infty)$. Dakle, trebalo bi odbaciti nultu hipotezu, odnosno, zaključiti da će ishod izbora zavisiti od pola birača koji budu izašli na izbore. Medjutim, primenom Jejcove korekcije na isti uzorak dolazimo do podatka $\chi^2_{\text{korigovano}} = 3,47$, što znači da se hipoteza $H_0$ prihvata, odnosno da izbor kandidata (muškarca ili žene) neće zavisiti od pola glasača koji budu izašli na izbore.

Osvrnimo se ovde još na činjenicu da je korigovana vrednost 3,47 relativno blizu granice kritične oblasti, što po pravilu nalaže dodatnu analizu. U ovoj situaciji primereno bi bilo tražiti veći obim uzorka, ili ako to nije moguće, zaključivati sa nekim drugim pragom značajnosti. Recimo, za prag značajnosti $\alpha = 0,1$, kritična oblast će biti $C = [2,71\,;+\infty)$, pa bi se na ovom nivou značajnosti nulta hipoteza odbacila. △

⋆ ⋆ ⋆

Prednost neparametarskih testova nad parametarskim je u tome što nisu neophodna nikakva prethodna znanja o obliku raspodele obeležja u vezi sa kojim se vrši testiranje (kod parametarskih testova su neophodne pretpostavke o raspodeli obeležja koje se testira). Medutim, ako su ispunjeni kriterijumi za primenu parametarskih testova, onda njih treba i primeniti jer su efikasniji od odgovarajućih neparametarskih.

Definicija 40. Od dva testa za testiranje iste nulte hipoteze protiv iste odgovarajuće alternativne, *efikasniji* je onaj za koji je potreban manji obim uzorka da bi se postigla jednaka verovatnoća odbacivanja nulte hipoteze ako je ona zaista pogrešna:

$$P_{H_1}\{(X_1,\ldots,X_n)\in C\} = 1 - P_{H_1}\{(X_1,\ldots,X_n)\in C^c\} = 1-\beta.$$

U nastavku su izloženi još neki od neparametarskih testova koji su često u upotrebi.

### 4.3.3 Binomni test (test znakova)

**Test znakova** se zasniva na posmatranju realizacije nekog dogadjaja $A$ u nizu nezavisnih opita, za koje je verovatnoća realizacije dogadjaja $A$ u svakom pojedinom opitu $P(A) = p$. Ostvarenje dogadjaja $A$ se označava sa "+", a neostvarenje sa "—".

Binomni test se može koristiti **za testiranje jednakosti raspodela dva obeležja** $X$ i $Y$:

$$H_0 : F_X = F_Y \quad \text{i} \quad H_1 : F_X \neq F_Y$$

na osnovu dva nezavisna uzorka $(X_1, \ldots, X_n)$ i $(Y_1, \ldots, Y_n)$ **istog obima**. Uzorci se ne smeju sredjivati u varijacioni niz niti intervalno, a uporedjivanje registrovanih (realizovanih) vrednosti iz dva uzorka ima suštinskog smisla.

**Primer 75.** U fabrici postoje dve nezavisne linije za proizvodnju jednog proizvoda. Testira se hipoteza da je kvalitet proizvoda proizvedenih na ovim linijama isti kroz registrovanje broja defektnih proizvoda u toku 10 dana na obe linije. Registrovani podaci o broju defektnih proizvoda su prikazani tabelom:

| Dan | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Linija I | 172 | 165 | 206 | 184 | 174 | 142 | 190 | 169 | 161 | 200 |
| Linija II | 201 | 179 | 159 | 192 | 177 | 170 | 182 | 179 | 169 | 210 |

.

Testiramo hipotezu o jednakoj raspodeli broja defektnih proizvoda na dve nezavisne proizvodne linije (ostvaren u istom periodu), protiv alternativne da su raspodele različite.

Za svaki od posmatranih 10 dana beležićemo znak "+" ukoliko je na liniji $I$ registrovano više defektnih proizvoda nego na liniji $II$, a znak "—" u suprotnom slučaju. Označimo sa $T$ broj znakova "+" u posmatranom uzorku. Za naš realizovani uzorak je $t = 2$. (Ukoliko bi se u nekom danu konstatovao jednak broj defektnih prizvoda na obe proizvodne linije, takav dan bi se u uzorku ignorisao, odnosno posmatrao bi se uzorak iz koga bi bili eliminisani ovakvi dani.)

Neka ostvarenje znaka "+" u nizu znači realizaciju dogadjaja $A$. U opštem slučaju, $T$ bi bila slučajna promenljiva sa binomnom raspodelom. Ukoliko je nulta hipoteza tačna, u nizu registrovanih znakova treba da je približno jednak broj znakova "+" i "—", odnosno, $P(+) = P(A) = 0,5$. Na taj način se nulta hipoteza $H_0 : F_X = F_Y$ prevodi u hipotezu

$$H_0 : \quad T : \mathcal{B}(10\,;\, 0,5)$$

a alternativna $H_1$ bi bila da $T$ nema naznačenu binomnu raspodelu. Baš iz ovog razloga sam test nosi naziv binomni test.

Iz prethodne analize sledi da su za nultu hipotezu problematične kako male, tako i velike vrednosti test statistike $T$.

Pretpostavimo da vršimo testiranje sa pragom značajnosti $\alpha = 0,05$ i $\alpha = 0,1$. Odredimo kritičnu oblast.

Ako kritičnu oblast čine vrednosti $T = 0$ i $T = 10$, pod uslovom da je nulta hipoteza tačna, verovatnoća greške prve vrste bi bila

$$\alpha = P\{T = 0 \vee T = 10\} = \binom{10}{0} \cdot 0,5^{10} + \binom{10}{10} \cdot 0,5^{10} = 0,002\,.$$

Medjutim, ova vrednost za $\alpha$ je suviše mala, pa znači da treba proširiti kritičnu oblast. Ako u kritičnu oblast uključimo i $T = 1$ i $T = 9$, dobijamo da je $\alpha = 0,022$, što je i dalje mala vrednost u odnosu na zadate vrednosti verovatnoće greške prve vrste za ovo testiranje. Dakle, još jednom proširimo kritičnu oblast i definišimo je sa $T = 0, 1, 2, 8, 9, 10$. U tom slučaju je $\alpha = 0,11$. Ukoliko smo zadovoljni ovom preciznošću, možemo da sprovedemo željeno testiranje. Prema tome, s obzirom da je realizovana vrednost test statistike $t = 2$ i da ona pripada kritičnoj oblasti, nultu hipotezu odbacujemo.△

Za mali obim uzorka, kao što je prikazano u prethodnom primeru, direktno se koristi binomna raspodela za odredjivanje kritične oblasti veličine $\alpha$. Medjutim, za grubu procenu ili za velike uzorke, koristi se normalna aproksimacija binomne raspodele, tj. statistika

$$Z_0 = \frac{T - np_0}{\sqrt{np_0(1-p_0)}}$$

gde je $T$–broj znakova " + " u nizu znakova dobijenom na osnovu uzorka i koja za $n > 50$ ima približno $\mathcal{N}(0,1)$ raspodelu. (Stvarna raspodela statistike $T$ je $\mathcal{B}(n, p_0)$.)

Kritične oblasti se odredjuju u zavisnosti od alternativne hipoteze, kao i kod parametarskog testa za testiranje parametra $p$ binomne raspodele (v. potpoglavlje 4.2.3).

Poznato je da za mali obim uzorka postoji i odgovarajuća modifikovana test statistika koja ima približno Fišerovu raspodelu, o čemu ovde neće biti reči.

Opisani test se često naziva i **test kvantila**, jer se njime za obeležje $X$ može da ispituje hipoteza

$$H_0 : P\{X \leq x_0\} = p_0,$$

gde su $x_0$ i $p_0$ zadate vrednosti i realizacija dogadjaja $\{X \leq x_0\}$ u uzorku se označi sa " + ", a njemu suprotnog sa " − ". Moguće alternativne hipoteze su:

$$H_1 : P\{X \leq x_0\} \neq p_0\,, \quad H_1 : P\{X \leq x_0\} > p_0 \quad \text{i} \quad H_1 : P\{X \leq x_0\} < p_0\,.$$

S obzirom na definiciju kvantila, nulta hipoteza se može da iskaže i kao:

$$H_0 : M_{p_0} = x_0,$$

a alternativne redom kao

$$H_1 : M_{p_0} \neq x_0\,, \quad H_1 : M_{p_0} < x_0 \quad \text{i} \quad H_1 : M_{p_0} > x_0\,.$$

**Primer 76.** Posmatra se grupa od 100 dvadesetogodišnjaka i registruje broj onih koji imaju konformizam. Za datu grupu dobijeno je da 68 njih ima datu osobinu. Testirati hipotezu da će konformizam imati 75% dvadesetogodišnjaka za $\alpha = 0,05$.

Obim uzorka jednak je $n = 100$, $p_0 = 0,75$ je verovatnoća dvadesetogodišnjaka sa konformizmom od ukupno posmatranih i $T = 68$ je broj onih koji imaju konformizam. Realizovana vrednost test statistike je $z_0 = -1,617$ a kritična oblast je $C = (-\infty; -1,96] \cup [1,96; +\infty)$. Kako realizovana vrednost ne pripada kritičnoj oblast, to se nulta hipoteza prihvata, tj. može se smatrati da 75% dvadesetogodišnjaka ima konformizam. △

Kada je $p_0 = 1/2$ za ovaj test se koristi naziv **test medijane**.

Test medijane se može primeniti i za dva nezavisna uzorka različitih obima kada se testira hipoteza o tome da uzorci potiču od dva obeležja sa istom medijanom, ili još bolje iz istog obeležja (na jednoj te istoj populaciji).

Test kvantila se koristi i za "sparene uzorke", odnosno za **testiranje hipoteze da nije došlo do promene raspodele obeležja** $X$ apsolutno neprekidnog tipa pod dejstvom različitih faktora uticaja na elemente jedne populacije. Za ovo testiranje se registruju vrednosti obeležja $X$ na istim elementima populacije dva puta, tj. pod različitim okolnostima i dobija se tzv. spareni uzorak

$$((X_1^{(1)}, X_1^{(2)}) \ldots, (X_n^{(1)}, X_n^{(2)}))\,.$$

**Primer 77.** Na 12 klijenata je primenjivana odgovarajuća grupna psihoterapija sa ciljem da se ublaži ili ukloni depresivno stanje u kome su se nalazili. Dati su kodirani nivoi psihičkog stanja svakog pacijenta pre ($X^{(1)}$) i posle ($X^{(2)}$) sprovedene terapije.

| Kl. | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $x^{(1)}$ | 5,6 | 7,1 | 6,4 | 5,8 | 4,9 | 4,7 | 5,0 | 4,9 | 3,6 | 5,4 | 4,7 | 3,1 |
| $x^{(2)}$ | 5,6 | 6,3 | 6,7 | 5,3 | 4,0 | 5,2 | 4,9 | 5,2 | 3,3 | 4,8 | 3,2 | 2,4 |
| $\Delta$ | 0 | + | − | + | + | − | + | − | + | + | + | + |

U cilju sprovodjenja statističkog testa definiše se statistika

$$\Delta = sgn(X^{(1)} - X^{(2)})\,,$$

čije su vrednosti: 0 – ukoliko je $X^{(1)} = X^{(2)}$, "+" – ukoliko je $X^{(1)} > X^{(2)}$ i "—" – ukoliko je $X^{(1)} < X^{(2)}$.

Ako se kod testa znakova u nizu "+" i "—" javi i 0, onda se ovakav element uzorka nadalje ignoriše i radi se sa uzorkom smanjenog obima. Ovo stoga što realizacija takvog dogadjaja ima verovatnoću 0, s obzirom na pretpostavku o neprekidnoj raspodeli.

Ako nema razlike u raspodelama, smatra se da je

$$P\{\Delta = " + "\} = P\{\Delta = " - "\} = 0,5\ ,$$

tj. da je jednako verovatno da je $X^{(1)} > X^{(2)}$ $(0,5)$ kao i $X^{(1)} < X^{(2)}$ $(0,5)$.

Testira se hipoteza da **nije** došlo do smanjenja depresije

$$H_0 : P\{X^{(1)} > X^{(2)}\} = 0,5$$

protiv alternativne

$$H_1 : P\{X^{(1)} > X^{(2)}\} > 0,5$$

da **je** došlo do smanjenja.

Prvi klijent se izostavlja iz razmatranja (jer je kod njega registrovano $x^{(1)} = x^{(2)}$), pa se radi sa $n = 11$ i $T = 8$. S obzirom da obim uzorka nije "dovoljno veliki", samo će se grubo proceniti odgovor na postavljeno pitanje primenom normalne aproksimacija test

statistike, pre svega radi demonstracije postupka, a ne za izvodjenje ozbiljnog i odgovornog zaključka (gde joj, zapravo, nije mesto), koja daje

$$z_0 = \frac{8 - 11 \cdot 0,5}{\sqrt{11 \cdot 0,5 \cdot 0,5}} = 1,51.$$

Za jednostranu alternativu $H_1$ kritične su velike vrednosti za $z_0$, tako da je za $\alpha = 0,05$ granica kritične oblasti $z_{0,5-0,05} = z_{0,495} = 1,645$, odnosno kritična oblast je $C = [1,645; +\infty)$. Kako $z_0 \notin C$ prihvata se $H_0$, što znači da se ne primećuje značajna razlika u psihičkom stanju klijenata pre i posle sprovedene terapije. $\triangle$

Umesto ponudjene test statistike, može se koristiti i statistika

$$Z = \frac{|D| - 1}{\sqrt{n}},$$

gde je $D$ =broj(+)−broj(−), koja takodje ima aproksimativno $\mathcal{N}(0,1)$ raspodelu.

### 4.3.4 Test serija (test koraka)

Test serija se zasniva na ispitivanju slučajnosti medjusobnog rasporeda dve vrste objekata, npr. 0 i 1.

Neka se 0 javlja $n_1$ puta, a 1 javlja $n_2$ puta, u zajedničkom nizu od $n = n_1 + n_2$ elemenata. Seriju čini podniz istih elemenata, bilo 0 bilo 1. Neka je $U$ ukupan broj serija u dobijenom nizu. Ako je hipoteza o slučajnom rasporedu 0 i 1 tačna, važe formule:

$$E(U) = \frac{2n_1n_2}{n} + 1\,, \quad D(U) = \frac{(E(U)-1)(E(U)-2)}{n-1}\,, \quad n = n_1 + n_2\,.$$

Za raspodelu statistike $U$ postoje posebne tablice, ali se pokazalo da se već za $n_1n_2 \geq 9$ može koristiti normalna aproksimacija za standardizovanu vrednost

$$Z_0 = \frac{U - E(U)}{\sqrt{D(U)}}\,.$$

Test koraka se koristi najčešće za testiranje slučajnosti niza podataka, kao i za ispitivanje jednakosti raspodela dvaju neprekidnih obeležja. U narednom primeru prikazano je **testiranje slučajnosti** niza podataka.

**Primer 78.** Razmatra se primer o zaradama radnika (primer 69). Uporedjuju se zarade radnika iz drugog pogona (poslednjih 20 radnika u uzorku) u odnosu na medijanu zarada svih radnika u uzorku (odnosno u odnosu na pretpostavljenu medijanu zarada svih zaposlenih u tom preduzeću), $M_{0,5} = 825$, na osnovu uzorka koji nije sredjen intervalno.

Veličina zarade manja od medijane označava se sa 0, a veća sa 1. Dobija se niz: 1, 0, 1, 1, 1, 1, 1, 0, 0, 1, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0. Testira se hipoteza da je dobijeni niz slučajan sa pragom značajnosti $\alpha = 0,05$. S obzirom da je problematičan mali broj serija u nizu podataka, radi se o jednostranoj alternativnoj hipotezi, pa je $C = (-\infty; -1,645]$.

$$n_1 = n_2 = 10, \quad E(U) = 2\frac{10 \cdot 10}{10 + 10} + 1 = 11,$$

$$D(U) = \frac{(11-1)(11-2)}{20-1} = (2,18)^2,$$
$$z_0 = \frac{10-11}{2,18} = -0,46\,.$$

Dakle, $z_0 \notin C$ pa se $H_0$ prihvata, odnosno, radnici drugog pogona bi činili reprezentativan uzorak pri ispitivanju obeležja "visina zarada zaposlenih" (u posmatranom preduzeću). Drugim rečima, o zaradi zaposlenih u preduzeću koje čine dva pomenuta pogona može se zaključivati samo na osnovu zarada radnika drugog pogona, jer je raspodela zarada ista u drugom pogonu i u celom preduzeću. △

Za normalnu aproksimaciju, kritična oblast je:

| $H_0$ | $H_1$ | $C$ |
|---|---|---|
| Niz je slučajan | U nizu je mali broj serija | $z_0 \leq -z_{0,5-\alpha}$ |

.

Ispitivanje slučajnosti se vrši i u odnosu na druge nivoe osim medijane.

Važno je da se za testiranje slučajnosti testom koraka ne sme sredjivati uzorak ni po kom kriterijumu. Dakle, realizovani uzorak se koristi onakav kakav je i nastao u seriji merenja (posmatranja).

Kod **testiranja jednakosti raspodela dva obeležja**, situacija je sledeća.

Razmatraju se dva obeležja apsolutno neprekidnog tipa, $X$ i $Y$, na nezavisnim uzorcima obima $n_1$ i $n_2$ redom. Testira se hipoteza da oba obeležja imaju istu raspodelu, tj.

$$H_0 : F_X = F_Y$$

protiv alternativne $H_1 : F_X \neq F_Y$.

Test serija primenjuje se tako što se elementi oba realizovana uzorka poredjaju u jedinstven neopadajući niz, pa se elementi prvog uzorka označavaju sa 0, drugog sa 1, čime se dobije niz od $n_1+n_2$ simbola 0 i 1. Ako je hipoteza $H_0$ tačna, raspored 0 i 1 je slučajan. Pri tome se kao sumnjiva za hipotezu $H_0$ smatra samo pojava malog broja serija (veća grupisanja 0 i 1) u nizu simbola. Prema tome, primenjuje se ista test statistika $Z_0$ kao i za prethodni test serija, važe iste formule za $E(U), D(U)$ i $n$. Kritična oblast data je u tabeli

| $H_0$ | $H_1$ | $C$ |
|---|---|---|
| $F_X = F_Y$ | $F_X \neq F_Y$ | $z_0 \leq -z_{0,5;-\alpha}$ |

,

a pre saznanja o granicama kritične oblasti, dobijene male vrednosti za $z_0$ znak su da sa statističkim zaključivanjem na osnovu takve serije treba biti obazriv.

### 4.3.5 Test rangova (test Vilkokson – Man – Vitnija)

Test rangova se, kao i test serija, zasniva na ispitivanju slučajnosti pojavljivanja 0 i 1 u nizu izvedenom iz realizovanog uzorka, a osetljiviji je od testa serija.

**Primer 79.** Posmatrajmo niz nula i jedinica iz primera 78. Testirajmo **hipotezu o slučajnosti** primenom testa rangova za $\alpha = 0,05$.

Utvrdjuje se koliko ukupno jedinica ima ispred svake pojedine nule. Za razmatrani niz: 1, 0, 1, 1, 1, 1, 1, 0, 0, 1, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0, dobija se sledeći broj inverzija (inverzijom u nizu se smatra pojava 1 ispred 0) 1, 6, 6, 7, 7, 8, 10, 10, 10, 10.

Kao test statistika koristi se $V$–ukupan broj jedinica levo od svake nule u nizu posebno, odnosno za razmatrani primer je

$$V = 1 + 6 + 6 + 7 + 7 + 8 + 10 + 10 + 10 + 10 = 75\,.$$

S obzirom na ukupan broj nula i jedinica u nizu:

$$0 \leq V \leq 100\,.$$

Za hipotezu o slučajnosti kritičan je kako mali tako i veliki broj inverzija u nizu nula i jedinica, jer to ukazuje na grupisanje nula ili jedinica na početku niza. U opštem slučaju, za $n_1$ nula i $n_2$ jedinica u nizu je $0 \leq V \leq n_1 n_2$. Ako je hipoteza o slučajnosti tačna i $n_1 \geq 10$ i $n_2 \geq 10$, statistika $V$ ima raspodelu veoma blisku normalnoj sa

$$E(V) = \frac{n_1 n_2}{2}\,, \quad D(V) = E(V) \cdot \frac{n+1}{6}\,, \quad n = n_1 + n_2\,.$$

U razmatranom primeru je $n = 20$, i

$$E(V) = \frac{10 \cdot 10}{2} = 50\,, \quad D(V) = 50\frac{20+1}{6} = 175\,, \quad \sqrt{D(V)} = 13,23\,.$$

Stoga se umesto tačne raspodele koristi normalna aproksimacija, tj. statistika

$$Z_0 = \frac{V - E(V)}{\sqrt{D(V)}}$$

koja ima približno $\mathcal{N}(0;1)$ raspodelu. Ovde je, znači,

$$z_0 = \frac{75 - 50}{13,23} = 1,89.$$

Kako je problematičan i veliki i mali broj inverzija, kritična oblast je:

| $H_0$ | $H_1$ | $C$ |
|---|---|---|
| Niz je slučajan | U nizu je premalo ili previše inverzija | $\lvert z_0 \rvert \geq z_{0,5-\frac{\alpha}{2}}$ |

.

Za razmatrani primer je

$$z_{0,475} = 1,96\,, \quad \text{te je} \quad C = (-\infty; -1,96] \cup [1,96; +\infty)\,,$$

pa se, s obzirom da $z_0 = 1,89$ ne pripada kritičnoj oblasti, prihvata hipoteza o slučajnosti razmatrane serije. Dakle, došli smo do istog zaključka kao i posle primene testa serija. Ovde je, medjutim, važno uočiti da je realizovana vrednost $z_0 = 1,89$ bliska granici kritične oblasti (1,96) čime je poljuljana "pouzdanost" zaključka. U praktičnim primenama nije poželjno oslanjati se na takve zaključke. Izlaz se može tražiti u povećanju obima uzorka ili u zaključivanju na nekom drugom nivou značajnosti. $\triangle$

Test rangova takodje se koristi i za **testiranje jednakosti raspodela**. Za neprekidna obeležja $X$ i $Y$ preko nezavisnih uzoraka obima $n_1$ i $n_2$ testira se hipoteza o jednakosti raspodela polazeći od niza 0 i 1 dužine $n = n_1 + n_2$, koji se formira na isti način kao kod testa serija. I nulta hipoteza je ista:

$$H_0 : F_X = F_Y \, .$$

Koristi se test statistika $Z_0$, ista i pod istim uslovima kao kod ispitivanja slučajnosti, a kritična oblast data je u tabeli

| $H_0$ | $H_1$ | $C$ |
|---|---|---|
| $F_X = F_Y$ | $F_X \neq F_Y$ | $\lvert z_0 \rvert \geq z_{0,5-\alpha/2}$ |

.

Osetljivost ovog testa, medjutim, sastoji se u tome što je pogodan za testiranje nulte hipoteze $H_0 : F_X = F_Y$, protiv alternativa $H_1 :$ "$X$ je češće veće od $Y$", i $H_1 :$ "$X$ je češće manje od $Y$". Odgovarajuće kritične oblasti su u tom slučaju:

| $H_0$ | $H_1$ | $C$ |
|---|---|---|
| $F_X = F_Y$ | $X$ je češće veće od $Y$ | $z_0 \geq z_{0,5-\alpha}$ |
| $F_X = F_Y$ | $X$ je češće manje od $Y$ | $z_0 \leq -z_{0,5-\alpha}$ |

.

**Primer 80.** Dve grupe, svaka od po 12 eksperimentalnih miševa obolelih od raka, izložene su hemoterapiji da bi se proverilo njeno dejstvo na ćelije raka. Osim toga, samo drugoj grupi miševa dat je istovremeno i antitoksin koji je imao za cilj da spreči uništavanje zdravih ćelija prilikom primene hemoterapije. Mereno je vreme (u satima) preživljavanja eksperimentalnih životinja u odnosu na početak primene terapije. Eksperiment je okončan nakon 480 sati, tj. nakon 20 dana, te je za životinje koje su preživele ovaj period registrovano vreme života 480 sati. Vremena preživljavanja eksperimentalnih životinja data su tabelom

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| I grupa | 84 | 128 | 168 | 92 | 184 | 92 | 76 | 104 | 72 | 180 | 144 | 120 |
| II grupa | 140 | 184 | 368 | 96 | 480 | 188 | 480 | 244 | 440 | 380 | 480 | 196 |

.

Da li se na osnovu ovog eksperimenta na nivou značajnosti $\alpha = 0,05$ može tvrditi da će primenom antitoksin terapije biti produžen život pacijenata koji se izlažu hemoterapiji u odnosu na pacijente kojima se ne uključuje antitoksin?

Sredjivanjem realizovanih uzoraka u jedinstven neopadajući niz i dodeljivanjem 0 svakom elementu I eksperimentalne grupe, a 1 svakom elementu II eksperimentalne grupe, dobijamo niz

$$0,0,0,0,0,1,0,0,0,1,0,0,0,0,1,1,1,1,1,1,1,1,1,1 \, .$$

Na osnovu ovog niza testiramo hipotezu

$$H_0 : \text{Nema razlike u dužini života I i II grupe} \ ,$$

protiv alternativne

$$H_1 : \text{Dužina života miševa II grupe je veća u odnosu na I grupu}\,.$$

Dakle,

$$V = 1+1+1+2+2+2+2 = 11\,, \quad E(V) = 72\,, \quad D(V) = 300\,,$$

dok je realizovana vrednost statistike $Z_0$

$$z_0 = \frac{11-72}{\sqrt{300}} = -3,523\,.$$

Kritična oblast ovog testa je $C = (-\infty\,;\,-1,645]$. Kako realizovana vrednost test statistike pripada kritičnoj oblasti, to nultu hipotezu odbacujemo u korist alternativne. Drugim rečima, na nivou značajnosti $\alpha = 0,05$ se može zaključiti da je primenom antitoksina u kombinaciji sa hemoterapijom značajno produžen život eksperimentalnih životinja.$\triangle$

# Glava 5

# Teorija odlučivanja

U mnogim slučajevima konačan cilj statističke analize se može interpretirati u obliku odlučivanja o odredjenom ponašanju ili delovanju. Evo nekoliko primera. Pri uzoračkoj kontroli proizvodnje, treba doneti jednu od dve odluke: prihvatiti ponudjenu partiju proizvoda ili je odbaciti. Zatim, lekar na osnovu analize simptoma bolesti kod odredjenog bolesnika mora da se ponese sa bolešću na jedan od konačno mnogo poznatih načina, tj. mora da donese jednu od konačno mnogo odluka kako da tretira bolesnika. Prilikom analize slučajnog procesa sa konačnim, ali nepoznatim očekivanjem, na osnovu rezultata posmatranja tog procesa, treba doneti odluku o veličini dejstva na proces (njegovu korekciju) za "pomeranje" očekivanja, na primer, u nulu. U poslednjem primeru to dejstvo može biti izraženo nekim realnim brojem $t$, pa je otuda u ovom slučaju broj mogućih odluka beskonačan.

U svim navedenim slučajevima odluka se donosi na osnovu analize posmatranja, uzorka $\mathbf{X}$, odnosno realizovanog uzorka $\mathbf{x}$ odgovarajućeg obeležja $X$ i kao posledica toga, **odluka** $d$ predstavlja vrednost funkcije $w(\mathbf{x})$ definisane na uzoračkom prostoru $\mathcal{X}$ čiji je kodomen **skup mogućih odluka** $\mathrm{D} = \{d\}$ u datoj situaciji. Na taj način, statistika $w(\mathbf{X})$,

$$w : \mathcal{X} \longrightarrow \mathrm{D}$$

je pravilo koje svaki rezultat posmatranja $\mathbf{x} \in \mathcal{X}$ dovodi u vezu sa odlukom $d = w(\mathbf{x}) \in \mathrm{D}$. Funkcija $w$ se zove **funkcija odluke (procedura)** i ona se bira na osnovu nekog kriterijuma optimalnosti. Princip rešavanja tog zadatka zove se teorija odlučivanja (definisao ju je Wald 1950. godine).

Neka je $(\mathcal{X}, \mathcal{B}, \mathrm{P})$ prostor verovatnoća koji odgovara statističkom modelu sa uzorkom fiksiranog obima. To znači da je $\mathcal{X}$ – konačnodimenzioni euklidski prostor, $\mathcal{B}$ – Borelovo $\sigma$–polje na $\mathcal{X}$, $\mathrm{P} \in \mathcal{P}$, gde je $\mathcal{P}$ – familija verovatnoća.

S obzirom na cilj ovog poglavlja, samo da predstavi koncept statisitčke teorije odlučivanja, nadalje ćemo pretpostavljati da je familija $\mathcal{P}$ opisana tako da $\mathrm{P} \in \mathcal{P}$ zavisi od parametra $\theta$ (jedno ili višedimenzionog), tj. $\mathcal{P} = \{\mathrm{P}_\theta; \theta \in \mathbf{\Theta}\}$ na $(\mathcal{X}, \mathcal{B})$, u zamenu za opšti model ove teorije.

Sa $\mathbf{\Theta}$ ćemo, kao i do sada, označavati parametarski prostor, pri čemu ćemo još smatrati da je $\mathbf{\Theta}$ – otvoren skup $k$–dimenzionog euklidskog prostora, $k \geq 1$.

Neka su zadati familija dopustivih raspodela $\{F(x;\theta), \theta \in \mathbf{\Theta}\}$, kojoj po pretpostavci pripada raspodela posmatranaog obeležja $X$, i skup odluka $\mathrm{D} = \{d\}$. Na osnovu uzorka

treba doneti odluku o izboru jedne od funkcija raspodele iz date familije. S obzirom da će se izbor funkcije raspodele izvršiti izborom vrednosti parametra $\theta$, to je u ovom slučaju kodomen funkcije odluke skup $\mathbf{\Theta}$. Prema tome,

$$w : \mathcal{X} \longrightarrow \mathbf{\Theta} \subseteq \mathrm{D}.$$

Dakle, na osnovu uzorka $\mathbf{x} \in \mathcal{X}$ treba doneti odluku $d \in \mathrm{D}$, gde je skup D sada skup raspoloživih vrednosti parametra $\theta$.

Da bismo postavili kriterijume izbora funkcije odluke, neophodno je uporediti rezultate korišćenja različitih pravila $w$. U tu svrhu se definiše **funkcija gubitka**:

$$\mathcal{L} : \mathbf{\Theta} \times D \longrightarrow [0, +\infty).$$

Funkcija gubitka meri grešku koju bismo načinili pri donošenju odluke $d$, a da je pri tome prava vrednost parametra baš $\theta$. Prema tome je $\mathcal{L}(\theta, \theta) = 0$. Funkciju $\mathcal{L}$ treba razumevati kao gubitak (na tačnosti) usled prihvatanja odluke $d$ pod uslovom da je raspodela obeležja $X$ tačno $F(x; \theta)$.

Po pravilu se funkcija gubitka traži u nekom odredjenom skupu funkcija, tj. skupu funkcija sa odredjenim svojstvom. Tako, na primer, funkcija gubitka se često definiše kao: $\mathcal{L}(\theta, d) = \lambda(\theta)W(|d - \theta|)$, gde je $W$ monotono rastuća realna funkcija realne promenljive $t \geq 0$, takva da je $W(0) = 0$. Što se tiče funkcije $\lambda$, pretpostavlja se da je pozitivna, konačna i $\mathcal{S}$– izmerljiva, gde je $\mathcal{S}$ Borelovo $\sigma$– polje na $\mathbf{\Theta}$.

Da bi se izbegle teškoće, često se uvodi pretpostavka da je $\mathcal{L}$ ograničena funkcija parametra $\theta$ za svako $d \in \mathrm{D}$. Ovu ćemo pretpostavku i mi usvojiti. Nekada se za funkciju $\mathcal{L}$ uzimaju samo neprekidne funkcije po oba argumenta, a nekada ravnomerno ograničene po $(\theta, d)$.

**Primer 81.** Pretpostavimo da parametarski prostor $\mathbf{\Theta}$ sadrži tačno dve tačke, tj. $\mathbf{\Theta} = \{0, 1\}$, a skup odluka neka je $\mathrm{D} = \{\mathrm{d}|0 \leq \mathrm{d} \leq 1\}$. Funkciju gubitka možemo izabrati na sledeći način

$$\mathcal{L}(\theta, w) = |w(\mathbf{x}) - \theta|^{\alpha} \quad , \quad \alpha \in N\,.$$

Ovde je $\lambda \equiv 1$, $W(t) = t^{\alpha}$, $t \geq 0$. $\triangle$

Za svaku proceduru $w$ definiše se dalje **funkcija rizika** u oznaci $R(\theta,w)$, takva da za svaku odluku $d{=}w(\mathbf{x}) \in \mathrm{D}$ ona predstavlja uslovno očekivanje funkcije gubitka pod uslovom da je $\theta$ prava vrednost parametra:

$$R(\theta, w) = E(\mathcal{L}(\theta, w(\mathbf{X}))|\theta) = E_{\theta}(\mathcal{L}(\theta, d)).$$

Funkcija rizika daje kriterijum uporedjivanja različitih pravila odlučivanja na sledeći način: ako imamo dve funkcije odluke (dva pravila odlučivanja) $w_1$ i $w_2$, takve da je

$$R(\theta, w_1) \leq R(\theta, w_2), \quad \forall \theta \in \mathbf{\Theta} \tag{5.1}$$

sa strogom nejednakošću za bar jedno $\theta$, tada je pravilo $w_1$ bolje od pravila $w_2$, jer $w_1$ dovodi do manjeg očekivanog gubitka.

S druge strane, moguće je da funkcije odluke $w_1$ i $w_2$ budu neuporedive po navedenom kriterijumu, tj. da za neke vrednosti $\theta$ bude $R(\theta, w_1) < R(\theta, w_2)$, dok je za ostale vrednosti $\theta$ znak nejednakosti suprotno usmeren. U takvim situacijama je neophodno pribaviti dodatne informacije o problemu, da bi se mogao izvršiti izbor procedure.

**Primer 82.** Neka je dat prost uzorak obima 25, $\mathbf{X} = (X_1, \ldots, X_{25})$, iz populacije sa obeležjem $X$ čija raspodela pripada familiji dopustivih raspodela $\{\mathcal{N}(\theta, 1), -\infty < \theta < \infty\}$, i neka je $Y = \overline{X}_{25}$. Zatim, neka je:

$$\mathcal{L}(\theta, w(y)) = (w(y) - \theta)^2$$

Razmotrićemo dve moguće funkcije odluke: $w_1(\mathbf{x}) = y$ i $w_2(\mathbf{x}) = 0$. One redom daju rizike

$$R(\theta, w_1) = E(Y - \theta)^2 = D(Y) = \frac{1}{25} \quad \text{i}$$

$$R(\theta, w_2) = E(0 - \theta)^2 = \theta^2 \,.$$

Vidimo da, ako je prava vrednost parametra $\theta = 0$, onda je $w_2$ odlična funkcija odluke jer je funkcija rizika jednaka nuli. Medjutim, ako se $\theta$ razlikuje od nule čak i vrlo malo, recimo da je $\theta = 2$, onda je:

$$R(2, w_1) = \frac{1}{25} \quad , \quad R(2, w_2) = 4 \,, \quad \text{pa je} \quad R(2, w_2) > R(2, w_1) \,.$$

Takodje vidimo da je

$$R(\theta, w_2) \leq R(\theta, w_1) \quad \text{ako je} \quad -\frac{1}{5} \leq \theta \leq \frac{1}{5} \,. \triangle$$

Funkcija odluke $w$ se naziva **nedopustivom** ako postoji $w'$ koja je bolja od $w$ u smislu navedenog kriterijuma (5.1). U suprotnom, odluka (funkcija odluke) $w$ je **dopustiva**.

**Primer 83.** Ako bismo se u primeru 82 ograničili na takve funkcije odluke (statistike) za koje je $E(w(\mathbf{X})) = \theta$, tada $w_2$ ne bi bila medju dopustivim funkcijama odluka. $\triangle$

Ako je klasa dopustivih funkcija odluke jednočlana, tada se može govoriti o optimalnoj (najboljoj) odluci .

U teoriji odlučivanja se tradicionalno primenjuju dva prilaza (rasudjivanja): *Bajesov* i *minimaksni*.

## 5.1 Minimaks odlučivanje

Minimaks princip se sastoji u tome da se donese odluka koja minimalizuje maksimalni rizik. Ova, inače jednostavna ideja, postavlja pred ocenjivača često neostvariv zahtev, jer funkcija odluke $w(\mathbf{x})$ koja minimalizuje rizik za jednu vrednost parametra $\theta$, za drugu to ne mora da čini, kao što smo videli u primeru 82.

**Primer 84.** Ako se još jednom vratimo na primer 82 i odredimo novi kriterijum za funkciju odluke, to može da bude ograničenje tipa da zahtevamo funkciju odluke koja minimalizuje maksimum funkcije rizika. U tom slučalu $w_2$ ne bi bila dobra zato što je $R(\theta, w_2) = \theta^2$ neograničena funkcija na parametarskom prostoru $\mathbf{\Theta} = (-\infty, +\infty)$. S druge strane:

$$\max_{\theta} R(\theta, w_1) = \max_{\theta} \left(\frac{1}{25}\right) = \frac{1}{25} \quad , \quad -\infty < \theta < \infty.$$

Kriterijum koji smo ovde primenili zove se minimaksni kriterijum. Prema ovom kriterijumu se može pokazati da je $w_1(\mathbf{x}) = y = \overline{x}_{25}$ najbolja funkcija odluke ako je funkcija gubitka $\mathcal{L}(\theta, w(\mathbf{x})) = (\theta - w(\mathbf{x}))^2$ .△

Ovim primerom smo ilustrovali sledeće:

a) Bez nekakvog ograničenja za funkciju odluke veoma je teško naći funkciju odluke koja ima funkciju rizika uniformno manju od funkcije rizika druge funkcije odluke.

b) Princip izbora najbolje funkcije odluke koji se zove **minimaks princip**.

Najzad definišimo precizno minimaksnu funkciju odluke :

Definicija 41 Ako je funkcija odluke $w_0(\mathbf{x})$ takva da za svaki $\theta \in \mathbf{\Theta}$

$$\max_{\theta \in \mathbf{\Theta}} R(\theta, w_0(\mathbf{x})) \leq \max_{\theta \in \mathbf{\Theta}} R(\theta, w(\mathbf{x}))$$

za bilo koju drugu funkciju odluke $w(\mathbf{x})$, tada se $w_0(\mathbf{x})$ zove *minimaksna funkcija odluke.*

Posvetimo sada malo pažnje drugom osnovnom principu statističkog zaključivanja, testiranju staističkih hipoteza, na bazi minimaks principa. Testiraćemo hipotezu o nepoznatom parametru $\theta$ kod dvočlanog parametarskog prostora $\mathbf{\Theta} = \{\theta_0, \theta_1\}$.

Ranije smo već istakli da se testiranje hipoteze $H_0 : \theta = \theta_0$ protiv $H_1 : \theta = \theta_1$ može da izrazi u terminima kritične oblasti u uzoračkom prostoru. Isto se može učiniti i kod minimaks postupka za testiranje hipoteza. Naime možemo da izaberemo podskup $C$ uzoračkog prostora $\mathcal{X}$ i ako $\mathbf{x} \in C$ prihvatamo hipotezu $H_1$, odnosno donosimo odluku $w(\mathbf{x}) = \theta_1$, a ukoliko $\mathbf{x} \in C^c$ odlučićemo $w(\mathbf{x}) = \theta_0$. Na taj način kritična oblast $C$ odredjuje funkciju odluke. U tom smislu funkciju rizika možemo označiti sa $R(\theta, C)$ umesto sa $R(\theta, w)$. Dakle,

$$R(\theta, C) = R(\theta, w) = \int_{C \cup C^c} \mathcal{L}(\theta, w) dF_\theta(\mathbf{x}).$$

Otuda

$$R(\theta, C) = \int_C \mathcal{L}(\theta, \theta_1) dF_\theta(\mathbf{x}) + \int_{C^c} \mathcal{L}(\theta, \theta_0) dF_\theta(\mathbf{x}).$$

Za $\theta = \theta_0$ dobija se

$$R(\theta_0, C) = \mathcal{L}(\theta_0, \theta_1) \int_C dF_{\theta_0}(\mathbf{x}),$$

a za $\theta = \theta_1$ se dobija

$$R(\theta_1, C) = \mathcal{L}(\theta_1, \theta_0) \int_{C^c} dF_{\theta_1}(\mathbf{x}).$$

Ako je $M(\theta)$ funkcija moći testa koja odgovara kritičnoj oblasti $C$, tada je

$$R(\theta_0, C) = \mathcal{L}(\theta_0, \theta_1) M(\theta_0) = \mathcal{L}(\theta_0, \theta_1) \alpha$$

i

$$R(\theta_1, C) = \mathcal{L}(\theta_1, \theta_0)\left(1 - \int_C dF_{\theta_1}(\mathbf{x})\right) = \mathcal{L}(\theta_1, \theta_0)(1 - M(\theta_1)) = \mathcal{L}(\theta_1, \theta_0)\beta,$$

gde su $\alpha$ i $\beta$ verovatnoća greške prve i druge vrste redom.

Minimaksno rešenje našeg problema bila bi kritična oblast $C$ za koju bi

$$max\{R(\theta_0, C), R(\theta_1, C)\}$$

bio minimalan. Dokazaćemo da, ako uvedemo dodatni uslov da je

$$R(\theta_0, C) = R(\theta_1, C)\,, \tag{5.2}$$

u tom slučaju minimaksno rešenje za testiranje proste nulte protiv takodje proste alternativne hipoteze, Nejman-Pirsonova najbolja kritična oblast veličine $\alpha$ za testiranje istih hipoteza. Dakle, dokazaćemo da je oblast

$$C = \left\{(x_1, x_2, \ldots, x_n)| \quad \frac{L(\theta_0; x_1, x_2, \ldots, x_n)}{L(\theta_1; x_1, x_2, \ldots, x_n)} \leq k\right\}, \tag{5.3}$$

gde je $k > 0$ izabrano tako da bude zadovoljen uslov (5.2). Takvo $k$ u apsolutno neprekidnom slučaju uvek postoji dok u diskretnom slučaju može da se desi da nam je potreban i pomoćni eksperiment za koji je

$$\frac{L(\theta_0; x_1, x_2, \ldots, x_n)}{L(\theta_1; x_1, x_2, \ldots, x_n)} = k,$$

da bismo postigli $R(\theta_0, C) = R(\theta_1, C)$. Da bismo dokazali da je $C$ definisana u (5.3) minimaksno rešenje, posmatrajmo proizvoljnu drugu kritičnu oblast $A \subset R^n$ veličine $\alpha$ za koju je

$$R(\theta_0, C) \geq R(\theta_0, A). \tag{5.4}$$

Oblasti $A$ u kojima je $R(\theta_0, C) < R(\theta_0, A)$ ne treba ni razmatrati, jer bi u tom slučaju očigledno bilo

$$R(\theta_0, C) = R(\theta_1, C) < max\{R(\theta_0, A), R(\theta_1, A)\}.$$

Uslov (5.4) je ekvivalentan sa

$$\mathcal{L}(\theta_0, \theta_1)\int_C dF_{\theta_0}(\mathbf{x}) \geq \mathcal{L}(\theta_0, \theta_1)\int_A dF_{\theta_0}(\mathbf{x}).$$

Otuda

$$\int_C dF_{\theta_0}(\mathbf{x}) \geq \int_A dF_{\theta_0}(\mathbf{x}).$$

Kako je

$$\alpha = \int_C dF_{\theta_0}(\mathbf{x}),$$

to je

$$\alpha = \int_A dF_{\theta_0}(\mathbf{x}).$$

Medjutim, prema Nejman-Pirsonovoj teoremi, $C$ je najbolja kritična oblast veličine $\alpha$, dakle, ona kritična oblast veličine $\alpha$ za koju je $\beta$ najmanje, tj.

$$\int_{C^c} dF_{\theta_1}(\mathbf{x}) \leq \int_{A^c} dF_{\theta_1}(\mathbf{x}).$$

Otuda,

$$\mathcal{L}(\theta_1, \theta_0) \int_{C^c} dF_{\theta_1}(\mathbf{x}) \leq \mathcal{L}(\theta_1, \theta_0) \int_{A^c} dF_{\theta_1}(\mathbf{x}).$$

Što znači da je

$$R(\theta_1, C) \leq R(\theta_1, A),$$

odnosno,

$$\max\{R(\theta_0, C), R(\theta_1, C)\} \leq \max\{R(\theta_0, A), R(\theta_1, A)\}$$

što je i trebalo dokazati.

## 5.2 Bajesovo odlučivanje

Kod Bajesovog ocenjivanja se parametar $\theta$ posmatra kao slučajna veličina. Dakle, za $\theta$ se definiše prostor verovatnoća $(\mathbf{\Theta}, \mathcal{S}, \mathrm{Q})$, gde je $\mathcal{S}$ – Borelovo $\sigma$-polje na $\mathbf{\Theta}$, a Q – mera, tj. verovatnoća na $(\mathbf{\Theta}, \mathcal{S})$. Meru Q nazivamo **apriornom raspodelom** parametra $\theta$. Pri tome se pretpostavlja da apriorna raspodela pripada nekoj familiji apriornih raspodela $\mathcal{H}$. Za svako fiksirano $\theta$ sa $f(\mathbf{x}; \theta)$ označavaćemo gustinu raspodele uzorka koja odgovara raspodeli verovatnoća P, a sa $h(\theta)$ apriornu gustinu raspodele koja odgovara raspodeli verovatnoća Q, sa funkcijom raspodele H. Dakle, $f(\mathbf{x}; \theta)$ je uslovna gustina raspodele za $\mathbf{X}$ pod uslovom da je prava vrednost parametra baš $\theta$. Ona je $\mathcal{S}$–izmerljiva po $\theta$ za svaki $\mathbf{x} \in \mathcal{X}$. Tada je

$$g(\mathbf{x}; \theta) = f(\mathbf{x}; \theta) h(\theta), \ \mathbf{x} \in \mathcal{X}, \ \theta \in \mathbf{\Theta}$$

gustina zajedničke raspodele za višedimenzionu slučajnu promenljivu $(\mathbf{X}, \theta)$ na prostoru $(\mathcal{X} \times \mathbf{\Theta}, \mathcal{B} \times \mathcal{S})$. Uslovna gustina raspodele

$$h(\theta | \mathbf{X} = \mathbf{x}) = \frac{f(\mathbf{x}; \theta) h(\theta)}{\int_{\mathbf{\Theta}} f(\mathbf{x}; \tau) dH(\tau)}$$

za svako $\mathbf{x} \in \mathcal{X}$ za koje je $f(\mathbf{x}) = \int_{\mathbf{\Theta}} f(\mathbf{x}; \tau) dH(\tau) > 0$, zove se **aposteriorna gustina raspodele** parametra $\theta$ .

Neka je $F_\theta(\mathbf{x})$ odgovarajuća funkcija raspodele za gustinu $f(\mathbf{x}; \theta)$. Označimo sa $\mathrm{H}(\theta)$, $\theta \in \mathbf{\Theta}$, apriornu funkciju raspodele na $\mathbf{\Theta}$ kojoj odgovara apriorna gustina raspodele $h(\theta)$. Neka je sa $\mathrm{H}(\theta | \mathbf{X} = \mathbf{x})$ označena aposteriorna raspodela (funkcija raspodele) parametra $\theta$ pri zadatom $\mathbf{X} = \mathbf{x}$. Marginalna (bezuslovna) gustina za $\mathbf{X}$ se tada može posmatrati kao

$$f_H(\mathbf{x}) = \int_{\mathbf{\Theta}} f(\mathbf{x}; \theta) dH(\theta) \,,$$

sa odgovarajućom bezuslovnom funkcijom raspodele u oznaci $F_H(\mathbf{x})$, a **aposteriorna gustina raspodele** za $\theta$ pri zadatom $\mathbf{X} = \mathbf{x}$ ima oblik

$$h(\theta | \mathbf{X} = \mathbf{x}) = \frac{f(\mathbf{x}; \theta) h(\theta)}{f_H(\mathbf{x})} \,.$$

Nadalje ćemo razmotriti neke osobine Bajesove ocene za funkciju gubitka oblika $\mathcal{L}(\theta, d) = \lambda(\theta)W(|d-\theta|)$, $d = w(\mathbf{x})$. Istaknimo ponovo da je funkcija rizika odluke $d$:

$$R(\theta, w) = \lambda(\theta)\int_{\mathcal{X}} W(|w(\mathbf{x}) - \theta|)dF_\theta(\mathbf{x}).$$

S obzirom da je $\mathcal{L}(\theta, d)$ ograničena funkcija po $\theta$ za svaki $d$, to je i $R(\theta, d)$ ograničena za svaki $d$. S druge strane, ako uvedemo pretpostavku o apriornoj raspodeli parametra $\theta$, **apriorni rizik ocene** $d$ **u odnosu na apriornu funkciju raspodele H** je apriorno očekivanje funkcije rizika definisano sa

$$R(H, w) = \int_{\mathbf{\Theta}} R(\theta, w)dH(\theta) = E(R(\theta, w)).$$

Dok je apriorni rizik uslovno matematičko očekivanje funkcije rizika pod uslovom da je učinjena odluka $d$, dotle je **aposteriorni rizik** funkcije $w(\mathbf{x}) = d$ pri zadanom $\mathbf{X} = \mathbf{x}$:

$$\int_{\mathbf{\Theta}} \lambda(\theta)W(|w(\mathbf{x}) - \theta|)dH(\theta|\mathbf{x})\,.$$

Dakle, aposteriorni rizik je očekivanje funkcije gubitka u odnosu na aposteriornu raspodelu parametra. Otuda je očigledno da je apriorni rizik $R(H, d)$ očekivanje aposteriornog rizika u odnosu na $F_H(\mathbf{x})$ – raspodelu uzorka $\mathbf{X}$ pri apriornoj raspodeli H. Zaista,

$$R(H, w) = \int_{\mathbf{\Theta}} R(\theta, w)dH(\theta) = \int_{\mathbf{\Theta}} \left(\int_{\mathcal{X}} \mathcal{L}(\theta, w(\mathbf{x}))dF_\theta(\mathbf{x})\right) dH(\theta)\,.$$

Ukoliko je u pitanju apsolutno neprekidna raspodela obeležja $X$, ali i apsolutno neprekidna raspodela (apriorna i aposteriorna) nepoznatog parametra $\theta$, apriorni rizik postaje:

$$R(H, w) = \int_{\mathbf{\Theta}} R(\theta, w)h(\theta)d\theta = \int_{\mathbf{\Theta}}\int_{\mathcal{X}} \mathcal{L}(\theta, w(\mathbf{x}))f_H(\mathbf{x})h(\theta|\mathbf{x})d\mathbf{x}d\theta =$$

$$= \int_{\mathcal{X}} \left(\int_{\mathbf{\Theta}} \mathcal{L}(\theta, w(\mathbf{x}))h(\theta|\mathbf{x})d\theta\right) f_H(\mathbf{x})d\mathbf{x}.$$

S bajesovske tačke gledišta, pošto je $\mathbf{X}$ realizovano, najpogodnija funkcija za dalje razmatranje je aposteriorni rizik, a ne apriorni. Znači, **bajesovska ocena** parametra $\theta$ u odnosu na apriornu raspodelu H je takva ocena (odluka) iz D koja minimizira aposteriorni rizik pri zadatom $\mathbf{X} = \mathbf{x}$. Ako označimo sa $\hat{\theta}_H(\mathbf{X})$ bajesovsku ocenu, onda je, za ranije komentarisanu funkciju gubitka,

$$\int_{\mathbf{\Theta}} \lambda(\theta)W(|\hat{\theta}_H(\mathbf{x}) - \theta|)dH(\theta|\mathbf{x}) = \inf_{d\in D}\int_{\mathbf{\Theta}} \lambda(\theta)W(|d-\theta|)dH(\theta|\mathbf{x}).$$

Za zadato H, Bajesova ocena ne mora biti jedinstvena. Ona će biti jedinstvena u slučaju da je funkcija gubitka strogo konveksna. Bajesova ocena minimizira i apriorni rizik (sledi iz leme Fatua). Mnogi autori je i definišu kao onaj element iz D koji minimizira apriorni rizik. Obe definicije vode ka istom rešenju.

**Primer 85.** Vratimo se primeru 81. Neka je apriorna raspodela za $\theta$:

$$P\{\theta = 0\} = \frac{3}{4} \quad , \quad P\{\theta = 1\} = \frac{1}{4}\,.$$

Neka je $\alpha = 1$, tada je apriorna funkcija rizika

$$R(\theta, w) = E\{\mathcal{L}(\theta, w(\mathbf{X}))|\theta\} = \mathcal{L}(0, w(\mathbf{x}))\cdot\frac{3}{4} + \mathcal{L}(1, w(\mathbf{x}))\cdot\frac{1}{4} = |d|\cdot\frac{3}{4} + |1-d|\cdot\frac{1}{4} = \frac{1}{2}d + \frac{1}{4}\,.$$

Dakle,

$$\inf_{\mathrm{d}\in\mathrm{D}} R(\theta, \mathrm{d}) = \frac{1}{4}\,,$$

tj. jedinstveno Bajesovo rešenje je $w(\mathbf{x}) = 0$. Skrenimo pažnju na činjenicu da u slučaju $D = \{d|0 < d \leq 1\}$ minimum funkcije rizika bio bi isti, $\frac{1}{4}$, ali nijedna odluka $\mathrm{d} \in \mathrm{D}$ ne bi bila Bajesovo rešenje.

Neka je sada $\alpha > 1$. Tada je funkcija rizika

$$R(\theta, \mathrm{d}) = \frac{3}{4}\mathrm{d}^\alpha + \frac{1}{4}\cdot(1-\mathrm{d})^\alpha.$$

Funkcija rizika ima minimum za odluku

$$\mathrm{d} = \left(1 + 3^{\frac{1}{\alpha-1}}\right)^{-1} \in (0,1)\,,$$

koja jeste Bajesovo rešenje. $\triangle$

Može se govoriti i o dovoljnim statistikama u Bajesovom smislu, o čemu ovde neće biti reči.

Na Bajesovom principu se zasnivaju ne samo tačkaste, već i ocene parametara oblastima. Ovde će biti reči samo o oceni jednodimenzionog parametra $\theta$, tj. o intervalu kao oceni jednodimenzionog parametra. Dakle, treba odrediti dve statistike, tj. dve procedure $w_1(\mathbf{X})$ i $w_2(\mathbf{X})$ takve da je

$$P\{w_1(\mathbf{X}) < \theta < w_2(\mathbf{X})|\mathbf{X} = \mathbf{x}\} = \gamma = 1 - \alpha$$

za unapred zadati nivo poverenja $\gamma$. To znači da

$$\int_{[w_1(\mathbf{x}),\ w_2(\mathbf{x})]} dH(\theta|\mathbf{x}) = \gamma\,. \tag{5.5}$$

**Primer 86.** Na osnovu uzorka $(X_1, X_2, \ldots, X_n)$ odrediti Bajesov interval poverenja za matematičko očekivanje obeležja $X$ čija raspodela pripada familiji dopustivih raspodela $\{\mathcal{N}(\theta, \sigma^2),\ -\infty < \theta < +\infty\}$ pod pretposravkom da je apriorna raspodela parametra $\theta$ takodje normalna: $\mathcal{N}(\mu,\, \nu^2)$. Parametri $\sigma^2, \mu$ i $\nu^2$ su poznati.

Ako kao proceduru $w$ izaberemo funkciju sredine uzorka, $w(\mathbf{X}) = \varphi(\overline{X}_n)$, označimo, sa $Y = \overline{X}_n$, treba doneti dve odluke $w_1(\mathbf{x})$ i $w_2(\mathbf{x})$, tj $\varphi_1(y)$ i $\varphi_2(y)$ koje će zadovoljiti gore navedeni uslov (5.5). Uslovna raspodela za $Y$ pod uslovom $\theta$ je $\mathcal{N}(\theta,\, \frac{\sigma^2}{n})$ što će dati odgovarajuću uslovnu gustinu raspodele

$$f(y;\theta) = \sqrt{\frac{n}{2\pi\sigma^2}}\exp\left(-\frac{(y-\theta)^2}{\frac{2\sigma^2}{n}}\right).$$

Zajednička gustina raspodele za vektor $(Y, \theta)$ je tada:

$$g(y;\theta) = \frac{\sqrt{n}}{2\pi\nu\sigma} \exp\left(-\frac{(\theta-\mu)^2}{2\nu^2} - \frac{n(y-\theta)^2}{2\sigma^2}\right),$$

tj. posmatrani vektor ima raspodelu

$$\mathcal{N}\left(\mu\,,\,\mu\,,\,\nu^2+\frac{\sigma^2}{n}\,,\,\nu^2\,,\,\frac{\nu}{\sqrt{\nu^2+\frac{\sigma^2}{n}}}\right),$$

gde je

$$\frac{\nu}{\sqrt{\nu^2+\frac{\sigma^2}{n}}}$$

koeficijent korelacije komponenata.

Aposteriorna gustina raspodele za $\theta$ pod uslovom $\mathbf{X} = \mathbf{x}$, odnosno, $Y = y$ biće:

$$h(\theta|Y=y) = \frac{1}{2\pi(\nu^2+\frac{\sigma^2}{n})} \exp\left(-\frac{(y-\mu)^2}{2(\nu^2+\frac{\sigma^2}{n})}\right),$$

što znači da je aposteriorna raspodela normalna

$$\mathcal{N}\left(\frac{\mu\sigma^2+ny\nu^2}{n\nu^2+\sigma^2},\ \frac{\nu^2\sigma^2}{n\nu^2+\sigma^2}\right).$$

Konačno, intervalna Bajesova ocena za $\theta$ odredjuje se iz

$$P\left(\frac{\mu\sigma^2+nY\nu^2}{n\nu^2+\sigma^2} - z\sqrt{\frac{\nu^2\sigma^2}{n\nu^2+\sigma^2}} < \theta < \frac{\mu\sigma^2+nY\nu^2}{n\nu^2+\sigma^2} + z\sqrt{\frac{\nu^2\sigma^2}{n\nu^2+\sigma^2}}\right) = \gamma\,,\ z \in R\,.\,\triangle$$

Bajesov koncept testiranja statističkih hipoteza izložićemo na primeru testiranja proste nulte hipoteze protiv alternativne takodje proste hipoteze. To je slučaj dvodimenzionog parametarskog prostora $\mathbf{\Theta} = \{\theta_0, \theta_1\}$. Dakle, slučajna promenljiva $\theta$ je diskretnog tipa i

$$h(\theta_0) + h(\theta_1) = 1.$$

U tom slučaju je marginalna gustina za $\mathbf{X}$:

$$\sum_{\mathbf{\Theta}} f(\mathbf{x};\theta)h(\theta) = f(\mathbf{x};\theta_0)h(\theta_0) + f(\mathbf{x};\theta_1)h(\theta_1),$$

odnosno, aposteriorna gustina za $\theta$ je:

$$h(\theta|\mathbf{X}=\mathbf{x}) = \frac{h(\theta)f(\mathbf{x};\theta)}{f(\mathbf{x};\theta_0)h(\theta_0) + f(\mathbf{x};\theta_1)h(\theta_1)}.$$

Testiraćemo hipotezu $H_0 : \theta = \theta_0$ protiv alternativne $H_1 : \theta = \theta_1$. Bajesovo rešenje će biti funkcija odluke $w(\mathbf{x})$ za koju je aposteriorni rizik minimalan. Dakle, treba minimalizovati

$$E(\mathcal{L}(\theta, w(\mathbf{X}))|\mathbf{X}=\mathbf{x}).$$

S obzirom da je skup mogućih odluka dvočlan, to treba razmotriti samo dve moguće vrednosti aposteriornog rizika i to:

1. $w = \theta_0$

$$\Rightarrow E(\mathcal{L}(\theta, w(\mathbf{X}))|\mathbf{X} = \mathbf{x}) = E(\mathcal{L}(\theta, \theta_0)|\mathbf{X} = \mathbf{x}) = \\ = \frac{\mathcal{L}(\theta_1, \theta_0)h(\theta_1)f(\mathbf{x}; \theta_1)}{f(\mathbf{x}; \theta_0)h(\theta_0) + f(\mathbf{x}; \theta_1)h(\theta_1)}$$

2. $w = \theta_1$

$$\Rightarrow E(\mathcal{L}(\theta, w(\mathbf{X}))|\mathbf{X} = \mathbf{x}) = E(\mathcal{L}(\theta, \theta_1)|\mathbf{X} = \mathbf{x}) = \\ = \frac{\mathcal{L}(\theta_0, \theta_1)h(\theta_0)f(\mathbf{x}; \theta_0)}{f(\mathbf{x}; \theta_0)h(\theta_0) + f(\mathbf{x}; \theta_1)h(\theta_1)}.$$

Prema tome, prihvatićemo hipotezu $H_i : \theta = \theta_i,\ i = 0, 1$ ako i samo ako je

$$\frac{\mathcal{L}(\theta_j, \theta_i)h(\theta_j)f(\mathbf{x}; \theta_j)}{f(\mathbf{x}; \theta_0)h(\theta_0) + f(\mathbf{x}; \theta_1)h(\theta_1)} < \frac{\mathcal{L}(\theta_i, \theta_j)h(\theta_i)f(\mathbf{x}; \theta_i)}{f(\mathbf{x}; \theta_0)h(\theta_0) + f(\mathbf{x}; \theta_1)h(\theta_1)},$$

za $j \neq i$. Ili što je isto

$$\frac{f(\mathbf{x}; \theta_j)}{f(\mathbf{x}; \theta_i)} < \frac{\mathcal{L}(\theta_i, \theta_j)h(\theta_i)}{\mathcal{L}(\theta_j, \theta_i)h(\theta_j)} \quad , \quad i, j = 0, 1\ ,\ i \neq j\ .$$

Ukoliko se dogodi da se umesto nejednakosti javlja jednakost, morali bismo da potražimo dodatne informacije, odnosno da obavimo neki pomoćni eksperiment koji bi nam pomogao da donesemo odluku.

Uočimo važnu činjenicu da se poslednjom nejednakošću koja je gore navedena opisuje baš najbolja kritična oblast testa prema tvrdjenju Nejman-Pirsonove teoreme s obzirom da se na levoj strani nejednakosti nalazi količnik verodostojnosti za koji se ova teorema vezuje, jer je

$$f(\mathbf{x}; \theta) \equiv L(\theta; x_1, x_2, \ldots, x_n) \quad , \quad \mathbf{x} = (x_1, x_2, \ldots, x_n) \in R^n\ ,$$

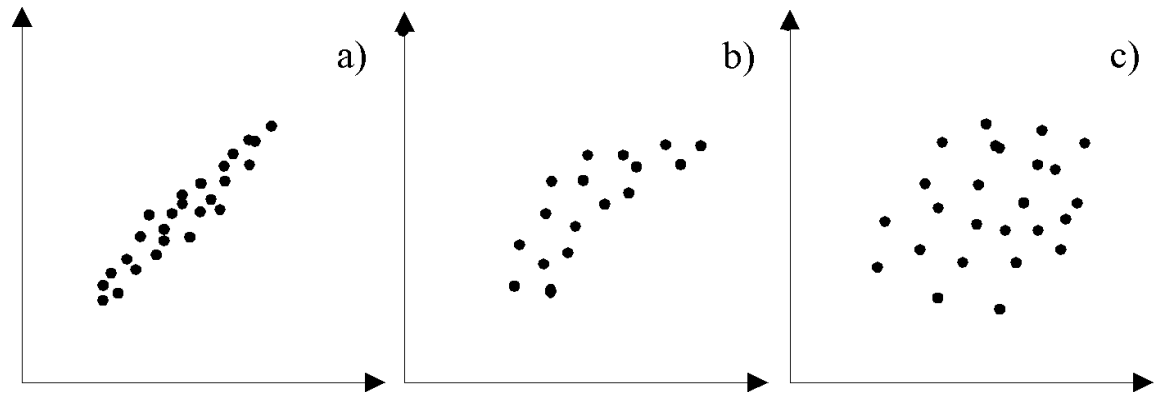
funkcija verodostojnosti.

# Glava 6

# Regresija

U velikom broju istraživanja ili eksperimenata uočava se veza izmedju dve ili više promenljivih veličina. Od istraživača se u tom slučaju očekuje da utvrdi da li postoji i kakva je direktna funkcionalna zavisnost medju tim veličinama. Na primeru dva svojstva $X$ i $Y$ koja se istražuju na nekom uzorku obima $n$, kao rezultat posmatranja dobija se $n$ uredjenih parova realizacija $(x_1, y_1)$, $(x_2, y_2)$, ..., $(x_n, y_n)$. Oni se mogu predstaviti u Dekartovoj ravni (slika 6.1), a grafička reprezentacija koja tom prilikom nastaje naziva se **dijagram rasturanja**, odnosno, **dijagram rasipanja**.

Ukoliko se posmatra $k$ obeležja na nekom uzorku obima $n$: $X_1, X_2, \ldots, X_k$, kao rezultat istraživanja javlja se $n$ uredjenih $k$-torki $(x_1^1, x_2^1, \ldots, x_k^1)$, $(x_1^2, x_2^2, \ldots, x_k^2)$, ..., $(x_1^n, x_2^n, \ldots, x_k^n)$, a dijagram rasturanja je skup od $n$ tačaka $k$-dimenzionalnog euklidskog prostora $E_k$. Na osnovu tih podataka (tačaka) pokušava se da se otkrije funkcionalna veza medju svojstvima, ako postoji. Opšti problem nalaženja funkcije koja dobro aproksimira dobijeni skup podataka, u statističkom žargonu se naziva "**fitovanje**[1] **krive**". Za odredjivanje odgovarajućeg tipa funkcionalne zavisnosti, u praksi se koristi upravo dijagram rasturanja. Radi ilustracije, to znači da za slučaj na slici 6.1 a) treba proveriti linearnu vezu $y = ax+b$, na istoj slici pod b) logaritamsku zavisnost tipa $y = a\ln(x + b)$, dok dijagram pod c) ne ukazuje ni na kakvu funkcionalnu zavisnost.




Slika 6.1: Dijagrami rasturanja tačaka.

[1]od engleskog glagola *to fit* – upasovati, prilagoditi, podesiti

Konstatujmo da se problem zavisnosti može posmatrati u dva pravca, koja će nadalje biti razjašnjena.

Posmatra se uticaj slučajnih veličina (obeležja) $X_1$, $X_2$, ..., $X_p$ na slučajnu veličinu $Y$, tj. uticaj slučajnog vektora $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ na slučajnu promenljivu $Y$. Pri tome svaka vrednost slučajnog vektora $\mathbf{X}$ proizvodi odgovarajuću vrednost slučajne promenljive (obeležja) $Y$. U tom slučaju postoji očigledna potreba da se utvrdi, što preciznije, oblik zavisnosti ovih slučajnih veličina. Kako svaka realizovana vrednost $\mathbf{x}$ slučajnog vektora $\mathbf{X}$ proizvodi realizovanu vrednost $y$ slučajne promenljive $Y$, zadatak se sastoji u nalaženju funkcije $f(\mathbf{x})$ koja "dobro" aproksimira vrednosti slučajne promenljive $Y$ pri svakoj realizovanoj vrednosti slučajnog vektora $\mathbf{X}$. Funkcija $f(\mathbf{X})$ koja zadovoljava uslov da je

$$E(Y - f(\mathbf{X}))^2$$

minimalno, uzima se kao dobra prognoza za $Y$ po $\mathbf{X}$. Najbolja prognoza $Y$ po $\mathbf{X}$ u smislu definisanog srednjekvadratnog odstupanja zove se **funkcija regresije $Y$ na X.** Ovaj tip regresije je **regresija prve vrste.**

Regresija druge vrste je drugi tip zavisnosti koji je takodje predmet statističkog proučavanja.

U većini eksperimentalnih istraživanja u laboratorijskim uslovima koncepcija je da se varira odredjeni broj neslučajnih veličina i posmatra njihov uticaj na ishod eksperimenta koji je slučajan. Neslučajne veličine o kojima je reč se nazivaju **kontrolisani faktori**. Ishod eksperimenta jeste slučajan, jer na posmatrano obeležje, osim kontrolisanih faktora, utiču po pravilu i slučajni faktori koji se ne mogu kontrolisati (na primer greške merenja), kao i neslučajni faktori koji su objektivno prisutni u eksperimentu, ali se njihov uticaj ne može sagledati ili pretpostaviti.

Slučajni ishod eksperimenta je obeležje $Y$ kojim se eksperiment opisuje. Neslučajna ulazna promenljiva se označava odgovarajućim malim slovom, recimo $x$, pri čemu se različiti posmatrani nivoi ovog faktora označavaju donjim ili gornjim indeksima, tj. $x_1$, $x_2$, ..., $x_n$ ili $x^1$, $x^2$, ..., $x^n$. Ovde ćemo koristiti drugu oznaku, a donjim indeksom ćemo označavati postojanje više neslučajnih faktora čije se dejstvo na obeležje $Y$ ispituje. Osim toga, gornji indeks ćemo stavljati u zagrade, da bi se vizuelno lakše razlikovao od stepena promenljive. Dakle, uticaj faktora $x_1$ na obeležje $Y$ ispituje se na $n$ nivoa: $x_1^{(1)}$, $x_1^{(2)}$, ..., $x_1^{(n)}$.

Za slučajni uticaj koji se pri eksperimentu ne može da kontroliše, koristi se oznaka $\varepsilon$. Ovaj koncept dovodi do konstrukcije modela **regresije druge vrste**. Kada se ispituje uticaj samo jednog faktora radi se o **jednostrukoj regresiji**, a kod ispitivanja istovremenog uticaja više faktora, reč je o **višestrukoj regresiji**. Kod ovog tipa regresije se pretpostavlja da se celokupan uticaj neslučajnih faktora sagledava kroz srednju vrednost obeležja $Y$, tj. da je slučajna komponenta aditivna i da joj je očekivanje nula.

Oba modela zavisnosti, regresija prve i regresija druge vrste, biće nadalje detaljnije razmatrana.

## 6.1 Linearna regresija druge vrste

Posebno mesto medju modelima regresije druge vrste imaju modeli linearne regresije kojima ćemo se nadalje baviti.

U opštem slučaju problem linearne regresije druge vrste polazi od pretpostavke da su matematička očekivanja opservacija $Y_i$, $i = 1, 2, \ldots, n$ linearne funkcije $\varphi_i(\boldsymbol{\beta})$ nepoznatih parametara $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_k)'$, koeficijenata regresije. Slučajnu veličinu $Y_i$ treba shvatiti kao ishod eksperimenta pri $i$–tom nivou posmatranih kontrolisanih faktora uticaja. Takodje se može da uvede pretpostavka o tome da posmatrani faktori utiču na ishod eksperimenta posredno, preko svojih funkcija $z_j = z_j(\mathbf{x})$, $j = 1, 2, \ldots, k$, $\mathbf{x} = (x_{j_1}, x_{j_2}, \ldots, x_{j_p})$, $j_l \in \{1, 2, \ldots, k\}$, $p \leq k$. Drugim rečima, u modelu linearne regresije koji je predmet proučavanja u ovom poglavlju, promenljive $z_1, \ldots, z_k$ mogu biti funkcionalno zavisne. Tako, sve promenljive $z_j$ mogu biti funkcije samo od jednog kontrolisanog faktora $x$. Na primer, $z_j = x^j$, $j > 1$ daje model koji je poznat pod imenom paraboličkа linearna regresija. Ovaj i još neke primere ćemo kasnije detaljnije obrazložiti.

Nivoi faktora $x_j$ će se tom prilikom ispoljiti kao "$i$"–ti nivo funkcije $z_j$ i u modelu ćemo ga označavati sa $z_j^{(i)}$. Nadalje ćemo razmatrati linearne funkcije od $\boldsymbol{\beta}$ oblika $\mathbf{z}^{(i)'}\boldsymbol{\beta}$, $\mathbf{z}^{(i)} = (z_1^{(i)}, \ldots, z_k^{(i)})'$, $i = 1, \ldots, n$ .

Pretpostavka koja se uvodi u model je, kao što smo rekli, da se celokupan uticaj neslučajnih faktora ostvaruje preko matematičkog očekivanja obeležja $Y$, tj. da je u modelu

$$Y_i = \mathbf{z}^{(i)'}\boldsymbol{\beta} + \varepsilon_i \quad , \quad i = 1, 2, \ldots, n \,, \tag{6.1}$$

$EY_i = \mathbf{z}^{(i)'}\boldsymbol{\beta}$ i $E\varepsilon_i = 0$, za svako $i$, i raspodele ostataka ("grešaka") $\varepsilon_i$ ne zavise od $\boldsymbol{\beta}$.

Planiranje eksperimenta u ovom slučaju podrazumeva uvodjenje matrice plana $\mathbf{Z} = ||\mathbf{z}^{(1)} \ldots \mathbf{z}^{(n)}||$ dimenzije $k \times n$, $k < n$ i vektora grešaka $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$ pa (6.1) dobija oblik

$$\mathbf{Y} = \mathbf{Z}'\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad , \quad E(\boldsymbol{\varepsilon}) = \mathbf{0} \,, \tag{6.2}$$

gde je $\mathbf{Y}$ vektor kolone slučajnih ishoda eksperimenta, $\mathbf{Y} = (Y_1, \ldots, Y_n)'$.

U model se obično uvodi i pretpostavka da su komponente vektora $\boldsymbol{\varepsilon}$ nekorelirane medju sobom i da imaju iste disperzije, što znači

$$D(Y_i) = D(\varepsilon_i) = \sigma^2 \quad , \quad i = 1, \ldots, n \quad \text{i} \quad Cov(\varepsilon_i, \varepsilon_j) = 0 \quad \text{za} \quad i \neq j \,.$$

U tom slučaju je kovarijansna matrica vektora $\mathbf{Y}$

$$\mathbf{D}(\mathbf{Y}) = \mathbf{D}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}_n \,, \tag{6.3}$$

gde je $\mathbf{I}_n$ jedinična matrica reda $n$. Ukoliko bi izostao uslov nekoreliranosti, autokovarijansna matrica vektora $\boldsymbol{\varepsilon}$ bi bila oblika

$$\mathbf{D}(\mathbf{Y}) = \mathbf{D}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{G} \,,$$

gde je matrica $\mathbf{G}$ simetrična pozitivno semidefinitna. Medjutim, slučaj nesingularne matrice $\mathbf{G}$ bi omogućio transformaciju vektora $\mathbf{Y}$ u vektor $\mathbf{W} = \mathbf{G}^{-1/2}\mathbf{Y}$ čija bi autokovarijansna matrica bila ponovo dijagonalna (6.3).

U modelu linearne regresije druge vrste važnu ulogu ima matrica

$$\mathbf{A} = \mathbf{Z}\mathbf{Z}'. \tag{6.4}$$

Ova matrica je kvadratna simetrična matrica reda $k$ i za nju važi tvrdjenje:

**Teorema 6.1.1** *Matrica* $\mathbf{A}$ *je pozitivno semidefinitna, a uslov* $rang\mathbf{Z} = k$ *je potreban i dovoljan da ona bude pozitivno definitna.*

**Dokaz.** Neka je $\mathbf{t} = (t_1, \ldots, t_k)'$ proizvoljan nenula vektor iz istog polja brojeva iz koga su elementi matrice $\mathbf{Z}$. Tada je

$$\mathbf{t}'\mathbf{A}\mathbf{t} = (\mathbf{Z}'\mathbf{t})'(\mathbf{Z}'\mathbf{t}) \geq 0\,,$$

pri čemu važi jednakost ako i samo ako je $\mathbf{Z}'\mathbf{t} = 0$, odnosno, ako i samo ako je $\sum_{j=1}^{k} t_j z_j^{(i)} = 0$ za svako $i = 1, \ldots, n$. Poslednji uslov je ekvivalentan sa uslovom

$$\sum_{j=1}^{k} t_j \mathbf{z}_j = \mathbf{0}\,, \tag{6.5}$$

gde je $\mathbf{z}_j$ vektor vrste matrice $\mathbf{Z}$. Poslednji iskaz izjednačava linearnu kombinaciju vektora vrsta matrice $\mathbf{Z}$ sa nula vektorom, što je svojstvo linearno zavisnih vrsta. Uslov (6.5) ekvivalentan je sa $rang\mathbf{Z} < k$. □

Osnovni zadatak statističkog postupka modela linearne regresije je ocenjivanje parametara regresionog modela. Pri tome se osim koeficijenata regresije, elemenata vektora $\boldsymbol{\beta}$, kao nepoznat parametar često javlja i disperzija $\sigma^2$. Za rešavanje ovog zadatka koristi se metod najmanjih kvadrata. Ovaj metod uveo je Gaus još 1809. godine.

### 6.1.1 Metod najmanjih kvadrata za ocenjivanje parametara modela linearne regresije

Primenićemo najpre metod najmanjih kvadrata na ocenjivanje vektora $\boldsymbol{\beta}$.

Definicija 42. Ocena nepoznatog vektora $\boldsymbol{\beta}$ dobijena metodom najmanjih kvadrata je vektor statistika $\widehat{\boldsymbol{\beta}}$ koje minimalizuju kvadratnu formu

$$S(\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{Z}'\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{Z}'\boldsymbol{\beta})\,, \tag{6.6}$$

dakle, vektor $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_k)'$ koji zadovoljava relaciju

$$S(\widehat{\boldsymbol{\beta}}) = \min_{\boldsymbol{\beta}} S(\boldsymbol{\beta})\,. \diamondsuit$$

Formula (6.6) predstavlja sumu kvadrata razlika slučajnih rezultata eksperimenta i njihovih matematičkih očekivanja.

Uobičajenim postupkom za odredjivanje minimuma diferencijabilne funkcije:

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \beta_j} = 0 \quad , \quad j = 1, \ldots, k \tag{6.7}$$

dobijamo sledeći sistem linearnih jednačina po nepoznatim parametrima $\beta_j$ , $j = 1, \ldots, k$

$$\sum_{i=1}^{n} z_j^{(i)} \left( \sum_{l=1}^{k} z_l^{(i)} \beta_l - Y_i \right) = 0 \quad , \quad j = 1, \ldots, k\,.$$

Koristeći matrični zapis i oznaku $\mathbf{ZY} = \mathbf{V}$ poslednji sistem se može zapisati kao

$$\mathbf{A}\boldsymbol{\beta} = \mathbf{V}\,. \tag{6.8}$$

Jednačina (6.8) nosi naziv **normalna jednačina metoda najmanjih kvadrata**, odnosno sistem linearnih jednačina koji se njome definiše nosi naziv **normalni sistem jednačina metoda najmanjih kvadrata**. O važnosti ovog sistema govori sledeća teorema.

**Teorema 6.1.2** *Neka je $\boldsymbol{\beta}^*$ proizvoljno rešenje normalne jednačine (6.8). Tada je*

$$S(\widehat{\boldsymbol{\beta}}) = S(\boldsymbol{\beta}^*) = \min_{\boldsymbol{\beta}} S(\boldsymbol{\beta})$$

*i minimum je isti za proizvoljno (svako) rešenje sistema (6.8).*

*Ako je $det\mathbf{A} \neq 0$, tada je ocena najmanjih kvadrata jedinstvena i jednaka*

$$\widehat{\boldsymbol{\beta}} = \mathbf{A}^{-1}\mathbf{V}\,.$$

**Dokaz.** Neka je $\boldsymbol{\beta}^*$ proizvoljno fiksirano rešenje jednačine (6.8). Tada je

$$\begin{aligned} S(\boldsymbol{\beta}) &= (\mathbf{Y} - \mathbf{Z}'\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{Z}'\boldsymbol{\beta}) = \\ &= (\mathbf{Y} - \mathbf{Z}'\boldsymbol{\beta}^* + \mathbf{Z}'\boldsymbol{\beta}^* - \mathbf{Z}'\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{Z}'\boldsymbol{\beta}^* + \mathbf{Z}'\boldsymbol{\beta}^* - \mathbf{Z}'\boldsymbol{\beta}) = \\ &= S(\boldsymbol{\beta}^*) + (\mathbf{V} - \mathbf{A}\boldsymbol{\beta}^*)'(\boldsymbol{\beta}^* - \boldsymbol{\beta}) + (\boldsymbol{\beta}^* - \boldsymbol{\beta})'(\mathbf{V} - \mathbf{A}\boldsymbol{\beta}^*) + (\boldsymbol{\beta}^* - \boldsymbol{\beta})'\mathbf{A}(\boldsymbol{\beta}^* - \boldsymbol{\beta}) = \\ &= S(\boldsymbol{\beta}^*) + (\boldsymbol{\beta}^* - \boldsymbol{\beta})'\mathbf{A}(\boldsymbol{\beta}^* - \boldsymbol{\beta}) \geq S(\boldsymbol{\beta}^*)\,, \end{aligned}$$

jer je matrica $\mathbf{A}$ pozitivno semidefinitna.

Za nesingularnu matricu $\mathbf{A}$ rešenje normalne jednačine je jedinstveno, pa je i ocena najmanjih kvadrata jedinstvena i ima oblik

$$\widehat{\boldsymbol{\beta}} = \mathbf{A}^{-1}\mathbf{V}\,. \quad \square$$

U mnogim praktičnim problemima primene od interesa su ne direktno koeficijenti regresije, već njihove linearne kombinacije. Reč je o statistikama za različite namene, a koje su funkcije od ocena koeficijenata regresije. Iz tog razloga ćemo nadalje razmatrati vektor linearnih kombinacija koeficijenata regresije $\mathbf{t} = \mathbf{T}\boldsymbol{\beta}$, gde je $\mathbf{T}$ zadata matrica reda $m \times k$, $m \leq k$. Ovo bi, na primer, bio slučaj kada se parametarski prostor $R^k$ za parametre koji predstavljaju koeficijente linearne regresije sužava nekim linearnim ograničenjima, tj. sistemom od $m$ linearnih ograničenja, sadržanih u vektoru $\mathbf{t}$ za neko $\mathbf{t} = \mathbf{t}_0$. U tom slučaju, kao ocenu najmanjih kvadrata vektora $\mathbf{t}$, u oznaci $\widehat{\mathbf{t}}$ imaćemo vektor statistika

$$\widehat{\mathbf{t}} = \mathbf{T}\widehat{\boldsymbol{\beta}}\,,$$

gde je, kao i do sada, $\widehat{\boldsymbol{\beta}}$ proizvoljno rešenje normalne jednačine. Jasno, ako je $det\mathbf{A} \neq 0$, $\widehat{\mathbf{t}}$ je jednoznačno odredjujen, tj.

$$\widehat{\mathbf{t}} = \mathbf{T}\mathbf{A}^{-1}\mathbf{V}\,. \tag{6.9}$$

Nadalje ćemo razmatrati svojstva ocena dobijenih metodom najmanjih kvadrata. Razmatraćemo zadatak ocenjivanja vektora $\mathbf{t}$ u klasi linearnih ocena, tj. razmatraćemo ocene oblika

$$\mathbf{l} = \mathbf{L}\mathbf{Y}$$

koje su statistike od posmatranja (vektora uzorka) $\mathbf{Y} = (Y_1, \ldots, Y_n)'$.

**Teorema 6.1.3** *Neka je matrica* $\mathbf{A}$ *nesingularna. Tada, za proizvoljan vektor* $\mathbf{t} = \mathbf{T}\boldsymbol{\beta}$ *ocena najmanjih kvadrata definisana relacijom (6.9) je nepristrasana ocena sa minimalnom disperzijom u odnosu na sve ostale linearne nepristrasne ocene za* $\mathbf{t}$*. Pri tome je kovarijansna matrica vektora* $\widehat{\mathbf{t}}$ *data sa*

$$\mathbf{D}(\widehat{\mathbf{t}}) = \sigma^2\mathbf{T}\mathbf{A}^{-1}\mathbf{T}',$$

*pod uslovom da je disperzija ostataka regresionog modela,* $\sigma^2$*, poznata.*

**Dokaz.** Nepristrasnost ocene $\widehat{\mathbf{t}}$ sledi direktno iz njene definicije i osobina matematičkog očekivanja.

Uporedimo sada proizvoljnu drugu linearnu nepristrasnu ocenu $\mathbf{l} = \mathbf{L}\mathbf{Y}$ vektora $\mathbf{t}$ sa ocenom $\widehat{\mathbf{t}}$. S obzirom da je $\mathbf{l}$ nepristrasna ocena, važi $E(\mathbf{l}) = \mathbf{t}$, odnosno $E(\mathbf{l}) = \mathbf{T}\boldsymbol{\beta}$. S druge strane,

$$E(\mathbf{l}) = \mathbf{L}E(\mathbf{Y}) = \mathbf{L}\mathbf{Z}'\boldsymbol{\beta}.$$

Otuda $\mathbf{L}\mathbf{Z}'\boldsymbol{\beta} = \mathbf{T}\boldsymbol{\beta}$. Poslednja jednakost mora da bude tačna za svaki vektor $\boldsymbol{\beta}$, pa odatle sledi da je $\mathbf{L}\mathbf{Z}' = \mathbf{T}$. Ostaje još da dokažemo da je u pitanju ocena sa najmanjom disperzijom medju svim linearnim nepristrasnim ocenama.

Disperzija ocene $\mathbf{l}$ je

$$\mathbf{D}(\mathbf{l}) = \mathbf{D}(\mathbf{L}\mathbf{Y}) = \mathbf{L}\mathbf{D}(\mathbf{Y})\mathbf{L}' = \mathbf{L}\sigma^2\mathbf{I}_n\mathbf{L}' = \sigma^2\mathbf{L}\mathbf{L}'.$$

Minimum disperzija ocena $l_1, l_2, \ldots, l_m$ komponenata vektora $\mathbf{l}$ će se ostvariti ukoliko su dijagonalni elementi glavne dijagonale kovarijansne matrice $\mathbf{D}(\mathbf{l})$ minimalni mogući. To znači da je potrebno minimalizovati elemente glavne dijagonale matrice $\mathbf{L}\mathbf{L}'$. S obzirom da za ovu matricu važi razlaganje

$$\begin{aligned}
\mathbf{L}\mathbf{L}' &= \mathbf{T}\mathbf{A}^{-1}\mathbf{T}' + \mathbf{L}\mathbf{L}' - \mathbf{T}\mathbf{A}^{-1}\mathbf{T}' - \mathbf{T}\mathbf{A}^{-1}\mathbf{T}' + \mathbf{T}\mathbf{A}^{-1}\mathbf{T}' = \\
&= \mathbf{T}\mathbf{A}^{-1}\mathbf{A}\mathbf{A}^{-1}\mathbf{T}' + \mathbf{L}\mathbf{L}' - \mathbf{L}\mathbf{Z}'\mathbf{A}^{-1}\mathbf{T}' - \mathbf{T}\mathbf{A}^{-1}\mathbf{Z}\mathbf{L}' + \mathbf{T}\mathbf{A}^{-1}\mathbf{A}\mathbf{A}^{-1}\mathbf{T}' = \\
&= (\mathbf{T}\mathbf{A}^{-1}\mathbf{Z})(\mathbf{T}\mathbf{A}^{-1}\mathbf{Z})' + (\mathbf{L} - \mathbf{T}\mathbf{A}^{-1}\mathbf{Z})(\mathbf{L} - \mathbf{T}\mathbf{A}^{-1}\mathbf{Z})',
\end{aligned}$$

to će njeni dijagonalni elementi dostići minimum ako i samo ako je drugi sabirak ovog razlaganja jednak nuli, tj. za

$$\mathbf{L} = \mathbf{T}\mathbf{A}^{-1}\mathbf{Z},$$

što je i trebalo dokazati. □

**Posledica 2.** *Kovarijansna matrica ocena najmanjih kvadrata vektora* $\boldsymbol{\beta}$ *u slučaju nesingularne matrice* $\mathbf{A}$ *je oblika*

$$\mathbf{D}(\widehat{\boldsymbol{\beta}}) = \sigma^2\mathbf{A}^{-1}.$$

**Dokaz.** Podjimo od kovarijansne matrice vektora $\widehat{\mathbf{t}}$

$$\mathbf{D}(\widehat{\mathbf{t}}) = \mathbf{D}(\mathbf{T}\widehat{\boldsymbol{\beta}}) = \mathbf{T}\mathbf{D}(\widehat{\boldsymbol{\beta}})\mathbf{T}'.$$

No, prema prethodnoj teoremi je

$$\mathbf{D}(\widehat{\mathbf{t}}) = \sigma^2\mathbf{T}\mathbf{A}^{-1}\mathbf{T}',$$

pa je

$$\sigma^2 \mathbf{T}\mathbf{A}^{-1}\mathbf{T}' = \mathbf{T}\mathbf{D}(\widehat{\boldsymbol{\beta}})\mathbf{T}' \,,$$

odnosno

$$\mathbf{D}(\widehat{\boldsymbol{\beta}}) = \sigma^2 \mathbf{A}^{-1} \,. \square$$

Dakle, poslednja teorema nam daje optimalnu linearnu ocenu za proizvoljnu linearnu kombinaciju koeficijenata linearne regresije – ocenu najmanjih kvadrata.

Preostalo je još da odredimo ocenu najmanjih kvadrata za disperziju ostataka regresionog modela, kada je ova nepoznata.

**Teorema 6.1.4** *Nepristrasna ocena disperzije $\sigma^2$ dobijena po metodu najmanjih kvadrata, kada je matrica $\mathbf{A}$ nesingularna, je statistika*

$$\widetilde{\sigma}^2 = \frac{1}{n-k} S(\widehat{\boldsymbol{\beta}}) = \frac{1}{n-k} (\mathbf{Y} - \mathbf{Z}'\widehat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{Z}'\widehat{\boldsymbol{\beta}}) \,,$$

*gde je, kao i do sada, $\widehat{\boldsymbol{\beta}}$ ocena najmanjih kvadrata vektora $\boldsymbol{\beta}$, odnosno proizvoljno rešenje normalne jednačine.*

**Dokaz.** Označimo sa $a_{ij}$ elemente matrice $\mathbf{A}$, a sa $a^{ij}$ elemente matrice $\mathbf{A}^{-1}$.

Kako je

$$S(\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{Z}'\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{Z}'\boldsymbol{\beta}) = (\mathbf{Y} - E(\mathbf{Y}))'(\mathbf{Y} - E(\mathbf{Y})) \,,$$

to je

$$E(S(\boldsymbol{\beta})) = \sum_{i=1}^{n} \left( E(Y_i^2) - (E(Y_i))^2 \right) = n\sigma^2 \,.$$

Osim toga je

$$\begin{aligned} E\left( (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{A}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \right) &= \sum_{j=1}^{k}\sum_{i=1}^{k} a_{ji} E\left( (\widehat{\beta}_i - \beta_i)(\widehat{\beta}_j - \beta_j) \right) = \\ &= \sigma^2 \sum_{j=1}^{k}\sum_{i=1}^{k} a_{ji} a^{ij} = \sigma^2 tr(\mathbf{A}\mathbf{A}^{-1}) = \\ &= \sigma^2 tr(\mathbf{I}_k) = k\sigma^2 \,. \end{aligned}$$

Koristeći sada rezultat

$$S(\boldsymbol{\beta}) = S(\widehat{\boldsymbol{\beta}}) + (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{A}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \,,$$

dobija se da je

$$E(S(\widehat{\boldsymbol{\beta}})) = E(S(\boldsymbol{\beta})) - E\left( (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{A}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \right) = n\sigma^2 - k\sigma^2 = (n-k)\sigma^2 \,,$$

odakle sledi tvrdjenje teoreme.□

**Primer 87.** Ilustrujmo prethodnu teoriju na primeru regresije sa jednim kontrolisanim faktorom, tj. na primeru jednostruke regresije. Dakle, neka je u pitanju kontrolisani faktor $x$ i njegov uticaj na slučajni ishod eksperimenta $Y$. Model linearne regresije podrazumeva odredjivanje najbolje linearne zavisnosti oblika

$$y = \beta_1 + \beta_2 x$$

na osnovu matrice plana reda $2 \times n$, čiji su vektori kolona $\mathbf{z}^{(i)} = (1, x^{(i)})'$, $i = 1, \ldots, n$. Uočimo da je u ovom slučaju $z(x) = x$, tj. da će rezultat eksperimenta biti registrovan u obliku $(x^{(i)}, y_i)$, $i = 1, \ldots, n$. Prema tome,

$$EY_i = \beta_1 + \beta_2 x^{(i)} \quad , \quad i = 1 \ldots, n\,.$$

Uobičajeno je da se koeficijent $\beta_1$ zove odsečak, a $\beta_2$ koeficijent pravca, zbog geometrijske interpretacije modela i naziva za ove koeficijente koji se koriste u analitičkoj geometriji.

Prateći dalje teoriju, treba definisati sve potrebne matrice i vektore:

$$\mathbf{Z} = \left\| \begin{array}{cccc} 1 & 1 & \cdots & 1 \\ x^{(1)} & x^{(2)} & \cdots & x^{(n)} \end{array} \right\| \qquad \mathbf{A} = \left\| \begin{array}{cc} n & \sum_{i=1}^n x^{(i)} \\ \sum_{i=1}^n x^{(i)} & \sum_{i=1}^n x^{(i)\,2} \end{array} \right\| \qquad \mathbf{V} = \left\| \begin{array}{c} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n x^{(i)} Y_i \end{array} \right\| .$$

Pretpostavimo da je eksperiment sproveden bar na dva različita nivoa kontrolisanog faktora $x$. To bi za posledicu imalo da je rang matrice plana 2, tj. $rang\mathbf{Z} = 2$, a samim tim bi značilo da je matrica $\mathbf{A}$ regularna. Otuda

$$det\mathbf{A} = n\sum_{i=1}^{n}(x^{(i)} - \overline{x}_n)^2 \; > \; 0 \quad , \quad \overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} x^{(i)}\,,$$

a njena inverzna matrica je

$$\mathbf{A}^{-1} = \frac{1}{det\mathbf{A}} \left\| \begin{array}{cc} \sum_{i=1}^n x^{(i)\,2} & -n\overline{x}_n \\ -n\overline{x}_n & n \end{array} \right\|$$

što znači da postoji jedinstveno rešenje sistema normalnih jednačina i ono je dato vektorom statistika

$$\widehat{\boldsymbol{\beta}} = \left\| \begin{array}{c} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{array} \right\| = \left\| \begin{array}{c} \dfrac{\sum_{i=1}^n x^{(i)\,2}\overline{Y}_n - \overline{x}_n \sum_{i=1}^n x^{(i)} Y_i}{\sum_{i=1}^n (x^{(i)} - \overline{x}_n)^2} \\ \dfrac{\sum_{i=1}^n (x^{(i)} - \overline{x}_n)(Y_i - \overline{Y}_n)}{\sum_{i=1}^n (x^{(i)} - \overline{x}_n)^2} \end{array} \right\| \qquad , \qquad \overline{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i\,.$$

Uočimo da su statistike $\widehat{\beta}_1$ i $\widehat{\beta}_2$ u sledećoj relaciji

$$\widehat{\beta}_1 = \overline{Y}_n - \overline{x}_n \widehat{\beta}_2\,.$$

Disperziona matrica vektora $\widehat{\boldsymbol{\beta}}$ je

$$\mathbf{D}(\widehat{\boldsymbol{\beta}}) = \sigma^2 \mathbf{A}^{-1}\,,$$

što znači da su disperzije komponenata

$$D(\widehat{\beta}_1) = \frac{\sum_{i=1}^n {x^{(i)}}^2}{n\sum_{i=1}^n (x^{(i)} - \overline{x}_n)^2}\sigma^2 \qquad \text{i} \qquad D(\widehat{\beta}_2) = \frac{1}{\sum_{i=1}^n (x^{(i)} - \overline{x}_n)^2}\sigma^2\,,$$

a njihova kovarijansa

$$Cov(\widehat{\beta}_1, \widehat{\beta}_2) = \frac{\overline{x}_n}{\sum_{i=1}^n (x^{(i)} - \overline{x}_n)^2}\sigma^2\,.$$

Ukoliko je disperzija grešaka, $\sigma^2$, nepoznata, treba i nju oceniti. Njena nepristrasna ocena data je relacijom

$$\widetilde{\sigma}^2 = \frac{S(\widehat{\boldsymbol{\beta}})}{n-2} \quad , \quad S(\widehat{\boldsymbol{\beta}}) = \sum_{i=1}^n (Y_i - \overline{Y}_n)^2 - \widehat{\beta}_2^2 \sum_{i=1}^n (x^{(i)} - \overline{x}_n)^2\,. \triangle$$

**Primer 88.** Koristeći rezultat prethodnog primera naći regresionu pravu kojom se može predvideti količina soli ($Y$) natrijumnitrata $NaNO_3$ koju je moguće rastvoriti u $100gr$ vode u zavisnosti od temperature ($x$) na osnovu sledećih eksperimentalnih podataka:

| $x^{(i)}$ | 0 | 4 | 10 | 15 | 21 | 29 | 36 | 51 | 68 |
|---|---|---|---|---|---|---|---|---|---|
| $y_i$ | 66,7 | 71,0 | 76,3 | 80,6 | 85,7 | 92,9 | 99,4 | 113,6 | 125,1 |

i oceniti disperziju grešaka merenja.

Rešenje se dobija rešavanjem sistema jednačina

$$\begin{aligned} 9\beta_1 \;+\; 234\beta_2 \;&=\; 811,3 \\ 234\beta_1 + 10144\beta_2 &= 24628,6 \end{aligned}$$

i ocene koeficijenata regresije su

$$\beta_1 = 67,5 \quad \text{i} \quad \beta_2 = 0,87\,.$$

Dakle, jednačina linearne regresije je

$$y = 67,5 + 0,87x\,,$$

a ocena disperzije slučajnih grešaka merenja je $\widetilde{\sigma}^2 = 0,92$. $\triangle$

**Primer 89.** Razmotrimo sada najjednostavniji primer linearne paraboličke regresije. Posmatraćemo ponovo uticaj samo jednog kontrolisanog faktora $x$ na slučajni ishod eksperimenta $Y$ i pretpostaviti da se radi o modelu kod koga je

$$EY = \beta_0 + \beta_1 x + \beta_2 x^2\,,$$

gde je sa $x^2$ označen kvadrat jedinog faktora $x$. Indekse za koeficijente $\beta$ smo, iz tradicionalnih razloga vezanih za označavanje koeficijenata polinoma, označili počev od 0.

Sistem normalnih jednačina je, u ovom slučaju,

$$\begin{aligned} \beta_0 n \;\;\;+\; \beta_1 \textstyle\sum x^{(i)} \;+ \beta_2 \sum {x^{(i)}}^2 &= \;\; \sum y_i \\ \beta_0 \textstyle\sum x^{(i)} \;+ \beta_1 \sum {x^{(i)}}^2 + \beta_2 \sum {x^{(i)}}^3 &= \sum x^{(i)} y_i \\ \beta_0 \textstyle\sum {x^{(i)}}^2 + \beta_1 \sum {x^{(i)}}^3 + \beta_2 \sum {x^{(i)}}^4 &= \sum {x^{(i)}}^2 y_i \end{aligned} \quad . \triangle$$

**Primer 90.** Navodimo još nekoliko modela regresije druge vrste koji ne spadaju u linearne, medjutim jednostavnim transformacijama se mogu svesti na model linearne regresije, te i optimizovati metodom najmanjih kvadrata.

Na linearni model mogu se svesti, na primer, sledeći modeli jednostruke regresije

- $Y = \beta_1 x^{\beta_2} + \varepsilon$
- $Y = \frac{1}{\beta_1 + \beta_2 x} + \varepsilon$
- $Y = e^{\beta_1 + \beta_2 x} + \varepsilon$ .

Označimo $EY = \overline{y}$, pa s obzirom na pretpostavku o očekivanju obeležja $Y$, posmatrajmo transformacije redom

- $\ln \overline{y} = v$, $\ln \beta_1 = b_1$, $\ln x = u$
- $\frac{1}{\overline{y}} = v$
- $\ln \overline{y} = v$

kojim dobijamo linearne modele

- $v = b_1 + \beta_2 u$
- $v = \beta_1 + \beta_2 x$
- $v = \beta_1 + \beta_2 x$ . $\triangle$

**Primer 91.** Najjednostavniji model višestruke regresije je model sa dva različita kontrolisana faktora čija je matrica plana

$$\mathbf{Z} = \left\| \begin{array}{cccc} 1 & 1 & \cdots & 1 \\ x_2^{(1)} & x_2^{(2)} & \cdots & x_2^{(n)} \\ x_3^{(1)} & x_3^{(2)} & \cdots & x_3^{(n)} \end{array} \right\| .$$

Odgovarajući sistem normalnih jednačina za ovaj model je

$$\begin{array}{ccccccc} \beta_1 n & + & \beta_2 \sum x_2^{(i)} & + & \beta_3 \sum x_3^{(i)} & = & \sum y_i \\ \beta_1 \sum x_2^{(i)} & + & \beta_2 \sum {x_2^{(i)}}^2 & + \beta_3 \sum x_2^{(i)} x_3^{(i)} & & = & \sum x_2^{(i)} y_i \\ \beta_1 \sum x_3^{(i)} & + \beta_2 \sum x_2^{(i)} x_3^{(i)} & + & \beta_3 \sum {x_3^{(i)}}^2 & & = & \sum x_3^{(i)} y_i \end{array} \quad . \triangle$$

Zadržimo se kratko na vektoru $\mathbf{U}$ definisanom sa

$$\mathbf{U} = \mathbf{Y} - \mathbf{Z}' \widehat{\boldsymbol{\beta}} \, ,$$

poznat pod nazivom **vektor ostataka**, a njegove komponente se nazivaju ostaci. Označimo sa

$$\mathbf{B} = \mathbf{I}_n - \mathbf{Z}' \mathbf{A}^{-1} \mathbf{Z} \, .$$

Tada je

$$\mathbf{U} = \mathbf{BY}\,.$$

Lako je konstatovati da je matrica $\mathbf{B}$ simetrična i idempotentna, kao i da je

$$\widetilde{\sigma}^2 = \frac{1}{n-k}\mathbf{Y}'\mathbf{BY} \quad , \quad E(\mathbf{U}) = \mathbf{0} \quad \text{i} \quad \mathbf{D}(\mathbf{U}) = \sigma^2\mathbf{B}\,.$$

Uloga vektora ostataka biće razjašnjena u poglavlju o analizi rasipanja.

Nadalje ćemo razmatrati specijalan slučaj modela linearne regresije druge vrste.

### 6.1.2 Model normalne regresije

Sa uvodjenjem dodatnih pretpostavki o raspodeli vektora grešaka $\boldsymbol{\varepsilon}$ moguće je dobiti još neke ocene u vezi sa koeficijentima linearne regresije. Najčešća pretpostavka je pretpostavka o normalnosti vektora $\boldsymbol{\varepsilon}$ tj. $\boldsymbol{\varepsilon} : \mathcal{N}(\mathbf{0}\,, \sigma^2\mathbf{I}_n)$. U tom slučaju se govori o normalnoj regresiji. Direktna posledica te pretpostavke i definicije modela linearne regresije je da vektor $\mathbf{Y}$ ima takodje normalnu raspodelu sa vektorom očekivanja $\mathbf{Z}'\boldsymbol{\beta}$ i kovarijansnom matricom $\sigma^2\mathbf{I}_n$ tj.

$$\mathbf{Y} : \mathcal{N}(\mathbf{Z}'\boldsymbol{\beta}\,,\, \sigma^2\mathbf{I}_n)\,. \tag{6.10}$$

Model normalne regresije će ovde biti razmatran isključivo na prostom uzorku.

### 6.1.3 Ocena maksimalne verodostojnosti parametara modela normalne regresije

Normalni model (6.10) sadrži $(k+1)$ parametar, odnosno, definisan je parametrom $\Theta$ dimenzije $(k+1)$:

$$\Theta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \\ \sigma^2 \end{pmatrix},$$

čije moguće vrednosti pripadaju Euklidovom poluprostoru $\boldsymbol{\Theta} \subset E_{k+1}$,

$$\boldsymbol{\Theta} = \{\Theta : -\infty < \beta_j < +\infty, j = 1, \ldots, k, \sigma^2 > 0\}.$$

Ako je $\boldsymbol{y} = (y_1, \ldots, y_n)'$ realizacija vektora $\mathbf{Y}$ tada će funkcija verodostojnosti biti:

$$L(\Theta; \boldsymbol{y}) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y} - \mathbf{Z}'\boldsymbol{\beta})'(\boldsymbol{y} - \mathbf{Z}'\boldsymbol{\beta})\right)\,.$$

Vidimo da se maksimum funkcije verodostojnosti po $\boldsymbol{\beta}$ pri svakom fiksiranom $\sigma^2$ postiže kada je kvadratna forma $(\boldsymbol{y} - \mathbf{Z}'\boldsymbol{\beta})'(\boldsymbol{y} - \mathbf{Z}'\boldsymbol{\beta})$ minimalna, tj. kada za $\boldsymbol{\beta}$ uzmemo ocenu najmanjih kvadrata dobijenu kod opšteg modela linearne regresije. Otuda se ocena dobijena

metodom maksimalne verodostojnosti i ocena dobijena metodom najmanjih kvadrata za $\boldsymbol{\beta}$ poklapaju kod modela normalne regresije.

Kod opšteg modela linearne regresije ocena najmanjih kvadrata je bila najbolja medju svim linearnim nepristrasnim ocenama vektora $\boldsymbol{\beta}$. Kod normalne linearne regresije ova ocena je (na osnovu ranije navedenog o oceni maksimalne verodostojnosti) najbolja medju svim nepristrasnim ocenama za $\boldsymbol{\beta}$.

Do ocene maksimalne verodostojnosti za nepoznatu disperziju $\sigma^2$ dolazi se na uobičajeni način:

$$\ln L(\Theta;\boldsymbol{y}) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}(\boldsymbol{y}-\mathbf{Z}'\boldsymbol{\beta})'(\boldsymbol{y}-\mathbf{Z}'\boldsymbol{\beta})\,,$$

$$\frac{\partial \ln L}{\partial \sigma^2} = 0\,,$$

$$-\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2}(\boldsymbol{y}-\mathbf{Z}'\boldsymbol{\beta})'(\boldsymbol{y}-\mathbf{Z}'\boldsymbol{\beta})\frac{1}{(\sigma^2)^2} = 0\,.$$

Rešavanjem poslednje jednačine po $\sigma^2$ dobija se

$$\sigma^2 = \frac{1}{n}(\boldsymbol{y}-\mathbf{Z}'\boldsymbol{\beta})'(\boldsymbol{y}-\mathbf{Z}'\boldsymbol{\beta})\,.$$

Otuda, ako umesto $\boldsymbol{\beta}$ koristimo ocenu $\widehat{\boldsymbol{\beta}}$ dobijamo statistiku

$$\widehat{\sigma}^2 = \frac{1}{n}S(\widehat{\boldsymbol{\beta}}) = \frac{1}{n}(\mathbf{Y}-\mathbf{Z}'\widehat{\boldsymbol{\beta}})'(\mathbf{Y}-\mathbf{Z}'\widehat{\boldsymbol{\beta}})$$

kao ocenu maksimalne verodostojnosti za $\sigma^2$.

Kao što znamo, nepristrasna ocena za $\sigma^2$ ima oblik:

$$\widetilde{\sigma}^2 = \frac{S(\widehat{\boldsymbol{\beta}})}{n-k}\,,$$

pa je ocena $\widehat{\sigma}^2$ pristrasna i njena pristrasnost iznosi:

$$\begin{aligned} E\widehat{\sigma}^2 - \sigma^2 &= E\left(\frac{S(\widehat{\boldsymbol{\beta}})}{n}\right) - \sigma^2 \\ &= \frac{1}{n}(n-k)\sigma^2 - \sigma^2 \\ &= -\frac{k}{n}\sigma^2\,. \end{aligned}$$

Dakle, pristrasnost opada sa porastom broja posmatranja n.

### 6.1.4 Osnovna teorema teorije normalne regresije

**Teorema 6.1.5** *Slučajne veličine $\widehat{\boldsymbol{\beta}}$ i $S(\widehat{\boldsymbol{\beta}})$ su nezavisne, a takve su i $S(\widehat{\boldsymbol{\beta}})$ i $Q = S(\boldsymbol{\beta}) - S(\widehat{\boldsymbol{\beta}})$. Pri tome su njihove raspodele redom:*

$$\widehat{\boldsymbol{\beta}} : \mathcal{N}(\boldsymbol{\beta}\,, \sigma^2\mathbf{A}^{-1})$$

$$\frac{S(\widehat{\boldsymbol{\beta}})}{\sigma^2} : \chi^2_{(n-k)}$$
$$\frac{Q}{\sigma^2} : \chi^2_k \,.$$

**Dokaz.**Kompletan dokaz teoreme može se naći u [14], a mi ćemo ovde dokazati samo nezavisnost navedenih slučajnih veličina.

Uvedimo normirani vektor grešaka

$$\boldsymbol{\varepsilon}^* = \left(\frac{\varepsilon_1}{\sigma}, \ldots, \frac{\varepsilon_n}{\sigma}\right)'$$

čija je raspodela data sa:

$$\boldsymbol{\varepsilon}^* : \mathcal{N}(\mathbf{0}\,, \mathbf{I}_n).$$

Tada vektor $\mathbf{Y}$ dobija oblik

$$\mathbf{Y} = \mathbf{Z}'\boldsymbol{\beta} + \sigma\boldsymbol{\varepsilon}^*.$$

Odavde se dobija da je:

$$\widehat{\boldsymbol{\beta}} = \mathbf{A}^{-1}\mathbf{Z}\mathbf{Y} = \mathbf{A}^{-1}\mathbf{Z}(\mathbf{Z}'\boldsymbol{\beta} + \sigma\boldsymbol{\varepsilon}^*) = \mathbf{A}^{-1}\mathbf{Z}\mathbf{Z}'\boldsymbol{\beta} + \mathbf{A}^{-1}\mathbf{Z}\sigma\boldsymbol{\varepsilon}^* = \boldsymbol{\beta} + \sigma\mathbf{A}^{-1}\mathbf{Z}\boldsymbol{\varepsilon}^*.$$

Kako je

$$\widetilde{\sigma}^2 = \frac{1}{n-k}\mathbf{Y}'\mathbf{B}\mathbf{Y} \quad ,$$

gde je, kao i do sada,

$$\mathbf{B} = \mathbf{I}_n - \mathbf{Z}'\mathbf{A}^{-1}\mathbf{Z}\,.$$

Sledi da je

$$\frac{S(\widehat{\boldsymbol{\beta}})}{n-k} = \widetilde{\sigma}^2 \;\; = \;\; \frac{1}{n-k}(\mathbf{Z}'\boldsymbol{\beta} + \sigma\boldsymbol{\varepsilon}^*)'\mathbf{B}(\mathbf{Z}'\boldsymbol{\beta} + \sigma\boldsymbol{\varepsilon}^*)\,.$$

Odavde je očigledno

$$S(\widehat{\boldsymbol{\beta}}) = \sigma^2\boldsymbol{\varepsilon}^{*\prime}\mathbf{B}\boldsymbol{\varepsilon}^*\,,$$

odnosno

$$\frac{S(\widehat{\boldsymbol{\beta}})}{\sigma^2} = \boldsymbol{\varepsilon}^{*\prime}\mathbf{B}\boldsymbol{\varepsilon}^*\,.$$

Dakle, treba utvrditi nezavisnost sledećih slučajnih promenljivih

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \sigma\mathbf{A}^{-1}\mathbf{Z}\boldsymbol{\varepsilon}^*,$$

i

$$S(\widehat{\boldsymbol{\beta}}) = \sigma^2\boldsymbol{\varepsilon}^{*\prime}\mathbf{B}\boldsymbol{\varepsilon}^*.$$

Njihova nezavisnost sledi iz činjenice da je $\mathbf{A}^{-1}\mathbf{Z}\mathbf{B} = \mathbf{0}$, odnosno činjenice da je proizvod matrica kvadratne i linearne forme vektora slučajnih grešaka $\boldsymbol{\varepsilon}^*$ jednak nula matrici. To ima za posledicu upravo nezavisnost kvadratne i linearne forme ([14], lema 1.2).

Kako slučajna veličina $Q$ zavisi od uzorka $\mathbf{Y}$ samo preko statistike $\widehat{\boldsymbol{\beta}}$, a slučajne promenljive $\widehat{\boldsymbol{\beta}}$ i $S(\widehat{\boldsymbol{\beta}})$ su nezavisne, to su i $Q$ i $S(\widehat{\boldsymbol{\beta}})$ takodje nezavisne. □

Kao direktne posledice prethodne teoreme mogu se navesti sledeća tvrdjenja:

Za svako $j = 1, \ldots, k$ je:

- $$\frac{\widehat{\beta}_j - \beta_j}{\sigma\sqrt{a^{jj}}} : \mathcal{N}(0,1) \tag{6.11}$$

  pri čemu je slučajna promenljiva $\widehat{\beta}_j$ nezavisna od slučajne promenljive $S(\widehat{\boldsymbol{\beta}})$, a $a^{jj}$ je $j$–ti element glavne dijagonale matrice $\mathbf{A}^{-1}$.

- $$\frac{\frac{\widehat{\beta}_j - \beta_j}{\sigma\sqrt{a^{jj}}}}{\sqrt{\frac{S(\widehat{\boldsymbol{\beta}})}{\sigma^2(n-k)}}} : t_{n-k}\,, \tag{6.12}$$

  dakle, u pitanju je statistika sa Studentovom raspodelom sa $n-k$ stepeni slobode.

- $$\frac{\frac{Q}{k\sigma^2}}{\frac{S(\widehat{\boldsymbol{\beta}})}{(n-k)\sigma^2}} = \frac{n-k}{k}\cdot\frac{Q}{S(\widehat{\boldsymbol{\beta}})} : F_{k,n-k}\,, \tag{6.13}$$

  tj. ova statistika ima Fišerovu raspodelu sa $k$ i $(n-k)$ stepeni slobode.

### 6.1.5 Skupovi poverenja za parametre normalne regresije

Odredićemo najpre interval poverenja za pojedini koeficijent $\beta_j$ linearne regresije.

Na osnovu posledice (6.11), slučajna promenljiva $\widehat{\beta}_j$ ima raspodelu $\mathcal{N}(\beta_j, \sigma^2 a^{jj})$, $j = 1,\ldots,k$. Dakle, zadatak se svodi na ocenjivanje nepoznatog matematičkog očekivanja normalne raspodele kada je disperzija nepoznata. U tu svrhu možemo koristiti centralnu statistiku:

$$\frac{\frac{\widehat{\beta}_j - \beta_j}{\sigma\sqrt{a^{jj}}}}{\sqrt{\frac{S(\widehat{\boldsymbol{\beta}})}{\sigma^2(n-k)}}} : t_{n-k}\,.$$

Uvedimo oznaku

$$\sqrt{\frac{n-k}{a^{jj}}}\cdot\frac{1}{\sqrt{S(\widehat{\boldsymbol{\beta}})}}(\widehat{\beta}_j - \beta_j) = t\,.$$

Odavde se dobija da je:

$$(\widehat{\beta}_j - \beta_j) = t\cdot\sqrt{\frac{S(\widehat{\boldsymbol{\beta}})a^{jj}}{n-k}}\,.$$

Interval poverenja sa nivoom poverenja $1-\alpha$, $0<\alpha<1$, sledi iz $P\left\{|t|\leq t_{n-k,\frac{1-\alpha}{2}}\right\}=1-\alpha$. Zaista, iz

$$\left|(\widehat{\beta}_j-\beta_j)\sqrt{\frac{n-k}{S(\widehat{\boldsymbol{\beta}})a^{jj}}}\right|\leq t_{n-k,\frac{1-\alpha}{2}}$$

sledi

$$I_{\beta_j}=\left[\widehat{\beta}_j-t_{n-k,\frac{1-\alpha}{2}}\cdot\sqrt{\frac{S(\widehat{\boldsymbol{\beta}})a^{jj}}{n-k}}\quad,\quad\widehat{\beta}_j+t_{n-k,\frac{1-\alpha}{2}}\cdot\sqrt{\frac{S(\widehat{\boldsymbol{\beta}})a^{jj}}{n-k}}\right].$$

Interval poverenja za nepoznatu disperziju $\sigma^2$ dobija se na osnovu centralne statistike

$$\frac{S(\widehat{\boldsymbol{\beta}})}{\sigma^2}$$

koja ima $\chi^2_{(n-k)}$ raspodelu. Dakle, dvostrani interval poverenja će biti oblika:

$$I_{\sigma^2}=\left[\frac{S(\widehat{\boldsymbol{\beta}})}{\chi^2_{n-k,1-\frac{\alpha}{2}}}\quad,\quad\frac{S(\widehat{\boldsymbol{\beta}})}{\chi^2_{n-k,\frac{\alpha}{2}}}\right].$$

Na taj način je moguće postaviti intervale poverenja za svaki od koeficijenata regresije $\beta_1,\ldots,\beta_k$. Ako nadjemo $k$ takvih intervala sa jednim istim nivoom poverenja $\gamma=1-\alpha$, tada će očekivana vrednost broja intervala koji prekrivaju vrednost $\beta_j$ biti $k\gamma$. Ocenimo sa kojom verovatnoćom svi intervali poverenja istovremeno pokrivaju svoj odgovarajući parametar $\beta_j$.

Označimo sa $A_j$ dogadjaj da slučajni interval

$$\left[\widehat{\beta}_j-t_{n-k,\frac{1-\alpha_j}{2}}\cdot\sqrt{\frac{S(\widehat{\boldsymbol{\beta}})a^{jj}}{n-k}}\quad,\quad\widehat{\beta}_j+t_{n-k,\frac{1-\alpha_j}{2}}\cdot\sqrt{\frac{S(\widehat{\boldsymbol{\beta}})a^{jj}}{n-k}}\right]$$

obuhvata pravu vrednost parametra $\beta_j$. Sledi da je

$$P(A_j)=1-\alpha_j.$$

Dogadjaj čija nas verovatnoća zanima je $A_1A_2\ldots A_k$ pa je:

$$P(A_1A_2\ldots A_k)=1-P(\cup_{j=1}^k A_j^c)$$

$$P(\cup_{j=1}^k A_j^c)\leq\sum_{j=1}^k P(A_j^c)=\sum_{j=1}^k\alpha_j$$

$$P(A_1A_2\ldots A_k)\geq 1-\sum_{j=1}^k\alpha_j.$$

Ako izaberemo $\alpha_1 = \alpha_2 = \ldots = \alpha_k = \frac{\alpha}{k}$, za neko $\alpha$, $0 < \alpha < 1$, dobijamo da je:

$$P(A_1 A_2 \ldots A_k) \geq 1 - k\frac{\alpha}{k} = 1 - \alpha.$$

Medjutim, moguće je postaviti i ovakav zadatak:

*U euklidskom prostoru $E_k$ naći oblast poverenja $G_\gamma$ sa nivoom poverenja $\gamma$ koja sa verovatnoćom $\gamma$ prekriva nepoznatu parametarsku tačku $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_k)$.*

U tu svrhu koristi se centralna statistika

$$\frac{n-k}{k} \cdot \frac{Q}{S(\widehat{\boldsymbol{\beta}})}$$

koja ima Fišerovu raspodelu $F_{k,n-k}$ na sledeći način:

$$\gamma = 1 - \alpha = P\{F \leq F_{k,n-k;1-\alpha}\} = P\{\boldsymbol{\beta} \in G_\gamma(\mathbf{Y})\},$$

gde je

$$\begin{aligned} G_\gamma(\mathbf{Y}) &= \left\{ \boldsymbol{\beta} \;\middle|\; \frac{n-k}{k} \cdot \frac{Q}{S(\widehat{\boldsymbol{\beta}})} \leq F_{k,n-k;1-\alpha} \right\} = \\ &= \left\{ \boldsymbol{\beta} \;\middle|\; \frac{S(\boldsymbol{\beta}) - S(\widehat{\boldsymbol{\beta}})}{S(\widehat{\boldsymbol{\beta}})} \leq \frac{k}{n-k} F_{k,n-k;1-\alpha} \right\} = \\ &= \left\{ \boldsymbol{\beta} \;\middle|\; (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \mathbf{A} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \leq \frac{k}{n-k} S(\widehat{\boldsymbol{\beta}}) F_{k,n-k;1-\alpha} \right\} \end{aligned}$$

i predstavlja oblast elipsoida sa centrom u $\boldsymbol{\beta}$ i granicom datom jednačinom

$$(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \mathbf{A} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \frac{k}{n-k} S(\widehat{\boldsymbol{\beta}}) F_{k,n-k;1-\alpha}\,.$$

**Primer 92.** Razmotrićemo intervale poverenja i oblast poverenja za model jednostruke linearne regresije

$$EY_i = \beta_1 + \beta_2 x^{(i)} \quad , \quad i = 1, \ldots, n\,.$$

Interval poverenja za $\beta_2$ sa nivoom poverenja $\gamma = 1 - \alpha$ ukoliko je disperzija nepoznata, biće:

$$I_{\beta_2} = \left[ \widehat{\beta}_2 - t_{n-2,\frac{1-\alpha}{2}} \sqrt{\frac{S(\widehat{\boldsymbol{\beta}}) a^{22}}{n-2}} \quad , \quad \widehat{\beta}_2 + t_{n-2,\frac{1-\alpha}{2}} \sqrt{\frac{S(\widehat{\boldsymbol{\beta}}) a^{22}}{n-2}} \right]$$

pri čemu je, u ovom specijalnom slučaju,

$$a^{22} = \frac{1}{\sum_{i=1}^n (x^{(i)} - \overline{x}_n)^2} \quad ,$$

pa je

$$I_{\beta_2} = \left[ \widehat{\beta}_2 - t_{n-2,1-\frac{\alpha}{2}} \sqrt{\frac{S(\widehat{\boldsymbol{\beta}})}{(n-2)\sum_{i=1}^n (x^{(i)} - \overline{x}_n)^2}} \quad , \widehat{\beta}_2 + t_{n-2,1-\frac{\alpha}{2}} \sqrt{\frac{S(\widehat{\boldsymbol{\beta}})}{(n-2)\sum_{i=1}^n (x^{(i)} - \overline{x}_n)^2}} \right].$$

Ako tražimo oblast poverenja $G_\gamma(\mathbf{Y})$ koja natkriva tačku $(\beta_1, \beta_2)$ u euklidskom prostoru $E_2$ sa verovatnoćom $\gamma$ , dobijamo oblast ograničenu elipsom:

$$G_\gamma(\mathbf{Y}) =$$
$$= \left\{ \boldsymbol{\beta} \left| \|\widehat{\beta}_1 - \beta_1, \widehat{\beta}_2 - \beta_2\| \left\| \begin{array}{cc} n & \sum_{i=1}^n x^{(i)} \\ \sum_{i=1}^n x^{(i)} & \sum_{i=1}^n x^{(i)^2} \end{array} \right\| \left\| \begin{array}{c} \widehat{\beta}_1 - \beta_1 \\ \widehat{\beta}_2 - \beta_2 \end{array} \right\| \leq \frac{2}{n-2} S(\widehat{\boldsymbol{\beta}}) F_{2,n-2;\gamma} \right. \right\} =$$
$$= \left\{ \boldsymbol{\beta} \left| (\widehat{\beta}_1 - \beta_1)^2 n + 2(\widehat{\beta}_1 - \beta_1)(\widehat{\beta}_2 - \beta_2) \sum x^{(i)} + (\widehat{\beta}_2 - \beta_2)^2 \sum x^{(i)^2} \leq \frac{2}{n-2} S(\widehat{\boldsymbol{\beta}}) F_{2,n-2;\gamma} \right. \right\} =$$
$$= \left\{ \boldsymbol{\beta} \left| (\widehat{\beta}_1 - \beta_1)^2 + 2\overline{x}_n (\widehat{\beta}_1 - \beta_1)(\widehat{\beta}_2 - \beta_2) + \frac{1}{n} (\widehat{\beta}_2 - \beta_2)^2 \sum x^{(i)^2} \leq \frac{2}{n(n-2)} S(\widehat{\boldsymbol{\beta}}) F_{2,n-2;\gamma} \right. \right\} . \triangle$$

Moguće je tražiti oblast poverenja i za linearnu kombinaciju parametara regresije, tj. za vektor $\boldsymbol{t} = \mathbf{T}\boldsymbol{\beta}$, gde je matrica $\mathbf{T}$ dimenzije $m \times k$, a $rang\mathbf{T} = m$. U tu svrhu koristi se činjenica da $\widehat{\boldsymbol{t}} = \mathbf{T}\widehat{\boldsymbol{\beta}}$ ima raspodelu $\mathcal{N}(\boldsymbol{t}, \sigma^2\mathbf{D})$, gde je matrica $\mathbf{D}$ definisana sa $\mathbf{D} = \mathbf{T}\mathbf{A}^{-1}\mathbf{T}'$. Odavde sledi da $Q_\mathbf{T} = (\widehat{\boldsymbol{t}} - \boldsymbol{t})'\mathbf{D}(\widehat{\boldsymbol{t}} - \boldsymbol{t})$ ne zavisi od $S(\widehat{\boldsymbol{\beta}})$ i količnik $\frac{Q_\mathbf{T}}{\sigma^2}$ ima $\chi^2$ raspodelu, $\frac{Q_\mathbf{T}}{\sigma^2} : \chi^2_m$.

### 6.1.6 Testiranje hipoteza o ocenama parametara normalne regresije

Testiraćemo nultu hipotezu

$$H_0 : (\beta_1, \beta_2, \ldots, \beta_k) \in \mathcal{B}_0 \subset E_k$$

protiv odgovarajuće alternativne. Najčešće je oblast $\mathcal{B}_0$ linearni potprostor od $E_k$ oblika:

$$\mathcal{B}_0 = \{\boldsymbol{\beta} \,|\, \mathbf{T}\boldsymbol{\beta} = \boldsymbol{t}_0\} ,$$

gde je $\mathbf{T}$, kao i u prethodnom, matrica dimenzije $m \times k$, a $\boldsymbol{t}_0$ vektor dimenzije $m \times 1$.

Preciznije, najopštiji oblik hipoteze vezane za nepoznati višedimenzioni parametar $\Theta = (\boldsymbol{\beta}' \vdots \sigma^2)'$ linearne regresije koju treba testirati je

$$H_0 : \Theta \in \boldsymbol{\Theta} = \{\Theta \,|\, \boldsymbol{\beta} \in \mathcal{B}_0, \sigma^2 > 0\} .$$

Kritična oblast veličine $\alpha$ za testiranje ove složene nulte hipoteze protiv alternativne

$$H_1 : \Theta \in \boldsymbol{\Theta} = \{\Theta \,|\, \boldsymbol{\beta} \in \mathcal{B}_0^c, \sigma^2 > 0\} ,$$

koja je takodje složena, je

$$C = \left\{ \boldsymbol{y} \quad \left| \quad \frac{n-k}{m} \frac{(\widehat{\boldsymbol{t}} - \boldsymbol{t}_0)'\mathbf{D}^{-1}(\widehat{\boldsymbol{t}} - \boldsymbol{t}_0)}{S(\widehat{\boldsymbol{\beta}})} \geq F_{m,n-k;1-\alpha} \quad \right. \right\} ,$$

gde je

$$\mathbf{D} = \mathbf{T}\mathbf{A}^{-1}\mathbf{T}' \quad , \quad \widehat{\boldsymbol{t}} = \mathbf{T}\widehat{\boldsymbol{\beta}} \quad ,$$

dok je oblast prihvatanja $H_0$ njen komplement.

**Primer 93.** Razmotrićemo ponovo model jednostruke linearne regresije $Y = \beta_1 + \beta_2 x + \varepsilon$. Testiraćemo hipotezu

$$H_0 : \beta_2 = \beta_{20}\,,$$

što znači da se testira samo koeficijent pravca prave kojom je predstavljen regresioni model.

S obzirom da je u ovom primeru $m = 1$ (dimenzija matrice $\mathbf{T}$ je $1 \times 2$), granica kritične oblasti će biti kvantil reda $1 - \alpha$ slučajne promenljive sa $F_{1,n-2}$ raspodelom. Medjutim, poznato je da se Fišerova raspodela sa $(1, n-k)$ stepeni slobode poklapa sa raspodelom kvadrata slučajne promenljive koja ima Studentovu raspodelu sa $n-k$ stepeni slobode. Zbog toga i na osnovu (6.12), za alternativnu hipotezu $H_1 : \beta_2 \neq \beta_{20}$, dobili bismo kritičnu oblast iz uslova:

$$|\widehat{\beta}_2 - \beta_{20}| \geq t_{n-2,\frac{1-\alpha}{2}} \sqrt{\frac{S(\widehat{\boldsymbol{\beta}})}{(n-2)\sum_{i=1}^{n}(x^{(i)} - \overline{x}_n)^2}} \quad ,$$

gde je

$$S(\widehat{\boldsymbol{\beta}}) = \sum_{i=1}^{n}(Y_i - \beta_{20}x^{(i)} - \widehat{\beta}_1)^2 \quad , \quad \widehat{\beta}_1 = \overline{Y}_n - \beta_{20}\overline{x}_n \,. \triangle$$

Razmotrimo i opšti problem testiranja hipoteza vezanih za koeficijente normalne linearne regresije.

Razmatraćemo samo osnovni model. Preciznije, ako na obeležje $Y$ utiče $l$ neslučajnih faktora $x_1$, $x_2$, ..., $x_l$ gde je $l > 2$, razmatraćemo model (A):

$$(A) \qquad Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_g x_g + \beta_{g+1} x_{g+1} + \ldots + \beta_l x_l + \varepsilon.$$

Podsetimo se da razmatramo ovaj model pod pretpostavkom da je $\varepsilon$ slučajna promenljiva sa normalnom raspodelom. To bi za posledicu imalo da $Y$ ima normalnu raspodelu, kao i to da statistike $\widehat{\beta}_j$, $j = 0, 1, \ldots, l$ kojima se ocenjuju parametri regresije $\beta_j$, $j = 0, 1, \ldots, l$ imaju normalnu raspodelu. Pri fitovanju obeležja $Y$ na ovaj način, važno je utvrditi da li svi pobrojani faktori utiču na obeležje sa istim utvrdjenim pragom značajnosti ili se za neke od njih može utvrditi da nisu od značaja, čime bi se smanjio broj sabiraka modela (A). Hipoteza o tome da neki od faktora nisu od dovoljnog značaja da bi bitno uticali na obeležje $Y$ postavlja se na sledeći način:

$$H_0 : \beta_{g+1} = \beta_{g+2} = \ldots = \beta_l = 0$$

za neko $g < l$, gde se bez smanjenja opštosti uzima da su krajnjih $l - g$ faktora modela (A) manje značajnosti od postavljenog praga.

Ukoliko se testiranjem postavljene hipoteze prihvati hipoteza $H_0$, prelazi se na model (B), tzv. redukovani model u odnosu na model (A), koji se smatra potpunim modelom:

$$(B) \qquad Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_g x_g + \varepsilon.$$

Odgovarajuće sume kvadrata odstupanja za uzorak obima $n$ za model (A) i model (B) su redom:

$$Q_A = \sum_{i=1}^{n}\left(Y_i - (\beta_0 + \beta_1 x_1^{(i)} + \beta_2 x_2^{(i)} + \ldots + \beta_g x_g^{(i)} + \beta_{g+1} x_{g+1}^{(i)} + \ldots + \beta_l x_l^{(i)})\right)^2,$$

$$Q_B = \sum_{i=1}^{n} \left( Y_i - (\beta_0 + \beta_1 x_1^{(i)} + \beta_2 x_2^{(i)} + \ldots + \beta_g x_g^{(i)}) \right)^2 ,$$

gde smo, da se podsetimo, $i$–ti nivo pojedinog faktora uticaja označili gornjim indeksom u zagradi.

Potpuni model, model (A), će davati prognozu sa manjom greškom u odnosu na redukovani model, model (B), tj. $Q_A < Q_B$. Otuda se $Q_B$ može predstaviti na sledeći način:

$$Q_B = Q_A + (Q_B - Q_A),$$

gde su $Q_A$ i $Q_B - Q_A$ nezavisne slučajne veličine, za koje, pri uslovu normalne regresije, slučajne promenljive

$$\frac{Q_B}{\sigma^2}, \quad \frac{Q_A}{\sigma^2}, \quad \text{i} \quad \frac{Q_B - Q_A}{\sigma^2}$$

imaju $\chi^2$ raspodele sa odgovarajućim brojem stepeni slobode, a statistike koje se dobijaju kao njihovi količnici imaju Fišerove raspodele. Otuda se za testiranje postavljene nulte hipoteze koristi test statistika

$$F_{\nu_1,\nu_2} = \frac{(n-l-1)(Q_B - Q_A)}{(l-g)Q_A},$$

koja ima Fišerovu raspodelu sa $\nu_1 = l - g$ i $\nu_2 = n - l - 1$ stepeni slobode, na osnovu koje se odredjuje kritična oblast testa.

## 6.2 Regresija prve vrste

Neka su dati obeležje $Y$ i slučajni vektor $\mathbf{X} = (X_1, \ldots, X_p)$ , gde su $X_1, \ldots, X_p$, obeležja koja utiču na obeležje $Y$, tj. obeležja $Y$ i $\mathbf{X}$ su povezana nekom statističkom zavisnošću. Pretpostavimo da nam je poznata i zajednička raspodela $F_{\mathbf{X}Y}(x_1, \ldots, x_p, y)$. Vektor $\mathbf{X}$ je dostupan ispitivanju, tj. možemo da registrujemo njegove vrednosti tokom eksperimenta, dok to nije slučaj sa veličinom $Y$. Sve informacije o obeležju $Y$ dobijamo preko vektora $\mathbf{X}$, pa u tom smislu komponente vektora $\mathbf{X}$ nazivamo **prediktori** ili **prediktorske (prognostičke) promenljive**. Ideja o prognozi se matematički ostvaruje pretpostavkom da je moguće odrediti statistiku $\psi(\mathbf{X})$ kojom ćemo na zadovoljavajući način ocenjivati vrednosti obeležja $Y$. Takva statistika se zove **predikcija** ili **prognoza** za obeležje $Y$ na osnovu $\mathbf{X}$.

Razradom metoda nalaženja optimalne prognoze prema pojedinim kriterijumima bavi se teorija statističke regresije.

### 6.2.1 Najbolja prognoza za obeležje $Y$ na osnovu vektora X

Vratimo se zajedničkoj raspodeli $F_{\mathbf{X}Y}(\mathbf{x}, y)$, $\mathbf{x} = (x_1, \ldots, x_p) \in R^p$, $y \in R$ za koju za sada pretpostavimo da nam je poznata. U tom slučaju je često moguće odrediti uslovnu raspodelu $F_{Y|\mathbf{X}=\mathbf{x}}(y|\mathbf{x})$, kao i odgovarajuće gustine za slučaj raspodele apsolutno neprekidnog, odnosno diskretnog tipa.

Model regresije prve vrste se bazira na uslovnom matematičkom očekivanju. Uslovno matematičko očekivanje $E(Y|\mathbf{X}=\mathbf{x})$ se može posmatrati kao funkcija od $\mathbf{x}$. Označimo je sa

$$M(\mathbf{x}) = E(Y|\mathbf{X}=\mathbf{x})\,.$$

Neka je $\psi(\mathbf{X})$ proizvoljna prognoza za $Y$ na osnovu $\mathbf{X}$. Srednjekvadratnom greškom te prognoze zvaćemo veličinu

$$E\left(Y-\psi(\mathbf{X})\right)^2.$$

**Optimalna (najbolja) prognoza** za $Y$ u oznaci $\psi^*(\mathbf{X})$ će biti ona koja minimizira srednjekvadratno odstupanje, tj.

$$E\left(Y-\psi^*(\mathbf{X})\right)^2 = \inf_{\psi} E\left(Y-\psi(\mathbf{X})\right)^2.$$

Dokažimo da je $\psi^*(\mathbf{X}) = M(\mathbf{X})$. Zaista,

$$\begin{aligned} E\left(Y-\psi(\mathbf{X})\right)^2 &= E\left(Y-M(\mathbf{X})+M(\mathbf{X})-\psi(\mathbf{X})\right)^2 = \\ &= E\left(Y-M(\mathbf{X})\right)^2 + 2E[(Y-M(\mathbf{X}))(M(\mathbf{X})-\psi(\mathbf{X}))] + \\ &+ E\left(M(\mathbf{X})-\psi(\mathbf{X})\right)^2 \geq E\left(Y-M(\mathbf{X})\right)^2, \end{aligned} \tag{6.14}$$

jer je,

$$\begin{aligned} E[(Y-M(\mathbf{X}))(M(\mathbf{X})-\psi(\mathbf{X}))] &= \\ &= E\{E[(Y-M(\mathbf{X}))(M(\mathbf{X})-\psi(\mathbf{X}))|\mathbf{X}]\} = \\ &= E\{(M(\mathbf{X})-\psi(\mathbf{X}))E[(Y-M(\mathbf{X}))|\mathbf{X}]\} = \\ &= E\left\{(M(\mathbf{X})-\psi(\mathbf{X}))[E(Y|\mathbf{X})-M(\mathbf{X})]\right\} = 0\,. \end{aligned}$$

U (6.14) važi jednakost ako i samo ako je

$$E\left(M(\mathbf{X})-\psi(\mathbf{X})\right)^2 = 0\,,$$

što će biti tačno ako i samo ako je

$$M(\mathbf{X}) = \psi(\mathbf{X})$$

skoro sigurno, što je i trebalo dokazati.

Ako sa $\Delta$ označimo minimalnu grešku prognoze po kriterijumu srednjekvadratnog odstupanja, biće $\Delta = E(Y-M(\mathbf{X}))^2$. Tada je

$$\Delta = E\left\{E[(Y-M(\mathbf{X}))^2|\mathbf{X}]\right\}.$$

Definicija 43. Funkcija po $\mathbf{x}$ u oznaci $D(Y|\mathbf{X}=\mathbf{x})$ ili kraće $\sigma^2_{Y\mathbf{X}}$, definisana kao

$$D(Y|\mathbf{X}=\mathbf{x}) = \sigma^2_{Y\mathbf{X}} = E\left((Y-M(\mathbf{X}))^2|\mathbf{X}=\mathbf{x}\right)$$

zove se *uslovna disperzija* za $Y$, na osnovu vektora $\mathbf{X}=\mathbf{x}$.

Važno svojstvo prognoze $M(\mathbf{X})$ daje sledeća teorema.

**Teorema 6.2.1** *Veličina $M(\mathbf{X})$ ima maksimalnu koreliranost sa $Y$ medju svim prognozama za $Y$ na osnovu vektora $\mathbf{X}$.*

**Dokaz.** Treba naći koeficijent korelacije slučajnih veličina $Y$ i $M(\mathbf{X})$. Da bismo to našli, podjimo od proizvoljne prognoze za $Y$ na osnovu $\mathbf{X}$, $\psi(\mathbf{X})$, i odredimo najpre kovarijansu

$$\begin{aligned} Cov(\psi(\mathbf{X}),Y) &= E[(\psi(\mathbf{X})-E\psi(\mathbf{X}))(Y-EY)] = \\ &= E\{E[(\psi(\mathbf{X})-E\psi(\mathbf{X}))(Y-EY)|\mathbf{X}]\} = \\ &= E\{(\psi(\mathbf{X})-E\psi(\mathbf{X}))[E(Y|\mathbf{X})-E(E(Y|\mathbf{X}))]\} = \\ &= Cov(\psi(\mathbf{X}),M(\mathbf{X}))\,. \end{aligned} \tag{6.15}$$

Sada možemo da procenimo traženi koeficijent korelacije

$$\rho^2_{\psi Y}=\frac{Cov^2(\psi(\mathbf{X}),Y)}{\sigma^2_\psi\sigma^2_Y}=\frac{Cov^2(\psi(\mathbf{X}),M)}{\sigma^2_\psi\sigma^2_Y}\cdot\frac{\sigma^2_M}{\sigma^2_M}=\rho^2_{\psi M}\frac{\sigma^2_M}{\sigma^2_Y}\,,$$

gde su $\sigma^2_\psi$, $\sigma^2_Y$ i $\sigma^2_M$ disperzije redom slučajnih promenljivih $\psi(\mathbf{X})$, $Y$ i $M(\mathbf{X})$.

Ako je

$$\psi(\mathbf{X})=M(\mathbf{X})$$

skoro sigurno, tada je na osnovu (6.15)

$$Cov(M,Y)=Cov(M,M)=\sigma^2_M\,.$$

Odavde je

$$\rho_{MY}=\frac{Cov(M,Y)}{\sigma_M\sigma_Y}=\frac{\sigma_M}{\sigma_Y}$$
$$\rho^2_{\psi Y}=\rho^2_{\psi M}\rho^2_{MY}\le\rho^2_{MY}\,.$$

Zaključujemo da $M$ ima maksimalnu koreliranost sa $Y$ tj. važi:

$$|\rho_{\psi Y}|\le|\rho_{MY}|.\ \square$$

Kvadrat koeficijenta korelacije prognoze $M(\mathbf{X})$ i prognozirane slučajne veličine $Y$ ima posebno mesto u regresionoj analizi, te se iz tog razloga uvodi sledeća definicija.

Definicija 44. Veličina

$$\eta^2_{Y\mathbf{X}}=\frac{\sigma^2_M}{\sigma^2_Y}=\rho^2_{MY}$$

se naziva *korelacioni količnik* ili *koeficijent determinacije*.

On daje udeo modelom objašnjenih varijacija u ukupnim varijacijama za $Y$.

Zadržimo se sada na linearnim prognozama za $Y$ na osnovu vektora $\mathbf{X}$.

**Teorema 6.2.2** *Funkcija*

$$\psi^*(\mathbf{X})=\beta^*_0+\boldsymbol{\beta}^{*'}\mathbf{X}$$

*će biti najbolja linearna prognoza za $Y$ na osnovu vektora $\mathbf{X}$, gde je*

$$\beta^*_0=EY-\boldsymbol{\beta}^{*'}E\mathbf{X}\quad,\quad\boldsymbol{\beta}^*=\boldsymbol{\Sigma}^{-1}\mathbf{a}\quad,\quad\boldsymbol{\Sigma}=\|Cov(X_i,X_j)\|_{p\times p}\quad,$$

$$\mathbf{a} = (Cov(Y, X_1), \ldots, Cov(Y, X_p))',$$

*medju svim linearnim prognozama, i imaće maksimalnu korelaciju sa $Y$ medju svim linearnim prognozama za $Y$ na osnovu vektora $\mathbf{X}$.*

**Dokaz.**Podjimo od linearne funkcije $\psi(\mathbf{X}) = \beta_0 + \boldsymbol{\beta}'\mathbf{X}$. Prema uvedenom kriterijumu minimalnog srednjekvadratnog odstupanja optimalne vrednosti za $\beta_i$ će biti $\beta_i^*$ koje minimalizuju srednjekvadratno odstupanje:

$$\begin{aligned} E(Y - \beta_0 - \boldsymbol{\beta}'\mathbf{X})^2 &= E(Y - EY + EY - \beta_0 - \boldsymbol{\beta}'\mathbf{X} + \boldsymbol{\beta}'E\mathbf{X} - \boldsymbol{\beta}'E\mathbf{X})^2 = \\ &= E[(Y - EY) + (EY - \beta_0 - \boldsymbol{\beta}'E\mathbf{X}) - \boldsymbol{\beta}'(\mathbf{X} - E\mathbf{X})]^2 . \end{aligned} \quad (6.16)$$

Uvedimo sledeće oznake

$$b_0 = \beta_0 - EY + \boldsymbol{\beta}'E\mathbf{X} \quad , \quad \sigma_Y^2 = E(Y - EY)^2.$$

Smenom u jednakosti (6.16), dobićemo da je

$$E(Y - \beta_0 - \boldsymbol{\beta}'\mathbf{X})^2 = \sigma_Y^2 + b_0^2 + \boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta} - 2\boldsymbol{\beta}'\mathbf{a}\,. \quad (6.17)$$

Dokažimo da su

$$\beta_0^* = EY - \boldsymbol{\beta}^{*'}E\mathbf{X} \quad \text{i} \quad \boldsymbol{\beta}^* = \boldsymbol{\Sigma}^{-1}\mathbf{a},$$

traženi optimalni koeficijenti linearne regresije. Posmatrajmo vektor $\boldsymbol{\beta}$ definisan sa

$$\boldsymbol{\beta} = \boldsymbol{\beta}^* + \boldsymbol{\delta}$$

i uvrstimo ga u kriterijumsku funkciju (6.17). Sada je

$$\begin{aligned} E(Y - \beta_0 - \boldsymbol{\beta}'\mathbf{X})^2 &= \sigma_Y^2 + b_0^2 + (\boldsymbol{\beta}^* + \boldsymbol{\delta})'\boldsymbol{\Sigma}(\boldsymbol{\beta}^* + \boldsymbol{\delta}) - 2(\boldsymbol{\beta}^* + \boldsymbol{\delta})'\mathbf{a} = \\ &= \sigma_Y^2 + b_0^2 + \mathbf{a}'\boldsymbol{\Sigma}^{-1}\mathbf{a} + \boldsymbol{\delta}'\boldsymbol{\Sigma}\boldsymbol{\delta} - 2\mathbf{a}'\boldsymbol{\Sigma}^{-1}\mathbf{a} = \\ &= \sigma_Y^2 + b_0^2 + \boldsymbol{\delta}'\boldsymbol{\Sigma}\boldsymbol{\delta} - \mathbf{a}'\boldsymbol{\Sigma}^{-1}\mathbf{a} \geq \\ &\geq \sigma_Y^2 - \mathbf{a}'\boldsymbol{\Sigma}^{-1}\mathbf{a}\,. \end{aligned}$$

Jednakost će se postići ako i samo ako je $b_0 = 0$ i $\boldsymbol{\delta} = \mathbf{0}$, tj. nula vektor, odnosno ako i samo ako je

$$\beta_0 = \beta_0^* \quad \text{i} \quad \boldsymbol{\beta} = \boldsymbol{\beta}^*\,.$$

Dokažimo još deo tvrdjenja koji se odnosi na koreliranost:

$$\begin{aligned} Cov(Y, \boldsymbol{\beta}'\mathbf{X}) &= E[(Y - EY)(\boldsymbol{\beta}'\mathbf{X} - E(\boldsymbol{\beta}'\mathbf{X}))] = \\ &= \boldsymbol{\beta}'E[(Y - EY)(\mathbf{X} - E\mathbf{X})] = \\ &= (\beta_1, \ldots, \beta_p) \begin{pmatrix} Cov(Y, X_1) \\ Cov(Y, X_2) \\ \vdots \\ Cov(Y, X_p) \end{pmatrix} = \\ &= \boldsymbol{\beta}'\mathbf{a} = \boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta}^*. \end{aligned}$$

S druge strane je

$$Cov(Y, \boldsymbol{\beta}^{*'}\mathbf{X}) = \boldsymbol{\beta}^{*'}\boldsymbol{\Sigma}\boldsymbol{\beta}^{*} = D(\boldsymbol{\beta}^{*'}\mathbf{X}) \geq 0.$$

Dakle, korelacija izmedju $Y$ i $\boldsymbol{\beta}^{*'}\mathbf{X}$ će iznositi:

$$\rho^2_{Y,\beta^{*\prime}\mathbf{X}} = \frac{Cov^2(Y, \boldsymbol{\beta}^{*'}\mathbf{X})}{\sigma^2_Y \sigma^2_{\beta^{*\prime}\mathbf{X}}},$$

a odavde sledi da je

$$\sigma^2_Y \rho^2_{Y,\beta^{*\prime}\mathbf{X}} = Cov(Y, \boldsymbol{\beta}^{*'}\mathbf{X}) = \boldsymbol{\beta}^{*'}\boldsymbol{\Sigma}\boldsymbol{\beta}^{*},$$

te je

$$\rho^2_{Y,\beta'\mathbf{X}} = \frac{(\boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta}^{*})^2}{\sigma^2_Y \boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta}}\,.$$

Ako primenimo nejednakost Koši–Švarc–Bunjakovskog, dobićemo da je:

$$\begin{aligned}\sigma^2_Y \rho^2_{Y,\beta'\mathbf{X}} &= \frac{(\boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta}^{*})^2}{\boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta}} \leq \\ &\leq \frac{(\boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta})(\boldsymbol{\beta}^{*'}\boldsymbol{\Sigma}\boldsymbol{\beta}^{*})}{\boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta}} = \\ &= \boldsymbol{\beta}^{*'}\boldsymbol{\Sigma}\boldsymbol{\beta}^{*} = \sigma^2_Y \rho^2_{Y,\beta^{*\prime}\mathbf{X}}\,.\end{aligned}$$

Dakle, dobija se da je:

$$\rho^2_{Y,\beta'\mathbf{X}} \leq \rho^2_{Y,\beta^{*\prime}\mathbf{X}} \quad \text{odnosno} \quad |\rho_{Y,\beta'\mathbf{X}}| \leq |\rho_{Y,\beta^{*\prime}\mathbf{X}}|\,,$$

a odavde je konačno

$$|\rho_{Y,\psi(\mathbf{X})}| \leq |\rho_{Y,\psi^{*}(\mathbf{X})}|$$

što je i trebalo dokazati. □

Imajući u vidu teoremu 6.2.1, da je uslovno matematičko očekivanje $M(\mathbf{X})$ statistika koja je maksimalno korelirana sa $Y$ u odnosu na sve druge prognoze za $Y$ na osnovu $\mathbf{X}$, to ako je $M(\mathbf{X})$ linearna funkcija, na osnovu teoreme 6.2.2 sledi da je $M(\mathbf{X})$ oblika

$$M(\mathbf{X}) = EY + \mathbf{a}'\boldsymbol{\Sigma}^{-1}(\mathbf{X} - E\mathbf{X})\,. \tag{6.18}$$

Zaista, ako je $M(\mathbf{X}) = \beta_0 + \boldsymbol{\beta}'\mathbf{X}$, a optimalne vrednosti za $\beta_0$ i $\boldsymbol{\beta}$ su date teoremom 6.2.2, zamenom je lako proveriti da je $M(\mathbf{X})$ dato sa (6.18).

**Primer 94.** Neka na slučajni ishod eksperimenta $Y$ utiče samo jedan slučajni faktor $X$ i neka je poznata njihova zajednička raspodela koja je normalna

$$(X,Y) : \mathcal{N}\left(\mu_X\,,\,\mu_Y\,,\,\sigma^2_X\,,\,\sigma^2_Y\,,\,\rho\right) \quad , \quad \rho = Cov\left(\frac{X-\mu_X}{\sigma_X}\,,\,\frac{Y-\mu_Y}{\sigma_Y}\right).$$

U tom slučaju je njihova zajednička gustina

$$f_{XY}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma^2_X} - 2\rho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \right.\right.$$
$$\left.\left. + \frac{(y-\mu_Y)^2}{\sigma^2_Y}\right]\right\},$$

a uslovna gustina za $Y$ pod uslovom $X = x$ je

$$f_{Y|X}(y|x) = \frac{1}{\sigma_Y\sqrt{2\pi(1-\rho^2)}} exp\left\{-\frac{(y-m(x))^2}{2\sigma_Y^2(1-\rho^2)}\right\},$$

gde je

$$m(x) = \mu_Y + \frac{\sigma_Y}{\sigma_X}\rho(x-\mu_X).$$

Dakle, uslovna raspodela je takodje normalna i to $\mathcal{N}\Big(m(x), \sigma_Y^2(1-\rho^2)\Big)$, pa je

$$M(x) = E(Y|X=x) = m(x) = \mu_Y + \frac{\sigma_Y}{\sigma_X}\rho(x-\mu_X).$$

Možemo da konstatujemo da je funkcija regresije $Y$ po $X$ linearna:

$$M(X) = EY + \frac{Cov(X,Y)}{\sigma_X^2}(X-EX).$$

Uslovna disperzija, odnosno srednjekvadratna greška prognoze, je

$$\sigma_{YX}^2 = \sigma_Y^2 - \sigma_M^2$$

a disperzija za $M$ i, kao posledica, korelacioni količnik su

$$\sigma_M^2 = E\left(M(X) - EM(X)\right)^2 = \frac{Cov^2(X,Y)}{\sigma_X^2} \quad , \quad \eta_{YX}^2 = \frac{\sigma_M^2}{\sigma_Y^2} = \rho^2.$$

Kao što se vidi, najbolja prognoza u smislu srednje kvadratnog odstupanja kod normalne raspodele je *linearna.* $\triangle$

# Glava 7

# Analiza rasipanja

Analiza rasipanja ili analiza disperzija ili disperziona analiza ili analiza varijansi ili analiza odstupanja je metod za razlučivanje ostvarenih (opserviranih) variranja (odstupanja, rasturanja,rasipanja, disperzije) u eksperimentalnim podacima u odnosu na izvore variranja. Metod se sastoji u tome što se ukupne varijacije predstavljaju kao zbir varijacija za koje se može odrediti izvor, odnosno uzrok ili faktor koji ih prouzrokuje i onih za koje se ne može odrediti izvor.

Ako već treba dovesti u vezu različite delove sveukupne varijacije sa uzročnicima, onda eksperiment mora da bude projektovan tako da se ovo povezivanje obavi na logički strog način.

Metod je razvio Fišer i izložio ga 1923. godine.

Okosnica ovog metoda je statistika $Q$ koja se koristi u ocenjivanju disperzije obeležja:

$$Q = \sum_{i=1}^{n} (X_i - \overline{X}_n)^2$$

na osnovu uzorka $\mathbf{X} = (X_1, \ldots, X_n)$. Ova statistika će se u disperzionoj analizi razbijati na sume iz kojih će se odredjivati uzroci variranja podataka u odnosu na očekivanu vrednost.

Analiza disperzija je skup statističkih postupaka koji se bave uglavnom analizom uticaja dejstva jednog ili više faktora na ishod eksperimenta – posmatrano obeležje. Samo ime **analiza disperzija** (ili **varijansi**) potiče otuda što se pre svega koriste statistike koje su zbirovi kvadrata nekih odstupanja. Uobičajena skraćenica za analizu disperzija je ANOVA što potiče od naziva ovog metoda na engleskom jeziku: Analysis of Variance.

Ovde će biti reči o jednofaktorskoj i dvofaktorskoj analizi.

## 7.1 Jednofaktorski problem

Ispituje se uticaj jednog neslučajnog faktora koji u eksperimentu ima $k$ različitih vrednosti, $k \geq 2$. Ove različite vrednosti se nazivaju nivoi uticaja posmatranog faktora na obeležje $X$ koje predstavlja ishod eksperimenta. Nivoi se opisuju kvalitativno ili kvantitativno.

**Primer 95.** Na tržištu se nude tri kvaliteta kafe: minas, santos i afrička vrsta. Statističkim postupkom na osnovu eksperimenta sprovedenog u više gradova treba utvrditi da

li će vrsta kafe uticati na prihod ostvaren prodajom kafe (ukoliko je cena svih vrsta kafe ista).

Dakle, posmatra se jedan kvalitativni faktor – vrsta kafe (na tri različita nivoa), na slučajni ishod eksperimenta – prihod.△

**Primer 96.** Eksperimentom treba utvrditi dozu leka u $mg$ koju treba primenjivati u terapiji odredjene bolesti kod pacijenata.

Ovde posmatramo kvantitativni faktor – doza leka, na slučajni ishod eksperimenta – poboljšanje stanja pacijenta u posmatranoj bolesti.△

Na nivou $j$, $1 \leq j \leq k$, uzima se prost uzorak obima $n_j$ : $(X_{j1}, X_{j2}, \ldots, X_{jn_j})$. Tako se dolazi do $k$ nezavisnih uzoraka koji ne moraju biti istog obima. Ako za obeležje $X$ važi $E(X) = m$, a pod uticajem uočenog faktora na nivou $j$ u populaciji se dobija $E(X_{ji}) = m_j$, $i = 1, 2, \ldots, n_j$, tada se veličina $\mu_j = m_j - m$ naziva efekat $j$–tog nivoa. Osim toga pretpostavlja se da elementi uzorka $X_{ji}$ imaju $\mathcal{N}(m_j, \sigma^2)$ raspodelu, $j = 1, 2, \ldots, k$, za svako $i$, tj. kao i kod modela normalne regresije druge vrste, matematički model za jednofaktorski problem je

$$X_{ji} = m + \mu_j + \varepsilon_{ij}, \quad i = 1, 2, \ldots n_j, \ j = 1, 2, \ldots, k$$

gde su $\varepsilon_{ij}$ nezavisne slučajne promenljive sa istom raspodelom, $\varepsilon_{ij} : \mathcal{N}(0, \sigma^2)$. S obzirom na definiciju efekta pojedinog nivoa, sledi da je $\sum_{j=1}^{k} \mu_j = 0$.

Testira se hipoteza da uočeni faktor ne utiče na obeležje $X$, što se najčešće izražava preko testiranja efekata pojedinačnih nivoa uočenog faktora

$$H_0(\mu_1 = \mu_2 = \ldots = \mu_k = 0)$$

ili direktno

$$H_0(m_1 = m_2 = \ldots = m_k),$$

protiv alternative

$$H_1(\exists j; j \in \{1, 2, \ldots, k\}, \ \mu_j \neq 0)$$

odnosno

$$H_1(\exists j, l; \ j, l \in \{1, 2, \ldots, k\}, \ m_j \neq m_l).$$

Kako se matematičko očekivanje može da oceni sredinom uzorka, uvode se statistike:

- sredina celog uzorka

$$\overline{X} = \frac{1}{n} \sum_{j=1}^{k} \sum_{i=1}^{n_j} X_{ji} = \frac{1}{n} \sum_{j=1}^{k} n_j \overline{X}_j, \quad n = \sum_{j=1}^{k} n_j,$$

- sredina poduzorka koji odgovara $j$-tom nivou

$$\overline{X}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} X_{ji}, \quad j = 1, 2, \ldots, k,$$

- totalna (ukupna) suma kvadrata

$$Q = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ji} - \overline{X})^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j} X_{ji}^2 - n\overline{X}^2,$$

- rezidualna suma kvadrata

$$Q_u = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ji} - \overline{X}_j)^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j} X_{ji}^2 - \sum_{j=1}^{k} n_j\overline{X}_j^2,$$

- suma kvadrata odstupanja od hipoteze $H_0$ – varijacije nastale usled dejstva posmatranog faktora

$$Q_s = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(\overline{X}_j - \overline{X})^2 = \sum_{j=1}^{k} n_j\overline{X}_j^2 - n\overline{X}^2.$$

Lako je proveriti da je $Q = Q_u + Q_s$. Statistika $Q$ predstavlja zbir svih varijacija, $Q_u$–varijacije u okviru istog nivoa (zbir kvadrata odstupanja unutar nivoa–"unutrašnje" varijacije), a $Q_s$–varijacije medju različitim nivoima (zbir kvadrata odstupanja medju nivoima–"spoljašnje" varijacije).

Pod uslovom da je nulta hipoteza tačna, statistika

$$F_{k-1,\ n-k} = \frac{(n-k)Q_s}{(k-1)Q_u}$$

ima Fišerovu raspodelu sa $(k-1, n-k)$ stepeni slobode. Ako se za realizovani uzorak $(x_{j1}, x_{j2}, \ldots, x_{jn_j})$, $j = 1, 2, \ldots, k$ realizovana vrednost statistike $F_{k-1,\ n-k}$ označi sa $f_{k-1,\ n-k}$, a kvantil reda $1-\alpha$ slučajne promenljive sa $F_{k-1,\ n-k}$ raspodelom označi sa $F_{k-1,\ n-k;\ 1-\alpha}$, najbolja kritična oblast veličine $\alpha$ je

| $H_0$ | $H_1$ | $C$ |
|---|---|---|
| $\mu_1 = \mu_2 = \ldots = \mu_k = 0$ | $\exists j, \quad j \in \{1, 2, \ldots, k\}$, $\mu_j \neq 0$ | $f_{k-1,\ n-k} \geq F_{k-1,\ n-k;\ 1-\alpha}$ |

.

**Primer 97.** Data je tabela uspeha 20 ispitanika koji su bili podeljeni u 4 grupe. Svaku od grupa podučavao je po jedan instruktor. Po obavljenoj obuci ispitanici su bili testirani, a rezultati koje su pokazali predstavljeni su tabelom:

| Instruktor | Uspeh ispitanika | | | | |
|---|---|---|---|---|---|
| I | 80 | 50 | 45 | 30 | 90 |
| II | 78 | 59 | 67 | 83 | 100 |
| III | 65 | 75 | 33 | 88 | 56 |
| IV | 72 | 99 | 51 | 86 | 23 |

Testirati hipotezu da različiti instruktori nisu uticali na uspeh ispitanika sa pragom značajnosti $\alpha = 0,05$.

Testira se hipoteza $H_0$(različiti instruktori ne utiču na uspeh ispitanika). Na osnovu datih podataka dobija se sledeća tabela suma:

| Instruktor | $\sum_i x_{ji}$ | $\sum_i x_{ji}^2$ | $n_j$ | $\bar{x}_j$ |
|---|---|---|---|---|
| I | 295 | 19925 | 5 | 59 |
| II | 387 | 30943 | 5 | 77,4 |
| III | 317 | 21819 | 5 | 63,4 |
| IV | 331 | 25511 | 5 | 66,2 |
| $\sum_j$ | 1330 | 98198 | 20 | 266 |

Sredina celog uzorka iznosi 66,5. Takodje, dobija se da je

$$\begin{aligned} Q_s &= \left(5 \cdot 59^2 + 5 \cdot 77,4^2 + 5 \cdot 63,4^2 + 5 \cdot 66,2^2\right) - 20 \cdot 66,5^2 = \\ &= 89368,8 - 88445 = 923,8 \\ Q_u &= 98198 - 89368,8 = 8829,2 \\ Q &= 98198 - 88445 = 9753. \end{aligned}$$

Realizovana vrednost statistike $F_{k-1,n-k}$ je

$$f_{3,16} = \frac{16 \cdot 923,8}{3 \cdot 8829,2} = 0,558,$$

koja, s obzirom da je $F_{3,16;0,95} = 3,24$, ne pripada kritičnoj oblasti, te se nulta hipoteza prihvata sa pragom značajnosti $\alpha = 0,05$. $\triangle$

Pretpostavka o normalnoj raspodeli je suštinska za primenu disperzione analize. Medjutim, u ovom primeru nije vršeno prethodno testiranje takve hipoteze s obzirom na opšteprihvaćenu činjenicu da kognitivna (saznajna) obeležja (za kakvo se može smatrati korišćeno u primeru) imaju normalnu raspodelu. Slično će biti i u nekoliko narednih primera.

Ako se statističkom analizom dodje do zaključka da treba odbaciti nultu hipotezu, tada treba izvršiti testiranje pojedinačnih hipoteza po svim parovima nivoa:

$$H_0(m_j = m_l) \quad \text{protiv} \quad H_1(m_j \neq m_l) \ \text{ za } \ j,l \in \{1,2,\ldots,k\},\ j \neq l,$$

postupkom o kome je bilo reči u delu 2, poglavlje 4.2.2, da bi se utvrdilo koji nivoi posmatranog faktora utiču na obeležje $X$. Treba napomenuti da postoje i preciznije metode za proveru ovih relacija od $t$–testa obradjenog u pomenutom poglavlju.

**Primer 98.** Prilikom ispitivanja da li buka utiče na pamćenje, učenici jednog odeljenja su dobili da uče jedan tekst koji je trebalo da reprodukuju posle izvesnog vremena. Dobijeni su sledeći rezultati procenata zapamćenog materijala:

| Tip učionice | Procenat zapamćenog materijala | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Tiha | 88 | 75 | 68 | 59 | 65 | 62 | 92 | 81 |
| Bučna | 50 | 54 | 62 | 48 | 25 | 36 | 32 | 60 |

.

Testirati hipotezu da buka ne utiče na pamćenje učenika sa pragom značajnosti $\alpha = 0,05$.

Dakle, nulta hipoteza je $H_0$(buka ne utiče na pamćenje učenika). Na osnovu datih podataka dobija se da je $Q_s = 3108,06$ , $Q_u = 2328,38$ i $Q = 5436,44$. Realizovana vrednost test statistike je

$$f_{1,14} = 18,688.$$

S obzirom da je $F_{1,14;0,95} = 4,60$ , to realizovana vrednost pripada kritičnoj oblasti, te se nulta hipoteza odbacuje sa pragom značajnosti $\alpha = 0,05$. Dakle, buka utiče na pamćenje (pa se dalje formalno proverava da li na smanjenje ili na poboljšanje pamćenja), ali s obzirom da su ispitivana samo dva nivoa faktora uticaja, nema drugih parova da bi bilo potrebe produžiti testiranje po parovima. $\triangle$

## 7.2 Dvofaktorski problem

### 7.2.1 Dvofaktorski problem na prostom uzorku

Dvofaktorski problem nastaje kada treba da se ispita uticaj dva faktora, recimo $A$ i $B$ na obeležje $X$. Neka se uticaj faktora $A$ ispituje na $k$ ($k \geq 2$) nivoa, a faktora $B$ na $l$ ($l \geq 2$) nivoa. Prost slučajni uzorak obima $k \times l$ predstavlja se dvodimenzionalno, tabelom:

| $A \downarrow \backslash B \rightarrow$ | 1 | 2 | ... | $l$ | |
|---|---|---|---|---|---|
| 1 | $X_{11}$ | $X_{12}$ | ... | $X_{1l}$ | $\overline{X}_{1\bullet}$ |
| 2 | $X_{21}$ | $X_{22}$ | ... | $X_{2l}$ | $\overline{X}_{2\bullet}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ |
| k | $X_{k1}$ | $X_{k2}$ | ... | $X_{kl}$ | $\overline{X}_{k\bullet}$ |
| | $\overline{X}_{\bullet 1}$ | $\overline{X}_{\bullet 2}$ | ... | $\overline{X}_{\bullet l}$ | $\overline{X}$ |

gde je $X_{ij}$ vrednost obeležja $X$ na elementu uzorka izloženom $i$–tom nivou faktora $A$ i $j$–tom nivou faktora $B$. Dakle, svaki ukršteni nivo $(i, j)$ ispitivanih faktora primenjuje se na tačno jedan element uzorka.

Za definisanje modela koriste se sledeće oznake. Kao i kod jednofaktorskog problema, $m = E(X)$, a zatim, $m_{i\bullet} = E(X_{ij})$, $j = 1, 2, \ldots, l$, $m_{\bullet j} = E(X_{ij})$, $i = 1, 2, \ldots, k$, tj. $m_{i\bullet}$ je matematičo očekivanje obeležja $X$ u populaciji koja je od faktora $A$ izložena samo $i$–tom nivou, a $m_{\bullet j}$ je matematičko očekivanje obeležja $X$ u populaciji koja je od faktora $B$ izložena samo $j$–tom nivou. Sa $\mu_i = m_{i\bullet} - m$ označava se efekat $i$–tog nivoa faktora $A$, a sa $\nu_j = m_{\bullet j} - m$ efekat $j$–tog nivoa faktora $B$.

Matematički linearni model dvofaktorske analize disperzija na prostom uzorku je

$$X_{ij} = m + \mu_i + \nu_j + \varepsilon_{ij},$$

gde su $\varepsilon_{ij}$ nezavisne identički raspodeljene $\mathcal{N}(0, \sigma^2)$ slučajne promenljive, pri čemu je $\sigma^2$ nepoznato. U okviru ovog modela testiraju se sledeće nulte hipoteze:

- $H_{0A}\,(\mu_1 = \mu_2 = \ldots = \mu_k = 0)$ — efekti nivoa faktora $A$ na obeležje $X$ su bez bitnih razlika;
- $H_{0B}\,(\nu_1 = \nu_2 = \ldots = \nu_l = 0)$ — efekti nivoa faktora $B$ na obeležje $X$ su bez bitnih razlika;
- $H_{0AB}\,(\mu_1 = \mu_2 = \ldots = \mu_k = \nu_1 = \nu_2 = \ldots = \nu_l = 0)$ — efekti nivoa oba posmatrana faktora $A$ i $B$ na slučajni ishod eksperimenta $X$ su bez bitnih razlika;

protiv alternativnih redom:

- $H_{1A}\,(\exists i,\ i \in \{1, 2, \ldots, k\},\ \mu_i \neq 0)$
- $H_{1B}\,(\exists j,\ j \in \{1, 2, \ldots, l\},\ \nu_j \neq 0)$
- $H_{1AB}\,(\exists (i, j)$ tako da $(i, j) \in \{1, 2, \ldots, k\} \times \{1, 2, \ldots, l\}, (\mu_i, \nu_j) \neq (0, 0))$.

Za sprovodjenje testova potrebne su sledeće statistike:

- sredina (celog) uzorka
$$\overline{X} = \frac{1}{kl} \sum_{i=1}^{k} \sum_{j=1}^{l} X_{ij},$$
- sredina poduzorka koji odgovara $i$-tom nivou faktora $A$
$$\overline{X}_{i\bullet} = \frac{1}{l} \sum_{j=1}^{l} X_{ij},$$
- sredina poduzorka koji odgovara $j$-tom nivou faktora $B$
$$\overline{X}_{\bullet j} = \frac{1}{k} \sum_{i=1}^{k} X_{ij},$$
- ukupna suma kvadrata odstupanja od srednje vrednosti, odnosno, uzoračke sredine celog uzorka
$$Q = \sum_{i=1}^{k} \sum_{j=1}^{l} (X_{ij} - \overline{X})^2,$$
- suma kvadrata odstupanja za faktor $A$
$$Q_A = l \sum_{i=1}^{k} (\overline{X}_{i\bullet} - \overline{X})^2,$$

- suma kvadrata odstupanja za faktor $B$

$$Q_B = k\sum_{j=1}^{l}(\overline{X}_{\bullet j} - \overline{X})^2,$$

- slučajna suma kvadrata

$$Q_S = Q - Q_A - Q_B.$$

Pod uslovom da su nulte hipoteze tačne, statistike

$$F_{k-1,(k-1)(l-1)} = \frac{(k-1)(l-1)Q_A}{(k-1)Q_S},$$

$$F_{l-1,(k-1)(l-1)} = \frac{(k-1)(l-1)Q_B}{(l-1)Q_S}, \text{ i}$$

$$F_{k+l-2,(k-1)(l-1)} = \frac{(k-1)(l-1)(Q_A+Q_B)}{(k+l-2)Q_S},$$

imaju Fišerove raspodele sa naznačenim brojem stepeni slobode. Ako se, kao i kod jednofaktorske analize, koriste oznake

$$f_{k-1,(k-1)(l-1)} \quad , \quad f_{l-1,(k-1)(l-1)} \quad \text{i} \quad f_{k+l-2,(k-1)(l-1)}$$

za realizovane vrednosti poslednjih statistika redom, najbolje kritične oblasti veličine $\alpha$ su:

| $H_0$ | $H_1$ | $C$ |
|---|---|---|
| $\mu_1 = \mu_2 = \ldots$ $\ldots = \mu_k = 0$ | $\exists i, i \in \{1,2,\ldots,k\}$, $\mu_i \neq 0$ | $f_{k-1,(k-1)(l-1)} \geq$ $F_{k-1,(k-1)(l-1);\ 1-\alpha}$ |
| $\nu_1 = \nu_2 = \ldots$ $\ldots = \nu_l = 0$ | $\exists j, j \in \{1,\ldots,l\}$, $\nu_j \neq 0$ | $f_{l-1,(k-1)(l-1)} \geq$ $F_{l-1,(k-1)(l-1);\ 1-\alpha}$ |
| $\mu_1 = \ldots = \mu_k =$ $= \nu_1 = \ldots = \nu_l = 0$ | $\exists (i,j), (i,j) \in$ $\{1,\ldots,k\} \times \{1,\ldots,l\}$ $\wedge(\mu_i,\ \nu_j) \neq (0,0)$ | $f_{k+l-2,(k-1)(l-1)} \geq$ $F_{k+l-2,(k-1)(l-1);\ 1-\alpha}$ |

.

**Primer 99.** Ispituje se da li buka i vrsta teksta utiču na pamćenje učenika. Dobijeni su sledeći procenti zapamćenog materijala:

| Učionica \ Materijal | Besmisleni slogovi | Proza | Poezija | Formule |
|---|---|---|---|---|
| Tiha | 58 | 85 | 73 | 61 |
| Bučna | 25 | 48 | 52 | 28 |

.

Testiraćemo hipotezu da buka ne utiče na pamćenje učenika, hipotezu da vrsta materijala ne utiče na pamćenje učenika i hipotezu da buka i vrsta materijala ne utiču na pamćenje učenika sve sa pragom značajnosti $\alpha = 0,05$.

Realizovane vrednosti odgovarajućih $F$ statistika su:

| Suma kvadrata | $f$ |
|---|---|
| $Q_A = 1922$ | 80,083 |
| $Q_B = 949,5$ | 13,188 |
| $Q_S = 72$ | /// |

Za prag značajnosti 0,05 granica kritične oblasti je kvantil $F_{1,3;0,95} = 10,1$. S obzirom na realizovanu vrednost test statistike 80,083 , $H_{0A}$ se odbacuje, tj. može se zaključiti da buka utiče na pamćenje učenika.

Što se tiče druge nulte hipoteze vezane za vrstu teksta, zna se da je $F_{3,3;0,95} = 9,28$ i kako je realizovana vrednost test statistike jednaka 13,188 , to se i druga nulta hipoteza odbacuje, tj. može da se zaključi da pamćenje učenika zavisi i od vrste materijala koji se uči.

Za treću hipotezu je $F_{4,3;0,95} = 9,12$ i kako je realizovana vrednost odgovarajuće test statistike jednaka 29,911 , to se i treća nulta hipoteza odbacuje, tj. zaključuje se da buka i vrsta materijala koji se uči itekako utiču na pamćenje učenika.

Zadržaćemo se ovde kratko na redosledu kojim smo vršili testiranje. U radu smo koristili redosled navodjenja hipoteza u prethodnom tekstu, medjutim, u praksi treba činiti upravo obrnuto. Dakle, kod modela dvofaktorske analize na prostom uzorku, najpre treba proveriti hipotezu $H_{0AB}$, jer njenim prihvatanjem bismo istovremeno utvrdili da će i $H_{0A}$ i $H_{0B}$ biti prihvaćene. Drugim rečima, $H_{0A}$ i $H_{0B}$ se proveravaju tek pošto se utvrdi da $H_{0AB}$ treba odbaciti. $\triangle$

Dvofaktorski problem ispitivan na prostom uzorku ne omogućava ispitivanje medjuzavisnosti dva ispitivana faktora. Testiranje medjuzavisnosti dva ispitivana kontrolisana faktora vrši se na uzorku sa ponavljanjem.

### 7.2.2 Dvofaktorski problem na uzorku sa ponavljanjem

Uzorak dvofaktorskog problema koji ima više elemenata koji odgovaraju svakom uredjenom paru nivoa $(i, j)$ posmatranih faktora je uzorak sa ponavljanjem. Pri tome broj elemenata (ponavljanja) ne mora biti jednak u svakoj "ćeliji", odnosno na svakom ukrštenom nivou $(i, j)$ posmatranih faktora. Tada je prikaz uzorka u tabeli trodimenzionalan.

Mi ćemo se ovde baviti samo slučajem jednakog broja elemenata uzorka u svakoj ćeliji:

| $A\downarrow \backslash B\rightarrow$ | 1 | 2 | ... | $l$ |
|---|---|---|---|---|
| 1 | $X_{111}$<br>$X_{112}$<br>$\vdots$<br>$X_{11r}$ | $X_{121}$<br>$X_{122}$<br>$\vdots$<br>$X_{12r}$ | ...<br>...<br>...<br>... | $X_{1l1}$<br>$X_{1l2}$<br>$\vdots$<br>$X_{1lr}$ |
| 2 | $X_{211}$<br>$X_{212}$<br>$\vdots$<br>$X_{21r}$ | $X_{221}$<br>$X_{222}$<br>$\vdots$<br>$X_{22r}$ | ...<br>...<br>$\ddots$<br>... | $X_{2l1}$<br>$X_{2l2}$<br>$\vdots$<br>$X_{2lr}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ |
| k | $X_{k11}$<br>$X_{k12}$<br>$\vdots$<br>$X_{k1r}$ | $X_{k21}$<br>$X_{k22}$<br>$\vdots$<br>$X_{k2r}$ | ...<br>...<br>$\ddots$<br>... | $X_{kl1}$<br>$X_{kl2}$<br>$\vdots$<br>$X_{klr}$ |

.

Koristićemo oznaku $E(X_{ija}) = m_{ij}$, $a = 1, 2, \ldots, r$, a medjusobni efekat faktora $A$ na nivou $i$ i faktora $B$ na nivou $j$ predstavljaćemo sa

$$\eta_{ij} = m_{ij} - (m + \mu_i + \nu_j)$$

i važi

$$\sum_{i=1}^{k} \mu_i = \sum_{j=1}^{l} \nu_j = \sum_{i=1}^{k} \eta_{ij} = \sum_{j=1}^{l} \eta_{ij} = 0.$$

Matematički linearni model dvofaktorske analize disperzija na uzorku sa ponavljanjem je

$$X_{ija} = m + \mu_i + \nu_j + \eta_{ij} + \varepsilon_{ija},$$

gde su $\varepsilon_{ija}$ nezavisne identički raspodeljene $\mathcal{N}(0, \sigma^2)$ slučajne promenljive, pri čemu se podrazumeva da je $\sigma^2$ nepoznato. U okviru ovog modela testiraju se sledeće nulte hipoteze:

- $H_{0A}\,(\mu_1 = \mu_2 = \ldots = \mu_k = 0)$ — efekti nivoa faktora $A$ na obeležje $X$ su bez bitnih razlika;
- $H_{0B}\,(\nu_1 = \nu_2 = \ldots = \nu_l = 0)$ — efekti nivoa faktora $B$ na obeležje $X$ su bez bitnih razlika;
- $H_{0AB}\,(\eta_{ij} = 0,\ \forall(i,j),\ i = 1, 2, \ldots, k;\ j = 1, 2, \ldots, l)$ — nema interaktivnog dejstva faktora $A$ i $B$ na obeležje $X$.

protiv alternativnih redom:

- $H_{1A}\,(\exists i,\ i \in \{1, 2, \ldots, k\},\ \mu_i \neq 0)$
- $H_{1B}\,(\exists j,\ j \in \{1, 2, \ldots, l\},\ \nu_j \neq 0)$
- $H_{1AB}\,(\exists(i,j),\ i \in \{1, 2, \ldots, k\},\ j \in \{1, 2, \ldots, l\},\ \eta_{ij} \neq 0.)$

Za sprovodjenje testova definišu se sledeće statistike:

- uzoračka sredina celog uzorka

$$\overline{X} = \frac{1}{k \cdot l \cdot r} \sum_{i=1}^{k} \sum_{j=1}^{l} \sum_{a=1}^{r} X_{ija},$$

- uzoračka sredina ćelije $(i, j)$

$$\overline{X}_{ij} = \frac{1}{r} \sum_{a=1}^{r} X_{ija},$$

- uzoračka sredina na nivou $i$ faktora $A$

$$\overline{X}_{i\bullet} = \frac{1}{l \cdot r} \sum_{j=1}^{l} \sum_{a=1}^{r} X_{ija},$$

- uzoračka sredina na nivou $j$ faktora $B$

$$\overline{X}_{\bullet j} = \frac{1}{k \cdot r} \sum_{i=1}^{k} \sum_{a=1}^{r} X_{ija},$$

- ukupna suma kvadrata odstupanja od srednje vrednosti celog uzorka

$$Q = \sum_{i=1}^{k} \sum_{j=1}^{l} \sum_{a=1}^{r} (X_{ija} - \overline{X})^2,$$

- suma kvadrata odstupanja za faktor $A$

$$Q_A = \sum_{i=1}^{k} \sum_{j=1}^{l} \sum_{a=1}^{r} (\overline{X}_{i\bullet} - \overline{X})^2 = l \cdot r \sum_{i=1}^{k} (\overline{X}_{i\bullet} - \overline{X})^2,$$

- suma kvadrata odstupanja za faktor $B$

$$Q_B = \sum_{i=1}^{k} \sum_{j=1}^{l} \sum_{a=1}^{r} (\overline{X}_{\bullet j} - \overline{X})^2 = k \cdot r \sum_{j=1}^{l} (\overline{X}_{\bullet j} - \overline{X})^2,$$

- suma kvadrata interaktivnog dejstva faktora $A$ i $B$

$$Q_I = \sum_{i=1}^{k} \sum_{j=1}^{l} \sum_{a=1}^{r} (\overline{X}_{ij} - \overline{X}_{i\bullet} - \overline{X}_{\bullet j} + \overline{X})^2 =$$

$$= r \sum_{i=1}^{k} \sum_{j=1}^{l} (\overline{X}_{ij} - \overline{X}_{i\bullet} - \overline{X}_{\bullet j} + \overline{X})^2,$$

- slučajna suma kvadrata

$$Q_S = \sum_{i=1}^{k}\sum_{j=1}^{l}\sum_{a=1}^{r}(X_{ija} - \overline{X}_{ij})^2.$$

Očigledno je

$$Q = Q_A + Q_B + Q_I + Q_S.$$

Pod uslovom da su nulte hipoteze tačne, statistike

$$F_{k-1,kl(r-1)} = \frac{kl(r-1)Q_A}{(k-1)Q_S},$$

$$F_{l-1,kl(r-1)} = \frac{kl(r-1)Q_B}{(l-1)Q_S} \quad \text{i}$$

$$F_{(k-1)(l-1),kl(r-1)} = \frac{kl(r-1)Q_I}{(k-1)(l-1)Q_S}$$

imaju Fišerove raspodele sa naznačenim brojem stepeni slobode. Za realizovane vrednosti poslednjih statistika, najbolje kritične oblasti veličine $\alpha$ su:

| $H_0$ | $H_1$ | $C$ |
|---|---|---|
| $\mu_1 = \mu_2 = \ldots$ $\ldots = \mu_k = 0$ | $\exists i, i \in \{1,2,\ldots,k\}$, $\mu_i \neq 0$ | $f_{k-1,kl(r-1)} \geq$ $F_{k-1,kl(r-1);\ 1-\alpha}$ |
| $\nu_1 = \nu_2 = \ldots$ $\ldots = \nu_l = 0$ | $\exists j, j \in \{1,2,\ldots,l\}$, $\nu_j \neq 0$ | $f_{l-1,kl(r-1)} \geq$ $F_{l-1,kl(r-1);\ 1-\alpha}$ |
| $\forall(i,j)$, $i = 1,2,\ldots,k$; $j = 1,2,\ldots,l$, $\eta_{ij} = 0$ | $\exists(i,j)$, $i \in \{1,2,\ldots,k\}$, $j \in \{1,2,\ldots,l\}$, $\eta_{ij} \neq 0$ | $f_{(k-1)(l-1),kl(r-1)} \geq$ $F_{(k-1)(l-1),kl(r-1);\ 1-\alpha}$ |

.

Ako se testiranje sprovodi bez upotrebe računara, formiraju se tabele za postepeno izračunavanje pomenutih suma kvadrata.

Moguće je analizirati disperzije i kada se ispituje istovremeni uticaj dva faktora a u svakoj ćeliji se ne nalazi isti broj podataka (elemenata uzorka).

Disperziona analiza se primenjuje i kod ispitivanja istovremenog uticaja više od dva faktora uticaja.

Na kraju naglasimo još jednom da se, u slučaju odbacivanja bilo koje od postavljenih nultih hipoteza, postupak testiranja po pravilu nastavlja. Ovo s toga što je odbacivanjem nulte hipoteze konstatovano da jedan ili više neslučajnih faktora utiče na ishod eksperimenta, pa je dalje od interesa utvrditi nivo na kome posmatrani faktor utiče na ishod eksperimenta. Postupak za ovo testiranje je već objašnjen kod jednofaktorskog problema.

# Glava 8

# Statistička analiza slučajnih procesa

Jedan od osnovnih zadataka pri obradi rezultata merenja, statističkih podataka, pojava koje imaju karakter slučajnih procesa je odredjivanje parametara ili funkcija koje karakterišu statistička svojstva tih procesa. U takve zadatke spadaju: ocenjivanje srednje vrednosti, ocenjivanje koeficijenata regresije, ocenjivanje disperzije, ocenjivanje korelacione funkcije, spektralne funkcije, spektralne gustine i dr. Pri tome je i u ovom slučaju jedno od najvažnijih pitanja kvalitet predložene ocene u odnosu na postavljeni kriterijum valjanosti. Medjutim, statistika slučajnih promenljivih, u najvećem delu razradjena za prost slučajni uzorak, ne može se direktno "preneti" na slučajne procese. Razlog za to je, pre svega taj, što su ordinate realizacija[1] slučajnog procesa $\{X(t), t \in T\}$, po pravilu, realizacije medjusobno zavisnih slučajnih veličina. U ovom kratkom osvrtu na statistiku slučajnih procesa bavićemo se uglavnom slučajnim procesima stacionarnim u širokom smislu ili slabo stacionarnim, tj. takvim koji imaju momente drugog reda i pri tome su zadovoljeni uslovi: a) $EX(t) = m$, tj. očekivanje je konstantno za svaki $t \in T$ i b) autokovarijansna funkcija je funkcija jedne promenljive, $K(t,s) = Cov(X(t), X(s)) = B(t-s)$, $t, s \in T$, odnosno, zavisi samo od razlike svojih argumenata. Posebnu pogodnost u postupku ocenjivanja pružaju ergodični slučajni procesi čiji se parametri mogu ocenjivati samo na osnovu jedne realizacije.

Ako je skup $T$ neprebrojiv (najčešće je u pitanju neki interval iz skupa realnih brojeva), kaže se da je dati proces sa neprekidnim vremenom. Ukoliko je $T \subseteq Z$, gde je sa $Z$ označen skup celih brojeva, kaže se da je u pitanju proces sa diskretnim vremenom. Za slučajni proces sa neprekidnim vremenom se, jednostavno, koristi termin **slučajni proces**, dok se za onaj sa diskretnim vremenom koristi termin **slučajni niz**, ili **vremenski niz**, ili **vremenska serija**.

Ovi i drugi specijalni slučajevi biće naglašavani nadalje.

Uočimo da su parametri slučajnih procesa u opštem slučaju funkcije od vremena $t \in T$. Nadalje ćemo se kratko baviti ocenama ovih funkcija.

[1] Realizacijom slučajnog procesa $X : \Omega \times T \longrightarrow R$ ili $X : \Omega \times T \longrightarrow C$ naziva se funkcija jednog argumenta u oznaci $X(t)$, $t \in T$, koja nastaje fiksiranjem elementa $\omega \in \Omega$, tj. za fiksirano $\omega \in \Omega$ u definiciji slučajnog procesa.

# 8.1 Slučajni procesi

Zadržimo se prvo na procesima neprekidnog vremena.

## 8.1.1 Ocene srednje vrednosti

Bavićemo se najpre opštim slučajem.
Neka je $\{X(t), t \in R\}$ realan slučajni proces čija je srednja vrednost

$$EX(t) = m(t)$$

nepoznata (funkcija jednog realnog argumenta). Neka je poznato $n$ **nezavisnih** realizacija ovog slučajnog procesa:

$$x_1(t), x_2(t), \ldots, x_n(t)$$

u nekom fiksiranom vremenskom intervalu pravom podskupu od $R$. Bez smanjenja opštosti možemo pretpostaviti da je u pitanju interval $[0, T_0]$, tj. $t \in [0, T_0]$. Izaberimo proizvoljan momenat $t_0 \in [0, T_0]$ i posmatrajmo zasek procesa $X(\omega, t_0)$, $\omega \in \Omega$ u oznaci $X(t_0)$. Na realizacije ordinata u tom momentu možemo gledati kao na $n$ realizovanih vrednosti slučajne veličine $X(t_0)$. Pod svim ovim uslovima se ocena vrednosti $m(t_0)$ može dobiti kao ocena nepoznatog matematičkog očekivanja slučajne promenljive $X(t_0)$ na osnovu prostog slučajnog uzorka. Dakle, uzećemo

$$\bar{x}(t_0) = \frac{1}{n} \sum_{j=1}^{n} x_j(t_0).$$

U tom slučaju za ocenu srednje vrednosti $m(t_0)$ imamo statistiku

$$\widetilde{m}(t_0) = \frac{1}{n} \sum_{j=1}^{n} X_j(t_0).$$

Jasno da je

$$E\widetilde{m}(t_0) = m(t_0),$$

odnosno da je statistika $\widetilde{m}(t_0)$ nepristrasna ocena srednje vrednosti $m(t_0)$. Srednjekvadratna greška ove ocene je

$$E[\widetilde{m}(t_0) - m(t_0)]^2 = D[\widetilde{m}(t_0)] = E\left\{\frac{1}{n}\sum_{j=1}^{n}[X_j(t_0) - m(t_0)]\right\}^2 = \frac{1}{n}D(t_0),$$

jer smo pretpostavili da su realizacije nezavisne medju sobom. Sa $D(t_0)$ je označena disperzija slučajne promenljive $X(t_0)$, gde je $t_0 \in [0, T_0]$ fiksirano.

Ukoliko je disperzija $D(t_0)$ konačna, statistika $\widetilde{m}(t_0)$ je postojana ocena srednje vrednosti.

Navedeni postupak bi trebalo ponoviti za svaku tačku $t_0 \in [0, T_0]$. To je, medjutim, po pravilu nemoguće, bilo zato što nam je u praksi najčešće dostupan samo mali broj merenja, bilo zato što u odredjenom vremenskom intervalu možemo da dodjemo do samo

jedne realizacije posmatranog slučajnog procesa. U takvim slučajevima se nameće pitanje da li možemo vršiti usrednjenje samo jedne realizacije po vremenu, umesto usrednjenja više realizacija. Ta pitanja su u tesnoj vezi sa pitanjima ergodičnosti. Naime, ergodični slučajni procesi imaju upravo to svojstvo da se može vršiti ocenjivanje njihovih parametara na osnovu samo jedne realizacije. Nadalje ćemo posmatrati samo ergodične slučajne procese.

Razmotrimo sada slabo stacionaran slučajni proces $\{X(t), t \in R\}$ sa nepoznatom srednjom vrednošću $EX(t) = m$ i nepoznatom korelacionom funkcijom $B(t)$. Neka su nam poznate vrednosti jedne realizacije procesa na intervalu $[0, T_0]$. Razbijmo taj interval na $n$ jednakih delova dužine $\Delta = \frac{T_0}{n}$. Ako za ocenu srednje vrednosti uzmemo

$$\widehat{\hat{m}} = \frac{1}{n+1} \sum_{j=0}^{n} x(j \cdot \Delta)\,,$$

dobićemo integralnu sumu funkcije $x(t)$ na intervalu $[0, T_0]$,

$$\widehat{\hat{m}} = \frac{1}{(n+1) \cdot \Delta} \sum_{j=0}^{n} x(j \cdot \Delta) \cdot \Delta = \frac{1}{T_0 + \Delta} \sum_{j=0}^{n} x(j \cdot \Delta) \cdot \Delta\,.$$

Dakle, za $n \to \infty$, odnosno, $\Delta \to 0$ dobijamo ocenu

$$\widehat{m} = \frac{1}{T_0} \int_0^{T_0} X(t)dt. \tag{8.1}$$

U opštem slučaju, za interval $[a, b]$, $a < b$,

$$\widehat{m} = \frac{1}{b-a} \int_a^b X(t)dt\,.$$

Ocena $\widehat{m}$ je nepristrasna, što se lako može pokazati, dok je srednjekvadratna greška funkcija kovarijansne funkcije procesa.

Nastavimo, bez smanjenja opštosti, da govorimo o ocenama isključivo na intervalu $[0, T_0]$.

Osim ocene (8.1), u praksi se često koriste i druge, složenije ocene srednje vrednosti, gde se pomenuta ocena javlja samo kao specijalan slučaj. Najčešće se koristi ocena

$$\widehat{m} = \int_0^{T_0} a(t)X(t)dt\,,$$

gde je $a(t)$ težinska funkcija (jezgro) koja zadovoljava uslove:

$$\int_0^{T_0} a(t)dt = 1 \tag{8.2}$$

i

$$a(t) = 0 \quad \text{za} \quad t \notin [0, T_0]\,. \tag{8.3}$$

Jezgro može biti i prekidna funkcija, a za $a(t) = \frac{1}{T_0}$ za $t \in [0, T_0]$ dobija se ocena (8.1).

Ovako definisana ocena srednje vrednosti je nepristrasna:

$$E\widehat{m} = E \int_0^{T_0} a(t)X(t)dt = \int_0^{T_0} a(t)EX(t)dt = m \int_0^{T_0} a(t)dt = m\,,$$

a njena srednjekvadratna greška zavisi od jezgra.

### 8.1.2 Ocena disperzije

O oceni za disperziju govorićemo samo kod slabo stacionarnih procesa. Takodje ćemo, bez smanjenja opštosti, pretpostaviti da je srednja vrednost procesa jednaka 0. To otuda što se slabo stacionaran proces sa očekivanjem $EX(t) = m$ može translacijom $Y(t) = X(t) - m$ prevesti u takodje slabo stacionaran proces, ali sa očekivanjem 0.

Ocena za nepoznatu disperziju, $K(t,t) = B(0) = \sigma^2$, $t \in R$, o kojoj će ovde biti reči, je statistika

$$\widehat{\sigma}^2 = \int_0^{T_0} a(t)X^2(t)dt\,,$$

gde je, kao i u prethodnom slučaju, $a(t)$, $t \in R$, jezgro. Da bi ova ocena bila nepristrasna potrebno je i dovoljno da jezgro zadovoljava iste uslove kao i ranije, tj. uslove (8.2) i (8.3). Zaista,

$$E\widehat{\sigma}^2 = E\int_0^{T_0} a(t)X^2(t)dt = \int_0^{T_0} a(t)EX^2(t)dt = \sigma^2\int_0^{T_0} a(t)dt = \sigma^2\,.$$

Srednjekvadratna greška ove ocene je

$$E(\widehat{\sigma}^2 - \sigma^2)^2 = D\widehat{\sigma}^2 = E(\widehat{\sigma}^2)^2 - \sigma^4 = \int_0^{T_0}\int_0^{T_0} a(t)a(s)K_{X^2}(t,s)dtds\,,$$

gde je $K_{X^2}(t,s) = Cov(X^2(t), X^2(s))$.

Dakle, da bi se odredila greška ocene, potrebno je poznavati kovarijansnu funkciju procesa $\{X^2(t), t \in R\}$, odnosno momente četvrtog reda posmatranog slučajnog procesa. To znači da bi za precizan odgovor na pitanje o srednjekvadratnoj grešci ocene za disperziju procesa bilo neophodno poznavanje četvorodimenzione raspodele procesa $\{X(t)\}$.

### 8.1.3 Ocena kovarijansne funkcije

Kovarijansna funkcija nam daje sliku o zavisnosti "u nizanju vrednosti" slučajnog procesa. To je jedan od važnih razloga zbog kojeg se iskazuje interesovanje za kovarijansnu funkciju. Osim toga, iz kovarijansne funkcije se na posredan način odredjuje disperzija procesa.

Razmatraćemo proizvoljan realan slučajni proces $\{X(t), t \in R\}$ sa kovarijansnom funkcijom

$$K(t,s) = E[(X(t) - m(t))(X(s) - m(s))]\,.$$

Pretpostavimo da nam je poznato $n$ nezavisnih realizacija ovog procesa na intervalu $[0, T_0]$. Za ocenu kovarijansne funkcije, za $t, s \in [0, T_0]$, može se uzeti statistika

$$\widehat{K}(t,s) = \frac{1}{n}\sum_{j=1}^{n}[X_j(t) - m(t)][X_j(s) - m(s)]\,,$$

ukoliko je poznata srednja vrednost procesa. Ukoliko pak srednja vrednost procesa nije poznata, koristi se statistika

$$\widetilde{K}(t,s) = \frac{1}{n-1}\sum_{j=1}^{n}[X_j(t) - \widetilde{m}(t)][X_j(s) - \widetilde{m}(s)]\,,$$

gde je umesto nepoznate srednje vrednosti korišćena njena ocena

$$\widetilde{m}(t) = \frac{1}{n}\sum_{j=1}^{n} X_j(t)\,.$$

Lako je pokazati da su i $\widehat{K}$ i $\widetilde{K}$ nepristrasne ocene kovarijansne funkcije posmatranog procesa.

Kod slabo stacionarnog procesa je kovarijansna funkcija oblika

$$K(t+\tau,t) = B(\tau) = E[(X(t+\tau)-m)(X(t)-m)]\,,$$

te se pod uslovom da je proces ergodičan, a po uzoru na (8.1), prelaskom na integralnu sumu na intervalu $[0,T_0]$, dobija ocena

$$\widehat{B}(\tau) = \frac{1}{T_0-\tau}\int_0^{T_0-\tau}(X(t+\tau)-m)(X(t)-m)dt$$

na osnovu dela jedne realizacije.

Posmatrajmo slučajni proces $\{Y(t), t \in R\}$ nastao translacijom slabo stacionarnog slučajnog procesa $\{X(t), t \in R\}$ za srednju vrednost $m$, $Y(t) = X(t) - m$. Ovaj se proces može koristiti za definisanje statistike za ocenu kovarijansne funkcije procesa $\{X(t), t \in R\}$:

$$\widehat{B}(\tau) = \int_0^{T_0-\tau} a(t,\tau)Y(t+\tau)Y(t)dt\,,$$

gde je jezgro takvo da je

$$\int_0^{T_0-\tau} a(t,\tau)dt = 1 \tag{8.4}$$

i

$$a(t,\tau) = 0 \quad \text{za} \quad t \notin [0,T_0]\,, \tag{8.5}$$

što obezbedjuje nepristrasnost ocene.

Srednjekvadratna greška ovih ocena je i sama funkcija od kovarijansne funkcije (koja je nepoznata), pa se u postupku ocenjivanja ne traže jezgra koja bi dala optimalne ocene, već se za unapred izabrano jezgro odredjuje srednjekvadratna greška. Tako je često u upotrebi jezgro

$$a(t,\tau) = \begin{cases} \frac{1}{T_0-\tau}, & 0 \le t \le T_0 \\ 0\,, & \text{van} \end{cases}\,.$$

U tom slučaju bi ocena kovarijacione funkcije bila:

$$\widehat{B}(\tau) = \frac{1}{T_0-\tau}\int_0^{T_0-\tau} Y(t+\tau)Y(t)dt.$$

Ocenjivanje koje je upravo vršeno, vršeno je pod pretpostavkom da je srednja vrednost procesa $m$ poznata. Medjutim, ako i nju treba oceniti na osnovu uzorka (dela realizacije ergodičnog slabo stacionarnog procesa), koristi se statistika

$$\widetilde{B}(\tau) = \frac{1}{T_0-\tau}\int_0^{T_0-\tau}[X(t+\tau)-\widetilde{m}][X(t)-\widetilde{m}]dt,$$

gde je $\widetilde{m}$ neka nepristrasna ocena srednje vrednosti procesa. Medjutim, ovako definisana statistika nije nepristrasna ocena kovarijansne funkcije, što se može lako pokazati.

## 8.2 Vremenske serije

Mnoge pojave, pre svega u prirodi, iako su po svojoj suštini slučajni procesi (na pr. meteo i klimatski uslovi, proticaj vode u reci i sl.) proučavamo svodeći ih na slučajne nizove. Mnoge društvene pojave, kao što su razni ekonomski indeksi, su po svojoj suštini baš slučajni nizovi.

Neka je dat prostor verovatnoća $(\Omega, \mathcal{F}, \mathcal{P})$. Ponovimo još jednom definiciju vremenske serije (slučajnog niza):

DEFINICIJA 45. *Slučajni niz* u oznaci $\{X_t, t \in Z\}$ je funkcija $X : \Omega \times Z \to R$, ili u skup kompleksnih brojeva $C$, koja je za svako fiksirano $t \in Z$ slučajna promenljiva u oznaci $X_t(\omega)$ u odnosu na prostor verovatnoća $(\Omega, \mathcal{F}, \mathcal{P})$. Za fiksirano $\omega \in \Omega$, funkcija $X(t) \equiv X_t$, $t \in Z$, se zove realizacija slučajnog niza.

Mi ćemo razamatrati samo realne vremenske nizove, tj. one čiji je kodomen skup realnih brojeva.

U praksi nam je poznat samo deo slučajnog niza nad nekim $T \subset Z$, odnosno samo deo realizacije $\{X_t,\, t \in T\}$. Zadatak statističke analize vremenskih nizova je da se na osnovu konačnog broja posmatranja, najčešće samo jedne realizacije, ocene nepoznati parametri posmatranog vremenskog niza ili proceni njegovo ponašanje "u prošlosti" ili "budućnosti", tj. za $t \in Z \setminus T$.

Definicije stroge i slabe stacionarnosti slučajnog procesa važe i za vremenske nizove.

Bavićemo se slabo stacionarnim realnim vremenskim nizovima, dakle, onim za koje važi

$$EX_t = m(= const.)$$

i

$$Cov(X_{t+k}, X_t) = E[(X_{t+k} - m)(X_t - m)] = R_k.$$

Kao što je naznačeno, kovarijansna funkcija zavisi samo od razlike "indekasa" slučajnih promenljivih. Otuda se kovarijansnom funkcijom stacionarnog niza $\{X_t\}$ smatra brojni niz $\{R_k\}_{k \in Z}$. Svojstva kovarijansne funkcije $\{R_k\}$ su analogna svojstvima funkcije $B$ proizvoljnog realnog slučajnog procesa. Uočićemo neka od svojstava o kojima do sada nije bilo reči.

Primetimo da je

$$R_0 = Cov(X_t, X_t) = E(X_t - m)^2 \equiv \sigma^2 = const.$$

Iz definicije korelacione funkcije ovog vremenskog niza sledi

$$Corr(X_{t+k}, X_t) = \frac{Cov(X_{t+k}, X_t)}{\sqrt{Cov(X_{t+k}, X_{t+k}) Cov(X_t, X_t)}} = \frac{R_k}{R_0}\,.$$

Otuda je očigledno da je

$$|R_k| \leq R_0$$

i

$$R_{-k} = R_k\,.$$

Autokovarijansna funkcija slabo stacionarnog vremenskog niza je pozitivno semidefinitna. Zaista, neka je $\{c_k\}$ proizvoljan niz realnih brojeva. Tada je

$$\sum_{k,j=1}^{n} R_{k-j}c_k c_j = \sum_{k,j=1}^{n} c_k c_j Cov(X_k, X_j) = D(\sum_{k=1}^{n} c_k X_k) \geq 0.$$

Od posebnog su interesa stacionarni vremenski nizovi za koje red $\sum_{k=-\infty}^{+\infty} R_k$ apsolutno konvergira, tj. konvergira red

$$\sum_{k=-\infty}^{+\infty} |R_k| = R_0 + 2\sum_{k=1}^{\infty} |R_k|.$$

Uočimo Furijeovu transformaciju kovarijansne funkcije $\{R_k\}$:

$$f(\lambda) = \frac{1}{2\pi}\sum_{k=-\infty}^{+\infty} R_k \cos k\lambda = \frac{1}{2\pi}R_0 + \frac{1}{\pi}\sum_{k=1}^{+\infty} R_k \cos k\lambda \quad , \quad \text{za } \lambda \in [-\pi, \pi].$$

Ako red kovarijansne funkcije $\{R_k\}$ apsolutno konvergira, onda njena Furijeova transformacija uniformno konvergira ka neprekidnoj funkciji argumenta $\lambda$. Suma reda, $f(\lambda)$, zove se *spektralna gustina* niza $\{X_t\}$. Postoji uzajamna jednoznačnost izmedju spektralne gustine i kovarijansne funkcije procesa koja omogućuje da se koeficijenti Furijeovog reda, $R_k$, odredjuju na osnovu njegove spektralne gustine formulom

$$R_k = \int_{-\pi}^{\pi} f(\lambda) \cos k\lambda d\lambda \,.$$

Ova uzajamna jednoznačnost u predstavljanju kovarijansne funkcije i spektralne gustine jedne preko druge, omogućava da se stacionarni vremenski niz može da opiše ravnopravno preko bilo koje od njih.

### 8.2.1 Ocena srednje vrednosti

Zadržimo se sada na ocenjivanju srednje vrednosti i kovarijansne funkcije realnog slabo stacionarnog i ergodičnog vremenskog niza $\{X_t\}$.

Neka nam je dato $n$ posmatranja jedne realizacije vremenskog niza $\{X_t\}$:

$$X_1, X_2, \ldots, X_n \,.$$

Tada će nepristrasna ocena srednje vrednosti $m$ biti

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \,,$$

a greška ove ocene u smislu srednjekvadratnog odstupanja

$$E(\overline{X} - m)^2 = \frac{1}{n^2}E\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i - m)(X_j - m) = \frac{1}{n}\left[R_0 + 2\sum_{k=1}^{n-1}\left(1 - \frac{k}{n}\right)R_k\right] \to 0 \quad (8.6)$$

kada $n \to +\infty$.

Do ovog zaključka se dolazi na osnovu apsolutne konvergencije reda kovarijansne funkcije. Otuda sledi da je $\overline{X}$ i postojana ocena srednje vrednosti vremenskog niza.

Na osnovu definicije spektralne gustine i zaključka (8.6) sledi

$$D(\overline{X}) = E(\overline{X} - m)^2 \sim \frac{2\pi}{n} f(0) \,.$$

### 8.2.2 Ocene kovarijansne funkcije

S obzirom na parnost kovarijansne funkcije, dovoljno je naći ocene ove funkcije samo za $k \geq 0$. I ovoga puta ćemo morati da razlikujemo dva slučaja i to, kada je srednja vrednost vremenskog niza poznata i kada to nije. Ocenjivanje ćemo vršiti na osnovu $n$ posmatranja samo jedne realizacije.

Kada je srednja vrednost vremenskog niza, $m$, poznata, statistika

$$\widehat{R}_k(n) = \frac{1}{n-k} \sum_{t=1}^{n-k} (X_t - m)(X_{t+k} - m)$$

je nepristrasna ocena nepoznate kovarijansne funkcije, što je lako pokazati. Pri tome je neophodno da imamo veći broj posmatranja nego što je korak kovarijanse, tj. ocenjivanje je moguće samo za $0 \leq k < n$.

Razmotrimo sada slučaj nepoznate srednje vrednosti. U tom slučaju se srednja vrednost mora oceniti, a takodje je neophodno da imamo veći broj posmatranja $n$ nego što je korak $k$. Statistika kojom se ocenjuje kovarijansna funkcija je

$$\widetilde{R}_k(n) = \frac{1}{n-k} \sum_{t=1}^{n-k} (X_t - \overline{X})(X_{t+k} - \overline{X}) \,.$$

Ova ocena je asimptotski nepristrasana.

Ispitivanje srednjekvadratnih grešaka u oba slučaja ocenjivanja srednje vrednosti zahteva složeniju tehniku i ovde neće biti sprovedeno.

### 8.2.3 Prognoza vremenske serije

Prilikom praktične primene analize vremenskih serija često je potrebno prognozirati ponašanje serije van vremenskog perioda u kome su vršena posmatranja. Nekada je to i osnovni cilj statističke analize. Vremenski trenuci u kojima nisu dostupne vrednosti realizacija vremenske serije se mogu odnositi, kako na "prošlost", tako i na trenutke "unutar" perioda posmatranja. Otkrivanje nepoznate "prošlosti", kao i predvidjanje "budućih" vrednosti realizacija vrši se ekstrapolacijom vremenske serije. Otkrivanje nepoznatih vrednosti unutar perioda posmatranja ostvaruje se interpolacijom. Ovde će biti reči o jednoj vrsti ekstrapolacije vrednosti vremenskog niza, prognozi. Naime, prognoza je ekstrapolacija sa korakom unapred. Ilustrovaćemo problem i metod za njegovo rešavanje samo u najjednostavnijem slučaju, tj. izložićemo samo prognozu sa korakom 1.

Pretpostavimo da su nam poznate vrednosti konačnog dela jedne realizacije procesa $\{X_t\}$ u "prošlosti", tj. u vremenskim trenucima $t = -n, -n+1, \ldots, -1, 0$. Zadatak je predvideti vrednost procesa u trenutku $t = 1$, tj. vrednost $X_1$ u "budućem trenutku". Jedan od mogućih načina za rešavanje ovog problema je nalaženje najbolje linearne prognoze u smislu kriterijuma srednjekvadratnog odstupanja od prave vrednosti.

Definišimo, dakle, linearni prediktor za $X_1$ koristeći se poznatom prošlošću vremenskog niza

$$X_{1n} = \sum_{t=-n}^{0} \beta_{tn} X_t \,,$$

gde koeficijente linearnog modela $\beta_{tn}$ treba odrediti tako da

$$E(X_1 - X_{1n})^2 = E\left(X_1 - \sum_{t=-n}^{0} \beta_{tn} X_t\right)^2$$

bude minimalno. Prema našem znanju o najboljem linearnom prediktoru (modeli regresije prve vrste), za slučaj vremenskog niza, dobijamo ocene $\beta^*_{tn}$ nepoznatih koeficijenata $\beta_{tn}$ u obliku

$$\beta^*_{tn} = \sum_{j=-n}^{0} R^{tj} R_j\,, \qquad t = -n, -n+1, \ldots, -1, 0\,, \qquad \|R^{tj}\|^0_{t,j=-n} = \|R_{|t-j|}\|^{-1}\,,$$

gde je matrica $\|R_{|t-j|}\|^0_{t,j=-n}$ matrica odgovarajućih kovarijansi.

Rešenje ovako postavljenog problema je prognoza

$$X^*_{1n} = \sum_{t=-n}^{0} \beta^*_{tn} X_t\,.$$

Srednjekvadratna greška ove prognoze je

$$E(X_1 - X^*_{1n})^2 = R_0 - \sum_{i=1}^{n+1}\sum_{j=1}^{n+1} R_j R^{ji} R_i$$

zbog parnosti kovarijansne funkcije. Problem se, naravno, usložnjava kada kovarijansna funkcija nije poznata pa je treba oceniti. O tome ovom prilikom neće biti reči.

### 8.2.4 Vremenske serije sa trendom i sezonskom komponentom

Poseban problem u statističkoj analizi vremenskih nizova predstavlja uočavanje i otklanjanje neslučajnih (determinističkih) komponenata vremenskog niza.

Na osnovu dijagrama procesa utvrdjuje se da li ima prekida u seriji kao što je nagla promena nivoa, odnosno vrednosti realizacije procesa. Preporučljivo je podeliti seriju na homogene segmente i analizirati da li postoje posmatranja koja odudaraju od ostatka serije pri čemu su moguća dva slučaja: prvo, da ta posmatranja potiču od neke druge serije i da su greškom registrovana i drugo, da proces ima u sebi neslučajne komponente koje zavise od vremena, u kom slučaju je možda moguće reprezentovati proces modelom

$$X_t = m_t + s_t + \varepsilon_t \tag{8.7}$$

gde je:

-$m_t$ sporo promenljiva funkcija od vremena tzv. **trend**,

-$s_t$ periodična funkcija od vremena sa poznatom periodom $d$ koju zovemo **sezonska komponenta**,

-$\varepsilon_t$ je slučajni šum koji je stacionaran u širokom smislu.

Linearni model (8.7) je najjednostavniji model vremenske serije sa neslučajnim komponentama.

Kod nekih vremenskih serija se uočavaju i dva tipa periodičnosti. Prvi tip se obično vezuje za jednogodišnji period i za njega se najčešće vezuje termin sezonska komponenta, dok drugi tip periodičnosti pretpostavlja period kraći od godinu dana, ili višegodišnji period. Za drugi oblik periodičnosti se obično koristi termin ciklična komponenta i označava sa $c_t$.

Cilj statističke analize vremenskih serija sa neslučajnim komponentama je da oceni i eliminiše determinističke komponente sa nadom da će se slučajni ostatak, tj. komponenta šuma, $\varepsilon_t$ pokazati stacionarnom i samim tim omogućiti primenu teorije stacionarnih slučajnih procesa. Medjutim, bilo da je posmatrana vremenska serija stacionarna ili ne, posle eliminacije neslučajnih komponenata moguće je utvrditi svojstva osnovnog procesa $\{X_t\}$ koji bi se nadalje, uz pomoć procesa $\{\varepsilon_t\}$ i neslučajnih komponenata mogao prognozirati i kontrolisati. Naglasimo, medjutim, da postoje razradjene statističke tehnike i za nestacionaran šum, ali se mi njima nećemo baviti.

Nadalje ćemo se baviti isključivo modelom (8.7) sa stacionarnim šumom.

Da bi se uopšte utvrdilo da posmatrana vremenska serija zadovoljava jednakost (8.7), sprovodi se postupak testiranja statističkih hipoteza nad elementima dela realizacije slučajnog niza. Navešćemo neke od testova.

## 8.3 Otkrivanje neslučajnih komponenata (analiza slučajnosti niza)

Dakle, potrebno je ispitati da li ima i neslučajnih komponenata u vremenskom nizu iz koga je uzet uzorak $X_1, X_2, \ldots, X_n$ na osnovu jedne realizacije tog niza.

Analiza slučajnosti vremenske serije obavlja se testiranjem nulte hipoteze o slučajnosti:

$H_0$: Vremenski niz iz koga je uzorak uzet je slučajan;

protiv alternativne hipoteze:

$H_1$: U vremenskom nizu iz koga je uzorak uzet postoji trend ili ciklična komponenta.

U poglavlju o testiranju statističkih hipoteza, već je bilo reči o testiranju slučajnosti uzorka. Dakle, nulta hipoteza $H_0$ je bila da je $(X_1, \ldots, X_n)$ prost uzorak, odnosno, da su sve slučajne promenljive $X_k, k = 1, 2, \ldots n$ nezavisne i sa istom raspodelom čija je funkcija raspodele $F(x)$, $x \in R$,

$$H_0 : F_{X_1,\ldots,X_n}(x_1, x_2, \ldots, x_n) = F(x_1)F(x_2)\ldots F(x_n)$$

protiv alternativne

$$H_1 : F_{X_1,\ldots,X_n}(x_1, x_2, \ldots, x_n) \neq F(x_1)F(x_2)\ldots F(x_n)\,.$$

U slučaju da se prihvati alternativna hipoteza $H_1$, naknadno se proverava o kom se faktoru neslučajnosti radi.

Kada se već zna da kod vremenske serije osim slučajne komponente mogu da se jave i neslučajne kao što su trend, ciklična i sezonska komponenta, onda je od interesa nultu hipotezu o slučajnosti hronološkog niza testirati protiv neke od alternativa, kao što je postojanje trenda, ciklične ili sezonske komponente.

Nadalje su navedeni neki od testova kojima se vrši testiranje ovih hipoteza.

### 8.3.1 Test tačaka zaokreta

Posmatra se **hronološki** uzorak $(X_1, X_2, \ldots, X_n)$. Na osnovu njega se definišu slučajne veličine $Y_2, Y_3, \ldots, Y_{n-1}$:

$$Y_j = \begin{cases} 1, & \text{ako je } X_j \text{ veće od oba suseda : } X_{j-1} \text{ i } X_{j+1} \\ & \text{ili ako je } X_j \text{ manje od oba suseda} \\ 0, & \text{u ostalim slučajevima} \end{cases} .$$

Očigledno da je $Y_j = 1$ kada uzorak ima ekstremnu tačku. Test statistika $Z_n$ se definiše kao ukupan broj ekstremnih tačaka u uzorku, tj. $Z_n = Y_2 + Y_3 + \ldots + Y_{n-1}$.

Ako je hipoteza o slučajnosti tačna, statistika $Z_n$ ima sledeće numeričke karakteristike

$$E(Z_n) = \frac{2(n-2)}{3}, \quad D(Z_n) = \frac{16n-29}{90}.$$

Ako se u realizovanom uzorku obima $n$ dobije sredina uzorka koja bitno odstupa od $E(Z_n)$, nulta hipoteza se odbacuje.

Za veliki obim uzorka ($n \geq 50$) koristi se test statistika

$$Z_n^* = \frac{Z_n - \frac{2(n-2)}{3}}{\sqrt{\frac{16n-29}{90}}}$$

koja ima približno normalnu normiranu raspodelu pomoću koje se nulta hipoteza testira na uobičajeni način:

| $H_0$ | $H_1$ | $C$ |
|---|---|---|
| Niz je slučajan | Postoji neslučajna komponenta | $\lvert z_n^* \rvert \geq z_{0,5-\alpha/2}$ |

.

Navedeni test tačaka zaokreta poznat je još i kao test povratnih ili ekstremnih tačaka. Ovaj test daje bolje rezultate kada je alternativna hipoteza postojanje periodične komponente u vremenskom nizu, nego kada je alternativa postojanje trenda.

### 8.3.2 Test rasta

Nasuprot testu tačaka zaokreta, postoji test pogodan za otkrivanje trenda u vremenskom nizu. Test statistika se tada definiše preko tačaka rasta.

Tačkom rasta naziva se element $X_j$ uzorka $(X_1, X_2, \ldots, X_n)$, za koji važi da je $X_j < X_{j+1}$, za $j = 1, 2, \ldots, n-1$.

Uz pomoć tačaka rasta definišu se slučajne veličine

$$Y_j = \begin{cases} 1, & \text{ako je } X_j < X_{j+1} \\ 0, & \text{u ostalim slučajevima} \end{cases} .$$

Test statistika $R_n$ je broj tačaka rasta u uzorku obima $n$ :
$R_n = Y_1 + Y_2 + \ldots + Y_{n-1}$.

Ova statistika ima sledeće numeričke karakteristike

$$E(R_n) = \frac{n-1}{2}\,, \quad D(R_n) = \frac{n+1}{12}\,,$$

odakle sledi da se nulta hipoteza odbacuje ako je broj tačaka rasta u realizovanom uzorku bitno različit od broja $E(R_n)$.

Pogodnost statistike $R_n$ je u tome što već za obim uzorka $n > 12$ statistika

$$R_n^* = \frac{R_n - \frac{n-1}{2}}{\sqrt{\frac{n+1}{12}}}$$

ima približno normalnu raspodelu $\mathcal{N}(0,1)$, pa se za testiranje nulte hipoteze koristi tablica normalne raspodele i uobičajena kritična oblast:

| $H_0$ | $H_1$ | $C$ |
|---|---|---|
| Niz je slučajan | Postoji neslučajna komponenta | $\lvert r_n^*\rvert \geq z_{0,5-\alpha/2}$ |

.

### 8.3.3 Test kvadrata uzastopnih razlika

Test kvadrata uzastopnih razlika primenjuje se kada je alternativna hipoteza postojanje trenda.

Ovaj test bazira se na statistici

$$D = \frac{\Delta^2}{2\widetilde{S}_n^2}\,,$$

gde je

$$\Delta^2 = \frac{1}{n-1}\sum_{i=1}^{n-1}(X_{i+1} - X_i)^2 \quad , \quad \widetilde{S}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \quad , \quad \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i\,.$$

Statistika $D$ ima očekivanje 1 i disperziju $\dfrac{n-2}{n^2-1}$, pa se za dovoljno veliki obim uzorka ($n \geq 20$) raspodela statistike

$$D^* = \frac{D-1}{\sqrt{\frac{n-2}{n^2-1}}}$$

može da aproksimira normalnom normiranom raspodelom, te je najbolja kritična oblast veličine $\alpha$:

| $H_0$ | $H_1$ | $C$ |
|---|---|---|
| Niz je slučajan | Postoji trend | $\lvert d^*\rvert \geq z_{0,5-\alpha/2}$ |

.

### 8.3.4 Test serijskih korelacija

Test serijskih korelacija odnosi se na proveru uzoračkih serijskih korelacija, tj. bazira se na statistici

$$r_k = \frac{\frac{1}{n-k}\sum_{i=1}^{n-k}[(X_i - \overline{X}_{n-k})(X_{i+k} - \overline{X}_{k+(n-k)})]}{\sqrt{\frac{1}{n-k}\sum_{i=1}^{n-k}(X_i - \overline{X}_{n-k})^2 \frac{1}{n-k}\sum_{i=1}^{n-k}(X_{i+k} - \overline{X}_{k+(n-k)})^2}}$$

gde je

$$\overline{X}_{n-k} = \frac{1}{n-k}\sum_{i=1}^{n-k} X_i \quad , \quad \overline{X}_{k+(n-k)} = \frac{1}{n-k}\sum_{i=1}^{n-k} X_{k+i}\,.$$

U praksi se , medjutim, statistika $r_k$ zamenjuje statistikom koja se najčešće isto označava:

$$r_k = \frac{\frac{1}{n-k}\sum_{i=1}^{n-k}[(X_i - \overline{X}_{n-k})(X_{i+k} - \overline{X}_{k+(n-k)})]}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}\,,$$

ili čak sa:

$$r_k = \frac{\frac{1}{n-k}\sum_{i=1}^{n-k}[(X_i - \overline{X}_n)(X_{i+k} - \overline{X}_n)]}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}\,,$$

za veliko $n$ i malo $k$ u odnosu na $n$, gde je $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Testiranje slučajnosti počinje izračunavanjem uzoračkih serijskih korelacija $r_1, r_2, \ldots$ i njihovom analizom. Kod slučajnog uzorka (kod koga nema trenda) je $r_1 = r_2 = \ldots = 0, (r_0 = 1)$ te je za prihvatanje nulte hipoteze potrebno da ove vrednosti za $k = 1, 2, \ldots$ kod realizovanog uzorka budu jednake (tj. veoma bliske) nuli. Ako je to ispunjeno, što znači da je nulta hipoteza tačna, tada je

$$E(r_1) = -\frac{1}{n-1}\,, \quad \text{a} \quad D(r_1) = \frac{(n-2)^2}{(n-1)^3}$$

pod pretpostavkom da je uzorak iz normalne raspodele.

U tom, vrlo uprošćenom slučaju, testiranje slučajnosti se sprovodi statistikom

$$r_1^* = \frac{r_1 + \frac{1}{n-1}}{\frac{n-2}{\sqrt{(n-1)^3}}}$$

koja ima približno normalnu normiranu raspodelu.

Odredjivanje najbolje kritične oblasti veličine $\alpha$ je uobičajeno:

| $H_0$ | $H_1$ | $C$ |
|---|---|---|
| Niz je slučajan | Postoji trend | $\|r_1^*\| \geq z_{0,5-\alpha/2}$ |

.

### 8.3.5 Test cikličnih korelacija

U vremenskim serijama u kojima se zapaža ciklični karakter, dakle u realizovanom uzorku,

$$(x_1, x_2, \ldots, x_m), (x_{m+1}, x_{m+2}, \ldots, x_{2m}), (x_{2m+1}, x_{2m+2}, \ldots, x_{3m}), \ldots$$

delovi serije za neki prirodan broj $m$ imaju isti karakter. Ako se zna koliko je $m$, definiše se statistika $c_m$, uzorački koeficijent ciklične korelacije sa korakom $m$, kao:

$$c_m = \frac{\sum_{j=1}^{n}(X_j - \overline{X}_n)(X_{m+j} - \overline{X}_{m+(n-m)})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \sum_{j=1}^{n}(X_{m+j} - \overline{X}_{m+(n-m)})^2}},$$

gde je

$$X_{m+j} = X_j \quad , \quad j = 1, 2, 3, \ldots,$$
$$\overline{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j \quad , \quad \overline{X}_{m+(n-m)} = \frac{1}{n}\sum_{j=1}^{n} X_{m+j}.$$

Na osnovu uzoračkih vrednosti koeficijenta $c_m$ za realizovani uzorak se vrši testiranje nulte hipoteze o slučajnosti, kao kod prethodnog testa, protiv alternativne da postoji cikličnost.

### 8.3.6 Metod pokretnih sredina

Metod pokretnih sredina ili kliznih proseka, primenjuje se u situacijama kada je alternativna hipoteza slučajnosti niza postojanje trenda. Ovaj metod pomaže da se trend lakše uoči. Najčešće, uz njega se crta i dijagram.

Za seriju od $n$ podataka vrši se izravnavanje pomoću usrednjavanja $m = 2k+1$ članova. Formiraju se proseci

$$Y_1 = \frac{X_1+X_2+\ldots+X_{k-1}+X_k+\ldots+X_{2k+1}}{2k+1},$$
$$Y_2 = \frac{X_2+X_3+\ldots+X_k+X_{k+1}+\ldots+X_{2k+2}}{2k+1},$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots,$$
$$Y_{n-2k} = \frac{X_{n-2k}+\ldots+X_{n-k}+\ldots+X_n}{2k+1}.$$

Formirani niz $Y_1, \ldots, Y_{n-2k}$ ima manja kolebanja nego originalni niz. To je razlog da se trend lakše uočava i odstranjuje. Treba uočiti da su $Y_i$ i $Y_j$ nekorelirani medju sobom ako je $|j - i| > m$.

Opšti način izravnavanja u okviru ovog metoda je pomoću pozitivnih pondera koji ispunjavaju uslov

$$\sum_{j=-k}^{k} c_j = 1 \quad , \quad c_j > 0 \quad , \quad j = -k, -k+1, \ldots, k\,.$$

U ovom slučaju, niz se formira na sledeći način:

$$\widetilde{Y}_t = \sum_{j=-k}^{k} c_j X_{t+j} \quad , \quad t = k+1, k+2, \ldots, n-k\,,$$

a opet je u pitanju izravnavanje sa po $m = 2k+1$ tačaka.

U oba slučaja izbor broja $k$ zaslužuje posebnu pažnju i neće biti predmet našeg detaljnijeg razmatranja. Ipak napomenimo da se polazeći od $k = 1$ pa nadalje proveravaju realizovane vrednosti uzoračkih disperzija statistika $Y_i$ i uočava momenat njihove stabilizacije oko neke konstante. Taj momenat odredjuje vrednost broja $k$ koju treba prihvatiti.

## 8.4 Otklanjanje neslučajnih komponenata

Pošto se gore navedenim testovima utvrdi postojanje neslučajnih komponenata u vremenskoj seriji, pristupa se njihovom otklanjanju. Navodimo neke od postupaka za tu namenu.

### 8.4.1 Eliminacija trenda kod procesa bez sezonske komponente

U odsustvu sezonske komponente, proces (8.7) postaje

$$X_t = m_t + \varepsilon_t \quad , \quad t = 1, \ldots, n \tag{8.8}$$

gde bez smanjenja opštosti možemo smatrati da je $E\varepsilon_t = 0$. Za eliminaciju trenda u ovom slučaju primenićemo nekoliko sledećih postupaka.

**I metod** (Ocena najmanjih kvadrata za $m_t$)

Ovaj metod pretpostavlja fitovanje neslučajne komponente parametarskom familijom funkcija, na primer

$$m_t = a_0 + a_1 t + a_2 t^2\,,$$

odredjujući parametre $a_0, a_1$ i $a_2$ po metodu najmanjih kvadrata iz uslova da $\sum_t (x_t - m_t)^2$ bude minimalno. Ocenjene vrednosti za $\varepsilon_t$ su razlike $x_t$ i $\hat{m}_t = \hat{a}_0 + \hat{a}_1 t + \hat{a}_2 t^2$.

**II metod** (Izgladjivanje pokretnom sredinom)

Neka je $q$ nenegativan ceo broj, i posmatrajmo tzv. **dvostranu pokretnu sredinu** procesa $\{X_t\}$

$$W_t = \frac{1}{2q+1} \sum_{j=-q}^{q} X_{t+j}\,.$$

Sledi da je

$$W_t = \frac{1}{2q+1} \sum_{j=-q}^{q} m_{t+j} + \frac{1}{2q+1} \sum_{j=-q}^{q} \varepsilon_{t+j} \approx m_t \quad \text{za} \quad q+1 \le t \le n-q\,,$$

ukoliko je tačna pretpostavka da je $m_t$ približno linearna na intervalu $[t-q, t+q]$ i da je $E\varepsilon_t \approx 0$ na ovom intervalu. Ocena trenda je

$$\hat{m}_t = \frac{1}{2q+1} \sum_{j=-q}^{q} X_{t+j} \quad , \quad q+1 \leq t \leq n-q\,.$$

Pošto nemamo opservacije $X_t$ za $t \leq 0$ i $t > n$, ne možemo ovu aproksimaciju da primenimo za $t \leq q$ ili $t > n-q$.

Umesto navedenih, moguće je koristiti jednostrane pokretne sredine

$$\hat{m}_t = \sum_{j=0}^{n-t} \alpha(1-\alpha)^j X_{t+j} \quad , \quad t = 1, \ldots, q$$

i

$$\hat{m}_t = \sum_{j=0}^{t-1} \alpha(1-\alpha)^j X_{t-j} \quad , \quad t = n-q+1, \ldots, n\,,$$

gde je $\alpha$ neki realan broj.

Ove ocene nisu previše osetljive na promenu $\alpha$. Empirijski je dokazano da se najbolje ocene trenda dobijaju za $\alpha$ izmedju $0,1$ i $0,3$.

**III metod** (Primena operatora razlike za generisanje podataka)

**Operator prve razlike** za posmatrani vremenski niz u oznaci $\nabla$ je definisan kao

$$\nabla X_t = X_t - X_{t-1} = (1-B)X_t$$

gde je B **operator pomeranja unazad** (za jedan).

$$BX_t = X_{t-1}$$

Stepeni operatora B i $\nabla$ se definišu na sledeći način

$$B^j(X_t) = X_{t-j} \quad , \quad \nabla^j(X_t) = \nabla(\nabla^{j-1}(X_t)) \quad , \quad j \geq 1 \quad , \quad \nabla^0(X_t) = X_t\,. \tag{8.9}$$

Sa polinomima od B i $\nabla$ se računa istovetno kao i sa realnim polinomima. Na primer,

$$\nabla^2 X_t = \nabla(\nabla X_t) = \nabla(X_t - X_{t-1}) = (X_t - X_{t-1}) - (X_{t-1} - X_{t-2}) = X_t - 2X_{t-1} + X_{t-2}\,.$$

Suština je u tome da ako se $\nabla$ primeni na linearni trend $m_t = at + b$, dobijamo konstantnu funkciju $\nabla m_t = a$. Na taj način se bilo koji polinomni trend stepena $k$ može redukovati na konstantu primenom $\nabla^k$.

Dakle, za $X_t = m_t + \varepsilon_t$ gde je $m_t = \sum_{j=0}^{k} a_j t^j$ i gde je $E\varepsilon_t = 0$ primenom operatora razlike dobijamo stacionaran proces (sa očekivanjem $k!a_k$):

$$\nabla^k X_t = k!a_k + \nabla^k \varepsilon_t\,.$$

To omogućava da polazeći od proizvoljnog zadatog niza $\{X_t\}$ podataka, višestrukom uzastopnom primenom operatora $\nabla$ dodjemo do niza $\{\nabla^k X_t\}$ koji se može modelirati kao da je jedna realizacija stacionarnog procesa. U praksi se pokazalo da k najčešće nije veliko tj. da je jedan ili dva.

### 8.4.2 Istovremena eliminacija trenda i sezonske komponente

Posmatrajmo ponovo model (8.7)

$$X_t = m_t + s_t + \varepsilon_t \quad , \quad E\varepsilon_t = 0\,,$$

kod koga je sezonska komponenta sa periodom $d$, tj.

$$s_{t+d} = s_t \quad , \quad \sum_{t=1}^{d} s_t = 0\,.$$

Kod pojave cikličnih komponenata zgodno je podatke indeksirati godinom i mesecom, tj. označićemo sa $X_{j,k}$ podatak koji odgovara $j$-toj godini, $j = 1, \ldots, n$ i $k$-tom mesecu $k = 1, \ldots, d$.

Ponovo nam je cilj uklanjanje neslučajnih komponenata, pri čemu sada imamo dve neslučajne komponente. Predstavićemo metode kojima se istovremeno uklanjaju obe.

**I metod** (Metod malog trenda)

Činjenica da je trend mali ukazuje na približno konstantnu vrednost za $m_t$ u okviru svakog pojedinog ciklusa, pri svakom $t$. Izložićemo u kratkim crtama metod koji se može da primeni na vremensku seriju kod koje se uočava mali trend.

Kako je $\sum_{k=1}^{d} s_k = 0$, to vodi do uobičajene nepristrasne ocene za $m_j$ kao matematičkog očekivanja za $X_j$:

$$\hat{m}_j = \frac{1}{d}\sum_{k=1}^{d} X_{j,k}\,,$$

a ocena za $s_k$ , $k = 1, \ldots, d$ je

$$\hat{s}_k = \frac{1}{n}\sum_{j=1}^{n}(X_{j,k} - \hat{m}_j).$$

Dakle, ocena slučajne komponente (greške, šuma) za $k$-ti mesec $j$-te godine je:

$$\hat{\varepsilon}_{j,k} = X_{j,k} - \hat{m}_j - \hat{s}_k \quad , \quad j = 1, \ldots, n \quad , \quad k = 1, \ldots, d\,.$$

**II metod** (Ocena pokretnim sredinama)

Ovaj metod je primenljiviji od predhodnog, jer ne pretpostavlja da je trend u toku jednog ciklusa približno konstantan.

Pretpostavimo da imamo posmatranja $\{X_1, \ldots, X_n\}$. U prvom koraku se ocenjuje trend primenom pokretnih sredina koje odstranjuju sezonsku komponentu i prigušuju šum.

Ako je period $d$ paran, $d = 2q$, tada koristimo

$$\hat{m}_t = \frac{1}{d}\left(\frac{1}{2}X_{t-q} + X_{t-q+1} + \cdots + X_{t+q-1} + \frac{1}{2}X_{t+q}\right) \quad , \quad q+1 \le t \le n-q\,.$$

Ako je period $d$ neparan, $d = 2q + 1$, koristimo običnu pokretnu sredinu:

$$\hat{m}_t = \frac{1}{d} \sum_{j=-q}^{q} X_{t+j} \quad , \quad q+1 \leq t \leq n-q \, .$$

U drugom koraku ocenjujemo sezonsku komponentu. Za svako $k = 1, \ldots, d$ izračunavamo sredinu $w_k$ odstupanja $\{ \, (X_{k+jd} - \hat{m}_{k+jd}) \, , \quad q < k + jd \leq n - q \, \}$:

$$w_k = \frac{1}{n-2q} \sum_{k+jd=q+1}^{n-q} (X_{k+jd} - \hat{m}_{k+jd}) \, .$$

Pošto ove aritmetičke sredine ne daju obavezno u zbiru nulu, sezonsku komponentu ocenjujemo sa

$$\hat{s}_k = w_k - \frac{1}{d} \sum_{i=1}^{d} w_i \quad \text{za} \quad k = 1, \ldots, d \quad \text{i} \quad \hat{s}_k = \hat{s}_{k-d} \quad \text{i} \quad k > d \, .$$

Uklanjanjem sezonskog uticaja iz podataka, dobijamo novi niz podataka koji nema sezonsku komponentu:

$$d_t = X_t - \hat{s}_t \quad , \quad t = 1, \ldots, n \, .$$

Konačno, ponovo ocenimo trend na osnovu niza $\{d_t\}$ nekim od već pomenutih metoda i dobijamo ocenu šuma:

$$\hat{\varepsilon}_t = X_t - \widehat{\hat{m}}_t - \hat{s}_t \, ,$$

gde je $\widehat{\hat{m}}_t$ novodobijena ocena trenda.

**III metod** (Primena razlike sa korakom $d$)

Kod ovog metoda se primenjuje **operator razlike sa korakom** $d$ u oznaci $\nabla_d$ definisan kao:

$$\nabla_d X_t = X_t - X_{t-d} = (1 - B^d) X_t \, ,$$

gde je $B$, kao i malopre, operator pomeranja unazad za jedan, a $B^d$ je njegov stepen. Pri tome treba razlikovati operator $\nabla_d$ od $d$–tog stepena $\nabla^d$ operatora prve razlike definisanog u (8.9). Primenom ovog operatora na model $X_t = m_t + s_t + \varepsilon_t$, gde $\{s_t\}$ ima period $d$, dobijamo

$$\nabla_d X_t = m_t - m_{t-d} + \varepsilon_t - \varepsilon_{t-d}$$

koji dovodi do toga da imamo seriju $\nabla_d X_t$ koja ima samo trend oblika $m_t - m_{t-d}$ i šum $\varepsilon_t - \varepsilon_{t-d}$. Sada se trend $m_t - m_{t-d}$ može da eliminiše nekim već ranije pomenutim metodom.

# Dodatak

# Glava 9

# Važnije raspodele verovatnoća

## 9.1 Raspodele diskretnog tipa

1. **Bernulijeva raspodela**
   $X : \mathcal{B}(1, p)\,, \quad 0 < p < 1$

$$P\{X = 0\} = 1 - p \qquad \text{i} \qquad P\{X = 1\} = p$$

$$E(X) = p$$
$$D(X) = p(1 - p)$$

2. **Binomna raspodela**
   $X : \mathcal{B}(n, p)\,, \quad 0 < p < 1\,, \quad n \in N$

$$P\{X = k\} = \binom{n}{k} p^k (1 - p)^{n-k} \quad , \quad 0 \leq k \leq n$$

$$E(X) = \sum_{k=0}^{n} k P\{X = k\} = np$$
$$D(X) = np(1 - p)$$

3. **Puasonova raspodela**
   $X : \mathcal{P}(\lambda)\,, \quad \lambda > 0$

$$P\{X = k\} = \frac{\lambda^k}{k!} e^{-\lambda} \quad , \quad k = 1, 2, \ldots$$

$$E(X) = \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} = \lambda$$
$$D(X) = \lambda$$

4. **Geometrijska raspodela**
   $0 < p < 1$

$$P\{X = k\} = (1-p)^{k-1} p \quad , \quad k = 1, 2, \ldots$$

$$E(X) = \sum_{k=1}^{\infty} k P\{X = k\} = \frac{1}{p}$$
$$D(X) = \frac{1}{p^2}$$

## 9.2 Raspodele apsolutno neprekidnog tipa

1. **Normalna (Gausova) raspodela**
   $X : \mathcal{N}(m, \sigma^2)\,, \quad -\infty < m < +\infty\,, \quad \sigma^2 > 0$

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-m)^2}{2\sigma^2}} \quad , \quad -\infty < x < +\infty$$
$$F(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x} e^{-\frac{(t-m)^2}{2\sigma^2}} dt$$

$$E(X) = m$$
$$D(X) = \sigma^2$$

2. **Log-normalna raspodela**

$X = \exp\{Y\}\,, \quad Y : \mathcal{N}(m, \sigma^2)\,, \quad -\infty < m < +\infty\,, \quad \sigma^2 > 0$

$$f(x) = \begin{cases} \frac{1}{x\sigma\sqrt{2\pi}} \exp\{-\frac{(\ln x - m)^2}{2\sigma^2}\}\,, & x > 0 \\ 0\,, & x \leq 0 \end{cases}$$

$$E(X) = \exp\{m + \frac{1}{2}\sigma^2\}$$

$$D(X) = \exp\{2m + \sigma^2\}\left(\exp\{\sigma^2\} - 1\right)$$

3. **Troparametarska log–normalna raspodela**

$X = \exp\{Y\} + x_0\,, \quad Y : \mathcal{N}(m, \sigma^2)\,, \quad -\infty < x_0 < +\infty$

$$f(x) = \begin{cases} \frac{1}{(x - x_0)\sigma\sqrt{2\pi}} \exp\left\{-\frac{(\ln(x - x_0) - m)^2}{2\sigma^2}\right\}, & x > x_0 \\ 0, & x \leq x_0 \end{cases}$$

4. **Uniformna raspodela**

$X : \mathcal{U}(a, b)\,, \quad a, b \in R\,, \quad a < b$

$$f(x) = \begin{cases} \frac{1}{b-a}, & x \in (a, b) \\ 0, & \text{inače} \end{cases}$$

$$F(x) = \begin{cases} 0, & x < a \\ \frac{x-a}{b-a}, & a \leq x < b \\ 1, & x \geq b \end{cases}$$

$$E(X) = \frac{a + b}{2}$$

$$D(X) = \frac{(b - a)^2}{12}$$

5. **Vejbulova raspodela**

   $a > 0\,, \quad b > 0$

$$f(x) = \begin{cases} \frac{b}{a}(\frac{x}{a})^{b-1} \exp\{-(\frac{x}{a})^b\}, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

$$E(X) = a^{-\frac{1}{b}}\Gamma\left(1 + \frac{1}{b}\right)$$

$$D(X) = a^{-\frac{2}{b}}\left(\Gamma\left(1 + \frac{2}{b}\right) - \Gamma^2\left(1 + \frac{1}{b}\right)\right)$$

6. **Eksponencijalna raspodela**

   $X : \mathcal{E}(\lambda)\,, \quad \lambda > 0$

$$f(x) = \begin{cases} \lambda \exp\{-\lambda x\}\,, & x > 0 \\ 0\,, & x \leq 0 \end{cases}$$

$$E(X) = \frac{1}{\lambda}$$

$$D(X) = \frac{1}{\lambda^2}$$

7. **Dvojna eksponencijalna raspodela**

$$f(x) = exp\left\{-e^{-x}\right\}\,, \quad -\infty < x < +\infty$$

8. **Dvostrana eksponencijalna raspodela**

   $\lambda > 0$

$$f(x) = \frac{\lambda}{2} exp\{-\lambda|x|\}\,, \quad -\infty < x < +\infty$$

9. **Košijeva raspodela**

$\alpha > 0$

$$f(x) = \frac{\alpha}{\pi(x^2+\alpha^2)}\,, \quad -\infty < x < +\infty$$

10. **Laplasova raspodela**

$\lambda > 0\,, \quad -\infty < m < +\infty$

$$f(x) = \frac{1}{2\lambda}\exp\left\{-\frac{|x-m|}{\lambda}\right\}, \quad -\infty < x < +\infty$$

11. **Gama raspodela**

$\alpha \geq 0\,, \quad \beta > 0$

$$f(x) = \begin{cases} \frac{1}{\beta^\alpha\Gamma(\alpha)}x^{\alpha-1}\exp\{-\frac{x}{\beta}\}\,, & x > 0 \\ 0\,, & x \leq 0 \end{cases}$$

$$E(X) = \alpha\beta$$
$$D(X) = \alpha\beta^2$$

12. **Troparametarska gama raspodela**

$\alpha > 0\,, \quad \beta > 0\,, \quad -\infty < x_0 < +\infty$

$$f(x) = \begin{cases} \frac{1}{\beta^\alpha\Gamma(\alpha)}(x-x_0)^{\alpha-1}\exp\left\{-\frac{x-x_0}{\beta}\right\}, & x > x_0 \\ 0\,, & x \leq x_0 \end{cases}$$

13. **Log–Pirson III raspodela**

$\alpha > 0\,, \quad \beta > 0\,, \quad -\infty < x_0 < +\infty$

$$f(x) = \begin{cases} \frac{1}{\beta^\alpha x\Gamma(\alpha)}(\ln x-x_0)^{\alpha-1}\exp\left\{-\frac{\ln x-x_0}{\beta}\right\}, & x > x_0 \\ 0\,, & x \leq x_0 \end{cases}$$

14. **Hi kvadrat ($\chi^2$) raspodela**

$X : \chi^2_\nu\,, \quad \nu \in N$

$$f(x) = \begin{cases} \dfrac{x^{\frac{\nu}{2}-1}}{2^{\frac{\nu}{2}}\Gamma\left(\dfrac{\nu}{2}\right)} \exp\{-\frac{x}{2}\}, & x > 0 \\ 0\,, & x \leq 0 \end{cases}$$

$$E(\chi^2_\nu) = \nu$$
$$D(\chi^2_\nu) = 2\nu$$

15. **Fišerova raspodela**

$X : F_{\nu_1,\nu_2}\,, \quad \nu_1\,, \nu_2 \in N$

$$X = \frac{\frac{\chi^2_{\nu_1}}{\nu_1}}{\frac{\chi^2_{\nu_2}}{\nu_2}}$$

$$f(x) = \begin{cases} \dfrac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)} \left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}} x^{\frac{\nu_1}{2}-1} \left(1 + \frac{\nu_1}{\nu_2}x\right)^{-\frac{\nu_1+\nu_2}{2}}, & x > 0 \\ 0\,, & x \leq 0 \end{cases}$$

$$E(X) = \frac{\nu_2}{\nu_2 - 2} \quad , \quad \nu_2 > 2$$
$$D(X) = \frac{2{\nu_2}^2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 4)(\nu_2 - 2)^2} \quad , \quad \nu_2 > 4$$

16. **Studentova (t) raspodela**

$t : t_\nu\,, \quad \nu \in N$

$$t_\nu = \frac{X}{\sqrt{\frac{\chi^2_\nu}{\nu}}}\,, \quad \text{gde je} \quad X : \mathcal{N}(0,1)$$

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu\pi}} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} \quad , \quad -\infty < t < +\infty$$

$$E(X) = 0$$
$$D(X) = \frac{\nu}{\nu - 2} \quad , \quad \nu > 2$$

17. **Gumbelova raspodela**

$\alpha > 0\,, \quad \beta > 0$

$$f(x) = \alpha \exp\{-\alpha(x-\beta) - \exp(-\alpha(x-\beta))\} \quad , \quad -\infty < x < +\infty \quad ,$$

$Z = \alpha(X - \beta)$

$$f(z) = \exp\{-z - \exp\{-z\}\} \quad , \quad -\infty < z < +\infty$$

18. **Pareto raspodela**

$\alpha > 0\,, \quad \beta > 0$

$$f(x) = \begin{cases} \alpha\beta^{\alpha}x^{-(\alpha+1)}\,, & x > \beta \\ 0\,, & x \leq \beta \end{cases}$$

# Glava 10

# Funkcija generatrisa momenata

Funkcija generatrisa momenata slučajne promenljive $X$ je funkcija po $t$ u oznaci

$$M_X(t) = E(e^{tX})$$

tj.

$$M_X(t) = \sum_{j=1}^{\infty} e^{tx_j} f(x_j) \quad \text{u diskretnom slučaju}$$

i

$$M_X(t) = \int_{-\infty}^{+\infty} e^{tx} f(x)dx \quad \text{u apsolutno neprekidnom slučaju.}$$

**Primer 100.**

$$I_A = \begin{cases} 1, & x \in A \\ 0, & \text{inače} \end{cases} \quad , \quad P(I_A = 1) = p \quad , \quad 0 \leq p \leq 1$$

$$M_{I_A}(t) = E(e^{I_A t}) = pe^t + 1 - p \,. \triangle$$

**Primer 101.**

$$X : \mathcal{N}(0, 1)$$

$$M_X(t) = E(e^{tX}) = \int_{-\infty}^{+\infty} e^{tx} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = e^{\frac{t^2}{2}} \,. \triangle$$

**Primer 102.**

$$X : \mathcal{N}(m, \sigma^2)$$

$$\begin{aligned} M_X(t) &= E(e^{tx}) = \int_{-\infty}^{+\infty} e^{tx} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-m)^2}{2\sigma^2}} dx = \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{+\infty} \exp\left\{ -\left( \frac{x^2}{2\sigma^2} - \left( t + \frac{m}{\sigma^2} \right) x + \frac{m^2}{2\sigma^2} \right) \right\} dx = \\ &= \exp\left\{ \left( tm + \frac{\sigma^2 t^2}{2} \right) \right\} \,. \triangle \end{aligned}$$

Iz činjenice da matematičko očekivanje ne postoji uvek, tj. da ima slučajnih promenljivih koje nemaju matematičko očekivanje, jasno je da funkcija generatrisa momenata ne postoji za svaku slučajnu promenljivu. Funkcija generatrisa momenata se, s toga definiše na sledeći način.

DEFINICIJA 46. Pretpostavimo da postoji pozitivan broj $h$ takav da za $-h < t < h$ očekivanje $E\left(e^{tX}\right)$ postoji, tada funkciju $M_X(t) = E\left(e^{tX}\right)$ za $t \in (-h, h)$ zovemo *funkcija generatrisa momenata slučajne promenljive* $X$.

**Primer 103.** Neka slučajna promenljiva $X$ ima gustinu raspodele

$$f(x) = \begin{cases} \frac{6}{\pi^2 x^2}\,, & x = 1, 2, 3, \ldots \\ 0\,, & \text{inače} \end{cases} .$$

Ako bi postojala funkcija generatrisa momenata slučajne promenljive $X$, bilo bi:

$$M(t) = E(e^{tX}) = \sum_x e^{tx} f(x) = \sum_{x=1}^{\infty} \frac{6e^{tx}}{\pi^2 x^2}\,.$$

Dakle, treba ispitati konvergenciju reda čiji je opšti član $a_n = \frac{6e^{tn}}{\pi^2 n^2}$. Očigledno

$$\lim_{n\to\infty} \frac{a_{n+1}}{a_n} = e^t > 1 \quad \text{za} \quad t > 0\,.$$

To znači da za svako $t > 0$, $M(t)$ divergira tj. ne postoji pozitivan broj h takav da za $-h < t < h$ očekivanje $M(t)$ postoji.△

Činjenica da funkcija generatrisa momenata ne postoji uvek ne umanjuje značaj ove funkcije zbog široke klase raspodela za koju ona postoji.

Lako je pokazati da je Mak-Loranov razvoj ove funkcije:

$$M_X(t) = 1 + E(X)t + E(X^2)\frac{t^2}{2!} + \ldots + E(X^r)\frac{t^r}{r!} + \ldots \quad ,$$

što razjašnjava i poreklo njenog imena – funkcija generatrisa momenata, jer sledi da je

$$E(X^r) = \frac{d^r}{dt^r} M_X(t)|_{t=0} \quad , \quad r = 0, 1, 2, \ldots \quad .$$

Značaj funkcije generatrisa momenata objašnjava sledeća teorema koju navodimo bez dokaza.

**Teorema 10.0.1** *Pretpostavimo da su $X$ i $Y$ slučajne promenljive koje imaju funkcije generatrisa momenata $M_X(t)$ i $M_Y(t)$ redom. Tada $X$ i $Y$ imaju istu raspodelu verovatnoća ako i samo ako su $M_X(t)$ i $M_Y(t)$ identički jednake.*

Sledeće teoreme su operativnog značaja.

**Teorema 10.0.2** *Ako je $M_X(t)$ funkcija generatrisa momenata slučajne promenljive $X$ i $a \neq 0$ i $b$ su konstante, tada je funkcija generatrisa momenata slučajne promenljive $Y = aX + b$*

$$M_Y(t) = e^{bt} M_X(at).$$

**Dokaz.**Direktno sledi iz osobina eksponencijalne funkcije i definicije funkcije generatrisa momenata. □

**Teorema 10.0.3** *Funkcija generatrisa momenata konačne sume nezavisnih slučajnih promenljivih jednaka je proizvodu funkcija generatrisa sabiraka.*

**Dokaz.** Neka su $X_1, \ldots, X_n$ nezavisne slučajne promenljive. Funkcija generatrisa momenata slučajne promenljive $Y = \sum_{i=1}^{n} X_i$ je

$$M_Y(t) = E\left(e^{tY}\right) = E\left(\exp\left\{t\sum_{i=1}^{n} X_i\right\}\right) = E\left(\prod_{i=1}^{n} e^{tX_i}\right) = \prod_{i=1}^{n} E\left(e^{tX_i}\right) = \prod_{i=1}^{n} M_{X_i}(t)\,.\square$$

**Primer 104.** Neka su $X_1, \ldots, X_n$ nezavisne slučajne promenljive sa raspodelama $X_i : \mathcal{N}(m,\, \sigma^2)$, $i = 1, 2, \ldots, n$. Primenom funkcije generatrisa momenata pokazati da slučajna promenljiva $U = \sum_{i=1}^{n} a_i X_i$, $a_i$ su konstante, ima $\mathcal{N}\left(\sum_{i=1}^{n} a_i m_i\,,\, \sum_{i=1}^{n} a_i^2 \sigma_i^2\right)$ raspodelu.

S obzirom da $X_i$ ima normalnu raspodelu sa parametrima $m_i$ i $\sigma_i^2$, to je njena funkcija generatrisa momenata

$$M_{X_i}(t) = \exp\left(m_i t + \frac{\sigma_i^2 t^2}{2}\right)\,.$$

Prema tome slučajna promenljiva $a_i X_i$ ima funkciju generatrisa momenata

$$M_{a_i X_i}(t) = M_{X_i}(a_i t) = \exp\left(m_i a_i t + \frac{\sigma_i^2 a_i^2 t^2}{2}\right)\,.$$

Najzad, s obzirom na nezavisnost slučajnih promenljivih $X_i$, $i = 1, 2, \ldots, n$, funkcija generatrisa momenata njihove linearne kombinacije je

$$M_U(t) = \prod_{i=1}^{n} M_{X_i}(a_i t) = \exp\left(t\sum_{i=1}^{n} a_i m_i + \frac{t^2}{n}\sum_{i=1}^{n} a_i^2 \sigma_i^2\right)\,,$$

što je funkcija generatrisa momenata normalno raspodeljene slučajne promenljive sa parametrima

$$E(U) = \sum_{i=1}^{n} a_i m_i \qquad \text{i} \qquad D(U) = \sum_{i=1}^{n} a_i^2 \sigma_i^2\,.\,\triangle$$

Za dvodimenzionu slučajnu promenljivu $(X, Y)$ funkcija generatrisa momenata, ukoliko postoji, definiše se kao funkcija od dva argumenta:

$$M(t, s) = E\left(e^{tX+sY}\right)\,.$$

Naravno, sve navedene teoreme važe i kada su u pitanju višedimenzione slučajne promenljive, čime se ovde nećemo posebno baviti, ali ćemo ipak ilustrovati jednim dobro poznatim primerom.

**Primer 105.** Naći funkciju generatrisa momenata dvodimenzionog vektora normalno raspodeljenih slučajnih promenljivih

$$(X,Y) : \mathcal{N}\left(m_X, m_Y, \sigma_X^2, \sigma_Y^2, \rho\right) .$$

Iz definicije sledi da je

$$\begin{aligned} &M(t,s) = \\ &= \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} e^{tx+sy} e^{-\frac{1}{2(1-\rho^2)}\left(\frac{(x-m_X)^2}{\sigma_X^2} - 2\rho\frac{(x-m_X)(y-m_Y)}{\sigma_X\sigma_Y} + \frac{(y-m_Y)^2}{\sigma_Y^2}\right)} dxdy = \\ &= \exp\left\{m_X t + m_Y s + \frac{\sigma_X^2 t^2 + 2\rho\sigma_X\sigma_Y ts + \sigma_Y^2 s^2}{2}\right\} . \triangle \end{aligned}$$

**Primer 106.** Neka je $(X_1,Y_1),(X_2,Y_2),\ldots,(X_n,Y_n)$ prost uzorak obima n iz dvodimenzionalne normalne raspodele $\mathcal{N}(m_X, m_Y, \sigma_X^2, \sigma_Y^2, \rho)$. Naći zajedničku raspodelu dveju statistika

$$\overline{X} = \sum_{i=1}^{n} \frac{X_i}{n} \qquad \text{i} \qquad \overline{Y} = \sum_{i=1}^{n} \frac{Y_i}{n} .$$

S obzirom da svaka od statistika $\overline{X}$ i $\overline{Y}$ ima normalnu raspodelu i to

$$\overline{X} : \mathcal{N}(m_X, \frac{\sigma_X^2}{n}) \qquad \text{i} \qquad \overline{Y} : \mathcal{N}(m_Y, \frac{\sigma_Y^2}{n}) ,$$

to je funkcija generatrisa momenata dvodimenzione statistike $(\overline{X}, \overline{Y})$ data sa

$$M(t,s) = E\left(e^{t\overline{X}+s\overline{Y}}\right) = E\left(\exp\left\{\frac{1}{n}\sum_{i=1}^{n}(tX_i + sY_i)\right\}\right) .$$

Kako je uzorak prost, to je

$$M(t,s) = \prod_{i=1}^{n} E\left(e^{\frac{t}{n}X_i + \frac{s}{n}Y_i}\right) .$$

Na osnovu primera 105 sledi da je tražena funkcija

$$\begin{aligned} M(t,s) &= \prod_{i=1}^{n} \exp\left\{\frac{m_X t}{n} + \frac{m_Y s}{n} + \frac{\sigma_X^2(\frac{t}{n})^2 + 2\rho\sigma_X\sigma_Y\frac{t}{n}\frac{s}{n} + \sigma_Y^2(\frac{s}{n})^2}{2}\right\} = \\ &= \exp\left\{m_X t + m_Y s + \frac{\frac{\sigma_X^2}{n}t^2 + 2\rho\frac{\sigma_X\sigma_Y}{n}ts + \frac{\sigma_Y^2}{n}s^2}{2}\right\} \end{aligned}$$

što je funkcija generatrisa momenata dvodimenzione normalne raspodele $\mathcal{N}(m_X\,,\, m_Y\,,\, \frac{\sigma_X^2}{n}\,,\, \frac{\sigma_Y^2}{n}\,,\, \rho)$. $\triangle$

# Glava 11

# Tačkasto ocenjivanje parametara obeležja konačne populacije

Kod beskonačne populacije (koja teorijski dozvoljava izbor uzoraka beskonačnog obima) primenjuju se granične teoreme teorije verovatnoće. Medjutim, na konačnoj populaciji ova teorija nema u potpunosti opravdanja. U okviru ovog poglavlja ćemo se u kratkim crtama upoznati sa ocenjivanjem parametara obeležja zadatog na populaciji konačnog obima. Razmatraćemo ocene očekivanja i disperzije.

## 11.1 Ocene matematičkog očekivanja i disperzije

Podsetimo se da je sredina uzorka nepristrasna ocena matematičkog očekivanja bez obzira na raspodelu obeležja.

Neka je data populacija $\Omega$ koja je konačan skup od N elemenata. Posmatraćemo izbor uzorka obima $n$, $n \leq N$, bez vraćanja iz populacije $\Omega$. Taj uzorak je slučajna veličina – vektor $(X_1, \ldots, X_n)$. Svaki element populacije ima jednaku verovatnoću izbora u uzorak u momentu izvlačenja. Dakle, $\omega \in s$ u prvom izvlačenju sa verovatnoćom $\frac{1}{N}$, u drugom $\frac{1}{N-1}$, itd. S obzirom da je populacija konačna, obeležje $X$ je diskretnog tipa tj.

$$P\{X = x_j\} = \frac{N_j}{N} \quad , \quad j = 1, \ldots, k \,,$$

gde su $x_j$, $j = 1, \ldots, k$ sve moguće vrednosti obeležja $X$, a $N_j$ je ukupan broj elemenata populacije kod kojih je vrednost posmatranog obeležja $X$ baš $x_j$. Lako je pokazati da slučajne promenljive $X_i$, $1 \leq i \leq n$ imaju istu raspodelu kao i $X$ iako su medjusobno zavisne. Naime,

$$P\{X_i = x_j\} = P\{\omega : X_i(\omega) = x_j\} = P\{\omega_1^{(i)}, \ldots, \omega_{N_j}^{(i)}\} \,,$$

gde je sa $\omega_l^{(i)}$, $l = 1, 2, \ldots, N_j$ označen element populacije, $\omega_l^{(i)} \in \Omega$, kod koga je vrednost obeležja $X$ baš $x_j$, a izabran je u uzorak u $i$–tom izvlačenju. Tada traženu verovatnoću odredjujemo kao

$$P\{\omega_1^{(i)}, \ldots, \omega_{N_j}^{(i)}\} = \sum_{l=1}^{N_j} P(\omega_l^{(i)}) \,,$$

gde je $P(\omega_l^{(i)})$ verovatnoća da je u $i$–tom opitu izabran baš $\omega_l^{(i)}$, što znači da se u prethodnih $i-1$ opita nije realizovao, jer je uzorak bez vraćanja. Dakle,

$$P(\omega_l^{(i)}) = \frac{\binom{N-1}{1}}{\binom{N}{1}} \cdot \frac{\binom{N-2}{1}}{\binom{N-1}{1}} \cdot \cdots \cdot \frac{\binom{N-i}{1}}{\binom{N-i+1}{1}} \cdot \frac{\binom{1}{1}}{\binom{N-i}{1}} = \frac{1}{N} \quad , \quad \forall i = 1, 2, \ldots, n\,.$$

To znači da je

$$P\{X_i = x_j\} = \sum_{l=1}^{N_j} \frac{1}{N} = \frac{N_j}{N}$$

za svako $i = 1, 2, \ldots, n$, tj. da $X_i$ ima istu raspodelu kao i $X$. Kod uzorka sa vraćanjem je ova činjenica očigledna.

Iz činjenice o raspodeli slučajnih promenljivih $X_i$, $i = 1, 2, \ldots, n$, tj. jednakosti te raspodele sa raspodelom samog obeležja $X$ sledi da je $E(X_i) = E(X)$, pa je sredina uzorka (kao i kod prostog uzorka sa vraćanjem) jednaka

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i\,,$$

a ranije smo pokazali da je ona nepristrasna ocena matematičkog očekivanja obeležja $X$.

S obzirom da je posmatrana ocena nepristrasna, srednje kvadratno odstupanje, kao kriterijum valjanosti, daje

$$E(\overline{X}_n - EX)^2 = D(\overline{X}_n)\,.$$

Kako su slučajne promenljive $X_1, \ldots, X_n$ kod uzorka bez vraćanja zavisne, biće

$$\begin{aligned} D(\sum_{i=1}^{n} X_i) &= E\left(\sum_{i=1}^{n} X_i - nE(X)\right)^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} E(X_i - E(X))(X_j - E(X)) = \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} Cov(X_i, X_j) = nD(X_i) + n(n-1)Cov(X_1, X_2)\,, \end{aligned}$$

jer svi parovi $(X_i, X_j)$ imaju istu raspodelu kao i par $(X_1, X_2)$.

Može se pokazati da je $P\{(X_i, X_j) = (x_l, x_r)\} = \frac{1}{N(N-1)}$ za svako $i, j = 1, 2, \ldots, n$ i svako $l, r = 1, 2, \ldots, k$. To ukazuje da raspodela za $(X_i, X_j)$ ne zavisi od $i$, $j$ ili $n$ pa sledi

$$Cov(X_i, X_j) = Cov(X_1, X_2)\,.$$

Označimo ovu kovarijansu sa $c(N)$. Dakle,

$$D(\overline{X}_n) = \frac{1}{n^2} D(\sum_{i=1}^{n} X_i) = \frac{1}{n^2}(nD(X) + n(n-1)c(N)) = \frac{1}{n}(D(X) + (n-1)c(N)).$$

Veličinu $c(N)$ moguće je odrediti iz uslova da je $n = N$, tj. da uzorkom bez vraćanja iscrpimo celu populaciju. Tada je

$$D(\overline{X}_N) = \frac{1}{N}(D(X) + (N-1)c(N))\,.$$

Medjutim, tada je $\overline{X}_N$ tačna vrednost očekivanja $E(X)$, jer je

$$\overline{X}_N = \frac{1}{N}(X(\omega_1) + \ldots + X(\omega_N)) = E(X)\,,$$

pa je $D(\overline{X}_N) = 0$. Otuda $D(X) + (N-1)c(N) = 0$, odnosno $c(N) = -\frac{1}{N-1}D(X)$. Najzad

$$D(\overline{X}_N) = \frac{1}{n}(D(X) - \frac{n-1}{N-1}D(X)) = \frac{D(X)}{n}\left(1 - \frac{n-1}{N-1}\right) = \frac{D(X)}{n} \cdot \frac{N-n}{N-1}\,.$$

Podsetimo se da je greška iste ocene kod uzorka sa vraćanjem

$$E(\overline{X}_N - E(X))^2 = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i - E(X)\right)^2 = \frac{1}{n}D(X)\,,$$

što znači da je greška ocene kod uzorka bez vraćanja manja u odnosu na onu kod uzorka sa vraćanjem.

Faktor $\frac{N-n}{N-1} = 1 - \frac{n-1}{N-1}$, koji je kod populacije velikog obima $N$ i obima uzorka $n$ koji je mali u odnosu na obim populacije približno jednak $1 - \frac{n}{N}$, naziva se *korekcija zbog konačnosti populacije*.

Kada je reč o oceni disperzije obeležja $D(X)$, već smo pokazali da je kod uzorka sa vraćanjem

$$\widetilde{S}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

nepristrasna ocena disperzije obeležja. Medjutim, ako je obim uzorka veliki, količnik $\frac{n}{n-1}$ je približno jednak 1, pa se za ocenu disperzije može uzeti i uzoračka disperzija

$$\overline{S}_n{}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

za koju je $E(\overline{S}_n{}^2) = \frac{n-1}{n}D(X)$.

Razmotrimo pristrasnost uzoračke disperzije za slučaj uzorka bez vraćanja iz konačne populacije. Tada je

$$\begin{aligned} E(\overline{S}_n{}^2) &= E\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}_n{}^2\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i^2) - E(\overline{X}_n{}^2) = \\ &= E(X^2) - \frac{D(X)}{n} \cdot \frac{N-n}{N-1} - (EX)^2 = D(X)\left(1 - \frac{1}{n} \cdot \frac{N-n}{N-1}\right) = \\ &= D(X)\frac{(n-1)N}{n(N-1)} = \frac{n-1}{n}D(X)\frac{N}{N-1}\,. \end{aligned}$$

Kako je obim populacije $N$ najčešće veliki, to je $\frac{N}{N-1}$ približno 1 pa se kao nepristrasna ocena disperzije obeležja može ponovo uzeti $\widetilde{S}_n^2$ kod uzorka malog obima $n$ (recimo $n \leq 30$), odnosno $\overline{S}_n{}^2$ kod uzorka velikog obima.

## 11.2 Ocene matematičkog očekivanja i disperzije kod stratifikovanog uzorka

Neka je ponovo reč o konačnoj populaciji sa obeležjem $X$ čija je raspodela

$$X : \begin{pmatrix} x_1 & x_2 & \dots & x_k \\ \frac{M_1}{N} & \frac{M_2}{N} & \dots & \frac{M_k}{N} \end{pmatrix}, \quad M_1+M_2+\ldots+M_k = N\,, \quad 0 \le M_i \le N\,, \quad i = 1,2,\ldots,k\,.$$

Označimo sa $X^{(h)}$ obeležje $X$ posmatrano na $h$-tom stratumu za koji ćemo pretpostaviti da je obima $N_h$. Pretpostavimo, takodje, da je populacija podeljena na $L$ stratuma, $L < N$. Tada slučajna promenljiva $X^{(h)}$ ima sledeću raspodelu

$$X^{(h)} : \begin{pmatrix} x_1 & x_2 & \dots & x_k \\ \frac{M_1^{(h)}}{N_h} & \frac{M_2^{(h)}}{N_h} & \dots & \frac{M_k^{(h)}}{N_h} \end{pmatrix}, \quad \sum_{i=1}^{k} M_i^{(h)} = N_h\,, \quad 0 \le M_i^{(h)} \le N_h\,, \quad i = 1,2,\ldots,k\,,$$
$$h = 1,2,\ldots,L\,.$$

Njegovo očekivanje je

$$E\left(X^{(h)}\right) = \sum_{j=1}^{k} x_j \frac{M_j^{(h)}}{N_h}\,,$$

a očekivanje obeležja $X$ je

$$\begin{aligned} E(X) &= \sum_{j=1}^{k} x_j \frac{M_j}{N} = \frac{1}{N}\sum_{j=1}^{k} x_j \left(M_j^{(1)} + \ldots + M_j^{(L)}\right) = \\ &= \frac{1}{N}\left(\sum_{j=1}^{k} x_j M_j^{(1)} + \ldots + \sum_{j=1}^{k} x_j M_j^{(L)}\right) = \\ &= \frac{N_1}{N}\sum_{j=1}^{k} x_j \frac{M_j^{(1)}}{N_1} + \ldots + \frac{N_k}{N}\sum_{j=1}^{k} x_j \frac{M_j^{(L)}}{N_L} = \\ &= \frac{N_1}{N} E\left(X^{(1)}\right) + \ldots + \frac{N_k}{N} E\left(X^{(L)}\right) = \frac{1}{N}\sum_{i=1}^{L} N_i E\left(X^{(i)}\right) = \sum_{i=1}^{L} w_i E\left(X^{(i)}\right)\,, \end{aligned}$$

gde smo sa $w_i = \frac{N_i}{N}$ označili udeo $i$–tog stratuma u populaciji. Relacija $E(X) = \sum_{i=1}^{L} w_i E\left(X^{(i)}\right)$ predstavlja vezu izmedju matematičkog očekivanja obeležja i očekivanja po stratumima.

Razmotrimo sada problem ocenjivanja na osnovu uzorka. Iz ukupno $L$ stratuma od kojih $j$–ti ima tačno $N_j$ elemenata populacije, bira se po $n_j \le N_j$, $j = 1,2,\ldots,L$, elemenata u uzorak. Pretpostavka je da su izvlačenja iz različitih stratuma nezavisna.

Deo uzorka iz $h$–tog stratuma je $(X_1^{(h)},\ldots,X_{n_h}^{(h)})$, a pripadajuća uzoračka sredina i disperzija su

$$\overline{X}_{n_h} = \frac{1}{n_h}\sum_{i=1}^{n_h} X_i^{(h)}$$

$$\overline{S}^2_{n_h} = \frac{1}{n_h}\sum_{i=1}^{n_h} (X_i^{(h)} - \overline{X}_{n_h})^2\,,$$

odnosno popravljena disperzija uzorka

$$\widetilde{S}_n^2 = \frac{n_h}{n_h - 1} \bar{S}_{n_h}^2.$$

Pri tome je $\overline{X}_{n_h}$ nepristrasna ocena srednje vrednosti obeležja $X^{(h)}$ u $h$–tom stratumu.

Statistika

$$\widetilde{X} = \frac{1}{N} \sum_{i=1}^{L} N_i \overline{X}_{n_i} = \sum_{i=1}^{L} w_i \overline{X}_{n_i}$$

nije sredina uzorka, ali je ipak nepristrasna ocena metemetičkog očekivanja obeležja $X$ bez obzira da li je uzorak sa vraćanjem ili bez vraćanja. Zaista, kako je

$$E\left(\overline{X}_{n_h}\right) = EX^{(h)}, \quad \text{to je} \quad E(\widetilde{X}) = \frac{1}{N} \sum_{i=1}^{L} N_i E(\overline{X}_{n_i}) = \frac{1}{N} \sum_{i=1}^{L} N_i E(X^{(i)}) = E(X).$$

Srednjekvadratna greška ove ocene je za uzorak sa vraćanjem

$$E(\widetilde{X} - E(X))^2 = D(\widetilde{X}) = \sum_{i=1}^{L} w_i^2 \frac{D(\overline{X}_{n_i})}{n_i},$$

a za uzorak bez vraćanja

$$E(\widetilde{X} - E(X))^2 = D(\widetilde{X}) = \sum_{i=1}^{L} w_i^2 D(\overline{X}_{n_i}) \approx \sum_{i=1}^{L} w_i^2 \frac{D(X)}{n_i} \left(1 - \frac{n_i}{N_i}\right).$$

## 11.3 Ocena parametra p binomne raspodele

U ovom poglavlju zadatak će nam biti da na osnovu uzorka obima $n$ ocenimo verovatnoću $p$ da pojedini element populacije $\omega \in \Omega$ ima svojstvo $A$. U tu svrhu uočićemo indikator dogadjaja $A$

$$I_A(\omega) = \begin{cases} 1, & \omega \in A \\ 0, & \text{inače} \end{cases}$$

Tada je $E(I_A) = p$, pa se ocenjivanje parametra $p$ svodi na ocenjivanje matematičkog očekivanja indikatorskog obeležja. Ako pretpostavimo da u uzorku obima $n$ ima $n_1$ elemenata sa svojstvom $A$, tada je

$$\overline{I}_{A,n} = \frac{1}{n} n_1 = \frac{n_1}{n}$$

relativna učestanost svojstva $A$ u uzorku i ona predstavlja nepristrasnu ocenu parametra $p$. Dakle, $\hat{p} = \overline{I}_{A,n}$. Srednjekvadratna greška ove ocene je

a) kod uzorka sa vraćanjem

$$E(\overline{I}_{A,n} - p)^2 = D(\overline{I}_{A,n}) = \frac{p(1-p)}{n},$$

b) kod uzorka bez vraćanja

$$E(\overline{I}_{A,n} - p)^2 = D(\overline{I}_{A,n}) = \frac{D(I_A)}{n} \cdot \frac{N-n}{N-1} = \frac{p(1-p)}{n} \cdot \frac{N-n}{N-1},$$

što je za veliki obim populacije $N$ približno $\frac{p(1-p)}{n}\left(1-\frac{n}{N}\right)$.

Jedan od ranije navedenih kriterijuma valjanosti ocene bio je količnik $\frac{(\hat{\theta}-\theta)^2}{\theta}$, gde je $\hat{\theta}$ statistika kojom ocenjujemo parametar $\theta$. Iskoristimo ovaj kriterijum za odredjivanje obima uzorka koji bi obezbedio da se postigne unapred zadata tačnost ocene parametra $p$. Dakle, za unapred zadato $\varepsilon > 0$ i $0 < \alpha < 1$ tražimo $n$ za koji važi

$$P\{Q_k \leq \varepsilon\} = 1-\alpha\,, \quad \text{za koji važi,} \quad Q_k = \sqrt{\frac{1}{k}\sum_{i=1}^{k}\frac{(\hat{\theta}_i-\theta_i)^2}{\theta_i}}\,.$$

**Primer 107.** Neka je obeležje $X$ diskretnog tipa sa raspodelom $P\{X = x_i\} = p_i$, $i = 1,\ldots,k$. Zadatak je oceniti parametre $p_i$ ove raspodele na osnovu prostog slučajnog uzorka.

Ako sa $n_i$ označimo broj pojavljivanja vrednosti $x_i$ u uzorku obima $n$, tada je $f_i = \frac{n_i}{n}$ relativna učestanost vrednosti $x_i$ u uzorku. Za ocenu verovatnoće $p_i$, kao što smo pokazali uzima se $\hat{p}_i = f_i$. Za meru odstupanja te ocene uzima se

$$\frac{(f_i-p_i)^2}{p_i}$$

odnosno, zajednička prosečna mera za ceo uzorak biće

$$Q_k^2 = \frac{1}{k}\sum_{i=1}^{n}\frac{(f_i-p_i)^2}{p_i} = \frac{1}{k}\sum_{i=1}^{k}\frac{(n_i-np_i)^2}{n^2p_i} = \frac{1}{n}\frac{1}{k}\sum_{i=1}^{k}\frac{(n_i-np_i)^2}{np_i}\,.$$

Kao što je poznato, statistika

$$\sum_{i=1}^{k}\frac{(n_i-np_i)^2}{np_i}$$

ima približno $\chi^2$–raspodelu sa $k$ stepeni slobode, što znači

$$knQ_k^2 \sim \chi^2_{k-1}\,, \quad \text{kada} n \longrightarrow \infty.$$

Dakle,

$$P\{Q_k \leq \varepsilon\} = 1-\alpha$$
$$P\{knQ_k^2 \leq kn\varepsilon^2\} = 1-\alpha$$
$$P\{\chi^2_{k-1} \leq kn\varepsilon^2\} = 1-\alpha\,.$$

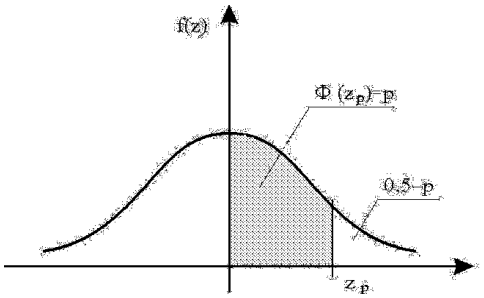
Sledi da je $kn\varepsilon^2$ kvantil reda $1-\alpha$ slučajne promenljive sa $\chi^2_{k-1}$ raspodelom odakle se dobija veličina $kn\varepsilon^2 = \chi^2_{k-1,1-\alpha}$ (što se čita iz tablice za $\chi^2$ raspodelu) i

$$n = \frac{\chi^2_{k-1,1-\alpha}}{k\varepsilon^2}\,.$$

$\triangle$

# Statističke tablice

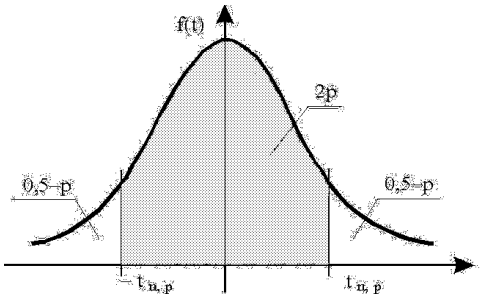
## 1. Normalna raspodela




$\Phi(z_p) = p$

| $z$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.0 | .0000 | .0040 | .0080 | .0120 | .0160 | .0199 | .0239 | .0279 | .0319 | .0359 |
| 0.1 | .0398 | .0438 | .0478 | .0517 | .0557 | .0596 | .0636 | .0675 | .0714 | .0754 |
| 0.2 | .0793 | .0832 | .0871 | .0910 | .0948 | .0987 | .1026 | .1064 | .1103 | .1141 |
| 0.3 | .1179 | .1217 | .1255 | .1293 | .1331 | .1368 | .1406 | .1443 | .1480 | .1517 |
| 0.4 | .1554 | .1591 | .1628 | .1664 | .1700 | .1736 | .1772 | .1808 | .1844 | .1879 |
| 0.5 | .1915 | .1950 | .1985 | .2019 | .2054 | .2088 | .2123 | .2157 | .2190 | .2224 |
| 0.6 | .2258 | .2291 | .2324 | .2357 | .2389 | .2422 | .2454 | .2486 | .2518 | .2549 |
| 0.7 | .2580 | .2612 | .2642 | .2673 | .2704 | .2734 | .2764 | .2794 | .2823 | .2852 |
| 0.8 | .2881 | .2910 | .2939 | .2967 | .2996 | .3023 | .3051 | .3078 | .3106 | .3133 |
| 0.9 | .3159 | .3186 | .3212 | .3238 | .3264 | .3289 | .3315 | .3340 | .3365 | .3389 |
| 1.0 | .3413 | .3438 | .3461 | .3485 | .3508 | .3531 | .3554 | .3577 | .3599 | .3621 |
| 1.1 | .3643 | .3665 | .3686 | .3708 | .3729 | .3749 | .3770 | .3790 | .3810 | .3830 |
| 1.2 | .3849 | .3869 | .3888 | .3907 | .3925 | .3944 | .3962 | .3980 | .3997 | .4015 |
| 1.3 | .4032 | .4049 | .4066 | .4082 | .4099 | .4115 | .4131 | .4147 | .4162 | .4177 |
| 1.4 | .4192 | .4207 | .4222 | .4236 | .4251 | .4265 | .4279 | .4292 | .4306 | .4319 |
| 1.5 | .4332 | .4345 | .4357 | .4370 | .4382 | .4394 | .4406 | .4418 | .4429 | .4441 |
| 1.6 | .4452 | .4463 | .4474 | .4484 | .4495 | .4505 | .4515 | .4525 | .4535 | .4545 |
| 1.7 | .4554 | .4564 | .4573 | .4582 | .4591 | .4599 | .4608 | .4616 | .4625 | .4633 |
| 1.8 | .4641 | .4649 | .4656 | .4664 | .4671 | .4678 | .4686 | .4693 | .4699 | .4706 |
| 1.9 | .4713 | .4719 | .4726 | .4732 | .4738 | .4744 | .4750 | .4756 | .4761 | .4767 |
| 2.0 | .4772 | .4778 | .4783 | .4788 | .4793 | .4798 | .4803 | .4808 | .4812 | .4817 |
| 2.1 | .4821 | .4826 | .4830 | .4834 | .4838 | .4842 | .4846 | .4850 | .4854 | .4857 |
| 2.2 | .4861 | .4864 | .4868 | .4871 | .4875 | .4878 | .4881 | .4884 | .4887 | .4890 |
| 2.3 | .4893 | .4896 | .4898 | .4901 | .4904 | .4906 | .4909 | .4911 | .4913 | .4916 |
| 2.4 | .4918 | .4920 | .4922 | .4925 | .4927 | .4929 | .4931 | .4932 | .4934 | .4936 |
| 2.5 | .4938 | .4940 | .4941 | .4943 | .4945 | .4946 | .4948 | .4949 | .4951 | .4952 |
| 2.6 | .4953 | .4955 | .4956 | .4957 | .4959 | .4960 | .4961 | .4962 | .4963 | .4964 |
| 2.7 | .4965 | .4966 | .4967 | .4968 | .4969 | .4970 | .4971 | .4972 | .4973 | .4974 |
| 2.8 | .4974 | .4975 | .4976 | .4977 | .4977 | .4978 | .4979 | .4979 | .4980 | .4981 |
| 2.9 | .4981 | .4982 | .4982 | .4983 | .4984 | .4984 | .4985 | .4985 | .4986 | .4986 |
| 3.0 | .4987 | .4987 | .4987 | .4988 | .4988 | .4989 | .4989 | .4989 | .4990 | .4990 |
| 3.1 | .4990 | .4991 | .4991 | .4991 | .4992 | .4992 | .4992 | .4992 | .4993 | .4993 |
| 3.2 | .4993 | .4993 | .4994 | .4994 | .4994 | .4994 | .4994 | .4995 | .4995 | .4995 |
| 3.3 | .4995 | .4995 | .4995 | .4996 | .4996 | .4996 | .4996 | .4996 | .4996 | .4997 |
| 3.4 | .4997 | .4997 | .4997 | .4997 | .4997 | .4997 | .4997 | .4997 | .4997 | .4998 |
| 3.5 | .4998 | .4998 | .4998 | .4998 | .4998 | .4998 | .4998 | .4998 | .4998 | .4998 |
| 3.6 | .4998 | .4998 | .4999 | .4999 | .4999 | .4999 | .4999 | .4999 | .4999 | .4999 |
| 3.7 | .4998 | .4998 | .4999 | .4999 | .4999 | .4999 | .4999 | .4999 | .4999 | .4999 |
| 3.8 | .4998 | .4998 | .4999 | .4999 | .4999 | .4999 | .4999 | .4999 | .4999 | .4999 |
| 3.9 | .5000 | .5000 | .5000 | .5000 | .5000 | .5000 | .5000 | .5000 | .5000 | .5000 |

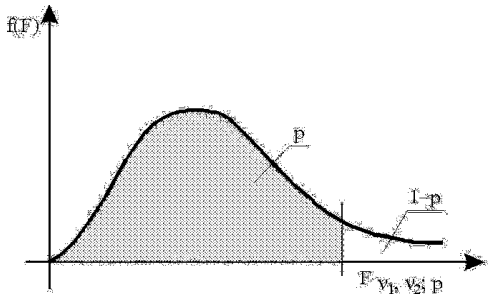
## 2. Studentova raspodela




$P\{|t_n| < t_{n,p}\} = 2p$

| $n \setminus p$ | 0.100 | 0.200 | 0.300 | 0.400 | 0.450 | 0.475 | 0.490 | 0.495 |
|---|---|---|---|---|---|---|---|---|
| 1 | .325 | .727 | 1.376 | 3.078 | 6.314 | 12.706 | 31.821 | 63.657 |
| 2 | .289 | .617 | 1.061 | 1.886 | 2.920 | 4.303 | 6.965 | 9.925 |
| 3 | .277 | .584 | .978 | 1.638 | 2.353 | 3.182 | 4.541 | 5.841 |
| 4 | .271 | .569 | .941 | 1.533 | 2.132 | 2.776 | 3.747 | 4.604 |
| 5 | .267 | .559 | .920 | 1.476 | 2.015 | 2.571 | 3.365 | 4.032 |
| 6 | .265 | .553 | .906 | 1.440 | 1.943 | 2.447 | 3.143 | 3.707 |
| 7 | .263 | .549 | .896 | 1.415 | 1.895 | 2.365 | 2.998 | 3.499 |
| 8 | .262 | .546 | .889 | 1.397 | 1.860 | 2.306 | 2.896 | 3.355 |
| 9 | .261 | .543 | .883 | 1.383 | 1.833 | 2.262 | 2.821 | 3.250 |
| 10 | .260 | .542 | .879 | 1.372 | 1.812 | 2.228 | 2.764 | 3.169 |
| 11 | .260 | .540 | .876 | 1.363 | 1.796 | 2.201 | 2.718 | 3.106 |
| 12 | .259 | .539 | .873 | 1.356 | 1.782 | 2.179 | 2.681 | 3.055 |
| 13 | .259 | .538 | .870 | 1.350 | 1.771 | 2.160 | 2.650 | 3.012 |
| 14 | .258 | .537 | .868 | 1.345 | 1.761 | 2.145 | 2.624 | 2.977 |
| 15 | .258 | .536 | .866 | 1.341 | 1.753 | 2.131 | 2.602 | 2.947 |
| 16 | .258 | .535 | .865 | 1.337 | 1.746 | 2.120 | 2.583 | 2.921 |
| 17 | .257 | .534 | .863 | 1.133 | 1.740 | 2.110 | 2.567 | 2.898 |
| 18 | .257 | .534 | .862 | 1.330 | 1.734 | 2.101 | 2.552 | 2.878 |
| 19 | .257 | .533 | .861 | 1.328 | 1.729 | 2.093 | 2.539 | 2.861 |
| 20 | .257 | .533 | .860 | 1.325 | 1.725 | 2.086 | 2.528 | 2.845 |
| 21 | .257 | .532 | .859 | 1.323 | 1.721 | 2.080 | 2.518 | 2.831 |
| 22 | .256 | .532 | .858 | 1.321 | 1.717 | 2.074 | 2.508 | 2.819 |
| 23 | .256 | .532 | .858 | 1.319 | 1.714 | 2.069 | 2.500 | 2.807 |
| 24 | .256 | .531 | .857 | 1.318 | 1.711 | 2.064 | 2.492 | 2.797 |
| 25 | .256 | .531 | .856 | 1.316 | 1.708 | 2.060 | 2.485 | 2.787 |
| 26 | .256 | .531 | .856 | 1.315 | 1.706 | 2.056 | 2.479 | 2.779 |
| 27 | .256 | .531 | .855 | 1.314 | 1.703 | 2.052 | 2.473 | 2.771 |
| 28 | .256 | .530 | .855 | 1.313 | 1.701 | 2.048 | 2.467 | 2.763 |
| 29 | .256 | .530 | .854 | 1.311 | 1.699 | 2.045 | 2.045 | 2.462 |
| 30 | .256 | .530 | .854 | 1.310 | 1.697 | 2.042 | 2.457 | 2.750 |
| 40 | .255 | .529 | .851 | 1.303 | 1.684 | 2.021 | 2.423 | 2.704 |
| 60 | .254 | .527 | .848 | 1.296 | 1.671 | 2.000 | 2.390 | 2.660 |
| 120 | .254 | .526 | .845 | 1.289 | 1.658 | 1.980 | 2.358 | 2.617 |
| $\infty$ | .253 | .524 | .842 | 1.282 | 1.645 | 1.960 | 2.326 | 2.576 |

### 3. Fišerova raspodela




$P\{F_{\nu_1,\nu_2} < F_{\nu_1,\nu_2;\ p}\} = p$

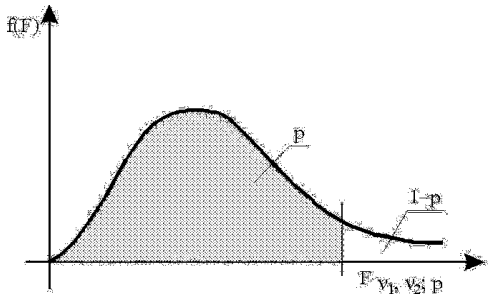
### 3. a) $p = 0{,}990$

| $\nu_2 \setminus \nu_1$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 4052 | 5000 | 5403 | 5625 | 5764 | 5859 | 5928 | 5981 | 6023 | 6056 |
| 2 | 98.5 | 99.0 | 99.2 | 99.2 | 99.3 | 99.3 | 99.4 | 99.4 | 99.4 | 99.4 |
| 3 | 34.1 | 30.8 | 29.5 | 28.7 | 28.2 | 27.9 | 27.7 | 27.5 | 27.3 | 27.2 |
| 4 | 21.2 | 18.0 | 16.7 | 16.0 | 15.5 | 15.2 | 15.0 | 14.8 | 14.7 | 14.5 |
| 5 | 16.3 | 13.3 | 12.1 | 11.4 | 11.0 | 10.7 | 10.5 | 10.3 | 10.2 | 10.1 |
| 6 | 13.7 | 10.9 | 9.78 | 9.15 | 8.75 | 8.47 | 8.26 | 8.10 | 7.98 | 7.87 |
| 7 | 12.2 | 9.55 | 8.45 | 7.85 | 7.46 | 7.19 | 6.99 | 6.84 | 6.72 | 6.62 |
| 8 | 11.3 | 8.65 | 7.59 | 7.01 | 6.63 | 6.37 | 6.18 | 6.03 | 5.91 | 5.81 |
| 9 | 10.6 | 8.02 | 6.99 | 6.42 | 6.06 | 5.80 | 5.61 | 5.47 | 5.35 | 5.26 |
| 10 | 10.0 | 7.56 | 6.55 | 5.99 | 5.64 | 5.39 | 5.20 | 5.06 | 4.94 | 4.85 |
| 11 | 9.65 | 7.21 | 6.22 | 5.67 | 5.32 | 5.07 | 4.89 | 4.74 | 4.63 | 4.54 |
| 12 | 9.33 | 6.93 | 5.95 | 5.41 | 5.06 | 4.82 | 4.64 | 4.50 | 4.39 | 4.30 |
| 13 | 9.07 | 6.70 | 5.74 | 5.21 | 4.86 | 4.62 | 4.44 | 4.30 | 4.19 | 4.10 |
| 14 | 8.86 | 6.51 | 5.56 | 5.04 | 4.70 | 4.46 | 4.28 | 4.14 | 4.03 | 3.94 |
| 15 | 8.68 | 6.36 | 5.42 | 4.89 | 4.56 | 4.32 | 4.14 | 4.00 | 3.89 | 3.80 |
| 16 | 8.53 | 6.23 | 5.29 | 4.77 | 4.44 | 4.20 | 4.03 | 3.89 | 3.78 | 3.69 |
| 17 | 8.40 | 6.11 | 5.19 | 4.67 | 4.34 | 4.10 | 3.93 | 3.79 | 3.68 | 3.59 |
| 18 | 8.29 | 6.01 | 5.09 | 4.58 | 4.25 | 4.01 | 3.84 | 3.71 | 3.60 | 3.51 |
| 19 | 8.18 | 5.93 | 5.01 | 4.50 | 4.17 | 3.94 | 3.77 | 3.63 | 3.52 | 3.43 |
| 20 | 8.10 | 5.85 | 4.94 | 4.43 | 4.10 | 3.87 | 3.70 | 3.56 | 3.46 | 3.37 |
| 21 | 8.02 | 5.78 | 4.87 | 4.37 | 4.04 | 3.81 | 3.64 | 3.51 | 3.40 | 3.31 |
| 22 | 7.95 | 5.72 | 4.82 | 4.31 | 3.99 | 3.76 | 3.59 | 3.45 | 3.35 | 3.26 |
| 23 | 7.88 | 5.66 | 4.76 | 4.26 | 3.94 | 3.71 | 3.54 | 3.41 | 3.30 | 3.21 |
| 24 | 7.82 | 5.61 | 4.72 | 4.22 | 3.90 | 3.67 | 3.50 | 3.36 | 3.26 | 3.17 |
| 25 | 7.77 | 5.57 | 4.68 | 4.18 | 3.86 | 3.63 | 3.46 | 3.32 | 3.22 | 3.13 |
| 26 | 7.72 | 5.53 | 4.64 | 4.14 | 3.82 | 3.59 | 3.42 | 3.29 | 3.18 | 3.09 |
| 27 | 7.68 | 5.49 | 4.60 | 4.11 | 3.78 | 3.56 | 3.39 | 3.26 | 3.15 | 3.06 |
| 28 | 7.64 | 5.45 | 4.57 | 4.07 | 3.75 | 3.53 | 3.36 | 3.28 | 3.12 | 3.03 |
| 29 | 7.60 | 5.42 | 4.54 | 4.04 | 3.73 | 3.50 | 3.33 | 3.20 | 3.09 | 3.00 |
| 30 | 7.56 | 5.39 | 4.51 | 4.02 | 3.70 | 3.47 | 3.30 | 3.17 | 3.07 | 2.98 |
| 40 | 7.31 | 5.18 | 4.31 | 3.83 | 3.51 | 3.29 | 3.12 | 2.99 | 2.89 | 2.80 |
| 60 | 7.08 | 4.98 | 4.13 | 3.65 | 3.34 | 3.12 | 2.95 | 2.82 | 2.72 | 2.63 |
| 120 | 6.85 | 4.79 | 3.95 | 3.48 | 3.17 | 2.96 | 2.79 | 2.66 | 2.56 | 2.47 |
| $\infty$ | 6.63 | 4.61 | 3.78 | 3.32 | 3.02 | 2.80 | 2.64 | 2.51 | 2.41 | 2.32 |

**nastavak 3. a)**

| 12 | 15 | 20 | 24 | 30 | 40 | 60 | 120 | $\nu_1/\nu_2$ |
|---|---|---|---|---|---|---|---|---|
| 6106 | 6157 | 6209 | 6235 | 6261 | 6287 | 6313 | 6339 | 1 |
| 99.4 | 99.4 | 99.4 | 99.5 | 99.5 | 99.5 | 99.5 | 99.5 | 2 |
| 27.1 | 26.9 | 26.7 | 26.6 | 26.5 | 26.4 | 26.3 | 26.2 | 3 |
| 14.4 | 14.2 | 14.0 | 13.9 | 13.8 | 13.7 | 13.7 | 13.6 | 4 |
| 9.89 | 9.72 | 9.55 | 9.47 | 9.38 | 9.29 | 9.20 | 9.11 | 5 |
| 7.72 | 7.56 | 7.40 | 4.31 | 7.23 | 7.14 | 7.06 | 6.97 | 6 |
| 6.47 | 6.31 | 6.16 | 6.07 | 5.99 | 5.91 | 5.82 | 5.74 | 7 |
| 5.67 | 5.52 | 5.36 | 5.28 | 5.20 | 5.12 | 5.03 | 4.95 | 8 |
| 5.11 | 4.96 | 4.81 | 4.73 | 4.65 | 4.57 | 4.48 | 4.40 | 9 |
| 4.71 | 4.56 | 4.41 | 4.33 | 4.25 | 4.17 | 4.08 | 4.00 | 10 |
| 4.40 | 4.25 | 4.10 | 4.02 | 3.94 | 3.86 | 3.78 | 3.69 | 11 |
| 4.16 | 4.01 | 3.86 | 3.78 | 3.70 | 3.62 | 3.54 | 3.45 | 12 |
| 3.96 | 3.82 | 3.66 | 3.59 | 3.51 | 3.43 | 3.34 | 3.25 | 13 |
| 3.80 | 3.66 | 3.51 | 3.43 | 3.35 | 3.27 | 3.18 | 3.09 | 14 |
| 3.67 | 3.52 | 3.37 | 3.29 | 3.21 | 3.13 | 3.05 | 2.96 | 15 |
| 3.55 | 3.41 | 3.26 | 3.18 | 3.10 | 3.02 | 2.93 | 2.84 | 16 |
| 3.46 | 3.31 | 3.16 | 3.08 | 3.00 | 2.92 | 2.83 | 2.75 | 17 |
| 3.37 | 3.23 | 3.08 | 3.00 | 2.92 | 2.84 | 2.75 | 2.66 | 18 |
| 3.30 | 3.15 | 3.00 | 2.92 | 2.84 | 2.76 | 2.67 | 2.58 | 19 |
| 3.23 | 3.09 | 2.94 | 2.86 | 2.78 | 2.69 | 2.61 | 2.52 | 20 |
| 3.17 | 3.03 | 2.88 | 2.80 | 2.72 | 2.64 | 2.55 | 2.46 | 21 |
| 3.12 | 2.98 | 2.83 | 2.75 | 2.67 | 2.58 | 2.50 | 2.40 | 22 |
| 3.07 | 2.93 | 2.78 | 2.70 | 2.62 | 2.54 | 2.45 | 2.35 | 23 |
| 3.03 | 2.89 | 2.74 | 2.66 | 2.58 | 2.49 | 2.40 | 2.31 | 24 |
| 2.99 | 2.85 | 2.70 | 2.62 | 2.54 | 2.45 | 2.36 | 2.27 | 25 |
| 2.96 | 2.82 | 2.66 | 2.58 | 2.50 | 2.42 | 2.33 | 2.23 | 26 |
| 2.93 | 2.78 | 2.63 | 2.55 | 2.47 | 2.38 | 2.29 | 2.20 | 27 |
| 2.90 | 2.75 | 2.60 | 2.52 | 2.44 | 2.35 | 2.26 | 2.17 | 28 |
| 2.87 | 2.73 | 2.57 | 2.49 | 2.41 | 2.33 | 2.23 | 2.14 | 29 |
| 2.84 | 2.70 | 2.55 | 2.47 | 2.39 | 2.30 | 2.21 | 2.11 | 30 |
| 2.66 | 2.52 | 2.37 | 2.29 | 2.20 | 2.11 | 2.02 | 1.92 | 40 |
| 2.50 | 2.35 | 2.20 | 2.12 | 2.03 | 1.94 | 1.84 | 1.73 | 60 |
| 2.34 | 2.19 | 2.03 | 1.95 | 1.86 | 1.76 | 1.66 | 1.53 | 120 |
| 2.18 | 2.04 | 1.88 | 1.79 | 1.70 | 1.59 | 1.47 | 1.32 | $\infty$ |

**Fišerova raspodela**




$P\{F_{\nu_1,\nu_2} < F_{\nu_1,\nu_2;\ p}\} = p$

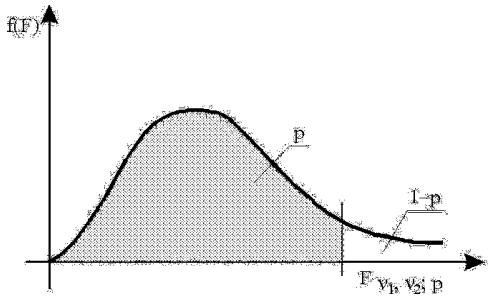
**3. b)** $p = 0{,}975$

| $\nu_2 \setminus \nu_1$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 648 | 799 | 864 | 900 | 922 | 937 | 948 | 957 | 963 | 969 |
| 2 | 38.5 | 39.0 | 39.2 | 39.2 | 39.3 | 39.3 | 39.4 | 39.4 | 39.4 | 39.4 |
| 3 | 17.4 | 16.0 | 15.4 | 15.1 | 14.9 | 14.7 | 14.6 | 14.5 | 14.5 | 14.5 |
| 4 | 12.2 | 10.6 | 9.98 | 9.60 | 9.36 | 9.20 | 9.07 | 8.98 | 8.90 | 8.84 |
| 5 | 10.0 | 8.43 | 7.76 | 7.39 | 7.15 | 6.98 | 6.85 | 6.76 | 6.68 | 6.62 |
| 6 | 8.81 | 7.26 | 6.60 | 6.23 | 5.99 | 5.82 | 5.70 | 5.60 | 5.52 | 5.46 |
| 7 | 8.07 | 6.54 | 5.89 | 5.52 | 5.29 | 5.12 | 4.99 | 4.90 | 4.82 | 4.76 |
| 8 | 7.57 | 6.06 | 5.42 | 5.05 | 4.82 | 4.65 | 4.53 | 4.43 | 4.36 | 4.30 |
| 9 | 7.21 | 5.71 | 5.08 | 4.72 | 4.48 | 4.32 | 4.20 | 4.10 | 4.03 | 3.96 |
| 10 | 6.94 | 5.46 | 4.83 | 4.47 | 4.24 | 4.07 | 3.95 | 3.85 | 3.78 | 3.72 |
| 11 | 6.72 | 5.26 | 4.63 | 4.28 | 4.04 | 3.88 | 3.76 | 3.66 | 3.59 | 3.53 |
| 12 | 6.55 | 5.10 | 4.47 | 4.12 | 3.89 | 3.73 | 3.61 | 3.51 | 3.44 | 3.37 |
| 13 | 6.41 | 4.97 | 4.35 | 4.00 | 3.77 | 3.60 | 3.48 | 3.39 | 3.31 | 3.25 |
| 14 | 6.30 | 4.86 | 4.24 | 3.89 | 3.66 | 3.50 | 3.38 | 3.29 | 3.21 | 3.15 |
| 15 | 6.20 | 4.77 | 4.15 | 3.80 | 3.58 | 3.41 | 3.29 | 3.20 | 3.12 | 3.06 |
| 16 | 6.12 | 4.69 | 4.08 | 3.73 | 3.50 | 3.34 | 3.22 | 3.12 | 3.05 | 2.99 |
| 17 | 6.04 | 4.62 | 4.01 | 3.66 | 3.44 | 3.28 | 3.16 | 3.06 | 2.98 | 2.92 |
| 18 | 5.98 | 4.56 | 3.95 | 3.61 | 3.38 | 3.22 | 3.10 | 3.01 | 2.93 | 2.87 |
| 19 | 5.92 | 4.51 | 3.90 | 3.56 | 3.33 | 3.17 | 3.05 | 2.96 | 2.88 | 2.82 |
| 20 | 5.87 | 4.46 | 3.86 | 3.51 | 3.29 | 3.13 | 3.01 | 2.91 | 2.84 | 2.77 |
| 21 | 5.83 | 4.42 | 3.82 | 3.48 | 3.25 | 3.09 | 2.97 | 2.87 | 2.80 | 2.73 |
| 22 | 5.79 | 4.38 | 3.78 | 3.44 | 3.22 | 3.05 | 2.93 | 2.84 | 2.76 | 2.70 |
| 23 | 5.75 | 4.35 | 3.75 | 3.41 | 3.18 | 3.02 | 2.90 | 2.81 | 2.73 | 2.67 |
| 24 | 5.72 | 4.32 | 3.72 | 3.38 | 3.15 | 2.99 | 2.87 | 2.78 | 2.70 | 2.64 |
| 25 | 5.69 | 4.29 | 3.69 | 3.35 | 3.13 | 2.97 | 2.85 | 2.75 | 2.68 | 2.61 |
| 26 | 5.66 | 4.27 | 3.67 | 3.33 | 3.10 | 2.94 | 2.82 | 2.73 | 2.65 | 2.59 |
| 27 | 5.63 | 4.24 | 3.65 | 3.31 | 3.08 | 2.92 | 2.80 | 2.71 | 2.63 | 2.57 |
| 28 | 5.61 | 4.22 | 3.63 | 3.29 | 3.06 | 2.90 | 2.78 | 2.69 | 2.61 | 2.55 |
| 29 | 5.59 | 4.20 | 3.61 | 3.27 | 3.04 | 2.88 | 2.76 | 2.67 | 2.59 | 2.53 |
| 30 | 5.57 | 4.18 | 3.59 | 3.25 | 3.03 | 2.87 | 2.75 | 2.65 | 2.57 | 2.51 |
| 40 | 5.42 | 4.05 | 3.46 | 3.13 | 2.90 | 2.74 | 2.62 | 2.53 | 2.45 | 2.39 |
| 60 | 5.29 | 3.93 | 3.34 | 3.01 | 2.79 | 2.63 | 2.51 | 2.41 | 2.33 | 2.27 |
| 120 | 5.15 | 3.80 | 3.23 | 2.89 | 2.67 | 2.52 | 2.39 | 2.30 | 2.22 | 2.16 |
| $\infty$ | 5.02 | 3.69 | 3.12 | 2.79 | 2.57 | 2.41 | 2.29 | 2.19 | 2.11 | 2.05 |

**nastavak 3. b)**

| 12 | 15 | 20 | 24 | 30 | 40 | 60 | 120 | $\nu_1/\nu_2$ |
|---|---|---|---|---|---|---|---|---|
| 977 | 985 | 993 | 997 | 1001 | 1006 | 1010 | 1014 | 1 |
| 39.4 | 39.4 | 39.4 | 39.5 | 39.5 | 39.5 | 39.5 | 39.5 | 2 |
| 14.3 | 14.3 | 14.2 | 14.1 | 14.1 | 14.0 | 14.0 | 13.9 | 3 |
| 8.75 | 8.65 | 8.56 | 8.51 | 8.46 | 8.41 | 8.36 | 8.31 | 4 |
| 6.52 | 6.43 | 6.33 | 6.28 | 6.23 | 6.18 | 6.12 | 6.07 | 5 |
| 5.37 | 5.27 | 5.17 | 5.12 | 5.07 | 5.01 | 4.96 | 4.90 | 6 |
| 4.67 | 4.57 | 4.47 | 4.42 | 4.36 | 4.31 | 4.25 | 4.20 | 7 |
| 4.20 | 4.10 | 4.00 | 3.95 | 3.89 | 3.84 | 3.78 | 3.71 | 8 |
| 3.87 | 3.77 | 3.67 | 3.61 | 3.56 | 3.51 | 3.45 | 3.39 | 9 |
| 3.62 | 3.52 | 3.42 | 3.37 | 3.31 | 3.26 | 3.20 | 3.14 | 10 |
| 3.43 | 3.33 | 3.23 | 3.17 | 3.12 | 3.06 | 3.00 | 2.94 | 11 |
| 3.28 | 3.18 | 3.07 | 3.02 | 2.96 | 2.91 | 2.85 | 2.79 | 12 |
| 3.15 | 3.05 | 2.95 | 2.89 | 2.84 | 2.78 | 2.72 | 2.66 | 13 |
| 3.05 | 2.95 | 2.84 | 2.79 | 2.73 | 2.67 | 2.61 | 2.55 | 14 |
| 2.96 | 2.86 | 2.76 | 2.70 | 2.64 | 2.59 | 2.52 | 2.46 | 15 |
| 2.89 | 2.79 | 2.68 | 2.63 | 2.57 | 2.51 | 2.45 | 2.38 | 16 |
| 2.82 | 2.72 | 2.62 | 2.56 | 2.50 | 2.44 | 2.38 | 2.32 | 17 |
| 2.77 | 2.67 | 2.56 | 2.50 | 2.44 | 2.38 | 2.32 | 2.26 | 18 |
| 2.72 | 2.62 | 2.51 | 2.45 | 2.39 | 2.33 | 2.27 | 2.20 | 19 |
| 2.68 | 2.57 | 2.46 | 2.41 | 2.35 | 2.29 | 2.22 | 2.16 | 20 |
| 2.64 | 2.53 | 2.42 | 2.37 | 2.31 | 2.25 | 2.18 | 2.11 | 21 |
| 2.60 | 2.50 | 2.39 | 2.33 | 2.27 | 2.21 | 2.14 | 2.08 | 22 |
| 2.57 | 2.47 | 2.36 | 2.30 | 2.24 | 2.18 | 2.11 | 2.04 | 23 |
| 2.54 | 2.44 | 2.33 | 2.27 | 2.21 | 2.15 | 2.08 | 2.01 | 24 |
| 2.51 | 2.41 | 2.30 | 2.24 | 2.18 | 2.12 | 2.05 | 1.98 | 25 |
| 2.49 | 2.39 | 2.28 | 2.22 | 2.16 | 2.09 | 2.03 | 1.95 | 26 |
| 2.47 | 2.36 | 2.25 | 2.19 | 2.13 | 2.07 | 2.00 | 1.93 | 27 |
| 2.45 | 2.34 | 2.23 | 2.17 | 2.11 | 2.05 | 1.98 | 1.91 | 28 |
| 2.43 | 2.32 | 2.21 | 2.15 | 2.09 | 2.03 | 1.96 | 1.89 | 29 |
| 2.41 | 2.31 | 2.20 | 2.14 | 2.07 | 2.01 | 1.94 | 1.87 | 30 |
| 2.29 | 2.18 | 2.07 | 2.01 | 1.94 | 1.88 | 1.80 | 1.72 | 40 |
| 2.17 | 2.06 | 1.94 | 1.88 | 1.82 | 1.74 | 1.67 | 1.58 | 60 |
| 2.05 | 1.94 | 1.82 | 1.76 | 1.69 | 1.61 | 1.53 | 1.43 | 120 |
| 1.94 | 1.83 | 1.71 | 1.64 | 1.57 | 1.48 | 1.39 | 1.27 | $\infty$ |

**Fišerova raspodela**




$P\{F_{\nu_1,\nu_2} < F_{\nu_1,\nu_2;\ p}\} = p$

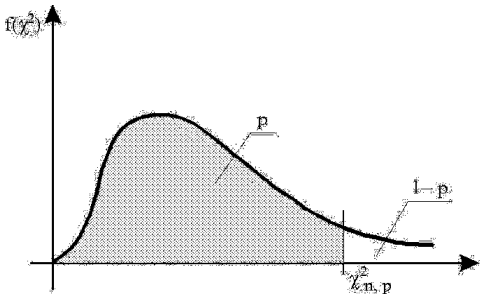
**3. c)** $p = 0{,}950$

| $\nu_2 \setminus \nu_1$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 161 | 200 | 216 | 225 | 230 | 234 | 237 | 239 | 241 | 242 |
| 2 | 18.5 | 19.0 | 19.2 | 19.2 | 19.3 | 19.3 | 19.4 | 19.4 | 19.4 | 19.4 |
| 3 | 10.1 | 9.55 | 9.28 | 9.12 | 9.01 | 8.94 | 8.89 | 8.85 | 8.81 | 8.79 |
| 4 | 7.71 | 6.94 | 6.59 | 6.39 | 6.26 | 6.16 | 6.09 | 6.04 | 6.00 | 5.96 |
| 5 | 6.61 | 5.79 | 5.41 | 5.19 | 5.05 | 4.95 | 4.88 | 4.82 | 4.77 | 4.74 |
| 6 | 5.99 | 5.14 | 4.76 | 4.53 | 4.39 | 4.28 | 4.21 | 4.15 | 4.10 | 4.06 |
| 7 | 5.59 | 4.74 | 4.35 | 4.12 | 3.97 | 3.87 | 3.79 | 3.73 | 3.68 | 3.64 |
| 8 | 5.32 | 4.46 | 4.07 | 3.84 | 3.69 | 3.58 | 3.50 | 3.44 | 3.39 | 3.35 |
| 9 | 5.12 | 4.26 | 3.86 | 3.63 | 3.48 | 3.37 | 3.29 | 3.23 | 3.18 | 3.14 |
| 10 | 4.96 | 4.10 | 3.71 | 3.48 | 3.33 | 3.22 | 3.14 | 3.07 | 3.02 | 2.98 |
| 11 | 4.84 | 3.98 | 3.59 | 3.36 | 3.20 | 3.09 | 3.01 | 2.95 | 2.90 | 2.85 |
| 12 | 4.75 | 3.89 | 3.49 | 3.26 | 3.11 | 3.00 | 2.91 | 2.85 | 2.80 | 2.75 |
| 13 | 4.67 | 3.81 | 3.41 | 3.18 | 3.03 | 2.92 | 2.83 | 2.77 | 2.71 | 2.67 |
| 14 | 4.60 | 3.74 | 3.34 | 3.11 | 2.96 | 2.85 | 2.76 | 2.70 | 2.65 | 2.60 |
| 15 | 4.54 | 3.68 | 3.29 | 3.06 | 2.90 | 2.79 | 2.71 | 2.64 | 2.59 | 2.54 |
| 16 | 4.49 | 3.63 | 3.24 | 3.01 | 2.85 | 2.74 | 2.66 | 2.59 | 2.54 | 2.49 |
| 17 | 4.45 | 3.59 | 3.20 | 2.96 | 2.81 | 2.70 | 2.61 | 2.55 | 2.49 | 2.45 |
| 18 | 4.41 | 3.55 | 3.16 | 2.93 | 2.77 | 2.66 | 2.58 | 2.51 | 2.46 | 2.41 |
| 19 | 4.38 | 3.52 | 3.13 | 2.90 | 2.74 | 2.63 | 2.54 | 2.48 | 2.42 | 2.38 |
| 20 | 4.35 | 3.49 | 3.10 | 2.87 | 2.71 | 2.60 | 2.51 | 2.45 | 2.39 | 2.35 |
| 21 | 4.32 | 3.47 | 3.07 | 2.84 | 2.68 | 2.57 | 2.49 | 2.42 | 2.37 | 2.32 |
| 22 | 4.30 | 3.44 | 3.05 | 2.82 | 2.66 | 2.55 | 2.46 | 2.40 | 2.34 | 2.30 |
| 23 | 4.28 | 3.42 | 3.03 | 2.80 | 2.64 | 2.53 | 2.44 | 2.37 | 2.32 | 2.27 |
| 24 | 4.26 | 3.40 | 3.01 | 2.78 | 2.62 | 2.51 | 2.42 | 2.36 | 2.30 | 2.25 |
| 25 | 4.24 | 3.39 | 2.99 | 2.76 | 2.60 | 2.49 | 2.40 | 2.34 | 2.28 | 2.24 |
| 26 | 4.23 | 3.37 | 2.98 | 2.74 | 2.59 | 2.47 | 2.39 | 2.32 | 2.27 | 2.22 |
| 27 | 4.21 | 3.35 | 2.96 | 2.73 | 2.57 | 2.46 | 2.37 | 2.31 | 2.25 | 2.20 |
| 28 | 4.20 | 3.34 | 2.95 | 2.71 | 2.56 | 2.45 | 2.36 | 2.29 | 2.24 | 2.19 |
| 29 | 4.18 | 3.33 | 2.93 | 2.70 | 2.55 | 2.48 | 2.35 | 2.28 | 2.22 | 2.18 |
| 30 | 4.17 | 3.32 | 2.92 | 2.69 | 2.53 | 2.42 | 2.33 | 2.27 | 2.21 | 2.16 |
| 40 | 4.08 | 3.23 | 2.84 | 2.61 | 2.45 | 2.34 | 2.25 | 2.18 | 2.12 | 2.08 |
| 60 | 4.00 | 3.15 | 2.76 | 2.53 | 2.37 | 2.25 | 2.17 | 2.10 | 2.04 | 1.99 |
| 120 | 3.92 | 3.07 | 2.68 | 2.45 | 2.29 | 2.18 | 2.09 | 2.02 | 1.96 | 1.91 |
| $\infty$ | 3.84 | 3.00 | 2.60 | 2.37 | 2.21 | 2.10 | 2.01 | 1.94 | 1.87 | 1.83 |

**nastavak 3. c)**

| 12 | 15 | 20 | 24 | 30 | 40 | 60 | 120 | $\nu_1/\nu_2$ |
|---|---|---|---|---|---|---|---|---|
| 244 | 246 | 248 | 249 | 250 | 251 | 252 | 253 | 1 |
| 19.4 | 19.4 | 19.4 | 19.5 | 19.5 | 19.5 | 19.5 | 19.5 | 2 |
| 8.74 | 8.70 | 8.66 | 8.64 | 8.62 | 8.59 | 8.57 | 8.55 | 3 |
| 5.91 | 5.86 | 5.80 | 5.77 | 5.75 | 5.72 | 5.69 | 5.66 | 4 |
| 4.68 | 4.62 | 4.56 | 4.53 | 4.50 | 4.46 | 4.43 | 4.40 | 5 |
| 4.00 | 3.94 | 3.87 | 3.84 | 3.81 | 3.77 | 3.74 | 3.70 | 6 |
| 3.57 | 3.51 | 3.44 | 3.41 | 3.38 | 3.34 | 3.30 | 3.27 | 7 |
| 3.28 | 3.22 | 3.15 | 3.12 | 3.08 | 3.04 | 3.01 | 2.97 | 8 |
| 3.07 | 3.01 | 2.94 | 2.90 | 2.86 | 2.83 | 2.79 | 2.75 | 9 |
| 2.91 | 2.85 | 2.77 | 2.74 | 2.70 | 2.66 | 2.62 | 2.58 | 10 |
| 2.79 | 2.72 | 2.65 | 2.61 | 2.57 | 2.53 | 2.49 | 2.45 | 11 |
| 2.69 | 2.62 | 2.54 | 2.51 | 2.47 | 2.43 | 2.38 | 2.34 | 12 |
| 2.60 | 2.53 | 2.46 | 2.42 | 2.38 | 2.34 | 2.30 | 2.25 | 13 |
| 2.53 | 2.46 | 2.39 | 2.35 | 2.31 | 2.27 | 2.22 | 2.18 | 14 |
| 2.48 | 2.40 | 2.33 | 2.29 | 2.25 | 2.20 | 2.16 | 2.11 | 15 |
| 2.42 | 2.35 | 2.28 | 2.24 | 2.19 | 2.15 | 2.11 | 2.06 | 16 |
| 2.38 | 2.31 | 2.23 | 2.19 | 2.15 | 2.10 | 2.06 | 2.01 | 17 |
| 2.34 | 2.27 | 2.19 | 2.15 | 2.11 | 2.06 | 2.02 | 1.97 | 18 |
| 2.31 | 2.23 | 2.16 | 2.11 | 2.07 | 2.03 | 1.98 | 1.93 | 19 |
| 2.28 | 2.20 | 2.12 | 2.08 | 2.04 | 1.99 | 1.95 | 1.90 | 20 |
| 2.25 | 2.18 | 2.10 | 2.05 | 2.01 | 1.96 | 1.92 | 1.87 | 21 |
| 2.23 | 2.15 | 2.07 | 2.03 | 1.98 | 1.94 | 1.89 | 1.84 | 22 |
| 2.20 | 2.13 | 2.05 | 2.01 | 1.96 | 1.91 | 1.86 | 1.81 | 23 |
| 2.18 | 2.11 | 2.03 | 1.98 | 1.94 | 1.89 | 1.84 | 1.79 | 24 |
| 2.16 | 2.09 | 2.01 | 1.96 | 1.92 | 1.87 | 1.82 | 1.77 | 25 |
| 2.15 | 2.07 | 1.99 | 1.95 | 1.90 | 1.85 | 1.80 | 1.75 | 26 |
| 2.13 | 2.06 | 1.97 | 1.93 | 1.88 | 1.84 | 1.79 | 1.73 | 27 |
| 2.12 | 2.04 | 1.96 | 1.91 | 1.87 | 1.82 | 1.77 | 1.71 | 28 |
| 2.10 | 2.03 | 1.94 | 1.90 | 1.85 | 1.81 | 1.75 | 1.70 | 29 |
| 2.09 | 2.01 | 1.93 | 1.89 | 1.84 | 1.79 | 1.74 | 1.68 | 30 |
| 2.00 | 1.92 | 1.84 | 1.79 | 1.74 | 1.69 | 1.64 | 1.58 | 40 |
| 1.92 | 1.84 | 1.75 | 1.70 | 1.65 | 1.59 | 1.53 | 1.47 | 60 |
| 1.83 | 1.75 | 1.66 | 1.61 | 1.55 | 1.50 | 1.43 | 1.35 | 120 |
| 1.75 | 1.67 | 1.57 | 1.52 | 1.46 | 1.39 | 1.32 | 1.22 | $\infty$ |

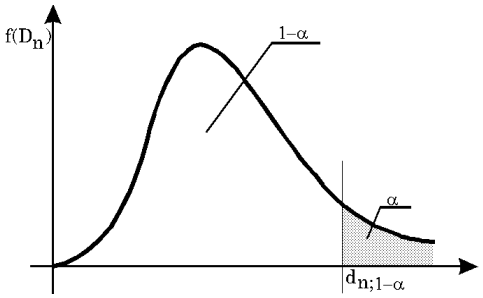
## 4. $\chi^2$ raspodela




$P\{\chi_n^2 < \chi_{n;\ p}^2\} = p$

| $n \setminus p$ | 0.005 | 0.010 | 0.025 | 0.050 | 0.95 | 0.975 | 0.990 | 0.995 |
|---|---|---|---|---|---|---|---|---|
| 1 | .0000 | .0002 | .0010 | .0039 | 3.84 | 5.02 | 6.63 | 7.88 |
| 2 | .0100 | .0201 | .0506 | .103 | 5.99 | 7.38 | 9.21 | 10.6 |
| 3 | .0717 | .115 | .216 | .352 | 7.81 | 9.35 | 11.3 | 12.8 |
| 4 | .207 | .297 | .484 | .711 | 9.49 | 11.1 | 13.3 | 14.9 |
| 5 | .412 | .554 | .831 | 1.15 | 11.1 | 12.8 | 15.1 | 16.7 |
| 6 | .676 | .872 | 1.24 | 1.64 | 12.6 | 14.4 | 16.8 | 18.5 |
| 7 | .989 | 1.24 | 1.69 | 2.17 | 14.1 | 16.0 | 18.5 | 20.3 |
| 8 | 1.34 | 1.65 | 2.18 | 2.73 | 15.5 | 17.5 | 20.1 | 22.0 |
| 9 | 1.73 | 2.09 | 2.70 | 3.33 | 16.9 | 19.0 | 21.7 | 23.6 |
| 10 | 2.16 | 2.56 | 3.25 | 3.94 | 18.3 | 20.5 | 23.2 | 25.2 |
| 11 | 2.60 | 3.05 | 3.82 | 4.57 | 19.7 | 21.9 | 24.7 | 26.8 |
| 12 | 3.07 | 3.57 | 4.40 | 5.23 | 21.0 | 23.3 | 26.2 | 28.3 |
| 13 | 3.57 | 4.11 | 5.01 | 5.89 | 22.4 | 24.7 | 27.7 | 29.8 |
| 14 | 4.07 | 4.66 | 5.63 | 6.57 | 23.7 | 26.1 | 29.1 | 31.3 |
| 15 | 4.60 | 5.23 | 6.26 | 7.26 | 25.0 | 27.5 | 30.6 | 32.8 |
| 16 | 5.14 | 5.81 | 6.91 | 7.96 | 26.3 | 28.8 | 32.0 | 34.3 |
| 17 | 5.70 | 6.41 | 7.56 | 8.67 | 27.6 | 30.2 | 33.4 | 35.7 |
| 18 | 6.26 | 7.01 | 8.23 | 9.39 | 28.9 | 37.5 | 34.8 | 37.2 |
| 19 | 6.84 | 7.63 | 8.91 | 10.1 | 30.1 | 32.9 | 36.2 | 38.6 |
| 20 | 7.43 | 8.26 | 9.59 | 10.9 | 31.4 | 34.2 | 37.6 | 40.0 |
| 21 | 8.03 | 8.90 | 10.3 | 11.6 | 32.7 | 35.5 | 38.9 | 41.4 |
| 22 | 8.64 | 9.54 | 11.0 | 12.3 | 33.9 | 36.8 | 40.3 | 42.8 |
| 23 | 9.26 | 10.2 | 11.7 | 13.1 | 35.2 | 38.1 | 41.6 | 44.2 |
| 24 | 9.89 | 10.9 | 12.4 | 13.8 | 36.4 | 39.4 | 43.0 | 45.6 |
| 25 | 10.5 | 11.5 | 13.1 | 14.6 | 37.7 | 40.6 | 44.3 | 46.9 |
| 26 | 11.2 | 12.2 | 13.8 | 15.4 | 38.9 | 41.9 | 45.6 | 48.3 |
| 27 | 11.8 | 12.9 | 14.6 | 16.2 | 40.1 | 43.2 | 47.0 | 49.6 |
| 28 | 12.5 | 13.6 | 15.3 | 16.9 | 41.3 | 44.5 | 48.3 | 51.0 |
| 29 | 13.1 | 14.3 | 16.0 | 17.7 | 42.6 | 45.7 | 49.6 | 52.3 |
| 30 | 13.8 | 15.0 | 16.8 | 18.5 | 43.8 | 47.0 | 50.9 | 53.7 |
| 40 | 20.7 | 22.2 | 24.4 | 26.5 | 55.8 | 59.3 | 63.7 | 66.8 |
| 50 | 28.0 | 29.7 | 32.4 | 34.8 | 67.5 | 71.4 | 76.2 | 79.5 |
| 60 | 35.5 | 37.5 | 40.5 | 43.2 | 79.1 | 83.3 | 88.4 | 92.0 |
| 70 | 43.3 | 45.4 | 48.8 | 51.7 | 90.5 | 95.0 | 100 | 104 |
| 80 | 51.2 | 53.5 | 57.2 | 60.4 | 102 | 107 | 112 | 116 |
| 90 | 59.2 | 61.8 | 65.6 | 69.1 | 113 | 118 | 124 | 128 |
| 100 | 67.3 | 70.1 | 74.2 | 77.9 | 124 | 130 | 136 | 140 |

## 5. Kritične vrednosti za test Kolmogorova




$P\{D_n \geq d_{n;\ 1-\alpha}\} = \alpha$

| $n \setminus \alpha$ | 0.200 | 0.150 | 0.100 | 0.050 | 0.010 |
|---|---|---|---|---|---|
| 1 | .900 | .925 | .950 | .975 | .995 |
| 2 | .684 | .726 | .776 | .842 | .929 |
| 3 | .565 | .597 | .642 | .708 | .829 |
| 4 | .494 | .525 | .564 | .624 | .734 |
| 5 | .446 | .474 | .510 | .563 | .669 |
| 6 | .410 | .436 | .470 | .521 | .618 |
| 7 | .381 | .405 | .438 | .486 | .577 |
| 8 | .358 | .381 | .411 | .457 | .543 |
| 9 | .339 | .360 | .388 | .432 | .514 |
| 10 | .322 | .342 | .368 | .409 | .486 |
| 11 | .307 | .326 | .352 | .391 | .468 |
| 12 | .295 | .313 | .338 | .375 | .450 |
| 13 | .284 | .302 | .325 | .361 | .433 |
| 14 | .274 | .292 | .314 | .349 | .418 |
| 15 | .266 | .283 | .304 | .338 | .404 |
| 16 | .258 | .274 | .295 | .328 | .391 |
| 17 | .250 | .266 | .286 | .318 | .380 |
| 18 | .244 | .259 | .278 | .309 | .370 |
| 19 | .237 | .252 | .272 | .301 | .361 |
| 20 | .231 | .246 | .264 | .294 | .352 |
| 25 | .210 | .220 | .240 | .264 | .320 |
| 30 | .190 | .200 | .220 | .242 | .290 |
| 35 | .180 | .190 | .210 | .230 | .270 |
| 40 | | | | .190 | .230 |
| 50 | | | | .190 | .230 |
| 60 | | | | .170 | .210 |
| 70 | | | | .160 | .190 |
| 80 | | | | .150 | .180 |
| 90 | | | | .140 | |
| 100 | | | | .140 | |
| asimptotska formula | $\frac{1.07}{\sqrt{n}}$ | $\frac{1.14}{\sqrt{n}}$ | $\frac{1.22}{\sqrt{n}}$ | $\frac{1.36}{\sqrt{n}}$ | $\frac{1.63}{\sqrt{n}}$ |

## 6. Slučajni brojevi

| | | | | |
|---|---|---|---|---|
| 51772 | 74640 | 42331 | 29044 | 46621 |
| 24033 | 23491 | 83587 | 06568 | 21960 |
| 45939 | 60173 | 52078 | 25424 | 11645 |
| 30586 | 02133 | 75797 | 45406 | 31041 |
| 03585 | 79353 | 81938 | 82322 | 96799 |
| 64937 | 03355 | 98683 | 20790 | 65304 |
| 15630 | 64759 | 51135 | 98527 | 62586 |
| 09448 | 56301 | 57683 | 30277 | 94623 |
| 21631 | 91157 | 77331 | 60710 | 52290 |
| 91097 | 17480 | 29414 | 06829 | 87843 |
| 62898 | 93582 | 04186 | 19640 | 87056 |
| 21387 | 76105 | 10863 | 97453 | 90581 |
| 55870 | 56974 | 37428 | 93507 | 94271 |
| 86707 | 12973 | 17169 | 88116 | 42187 |
| 85659 | 36081 | 50884 | 14070 | 74950 |
| 55189 | 00745 | 65253 | 11822 | 15804 |
| 41889 | 25439 | 88036 | 24034 | 67283 |
| 85418 | 68829 | 06652 | 41982 | 49159 |
| 16835 | 48653 | 71590 | 16159 | 14676 |
| 28195 | 27279 | 47152 | 35683 | 47280 |

# Literatura


1. Anděl J.: **Matematická statistika**, SNTZ/Alfa, Praha, 1985

2. Anderson T.: **The statistical analysis of time series**, John Wiley & Sons, New York, 1971

3. Borovkov A.: **Matematičeskaja statistika**, Nauka, Moskva, 1984

4. Box G., Jenkins G.: **Time series analysis**, Holden-Day, 1970

5. Brockwell P., Davis R.: **Time series: Theory and methods**, Springer-Verlag, New York, 1987

6. Brownlee K.: **Statistical theory and methodology in science and engineering**, John Wiley & Sons, New York, 1977

7. Cochran W.: **Sampling techniques**, John Wiley & Sons, New York, (1953), 1963

8. Deivid G.: **Porjadkovie statistiki**, Nauka, Moskva, 1979 (prevod sa engleskog)

9. DeGroot M.: **Optimal statistical decisions**, McGrow-Hill Co., New York, 1970

10. Djuge D.: **Teoretičeskaja i prikladnaja statistika**, Nauka, Moskva, 1972

11. Ermakov S. i drugi: **Matematičeskaja teorija planirovanija eksperimenta**, Nauka, Moskva, 1983

12. Hadžić O.: **Numeričke i statističke metode u obradi eksperimentalnih podataka**, Univerzitet u Novom Sadu, Institut za matematiku, Novi Sad, 1992

13. Hogg R., Craig A.: **Introduction to mathematical statistics**, Macmillan Co., New York, 1965

14. Ivčenko G. I., Medvedev J. I.: **Matematičeskaja statistika**, Višaja škola, Moskva, 1984

15. Kendall M. G., Steward A.: **The Advanced Theory of Statistics**, vol. 3: Design and Analysis, and Time Series, Griffin company, London, 1968

16. Mališić J.: **Vremenske serije**, Matematički fakultet, Beograd, 2002

17. Popović B., Blagojević B.: **Matematička statistika sa primenama u hidrotehnici**, Univerzitet u Nišu, Niš, (1997, 1999), 2003

18. Popović B., Ristić M.: **Statistika u psihologiji**, Mrlješ, Beograd, 2001

19. Rao S.R.: **Linejnije statističeskie metodi i ih primenenija**, Nauka, Moskva, 1968 (prevod sa engleskog)

20. Spiegel M. R.: **Probability and Statistics**, McGrow–Hill Book company, New York, 1975

21. Stojanović S.: **Matematička statistika**, Naučna knjiga, Beograd, 1979

22. Šmeterer L.: **Vvedenie v matematičeskuju statistiku**, Nauka, Moskva, 1976 (prevod sa nemačkog)

23. Thompson S.: **Sampling**, John Wiley & Sons, New York, 1992

24. Wackerly D., Mendenhall W., Scheaffer R.: **Mathematical statistics with applications**, Duxbury Press, Belmont, 1996

25. Zacks S.: **The theory of statistical inference**, John Wiley & Sons, New York, 1971

# Indeks pojmova