Sadržaj

1

Elementi teorije verovatnoće

Dugo su naučni zakoni iskazivani tako da određeni uslovi nekog opita (pojave) jednoznačno određuju rezultat (ishod) tog opita. Međutim, ako se posmatra, na primer, bacanje kocke za igru, pokazuje se da rezultat ovog opita nije jednoznačno određen, budući da može da padne jedan, dva, tri, četiri, pet ili šest. Naučna analiza ovakvih i sličnih pojava i njihovih zakonitosti počinje od 17. veka. Matematička teorija ovih pojava jeste Matematička verovatnoća. Ona je podloga Matematičke statistike.

Posmatrajmo bacanje kocke. Pojavljivanje jednog od brojeva 1, 2, 3, 4, 5, 6 nazivamo ishod eksperimenta ili elementarni događaj. Na primer, "kocka pokazuje broj 1" je elementaran događaj. Obeležimo ga sa A_1. Slično imamo A_2, A_3, A_4, A_5, A_6 . Međutim, postoje događaji koji nisu elementarni. Na primer, "kocka pokazuje paran broj" je događaj koji se sastoji od više događaja. Naime, on će se ostvariti ako se ostvari bilo koji od događaja A_2, A_4, A_6. Pri (poštenom) bacanju kocke ne možemo uticati da li će se, na primer, ostvariti A_3 ili neće. Zbog toga takve događaje zovemo slučajnim događajima a sam eksperiment slučajnim eksperimentom. Ipak, sa sigurnošću možemo tvrditi da će se u slučajnom eksperimentu bacanja kocke događaj "kocka pokazuje jedan od brojeva 1, 2, 3, 4, 5, 6" uvek ostvariti. Takav događaj se zove siguran događaj. Međutim, događaj "kocka pokazuje broj 7" ne može da se ostvari, pa takav događaj nazivamo nemoguć . Događaj "kocka pokazuje bilo koji od brojeva 1, 2, 3, 4, 5" je suprotan događaju "kocka pokazuje broj 6" tj. događaju A_6 . Označavamo ga sa $\overline{A}_6$. Uopšte, događaj $\overline{A}$ je suprotan događaju A ako i samo ako se on realizuje kada se A ne realizuje. Skup svih mogućih ishoda

(elementarnih događaja) u slučajnom eksperimentu E obeležava se sa S_E ili kraće sa S.

Definicija 1.1. *Slučajni događaj je bilo koji podskup skupa svih elementarnih događaja eksperimenta E.*

Primer 1.1. Neka je E bacanje kocke. Znamo da je $S = \{1, 2, 3, 4, 5, 6\}$. Jednočlani podskupovi $\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}$ su elementarni događaji. Podskup $\{2, 4, 6\}$ odgovara događaju "kocka pokazuje paran broj". Podskup $\{1, 2, 3, 4\}$ odgovara događaju "kocka pokazuje broj manji od 5".

Ako neki elementarni događaj pripada nekom skupu A kažemo da je taj elementarni događaj povoljan za događaj A.

Definicija 1.2. *Ako neki eksperiment ima konačno mnogo elementarnih događaja i ako su svi elementarni događaji jednako verovatni, tada je verovatnoća nekog događaja A jednaka količniku broja povoljnih elementarnih događaja i broja svih elementarnih događaja.*

Iz ove definicije sledi da je verovatnoća nemogućeg događaja jednaka 0, a da je verovatnoća sigurnog događaja jednaka 1.

Primer 1.2. Naći verovatnoću događaja "kocka pokazuje broj manji od 5".

Rešenje. U ovom slučaju je $S = \{1, 2, 3, 4, 5, 6\}$ a događaj čija se verovatnoća traži je $A = \{1, 2, 3, 4\}$. Verovatnoća događaja A je $P(A) = \frac{4}{6} = \frac{2}{3}$.

Primer 1.3. Jedno naselje broji 6200 stanovnika od kojih je 434 obolelo od gripa. Ako na slučajan način iz ovog naselja izaberemo jednog stanovnika, kolika je verovatnoća da je on oboleo od gripa?

Rešenje. Ovde je broj elementarnih događaja 6200 a broj "povoljnih" događaja 434, pa je $P(A) = \frac{434}{6200} = 0,07$.

1.1 Raspodele verovatnoća

Neka je S skup elementarnih događaja nekog eksperimenta E. U primenama smo često u situaciji da svakom elementu skupa S dodeljujemo realan broj, odnosno da, na neki način, kodiramo taj elementarni događaj.

Primer 1.4. Novčić se baca dva puta. Neka je X broj registrovanih pisama t.j. $X(\Pi,\Pi)=2,\ X(\Pi,\Gamma)=X(\Gamma,\Pi)=1,\ X(\Gamma,\Gamma)=0$. Ovakva funkcija X je primer takozvane slučajne promenljive.

Definicija 1.3. *Funkcija X koja svakom elementarnom događaju dodeljuje realan broj zove se slučajna promenljiva.*

Nad skupom elementarnih događaja jednog eksperimenta može se definisati više slučajnih promenljivih. Tako u prethodnom primeru možemo da definišemo slučajnu promenljivu Y na sledeći način:

$Y(\Pi,\Pi)=Y(\Gamma,\Gamma)=1$ i $Y(\Pi,\Gamma)=Y(\Gamma,\Pi)=0$,

t.j. vrednost funkcije Y je jednaka 1 ako dva puta padne ista strana, odnosno 0 ako padnu različite strane.

Verovatnoća da slučajna promenljiva X uzme vrednost x obeležava se sa $P\{X=x\}$.

Definicija 1.4. *Slučajna promenljiva X je diskretnog tipa ako i samo ako postoji skup $\{x_1,x_2,\ldots,x_n\}$ realnih brojeva takav da je*

$$\sum_{k=1}^{n} P\{X=x_k\}=1.$$

Skup uređenih parova (x_k,p_k) $(k=1,2,3,\ldots,n)$, gde je $p_k=P\{X=x_k\}$, zove se raspodela verovatnoća diskretne slučajne promenljive X i obično se prikazuje šemom

$$\begin{pmatrix} x_1 & x_2 & \cdots & x_n \\ p_1 & p_2 & \cdots & p_n \end{pmatrix}$$

Ako u ravni odredimo tačke (x_k,p_k), onda pomoću njih određujemo poligon raspodele verovatnoća diskretne slučajne promenljive.

Specijalno, ako je raspodela oblika

$$\begin{pmatrix} x_1 & x_2 & \cdots & x_n \\ \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \end{pmatrix}$$

tada se govori o diskretnoj **uniformnoj raspodeli**.

Primer 1.5. U eksperimentu bacanja novčića sa skupom elementarnih događaja $S = \{\Pi, \Gamma\}$ definišimo slučajnu promenljivu X sa: $X(\Gamma) = 0$ i $X(\Pi\} = 1$. Ovde je $X(S) = \{0, 1\}$. Ako smatramo da su ishodi eksperimenta jednako verovatni, onda imamo $p_0 = P\{X = 0\} = \frac{1}{2}$ i $p_1 = P\{X = 1\} = \frac{1}{2}$. Ova raspodela verovatnoća se može prikazati šemom

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

Primetimo da se u prethodnom primeru radilo o uniformonoj raspodeli. Slično imamo kod bacanja kocke, gde se raspodela verovatnoća može prikazati šemom

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \end{pmatrix}$$

Primer 1.6. Bacamo istovremeno dve kocke i posmatramo zbir brojeva koje pokazuju kocke. Neka je $X(i, j) = i + j$, gde je i broj koji pokazuje prva kocka a j broj koji pokazuje druga kocka. Odredimo raspodelu verovatnoća za X.

Rešenje. Ovde je skup elementarnih događaja skup uređenih parova (i, j) gde je $i, j \in \{1, 2, 3, 4, 5, 6\}$ t.j.

$$\begin{aligned} S = \{\ & (1,1),(1,2),(1,3),(1,4),(1,5),(1,6), \\ & (2,1),(2,2),(2,3),(2,4),(2,5),(2,6), \\ & (3,1),(3,2),(3,3),(3,4),(3,5),(3,6), \\ & (4,1),(4,2),(4,3),(4,4),(4,5),(4,6), \\ & (5,1),(5,2),(5,3),(5,4),(5,5),(5,6) \\ & (6,1),(6,2),(6,3),(6,4),(6,5),(6,6)\}. \end{aligned}$$

Događaj "zbir brojeva koji pokazuju kocke je 2" je $\{(1, 1)\}$. Kao što se vidi broj povoljnih elementarnih događaja je 1, dok je broj mogućih elementarnih događaja 36, pa je $p_2 = P\{X = 2\} = \frac{1}{36}$.

Događaj "zbir brojeva koji pokazuju kocke je 3" je $\{(1, 2), (2, 1)\}$, t.j. broj povoljnih elementarnih događaja je 2 a broj mogućih je, opet, 36, pa je $p_3 = P\{X = 3\} = \frac{2}{36}$.

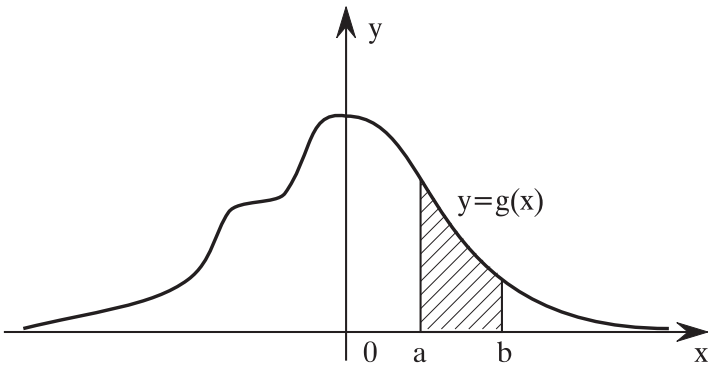
Na sličan način dobijamo i ostale verovatnoće. Tako, na primer, događaj "zbir brojeva koje pokazuju kocke je 8" je $\{(2,6),(3,5),(4,4),(5,3),(6,2)\}$, t.j. broj povoljnih elementarnih događaja je 5 a mogućih je 36, pa se dobija $p_8 = P\{X = 8\} = \frac{5}{36}$.

Raspodela verovatnoća slučajne promenljive X sada može da se prikaže šemom

$$\begin{pmatrix} 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \\ \frac{1}{36} & \frac{2}{36} & \frac{3}{36} & \frac{4}{36} & \frac{5}{36} & \frac{6}{36} & \frac{5}{36} & \frac{4}{36} & \frac{3}{36} & \frac{2}{36} & \frac{1}{36} \end{pmatrix}$$

Primetimo da je zbir ovih verovatnoća jednak 1 i da odgovara sigurnom događaju S.

Definicija 1.5. *Slučajna promenljiva X je neprekidnog tipa ako i samo ako postoji nenegativna funkcija $y = g(x)$ takva da je $P(a \leq X \leq b)$ jednaka površini krivolinijskog trapeza ograničenog krivom $y = g(x)$, pravom $x = a$, pravom $x = b$ i x-osom (Slika 1.1). Funkcija $y = g(x)$ se zove gustina raspodele verovatnoća slučajne promenljive X.*



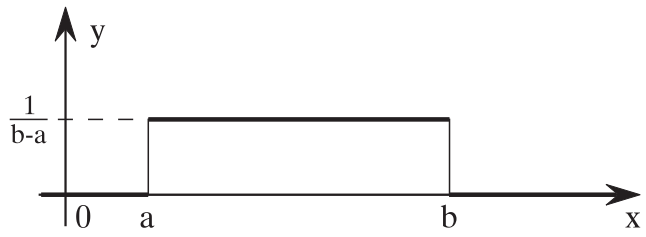
Slika 1.1.

Svaka gustina raspodele verovatnoća ima osobinu da je površina ograničena grafikom gustine i x-osom jednaka 1.

Neka je $a < b$. Može se pokazati da je funkcija

$$g(x) = \begin{cases} \frac{1}{b-a}, & x \in [a,b] \\ 0, & x \notin [a,b] \end{cases}$$

jedna gustina. Za slučajnu promenljivu koja je određena ovom gustinom kaže se da ima **uniformnu raspodelu** na intervalu $[a,b]$ i ta raspodela se obeležava sa $\mathcal{U}(a,b)$. Jedna takva gustina je prikazana na Slici 1.2.



Slika 1.2.

Ako je X neprekidna slučajna promenljiva, tada je $P\{X = a\}$, gde je a bilo koji realan broj. Ovo znači da je, za neprekidnu slučajnu promenljivu X, verovtnoća bilo kog elementarnog događaja $\{X = a\}$ jednaka nuli. Pirodno je da ova verovatnoća bude nula, jer bi u suprotnom zbir verovatnoća elementarnih događaja bio beskonačan, a po definiciji verovatnoće on treba da bude jednak 1. Činjenica da je verovatnoća nekog elementarnog događaja jednaka nuli ne znači da se taj događaj neće nikada ostvariti, već da je ta verovatnoća veoma mala.

Iz prethodnog proizilazi da je

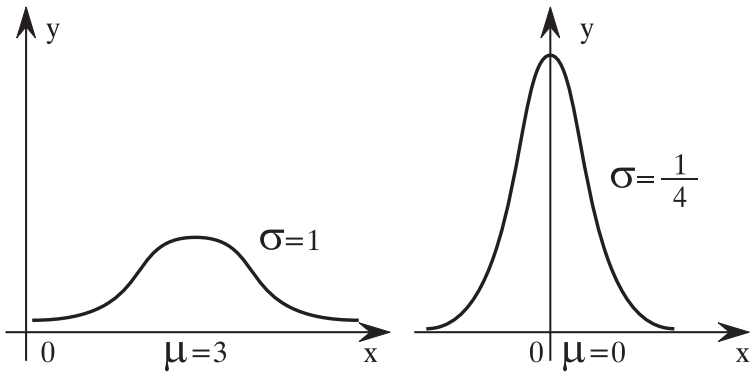
$$P\{a \leq X \leq b\} = P\{a < X \leq b\} = P\{a \leq X < b\} = P\{a < X < b\}.$$

1.2 Normalna raspodela

Za slučajnu promenljivu X kažemo da ima Gausovu ili **normalnu raspodelu** verovatnoća s parametrima μ i σ ako je njena gustina

$$g(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\,\sigma^2}}. \tag{1.1}$$

Ovu raspodelu označavaćemo sa $N(\mu, \sigma^2)$. Sve krive date pomoću poslednje jednakosti su simetrične u odnosu na pravu $x = \mu$ i imaju oblik osnog preseka zvona. Na Slici 1.6 date su krive oblika (1.1), za različite vrednosti parametara μ i σ.



Slika 1.3.

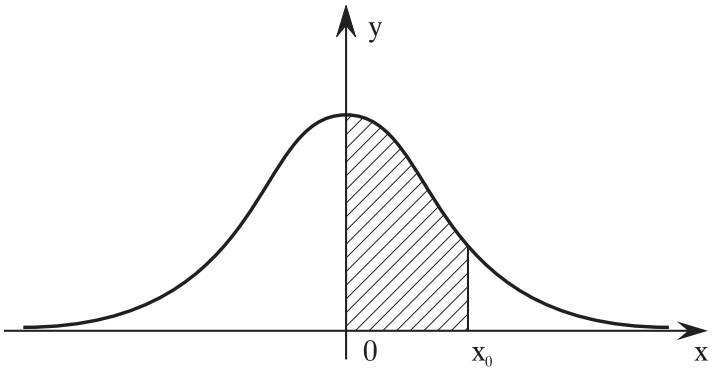
Interesantan je uticaj parametra σ na oblik krive raspodele. Što je σ veće, to je kriva raspodele spljoštenija i šira. Obratno, što je σ manje kriva je špicastija i uža.

Specijano, ako je $\mu = 0$ i $\sigma = 1$, onda kažemo da slučajna promenljiva ima **standardizovanu normalnu raspodelu**, koju označavamo sa $N(0, 1)$. Kod ove slučajne promenljive verovatnoća se izračunava na sledeći način:

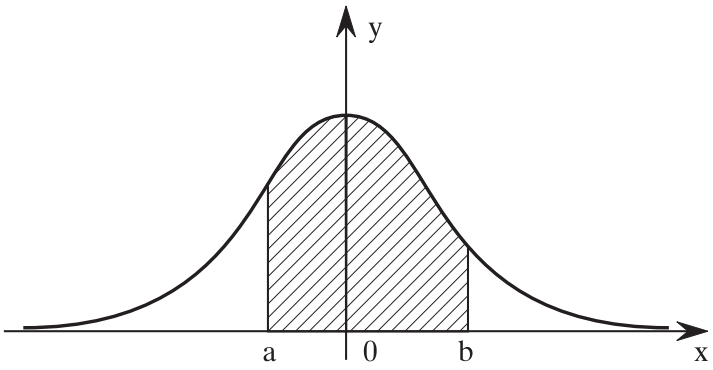
$$P\{a \leq X \leq b\} = \Phi(b) - \Phi(a),$$

gde je Φ takozvana Laplasova funkcija. Geometrijski tumačeno, $\Phi(x_0)$ predstavlja površinu "krivolinijskog trapeza" označenog na Slici 1.4. Razlika $\Phi(b) - \Phi(a)$ prikazana je na Slici 1.5. Vrednosti Laplasove funkcije $\Phi(x)$, za različite pozitivne vrednosti x, nalaze se u Tabeli II. Može da se dokaže da Laplasova funkcija $\Phi(x)$ ima osobine:

1. $\Phi(0) = 0$,
2. $\Phi(+\infty) = 0,5$,
3. $\Phi(-x) = -\Phi(x)$.



Slika 1.4.



Slika 1.5.

Verovatnoće oblika $P(a \leq Z \leq b)$ za slučajnu promenljivu X sa raspodelom $N(\mu, \sigma^2)$ mogu se odrediti pomoću funkcije $\Phi(x)$ i Tabele II. Naime, ako slučajna promenljiva X ima normalnu raspodelu $N(\mu, \sigma^2)$ tada slučajna promenjiva

$$Z = \frac{X - \mu}{\sigma}$$

ima standardizovanu normalnu raspodelu $N(0, 1)$.

Primer 1.7. Slučajna promenljiva X ima normalnu raspodelu $N(0, 1)$. Izračunati

(a) $P\{0 \leq X \leq 1, 42)\}$

(b) $P\{-1, 37 \leq X \leq 2, 01)\}$.

Rešenje.

(a) $P\{0 \leq X \leq 1, 42\} = \Phi(1, 42) - \Phi(0) = 0, 4222 - 0 = 0, 4222,$

(b) $P\{-1, 37 \leq X \leq 2, 01)\} = \Phi(2, 01) - \Phi(-1, 37) = \Phi(2, 01) + \Phi(1, 37)$

$= 0, 4778 + 0, 4147 = 0, 8925.$

Primer 1.8. Slučajna promenljiva X ima normalnu raspodelu $N(8, 4^2)$. Izračunati $P\{5 \leq X \leq 10\}$.

Rešenje.

$$\begin{aligned} P\{5 \leq X \leq 10\} &= P(\tfrac{5-8}{4} \leq \tfrac{X-8}{4} \leq \tfrac{10-8}{4}) = P(-\tfrac{3}{4} \leq Z \leq \tfrac{1}{2}) \\ &= \Phi(\tfrac{1}{2}) - \Phi(-\tfrac{3}{4}) = \Phi(0, 5) + \Phi(0, 75) \\ &= 0, 2737 + 0, 1915 = 0, 4649. \end{aligned}$$

Primer 1.9. Slučajna promenljiva X ima normalnu raspodelu $N(0, 1)$. Odrediti broj c tako da je

(a) $P(|X| \geq c) = 0, 05$

(b) $P(X \leq c) = 0, 05$

(c) $P(X \geq c) = 0, 05.$

Rešenje.

(a) $P(|X| \geq c) = 1 - P(|X| \leq c) = 1 - P(-c \leq X \leq c) = 1 - (\Phi(c) - \Phi(-c)) = 1 - 2\Phi(c)$. Iz uslova $P(|X| \geq c) = 0, 05$ dobijamo $1 - 2\Phi(c) = 0, 05$, odakle je $\Phi(c) = 0, 475$. Iz tabele normalne raspodele dobija se $c = 1, 96$.

(b) $P(X \leq c) = P(-\infty < X \leq c) = \Phi(c) - \Phi(\infty) = \Phi(c) - 0,5$. Iz uslova $\Phi(c) - 0,5 = 0,05$ dobijamo $\Phi(c) = -0,45$. Iz tablice normalne raspodele dobija se $c = -1,65$.

(c) $P(X \geq c) = P(c \leq X < \infty) = \Phi(\infty) - \Phi(c) = 0,5 - \Phi(c)$. Iz uslova $0,5 - \Phi(c) = 0,05$ dobijamo $\Phi(c) = 0,45$. Iz tablice normalne raspodele dobija se $c = 1,65$.

Primer 1.10. Slučajna promenljiva X ima normalnu raspodelu $N(0,1)$. Odrediti broj c tako da je

(a) $P(|X| \geq c) = 0,01$

(b) $P(X \leq c) = 0,01$

(c) $P(X \geq c) = 0,01$.

Rešenje. Postupkom koji je korišćen u prethodnom primeru dobija se

(a) $c = 2,58$, (b) $c = -2,32$, (c) $c = 2,32$.

Normalna raspodela ima važnu ulogu u teoriji verovatnoće i matematičkoj statistici. U praksi se često srećemo sa slučajnim promenljivim čije su raspodele verovatnoća normalne ili veoma bliske normalnoj raspodeli. Neke slučajne promenljive, koje nemaju normalnu raspodelu, mogu da se transformišu u slučajne promenljive sa normalnom raspodelom. Takođe se iz normalne raspodele izvode i druge važne raspodele.

1.3 χ^2 raspodela

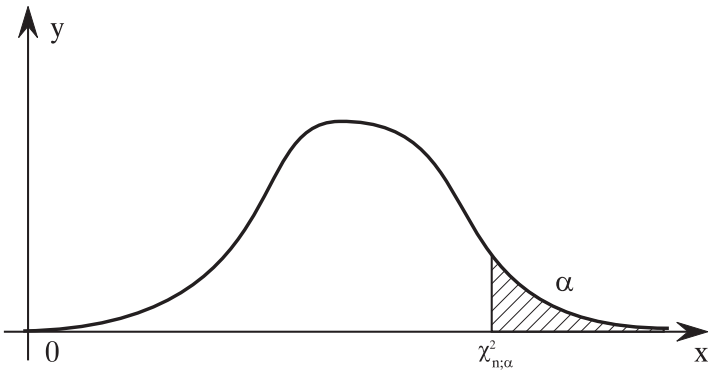
Slučajna promenljiva χ^2_n ("hi kvadrat") koja ima χ^2 raspodelu je neprekidnog tipa. Tipičan grafik gustine za ovu raspodelu prikazan je na Slici 1.6. Karakteristika χ^2 raspodele je da zavisi od parametra n $(n = 1, 2, \ldots)$. Raspodela za slučajnu promenljivu χ^2_n zove se χ^2 raspodela sa n stepena slobode.

Slučajna promenljiva χ^2 ima samo nenegativne vrednosti a njena gustina je jednaka 0 za negativne vrednosti argumenta.

Verovatnoće vezane za χ^2 raspodelu daju se tabelarno. Naime, za dati stepen slobode n i za dati broj a $(0 < a < 1)$, iz Tabele III se čita broj $\chi^2_{n;\,\alpha}$ takav da je

$$P\{\chi_n^2 \geq \chi_{n;\,\alpha}^2\} = \alpha.$$

Na primer, $\chi_{8;\,0,05}^2 = 15,507$. Na Slici 1.7. prikazan je broj $\chi_{n;\,\alpha}^2$ i verovatnoća α, koja je predstavljena kao površina šrafirane površi.



Slika 1.6.

U tabelama se, obično, daju vrednosti χ^2-raspodele za $n \leq 30$. Ako je $n > 30$, onda χ_n^2 ima približno normalnu raspodelu $N(n, \sqrt{2n})$.

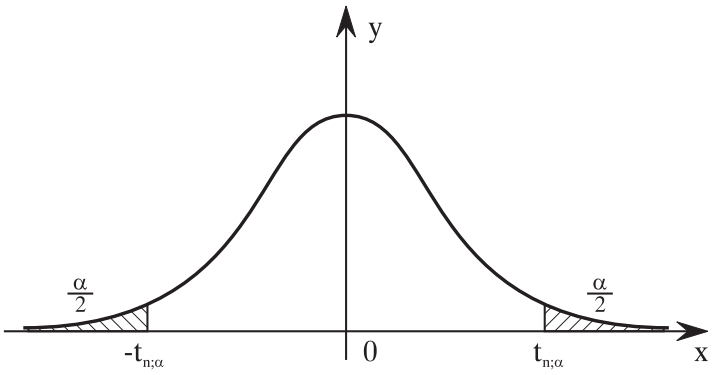
1.4 Studentova t raspodela

Slučajna promenljiva t_n koja ima Studentovu t raspodelu je neprekidnog tipa i zavisi od parametra n $(n = 1, 2, \ldots)$. Tipičan grafik gustine za ovu raspodelu prikazan je na Slici 1.7. Grafik gustine je simetričan u odnosu na $y-$osu. Raspodela za slučajnu promenljivu t_n zove se Studentova t raspodela sa n stepena slobode.

Verovatnoće vezane za Studentovu t raspodelu date su u Tabeli IV. Za određeni stepen slobode n i određeni broj α $(0 < a < 1)$ iz Tabele IV se čita broj $t_{n;\,\alpha}$ takav da je

$$P\{|t_n| \geq t_{n;\,\alpha}\} = \alpha.$$

Na primer, $t_{12;\,0,02} = 1,356$.



Slika 1.7.

Na Slici 8 verovatnoća α je prikazana kao zbir šrafiranih površina koje odgovaraju verovatnoći događaja $\{t_n \leq -t_{n;\,\alpha}\} \cup \{t_n \geq t_{n;\,\alpha}\}$. Primetimo da je, zbog simetrije,

$$P\{|t_n| \geq t_{n;\,\alpha}\} = P\{t_n \ \leq -t_{n;\,\alpha}\} + P\{t_n \geq t_{n;\,\alpha}\} + P\{t_n \ \leq -t_{n;\,\alpha}\} = \frac{\alpha}{2} + \frac{\alpha}{2} = \alpha.$$

Ako umesto α stavimo 2α dobijamo

$$P\{|t_n| \geq t_{n;2\,\alpha}\} = P\{t_n \geq t_{n;\,2\alpha}\} + P\{t_n \ \leq -t_{n;\,2\alpha}\} = \alpha + \alpha = 2\alpha,$$

odakle dobijamo

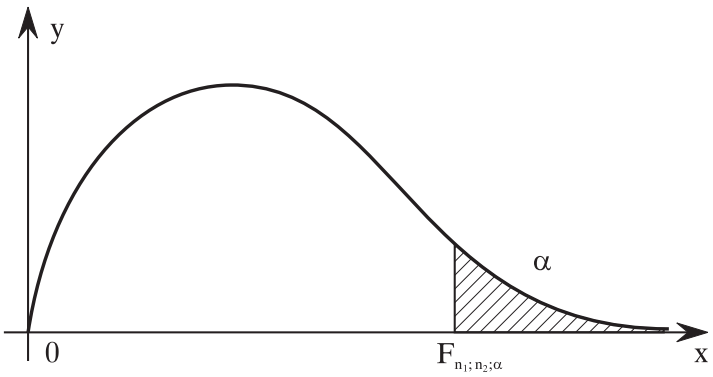
$$P\{t_n \geq t_{n;2\,\alpha}\} = \alpha \quad \text{i} \quad P\{t_n \leq -t_{n;\,2\alpha}\} = \alpha.$$

Kada n teži beskonačnosti tada Studentova t raspodela teži normalnoj raspodeli $N(0,1)$. Ako je $n > 30$, onda se Studentova raspodela dobro aproksimira raspodelom $N(0,1)$. Otuda u Tabeli IV za Studentovu t raspodelu nisu date vrednosti verovatnoća za $n > 30$.

1.5 Fišerova raspodela

Slučajna promenljiva $F_{n_1;\,,n_2;\,\alpha}$ koja ima Fisherovu raspodelu je neprekidnog tipa i zavisi od dva parametra n_1 i n_2. Na Slici 1.8. prikazana je gustina

jedne Fišerove slučajne promenljive, broj $F_{n_1;,n_2;\alpha}$ kao i verovatnoća α, koja je predstavljena kao površina šrafirane površi.



Slika 1.8.

Kako Fišerova raspodela zavisi od dva parametra n_1 i n_2 , njeno tabeliranje je nešto komplikovanije. U Tabeli V daju se, za razne vrednosti n_1 i n_2 i za $\alpha = 0,05$, brojevi $F_{n_1;\,n_2;\,0,05}$ takvi da je

$$P\{F_{n_1;\,n_2} \geq F_{n_1;\,n_2;\,0,05}\} = 0,05.$$

Na primer, za $n_1 = 5,\ n_2 = 10,\ \alpha = 0,05$ iz tabele V čitamo $F_{5;10;0,05} = 3,48$.

1.6 Matematičko očekivanje

Matematičko očekivanje je, na neki način, srednja vrednost slučajne promenljive. Ovde dajemo formalnu definiciju ovog pojma.

Ako je X diskretna slučajna promenljiva sa raspodelom

$$\begin{pmatrix} x_1 & x_2 & \cdots & x_n \\ p_1 & p_2 & \cdots & p_n \end{pmatrix}$$

tada je matematičko očekivanje slučajne promenljive X jednako

$$E(X) = x_1p_1 + x_2p_2 + \cdots x_np_n = \sum_{k=1}^{n} x_kp_k.$$

Ako je X neprekidna slučajna promenljiva sa gustinom $y = g(x)$ $(-\infty < x < +\infty)$ njeno matematičko očekivanje je jednako

$$E(X) = \int_{-\infty}^{+\infty} xg(x)dx$$

t.j. jednako je površini ograničenom krivom $y = xg(x)$ i x-osom.

Primer 1.11. Neka je X slučajna promenljiva definisana kao broj koji pokazuje kocka prilikom bacanja. Raspodela ove slučajne promenljive data je sa

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \end{pmatrix}$$

Njeno matematičko očekivanje je

$$E(X) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = \frac{7}{2} = 3,5$$

.

Primer 1.12. Događaj A ostvaruje se sa verovatnoćom $\frac{1}{4}$. Neki čovek se kladi na taj događaj na sledeći način. On ulaže 1 dinar, s tim što gubi svoj ulog ako se događaj A ne ostvari, a dobija 3 dinara (dakle svoj ulog i još dva dinara) ako se događaj A ostvari. Da li je umesno kladiti se na ovaj način?

Rešenje. Slučajnu promenljivu X definišemo kao broj dobijenih dinara. Dobitak može da iznosi 2 dinara ili -1 dinar. Dakle, X uzima vrednosti iz skupa $\{2, -1\}$ i ima raspodelu verovatnoća

$$\begin{pmatrix} 2 & -1 \\ \frac{1}{4} & \frac{3}{4} \end{pmatrix}$$

Matematičko očekivanje je $E(X) = 2 \cdot \frac{1}{4} + (-1) \cdot \frac{3}{4} = -\frac{1}{4}$. Znači, nije umesno kladiti se na opisani način.

Jedna slučajna promenljiva X je potpuno određena svojom raspodelom. Matematičko očekivanje, kao "srednja vrednost" slučajne promenljive je važna informacija o slučajnoj promenljivoj, ali ne može da zameni kompletnu informaciju koju daje raspodela. Pre svega, matematičko očekivanje ne daje podatak o raspršivanju mogućih vrednosti slučajne promenljive oko "srednje vrednosti". Na primer, slučajne promenljive X i Y sa raspodelama

$$X : \begin{pmatrix} -1 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix} \qquad Y : \begin{pmatrix} -100 & 100 \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

imaju $E(X) = E(Y) = 0$, ali je raspršivanje oko 0 veće kod Y nego kod X. Jedna od mera raspršivanja ,koja inače ima najveći teorijski i praktični značaj, poznata je pod nazivom varijansa (disperzija) i obeležava se sa $D(X)$.

Definicija 1.6. *Varijansa (disperzija), u oznaci $D(X)$, slučajne promenljive X definiše se jednakošću*

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

Može se pokazati da je

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

Ova druga formula za disperziju je često lakša nego prva.

Definicija 1.7. *Standardno odstupanje slučajne promenljive X, u oznaci $\boldsymbol{s}$, definisano je jednakošću $\boldsymbol{s} = \sqrt{D(X)}$.*

Primer 1.13. Izračunajmo disperziju i standardno odstupanje slučajne promenljive X sa raspodelom

$$\begin{pmatrix} 0 & 1 & 2 \\ \frac{1}{2} & \frac{3}{8} & \frac{1}{8} \end{pmatrix}$$

Rešenje.

$$\begin{aligned} E(X) &= 0 \cdot \tfrac{1}{2} + 1 \cdot \tfrac{3}{8} + 2 \cdot \tfrac{1}{8} = \tfrac{5}{8} \\ D(X) &= E((X - \tfrac{5}{8})^2) = (0 - \tfrac{5}{8})^2 \cdot \tfrac{1}{2} + (1 - \tfrac{5}{8})^2 \cdot \tfrac{3}{8} + (2 - \tfrac{5}{8})^2 \cdot \tfrac{1}{8} \\ &= \tfrac{25}{64} \cdot \tfrac{1}{2} + \tfrac{9}{64} \cdot \tfrac{3}{8} + \tfrac{121}{64} \cdot \tfrac{1}{8} = \tfrac{31}{64} \\ \mathbf{s} &= \sqrt{\tfrac{31}{64}} = \tfrac{\sqrt{31}}{8}. \end{aligned}$$

Sada navodimo, bez dokaza, matematička očekivanja i disperzije za neke slučajne promenljive sa poznatim raspodelama.

1. Ako slučajna promenljiva X ima binomnu raspodelu $B(n,p)$, tada je

$$E(X) = n \cdot p,\ D(X) = n \cdot p \cdot q.$$

2. Ako slučajna promenljiva X ima Puasonovu raspodelu $\mathcal{P}(\lambda)$, tada je

$$E(X) = \lambda,\ D(X) = \lambda.$$

3. Ako neprekidna slučajna promenljiva X ima normalnu raspodelu $N(\mu, \sigma^2)$, tada je

$$E(X) = \mu,\ \ D(X) = \sigma^2.$$

2

Deskriptivna statistika i karakteristike empirijske raspodele

2.1 Uvod

Statistika se bavi prikupljanjem određenih podataka, obradom tih podataka i donošenjem odluka (zaključaka) na osnovi dobijenih rezultata.

Posmatrajmo neki skup S. U matematičkoj statistici takav skup nazivamo **osnovni skup**, **statistički skup** ili **populacija**. Osobine elemenata određene populacije nazivaju se statističkim obeležjima. Obeležja mogu biti **opisna** (kategorijska, kvalitativna) ili **numerička**. Opisna obeležja mogu biti nominalna i ordinalna. Nominalna obeležja su neuređena (na primer, krvne grupe A, B, AB i 0), dok su ordinalna obeležja na neki način uređena (na primer, slab, umeren, jak, vrlo jak). "Vrednosti" opisnih obeležja se nazivaju kategorije ili modaliteti. Ako opisno obeležje ima samo dve "vrednosti" (da - ne, pušač - nepušač i slično) onda se naziva binarno ili dihotomno. Obeležja koja se izražavaju brojčano nazivaju se numerička obeležja. Naime, ako svakom elementu osnovnog skupa S dodelimo jedan realan broj, time je određeno preslikavanje $X : S \to R$ koje se naziva numeričko obeležje skupa S. Ako X preslikava skup S u konačan ili beskonačan niz realnih brojeva, onda je to obeležje diskretno (broj članova porodice, broj vizita i slično). Ako X preslikava skup S u neki interval realnih

brojeva onda kažemo da je to obeležje neprekidno (visina, težina, nivo šećera u krvi i slično).

Bilo koji podskup skupa S naziva se uzorak. Ako jedan element biramo slučajno iz populacije, onda populaciju možemo shvatiti kao skup svih elementarnih događaja. Kako se svakom elementu populacije dodeljuje jedan broj, njegovo obeležje, to obeležje je jedna slučajna promenljiva X.

Ako slučajno biramo n elemenata iz neke populacije onda imamo n-dimenzionalnu slučajnu promenljivu $(X_1, \ldots, X_n)$, koja se naziva i **slučajni uzorak** obima n. Ako su slučajne promenljive $X_1, \ldots, X_n$ nezavisne i sve sa istom raspodelom kao i slučajna promenljiva X, tada se takav uzorak naziva prost slučajni uzorak.

Često se radi sa raznim funkcijama slučajnog uzorka $(X_1, \ldots, X_n)$. Funkcija oblika $Z = f(X_1, \ldots, X_n)$ naziva se **statistika**. Napomenimo da je statistika slučajna promenljiva.

Jedan od zadataka statistike se sastoji u tome da se ispitivanje nekog obeležja u populaciji zameni ispitivanjem tog obeležja na uzorku i da se na osnovi osobina tog uzorka donesu određeni zaključci ili predviđanja za čitavu populaciju. Osnovni problem je da se, na osnovi uzorka, odredi raspodela verovatnoća slučajne promenljive X.

Predviđanja na osnovi uzoraka ne moraju uvek biti pouzdana, ali je praksa pokazala da su statističke metode veoma korisne, pa je statistika prisutna u mnogim oblastima ljudske delatnosi.

Podaci, koji predstavljaju vrednosti posmatranog obeležja za elemente populacije ili uzorka, mogu se srediti tako da raspodela njihovih obeležja bude podesna i pregledna. Obično se u te svrhe koristi tablični i grafički način prikazivanja. Neka posmatrano obeležje X uzima vrednosti $x_1, x_2, \ldots, x_k$, pri čemu je $x_1 < x_2 < \cdots < x_k$. Neka se u populaciji od N elemenata vrednosti obeležja $x_1, x_2, \ldots, x_k$ pojavljuju redom $f_1, f_2, \ldots, f_k$ puta. Brojevi $f_1, f_2, \ldots, f_k$, koji se zovu apsolutne frekvencije (učestanosti) vrednosti obeležja $x_1, x_2, \ldots, x_k$ zadovoljavaju, prirodno, uslov $f_1 + f_2 + \cdots + f_k = N$. Često se koriste i relativne frekvencije $r_1 = \frac{f_1}{N}, r_2 = \frac{f_2}{N}, \ldots, r_k = \frac{f_k}{N}$ koje, očigledno, zadovoljavaju uslov $r_1 + r_2 + \cdots + r_k = 1$. Vrednosti obeležja $x_1, x_2, \ldots, x_k$ sa odgovarajućim frekvencijama formiraju statističku tabelu

X	x_1	x_2	$\cdots$	x_k	Σ
f_i	f_1	f_2	$\cdots$	f_k	N
f_i	r_1	r_2	$\cdots$	r_k	1

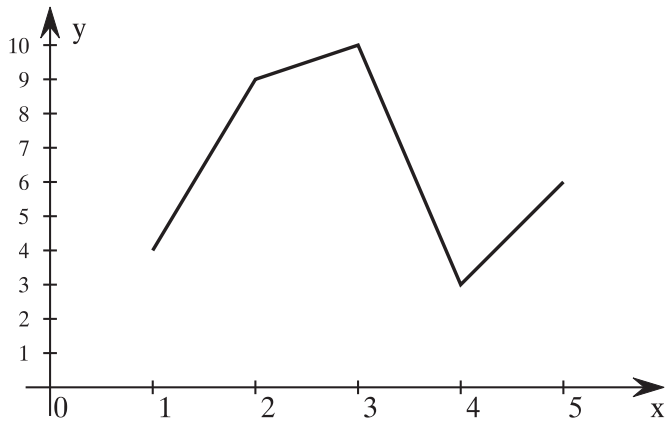
Tabela 2.1.

Podatke iz tabele predstavljamo u koordinatnom sistemu unošenjem tačaka čije su apscise vrednosti obeležja, a ordinate frekvencije vrednosti obeležja (apsolutne ili relativne). Spajanjem dobijenih tačaka dobijamo izlomljenu liniju koja se naziva **poligon raspodele učestanosti**.

Primer 2.1. U jednom odeljenju ima 6 petica iz biologije, 3 četvorke, 10 trojaka, 9 dvojaka i 4 jedinice. U Tabeli 2.2. su prikazane apsolutne i relativne frekvencije ocena iz biologije dok je poligon raspodele frekvencija ocena prikazan na Slici 2.1.

X	1	2	3	4	5	Σ
f_i	4	9	10	3	6	32
r_i	$\frac{4}{32}$	$\frac{9}{32}$	$\frac{10}{32}$	$\frac{3}{32}$	$\frac{6}{32}$	$\frac{32}{32}$

Tabela 2.2.



Slika 2.1.

Kako su brojevi f_i i r_i proporcionalni t.j. $\frac{f_i}{r_i} = N$ $(i = 1, 2, \ldots, k)$ poligone

raspodele apsolutnih i relativnih frekvencija možemo predstaviti istim tačkama, s tim što treba pogodno odabrati jedinice mere na ordinatnoj osi.

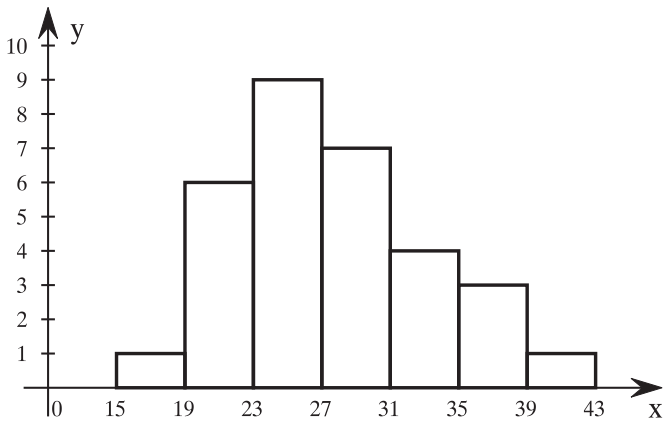
Često je broj vrednosti koje neko obeležje uzima veliki. Tada se interval [a,b], unutar koga se nalaze sve posmatrane vrednosti obeležja, deli na klase. Naime, ovaj interval se podeli na određeni broj k (najčešće jednakih) podintervala: $[a, a_1), [a_1, a_2), \ldots, [a_{k-1}, b]$. Frekvencije $f_1, f_2, \ldots, f_k$ sada označavaju koliko vrednosti obeležja pada u prvu, drugu,..., k-tu klasu. Grafički se ova podela predstavlja tako što se nad podintervalom (klasom) crta pravougaonik sa visinom jednakom frekvenciji (učestanosti) podataka u toj klasi. Ovako dobijeni dijagram se zove histogram frekvencija. Ako se tačke $(x_1, f_1), (x_2, f_2), \ldots, (x_k, f_k)$, gde su $x_1, x_2, \ldots, x_k$ sredine klasa $[a, a_1), [a_1, a_2), \ldots, [a_{k-1}, b]$, spoje dužima, dobija se poligon raspodele frekvencija. Klase su obično jednake dužine a njihov broj se određuje tako da se što lakše i jasnije uoči raspodela frekvencija posmatranog obeležja.

Primer 2.2. Populaciju čine životinje jedne vrste na jednoj farmi a obeležje je težina životinja data u intervalima težine 4 kg. Raspodela frekvencija težina životinja data je u Tabeli 2.3. a odgovarajući histogram raspodele na Slici 2.2.

X	$15-19$	$19-23$	$23-27$	$27-31$	$31-35$	$35-39$	$39-43$
f_i	1	6	9	7	4	3	1

Tabela 2.3.

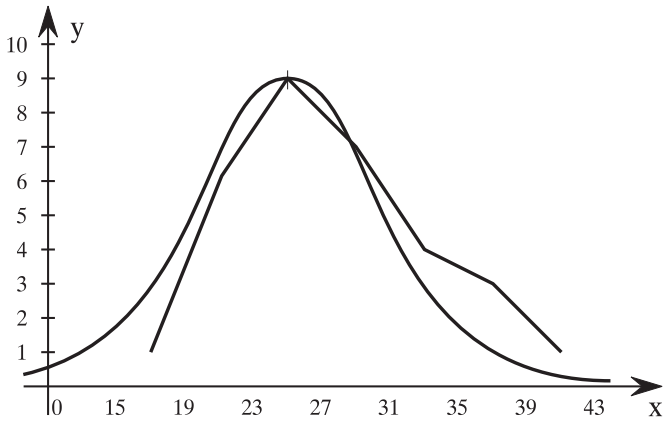
Neka je X neprekidno obeležje čije se sve vrednosti nalaze u intervalu $[a, b]$. Podelom ovog intervala na k klasa možemo da nacrtamo odgovarajući histogram frekvencija kao i poligon raspodele frekvencija. Ako uzmemo gušću podelu intervala $[a, b]$, odgovarajući poligon raspodele biće sastavljen od većeg broja manjih duži. Kada je broj podeonih tačaka veliki, poligon raspodele liči na glatku krivu liniju, koja se naziva kriva raspodele frekvencija. Od velikog je značaja da empirijsku izlomljenu liniju zamenimo nekom teorijskom, neprekidnom krivom, koja se u što većoj meri prilagođava datom histogramu, odnosno poligonu. U praksi se dosta često sreću neprekidna obeležja čije vrednosti imaju raspodelu frekvancija veoma blisku normalnoj raspodeli. Ovo je u vezi sa činjenicom da ima najviše "prosečnih", a znatno manje "ekstremnih" slučajeva. Tako, na primer, podaci o težinama životinja ukazuju da je najveći broj životinja "sred-



Slika 2.2.

nje" težine, a da je mnogo manji broj lakih životinja i teških životinja.

Na Slici 2.3. prikazan je poligon raspodele relativnih frekvencija iz Primera 2.2. kao i odgovarajuća kriva (normalne) raspodele.



Slika 2.3.

Raspodela frekvencija jednog obeležja daje dobre mogućnosti za analizu karakteristika tog obeležja. Statističke metode nam služe da na podesan način ceo skup zamenimo jednim relativno malim podskupom koji će reprezentovati ceo skup i sadržati najveći mogući deo informacije sadržane u početnom skupu.

Često je potrebno uporediti dva ili više uzoraka, odnosno populacija u odnosu na neko obeležje. Ako su vrednosti obeležja tih uzoraka podeljene u iste grupe, odnosno intervale, onda se uzorci porede tako što se porede odgovarajuće apsolutne frekvencije (u slučaju da uzorci imaju jednake obime) ili relativne frekvencije (u slučaju da uzorci imaju različite obime). Međutim, ako se grupe, odnosno intervali, ne poklapaju, onda treba naći broj koji bi, na neki način, zamenio čitavu raspodelu frekvencija jednog uzorka, što bi omogućilo poređenje dva ili više uzoraka, odnosno populacija. Takav broj se naziva srednja vrednost.

Srednja vrednost obeležja je takva vrednost obeležja koja, na neki način, reprezentuje čitav skup i omogućava upoređivanje između raznih skupova. Često su vrednosti obeležja tako raspoređene da se njihove frekvencije grupišu negde oko sredine, između najveće i najmanje vrednosti obeležja. Ukoliko su udaljenije od te vrednosti, frekvencije su manje. Takva vrednost postaje reprezentativna za ceo skup podataka i naziva se srednja vrednost. Ona se može odrediti prema različitim kriterijumima. To može da bude aritmetička, geometrijska, harmonijska ili neka druga sredina. U statistici se najviše koristi aritmetička sredina.

2.2 Aritmetička sredina

Ako je X dato obeležje a $x_1, x_2, \ldots, x_N$ vrednosti tog obeležja u populaciji, pri čemu je N broj elemenata populacije, tada se aritmetička sredina obeležja X populacije definiše sa

$$\mu = \frac{1}{N}(x_1 + x_2 + \cdots + x_N) = \frac{1}{N}\sum_{i=1}^{N} x_i.$$

Retko kada raspolažemo sa vrednostima obeležja čitave populacije, već obično radimo sa uzorcima.

Ako je X posmatrano obeležje, a $(X_1, \ldots, X_n)$ prost slučajni uzorak od n elemenata tada se statistika

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

naziva aritmetička sredina uzorka $(X_1, \ldots, X_n)$.

Ako smo iz neke populacije izabrali jedan uzorak, onda vrednosti obeležja $x_1, \ldots, x_n$ tog izabranog uzorka predstavljaju realizovane vrednosti slučajne promenljive $(X_1, \ldots, X_n)$. Aritmetička sredina slučajnog uzorka $\overline{X}_n$ je slučajna promenljiva, dok je aritmetička sredina izabranog uzorka

$$\overline{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i$$

konstanta.

Primer 2.3. Ako je iz jednog ribnjaka uzet uzorak od pet riba i konstatovano da imaju, redom, dužine 30, 28, 32, 29, 31 cm, tada je prosečna dužina riba ovog uzorka

$\overline{x}_n = \frac{1}{5}(30 + 28 + 32 + 29 + 31) = 30.$

Ako treba da izračunamo aritmetičku sredinu brojeva 4, 4, 4, 5, 5, 5, 5, 5, 5, 6, 6, 7, 7, 7, 7, onda nije uputno, na primer, broj 5 sabirati 6 puta, već pomnožiti brojeve 5 i 6, odnosno aritmetičku sredinu izračunati kao

$$\overline{x}_n = \frac{1}{15}(3 \cdot 4 + 6 \cdot 5 + 2 \cdot 6 + 4 \cdot 7) = \frac{112}{15} = 7,47.$$

Uopšte, ako se vrednosti obeležja $x_1, x_2, \ldots, x_k$ javljaju sa različitim apsolutnim frekvencijama $f_1, f_2, \ldots, f_k$ u takvim slučajevima aritmetička sredina se izračunava prema formuli

$$\overline{x}_n = \frac{1}{n} \sum_{i=1}^{k} f_i x_i,$$

gde je $n = f_1 + \cdots + f_k$, i naziva se ponderisana aritmetička sredina.

Ako su podaci dati po intervalima, onda se aritmetička sredina izračunava kao ponderisana aritmetička sredina, pri čemu se za $x_1, \ldots, x_k$ uzimaju sredine intervala.

Pored aritmetičke sredine, nekada se kao srednja vrednost koristi geometrijska sredina, prema formuli

$$g = \sqrt[n]{x_1 \cdot x_2 \cdots x_n}$$

i harmonijska sredina, prema formuli

$$h = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots \frac{1}{x_n}}$$

.

2.3 Medijana

Nekada, naročito kada vrednosti obeležja dosta odstupaju od aritmetičke sredine ili ako raspodela nije normalna, aritmetička sredina ne predstavlja dobar reprezent tog obeležja.

Primer 2.4. U jednom preduzeću plate radnika (u hiljadama) su

$$7, 8, 10, 11, 12, 12, 94.$$

Aritmetička sredina je

$$\overline{x}_n = \frac{1}{7}(7 + 8 + 10 + 11 + 12 + 12 + 94) = 22.$$

Očigledno je da dobijamo pogrešnu predstavu o prosečnoj plati u ovom preduzeću ako kažemo da su plate oko 22.000. Mnogo bolju predstavu o prosečnoj plati u ovom preduzeću imaćemo ako umesto aritmetičke sredine koristimo medijanu.

Definicija 2.1. *Medijana je ona vrednost obeležja koja se nalazi u sredini niza vrednosti obeležja poređanih u rastući poredak.*

Prilikom određivanja medijane razlikujemo slučajeve kada je broj članova niza n neparan i paran. Ako je n neparan, tada srednji član (medijana) deli ovaj niz na dva jednaka dela. U primeru sa platama, u nizu $7, 8, 10, 11, 12, 12, 94$ taj srednji član, odnosno medijana je 11. U slučaju kada je n paran, u uređenom nizu vrednosti obeležja postoje dva srednja člana. U tom slučaju se uzima aritmetička sredina ta dva srednja člana. Tako, na primer, ako je dat niz podataka $4, 5, 8, 10, 14, 19$, uzima se da je medijana $\frac{8+10}{2} = 9$.

Može se reći da je medijana uzorka $(X_1, \ldots, X_n)$ statistika $Me(X)$ definisana na sledeći način:

$$Me(X) = \begin{cases} X_{\frac{n+1}{2}}, & \text{ako je } n \text{ neparno} \\ \frac{1}{2}(X_{\frac{n}{2}} + X_{\frac{n}{2}+1}), & \text{ako je } n \text{ parno} \end{cases}$$

Primetimo da u primeru sa platama medijana, koja je 11.000, mnogo bolje reprezentuje plate nego aritmetička sredina, koja je 22.000.

Ako se niz podataka, rangiranih po veličini, podeli u četiri jednaka dela, vrednosti obeležja koje ih dele nazivaju se **kvartili**: prvi kvartil Q_1, drugi kvartil Q_2 (medijana) i treći kvartil Q_3. Prvi kvartil Q_1 je vrednost obeležja od koje 25% elemenata skupa uređenih po veličini ima manju ili jednaku vrednost. Treći kvartil Q_3 je ona vrednost obeležja od koje 75% elementa skupa ima manju ili jednaku vrednost.

Neka, na primer, imamo uređeni niz brojeva

$3, 5, 7, 8, 10, 12, 15, 17, 19, 20, 25, 27, 30, 31, 34, 36$

Ovaj niz ima 16 članova. Medijana, odnosno drugi kvartil Q_2, je broj 18 (aritmetička sredina brojeva 17 i 19), jer se ispred tog broja nalazi osam t.j. 50% članova niza. Prvi kvartil Q_1 je broj 9, koji se dobija kao aritmetička sredina $\frac{8+10}{2}$ i 25% članova niza (t.j. četiri člana) je manje od broja 9. Treći kvartil Q_3 je 28,5 (dobija se kao $\frac{27+30}{2}$) i 75% članova niza (t.j. dvanaest članova) je manje od tog broja.

2.4 Standardno odstupanje

Prilikom izračunavanja srednjih vrednosti često se zapaža da unutar jednog uzorka postoje velike razlike u vrednostima obeležja koja se posmatraju. Ove razlike nastaju usled delovanja raznih faktora. Ta promenljivost posmatranih vrednosti u jednom uzorku naziva se varijabilitetom i može biti veći ili manji, što zavisi od homogenosti uzorka. Iako srednja vrednost može da da izvesnu sliku o nizu datih vrednosti ili o raspodeli frekvencija, ona nije u mogućnosti da bliže opiše pojedinačne varijabilitete u posmatranom uzorku. Kao ilustracija mogu da posluže dva uzorka:

$\{30, 60, 90\}$ i $\{0, 60, 120\}$.

Očigledno, oba uzorka imaju aritmetičku sredinu 60, ali na osnovi aritmetičke sredine ne može da se dobije slika varijabiliteta podataka koji su rasuti oko nje. Da bi mogao da se meri varijabilitet t.j. grupisanost podataka oko aritmetičke sredine, mora da se odredi koliko se svaki podatak razlikuje, odnosno odstupa od aritmetičke sredine. Za obeležje X jedne populacije čije su vrednosti $x_1, x_2, \ldots, x_N$ i čija je aritmetička sredina μ, razlike $x_1 - \mu, x_2 - \mu, \ldots, x_N - \mu$ nazivaju se odstupanja vrednosti obeležja od njihove aritmetičke sredine. Što su odstupanja manja, podaci su jednoličniji, t.j. srednja vrednost obeležja bolje reprezentuje skup. Može da se pokaže da je zbir odstupanja pojedinačnih vrednosti obeležja od aritmetičke sredine jednak nuli. Da bi se izrazila ukupna odstupanja, uvode se razni pokazatelji odstupanja. Aritmetička sredina apsolutnih vrednosti odstupanja

$$A_N = \frac{|x_1 - \mu| + |x_2 - \mu| + \cdots |x_N - \mu|}{N} = \frac{1}{N} \sum_{i=1}^{N} |x_i - \mu|$$

predstavlja srednje apsolutno odstupanje vrednosti $x_1, x_2, \ldots, x_N$ od aritmetičke sredine μ. Da bi se izbegla računanja sa apsolutnim vrednostima korisnije je za meru odstupanja vrednosti obeležja X uzeti srednji kvadrat odstupanja vrednosti $x_1, x_2, \ldots, x_N$ od μ.

Ako su vrednosti obeležja X u populaciji $x_1, x_2, \ldots, x_N$, gde je N broj elemenata populacije, a aritmetička sredina tih vrednosti μ, tada se

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2$$

naziva **varijansa** obeležja X populacije. Kako μ i σ^2 imaju različite dimenzije (ako je μ izraženo u centimetrima, onda je σ^2 izraženo u centimetrima na kvadrat), često se koristi kvadratni koren iz σ^2 t.j

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2}$$

koji se naziva **standardno odstupanje**. Veličina σ pokazuje koliko vrednosti obeležja, u proseku, odstupaju od aritmetičke sredine.

Ako je X posmatrano obeležje a $(X_1, \ldots, X_n)$ prost slučajni uzorak obima n, čija je aritmetička sredina $\overline{X}$, tada se statistika

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

naziva **disperzija** ili **varijansa** uzorka $(X_1, \ldots, X_n)$, a

$$S_n = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2}$$

se naziva **standardno odstupanje** uzorka $(X_1, \ldots, X_n)$.

Standardno odstupanje izabranog uzorka $(x_1, \ldots, x_n)$ je

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

Primer 2.5. Iz populacije bolesnika koji boluju od jedne bolesti uzet je uzorak od šestoro ispitanika i meren im broj leukocita u krvi (u hiljadama): 17, 18, 20, 22, 24 i 25. Izračunati standardno odstupanje broja leukocita ovog uzorka.

Rešenje. Aritmetička sredina je

$$\overline{x}_n = \frac{1}{6}(17 + 18 + 20 + 22 + 24 + 25) = 21$$

a varijansa

$$\begin{aligned} s_n^2 &= \tfrac{1}{6-1}\left((17-21)^2 + (18-21)^2 + (20-21)^2 + (22-21)^2\right. \\ &+ \left.(24-21)^2 + (25-21)^2\right) = \tfrac{52}{5} = 10,40. \end{aligned}$$

odakle je $s_n = \sqrt{10,40} = 3,22$.

Ako se vrednosti obeležja $x_1, x_2, \ldots, x_k$ javljaju sa različitim apsolutnim frekvencijama $f_1, f_2, \ldots, f_k$, u takvim slučajevima varijansa se izračunava prema formuli:

$$s_n = \sqrt{\frac{1}{n-1} \sum_{i=1}^{k} f_i (x_i - \overline{x}_n)^2},$$

gde je $n = f_1 + \cdots + f_k$.

Za izračunavanje varijanse može se koristiti i jednostavnija formula. Naime, imamo

$$\begin{aligned} s_n^2 &= \tfrac{1}{n-1}\textstyle\sum_{i=1}^k f_i(x_i - \overline{x}_n)^2 = \tfrac{1}{n-1}(\sum_{i=1}^k f_i x_i^2 \,-\, 2\cdot\overline{x}_n \sum_{i=1}^k f_i x_i \,+ (\overline{x}_n)^2 \sum_{i=1}^k f_i) \\ &= \tfrac{1}{n-1}(\textstyle\sum_{i=1}^k f_i x_i^2 - 2\overline{x}_n \cdot n \cdot \overline{x}_n \,+\, n\cdot(\overline{x}_n)^2) = \tfrac{1}{n-1}(\sum_{i=1}^k f_i x_i^2 - n(\overline{x}_n)^2), \end{aligned}$$

odakle se dobija

$$s_n = \sqrt{\frac{1}{n-1}(\sum_{i=1}^{k} f_i x_i^2 - n(\overline{x}_n)^2)}.$$

Primetimo da u slučaju kada se ne radi sa frekvencijama poslednja formula postaje

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

Primer 2.6. Odrediti standardno odstupanje s_n za skup brojeva 2, 2, 2, 3, 4, 4, 4, 4, 4, 5, 5, 5.

Rešenje. Aritmetička sredina ovih brojeva je

$$\overline{x}_n = \frac{1}{12}\,(3\cdot 2 + 1\cdot 3 + 5\cdot 4 + 3\cdot 5) = \frac{44}{12} = 3,67$$

dok je varijansa

$$s_n^2 = \frac{1}{12-1}\,(3\cdot 2^2 + 1\cdot 3^2 + 5\cdot 4^2 + 3\cdot 5^2 - 12\cdot(3,67)^2) = \frac{1}{11}\,(176 - 161,63) = 1,31.$$

Otuda je $s_n \approx 1,14$.

2.5 Standardna greška aritmetičke sredine

Neka je X obeležje neke populacije a μ aritmetička sredina tog obeležja u populaciji. Uzmimo uzorak obima n i izračunajmo njegovu aritmetičku sredinu. Ako ponovimo ovaj postupak više puta dobićemo razne vrednosti aritmetičkih sredina. Ako bi se nacrtao histogram vrednosti aritmetičkih sredina, videlo bi se da on ima zvonast oblik.

Ako je broj uzoraka relativno veliki (recimo veći od 30), raspodela aritmetičkih sredina obeležja X je približno normalna, bez obzira na raspodelu obeležja X u populaciji. Ako je broj uzoraka mali a obeležje X ima normalnu raspodelu, onda raspodela aritmetičkih sredina obeležja X ima približno normalnu raspodelu,

Može se pokazati da je aritmetička sredina aritmetičkih sredina obeležja X svih uzoraka obima n koji se mogu izabrati iz jedne populacije jednaka aritmetičkoj sredini μ obeležja populacije.

Varijabilitet aritmetičkih sredina se meri varijansom $\sigma^2_{\overline{x}}$. Ako se zna varijansa σ^2 obeležja X, onda je

$$\sigma^2_{\overline{x}} = \frac{\sigma^2}{n}$$

odnosno varijansa aritmetičkih sredina uzoraka se dobija kao količnik varijanse obeležja X i obima uzorka. Standardna devijacija aritmetičkih sredina uzoraka je

$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$$

i naziva se **standardna greška aritmetičke sredine**. Obeležava se sa SEM. Ako se ne zna varijansa obeležja X, tada se standardna greška aritmetičke sredine ocenjuje sa

$$\sigma_{\overline{x}} = \frac{s_n}{\sqrt{n}},$$

gde je s_n standardna devijacija izračunata iz uzorka.

Mala standardna greška aritmetičke sredine ukazuje da je ocena aritmetičke sredine obeležja osnovnog skupa pomoću aritmetičke sredine uzorka dosta precizna.

Standardna devijacija pokazuje variranje vrednosti obeležja, dok standardna greška opisuje preciznost uzoračke aritmetičke sredine.

2.6 Koeficijent varijacije

Ako su iz dva uzorka izračunati

$$\overline{x}_1 = 100 \text{ cm}, \qquad s_1 = 10 \text{ cm},$$
$$\overline{x}_2 = 10 \text{ cm}, \qquad s_2 = 2 \text{ cm},$$

teško je na prvi pogled ustanoviti koji podaci relativno više variraju. Budući da je $s_1 > s_2$, može se, u prvi mah, zaključiti da podaci iz prvog uzorka više variraju. Međutim, s_1 iznosi samo 10% od odgavarajuće aritmetičke sredine, dok s_2 iznosi 20%. Iz ovih razloga se definiše koeficijent varijacije

$$C_v = \frac{s}{\overline{x}} \cdot 100\%.$$

To je relativna mera odstupanja i pokazuje koje se obeležje više menja u odnosu na aritmetičku sredinu. Koeficijent varijacije se upotrebljava za merenje promenljivosti različitih obeležja (na primer, visine i težine) ili istih obeležja sa različitim aritmetičkim sredinama. Koeficijent varijacije se ne preporučuje kada je aritmetička sredina blizu nule. Pored toga što je koeficijent varijacije mera promenljivosti, često se upotrebljava kao mera za homogenost. Naime, što je koefijent varijacije manji , homogenost statističkog skupa je veća, i obratno, što je koeficijent varijacije veći, veće je i raspršivanje oko aritmetičke sredine. Obično se smatra da je neka pojava homogena ako je vrednost koeficijenta varijacije do 30%. U suprotnom, kaže se da je pojava nehomogena.

Primer 2.7. Merenjem visine i težine jedne grupe studenata dobijena je prosečna visina $\overline{x} = 180$ cm sa standardnim odstupanjem $s_1 = 5,4$ cm dok je prosečna težina bila $\overline{y} = 80$ kg sa standarnim odstupanjem $s_2 = 4$ kg. Ispitati koji podaci više variraju.

Rešenje. Kako je $C_{v1} = \frac{5,4}{180} \cdot 100\% = 3\%$ a $C_{v2} = \frac{4}{80} \cdot 100\% = 5\%$,
više variraju podaci o težini nego podaci o visini.

2.7 Proporcija

Nekada je za populaciju važan odnos broja elemenata populacije koji imaju određenu osobinu $\mathcal{O}$ prema ukupnom broju elemenata populacije. Taj odnos se

naziva **proporcija** ili relativna frekvencija i obeležava se sa π. Imajući u vidu Definiciju 1.2, može se reći da je π verovatnoća da neki element populacije ima osobinu $\mathcal{O}$. Na primer, od 320 radnika jednog preduzeća 112 su pušači. Ovde je proporcija $\pi = \frac{112}{320} = 0,35$ t.j. 35% radnika preduzeća su pušači.

Ako ne znamo proporciju π (koja se odnosi na određenu osobinu) populacije, onda je ocenjujemo proporcijom uzetog uzorka

$$p = \frac{m}{n},$$

gde je n broj elemenata uzorka a m broj elemenata uzorka koji imaju osobinu $\mathcal{O}$. Ako uzimamo uzorak od n elemenata više puta, dobijamo razne proporcije. Raspodela tih proporcija je približno jednaka normalnoj rasodeli, gde je aritmetička sredina tih proporcija jednaka proporciji π osnovnog skupa. Standardna devijacija ove raspodele se naziva **standardna greška proporcije**. Ako uzmemo jedan uzorak obima n, onda se standardna greška proporcije ocenjuje sa:

$$SE(p) = \sqrt{\frac{p(1-p)}{n}}.$$

Standardna greška proporcije služi kao mera preciznosti ocene za π. Mala standardna greška ukazuje na precizniju ocenu.

2.8 Intervali poverenja

Ako nije poznata aritmetička sredina nekog obeležja X populacije, onda je ocenjujemo aritmetičkom sredinom uzorka. Neka obeležje X ima normalnu raspodelu. Interval poverenja za aritmetičku sredinu je

$$[\overline{x}_n - t_{n-1;\,1-\beta} \cdot \frac{s_n}{\sqrt{n}}\,,\ \overline{x}_n + t_{n-1;\,1-\beta} \cdot \frac{s_n}{\sqrt{n}}],$$

gde su $\overline{x}$ i s_n aritmetička sredina i standardna devijacija obeležja izračunati iz uzorka, a $t_{n-1;\,1-\beta}$ vrednost koja se nalazi u tabeli Studentove raspodele.

Primer 2.8. Neka je na osnovi uzorka od 8 elemenata izračunata aritmetička sredina $\overline{x} = 556,25$ i i standardna devijacija $s_n = 35,03$ nekog obeležja X. Odrediti 95% interval poverenja za aritmetičku sredinu ovog obeležja X.

Rešenje. Kako je $t_{n-1;\,1-\beta} = t_{8-1;\,1-0,95} = t_{7;\,0,05} = 2,365$ to je interval poverenja

$$[556,25 - 2,365 \cdot \frac{35,03}{\sqrt{8}};\; 556,25 + 2,365 \cdot \frac{35,03}{\sqrt{8}}] = [526,98;\; 585,52].$$

Ovo znači da se, sa verovatnoćom od 95%, aritmetička sredina obeležja X osnovnog skupa nalazi u intervalu $[526,98;\; 585,52]$.

Ako nije poznata proporcija π (koja se odnosi na određenu osobinu) osnovnog skupa, onda je ocenjujemo proporcijom uzetog uzorka $p = \frac{m}{n}$. **Interval poverenja za proporciju** π osnovnog skupa je

$$\Big[p - c_\beta \cdot \sqrt{\frac{p\,(1-p)}{n}}\,;\; p + c_\beta \cdot \sqrt{\frac{p\,(1-p)}{n}}\;\Big]$$

gde je $c_\beta = 1,96$, ako je $\beta = 0,95$ a $c_\beta = 2,58$, ako je $\beta = 0,99$.

Intervali poverenja se određuju i za druge parametre raspodele.

3

Testiranje statističkih hipoteza

Statistička hipoteza je tvrđenje ili pretpostavka o nekoj važnoj osobini jednog ili više skupova. Postupak verifikovanja hipoteze pomoću uzorka naziva se (statistički) test.

Ako se hipoteza odnosi na parametre raspodele (aritmetička sredina, standardna devijacija), onda je to parametarski test. U ostalim slučajevima testovi su neparametarski. Osnovni zadatak, kod provere statističkih hipoteza, je određivanje pravila, odnosno kriterijuma po kome se, na osnovi eksperimentalnih podataka, odnosno uzorka, može odgovoriti na pitanje da li se hipoteza prihvata ili odbacuje.

Ako, na primer, želimo da proverimo da li je prosečna težina boca fiziološkog rastvora jedne fabrike (zajedno sa ambalažom) 600 grama, onda možemo postaviti hipotezu da je $\mu = 600$, gde je μ aritmetička sredina težina boca fiziološkog rastvora u toj fabrici. Takva hipoteza se naziva nulta hipoteza i označava sa H_0, t.j. $H_0(\mu = 600)$.

Proizvoljna druga hipoteza, koja se razlikuje od nulte, naziva se alternativna hipoteza i označava se sa H_1. Ako je $H_0(\mu = 5)$, alternativna hipoteza može biti $H_1(\mu \neq 600)$, $H_1(\mu > 600)$, $H_1(\mu < 600)$, $H_1(\mu = 500)$ i slično.

3.1 Testovi značajnosti

Postupak testiranja statističkih hipoteza se vrši u šest koraka

- formulisanje nulte i alternativne hipoteze

- izbor statistike testa i određivanje njene raspodele verovatnoća

- izbor praga značajnosti testa

- formulisanje pravila pri kome se odbacuje ili prihvata nulta hipoteza

- uzimanje uzorka i izračunavanje vrednosti statistike testa

- donošenje odluke o odbacivanju ili prihvatanju nulte hipoteze.

Ako je nulta hipoteza H_0 tačna i ako je postupkom testiranja, na osnovi uzetog uzorka, ona prihvaćena, onda je zaključak testiranja ispravan. Međutim, može se dogoditi da je nulta hipoteza H_0 tačna, a da je postupkom testiranja odbačena. Tada je zaključak pogrešan a greška koja je napravljena naziva se **greška prvog tipa**. Ako se prihvati netačna nulta hipoteza onda se čini **greška drugog tipa**.

Verovatnoća da ćemo odbaciti tačnu nultu hipotezu (verovatnoća da ćemo napraviti grešku prve vrste) naziva se rizik prve vrste ili **prag značajnosti** i obeležava se sa α. Uobičajeno je da rizik greške prve vrste unapred biramo t.j sami biramo verovatnoću sa kojom se može dogoditi da odbacimo tačnu nultu hipotezu. Ako je, na primer, $\alpha = 0,05$, onda svesno prihvatamo da ćemo u 5% uzoraka odbaciti nultu hipotezu, iako je ona tačna.

Najčešće se uzima da je $\alpha = 0,05$ ili $\alpha = 0,01$.

Ako imamo nultu hipotezu $H_0(Q = Q_0)$ i odgovarajuću alternativnu hipotezu $H_1(Q \neq Q_0)$, tada hipotezu H_1 zovemo dvostranom hipotezom a odgovarajući test dvostranim testom. Alternativnu hipotezu $H_1(Q > Q_0)$ zovemo desnostranom hipotezom a odgovarajući test desnostranim testom. Analogno, alternativnu hipotezu $H_1(Q < Q_0)$ nazivamo levostranom hipotezom a odgovarajući test nazivamo levostranim testom. Oblast vrednosti statistike testa za koju odbacujemo nultu hipotezu nazivamo **kritična oblast** ili oblast odbacivanja hipoteze. Kritična oblast se određuje tako da verovatnoća da vrednost statistike testa pripadne toj oblasti bude mala kada je H_0 tačna. Preostali podskup mogućih vrednosti statistike testa čini oblast prihvatanja nulte hipoteze.

Vrednosti statistike testa koje razdvajaju oblast prihvatanja od oblasti odbacivanja hipoteze nazivaju se **kritičnim vrednostima**.

Ako imamo alternativnu hipotezu $H_1(Q < Q_0)$, oblast odbacivanja hipoteze $H_0(Q = Q_0)$ je oblika $(-\infty, c]$, gde je c kritična vrednost.

Kada je alternativna hipoteza $H_1(Q > Q_0)$, oblast odbacivanja hipoteze $H_0(Q = Q_0)$ je oblika $[c, +\infty)$, gde je c kritična vrednost.

U slučaju alternativne hipoteze $H_1(Q \neq Q_0)$, postoje dve kritične vrednosti c_1 i c_2 a oblast odbacivanja hipoteze $H_0(Q = Q_0)$ je $(-\infty, c_1] \cup [c_2, +\infty)$.

4

Parametarski testovi

4.1 Testiranje hipoteze o srednjoj vrednosti

Nekada je potrebno da se utvrdi da li se aritmetička sredina određenog obeležja neke populacije razlikuje od neke unapred date vrednosti. To je slučaj kada se neka vrednost iz iskustva ili iz literature smatra "normalnom", odnosno referentnom.

Studentov t test

Osnovna pretpostavka za primenu ovog testa je da dato obeležje ima normalnu raspodelu u populaciji. Testira se hipoteza $H_0(\mu = \mu_0)$, gde je μ aritmetička sredina obeležja u populaciji a μ_0 je data vrednost. Ako je hipoteza $H_0(\mu = \mu_0)$ tačna, onda statistika

$$t_{n-1} = \frac{\overline{X}_n - \mu_0}{\frac{S_n}{\sqrt{n}}},$$

ima Studentovu raspodelu sa $n-1$ stepenom slobode. Ako je $n-1$ veće od 30 koristi se normalana raspodela.

Neka je iz uzetog uzorka $(x_1, \ldots, x_n)$ izračunata vrednost

$$t^*_{n-1} = \frac{\overline{x}_n - \mu_0}{\frac{s_n}{\sqrt{n}}},$$

gde su $\overline{x}_n$ i s_n aritmetička sredina i standardna devijacija dobijene iz uzorka.

Ako je alternativna hipoteza oblika $H_1(\mu \neq \mu_0)$, onda, za dati prag značajnosti α, u tabeli Studentove raspodele nalazimo kritičnu vrednost $t_{n-1;\,\alpha}$. Primetimo da ova vrednost zadovoljava uslov $P_{H_0}(|t_{n-1}| \geq t_{n-1;\,\alpha}) = \alpha$. Ako u uzetom uzorku $(x_1, \ldots, x_n)$ konstatujemo da je $|t^*_{n-1}| \geq t_{n-1;\,\alpha}$ onda odbacujemo hipotezu $H_0(\mu = \mu_0)$. Ako je $|t^*_{n-1}| < t_{n-1;\,\alpha}$ onda prihvatamo hipotezu $H_0(\mu = \mu_0)$, odnosno kažemo da uzeti uzorak ne protivureči hipotezi $H_0(\mu = \mu_0)$.

Ako je alternativna hipoteza oblika $H_1(\mu < \mu_0)$, onda, za dati prag značajnosti α, u tabeli Studentove raspodele nalazimo kritičnu vrednost $t_{n-1;2\,\alpha}$. Ova vrednost zadovoljava uslov $P_{H_0}(t_{n-1} \leq -t_{n-1;2\,\alpha}) = \alpha$. Ako u uzetom uzorku $(x_1, \ldots, x_n)$ konstatujemo da je $t^*_{n-1} \leq -t_{n-1;2\,\alpha}$ onda odbacujemo hipotezu $H_0(\mu = \mu_0)$. Ako je $t^*_{n-1} > -t_{n-1;2\,\alpha}$ onda prihvatamo hipotezu $H_0(\mu = \mu_0)$.

Ako je alternativna hipoteza oblika $H_1(\mu > \mu_0)$, onda, za dati prag značajnosti α, u tabeli Studentove raspodele nalazimo kritičnu vrednost $t_{n-1;\,2\,\alpha}$. Ova vrednost zadovoljava uslov $P_{H_0}(t_{n-1} \geq t_{n-1;2\,\alpha}) = \alpha$. Ako u uzetom uzorku $(x_1, \ldots, x_n)$ konstatujemo da je $t^*_{n-1} \geq t_{n-1;2\,\alpha}$, onda odbacujemo hipotezu $H_0(\mu = \mu_0)$. Ako je $t^*_{n-1} < t_{n-1;2\,\alpha}$, onda prihvatamo hipotezu $H_0(\mu = \mu_0)$.

Ako je je $n - 1 > 30$ onda se Studentova raspodela aproksimira normalnom raspodelom $N(0, 1)$. Ako je alternativna hipoteza oblika $H_1(\mu \neq \mu_0)$, kritična vrednost c se dobija iz uslova $P(|t_{n-1}| \geq c) = \alpha$. Ako je alternativna hipoteza oblika $H_1(\mu < \mu_0)$, kritična vrednost c se dobija iz uslova $P(t_{n-1} \leq c) = \alpha$. Ako je alternativna hipoteza oblika $H_1(\mu > \mu_0)$, kritična vrednost c se dobija iz uslova $P(t_{n-1} \geq c) = \alpha$. Određivanje ovih kritičnih vrednosti pokazano je u Primeru 1.9 i Primeru 1.10.

Primer 4.1. Mašina je podešena da proizvodi tablete težine $0,50$ gr. Radi provere da li mašina proizvodi tablete propisane težine, uzet je uzorak od 11 tableta:

$0,57 \quad 0,49 \quad 0,51 \quad 0,55 \quad 0,56 \quad 0,51 \quad 0,57 \quad 0,49 \quad 0,55 \quad 0,52 \quad 0,51$ (gr).

Testirati hipotezu da mašina proizvodi tablete propisane težine s pragom značajnosti $\alpha = 0,05$.

Rešenje. Testira se nulta hipoteza $H_0(\mu = 0,50)$ protiv alternativne hipoteze $H_1(\mu \neq 0,50)$. Može se smatrati da težina tableta ima normalnu raspodelu. Najpre izračunavamo:

$$\begin{aligned}
\overline{x}_n &= \tfrac{1}{11}\,(0,57+0,49+0,51+0,55+0,56+0,51+0,57+0,49+0,55\\
&+ 0,52+0,51) = 0,53
\end{aligned}$$

$$\begin{aligned}
s_n^2 &= \tfrac{1}{11-1}\,((0,57-0,53)^2+(0,49-0,53)^2+(0,51-0,53)^2\\
&+ (0,55-0,53)^2+(0,56-0,53)^2+(0,51-0,53)^2+(0,57-0,53)^2\\
&+ (0,49-0,53)^2+(0,55-0,53)^2+(0,52-0,53)^2+(0,51-0,53)^2)\\
&= 0,00094
\end{aligned}$$

$$s_n = \sqrt{0,00094} = 0,03$$

odakle dobijamo

$$t^*_{n-1} = \frac{\overline{x}_n - \mu_0}{\frac{s_n}{\sqrt{n}}} = \frac{0,53-0,50}{\frac{0,03}{\sqrt{11}}} = 3,32.$$

Iz tabele Studentove raspodele čitamo $t_{n-1;\,\alpha} = t_{10;\,0,05} = 2,228$. Kako je

$$|t^*_{n-1}| = 3,32 > 2,228 = t_{10;\,0,05},$$

odbacujemo nultu hipotezu, odnosno odstupanje od propisane težine od 0,50 gr je statistički značajno.

Primer 4.2. Propisana težina boce fiziološkog rastvora (zajedno sa ambalažom) je 600 gr. Pošto se javila sumnja da su boce lakše od propisane težine, uzet je uzorak od 8 boca: 540, 580, 610, 530, 600, 520, 540, 530 (gr). Na osnovi uzorka treba utvrditi da li je sumnja opravdana.

Rešenje. Treba testirati hipotezu $H_0(\mu = 600)$. Zbog postojanja sumnje da su boce lakše od 600 grama, uzima se da je alternativna hipoteza $H_1(\mu < 600)$. Može se smatrati da težina boca ima normalnu raspodelu.

Izračunavanjem se dobija

$$\overline{x}_n = \tfrac{1}{8}(540+580+610+530+600+520+540+530) = 556,25$$

$$\begin{aligned}
s_n^2 &= \tfrac{1}{8-1}((540-556,25)^2+(580-556,25)^2+(610-556,25)^2\\
&+ (530-556,25)^2+(600-556,25)^2+(520-556,25)^2\\
&+ (540-556,25)^2+(530-556,25)^2)\\
&= 1226,79
\end{aligned}$$

$$s_n = \sqrt{1226,79} = 35,03$$

pa je

$$t^*_{n-1} = \frac{\overline{x}_n - \mu_0}{\frac{s_n}{\sqrt{n}}} = \frac{556,25 - 600}{\frac{35,79}{\sqrt{8}}} = -3,51.$$

Neka je $\alpha = 0,01$. Pošto imamo levostrani test, iz tabele Studentove raspodele čitamo $t_{n-1;\,2\,\alpha} = t_{7;\,0,02} = 2,295$. Kako je

$$t^*_{n-1} = -3,51 < -2,295 = -t_{7;\,0,02}$$

odbacujemo hipotezu $H_0(\mu = 600)$ i kažemo da je odstupanje visoko značajno, odnosno da su sumnje bile opravdane.

Zapažanje. Vratimo se Primeru 2.8. u kome je bilo $\overline{x}_n = 556,25$ i $s_n = 35,03$, odnosno aritmetička sredina i standardno odstupanje su uzeti upravo iz Primera 4.2. U pomenutom primeru 95% interval poverenja je bio $[526,98;\ 585,52]$. Ovo znači da se sa verovatnoćom od 0,95 može tvrditi da se aritmetička sredina težina boca cele populacije nalazi u intervalu $[526,98;\ 585,52]$, odnosno da je verovatnoća da se aritmetička sredina težina boca osnovnog skupa nađe van ovog intervala jednaka 0,05. Pošto propisana težina boce od 600 gr ne pripada intervalu povrenja, mala je verovatnoća da je to aritmetička sredina težina boca u populaciji, pa je logično da se hipoteza $H_0(\mu = 600)$ odbaci.

Pokazuje se da se testiranje hipoteze $H_0(\mu = \mu_0)$ protiv alternativne hipoteze $H_1(\mu \neq \mu_0)$ može sprovesti i pomoću intervala poverenja.

Primer 4.2, kao i slični primeri koji se odnose na mali uzorak, daje se zbog jednostavnosti računanja, ali je njegova verodostojnost zaključivanja dosta mala. Verodostojnost zaključivanja u prethodnom primeru bi bila svakako veća kada bi uzorak bio veliki, odnosno kada bi se uzelo više od 30 boca fiziološkog rastvora. U tom slučaju statistika t_{n-1} bi imala normalnu raspodelu.

Primer 4.3. Predviđena norma za proizvodnju jednog proizvoda je 55 sekundi. Radnici su se žalili da je norma nerealna. Da bi se utvrdilo da li je norma realna, mereno je vreme kod 60 radnika, pri čemu je dobijeno $\overline{x}_n = 72$ sekundi i $s_n = 20$ sekundi. Da li se, sa pragom značajnosti $\alpha = 0,01$, može prihvatiti hipoteza da je propisana norma saglasna sa realnim vremenom potrebnim za proizvodnju jednog proizvoda?

Rešenje. Ovde se testira nulta hipoteza $H_0(\mu = 55)$ protiv alternativne hipoteze $H_1(\mu > 55)$. Ovakva alternativna hipoteza se uzima zbog toga što

radnici smatraju da je za proizvodnju jednog proizvoda potrebno više od 55 sekundi . Za testiranje se koristi statistika

$$t_{n-1} = \frac{\overline{X}_n - \mu_0}{\frac{S_n}{\sqrt{n}}}$$

koja ima Studentovu raspodelu sa $k = n - 1 = 59$ stepena slobode i koja se, zbog $n-1 > 30$, dobro aproksimira normalnom raspodelom. Vrednost statistike t se računa na uobičajeni način:

$$t^* = \frac{\overline{x}_n - \mu_0}{\frac{s_n}{\sqrt{n}}} = \frac{72 - 55}{\frac{20}{\sqrt{60}}} = 6,58.$$

Pošto je prag značajnosti $\alpha = 0,01$, kritičnu vrednost c određujemo iz uslova $P(t_{n-1} \geq c) = 0,01$. Iz primera 1.10 se vidi da je $c = 2,32$. Kako je $t^* = 6,58 > 2,32$, s pragom značajnosti 0,01 odbacujemo nultu hipotezu H_0 da je realna norma 55 sekundi, odnosno zaključujemo da realna norma značajno odstupa od predviđenih 55 sekundi.

4.2 Testiranje hipoteze o jednakosti srednjih vrednosti

U praksi često treba uporediti aritmetičke sredine nekog obeležja dveju populacija. Potrebno je, naime, testirati hipotezu $H_0(\mu_1 = \mu_2)$, gde je μ_1 aritmetička sredina obeležja u prvoj popuaciji a μ_2 aritmetička sredina obeležja u drugoj populaciji.

t test

Osnovna pretpostavka za primenu ovog testa je da dato obeležje ima normalnu raspodelu u obe populacije. Takođe se pretpostavlja da su disperzije obeležja u populacijama σ_1^2 i σ_2^2 jednake.

Ako je hipoteza $H_0(\mu_1 = \mu_2)$ tačna onda statistika

$$t = \frac{\overline{X}_{n_1} - \overline{X}_{n_2}}{\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}(\frac{1}{n_1} + \frac{1}{n_2})}}$$

ima Studentovu raspodelu sa $k = n_1 + n_2 - 2$ stepena slobode. Iz populacija se uzimaju dva uzorka obima n_1 odnosno n_2 . Iz uzoraka se izračunava vrednost

$$t^* = \frac{\overline{x}_{n_1} - \overline{x}_{n_2}}{\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}(\frac{1}{n_1}+\frac{1}{n_2})}}.$$

gde su $\overline{x}_{n_1}$ i s_1 aritmetička sredina i standardna devijacija obeležja prvog uzorka a $\overline{x}_{n_2}$ i s_2 aritmetička sredina i standardna devijacija drugog uzorka.

Ako je alternativna hipoteza oblika $H_1(\mu_1 \neq \mu_2)$ onda, za dati prag značajnosti α, u tabeli Studentove raspodele čitamo kritičnu vrednost $t_{n_1+n_2-2;\,\alpha}$. Ako u uzetim uzorcima konstatujemo da je $|t^*| \geq t_{n_1+n_2-2;\,\alpha}$ onda odbacujemo hipotezu $H_0(\mu_1 = \mu_2)$. Ako je $|t^*| < t_{n_1+n_2-2;\,\alpha}$, onda prihvatamo hipotezu $H_0(\mu_1 = \mu_2)$.

Ako je alternativna hipoteza oblika $H_1(\mu_1 < \mu_2)$ onda, za dati prag značajnosti α, u tabeli Studentove raspodele čitamo kritičnu vrednost $t_{n_1+n_2-2;2\,\alpha}$. Ako u uzetim uzorcima konstatujemo da je $t^* < -t_{n_1+n_2-2;\,2\ \alpha}$ onda odbacujemo hipotezu $H_0(\mu_1 = \mu_2)$. Ako je $t^* > -t_{n_1+n_2-2;\,2\,\alpha}$, onda prihvatamo hipotezu $H_0(\mu_1 = \mu_2)$.

Ako je alternativna hipoteza oblika $H_1(\mu_1 > \mu_2)$ onda, za dati prag značajnosti α, u tabeli Studentove raspodele čitamo kritičnu vrednost $t_{n_1+n_2-2;2\,\alpha}$. Ako u uzetim uzorcima konstatujemo da je $t^* \geq t_{n_1+n_2-2;\,2\,\alpha}$ onda odbacujemo hipotezu $H_0(\mu_1 = \mu_2)$. Ako je $t^* < t_{n_1+n_2-2;\,2\,\alpha}$ onda prihvatamo hipotezu $H_0(\mu_1 = \mu_2)$.

Ako je $n_1+n_2-2 > 30$ onda se Studentova raspodela aproksimira normalnom raspodelom $N(0,1)$. Kritične vrednosti se određuju kao kod Studentovog t testa, odnosno kao u Primeru 1.9 i Primeru 1.10.

Primer 4.4. Pretpostavlja se da pacijenti koji boluju od bolesti A imaju veći broj leukocita nego pacijenti koji boluju od bolesti B. Zbog toga su na slučajan način iz ovih populacija uzeti ispitanici i meren im je broj leukocita:

bolest A	17	11	22	18	19	13	14	16
bolest B	15	12	10	18	14	15	13	

Tabela 4.1.

Testirati postavljenu hipotezu sa pragom značajnosti $\alpha = 0,05$.

Rešenje. U ovom primeru testiramo hipotezu $H_0(\mu_1 = \mu_2)$ protiv alternativne hipoteze $H_1(\mu_1 > \mu_2)$. Može se prihvatiti da broj leukocita ima normalnu raspodelu, pa možemo da koristimo statistiku

$$t = \frac{\overline{X}_{n_1} - \overline{X}_{n_2}}{\sqrt{\frac{(n_1-1)S_1^2+(n_2-2)S_2^2}{n_1+n_2-2}(\frac{1}{n_1}+\frac{1}{n_2})}},$$

koja ima Studentovu raspodelu sa $k = n_1 + n_2 - 2 = 8 + 7 - 2 = 13$ stepena slobode. Računanjem dobijamo

$$\begin{aligned}
\overline{x}_1 &= \tfrac{1}{8}(17+11+22+18+19+13+14+16) = \tfrac{130}{8} \approx 16,25 \\
\overline{x}_2 &= \tfrac{1}{7}(15+12+10+18+14+15+13) = \tfrac{97}{7} = 13,86, \\
(n_1-1)s_1^2 &= (17-16,25)^2+(11-16,25)^2+(22-16,25)^2+(18-16,25)^2 \\
&+ (19-16,25)^2+(13-16,25)^2+(14-16,25)^2+(16-16,25)^2 = 87,50 \\
(n_2-1)s_2^2 &= (15-13,86)^2+(12-13,86)^2+(10-13,86)^2+(18-13,86)^2 \\
&+ (14-13,86)^2+(15-13,86)^2+(13-13,86)^2 = 38,86
\end{aligned}$$

$$t^* = \frac{\overline{x}_{n_1} - \overline{x}_{n_2}}{\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}(\frac{1}{n_1}+\frac{1}{n_2})}} = \frac{16,25-13,86}{\sqrt{\frac{87,50+38,86}{8+7-2}(\frac{1}{8}+\frac{1}{7})}} = 1,48.$$

U tabeli Studentove raspodele čitamo broj $t_{n_1+n_2-2;\,2\,\alpha} = t_{13;\,0,10} = 1,771$. Kako je

$$t^* = 1,48 < 1,771 = t_{13;\,0,10},$$

nemamo razloga da odbacimo nultu hipotezu $H_0(\mu_1 = \mu_2)$, o jednakosti broja leukocita kod bolesti A i B, t.j. veći broj leukocita kod bolesti A nije statistički značajan i može se smatrati da je nastao slučajno.

Primer 4.5. Jedna grupa ispitanika sa povišenim krvnim pritiskom uzimala je lek A, dok je druga grupa uzimala lek B. Zabeleženi su sledeći rezultati:

Lek A

pre terapije	170	185	190	160	150	180	145	170	185	190
posle terapije	140	160	150	175	120	140	120	135	140	150

Tabela 4.2.

Lek B

pre terapije	180	160	150	175	190	170	155
posle terapije	170	160	155	155	180	140	130

Tabela 4.3.

Ispitati koji je lek efikasniji.

Rešenje. Pod efektom leka podrazumevama se razlika između krvnog pritiska pre uzimanja leka i krvnog pritiska posle uzimanja leka, odnosno za koliko je smanjen krvni pritisak posle uzimanja leka. Razliku ćemo izračunati za svakog ispitanika i uporediti srednje vrednosti razlika grupe koja uzima lek A i grupe koja uzima lek B, odnosno testiramo ćemo nultu hipotezu $H_0(\mu_1 = \mu_2)$ protiv alternativne hipoteze $H_1(\mu_1 \neq \mu_2)$, gde su μ_1 i μ_2 aritmetičke sredine razlika. Razlike su date u Tabeli 4.4. i Tabeli 4.5.

Lek A

pre terapije	170	185	190	160	150	180	145	170	185	190
posle terapije	140	160	150	175	120	140	120	135	140	150
razlike	30	25	40	-15	30	40	25	35	45	40

Tabela 4.4.

Lek B

pre terapije	180	160	150	175	190	170	155
posle terapije	170	160	155	155	180	140	130
razlike	10	0	-5	20	10	30	25

Tabela 4.5.

Računanjem dobijamo

$$\overline{x}_{n_1} = \tfrac{1}{10}(30+25+40-15+30+40+25+35+45+40) = 29,50$$
$$\overline{x}_{n_2} = \tfrac{1}{7}(10+0-5+20+10+30+25) = 12,86$$

$$\begin{aligned}(n_1-1)s_1^2 &= (30-29,50)^2+(25-29,50)^2+(40-29,50)^2\\ &+ (-15-29,50)^2+(30-29,50)^2+(40-29,50)^2\\ &+ (25-29,50)^2+(35-29,50)^2+(45-29,50)^2\\ &+ (40-29,50)^2 = 2622,50\\ (n_2-1)s_2^2 &= (10-12,86)^2+(0-12,86)^2+(-5-12,86)^2\\ &+ (20-12,86)^2+(10-12,86)^2+(30-12,86)^2\\ &+ (25-12,86)^2 = 992,28\end{aligned}$$

$$t^* = \frac{\overline{x}_{n_1}-\overline{x}_{n_2}}{\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}(\frac{1}{n_1}+\frac{1}{n_2})}} = \frac{29,50-12,86}{\sqrt{\frac{2622,50+992,28}{10+7-2}(\frac{1}{10}+\frac{1}{7})}} = 2,175.$$

Iz tabele Studentove raspodele nalazimo broj $t_{n_1+n_2-2;\,0,05} = t_{15;\,0,05} = 2,131$. Kako je $|t^*| = 2,175 > 2,131 = t_{15;\,0,05}$ odbacujemo nultu hipotezu i kažemo da je razlika između efekata leka A i leka B statistički značajna. Lek A, u proseku, više smanjuje krvni pritisak od leka B.

U prethodnom primeru, kao i sličnim primerima koji se odnose na mali uzorak, verodostojnost zaključivanja je dosta mala. Verodostojnost zaključivanja u prethodnom primeru bi bila, svakako, veća kada bi uzorak bio veliki, odnosno kada bi bilo $n_1+n_2-2 > 30$. U tom slučaju statistika t se aproksimira normalnom raspodelom $N(0,1)$.

Primer 4.6. Iz jedne populacije uzet je uzorak od 20 ispitanika i izmerena im je vrednost šećera u krvi, pri čemu je dobijeno $\overline{x}_1 = 9,40$ i $s_1 = 1,85$, dok je iz druge populacije uzet uzorak od 18 ispitanika , pri čemu je dobijeno $\overline{x}_2 = 11,00$ i $s_2 = 1,49$. Ispitati, s pragom značajnosti $\alpha = 0,05$, da li je razlika između srednjih vrednosti šećera u krvi ovih dveju populacija značajna.

Rešenje. Iz dobijenih vrednosti izračunava se vrednost statistike

$$t^* = \frac{\overline{x}_{n_1}-\overline{x}_{n_2}}{\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}(\frac{1}{n_1}+\frac{1}{n_2})}} = \frac{9,40-11,00}{\sqrt{\frac{(20-1)\cdot 1,85^2+(18-1)\cdot 1,49^2}{20+18-2}(\frac{1}{20}+\frac{1}{18})}} = -2,91.$$

Pošto je stepen slobode $k = 20+18-2 = 36$ veći od 30, Studentova raspodela se dobro aproksimira normalnom raspodelom $N(0,1)$. Kritična vrednost se dobija

iz uslova $P(|t| \geq c) = 0,05$. Iz primera 1.9 se vidi da je $c = 1,96$. Kako je $|t^*| = 2,91 > 1,96$, odbacujemo hipotezu H_0 i kažemo da je razlika između srednjih vrednosti šećera u krvi ovih dveju populacija značajna.

Ako bi se uzelo da je $\alpha = 0,01$, saglasno Primeru 1.10, dobilo bi se $c = 2,58$. Kako je $|t^*| = 2,91 > 2,58$, nulta hipoteza bi bila odbačena i sa ovim pragom značajnosti, što znači da je razlika između srednjih vrednosti šećera u krvi ovih dveju populacija visoko značajna.

4.3 Upareni t-test

Često se dešava da se dva puta (različitim metodama ili u različitim vremenskim intervalima) vrše merenja nekog obeležja nad istim skupom elemenata. Na primer, nekoj grupi ispitanika se izmeri krvni pritisak i propiše određena terapija. Posle izvesnog vremena istoj grupi ispitanika se meri krvni pritisak kako bi se utvrdilo da li su razlike između prvog i drugog merenja statistički značajne, odnosno da li propisana terapija ima efekta. Pretpostavimo da su prilikom prvog merenja dobijene vrednosti $x_1, x_2, \ldots, x_n$ a prilikom drugog merenja vrednosti $y_1, y_2, \ldots, y_n$, odnosno imamo parove $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Kada se izračunaju razlike $d_1 = x_1 - y_1, d_2 = x_2 - y_2, \ldots, d_n = x_n - y_n$, testira se hipoteza da je aritmetička sredina ovih razlika u populaciji jednaka nuli, odnosno hipoteza $H_0(\mu = 0)$. Uslov za primenu ovog testa je da razlike parova imaju normalnu raspodelu.

Primer 4.7. Grupi od 7 pacijenata meren je krvni pritisak. Posle uzimanja određenog leka istim pacijentima je ponovo meren krvni pritisak. Dobijeni rezultati merenja prikazani su u Tabeli 4.6.

pre terapije	180	160	150	175	190	170	155
posle terapije	170	160	155	155	180	140	130
razlike	10	0	-5	20	10	30	25

Tabela 4.6.

Ispitati efikasnost ovog leka.

Rešenje. Ovde se testira nulta hipoteza $H_0(\mu = 0)$ protiv alternativne hipoteze $H_1(\mu \neq 0)$. Računanjem dobijamo

$$\begin{aligned}
\overline{d}_n &= \tfrac{1}{7}(10+0-5+20+10+30+25) = 12,86 \\
s_n^2 &= \tfrac{1}{7-1}((10-12,86)^2+(0-12,86)^2+(-5-12,86)^2+(20-12,86)^2 \\
&+ (10-12,86)^2+(30-12,86)^2+(25-12,86)^2) = 165,48 \\
s_n &= \sqrt{165,48} \approx 12,86
\end{aligned}$$

$$t^* = \frac{\overline{d_n}-0}{\frac{s_n}{\sqrt{n}}} = \frac{12,86}{\frac{12,86}{\sqrt{7}}} \approx 2,645.$$

U tabeli Studentove raspodele nalazimo broj $t_{n-1;\,\alpha} = t_{6;\,0,05} = 2,447$. Kako je $|t^*| = 2,645 > 2,447 = t_{6;\,0,05}$ odbacujemo nultu hipotezu i kažemo da je rzlika između krvnog pritiska pre i posle uzimanja leka statistički značajna, odnosno da je lek efikasan.

Napominjemo da se ovde radi o leku B iz Primera 4.5. Kao što je pokazano u Primeru 4.5., lek B je manje efikasan od leka A, ali ovaj primer pokazuje da je i lek B efikasan.

4.4 Testovi o procentualnoj zastupljenosti

Testiranje hipoteze $H_0(\pi = \pi_0)$

Neka je π proporcija, odnosno verovatnoća da neki element date populacije ima određeno svojstvo $\mathcal{S}$. Na primer, verovatnoća da je stanovnik određenog grada levoruk je 0,12. Primetimo da se množenjem verovatnoće π brojem 100 dobija procenat zastupljenosti elemenata sa svojstvom $\mathcal{S}$ u populaciji. Testiraćemo hipotezu $H_0(\pi = \pi_0)$, gde je π proporcija, odnosno verovatnoća da neki element date populacije ima svojstvo $\mathcal{S}$. Neka je n broj elemenata uzorka uzetog iz populacije a m broj elemenata iz uzorka koji imaju određeno svojstvo $\mathcal{S}$. Pomoću statistike $\overline{p} = \frac{m}{n}$, koja predstavlja verovatnoću da neki element uzorka ima svojstvo $\mathcal{S}$, formiramo statistiku

$$z = \frac{\overline{p}-\pi_0}{\sqrt{\frac{\pi_0\cdot(1-\pi_0)}{n}}}.$$

Pod pretpostavkom da je hipoteza $H_0(\pi = \pi_0)$ tačna, ova statistika ima približno normalnu raspodelu $N(0,1)$. Uzimamo veliki uzorak ($n > 100$) i iz

uzorka izračunavamo

$$z^* = \frac{\frac{m}{n} - \pi_0}{\sqrt{\frac{\pi_0 \cdot (1-\pi_0)}{n}}}.$$

Ako je alternativna hipoteza oblika $H_1(\pi \neq \pi_0)$ onda, za dati prag značajnosti α, u tabeli normalne raspodele nalazimo kritičnu vrednost c takvu da je

$$P(|z| \geq c) = \alpha.$$

Ako je u uzetom uzorku $|z^*| \geq c$, onda odbacujemo hipotezu $H_0(\pi = \pi_0)$. U suprotnom, prihvatamo hipotezu $H_0(\pi = \pi_0)$.

Ako je alternativna hipoteza oblika $H_1(\pi < \pi_0)$ onda, za dati prag značajnosti α, u tabeli normalne raspodele nalazimo kritičnu vrednost c takvu da je

$$P(z \leq c) = \alpha.$$

Ako je u uzetom uzorku $z^* \leq c$, onda odbacujemo hipotezu $H_0(\pi = \pi_0)$. U suprotnom, prihvatamo hipotezu $H_0(\pi = \pi_0)$.

Ako je alternativna hipoteza oblika $H_1(\pi > \pi_0)$ onda, za dati prag značajnosti α, u tabeli normalne raspodele nalazimo kritičnu vrednost c takvu da je

$$P_{H_0}(z \geq c) = \alpha.$$

Ako je u uzetom uzorku $z^* > c$, onda odbacujemo hipotezu $H_0(\pi = \pi_0\ )$. U suprotnom, prihvatamo hipotezu $H_0(\pi = \pi_0)$.

Kritična vrednost c se određuje kao u Primeru 1.9 i Primeru 1.10.

Primer 4.8. Jedan institut tvdi da je najviše 1% stanovnika jednog grada obolelo od gripa. Na slučajan način je uzeto 1600 stanovnika tog grada i utvrđeno je da je među njima ima 18 obolelih od gripa.

(a) Da li se na osnovi ovog uzorka može zaključiti, sa pragom značajnosti $\alpha = 0,05$, da je tvrđenje instituta tačno?

(b) Koliko bi obolelih građana trebalo da se nađe u uzorku obima 1600 da bi se odbacilo tvrđenje instituta?

Rešenje.

(a) Označimo sa π verovatnoću da je neki građanin pomenutog grada oboleo od gripa. Pošto institut tvrdi da je $\pi \leq 0,01$, za alternativnu hipotezu se uzima suprotno tvrđenje, odnosno hipoteza $H_1(\pi > 0,01)$. Nulta hipoteza je $H_0(\pi = 0,01)$. Kako je $\pi = 0,01$, $m = 18$ and $n = 1600$ imamo

$$z^* = \frac{\frac{m}{n} - \pi_0}{\sqrt{\frac{\pi_0 \cdot (1-\pi_0)}{n}}} = \frac{\frac{18}{1600} - 0,01}{\sqrt{\frac{0,01 \cdot 0,99}{1600}}} = 0,5025.$$

Kritičnu vrednost c određujemo iz uslova $P(z \geq c) = 0,05$. Iz Primera 1.9. imamo da je $c = 1,645$. Kako je

$$z^* = 0,5025 < 1,645 = c$$

nemamo razloga da odbacimo hipotezu H_0, odnosno smatramo da je tvrđenje instituta tačno.

(b) Pošto je kritična vrednost $c = 1,645$, vrednost statistike z bi trebalo da bude veća od 1,645 da bi se nulta hipoteza odbacila t.j.

$$\frac{\frac{m}{n} - \pi_0}{\sqrt{\frac{\pi_0 \cdot (1-\pi_0)}{n}}} = \frac{\frac{m}{1600} - 0,01}{\sqrt{\frac{0,01 \cdot 0,99}{1600}}} > 1,645$$

$$\Rightarrow \frac{m}{1600} - 0,01 > 1,645\sqrt{\frac{0,01 \cdot 0,99}{1600}}$$

$$\Rightarrow m > 1600(0,01 + 1,645\sqrt{\frac{0,01 \cdot 0,99}{1600}}) \approx 23.$$

Znači, tvrđenje instituta bi se odbacilo, ako bi u uzorku broj obolelih od gripa bio veći od 23.

Testiranje hipoteze $H_0(\pi_1 = \pi_2)$

Date su dve populacije. Neka je π_1 verovatnoća da neki element prve populacije ima određeno svojstvo $\mathcal{S}$, a π_2 verovatnoća da neki element druge populacije ima svojstvo $\mathcal{S}$. Testiramo hipotezu $H_0(\pi_1 = \pi_2)$. Neka je n_1 broj elemenata uzorka uzetog iz prve populacije a m_1 broj elemenata iz uzorka koji imaju svojstvo $\mathcal{S}$. Neka je n_2 broj elemenata uzorka uzetog iz druge populacije a m_2 broj elemenata iz uzorka koji imaju svojstvo $\mathcal{S}$. Brojevi n_1 i n_2 su veći od 100. Pod pretpostavkom da je hipoteza $H_0(\pi_1 = \pi_2)$ tačna, statistika

$$z = \frac{\overline{p}_1 - \overline{p}_2}{\sqrt{\frac{\overline{p} \cdot \overline{q}}{n}}},$$

gde je

$$\overline{p} = \frac{m_1 + m_2}{n_1 + n_2}, \quad n = \frac{n_1 \cdot n_2}{n_1 + n_2}, \quad \overline{p}_1 = \frac{m_1}{n_1}, \quad \overline{p}_2 = \frac{m_2}{n_2}, \quad \overline{q} = 1 - \overline{p},$$

ima približno normalnu raspodelu $N(0,1)$. Iz uzoraka izračunavamo

$$z^* = \frac{\overline{p}_1 - \overline{p}_2}{\sqrt{\frac{\overline{p} \cdot \overline{q}}{n}}}.$$

Kritična vrednost c, pomoću koje prihvatamo ili odbacujemo nultu hipotezu $H_0(\pi_1 = \pi_2)$, dobija se kao i u slučaju testiranja hipoteze $H_0(\pi = \pi_0)$.

Primer 4.9. Na slučajan način je odabrano $n_1 = 1200$ stanovnika sa sela i među njima je nađeno $m_1 = 40$ obolelih od povišenog krvnog pritiska a među $n_2 = 1500$ stanovnika iz grada nađeno je $m_2 = 100$ stanovnika sa povišenim krvnim pritiskom. S pragom značajnosti $\alpha = 0,05$ testirati hipotezu o jednakosti procenata obolelih od povišenog krvnog pritiska u populaciji ljudi sa sela i populaciji ljudi iz grada.

Rešenje. Koristićemo dvostrani test, t.j. nulta hipoteza je $H_0(\pi_1 = \pi_2)$, a alternativna je $H_1(\pi_1 \neq \pi_2)$, gde π_1 i π_2 označavaju verovatnoće oboljevanja od povišenog krvnog pritiska u populacijama ljudi iz sela i iz grada. Na osnovi datih podataka izračunavamo

$\overline{p_1} = \frac{m_1}{n_1} = \frac{40}{1200} = 0,033 = 3,3\%,$

$\overline{p_2} = \frac{m_2}{n_2} = \frac{100}{1500} = 0,067 = 6,7\%,$

$\overline{p} = \frac{m_1+m_2}{n_1+n_2} = \frac{140}{2700} = 0,052, \qquad \overline{q} = 1 - \overline{p} = 0,948,$

$n = \frac{n_1 \cdot n_2}{n_1+n_2} = \frac{1200 \cdot 1500}{1200+1500} = \frac{1800000}{2700} = \frac{18000}{27} = 667,$

odakle se dobija

$$z^* = \frac{\overline{p}_1 - \overline{p}_2}{\sqrt{\frac{\overline{p} \cdot \overline{q}}{n}}} = \frac{0,033 - 0,067}{\sqrt{\frac{0,052 \cdot 0,948}{667}}} = \frac{-0,033}{\sqrt{0,0000739}} = -3,9.$$

Kritična vrednost c se određuje iz uslova $P(|z| \geq c) = 0,05$. Iz Primera 1.9. imamo da je $c = 1,96$. Pošto je $|z^*| = |-3,91| = 3,9 > 1,96 = c$ odbacujemo nultu hipotezu H_0, odnosno procenti obolelih od povišenog krvnog pritiska kod seoskog i gradskog stanovništva se statistički značajno razlikuju.

Primer 4.10. Uzmimo prethodni primer, samo broj osoba umanjimo 10 puta t.j. $n_1 = 120$, $m_1 = 4$, $n_2 = 150$, $m_2 = 10$. Računanjem se, pomoću iste formule, dobija $z^* = -1,23$. Na osnovi ovog rezultata ne možemo odbaciti nultu hipotezu, jer je $|z^*| = 1,23 < 1,96$, već konstatujemo da je razlika procenata slučajnog karaktera.

Objašnjenje za to što ova dva primera daju različite rezultate leži u činjenici da su u drugom primeru uzorci malog obima i da odstupanja procentualnih zastupljenosti nisu mogla da se ispolje.

4.5 Test o jednakosti disperzija

Jedna od pretpostavki za primenu t testa je da su disperzije obeležja u populacijama jednake. Zbog toga je potrebno da se testira hipoteza $H_0(\sigma_1^2 = \sigma_2^2)$, gde je σ_1^2 disperzija obeležja prve populacije a σ_2^2 disperzija obeležja druge populacije.

Pretpostavimo da dato obeležje ima normalnu raspodelu u obe populacije. Iz prve populacije se uzima uzorak od n_1 elemenata a iz druge populacije uzorak od n_2 elemenata. Neka su iz uzoraka izračunate disperzije $s_{n_1}^2$ i $s_{n_2}^2$. Hipoteza $H_0(\sigma_1^2 = \sigma_2^2)$ se, obično, testira sa pragom značajnosti $\alpha = 0,05$. Može se dokazati da statistika

$$F = \frac{n_1(n_2-1)S_{n_1}^2}{n_2(n_1-1)S_{n_2}^2} \approx \frac{S_{n_1}^2}{S_{n_2}^2}$$

ima Fišerovu raspodelu čiji su stepeni slobode $k_1 = n_1 - 1$ i $k_2 = n_2 - 1$. Iz uzoraka izračunavamo

$$F^* = \frac{s_{n_1}^2}{s_{n_2}^2}$$

a iz tabele Fišerove raspodele (Tabela V) čitamo $F_{k_1;\,k_2;\,0,05}$. Ako je

$$F^* \geq F_{k_1;\,k_2;\,0,05},$$

onda odbacujemo hipotezu o jednakosti disperzija i kažemo da je razlika između disperzija statistčki značajna. U suprotnom, hipotezu prihvatamo i kažemo da razlika između disperzija nije statistčki značajna. U slučaju da je razlika između disperzija značajna, treba odbaciti pretpostavku da uzeti uzorci pripadaju istoj osnovnoj populaciji. Ovo je naročito važno kod malih uzoraka.

Valja napomenuti da, kada računamo veličinu F^* , uvek treba staviti u brojilac veću od vrednosti $s^2_{n_1}$ i $s^2_{n_2}$.

Primer 4.11. Iz populacije ispitanika koji boluju od bolesti A uzet je uzorak od 8 ispitanika, iz populacije koji boluju od bolesti B uzorak od 7 ispitanika i meren im je broj leukocita:

bolest A	17	11	22	18	19	13	14	16
bolest B	15	12	10	18	14	15	13	

Tabela 4.1.

Testirati hipotezu o jednakosti disperzija broja leukocita u populacijama A i B.

Rešenje. U ovom primeru se testira nulta hipoteza o jednakosti disperzija $H_0(\sigma_1^2 = \sigma_2^2)$ protiv alternativne hipoteze $H_1(\sigma_1^2 \neq \sigma_2^2)$. Može se prihvatiti da broj leukocita ima normalnu raspodelu. Računanjem dobijamo

$$\begin{aligned}
\overline{x}_1 &= \tfrac{1}{8}(17+11+22+18+19+13+14+16) = \tfrac{100}{8} \approx 16,25 \\
\overline{x}_2 &= \tfrac{1}{7}(15+12+10+18+14+15+13) = \tfrac{97}{7} = 13,86, \\
s^2_{n_1} &= \tfrac{1}{8-1}((17-16,25)^2 + (11-16,25)^2 + (22-16,25)^2 + (18-16,25)^2) \\
&+ (19-16,25)^2 + (13-16,25)^2 + (14-16,25)^2 + (16-16,25)^2 \\
&= \tfrac{1}{7} \cdot 87,50 = 12,50 \\
s^2_{n_2} &= \tfrac{1}{7-1}(15-13,86)^2 + (12-13,86)^2 + (10-13,86)^2 + (18-13,86)^2) \\
&+ (14-13,86)^2 + (15-13,86)^2 + (13-13,86)^2 = \tfrac{1}{6} \cdot 38,86 = 6,48
\end{aligned}$$

$$F^* = \frac{s_8^2}{s_7^2} = \frac{12,50}{6,48} = 1,93.$$

Iz tabele Fišerove raspodele čitamo $F_{7;\,6;\,0,05} \approx F_{8;\,6;\,0,05} = 4,15$. Kako je $F^* < F_{8;\,6;\,0,05}$, nemamo razloga da odbacimo hipotezu o jednakosti disperzija broja leukocita u populacijama A i B. Primetimo da su podaci u ovom primeru jednaki podacima iz Primera 4.4., u kome je korišćen test o jednakosti aritmetičkih sredina a koji pretpostavlja jednakost disperzija. Kao što se vidi, ova pretpostavka je zadovoljena.

Primer 4.12. Jedna mašina proizvodi tablete određene težine. Uzet je uzorak od $n_1 = 10$ tableta i konstatovano je da je disperzija uzorka $s^2_{10} = 5,7\,mg^2$.

Posle određenog vremena uzet je drugi uzorak obima $n_2 = 15$ proizvoda iste mašine i konstatovano da je disperzija $s_{15}^2 = 9,6\,mg^2$. Da li se sa pragom značajnosti $\alpha = 0,05$ može smatrati da je došlo do značajnog porasta disperzije težine tableta kod posmatrane mašine, odnosno da se mašina "raštelovala"?

Rešenje. Iz dobijenih podataka izračunavamo

$$F^* = \frac{s_{15}^2}{s_{10}^2} = \frac{9,6}{5,7} = 1,68$$

Pošto se broj s_{15}^2 nalazi u brojiocu (jer je veći od s_{10}^2), to je $k_1 = 15-1 = 14$ dok je $k_2 = 10 - 1 = 9$. Iz tabele Fišerove raspodele čitamo $F_{14;\,9;\,0,05} \approx F_{12;\,9;\,0,05} = 3,07$. Kako je

$$F^* = 1,68 < 3,07 = F_{14;\,9;\,0,05}$$

prihvatamo nultu hipotezu $H_0(\sigma_1^2 = \sigma_2^2)$, odnosno ne možemo smatrati porast disperzije značajnim.

4.6 Analiza varijansi

Nekada je potrebno uporediti aritmetičke sredine nekog obeležja više od dve populacije. Na primer, ako imamo četiri populacije ljudi koji uzimaju četiri vrste leka protiv povišenog krvnog pritiska i želimo da uporedimo efekte tih lekova (smanjenje krvnog pritiska). To je moguće uraditi i na taj način što bi se pomoću t-testa poredile populacije svaka sa svakom. U slučaju četiri populacije trebalo bi uraditi 6 t-testova, u slučaju 5 populacija trebalo bi uraditi 10 t-testova itd. Ovde nije problem samo u tome što treba uraditi veliki broj testova (što podrazumeva dosta vremena), već što se verovatnoća da se napravi greška prve vrste višestruko uvećava. Naime, ako se opredelimo za prag značajnosti $\alpha = 0,05$, poređenjem aritmetičkih sredina pet populacija stvarni rizik da bar u jednom testu pogrešimo (odbacimo tačnu hipotezu), prema nekim autorima, iznosi oko 0,29. Da bi seo istovremeno, jednim postupkom, ispitala jednakost aritmetičkih sredina nekog obeležja više populacija koristi se statistički metod koji se zove analiza varijansi.

Jednofaktorska analiza varijansi

Posmatrajmo uticaj jednog promenljivog faktora A na neko obeležje X osnovnog skupa. Na primer, neka faktor A bude vrsta leka a obeležje X krvni pritisak. Neka su u pitanju lekovi A_1, A_2 i A_3. "Vrednosti" nekog faktora (u našem slučaju A_1, A_2 i A_3) nazivaju se nivoi ili tretmani. Postavlja se pitanje da li različiti nivoi faktora A dovode do bitnih ili slučajnih razlika vrednosti obeležja X.

Neka na obeležje X deluje jedan promenljivi faktor A, sa svojim nivoima $A_1, A_2 \ldots, A_r$. Iz populacije na koju je delovao nivo A_1 uzimamo uzorak obima n_1, iz populacije na koju je delovao nivo A_2 uzimamo uzorak obima n_2 itd. Na taj način dobijamo r uzoraka:

$$\begin{array}{cccc} X_{1,1}, & X_{1,2}, & \ldots, & X_{1,n_1} \\ X_{2,1}, & X_{2,2}, & \ldots, & X_{2,n_2} \\ . & . & \ldots & . \\ X_{r,1}, & X_{r,2}, & \ldots, & X_{r,n_r} \end{array}$$

gde je $n_1 + n_2 + \cdots + n_r = n$.

Pretpostavimo da obeležje X **unutar svake populacije ima normalnu raspodelu** $N(\mu_i, \sigma_i^2)$ $(i = 1, 2, \ldots, r)$, pri čemu se pretpostavlja da su varijanse jednake t.j. $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_r^2$, ali ne mora da budu poznate. Treba testirati hipotezu:

$$H_0(\mu_1 = \mu_2 = \cdots = \mu_r) \qquad (r > 2),$$

gde je μ_1 aritmetička sredina obeležja prve populacije, μ_2 aritmetička sredina obeležja druge populacije itd. Uvedimo oznake:

$\overline{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{i,j}$ $(i = 1, 2, \ldots, r)$ (aritmetička sredina obeležja i-tog uzorka)

$\overline{X} = \frac{1}{n} \sum_{i=1}^{r} \sum_{j=1}^{n_i} X_{i,j}$ $(j = 1, 2, \ldots, r)$ (aritmetička sredina celog uzorka)

$Q_1 = \sum_{i=1}^{r} n_i (\overline{X}_i - \overline{X})^2$ (zbir kvadrata među uzorcima)

$Q_2 = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (X_{i,j} - \overline{X}_i)^2$ (zbir kvadrata unutar uzoraka).

Ako je hipoteza $H_0(\mu_1 = \mu_2 = \cdots = \mu_r)$ tačna, može da se dokaže da statistika

$$F = \frac{(n-r)\,Q_1}{(r-1)\,Q_2}$$

ima Fišerovu raspodelu F_{k_1,k_2}, gde je $k_1 = r-1$ i $k_2 = n-r$. Za uzeti uzorak

$$\begin{array}{cccc} x_{1,1}, & x_{1,2}, & \ldots, & x_{1,n_1} \\ x_{2,1}, & x_{2,2}, & \ldots, & x_{2,n_2} \\ \cdot & \cdot & \cdots & \cdot \\ x_{r,1}, & x_{r,2}, & \ldots, & x_{r,n_r} \end{array}$$

svaka vrednost $x_{i,j}$ može se smatrati realizacijom slučajne promenljive $X_{i,j}$. Neka su q_1 i q_2 vrednosti statistika Q_1 i Q_2 izračunatih iz uzorka. Tada je

$$F^* = \frac{(n-r)\,q_1}{(r-1)\,q_2}$$

realizovana vrednost statistike F. Iz tabele Fišerove raspodele (Tabela V) čitamo vrednost $F_{r-1;\,n-r;\,0,05}$. Ako je

$$F^* \geq F_{r-1;\,n-r;\,0,05},$$

onda odbacujemo hipotezu $H_0(\mu_1 = \mu_2 = \cdots = \mu_r)$ i kažemo da je razlika između aritmetičkih sredina ovih r populacija statistički značajna. U suprotnom, hipotezu prihvatamo.

U slučaju odbacivanja hipoteze H_0, ovaj test nam ukazuje na to da se aritmetičke sredine bar dve populacije statistički značajno razlikuju.

Primer 4.13. U proizvodnji jedne vrste proizvoda primenjene su tri metode. Mereno je vreme (u minutima) potrebno da se proizvede jedan proizvod. Dobijeni rezultati prikazani su u Tabeli 4.7.

prva metoda	25	15	20	30	20		
druga metoda	40	20	25	50	10	35	
treća metoda	5	15	20	20	40	10	30

Tabela 4.7.

Testirati hipotezu, s pragom značajnosti $\alpha = 0,05$, da su vremena potrebna za proizvodnju jednog proizvoda kod ove tri metode jednaka.

Rešenje. Potrebno je testirati hipotezu $H_0(\mu_1 = \mu_2 = \mu_3)$. Da bismo primenili analizu varijansi, najpre izračunavamo:

$$\overline{x}_1 = \tfrac{25+15+20+30+20}{5} = 22$$

$$\overline{x}_2 = \tfrac{40+20+25+50+10+35}{6} = 30$$

$$\overline{x}_3 = \tfrac{5+15+20+20+40+10+30}{7} = 20$$

$$\overline{x} = \tfrac{25+15+20+30+20+40+20+25+50+10+35+5+15+20+20+40+10+30}{5+6+7} = 23,89.$$

$$\begin{aligned} q_1 &= \textstyle\sum_{i=1}^{r} n_i(\overline{x}_i - \overline{x})^2 = 5\,(22 - 23,89)^2 + 6\,(30 - 23,89)^2 \\ &+ 7\,(20 - 23,89)^2 = 347,78 \end{aligned}$$

$$\begin{aligned} q_2 &= \textstyle\sum_{i=1}^{r} \sum_{j=1}^{n_i}(x_{i,j} - \overline{x}_i)^2 \\ &= (25-22)^2 + (15-22)^2 + (20-22)^2 + (30-22)^2 + (20-22)^2 \\ &+ (40-30)^2 + (20-30)^2 + (25-30)^2 + (50-30)^2 + (10-30)^2 \\ &+ (35-30)^2 + (5-20)^2 + (15-20)^2 + (20-20)^2 + (20-20)^2 \\ &+ (40-20)^2 + (10-20)^2 + (30-20)^2 = 2030 \end{aligned}$$

Iz prethodnih podataka izračunavamo

$$F^* = \frac{(n-r)\,q_1}{(r-1)\,q_2} = \frac{(18-3)\cdot 347,78}{(3-1)\cdot 2030} = 1,285.$$

Iz tabele Fišerove raspodele čitamo $F_{2;\,15;\,0,05} = 3,68$. Kako je

$$F^* = 1,285 < 3,68 = F_{2;\,15;\,0,05}$$

nemamo razloga da odbacimo hipotezu $H_0(\mu_1 = \mu_2 = \mu_3)$ t.j. između primenjenih metoda ne postoje bitne razlike u vremenu potrebnom za proizvodnju jednog proizvoda.

Primer 4.14. Sa četiri fakulteta su, na slučajan način, izabrani studenti koji su radili test iz hemije pri čemu su dobijeni sledeći pojedinačni rezultati

1. fakultet	72	64	85	87	
2. fakultet	65	54	34	67	55
3. fakultet	94	89	85	95	
4. fakultet	84	87	89		

Tabela 4.8.

Da li se sa pragom značajnosti $\alpha = 0,05$ može tvrditi da su rezultati testa na ovim fakultetima jednaki?

Rešenje. Ovde testiramo hipotezu $H_0(\mu_1 = \mu_2 = \mu_3 = \mu_4)$. Iz datih podataka izračunavamo

$$\overline{x}_1 = \tfrac{1}{4}(72 + 64 + 85 + 87) = 77,$$
$$\overline{x}_2 = \tfrac{1}{5}(65 + 54 + 34 + 67 + 55) = 55,$$
$$\overline{x}_3 = \tfrac{1}{4}(94 + 89 + 85 + 95) = 90,75,$$
$$\overline{x}_4 = \tfrac{1}{3}(84 + 87 + 89) = 86,67$$
$$\overline{x} = \tfrac{1}{16}(72+63+85+87+65+54+34+67+55+94+89+85+95+84+87+89) = 75,37,$$

$$\begin{aligned} q_1 &= \textstyle\sum_{i=1}^{4} n_i(\overline{x}_i - \overline{x})^2 \\ &= 4(77 - 75,37)^2 + 5(55 - 75,37)^2 + 4(90,75 - 75,37)^2 \\ &+ 3(86,67 - 75,37)^2 = 3414,33 \end{aligned}$$

$$\begin{aligned} q_2 &= \textstyle\sum_{i=1}^{4} \sum_{j=1}^{n_i}(x_{ij} - \overline{x}_i)^2 \\ &= (72 - 77)^2 + (64 - 77)^2 + (85 - 77)^2 + (87 - 77)^2 \\ &+ (65 - 55)^2 + (54 - 55)^2 + (34 - 55)^2 + (67 - 55)^2 \\ &+ (55 - 55)^2 + (94 - 90,75)^2 + (89 - 90,75)^2 + (85 - 90,75)^2 \\ &+ (95 - 90,75)^2 + (84 - 86,67)^2 + (87 - 86,67)^2 + (89 - 86,67)^2 \\ &= 1121,42 \end{aligned}$$

$$F^* = \tfrac{(n-r)q_1}{(r-1)q_2} = \tfrac{(16-4)3414,33}{(4-1)1121,42} = 12,18.$$

U Tabeli V nalazimo $F_{r-1;\,n-r;\,0,05} = F_{3;\,12;\,0,05} = 3,49$. Pošto je $F^* = 12,18 > 3,49 = F_{3;\,12;\,0,05}$ odbacujemo hipotezu $H_0(\mu_1 = \mu_2 = \mu_3 = \mu_4)$, odnosno rezultati testa na ovim fakultetima se statistički značajno razlikuju.

Višestruka komparacija

Ako je primenom analize varijansi odbačena nulta hipoteza, odnosno ako je pokazano da je razlika između nekih populacija statistički značajna, onda je potrebno ispitati koje su to populacije. U tom smislu se može viputa primeniti

t-test, ali to često dovodi, kao što je rečeno, do velike verovatnoće da se bar u nekom od tih t-testova napravi greška prve vrste. Zbog toga su napravljeni testovi koji vrše testiranja između svake dve populacije i istovremeno "kontrolišu" grešku prve vrste. Naime, ovi testovi pokazuju između kojih populacija su razlike statistički značajne, pri čemu ukupna verovatnoća da se napravi greška prve vrste ne prelazi 0,05. Poznati su Bonferroni-jev, Tuckey-ev i Dunnet-ov T3 test. Prva dva se koriste kada važi hipoteza $H_0(\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_r^2)$, odnosno kada je zadovoljen takozvani uslov homogenosti. Ova hipoteza se testira na poseban način, o čemu ovde neće biti reči, već se čitalac upućuje na programske pakete za statistiku. Dunnet-ov T3 test se koristi kada ne važi hipoteza o jednakosti varijansi.

4.7 Realizovani nivo značajnosti testa

Do sada smo prilikom testiranja neke hipoteze unapred određivali nivo značajnosti testa. To je obično bilo 0,05 ili 0,01. Ako, na primer, testiramo hipotezu $H_0(\mu = \mu_0)$ protiv alternativne $H_1(\mu \neq \mu_0)$ pri čemu je uzorak veliki ($n > 30$), onda koristimo statistiku z koja ima normalnu raspodelu. Ako je izračunata vrednost statistike $z^* = 2,35$, tada, sa pragom značajnosti $\alpha = 0,05$, odbacujemo nultu hipotezu (jer je $|z^*| = 2,35 > 1,96$), dok sa pragom značajnosti $\alpha = 0,01$ prihvatamo nultu hipotezu (jer je $|z^*| = 2,35 < 2,58$). Zbog toga se u novijoj literaturi, a posebno u programskim paketima za statistiku, umesto praga značajnosti α sve više koristi takozvana $p-$vrednost. $P-$vrednost je verovatnoća da će se realizovati ona vrednost statistike testa koja je upravo izračunata iz uzorka ili neka vrednost koja je još manje verovatna, ako je nulta hipoteza tačna. Ako je, u navedenom primeru, recimo $z^* = 2,94$, može da se pokaže da je $p = P(|z| \geq 2,94) = 0,0032$. Ova $p-$vrednost pokazuje da, prosečno, 32 od 10000 uzoraka od n elemenata ima ovakvu vrednost statistike testa, ako je nulta hipoteza tačna. Ovaj uzorak pokazuje jake dokaze protiv nulte hipoteze, pa se ona odbacuje sa rizikom $p = 0,0032$.

Može se reći da je $p-$vrednost najmanji prag značajnosti sa kojim se nulta hipoteza može odbaciti na osnovi podataka iz uzorka.

Izračunavanje $p-$vrednosti za statistiku koja nema normalnu raspodelu je znatno složenije. U programskim paketima za statistiku se prilikom testiranja hipoteze izračunava $p-$vrednost na osnovi koje se donosi zaključak

da li se hipoteza prihvata ili odbacuje. Uobičajeno je da se za $p < 0,05$ hipoteza odbacuje. Naravno da nije svejedno da li je, na primer, $p = 0,0497$ ili $p = 0,0001$. U drugom slučaju hipoteza se odbacuje sa vrlo malim rizikom da je ona tačna.

5

Neparametarski testovi

Većina testova koje smo do sada razmatrali, a svi su se odnosili na parametre raspodele nekog obeležja, imala je pretpostavku da posmatano obeležje ima normalnu raspodelu. U slučaju da ovaj uslov nije ispunjen ovi testovi ne mogu da se koriste. S druge strane, neka obeležja su nenumerička, pa nema smisla računati, na primer, aritmetičku sredinu ili standardnu devijaciju. Može se, recimo, testitati hipoteza da su dva nenumerička obeležja nezavisna. Nekada je, pored parametara raspodele, potrebno znati i oblik raspodele, odnosno testirati hipotezu da određeno obeležje ima, na primer, normalnu raspodelu.

Za sve parametarske testove, koje smo do sada razmatrali, a koji su zahtevali normalnu raspodelu, postoje odgovarajući neparametarski testovi koji ne zahtevaju normalnu raspodelu.

5.1 χ^2 test za tabele kontigencije

Ovaj test služi za testiranje hipoteze H_0 da su dva opisna obeležja X i Y, nezavisni, protiv alternativne hipoteze da nisu nezavisni.

Neka su $x_1, x_2, \ldots, x_r$ kategorije ("vrednosti") obeležja X, a $y_1, y_2, \ldots, y_r$ kategorije obeležja Y. Neka je u uzorku obima n konstatovano $f_{i,j}$ slučajeva kod kojih je $X = x_i$ i $Y = y_j$ $(i = 1, \ldots, r;\ j = 1, \ldots, s)$. Rezultati se prikazuju u tabeli kontigencije

	y_1	y_2	. . .	y_s	V_i
x_1	f_{11}	f_{12}	. . .	f_{1s}	V_1
x_2	f_{21}	f_{22}	. . .	f_{2s}	V_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_r	f_{r1}	f_{r2}	. . .	f_{rs}	V_r
K_j	K_1	K_2	. . .	K_s	n

Tabela 5.4.

gde je V_i zbir elemenata i-te vrste a K_j zbir elemenata j-te kolone. Primetimo da kada se saberu zbirovi vrsta dobija se broj n a isto važi kada se saberu zbirovi kolona. "Teorijsku" frekvenciju $f_{t_{i,j}}$ izračunavamo kada proizvod zbira i-te vrste i zbira j-te kolone podelimo sa ukupnim broj elemenata uzorka t.j.

$$f_{t_{i,j}} = \frac{V_i \cdot K_j}{n} \tag{5.2}$$

Može da se pokaže da statistika

$$\chi^2_{(r-1)(s-1)} = \frac{(f_{1,1} - f_{t_{1,1}})^2}{f_{t_{1,1}}} + \frac{(f_{1,2} - f_{t_{1,2}})^2}{f_{t_{1,2}}} + \cdots + \frac{(f_{r,s} - f_{t_{r,s}})^2}{f_{t_{r,s}}} \tag{5.3}$$

ima približno χ^2 raspodelu sa $(r-1)(s-1)$ stepena slobode, pod pretpostavkom da je hipoteza H_0 tačna i da je n veliko.

Neka je χ^2_* vrednost izračunata iz uzorka, prema formuli (5.3). Iz Tabele III čitamo broj $\chi^2_{(r-1)(s-1);\,\alpha}$. Ako je $\chi^2_* \geq \chi^2_{(r-1)(s-1);\,\alpha}$, onda se odbacuje hipoteza H_0 o nezavisnosti obeležja X i Y. Ako je $\chi^2_* < \chi^2_{(r-1)(s-1);\,\alpha}$, nema razloga za odbacivanje hipoteze H_0 .

Napominjemo da empirijske frekvencije $f_{i,j}$ ne treba da budu manje od 5.

Primer 5.1. Među bolesnicima odeljenja A i B sprovedena je anketa o tome da li su zadovoljni negom i dobijeni su sledeći rezultati:

	zadovoljni	nezadovoljni	V_i
odeljenje A	17	7	24
odeljenje B	10	6	16
K_j	27	13	40

Tabela 5.5

Da li su mišljenje bolesnika (zadovoljan, nezadovoljan) i odeljenje na kome se nalaze (A, B) zavisni?

Rešenje. Testira se nulta hipoteza da su mišljenje bolesnika i odeljenje nezavisni. Teorijske frekvencije se izračunavaju korišćenjem formule (5.2):

$$f_{t_{1,1}} = \frac{V_1 \cdot K_1}{n} = \frac{24 \cdot 27}{40} = 16,2 \qquad f_{t_{1,2}} = \frac{V_1 \cdot K_2}{n} = \frac{24 \cdot 13}{40} = 7,8$$

$$f_{t_{2,1}} = \frac{V_2 \cdot K_1}{n} = \frac{16 \cdot 27}{40} = 10,8 \qquad f_{t_{2,2}} = \frac{V_2 \cdot K_2}{n} = \frac{16 \cdot 13}{40} = 5,2$$

Koristeći formulu (5.3) izračunavamo vrednost statistike

$$\chi^2_* = \frac{(17-16,2)^2}{16,2} + \frac{(7-7,8)^2}{7,8} + \frac{(10-10,8)^2}{10,8} + \frac{(6-5,2)^2}{5,2} = 0,30.$$

Broj stepena slobode je $k = (r-1)(s-1) = (2-1)(2-1) = 1$. Neka je prag značajnosti $\alpha = 0,05$. Iz tabele χ^2 raspodele čitamo broj $\chi^2 = 3,841$. Kako je $\chi^2_* = 0,30 < 3,84 = \chi^2_{1;\,0,05}$ prihvatamo hipotezu da odgovori bolesnika ne zavise od odeljenja.

Primer 5.2. Ispitati da li su boja očiju i boja kose ljudi nezavisni, na osnovi uzorka od $n = 691$ osobe za koje su podaci dati u Tabeli 5.6.

	svetla	smeđa	crna	crvena	V_i
plave	176	81	19	9	285
zelene	95	139	75	8	317
tamne	11	44	29	5	89
K_j	282	264	123	22	691

Tabela 5.6.

Teorijske frekvencije, koje se izračunavaju prema formuli (5.2), date su u Tabeli 5.7.

	svetla	smeđa	crna	crvena
plave	116	109	51	9
zelene	129	121	56	10
tamne	36	34	16	3

Tabela 5.7.

Koristeći vrednosti iz Tabele 5.6. i Tabele 5.7. izračunavamo vrednost statistike

$$\begin{aligned}\chi^2_* &= \frac{(176-116)^2}{116} + \frac{(81-109)^2}{109} + \frac{(19-51)^2}{51} + \frac{(9-9)^2}{9} + \frac{(95-129)^2}{129} + \frac{(139-121)^2}{121} \\ &+ \frac{(75-56)^2}{56} + \frac{(8-10)^2}{10} + \frac{(11-36)^2}{36} + \frac{(44-34)^2}{34} + \frac{(29-16)^2}{16} + \frac{(5-3)^2}{3} \\ &= 108,99.\end{aligned}$$

Broj stepena slobode je $k = (r-1)(s-1) = (3-1)(4-1) = 6$. Neka je prag značajnosti $\alpha = 0,01$. Iz tabele χ^2 raspodele čitamo broj $\chi^2_{6;\,0,01} = 16,81$. Kako je $\chi^2_* = 108,99 > 16,812 = \chi^2_{6;\,0,01}$ odbacujemo hipotezu o nezavisnosti boje očiju i boje kose.

Ako je utvđeno da su dva obeležja zavisna, postavlja se pitanje jačine te zavisnosti. Intezitet međusobne veze posmatranih obeležja meri se **koeficijentom kontigencije** koji se izračunava po formuli

$$C = \sqrt{\frac{\chi^2_*}{n + \chi^2_*}}$$

Vrednosti koeficijenta kontigencije se nalaze između 0 i 1. Što je koeficijent kontigencije bliži jedinici to je veza između posmatranih obeležja jača.

U prethodnom primeru koeficijent kontigencije je

$$C = \sqrt{\frac{\chi^2_*}{n + \chi^2_*}} = \sqrt{\frac{108,99}{691 + 108,99}} = 0,369.$$

Kada su frekvencije u Tabeli kontigencije tipa 2×2 male (manje od 5, uključujući i nulu) koristi se **test stvarne verovatnoće** ili **Fisher-ov test**.

Nekada se upoređuju frekvencije jedne grupe pri različitim merenjima. Na primer, među bolesnicima jednog odeljenja sprovedena je anketa o tome da li

su zadovoljni negom u jednoj smeni a zatim su isti bolesnici anketirani da li su sadovoljni negom u drugoj smeni. Treba ispitati da li odgovori pacijenata zavise od smene. U ovakvim slučajevima treba primeniti χ^2 test za dva zavisna uzorka koji se zove **Mac Nemar-ov test**.

5.2 Mann-Whitney-ev test

Kada se porede srednje vrednosti nekog obeležja dveju populacija, najčešće se koristi $t-$test, pri čemu se porede aritmetičke sredine. Uslov za primenu t-testa je da obeležje ima normalnu raspodelu u obe populacije. Međutim, često se dešava da obeležje nema normalnu raspodelu. U tom slučaju se koristi **test sume rangova** ili **Mann-Whitney-ev test**. Ovaj test se koristi i kada su vrednosti obeležja date u vidu rangova. Mann-Whitney-evim testom se testira nulta hipoteza da dato obeležje ima istu raspodelu u obe populacije. Ova nulta hipoteza znači da su vrednosti obeležja jedne i druge populacije slično raspoređene, odnosno da nema bitne razlike u vrednostima obeležja jedne i druge populacije. Odbacivanje nulte hipoteze bi značilo da se vrednosti obeležja jedne populacije, većim delom, nalaze ispred obeležja druge populacije, odnosno da je razlika u vrednostima obeležja jedne i druge populacije statistički značajna.

Kada neko obeležje nema normalnu raspodelu, onda medijana dobro reprezentuje vrednosti tog obeležja. Zato se kod primene Mann-Whitney-evog testa, umesto aritmetičkih sredina, obično prikazuju medijane.

Iz dveju populacija uzimamo po jedan uzorak. Neka je n_1 broj elemenata manjeg uzorka a n_2 broj elemenata većeg uzorka. Ako je $n_1 = n_2$, tada se za izračunavanje statistike testa može uzeti jedan od dva uzorka. Od elemenata ova dva uzorka (kojih ima $n_1 + n_2$) formira se niz čiji su elementi poređani po veličini, pri čemu znamo koji je element iz kog uzorka. Najmanjoj vrednosti obeležja dodeljuje se rang 1, sledećoj po veličini vrednosti rang 2 itd. a rang $n_1 + n_2$ dobija najveće obeležje. Ukoliko postoje jednake vrednosti obeležja, onda se uzima njihov prosečni rang.

Zbir rangova manjeg uzorka obeležavamo sa T_{n_1} a zbir rangova većeg uzorka obeležavamo sa T_{n_2}.

Ukupan broj elemenata koji se rangiraju je $n = n_1 + n_2$. Njihovi rangovi su

$1, 2, \ldots, n$. Koristeći formulu

$$1 + 2 + \cdots + n = \frac{n(n+1)}{2}$$

za zbir prvih n prirodnih brojeva, dolazimo do zaključka da zbirovi T_{n_1} i T_{n_2} zadovoljavaju uslov

$$T_{n_1} + T_{n_2} = \frac{n(n+1)}{2}.$$

Napominjemo da se ovaj uslov koristi da bi se proverilo da li su zbirovi T_{n_1} i T_{n_2} tačno izračunati.

Statistika testa sume rangova jednaka je zbiru rangova u manjem uzorku t.j. T_{n_1}.

Iz Tabele VII, u preseku n_1-te kolone i n_2-te vrste, nalaze se dve vrednosti $T^a_{n_1;\,n_2}$ i $T^b_{n_1;\,n_2}$. Kritična oblast testa je $(-\infty, T^a_{n_1;\,n_2}] \cup [\,T^b_{n_1;\,n_2}, +\infty)$. Drugim rečima, ako se dobijena vrednost T_{n_1} nalazi **između** $T^a_{n_1;\,n_2}$ i $T^b_{n_1;\,n_2}$, **nemamo** razloga da odbacimo nultu hipotezu da dato obeležje ima istu raspodelu u obe populacije. U suprotnom, odbacujemo nultu hipotezu i kažemo da je razlika u raspodelama datog obeležja između populacija statistički značajna.

Napominjemo da je Tabela VII napravljena samo za prag značajnosti $\alpha = 0,05$.

Ako su n_1 i n_2 dovoljno veliki, statistika T_{n_1} ima približno normalnu raspodelu $N(\mu, \sigma^2)$, gde je

$$\mu = \frac{n_1(n_1 + n_2 + 1)}{2}, \qquad \sigma^2 = \frac{n_1 \cdot n_2(n_1 + n_2 + 1)}{12},$$

odnosno statistika

$$z = \frac{T_{n_1} - \frac{n_1(n_1+n_2+1)}{2}}{\sqrt{\frac{n_1\, n_2\, (n_1+n_2+1)}{12}}} \tag{5.4}$$

ima približno standardizovanu normalnu raspodelu $N(0, 1)$. Iz populacija se uzimaju uzorci obima n_1 i n_2 a zatim se izračunava vrednost z^*, prema formuli (5.4). Za dati prag značajnosti α, koristeći tabelu normalne raspodele određujemo broj c takav da je

$$P(|z^*| \geq c) = \alpha.$$

Ako konstatujemo da je $|z^*| \geq c$, odbacujemo nultu hipotezu i kažemo da je razlika u raspodelama datog obeležja između populacija statistički značajna. U suprotnom nultu hipotezu prihvatamo.

Primer 5.3. Iz populacije ispitanika koji boluju od bolesti A izabran je uzorak od 7 ispitanika i određen im je broj leukocita. Iz populacije ispitanika koji boluju od bolesti B izabran je uzorak od 9 ispitanika i takođe im je određen broj leukocita. Rezultati (u hiljadama) su dati u Tabeli 5.8. Ispitati da li je razlika u broju leukocita između ispitanika koji boluju od bolesti A i ispitanika koji boluju od bolesti B statistički značajna.

A	3,2	9,4	4,2	6,2	6,2	9,4	3,7		
B	9,4	5,1	10,1	7,8	12,4	8,4	10,2	12,4	54,3

Tabela 5.8.

Rešenje. Može se pokazati (na primer, korišćenjem programskog paketa za statistiku) da broj leukocita u populaciji bolesnika koji boluju od bolesti B nema normalnu raspodelu. Zato testiramo nultu hipotezu da obe populacije imaju istu raspodelu broja leukocita.

U Tabeli 5.9. u koloni "(A, B)" poređani su po veličini elementi oba uzorka. U koloni "$rang(A, B)$" određeni su rangovi združenog uzorka. Pošto su vrednosti obeležja pod rednim brojem 5 i 6 jednake (6,2), za rang se uzima aritmetička sredina $\frac{5+6}{2}$ t.j. 5,5. Za redne brojeve 9, 10 i 11 vrednosti obeležja su takođe jednake (9,4) pa se za rang ovih elemenata uzima aritmetička sredina $\frac{9+10+11}{3}$ t.j. broj 10. Slično se radi za redne brojeve 14 i 15. U koloni "$rang(A)$" dati su rangovi ispitanika koji boluju od bolesti A a u koloni "$rang(B)$" su dati rangovi ispitanika koji boluju od bolesti B. U Tabeli 5.9. su dati i odgovajući zbirovi rangova $T_7 = 37$ i $T_9 = 99$.

Ukupan broj elemenata koji se rangiraju je $n = n_1 + n_2 = 7 + 9 = 16$. Otuda je $\frac{n(n+1)}{2} = \frac{16(16+1)}{2} = 136$. S druge strane, zbir rangova je $T_7 + T_9 = 37 + 99 = 136$, pa je uslov $T_{n_1} + T_{n_2} = \frac{n(n+1)}{2}$ zadovoljen.

r.b.	(A,B)	rang(A,B)	bolest	rang(A)	rang(B)
1	3,2	1	A	1	
2	3,7	2	A	2	
3	4,2	3	A	3	
4	5,1	4	B		4
5	6,2	5,5	A	5,5	
6	6,2	5,5	A	5,5	
7	7,8	7	B		7
8	8,4	8	B		8
9	9,4	10	A	10	
10	9,4	10	A	10	
11	9,4	10	B		10
12	10,1	12	B		12
13	10,2	13	B		13
14	12,4	14,5	B		14,5
15	12,4	14,5	B		14,5
16	54,3	16	B		16
T_{n_i}				37	99

Tabela 5.9.

U Tabeli VII nalazimo $T^a_{7;\,9} = 40$ i $T^b_{7;\,9} = 79$. Pošto se dobijena vrednost $T_7 = 37$ ne nalazi između 40 i 79, odbacujemo nultu hipotezu da obe populacije imaju istu raspodelu broja leukocita, odnosno da je razlika u broju leukocita između ispitanika koji boluju od bolesti A i ispitanika koji boluju od bolesti B statistički značajna. Primetimo da su medijane $M_1 = 6,2$ i $M_2 = 10,1$. Znači da ispitanici koji boluju od bolesti B imaju značajno veći broj leukocita nego ispitanici koji boluju od bolesti A.

Ako koristimo aproksimativnu formulu statistike T_{n_1} t.j formulu (5.4) dobijamo

$$z^* = \frac{T_{n_1} - \frac{n_1(n_1+n_2+1)}{2}}{\sqrt{\frac{n_1\, n_2\, (n_1+n_2+1)}{12}}} = \frac{37 - \frac{7\cdot(7+9+1)}{2}}{\sqrt{\frac{7\cdot 9\cdot(7+9+1)}{12}}} = -2,38.$$

Iz uslova $P(|z| \geq c) = 0,05$, koristeći funkciju Φ, dobijamo $c = 1,96$. Kako je $|z^*| = 2,38 > 1,96$, odbacujemo nultu hipotezu.

Ovo znači da korišćenjem statistike (5.4) dobijamo isti rezultat.

5.3 Wilcoxon-ov test ekvivalentnih parova

Ovaj test odgovara uparenom t-testu, gde se radi sa parovima vrednosti obeležja $(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)$ i koristi se kada nije ispunjen uslov za primenu uparenog t-testa (da razlike parova imaju normalnu raspodelu). Nulta hipoteza je da su razlike između parova u populaciji jednake nuli.

Postupak za primenu Wilcoxonovog testa počinje izračunavanjem razlika $d_1 = x_1 - y_1, d_2 = x_2 - y_2, \ldots, d_m = x_n - y_n$. Ako je neka razlika nula, ona se izostavlja iz daljeg postupka. Razlike se rangiraju prema svojim apsolutnim vrednostima. Ako je razlika pozitivna, dodeljuje joj se pozitivan rang a ako je negativna, dodeljuje joj se negativan rang. Zatim se izračunava zbir pozitivnih rangova (Σ^+) i zbir apsolutnih vrednosti negativnih rangova (Σ^-). Neka je T manji od ova dva zbira. Iz Tabele VIII čitamo broj $T_{n;\,\alpha}$, gde je $n (n \leq m)$ broj parova kod kojih razlike nisu jednake nuli. Ako je $T \leq T_{n;\,\alpha}$ odbacujemo hipotezu da su razlike između parova u populaciji jednake nuli. Ako je $T > T_{n;\,\alpha}$ nemamo razloga da odbacimo nultu hipotezu.

Ako je broj parova koji su ostali u postupku veliki ($n > 25$) statistika

$$z = \frac{T - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$$

ima standardizovanu normalnu raspodelu $N(0, 1)$. Kritična vrednost c se određuje kao i kod drugih testova kod kojih statistika testa ima standardizovanu normalnu raspodelu.

Primer 5.4. Grupi od 10 pacijenata meren je nivo depresivnosti na Hamiltonovoj skali pre uzimanja terapije. Posle tronedeljnog uzimanja određenog antidepresiva istim pacijentima je ponovo meren nivo depresivnosti. Dobijeni rezultati merenja su prikazani u Tabeli 5.10. (HAMD-pre i HAMD-posle).

Ispitanik	1	2	3	4	5	6	7	8	9	10
HAMD-pre	21	25	19	41	24	20	29	27	20	29
HAMD-posle	18	25	23	15	26	17	23	22	14	20

Tabela 5.10.

Koristeći Wilcoxon-ov test, ispitati da li je efekat ovog antidepresiva statistički značajan?

Rešenje. Testiraćemo nultu hipotezu da su razlike u nivoima depresivnosti pre i posle terapije jednake nuli. Najpre se izračunaju razlike nivoa depresivnosti pre i posle uzimanja antidepresiva (kolona "Razlike"). Zatim se odrede rangovi tih razlika, vodeći računa o tome da negativnoj razlici odgovara rang sa znakom minus (kolona "Rang").

Ispitanik	Hamd-pre	HAMD-posle	Razlike	Rang
2	21	18	3	2,5
2	25	25	0	-
3	19	23	-4	-4
4	41	15	26	9
5	24	26	-2	-1
6	20	17	3	2,5
7	29	23	6	6,5
8	27	22	5	5
9	20	14	6	6,5
10	29	20	9	8

Tabela 5.11.

Zbir rangova je

$\sum^+ = 2,5 + 9 + 2,5 + 6,5 + 5 + 6,5 + 8 = 40$

$\sum^- = |-4| + |-1| = 5.$

Manji od ova dva zbira je $T = 5$. Neka je $\alpha = 0,05$. Iz Tabele VIII čitamo $T_{9;\,0,05} = 6$. Kako je

$$T = 5 < 6 = T_{9;\,0,05},$$

odbacujemo hipotezu da su razlike u nivoima depresivnosti pre i posle terapije jednake nuli. Ovo znači da su razlike u nivoima depresivnosti pre i posle uzimanja antidepresiva statistički značajne, odnosno da je efekat ovog antidepresiva statistički značajan.

Primer 5.5. Grupi od 32 sportista mereno je vreme za koje pretrče 100 metara a zatim je grupa provela dve nedelje na visinskim pripremama. Ponovo im je

mereno vreme za koje pretrče 100 metara. Izračunate su razlike u postignutim vremenima pre i posle visinskih priprema, od kojih su dve jednake nuli. Razlike su rangirane i izračunati su zbir pozitivnih rangova i zbir apsolutnih vrednosti negativnih rangova: $\sum^+ = 159$ i $\sum^- = 306$. Ispitati da li je došlo do značajnih promena rezultata na 100 metara posle visinskih priprema.

Rešenje. Ovde je $T = 159$, jer je 159 manji zbir. Pošto su dve razlike bile jednake nuli, one su izbačene iz daljeg postupka tako da je $n = 32 - 2 = 30$. Kako je $n > 25$ izračunavamo vrednost statistike

$$z^* = \frac{T - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} = \frac{159 - \frac{30(30+1)}{4}}{\sqrt{\frac{30(30+1)(2\cdot 30+1)}{24}}} = -1,51.$$

Neka je $\alpha = 0,05$. Kako z ima standardizovanu normalnu raspodelu, iz uslova $P(|z| \geq c) = 0,05$, korišćenjem Tabele II, dobijamo $c = 1,96$. Pošto je

$$|z^*| = 1,51 < 1,96 = c$$

nemamo osnova da odbacimo hipotezu H_0, odnosno promene (razlike) u rezultatima trčanja na 100 metara posle visinskih priprema nisu statistički značajne.

5.4 Kruskal-Wallis-ov test

Ovaj test odgovara analizi varijansi, kojom se testira hipoteza $H_0(\mu_1 = \mu_2 = \cdots = \mu_k)$ o jednakosti aritmetičkih sredina obeležja u populacijama. Osnovna pretpostavka za primenu analize varijansi je da obeležje u svim populacijama ima normalnu raspodelu. Ako ovaj uslov nije ispunjen ili je obeležje dato tako da se njegove "vrednosti" mogu rangirati, primenjuje se Kruskal-Wallisov-ov test kojim se testira nulta hipoteza da dato obeležje ima istu raspodelu u svih r populacija, odnosno da nema razlika u vrednostima obeležja između populacija. I ovde se, umesto aritmetičkih sredina, kao dobri reprezenti vrednosti obeležja prikazuju medijane.

Neka su, kao kod analize varijansi, iz r uzoraka dobijeni podaci:

$$\begin{array}{cccc} x_{1,1}, & x_{1,2}, & \ldots, & x_{1,n_1} \\ x_{2,1}, & x_{2,2}, & \ldots, & x_{2,n_2} \\ \cdot & \cdot & \cdots & \cdot \\ x_{r,1}, & x_{r,2}, & \ldots, & x_{r,n_r} \end{array}$$

gde je $n_1 + n_2 + \cdots + n_r = n$.

Dobijeni podaci iz svih grupa (njih n) se zajedno rangiraju a zatim se izračunaju zbirovi rangova za svaki uzorak. Obeležimo ove zbirove sa $T_1, T_2, \ldots, T_r$. Ako su uzorci dovoljno veliki (veći od 5), onda statistika

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{r} \frac{T_i^2}{n_i} - 3\,(n+1)$$

ima χ^2 raspodelu sa $r-1$ stepena slobode. Za dati prag značajnosti α iz tablice za χ^2 raspodelu čita se broj $\chi^2_{r-1;\,\alpha}$. Ako je $H \geq \chi^2_{r-1;\,\alpha}$, odbacujemo nultu hipotezu da dato obeležje ima istu raspodelu u svim populacijama. Ako je $H < \chi^2_{r-1;\,\alpha}$ nemamo razloga da odbacimo nultu hipotezu.

U slučaju manjih uzoraka koriste se posebne tabele, koje ovde nećemo davati.

Primer 5.6. Iz populacija ispitanika koje su određene prema NYHA klasifikaciji, na slučajan način su izabrani ispitanici i meren je im je nivo hormona BNP. Rezultati merenja dati su u Tabeli 5.12.

NYHA = 1	21	23	1	4	5	11		
NYHA = 2	48	2	14	17	12			
NYHA = 3	25	13	15	22	16	111	19	7
NYHA = 4	106	45	127	96	128	109		

Tabela 5.12.

S pragom značajnosti $\alpha = 0,01$ ispitati da li su razlike u vrednostima BNP-a između populacija određenim NYHA klasifikacijom statistički značajne.

Rešenje. Poznato je da BNP nema normalnu raspodelu, pa primenjujemo Kruskal-Wallisov-ov test. U Tabeli 5.13 prikazane su vrednosti hormona BNP, odgovarajući rangovi, kao i zbirovi rangova sve četiri grupe.

NYHA=1	rang	NYHA=2	rang	NYHA=3	rang	NYHA=4	rang
21	14	48	19	25	17	106	21
23	16	2	2	13	8	45	18
1	1	14	9	15	10	127	24
4	3	17	12	22	15	96	20
5	4	12	7	16	11	128	25
11	6			111	23	109	22
				19	13		
				7	5		
T_i	44		49		102		130
n_i	6		5		8		6

Tabela 5.13.

Kako je $n = n_1 + n_2 + n_3 + n_4 = 6 + 5 + 8 + 6 = 25$, imamo

$$\begin{aligned} H &= \frac{12}{n(n+1)} \sum_{i=1}^{r} \frac{T_i^2}{n_i} - 3\,(n+1) \\ &= \frac{12}{25(25+1)}\left(\frac{44^2}{6} + \frac{49^2}{5} + \frac{102^2}{8} + \frac{130^2}{6}\right) - 3\,(25+1) = 12,83. \end{aligned}$$

Statistika H ima χ^2 raspodelu sa $r-1 = 4-1 = 3$ stepena slobode, jer u svakoj od četiri grupe ima više od 5 elemenata. Iz tabele χ^2 raspodele čitamo broj $\chi^2_{3;\,0,01} = 11,341$. Pošto je

$$H = 12,83,75 > 11,341 = \chi^2_{3;\,0,01}$$

odbacujemo hipotezu da dato obeležje ima istu raspodelu u sve četiri populacije, odnosno razlike u vrednostima BNP-a između populacija određenim NYHA klasifikacijom su statistički značajne.

Iz Tabele 5.12. mogu da se odrede medijane četiri grupe

$$M_1 = 8 \qquad M_2 = 14 \qquad M_3 = 17,5 \qquad M_4 = 107,5.$$

Imajući u vidu rezultat testa i vrednosti medijana, može se zaključiti da većem broju, prema NYHA klasifikaciji, (što znači težoj bolesti) odgovara veća vrednost hormona BNP.

5.5 Friedman-ov test

Kod Wilcoxonovog testa se dva puta (različitim metodama ili u različitim vremenskim intervalima) vrše merenja nekog obeležja nad istim skupom elemenata. Nekada se ta merenja vrše više od dva puta. Treba testirati nultu hipotezu da su razlike između ovih merenja jednake nuli.

Neka je nad uzorkom od n elemenata izvršeno k merenja pri čemu su dobijeni rezultati:

$$\begin{array}{cccc} x_{11}, & x_{12}, & \ldots, & x_{1k} \\ x_{21}, & x_{22}, & \ldots, & x_{2k} \\ . & . & \ldots & . \\ x_{n1}, & x_{n2}, & \ldots, & x_{nk} \end{array}$$

Prvo se vrši rangiranje elemenata za svaku vrstu pojedinačno a zatim se izračuna zbir rangova svake kolone. Te rangove obeležavamo sa $T_1, T_2, \ldots, T_k$. Statistika

$$H = \frac{12}{nk(k+1)} \sum_{i=1}^{k} T_i^2 - 3\,n\,(k+1)$$

ima χ^2 raspodelu sa $k-1$ stepena slobode. Za dati prag značajnosti α iz tablice za χ^2 raspodelu čita se broj $\chi^2_{k-1;\,\alpha}$. Ako je $H \geq \chi^2_{k-1;\,\alpha}$, odbacujemo nultu hipotezu da su razlike između merenja jednake nuli. Ako je $H < \chi^2_{k-1;\,\alpha}$ nemamo razloga da odbacimo nultu hipotezu.

Primer 5.7. Studenti su radili četiri zadatka iz statistike. Na slučajan način je odabrano 7 studenata i njihovi rezultati su prikazani su u Tabeli 5.14. Ispitati da li su ova četiri zadatka iste težine.

Rešenje. Napominjemo da jedna vrsta u Tabeli 5.14. odgovara jednom studentu. Elementi vrsta se rangiraju. Tako, na primer, u prvoj vrsti, prvom elementu (25), koji je najmanji, odgovara rang 1, drugom elementu (26) odgovara rang 2, a trećem i četvrtom elementu (29 i 29), pošto su jednaki, odgovara rang 3,5. Slično se određuju rangovi ostalih vrsta.

	Prvi test	Drugi test	Treći test	Četvrti test
Student	broj poena	broj poena	broj poena	broj poena
1	25	26	29	29
2	38	39	45	46
3	45	44	49	48
4	41	42	42	43
5	28	27	34	34
6	21	22	25	25
7	27	28	29	30

Tabela 5.14.

U Tabeli 5.15. dati su rangovi (određeni prema vrstama) i zbirovi rangova kolona.

	Prvi test	Drugi test	Treći test	Četvrti test
Student	rang	rang	rang	rang
1	1	2	3,5	3,5
2	1	2	3	4
3	2	1	4	3
4	1	2,5	2,5	4
5	2	1	3,5	3,5
6	1	2	3,5	3,5
7	1	2	3	4
T_i	9	12,5	23	25,5

Tabela 5.15.

Na osnovi dobijenih podataka izračunavamo vrednost statistike

$$\begin{aligned} H &= \tfrac{12}{nk(k+1)}(T_1^2+T_2^2+T_3^2+T_4^2) - 3\,n\,(k+1) \\ &= \tfrac{12}{7\cdot 4\cdot (4+1)}\Big(9^2+(12,5)^2+23^2+(25,5)^2\Big) - 3\cdot 7\cdot (4+1) = 16,41. \end{aligned}$$

U Tabeli III nalazimo vrednost $\chi^2_{r-1;\,\alpha} = \chi^2_{3;\,0,01} = 11,341$. Kako je

$$H = 16,41 > 11,341 = \chi^2_{3;\,0,01}$$

odbacujemo hipotezu da su razlike između zadataka jednake nuli t.j. zadaci nisu iste težine.

Ako bismo hteli da vidimo između kojih zadataka je razlika značajna, trebalo bi da se primeni Wilcoxon-ov test.

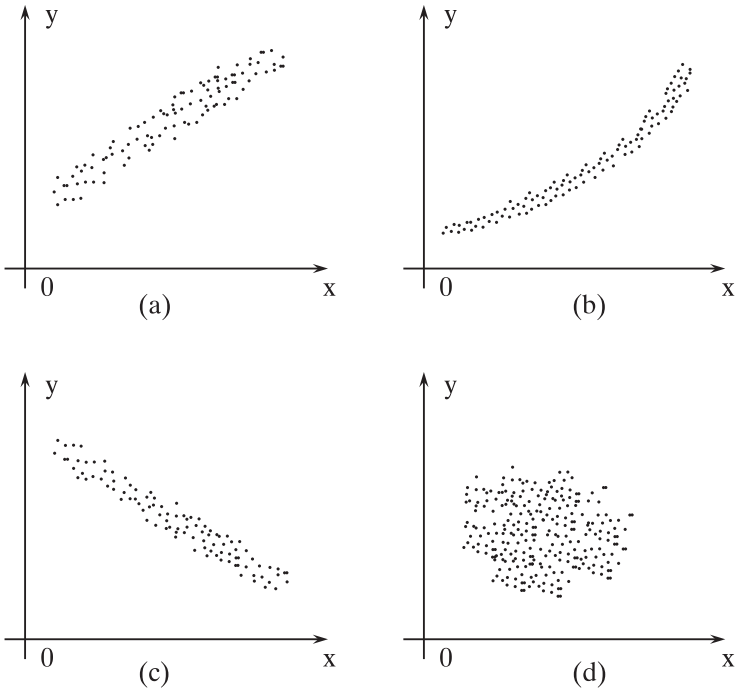
6

Linearna regresija i korelacija

6.1 Linearna regresija

Ako su u određenoj populaciji data dva obeležja X i Y, onda može da se proučava povezanost između ta dva obeležja. Naime, često promena jednog obeležja utiče na promenu drugog, zbog međusobne povezanosti. Ova povezanost može da se razlikuje po jačini, smeru i obliku povezanosti. Smer povezanosti je pozitivan ako pri rastu vrednosti jednog obeležja rastu i vrednosti drugog obeležja. Na primer, ako je X visina a Y težina ljudi, tada, po pravilu, viši ljudi imaju i veću težinu, odnosno između ovih obeležja postoji pozitivan smer povezanosti. Povezanost između obeležja može se posmatrati i po jačini poveznosti. Najjača veza između dva obeležja je funkcionalna veza, kada svakoj vrednosti jednog obeležja odgovara tačno jedna određena vrednost drugog obeležja. Labavija veza između obeležja je ona koja je podložna manjim ili većim odstupanjima i naziva se korelativna ili stohastička veza. Ima i takvih obeležja koja ne pokazuju nikakvu vezu i za njih kažemo da su nekorelativna. Postoje takođe različiti matematički oblici povezanosti - linearan, kvadratni, eksponencijalni itd. Ovde će biti reči o linearnom obliku. Prvu orijentaciju o obliku povezanosti među obeležjima X i Y daje grafičko prikazivanje uređenih parova $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ u koordinatnom sistemu, gde su $x_1, x_2, \ldots, x_n$ vrednosti obeležja X a $y_1, y_2, \ldots, y_n$ su vrednosti obeležja Y. Skup tačaka koje odgovaraju parovima (x_i, y_i) $(i = 1, \ldots, n)$ naziva se dijagram raspršivanja. Prema rasporedu tačaka na dijagramu može se utvrditi

oblik povezanosti, smer a donekle i jačina povezanosti. Na Slici 6.1.(a) i Slici 6.1.(b) vide se pozitivne korelacije (većoj vrednosti x odgovara veća vrednost y). Prva od njih ima linearan oblik a druga eksponencijalni. Na Slici 6.1.(c) se vidi negativna linearna korelacija (većoj vrednosti x odgovara manja vrednost y), dok na Slici 6.1.(d) vidimo da su obeležja nekorelativna. Linija koja najbolje reprezentuje raspored tačaka naziva se linija regresije.



Slika 6.1.

Regresiona prava

Za najbolju linearnu vezu, takozvanu regresionu pravu, uzima se ona prava $y = a+bx$ za koju je zbir kvadrata "vertikalnih rastojanja" tačaka od regresione prave najmanji t.j. za koju je veličina

$$S(a,b) = (a + bx_1 - y_1)^2 + (a + bx_2 - y_2)^2 + \cdots + (a + bx_n - y_n)^2$$

najmanja. Može da se pokaže (kada se parcijalni izvodi funkcije $S(a,b)$ po a i

po b izjednače sa nulom) da je $S(a,b)$ najmanje kada je

$$na + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$
$$a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

odakle se, rešavanjem sistema jednačina po a i b, dobija

$$b = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x}\,\overline{y}}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2}, \qquad a = \overline{y} - b\overline{x}.$$

Primer 6.1. Iz populacije jedne vrste životinja na slučajan način je uzet uzorak od 10 životinja. U Tabeli 6.1. su dati starost životinja i težina životinja.

Životinja	starost	težina
A	2	4
B	3	3
C	3	5
D	4	6
E	5	7
F	5	8
G	6	9
H	7	12
I	8	10
J	10	14

Tabela 6.1.

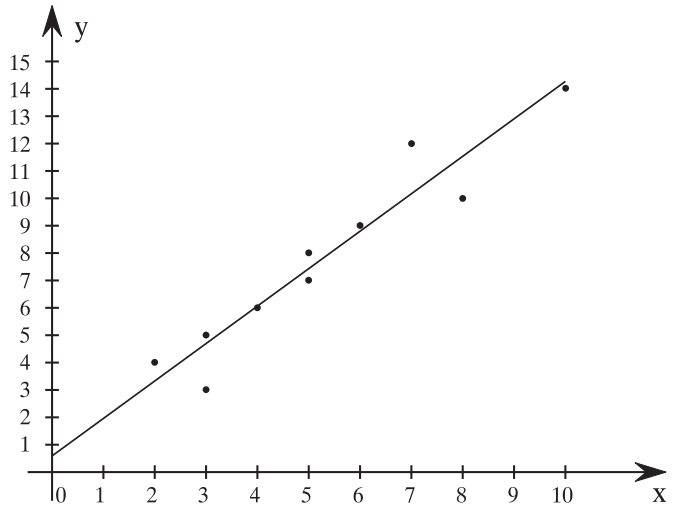
Nacrtati dijagram raspršivanja i napisati jednačinu regresione prave.

Rešenje. Obeležimo starost životinje sa X a težinu sa Y. U Tabeli 6.2. su dati zbirovi potrebni za određivanje regresione prave.

Životinja	x_i	y_i	x_i^2	$x_i\, y_i$
A	2	4	4	8
B	3	3	9	9
C	3	5	9	15
D	4	6	16	24
E	5	7	25	35
F	5	8	25	40
G	6	9	36	54
H	7	12	49	84
I	8	10	64	80
J	10	14	100	140
Σ	53	78	337	489

Tabela 6.2.

Dijagram raspršivanja dat je na Slici 6.2.



Slika 6.2.

Iz Tabele 6.2. dobijamo

$\overline{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = \frac{1}{10}\cdot 53 = 5,3 \qquad \overline{y} = \frac{1}{10}\sum_{i=1}^{10} y_i = \frac{1}{10}\cdot 78 = 7,8$

$\sum_{i=1}^{10} x_i^2 = 337, \qquad \sum_{i=1}^{10} x_i y_i = 489,$

odakle se izračunavaju parametri regresione prave

$$b = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x}\,\overline{y}}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2} = \frac{\frac{1}{10}\cdot 489 - 5,3\cdot 7,8}{\frac{1}{10}\cdot 337 - 5,3^2} = 1,3476$$

$$a = \overline{y} - b\overline{x} = 7,8 - 1,3476\cdot 5,3 = 0,6577.$$

Jednačina regresione prave je $y = 0,6577 + 1,3476\,x$. Grafik ove prave prikazan je na Slici 6.2.

Testiranje značajnosti regresione veze

Kod regresione prave vrednost a predstavlja odsečak na y-osi i nema poseban značaj. Ako bi između promenljivih x i y u populaciji postojala funkcionalna veza oblika $y = \alpha + \beta\,x$, onda bi koeficijent β pokazivao za koliko se promeni y kada se x promeni za jednu jedinicu. Međutim, regresioni koeficijent b predstavlja ocenjenu vrednost **prosečne** promene zavisne promenljive (dobijene iz uzorka) kada se nezavisna promenljiva promeni za jednu jedinicu. To bi u Primeru 6.1 značilo da ako je zivotinja starija za jedan mesec, onda je njena težina, u proseku, veća za 1,35 kg. Na ovaj način bi se vršilo predviđanje težine životinje u zavisnosti od starosti životinje. Postavlja se pitanje koliko je to predviđanje pouzdano.

Koeficijenti regresione prave a i b mogu da se izračunaju i kada praktično ne postoji nikakva linearna veza između obeležja X i obeležja Y. Da bi primena regresione prave (u smislu predviđanja vrednosti y) bila korektna, potrebno je da se ispita da li u populaciji postoji linearna veza između obeležja X i Y. Zbog toga treba testirati hipotezu $H_0(\beta = 0)$, gde je $y = \alpha x + \beta$ jednačina koja važi za populaciju. Naime, koeficijent β je jednak nuli ako je Y konstanta ili ako X ne utiče na Y. Za testiranje ove hipoteze koristi se statistika

$$t_{n-2} = \frac{b\cdot\sqrt{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}}{\sqrt{\frac{\sum_{i=1}^{n} y_i^2 - a\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i y_i}{n-2}}} \tag{6.1}$$

koja ima Studentovu raspodelu sa $n-2$ stepena slobode. Za dati prag značajnosti α iz tabele za Studentovu raspodelu nalazimo $t_{n-2;\,\alpha}$. Ako je $|t^*_{n-2}| \geq t_{n-2;\,\alpha}$ odbacujemo nultu hipotezu $H_0(\beta = 0)$, odnosno zaključujemo da postoji linearna veza t.j da X utiče na Y. U slučaju $|t_{n-2}|^* < t_{n-2;\,\alpha}$ prihvatamo nultu hipotezu, odnosno zaključujemo da nema linearnog uticaja X na Y.

Primer 6.2. Ispitati linearnu zavisnost između starosti životinja i težine životinja iz Primera 6.1.

Rešenje. Iz Primera 6.1. imamo

$\overline{x} = 5,3 \qquad \sum_{i=1}^{10} y_i = 78 \qquad \sum_{i=1}^{10} x_i^2 = 337$

$\sum_{i=1}^{10} x_i y_i = 489 \qquad a = 0,6577 \qquad b = 1,3476.$

Za korišćenje formule (6.1) treba još izračunati

$$\sum_{i=1}^{10} y_i^2 = 4^2 + 3^2 + 5^2 + 6^2 + 7^2 + 8^2 + 9^2 + 12^2 + 10^2 + 14^2 = 720.$$

Iz prethodnih vrednosti se dobija

$$t^*_{n-2} = \frac{b \cdot \sqrt{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}}{\sqrt{\frac{\sum_{i=1}^{n} y_i^2 - a\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i y_i}{n-2}}} = \frac{1,3476 \cdot \sqrt{337 - 10 \cdot (5,3)^2}}{\sqrt{\frac{720 - 0,6577 \cdot 78 - 1,3476 \cdot 489}{10-2}}} = 9,17.$$

Kako je $|t^*_{n-2}| = 9,17 > 3,355 = t_{8;\,0,01}$ odbacujemo hipotezu $H_0(\beta = 0)$, što znači da postoji linearna veza, odnosno da starost životinje utiče na njenu težinu.

Višestruka linearna regresija

Nekada nas interesuje istovremeni uticaj više obeležja na neko obeležje. Na primer, interesuje nas uticaj visine, težine i pola na sistolni pritisak dece. Uopšte u pitanju su veze oblika

$$y = a + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n,$$

gde su $x_1, x_2, \ldots, x_n$ nezavisne promenljive, y je zavisna promenljiva, a je konstanta a $b_1, b_2, \ldots, b_n$ su parcijalni regresioni koeficijenti.

Ovde neće biti reči o tome kako se određuju $a, b_1, b_2, \ldots, b_n$, već se čitalac upućuje na programske pakete za statistiku. Tako se za jedan konkretan primer od desetoro dece dobija

$$\text{pritisak} = 79,44 - 0,03 \cdot \text{visina} + 1,18 \cdot \text{težina} + 4,23 \cdot \text{pol}$$

gde se sa 0 obelažava pol dečaka a sa 1 pol devojčica.

Kod višestruke regresije, slično jednostrukoj regresiji, se testira značajnost parcijalnih koeficijenata regresije $b_1, b_2, \ldots, b_n$, kako bi se videlo koje promenljive značajno utiču na promenljivu y. Statistički paketi daju verovatnoće za svaki parcijalni koeficijent regresije. Ako je verovatnoća koja odgovara koeficijentu b_i manja od 0,05 to znači da je uticaj promenljive x_i na zavisno promenljivu y značajan. Ako je uticaj neke promenljive x_i na y značajan, onda to znači ako se x_i poveća za 1 tada se se y, u proseku, poveća za b_i, pod uslovom da se vrednosti ostalih promenljivih ne promene.

U primeru sa sistolnim pritiskom kod dece, uticaj težine dece na sistolni pritisak je značajan ($p = 0,001$). Sa porastom težine od jednog kilograma sistolni pritisak se, u proseku, povećava za 1,18 mmHg, pod uslovom da su visina i pol konstantni. Uticaj visine dece na sistolni pritisak dece nije značajan ($p = 0,860$). Ovo je posledica činjenice da su visina i težina dece u korelaciji (deca sa većom visinom imaju i veću težinu), pa se uticaj visine odražava na sistolni pritisak preko težine. Postoji značajan uticaj pola deteta na sistolni pritisak ($p = 0,012$). Sistolni pritisak devojčica je, u proseku, viši od sistolnog pritiska dečaka (primetimo da je pol devojčica označen sa 1 a dečaka sa 0) za 4,23 mmHg.

U programskim paketima za statistiku postoje programi koji, kroz određeni broj iteracija, izbacuju one promenljive x_i koje nemaju značajan uticaj na y, a ostavljaju samo one čiji je uticaj značajan.

Binarna logistička regresija

Binarana logistička regresija je slična linearnoj regresiji, s tim što su vrednosti zavisne promenljive y binarne t.j. 0 i 1. Ove vrednosti mogu, na primer, da znače: prisustnost ili odsutnost simptoma, osoba ima ili nema bolest i slično.

Na primer, ako nas interesuje uticaj promenljivih: sistolni pritisak, nivo šećera u krvi i nivo holesterola na pojavu angine pectoris (ima anginu pectoris 1, nema anginu pectoris 0) onda možemo da koristimo logističku regresiju.

Jednačina binarne logističke regresije je

$$ln(p) = a + b_1x_1 + b_2x_2 + \cdots + b_nx_n,$$

gde su $b_1, b_2, \ldots, b_n$ logistički regresioni koeficijenti a p je procenjena vrednost verovatnoće da, na primer, neka osoba, koja ima određene vrednosti promenljivih $x_1, x_2, \ldots, x_n$, ima bolest. Veća vrednost p znači veću verovatnoću prisustva bolesti. I kod ove vrste regresije se testira značajnost parcijalnih koeficijenata regresije $b_1, b_2, \ldots, b_n$, kako bi se videlo koje promenljive značajno utiču na promenljivu y.

U programskim paketima za statistiku, pored regresionih koeficijenata b_i, daju se i vrednosti $exp(b_i)$ t.j. e^{b_i}. Veličina $exp(b_i)$ je takozvani količnik šanse (odds ratio), o kome će biti reči u poslednjem poglavlju. Za neku promenljivu x_i to je ocenjena šansa za $(x_i + 1)$ u odnosu na ocenjenu šansu za x_i, kada su vrednosti ostalih promenljivih konstantne. Ako je, na primer, $exp(b_i)$ jednak 2, to znači da ako se promenljiva x_i poveća za 1, onda se šansa da se, na primer, oboli od date bolesti, u proseku, povećava dva puta, pod uslovom da vrednosti ostalih promenljivih ostanu nepromenjene. Ako je, na primer, $exp(b_i)$ jednak $\frac{1}{3}$, šansa se smanjuje tri puta.

6.2 Linearna korelacija

Cilj linearne korelacije je da se utvrdi koliko je linearna veza između neka dva obeležja X i Y jaka. Pretpostavimo da bar jedno od obeležja X i Y ima normalnu raspodelu. Tada se za merenje jačine linearne veze između obeležja X i Y koristi **Pearsonov koeficijent korelacije** r koji se izračunava prema formuli:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \cdot \sum_{i=1}^{n}(y_i - \overline{y})^2}}.$$

Koeficijent korelacije r uzima vrednosti od -1 do 1. Ako je r pozitivno, to ukazuje da kada jedno obeležje raste onda raste i drugo. Ako je r negativno, to znači da kada jedno obeležje raste drugo opada.

Veličina $|r|$ ukazuje na to koliko su tačke na dijagramu raspršivanja bliske pravoj liniji. Što je vrednost $|r|$ bliža nuli, to je linearna veza između X i Y slabija. Što je $|r|$ bliže broju 1, to je linearna veza između X i Y jača. Ako je $|r|$ između 0,8 i 0,9 veza se naziva jaka a ako je $|r|$ između 0,9 i 1 veza se naziva vrlo jaka. Ako je $|r|$ manje od 0,8 veza takođe može da bude značajna.

U slučaju da između obeležja X i Y postoji funkcionalna linearna veza, onda je koeficijent korelacije $r = 1$.

Primer 6.3. U Tabeli 6.3 su date vrednosti za X i Y.

X	2	3	3	4	5	5	6	7	8	10
Y	4	3	5	6	7	8	9	12	10	14

Tabela 6.3.

Izračunati koeficijent linearne korelacije.

Rešenje. Rezulatati računanja dati su u Tabeli 6.4.

	x_i	y_i	$x_i - \overline{x}$	$y_i - \overline{y}$	$(x_i - \overline{x})^2$	$(y_i - \overline{y})^2$	$(x_i - \overline{x})(y_i - \overline{y})$
	2	4	-3,30	-3,80	10,89	14,14	12,54
	3	3	-2,30	-4,80	5,29	23,04	11,04
	3	5	-2,30	-2,80	5,29	7,84	6,44
	4	6	-1,30	-1,80	1,69	3,24	2,34
	5	7	-0,30	-0,80	0,09	0,64	0,24
	5	8	-0,30	0,20	0,09	0,04	-0,06
	6	9	0,70	1,20	0,49	1,44	0,84
	7	12	1,70	4,20	2,89	17,64	7,14
	8	10	2,70	2,20	7,29	4,84	5,94
	10	14	4,70	6,20	22,09	38,44	29,14
$\sum$	53	78			56,10	111,60	75,60

Tabela 6.4.

Deljenjem sa 10 iz prve kolone se dobija $\overline{x} = 5,3$ a iz druge $\overline{y} = 7,8$. Dalje se,

iz Tabele 6.4, dobija

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}} = \frac{75,60}{\sqrt{56,10 \cdot 111,60}} = 0,955.$$

Ovo pokazuje da između obeležja X i Y postoji vrlo jaka linearna veza.

Testiranje koeficijenta korelacije

Neka su X i Y obeležja neke populacije. Koeficijent linearne korelacije u populaciji se označava sa ϱ i njegova vrednost bi se mogla izračunati samo kada bi se znale vrednosti obeležja X i Y za sve elemente populacije. Postavlja se pitanje kako na osnovi koeficijenta korelacije r, dobijenog iz uzorka, doneti korektan zaključak da li u populaciji postoji linearna korelacija između X i Y.

Pretpostavimo da bar jedno od obeležja X i Y ima normalnu raspodelu.

Nulta hipoteza je $H_0(\varrho = 0)$, odnosno da u populaciji nema linearne korelacije između X i Y. Za testiranje ove hipoteze koristimo statistiku

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}.$$

Ako je hipoteza $H_0(\varrho = 0)$ tačna onda statistika t ima Studentovu raspodelu sa $n-2$ stepena slobode. Ako je $n > 30$, tada statistika t ima približno normalnu raspodelu $N(0,1)$. U zavisnosti od toga da li je alternativna hipoteza oblika $H_1(\varrho \neq 0)$, $H_1(\varrho > 0)$ ili $H_1(\varrho < 0)$ imamo dvostranu kritičnu oblast, desnostranu kritičnu oblast i levostranu kritična oblast. Kritična oblast se određuje slično kao kod t testa. Ako izračunata vrednost t^* pripada kritičnoj oblasti, onda, sa pragom značajnosti α, odbacujemo hipotezu $H_0(\varrho = 0)$ i kažemo da se koeficijent korelacije ϱ značajno (visoko značajno) razlikuje od 0, odnosno da u populaciji postoji značajna linearna veza između obeležja X i Y. U suprotnom, hipotezu prihvatamo i kažemo da u populaciji ne postoji značajna linearna veza između obeležja X i Y.

Primer 6.4. Testirati koeficijent linearne korelacije iz Primera 6.3.

Rešenje. Kako je $n = 10$ i $r = 0,955$ imamo

$$t^* = \frac{0,955 \cdot \sqrt{10-2}}{\sqrt{1-0,955^2}} = 9,11.$$

U Tabeli IV nalazimo $t_{n-2;\,\alpha} = t_{8;\,0,01} = 3,355$. Kako je

$$t^* = 9,11 > 3,355 = t_{8;\,0,01}$$

odbacujemo hipotezu $H_0(\varrho = 0)$ i zaključujemo da postoji značajna linearna veza između posmatranih obeležja.

Primer 6.5. Iz uzorka obima $n = 27$ izračunat je koeficijent korelacije $r = 0,60$. Testirati hipotezu $H_0(\varrho = 0)$ protiv alternativne hipoteze $H_1(\varrho > 0)$ s pragom značajnosti $\alpha = 0,01$.

Rešenje. Izračunajmo, najpre,

$$t^* = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0,60\sqrt{27-2}}{\sqrt{1-(0,60)^2}} = 3,75.$$

Pošto imamo desnostrani test, iz tablice za Studentovu raspodelu (Tabela IV) čitamo $t_{n-2;\,2\cdot\alpha} = t_{27-2;\,2\cdot 0,01} = t_{25;\,0,02} = 2,485$. Kako je $t^* = 3,75 > 2,485 = t_{25;\,0,02}$, odbacujemo hipotezu $H_0(\varrho = 0)$ i zaključujemo da postoji značajna linearna veza između posmatranih obeležja.

Primer 6.6. Iz uzorka od 163 elementa izračunat je koeficijent linearne korelacije $r = -0,23$. Može li se zaključiti da se odgovarajući koeficijent korelacije u populaciji bitno razlikuje od nule?

Rešenje. Najpre izračunavamo

$$t^* = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{-0,23\sqrt{163-2}}{\sqrt{1-(-0,23)^2}} = -2,89.$$

Neka je prag značajnosti $\alpha = 0,01$. Kritičnu vrednost c određujemo iz uslova $P(|t| \geq c) = 0,01$. Pošto imamo veliki uzorak ($n > 30$), koristimo normalnu raspodelu. Saglasno Primeru 1.10. dobija se $c = 2,58$.

Pošto je $|t^*| = 2,89 > 2,58 = c$ odbacujemo hipotezu $H_0(\varrho = 0)$ i kažemo da se koeficijent korelacije ϱ visoko značajno razlikuje od nule. Primetimo da je u ovom primeru apsolutna vrednost koeficijenta linearne korelacije uzorka relativno mala, a da se koeficijenta linearne korelacije populacije, ipak, statistički visoko značajno razlikuje od nule. To je zato što je broj elemenata $n = 163$ dosta veliki, što bitno utiče na statističku značajnost testa.

Spearmanov koeficijent korelacije ranga

Pretpostavka za primenu koeficijenta linearne korelacije je bila da bar jedno od obeležja X i Y ima normalnu raspodelu. U slučaju da ova pretpostavka nije zadovoljena, koriste se neparametarske metode korelacije. Najčešće se koristi Spearmanov koeficijent korelacije ranga. Ovaj koeficijent se takođe koristi kada podaci nisu dati kao numeričke vrednosti, već kao rangirani podaci.

Neka je dato n parova podataka koji se odnose na obeležja X i Y tog uzorka. Pretpostavlja se da se vrednosti obeležja X i Y mogu rangirati. Svakoj vrednosti x_i obeležja X dodeljujemo njen rang a_i i svakoj vrednosti y_i dodeljujemo njen rang b_i. Neka je $d_i = a_i - b_i$, $(i = 1, \ldots, n)$. Spearmanov koeficijent korelacije ranga datih podataka se izračunava prema formuli

$$r_S = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}.$$

Vrednosti Spearmanovog koeficijenta korelacije mogu biti između -1 i 1. Što je $|r_S|$ bliži broju 1 to je veći stepen povezanosti između X i Y.

Slično kao kod koeficijenta linearne korelacije, da bi se utvdilo da li je veza između obeležja X i Y značajna, testira se hipoteza $H_0(\varrho_S = 0)$, gde je ϱ_S Spearmanov koeficijent korelacije obeležja X i Y u populaciji. Ovde nećemo navoditi postupak testiranja ove hipoteze, već upućujemo čitaoca na programske pakete za statistiku, gde se prilikom izračunavanja Spearmanovog koeficijenta korelacije ranga istovremeno vrši i njegovo testiranje.

Primer 6.7. Iz uzorka od 10 studenata dobijeni su podaci o mestu na rang listi sa prijemnog ispita i broju poena na ispitu iz statistike. Podaci su prikazani u Tabeli 6.5. Ispitati korelaciju između mesta na rang listi sa prijemng ispita i broja poena iz statistike.

Rešenje. Pošto su podaci, koji se odnose na uspeh studenata na prijemnom ispitu, dati kao mesta na rang listi, korelacija se ispituje pomoću Spearmanovog koeficijenta korelacije ranga. Primetimo da su studenti u Tabeli 6.5 već rangirani prema uspehu na prijemnom ispitu pa je potrebno rangirati samo uspeh na ispitu iz statistike. Pošto studenti E, F i G imaju isti broj poena iz statistike (79) a njihovi rangovi su redom 5, 6 i 7, uzima se aritmetička sredina rangova t.j. $\frac{1}{3}(5 + 6 + 7) = 6$.

Student	Mesto na rang listi	Broj poena iz stat.
A	1	95
B	2	95
C	3	85
D	4	91
E	5	79
F	6	79
G	7	79
H	8	63
I	9	72
J	10	52

Tabela 6.5.

U Tabeli 6.6. su dati rangovi sa prijemnog ispita (a_i), ocene iz statistike (o_i), rangovi ocena iz statistike (b_i), razlike rangova rangova a_i i b_i (d_i), kvadrati razlika rangova (d_i^2) i zbir kvadrata razlika rangova (Σ).

Student	a_i	o_i	b_i	d_i	d_i^2
A	1	95	1,5	-0,5	0,25
B	2	95	1,5	0,5	0,25
C	3	85	4	-1	1
D	4	91	3	1	1
E	5	79	6	-1	1
F	6	79	6	0	0
G	7	79	6	1	1
H	8	63	9	-1	1
I	9	72	8	1	1
J	10	52	10	0	0
Σ					6,5

Tabela 6.6.

Iz Tabele 6.6. izračunavamo

$$r_S = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} = 1 - \frac{6 \cdot 6,5}{10 \cdot (100-1)} = 0,96,$$

što znači da postoji vrlo jaka koreacija između mesta na rang listi sa prijemnog ispita i broja poena na ispitu iz statistike.

7

Medicinski dodatak

7.1 Dijagnostički alati

Senzitivnost i specifičnost

Nekada nismo u mogućnosti da brzo i tačno utvrdimo da li neka osoba ima određenu bolest ili ne. Želeli bismo da imamo jednostavan test koji to može dosta dobro da proceni. Pretpostavimo da imamo grupu ispitanika za koju je tačno utvrđeno koji od njih boluju od određene bolesti, a koji ne. Na ove bolesnike primenjujemo test čiji rezultati mogu da budu pozitivni (ukazuju na prisustvo bolesti) ili negativni (ukazuju na odsustvo bolesti) i ne moraju da se poklapaju sa tačnom dijagnozom. To može da se prikaže u Tabeli 7.1.

Rezultat testa	bolest (da)	bolest (ne)	zbir
pozitivan	a	b	a+b
negativan	c	d	c+d
Zbir	a+ c	b+ d	a+b+c+d

Tabela 7.1.

Primetimo da $a + c$ osoba ima bolest, od kojih a osoba ima pozitivan test (**tačno pozitivni**), a c osoba ima negativan test (**lažno negativni**). Od $b + d$

osoba koje nemaju bolest, d osoba ima negativan test (**tačno negativni**), dok b osoba ima pozitivan test (**lažno pozitivni**).

Proporcija osoba kod kojih bolest postoji i koje su tačno identifikovane t.j. količnik

$$\frac{a}{a+c}$$

naziva se **senzitivnost**.

Proporcija osoba kod kojih bolest ne postoji i koje su tačno identifikovane t.j. količnik

$$\frac{d}{b+d}$$

naziva se **specifičnost**. Specifičnost i senzitivnost se obično izražavaju u procentima.

Poželjno je da specifičnost i senzitivnost budu što bliže broju 1, odnosno 100 %. Međutim, u praksi višu senzitivnost dobijamo na račun niže specifičnosti i obratno.

Ako su u pitanju bolesti koje se lako leče, onda prednost ima senzitivnost. Ako su u pitanju ozbiljne i neizlečive bolesti, onda se daje prednost visokoj specifičnosti, da bi se izbegla lažno pozitivna dijagnoza.

Proporcija osoba sa pozitivnim testom kod kojih je bolest prisutna t.j. količnik

$$\frac{a}{a+b}$$

naziva se **pozitivna prediktivna vrednost**.

Proporcija osoba sa negativnim testom kod kojih bolest nije prisutna t.j. količnik

$$\frac{d}{c+d}$$

naziva se **negativna prediktivna vrednost**.

Prediktivne vrednosti daju informaciju o tome kolika je verovatnoća da neka osoba ima bolest ako znamo rezultat njenog testa.

Primer 7.1. Grupi od 273 ispitanika urađen je tumorski marker a kasnije je biopsijom utvrđeno koji od njih imaju tumor, što je prikazano u Tabeli 7.2.

	postoji tumor	ne postoji tumor
pozitivan test	137	32
negativan test	14	90

Tabela 7.2.

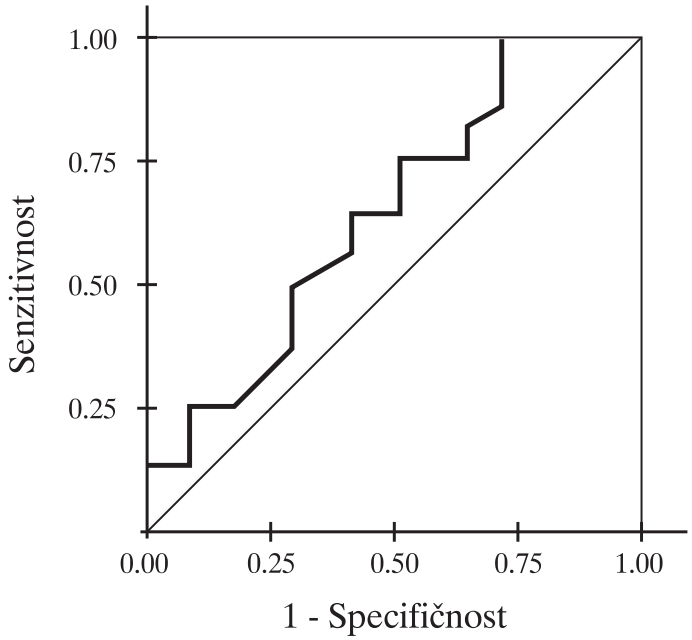
Iz Tabele 7.2. dobijamo

senzitivnost $= \frac{a}{a+c} = \frac{137}{137+14} = 0,91 = 91\,\%$

specifičnost $= \frac{d}{b+d} = \frac{90}{32+90} = 0,74 = 74\,\%$

pozitivna prediktivna vrednost $= \frac{a}{a+b} = \frac{137}{137+32} = 0,81 = 81\,\%$

negativna prediktivna vrednost $= \frac{d}{c+d} = \frac{98}{14+98} = 0,87 = 87\,\%$.



Slika 7.1.

Nekada treba da se da dijagnoza na osnovi vrednosti neprekidne slučajne promenjive (na primer, broj leukocita, vrednost hormona FT4, vrednost tumorskih markera). Često ne postoji precizan prag takav da vrednosti promenljive iznad (ispod) njega znače prisustvo bolesti. U takvim situacijama je potrebno

da sami odredimo taj prag (cut-off). Određivanjem raznih pragova dobijaju se različite senzitivnosti i specifičnosti. Ovim može da se dobije prag za koji smatramo da je najuverljiviji.

Uzimajući za prag sve vrednosti date promenljive (od najmanje do najveće) i izračunavanjem senzitivnosti i specifičnosti, može da se nacrta grafik takozvane ROC (receiver operating characteristic) krive. Na Slici 7.1. data je tipična ROC kriva. Neka promenljiva može utoliko bolje da posluži za testiranje (tačnije razlikuje zdrave od bolesnih) ukoliko je oblast ispod krive (area under ROC curve - AUROC) veća. Što je oblast bliža broju 0,5 promenljiva je manje pogodna kao marker, a što je bliža broju 1 promenljiva je pogodnija kao marker. U slučaju dve promenljive, bolja je ona promenljiva kod koje je oblast ispod krive veća. Na slici 7.1 se vidi da je površina ispod krive (AUROC) nešto veća od 0,5 (od površine ispod dijagonale). Međutim, da bi se videlo da li je ta razlika statistički značajna potebno je da se izvrši testiranje, o kome ovde neće biti reći.

Relativni rizik i količnik šanse (odds ratio)

Pretpostavimo da imamo dve grupe ispitanika, od kojih je prva grupa izložena dejstvu nekog faktora (pušenje, zračenje, rekreacija, uzimanje određenog leka i slično). Utvrđeno je koji od ispitanika iz ovih grupa boluje od određene bolesti a koji ne, što se može predstaviti Tabelom 7.3.

	izloženost (da)	izloženost (ne)
bolest (da)	a	b
bolest (ne)	c	d

Tabela 7.3.

U vezi sa Tabelom 7.3. definišu se dva pojma: relativni rizik i količnik šanse.

Relativni rizik (Relative risk) da se ima bolest predstavlja količnik verovatnoće da ispitanik iz grupe izloženih ima bolest i verovatnoće da ispitanik iz grupe neizloženih ima bolest. Prema Definiciji 1.2, verovatnoća da ispitanik iz grupe izloženih ima bolest je $\frac{a}{a+c}$ a verovatnoća da ispitanik iz grupe neizloženih

ima bolest je $\frac{b}{b+d}$. Otuda je

$$RR = \frac{\frac{a}{a+c}}{\frac{b}{b+d}}$$

Relativni rizik ukazuje na porast, odnosno smanjenje rizika od pojave bolesti u zavisnosti od izloženosti dejstvu faktora. Ako je relativni rizik jednak 1 to znači isti rizik da se ima bolest u grupi izloženih i u grupi neizloženih. Ako je relativni rizik veći od 1 onda to ukazuje na povećani rizik od pojave bolesti u grupi izloženih u odnosu na grupu neizloženih. Na primer, relativni rizik koji je jednak 3 ukazuje da osobe u grupi izloženih imaju tri puta veći rizik da obole od onih u grupi neizloženih, dok relativni rizik jednak $\frac{1}{2}$ ukazuje da osobe u grupi izloženih imaju dva puta manji rizik da obole od onih u grupi neizloženih.

Primer 7.2. U Velikoj Britaniji je rađena studija kojom je obuhvaćeno 7735 muškaraca starih između 40 i 59 godina. Od 7735 muškaraca njih 5899 je pušilo u nekom periodu života (pušači i bivši pušači). U periodu od 10 godina , 650 od ovih 7718 ljudi je doživelo infarkt miokarda. Rezultati su prikazani u Tabeli 7.4.

	Pušač	nepušač
Infarkt (da)	563	87
Infarkt (ne)	5336	1732

Tabela 7.4.

Ocenjeni relativni rizik je

$$RR = \frac{\frac{a}{a+c}}{\frac{b}{b+d}} = \frac{\frac{563}{563+5336}}{\frac{87}{87+1732}} = 2,00.$$

Ovo znači da ljudi srednjih godina koji su bilo kada bili pušači imaju, u proseku, dva puta veću verovatnoću da će u narednih 10 godina doživeti infarkt nego oni koji nikada nisu bili pušači.

Šansa je količnik verovatnoća dva suprotna događaja. Na primer, verovatnoća da se ima bolest i verovatnoća da se nema bolest.

Verovatnoća da se ima bolest u grupi izloženih uticaju faktora (Tabela 7.3.) je $\frac{a}{a+c}$ (prema Definiciji 1.2) dok je verovatnoća da se nema bolest u grupi

izloženih jednaka $\frac{c}{a+c}$. Količnik ove dve verovatnoće, odnosno šansa oboljevanja u grupi izloženih je

$$\frac{\frac{a}{a+c}}{\frac{c}{a+c}} = \frac{a}{c}$$

Slično, šansa oboljevanja u grupi neizloženih je $\frac{b}{d}$.

Količnik šanse (Odds ratio) da se ima bolest predstavlja količnik dve šanse: šanse da se ima bolest u grupi izloženih ($\frac{a}{c}$) i šanse da se ima bolest u grupi neizloženih ($\frac{b}{d}$). Otuda je

$$OR = \frac{\frac{a}{c}}{\frac{b}{d}} = \frac{a \cdot d}{b \cdot c}$$

Kada je bolest retka, odnosno kada su brojevi a i b mali, onda je količnik šanse približno jednak relativnom riziku i tako se i interpretira.

Ako je količnik šanse (odds ratio) jednak jedan, onda to ukazuje da su šanse da se ima bolest u izloženoj i neizloženoj grupi jednake. Količnik šanse veći od 1 ukazuje da je šansa da se ima bolest veća u izloženoj nego u neizloženoj grupi.

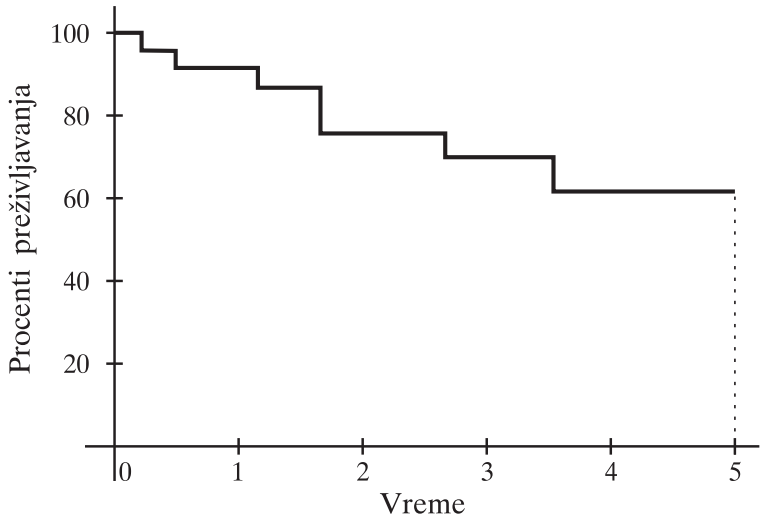
Ako je, na primer $OR = 2,3$, to znači da je u grupi izloženih šansa da se ima bolest 2,3 puta veća nego u grupi neizloženih. Ako je, na primer, $OR = \frac{1}{3}$, onda to znači da je šansa da se ima bolest u grupi izloženih tri puta manja nego u grupi neizloženih.

7.2 Metode preživljavanja

Podaci koji se tiču preživljavnja pre svega se odnose na vreme koje je potrebno da prođe do ishoda, koji nas interesuje (to je najčešće, ali ne uvek, smrt). Najvažniji podaci su da li je došlo do ishoda ili nije i kolika je dužina vremena koje prođe do ishoda. Na primer, interesuje nas preživljavanje kod pacijenata obolelih od ciroze.

Vreme preživljavanja se računa od neke polazne tačke (od operacije, od uspostavljanja dijagnoze) do ishoda. Često se ne zna kada dolazi do ishoda, jer se studija koju radimo prekida a do ishoda kod pojedinih pacijenata nije došlo. Tada se kaže da je vreme preživljavanja cenzurisano.

Kada se raspolaže podacima o dužini vremena i o tome da li je došlo do ishoda ili ne, onda može da se nacrta takozvana **Kaplan-Meier-ova kriva** kumulativnih verovatnoća. Na $x-$osi se prikazuje vreme a na $y-$osi kumulativne verovatnoće, date kao procenti. Na početku je kumulativna verovatnoća 100 %, jer su tada svi ispitanici prisutni. Kumulativna verovatnoća se menja kroz vreme (kriva ide nadole) kada dođe do ishoda. Na primer, ako u istraživanju učestvuje 50 ispitanika i posle $1,5$ godine dođe do ishoda kod dva ispitanika (ova dva ispitanika čine 4% od ukupnog broja ispitanika), onda kriva ide nadole za 4%. Na Slici 7.2. prikazana je jedna Kaplan-Meier-ova kriva, koja se odnosi na jednu grupu ispitanika praćenih 5 godina. Ako su na grafiku prikazane dve Kaplan-Meier-ove krive i ako se prva kriva nalazi ispod druge, onda to ukazuje da je preživljavanje u grupi koja odgovara drugoj krivoj veće. Međutim, ipak treba i formalno testirati hipotezu da li su razlike u preživljavanju između dve ili više grupa signifikantne. **Log rank test** testira nultu hipotezu da nema razlike u preživljavanju između grupa koje se posmatraju. Ako se dobije da je $p < 0,05$, onda to znači da je razlika u preživljavanju između grupa statistički značajna.



Slika 7.2.

Da bi se ispitao uticaj pojedinih promenljivih na preživljavanje, koristi se **Cox-ova regresija**. Za svaku promenljivu dobija se određena verovatnoća.

Ako je verovatnoća manja od 0,05 onda ta promenljiva značajno utiče na preživljavanje. U suprotnom se smatra da uticaj te promenljive nije značajan. Slično binarnoj regresiji i ovde se, kao rezultat, pojavljuje $exp(b_i)$, koji se naziva količnik rizika (hazard ratio) i koji se tumači slično kao odds ratio kod binarne logističke regresije. Naime, količnik rizika predstavlja procenu za koliko se povećava (smanjuje) rizik od lošeg ishoda kada se vrednost određene promenljive poveća za 1, pod uslovom da se ostale promenljive ne menjaju. Ako je, recimo, taj broj 3, onda to znači da se rizik povećava tri puta. Ako je, na primer, 0,2 onda znači da se rizik smanjio 5 puta.