Sveučilište u Zagrebu
PMF-Matematički odjel
Poslijediplomski specijalistički sveučilišni studij
aktuarske matematike

Miljenko Huzak

Vjerojatnost i matematička statistika

Predavanja

Travanj 2006.

Sadržaj

Poglavlje 1

Opisna analiza podataka

U ovom poglavlju bavit ćemo se *opisnom* ili *deskriptivnom statistikom*: metodama prikaza skupova podataka pomoću tablica, grafikona i numeričkih pokazatelja.

1.1 Vrste podataka

U statistici pod skupom podataka razumijevamo vrijednosti dobivene mjerenjem (ili opažanjem) nekog *statističkog obilježja* (ili *varijable*) promatrane (izučavane) skupine objekata ili osoba. Varijabla može biti jednodimenzionalna ili višedimenzionalna.

Primjer 1.1 Svi osiguranici od autoodgovornosti nekog osiguravajućeg društva predstavljaju skupinu koju promatramo. Statističko obilježje koje nas zanima je

$$X = \text{broj šteta po polici u proteklih godinu dana.}$$

Druga varijabla koja nas zanima je

$$Y = \text{ukupan iznos šteta (u kunama) po polici u proteklih godinu dana.}$$

Vektor $\mathbf{Z} = (X, Y)$ čije komponente su varijable X i Y je dvodimenzionalno statističko obilježje. □

Grupa objekata ili osoba koju promatramo, odnosno za koju izučavamo odabrano statističko obilježje, zove se *populacija*. Često nije moguće popisati (izmjeriti, opaziti) sve vrijednosti izučavanoga statističkog obilježja. U tom slučaju odabiremo (reprezentativni) *uzorak* iz populacije i iz njega popisujemo vrijednosti statističkog obilježja.

Primjer 1.2 Pomoću računala na slučajan način odabran je uzorak od 100 osiguranika (nekog osiguravajućeg društva) s policom mješovitog osiguranja života. Računalni program je u datoteku pohranio podatke o njihovim osiguranim svotama. □

Dakle, razlikujemo skupove podataka dobivene mjerenjem (opažanjem) odabranog statističkog obilježja na populaciji (*populacijski podaci*), od onih dobivenih na uzorku iz populacije (*uzorački podaci*).

Cilj statističke analize skupova populacijskih podataka je izdvojiti važne značajke tih podataka, na primjer, distribuciju izučavanog statističkog obilježja, njegovu srednju vrijednost i sl. Pri tome se koriste metode deskriptivne statistike. U slučaju skupova uzoračkih podataka, cilj statističke analize je da se na osnovi podataka iz uzorka donesu određeni zaključci o populacijskoj razdiobi promatranog statističkog obilježja. U tu svrhu, uz metode

deskriptivne statistike, koriste se i metode *inferencijalne statistike* o kojima će biti riječi u sljedećim poglavljima.

Osim po porijeklu (izvoru, obuhvatu), podatke možemo podijeliti po tipu vrijednosti opažanog statističkog obilježja, odn. varijable. Razlikujemo *numeričke* i *kategorijalne* varijable. Varijable iz primjera 1.1 i 1.2 su numeričke, dok su, na primjer, spol, mjesto rođenja ili kategorija vozača kategorijalne varijable. Vrijednosti kategorijalnih varijabli zovemo *razredima*.

Numeričke se varijable dijele na:

- *diskretne* (obično predstavljaju neko prebrojavanje). Na primjer, broj šteta po polici osiguranja (X iz primjera 1.1), broj ovlaštenih aktuara u HAD-u.
- *neprekidne* (obično predstavljaju rezultat mjerenja neke fizikalne ili novčane veličine). Na primjer, visina, težina, iznos šteta po polici osiguranja (Y iz primjera 1.1).

Kategorijalne se varijable dijele na:

- *dihotomne* (imaju samo dva razreda). Na primjer, spol, odgovori sa da/ne.
- *nominalne* (razredi su neuređeni). Na primjer, tip police, priroda (uzrok) štete.
- *ordinalne* (vrijednosti su im uređene). Na primjer, rangiranje hrane, školske ocjene.

1.2 Frekvencijske distribucije

Skupovi diskretnih numeričkih i kategorijalnih podataka opisuju se svojim frekvencijskim distribucijama. Frekvencijske distribucije prikazuju se tabelarno pomoću *frekvencijskih tablica* ili grafički pomoću *stupčastih* ili *strukturnih dijagrama*.

Frekvencija ili *učestalost* vrijednosti varijable (odnosno njenog razreda) je broj pojavljivanja te vrijednosti u skupu podataka, a njena *relativna frekvencija* je omjer frekvencije i ukupnog broja podataka.

Primjer 1.3 Navedena frekvencijska tablica predstavlja frekvencijsku distribuciju skupa podataka dobivenih mjerenjem varijable X, koja predstavlja broj djece u obitelji mlađe od 16 godina, na uzorku od 80 obitelji.

broj djece	frekvencija	relativna frekvencija
0	8	0.1
1	12	0.15
2	28	0.35
3	19	0.2375
4	7	0.0875
5	4	0.05
6	1	0.0125
7	1	0.0125
8 ili više	0	0
Σ	80	1.

Na primjer, frekvencija vrijednosti"1" varijable X je 12, a njena relativna frekvencija je $12/80 = 0.15$. Ista frekvencijska distribucija grafički je prikazana na slici 1.1 kao stupčasti

dijagram frekvencija (tj. visine stupaca predstavljaju iznose frekvencija), na slici 1.2 pomoću stupčastog dijagrama relativnih frekvencija, a na slici 1.3 pomoću strukturnog dijagrama. □

Stupčasti dijagrami relativnih frekvencija koristi se za grafičku usporedbu frekvencijskih distribucija više skupova podataka istoga tipa, na primjer, dobivenih mjerenjem istog statističkog obilježja na raznim uzorcima. Strukturni dijagrami se koriste za prikaz frekvencijskih distribucija varijabli s (relativno) malo razreda. Za prikaz distribucija nominalnih varijabli s (relativno) mnogo razreda najčešće se koriste položeni stupčasti dijagrami s razredima sortiranima po veličini frekvencije.

1.3 Histogrami i frekvencijske distribucije grupiranih vrijednosti

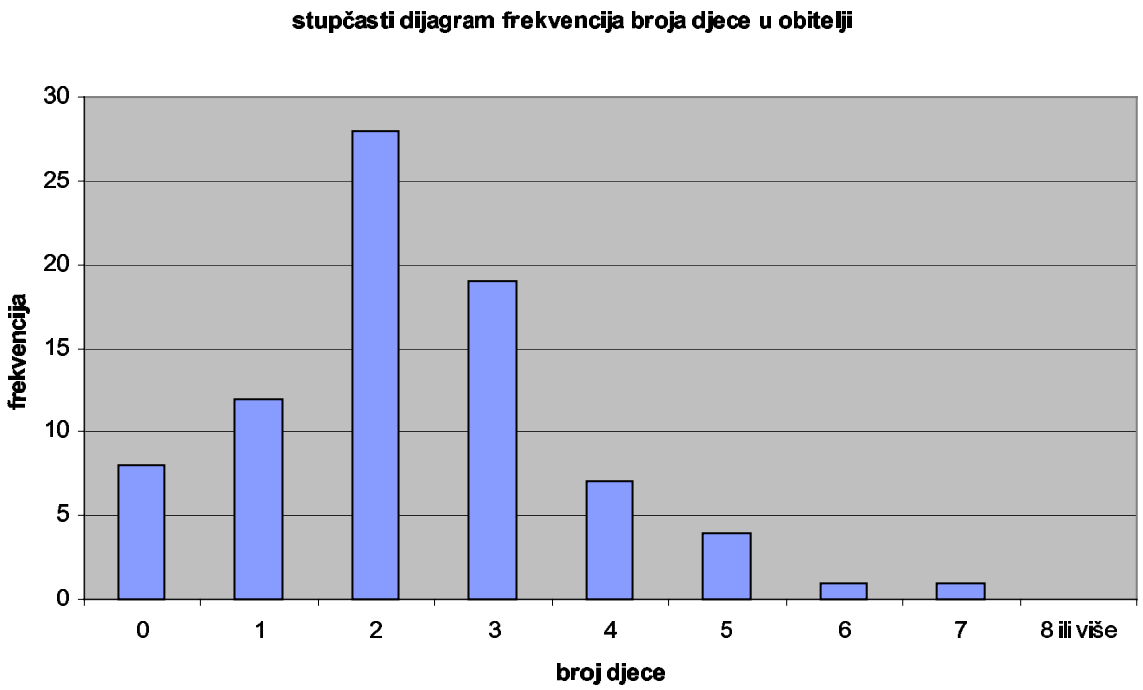
Za razliku od diskretnih numeričkih i kategorijalnih varijabli, vrijednosti se neprekidnih varijabli (u pravilu) ne ponavljaju, pa se skupovi takvih podataka ne mogu prikazivati pomoću frekvencijske distribucije na način opisan u 1.2. Za njihov prikaz koristimo frekvencijsku distribuciju *grupiranih vrijednosti*. Preciznije, vrijednosti varijable grupiramo u konačno mnogo razreda, a zatim odredimo frekvencije (i/ili relativne frekvencije) tih razreda. Razredi su predstavljeni sa međusobno disjunktnim intervalima kojima su obuhvaćane sve vrijednosti varijable (tj. razredi čine konačnu particiju područja vrijednosti varijable).

Frekvencijska distribucija grupiranih vrijednosti varijable grafički se prikazuje *histogramom*. Histogram je sličan stupčastom dijagramu, ali, za razliku od stupčastog dijagrama, prikazuje se u Kartezijevom koordinatnom sustavu. Sastoji se od onoliko pravokutnika koliko ima razreda, s osnovicama nad intervalima koji reprezentiraju razrede na osi apscisa. Površina svakog takvog pravokutnika jednaka je relativnoj frekvenciji razreda kojeg predstavlja. Dakle, ukupan zbroj površina pravokutnika histograma je jednak jedan.

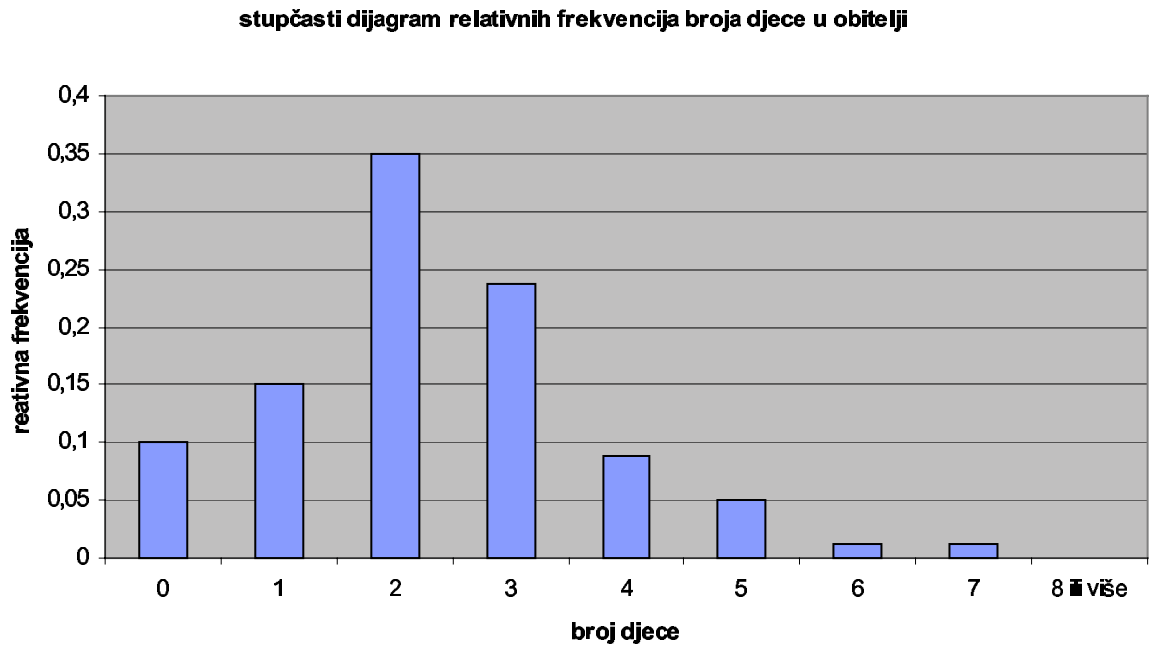
Primjer 1.4 Raspolažemo sa 100 podataka o iznosima šteta zbog popuštanja vodovodnih instalacija po policama osiguranja kućanstava.

243	306	271	396	287	399	466	269	295	330
425	324	228	113	226	176	320	230	404	487
127	74	523	164	366	343	330	436	141	388
293	464	200	392	265	403	372	259	426	262
221	355	324	374	347	261	278	113	135	291
176	342	443	239	302	483	231	292	373	346
293	236	223	371	287	400	314	464	337	308
359	352	273	267	277	184	286	214	351	270
330	238	248	419	330	319	440	427	343	414
291	299	265	318	415	372	238	323	411	494

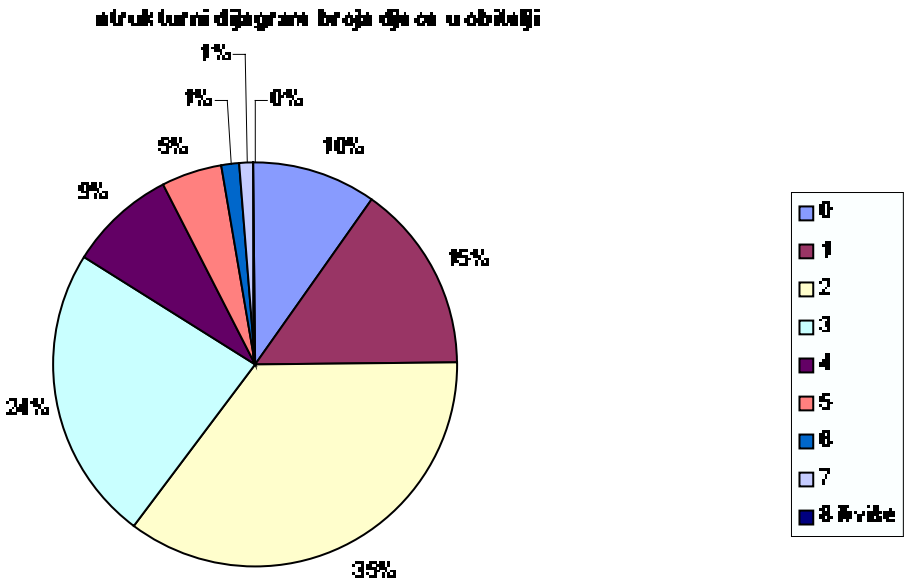
Minimalna vrijednost opažane varijable je 74, a maksimalna 523. U nedostatku dodatnih informacija o području mogućih vrijednosti te varijable, pretpostavit ćemo da se one kreću u rasponu od 50 do 550 novčanih jedinica. To područje particioniramo u 10 razreda kako je prikazano u frekvencijskoj tablici grupiranih vrijednosti.



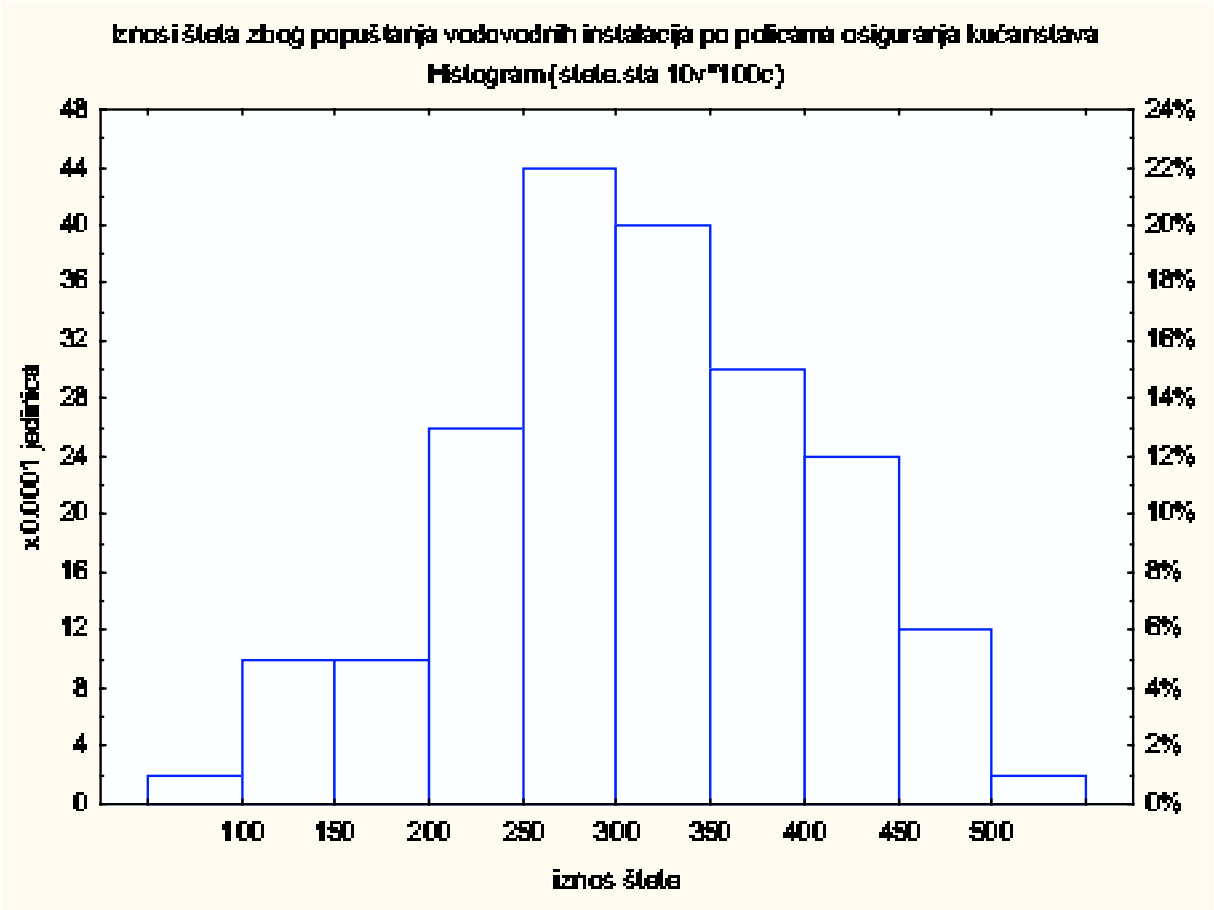
Slika 1.1



Slika 1.2



Slika I.3



Slika I.4

razred	frekvencija	relativna frekvencija	visina pravokutnika
$[50, 100\rangle$	1	0.01	0.0002
$[100, 150\rangle$	5	0.05	0.0010
$[150, 200\rangle$	4	0.04	0.0008
$[200, 250\rangle$	14	0.14	0.0028
$[250, 300\rangle$	22	0.22	0.0044
$[300, 350\rangle$	20	0.20	0.0040
$[350, 400\rangle$	14	0.14	0.0028
$[400, 450\rangle$	13	0.13	0.0026
$[450, 500\rangle$	6	0.06	0.0012
$[500, 550\rangle$	1	0.01	0.0002
Σ	100	1.	—

Histogram tog skupa podataka nalazi se na slici 1.4. Budući da je širina svakog intervala (razreda) jednaka 50, a ujedno je to i duljina osnovica pripadnih pravokutnika, primijetite da visine pravokutnika nisu jednake relativnim frekvencijama. □

1.4 *Stem and leaf* dijagram

Stem and leaf dijagram je, u stvari, histogram prikazan pomoću nizova brojeva. Formira se na sljedeći način. Na početku svakog retka, odijeljen vertikalnom crtom zdesna, nalazi se broj koji reprezentira razred, tzv. *stabljika* (engl. *stem*). Desno od vertikalne crte slijede ga u nizu druge po značaju znamenke brojeva koji pripadaju tom razredu, tzv. *lišće*. Dakle, svaka znamenka desno od crte je list (engl. *leaf*). Dijagram se sastoji od onoliko redaka koliko ima stabljika (razreda).

Primjer 1.5 Naveden je *stem and leaf* dijagram za skup podataka iz primjera 1.4. Stabljike predstavljaju znamenke stotice, a lišće znamenke desetice svakog od brojeva iz uzorka.

0	7
1	112346778
2	012222333333445666666777778889999999
3	000111222233333444445555677777899 9
4	0001111222344666889
5	2

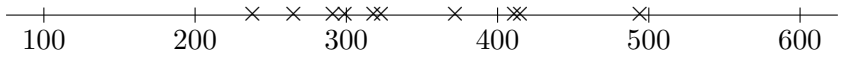
□

1.5 Linijski dijagram i dijagram točaka

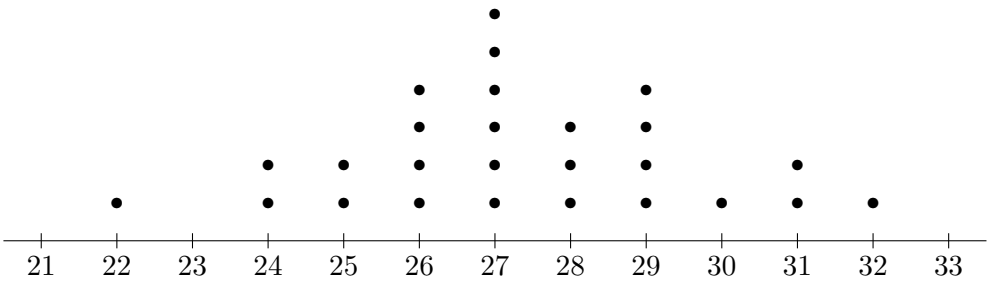
Za prikaz malog skupa numeričkih podataka koriste se linijski i dijagram točaka.

Linijski dijagram sastoji se od brojevnog pravca na kojemu su, na primjer križićem, naznačene vrijednosti iz skupa podataka. U slučaju da se neki podaci više puta ponavljaju, koristi se *dijagram točaka*. Taj dijagram se također sastoji od brojevnog pravca. Podaci se reprezentiraju sa po jednom točkom koja se ucrtava iznad njihove vrijednosti na brojevnom pravcu. Svaka ponovljena vrijednost naznačava se novom točkom koja se ucrtava nad prethodnom točkom. Dakle, dijagram točaka ima oblik histograma.

Primjer 1.6 Linijski dijagram skupa podataka koji se sastoji od zadnjih 10 brojeva iz primjera 1.4 (zadnji redak):



Primjer 1.7 Navedeni dijagram točaka predstavlja uzorak dobiven nezavisnim mjerenjem vremena izvođenja određene radne operacije (u sekundama).



1.6 Mjere lokacije

Postoji više različitih mjera centralnih tendencija skupova podataka. Navest ćemo tri najvažnije: *aritmetičku sredinu*, *medijan* i *mod*.

Neka su

$$x_1, x_2, \ldots, x_n \tag{1.1}$$

n vrijednosti varijable X koje čine skup podataka. Ako je X numerička varijabla, tada je (1.1) niz brojeva.

1.6.1 Aritmetička sredina

Neka je X numerička varijabla. *Aritmetička sredina* brojeva (1.1) je broj

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{1.2}$$

Ako X poprima samo par različitih vrijednosti

$$a_1, a_2, \ldots, a_k, \tag{1.3}$$

koje se u nizu (1.1) pojavljuju više puta (vrijednost a_1 s frekvencijom f_1, vrijednost a_2 s frekvencijom f_2 itd.), tada se formula (1.2) za aritmetičku sredinu može zapisati u sažetom obliku:

$$\bar{x} = \frac{1}{n}(f_1 a_1 + f_2 a_2 + \cdots + f_k a_k) = \frac{1}{n}\sum_{j=1}^{k} f_j a_j. \tag{1.4}$$

Primijetite da je $n = f_1 + f_2 + \cdots + f_k$.

Primjer 1.8 Aritmetička sredina podataka iz primjera 1.3 je:

$$\bar{x} = \frac{8 \cdot 0 + 12 \cdot 1 + 28 \cdot 2 + 19 \cdot 3 + 7 \cdot 4 + 4 \cdot 5 + 1 \cdot 6 + 1 \cdot 7}{8 + 12 + 28 + 19 + 7 + 4 + 1 + 1} = \frac{186}{80} = 2.325.$$

□

1.6.2 Medijan

Neka je X numerička ili ordinalna varijabla. Tada je njene vrijednosti (1.1) moguće urediti:

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}. \tag{1.5}$$

Medijan skupa podataka (1.1) je vrijednost od X za koju vrijedi da je 50% svih podataka u skupu manje od ili jednako toj vrijednosti i 50% svih podataka je veće od nje ili jednako joj. Kada je broj podataka n u (1.1) (odn. (1.5)) neparan broj, $n = 2k - 1$, medijan m od (1.1) je jednak $x_{(k)}$. U stvari, to je vrijednost koja se nalazi u sredini niza (1.5). Dakle, medijan se može odrediti za neparne skupove ordinalnih podataka. Uz takvu općenitost teško je odrediti medijan parnog skupa podataka. Zato pretpostavimo da su (1.1) numerički podaci. Tada je medijan skupa s parnim brojem podataka $n = 2k$ jednak $m = (x_{(k)} + x_{(k+1)})/2$, tj. aritmetičkoj sredini dva srednja broja u (1.5).

Primjer 1.9 Medijan uzorka iz primjera 1.3 je

$$m = \frac{x_{(40)} + x_{(41)}}{2} = \frac{2 + 2}{2} = 2.$$

□

1.6.3 Mod

Mod je vrijednost obilježja X koja su u skupu podataka (1.1) pojavljuje najviše puta, dakle, ima najveću frekvenciju. Mod se može opisati i kao *najtipičnija* vrijednost promatrane varijable. Na primjer, osiguravajuće društvo može zanimati najtipičnija vrsta osiguranika po zanimanju. Jasno je da mod općenito ne mora postojati.

Primjer 1.10 Mod uzorka iz primjera 1.3 je 2 jer ta vrijednost ima najveću frekvenciju (28). Dakle, najtipičnija obitelj (u uzorku) ima dvoje djece mlađe od 16 godina. □

1.7 Mjere raspršenja

Uz mjere lokacije, odnosno srednje vrijednosti skupa podataka, važno svojstvo distribucije tih podataka je i kako su podaci raspršeni, često u odnosu na neku srednju vrijednost.

1.7.1 Standardna devijacija

Najčešće korištena mjera raspršenja skupa numeričkih podataka je standardna devijacija. *Standardna devijacija* je srednje kvadratno odstupanje podataka od njihove aritmetičke sredine. Formulom,

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}, \tag{1.6}$$

odnosno

$$s = \sqrt{\frac{1}{n-1}\sum_{j=1}^{k} f_j(a_j - \bar{x})^2}, \tag{1.7}$$

u slučaju da se k različitih vrijednosti (1.3) od X u (1.1) ponavljaju. Broj

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, \text{ odnosno } s^2 = \frac{1}{n-1}\sum_{j=1}^{k} f_j(a_j - \bar{x})^2 \tag{1.8}$$

zovemo *varijanca* skupa podataka (1.1). Formule (1.8) se mogu pojednostaviti:

$$s^2 = \frac{1}{n-1}(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2), \text{ odnosno } s^2 = \frac{1}{n-1}(\sum_{j=1}^{k} f_j a_j^2 - n\bar{x}^2). \tag{1.9}$$

Primjer 1.11 Za uzorak iz primjera 1.3, uzoračka varijanca je:

$$s^2 = \frac{1}{79}(592 - 80 \cdot (\frac{186}{80})^2) = 2.02.$$

□

1.7.2 Momenti

Aritmetička sredina i varijanca su specijalni slučajevi mjera koje zovemo momentima skupa (numeričkih) podataka. *k-ti moment* skupa podataka (1.1) *oko vrijednosti* α je broj

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \alpha)^k.$$

Momente oko ishodišta ($\alpha = 0$) zovemo jednostavno momentima. Centralnim momentima zovemo momente oko aritmetičke sredine ($\alpha = \bar{x}$). Prema tome, aritmetička sredina je prvi moment, a varijanca drugi centralni moment normiran sa $n-1$ umjesto sa n.

1.7.3 Raspon

Raspon R skupa podataka (1.1) definira se kao razlika maksimalne i minimalne vrijednosti u uzorku:

$$R = \max_{1\le i\le n} x_i - \min_{1\le i\le n} x_i = x_{(n)} - x_{(1)}.$$

Primjer 1.12 Raspon uzorka iz primjera 1.3 je $R = 7 - 0 = 7$. □

1.7.4 Interkvartil

Interkvartil je mjera raspršenja slična rasponu skupa podataka, ali je, za razliku od raspona, robustna na ekstremne vrijednosti.

Neka su numerički podaci (1.1) uređeni kao u (1.5). Interkvartil definiramo u više koraka. Prvo, definiramo interpolacijske vrijednosti u (1.5). Za realan pozitivan broj $r = k + \alpha$, gdje je k prirodan broj takav da je $k < n$ i $0 \le \alpha < 1$, definiramo interpolacijsku vrijednost na mjestu r, $x_{(r)}$, (koji još zovemo r-tim kvantilom) formulom:

$$x_{(r)} = x_{(k+\alpha)} := x_{(k)} + \alpha(x_{(k+1)} - x_{(k)}).$$

U sljedećem koraku definiramo *donji kvartil* q_L i *gornji kvartil* q_U skupa podataka formulama

$$q_L := x_{(\frac{n+1}{4})}, \quad q_U := x_{(\frac{3(n+1)}{4})}.$$

Opisno, za donji kvartil vrijedi da je 25% podataka u skupu manje od ili jednako njemu, a 75% podataka je veće od ili jednako njemu. Slično se objašnjava i gornji kvartil (u prethodnoj rečenici treba zamijeniti 25% sa 75% i obratno). Konačno, *interkvartil* IQR skupa podataka (1.1) je razlika gornjeg i donjeg kvartila:

$$IQR = q_U - q_L.$$

Primjer 1.13 Za uzorak iz primjera 1.3 računamo

$$\begin{aligned} q_L &= x_{(\frac{81}{4})} = x_{(20+\frac{1}{4})} = x_{(20)} + \frac{1}{4}(x_{(21)} - x_{(20)}) = 1 + \frac{1}{4}(2-1) = \frac{5}{4} = 1.25 \\ q_U &= x_{(\frac{243}{4})} = x_{(60+\frac{3}{4})} = x_{(60)} + \frac{3}{4}(x_{(61)} - x_{(60)}) = 3 + \frac{3}{4}(3-3) = 3. \end{aligned}$$

Dakle, interkvartil tog uzorka je $IQR = 3 - 1.25 = 1.75$. □

1.8 Mjere asimetričnosti

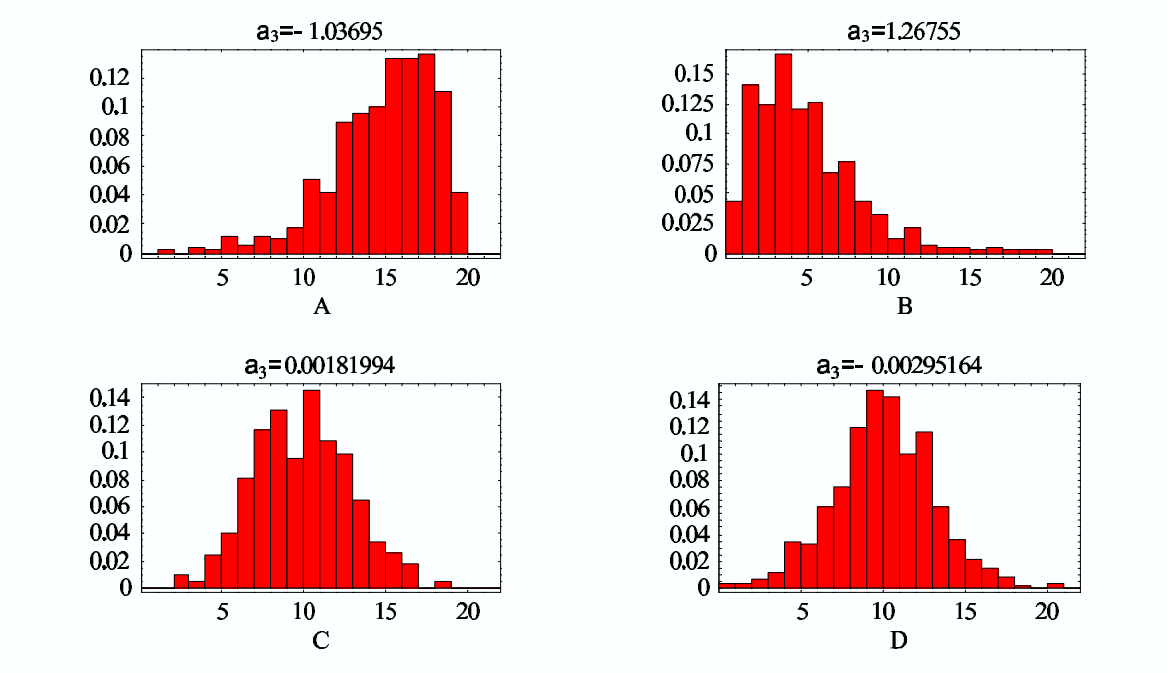
Sljedeća važna značajka distribucije skupa podataka je njezin oblik. Jedna komponenta oblika je simetričnost. Pretpostavimo da su podaci (1.1) numerički. Simetričnost skupa podataka opisuje se trećim centralnim momentom. Kažemo da je skup podataka *simetričan* ako je njihov treći centralni moment jednak nuli, drugim riječima, ako je taj skup brojeva simetričan u odnosu na svoju aritmetičku sredinu. Odstupanje od simetričnosti mjeri se koeficijentom asimetrije. *Koeficijent asimetrije* α_3 je treći moment skupa *standardiziranih* podataka normiran sa $n-1$ umjesto n:

$$\alpha_3 := \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^3.$$

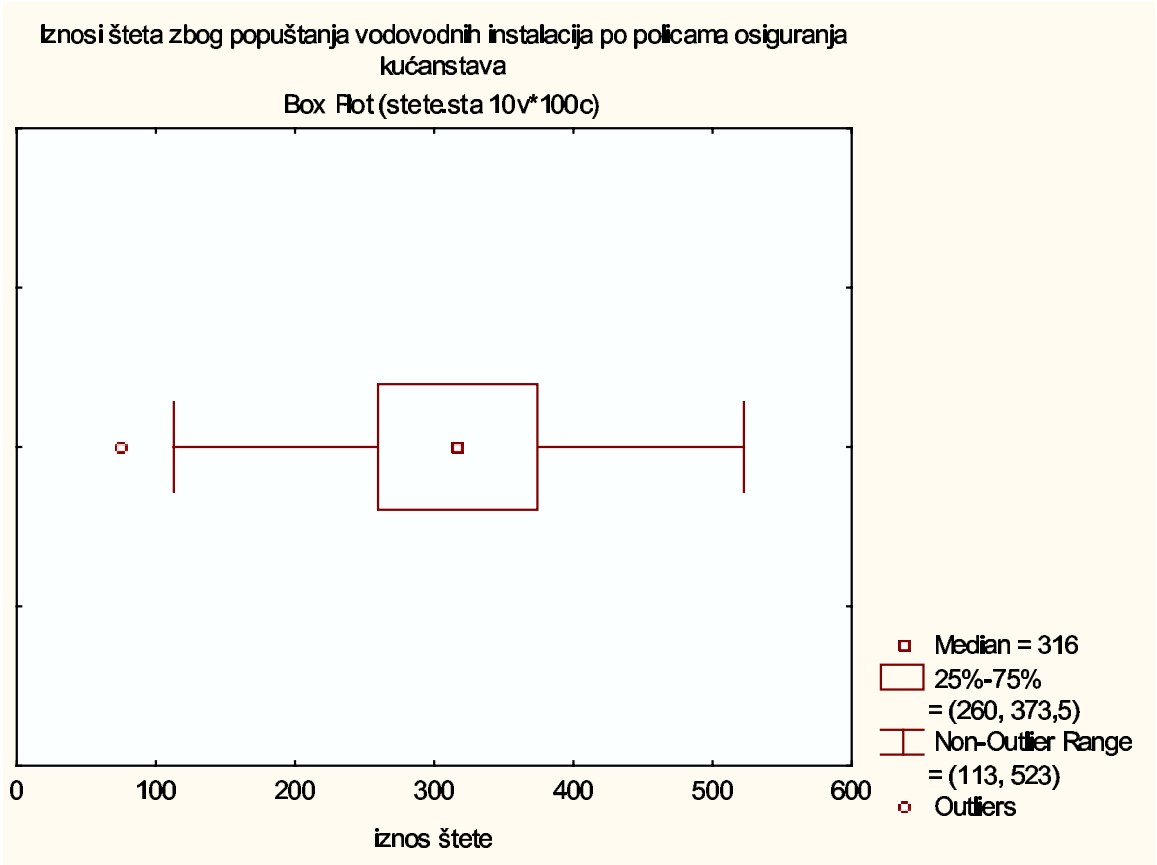
Kažemo da je skup podataka *negativno asimetričan* ako je $\alpha_3 < 0$, a *pozitivno asimetričan* ako je $\alpha_3 > 0$. $\alpha_3 = 0$ znači da je skup podataka simetričan. Grafički, histogram simetričnog skupa podataka je simetričan u odnosu na vertikalni pravac koji prolazi aritmetičkom sredinom. Na primjer, histogrami B i C prikazani na slici 1.5 prikazuju pozitivno asimetrične, a histogrami A i D na istoj slici, negativno asimetrične skupove podataka. Budući da su koeficijenti asimetrije za skupove podataka čiji su histogrami C i D, vrlo mali, gotovo jednaki nuli, uzimamo da su ti skupovi podataka (gotovo) simetrični.

1.9 Dijagram pravokutnika

Dijagram pravokutnika (engl. *box and whisker*) koristi se za grafički prikaz distribucije velikog i malog skupa numeričkih podataka. Iz njega se direktno može očitati medijan, donji i gornji kvartil, interkvartil, raspon, ekstremne vrijednosti i simetrija. Na slici 1.6 nalazi se dijagram pravokutnika za podatke iz primjera 1.4.



Slika 1.5



Slika 1.6

Poglavlje 2

Slučajne varijable

2.1 Vjerojatnosni prostor. Uvjetna vjerojatnost. Nezavisnost događaja.

Vjerojatnosni prostor predstavlja matematički model za slučajni pokus. *Slučajni pokus* je pokus koji ima više mogućih ishoda. Ishode slučajnog pokusa zovemo *događajima*. Intuitivno, slučajni pokus je svaki pokus (proces, opažanje, mjerenje) kojemu se ishod ne može sa sigurnošću predvidjeti.

Primjer 2.1 Bacanje igraće kocke je slučajni pokus jer ima više mogućih ishoda, na primjer, može se dogoditi da se je okrenuo paran broj ili da se okrenula jedinica itd. Specijalno, 6 događaja: okrenuo se broj 1, okrenuo se broj 2, itd., okrenuo se broj 6, zovemo *elementarnim događajima* jer se, svaki za sebe, ne mogu razložiti na još jednostavnije događaje za razliku od, na primjer, događaja "okrenuo se paran broj" koji se može razložiti na događaje 2, 4 i 6. Za ta tri elementarna događaja kažemo da su povoljni za događaj "okrenuo se paran broj". □

Označimo sa Ω skup elementarnih događaja ω_1, ω_2 itd. Ω zovemo *prostor elementarnih događaja*. Budući da se svaki događaj sastoji od elementarnih događaja koji su povoljni za njega, događaji su podskupovi od Ω. Specijalno, Ω i prazan skup $\emptyset$ su događaji. Ω zovemo *sigurnim*, a $\emptyset$ *nemogućim* događajem.

Neka su A i B događaji. Tada je jasno da su i

$$A \cap B,\ A \cup B,\ A \setminus B,\ A^c = \Omega \setminus A$$

događaji. Još više, prebrojive unije i prebrojivi presjeci događaja su događaji. Ako sa $\mathcal{F}$ označimo familiju događaja, tada $\mathcal{F}$ mora biti zatvorena na komplementiranje i prebrojive unije i presjeke, te mora sadržavati Ω i $\emptyset$. Očito je da je skup svih podskupova od Ω, u oznaci $\mathcal{P}(\Omega)$, jedna takva familija. Postoje i manje familije s takvim svojstvima.

Vjerojatnost je normirana mjera na događajima. Preciznije, neka je $\mathcal{F}$ familija događaja na prostoru elementarnih događaja Ω. Preslikavanje $\mathbb{P}$ koje svakom događaju A iz $\mathcal{F}$ pridružuje realan broj $\mathbb{P}(A)$ tako da vrijedi

$(P1)$ $0 \leq \mathbb{P}(A) \leq 1$ za sve događaje $A \in \mathcal{F}$,

$(P2)$ $\mathbb{P}(\Omega) = 1$,

$(P3)$ Za međusobno disjunktne događaje $A_1, A_2, \ldots$ iz $\mathcal{F}$ vrijedi:

$$\mathbb{P}(A_1 \cup A_2 \cup \ldots) = \mathbb{P}(A_1) + \mathbb{P}(A_2) + \cdots,$$

zovemo vjerojatnost na $(\Omega, \mathcal{F})$. Broj $\mathbb{P}(A)$ zovemo vjerojatnost događaja A. U slučaju kada je Ω prebrojiv skup, $\mathcal{F} = \mathcal{P}(\Omega)$. Uređenu trojku $(\Omega, \mathcal{F}, \mathbb{P})$ zovemo *vjerojatnosni prostor*.

Neka je $\Omega = \{\omega_1, \omega_2, \ldots\}$ diskretan skup. Ako su zadane vjerojatnosti p_1, p_2,... elementarnih događaja ω_1, ω_2,..., tada je sa

$$\mathbb{P}(A) := \sum_{\omega_i \in A} p_i \tag{2.1}$$

zadana vjerojatnost na $(\Omega, \mathcal{P}(\Omega))$. Vjerojatnosni prostor $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ zovemo *diskretnim vjerojatnosnim prostorom*.

Primjer 2.1 (*nastavak*) Ako je igraća kocka simetrična, tada su vjerojatnosti p_1,..., p_6 elementarnih događaja 1,..., 6 sve jednake 1/6. U tom slučaju za vjerojatnost bilo kojeg događaja A po formuli (2.1) vrijedi

$$\mathbb{P}(A) = \sum_{\omega_i \in A} \frac{1}{6} = \frac{|A|}{6},$$

gdje je sa $|A|$ označen broj elemenata skupa A. Na taj način dobiveni vjerojatnosni prostor $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ je matematički model za bacanje simetrične (fer) igraće kocke. Ako kocka nije simetrična, na primjer, ako vrijedi da je $p_1 = 1/12$, $p_2 = p_3 = p_4 = p_5 = 1/6$ i $p_6 = 1/4$, tada se i dalje vjerojatnost svakog događaja A definira po formuli (2.1), ali $\mathbb{P}(A) \neq |A|/6$. Dakle, dobili smo novi model za bacanje igraće kocke različit od prethodnog. □

Brojevi p_1, p_2,... kojima zadajemo vjerojatnosti elementarnih događaja nisu sasvim proizvoljni. Naime, iz svojstva vjerojatnosti $(P1)$ nužno slijedi da je za sve i, $0 \leq p_i \leq 1$, a iz $(P2)$ i $(P3)$ da mora vrijediti $p_1 + p_2 + \cdots = 1$.

Neka je sada $(\Omega, \mathcal{F}, \mathbb{P})$ bilo koji vjerojatnosni prostor, te neka su A, B dva događaja. Pretpostavimo da znamo da se dogodio B ($\mathbb{P}(B) > 0$), pa nas zanima kolika je vjerojatnost da se dogodio A. Dakle, zanima nas kolika je *uvjetna* vjerojatnost od A uz uvjet da se dogodio B. Jasno je da ta vjerojatnost ne mora biti jednaka $\mathbb{P}(A)$ jer su njome (možda) obuhvaćeni i oni elementarni događaji za koje već znamo da se nisu dogodili, dakle, oni koji nisu povoljni za događaj B. Isto tako, ta vjerojatnost ne mora biti jednaka $\mathbb{P}(A \cap B)$ jer je, zbog informacije da se B dogodio, B postao novi prostor elementarnih događaja, pa bi nova vjerojatnost na B morala biti normirana (svojstvo $(P2)$). Stoga se *uvjetna vjerojatnost događaja A uz uvjet da se dogodio B* definira kao broj

$$\mathbb{P}(A|B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}. \tag{2.2}$$

Kažemo da su događaji A i B *nezavisni* ako je

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B).$$

Primijetite da je ta relacija ekvivalentna sa svakom od sljedećih, dolje navedenih relacija (ako je $0 < \mathbb{P}(A) < 1$ i $0 < \mathbb{P}(B) < 1$):

$$\mathbb{P}(A|B) = \mathbb{P}(A),\ \mathbb{P}(A|B^c) = \mathbb{P}(A),\ \mathbb{P}(B|A) = \mathbb{P}(B),\ \mathbb{P}(B|A^c) = \mathbb{P}(B).$$

Odavde odmah slijedi da su i događaji A i B^c, A^c i B, te A^c i B^c nezavisni. Dakle, događaji su nezavisni ako događanje ili nedogađanje jednog ne utječe na vjerojatnost drugog događaja i obratno.

2.2 Diskretne slučajne varijable

Slučajna varijabla je funkcija koja svakom elementarnom događaju pridružuje broj. Označavamo ih velikim slovima abecede: X, Y, Z,... Nadalje, sa ImX označavamo sliku slučajne varijable X. Dakle, ImX je skup brojeva, vrijednosti koje X poprima.

Slučajna varijabla je *diskretna* ako je ImX prebrojiv skup. Pri tome mora vrijediti da su skupovi

$$\{X = x\} = (\text{oznaka}) = \{\omega \in \Omega : X(\omega) = x\}$$

događaji za svaki $x \in \text{Im}X$. Primijetimo da je za $x \notin \text{Im}X$, $\{X = x\} = \emptyset$ što je također događaj. Diskretnoj slučajnoj varijabli X pridružujemo funkciju $f_X : \mathbb{R} \to \mathbb{R}$, definiranu sa

$$f_X(x) := \mathbb{P}(X = x),$$

koju zovemo *funkcijom vjerojatnosti* od X ili *funkcijom gustoće* razdiobe od X ili, jednostavno, *gustoćom* razdiobe od X. Primijetimo da je za $x \notin \text{Im}X$, $f_X(x) = 0$. Nadalje, vrijedi:

$(G1)$ $f_X(x) \geq 0$ za sve x

$(G2)$ $\sum_{x \in \text{Im}X} f_X(x) = 1$.

Ako neka realna funkcija s diskretnom slikom ima svojstva $(G1-2)$, tada kažemo da je ona *funkcija gustoće neke diskretne vjerojatnosne razdiobe*.

Svakoj slučajnoj varijabli pridružujemo *funkciju distribucije* F_X koja se definira kao funkcija $F_X : \mathbb{R} \to \mathbb{R}$,

$$F_X(x) := \mathbb{P}(X \leq x).$$

Za diskretnu slučajnu varijablu X s gustoćom razdiobe f_X vrijedi

$$F_X(x) = \sum_{\{y \in \text{Im}X : y \leq x\}} f_X(y). \tag{2.3}$$

Naime,

$$F_X(x) = \mathbb{P}(X \leq x) = \mathbb{P}(\bigcup_{y \leq x} \{X = y\}) \overset{(P3)}{=} \sum_{y \leq x} \mathbb{P}(X = y) = \sum_{y \leq x} f_X(y).$$

Primijetite da je funkcija (2.3) stepenasta, rastuća, neprekidna zdesna i da vrijedi

$$\lim_{x \to -\infty} F_X(x) = 0, \quad \lim_{x \to +\infty} F_X(x) = 1.$$

2.3 Neprekidne slučajne varijable

Slučajna varijabla X je *neprekidna* ako vrijedi:

(i) ImX je interval u $\mathbb{R}$,

(ii) Skup $\{a \leq X \leq b\}$ je događaj za sve realne brojeve $a < b$,

(iii) Postoji funkcija $f_X : \mathbb{R} \to \mathbb{R}$ takva da je za sve $a < b$,

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x)\, dx.$$

Funkciju f_X zovemo *funkcijom gustoće razdiobe* od X ili, jednostavno, *gustoćom* razdiobe od X.

Budući da za neprekidnu slučajnu varijablu X vrijedi da je za sve x (uključujući i $x \in \mathrm{Im}X$), $\mathbb{P}(X = x) = 0$, slijedi da je za sve $a < b$,

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

Nadalje, za gustoću vrijedi:

$(G1)$ $f_X(x) \geq 0$ za sve x

$(G2)$ $\int_{-\infty}^{+\infty} f_X(x)\,dx = 1.$

Slično kao i u diskretnom slučaju, ako neka realna funkcija kojoj je slika interval realnih brojeva ima svojstva $(G1-2)$, tada kažemo da je to *funkcija gustoće neke neprekidne vjerojatnosne razdiobe*.

Za funkciju distribucije neprekidne slučajne varijable X s gustoćom f_X vrijedi:

$$F_X(x) = \int_{-\infty}^{x} f_X(y)\,dy. \tag{2.4}$$

Funkcija (2.4) je neprekidna, rastuća, a u $-$ i $+$ beskonačnosti teži vrijednostima 0 i 1, redom. Iz (2.4) i svojstva (iii) iz definicije, slijedi

$$\mathbb{P}(a \leq X \leq b) = F_X(b) - F_X(a).$$

Pomoću te relacije može se pokazati da je

$$\frac{dF_X}{dx}(x) = f_X(x)$$

u vrijednostima x za koje je funkcija distribucije F_X derivabilna.

2.4 Matematičko očekivanje

Matematičko očekivanje slučajne varijable X interpretira se kao srednja (očekivana) vrijednost od X. Definira se kao broj $\mathbb{E}[X]$:

$$\begin{aligned}
\mathbb{E}[X] &:= \sum_{x \in \mathrm{Im}X} x f_X(x) \quad \text{(ako je } X \text{ diskretna)} \\
\mathbb{E}[X] &:= \int_{-\infty}^{+\infty} x f_X(x)\,dx \quad \text{(ako je } X \text{ neprekidna),}
\end{aligned}$$

pod pretpostavkom da desne strane postoje u smislu da (red/integral) apsolutno konvergiraju.

Ako je $g : \mathbb{R} \to \mathbb{R}$ neka (dobra) realna funkcija (npr. po dijelovima neprekidna) i $X : \Omega \to \mathbb{R}$ slučajna varijabla, tada je $g(X) = g \circ X : \Omega \to \mathbb{R}$ također slučajna varijabla, pa ima smisla računati $\mathbb{E}[g(X)]$. Vrijedi:

$$\begin{aligned}
\mathbb{E}[g(X)] &= \sum_{x \in \mathrm{Im}X} g(x) f_X(x) \quad \text{(ako je } X \text{ diskretna)} \\
\mathbb{E}[g(X)] &= \int_{-\infty}^{+\infty} g(x) f_X(x)\,dx \quad \text{(ako je } X \text{ neprekidna),}
\end{aligned}$$

pod pretpostavkom da desne strane postoje u smislu da (red/integral) apsolutno konvergiraju. Pomoću tih formula može se pokazati da matematičko očekivanje ima svojstvo *linearnosti*. Naime, za realne funkcije g_1, g_2,..., g_k i brojeve c_1, c_2,..., c_k vrijedi

$$\mathbb{E}[\sum_{i=1}^{k} c_i g_i(X)] = \sum_{i=1}^{k} c_i \mathbb{E}[g_i(X)].$$

2.5 Varijanca i standardna devijacija

Varijanca slučajne varijable je mjera raspršenja njenih vrijednosti od matematičkog očekivanja. Preciznije, varijanca od X je srednje kvadratno odstupanje X od $\mathbb{E}[X]$:

$$\mathbb{V}\mathrm{ar}[X] := \mathbb{E}[(X - \mathbb{E}[X])^2].$$

Korištenjem linearnosti matematičkog očekivanja može se pokazati da je

$$\mathbb{V}\mathrm{ar}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2. \tag{2.5}$$

Standardna devijacija od X je drugi korijen varijance:

$$\sigma(X) := \sqrt{\mathbb{V}\mathrm{ar}[X]}.$$

Standardnom devijacijom se raspršenje izražava u istim fizikalnim jedinicama u kojima se izražavaju vrijednosti od X.

2.6 Matematičko očekivanje i varijanca linearne transformacije od X

Neka je X slučajna varijabla s matematičkim očekivanjem μ i varijancom σ^2, te neka su $a \neq 0$, b realni brojevi. Tada je $Y := aX + b$ također slučajna varijabla. Korištenjem svojstva linearnosti očekivanja dobijamo:

$$\begin{aligned} \mathbb{E}[Y] &= \mathbb{E}[aX + b] = a\mathbb{E}[X] + b = \\ &= a\mu + b \qquad (2.6) \\ \mathbb{V}\mathrm{ar}[Y] &= \mathbb{E}[(Y - a\mu - b)^2] = \mathbb{E}[(aX + b - a\mu - b)^2] = \\ &= \mathbb{E}[a^2(X - \mu)^2] = a^2\mathbb{E}[(X - \mu)^2] = a^2\mathbb{V}\mathrm{ar}[X] = \\ &= a^2\sigma^2 \qquad (2.7) \end{aligned}$$

Slučajnoj varijabli X pridružimo njenu *standardiziranu verziju*

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

Iz (2.6) i (2.7) slijedi da je standardizirana verzija centrirana ($\mathbb{E}[Z] = 0$) i da ima jediničnu varijancu ($\mathbb{V}\mathrm{ar}[Z] = 1$).

2.7 Momenti

Neka je X slučajna varijabla, k prirodan, a c neki realan broj. *k-ti moment od X oko c* je broj

$$\mathbb{E}[(X-c)^k].$$

Momenti oko ishodišta ($c=0$) jednostavno se zovu *momentima*. *Centralni momenti* su momenti oko matematičkog očekivanja. Na primjer, matematičko očekivanje je prvi moment, a varijanca drugi centralni moment. Prvi centralni moment je identički jednak nuli. Korištenjem linearnosti očekivanja može se pokazati da za treći centralni moment μ_3 vrijedi

$$\mu_3 = \mathbb{E}[X^3] - 3\mu\mathbb{E}[X^2] + 2\mu^3,$$

gdje je $\mu = \mathbb{E}[X]$. Treći centralni moment standardizirane verzije Z od X, u oznaci $\alpha_3(X)$, zove se *koeficijent asimetrije* od X. Dakle,

$$\alpha_3(X) = \mathbb{E}[\left(\frac{X-\mu}{\sigma}\right)^3],$$

gdje su $\mu = \mathbb{E}[X]$ i $\sigma = \sigma(X)$. Distribucija od X je *simetrična* ako je $\alpha_3(X) = 0$, *negativno* je *asimetrična* ako je $\alpha_3(X) < 0$, a *pozitivno asimetrična* ako je $\alpha_3(X) > 0$.

2.8 Primjeri važnih distribucija

Najvažnije svojstvo slučajnih varijabli za primjene je njihova distribucija opisana gustoćom ili funkcijom distribucije. U ovoj točki navedene distribucije prirodno se pojavljuju u mnogim područjima primjene. Uz precizno opisane razdiobe, navedena je i njihova interpretacija.

2.8.1 Diskretne razdiobe

Diskretne slučajne varijable najčešće se interpretiraju kao broj nečega, odnosno kao rezultat nekog procesa prebrojavanja (broj uspjeha, broj smrti, broj šteta po polici osiguranja i sl.).

Uniformna razdioba

Slučajna varijabla X ima *uniformnu razdiobu* na skupu $S = \{1, 2, \ldots, k\}$ (k je prirodni broj) ako je

$$f_X(x) = \mathbb{P}(X = x) = \frac{1}{k} \text{ za } x \in S = \mathrm{Im}X.$$

Tada je

$$\begin{aligned}
\mathbb{E}[X] &= 1\cdot\frac{1}{k} + 2\cdot\frac{1}{k} + \cdots + k\cdot\frac{1}{k} = \frac{k+1}{2} \\
\mathbb{E}[X^2] &= 1^2\cdot\frac{1}{k} + 2^2\cdot\frac{1}{k} + \cdots + k^2\cdot\frac{1}{k} = \frac{(k+1)(2k+1)}{6} \\
\Rightarrow\ \mathbb{V}\mathrm{ar}[X] &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{(k+1)(2k+1)}{6} - (\frac{k+1}{2})^2 = \frac{k^2-1}{12}.
\end{aligned}$$

Na primjer, u pokusu bacanja simetrične kocke neka je X jednako broju koji se okrenuo. Tada X ima uniformnu razdiobu na skupu $\{1, 2, 3, 4, 5, 6\}$.

Bernoullijeva razdioba

Bernoullijeva slučajna varijabla X indicira je li rezultat slučajnog pokusa bio "uspjeh" ili ne. Preciznije, X će imati vrijednost 1 ako se dogodio elementaran ishod koji je povoljan za događaj "uspjeh", inače će imati vrijednost 0. Dakle, $\mathrm{Im}X = \{0, 1\}$. Označimo sa $\theta = \mathbb{P}(X = 1)$ *vjerojatnost uspjeha*. Primijetimo da je nužno $\theta \in [0, 1]$. Tada je funkcija vjerojatnosti od X jednaka

$$f_X(x) = \theta^x \cdot (1-\theta)^{1-x} \text{ za } x \in \mathrm{Im}X = \{0, 1\}$$

i vrijedi

$$\begin{aligned} \mathbb{E}[X] &= 0 \cdot (1-\theta) + 1 \cdot \theta = \theta \\ \mathbb{E}[X^2] &= 0^2 \cdot (1-\theta) + 1^2 \cdot \theta = \theta \\ \Rightarrow \mathbb{V}\mathrm{ar}[X] &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \theta - \theta^2 = \theta(1-\theta). \end{aligned}$$

Na primjer, recimo da se u pokusu bacanja simetrične kocke dogodio uspjeh ako se okrenula šestica. Tada slučajna varijabla X koja ima vrijednost 1 kada se okrene šestica, a inače je jednaka 0, ima Bernullijevu razdiobu s vjerojatnosti uspjeha 1/6.

Binomna razdioba

Zamislimo slučajni pokus koji se sastoji od n *nezavisnih jednako distribuiranih Bernoullijevih pokusa*. Bernoullijevim pokusima zvat ćemo one slučajne pokuse kod kojih nas samo zanima je li se dogodio ili se nije dogodio uspjeh. Jednaka distribuiranost takvih pokusa znači da je u svima njima vjerojatnost uspjeha θ jednaka, a da su nezavisni znači da ishod svakog od njih ne ovisi o ishodima ostalih pokusa. Matematički, događaji su *nezavisni* ako je za svaku njihovu kombinaciju (ili kombinaciju njihovih komplemenata, ili zajedničku kombinaciju s komplementima) vjerojatnost istovremenog događanja te kombinacije događaja jednako produktu vjerojatnosti svakog događaja u toj kombinaciji posebno. Tada slučajna varijabla X koja je jednaka ukupnom broju uspjeha u tih n pokusa ima *binomnu razdiobu* s parametrima n i θ ($0 \leq \theta \leq 1$). Vrijedi:

$$f_X(x) = \binom{n}{x} \theta^x (1-\theta)^{n-x} \text{ za } x \in \mathrm{Im}X = \{0, 1, \ldots, n\},$$

$$\mathbb{E}[X] = n\theta, \quad \mathbb{V}\mathrm{ar}[X] = n\theta(1-\theta).$$

Iz navedene interpretacije binomne slučajne varijable slijedi da se ta varijabla može prikazati kao zbroj od n nezavisnih jednako distribuiranih Bernoullijevih slučajnih varijabli. Svaka od tih n Bernoullijevih varijabli interpretira se kao indikator uspjeha u jednom od Bernoullijevih pokusa.

Geometrijska razdioba

Izvodi se niz nezavisnih jednako distribuiranih Bernoullijevih pokusa s vjerojatnosti uspjeha θ sve dok se ne dogodi (prvi put) uspjeh. Tada je broj pokusa do prvog uspjeha X slučajna varijabla koja ima *geometrijsku razdiobu* s parametrom θ, $0 < \theta < 1$. Ako redni broj pokusa interpretiramo kao vrijeme, tada X pripada klasi *vremena čekanja*. Vrijedi:

$$f_X(x) = \theta(1-\theta)^{x-1} \text{ za } x \in \mathrm{Im}X = \{1, 2 \ldots\},$$

$$\mathbb{E}[X] = \frac{1}{\theta}, \quad \mathbb{V}\mathrm{ar}[X] = \frac{1-\theta}{\theta^2}.$$

Geometrijska slučajna varijabla ima svojstvo *neimanja memorije*. Naime, za sve prirodne brojeve n i k vrijedi:

$$\mathbb{P}(X > n + k \mid X > n) = \mathbb{P}(X > k).$$

Riječima, vjerojatnost da treba čekati više od $n + k$ pokusa do uspjeha ako se zna da u prethodnih n pokusa nije bilo uspjeha, jednaka je vjerojatnosti da do prvog uspjeha treba čekati više od k pokusa. Dakle, irelevantno je što u prvih n pokusa nije bilo uspjeha. Šansa za uspjeh nakon n neuspjeha nije ni veća, ni manja.

Geometrijska slučajna varijabla se može definirati i kao broj neuspjeha do prvog uspjeha. Označimo tu varijablu sa Y. Tada je $Y = X - 1$ i

$$f_Y(x) = \theta(1-\theta)^x \text{ za } x \in \mathrm{Im}Y = \{0, 1, 2 \ldots\},$$

$$\mathbb{E}[Y] = \frac{1-\theta}{\theta}, \quad \mathbb{V}\mathrm{ar}[Y] = \frac{1-\theta}{\theta^2}.$$

Negativna binomna razdioba

Negativna binomna distribucija je poopćenje geometrijske distribucije. Broj X nezavisnih jednako distribuiranih Bernoullijevi pokusa s vjerojatnosti uspjeha θ do uključivo k-tog uspjeha je slučajna varijabla koja ima *negativnu binomnu razdiobu* s parametrima k i θ, $0 < \theta < 1$. Vrijedi:

$$f_X(x) = \binom{x-1}{k-1}\theta^k(1-\theta)^{x-k} \text{ za } x \in \mathrm{Im}X = \{k, k+1, \ldots\},$$

$$\mathbb{E}[X] = \frac{k}{\theta}, \quad \mathbb{V}\mathrm{ar}[X] = k\frac{1-\theta}{\theta^2}.$$

Za računanje funkcije vjerojatnosti $f_X(x) = \mathbb{P}(X = x)$ koristi se rekurzivna relacija

$$f_X(x) = \frac{x-1}{x-k}(1-\theta)f_X(x-1), \text{ za } x = k+1, k+2, \ldots \text{ i } f_X(k) = \theta^k.$$

Negativna binomna slučajna varijabla X može se prikazati kao zbroj k nezavisnih jednako distribuiranih geometrijskih slučajnih varijabli od kojih svaka predstavlja vrijeme čekanja između dva uspjeha.

Kao u slučaju geometrijske razdiobe, negativna binomna slučajna varijabla može se definirati i kao broj neuspjeha do k-tog uspjeha. Označimo tu varijablu sa Y. Tada je $Y = X - k$ i

$$f_Y(x) = \binom{k+x-1}{k-1}\theta^k(1-\theta)^x \text{ za } x \in \mathrm{Im}Y = \{0, 1, 2, \ldots\},$$

$$\mathbb{E}[Y] = k\frac{1-\theta}{\theta}, \quad \mathbb{V}\mathrm{ar}[Y] = k\frac{1-\theta}{\theta^2}.$$

Hipergeometrijska distribucija

Promotrimo sljedeći primjer. Od N kuglica u kutiji, njih K su bijele, a ostale su crne. Na slučajan način izvlačimo jednu za drugom n kuglica *bez vraćanja*. Uspjeh je kada izvučemo bijelu kuglicu. Pretpostavimo da je $n \leq K \leq N$ i označimo sa X ukupan broj bijelih kuglica među n izvučenih. Drugim riječima, X je ukupan broj uspjeha tijekom izvođenja n

Bernoullijevih pokusa koji nisu niti nezavisni, niti jednako distribuirani. Tada je X slučajna varijabla kojoj je funkcija vjerojatnosti:

$$f_X(x) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}} \text{ za } x \in \mathrm{Im}X = \{0, 1, \ldots, n\}.$$

Kažemo da slučajna varijabla ima *hipergeometrijsku razdiobu* ako joj je razdioba opisana gore navedenom funkcijom vjerojatnosti. Ako sa $\theta = K/N$ označimo vjerojatnost uspjeha u prvom izvlačenju, tada je $\mathbb{E}[X] = n\theta$.

Ako je N dovoljno velik (u apsolutnom smislu i u odnosu na n), tada je binomna razdioba s parametrima n i θ dobra aproksimacija za razdiobu od X. Primijetimo da je ta binomna razdioba dobar model za ukupan broj uspjeha u gore navedenom primjeru, ali kada kuglice izvlačimo *sa vraćanjem*.

Poissonova razdioba

Poissonova distribucija modelira broj slučajnih događaja koji se realiziraju tijekom nekog vremenskog intervala, a koji zadovoljavaju sljedeće uvjete:

(*i*) vjerojatnost pojavljivanja jednog događaja tijekom nekog vremenskog intervala proporcionalna je duljini tog intervala s konstantom proporcionalnosti neovisnoj o vremenskom intervalu;

(*ii*) vjerojatnost istovremenog pojavljivanja dva i više događaja je jednaka nuli;

(*iii*) brojevi pojavljivanja događaja tijekom međusobno disjunktnih vremenskih intervala su nezavisni.

Kažemo da se događaji pojavljuju u skladu sa zakonom *Poissonovog procesa*.

Druga interpretacija Poissonove distribucije je kao granične distribucije binomne razdiobe s parametrima n i θ kada n teži beskonačnosti ($n \to +\infty$), a θ teži nuli ($\theta \to 0$) na način da je broj $\lambda = n\theta$ konstantan.

Kažemo da diskretna slučajna varijabla ima *Poissonovu distribuciju* s parametrom λ, $\lambda > 0$, i pišemo $X \sim P(\lambda)$, ako je njena funkcija vjerojatnosti oblika:

$$f_X(x) = \frac{\lambda^x}{x!}e^{-\lambda} \text{ za } x \in \mathrm{Im}X = \{0, 1, \ldots\}.$$

Vrijedi:

$$\mathbb{E}[X] = \mathbb{V}\mathrm{ar}[X] = \lambda.$$

Poissonovu razdiobu koristimo za aproksimaciju binomne s parametrima n i θ kada je n veliko, a θ malo (na primjer, $n \geq 100$ i $\theta \leq 0.05$). Za parametar Poissonove razdiobe uzimamo $\lambda = n\theta$.

Kada se događaji pojavljuju u skladu sa zakonom Poissonovog procesa *s intenzitetom* λ (kaže se još da su događaji *slučajni s intenzitetom* λ *po jedinici vremena*), tada broj događaja tijekom vremenskog intervala duljine t ima Poissonovu razdiobu $P(\lambda t)$.

2.8.2 Neprekidne razdiobe

Uniformna razdioba

Kažemo da neprekidna slučajna varijabla X ima *uniformnu razdiobu* na intervalu $\langle \alpha, \beta \rangle$ ako joj je gustoća razdiobe

$$f_X(x) = \begin{cases} \frac{1}{\beta - \alpha} & \text{za } x \in \langle \alpha, \beta \rangle \\ 0 & \text{inače.} \end{cases}$$

Vrijedi:

$$\mathbb{E}[X] = \frac{\alpha+\beta}{2}, \;\; \mathbb{V}\mathrm{ar}[X] = \frac{(\beta-\alpha)^2}{12}.$$

Svojstvo te razdiobe je da su vjerojatnosti podintervala iste duljine jednake.

Gama distribucija

Kažemo da slučajna varijabla X ima *gama distribuciju* s parametrima $\alpha > 0$ i $\lambda > 0$, i pišemo $X \sim \Gamma(\alpha, 1/\lambda)$, ako je strogo pozitivna ($\mathrm{Im}X = \langle 0, +\infty\rangle$) i gustoća razdiobe je

$$f_X(x) = \begin{cases} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} & \text{za } \; x > 0 \\ 0 & \text{inače,} \end{cases}$$

gdje je $\Gamma(\alpha) = \int_0^{+\infty} t^{\alpha-1} e^{-t}\, dt$ Γ-funkcija. Svojstva te funkcije su:

(*i*) $\Gamma(1) = 1$, $\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$ za $\alpha > 1$, odakle slijedi da je $\Gamma(n) = (n-1)!$ za prirodan broj n;

(*ii*) $\Gamma(1/2) = \sqrt{\pi}$.

Vrijedi:

$$\mathbb{E}[X] = \frac{\alpha}{\lambda}, \;\; \mathbb{V}\mathrm{ar}[X] = \frac{\alpha}{\lambda^2}.$$

Eksponencijalna distribucija

Ako je $X \sim \Gamma(1, 1/\lambda)$, tada kažemo da X ima *eksponencijalnu distribuciju* s parametrom $\lambda > 0$, i pišemo $X \sim Exp(\lambda)$. Dakle, eksponencijalna distribucija je specijalni slučaj gama distribucije kada je parametar $\alpha = 1$. Prema tome, gustoća razdiobe i funkcija distribucije od X, te matematičko očekivanje i varijanca, su:

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{za } \; x > 0 \\ 0 & \text{inače,} \end{cases} \qquad F_X(x) = \begin{cases} 1 - e^{-\lambda x} & \text{za } \; x > 0 \\ 0 & \text{inače,} \end{cases}$$

$$\mathbb{E}[X] = \frac{1}{\lambda}, \;\; \mathbb{V}\mathrm{ar}[X] = \frac{1}{\lambda^2}.$$

Eksponencijalna distribucija se koristi kao jednostavan model za vrijeme trajanja nekih vrsta uređaja.

Važna interpretacija te distribucije je kao modela za vrijeme čekanja T između pojavljivanja dva događaja u Poissonovom procesu. Preciznije, ako Poissonov proces ima intenzitet λ, te ako je X broj događaja u vremenskom intervalu $[0, t]$, tada je (zbog $X \sim P(\lambda t)$):

$$\mathbb{P}(T > t) = \mathbb{P}(X = 0) = e^{-\lambda t} \;\Rightarrow\; F_T(t) = 1 - \mathbb{P}(T > t) = 1 - e^{-\lambda t} \;\Rightarrow\; T \sim Exp(\lambda).$$

Nadalje, za sve pozitivne s i t vrijedi:

$$\mathbb{P}(T > t + s \,|\, T > t) = \mathbb{P}(T > s)$$

što znači da ako vrijeme mjerimo od bilo koje ishodišne točke (ne nužno od vremena kada se zadnji događaj pojavio), da vrijeme čekanja ima istu eksponencijalnu razdiobu. Prema tome, kao i u slučaju geometrijske razdiobe, eksponencijalna razdioba ima svojstvo neimanja memorije.

Može se pokazati da se $\Gamma(k, 1/\lambda)$-razdioba, gdje je k prirodan broj, može interpretirati kao zbroj od k nezavisnih $Exp(\lambda)$-distributiranih slučajnih varijabli. Drugim riječima, slučajna varijabla s tom gama razdiobom se interpetira kao vrijeme čekanja da se dogodi točno k događaja u Poissonovom procesu s intenzitetom λ.

χ^2-razdioba

Ako je $X \sim \Gamma(\frac{n}{2}, 2)$ za n prirodan broj, tada kažemo da X ima *χ^2-razdiobu* s n *stupnjeva slobode*, i pišemo $X \sim \chi^2(n)$. Vrijedi:

$$\mathbb{E}[X] = n, \;\; \mathbb{V}\mathrm{ar}[X] = 2n.$$

Primijetimo da je $\chi^2(2) = Exp(\frac{1}{2})$.

Beta distribucija

Neprekidna slučajna varijabla X ima *beta distribuciju* s parametrima $\alpha > 0$ i $\beta > 0$, i pišemo $X \sim B(\alpha, \beta)$, ako prima vrijednosti u intervalu $\langle 0, 1\rangle$ i gustoća joj je:

$$f_X(x) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} & \text{za } \; 0 < x < 1 \\ 0 & \text{inače,} \end{cases}$$

Primijetite da vrijedi

$$\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\, dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

Desna strana te jednakosti se označava sa $B(\alpha, \beta)$. Funkcija koja paru pozitivnih brojeva (α, β) pridružuje tu vrijednost zove se *beta funkcija*.

Vrijedi:

$$\mathbb{E}[X] = \frac{\alpha}{\alpha+\beta}, \;\; \mathbb{V}\mathrm{ar}[X] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

Uniformna razdioba na $\langle 0, 1\rangle$ je poseban slučaj beta razdiobe kada su parametri $\alpha = \beta = 1$.

Normalna razdioba

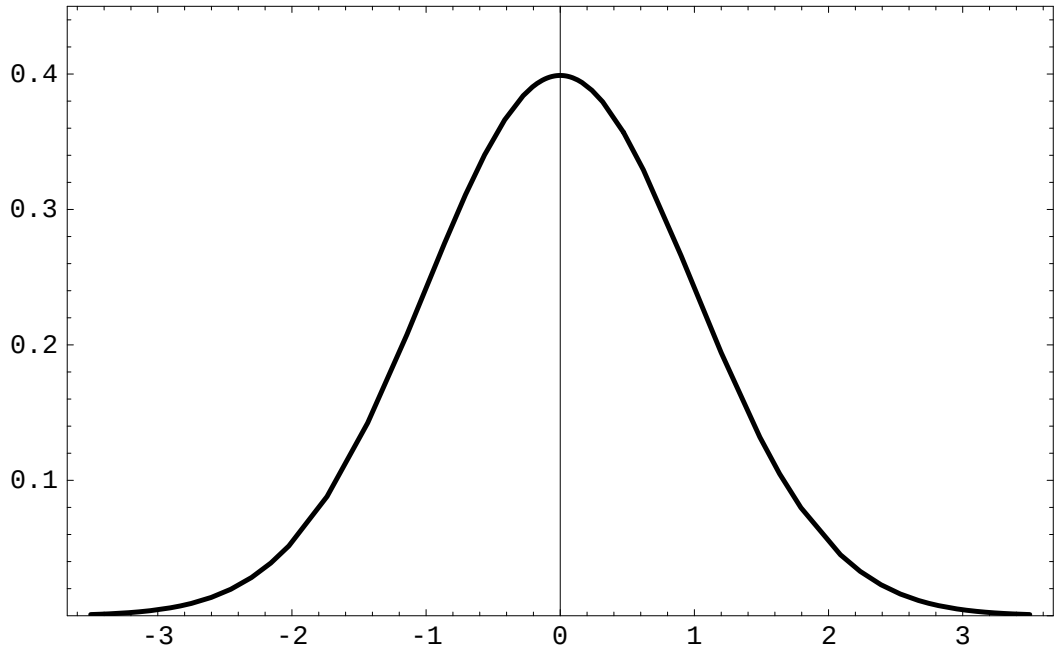
Kažemo da X ima *normalnu razdiobu* s parametrima μ i $\sigma^2 > 0$, i pišemo $X \sim N(\mu, \sigma^2)$, ako je $\mathrm{Im}X = \mathbb{R}$ i gustoća joj je

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Ta razdioba je jedna od najvažnijih jer:

1. dobar je model za veliku većinu fizikalnih mjerenja
2. dobra je aproksimacija velike klase drugih distribucija (na primjer, binomne)
3. dobar je model za uzoračku razdiobu raznih statistika
4. zaključivanje na osnovi velikih uzoraka i neki statistički postupci zasnivaju se na pretpostavci normalnosti
5. pomoću nje se izvode mnoge druge distribucije

Graf funkcije gustoće se zove *Gaussova krivulja* i ona je zvonolikog oblika.



Interpretacija parametara normalne razdiobe je da je $\mu = \mathbb{E}[X]$ i $\sigma^2 = \mathbb{V}\mathrm{ar}[X]$.

Linearna transformacija normalno distribuirane varijable je opet normalno distribuirana varijabla. Preciznije, ako je $X \sim N(\mu, \sigma^2)$, te ako su $a \neq 0$ i b realni brojevi, tada je $Y := aX + b \sim N(a\mu + b, a^2\sigma^2)$. Specijalno, standardizirana verzija normalne varijable X, $Z = (X - \mu)/\sigma$, je normalno distribuirana s očekivanjem 0 i varijancom 1. Kažemo da Z ima *jediničnu normalnu razdiobu*. Vrijednosti od Z su bezdimenzionalne (u smislu da nisu izražene nekim fizikalnim jedinicama) i njima izražavamo koliko je standardnih devijacija pripadna vrijednost X udaljena (i na koju stranu) od svoje očekivane vrijednosti μ. Ako je $Z < 0$, tada je X za $|Z|$ standardnih devijacija manji od μ, a ako je $Z > 0$, tada je X za Z standardnih devijacija veći od μ.

Vrijednosti jedinične normalne razdiobe su tabelirane. Preciznije, ako je Φ funkcija distribucije od $N(0, 1)$,

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\, dt,$$

tada su tabelirane vrijednosti funkcije

$$\Phi_0(x) = \int_{0}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\, dt, \quad \text{za } \; x > 0.$$

Ta se funkcija može proširiti (po neparnosti) na sve realne brojeve x relacijom

$$\Phi_0(x) := -\Phi_0(-x), \quad \text{za } \; x < 0. \tag{2.8}$$

Očito mora biti $\Phi_0(0) = 0$. Φ i Φ_0 su vezane relacijom

$$\Phi(x) = \frac{1}{2} + \Phi_0(x), \quad \text{za } \; x \in \mathbb{R}. \tag{2.9}$$

Na primjer, iz tablica čitamo da je

$$\mathbb{P}(0 < Z < 1.96) = \Phi_0(1.96) = 0.475,$$

pa je

$$\begin{aligned}\mathbb{P}(Z<1.96) &= \Phi(1.96) \overset{(2.9)}{=} 0.5+0.475=0.975\\ \mathbb{P}(-1.96<Z<1.96) &= \Phi(1.96)-\Phi(-1.96) \overset{(2.9)}{=} \Phi_0(1.96)-\Phi_0(-1.96) \overset{(2.8)}{=} 2\cdot 0.475 = \\ &= 0.950.\end{aligned}$$

Slično se izračuna

$$\mathbb{P}(-2.576<Z<2.576)=0.99 \text{ i } \mathbb{P}(-3<Z<3)=0.997$$

Dakle, u 95% slučajeva će se vrijednosti normalne varijable od svoje očekivane vrijednosti razlikovati za ne više od 1.96 standardnih devijacija, a u 99% slučajeva ta razlika neće biti veća od 2.576 standardnih devijacija. Zadnji izračun kaže da se 99.7% svih realiziranih vrijednosti normalno distribuirane varijable od matematičkog očekivanja neće razlikovati za više od tri standardne devijacije. Ta tvrdnja se zove *pravilo* 3σ.

2.9 Funkcije slučajnih varijabli

Ako je X slučajna varijabla s poznatom distribucijom (zna se f_X ili F_X) te ako je zadana (dobra) realna funkcija g realne varijable, tada je $Y=g(X)$ slučajna varijabla. Cilj nam je odrediti razdiobu od Y, tj. izračunati gustoću f_Y ili funkciju distribucije F_Y.

2.9.1 Diskretne razdiobe

Ako je X diskretna, tada je i Y diskretna slučajna varijabla. Nadalje, $\mathrm{Im}Y \subseteq \mathrm{Im}\, g = g(\mathbb{R})$. Budući da je za $y \in \mathrm{Im}Y$,

$$\mathbb{P}(Y=y)=\mathbb{P}\Big(\bigcup_{\{x\in \mathrm{Im}X: g(x)=y\}} \{X=x\}\Big) \overset{(P3)}{=} \sum_{\{x\in \mathrm{Im}X: g(x)=y\}} \mathbb{P}(X=x),$$

slijedi formula za gustoću od Y:

$$f_Y(y)=\sum_{\{x\in \mathrm{Im}X: g(x)=y\}} f_X(x), \quad y\in \mathrm{Im}Y.$$

Primjer 2.2 Neka je X binomna varijabla s parametrima $n=10$ i $\theta\in\langle 0,1\rangle$, i $Y=\sin\frac{\pi}{2}X$. Tada je $\mathrm{Im}Y=\{-1,0,1\}$ i $g(x)=\sin\frac{\pi}{2}x$. Prema gornjoj formuli je

$$\begin{aligned}f_Y(-1) &= \sum_{\{0\le k\le 10:\sin\frac{\pi}{2}k=-1\}} f_X(k)=\binom{10}{3}\theta^3(1-\theta)^7+\binom{10}{7}\theta^7(1-\theta)^3\\ f_Y(0) &= \sum_{\{0\le k\le 10:\sin\frac{\pi}{2}k=0\}} f_X(k)=\sum_{i=0}^{5}\binom{10}{2i}\theta^{2i}(1-\theta)^{2(5-i)}=\frac{1+(1-2\theta)^{10}}{2}\\ f_Y(1) &= \sum_{\{0\le k\le 10:\sin\frac{\pi}{2}k=1\}} f_X(k)=\binom{10}{1}\theta(1-\theta)^9+\binom{10}{5}\theta^5(1-\theta)^5+\binom{10}{9}\theta^9(1-\theta).\end{aligned}$$

□

2.9.2 Neprekidne razdiobe

Neka je X neprekidna slučajna varijabla, a F_x i f_X njene funkcije distribucije i gustoće. Pretpostavimo da je funkcija g strogo rastuća na intervalu $\mathrm{Im}X$. Tada je za $y \in \mathrm{Im}Y$,

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(g(X) \leq y) = \mathbb{P}(X \leq g^{-1}(y)) = F_X(g^{-1}(y)). \tag{2.10}$$

Ako je g strogo padajuća na $\mathrm{Im}X$, tada je za sve $y \in \mathrm{Im}Y$,

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(g(X) \leq y) = \mathbb{P}(X \geq g^{-1}(y)) = 1 - F_X(g^{-1}(y)). \tag{2.11}$$

U slučaju da je g^{-1} diferencijabilna u $y \in \mathrm{Im}Y$,

$$f_Y(y) = \frac{dF_Y}{dy}(y).$$

Budući da za rastuću funkciju g vrijedi (2.10), slijedi

$$\frac{dF_Y}{dy}(y) = \frac{dF_X}{dx}(g^{-1}(y))\frac{dg^{-1}}{dy}(y) = f_X(g^{-1}(y))\frac{dg^{-1}}{dy}(y).$$

Slično, budući da za padajuću funkciju g rijedi (2.11), slijedi

$$\frac{dF_Y}{dy}(y) = -\frac{dF_X}{dx}(g^{-1}(y))\frac{dg^{-1}}{dy}(y) = -f_X(g^{-1}(y))\frac{dg^{-1}}{dy}(y).$$

Dakle, ako je inverz strogo monotone funkcije g diferencijabilan, vrijedi formula:

$$f_Y(y) = f_X(g^{-1}(y))\left|\frac{dg^{-1}}{dy}(y)\right| \quad \text{za} \quad y \in \mathrm{Im}Y. \tag{2.12}$$

Primjer 2.3 Kažemo da slučajna varijabla Y ima *log-normalnu razdiobu* s parametrima μ i σ^2, ako je strogo pozitivna i $\log Y$ ima normalnu distribuciju $N(\mu, \sigma^2)$. Označimo $X = \log Y$. Tada je $Y = e^X$, $X \sim N(\mu, \sigma^2)$ i $g(x) = e^x$. g je strogo rastuća funkcija na $\mathbb{R} = \mathrm{Im}X$, a njen inverz $g^{-1}(y) = \log y$ je diferencijabilna funkcija na $\langle 0, +\infty\rangle = \mathrm{Im}Y$ s derivacijom $(g^{-1})'(y) = 1/y$. Prema formuli (2.12),

$$f_Y(y) = f_X(\log y) \cdot \frac{1}{y} = \frac{1}{y\sigma\sqrt{2\pi}} e^{-\frac{(\log y - \mu)^2}{2\sigma^2}} \quad \text{za} \quad y > 0.$$

□

Često puta u primjenama funkcija g nije strogo monotona na intervalu $\mathrm{Im}X$, ali je takva po dijelovima. Na sljedećem primjeru ilustrirat ćemo kako se u tom slučaju izvodi formula za gustoću od $Y = g(X)$.

Primjer 2.4 Neka je $Y = X^2$. Tada je $g(x) = x^2$ i g očito nije strogo monotona na svakom intervalu realnih brojeva. Na primjer, nije monotona na $\mathbb{R}$. Nadalje, očito je $\mathrm{Im}Y \subseteq [0, +\infty\rangle$. Za $y > 0$ računamo

$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y \leq y) = \mathbb{P}(X^2 \leq y) = \mathbb{P}(|X| \leq \sqrt{y}) = \mathbb{P}(-\sqrt{y} \leq X \leq \sqrt{y}) = \\ &= F_X(\sqrt{y}) - F_X(-\sqrt{y}). \end{aligned}$$

Deriviranjem tog izraza dobijamo formulu:

$$f_Y(y) = (f_X(-\sqrt{y}) + f_X(\sqrt{y})) \cdot \frac{1}{2\sqrt{y}}, \quad y > 0. \tag{2.13}$$

Specijalno, ako je $X \sim N(0,1)$, tada $Y = X^2$ ima gustoću

$$f_Y(y) = \frac{2}{\sqrt{2\pi}} e^{-\frac{y}{2}} \frac{1}{2\sqrt{y}} = \frac{(2^{-1})^{\frac{1}{2}}}{\Gamma(\frac{1}{2})} y^{\frac{1}{2}-1} e^{-\frac{y}{2}}, \quad y > 0,$$

Dakle, $Y \sim \chi^2(1)$. □

U svim primjerima do sada Y je bila neprekidna slučajna varijabla. To općenito ne mora biti tako. Naime, iako je X neprekidna slučajna varijabla, funkcija g može biti takva da $Y = g(X)$ bude, na primjer, diskretna.

Primjer 2.5 Neka su $X \sim N(0,1)$ i $Y = \text{sign}X$. Tada je $Y = g(X)$, gdje je

$$g(x) = \text{sign}\, x = \begin{cases} -1, & x < 0 \\ 0, & x = 0 \\ 1, & x > 0. \end{cases}$$

Budući da je $\text{Im}Y \subseteq \text{Im}\, g = \{-1, 0, 1\}$, Y je diskretna slučajna varijabla s funkcijom vjerojatnosti:

$$f_Y(-1) = \mathbb{P}(Y = -1) = \mathbb{P}(X < 0) = \frac{1}{2}, \quad f_Y(1) = \mathbb{P}(Y = 1) = \mathbb{P}(X > 0) = \frac{1}{2}.$$

□

Poglavlje 3

Funkcije izvodnice

3.1 Funkcije izvodnice vjerojatnosti

Neka je X diskretna slučajna varijabla s vrijednostima u skupu prirodnih brojeva s nulom. Takvu varijablu zvat ćemo *brojeća* slučajna varijabla. Označimo sa p_0, p_1, itd. vjerojatnosti događaja $\{X = 0\}$, $\{X = 1\}$, itd., dakle,

$$p_k := \mathbb{P}(X = k) = f_X(k), \quad k = 0, 1, \ldots$$

Tada je *funkcija izvodnica vjerojatnosti* od X, kraće f.i.v., realna funkcija G_X definirana sa

$$G_X(t) := \mathbb{E}[t^X] = p_0 + p_1 t + p_2 t^2 + \cdots,$$

za one realne brojeve t za koje to očekivanje postoji. Uvijek vrijedi

$$G_X(1) = 1, \quad G_X(0) = p_0 = \mathbb{P}(X = 0).$$

Primijetite da red potencija iz definicije f.i.v. apsolutno konvergira za sve $t \in \mathbb{R}$ za koje je $|t| \leq 1$. F.i.v. je *jedinstvena* u smislu da dvije brojeće slučajne varijable X i Y imaju iste f.i.v ako i samo ako su X i Y *po distribuciji jednake*, tj. ako je

$$f_X(x) = f_Y(x) \text{ za sve } x \in \{0, 1, \ldots\}.$$

Primjer 3.1 Ako X ima uniformnu razdiobu na $\{1, 2, \ldots, k\}$, tada joj je f.i.v.

$$G_X(t) = \frac{1}{k}(t + t^2 + \cdots + t^k) = \frac{t(1 - t^k)}{k(1 - t)} \text{ ako } t \neq 1.$$

□

Primjer 3.2 Ako X ima binomnu razdiobu s parametrima (n, θ), tada joj je f.i.v.

$$G_X(t) = \sum_{k=0}^{n} \binom{n}{k} t^k \theta^k (1 - \theta)^{n-k} = (\theta t + 1 - \theta)^n.$$

□

Primjer 3.3 Za negativnu binomnu varijablu X s parametrima (k, θ), f.i.v. je

$$G_X(t) = \sum_{m=k}^{\infty} \binom{m-1}{k-1} t^m \theta^k (1 - \theta)^{m-k} = \left(\frac{\theta t}{1 - t(1 - \theta)}\right)^k,$$

pri čemu red apsolutno konvergira za $|t(1 - \theta)| < 1$. □

Primjer 3.4 Za $X \sim P(\lambda)$, f.i.v. je

$$G_X(t) = \sum_{k=0}^{\infty} t^k \frac{\lambda^k}{k!} e^{-\lambda} = e^{-\lambda(1-t)},$$

pri čemu red apsolutno konvergira za sve $t \in \mathbb{R}$. □

3.2 Računanje momenata pomoću f.i.v.

Pomoću f.i.v. $G_X(t)$ brojeće slučajne varijable X možemo računati momente od X nižeg reda. Rastavimo funkciju $t \mapsto t^X$ u Taylorov red u okolini $t = 1$:

$$t^X = 1 + X(t-1) + X(X-1)\frac{(t-1)^2}{2!} + X(X-1)(X-2)\frac{(t-1)^3}{3!} + \cdots,$$

a zatim izračunajmo matematičko očekivanje obje strane. Vrijedi da je

$$G_X(t) = \mathbb{E}[t^X] = 1 + \mathbb{E}[X](t-1) + \mathbb{E}[X(X-1)]\frac{(t-1)^2}{2!} + \mathbb{E}[X(X-1)(X-2)]\frac{(t-1)^3}{3!} + \cdots$$

za $t \geq 1$. Dakle,

$$\begin{aligned} \mathbb{E}[X] &= G'_X(1) \\ \mathbb{E}[X(X-1)] &= G''_X(1) \quad \Rightarrow \quad \mathbb{E}[X^2] = G'_X(1) + G''_X(1) \\ \Rightarrow \quad \mathbb{V}\mathrm{ar}[X] &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 = G''_X(1) + G'_X(1)(1 - G'_X(1)). \end{aligned}$$

3.3 Funkcija izvodnica momenata

Funkcija izvodnica momenata koristi se za računanje momenata slučajne varijable oko nule. Neka je X (diskretna ili neprekidna) slučajna varijabla. Tada je *funkcija izvodnica momenata* (kraće f.i.m.) od X, funkcija M_X definirana sa

$$M_X(t) := \mathbb{E}[e^{tX}],$$

za sve realne brojeve t za koje to očekivanje postoji. Odmah se vidi da je $M_X(0) = 1$. Razvijmo funkciju $t \mapsto e^{tX}$ u Taylorov red oko nule i (formalno) izračunajmo matematičko očekivanje dobivenog reda potencija kao red matematičkih očekivanja:

$$M_X(t) = \mathbb{E}[e^{tX}] = \mathbb{E}[\sum_{k=0}^{\infty} X^k \frac{t^k}{k!}] = \sum_{k=0}^{\infty} \mathbb{E}[X^k]\frac{t^k}{k!}.$$

Na primjer, to možemo učiniti u slučaju da je X nenegativna varijabla i f.i.m. je definirana za t u okolini nule. Tada očitavamo da je k-ti moment od X jednak derivaciji k-tog reda f.i.m. u $t = 0$:

$$\mathbb{E}[X^k] = M_X^{(k)}(0), \quad k = 1, 2, \ldots$$

Općenito, ta jednakost vrijedi ako postoje obje strane.

Ako znamo zakon razdiobe od X, onda možemo izračunati sve njene momente koje postoje. Obratno, ako postoje svi momenti od X (i poznati su), te ako su zadovoljeni još neki dodatni uvjeti na njih, tada taj niz momenata jednoznačno određuje razdiobu od X. Nadalje, dvije različite razdiobe ne mogu imati istu f.i.m. Dakle, svaka se f.i.m. može prepoznati kao f.i.m neke točno određene razdiobe.

F.i.m. brojeće slučajne varijable X može se jednostavno odrediti pomoću njene f.i.v. jer vrijedi (ako obje strane postoje za dani t):

$$M_X(t) = G_X(e^t).$$

Primjer 3.5 Pomoću f.i.v., izračunane su f.i.m. sljedećih brojećih varijabli:
(a) za X binomnu (n, θ):

$$M_X(t) = G_X(e^t) = (\text{pr.3.2}) = (\theta e^t + 1 - \theta)^n = (1 + \theta(e^t - 1))^n;$$

specijalno, f.i.m. Bernoullijeve distribucije dobije se za $n = 1$;
(b) za X negativno binomnu (k, θ):

$$M_X(t) = G_X(e^t) = (\text{pr.3.3}) = \left(\frac{\theta e^t}{1 - e^t(1-\theta)}\right)^k;$$

(c) za $X \sim P(\lambda)$:

$$M_X(t) = G_X(e^t) = (\text{pr.3.4}) = e^{\lambda(e^t - 1)}.$$

□

Primjer 3.6 Za $X \sim \Gamma(\alpha, \frac{1}{\lambda})$, f.i.m. je:

$$\begin{aligned} M_X(t) &= \mathbb{E}[e^{tX}] = \int_0^{+\infty} e^{tx} \cdot \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}\, dx = \frac{\lambda^\alpha}{(\lambda - t)^\alpha} \int_0^{+\infty} \frac{(\lambda - t)^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-(\lambda - t)x}\, dx = \\ &= \left(\frac{\lambda}{\lambda - t}\right)^\alpha \text{ za } t < \lambda. \end{aligned}$$

Odavde je

$$\begin{aligned} M_X'(t) &= \alpha\lambda^\alpha(\lambda - t)^{-(\alpha+1)} \Rightarrow \mathbb{E}[X] = M_X'(0) = \frac{\alpha}{\lambda} \\ M_X''(t) &= \alpha(\alpha+1)\lambda^\alpha(\lambda - t)^{-(\alpha+2)} \Rightarrow \mathbb{E}[X^2] = M_X''(0) = \frac{\alpha(\alpha+1)}{\lambda^2} \\ \Rightarrow \mathbb{V}\text{ar}[X] &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{\alpha(\alpha+1)}{\lambda^2} - \frac{\alpha^2}{\lambda^2} = \frac{\alpha}{\lambda^2}. \end{aligned}$$

Specijalno, za $X \sim Exp(\lambda)$,

$$\mathbf{M}_X(t) = \frac{\lambda}{\lambda - t}, \text{ za } t < \lambda.$$

Ako je $\theta = \mathbb{E}[X] = 1/\lambda$,

$$\mathbf{M}_X(t) = \frac{1}{1 - \theta t}, \text{ za } t < \frac{1}{\theta}.$$

Specijalno, za $X \sim \chi^2(n)$,

$$\mathbf{M}_X(t) = \frac{1}{(1-2t)^{\frac{n}{2}}}, \text{ za } t < \frac{1}{2}.$$

□

Primjer 3.7 Neka je $X \sim N(\mu, \sigma^2)$. Bez izvoda navodimo da je f.i.m. od X:

$$M_X(t) = e^{\mu t + \frac{\sigma^2 t^2}{2}}.$$

Odavde je

$$\begin{aligned} M_X'(t) &= (\mu + \sigma^2 t) M_X(t) \;\Rightarrow\; \mathbb{E}[X] = M_X'(0) = \mu \\ M_X''(t)) &= (\sigma^2 + (\mu + \sigma^2 t)^2) M_X(t) \;\Rightarrow\; \mathbb{E}[X^2] = M_X''(0) = \sigma^2 + \mu^2 \\ \Rightarrow\; \mathbb{V}\mathrm{ar}[X] &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \sigma^2 + \mu^2 - \mu^2 = \sigma^2. \end{aligned}$$

Specijalno je za standardiziranu verziju $Z = (X - \mu)/\sigma$ od X, $M_Z(t) = e^{t^2/2}$. Razvojem te funkcije u Taylorov red oko $t = 0$

$$M_X(t) = 1 + \frac{t^2}{2} + \frac{t^4}{8} + \cdots,$$

zaključujemo da je

$$\mathbb{E}[Z^{2k+1}] = 0, \quad \mathbb{E}[Z^{2k}] = \frac{(2k)!}{2^k k!}, \quad \text{za } \; k = 0, 1, 2, \ldots$$

Specijalno je

$$\mathbb{E}[Z] = 0,\ \mathbb{E}[Z^2] = 1,\ \mathbb{E}[Z^3] = 0,\ \mathbb{E}[Z^4] = 3.$$

Odavde slijedi da su prva tri momenta od X:

$$\mathbb{E}[X] = \mathbb{E}[\mu + \sigma Z] = \mu,\ \mathbb{E}[X^2] = \mathbb{E}[(\mu + \sigma Z)^2] = \mu^2 + \sigma^2,\ \mathbb{E}[X^3] = \mathbb{E}[(\mu + \sigma Z)^3] = \mu^3 + 3\sigma^2\mu,$$

a treći i četvrti centralni momenti:

$$\mathbb{E}[(X - \mu)^3] = \mathbb{E}[(\sigma Z)^3] = 0, \;\; \mathbb{E}[(X - \mu)^4] = \mathbb{E}[(\sigma Z)^4] = 3\sigma^4.$$

□

3.4 Funkcije izvodnice kumulanata

Za računanje matematičkog očekivanja i varijance slučajne varijable, funkcija izvodnica kumulanata je prirodnija od f.i.m. *Funkcija izvodnica kumulananata* (kraće f.i.k.) od X je funkcija C_X definirana sa

$$C_X(t) = \log M_X(t)$$

ako $M_X(t)$ postoji. Jasno je da vrijedi

$$M_X(t) = e^{C_X(t)}.$$

Nadalje,

$$\begin{aligned} C_X'(t) &= \frac{M_X'(t)}{M_X(t)} \\ C_X''(t) &= \frac{M_X''(t) M_X(t) - M_X'(t)^2}{M_X^2(t)} \end{aligned}$$

Budući da je $M_X(0) = 1$, $M_X'(0) = \mathbb{E}[X]$ i $\mathbf{M}_X''(0) = \mathbb{E}[X^2]$,

$$\begin{aligned} C_X'(0) &= \frac{M_X'(0)}{M_X(0)} = \frac{\mathbb{E}[X]}{1} = \mathbb{E}[X] \\ C_X''(0) &= \frac{M_X''(0) M_X(0) - M_X'(0)^2}{M_X^2(0)} = \frac{\mathbb{E}[X^2] \cdot 1 - \mathbb{E}[X]^2}{1} = \mathbb{V}\mathrm{ar}[X]. \end{aligned}$$

Koeficijent uz $t^r/r!$ u razvoju funkcije $t \mapsto C_{X - \mathbb{E}[X]}(t)$ u Taylorov red oko $t = 0$ (tzv. Maclaurinov red) zove se *r-ti kumulant* od X i označava se sa κ_r.

3.5 Funkcije izvodnice linearnih funkcija od X

Pretpostavimo da (brojeća) slučajna varijabla X ima f.i.v. $G_X(t)$. Tada je za $Y = aX + b$ (a, b realni brojevi, $a \neq 0$),

$$G_Y(t) = \mathbb{E}[t^Y] = \mathbb{E}[t^{aX+b}] = t^b \mathbb{E}[(t^a)^X] = t^b G_X(t^a).$$

Ako je $M_X(t)$ f.i.m. od X, tada je

$$M_Y(t) = \mathbb{E}[e^{tY}] = \mathbb{E}[e^{taX+tb}] = e^{tb} \mathbb{E}[e^{(ta)X}] = e^{tb} M_X(ta).$$

Poglavlje 4

Zajednička razdioba slučajnih varijabli

Ako imamo više slučajnih varijabli definiranih na istom vjerojatnosnom prostoru, tada možemo govoriti o njihovoj zajedničkoj distribuciji. Na te varijable možemo gledati kao na komponente nekog slučajnog vektora. Razdiobe slučajnih vektora zovemo *višedimenzionalnim razdiobama* (distribucijama), a to su, u stvari, njihove *zajedničke razdiobe* (distribucije). Zajedničku razdiobu dviju slučajnih varijabli zovemo *bivarijatnom* razdiobom.

4.1 Zajednička gustoća i funkcija distribucije

Neka su X i Y dvije slučajne varijable definirane na istom vjerojatnosnom prostoru.

Pretpostavimo da su X, Y diskretne slučajne varijable (tj. slučajni vektor (X, Y) je diskretan), te da je

$$\mathrm{Im}X = \{a_1, a_2, \ldots\}, \quad \mathrm{Im}Y = \{b_1, b_2, \ldots\}.$$

Tada je

$$\mathrm{Im}(X, Y) = \{(a_1, b_1), (a_1, b_2), \ldots, (a_2, b_1), \ldots\} = \{(a_i, b_j) : a_i \in \mathrm{Im}X, b_j \in \mathrm{Im}Y\}.$$

U tablici zajedničke razdiobe od X, Y:

			Y		
X	b_1	b_2	$\cdots$	b_j	$\cdots$
a_1	p_{11}	p_{12}	$\cdots$	p_{1j}	$\cdots$
a_2	p_{21}	p_{22}	$\cdots$	p_{2j}	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	
a_i	p_{i1}	p_{i2}	$\cdots$	p_{ij}	$\cdots$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\ddots$

broj p_{ij} označava vjerojatnost događaja da je istovremeno $X = a_i$ i $Y = b_j$:

$$p_{ij} = \mathbb{P}(X = a_i, Y = b_j) \text{ za sve } i, j.$$

Zajednička funkcija vjerojatnosti diskretnih slučajnih varijabli X, Y (ili *gustoća diskretnog slučajnog vektora* (X, Y) ili, jednostavno, *zajednička gustoća* tih varijabli) je funkcija $f_{X,Y} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ definirana sa

$$f_{X,Y}(x, y) := \mathbb{P}(X = x, Y = y) = \begin{cases} p_{ij} & \text{za } x = a_i, y = b_j \\ 0 & \text{inače.} \end{cases}$$

Svojstva te funkcije:

$(G1)$ $f_{X,Y}(x,y) \geq 0$ za sve x, y

$(G2)$ $\sum_{x \in \mathrm{Im}X, y \in \mathrm{Im}Y} f_{X,Y}(x,y) = 1.$

Zajednička funkcija distribucije od X i Y (ili *funkcija distribucije slučajnog vektora* (X,Y)) definira se kao funkcija $F_{X,Y} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$,

$$F_{X,Y}(x,y) := \mathbb{P}(X \leq x, Y \leq y).$$

Ako je (X,Y) diskretan slučajni vektor s funkcijom vjerojatnosti $f_{X,Y}$, tada vrijedi:

$$F_{X,Y}(x,y) = \sum_{\{a \in \mathrm{Im}X : a \leq x\}} \sum_{\{b \in \mathrm{Im}X : b \leq y\}} f_{X,Y}(a,b) \text{ za sve realne } x, y.$$

Primjer 4.1 Bacamo dvije simetrične igraće kocke: crvenu i plavu. Naka je X broj koji se okrenuo na crvenoj kocki, a Y manji od brojeva koji su se okrenuli na obje kocke. Tada je zajednička razdioba od X i Y prikazana tablicom:

	Y						
X	1	2	3	4	5	6	Σ
1	$\frac{6}{36}$	0	0	0	0	0	$\frac{1}{6}$
2	$\frac{1}{36}$	$\frac{5}{36}$	0	0	0	0	$\frac{1}{6}$
3	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{4}{36}$	0	0	0	$\frac{1}{6}$
4	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{3}{36}$	0	0	$\frac{1}{6}$
5	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{2}{36}$	0	$\frac{1}{6}$
6	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
Σ	$\frac{11}{36}$	$\frac{9}{36}$	$\frac{7}{36}$	$\frac{5}{36}$	$\frac{3}{36}$	$\frac{1}{36}$	1

□

Zajednička razdioba dviju *neprekidnih* slučajnih varijabli X i Y opisana je *zajedničkom gustoćom* $f_{X,Y} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ za koju vrijedi da je

$$\mathbb{P}(a \leq X \leq b, c \leq Y \leq d) = \int_a^b dx \int_c^d dy \, f_{X,Y}(x,y) \text{ za sve } a < b, c < d.$$

$f_{X,Y}$ još zovemo *gustoćom* (ili *funkcijom gustoće*) *neprekidnog slučajnog vektora* (X,Y). Za njihovu zajedničku funkciju distribucije vrijedi:

$$F_{X,Y}(x,y) = \int_{-\infty}^x du \int_{-\infty}^y dv \, f_{X,Y}(u,v) \text{ za sve realne } x, y,$$

odnosno

$$f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}}{\partial x \partial y}(x,y),$$

za točke (x,y) za koje navedena parcijalna derivacija postoji.

Svojstva zajedničke gustoće neprekidnih varijabli:

$(G1)$ $f_{X,Y}(x,y) \geq 0$ za sve x, y

$(G2)$ $\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f_{X,Y}(x,y)\, dx\, dy = 1.$

4.2 Marginalne gustoće

Za diskretan slučajni vektor (X, Y) s gustoćom razdiobe $f_{X,Y}$, *marginalna gustoća od* X je funkcija jedna varijable dana izrazom

$$f_X(x) = \sum_{y \in \mathrm{Im}Y} f_{X,Y}(x, y), \quad x \in \mathbb{R}.$$

To je ujedno i gustoća vjerojatnosne razdiobe diskretne slučajne varijable X. Analogno, *marginalna gustoća od* Y je funkcija jedne varijable

$$f_Y(y) = \sum_{x \in \mathrm{Im}X} f_{X,Y}(x, y), \quad y \in \mathbb{R}$$

i gustoća je razdiobe diskretne slučajne varijable Y.

Primjer 4.2 Iz tablice zajedničke razdiobe slučajnih varijabli X, Y iz primjera 4.1, možemo očitati (marginalne) razdiobe od X i od Y. Naime, u osmom stupcu se nalaze vjerojatnosti vrijednosti od X iz prvog stupca, a u devetom retku se nalaze vjerojatnosti vrijednosti od Y koje se nalaze u drugom retku. □

Marginalna gustoća od X *neprekidnog* slučajnog vektora (X, Y) sa gustoćom razdiobe $f_{X,Y}$ je funkcija jedne varijable dana izrazom

$$f_X(x) = \int_{-\infty}^{+\infty} f_{X,Y}(x, y)\, dy, \quad x \in \mathbb{R}.$$

To je ujedno gustoća vjerojatnosne razdiobe neprekidne slučajne varijable X. Analogno, marginalna gustoća od Y je funkcija

$$f_Y(y) = \int_{-\infty}^{+\infty} f_{X,Y}(x, y)\, dx, \quad y \in \mathbb{R}$$

i gustoća je razdiobe neprekidne slučajne varijable Y.

4.3 Uvjetna razdioba

Neka je (X, Y) slučajni vektor. Distribuciju slučajne varijable Y za danu vrijednost x slučajne varijable X zovemo *uvjetnom distribucijom od* Y *za dano* $X = x$ (ili *uz uvjet* $X = x$). Analogno određujemo *uvjetnu distribuciju od* X *za dano* (ili *uz uvjet*) $Y = y$. Uvjetne distribucije zadaju se *uvjetnim gustoćama*.

Neka je (X, Y) diskretan slučajni vektor sa zajedničkom funkcijom vjerojatnosti $f_{X,Y}$ i marginalnim funkcijama vjerojatnosti f_X, f_Y. Ako je $f_Y(y) = \mathbb{P}(Y = y) > 0$ za $y \in \mathrm{Im}Y$, tada je *uvjetna funkcija vjerojatnosti* (ili *uvjetna gustoća*) *od* X *za dano* $Y = y$ funkcija $x \mapsto f_{X|Y}(x|y)$ dana sa

$$f_{X|Y}(x|y) := \mathbb{P}(X = x|Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)} = \frac{f_{X,Y}(x, y)}{f_Y(y)}, \quad x \in \mathbb{R},$$

i to je gustoća diskretne vjerojatnosne razdiobe koja odgovara uvjetnoj razdiobi od X uz dano $Y = y$. Analogno se definira *uvjetna funkcija vjerojatnosti* (ili *uvjetna gustoća*) *od* Y *uz dano* $X = x$.

Primjer 4.3 Za slučajni vektor (X, Y) iz primjera 4.1, uvjetna gustoća od X uz dano $Y = 3$ je dana tablicom:

x	1	2	3	4	5	6
$f_{X\|Y}(x\|3)$	0	0	$\frac{4}{7}$	$\frac{1}{7}$	$\frac{1}{7}$	$\frac{1}{7}$

Formula za istu uvjetnu gustoću:

$$f_{X|Y}(x|3) = \frac{f_{X,Y}(x,3)}{f_Y(3)} = \frac{f_{X,Y}(x,3)}{\frac{7}{36}} = \frac{36 f_{X,Y}(x,3)}{7}, \quad x = 1,2,3,4,5,6.$$

Na primjer,

$$f_{X|Y}(4|3) = \frac{36 f_{X,Y}(x,3)}{7} = \frac{36 \cdot \frac{1}{36}}{7} = \frac{1}{7}.$$

□

Ako je (X, Y) neprekidan slučajni vektor sa zajedničkom gustoćom $f_{X,Y}$ i marginalnim gustoćama f_X, f_Y, te ako je $f_Y(y) > 0$ za $y \in \mathrm{Im}Y$, tada je *uvjetna gustoća* od X za dano $Y = y$ funkcija $x \mapsto f_{X|Y}(x|y)$ dana sa

$$f_{X|Y}(x|y) := \frac{f_{X,Y}(x,y)}{f_Y(y)}, \quad x \in \mathbb{R}.$$

To je ujedno funkcija gustoće neprekidne vjerojatnosne razdiobe koja odgovara uvjetnoj distribuciji od X uz dano $Y = y$. Naime, neka je sa $\mathbb{P}(a \le X \le b|Y = y)$ označena uvjetna vjerojatnost da će X poprimiti vrijednosti u intervalu $[a, b]$ uz dano $Y = y$. Tada je

$$\mathbb{P}(a \le X \le b|Y = y) := \int_a^b f_{X|Y}(x|y)\, dx.$$

Primijetite da se lijeva strane gornje jednakosti *ne* računa po formuli za uvjetnu vjerojatnost danu formulom (2.2) (jer je to nemoguće zbog $\mathbb{P}(Y = y) = 0$) iako ima istu interpretaciju. Analogno se definira *uvjetna gustoća* od Y uz dano $X = x$.

4.4 Nezavisnost slučajnih varijabli

Neka je (X, Y) slučajni vektor sa zajedničkom gustoćom $f_{X,Y}$ i marginalnim gustoćama f_X, f_Y. Da uvjetna razdioba od Y za dano $X = x$ ne ovisi o x za sve x, znači da je uvjetna razdioba jednaka marginalnoj, dakle, da vrijedi

$$f_{Y|X}(y|x) = f_Y(y) \ \text{ za sve } \ y \in \mathrm{Im}Y \ \text{ i sve } \ x \in \mathrm{Im}X \ \text{ za koje je } \ f_X(x) > 0. \tag{4.1}$$

Odavde odmah slijedi, po definiciji uvjetne gustoće, da je

$$f_{X,Y}(x,y) = f_X(x) \cdot f_y(y) \ \text{ za sve } \ y \in \mathrm{Im}Y, \ x \in \mathrm{Im}X. \tag{4.2}$$

Opet, po definiciji uvjetne gustoće, ako vrijedi (4.2), tada ni uvjetna razdioba od Y uz dano $Y = y$ ne ovisi o y za sve y. Po definiciji kažemo da su slučajne varijable X i Y *nezavisne* ako vrijedi (4.2).

U slučaju diskretnih slučajnih varijabli X i Y, (4.2) je ekvivalentno sa

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x) \cdot \mathbb{P}(Y = y) \ \text{ za sve } \ x, y,$$

odnosno, uz oznake kao u 4.1,

$$p_{ij} = \mathbb{P}(X = a_i) \cdot \mathbb{P}(Y = b_j) \ \text{ za sve } \ i, j.$$

Dakle, p_{ij} se dobije kao produkt pripadnih vjerojatnosti na margini.

Primjer 4.4 Slučajne varijable X, Y iz primjera 4.1 nisu nezavisne jer je, na primjer,

$$p_{11} = \frac{6}{36} \neq \frac{11}{36} \cdot \frac{1}{6} = \mathbb{P}(X = 1) \cdot \mathbb{P}(Y = 1).$$

S druge strane, za isti slučajni pokus, slučajne varijable X i Z, gdje je Z broj koji se okrenuo na plavoj kocki, su nezavisne. □

Ako su X i Y neprekidne slučajne varijable, tada je uvjet (4.2) za njihovu nezavisnost ekvivalentan sa

$$\mathbb{P}(a \leq X \leq b, c \leq Y \leq d) = \mathbb{P}(a \leq X \leq b) \cdot \mathbb{P}(c \leq Y \leq d) \text{ za sve } a < b, c < d.$$

I u diskretnom, i u neprekidnom slučaju, (4.2) je ekvivalentno sa

$$F_{X,Y}(x, y) = F_X(x) \cdot F_Y(y) \text{ za sve } x, y,$$

gdje je $F_{X,Y}$ funkcija distribucije zajedničke razdiobe od X, Y i F_X, F_Y su marginalne funkcije distribucije.

Ako su slučajne varijable X i Y nezavisne, tada su i $g(X)$, $h(Y)$ nezavisne slučajne varijable za sve funkcije g i h.

Općenito, kažemo da su slučajne varijable X_1, X_2,... *nezavisne* ako za svaki izbor broja k ($k \geq 2$) i svaki izbor k-člane kombinacije X_{i_1}, X_{i_2},..., X_{i_k} tog niza varijabli vrijedi da je

$$f_{X_{i_1},\ldots,X_{i_k}}(x_1, \ldots, x_k) = f_{X_{i_1}}(x_1) \cdots f_{X_{i_k}}(x_k) \text{ za sve } x_1, \ldots, x_k.$$

4.5 Matematičko očekivanje funkcije dviju slučajnih varijabli

Neka je (X, Y) slučajni vektor sa zajedničkom gustoćom $f_{X,Y}$ i neka je g funkcija u dvije varijable takva da je $g(X, Y) = g \circ (X, Y)$ slučajna varijabla. Ako je (X, Y) diskretan slučajni vektor, tada vrijedi

$$\mathbb{E}[g(X, Y)] = \sum_{x \in \mathrm{Im}X} \sum_{y \in \mathrm{Im}Y} g(x, y) f_{X,Y}(x, y) = \sum_{i,j} g(a_i, b_j) p_{ij},$$

a ako je (X, Y) neprekidan slučajni vektor, tada je

$$\mathbb{E}[g(X, Y)] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g(x, y) f_{X,Y}(x, y)\, dx\, dy.$$

Pomoću tih relacija može se dokazati da matematičko očekivanje ima svojstvo *linearnosti*, općenitije od navedenog u 2.4. Neka su X i Y dvije slučajne varijable definirane na istom vjerojatnosnom prostoru, neka su g, h dvije funkcije jedne varijable takve da su $g(X)$ i $h(Y)$ slučajne varijable koje imaju matematičko očekivanje i neka su α, β dva broja. Tada slučajna varijabla $\alpha g(X) + \beta h(Y)$ ima matematičko očekivanje i vrijedi

$$\mathbb{E}[\alpha g(X) + \beta h(Y)] = \alpha \mathbb{E}[g(X)] + \beta \mathbb{E}[h(Y)].$$

Dakle, da bi izračunali matematičko očekivanje varijable $\alpha g(X) + \beta h(Y)$, dovoljno je znati marginalne razdiobe od X i Y. Taj rezultat se može proširiti na bilo koju linearnu kombinaciju bilo kojeg broja funkcija slučajnih varijabli.

Za nezavisne slučajne varijable X, Y takve da postoje matematička očekivanja od $g(X)$ i $h(Y)$ vrijedi da postoji matematičko očekivanje slučajne varijable $g(X) \cdot h(Y)$ i da je

$$\mathbb{E}[g(X) \cdot h(Y)] = \mathbb{E}[g(X)] \cdot \mathbb{E}[h(Y)]. \tag{4.3}$$

Također se i taj rezultat može poopćiti na produkt bilo kojeg broja funkcija nezavisnih slučajnih varijabli.

4.6 Kovarijanca i koeficijent korelacije

Kovarijanca dviju slučajnih varijabli X, Y definira se kao broj

$$\mathrm{cov}[X,Y] := \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] \tag{4.4}$$

ako desna strana postoji. (4.4) je ekvivalentno sa

$$\mathrm{cov}[X,Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]. \tag{4.5}$$

Fizikalna jedinica kovarijance jednaka je produktu fizikalnih jedinica od X i Y. Primijetite da je $\mathrm{cov}[X,X] = \mathbb{V}\mathrm{ar}[X]$.

Primjer 4.5 Za slučajne varijable X, Y iz primjera 4.1 je $\mathbb{E}[X] = 7/2$, $\mathbb{E}[Y] = 91/36$ i

$$\mathbb{E}[XY] = \sum_{i=1}^{6}\sum_{j=1}^{6} ijp_{ij} = \sum_{i=1}^{6}\frac{i^2(7-i)}{36} + \sum_{i=1}^{6}\sum_{j=i+1}^{6}\frac{ij}{36} = \frac{371}{36},$$

pa je

$$\mathrm{cov}[X,Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \frac{371}{36} - \frac{7}{2}\cdot\frac{91}{36} = \frac{35}{24} = 1.45833.$$

□

Navedimo neka svojstva kovarijance. Prvo, vrijedi

$$\mathrm{cov}[aX+b, cY+d] = ac\mathrm{cov}[X,Y] \tag{4.6}$$

za sve brojeve a, b, c, d i sve slučajne varijable X,Y za koje navedene kovarijance postoje.

Dokaz. $\mathbb{E}[aX+b] = a\mathbb{E}[X]+b$ i $\mathbb{E}[cY+d] = c\mathbb{E}[Y]+d$ pa je $aX+b-\mathbb{E}[aX+b] = a(X-\mathbb{E}[X])$ i $cY+d-\mathbb{E}[cY+d] = c(Y-\mathbb{E}[Y])$, dakle,

$$\mathrm{cov}[aX+b, cY+d] = \mathbb{E}[a(X-\mathbb{E}[X])\cdot c(Y-\mathbb{E}[Y])] = ac\mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] = ac\mathrm{cov}[X,Y].$$

□

Nadalje, za sve slučajne varijable X, Y, Z za koje navedene kovarijance postoje je

$$\mathrm{cov}[X, Y+Z] = \mathrm{cov}[X,Y] + \mathrm{cov}[X,Z]. \tag{4.7}$$

Dokaz.

$$\begin{aligned}
\mathbb{E}[X(Y+Z)] &= \mathbb{E}[XY] + \mathbb{E}[XZ] \ \text{(linearnost)}\\
\mathbb{E}[Y+Z] &= \mathbb{E}[Y] + \mathbb{E}[Z] \ \text{(linearnost)}\\
\Rightarrow\ \mathrm{cov}[X,Y+Z] &\overset{(4.5)}{=} \mathbb{E}[X(Y+Z)] - \mathbb{E}[X]\mathbb{E}[Y+Z] =\\
&= \mathbb{E}[XY] + \mathbb{E}[XZ] - \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Z] =\\
&\overset{(4.5)}{=} \mathrm{cov}[X,Y] + \mathrm{cov}[X,Z].
\end{aligned}$$

□

Ako su X i Y nezavisne slučajne varijable, tada je nužno $\mathrm{cov}[X,Y] = 0$. Ta činjenica slijedi iz (4.5) i svojstva matematičkog očekivanja produkta nezavisnih varijabli (4.3).

Kovarijanca dviju slučajnih varijabli mjeri stupanj njihove *linearne povezanosti*. Istu stvar mjeri *koeficijent korelacije* koji se definira kao kovarijanca njihovih standardiziranih verzija. Prema tome, za razliku od kovarijance, koeficijent korelacije je bezdimenzionalna

mjera stupnja njihove linearne povezanosti. Preciznije, koeficijent korelacije varijabli X, Y je broj

$$\rho = \mathrm{corr}[X,Y] := \mathbb{E}[\frac{X-\mathbb{E}[X]}{\sigma(X)} \cdot \frac{Y-\mathbb{E}[Y]}{\sigma(Y)}] \overset{(4.6)}{=} \frac{\mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])]}{\sigma(X)\cdot\sigma(Y)} \overset{(4.4)}{=} \frac{\mathrm{cov}[X,Y]}{\sigma(X)\cdot\sigma(Y)}.$$

Za koeficijent korelacije vrijedi da je

$$-1 \leq \rho \leq 1.$$

Ako je $\rho = \pm 1$, tada s vjerojatnosti jedan slijedi da su X i Y u linearnoj vezi. Ako je $\rho = 0$, tada kažemo da su X, Y *nekorelirane* slučajne varijable. Dakle, nezavisne slučajne varijable su nekorelirane. Obrat ne vrijedi općenito.

4.7 Varijanca zbroja slučajnih varijabli

Za svake dvije slučajne varijable X i Y koje imaju konačnu varijancu, je

$$\mathbb{V}\mathrm{ar}[X+Y] = \mathbb{V}\mathrm{ar}[X] + \mathbb{V}\mathrm{ar}[Y] + 2\mathrm{cov}[X,Y]. \tag{4.8}$$

Ako su X i Y nezavisne slučajne varijable, tada je

$$\mathbb{V}\mathrm{ar}[X+Y] = \mathbb{V}\mathrm{ar}[X] + \mathbb{V}\mathrm{ar}[Y]. \tag{4.9}$$

Matematičkom indukcijom se može pokazati da za niz slučajnih varijabli X_1, X_2,..., X_n koje imaju konačnu varijancu, vrijedi

$$\mathbb{V}\mathrm{ar}[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} \mathbb{V}\mathrm{ar}[X_i] + 2 \sum_{1\leq i<j\leq n} \mathrm{cov}[X_i, X_j]. \tag{4.10}$$

Ako su te varijable nezavisne, tada je

$$\mathbb{V}\mathrm{ar}[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} \mathbb{V}\mathrm{ar}[X_i]. \tag{4.11}$$

Primjer 4.6 Binomna slučajna varijabla X s parametrima (n,θ) se interpretira kao ukupan broj uspjeha u nizu od n nezavisnih jednako distribuiranih Bernoullijevih pokusa s vjerojatnosti uspjeha θ. Dakle, X se može prikazati kao zbroj od n nezavisnih jednako distribuiranih Bernoullijevih slučajnih varijabli X_1, X_2,..., X_n takvih da X_i indicira uspjeh u i-tom pokusu ($i = 1, 2, \ldots, n$). Budući da je $\mathbb{E}[X_i] = \theta$ i $\mathbb{V}\mathrm{ar}[X_i] = \theta(1-\theta)$ za sve i, zbog linearnosti matematičkog očekivanja je

$$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X_i] = (\text{jednaka distribuiranost}) = n\theta,$$

a zbog nezavisnosti je

$$\mathbb{V}\mathrm{ar}[X] \overset{(4.11)}{=} \sum_{i=1}^{n} \mathbb{V}\mathrm{ar}[X_i] = (\text{jednaka distribuiranost}) = n\theta(1-\theta).$$

□

4.8 Konvolucije

Neka su X i Y slučajne varijable sa zajedničkom gustoćom $f_{X,Y}$ i $Z = X + Y$ njihov zbroj što je također slučajna varijabla. Ako je vektor (X, Y) diskretan, tada je Z diskretna slučajna varijabla s funkcijom gustoće

$$\begin{aligned} f_Z(z) &= \mathbb{P}(Z = z) = \mathbb{P}\Big(\bigcup_{x\in \mathrm{Im}X} \{X = x, Y = z - x\}\Big) \overset{(P3)}{=} \sum_{x\in \mathrm{Im}X} \mathbb{P}(X = x, Y = z - x) = \\ &= \sum_{x\in \mathrm{Im}X} f_{X,Y}(x, z - x). \end{aligned}$$

Ako su X, Y nezavisne slučajne varijable, tada je

$$f_Z(z) = \sum_{x\in \mathrm{Im}X} f_X(x) f_Y(z - x), \quad z \in \mathbb{R}. \tag{4.12}$$

Funkciju f_Z izraženu u formi (4.12) zovemo *konvolucijom* funkcija f_X i f_Y, i pišemo

$$f_Z = f_X * f_Y.$$

Dakle, gustoća zbroja nezavisnih slučajnih varijabli jednaka je konvoluciji njihovih gustoća. U slučaju da su X, Y neprekidne slučajne varijable, gustoća njihovog zbroja $Z = X + Y$ je

$$f_Z(z) = \int_{-\infty}^{+\infty} f_{X,Y}(x, z - x)\, dx, \quad z \in \mathbb{R}.$$

Ako su X i Y nezavisne, tada je

$$f_Z(z) = \int_{-\infty}^{+\infty} f_X(x) f_Y(z - x)\, dx, \quad z \in \mathbb{R}. \tag{4.13}$$

Ta funkcija je također *konvolucija* funkcija f_X i f_Y.

Konvolucija više funkcija f_1, f_2,..., f_n definira se rekurzivno:

$$f_1 * f_2 * \cdots * f_n := (f_1 * f_2 * \cdots * f_{n-1}) * f_n.$$

Može se pokazati da je gustoća zbroja n nezavisnih slučajnih varijabli jednaka konvoluciji njihovih gustoća. Ako su pri tome svih n varijabli jednako distribuirane s gustoćom f, tada se n-terostruka konvolucija tih funkcija označava sa f^{n*}. Dakle, f^{n*} je gustoća zbroja od n nezavisnih jednako distribuiranih (kraće n.j.d.) slučajnih varijabli sa gustoćom f. Funkciju distribucije zbroja od n n.j.d. slučajnih varijabli X_1, X_2,..., X_n sa funkcijom distribucije F označavamo sa F^{n*}:

$$F^{n*}(x) := \mathbb{P}(X_1 + X_2 + \cdots + X_n \leq x), \quad x \in \mathbb{R}.$$

4.9 Razdiobe linearnih kombinacija nezavisnih slučajnih varijabli pomoću funkcija izvodnica

Neka su X_1, X_2,..., X_n nezavisne slučajne varijable i α_1, α_2,..., α_n realni brojevi. Za njihovu linearnu kombinaciju

$$Y = \alpha_1 X_1 + \alpha_2 X_2 + \cdots + \alpha_n X_n,$$

zbog linearnosti matematičkog očekivanja, nezavisnosti i svojstava varijance (2.7) i (4.11), vrijedi da je

$$\mathbb{E}[Y] = \sum_{i=1}^{n} \alpha_i \mathbb{E}[X_i] \tag{4.14}$$

$$\mathbb{V}\mathrm{ar}[Y] = \sum_{i=1}^{n} \alpha_i^2 \mathbb{V}\mathrm{ar}[X_i]. \tag{4.15}$$

Distribucija slučajne varijable Y može se dobiti pomoću konvolucija, ali jednostavnije je koristiti f.i.v. ili f.i.m. budući da postoji jednoznačna veza između njih i razdioba.

Neka su X, Y nezavisne brojeće slučajne varijable sa f.i.v. $G_X(t)$ i $G_Y(t)$, i neka je $S = \alpha X + \beta Y$ jedna njihova linearna kombinacija sa koeficijentima α, β. Tada je

$$G_S(t) = \mathbb{E}[t^S] = \mathbb{E}[t^{\alpha X+\beta Y}] = \mathbb{E}[t^{\alpha X} \cdot t^{\beta Y}] \overset{(4.3)}{=} \mathbb{E}[t^{\alpha X}] \cdot \mathbb{E}[t^{\beta Y}] = G_X(t^\alpha) \cdot G_Y(t^\beta).$$

U slučaju $\alpha = \beta = 1$,

$$G_{X+Y}(t) = G_X(t) \cdot G_Y(t). \tag{4.16}$$

Taj rezultat se može poopćiti na zbroj $Y = X_1 + X_2 + \cdots + X_n$ n nezavisnih slučajnih varijabli X_1, X_2,..., X_n:

$$G_Y(t) = G_{X_1}(t) \cdot G_{X_2}(t) \cdots G_{X_n}(t). \tag{4.17}$$

U slučaju da su te varijable i jednako distribuirane, vrijedi:

$$G_Y(t) = (G_{X_1}(t))^n. \tag{4.18}$$

Primjer 4.7 Neka su X_1, X_2,..., X_n n.j.d. Bernoullijeve slučajne varijable s parametrom θ. Tada je $G_{X_i}(t) = \theta t + 1 - \theta$. Za $Y := X_1 + X_2 + \cdots + X_n$ je

$$G_Y(t) = (G_{X_1}(t))^n = (\theta t + 1 - \theta)^n$$

što je f.i.v. binomne razdiobe (n, θ). Dakle, svaka se binomna (n, θ) slučajna varijabla može reprezentirati kao zbroj od n n.j.d. Bernoullijevih varijabli s parametrom θ. Nadalje, neka su $X \sim$ binomna (m, θ) i $Y \sim$ binomna (n, θ) nezavisne slučajne varijable. Tada je

$$G_{X+Y}(t) = G_X(t) \cdot G_Y(t) = (\theta t + 1 - \theta)^m \cdot (\theta t + 1 - \theta)^n = (\theta t + 1 - \theta)^{m+n},$$

dakle, zbroj dviju nezavisnih binomnih slučajnih varijabli s istim parametrom θ je opet binomna slučajna varijabla. □

Primjer 4.8 Neka je X_1, X_2,..., X_k niz n.j.d. geometrijskih slučajnih varijabli s parametrom uspjeha θ. Tada je

$$G_{X_i}(t) = \frac{\theta t}{1 - (1 - \theta)t}, \quad i = 1, 2, \ldots, k.$$

Odavde slijedi da je f.i.v. za $Y := X_1 + X_2 + \cdots + X_k$ jednaka

$$G_Y(t) = \left(\frac{\theta t}{1 - (1 - \theta)t}\right)^k$$

što je f.i.v. negativne binomne razdiobe s parametrima (k, θ). Nadalje, budući da geometrijska razdioba ima očekivanje $1/\theta$ i varijancu $(1-\theta)/\theta^2$, negativna (k, θ)-binomna razdioba ima očekivanje k/θ i varijancu $k(1 - \theta)/\theta^2$. Još više, na isti način možemo zaključiti da je zbroj dviju nezavisnih negativnih binomnih slučajnih varijabli s parametrima (k, θ) i (m, θ) opet negativna binomna razdioba s parametrima $(k + m, \theta)$. □

Primjer 4.9 Neka su $X \sim P(\lambda)$ i $Y \sim P(\mu)$ nezavisne Poissonove slučajne varijable. Tada za njihov zbroj $Z = X + Y$ vrijedi:

$$G_Z(t) = G_X(t) \cdot G_Y(t) = e^{-\lambda(t-1)} \cdot e^{-\mu(t-1)} = e^{-(\lambda+\mu)(t-1)}$$

što je f.i.v. Poissonove razdiobe $P(\lambda + \mu)$. Dakle, $Z \sim P(\lambda + \mu)$. □

Neka su sada X, Y dvije nezavisne slučajne varijable sa f.i.m. $M_X(t)$ i $M_Y(t)$, i neka je za zadane brojeve α, β, $S := \alpha X + \beta Y$. Tada je

$$M_S(t) = \mathbb{E}[e^{tS}] = \mathbb{E}[e^{t(\alpha X+\beta Y)}] = \mathbb{E}[e^{t\alpha X} \cdot e^{t\beta Y}] \overset{(4.3)}{=} \mathbb{E}[e^{t\alpha X}] \cdot \mathbb{E}[e^{t\beta Y}] = M_X(\alpha t) \cdot M_Y(\beta t).$$

U slučaju $\alpha = \beta = 1$, za nezavisne slučajne varijable X i Y vrijedi:

$$M_{X+Y}(t) = M_X(t) \cdot M_Y(t). \tag{4.19}$$

Slično, za $Y = X_1 + X_2 + \cdots + X_n$, gdje su X_1, X_2,..., X_n n nezavisnih slučajnih varijabli:

$$M_Y(t) = M_{X_1}(t) \cdot M_{X_2}(t) \cdots M_{X_n}(t). \tag{4.20}$$

U slučaju da su te varijable i jednako distribuirane, vrijedi:

$$M_Y(t) = (M_{X_1}(t))^n. \tag{4.21}$$

Primjer 4.10 Neka je X_1, X_2,..., X_k niz n.j.d. eksponencijalnih $Exp(\lambda)$-slučajnih varijabli. Tada je

$$M_{X_i}(t) = \frac{\lambda}{\lambda - t}, \quad \text{za } t < \lambda, \quad i = 1, 2, \ldots, k.$$

Odavde slijedi da je f.i.m. za $Y = X_1 + X_2 + \cdots + X_k$ jednaka

$$M_Y(t) = \left(\frac{\lambda}{\lambda - t}\right)^k$$

što je f.i.m. $\Gamma(k, 1/\lambda)$-razdiobe. Dakle, svaka se $\Gamma(k, 1/\lambda)$-razdioba može reprezentirati pomoću zbroja k n.j.d. $Exp(\lambda)$-slučajnih varijabli. Budući da je $\mathbb{E}[X_i] = 1/\lambda$ i $\mathbb{V}\mathrm{ar}[X_i] = 1/\lambda^2$, $\mathbb{E}[Y] = k/\lambda$ i $\mathbb{V}\mathrm{ar}[Y] = k/\lambda^2$. U interpretaciji, vrijeme do pojave k-tog događaja u Poissonovom procesu s intenzitetom λ je zbroj od k međuvremena pojavljivanja individualnih događaja.
Općenito, neka su $X \sim \Gamma(\alpha, 1/\lambda)$ i $Y \sim \Gamma(\beta, 1/\lambda)$ nezavisne slučajne varijable i $Z = X + Y$. Tada je

$$M_Z(t) = M_X(t) \cdot M_Y(t) = \left(\frac{\lambda}{\lambda - t}\right)^\alpha \cdot \left(\frac{\lambda}{\lambda - t}\right)^\beta = \left(\frac{\lambda}{\lambda - t}\right)^{\alpha+\beta},$$

dakle, $Z \sim \Gamma(\alpha + \beta, 1/\lambda)$. Odavde slijedi da ako su $X \sim \chi^2(n)$ i $Y \sim \chi^2(m)$ nezavisne slučajne varijable, da je tada $X + Y \sim \chi^2(n + m)$. □

Primjer 4.11 Neka su $X \sim N(\mu_X, \sigma_X^2)$ i $Y \sim N(\mu_Y, \sigma_Y^2)$ nezavisne normalne slučajne varijable i $Z = X + Y$ njihov zbroj. Tada je:

$$M_Z(t) = M_X(t) \cdot M_Y(t) = e^{\mu_X t + \frac{\sigma_X^2}{2} t^2} \cdot e^{\mu_Y t + \frac{\sigma_Y^2}{2} t^2} = e^{(\mu_X + \mu_Y) t + \frac{(\sigma_X^2 + \sigma_Y^2)}{2} t^2}$$

što je f.i.v. normalne razdiobe $N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$. Dakle, zbroj dviju nezavisnih normalno distribuiranih slučajnih varijabli je opet normalna slučajna varijabla $Z \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$. □

4.10 Uvjetno očekivanje

Neka je (X, Y) slučajni vektor. *Uvjetno očekivanje od Y uz dano $X = x$* je matematičko očekivanje uvjetne distribucije od Y uz dano $X = x$, u oznaci $\mathbb{E}[Y|X = x]$. Analogno se definira *uvjetno očekivanje od X uz dano $Y = y$*, $\mathbb{E}[X|Y = y]$. Dakle, ako je (X, Y) diskretan slučajni vektor, tada je

$$\mathbb{E}[Y|X = x] := \sum_{y \in \mathrm{Im}Y} y f_{Y|X}(y|x),$$

a ako je neprekidan vektor, tada je

$$\mathbb{E}[Y|X = x] := \int_{-\infty}^{+\infty} y f_{Y|X}(y|x)\, dy,$$

uz uvjet da je u oba slučaja $f_X(x) > 0$. Analogne definicijske formule vrijede za $\mathbb{E}[X|Y = y]$. Ukoliko su X i Y nezavisne slučajne varijable, primijetimo da je, zbog (4.1),

$$\mathbb{E}[Y|X = x] = \mathbb{E}[Y] \ \text{ i } \ \mathbb{E}[X|Y = y] = \mathbb{E}[X] \tag{4.22}$$

za sve x i y takve da su pripadajuća uvjetna očekivanja definirana.

Funkcija $x \mapsto \mathbb{E}[Y|X = x]$, koja je definirana za brojeve x za koje je $f_X(x) > 0$, zove se *regresijska funkcija od Y na $X = x$*. Ta se funkcija može proširiti na sve realne brojeve x na sljedeći način. Definiramo funkciju $g : \mathbb{R} \to \mathbb{R}$ sa

$$g(x) := \begin{cases} \mathbb{E}[Y|X = x] & \text{ako je } \ f_X(x) > 0 \\ 0 & \text{inače.} \end{cases}$$

Ako je x opažena vrijednost slučajne varijable X, tada je $g(x)$ regresijska funkcija od Y na toj vrijednosti od X. Označimo slučajnu varijablu $g(X) = g \circ X$ sa $\mathbb{E}[Y|X]$. Tu slučajnu varijablu zovemo *uvjetno očekivanje od Y za dano X*. Analogno se definiraju *regresijska funkcija od X na $Y = y$* i slučajna varijabla $\mathbb{E}[X|Y]$, *uvjetno očekivanje od X za dano Y*.

Budući da je $\mathbb{E}[Y|X]$ slučajna varijabla, $\mathbb{E}[Y|X]$ ima svoju (vjerojatnosnu) razdiobu. Posebno nas zanima matematičko očekivanje i varijanca te varijable. Vrijedi:

$$\mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}[Y]. \tag{4.23}$$

Dokaz. Prvo, pretpostavimo da je (X, Y) diskretan slučajni vektor. Tada je

$$\begin{aligned}
\mathbb{E}[\mathbb{E}[Y|X]] &= \mathbb{E}[g(X)] = \sum_{x \in \mathrm{Im}X} g(x) f_X(x) = \sum_{x \in \mathrm{Im}X} \mathbb{E}[Y|X = x] f_X(x) = \\
&= \sum_{x \in \mathrm{Im}X} \Big(\sum_{y \in \mathrm{Im}Y} y f_{Y|X}(y|x) \Big) f_X(x) = \sum_{x \in \mathrm{Im}X} \sum_{y \in \mathrm{Im}Y} y \frac{f_{X,Y}(x,y)}{f_X(x)} f_X(x) = \\
&= \sum_{x \in \mathrm{Im}X} \sum_{y \in \mathrm{Im}Y} y f_{X,Y}(x,y) = \sum_{y \in \mathrm{Im}Y} y \Big(\sum_{x \in \mathrm{Im}X} f_{X,Y}(x,y) \Big) = \sum_{y \in \mathrm{Im}Y} y f_Y(y) = \\
&= \mathbb{E}[Y].
\end{aligned}$$

Ako je (X, Y) neprekidan slučajni vektor, tada je dokaz isti kao za diskretan slučaj samo što se svuda sume zamijene integralima. □

Označimo sa $\mathbb{V}\mathrm{ar}[Y|X = x]$ varijancu uvjetne razdiobe od Y uz dano $X = x$. Tada je ta varijanca također funkcija vrijednosti x od X takvih da je $f_X(x) > 0$. Nadalje, vrijedi da je

$$\mathbb{V}\mathrm{ar}[Y|X = x] = \mathbb{E}[Y^2|X = x] - \mathbb{E}[Y|X = x]^2. \tag{4.24}$$

Kompozicaja te funkcije i slučajne varijable X je također slučajna varijabla koju označavamo $\mathbb{V}\mathrm{ar}[Y|X]$ i zovemo *uvjetnom varijancom od Y uz dano X*. Vrijedi:

$$\mathbb{V}\mathrm{ar}[\mathbb{E}[Y|X]] = \mathbb{V}\mathrm{ar}[Y] - \mathbb{E}[\mathbb{V}\mathrm{ar}[Y|X]]. \tag{4.25}$$

Dokaz. Iz (4.24) je $\mathbb{V}\mathrm{ar}[Y|X] = \mathbb{E}[Y^2|X] - \mathbb{E}[Y|X]^2$ pa je

$$\mathbb{E}[\mathbb{V}\mathrm{ar}[Y|X]] \overset{\text{lin.}}{=} \mathbb{E}[\mathbb{E}[Y^2|X]] - \mathbb{E}[\mathbb{E}[Y|X]^2] \overset{(4.23)}{=} \mathbb{E}[Y^2] - \mathbb{E}[\mathbb{E}[Y|X]^2].$$

S druge strane je

$$\mathbb{V}\mathrm{ar}[\mathbb{E}[Y|X]] = \mathbb{E}[\mathbb{E}[Y|X]^2] - \mathbb{E}[\mathbb{E}[Y|X]]^2 \overset{(4.23)}{=} \mathbb{E}[\mathbb{E}[Y|X]^2] - \mathbb{E}[Y]^2.$$

Zbrajanjem lijevih, te desnih strana dobivenih jednakosti imamo

$$\mathbb{V}\mathrm{ar}[\mathbb{E}[Y|X]] + \mathbb{E}[\mathbb{V}\mathrm{ar}[Y|X]] = \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 = \mathbb{V}\mathrm{ar}[Y]$$

odakle slijedi jednakost (4.25). □

Primjer 4.12 Neka su $N, X_1, X_2, \ldots$ nezavisne slučajne varijable takve da su $X_1, X_2, \ldots$ jednako distribuirane s f.i.m. jednakom $M_X(t)$, a N je brojeća slučajna varijabla s f.i.v. $G_N(t)$. Nadalje, neka je

$$S = X_1 + X_2 + \cdots + X_N,$$

pri čemu uzimamo da je $S = 0$ na događaju $\{N = 0\}$. Tada je f.i.m. slučajne varijable S jednaka

$$\begin{aligned}
M_S(t) &= \mathbb{E}[e^{tS}] = (4.23) = \mathbb{E}[\mathbb{E}[e^{tS}|N]] = \\
&= \sum_{n=0}^{\infty} \mathbb{E}[e^{tS}|N=n]\mathbb{P}(N=n) = \\
&= \sum_{n=0}^{\infty} \mathbb{E}[e^{t(X_1+X_2+\cdots+X_n)}|N=n]\mathbb{P}(N=n) = (\text{nez. i } (4.22)) \\
&= \sum_{n=0}^{\infty} \mathbb{E}[e^{t(X_1+X_2+\cdots+X_n)}]\mathbb{P}(N=n) = (\text{nez. i } (4.21)) \\
&= \sum_{n=0}^{\infty} M_X(t)^n \mathbb{P}(N=n) = (\text{def. f.i.v.}) \\
&= G_N(M_X(t)). \qquad (4.26)
\end{aligned}$$

□

Poglavlje 5

Centralni granični teorem

Centralni granični teorem (kraće, CGT) jedan je od najvažnijih rezultata teorije vjerojatnosti koji je našao primjene u statistici. Na njemu se zasniva statističko zaključivanje (inferencijalna statistika) o populacijskim srednjim vrijednostima i proporcijama na osnovi velikih uzoraka i kada nije poznata populacijska distribucija. Jedan je od uzroka važnosti normalne razdiobe u statsitici.

5.1 CGT

Teorem. (CGT) *Neka je X_1, X_2,... niz n.j.d. slučajnih varijabli s konačnim matematičkim očekivanjem μ i konačnom varijancom $\sigma^2 > 0$. Nadalje, neka je $\overline{X}_n := (X_1 + X_2 + \cdots + X_n)/n$ za sve prirodne brojeve n. Tada za sve $a < b$ vrijedi*

$$\lim_{n\to+\infty} \mathbb{P}\left(a \leq \frac{\overline{X}_n - \mu}{\sigma}\sqrt{n} \leq b\right) = \Phi(b) - \Phi(a),$$

gdje je $\Phi(x)$ funkcija distribucije jedinične normalne razdiobe.

Drugim riječima, kažemo da niz slučajnih varijabli $(\overline{X}_n - \mu)\sqrt{n}/\sigma$ *konvergira po distribuciji* jedničnoj normalnoj razdiobi kada n teži u beskonačnost, i pišemo

$$\frac{\overline{X}_n - \mu}{\sigma}\sqrt{n} \xrightarrow{D} N(0,1), \quad n \to \infty.$$

U sljedećem poglavlju pokazat ćemo da je slučajna varijabla

$$Z = \frac{\overline{X}_n - \mu}{\sigma}\sqrt{n}$$

standardizirana verzija od aritmetičke sredine $\overline{X}_n =$ (kraće) $= \overline{X}$. Nadalje, Z je i standardizirana verzija od $\sum_{i=1}^n X_i$ jer je

$$\frac{\overline{X} - \mu}{\sigma}\sqrt{n} = \frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}}.$$

Dakle, CGT kaže da standardizirana verzija aritmetičke sredine (odnosno, zbroja) od n n.j.d. slučajnih varijabli s konačnom ne-nul varijancom ima aproksimativno jediničnu normalnu razdiobu za velike n. Ako sa "$\approx$" označimo izraz "aproksimativno distribuirano", tada pišemo

$$\frac{\overline{X} - \mu}{\sigma}\sqrt{n} \approx N(0,1) \text{ za veliko } n, \quad \text{odnosno} \quad \frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} \approx N(0,1) \text{ za veliko } n.$$

Alternativno se koristi nestandardizirana verzija CGT-a:

$$\overline{X} \overset{\cdot}{\sim} N(\mu, \frac{\sigma^2}{n}) \text{ za veliko } n, \quad \text{odnosno} \quad \sum_{i=1}^{n} X_i \overset{\cdot}{\sim} N(n\mu, n\sigma^2) \text{ za veliko } n.$$

Prirodno se postavlja pitanje koliko n mora biti velik da bi aproksimacija normalnom razdiobom bila zadovoljavajuća. Obično se uzima $n \geq 30$, ali potpuni odgovor bi glasio da veličina od n ovisi o obliku razdiobe slučajnih varijabli X_i, preciznije, je li simetrična, a ako nije, koliko je asimetrična. Ako je razdioba od X_i približno simetrična, tada i $n = 10$ može biti dovoljno velik, a ako je distrbucija znatno asimetrična, tada se mora uzeti barem $n = 50$. Na slikama 5.1, 5.2 i 5.3 ilustrirano je koliko dobro normalna razdioba aproksimira distribuciju zbroja n.j.d. slučajnih varijabli (ili njihovih standardiziranih verzija) u ovisnosti o n.

5.2 Normalna aproksimacija

Navedimo primjene CGT-a na neke od važnih distribucija.

5.2.1 Binomna razdioba

Neka je X binomna (n, θ) slučajna varijabla. Tada znamo da se X može reprezentirati pomoću zbroja $X = \sum_{i=1}^{n} X_i$ od n n.j.d. Bernoullijevih slučajnih varijabli X_i $(i = 1, 2, \ldots, n)$ s parametrom θ. Budući da je

$$\mu = \mathbb{E}[X_i] = \theta, \quad \sigma^2 = \mathbb{V}\text{ar}[X_i] = \theta(1 - \theta),$$

varijable X_i $(i = 1, 2, \ldots, n)$ zadovoljavaju uvjete CGT-a. Dakle,

$$X \overset{\cdot}{\sim} N(n\theta, n\theta(1 - \theta)) \text{ za velike } n.$$

n je dovoljno velik ako je istovremeno $n\theta \geq 5$ i $n(1 - \theta) \geq 5$. Na primjer, ako je $\theta = 0.5$, tada je dovoljno uzeti $n = 10$, a ako je $\theta = 0.2$ ili $\theta = 0.8$ treba uzeti barem $n = 25$. Za manje (ili veće) θ treba uzeti veći n. Na slici 5.1 grafički je prikazana aproksimacija binomne pomoću normalne razdiobe za razne vrijednosti parametra n.

5.2.2 Poissonova razdioba

Neka je X_1, X_2,..., X_n niz n.j.d. $P(\lambda)$-distribuiranih slučajnih varijabli. Dakle,

$$\mu = \mathbb{E}[X_i] = \lambda, \quad \sigma^2 = \mathbb{V}\text{ar}[X_i] = \lambda,$$

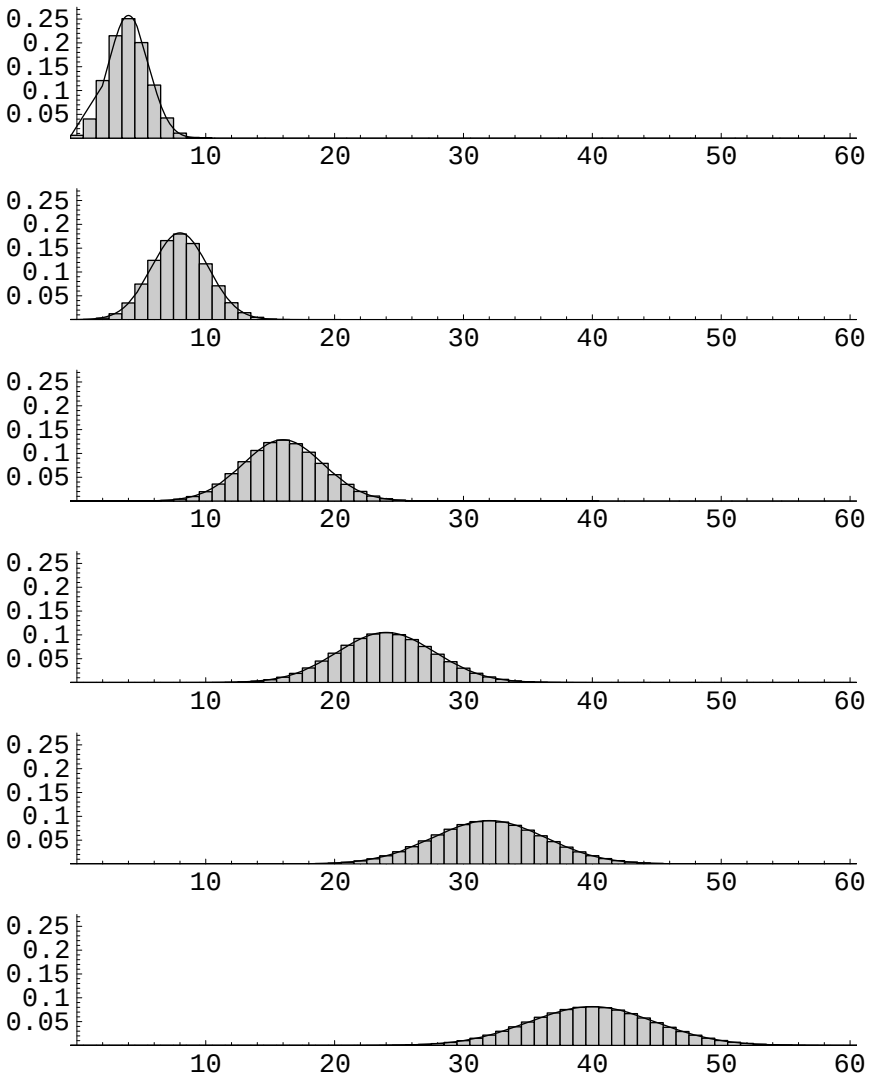
pa je prema CGT-u za $X = \sum_{i=1}^{n} X_i$,

$$X \overset{\cdot}{\sim} N(n\lambda, n\lambda) \text{ za velike } n.$$

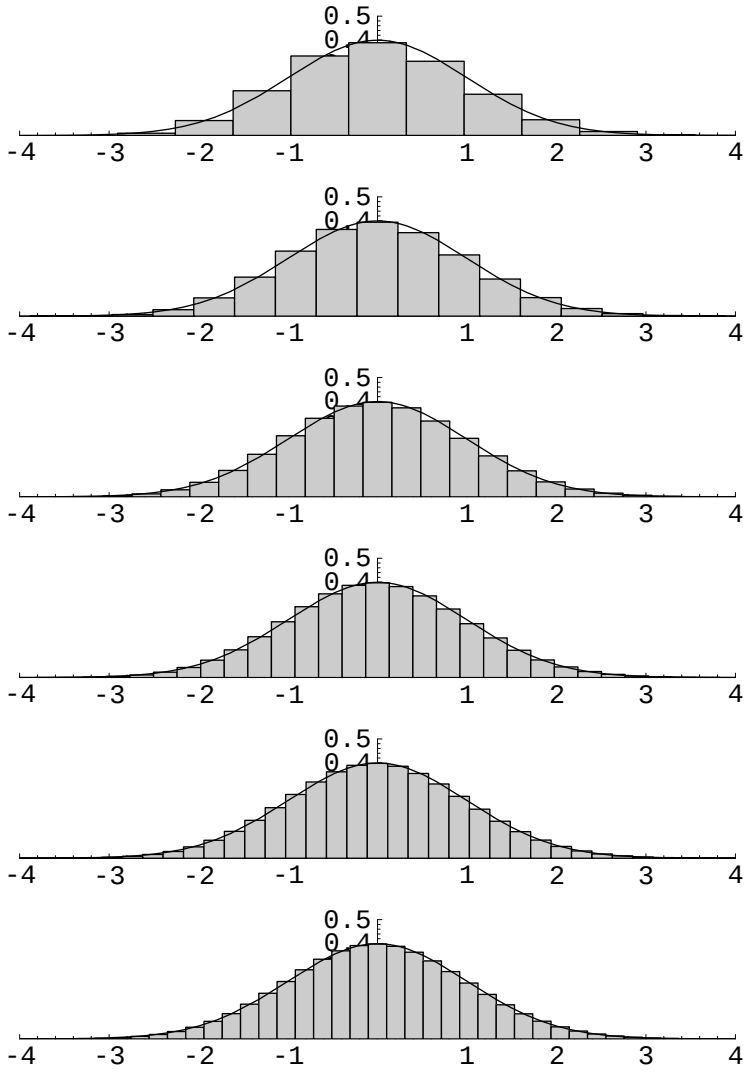
S druge strane, $X \sim P(n\lambda)$, pa gornja relacija za granično ponašanje od X povlači da je

$$P(\lambda) \overset{\cdot}{\sim} N(\lambda, \lambda) \text{ za veliko } \lambda.$$

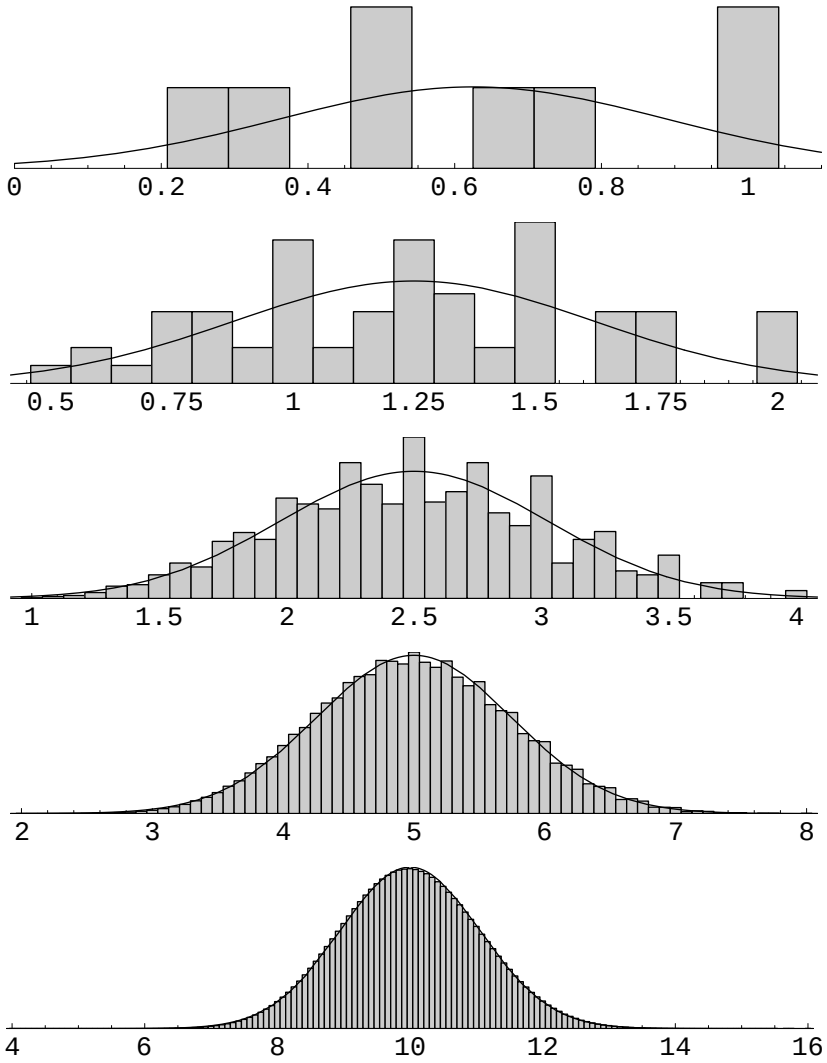
Ta aproksimacija je dobra za $\lambda > 5$.



Slika 5.1: Usporedba histograma binomne $(n, \frac{2}{5})$ razdiobe i gustoće pripadne normalne $N(\frac{2}{5}n, \frac{6}{5}n)$ razdiobe za $n = 10, 20, 40, 60, 80, 100$.



Slika 5.2: Usporedba histograma standardizirane verzije binomne $(n, \frac{2}{5})$ razdiobe i gustoće jedinične normalne razdiobe za $n = 10, 20, 40, 60, 80, 100$.



Slika 5.3: Usporedba histograma konvolucija f^{n*} gustoće f diskretne vjerojatnosne razdiobe s gustoćama pripadnih normalnih razdioba (s istim matematičkim očekivanjem i varijancom) za $n = 1, 2, 4, 8, 16$.

5.2.3 Gama razdioba

Neka je X_1, X_2,..., X_n niz n.j.d. $Exp(\lambda)$-distribuiranih slučajnih varijabli. Dakle,

$$\mu = \mathbb{E}[X_i] = \frac{1}{\lambda}, \quad \sigma^2 = \mathbb{V}\mathrm{ar}[X_i] = \frac{1}{\lambda^2},$$

pa je prema CGT-u za $X = \sum_{i=1}^{n} X_i \sim \Gamma(k, 1/\lambda)$,

$$X \approx N(\frac{n}{\lambda}, \frac{n}{\lambda^2}) \text{ za velike } n.$$

Slično,

$$\chi^2(n) \equiv \Gamma(\frac{n}{2}, 2) \approx N(n, 2n) \text{ za veliki broj stupnjeva slobode } n.$$

5.3 Korekcija zbog neprekidnosti

Na primjer, u slučaju aproksimacije binomne ili Poissonove razdiobe normalnom, diskretnu razdiobu aproksimiramo neprekidnom razdiobom. Iako je za diskretnu slučajnu varijablu X sasvim legitimno računati vjerojatnosti diskretnih događaja oblika $\{X = x\}$, u slučaju neprekidnih varijabli, te vjerojatnosti su uvijek jednake nula. Dakle, problem je kako računati aproksimativne vrijednosti vjerojatnosti takvih događaja pomoću neprekidnih razdioba. Rješenje je da se svaka cjelobrojna vrijednost tretira kao da je dobivena zaokruživanjem varijable na najbliži cijeli broj, odnosno da se zapis događaja pomoću diskretnih vrijednosti izrazi u ekvivalentnom obliku izraženom pomoću poprimanja vrijednosti varijable u intervalu. Na primjer, sljedeći događaji su ekvivalentni:

$$\begin{aligned} \{X = 4\} &= \{3.5 < X < 4.5\} \\ \{X > 15\} &= \{X > 15.5\} \\ \{X \geq 15\} &= \{X > 14.5\}. \end{aligned}$$

Kažemo da radimo *korekciju zbog neprekidnosti*.

Primjer 5.1 Neka je $X \sim P(20)$. Izračunajmo $\mathbb{P}(X \leq 15)$ egzaktno i pomoću normalne aproksimacije $X \approx N(20, 20)$ i usporedimo rezultate. Egzaktnu vrijednost vjerojatnosti navedenog događaja možemo očitati iz tablica za Poissonovu razdiobu ili izračunati po formuli

$$\mathbb{P}(X \leq 15) = \sum_{k=0}^{15} f_X(k)$$

gdje su vjerojatnosti $f_X(k)$ $(k = 0, 1, \ldots, 15)$ izračunate pomoću rekurzivne relacije:

$$f_X(k) = \frac{\lambda}{k} f_X(k-1), \text{ za } k \geq 1 \text{ i uz početnu vrijednost } f_X(0) = e^{-\lambda},$$

za $\lambda = 20$. Izlazi da je egzatno $\mathbb{P}(X \leq 15) = 0.15651$. Pomoću normalne aproksimacije račun je jednostavniji:

$$\begin{aligned} \mathbb{P}(X \leq 15) &= \mathbb{P}(X < 15.5) \quad \text{(korekcija zbog neprekidnosti)} \\ &= \mathbb{P}(\frac{X - 20}{\sqrt{20}} < \frac{15.5 - 20}{\sqrt{20}}) = \mathbb{P}(Z < -1.006) \quad \text{(standardizacija)} \\ &\approx \Phi(-1.006) \quad \text{(normalna aproksimacija)} \\ &= \frac{1}{2} - \Phi_0(1.006) = \text{(tablice)} = 0.15721. \end{aligned}$$

Pogreška je $|0.15721 - 0.15651| = 0.0007$, odnosno relatina pogreška je $0.0007/0.15651 = 0.45\%$. □

Poglavlje 6

Uzorkovanje i statističko zaključivanje

O populacijskoj razdiobi varijable koju izučavamo informacije dobivamo iz uzorka uzetog iz te populacije. Na primjer, procjenu populacijske srednje vrijednosti ili proporcije, odnosno utvrđivanje istinitosti hipoteza o populacijskoj razdiobi donosimo na osnovi vrijednosti iz uzorka.

6.1 Osnovne definicije

Sve definicije u ovom poglavlju temelje se na pretpostavci da su populacije beskonačne. Iako je većina populacija koje su u fokusu interesa aktuara u stvarnosti konačna, na primjer, osiguranici, police, zaposlenici, zgrade itd., njihova brojnost je dovoljno velika da se na njih mogu primijeniti metode statističkog zaključivanja (inferencijalne statistike) o beskonačnim populacijama.

Slučajni uzorak je niz n.j.d. slučajnih varijabli. Označavamo ga sa $\underline{X}$. Na primjer, ako se sastoji od n n.j.d. varijabli X_1, X_2,..., X_n, $\underline{X}$ je slučajni vektor

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

Intuitivno, slučajni uzorak predstavlja niz mjerenja (opažanja) slučajnih vrijednosti izučavane varijable X na članovima (jedinicama) odabranih u uzorak na slučajan način iz populacije. Kažemo da se članovi biraju u uzorak na *slučajan način* ako svaki element iz populacije ima jednaku šansu da bude izabran u uzorak i neovisno od drugih članova u uzorku. Uz tu interpretaciju, dakle, varijable X_1, X_2,..., X_n su nezavisne i imaju distribuciju jednaku populacijskoj distribuciji varijable X. Označimo sa $f(x|\theta)$ gustoću populacijske razdiobe od X, dakle, razdiobu svake od varijabli u slučajnom uzorku. θ označava parametre te razdiobe.

Uređenu n-torku brojeva $\underline{x} = (x_1, x_2, \ldots, x_n)$ koja predstavlja realizaciju slučajnog uzorka $\underline{X}$ zovemo *opaženi uzorak*.

Statistika je funkcija slučajnog uzorka koja ne sadrži nepoznate parametre. Na primjer, *uzoračka sredina*

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

i *uzoračka varijanca*

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

su statistike, dok uzoračka drugi moment oko parametra očekivanja $\mu = \mathbb{E}[X]$,

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

nije statistika, ukoliko nam vrijednost od μ nije poznata. Općenito statistike označavamo sa $g(\underline{X})$.

Budući da je statistika funkcija slučajnih varijabli, slučajna je varijabla pa ima svoju razdiobu koju zovemo *uzoračkom razdiobom*. Primijetite da je zbog CGT-a uzoračka razdioba uzoračke sredine $\overline{X}$ za velike uzorke aproksimativno normalna bez obzira koja je populacijska razdioba izučavane varijable X (uz jedini uvjet da je populacijska varijanca konačna i nije jednaka nuli).

6.2 Momenti uzoračke sredine i varijance

Neka je $\underline{X} = (X_1, X_2, \ldots, X_n)$ slučajni uzorak duljine n iz populacije opisane varijablom X čija populacijska razdioba ima matematičko očekivanja μ i varijancu σ^2. Kraće kažemo da je $\underline{X}$ slučajni uzorak iz populacije s parametrima očekivanja μ i varijance σ^2.

6.2.1 Uzoračka sredina

Uzoračka sredina je statistika

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

Računamo matematičko očekivanje i varijancu uzoračke razdiobe te statistike:

$$\begin{aligned}
\mathbb{E}[\overline{X}] &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] \quad \text{(linearnost očekivanja)} \\
&= \frac{1}{n}\cdot n\mu = \mu \quad \text{(jednaka distribuiranost)} \\
\mathbb{V}\mathrm{ar}[\overline{X}] &\overset{(2.7)}{=} \frac{1}{n^2}\mathbb{V}\mathrm{ar}[\sum_{i=1}^{n} X_i] = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}\mathrm{ar}[X_i] \quad \text{(nezavisnost)} \\
&= \frac{1}{n^2}\cdot n\sigma^2 = \frac{\sigma^2}{n} \quad \text{(jednaka distribuiranost).}
\end{aligned}$$

Dakle,

$$\mathbb{E}[\overline{X}] = \mu \tag{6.1}$$

$$\mathbb{V}\mathrm{ar}[\overline{X}] = \frac{\sigma^2}{n}. \tag{6.2}$$

Standardna devijacija od $\overline{X}$ zove se *standardna greška* uzoračke sredine, i označava se sa s.e.$(\overline{X})$. Dakle,

$$\text{s.e.}(\overline{X}) := \sigma(\overline{X}) = \frac{\sigma}{\sqrt{n}}.$$

6.2.2 Uzoračka varijanca

Uzoračka varijanca je statistika

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

Koristeći identitete

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n} X_i^2 - \frac{n}{n-1}\overline{X}^2, \quad \mathbb{E}[Y^2] = \mathbb{V}\mathrm{ar}[Y] + \mathbb{E}[Y]^2,$$

računamo matematičko očekivanje od S^2:

$$\begin{aligned}
\mathbb{E}[S^2] &= \frac{1}{n-1}\sum_{i=1}^{n}\mathbb{E}[X_i^2] - \frac{n}{n-1}\mathbb{E}[\overline{X}^2] \quad \text{(linearnost očekivanja)} \\
&= \frac{1}{n-1}\sum_{i=1}^{n}(\mathbb{V}\mathrm{ar}[X_i] + \mathbb{E}[X_i]^2) - \frac{n}{n-1}(\mathbb{V}\mathrm{ar}[\overline{X}] + \mathbb{E}[\overline{X}]^2) = \\
&= \frac{1}{n-1}\cdot n(\sigma^2+\mu^2) - \frac{n}{n-1}(\frac{\sigma^2}{n}+\mu^2) \quad \text{(jednaka distribuiranost, (6.1), (6.2))} \\
&= \frac{1}{n-1}\cdot(n-1)\sigma^2 = \sigma^2.
\end{aligned}$$

Dakle,

$$\mathbb{E}[S^2] = \sigma^2. \tag{6.3}$$

Varijancu te statistike izračunat ćemo u sljedećem potpoglavlju samo u jednom specijalnom, ali važnom, slučaju: za uzorke iz normalno distribuiranih populacija.

6.3 Uzoračke razdiobe statistika normalnog uzorka

Neka je $\underline{X} = (X_1, X_2, \ldots, X_n)$ slučajni uzorak duljine n iz populacije s normalnom distribucijom (kraće, normalne populacije) $N(\mu, \sigma^2)$.

6.3.1 Uzoračka sredina

Zbog svojstva invarijantnosti normalne razdiobe na linearne kombinacije nezavisnih normalno distribuiranih slučajnih varijabli, $\overline{X}$ ima normalnu (uzoračku) razdiobu, preciznije:

$$\overline{X} \sim N(\mu, \frac{\sigma^2}{n}).$$

Prema tome, sandardizirana verzija Z od $\overline{X}$ ima jediničnu normalnu razdiobu:

$$Z = \frac{\overline{X}-\mu}{\sigma}\sqrt{n} \sim N(0,1).$$

Usporedimo li taj zaključak sa CGT-om, vidimo da u slučaju normalno distribuiranog uzorka, konvergencija standardiziranih verzija aritmetičkih sredina iz CGT-a prelazi u identitet.

6.3.2 Uzoračka varijanca

Za uzoračku varijancu S^2 normalnog uzorka vrijedi da je

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1).$$

Dakle, S^2 ima χ^2-distribuciju koja je pozitivno asimetrična i čiji se koeficijent asimetrije smanjuje povećanjem duljine uzorka (primijetite da je $\chi^2(k) \approx N(k, 2k)$ za velike k). Vrijedi:

$$\mathbb{E}[\frac{(n-1)S^2}{\sigma^2}] = \mathbb{E}[\chi^2(n-1)] = n-1 \quad \Rightarrow \quad \mathbb{E}[S^2] = \sigma^2$$
$$\mathbb{V}\mathrm{ar}[\frac{(n-1)S^2}{\sigma^2}] = \mathbb{V}\mathrm{ar}[\chi^2(n-1)] = 2(n-1) \quad \Rightarrow \quad \mathbb{V}\mathrm{ar}[S^2] = \frac{2\sigma^4}{n-1}$$

Za obje statistike $\overline{X}$ i S^2 vrijedi da im varijance teže u nulu porastom uzorka. Budući da je $\mathbb{E}[\overline{X}] = \mu$ i $\mathbb{E}[S^2] = \sigma^2$, to znači da $\overline{X}$ teži vrijednosti parametra μ i isto tako S^2 teži vrijednosti parametra σ^2 kako veličina uzorka raste. To su poželjna svojstva tih statistika.

6.3.3 Nezavisnost uzoračke sredine i varijance

Sljedeće važno svojstvo uzorka iz normalne razdiobe je nezavisnost statistika $\overline{X}$ i S^2. Dokaz te činjenice nije jednostavan, ali je njezina vjerodostojnost vidljiva iz sljedećeg promišljanja. Ako uzorak dolazi iz normalne razdiobe, tada vrijednost $\bar{x}$ statistike $\overline{X}$ na opaženom uzorku ne daje nikakve informacije o parametru σ^2. Isto tako ni vrijednost s^2 statistike S^2 ne daje nikakvu informaciju o parametru μ. Naprotiv, ako uzorak dolazi iz, na primjer, eksponencijalno distribuirane populacije, tada $\bar{x}$ nosi informaciju i o populacijskoj varijanci jer su parametri μ i σ^2 za takvu populaciju povezani relacijom $\sigma^2 = \mu^2$.

6.4 Studentova t-distribucija

Za statističko zaključivanje o parametru očekivanja μ koristi se standardizirana verzija Z statistike $\overline{X}$:

$$Z = \frac{\overline{X} - \mu}{\sigma}\sqrt{n},$$

ako je poznata vrijednost parametra standardne devijacije σ. U tom slučaju je uzoračka razdioba od Z:

$$\begin{array}{rcll} Z & \sim & N(0,1) & \text{ako je } \underline{X} \text{ uzet iz normalne populacije,} \\ Z & \approx & N(0,1) & \text{ako je } n \text{ dovoljno velik i } 0 < \sigma^2 < +\infty. \end{array}$$

Ako je vrijednost od σ nepoznata, za zaključivanje o μ koristi se *studentizirana* verzija T od $\overline{X}$, dakle, slučajna varijabla

$$T := \frac{\overline{X} - \mu}{S}\sqrt{n}$$

gdje je $S = \sqrt{S^2}$ *uzoračka standardna devijacija*.

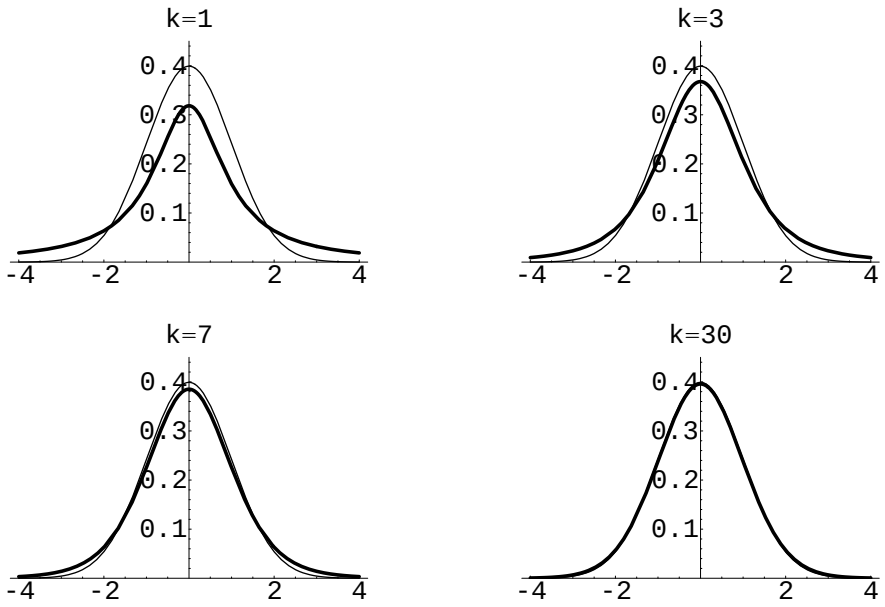
Pokazuje se da za uzorak $\underline{X}$ iz normalne populacije uzoračka razdioba od T ne ovisi o parametrima μ i σ^2 te populacije, već samo o duljini uzorka. Kažemo da T ima *Studentovu* ili *t-razdiobu* s $n-1$ stupnjem slobode. Tu činjenicu zapisujemo na sljedeći način:

$$T \sim t(n-1).$$

Ako populacijska razdioba nije normalna, taj rezutat ne mora vrijediti.

Općenito, ako su Z i V dvije nezavisne slučajne varijable, $Z \sim N(0,1)$ i $V \sim \chi^2(k)$, tada slučajna varijabla $Z/\sqrt{V/k}$ ima Studentovu ili t-distribuciju s k stupnjeva slobode, i pišemo

$$\frac{Z}{\sqrt{V/k}} \sim t(k).$$



Slika 6.1: Usporedba gustoća $t(k)$-razdioba s gustoćom $N(0,1)$-razdiobe za $k = 1, 3, 7, 30$.

Vrijednosti t-razdiobe su tabelirane[1].

Za $k > 1$, $t(k)$-razdioba ima matematičko očekivanje, a za $k > 2$ ima i varijancu, i vrijedi

$$\mathbb{E}[t(k)] = 0, \quad \mathbb{V}\mathrm{ar}[t(k)] = \frac{k}{k-2}.$$

Studentova $t(1)$-razdioba zove se još *Cauchyjeva* distribucija. Za tu razdiobu vrijedi da nema niti jedan moment. Dakle, za Cauchyjevu razdiobu ne vrijedi CGT. Cauchyjeva razdioba se hipotetski može pojaviti kao uzoračka razdioba statistike T koja se računa na normalnom uzorku duljine $n = 2$, ali tako mali uzorci se u pravilu ne pojavljuju u primjenama.

Nadalje, vrijedi

$$t(k) \xrightarrow{D} N(0,1), \quad k \to \infty.$$

Brzina te konvergencije ilustrirana je na slici 6.1 usporedbom grafova gustoća $t(k)$-razdioba za $k = 1, 3, 7, 30$ sa Gaussovom krivuljom jedinične normalne radiobe.

Za populaciju koja nije normalno distribuirana studentizirana verzija T od $\overline{X}$ ne mora imati t-razdiobu, ali prema CGT-u i zbog činjenice da S teži ka σ kada n raste,

$$T = \frac{\overline{X} - \mu}{S}\sqrt{n} \overset{\cdot}{\sim} N(0,1) \text{ za velike } n,$$

što se koristi za statističko zaključivanje o parametru μ na osnovi velikih uzoraka.

[1] Najčešće su tabelirane vrijednosti $t_\alpha(k)$ $(1-\alpha)$-kvantila $t(k)$-razdiobe. Brojevi $t_\alpha(k)$ još se zovu *kritične vrijednosti* i opisuju se relacijom $\mathbb{P}(T \geq t_\alpha(k)) = \alpha$ ako je $T \sim t(k)$.

6.5 Fisherova F-razdioba

Ako su U i V dvije nezavisne slučajne varijable, $U \sim \chi^2(\nu_1)$, $V \sim \chi^2(\nu_2)$, tada slučajna varijabla

$$F := \frac{U/\nu_1}{V/\nu_2}$$

ima *Fisherovu* ili *F-razdiobu* s (ν_1, ν_2) stupnjeva slobode. Pišemo

$$F \sim F(\nu_1, \nu_2).$$

Neka su S_1^2 i S_2^2 uzoračke varijance dvaju nezavisnih uzoraka duljine n_1, odnosno n_2 iz normalno distribuiranih populacija s varijancama σ_1^2, odnosno σ_2^2. Tada vrijedi:

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1). \tag{6.4}$$

Pri tome je svejedno kako smo uzorke numerirali. Odavde odmah slijedi ekvivalencija:

$$F \sim F(\nu_1, \nu_2) \quad \Leftrightarrow \quad \frac{1}{F} \sim F(\nu_2, \nu_1).$$

Ta se činjenica koristi kod izračunavanja kritičnih vrijednosti iz tablica za F-distribuciju, budući da su tabelirani samo neki kvantili[2].

Na kraju, spomenimo da (6.4) ne mora vrijediti ako obje populacije nisu normalno distribuirane ili ako uzorci nisu nezavisni jedan od drugoga.

[2]U tablicama se najčešće nalaze samo 0.95 i 0.99-kvantili F-razdioba, dakle kritične vrijednosti $f_\alpha(\nu_1, \nu_2)$ za $\alpha = 0.05$ ili 0.01 opisane relacijom $\mathbb{P}(F \geq f_\alpha(\nu_1, \nu_2)) = \alpha$ za $F \sim F(\nu_1, \nu_2)$. 0.05 i 0.01-kvantili, odnosno kritične vrijednosti $f_{1-\alpha}(\nu_1, \nu_2)$ računaju se pomoću formule $f_{1-\alpha}(\nu_1, \nu_2) = 1/f_\alpha(\nu_2, \nu_1)$.

Poglavlje 7

Točkovne procjene

U ovom poglavlju navest ćemo metode za određivanje *procjenitelja* parametara populacijske razdiobe, odnosno statistika koje služe za procjenjivanje njihovih vrijednosti, i navest ćemo neka njihova svojstva.

7.1 Metoda momenata

Osnovni princip metode momenata je da se izjednače populacijske vrijednosti momenata (koje su funkcije nepoznatih parametara) s odgovarajućim uzoračkim momentima. Procjenitelji parametara su rješenja tog sustava jednadžbi po parametrima kao nepoznanicama, dakle funkcije su uzoračkih momenata, a time i slučajnog uzorka. Procjena populacijskog parametra je vrijednost procjenitelja na opaženom uzorku. Dakle, procjenitelj je statistika, a procjena je broj, njena opažena vrijednost.

7.1.1 Jednoparametarski slučaj

Neka je $\underline{x}$ opaženi uzorak za varijablu X čija je populacijska razdioba opisana gustoćom $f(x|\theta)$. Ako imamo samo jedan nepoznati parametar θ, izračunajmo populacijsko očekivanje

$$\mu(\theta) = \mathbb{E}[X] = \begin{cases} \sum_{x \in \mathrm{Im}X} x f(x|\theta) & \text{ako je } X \text{ diskretna slučajna varijabla} \\ \int_{-\infty}^{+\infty} x f(x|\theta)\, dx & \text{ako je } X \text{ neprekidna slučajna varijabla} \end{cases}$$

kao funkciju od θ, te uzoračku sredinu $\bar{x}$. *Procjena od θ metodom momenata* je rješenje jednadžbe

$$\mu(\theta) = \bar{x}.$$

Označimo dobivenu procjenu sa $\hat{\theta} = \hat{\theta}(\underline{x})$. Tada je *procjenitelj za θ metodom momenata* statistika $\hat{\theta}(\underline{X})$. Kada je iz konteksta jasno da se radi o procjenitelju (a ne o njegovoj realizaciji, procjeni), kraće ćemo ga označavati sa $\hat{\theta}$. Primijetite da je $\hat{\theta}$, u stvari, funkcija prvog uzoračkog momenta $\bar{X}$.

Primjer 7.1 Neka je $\underline{X} = (X_1, X_2, \ldots, X_n)$ slučajni uzorak za varijablu X s populacijskom $Exp(\lambda)$-razdiobom, pri čemu je $\lambda > 0$ nepoznati parametar. Znamo da je $\mu(\lambda) = \mathbb{E}[X] = 1/\lambda$. Izjednačavanje prvih momenata, populacijskog i opaženog uzoračkog, dobijemo:

$$\frac{1}{\lambda} = \bar{x} \quad \Rightarrow \quad \hat{\lambda} = \frac{1}{\bar{x}}.$$

Dakle, procjenitelj metodom momenata za parametar λ eksponencijalne populacijske razdiobe je $\hat{\lambda} = \hat{\lambda}(\underline{X}) = 1/\overline{X}$. □

Može se dogoditi da populacijsko očekivanje nije funkcija nepoznatog parametra. U tom slučaju izjednačavaju se momenti prvog višeg reda za koji je populacijski moment funkcija parametra.

Primjer 7.2 Neka je $\underline{X} = (X_1, X_2, \ldots, X_n)$ slučajni uzorak za uniformnu razdiobu na intervalu $[-\theta, \theta]$ s nepoznatim parametrom $\theta > 0$. Budući da je $\mu = 0$, a $\sigma^2 = \mathbb{V}\mathrm{ar}[X] = \theta^2/3$, izjednačimo populacijsku i uzoračku varijancu i dobivenu jednadžbu riješimo po θ:

$$\frac{\theta^2}{3} = s^2 \quad \Rightarrow \quad \hat{\theta} = s\sqrt{3}.$$

Dakle, procjenitelj metodom momenata za parametar θ uniformne populacijske razdiobe na $[-\theta, \theta]$ je $\hat{\theta} = \hat{\theta}(\underline{X}) = S\sqrt{3}$, gdje je S uzoračka standardna devijacija. □

7.1.2 Dvoparametarski slučaj

Ako su dva populacijska parametra nepoznata, odnosno ako je $\theta = (\theta_1, \theta_2)$ dvodimenzionalan populacijski parametar, tada izjednačimo prve i druge populacijske momente $\mu(\theta_1, \theta_2) = \mathbb{E}[X]$ i $\mu_2(\theta_1, \theta_2) := \mathbb{E}[X^2]$ sa pripadnim uzoračkim momentima:

$$\begin{array}{rcl} \mu(\theta_1, \theta_2) & = & \bar{x} \\ \mu_2(\theta_1, \theta_2) & = & \frac{1}{n}\sum_{i=1}^{n} x_i^2 \end{array} \tag{7.1}$$

Rješenje $(\hat{\theta}_1, \hat{\theta}_2) = (\hat{\theta}_1(\underline{x}), \hat{\theta}_2(\underline{x}))$ tog sustava jednadžbi čine *procjene metodom momenata parametara* θ_1, θ_2. Prema tome, *procjenitelji metodom momenata* za te parametre su statistike $\hat{\theta}_1(\underline{X})$ i $\hat{\theta}_2(\underline{X})$.

Često se procjenitelji metodom momenata za dva nepoznata parametra, umjesto rješavanjem sustava (7.1), traže rješavanjem sustava:

$$\begin{array}{rcl} \mu(\theta_1, \theta_2) & = & \bar{x} \\ \sigma^2(\theta_1, \theta_2) & = & s^2, \end{array}$$

gdje je $\sigma^2(\theta_1, \theta_2) = \mathbb{V}\mathrm{ar}[X]$ populacijska, a s^2 opažena uzoračka varijanca.

Primjer 7.3 Neka je $\underline{X}$ slučajni uzorak iz normalno $N(\mu, \sigma^2)$-distribuirane populacije s nepoznatim parametrima $\theta = (\mu, \sigma^2)$. Budući da su $\mathbb{E}[X] = \mu$ i $\mathbb{V}\mathrm{ar}[X] = \sigma^2$, procjenitelji metodom momenata za μ i σ^2 su trivijalno

$$\hat{\mu} = \bar{X}, \quad \hat{\sigma}^2 = S^2.$$

□

U slučaju većeg broja nepoznatih parametara, sustave jednadžbi treba tvoriti izjednačavanjem populacijskih i uzoračkih momenata oko nule. Primijetite da su procjenitelji metodom momenata uvijek funkcije uzoračkih momenata.

7.2 Metoda najveće vjerodostojnosti

Metoda maksimalne vjerodostojnosti (kraće ML) smatra se najboljom općom metodom za nalaženje dobrih procjenitelja populacijskih parametara. Procjenitelji dobiveni ovom metodom imaju dobra i lako odredljiva asimptotska svojstva, dakle, posebno su dobri za procjenjivanje na osnovi velikih uzoraka.

7.2.1 Jednoparametarski slučaj

Neka je $\underline{x} = (x_1, x_2, \ldots, x_n)$ opaženi uzorak za varijablu X s populacijskom gustoćom $f(x|\theta)$, te neka je θ jednodimenzionalan parametar. Tada je *vjerodostojnost* parametra θ funkcija

$$L(\theta) := \prod_{i=1}^{n} f(x_i|\theta).$$

Vjerodostojnost nepoznatog populacijskog parametra je, dakle, vjerojatnost da se dogodi realizacija uzorka koja je opažena ako je populacijska distribucija diskretna, odnosno proporcionalna je vjerojatnosti okoline opažene realizacije u neprekidnom slučaju, izražena kao funkcija nepoznatog parametra.

Procjena metodom maksimalne vjerodostojnosti parametra θ je ona vrijednost $\hat{\theta}$ za koju funkcija $\theta \mapsto L(\theta)$ poprima maksimalnu vrijednost:

$$L(\hat{\theta}) = \max_{\theta} L(\theta).$$

Očito je $\hat{\theta} = \hat{\theta}(\underline{x})$. Dakle, *procjenitelj metodom maksimalne vjerodostojnosti*, kraće MLE, parametra θ je statistika $\hat{\theta}(\underline{X})$.

Često se ne traži maksimum vjerodostojnosti nego njenog logaritma, tzv. *log-vjerodostojnosti*:

$$\ell(\theta) := \log L(\theta),$$

budući da se maksimumi tih dviju funkcija postižu u istoj vrijednosti za θ, a log-vjerodostojnost je u pravilu jednostavnija funkcija za maksimiziranje. Ako je $\theta \mapsto \ell(\theta)$ diferencijabilna funkcija, tada je MLE $\hat{\theta}$ jedno od rješenja stacionarne jednadžbe:

$$\ell'(\theta) = 0.$$

Dakle, MLE efektivno tražimo tako da riješimo stacionarnu jednadžbu, te kao MLE izdvojimo ono rješenje koje maksimizira log-vjerodostojnost. Na taj način možemo MLE tražiti jedino ako $\mathrm{Im}X$ ne ovisi o parametru θ.

Često nas osim parametra θ zanima neka funkcija $g(\theta)$ tog parametra. Za bilo koju funkciju g vrijedi da je MLE $\hat{g}(\theta)$ od $g(\theta)$ jednak $g(\hat{\theta})$ gdje je $\hat{\theta}$ MLE za θ. Kažemo da procjenitelj metodom maksimalne vjerodostojnosti ima svojstvo *invarijantnosti* na funkcijske transformacije parametara.

Primjer 7.4 Neka je $\underline{X} = (X_1, X_2, \ldots, X_n)$ slučajni uzorak iz populacije s $Exp(\lambda)$-razdiobom, pri čemu je $\lambda > 0$ nepoznati parametar. Tada je na osnovi opaženog uzorka $\underline{x} = (x_1, x_2, \ldots, x_n)$, vjerodostojnost od λ:

$$L(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i} = \lambda^n e^{-n\lambda \bar{x}}.$$

Nadalje, budući da $\mathrm{Im}X = \langle 0, +\infty \rangle$ ne ovisi o λ, MLE za λ tražimo rješavajući stacionarnu jednadžbu od log-vjerodostojnosti:

$$\begin{aligned} \ell(\lambda) &= \log L(\lambda) = n \log \lambda - n\lambda \bar{x} \\ \ell'(\lambda) &= 0 \quad \Leftrightarrow \quad \frac{n}{\lambda} - n\bar{x} = 0 \quad \Leftrightarrow \quad \lambda = \frac{1}{\bar{x}}. \end{aligned}$$

Dakle, MLE od λ je $\hat{\lambda} = \hat{\lambda}(\underline{X}) = 1/\overline{X}$. Nadalje, budući da je $\mu(\lambda) = 1/\lambda$ populacijsko očekivanje, zbog invarijantnosti od MLE vrijedi da je MLE populacijskog očekivanja jednak $\hat{\mu}(\lambda) = \mu(\hat{\lambda}) = \overline{X}$. □

Primjer 7.5 Neka je $\underline{x} = (x_1, x_2, \ldots, x_n)$ opaženi uzorak s populacijskom gustoćom

$$f(x|\theta) = \begin{cases} \frac{2}{\theta^2}x & \text{ako je } 0 \le x \le \theta, \\ 0 & \text{inače} \end{cases}$$

i nepoznatim parametrom $\theta > 1$. Vjerodostojnost od θ je:

$$L(\theta) = \prod_{i=1}^{n} f(x_i|\theta) = \begin{cases} \frac{2^n}{\theta^{2n}} \prod_{i=1}^{n} x_i & \text{ako je } \theta \ge x_{(n)} \\ 0 & \text{inače,} \end{cases}$$

gdje je $x_{(n)}$ maksimalna vrijednost u nizu $\underline{x}$. Budući da $\mathrm{Im}X = [0, \theta]$ ovisi o parametru, MLE ne možemo tražiti rješavajući stacionarnu jednadžbu log-vjerodostojnosti. Iz oblika funkcije $\theta \mapsto L(\theta)$ očito je da je $L(\theta) > 0$ samo za $\theta \ge x_{(n)}$, a na tom intervalu funkcija strogo pada. Dakle, maksimum vjerodostojnosti se postiže u točki $\hat{\theta} = x_{(n)}$, pa je MLE za θ statistika $\hat{\theta}(\underline{X}) = X_{(n)}$. □

7.2.2 Višeparametarski slučaj

Metoda je ista kao za jednoparametarski slučaj, samo što se kod traženja maksimuma log-vjerodostojnosti stacionarna jednadžba svodi na sustav više jednadžbi s više nepoznanica. Naime, taj sustav čine parcijalne derivacije log-vjerodostojnosti po nepoznatim parametrima izjednačene sa nulom. Na taj način tražimo MLE samo ako slika varijable X, $\mathrm{Im}X$, ne ovisi o parametrima.

Primjer 7.6 Neka je $\underline{x} = (x_1, x_2, \ldots, x_n)$ opaženi uzorak iz normalne populacije $N(\mu, \sigma^2)$ s nepoznatim parametrima μ, σ^2. Vjerodostojnost i pripadna log-vjerodostojnost tih parametara su

$$\begin{aligned} L(\mu, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{\sigma^2 2\pi}} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2} = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2} \\ \ell(\mu, \sigma^2) &= \log L(\mu, \sigma^2) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 - \frac{n}{2}\log\sigma^2 - \frac{n}{2}\log(2\pi). \end{aligned}$$

Budući da $\mathrm{Im}X = \mathbb{R}$ ne ovisi o parametrima, MLE se nalazi rješavanjem sustava:

$$\frac{\partial \ell}{\partial \mu}(\mu, \sigma^2) = 0, \quad \frac{\partial \ell}{\partial \sigma^2}(\mu, \sigma^2) = 0 \quad \Leftrightarrow \quad \hat{\mu} = \bar{x}, \quad \hat{\sigma}^2 = \frac{n-1}{n}s^2.$$

Dakle, MLE za (μ, σ^2) je $(\overline{X}, \frac{n-1}{n}S^2)$. □

7.2.3 Nepotpuni uzorci

Metoda ML se može primijeniti u situacijama kada je uzorak nepotpun. Na primjer, kada imamo rezane podatke ili kada imamo cenzurirane podatke za koje znamo jedino da je opažena vrijednost veća od neke fiksne vrijednosti. Preciznije, neka su opažene vrijednosti $x_1, x_2, \ldots, x_n$, a za ostalih m vrijednosti u uzorku zna se samo da su veće od broja y. Ako označimo sa $\mathbb{P}_\theta(X > y)$ vjerojatnost da će vrijednost varijable X (dakle, opažena vrijednost) biti veća od y (budući da je to funkcija parametra θ, da bi to naglasili, θ pišemo kao indeks), vjerodostojnost od θ je

$$L(\theta) := \prod_{i=1}^{n} f(x_i|\theta) \cdot (\mathbb{P}_\theta(X > y))^m.$$

MLE se kao i prije dobije maksimiziranjem te funkcije ili njenog logaritma. Pri tome je često nemoguće dobiti eksplicitno rješenje, pa se do procjena dolazi numeričkim metodama.

Primjer 7.7 Pretpostavimo da se u opaženom uzorku iz $Exp(\lambda)$-distribuirane populacije nalaze vrijednosti x_1, x_2,..., x_n, a da se za ostalih m zna da su veće od broja y. Tada je vjerodostojnost

$$\begin{aligned} L(\lambda) &= \prod_{i=1}^{n} f(x_i|\theta)\cdot(\mathbb{P}_\theta(X>y))^m = \prod_{i=1}^{n}\lambda e^{-\lambda x_i}\cdot(e^{-\lambda y})^m = \lambda^n e^{-\lambda(n\bar{x}+my)} \\ \Rightarrow\quad \ell(\lambda) &= n\log\lambda - \lambda(n\bar{x}+my) \\ \Rightarrow\quad \ell'(\lambda) &= \frac{n}{\lambda} - (n\bar{x}+my). \end{aligned}$$

Dakle, procjena za λ je rješenje:

$$\ell'(\lambda)=0 \quad\Leftrightarrow\quad \frac{n}{\lambda}-(n\bar{x}+my)=0 \quad\Leftrightarrow\quad \hat{\lambda}=\frac{1}{\bar{x}+\frac{m}{n}y}.$$

Zapis procjenitelja $\hat{\lambda}(\underline{X})$ je nešto složeniji. Naime, primijetite da su n i m slučajne vrijednosti (jedino je njihov zbroj, $n+m$, fiksan i jednak duljini slučajnog uzorka). □

Primjer 7.8 Iz 100000 polica autoodgovornosti izvučeni su podaci o štetama u jednoj godini. Frekvencijska tablica pokazuje nam brojeve polica po kojima je bilo 1, 2, 3, 4, te 5 i više šteta.

broj šteta	broj polica
0	81056
1	16174
2	2435
3	295
4	36
≥ 5	4
Σ	100000

Uz pretpostavku da se brojevi šteta po polici autoodgovornosti u godini dana ravnaju po Poissonovom zakonu razdiobe $P(\lambda)$, treba procijeniti nepoznati parametar λ. Primijetite da se radi o nepotpunom uzorku jer za 4 vrijednosti iz razreda "≥ 5" ne znate kolike su, ali znate da su veće od ili jednake $y=5$. Budući da je

$$\begin{aligned} f(k|\lambda) &= \mathbb{P}_\lambda(X=k)=\frac{\lambda^k}{k!}e^{-\lambda} \quad \text{za} \quad k=0,1,\ldots \\ \mathbb{P}_\lambda(X\geq 5) &= 1-\mathbb{P}_\lambda(X\leq 4)=1-e^{-\lambda}(1+\lambda+\frac{\lambda^2}{2}+\frac{\lambda^3}{6}+\frac{\lambda^4}{24}), \end{aligned}$$

vjerodostojnost, odnosno log-vjerodostojnost je

$$\begin{aligned} L(\lambda) &= f^{81056}(0|\lambda)\cdot f^{16174}(1|\lambda)\cdot f^{2435}(2|\lambda)\cdot f^{295}(3|\lambda)\cdot f^{36}(4|\lambda)\cdot(\mathbb{P}_\lambda(X\geq 5))^4 = \\ &= \frac{1}{2^{2435}\cdot 6^{295}\cdot 24^{36}}e^{-99996\lambda}\lambda^{22073}(1-e^{-\lambda}(1+\lambda+\frac{\lambda^2}{2}+\frac{\lambda^3}{6}+\frac{\lambda^4}{24}))^4 \\ \ell(\lambda) &= \log L(\lambda) = -100000\lambda+22073\log\lambda+4\log(e^\lambda-(1+\lambda+\frac{\lambda^2}{2}+\frac{\lambda^3}{6}+\frac{\lambda^4}{24}))-C, \end{aligned}$$

gdje je $C=\log(2^{2435}\cdot 6^{295}\cdot 24^{36})$. Procjena $\hat{\lambda}$ metodom ML je rješenje stacionarne jednadžbe:

$$\ell'(\lambda)=-100000+\frac{22073}{\lambda}+\frac{e^\lambda-(1+\lambda+\frac{\lambda^2}{2}+\frac{\lambda^3}{6})}{e^\lambda-(1+\lambda+\frac{\lambda^2}{2}+\frac{\lambda^3}{6}+\frac{\lambda^4}{24})}=0.$$

Očito se ta jednadžba mora rješavati numeričkim metodama. Da bi je riješili, na primjer, Newtonovom metodom, treba nam početna vrijednost za iteracije λ_0 koja se nalazi u okolini tražene točke maksimuma. Do nje možemo doći grubom procjenom za λ, a nju možemo dobiti ako 4 vrijednosti iz razreda "≥ 5" tretiramo kao da su točno jednake 5, pa primijenimo metodu momenata. Budući da je λ populacijsko očekivanje, (gruba) procjena $\tilde{\lambda}$ metodom momenata iznosi

$$\tilde{\lambda} = \bar{x} \approx \frac{81056 \cdot 0 + 16174 \cdot 1 + 2435 \cdot 2 + 295 \cdot 3 + 36 \cdot 4 + 4 \cdot 5}{100000} = 0.22093.$$

Dakle, uzmimo $\lambda_0 = 0.22093$. Newtonovom metodom dobijemo da je rješenje stacionarne jednadžbe log-vjerodostojnosti

$$\hat{\lambda} = 0.22078.$$

Primijetite da se gruba procjena dobivena metodom momenata i MLE razlikuju tek na četvrtoj decimali. Dakle, ako želimo točnost do na prve tri decimale, gruba procjena metodom momenata, za ovako veliki uzorak, je zadovoljavajuća. □

7.2.4 Nezavisni uzorci

Pretpostavimo da imamo dva nezavisna uzorka iz populacija čije razdiobe ovise o istom parametru. Tada je vjerodostojnost tog parametra na osnovi oba uzorka jednaka produktu vjerodostojnosti istog parametra, ali na osnovi svakog uzorka posebno.

7.3 Nepristranost

Dobra svojstva procjenitelja parametra su da je lociran blizu prave vrijednosti parametra, te da ima malu raspršenost. Ta svojstva proizlaze iz njegove uzoračke razdiobe.

Neka je $\underline{X} = (X_1, X_2, \ldots, X_n)$ slučajni uzorak iz populacije sa gustoćom $f(x|\theta)$. Kažemo da je procjenitelj $\hat{\theta}(\underline{X})$ *nepristran* za parametar θ ako je

$$\mathbb{E}[\hat{\theta}(\underline{X})] = \theta.$$

Svojstvo nepristranosti procjenitelja nije invarijantno na nelinearne transformacije parametara.

Čini se da je to dobro svojstvo procjenitelja. S druge strane, to nije bitno svojstvo za procjenitelja. Naime, u nekim situacijama pristran procjenitelj je bolji od nepristranog, pa i od najboljeg nepristranog procjenitelja. Od nepristranosti, bitnije je da procjenitelj ima malu *srednjekvadratnu grešku*.

7.4 Srednjekvadratna pogreška

Da bi procjenitelje mogli uspoređivati, trebamo mjeriti njihovu efikasnost. Mjera za efikasnost procjenitelja je srednjekvadratna pogreška.

Srednjekvadratna pogreška, kraće MSE, procjenitelja $\hat{\theta} = \hat{\theta}(\underline{X})$ za parametar θ je broj

$$\mathrm{MSE}(\hat{\theta}) := \mathbb{E}[(\hat{\theta}(\underline{X}) - \theta)^2].$$

Primijetite da je $\mathrm{MSE}(\hat{\theta})$ funkcija parametra θ. Nadalje, MSE je drugi moment od $\hat{\theta}(\underline{X})$ oko θ. U slučaju da je $\hat{\theta}(\underline{X})$ nepristrani procjenitelj za θ, $\mathrm{MSE}(\hat{\theta}) = \mathbb{V}\mathrm{ar}[\hat{\theta}(\underline{X})]$.

Ako *pristranost* od $\hat{\theta}(\underline{X})$ definiramo kao broj

$$b(\hat{\theta}) := \mathbb{E}[\hat{\theta}(\underline{X})] - \theta,$$

tada vrijedi da je

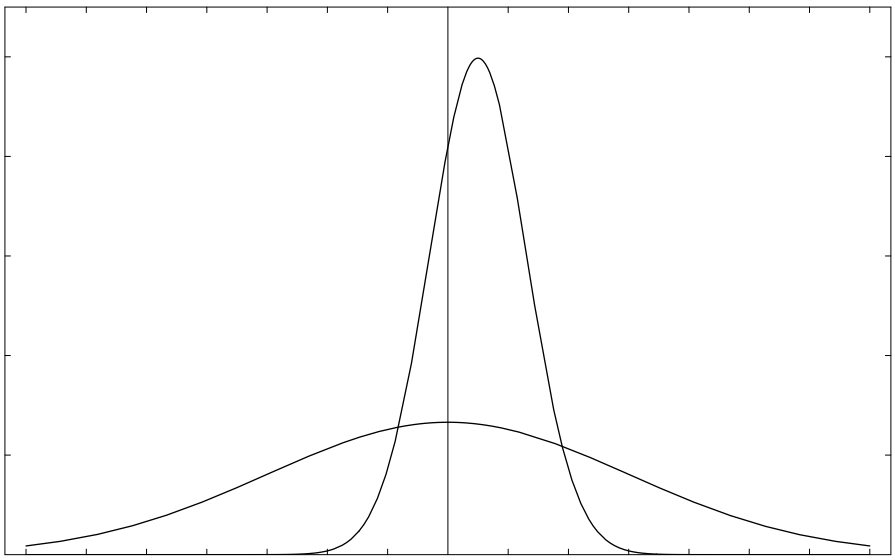
$$\mathrm{MSE}(\hat{\theta}) = \mathbb{V}\mathrm{ar}[\hat{\theta}(\underline{X})] + b^2(\hat{\theta}).$$

Dokaz.

$$\begin{aligned}
\mathrm{MSE}(\hat{\theta}) &= \mathbb{E}[(\hat{\theta}(\underline{X}) - \theta)^2] = \mathbb{E}[((\hat{\theta}(\underline{X}) - \mathbb{E}[\hat{\theta}(\underline{X})]) + (\mathbb{E}[\hat{\theta}(\underline{X})] - \theta))^2] = \\
&= \mathbb{E}[(\hat{\theta}(\underline{X}) - \mathbb{E}[\hat{\theta}(\underline{X})])^2 + 2(\hat{\theta}(\underline{X}) - \mathbb{E}[\hat{\theta}(\underline{X})])(\mathbb{E}[\hat{\theta}(\underline{X})] - \theta) + (\mathbb{E}[\hat{\theta}(\underline{X})] - \theta)^2] = \\
&= \mathbb{V}\mathrm{ar}[\hat{\theta}(\underline{X})] + 0 + b^2(\hat{\theta}) = \\
&= \mathbb{V}\mathrm{ar}[\hat{\theta}(\underline{X})] + b^2(\hat{\theta}).
\end{aligned}$$

□

Na slici se vidi odnos uzoračkih distribucija jednog pristranog procjenitelja male srednjekvadratne pogreške i jednog nepristranog procjenitelja velike varijance (odnosno, srednjekvadratne pogreške). Vertikalna crta označava vrijednost parametra. To je primjer kada je pristrani procjenitelj bolji od nepristranog.



Kažemo da je procjenitelj *konzistentan*, odnosno *asimptotski nepristran*, ako njegova srednjekvadratna greška teži ka nuli kada veličina uzorka raste u beskonačnost:

$$\mathrm{MSE}(\hat{\theta}) \to 0, \quad n \to \infty.$$

Na primjer, $\overline{X}$ je konzistetan procjenitelj za populacijsko očekivanje.

7.5 Asimptotska razdioba od MLE

Za MLE $\hat{\theta}$ parametra θ na osnovi uzorka $\underline{X}$ duljine n iz populacije s populacijskom gustoćom $f(x|\theta)$ varijable X, vrijedi

$$\hat{\theta} \approx N(\theta, \mathrm{CRlb}) \text{ za veliko } n, \tag{7.2}$$

gdje je

$$\mathrm{CRlb} = \frac{1}{n\mathbb{E}[(\frac{\partial}{\partial\theta}\log f(X|\theta))^2]}$$

Cramer-Raova donja granica. Primijetite da je slučajna variabla $\frac{\partial}{\partial\theta}\log f(X|\theta)$ dobijena kao kompozicija funkcije $x \mapsto \frac{\partial}{\partial\theta}\log f(x|\theta)$ i varijable X. Taj rezultat vrijedi pod vrlo općenitim uvjetima na populacijsku razdiobu. Jedina bitna restrikcija je da $\mathrm{Im}X$ ne smije ovisiti o θ. Nadalje, pokazuje se da je pod skoro istim uvjetima Cramer-Raova donja granica najmanja asimptotska varijanca koju konzistentan procjenitelj može imati. Kažemo da je uz te uvjete MLE *asimptotski efikasan* procjenitelj parametara.

Cramer-Raovu donju granicu možemo izraziti i pomoću *slučajne* log-vjerodostojnosti $\ell(\theta|\underline{X}) = \sum_{i=1}^{n}\log f(X_i|\theta)$:

$$\mathrm{CRlb} = \frac{1}{\mathbb{E}[(\frac{\partial\ell}{\partial\theta}(\theta|\underline{X}))^2]} = \frac{1}{-\mathbb{E}[\frac{\partial^2\ell}{\partial\theta^2}(\theta|\underline{X})]}.$$

Asimptotski rezultat (7.2) koristi se za konstrukciju aproksimativnih *pouzdanih intervala*.

7.6 Završne napomene

Općenito se smatra da je metoda maksimalne vjerodostojnosti bolja od metode momenata. Naime, procjenitelji dobiveni metodom momenata su uvijek funkcije uzoračkih momenata što je veliko ograničenje na klasu procjenitelja. Nadalje, ta metoda može dovesti do neprihvatljivih procjenitelja u smislu da procjene mogu biti izvan područja razumnih vrijednosti parametara. S druge strane, u primjenama na najčešće modele razdioba (binomni, Poissonov, eksponencijalni, normalni model), obje metode daju iste procjenitelje.

U nekim situacijama, na primjer u slučaju gama razdiobe $\Gamma(\alpha,\beta)$ kada su oba parametra nepoznata, metoda momenta je u prednosti jer daje procjenitelje na jednostavan način i u eksplicitnoj formi, dok se MLE mora računati numeričkim metodama.

Poglavlje 8

Pouzdani intervali

Pouzdanim intervalim mjeri se točnost (preciznost) procjenitelja. Neka je $\underline{X}$ slučajni uzorak iz populacije s jednodimenzionalnim parametrom θ. $(1-\alpha)\cdot 100\%$-*pouzdan interval* za θ je slučajni interval $[\hat{\theta}_1(\underline{X}), \hat{\theta}_2(\underline{X})]$ takav da je

$$\mathbb{P}(\hat{\theta}_1(\underline{X}) \leq \theta \leq \hat{\theta}_2(\underline{X})) = 1-\alpha.$$

Kažemo da je *vjerojatnost pokrivanja* intervala $[\hat{\theta}_1(\underline{X}), \hat{\theta}_2(\underline{X})]$ jednaka $1-\alpha$. U primjenama se najčešće koriste 95%-pouzdani intervali ($\alpha = 0.05$). Dakle, 95%-pouzdani intervali za θ $[\hat{\theta}_1(\underline{X}), \hat{\theta}_2(\underline{X})]$ određeni su vjerojatnošću pokrivanja od 0.95:

$$\mathbb{P}(\hat{\theta}_1(\underline{X}) \leq \theta \leq \hat{\theta}_2(\underline{X})) = 0.95.$$

Primijetite da ovdje θ predstavlja pravu (dakle, ne slučajnu) vrijednost parametra, dok su granice pouzdanog intervala statistike. 95%-pouzdan interval interpretiramo na sljedeći način. U 95% svih realizacija tog intervala, prava vrijednost od θ će se nalaziti unutar njihovih granica, a u 5% realizacija to neće biti slučaj. Ovdje se riječ "svih" odnosi na vrlo velik broj realizacija (opažaja).

Pouzdan interval nije jedinstven. Općenito, konstruira se pomoću uzoračke distribucije dobrog procjenitelja parametra, na primjer, pomoću MLE. I u tom slučaju treba izabrati između *jednostranog* ili *dvostranog* pouzdanog intervala, *jednakorepnog* i/ili najkraće duljine. U slučaju simetrične uzoračke distribucije oko prave vrijednosti parametra, jednakorepni i pouzdani intervali najkraće duljine se podudaraju.

8.1 Konstrukcija pouzdanih intervala

8.1.1 Pivotna metoda

Opća metoda konstrukcije pouzdanih intervala zove se *pivotna* metoda. Osnovna pretpostavka za primjenu te metode je da postoji *pivotna veličina* $g(\underline{X}, \theta)$ sa sljedećim svojstvima:

1. funkcija je uzorka i parametra;
2. njen zakon razdiobe je poznat;
3. strogo je monotona kao funkcija po parametru.

Budući da je razdioba od $g(\underline{X}, \theta)$ poznata, odredimo brojeve g_1, g_2 tako da vrijedi:

$$\mathbb{P}(g_1 \leq g(\underline{X}, \theta) \leq g_2) = 0.95. \tag{8.1}$$

Na primjer, ako je $\theta \mapsto g(\underline{X}, \theta)$ strogo rastuća funkcija, tada je

$$\begin{aligned} g_1 \leq g(\underline{X}, \theta) &\Leftrightarrow \hat{\theta}_1(\underline{X}) \leq \theta \\ g_2 \geq g(\underline{X}, \theta) &\Leftrightarrow \hat{\theta}_2(\underline{X}) \geq \theta, \end{aligned}$$

pa je (8.1) ekvivalentno sa

$$\mathbb{P}(\hat{\theta}_1(\underline{X}) \leq \theta \leq \hat{\theta}_2(\underline{X})) = 0.95.$$

Dakle, $[\hat{\theta}_1(\underline{X}), \hat{\theta}_2(\underline{X})]$ je 95%-pouzdani interval za θ.

U gotovo svim važnim slučajevima koji se koriste u primjenama, pivotna veličina postoji. U nekim slučajevima, na primjer, za populacije s binomnom i Poissonovom razdiobom, do pivotne veličine dolazimo jedino kada egzaktnu uzoračku razdiobu aproksimiramo normalnom (npr. kada su uzorci veliki).

Primjer 8.1 Neka je $\underline{X}$ slučajni uzorak duljine $n = 20$ iz normalno $N(\mu, 10^2)$-distribuirane populacije. Opažena vrijednost uzoračke sredine je $\bar{x} = 62.75$. Za pivotnu veličinu uzmemo standardiziranu verziju od $\overline{X}$:

$$g(\underline{X}, \mu) = \frac{\overline{X} - \mu}{10}\sqrt{20}.$$

Znamo da ta slučajna varijabla ima jediničnu normalnu razdiobu i očito je strogo padajuća po μ. Iz tablica slijedi

$$\begin{aligned} \mathbb{P}(-1.96 \leq \frac{\overline{X} - \mu}{10}\sqrt{20} \leq 1.96) &= 0.95 \\ \Leftrightarrow \quad \mathbb{P}(\overline{X} - 1.96 \cdot \frac{10}{\sqrt{20}} \leq \mu \leq \overline{X} + 1.96 \cdot \frac{10}{\sqrt{20}}) &= \mathbb{P}(\overline{X} - 4.38 \leq \mu \leq \overline{X} + 4.38) = 0.95 \end{aligned}$$

Dakle, 95%-pouzdan interval za μ je $[\overline{X} - 4.38, \overline{X} + 4.38]$. Taj interval je jednakorepni jer je

$$\mathbb{P}(\mu < \overline{X} - 4.38) = \mathbb{P}(\mu > \overline{X} + 4.38) = 0.0025.$$

Budući da je uzoračka razdioba od $g(\underline{X}, \mu)$ simetrična, dobiveni interval je najkraće duljine. Kraće taj interval možemo zapisati pomoću njegovih granica:

$$\overline{X} \pm 4.38.$$

Opažena vrijednost tog intervala je $[58.37, 67.13]$ ili, pomoću granica, 62.75 ± 4.38. Drugi 95%-pouzdan interval za μ možemo dobiti ako uzmemo

$$\begin{aligned} \mathbb{P}(-1.881 \leq \frac{\overline{X} - \mu}{10}\sqrt{20} \leq 2.054) &= 0.95 \\ \Rightarrow \quad \mathbb{P}(\overline{X} - 4.21 \leq \mu \leq \overline{X} + 4.59) &= 0.95. \end{aligned}$$

Novodobiveni pouzdani interval nije najkraće duljine, pa u obzir uzimamo samo prvoizvedeni jednakorepni 95%-pouzdani interval. Spomenimo još da su jednostrani 95%-pouzdani intervali za μ:

$$\langle -\infty,\ \overline{X} + 1.64 \cdot \frac{10}{\sqrt{20}}], \quad [\overline{X} - 1.64 \cdot \frac{10}{\sqrt{20}},\ +\infty\rangle.$$

□

8.1.2 Pouzdane granice

95%-pouzdani interval za parametar μ normalno $N(\mu, \sigma^2)$-distribuirane populacija s poznatom varijancom σ^2, ili aproksimativni 95%-pouzdani interval za parametar očekivanja μ s konzistentno procjenjenom vrijednosti za populacijsku standardnu devijaciju σ je

$$[\overline{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}}, \overline{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}].$$

Taj interval se može alternativno zapisati pomoću svojih *pouzdanih granica*:

$$\overline{X} \pm 1.96 \cdot \frac{\sigma}{\sqrt{n}}.$$

Prednost takvog zapisa je njegova informativnost. Naime, iz njega čitamo točkovnu procjenu za μ i preciznost te procjene (uz pouzdanost od 95%). Nadalje, jednostrani pouzdani intervali odgovaraju gornjoj, odnosno donjoj pouzdanoj granici.

8.1.3 Veličina uzorka

Statističarima se često postavlja pitanje o potrebnoj veličini uzorka. Za odgovor na to pitanje, potrebne su sljedeće informacije:

1. kolika je preciznost procjene potrebna;
2. koliko iznosi (barem približno) standardna devijacija σ.

Informacija o standardnoj devijaciji nije uvijek na raspolaganju. Često se do nje dolazi procjenom na osnovi izv. pilot-uzoraka.

Primjer 8.2 Osiguravajuće društvo želi procijeniti srednju vrijednost iznosa šteta po policama određene klase, a koje su nastale tijekom prošle godine. Iscrpni podaci o istovrsnim štetama iz prethodnih godina sugeriraju da bi standardna devijacija prošlogodišnjih iznosa šteta mogla biti oko 450 kn. Srednju vrijednost prošlogodišnjih iznosa šteta treba procijeniti do na ± 50 kn točnosti uz 95% pouzdanosti. Koliko veliki uzorak treba uzeti? Označimo sa μ srednju vrijednost koju želimo procijeniti. 95%-pouzdani interval za μ će imati granice $\overline{X} \pm 1.96 \cdot \frac{\sigma}{\sqrt{n}}$, gdje je σ standardna devijacija iznosa prošlogodišnjih šteta. Prema informacijama kojima raspolažemo, $\sigma \approx 450$. Tada iz zahtjeva na preciznost imamo:

$$1.96 \cdot \frac{450}{\sqrt{n}} \leq 50 \quad \Rightarrow \quad \sqrt{n} \geq 1.96 \cdot \frac{450}{50} \quad \Rightarrow \quad n \geq 312.$$

Dakle, odabiremo slučajni uzorak veličine $n = 320$ (malo veći od dobivene vrijednosti jer je procjena od σ gruba). □

8.2 Pouzdani intervali za parametre normalno distribuirane populacije

8.2.1 Populacijska sredina

U proslom smo potpoglavlju konstruirali pouzdani interval za parametar očekivanja normalne populacije kada je populacijska standardna devijacija poznata. To je nerealna situacija. Dakle, trebamo pivotnu veličinu za situaciju kada su oba parametra nepoznata.

Ta veličina je studentizirana verzija uzoračke sredine za koju znamo da za normalnu populaciju ima Studentovu t-distribuciju sa $n-1$ stupnjem slobode (n je veličina uzorka) i očito je strogo padajuća po μ:

$$\frac{\overline{X}-\mu}{S}\sqrt{n} \sim t(n-1).$$

Rezultirajući 95%-pouzdani interval za μ najkraće duljine je

$$\overline{X} \pm t_{0.025}(n-1) \cdot \frac{S}{\sqrt{n}}, \tag{8.2}$$

gdje je $t_{0.025}(n-1)$ tablična kritična vrijednost opisana u 6.3. Budući da je za velike k $t(k) \rightsquigarrow N(0,1)$, za velike uzorke možemo uzeti $t_{0.025}(n-1) \approx 1.96$. Normalna distribuiranost populacije je bitna pretpostavka da bi sa (8.2) bio dan 95%-pouzdan interval za parametar očekivanja, pogotovo za male uzorke. S druge strane, pokazuje se da je taj interval robustan za populacijske distribucije koje odstupaju od normalnosti, pogotovo za velike uzorke. Normalnost uzorka se može provjeriti uvidom u, na primjer, dijagram točaka. Na taj se način može detektirati velika asimetrija i prisustvo tzv. "outliera" (stršećih vrijednosti) koje mogu bitno utjecati na analizu.

8.2.2 Populacijska varijanca

Za konstrukciju pouzdanih intervala za parametar varijance σ^2 (i standardne devijacije σ) koristimo statistiku

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

za pivotnu veličinu. Rezultirajući jednakorepni 95%-pouzdani intervali su:

$$\text{za } \sigma^2: \quad \left[\frac{(n-1)S^2}{\chi^2_{0.025}(n-1)}, \frac{(n-1)S^2}{\chi^2_{0.975}(n-1)}\right], \quad \text{za } \sigma: \quad \left[\sqrt{\frac{(n-1)S^2}{\chi^2_{0.025}(n-1)}}, \sqrt{\frac{(n-1)S^2}{\chi^2_{0.975}(n-1)}}\right],$$

gdje su tablične kritične vrijednosti $\chi^2_{0.025}(k)$ i $\chi^2_{0.975}(k)$ definirane na sličan načina kao kritične vrijednosti za t i F-razdiobu u poglavlju 6 [1]. Primijetite da zbog pozitvne asimetrije od χ^2-razdiobe ti intervali nisu simetrični oko procjena, pa ne moraju biti najmanje duljine.

Za konstrukciju ovih pouzdanih intervala, normalna distribuiranost populacije je bitna pretpostavka za relativno male uzorke. Slično kao i u slučaju pouzdanih intervala za parametar očekivanja, i ovi intervali su robustni na odstupanja populacije od normalnosti za dovoljno velike uzorke.

8.3 Pouzdani intervali za parametre binomne i Poissonove razdiobe

Budući da su binomna i Poissonova diskretne razdiobe, nije uvijek moguće postići da vjerojatnost pokrivanja $(1-\alpha)\cdot 100\%$-pouzdanog intervala $[\hat{\theta}_1(\underline{X}), \hat{\theta}_2(\underline{X})]$ za neki njihov parametar θ bude točno jednaka $1-\alpha$. Zato je dovoljno zahtjevati da iznosi *barem* $1-\alpha$:

$$\mathbb{P}(\hat{\theta}_1(\underline{X}) \leq \theta \leq \hat{\theta}_2(\underline{X})) \geq 1-\alpha.$$

[1] Općenito, $\chi^2_\alpha(k)$ je $(1-\alpha)$-kvantil χ^2-razdiobe s k stupnjeva slobode, odnosno $\mathbb{P}(\chi^2_\alpha(k) \geq H) = \alpha$ ako je $H \sim \chi^2(k)$.

8.3.1 Vjerojatnost uspjeha u binomnoj razdiobi

Ako X ima binomnu razdiobu s parametrima (n, θ), gdje je θ vjerojatnost uspjeha, MLE za θ je relativna frekvencija uspjeha:

$$\hat{\theta} = \frac{X}{n}.$$

X je, u stvari, zbroj n.j.d. Bernoullijevih varijabli koje čine slučajni uzorak iz populacije s Bernoullijevom distribucijom.

Konstrukcija pouzdanog intervala za parametar θ zasniva se na statistici X koju ne možemo uzeti za pivotnu veličinu jer ne ovisi o θ. Ali zato njezina funkcija vjerojatnosti ovisi o θ. Neka je x opažena vrijednost od X. Tada 95%-pouzdani interval za θ odredimo iz uvjeta da je

$$\mathbb{P}_\theta(X \leq x) \geq 0.025, \quad \mathbb{P}_\theta(X \geq x) \geq 0.025. \tag{8.3}$$

Preciznije, realizacija $[\hat{\theta}_1(x), \hat{\theta}_2(x)]$ 95%-pouzdanog intervala je rješenje tog sustava nejednadžbi. Dakle, $[\hat{\theta}_1(X), \hat{\theta}_2(X)]$ je 95%-pouzdani interval za θ. Neka je

$$F(x|\theta) = \sum_{k=0}^{x} \binom{n}{k} \theta^k (1-\theta)^k, \;\; x = 0, 1, \ldots, n,$$

funkcija distribucije od X. Tada se sustav nejednadžbi (8.3) može zapisati u obliku:

$$F(x|\theta) \geq 0.025, \quad 1 - F(x-1|\theta) \geq 0.025.$$

Nije teško pokazati da je za $0 < x < n$ funkcija $\theta \mapsto F(x|\theta)$ strogo padajuća, a $\theta \mapsto 1 - F(x-1|\theta)$ strogo rastuća funkcija za $\theta \in \langle 0, 1\rangle$. Tada su granice pouzdanog intervala $[\hat{\theta}_1, \hat{\theta}_2]$ rješenja jednadžbi:

$$F(x|\hat{\theta}_2) = 0.025, \quad 1 - F(x-1|\hat{\theta}_1) = 0.025.$$

Očito je da se te jednadžbe moraju numerički rješavati.

Ako je n velik, pouzdani se intervali mogu dobiti pomoću normalne aproksimacije binomne razdiobe. Tada je pivotna veličina standardizirana verzija od X:

$$\frac{X - n\theta}{\sqrt{n\theta(1-\theta)}} \approx N(0, 1).$$

U primjenama se konstrukcija sprovodi pomoću nešto jednostavnije verzije navedene pivotne veličine:

$$\frac{X - n\theta}{\sqrt{n\hat{\theta}(1-\hat{\theta})}} \approx N(0, 1),$$

dakle, one koja je dobivena iz početne zamjenom nepoznate vrijednosti od θ u izrazu za standardnu devijaciju (nazivnik standardizirane verzije) sa procjenom $\hat{\theta}$. Na taj način dolazimo do aproksimativnog 95%-pouzdanog intervala za θ:

$$\hat{\theta} \pm 1.96 \cdot \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}.$$

8.3.2 Parametar Poissonove razdiobe

Konstrukcije pouzdanih intervala za slučajeve malog i velikog uzorka su slične kao za vjerojatnost uspjeha u binomnoj razdiobi.

Neka je $\underline{X} = (X_1, X_2, \ldots, X_n)$ slučajni uzorak iz $P(\lambda)$-distribuirane populacije. Tada je $Y := X_1 + X_2 + \cdots + X_n \sim P(n\lambda)$. MLE za λ je

$$\hat{\lambda} = \frac{Y}{n} = \overline{X}.$$

Ako je uzorak mali (n je mali broj), 95%-pouzdani interval za λ dobije se rješavanjem sustava nejednadžbi

$$F_Y(y|\lambda) \geq 0.025, \quad 1 - F_Y(y-1|\lambda) \geq 0.025,$$

gdje je y opažena vrijednost od Y, a F_Y funkcija distribucije od Y:

$$F_Y(y|\lambda) = \sum_{k=0}^{y} \frac{(n\lambda)^k}{k!} e^{-n\lambda}, \;\; y = 0, 1, 2, \ldots$$

Pokazuje se da je funkcije $\lambda \mapsto F_Y(y|\lambda)$ strogo padajuća na $\langle 0, +\infty \rangle$, pa za granice $\hat{\lambda}_1 = \hat{\lambda}_1(y)$, $\hat{\lambda}_2 = \hat{\lambda}_2(y)$ 95%-pouzdanog intervala $[\hat{\lambda}_1(Y), \hat{\lambda}_2(Y)]$ vrijedi da su rješenja jednadžbi:

$$F_Y(y|\hat{\lambda}_2) = 0.025, \quad 1 - F_Y(y-1|\hat{\lambda}_1) = 0.025.$$

Ako je n velik broj, tada za konstrukciju aproksimativnog 95%-pouzdanog intervala za λ koristimo standardiziranu verziju statistike $\overline{X}$,

$$\frac{\overline{X} - \lambda}{\sqrt{\lambda}} \sqrt{n} \overset{\cdot}{\sim} N(0, 1),$$

kao pivotne veličine, odnosno njenu modificiranu verziju

$$\frac{\overline{X} - \lambda}{\sqrt{\hat{\lambda}}} \sqrt{n} \overset{\cdot}{\sim} N(0, 1).$$

Dakle, aproksimativni 95%-pouzdani interval za λ je

$$\hat{\lambda} \pm 1.96 \cdot \sqrt{\frac{\hat{\lambda}}{n}}.$$

8.4 Pouzdani intervali za probleme s dva uzorka

Usporedba parametara dviju populacija obično se bazira na nezavisnim uzorcima iz tih populacija. Na primjer, ako se radi o parametrima očekivanja, jasno je da ćemo ih uspoređivati pomoću njihovih procjena, odnosno statistika $\overline{X}_1$ i $\overline{X}_2$. Važnost pretpostavke o nezavisnosti uzoraka sada se vidi iz sljedećih relacija. Ako su uzorci nezavisni, tada je

$$\mathbb{V}\mathrm{ar}[\overline{X}_1 - \overline{X}_2] = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2},$$

a ako su zavisni moramo uzeti u obzir njihovu korelaciju:

$$\mathbb{V}\mathrm{ar}[\overline{X}_1 - \overline{X}_2] = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} - 2\,\mathrm{cov}[\overline{X}_1, \overline{X}_2].$$

Brojevi n_1 i n_2 označavaju duljine prvog, odnosno drugog uzorka, a σ_1^2 i σ_2^2 su populacijske varijance.

Najčešća forma zavisnosti je kada imamo *sparene* podatke.

8.4.1 Usporedba očekivanja normalno distribuiranih populacija

Ako su $\overline{X}_1$ i $\overline{X}_2$ uzoračke sredine dvaju nezavisnih uzoraka duljina n_1, n_2 iz normalno distribuiranih populacija s poznatim varijancama σ_1^2 i σ_2^2, tada je jednakorepni (a time i najkraće duljine) 95%-pouzdani interval za razliku $\mu_1 - \mu_2$ populacijskih očekivanja, jednak

$$\overline{X}_1 - \overline{X}_2 \pm 1.96 \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

Navedeni interval će biti aproksimativni 95%-pouzdani interval za razliku parametara očekivanja općenito bilo kojih populacija, pri čemu umjesto populacijskih varijanci mogu stajati njihove konzistentne procjene, uz uvjet da je uzorak dovoljno velik.

Ako su populacijske varijance nepoznate, ali pretpostavljamo da su jednake, dakle da je $\sigma_1^2 = \sigma_2^2 = \sigma^2$, tada je jednakorepni 95%-pouzdani interval najkraće duljine za razliku $\mu_1 - \mu_2$ normalnih populacija:

$$\overline{X}_1 - \overline{X}_2 \pm t_{0.025}(n_1 + n_2 - 2) \cdot S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

gdje je

$$S_p^2 := \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

procjenitelj zajedničke varijance σ^2. S_p^2 je dobiven metodom ML, ali je normiran tako da bude nepristrani procjenitelj za σ^2. Statistiku S_p^2 zovemo *zajednička uzoračka varijanca*.

8.4.2 Usporedba varijanci normalno distribuiranih populacija

Za usporedbu populacijskih varijanci, gledamo njihov omjer σ_1^2/σ_2^2, a ne razliku. To se može opravdati koncepcijom varijance, ali i tehničkim razlozima. Naime, postoji pivotna veličina pomoću koje se mogu konstruirati pouzdani intervali za omjer σ_1^2/σ_2^2, a na osnovi dva nezavisna uzorka duljina n_1, n_2 iz normalno distribuiranih populacija:

$$\frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1).$$

Jednakorepni (ne nužno i najkraći) 95%-pouzdani interval za σ_1^2/σ_2^2 je

$$\left[\frac{S_1^2}{S_2^2} \cdot \frac{1}{f_{0.025}(n_1 - 1, n_2 - 1)},\ \frac{S_1^2}{S_2^2} \cdot f_{0.025}(n_2 - 1, n_1 - 1) \right].$$

Kritične su vrijednosti $f_{0.025}(\nu_1, \nu_2)$ objašnjene u potpoglavlju 6.5.

8.4.3 Usporedba populacijskih proporcija

Usporedba populacijskih proporcija odgovara usporedbi vjerojatnosti uspjeha dviju Bernoullijevih populacija čiji su nezavisni uzorci opisani dvama nezavisnim binomnom slučajnim varijablama, X_1 s parametrima (n_1, θ_1) i X_2 s parametrima (n_2, θ_2). Razmotrit ćemo samo slučaj velikih uzoraka (n_1 i n_2 su veliki brojevi) i konstrukcije aproksimativnih pouzdanih intervala. Konstrukcija se bazira na pivotnoj veličini

$$\frac{(\hat{\theta}_1 - \hat{\theta}_2) - (\theta_1 - \theta_2)}{\sqrt{\frac{\hat{\theta}_1(1-\hat{\theta}_1)}{n_1} + \frac{\hat{\theta}_2(1-\hat{\theta}_2)}{n_2}}} \approx N(0, 1),$$

gdje su $\hat{\theta}_1 = X_1/n_1$ i $\hat{\theta}_2 = X_2/n_2$ MLE parametara θ_1 i θ_2. Dakle, aproksimativni 95%-pouzdani interval za $\theta_1 - \theta_2$ je

$$\hat{\theta}_1 - \hat{\theta}_2 \pm 1.96 \cdot \sqrt{\frac{\hat{\theta}_1(1-\hat{\theta}_1)}{n_1} + \frac{\hat{\theta}_2(1-\hat{\theta}_2)}{n_2}}.$$

8.4.4 Usporedba dva Poissonova parametra

Slično kao i za proporcije, razmotrit ćemo samo slučaj velikih uzoraka i konstrukcije aproksimativnih pouzdanih intervala. Neka su $\hat{\lambda}_1 = \overline{X}_1$ i $\hat{\lambda}_2 = \overline{X}_2$ MLE parametara λ_1 i λ_2 dviju populacija s Poissonovim razdiobama, a na osnovi nezavisnih uzoraka uzetih iz tih populacija. Konstrukcija se bazira na pivotnoj veličini

$$\frac{(\hat{\lambda}_1 - \hat{\lambda}_2) - (\lambda_1 - \lambda_2)}{\sqrt{\frac{\hat{\lambda}_1}{n_1} + \frac{\hat{\lambda}_2}{n_2}}} \dot{\sim} N(0,1).$$

Dakle, aproksimativni 95%-pouzdani interval za $\lambda_1 - \lambda_2$ je

$$\hat{\lambda}_1 - \hat{\lambda}_2 \pm 1.96 \cdot \sqrt{\frac{\hat{\lambda}_1}{n_1} + \frac{\hat{\lambda}_2}{n_2}}.$$

8.5 Spareni podaci

Spareni su podaci uobičajen primjer zavisnih uzoraka. Pretpostavimo da imamo slučajan uzorak iz dvodimenzionalne razdiobe vektora (X, Y):

$$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n).$$

Interpretacija varijabli X_i i Y_i u paru (X_i, Y_i) je da su to po tipu iste varijable, X_i se mjeri prije nečega, recimo nekog tretmana, a Y_i nakon tretmana, na istoj statističkoj jedinici i $(i = 1, 2, \ldots, n)$. Zbog toga je prirodno na sparene podatke gledati kao na jedan dvodimenzionalan uzorak, a ne kao na dva odvojena uzorka $(X_1, X_2, \ldots, X_2)$ i $(Y_1, Y_2, \ldots, Y_n)$. Budući da nas zanima *srednji* doprinos tretmana, analiziraju se razlike $D_i := X_i - Y_i$, $i = 1, 2, \ldots, n$. Dakle, ako sa $D := X - Y$ označimo varijablu čiju populacijsku srednju vrijednost $\mu_D := \mu_1 - \mu_2$ želimo procijeniti, tada je baza za procjenu izvedeni slučajni uzorak $\underline{D} = (D_1, D_2, \ldots, D_n)$.

Ako je $\underline{D}$ uzorak iz normalno distribuirane populacije, tada je pivotna veličina za konstrukciju pouzdanih intervala za μ_D:

$$\frac{\overline{D} - \mu_d}{S_D}\sqrt{n} \sim t(n-1).$$

Dakle, 95%-pouzdani interval za μ_D je

$$\overline{D} \pm t_{0.025}(n-1)\frac{S_D}{\sqrt{n}}.$$

Za velike n je $t_{0.025}(n-1) \approx 1.96$. U tom slučaju navedeni interval dobro aproksimira 95%-pouzdani interval za μ_D i ako populacija nije normalno distribuirana.

Praktične napomene:

1. Uvijek je dobro ispitati da li su podaci koje čine dva uzorka u stvari spareni podaci. To se može učiniti grafički pomoću *dijagrama raspršenja* ili računajući koeficijent korelacije. Ukoliko je zavisnost znatna, treba provjeriti izvor podataka da se vidi jesu li podaci spareni u skladu sa dizajnom uzorkovanja.

2. Analiza sparenih podataka kao nezavisnih vodi pogrešnim zaključcima jer se bazira na krivim pretpostavkama. S druge strane, ako nezavisne podatke analiziramo kao sparene, grešaka neće biti, ali će metode biti neefikasne.

Poglavlje 9

Testiranje statističkih hipoteza

9.1 Hipoteze, testne statistike, odluke i pogreške

Statistička hipoteza je pretpostavka o populacijskoj razdiobi promatrane varijable. U slučaju *parametarskih* modela za populacijske razdiobe, to će biti bilo koja izjava o vrijednostima parametara. U ovom poglavlju uglavnom ćemo razmatrati takve statističke hipoteze. Osnovna hipoteza koja se testira zove se *nulhipoteza* i označava se sa H_0. Nulhipoteza često predstavlja aktualno znanje o vrijednostima parametara ili neutralnu izjavu. Na primjer, želimo li testirati postojanje razlike između dva populacijska parametra, nulhipoteza bi bila da nema nikakve razlike. Uz nulhipotezu, postavlja se i njoj *alternativna hipoteza* koju označavamo sa H_1. Ako se hipotezom u potpunosti (jednoznačno) određuje populacijska razdioba, tada se takva hipoteza zove *jednostavnom*, u suprotnom se zove *složena* hipoteza.

Statistički test je pravilo podjele prostora vrijednosti uzoraka na dva podskupa: na područje vrijednosti uzoraka koji su konzistentni sa H_0, i na njegov komplement u kojemu se nalaze vrijednosti nekonzistentne sa H_0.

Testovi kojima ćemo se baviti dizajnirani su za odgovaranje na pitanje: "Da li podaci daju dovoljno dokaza za odbacivanje H_0?" Odluka o odbacivanju ili ne odbacivanju nulhipoteze donosi se na osnovi vrijednosti *testne statistike*. Područje vrijednosti koje testna statistika poprima (dakle, njezina slika) dijeli se na područje vrijednosti koje su konzistentne sa H_0 i na područje nekonzistentno sa H_0. Područje testne statistike koje je nekozistentno sa H_0 zove se *kritično područje*. Dakle, ako se opažena vrijednost testne statistike nalazi u kritičnom području, H_0 se odbacuje (u korist H_1).

Razina značajnosti testa α je vjerojatnost odbacivanja H_0 ako je H_0 istinita hipoteza. Za slučaj kada odbacimo H_0 ako je H_0 istinito, kažemo da smo počinili *pogrešku prve vrste*. *Pogrešku druge vrste* činimo kada ne odbacujemo H_0, a H_1 je istinito. Vjerojatnost pogreške druge vrste označavamo sa β. Idealni test bi bio onaj za koji bi bilo moguće vjerojatnosti obiju grešaka učiniti po volji malima. Takav test ne postoji.

9.2 Klasično testiranje, značajnost i p-vrijednosti

9.2.1 "Najbolji" testovi

Klasični pristup nalaženja "dobrog" testa, tzv. *Neyman-Pearsonova teorija*, polazi od fiksne razine značajnosti α i konstruira test za koji vrijedi da je pogreška druge vrste β najmanja moguća za sve parametre specificirane alternativnom hipotezom H_1. Ključni rezultat u toj teoriji je Neyman-Pearsonova lema koja daje najbolji test (najmanji β uz fiksno α) u

slučaju kada su obje hipoteze, nulhipoteza i alternativna, jednostavne hipoteze. Za zadanu razinu značajnosti, kritično se područje (a time i testna statistika) za najbolji test, odredi kao skup onih vrijednosti uzoraka za koje vrijedi da je omjer L_0/L_1 vjerodostojnosti L_0 uz H_0 i L_1 uz H_1, izraženih kao funkcije uzoraka, ograničen odozgo nekom konstantom. Preciznije, ako je C kritično područje veličine α, te ako postoji konstanta k takva da za sve točke iz C vrijedi $\frac{L_0}{L_1} \leq k$, a da za točke izvan C vrijedi $\frac{L_0}{L_1} > k$, tada je C najjače kritično područje veličine α za testiranje jednostavne osnovne hipoteze $H_0 : \theta = \theta_0$ u odnosu na jednostavnu alternativnu hipotezu $H_1 : \theta = \theta_1$ ($L_0 \equiv L(\theta_0), L_1 \equiv L(\theta_1)$).

Ako je barem jedna od hipoteza H_0 i/ili H_1 složena, tada samo u specijalnim slučajevima, na primjer kod jednostranih testova, postoji test koji je najbolji za sve parametre. U slučajevima kada najbolji test u smislu Neymman-Pearsonove teorije ne postoji, koristi se drugi pristup za nalaženje dobrih testova: metoda *omjera vjerodostojnosti*. Testovi dobiveni metodom omjera vjerodostojnosti su, na neki način, poopćenja testova dobivenih Neyman-Pearsonovim pristupom. Kritično područje testa omjera vjerodostojnosti zadovoljava nejednadžbu:

$$\frac{\max_{\theta \in H_0} L(\theta|\underline{x})}{\max_{\theta} L(\theta|\underline{x})} \leq k$$

za neki k. Maksimum u brojniku se uzima samo po vrijednostima parametra koja zadovoljavaju hipotezu H_0, dok se maksimum u nazivniku uzima u odnosu na sve moguće vrijednosti parametara. Na primjer, primjenom tog pristupa na slučaj uzorkovanja iz normalne populacije $N(\mu, \sigma^2)$ i testiranje hipoteze $H_0 : \mu = \mu_0$, gdje je μ_0 zadani broj, dolazimo do testa opisanoga testnom statistikom

$$T = \frac{\overline{X} - \mu_0}{S}\sqrt{n}$$

za koju vrijedi da, ako je H_0 ispunjeno, ima Studentovu $t(n-1)$-razdiobu. Tu činjenicu pišemo:

$$T \overset{H_0}{\sim} t(n-1).$$

Istom metodom dolazimo do testne statistike za testiranje $H_0 : \sigma^2 = \sigma_0^2$ (σ_0^2 je zadani pozitivni realni broj):

$$\frac{(n-1)s^2}{\sigma_0^2} \overset{H_0}{\sim} \chi^2(n-1).$$

Primjer 9.1 $\underline{X}$ je slučajni uzorak iz $N(\mu, \sigma^2)$-distribuirane populacije i oba parametra su nepoznata. želimo sprovesti *jednostrani* test (μ_0 je zadani broj):

$$H_0 : \mu = \mu_0, \quad H_1 : \mu < \mu_0$$

uz razinu značajnosti od 5%. Testna statistika je

$$T = \frac{\overline{X} - \mu_0}{S}\sqrt{n} \overset{H_0}{\sim} t(n-1),$$

a kritično područje je $\langle -\infty, -t_{0.05}(n-1)]$. Dakle, ako je vrijednost $t = T(\underline{x})$ testne statistike T na opaženom uzorku $\underline{x}$ takva da je $t \leq -t_{0.05}(n-1)$, tada odbacujemo H_0 uz značajnost od 5% (tj. uz rizik od 5% da smo pogriješili)[1]. Da smo imali suprotnu jednostranu alternativu $H_1' : \mu > \mu_0$, tada bi kritično područje bio interval $[t_{0.05}(n-1), +\infty\rangle$.

[1]Korektna interpretacija ovog jednostranog testa (i gotovo svakog drugog jednostranog testa) je da, u stvari, testiramo $H_0' : \mu \geq \mu_0$ u odnosu na $H_1 : \mu < \mu_0$. U tom slučaju, razina značajnosti se interpretira kao *najveća* vjerojatnost pogreške prve vrste. Budući da nije teško pokazati da se ta najveća vjerojatnost postiže upravo za graničnu vrijednost $\mu = \mu_0$, jednostrani test iz primjera 9.1 je ekvivalentan testu sa korektnim zapisom hipoteza.

Ako bi htjeli sprovesti *dvostrani* test iste osnovne hipoteze (ali drugačije, tzv. *dvostrane* alternative):

$$H_0 : \mu = \mu_0, \quad H_1 : \mu \neq \mu_0,$$

tada bi koristili istu testnu statistiku, ali bi kritično područje bila unija dva simetrična intervala:

$$\langle -\infty, -t_{0.025}(n-1)] \cup [t_{0.025}(n-1), +\infty\rangle.$$

Dakle, H_0 bi odbacili u korist dvostrane alternative (uz rizik od 5%) ako bi za opaženu vrijednost t testne statistike T vrijedilo da je $|t| \geq t_{0.025}(n-1)$. □

9.2.2 p-vrijednosti

Klasičan pristup u kojemu se za zadanu razinu značajnosti donosi odluka odbaciti ili ne odbaciti H_0 u odnosu na zadanu alternativu, ne daje informacije o tome koliko su jaki argumenti za odbacivanje ili ne odbacivanje H_0. Puno je informativniji pristup u kojemu se uz vrijednost testne statistike računa i izražava njena p-vrijednost. *p-vrijednost* je vjerojatnost pogreške prve vrste u odnosu na kritično područje kojemu je opažena vrijednost testne statistike granična vrijednost. Drugim riječima, p-vrijednost je najmanja razina značajnosti uz koju bi H_0 bila odbačena u korist zadane alternative H_1, uz vrijednost testne statistike koja je opažena. Dakle, što je p-vrijednost manja to su dokazi protiv H_0 jači.

Primjer 9.2 Za jednostrani test iz primjera 9.1 i opaženu vrijednost t testne statistike T, p-vrijednost je

$$\mathbb{P}(T \leq t|H_0) = \int_{-\infty}^{t} f_{t(n-1)}(u)\, du,$$

gdje je $f_{t(n-1)}$ gustoća Studentove $t(n-1)$-razdiobe. p-vrijednost za dvostrani test s istom nulhipotezom je

$$\mathbb{P}(|T| \geq |t|\,|H_0) = \int_{-\infty}^{-|t|} f_{t(n-1)}(u)\, du + \int_{|t|}^{+\infty} f_{t(n-1)}(u)\, du = 2\cdot\mathbb{P}(T \leq -|t|\,|H_0) = 2\cdot\mathbb{P}(T \geq |t|\,|H_0),$$

jer je t-razdioba simetrična oko nule. □

Primjer 9.3 Opažena vrijednost binomne varijable X s parametrima $(n = 200, \theta)$ je $x = 82$. Želimo testirati

$$H_0 : \theta = 0.5, \quad H_1 : \theta = 0.4.$$

Za testnu statistiku uzet ćemo upravo X. Budući da bi kritično područje trebali tražiti među manjim vrijednostima od ImX (dakle, u smjeru lijevog repa razdiobe od X pod H_0), p-vrijednost je

$$\mathbb{P}(X \leq 82|H_0) = \mathbb{P}(X < 82.5|H_0) = \mathbb{P}(Z < \frac{82.5 - 100}{\sqrt{50}}) \approx \Phi(-2.475) = 0.0067.$$

Dakle, H_0 je vrlo nevjerodostojna pretpostavka. Drugim riječima, imamo jak dokaz protiv H_0, a u korist H_1. □

Statističkim testom ne dokazujemo istinitost hipoteza. Neodbacivanje H_0 ne znači da je ta hipoteza stvarno istinita već samo da nemamo jake dokaze protiv nje. Jedino se u tom smislu može reći da prihvaćamo H_0. Naime, takav stav prema "prihvaćanju" H_0 proizlazi iz činjenice da je H_0 vrlo precizna tvrdnja (uglavnom jednostavna hipoteza), pa kao takva gotovo sigurno nije istinita.

9.3 Osnovni testovi bazirani na jednom uzorku

9.3.1 Testovi o parametru očekivanja

Zadan je slučajni uzorak duljine n iz $N(\mu, \sigma^2)$-distribuirane populacije. Testiramo

$$H_0 : \mu = \mu_0.$$

Imamo dvije situacije:

1. σ je poznata. Tada je testna statistika

$$\frac{\overline{X} - \mu_0}{\sigma}\sqrt{n} \overset{H_0}{\sim} N(0,1).$$

2. σ je nepoznata. U tom slučaju je testna statistika

$$\frac{\overline{X} - \mu_0}{S}\sqrt{n} \overset{H_0}{\sim} t(n-1).$$

   Za velike uzorke je

$$\frac{\overline{X} - \mu_0}{S}\sqrt{n} \overset{H_0}{\sim} N(0,1).$$

Budući da po CGT-u, $\overline{X}$ ima asimptotsku normalu razdiobu, za velike uzorke nije bitna normalnost populacije.

9.3.2 Testovi o populacijskoj varijanci

Zadan je slučajni uzorak duljine n iz $N(\mu, \sigma^2)$-distribuirane populacije. Testiramo

$$H_0 : \sigma^2 = \sigma_0^2.$$

Testna statistika je

$$\frac{(n-1)S^2}{\sigma_0^2} \overset{H_0}{\sim} \chi^2(n-1).$$

9.3.3 Testovi o populacijskoj proporciji

Zadan je slučajni uzorak duljine n iz Bernoullijeve populacije s parametrom θ. Testiramo

$$H_0 : \theta = \theta_0.$$

Neka je X frekvencija uspjeha u tom uzorku. Tada je X testna statistika i

$$X \overset{H_0}{\sim} \text{ binomna } (n, \theta_0).$$

Za veliko n koristi se normalna aproksimacija razdiobe od X.

9.3.4 Testovi o parametru Poissonove populacije

Zadan je slučajni uzorak $\underline{X} = (X_1, X_2, \ldots, X_n)$ duljine n iz populacije s Poissonovom $P(\lambda)$-razdiobom. Testiramo

$$H_0 : \lambda = \lambda_0.$$

Testna statistika je $Y := X_1 + X_2 + \cdots + X_n$ i za nju vrijedi

$$Y \overset{H_0}{\sim} P(n\lambda_0).$$

Za veliko n koristi se normalna aproksimacija razdiobe od Y ili od $\overline{X} = Y/n$:

$$\frac{Y - n\lambda_0}{\sqrt{n\lambda_0}} \overset{H_0}{\sim} N(0,1) \quad \text{ili} \quad \frac{\overline{X} - \lambda_0}{\sqrt{\lambda_0}}\sqrt{n} \overset{H_0}{\sim} N(0,1).$$

Prva statistika je primjerenija za primjenu korekcije zbog neprekidnosti.

9.4 Osnovni testovi bazirani na dva uzorka

9.4.1 Test o razlici populacijskih očekivanja

Zadana su dva nezavisna uzorka duljina n_1 i n_2 iz normalno distribuiranih populacija: $N(\mu_1, \sigma_1^2)$ i $N(\mu_2, \sigma_2^2)$. Testiramo (δ_0 je zadani broj)

$$H_0 : \mu_1 - \mu_2 = \delta_0.$$

Imamo sljedeće situacije:

1. σ_1^2 i σ_2^2 su poznati. Tada je testna statistika

$$Z = \frac{\overline{X}_1 - \overline{X}_2 - \delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \overset{H_0}{\sim} N(0,1).$$

2. σ_1^2 i σ_2^2 su nepoznati. Ako imamo velike uzorke, σ_1^2 i σ_2^2 procijenimo iz uzoraka pomoću S_1^2 i S_2^2. U tom slučaju je $Z \overset{H_0}{\sim} N(0,1)$. Ako imamo male uzorke, tada moramo još pretpostaviti da su $\sigma_1^2 = \sigma_2^2 = \sigma^2$, te zajedničku varijancu procijeniti pomoću zajedničke uzoračke varijance S_p^2 (vidjeti 8.4.1). U tom slučaju, testna statistika je

$$T = \frac{\overline{X}_1 - \overline{X}_2 - \delta_0}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \overset{H_0}{\sim} t(n_1 + n_2 - 2).$$

9.4.2 Test o kvocijentu populacijskih varijanci

Kao i u prethodnom potpoglavlju, zadana su dva nezavisna uzorka duljina n_1 i n_2 iz normalno distribuiranih populacija: $N(\mu_1, \sigma_1^2)$ i $N(\mu_2, \sigma_2^2)$. Testiramo

$$H_0 : \sigma_1^2 = \sigma_2^2.$$

Testna statistika je

$$\frac{S_1^2}{S_2^2} \overset{H_0}{\sim} F(n_1 - 1, n_2 - 1).$$

Dvostrani se test hipoteze H_0 (dakle, uz alternativu $H_1 : \sigma_1^2 \neq \sigma_2^2$) koristi da bi se ispitala pretpostavka o jednakosti populacijskih varijanci za primjenu t-testa usporedbe populacijskih očekivanja (mali uzorak, potpoglavlje 9.4.1, situacija 2.) Kojiput je dovoljno sprovest grafički test za provjeru te pretpostavke. Postupak za provedbu tog testa je sljedeći. Izračunajte uzoračke varijance i po potrebi renumerirajte uzorke tako da prvi uzorak ima veću opaženu vrijednost uzoračke varijance. Izračunajte vrijednost testne statistike i uvidom u tablice izračunajte vjerojatnost da testna statistika, uz H_0, poprimi vrijednosti veće ili jednake opaženoj. Tada je p-vrijednost jednaka dvostrukom iznosu te vjerojatnosti.

9.4.3 Test razlike između populacijskih proporcija

Imamo velike i nezavisne uzorke duljina n_1 i n_2 iz Bernoullijevih populacija. Testiramo

$$H_0 : \theta_1 = \theta_2.$$

Testna statistika je

$$\frac{\hat{\theta}_1 - \hat{\theta}_2}{\sqrt{\hat{\theta}(1-\hat{\theta})(\frac{1}{n_1} + \frac{1}{n_2})}} \overset{H_0}{\sim} N(0,1),$$

pri čemu su $\hat{\theta}_1$ i $\hat{\theta}_2$ MLE za parametre θ_1 i θ_2 na bazi svakog uzorka posebno, te $\hat{\theta} = \frac{n_1\hat{\theta}_1 + n_2\hat{\theta}_2}{n_1+n_2}$ je procjena zajedničke proporcije (uz H_0) na bazi združenih uzoraka.

9.4.4 Test razlike između parametara Poissonovih razdioba

Zadana su dva velika i nezavisna uzorka duljina n_1 i n_2 iz populacija s Poissonovim razdiobama $P(\lambda_1)$ i $P(\lambda_2)$. Testiramo

$$H_0 : \lambda_1 = \lambda_2.$$

Testna statistika je

$$\frac{\hat{\lambda}_1 - \hat{\lambda}_2}{\sqrt{\hat{\lambda}(\frac{1}{n_1} + \frac{1}{n_2})}} \overset{H_0}{\sim} N(0,1),$$

pri čemu su $\hat{\lambda}_1$ i $\hat{\lambda}_2$ MLE za parametre λ_1 i λ_2 na bazi svakog uzorka posebno, te $\hat{\lambda} = \frac{n_1\hat{\lambda}_1 + n_2\hat{\lambda}_2}{n_1+n_2}$ je procjena zajedničkog parametra $\lambda = \lambda_1 = \lambda_2$ (uz H_0) na bazi združenih uzoraka.

9.5 Osnovni test za sparene podatke

Pretpostavimo da slučajni uzorak $\underline{D} = (D_1, D_2, \ldots, D_n)$ razlika komponenti parova sparenih podataka $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ čine uzorak iz normalne populacije s nepoznatim parametrima očekivanja $\mu_D = \mu_1 - \mu_2$ i varijance. Za zadani broj δ_0 testiramo

$$H_0 : \mu_D = \delta_0.$$

Testna statistika je

$$T_D = \frac{\overline{D} - \delta_0}{S_D}\sqrt{n} \overset{H_0}{\sim} t(n-1).$$

U slučaju velikog uzorka

$$T_D \overset{H_0}{\sim} N(0,1).$$

Ista asimptotska razdioba vrijedi za populacije koje nisu normalno distribuirane.

9.6 Testovi i pouzdani intervali

Obzirom na način kako su konstruirani, postoji direktna veza između pouzdanih intervala i testova. Ako je testna statistika (uz H_0) ista kao pivotna veličina i ako razina značajnosti testa α odgovara pouzdanosti intervala od $1-\alpha$, onda je veza sljedeća. Za sve one opažene vrijednosti testne statistike koje se nalaze unutar granica pouzdanog intervala, H_0 se ne odbacuje. U suprotnome se H_0 odbacuje. Pri tome dvostranim testovima odgovaraju dvostrani pouzdani intervali, a jednostranim testovima odgovarajući jednostrani intervali.

U nekim slučajevima testna statistika i pivotna veličina nisu sasvim istoga oblika. Na primjer, pri konstrukciju pouzdanog intervala za razliku proporcija, u nazivniku pivotne veličine nalaze se procjene proporcija svake od populacija posebno, dok se u testnoj statistici kojom se testira jednakost proporcija, u nazivniku nalazi procjena zajedničke proporcije. Ipak, aproksimativno su ti nazivnici bliski pa i u tom slučaju možemo donositi zaključke o nulhipotezama na bazi pouzdanih intervala kao i u ostalim slučajevima.

9.7 χ^2-testovi

χ^2-testovi se odnose na populacijske razdiobe kategorijalnih ili diskretnih numeričkih varijabli. Testiranje se zasniva na usporedbi opaženih frekvencija f_i i očekivanih (u skladu s nulhipotezom) frekvencija e_i i-tog razreda. Testna statistika je, u stvari, težinska mjera udaljenosti tih frekvencija:

$$H = \sum_i \frac{(f_i - e_i)^2}{e_i}.$$

Ako vrijedi nulhipoteza, tada H ima aproksimativno χ^2-razdiobu kada je uzorak velik. Nulhipotezu odbacujemo ako je opažena vrijednost od H prevelika.

9.7.1 Test prilagodbe modela podacima

χ^2-testom testiramo da li predloženi diskretni model za populacijsku razdiobu dobro objašnjava opažene podatke. Drugim riječima, testiramo je li model dobro prilagođen podacima. U tom slučaju, nulhipoteza H_0 je da se populacijska razdioba opažane varijable ravna po zakonu razdiobe pretpostavljenog modela. Još kažemo da se podaci ravnaju po zakonu razdiobe pretpostavljenom nulhipotezom H_0.

Odredimo broj stupnjeva slobode razdiobe testne statistike. Pretpostavimo da opažana varijabla ima k razreda. Budući da su frekvencije zavisne u smislu da je njihov zbroj fiksan i jednak n (duljini uzorka), imamo jedan stupanj slobode manje. Ako pri tome još i po H_0 pretpostavljena populacijska razdioba ima r nepoznatih parametara koje treba procijeniti iz uzorka (da bi se mogle izračunati očekivane frekvencije), tada gubimo još r stupnjeva slobode. Dakle testna statistika H uz H_0 ima χ^2-razdiobu s $k-r-1$ stupnjem slobode. Nepoznati parametri procjenjuju se metodom maksimalne vjerodostojnosti.

Budući da je razdioba testne statistike aproksimativno jednaka χ^2-razdiobi, treba paziti da ta aproksimacija bude zadovoljavajuće točna. To će biti ispunjeno ako nazivnici u izrazu za H, dakle očekivane frekvencije, ne budu premale. Obično se zahtjeva da budu barem 5. U nekim situacijama se prihvaća da je dovoljno da su veće od 1, ali uz uvjet da je barem 80% razreda frekvencije barem 5.

Primjer 9.4 Želim li testirati je li neka igraća kocka fer, prikladni model je uniformna razdioba na skupu prvih šest prirodnih brojeva. Dakle, ako je X broj koji se okrene na

kocki nakon bacanja, za nulhipotezu uzimamo

$$H_0: \ \mathbb{P}(X=i)=\frac{1}{6} \ \text{ za } \ i=1,2,3,4,5,6.$$

Rezltati $n=300$ bacanja te kocke dani su u frekvencijskoj tablici:

i	1	2	3	4	5	6
f_i	43	56	54	47	41	59

Uz H_0, očekivane frekvencije svih šest razreda su jednake $n\cdot\frac{1}{6}=300/6=50$. Izračun vrijednosti h testne statistike H prikazan je u tablici:

i	f_i	e_i	$\frac{(f_i-e_i)^2}{e_i}$
1	43	50	49/50
2	56	50	36/50
3	54	50	16/50
4	47	50	9/50
5	41	50	81/50
6	59	50	81/50
Σ	300	300	272/50

Dakle, $h=272/50=5.44$. Budući da je $H \overset{H_0}{\sim} \chi^2(6-0-1)=\chi^2(5)$, p-vrijednost je

$$\mathbb{P}(H\geq 5.44|H_0)=0.365.$$

Dakle, nemamo jakih dokaza da kocka nije fer. □

Primjer 9.5 U primjeru 7.8 podacima o brojevima šteta po 100000 polica od autoodgovornosti prilagođena je Poissonova razdioba. Sprovedimo test je li prilagodba Poissonovog modela tim podacima dobra. Dakle, nulhipoteza je da se brojevi šteta X po polici autoodgovornosti u protekloj godini ravnaju po Poissonovom zakonu razdiobe. Metodom maksimalne vjerodostojnosti procijenjen je nepoznati parametar razdiobe i ta procjena je $\hat{\lambda}=0.22078$. Izračunajmo očekivane frekvencije i vrijednost h testne statistike H. Očekivane frekvencije za prvih pet razreda

$$e_i=n\cdot\mathbb{P}(X=i|H_0)=n\cdot\frac{\hat{\lambda}^i}{i!}e^{-\hat{\lambda}}, \ \ i=0,1,2,3,4$$

zadovoljavaju rekurzivnu relaciju

$$e_i=\frac{\hat{\lambda}}{i}\cdot e_{i-1} \ \text{ za } \ i=1,2,3,4 \ \text{ i } \ e_0=ne^{-\hat{\lambda}},$$

gdje je $n=100000$ i $\hat{\lambda}=0.22078$. Očekivanu frekvenciju zadnjeg razreda računamo po formuli $e_{\geq 5}=n-\sum_{i=0}^{4}e_i$. Rezultat:

i	f_i		e_i		$\frac{(f_i-e_i)^2}{e_i}$
0	81056		80177.3		9.6
1	16174		17713.6		133.8
2	2435		1956.7		116.9
3	295		144.0		158.3
4	36	40	8.0	8.4	118.9
≥ 5	4		0.4		
Σ	100000		100000.0		537.5

Zadnja smo dva razreda u tablici morali združiti u jedan budući da očekivana frekvencija u zadnjem razredu (0.4) nije bila veća od 5. Dakle, imamo ukupno $k = 5$ razreda. Budući da smo jedan parametar morali procijeniti iz podataka, ukupan broj stupnjeva slobode je $5 - 1 - 1 = 3$, pa je

$$H \overset{H_0}{\approx} \chi^2(3).$$

Opažena vrijednost te statistike je $h = 537.5$, pa je p-vrijednost $\mathbb{P}(H \geq 537.5|H_0)$ vrlo mala, gotovo jednaka nuli. Dakle, Poissonova razdioba ne opisuje zakon razdiobe broja šteta prezentiranih danim uzorkom. □

9.7.2 Kontingencijske tablice

Kontingencijskim tablicama prikazujemo frekvencije uzorka dobivenog mjerenjem dvodimenzionalnog kategorijalnog ili diskretnog numeričkog obilježja (X, Y). Cilj nam je testirati nulhipotezu da su X i Y nezavisne varijable ili nulhipotezu da su populacijske razdiobe jedne varijable, recimo X, homogene obzirom na populacije klasificirane po drugoj komponenti (Y). Uz te pretpostavke, očekivane frekvencije svakog polja u tablici računaju se po formuli:

$$\frac{\text{ukupan zbroj toga retka} \times \text{ukupan zbroj toga stupca}}{\text{veličina uzorka}}.$$

Ako u tablici imamo r redaka i c stupaca, tada je broj stupnjeva slobode testne statistike

$$rc - (r - 1 + c - 1) - 1 = (r - 1)(c - 1).$$

Primjer 9.6 Za svako od osiguravajućih društava A, B i C uzet je po slučajni uzorak polica neživotnih osiguranja određenog tipa. Rezultat dobiven opažanjima tih uzoraka je da je šteta u prošloj godini bilo po 23% polica od A, po 28% polica od B i po 20% polica od C. Testirajte ima li značajnih razlika između tih proporcija ako su veličine uzoraka

(a) 100, 100, 200

(b) 300, 300, 600

redom za A, B, C.

Rješenje. Nulhipoteza je da je obilježje "ima ili nema štete po polici" homogeno po distribuciji obzirom na pripadnost police osiguravajućem društvu A, B ili C.

Za slučaj (a) kontingencijske tablice opaženih i očekivanih frekvencija su:

f_{ij}	A	B	C	Σ
ima štete	23	28	40	91
nema štete	77	72	160	309
Σ	100	100	200	400

e_{ij}	A	B	C	Σ
ima štete	22.75	22.75	45.50	91
nema štete	77.25	77.25	154.50	309
Σ	100	100	200	400

Vrijednost testne statistike

$$H = \sum_{i,j} \frac{(f_{ij} - e_{ij})^2}{e_{ij}} \overset{H_0}{\approx} \chi^2((2-1)\cdot(3-1)) = \chi^2(2)$$

je $h = 2.43$, pa je p-vrijednost $\mathbb{P}(H \geq 2.43|H_0) = 0.30$. Dakle, nemamo jakih dokaza za odbacivanje hipoteze o homogenosti proporcija polica sa štetama u tri navedena osiguravajuća društva.

Budući da su opažene proporcije iste, u slučaju (b) su sve frekvencije uvećane tri puta u odnosu na slučaj (a). To znači da je i vrijednost testne statistike uvećana tri puta i iznosi $h = 3 \cdot 2.43 = 7.29$. Primijetite da je aproksimativna razdioba testne statistika ista. p-vrijednost je $\mathbb{P}(H \geq 7.29|H_0) = 0.026$, pa ovoga puta imamo jake dokaze za odbacivanje nulhipoteze o homogenosti proporcija broja šteta. □

Poglavlje 10

Korelacija i regresija

U ovom poglavlju bavimo se statističkom analizom *povezanosti* među varijablama. Iako se može izučavati povezanost više varijabli, mi ćemo se ograničiti na bivarijatni slučaj (X, Y). Nadalje, ograničit ćemo se samo na *linearnu* povezanost, dakle, na modele koji pretpostavljaju da je uvjetno očekivanje od Y za svaku danu vrijednost x od X, linearna funkcija od x (tj. regresijska funkcija od Y na $X = x$ je linearna funkcija):

$$\mathbb{E}[Y|X = x] = \alpha + \beta x.$$

U *korelacijskoj analizi* dviju varijabli naglasak je na problemu određivanja mjere jakosti linearne povezanosti među njima.

U *regresijskoj analizi* varijabli X i Y naglasak je na prirodi veze između varijable Y kao zavisne varijable (odziv) i X kao nezavisne varijable (poticaj). Analiza se sastoji u odabiru i prilagodbi primjerenog modela u svrhu predviđanja odziva (individualne vrijednosti od Y ili očekivane vrijednosti od Y) za zadani poticaj (za zadanu vrijednost x od X). Kao što smo već spomenuli, ograničit ćemo se samo na linearnu vezu između X i Y.

Pretpostavimo da je mjerenje bivarijatnog vektora (X, Y) dalo sljedeće podatke:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) \tag{10.1}$$

Prije nego odaberemo i primijenimo metode inferencijalne statistike, analizirajmo podatke grafički, pomoću dijagrama raspršenja, da bi vidjeli postoji li uopće ikakva veza između varijabli X i Y, te, ako postoji, koja je priroda povezanosti.

Ako je linearna povezanost od X i Y plauzibilna, analizira se varijabilnost od Y za fiksnu vrijednost x od X da bi se ocijenila jakost linearne veze između X i Y.

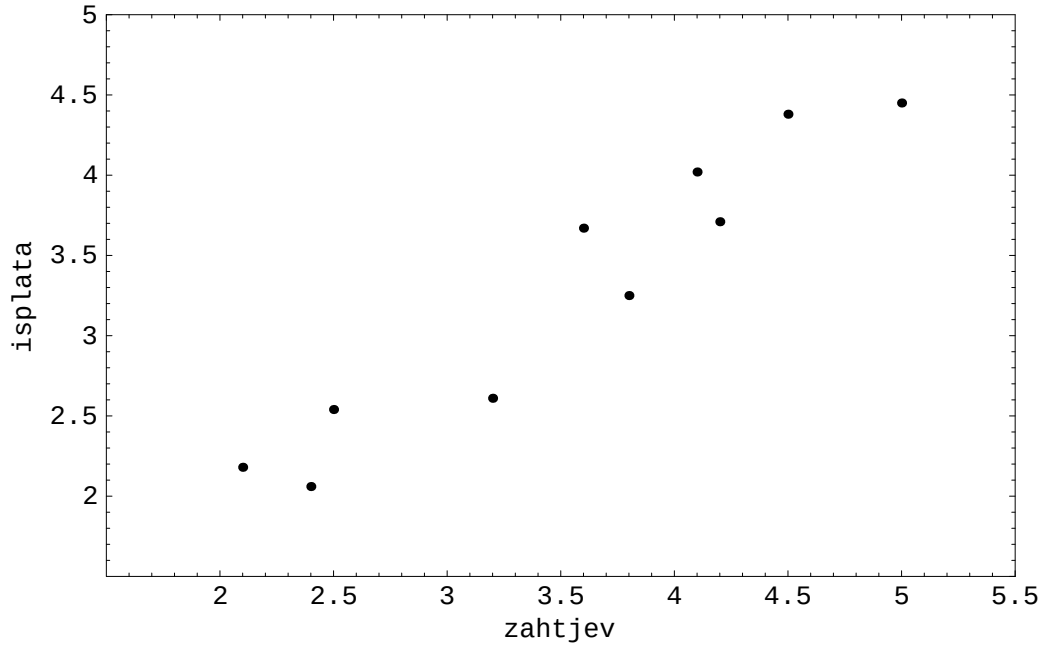
Ako se pokazalo da X i Y nisu povezane ili da veza nije linearna, tada se metode koje ćemo diskutirati u ovom poglavlju ne mogu primijeniti. S druge strane, može se dogoditi da dobro odabrana transformacija originalnih podataka pokaže linearnu povezanost između transformiranih podataka. U tom slučaju se opisane metode mogu primijeniti na transformirane varijable.

Dijagram raspršenja je prikaz podataka (10.1) kao točaka u Kartezijevom koordinatnom sustavu.

Primjer 10.1 Uzorak se sastoji od 10 podataka o iznosima zahtjeva za naknadu šteta i korespodentnih iznosa koje je osiguravajuće društvo stvarno platilo (u jedinicama od po 100 kn):

zahtjev	(x)	2.10	2.40	2.50	3.20	3.60	3.80	4.10	4.20	4.50	5.00
isplata	(y)	2.18	2.06	2.54	2.61	3.67	3.25	4.02	3.71	4.38	4.45

Dijagram raspršenja:



Očito se radi o linearnoj povezanosti između opaženih vrijednosti varijabli $X =$ "zahtjev" (za isplatom šteta) i $Y =$ "isplata" (od strane društva). □

U analizi linearne zavisnosti dviju varijabli, sljedeće se statistike koriste:

$$\begin{aligned} S_{XX} &:= \sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n} X_i^2 - n \cdot \overline{X}^2 \\ S_{XY} &:= \sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = \sum_{i=1}^{n} X_i Y_i - n\overline{XY} \\ S_{YY} &:= \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n} Y_i^2 - n \cdot \overline{Y}^2. \end{aligned}$$

Opažene vrijednosti tih statistika označavat ćemo S_{xx}, S_{xy}, S_{yy}.

Primjer 10.1 (*nastavak*) Izračunajmo vrijednosti S_{xx}, S_{xy}, S_{yy} navedenih statistika.

i	x_i	y_i	x_i^2	$x_i y_i$	y_i^2
1	2.10	2.18	4.41	4.578	4.7524
2	2.40	2.06	5.76	4.944	4.2436
3	2.50	2.54	6.25	6.350	6.4516
4	3.20	2.61	10.24	8.352	6.8121
5	3.60	3.67	12.96	13.212	13.4689
6	3.80	3.25	14.44	12.350	10.5625
7	4.10	4.02	16.81	16.482	16.1604
8	4.20	3.71	17.64	15.582	13.7641
9	4.50	4.38	20.25	19.710	19.1844
10	5.00	4.45	25.00	22.250	19.8025
Σ	35.40	32.87	133.76	123.810	115.2025

Budući da je $n = 10$, iz tablice čitamo

$$\bar{x} = \frac{35.40}{10} = 3.540, \quad \bar{y} = \frac{32.87}{10} = 3.287,$$

$$\sum_{i=1}^{10} x_i^2 = 133.76, \quad \sum_{i=1}^{10} x_i y_i = 123.810, \quad \sum_{i=1}^{10} y_i^2 = 115.2025,$$

odakle slijedi

$$\begin{aligned} S_{xx} &= \sum_{i=1}^{n} x_i^2 - n \cdot \bar{x}^2 = 133.76 - 10 \cdot 3.540^2 = 8.4440 \\ S_{xy} &= \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} = 123.810 - 10 \cdot 3.540 \cdot 3.287 = 7.4502 \\ S_{yy} &= \sum_{i=1}^{n} y_i^2 - n \cdot \bar{y}^2 = 115.2025 - 10 \cdot 3.287^2 = 7.1588. \end{aligned}$$

□

10.1 Korelacijska analiza

10.1.1 Uzorački koeficijent korelacije

Povezanost među komponentama podataka (10.1) u uzorku za (X, Y) mjeri se *uzoračkim* ili *Pearsonovim koeficijentom korelacije*:

$$r := \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}. \tag{10.2}$$

Primijetite da se r može pomoću aritmetičkih sredina $\bar{x}$, $\bar{y}$ i standardnih devijacija s_x, s_y podataka za X, odnosno Y, zapisati na sljedeći način:

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \frac{x_i - \bar{x}}{s_x} \cdot \frac{y_i - \bar{y}}{s_y},$$

dakle, kao srednja vrijednost produkata standardiziranih verzija podataka za X, odnosno Y iz (10.1). Nadalje, uvijek vrijedi da je

$$-1 \leq r \leq 1.$$

Koeficijent korelacije r je mjera stupnja linearne povezanosti i nije indikator uzročnosti u smislu da mjeri koliko X uzrokuje Y. Naime, varijable X i Y mogu biti jako povezane, a da jedna ne uzrokuje drugu, već su, možda, i jedna i druga uzrokovane djelovanjem nekog trećeg faktora.

Primjer 10.2 Za podatke iz primjera 10.1 Pearsonov koeficijent korelacije iznosi

$$r = \frac{7.4502}{\sqrt{8.444 \cdot 7.1588}} = 0.958$$

što indicira jaku pozitivnu linearnu povezanost među komponentama u uzorku. □

10.1.2 Normalni model i inferencija

Odgovarajući model za populacijsku razdiobu od (X, Y) je *bivarijatna normalna razdioba* s parametrima μ_X, μ_Y (populacijske sredine komponenata), σ_X^2, σ_Y^2 (populacijske varijance komponenata) i ρ (koeficijent korelacije komponenata). Podatke (10.1) možemo interpretirati kao realizaciju slučajnog uzorka

$$\underline{(X,Y)} = ((X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n))$$

za vektor (X, Y), a Pearsonov koeficijent korelacije (10.2) kao opaženu vrijednost statistike

$$R = \frac{S_{XY}}{\sqrt{S_{XX} \cdot S_{YY}}} \quad \text{(uzorački koeficijent korelacije).}$$

R je i MLE za parametar ρ, populacijski koeficijent korelacije.

Za konstrukciju pouzdanih intervala za ρ, odnosno za testiranje hipoteza o ρ, treba nam uzoračka razdioba od R. Pokazuje se da je ta razdioba asimetrična s velikom varijancom.

Želimo li testirati jesu li X i Y korelirane varijable, dakle, nulhipotezu $H_0 : \rho = 0$, tada je tesna statistika

$$\frac{R}{\sqrt{1-R^2}}\sqrt{n-2} \overset{H_0}{\sim} t(n-2).$$

Postoji općenitiji, ali asimptotski rezultat koji nam omogućava testiranje nulhipoteza oblika $H_0 : \rho = \rho_0$, gdje je ρ_0 zadani broj takav da je $|\rho_0| < 1$. Naime, vrijedi

$$W := \frac{1}{2}\log\frac{1-R}{1+R} \overset{\cdot}{\sim} N\left(\frac{1}{2}\log\frac{1-\rho}{1+\rho}, \frac{1}{n-3}\right) \text{ za veliko } n.$$

Statistika W je *Fisherova transformacija od* R. Iz nje izvodimo testnu statistiku

$$Z = \frac{\sqrt{n-3}}{2}\left(\log\frac{1-R}{1+R} - \log\frac{1-\rho_0}{1+\rho_0}\right) \overset{H_0}{\overset{\cdot}{\sim}} N(0,1) \text{ za veliko } n.$$

Standardizirana verzija od W se koristi kao pivotna veličina u konstrukciji aproksimativnih pouzdanih intervala za ρ na osnovi velikih uzoraka.

Primjer 10.3 Na osnovi podataka iz primjera 10.1, sprovedimo jednostrani test

$$H_0 : \rho = 0.9, \quad H_1 : \rho > 0.9.$$

Budući da je $r = 0.958$, $n = 10$, opažena wrijednost od W je $w = 1.921$. Dakle, opažena vrijednost testne statistike Z je $z = (1.921 - 1.472)/0.378 = 1.19$, pa je p-vrijednost $\mathbb{P}(Z \geq 1.19|H_0) \approx 0.12$. Prema tome, dokazi za odbacivanje H_0 su nedostatni, odnosno procijenjena vrijednost za ρ nije značajno veća od 0.9. □

Napomene:

1. Postojanje "outliera" indicira da je adekvatnost pretpostavke o normalnoj distribuiranosti bivarijatne populacije upitna.

2. Kako samo jedna opažena vrijednost može imati značajan utjecaj na procjene populacijskih sredina i varijance, isto tako može imati i znatan utjecaj na procjenu koeficijenta korelacije.

10.2 Regresijska analiza. Jednostavni linearni regresijski model.

10.2.1 Uvod

U ovom je potpoglavlju Y slučajna varijabla koja ovisi o vrijednostima x nezavisne varijable X. S druge strane, pretpostavljamo da imamo kontrolu nad vrijednostima x nezavisne varijable, odnosno, da ih mi na neki način zadajemo. Dakle, ovdje X nije slučajna varijabla. Regresijska analiza se sastoji od odabira i prilagodbe odgovarajućeg modela (u našem slučaju, linearnog) opaženim podacima, u svrhu predviđanja individualnih ili očekivanih vrijednosti od Y za zadanu vrijednost x od X. Prije postavljanja potpunog modela, podatke treba ispitati deskriptivnim metodama (npr. grafički) da se vidi koje su sve pretpostavke na model razumne.

Podatke (10.1) interpretiramo kao realizaciju slučajnog uzorka

$$(x_1, Y_1), (x_2, Y_2), \ldots, (x_n, Y_n)$$

pri čemu pretpostavljamo da vrijedi

$$Y_i = \alpha + \beta x_i + \varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{10.3}$$

gdje su α, β parametri modela (α je *slobodni član*, a β je *koeficijent smjera* pravca), a ε_i ($i = 1, 2, \ldots, n$) su slučajne varijable koje zovemo *slučajnim pogreškama* ili *šumovima*. Model (10.3) za vezu između komponenti od (X, Y) zovemo *jednostavni linearni regresijski model*. Dodatne su pretpostavke na jednostavni linearni regresijski model, preciznije, na slučajne pogreške, da su:

$(A1)$ *centrirane*: $\mathbb{E}[\varepsilon_i] = 0$ za sve i;

$(A2)$ *jednake varijance*: $\mathbb{V}\mathrm{ar}[\varepsilon_i] = \sigma^2$ za sve i;

$(A3)$ *nekorelirane*: $\mathrm{cov}[\varepsilon_i, \varepsilon_j] = 0$ za sve $i \neq j$.

Uvjete $(A1 - 3)$ zovemo *Gauss-Markovljevim uvjetima*.

10.2.2 Prilagodba modela

Prilagodba linearnog regresijskog modela za koji vrijede Gauss-Markovljevi uvjeti sastoji se od:

(a) procjene parametara α i β;

(b) procjene zajedničke varijance grešaka σ^2.

Regresijski parametri α i β procjenjuju se *metodom najmanjih kvadrata*. Neka je

$$q(\alpha, \beta) := \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2 \tag{10.4}$$

funkcija kojom se mjeri ukupan zbroj kvadrata odstupanja opaženih vrijednosti y_i od vrijednosti predviđenih pravcem $y = \alpha + \beta x$ u točkama $x = x_i$ za $i = 1, 2, \ldots, n$. Procjene $\hat{\alpha}$ i $\hat{\beta}$ parametara regresijskog pravca su one vrijednosti α i β za koje funkcija $q(\alpha, \beta)$ postiže svoj minimum:

$$q(\hat{\alpha}, \hat{\beta}) = \min_{\alpha, \beta} q(\alpha, \beta).$$

Nije teško pokazati da se te procjene mogu izračunati po formulama

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}}, \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}.$$

Naime, $\hat{\alpha}$ i $\hat{\beta}$ su rješenja sustava od dvije jednadžbe koje se dobiju kada se parcijalne derivacije od $q(\alpha, \beta)$ izjednače s nulom. Dakle, *procjenitelji metodom najmanjih kvadrata* od α i β su statistike

$$\hat{\beta} = \frac{S_{x\underline{Y}}}{S_{xx}}, \quad \hat{\alpha} = \overline{Y} - \hat{\beta}\bar{x}.$$

Za uzoračke razdiobe tih statistika vrijedi:

$$\mathbb{E}[\hat{\beta}] = \beta, \quad \mathbb{V}\mathrm{ar}[\hat{\beta}] = \sigma^2 \cdot \frac{1}{S_{xx}},$$

$$\mathbb{E}[\hat{\alpha}] = \alpha, \quad \mathbb{V}\mathrm{ar}[\hat{\alpha}] = \sigma^2 \cdot (\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}).$$

Označimo sa $\hat{Y}_i := \hat{\alpha} + \hat{\beta}x_i$ procjenitelj za varijablu Y_i, a sa $\hat{y}_i$ opaženu vrijednost tog procjenitelja ($i = 1, 2, \ldots, n$). Tada je nepristrani procjenitelj zajedničke varijance slučajnih grešaka statistika

$$\hat{\sigma}^2 := \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

Primijetite da je opažena vrijednost $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ od $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ upravo jednaka minimumu funkcije (10.4)

10.2.3 Rastav varijance odziva

Ukupna varijabilnost u slučajnom uzorku $\underline{Y}$ varijable odziva Y je

$$\mathrm{SSTOT} := \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = S_{YY}.$$

Dio te varijabilnosti se objašnjava postojanjem linearne ovisnosti varijabli Y_i o vrijednostima x_i od X, a dio postojanjem slučajnih pogrešaka. Udio varijabilnosti zbog linearne ovisnosti Y o X u ukupnoj varijabilnosti SSTOT predstavlja mjeru prilagodbe linearnog regresijskog modela podacima.

Kvadriranjem izraza

$$Y_i - \overline{Y} = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \overline{Y})$$

i zbrajanjem po svim $i = 1, 2, \ldots, n$ dobijamo

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2,$$

budući da je zbroj srednjih članova kvadrata binoma jednak nuli. Član s lijeve strane dobivene jednakosti je SSTOT, ukupna varijabilnost u podacima od Y ili *ukupna suma kvadrata*. Drugi član s desne strane jednakosti predstavlja zbroj kvadrata odstupanja vrijednosti prilagođene regresijske funkcije od uzoračke srednje vrijednosti od Y, dakle, onaj dio ukupne varijabilnosti koji je objašnjen linearnom zavisnošću Y o X. Taj zbroj zovemo *sumom kvadrata zbog regresije* i označavamo sa SSR. Na kraju, prvi član zdesna je zbroj kvadrata procijenjenih pogrešaka ili *reziduala* $\hat{\varepsilon}_i := Y_i - \hat{Y}_i$ ($i = 1, 2, \ldots, n$), koji još zovemo

suma kvadrata pogrešaka ili *reziduala* i označavamo sa SSE. SSE je onaj dio ukupne varijabilnosti koji se objašnjava slučajnim pogreškama.

Dakle,

$$\text{SSTOT} = \text{SSR} + \text{SSE}.$$

Opažene vrijednosti računaju se na sljedeći način:

$$\begin{aligned} \text{SSTOT} &= S_{yy} \\ \text{SSR} &= \sum_{i=1}^{n} \Big((\hat{\alpha} + \hat{\beta} x_i) - (\hat{\alpha} + \hat{\beta}\bar{x}) \Big)^2 = \hat{\beta}^2 S_{xx} = \frac{S_{xy}^2}{S_{xx}} \\ \Rightarrow \quad \text{SSE} &= S_{yy} - \frac{S_{xy}^2}{S_{xx}}. \end{aligned}$$

Nadalje, može se pokazati da vrijedi:

$$\mathbb{E}[\text{SSTOT}] = (n-1)\sigma^2 + \beta^2 S_{xx}, \quad \mathbb{E}[\text{SSR}] = \sigma^2 + \beta^2 S_{xx},$$

odakle slijedi da je

$$\mathbb{E}[\text{SSE}] = (n-2)\sigma^2.$$

Primijetite da zadnja relacija pokazuje da je procjenitelj $\hat{\sigma}^2$ nepristran za σ^2, budući da je očito $\hat{\sigma}^2 = \text{SSE}/(n-2)$.

U slučaju kada su podaci "blizu" pravca (dakle, kada je $|r|$ blisko jedinici što je indikacija jake linearne povezanosti), prilagodba linearnog regresijskog modela je dobra, dakle, procjene $\hat{y}_i$ su bliske opaženim vrijednostima y_i, pa je SSE relativno malo, a SSR relativo veliko. Obratno, kada podaci nisu "bliski" pravcu ($|r|$ je bliže nuli indicirajući slabu linearnu vezu), prilagodba linearnog regresijskog modela nije tako dobra, dakle, procjene $\hat{y}_i$ nisu bliske opaženim vrijednostima y_i, pa je SSE relativno veliko, a SSR relativo malo.

Omjer varijabilnosti objašnjene linearnom vezom i ukupne varijabilnosti:

$$R^2 := \frac{\text{SSR}}{\text{SSTOT}} \cdot 100\% = \frac{S_{xy}^2}{S_{xx} \cdot S_{yy}} \cdot 100\%$$

zove se *koeficijent determinacije*. Primijetite da je (za jednostavni linearni model) koeficijent determinacije, u stvari, kvadrat Pearsonovog koeficijenta korelacije r izražen u formi postotka.

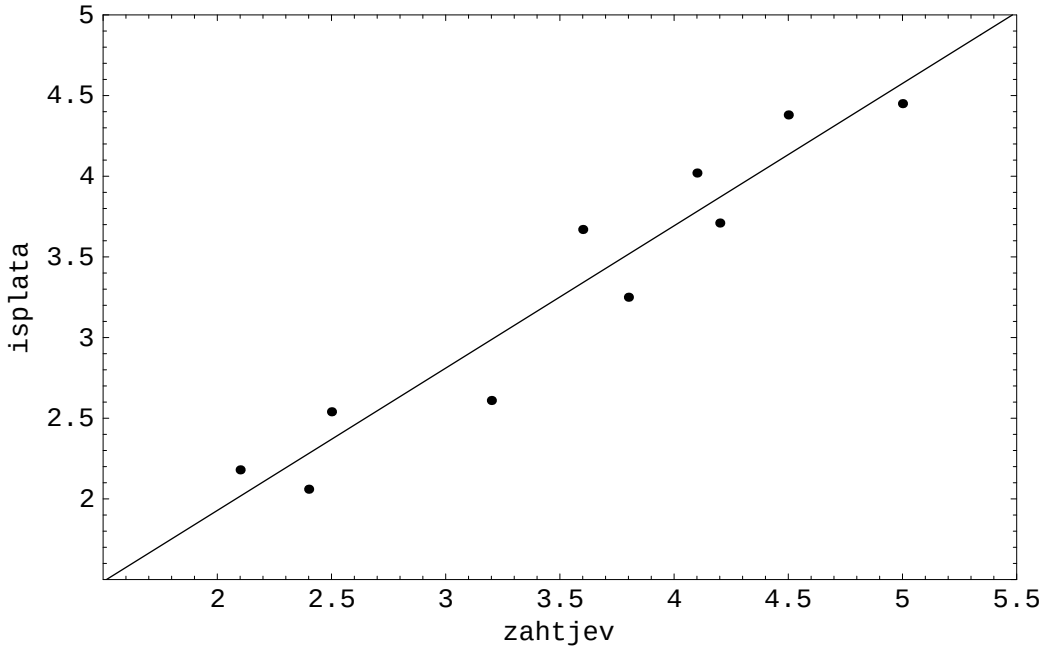
Primjer 10.4 Podacima iz primjera 10.1 prilagodimo jednostavni linearni regresijski model (10.3) pretpostavljajući da su zadovoljeni Gauss-Markovljevi uvjeti. Na osnovi opaženih vrijednosti statistika iz primjera 10.2, procjene koeficijenta smjera β i slobodnog člana α regresijskog pravca su

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{7.4502}{8.4440} = 0.8823, \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 3.287 - 0.8823 \cdot 3.54 = 0.1636.$$

Dakle, procijenjeni pravac je $\hat{y} = 0.1636 + 0.8823x$ i njegov je graf, zajedno s podacima, prikazan na slici (sljedeća stranica). Nadalje,

$$\text{SSTOT} = S_{yy} = 7.1588, \quad \text{SSR} = \frac{S_{xy}^2}{S_{xx}} = \frac{7.4502^2}{8.440} = 6.5734,$$

odakle slijedi da je $\text{SSE} = \text{SSTOT} - \text{SSR} = 0.5854$ i $\hat{\sigma}^2 = \text{SSE}/8 = 0.0732$. Koeficijent determinacije iznosi $R^2 = \text{SSR}/\text{SSTOT} = 91.8\%$ što pokazuje da je prilagodba modela dobra. □



10.2.4 Potpuni normalni model i inferencija

Želimo li predviđati individualni ili srednji odziv na osnovi prilagođenog modela ili konstruirati pouzdane intervale za parametre i sprovoditi testove o njihovim vrijednostima, model treba u potpunosti specificirati. To znači da nam treba pretpostavka o populacijskoj razdiobi slučajnih varijabli Y_i, $i = 1, 2, \ldots, n$, odnosno slučajnih grešaka. Dodatno pretpostavljamo da su slučajne greške

$(A4)$ *nezavisne* i *normalno distribuirane*: $\varepsilon_i \sim N(0, \sigma^2)$ za sve i.

Uz takav potpuni model, slučajne greške $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ su n.j.d. sa $N(0, \sigma^2)$-razdiobom. Slijedi da su varijable $Y_1, Y_2, \ldots, Y_n$ nezavisne i normalno distribuirane, $Y_i \sim N(\alpha + \beta x_i, \sigma^2)$ za $i = 1, 2, \ldots, n$.

Budući da se procjenitelj $\hat{\beta}$ za koeficijent smjera β može prikazati kao linearna kombinacija nezavisnih normalnih varijabli Y_i, $i = 1, 2, \ldots, n$, normalno je distribuiran s očekivanjem i varijancom kao što je ranije navedeno. Nadalje, može se pokazati da su statistike $\hat{\beta}$ i $\hat{\sigma}^2$ nezavisne. Isti rezultati vrijede i za statistiku $\hat{\alpha}$. Još vrijedi:

$$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2).$$

Za takav, potpuno specificirani model mogu se tražiti MLE nepoznatih parametara. Uz dane pretpostavke proizlazi da su procjenitelji $\hat{\alpha}$ i $\hat{\beta}$ dobiveni metodom najmanjih kvadrata ujedno i MLE za te parametre, a da je MLE za σ^2 jednak $(n-2)\hat{\sigma}^2/n$.

10.2.5 Zaključivanje o koeficijentu smjera

Budući da su standardizirana verzija Z od $\hat{\beta}$,

$$Z = \frac{\hat{\beta} - \beta}{\sigma\sqrt{\frac{1}{S_{xx}}}} \sim N(0, 1),$$

i varijabla

$$U = \frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2)$$

nezavisne, studentizirana verzija T_β od $\hat{\beta}$ ima Studentovu razdiobu (vidjeti 6.4):

$$T_\beta = \frac{\hat{\beta} - \beta}{\hat{\sigma}\sqrt{\frac{1}{S_{xx}}}} = \frac{Z}{\sqrt{U/(n-2)}} \sim t(n-2). \tag{10.5}$$

Slučajna se varijabla T_β koristi kao pivotna veličina za konstrukciju pouzdanih intervala za parametar β, te za testiranje hipoteza o vrijednostima od β. Na primjer, želimo li testirati nulhipotezu da nema ovisnosti Y o X, tj. $H_0 : \beta = 0$, testna statistika će biti

$$\frac{\hat{\beta}}{\hat{\sigma}\sqrt{\frac{1}{S_{xx}}}} \overset{H_0}{\sim} t(n-2).$$

Primjer 10.5 Ponovo se vratimo primjeru 10.1. Na osnovi podataka iz tog primjera,

(a) procijenite 95%-pouzdan interval za koeficijent smjera regresijskog pravca β;

(b) testirajte

$$H_0 : \beta = 1, \quad H_1 : \beta \neq 1.$$

95%-pouzdan interval za β je

$$\hat{\beta} \pm t_{0.025}(n-2) \cdot \hat{\sigma}\sqrt{\frac{1}{S_{xx}}}.$$

Budući da je $t_{0.025}(8) = 2.306$, opažena vrijednost tog intervala je

$$0.8823 \pm 2.306 \cdot \sqrt{\frac{0.0732}{8.4440}} = 0.8823 \pm 0.2147.$$

Nadalje, budući da taj interval sadrži vrijednost "1", nulhipotezu H_0 iz (b) ne odbacujemo uz značajnost od 5%. □

10.2.6 Procjena i predviđanje srednjeg i individualnog odziva

Prvo, diskutirajmo problem predviđanja srednjeg odziva. Dakle, želimo procijeniti *očekivanu* vrijednost od Y za zadanu vrijednost x_0 od X,

$$\mathbb{E}[Y|X = x_0] = (\text{kraće}) = \mathbb{E}[Y|x_0] = \alpha + \beta x_0,$$

na osnovi podataka iz uzorka (10.1). Nepristrani procjenitelj za tu vrijednost je

$$\hat{\mathbb{E}}[Y|x_0] := \hat{\alpha} + \hat{\beta} x_0.$$

Varijanca tog procjenitelja je:

$$\mathbb{V}\text{ar}[\hat{\mathbb{E}}[Y|x_0]] = \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right).$$

Nije teško pokazati da studentizirana verzija od $\hat{\mathbb{E}}[Y|x_0]$ ima Studentovu razdiobu:

$$\frac{\hat{\mathbb{E}}[Y|x_0] - \mathbb{E}[Y|x_0]}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}}} = \frac{(\hat{\alpha} + \hat{\beta}x_0) - (\alpha + \beta x_0)}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}}} \sim t(n-2).$$

Ta se slučajna varijabla kristi kao pivotna veličina za konstrukciju pouzdanih intervala od $\mathbb{E}[Y|x_0]$.

Pretpostavimo sada da želimo procijeniti koliko bi iznosila jedna opservacija od Y za dano $X = x_0$, dakle, koliki bi bio iznos *individualnog* odziva Y, u oznaci Y_0, na poticaj $X = x_0$. Procjenitelj za tu *slučajnu* vrijednost (na osnovi podataka iz uzorka) je

$$\hat{Y}_0 := \hat{\alpha} + \hat{\beta}x_0.$$

Varijanca (slučajne) pogreške koja nastaje tom procjenom je:

$$\mathbb{V}\mathrm{ar}[\hat{Y}_0 - Y_0] = \mathbb{V}\mathrm{ar}[(\hat{\alpha} + \hat{\beta}x_0) - (\alpha + \beta x_0 + \varepsilon_0)] = \sigma^2(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}).$$

Nadalje, za studentiziranu verziju te pogreške procjene vrijedi:

$$\frac{\hat{Y}_0 - Y_0}{\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \sim t(n-2).$$

Ta se varijabla koristi kao pivotna veličina za konstrukciju pouzdanih intervala od Y_0. Primijetite da je dobiveni pouzdani interval za individualni odziv širi od odgovarajućeg pouzdanog intervala za srednji odziv.

Primjer 10.6 Ponovo se vratimo primjeru 10.1. Na osnovi podataka iz tog primjera,

(a) procijenite 95%-pouzdan interval za očekivanu vrijednost isplata za zahtjeve s iznosom jednakim 460 kn;

(b) procijenite 95%-pouzdan interval za vrijednost isplate ako je iznos zahtjeva jednak 460 kn.

Procjena očekivane (i individualne) vrijednosti isplate za dani iznos štete ($x_0 = 4.6$) je jednaka

$$\hat{\alpha} + \hat{\beta}x_0 = 0.1636 + 0.88231 \cdot 4.6 = 4.222,$$

dakle, 422.20 kn. Opažena vrijednost 95%-pouzdanog intervala za srednji iznos isplate (a) je

$$\hat{\mathbb{E}}[Y|4.6]] \pm t_{0.025}(8) \cdot \hat{\sigma}\sqrt{\frac{1}{10} + \frac{(4.6 - \bar{x})^2}{S_{xx}}} = 4.222 \pm 2.306 \cdot 0.1306 = 4.222 \pm 0.301,$$

a za individualni iznos isplate (b),

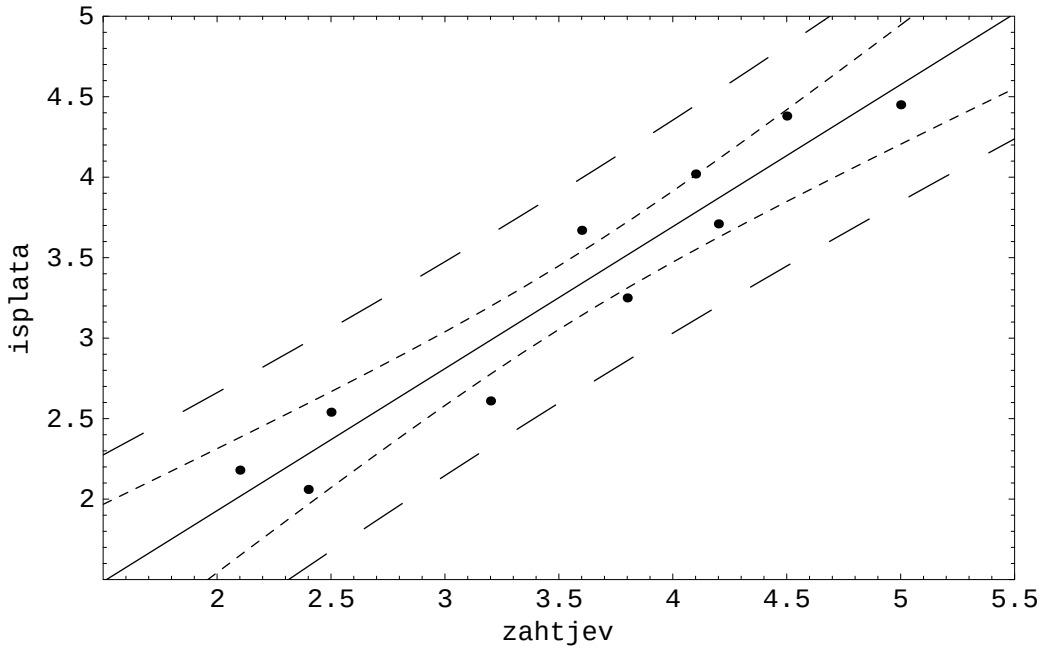
$$\hat{Y}_0 \pm t_{0.025}(8) \cdot \hat{\sigma}\sqrt{1 + \frac{1}{10} + \frac{(4.6 - \bar{x})^2}{S_{xx}}} = 4.222 \pm 2.306 \cdot 0.3004 = 4.222 \pm 0.693.$$

Dakle,

(a) uz 95% pouzdanosti očekivana (srednja) vrijednost isplata za štete od 460 kn bit će u intervalu od 392 do 452 kune, a

(b) uz 95% pouzdanosti vrijednost isplate za štetu od 460 kn bit će u intervalu od 353 do 492 kune.

Na slici (sljedeća stranica) navedene su granice 95%-pouzdanih intervala za srednju (točkasto) i individualnu (crtkano) vrijednost isplata. □



10.2.7 Provjera modela

Kako je već prije spomenuto, slučajne se pogreške u modelu (10.3) procjenjuju pomoću reziduala

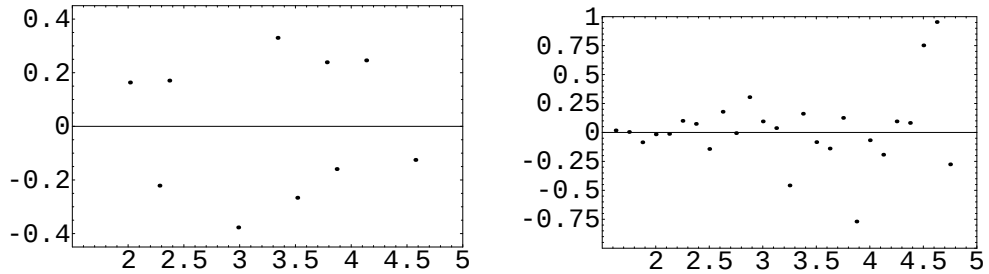
$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i, \quad i = 1, 2, \ldots, n.$$

Uvidom u opažene vrijednosti reziduala (dakle, na osnovi podataka (10.1), može se ispitati opravdanost pojedinih pretpostavki na model, posebno

(a) pretpostavke na slučajne pogreške ($A1 - 4$);

(b) pretpostavka na prirodu veze između varijabli X i Y.

Na primjer, prikazom reziduala pomoću linijskog grafa može se ispitati pretpostavka ($A4$) o normalnosti pogrešaka. Nadalje, adekvatnost modela (pretpostavka linearnost i/ili Gauss-Markovljevi uvjeti) može se ispitati prikazom reziduala u Kartezijevom koordinatnom sustavu u odnosu na pripadne procjene $\hat{y}_i$, odnosno uvidom u dijagram raspršenja točaka $(\hat{y}_i, \hat{\varepsilon}_i)$, $i = 1, 2, \ldots, n$. Na slici su prikazana dva dijagrama raspršenja reziduala. Lijevi dijagram prikazuje reziduale jednostavnog linearnog modela prilagođenog podacima iz primjera 10.1. Odsustvo bilo kakvog uzorka pokazuje da je pretpostavljeni model dobar. S druge strane, desni dijagram prikazuje reziduale neke prilagodbe koji indiciraju da je pretpostavka ($A2$) o homogenosti varijance pogrešaka neadekvatna.



10.2.8 Transformirani podaci

U nekim *modelima rasta* pretpostavlja se da očekivani odziv Y eksponencijalno ovisi o vrijednosti x varijable poticaja X:

$$\mathbb{E}[Y|x] = \alpha e^{\beta x}.$$

U tom slučaju se varijabla odziva Y transformira primjenom logaritamske funkcije $W = \log Y$. Na taj način dolazimo do linearnog modela za transformirane podatke (x_i, w_i), $i = 1, 2, \ldots, n$, koji želimo prilagoditi:

$$W_i = \eta + \beta x_i + \varepsilon_i, \quad i = 1, 2, \ldots, n.$$

Ovdje su $\eta = \log \alpha$ i β parametri regresijske funkcije. Primijetite da pretpostavka o aditivnosti slučajne pogreške u modelu za transformirane podatke povlači multiplikativnost pogreške u modelu za originalne, netransformirane podatke. Ako netransformirani podaci ne podržavaju pretpostavku o multiplikativnosti slučajnih pogrešaka, tada se druge od u ovom poglavlju opisanih metoda moraju primijeniti na njih.

10.3 Višestruki linearni regresijski model

U mnogim se problemima varijabla odziva Y može predviđati na osnovi više nezavisnih varijabli, recimo X_1, X_2,..., X_k. Najjednostavnija veza između Y i varijabli poticaja X_1, X_2,..., X_k je linearna:

$$\mathbb{E}[Y|X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k.$$

Parametri $\beta_1, \beta_2, \ldots, \beta_k$ zovu se *koeficijenti višestruke regresije*, a α je slobodni član. Njihove vrijednosti se procjenjuju iz uzoraka koji se sastoje od $k+1$-dimenzionalnih podataka.

Višestruki linearni regresijski model je:

$$Y_i = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon_i, \quad i = 1, 2, \ldots, n.$$

Kao i u slučaju jednostavnog linearnog regresijskog modela, parametri se procjenjuju metodom najmanjih kvadrata. Već za $k = 3$ ta metoda postaje dovoljno složena da se u efektivnim izračunima moraju koristiti računala.

Poglavlje 11

Analiza varijance

Analiza varijance se koristi u situacijama kada želimno uspoređivati parametre očekivanja više normalno distribuiranih populacija. Dakle, radi se o poopćenju problema opisanog u potpoglavlju 9.4.1.: da li su opažene razlike između sredina dvaju uzoraka slučajne ili odražavaju stvarnu razliku između populacijskih očekivanja.

Pretpostavka je da nas zanima kako k različitih *tretmana* djeluje na populacijsko očekivanje neke varijable Y. Budući da se u mnogim situacijama može dogoditi da stvarni efekt tretmana bude potisnut djelovanjem nekih vanjskih faktora koji nam nisu od interesa, važno je kako *dizajniramo* uzorak. Dobar dizajn se sastoji od *slučajnog* (vidjeti 6.1) dodjeljivanja tretmana ekperimentalnim jedinicama pri čemu se teži da broj jedinica podvrgnutih istom tretmanu bude što veći. Tako dizajnirani slučajni uzorak omogućit će analizu efekata i procjenu doprinosa raznih tretmana na parametar očekivanja promatrane varijable, a utjecaj raznih vanjskih faktora imat će efekt slučajnog šuma.

Tehnički, analiza varijance se sastoji od rastava ukupne varijabilnosti uzorka od Y na varijabilnost koja se može opisati utjecajem tretmana i na varijabilnost do koje dolazi zbog slučajnog šuma. Usporedba tih komponenti varijabilnosti omogućava testiranje nulhipoteze da promatrani tretmani ne utječu na populacijsko očekivanje od Y.

11.1 Jednofaktorska analiza varijance

11.1.1 Model

Uspoređujemo djelovanje k tretmana na razdiobu varijable Y. Slučajni uzorak se sastoji od varijabli Y_{ij} pri čemu indeks j označava da se radi o j-toj varijabli u poduzorku duljine n_i ($j = 1, 2, \ldots, n_i$) koji se odnosi na potpopulaciju opisanu djelovanjem i-tog tretmana ($i = 1, 2, \ldots, k$). Dakle, slučajni uzorak ukupne duljine $n = \sum_{i=1}^{k} n_i$ se sastoji od k nezavisnih poduzoraka duljina n_1,..., n_k, svaki iz populacije opisane djelovanjem jednog od k tretmana. Opaženu vrijednost varijable Y_{ij} označavamo sa y_{ij}.

Matematički model je

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}, \quad j = 1, 2, \ldots, n_i, \quad i = 1, 2, \ldots, k. \tag{11.1}$$

Pretpostavka je da su slučajne greške ε_{ij}, $j = 1, 2, \ldots, n_i$, $i = 1, 2, \ldots, k$, nezavisne, $N(0, \sigma^2)$-distribuirane. Dakle, prema tom modelu, slučajne greške ne ovise o tretmanu, varijable Y_{ij} su nezavisne i $Y_{ij} \sim N(\mu + \tau_i, \sigma^2)$. Parametri modela su *sveukupna populacijska sredina* μ za koju vrijedi da je

$$\mu = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \mathbb{E}[Y_{ij}], \tag{11.2}$$

te τ_i, odstupanje i-tog tretmana od μ ili *doprinos* (*efekt*) *i-tog tretmana*, za $i = 1, 2, \ldots, k$. Veza među efektima je

$$\sum_{i=1}^{k} n_i \tau_i = 0 \tag{11.3}$$

i ona je posljedica (11.1), pretpostavki na model i (11.2).

11.1.2 Procjena parametara

Parametri μ, τ_i, $i = 1, 2, \ldots, k$, modela (11.1) procjenjuju se na osnovi opaženog uzorka metodom najmanjih kvadrata, dakle, minimizacijom funkcije

$$q(\mu, \tau_1, \ldots, \tau_k) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \mu - \tau_i)^2$$

uz uvjet (11.3). Primijetite da zbog veze (11.3) slijedi da, u stvari, imamo k nepoznatih parametara, a ne $k + 1$. Dobiveni procjenitelji su:

$$\hat{\mu} = \overline{Y}_{..}, \quad \hat{\tau}_i = \overline{Y}_{i.} - \overline{Y}_{..}, \;\; i = 1, 2, \ldots, k,$$

gdje su statistike $\overline{Y}_{i.}$ i $\overline{Y}_{..}$ definirane izrazima:

$$\begin{aligned} \overline{Y}_{i.} &:= \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} \quad (\textit{uzoračka sredina za i-ti tretman}), \quad i = 1, 2, \ldots, k \\ \overline{Y}_{..} &:= \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij} = \frac{1}{n} \sum_{i=1}^{k} n_i \overline{Y}_{i.} \quad (\textit{sveukupna uzoračka sredina}). \end{aligned}$$

Primijetite da relacija (11.3) vrijedi i za procjenitelje efekata, dakle,

$$\sum_{i=1}^{k} n_i \hat{\tau}_i = 0.$$

Uzoračka varijanca poduzorka koji odgovara i-tom tretmanu je

$$S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i.})^2.$$

Za svaki i, S_i^2 je nepristrani procjenitelj za σ^2 i $(n_i - 1)S_i^2/\sigma^2 \sim \chi^2(n_i - 1)$. Nadalje, slučajne varijable S_1^2, S_2^2,..., S_k^2 su nezavisne, pa

$$\frac{1}{\sigma^2} \sum_{i=1}^{k} (n_i - 1) S_i^2 \sim \chi^2(n - k).$$

Budući da je

$$\mathbb{E}[\sum_{i=1}^{k} (n_i - 1) S_i^2] = \sum_{i=1}^{k} (n_i - 1) \mathbb{E}[S_i^2] = (n - k)\sigma^2,$$

zajednička uzoračka varijanca

$$\hat{\sigma}^2 := \frac{1}{n - k} \sum_{i=1}^{k} (n_i - 1) S_i^2 = \frac{1}{n - k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i.})^2$$

je nepristrani procjenitelj za varijancu pogrešaka σ^2.

Želimo testirati nulhipotezu da su populacijska očekivanja tretmana jednaka, tj. da tretmani nemaju utjecaja na populacijsko očekivanje od Y, u odnosu na alternativu da to nije tako. Dakle,

$$H_0 : \tau_i = 0 \text{ za svaki } i = 1, 2, \ldots, k, \quad H_1 : \tau_i \neq 0 \text{ za barem jedan } i \text{ od } 1, 2, \ldots, k.$$

11.1.3 Rastav varijance

Ukupna se uzoračka varijabilnost može rastaviti na dvije komponente. Jednu komponentu čini varijabilnost unutar svakog od poduzoraka određenih pojedinim tretmanom, a drugu varijabilnost do koje dolazi zbog razlika među tretmanima, preciznije, između njihovih uzoračkih sredina. Naime, vrijedi

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{..})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i.})^2 + \sum_{i=1}^{k} n_i(\overline{Y}_{i.} - \overline{Y}_{..})^2.$$

Ako označimo sa

$$\begin{aligned}
\text{SSTOT} &:= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{..})^2 \quad (\textit{ukupnu sumu kvadrata}) \\
\text{SST} &:= \sum_{i=1}^{k} n_i(\overline{Y}_{i.} - \overline{Y}_{..})^2 \quad (\textit{sumu kvadrata zbog razlike među tretmanima}) \\
\text{SSE} &:= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i.})^2 \quad (\textit{sumu kvadrata pogrešaka} \text{ ili } \textit{reziduala}),
\end{aligned}$$

tada se gornja relacija može zapisati i na način:

$$\text{SSTOT} = \text{SSE} + \text{SST}.$$

Primijetite da je $\hat{\sigma}^2 = \text{SSE}/(n-k)$. Statistiku $\hat{\sigma}^2$ još označavamo sa MSE i zovemo *srednjekvadratna greška*. Slično, statistiku $\text{MST} := \text{SST}/(k-1)$ zovemo *srednjekvadratno odstupanje zbog tretmana*.

Ako vrijedi nulhipoteza H_0, tada je $\text{SSTOT}/(n-1)$ uzoračka varijanca združenog uzorka koji reprezentira jednu normalno distribuiranu populaciju, pa je $\text{SSTOT}/\sigma^2 \sim \chi^2(n-1)$. U tom slučaju može se pokazati da su statistike MSE i MST nezavisne i $\text{SST}/\sigma^2 \sim \chi^2(k-1)$. Primijetite da je tada MST također nepristrani procjenitelj za σ^2. Dakle, testna statistika je

$$F = \frac{\text{MST}}{\text{MSE}} \overset{H_0}{\sim} F(k-1, n-k).$$

H_0 odbacujemo ako je opažena vrijednost f testne statistike F prevelika.

Razultati izračuna opaženih vrijednosti navedenih statistika prikazuju se u *ANOVA-tablici*:

izvor varijabilnosti	stupnjevi slobode	sume kvadrata	srednji kvadrati	test-stat.
zbog tretmana	$k-1$	SST	MST	f
slučajne greške	$n-k$	SSE	MSE	—
ukupno	$n-1$	SSTOT	—	—

Primjer 11.1 Iz svakog od tri osiguravajućeg društva A, B i C na slučajan način uzet je po uzorak polica osiguranja privatnih kuća. Zabilježene su osigurane svote po svakoj polici (u iznosima od po 100 kn):

društvo A: 36, 28, 32, 43, 30, 21, 33, 37, 26, 34
društvo B: 26, 21, 31, 29, 27, 35, 23, 33
društvo C: 39, 28, 45, 37, 21, 49, 34, 38, 44.

Želimo testirati nulhipotezu da su populacijske srednje vrijednosti osiguranih svota po policama osiguranja privatnih kuća jednake, odnosno, da izbor osiguravajućeg društva ne utječe na očekivani iznos osigurane svote po tim policama.

Duljine poduzoraka su $n_A = 10$, $n_B = 8$, $n_c = 9$, a ukupna duljina je $n = n_A+n_B+n_C = 10 + 8 + 9 = 27$. Uzoračke sredine i varijance svakog od poduzoraka su:

$$\begin{array}{ccc} \bar{y}_{A.} = 32.0000, & \bar{y}_{B.} = 28.1250, & \bar{y}_{C.} = 37.2222, \\ s_A^2 = 38.2222, & s_B^2 = 23.2679, & s_C^2 = 75.9444. \end{array}$$

Odavde slijedi da je sveukupna uzoračka sredina

$$\bar{y}_{..} = \frac{n_A\bar{y}_{A.} + n_B\bar{y}_{B.} + n_C\bar{y}_{C.}}{n} = \frac{10 \cdot 32.0000 + 8 \cdot 28.1250 + 9 \cdot 37.2222}{27} = 32.5926.$$

Nadalje, računamo:

$$\begin{aligned} \text{SST} &= n_A(\bar{y}_{A.} - \bar{y}_{..})^2 + n_B(\bar{y}_{B.} - \bar{y}_{..})^2 + n_C(\bar{y}_{C.} - \bar{y}_{..})^2 = \\ &= 10 \cdot (32. - 32.5926)^2 + 8 \cdot (28.125 - 32.5926)^2 + 9 \cdot (37.2222 - 32.5926)^2 = \\ &= 356.088 \\ \text{MST} &= \frac{\text{SST}}{k-1} = \frac{356.088}{3-1} = 178.044 \\ \text{SSE} &= (n_A - 1)s_A^2 + (n_B - 1)s_B^2 + (n_C - 1)s_C^2 = \\ &= 9 \cdot 38.2222 + 7 \cdot 23.2679 + 8 \cdot 75.9444 = \\ &= 1114.43 \\ \text{MSE} &= \frac{\text{SSE}}{n-k} = \frac{1114.43}{27-3} = 46.4346 \\ f &= \frac{\text{MST}}{\text{MSE}} = 3.8343. \end{aligned}$$

ANOVA-tablica:

izvor varijabilnosti	stupnjevi slobode	sume kvadrata	srednji kvadrati	test-stat.
zbog osig. društva	2	356.09	178.044	3.83
slučajne greške	24	1114.43	46.435	—
ukupno	26	1470.52	—	—

Uz pretpostavku da su ispunjeni uvjeti na model (11.1), želimo testirati nulhipotezu

$$H_0 : \tau_A = \tau_B = \tau_C = 0$$

u odnosu na alternativu da to nije tako. Budući da je $F \overset{H_0}{\sim} F(2, 24)$ i $f = 3.83$, p-vrijednost je $\mathbb{P}(F \geq 3.83|H_0) = 0.036$ pa možemo odbaciti H_0 uz razinu značajnosti od 5%. □

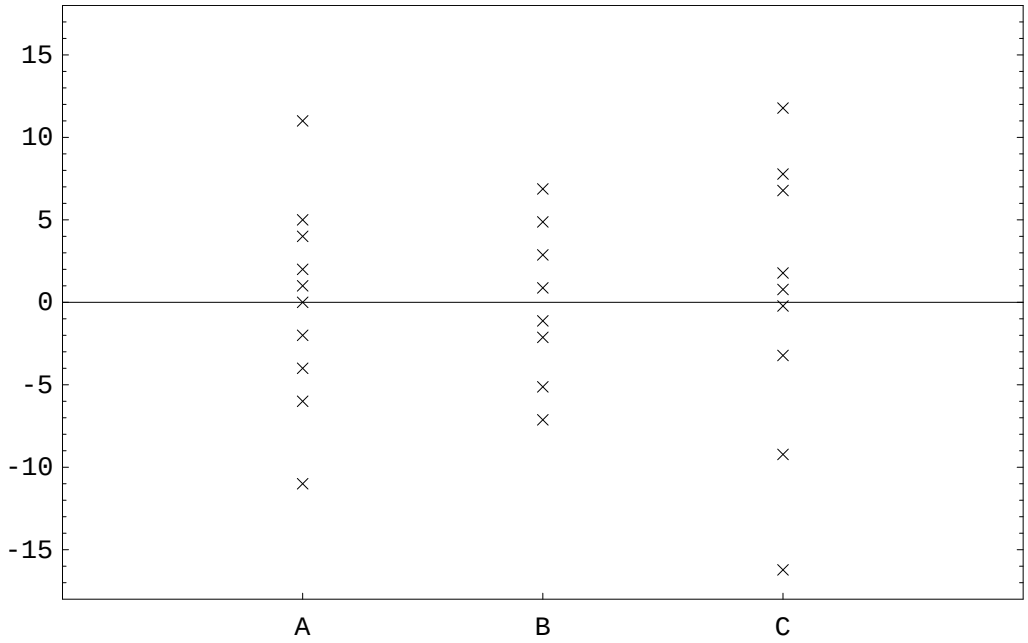
11.1.4 Provjera modela

Analizom procijenjenih veličina pogrešaka na osnovi prilagođenog ANOVA-modela, dakle, reziduala

$$\hat{\varepsilon}_{ij} := y_{ij} - \hat{y}_{ij} = y_{ij} - \hat{\mu} - \hat{\tau}_i = y_{ij} - \bar{y}_{i.}, \quad j = 1, 2, \ldots, n_i,$$

za svaki i posebno ($i = 1, 32, \ldots, k$), mogu se uočiti neadekvatnosti u pretpostavkama na model: pristranost u pogreškama (tj. sistematski utjecaj nekog vanjskog faktora), odstupanje od normalnosti slučajnih pogrešaka, te nehomogenost populacijskih varijanci pogrešaka. Na primjer, ako linijski dijagram reziduala ukazuje na postojanje pravilnosti (uzorka) u raspodjeli reziduala, vjerojatno se radi o pristranom uzorku. Nadalje, ako reziduali pokazuju odstupanje od normalnosti, adekvatnom transformacijom podataka se može postići normalnost pogreške. Na primjer, u slučaju pozitivno asimetričnog uzorka reziduala, obično je dobar izbor logaritamska transformacija. Primjenom adekvatne transformacije obično se rješava i problem nehomogenih varijanci pogrešaka kada veličina varijance za pojedini tretman ovisi o njegovoj srednjoj vrijednosti. Spomenimo da je F-test koji primjenjujemo za testiranje nulhipoteze o jednakosti srednjih vrijednosti tretmana, robustan na odstupanja od populacijske normalnosti i homogenosti varijance pogrešaka.

Primjer 11.1 (*nastavak*) Uvidom u usporedne linijske dijagrame reziduala, možemo zaključiti da je pretpostavka o homogenosti varijanci osiguranih svota promatranih polica između osiguravajućih društava A, B i C, neadekvatna. Neadekvatna je i pretpostavka o normalnosti osiguranih svota za društvo B.



Nadalje, uočimo da je

$$\bar{y}_{B.} < \bar{y}_{A.} < \bar{y}_{C.} \quad \text{i} \quad s_B^2 < s_A^2 < s_C^2,$$

dakle, da veličina procijenjih varijanci ovisi o procijenjenim srednjim vrijednostima u smislu da većim srednjim vrijednostima odgovaraju veće varijance. □

11.2 Analiza sredina tretmana

Pretpostavimo da nas zanima samo populacijsko očekivanje $\mu + \tau_i$ varijable Y za tretman i ($i = 1, 2, \ldots, k$). Pouzdani se interval za taj parametar konstruira pomoću studentizirane verzije statistike $\overline{Y}_{i.}$, a na osnovi sveukupnih podataka u uzorku. Dakle, 95%-pouzdan interval za $\mu + \tau_i$ je

$$\overline{Y}_{i.} \pm t_{0.025}(n-k)\frac{\hat{\sigma}}{\sqrt{n_i}}.$$

Želimo li uspoređivati parametre očekivanja dviju (od k različitih) potpopulacija, recimo za tretmane i i l ($i, l \in \{1, 2, \ldots, k\}$ i $i \neq l$), tada to činimo pomoću razlike $\overline{Y}_{i.} - \overline{Y}_{l.}$ njihovih procjenitelja. Vrijedi:

$$\mathbb{V}\mathrm{ar}[\overline{Y}_{i.} - \overline{Y}_{l.}] = \sigma^2(\frac{1}{n_i} + \frac{1}{n_l}).$$

95%-pouzdani interval za parametar $\tau_i - \tau_l$ je

$$\overline{Y}_{i.} - \overline{Y}_{l.} \pm t_{0.025}(n-k)\hat{\sigma}\sqrt{\frac{1}{n_i} + \frac{1}{n_l}}.$$

Konstrukcija 95%-pouzdanih intervala za sve moguće razlike parametara očekivanja tretmana nije preporučljiva jer je vjerojatnost istovremenog pokrivanja svih takvih intervala (a time i pouzdanost takve kombinacije intervala) manja od 0.95. S druge strane, budući da F-test kojim se testira nulhipoteza

$$H_0 : \tau_i = 0 \text{ za sve } i = 1, 2, \ldots, k,$$

u slučaju odbacivanja te hipoteze, ne kaže koje se dvije (ili više) srednjih vrijednosti tretmana značajno razlikuje, sasvim je prirodno zapitati se za koje grupe tretmana se njihove srednje vrijednosti značajno ne razlikuju. Način na koji se takve grupe određuju ilustrirat ćemo na primjeru.

Primjer 11.2 Na osnovi podataka iz primjera 11.1 odbacili smo nulhipoteza o homogenosti srednjih vrijednosti osiguranih svota polica osiguranja privatnih kuća u tri osiguravajuća društva. Želimo utvrditi koje se grupe osiguravajućih društava značajno razlikuju po srednjim vrijednostima izučavane varijable osigurane svote. Prvo, uredimo procjene tih srednjih vrijednosti po veličini:

$$\bar{y}_{B.} < \bar{y}_{A.} < \bar{y}_{C.}.$$

Zatim promotrimo prvi par po veličini u gornjem uređenom nizu. To su $\bar{y}_{B.} < \bar{y}_{A.}$. Za zadanu razinu značajnosti, recimo 5%, izračunajmo najmanju razliku između $\bar{y}_{A.}$ i $\bar{y}_{B.}$ za koju će razlika $\tau_A - \tau_B$ biti značajno različita od nule. Ta razlika je

$$t_{0.025}(n-k)\hat{\sigma}\sqrt{\frac{1}{n_A} + \frac{1}{n_B}} = 2.064 \cdot \sqrt{46.44} \cdot \sqrt{\frac{1}{10} + \frac{1}{8}} = 6.67.$$

Budući da je $\bar{y}_{A.} - \bar{y}_{B.} = 3.9 < 6.67$, $\mu + \tau_B$ i $\mu + \tau_A$ se značajno ne razlikuju. Tu činjenicu možemo grafički predočiti podcrtavanjem njihovih procjena:

$$\underline{\bar{y}_{B.} < \bar{y}_{A.}} < \bar{y}_{C.}.$$

Sada isti postupak ponovimo na sljedećem paru u uređenom nizu: $\bar{y}_{A.} < \bar{y}_{C.}$. Uz istu razinu značajnosti, najmanja razlika uz koju će razlika $\tau_C - \tau_A$ biti značajno različita od nule je:

$$t_{0.025}(n-k)\hat{\sigma}\sqrt{\frac{1}{n_C} + \frac{1}{n_A}} = 2.064 \cdot \sqrt{46.44} \cdot \sqrt{\frac{1}{9} + \frac{1}{10}} = 6.46.$$

Budući da je $\bar{y}_{C.} - \bar{y}_{A.} = 5.2 < 6.46$, $\mu + \tau_A$ i $\mu + \tau_C$ se značajno ne razlikuju. Dakle,

$$\underline{\bar{y}_{B.} < \bar{y}_{A.}} < \bar{y}_{C.}.$$

Primijetite da dobiveni rezultat nije u kontradikciji sa zaključkom testa budući da se $\mu+\tau_B$ i $\mu + \tau_C$ značajno razlikuju. Naime,

$$t_{0.025}(n-k)\hat{\sigma}\sqrt{\frac{1}{n_C} + \frac{1}{n_B}} = 2.064 \cdot \sqrt{46.44} \cdot \sqrt{\frac{1}{8} + \frac{1}{9}} = 6.8 \; < \; 9.1 = \bar{y}_{C.} - \bar{y}_{B.}.$$

□

11.3 Dodatne napomene

F-test u analizi varijance za usporedbu $k = 2$ tretmana je ekvivalentan t-testu iz potpoglavlja 9.4.1 (situacija 2.) za usporedbu dviju normalno distribuiranih populacija. Naime, veza između testnih statistika F iz ANOVA-e i T iz t-testa je $T^2 = F$.

Nadalje, analiza rastava varijance odziva u regresijskoj analizi linearnog modela (10.3) može se također prikazati u ANOVA-tablici:

izvor varijabilnosti	stupnjevi slobode	sume kvadrata	srednji kvadrati	test-stat.
zbog regresije	1	SSR	$\frac{\text{SSR}}{1}$	$\frac{\text{SSR}}{\text{SSE}/(n-2)}$
slučajne greške	$n-2$	SSE	$\frac{\text{SSE}}{n-2}$	—
ukupno	$n-1$	SSTOT	—	—

Primijetite da se nulhipoteza da odziv ne ovisi o nezavisnoj varijabli zapisuje u terminima iz poglavlja 9 kao $H_0 : \beta = 0$. t-test koji se koristi za testiranje te nulhipoteze je ekvivalentan F-testu iz ANOVA-e budući da za testnu statistiku

$$T = \frac{\hat{\beta}}{\hat{\sigma}\sqrt{\frac{1}{S_{xx}}}}$$

vrijedi relacija

$$T^2 = \frac{\text{SSR}}{\text{SSE}/(n-2)} \overset{H_0}{\sim} F(1, n-2).$$

S druge strane, analiza varijance k tretmana ekvivalentna je regresijskoj analizi modela u kojoj je varijabla Y zavisna u odnosu na $k-1$ nezavisnu varijablu koje sve poprimaju samo vrijednosti "0" ili "1".

Literatura

[1] *Subject* 101: *Statistical Modelling, Core Reading* 2000., Faculty and Institute of Actuaries

[2] *Subjects C*1/2: *Statistics, Core Reading* 1996., Faculty and Institute of Actuaries

[3] F. Daly, D.L. Hand, M.C. Jones, A.D. Lunn, K.J. McConway, *Elements of Statistics*, Addison-Wesley, 1995.

[4] E.L. Lehmann, *Testing Statistical Hypotheses*, 2nd edition, Springer, 1997.

[5] E.L. Lehmann, G. Casella, *Theory of Point Estimation*, 2nd edition, Springer, 1998.

[6] Ž. Pauše, *Uvod u matematičku statistiku*, Školska knjiga, Zagreb, 1993.

[7] I. Šošić, V. Serdar, *Uvod u statistiku*, Školska knjiga, Zagreb, 1992.

[8] J.E. Freund, *Mathematical Statistics*, Prentice Hall International, 1992.