Nikola Adžaga
Ana Martinčić Špoljarić
Nikola Sandrić

Vjerojatnost i statistika

Građevinski fakultet
Sveučilište u Zagrebu

Sadržaj

Predgovor

Poštovani čitatelji, pred vama se nalazi nastavni materijal za kolegij *Vjerojatnost i statistika* koji se predaje u trećem semestru Preddiplomskog studija na Građevinskom fakultetu Sveučilišta u Zagrebu. Cilj kolegija, pa samim time i ovog priručnika, je upoznati studente nematematičkih fakulteta s osnovnim pojmovima teorije vjerojatnosti i matematičke statistike te njihovim primjenama. Sam izbor tema obrađenih u priručniku je standardan i prati sadržaj uglavnom svih klasičnih udžbenika iz teorije vjerojatnosti i matematičke statistike.

Prvi dio priručnika posvećen je teoriji vjerojatnosti. Započinje osnovnim pojmom teorije vjerojatnosti – vjerojatnosnim prostorom (matematički model slučajnog pokusa). Nakon toga definiraju se pojmovi uvjetne vjerojatnosti i nezavisnosti događaja. U nastavku se uvodi pojam slučajnih varijabli (diskretnih i neprekidnih) te se diskutiraju njihova najosnovnija vjerojatnosna obilježja (funkcije vjerojatnosti, gustoće i raspodjele, očekivanje, varijanca i standardna devijacija). Također, daju se pregled i svojstva najbitnijih diskretnih i neprekidnih slučajnih varijabli. Nakon toga definiraju se pojmovi nezavisnosti i jednake distribuiranosti slučajnih varijabli te se interpretiraju zakoni velikih brojeva i centralni granični teorem. Na kraju, diskutiraju se slučajni vektori (s naglaskom na dvodimenzionalne slučajne vektore) te se uvode pojmovi kovarijance, koeficijenta korelacije i pravaca regresije.

Drugi dio priručnika posvećen je matematičkoj statistici. Započinje pregledom i diskusijom osnovnih pojmova deskriptivne statistike (mjere cetralne tendencije, raspršenja i oblika). Nastavlja se inferencijalnom statistikom gdje se ograničava na procjenjivanje dvije osnovne numeričke karakteristike varijabli (očekivanja i varijance) koristeći točkovne i intervalne procjene te testiranje statističkih hipoteza. Na kraju, obrađuju se i višedimenzionalne varijable (ponovno, naglasak je na dvodimenzionalnim varijablama) i analizira se povezanost varijabli primjenjujući Pearsonov koeficijent korelacije i linearna regresija.

Priručnik završava dodatkom u kojem je dan pregled osnovnih pojmova teorije skupova i kombinatorike, koji se koriste kroz čitav materijal.

Na kraju priručnika dan je pregled literature koji je korišten u pripremi ovog materijala. Također, isti može poslužiti zainteresiranom čitatelju koji želi produbiti znanje o temama obrađenim u ovom priručniku (a i više). Za strogo matematički pristup svakako preporučamo referencu [9], dok za ne previše matematički rigorozan pristup preporučamo reference [5, 6, 7, 8, 10].

U Zagrebu, 2017.

Nikola Adžaga
Ana Martinčić Špoljarić
Nikola Sandrić

Dio I

Teorija vjerojatnosti

Poglavlje 1

Vjerojatnosni prostor

Teorija vjerojatnosti je grana matematike koja proučava slučajne pojave. Centralni problem teorije vjerojatnosti jest opis i analiza slučajnih (nedeterminističkih) pokusa, koristeći stroge matematičke alate. Deterministički pokusi oni su pokusi čiji je ishod određen uvjetima u kojima se pokus izvršava. Primjerice, hlađenjem vode ispod 0° C voda mijenja agregatno stanje. S druge strane, slučajni pokusi nisu u potpunosti određeni uvjetima u kojima se pokus izvršava. Na primjer, ako bacamo igraću kocku ne znamo unaprijed ishod samog pokusa. Teorija vjerojatnosti proučava takve pokuse i pokušava odrediti (izmjeriti) neizvjesnosti događaja vezanih uz određeni slučajan pokus.

Prve snažne ideje (a i potrebe) za teorijom vjerojatnosti javile su se u kontekstu igara na sreću (bacanje novčića, bacanje igraće kocke, igre s kartama, itd.). U tim igrama ljudi su pokušavali izmjeriti neizvjesnosti pojedinih događaja i na osnovu toga okušati sreću u igri. Mnoge takve igre (slučajni pokusi) imaju zajedničke sljedeće dvije karakteristike:

(i) imaju najviše konačno mnogo ishoda

(ii) svi ishodi su jednako vjerojatni (izvjesni).

Primjerice, bacanje tri simetrična novčića, bacanje dvije simetrične kocke ili izvlačenje karte iz špila karata. Novčić i kocka su simetrični ako su svi ishodi jednako vjerojatni.

Definicija 1.1 (Vjerojatnost *a priori*)**.** Izvršavamo pokus koji ima najviše konačno mnogo ishoda koji su svi jednako vjerojatni. **Vjerojatnost *a priori*** događaja vezanog uz ovaj pokus se definira kao omjer broja povoljnih ishoda i broja svih ishoda.

Primjer 1.1. Slučajan pokus bacanja simetrične igraće kocke ima konačno ishoda (točnije šest) i svi su jednako vjerojatni (kocka je simetrična). Dakle, vjerojatnost *a priori* da je prilikom jednog bacanja kocke pao paran broj iznosi $3/6 = 1/2$. □

Problemi vjerojatnosti *a priori* su sljedeći:

(i) primjenjiva je samo na slučajne pokuse s konačno mnogo jednako vjerojatnih ishoda

(ii) sama definicija je kružna, tj. u definiciji koristimo pojam "vjerojatnost".

Definicija 1.2 (Vjerojatnost *a posteriori*)**.** Izvršavamo neki slučajan pokus. Neka je A događaj vezan uz taj pokus. **Vjerojatnost *a posteriori*** događaja A se definira kao

$$\lim_{n\to\infty} \frac{n_A}{n},$$

gdje je n_A broj pojavljivanja događaja A u n ponavljanja pokusa.

Napomenimo da je vjerojatnost *a posteriori* zapravo statistički pristup definiranju vjerojatnosti, tj. podrazumijeva svojstvo statističke stabilnosti relativnih frekvencija na čemu se temelji procjena vjerojatnosti događaja, o čemu će biti riječ u drugom dijelu priručnika. Vjerojatnost *a posteriori* rješava neke nedostatke vjerojatnosti *a priori*. Međutim, generira sljedeće probleme:

(i) oslanja se na beskonačno ponavljanje pokusa, tj. nepraktična je

(ii) ne možemo biti sigurni da gornji limes uopće postoji.

Konačno 1933. g. ruski matematičar A. N. Kolmogorov uvodi matematički ispravnu definiciju vjerojatnosti i postavlja fundamente teorije vjerojatnosti kao matematičke discipline. Izvršavamo neki slučajan pokus. Označimo s Ω skup svih mogućih ishoda tog pokusa. Elemente od Ω nazivamo **elementarnim događajima**, a sam skup Ω nazivamo **prostorom elementarnih događaja**. Elementarne događaje ćemo označavati s ω. Primjerice, u pokusu bacanja igrače kocke je $\Omega = \{1, 2, 3, 4, 5, 6\}$ i brojevi $1, 2, 3, 4, 5$ i 6 su elementarni događaji tog pokusa. Osim elementarnih događaja od interesa su i neki drugi objekti čije vjerojatnosti želimo računati, podskupovi od Ω. U Primjeru 1.1 $A = \{2, 4, 6\}$ je upravo skup čija vjerojatnost nas zanima. Prirodno bi bilo za pretpostaviti da je to moguće učiniti za svaki podskup od Ω, tj. za svaki element partitivnog skupa $\mathcal{P}(\Omega)$ od Ω. Međutim, to nije uvijek moguće. U slučaju kada je Ω diskretan (konačan ili prebrojiv) to

možemo napraviti, ali inače ne. Pokazuje se da kada je Ω neprebrojiv (primjerice $\Omega = \mathbb{R}$), $\mathcal{P}(\Omega)$ ima "previše" elemenata i vrlo je teško pojmiti kakve sve podskupove možemo konstruirati (vidi [9, str. 13]). Stoga, kao skup podskupova od Ω čiju vjerojatnost želimo računati nećemo uzeti najveći mogući skup (tj. $\mathcal{P}(\Omega)$), već najmanji skup koji ima svojstva/strukturu koja su u skladu s prirodom problema koje proučavamo. Neka je $\mathcal{F}$ skup podskupova od Ω koja zadovoljava sljedeće:

(i) $\emptyset \in \mathcal{F}$

(ii) $A \in \mathcal{F} \implies A^c \in \mathcal{F}$

(iii) $A_1, A_2, \ldots \in \mathcal{F} \implies \cup_{n=1}^{\infty} A_n \in \mathcal{F}$.

Skup $\mathcal{F}$ nazivamo **prostorom događaja**, a njegove elemete **događajima**. Napomenimo ovdje da bilo koju familiju skupova koja zadovoljava svojstva (i), (ii) i (iii) nazivamo σ-algebrom. Iz gornjih svojstava lagano se izvede sljedeće:

(i) $\Omega \in \mathcal{F}$

(ii) $A_1, A_2, \ldots \in \mathcal{F} \implies \cap_{n=1}^{\infty} A_n \in \mathcal{F}$

(iii) $A, B \in \mathcal{F} \implies A \setminus B \in \mathcal{F}$.

Dakle, $(\Omega, \mathcal{F})$ opisuje događaje vezane uz naš pokus. U primjeru bacanja igraće kocke imamo

$$\Omega = \{1, 2, 3, 4, 5, 6\} \quad \text{i} \quad \mathcal{F} = \mathcal{P}(\Omega) = \{\emptyset, \{1\}, \{2\}, \ldots, \{6\}, \{1, 2\}, \ldots, \Omega\}.$$

Sada želimo uvesti definiciju vjerojatnosti, tj. jedne mjere neizvjesnosti događaja.

Definicija 1.3. Vjerojatnost je funkcija $\mathbb{P} : \mathcal{F} \to [0, 1]$ koja zadovoljava sljedeće:

(i) $\mathbb{P}(\Omega) = 1$

(ii) $A_1, A_2, \ldots \in \mathcal{F}$ disjunktni $\implies \mathbb{P}(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mathbb{P}(A_n)$.

Trojku $(\Omega, \mathcal{F}, \mathbb{P})$ nazivamo **vjerojatnosnim prostorom**.

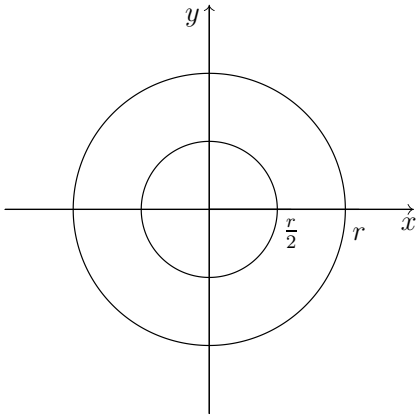
Problem ovako definirane vjerojatnosti, za razliku od vjerojatnosti *a priori* i vjerojatnosti *a posteriori*, je što za dani slučajni pokus ne daje definiciju funkcije $\mathbb{P}$. Drugim riječima, sama definicija kaže da je $\mathbb{P}$ nešto što zadovoljava (i) i (ii). Međutim, ovaj problem rješava matematička statistika i time ćemo se baviti u drugom dijelu kolegija.

Primjer 1.2. Bacamo simetričan novčić. Označimo s P padanje pisma, a s G padanje glave. Dakle, $\Omega = \{P, G\}$, $\mathcal{F} = \mathcal{P}(\Omega)$ i zbog simetrije za $\mathbb{P}$ je prirodno uzeti vjerojatnost *a priori*, tj. $\mathbb{P}(\{\omega\}) = 1/2$ za $\omega \in \Omega$. □

Primjer 1.3. U pokusu bacanja simetrične kocke imamo $\Omega = \{1, 2, 3, 4, 5, 6\}$, $\mathcal{F} = \mathcal{P}(\Omega)$ te, ponovno zbog simetrije, $\mathbb{P}$ definiramo kao $\mathbb{P}(\{\omega\}) = 1/6$ za $\omega \in \Omega$. □

Napomenimo ovdje da je za definiranje vjerojatnosti $\mathbb{P}$ vezane uz diskretan Ω dovoljno definirati $\mathbb{P}$ za elementarne događaje, a da vjerojatnosti ostalih događaja onda slijede iz svojstva (iii) u definiciji prostora događaja.

Primjer 1.4. Unutar kruga radijusa $r > 0$ biramo točku na slučajan način. Odredimo vjerojatnost da se odabrana točka nalazi bliže rubu kruga nego centru. U ovom slučaju je $\Omega = \{\omega = (x, y) \in \mathbb{R}^2 : x^2 + y^2 \leqslant r^2\}$. Međutim za $\mathcal{F}$ ne možemo uzeti $\mathcal{P}(\Omega)$, ali se time sad nećemo zamarati.



Slika 1.1: Koncentrični krugovi radijusa r i $r/2$ iz Primjera 1.4

Zaima nas vjerojatnost događaja

$$A = \left\{\omega = (x, y) \in \Omega : x^2 + y^2 > \frac{r^2}{4}\right\}$$

(točka se nalazi bliže rubu ako se nalazi izvan kruga radijusa $r/2$, vidi Sliku 1.1). Stoga možemo zaključiti da je tražena vjerojatnost

$$\mathbb{P}(A) = \frac{r^2\pi - \frac{r^2}{4}\pi}{r^2\pi} = \frac{3}{4}.$$

Uočimo da smo ovdje, u duhu definicije vjerojatnosti *a priori*, za $A \in \mathcal{F}$ stavili

$$\mathbb{P}(A) = \frac{\text{pov}(A)}{\text{pov}(\Omega)},$$

gdje pov(A) predstavlja površinu skupa A. To je opravdano uz pretpostavku uniformnosti (svaka točka je "jednako vjerojatna"). □

Gornji primjer sugerira da na vjerojatnost trebamo gledati kao na mjeru, jer u biti ona i jest mjera, mjera neizvjesnosti. Dajmo sada neka svojstva vjerojatnosti. Za sve $A, B \in \mathcal{F}$ vrijedi:

(i) $A \subseteq B \implies \mathbb{P}(A) \leqslant \mathbb{P}(B)$

Dokaz. Kako je $B = A \cup (B \setminus A)$ te su A i $B \setminus A$ disjunktni, iz svojstva (ii) vjerojatnosti imamo

$$\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A) \geqslant \mathbb{P}(A).$$ □

(ii) $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$

Dokaz. Očito, $\Omega = A \cup A^c$. Sada, zbog disjunktnosti od A i A^c imamo

$$1 = \mathbb{P}(\Omega) = \mathbb{P}(A) + \mathbb{P}(A^c).$$ □

(iii) $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$

Dokaz. Prvo uočimo da je $A \cup B = A \cup (B \setminus A)$ i $B = (A \cap B) \cup (B \setminus A)$. Sada, iz disjunktnosti od A i $(B \setminus A)$ te $(A \cap B)$ i $(B \setminus A)$, imamo

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A) \quad \text{i} \quad \mathbb{P}(B) = \mathbb{P}(A \cap B) + \mathbb{P}(B \setminus A).$$

Konačno, oduzimajući ove relacije dobivamo

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).$$ □

Poglavlje 2

Uvjetna vjerojatnost i nezavisnost događaja

Promotrimo sljedeći slučajan pokus. Bacamo simetričnu kocku. Dakle, naš vjerojatnostni prostor je $(\Omega, \mathcal{F}, \mathbb{P})$, gdje je $\Omega = \{1, 2, 3, 4, 5, 6\}$, $\mathcal{F} = \mathcal{P}(\Omega)$ i $\mathbb{P}(\{\omega\}) = 1/6$ za $\omega \in \Omega$. Pretpostavimo sad da imamo dodatnu informaciju vezanu za ishode ovog pokusa. Primjerice, netko nam je iz budućnosti javio da će prilikom sljedećeg bacanja kocke pasti paran broj. Dakle, uz tu informaciju $(\Omega, \mathcal{F}, \mathbb{P})$ više ne opisuje dobro naš pokus. Pronađimo adekvatnu zamjenu. Stavimo $A = \{2, 4, 6\}$ i $\mathcal{F}_A = \mathcal{P}(A)$. Očito par $(A, \mathcal{F}_A)$ rješava problem prostora elementarnih događaja i događaja. Kako odabrati $\mathbb{P}_A$ na $\mathcal{F}_A$? Logičan izbor je

$$\mathbb{P}_A(B) = \frac{\mathbb{P}(B)}{\mathbb{P}(A)}, \qquad B \in \mathcal{F}_A.$$

Sada vjerojatnosni prostor $(A, \mathcal{F}_A, \mathbb{P}_A)$ adekvatno opisuje slučajni pokus bacanja simetrične igraće kocke uz dodatnu informaciju da će pri bacanju pasti paran broj. Vjerojatnost $\mathbb{P}_A(B)$, koju još označavamo i s $\mathbb{P}(B|A)$, čitamo kao vjerojatnost događaja B uz uvjet da se dogodio događaj A. U općenitoj situaciji činimo analognu zamjenu, tj. za vjerojatnosni prostor $(\Omega, \mathcal{F}, \mathbb{P})$ i $A \in \mathcal{F}$ t.d. $\mathbb{P}(A) > 0$, **uvjetni vjerojatnosni prostor** je $(A, \mathcal{F}_A, \mathbb{P}(\cdot|A))$, gdje je

$$\mathcal{F}_A = \{B \cap A : B \in \mathcal{F}\} \qquad \text{i} \qquad \mathbb{P}(C|A) = \frac{\mathbb{P}(C)}{\mathbb{P}(A)}, \quad C \in \mathcal{F}_A.$$

Provjerite da je $\mathbb{P}(\cdot|A)$ zaista vjerojatnost. Uočimo da $(A, \mathcal{F}_A, \mathbb{P}(\cdot|A))$ možemo zamijeniti s $(A, \mathcal{F}, \mathbb{P}(\cdot|A))$, gdje je

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)}, \qquad B \in \mathcal{F}.$$

Primjer 2.1. Na grupi od 60 muškaraca i 40 žena provedeno je istraživanje o sklonosti određenom proizvodu i dobiveni su sljedeći podaci (V = voli, N = ne voli), prikazani u Tablici 2.1:

	V	N	Σ
M	15	45	60
Ž	4	36	40
Σ	19	81	100

Tablica 2.1: Podaci o sklonosti proizvodu ovisno o spolu

Na slučajan način biramo osobu iz grupe. Vjerojatnost odabira bilo koje osobe jednaka je i iznosi 1/100 (vjerojatnost *a priori*). Kolika je vjerojatnost da slučajno odabrana osoba voli taj proizvod? Označimo s V skup svih osoba koje vole taj proizvod. Tada imamo

$$\mathbb{P}(V) = \frac{19}{100}.$$

Kolika je vjerojatnost da osoba voli taj proizvod ako znamo da je osoba muškarac? Označimo s M skup svih muškaraca. Tada

$$\mathbb{P}(V|M) = \frac{\mathbb{P}(V \cap M)}{\mathbb{P}(M)} = \frac{\frac{15}{100}}{\frac{60}{100}} = \frac{15}{60}.$$

A ako je žena? Označimo s $\check{Z}$ skup svih žena. Tada imamo

$$\mathbb{P}(V|\check{Z}) = \frac{\mathbb{P}(V \cap \check{Z})}{\mathbb{P}(\check{Z})} = \frac{\frac{4}{100}}{\frac{40}{100}} = \frac{4}{40}.$$

Kolika je vjerojatnost da je osoba koja voli taj proizvod žena, a kolika da je osoba koja ne voli taj proizvod muškarac? Označimo s N skup svih osoba koje ne vole taj proizvod. Dakle,

$$\mathbb{P}(\check{Z}|V) = \frac{\mathbb{P}(V \cap \check{Z})}{\mathbb{P}(V)} = \frac{\frac{4}{100}}{\frac{19}{100}} = \frac{4}{19} \quad \text{i} \quad \mathbb{P}(M|N) = \frac{\mathbb{P}(N \cap M)}{\mathbb{P}(N)} = \frac{\frac{45}{100}}{\frac{81}{100}} = \frac{45}{81}. \quad \square$$

Definicija 2.1. Za događaje A i B kažemo da su **nezavisni** ako je

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

Uočimo, ako je $\mathbb{P}(A) > 0$ (ili $\mathbb{P}(B) > 0$), onda nezavisnost od A i B povlači $\mathbb{P}(B|A) = \mathbb{P}(B)$ (ili $\mathbb{P}(A|B) = \mathbb{P}(A)$). Dakle, ako su A i B nezavisni, informacija o A (ili B) nam ne donosi ništa važno prilikom računanja vjerojatnosti događaja B (ili A).

Primjer 2.2. Jesu li sklonosti proizvodu iz Primjera 2.1 neovisne o spolu? Imamo

$$\mathbb{P}(V) = \frac{19}{100}, \quad \mathbb{P}(\check{Z}) = \frac{40}{100}, \quad \mathbb{P}(M) = \frac{60}{100},$$
$$\mathbb{P}(V \cap \check{Z}) = \frac{4}{100} \qquad \text{i} \qquad \mathbb{P}(V \cap M) = \frac{15}{100},$$

iz čega slijedi

$$\mathbb{P}(V \cap \check{Z}) = \frac{4}{100} \neq \mathbb{P}(V)\mathbb{P}(\check{Z}) = \frac{38}{500}$$
$$\mathbb{P}(V \cap M) = \frac{15}{100} \neq \mathbb{P}(V)\mathbb{P}(M) = \frac{57}{500}.$$

Dakle, sklonosti (događaji) nisu nezavisne. □

Iz nezavisnosti događaja A i B lagano se može zaključiti nezavisnost događaja A^c i B, A i B^c te A^c i B^c. Pokažimo da nezavisnost od A i B povlači nezavisnost od A^c i B. Preostala dva slučaja se pokazuju analogno. Imamo

$$\begin{aligned}\mathbb{P}(A^c \cap B) &= \mathbb{P}(B \setminus A)\\ &= \mathbb{P}(B \setminus (A \cap B))\\ &= \mathbb{P}(B) - \mathbb{P}(A \cap B)\\ &= \mathbb{P}(B) - \mathbb{P}(A)\mathbb{P}(B)\\ &= \mathbb{P}(A^c)\mathbb{P}(B),\end{aligned}$$

gdje smo u prvom koraku koristili činjenicu da je $A^c \cap B = B \setminus A$, u drugom da je $B \setminus A = B \setminus (A \cap B)$, u trećem da je $B = (A \cap B) \cup (B \setminus (A \cap B))$ i da su $A \cap B$ i $B \setminus (A \cap B)$ disjunktni, u četvrtom smo iskoristili pretpostavku nezavisnosti od A i B te, konačno, u zadnjem koraku smo iskoristili svojstvo vjerojatnosti $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.

Primjer 2.3. Bacamo dvije simetrične igraće kocke. Dakle, prostor elementarnih događaja je $\Omega = \{\omega = (i,j) : i,j = 1,\ldots,6\}$ i ima 36 elemenata. Definirajmo sljedeće događaje:

$$\begin{aligned}A &= \{\omega = (i,j) \in \Omega : i = 4\}\\ B_1 &= \{\omega = (i,j) \in \Omega : j = 2\}\\ B_2 &= \{\omega = (i,j) \in \Omega : i + j = 3\}\\ B_3 &= \{\omega = (i,j) \in \Omega : i + j = 9\}\\ B_4 &= \{\omega = (i,j) \in \Omega : i + j = 7\}.\end{aligned}$$

Je li koji od događaja B_1, B_2, B_3 i B_4 zavisan o događaju A? Zapišimo u Tablicu 2.2 zbroj brojeva za svaki od mogućih ishoda:

1. / 2.	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Tablica 2.2: Zbroj brojeva na kockama za svaki mogući ishod

Vrijedi da je

$$\mathbb{P}(A) = \frac{1}{6}, \quad \mathbb{P}(B_1) = \frac{1}{6}, \quad \mathbb{P}(B_2) = \frac{1}{18}, \quad \mathbb{P}(B_3) = \frac{1}{9} \quad \text{i} \quad \mathbb{P}(B_4) = \frac{1}{6}.$$

Odredimo sad vjerojatnosti presjeka. Imamo

$$\mathbb{P}(A \cap B_1) = \frac{1}{36} = \mathbb{P}(A)\mathbb{P}(B_1), \quad \mathbb{P}(A \cap B_2) = 0 \neq \mathbb{P}(A)\mathbb{P}(B_2),$$
$$\mathbb{P}(A \cap B_3) = \frac{1}{36} \neq \mathbb{P}(A)\mathbb{P}(B_3) \quad \text{i} \quad \mathbb{P}(A \cap B_4) = \frac{1}{36} = \mathbb{P}(A)\mathbb{P}(B_4).$$

Dakle, događaji A i B_1 te A i B_4 su nezavisni, dok su događaji A i B_2 te A i B_3 zavisni. □

Neka su A i B neki događaji. Mogući su sljedeći slučajevi:

(i) barem jedan od događaja ima vjerojatnost 0, tj. $\mathbb{P}(A) = 0$ ili $\mathbb{P}(B) = 0$, i onda su A i B uvijek nezavisni.

(ii) oba imaju pozitivnu vjerojatnost, tj. $\mathbb{P}(A) > 0$ i $\mathbb{P}(B) > 0$. Tada A i B mogu biti disjunktni (onda su nužno zavisni), nezavisni (onda nisu disjunktni) i mogu biti zavisni (onda mogu, ali ne moraju biti disjunktni).

Definicija 2.2. Za događaje A_i, $i \in \mathbb{N}$, kažemo da su **nezavisni** ako za svaki izbor indeksa $1 \leqslant i_1 < i_2 < \cdots < i_n$, $n \in \mathbb{N}$, vrijedi

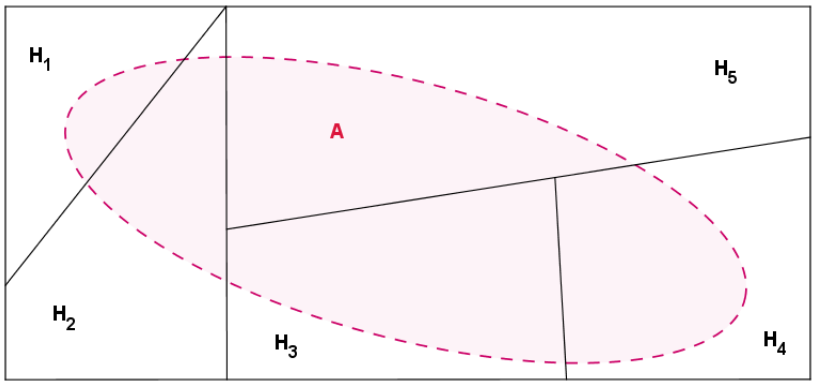
$$\mathbb{P}(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_n}) = \mathbb{P}(A_{i_1})\mathbb{P}(A_{i_2}) \cdots \mathbb{P}(A_{i_n}).$$

Koristeći definiciju uvjetne vjerojatnosti lagano možemo izvesti često puta vrlo korisnu tzv. **formulu produkta vjerojatnosti**. Neka su A_i, $i = 1, \ldots, n$, događaji. Tada vrijedi

$$\mathbb{P}(A_1 \cap A_2 \cap \cdots \cap A_n) = \mathbb{P}(A_1)\mathbb{P}(A_2|A_1)\cdots\mathbb{P}(A_n|A_1 \cap \cdots \cap A_{n-1}).$$

Neka je sada $(\Omega, \mathcal{F}, \mathbb{P})$ vjerojatnosni prostor i neka su H_i, $i = 1 \ldots, n$, disjunktni događaji takvi da vrijedi $\cup_{i=1}^n H_i = \Omega$ i $\mathbb{P}(H_i) > 0$, $i = 1, \ldots, n$. Takvu familiju događaja nazivamo **potpun sustav događaja**. Uočimo da za svaki $A \in \mathcal{F}$ vrijedi

$$\begin{aligned}\mathbb{P}(A) &= \mathbb{P}((A \cap H_1) \cup (A \cap H_2) \cup \cdots \cup (A \cap H_n)) \\ &= \mathbb{P}(A \cap H_1) + \cdots + \mathbb{P}(A \cap H_n) \\ &= \mathbb{P}(A|H_1)\mathbb{P}(H_1) + \cdots + \mathbb{P}(A|H_n)\mathbb{P}(H_n).\end{aligned}$$



Slika 2.1: Grafički prikaz potpunog sustava događaja i formule potpune vjerojatnosti

Dakle, gornja formula daje vjerojatnost događaja A ako znamo njegovu vjerojatnost uvjetovanu događajima H_i, $i = 1, \ldots, n$. Formulu zovemo **formula potpune vjerojatnosti** (vidi Sliku 2.1). Iz formule potpune vjerojatnosti izravno slijedi i tzv. **Bayesova formula**:

$$\mathbb{P}(H_i|A) = \frac{\mathbb{P}(H_i \cap A)}{\mathbb{P}(A)} = \frac{\mathbb{P}(H_i)\mathbb{P}(A|H_i)}{\sum_{j=1}^n \mathbb{P}(H_j)\mathbb{P}(A|H_j)}, \quad i = 1, \ldots, n.$$

Gornja formula daje vjerojatnost uzroka H_i uz danu posljedicu A.

Primjer 2.4. Neki proizvod izrađuje se na tri stroja. Na prvom se izrađuje 40% ukupne proizvodnje i od toga je 0.1% neispravnih proizvoda, na drugom

stroju izrađuje se 35% ukupne proizvodnje i od toga je 0.2% neispravnih proizvoda, a na trećem stroju se izrađuje 25% ukupne proizvodnje i od toga je 0.25% neispravnih proizvoda. Kolika je vjerojatnost da nasumično odabrani proizvod bude neispravan? Kolika je vjerojatnost da uzrok neispravnosti bude drugi stroj? Stavimo

$$\begin{aligned} H_1 &\ldots \text{“proizvod je izrađen na prvom stroju”} \\ H_2 &\ldots \text{“proizvod je izrađen na drugom stroju”} \\ H_3 &\ldots \text{“proizvod je izrađen na trećem stroju”} \\ A &\ldots \text{“proizvod je neispravan”}. \end{aligned}$$

Očito je

$$\begin{aligned} &\mathbb{P}(H_1) = 0.4, \quad \mathbb{P}(H_2) = 0.35, \quad \mathbb{P}(H_3) = 0.25, \\ &\mathbb{P}(A|H_1) = 0.001, \quad \mathbb{P}(A|H_2) = 0.002, \quad \text{i} \quad \mathbb{P}(A|H_3) = 0.0025. \end{aligned}$$

Sada, koristeći formulu potpune vjerojatnosti i Bayesovu formulu dobijemo

$$\begin{aligned} &\mathbb{P}(A) = \mathbb{P}(A|H_1)\mathbb{P}(H_1) + \mathbb{P}(A|H_2)\mathbb{P}(H_2) + \mathbb{P}(A|H_3)\mathbb{P}(H_3) = 0.001725 \quad \text{i} \\ &\mathbb{P}(H_2|A) = \frac{\mathbb{P}(H_2)\mathbb{P}(A|H_2)}{\mathbb{P}(A)} = 0.4058. \end{aligned}$$

□

Poglavlje 3

Slučajne varijable

Teorija vjerojatnosti bazirana samo na definiciji vjerojatnosnog prostora nije dovoljno moćan alat. Željeli bismo dalje "matematizirati" vjerojatnosni prostor. Stoga na prostoru elementarnih događaja definiramo funkciju koja elementarne događaje preslikava u relane brojeve i koju, uz određena dodatna svojstva u koja ovdje nećemo ulaziti, zovemo slučajna varijabla na vjerojatnosnom prostoru $(\Omega, \mathcal{F}, \mathbb{P})$.

U matematičkoj analizi varijabla predstavlja proizvoljnu (promjenjivu) deterministčku veličinu, dok u teoriji vjerojatnosti nemamo determinističnost pa je varijabla koju pridružujemo nekom pokusu slučajna. Također, u matematičkoj analizi za određeni x_0 iz domene funkcije $f(x)$ znamo izračunati $f(x_0)$, tj. odrediti vrijednost funkcije $f(x)$ u točki x_0. S druge strane, slučajna varijabla u pravilu nema to svojstvo. Naime, nju najčešće ne zadajemo analitički (formulom) pa njezine vrijednosti za konkretne elementarne događaje ne računamo, ali uvijek nam je poznat skup svih njezinih vrijednosti (slika slučajne varijable) i samim time ono što u pravilu možemo izračunati je vjerojatnost da slučajna varijabla poprima neku određenu vrijednost ili da je njena vrijednost u nekom skupu.

Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ vjerojatnosni prostor. **Slučajna varijabla** je funkcija

$$X : \Omega \to \mathbb{R}.$$

Napomenimo ovdje da u slučaju kada je Ω diskretan, X zaista može biti bilo koja funcija sa Ω u $\mathbb{R}$. S druge strane, kada je Ω neprebrojiv pokazuje se da svih mogućih funkcija sa Ω u $\mathbb{R}$ ima "previše", što je usko povezano s činjenicom da u slučaju neprebrojivog Ω za prostor događaja ne možemo uzeti $\mathcal{P}(\Omega)$. Uočimo da u slučaju diskretnog Ω prostor događaja je uvijek $\mathcal{P}(\Omega)$. Da bi razriješili ovaj problem postavljamo dodatni zahtjev na X, a to je "izmjerivost" obzirom na $\mathcal{F}$ ($\mathcal{F}$ je manji od $\mathcal{P}(\Omega)$). Drugim riječima, slučajna varijabla je "izmjeriva" (obzirom na $\mathcal{F}$) funkcija $X : \Omega \to \mathbb{R}$. Pojam

izmjerivosti funkcija izlazi van okvira ovog priručnika te se istime nećemo baviti (vidi [9, str. 235] za detalje).

Slučajne varijable obzirom na njihovu sliku $R(X)$ dijelimo na:

(i) **diskretne** – slika $R(X)$ je konačan ili prebrojiv skup (npr. broj automobila, broj ljudi, itd.)

(ii) **neprekidne** – slika $R(X)$ je neprebrojiv skup (npr. količina vode, vrijeme, itd.).

3.1 Diskretne slučajne varijable

Kao što smo već spomenuli, slučajna varijabla X je diskretna ako je $R(X)$ diskretan skup. Uočimo da, budući je $R(X)$ diskretan, vjerojatnosti svih događaja vezanih uz X možemo u potpunosti opisati vjerojatnostima događaja da X poprimi baš vrijednost $x \in R(X)$, tj. $\{\omega \in \Omega : X(\omega) = x\}$. Taj događaj ćemo skraćeno zapisivati s $\{X = x\}$. Općenito, ako je $X : \Omega \to \mathbb{R}$ neka slučajna varijabla (diskretna ili neprekidna) i ako je A ("izmjeriv") podskup od $\mathbb{R}$, onda događaj $\{\omega \in \Omega : X(\omega) \in A\}$ skraćeno zapisujemo kao $\{X \in A\}$.

Promotrimo sada nekoliko primjera.

Primjer 3.1. Promatramo pokus bacanja simetrične kocke. Neka je $X : \Omega \to \mathbb{R}$ slučajna varijabla koja poprima vrijednosti koje se okrenu pri bacanju kocke. Dakle,

$$R(X) = \{1, 2, 3, 4, 5, 6\} \qquad \text{i} \qquad X(\omega) = \omega, \quad \omega \in \Omega.$$

Uočimo da je npr. $\mathbb{P}(X = 2) = \mathbb{P}(\{2\}) = 1/6$. □

Primjer 3.2. Promatramo pokus bacanja simetrične kocke. Zanima nas vjerojatnost događaja "pao je paran broj". Neka je $Y : \Omega \to \mathbb{R}$ definirana s

$$Y(\omega) = \begin{cases} 1, & \omega \text{ je paran broj} \\ 0, & \omega \text{ je neparan broj.} \end{cases}$$

Dakle, $R(Y) = \{0, 1\}$ i $\mathbb{P}(\{2, 4, 6\}) = \mathbb{P}(Y = 1) = 1/2$. □

Primjer 3.3. Promatramo pokus bacanja dvije simetrične kocke. Neka je $Z : \Omega \to \mathbb{R}$ slučajna varijabla koja poprima vrijednost većeg od brojeva koji su pali na kockama. Dakle,

$$R(Z) = \{1, 2, 3, 4, 5, 6\} \qquad \text{i} \qquad Z(\omega) = \max\{i, j\}, \quad \omega = (i, j) \in \Omega.$$

Uočimo da je npr. $\mathbb{P}(Z = 2) = 1/12$. □

Primjer 3.4. U mjestu s 2000 obitelji ispitano je koliko koja obitelj posjeduje mobitela. Dobiveni su sljedeći rezultati (prikazani u Tablici 3.1):

broj mobitela	frekvencija	relativna frekvencija
0	30	0.015
1	470	0.235
2	850	0.425
3	490	0.245
4	160	0.08

Tablica 3.1: Frekvencije obitelji s različitim brojem mobitela

Nasumično biranje jedne od 2000 obitelji je slučajan pokus (vjerojatnost *a priori*). Neka je $U : \Omega \to \mathbb{R}$ slučajna varijabla koja označava broj mobitela u obitelji. Dakle,

$$R(U) = \{0, 1, 2, 3, 4\}.$$

Uočimo da je npr. $\mathbb{P}(U = 2) = 0.425$. □

Iz gornjih primjera vidimo da je svaka diskretna slučajna varijabla u potpunosti određena svojom slikom $R(X)$ i brojevima $p_i = \mathbb{P}(X = x_i)$ za $x_i \in R(X)$. Jasno je da je

$$\sum_{x_i \in R(X)} p_i = \sum_{x_i \in R(X)} \mathbb{P}(X = x_i) = \mathbb{P}(X \in R(X)) = \mathbb{P}(\Omega) = 1.$$

Svaku diskretnu slučajnu varijablu zapisujemo tablično s

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

i tu tablicu nazivamo **raspodjela** (ili distribucija) od X i pišemo

$$X \sim \begin{pmatrix} x_1 & x_2 & \cdots \\ p_1 & p_2 & \cdots \end{pmatrix}.$$

U gornjim primjerima imamo

$$\text{Primjer 3.1: } X \sim \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \end{pmatrix}$$

$$\text{Primjer 3.2: } Y \sim \begin{pmatrix} 0 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

$$\text{Primjer 3.3: } Z \sim \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ \frac{1}{36} & \frac{3}{36} & \frac{5}{36} & \frac{7}{36} & \frac{9}{36} & \frac{11}{36} \end{pmatrix}$$

$$\text{Primjer 3.4: } U \sim \begin{pmatrix} 0 & 1 & 2 & 3 & 4 \\ 0.015 & 0.235 & 0.425 & 0.245 & 0.08 \end{pmatrix}.$$

Uočimo da pomoću slučajnih varijabli možemo opisati sve događaje vezane za slučajni pokus. Na primjer, u kontekstu Primjera 3.1, kolika je vjerojatnost da je prilikom bacanja jedne simetriče kocke pao broj manji od 3? Imamo

$$\mathbb{P}(X < 3) = \mathbb{P}(X = 1) + \mathbb{P}(X = 2) = p_1 + p_2 = \frac{2}{6} = \frac{1}{3}.$$

U kontekstu Primjera 3.2, kolika je vjerojatnost da je prilikom bacanja jedne simetrične kocke pao neparan broj? Imamo

$$\mathbb{P}(Y = 0) = p_1 = \frac{1}{2}.$$

U kontekstu Primjera 3.3, kolika je vjerojatnost da je prilikom bacanja dvije kocke veći od brojeva koji je pao veći ili jednak 4? Imamo

$$\mathbb{P}(Z \geqslant 4) = \mathbb{P}(Z = 4) + \mathbb{P}(Z = 5) + \mathbb{P}(Z = 6) = p_4 + p_5 + p_6 = \frac{27}{36}.$$

U kontekstu Primjera 3.4, kolika je vjerojatnost da slučajno odabrana obitelj posjeduje barem dva mobitela? Imamo

$$\mathbb{P}(U \geqslant 2) = \mathbb{P}(U = 2) + \mathbb{P}(U = 3) + \mathbb{P}(U = 4) = p_2 + p_3 + p_4 = 0.75.$$

Kao što smo rekli, diskretna slučajna varijabla je u potpunosti zadana i određena svojom slikom i raspodjelom. Alternativna karakterizacija diskretne slučajne varijable može se dati pomoću funkcija vjerojatnosti i raspodjele. Neka je

$$X \sim \begin{pmatrix} x_1 & x_2 & \cdots \\ p_1 & p_2 & \cdots \end{pmatrix}.$$

Funkcija vjerojatnosti od X je funkcija $f : \mathbb{R} \to \mathbb{R}$ dana s

$$f(x) = \begin{cases} p_i, & x = x_i \\ 0, & \text{inače.} \end{cases}$$

Funkcija raspodjele od X je funkcija $F : \mathbb{R} \to \mathbb{R}$ dana s $F(x) = \mathbb{P}(X \leqslant x)$. Vrijedi sljedeće:

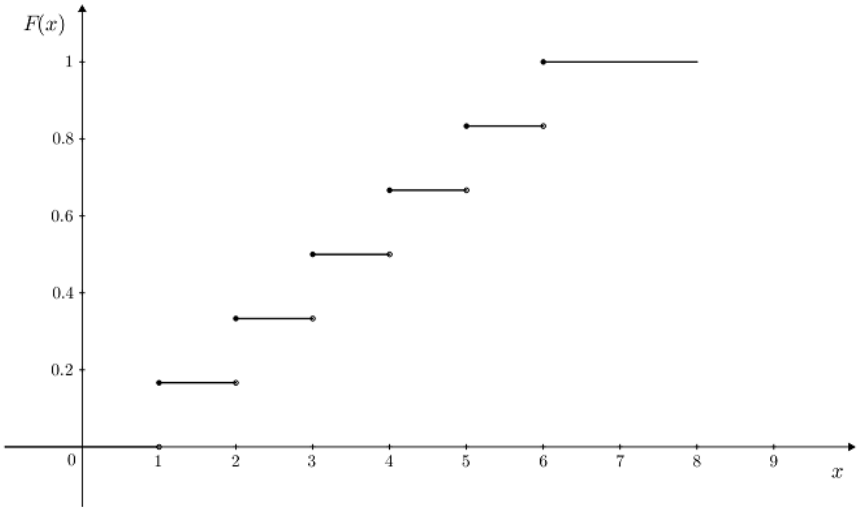
(i) $\lim_{x \to -\infty} F(x) = 0$ i $\lim_{x \to \infty} F(x) = 1$

(ii) $F(x)$ je neopadajuća funkcija

(iii) $F(x) = \sum_{x_i \in R(X),\, x_i \leqslant x} f(x) = \sum_{x_i \in R(X),\, x_i \leqslant x} p_i$

(iv) $f(x_i) = p_i = F(x_i) - F(x_{i-1})$.

Dakle, izravno iz funkcije raspodjele možemo iščitati $R(X)$ i sve p_i. Preciznije, funkcija $F(x)$ ima skokove u točkama x_i za visine p_i, lijevo od najmanjeg x_i-a (ako postoji) jednaka je 0 i desno od najvećeg x_i-a (ako postoji) jednaka je 1. U Primjeru 3.1 bacanja simetrične kocke funkcija raspodjele ima graf kao na Slici 3.1.



Slika 3.1: Funkcija raspodjele za simetričnu kocku iz Primjera 3.1

Slučajna varijabla je nedeterministički objekt. Međutim, svakoj slučajnoj varijabli možemo pridružiti neke deterministčke karakteristike koje ju opisuju, pomoću kojih slučajnu varijablu bolje razumijemo i s kojima je lakše raditi nego sa samom slučajnom varijablom. Mi ćemo obraditi dvije, odnosno tri takve veličine:

(i) očekivanje (srednja vrijednost)

(ii) varijancu (kvadratno raspršenje slučajne varijable oko njenog očekivanja)

(iii) standardnu devijaciju (linearno raspršenje slučajne varijable oko njenog očekivanja).

Neka je

$$X \sim \begin{pmatrix} x_1 & x_2 & \cdots \\ p_1 & p_2 & \cdots \end{pmatrix}$$

diskretna slučajna varijabla. **Očekivanje** od X, u oznaci $\mathbb{E}(X)$, broj je (ako donja suma postoji) definiran s

$$\mathbb{E}(X) = \sum_{x_i \in R(X)} x_i p_i.$$

Svojstva očekivanja jesu:

(i) $\mathbb{E}(\lambda X) = \lambda \mathbb{E}(X), \quad \lambda \in \mathbb{R}$,

(ii) $\mathbb{E}(X+Y) = \mathbb{E}(X)+\mathbb{E}(Y)$ (gdje su X i Y dvije diskretne slučajne varijable definirane na istom vjerojatnosnom prostoru koje imaju očekivanje).

Varijanca od X, u oznaci $\mathrm{Var}(X)$, broj je (ako donja suma postoji) definiran s

$$\mathrm{Var}(X) = \mathbb{E}((X - \mathbb{E}(X))^2) = \sum_{x_i \in R(X)} (x_i - \mathbb{E}(X))^2 p_i.$$

Uočimo da vrijedi

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \sum_{x_i \in R(X)} x_i^2 p_i - \mathbb{E}(X)^2.$$

Ovaj broj mjeri kvadratnu prosječnu udaljenost (raspršenje) slučajne varijble X (vrijednosti od X) od $\mathbb{E}(X)$. Svojstva varijance jesu:

(i) $\mathrm{Var}(\lambda X) = \lambda^2 \mathrm{Var}(X), \qquad \lambda \in \mathbb{R}$

(ii) $\mathrm{Var}(X + \lambda) = \mathrm{Var}(X), \qquad \lambda \in \mathbb{R}$.

Napomenimo sada da $\mathbb{E}(X)$ je broj koji je, u smislu mjerenja udaljenosti varijancom, najbliži slučajnoj varijabli X, tj. minimizira očekivano kvadratno odstupanje od X. Zaista, za $a \in \mathbb{R}$ imamo

$$\begin{aligned} \mathbb{E}[(X-a)^2] &= \mathbb{E}(((X - \mathbb{E}(X)) + (\mathbb{E}(X) - a))^2) \\ &= \mathbb{E}((X - \mathbb{E}(X))^2) + 2(\mathbb{E}(X) - a)\mathbb{E}(X - \mathbb{E}(X)) + (\mathbb{E}(X) - a)^2 \\ &= \mathrm{Var}(X) + (\mathbb{E}(X) - a)^2. \end{aligned}$$

Dakle, izborom $a = \mathbb{E}(X)$ vidimo da funkcija $a \mapsto \mathbb{E}((X-a)^2)$ poprima minimum.

Standardna devijacija od X, u oznaci $\sigma(X)$, broj je (ako X ima varijancu) dan formulom

$$\sigma(X) = \sqrt{\mathrm{Var}(X)}.$$

Standardna devijacija mjeri linearnu prosječnu udaljenost (raspršenje) od $\mathbb{E}(X)$. Mali $\sigma(X)$ znači da je većina vrijednosti od X grupirana oko $\mathbb{E}(X)$, a veliki $\sigma(X)$ znači da su vrijednosti od X razvučene preko većeg raspona.

Primjer 3.5. U Primjeru 3.1 imamo

$$\begin{aligned}
\mathbb{E}(X) &= 1\cdot\frac{1}{6} + 2\cdot\frac{1}{6} + 3\cdot\frac{1}{6} + 4\cdot\frac{1}{6} + 5\cdot\frac{1}{6} + 6\cdot\frac{1}{6} = 3.5\\
\mathrm{Var}(X) &= (1-3.5)^2\cdot\frac{1}{6} + (2-3.5)^2\cdot\frac{1}{6} + (3-3.5)^2\cdot\frac{1}{6} + (4-3.5)^2\cdot\frac{1}{6}\\
&\quad + (5-3.5)^2\cdot\frac{1}{6} + (6-3.5)^2\cdot\frac{1}{6}\\
&= 2.91\\
\sigma(X) &= \sqrt{2.91}.
\end{aligned}$$

□

Primjer 3.6. U kontekstu Primjera 3.3 definirajmo

$$Z(\omega) = \begin{cases} -1, & \max\{i,j\} \leqslant 4\\ 1, & \max\{i,j\} > 4, \end{cases} \qquad \omega = (i,j) \in \Omega.$$

Dakle,

$$Z \sim \begin{pmatrix} -1 & 1\\ \frac{16}{36} & \frac{20}{36} \end{pmatrix}.$$

Slučajnu varijablu Z možemo interpretirati kao igru bacanja dvije simetrične kocke u kojoj dobivamo kunu ako je veći od brojeva na obje kocke veći od četiri, a inače gubimo kunu. Imamo

$$\begin{aligned}
\mathbb{E}(Z) &= -1\cdot\frac{16}{36} + 1\cdot\frac{20}{36} = \frac{1}{9}\\
\mathrm{Var}(Z) &= (-1-1/9)^2\cdot\frac{16}{36} + (1-1/9)^2\cdot\frac{20}{36} = \frac{80}{81}\\
\sigma(Z) &= \frac{\sqrt{80}}{9}.
\end{aligned}$$

Ako se nudi posao plaćen 10 kn/h, isplati li se raditi ili kladiti ako pretpostavimo da imamo jedno klađenje u minuti? Nakon jednog klađenja (jedne minute) očekivani dobitak je 1/9 kn. Dakle, nakon jednog sata (60) klađenja očekivani dobitak je $60/9$ kn ≈ 6.66 kn. □

Primjer 3.7. U Primjeru 3.4 imamo

$$\begin{aligned}\mathbb{E}(U) &= 0 \cdot 0.015 + 1 \cdot 0.235 + 2 \cdot 0.425 + 3 \cdot 0.245 + 4 \cdot 0.08 = 2.14\\ \mathrm{Var}(U) &= (0-2.14)^2 \cdot 0.015 + (1-2.14)^2 \cdot 0.235 + (2-2.14)^2 \cdot 0.425\\ &\quad + (3-2.14)^2 \cdot 0.245 + (4-2.14)^2 \cdot 0.08\\ &= 0.84\\ \sigma(U) &= \sqrt{0.84}.\end{aligned}$$

Interpretacija broja $\mathbb{E}(U) = 2.14$ je da prosječan broj mobitela u slučajno odabranoj obitelji je 2.14. □

Primjer 3.8. Bacamo simetričan novčić. Dakle, $\Omega = \{P, G\}$, $\mathcal{F} = \mathcal{P}(\Omega)$ i $\mathbb{P}(\{\omega\}) = 1/2$, $\omega \in \Omega$. Definirajmo slučajnu varijablu $V : \Omega \longrightarrow \mathbb{R}$ s $V(P) = -1$ i $V(G) = 1$. Imamo

$$V \sim \begin{pmatrix} -1 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$

Slučajnu varijablu V možemo interpretirati kao igru u kojoj gubimo kunu ako padne pismo i dobivamo kunu ako padne glava. Očito je $\mathbb{E}(V) = 0$. Dakle, nakon npr. 100 bacanja očekujemo približno 50 pobjeda i 50 poraza. Stoga možemo reći da je igra "poštena". □

Primjer 3.9. Koliki mora biti dobitak na lotu 7/39 da bi igra bila "poštena"? Uplata jedne kombinacije iznosi 3 kn. Imamo,

$$\Omega \ldots \text{sedmeročlani podskupovi skupa } \{1, \ldots, 39\}, \quad \mathcal{F} = \mathcal{P}(\Omega) \quad \text{i}$$

$$\mathbb{P}(\{\omega\}) = \frac{1}{\binom{39}{7}}, \qquad \omega \in \Omega.$$

Označimo s λ "pošteni" dobitak te s ω_0 dobitnu kombinaciju. Definirajno slučajnu varijablu $W : \Omega \to \mathbb{R}$ s

$$W(\omega) = \begin{cases} \lambda - 3, & \omega = \omega_0 \\ -3, & \omega \neq \omega_0. \end{cases}$$

Dakle,

$$W \sim \begin{pmatrix} -3 & \lambda - 3 \\ \frac{\binom{39}{7}-1}{\binom{39}{7}} & \frac{1}{\binom{39}{7}} \end{pmatrix}.$$

Sada imamo

$$\mathbb{E}(W) = \frac{\lambda - 3}{\binom{39}{7}} - 3\frac{\binom{39}{7}}{\binom{39}{7}} = \frac{\lambda}{\binom{39}{7}} - 3.$$

Igra je "poštena" ako je $\mathbb{E}(W) = 0$. Dakle,

$$\lambda = 3\binom{39}{7} = 46142811\,\text{kn}. \qquad \square$$

Napomenimo da smo prilikom uvođenja pojmova očekivanja i varijance koristili sljedeće činjenice:

(i) ako je $f : \mathbb{R} \to \mathbb{R}$ neka funkcija i

$$X \sim \begin{pmatrix} x_1 & x_2 & \cdots \\ p_1 & p_2 & \cdots \end{pmatrix},$$

onda je $f(X)$ diskretna slučajna varijabla definirana na istom vjerojatnosnom prostoru kao i X te vrijedi

$$f(X) \sim \begin{pmatrix} f(x_1) & f(x_2) & \cdots \\ p_1 & p_2 & \cdots \end{pmatrix}.$$

(ii) ako su X i Y dvije diskretne slučajne varijable definirane na istom vjerojatnosnom prostoru, sa slikama $R(X) = \{x_1, x_2, \dots\}$ i $R(Y) = \{y_1, y_2, \dots\}$, onda je $X + Y$ diskretna slučajna varijabla definira na istom vjerojatnosnom prostoru kao X i Y te vrijedi

$$R(X+Y) \subseteq \{x_i + y_j : x_i \in R(X),\ y_j \in R(Y)\}$$
$$\mathbb{P}(X+Y=z_k) = \sum_{\substack{x_i \in R(X),\, y_j \in R(Y) \\ x_i + y_j = z_k}} \mathbb{P}(X = x_i, Y = y_j).$$

Vrijedi sljedeće (ako donje sume postoje):

$$\mathbb{E}(f(X)) = \sum_{x_i \in R(X)} f(x_i) p_i$$

$$\text{Var}(f(X)) = \sum_{x_i \in R(X)} (f(x_i) - \mathbb{E}(f(X)))^2 p_i = \sum_{x_i \in R(X)} f(x_i)^2 p_i - \mathbb{E}(f(X))^2.$$

Primjer 3.10. Neka je

$$X \sim \begin{pmatrix} -1 & 0 & 1 \\ 0.2 & 0.3 & 0.5 \end{pmatrix}.$$

Odredimo funkciju vjerojatnosti slučajnih varijabli $2X + 1$ i X^2. Imamo

$$2X+1 \sim \begin{pmatrix} -1 & 1 & 3 \\ 0.2 & 0.3 & 0.5 \end{pmatrix} \quad \text{i} \quad X^2 \sim \begin{pmatrix} 1 & 0 & 1 \\ 0.2 & 0.3 & 0.5 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 0.3 & 0.7 \end{pmatrix}. \qquad \square$$

Pogledajmo sada primjere nekih standardnih diskretnih slučajnih varijabli.

3.1.1 Diskretna uniformna slučajna varijabla

Diskretna slučajna varijabla X je **uniformna** na diskretnom n-članom skupu $\{x_1, \ldots, x_n\} \subseteq \mathbb{R}$ ako je njena raspodjela oblika

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

Očito je gornjom tablicom zadana vjerojatnosna raspodjela. Uniformna slučajna varijabla koristi se kod pokusa s konačno mnogo jednako vjerojatnih ishoda. Očito je

$$\mathbb{E}(X) = \frac{x_1 + \cdots + x_n}{n} \quad \text{i} \quad \mathrm{Var}(X) = \frac{x_1^2 + \cdots + x_n^2}{n} - \mathbb{E}(X)^2.$$

Primjer 3.11. Bacamo simetričnu kocku. Neka je

$$X(\omega) = \begin{cases} 1, & \omega \in \{1, 2\} \\ 2, & \omega \in \{3, 4\} \\ 3, & \omega \in \{5, 6\}. \end{cases}$$

Tada imamo

$$X \sim \begin{pmatrix} 1 & 2 & 3 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix}.$$ □

3.1.2 Bernoullijeva slučajna varijabla

Diskretna slučajna varijabla X je **Bernoullijeva** s parametrom $0 \leqslant p \leqslant 1$ ako je njena raspodjela oblika

$$X \sim \begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}.$$

Očito je gornjom tablicom zadana vjerojatnosna raspodjela. Ova slučajna varijabla se može interpretirati kao ishod pokusa koji može rezultirati samo uspjehom ili neuspjehom (vjerojatnost uspjeha je p). Očito je

$$\mathbb{E}(X) = p \quad \text{i} \quad \mathrm{Var}(X) = p - p^2 = p(1-p).$$

Primjer 3.12. Bacamo par simetričnih kocki. Neka je

$$X(\omega) = \begin{cases} 0, & i + j > 6 \\ 1, & i + j \leqslant 6, \end{cases} \qquad \omega = (i, j) \in \Omega.$$

Tada imamo

$$X \sim \begin{pmatrix} 0 & 1 \\ \frac{7}{12} & \frac{5}{12} \end{pmatrix}.$$

Dakle, X je Bernoullijeva slučajna varijabla s parametrom $p = 5/12$. □

3.1.3 Binomna slučajna varijabla

Diskretna slučajna varijabla X je **binomna** s parametrima $0 \leqslant p \leqslant 1$ i $n \in \mathbb{N}$, u oznaci $X \sim B(n,p)$, ako je njena raspodjela oblika

$$X \sim \begin{pmatrix} 0 & 1 & \cdots & n \\ p_0 & p_1 & \cdots & p_n \end{pmatrix},$$

gdje je

$$p_i = \binom{n}{i} p^i (1-p)^{n-i}, \quad i = 0, 1, \ldots, n.$$

Pokažite da je gornjom tablicom zadana vjerojatnosna raspodjela. Binomna slučajna varijabla predstavlja broj uspjeha $i \in \{0, 1, \ldots, n\}$ u n nezavisnih ponavljanja pokusa kod kojeg kao ishod imamo uspjeh s vjerojatnošću p i neuspjeh s vjerojatnošću $1-p$. Vjerojatnost bilo kojeg fiksnog niza od i uspjeha i $n-i$ neuspjeha u n nezavisnih ponavljanja pokusa je $p^i(1-p)^{n-i}$, a broj takvih nizova je $\binom{n}{i}$. Nije teško izračunati (što ćemo učiniti u Poglavlju 4) da je

$$\mathbb{E}(X) = np \qquad \text{i} \qquad \mathrm{Var}(X) = np(1-p).$$

Uočimo također da je $X \sim B(1,p)$ u biti Bernoullijeva slučajna varijabla s parametrom p. Pravu vezu između binomne i Bernoullijeve slučajne varijable ćemo također komentirati kasnije.

Primjer 3.13. Marko i Ivan igraju sljedeću igru. Ivan izabere broj između $\{1, \ldots, 6\}$ te nakon toga baca igraću kocku tri puta. Ako prilikom bacanja ne dobije prethodno izabran broj, onda Marku isplaćuje 100 kn. Međutim, ako se prethodno izabrani broj pojavi jednom na kocki, onda Marko isplaćuje Ivanu 100 kn, ali ako se pojavi dva puta 200 kn i ako se pojavi tri puta 300 kn. Koga ova igra favorizira? Broj pojavljivanja izabranog broja opisuje binomna slučajna varijabla

$$X \sim B\left(3, \frac{1}{6}\right).$$

Sada, očekivani dobitak za Marka je

$$100p_0 - 100p_1 - 200p_2 - 300p_3 = 7.37 \text{ kn}.$$

Dakle, igra favorizira Marka. □

3.1.4 Geometrijska slučajna varijabla

Neka je u jednom pokusu vjerojatnost uspjeha p. Pokus se ponavlja nezavisno dok se ne dogodi uspjeh. Geometrijska slučajna varijabla je broj ponavljanja pokusa do prvog uspjeha. Preciznije, diskretna slučajna varijabla X je **geometrijska** s parametrom $0 < p \leqslant 1$, u oznaci $X \sim G(p)$, ako je njena raspodjela oblika

$$X \sim \begin{pmatrix} 1 & 2 & 3 & \cdots \\ p_1 & p_2 & p_3 & \cdots \end{pmatrix},$$

gdje je

$$p_i = (1-p)^{i-1}p, \quad i \in \mathbb{N}.$$

Pokažite da je gornjom tablicom zadana vjerojatnosna raspodjela. Dakle, $(1-p)^{i-1}p$ znači da u prvih $i-1$ ponavljanja pokusa imamo neuspjehe, a u i-tom ponavljanju dogodi se uspjeh. Može se pokazati da je

$$\mathbb{E}(X) = \frac{1}{p} \qquad \text{i} \qquad \mathrm{Var}(X) = \frac{1-p}{p^2}$$

(vidi [9, str. 154]).

Primjer 3.14. Bacamo simetričnu kocku sve dok ne padne broj 6. Kolika je vjerojatnost da će se to dogoditi u petom bacanju? Ovo je primjer slučajnog pokusa čiji ishod opisujemo geometrijskom slučajnom varijablom $X \sim G(\frac{1}{6})$. Dakle, tražena vjerojatnost je

$$p_5 = \left(1 - \frac{1}{6}\right)^4 \frac{1}{6} = \frac{5^4}{6^5}.$$ □

Uočimo sljedeće svojstvo geometrijske raspodjele: za sve $i, j \geq 1$ vrijedi

$$\mathbb{P}(X \geq i + j | X > j) = \mathbb{P}(X \geq i).$$

Drugim riječima, gornja relacija znači da geometrijska raspodijela nema memoriju. Kako X označava broj (nezavisnih) ponavljanja pokusa do pojavljivanja prvog uspjeha, gornja relacija govori da ako je X uvjetovano na nepojavljivanje uspjeha u prvih j ponavljanja, raspodjela pojavljivanja uspjeha u preostalim ponavljanjima jednaka je kao i neuvjetovana raspodjela.

3.1.5 Poissonova slučajna varijabla

Diskretna slučajna varijabla X je **Poissonova** s parametrom $\lambda > 0$, u oznaci $X \sim \mathrm{Poi}(\lambda)$, ako je njena raspodjela oblika

$$X \sim \begin{pmatrix} 0 & 1 & 2 & 3 & \cdots \\ p_0 & p_1 & p_2 & p_3 & \cdots \end{pmatrix},$$

gdje je

$$p_i = e^{-\lambda}\frac{\lambda^i}{i!}, \qquad i \in \mathbb{N} \cup \{0\}.$$

Pokažite da je gornjom tablicom zadana vjerojatnosna raspodjela. Intuitivno, ova slučajna varijabla predstavlja broj uspjeha prilikom velikog broja nezavisnih ponavljanja pokusa koji mogu rezultirati uspjehom ili neuspjehom i vjerojatnost uspjeha je p_n tako da je

$$\lim_{n\to\infty} np_n = \lambda.$$

Na primjer, za velike n, za p_n možemo uzeti baš λ/n. Nadalje, vrijedi

$$\mathbb{E}(X) = \mathrm{Var}(X) = \lambda.$$

Imamo,

$$\mathbb{E}(X) = \sum_{i=0}^{\infty} ie^{-\lambda}\frac{\lambda^i}{i!} = \lambda e^{-\lambda}\sum_{i=1}^{\infty}\frac{\lambda^{i-1}}{(i-1)!} = \lambda e^{-\lambda}e^{\lambda} = \lambda.$$

Analogno,

$$\mathbb{E}[X(X-1)] = \sum_{i=0}^{\infty} i(i-1)e^{-\lambda}\frac{\lambda^i}{i!} = \lambda^2 e^{-\lambda}\sum_{i=2}^{\infty}\frac{\lambda^{i-2}}{(i-2)!} = \lambda^2,$$

što daje $\mathbb{E}(X^2) = \lambda^2 + \lambda$, tj. $\mathrm{Var}(X) = \lambda$.

Prisjetimo se, binomna slučajna varijabla predstavlja broj uspjeha prilikom n nezavisnih ponavljanja pokusa koji mogu rezultirati uspjehom ili neuspjehom. Dakle, na Poissonovu slučajnu varijablu možemo gledati kao na granični slučaj binomne kad $n \to \infty$. Zaista, neka je $p_n = \lambda/n$ i $X_n \sim B(n, p_n)$. Tada imamo

$$\mathbb{P}(X_n = i) = \binom{n}{i}p_n^i(1-p_n)^{n-i} = \frac{\lambda^i}{i!}\left(1-\frac{\lambda}{n}\right)^n\frac{n!}{n^i(n-i)!}\left(1-\frac{\lambda}{n}\right)^{-i}.$$

Sada je

$$\lim_{n\to\infty}\mathbb{P}(X_n = i) = \frac{\lambda^i}{i!}e^{-\lambda}.$$

Iz gornjeg razmatranja zaključujemo da binomnu slučajnu varijablu možemo aproksimirati Poissonovom i za tu aproksimaciju nije potrebno znati n i p_n već samo λ, a to je očekivani broj uspjeha.

Primjer 3.15. U nekoj telefonskoj centrali stižu pozivi po Poissonovoj raspodjeli s očekivanjem od 10 poziva po minuti. Kolika je vjerojatnost da će sljedeću minutu stići više od 20 poziva? Broj poziva u minuti opisan je Poissonovom slučajnom varijablom $X \sim \text{Poi}(10)$. Dakle, odgovor je

$$\mathbb{P}(X > 20) = \sum_{i=21}^{\infty} \mathbb{P}(X = i) = \sum_{i=21}^{\infty} p_i = 1 - \sum_{i=0}^{20} p_i. \qquad \square$$

Primjer 3.16. Igra loto 7/39 izvlači se dva puta tjedno. Odlučili smo igrati narednih 50 tjedana (100 igara). U svakoj igri uplaćujemo listić s 10 kombinacija. Kolika je vjerojatnost da ćemo osvojiti Jackpot dva ili više puta? Uočimo da je vjerojatnost dobitka u jednoj igri $p = 10/\binom{39}{7}$. Slučaj je opisan binomnom slučajnom varijablom $X_b \sim B(100, p)$. Stavimo

$A \ldots$ "dobili smo Jackpot dva ili više puta"
$B \ldots$ "nismo dobili Jackpot"
$C \ldots$ "dobili smo Jackpot samo jednom".

Sada imamo

$$\mathbb{P}(A) = 1 - \mathbb{P}(B) - \mathbb{P}(C) = 1 - p_0^b - p_1^b.$$

Međutim, u ovom problemu nije lako izračunati vjerojatnosti p_i^b jer baratamo s jako malim brojevima. Ideja je aproksimirati X_b Poissonovom slučajnom varijablom. Stavimo $\lambda = 100p$ i neka je X_p Poissonova slučajna varijabla $X_p \sim \text{Poi}(\lambda)$. Sada prema prethodnoj diskusiji imamo

$$\mathbb{P}(A) = 1 - \mathbb{P}(B) - \mathbb{P}(C) = 1 - p_0^b - p_1^b \approx 1 - p_0^p - p_1^p = 1 - e^{-\lambda} - \lambda e^{-\lambda}. \quad \square$$

3.1.6 Hipergeometrijska slučajna varijabla

Diskretna slučajna varijabla X je **hipergeometrijska** s parametrima $n \geqslant 1, 1 \leqslant m \leqslant n$ i $1 \leqslant k \leqslant n$, u oznaci $X \sim \text{Hip}(n, m, k)$, ako vrijedi

$$X \sim \begin{pmatrix} 0 & 1 & \cdots & \min\{k, m\} \\ p_0 & p_1 & \cdots & p_{\min\{k,m\}} \end{pmatrix},$$

gdje je

$$p_i = \frac{\binom{m}{i}\binom{n-m}{k-i}}{\binom{n}{k}}, \quad i = 0, 1, \ldots, \min\{k, m\}.$$

Pokažite da je gornjom tablicom zadana vjerojatnosna raspodjela (vidi [9, str. 95]). Ova slučajna varijabla predstavlja broj izvučenih članova prve vrste

prilikom izvlačenja k-članog uzorka iz n-člane populacije koja je sačinjena od m članova prve vrste i $n-m$ članova druge vrste. Može se pokazati da vrijedi

$$\mathbb{E}(X) = \frac{km}{n} \qquad \text{i} \qquad \mathrm{Var}(X) = \frac{km}{n}\left(1 - \frac{m}{n}\right)\frac{n-k}{n-1}$$

(vidi [9, str. 125 i 142]).

Primjer 3.17. Kolika je vjerojatnost da ćemo imati točno 5 pogodaka u igri loto 7/39? Slučaj je opisan hipergeometrijskom slučajnom varijablom

$$X \sim \mathrm{Hip}(39, 7, 7).$$

Dakle, odgovor je

$$p_5 = \frac{\binom{7}{5}\binom{39-7}{7-5}}{\binom{39}{7}}.$$ □

3.2 Neprekidne slučajne varijable

Diskretne slučajne varijable u potpunosti su određene svojom slikom i funkcijom vjerojatnosti (tj. vjerojatnostima $p_i = \mathbb{P}(X = x_i)$, $x_i \in R(X)$). Neprekidne slučajne varijable kao sliku imaju neprebrojiv skup u $\mathbb{R}$ i one uglavnom modeliraju probleme koji su po svojoj prirodi neprekidni: masa ili volumen nečega, količina proteklog vremena itd. Vidjet ćemo da u tom slučaju vjerojatnosti $\mathbb{P}(X = x)$, $x \in R(x)$, neće biti od prevelike koristi, tj. uvijek će biti nula. Na primjer, postavimo si pitanje kolika je vjerojatnost da će tramvaj broj 17 doći na stanicu *Trg bana Josipa Jelačića* točno u 17 sati, 35 minuta i 13 sekundi. Naravno, odgovor je nula. Ima smisla računati i interpretirati samo vjerojatnosti događaja reprezentiranih intervalima realnih brojeva. Međutim, napomenimo da imamo dva tipa neprekidnih slučajnih varijabli: prave neprekidne (u daljnjem tekstu samo neprekidne) i singularne neprekidne. Analiza ovih drugih je nešto kompleksnija i njima se nećemo baviti. Napomenimo samo da se svaka neprekidna slučajna varijabla može (na jedinstven način) zapisati kao suma jedna prave neprekidne i jedne singularne neprekidne slučajne varijable, ili još općenitije svaka slučajna varijabla se može (na jedinstven način) zapisati kao suma jedne diskretne, jedne prave neprekidne i jedne singularne neprekidne slučajne varijable (vidi [9, str. 264]).

Slučajna varijabla $X : \Omega \to \mathbb{R}$ je **neprekidna** ako postoji ("izmjeriva") funkcija $f : \mathbb{R} \to \mathbb{R}$ za koju vrijedi:

(i) $f(x) \geqslant 0, \quad x \in \mathbb{R},$

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

(iii) $\mathbb{P}(X \leqslant a) = \displaystyle\int_{-\infty}^{a} f(x)\,dx, \qquad a \in \mathbb{R}.$

Funkciju $f(x)$ zovemo **funkcija gustoće** od X. Uočimo da iz svojstva (iii) slijedi da za sve $a, b \in \mathbb{R}$, $a \leqslant b$, vrijedi

$$\mathbb{P}(a < X \leqslant b) = \int_{a}^{b} f(x)\,dx.$$

Također, može se pokazati da je $\mathbb{P}(X = x) = 0$ za sve $x \in \mathbb{R}$. Pokažite prethodnu tvrdnju u slučaju kada je funkcija gustoće od X neprekidna funkcija. Nadalje, iz gornjeg svojstva zaključujemo

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

Analogno kao i u diskrentnom slučaju uvodimo pojam funkcije raspodjele. **Funkcija raspodjele** od X je funkcija $F : \mathbb{R} \to \mathbb{R}$ dana s

$$F(x) = \mathbb{P}(X \leqslant x).$$

Vrijedi sljedeće:

(i) $\lim_{x\to-\infty} F(x) = 0$ i $\lim_{x\to\infty} F(x) = 1$,

(ii) $F(x)$ je neopadajuća,

(iii) $F(x) = \int_{-\infty}^{x} f(t)\,dt, \qquad x \in \mathbb{R}$,

(iv) $\mathbb{P}(a < X \leqslant b) = F(b) - F(a)$,

(v) ako je $f(x)$ po dijelovima neprekidna, onda $F'(x) = f(x)$ osim možda u točkama prekida od $f(x)$.

Uočimo da kao i u diskretnom slučaju (ako je $f(x)$ po dijelovima neprekidna) $F(x)$ daje punu informaciju o X, tj.

$$f(x) = F'(x) \text{ (uz komentar kao i u (v))} \quad \text{i} \quad R(X) = \{x \in \mathbb{R} : f(x) \neq 0\}.$$

Uvodimo i determinističke karakteristike neprekidnih slučajnih varijabli. Za neprekidnu slučajnu varijablu X definiramo **očekivanje** od X (ako donji integral postoji) s

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

Vrijedi sljedeće:

(i) $\mathbb{E}(\lambda X) = \lambda \mathbb{E}(X), \quad \lambda \in \mathbb{R}$,

(ii) $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$.

Varijancu od X (ako donji integral postoji) definiramo s

$$\mathrm{Var}(X) = \mathbb{E}((X - \mathbb{E}(X))^2) = \int_{-\infty}^{\infty} (x - \mathbb{E}(X))^2 f(x) dx.$$

Također, lagano se pokaže

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \int_{-\infty}^{\infty} x^2 f(x) dx - \mathbb{E}(X)^2.$$

Vrijedi sljedeće:

(i) $\mathrm{Var}(\lambda X) = \lambda^2 \mathrm{Var}(X), \quad \lambda \in \mathbb{R},$

(ii) $\mathrm{Var}(X + \lambda) = \mathrm{Var}(X).$

Analogno kao i u diskretnom slučaju (uz isti dokaz), $\mathbb{E}(X)$ je broj koji je, u smislu mjerenja udaljenosti varijancom, najbliži slučajnoj varijabli X, tj. minimizira očekivano kvadratno odstupanje od X.

Standardnu devijaciju od X (ako X ima varijancu) definiramo s

$$\sigma(X) = \sqrt{\mathrm{Var}(X)}.$$

Napomenimo sljedeće:

(i) neka je sada X neprekidna slučajna varijabla sa slikom $R(X)$ i funkcijom gustoće $f(x)$ te neka je $g : \mathbb{R} \longrightarrow \mathbb{R}$ neka funkcija (u definicijama očekivanja i varijance od X smo imali $g(x) = \lambda x$, $\lambda \in \mathbb{R}$, i $g(x) = (x - \mathbb{E}(X))^2$). Tada je $g(X)$ slučajna varijabla definirana na istom vjerojatnosnom prostoru kao i X, ima sliku $g(R(X))$ i vrijedi (ako donji integrali postoje)

$$\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x) f(x) dx,$$

$$\begin{aligned} \mathrm{Var}(g(X)) &= \int_{-\infty}^{\infty} (g(x) - \mathbb{E}(g(X))^2 f(x) dx \\ &= \int_{-\infty}^{\infty} g^2(x) f(x) dx - \mathbb{E}(g(X))^2. \end{aligned}$$

Napomenimo ovdje dvije bitne stvari. Prvo, u duhu kometara s početka ovog poglavlja, budući da je X nerekidna slučajna varijabla za funkciju $g(x)$ ne možemo uzeti baš bilo što. Funkcija $g(x)$ mora biti dovoljno "lijepa" (preciznije, "izmjeriva", primjerice svaka po dijelovima neprekidna funkcija je uvijek "izmjeriva") kako bi osigurali da je $g(X)$ zaista slučajna varijabla. Drugo, u slučaju da je $g(X)$ slučajna varijabla, ona ne mora biti nužno neprekidna (npr. uzmimo $g(x) = 0$). Dovoljan uvjet je da je $g : \mathbb{R} \to \mathbb{R}$ ("izmjeriva") bijekcija. Specijalno, ako je $g(x) = ax + b, a \neq 0$ i $b \in \mathbb{R}$, onda je funkcija gustoće od $g(X)$ dana s $f_g(x) = \frac{1}{|a|} f(\frac{x-b}{a})$, a funkcija raspodjele

$$F_g(x) = \begin{cases} F(\frac{x-b}{a}), & a > 0 \\ 1 - F(\frac{x-b}{a}), & a < 0. \end{cases}$$

Očito, očekivanje, varijanca i standardna devijacija od $g(X)$ su onda dani, redom, sa $\mathbb{E}(g(X)) = a\mathbb{E}(X) + b$, $\mathrm{Var}(g(X)) = a^2 \mathrm{Var}(X)$ i $\sigma(g(X)) = |a|\sigma(X)$.

(ii) ako su X i Y dvije neprekidne slučajne varijable definirane na istom vejrojatnosnom prostoru, onda je i $X+Y$ slučajna varijabla definirana na istom vjerojatnosnom prostoru kao X i Y te vrijedi

$$R(X+Y) \subseteq \{x+y : x \in R(X),\ y \in R(Y)\}.$$

Međutim, uočimo da $X+Y$ nije nužno neprekidna (npr. ako je $Y = -X$, onda je $X+Y=0$).

(iii) rekli smo da ako su X i Y iste vrste (obje diskretne ili obje neprekidne), onda vrijedi $\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$. Međutim, ta relacija vrijedi i ako nisu iste vrste.

Primjer 3.18. Neka je $[a,b]$ neki segment u $\mathbb{R}$. Promotrimo slučajni pokus biranja jedne točke iz $[a,b]$. Neka je X slučajna varijabla čija je vrijednost izabrana točka. Odredimo funkcije gustoće i raspodjele od X te izračunajmo pripadno očekivaje, varijancu i standardnu devijaciju. Očito je $\Omega = [a,b]$ i $X : \Omega \to [a,b]$ je dana s $X(\omega) = \omega$. Nadalje, pretpostavljamo da su sve točke "jednako vjerojatne", tj. posrijedi je geometrijska vjerojatnost

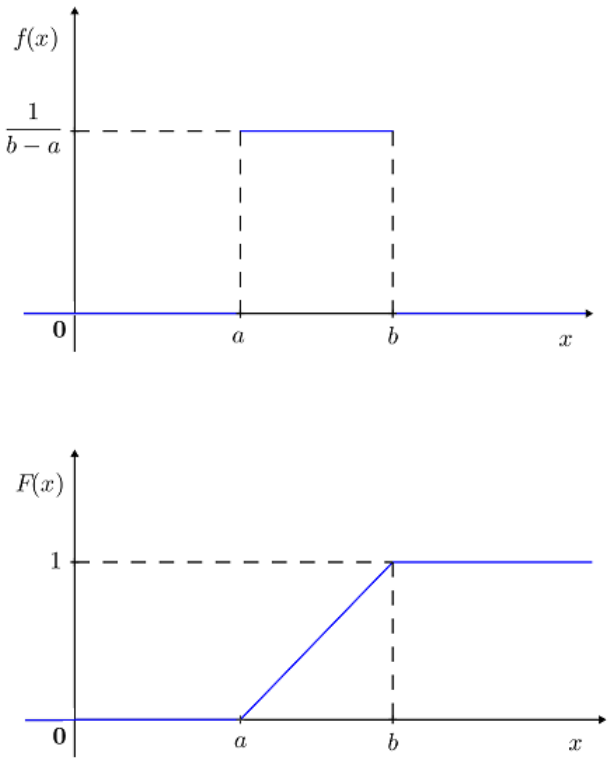
$$\mathbb{P}([c,d]) = \frac{d-c}{b-a}, \quad [c,d] \subseteq [a,b].$$

Dakle,

$$F(x) = \mathbb{P}(X \leqslant x) = \begin{cases} 0, & x < a \\ \frac{x-a}{b-a}, & a \leqslant x < b \\ 1, & x \geqslant b \end{cases}$$

$$f(x) = F'(x) = \begin{cases} 0, & x < a \\ \frac{1}{b-a}, & a < x < b \\ 0, & x > b \end{cases}$$

Uočimo da $F(x)$ nije derivabilna u $x = a$ i $x = b$.



Slika 3.2: Funkcija gustoće i funkcija raspodjele slučajne varijable X iz Primjera 3.18

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x)\,dx = \int_a^b \frac{x}{b-a}\,dx = \frac{a+b}{2}$$
$$\mathrm{Var}(X) = \int_{-\infty}^{\infty} x^2 f(x)\,dx - \mathbb{E}(X)^2 = \int_a^b \frac{x^2}{b-a}\,dx - \left(\frac{a+b}{2}\right)^2 = \frac{(b-a)^2}{12}$$
$$\sigma(X) = \sqrt{\mathrm{Var}(X)} = \frac{b-a}{2\sqrt{3}}.$$ □

Slika 3.2 prikazuje funkciju gustoće i funkciju raspodjele od X. Uočimo da funkcija gustoće ne mora biti neprekidna već samo funkcija raspodjele (zadana je integralom s nezavisnom varijablom u gornoj granici integracije). Funkcija raspodjele diskretnih slučajnih varijabli je stepenastog oblika i očito ima prekid u svakoj točki slike (zadana je sumom). Dakle, slučajne varijable nazivamo diskretnim ili neprekidnim zbog svojstva njihove funkcije raspodjele (skokovita ili neprekidna).

Primjer 3.19. Biramo točku na slučajan način unutar kruga radijusa $r > 0$. Neka je X slučajna varijabla čija je vrijednost udaljenost odabrane točke od središta kruga. Odredimo funkcije gustoće i raspodjele od X te izračunajmo pripadno očekivaje, varijancu i standardnu devijaciju. Očito je $\Omega = \{\omega = (x, y) \in \mathbb{R}^2 : x^2 + y^2 \leqslant r^2\}$ i $X : \Omega \to \mathbb{R}$ je dana s $X(\omega) = \sqrt{x^2 + y^2}$, $\omega = (x, y) \in \Omega$. Slično kao i u prethodnom primjeru sve točke su "jednako vjerojatne" i u pozadini je ponovno geometrijska vjerojatnost, tj.

$$\mathbb{P}(A) = \frac{\text{pov}(A)}{r^2\pi}, \qquad A \subseteq \Omega,$$

gdje pov(A) predstavlja površinu skupa A. Dakle,

$$F(x) = \mathbb{P}(X \leqslant x) = \begin{cases} 0, & x < 0 \\ \frac{x^2}{r^2}, & 0 \leqslant x < r \\ 1, & x \geqslant r \end{cases}$$

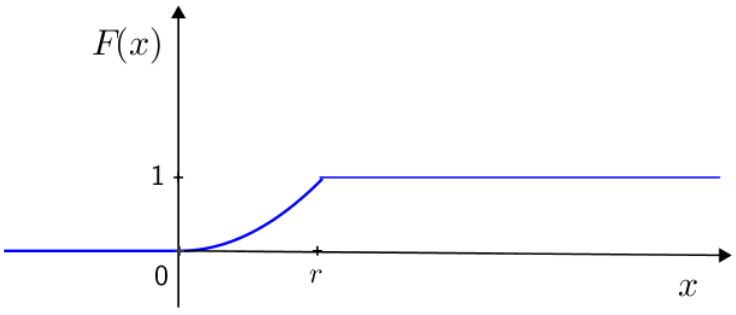
$$f(x) = F'(x) = \begin{cases} 0, & x < 0 \\ \frac{2x}{r^2}, & 0 \leqslant x < r \\ 0, & x > r \end{cases}$$

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x)\, dx = \int_0^r \frac{2x^2}{r^2}\, dx = \frac{2}{3}r$$

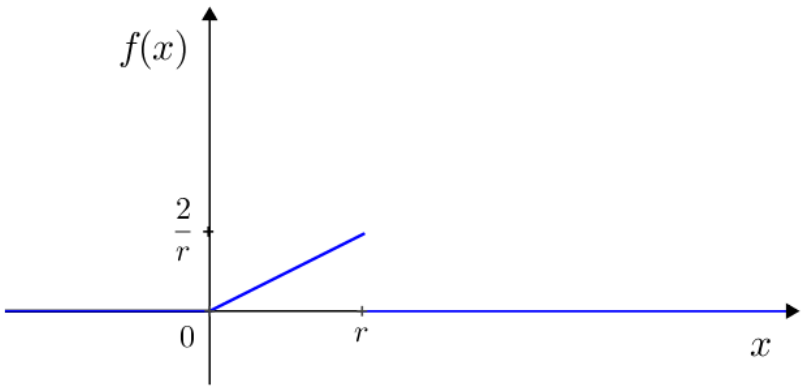
$$\text{Var}(X) = \int_{-\infty}^{\infty} x^2 f(x)\, dx - \mathbb{E}(X)^2 = \int_0^r \frac{2x^3}{r^2}\, dx - \frac{4}{9}r^2 = \frac{1}{18}r^2$$

$$\sigma(X) = \sqrt{\text{Var}(X)} = \frac{r}{3\sqrt{2}}.$$

Uočimo da $F(x)$ nije derivabilna u $x = r$. Grafovi funkcije gustoće i funkcije raspodjele prikazani su na Slikama 3.3 i 3.4.



Slika 3.3: Funkcija raspodjele iz Primjera 3.19



Slika 3.4: Funkcija gustoće iz Primjera 3.19

□

Primjer 3.20. Neka je X slučajna varijabla čija je vrijednost vrijeme trajanja akumulatora i dana je funkcijom raspodjele

$$F(x) = \mathbb{P}(X \leqslant x) = \begin{cases} 0, & x < 0 \\ \frac{1}{4}x^2, & 0 \leqslant x < 2 \\ 1, & x \geqslant 2. \end{cases}$$

Odredimo funkciju gustoće od X, izračunajmo pripadno očekivanje, varijancu i standardnu devijaciju te odredimo vjerojatnost da se akumulator potrošio u razdoblju od 1.5 godine do 2 godine. Imamo

$$f(x) = F'(x) = \begin{cases} 0, & x < 0 \\ \frac{x}{2}, & 0 \leqslant x < 2 \\ 0, & x > 2. \end{cases}$$

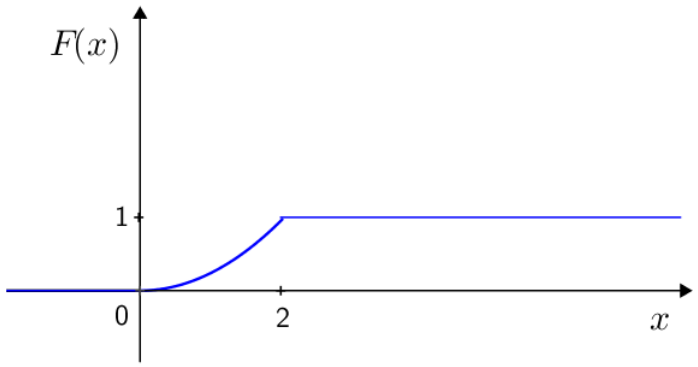
Uočimo da $F(x)$ nije derivabilna u $x = 2$. Grafovi funkcije gustoće i funkcije raspodjele prikazani su na Slikama 3.5 i 3.6.

Dakle,

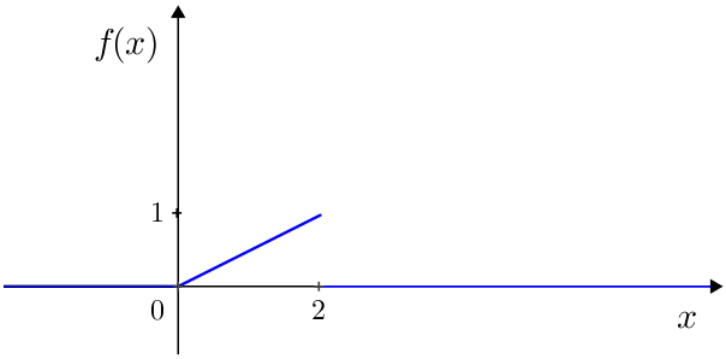
$$\begin{aligned} R(X) &= [0, 2] \\ \mathbb{E}(X) &= \int_{-\infty}^{\infty} x f(x)\, dx = \int_0^2 \frac{x^2}{2}\, dx = \frac{4}{3} \\ \mathrm{Var}(X) &= \int_{-\infty}^{\infty} x^2 f(x)\, dx - \mathbb{E}(X)^2 = \int_0^2 \frac{x^3}{2}\, dx - \frac{16}{9} = \frac{2}{9} \\ \sigma(X) &= \sqrt{\mathrm{Var}(X)} = \frac{\sqrt{2}}{3}. \end{aligned}$$

Tražena vjerojatnost je

$$\mathbb{P}(1.5 < X \leqslant 2) = F(2) - F(1.5) = 0.4375.$$



Slika 3.5: Funkcija raspodjele iz Primjera 3.20



Slika 3.6: Funkcija gustoće iz Primjera 3.20

□

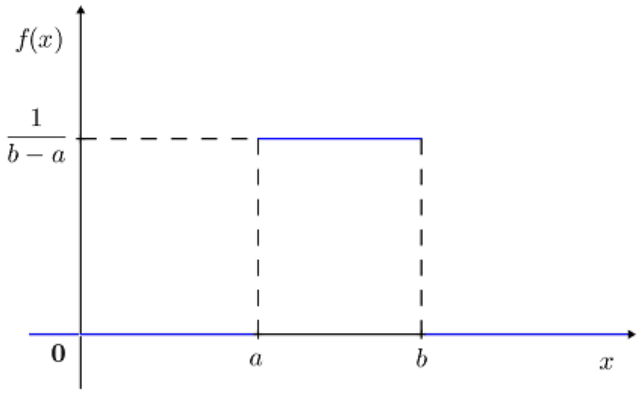
Pogledajmo sada primjere nekih neprekidnih slučajnih varijabli.

3.2.1 Uniformna slučajna varijabla

Uniformna slučajna varijabla X na segmentu $[a, b]$, u oznaci $X \sim U(a, b)$, je neprekidna slučajna varijabla za koju vrijedi $R(X) = [a, b]$ i

$$f(x) = \begin{cases} \frac{1}{b-a}, & x \in (a, b) \\ 0, & \text{inače.} \end{cases}$$

Očito, $f(x)$ je funkcija gustoće. Njen graf prikazan je na Slici 3.7.

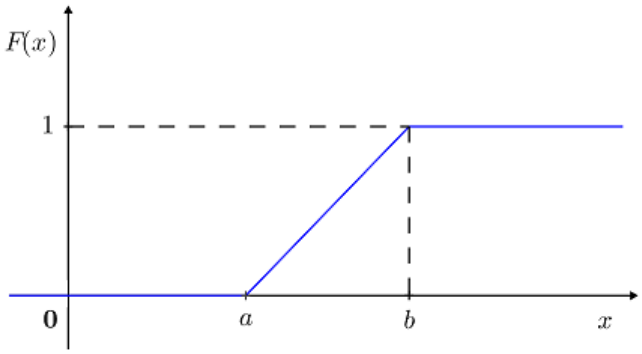


Slika 3.7: Funkcija gustoće uniformne slučajne varijable

Uočimo da je

$$F(x) = \int_{-\infty}^{x} f(t)\,dt = \begin{cases} 0, & x < a \\ \frac{x-a}{b-a}, & a \leqslant x < b \\ 1, & x \geqslant b. \end{cases}$$

Funkcija raspodjele prikazana je na Slici 3.8.



Slika 3.8: Funkcija raspodjele uniformne slučajne varijable

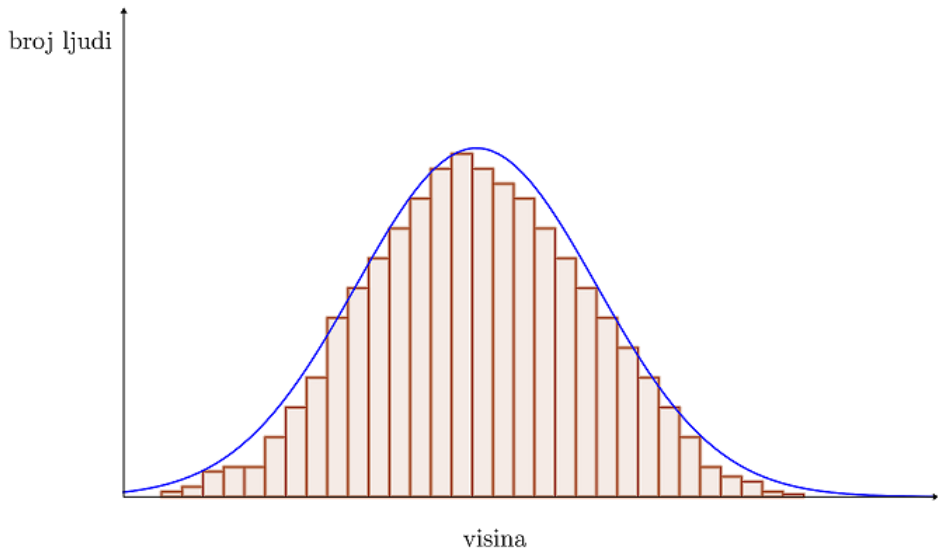
U Primjeru 3.18 smo izračunali

$$\mathbb{E}(X) = \frac{a+b}{2} \quad \text{i} \quad \text{Var}(X) = \frac{(b-a)^2}{12}.$$

Ova slučajna varijabla koristi se kod eksperimenata kod kojih je pripadnost ishoda jednako dugim podintervalima od $[a, b]$ jednako vjerojatna.

3.2.2 Normalna (Gaussova) slučajna varijabla

Normalna slučajna varijabla jedna je od najvažnijih i najčešće korištenih slučajnih varijabli. Ona opisuje mnoge slučajne pojave u prirodi i društvu. Primjerice, ako mjerimo visinu svih građana Republike Hrvatske dobit ćemo nešto ovog tipa:



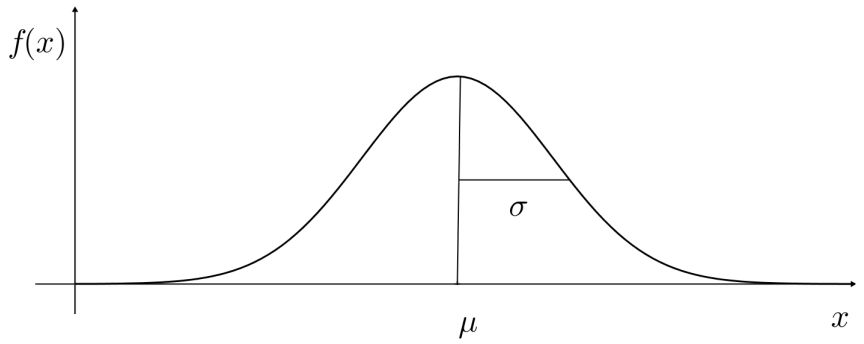
Slika 3.9: Ilustracija učestalosti različitih visina

Krivulja koja "uokviruje stupiće" je zvonolikog oblika kod kojega su (na neki način) bitne samo dvije informacije: srednja vrijednost i raspršenje. Postavlja se pitanje koja to slučajna varijabla dobro aproksimira? Sličan graf daju i slučajne pojave tipa: rezultati na ispitima, stvarna težina nekog proizvoda, greška koja nastaje prilikom mjerenja neke veličine, itd. Odgovor na gornje pitanje je normalna ili Gaussova slučajna varijabla.

Normalna slučajna varijabla X s parametrima $\mu \in \mathbb{R}$ i $\sigma > 0$, u oznaci $X \sim N(\mu, \sigma^2)$, je neprekidna slučajna varijabla dana s $R(X) = \mathbb{R}$ i funkcijom gustoće

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

Pokažite da je $f(x)$ funkcija gustoće (vidi [9, str. 109]). Njen graf prikazan je na Slici 3.10.



Slika 3.10: Funkcija gustoće normalne slučajne varijable

Dakle, $f(x)$ je zvonolikog oblika, simetrična je oko μ i

$$\int_{-\infty}^{\mu} f(x)\,dx = \int_{\mu}^{\infty} f(x)\,dx = \frac{1}{2}.$$

Parametar μ zovemo parametrom lokacije, a σ^2 zovemo parametrom raspršenja. Naime, vrijedi

$$\mathbb{E}(X) = \mu \qquad \text{i} \qquad \mathrm{Var}(X) = \sigma^2.$$

Zaista,

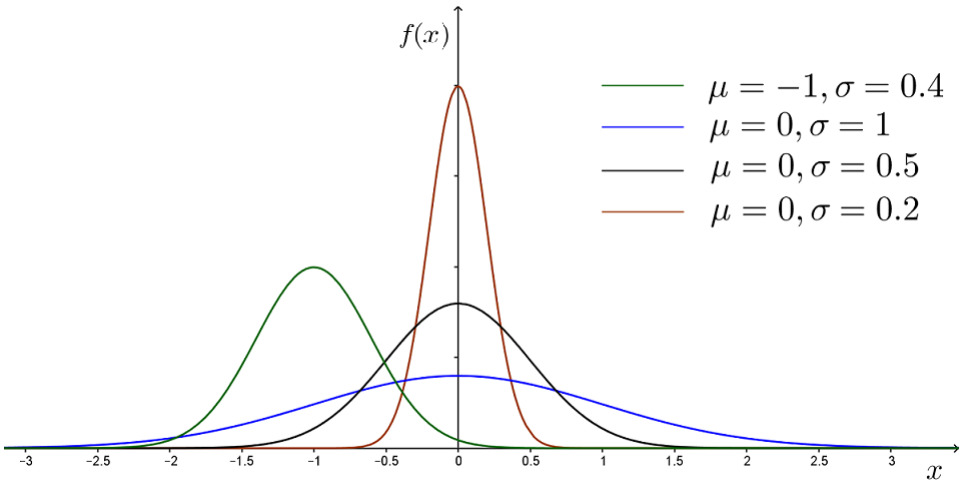
$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x) dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (y+\mu) e^{-\frac{y^2}{2\sigma^2}} dy = \mu,$$

gdje smo u prvom koraku napravili zamjenu varijabli $y = x - \mu$, a u drugom smo iskoristili neparnost funkcije $ye^{-y^2/2\sigma^2}$ na području integracije. Analogno,

$$\begin{aligned}
\mathrm{Var}(X) &= \int_{-\infty}^{\infty} (x - \mathbb{E}(X))^2 f(x) dx \\
&= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} y^2 e^{-\frac{y^2}{2\sigma^2}} dy \\
&= \frac{2}{\sigma\sqrt{2\pi}} \int_{0}^{\infty} y^2 e^{-\frac{y^2}{2\sigma^2}} dy \\
&= \frac{2}{\sigma\sqrt{2\pi}} \left(-\sigma^2 y e^{-\frac{y^2}{2\sigma^2}} \Big|_0^{\infty} + \sigma^2 \int_0^{\infty} e^{-\frac{y^2}{2\sigma^2}} dy \right) \\
&= \sigma^2,
\end{aligned}$$

gdje smo u prvom koraku napravili zamjenu varijabli $y = x - \mathbb{E}(X) = x - \mu$, u drugom koraku smo iskoristili parnost funkcije $y^2 e^{-y^2/2\sigma^2}$ na području integracije, a u trećem smo koristili metodu parcijalne integracije.

Na Slici 3.11 prikazani su grafovi funkcija gustoća normalnih slučajnih varijabli različitih varijanci i očekivanja.



Slika 3.11: Funkcije gustoća različitih normalnih slučajnih varijabli

Prvi koji je uočio značajnost normalne slučajne varijable je C. F. Gauss koji je 1809. godine pokazao da se greške načinjene prilikom astronomskih mjerenja mogu modelirati normalnom raspodjelom. Postoje i druge zvonolike funkcije gustoće, na primjer $f(x) = 1/\pi(1 + x^2)$ (pokažite da je to zaista fukcija gustoća neke neprekidne slučajne varijable). Međutim, pokazuje se da se većina prirodnih pojava ponaša baš po normalnoj raspodjeli. Napomenimo da se raspodjela sa gore spomenutom funkcijom gustoće naziva Cauchyeva raspodjela. Uočimo da ta raspodjela nema očekivanje i varijancu. Naime, $|x|f(x)$ i $x^2 f(x)$ se za velike $|x|$ ponašaju, redom, kao $1/\pi|x|$ i $1/\pi$, što nije integrabilno.

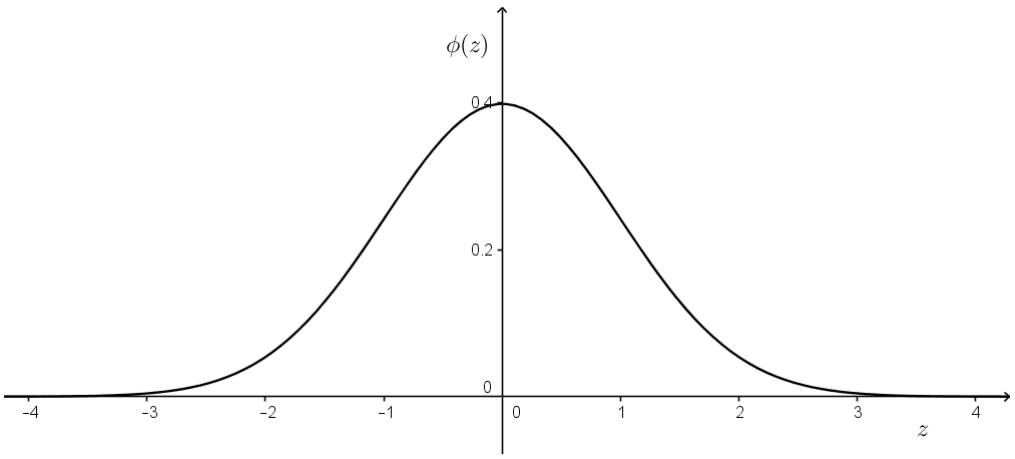
Funkcija raspodjele od $X \sim N(\mu, \sigma^2)$ dana je s

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

Ovaj integral se ne da elementarno riješiti i zato su vrijednosti od $F(x)$ tabelirane. Međutim, bilo bi nepraktično tabelirati $F(x)$ za sve μ i σ pa to činimo samo za jedan slučaj, za takozvanu **jediničnu normalnu slučajnu varijablu**, a ostale dobivamo iz ovog. Jedinična normalna slučajna varijabla je

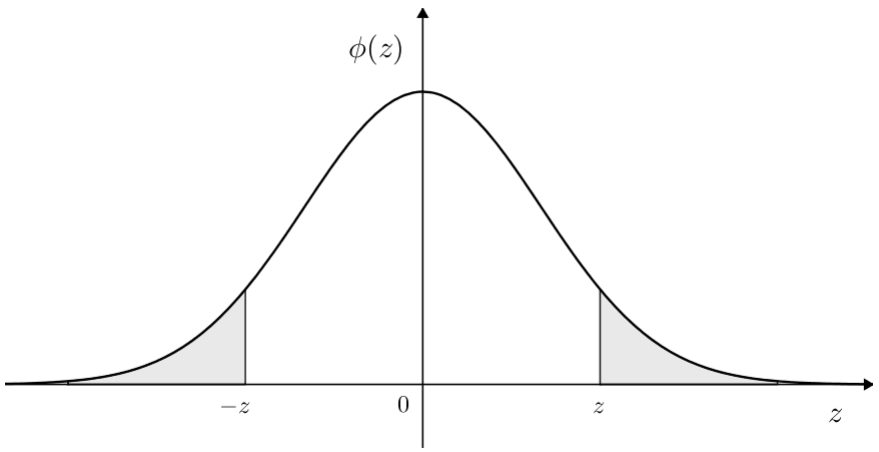
$$Z \sim N(0,1), \quad \phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \quad \text{i} \quad \Phi(z) = \mathbb{P}(Z \leqslant z) = \int_{-\infty}^{z} \phi(x)\, dx.$$

Graf funkcije gustoće jedinične normalne slučajne varijable, $\phi(z)$, prikazan je na Slici 3.12.



Slika 3.12: Funkcija gustoće jedinične normalne slučajne varijable

Vrijednosti na osi apscisa, u slučaju standardne normalne slučajne varijable, interpretiraju se u jedinicama standardnih devijacija i nazivaju se **z-vrijednosti**. Na primjer, izraz $z = 2$ označava da je točka apscise udaljena za dvije standardne devijacije u desno od očekivanja (vidi Sliku 3.13).



Slika 3.13: z-vrijednosti označavaju udaljenost od očekivanja u broju standardnih devijacija

Imajući tabeliranu $Z \sim N(0,1)$, slučajnu varijablu $X \sim N(\mu, \sigma^2)$ dobijemo iz

$$X = \sigma Z + \mu.$$

Dakle,

$$\begin{aligned} Z \sim N(0,1) &\implies \sigma Z + \mu \sim N(\mu, \sigma^2) \\ X \sim N(\mu, \sigma^2) &\implies \frac{X-\mu}{\sigma} \sim N(0,1). \end{aligned}$$

Sada imamo

$$F(x) = \mathbb{P}(X \leqslant x) = \mathbb{P}\left(\frac{X-\mu}{\sigma} \leqslant \frac{x-\mu}{\sigma}\right) = \mathbb{P}(Z \leqslant z) = \Phi(z),$$

gdje je $z = (x-\mu)/\sigma$ i $\Phi(z)$ iščitamo iz tablice. Uočimo još da je dovoljno tabelirati vrijednosti od $\Phi(z)$ samo za $z \geqslant 0$, jer vrijedi

$$\Phi(-z) = \mathbb{P}(Z \leqslant -z) = \mathbb{P}(Z \geqslant z) = 1 - \mathbb{P}(Z \leqslant z) = 1 - \Phi(z).$$

Primjer 3.21. Neka je $X \sim N(5, 25)$. Odredimo $\mathbb{P}(-2 < X < 10)$. Imamo

$$\begin{aligned} \mathbb{P}(-2 < X < 10) &= F(10) - F(-2) \\ &= \Phi\left(\frac{10-5}{5}\right) - \Phi\left(\frac{-2-5}{5}\right) \\ &= \Phi(1) - \Phi(-1.4) \\ &= \Phi(1) - 1 + \Phi(1.4) \\ &= 0.761. \end{aligned}$$

□

3.2.3 Eksponencijalna slučajna varijabla

Eksponencijalna slučajna varijabla X s parametrom $\lambda > 0$, u oznaci $X \sim \text{Exp}(\lambda)$, je neprekidna slučajna varijabla dana s $R(X) = (0, \infty)$ i

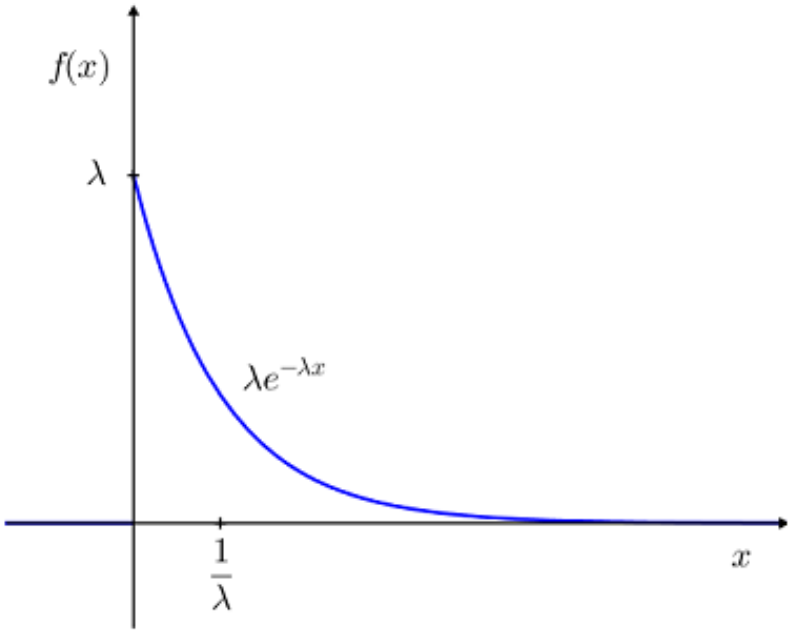
$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0, & \text{inače.} \end{cases}$$

Pokažite da je $f(x)$ funkcija gustoće. Njen graf prikazan je na Slici 3.14. Uočimo da vrijedi

$$F(x) = \int_{-\infty}^{x} f(t)\,dt = \begin{cases} 1 - e^{-\lambda x}, & x > 0 \\ 0, & \text{inače.} \end{cases}$$

Također, koristeći metodu parcijalne integracije, lagano zaključimo da vrijedi

$$\mathbb{E}(X) = \frac{1}{\lambda} \qquad \text{i} \qquad \text{Var}(X) = \frac{1}{\lambda^2}.$$



Slika 3.14: Funkcija gustoće eksponencijalne slučajne varijable

Ova slučajna varijabla usko je povezana s Poissonovom slučajnom varijablom koja broji slučajne događaje u jedinici vremena (frekvenciju), dok eksponencijalna mjeri vrijeme između dva slučajna događaja: dolazak telefonskih poziva u centralu, dolazak mušterija u trgovinu, itd. Parametar λ, slično kao i kod Poissonove slučajne varijable, označava prosječan broj pojavljivanja u jedinici vremena, tj. prosječnu frekvenciju.

Primjer 3.22. Službenik na šalteru posluži u prosjeku 30 stranaka na sat. Ako je vrijeme posluživanja eksponencijalna slučajna varijabla, kolika je vjerojatnost da će iduća stranka potrošiti više od 5 minuta na posluživanju (i čekanju)? Kolika je vjerojatnost da će potrošiti manje od 2 minute? Imamo $X \sim \text{Exp}(0.5)$. Dakle,

$$\mathbb{P}(X > 5) = 1 - \mathbb{P}(X < 5) = 1 - F(5) = e^{-0.5 \cdot 5} = 0.082$$
$$\mathbb{P}(X < 2) = F(2) = 1 - e^{-0.5 \cdot 2} = 0.632. \qquad \square$$

Eksponencijalna slučajna varijable može poslužiti za aproksimiranje geometrijske slučajne varijable. Zaista, neka je $X_g \sim G(p)$. Pretpostavimo da je p "mali", recimo $p = 1/n$ za veliki $n \in \mathbb{N}$. Sada, za $i \geq 1$, imamo

$$\mathbb{P}(X_g \leq i) = 1 - \mathbb{P}(X_g > i) = 1 - (1-p)^i = 1 - \left(1 - \frac{1}{n}\right)^{n\frac{i}{n}} \approx 1 - e^{-\frac{i}{n}},$$

što nije ništa drugo nego raspodjela (funkcija raspodjele) eksponencijalne slučajne varijable s parametrom $1/n$.

Slično kao i geometrijska slučajna varijabla, eksponencijalna slučajna varijabla ima svojstvo zaboravljivosti: za sve $s, t \geq 0$ vrijedi

$$\mathbb{P}(X > s + t | X > s) = \mathbb{P}(X > t).$$

Drugim riječima, ako X označava vrijeme čekanja nekog događaja, onda gornja relacija govori da ako je X uvjetovano na nepojavljivanje događaja kroz neki početni vremenski interval duljine s, raspodjela preostalog vremena jednaka je kao i neuvjetovana raspodjela.

3.2.4 Paretova slučajna varijabla

Paretova slučajna varijabla X s parametrima $\alpha > 0$ i $x_m > 0$, u oznaci $X \sim \mathrm{Par}(\alpha, x_m)$, je neprekidna slučajna varijabla dana s $R(X) = [x_m, \infty)$ i

$$f(x) = \begin{cases} \frac{\alpha x_m^\alpha}{x^{\alpha+1}}, & x \geqslant x_m \\ 0, & \text{inače.} \end{cases}$$

Pokažite da je $f(x)$ funkcija gustoće. Uočimo da vrijedi

$$F(x) = \int_{-\infty}^{x} f(t)\,dt = \begin{cases} 1 - \frac{x_m^\alpha}{x^\alpha}, & x \geqslant x_m \\ 0, & \text{inače,} \end{cases}$$

i

$$\mathbb{E}(X) = \begin{cases} \frac{\alpha x_m}{\alpha - 1}, & \alpha > 1 \\ \infty, & \text{inače} \end{cases} \qquad \text{i} \qquad \mathrm{Var}(X) = \begin{cases} \frac{\alpha x_m^2}{(\alpha-2)(\alpha-1)^2}, & \alpha > 2 \\ \infty, & \text{inače.} \end{cases}$$

V. Pareto je uočio da se broj ljudi čiji su prihodi veći od x_m može dobro aproksimirati Paretovom slučajnom varijablom s određenim parametrom α. Slično, ovom slučajnom varijablom dobro se može aproksimirati populacija gradova, veličine tvrtki, magnitude potresa, itd. Uočimo da za $0 < \alpha \leqslant 1$ imamo $\mathbb{E}(X) = \infty$ i za $0 < \alpha \leqslant 2$ imamo $\mathrm{Var}(X) = \infty$. To znači da u tim situacijama za danu slučajnu varijablu ne postoje standardne determinističke karakteristike (očekivanje i varijanca).

Poglavlje 4

Čebiševljeva nejednakost, nezavisnost slučajnih varijabli i granični teoremi

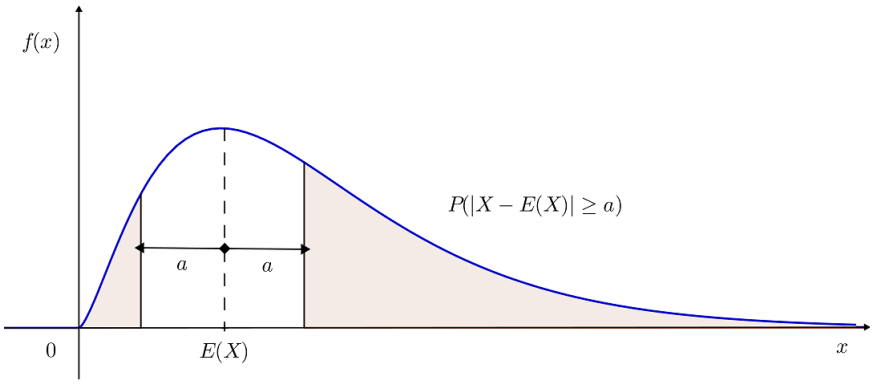
Pogledajmo sljedeći problem. Za slučajnu varijablu X znamo da njena standardna devijacija $\sigma(X)$ (ako postoji) daje informaciju o prosječnoj raspršenosti vrijednosti od X oko $\mathbb{E}(X)$. Dakle, veća $\sigma(X)$ znači da su vrijednosti raspršenije, a manja $\sigma(X)$ znači da su lokalizirane oko $\mathbb{E}(X)$. Međutim, htjeli bismo odrediti proporciju vrijednosti slučajne varijable koja upada u određeni interval oko $\mathbb{E}(X)$. Odgovor na to pitanje nam daje tzv. **Čebiševljeva nejednakost**: za slučajnu varijablu X (koja ima varijancu) i proizvoljan $a > 0$ vrijedi

$$\mathbb{P}(|X - \mathbb{E}(X)| \geqslant a) \leqslant \frac{\mathrm{Var}(X)}{a^2}$$

ili

$$\mathbb{P}(|X - \mathbb{E}(X)| < a) \geqslant 1 - \frac{\mathrm{Var}(X)}{a^2}$$

(vidi Sliku 4.1 te [9, str. 313]).



Slika 4.1: Ilustracija Čebiševljeve nejednakosti

Specijalno, za $a = n\sigma(X)$ imamo

$$\mathbb{P}(|X - \mathbb{E}(X)| \geqslant n\sigma(X)) \leqslant \frac{1}{n^2}$$

ili

$$\mathbb{P}(|X - \mathbb{E}(X)| < n\sigma(X)) \geqslant 1 - \frac{1}{n^2}.$$

Primjerice za $n = 3$ dobivamo

$$\mathbb{P}(|X - \mathbb{E}(X)| \geqslant 3\sigma(X)) \leqslant \frac{1}{9} \approx 0.11$$

ili

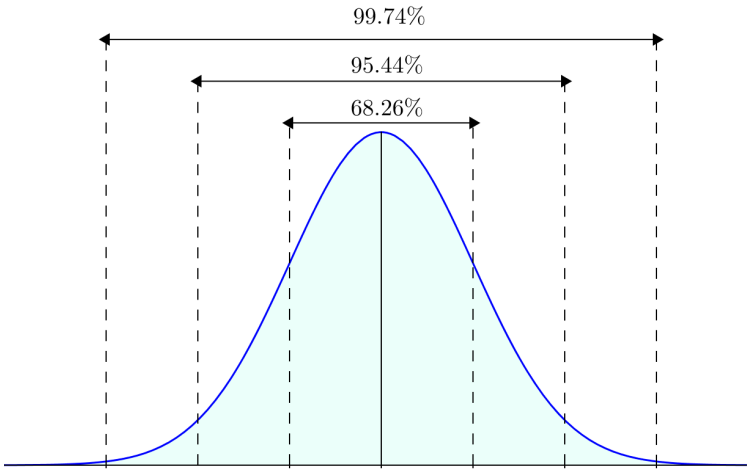
$$\mathbb{P}(|X - \mathbb{E}(X)| < 3\sigma(X)) \geqslant 1 - \frac{1}{9} = \frac{8}{9} \approx 0.89.$$

Dakle, barem 89% realizacija slučajne varijable X se nalazi u intervalu tri standardne devijacije lijevo i tri standardne devijacije desno od njezinog očekivanja.

Čebiševljeva nejednakost vrijedi za svaku slučajnu varijablu (koja ima varijancu). U slučaju normalne slučajne varijable $X \sim N(\mu, \sigma^2)$ možemo biti precizniji, tj. raspršenje možemo izračunati egzaktno:

$$\begin{aligned}\mathbb{P}(|X - \mu| < \sigma) &\approx 0.682\\ \mathbb{P}(|X - \mu| < 2\sigma) &\approx 0.954\\ \mathbb{P}(|X - \mu| < 3\sigma) &\approx 0.997.\end{aligned}$$

Ova raspršenja prikazana su i na Slici 4.2.



Slika 4.2: Udjeli normalne raspodjele udaljeni σ, 2σ i 3σ od očekivanja

Prisjetimo se, dva događaja A i B su nezavisna ako je $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. Analogno definiramo nezavisnost slučajnih varijabli.

Definicija 4.1. Familija slučajnih varijabli X_i, $i \in I$, je **nezavisna** ako za svaki konačan podskup različitih indeksa $\{i_1, \ldots, i_n\} \subseteq I$ i sve $x_1, x_2, \ldots, x_n \in \mathbb{R}$ vrijedi

$$\begin{aligned}&\mathbb{P}(X_{i_1} \leq x_1, X_{i_2} \leq x_2, \cdots, X_{i_n} \leq x_n)\\ &= \mathbb{P}(X_{i_1} \leq x_1)\mathbb{P}(X_{i_2} \leq x_2) \cdots \mathbb{P}(X_{i_n} \leq x_3).\end{aligned}$$

Uvedimo i pojam jednakosti po raspodjeli među slučajnim varijablama.

Definicija 4.2. Slučajne varijable X i Y su **jednako distribuirane** ako za svaki $x \in \mathbb{R}$ vrijedi

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

Uočimo sljedeće:

(i) u slučaju nezavisnosti slučajne varijable moraju biti definirane na istom vjerojatnosnom prostoru.

(ii) u slučaju jednake distribuiranosti slučajne varijable ne moraju biti definirane na istom vjerojatnosnom prostoru.

(iii) jednaka distribuiranost znači da slučajne varijable imaju jednake slike, funkcije gustoće, funkcije raspodjele pa samim time i očekivanja, varijance i standardne devijacije. Međutim, to ne znači da su jednake kao funkcije jer, kao što smo već napomenuli u (ii), ne moraju niti imati iste domene. Čak iako su definirane na istom vjerojatnosnom prostoru ne moraju biti jednake. Primjerice, neka je $(\Omega, \mathcal{F}, \mathbb{P})$ vjerojatnosni prostor vezan za pokus bacanja simetričnog novčića. Dakle, $\Omega = \{P, G\}$. Neka je $X : \Omega \longrightarrow \mathbb{R}$ dana sa $X(P) = 0$ i $X(G) = 1$ te neka je $Y : \Omega \longrightarrow \mathbb{R}$ dana sa $Y(P) = 1$ i $Y(G) = 0$. Očito, X i Y nisu jednake te su jednako dstribuirane (obje imaju Bernoullijevu raspodjelu s parametrom $1/2$). Također, uočimo da ako dvije slučajne varijable imaju ista očekivanja i varijance to ne znači da su jednako distribuirane. Neka je $X \sim N(0, 1)$ i

$$Y \sim \begin{pmatrix} -1 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$

Tada, $\mathbb{E}(X) = \mathbb{E}(Y) = 0$ i $\mathrm{Var}(X) = \mathrm{Var}(Y) = 1$, ali X i Y nisu jednako distribuirane.

(iv) kasnije ćemo vidjeti da nezavisnost od X i Y povlači $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$ (obrat ne mora vrijediti).

(v) znamo da za sve slučajne varijable X i Y vrijedi $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$ (ako očekivanja postoje). Ako su X i Y još i nezavisne slična relacija vrijedi i za varijancu, tj. $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$. Napomenimo da za zavisne slučajne varijable prethodna relacija ne mora vrijediti, što ćemo vidjeti kasnije.

Nas će zanimati slučajne varijable koje su nezavisne i jednako distribuirane. Promotrimo sljedeću situaciju. Neka su

$$X_i \sim \begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}, \quad i = 1, \ldots, n,$$

nezavisne. Tada, $X := X_1 + \cdots + X_n$ je $B(n, p)$ (vidi [9, str. 93]). Primjerice, za $n = 3$ imamo

$$X = X_1 + X_2 + X_3 \sim \begin{pmatrix} 0 & 1 & 2 & 3 \\ p_0 & p_1 & p_2 & p_3 \end{pmatrix},$$

gdje su

$$\begin{aligned} p_0 &= \mathbb{P}(X=0) \\ &= \mathbb{P}(X_1+X_2+X_3=0) \\ &= \mathbb{P}(X_1=0, X_2=0, X_3=0) \\ &= \mathbb{P}(X_1=0)\mathbb{P}(X_2=0)\mathbb{P}(X_3=0) \\ &= (1-p)^3 \end{aligned}$$

$$\begin{aligned} p_1 &= \mathbb{P}(X=0) \\ &= \mathbb{P}(X_1+X_2+X_3=1) \\ &= \mathbb{P}(X_1=1, X_2=0, X_3=0) + \mathbb{P}(X_1=0, X_2=1, X_3=0) \\ &\quad + \mathbb{P}(X_1=0, X_2=0, X_3=1) \\ &= \mathbb{P}(X_1=1)\mathbb{P}(X_2=0)\mathbb{P}(X_3=0) + \mathbb{P}(X_1=0)\mathbb{P}(X_2=1)\mathbb{P}(X_3=0) \\ &\quad + \mathbb{P}(X_1=0)\mathbb{P}(X_2=0)\mathbb{P}(X_3=1) \\ &= 3p^2(1-p)^2 \end{aligned}$$

$$p_2 = 3p^2(1-p)$$

$$p_3 = p^3$$

Nadalje, iz $X = X_1 + X_2 + \cdots + X_n$ jednostavno slijedi da je $\mathbb{E}(X) = n\mathbb{E}(X_1) = np$ i $\mathrm{Var}(X) = n\,\mathrm{Var}(X_1) = np(1-p)$, čime su pokazane relacije za očekivanje i varijancu binomne slučajne varijable komentirane u Poglavlju 3.1.3.

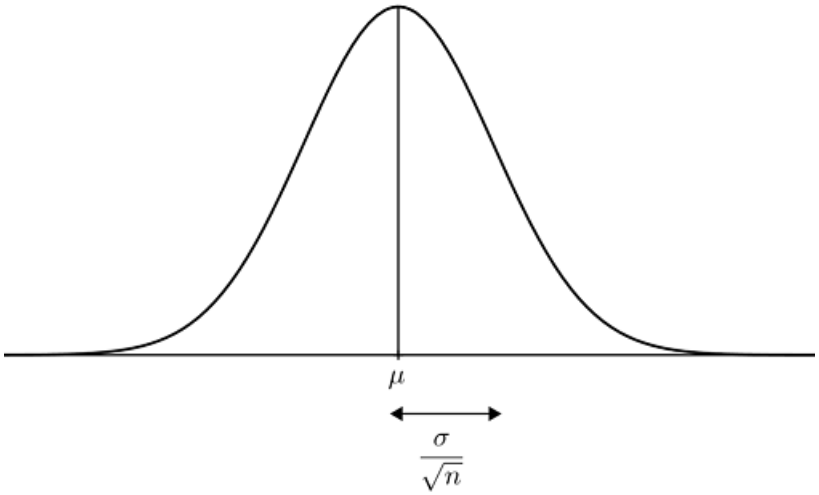
Prema definiciji vjerojatnosti *aposteriori* vjerojatnost događaja A je dana kao limes omjera broja pojavljivanja događaja A u n ponavljanja pokusa:

$$\mathbb{P}(A) = \lim_{n\to\infty} \frac{n_A}{n}.$$

Analogna tvrdnja vrijedi za slučajne varijable. Neka su $X_1, X_2, \ldots$ nezavisne i jednako distribuirane slučajne varijable s očekivanjem μ i varijancom σ^2. Stavimo $S_n = X_1 + \cdots + X_n$. Tada vrijedi

$$\mathbb{E}\left(\frac{S_n}{n}\right) = \mu \qquad \text{i} \qquad \mathrm{Var}\left(\frac{S_n}{n}\right) = \frac{\sigma^2}{n},$$

tj. $\sigma(S_n/n) = \sigma/\sqrt{n}$. Dakle, nakon n nezavisnih ponavljanja pokusa, standardna devijacija se smanjuje s faktorom $1/\sqrt{n}$. Što više ponavljamo pokus, vrijednost od S_n/n se u prosjeku sve više koncentrira oko μ (vidi Sliku 4.3).



Slika 4.3: Raspodjela aritmetičkih sredina S_n/n oko očekivanja μ

Primjena Čebiševljeve nejednakosti nam daje tzv. **slabi zakon velikih brojeva**: za sve $a > 0$ vrijedi

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| \geqslant a\right) \leqslant \lim_{n\to\infty} \frac{\sigma^2}{a^2 n} = 0.$$

Dakle, vjerojatnost da prosjek nezavisnih i jednako distribuiranih slučajnih varijabli odstupa od svog očekivanja za proizvoljni $a > 0$ možemo učiniti proizvoljno malom odabirom dovoljno velikog $n \in \mathbb{N}$. U istoj situaciji može se pokazati i jači rezultat, tzv. **jaki zakon velikih brojeva** (vidjeti [9, str. 416]):

$$\mathbb{P}\left(\left\{\omega \in \Omega : \lim_{n\to\infty} \frac{S_n(\omega)}{n} = \mu\right\}\right) = 1.$$

Uočimo da nam slabi zakon velikih brojeva daje informaciju o prosjeku niza nezavisnih i jednako distribuiranih slučajnih variabli S_n/n, što je ponovno slučajna varijabla, dok jaki zakon velikih brojeva daje informaciju o prosjeku pojedinačnih ishoda niza nezavisnih i jednako distribuiranih slučajnih variabli $S_n(\omega)/n$, $\omega \in \Omega$, što je broj. Može se pokazati da je jaki zakon velikih brojeva jači od slabog zakona (od tuda i naziv), tj. implicira ga (vidi [9, str. 322]). Pogledajmo sljedeću situaciju. Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ vjerojatnosni prostor i neka je $A \in \mathcal{F}$. Nadalje, neka je X_i, $i \in \mathbb{N}$, niz nezavisnih i jedako

distribuiranih slučajnih varijabli s Bernoullijevom raspodjelom

$$X_i \sim \begin{pmatrix} 0 & 1 \\ 1-\mathbb{P}(A) & \mathbb{P}(A) \end{pmatrix}.$$

Sada, po jakom zakonu velikih brojeva, imamo

$$\mathbb{P}\left(\left\{\omega \in \Omega : \lim_{n\to\infty} \frac{S_n(\omega)}{n} = \mathbb{E}(X_i) = \mathbb{P}(A)\right\}\right) = 1.$$

Usporedite gornju situaciju s definicijom vjerojatnosti *aposteriori*.

Izračunajmo grešku zakona velikih brojeva, tj. $S_n/n - \mu$. Za dani $a > 0$, iz Čebiševljeve nejednakosti imamo

$$\mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| < a\right) \geqslant 1 - \frac{\sigma^2}{a^2 n}.$$

Međutim, možemo dobiti i bolji opis greške. Uočimo da je

$$\mathbb{E}\left(\frac{S_n}{n} - \mu\right) = 0 \qquad \text{i} \qquad \operatorname{Var}\left(\frac{S_n}{n} - \mu\right) = \frac{\sigma^2}{n},$$

tj. očekivana greška je 0 s raspršenjem $\sigma/\sqrt{n}$. Pomnožimo sada $S_n/n - \mu$ sa $\sqrt{n}$. Tada imamo

$$\mathbb{E}\left(\frac{S_n}{\sqrt{n}} - \sqrt{n}\mu\right) = 0 \qquad \text{i} \qquad \operatorname{Var}\left(\frac{S_n}{\sqrt{n}} - \sqrt{n}\mu\right) = \sigma^2.$$

Dakle, raspršenje od $S_n/\sqrt{n} - \sqrt{n}\mu$ oko 0 (očekivanje od $S_n/\sqrt{n} - \sqrt{n}\mu$) ne konvergira u 0, tj. daje nam za naslutiti da možemo dobiti dodatnu informaciju o samoj grešci od $S_n/n - \mu$. Zaista, vrijedi tzv. **centralni granični teorem**:

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{S_n - n\mu}{\sqrt{n}} \in (a,b)\right) = \frac{1}{\sigma\sqrt{2\pi}} \int_a^b e^{-\frac{x^2}{2\sigma^2}}\, dx, \qquad a, b \in \mathbb{R}$$

(vidi [9, str. 507]). Drugim riječima, za velike n, $(S_n - n\mu)/\sqrt{n}$ je približno distribuirano kao $N(0, \sigma^2)$, tj. greška procjene S_n/n sa μ je približno normalno distribuirana s očekivanjem 0 i varijancom σ^2/n.

Vidjeli smo da binomnu slučajnu varijablu možemo aproksimirati Poissonovom. Primjenjujući centralni granični teorem to možemo učiniti i normalnom. Neka je $X \sim B(n,p)$. Vidjeli smo da je $X = X_1 + \cdots + X_n$, gdje su $X_1, \ldots, X_n$ nezavisne i jednako distribuirane Bernoullijeve slučajne varijable s raspodjelom

$$X_i \sim \begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}, \qquad i = 1, \ldots n.$$

Dakle, $\mathbb{E}(X_i) = p$ i $\mathrm{Var}(X_i) = p(1-p)$ za $i = 1, \dots n$. Sada, za velike n, iz centralnog graničnog teorema imamo da je

$$\frac{X_1 + X_2 + \cdots + X_n - np}{\sqrt{p(1-p)}\sqrt{n}}$$

približno distribuirano kao $N(0,1)$. U praksi gornju aproksimaciju koristimo kad je $np > 5$ i $n(1-p) > 5$ (vidi [10, str. 167]).

Primjer 4.1. Neka je $X \sim B(50, 0.237)$. Izračunajmo $\mathbb{P}(7 \leqslant X \leqslant 9)$. Znamo da je $X = X_1 + \cdots + X_{50}$, gdje su $X_1, \dots, X_{50}$ nezavisne jednako distribuirane Bernoullijeve slučajne varijable

$$X_i \sim \begin{pmatrix} 0 & 1 \\ 0.763 & 0.237 \end{pmatrix}, \qquad i = 1, \dots, 50.$$

Kako je $\mathbb{E}(X_i) = 0.237$, $\mathrm{Var}(X_i) = 0.18$ i $\sigma(X_i) = 0.42$ za $i = 1, \dots, 50$, iz centralnog graničnog teorema imamo

$$\begin{aligned}
\mathbb{P}(7 \leqslant X \leqslant 9) &= \mathbb{P}(7 \leqslant X_1 + \cdots + X_{50} \leqslant 9) \\
&= \mathbb{P}(-4.85 \leqslant X_1 + \cdots + X_{50} - 11.85 \leqslant -2.85) \\
&= \mathbb{P}(-1.613 \leqslant \frac{X_1 + \cdots + X_{50} - 11.85}{3} \leqslant -0.94) \\
&\approx \mathbb{P}(-1.613 \leqslant Z \leqslant -0.94) \\
&= \Phi(-0.94) - \Phi(-1.613) \\
&= \Phi(1.613) - \Phi(0.94) \\
&= 0.12.
\end{aligned}$$

□

Poglavlje 5

Slučajni vektori

Često se događa da nas zanima više slučajnih varijabli vezanih uz istu pojavu. Primjer je prihod, obrazovanje i spol kod popisa stanovništva. Svaka slučajna varijabla za sebe je zanimljiva, no kad ih gledamo zajedno možemo više tog zaključiti o populaciji koju proučavamo. Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ vjerojatnosni prostor i neka su $X, Y : \Omega \to \mathbb{R}$ dvije slučajne varijable. **Dvodimenzionalni slučajni vektor** je funkcija $(X, Y) : \Omega \to \mathbb{R}^2$, dana s $(X, Y)(\omega) = (X(\omega), Y(\omega))$. Uočimo da za sliku od (X, Y) vrijedi $R(X, Y) \subseteq R(X) \times R(Y)$. U slučaju da je

$$X \sim \begin{pmatrix} x_1 & x_2 & \dots \\ p_1 & p_2 & \dots \end{pmatrix} \qquad \text{i} \qquad Y \sim \begin{pmatrix} y_1 & y_2 & \dots \\ q_1 & q_2 & \dots \end{pmatrix},$$

onda je (X, Y) **diskretni dvodimenzionalni slučajni vektor** i njegovu raspodjelu zapisujemo pomoću sheme:

$$(X, Y) \sim \begin{pmatrix} X \backslash Y & y_1 & y_2 & \dots & y_m \\ x_1 & p_{11} & p_{12} & \dots & p_{1m} \\ x_2 & p_{21} & p_{22} & \dots & p_{2m} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_n & p_{n1} & p_{n2} & \dots & p_{nm} \end{pmatrix},$$

gdje su $R(X, Y) \subseteq R(X) \times R(Y) = \{(x_i, y_j) : x_i \in R(X), y_j \in R(Y)\}$ i

$$p_{ij} = \mathbb{P}((X, Y) = (x_i, y_j)) = \mathbb{P}(X = x_i, Y = y_j)$$

za $x_i \in R(X)$ i $y_j \in R(Y)$.

Funkcija vjerojatnosti $f : \mathbb{R}^2 \to \mathbb{R}$ od (X, Y) je dana s

$$f(x, y) = \begin{cases} p_{ij}, & x = x_i, y = y_j \\ 0, & \text{inače,} \end{cases}$$

a **funkcija raspodjele** $F : \mathbb{R}^2 \to [0,1]$ je dana s

$$F(x,y) = \mathbb{P}(X \leq x, Y \leq y) = \sum_{\substack{x_i \leq x,\, x_i \in R(X) \\ y_j \leq y,\, y_j \in R(Y)}} p_{ij}.$$

Očito, $p_{ij} = f(x_i, y_j)$ i $R(X,Y) = \{(x,y) \in \mathbb{R}^2 : f(x,y) > 0\}$. Nadalje, uočimo sljedeće

$$\sum_{y_j \in R(Y)} p_{ij} = \sum_{y_j \in R(Y)} \mathbb{P}(X = x_i, Y = y_j) = \mathbb{P}(X = x_i) = p_i$$

i slično

$$\sum_{x_i \in R(X)} p_{ij} = \sum_{x_i \in R(X)} \mathbb{P}(X = x_i, Y = y_j) = \mathbb{P}(Y = y_j) = q_j.$$

Dakle, ako zbrojimo sve elemente po y_j dobijemo raspodjelu od X i ako zbrojimo sve elemente po x_i dobijemo raspodjelu od Y. Te se raspodjele zovu **marginalne raspodjele** od (X,Y). Također vrijedi

$$\lim_{y \to \infty} F(x,y) = \lim_{y \to \infty} \mathbb{P}(X \leq x, Y \leq y) = \mathbb{P}(X \leq x) = F_X(x)$$

i slično $\lim_{x \to \infty} F(x,y) = F_Y(y)$. Uočimo, slučajne varijable X i Y su nezavisne ako, i samo ako, vrijedi $p_{ij} = p_i q_j$ i $F(x,y) = F_X(x) F_Y(y)$.

Primjer 5.1. Bacamo dvije kocke. Neka slučajne varijable X i Y poprimaju vrijednost, redom, zbroja brojeva koji su pali i većeg od brojeva koji su pali. Raspodjela od (X,Y) je dana s

$$\begin{pmatrix}
X \backslash Y & 1 & 2 & 3 & 4 & 5 & 6 & p_i \\
2 & \frac{1}{36} & 0 & 0 & 0 & 0 & 0 & \frac{1}{36} \\
3 & 0 & \frac{2}{36} & 0 & 0 & 0 & 0 & \frac{2}{36} \\
4 & 0 & \frac{1}{36} & \frac{2}{36} & 0 & 0 & 0 & \frac{3}{36} \\
5 & 0 & 0 & \frac{2}{36} & \frac{2}{36} & 0 & 0 & \frac{4}{36} \\
6 & 0 & 0 & \frac{1}{36} & \frac{2}{36} & \frac{2}{36} & 0 & \frac{5}{36} \\
7 & 0 & 0 & 0 & \frac{2}{36} & \frac{2}{36} & \frac{2}{36} & \frac{6}{36} \\
8 & 0 & 0 & 0 & \frac{1}{36} & \frac{2}{36} & \frac{2}{36} & \frac{5}{36} \\
9 & 0 & 0 & 0 & 0 & \frac{2}{36} & \frac{2}{36} & \frac{4}{36} \\
10 & 0 & 0 & 0 & 0 & \frac{1}{36} & \frac{2}{36} & \frac{3}{36} \\
11 & 0 & 0 & 0 & 0 & 0 & \frac{2}{36} & \frac{2}{36} \\
12 & 0 & 0 & 0 & 0 & 0 & \frac{1}{36} & \frac{1}{36} \\
q_j & \frac{1}{36} & \frac{3}{36} & \frac{5}{36} & \frac{7}{36} & \frac{9}{36} & \frac{11}{36} & 1
\end{pmatrix},$$

gdje su $p_{ij} = \mathbb{P}(X = x_i, Y = y_j)$, $p_i = \mathbb{P}(X = x_i)$ i $q_j = \mathbb{P}(Y = y_j)$. □

Neka su $X, Y : \Omega \to \mathbb{R}$ neprekidne slučajne varijable. Dvodimenzionalni slučajni vektor $(X, Y) : \Omega \to \mathbb{R}^2$ je **neprekidni slučajni vektor** ako postoji ("izmjeriva") funkcija $f : \mathbb{R}^2 \to \mathbb{R}$ za koju vrijedi:

(i) $f(x, y) \geq 0$

(ii) $\displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)dxdy = 1$

(iii) $\mathbb{P}((X, Y) \in (-\infty, a] \times (-\infty, b]) = \mathbb{P}(X \in (-\infty, a], Y \in (-\infty, b]) = \displaystyle\int_{-\infty}^{a}\int_{-\infty}^{b} f(x, y)dxdy.$

Funkciju $f(x, y)$ zovemo **funkcija gustoće** od (X, Y). Uočimo, $R(X, Y) = \{(x, y) \in \mathbb{R}^2 : f(x, y) > 0\} \subseteq R(X) \times R(Y)$. **Funkcija raspodjele** $F : \mathbb{R}^2 \to [0, 1]$ od (X, Y) je dana s

$$F(x, y) = \mathbb{P}(X \leq x, Y \leq y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(s, t)dsdt.$$

Analogno kao i u jednodimenzionalnom slučaju imamo (kada je funkcija gustoće neprekidna)

$$\frac{\partial^2}{\partial x \partial y}F(x, y) = \frac{\partial^2}{\partial y \partial x}F(x, y) = f(x, y).$$

Kao i u diskretnom slučaju imamo

$$\int_{-\infty}^{\infty} f(x, y)dy = f_X(x) \qquad \text{i} \qquad \int_{-\infty}^{\infty} f(x, y)dx = f_Y(y)$$

te

$$\lim_{y\to\infty} F(x, y) = F_X(x) \qquad \text{i} \qquad \lim_{x\to\infty} F(x, y) = F_Y(y).$$

Dakle, integriranjem po y dobivamo raspodjelu od X i integriranjem po x dobivamo raspodjelu od Y, tj. dobivamo **marginalne raspodjele** od (X, Y). Analogno kao i u diskretnom slučaju, slučajne varijable X i Y su nezavisne ako, i samo ako, vrijedi $f(x, y) = f_X(x)f_Y(y)$ i $F(x, y) = F_X(x)F_Y(y)$.

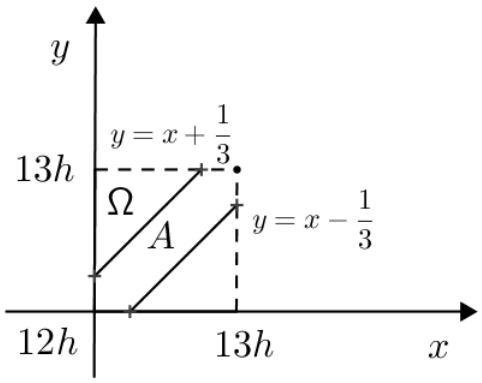
Napomenimo još da činjenica da su X i Y neprekidne slučajne varijable ne znači da je (X, Y) neprekidan slučajni vektor. Primjerice, neka je $X \sim U(0, 1)$ i $Y = X$. Tada je (X, Y) dvodimenzionalni slučajni vektor za koji vrijedi $R(X, Y) = \{(x, x) : x \in [0, 1]\}$ (dijagonala kvadrata $[0, 1] \times [0, 1]$). Budući da je $R(X, Y)$ skup površine nula zaključujemo da (X, Y) ne može biti neprekidni slučajni vektor, već je singularni neprekidni (vidjeti početak

Poglavlja 3.2 za kratak komentar i referencu o singularnim neprekidnim raspodjelama). Međutim, ako su X i Y još i nezavisne, onda je par (X, Y) uvijek neprekidan jer je pripadna funkcija gustoće baš $f(x, y) = f_X(x)f_Y(y)$. Također, ima smisla gledati i (dvodimenzionalne) slučajne vektore kod kojih je jedna komponenta neprekidna, a druga diskretna (recimo jedna komponenta označava krvni tlak, a druga broj otkucaja srca u minuti). Naravno, takav vektor ne može biti neprekidan niti diskretan, već je uvijek singularan neprekidan.

Primjer 5.2. Marko i Ivan dogovorili su sastanak na *Trgu bana Josipa Jelačića* u 12 h. Pretpostavimo da je njihov dolazak slučajan trenutak između 12 i 13 h. Neka slučajne varijable X i Y označavaju dolazak Marka i Ivana, redom. Pretpostavimo da je funkcija gustoće od (X, Y) dana s

$$f(x, y) = \begin{cases} 1, & x, y \in [12, 13] \\ 0, & \text{inače.} \end{cases}$$

Uočimo, $f_X(x) = \int_{12}^{13} f(x, y)dy = 1$ i $f_Y(y) = \int_{12}^{13} f(x, y)dx = 1$. Dakle, $X \sim U(12, 13)$ i $Y \sim U(12, 13)$. Nadalje, pretpostavimo da onaj koji stigne na dogovoreno mjesto prvi čeka onog drugog 20 min. Kolika je vjerojatnost da se sretnu? Neka je $\Omega = [12, 13]^2$. Označimo s A događaj "Marko i Ivan su se sreli", tj. $A = \{\omega \in \Omega : |X(\omega) - Y(\omega)| \leq 1/3\}$ (vidi Sliku 5.1).



Slika 5.1: Grafički prikaz skupa Ω i događaja A iz Primjera 5.2

Dakle,

$$\mathbb{P}(A) = \iint_A f(x, y)dxdy = \frac{5}{9}. \qquad \square$$

Neka je (X, Y) dvodimenzionalan slučajni vektor i neka je $g : \mathbb{R}^2 \to \mathbb{R}$ neka funkcija. Tada je $g(X, Y)$ jednodimenzionalna slučajna varijabla. Često nas zanima prosječna vrijednost od $g(X, Y)$. Pokazuje se da to možemo izravno izračunati iz zajedničke raspodjele, bez računanja raspodjele za $g(X, Y)$. Ako donja suma i integral postoje, onda imamo

$$\mathbb{E}(g(X,Y)) = \sum_{x_i \in R(X)} \sum_{y_j \in R(Y)} g(x_i, y_j) p_{ij}$$

za diskretan slučajni vektor te

$$\mathbb{E}(g(X,Y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f(x,y) dxdy$$

za neprekidan slučajni vektor. Napomenimo ovdje, kao i u jednodimenzionalnoj situaciji, u slučaju neprekidnih slučajnih vektora funkcija $g(x, y)$ mora biti "izmjeriva" da bi $g(X, Y)$ uopće bila slučajna varijabla. U slučaju diskretnih slučajnih vektora funkcija $g(x, y)$ može biti proizvoljna.

Primjer 5.3. Poludjeli stroj za proizvodnju vaza na slučajan način proizvodi vaze radijusa $R \sim U(7.5, 12.5)$ i visine $H \sim U(25, 35)$, gdje su obje dimenzije izražene u centimetrima. Pretpostavimo da su R i H nezavisne. Odredimo očekivani volumen slučajno odabrane vaze. Označimo s V volumen slučajno proizvedene vaze. Imamo

$$\mathbb{E}(V) = \mathbb{E}(R^2 \pi H) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} r^2 \pi h f_H(h) f_r(r) dr dh = 9621.127 \text{ cm}^3. \quad \square$$

Prisjetimo se, za slučajne varijable X i Y (definirane na istom vjerojatnosnom prostoru) vrijedi $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$. Ako su X i Y još i nezavisne, onda je $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$ (što lagano možemo zaključiti iz donjeg računa). Što ako X i Y nisu nezavisne? Imamo

$$\begin{aligned}
\text{Var}(X+Y) &= \mathbb{E}((X+Y-\mathbb{E}(X+Y))^2) \\
&= \mathbb{E}(((X-\mathbb{E}(X)) + (Y-\mathbb{E}(Y)))^2) \\
&= \mathbb{E}((X-\mathbb{E}(X))^2 + (Y-\mathbb{E}(Y))^2 + 2(X-\mathbb{E}(X))(Y-\mathbb{E}(Y))) \\
&= \text{Var}(X) + \text{Var}(Y) + 2\mathbb{E}((X-\mathbb{E}(X))(Y-\mathbb{E}(Y))).
\end{aligned}$$

Izraz $\mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y)))$, u oznaci $\text{Cov}(X, Y)$, nazivamo **kovarijancom** od X i Y. Uočimo

$$\text{Cov}(X,Y) = \mathbb{E}(XY - \mathbb{E}(X)Y - \mathbb{E}(Y)X + \mathbb{E}(X)\mathbb{E}(Y)) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$

Lagano se vidi da je u slučaju nezavisnosti $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$, iz čega slijedi da je $\mathrm{Cov}(X,Y) = 0$. Ako je $\mathrm{Cov}(X,Y) > 0$ kažemo da su X i Y **pozitivno korelirane**, za $\mathrm{Cov}(X,Y) < 0$ su **negativno korelirane**, a za $\mathrm{Cov}(X,Y) = 0$ su **nekorelirane**. Pozitivna koreliranost znači da X i Y imaju sklonost odstupanja od očekivanja u istu stranu, a u slučaju negativne koreliranosti imaju sklonost odstupanja na različite strane od očekivanja. Napomenimo da nekoreliranost ne znači nezavisnost.

Primjer 5.4. Neka je $X \sim \begin{pmatrix} -1 & 0 & 1 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix}$. Tada imamo $X^2 \sim \begin{pmatrix} 0 & 1 \\ \frac{1}{3} & \frac{2}{3} \end{pmatrix}$ i $X^3 \sim \begin{pmatrix} -1 & 0 & 1 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix}$. Dakle, $\mathrm{Cov}(X, X^2) = \mathbb{E}(X^3) - \mathbb{E}(X)\mathbb{E}(X^2) = 0$ pa su X i X^2 nekorelirane. S druge strane, vrijedi

$$\mathbb{P}(X = 0, X^2 = 0) = \mathbb{P}(X = 0) = \frac{1}{3} \neq \mathbb{P}(X = 0)\mathbb{P}(X^2 = 0) = \frac{1}{9}$$

pa X i X^2 nisu nezavisne. □

Vrijedi sljedeće:

$$\mathrm{Cov}(rX + s, tY + u) = rt\,\mathrm{Cov}(X,Y), \quad r,s,t,u \in \mathbb{R}.$$

Dakle, kovarijanca je osjetljiva na linearne transformacije. Problem se rješava definiranjem

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma(X)\sigma(Y)}.$$

Broj $\rho(X,Y)$ nazivamo **koeficijent korelacije**. Sada je $\rho(yX+s, tY+u) = \rho(X,Y)$. Uočimo da $\rho(X,Y)$ poprima vrijednosti između -1 i 1, a značenje predznaka je isto kao kod kovarijance. Zaista, može se pokazati da je relacijom

$$\langle X, Y\rangle = \mathbb{E}(XY)$$

definiran skalarni produkt na skupu svih slučajnih varijabli (defiranih na istom vjerojatnosnom prostoru) koje imaju varijancu. Dakle, kao i za svaki skalarni produkt, vrijedi Cauchy-Schwartz-Bunjakovski nejednakost:

$$|\langle X, Y\rangle| = |\mathbb{E}(XY)| \leqslant \sqrt{|\langle X, X\rangle|}\sqrt{|\langle Y, Y\rangle|} = \sqrt{\mathbb{E}(X^2)}\sqrt{\mathbb{E}(Y^2)},$$

(vidi [9, str. 314]). Sada imamo

$$\rho(X,Y) = \frac{\mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y)))}{\sqrt{\mathbb{E}((X - \mathbb{E}(X))^2)}\sqrt{\mathbb{E}((Y - \mathbb{E}(Y))^2)}},$$

što pokazuje tvrdnju. Nadalje, uočimo da kovarijancom zapravo mjerimo stupanj linearne zavisnosti dviju slučajnih varijabli. Naime, nejednakost Cauchy-Schwartz-Bunjakovski postaje jednakost ako, i samo ako, su slučajne varijable X i Y kolinearne s vjerojatnošću 1, tj. ako postoji $\lambda \in \mathbb{R}$ t.d.

$$\mathbb{P}(\{\omega \in \Omega : Y(\omega) = \lambda X(\omega)\}) = 1$$

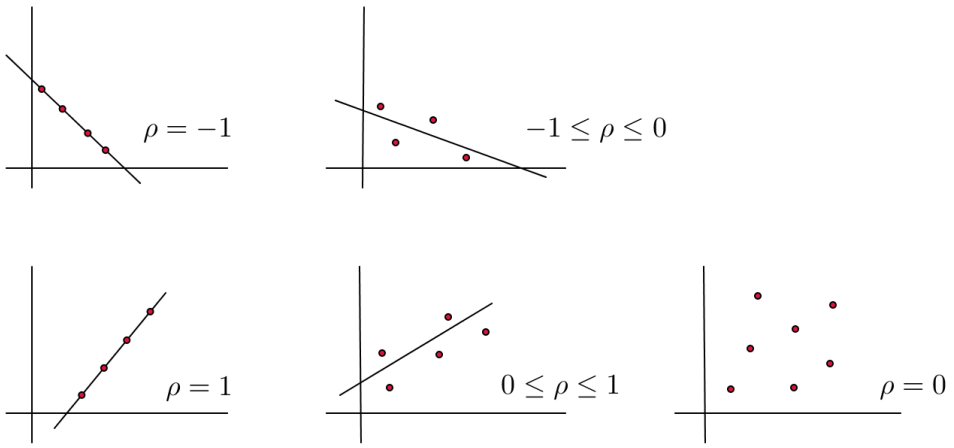
(vidi [9, str. 315]). Sada zaključujemo

(i) $|\rho(X,Y)| = 1$ ako, i samo ako, $\sigma(X) > 0$, $\sigma(Y) > 0$ i postoje $a, b \in \mathbb{R}$ t.d. $Y = aX + b$ s vjerojatnošću 1 (uočimo da je nužno $a \neq 0$)

(ii) manji $|\rho(X,Y)|$ znači "manji stupanj linearne zavisnosti" od X i Y

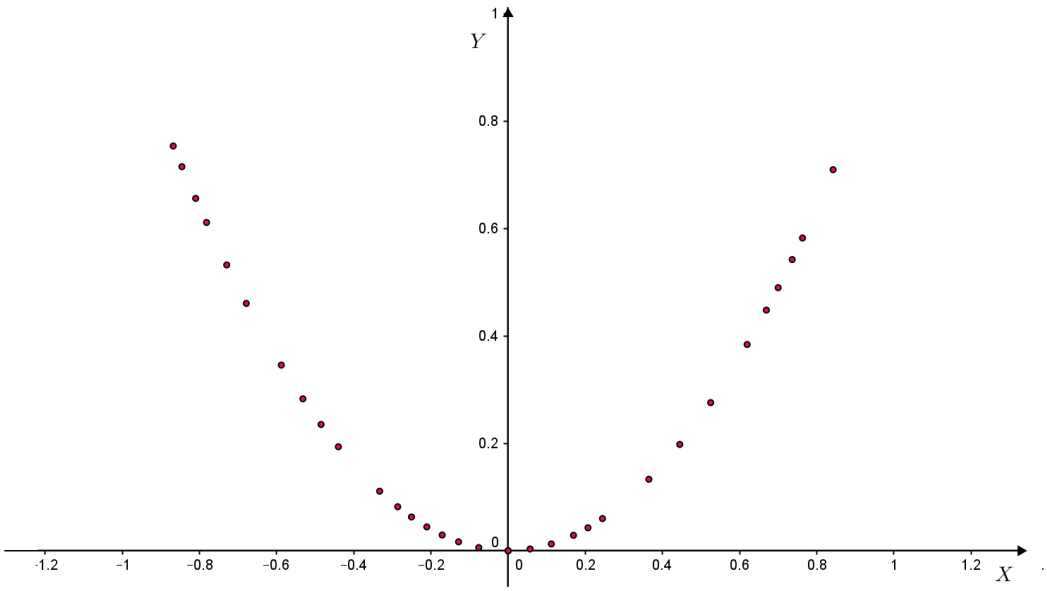
(iii) $\rho(X,Y) = 0$ znači da su $X - \mathbb{E}(X)$ i $Y - \mathbb{E}(Y)$ "ortogonalne".

Na Slici 9.1 prikazani su dijagrami raspršenja za različite koeficijente korelacije.



Slika 5.2: Parovi slučajnih varijabli različitih koeficijenata korelacije

Međutim, kao što i Primjer 5.4 pokazuje (gdje je $Y = X^2$), ne mora uvijek posrijedi biti linearna zavisnost (vidi Sliku 5.3).



Slika 5.3: Zavisnost slučajnih varijabli $Y = X^2$ koja nije linearna

Potpuna korelacija postoji kada svakoj vrijednosti od X odgovara samo jedna vrijednost od Y. Djelomična korelacija znači da vrijednosti prve varijable odgovara više vrijednosti druge. Korelacija je manja što ima više različitih vrijednosti od Y koje vežemo uz određenu vrijednost od X. Sada želimo odrediti tu funkcijsku zavisnost, tj. tražimo funkciju $g : \mathbb{R} \to \mathbb{R}$ tako da Y i $g(X)$ imaju "približno" jednaku raspodjelu. Ideja je $g(x)$ uzeti tako da $\mathbb{E}((Y - g(X))^2)$ bude minimalno. U slučaju kad $g(x)$ tražimo u klasi linearnih funkcija, dobivamo

$$g(x) = \rho(X, Y)\frac{\sigma(Y)}{\sigma(X)}(x - \mathbb{E}(X)) + \mathbb{E}(Y).$$

Ako želimo da X i $g(Y)$ imaju "približno" jednaku raspodjelu i $g(y)$ tražimo u klasi linearnih funkcija, onda imamo

$$g(y) = \rho(X, Y)\frac{\sigma(X)}{\sigma(Y)}(y - \mathbb{E}(Y)) + \mathbb{E}(X)$$

(vidi [9, str. 141]). Te pravce nazivamo **pravcima regresije** i oni daju najbolju aproksimaciju slučajne varijable Y linearnom funkcijom slučajne varijable X i najbolju aproksimaciju slučajne varijable X linearnom funkcijom slučajne varijable Y. Specijalno, kada su X i Y nekorelirane, onda dobijemo $g(x) = \mathbb{E}(Y)$ i $g(y) = \mathbb{E}(X)$.

Primjer 5.5. Dvodimenzionalni diskretni slučajni vektor (X, Y) dan je shemom

$$\begin{pmatrix} X\backslash Y & 0 & 1 & 2 \\ -2 & 0.05 & 0.1 & 0.03 \\ -1 & 0.05 & 0.05 & 0.12 \\ 0 & 0.1 & 0.05 & 0.07 \\ 1 & 0 & 0.1 & 0.06 \\ 2 & 0.05 & 0 & 0.03 \\ 3 & 0.05 & 0.05 & 0.04 \end{pmatrix}.$$

Odredimo pripadne pravce regresije. Iz gornje sheme jednostavno iščitamo

$$X \sim \begin{pmatrix} -2 & -1 & 0 & 1 & 2 & 3 \\ 0.18 & 0.22 & 0.22 & 0.16 & 0.08 & 0.14 \end{pmatrix} \quad \text{i} \quad Y \sim \begin{pmatrix} 0 & 1 & 2 \\ 0.3 & 0.35 & 0.35 \end{pmatrix}.$$

Sada izračunamo $\mathbb{E}(X) = 0.16$, $\mathbb{E}(Y) = 1.05$, $\mathbb{E}(X^2) = 2.68$ i $\mathbb{E}(Y^2) = 1.75$, što daje $\sigma(X) = 1.629$ i $\sigma(Y) = 0.804$. Nadalje, imamo $\mathbb{E}(XY) = 0.12$, što u konačnici daje $\rho(X, Y) = -0.036$, odnosno

$$g(x) = -0.018x + 1.052 \qquad \text{i} \qquad g(y) = -0.074y + 0.237. \qquad \square$$

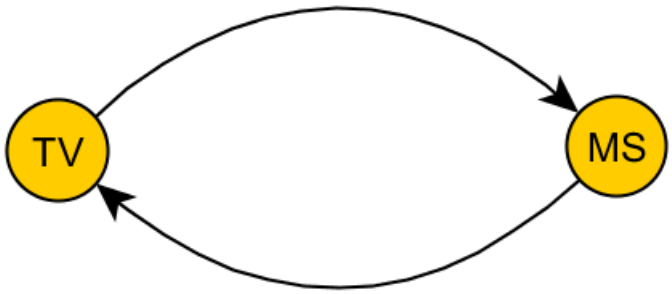
Dio II

Matematička statistika

Poglavlje 6

Statistika

Matematički model slučajnog pokusa dan je vjerojatnosnim prostorom $(\Omega, \mathcal{F}, \mathbb{P})$. Kao što smo već rekli, problem je što u pravilu ne znamo mjeru neizvjesnosti $\mathbb{P} : \mathcal{F} \to [0, 1]$, tj. nemamo "recept" kako ju odrediti, već samo znamo svojstva koja ona mora zadovoljavati. Dakle, da bismo vjerodostojno modelirali funkciju $\mathbb{P} : \mathcal{F} \to [0, 1]$, trebamo dobro poznavati prirodu fenomena kojeg proučavamo. U tu svrhu pomaže nam matematička disciplina koju nazivamo matematička statistika ili samo statistika. Ukratko, vjerojatnost daje informaciju o (ne)izvjesnosti pojedinog događaja vezanog za slučajni pokus, tj. informaciju o ishodu pokusa, a statistika na osnovu ishoda pokusa pokušava procijeniti mjeru (ne)izvjesnost događaja, tj. pokušava dati kvalitetnu (dobru) procjenu vjerojatnosti $\mathbb{P} : \mathcal{F} \to [0, 1]$ vezanu za promatrani slučajni pokus.



Slika 6.1: Odnos teorije vjeorjatnosti i matematičke statistike

Statistika je skup metoda koje se koriste za prikupljanje, analizu, prikaz i interpretaciju podataka te za donošenje odluka. Statistika se dijeli na:

(i) **dizajn eksperimenta** – bavi se planiranjem eksperimenata i sakupljanjem podataka. Primjerice, u mnogim područjima znanosti eksperi-

menti su skupi te je unaprijed potrebno odrediti tip i količinu potrebnih podataka.

(ii) **deskriptivna statistika** – bavi se organizacijom, predočavanjem i opisivanjem sakupljenih podataka (tablice, grafikoni i mjere deskriptivne statistike).

(iii) **inferencijalna statistika** – bavi se vrednovanjem infomacija sadržanih u podacima i ocjenom novog znanja dobivenog iz tih podataka (procjena parametara i testiranje statističkih hipoteza) s ciljem donošenja utemeljenih zaključaka vezanih za promatrani slučajni pokus.

Sljedeći primjeri ilustriraju neke od problema kojima se bavi statistika.

Primjer 6.1. Bacamo novčić 100 puta i bilježimo rezultat. Kao rezultat dobivamo 60 glava i 40 pisama. Sumnjamo da je novčić neispravan i želimo testirati tu pretpostavku. □

Primjer 6.2. U jednom istraživanju izmjerene su gustoće i tlačne čvrstoće 19 betonskih kocki različitih (standardiziranih) tipova betona (Mattaacchione i Mattacchione, "Correlation Between 28-Day Strength and Density", *Concrete International*, 1995 (3), str. 37-41) i dobiveni podaci prikazani su u tablici 6.1.

Gustoća (kg/m^3)	Tlačna čvrstoća (MPa)
2236	25.2
2244	25.2
2244	25.2
2244	25.3
2244	25.7
2253	25.4
2253	25.8
2262	25.4
2272	27.5
2280	27.5
2290	27.1
2290	28.1
2290	28.8
2295	27.1
2295	29.7
2307	29.9
2315	29.5
2325	29.3
2334	28.7

Tablica 6.1: Podaci o gustoći i tlačnoj čvrstoći betona

Bez nekakve duboke analize možemo zaključiti da podaci mjerenja sugeriraju zavisnost gustoće i tlačne čvrstoće betona. Štoviše, sugeriraju da beton veće gustoće ima i veću tlačnu čvrstoću, što je i za očekivati. Postavlja se pitanje koji je tip i stupanj zavisnosti posrijedi. □

U gornjim primjerima prvo je bilo potrebno sakupiti podatke (pokusom, popisom, promatranjem). Međutim, u većini slučajeva, fizički ili praktično, neizvedivo je sakupiti iscrpni i sveobuhvatni skup podataka. Primjerice, koliko god podataka sakupili eksperimentiranjem, u principu je moguće eksperiment i dalje ponavljati i dobivati nove podatke. Skup svih subjekata ili objekata koji se žele obuhvatiti istraživanjem naziva se **populacija**. Dio populacije odabran za proučavanje naziva se **uzorak**. **Element** populacije ili uzorka je subjekt ili objekt o kojem se prikuplja informacija. Zadatak statistike je ograničiti proučavanje na neki uzorak te donositi globalne zaključke (za cijelu populaciju) na osnovu zaključaka dobivenih analiziranjem uzorka. Uočimo da su zaključci izvedeni statističkom analizom nesigurni jer se zasnivaju na uzorku. Primjerice, zaključak o vezi gustoće i tlačne čvrstotće

betona donosimo na osnovu uzorka od 19 betonskih kocki različitih (standardiziranih) tipova betona. **Varijabla** je svojstvo koje se proučava i koje poprima razne vrijednosti iz skupa dozvoljenih vrijednosti za razne elemente. Ovdje još ne govorimo o slučajnim varijablama jer pokus koji promatramo ne mora imati slučajni karakter. **Skup podataka** je kolekcija opažanja jedne ili više varijabli. Napomenimo još da se matematička statistika bavi **slučajnim uzorcima**, tj. uzorcima kod kojih je svaki element populacije imao šansu biti izabran. U neslučajnom uzorku neki elementi populacije nemaju nikakvu šansu biti izabrani. **Jednostavni slučajni uzorak** je uzorak pri čijem je izboru svaki član populacije imao iste izglede biti uključen i nezavisno je biran od ostalih članova. U Primjeru 6.1 populacija su sva bacanja novčića (beskonačan skup), uzorak je prvih 100 bacanja i pripadna varijabla označava ishod (pismo/glava) pojedinog bacanja. U Primjeru 6.2 populacija su svi mogući (standardizirani) tipovi betona, uzorak je odabranih 19 tipova betona i (dvije) pripadne varijable označavaju iznos gustoće i tlačne čvrstoće pojedinog tipa betona.

Zadatak statistike, među ostalim, je donijeti zaključak o populaciji na temelju uzorka uz određenu, unaprijed zadanu, toleranciju pogreške da smo pogriješili. Kao i u teoriji vjerojatnosti, fokus statistike su varijable. Varijable koje proučavamo mogu biti **kvantitativne** (numeričke) i **kvalitativne** (kategorijalne). Kvantitativna varijabla poprima brojčanu vrijednost i dijeli se na:

(i) **diskretne** (npr. predstavlja neko prebrojavanje)

(ii) **neprekidne** (npr. rezultat mjerenja neke fizikalne veličine).

Kvalitativna varijabla ne poprima brojčane vrijednosti, nego može biti razvrstana u dvije ili više nebrojčanih **kategorija** (npr. spol, mjesto rođenja, zanimanje). Napomenimo da te kategorije mogu biti označene brojevima, no to varijablu ne čini numeričkom (npr. spol možemo označavati kao: 0 - muški spol i 1 - ženski spol). Varijable također možemo razlikovati i obzirom na **skalu mjerenja**:

(i) **nominalna** (vrijednosti su neuređene, npr. zanimanje)

(ii) **ordinalna** (vrijednosti su uređene, npr. školske ocjene: odličan, vrlo dobar, dobar, dovoljan, nedovoljan)

(iii) **intervalna** (vrijednosti su uređene i za njih se može izračunati razlika, npr. IQ)

(iv) **omjerna** (vrijednosti su uređene i za njih imaju smisla sve osnovne aritmetičke operacije, npr. masa, cijena, zarada).

Poglavlje 7

Deskriptivna statistika

Izvodimo eksperiment i bilježimo realizacije varijable X kojom modeliramo neko promatrano obilježje na uzorku iz populacije od interesa. Rezultat opažanja varijable X na elementu populacije označavamo s x. Opažene vrijednosti od X na uzorku veličine n označavamo s $x_1, \ldots, x_n$.

Primjer 7.1. Ocjene iz matematike jednog razreda od 30 učenika na kraju školske godine su:

$$1, 4, 2, 3, 1, 1, 2, 4, 3, 4, 5, 3, 2, 2, 3, 2, 5, 3, 2, 3, 3, 4, 2, 3, 2, 3, 3, 2, 2, 2.$$

Zanima nas ocjena iz matematike kao obilježje točno tog razreda. U ovom primjeru populacija i uzorak su učenici danog razreda, element je pojedini učenik, a varijabla X predstavlja ocjenu iz matematike pojedinog učenika na kraju školske godine. Dakle, X je diskretna kvantitativna (ordinalna) varijabla (slika od X je $R(X) = \{1, 2, 3, 4, 5\}$). □

Primjer 7.2. Na Građevinskom fakultetu Sveučilišta u Zagrebu provedena je anketa o županiji rođenja. Jedan uzorak daje sljedeće informacije:

student (OIB)	županija rođenja
12890567891	Zagrebačka
89734087637	Koprivničko-križevačka
23890651073	Splitsko-dalmatinska
28761034891	Koprivničko-križevačka
78902346529	Krapinsko-zagorska
29734509725	Koprivničko-križevačka
39475012784	Ličko-senjska

Tablica 7.1: Županije rođenja sedam studenata

Zanima nas informacija o županiji rođenja studenata kao obilježje točno tog fakulteta. U ovom primjeru populacija su svi studenti spomenutog fakulteta, uzorak čine gore odabranih sedam studenata, element populacije je pojedini student, a varijabla X predstavlja županiju rođenja pojedinog studenta. Uočimo da sliku od X čine sve županije u RH. Dakle, X je kvalitativna (nominalna) varijabla. □

Osnovna zadaća deskriptivne statistike jest opisivati opažene vrijednosti (na uzorku ili populaciji) od X. **Raspodjela** opaženih vrijednosti $x_1, \ldots, x_n$ od X (na uzorku ili populaciji veličine n) opisuje se frekvencijama, relativnim frekvencijama i kumulativnim frekvencijama te kumulativnim relativnim frekvencijama (u slučaju kvantitativnih varijabli). **Frekvencija** je broj pojavljivanja pojedinog x_i u nizu $x_1, \ldots, x_n$, a **relativna frekvencija** od x_i je omjer frekvencije te vrijednosti i veličine niza $x_1, \ldots, x_n$, tj. n. **Kumulativna frekvencija** u vrijednosti x_i (pri čemu je niz $x_1, \ldots, x_n$ poredan po veličini od najmanje do najveće vrijednosti) je suma frekvencija vrijednosti manjih ili jednakih od x_i. **Kumulativna relativna frekvencija** u vrijednosti x_i je omjer kumulativne frekvencije u x_i i veličine niza $x_1, \ldots, x_n$.

Primjerice, u prethodna dva primjera, redom, imamo:

a_i	f_i	$\frac{f_i}{30}$	K_{f_i}	$\frac{K_{f_i}}{30}$
1	3	0.1	3	0.1
2	11	0.37	14	0.47
3	10	0.33	24	0.8
4	4	0.13	28	0.93
5	2	0.07	30	1
$\sum$	30	1		

Tablica 7.2: Frekvencije, relativne frekvencije, kumulativne frekvencije i kumulativne relativne frekvencije ocjena u razredu iz Primjera 7.1

županije	f_i	$\frac{f_i}{7}$
Zagrebačka	1	0.143
Koprivničko-križevačka	3	0.428
Splitsko-dalmatinska	1	0.143
Krapinsko-zagorska	1	0.143
Ličko-senjska	1	0.143
$\sum$	7	1

Tablica 7.3: Frekvencije i relativne frekvencije podataka o županiji rođenja studenata iz Primjera 7.2

Neprekidne varijable, odnosno raspodjelu njihovih opaženih vrijednosti, diskutiramo na analogan način. Napomenimo samo da su zbog prirode problema vrijednosti neprekidnih varijabli, za razliku od diskretnih, dane u terminima intervala.

Primjer 7.3. Mjesečna zarada (u kunama) zaposlenih u nekom poduzeću je dana u Tablici 7.4.

Mjesečna zarada razredi	Broj radnika f_i	K_{f_i}	$\frac{f_i}{250}$	$\frac{K_{f_i}}{250}$
[1000,2000)	16	16	0.064	0.064
[2000,3000)	38	54	0.152	0.216
[3000,4000)	66	120	0.264	0.48
[4000,5000)	70	190	0.28	0.76
[5000,6000)	41	231	0.164	0.924
[6000,7000)	18	249	0.072	0.996
[7000,8000)	1	250	0.04	1
$\sum$	250		1	

Tablica 7.4: Frekvencije, kumulativne frekvencije, relativne frekvencije i kumulativne relativne frekvencije mjesečnih zarada po razredima

□

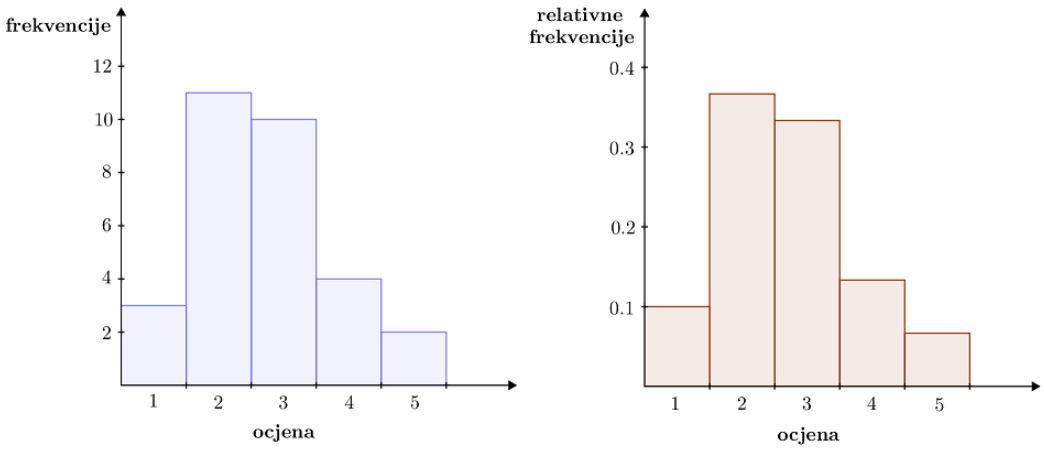
Napomenimo da diskretne kvantitativne varijable, tj. njihove opažene vrijednosti, mogu biti i grupirane. Grupiranje radimo kada to ima smisla u kontekstu problema koji proučavamo i prema kriteriju koji nam se u tom trenutku s obzirom na pitanja na koja želimo dobiti odgovor čini prikladan. Na primjer, broj automobila po kućanstvu u pravilu ne grupiramo (jer ne

baratamo velikim brojevima) dok starost stanovništva ponekad ima smisla grupirati jer, primjerice, nema velike razlike između 44 i 45 godina starosti. U slučaju da želimo razrede jednake duljine, veličinu razreda ili broj razreda biramo sami i određujemo iz formule

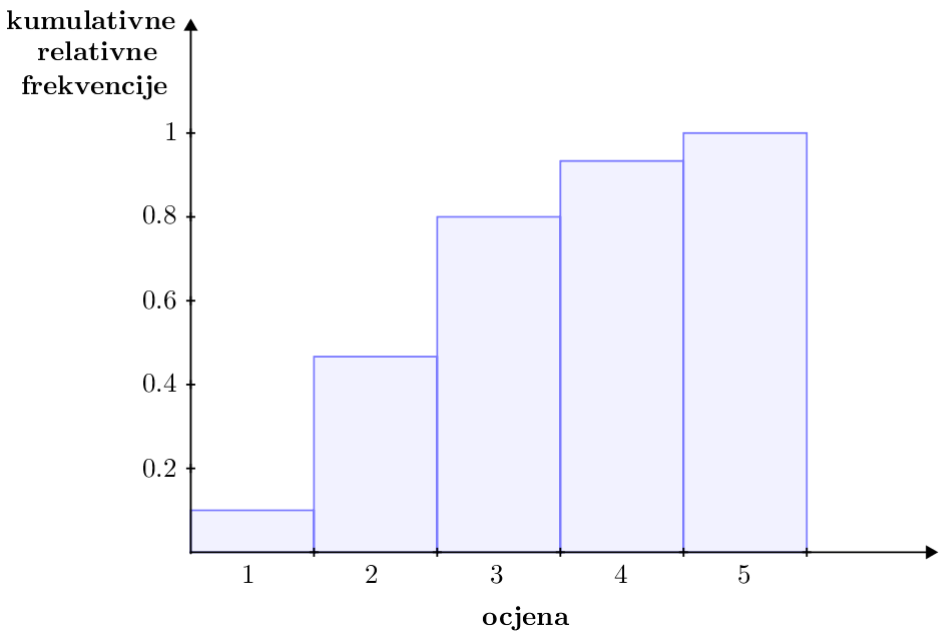
$$\text{širina razreda} = \frac{\max\{x_1, \ldots, x_n\} - \min\{x_1, \ldots, x_n\}}{\text{broj razreda}}.$$

Pojmovi frekvencije, relativne frekvencije, kumulativne frekvencije i kumulativne relativne frekvencije definiraju se analogno kao i kod neprekidnih varijabli.

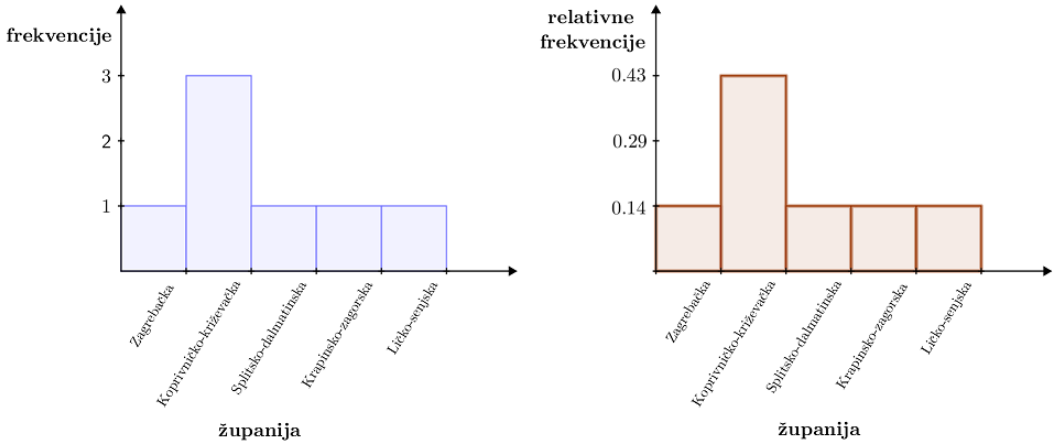
Raspodjelu opaženih vrijednosti od X prikazujemo grafički preko stupčastih dijagrama i histograma (frekvencija, relativnih frekvencija, kumulativnih frekvencija i relativnih kumulativnih frekvencija). **Stupčasti dijagram** je grafički prikaz (frekvencija, relativnih frekvencija, kumulativnih frekvencija i kumulativnih relativnim frekvencija) opaženih vrijednosti kvalitativnih varijabli i negrupiranih opaženih vrijednosti diskretnih kvantitativnih varijabli. **Histogram** je prikaz (frekvencija, relativnih frekvencija, kumulativnih frekvencija i kumulativnih relativnih frekvencija) opaženih vrijednosti neprekidnih kvantitativnih varijabli i grupiranih opaženih vrijednosti diskretnih kvantitativnih varijabli. Napomenimo da "stupići" u histogramu smiju biti različite širine. U gornjim primjerima imamo stupčaste dijagrame i histograme kao na sljedećim slikama (od Slike 7.1 do Slike 7.6).



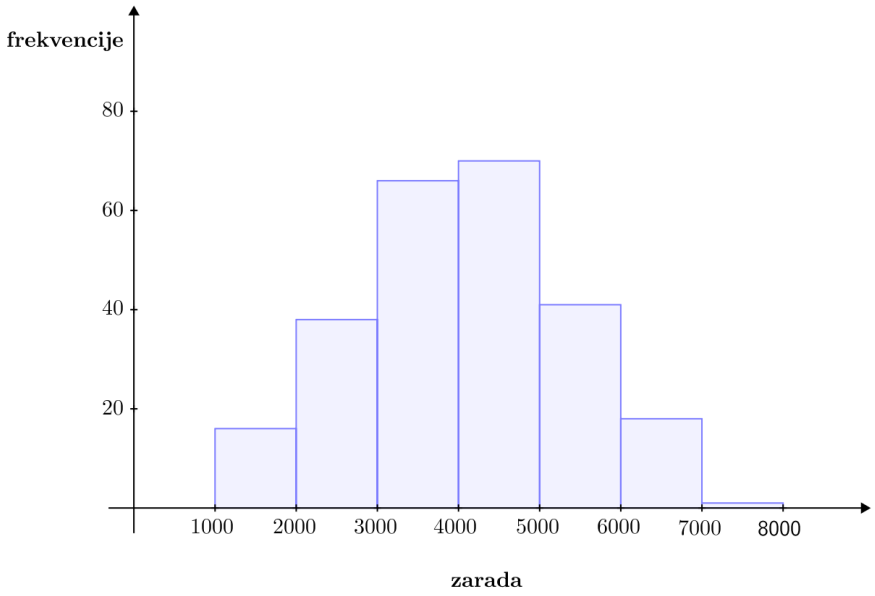
Slika 7.1: Stupčasti dijagrami frekvencija i relativnih frekvencija ocjena iz matematike iz Primjera 7.1



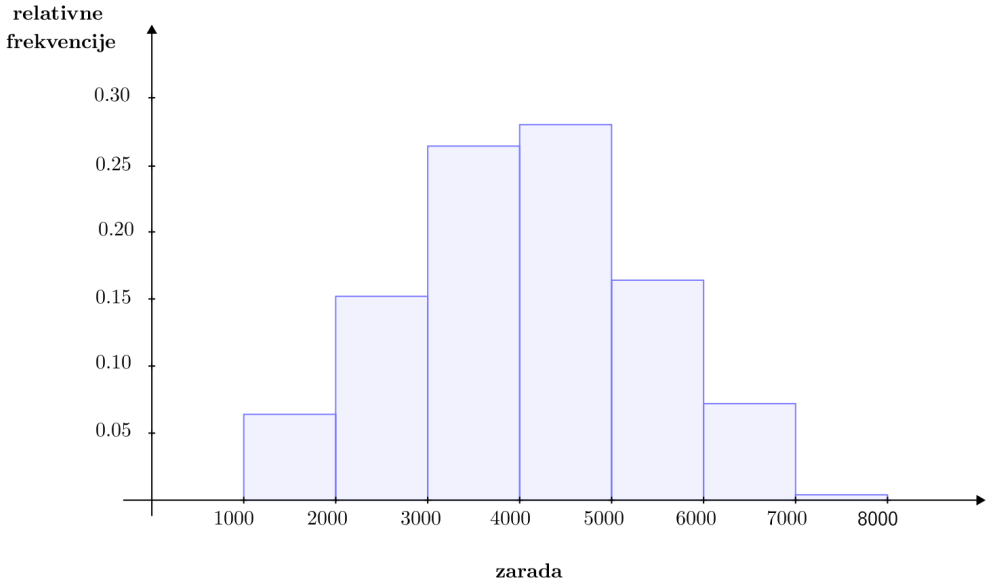
Slika 7.2: Stupčasti dijagram kumulativnih relativnih frekvencija ocjena iz matematike iz Primjera 7.1



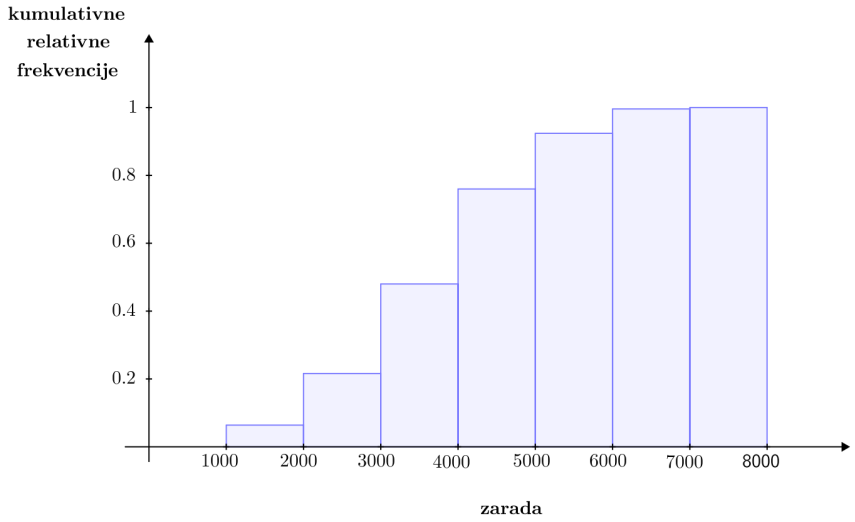
Slika 7.3: Stupčasti dijagrami frekvencija i relativnih frekvencija županija studenata iz Primjera 7.2



Slika 7.4: Histogram frekvencija plaća po razredima iz Primjera 7.3

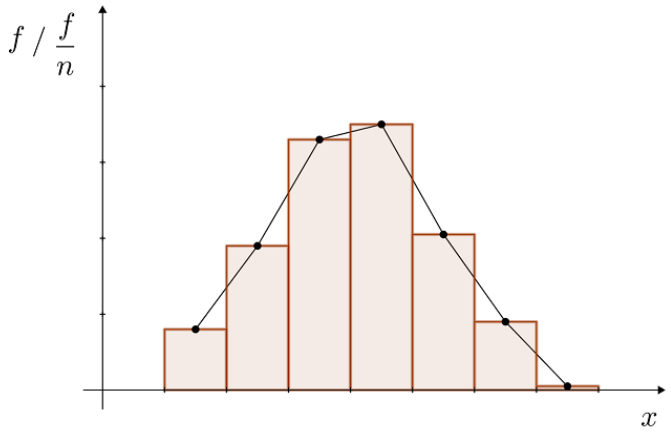


Slika 7.5: Histogram relativnih frekvencija plaća po razredima iz primjera 7.3



Slika 7.6: Histogram kumulativnih relativnih frekvencija plaća po razredima iz Primjera 7.3

Ako spojimo sredine gornjih stranica uzastopnih pravokutnika u histogramima ravnim crtama, dobivamo **poligon**. Dobivena krivulja naziva se još i **frekvencijskom krivuljom**, **krivuljom relativnih frekvencija**, **krivuljom kumulativnih frekvencija** i **krivuljom kumulativnih relativnih frekvencija**. Na Slici 7.7 je primjer histograma s frekvencijskom krivuljom.



Slika 7.7: Histogram s frekvencijskom krivuljom

U nastavku usredotočit ćemo se na deskriptivnu statistiku kvantitativnih varijabli. Analiziramo numeričke deskriptivne mjere:

(i) **mjere centralne tendencije**

(ii) **mjere raspršenja**

(iii) **mjere oblika**.

Neka je X kvantitativna varijabla s vrijednostima opažanja $x_1, \ldots, x_n$ na nekom uzorku veličine n. Pretpostavimo da su vrijednosti $x_1, \ldots, x_n$ poredane po veličini od najmanje do najveće. U slučaju da je X diskretna kvantitativna varijabla čije su vrijednosti grupirane u n razreda ili neprekidna varijabla dana s n razreda (intervala), onda za vrijednosti $x_1, \ldots, x_n$ uzimamo sredine razreda.

7.1 Mjere centralne tendencije

Aritmetičku sredinu vrijednosti $x_1, \ldots, x_n$ računamo kao

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

Ako se među brojevima $x_1, \ldots, x_n$ pojavljuju brojevi $a_1, \ldots, a_k$, $k < n$, s frekvencijama $f_1, \ldots, f_k$, onda je

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

Napomenimo da je aritmetička sredina osjetljiva na stršeće vrijednosti, tzv. **izdvojenice** (engl. *outliere*). Izdvojenice se najčešće javljaju kao posljedica krivog mjerenja ili je podatak točno izmjeren ali predstavlja rijetku pojavu ili dolazi iz neke druge populacije. Također, napomenimo da je $\overline{x}_n$ jedinstveni broj u kojem funkcija

$$f(x) = \sum_{i=1}^{n} (x_i - x)^2$$

postiže svoj minimum. Dokaz ove tvrdnje slijedi na potpuno analogan način kao i u slučaju odnosa očekivanja i varijance slučajne varijable.

Medijan vrijednosti $x_1, \ldots, x_n$ je vrijednost za koju vrijedi da je 50% podataka manje ili jednako toj vrijednosti, a 50% podataka je veće ili jednako navedenoj vrijednosti. Ako je broj podataka neparan, medijan je podatak na

centralnoj (srednjoj) poziciji. Ako je broj podataka paran, onda je medijan prosjek vrijednosti dva srednja člana:

$$\mathrm{M_e} = \begin{cases} x_{\lfloor \frac{n}{2}+1 \rfloor}, & \frac{n}{2} \text{ nije prirodan broj} \\ \dfrac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}, & \frac{n}{2} \text{ je prirodan broj.} \end{cases}$$

Primjer 7.4. U nizu $1, 7, 9, 14, 18$ medijan je $\mathrm{M_e} = 9$, a u nizu $1, 9, 12, 15, 22, 23$ medijan je $\mathrm{M_e} = (12 + 15)/2 = 13.5$. □

Napomenimo da je $\mathrm{M_e}$ broj u kojem funkcija

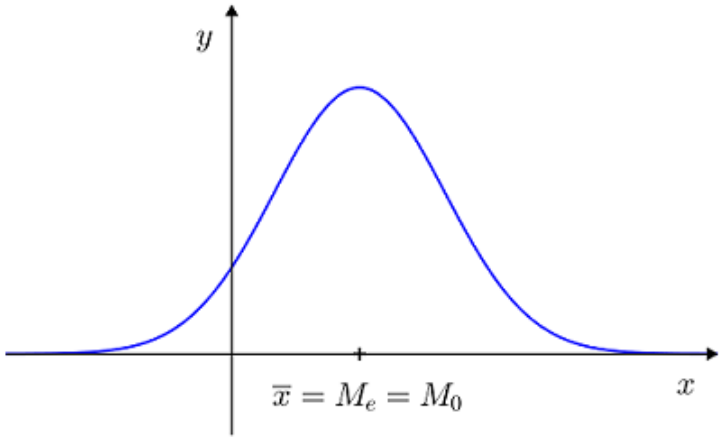
$$f(x) = \sum_{i=1}^{n} |x_i - x|$$

postiže svoj minimum. Ako je n neparan, lagano je za vidjeti da je $\mathrm{M_e}$ nužno jedinstveni minimum od $f(x)$. Ako je n paran, nije teško vidjeti da $f(x)$ nužno postiže minimum na $[x_{\frac{n}{2}}, x_{\frac{n}{2}+1}]$ te da je konstantna na $[x_{\frac{n}{2}}, x_{\frac{n}{2}+1}]$. Samim time $\mathrm{M_e}$ je (nejedinistveni) minimum od $f(x)$.

Mod vrijednosti $x_1, \ldots, x_n$ je vrijednost s najvećom frekvencijom. Varijabla može imati i više modova, što pokazujemo u Primjeru 7.5.

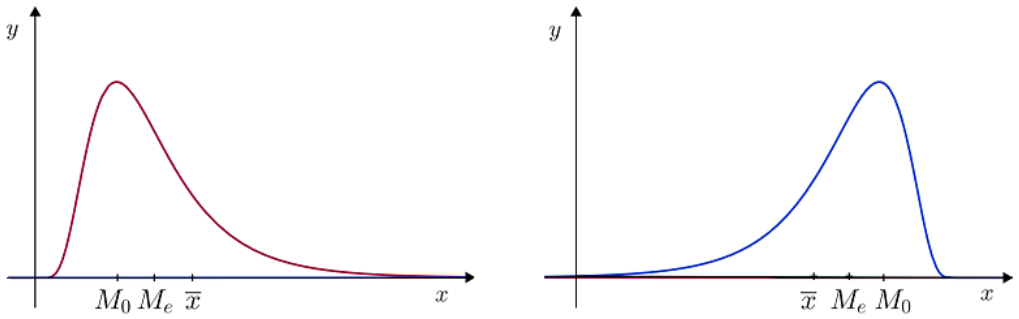
Primjer 7.5. U nizu $1, 1, 2, 3, 4$ mod je $\mathrm{M_0} = 1$, a u nizu $1, 1, 2, 3, 3, 4$ modovi su $\mathrm{M_0} = 1, 3$. □

Uočimo da mod i medijan nisu osjetljivi na izdvojenice, za razliku od aritmetičke sredine. Nadalje, uočimo da u slučaju simetričnih ($x_i = -x_{n-i+1}$ za $i = 1, \ldots, \lfloor n/2 \rfloor$) i unimodalnih ($x_1, \ldots, x_n$ imaju samo jedan mod) vrijednosti, aritmetička sredina, mod i medijan se podudaraju (vidi Sliku 7.8).



Slika 7.8: Odnos aritmetičke sredine, medijana i moda u slučaju simetričnih unimodalnih vrijednosti

U slučaju asimetričnih unimodalnih vrijednosti (vrijednosti kod kojih nije zadovoljena gornja relacija simetričnosti), medijan je uvijek između aritmetičke sredine i moda. Mod se nalazi na mjestu gdje je frekvencijska krivulja najviša, a aritmetička sredina je uvijek na strani na kojoj se nalazi dulji rep krivulje (vidi Sliku 7.9).



Slika 7.9: Odnos aritmetičke sredine, medijana i moda u slučaju asimetričnih unimodalnih vrijednosti

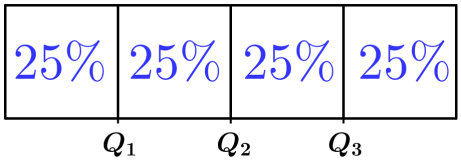
7.2 Mjere raspršenja

Kvartili vrijednosti $x_1, \ldots, x_n$ su vrijednosti koje dijele podatke u četiri jednakobrojna dijela. Računamo ih po principu sličnom računanju medijana. Primijetimo odmah da je drugi kvartil vrijednost za koju vrijedi da je 50% podataka manje ili jednako toj vrijednosti, a 50% podataka veće ili jednako navedenoj vrijednosti, odnosno $Q_2 = \mathrm{M_e}$. Prvi i treći kvartil dobivamo na sljedeći način:

$$Q_1 = \begin{cases} x_{\lfloor \frac{n}{4}+1 \rfloor}, & \frac{n}{4} \text{ nije prirodan broj} \\ \dfrac{x_{\frac{n}{4}} + x_{\frac{n}{4}+1}}{2}, & \frac{n}{4} \text{ je prirodan broj.} \end{cases}$$

$$Q_3 = \begin{cases} x_{\lfloor \frac{3n}{4}+1 \rfloor}, & \frac{3n}{4} \text{ nije prirodan broj} \\ \dfrac{x_{\frac{3n}{4}} + x_{\frac{3n}{4}+1}}{2}, & \frac{3n}{4} \text{ je prirodan broj.} \end{cases}$$

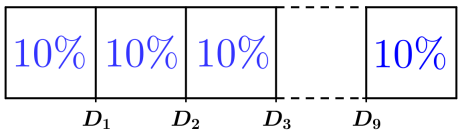
Kvartili su ilustrirani na Slici 7.10.



Slika 7.10: Kvartili

Analogno se definiraju **decili** (vidi Sliku 7.11), tj. za $i = 1, 2, \ldots, 9$ vrijedi:

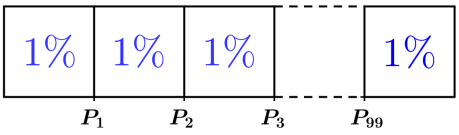
$$D_i = \begin{cases} x_{\lfloor \frac{in}{10}+1 \rfloor}, & \frac{in}{10} \text{ nije prirodan broj} \\ \dfrac{x_{\frac{in}{10}} + x_{\frac{in}{10}+1}}{2}, & \frac{in}{10} \text{ je prirodan broj.} \end{cases}$$



Slika 7.11: Decili

Na jednak način definiramo i **percentile**, vrijednosti koje dijele podatke na sto jednakobrojnih dijelova (vidi Sliku 7.12). Za $i = 1, 2, \ldots, 99$ vrijedi:

$$P_i = \begin{cases} x_{\lfloor \frac{in}{100}+1 \rfloor}, & \frac{in}{100} \text{ nije prirodan broj} \\ \frac{x_{\frac{in}{100}} + x_{\frac{in}{100}+1}}{2}, & \frac{in}{100} \text{ je prirodan broj.} \end{cases}$$



Slika 7.12: Percentili

Primjer 7.6. Na prvom zimskom ispitnom roku, pismenom ispitu iz kolegija Vjerojatnost i statistika pristupilo je 24 studenata. Broj bodova (od maksimalno mogućih 100) koje su studenti ostvarili na ispitu je sljedeći:

$$53, 72, 82, 45, 36, 88, 20, 31, 81, 68, 75, 58, 24, 67, 54, 93, 98, 70, 30, 5, 34, 7, 2, 61.$$

Da bismo mogli primijeniti gornje formule prvo poredajmo podatke po veličini (od najmanjeg do najvećeg):

$$2, 5, 7, 20, 24, 30, 31, 34, 36, 45, 53, 54, 58, 61, 67, 68, 70, 72, 75, 81, 82, 88, 93, 98.$$

Izračunajmo prvi i treći kvartil, prvi i deveti decil te prokomentirajmo neke od dobivenih rezultata. Budući da su $n/4 = 6$ i $3n/4 = 18$ prirodni brojevi, prvi i treći kvartil računamo na sljedeći način:

$$Q_1 = \frac{x_{\frac{24}{4}} + x_{\frac{24}{4}+1}}{2} = \frac{x_6 + x_7}{2} = \frac{30 + 31}{2} = 30.5,$$

$$Q_3 = \frac{x_{\frac{3\cdot 24}{4}} + x_{\frac{3\cdot 24}{4}+1}}{2} = \frac{x_{18} + x_{19}}{2} = \frac{72 + 75}{2} = 73.5.$$

Vrijednost prvog kvartila nam govori da je 25% studenata na završnom ispitu iz Statistike ostvarilo 30.5 bodova ili manje, dok je 75% studenata ostvarilo 30.5 bodova ili više. Da bismo izračunali prvi i deveti decil podataka potrebno je provjeriti jesu li $n/10$ i $9n/10$ prirodni brojevi. Budući da je $n/10 = 2.4$ i $9n/10 = 21.6$, decile računamo na sljedeći način:

$$D_1 = x_{\lfloor \frac{24}{10}+1 \rfloor} = x_{\lfloor 3.4 \rfloor} = x_3 = 7,$$

$$D_9 = x_{\left\lfloor \frac{9\cdot 24}{10}+1 \right\rfloor} = x_{\lfloor 22.6 \rfloor} = x_{22} = 88.$$

Vrijednost devetog decila nam govori da je 90% studenata na završnom ispitu dobilo 88 bodova ili manje, dok je najboljih 10% studenata ostvarilo 88 bodova ili više. □

Raspon vrijednosti $x_1, \ldots, x_n$ je razlika najveće i najmanje vrijednosti: $R = x_n - x_1$. **Interkvartilni raspon** vrijednosti $x_1, \ldots, x_n$ je razlika između trećeg i prvog kvartila: $I_Q = Q_3 - Q_1$. Interkvartilni raspon nam daje raspon u kojem se nalazi srednjih 50% podataka. **Uzoračka varijanca** vrijednosti $x_1, \ldots, x_n$ daje nam približno srednje kvadratno odstupanje podataka od aritmetičke sredine:

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

ili, kada su podaci dani svojim frekvencijama,

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

Uzoračka standardna devijacija vrijednosti $x_1, \ldots, x_n$, koja daje približno srednje odstupanje podataka od $\overline{x}_n$, dana je sa $s_n = \sqrt{s_n^2}$. Napomenimo ovdje da bi na prvi pogled bilo puno prirodnije definirati uzoračku varijancu s faktorom $1/n$, nego s faktorom $1/(n-1)$. Razlog tome je dan u sljedećem poglavlju.

Ako imamo dva uzorka ili čak dva eksperimenta i ako želimo usporediti njihove aritmetičke sredine i varijance (standardne devijacije), koristimo **koeficijent varijacije**:

$$V = \frac{s_n}{\overline{x}_n}.$$

Kao i u teoriji vjerojatnosti (uz analogan dokaz), i ovdje vrijedi Čebiševljeva nejednakost: za proizvoljan $a > 0$, broj vrijednosti $x_1, \ldots, x_n$ za koje je $|x_i - \overline{x}_n| \geqslant a$ je manji ili jednak od $(n-1)s_n^2/a^2$. Specijalno, za $a = ks_n$, $k \in \mathbb{N}$, imamo:

$$\frac{\#\{i = 1, \ldots, n : |x_i - \overline{x}_n| \geqslant a\}}{n-1} \leqslant \frac{1}{k^2},$$

odnosno

$$\frac{\#\{i = 1, \ldots, n : |x_i - \overline{x}_n| < a\}}{n-1} \geqslant 1 - \frac{1}{k^2}.$$

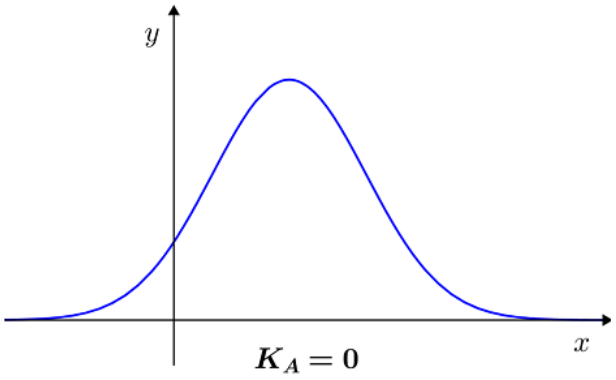
Primjerice, za $k = 3$, otprilike 89% vrijednosti se nalazi u intervalu $[\overline{x}_n - 3s_n, \overline{x}_n + 3s_n]$.

7.3 Mjere oblika

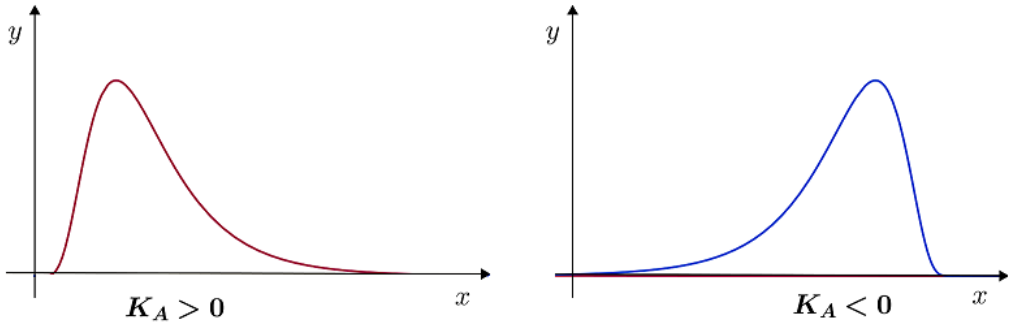
Koeficijent asimetrije vrijednosti $x_1, \ldots, x_n$ daje nam podatak o simetričnosti pripadne frekvencijske krivulje. Računa se po sljedećoj formuli:

$$\mathrm{K_A} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_n)^3}{s_n^3} = \frac{\frac{1}{n}\sum_{i=1}^{k} f_i(a_i - \overline{x}_n)^3}{s_n^3}.$$

Ako je $\mathrm{K_A} = 0$, podaci su simetrični (vidi Sliku 7.13). Za $\mathrm{K_A} < 0$ podaci su negativno asimetrični, a za $\mathrm{K_A} > 0$ podaci su pozitivno asimetrični (vidi Sliku 7.14).



Slika 7.13: Simetrični podaci

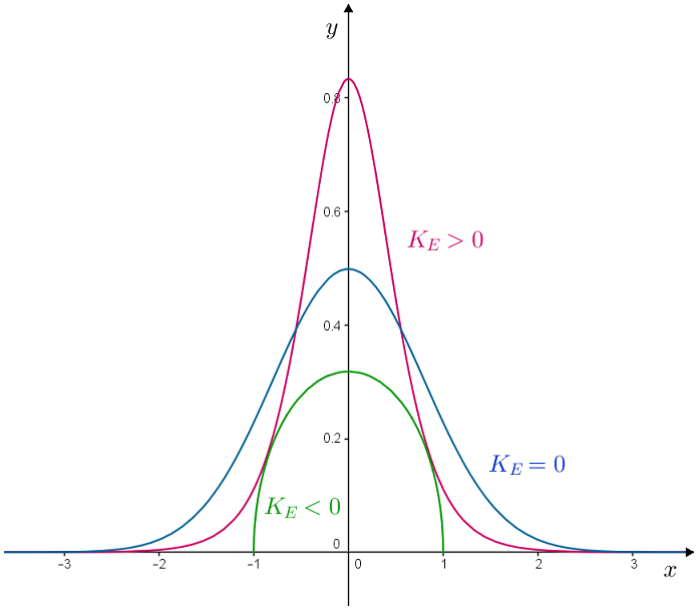


Slika 7.14: Asimetrični podaci (lijevo pozitivno asimetrični, a desno negativno asimetrični)

Koeficijent zaobljenosti vrijednosti $x_1, \ldots, x_n$ daje nam podatak o "zaobljenosti" (spljoštenosti) pripadne frekvencijske krivulje oko aritmetičke sredine. Računa se po formuli:

$$\mathrm{K_E} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_n)^4}{s_n^4} - 3 = \frac{\frac{1}{n}\sum_{i=1}^{k} f_i(a_i - \overline{x}_n)^4}{s_n^4} - 3.$$

Ako su vrijednosti $x_1, \ldots, x_n$ dobivene opažanjem normalne slučajne varijable, onda se može pokazati da je $\mathrm{K_E} \approx 0$. Dakle, koeficijent zaobljenosti uspoređuje razdiobu (frekvencijsku krivulju) podataka $x_1, \ldots, x_n$ s (gustoćom) normalnom razdiobom $N(\overline{x}_n, s_n^2)$. Na Slici 7.15 ove usporedbe grafički su prikazane.



Slika 7.15: Različito zaobljene frekvencijske krivulje podataka

Primjer 7.7. U Tablici 7.5 je dana raspodjela broja kvarova nekog uređaja dobivena na uzorku veličine 114.

Broj kvarova a_i	Broj uređaja f_i	$\frac{f_i}{114}$
0	3	0.026
1	9	0.078
2	15	0.131
3	26	0.228
4	38	0.333
5	18	0.158
6	5	0.044
$\sum$	114	1

Tablica 7.5: Frekvencije i relativne frekvencije broja kvarova

Da bismo izračunali koeficijent asimetrije i zaobljenosti prvo će nam trebati aritmetička sredina i uzoračka standarna devijacija navedenih podataka. U našem slučaju imamo

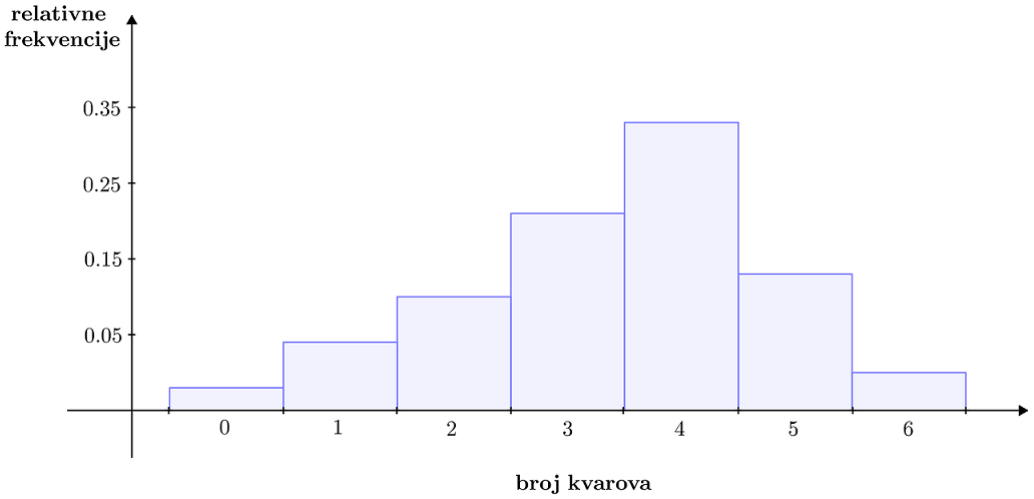
$$\overline{x}_{114} = 3.412 \qquad \text{i} \qquad s_{114} = 0.997.$$

Izračunajmo sada koeficijent asimetrije i koeficijent zaobljenosti danih podataka:

$$\begin{aligned} \mathrm{K_A} &= \frac{\frac{1}{114}\sum_{i=1}^{7} f_i(a_i - \overline{x}_{114})^3}{s_{114}^3} \\ &= \frac{\frac{1}{114}(3(0-3.412)^3 + \cdots + 5(6-3.412)^3)}{0.997^3} \\ &= \frac{\frac{1}{114}\cdot(-123.207)}{0.997^3} \\ &= -1.091, \end{aligned}$$

$$\begin{aligned} \mathrm{K_E} &= \frac{\frac{1}{114}\sum_{i=1}^{7} f_i(a_i - \overline{x}_{114})^4}{s_{114}^4} - 3 \\ &= \frac{\frac{1}{114}(3(0-3.412)^4 + \cdots + 5(6-3.412)^4)}{0.997^4} - 3 \\ &= \frac{\frac{1}{114}\cdot 1115.025}{0.997^4} - 3 \\ &= 6.909. \end{aligned}$$

Po negativnom predznaku koeficijenta asimetrije zaključujemo da su podaci negativno asimetrični, što možemo potvrditi i stupčastim dijagramom njihovih relativnih frekvencija (vidi Sliku 7.16).



Slika 7.16: Stupčasti dijagram relativnih frekvencija broja uređaja prema broju kvarova iz Primjera 7.7

□

Poglavlje 8

Inferencijalna statistika

Nakon organiziranja, predočavanja i opisivanja podataka pomoću metoda deskriptivne statistike možemo primijeniti inferencijalnu statistiku i vrednovati informaciju sadržanu u tim podacima. Neka je X varijabla koja opisuje ishode eksperimenata. Ako znamo sve vrijednosti varijable X (na svim elementima populacije), imamo punu informaciju o X i samom eksperimentu. Međutim, to nije uvijek moguće. Sama populacija može imati veliki broj elemenata i određivanje vrijednosti od X na svakom elementu može biti vrlo skupo i zahtjevno. Također, populacija može imati beskonačno elemenata. Cilj inferencijalne statistike je na osnovu uzorka izvesti određene zaključke o varijabli X, tj. njezinoj raspodjeli. S druge strane, sjetimo se da je slučajni pokus matematički opisan vjerojatnosnim prostorom $(\Omega, \mathcal{F}, \mathbb{P})$, gdje vjerojatnost $\mathbb{P} : \mathcal{F} \to [0, 1]$ u pravilu ne znamo. Metodama inferencijalne statistike možemo procijeniti vjerojatnost $\mathbb{P}$ jer na nju možemo gledati kao na raspodjelu od X.

Neka je X kvantitativna varijabla sa slikom $R(X)$ koja opisuje ishode eksperimenta na elementima populacije. Dakle, za svaki element populacije X se realizira nekim brojem x iz $R(X)$. Pretpostavka je da X ima slučajan karakter, tj. X je slučajna varijabla. **Slučajni uzorak** duljine n iz raspodjele od X je n-dimenzionalni slučajni vektor $(X_1, \ldots, X_n)$ čije su komponente nezavisne i jednako distribuirane slučajne varijable s raspodjelom od X. Realizaciju slučajnog uzorka $(X_1, \ldots, X_n)$ označavamo s $(x_1, \ldots, x_n)$. Deskriptivna statistika bavi se opisivanjem vrijednosti opažanja slučajnog uzorka, a inferencijalna zaključivanjem o raspodjeli iz koje dolazi slučajni uzorak. **Statistika** je svaka funkcija (transformacija) slučajnog uzorka koja je i sama slučajna varijabla. Primjerice,

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

su dvije statistike od $(X_1, \ldots, X_n)$. Samu raspodjelu slučajne varijable X nije lako odrediti. Da bismo je približe razumjeli, prvo ćemo odrediti neke njezine determinističke karakteristike: očekivanje $\mu = \mathbb{E}(X)$ i varijancu $\sigma^2 = \mathrm{Var}(X)$ (u slučaju da X ima iste). Dakle, želimo procijeniti μ i σ^2 na temelju informacija koje dobivamo iz realizacije slučajnog uzorka. **Procjenjivanje** je pridruživanje vrijednosti parametru populacije na temelju procjenitelja tog parametra. **Procjenitelj parametra** kojeg promatramo je funkcija slučajnog uzorka, dakle statistika. Primjerice, procjenitelji za μ i σ^2 su, redom, $\overline{X}_n$ i S_n^2. Međutim, kao što smo gore rekli, svaka funkcija slučajnog uzorka je procijenitelj za μ i σ^2. Primjerice, S_n^2 je i procjenitelj za μ, kao i $\overline{X}_n$ za σ^2. Pitanje je kako definirati pojam "dobrog" procjenitelja za određeni parametar te kako ga odrediti. U ovoj situaciji to je relativno lagano. Iz jakog zakona velikih brojeva slijedi da s vjerojatnošću 1 vrijedi

$$\lim_{n\to\infty} \overline{X}_n = \lim_{n\to\infty} \frac{\sum_{i=1}^n X_i}{n} = \mathbb{E}(X) = \mu$$

i

$$\begin{aligned}
\lim_{n\to\infty} S_n^2 &= \lim_{n\to\infty} \frac{\sum_{i=1}^n (X_i - \overline{X}_n)^2}{n-1} \\
&= \lim_{n\to\infty} \frac{\sum_{i=1}^n (X_i - \overline{X}_n - \mu + \mu)^2}{n-1} \\
&= \lim_{n\to\infty} \Big(\frac{\sum_{i=1}^n (X_i - \mu)^2}{n-1} + 2(\mu - \overline{X}_n) \frac{\sum_{i=1}^n (X_i - \mu)}{n-1} \\
&\qquad + \frac{n}{n-1} (\mu - \overline{X}_n)^2 \Big) \\
&= \lim_{n\to\infty} \frac{n}{n-1} \Big(\frac{\sum_{i=1}^n (X_i - \mu)^2}{n} - (\mu - \overline{X}_n)^2 \Big) \\
&= \mathbb{E}(X - \mu)^2 \\
&= \sigma^2,
\end{aligned}$$

gdje smo u petom koraku opet primjenili jaki zakon velikih brojeva. Nadalje, uočimo da bi

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

mogli biti "dobri" procjenitelji za, redom, μ i σ^2. Međutim, pokazuje sa da je u većini situacija (ali ne uvijek) "dobre" procjenitelje bolje tražiti među **nepristranim i konzistetnim procjeniteljima**. U konkretnoj situaciji nepristranost znači da vrijedi $\mathbb{E}(\overline{X}_n) = \mu$ i $\mathbb{E}(S_n^2) = \sigma^2$. S druge

strane, za predložene procjenitelje $\hat{X}_n$ i $\hat{S}_n^2$ imamo $\mathbb{E}(\hat{X}_n) = \mu n/(n-1)$ i $\mathbb{E}(\hat{S}_n^2) = \sigma^2(n-1)/n$, dakle ne zadovoljava potrebnu relaciju. Konzistetnost znači da procijenitelji $\overline{X}_n$ i S_n^2 kovergiraju u smislu (slabog ili jakog) zakona velikih brojeva ka, redom, μ i σ^2. U slučaju da konvergiraju u smislu slabog zakona govorimo o **slaboj konzistentnosti**, a u slučaju jakog zakona o **jakoj konzistentnosti**. U Poglavlju 4 smo komentirali da konvergencija u smislu jakog zakona velikih brojeva implicira konvergenciju u smislu slabog zakona pa samim time jaka konzistentost implicira slabu. Gore smo pokazali da $\overline{X}_n$ i S_n^2 kovergiraju ka, redom, μ i σ^2 u smislu jakog zakona velikih brojeva pa zaključujemo da su oni jako konzistentni procjenitelji za, redom, μ i σ^2. Uočimo da i $\hat{X}_n$ te $\hat{S}_n^2$ konvergiraju (u smislu jakog zakona velikih brojeva) ka, redom, μ i σ^2 pa su i oni jako konzistentni procijenitelji za, redom, μ i σ^2.

Promotrimo sljedeću situaciju. Na promatranoj populaciji promatramo neko dihotomno svojstvo, tj. svaki element populacije zadovoljava jedno, i samo jedno, od dva ponuđena svojstva (npr. spol). Želimo odrediti "dobar" procijenitelj za **populacijsku proporciju** koja ima prvo svojstvo (ili, ekvivalento, drugo svojstvo). U ovom slučaju prirodno je danu situaciju modelirati slučajnom varijablom X koja ima Bernoullijevu raspodjelu

$$\begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix},$$

gdje 1 označava prvo svojstvo, a 0 drugo svojstvo. Parametar p ozačava nepoznatu populacijsku proporciju sa svojstvom jedan. Kako je $p = \mathbb{E}(X)$, prema prethodnoj diskusiji, "dobar" procijenitelj za p je $\overline{X}_n$, gdje je $(X_1, \ldots, X_n)$ slučajni ozorak iz raspodjele od X.

U procjeni parametara μ i σ^2 od X imamo dva pristupa:

(a) točkovne procjene

(b) intervalne procjene.

Također, u tu svrhu diskutirat ćemo i postupak testiranja statističkih hipoteza, što je generalno puno općenitiji pristup koji ne služi samo za procjenu parametara raspodjele.

8.1 Točkovne procjene

Točkovna procjena je procjena kod koje se vrijednost (realizacija) procjenitelja uzima kao procjena parametra promatrane raspodjele. Ranije smo naveli da je $\overline{X}_n$ "dobar" izbor za procjenitelja od μ, a S_n^2 od σ^2. Također

uočimo da možemo odrediti brzinu (u smislu raspršenja/standardne devijacije) kojom $\overline{X}_n$ konvergira ka μ:

(i) ako je $(X_1, \ldots, X_n)$ slučajni uzorak od $N(\mu, \sigma^2)$, onda nije teško vidjeti da vrijedi $\overline{X}_n \sim N(\mu, \sigma^2/n)$. Dakle, u ovom slučaju, jer je $\overline{X}_n - \mu \sim N(0, \sigma^2/n)$, brzina konvergencije (u smislu raspršenja) od $\overline{X}_n$ ka μ je $\sigma(\overline{X}_n - \mu) = \sigma/\sqrt{n}$. Iz slabog zakona velikih brojeva zaključujemo da je za proizvoljni $\varepsilon > 0$ proporcija realizacija od $(X_1, \ldots, X_n)$ za koje $\overline{X}_n$ ne upada u interval $(\mu - \varepsilon, \mu + \varepsilon)$ najviše $\sigma^2/n\varepsilon$. Dakle, $\mu \approx \overline{x}_n$.

(ii) ako $(X_1, \ldots, X_n)$ nije slučajni uzorak normalne raspodjele, onda po centralnom graničnom teoremu znamo da je za velike n raspodjela od $\overline{X}_n$ blizu $N(\mu, \sigma^2/n)$, tj. raspodjela od $\overline{X}_n - \mu$ je blizu $N(0, \sigma^2/n)$. Ponovno, $\overline{X}_n$ se približava ka μ brzinom (u smislu raspršenja) $\sigma(\overline{X}_n - \mu) = \sigma/\sqrt{n}$. Također, kao i gore, zaključujemo da je za svaki $\varepsilon > 0$ proporcija realizacija od $(X_1, \ldots, X_n)$ za koje $\overline{X}_n$ ne upada u interval $(\mu - \varepsilon, \mu + \varepsilon)$ najviše $\sigma^2/n\varepsilon$. Dakle, $\mu \approx \overline{x}_n$.

Napomenimo da svaki slučajni uzorak daje drugačiju točkovnu procjenu parametara promatrane raspodjele i točkovna procjena se gotovo uvijek razlikuje od vrijednosti traženog parametra.

Primjer 8.1. Mjerimo visinu čovjeka u populaciji ljudi. Poznato je da visina čovjeka ima normalnu raspodjelu s poznatom varijancom $\sigma^2 = 64$ cm^2 i nepoznatim očekivanjem μ. Na slučajan način je izabran uzorak od 100 ljudi i izmjerena im je visina. Zbroj svih dobivenih visina iznosi 16910 cm. Odredimo točkovnu procjenu za μ. Neka slučajna varijabla X opisuje visinu čovjeka u populaciji ljudi. Po pretpostavci vrijedi $X \sim N(\mu, 64)$. Parametar μ procjenjujemo s

$$\overline{X}_{100} \sim N(\mu, 64/100).$$

Kako je realizacija od $X_1 + \cdots + X_{100}$ jednaka 16910, točkovna procjena od μ iznosi $16910/100 = 169.1$ cm. □

Primjer 8.2. U nekom gradu u glasačke listiće upisano je 10000 glasača. Ispitivanjem slučajnog uzorka od 1000 osoba želimo procijeniti postotak glasača stranke S. Nakon ispitivanja pokazuje se da u uzorku ima 700 glasača S-a. Odredimo točkovnu procjenu populacijske proporcije (postotka) glasača stranke S. Uočimo prvo da je na populaciji od 10000 glasača raspodjela slučajne varijable kojom modeliramo glasačku preferenciju prema S-u dana s

$$X \sim \begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix},$$

gdje 0 znači da glasač nije glasao za S, a 1 znači da je glasao za S. Populacijska proporcija glasača stranke S koju trebamo procijeniti je $\mathbb{E}(X) = p$. Kako je realizacija od $X_1 + \cdots + X_{1000}$ jednaka 700, procjena od p je $700/1000 = 0.7$. □

8.2 Intervalne procjene

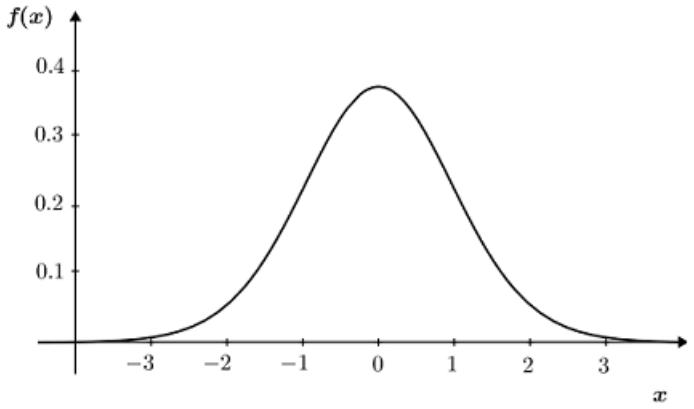
U ovom poglavlju susrest ćemo se s dvije neprekidne raspodjele koje do sada nismo spominjali: t-raspodjelom i χ^2-raspodjelom. U funkcijama gustoća navedenih raspodjela pojavljuje se tzv. Gama funkcija koju možemo definirati kroz konvergentni nepravi integral:

$$\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt, \qquad x > 0.$$

Napomenimo da za $n \in \mathbb{N}$ vrijedi $\Gamma(1) = 1$ i $\Gamma(n+1) = n\Gamma(n) = n!$.

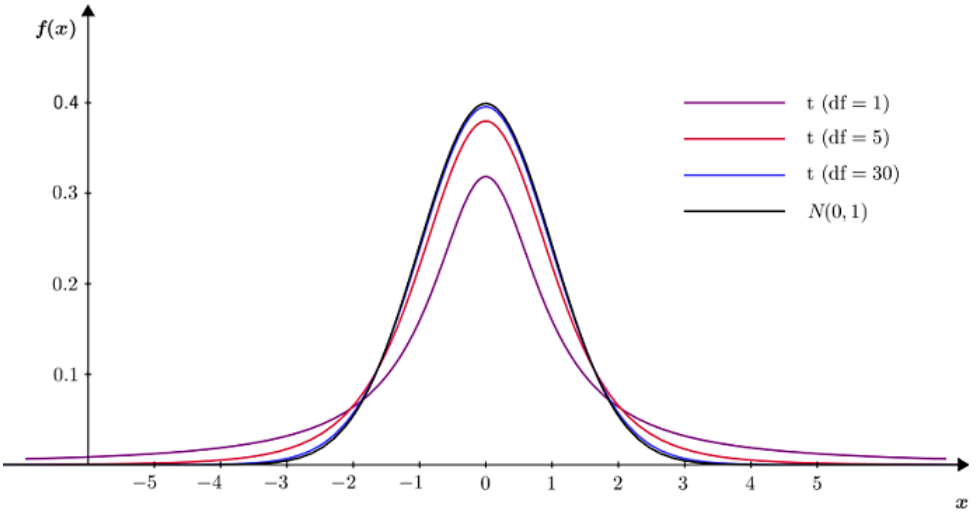
Za neprekidnu slučajnu varijablu X kažemo da ima t-**raspodjelu** (ili **Studentovu** raspodjelu) s n stupnjeva slobode (engl. *degrees of freedom*), $n \in \mathbb{N}$, u oznaci $X \sim t(n)$, ako je $R(X) = \mathbb{R}$ i funkcija gustoće od X je dana s

$$f(x) = \frac{1}{\sqrt{n\pi}} \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}.$$



Slika 8.1: Graf funkcije gustoće t-raspodjele s 20 stupnjeva slobode

Vrijedi $\mathbb{E}(X) = 0$ za $n \geqslant 2$ (za $n = 1$ očekivanje ne postoji) i $\mathrm{Var}(X) = n/(n-2)$ za $n \geqslant 3$ (za $n = 1$ i 2 varijanca ne postoji). Za praktične potrebe vrijednosti pripadne funkcije raspodjele tabelirane su za $n = 1, \ldots, 30$, dok za $n > 30$ istu možemo dobro aproksimirati funkcijom raspodjele od $N(0,1)$ (vidi Slike 8.1 i 8.2 te [6, str. 117]).



Slika 8.2: Usporedba t-raspodjele s različitim stupnjevima slobode i jedinične normalne raspodjele

Zašto parametar n u t-raspodjeli nazivamo stupnjem slobode? Ako je aritmetička sredina četiriju vrijednosti 10, tri od njih možemo odabrati po volji a četvrta je onda jedinstveno određena aritmetičkom sredinom. Npr. ako odaberemo brojeve 5, 8 i 12, četvrta vrijednost mora biti jednaka 15 (uz pretpostavku da je aritmetička sredina 10). U ovom primjeru imamo tri stupnja slobode. Dakle, ako imamo $n+1$ opažanja, n opažanja možemo odabrati po volji, a opažanje $(n+1)$ je jedinstveno određeno prethodnim opažanjima.

Napomenimo da za niz nezavisnih i jednako distribuiranih slučajnih varijabli $X_1, \ldots, X_n$ s raspodjelom $N(\mu, \sigma^2)$ vrijedi da slučajna varijabla

$$\frac{\overline{X}_n - \mu}{\sqrt{\frac{S_n^2}{n}}}$$

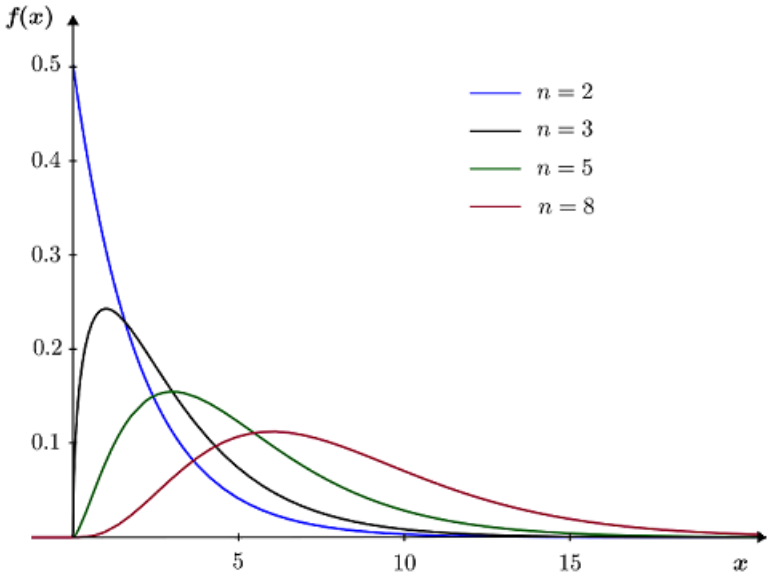
ima $t(n-1)$ raspodjelu (vidi [10, str. 307]). Stoga se t-raspodjela prirodno javlja u procjeni očekivanja.

Neprekidna slučajna varijabla X ima χ^2-**raspodjelu** s n stupnjeva slobode, $n \in \mathbb{N}$, u oznaci $X \sim \chi^2(n)$, ako je $R(X) = (0, \infty)$, a funkcija gustoće

od X je dana s:

$$f(x) = \begin{cases} 0, & x \leq 0 \\ \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{1}{2}x}, & x > 0. \end{cases}$$

Primjeri grafova ovih funkcija dani su na Slici 8.3.



Slika 8.3: Graf funkcije gustoće χ^2-raspodjele s različitim stupnjevima slobode

Vrijedi $\mathbb{E}(X) = n$ i $\mathrm{Var}(X) = 2n$ za sve $n \in \mathbb{N}$. Za praktične potrebe vrijednosti pripadne funkcije raspodjele tabelirane su za $n = 1, \ldots, 30$, dok za $n > 30$ istu možemo dobro aproksimirati funkcijom raspodjele od $N(n, 2n)$ (vidi [6, str. 116]). Napomenimo da za niz nezavisnih i jednako distribuiranih slučajnih varijabli $X_1, \ldots, X_n$ s raspodjelom $N(\mu, \sigma^2)$ vrijedi da slučajne varijable

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

imaju, redom, $\chi^2(n)$ i $\chi^2(n-1)$ raspodjelu (vidi [10, str. 305]). Stoga se χ^2-raspodjela prirodno javlja u procjeni varijance.

Kao što smo već naglasili, točkovna procjena se gotovo nikad ne podudara s pravom vrijednošću parametra. Međutim, uzmemo li određeni interval oko točkovne procjene, prava vrijednost bi se mogla nalaziti unutar tog intervala. Dakle, ideja intervalne procjene je konstruirati interval oko točkovne procjene, pri čemu želimo da navedeni interval sadrži nepoznatu vrijednost parametra promatrane raspodjele s određenom pouzdanošću. **Koeficijent pouzdanosti** je vjerojatnost da promatrani interval sadrži vrijednost nepozatog parametra promatrane raspodjele i zato taj interval konstruiramo kao slučajni interval, tj. interval čije su granice slučajne varijable. Uobičajeno uzimamo da je razina pouzdanosti $1-\alpha$ jednaka 0.9, 0.95 ili 0.99 (tj. α je 0.01, 0.05 ili 0.1). Za danu pouzdanost $1-\alpha \in (0,1)$, slučajni uzorak $(X_1,\ldots,X_n)$ iz X i statistike $L_n = f(X_1,\ldots,X_n)$ i $D_n = g(X_1,\ldots,X_n)$, kažemo da je $[L_n, D_n]$ **interval pouzdanosti** (pouzdanosti $1-\alpha$) za parametar τ (μ ili σ^2) ako $\mathbb{P}(L_n \leq \tau \leq D_n) \geqslant 1-\alpha$. Dakle, za barem $(1-\alpha)\cdot 100\%$ realizacija od $(X_1,\ldots,X_n)$ interval pouzdanosti $[L_n, D_n]$ sadrži stvarnu vrijednost τ nepoznatog parametra.

Neka je X slučajna varijabla s funkcijom raspodjele $F(x)$ i neka je $q \in (0,1)$. Broj $x_q \in \mathbb{R}$ zove se **q-ti kvantil** od X ako vrijedi $F(x_q) = q$.

8.2.1 Intervali pouzdanosti za μ

(a) Neka je $(X_1,\ldots,X_n)$ slučajni uzorak iz $N(\mu,\sigma^2)$.

(i) Varijanca σ^2 je poznata. U ovoj situaciji interval pouzdanosti (pouzdanosti $1-\alpha$) za μ je dan s

$$\left[\overline{X}_n - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \overline{X}_n + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right],$$

gdje je $z_{1-\frac{\alpha}{2}}$ $(1-\alpha/2)$-ti kvantil od $N(0,1)$. Uočimo da je $\sqrt{n}(\overline{X}_n - \mu)/\sigma \sim N(0,1)$ pa je

$$\begin{aligned}
&\mathbb{P}\left(\overline{X}_n - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leqslant \mu \leqslant \overline{X}_n + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)\\
&= \mathbb{P}\left(-z_{1-\frac{\alpha}{2}} \leqslant \frac{\overline{X}_n - \mu}{\sigma}\sqrt{n} \leqslant z_{1-\frac{\alpha}{2}}\right)\\
&= \Phi(z_{1-\frac{\alpha}{2}}) - \Phi(-z_{1-\frac{\alpha}{2}})\\
&= 1 - \frac{\alpha}{2} - 1 + 1 - \frac{\alpha}{2}\\
&= 1-\alpha,
\end{aligned}$$

gdje je $\Phi(z)$ funkcija raspodjele od $N(0,1)$. Dakle, za $(1-\alpha)\cdot 100\%$ realizacija slučajnog uzorka $(X_1,\ldots,X_n)$, interval pouzdanosti sadrži stvarnu vrijednost μ. Također, iz intervala pouzdanosti možemo iščitati da je širina intervala pouzdanosti dana s $\delta = 2z_{1-\frac{\alpha}{2}}\sigma/\sqrt{n}$, dok je minimalna veličina uzorka potrebna za postizanje intervala pouzdanosti (pouzdanosti $1-\alpha$) širine δ dana s $n \geqslant 4z^2_{1-\frac{\alpha}{2}}\sigma^2/\delta^2$.

(ii) Varijanca σ^2 nije poznata. U ovoj situaciji interval pouzdanosti (pouzdanosti $1-\alpha$) za μ je dan s

$$\left[\overline{X}_n - t_{1-\frac{\alpha}{2}}\frac{S_n}{\sqrt{n}}, \overline{X}_n + t_{1-\frac{\alpha}{2}}\frac{S_n}{\sqrt{n}}\right],$$

gdje je $t_{1-\frac{\alpha}{2}}$ $(1-\alpha/2)$-ti kvantil od $t(n-1)$ raspodjele. Ranije smo komentirali da vrijedi $\sqrt{n}(\overline{X}_n-\mu)/S_n \sim t(n-1)$ pa je

$$\begin{aligned}
&\mathbb{P}\left(\overline{X}_n - t_{1-\frac{\alpha}{2}}\frac{S_n}{\sqrt{n}} \leqslant \mu \leqslant \overline{X}_n + t_{1-\frac{\alpha}{2}}\frac{S_n}{\sqrt{n}}\right)\\
&= \mathbb{P}\left(-t_{1-\frac{\alpha}{2}} \leqslant \frac{\overline{X}_n-\mu}{S_n}\sqrt{n} \leqslant t_{1-\frac{\alpha}{2}}\right)\\
&= \mathbb{P}\left(\frac{\overline{X}_n-\mu}{S_n}\sqrt{n} \leqslant t_{1-\frac{\alpha}{2}}\right) - \mathbb{P}\left(\frac{\overline{X}_n-\mu}{S_n}\sqrt{n} \leqslant -t_{1-\frac{\alpha}{2}}\right)\\
&= 1-\frac{\alpha}{2}-1+1-\frac{\alpha}{2}\\
&= 1-\alpha.
\end{aligned}$$

Dakle, za $(1-\alpha)\cdot 100\%$ realizacija slučajnog uzorka $(X_1,\ldots,X_n)$, interval pouzdanosti sadrži stvarnu vrijednost μ.

(b) Imamo veliki (u praksi $n > 30$) slučajni uzorak $(X_1,\ldots,X_n)$ iz raspodjele koja ne mora biti normalna. Prisjetimo se, iz centralnog graničnog teorema znamo da za veliki n raspodjelu od $\sqrt{n}(\overline{X}_n-\mu)/\sigma$ možemo aproksimirati s $N(0,1)$.

(i) Varijanca σ^2 je poznata. U ovoj situaciji interval pouzdanosti (pouzdanosti $1-\alpha$) za μ je dan s

$$\left[\overline{X}_n - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \overline{X}_n + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right],$$

gdje je $z_{1-\frac{\alpha}{2}}$ $(1-\alpha/2)$-ti kvantil od $N(0,1)$. Budući da je uzorak velik, zaključujemo

$$\begin{aligned}
&\mathbb{P}\left(\overline{X}_n - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leqslant \mu \leqslant \overline{X}_n + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)\\
&= \mathbb{P}\left(\frac{\overline{X}_n - \mu}{\sigma}\sqrt{n} \leqslant z_{1-\frac{\alpha}{2}}\right) - \mathbb{P}\left(\frac{\overline{X}_n - \mu}{\sigma}\sqrt{n} \leqslant -z_{1-\frac{\alpha}{2}}\right)\\
&\approx \Phi(z_{1-\frac{\alpha}{2}}) - \Phi(-z_{1-\frac{\alpha}{2}})\\
&= 1-\alpha.
\end{aligned}$$

Dakle, za otprilike $(1-\alpha)\cdot 100\%$ realizacija slučajnog uzorka $(X_1,\ldots,X_n)$, interval pouzdanosti sadrži stvarnu vrijednost μ. Kao i ranije, širina intervala pouzdanosti je $\delta = 2z_{1-\frac{\alpha}{2}}\sigma/\sqrt{n}$ i minimalna veličina uzorka potrebna za postizanje intervala pouzdanosti (pouzdanosti $1-\alpha$) širine δ zadovoljava $n \geqslant 4z^2_{1-\frac{\alpha}{2}}\sigma^2/\delta^2$.

(ii) Varijanca σ^2 nije poznata. U ovoj situaciji interval pouzdanosti (pouzdanosti $1-\alpha$) za μ je dan s

$$\left[\overline{X}_n - z_{1-\frac{\alpha}{2}}\frac{S_n}{\sqrt{n}}, \overline{X}_n + z_{1-\frac{\alpha}{2}}\frac{S_n}{\sqrt{n}}\right],$$

gdje je $z_{1-\frac{\alpha}{2}}$ $(1-\alpha/2)$-ti kvantil od $N(0,1)$. Budući da je uzorak velik može se pokazati da raspodjelu od $\sqrt{n}(\overline{X}_n-\mu)/S_n$ možemo aproksimirati s $N(0,1)$. Stoga imamo

$$\begin{aligned}
&\mathbb{P}\left(\overline{X}_n - z_{1-\frac{\alpha}{2}}\frac{S_n}{\sqrt{n}} \leqslant \mu \leqslant \overline{X}_n + z_{1-\frac{\alpha}{2}}\frac{S_n}{\sqrt{n}}\right)\\
&= \mathbb{P}\left(\frac{\overline{X}_n - \mu}{S_n}\sqrt{n} \leqslant z_{1-\frac{\alpha}{2}}\right) - \mathbb{P}\left(\frac{\overline{X}_n - \mu}{S_n}\sqrt{n} \leqslant -z_{1-\frac{\alpha}{2}}\right)\\
&\approx \Phi(z_{1-\frac{\alpha}{2}}) - \Phi(-z_{1-\frac{\alpha}{2}})\\
&= 1-\alpha.
\end{aligned}$$

Dakle, za otprilike $(1-\alpha)\cdot 100\%$ realizacija slučajnog uzorka $(X_1,\ldots,X_n)$ interval pouzdanosti sadrži stvarnu vrijednost μ.

Kao posljedicu možemo odrediti interval pouzdanosti (pouzdanosti $1-\alpha$) za parametar Bernoullijeve raspodjele. Neka je

$$X \sim \begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}$$

i $\alpha \in (0,1)$. Kako je $\mathbb{E}(X) = p$ i $\mathrm{Var}(X) = p(1-p)$, iz centralnog graničnog teorema znamo da za velike n raspodjelu od $\sqrt{n}(\overline{X}_n - p)/\sqrt{p(1-p)}$ možemo aproksimirati s $N(0,1)$. Dakle, interval pouzdanosti (pouzdanosti $1-\alpha$) za p je dan s

$$\left[\overline{X}_n - z_{1-\frac{\alpha}{2}}\frac{S_n}{\sqrt{n}}, \overline{X}_n + z_{1-\frac{\alpha}{2}}\frac{S_n}{\sqrt{n}}\right],$$

gdje je $z_{1-\frac{\alpha}{2}}$ $(1-\alpha/2)$-ti kvantil od $N(0,1)$. Nadalje, lagano se pokaže da u ovoj situaciji vrijedi

$$S_n^2 = \frac{n}{n-1}\overline{X}_n(1-\overline{X}_n).$$

Uočimo da za sve $x \in [0,1]$ vrijedi $x(1-x) \leqslant 1/4$. Dakle, u ovoj situaciji, širina intervala pouzdanosti zadovoljava

$$\delta = 2z_{1-\frac{\alpha}{2}}\frac{S_n}{\sqrt{n}} = 2z_{1-\frac{\alpha}{2}}\frac{\sqrt{\overline{X}_n(1-\overline{X}_n)}}{\sqrt{n-1}} \leqslant \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n-1}}.$$

Primjer 8.3. U jednom istraživanju kontrole kvalitete boca na uzorku veličine 25 izmjerene su debljine stijenki dvolitrenih staklenih boca. Uzoračka aritmetička sredina iznosila je 4.05 milimetara, a uzoračka standardna devijacija 0.08 milimetara. Poznato je da debljina stijenki boca ima normalnu raspodjelu s nepoznatim očekivanjem i nepoznatom varijancom. Pronađimo 95%-tni pouzdani interval za očekivanu debljinu stijenke. Budući da nam varijanca populacije nije poznata, tražimo interval oblika

$$\left[\overline{X}_{25} - t_{0.975}\frac{S_{25}}{\sqrt{25}}, \overline{X}_{25} + t_{0.975}\frac{S_{25}}{\sqrt{25}}\right].$$

Pritom, realizacije od $\overline{X}_{25}$ i S_{25} su, redom, $\overline{x}_{25} = 4.05$ i $s_{25} = 0.08$. Iz statističkih tablica iščitavamo da je za $n-1 = 24$ stupnja slobode $t_{0.975} = 2.06$. Dobivamo interval

$$\left[4.05 - 2.06 \cdot \frac{0.08}{\sqrt{25}}, 4.05 + 2.06 \cdot \frac{0.08}{\sqrt{25}}\right] = [4.01704, 4.08296].$$

Dakle, s 95% pouzdanosti možemo tvrditi da stvarno očekivanje debljine stijenke leži u dobivenom intervalu. □

8.2.2 Intervali pouzdanosti za σ^2

Neka je $(X_1, \ldots, X_n)$ slučajni uzorak iz $N(\mu, \sigma^2)$.

(a) Očekivanje μ je poznato. U ovoj situaciji interval pouzdanosti (pouzdanosti $1-\alpha$) za σ^2 je dan s

$$\left[\frac{1}{\chi^2_{1-\frac{\alpha}{2}}}\sum_{i=1}^{n}(X_i-\mu)^2, \frac{1}{\chi^2_{\frac{\alpha}{2}}}\sum_{i=1}^{n}(X_i-\mu)^2\right],$$

gdje su $\chi^2_{\frac{\alpha}{2}}$ i $\chi^2_{1-\frac{\alpha}{2}}$, redom, $\alpha/2$-ti kvanil i $(1-\alpha/2)$-ti kvantil od $\chi^2(n)$. Ranije smo komentirali da vrijedi

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

Sada imamo

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{\chi^2_{1-\frac{\alpha}{2}}}\sum_{i=1}^{n}(X_i-\mu)^2 \leqslant \sigma^2 \leqslant \frac{1}{\chi^2_{\frac{\alpha}{2}}}\sum_{i=1}^{n}(X_i-\mu)^2\right)\\
&= \mathbb{P}\left(\chi^2_{\frac{\alpha}{2}} \leqslant \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2} \leqslant \chi^2_{1-\frac{\alpha}{2}}\right)\\
&= \mathbb{P}\left(\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2} \leqslant \chi^2_{1-\frac{\alpha}{2}}\right) - \mathbb{P}\left(\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2} \leqslant \chi^2_{\frac{\alpha}{2}}\right)\\
&= 1-\frac{\alpha}{2}-\frac{\alpha}{2}\\
&= 1-\alpha.
\end{aligned}$$

Dakle, za $(1-\alpha)\cdot 100\%$ realizacija slučajnog uzorka $(X_1,\ldots,X_n)$, interval pouzdanosti sadrži stvarnu vrijednost σ^2. Širina intervala pouzdanosti je

$$\delta = \left(\frac{1}{\chi^2_{1-\frac{\alpha}{2}}} - \frac{1}{\chi^2_{\frac{\alpha}{2}}}\right)\sum_{i=1}^{n}(X_i-\mu)^2.$$

(b) Očekivanje μ je nepoznato. U ovoj situaciji interval pouzdanosti (pouzdanosti $1-\alpha$) za σ^2 je dan s

$$\left[\frac{n-1}{\chi^2_{1-\frac{\alpha}{2}}}S_n^2, \frac{n-1}{\chi^2_{\frac{\alpha}{2}}}S_n^2\right],$$

gdje su $\chi^2_{\frac{\alpha}{2}}$ i $\chi^2_{1-\frac{\alpha}{2}}$, redom, $\alpha/2$-ti kvantil i $(1-\alpha/2)$-ti kvantil od $\chi^2(n-1)$. Ranije smo napomenuli da vrijedi $(n-1)S_n^2/\sigma^2 \sim \chi^2(n-1)$.

Sada imamo

$$\begin{aligned}
&\mathbb{P}\left(\frac{n-1}{\chi^2_{1-\frac{\alpha}{2}}}S_n^2 \leqslant \sigma^2 \leqslant \frac{n-1}{\chi^2_{\frac{\alpha}{2}}}S_n^2\right)\\
&= \mathbb{P}\left(\chi^2_{\frac{\alpha}{2}} \leqslant \frac{(n-1)S_n^2}{\sigma^2} \leqslant \chi^2_{1-\frac{\alpha}{2}}\right)\\
&= \mathbb{P}\left(\frac{(n-1)S_n^2}{\sigma^2} \leqslant \chi^2_{1-\frac{\alpha}{2}}\right) - \mathbb{P}\left(\frac{(n-1)S_n^2}{\sigma^2} \leqslant \chi^2_{\frac{\alpha}{2}}\right)\\
&= 1 - \frac{\alpha}{2} - \frac{\alpha}{2}\\
&= 1 - \alpha.
\end{aligned}$$

Dakle, za $(1-\alpha)\cdot 100\%$ realizacija slučajnog uzorka $(X_1, \ldots, X_n)$, interval pouzdanosti sadrži stvarnu vrijednost σ^2.

Primjer 8.4. Mjerimo visinu čovjeka u populaciji ljudi. Poznato je da visina čovjeka ima normalnu raspodjelu s nepoznatim očekivanjem μ i varijancom σ^2. Na slučajan način izabran je uzorak od 100 ljudi i izmjerena je aritmetička sredina tog uzorka $\overline{x}_n = 170$ cm i uzoračka varijanca $s_n^2 = 64$ cm^2. Odredimo intervale povjerenja pouzdanosti $1-\alpha = 0.95$ za μ i σ^2. Kako je $(X_1, \ldots, X_n)$ slučajni uzorak iz $N(\mu, \sigma^2)$ i uzorak je velik, interval pouzdanosti (pouzdanosti 0.95) za μ je dan s

$$\left[\overline{X}_{100} - 1.96\frac{S_{100}^2}{10}, \overline{X}_{100} + 1.96\frac{S_{100}^2}{10}\right],$$

a za σ^2 je dan s

$$\left[\frac{99}{126.58}S_{100}^2, \frac{99}{71.42}S_{100}^2\right].$$

Realizacije od $\overline{X}_{100}$ i S_{100}^2 su, redom, 170 i 64. Dakle, realizacije 95%-tnih pouzdanih intervala su $[168.432, 171.568]$ i $[50.05, 88.71]$. □

Kao što smo već rekli, raspodjele $t(n)$ i $\chi^2(n)$ su tabelirane za vrijednosti $n = 1, \ldots, 30$, a za vrijednosti $n > 30$, $t(n)$ i $\chi^2(n)$ mogu se dobro aproksimirati s, redom, $N(0,1)$ i $N(n, 2n)$. Za detaljniju diskusiju o svim gore spomenutim rezultatima čitatelju preporučamo reference [5, 6, 7, 8, 10].

8.3 Testiranje statističkih hipoteza

Promotrimo sljedeći primjer. Razumno je pomisliti da beton s čeličnim vlaknima ima bolja svojstva od običnog betona, npr. veću tlačnu čvrstoću. Ako želimo provjeriti ovu pretpostavku, možemo izmjeriti tlačnu čvrstoću nekog broja, recimo 7, betonskih kocki s čeličnim vlaknima. Zamislimo da dobijemo sljedeće vrijednosti (u $\mathrm{N/mm^2}$):

$$68, 70, 62, 65, 70, 66, 67.$$

Naravno, ova mjerenja ne mogu nam reći ništa o našoj pretpostavci, ako nemamo nikakve podatke o tlačnoj čvrstoći betona bez čeličnih vlakana. Najbolji način je, ako to možemo, u istim uvjetima izmjeriti tlačnu čvrstoću nekog broja, recimo ponovo 7, betonskih kocki bez čeličnih vlakana. Ako dobijemo sljedeće vrijednosti:

$$63, 61, 59, 64, 64, 62, 66,$$

vidimo da su one zaista uglavnom niže od vrijednosti izmjerenih za betonske kocke s čeličnim vlaknima. Ipak, ponekad dolazi i do preklapanja. Stoga je razumno izračunati aritmetičke sredine dvaju mjerenja – one iznose $\overline{x}_7^{(1)} = 66.8571$ i $\overline{x}_7^{(2)} = 62.1428$. No, ni ovo nije dovoljno da dođemo do željenog zaključka. Osim prosjeka, bitno je i koliko se podaci raspršuju (variraju). Ako su jako raspršeni i imaju iste aritmetičke sredine, onda se ova razlika, na broju mjerenja te veličine, vrlo lako mogla dogoditi i slučajno. Stoga ćemo razliku između tih prosjeka usporediti (podijeliti) s nekom mjerom raspršenja tih mjerenja, uzimajući u obzir i broj mjerenja. Uz pretpostavku da su tlačne čvrstoće koje mjerimo normalno distribuirane, to će nam omogućiti da dobiveni rezultat usporedimo s nekim poznatim vrijednostima (koje se nalaze u statističkim tablicama) i zaključimo koliko je (mala) vjerojatnost da se naša razlika, na broju mjerenja te veličine, pojavi slučajno, iako su stvarni prosjeci jednaki. Ako je ta vjerojatnost dovoljno mala (npr. manja od 5%), odbacit ćemo hipotezu da su stvarna očekivanja jednaka i doći do željenog zaključka da beton s čeličnim vlaknima ima veću tlačnu čvrstoću.

Neka je X slučajna varijabla koja modelira ishode slučajnog eksperimenta. **Statistička hipoteza** je bilo koja pretpostavka o raspodjeli od X. Statističke hipoteze označavamo s H_0 i H_1 te ih zovemo **nul-hipoteza** i **alternativna hipoteza**, redom. Hipoteza H_1 je alternativna (u nekom smislu suprotna) hipotezi H_0. **Neparametarska hipoteza** je pretpostavka o raspodjeli od X. Primjerice, H_0: X ima normalnu raspodjelu. **Parametarska hipoteza** je pretpostavka o parametrima od X. Primjerice, $H_0 : \mu = \mu_0$, gdje je μ očekivanje od X, a μ_0 fiksan broj. Mi ćemo se baviti samo testiranjem

parametarskih hipoteza. **Testiranje statističkih hipoteza** je postupak donošenja odluke o odbacivanju ili neodbacivanju H_0 na osnovu informacije dobivene iz opažanja slučajnog uzorka. U slučaju neodbacivanja H_0 ne možemo tvrditi da je H_1 točna, dok u slučaju odbacivanja H_0 tvrdimo da je H_1 točna. Prethodnu situaciju možemo usporediti sa slučajem sudskog postupka: H_0: optuženi je nevin i H_1: optuženi je kriv. Ako se optuženom ne dokaže krivica, to ne znači da je on nevin nego samo da nemamo dovoljno dokaza koji bi potvrdili njegovu krivicu. S druge strane, ako imamo dovoljno dokaza koji potvrđuju njegovu krivicu, onda možemo tvrditi da je optuženi kriv. Statistika slučajnog uzorka pomoću koje se donosi odluka o odbacivanju ili neodbacivanju H_0 zove se **test-statistika**. Budući da niti jedna odluka bazirana na uzorcima nije 100% pouzdana, ni zaključci statističkog testa nisu 100% pouzdani. Dakle, može se dogoditi da je odluka donesena temeljem statističkog testa pogrešna. Test-statistikom želimo odrediti granice područja odbacivanja i neodbacivanja H_0, tj. **kritičnog područja testa**. Odluku o odbacivanju ili neodbacivanju H_0 donosimo na temelju pripadanja ili nepripadanja vrijednosti test-statistike kritičnom području. Kao što smo rekli, u tom postupku javljaju se moguće pogreške, koje su prikazane u tablici 8.1.

Stanje \ Odluka	H_0 se ne odbacuje	H_0 se odbacuje
H_0 istinita	ispravna odluka	α – pogreška 1. tipa
H_0 lažna	β – pogreška 2. tipa	ispravna odluka

Tablica 8.1: Mogući zaključci testiranja

Broj $\alpha \in (0,1)$ predstavlja maksimalnu pogrešku koju činimo kada odbacujemo istinitu H_0:

$$\alpha = \mathbb{P}(H_0 \text{ se odbacuje} | H_0 \text{ je istinita}).$$

Navedenu pogrešku zovemo **nivo značajnosti**. Broj $\beta \in (0,1)$ predstavlja maksimalnu pogrešku koju činimo kada ne odbacujemo H_0, a istinita je H_1:

$$\beta = \mathbb{P}(H_0 \text{ se ne odbacuje} | H_0 \text{ nije istinita}).$$

Broj $1-\beta$ naziva se **jakost testa**. Razumno je zahtjevati test kojim se mogu kontrolirati (smanjiti) obje pogreške. Međutim, to nije moguće jer ako se α smanji, onda se smanjuje kritično područje testa, što rezultira povećanjem β. U praksi se prije provođenja testa izabere značajnost (najčešće $\alpha = 0.01$ ili 0.05), odredi se kritično područje, pa se naknadno izračuna β. Ako je β prevelik, test ponavljamo s većim uzorkom.

8.3.1 Testiranje statističkih hipoteza za μ

(1) Neka je $(X_1, \ldots, X_n)$ slučajni uzorak iz $N(\mu, \sigma^2)$.

(a) Varijanca σ^2 je poznata.

(i) Dvostrani test uz nivo značajnosti $\alpha \in (0, 1)$. Testiramo:

$$H_0 \ : \ \mu = \mu_0$$
$$H_1 \ : \ \mu \neq \mu_0.$$

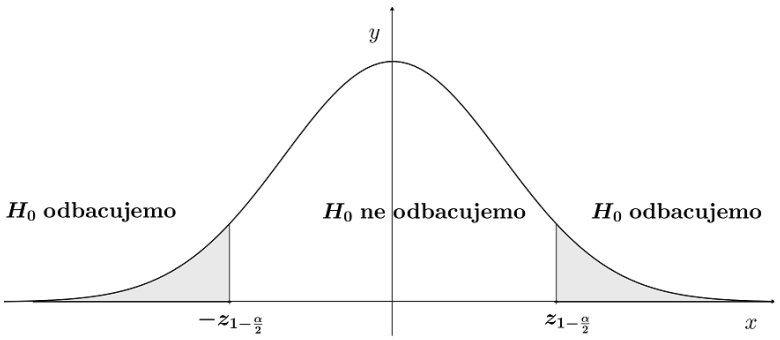
U ovom slučaju test-statistika je

$$T = \sqrt{n}\frac{\overline{X}_n - \mu_0}{\sigma} \sim N(0, 1).$$

Nul-hipotezu ćemo odbaciti ukoliko vrijednost test-statistike

$$t = \sqrt{n}\frac{\overline{x}_n - \mu_0}{\sigma}$$

za realizaciju $(x_1, \ldots, x_n)$ slučajnog uzorka $(X_1, \ldots, X_n)$ pripada kritičnom području $(-\infty, -z_{1-\frac{\alpha}{2}}] \cup [z_{1-\frac{\alpha}{2}}, \infty)$, gdje je $z_{1-\frac{\alpha}{2}}$, kao i do sada, $(1-\alpha/2)$-ti kvantil od $N(0, 1)$ (vidi Sliku 8.4).



Slika 8.4: Kritično područje kod dvostranog testa

Odbacivanjem H_0 na nivou značajnosti α zaključili smo da je razlika između realizacije od $\overline{X}_n$ i μ_0 statistički značajna. Neodbacivanjem H_0 zaključujemo da razlika nije statistički značajna. Uočimo da je površina ispod grafa funkcije gustoće od $N(0, 1)$ nad kritičnim područjem α, a nad njegovim komplementom $1 - \alpha$. Dakle, $\alpha \cdot 100\%$ realizacija od $(X_1, \ldots, X_n)$ uzrokuje odbacivanje H_0 ako je H_0 istinita.

(ii) Lijevi jednostrani test uz nivo značajnosti $\alpha \in (0,1)$. Testiramo:

$$H_0 \ : \ \mu \geqslant \mu_0$$
$$H_1 \ : \ \mu < \mu_0.$$

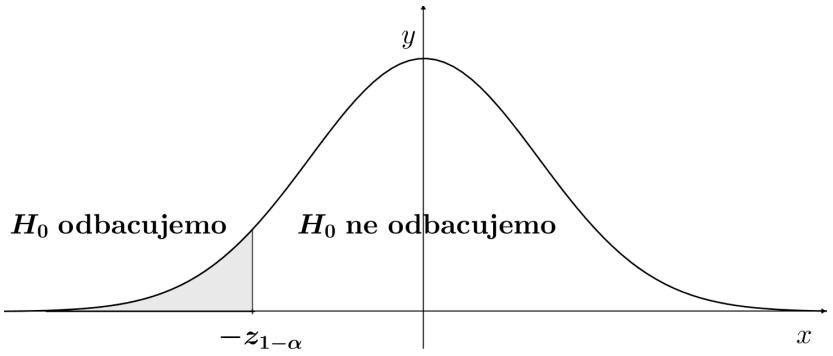
Test-statistika je ponovno

$$T = \sqrt{n}\frac{\overline{X}_n - \mu_0}{\sigma} \sim N(0,1),$$

a kritično područje je $(-\infty, -z_{1-\alpha}]$, gdje je $z_{1-\alpha}$ $(1-\alpha)$-ti kvantil od $N(0,1)$. Nul-hipotezu ćemo odbaciti ukoliko vrijednost test-statistike

$$t = \sqrt{n}\frac{\overline{x}_n - \mu_0}{\sigma}$$

za realizaciju $(x_1, \ldots, x_n)$ slučajnog uzorka $(X_1, \ldots, X_n)$ pripada kritičnom području $(-\infty, -z_{1-\alpha}]$ (vidi Sliku 8.5).



Slika 8.5: Kritično područje kod lijevog jednostranog testa

Ponovno, $\alpha \cdot 100\%$ realizacija od $(X_1, \ldots, X_n)$ uzrokuje odbacivanje H_0 ako je H_0 istinita.

(iii) Desni jednostrani test uz nivo značajnosti $\alpha \in (0,1)$. Testiramo:

$$H_0 \ : \ \mu \leqslant \mu_0$$
$$H_1 \ : \ \mu > \mu_0.$$

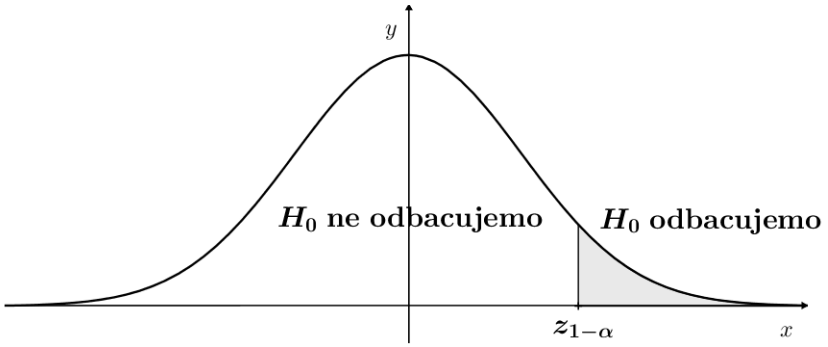
Test-statistika je

$$T = \sqrt{n}\frac{\overline{X}_n - \mu_0}{\sigma} \sim N(0,1),$$

a kritično područje je $[z_{1-\alpha}, \infty)$, pri čemu je $z_{1-\alpha}$ $(1-\alpha)$-ti kvantil od $N(0,1)$. Nul-hipotezu ćemo odbaciti ukoliko vrijednost test-statistike

$$t = \sqrt{n}\frac{\overline{x}_n - \mu_0}{\sigma}$$

za realizaciju $(x_1, \ldots, x_n)$ slučajnog uzorka $(X_1, \ldots, X_n)$ pripada kritičnom području $[z_{1-\alpha}, \infty)$ (vidi Sliku 8.6).



Slika 8.6: Kritično područje kod desnog jednostranog testa

Dakle, najviše $\alpha \cdot 100\%$ realizacija od $(X_1, \ldots, X_n)$ uzrokuje odbacivanje H_0 ako je H_0 istinita.

(b) Varijanca σ^2 nije poznata. U ovoj situaciji procedura je analogna kao u slučaju poznate varijance, samo test-statistiku zamijenimo s

$$T = \sqrt{n}\frac{\overline{X}_n - \mu_0}{S_n} \sim t(n-1),$$

vrijednost test-statistike s

$$t = \sqrt{n}\frac{\overline{x}_n - \mu_0}{s_n}$$

za realizaciju $(x_1, \ldots, x_n)$ slučajnog uzorka $(X_1, \ldots, X_n)$, a kvantile koji se javljaju za određivanje kritičnog područja određujemo iz $t(n-1)$ raspodjele.

(2) Imamo veliki (u praksi $n > 30$) slučajni uzorak $(X_1, \ldots, X_n)$ iz raspodjele koja ne mora biti normalna.

(a) Varijanca σ^2 je poznata.

(i) Dvostrani test uz nivo značajnosti $\alpha \in (0,1)$. Testiramo:

$$H_0 \ : \ \mu = \mu_0$$
$$H_1 \ : \ \mu \neq \mu_0.$$

U ovom slučaju test-statistika je $T = \sqrt{n}(\overline{X}_n - \mu_0)/\sigma$ i iz centralnog graničnog teorema znamo da za velike n raspodjelu od T možemo aproksimirati s $N(0,1)$. Vrijednost test-statistike je $t = \sqrt{n}(\overline{x}_n - \mu_0)/\sigma$, a kritično područje je dano s $(-\infty, -z_{1-\frac{\alpha}{2}}] \cup [z_{1-\frac{\alpha}{2}}, \infty)$, gdje je $z_{1-\frac{\alpha}{2}}$ $(1-\alpha/2)$-ti kvantil od $N(0,1)$.

(ii) Lijevi jednostrani test uz nivo značajnosti $\alpha \in (0,1)$. Testiramo:

$$H_0 \ : \ \mu \geqslant \mu_0$$
$$H_1 \ : \ \mu < \mu_0$$

Test-statistika je ponovno $T = \sqrt{n}(\overline{X}_n - \mu_0)/\sigma$, čiju raspodjelu za velike n možemo aproksimirati s $N(0,1)$, vrijednost test-statistike je $t = \sqrt{n}(\overline{x}_n - \mu_0)/\sigma$, a kritično područje je $(-\infty, -z_{1-\alpha}]$, gdje je $z_{1-\alpha}$ $(1-\alpha)$-ti kvantil od $N(0,1)$.

(iii) Desni jednostrani test uz nivo značajnosti $\alpha \in (0,1)$. Testiramo:

$$H_0 \ : \ \mu \leqslant \mu_0$$
$$H_1 \ : \ \mu > \mu_0.$$

Test-statistika je $T = \sqrt{n}(\overline{X}_n - \mu_0)/\sigma$ (čiju raspodjelu za velike n možemo aproksimirati s $N(0,1)$), vrijednost test-statistike je $t = \sqrt{n}(\overline{x}_n - \mu_0)/\sigma$ i kritično područje je $[z_{1-\alpha}, \infty)$, gdje je $z_{1-\alpha}$ $(1-\alpha)$-ti kvantil od $N(0,1)$.

b) Varijanca σ^2 nije poznata. U ovoj situaciji procedura je analogna kao u slučaju poznate varijance samo test-statistiku zamijenimo s $T = \sqrt{n}(\overline{X}_n - \mu_0)/S_n$ i vrijednost test-statistike s $t = \sqrt{n}(\overline{x}_n - \mu_0)/s_n$. Primjenom centralnog graničnog teorema može se pokazati da za velike n raspodjelu od T možemo aproksimirati s $N(0,1)$. Dakle, kvantile koji određuju kritično područje ponovno određujemo iz $N(0,1)$.

Koristeći gore navedene rezultate, primjenom centralnog graničnog teorema, možemo testirati hipoteze o parametru Bernoullijeve slučajne varijable

$$X \sim \begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}.$$

Neka je $(X_1, \ldots, X_n)$ veliki slučajni uzorak za X. Kako je $\mathbb{E}(X) = p$ i $\mathrm{Var}(X) = p(1-p)$, za test-statistiku uzimamo

$$T = \sqrt{n}\frac{\overline{X}_n - p_0}{S_n}.$$

Uočimo da vrijedi

$$S_n^2 = \frac{n\overline{X}_n(1-\overline{X}_n)}{n-1} \qquad \text{i} \qquad s_n^2 = \frac{n\overline{x}_n(1-\overline{x}_n)}{n-1}$$

pa je

$$T = \sqrt{n-1}\frac{\overline{X}_n - p_0}{\sqrt{\overline{X}_n(1-\overline{X}_n)}} \qquad \text{i} \qquad t = \sqrt{n-1}\frac{\overline{x}_n - p_0}{\sqrt{\overline{x}_n(1-\overline{x}_n)}}.$$

Sada možemo testirati hipoteze:

(i) Dvostrani test

$$\begin{aligned} H_0 &: p = p_0 \\ H_1 &: p \neq p_0. \end{aligned}$$

(ii) Lijevi jednostrani test

$$\begin{aligned} H_0 &: p \geqslant p_0 \\ H_1 &: p < p_0. \end{aligned}$$

(iii) Desni jednostrani test

$$\begin{aligned} H_0 &: p \leqslant p_0 \\ H_1 &: p > p_0. \end{aligned}$$

Primjer 8.5. Komentirajmo sada Primjer 6.1 s početka ovog poglavlja. Danu situaciju modeliramo Bernoullijevom raspodjelom

$$\begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix},$$

gdje 0 označava da je prilikom bacanja novčića palo pismo, a 1 da je pala glava. Testiramo

$$H_0 \ : \ p = \frac{1}{2}$$
$$H_1 \ : \ p \neq \frac{1}{2}$$

uz nivo značajnosti $\alpha = 0.05$. Na osnovu 100 nezavisnih bacanja novčića (slučajnog uzorka veličine $n = 100$) imamo $\overline{x}_{100} = 0.6$. Test-statistika je

$$T = \sqrt{99}\frac{\overline{X}_{100} - 1/2}{\sqrt{\overline{X}_{100}(1 - \overline{X}_{100})}},$$

a kritično područje $(-\infty, -z_{0.975}] \cup [z_{0.975}, \infty) = (-\infty, -1.96] \cup [1.96, \infty)$, gdje je $z_{0.975}$ 0.975-ti kvantil od $N(0, 1)$. Sada, kako vrijednost test-statistike $t = 2.0307$ pripada kritičnom području, odbacujemo H_0 s mogućnošću pogreške od 5%. Drugim riječima, na osnovu gornjeg testa zaključujemo da novčić nije simetričan, s mogućnošću pogreške od 5%. □

8.3.2 Testiranje statističkih hipoteza za σ^2

Neka je $(X_1, \ldots, X_n)$ slučajni uzorak iz $N(\mu, \sigma^2)$. Parametar očekivanja μ može biti nepoznat.

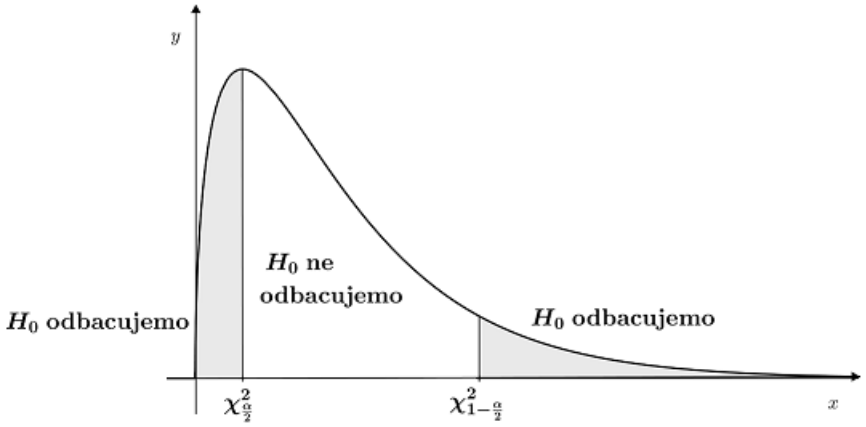
(i) **Dvostrani test** uz nivo značajnost $\alpha \in (0, 1)$. Testiramo:

$$H_0 \ : \ \sigma^2 = \sigma_0^2$$
$$H_1 \ : \ \sigma^2 \neq \sigma_0^2.$$

U ovom slučaju test-statistika je

$$T = \frac{n-1}{\sigma_0^2} S_n^2 \sim \chi^2(n-1).$$

Kritično područje je dano s $[0, \chi^2_{\frac{\alpha}{2}}] \cup [\chi^2_{1-\frac{\alpha}{2}}, \infty)$, gdje su $\chi^2_{\frac{\alpha}{2}}$ i $\chi^2_{1-\frac{\alpha}{2}}$, redom, $\alpha/2$-ti i $(1-\alpha/2)$-ti kvantili od $\chi^2(n-1)$. Ovo kritično područje prikazano je na Slici 8.7.



Slika 8.7: Kritično područje kod dvostranog testa

Nul-hipotezu ćemo odbaciti ukoliko vrijednost test-statistike

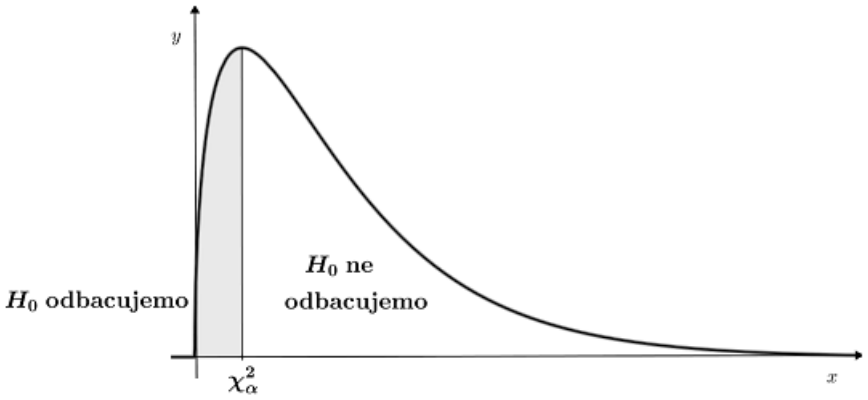
$$t = \frac{n-1}{\sigma_0^2} s_n^2$$

za realizaciju $(x_1, \ldots, x_n)$ slučajnog uzorka $(X_1, \ldots, X_n)$ pripada kritičnom području. Površina ispod grafa funkcije gustoće od $\chi^2(n-1)$ raspodjele nad kritičnim područjem je α, a nad njegovim komplementom $1-\alpha$. Dakle, $\alpha \cdot 100\%$ realizacija od $(X_1, \ldots, X_n)$ uzrokuje odbacivanje H_0 ako je H_0 istinita.

(ii) **Lijevi jednostrani test** uz nivo značajnosti $\alpha \in (0,1)$. Testiramo:

$$\begin{aligned} H_0 &: \sigma^2 \geqslant \sigma_0^2 \\ H_1 &: \sigma^2 < \sigma_0^2. \end{aligned}$$

Test-statistika je ponovno $T = (n-1)S_n^2/\sigma_0^2 \sim \chi^2(n-1)$. Kritično područje je dano s $[0, \chi_\alpha^2]$ (vidi Sliku 8.8), gdje je χ_α^2 α-ti kvantil od $\chi^2(n-1)$.



Slika 8.8: Kritično područje kod lijevog jednostranog testa

Nul-hipotezu ćemo odbaciti ukoliko vrijednost test-statistike

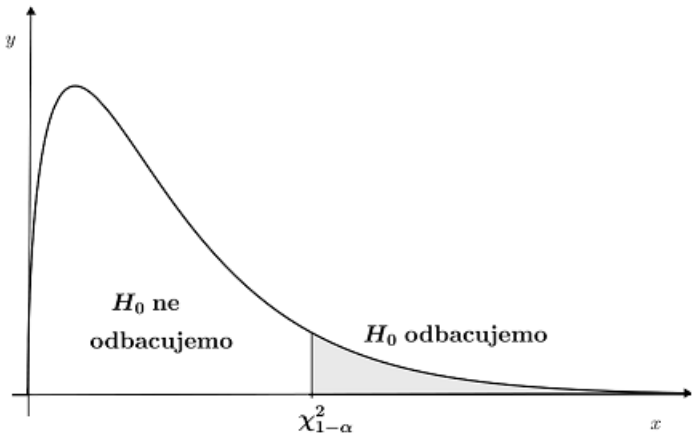
$$t = \frac{n-1}{\sigma_0^2} s_n^2$$

za realizaciju $(x_1, \ldots, x_n)$ slučajnog uzorka $(X_1, \ldots, X_n)$ pripada kritičnom području. Ponovno, $\alpha \cdot 100\%$ realizacija od $(X_1, \ldots, X_n)$ uzrokuje odbacivanje H_0 ako je H_0 istinita.

(iii) **Desni jednostrani test** uz nivo značajnosti $\alpha \in (0, 1)$. Testiramo:

$$\begin{aligned} H_0 &: \sigma^2 \leqslant \sigma_0^2 \\ H_1 &: \sigma^2 > \sigma_0^2. \end{aligned}$$

Test-statistika je $T = (n-1)S_n^2/\sigma_0^2 \sim \chi^2(n-1)$, a kritično područje je $[\chi^2_{1-\alpha}, \infty)$ (vidi Sliku 8.9), gdje je $\chi^2_{1-\alpha}$ $(1-\alpha)$-ti kvantil od $\chi^2(n-1)$.



Slika 8.9: Kritično područje kod desnog jednostranog testa

Nul-hipotezu ćemo odbaciti ukoliko vrijednost test-statistike

$$t = \frac{n-1}{\sigma_0^2} s_n^2$$

za realizaciju $(x_1, \ldots, x_n)$ slučajnog uzorka $(X_1, \ldots, X_n)$ pripada kritičnom području. Dakle, $\alpha \cdot 100\%$ realizacija od $(X_1, \ldots, X_n)$ uzrokuje odbacivanje H_0 ako je H_0 istinita.

Primjer 8.6. Na vrećama cementa deklarirano je da je sadržaj vreća distribuiran kao $N(50\,\text{kg}, 0.01\,\text{kg}^2)$. Na slučajnom uzorku od 100 vreća ustanovljeno je da je u njima prosječno 49 kg uz standardnu devijaciju od 0.15 kg. Možemo li na temelju tog nalaza optužiti proizvođača da vara kupce? Dakle, $X \sim N(\mu\,\text{kg}, \sigma^2\,\text{kg}^2)$, $\mu_0 = 50$ kg, $\sigma_0^2 = 0.01$ kg^2, $n = 100$ (uzorak je velik), $x_{100} = 49$ kg i $s_{100} = 0.15$ kg. Uzmimo nivo značajnosti $\alpha = 0.05$. Prvo testiramo

$$\begin{aligned} H_0 &: \mu = 50 \\ H_1 &: \mu \neq 50. \end{aligned}$$

Test-statistika je

$$T = 10\frac{\overline{X}_{100} - 50}{S_{100}} \sim t(99).$$

Kritično područje je $(-\infty, -t_{0.975}] \cup [t_{0.975}, \infty)$. Zbog veličine uzorka, $t(99)$ možemo zamijeniti (aproksimirati) s $N(0,1)$. Sada dobivamo da je $t_{0.975} \approx$

1.96. Konačno, kako vrijednost pripadne test-statistike $10(49-50)/0.15$ pripada kritičnom području, odbacujemo H_0 s mogućnošću pogreške od 5%. Dalje, testiramo

$$\begin{aligned} H_0 &: \sigma^2 = 0.01 \\ H_1 &: \sigma^2 \neq 0.01. \end{aligned}$$

Test-statistika je

$$T = \frac{99}{0.01} S_{100}^2 \sim \chi^2(99),$$

a kritično područje je $[0, \chi^2_{0.025}] \cup [\chi^2_{0.975}, \infty)$. Ponovo, kako je uzorak velik $\chi^2(99)$ možemo aproksimirati s $N(99, 198)$. Sada imamo $\chi^2_{0.025} = 71.42$ i $\chi^2_{0.975} = 126.58$. Koačno, kako vrijednost pripadne test-statistike $0.0225 \cdot 99/0.01$upada u kritično područje, odbacujemo H_0 s mogućnošću pogreške od 5%. Dakle, možemo proizvođača optužiti za prijevaru. □

8.3.3 Uspoređivanje očekivanja dviju slučajnih varijabli

Pretpostavimo da želimo ispitati razlikuje li se neko statističko obilježje u dvije različite populacije, kao u primjeru s početka ove točke. Dakle, imamo dvije slučajne varijable $X^{(1)}$ i $X^{(2)}$, njihova očekivanja μ_1 i μ_2 te varijance σ_1^2 i σ_2^2. Zanima nas postoji li razlika u očekivanjima ove dvije slučajne varijable.

(a) Neka su $(X_1^{(1)}, \ldots, X_{n_1}^{(1)})$ i $(X_1^{(2)}, \ldots, X_{n_2}^{(2)})$ slučajni uzorci (koji su međusobno nezavisni) iz, redom, $N(\mu_1, \sigma_1^2)$ i $N(\mu_2, \sigma_2^2)$.

(i) Pretpostavimo da su σ_1^2 i σ_2^2 poznate. Radimo **dvostrani test** uz nivo značajnost $\alpha \in (0, 1)$. Testiramo:

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 \\ H_1 &: \mu_1 \neq \mu_2. \end{aligned}$$

U ovom slučaju test-statistika je

$$T = \frac{\overline{X}_{n_1}^{(1)} - \overline{X}_{n_2}^{(2)}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$$

Nije teško vidjeti da je $\overline{X}_{n_1}^{(1)} - \overline{X}_{n_2}^{(2)} \sim N(0, (\sigma_1^2/n_1 + \sigma_2^2/n_2))$. Dakle, $T \sim N(0, 1)$. Kritično područje je dano s $(-\infty, -z_{1-\frac{\alpha}{2}}] \cup$

$[z_{1-\frac{\alpha}{2}}, \infty)$, gdje je $z_{1-\frac{\alpha}{2}}$ $(1-\alpha/2)$-ti kvantili od $N(0,1)$. Nul-hipotezu ćemo odbaciti ukoliko vrijednost test-statistike

$$t = \frac{\overline{x}_{n_1}^{(1)} - \overline{x}_{n_2}^{(2)}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

za realizacije $(x_1^{(1)}, \ldots, x_{n_1}^{(1)})$ i $(x_1^{(2)}, \ldots, x_{n_2}^{(2)})$ od, redom, $(X_1^1, \ldots, X_{n_1}^{(1)})$ i $(X_1^{(2)}, \ldots, X_{n_2}^{(2)})$ pripada kritičnom području. Dakle, $\alpha \cdot 100\%$ realizacija od $(X_1, \ldots, X_n)$ uzrokuje odbacivanje H_0 ako je H_0 istinita. Napomenimo da analogno kao i gore vršimo lijevi i desni jednostrani test, pri čemu $z_{1-\frac{\alpha}{2}}$ zamijenjujemo s $z_{1-\alpha}$, gdje je $z_{1-\alpha}$ $(1-\alpha)$-ti kvantil od $N(0,1)$.

(ii) Pretpostavimo da je $\sigma_1^2 = \sigma_2^2 = \sigma^2$ i da σ^2 nije poznata. Radimo **dvostrani test** uz nivo značajnost $\alpha \in (0,1)$. Testiramo:

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 \\ H_1 &: \mu_1 \neq \mu_2. \end{aligned}$$

U ovom slučaju test-statistika je

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

gdje je

$$S_D = \sqrt{\frac{(n_1-1)(S_{n_1}^{(1)})^2 + (n_2-1)(S_{n_2}^{(2)})^2}{n_1+n_2-2}}.$$

Može se pokazati da je $T \sim t(n_1+n_2-2)$. Kritično područje je dano s $(-\infty, -t_{1-\frac{\alpha}{2}}] \cup [t_{1-\frac{\alpha}{2}}, \infty)$, gdje je $t_{1-\frac{\alpha}{2}}$ $(1-\alpha/2)$-ti kvantili od $t(n_1+n_2-2)$. Nul-hipotezu ćemo odbaciti ukoliko vrijednost test-statistike

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

za realizacije $(x_1^{(1)}, \ldots, x_{n_1}^{(1)})$ i $x_1^{(2)}, \ldots, x_{n_2}^{(2)}$ od, redom, $(X_1^{(1)}, \ldots, X_{n_1}^{(1)})$ i $(X_1^{(2)}, \ldots, X_{n_2}^{(2)})$ pripada kritičnom području. Ovdje je

$$s_D = \sqrt{\frac{(n_1-1)(s_{n_1}^{(1)})^2 + (n_2-1)(s_{n_2}^{(2)})^2}{n_1+n_2-2}}.$$

Dakle, $\alpha \cdot 100\%$ realizacija od $(X_1, \ldots, X_n)$ uzrokuje odbacivanje H_0 ako je H_0 istinita. Napomenimo da analogno kao i gore provodimo lijevi i desni jednostrani test, pri čemu $t_{1-\frac{\alpha}{2}}$ zamijenjujemo s $t_{1-\alpha}$, gdje je $t_{1-\alpha}$ $(1-\alpha)$-ti kvantil od $t(n_1+n_2-2)$.

(b) Imamo velike (u praksi $n_1, n_2 > 30$) slučajne uzorke $(X_1^{(1)}, \ldots, X_{n_1}^{(1)})$ i $(X_1^{(2)}, \ldots, X_{n_2}^{(2)})$ (koji su međusobno nezavisni) iz raspodjela koje ne moraju biti normalne.

(i) Pretpostavimo da su σ_1^2 i σ_2^2 poznate. Radimo **dvostrani test** uz nivo značajnost $\alpha \in (0,1)$. Testiramo:

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 \\ H_1 &: \mu_1 \neq \mu_2. \end{aligned}$$

U ovom slučaju test-statistika je

$$T = \frac{\overline{X}_{n_1}^{(1)} - \overline{X}_{n_2}^{(2)}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$$

Iz centralnog graničnog teorema možemo zaključiti da zbog veličine uzoraka raspodjelu od T možemo aproksimirati s $N(0,1)$. Kritično područje je dano s $(-\infty, -z_{1-\frac{\alpha}{2}}] \cup [z_{1-\frac{\alpha}{2}}, \infty)$, gdje je $z_{1-\frac{\alpha}{2}}$ $(1-\alpha/2)$-ti kvantili od $N(0,1)$. Nul-hipotezu ćemo odbaciti ukoliko vrijednost test-statistike

$$t = \frac{\overline{x}_{n_1}^{(1)} - \overline{x}_{n_2}^{(2)}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

za realizacije $(x_1^{(1)}, \ldots, x_{n_1}^{(1)})$ i $(x_1^{(2)}, \ldots, x_{n_2}^{(2)})$ od, redom, $(X_1^{(1)}, \ldots, X_{n_1}^{(1)})$ i $(X_1^{(2)}, \ldots, X_{n_2}^{(2)})$ pripada kritičnom području. Dakle, $\alpha \cdot 100\%$ realizacija od $(X_1, \ldots, X_n)$ uzrokuje odbacivanje H_0 ako je H_0 istinita. Napomenimo da analogno kao i gore vršimo lijevi i desni jednostrani test, pri čemu $z_{1-\frac{\alpha}{2}}$ zamijenjujemo s $z_{1-\alpha}$, gdje je $z_{1-\alpha}$ $(1-\alpha)$-ti kvantil od $N(0,1)$.

(ii) Pretpostavimo da σ_1^2 i σ_2^2 nisu poznate. Radimo **dvostrani test** uz nivo značajnost $\alpha \in (0,1)$. Testiramo:

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 \\ H_1 &: \mu_1 \neq \mu_2. \end{aligned}$$

U ovom slučaju test-statistika je

$$T = \frac{\overline{X}^{(1)}_{n_1} - \overline{X}^{(2)}_{n_2}}{\sqrt{\frac{(S^{(1)}_{n_1})^2}{n_1} + \frac{(S^{(2)}_{n_1})^2}{n_2}}}.$$

Ponovno, zbog veličine uzoraka, iz centralnog graničnog teorema možemo zaključiti da raspodjelu od T možemo aproksimirati s $N(0,1)$. Kritično područje je dano s $(-\infty, -z_{1-\frac{\alpha}{2}}] \cup [z_{1-\frac{\alpha}{2}}, \infty)$, gdje je $z_{1-\frac{\alpha}{2}}$ $(1-\alpha/2)$-ti kvantil od $N(0,1)$. Nul-hipotezu ćemo odbaciti ukoliko vrijednost test-statistike

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

za realizacije $(x^{(1)}_1, \ldots, x^{(1)}_{n_1})$ i $(x^{(2)}_1, \ldots, x^{(2)}_{n_2})$ od, redom, $(X^{(1)}_1, \ldots, X^{(1)}_{n_1})$ i $(X^{(2)}_1, \ldots, X^{(2)}_{n_2})$ pripada kritičnom području. Dakle, $\alpha \cdot 100\%$ realizacija od $(X_1, \ldots, X_n)$ uzrokuje odbacivanje H_0 ako je H_0 istinita. Napomenimo da analogno kao i gore provodimo lijevi i desni jednostrani test, pri čemu $z_{1-\frac{\alpha}{2}}$ zamijenjujemo s $z_{1-\alpha}$, gdje je $z_{1-\alpha}$ $(1-\alpha)$-ti kvantil od $N(0,1)$.

Koristeći gore navedene rezultate, primjenom centralnog graničnog teorema, možemo testirati hipoteze o razlici parametara Bernoullijevih raspodjela

$$\begin{pmatrix} 0 & 1 \\ 1-p_1 & p_1 \end{pmatrix} \qquad \text{i} \qquad \begin{pmatrix} 0 & 1 \\ 1-p_2 & p_2 \end{pmatrix}.$$

Neka su $(X^{(1)}_1, \ldots, X^{(1)}_{n_1})$ i $(X^{(2)}_1, \ldots, X^{(2)}_{n_2})$, redom, veliki slučajni uzorci iz gornjih raspodjela. Kako je $\mathbb{E}(X^{(1)}_i) = p_1$, $\mathrm{Var}(X^{(1)}_i) = p_1(1-p_1)$, $\mathbb{E}(X^{(2)}_i) = p_2$ i $\mathrm{Var}(X^{(2)}_i) = p_2(1-p_2)$, za test-statistiku uzimamo

$$T = \frac{\overline{X}^{(1)}_{n_1} - \overline{X}^{(2)}_{n_2}}{\sqrt{\frac{(S^{(1)}_{n_1})^2}{n_1} + \frac{(S^{(2)}_{n_1})^2}{n_2}}}.$$

Uočimo da vrijedi

$$(S^{(1)}_{n_1})^2 = \frac{n_1 \overline{X}^{(1)}_{n_1}(1 - \overline{X}^{(1)}_{n_1})}{n_1 - 1}, \qquad (s^{(1)}_{n_1})^2 = \frac{n_1 \overline{x}^{(1)}_{n_1}(1 - \overline{x}^{(1)}_{n_1})}{n_1 - 1},$$

$$(S_{n_2}^{(2)})^2 = \frac{n_2 \overline{X}_{n_2}^{(2)}(1-\overline{X}_{n_2}^{(2)})}{n_2-1} \qquad \text{i} \qquad (s_{n_2}^{(2)})^2 = \frac{n_2 \overline{x}_{n_2}^{(2)}(1-\overline{x}_{n_2}^{(2)})}{n_2-1}$$

pa je

$$T = \frac{\overline{X}_{n_1}^{(1)} - \overline{X}_{n_2}^{(2)}}{\sqrt{\frac{\overline{X}_{n_1}^{(1)}(1-\overline{X}_{n_1}^{(1)})}{n_1-1} + \frac{\overline{X}_{n_2}^{(2)}(1-\overline{X}_{n_2}^{(2)})}{n_2-1}}} \qquad \text{i} \qquad t = \frac{\overline{x}_{n_1}^{(1)} - \overline{x}_{n_2}^{(2)}}{\sqrt{\frac{\overline{x}_{n_1}^{(1)}(1-\overline{x}_{n_1}^{(1)})}{n_1-1} + \frac{\overline{x}_{n_2}^{(2)}(1-\overline{x}_{n_2}^{(2)})}{n_2-1}}}.$$

Sada možemo testirati hipoteze:

(i) Dvostrani test

$$\begin{aligned} H_0 &: p_1 = p_2 \\ H_1 &: p_1 \neq p_2. \end{aligned}$$

(ii) Lijevi jednostrani test

$$\begin{aligned} H_0 &: p_1 \geqslant p_2 \\ H_1 &: p_1 < p_2. \end{aligned}$$

(iii) Desni jednostrani test

$$\begin{aligned} H_0 &: p_1 \leqslant p_2 \\ H_1 &: p_1 > p_2. \end{aligned}$$

Primjer 8.7. Vratimo se na primjer u kojem su izmjerene tlačne čvrstoće betona s čeličnim vlaknima i betona bez čeličnih vlakana. Podaci su prikazani u Tablici 8.2.

beton s čeličnim vlaknima	68	70	62	65	70	66	67
beton bez čeličnih vlakana	63	61	59	64	62	66	60

Tablica 8.2: Podaci o tlačnoj čvrstoći različitih vrsta betona

Pretpostavimo da su tlačne čvrstoće koje mjerimo normalno distribuirane s raspodjelama $N(\mu_1, \sigma^2)$ (za beton s čeličnim vlaknima) i $N(\mu_2, \sigma^2)$ (za beton bez čeličnih vlakana). Na osnovu gornjih mjerenja zaključujemo da je $\overline{x}_7^{(1)} = 66.8571, s_7^{(1)} = 2.8535$, $\overline{x}_7^{(2)} = 62.1428$ i $s_7^{(2)} = 2.4103$. Budući da se radi o malim slučajnim uzorcima iz normalnih raspodjela s jednakom varijancom koja nije poznata, test-statistika je

$$T = \sqrt{7}\frac{\overline{X}_7^{(1)} - \overline{X}_7^{(2)}}{S_D},$$

gdje je

$$S_D = \sqrt{\frac{(S_7^{(1)})^2 + (S_7^{(2)})^2}{2}}.$$

Vrijednost test-statistike za gornje realizacije iznosi $t = 3.3391$. Kritično područje, uz nivo značajnosti $\alpha = 0.05$, je dano s $(-\infty, -t_{0.975}] \cup [t_{0.975}, \infty) = (-\infty, -2.18] \cup [2.18, +\infty)$, gdje je $t_{0.975}$ 0.975-ti kvantili od $t(12)$. Dakle, budući da t pripada kritičnom području, odbacujemo nul-hipotezu da su očekivane vrijednosti jednake s mogućnošću pogreške od 5%. Praktično interpretirajući, postoji statistički značajna razlika u tlačnoj čvrstoći betona s čeličnim vlaknima i betona bez tih vlakana. □

Za detaljniju diskusiju o svim gore spomenutim rezultatima čitatelju preporučamo reference [5, 6, 7, 8, 10].

Poglavlje 9

Dvodimenzionalne varijable

Mnogi inženjerski i znanstveni problemi uključuju analizu zavisnosti dviju varijabli. Na primjer, tlačna čvrstoća betona ovisi o omjeru vode i cementa. Ako mješavina ima previše vode, ne očekujemo da će beton biti čvrst. Stoga možemo predložiti neki model koji bi konkretnije opisao tu zavisnost. Taj model je nedeterministički, a ne deterministički, tj. pojavljivat će se greška u predviđanju jedne varijable na temelju druge (tzv. prediktora), jer npr. na čvrstoću betona utječu i neke druge varijable koje je teško ili nemoguće opservirati (izmjeriti). Čak i kad bismo u predviđanje uključili sve varijable koje možemo zamisliti, postojala bi greška mjerenja. Međutim, ako je ta greška slučajna i nesistematična, modeli koje dobivamo bit će iskoristivi, čak i kad radimo samo s jednom varijablom kao prediktorom.

U pozadini ovakvog predviđanja je zavisnost dviju varijabli. Postoje različiti tipovi zavisnosti među varijablama i zavisnost možemo opisivati na više načina. Mi ćemo se ovdje fokusirati na linearnu zavisnost. U tu svrhu uvest ćemo pojmove Pearsonovog koeficijenta korelacije (koji nam daje informaciju o tome koliko je smisleno funkcijsku vezu među dvjema varijablama aproksimirati linearnom funkcijom) i linearne regresije (pomoću koje vršimo procjenu koeficijenata te linearne funkcije).

9.1 Pearsonov koeficijent korelacije

Dvodimenzionalna varijabla je uređeni par dviju varijabli: (X, Y). Opažene vrijednosti (podatke dobivene mjerenjem) od (X, Y) na uzorku veličine n, $(x_1, y_1), \ldots, (x_n, y_n)$, obično prikazujemo tablicom:

X	x_1	x_2	x_3	$\ldots$	x_n
Y	y_1	y_2	y_3	$\ldots$	y_n

Tablica 9.1: Tablični prikaz opaženih vrijednosti od (X, Y)

Cilj nam je donijeti zaključak o potenijalnoj zavisnosti (ili procijeniti jednu mjeru zavisnosti) varijabli X i Y. Međutim, napomenimo da prilikom procijenjivanja mjere zavisnosti između dvije varijable treba biti oprezan. Nema smisla procijenjivati mjeru zavisnosti između bilo koje dvije varijable, moramo imati teoretsku podlogu. Jedna od najčešćih pogrešaka je procijenjivanje mjere zavisnosti između varijabli koje imaju zajednički uzrok.

Jedan od načina mjerenja zavisnosti dviju varijabli, analogno kao i u teoriji vjerojatnosti, je preko koeficijenta korelacije. **Pearsonov koeficijent korelacije** opaženih vrijednosti $(x_1, y_1), \ldots, (x_n, y_n)$ od (X, Y) definiramo kao

$$r_{XY} = \frac{S_{XY}}{\sqrt{S_{XX}}\sqrt{S_{YY}}},$$

gdje su

$$S_{XX} = \sum_{i=1}^{n}(x_i - \overline{x}_n)^2, \quad S_{YY} = \sum_{i=1}^{n}(y_i - \overline{y}_n)^2, \quad S_{XY} = \sum_{i=1}^{n}(x_i - \overline{x}_n)(y_i - \overline{y}_n).$$

Ovaj broj možemo interpretirati kao procjenu koeficijenta korelacije varijabli X i Y. Uočimo da vrijedi

$$-1 \leq r_{XY} \leq 1.$$

Da bi to pokazali, prvo se prisjetimo da je standardni skalarni produkt na n-dimenzionalnom Euklidskom prostoru $\mathbb{R}^n$ definiran s

$$\langle x, y\rangle = \sum_{i=1}^{n} x_i y_i, \qquad x, y \in \mathbb{R}^n,$$

($x = (x_1, \ldots, x_n)$ i $y = (y_1, \ldots, y_n)$). Nadalje, dobro poznata Cauchy-Schwartz-Bunjakovski nejednakost kaže da za svaki skalarni produkt vrijedi

$$|\langle x, y\rangle| \leqslant \sqrt{|\langle x, x\rangle|}\sqrt{|\langle y, y\rangle|}.$$

Dakle, kako S_{XY}, S_{XX} i S_{YY} predstavljaju, redom, skalarne produkte između vektora $x - \overline{x}_n$ i $y - \overline{y}_n$, $x - \overline{x}_n$ i $x - \overline{x}_n$ te $y - \overline{y}_n$ i $y - \overline{y}_n$, tvrdnja slijedi. Nadalje, uočimo da Pearsonovim koeficijentom korelacije zapravo mjerimo stupanj linearne zavisnosti od opažanja $x_1, \ldots, x_n$ i $y_1, \ldots, y_n$. Naime, dobro je poznato da nejednakost Cauchy-Schwartz-Bunjakovski postaje jednakost ako, i samo ako, su vektori x i y kolinearni, tj. ako postoji $\lambda \in \mathbb{R}$ t.d. $y = \lambda x$. Sada zaključujemo

(i) $|r_{XY}| = 1$ ako, i samo ako, $S_{XX} > 0$, $S_{YY} > 0$ i opažanja $(x_1, y_1), \ldots, (x_n, y_n)$ leže na istom pravcu

(ii) manji $|r_{XY}|$ znači "manji stupanj linearne zavisnosti" od opažanja $x_1, \ldots, x_n$ i $y_1, \ldots, y_n$

(iii) $r_{XY} = 0$ znači da su $x - \overline{x}_n$ i $y - \overline{y}_n$ "ortogonalni".

Nadalje, slično kao i u teoriji vjerojatnosti,

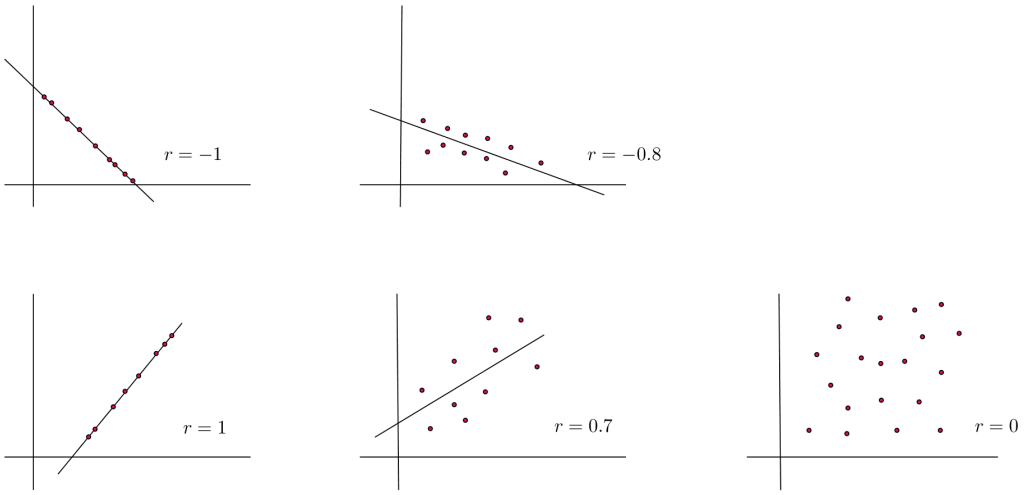
(i) ako je $r_{XY} < 0$, kažemo da su opažene vrijednosti $x_1, \ldots, x_n$ i $y_1, \ldots, y_n$ **negativno korelirane**.

(ii) ako je $r_{XY} > 0$, kažemo da su opažene vrijednosti $x_1, \ldots, x_n$ i $y_1, \ldots, y_n$ **pozitivno korelirane**.

(iii) ako je $r_{XY} = 0$, kažemo da su opažene vrijednosti $x_1, \ldots, x_n$ i $y_1, \ldots, y_n$ **nekorelirane**.

Pozitivna koreliranost znači da opažene vrijednosti $x_1, \ldots, x_n$ i $y_1, \ldots, y_n$ imaju sklonost odstupanja od svojih aritmetičkih sredina u istu stranu, a u slučaju negativne koreliranosti imaju sklonost odstupanja na različite strane od svojih aritmetičkih sredina.

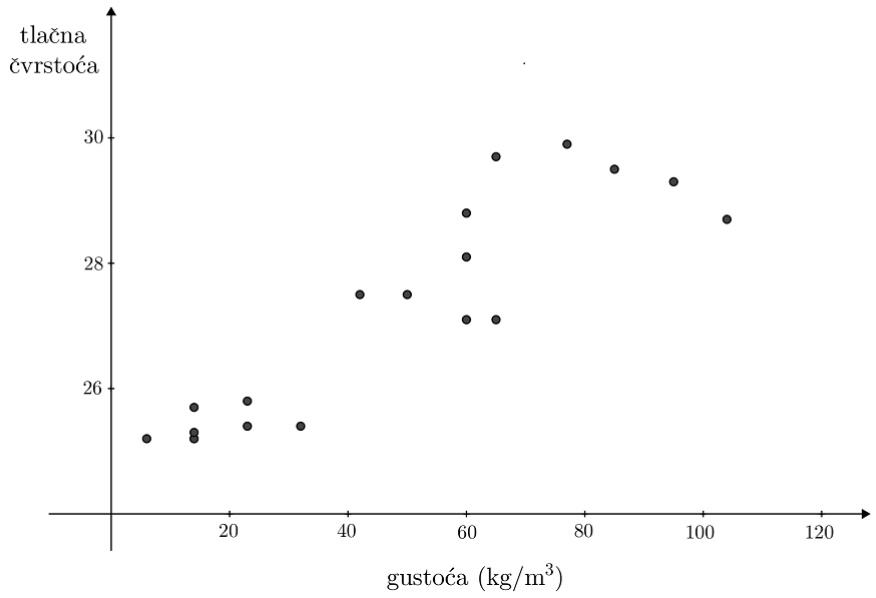
Kad grafički skiciramo mjerenja, možemo dobiti razne dijagrame raspršenja za različite Pearsonove koeficijente korelacije (vidi Sliku 9.1).



Slika 9.1: Primjeri podataka za različite Pearsonove koeficijente korelacije

Dakle, što je Pearsonov koeficijent korelacije veći (po apsolutnoj vrijednosti), možemo zaključiti da je veći stupanj linearne zavisnosti između opaženih vrijednosti $x_1, \ldots, x_n$ i $y_1, \ldots, y_n$ pa samim time i varijabli X i Y.

Primjer 9.1. Komentirajmo sada Primjer 6.2 s početka ovog poglavlja. Ako podatke iz Tablice 6.1 prikažemo grafički, tako da svaki par mjerenja odgovara jednoj točki na grafu (pri čemu je na x-osi gustoća, a na y-osi tlačna čvrstoća), dobivamo dijagram raspršenja prikazan na Slici 9.2. Na tom dijagramu jasno možemo uočiti da, iako točke ne leže na pravcu, beton veće gustoće obično ima i veću tlačnu čvrstoću.



Slika 9.2: Dijagram raspršenja podataka (gustoća i tlačna čvrstoća betona iz Tablice 6.1)

Računamo Pearsonov koeficijent korelacije

$$r_{XY} = \frac{S_{XY}}{\sqrt{S_{XX}}\sqrt{S_{YY}}} = \frac{858.7105}{\sqrt{16494.7368 \cdot 55.3516}} = 0.8987.$$

Ovako veliki Pearsonov koeficijent korelacije znači da su gustoća i tlačna čvrstoća betona jako pozitivno korelirane i potvrđuje naš dojam da beton veće gustoće ima i veću tlačnu čvrstoću.

Važno je znati da ovakav rezultat ne treba olako interpretirati, npr. iz njega ne možemo zaključiti da veća gustoća betona uzrokuje veću tlačnu čvrstoću. Korelacija nam govori o (linearnoj) zavisnosti varijabli, ali ne i o uzročno-posljedičnoj vezi. Moguće je da neka treća varijabla, npr. različita vrsta korištenog cementa, uzrokuje i veću gustoću i veću tlačnu čvrstoću dobivenog betona. □

9.2 Linearna regresija

Koristeći Pearsonov koeficijent korelacije možemo odrediti stupanj linearne zavisnosti dviju varijabli. Ako želimo bolje razumijeti tu zavisost, u tu svrhu nam pomaže linearna regresija. **Regresijska analiza** je statistički proces analize odnosa dviju ili više varijabli. Za razliku od Pearsonovog koeficijenta korelacije gdje varijable X i Y tretiramo "ravnopravno" – simetrično, u regresijskoj analizi pretpostavljamo odnos između varijabli, tj. varijable dijelimo na zavisnu varijablu i jednu ili više nezavisnih varijabli. Pod takvom pretpostavkom, regresijska analiza nam pomaže razumijeti kako se mijenja vrijednost zavisne varijable pri promijeni bilo koje od nezavisnih varijabli. Iako se u praktičnim problemima češće pojavljuju modeli u kojima se varijacije zavisne varijable opisuju ponašanjem više nezavisnih varijabli, mi ćemo se usredotočiti na model **jednostavne regresije** u kojoj se pretpostavlja povezanost između zavisne varijable Y i jedne nezavisne varijable X. Varijablu X smatramo neslučajnom te ćemo ju u nastavku označavati s x. Dakle, njezine vrijednosti na elementima populacije znamo (dob, spol, gustoća, viskoznost, ...). S druge strane, na istoj populaciji imamo i varijablu Y koja je u pravilu slučajna. Nas će zanimati informacija o raspodjeli od Y za fiksni x, tj. $\mathbb{P}(Y \in \cdot | x = x_i)$. U tu svrhu uvodimo oznaku Y_x. Primjerice, ako x označava težinu ispitanika, a Y vrijednost kolesterola u krvi, Y_x označava vrijednost kolesterola u krvi kod ispitanika s težinom x. Regresijska metoda pretpostavlja da možemo uspostaviti funkcijsku vezu između Y_x i x:

$$Y_x = f(x)$$

za neku funkciju $f : \mathbb{R} \to \mathbb{R}$. Međutim, uočimo da to nema baš smisla jer Y_x je u pravilu slučajna varijabla. Dakle, pretpostavljamo vezu oblika

$$Y_x = f(x) + \varepsilon_x,$$

gdje je $f : \mathbb{R} \to \mathbb{R}$ neka funkcija a varijabla ε_x je slučajna (nemjerljiva) i predstavlja pogrešku koju akumulira utjecaj ostalih faktora na varijablu Y_x. Dakle, uz dani x, ε_x određuje svojstva varijable Y_x. Prirodno je za pretpostaviti da je $\varepsilon_x \sim N(0, \sigma_x^2)$. Fiksirajmo $x_1, \ldots, x_n$. Želimo odrediti prirodu ovisnosti od Y_{x_i} o x_i. U nastavku ćemo umjesto Y_{x_i} i ε_{x_i} pisati samo Y_i i ε_i. Dakle, imamo

$$Y_i = f(x_i) + \varepsilon_i, \quad i = 1, \ldots, n.$$

U nastavku pretpostavljamo:

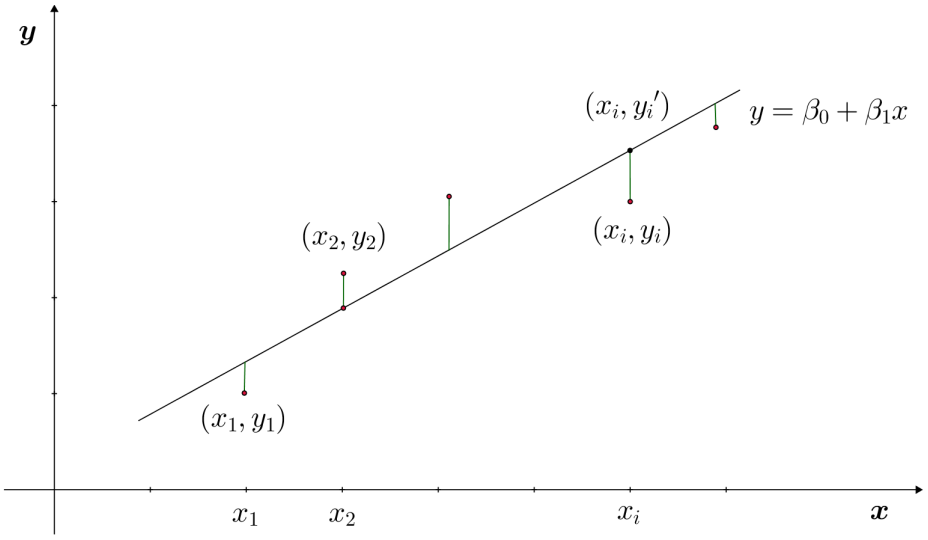
(i) $f(x) = \beta_0 + \beta_1 x$, tj. usredotočit ćemo se samo na model **jednostavne linearne regresije** u kojoj se pretpostavlja linearna povezanost između Y_x i x.

(ii) varijable $\varepsilon_1, \ldots, \varepsilon_n$ su nezavisne i jednako distribuirane s raspodjelom $N(0, \sigma^2)$.

Uočimo,

$$\mathbb{E}(Y_i) = \mathbb{E}(\beta_0 + \beta_1 x_i + \varepsilon_i) = \beta_0 + \beta_1 x_i$$
$$\mathrm{Var}(Y_i) = \mathrm{Var}(\beta_0 + \beta_1 x_i + \varepsilon_i) = \sigma^2.$$

Koeficijenti β_0 i β_1 su nepoznati **regresijski parametri** koje ćemo u postupku modeliranja procijeniti na osnovi opažanja. Neka je $y_1, \ldots, y_n$ realizacija od $Y_1, \ldots, Y_n$. Označimo li s $y_i' = \mathbb{E}(Y_i)$, iz gornje relacije vidimo da na pravcu $y = \beta_0 + \beta_1 x$ zapravo leže parovi (x_i, y_i'). Dakle, na temelju postavljenog modela, vrijednost koju bismo očekivali izmjeriti za određeni x_i bi bila y_i', dok je stvarna vrijednost koju smo izmjerili y_i. Želimo li kroz parove vrijednosti (x_i, y_i) provući pravac koji bi najbolje pristajao svim podacima (bio "najbliži" točkama (x_i, y_i)) morat ćemo minimizirati udaljenost između y_i i y_i'.



Slika 9.3: Pravac najboljeg pristajanja

Na toj ideji se zasniva i najčešće korištena metoda u procjeni regresijskih parametara – **metoda najmanjih kvadrata**. Preciznije, ideja metode je odrediti parametre $\hat{\beta}_0$ i $\hat{\beta}_1$ za koje je suma kvadrata odstupanja pogrešaka minimalna, tj.

$$\sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2 \to \min.$$

Definirajmo funkciju

$$D(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

i pronađimo njezin minimum. Nužni uvjeti za ekstrem funkcije $D(\beta_0, \beta_1)$ su

$$\frac{\partial D}{\partial \beta_0} = 0 \qquad \text{i} \qquad \frac{\partial D}{\partial \beta_1} = 0.$$

Dakle, parcijalnim deriviranjem funkcije $D(\beta_0, \beta_1)$ i izjednačavanjem s nulom dolazimo do sljedećeg sustava jednadžbi

$$\begin{aligned} \frac{\partial D}{\partial \beta_0} &= 2 \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i)) (-1) = 0, \\ \frac{\partial D}{\partial \beta_1} &= 2 \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i)) (-x_i) = 0. \end{aligned}$$

Lagano se vidi da su prethodne relacije ekvivalentne s

$$\begin{aligned} &\sum_{i=1}^{n} y_i - n\beta_0 - \beta_1 \sum_{i=1}^{n} x_i = 0, \\ &\sum_{i=1}^{n} y_i x_i - \beta_0 \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 = 0. \end{aligned}$$

Rješavanjem prethodnog sustava dolazimo do traženih procjena $\hat{\beta}_0$, $\hat{\beta}_1$ parametara β_0, β_1:

$$\hat{\beta}_0 = \overline{y}_n - \overline{x}_n \hat{\beta}_1, \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n \overline{x}_n \overline{y}_n}{\sum_{i=1}^{n} x_i^2 - n \overline{x}_n^2}.$$

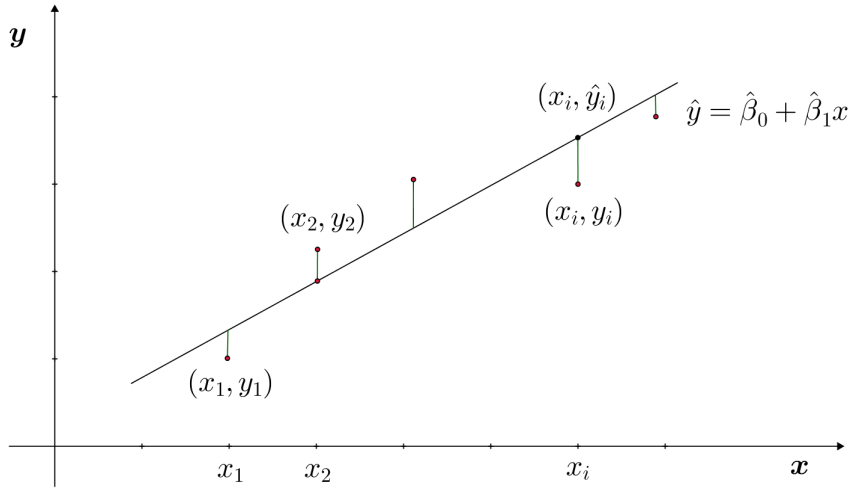
Izraz za $\hat{\beta}_1$ možemo napisati i u sljedećem obliku (koristeći već poznate oznake):

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \overline{x}_n)(y_i - \overline{y}_n)}{\sum_{i=1}^{n} (x_i - \overline{x}_n)^2} = \frac{S_{XY}}{S_{XX}}.$$

Iako bismo formalno trebali provjeriti o kojem se ekstremu radi, iz prirode problema jasno je da su dobiveni koeficijenti točka minimuma funkcije $D(\beta_0, \beta_1)$. Dakle, procijenjena jednadžba regresijskog pravca $y = \beta_0 + \beta_1 x$ glasi:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x.$$

Za navedeni pravac procijenjena vrijednost $\hat{y}_i$ za odgovarajući x_i iznosi $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, dok odstupanje od izmjerene (stvarne) vrijednosti, $y_i - \hat{y}_i$, nazivamo **rezidualnim odstupanjem**. Na vrijednosti $y_1 - \hat{y}_1, \ldots, y_n - \hat{y}_n$ možemo gledati kao na procjenu ili realizacije grešaka $\varepsilon_1, \ldots, \varepsilon_n$. Rezidualna odstupanja ilustrirana su na Slici 9.4.



Slika 9.4: Regresijski pravac i reziduali

Nakon postavljanja modela i procjene parametara, postavlja se pitanje adekvatnosti modela, tj. njegove sposobnosti da na temelju nezavisne varijable x objasni kretanje zavisne varijable Y. U svrhu određivanja reprezentativnosti modela najčešće koristimo **koeficijent determinacije**:

$$r_{XY}^2 = \frac{S_{XY}^2}{S_{XX} S_{YY}}, \qquad 0 \leq r_{XY}^2 \leq 1.$$

Nije teško vidjeti da vrijedi

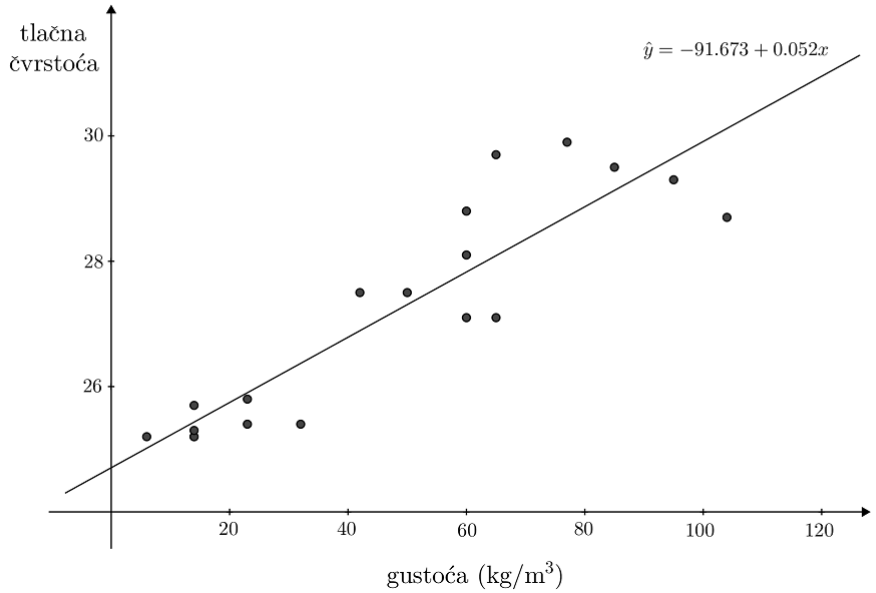
$$r_{XY}^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \overline{y})^2}{\sum_{i=1}^n (y_i - \overline{y})^2}.$$

Uočimo, što je r_{XY}^2 bliži 1, to je model reprezentativniji.

Primjer 9.2. Vratimo se na Primjer 6.2 u kojem su mjerene gustoća i tlačna čvrstoća betona i odredimo regresijski pravac. Računamo koeficijente

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} = \frac{858.7105}{16494.7368} = 0.052,$$
$$\hat{\beta}_0 = \overline{y}_{19} - \hat{\beta}_1 \overline{x}_{19} = 27.2 - 0.052 \cdot 2277.5263 = -91.3673.$$

Dakle, pravac koji najbolje pristaje izmjerenim podacima je $y = -91.673 + 0.052x$, gdje je x gustoća, a y tlačna čvrstoća.



Slika 9.5: Regresijski pravac za podatke o gustoći i tlačnoj čvrstoći betona iz Primjera 6.2

Primjer 9.3. U jednom istraživanju (Neyman, "The influence of oil properties on the fretting wear of mild steel", *Wear*, Vol. 152, 1992, str. 171-181) dobiveni su podaci o trošenju mekog čelika (zavisna varijabla, 10^{-4} kubičnih mililitara) i viskoznosti ulja (nezavisna varijabla), prikazane u Tablici 9.2.

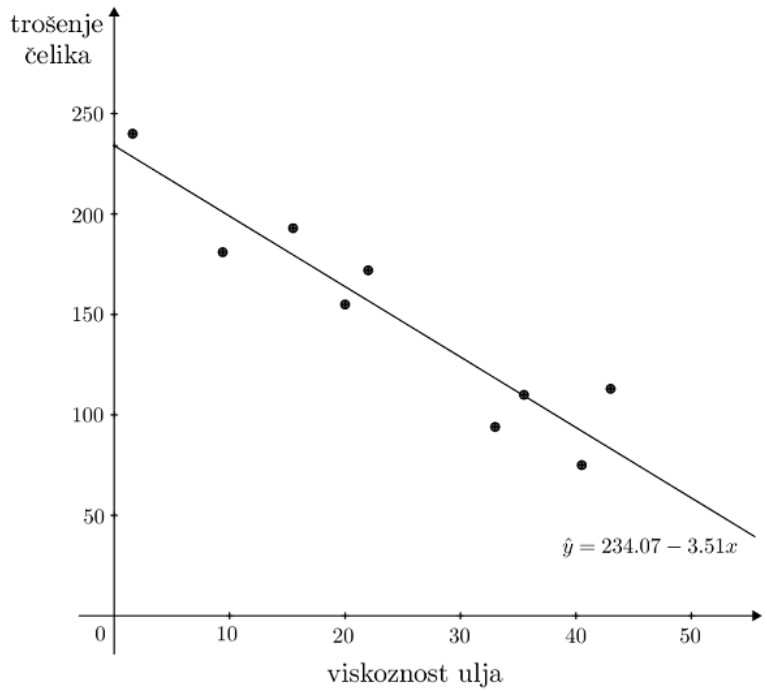
Viskoznost ulja	1.6	9.4	15.5	20	22	35.5	43	40.5	33
Trošenje čelika	240	181	193	155	172	110	113	75	94

Tablica 9.2: Podaci o mekom čeliku i viskoznosti korištenog ulja

Odredit ćemo jednadžbu regresijskog pravca te pomoću dobivenog modela procijeniti koliko će biti trošenje mekog čelika uz viskoznost $x = 21$.

Računamo procjene koeficijenata regresijskog pravca:

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} = \frac{-5794.1}{1651.420} = -3.5086,$$
$$\hat{\beta}_0 = \overline{y}_9 - \hat{\beta}_1 \overline{x}_9 = 148.111 - (-3.5086) \cdot 24.5 = 234.07.$$



Slika 9.6: Regresijski pravac za podatke o trošenju mekog čelika i viskoznosti korištenog ulja

Sada, za viskoznost $x = 21$ procjenjujemo trošenje čelika: $\hat{y} = -3.5086 \cdot 21 + 234.07 = 160.389$ mm^3. □

Za detaljniju diskusiju o svim gore spomenutim rezultatima čitatelju preporučamo reference [5, 6, 7, 8, 10].

Dodatak A

Skupovi i osnovne operacije sa skupovima

U ovom odjeljku ponovit ćemo neke osnovne pojmove vezane za skupove. Skup je osnovni matematički pojam. To je cjelina sastavljena od njenih osnovnih dijelova koje nazivamo **elementima**. Skupove ćemo označavati tiskanim slovima $A, B, C, \ldots$, a njihove elemente pisanim slovima $a, b, c, \ldots$ Ako se element a nalazi u skupu A, onda kažemo da je a **element** od A i pišemo $a \in A$. Ako se element a ne nalazi u skupu A, onda kažemo da a nije element od A i pišemo $a \notin A$. Skup možemo zadati tako da navedemo sve njegove elemente ($A = \{2, 4, 6, 8, \ldots\}$) ili isticanjem karakterističnog svojstva koje veže elemente tog skupa ($A = \{n \in \mathbb{N} : n \text{ je djeljiv brojem } 2\}$). Ako je svaki element skupa A ujedno i element skupa B, onda kažemo da je A **podskup** od B i pišemo $A \subseteq B$. Skupovi A i B su **jednaki**, u oznaci $A = B$, ako je $A \subseteq B$ i $B \subseteq A$. Ako se svi skupovi koji se promatraju smatraju podskupovima nekog skupa $\mathcal{U}$, onda se $\mathcal{U}$ naziva **univerzalni skup**. Skup koji ne sadrži niti jedan element nazivamo **praznim skupom** i označavamo $\emptyset$. Uočimo da za proizvoljni skup A vrijedi $\emptyset \subseteq A \subseteq \mathcal{U}$. **Partitivni skup** skupa A, u oznaci $\mathcal{P}(A)$, je skup svih njegovih podskupova $\mathcal{P}(A) = \{B : B \subseteq A\}$. Dakle, elementi od $\mathcal{P}(A)$ su podskupovi od A. Primjerice, ako je $A = \{1, 2, 3\}$, onda imamo $\mathcal{P}(A) = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}$. Uočimo da u prethodnom primjeru skup $\mathcal{P}(A)$ ima $8 = 2^3$ elemenata. Općenito, ako skup A ima n elemenata, onda $\mathcal{P}(A)$ ima 2^n elemenata. **Presjek** skupova A i B je skup, u oznaci $A \cap B$, koji sadrži elemente koji su u A i B, $A \cap B = \{c : c \in A \text{ i } c \in B\}$. Za skupove A i B kažemo da su **disjunktni** ako je $A \cap B = \emptyset$. **Unija** skupova A i B je skup, u oznaci $A \cup B$, koji sadrži elemente koji su u A ili B, $A \cup B = \{c : c \in A \text{ ili } c \in B\}$. **Razlika** skupova A i B je skup, u oznaci $A \setminus B$, koji sadrži elemente koji su u A ali nisu u B, $A \setminus B = \{c : c \in A \text{ i } c \notin B\}$. Uočimo da $A \setminus B \neq B \setminus A$. **Komplement** skupa

A je skup, u oznaci A^c, koji sadrži elemente koji nisu u A, $A^c = \{a : a \notin A\}$. Dajmo sada neke identitete koji vrijede među skupovima. Za sve skupove A, B i C vrijedi:

(i) $A \cap B = B \cap A$ i $A \cup B = B \cup A$.

(ii) $(A \cap B) \cap C = A \cap (B \cap C)$ i $(A \cup B) \cup C = A \cup (B \cup C)$.

(iii) $A \cup (A \cap B) = A$ i $A \cap (A \cup B) = A$.

(iv) $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ i $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.

(v) $A \cap A^c = \emptyset$ i $A \cup A^c = \mathcal{U}$.

(vi) $A \cap \emptyset = \emptyset$ i $A \cup \emptyset = A$.

(vii) $A \cap \mathcal{U} = A$ i $A \cup \mathcal{U} = \mathcal{U}$.

(viii) $(A \cap B)^c = A^c \cup B^c$ i $(A \cup B)^c = A^c \cap B^c$.

(ix) $(A^c)^c = A$, $\emptyset^c = \mathcal{U}$ i $\mathcal{U}^c = \emptyset$.

Dodatak B

Elementi kombinatorike

Kombinatorika je grana matematike koja se bavi diskretnim (uglavnom konačnim) strukturama. Jedno od osnovnih pitanja/problema u kombinatorici je prebrojavanje konačnih stuktura i stoga u kombinatornim problemima često nailazimo na pitanja tipa "na koliko načina se može nešto napraviti"?

Primjer B.1. Na koliko načina možemo ispuniti loto 7/39 listić?

U ovom odjeljku obradit ćemo samo neke osnovne principe i tehnike prebrojavanja koje su usko vezane s teorijom vjerojatnosti.

Pravilo sume

Ako neka familija broji n_1 članova, neka druga familija n_2 članova, treća n_3 članova, ... i zadnja, k-ta familija broji n_k članova, onda populacija sačinjena od tih familija, uz pretpostavku da su familije disjunktne, broji $n_1 + n_2 + \cdots + n_k$ članova.

Primjer B.2. Na stolu se nalazi 7 jabuka, 5 kruški i 5 banana. Na koliko načina možemo odabrati jednu voćku? Voćku možemo odabrati na $7+5+5 = 17$ načina. □

Pravilo produkta

Imamo istu populaciju kao i u prethodnom odjeljku. Tada, ako želimo izabrati jednog člana iz prve familije, jednog iz druge, ... i jednog iz k-te familije, to možemo napraviti na $n_1 \cdot n_2 \cdots n_k$ načina.

Primjer B.3. Registarska oznaka vozila u Zagrebu sastoji se od dva slova (od 22 moguća) i četiri znamenke. Koliko različitih registarskih oznaka može izdati PU Zagrebačka? PU Zagrebačka može izdati $22 \cdot 22 \cdot 10 \cdot 10 \cdot 10 \cdot 10 = 4840000$ različitih registarskih oznaka. □

Permutacije

Permuacija je poredak konačnog broja objekata u dani redoslijed. Ako imamo n objekata, onda je broj mogućih poredaka $n! = n \cdot (n-1) \cdot (n-2) \cdots 3 \cdot 2 \cdot 1$.

Primjer B.4. Koliko različitih četveroznamenkastih brojeva možemo sastaviti od znamenaka $1, 2, 3$ i 4? Možemo sastaviti $4! = 24$ različitih četveroznamenkastih brojeva. □

Permutacije s ponavljanjem

Neka je dana familija n objekata, od čega je n_1 objekata prve vrste, n_2 objekata druge vrste, ... i n_k objekata k-te vrste ($n = n_1 + \cdots + n_k$). **Permutacija s ponavljanjem** je broj različitih poredaka ove familije. Broj takvih poredaka je

$$\frac{n!}{n_1! \cdot n_2! \cdots n_k!}.$$

Primjer B.5. Koliko različitih četveroznamenkastih brojeva možemo sastaviti od znamenaka $1, 1, 2$ i 2? Možemo sastaviti $4!/(2! \cdot 2!) = 6$ različitih četveroznamenkastih brojeva. □

Varijacije

Imamo familiju od n različitih objekata. **Varijacija** duljine k je poredak bilo kojih k objekata (od n) u dani redoslijed duljine k. Broj takvih poredaka je

$$n \cdot (n-1) \cdots (n-k+1).$$

Specijalno, ako je $k = n$, onda se radi o permutaciji.

Primjer B.6. Koliko različitih troznamenkastih brojeva možemo sastaviti od znamenaka $1, 2, 3$ i 4? Možemo sastaviti $4 \cdot 3 \cdot 2 = 24$ različitih troznamenkastih brojeva. □

Varijacije s ponavljanjem

Imamo familiju od n različitih objekata. **Varijacija s ponavljanjem** duljine k je poredak bilo kojih k objekata (možemo ponavljati/uzimati isti objekt) u dani redoslijed duljine k. Broj takvih poredaka je

$$n^k.$$

Primjer B.7. Koliko različitih troznamenkastih brojeva možemo sastaviti od znamenaka $1, 2, 3$ i 4 ako dozvoljavamo ponavljanje? Možemo sastaviti $4^3 = 64$ različitih troznamenkastih brojeva. □

Kombinacije

Neka su $n, k \in \mathbb{N}_0$, $k \leqslant n$. Binomni koeficijent, u oznaci $\binom{n}{k}$, je broj dan formulom

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

Vrijede sljedeća svojstva:

(i) $\binom{n}{0} = \binom{0}{0} = 1$, jer definiramo $0! = 1$.

(ii) $\binom{n}{r} = \binom{n}{n-r}$.

(iii) $\binom{n+1}{r} = \binom{n}{r-1} + \binom{n}{r}$.

Binomni koeficijenti usko su vezani s kombinacijama. Imamo familiju od n različitih objekata. **Kombinacija** veličine k je bilo koji k-člani podskup ove familije. Takvih podskupova imamo

$$\binom{n}{k}.$$

Uočite vezu s varijacijama, tj.

$$\binom{n}{k} = \frac{n \cdot (n-1) \cdots (n-k+1)}{k!},$$

te s permutacijama s ponavljanjem

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

Primjer B.8. Koliko ima tročlanih podskupova skupa $\{1, 2, 3, 4\}$? Ima ih $\binom{4}{3} = 4$. □

Sada smo spremni dati odgovor na pitanje s početka. Loto 7/39 listić možemo ispuniti na $\binom{39}{7} = 15380937$ različitih načina.

Kombinacije s ponavljanjem

Imamo familiju od n različitih objekata. **Kombinacija s ponavljanjem** veličine k je bilo koji k-člani podskup ove familije (elementi se mogu ponavljati). Takvih podskupova imamo

$$\binom{n+k-1}{k}.$$

Uočite vezu s permutacijama s ponavljanjem, tj.

$$\binom{n+k-1}{k} = \frac{(n+k-1)}{k!(n-1)!}.$$

Primjer B.9. Koliko ima peteročlanih skupova čiji su elementi iz skupa $\{1,2,3,4\}$? Ima ih $\binom{4+5-1}{5} = \binom{8}{5} = 56$. □

Literatura

[1] M. Benšić, N. Šuvak, *Primijenjena statistika*. Web skripta, dostupno na `http://www.mathos.unios.hr/ptfstatistika/00_statistika.pdf`, 2013.

[2] M. Benšić, N. Šuvak, *Uvod u vjerojatnost i statistiku*. Web skripta, dostupno na `http://www.mathos.unios.hr/uvis/UVIS_knjiga_final/UVIS_knjiga_web.pdf`, 2014.

[3] V. Čuljak, *Vjerojatnost i statistika*. Web skripta, dostupno na `http://www.grad.hr/vera/webnastava/vjerojatnostistatistika/vis-pdf.pdf`, 2011.

[4] T. Došlić, D. Vrgoč, *Poslovna statistika I*. Web skripta, dostupno na `http://www.agr.unizg.hr/multimedia/pdf/26711_skripta.pdf`, 2011.

[5] N. T. Kottegoda, R. Rosso, *Applied Statistics for Civil and Environmental Engineers*, Blackwell Publishing, 2008.

[6] M. R. Spiegel, *Schaums's Outline of Theory and Problems of Probability and Statistics*, Schaum's Outline Series, 1998.

[7] D. C. Montgomery, G. C. Runger, *Applied Statistics and Probability for Engineers*, John Wiley & Sons, New York, 2003.

[8] B. Petz, *Osnovne statističke metode za nematematičare*, Naklada Slap, Zagreb, 2007.

[9] N. Sarapa, *Teorija vjerojatnosti*, Školska knjiga, Zagreb, 2002.

[10] H. Tijms, *Understanding probability*, Cambridge University Press, Cambridge, 2007.