Vjerojatnost i matematička statistika

Ante Mimica

Poslijediplomski specijalistički studij
aktuarske matematike

29. siječnja 2016.

Sadržaj kolegija

1. Opisna analiza podataka
2. Slučajne varijable
3. Funkcije izvodnice
4. Zajednička razdioba slučajnih varijabli
5. Centralni granični teorem i primjena
6. Uzorkovanje i statističko zaključivanje
7. Točkovno procjenjivanje
8. Pouzdani intervali
9. Testiranje statističkih hipoteza
10. Korelacijska i regresijska analiza
11. Analiza Varijance

Literatura

1. M. Huzak, Vjerojatnost i matematička statistika, skirpta, 2006.
2. *Subject* 101: *Statistical Modelling, Core Reading* 2000, Faculty and Institute of Actuaries
3. *Subjects C*1/2: *Statistics, Core Reading* 1996, Faculty and Institute of Actuaries
4. F. Daly, D.L. Hand, M.C. Jones, A.D. Lunn, K.J. McConway, *Elements of Statistics*, Addison-Wesley, 1995.
5. Ž. Pauše, *Uvod u matematičku statistiku*, Školska knjiga, Zagreb, 1993.
6. J.E. Freund, *Mathematical Statistics*, Prentice Hall International, 1992.

Outline

1. Opisna analiza podataka

1.1 Vrste podataka

Primjer 1.1

Osiguranici od autoodgovornosti nekog osiguravajućeg društva i

$$\begin{aligned} X &= \text{broj šteta po polici u proteklih godinu dana,} \\ Y &= \text{ukupan iznos šteta po polici u prošloj godini.} \end{aligned}$$

$\mathbf{Z} = (X, Y)$ je dvodimenzionalno statističko obilježje.

- populacija $\rightarrow$ grupa objekata koje proučavamo
- (reprezentativni) uzorak

Primjer 1.2

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

Razlikujemo:

- populacijske podatke
- uzoračke podatke

Podjela podataka po *tipu* varijable (stat. obilježja):

- numeričke $\longrightarrow$ vrijednosti: brojevi
- kategorijalne $\longrightarrow$ vrijednosti: razredi
  (npr. spol, mjesto rođenja, kategorija vozača)

Numeričke varijable:

- *diskretne* (obično predstavljaju neko prebrojavanje).
  Npr. broj šteta po polici osiguranja, broj ovlaštenih aktuara u HAD-u.
- *neprekidne* (obično predstavljaju rezultat mjerenja neke fizikalne ili novčane veličine)
  Npr. visina, težina, iznos šteta po polici osiguranja

1.2 Frekvencijske distribucije

Frekvencijskim distribucijama opisuju se skupovi:

- diskretnih numeričkih podataka
- kategorijalnih podataka

Frekvencijske distribucije prikazuju se

- tabelarno pomoću *frekvencijskih tablica*
- grafički pomoću *stupčastih dijagrama, strukturnih dijagrama*

Primjer 1.3

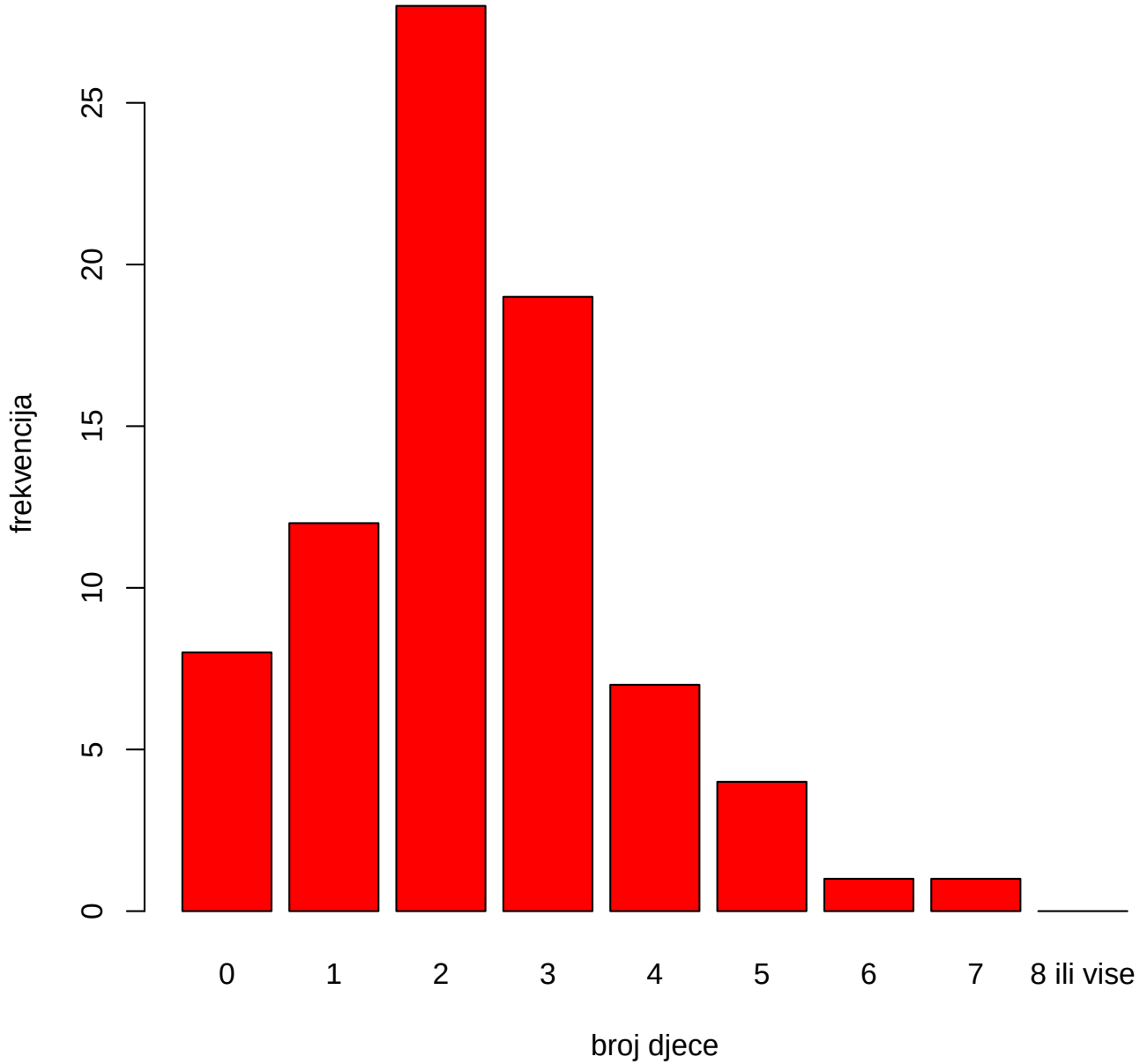
Uzorak od 80 obitelji.

$$X = \text{broj djece u obitelji mlađe od 16 god.}$$

Frekvencijska tablica:

broj djece	frekvencija	relativna frekvencija
0	8	0.1
1	12	0.15
2	28	0.35
3	19	0.2375
4	7	0.0875
5	4	0.05
6	1	0.0125
7	1	0.0125
8 ili više	0	0
Σ	80	1.0

Stupčasti dijagram frekvencija broja djece u obitelji:

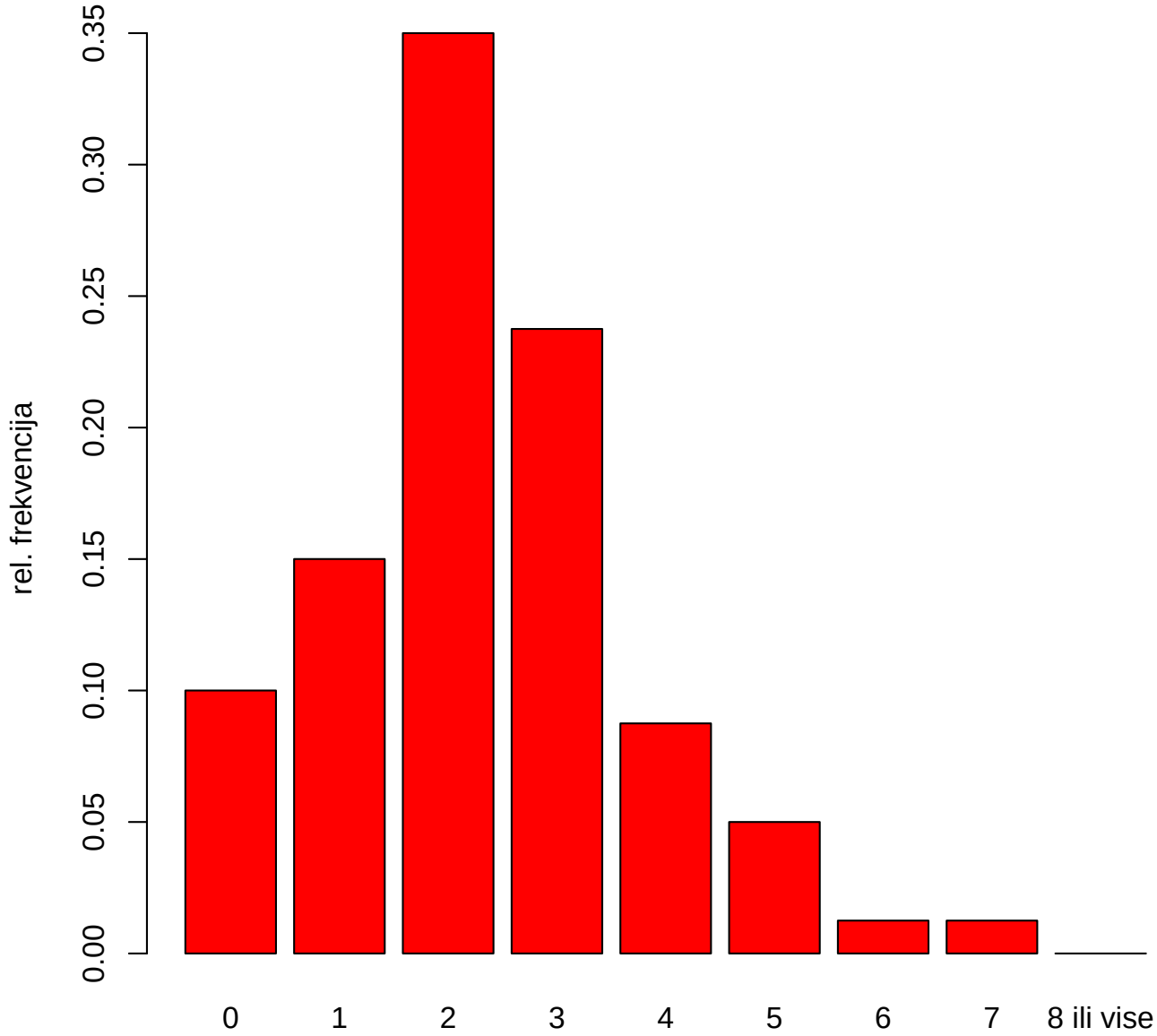


U R-u:

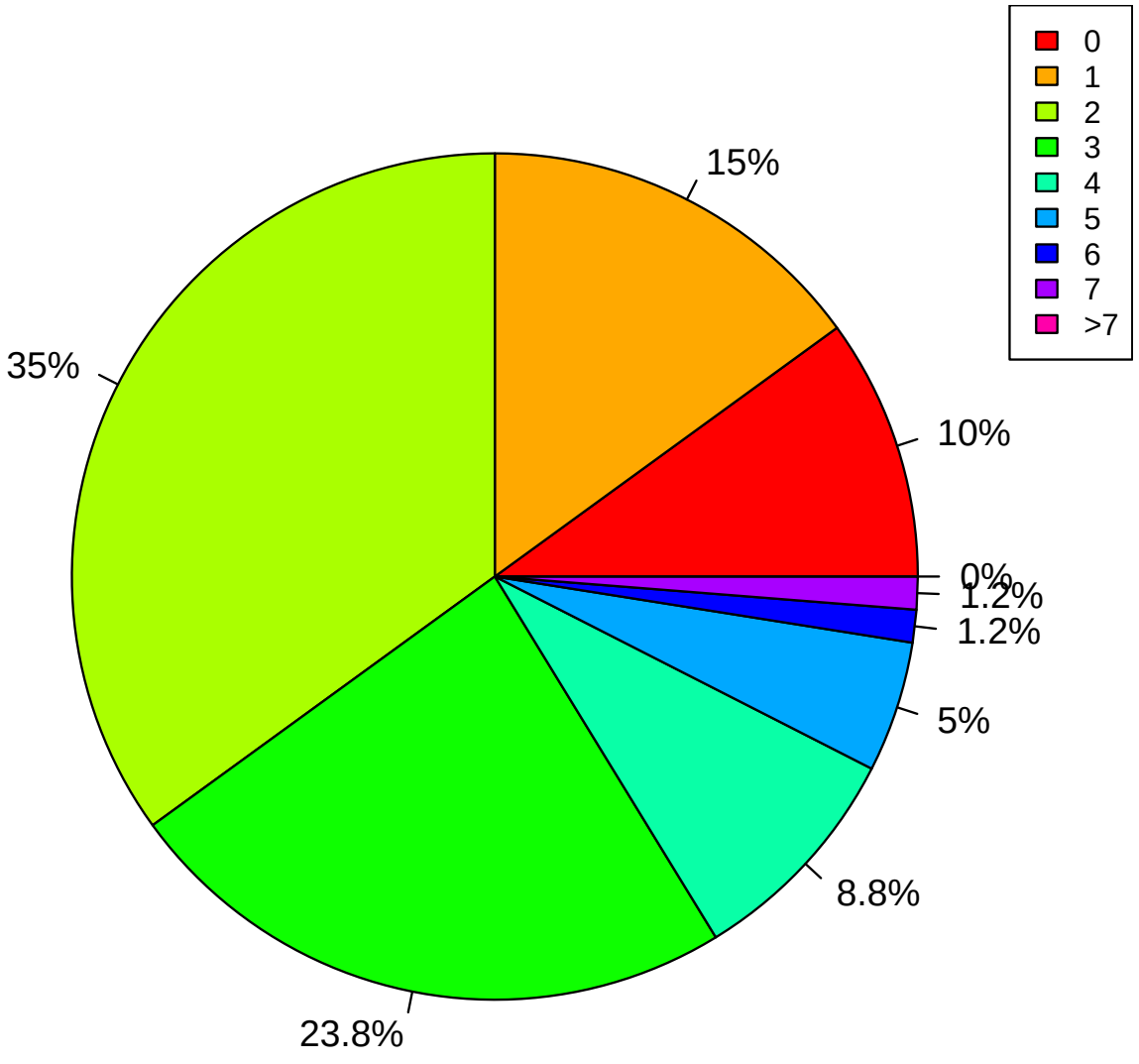
```
> podaci<-data.frame(levels=c(0,1,2,3,4,5,6,7,"8
ili vise"),frekv=c(8,12,28,19,7,4,1,1,0))
> barplot(c$frekv,names=c$levels,xlab="broj
djece",ylab="frekvencija",col="red")

>
podaci<-data.frame(podaci,podaci$frekv/sum(podaci
$frekv))
> names(podaci)[3]<-"relfrekv"
>
barplot(podaci$relfrekv,names=podaci$levels,xlab="broj
djece",ylab="rel.  frekvencija",col="red")
```

Stupčasti dijagram relativnih frekvencija broja djece u obitelji:



Strukturni dijagram relativnih frekvencija broja djece u obitelji:



U R-u:

```
> pie(podaci$rf,labels=paste(round(100*podaci$rf,
1),"%",sep=""),col=rainbow(length(podaci$rf)))
> legend("topright",
c("0","1","2","3","4","5","6",
"7",">7"),fill=rainbow(length(podaci$rf)),cex=0.8)
```

1.3 Histogrami i frekvencijske distribucije grupiranih vrijednosti

Frekvencijskim distribucijama grupiranih vrijednosti opisuju se skupovi neprekidnih numeričkih podataka.

Prikazuju se

- tabelarno pomoću *frekvencijskih tablica grupiranih vrijednosti*
- grafički pomoću *histograma*

Primjer 1.4

Raspolažemo sa 100 podataka o iznosima šteta zbog popuštanja vodovodnih instalacija po policama osiguranja kućanstava:

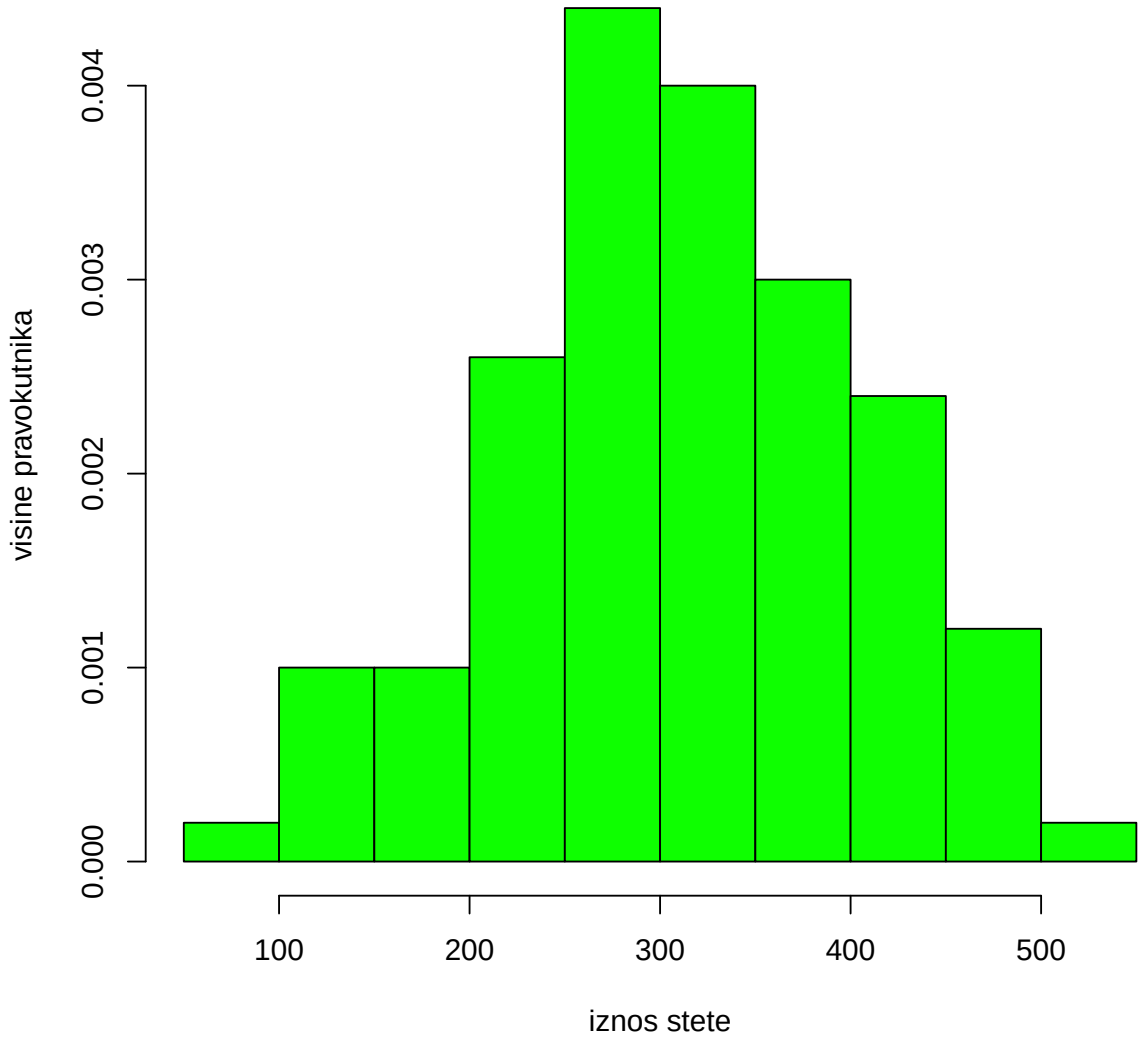
243	306	271	396	287	399	466	269	295	330
425	324	228	113	226	176	320	230	404	487
127	74	523	164	366	343	330	436	141	388
293	464	200	392	265	403	372	259	426	262
221	355	324	374	347	261	278	113	135	291
176	342	443	239	302	483	231	292	373	346
293	236	223	371	287	400	314	464	337	308
359	352	273	267	277	184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

Frekvencijska tablica grupiranih vrijednosti:

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

Histogram:

- ukupna površina je jednaka 1



1.4 Stem and leaf dijagram

- stabljika (eng. *stem*) reprezentira razred (npr. znamenka stotica)
- list (eng. *leaf*) znamenka koja reprezentira broj iz razreda (npr. znamenka desetica)

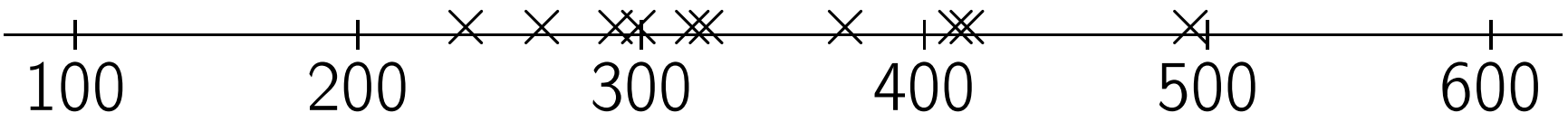
Npr. za skup podataka iz Primjera 1.4 dobijemo sljedeći *stem and leaf dijagram*:

0	7
1	112346778
2	012222333334456666667777788899999999
3	000111222233333444445555677777899 9
4	00011112223446668 89
5	2

1.5 Linijski dijagram i dijagram točaka

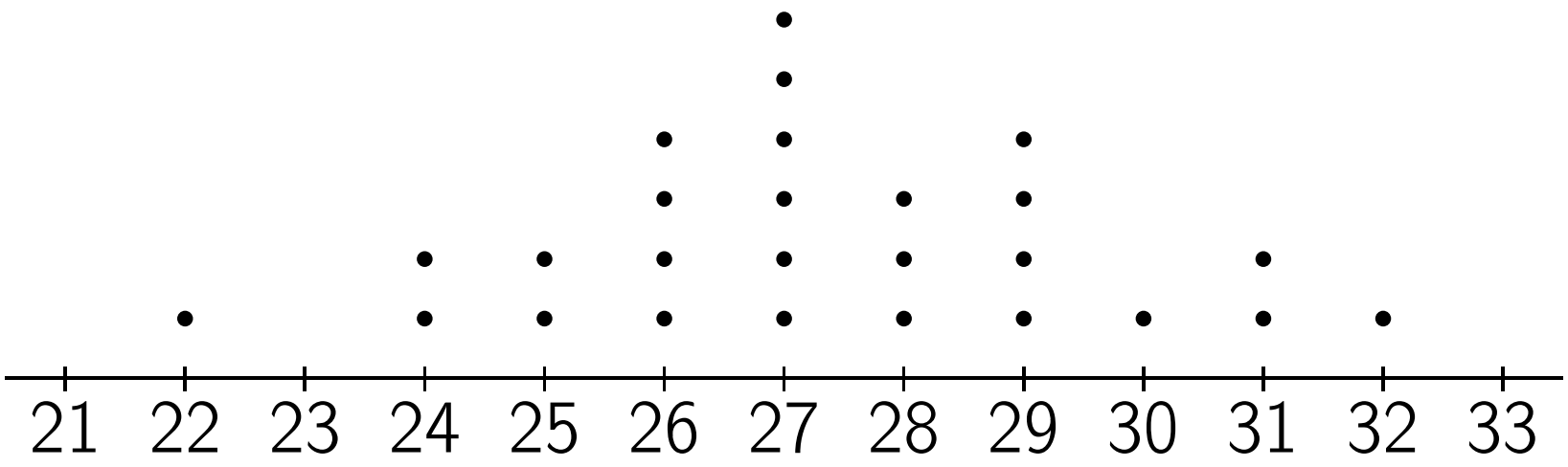
- *linijski dijagram* se koristi za prikaz vrijednosti koje se ne ponavljaju previše
- inače se koristi *dijagram točaka*

Npr. linijski dijagram skupa podataka koji se sastoji od zadnjih 10 brojeva iz Primjera 1.4:



Primjer 1.5

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



1.6 Mjere lokacije

Mjere centralnih tendencija:

- *aritmetička sredina*
- *medijan*
- *mod*

Podaci (realizacije varijable X):

$$x_1, x_2, \ldots, x_n \tag{1}$$

Ako je X ordinalna ili numerička varijabla, podaci se mogu urediti

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

1.6.1 Aritmetička sredina

X je numerička varijabla.

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

Ako se u (1) ponavljaju brojevi:

$$a_1, a_2, \ldots, a_k \tag{3}$$

s frekvencijama

$$f_1, f_2, \ldots, f_k,$$

tada je

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

Npr. aritmetička sredina podataka iz Primjera 1.3 je:

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

1.6.2 Medijan

X je numerička ili ordinalna varijabla. Uređeni podaci iz (1):

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}. \tag{4}$$

Medijan je vrijednost m takva da je:

- točno pola (50%) svih podataka manje ili jednako od m i
- točno pola svih podataka veće li jednako od m.

Dakle,

$$\begin{aligned} m &= x_{(k)} && \text{ako je } n = 2k-1 \\ m &= \frac{1}{2}(x_{(k)} + x_{(k+1)}) && \text{ako je } n = 2k \end{aligned}$$

Npr. u Primjeru 1.3 je $n = 80$ pa je

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

1.6.3 Mod

- vrijednost od X s najvećom frekvencijom

Npr. mod uzorka iz Primjera 1.3 je 2 jer ima najveću frekvenciju (28).

1.7 Mjere raspršenja

1.7.1 Standardna devijacija

Standardna devijacija:

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

Varijanca: s^2

Alternativne formule za varijancu (standardnu devijaciju):

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

Za uzorak iz Primjera 1.3, uzoračka varijanca je:

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

1.7.2 Momenti

k-ti moment oko α:

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

Moment je moment oko $\alpha = 0$.

Centralni moment je moment oko $\alpha = \bar{x}$.

1.7.3 Raspon

Raspon:

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

Raspon uzorka iz Primjera 1.3 je

$$R = 7 - 0 = 7.$$

1.7.4 Interkvartil

r-ti kvantil:

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

$$(r = k + \alpha,\ k \in \mathbb{N},\ k < n,\ 0 \leq \alpha < 1)$$

Donji (q_L) i gornji (q_U) kvartili:

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

Interkvartil:

$$IQR = q_U - q_L$$

Za uzorak iz Primjera 1.3:

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

$$\Rightarrow IQR = 3 - 1.25 = 1.75.$$

1.8 Mjere asimetričnosti

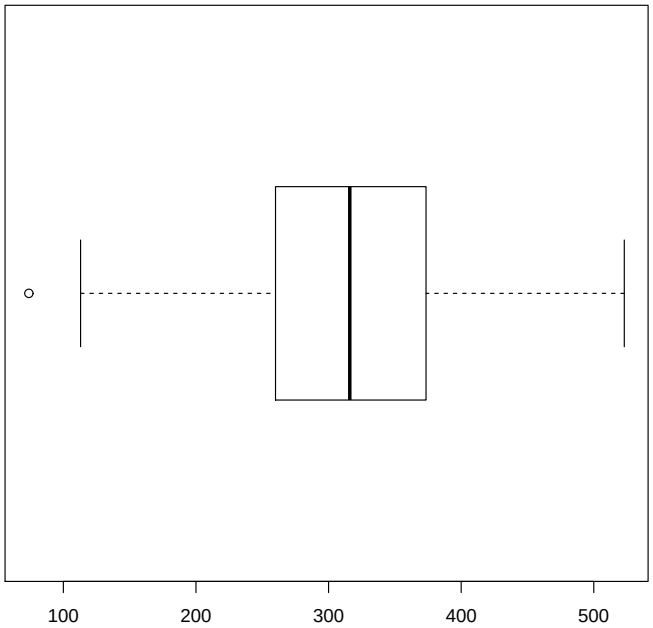
Koeficijent asimetrije:

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

Ako je

- $\alpha_3 = 0$ podaci su *simetrični*
- $\alpha_3 < 0$ podaci su *negativno asimetrični*
- $\alpha_3 > 0$ podaci su *pozitivno asimetrični*

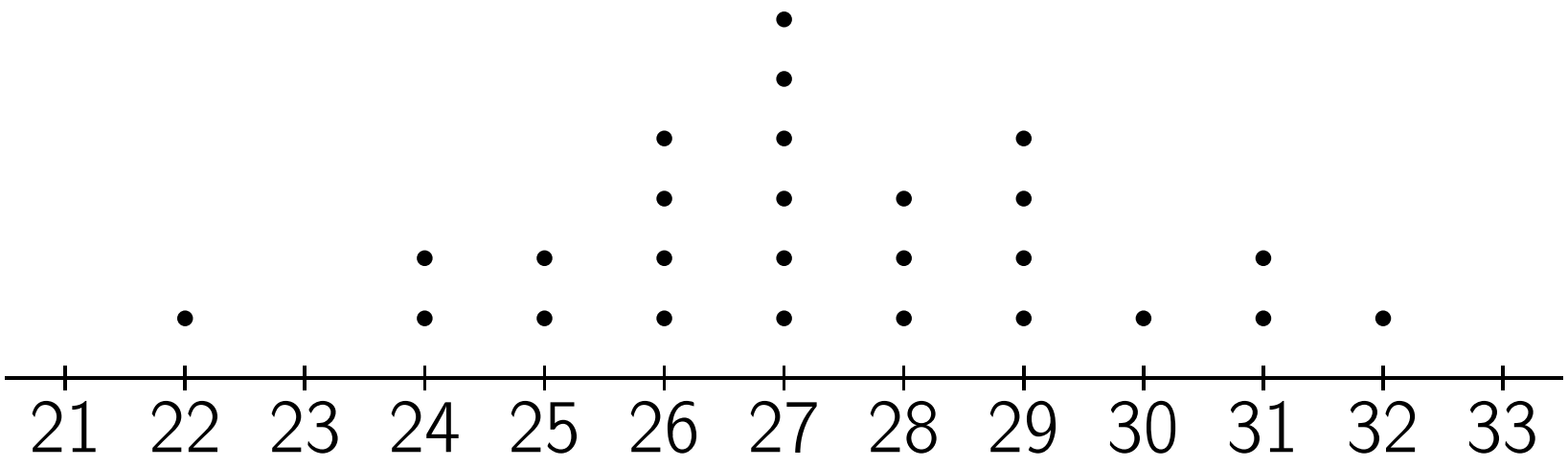
1.9 Dijagram pravokutnika



- *box and whisker plot*
- "brkovi" - najmanja i najveća vrijednost unutar intervala $[q_L - 1.5 \cdot IQR, q_U + 1.5 \cdot IQR]$
- *outlier* - vrijednost koja se nalazi izvan "brkova"

Zadatak 1.1

Zadan je dijagram točaka kao u Primjeru 1.5 koji opisuje mjerenja vezana uz vrijeme potrebno za izvođenje neke operacije u sekundama:



(a) Izračunajte aritmetičku sredinu i varijancu.

(b) Izračunajte medijan i interkvartil.

(c) Skicirajte dijagram pravokutnika.

Outline

2. Slučajne varijable

Primjer 2.1 Bacanje igraće kocke.

Događaji: pao je paran broj, pala je 6,...

Elementarni događaji: 1,2,3,4,5,6

A, B događaji $\Rightarrow$ događaji su i

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

Prostor elementarnih događaja: Ω

Familija događaja: $\mathcal{F}$

Vjerojatnost: Preslikavanje $\mathbb{P} : \mathcal{F} \to \mathbb{R}$ sa svojstvima:

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

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

$(P3)$ $A_1, A_2, \ldots$ iz $\mathcal{F}$ i $A_i \cap A_j = \emptyset$ za $i \leq j \Rightarrow$

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

Vjerojatnosni prostor: $(\Omega, \mathcal{F}, \mathbb{P})$

Vrijedi:

- $$A \subset B \Rightarrow \mathbb{P}(B \setminus A) = \mathbb{P}(B) - \mathbb{P}(A)$$

  (za $A \not\subset B$ formula općenito ne vrijedi!)

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

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

DZad

$$\begin{aligned}\mathbb{P}(A \cup B \cup C) = \quad & \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) \\ & - \mathbb{P}(A \cap B) - \mathbb{P}(A \cap C) - \mathbb{P}(B \cap C) \\ & + \mathbb{P}(A \cap B \cap C)\end{aligned}$$

Primjer 2.1(nastavak)

Igraća kocka je simetrična.

$$p_1 = p_2 = \cdots = p_6 = \frac{1}{6} \quad \Rightarrow$$

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

Što ako kocka nije simetrična?

Uvjetna vjerojatnost

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

Primjer 2.3

$A = \{\text{pala je šestica}\}$
$B = \{\text{pao je paran broj na kocki}\}$

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\frac{1}{6}}{\frac{3}{6}} = \frac{1}{3}.$$

$\mathbb{P}(\cdot|B)$ je isto vjerojatnost

Nezavisnost događaja

A i B su *nezavisni događaji* ako je

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

$$\begin{aligned} A \text{ i } B \text{ su nezavisni} &\iff \mathbb{P}(A|B) = \mathbb{P}(A) \\ &\iff \mathbb{P}(A|B^c) = \mathbb{P}(A) \\ &\iff \mathbb{P}(B|A) = \mathbb{P}(B) \\ &\iff \mathbb{P}(B|A^c) = \mathbb{P}(B). \end{aligned}$$

DZad A, B su nezavisni ako i samo ako su A^c i B nezavisni.

2.2 Diskretne slučajne varijable

$X : \Omega \to \mathbb{R}$, $X, Y, Z, \ldots$

Slučajna varijabla je *diskretna* ako je $\mathrm{Im}X := f(\Omega)$ prebrojiv skup i

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

je događaj za svaki $x \in \mathrm{Im}X$.

Funkcija vjerojatnosti (gustoće) od X:

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

Vrijedi:

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

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

Posebno, $f_X(x) = 0$ za $x \notin \mathrm{Im}X$.

Funkcija distribucije od X:

$$F_X : \mathbb{R} \to \mathbb{R}, \quad F_X(x) := \mathbb{P}(X \leq x)$$

Vrijedi:

$$F_X(x) = \sum_{\{y \in \mathsf{Im} X : y \leq x\}} f_X(y)$$

Stepenasta je, rastuća, neprekidna zdesna i

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

2.3 Neprekidne slučajne varijable

Slučajna varijabla X je *neprekidna* ako:

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

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

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

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

f_X zovemo *funkcijom gustoće razdiobe* od X.

Za sve $a, b \in \mathrm{Im}X$,

$$\mathbb{P}(X = a) = 0,$$

i ako je $a < b$,

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

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

Za gustoću vrijedi:

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

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

Za funkciju distribucije neprekidne s.v. vrijedi:

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

Neprekidna je, rastuća, $F_X(-\infty) = 0$, $F_X(+\infty) = 1$.
Vrijedi:

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

Ako je F_X derivabilna,

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

2.4 Matematičko očekivanje

$$\mathbb{E}[X] := \sum_{x \in \mathrm{Im} X} x f_X(x) \quad \text{(ako je } X \text{ diskretna)}$$

$$\mathbb{E}[X] := \int_{-\infty}^{+\infty} x f_X(x)\, dx \quad \text{(ako je } X \text{ neprekidna)}$$

(ako red/integral zdesna apsolutno konvergira)

Zadatak 2.1

Slučajno se bira točka unutar kvadrata duljine stranice 2. Označimo s X najmanju udaljenost te točke od stranica kvadrata. Nađite funkciju gustoće i matematičko očekivanje od X.

Za funkciju $g : \mathbb{R} \to \mathbb{R}$ vrijedi

$$\mathbb{E}[X] := \sum_{x \in \mathsf{Im} X} g(x) f_X(x) \quad \text{(ako je } X \text{ diskretna)}$$

$$\mathbb{E}[X] := \int_{-\infty}^{+\infty} g(x) f_X(x)\, dx \quad \text{(ako je } X \text{ neprekidna)}$$

Varijanca sl. var. X je definirana s

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

$$\mathrm{Var} X = \mathbb{E}[X^2] - (\mathbb{E}X)^2$$

2.6 Očekivanje i varijanca linearne transformacije s.v. $(\mathbb{E}X = \mu,\ \mathrm{Var}X = \sigma^2)$

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

Za *standardiziranu verziju* od X:

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

vrijedi: $\mathbb{E}Z = 0,\ \mathrm{Var}Z = 1.$

2.7 Momenti *k-ti moment od X oko c* je broj:

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

momenti $(c=0)$,

centralni momenti $(c=\mathbb{E}X)$

Koeficijent asimetrije od X:

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

$(\mu = \mathbb{E}X,\ \sigma = \sigma(X))$

Distribucija od X je:
simetrična ako je $\alpha_3(X) = 0$,

negativno asimetrična ako je $\alpha_3(X) < 0$
$\rightarrow$ lijevi rep, asimetričnost slijeva

pozitivno asimetrična ako je $\alpha_3(X) > 0$.
$\rightarrow$ desni rep, asimetričnost zdesna

2.8 Primjeri važnih distribucija

2.8.1 Diskretne razdiobe

Uniformna razdioba

– na skupu $S = \{1, 2, \ldots, k\}$ $(k \in \mathbb{N})$

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

$$\mathbb{E}X = \frac{k+1}{2} \qquad \mathrm{Var}X = \frac{k^2 - 1}{12}$$

Npr. bacanje igraće kocke $\rightarrow k = 6$, $X =$broj na kocki

$$\mathbb{E}X = \frac{7}{2} \qquad \mathrm{Var}X = \frac{35}{12}$$

Bernoullijeva razdioba

$X = 1$ ako je *uspjeh*, inače je $X = 0 \Rightarrow$
$\mathrm{Im}X = \{0, 1\}$

$\theta = \mathbb{P}(X = 1)$ je *vjerojatnost uspjeha* $(\theta \in [0, 1])$

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

$$\mathbb{E}X = \theta \quad \mathrm{Var}X = \theta(1 - \theta)$$

Binomna razdioba

$X =$ broj uspjeha u nizu on n njd Bernoullijevih pokusa

$X \sim b(n, \theta)$ $(0 \leq \theta \leq 1)$.

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

$$\mathbb{E}X = n\theta \quad \mathrm{Var}X = n\theta(1-\theta)$$

Geometrijska razdioba

$X =$ broj njd Bernoullijevih pokusa do prvog uspjeha

$X \sim$ geometrijska (θ) $(0 < \theta < 1)$

X je *vrijeme čekanja*

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

$$\mathbb{E}X = \frac{1}{\theta} \quad \text{Var}X = \frac{1-\theta}{\theta^2}$$

$Y = X - 1 =$ broj neuspjeha do prvog uspjeha

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

$$\mathbb{E}Y = \frac{1-\theta}{\theta} \quad \text{Var}Y = \frac{1-\theta}{\theta^2}$$

Negativna binomna razdioba

$X =$ broj njd Bernoullijevih pokusa do uključivo k-tog uspjeha

$X \sim$ negativna bin. (k, θ) $(0 < \theta < 1)$

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

$$\mathbb{E}X = \frac{k}{\theta} \quad \text{Var}[X] = k\frac{1-\theta}{\theta^2}$$

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

i $f_X(k) = \theta^k$.

$Y = X - k =$ broj neuspjeha do k-tog uspjeha

$$f_Y(x) = \binom{k+x-1}{k-1} \theta^k (1-\theta)^x$$

za $x \in \mathsf{Im} Y = \{0, 1, 2, \ldots\}$,

$$\mathbb{E}[Y] = k\frac{1-\theta}{\theta} \quad \mathrm{Var}[Y] = k\frac{1-\theta}{\theta^2}$$

Hipergeometrijska distribucija

Kutija: N kuglica $= K$ bijelih $+ (N - K)$ crnih

$X =$ broj bijelih kuglica među n izvučenih *bez vraćanja*

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

$$\theta = K/N \quad \Rightarrow \quad \mathbb{E}[X] = n\theta$$

Poissonova razdioba

1. Model za broj slučajnih događaja koji se realiziraju tijekom nekog vremenskog intervala uz uvjete:

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

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

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

$\Rightarrow$ Događaji se pojavljuju u skladu sa zakonom *Poissonovog procesa*.

2. Granična je distribucija $b(n, \theta)$-razdiobe kada $n \to +\infty$, $\theta \to 0$ t.d. je $\lambda = n\theta =$ konstantno.

$X \sim P(\lambda)$:

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

$$\mathbb{E}X = \mathrm{Var}X = \lambda$$

2.8.2 Neprekidne razdiobe

Uniformna razdioba $X \sim U(\alpha, \beta)$

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

$$\mathbb{E}X = \frac{\alpha + \beta}{2} \quad \mathrm{Var}X = \frac{(\beta - \alpha)^2}{12}$$

Gama distribucija

$X \sim \Gamma(\alpha, 1/\lambda)$, $(\alpha > 0,\ \lambda > 0)$, $\mathsf{Im}X = \langle 0, +\infty \rangle$

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

$\Gamma(\alpha) = \int_0^{+\infty} t^{\alpha-1} e^{-t}\, dt$ (Γ-funkcija)

(i) $\Gamma(1) = 1$, $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$ za $\alpha > 1$
$\Rightarrow \Gamma(n) = (n-1)!$ za $n \in \mathbb{N}$;

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

$$\mathbb{E}X = \frac{\alpha}{\lambda} \quad \mathrm{Var}X = \frac{\alpha}{\lambda^2}$$

Eksponencijalna distribucija

$X \sim Exp(\lambda) \equiv \Gamma(1, 1/\lambda)$

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

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

$$\mathbb{E}X = \frac{1}{\lambda} \quad \mathrm{Var}X = \frac{1}{\lambda^2}$$

X je vrijeme čekanja između pojavljivanja dva događaja u Poissonovom procesu

χ^2-razdioba

$$X \sim \chi^2(n) \equiv \Gamma(\tfrac{n}{2}, 2) \text{ za } n \in \mathbb{N}$$

$$\mathbb{E}[X] = \frac{n}{2} \cdot 2 = n \quad \mathrm{Var}[X] = \frac{n}{2} \cdot 2^2 = 2n$$

Beta distribucija

$$X \sim B(\alpha, \beta),\ (\alpha > 0,\ \beta > 0),\ \mathrm{Im}X = \langle 0, 1 \rangle$$

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

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

$$\mathbb{E}[X] = \frac{\alpha}{\alpha+\beta} \quad \mathrm{Var}[X] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Normalna razdioba

$X \sim N(\mu, \sigma^2)$, $(\mu,\ \sigma^2 > 0)$, $\mathrm{Im}X = \mathbb{R}$

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

$$\mu = \mathbb{E}X \quad \sigma^2 = \mathrm{Var}X$$

$$X \sim N(\mu, \sigma^2) \quad \Rightarrow \quad Y := aX+b \sim N(a\mu+b, a^2\sigma^2)$$

Važna je jer:

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

Zadatak 2.2

Neka je $X \sim N(0,1)$. Dokažite da je $X^2 \sim \chi^2(1)$.

Standardizirana verzija od X: $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$

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

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

$$\Phi_0(x) := -\Phi_0(-x), \quad \text{za} \quad x < 0, \qquad \Phi_0(0) = 0$$

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

Na primjer, iz tablica:

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

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

Slično,

$$\begin{aligned}
\mathbb{P}(-2.576 < Z < 2.576) &= 0.99 \\
\mathbb{P}(-3 < Z < 3) &= 0.997 \quad \text{(pravilo } 3\sigma\text{)}
\end{aligned}$$

3. Funkcije izvodnice

3.1 Funkcije izvodnice vjerojatnosti

X diskretna s.v., $\mathrm{Im}X = \{0, 1, 2, 3, \ldots\}$

$$p_k := \mathbb{P}(X = k), \quad k = 0, 1, 2, \ldots.$$

Funkcija izvodnica vjerojatnosti od X

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

(definirana je za $t \in \mathbb{R}$ za koje gornje očekivanje postoji, npr. uvijek je definirana za $|t| \leq 1$).

Teorem jedinstvenosti za f.i.v.

$X \stackrel{d}{=} Y$ ako i samo ako je $G_X = G_Y$.

Primjer 3.1

(a) X uniformna na $\{1, 2, \ldots, k\}$

$$G_X(t) = \begin{cases} \frac{t(1-t^k)}{k(1-t)} & t \neq 1 \\ 1 & t = 1. \end{cases}$$

(b) $X \sim b(n, \theta)$

$$G_X(t) = (\theta t + 1 - \theta)^n, \quad t \in \mathbb{R}.$$

(c) $X \sim$ geometrijska(θ)

$$G_X(t) = \frac{\theta t}{1 - t(1-\theta)}, \quad |t| \leq \frac{1}{1-\theta}$$

(d) $X \sim P(\lambda)$

$$G_X(t) = e^{-\lambda(1-t)}, \quad t \in \mathbb{R}$$

(e) $X \sim$ negativna binomna(k, θ)

$$G_X(t) = \left(\frac{\theta t}{1 - t(1-\theta)}\right)^k, \quad |t| \leq \frac{1}{1-\theta}$$

Računanje momenata

Razvijmo $t \mapsto t^X$ u Taylorov red oko 1:

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

Računanjem mat. očekivanja dobijemo

$$G_X(t) = \mathbb{E}[t^X] = 1 + \underbrace{\mathbb{E}X}_{=G'_X(1)}(t-1) + \underbrace{\mathbb{E}[X(X-1)]}_{=G''_X(1)}\frac{(t-1)^2}{2!}$$

$$+ \underbrace{\mathbb{E}[X(X-1)(X-2)]}_{=G'''_X(1)}\frac{(t-1)^3}{3!} + \dots$$

$$\Longrightarrow$$

$$\mathbb{E}X = G'_X(1)$$

$$\mathbb{E}[X^2] = \mathbb{E}[X(X-1)] + \mathbb{E}X = G'_X(1) + G''_X(1)$$

$$\mathrm{Var}X = \mathbb{E}[X^2] - (\mathbb{E}X)^2 = G''_X(1) + G'_X(1)(1 - G'_X(1))$$

Npr. za $X \sim$ geometrijska(θ)

$$\begin{aligned}
\mathbb{E}X = G'_X(1) &= \frac{d}{dt}\frac{\theta t}{1-t(1-\theta)}\Big|_{t=1} \\
&= \frac{\theta}{(1-t(1-\theta))^2}\Big|_{t=1} = \frac{1}{\theta} \\
\mathbb{E}[X(X-1)] = G''_X(1) &= \frac{2\theta(1-\theta)}{(1-t(1-\theta))^3}\Big|_{t=1} \\
&= \frac{2(1-\theta)}{\theta^2} \\
\mathrm{Var}X &= \frac{2(1-\theta)}{\theta^2} + \frac{1}{\theta} - \frac{1}{\theta^2} = \frac{1}{\theta^2}
\end{aligned}$$

3.3 Funkcije izvodnice momenata

X diskretna ili neprekidna sl. var.

Funkcija izvodnica momenata je definirana s

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

za $t \in \mathbb{R}$ za koje gornje očekivanje postoji.

Teorem jedinstvenosti

Funkcija izvodnica momenata jedinstveno određuje razdiobu: $X \stackrel{d}{=} Y$ ako i samo ako je $M_X = M_Y$

$t \mapsto e^{tX}$ razvijemo u Taylorov red oko 0 i formalno izračunamo očekivanje (npr. ako je M_X definirana na okolini 0 ili ako je X nenegativna):

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

Zašto ime f.i. *momenata*?

Ako znamo sve momente $\mathbb{E}[X^k]$, onda znamo i M_X pa je razdioba od X jednoznačno određena.

Funkcija izvodnica momenata linearne transformacije

$Y = aX + b,\ a, b \in \mathbb{R}$

$$M_Y(t) = \mathbb{E}[e^t(aX + b)] = e^{bt}\mathbb{E}[e^{atX}] = e^{bt}M_X(at).$$

Primjer 3.2

U slučaju $\mathrm{Im}X = \{0, 1, 2, \ldots\}$ vrijedi

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

Npr. za $X \sim b(n, \theta)$ dobijemo

$$M_X(t) = (\theta e^t + 1 - \theta)^n = (1 + \theta(e^t - 1))^n.$$

Primjer 3.3

(a) $X \sim \Gamma(\alpha, \frac{1}{\lambda})$, $\alpha, \lambda > 0$

$$M_X(t) = \left(\frac{\lambda}{\lambda - t}\right)^{\alpha}, \quad t < \lambda$$

$$M_X'(t) = \alpha\lambda^{\alpha}(\lambda - t)^{-(\alpha+1)}$$

$$\Longrightarrow \mathbb{E}X = M_X'(0) = \frac{\alpha}{\lambda}$$

$$M_X''(t) = \alpha(\alpha+1)\lambda^{\alpha}(\lambda - t)^{-(\alpha+2)}$$

$$\Longrightarrow \mathbb{E}[X^2] = M_X''(0) = \frac{\alpha(\alpha+1)}{\lambda^2}$$

$$\Longrightarrow \mathrm{Var}X = \mathbb{E}[X^2] - (\mathbb{E}X)^2 = \frac{\alpha^2}{\lambda^2} - \frac{\alpha(\alpha+1)}{\lambda^2} = \frac{\alpha}{\lambda^2}$$

Specijalno,

- za $X \sim Exp(\lambda) \sim \Gamma(1, \frac{1}{\lambda})$, $\lambda > 0$ dobijemo

$$M_X(t) = \frac{\lambda}{\lambda - t}, \quad t < \lambda.$$

- za $X \sim \chi^2(n) \sim \Gamma(\frac{n}{2}, \frac{1}{\frac{1}{2}})$, $n \in \mathbb{N}$ dobijemo

$$M_X(t) = \left(\frac{\frac{1}{2}}{\frac{1}{2} - t}\right)^{\frac{n}{2}} = \frac{1}{(1-2t)^{\frac{n}{2}}}, \quad t < \frac{1}{2}.$$

(b) $X \sim N(\mu, \sigma^2)$, $\mu \in \mathbb{R}$, $\sigma^2 > 0$

$$M_X(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}, \quad t \in \mathbb{R}.$$

$$M_X'(t) = (\mu + \sigma^2 t) M_X(t) \implies \mathbb{E}X = M_X'(0) = \mu$$

$$M_X''(t) = \sigma^2 M_X(t) + (\mu + \sigma^2 t)^2 M_X(t)$$
$$\implies \mathbb{E}[X^2] = M_X''(0) = \sigma^2 + (\mu + \sigma^2)^2$$
$$\implies \mathrm{Var}X = \mathbb{E}[X^2] - (\mathbb{E}X)^2 = \sigma^2$$

Neka je $X \sim N(\mu, \sigma^2)$. Tada je $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$ pa je

$$M_Z(t) = e^{\frac{1}{2}t^2} = 1 + \underbrace{\frac{2^{-1}}{1!}}_{=\frac{\mathbb{E}[Z^2]}{2!}} t^2 + \underbrace{\frac{2^{-2}}{2!}}_{=\frac{\mathbb{E}[Z^4]}{4!}} t^4 + \ldots + \underbrace{\frac{2^{-n}}{n!}}_{=\frac{\mathbb{E}[Z^{2n}]}{(2n)!}} t^{2n} + \ldots$$

odakle slijedi

$$\mathbb{E}[Z^{2n}] = \frac{(2n)!}{2^n n!} \qquad \mathbb{E}[Z^{2n+1}] = 0, \quad n = 0, 1, 2, \ldots.$$

Posebno,

$$\mathbb{E}Z = \mathbb{E}[Z^3] = \mathbb{E}[Z^5] = 0,$$
$$\mathbb{E}[Z^2] = 1, \mathbb{E}[Z^4] = 3, \mathbb{E}[Z^6] = 15$$

i, budući da je $X = \mu + \sigma Z$,

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

Također, treći i četvrti centralni momenti su

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

3.4 Funkcije izvodnice kumulanata

Funkcija izvodnica kumulanata sl. var. X je definirana s

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

za $t \in \mathbb{R}$ za koje je $M_X(t)$ definirana.

r-ti kumulant κ_r je definiran preko

$$C_X(t) = \sum_{r=0}^{\infty} \kappa_r \frac{t^r}{r!}$$

Uočimo da vrijedi

$$C'_X(t) = \frac{M'_X(t)}{M_X(t)}$$

$$C''_X(t) = \frac{M''_X(t)M_X(t) - M'_X(t)^2}{M_X(t)^2}$$

Koristeći

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

dobijemo

$$\kappa_1 = C'_X(0) = \frac{M'_X(0)}{M_X(0)} = \mathbb{E}X$$

$$\kappa_2 = C''_X(0) = \frac{M''_X(0)M_X(t) - M'_X(0)^2}{M_X(0)^2}$$

$$= \mathbb{E}[X^2] - (\mathbb{E}X)^2 = \mathrm{Var}X$$

Zadatak 3.1

Funkcija izvodnica kumulanata slučajne varijable X je

$$C_X(t) = 2\left(\frac{1}{(1-t)^{10}} - 1\right).$$

Izračunajte matematičko očekivanje, drugi moment i varijancu sl. var. X.

Zadatak 3.2

Neka je $X \sim U(0,1)$.

(a) Izračunajte funkciju izvodnicu momenata sl. var. $Y = -\ln X$

(b) Odredite razdiobu od X.

4. Zajednička razdioba slučajnih varijabli

4.1 Zajednička gustoća i funkcija distribucije

X i Y su s.v. definirane na istom vjerojatnosnom prostoru.

Pretpostavimo: (X, Y) je diskretan s. vektor

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

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

Tablica zajedničke razdiobe od (X, Y):

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

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

Zajednička funkcija vjerojatnosti (*gustoća*) od X, Y:

$f_{X,Y} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$,

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

Svojstva:

(*G*1) $f_{X,Y}(x,y) \geq 0$ za sve x, y

(*G*2) $\sum_{x \in \mathsf{Im}X, y \in \mathsf{Im}Y} f_{X,Y}(x,y) = 1.$

Marginalne razdiobe:

- gustoća od X je

$$f_X(x) = \sum_{y \in \mathsf{Im} Y} f_{X,Y}(x, y)$$

- gustoća od Y je

$$f_Y(y) = \sum_{x \in \mathsf{Im} X} f_{X,Y}(x, y)$$

Kovarijanca slučajnih varijabli

$$\begin{aligned}\mathsf{Cov}(X,Y) &= E[(X-\mathbb{E}X)(Y-\mathbb{E}Y)] \\ &= \mathbb{E}[XY]-\mathbb{E}X\mathbb{E}Y\end{aligned}$$

Zajednička funkcija distribucije od X i Y:

$F_{X,Y} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$,

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

(X, Y) diskretan s. vektor $\Rightarrow$

$$F_{X,Y}(x, y) = \sum_{\{a \in \mathsf{Im} X : a \leq x\}} \sum_{\{b \in \mathsf{Im} X : b \leq y\}} f_{X,Y}(a, b)$$

za sve $x, y \in \mathbb{R}$.

Primjer 4.1

Bacamo dvije simetrične igraće kocke: crvenu i plavu.

$X =$ broj koji se okrenuo na crvenoj kocki
$Y =$ manji od okrenutih brojeva

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

Neprekidni s. vektor (X, Y):

Za funkciju gustoće $f_{X,Y} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ je

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

za sve $a < b, c < d$.

Svojstva:

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

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

Zadatak 4.1

Je li $f(x, y) = 6x^2y$, $0 < x, y < 1$ funkcija gustoće neprekidnog slučajnog vektora (X, Y)? Ako jest, izračunajte $P(0 < X < \frac{1}{2}, \frac{1}{2} < Y < 1)$.

Marginalne razdiobe:

- gustoća od X je

$$f_X(x) = \int_{-\infty} f_{X,Y}(x, y)\, dy$$

- gustoća od Y je

$$f_Y(y) = \int_{-\infty} f_{X,Y}(x, y)\, dx$$

Za funkciju distribucije vrijedi:

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

za sve $x, y \in \mathbb{R}$, i

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

Zadatak 4.2

Zadana je funkcija

$$F(x,y) = 1 - e^{-x} - e^{-2y} + e^{-(x+2y)}, \quad x, y > 0.$$

Je li F funkcija distribucije neprekidnog sl. vektora (X, Y)? U slučaju da jest odredite distribuciju sl. var. X i Y.

4.3 Uvjetna razdioba

– zadaje se *uvjetnim* gustoćama

Neka je (X, Y) diskretan s. vektor:
Uvjetna funkcija vjerojatnosti (ili *uvjetna gustoća*) *od* X *za dano* $Y = y$:

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

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

(ukoliko je $f_Y(y) > 0$)
Analogno: $f_{Y|X}(y|x)$

Primjer 4.3

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

$\rightarrow$

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

Za neprekidni s. vektor (X, Y),
uvjetna gustoća od X *za dano* $Y = y$:

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

(ukoliko je $f_Y(y) > 0$)

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

4.4 Nezavisnost slučajnih varijabli

X i Y su *nezavisne* s.v. ako

$$f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y)$$

za sve $y \in \mathrm{Im}Y, x \in \mathrm{Im}X$

$\Longleftrightarrow$

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

Diskretne s.v. X, Y su nezavisne akko

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

Neprekidne s.v. X, Y su nezavisne akko

$$\mathbb{P}(a \leq X \leq b, c \leq Y \leq d) = \mathbb{P}(a \leq X \leq b) \cdot \mathbb{P}(c \leq Y \leq d)$$

za sve $a < b, c < d$.

X, Y nezavisne s.v. $\Rightarrow$ $g(X)$, $h(Y)$ su nezavisne s.v.

Def. X_1, X_2,... su *nezavisne* s.v. ako
$(\forall k \geq 2)\ (\forall i_1, i_2, \ldots, i_k)\ (\forall x_1, \ldots, x_k)$

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

(X, Y) s. vektor, $g : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$
$\Rightarrow g(X, Y) = g \circ (X, Y)$ je s.v.

Za (X, Y) diskretan s. vektor:

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

Za (X, Y) neprekidan s. vektor:

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

Vrijedi:

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

X, Y nezavisne s.v. $\Rightarrow$

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

$$X, Y \text{ nezavisne} \implies \operatorname{Var}(X+Y) = \operatorname{Var}X + \operatorname{Var}Y$$

$$\begin{aligned}
\operatorname{Var}(X+Y) &= \mathbb{E}[(X+Y-\mathbb{E}[X+Y])^2] \\
&= \mathbb{E}[((X-\mathbb{E}X)+(Y-\mathbb{E}Y))^2] \\
&= \mathbb{E}[(X-\mathbb{E}X)^2] + 2\mathbb{E}[(X-\mathbb{E}X)(Y-\mathbb{E}Y)] \\
&\quad + \mathbb{E}[(Y-\mathbb{E}Y)^2] \\
&\overset{\text{nez.}}{=} \operatorname{Var}X + 2\underbrace{\mathbb{E}[X-\mathbb{E}X]}_{=\mathbb{E}X-\mathbb{E}X=0}\mathbb{E}[Y-\mathbb{E}Y] + \operatorname{Var}Y
\end{aligned}$$

Dokaz pomoću f.i. kumulanata:

$$C_{X+Y}(t) = \ln(M_X(t)M_Y(t))) = C_X(t) + C_Y(t)$$

$$\begin{aligned}
\implies \operatorname{Var}(X+Y) &= C''_{X+Y}(0) = C''_X(0) + C''_Y(0) \\
&= \operatorname{Var}X + \operatorname{Var}Y.
\end{aligned}$$

- $X_1, \ldots, X_n$ nezavisne

$$\begin{aligned}\mathrm{Var}(X_1 + \ldots + X_n) &= \mathrm{Var}X_1 + \mathrm{Var}(X_2 + X_3 + \ldots + X_n) \\ &= \mathrm{Var}X_1 + \mathrm{Var}X_2 + \mathrm{Var}(X_3 + \ldots + X_n) \\ &= \ldots = \\ &= \mathrm{Var}X_1 + \mathrm{Var}X_2 + \mathrm{Var}X_n + \ldots + \mathrm{Var}X_n\end{aligned}$$

Zadatak 4.3

Neka su $X \sim Exp(1)$ i $Y \sim U(0,1)$ nezavisne slučajne varijable. Izračunajte $\mathbb{P}(X+Y \geq 1)$.

Zadatak 4.4

Slučajni vektor (X,Y) ima gustoću

$$f(x,y) = xe^{-x-xy}, \quad x,y > 0.$$

Izračunajte $\mathbb{E}[\frac{1}{X(Y+1)}]$. Jesu li slučajne varijable X i Y nezavisne?

Zadatak 4.5

Simetrična kocka se baca 2 puta. Označimo s X manji, a s Y veći od brojeva koji su pali. Jesu li X i Y nezavisne sl. var.?

Nezavisnost i funkcije izvodnice

Neka su $X_1, \ldots, X_n$ nezavisne slučajne varijable i $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$. Tada je

$$M_{\alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_n X_n}(t) = M_{X_1}(\alpha_1 t) M_{X_2}(\alpha_2 t) \cdots M_{X_n}(\alpha_n t)$$

(za sve $t \in \mathbb{R}$ za koje su sve f.i.m. definirane).

$$L.S. = \mathbb{E}[e^{t(\alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_n X_n)}] = \mathbb{E}[e^{t\alpha_1 X_1} e^{t\alpha_2 X_2} \cdots e^{t\alpha_n X_n}]$$

$$\overset{\text{nez.}}{=} \underbrace{\mathbb{E}[e^{\alpha_1 t X_1}]}_{M_{X_1}(\alpha_1 t)} \underbrace{\mathbb{E}[e^{\alpha_2 t X_2}]}_{M_{X_2}(\alpha_2 t)} \cdots \underbrace{\mathbb{E}[e^{\alpha_n t X_n}]}_{M_{X_n}(\alpha_n t)} = D.S.$$

Neka su $X_1, \ldots, X_n$ nezavisne slučajne varijable s vrijednostima u skupu $\{0, 1, 2, \ldots\}$. Tada je

$$G_{X_1+X_2+\ldots+X_n}(t) = G_{X_1}(t)G_{X_2}(t)\cdots G_{X_n}(t)$$

Primjer 4.4

$X_1, \ldots, X_n \sim$ Bernoullijeva(θ) nezavisne

$$\begin{aligned} G_{X_1+\ldots+X_k}(t) &= G_{X_1}(t)G_{X_2}(t)\cdots G_{X_n}(t) \\ &= (1-\theta+\theta t)^n \end{aligned}$$

$$\Rightarrow X_1 + \ldots + X_k \sim b(n, \theta)$$

Primjer 4.5

$X_1, \ldots, X_n \sim$ geometrijska(θ) nezavisne

$$
\begin{aligned}
G_{X_1+\ldots+X_k}(t) &= G_{X_1}(t) \cdots G_{X_k}(t) \\
&= \frac{\theta t}{1-t(1-\theta)} \cdots \frac{\theta t}{1-t(1-\theta)} \\
&= \left(\frac{\theta t}{1-t(1-\theta)} \right)^k
\end{aligned}
$$

$\Rightarrow X_1 + \ldots + X_k \sim$ negativna binomna(k, θ)

Zadatak 4.6

Neka su $X \sim P(\lambda)$ i $Y \sim P(\nu)$ nezavisne slučajne varijable, $\lambda, \mu > 0$.

(a) Dokažite da $S = X + Y \sim P(\lambda + \mu)$.

(b) Dokažite da je uvjetna distribucija od X uz uvjet $S = s$ binomna. Odredite joj parametre.

4.10. Uvjetno očekivanje

(X, Y) slučajni vektor

Uvjetno očekivanje od Y uz dano $X = x$ je definirano:

- za diskretni sl. vektor s

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

- za neprekidni sl. vektor s

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

Uz $g(x) = \mathbb{E}[Y|X = x]$ definiramo *uvjetno očekivanje*

$$\mathbb{E}[Y|X] = g(X).$$

Nap. (a) Ako su X i Y nezavisne, onda je

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

(b) $\mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}X$

$$\begin{aligned}
\rightarrow L.S. &= \int_{\{x:f_X(x)>0\}} \mathbb{E}[Y|X=x] f_X(x)\,dx \\
&= \int_{\{x:f_X(x)>0\}} \int_{-\infty}^{\infty} y \underbrace{f_{Y|X}(y|x) f_X(x)}_{=f_{X,Y}(x,y)}\,dy\,dx \\
&= \int_{-\infty}^{\infty} y \underbrace{\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx}_{=f_Y(y)}\,dy \\
&= \int_{-\infty}^{\infty} y f_Y(y)\,dy = D.S.
\end{aligned}$$

Uvjetna varijanca

$$g(x) := \mathrm{Var}[Y|X=x] = \mathbb{E}[Y^2|X=x] - \mathbb{E}[Y|X=x]^2$$

Uvjetna varijanca je definirana s $Var[Y|X] = g(X)$ i tada vrijedi

$$\mathrm{Var}[\mathbb{E}[Y|X]] = \mathrm{Var}Y - \mathbb{E}[\mathrm{Var}[Y|X]].$$

Dokaz.

$$\mathbb{E}[\mathrm{Var}[Y|X]] = \underbrace{\mathbb{E}[\mathbb{E}[Y^2|X]]}_{=\mathbb{E}[Y^2]} - \mathbb{E}[\mathbb{E}[Y|X]^2]$$

$$\mathrm{Var}[E[Y|X]] = \mathbb{E}[\mathbb{E}[Y|X]^2] - \left(\underbrace{\mathbb{E}[\mathbb{E}[Y|X]]}_{=\mathbb{E}[Y]}\right)^2$$

Zadatak 4.7

Broj odlazaka aktuara s posla nakon redovnog radnog vremena tijekom radnog tjedna modelira se pomoću binomne slučajne varijable X s parametrima (n, θ) gdje je $n = 5$, a $\theta = \frac{4}{5}$. Za uvjetnu razdiobu ukupnog vremena Y koje je aktuar proveo na poslu tijekom tjedna (u satima) ako je taj tjedan morao na poslu ostati dulje x dana, vrijedi:

$$\mathbb{E}[Y|X = x] = 4(x + 10), \quad \mathrm{Var}[Y|X = x] = x.$$

(a) Koliko u srednjem sati aktuar provodi u uredu tijekom tjedna?

(b) Izračunajte $\mathrm{Var}Y$.

Funkcija izvodnica momenata slučajne sume

$X_1, X_2, \ldots$ nezavisne i jednako distribuirane slučajne varijable s f.i.m. $M(t)$ i N sl. var. s vrijednostima u $\{0, 1, 2, \ldots\}$, f.i.v. $G(t)$ nezavisna od $X_1, X_2, \ldots$.
Tada je f.i.m. slučajne sume

$$S = X_1 + X_2 + \ldots + X_N \quad (\text{konvencija: } S = 0 \text{ za } N = 0).$$

dana s

$$M_S(t) = G(M(t)).$$

Zadatak 4.8

Broj šteta N po portfelju istovrsnih nezavisnih polica osiguranja ima Poissonovu razdiobu s očekivanjem $\mu > 0$. Kada se šteta dogodi, njezin iznos $X_i (i = 1, 2, \ldots)$ ima gama razdiobu $\Gamma(\alpha, \frac{1}{\lambda})$, $\alpha, \lambda > 0$ i iznosi šteta su međusobno nezavisni te nezavisni od broja šteta.

Označimo sa $S = X_1 + \ldots + X_N$ ukupni iznos šteta u tom portfelju.

Izrazite $\mathbb{E}S$ i $\mathrm{Var}S$ preko parametara μ, α, λ.

Zadatak 4.9

Neka su $X_1, \ldots, X_n \sim Exp(\lambda)$, $\lambda > 0$ nezavisne slučajne varijable. Dokažite:

$$S = X_1 + \ldots + X_n \sim \Gamma(n, \frac{1}{\lambda}).$$

Zadatak 4.10

Neka su $X \sim N(\mu_1, \sigma_1^2)$ i $Y \sim N(\mu_2, \sigma_2^2)$ nezavisne slučajne varijable. Dokažite:

$$X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2).$$

Zadatak 4.11

Neka su $X \sim \Gamma(\alpha, \frac{1}{\lambda})$ i $Y \sim \Gamma(\beta, \frac{1}{\lambda})$ nezavisne slučajne varijable, $\alpha, \beta, \lambda > 0$.

(a) Izračunajte

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

(b) Odredite razdiobu od $Z = X + Y$.

5. Centralni granični teorem

Neka je $X_1, X_2, \ldots$ *niz n.j.d. s. v.,*

$$\mu = \mathbb{E}X_1, 0 < \mathrm{Var}X_1 = \sigma^2 < +\infty$$

i

neka je

$$\overline{X}_n := \frac{X_1 + X_2 + \cdots + X_n}{n}, \ n \in \mathbb{N}.$$

Tada za sve $a < b$ *vrijedi*

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

gdje je $\Phi(x)$ *funkcija distribucije od* $N(0,1)$.

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

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

$$\frac{\overline{X} - \mu}{\sigma}\sqrt{n} \dot{\sim} N(0,1) \text{ za veliko } n,$$

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} \dot{\sim} N(0,1) \text{ za veliko } n.$$

$$\overline{X} \dot{\sim} N(\mu, \frac{\sigma^2}{n}) \text{ za veliko } n,$$

$$\sum_{i=1}^{n} X_i \dot{\sim} N(n\mu, n\sigma^2) \text{ za veliko } n.$$

5.2 Normalna aproksimacija

Primjer 5.1(binomna razdioba)

$X \sim b(n, \theta)$

$X \stackrel{d}{=} X_1 + \ldots + X_n,$

$X_1, \ldots, X_n \sim$ Bernoullijeva(θ) nezavisne

$$\mu = \mathbb{E}X_i = \theta \quad \sigma^2 = \mathrm{Var}X_i = \theta(1-\theta)$$

CGT $\Longrightarrow$

$$X \stackrel{\cdot}{\sim} N(n\theta, n\theta(1-\theta))$$

Nap. Aproksimacija je dobra ako je

$$n\theta \geq 5 \quad \text{i} \quad n(1-\theta) \geq 5.$$

Primjer 5.2(Poissonova razdioba)

$X_1, \ldots, X_n \sim P(\lambda)$ nezavisne

$$\mu = \mathbb{E}X_i = \lambda \quad \text{i} \quad \sigma^2 = \mathrm{Var}X_i = \lambda$$

CGT $\Longrightarrow$

$$X_1 + \ldots + X_n \approx N(n\lambda, n\lambda)$$

Uočimo da je $X := X_1 + \ldots + X_n \sim P(n\lambda)$ pa slijedi

$$P(\lambda) \approx N(\lambda, \lambda) \quad \text{za velike } \lambda > 0.$$

Nap. Aproksimacija je dobra za $\lambda > 5$.

Primjer 5.2(Gama razdioba)

$X_1, \ldots, X_n \sim Exp(\lambda)$ nezavisne

$$\mathbb{E}X_i = \frac{1}{\lambda} \qquad \mathrm{Var}X_i = \frac{1}{\lambda^2}.$$

Po Zadatku 4.8,

$$X = X_1 + \ldots + X_n \sim \Gamma(n, \frac{1}{\lambda}).$$

$$\text{CGT} \implies X \stackrel{\cdot}{\sim} N(\frac{n}{\lambda}, \frac{n}{\lambda^2}).$$

Slično se pokaže (za veliki n):

$$\chi^2(n) \sim \Gamma(\frac{n}{2}, 2) \stackrel{\cdot}{\sim} N(n, 2n).$$

5.3 Korekcija zbog neprekidnosti

Kod aproksimacije diskretnih slučajnih varijabli aproksimiramo vjerojatnosti događaja

$$\{X = x\}.$$

Aproksimativna vjerojatnost se računa tako da se promatra vjerojatnost da X upadne u neki interval. Npr. za $X \sim P(\lambda)$

$$\mathbb{P}(X = 5) = \mathbb{P}(4.5 < X < 5.5)$$

$$\mathbb{P}(X \geq 10) = \mathbb{P}(X > 9.5).$$

Ovakav postupak zovemo *korekcija zbog neprekidnosti*.

Zadatak 5.1

Iz portfelja istovrsnih polica na slučajan način je izabrano njih 500. Poznato je da se šteta po jednoj polici tijekom godine pojavljuje s vjerojatnosti 0.04 neovisno o ostalim policama. Po jednoj polici osiguranja moguća je najviše jedna šteta. Izračunajte (približno) vjerojatnost da na kraju godine u uzorku neće biti više od 30 šteta.

Outline

6. Uzorkovanje

- populacija je beskonačna (iako su populacije konačne, ali velike: osiguranici, police osiguranja, ...)
- želimo zaključiti nešto o populaciji (npr. procijeniti neki parametar populacije) uzimanjem slučajnog uzorka

Def. *Slučajni uzorak* je niz nezavisnih i jednako distribuiranih slučajnih varijabli $X_1, \ldots, X_n$. Tada je $\underline{X} = (X_1, \ldots, X_n)$ slučajni vektor.

- slučajni uzorak $\rightarrow$ mjerenja (opažanja) sl. veličine X vezane uz populaciju koja se proučava
- svaki element populacije ima jednaku šansu da bude odabran u sl. uzorak
- θ parametar o kojem ovisi populacija (nepoznat) $\rightarrow X$ ovisi o θ

Def. Uređena n-torka $\underline{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$, koja je realizacija slučajnog uzorka $\underline{X}$ se zove *opaženi uzorak*.

Def. *Statistika* je funkcija slučajnog uzorka koja ne sadrži nepoznate parametre.
Npr.

- *uzoračka sredina*

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

- *uzoračka varijanca*

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

Dakle, statistika je općenito oblika $g(\underline{X})$.

Ako je $\mu = \mathbb{E}X$, onda μ ovisi o parametru populacije pa npr.

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

nije statistika! Zato promatramo $\overline{X}$.

Uočimo (ako je populacijska varijanca konačna):

$$\overline{X} \dot{\sim} N(\mathbb{E}X, \frac{\mathrm{Var}X}{n})(\text{asimptotska normalnost!})$$

$\underline{X} = (X_1, \ldots, X_n)$ slučajni uzorak (duljine n) iz populacije u kojoj populacijska razdioba X ima očekivanje μ i varijancu σ^2

Vrijedi

$$\mathbb{E}[\overline{X}] = \mu \qquad \mathrm{Var}\overline{X} = \frac{\sigma^2}{n}$$

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

Uzoračke razdiobe statistika normalnog uzorka

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

Uzoračka sredina

Vrijedi

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

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

Uočimo:

$$\mathbb{E}[(\overline{X} - \underbrace{\mu}_{=\mathbb{E}\overline{X}})^2] = \mathrm{Var}X = \frac{\sigma^2}{n} \underset{n\to\infty}{\longrightarrow} 0$$

Uzoračka varijanca

Vrijedi

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

Uočimo:

$$\mathbb{E}[(S^2-\sigma^2)^2] = \mathrm{Var}S^2 = \frac{\sigma^4}{(n-1)^2}\mathrm{Var}(\frac{(n-1)S^2}{\sigma^2})$$
$$= \frac{\sigma^4}{(n-1)^2}(n-1) = \frac{\sigma^4}{n-1} \underset{n\to\infty}{\longrightarrow} 0$$

Pokazuje se da su sl. var. $\overline{X}$ i S^2 nezavisne.

Studentova razdioba

Ako su $Z \sim N(0,1)$ i $V \sim \chi^2(k)$ nezavisne, onda slučajna varijabla

$$\frac{Z}{\sqrt{\frac{V}{k}}}$$

ima *Studentovu* ili *t-razdiobu s k stupnjeva slobode*. Oznaka za ovu razdiobu je $t(k)$.

Pokazuje se da je funkcija gustoće dana s:

$$\underbrace{\frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k\pi}\Gamma\left(\frac{k}{2}\right)}}_{\underset{k\to\infty}{\to}\frac{1}{\sqrt{2\pi}}}\underbrace{\left(1+\frac{x^2}{k}\right)^{-\frac{k+1}{2}}}_{\underset{k\to\infty}{\to}e^{-\frac{x^2}{2}}},\quad x\in\mathbb{R}.$$

Može se pokazati da vrijedi:

$$t(n)\xrightarrow{d}N(0,1), n\to\infty.$$

Sl. var. $X \sim t(k)$ ima očekivanje za $k > 1$, a varijancu za $k > 2$ i tada je:

$$\mathbb{E}X = 0 \qquad \mathrm{Var}X = \frac{k}{k-2}.$$

Specijalni slučaj $k = 1$. Tada X ima (jediničnu) *Cauchyjevu razdiobu*: gustoća je

$$f_X(x) = \frac{1}{\pi(1+x^2)}.$$

Ako je parametar σ poznat, onda je

$$\frac{\overline{X}-\mu}{\sigma}\sqrt{n}\begin{cases}\sim N(0,1) & \text{ako je } \underline{X} \text{ iz normalne populacije}\\ \dot{\sim} N(0,1) & \text{ako je } 0<\sigma^2<\infty\end{cases}$$

Što ako je parametar σ nepoznat?

Tada koristimo

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

gdje je $S = \sqrt{S^2}$ *uzoračka standardna devijacija*.
Za uzorak iz normalne populacije vrijedi

$$\frac{\overline{X} - \mu}{\sigma}\sqrt{n} \sim N(0,1) \quad \text{i} \quad \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

pa iz nezavisnosti zaključujemo da

$$T = \frac{\frac{\overline{X}-\mu}{\sigma}\sqrt{n}}{\sqrt{\frac{\frac{(n-1)S^2}{\sigma^2}}{n-1}}} \sim t(n-1).$$

Ako populacija nije normalna, ali ima konačnu varijancu, onda je

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

jer je po CGT

$$T = \underbrace{\frac{X_1 + \ldots + X_n - n\mu}{\sigma\sqrt{n}}}_{\xrightarrow{d} N(0,1)} \sqrt{\underbrace{\frac{S^2}{\sigma^2}}_{\to 1}} \xrightarrow{d} N(0,1)$$

Fisherova F-razdioba

Ako su $U \sim \chi^2(\nu_1)$ i $V \sim \chi^2(\nu_2)$ nezavisne, onda slučajna varijabla

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

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

Oznaka za ovu razdiobu je $F(\nu_1, \nu_2)$.

Promotrimo dva nezavisna slučajna uzorka duljina n_1 i n_2 iz normalno distribuiranih populacija s varijancama σ_1^2 i σ_2^2.

Tada je $S_i^2/\sigma_i^2 \sim \chi^2(n_i - 1)$ pa je

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

(Ako populacije nisu normalne, onda ovo ne mora vrijediti.)

Kvantil

Za sl. var. X i $\alpha \in (0,1)$ definiramo $(1-\alpha)$-*kvantil* x_α s

$$\mathbb{P}(X \geq x_\alpha) = \alpha.$$

Kvantili su obično tabelirani:

- $X \sim N(0,1)$

  $$\mathbb{P}(X \geq z_\alpha) = \alpha \qquad \text{npr. } z_{0.05} = 1.64$$

- $X \sim t(k)$

  $$\mathbb{P}(X \geq t_\alpha(k)) = \alpha \qquad \text{npr. } t_{0.025}(10) = 2.281$$

- $X \sim F(n_1, n_2)$

  $$\mathbb{P}(X \geq f_\alpha(n_1, n_2)) = \alpha \qquad \text{npr. } f_{0.1}(15,5) = 2.27$$

Vrijedi:

$$X \sim F(\nu_1, \nu_2) \iff Y := \frac{1}{X} \sim F(\nu_2, \nu_1)$$

pa je

$$\begin{aligned}\alpha &= \mathbb{P}(X \geq f_\alpha(\nu_1, \nu_2)) = \mathbb{P}(\frac{1}{Y} \geq f_\alpha(\nu_1, \nu_2)) \\ &= \mathbb{P}(Y \leq \frac{1}{f_\alpha(\nu_1, \nu_2)}) = 1 - \mathbb{P}(Y > \frac{1}{f_\alpha(\nu_1, \nu_2)}),\end{aligned}$$

odakle zaključujemo

$$f_{1-\alpha}(\nu_2, \nu_1) = \frac{1}{f_\alpha(\nu_1, \nu_2)}.$$

7. Točkovne procjene

- procjena parametara populacijske razdiobe
- pomoću statistika
- populacijska razdioba je opisana gustoćom

$$f(x|\theta) \qquad \theta \text{ nepoznati parametar}$$

- 2 metode:
  - metoda momenata
  - metoda maksimalne vjerodostojnosti

7.1 Metoda momenata

- izjednačavanje populacijskih momenata s odgovarajućim uzoračkim momentima i rješavanje sustava
- procjenitelj je statistika
- procjena ce biti realizacija procjenitelja na opaženom uzorku

7.1.1 Slučaj jednog parametra

Populacijska razdioba ovisi samo o jednom parametru θ: gustoća je $f(x|\theta)$.

Ako je $\underline{x}$ opaženi uzorak, onda je *procjena* od θ *metodom momenata* rješenje jednadžbe

$$\overline{x} = \mu(\theta)\,,$$

gdje je

$$\mu(\theta) = \mathbb{E}X = \begin{cases} \sum\limits_{x \in \mathsf{Im} X} x f(x|\theta) & X \text{ diskretna} \\ \int_{-\infty}^{\infty} x f(x|\theta)\, dx & X \text{ neprekidna} \end{cases}$$

$\hat{\theta} = \hat{\theta}(\underline{x})$ procjena

$\hat{\theta} = \hat{\theta}(\underline{X})$ procjenitelj

Primjer 7.1

Procijenimo parametar $\lambda > 0$ iz populacije s populacijskom razdiobom koja je $Exp(\lambda)$. Neka je $\underline{X} = (X_1, \ldots, X_n)$ slučajni uzorak.

$$\mu(\lambda) = \underbrace{\mathbb{E}X}_{=\frac{1}{\lambda}} = \overline{x} \implies \lambda = \frac{1}{\overline{x}}$$

Procjenitelj metodom momenata je

$$\hat{\lambda} = \hat{\lambda}(\underline{X}) = \frac{1}{\overline{X}}.$$

Primjer 7.2

Populacijska razdioba je $U(-\theta, \theta)$, $\theta > 0$ nepoznati parametar.

- $\mu(\theta) = \mathbb{E}X = \int_{-\theta}^{\theta} x \frac{dx}{2\theta} = 0$
  $\rightarrow$ parametar θ se ne pojavljuje u 1. momentu
- $VarX = \frac{\theta^2}{3}$ izjednačimo s opaženom uzoračkom varijancom:

$$\frac{\theta^2}{3} = s^2 \implies \theta = s\sqrt{3}.$$

Procjenitelj metodom momenata je
$\hat{\theta} = \hat{\theta}(\underline{X}) = S\sqrt{3}$,
gdje je $S = \sqrt{S^2}$ uzoračka standardna devijacija.

7.1.2 Slučaj dva parametra

$\theta = (\theta_1, \theta_2)$ dvodimenzionalni populacijski parametar
Izjednačavanjem prva dva momenta se dobije sustav

$$\mathbb{E}X = \overline{x}$$

$$\mathbb{E}[X^2] = \frac{1}{n}\sum_{i=1}^{n} x_i^2 \quad (\text{ ili } \operatorname{Var}X = s^2)$$

Primjer 7.3

$N(\mu, \sigma^2)$ populacija $\implies \mathbb{E}X = \mu, \operatorname{Var}X = \sigma^2$

$$\implies \hat{\mu} = \overline{X} \quad \hat{\sigma}^2 = S^2$$

7.2 Metoda maksimalne vjerodostojnosti
Jednoparametarski slučaj

$\underline{x} = (x_1, x_2, \ldots, x_n)$ opaženi uzorak iz populacije s gustoćom $f(x|\theta)$.

Vjerodostojnost

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

Npr. $L(\theta)$ je vjerojatnost realizacije opaženog uzorka u diskretnom slučaju

Procjena metodom maksimalne vjerodostojnosti parametra θ je vrijednost $\hat{\theta}$ koja maksimizira funkciju $\theta \mapsto L(\theta)$, tj.

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

Procjenitelj metodom maksimalne vjerodostojnosti (MLE) je statistika $\hat{\theta}(\underline{X})$.

Dovoljno je maksimizirati *log-vjerodostojnost*

$$\ell(\theta) = \ln L(\theta).$$

Kandidati (u slučaju derivabilne funkcije ℓ) za $\hat{\theta}$ su rješenja jednadžbe

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

(ako ImX ne ovisi o θ). Može se pokazati da je za funkciju $g(\theta)$ od parametra

$$\text{MLE}\hat{g}(\theta) = g(\hat{\theta}).$$

Def. Procjenitelj $\hat{\theta} = \hat{\theta}(\underline{X})$ za parametar θ je *nepristran* ako je

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

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

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

Procjenitelj je *konzistentan* ako vrijedi

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

Npr. ako postoje i konačni su $\mu = \mathbb{E}X$ i $\sigma^2 = \mathrm{Var}X$, onda je $\overline{X}$ je nepristrani procjenitelj za populacijsko očekivanje μ

$$\mathbb{E}[\overline{X}] = \frac{1}{n}\sum_{i=1}^{n}\underbrace{\mathbb{E}[X_i]}_{=\mu} = \mu.$$

Također je i konzistentan:

$$MSE(\overline{X}) = \mathbb{E}[(\overline{X} - \mu)^2] = \mathrm{Var}\overline{X} =$$
$$= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}X_i = \frac{\sigma^2}{n} \to 0, \quad n \to \infty,$$

Zadatak 7.1

Nađite procjenitelja maksimalne vjerodostojnosti za parametar $\lambda > 0$ iz populacije s $Exp(\lambda)$-razdiobom.

Zadatak 7.2

Zadana je populacija s populacijskom gustoćom

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

i nepoznatim parametrom $\theta > 1$. Nađite MLE za θ.

Zadatak 7.3

Populacijska gustoća je Bernoullijeva s parametrom uspjeha $p \in (0, 1)$. Nađite MLE za p.
Kako biste procijenili parametar uspjeha binomne populacijske razdiobe s poznatim parametrom $m \in \mathbb{N}$?

7.2.3 Nepotpuni uzorci

- nepotpuni uzorak: rezani podaci ili cenzurirani podaci
- ako su npr. opažene vrijednosti

$$x_1, \ldots, x_n$$

i još znamo da je m opaženih vrijednosti veće od y

Vjerodostojnost je

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

Zadatak 7.4

U opaženom uzorku iz $Exp(\lambda)$-distribucije se nalaze vrijednosti $x_1, \ldots, x_n$ i za m vrijednosti se zna da je veće od $y > 0$. Nađite MLE za λ.

Zadatak 7.5

Podaci o štetama po 4000 polica osiguranja koje su bile pod rizikom točno godinu dana su prikazani frekvencijskom tablicom:

broj šteta i	frekvencija f_i
0	3288
1	642
2	66
≥ 3	4
ukupno	4000

Pretpostavimo da je broj šteta $X \sim P(\lambda)$. Odredite funkciju vjerodostojnosti te provjerite da je $\hat{\lambda} = 0.196551$ procjena maksimalne vjerodostojnosti na temelju danog opaženog uzorka.

8. Pouzdani intervali

- mjerenje točnosti (preciznosti) procjenitelja
- slučajni interval, ne mora biti jedinstven

Def. $(1-\alpha)\cdot 100\%$ *pouzdani interval* za θ je slučajni interval $[\hat{\theta}_1(\underline{X}), \hat{\theta}_2(\underline{X})]$ takav da je

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

Uočimo:

- θ je stvarna(prava) vrijednost parametra
- $\hat{\theta}_i(\underline{X})$ su statistike

8.1 Konstrukcija pouzdanih intervala

Pivotna metoda daje općenit postupak konstrukcije pouzdanog intervala.

Pretpostavimo da postoji *pivotna veličina* $g(\underline{X}, \theta)$ takva da je:

- funkcija uzorka i parametra
- ima poznat zakon razdiobe
- $\theta \mapsto g(\underline{X}, \theta)$ strogo monotona.

Odredimo $g_1 \leq g_2$ takve da je

$$\mathbb{P}(g_1 \leq g(\underline{X}, \theta) \leq g_2) = 1 - \alpha.$$

Ako je $h(\theta) = g(\underline{X}, \theta)$ str. rastuća, onda je

$$g_1 \leq g(\underline{X}, \theta) = h(\theta) \iff \underbrace{h^{-1}(g_1)}_{=:\hat{\theta}_1(\underline{X})} \leq \theta$$

$$g_2 \geq g(\underline{X}, \theta) = h(\theta) \iff \underbrace{h^{-1}(g_2)}_{=:\hat{\theta}_2(\underline{X})} \geq \theta$$

pa vrijedi

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

čime smo dobili $(1 - \alpha) \cdot 100\%$ pouzdani interval $[\hat{\theta}_1(\underline{X}), \hat{\theta}_2(\underline{X})]$.

Primjer 8.1

(a) $\underline{X}$ sl. uzorak duljine 20 iz $N(\mu, 10^2)$ populacije, opažena vrijednost $\overline{x} = 62.75$.

Pivotna veličina:

$$g(\underline{X}, \mu) = \frac{\overline{X} - \mu}{10}\sqrt{20} \sim N(0, 1)$$

Tada:

- $g(\underline{X}, \mu) \sim N(0, 1)$
- $\mu \mapsto g(\underline{X}, \mu)$ je strogo padajuća

Budući da je $\Phi(1.96) = 0.975$ i $\Phi(-1.96) = 0.025$, slijedi

$$\mathbb{P}(-1.96 \leq \frac{\overline{X} - \mu}{10}\sqrt{20} \leq 1.96) = 0.975 - 0.025 = 0.95$$

pa je

$$\begin{aligned} 0.95 &= \mathbb{P}(\overline{X} - 1.96\frac{10}{\sqrt{20}} \leq \mu \leq \overline{X} + 1.96\frac{10}{\sqrt{20}}) \\ &= \mathbb{P}(\overline{X} - 4.21 \leq \overline{X} + 4.59) \end{aligned}$$

Dakle, 95% pouzdani interval za μ je

$$[\overline{X} - 4.21, \overline{X} + 4.59].$$

Nap. Ovaj sl. interval je najkraće duljine (zbog oblika funkcije gustoće jedinične normalne razdiobe.

(b) Općenito, $(1-\alpha)\cdot 100\%$ pouzdani interval za parametar očekivanja μ iz $N(\mu, \sigma^2)$ populacije je dan s

$$[\overline{X} - z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}],$$

gdje je $z_{\alpha/2} > 0$ takav da je $\Phi(z_{\alpha/2}) = 1 - \alpha/2$.

Zadatak 8.1

Osiguravajuće društvo treba procjenu srednje vrijednosti šteta po policama određene klase koje su nastale tijekom prošle godine. Detaljni podaci o tim štetama sugeriraju da bi standardna devijacija mogla biti oko 450 kn. Ako se želi procijeniti srednja vrijednost iznosa šteta do na ± 80 kn točnosti uz 90% pouzdanosti, kolika je veličina uzorka potrebna?

Pouzdani intervali za parametre normalno distribuirane populacije

- populacijska sredina

$$\text{pivotna veličina} : \frac{\overline{X} - \mu}{S}\sqrt{n} \sim t(n-1)$$

Npr. 95%-pouzdani interval za μ je

$$[\overline{X} - t_{0.025}(n-1), \overline{X} + t_{0.025}(n-1)]$$

gdje je $\mathbb{P}(t(n-1) \geq t_{0.025}(n-1)) = 0.025.$

- populacijska varijanca

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

Tada je 95% pouzdani interval za σ^2

$$\left[\frac{(n-1)S^2}{\chi^2_{0.025}(n-1)}, \frac{(n-1)S^2}{\chi^2_{0.975}(n-1)}\right]$$

- asimetrija od $\chi^2(n-1) \implies$ pouzdani interval ne mora biti najkraći

Pouzdani intervali za parametre diskretnih populacija

- vjerojatnost pokrivanja $[\hat{\theta}_1(\underline{X}), \hat{\theta}_1(\underline{X})]$ ne mora biti točno $1-\alpha$ pa tražimo da bude $\geq 1-\alpha$

Primjer 8.2

Pouzdani intervali za binomnu razdiobu $X \sim b(n, \theta)$
MLE za θ je

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

- X ne sadrži θ (nije kandidat za pivotnu veličinu)
- npr. ako je x opažena vrijednost, 95% pouzdani interval za θ možemo odrediti iz uvjeta

$$\mathbb{P}_\theta(X \leq x) \geq 0.025 \quad \text{i} \quad \mathbb{P}_\theta(X \geq x) \geq 0.025.$$

Granice pouzdanog intervala određujemo iz ekvivalentnog uvjeta:

$$F(x|\theta) \geq 0.025 \quad \text{i} \quad 1 - F(x-1|\theta) \geq 0.025,$$

što možemo, jer je

$$\theta \mapsto F(x|\theta) \quad \text{strogo rastuća}$$

$$\Longrightarrow \ \theta \mapsto 1 - F(x-1|\theta) \quad \text{strogo rastuća.}$$

pa su granice pouzd. int. $[\hat{\theta}_1, \hat{\theta}_2]$ rješenja jednadžbi:

$$1 - F(x-1|\hat{\theta}_1) = 0.025 \quad \text{i} \quad F(x|\hat{\theta}_2) = 0.025$$

(numeričko rješavanje!).

Ako je n velik, onda

$$\frac{X - n\theta}{\sqrt{n\theta(1-\theta)}} \approx N(0,1),$$

ali i

$$\frac{X - n\theta}{\sqrt{n\hat{\theta}(1-\hat{\theta})}} \approx N(0,1),$$

odakle iz

$$1 - \alpha = \mathbb{P}(-z_{\alpha/2} \leq \frac{X - n\theta}{\sqrt{n\hat{\theta}(1-\hat{\theta})}} \leq z_{\alpha/2})$$

$$= \mathbb{P}(\frac{X}{n} - z_{\alpha/2}\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} \leq \theta \leq \frac{X}{n} + z_{\alpha/2}\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}})$$

dobijemo granice $(1-\alpha)\cdot 100\%$ pouzdanog intervala za θ:

$$\hat{\theta} \pm z_{\alpha/2}\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}.$$

Parametar Poissonove razdiobe

$\underline{X} = (X_1, \ldots, X_n)$ sl. uzorak iz $P(\lambda)$-distribuirane populacije

Budući da je $Y = X_1 + \ldots + X_n \sim P(n\lambda)$, MLE za λ je

$$\hat{\lambda} = \frac{Y}{n} = \overline{X}.$$

U slučaju malog n npr. 95% pouzdani interval dobijemo rješavanjem

$$F_Y(y|\lambda) \geq 0.025, \qquad 1 - F_Y(y-1|\lambda) \geq 0.025\,,$$

gdje je y opažena vrijednost od Y i

$$F_Y(y|\lambda) = \sum_{k=0}^{y} \frac{(n\lambda)^k}{k!} e^{-n\lambda}, \; y \in \{0, 1, 2, \ldots\}.$$

Može se pokazati da je

$$\lambda \mapsto F(y|\lambda) \quad \text{strogo padajuća na } (0, \infty)$$

pa su granice traženog pouzdanog intervala rješenja $\hat{\lambda}_1$ i $\hat{\lambda}_2$ jednadžbi

$$F(y|\hat{\lambda}_1) = 0.025, 1 - F(y-1|\hat{\lambda}_2) = 0.025$$

Za veliki n koristimo

$$\frac{\overline{X} - \lambda}{\sqrt{\lambda}}\sqrt{n} \approx N(0,1), \text{ tj.} \frac{\overline{X} - \lambda}{\sqrt{\hat{\lambda}}}\sqrt{n} \approx N(0,1)$$

za konstrukciju 95% pouzdanog intervala za λ

$$\hat{\lambda} \pm 1.96\sqrt{\frac{\hat{\lambda}}{n}}$$

Usporedba očekivanja normalnih populacija

$\overline{X_1}$ i $\overline{X_2}$ uzoračke sredine dvaju nezavisnih sl. uzoraka duljine n_1 i n_2 iz dviju normalnih populacija s poznatim varijancama σ_1^2 i σ_2^2.
Budući da su $\overline{X_1} \sim N(\mu_1, \frac{\sigma_1^2}{n_1^2})$ i $\overline{X_2} \sim N(\mu_2, \frac{\sigma_2^2}{n_2^2})$ nezavisne, slijedi da je

$$\overline{X_1} - \overline{X_2} \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1^2} + \frac{\sigma_2^2}{n_2^2})$$

pa je $(1-\alpha) \cdot 100\%$ pouzdani interval za $\mu_1 - \mu_2$ oblika

$$\overline{X_1} - \overline{X_2} \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

Ako su populacijske varijance nepoznate, ali ako pretpostavimo da su jednake:

$$\sigma_1^2 = \sigma_2^2 = \sigma^2\,,$$

onda je npr. 95% pouzdani interval za razliku očekivanja jednak

$$\overline{X_1} - \overline{X_2} \pm t_{0.025}(n_1 + n_2 - 2) \cdot S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\,,$$

gdje je

$$S_p^2 := \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

procjenitelj zajedničke varijance σ^2.

Usporedba varijanci normalnih populacija

Pivotna veličina: $\frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \sim F(n_1-1, n_2-2)$

$(1-\alpha)\cdot 100\%$ pouzdani interval za $\frac{\sigma_1^2}{\sigma_2^2}$ je

$$[\frac{S_1^2}{S_2^2}\cdot\frac{1}{f_{\alpha/2}(n_1-1,n_2-1)}, \frac{S_1^2}{S_2^2}f_{\alpha/2}(n_2-1,n_1-1)]$$

Spareni podaci

Sl. uzorak iz dvodimenzionalne razdiobe vektora (X, Y):

$$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n).$$

Analiziramo razlike

$$D_1 := X_1 - Y_1, D_2 := X_2 - Y_2, \ldots, D_n := X_n - Y_n$$

i procjenjujemo vrijednost $\mu_D := \mu_1 - \mu_2$.

Ako $\underline{D} = (D_1, \ldots, D_n)$ shvatimo kao sl. uzorak, onda koristimo

$$\frac{\overline{D} - \mu_D}{S_D}\sqrt{n} \sim t(n-1)$$

za konstrukciju 95%-pouzdanih intervala za μ_D:

$$\overline{D} \pm t_{0.025}(n-1)\frac{S_D}{\sqrt{n}}$$

Zadarak 8.2

Za reazlizaciju $x_1, x_2, \ldots, x_{16}$ slučajnog uzorka iz normalno distribuirane populacije vrijedi

$$\sum_{i=1}^{16} x_i = 15.2 \quad \text{i} \quad \sum_{i=1}^{16} x_i^2 = 243.19.$$

(a) Procijenite 95% pouzdani interval za populacijsku srednju vrijednost.

(b) Koliki bi uzorak trebali uzeti da uz 95% pouzdanosti populacijsku srednju vrijednost procijenimo s točnosti od $\varepsilon = 0.5$?

9. Testiranje statističkih hipoteza

- *statistička hipoteza* - pretpostavka o populacijskoj razdiobi - izjava o vrijednostima parametara
- *nulhipoteza* H_0 - aktualno znanje o vrijednostim parametara
  - jednostavna - populacijska razdioba jednoznačno određena
  - inače je složena
- *alternativna hipoteza*
- *testna statistika* - odluka u testu
- *statistički test* - pravilo raspodjele područja vrijednosti testne statistike na
  - područje konzistentno s H_0
  - područje nekonzistentno s H_0 - *kritično područje*

razina značajnosti testa α - vjerojatnost odbacivanja H_0, ako je H_0 istinita

	H_0 istinita	H_0 nije istinita
odbacili H_0	pogreška 1. vrste	✓
nismo odbacili H_0	✓	pogreška 2. vrste

$\beta =$ vjerojatnost pogreške 2. vrste

Primjer 9.1

$\underline{X}$ slučajni uzorak iz $N(\mu, \sigma^2)$-populacije s nepoznatim parametrima
Provodimo *jednostrani test*

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

uz razinu značajnosti 5%.

Testna statistika:

$$T = \frac{\overline{X} - \mu_0}{S} \overset{H_0}{\sim} t(n-1)$$

Kritično područje: $(-\infty, -t_{0.05}(n-1)]$

(H_0 odbacujemo u koristi H_1 ako opažena vrijednost $t = T(\underline{x})$ upadne u kritično područje).

Za *dvostrani test*

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

koristimo istu statistiku T i kritično područje

$$(-\infty, -t_{0.025}(n-1)] \cup [t_{0.025}(n-1), \infty).$$

p-vrijednost

- Koliko su jaki argumenti za odbacivanje (ne odbacijavnje) nul-hipoteze?
- *p-vrijednost* - vjerojatnost pogreške 1. vrste, ako je granica kritičnog područja opažena vrijednost statistike - najmanja značajnost uz koju bi H_0 bila odbačena u korist H_1 uz vrijednost opažene testne statistike

Primjer 9.2

Promatramo populaciju s razdiobom $X \sim B(200, \theta)$ uz opaženu vrijednost $x = 82$. Provodimo test

$$H_0 : \mu = 0.5$$
$$H_1 : \mu = 0.4$$

Testna statistika je X, a p-vrijednost je

$$\begin{aligned}\mathbb{P}(X \leq 82|H_0) &= \mathbb{P}(X < 82.5|H_0) \\ &= \mathbb{P}(\frac{X - 100}{\sqrt{50}} < \frac{82.5 - 100}{\sqrt{50}}) \\ &\approx \Phi(-2.475) = 0.0067\end{aligned}$$

- H_0 odbacujemo kad god je razina značajnosti barem 0.67%

9.3 Osnovni testovi bazirani na jednom uzorku

9.3.1 Testovi o parametru očekivanja

Zadan: sl. uzorak iz $N(\mu, \sigma^2)$-populacije
Testiramo nul-hipotezu:

$$H_0 : \mu = \mu_0$$

u odnosu na uobčajene alternative
(obje jednostrane i dvostrane)

Imamo dvije situacije:

1. σ je poznata. Tada je testna statistika

$$\frac{\overline{X} - \mu_0}{\sigma}\sqrt{n} \overset{H_0}{\sim} N(0,1).$$

2. σ je nepoznata. U tom slučaju je testna statistika

$$\frac{\overline{X} - \mu_0}{S}\sqrt{n} \overset{H_0}{\sim} t(n-1).$$

Za velike uzorke je

$$\frac{\overline{X} - \mu_0}{S}\sqrt{n} \overset{H_0}{\dot{\sim}} N(0,1).$$

9.3.2 Testovi o populacijskoj varijanci

Zadan: sl. uzorak iz $N(\mu, \sigma^2)$-populacije
Testiramo nul-hipotezu:

$$H_0 : \sigma^2 = \sigma_0^2.$$

Testna statistika je

$$\frac{(n-1)S^2}{\sigma_0^2} \overset{H_0}{\sim} \chi^2(n-1).$$

9.3.3 Testovi o populacijskoj proporciji

Zadan: sl. uzorak iz Bernoullijeve populacije $\text{bin}(1, \theta)$. Testiramo nul-hipotezu:

$$H_0 : \theta = \theta_0.$$

Testna statistika:

$$X = \text{frekvencija uspjeha u uzorku duljine } \ n$$

$$X \overset{H_0}{\sim} b(n, \theta_0).$$

Za veliko n koristi se normalna aproksimacija:

$$\frac{X - n\theta_0}{\sqrt{n\theta_0(1-\theta_0)}} \overset{H_0}{\approx} N(0,1).$$

9.3.4 Testovi o parametru Poissonove populacije

Zadan: sl. uzorak duljine n iz $P(\lambda)$-populacije
Testiramo hul-hipotezu

$$H_0 : \lambda = \lambda_0.$$

Testna statistika:

$$Y := X_1 + X_2 + \cdots + X_n \overset{H_0}{\sim} P(n\lambda_0).$$

Za veliko n koristi se normalna aproksimacija:

$$\frac{Y - n\lambda_0}{\sqrt{n\lambda_0}} \overset{H_0}{\dot{\sim}} N(0,1) \quad \text{ili} \quad \frac{\overline{X} - \lambda_0}{\sqrt{\lambda_0}}\sqrt{n} \overset{H_0}{\dot{\sim}} N(0,1).$$

9.4 Osnovni testovi bazirani na dva uzorka

9.4.1 Test o razlici populacijskih očekivanja

Zadano: 2 nezavisna uzorka duljina n_1 i n_2 iz $N(\mu_1, \sigma_1^2)$ i $N(\mu_2, \sigma_2^2)$-populacija.
Testiramo nul-hipotezu:

$$H_0 : \mu_1 - \mu_2 = \delta_0$$

(δ_0 je zadani broj)

Imamo sljedeće situacije:

1. σ_1^2 i σ_2^2 su poznati. Tada je testna statistika

$$Z = \frac{\overline{X}_1 - \overline{X}_2 - \delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \overset{H_0}{\sim} N(0,1).$$

2. σ_1^2 i σ_2^2 su nepoznati.
   Ako imamo velike uzorke,

$$Z = \frac{\overline{X}_1 - \overline{X}_2 - \delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \overset{H_0}{\sim} N(0,1);$$

Ako imamo male uzorke,
uz pretpostavku $\sigma_1^2 = \sigma_2^2 = \sigma^2$,
testna statistika je

$$T = \frac{\overline{X}_1 - \overline{X}_2 - \delta_0}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \overset{H_0}{\sim} t(n_1 + n_2 - 2),$$

gdje je

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}.$$

9.4.2 Test o kvocijentu populacijskih varijanci

Zadano: 2 nezavisna uzorka duljina n_1 i n_2 iz $N(\mu_1, \sigma_1^2)$ i $N(\mu_2, \sigma_2^2)$-populacija.
Testiramo nul-hipotezu:

$$H_0 : \sigma_1^2 = \sigma_2^2.$$

Testna statistika:

$$\frac{S_1^2}{S_2^2} \overset{H_0}{\sim} F(n_1 - 1, n_2 - 1).$$

9.4.3 Test razlike između popul. proporcija

Zadano: nezavisni uzorci velikih duljina n_1 i n_2 iz Bernoullijevih populacija.
Testiramo nul-hipotezu:

$$H_0 : \theta_1 = \theta_2.$$

Testna statistika:

$$\frac{\hat{\theta}_1 - \hat{\theta}_2}{\sqrt{\hat{\theta}(1-\hat{\theta})(\frac{1}{n_1}+\frac{1}{n_2})}} \overset{H_0}{\dot{\sim}} N(0,1),$$

$\hat{\theta}_1$ i $\hat{\theta}_2$ su relativne frekvencije uspjeha,
$\hat{\theta} = \frac{n_1\hat{\theta}_1 + n_2\hat{\theta}_2}{n_1+n_2}$ je procjena zajedničke proporcije

9.4.4 Test razlike između parametara Poissonovih razdioba

Zadano: nezavisni uzorci velikih duljina n_1 i n_2 iz $P(\lambda_1)$ i $P(\lambda_2)$ populacija.
Testiramo nul-hipotezu:

$$H_0 : \lambda_1 = \lambda_2.$$

Testna statistika:

$$\frac{\hat{\lambda}_1 - \hat{\lambda}_2}{\sqrt{\hat{\lambda}(\frac{1}{n_1} + \frac{1}{n_2})}} \overset{H_0}{\sim} N(0,1),$$

$\hat{\lambda}_1$ i $\hat{\lambda}_2$ su MLE,
$\hat{\lambda} = \frac{n_1\hat{\lambda}_1 + n_2\hat{\lambda}_2}{n_1+n_2}$ je procjena zajedničkog parametra

9.5 Osnovni test za sparene podatke

Zadan: sl. uzorak razlika sparenih vrijednosti iz normalne populacije (X_i, Y_i),

$$D_i = X_i - Y_i, \quad \mu_D = \mu_1 - \mu_2.$$

Testiramo nul-hipotezu:

$$H_0 : \mu_D = \delta_0.$$

Testna statistika:

$$T_D = \frac{\overline{D} - \delta_0}{S_D}\sqrt{n} \overset{H_0}{\sim} t(n-1).$$

Za veliki uzorak iz općenite ne-normalne popul.:

$$T_D \overset{H_0}{\dot{\sim}} N(0,1).$$

9.7 χ^2-testovi

- za kategorijalne i diskretne numeričke varijable
- usporedba frekvencija i očekivanih frekvencija (koje su u skladu s H_0)
- testna statistika

$$H = \sum_i \frac{(f_u - e_i)^2}{e_i} \overset{H_0}{\sim} \chi^2$$

9.7.1 Test prilagodbe modela podacima

- objašnjava li predloženi model za populacijsku razdiobu dobro poažene podatke
- nepoznati parametri se procjenjuju iz uzorka MLE metodom i ima ih r
- varijabla koju opažamo ima k razreda
  $\Longrightarrow$ testna statistika H uz H_0 ima $k - r - 1$ stupnjeva slobode, tj. $\chi^2(k - r - 1)$ razdiobu

Primjer 9.2

Je li igraća kocka fer?

$$H_0: \quad X = \text{broj na kocki} \sim \text{diskr. uniformna}$$
$$H_1: \quad \text{ne } H_0$$

Empirijski rezulatati $n = 300$ bacanja:

i	1	2	3	4	5	6
f_i	43	56	54	47	41	59

i	f_i	e_i	$\frac{(f_i-e_i)^2}{e_i}$
1	43	50	$49/50$
2	56	50	$36/50$
3	54	50	$16/50$
4	47	50	$9/50$
5	41	50	$81/50$
6	59	50	$81/50$
Σ	300	300	$272/50$

$h = 272/50 = 5.44.$

$H \overset{H_0}{\dot{\sim}} \chi^2(6-0-1) = \chi^2(5) \Rightarrow$

$$\text{pv} = \mathbb{P}(H \geq 5.44 | H_0) = 0.365.$$

$\Longrightarrow$ nema jakih argumenata za odbacivanje H_0

Zadatak 9.1

Podaci o štetama po 4000 polica osiguranja koje su bile pod rizikom točno godinu dana iz Zadatka 7.5 su prikazani frekvencijskom tablicom:

broj šteta i	frekvencija f_i
0	3288
1	642
2	66
≥ 3	4
ukupno	4000

Pretpostavimo da je broj šteta $X \sim P(\lambda)$ i MLE procjena parametra je bila $\hat{\lambda} = 0.196551$. Provedite χ^2-test prilagodbe Poissonovog modela navedenim

Kontingencijske tablice

(X, Y) diskretno numeričko obilježje

- testiraju se nul-hipoteze:
  - X i Y su nezavisne
  - da su populacijske razdiobe (npr. X) homogene obzirom na klasifikaciju po drugoj komponenti
- očekivane frekvencije se računaju po formuli:

$$\frac{\text{ukupan zbroj tog retka} \times \text{ukupan zbroj tog stupca}}{\text{veličina uzorka}}$$

- ako je u tablici r redaka i c stupaca, onda je broj stupnjeva slobode testne statistike:

$$rc - (r - 1 + c - 1) - 1 = (r - 1)(c - 1)$$

Primjer 9.3

Za svako od osiguravajućih društava A,B i C je uzet slučajni uzorak polica neživotnih osiguranja određenog tipa. Opažanjima je dobiveno da je u prošloj godini šteta bilo po 23% polica od A, 28% polica od B i 20% polica od C. Testirajte ima li značajnih razlika između tih proporcija ako su veličine uzoraka:

(a) $100, 100, 200$

(b) $300, 300, 600.$

Zadatak 9.2

Štete se mogu klasificirati na jednostavne, standardne i složene. Prošle je godine medu svim štetama bilo 18.4% jednostavnih, 70.3% standardnih i 11.3% složenih. U slučajnom uzorku od 120 ovogodišnih šteta opaženo je 15 jednostavnih, 87 standardnih i 18 složenih šteta. Pomoću χ^2-testa testirajte da li se raspodjela ovogodišnjih šteta značajno razlikuje od razdiobe prošlogodišnjih šteta.

Zadatak 9.3

U svrhu usporedbe iznosa premija osiguranja kućanstava koje naplaćuju dva osiguravajuća društva A i B, na slučajan način i nezavisno jedan od drugoga, odabrana su dva uzorka od po pet polica tog tipa iz svakog od navedenih društva. Opaženi iznosi premija su:
društvo A: 175 155 162 186 148
društvo B: 152 141 129 120 115
Pretpostavljamo da su iznosi premija normalno distribuirani s istim varijancama: redom $N(\mu_A, \sigma^2)$ i $N(\mu_B, \sigma^2)$.

(a) Procijenite zajedničku varijancu oba uzorka.

(b) Konstruirajte i izračunajte opaženi 95% pouzdani interval za razliku parametara očekivanja $\mu_A - \mu_B$.

(c) Kolika je p-vrijednost u jednostranom testu:

$$H_0 : \mu_A = \mu_B \qquad H_1 : \mu_A > \mu_B.$$

Je li (uz razinu značajnosti 5%) opažena uzoračka sredina iznosa premija osiguranja kućanstva A značajno veća od odgovarajuće uzoračke sredine za društvo B?

10. Korelacija i regresija

Mjerenja iz populacije (X, Y):

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

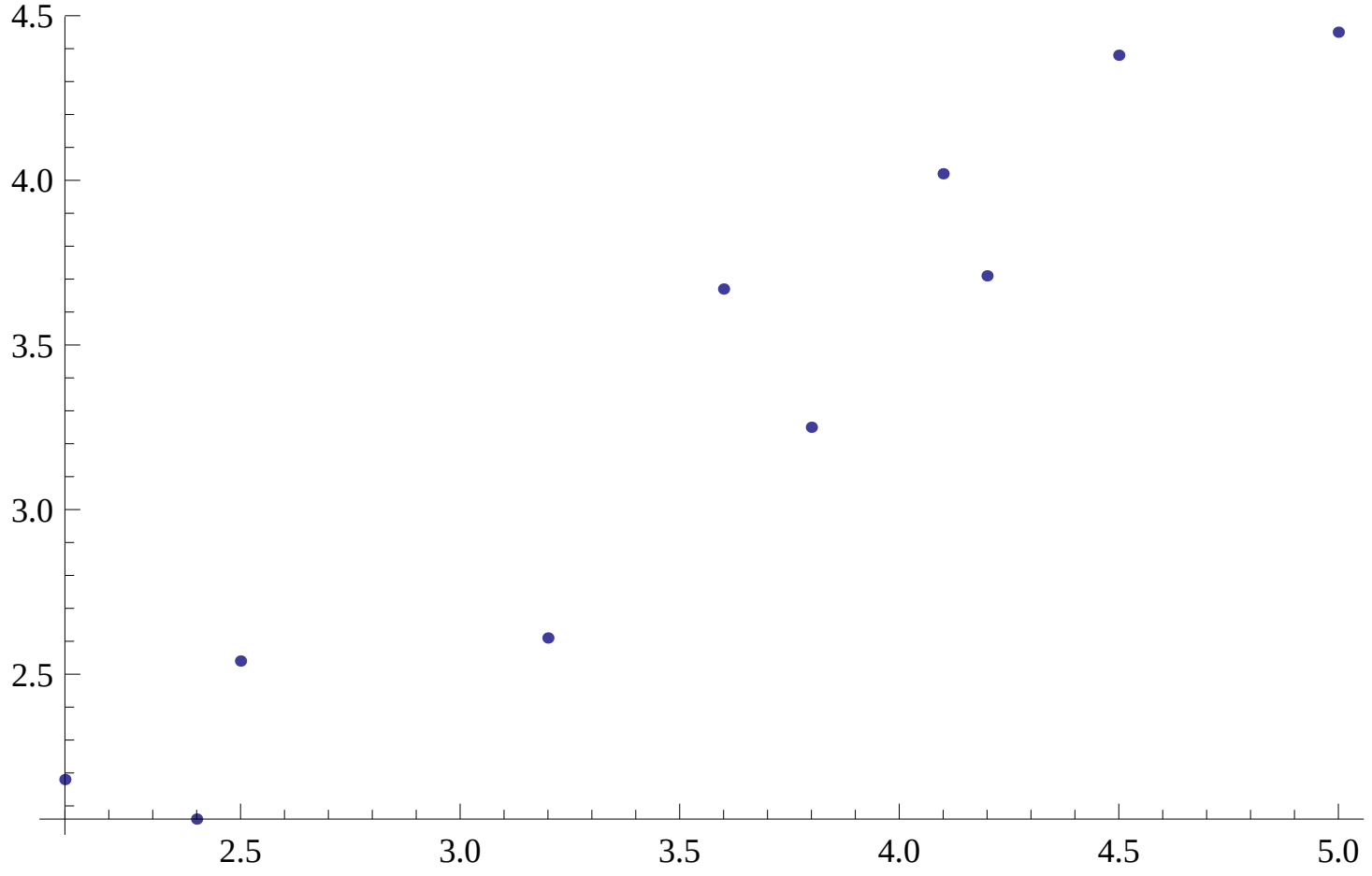
- korelacijska analiza: određivanje jakosti linearne povezanosti izmedju X i Y
- regresijska analiza: Y odziv (zavisna varijabla), X poticaj(nezavisna varijabla)

Primjer 10.1

Uzorak se sastoji od 10 podataka o iznosima zahtjeva za naknadu šteta i korespodentnih iznosa koje je osiguravajuće društvo stvarno platilo (u jedinicama od po 100 kn):

zahtjev	(x)	2.10	2.40	2.50	3.20	3.60
isplata	(y)	2.18	2.06	2.54	2.61	3.67
zahtjev	(x)	3.80	4.10	4.20	4.50	5.00
isplata	(y)	3.25	4.02	3.71	4.38	4.45

Dijagram raspršenja



Linearna zavisnost?

Koriste se statistike :

$$S_{XX} := \sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n} X_i^2 - n \cdot \overline{X}^2$$

$$S_{XY} := \sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = \sum_{i=1}^{n} X_i Y_i - n\overline{XY}$$

$$S_{YY} := \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n} Y_i^2 - n \cdot \overline{Y}^2.$$

Opažene vrijednosti: s_{xx}, s_{xy}, s_{yy}.

Primjer 10.1(nastavak)

i	x_i	y_i	x_i^2	$x_i y_i$	y_i^2
1	2.10	2.18	4.41	4.578	4.7524
2	2.40	2.06	5.76	4.944	4.2436
3	2.50	2.54	6.25	6.350	6.4516
4	3.20	2.61	10.24	8.352	6.8121
5	3.60	3.67	12.96	13.212	13.4689
6	3.80	3.25	14.44	12.350	10.5625
7	4.10	4.02	16.81	16.482	16.1604
8	4.20	3.71	17.64	15.582	13.7641
9	4.50	4.38	20.25	19.710	19.1844
10	5.00	4.45	25.00	22.250	19.8025
Σ	35.40	32.87	133.76	123.810	115.2025

Iz tablice ($n = 10$):

$$\bar{x} = \frac{35.40}{10} = 3.540, \quad \bar{y} = \frac{32.87}{10} = 3.287,$$

$$\sum_{i=1}^{10} x_i^2 = 133.76, \quad \sum_{i=1}^{10} x_i y_i = 123.810, \quad \sum_{i=1}^{10} y_i^2 = 115.2025,$$

$$\begin{aligned} S_{xx} &= \sum_{i=1}^{n} x_i^2 - n \cdot \bar{x}^2 = 133.76 - 10 \cdot 3.540^2 = 8.4440 \\ S_{xy} &= \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} = 123.810 - 10 \cdot 3.540 \cdot 3.287 = 7.4502 \\ S_{yy} &= \sum_{i=1}^{n} y_i^2 - n \cdot \bar{y}^2 = 115.2025 - 10 \cdot 3.287^2 = 7.1588. \end{aligned}$$

10.1 Korelacijska analiza

Pearsonov koeficijent korelacije:

$$r := \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}$$

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \frac{x_i - \bar{x}}{s_x} \cdot \frac{y_i - \bar{y}}{s_y},$$

$$-1 \leq r \leq 1$$

U Primjeru 10.1:

$$r = \frac{7.4502}{\sqrt{8.444 \cdot 7.1588}} = 0.958$$

$\rightarrow$ jaka linearna povezanost

10.1.2 Normalni model i inferencija

Zadan: sl. uzorak iz bivarijatnog normalnog modela

$$\underline{(X,Y)} = ((X_1,Y_1),(X_2,Y_2),\ldots,(X_n,Y_n))$$

$$R = \frac{S_{XY}}{\sqrt{S_{XX}\cdot S_{YY}}} \quad \text{(uzorački koeficijent korelacije).}$$

R je MLE za parametar ρ, populacijski koeficijent korelacije.

Test koreliranosti X i Y:

$$H_0 : \rho = 0$$

Testna statistika:

$$\frac{R}{\sqrt{1-R^2}}\sqrt{n-2} \overset{H_0}{\sim} t(n-2).$$

Vrijedi:

$$W := \frac{1}{2}\log\frac{1+R}{1-R} \approx N(\frac{1}{2}\log\frac{1+\rho}{1-\rho}, \frac{1}{n-3}) \text{ za veliko } n.$$

Testiramo nul-hipotezu:

$$H_0 : \rho = \rho_0$$

Testna statistika (i pivotna vel. za p.i. od ρ):

$$Z = \frac{\sqrt{n-3}}{2}(\ln\frac{1+R}{1-R} - \ln\frac{1+\rho_0}{1-\rho_0}) \overset{H_0}{\approx} N(0,1)$$

za veliko n.

Primjer 10.3

Na osnovi podataka iz Primjera 10.1, sprovedimo jednostrani test:

$$H_0 : \rho = 0.9, \quad H_1 : \rho > 0.9.$$

$r = 0.958$, $n = 10 \Rightarrow$
$z = (1.921 - 1.472)\sqrt{7} = 1.19$

$$\begin{aligned} \text{pv} &= \mathbb{P}(Z \geq 1.19 | H_0) \\ &= 1 - \Phi(1.19) \approx 1 - 0.8830 = 0.1170 \end{aligned}$$

$\Longrightarrow$ nema razloga za odbacivanje H_0

10.2 Regresijska analiza. Jednostavni linearni regresijski model

Podaci:

$$(x_1, Y_1), (x_2, Y_2), \ldots, (x_n, Y_n)$$

Jednostavni linearni regresijski model:

$$Y_i = \alpha + \beta x_i + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

Pretp. da su ispunjeni *Gauss-Markovljevi uvjeti* na pogreške:

$(A1)$ *centriranost*: $\mathbb{E}[\varepsilon_i] = 0$ za sve i;

$(A2)$ *jednakost varijanci*: $\mathrm{Var}[\varepsilon_i] = \sigma^2$ za sve i;

$(A3)$ *nekoreliranost*: $\mathrm{Cov}[\varepsilon_i, \varepsilon_j] = 0$ za sve $i \neq j$.

10.2.2 Prilagodba modela

Sastoji se od:

(a) procjene parametara α i β;

(b) procjene zajedničke varijance grešaka σ^2.

α i β se procjenjuju *metodom najmanjih kvadrata*:

$$q(\alpha, \beta) := \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2$$

$$q(\hat{\alpha}, \hat{\beta}) = \min_{\alpha, \beta} q(\alpha, \beta)$$

$$\hat{\beta} = \frac{S_{xY}}{S_{xx}}, \quad \hat{\alpha} = \overline{Y} - \hat{\beta}\bar{x}$$

Iz jednadžbi:

$$0 = \frac{\partial q}{\partial \alpha} = -2 \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))$$

$$0 = \frac{\partial q}{\partial \beta} = -2 \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i)) x_i$$

Vrijedi:

$$\mathbb{E}[\hat{\beta}] = \beta, \quad \mathrm{Var}[\hat{\beta}] = \sigma^2 \cdot \frac{1}{S_{xx}},$$

$$\mathbb{E}[\hat{\alpha}] = \alpha, \quad \mathrm{Var}[\hat{\alpha}] = \sigma^2 \cdot (\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}).$$

Procjenitelj za Y_i:

$$\hat{Y}_i := \hat{\alpha} + \hat{\beta} x_i$$

Nepristrani procjenitelj zajedničke varijance sl. grešaka:

$$\hat{\sigma}^2 := \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \frac{1}{n-2}q(\hat{\alpha}, \hat{\beta}).$$

$$\text{SSTOT} := \sum_{i=1}^{n} (Y_i - \overline{Y})^2 = S_{YY}$$

$$\text{SSE} := \sum_{i=1}^{n} (\underbrace{Y_i - \hat{Y}_i}_{\text{rezidual}})^2$$

$$\text{SSR} := \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2$$

$$Y_i - \overline{Y} = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \overline{Y}) \quad \Rightarrow$$

$$\text{SSTOT} = \text{SSR} + \text{SSE}.$$

Račun:

$$\begin{aligned} \mathrm{SSTOT} &= S_{yy} \\ \mathrm{SSR} &= \sum_{i=1}^{n}\Big((\hat{\alpha}+\hat{\beta}x_i)-(\hat{\alpha}+\hat{\beta}\bar{x})\Big)^2 = \hat{\beta}^2 S_{xx} \\ &= \frac{S_{xy}^2}{S_{xx}} \\ \Rightarrow \quad \mathrm{SSE} &= S_{yy} - \frac{S_{xy}^2}{S_{xx}}. \end{aligned}$$

Vrijedi:

$$\mathbb{E}[\mathrm{SSTOT}] = (n-1)\sigma^2+\beta^2 S_{xx}, \quad \mathbb{E}[\mathrm{SSR}] = \sigma^2+\beta^2 S_{xx},$$

$$\Rightarrow \quad \mathbb{E}[\mathrm{SSE}] = (n-2)\sigma^2.$$

Koeficijent determinacije:

$$R^2 := \frac{\mathrm{SSR}}{\mathrm{SSTOT}} \cdot 100\% = \frac{S_{xy}^2}{S_{xx} \cdot S_{yy}} \cdot 100\%$$

Podacima iz primjera 10.1 prilagodimo jednostavni linearni regresijski model.

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{7.4502}{8.4440} = 0.8823, \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 3.287 - 0.8823 \cdot 3.54 = 0.1636.$$

Procijenjeni pravac: $\hat{y} = 0.1636 + 0.8823x$

$$\text{SSTOT} = S_{yy} = 7.1588, \quad \text{SSR} = \frac{S_{xy}^2}{S_{xx}} = \frac{7.4502^2}{8.440} = 6.5734,$$

$$\Rightarrow \text{SSE} = \text{SSTOT} - \text{SSR} = 0.5854$$

$$\Rightarrow \hat{\sigma}^2 = \text{SSE}/8 = 0.0732$$

Koeficijent determinacije:

$$R^2 = \text{SSR}/\text{SSTOT} = 91.8\%$$

10.2.4 Potpuni normalni model i inferencija

Pretpostavimo da su još greške i:

$(A4)$ *nezavisne* i *normalno distribuirane*:
$\varepsilon_i \sim N(0, \sigma^2)$ za sve i.

Vrijedi:

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

$$T_\beta = \frac{\hat{\beta} - \beta}{\hat{\sigma}\sqrt{\frac{1}{S_{xx}}}} =\sim t(n-2)$$

Testiramo nul-hipotezu:

$$H_0 : \beta = 0$$

Testna statistika:

$$\frac{\hat{\beta}}{\hat{\sigma}\sqrt{\frac{1}{S_{xx}}}} \overset{H_0}{\sim} t(n-2)$$

Primjer 10.1(nastavak)

Na osnovi podataka iz Primjera 10.1,

(a) procijenimo 95%-pouzdan interval za koeficijent smjera regresijskog pravca β;

(b) testirajmo

$$H_0 : \beta = 1, \quad H_1 : \beta \neq 1.$$

95%-pouzdan interval za β:

$$\hat{\beta} \pm t_{0.025}(n-2) \cdot \hat{\sigma} \sqrt{\frac{1}{S_{xx}}}.$$

Opažena vrijednost tog intervala ($t_{0.025}(8) = 2.306$):

$$0.8823 \pm 2.306 \cdot \sqrt{\frac{0.0732}{8.4440}} = 0.8823 \pm 0.2147.$$

Budući da taj interval sadrži vrijednost "1",
nulhipotezu H_0 ne odbacujemo uz značajnost od 5%.

10.2.6 Procjena i predviđanje srednjeg i individualnog odziva

Očekivana vrijednost od Y uz dano $X = x_0$:

$$\mathbb{E}[Y|X = x_0] = (\text{kraće}) = \mathbb{E}[Y|x_0] = \alpha + \beta x_0$$

Procjenitelj:

$$\hat{\mathbb{E}}[Y|x_0] := \hat{\alpha} + \hat{\beta} x_0$$

$$\text{Var}[\hat{\mathbb{E}}[Y|x_0]] = \sigma^2 (\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}})$$

$$\frac{\hat{\mathbb{E}}[Y|x_0] - \mathbb{E}[Y|x_0]}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} = \frac{(\hat{\alpha} + \hat{\beta} x_0) - (\alpha + \beta x_0)}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \sim t(n-2)$$

$$\hat{Y}_0 := \hat{\alpha} + \hat{\beta} x_0$$

$$\begin{aligned} \mathrm{Var}[\hat{Y}_0 - Y_0] &= \mathrm{Var}[(\hat{\alpha} + \hat{\beta} x_0) - (\alpha + \beta x_0 + \varepsilon_0)] = \\ &= \sigma^2 (1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}) \end{aligned}$$

$$\frac{\hat{Y}_0 - Y_0}{\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \sim t(n-2)$$

Na osnovi podataka iz primjera

(a) procijenite 95%-pouzdan interval za očekivanu vrijednost isplata za zahtjeve s iznosom jednakim 460 kn;

(b) procijenite 95%-pouzdan interval za vrijednost isplate ako je iznos zahtjeva jednak 460 kn.

$$\hat{\alpha} + \hat{\beta} x_0 = 0.1636 + 0.88231 \cdot 4.6 = 4.222 = 422.20 \text{ kn},$$

$$\begin{aligned} \hat{\mathbb{E}}[Y|4.6]] \pm t_{0.025}(8) \cdot \hat{\sigma} \sqrt{\frac{1}{10} + \frac{(4.6 - \bar{x})^2}{S_{xx}}} &= 4.222 \pm 2.306 \cdot 0.1306 = \\ &= 4.222 \pm 0.301, \end{aligned}$$

$$\begin{aligned} \hat{Y}_0 \pm t_{0.025}(8) \cdot \hat{\sigma} \sqrt{1 + \frac{1}{10} + \frac{(4.6 - \bar{x})^2}{S_{xx}}} &= 4.222 \pm 2.306 \cdot 0.3004 = \\ &= 4.222 \pm 0.693. \end{aligned}$$

Zadatak 10.1

Za zadanih 12 vrijednosti varijable poticaja X izmjerene su pripadne vrijednosti $y_1, y_2, \ldots, y_{12}$ varijable odziva. Na taj način je dobiven uzorak $(x_i, y_i), i = 1, 2, \ldots, 12$ za koji vrijedi

$$\sum_{i=1}^{12} x_i = 516.4 \qquad \sum_{i=1}^{12} x_i^2 = 22741.34 \qquad \sum_{i=1}^{12} y_i = 14821$$

$$\sum_{i=1}^{12} y_i^2 = 18695125 \qquad \sum_{i=1}^{12} x_i y_i = 650264.8.$$

(a) Uz pretpostavku da je model regresijski, procijenite pravac regresije.

(b) Konstruirajte i procijenite 95% pouzdani interval za koeficijent smjera regresijskog pravca.

(c) Testirajte hipotezu da je koeficijent smjera jednak 0 (uz alternativu da nije tako).

(d) Konstruirajte i procijenite 95% pouzdani interval za srednju vrijednost varijable Y ako je $X = 50$.

10.2.8 Transformirani podaci

Modeli rasta:

$$\mathbb{E}[Y|x] = \alpha e^{\beta x}$$

$W = \log Y \Rightarrow$

$$W_i = \eta + \beta x_i + \varepsilon_i, \quad i = 1, 2, \ldots, n,$$

$\eta = \log \alpha$

$X_1, X_2, \ldots, X_k$ – varijable poticaja
Y – varijabla odziva

$$\begin{aligned} & \mathbb{E}[Y|X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n] = \\ = \quad & \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \end{aligned}$$

Višestruki linearni regresijski model:

$$Y_i = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

11. Analiza varijance
10.1 Jednofaktorska ANOVA

- usporedba djelovanja tretmana na razdiobu varijable Y

Model:

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}, \quad j = 1, 2, \ldots, n_i, \quad i = 1, 2, \ldots, k,$$

Pretp. $\varepsilon_{ij} \sim N(0, \sigma^2)$ nezavisne

Parametri modela: μ, τ_i, $i = 1, , 2, \ldots, k$, σ^2

$$\mu = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \mathbb{E}[Y_{ij}]$$

(ukupna populacijska sredina)

model $\Longrightarrow$

$$\sum_{i=1}^{k} n_i \tau_i = 0$$

11.1.2 Procjena parametara

– metodom najmanjih kvadrata:

$$q(\mu, \tau_1, \ldots, \tau_k) := \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \mu - \tau_i)^2 \to \min$$

(uz uvjet $\sum_{i=1}^{k} n_i \tau_i = 0$)

$$0 = \frac{\partial q}{\partial \mu} = -2 \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \mu - \tau_i)$$

$$= -2 \left(\sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij} - n\mu \right)$$

$$0 = \frac{\partial q}{\partial \tau_i} = -2 \sum_{j=1}^{n_i} (y_{ij} - \mu - \tau_i)$$

$$\hat{\mu} = \overline{Y}_{..}, \quad \hat{\tau}_i = \overline{Y}_{i.} - \overline{Y}_{..}, \quad i = 1, 2, \ldots, k,$$

$$\overline{Y}_{i.} := \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} \quad (\textit{uzoračka sredina za } i\textit{-ti tretman}), \quad i = 1, 2, \ldots, k$$

$$\overline{Y}_{..} := \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij} = \frac{1}{n} \sum_{i=1}^{k} n_i \overline{Y}_{i.} \quad (\textit{sveukupna uzoračka sredina}).$$

Vrijedi:

$$\sum_{i=1}^{k} n_i \hat{\tau}_i = 0.$$

Za

$$S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i.})^2, \quad i = 1, 2, \ldots, k$$

vrijedi: $(n_i - 1)S_i^2/\sigma^2 \sim \chi^2(n_i - 1)$ i nezavisne su

$$\frac{1}{\sigma^2} \sum_{i=1}^{k} (n_i - 1)S_i^2 \sim \chi^2(n - k).$$

Zajednička uzoračka varijanca:

$$\hat{\sigma}^2 := \frac{1}{n-k} \sum_{i=1}^{k} (n_i - 1)S_i^2 = \frac{1}{n-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i.})^2$$

je nepristrani procjenitelj za σ^2.

11.1.3 Rastav varijance

$$\text{SSTOT} := \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{..})^2 \quad \text{(ukupna suma kvadrata)}$$

$$\text{SST} := \sum_{i=1}^{k} n_i(\overline{Y}_{i.} - \overline{Y}_{..})^2 \quad \text{(suma kvadrata zbog razlike u tretmanima)}$$

$$\text{SSE} := \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i.})^2. \quad \text{(suma kvadrata pogrešaka(reziduala))}$$

Vrijedi:

$$\text{SSTOT} = \text{SSE} + \text{SST}$$

Test:

$$H_0 : \tau_i = 0 \ \text{ za svaki } \ i = 1, 2, \ldots, k,$$

$$H_1 : \tau_i \neq 0 \ \text{ za barem jedan } \ i \ \text{ od } \ 1, 2, \ldots, k$$

Testna statistika:

$$F = \frac{\mathrm{MST}}{\mathrm{MSE}} \overset{H_0}{\sim} F(k-1, n-k)$$

gdje su

$$\mathrm{MST} := \frac{\mathrm{SST}}{k-1} (\text{srednjekvadratno odstupanje zbog tretmana})$$

$$\mathrm{MSE} := \frac{\mathrm{SSE}}{n-k} (\text{srednjekvadratna greška})$$

ANOVA tablica:

izvor var.	stup. slob.	sume kv.	srednji kv.	test-stat.
zbog tretmana	$k-1$	SST	MST	f
sl. greške	$n-k$	SSE	MSE	—
ukupno	$n-1$	SSTOT	—	—

Primjer 11.1

Iz svakog od tri osiguravajućeg društva A, B i C na slučajan način uzet je po uzorak polica osiguranja privatnih kuća. Zabilježene su osigurane svote po svakoj polici (u iznosima od po 100 kn):

društvo A: 36, 28, 32, 43, 30, 21, 33, 37, 26, 34
društvo B: 26, 21, 31, 29, 27, 35, 23, 33
društvo C: 39, 28, 45, 37, 21, 49, 34, 38, 44.

Želimo testirati nulhipotezu da su populacijske srednje vrijednosti osiguranih svota po policama osiguranja privatnih kuća jednake, odnosno, da izbor osiguravajućeg društva ne utječe na očekivani iznos osigurane svote po tim policama.

$$n_A = 10,\ n_B = 8,\ n_c = 9,$$
$$n = n_A + n_B + n_C = 10 + 8 + 9 = 27$$

$$\bar{y}_{A.} = 32.0000,\ \ \bar{y}_{B.} = 28.1250,\ \ \bar{y}_{C.} = 37.2222,$$
$$s_A^2 = 38.2222,\ \ \ s_B^2 = 23.2679,\ \ \ s_C^2 = 75.9444.$$

$$\begin{aligned} \bar{y}_{..} &= \frac{n_A\bar{y}_{A.} + n_B\bar{y}_{B.} + n_C\bar{y}_{C.}}{n} = \\ &= \frac{10 \cdot 32.0000 + 8 \cdot 28.1250 + 9 \cdot 37.2222}{27} = \\ &= 32.5926. \end{aligned}$$

$$
\begin{aligned}
\text{SST} &= n_A(\bar{y}_{A.} - \bar{y}_{..})^2 + n_B(\bar{y}_{B.} - \bar{y}_{..})^2 + n_C(\bar{y}_{C.} - \bar{y}_{..})^2 = \\
&= 10 \cdot (32. - 32.5926)^2 + 8 \cdot (28.125 - 32.5926)^2 + \\
&\quad +9 \cdot (37.2222 - 32.5926)^2 = \\
&= 356.088 \\
\text{MST} &= \frac{\text{SST}}{k-1} = \frac{356.088}{3-1} = 178.044 \\
\text{SSE} &= (n_A - 1)s_A^2 + (n_B - 1)s_B^2 + (n_C - 1)s_C^2 = \\
&= 9 \cdot 38.2222 + 7 \cdot 23.2679 + 8 \cdot 75.9444 = \\
&= 1114.43 \\
\text{MSE} &= \frac{\text{SSE}}{n-k} = \frac{1114.43}{27-3} = 46.4346 \\
f &= \frac{\text{MST}}{\text{MSE}} = 3.8343
\end{aligned}
$$

ANOVA tablica:

izvor var.	st. slob.	sume kv.	sr. kv.	test-stat.
zbog o. d.	2	356.09	178.044	3.83
sl. greške	24	1114.43	46.435	—
ukupno	26	1470.52	—	—

$$H_0 : \tau_A = \tau_B = \tau_C = 0$$

$F \overset{H_0}{\sim} F(2, 24)$ i $f = 3.83 \Rightarrow$
pv= $\mathbb{P}(F \geq 3.83 | H_0) = 0.042 \implies$ možemo odbaciti H_0 uz razinu značajnosti 5%

Zadatak 11.1

27 zaposlenika jednog poduzeća podijeljeno je u tri jednake grupe. Jedna grupa je pohađala tečaj A, druga tečaj B, a treća je kontrolna skupina (nije pohađala nikakav tečaj). Oba tečaja su istog tipa i nakon završenog tečaja zaposlenici su pisali test. Rezultati su sljedeći:

kontrola:	55 74 64 62 37 78 50 44
tečaj A:	63 79 60 75 89 58 75 72 84 69
tečaj B:	64 55 57 73 51 60 62 78 68.

Sprovedite test nulhipoteze da nema razlike u distribuciji rezultata testa izmedu tri navedena skupine

Analiza sredina tretmana

Zanima li nas pouzdani interval za očekivanje $\mu + \tau_i$ i-tog tretmana, onda koristimo

$$\frac{\overline{Y}_{i.} - (\mu + \tau_i)}{\hat{\sigma}} \sqrt{n_i} \sim t(n-k).$$

pa je npr. 95% pouzdani interval za $\mu + \tau_i$ jednak

$$\overline{Y}_{i.} \pm t_{0.025}(n-k) \frac{\hat{\sigma}}{\sqrt{n_i}}$$