Statistika

primjeri i zadaci

Ante Mimica, Marina Ninčević

16. listopada 2009.

Sadržaj

1

Opisna statistika

Zadatak 1.1 Kocku smo bacali 20 puta i zabilježili smo sljedeće rezultate:

$$\begin{array}{cccccccccc} 6 & 3 & 3 & 6 & 3 & 5 & 6 & 1 & 4 & 6 \\ 3 & 5 & 5 & 2 & 2 & 2 & 2 & 3 & 2 & 3 \end{array}$$

(a) Nacrtajte histogram relativnih frekvencija.

(b) Odredite aritmetičku sredinu, mod i medijan uzorka.

(c) Odredite varijancu i standardnu devijaciju uzorka.

(d) Odredite raspon uzorka.

(e) Odredite donji i gornji kvartil te interkvartil uzorka.

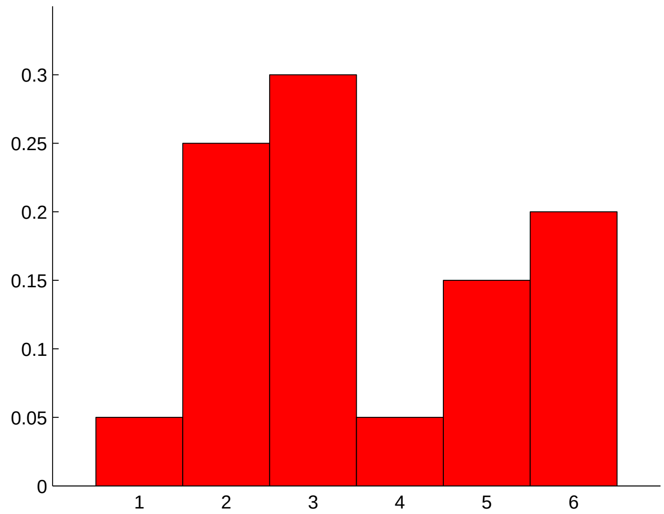
(f) Nacrtajte dijagram pravokutnika ("box and whisker plot").

Rješenje:

(a) Frekvencijska tablica:

i	y_i	f_i	$r_i = f_i/20$
1	1	1	0.05
2	2	5	0.25
3	3	6	0.3
4	4	1	0.05
5	5	3	0.15
6	6	4	0.2
Σ		20	1.00

Histogram relativnih frekvencija:



(b) **Aritmetička sredina** iznosi

$$\overline{x} = \frac{f_1 y_1 + \cdots + f_6 y_6}{f_1 + \cdots + f_6} = \frac{1 \cdot 1 + 5 \cdot 2 + 6 \cdot 3 + 1 \cdot 4 + 3 \cdot 5 + 4 \cdot 6}{1 + 5 + 6 + 1 + 3 + 4} = 3.6.$$

Mod uzorka je 3.

Podatke $x_1, \ldots, x_n$ poredane po veličini označavamo s $x_{(1)}, \ldots, x_{(n)}$. Pri tome za $s = k + r$, gdje je $k \in \mathbb{Z}, r \in [0, 1\rangle$, vrijedi formula

$$x_{(s)} = x_{(k)} + r \left[x_{(k+1)} - x_{(k)} \right]. \tag{1.1}$$

Uredimo podatke po veličini: $1, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6$.

Medijan uzorka je

$$m = x_{\left(\frac{n+1}{2}\right)} = x_{\left(\frac{20+1}{2}\right)} = x_{\left(10+\frac{1}{2}\right)} = x_{(10)} + \frac{1}{2} \left[x_{(11)} - x_{(10)} \right] = 3 + \frac{1}{2}(3 - 3) = 3.$$

(c) **Varijanca** uzorka iznosi

$$s^2 = \frac{1 \cdot 1^2 + 5 \cdot 2^2 + 6 \cdot 3^2 + 1 \cdot 4^2 + 3 \cdot 5^2 + 4 \cdot 6^2 - 20 \cdot 3.6^2}{1 + 5 + 6 + 1 + 3 + 4 - 1} = 2.673684$$

pa je standardna devijacija $s = 1.64$.

(d) **Raspon** uzorka je $R = x_{(20)} - x_{(1)} = 6 - 1 = 5$.

(e) **Donji kvartil** je

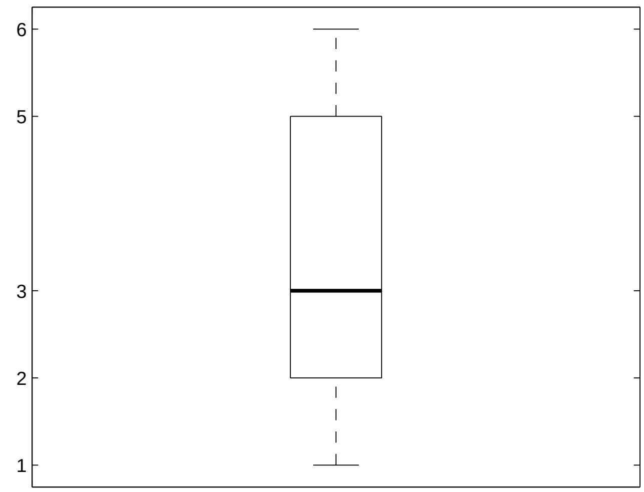
$$q_L = x_{\left(\frac{n+1}{4}\right)} = x_{\left(\frac{20+1}{4}\right)} = x_{\left(5+\frac{1}{4}\right)} = x_{(5)} + \frac{1}{4} \left[x_{(6)} - x_{(5)} \right] = 2 + \frac{1}{4}(2 - 2) = 2.$$

Gornji kvartil je

$$q_U = x_{\left(\frac{3(n+1)}{4}\right)} = x_{\left(\frac{3(20+1)}{4}\right)} = x_{\left(15+\frac{3}{4}\right)} = x_{(15)} + \frac{3}{4}\left[x_{(16)} - x_{(15)}\right] = 5 + \frac{3}{4}(5-5) = 5.$$

Interkvartil iznosi $d_q = q_U - q_L = 5 - 2 = 3$.

(f) Dijagram pravokutnika:



Riješimo sada ovaj zadatak u R-u.

Ubacimo podatke u vektor `rezultati`:

```
> rezultati<-c(6,3,3,6,3,5,6,1,4,6,3,5,5,2,2,2,2,3,2,3)
```

Ispišimo podatke

```
> rezultati
[1] 6 3 3 6 3 5 6 1 4 6 3 5 5 2 2 2 2 3 2 3
```

Duljina uzorka je

```
> n<-length(rezultati)
> n
[1] 20
```

Tablica frekvencija je

```
> frekv<-data.frame(table(rezultati))
> frekv
  rezultati Freq
1         1    1
2         2    5
3         3    6
4         4    1
5         5    3
6         6    4
```

Promijenimo imena stupaca i dodajmo stupac relativnih frekvencija:

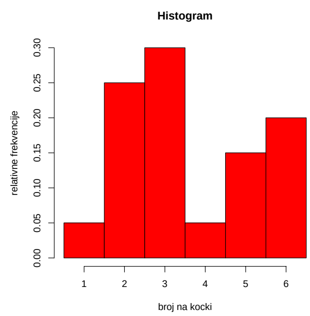
```
> frekv<-data.frame(frekv,frekv[2]/sum(frekv[2]))
> frekv
  rezultati frekvencije frekvencije.1
1         1           1          0.05
2         2           5          0.25
3         3           6          0.30
4         4           1          0.05
5         5           3          0.15
6         6           4          0.20
> names(frekv)[3]<-"rel. frekvencije"
> frekv
  rezultati frekvencije rel. frekvencije
1         1           1             0.05
2         2           5             0.25
3         3           6             0.30
4         4           1             0.05
5         5           3             0.15
6         6           4             0.20
```

Histogram dobijemo na sljedeći način

```
> hist(rezultati)
```

Ako naredbu `hist` upotrijebimo bez dodatnih argumenata, onda se razredi određuju automatski i možda nećemo dobiti željeni histogram, kao što vidimo na gornjem histogramu.

```
> hist(rezultati,probability=TRUE,breaks=c(0.5,1.5,2.5,3.5,4.5,5.5,6.5),
xlab="broj na kocki",ylab="relativne frekvencije",main="Histogram",col="red")
```



Dodatnim argumentom `prob=TRUE` dobijemo površinu histograma jednaku 1. Koristeći

```
breaks=c(0.5,1.5,2.5,3.5,4.5,5.5,6.5)
```

određujemo granice razreda i u ovom slučaju razredu odgovara točno jedna kvalitativna vrijednost te su na y-osi zaista relativne frekvencije.

Možemo izračunati:

- aritmetičku sredinu

```
 > mean(rezultati)
[1] 3.6
```

- uzoračku varijancu

```
> var(rezultati)
[1] 2.673684
```

- mod

  Prvo trebamo definirati funkciju

```
 > statmod <- function(x) {
+    z <- table(as.vector(x))
+    names(z)[z == max(z)]
+ }
```

  i tada je mod

```
 > statmod(rezultati)
[1] "3"
```

- raspon uzorka

```
 > max(rezultati)-min(rezultati)
[1] 5
```

- medijan

```
 > median(rezultati)
[1] 3
```

  Medijan možemo dobiti i preko funkcije quantile:

```
 > quantile(rezultati,0.5,type=6)
50%
  3
```

- donji i gornji kvartil

```
 > quantile(rezultati,0.25,type=6)
25%
  2
> quantile(rezultati,0.75,type=6)
75%
  5
```

  ili sve zajedno

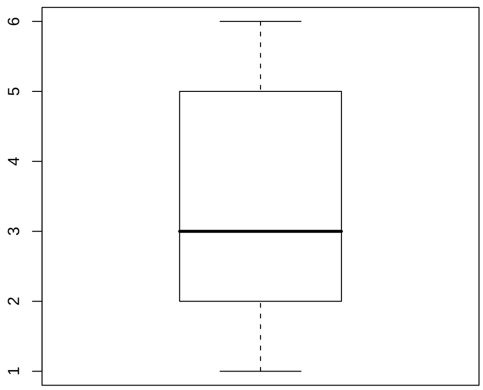
```
 > quantile(rezultati,type=6)
  0%  25%  50%  75% 100%
   1    2    3    5    6
```

- interkvartil

```
 > IQR(rezultati)
[1] 3
```

- dijagram pravokutnika

```
 > boxplot(rezultati)
```



△

Napomena: $\left(x_{(1)}, q_L, m, q_U, x_{(n)}\right)$ se zove **karakteristična petorka uzorka**.

Napomena: Pri formiranju dijagrama pravokutnika **outlieri** su sve vrijednosti koje su od donjeg ili gornjeg kvartila udaljene za više od $\frac{3}{2}\, d_q$. "**Brkovi**" su najmanja i najveća vrijednost koje nisu outlieri. Outlieri se posebno naznačavaju na dijagramu pravokutnika.

Primijetite da u prethodnom zadatku nema outliera.

Zadatak 1.2 Na nekom fakultetu je odabran uzorak od 40 studenata i izmjerene su im visine:

140	188	175	176	177	168	162	181
183	187	187	162	184	161	180	169
195	171	170	199	181	169	189	191
172	182	183	178	180	165	185	205
183	187	188	182	163	179	178	188

(a) Odredite karakterističnu petorku uzorka.

(b) Nacrtajte dijagram pravokutnika.

Rješenje:

(a) Uredimo podatke:

140	161	162	162	163	165	168	169
169	170	171	172	175	176	177	178
178	179	180	181	181	182	182	183
183	183	184	185	187	187	187	188
188	188	188	189	191	195	199	205

Tada je

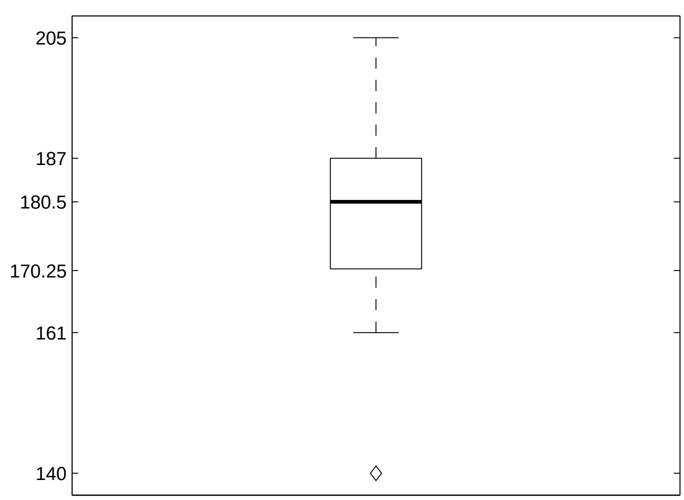
$$
\begin{aligned}
&n = 40,\\
&x_{(1)} = 140,\\
&q_L = x_{\left(\frac{40+1}{4}\right)} = x_{\left(10+\frac{1}{4}\right)} = x_{(10)} + \frac{1}{4}\left[x_{(11)} - x_{(10)}\right] = 170 + \frac{1}{4}(171 - 170) = 170.25,\\
&m = x_{\left(\frac{40+1}{2}\right)} = x_{\left(20+\frac{1}{2}\right)} = x_{(20)} + \frac{1}{2}\left[x_{(21)} - x_{(20)}\right] = 180 + \frac{1}{2}(181 - 180) = 180.5,\\
&q_U = x_{\left(\frac{3(40+1)}{4}\right)} = x_{\left(30+\frac{3}{4}\right)} = x_{(30)} + \frac{3}{4}\left[x_{(31)} - x_{(30)}\right] = 187 + \frac{3}{4}(187 - 187) = 187,\\
&x_{(40)} = 205.
\end{aligned}
$$

Karakteristična petorka je $(140, 170.25, 180.5, 187, 205)$.

(b) Interkvartil iznosi $d_q = q_U - q_L = 187 - 170.25 = 16.75$ pa je

$$q_L - \frac{3}{2}\, d_q = 145.125 \quad \text{i} \quad q_U + \frac{3}{2}\, d_q = 212.125.$$

Dijagram pravokutnika:



Riješimo zadatak u R-u. Na `www.math.hr/nastava/stat` pronađite datoteku `visine.dat` i snimite je u radni direktorij.

Učitajmo podatke iz datoteke:

```
> visine<-read.table("visine.dat",header=FALSE)
> visine
```

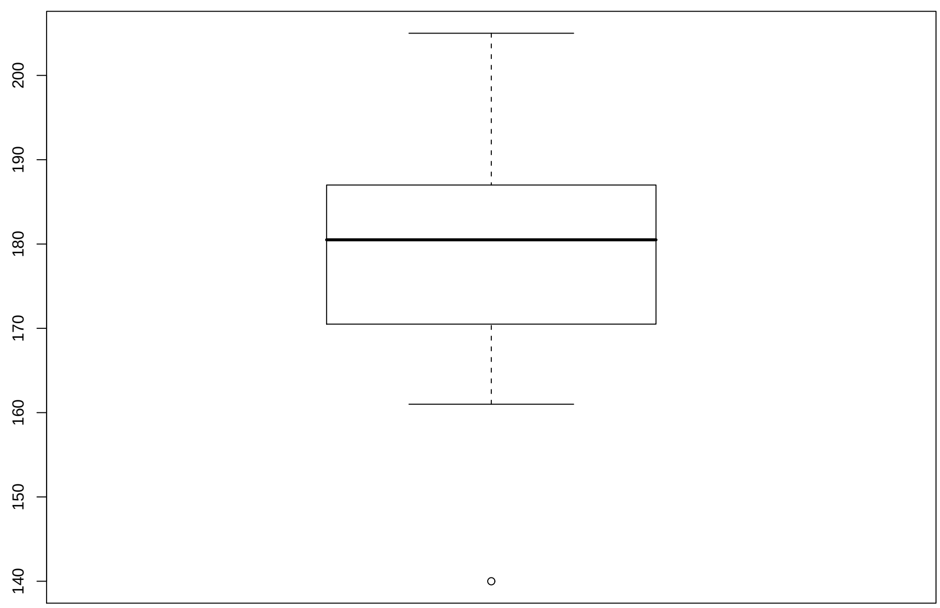
```
    V1
1  140
2  188
3  175
4  176
5  177
6  168
7  162
8  181
9  183
10 187
11 187
12 162
13 184
14 161
15 180
16 169
17 195
18 171
19 170
20 199
21 181
22 169
23 189
24 191
25 172
26 182
27 183
28 178
29 180
30 165
31 185
32 205
33 183
34 187
35 188
36 182
37 163
38 179
39 178
40 188
```

Naredbom `quantile` dobijemo karakterističnu petorku uzorka:

```
 > quantile(visine$V1,type=6)
    0%    25%    50%    75%   100%
140.00 170.25 180.50 187.00 205.00
```

Dijagram pravokutnika:

```
> boxplot(visine$V1)
```



△

Zadatak 1.3 U 15 država izmjerena je prosječna konzumacija alkohola jednog stanovnika tijekom jedne godine. Nacrtajte histogram dobivenih sljedećih podataka u litrama:

4.7	(Meksiko)	8.4	(SAD)	11.8	(Španjolska)	12.0	(Njemačka)
11.9	(Mađarska)	15.8	(Luksemburg)	7.9	(Japan)	9.0	(Australija)
11.9	(Danska)	13.0	(Češka)	13.9	(Irska)	8.0	(Italija)
11.6	(Francuska)	11.8	(V. Britanija)	8.3	(Kanada)		

Rješenje: Sada imamo neprekidno statističko obilježje, tj. podaci dolaze iz $[0, +\infty\rangle$.

- Odredimo **broj razreda**. Uzmimo na primjer $k = 6$.

- Odredimo **širinu razreda**. Vrijedi $\dfrac{x_{(15)} - x_{(1)}}{k} = \dfrac{15.8 - 4.7}{6} = 1.85.$

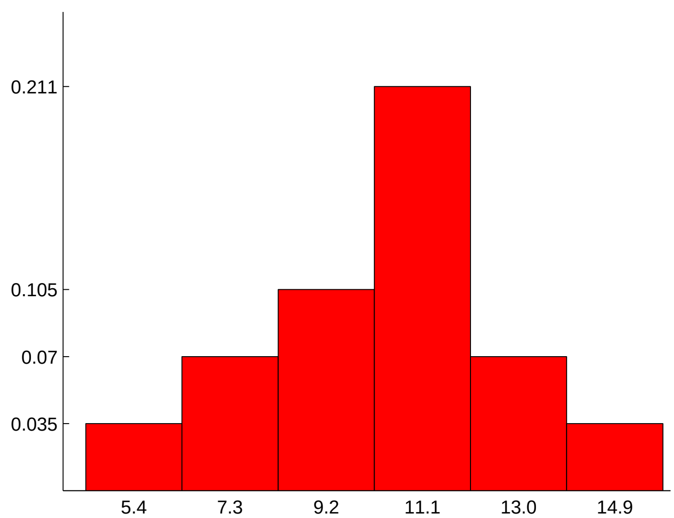
  Ako dobijemo broj koji ima veći broj decimalnih mjesta od broja decimalnih mjesta na koji su zaokruženi podaci, onda za širinu razreda c uzimamo broj koji je za jednu značajnu decimalu veći. U ovom slučaju je $c = 1.9$.

- Odredimo **granice razreda**. One moraju biti određene za jedno decimalno mjesto više (zbog jednoznačnog svrstavanja podataka u razrede).

Frekvencijska tablica:

i	I_i	f_i	r_i	r_i/c	$\overline{x}_i$
1	$[4.45, 6.35\rangle$	1	1/15	0.035	5.4
2	$[6.35, 8.25\rangle$	2	2/15	0.07	7.3
3	$[8.25, 10.15\rangle$	3	3/15	0.105	9.2
4	$[10.15, 12.05\rangle$	6	6/15	0.211	11.1
5	$[12.05, 13.95\rangle$	2	2/15	0.07	13.0
6	$[13.95, 15.85\rangle$	1	1/15	0.035	14.9
Σ		15	1		

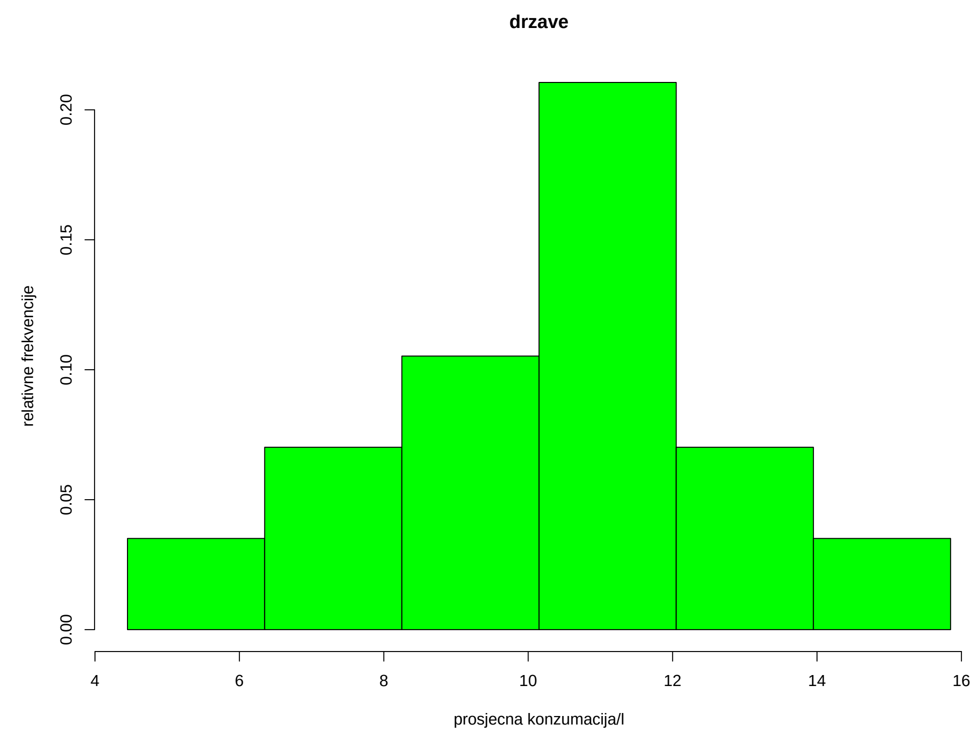
Histogram:



U R-u:

```
> drzave<-c(4.7,8.4,11.8,12.0,11.9,15.8,7.9,9.0,11.9,13.0,13.9,8.0,11.6,11.8,8.3)
> drzave
 [1]  4.7  8.4 11.8 12.0 11.9 15.8  7.9  9.0 11.9 13.0 13.9  8.0 11.6 11.8  8.3
> k<-6
```

```
> c<-round((max(drzave)-min(drzave))/k,1)
> c
[1] 1.9
> hist(drzave,probability=TRUE,breaks=seq(4.45,by=c,length=k+1),
+xlab="prosjecna konzumacija/l",ylab="relativne frekvencije",
+main="drzave",col="green")
```



△

Zadatak 1.4 Mjereno je vrijeme života litijske baterija koja se upotrebljava u kalkulatorima. Dobiveni su sljedeći podaci u satima:

4285	564	1278	205	3920
2066	604	209	602	1379
2584	14	349	3770	99
1009	4152	478	726	510
318	737	3032	3894	582
1429	852	1461	2662	308
981	1560	701	497	3367
1402	1786	1406	35	99
1137	520	261	2778	373
414	396	83	1379	454

(a) Odredite ”stem and leaf” dijagram.

(b) Nacrtajte histogram podataka.

Rješenje:

(a) ”Stem and leaf” dijagram:

$$\begin{array}{r|l} 0 & 39456078532347202674005530 34 \\ 1 & 04415724433 \\ 2 & 0567 \\ 3 & 07893 \\ 4 & 21 \end{array}$$

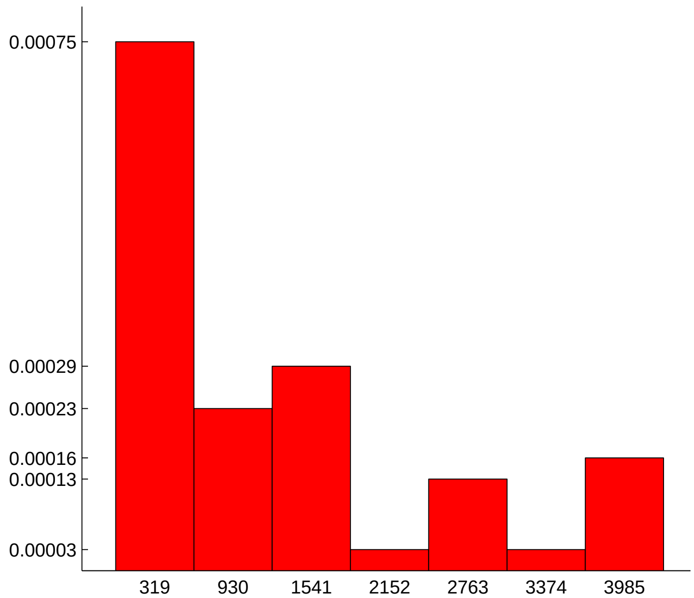
(b) Vrijedi $n = 50, \quad x_{(1)} = 14, \quad x_{(50)} = 4285.$

Za $k = 7 \quad \Rightarrow \quad \dfrac{x_{(50)} - x_{(1)}}{k} = \dfrac{4285 - 14}{7} = 610.14 \quad \Rightarrow \quad c = 611.$

Frekvencijska tablica:

i	I_i	f_i	r_i	r_i/c	$\overline{x}_i$
1	$[13.5, 624.5\rangle$	23	0.46	0.00075	319
2	$[624.5, 1235.5\rangle$	7	0.14	0.00023	930
3	$[1235.5, 1846.5\rangle$	9	0.18	0.00029	1541
4	$[1846.5, 2457.5\rangle$	1	0.02	0.00003	2152
5	$[2457.5, 3068.5\rangle$	4	0.08	0.00013	2763
6	$[3068.5, 3679.5\rangle$	1	0.02	0.00003	3374
7	$[3679.5, 4290.5\rangle$	5	0.10	0.00016	3985
Σ		50	1.00		

Histogram:



Riješimo zadataku R-u. Na `www.math.hr/nastava/stat` pronađite datoteku `baterije.dat` i snimite je u radni direktorij.

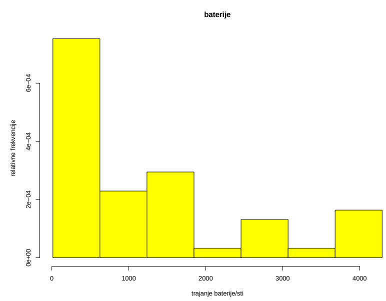
```
> baterije<-read.table("baterije.dat",header=FALSE)
> baterije
     V1
1  4285
2  2066
3  2584
4  1009
5   318
6  1429
7   981
8  1402
9  1137
10  414
11  564
12  604
13   14
14 4152
15  737
16  852
17 1560
```

```
18 1786
19  520
20  396
21 1278
22  209
23  349
24  478
25 3032
26 1461
27  701
28 1406
29  261
30   83
31  205
32  602
33 3770
34  726
35 3894
36 2662
37  497
38   35
39 2778
40 1379
41 3920
42 1379
43   99
44  510
45  582
46  308
47 3367
48   99
49  373
50  454
> n<-length(baterije$V1)
> n
[1] 50
> stem(baterije$V1,scale=0.5)

  The decimal point is 3 digit(s) to the right of the |

  0 | 00111223333444555556666667779
  1 | 001344444568
  2 | 1678
  3 | 04899
```

```
  4 | 23
> k<-7
> c=ceiling((max(baterije$V1)-min(baterije$V1))/k)
> c
[1] 611
> hist(baterije$V1,probability=TRUE,breaks=seq(13.5,by=c,length=k+1),
+xlab="trajanje baterije/sti",ylab="relativne frekvencije",
+main="baterije",col="yellow")
```



△

Zadatak 1.5 Izmjeren je kapacitet 485 istovrsnih kondenzatora. Rezultati mjerenja su izraženi u μF:

i	razred	frekvencija
1	19.58 - 19.62	3
2	19.63 - 16.67	5
3	19.68 - 19.72	5
4	19.73 - 19.77	20
5	19.78 - 19.82	35
6	19.83 - 19.87	74
7	19.88 - 19.92	92
8	19.93 - 19.97	83
9	19.98 - 20.02	70
10	20.03 - 20.07	54
11	20.08 - 20.12	27
12	20.13 - 20.17	12
13	20.18 - 20.22	2
14	20.23 - 20.27	3
Σ		485

(a) Izračunajte aritmetičku sredinu i standardnu devijaciju uzorka.

(b) Izračunajte medijan te donji i gornji kvartil.

Rješenje: Budući da nemamo točno izmjerene podatke, već samo razrede i pripadne frekvencije, aritmetičku sredinu i standardnu devijaciju računamo tako da umjesto y_i promatramo sredine razreda $\overline{x}_i$.

Kod računanja s^2 u formuli možemo umjesto $n-1$ uzeti n jer je $n = 485$ velik. Tada je

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{k} f_i\overline{x}_i \quad \text{i} \quad s^2 = \frac{1}{n}\sum_{i=1}^{k} f_i(\overline{x}_i - \overline{x})^2.$$

Kako sredine razreda $\overline{x}_i$ čine aritmetički niz, možemo pojednostavniti formule za $\overline{x}$ i s^2. Odaberemo referentnu vrijednost, najčešće uzmemo mod uzorka, ovdje $\overline{x}_0 = 19.90$. Stavimo

$$d_i = \frac{\overline{x}_i - \overline{x}_0}{c} \quad \Rightarrow \quad \overline{x}_i = \overline{x}_0 + cd_i, \quad i = 1, 2, \ldots, k.$$

Sada je

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{k} f_i\overline{x}_i = \frac{1}{n}\sum_{i=1}^{k} f_i\left(\overline{x}_0 + cd_i\right) = c\cdot\frac{1}{n}\sum_{i=1}^{k} f_i d_i + \overline{x}_0 \underbrace{\frac{1}{n}\sum_{i=1}^{k} f_i}_{=1}$$

$$\Rightarrow \quad \overline{x} = c\cdot\frac{1}{n}\sum_{i=1}^{k} f_i d_i + \overline{x}_0,$$

$$\begin{aligned}
s^2 &= \frac{1}{n}\sum_{i=1}^{k} f_i(\overline{x}_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{k} f_i\left(\overline{x}_0 + cd_i - c\cdot\frac{1}{n}\sum_{j=1}^{k} f_j d_j - \overline{x}_0\right)^2 = \\
&= c^2\cdot\frac{1}{n}\sum_{i=1}^{k} f_i\left[d_i^2 - 2\left(\frac{1}{n}\sum_{j=1}^{k} f_j d_j\right)d_i + \left(\frac{1}{n}\sum_{j=1}^{k} f_j d_j\right)^2\right] = \\
&= c^2\left[\frac{1}{n}\sum_{i=1}^{k} f_i d_i^2 - 2\left(\frac{1}{n}\sum_{j=1}^{k} f_j d_j\right)^2 + \left(\frac{1}{n}\sum_{j=1}^{k} f_j d_j\right)^2\right]
\end{aligned}$$

$$\Rightarrow \quad s^2 = c^2\left[\frac{1}{n}\sum_{i=1}^{k} f_i d_i^2 - \left(\frac{1}{n}\sum_{j=1}^{k} f_j d_j\right)^2\right].$$

(a) Vratimo se zadatku. Širina razreda je $c = 0.05$ i tablica frekvencija glasi:

i	I_i	f_i	d_i	$f_i d_i$	$f_i d_i^2$	r_i	$\bar{x}_i$	F_i
1	$[19.575, 19.625\rangle$	3	-6	-18	108	0.006	19.60	0.006
2	$[19.625, 19.675\rangle$	5	-5	-25	125	0.010	19.65	0.016
3	$[19.675, 19.725\rangle$	5	-4	-20	80	0.010	19.70	0.026
4	$[19.725, 19.775\rangle$	20	-3	-60	180	0.041	19.75	0.067
5	$[19.775, 19.825\rangle$	35	-2	-70	140	0.072	19.80	0.139
6	$[19.825, 19.875\rangle$	74	-1	-74	74	0.153	19.85	0.292
7	$[19.875, 19.925\rangle$	92	0	0	0	0.190	19.90	0.482
8	$[19.925, 19.975\rangle$	83	1	83	83	0.171	19.95	0.653
9	$[19.975, 20.025\rangle$	70	2	140	280	0.144	20.00	0.797
10	$[20.025, 20.075\rangle$	54	3	162	486	0.111	20.05	0.908
11	$[20.075, 20.125\rangle$	27	4	108	432	0.056	20.10	0.964
12	$[20.125, 20.175\rangle$	12	5	60	300	0.025	20.15	0.989
13	$[20.175, 20.225\rangle$	2	6	12	72	0.004	20.20	0.993
14	$[20.225, 20.275\rangle$	3	7	21	147	0.007	20.25	1.000
Σ		485		319	2507	1.000		

Pri tome su F_i **kumulativne relativne frekvencije**, odnosno

$$\begin{aligned} F_1 &= r_1, \\ F_i &= F_{i-1} + r_i, \quad i = 2, \ldots, k. \end{aligned}$$

Sada je

$$\begin{aligned} \overline{x} &= 0.05 \cdot \frac{319}{485} + 19.90 = 19.93\,\mu F, \\ s^2 &= 0.05^2 \cdot \left[\frac{2507}{485} - \left(\frac{319}{485} \right)^2 \right] = 0.012 \quad \Rightarrow \quad s = 0.11\,\mu F. \end{aligned}$$

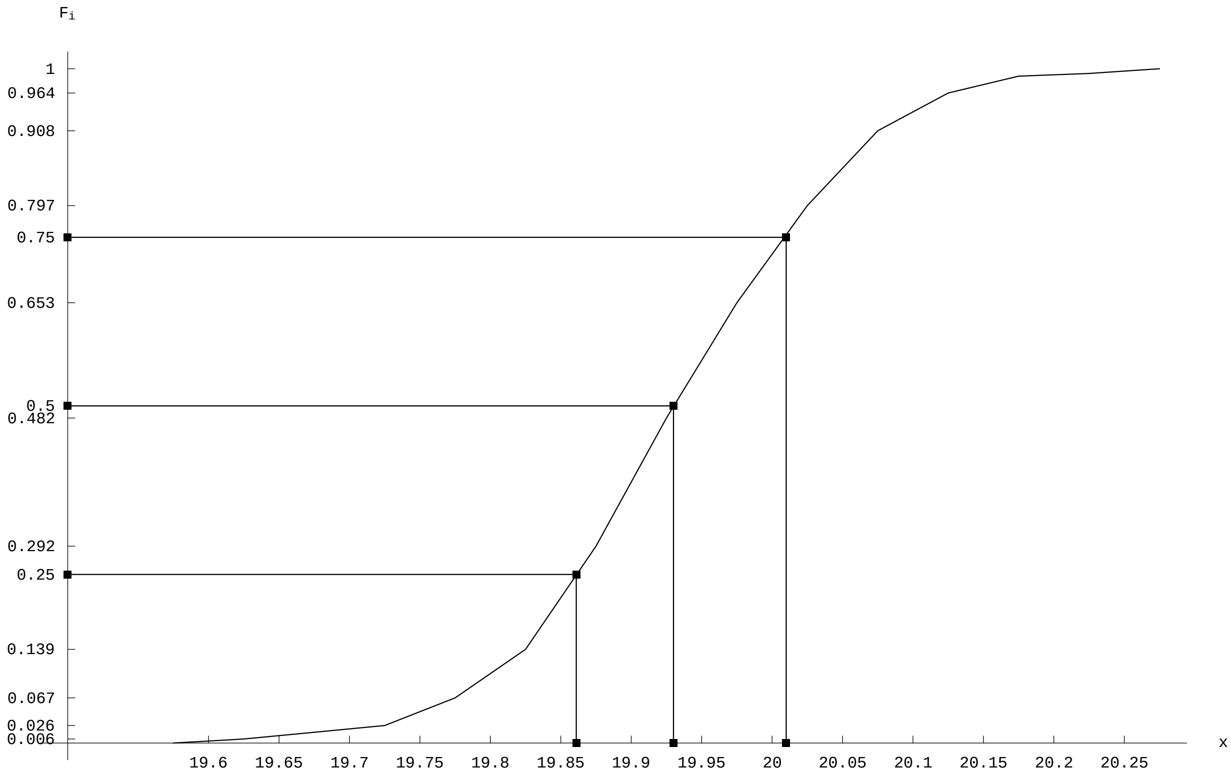
(b) Za m, q_L i q_U crtamo graf kumulativnih relativnih frekvencija (vidi sliku 1.1) i onda m, q_L i q_U odredimo linearnom interpolacijom:

$$\frac{m - 19.925}{19.975 - 19.925} = \frac{0.5 - 0.482}{0.653 - 0.482} \quad \Rightarrow \quad m = 19.925 + \frac{0.05}{0.171} \cdot 0.018 = 19.93,$$

$$\frac{q_L - 19.825}{19.875 - 19.825} = \frac{0.25 - 0.139}{0.292 - 0.139} \quad \Rightarrow \quad q_L = 19.825 + \frac{0.05}{0.153} \cdot 0.111 = 19.86,$$

$$\frac{q_U - 19.975}{20.025 - 19.975} = \frac{0.75 - 0.653}{0.797 - 0.653} \quad \Rightarrow \quad q_U = 19.975 + \frac{0.05}{0.144} \cdot 0.097 = 20.01.$$

△



Slika 1.1: Graf kumulativnih relativnih frekvencija

Zadatak 1.6 Pokažite da za varijancu uzorka $x_1, \ldots, x_n$ vrijedi

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right).$$

Rješenje: Po definiciji je

$$\begin{aligned} s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i^2 - 2x_i\cdot\overline{x} + \overline{x}^2\right) = \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - 2\overline{x}\sum_{i=1}^{n} x_i + n\overline{x}^2\right) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - 2\overline{x}\cdot n\overline{x} + n\overline{x}^2\right) = \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right). \end{aligned}$$

$\triangle$

Napomena: Ako u nizu ima više istovrsnih mjerenja tako da se vrijednosti $y_1, \ldots y_k$

pojavljuju s frekvencijama $f_1, \ldots, f_k$, onda iz 1.6 slijedi

$$s^2 = \frac{1}{f_1 + \cdots + f_k - 1}\left(\sum_{i=1}^{k} f_i y_i^2 - n\overline{x}^2\right), \text{ gdje je } \overline{x} = \frac{1}{f_1 + \cdots + f_k}\sum_{i=1}^{k} f_i y_i.$$

Zadatak 1.7 Pokažite da je aritmetička sredina podataka $x_1, \ldots, x_n$ jedinstveni broj u kojem funkcija

$$v\colon \mathbb{R} \to \mathbb{R}, \quad v(\mu) = \sum_{i=1}^{n}(x_i - \mu)^2$$

postiže minimum.

Rješenje: Neka je $\mu \in \mathbb{R}$.

$$\begin{aligned}
\Rightarrow \quad v(\mu) &= \sum_{i=1}^{n}(x_i - \mu)^2 = \sum_{i=1}^{n}[(x_i - \overline{x}) + (\overline{x} - \mu)]^2 = \\
&= \sum_{i=1}^{n}(x_i - \overline{x})^2 + 2\sum_{i=1}^{n}(x_i - \overline{x})(\overline{x} - \mu) + \sum_{i=1}^{n}(\overline{x} - \mu)^2 = \\
&= \sum_{i=1}^{n}(x_i - \overline{x})^2 + 2(\overline{x} - \mu)\underbrace{\left(\sum_{i=1}^{n} x_i - n\overline{x}\right)}_{=0} + n(\overline{x} - \mu)^2 \\
\Rightarrow \quad v(\mu) &= \sum_{i=1}^{n}(x_i - \overline{x})^2 + \underbrace{n(\overline{x} - \mu)^2}_{\geq 0} \geq \sum_{i=1}^{n}(x_i - \overline{x})^2 = v(\overline{x})
\end{aligned}$$

Dakle, funkcija v postiže minimum u $\mu = \overline{x}$. Nadalje,

$$v(\mu) = v(\overline{x}) \quad \Leftrightarrow \quad n(\overline{x} - \mu)^2 = 0 \quad \Leftrightarrow \quad \overline{x} = \mu.$$

Stoga je točka minimuma jedinstvena.

△

Zadatak 1.8 Ako su podaci dobiveni afinom transformacijom

$$y_i = ax_i + b, \quad i = 1, \ldots, n, \quad a \neq 0,$$

podataka $x_1, \ldots, x_n$, pokažite da je aritmetička sredina transformiranih podataka jednaka $\overline{y} = a\overline{x} + b$, a uzoračka varijanca $s^2(y) = a^2 s^2$.

Rješenje: Aritmetička sredina je

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n}(ax_i + b) = a \cdot \frac{1}{n}\sum_{i=1}^{n} x_i + \frac{1}{n} \cdot nb = a\overline{x} + b,$$

a uzoračka varijanca

$$
\begin{aligned}
s^2(y) &= \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - n\overline{y}^2\right) = \frac{1}{n-1}\left[\sum_{i=1}^{n}(ax_i+b)^2 - n(a\overline{x}+b)^2\right] = \\
&= \frac{1}{n-1}\left(a^2\sum_{i=1}^{n} x_i^2 + 2ab\underbrace{\sum_{i=1}^{n} x_i}_{=n\overline{x}} + nb^2 - na^2b^2 - 2abn\overline{x} - nb^2\right) = \\
&= a^2\cdot\frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right) = a^2s^2.
\end{aligned}
$$

△

Napomena: Za standardnu devijaciju slijedi $s(y) = \sqrt{s^2(y)} = \sqrt{a^2s^2} = |a|s$.

Zadatak 1.9 (Čebievljeva nejednakost) Neka je $x_1,\ldots,x_n$ skup podataka i $\varepsilon > 0$. Pokažite da tada vrijedi:

$$
\#\{i\colon |x_i-\overline{x}| \geq \varepsilon\} \leq \frac{(n-1)s^2}{\varepsilon^2}.
$$

Rješenje: Vrijedi

$$
\begin{aligned}
s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2 = \frac{1}{n-1}\sum_{\{i\colon |x_i-\overline{x}|\geq\varepsilon\}}\underbrace{(x_i-\overline{x})^2}_{\geq\varepsilon^2} + \frac{1}{n-1}\sum_{\{i\colon |x_i-\overline{x}|<\varepsilon\}}\underbrace{(x_i-\overline{x})^2}_{\geq 0} \\
&\geq \frac{1}{n-1}\sum_{\{i\colon |x_i-\overline{x}|\geq\varepsilon\}}\varepsilon^2 = \frac{1}{n-1}\#\{i\colon |x_i-\overline{x}|\geq\varepsilon\}\cdot\varepsilon^2 \\
\Rightarrow \quad & \#\{i\colon |x_i-\overline{x}|\geq\varepsilon\} \leq \frac{(n-1)s^2}{\varepsilon^2}.
\end{aligned}
$$

△

Zadatak 1.10 Neka je $x_1,\ldots,x_n$ skup podataka. Tada postoji jedinstveni broj $m\in\mathbb{R}$ koji minimizira funkciju

$$
d(\mu) = \sum_{i=1}^{n}|x_i-\mu|
$$

ako i samo ako je n neparan ili je $n = 2l$ i vrijedi $x_{(l)} = x_{(l+1)}$.

Rješenje: Poredajmo podatke: $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$.

- Ako je $\mu \leq x_{(1)}$, onda je $d(\mu) = \sum_{i=1}^{n}|x_i-\mu| = \sum_{i=1}^{n}(x_{(i)}-\mu) = \sum_{i=1}^{n}x_{(i)} - n\mu$.

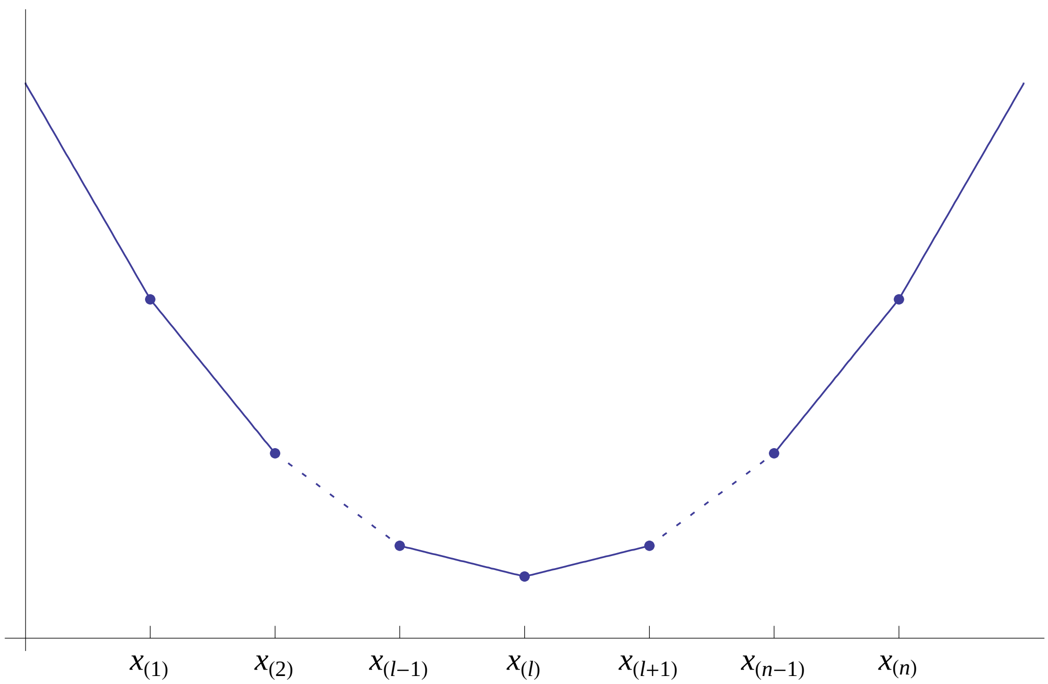
- Ako je $\mu \geq x_{(n)}$, onda je $d(\mu) = \sum_{i=1}^{n} |x_i - \mu| = \sum_{i=1}^{n} (\mu - x_{(i)}) = n\mu - \sum_{i=1}^{n} x_{(i)}$.
- Ako je $x_{(k)} \leq \mu \leq x_{(k+1)}$, za neki $1 \leq k \leq n-1$, onda je

$$\begin{aligned} d(\mu) &= \sum_{i=1}^{n} |x_{(i)} - \mu| = \sum_{i=1}^{k} |x_{(i)} - \mu| + \sum_{i=k+1}^{n} |x_{(i)} - \mu| = \\ &= \sum_{i=1}^{k} (\mu - x_{(i)}) + \sum_{i=k+1}^{n} (x_{(i)} - \mu) = (2k-n)\mu - \sum_{i=1}^{k} x_{(i)} + \sum_{i=k+1}^{n} x_{(i)}. \end{aligned}$$

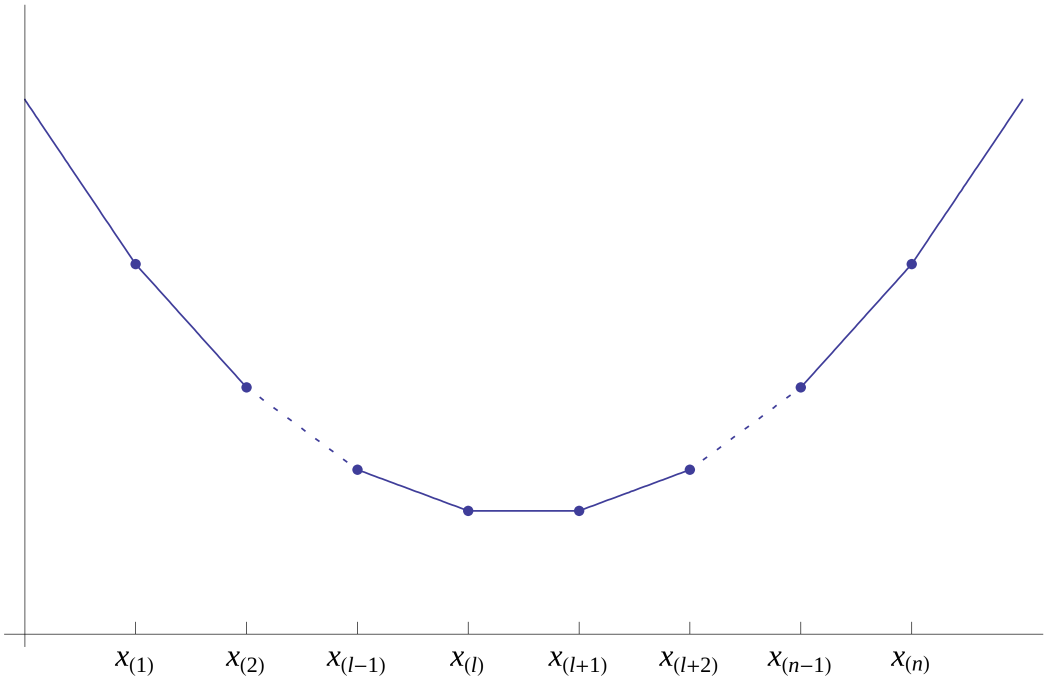
Sada možemo skicirati grafove funkcije d za n neparan i n paran.

Ako je $n = 2l - 1$, iz slike 1.2 vidimo da je jedinstveni minimum funkcije $m = x_{(l)}$.

Ako je $n = 2l$, iz slike 1.3 vidimo da funkcija d poprima minimum za sve $\mu \in [x_{(l)}, x_{(l+1)}]$ pa je minimum m jedinstven ako i samo ako je $x_{(l)} = x_{(l+1)}$.



Slika 1.2: Graf funkcije d za n=2l-1.



Slika 1.3: Graf funkcije d za n=2l.

Pretpostavimo da imamo dvodimenzionalno obilježje dano kontingencijskom tablicom:

$X\backslash Y$	b_1	b_2	$\cdots$	b_c	Σ
a_1	f_{11}	f_{12}	$\cdots$	f_{1c}	f_1
a_2	f_{21}	f_{22}	$\cdots$	f_{2c}	f_2
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
a_r	f_{r1}	f_{r2}	$\cdots$	f_{rc}	f_r
Σ	g_1	g_2	$\cdots$	g_c	n

Mjera statističke nezavisnosti obilježja je definirana sa

$$f^2 := \frac{1}{n^2}\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(nf_{ij}-f_ig_j)^2}{f_ig_j}.$$

Zadatak 1.11 Pokažite da je

(a) $f^2 = \displaystyle\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{f_{ij}^2}{f_ig_j} - 1,$

(b) $0 \le f^2 \le \min\{r,c\} - 1,$

(c) $f^2 = \min\{r,c\} - 1$ ako i samo ako je $r \ge c$ ($r \le c$) i u svakom retku (stupcu) kontingencijske tablice je točno jedna frekvencija različita od nule.

Rješenje:

(a)

$$\begin{aligned}
f^2 &= \frac{1}{n^2}\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(nf_{ij}-f_ig_j)^2}{f_ig_j} = \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{n^2f_{ij}^2}{f_ig_j} - \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{2nf_{ij}f_ig_j}{f_ig_j} + \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(f_ig_j)^2}{f_ig_j}\right) = \\
&= \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{f_{ij}^2}{f_ig_j} - \frac{2}{n}\underbrace{\sum_{i=1}^{r}\sum_{j=1}^{c}f_{ij}}_{=n} + \frac{1}{n^2}\underbrace{\left(\sum_{i=1}^{r}f_i\right)}_{=n}\underbrace{\left(\sum_{j=1}^{c}g_j\right)}_{=n} = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{f_{ij}^2}{f_ig_j} - 1.
\end{aligned}$$

(b) Očito je $f^2 \geq 0$.

- $$\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{f_{ij}^2}{f_ig_j} = \sum_{i=1}^{r}\sum_{j=1}^{c}\underbrace{\frac{f_{ij}}{f_i}}_{\leq 1}\frac{f_{ij}}{g_j} \leq \sum_{j=1}^{c}\frac{1}{g_j}\underbrace{\sum_{i=1}^{r}f_{ij}}_{=g_j} = c \quad \Rightarrow \quad f^2 \leq c-1$$
- $$\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{f_{ij}^2}{f_ig_j} = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{f_{ij}}{f_i}\underbrace{\frac{f_{ij}}{g_j}}_{\leq 1} \leq \sum_{i=1}^{r}\frac{1}{f_i}\underbrace{\sum_{j=1}^{c}f_{ij}}_{=f_i} = r \quad \Rightarrow \quad f^2 \leq r-1$$

Dakle, $f^2 \leq \min\{r,c\} - 1$.

(c) Pretpostavimo da je $r \geq c$.

$\Rightarrow$ Ako u nekom retku i_0 postoje barem dvije frekvencije različite od nula, tada je $f_{i_0j} < f_{i_0}$, za svaki $j = 1, \ldots, c$ pa vrijedi

$$\begin{aligned}
f^2 &= \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{f_{ij}^2}{f_ig_j} - 1 = \sum_{i=1, i\neq i_0}^{r}\sum_{j=1}^{c}\underbrace{\frac{f_{ij}}{f_i}}_{\leq 1}\frac{f_{ij}}{g_j} + \sum_{j=1}^{c}\underbrace{\frac{f_{i_0j}}{f_{i_0}}}_{<1}\frac{f_{ij}}{g_j} - 1 \\
&< \sum_{i=1, i\neq i_0}^{r}\sum_{j=1}^{c}\frac{f_{ij}}{g_j} + \sum_{j=1}^{c}\frac{f_{ij}}{g_j} - 1 = \sum_{j=1}^{c}\frac{1}{g_j}\underbrace{\sum_{i=1}^{r}f_{ij}}_{=g_j} - 1 = c - 1 = \min\{r,c\} - 1.
\end{aligned}$$

$\Leftarrow$ Neka je u svakom retku točno jedna frekvencija različita od nula. Označimo ih sa $f_{1j_1}, f_{2j_2}, \ldots, f_{rj_r}$. Tada vrijedi $f_i = f_{ij_i}$, za svaki $i = 1, \ldots, r$, pa je

$$f^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{f_{ij}^2}{f_i g_j} - 1 = \sum_{i=1}^{r}\frac{f_{ij_i}^2}{f_i g_{j_i}} - 1 = \sum_{i=1}^{r}\frac{f_{ij_i}}{g_{j_i}} - 1 =$$
$$= \sum_{i=1}^{r}\sum_{k=1}^{c}\frac{f_{ik}}{g_k} - 1 = \sum_{k=1}^{c}\frac{1}{g_k}\underbrace{\sum_{i=1}^{r} f_{ik}}_{=g_k} - 1 = c - 1 = \min\{r, c\} - 1.$$

Slučaj $r \leq c$ se pokazuje na isti način.

$\triangle$

Napomena: $f^2 = \min\{r, c\} - 1 \quad \Leftrightarrow \quad$ postoji funkcijska veza između obilježja.

Zadatak 1.12 U cilju istraživanja imaju li muškarci i žene isti stav prema boksu, anketirano je 350 osoba i dobiveni su sljedeći podaci:

spol \ stav	pozitivan	negativan	ravnodušan
muškarci	130	40	30
žene	42	80	28

(a) Odredite marginalne distribucije.

(b) Izračunajte stupanj statističke zavisnosti.

Rješenje:

(a) Vrijedi

spol \ stav	pozitivan	negativan	ravnodušan	Σ
muškarci	130	40	30	200
žene	42	80	28	150
Σ	172	120	58	350

Marginalne distribucije su: $f_1 = 200, f_2 = 150, g_1 = 172, g_2 = 120, g_3 = 58$.

(b) Mjera statističke nezavisnosti obilježja iznosi

$$f^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{f_{ij}^2}{f_i g_j} - 1 =$$
$$= \frac{130^2}{200\cdot 172} + \frac{40^2}{200\cdot 120} + \frac{30^2}{200\cdot 58} + \frac{42^2}{150\cdot 172} + \frac{80^2}{150\cdot 120} + \frac{28^2}{150\cdot 58} - 1 =$$
$$= 0.1495745.$$

Stupanj statističke zavisnosti je

$$o = \frac{f^2}{\min\{r,c\}-1} = \frac{0.1495745}{\min\{2,3\}-1} = 0.1495745,$$

tj. oko 14.95% pa je statistička zavisnost stava o boksu i spola slaba.

U R-u:

```
> r<-2
> c<-3
> spol<-matrix(c(130,40,30,42,80,28),nrow=r,ncol=c,byrow=T)
> spol
     [,1] [,2] [,3]
[1,]  130   40   30
[2,]   42   80   28
> sume_stupaca<-colSums(spol)
> sume_redaka<-rowSums(spol)
> sume_stupaca
[1] 172 120  58
> sume_redaka
[1] 200 150
> marginalne<-matrix(kronecker(sume_redaka,sume_stupaca),nrow=r,ncol=c,byrow=T)
> marginalne
      [,1]  [,2]  [,3]
[1,] 34400 24000 11600
[2,] 25800 18000  8700
> f<-sum(spol^2/marginalne)-1
> f
[1] 0.1495745
> o=f/(min(r,c)-1)
> o
[1] 0.1495745
```

$\triangle$

Zadatak 1.13 Visine i udaljenosti koje su bile potrebne za osvajanje zlatne medalje na olimpijskim igrama u disciplinama skoka u vis i bacanju diska za atletičare dane su u metrima u sljedećoj tablici. Odredite uzoračku kovarijancu i Pearsonov koeficijent korelacije.

godina	skok u vis	bacanje diska	godina	skok u vis	bacanje diska
1896	1.81	29.15	1956	2.12	56.36
1900	1.90	36.04	1960	2.16	59.18
1904	1.80	39.28	1964	2.18	61.00
1906	1.775	41.46	1968	2.24	64.78
1908	1.905	40.89	1972	2.23	64.40
1912	1.93	45.21	1976	2.25	67.50
1920	1.935	44.685	1980	2.36	66.64
1924	1.98	46.155	1984	2.35	66.60
1928	1.94	47.32	1988	2.38	68.82
1932	1.97	49.49	1992	2.34	65.12
1936	2.03	50.48	1996	2.39	69.40
1948	1.98	52.78	2000	2.35	69.30
1952	2.04	55.03	2004	2.36	69.89

Rješenje: Stavimo X=visina u skoku u vis, Y=udaljenost u bacanju diska.

i	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	1.81	29.15	3.2761	849.723	52.7615
2	1.90	36.04	3.61	1298.88	68.4760
3	1.80	39.28	3.24	1542.92	70.7040
4	1.775	41.46	3.15063	1718.93	73.5915
5	1.905	40.89	3.62903	1671.99	77.8955
6	1.93	45.21	3.7249	2043.94	87.2553
7	1.935	44.685	3.74423	1996.75	86.4655
8	1.98	46.155	3.9204	2130.28	91.3869
9	1.94	47.32	3.7636	2239.18	91.8008
10	1.97	49.49	3.8809	2449.26	97.4953
11	2.03	50.48	4.1209	2548.23	102.474
12	1.98	52.78	3.9204	2785.73	104.504
13	2.04	55.03	4.1616	3028.30	112.261
14	2.12	56.36	4.4944	3176.45	119.483
15	2.16	59.18	4.6656	3502.27	127.829
16	2.18	61.00	4.7524	3721.00	132.980
17	2.24	64.78	5.0176	4196.45	145.107
18	2.23	64.40	4.9729	4147.36	143.612
19	2.25	67.50	5.0625	4556.25	151.875
20	2.36	66.64	5.5696	4440.89	157.270
21	2.35	66.60	5.5225	4435.56	156.510
22	2.38	68.82	5.6644	4736.19	163.792
23	2.34	65.12	5.4756	4240.61	152.381
24	2.39	69.40	5.7121	4816.36	165.866
25	2.35	69.30	5.5225	4802.49	162.855
26	2.36	69.89	5.5696	4884.61	164.940
Σ	54.705	1426.96	116.144	81960.62	3061.573

Radi se o dvodimenzionalnom neprekidnom statističkom obilježju. Iz prethodne tablice slijedi $n = 26, \overline{x} = 2.104, \overline{y} = 54.883$ i vrijedi

$$S_{XX} = \sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2 = 116.144 - 26 \cdot 2.104^2 = 1.043,$$

$$S_{YY} = \sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n} y_i^2 - n\overline{y}^2 = 81960.62 - 26 \cdot 54.883^2 = 3644.669,$$

$$S_{XY} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n} x_i y_i - n\overline{x}\overline{y} = 3061.573 - 26 \cdot 2.104 \cdot 54.883 = 59.193.$$

Dakle, uzoračka kovarijanca je

$$C(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \frac{1}{n-1} S_{XY} = 2.368.$$

Nadalje, Pearsonov koeficijent korelacije je

$$r_{XY} = \frac{S_{XY}}{\sqrt{S_{XX} S_{YY}}} = 0.96$$

pa su X i Y pozitivno korelirane.

Riješimo zadatak u R-u.

Pronađite datoteku `olimpijade.dat` na

`http://www.math.hr/nastava/stat/vjezbe.html`

i spremite je u radni direktorij. U datoteci su podaci o visinama i udaljenostima koje su bile potrebne za osvajanje zlatne medalje na olimpijskim igrama od 1896. godine. Učitajmo podatke:

```
> olimp<-read.table("olimpijade.dat",header=TRUE)
> olimp
   godina skok_u_vis bacanje_diska
1    1896      1.810        29.150
2    1900      1.900        36.040
3    1904      1.800        39.280
4    1906      1.775        41.460
5    1908      1.905        40.890
6    1912      1.930        45.210
7    1920      1.935        44.685
8    1924      1.980        46.155
9    1928      1.940        47.320
10   1932      1.970        49.490
```

```
11   1936    2.030    50.480
12   1948    1.980    52.780
13   1952    2.040    55.030
14   1956    2.120    56.360
15   1960    2.160    59.180
16   1964    2.180    61.000
17   1968    2.240    64.780
18   1972    2.230    64.400
19   1976    2.250    67.500
20   1980    2.360    66.640
21   1984    2.350    66.600
22   1988    2.380    68.820
23   1992    2.340    65.120
24   1996    2.390    69.400
25   2000    2.350    69.300
26   2004    2.360    69.890
```

Tada je Pearsonov koeficijent korelacije

```
> cor(olimp$skok_u_vis,olimp$bacanje_diska,method="pearson")
[1] 0.9600989
```

Drugi način:

```
> attach(olimp)
> SXX=sum((skok_u_vis-mean(skok_u_vis))^2)
> SYY=sum((bacanje_diska-mean(bacanje_diska))^2)
> SXY=sum((skok_u_vis-mean(skok_u_vis))*(bacanje_diska-mean(bacanje_diska)))
> pearsonov_koef<-SXY/sqrt(SXX*SYY)
> pearsonov_koef
[1] 0.9600989
```

$\triangle$

Zadatak 1.14 Pokažite da je Pearsonov koeficijent korelacije podataka

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

jednak uzoračkoj kovarijanci standardiziranih podataka

$$\left(\frac{x_1 - \overline{x}}{s_x}, \frac{y_1 - \overline{y}}{s_y}\right), \ldots, \left(\frac{x_n - \overline{x}}{s_x}, \frac{y_n - \overline{y}}{s_y}\right).$$

Rješenje: Ako stavimo $x_i' = \dfrac{x_i - \overline{x}}{s_x}$ i $y_i' = \dfrac{y_i - \overline{x}}{s_x}$ za $i = 1, \ldots, n$, uzoračka kovarijanca standardiziranih podataka je jednaka

$$S_{X'Y'} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i' - \overline{x'})(y_i' - \overline{y'}) = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s_x} - \frac{\overline{x} - \overline{x}}{s_x}\right)\left(\frac{y_i - \overline{y}}{s_y} - \frac{\overline{y} - \overline{y}}{s_y}\right) =$$
$$= \frac{1}{s_x s_y}\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \frac{1}{\sqrt{\frac{S_{XX}}{n-1}}\sqrt{\frac{S_{YY}}{n-1}}}\frac{1}{n-1}S_{XY} = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}} = r_{XY}.$$

△

Zadatak 1.15 Pokažite da je Pearsonov koeficijent korelacije invarijantan na afine transformacije s istim predznakom koeficijenata smjera, odnosno ako su $a, b, c, d \in \mathbb{R}, a \cdot c \neq 0$, onda je

$$r_{aX+b,cY+d} = \operatorname{sgn}(ac) r_{XY}.$$

Rješenje: Prema zadatku 1.8 vrijedi

$$\begin{aligned} r_{aX+b,cY+d} &= \frac{S_{aX+b,cY+d}}{\sqrt{S_{aX+b,aX+b} \cdot S_{cY+d,cY+d}}} = \\ &= \frac{\sum_{i=1}^{n}(ax_i + b - (a\overline{x} + b))(cy_i + d - (c\overline{y} + d))}{\sqrt{\sum_{i=1}^{n}(ax_i + b - (a\overline{x} + b))^2 \cdot \sum_{j=1}^{n}(cy_j + d - (c\overline{y} + d))^2}} = \\ &= \frac{ac\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{a^2c^2\sum_{i=1}^{n}(x_i - \overline{x})^2 \cdot \sum_{j=1}^{n}(y_j - \overline{y})^2}} = \\ &= \frac{acS_{XY}}{|ac|\sqrt{S_{XX}S_{YY}}} = \operatorname{sgn}(ac) r_{XY}. \end{aligned}$$

△

Zadatak 1.16 Aproksimirajte podatke $(x_1, y_1), \ldots, (x_n, y_n)$ pravcem

$$y = \alpha + \beta x$$

metodom najmanjih kvadrata, tj. odredite točku $(\hat{\alpha}, \hat{\beta}) \in \mathbb{R}^2$ u kojoj funkcija

$$L(\alpha, \beta) = \sum_{i=1}^{n}(\alpha + \beta x_i - y_i)^2, \alpha, \beta \in \mathbb{R}$$

ima jedinstveni globalni minimum.

Napomena: Ako je $x_1 = x_2 = \cdots = x_n$, onda je traženi pravac $x = x_1$.

Rješenje: Pretpostavimo da su barem dvije vrijednosti x_i-ova različite. Iz

$$\partial_\alpha L = 2\sum_{i=1}^{n}(\alpha + \beta x_i - y_i) \quad \text{i} \quad \partial_\beta L = 2\sum_{i=1}^{n}(\alpha + \beta x_i - y_i)x_i,$$

nužan uvjet za lokalni ekstrem $(\hat{\alpha}, \hat{\beta})$ glasi

$$\begin{array}{c}
2\sum_{i=1}^{n}(\hat{\alpha}+\hat{\beta}x_i-y_i)=0 \\
2\sum_{i=1}^{n}(\hat{\alpha}+\hat{\beta}x_i-y_i)x_i=0 \\
\hline
\hat{\alpha}n+\hat{\beta}\sum_{i=1}^{n}x_i-\sum_{i=1}^{n}y_i=0 \\
\hat{\alpha}\sum_{i=1}^{n}x_i+\hat{\beta}\sum_{i=1}^{n}x_i^2-\sum_{i=1}^{n}x_iy_i=0 \\
\hline
\end{array}$$

$$\begin{array}{c}
\hat{\alpha}n+\hat{\beta}n\overline{x}-n\overline{y}=0 \quad / \cdot \overline{x} \\
\hat{\alpha}n\overline{x}+\hat{\beta}(S_{XX}+n\overline{x}^2)-(S_{XY}+n\overline{xy})=0
\end{array}$$

odakle je

$$\hat{\beta}=\frac{S_{XY}}{S_{XX}} \quad \text{i} \quad \hat{\alpha}=\overline{y}-\hat{\beta}\overline{x}.$$

Provjerimo sada dovoljan uvjet. Hesseova matrica glasi

$$HL(\hat{\alpha},\hat{\beta})=\begin{bmatrix}\partial_\alpha^2 L(\hat{\alpha},\hat{\beta}) & \partial_\alpha\partial_\beta L(\hat{\alpha},\hat{\beta}) \\ \partial_\beta\partial_\alpha L(\hat{\alpha},\hat{\beta}) & \partial_\beta^2 L(\hat{\alpha},\hat{\beta})\end{bmatrix}=\begin{bmatrix}2n & 2\sum_{i=1}^n x_i \\ 2\sum_{i=1}^n x_i & 2\sum_{i=1}^n x_i^2\end{bmatrix}.$$

Vrijedi $n>0$ i

$$\begin{vmatrix}2n & 2\sum_{i=1}^n x_i \\ 2\sum_{i=1}^n x_i & 2\sum_{i=1}^n x_i^2\end{vmatrix}=4n\sum_{i=1}^{n}x_i^2-4\left(\sum_{i=1}^{n}x_i\right)^2>0$$

zbog aritmetičko-kvadratne nejednakosti i pretpostavke s početka. Stoga, prema Sylvesterovom kriteriju, funkcija L ima u točki $(\hat{\alpha}, \hat{\beta})$ strogi lokalni minimum. Budući da je jedina stacionarna točka, $(\hat{\alpha}, \hat{\beta})$ je jedinstveni globalni minimum.

△

Zadatak 1.17 Za podatke iz zadatka 1.13 procijenite metodom najmanjih kvadrata visinu u skoku u vis i udaljenost u bacanju diska potrebne za osvajanje zlatne medalje na olimpijskim igrama u Pekingu 2008. godine.

Rješenje:

(a) Stavimo X = vrijeme (godine) i Y = visina u skoku u vis (m).

Tablica podataka:

i	x_i	y_i	x_i^2	$x_i\,y_i$
1	1896	1.81	3594816	3431.76
2	1900	1.90	3610000	3610.00
3	1904	1.80	3625216	3427.20
4	1906	1.775	3632836	3383.15
5	1908	1.905	3640464	3634.74
6	1912	1.93	3655744	3690.16
7	1920	1.935	3686400	3715.20
8	1924	1.98	3701776	3809.52
9	1928	1.94	3717184	3740.32
10	1932	1.97	3732624	3806.04
11	1936	2.03	3748096	3930.08
12	1948	1.98	3794704	3857.04
13	1952	2.04	3810304	3982.08

14	1956	2.12	3825936	4146.72
15	1960	2.16	3841600	4233.60
16	1964	2.18	3857296	4281.52
17	1968	2.24	3873024	4408.32
18	1972	2.23	3888784	4397.56
19	1976	2.25	3904576	4446.00
20	1980	2.36	3920400	4672.80
21	1984	2.35	3936256	4662.40
22	1988	2.38	3952144	4731.44
23	1992	2.34	3968064	4661.28
24	1996	2.39	3984016	4770.44
25	2000	2.35	4000000	4700.00
26	2004	2.36	4016016	4729.44
Σ	50706	54.705	98918276	106858.8

Slijedi $n = 26, \overline{x} = 1950.231, \overline{y} = 2.104038$ pa je

$$S_{XX} = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2 = 98918276 - 26 \cdot 1950.231^2 = 29874.62,$$
$$S_{XY} = \sum_{i=1}^{n} x_i y_i - n\overline{x}\overline{y} = 106858.8 - 26 \cdot 1950.231 \cdot 2.104038 = 171.4358.$$

Prema prethodnom zadatku je

$$\hat{\beta} = \frac{S_{XY}}{S_{XX}} = 0.00573851 \quad \text{i} \quad \hat{\alpha} = \overline{y} - \hat{\beta}\overline{x} = -9.08738.$$

Stoga je traženi pravac

$$y = -9.08738 + 0.00573851x$$

pa je procjena visine u skoku uvis za olimpijske igre 2008. godine $y(2008) = 2.436$m.

(b) Stavimo $X =$ vrijeme (godine) i $Y =$udaljenost u bacanju diska (m).

Tablica podataka:

i	x_i	y_i	x_i^2	$x_i\, y_i$
1	1896	29.150	3594816	55268.40
2	1900	36.040	3610000	68476.00
3	1904	39.280	3625216	74789.12
4	1906	41.460	3632836	79022.76
5	1908	40.890	3640464	78018.12
6	1912	45.210	3655744	86441.52
7	1920	44.685	3686400	85795.20
8	1924	46.155	3701776	88802.22
9	1928	47.320	3717184	91232.96
10	1932	49.490	3732624	95614.68
11	1936	50.480	3748096	97729.28
12	1948	52.780	3794704	102815.44
13	1952	55.030	3810304	107418.56
14	1956	56.360	3825936	110240.16
15	1960	59.180	3841600	115992.80
16	1964	61.000	3857296	119804.00
17	1968	64.780	3873024	127487.04
18	1972	64.400	3888784	126996.80
19	1976	67.500	3904576	133380.00
20	1980	66.640	3920400	131947.20
21	1984	66.600	3936256	132134.40
22	1988	68.820	3952144	136814.16
23	1992	65.120	3968064	129719.04
24	1996	69.400	3984016	138522.40
25	2000	69.300	4000000	138600.00
26	2004	69.890	4016016	140059.56
Σ	50706	1426.96	98918276	2793122

Dakle, $n = 26, \overline{x} = 1950.231, \overline{y} = 54.883$ i vrijedi

$$S_{XX} = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2 = 98918276 - 26 \cdot 1950.231^2 = 29874.62,$$
$$S_{XY} = \sum_{i=1}^{n} x_i y_i - n\overline{xy} = 2793122 - 26 \cdot 1950.231 \cdot 54.883 = 10220.52.$$

Sada je

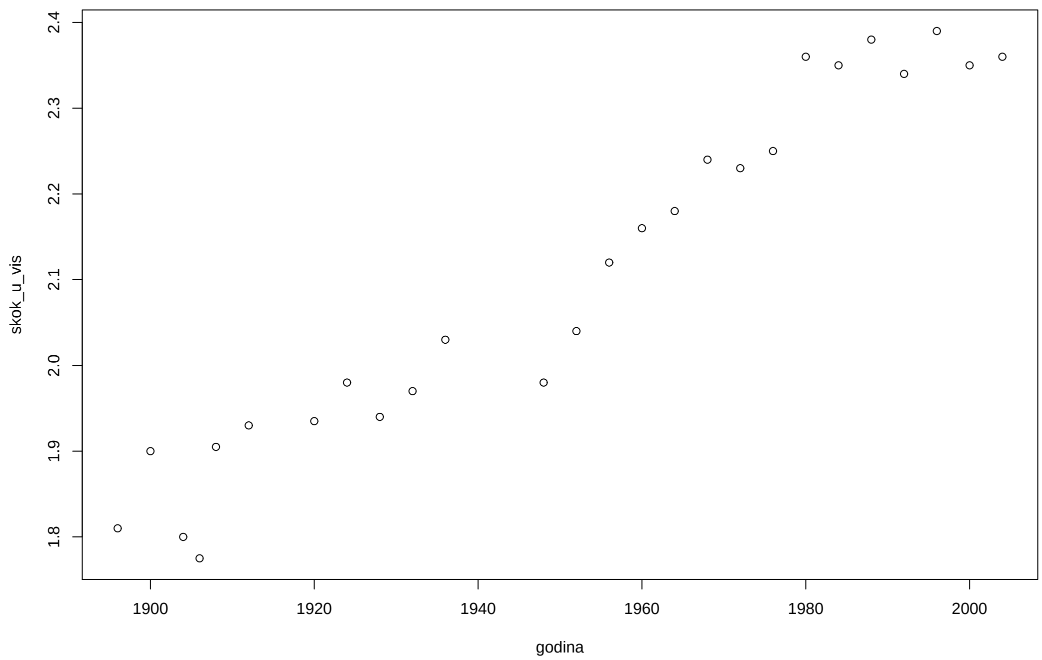
$$\hat{\beta} = \frac{S_{XY}}{S_{XX}} = 0.3421138 \quad \text{i} \quad \hat{\alpha} = \overline{y} - \hat{\beta}\overline{x} = -612.318.$$

Traženi pravac je $y = -612.318 + 0.3421138x$, a procjena $y(2008) = 74.647$m.

U R-u:

Pogledajmo kako se visina u skoku u vis mijenjala s godinama:

```
> plot(godina,skok_u_vis)
```

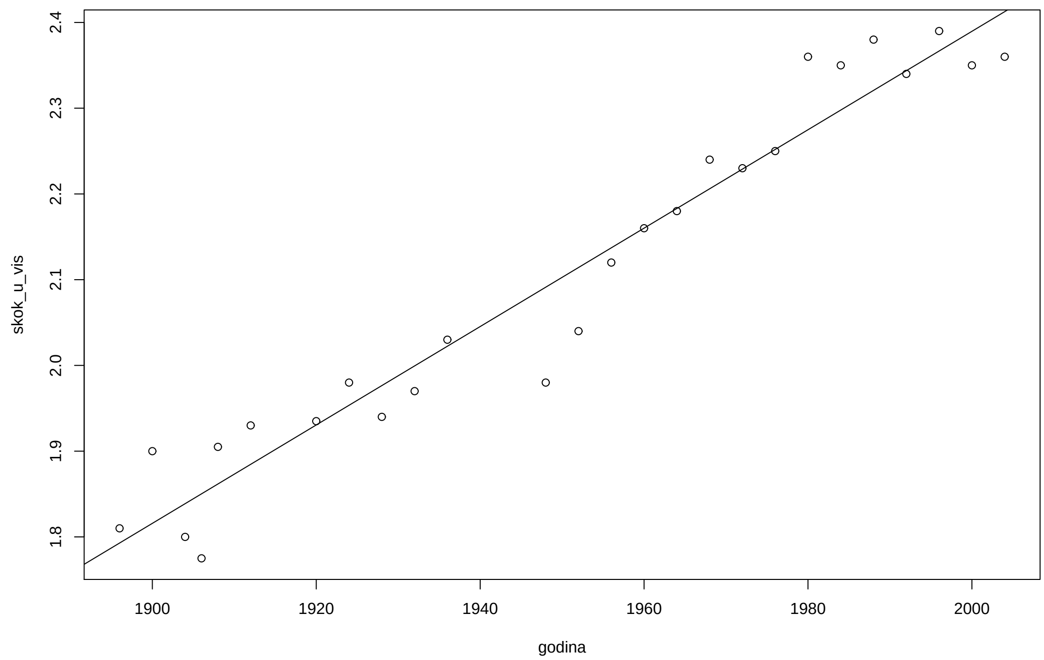


Procjena za 2008. godinu metodom najmanjih kvadrata:

```
> n=length(godina)
> SXX<-sum(godina^2)-n*mean(godina)^2
> SXY<-sum(godina*skok_u_vis)-n*mean(godina)*mean(skok_u_vis)
> beta<-SXY/SXX
> alpha=mean(skok_u_vis)-beta*mean(godina)
> f<-function(t)
+ alpha+beta*t
> f(2008)
[1] 2.435548
```

Pogledajmo kako pravac opisuje podatke:

```
> plot(godina,skok_u_vis)
> abline(c(alpha,beta))
```



Do pravca regresije možemo doći i koristeći naredbu `lm(podaci1~podaci2)`:

```
> regr<-lm(skok_u_vis~godina)
> regr

Call:
lm(formula = skok_u_vis ~ godina)

Coefficients:
(Intercept)       godina
  -9.087380     0.005739
> plot(godina,skok_u_vis)
> abline(regr$coefficients)
```

△

Zadatak 1.18 U tablici su prikazane eksperimentalne vrijednosti tlaka nekog plina konstantne mase koje odgovaraju različitim volumenima V.

volumen (V/m^3)	0.543	0.618	0.724	0.887	1.186	1.94
tlak (p/Pa)	0.612	0.492	0.376	0.284	0.192	0.101

Iz termodinamike je poznata jednadžba $pV^\gamma = C$, gdje su γ i C konstante.

(a) Odredite konstante γ i C.

(b) Odredite p za $V = 1\text{m}^3$.

Rješenje: Logaritmiranjem jednadžbe $pV^\gamma = C$ slijedi $\ln p = \ln C - \gamma \ln V$. Stavimo li $x = \ln V, y = \ln p, \alpha = \ln C$ i $\beta = -\gamma$, dobivamo pravac $y = \alpha + \beta x$. Procijenimo sada α i β metodom najmanjih kvadrata.

Podaci su dani sljedećom tablicom:

i	x_i	y_i	x_i^2	$x_i y_i$
1	-0.6106	-0.4910	0.3728	0.2998
2	-0.4812	-0.7092	0.2316	0.3413
3	-0.3229	-0.9781	0.1043	0.3159
4	-0.1199	-1.2587	0.0143	0.1509
5	0.1705	-1.6502	0.0291	-0.2815
6	0.6626	-2.2926	0.4391	-1.5193
Σ	-0.701513	-7.38014	1.19145	-0.692768

pa je $n = 6, \overline{x} = -0.116919, \overline{y} = -1.23002$ i vrijedi

$$S_{XX} = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2 = 1.19145 - 6 \cdot (-0.116919)^2 = 1.10943,$$

$$S_{XY} = \sum_{i=1}^{n} x_i y_i - n\overline{xy} = -0.692768 - 6 \cdot (-0.116919) \cdot (-1.23002) = -1.55564.$$

Prema zadatku 1.16 je

$$\hat{\beta} = \frac{S_{XY}}{S_{XX}} = -1.4 \quad \text{i} \quad \hat{\alpha} = \overline{y} - \hat{\beta}\overline{x} = -1.39$$

i traženi pravac $y = -1.39 - 1.4x$.

(a) Za konstante γ i C vrijedi $\hat{\alpha} = \ln C$ i $\hat{\beta} = -\gamma$ pa je $C = e^{\hat{\alpha}} = 0.25$ i $\gamma = -\hat{\beta} = 1.4$. Dakle, plin zadovoljava jednadžbu $pV^{1.4} = 0.25$.

(b) Za $V = 1\text{m}^3$ iz dobivene jednadžbe slijedi

$$p \cdot 1^{1.4} = 0.25 \quad \Rightarrow \quad p = 0.25.$$

$\triangle$

1.1 Zadaci za vježbu

1.19 Izabran je slučajan uzorak od 33 osobe koje slušaju radio i zabilježeno koliko sati slušaju radio tjedno. Podaci su sljedeći:

9	8	7	4	8	6	8	8	7	10	8
10	6	7	7	8	9	6	5	8	5	6
8	7	8	5	5	8	7	6	6	4	5

(a) Nacrtajte histogram relativnih frekvencija.

(b) Odredite aritmetičku sredinu, mod i medijan uzorka.

(c) Odredite varijancu i standardnu devijaciju uzorka.

(d) Odredite raspon uzorka.

(e) Odredite donji i gornji kvartil te interkvartil uzorka.

(f) Nacrtajte dijagram pravokutnika (”box and whisker plot”).

1.20 Pet novčića smo bacali 1000 puta i zabilježili dobiveni broj glava. Broj bacanja u kojima je palo 0, 1, 2, 3, 4 i 5 glava dan je u sljedećoj tablici:

broj glava	broj bacanja
0	38
1	144
2	342
3	287
4	164
5	25
Σ	1000

(a) Nacrtajte histogram podataka.

(b) Odredite aritmetičku sredinu, varijancu i standardnu devijaciju uzorka.

1.21 Izabran je slučajan uzorak od 400 brucoša nekog fakulteta i zabilježeno koliko minuta gledaju televiziju tjedno. Podaci su dani u sljedećoj tablici:

vrijeme gledanja	broj studenata
300 - 399	14
400 - 499	46
500 - 599	58
600 - 699	76
700 - 799	68
800 - 899	62
900 - 999	48
1000 - 1099	22
1100 - 1199	6
Σ	400

(a) Nacrtajte histogram podataka.

(b) Izračunajte aritmetičku sredinu i standardnu devijaciju uzorka.

(c) Izračunajte medijan te donji i gornji kvartil.

1.22 Dokažite da je $\overline{x}$ jedinstveni minimum funkcije v iz zadatka 1.7 koristeći prvu i drugu derivaciju.

1.23 Od grupe pacijenata koji su se žalili na probleme sa spavanjem nekima je dana tableta za spavanje dok je drugima dana tableta sa šećerom (makar su svi mislili da su dobili tabletu za spavanje). Kasnije su upitani je li im tableta pomogla ili ne. Dobiveni su sljedeći podaci:

	Spavali dobro	Nisu spavali dobro
Uzeli tabletu za spavanje	44	10
Uzeli tabletu sa šećerom	81	35

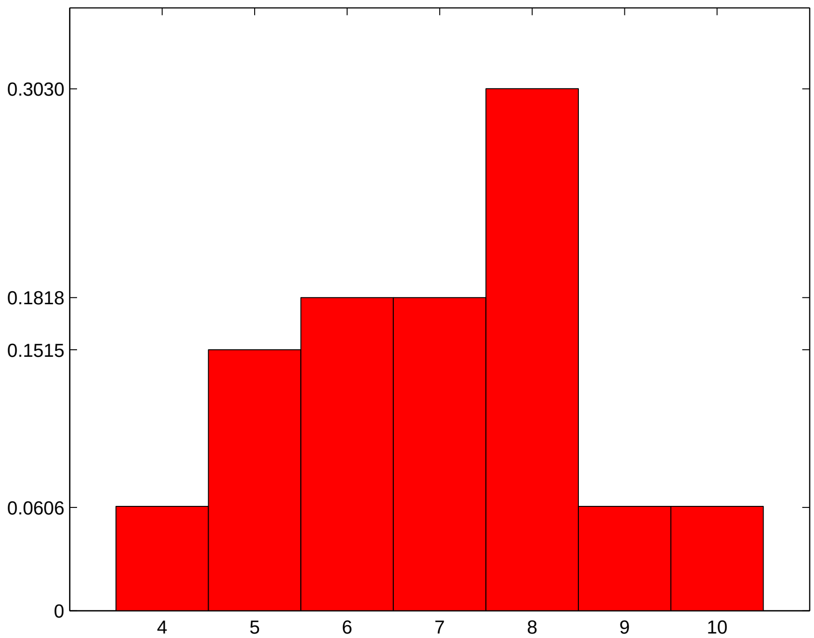
(a) Odredite marginalne distribucije.

(b) Izračunajte stupanj statističke zavisnosti.

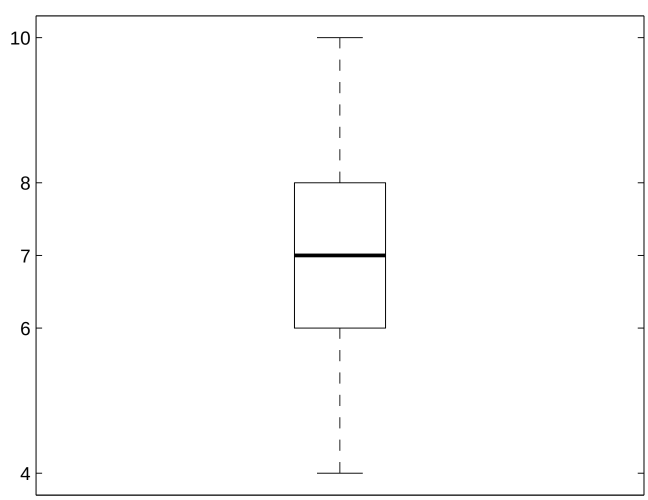
1.24 Bez upotrebe Hesseove matrice dokažite da je točka $(\hat{\alpha}, \hat{\beta})$ iz zadatka 1.16 jedinstveni globalni minimum funkcije L.

1.25 Uzimajući u obzir podatke iz zadatka 1.13 samo od 1896. do 1988. godine, odredite Pearsonov koeficijent korelacije i uzoračku kovarijancu. Nadalje, procijenite metodom najmanjih kvadrata visinu u skoku u vis i udaljenost u bacanju diska potrebne za osvajanje zlatne medalje na olimpijskim igrama u Pekingu 2008. godine, te usporedite dobiveni rezultat sa zadatkom 1.17.

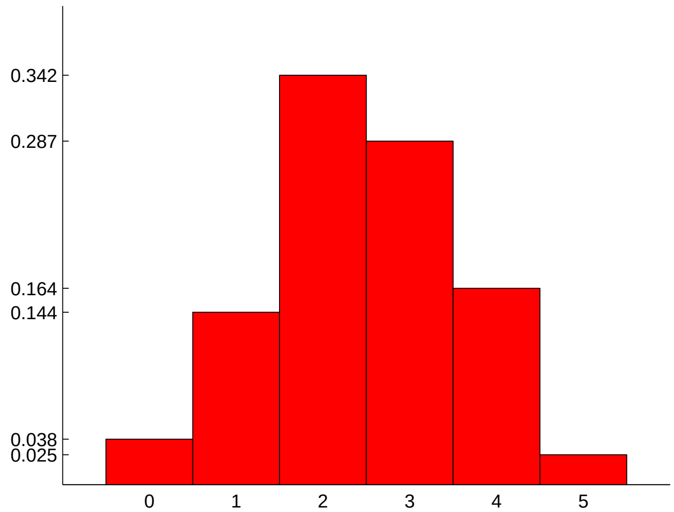
Rješenja: **1.19.** (a) Vidi sliku 1.4. (b) $\overline{x} = 6.9394$, mod=8, m=7 (c) $s^2 = 2.4962, s = 1.5799$ (d) $R = 6$ (e) $g_L = 6, q_U = 8, d_q = 2$ (f) Vidi sliku 1.5. **1.20.** (a) Vidi sliku 1.6. (b) $\overline{x} = 02.47, s^2 = 1.2443, s = 1.1155$ **1.21.** (a) Vidi sliku 1.7. (b) $\overline{x} = 715, s = 190.3953$ (c) Vidi sliku 1.8: $m = 708.324, q_L = 568.466, q_U = 860.79$. **1.23.** (a) $f_1 = 54, f_2 = 116, g_1 = 125, g_2 = 45$ (b) $o = 0.0151242$ **1.25.** Za X=visina i Y=udaljenost je $r_{XY} = 0.950104, C(X, Y) = 2.03447$. Procjena visine je 2.45465, a udaljenosti 77.3602.



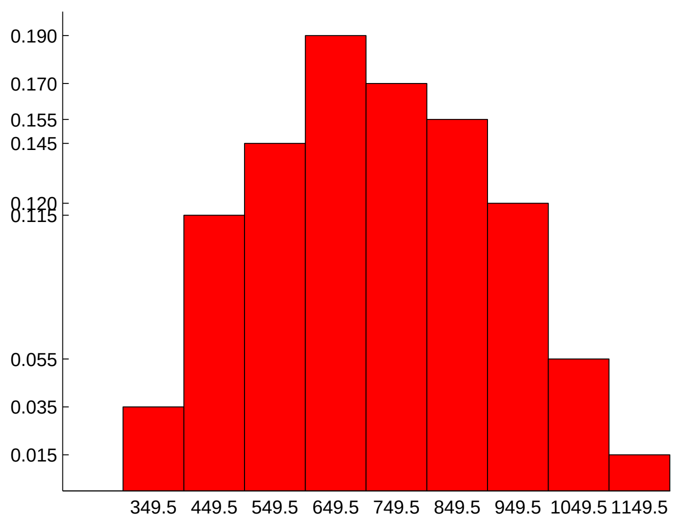
Slika 1.4: Histogram relativnih frekvencija za zadatak 1.19.



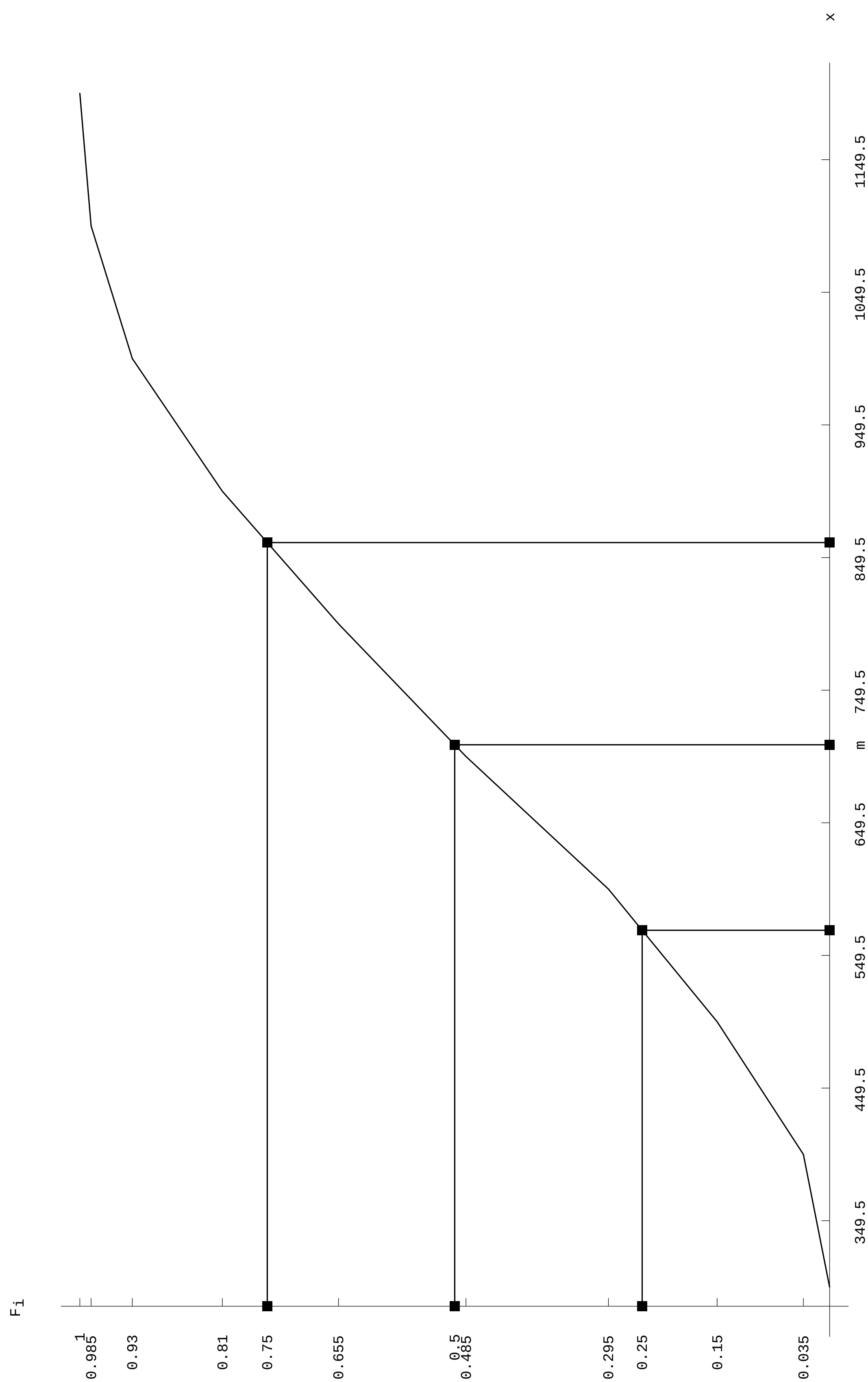
Slika 1.5: Dijagram pravokutnika za zadatak 1.19.



Slika 1.6: Histogram relativnih frekvencija za zadatak 1.20.



Slika 1.7: Histogram relativnih frekvencija za zadatak 1.21.



Slika 1.8: Graf kumulativnih frekvencija za zadatak 1.21.

2

Neprekidne slučajne varijable

Definicija: Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ vjerojatnosni prostor. Funkcija $X\colon \Omega \to \mathbb{R}$ je **slučajna varijabla** ako za sve $a, b \in \mathbb{R}$, $a < b$ vrijedi

$$\{a \leq X \leq b\} \in \mathcal{F}.$$

Funkciju $F = F_X\colon \mathbb{R} \to [0,1]$ definiranu sa

$$F(a) := \mathbb{P}(X \leq a), \quad a \in \mathbb{R}$$

zovemo **funkcija distribucije** slučajne varijable X.

Napomena: Primjetite da je

$$\{X \leq a\} = \bigcup_{n=1}^{\infty} \{-n \leq X \leq a\} \in \mathcal{F}$$

pa je funkcija distribucije dobro definirana.

Definicija: Slučajna varijabla X je **neprekidna slučajna varijabla** ako postoji nenegativna (izmjeriva) funkcija $f\colon \mathbb{R} \to \mathbb{R}$ takva da je

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f(x)\,dx, \qquad \text{za sve } a, b \in \mathbb{R},\ a < b.$$

Funkciju f zovemo **(vjerojatnosna) funkcija gustoće** slučajne varijable X.

Napomena:

(a) Vrijedi

$$F(x) = \mathbb{P}(X \leq x) = \mathbb{P}\left(\bigcup_{n=1}^{\infty} \{-n \leq X \leq x\}\right) =$$

$$= \lim_{n\to\infty} \mathbb{P}(-n \leq X \leq x) = \lim_{n\to\infty} \int\limits_{-n}^{x} f(t)\,dt = \int\limits_{-\infty}^{x} f(t)\,dt.$$

Dakle,

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

(b) Zbog $1 = \lim_{x\to+\infty} F(x) = \lim_{x\to+\infty} \int_{-\infty}^{x} f(t)\,dt = \int_{-\infty}^{\infty} f(t)\,dt$, da bi funkcija f bila vjerojatnosna funkcija gustoće, mora vrijediti

(i) $f(x) \geq 0, \forall x \in \mathbb{R}$,

(ii) $\int_{-\infty}^{\infty} f(t)\,dt = 1$.

(c) $\mathbb{P}(X = a) = \int_{a}^{a} f(t)\,dt = 0$

(d) Vrijedi

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

Slijedi

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

Analogno dobijemo $\mathbb{P}(a \leq X < b) = \mathbb{P}(a < X < b) = F(b) - F(a)$.

Definicija: Neka je X neprekidna slučajna varijabla s gustoćom f. Ako nepravi integral

$$\int_{-\infty}^{+\infty} |x| f(x)\,dx$$

konvergira, onda broj

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

zovemo **(matematičko) očekivanje** slučajne varijable X i kažemo da slučajna varijabla X ima očekivanje.

Napomena: Ako je $g\colon \mathbb{R} \to \mathbb{R}$ (izmjeriva) funkcija, onda je i $g(X)$ slučajna varijabla i vrijedi

$$\mathbb{E}[g(X)] := \int_{-\infty}^{+\infty} g(x) f(x)\,dx.$$

Definicija: Za $r \in \mathbb{N}$ sa $M_r := \mathbb{E}[X^r]$ definiramo **r-ti moment** slučajne varijable X ukoliko očekivanje postoji. Ako postoji drugi moment slučajne varijable X, onda definiramo **varijancu** slučajne varijable X sa

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

Broj $\sigma := \sqrt{\operatorname{Var} X} \geq 0$ zovemo **standardna devijacija** slučajne varijable X.

2.1 Normalna distribucija

Definicija: Slučajna varijabla X ima **normalnu distribuciju** s parametrima μ i $\sigma^2 > 0$ ako joj je funkcija gustoće

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

Pišemo: $X \sim N(\mu, \sigma^2)$

Zadatak 2.1 Neka je $X \sim N(\mu, \sigma^2)$.

(a) Pokažite da je X dobro definirana.

(b) Izračunajte prvi moment, drugi moment i varijancu od X.

Rješenje:

(a) Trebamo pokazati da je pripadna funkcija f vjerojatnosna funkcija gustoće:

(i) $f(x) \geq 0, \quad \forall x \in \mathbb{R}$.

(ii) Vrijedi $\int\limits_{-\infty}^{+\infty} f(x)\,dx = \frac{1}{\sigma\sqrt{2\pi}} \int\limits_{-\infty}^{+\infty} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx = \left\{ \begin{array}{l} t = \frac{x-\mu}{\sigma} \\ dt = \frac{dx}{\sigma} \end{array} \right\} = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{+\infty} e^{-\frac{t^2}{2}}\,dt.$

Izračunajmo $I := \int\limits_{-\infty}^{+\infty} e^{-\frac{t^2}{2}}\,dt.$

$$\Rightarrow \quad I^2 = I \cdot I = \left(\int\limits_{-\infty}^{+\infty} e^{-\frac{t^2}{2}}\,dt \right) \cdot \left(\int\limits_{-\infty}^{+\infty} e^{-\frac{t^2}{2}}\,dt \right) = \{\text{Fubinijev teorem}\} =$$

$$= \int\limits_{-\infty}^{+\infty} \int\limits_{-\infty}^{+\infty} e^{-\frac{s^2+t^2}{2}}\,ds\,dt = \left\{ \begin{array}{l} \begin{array}{l} s = r\cos\varphi \\ t = r\sin\varphi \end{array} \Rightarrow s^2 + t^2 = r^2 \\ \\ J = \begin{vmatrix} \cos\varphi & -r\sin\varphi \\ \sin\varphi & r\cos\varphi \end{vmatrix} = r \end{array} \right\} =$$

$$= \int\limits_0^{2\pi} \int\limits_0^{\infty} re^{-\frac{r^2}{2}}\,dr\,d\varphi = 2\pi \int\limits_0^{\infty} re^{-\frac{r^2}{2}}\,dr = 2\pi(-e^{-\frac{r^2}{2}})\Big|_0^{\infty} = 2\pi$$

$$\Rightarrow \quad I = \sqrt{2\pi} \quad \Rightarrow \quad \int\limits_{-\infty}^{+\infty} f(x)\,dx = \frac{1}{\sqrt{2\pi}} \cdot I = 1.$$

(b)

$$\mathbb{E}X = \int\limits_{-\infty}^{+\infty} xf(x)\,dx = \frac{1}{\sigma\sqrt{2\pi}} \int\limits_{-\infty}^{+\infty} xe^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx = \begin{Bmatrix} t = \frac{x-\mu}{\sigma} \\ dt = \frac{dx}{\sigma} \end{Bmatrix} =$$

$$= \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{+\infty} (\mu + \sigma t)e^{-\frac{t^2}{2}}\,dt = \frac{\mu}{\sqrt{2\pi}} \cdot \underbrace{\int\limits_{-\infty}^{+\infty} e^{-\frac{t^2}{2}}\,dt}_{=\sqrt{2\pi}} + \frac{\sigma}{\sqrt{2\pi}} \cdot \underbrace{\int\limits_{-\infty}^{+\infty} te^{-\frac{t^2}{2}}\,dt}_{\substack{=0 \text{ jer je } t\mapsto te^{-\frac{t^2}{2}} \\ \text{neparna funkcija na} \\ \text{simetričnoj domeni}}} = \mu$$

$$\operatorname{Var} X = \mathbb{E}[(X-\mu)^2] = \int\limits_{-\infty}^{+\infty} (x-\mu)^2 f(x)\,dx = \frac{1}{\sigma\sqrt{2\pi}} \int\limits_{-\infty}^{+\infty} (x-\mu)^2 e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx =$$

$$= \begin{Bmatrix} t = \frac{x-\mu}{\sigma} \\ dt = \frac{dx}{\sigma} \end{Bmatrix} = \frac{\sigma^2}{\sqrt{2\pi}} \int\limits_{-\infty}^{+\infty} t^2 e^{-\frac{t^2}{2}}\,dt = \begin{Bmatrix} u = t & \Rightarrow & du = dt \\ dv = te^{-\frac{t^2}{2}} & \Rightarrow & v = -e^{-\frac{t^2}{2}} \end{Bmatrix} =$$

$$= \frac{\sigma^2}{\sqrt{2\pi}} \cdot \underbrace{(-te^{-\frac{t^2}{2}})\Big|_{-\infty}^{+\infty}}_{=0} + \frac{\sigma^2}{\sqrt{2\pi}} \cdot \underbrace{\int\limits_{-\infty}^{+\infty} e^{-\frac{t^2}{2}}\,dt}_{=\sqrt{2\pi}} = \sigma^2.$$

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

△

Zadatak 2.2 Neka je $X \sim N(\mu, \sigma^2)$. Pokažite da je $\dfrac{X-\mu}{\sigma} \sim N(0,1)$.

Rješenje: Stavimo $Y := \dfrac{X-\mu}{\sigma}$.

$$\Rightarrow \quad F_Y(x) = \mathbb{P}(Y \le x) = P\left(\frac{X-\mu}{\sigma} \le x\right) = \mathbb{P}(X \le \mu + \sigma x) =$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int\limits_{-\infty}^{\mu+\sigma x} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dt = \begin{Bmatrix} s = \frac{t-\mu}{\sigma} \\ ds = \frac{dt}{\sigma} \end{Bmatrix} = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{x} e^{-\frac{s^2}{2}}\,ds$$

$$\Rightarrow \quad Y \sim N(0,1).$$

△

Napomena: Funkciju distribucije jedinične normalne razdiobe označavamo sa

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

Vrijedi $\Phi(-x) = 1 - \Phi(x)$.

Zadatak 2.3 Neka je $X \sim N(\mu, \sigma^2)$. Izračunajte

$$\mathbb{P}(|X - \mu| \le k\sigma), \text{ za } k = 1, 2, 3.$$

Rješenje: Zbog $\dfrac{X-\mu}{\sigma} \sim N(0,1)$ je

$$\begin{aligned}\mathbb{P}(|X - \mu| \le k\sigma) = \mathbb{P}(-k\sigma \le X - \mu \le k\sigma) = \mathbb{P}\left(-k \le \frac{X-\mu}{\sigma} \le k\right) = \\ = \Phi(k) - \Phi(-k) = 2\Phi(k) - 1.\end{aligned}$$

Slijedi

$$\begin{aligned}\mathbb{P}(|X - \mu| \le \sigma) &= 2\Phi(1) - 1 \approx 2 \cdot 0.8413 - 1 = 0.6826,\\ \mathbb{P}(|X - \mu| \le 2\sigma) &= 2\Phi(2) - 1 \approx 2 \cdot 0.9772 - 1 = 0.9544,\\ \mathbb{P}(|X - \mu| \le 3\sigma) &= 2\Phi(3) - 1 \approx 2 \cdot 0.9987 - 1 = 0.9974.\end{aligned}$$

$\triangle$

Zadatak 2.4 Neka je $Y \sim N(1, 4)$. Izračunajte vjerojatnost da jednadžba

$$4x^2 + 4Yx + Y + 2 = 0$$

ima samo realna rješenja.

Rješenje: Zbog $\mu = 1$ i $\sigma = 2$ je $\dfrac{Y-1}{2} \sim N(0,1)$. Jednadžba ima samo realna rješenja ako je diskriminanta

$$\begin{aligned}D = (4Y)^2 - 4 \cdot 4(Y + 2) &\ge 0,\\ Y^2 - Y - 2 &\ge 0,\\ (Y + 1)(Y - 2) &\ge 0.\end{aligned}$$

Tražena vjerojatnost je

$$\begin{aligned}\mathbb{P}(D \ge 0) &= \mathbb{P}((Y + 1)(Y - 2) \ge 0) = \mathbb{P}(\{Y \le -1\} \cup \{Y \ge 2\}) = \{\sigma - \text{aditivnost}\} =\\ &= \mathbb{P}(Y \le -1) + \mathbb{P}(Y \ge 2) = \mathbb{P}\left(\frac{Y-1}{2} \le \frac{-1-1}{2}\right) + \mathbb{P}\left(\frac{Y-1}{2} \ge \frac{2-1}{2}\right) =\\ &= \Phi(-1) + (1 - \Phi(1/2)) = 1 - \Phi(1) + 1 - \Phi(1/2) \approx\\ &\approx 1 - 0.8413 + 1 - 0.6915 = 0.4672.\end{aligned}$$

$\triangle$

2.2 Eksponencijalna distribucija

Definicija: Slučajna varijabla X ima **eksponencijalnu distribuciju** s parametrom $\lambda > 0$ ako joj je funkcija gustoće

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0, & \text{inače} \end{cases} = \lambda e^{-\lambda x}\, \mathbb{1}_{\langle 0,+\infty\rangle}(x).$$

Pišemo: $X \sim \text{Exp}(\lambda)$.

Zadatak 2.5 Neka je $X \sim \text{Exp}(\lambda)$.

(a) Pokažite da je X dobro definirana.

(b) Odredite funkciju distribucije slučajne varijable X.

(c) Izračunajte $\mathbb{E}X, \mathbb{E}[X^2]$ i $\text{Var}\, X$.

Rješenje:

(a) (i) $f(x) \geq 0, \quad \forall x \in \mathbb{R}$.

(ii) $$\int_{-\infty}^{+\infty} f(x)\, dx = \lambda \int_{0}^{+\infty} e^{-\lambda x}\, dx = \lambda \left(-\frac{1}{\lambda}\, e^{-\lambda x} \right) \Big|_{0}^{+\infty} = 1.$$

(b) $$\mathbb{P}(X \leq 0) = \int_{-\infty}^{0} f(x)\, dx = 0 \quad \Rightarrow \quad X > 0 \text{ (g.s.)}$$

- $x \leq 0 \quad \Rightarrow \quad F(x) = \mathbb{P}(\underbrace{X}_{>0} \leq \underbrace{x}_{\leq 0}) = \mathbb{P}(\emptyset) = 0.$

- $x > 0 \quad \Rightarrow \quad F(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f(t)\, dt = \lambda \int_{0}^{x} e^{-\lambda t}\, dt =$

$$= \lambda \left(-\frac{1}{\lambda}\, e^{-\lambda t} \right) \Big|_{0}^{x} = 1 - e^{-\lambda x}.$$

$$\Rightarrow \quad F(x) = \begin{cases} 1 - e^{-\lambda x}, & x > 0 \\ 0, & x \leq 0 \end{cases}.$$

(c)

$$\mathbb{E}X = \int_{-\infty}^{+\infty} x f(x)\,dx = \lambda \int_{0}^{+\infty} x e^{-\lambda x}\,dx = \left\{ \begin{matrix} u = x & \Rightarrow & du = dx \\ dv = e^{-\lambda x}dx & \Rightarrow & v = -\frac{1}{\lambda}e^{-\lambda x} \end{matrix} \right\} =$$

$$= \underbrace{\left(-xe^{-\lambda x}\right)\Big|_0^{+\infty}}_{=0} + \int_0^{+\infty} e^{-\lambda x}\,dx = \frac{1}{\lambda}\cdot \underbrace{\lambda \int_0^{+\infty} e^{-\lambda x}\,dx}_{=1} = \frac{1}{\lambda},$$

$$\mathbb{E}[X^2] = \int_{-\infty}^{+\infty} x^2 f(x)\,dx = \lambda \int_{0}^{+\infty} x^2 e^{-\lambda x}\,dx = \left\{ \begin{matrix} u = x^2 & \Rightarrow & du = 2xdx \\ dv = e^{-\lambda x}dx & \Rightarrow & v = -\frac{1}{\lambda}e^{-\lambda x} \end{matrix} \right\} =$$

$$= \underbrace{\left(-x^2e^{-\lambda x}\right)\Big|_0^{+\infty}}_{=0} + 2\cdot \int_0^{+\infty} x e^{-\lambda x}\,dx = \frac{2}{\lambda}\cdot \underbrace{\lambda \int_0^{+\infty} x e^{-\lambda x}\,dx}_{=\mathbb{E}X} = \frac{2}{\lambda^2},$$

$$\operatorname{Var} X = \mathbb{E}[X^2] - (\mathbb{E}X)^2 = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}.$$

$\triangle$

2.3 Γ-distribucija

Definicija: **Γ-funkcija** je funkcija $\Gamma \colon \langle 0, +\infty\rangle \to \mathbb{R}$ definirana sa

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

Zadatak 2.6 Pokažite sljedeća svojstva Γ-funkcije:

(a) $\Gamma(x+1) = x\Gamma(x), \quad$ za $x > 0$.

(b) $\Gamma(1) = 1, \quad \Gamma(n+1) = n!, \quad$ za $n \in \mathbb{N}$.

(c) $\Gamma(\frac{1}{2}) = \sqrt{\pi}, \quad \Gamma(\frac{3}{2}) = \frac{\sqrt{\pi}}{2}$.

Rješenje:

(a)

$$\Gamma(x+1) = \int_0^\infty t^x e^{-t}\,dt = \left\{\begin{array}{ccc} u = t^x & \Rightarrow & du = xt^{x-1}\,dx \\ dv = e^{-t} & \Rightarrow & v = -e^{-t}\,dt \end{array}\right\} =$$

$$= -t^x e^{-t}\Big|_0^\infty + x\int_0^\infty t^{x-1}e^{-t}\,dt = x\Gamma(x).$$

(b)

$$\Gamma(1) = \int_0^\infty e^{-t}\,dt = -e^{-t}\Big|_0^\infty = 1,$$

$$\Gamma(n+1) \overset{\text{(a)}}{=} n\Gamma(n) \overset{\text{(a)}}{=} n(n-1)\Gamma(n-1) = \cdots = n!\Gamma(1) = n!.$$

(c)

$$\Gamma\left(\frac{1}{2}\right) = \int_0^\infty t^{-\frac{1}{2}}e^{-t}\,dt = \left\{\begin{array}{c} t = \frac{u^2}{2} \\ dt = u\,du \end{array}\right\} = \int_0^\infty \frac{\sqrt{2}}{u}e^{-\frac{u^2}{2}}u\,du =$$

$$= \sqrt{2}\int_0^\infty e^{-\frac{u^2}{2}}\,du = \frac{\sqrt{2}}{2}\underbrace{\int_{-\infty}^{+\infty} e^{-\frac{u^2}{2}}\,du}_{=\sqrt{2\pi}} = \sqrt{\pi},$$

$$\Gamma\left(\frac{3}{2}\right) \overset{\text{(a)}}{=} \frac{1}{2}\Gamma\left(\frac{1}{2}\right) = \frac{\sqrt{\pi}}{2}.$$

△

Definicija: Slučajna varijabla X ima **Γ-distribuciju** s paramerima $\alpha > 0$ i $\beta > 0$ ako joj je funkcija gustoće

$$f(x) = \begin{cases} \dfrac{1}{\Gamma(\alpha)\beta^\alpha}\,x^{\alpha-1}e^{-\frac{x}{\beta}}, & x > 0 \\ 0, & \text{inače} \end{cases} = \frac{1}{\Gamma(\alpha)\beta^\alpha}\,x^{\alpha-1}e^{-\frac{x}{\beta}}\,\mathbb{1}_{\langle 0,\infty\rangle}(x).$$

Pišemo: $X \sim \Gamma(\alpha, \beta)$.

Napomena:

$$X \sim \text{Exp}(\lambda) \;\Rightarrow\; f_X(x) = \lambda e^{-\lambda}\,\mathbb{1}_{\langle 0,\infty\rangle}(x) = \frac{1}{\Gamma(1)\;1/\lambda}e^{-\frac{x}{1/\lambda}}\,\mathbb{1}_{\langle 0,\infty\rangle}(x) \;\Rightarrow\; X \sim \Gamma(1, 1/\lambda).$$

Zadatak 2.7 Neka je $X \sim \Gamma(\alpha, \beta)$.

(a) Pokažite da je X dobro zadana.

(b) Izračunajte $\mathbb{E}X$ i $\operatorname{Var} X$.

Rješenje:

(a) (i) $f(x) \geq 0, \quad \forall x \in \mathbb{R}$.

(ii) $$\int_{-\infty}^{+\infty} f(x)\,dx = \frac{1}{\Gamma(\alpha)\beta^\alpha}\int_0^{+\infty} x^{\alpha-1}e^{-\frac{x}{\beta}}\,dx = \left\{\begin{matrix} t = \frac{x}{\beta} \\ dt = \frac{dx}{\beta}\end{matrix}\right\} = \frac{1}{\Gamma(\alpha)}\underbrace{\int_0^{\infty} t^{\alpha-1}e^{-t}\,dt}_{=\Gamma(\alpha)} = 1.$$

(b)

$$\mathbb{E}X = \int_{-\infty}^{+\infty} xf(x)\,dx = \frac{1}{\Gamma(\alpha)\beta^\alpha}\int_0^{+\infty} x\cdot x^{\alpha-1}e^{-\frac{x}{\beta}}\,dx = \frac{1}{\Gamma(\alpha)\beta^\alpha}\int_0^{+\infty} x^{\alpha}e^{-\frac{x}{\beta}}\,dx =$$

$$= \frac{\Gamma(\alpha+1)\beta^{\alpha+1}}{\Gamma(\alpha)\beta^\alpha}\underbrace{\frac{1}{\Gamma(\alpha+1)\beta^{\alpha+1}}\int_0^{+\infty} x^{(\alpha+1)-1}e^{-\frac{x}{\beta}}\,dx}_{=1} \overset{2.6}{=} \beta\,\frac{\alpha\Gamma(\alpha)}{\Gamma(\alpha)} = \alpha\beta,$$

$$\mathbb{E}[X^2] = \int_{-\infty}^{+\infty} x^2f(x)\,dx = \frac{1}{\Gamma(\alpha)\beta^\alpha}\int_0^{+\infty} x^2\cdot x^{\alpha-1}e^{-\frac{x}{\beta}}\,dx = \frac{1}{\Gamma(\alpha)\beta^\alpha}\int_0^{+\infty} x^{\alpha+1}e^{-\frac{x}{\beta}}\,dx =$$

$$= \frac{\Gamma(\alpha+2)\beta^{\alpha+2}}{\Gamma(\alpha)\beta^\alpha} \overset{2.6}{=} \frac{(\alpha+1)\alpha\Gamma(\alpha)\beta^2}{\Gamma(\alpha)} = \alpha(\alpha+1)\beta^2,$$

$$\operatorname{Var} X = \mathbb{E}[X^2] - (\mathbb{E}X)^2 = \alpha(\alpha+1)\beta^2 - (\alpha\beta)^2 = \alpha\beta^2.$$

$\triangle$

2.4 χ^2-distribucija

Definicija: Slučajna varijabla X ima **χ^2-distribuciju** s n stupnjeva slobode ako je $X \sim \Gamma\left(\frac{n}{2}, 2\right)$. Dakle,

$$f(x) = \begin{cases} \dfrac{1}{\Gamma(\frac{n}{2})\,2^{\frac{n}{2}}}\,x^{\frac{n}{2}-1}e^{-\frac{x}{2}}, & x > 0 \\ 0, & \text{inače} \end{cases} = \frac{1}{\Gamma(\frac{n}{2})\,2^{\frac{n}{2}}}\,x^{\frac{n}{2}-1}e^{-\frac{x}{2}}\,\mathbb{1}_{\langle 0,+\infty\rangle}(x).$$

Pišemo: $X \sim \chi^2(n)$.

Zadatak 2.8 Neka je $X \sim N(0,1)$. Odredite distribuciju slučajne varijable $Y = X^2$.

Rješenje:

- $x < 0 \quad \Rightarrow \quad F_Y(x) = \mathbb{P}(Y \leq x) = \mathbb{P}(\underbrace{X^2}_{\geq 0} \leq \underbrace{x}_{<0}) = \mathbb{P}(\emptyset) = 0.$

- $x \geq 0 \quad \Rightarrow \quad F_Y(x) = \mathbb{P}(Y \leq x) = \mathbb{P}(X^2 \leq x) = \mathbb{P}(|X| \leq \sqrt{x}) =$

$$= \mathbb{P}(-\sqrt{x} \leq X \leq \sqrt{x}) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\sqrt{x}}^{\sqrt{x}} e^{-\frac{t^2}{2}}\, dt =$$

$$= \frac{2}{\sqrt{2\pi}} \int\limits_0^{\sqrt{x}} e^{-\frac{t^2}{2}}\, dt = \begin{Bmatrix} t = \sqrt{u} \\ dt = \frac{du}{2\sqrt{u}} \end{Bmatrix} = \frac{2}{\sqrt{2\pi}} \int\limits_0^x e^{-\frac{u}{2}} \frac{du}{2\sqrt{u}} =$$

$$= \int\limits_0^x \frac{1}{\Gamma(\frac{1}{2})2^{\frac{1}{2}}}\, u^{\frac{1}{2}-1} e^{-\frac{u}{2}}\, du = \int\limits_{-\infty}^x \frac{1}{\Gamma(\frac{1}{2})2^{\frac{1}{2}}}\, u^{\frac{1}{2}-1} e^{-\frac{u}{2}}\, \mathbb{1}_{\langle 0,+\infty\rangle}(u)\, du.$$

$\Rightarrow \quad Y \sim \chi^2(1).$

△

2.5 Studentova t-distribucija

Definicija: Slučajna varijabla ima **(Studentovu) t-distribuciju** s n stupnjeva slobode ako joj je gustoća

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

Pišemo: $X \sim t(n)$.

Napomena: Slučajna varijabla X ima **Cauchyevu distribuciju** ako je $X \sim t(1)$. Dakle,

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

Zadatak 2.9 Neka je $X \sim t(n)$. Odredite $\mathbb{E}X$ i $\operatorname{Var} X$ kada postoje.

Rješenje:

- $n = 1 \quad \Rightarrow \quad \displaystyle\int\limits_{-\infty}^{+\infty} |x| f(x)\, dx = \frac{2}{\pi} \int\limits_0^{+\infty} \frac{x}{1+x^2}\, dx = \frac{2}{\pi} \ln(1+x^2)\Big|_0^{+\infty} = +\infty$

  $\Rightarrow \quad X$ nema očekivanje ni varijancu.

- $n \geq 2 \quad \Rightarrow \quad \mathbb{E}X = \int\limits_{-\infty}^{+\infty} x f(x)\, dx = \frac{1}{\sqrt{n\pi}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \underbrace{\int\limits_{-\infty}^{+\infty} \frac{x}{\left(1+\frac{x^2}{n}\right)^{\frac{n+1}{2}}}\, dx}_{\substack{=0 \text{ jer je podintegralna} \\ \text{funkcija neparna na} \\ \text{simetričnoj domeni}}} = 0$

$$\Rightarrow \quad \operatorname{Var} X = \mathbb{E}[X^2] = \int\limits_{-\infty}^{+\infty} x^2 f(x)\, dx = \frac{1}{\sqrt{n\pi}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \underbrace{\int\limits_{-\infty}^{+\infty} \frac{x^2}{\left(1+\frac{x^2}{n}\right)^{\frac{n+1}{2}}}\, dx}_{\substack{\text{divergira za } n=2, \\ \text{konvergira za } n>2}} =$$

$$\overset{n\geq 2}{=} \frac{1}{\sqrt{n\pi}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \int\limits_{-\infty}^{+\infty} n \cdot \frac{\left(1+\frac{x^2}{n}\right) - 1}{\left(1+\frac{x^2}{n}\right)^{\frac{n+1}{2}}}\, dx = I_1 - I_2, \text{ gdje je}$$

$$I_1 = n \frac{1}{\sqrt{n\pi}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \int\limits_{-\infty}^{+\infty} \frac{1}{\left(1+\frac{x^2}{n}\right)^{\frac{n+1}{2}-1}}\, dx = \left\{ \begin{array}{c} \frac{t}{\sqrt{n-2}} = \frac{x}{\sqrt{n}} \\ \frac{dt}{\sqrt{n-2}} = \frac{dx}{\sqrt{n}} \end{array} \right\} =$$

$$= n \frac{1}{\sqrt{(n-2)\pi}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \int\limits_{-\infty}^{+\infty} \frac{1}{\left(1+\frac{t^2}{n-2}\right)^{\frac{n-1}{2}}}\, dt \overset{2.6}{=}$$

$$= n \frac{\frac{n-1}{2}}{\frac{n-2}{2}} \underbrace{\frac{1}{\sqrt{(n-2)\pi}} \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n-2}{2}\right)} \int\limits_{-\infty}^{+\infty} \frac{1}{\left(1+\frac{t^2}{n-2}\right)^{\frac{(n-2)+1}{2}}}\, dt}_{=1} = \frac{n(n-1)}{n-2},$$

$$I_2 = n \underbrace{\frac{1}{\sqrt{n\pi}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \int\limits_{-\infty}^{+\infty} \frac{1}{\left(1+\frac{x^2}{n}\right)^{\frac{n+1}{2}}}\, dx}_{=1} = n.$$

$$\Rightarrow \quad \operatorname{Var} X = \frac{n(n-1)}{n-2} - n = \frac{n}{n-2}$$

$\Rightarrow$ X ima očekivanje za $n \geq 2$ i varijancu za $n \geq 3$.

2.6 Fisherova F-distribucija

Definicija: Slučajna varijabla X ima **(Fisherovu) F-distribuciju** s parametrima $m \in \mathbb{N}$ i $n \in \mathbb{N}$ ako joj je funkcija gustoće

$$f(x) = \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} m^{\frac{m}{2}} n^{\frac{n}{2}} \frac{x^{\frac{m}{2}-1}}{(mx+n)^{\frac{m+n}{2}}} \mathbb{1}_{\langle 0,+\infty\rangle}(x).$$

Pišemo: $X \sim F(m,n)$.

Zadatak 2.10 Neka je $X \sim F(m,n)$. Odredite distribuciju slučajne varijable $Y = \dfrac{1}{X}$.

Rješenje: Y je dobro definirana jer je $X > 0$ (g.s.).

- $x \leq 0 \quad \Rightarrow \quad F_Y(x) = \mathbb{P}(Y \leq x) = \mathbb{P}\left(\underbrace{\frac{1}{X}}_{>0} \leq \underbrace{x}_{\leq 0}\right) = \mathbb{P}(\emptyset) = 0.$

- $x > 0 \quad \Rightarrow \quad F_Y(x) = \mathbb{P}(Y \leq x) = \mathbb{P}\left(\frac{1}{X} \leq x\right) = \left(X \geq \frac{1}{x}\right) =$

$$= \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} m^{\frac{m}{2}} n^{\frac{n}{2}} \int_{\frac{1}{x}}^{+\infty} \frac{u^{\frac{m}{2}-1}}{(mu+n)^{\frac{m+n}{2}}}\, du = \left\{\begin{matrix} u = \frac{1}{v} \\ du = -\frac{dv}{v^2} \end{matrix}\right\} =$$

$$= \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} m^{\frac{m}{2}} n^{\frac{n}{2}} \int_{0}^{x} \frac{v^{-\frac{m}{2}+1}}{\left(\frac{m}{v}+n\right)^{\frac{m+n}{2}}} \frac{v^{\frac{m+n}{2}}}{v^{\frac{m+n}{2}}} \frac{dv}{v^2} =$$

$$= \frac{\Gamma\left(\frac{n+m}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{m}{2}\right)} n^{\frac{n}{2}} m^{\frac{m}{2}} \int_{-\infty}^{x} \frac{v^{\frac{n}{2}-1}}{(nv+m)^{\frac{n+m}{2}}} \mathbb{1}_{\langle 0,+\infty\rangle}(v)\, dv.$$

$\Rightarrow \quad Y \sim F(n,m)$.

△

Zadatak 2.11 Na slučajan način biramo točku iz pravokutnika stranica duljine 3 cm i 4 cm. Izračunajte očekivanu udaljenost točke od najbliže stranice pravokutnika.

Rješenje: Označimo s X udaljenost točke od najbliže stranice pravokutnika. Tada je za $x \in [0, 3/2]$

$$F_X(x) = \mathbb{P}(X \leq x) = 1 - \mathbb{P}(X > x) = 1 - \frac{(4-2x)(3-2x)}{4 \cdot 3} = -\frac{x^2}{3} + \frac{7x}{6}.$$

$$\Rightarrow \quad F_X(x) = \begin{cases} 0, & x < 0 \\ -\frac{x^2}{3} + \frac{7x}{6}, & 0 \leq x \leq 3/2 \\ 1, & x > 3/2 \end{cases}$$

$$\Rightarrow \quad f_X(x) = F_X'(x) = \left(-\frac{2x}{3} + \frac{7}{6}\right) 1_{[0,3/2]}(x)$$

$$\Rightarrow \quad \mathbb{E}X = \int_0^{3/2} x\left(-\frac{2x}{3} + \frac{7}{6}\right) dx = \frac{27}{48} = 0.5625$$

△

2.7 Zadaci za vježbu

2.12 Slučajna varijabla X ima **uniformnu distribuciju** s parametrima $a, b \in \mathbb{R}$ ako joj je funkcija gustoće

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

Pišemo $X \sim U(a,b)$. Odredite funkciju distribucije i izračunajte $\mathbb{P}\left(a \leq X \leq \frac{a+b}{2}\right)$.

2.13 Neka je $X \sim N(\mu, \sigma^2)$. Izračunajte treći i četvrti moment od X.

2.14 Neka je $X \sim N(3,4)$. Odredite $c \in \mathbb{R}$ t. d. je $\mathbb{P}(X > c) = 2\mathbb{P}(X \leq c)$.

2.15 Neka je $X \sim \text{Exp}(\lambda)$. Izračunajte $\mathbb{P}\left(X \leq \frac{1}{\lambda}\right)$.

2.16 Neka je $X \sim \Gamma(\alpha, \beta)$. Izračunajte treći i četvrti moment od X.

2.17 Neka X ima Cauchyevu distribuciju. Odredite funkciju distribucije.

2.18 Neka je X slučajna varijabla s funkcijom gustoće

$$f(x) = \frac{1}{2}e^{-|x|}, \quad x \in \mathbb{R}.$$

Odredite funkciju distribucije te izračunajte $\mathbb{P}(X > 0)$ i $\mathbb{P}(1 < X < 2)$.

2.19 **Beta-funkcija** je funkcija $B\colon \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$ definirana sa

$$B(p,q) := \int_0^1 x^{p-1}(1-x)^{q-1}\,dx.$$

Dokažite da je $B(p,q) = \dfrac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}$.

2.20 Slučajna varijabla X ima **beta-distribuciju** s parametrima $p > 0$ i $q > 0$ ako joj je funkcija gustoće

$$f(x) = \begin{cases} \dfrac{1}{B(p,q)} x^{p-1}(1-x)^{q-1}, & x \in \langle 0,1 \rangle \\ 0, & x \notin \langle 0,1 \rangle \end{cases}.$$

Pišemo $X \sim B(p,q)$. Izračunajte $\mathbb{E}X, \operatorname{Var} X$ i $\mathbb{E}[X^n], n \in \mathbb{N}$.

2.21 Neka je $X \sim U(a,b)$. Odredite funkciju distribucije slučajne varijable Y ako je:

(a) $Y = \alpha X + \beta, \alpha \neq 0$,

(b) $Y = e^X$.

2.22 Neka je $X \sim U(-1, 1)$. Odredite funkciju distribucije slučajne varijable $Y = X^2$.

2.23 Neka je $X \sim N(\mu, \sigma^2)$. Odredite funkciju distribucije slučajne varijable $Y = \alpha X + \beta$.

2.24 Neka je $X \sim N(0, 1)$. Odredite funkciju distribucije slučajne varijable Y ako je:

(a) $Y = 1/X$,

(b) $Y = X^2$,

(c) $Y = X^3$.

2.25 Neka je $X \sim \text{Exp}(\lambda)$. Odredite funkciju distribucije slučajne varijable Y ako je:

(a) $Y = \alpha X + \beta$,

(b) $Y = X^3$.

2.26 Neka je $X \sim \text{Exp}(1)$. Odredite funkciju distribucije slučajne varijable $Y = \sqrt{X}$.

2.27 Ako je $X \sim U(0, 1)$, kojom transformacijom od X se dobije $Y \sim \text{Exp}(\lambda)$?

2.28 Ako je $X \sim \text{Exp}(\lambda)$, kojom transformacijom od X se dobije Y sa Cauchyevom distribucijom?

2.29 Slučajna varijabla X ima funkciju distribucije

$$F(x) = \begin{cases} e^{-x^{-\alpha}}, & x > 0 \\ 0, & x \le 0 \end{cases},$$

gdje je $\alpha > 0$. Odredite funkciju distribucije slučajne varijable $Y = -1/X$.

Rješenja: **2.12.** $F_X(x) = \begin{cases} 0, & x < a \\ \frac{x-a}{b-a}, & a \le x \le b, \\ 1, & x > b \end{cases}$ $\mathbb{P}(a \le X \le \frac{a+b}{2}) = 1/2$ **2.13.** $\mathbb{E}[X^3] = \mu^3 + 3\mu\sigma^2, \mathbb{E}[X^4] = 3\sigma^4 + \mu^4 + 6\mu^2\sigma^2$ **2.14.** $c \approx 2.14$ **2.15.** $1 - e^{-1}$ **2.16.** $\mathbb{E}[X^3] = \alpha(\alpha+1)(\alpha+2)\beta^3, \mathbb{E}[X^4] = \alpha(\alpha+1)(\alpha+2)(\alpha+3)\beta^4$ **2.17.** $F(x) = \frac{1}{\pi}\operatorname{arctg} x + \frac{1}{2}$ **2.18.** $F(x) = \begin{cases} 1 - e^{-x}/2, & x \ge 0 \\ e^x/2, & x < 0 \end{cases}, \mathbb{P}(X > 0) = 1/2, \mathbb{P}(1 < X < 2) = (e^{-1} - e^{-2})/2$ **2.20.** $\mathbb{E}X = \frac{p}{p+q}, \operatorname{Var} X = \frac{pq}{(p+q)^2(p+q+1)}, \mathbb{E}[X^n] = \frac{B(p+n,q)}{B(p,q)}$ **2.21.** (a) Za $\alpha > 0$ je $F_Y(x) = F_X(\frac{x-\beta}{\alpha})$, za $\alpha < 0$ je $F_Y(x) = 1 - F_X(\frac{x-\beta}{\alpha})$ (b) $F_Y(x) = \begin{cases} F_X(\ln y), & y > 0 \\ 0, & y \le 0 \end{cases}$ **2.22.** $F_Y(x) = \begin{cases} 0, & x < 0 \\ \sqrt{x}, & 0 \le x \le 1, \\ 1, & x > 1 \end{cases}$ **2.23.** $Y \sim N(\alpha\mu + \beta, \alpha^2\sigma^2)$ **2.24.** (a) $F_Y(x) = \begin{cases} 3/2 - F_X(1/y), & x > 0 \\ 1/2, & x = 0 \\ 1/2 + F_X(1/y), & x < 0 \end{cases}$ (b) $F_Y(x) = \begin{cases} F_X(\sqrt{x}) - F_X(-\sqrt{x}), & x > 0 \\ 0, & x \le 0 \end{cases}$ (c) $F_Y(x) = F_X(\sqrt[3]{x})$ **2.25.** (a) Za $\alpha > 0$ je $F_Y(x) = F_X(\frac{x-\beta}{\alpha})$, za $\alpha < 0$ je $F_Y(x) = 1 - F_X(\frac{x-\beta}{\alpha})$ (b) $F_Y(x) = \begin{cases} 1 - e^{-\lambda\sqrt[3]{x}}, & x > 0 \\ 0, & x \le 0 \end{cases}$ **2.26.** $F_Y(x) = \begin{cases} 1 - e^{-x^2}, & x > 0 \\ 0, & x \le 0 \end{cases}$ **2.27.** $Y = -\frac{1}{\lambda}\ln(1 - X)$ **2.28.** $Y = \operatorname{tg}[\pi(\frac{1}{2} - e^{-\lambda X})]$ **2.29.** $F_Y(x) = \begin{cases} 1, & x \ge 0 \\ e^{-(-x)^\alpha}, & x < 0 \end{cases}$

3

Neprekidni slučajni vektori. Uvjetno očekivanje

3.1 Neprekidni slučajni vektori

Definicija: Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ vjerojatnosni prostor. Funkcija $(X, Y)\colon \Omega \to \mathbb{R}^2$ je **slučajni vektor** ako za sve $a, b, c, d \in \mathbb{R}$, $a < b, c < d$ vrijedi

$$\{a \le X \le b, c \le Y \le d\} \in \mathcal{F}.$$

Funkciju $F = F_{X,Y}\colon \mathbb{R}^2 \to [0, 1]$ definiranu sa

$$F(a, b) := \mathbb{P}(X \le a, Y \le b), \quad a, b \in \mathbb{R}$$

zovemo **funkcija distribucije** slučajnog vektora (X, Y).

Definicija: Slučajni vektor (X, Y) je **neprekidni slučajni vektor** ako postoji nenegativna (izmjeriva) funkcija $f\colon \mathbb{R}^2 \to \mathbb{R}$ takva da je

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

Funkciju f zovemo **(vjerojatnosna) funkcija gustoće** slučajnog vektora (X, Y).

Napomena:

(a) $\mathbb{P}(X = a, Y = b) = \int\limits_a^a \int\limits_b^b f(x, y)\, dx\, dy = 0$

(b) Ako je $g\colon \mathbb{R}^2 \to \mathbb{R}$ (izmjeriva) funkcija, onda je

$$\mathbb{E}[g(X, Y)] := \iint\limits_{\mathbb{R}^2} g(x, y) f(x, y)\, dx\, dy.$$

(c) Za $C \subset \mathbb{R}^2$ izmjeriv, uz $g(x,y) := \mathbb{1}_C(x,y)$ je $\mathbb{P}((X,Y) \in C) = \iint\limits_C f(x,y)\,dx\,dy$.

(d) $\mathbb{P}(a \le X \le b) = \mathbb{P}(a \le X \le b, Y \in \mathbb{R}) = \int\limits_a^b \int\limits_{-\infty}^{+\infty} f(x,y)\,dx\,dy = \int\limits_a^b \left(\int\limits_{-\infty}^{+\infty} f(x,y)\,dy \right) dx$

$\Rightarrow X$ je neprekidna slučajna varijabla i $f_X(x) = \int\limits_{-\infty}^{+\infty} f(x,y)\,dy$.

Analogno, Y je neprekidna slučajna varijabla i $f_Y(y) = \int\limits_{-\infty}^{+\infty} f(x,y)\,dx$.

X i Y zovemo **marginalne distribucije** slučajnog vektora (X,Y).

Zadatak 3.1 Funkcija gustoće slučajnog vektora (X,Y) je dana s

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

(a) Odredite marginalne distribucije.

(b) Odredite funkciju gustoće slučajne varijable $\dfrac{X}{Y}$.

Rješenje:

(a) Vrijedi

$$f_X(x) = \int\limits_{-\infty}^{+\infty} f(x,y)\,dy = f_X(x) = \int\limits_{-\infty}^{+\infty} e^{-(x+y)}\, \mathbb{1}_{\langle 0,\infty\rangle \times \langle 0,\infty\rangle}(x,y)\,dy =$$

$$= \begin{cases} \int\limits_0^{+\infty} e^{-(x+y)}\,dy, & x > 0 \\ 0, & x \le 0 \end{cases} = \begin{cases} e^{-x}, & x > 0 \\ 0, & x \le 0 \end{cases} \quad \Rightarrow \quad X \sim \text{Exp}(1).$$

Analogno, $Y \sim \text{Exp}(1)$.

(b) • $a \le 0 \quad \Rightarrow \quad F_{\frac{X}{Y}}(a) = \mathbb{P}\left(\underbrace{\frac{X}{Y}}_{>0} \le \underbrace{a}_{\le 0} \right) = \mathbb{P}(\emptyset) = 0.$

• $a > 0 \quad \Rightarrow$

$$
\begin{aligned}
F_{\frac{X}{Y}}(a) &= \mathbb{P}\left(\frac{X}{Y} \le a\right) = \iint\limits_{\frac{x}{y}\le a} f(x,y)\,dx\,dy = \iint\limits_{\frac{x}{y}\le a,\,x>0,\,y>0} e^{-(x+y)}\,dx\,dy = \\
&= \int\limits_0^\infty \int\limits_0^{ay} e^{-(x+y)}\,dx\,dy = \int\limits_0^\infty e^{-y}\left(\int\limits_0^{ay} e^{-x}\,dx\right) dy = \\
&= \int\limits_0^\infty e^{-y}\left(-e^{-x}\Big|_0^{ay}\right) dy = \int\limits_0^\infty e^{-y}\left(1-e^{-ay}\right) dy = \\
&= \left(-e^{-y} + \frac{e^{-(a+1)y}}{a+1}\right)\Bigg|_0^\infty = 1 - \frac{1}{a+1}
\end{aligned}
$$

$$
\Rightarrow \quad f_{\frac{X}{Y}}(a) = F'_{\frac{X}{Y}} = \begin{cases} \dfrac{1}{(a+1)^2}, & a > 0 \\ 0, & a \le 0 \end{cases}.
$$

△

Zadatak 3.2 Na slučajan način biramo točku iz kruga radijusa R. Ako točku označimo s (X,Y), onda je (X,Y) neprekidni slučajni vektor s uniformnom distribucijom na krugu radijusa R, tj.

$$
f(x,y) = \begin{cases} c, & x^2+y^2 \le R \\ 0, & x^2+y^2 > R \end{cases}.
$$

(a) Odredite c.

(b) Odredite marginalne distribucije X i Y.

(c) Izračunajte vjerojatnost da je udaljenost D točke (X,Y) od ishodišta manja ili jednaka a.

(d) Izračunajte $\mathbb{E}D$.

Rješenje:

(a) $1 = \iint\limits_{\mathbb{R}^2} f(x,y)\,dx\,dy = \iint\limits_{x^2+y^2\le R^2} c\,dx\,dy = c\cdot \iint\limits_{x^2+y^2\le R^2} dx\,dy = c\cdot R^2\pi \quad \Rightarrow \quad c = \dfrac{1}{R^2\pi}.$

(b) Vrijedi

$$f_X(x) = \int\limits_{-\infty}^{+\infty} f(x,y)\,dy = \frac{1}{R^2\pi} \int\limits_{x^2+y^2\le R^2} dy = \begin{cases} \dfrac{1}{R^2\pi} \displaystyle\int\limits_{-\sqrt{R^2-x^2}}^{\sqrt{R^2-x^2}} dy, & |x| \le R \\ 0, & |x| > R \end{cases}$$

$$= \begin{cases} \dfrac{2\sqrt{R^2-x^2}}{R^2\pi} & |x| \le R \\ 0, & |x| > R \end{cases}.$$

Analogno, $f_Y(y) = \begin{cases} \dfrac{2\sqrt{R^2-y^2}}{R^2\pi} & |y| \le R \\ 0, & |y| > R \end{cases}.$

(c) Tražena vjerojatnost je $F_D(a)$, gdje je $D = \sqrt{X^2+Y^2}$. Vrijedi $0 \le D \le R$.

- $a < 0 \quad \Rightarrow \quad F_D(a) = \mathbb{P}(\underbrace{D}_{\ge 0} \le \underbrace{a}_{<0}) = \mathbb{P}(\emptyset) = 0$
- $a > R \quad \Rightarrow \quad F_D(a) = \mathbb{P}(\underbrace{D}_{\le R} \le \underbrace{a}_{>R}) = \mathbb{P}(\Omega) = 1$
- $0 \le a \le R$

$$\Rightarrow \quad F_D(a) = \mathbb{P}(D \le a) = \mathbb{P}(\sqrt{X^2+Y^2} \le a) = \mathbb{P}(X^2+Y^2 \le a^2) =$$
$$= \iint\limits_{x^2+y^2\le a^2} f(x,y)\,dx\,dy = \frac{1}{R^2\pi} \iint\limits_{x^2+y^2\le a^2} dx\,dy = \frac{a^2\pi}{R^2\pi} = \frac{a^2}{R^2}.$$

(d) Dakle,

$$f_D(a) = \begin{cases} \dfrac{2a}{R^2}, & 0 \le a \le R \\ 0, & \text{inače} \end{cases} \quad \Rightarrow \quad \mathbb{E}D = \frac{2}{R^2} \int\limits_0^R a^2\,da = \frac{2R}{3}.$$

△

Zadatak 3.3 Neka je (X, Y) neprekidni slučajni vektor s funkcijom gustoće f. Odredite funkciju gustoće slučajne varijable:

(a) $X + Y$

(b) $X \cdot Y$

Rješenje:

(a)

$$
\begin{aligned}
F_{X+Y}(a) = \mathbb{P}(X+Y \le a) &= \iint\limits_{x+y\le a} f(x,y)\,dx\,dy = \\
&= \int_{-\infty}^{+\infty} \left(\int_{-\infty}^{a-x} f(x,y)\,dy \right) dx = \begin{Bmatrix} z = x+y \\ dz = dy \end{Bmatrix} = \\
&= \int_{-\infty}^{+\infty} \left(\int_{-\infty}^{a} f(x,z-x)\,dz \right) dx = \int_{-\infty}^{a} \left(\int_{-\infty}^{+\infty} f(x,z-x)\,dx \right) dz
\end{aligned}
$$

$$
\Rightarrow \quad f_{X+Y}(z) = \int_{-\infty}^{+\infty} f(x,z-x)\,dx = \begin{Bmatrix} y = z-x \\ dy = -dx \end{Bmatrix} = \int_{-\infty}^{+\infty} f(z-y,y)\,dy.
$$

(b)

$$
\begin{aligned}
F_{X\cdot Y}(a) = \mathbb{P}(X\cdot Y \le a) &= \iint\limits_{x\cdot y\le a} f(x,y)\,dx\,dy = \\
&= \int_{-\infty}^{0} \left(\int_{\frac{a}{x}}^{+\infty} f(x,y)\,dy \right) dx + \int_{0}^{+\infty} \left(\int_{-\infty}^{\frac{a}{x}} f(x,y)\,dy \right) dx = \begin{Bmatrix} z = x\cdot y \\ dz = x\,dy \end{Bmatrix} = \\
&= \int_{-\infty}^{0} \left(\int_{a}^{-\infty} f\left(x,\frac{z}{x}\right) \frac{dz}{x} \right) dx + \int_{0}^{+\infty} \left(\int_{-\infty}^{a} f\left(x,\frac{z}{x}\right) \frac{dz}{x} \right) dx = \\
&= \int_{-\infty}^{0} \left(\int_{-\infty}^{a} f\left(x,\frac{z}{x}\right) \frac{dz}{|x|} \right) dx + \int_{0}^{+\infty} \left(\int_{-\infty}^{a} f\left(x,\frac{z}{x}\right) \frac{dz}{|x|} \right) dx = \\
&= \int_{-\infty}^{+\infty} \left(\int_{-\infty}^{a} f\left(x,\frac{z}{x}\right) \frac{dz}{|x|} \right) dx = \int_{-\infty}^{a} \left(\int_{-\infty}^{+\infty} f\left(x,\frac{z}{x}\right) \frac{dx}{|x|} \right) dz
\end{aligned}
$$

$$
\Rightarrow \quad f_{X\cdot Y} = \int_{-\infty}^{+\infty} f\left(x,\frac{z}{x}\right) \frac{dx}{|x|}.
$$

$\triangle$

Napomena: Ako je (X,Y) slučajni vektor s funkcijom gustoće $f_{X,Y}$ i marginalnim gustoćama f_X i f_Y, onda vrijedi

$$
X \text{ i } Y \text{ su nezavisne} \quad \Leftrightarrow \quad f_{X,Y}(x,y) = f_X(x)f_Y(y), \forall x,y \in \mathbb{R}.
$$

Zadatak 3.4 Neka su $X \sim \Gamma(\alpha_1, \beta)$ i $Y \sim \Gamma(\alpha_2, \beta)$ nezavisne slučajne varijable. Pokažite da je tada $X + Y \sim \Gamma(\alpha_1 + \alpha_2, \beta)$.

Rješenje: Zbog nezavisnosti slučajnih varijabli X i Y vrijedi

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

Slijedi

$$f_{X+Y}(z) = \int_{-\infty}^{+\infty} f_{X,Y}(z-y, y)\, dy = \int_{-\infty}^{+\infty} f_X(z-y) f_Y(y)\, dy =$$

$$= \frac{1}{\Gamma(\alpha_1)\beta^{\alpha_1}} \cdot \frac{1}{\Gamma(\alpha_2)\beta^{\alpha_2}} \int_{-\infty}^{+\infty} e^{-\frac{z-y}{\beta}} (z-y)^{\alpha_1 - 1}\, \mathbb{1}_{\langle 0,\infty\rangle}(z-y) e^{-\frac{y}{\beta}} y^{\alpha_2 - 1}\, \mathbb{1}_{\langle 0,\infty\rangle}(y)\, dy$$

- $z \leq 0 \quad \Rightarrow \quad f_{X+Y}(z) = 0.$
- $z > 0 \quad \Rightarrow$

$$f_{X+Y}(z) = \frac{1}{\Gamma(\alpha_1)\Gamma(\alpha_2)\beta^{\alpha_1+\alpha_2}} e^{-\frac{z}{\beta}} \int_0^z (z-y)^{\alpha_1-1} y^{\alpha_2-1}\, dy = \left\{ \begin{array}{c} x = \frac{y}{z} \\ dx = \frac{dy}{z} \end{array} \right\} =$$

$$= \frac{1}{\Gamma(\alpha_1)\Gamma(\alpha_2)\beta^{\alpha_1+\alpha_2}} e^{-\frac{z}{\beta}} z^{\alpha_1+\alpha_2-1} \int_0^1 (1-x)^{\alpha_1-1} x^{\alpha_2-1}\, dx = C \cdot e^{-\frac{z}{\beta}} z^{\alpha_1+\alpha_2-1}$$

Zbog toga što je f_{X+Y} funkcija gustoće, vrijedi

$$1 = \int_{-\infty}^{+\infty} f_{X+Y}(z)\, dz = C \int_0^{+\infty} e^{-\frac{z}{\beta}} z^{\alpha_1+\alpha_2-1}\, dz = \left\{ \begin{array}{c} w = \frac{z}{\beta} \\ dw = \frac{dz}{\beta} \end{array} \right\} =$$

$$= C\beta^{\alpha_1+\alpha_2-1} \int_0^{+\infty} w^{\alpha_1+\alpha_2-1} e^{-w}\, dw = C\beta^{\alpha_1+\alpha_2}\Gamma(\alpha_1+\alpha_2)$$

$$\Rightarrow \quad C = \frac{1}{\Gamma(\alpha_1+\alpha_2)\beta^{\alpha_1+\alpha_2}}$$

$\Rightarrow \quad X + Y \sim \Gamma(\alpha_1 + \alpha_2, \beta)$

△

Napomena: Ako su $X_i \sim \Gamma(\alpha_i, \beta), i = 1, \ldots, n$, nezavisne slučajne varijable, onda je

$$X_1 + \cdots + X_n \sim \Gamma(\alpha_1 + \cdots + \alpha_n, \beta).$$

Ako su $X_i \sim \chi^2(\alpha_i), i = 1, \ldots, n$, nezavisne slučajne varijable, onda je

$$X_1 + \cdots + X_n \sim \chi^2(\alpha_1 + \cdots + \alpha_n).$$

Ako su $X_i \sim \text{Exp}(\lambda), i = 1, \ldots, n$, nezavisne slučajne varijable, onda je

$$X_1 + \cdots + X_n \sim \Gamma\left(n, \frac{1}{\lambda}\right).$$

Ako su $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ nezavisne slučajne varijable, onda su

$\frac{X_1 - \mu}{\sigma}, \ldots, \frac{X_n - \mu}{\sigma} \sim N(0,1)$ nezavisne slučajne varijable pa su
$\left(\frac{X_1 - \mu}{\sigma}\right)^2, \ldots, \left(\frac{X_n - \mu}{\sigma}\right)^2 \sim \chi^2(1)$ nezavisne slučajne varijable, odakle je

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

Zadatak 3.5 Neka su $X \sim \chi^2(n)$ i $Y \sim \text{N}(0,1)$ nezavisne slučajne varijable i neka je

$$Z = \frac{Y}{\sqrt{\frac{X}{n}}}.$$

Dokažite da je $Z \sim \text{t}(n)$.

Rješenje: Prvo odredimo funkciju gustoće slučajne varijable $\sqrt{\frac{X}{n}}$.

- $x < 0 \quad \Rightarrow \quad F_{\sqrt{\frac{X}{n}}}(x) = \mathbb{P}\left(\sqrt{\frac{X}{n}} \leq x\right) = \mathbb{P}(\emptyset) = 0.$

- $x \geq 0 \quad \Rightarrow$

$$F_{\sqrt{\frac{X}{n}}}(x) = \mathbb{P}(X \leq nx^2) = \int_0^{nx^2} f_X(t)\, dt = \left\{ \begin{matrix} t = ny^2 \\ dt = 2ny\, dy \end{matrix} \right\} = \int_0^x f_X(ny^2) 2ny\, dy$$

$\Rightarrow \quad f_{\sqrt{\frac{X}{n}}}(x) = 2nx f_X(nx^2)\, \mathbb{1}_{\langle 0,\infty\rangle}(x).$

Zbog nezavisnosti slučajnih varijabli X i Y slijedi

$$\begin{aligned}
f_Z(z) &= \int\limits_{-\infty}^{\infty} f_{\sqrt{\frac{X}{n}},Y}(x, zx)|x|\,dx = \int\limits_{-\infty}^{\infty} f_{\sqrt{\frac{X}{n}}}(x) f_Y(zx)|x|\,dx \\
&= \int\limits_{0}^{\infty} 2nx f_X(nx^2) f_Y(zx) x\,dx = \frac{2^{-\frac{n}{2}+1} n^{\frac{n}{2}}}{\sqrt{2\pi}\Gamma(\frac{n}{2})} \int\limits_{0}^{\infty} x^n e^{-\frac{n}{2}\left(1+\frac{z^2}{n}\right)x^2}\,dx = \\
&= \left\{ \begin{array}{ll} u = \frac{n}{2}\left(1+\frac{z^2}{n}\right)x^2 \;\Rightarrow & x = \left(\frac{2u}{n\left(1+\frac{z^2}{n}\right)}\right)^{1/2} \\ & dx = \frac{\sqrt{2}}{\sqrt{n\left(1+\frac{z^2}{n}\right)}} \frac{du}{2\sqrt{u}} \end{array} \right\} \\
&= \frac{2^{-\frac{n}{2}+1} n^{\frac{n}{2}}}{\sqrt{2\pi}\Gamma(\frac{n}{2})} \int\limits_{0}^{\infty} \frac{2^{\frac{n}{2}} u^{\frac{n}{2}}}{n^{\frac{n}{2}} \left(1+\frac{z^2}{n}\right)^{\frac{n}{2}}} e^{-u} \frac{\sqrt{2}}{\sqrt{n\left(1+\frac{z^2}{n}\right)}} \frac{du}{2\sqrt{u}} = \\
&= \frac{1}{\sqrt{n\pi}\Gamma(\frac{n}{2})} \frac{1}{\left(1+\frac{z^2}{n}\right)^{\frac{n+1}{2}}} \int\limits_{0}^{\infty} u^{\frac{n+1}{2}-1} e^{-u}\,du = \frac{1}{\sqrt{n\pi}} \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \frac{1}{\left(1+\frac{z^2}{n}\right)^{\frac{n+1}{2}}}
\end{aligned}$$

$\Rightarrow \quad Z \sim t(n)$

$\triangle$

Zadatak 3.6 Neka su $X \sim \chi^2(n)$ i $Y \sim \chi^2(m)$ nezavisne slučajne varijable te neka je

$$Z = \frac{\frac{Y}{m}}{\frac{X}{n}}.$$

Dokažite da je $Z \sim \mathrm{F}(m, n)$.

Rješenje: Prvo odredimo funkciju gustoće slučajne varijable $\frac{Y}{m}$.

$$F_{\frac{Y}{m}}(y) = \mathbb{P}\left(\frac{Y}{m} \le y\right) = \mathbb{P}(Y \le my) = \int\limits_{0}^{my} f_Y(t)\,dt = \left\{ \begin{array}{c} t = mu \\ dt = m\,du \end{array} \right\} = \int\limits_{0}^{y} f_Y(mu)\,m\,du$$

$\Rightarrow \quad f_{\frac{Y}{m}}(y) = m f_Y(my)$

Analogno je $f_{\frac{X}{n}}(x) = n f_X(nx)$.

Sada je zbog nezavisnosti

$$f_Z(z) = \int\limits_{-\infty}^{\infty} f_{\frac{X}{n},\frac{Y}{m}}(x, zx)|x|\,dx = nm \int\limits_{-\infty}^{\infty} f_X(nx)\, f_Y(mzx)\, |x|\,dx.$$

- $z \le 0 \quad \Rightarrow \quad f_Z(z) = 0.$

- $z > 0 \quad \Rightarrow$

$$
\begin{aligned}
f_Z(z) &= \frac{m^{\frac{m}{2}} n^{\frac{n}{2}}}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2}) 2^{\frac{m+n}{2}}} z^{\frac{m}{2}-1} \int_0^\infty x^{\frac{m+n}{2}-1} e^{-\frac{(mz+n)x}{2}}\, dx = \left\{ \begin{matrix} u = \frac{(mz+n)x}{2} \\ du = \frac{mz+n}{2}\, dx \end{matrix} \right\} = \\
&= \frac{m^{\frac{m}{2}} n^{\frac{n}{2}}}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2}) 2^{\frac{m+n}{2}}} z^{\frac{m}{2}-1} \frac{2^{\frac{m+n}{2}}}{(mz+n)^{\frac{m+n}{2}}} \int_0^\infty u^{\frac{m+n}{2}-1} e^{-u}\, du = \\
&= \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} m^{\frac{m}{2}} n^{\frac{n}{2}} \frac{z^{\frac{m}{2}-1}}{(mz+n)^{\frac{m+n}{2}}}
\end{aligned}
$$

$\Rightarrow \quad Z \sim \mathrm{F}(m, n)$

$\triangle$

3.2 Uvjetno očekivanje

Definicija: Neka je (X, Y) slučajni vektor s funkcijom gustoće $f = f_{X,Y} \colon \mathbb{R}^2 \to \mathbb{R}$. Ako je $y \in \mathbb{R}$ takav da je $f_Y(y) > 0$, onda definiramo **uvjetnu funkciju gustoće**

$$f(\cdot|y) = f_{X|Y}(\cdot|y) \colon \mathbb{R} \to \mathbb{R}$$

slučajne varijable X **uz uvjet** Y=y formulom

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

Napomena: Ako je $y \in \mathbb{R}$ takav da je $f_Y(y) > 0$, onda definiramo

$$\mathbb{P}(X \in A | Y = y) := \int_A f_{X|Y}(x|y)\, dx.$$

Sjetite se da u tom slučaju za diskretni slučajni vektor (X, Y) vrijedi

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

Napomena: Ako su X i Y nezavisne slučajne varijable, onda je

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{f_X(x) f_Y(y)}{f_Y(y)} = f_X(x).$$

Definicija: Neka su X i Y slučajne varijable i neka slučajna varijabla X ima konačno očekivanje. **Uvjetno (matematičko) očekivanje slučajne varijable X uz dano $Y = y$** je definirano za sve $y \in \mathbb{R}$ takve da je $f_Y(y) > 0$ formulom

$$\mathbb{E}[X|Y = y] := \begin{cases} \sum_x x f_{X|Y}(x|y), & \text{za diskretne slučajne varijable} \\ \int\limits_{-\infty}^{+\infty} x f_{X|Y}(x|y)\, dx, & \text{za neprekidne slučajne varijable} \end{cases}.$$

Napomena: Ako je $h\colon \mathbb{R} \to \mathbb{R}$ (izmjeriva) funkcija i slučajna varijabla $h(X)$ ima konačno očekivanje, onda za sve $y \in \mathbb{R}$ takve da je $f_Y(y) > 0$ vrijedi

$$\mathbb{E}[h(X)|Y = y] := \begin{cases} \sum_x h(x) f_{X|Y}(x|y), & \text{za diskretne slučajne varijable} \\ \int\limits_{-\infty}^{+\infty} h(x) f_{X|Y}(x|y)\, dx, & \text{za neprekidne slučajne varijable} \end{cases}.$$

Napomena: Na ovaj način je dobro definirana funkcija

$$g\colon \{y \in \mathbb{R}\colon f_Y(y) > 0\} \to \mathbb{R}, \quad g(y) = \mathbb{E}[X|Y = y].$$

Ako definiramo $g(y) := 0$, za $y \in \mathbb{R}$ takve da je $f_Y(y) = 0$, onda sa $\mathbb{E}[X|Y]$ označavamo funkciju slučajne varijable Y koja je za $Y = y$ jednaka $g(y)$, odnosno

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

Tada je $\mathbb{E}[X|Y]$ slučajna varijabla i vrijedi

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

Zadatak 3.7 Bacamo dvije simetrične kocke. Neka je X manji, a Y veći od brojeva koji su pali na kockama.

(a) Odredite distribuciju slučajnog vektora (X, Y).

(b) Odredite marginalne distribucije.

(c) Odredite $f_{X|Y}(x|4)$.

(d) Odredite $\mathbb{E}[X|Y = y]$.

(e) Odredite $\mathbb{E}[\mathbb{E}[X|Y]]$.

(f) Odredite $\mathbb{E}X$.

Rješenje:

(a) Distribucija slučajnog vektora (X, Y) je dana sljedećom tablicom:

$X \setminus Y$	1	2	3	4	5	6
1	1/36	2/36	2/36	2/36	2/36	2/36
2	0	1/36	2/36	2/36	2/36	2/36
3	0	0	1/36	2/36	2/36	2/36
4	0	0	1/36	2/36	2/36	2/36
5	0	0	0	1/36	1/36	2/36
6	0	0	0	0	0	1/36

(b) Marginalne distribucije su:

$$X \sim \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 11/36 & 9/36 & 7/36 & 5/36 & 3/36 & 1/36 \end{pmatrix}$$

$$Y \sim \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 1/36 & 3/36 & 5/36 & 7/36 & 9/36 & 11/36 \end{pmatrix}$$

(c) $f_{X|Y}(x|4) = \dfrac{f_{X,Y}(x,4)}{f_Y(4)} = \dfrac{f_{X,Y}(x,4)}{7/36}$

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

(d) Stavimo $g(y) = \mathbb{E}[X|Y=y]$. Tada je

y	1	2	3	4	5	6
$g(y)$	1	4/3	9/5	16/7	25/9	36/11

(e) $\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[g(Y)] = \sum_{y=1}^{6} g(y) f_Y(y) = 91/36.$

(f) $\mathbb{E}X = \sum_{x=1}^{6} x f_X(x) = 91/36.$

△

Zadatak 3.8 Slučajni vektor (X,Y) ima funkciju gustoće

$$f_{X,Y}(x,y) = \begin{cases} \frac{e^{-\frac{x}{y}} e^{-y}}{y}, & x > 0, y > 0 \\ o, & \text{inače} \end{cases}.$$

(a) Odredite $f_{X|Y}(x|y)$, za $y > 0$.

(b) Izračunajte $\mathbb{P}(X > 1|Y = y)$, za $y > 0$.

(c) Odredite $\mathbb{E}[X|Y=y]$, za $y>0$.

Rješenje:

(a) Odredimo prvo marginalnu funkciju gustoće slučajne varijable Y.

- $y \leq 0 \quad \Rightarrow \quad f_Y(y)=0.$
- $y > 0 \quad \Rightarrow$

$$f_Y(y)=\int\limits_{-\infty}^{+\infty} f_{X,Y}(x,y)\,dx=\int\limits_{0}^{+\infty}\frac{e^{-\frac{x}{y}}e^{-y}}{y}\,dx=e^{-y}\left(-e^{-\frac{x}{y}}\right)\Big|_0^{+\infty}=e^{-y}$$

$$\Rightarrow \quad f_Y(y)=\begin{cases} e^{-y}, & y>0\\ 0, & y\leq 0\end{cases}$$

Slijedi da je uvjetna funkcija gustoće slučajne varijable X uz uvjet $Y=y$ definirana za $y>0$ i jednaka

$$f_{X|Y}(x|y)=\frac{f_{X,Y}(x,y)}{f_Y(y)}=\begin{cases}\frac{e^{-\frac{x}{y}}}{y}, & x>0\\ 0 & x\leq 0\end{cases}$$

(b) $\mathbb{P}(X>1|Y=y)=\int\limits_{1}^{+\infty} f_{X|Y}(x|y)\,dx=\int\limits_{1}^{+\infty}\frac{e^{-\frac{x}{y}}}{y}\,dx=\left(-e^{-\frac{x}{y}}\right)\Big|_1^{+\infty}=e^{-\frac{1}{y}}$, za $y>0$.

(c) Uvjetno očekivanje od X uz dano $Y=y$ je također definirano za $y>0$ i iznosi

$$\mathbb{E}[X|Y=y]=\int\limits_{-\infty}^{+\infty} x f_{X|Y}(x|y)\,dx=\int\limits_{0}^{+\infty} x\,\frac{e^{-\frac{x}{y}}}{y}\,dx=\left\{\begin{matrix} u=x & dv=\frac{e^{-\frac{x}{y}}}{y}\,dx\\ du=dx & v=-e^{-\frac{x}{y}}\end{matrix}\right\}=$$

$$=\left(-xe^{-\frac{x}{y}}\right)\Big|_0^{+\infty}+\int\limits_0^{+\infty} e^{-\frac{x}{y}}\,dx=0+\left(-ye^{-\frac{x}{y}}\right)\Big|_0^{+\infty}=y.$$

△

Zadatak 3.9 Funkcija gustoće slučajnog vektora (X,Y) je

$$f(x,y)=\begin{cases} C(x^2-y^2)e^{-x}, & x>0, -x\leq y\leq x\\ 0, & \text{inače}\end{cases}.$$

Odredite $\mathbb{E}[Y^2|X]$.

Rješenje: Odredimo prvo marginalnu funkciju gustoće slučajne varijable X.

- $x \leq 0 \quad \Rightarrow \quad f_X(x) = 0.$
- $x > 0 \quad \Rightarrow \quad f_X(x) = \int\limits_{-\infty}^{+\infty} f_{X,Y}(x,y)\,dy = \int\limits_{-x}^{x} C(x^2 - y^2)e^{-x}\,dy = \frac{4C}{3}x^3e^{-x}$

$$\Rightarrow \quad f_X(x) = \begin{cases} \frac{4C}{3}x^3e^{-x}, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

Slijedi da je uvjetna funkcija gustoće slučajne varijable Y uz uvjet $X = x$ definirana za $x > 0$ i jednaka

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \begin{cases} \frac{3}{4}\left(\frac{1}{x} - \frac{y^2}{x^3}\right), & -x \leq y \leq x \\ 0, & \text{inače} \end{cases}.$$

Uvjetno očekivanje od Y^2 uz dano $X = x$ je također definirano za $x > 0$ i iznosi

$$\mathbb{E}[Y^2|X = x] = \int\limits_{-\infty}^{+\infty} y^2 f_{Y|X}(y|x)\,dy = \int\limits_{-x}^{x} y^2\,\frac{3}{4}\left(\frac{1}{x} - \frac{y^2}{x^3}\right)dy = \frac{x^2}{5}$$

$$\Rightarrow \quad \mathbb{E}[Y^2|X] = \frac{X^2}{5}.$$

△

Zadatak 3.10 Funkcija gustoće slučajnog vektora (X,Y) je

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

Odredite $\mathbb{E}[X^3|Y]$.

Rješenje: Odredimo prvo marginalnu funkciju gustoće slučajne varijable Y.

- $y \leq 0 \quad \Rightarrow \quad f_Y(y) = 0.$
- $y > 0 \quad \Rightarrow \quad f_Y(y) = \int\limits_{-\infty}^{+\infty} f_{X,Y}(x,y)\,dx = \int\limits_{0}^{y} \frac{e^{-y}}{y}\,dx = e^{-y}$

$$\Rightarrow \quad f_Y(y) = \begin{cases} e^{-y}, & y > 0 \\ 0, & y \leq 0 \end{cases}$$

Slijedi da je uvjetna funkcija gustoće slučajne varijable X uz uvjet $Y = y$ definirana za $y > 0$ i jednaka

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \begin{cases} \frac{1}{y}, & 0 < x < y \\ 0 & \text{inače} \end{cases}.$$

Uvjetno očekivanje od X^3 uz dano $Y = y$ je također definirano za $y > 0$ i iznosi

$$\mathbb{E}[X^3|Y = y] = \int_{-\infty}^{+\infty} x^3 f_{X|Y}(x|y)\,dx = \int_0^y x^3\,\frac{1}{y}\,dx = \frac{y^3}{4}$$

$$\Rightarrow \quad \mathbb{E}[X^3|Y] = \frac{Y^3}{4}.$$

△

Zadatak 3.11 Neka je (X, Y) neprekidni slučajni vektor s funkcijom gustoće

$$f(x,y) = \begin{cases} \frac{C}{x}, & 27 \le y \le x \le 33 \\ 0, & \text{inače} \end{cases}.$$

(a) Odredite C.

(b) Odredite marginalne distribucije.

(c) Odredite $\mathbb{E}X$ i $\mathbb{E}Y$.

(d) Odredite $\mathbb{E}[XY]$ i $\mathrm{Cov}(X, Y)$.

(e) Odredite $f_{X|Y}(x|y)$ i $\mathbb{E}[X|Y = y]$ za $y \in [27, 33\rangle$.

(f) Odredite $f_{Y|X}(y|x)$ i $\mathbb{E}[Y|X = x]$ za $x \in \langle 27, 33]$.

Rješenje:

(a) $$1 = \iint_{\mathbb{R}^2} f(x,y)\,dx\,dy = \int_{27}^{33}\left(\int_y^{33}\frac{C}{x}\,dx\right)\,dy = C\int_{27}^{33}(\ln 33 - \ln y)\,dy = C\left(6 - 27\ln\frac{33}{27}\right)$$

$$\Rightarrow \quad C = \frac{1}{6 - 27\ln\frac{33}{27}}.$$

(b) $$f_X(x) = \int_{-\infty}^{+\infty} f(x,y)\,dy = \begin{cases} 0, & x \notin [27,33] \\ \displaystyle\int_{27}^{x}\frac{C}{x}\,dy = C\,\frac{x-27}{x}, & x \in [27,33] \end{cases}$$

$$f_Y(y) = \int_{-\infty}^{+\infty} f(x,y)\,dx = \begin{cases} 0, & y \notin [27,33] \\ \displaystyle\int_{y}^{3} 3\frac{C}{x}\,dx = C(\ln 33 - \ln y), & y \in [27,33] \end{cases}$$

(c) $\mathbb{E}X = \int\limits_{-\infty}^{+\infty} x f_X(x)\,dx = \int\limits_{27}^{33} C(x-27)\,dx = 18C$

$$\mathbb{E}Y = \int\limits_{-\infty}^{+\infty} y f_Y(y)\,dy = \int\limits_{27}^{33} Cy(\ln 33 - \ln y)\,dy = C\left(\frac{729}{2}\ln\frac{27}{33} + 90\right)$$

(d) $\mathbb{E}[XY] = \iint\limits_{\mathbb{R}^2} xyf(x,y)\,dx\,dy = \int\limits_{27}^{33}\left(\int\limits_{27}^{x} xy\frac{C}{x}\,dy\right)dx = \frac{C}{2}\int\limits_{27}^{33}\left(x^2-729\right)dx = 522C$

$$\mathrm{Cov}(X,Y) = \mathbb{E}[(X-\mathbb{E}X)(Y-\mathbb{E}Y)] = \mathbb{E}[XY]-\mathbb{E}X\mathbb{E}Y = 522C-18C^2\left(\frac{729}{2}\ln\frac{27}{33}+90\right).$$

(e) Za $y \in [27,33\rangle$ je $f_{X|Y}(x|y) = \dfrac{f(x,y)}{f_Y(y)} = \dfrac{1}{x(\ln 33 - \ln y)}\mathbb{1}_{[y,33]}(x)$

$$\Rightarrow \quad \mathbb{E}[X|Y=y] = \int\limits_{-\infty}^{+\infty} x f_{X|Y}(x|y)\,dx = \int\limits_{y}^{33}\frac{1}{\ln 33 - \ln y}\,dx = \frac{33-y}{\ln 33 - \ln y}.$$

(f) Za $x \in \langle 27,33]$ je $f_{Y|X}(y|x) = \dfrac{f(x,y)}{f_X(x)} = \dfrac{1}{x-27}\mathbb{1}_{[27,x]}(y)$

$$\Rightarrow \quad X|Y=y \sim U(27,x) \quad \Rightarrow \quad \mathbb{E}[Y|X=x] = \frac{27+x}{2}.$$

△

Zadatak 3.12 Neka je (X,Y) dvodimenzionalni normalni slučajni vektor s gustoćom

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho\frac{x-\mu_x}{\sigma_x}\frac{y-\mu_y}{\sigma_y} + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right]\right\},$$

gdje su $\mu_x, \mu_y\sigma_x\sigma_y\rho \in \mathbb{R}, -1 < \rho < 1, \sigma_x, \sigma_y > 0$.

(a) Odredite marginalne razdiobe.

(b) Izračunajte koeficijent korelacije slučajnih varijabli X i Y.

(c) Odredite $f_{X|Y}(x|y)$ i $\mathbb{E}[X|Y=y]$.

(d) Odredite $f_{Y|X}(y|x)$ i $\mathbb{E}[Y|X=x]$.

Rješenje: Stavimo $C := \dfrac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}$

(a) $f_X(x) = \int\limits_{-\infty}^{+\infty} f(x,y)\,dy =$

$$= C\int\limits_{-\infty}^{+\infty} exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho\frac{x-\mu_x}{\sigma_x}\frac{y-\mu_y}{\sigma_y} + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right]\right\} dy =$$

$$= \left\{\begin{array}{c} z = \dfrac{y-\mu_y}{\sigma_y} \\ dz = \dfrac{dy}{y} \end{array}\right\} =$$

$$= C\sigma_y exp\left\{-\frac{1}{2(1-\rho^2)}\left(\frac{x-\mu_x}{\sigma_x}\right)^2\right\}\int\limits_{-\infty}^{+\infty} exp\left\{-\frac{1}{2(1-\rho^2)}\left(z^2 - 2\rho\frac{x-\mu_x}{\sigma_x}z\right)\right\} dz =$$

$$= C\sigma_y exp\left\{-\frac{1}{2(1-\rho^2)}\left(\frac{x-\mu_x}{\sigma_x}\right)^2(1-\rho^2)\right\}\int\limits_{-\infty}^{+\infty} exp\left\{-\frac{1}{2(1-\rho^2)}\left(z - \rho\frac{x-\mu_x}{\sigma_x}\right)^2\right\} dz =$$

$$= C\sigma_y e^{-\frac{(x-\mu_x)^2}{2\sigma_x^2}}\sqrt{2\pi}\sqrt{1-\rho^2} = \frac{1}{\sigma_x\sqrt{2\pi}}e^{-\frac{(x-\mu_x)^2}{2\sigma_x^2}}$$

$\Rightarrow \quad X \sim N(\mu_x, \sigma_x^2)$ i analogno $Y \sim N(\mu_y, \sigma_y^2)$

(b) $\mathrm{Cov}(X,Y)\mathbb{E}[(X-\mathbb{E}X)(Y-\mathbb{E}Y)] = \mathbb{E}[(X-\mu_x)(Y-\mu_y)] = \iint\limits_{\mathbb{R}^2}(x-\mu_x)(y-\mu_y)f(x,y)\,dx\,dy =$

$$= C\iint\limits_{\mathbb{R}^2}(x-\mu_x)(y-\mu_y)exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho\frac{x-\mu_x}{\sigma_x}\frac{y-\mu_y}{\sigma_y} + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right]\right\} dx\,dy =$$

$$= \left\{\begin{array}{c} z = \frac{x-\mu_x}{\sigma_x} \Rightarrow x = \mu_x + \sigma_x z \\ w = \frac{y-\mu_y}{\sigma_y} \Rightarrow y = \mu_y + \sigma_y w \\ J = \begin{vmatrix}\sigma_x & 0 \\ 0 & \sigma_y\end{vmatrix} = \sigma_x\sigma_y \end{array}\right\} = C\sigma_x^2\sigma_y^2\int\limits_{-\infty}^{+\infty}\int\limits_{-\infty}^{+\infty} zw\exp\left\{-\frac{z^2-2\rho zw+w^2}{2(1-\rho^2)}\right\} dz\,dw =$$

$$= C\sigma_x^2\sigma_y^2\int\limits_{-\infty}^{+\infty} we^{-\frac{w^2}{2}}\left(\int\limits_{-\infty}^{+\infty} z\exp\left\{-\frac{(z-\rho w)^2}{2(1-\rho^2)}\right\}\right) dz\,dw = C\sigma_x^2\sigma_y^2\int\limits_{-\infty}^{+\infty} we^{-\frac{w^2}{2}}\rho w\sqrt{2\pi}\sqrt{1-\rho^2}\,dw =$$

$$= C\sigma_x^2\sigma_y^2\rho\sqrt{2\pi}\sqrt{1-\rho^2}\cdot\sqrt{2\pi}\quad\underbrace{\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty} w^2e^{-\frac{w^2}{2}}\,dw}_{=\text{ varijanca N(0,1) razdiobe }=1}\quad = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\sigma_x^2\sigma_y^2\rho 2\pi\sqrt{1-\rho^2} = \sigma_x\sigma_y\rho$$

$$\Rightarrow \quad \rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}} = \frac{\sigma_x\sigma_y\rho}{\sigma_x\sigma_y} = \rho.$$

(c) Prema (a) je $f_Y(y) = \frac{1}{\sigma_y\sqrt{2\pi}} e^{-\frac{(y-\mu_y)^2}{2\sigma_y^2}}$. Stoga je $f_{X|Y}(x|y) =$

$$= \frac{f(x,y)}{f_Y(y)} = \frac{\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho\frac{x-\mu_x}{\sigma_x}\frac{y-\mu_y}{\sigma_y} + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right]\right\}}{\frac{1}{\sigma_y\sqrt{2\pi}} e^{-\frac{(y-\mu_y)^2}{2\sigma_y^2}}} =$$

$$= \frac{1}{\sigma_y\sqrt{1-\rho^2}\sqrt{2\pi}} exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho\frac{x-\mu_x}{\sigma_x}\frac{y-\mu_y}{\sigma_y} + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\rho^2\right]\right\} =$$

$$= \frac{1}{\sigma_y\sqrt{1-\rho^2}\sqrt{2\pi}} exp\left\{-\frac{1}{2(1-\rho^2)}\left(\frac{x-\mu_x}{\sigma_x} - \rho\frac{y-\mu_y}{\sigma_y}\right)^2\right\} =$$

$$= \frac{1}{\sigma_y\sqrt{1-\rho^2}\sqrt{2\pi}} exp\left\{-\frac{1}{2\sigma_x^2(1-\rho^2)}\left[x - \left(\mu_x + \rho\frac{\sigma_x}{\sigma_y}(y-\mu_y)\right)\right]^2\right\}$$

$$\Rightarrow \quad X|Y = y \sim N(\mu_x + \rho\frac{\sigma_x}{\sigma_y}(y-\mu_y), \sigma_x^2(1-\rho^2))$$

$$\Rightarrow \quad \mathbb{E}[X|Y=y] = \mu_x + \rho\frac{\sigma_x}{\sigma_y}(y-\mu_y)$$

(d) Analogno je

$Y|X = x \sim N(\mu_y + \rho\frac{\sigma_y}{\sigma_x}(x-\mu_x), \sigma_y^2(1-\rho^2))$

$$\Rightarrow \quad \mathbb{E}[Y|X=x] = \mu_y + \rho\frac{\sigma_y}{\sigma_x}(x-\mu_x)$$

$\triangle$

Zadaci za vježbu

3.13 Neka je (X,Y) slučajni vektor s funkcijom gustoće

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

(a) Odredite C.

(b) Izračunajte $\mathbb{E}X$ i $\mathbb{E}Y$.

3.14 Bacate nesimetričnu kocku: svaki od brojeva 1, 3 ili 5 pada s vjerojatnošću C, a svaki od brojeva 2,4 ili 6 pada s vjerojatnošću $2C$.

(a) Odredite C.

(b) Neka je X jednak 1 ako je pao neparan broj, a 0 inače. Neka je Y jednak 1 ako je pao broj veći od 3, a inače je 0. Odredite distribuciju slučajnog vektora (X,Y).

3.15 Neka je (X,Y) slučajni vektor s funkcijom gustoće

$$f(x,y)=\begin{cases} \frac{x}{5}+cy, & 0<x<1, 1<y<5,\\ 0, & \text{inače.}\end{cases}$$

(a) Odredite c.

(b) Jesu li X i Y nezavisne slučajne varijable?

(c) Odredite $\mathbb{P}(X+Y>3)$.

3.16 Neka je (X,Y) neprekidni slučajni vektor s funkcijom gustoće $f(x,y)$.

(a) Pokažite da je $U=X+Y$ neprekidna slučajna varijabla i da je

$$f_U(z)=\int_{-\infty}^{+\infty} f(x,z-x)\,dx.$$

(b) Pokažite da je $V=Y-X$ neprekidna slučajna varijabla i da je

$$f_V(z)=\int_{-\infty}^{+\infty} f(x,z+x)\,dx.$$

(c) Pokažite da je $W=X\cdot Y$ neprekidna slučajna varijabla i da je

$$f_W(z)=\int_{-\infty}^{+\infty} f(x,\frac{z}{x})\frac{1}{|x|}\,dx.$$

(c) Pokažite da je $Z = \frac{Y}{X}$ neprekidna slučajna varijabla i da je

$$f_Z(z) = \int_{-\infty}^{+\infty} f(x, zx)|x|\, dx.$$

3.17 Neka su X i Y nezavisne uniformne slučajne varijable na $(0, 1)$. Izračunajte funkciju gustoće slučajne varijable $Z = X + Y$.

3.18 Neka je (X, Y) slučajni vektor s funkcijom gustoće

$$f(x, y) = \begin{cases} axye^{-(x^2+y^2)}, & x > 0, y > 0, \\ 0, & \text{inače.} \end{cases}$$

(a) Odredite a.

(b) Odredite marginalne gustoće.

(c) Odredite uvjetne gustoće.

3.19 Ako je (X, Y) neprekidni slučajni vektor s funkcijom gustoće

$$f(x, y) = \frac{3\sqrt{3}}{\pi} e^{-4x^2-6xy-9y^2},$$

odredite $f_{X|Y}(x|y)$ i $f_{Y|X}(y|x)$.

3.20 Neka su X i Y nezavisne binomne slučajne varijable s parametrima n i p. Odredite

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

3.21 Neka su $X \sim P(\lambda_1)$ i $Y \sim P(\lambda_2)$ nezavisne slučajne varijable. Odredite $f_{X|X+Y=n}(x|n)$.

3.22 Neka su X i Y nezavisne uniformne slučajne varijable na skupu $\{1, 2, \ldots, m\}$. Pokažite da je

$$\mathbb{E}[|X - Y|] = \frac{m^2 - 1}{3m}.$$

3.23 Simetrični novčić bacamo 2 puta. Označimo s X broj pisama koji su pali. Neka je Y jednaka 0 ako glava uopće nije pala. Inače, neka je Y jednaka rednom broju bacanja kada je prvi put pala glava.

(a) Odredite razdiobu slučajnog vektora (X, Y).

(b) Jesu li slučajne varijable X i Y nezavisne?

(c) Jesu li slučajne varijable X i Y nekorelirane?

(d) Odredite koeficijent korelacije između slučajnih varijabli X i Y.

3.24 Neka su $X \sim N(\mu_1, \sigma_1^2)$ i $Y \sim N(\mu_2, \sigma_2^2)$ nezavisne slučajne varijable. Pokažite da je

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

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

3.25 Neka su $X, Y \sim N(0, 1)$ nezavisne slučajne varijable. Pokažite da $\frac{X}{Y}$ ima Cauchyevu razdiobu.

Rješenja: **3.13.** (a) 0.25 (b) $\mathbb{E}X = -1$ i $\mathbb{E}Y = 3$ **3.14.** (a) 1/9 (b) p(1,1)=4/9, P(1,0)=2/9, p(0,1)=1/9, p(0,0)=2/9 **3.15.** (a) 1/20 (b) Ne (c)11/15 **3.17.** $f_Z(x) = \begin{cases} x, & 0 \leq x \leq 1, \\ 2 - x, & 1 < x < 2, \\ 0, & \text{inače.} \end{cases}$ **3.18.** (a) 4 (b)$f_X(x) = \begin{cases} 2xe^{-x^2}, & x \geq 0, \\ 0, & \text{inače.} \end{cases}$, $f_Y(x) = \begin{cases} 2ye^{-y^2}, & y \geq 0, \\ 0, & \text{inače.} \end{cases}$ (c) $f_{X|Y}(x|y) = f_X(x),\ y \geq 0, f_{Y|X}(y|x) = f_Y(y),\ x \geq 0$ **3.19.** $f_{X|Y}(x|y) = \frac{2}{\sqrt{\pi}} e^{-(2x+1.5y)^2}$, $f_{Y|X}(y|x) = \frac{3}{\sqrt{\pi}} e^{-(x+3y)^2}$ **3.20.** $m/2$ **3.21.** $f_{X|X+Y=n}(x|n) = \binom{n}{x} \left(\frac{\lambda_1}{\lambda_1+\lambda_a}\right)^k \left(\frac{\lambda_2}{\lambda_1+\lambda_a}\right)^{n-k}$, za $x \in \{0, 1, \ldots, n\}$, a 0 inače **3.23.** (a) $p(0,1) = p(1,1) = p(1,2) = p(2,0) = 0.25$, ostale vrijednosti su 0 (b) Ne (c) Ne (d) -0.5

4

Procjena parametara

Neka je X slučajna varijabla čiju distribuciju proučavamo.

Definicija: Slučajni uzorak duljine n za X je niz od n nezavisnih i jednako distribuiranih slučajnih varijabli $X_1, X_2, \dots, X_n$ koje imaju istu razdiobu kao i X.

Za $\omega \in \Omega$ je $x_1 = X_1(\omega), x_2 = X_2(\omega), \dots, x_n = X_n(\omega)$ jedna realizacija slučajnog uzorka i zovemo je **uzorak**. **Statistika** je funkcija slučajnog uzorka.

Razdioba slučajne varijable X je često opisana parametrima koje pokušavamo procijeniti.

Definicija: Procjenitelj $T_n = f_n(X_1, \dots, X_n)$ je **nepristrani procjenitelj** za parametar τ ako vrijedi $\mathbb{E}T_n = \tau$.

Zadatak 4.1 Neka je $X_1, X_2, \dots, X_n$ slučajni uzorak iz populacije s konačnim očekivanjem μ i varijancom σ^2. Pokažite da je

(a) $\overline{X}_n := \dfrac{X_1 + \cdots + X_n}{n}$ nepristrani procjenitelj za μ,

(b) $S_n^2 := \dfrac{1}{n-1} \displaystyle\sum_{i=1}^{n} (X_i - \overline{X}_n)^2$ nepristrani procjenitelj za σ^2.

Rješenje:

(a) $\mathbb{E}[\overline{X}_n] = \mathbb{E}\left[\dfrac{X_1 + \cdots + X_n}{n}\right] = \dfrac{\overbrace{\mathbb{E}X_1}^{=\mu} + \cdots + \overbrace{\mathbb{E}X_n}^{=\mu}}{n} = \mu$

(b)

$$\begin{aligned}
\mathbb{E}[S_n^2] &= \mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X}_n)^2\right] = \frac{1}{n-1}\,\mathbb{E}\left[\sum_{i=1}^{n}X_i^2 - 2\overline{X}_n\sum_{i=1}^{n}X_i + n\overline{X}_n^2\right] = \\
&= \frac{1}{n-1}\,\mathbb{E}\left[\sum_{i=1}^{n}X_i^2 - 2n\overline{X}_n^2 + n\overline{X}_n^2\right] = \frac{1}{n-1}\,\mathbb{E}\left[\sum_{i=1}^{n}X_i^2 - n\overline{X}_n^2\right] = \\
&= \frac{1}{n-1}\,\mathbb{E}\left[\sum_{i=1}^{n}X_i^2 - \frac{1}{n}(\sum_{i=1}^{n}X_i)^2\right] = \frac{1}{n-1}\left\{\sum_{i=1}^{n}\mathbb{E}[X_i^2] - \frac{1}{n}\,\mathbb{E}[(\sum_{i=1}^{n}X_i)^2]\right\} = \\
&= \frac{1}{n-1}\left\{\sum_{i=1}^{n}[\underbrace{\operatorname{Var}X_i}_{\sigma^2} + (\underbrace{\mathbb{E}X_i}_{=\mu})^2] - \frac{1}{n}\,[\underbrace{\operatorname{Var}(\sum_{i=1}^{n}X_i)}_{=\sum_{i=1}^{n}\operatorname{Var}X_i} + (\underbrace{\mathbb{E}[\sum_{i=1}^{n}X_i]}_{=\sum_{i=1}^{n}\mathbb{E}X_i})^2]\right\} = \\
&= \frac{1}{n-1}\left\{n\sigma^2 + n\mu^2 - \frac{1}{n}\,[n\sigma^2 + (n\mu)^2]\right\} = \frac{1}{n-1}\,(n-1)\sigma^2 = \sigma^2.
\end{aligned}$$

△

4.1 Metoda maksimalne vjerodostojnosti

Neka je $(x_1,\dots,x_n)$ opaženi uzorak za slučajnu varijablu X s gustoćom $f(x|\theta)$, gdje je $\theta = (\theta_1,\dots,\theta_k) \in \Theta \subseteq \mathbb{R}^k$ nepoznati parametar.

Definiramo **funkciju vjerodostojnosti** $L\colon \Theta \to \mathbb{R}$ sa

$$L(\theta) := f(x_1|\theta)\cdots f(x_n|\theta),\ \theta \in \Theta.$$

Vrijednost $\hat{\theta} = \hat{\theta}(x_1,\dots,x_n) \in \Theta$ za koju je

$$L(\hat{\theta}) = \max_{\theta\in\Theta} L(\theta)$$

zovemo **procjena metodom maksimalne vjerodostojnosti**. Statistika $\hat{\theta}(X_1,\dots,X_n)$ je **procjenitelj metodom maksimalne vjerodostojnosti** ili kraće **MLE**.

Zadatak 4.2 Neka je $X_1, X_2, \dots, X_n$ slučajni uzorak iz geometrijske razdiobe s parametrom $p \in \langle 0,1\rangle$. Nađite MLE za p.

Rješenje:

$$X \sim G(p) \quad\Rightarrow\quad \mathbb{P}(X=k) = (1-p)^{k-1}p, k\in\mathbb{N} \quad\Rightarrow\quad f(x|p) = \begin{cases}(1-p)^{x-1}p, x\in\mathbb{N}\\ 0, \text{inače}\end{cases}$$

Neka je $x_1, x_2, \ldots, x_n$ opaženi uzorak. Tada vrijedi $x_1, x_2, \ldots, x_n \in \mathbb{N}$ i

$$L(p) = \prod_{i=1}^{n} f(x_i|p) = (1-p)^{\sum\limits_{i=1}^{n} x_i - n} p^n, \ p \in \langle 0, 1 \rangle$$

Funkcija $x \mapsto \ln x$ je strogo rastuća pa je dovoljno maksimizirati funkciju

$$l(p) = \ln L(p) = \left(\sum_{i=1}^{n} x_i - n \right) \ln(1-p) + n \ln p.$$

Vrijedi

$$l'(p) = -\frac{\sum\limits_{i=1}^{n} x_i - n}{1-p} + \frac{n}{p}$$

$$l'(p) = 0 \Leftrightarrow \frac{n}{p} = \frac{\sum\limits_{i=1}^{n} x_i - n}{1-p} \Leftrightarrow p = \frac{n}{\sum\limits_{i=1}^{n} x_i}$$

$$l''(p) = -\frac{\sum\limits_{i=1}^{n} x_i - n}{(1-p)^2} - \frac{n}{p^2} < 0$$

pa funkcija l poprima maksimum u $\hat{p} = \frac{n}{\sum\limits_{i=1}^{n} x_i}$. Dakle, MLE za parametar p je $\frac{1}{\overline{X}_n}$.

△

Zadatak 4.3 Neka je $X_1, X_2, \ldots, X_n$ slučajni uzorak iz Poissonove razdiobe s parametrom $\lambda > 0$.

(a) Nađite MLE za λ.

(b) Ispitajte je li MLE nepristrani procjenitelj za λ.

Rješenje:

(a) Neka je $x_1, x_2, \ldots, x_n$ opaženi uzorak. Tada je

$$L(\lambda) = \prod_{i=1}^{n} f(x_i|\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} = \frac{\lambda^{\sum\limits_{i=1}^{n} x_i}}{x_1! \cdots x_n!} e^{-n\lambda} = c \cdot \lambda^{\sum\limits_{i=1}^{n} x_i} e^{-n\lambda},$$

gdje $c > 0$ ne ovisi o λ. Stoga je dovoljno maksimizirati funkciju

$$l(\lambda) = \ln \left(\lambda^{\sum\limits_{i=1}^{n} x_i} e^{-n\lambda} \right) = \left(\sum_{i=1}^{n} x_i \right) \ln \lambda - n\lambda, \ \lambda > 0.$$

Vrijedi

$$l'(\lambda) = \frac{\sum_{i=1}^{n} x_i}{\lambda} - n$$

$$l'(\lambda) = 0 \Leftrightarrow \lambda = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$l''(\lambda) = -\frac{\sum_{i=1}^{n} x_i}{\lambda^2} \leq 0$$

pa funkcija l poprima maksimum u $\hat{\lambda} = \frac{\sum_{i=1}^{n} x_i}{n}$ i MLE za λ je $\overline{X}_n$.

(b) Prema zadatku 4.1 je $\mathbb{E}[\overline{X}_n] = \mathbb{E}X_1 = \lambda \quad \Rightarrow \quad \overline{X}_n$ je nepristrani procjenitelj za λ.

$\triangle$

Napomena: Ako je $\hat{\theta}$ MLE za θ i $g\colon \Theta \to g(\Theta)$ neka funkcija, onda je $g(\hat{\theta})$ MLE za $g(\theta)$.

Zadatak 4.4 Nađite MLE parametara $\theta = (\mu, \sigma^2)$ normalnog modela $N(\mu, \sigma^2)$. Ispitajte nepristranost procjenitelja.

Rješenje: Stavimo $\theta_1 = \mu$ i $\theta_2 = \sigma^2$. Neka je $x_1, x_2, \ldots, x_n$ opaženi uzorak. Tada je

$$L(\theta_1, \theta_2) = \prod_{i=1}^{n} f(x_i|\theta_1, \theta_2) = \prod_{i=1}^{n} \frac{1}{\sqrt{\theta_2}\sqrt{2\pi}} e^{-\frac{(x_i-\theta_1)^2}{2\theta_2}} = c\ \theta_2^{-\frac{n}{2}} e^{-\frac{1}{2\theta_2}\sum_{i=1}^{n}(x_i-\theta_1)^2}$$

za $(\theta_1, \theta_2) \in \mathbb{R} \times \langle 0, +\infty\rangle$. Budući da je $c > 0$ i funkcija $x \mapsto \ln x$ strogo rastuća, dovoljno je maksimizirati funkciju

$$l(\theta_1, \theta_2) = \ln\left(\theta_2^{-\frac{n}{2}} e^{-\frac{1}{2\theta_2}\sum_{i=1}^{n}(x_i-\theta_1)^2}\right) = -\frac{n}{2}\ln\theta_2 - \frac{1}{2\theta_2}\sum_{i=1}^{n}(x_i - \theta_1)^2.$$

Vrijedi

$$\frac{\partial l}{\partial \theta_1} = \frac{1}{\theta_2}\sum_{i=1}^{n}(x_i - \theta_1) = 0 \quad \Leftrightarrow \quad \theta_1 = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\frac{\partial l}{\partial \theta_2} = -\frac{n}{2}\frac{1}{\theta_2} + \frac{1}{2\theta_2^2}\sum_{i=1}^{n}(x_i - \theta_1)^2 = 0 \quad \Leftrightarrow \quad \theta_2 = \frac{\sum_{i=1}^{n}(x_i - \theta_1)^2}{n}$$

Zbog prethodne napomene stavimo

$$\hat{\theta}_1 = \frac{\sum\limits_{i=1}^{n} x_i}{n} = \bar{x}, \quad \hat{\theta}_2 = \frac{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}{n} = \frac{n-1}{n}s^2.$$

Budući da je Hesseova matrica

$$Hl(\hat{\theta}_1, \hat{\theta}_2) = \begin{bmatrix} -\frac{n}{\hat{\theta}_2} & -\frac{1}{\hat{\theta}_2^2}\sum\limits_{i=1}^{n}(x_i - \hat{\theta}_1) \\ -\frac{1}{\hat{\theta}_2^2}\sum\limits_{i=1}^{n}(x_i - \hat{\theta}_1) & \frac{n}{2}\frac{1}{\hat{\theta}_2^2} - \frac{1}{\hat{\theta}_2^3}\sum\limits_{i=1}^{n}(x_i - \hat{\theta}_1)^2 \end{bmatrix} = \begin{bmatrix} -\frac{n}{\hat{\theta}_2} & 0 \\ 0 & -\frac{n}{2}\frac{1}{\hat{\theta}_2^2} \end{bmatrix}$$

negativno definitna, MLE za μ je $\overline{X}_n$, a MLE za σ^2 je $\frac{n-1}{n}S_n^2$.

Prema zadatku 4.1 slijedi

$$\mathbb{E}[\overline{X}_n] = \mu \quad \text{i} \quad \mathbb{E}[\frac{n-1}{n}S_n^2] = \frac{n-1}{n}\mathbb{E}[S_n^2] = \frac{n-1}{n}\sigma^2 \neq \sigma^2$$

Dakle, $\overline{X}_n$ je nepristrani procjenitelj za μ, ali $\frac{n-1}{n}S_n^2$ nije nepristrani procjenitelj za σ^2.

$\triangle$

Zadatak 4.5 Neka je $X_1, \ldots, X_n$ slučajni uzorak iz modela s funkcijom gustoće

$$f(x|t) = \begin{cases} 0, & x \leq t \\ e^{-(x-t)}, & x > t \end{cases}.$$

Nađite MLE za parametar $t \in \mathbb{R}$.

Rješenje: Neka je $x_1, x_2, \ldots, x_n$ opaženi uzorak. Tada je

$$L(t) = \prod_{i=1}^{n} f(x_i|t) = \begin{cases} e^{-\sum\limits_{i=1}^{n} x_i + nt}, & x_{(1)} > t \\ 0, & \text{inače} \end{cases}.$$

Funkcija $t \mapsto L(t)$ strogo raste na $\langle -\infty, x_{(1)} \rangle$, a dalje je jednaka 0. Stoga se maksimum postiže u $\hat{t} = x_{(1)}$ i MLE za t je $X_{(1)} = \min\{X_1, \ldots, X_n\}$.

$\triangle$

Zadatak 4.6 Neka je $X_1, X_2, \ldots, X_n$ slučajni uzorak iz diskretne uniformne razdiobe na skupu $\{1, 2, \ldots, m\}, m \in \mathbb{N}$.

(a) Nađite MLE za m.

(b) Ispitajte je li MLE nepristrani procjenitelj za m.

Rješenje:

(a) Neka je $x_1, x_2, \dots, x_n$ opaženi uzorak. Vrijedi

$$f(x|m) = \begin{cases} \frac{1}{m}, & x \in \{1, \dots, m\} \\ 0, & \text{inače} \end{cases}$$

pa je

$$L(m) = \prod_{i=1}^{n} f(x_i|m) = \begin{cases} \frac{1}{m^n}, & x_{(n)} \leq m \\ 0, & \text{inače} \end{cases}.$$

Funkcija $m \mapsto L(m)$ strogo pada na $[x_{(n)}, +\infty\rangle$, a inače je jednaka 0. Stoga se maksimum postiže u $\hat{m} = x_{(n)}$ pa je MLE za m jednak $X_{(n)} = \max\{X_1, \dots, X_n\}$.

(b) MLE nije nepristrani procjenitelj za m jer je

$$\begin{aligned} \mathbb{E}[X_{(n)}] &= \mathbb{E}[\max\{X_1, \dots, X_n\}] = \\ &= \sum_{k=1}^{\infty} \mathbb{P}(\max\{X_1, \dots, X_n\} \geq k) = \sum_{k=1}^{\infty} (1 - \mathbb{P}(\max\{X_1, \dots, X_n\} < k)) = \\ &= \sum_{k=1}^{m} (1 - \mathbb{P}(X_1 < k, \dots, X_n < k)) = \sum_{k=1}^{m} (1 - \mathbb{P}(X_1 < k) \cdots \mathbb{P}(X_n < k)) = \\ &= \sum_{k=1}^{m} (1 - \mathbb{P}(X_1 < k)^n) = \sum_{k=0}^{m-1} (1 - \mathbb{P}(X_1 \leq k)^n) = \sum_{k=0}^{m-1} \left[1 - \left(\frac{k}{m}\right)^n\right] < m. \end{aligned}$$

4.2 Metoda momenata

Procjenitelje metodom momenata dobivamo izjednačavanjem **momenata razdiobe** s odgovarajućim **uzoračkim momentima**.

Zadatak 4.7 Neka je $X_1, \dots, X_n$ slučajni uzorak za $X \sim \Gamma(\alpha, \beta), \alpha, \beta > 0$. Nađite procjenitelje za α i β metodom momenata.

Rješenje: Neka je $x_1, x_2, \dots, x_n$ opaženi uzorak. Prema zadatku 2.7 je $\mathbb{E}X = \alpha\beta$ i $\operatorname{Var} X = \alpha\beta^2$. Izjednačavanjem populacijskog i uzoračkog očekivanja, te populacijske i uzoračke varijance, dobivamo

$$\begin{aligned} \alpha\beta &= \overline{x} \\ \alpha\beta^2 &= s^2 \end{aligned}$$

pa je

$$\alpha = \frac{\bar{x}^2}{s^2}, \beta = \frac{s^2}{\bar{x}}.$$

Dakle, procjenitelj za α metodom momenata je $\frac{\overline{X}_n^2}{S_n^2}$, a procjenitelj za β je $\frac{S_n^2}{\overline{X}_n}$.

△

4.3 Pouzdani intervali za parametre normalne razdiobe

Definicija: Neka su $L_n = l_n(X_1, \dots, X_n)$ i $D_n = d_n(X_1, \dots, X_n)$ statistike slučajnog uzorka $X_1, \dots, X_n$. Za $[L_n, D_n]$ kažemo da je $(1-\alpha) \cdot 100\%$ **pouzdani interval** za parametar τ ako vrijedi

$$\mathbb{P}(L_n \leq \tau \leq D_n) \geq 1 - \alpha, \quad \alpha \in \langle 0, 1 \rangle.$$

Napomena: Neka su $X_i \sim N(\mu, \sigma^2), i = 1, \dots, n$ nezavisne slučajne varijable. Tada je:

$$\frac{\overline{X}_n - \mu}{\sigma} \cdot \sqrt{n} \sim N(0,1) \tag{4.1}$$

$$\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2(n-1) \tag{4.2}$$

$$\frac{\overline{X}_n - \mu}{S_n} \cdot \sqrt{n} \sim t(n-1) \tag{4.3}$$

$$\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2} \sim \chi^2(n) \tag{4.4}$$

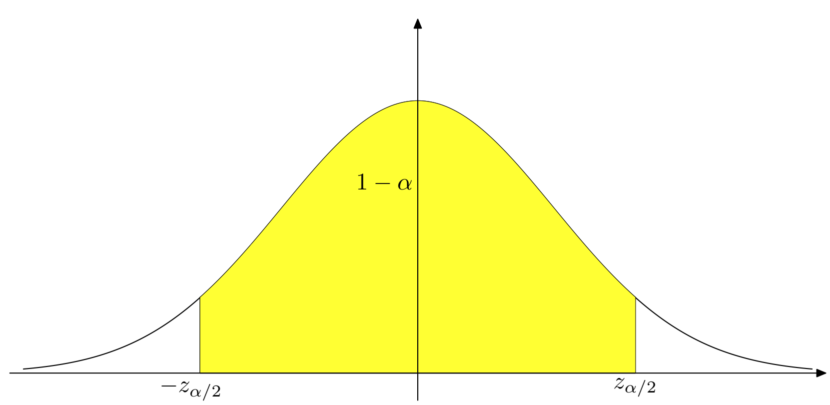
Zadatak 4.8 Neka je $X_1, \dots, X_{100}$ slučajni uzorak iz normalne razdiobe s varijancom 8 i nepoznatim očekivanjem μ. Aritmetička sredina uzorka je $\bar{x} = 42.7$. Nađite 95% pouzdani interval za parametar μ.

Rješenje: Budući da je $X_1, \dots, X_{100} \sim N(\mu, 8)$, prema 4.1 je

$$Z := \frac{\overline{X}_{100} - \mu}{\sqrt{8}} \cdot \sqrt{100} \sim N(0,1).$$

Iz $1 - \alpha = 0.95$ dobivamo danu pouzdanost $\alpha = 0.05$. Iz tablice normalne distribucije odredimo $z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$ za koji vrijedi

$$\mathbb{P}(-z_{0.025} \leq Z \leq z_{0.025}) = 0.95.$$

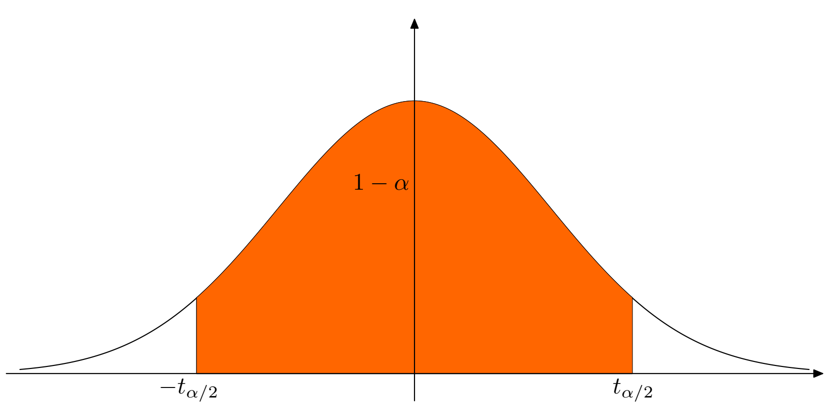


Stoga je
$$\mathbb{P}\left(-1.96 \leq \frac{\overline{X}_{100} - \mu}{\sqrt{8}} \cdot \sqrt{100} \leq 1.96\right) = 0.95,$$
odnosno
$$\mathbb{P}\left(\overline{X}_{100} - 1.96 \cdot \frac{\sqrt{8}}{\sqrt{100}} \leq \mu \leq \overline{X}_{100} + 1.96 \cdot \frac{\sqrt{8}}{\sqrt{100}}\right) = 0.95.$$
Dakle, 95% pouzdani interval za parametar μ je
$$\left[\overline{X}_{100} - 1.96 \cdot \frac{\sqrt{8}}{\sqrt{100}}, \overline{X}_{100} + 1.96 \cdot \frac{\sqrt{8}}{\sqrt{100}}\right]$$
Aritmetička sredina uzorka je $\bar{x} = 42.7$ pa je procjena 95% pouzdanog intervala za μ na osnovi opaženog uzorka
$$\left[42.7 - 1.96 \cdot \frac{\sqrt{8}}{\sqrt{100}}, 42.7 + 1.96 \cdot \frac{\sqrt{8}}{\sqrt{100}}\right] = [42.14, 43.25].$$

$\triangle$

Zadatak 4.9 Razvijena je nova slitina metala za izgradnju svemirskih letjelica. Izvršeno je 15 mjerenja koeficijenta napetosti te je izračunata sredina uzorka $\bar{x} = 39.3$ i standardna devijacija $s = 2.6$. Pretpostavljamo da je mjerena veličina normalno distribuirana. Nađite 90% pouzdani interval za očekivanje populacije.

Rješenje: Budući da je $X_1, \ldots, X_{15} \sim N(\mu, \sigma^2)$, prema 4.3 je
$$T := \frac{\overline{X}_{15} - \mu}{S_{15}} \cdot \sqrt{15} \sim t(14).$$
Iz $1 - \alpha = 0.90$ dobivamo danu pouzdanost $\alpha = 0.01$. Iz tablice t-distribucije odredimo $t_{\frac{\alpha}{2}}(n-1) = t_{0.05}(14) = 1.761$ za koji vrijedi
$$\mathbb{P}(-t_{0.05}(14) \leq T \leq t_{0.05}(14)) = 0.90.$$



Stoga je

$$\mathbb{P}\left(-1.761 \leq \frac{\overline{X}_{15} - \mu}{S_{15}} \cdot \sqrt{15} \leq 1.761\right) = 0.90,$$

odnosno

$$\mathbb{P}\left(\overline{X}_{15} - 1.761 \cdot \frac{S_{15}}{\sqrt{15}} \leq \mu \leq \overline{X}_{15} + 1.761 \cdot \frac{S_{15}}{\sqrt{15}}\right) = 0.90.$$

Dakle, 90% pouzdani interval za parametar μ je

$$\left[\overline{X}_{15} - 1.761 \cdot \frac{S_{15}}{\sqrt{15}}, \overline{X}_{15} + 1.761 \cdot \frac{S_{15}}{\sqrt{15}}\right]$$

Uvrštavanjem aritmetičke sredine uzorka $\bar{x} = 39.3$ i standardne devijacije uzorka $s = 2.6$ dobivamo da je procjena 90% pouzdanog intervala za μ na osnovi opaženog uzorka jednaka

$$\left[39.3 - 1.761 \cdot \frac{2.6}{\sqrt{15}}, 39.3 + 1.761 \cdot \frac{2.6}{\sqrt{15}}\right] = [38.12, 40.48].$$

△

Zadatak 4.10 Na slučajnom uzorku od 25 elemenata iz normalne razdiobe izračunata je varijanca uzorka $s^2 = 12.5$. Procijenite 90% pouzdani interval za varijancu populacije.

Rješenje: Budući da je $X_1, \ldots, X_{25} \sim N(\mu, \sigma^2)$, prema 4.2 je

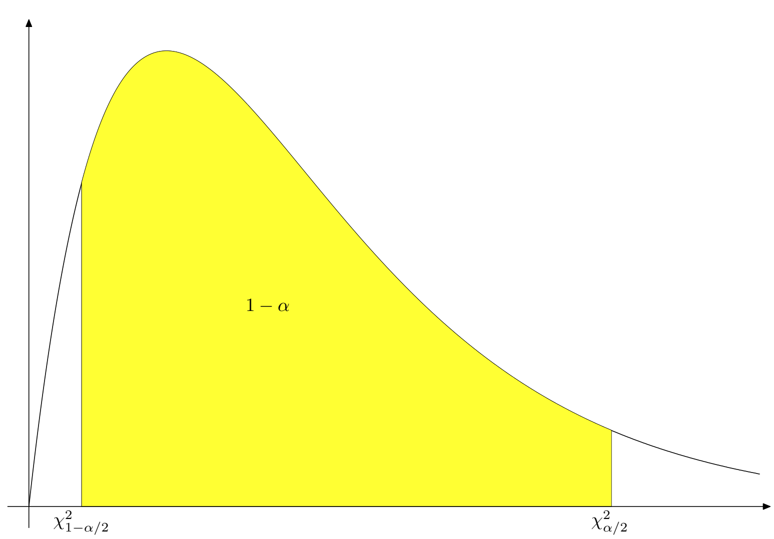
$$V := \frac{24 \cdot S_{24}^2}{\sigma^2} \sim \chi^2(24).$$

Za danu pouzdanost $\alpha = 0.01$, iz tablice χ^2-distribucije odredimo

$$\chi^2_{\frac{\alpha}{2}}(n-1) = \chi^2_{0.05}(24) = 36.415 \quad \text{i}$$
$$\chi^2_{1-\frac{\alpha}{2}}(n-1) = \chi^2_{0.95}(24) = 13.848$$

za koje vrijedi

$$\mathbb{P}(\chi^2_{0.95}(24) \leq V \leq \chi^2_{0.05}(24)) = 0.90.$$



Stoga je

$$\mathbb{P}\left(13.848 \leq \frac{24 \cdot S_{24}^2}{\sigma^2} \leq 36.415\right) = 0.90,$$

odnosno

$$\mathbb{P}\left(\frac{24 \cdot S_{24}^2}{36.415} \leq \sigma^2 \leq \frac{24 \cdot S_{24}^2}{13.848}\right) = 0.90.$$

Dakle, 90% pouzdani interval za parametar μ je

$$\left[\frac{24 \cdot S_{24}^2}{36.415}, \frac{24 \cdot S_{24}^2}{13.848}\right]$$

Uvrštavanjem varijance uzorka $s^2 = 12.5$ dobivamo da je procjena 90% pouzdanog intervala za σ^2 na osnovi opaženog uzorka jednaka

$$\left[\frac{24 \cdot 12.5}{36.415}, \frac{24 \cdot 12.5}{13.848}\right] = [8.24, 21.66].$$

$\triangle$

Napomena:

(a) Za $n \to \infty$ t-distribucija se asimptotski ponaša kao jedinična normalna distribucija pa za velike n (u praksi $n \geq 31$) umjesto $t_\alpha(n)$ možemo uzeti z_α iz tablice normalne distribucije.

(b) Neka je alpha=0.025. U R-u z_α dobijemo naredbom

```
> qnorm(1-alpha)
[1] 1.959964
```

Neka je alpha=0.05 i $n = 14$. Tada $t_\alpha(n)$ dobivamo naredbom

```
> qt(1-alpha,n)
[1] 1.76131
```

Neka je sada alpha=0.05 i $n = 24$. Naredbe za $\chi^2_\alpha(n)$ i $\chi^2_{1-\alpha}(n)$ su redom:

```
> qchisq(1-alpha,n)
[1] 36.41503
> qchisq(alpha,n)
[1] 13.84843
```

4.4 Aproksimativni pouzdani intervali

Definicija: Niz statistika $\{Z_n\colon n \in \mathbb{N}\}$ je asimptotski normalan ako konvergira po distribuciji slučajnoj varijabli $Z \sim N(0,1)$, odnosno ako je

$$\lim_{n\to\infty} \mathbb{P}(Z_n \leq x) = \Phi(x), \quad \forall x \in \mathbb{R}.$$

Pišemo: $Z_n \sim AN(0,1)$.

Napomena: Neka je $X_1,\dots,X_n$ slučajni uzorak s konačnim očekivanjem $\mu = \mathbb{E}X_1$ i varijancom $\sigma^2 = \operatorname{Var} X_1$, te neka je $S_n := X_1 + \cdots + X_n$. Prema centralnom graničnom teoremu vrijedi

$$\frac{S_n - \mathbb{E}S_n}{\sqrt{\operatorname{Var} S_n}} \sim AN(0,1),$$

odakle slijedi

$$\frac{\overline{X}_n - \mu}{\sigma} \cdot \sqrt{n} \sim AN(0,1). \tag{4.5}$$

Definicija: Procjenitelj T_n je **(slabo) konzistentan** ako je

$$(\mathbb{P}) \lim_n T_n = \tau, \quad \text{tj.} \quad \forall \varepsilon > 0 \ \lim_n \mathbb{P}(|T_n - \tau| > \varepsilon) = 0.$$

Napomena:

(a) S_n je konzistentan procjenitelj za standardnu devijaciju σ.

(b) Ako je $\hat{\sigma}_n$ konzistentan procjenitelj za standardnu devijaciju σ, tada vrijedi

$$\frac{\overline{X}_n - \mu}{\hat{\sigma}_n} \cdot \sqrt{n} \sim AN(0,1). \tag{4.6}$$

Zadatak 4.11 Nađite 95% pouzdani interval za nepoznati parametar θ iz $\operatorname{Exp}(\theta)$, $\theta > 0$, modela ako je mjeren uzorak duljine $n = 190$ i $\sum_{i=1}^{190} x_i = 40549$ dana.

Rješenje: Budući da je $X_1,\dots,X_{190} \sim \operatorname{Exp}(\theta)$, vrijedi $\mu = \mathbb{E}X_1 = \frac{1}{\theta}$ i $\sigma^2 = \operatorname{Var} X_1 = \frac{1}{\theta^2}$. Prema 4.5 je

$$Z := \frac{\overline{X}_{190} - \mu}{\sigma} \cdot \sqrt{190} \sim AN(0,1).$$

Za danu pouzdanost $\alpha = 0.05$ je $z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$ i vrijedi $\mathbb{P}(-z_{0.025} \leq Z \leq z_{0.025}) \approx 0.95$. Stoga je

$$\mathbb{P}\left(-1.96 \leq \frac{\overline{X}_{190} - \frac{1}{\theta}}{\sqrt{\frac{1}{\theta^2}}} \cdot \sqrt{190} \leq 1.96\right) \approx 0.95,$$

odnosno
$$\mathbb{P}\left(\frac{-1.96/\sqrt{190}+1}{\overline{X}_{190}} \leq \theta \leq \frac{1.96/\sqrt{190}+1}{\overline{X}_{190}}\right) \approx 0.95.$$

Dakle, aproksimativni 95% pouzdani interval za parametar θ je

$$\left[\frac{-1.96/\sqrt{190}+1}{\overline{X}_{190}}, \frac{1.96/\sqrt{190}+1}{\overline{X}_{190}}\right]$$

Aritmetička sredina uzorka je

$$\bar{x} = \frac{\sum_{i=1}^{190} x_i}{n} = \frac{40549}{190} = 213.416$$

pa je procjena tog pouzdanog intervala za θ na osnovi opaženog uzorka

$$\left[\frac{-1.96/\sqrt{190}+1}{213.416}, \frac{1.96/\sqrt{190}+1}{213.416}\right] = [0.00402, 0.00535].$$

△

Zadatak 4.12 Obavljeno je 100 mjerenja težine čokoladnih pločica od 100 grama. Dobiveni su podaci:

tezina u g	frekvencija
92-93	5
94-95	7
96-97	21
98-99	33
100-101	19
102-103	11
104-105	4

Nađite 95% pouzdani interval za očekivanu težinu čokoladnih pločica.

Rješenje: Neka je μ očekivanje distribucije težine čokoladnih pločica. Budući da je $n = 100$ velik, prema prethodnoj napomeni vrijedi

$$Z := \frac{\overline{X}_{100} - \mu}{S_{100}} \cdot \sqrt{100} \sim AN(0,1).$$

Za danu pouzdanost $\alpha = 0.05$ je $z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$ i vrijedi $\mathbb{P}(-z_{0.025} \leq Z \leq z_{0.025}) \approx 0.95$. Stoga je

$$\mathbb{P}\left(-1.96 \leq \frac{\overline{X}_{100} - \mu}{S_{100}} \cdot \sqrt{100} \leq 1.96\right) \approx 0.95,$$

odnosno

$$\mathbb{P}\left(\overline{X}_{100} - 1.96 \cdot \frac{S_{100}}{\sqrt{100}} \leq \mu \leq \overline{X}_{100} + 1.96 \cdot \frac{S_{100}}{\sqrt{100}}\right) \approx 0.95.$$

Dakle, aproksimativni 95% pouzdani interval za parametar μ je

$$\left[\overline{X}_{100} - 1.96 \cdot \frac{S_{100}}{\sqrt{100}}, \overline{X}_{100} + 1.96 \cdot \frac{S_{100}}{\sqrt{100}}\right]$$

Aritmetičku sredinu i varijancu uzorka računamo kao u zadatku 1.5. Širina razreda je $c = 2$, a referentna vrijednost $\bar{x}_0 = 98.5$. Tablica frekvencija glasi:

i	I_i	f_i	$\bar{x}_i$	$d_i = (\bar{x}_i - \bar{x}_0)/c$	$f_i d_i$	$f_i d_i^2$
1	$[91.5, 93.5\rangle$	5	92.5	-3	-15	45
2	$[93.5, 95.5\rangle$	7	94.5	-2	-14	28
3	$[95.5, 97.5\rangle$	21	96.5	-1	-21	21
4	$[97.5, 99.5\rangle$	33	98.5	0	0	0
5	$[99.5, 101.5\rangle$	19	100.5	1	19	19
6	$[101.5, 103.5\rangle$	11	102.5	2	22	44
7	$[103.5, 105.5\rangle$	4	104.5	3	13	36
Σ		100			3	193

Sada je

$$\bar{x} = c \cdot \frac{1}{n}\sum_{i=1}^{k} f_i d_i + \bar{x}_0 = 2 \cdot \frac{3}{100} + 98.5 = 98.56$$

$$s^2 = c^2 \left[\frac{1}{n}\sum_{i=1}^{k} f_i d_i^2 - \left(\frac{1}{n}\sum_{i=1}^{k} f_i d_i\right)^2\right] = 4 \cdot \left[\frac{193}{100} - \left(\frac{3}{100}\right)^2\right] = 7.72 \quad \Rightarrow \quad s = 2.79$$

pa je procjena aproksimativnog 95% pouzdanog intervala za parametar μ na osnovi opaženog uzorka

$$\left[98.56 - 1.96 \cdot \frac{2.79}{\sqrt{100}}, 98.56 + 1.96 \cdot \frac{2.79}{\sqrt{100}}\right] = [98.01, 99.11].$$

$\triangle$

Zadatak 4.13 U 400 izvedenih pokusa događaj A nastupio je 280 puta. Procijenite 95% pouzdani interval za $p = \mathbb{P}(A)$.

Rješenje: Budući da je $X_1, \ldots X_{400}$ slučajni uzorak iz Bernoullijevog modela s parametrom p, slijedi $\mu = \mathbb{E}X_1 = p$ i $\sigma^2 = \operatorname{Var} X_1 = p(1-p)$. Dana pouzdanost je $\alpha = 0.05$ i $z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$.

1. Prema 4.5 je

$$\frac{\overline{X}_{400} - \mu}{\sigma} \cdot \sqrt{400} \sim AN(0,1)$$

pa vrijedi

$$\mathbb{P}\left(\left|\frac{\overline{X}_{400} - p}{\sqrt{p(1-p)}} \cdot \sqrt{400}\right| \leq 1.96\right) \approx 0.95.$$

Kvadriranjem dobivamo da uz 95% pouzdanosti vrijedi

$$(\overline{X}_{400} - p)^2 \cdot 400 \leq 1.96^2 \cdot p(1-p),$$

odnosno

$$(\overline{X}_{400}^2 - 2\overline{X}_{400}p + p^2) \cdot 400 - 1.96^2(p - p^2) \leq 0.$$

Uvrštavanjem $\bar{x} = 280/400 = 0.7$ i rješavanjem kvadratne jednadžbe

$$p^2(400 + 1.96^2) - p(800 \cdot 0.7 + 1.96^2) + 400 \cdot 0.7^2 \leq 0,$$

dobivamo da je procjena 95% pouzdanog intervala za p jednaka $[0.653, 0.743]$.

2. $\overline{X}_n$ je konzistentan procjenitelj za p pa je $\hat{\sigma}_n = \sqrt{\overline{X}_n(1 - \overline{X}_n)}$ konzistentan procjenitelj za $\sigma = \sqrt{p(1-p)}$. Prema 4.6 je

$$\frac{\overline{X}_{400} - \mu}{\hat{\sigma}_{400}} \cdot \sqrt{400} \sim AN(0,1),$$

pa je

$$\mathbb{P}\left(-1.96 \leq \frac{\overline{X}_{400} - p}{\hat{\sigma}_{400}} \cdot \sqrt{400} \leq 1.96\right) \approx 0.95,$$

odnosno

$$\mathbb{P}\left(\overline{X}_{400} - 1.96 \cdot \frac{\hat{\sigma}_{400}}{\sqrt{400}} \leq p \leq \overline{X}_{400} + 1.96 \cdot \frac{\hat{\sigma}_{400}}{\sqrt{400}}\right) \approx 0.95,$$

odakle uvrštavanjem $\bar{x} = 0.7$ i $\hat{\sigma}_{400} = \sqrt{0.7 \cdot 0.3}$ dobivamo da je procjena 95% pouzdanog intervala za p jednaka $[0.655, 0.745]$.

$\triangle$

Zadaci za vježbu

4.14 Neka je $X \sim U(0, \tau), \tau > 0$. Pokažite da je $2\overline{X}_n$ nepristrani procjenitelj za τ.

4.15 Neka je $X_1, \ldots, X_n$ sl. uzorak iz Bernoullijevog modela s parametrom $\mathbb{P}(X_i = 1) = p$ i $n \geq 3$. Nađite nepristrani procjenitelj za: (a) $\tau(p) = p(1-p)$ i (b) $\tau(p) = p^2(1-p)$.

4.16 Neka je $X_1, \ldots, X_n$ sl. uzorak iz Bernoullijevog modela s parametrom $\mathbb{P}(X_i = 1) = p \in \langle 0, 1 \rangle$. Nađite MLE za p.

4.17 Neka je $X \sim U(\tau, \tau + 2), \tau > 0$. Odredite sve procjene za τ metodom ML.

4.18 Neka je $X_1, \ldots, X_n$ sl. uzorak iz eksponencijalnog modela s parametrom $\lambda > 0$.

(a) Nađite MLE procjenitelj za λ.

(b) Nađite bar jedan nepristrani procjenitelj za λ.

(c) Nađite bar jedan nepristrani procjenitelj za λ^2.

4.19 Nađite MLE procjenitelj parametra $\theta = (\alpha, \beta)$ za model $\Gamma(\alpha, \beta), \alpha, \beta > 0$. Uputa: funkcija $f : \mathbb{R}_+ \to \mathbb{R}_+$ definirana sa $f(x) = \ln x - \frac{\Gamma'(x)}{\Gamma(x)}$ je bijekcija.

4.20 Neka je $X_1, \ldots, X_n$ sl. uzorak iz Bernoullijevog modela s parametrom $\mathbb{P}(X_i = 1) = p$. Nađite procjenitelj metodom momenata za: (a) p, (b) $p(1-p)$ i (c) $\cos p$.

4.21 Da bi se ispitala čvrstoća jedne vrste čelika, obavljeno je mjerenje prijelomne sile na 48 epruveta. Pretpostavljamo da je prijelomna sila normalno distribuirana. Rezultati mjerenja daju $\overline{x} = 70.2$ i $s = 10.1$. Nađite 95% pouzdani interval za parametar μ, očekivanje populacije.

4.22 Promatran je slučajan uzorak od 100 stabala između njih 1546 na nekoj farmi i izmjerena aritmetička sredina $\overline{x} = 59.22$ te standardna devijacija $s = 10.11$. Odredite 95% pouzdani interval za srednju visinu svih stabala na toj farmi.

4.23 Na uzorku od 20 elemenata dobivene su sljedeće vrijednosti normalne varijable:

72 74 73 73 70 74 72 76 72 71 71 74 74 76 73 76 70 75 71 73

Procijenite 80% p. i. za: (a) očekivanje populacije i (b) standardnu devijaciju populacije.

4.24 Nađite 95% pouzdani interval za nepoznati parametar λ Poissonovog modela $P(\lambda)$ ako je mjerenje slučajnog uzorka duljine $n = 621$ dalo iznos od $\sum_{i=1}^{621} x_i = 1522$.

4.25 U 40 bacanja novčića 24 puta je palo pismo. Odredite 95% pouzdani interval za očekivani broj pisama u neograničenom broju bacanja novčića.

Rješenja: **4.15.** (a) S_n^2, (b) $\mathbb{1}_{\{X_1=1, X_2=1, X_3=0\}}$. **4.16.** $\overline{X}_n$ **4.17.** $[x_{(n)} - 2, x_{(1)}]$ **4.18.** Uz oznaku $Y = X_1 + \cdots + X_n$: (a) n/Y, (b) $(n-1)/Y$, (c) $(n-1)(n-2)/Y^2$. **4.19.** $\hat{\alpha} = f^{-1}(\ln \overline{X}_n - (\sum_{i=1}^n \ln X_i)/n), \hat{\beta} = \overline{X}_n / \hat{\alpha}$ **4.20.** (a) $\overline{X}_n$, (b) $\overline{X}_n(1 - \overline{X}_n)$, (c) $\cos \overline{X}_n$. **4.21.** $[67.3, 73.1]$ **4.22.** $[57.24, 61.20]$ **4.23.** (a) $[72.43, 73.56]$, (b) $[1.58, 2.41]$. **4.24.** $[2.33, 2.57]$ **4.25.** $[0.45, 0.74]$

5

Testiranje statističkih hipoteza

Neka je $X_1, \ldots, X_n$ slučajni uzorak iz populacije s razdiobom

$$f(x|\theta),\ \theta \in \Theta \subseteq \mathbb{R}^d$$

i neka je za opaženi uzorak $x_1, \ldots, x_n$ definirana funkcija vjerodostojnosti

$$\begin{aligned} &L\colon \Theta \to \mathbb{R}, \\ &L(\theta|\mathbf{x}) = \prod_{i=1}^{n} f(x_i|\theta), \end{aligned}$$

gdje je $\mathbf{x} = (x_1, \ldots, x_n)$. Promotrimo test $\tau\colon \mathbb{R}^n \to \{0, 1\}$ za testiranje hipoteza

$$\begin{aligned} &H_0\colon \theta \in \Theta_0 \\ &H_1\colon \theta \in \Theta_1 \end{aligned}$$

s kritičnim područjem $C = \tau^{-1}(\{1\})$. Neka je $\mathbf{X} = (X_1, \ldots, X_n)$.

Definicija: Preslikavanje $\gamma\colon \Theta \to [0, 1]$ definirano s

$$\gamma(\theta) = \mathbb{P}(\mathbf{X} \in C|\theta) = \int_C L(\theta|\mathbf{x})\, dx$$

zovemo **jakost testa**.

Preslikavanje $\alpha\colon \Theta_0 \to [0, 1]$ definirano s

$$\alpha(\theta) = \gamma(\theta) = \mathbb{P}(\mathbf{X} \in C|\theta),$$

zovemo **vjerojatnost pogreške 1. vrste**, dok preslikavanje $\beta\colon \Theta_1 \to [0, 1]$ definirano s

$$\beta(\theta) = 1 - \gamma(\theta) = \mathbb{P}(\mathbf{X} \notin C|\theta),$$

zovemo **vjerojatnost pogreške 2. vrste**. **Značajnost testa** se definira s

$$\alpha_\tau = \sup_{\theta \in \Theta_0} \alpha(\theta).$$

Kažemo da test ima razinu značajnosti α, ako je $\alpha_\tau \leq \alpha$.

Definicija: Test τ je **uniformno najjači** ako za sve druge testove τ' takve da je $\alpha_{\tau'} \leq \alpha_\tau$ vrijedi

$$\gamma_{\tau'}(\theta) \leq \gamma_\tau(\theta), \ \forall \theta \in \Theta.$$

Uniformno najjače testove oblika

$$H_0 \colon \theta = \theta_0$$
$$H_1 \colon \theta = \theta_1$$

je ponekad moguće naći koristeći sljedeći rezultat.

Teorem: (**Neyman-Pearsonova lema**) Neka je $\alpha \in (0,1)$ i neka je $k \geq 0$ takav da za skup

$$C = \{\mathbf{x} \in \mathbb{R}^n \colon L(\theta_0|\mathbf{x}) \leq k\, L(\theta_1|\mathbf{x})\} \tag{5.1}$$

vrijedi

$$\int_C L(\theta_0|\mathbf{x})\, dx = \alpha.$$

Ako je $B \subseteq \mathbb{R}^n$ takav da je

$$\int_B L(\theta_0|\mathbf{x})\, dx \leq \alpha,$$

onda je

$$\int_B L(\theta_1|\mathbf{x})\, dx \leq \int_C L(\theta_1|\mathbf{x})\, dx.$$

$\triangle$

Napomena: Iz Neyman-Pearsonove leme slijedi da ako zadamo razinu značajnosti $\alpha \in (0,1)$, onda je sa (5.1) definirano kritično područje uniformno najjačeg testa.

Zadatak 5.1 Neka je $X_1, \ldots, X_n$ slučajni uzorak iz populacije s razdiobom $N(\mu, \sigma^2)$, pri čemu je σ^2 poznato. Pokažite da je uniformno najjači test značajnosti $\alpha \in (0,1)$ za testiranje hipoteza

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

dan testnom statistikom

$$Z = \frac{\overline{X}_n - \mu_0}{\sigma}\sqrt{n}$$

i kritičnim područjem

$$(-\infty, -z_\alpha].$$

Rješenje:

(a) Neka je $\mu_1 < \mu_0$. Pronađimo uniformno najjači test za testiranje hipoteza

$$H_0\colon \mu = \mu_0$$
$$H_1\colon \mu = \mu_1$$

uz razinu značajnosti $\alpha \in (0,1)$. Neka je $x_1, \ldots, x_n$ opaženi uzorak i $\mathbf{x} = (x_1, \ldots, x_n)$. Funkcija vjerodostojnosti je

$$L(\mu|\mathbf{x}) = \frac{1}{(2\pi)^{d/2}\sigma^n} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2}.$$

Iz Neyman-Pearsonove leme slijedi da je kritično područje uniformno najjačeg testa oblika

$$C = \{\mathbf{x} \in \mathbb{R}^d \colon L(\mu_0|\mathbf{x}) \leq kL(\mu_1|\mathbf{x})\},$$

za neki $k > 0$. Računamo:

$$\begin{aligned}\frac{L(\mu_0|\mathbf{x})}{L(\mu_1|\mathbf{x})} &= e^{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n (x_i-\mu_1)^2 - \sum_{i=1}^n (x_i-\mu_0)^2\right)} \\ &= e^{\frac{1}{2\sigma^2}(2n(\mu_0-\mu_1)\overline{x} + n(\mu_1^2-\mu_0^2))}\end{aligned}$$

Zbog $\mu_0 - \mu_1 > 0$ vrijedi

$$\begin{aligned}\mathbf{x} \in C \iff \frac{L(\mu_0|\mathbf{x})}{L(\mu_1|\mathbf{x})} \leq k \iff 2n(\mu_0-\mu_1)\overline{x} + n(\mu_1^2 - \mu_0^2) \leq 2\sigma^2 \ln k \\ \iff \overline{x} \leq \frac{\sigma^2 \ln k}{n(\mu_0-\mu_1)} + \frac{\mu_0+\mu_1}{2} =: k_1\end{aligned}$$

pa je kritično područje oblika

$$C = \{\mathbf{x} \in \mathbb{R}^n \colon \overline{x} \leq k_1\}.$$

Mora vrijediti

$$\alpha = \mathbb{P}(\mathbf{X} \in C | H_0) = \mathbb{P}(\overline{X}_n \leq k_1 | H_0) = \mathbb{P}(\frac{\overline{X}_n - \mu_0}{\sigma}\sqrt{n} \leq \frac{k_1 - \mu_0}{\sigma}\sqrt{n} | H_0)$$
$$= \mathbb{P}(\frac{\overline{X}_n - \mu_0}{\sigma}\sqrt{n} \leq -z_\alpha | H_0),$$

jer je uz $\mu = \mu_0$

$$\frac{\overline{X}_n - \mu_0}{\sigma}\sqrt{n} \sim N(0,1). \tag{5.2}$$

Dakle, $\frac{k_1-\mu_0}{\sigma}\sqrt{n} = -z_\alpha$ pa je

$$C = \{\mathbf{x} \in \mathbb{R}^n : \overline{x} \leq k_1\} = \{\mathbf{x} \in \mathbb{R}^n : \frac{\overline{x} - \mu_0}{\sigma}\sqrt{n} \leq -z_\alpha\}. \tag{5.3}$$

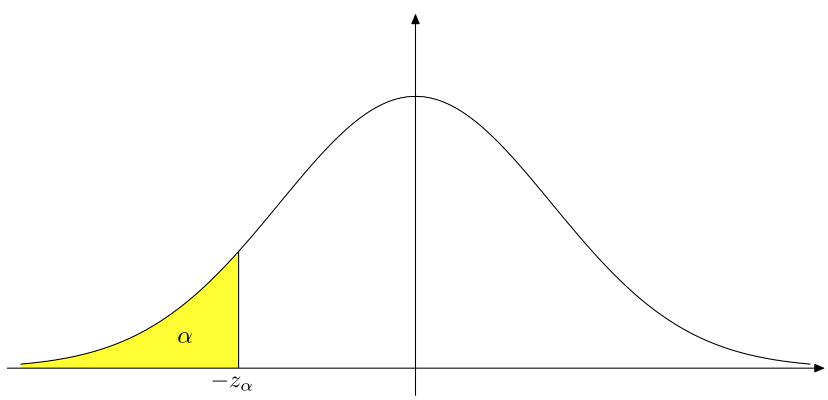
(b) Promotrimo sada test

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

s kritičnim područjem C. Trebamo još pokazati da je to uniformno najjači test. Neka je $B \subseteq \mathbb{R}^n$ takav da je $\mathbb{P}(\mathbf{X} \in B | \mu_0) = \alpha$. Iz (a) slijedi da za $\mu < \mu_0$ vrijedi

$$\gamma'(\mu) = \mathbb{P}(\mathbf{X} \in B | \mu) \leq \mathbb{P}(\mathbf{X} \in C | \mu) = \gamma(\mu)$$

pa je test s kritičnim područjem C uniformno najjači.



Iz (5.2) i (5.3) slijedi da možemo uzeti testnu statistiku

$$Z = \frac{\overline{X}_n - \mu_0}{\sigma}\sqrt{n}$$

i kritično područje

$$(-\infty, -z_\alpha].$$

$\triangle$

Zadatak 5.2 Neka je $X_1, \ldots, X_n$ slučajni uzorak iz populacije s razdiobom $N(\mu, \sigma^2)$, pri čemu je σ^2 poznato. Pokažite da je uniformno najjači test značajnosti $\alpha \in (0,1)$ za testiranje hipoteza

$$\begin{aligned} H_0 &\colon \mu \leq \mu_0 \\ H_1 &\colon \mu > \mu_0 \end{aligned}$$

dan testnom statistikom

$$Z = \frac{\overline{X}_n - \mu_0}{\sigma} \sqrt{n}$$

i kritičnim područjem

$$[z_\alpha, \infty).$$

Rješenje:

(a) Uzmimo $\mu_1 \leq \mu_0$ i $\mu_2 > \mu_0$. Promotrimo prvo test za testiranje hipoteza

$$\begin{aligned} H_0 &\colon \mu = \mu_1 \\ H_1 &\colon \mu = \mu_2 \end{aligned}$$

Slično kao u Zadatku 5.1 se pokaže da je kritično područje uniformno najjačeg testa dano s

$$C = \{x \in \mathbb{R}^n \colon \overline{x} \geq k_1\},$$

gdje je

$$\frac{k_1 - \mu_0}{\sigma} \sqrt{n} = z_\alpha.$$

(b) Neka je $B \subseteq \mathbb{R}^n$ takav da je

$$\sup_{\mu \leq \mu_0} \mathbb{P}(\mathbf{X} \in B | \mu_0) \leq \alpha.$$

Tada iz (a) dobijemo

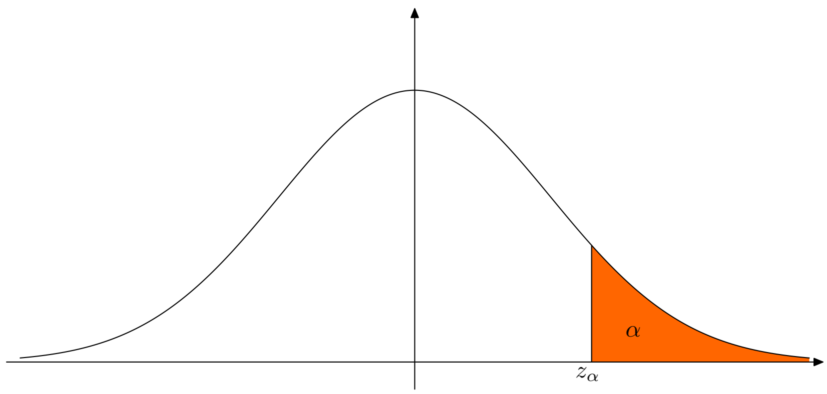
$$\mathbb{P}(\mathbf{X} \in B | \mu) \leq \mathbb{P}(\mathbf{X} \in C | \mu), \quad \text{za sve } \mu > \mu_0.$$

Dakle,

$$\begin{aligned} &\alpha'(\mu) \leq \alpha = \mathbb{P}(\mathbf{X} \in C | H_0), \ \forall \mu \leq \mu_0 \\ &\gamma'(\mu) = \mathbb{P}(\mathbf{X} \in B | \mu) \leq \mathbb{P}(\mathbf{X} \in C | \mu) = \gamma(\mu), \ \forall \mu > \mu_0, \end{aligned}$$

pa je C kritično područje uniformno najjačeg testa za testiranje hipoteza

$$\begin{aligned} H_0&\colon \mu \leq \mu_0 \\ H_1&\colon \mu > \mu_0. \end{aligned}$$



Dakle, uz testnu statistiku

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

imamo kritično područje

$$[z_\alpha, \infty).$$

△

Promotrimo test za testiranje hipoteza

$$\begin{aligned} H_0&\colon \theta \in \Theta_0 \\ H_1&\colon \theta \in \Theta_1 \end{aligned}$$

i pretpostavimo da postoje $\theta_0 \in \Theta_0$ i $\hat{\theta} \in \Theta$ takvi da je

$$L(\theta_0|\mathbf{x}) = \max_{\theta\in\Theta_0} L(\theta|\mathbf{x}) \ \text{ i } \ L(\hat{\theta}|\mathbf{x}) = \max_{\theta\in\Theta} L(\theta|\mathbf{x}).$$

Definicija: **Omjer vjerodostojnosti** je funkcija $\lambda\colon \mathbb{R}^n \to \mathbb{R}$ definirana s

$$\lambda(\mathbf{x}) = \frac{\max_{\theta\in\Theta_0} L(\theta|\mathbf{x})}{\max_{\theta\in\Theta} L(\theta|\mathbf{x})} = \frac{L(\theta_0|\mathbf{x})}{L(\hat{\theta}|\mathbf{x})}.$$

Ako je kritično područje testa oblika

$$C = \{\mathbf{x} \in \mathbb{R}^n\colon \lambda(\mathbf{x}) \leq c\},$$

za neku konstantu $c \in (0,1)$, onda pripadni test zovemo **test omjera vjerodostojnosti**.

Zadatak 5.3 Neka je $X_1, \dots, X_n$ slučajni uzorak iz populacije s razdiobom $N(\mu, \sigma^2)$ pri čemu su oba parametra nepoznata. Pokažite da je test omjera vjerodostojnosti za testiranje hipoteza

$$\begin{aligned} H_0 &\colon \mu = \mu_0 \\ H_1 &\colon \mu \neq \mu_0 \end{aligned}$$

dan testnom statistikom

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

i kritičnim područjem

$$(-\infty, -t_{\alpha/2}(n-1)] \cup [t_{\alpha/2}(n-1), \infty).$$

Rješenje: Ovdje je $\Theta = \{(\mu, \sigma^2) \colon \mu \in \mathbb{R},\ \sigma^2 > 0\}$ i

$$\Theta_0 = \{(\mu_0, \sigma^2) \colon \sigma^2 > 0\}, \quad \Theta_1 = \Theta \setminus \Theta_0.$$

Neka je $x_1, \dots, x_n$ opaženi uzorak i $\mathbf{x} = (x_1, \dots, x_n)$. Znamo da je

$$\max_{(\mu,\sigma^2)\in\Theta} L((\mu, \sigma^2)|\mathbf{x}) = (\overline{x}, \frac{n-1}{n}s^2).$$

Također

$$\max_{(\mu,\sigma^2)\in\Theta_0} L((\mu, \sigma^2)|\mathbf{x}) = \max_{\sigma^2>0} L((\mu_0, \sigma^2)|\mathbf{x}) = (\mu_0, \frac{n-1}{n}s^2).$$

Stoga je

$$\lambda(\mathbf{x}) = \left(\frac{\sum_{i=1}^n (x_i - \overline{x})^2}{\sum_{i=1}^n (x_i - \mu_0)^2}\right)^{-\frac{n}{2}}$$

pa je

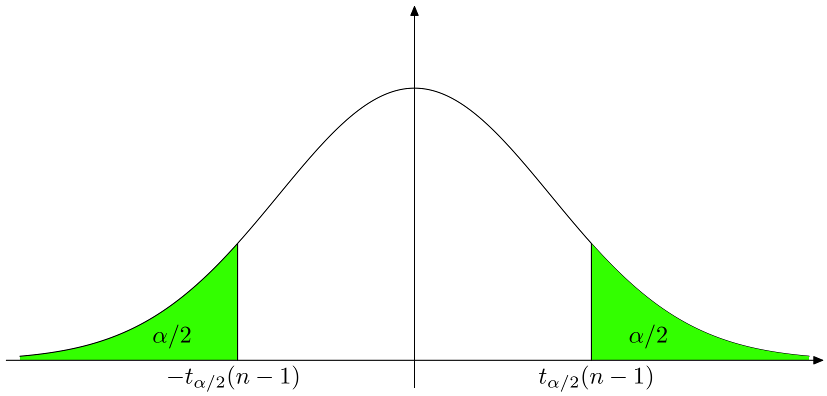
$$\begin{aligned} \mathbf{x} \in C \iff \left(\frac{\sum_{i=1}^n (x_i - \overline{x})^2}{\sum_{i=1}^n (x_i - \mu_0)^2}\right)^{-\frac{n}{2}} \le c &\iff \frac{n}{n-1}\frac{(\overline{x} - \mu_0)^2}{s^2} \ge c^{-\frac{2}{n}} - 1 \\ &\iff \left|\frac{\overline{x} - \mu_0}{s}\sqrt{n}\right| \ge c_0, \end{aligned}$$

gdje je

$$c_0 = \sqrt{(n-1)(c^{-\frac{2}{n}} - 1)} \ge 0.$$

Dakle,

$$C = \{\mathbf{x} \in \mathbb{R}^n \colon \left|\frac{\overline{x} - \mu_0}{s}\sqrt{n}\right| \ge c_0\}$$



pa možemo uzeti testnu statistiku

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

s kritičnim područjem

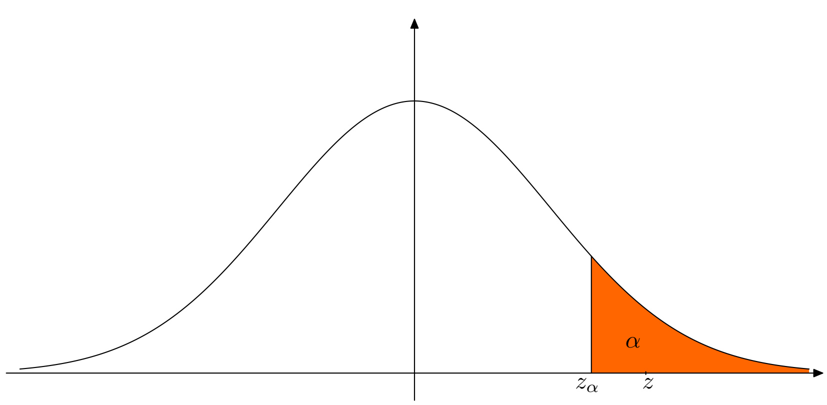
$$(-\infty, -t_{\alpha/2}(n-1)] \cup [t_{\alpha/2}(n-1), \infty).$$

$\triangle$

Zadatak 5.4 Neki proizvođač proizvodi sajle čija je izdržljivost u prosjeku jednaka 1800 kg uz standardnu devijaciju od 100 kg. Nedavno je proizvođač uveo novu tehniku proizvodnje i tvrdi da se na taj način mogu dobiti sajle veće izdržljivosti. Odabran je slučajni uzorak od 50 sajli proizvedenih novom tehnikom i izračunata je prosječna izdržljivost od 1850 kg. Uz pretpostavku da je izdržljivost sajli normalno distribuirana, može li se na nivou značajnosti od 1 % zaključiti da se novom tehnikom mogu dobiti izdržljivije sajle?

Rješenje: Testiramo hipoteze:

$$H_0 : \mu = 1800 \qquad \text{(nema promjene u izdržljivosti)}$$
$$H_1 : \mu > 1800 \qquad \text{(izdržljivost se povećala)}$$



Testna statistika je:

$$Z = \frac{\overline{X}_n - \mu}{\sigma}\sqrt{n} \overset{H_0}{\sim} N(0,1)$$

Iz tablice za normalnu razdiobu odredimo z_α takav da je $\mathbb{P}(Z \geq z_\alpha) = 0.01$ i dobijemo $z_\alpha = 2.33$.

$$\overline{x} = 1850,\ n = 50 \implies z = \frac{1850 - 1800}{100}\sqrt{50} = 3.536 > 2.33 = z_\alpha$$

$\implies$ odbacujemo hipotezu H_0 u korist hipoteze H_1 na nivou značajnosti od 1 %

Dakle, novom tehnikom se dobiju izdržljivije sajle.

$\triangle$

Zadatak 5.5 Mjerenjem mase 20 istovrsnih čokolada dobiveni su sljedeći rezultati u gramima:

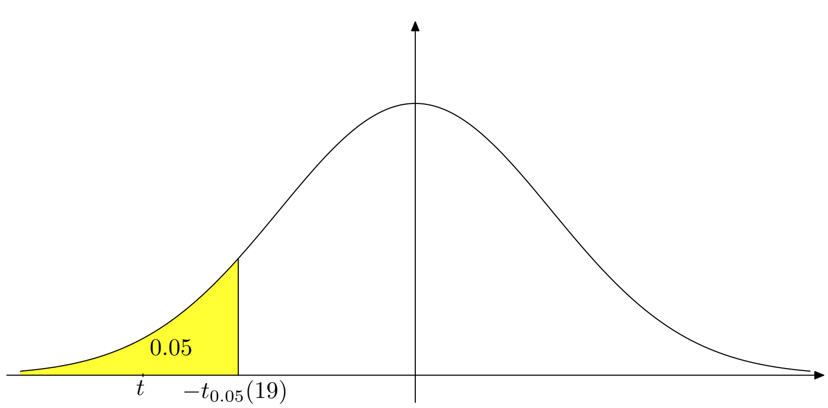
97	99	98	96	98	101	98	95	97	99
98	96	97	98	98	100	99	97	101	98

Pretpostavimo da se mase čokolada podvrgavaju normalnoj razdiobi. Ako na omotu čokolade piše da je njena masa 100 g, možemo li to zaključiti na nivou značajnosti od 5 % ?

Rješenje: Testiramo hipoteze:

$$H_0 : \mu = 100$$
$$H_1 : \mu < 100$$

Provodimo t-test, jer je populacija normalna s nepoznatom standardnom devijacijom.



Testna statistika je:

$$T = \frac{\overline{X}_n - \mu}{S_n}\sqrt{n} \overset{H_0}{\sim} t(n-1)$$

Uzorak je veličine $n = 20$ i nivo značajnosti je $\alpha = 0.05$ pa je

$$T = \frac{\overline{X}_{20} - \mu}{S_{20}}\sqrt{20} \overset{H_0}{\sim} t(19)$$

Iz tablice Studentove t-razdiobe odredimo $t_{0.05}(19)$ takav da je $\mathbb{P}(T \leq -t_{0.05}(19)) = 0.05$ i dobijemo $t_{0.05}(19) = 1.729$.

$$\overline{x} = 98, s = 1.55 \implies t = \frac{98-100}{1.55}\sqrt{20} = -5.7 < -1.729 = -t_{0.05}(19)$$

$\implies$ odbacujemo hipotezu H_0 u korist hipoteze H_1 na nivou značajnosti od 5 %

Dakle, proizvođač malo 'vara'.

Pokušajmo riješiti isti zadatak u R-u:

```
> cokolade<-c(97,99,98,96,98,101,98,95,97,99,98,96,97,98,98,100,99,97,101,98)
> t.test(cokolade, mu=100,alternative="less",conf.level=0.95)

One Sample t-test

data:  cokolade
t = -5.7483, df = 19, p-value = 7.69e-06
alternative hypothesis: true mean is less than 100
95 percent confidence interval:
     -Inf 98.60161
sample estimates:
mean of x
       98
```

U retku

```
t = -5.7483, df = 19, p-value = 7.69e-06
```

je izračunata realizacija t-statistike i broj stupnjeva slobode df. Zadnja veličina je **p-vrijednost**, koja je jednaka

$$p = \mathbb{P}(T \geq t | H_0),$$

odakle možemo zaključiti sljedeće:

- ako je $p \leq \alpha$, onda se t nalazi u kritičnom području pa odbacujemo H_0 u korist H_1 na nivou značajnosti α,
- ako je $p > \alpha$, onda se t ne nalazi u kritičnom području pa ne možemo odbaciti H_0 u korist H_1 na nivou značajnosti α.

U ovom slučaju je $p = 0.00000769 < 0.05$ pa odbacujemo H_0 u korist H_1. Primijetimo da iz p-vrijednosti odmah možemo vidjeti na kojim nivoima značajnosti odbacujemo H_0. Npr. u ovom slučaju vidimo da bi H_0 odbacili čim je $\alpha > 0.00000769$.

Dakle, p-vrijednost mjeri koliko je t 'daleko' od ruba kritičnog područja.

U retku

```
alternative hypothesis: true mean is less than 100
```

je opisana alternativna hipoteza da je $\mu < 100$,

dok je s

```
 95 percent confidence interval:
     -Inf 98.60161
```

dan 95 % pouzdani interval za očekivanje μ i njegova procjena na temelju uzorka je

$$(-\infty, 98.60161].$$

Procjena aritmetičke sredine je dana s

```
 mean of x
       98
```

pa je $\overline{x} = 98$.

$\triangle$

Zadatak 5.6 U nekom pjevačkom zboru su mjerene visine (u cm) soprana i altova. Dobiveni su sljedeći podaci

glas	visina								
sopran	162	157	168	165	152	155	160		
alt	165	157	172	170	170	160	170	168	160

Uz pretpostavku da su visine normalno distribuirane s jednakom varijancom, možemo li na nivou značajnosti od 5 % tvrditi da su u prosjeku altovi viši od soprana?

Rješenje:

Uspoređujemo prosječne visine dviju populacija: soprana i altova. Uz pretpostavku o normalnosti i jednakosti varijanci možemo provest t-test za usporedbu očekivanja populacija. Neka su

$$\begin{aligned} X_1 = &\text{ visine soprana}, \quad \mu_1 = \mathbb{E}X_1, \\ X_2 = &\text{ visine altova}, \quad \mu_2 = \mathbb{E}X_2. \end{aligned}$$

Testiramo sljedeće hipoteze:

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 < \mu_2$$

Testna statistika je:

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

gdje je

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

Ovdje je

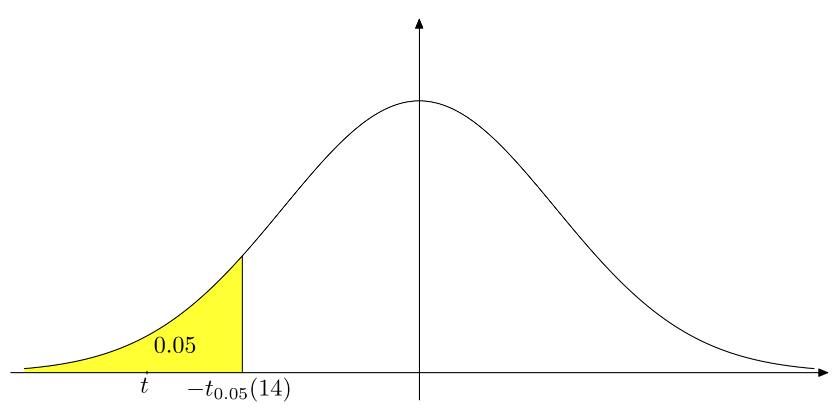
$n_1 = 7$	$n_2 = 9$
$\overline{x}_1 = 159.86$	$\overline{x}_2 = 165.78$
$s_1^2 = 31.81$	$s_2^2 = 30.19$

i

$$s_d^2 = \frac{(7-1) \cdot 31.81 + (9-1) \cdot 30.19}{7 + 9 - 2} = 30.88.$$

Znamo da je

$$T = \frac{\overline{X}_1 - \overline{X}_2}{S_d} \frac{1}{\sqrt{\frac{1}{7} + \frac{1}{9}}} \overset{H_0}{\sim} t(14).$$



Za $\alpha = 0.05$ nađemo iz tablica Studentove t-razdiobe $t_{0.05}(14)$ takav da je

$$\mathbb{P}(T \leq -t_{0.05}(14) | H_0) = 0.05$$

i dobijemo $t_{0.05}(14) = 1.761$. Realizacija statistike T je

$$t = \frac{159.88 - 165.778}{\sqrt{30.88}} \cdot \frac{1}{\sqrt{\frac{1}{7} + \frac{1}{9}}} = -2.099 < -1.761 = -t_{0.05}(14).$$

Odbacujemo H_0 u korist H_1 pa možemo zaključiti da su u prosjeku altovi viši od soprana.

Riješimo zadatak u R-u:

```
> soprani<-c(162,157,168,165,152,155,160)
> altovi<-c(165,157,172,170,170,160,170,168,160)
> t.test(soprani,altovi,var.equal=T,alternative="less")

Two Sample t-test

data:  soprani and altovi
t = -2.1139, df = 14, p-value = 0.02647
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -0.987639
sample estimates:
mean of x mean of y
 159.8571  165.7778
```

Vidimo da je p-vrijednost $p = 0.02647 < 0.05$ pa odbacujemo H_0 u korist alternative H_1 na nivou značajnosti od 5 %.

△

Zadatak 5.7 Mjereni su prinosi kukuruza na poljima sličnog sastava tla s dvije vrste sjemena:

sjeme	prinos	(kg/m^2)					
standardno	0.24	0.22	0.24	0.28	0.24	0.21	0.23
specijalno	0.23	0.21	0.23	0.27	0.25		

Možemo li na temelju gornjih podataka zaključiti da prinosi na poljima zasijanim različitim vrstama sjemena imaju istu varijancu na nivou značajnosti 0.05?

Rješenje:

Koristimo F-test za usporedbu varijanci dviju normalno distribuiranih populacija. Neka su

$$X_1 = \text{ prinos na polju sa standardnim sjemenom}, \quad \sigma_1^2 = \operatorname{Var} X_1,$$
$$X_2 = \text{ prinos na polju sa specijalnim sjemenom}, \quad \sigma_2^2 = \operatorname{Var} X_2.$$

Testiramo sljedeće hipoteze:

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

Testna statistika je:

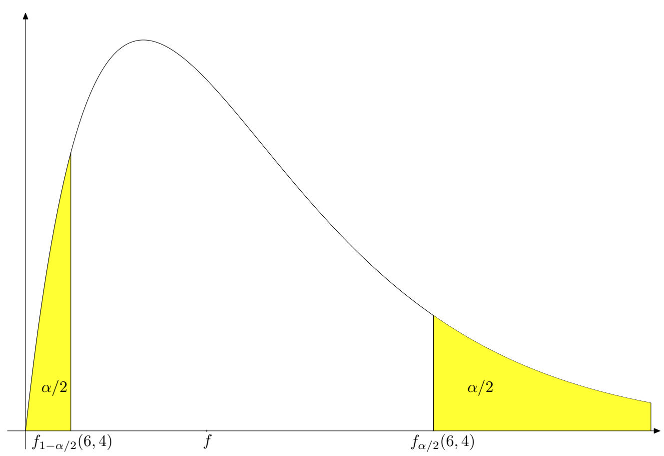
$$F = \frac{S_1^2}{S_2^2} \overset{H_0}{\sim} F(n_1 - 1, n_2 - 1).$$

Ovdje je

$$n_1 = 7 \qquad n_2 = 5$$
$$s_1^2 = 0.0004904 \qquad s_2^2 = 0.00052.$$

Dakle,

$$F = \frac{S_1^2}{S_2^2} \overset{H_0}{\sim} F(6, 4)$$



Za $\alpha = 0.05$ nađemo iz tablica Fisherove F-razdiobe $f_{1-\frac{\alpha}{2}}(6, 4)$ i $f_{\frac{\alpha}{2}}(6, 4)$ takav da je

$$\mathbb{P}(f_{1-\frac{\alpha}{2}}(6, 4) \leq F \leq f_{\frac{\alpha}{2}}(6, 4) | H_0) = 0.05.$$

koristeći i činjenicu

$$f_\alpha(m, n) = \frac{1}{f_{1-\alpha}(n, m)}.$$

Dobijemo

$$f_{0.975}(6, 4) = \frac{1}{f_{0.025}(4, 6)} = \frac{1}{6.23} = 0.1605$$
$$f_{0.025}(6, 4) = 9.20.$$

i realizaciju statistike F

$$f = \frac{s_1^2}{s_2^2} = 0.94322 \in \langle 0.1605, 9.20 \rangle$$

Dakle, ne odbacujemo hipotezu H_0.

Riješimo zadatak u R-u:

```
> polje1<-c(0.24,0.22,0.24,0.28,0.24,0.21,0.23)
> polje2<-c(0.23,0.21,0.23,0.27,0.25)
> var.test(polje1,polje2,conf.level=0.95)

F test to compare two variances

data:  polje1 and polje2
F = 0.9432, num df = 6, denom df = 4, p-value = 0.902
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1025543 5.8736044
sample estimates:
ratio of variances
         0.9432234
```

Vidimo da je p-vrijednost $p = 0.902 > 0.05$ pa ne odbacujemo H_0 na nivou značajnosti od 5 %.

$\triangle$

Zadatak 5.8 Proizvođač tvrdi da njegove pošiljke sadrže najviše 7 % defektnih proizvoda. Uzet je slučajni uzorak od 200 proizvoda iz jedne velike pošiljke i ustanovljeno je da je u njemu 11 % defektnih prozivoda. Ima li proizvođač pravo? ($\alpha = 0.05$)

Rješenje:

Slučajni uzorak $X_1, \ldots, X_n$ je iz Bernoullijevog modela s parametrom p. Ovdje je $n = 200$ pa je uzorak velik i onda je

$$\frac{\overline{X}_n - \mathbb{E}\overline{X}_n}{\sqrt{\operatorname{Var} \overline{X}_n}} \overset{H_0}{\sim} AN(0,1),$$

tj.

$$Z = \frac{\overline{X}_n - p}{\sqrt{p(1-p)}} \overset{H_0}{\sim} AN(0,1).$$

Testiramo sljedeće hipoteze:

$$H_0 : p = 0.07$$
$$H_1 : p > 0.07$$

Za $\alpha = 0.05$ tražimo z_α takav da je

$$\mathbb{P}(Z \geq z_\alpha) = 0.05.$$

Ovdje je $n = 200$, $\overline{x} = 0.11$ i $z_{0.05} = 1.64$.

$$z = \frac{0.11 - 0.07}{\sqrt{0.07 \cdot (1 - 0.07)}}\sqrt{200} = 2.217 > 1.64 = z_{0.05}$$

pa odbacujemo H_0 u korist alternative H_1. Dakle proizvođač nije u pravu.

△

Zadatak 5.9 U nekom gradu su se dvije osobe kandidirale za gradonačelnika. Grad je podijeljen na dva dijela: A i B. U dijelu A je uzet uzorak od 300 glasača i među njima je 168 glasovalo za prvog kandidata, dok je u dijelu B iz uzorka od 200 glasača njih 96 glasovalo za prvog kandidata. Je li prvi kandidat popularniji u dijelu A? ($\alpha = 0.05$)

Rješenje:

Provodimo test proporcija:

$$p_A = \text{omjer glasača koji su glasovali za prvog kandidata u dijelu A},$$
$$p_B = \text{omjer glasača koji su glasovali za prvog kandidata u dijelu B}$$

.

Testiramo sljedeće hipoteze:

$$H_0 : p_A = p_B$$
$$H_1 : p_A > p_B$$

Testna statistika je

$$\frac{\hat{p}_A - \hat{p}_B}{\sqrt{\hat{p}(1-\hat{p})}} \frac{1}{\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}} \overset{H_0}{\sim} AN(0,1),$$

gdje je

$$\hat{p}_A = \text{ procjenitelj za } p_A,$$
$$\hat{p}_B = \text{ procjenitelj za } p_B,$$
$$\hat{p} = \frac{n_A \hat{p}_a + n_b \hat{p}_B}{n_A + n_B}.$$

Za $\alpha = 0.05$ tražimo z_α takav da je

$$\mathbb{P}(Z \geq z_\alpha) = 0.05$$

i dobijemo $z_{0.05} = 1.64$. Ovdje je

$$n_A = 300 \qquad\qquad n_B = 200$$
$$\hat{p}_A = \frac{168}{300} = 0.56 \qquad\qquad \hat{p}_B = \frac{96}{200} = 0.48$$
$$\hat{p} = \frac{300 \cdot 0.56 + 200 \cdot 0.48}{300 + 200} = 0.528.$$

Dakle,

$$t = \frac{0.56 - 0.48}{\sqrt{0.528 \cdot (1 - 0.528)}} \frac{1}{\sqrt{\frac{1}{300} + \frac{1}{200}}} = 1.756 > 1.64 = z_{0.05}$$

pa odbacujemo H_0 u korist alternative H_1. Zaključujemo da je prvi kandidat popularniji u dijelu A nego u dijelu B.

△

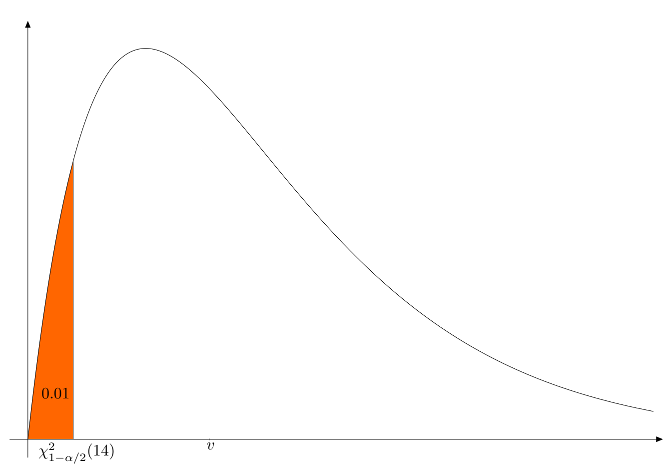
Zadatak 5.10 Standardna devijacija godišnjih temperatura u nekom gradu mjerena u periodu od 100 godina je bila $8°C$. Mjerena je srednja dnevna temperatura 15. dana u mjesecu u zadnjih 15 godina i izračunata je standardna devijacija godišnjih temperatura od $5°C$. Uz pretpostavku o normalnosti temperatura, možemo li na razini značajnosti od 1 % zaključiti da je temperatura u zadnjih 15 godina postala manje varijabilna?

Rješenje: Uzorak dolazi iz normalne razdiobe. Testiramo sljedeće hipoteze:

$$H_0 : \sigma = 8°C$$
$$H_1 : \sigma < 8°C$$

Testna statistika je

$$V = \frac{(n-1)S_n^2}{\sigma^2} \overset{H_0}{\sim} \chi^2(n-1).$$



Za $\alpha = 0.01$ tražimo $\chi^2_{1-\alpha}(n-1)$ takav da je

$$\mathbb{P}(V \leq \chi^2_{1-\alpha}(n-1)) = 0.01$$

i dobijemo $\chi^2_{0.99}(14) = 4.7$. Ovdje je

$$n = 15,\ s^2 = 5^2 \implies v = \frac{14 \cdot 5^2}{8^2} = 5.46 > 4.6 = \chi^2_{0.99}(14)$$

pa ne odbacujemo H_0. Zaključujemo da temperatura u zadnjih 15 godina nije postala manje varijabilna.

△

Zadatak 5.11 Proveden je eksperiment koji se sastoji od uzimanja 103 uzorka vode i određivanja broja pijavica *Helobdella* u svakom uzorku vode. Dobiveni su sljedeći podaci:

# pijavica	0	1	2	3	4	5	6	7	8	≥ 9
frekvencija	58	25	13	2	2	1	1	0	1	0

Iz literature je poznato da je očekivani broj pijabvica u uzorku vode jednak 1. Potvrđuje li to dobiveni uzorak na nivou značajnosti 0.05? (Prepostavite da podaci dolaze iz Poissonove razdiobe.)

Rješenje: Slučajni uzorak je iz Poissonovog modela s parametrom $\lambda > 0$. Testiramo sljedeće hipoteze:

$$H_0 : \lambda = 1$$
$$H_1 : \lambda \neq 1$$

$n = 103$ pa je uzorak velik i iz centralnog graničnog teorema slijedi

$$Z = \frac{\overline{X}_n - \mathbb{E}\overline{X}_n}{\sqrt{\operatorname{Var}\overline{X}_n}} \overset{H_0}{\sim} AN(0,1).$$

Testna statistika je

$$Z = \frac{\overline{X}_n - \lambda}{\sqrt{\lambda}}\sqrt{n} \overset{H_0}{\sim} AN(0,1).$$

Za $\alpha = 0.05$ tražimo $z_{\alpha/2}$ takav da je

$$\mathbb{P}(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 0.05$$

i dobijemo $z_{0.025} = 1.96$. Ovdje je

$$n = 103,\ \overline{x} = \frac{84}{103} \implies z = \frac{\frac{84}{103} - 1}{\sqrt{1}}\sqrt{103} = -1.8721 \in \langle -1.96, 1.96\rangle = \langle -z_{0.025}, z_{0.025}\rangle$$

pa ne odbacujemo H_0. Zaključujemo da podaci potvrđuju teoretske rezultate.

$\triangle$

Zadatak 5.12 Na Cipru su pronađeni bizantski novčići iz razdoblja vladavine kralja Manuela I. Komnenusa (1118.-1180.). Otkriveno je da novčići dolaze iz 4 serije. Mjeren je postotak srebra (Ag) prisutnog u svakom novčiću i dobiveni su sljedeći podaci:

serija	1.	2.	3.	4.
% Ag	5.9	6.9	4.9	5.9
	6.8	9.0	5.5	5.6
	6.4	6.6	4.6	5.5
	7.0	8.1	4.5	5.1
	6.6	9.3		6.2
	7.7	9.2		5.8
	7.2	8.6		5.8
	6.9			
	6.2			

Postoji li razlika u postotku srebra među novčićima iz različitih serija? ($\alpha = 0.05$)

Rješenje: Uspoređujemo 4 populacije iz normalne razdiobe uz pretpostavku o jednakosti varijanci. Koristimo ANOVA-u. Neka je

$$X_i = \text{ postotak srebra u novčiću iz } i\text{-te serije }, \ \mu_i = \mathbb{E}X_i, \quad i = 1, 2, 3, 4.$$

Testiramo sljedeće hipoteze:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$
$$H_1 : \text{ postoje } i, j \in \{1, 2, 3, 4\} \text{ takvi da je } \mu_i \neq \mu_j$$

i	n_i	$\overline{x}_{i\cdot}$	$n_i\overline{x}_{i\cdot}$	$n_i(\overline{x}_{i\cdot} - \overline{x})^2$	$s^2_{i\cdot}$	$(n_i - 1)s^2_{i\cdot}$
1	9	6.744	60.7	0.296	0.295	2.362
2	7	8.243	57.7	19.573	1.21	7.257
3	4	4.875	19.5	11.397	0.2025	0.607
4	7	5.614	39.3	6.3	0.131	0.789
$\sum$	27		177.2	$\underbrace{37.748}_{\text{SST}}$		$\underbrace{11.015}_{\text{SSE}}$

Ovdje je $k = 4$, $n = 27$,

$$\overline{x} = \frac{177.2}{27} = 6.5629, \ \text{MST} = \frac{\text{SST}}{k-1} = \frac{37.748}{4-1} = 12.583,$$

$$\text{MSE} = \frac{\text{SSE}}{n-k} = \frac{11.015}{27-4} = 0.479, \ f = \frac{\text{MST}}{\text{MSE}} = 26.272.$$

ANOVA tablica:

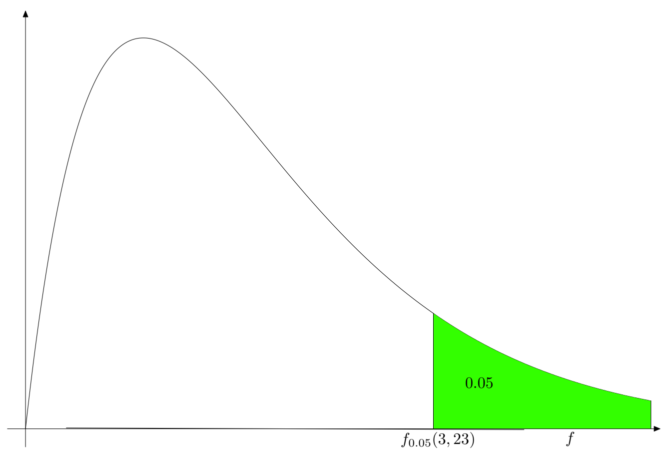
izvor varijabilnosti	broj st. slobode	suma kvadrata odstupanja	srednje kvadratno odstupanje	F-statistika
zbog tretmana	3	37.748	12.583	26.272
sl. greška	23	11.015	0.479	
$\sum$	26	48.763		

Testna statistika je

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

Iz tablica pročitamo

$$f_{0.05}(3,23) = 3.03.$$



Zbog

$$f = 26.272 > 3.03 = f_{0.05}(3,23)$$

odbacujemo H_0 u korist H_1. Dakle, na nivou značajnosti 5 % postoji razlika među novčićima iz različitih serija.

U R-u:

```
> serija1<-c(5.9,6.8,6.4,7.0,6.6,7.7,7.2,6.9,6.2)
> serija2<-c(6.9,9.0,6.6,8.1,9.3,9.2,8.6)
> serija3<-c(4.9,5.5,4.6,4.5)
> serija4<-c(5.3,5.6,5.5,5.1,6.2,5.8,5.8)
> novcic<-c(serija1,serija2,serija3,serija4)
> novcic
 [1] 5.9 6.8 6.4 7.0 6.6 7.7 7.2 6.9 6.2 6.9 9.0 6.6 8.1 9.3 9.2 8.6 4.9 5.5 4.6
[20] 4.5 5.3 5.6 5.5 5.1 6.2 5.8 5.8
> grupe<-c(rep(1,length(serija1)),rep(2,length(serija2)),rep(3,length(serija3)),
 +rep(4,length(serija4)))
> grupe
 [1] 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 3 3 3 3 4 4 4 4 4 4 4
```

```
> grupe=factor(grupe)
> grupe
 [1] 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 3 3 3 3 4 4 4 4 4 4 4
Levels: 1 2 3 4
> anova(lm(novcic~grupe))
Analysis of Variance Table

Response: novcic
          Df Sum Sq Mean Sq F value    Pr(>F)
grupe      3 37.748  12.583  26.272 1.306e-07 ***
Residuals 23 11.015   0.479
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Vidimo da je p-vrijednost jako mala.

$\triangle$

5.1 Zadaci za vježbu

5.13 Iz Mendelove teorije o nasljeđivanju slijedi da se određena vrsta graška pojavljuje u žutoj i zelenoj boji u omjeru 3:1. Provedeno je križanje i na kraju se dobilo 179 žutih i 49 zelenih grašaka. Potvrđuju li podaci teoretski model na razini značajnosti od 0.05?

5.14 Promatra se potrošnja goriva dviju vrsta automobila. Provedeno je mjerenje i za 50 automobila tipa A prosječna potrošnja je bila 17.0 milja po galonu uz standardnu devijaciju od 2.5 milja po galonu, dok je za 70 automobila tipa B prosječna potrošnja bila 18.6 milja po galonu uz standardnu devijaciju od 3.0 milja po galonu. Možemo li na razini značajnosti od 0.05 zaključiti da automobili tipa A troše u prosjeku manje goriva?

5.15 Kocka je bacana 144 puta i palo je ukupno 32 šestice. Je li kocka simetrična? ($\alpha = 0.05$)

5.16 Neka kompanija koja proizvodi perece koristi stroj za pakiranje pereca u vrećice na kojima piše da je masa 454 g. Uzet je slučajni uzorak vrećica i dobiveni su rezultati

464	450	450	456	452	433
446	446	450	447	442	438

Je li proizvođač u pravu? ($\alpha = 0.05$)

5.17 Istražuje se pospješuje li nova tehnika sposobnost čitanja s razumijevanjem kod učenika u nekom razredu. Dio učenika nije korisitio novu tehniku, dok dio jest. Kasnije su učenici pisali test. U tablici su prikazani bodovi:

Grupa 1 (bez nove tehnike)	42	43	55	26	62	37	33	41	19
Grupa 2 (s novom tehnikom)	24	43	58	71	43	49			

Je li nova tehnika bolja? ($\alpha = 0.05$)

5.18 Di bi se ispitao utjecaj vrste hrane na prirast težine krave, uzete su tri hranjive smjese A, B i C. Smjesom A su hranjene 4 krave, smjesom B 3 krave i smjesom C 5 krava. Nakon mjesec dana, izmjerena je težina krava i izračunat je prirast (u %):

Smjesa A	13.7	14.2	12.8	13.7	
Smjesa B	14.0	13.9	11.7		
Smjesa C	13.7	14.2	13.3	14.0	14.1

Utječe li vrsta hrane na prirast težine? ($\alpha = 0.05$)

Rješenja: **5.13.** Da. **5.14.** Da. **5.15.** Da. **5.16.** Proizođač nije u pravu, jer je $t = -2.64$ **5.17.** Ne, jer je $t = -1.08$ **5.18.** Ne, jer je $F = 0.7456$

6

Linearni regresijski modeli

Zadatak 6.1 Specijalizirani magazin je objavio listu cijena rabljenih automobila. x predstavlja starost u godinama, a y pripadnu cijenu u tisućama \$. Podaci su dani tablicom:

x	1	2	2	3	3	4	6	7	8	10
y	2.45	1.80	2.00	2.00	1.70	1.20	1.15	0.69	0.60	0.47

(a) Procijenite parametre linearne regresije i nacrtajte regresijski pravac zajedno s podacima.

(b) Konstruirajte 95% pouzdane intervale za parametre regresije.

(c) Uz pretpostavku o normalnosti pogreške, odredite 90% pouzdani interval za σ^2.

(d) Pomoću prilagođenog regresijskog pravca procijenite srednju cijenu pet godina starog automobila i konstruirajte 95% pouzdani interval.

(e) Pretpostavite da imate auto star 5 godina. Odredite 95% pouzdani interval za njegovu cijenu.

(f) Provedite test značajnosti linearnog regresijskog modela ($\alpha = 0.05$).

Rješenje:

(a) Iz $\bar{x} = 4.6, \bar{y} = 1.406$ i

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = 80.4$$

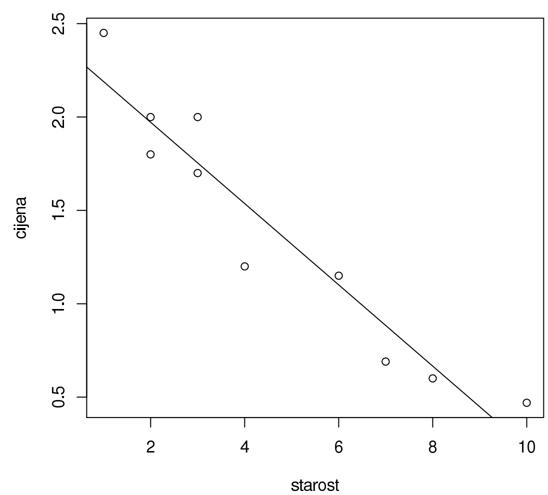
$$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = 4.18364$$

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\,\bar{y} = -17.496$$

dobivamo prema zadatku 1.16 parametre linearne regresije

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = -0.218, \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 2.407.$$

Regresijski pravac je $y = \hat{\alpha} + \hat{\beta}x = 2.407 - 0.218x$.



(b) Vrijedi

$$\frac{\hat{\alpha} - \alpha}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}} \sim t(n-2), \quad \frac{\hat{\beta} - \beta}{\hat{\sigma}\sqrt{\frac{1}{S_{xx}}}} \sim t(n-2),$$

gdje su

$$\mathrm{SSE} = S_{yy} - \hat{\beta}^2 S_{xx} = 0.3763, \quad \hat{\sigma} = \sqrt{\frac{\mathrm{SSE}}{n-2}} = 0.217.$$

Za danu pouzdanost $\alpha = 0.05$ i $n = 10$ je $t_{\frac{\alpha}{2}}(n-2) = t_{0.025}(8) = 2.306$. Dakle,

$$\mathbb{P}\left(-2.306 \leq \frac{\hat{\alpha} - \alpha}{\hat{\sigma}\sqrt{\frac{1}{10} + \frac{\bar{x}^2}{S_{xx}}}} \leq 2.306\right) = 0.95,$$

odnosno

$$\mathbb{P}\left(\hat{\alpha} - 2.306 \cdot \hat{\sigma}\sqrt{\frac{1}{10} + \frac{\bar{x}^2}{S_{xx}}} \leq \alpha \leq \hat{\alpha} + 2.306 \cdot \hat{\sigma}\sqrt{\frac{1}{10} + \frac{\bar{x}^2}{S_{xx}}}\right) = 0.95$$

pa je procjena 95% pouzdanog intervala za α

$$\left[2.407 - 2.306 \cdot 0.217\sqrt{\frac{1}{10} + \frac{4.6^2}{80.4}}, 2.407 + 2.306 \cdot 0.217\sqrt{\frac{1}{10} + \frac{4.6^2}{80.4}}\right] = [2.105, 2.708].$$

Slično,

$$\mathbb{P}\left(-2.306 \leq \frac{\hat{\beta}-\beta}{\hat{\sigma}\sqrt{\frac{1}{S_{xx}}}} \leq 2.306\right) = 0.95,$$

odnosno

$$\mathbb{P}\left(\hat{\beta} - 2.306 \cdot \hat{\sigma}\sqrt{\frac{1}{S_{xx}}} \leq \beta \leq \hat{\beta} + 2.306 \cdot \hat{\sigma}\sqrt{\frac{1}{S_{xx}}}\right) = 0.95$$

pa je procjena 95% pouzdanog intervala za β

$$\left[-0.218 - 2.306 \cdot 0.217\sqrt{\frac{1}{80.4}}, -0.218 + 2.306 \cdot 0.217\sqrt{\frac{1}{80.4}}\right] = [-0.236, -0.2].$$

(c) Vrijedi

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

Za danu pouzdanost $\alpha = 0.10$ i $n = 10$ su

$$\chi^2_{\frac{\alpha}{2}}(n-2) = \chi^2_{0.05}(8) = 15.5073,$$
$$\chi^2_{1-\frac{\alpha}{2}}(n-2) = \chi^2_{0.95}(8) = 2.7326.$$

Dakle,

$$\mathbb{P}\left(2.7326 \leq 8 \cdot \frac{\hat{\sigma}^2}{\sigma^2} \leq 15.5073\right) = 0.90,$$

odnosno

$$\mathbb{P}\left(\frac{8 \cdot \hat{\sigma}^2}{15.5073} \leq \sigma^2 \leq \frac{8 \cdot \hat{\sigma}^2}{2.7326}\right) = 0.90$$

pa je procjena 90% pouzdanog intervala za σ^2

$$\left[\frac{8 \cdot 0.217^2}{15.5073}, \frac{8 \cdot 0.217^2}{2.7326}\right] = [0.000143, 0.13786].$$

(d) Ovdje procjenjujemo $\mathbb{E}[Y|x=5] = \alpha + \beta \cdot 5$ sa $\hat{Y} = \hat{\alpha} + \hat{\beta} \cdot 5$.
Vrijedi

$$\sqrt{\frac{n-2}{\mathrm{SSE}}} \cdot \frac{\hat{\alpha} + \hat{\beta}x - (\alpha + \beta x)}{\sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{S_{xx}}}} \sim t(n-2).$$

Za $\alpha = 0.05$ i $n = 10$ je $t_{\frac{\alpha}{2}}(n-2) = t_{0.025}(8) = 2.306$. Dakle,

$$\mathbb{P}\left(-2.306 \leq \sqrt{\frac{8}{\mathrm{SSE}}} \cdot \frac{\hat{\alpha} + \hat{\beta} \cdot 5 - (\alpha + \beta \cdot 5)}{\sqrt{\frac{1}{10} + \frac{(5-\overline{x})^2}{S_{xx}}}} \leq 2.306\right) = 0.95,$$

odnosno

$$\mathbb{P}\left(\hat{\alpha}+\hat{\beta}\cdot 5-2.306\cdot\sqrt{\frac{\mathrm{SSE}}{8}}\sqrt{\frac{1}{10}+\frac{(5-\overline{x})^2}{S_{xx}}}\leq \alpha+\beta\cdot 5\leq \hat{\alpha}+\hat{\beta}\cdot 5+2.306\cdot\sqrt{\frac{\mathrm{SSE}}{8}}\sqrt{\frac{1}{10}+\frac{(5-\overline{x})^2}{S_{xx}}}\right)=0.95.$$

Procjena srednje cijene automobila starog 5 godina je $\hat{y}(5)=2.407-0.218\cdot 5=1.317$, a procjena 95% pouzdanog intervala za srednju cijenu automobila starog 5 godina je

$$\left[1.317-2.306\cdot\sqrt{\frac{0.3763}{8}}\sqrt{0.1+\frac{(5-4.6)^2}{80.4}}, 1.317+2.306\cdot\sqrt{\frac{0.3763}{8}}\sqrt{0.1+\frac{(5-4.6)^2}{80.4}}\right],$$

odnosno [1.157,1.477].

(e) Ovdje procjenjujemo $Y=\alpha+\beta\cdot 5+e$ sa $\hat{Y}=\hat{\alpha}+\hat{\beta}\cdot 5$, gdje je $e\sim N(0,\sigma^2)$. Vrijedi

$$\sqrt{\frac{n-2}{\mathrm{SSE}}}\cdot\frac{\hat{\alpha}+\hat{\beta}x-Y}{\sqrt{1+\frac{1}{n}+\frac{(x-\overline{x})^2}{S_{xx}}}}\sim t(n-2).$$

Za $\alpha=0.05$ i $n=10$ je $t_{\frac{\alpha}{2}}(n-2)=t_{0.025}(8)=2.306$. Dakle,

$$\mathbb{P}\left(-2.306\leq\sqrt{\frac{8}{\mathrm{SSE}}}\cdot\frac{\hat{\alpha}+\hat{\beta}\cdot 5-Y}{\sqrt{1+\frac{1}{10}+\frac{(5-\overline{x})^2}{S_{xx}}}}\leq 2.306\right)=0.95,$$

odnosno

$$\mathbb{P}\left(\hat{\alpha}+\hat{\beta}\cdot 5-2.306\cdot\sqrt{\frac{\mathrm{SSE}}{8}}\sqrt{1+\frac{1}{10}+\frac{(5-\overline{x})^2}{S_{xx}}}\leq Y\leq \hat{\alpha}+\hat{\beta}\cdot 5+2.306\cdot\sqrt{\frac{\mathrm{SSE}}{8}}\sqrt{1+\frac{1}{10}+\frac{(5-\overline{x})^2}{S_{xx}}}\right)=0.95$$

Procjena 95% pouzdanog intervala za cijenu automobila starog 5 godina je

$$\left[1.317-2.306\cdot\sqrt{\frac{0.3763}{8}}\sqrt{1+0.1+\frac{(5-4.6)^2}{80.4}}, 1.317+2.306\cdot\sqrt{\frac{0.3763}{8}}\sqrt{1+0.1+\frac{(5-4.6)^2}{80.4}}\right],$$

odnosno [0.792,1.842].

(f) Testiramo hipoteze:

$$H_0 \colon \beta = 0$$
$$H_1 \colon \beta \neq 0$$

Prema (b) je procjena 95% pouzdanog intervala za β jednaka $[-0.236, -0.2]$. Budući da $0 \notin [-0.236, -0.2]$, odbacujemo H_0 u korist H_1 na nivou značajnosti 0.05.

△

Zadatak 6.2 Proučava se količina etilena koju sadrži sjeme salate kao funkcija vremena izlaganja tvari koja apsorbira etilen. Podaci sa dani tablicom:

x (min)	2	10	20	30	40	50	60	70	80	90	100
y (nl/g)	408	274	196	137	90	78	51	40	30	22	15

Neka je Y količina etilena u sjemenu salate i $Z = \ln Y$.

(a) Prikažite točke (x, y) u Kartezijevom koordinatnom sustavu.

(b) Prikažite točke (x, z) u Kartezijevom koordinatnom sustavu, gdje su sa $z = \ln y$ transformirani originalni podaci.

(c) Procijenite parametre linearne regresije i nacrtajte regresijski pravac za transformirane podatke.

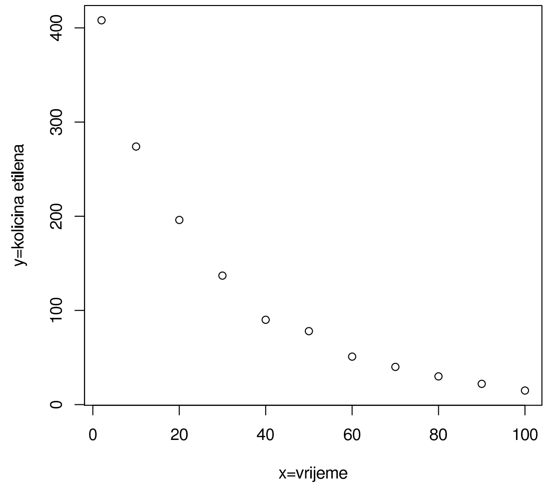
(d) Koristeći linearan regresijski model transformiranih podataka, napišite kako glasi model za originalne podatke. Nacrtajte pripadnu regresijsku funkciju zajedno s originalnim podacima.

(e) Odredite gornju i donju krivulju koje definiraju 95% pouzdani interval za srednju vrijednost od Z uz dano $x = x_0$ te ih prikažite zajedno s transformiranim podacima (x, z) i regresijskim pravcem.

(f) Odredite gornju i donju krivulju koje definiraju 95% pouzdani interval za srednju vrijednost od Y uz dano $x = x_0$ te ih prikažite zajedno s originalnim podacima (x, y) i regresijskom funkcijom.

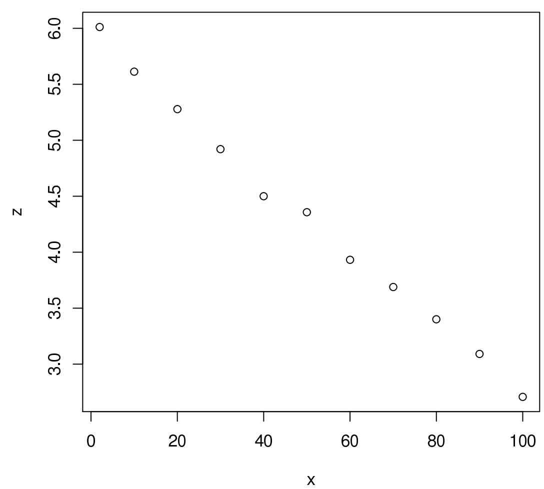
Rješenje:

(a) Originalni podaci su prikazani na sljedećoj slici:



(b) Transformirani podaci su:

x	2	10	20	30	40	50	60	70	80	90	100
z	6.01	5.61	5.28	4.92	4.5	4.36	3.93	3.69	3.4	3.09	2.71



(c) Iz $\bar{x} = 50.18, \bar{z} = 4.32$ i

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = 10803.64$$

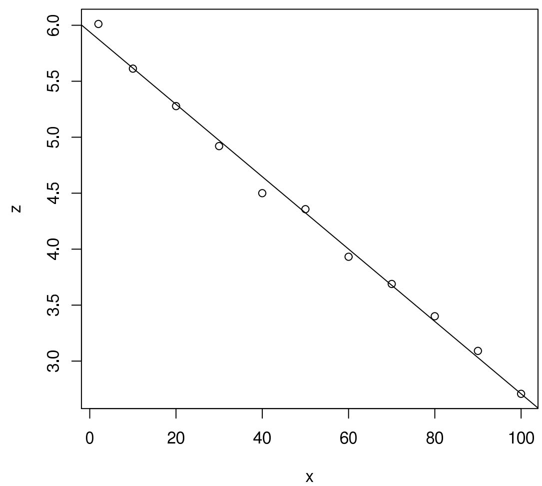
$$S_{zz} = \sum_{i=1}^{n} z_i^2 - n\bar{z}^2 = 11.35$$

$$S_{xz} = \sum_{i=1}^{n} x_i z_i - n\bar{x}\,\bar{z} = -349.27$$

dobivamo prema zadatku 1.16 parametre linearne regresije

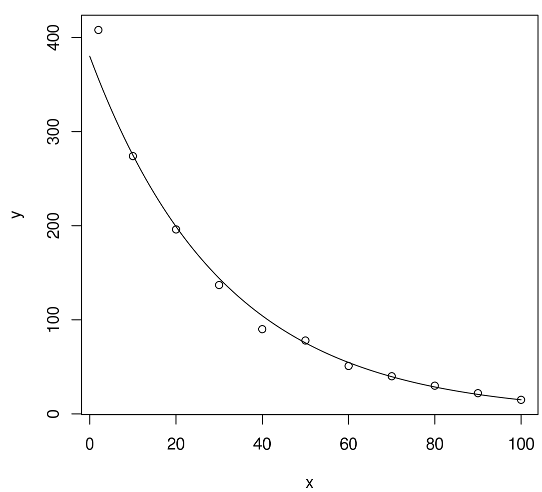
$$\hat{\beta} = \frac{S_{xz}}{S_{xx}} = -0.0323, \quad \hat{\alpha} = \bar{z} - \hat{\beta}\bar{x} = 5.9405.$$

Regresijski pravac je $z = \hat{\alpha} + \hat{\beta}x = 5.9405 - 0.0323x$:



(d) Iz $\ln y = z$ dobivamo regresijsku krivulju za originalne podatke:

$$y = e^z = e^{5.9405 - 0.0323x}$$

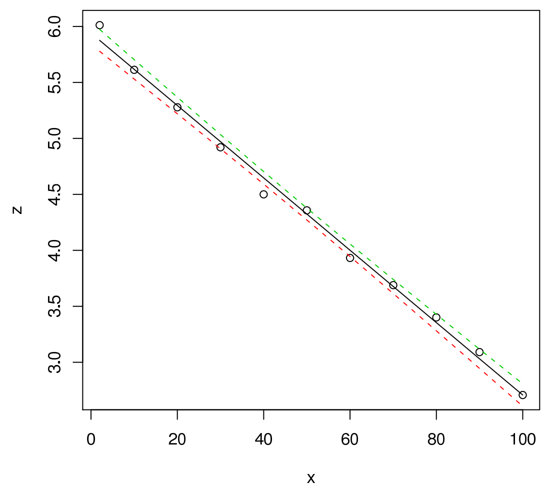


(e) 95% pouzdani interval za srednju vrijednost od Z uz dano $x = x_0$, odnosno za $\mathbb{E}[Z|x = x_0]$ je dan sa:

$$\hat{\alpha} + \hat{\beta} \cdot x_0 \pm t_{\frac{\alpha}{2}}(n-2) \cdot \sqrt{\frac{\mathrm{SSE}}{n-2}} \sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}}.$$

Za $\alpha = 0.05$ i $n = 11$ je $t_{0.025}(9) = 2.2622$, a SSE $= S_{zz} - \hat{\beta}^2 S_{xx} = 0.0547$. Uvrštavanjem svih podataka dobivamo procjenu traženog pouzdanog intervala:

$$5.9405 - 0.0323 \cdot x_0 \pm \underbrace{2.2622 \cdot \sqrt{\frac{0.0547}{9}}}_{=0.1716} \sqrt{\frac{1}{11} + \frac{(x_0 - 50.18)^2}{10803.64}},$$



Budući da je $Z = \ln Y$, tada iz $Y = e^Z$ slijedi da je donja krivulja koja definira procjenu 95% pouzdanog intervala za $\mathbb{E}[Y|x = x_0]$ jednaka

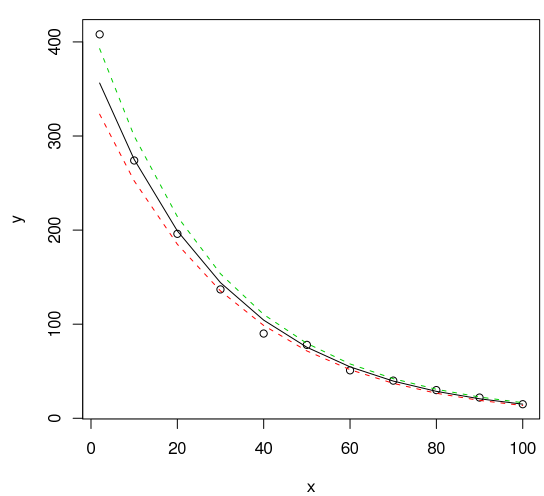
$$d(x_0) = \exp\left\{5.9405 - 0.0323 \cdot x_0 - 0.1716 \cdot \sqrt{\frac{1}{11} + \frac{(x_0 - 50.18)^2}{10803.64}}\right\},$$

dok je gornja krivulja jednaka

$$g(x_0) = \exp\left\{5.9405 - 0.0323 \cdot x_0 + 0.1716 \cdot \sqrt{\frac{1}{11} + \frac{(x_0 - 50.18)^2}{10803.64}}\right\}.$$

Dobivene krivulje su prikazane na sljedećoj slici:

$\triangle$



Nacrtajmo slike iz prethodnog zadatka u R-u:

```
> x=c(2,10,20,30,40,50,60,70,80,90,100)
> y=c(408,274,196,137,90,78,51,40,30,22,15)
> z=log(y)
```

(c) Odredimo pravac regresije za transformirane podatke i nacrtajmo ga:

```
> regr=lm(z~x)
> plot(x,z)
> abline(regr$coefficients)
```

(d) Definirajmo i nacrtajmo funkciju regresije f za originalne podatke:

```
> f=function(t)
+ exp(regr$coefficients[1]+regr$coefficients[2]*t)
> plot(x,y)
> curve(f,add=T)
```

(e) Procijenjene vrijednosti se mogu dobiti naredbom `predict(lm(podaci1~podaci2))`:

```
> predict(regr)
       1        2        3        4        5        6        7        8
5.875838 5.617208 5.293921 4.970634 4.647347 4.324060 4.000773 3.677486
       9       10       11
3.354199 3.030912 2.707625
```

Ako dodamo `interval="confidence"`, tada dobivamo i stupce s granicama procjene 95% pouzdanog intervala za $\mathbb{E}[Z|x = x_0]$, za sve očitane x_0:

```
> predict(regr,interval="confidence")
        fit      lwr      upr
1  5.875838 5.778302 5.973374
2  5.617208 5.530735 5.703681
...
11 2.707625 2.607750 2.807501
```

Krivulje skiciramo naredbom `matlines` koja crta stupce matrice:

```
> pc=predict(regr,interval="confidence")
> plot(x,z)
> matlines(x,pc,lty=c(1,2,2))
```

Pomoću `interval="prediction"` dobivamo stupce s granicama procjene 95% pouzdanog intervala za Z uz dano $x = x_0$, za sve očitane x_0. Razinu značajnosti zadajemo argumentom `conf.level`.

(f)
```
> plot(x,y)
> matlines(x,exp(pc),lty=c(1,2,2))
```

Zadaci za vježbu

6.3 Dani su sljedeći podaci:

x	1	2	3	4	5
y	0.9	2.1	2.5	3.3	3.8

Pretpostavljamo da su pogreške normalne, nezavisne i jednako distribuirane.

(a) Prikažite podatke grafički.

(b) Procijenite i nacrtajte regresijski pravac zajedno s podacima.

(c) Izračunajte reziduale $e_i = y_i - \hat{y}_i$ i uvjerite se da im je suma jednaka 0.

(d) Izračunajte sumu kvadrata reziduala SSE.

(e) Procijenite varijancu σ^2 i nađite 95% pouzdani interval.

(f) Konstruirajte 95% pouzdane intervale za parametre regresije.

(g) Provedite test značajnosti linearnog regresijskog modela ($\alpha = 0.05$).

6.4 Mjerenje dviju varijabli (x, Y) dalo je sljedeće rezultate:

$$\begin{array}{lll} n = 15 & \overline{x} = 10.8 & \overline{y} = 122.7 \\ S_{xx} = 70.6 & S_{yy} = 98.5 & S_{xy} = 68.3 \end{array}$$

(a) Procijenite parametre linearne regresije te pravac regresije.

(b) Izračunajte SSE i procijenite σ^2.

(c) Da li su podaci u kontradikciji s eksperimentalnom pretpostavkom da je srednja procjena varijable Y u jedinici od x barem 1.5?

(d) Konstruirajte 95% pouzdani interval za očekivanu vrijednost od Y u $x = 12$.

6.5 Dani su sljedeći podaci:

x	1	2	3	4	5	6
y	6	18	55	160	485	1460

Prikažite podatke grafički i odredite eksponencijalnu krivulju koja najbolje aproksimira podatke metodom najmanjih kvadrata.

Rješenja: **6.3.** (b) $\hat{\alpha} = 0.42, \hat{\beta} = 0.7$ (d) SSE=0.148, (e) $\hat{\sigma}^2 = 0.047, \sigma^2 \in [0.0158, 0.6858]$ (f) $\alpha \in [-0.32, 1.16], \beta \in [0.48, 0.92]$ (g) Odbacujemo H0: $\beta = 0$ u korist H1: $\beta \neq 0$. **6.4.** (a) $\hat{\alpha} = 112.252$, $\hat{\beta} = 0.967$, (b) SSE=32.4251, $\hat{\sigma}^2 = 2.49$, (c) Odbacujemo H0: $\beta = 1.5$ u korist H1: $\beta < 1.5$. (d) $[122.849, 124.863]$ **6.5.** $y = 2 \cdot 3^x$

7

χ^2-test i Kolmogorov-Smirnovljev test

7.1 χ^2-test o pripadnosti distribuciji

Zadatak 7.1 Tri novčića se bacaju 250 puta i broji se broj pisama koji su pali. Dobiveni su sljedeći podaci:

Broj pisama	0	1	2	3
Frekvencija	24	108	95	23

Provjerite je li novčić simetričan na nivou značajnosti od $\alpha = 0.05$?

Rješenje: Broj podataka je: $n = 24 + 108 + 95 + 23 = 250$.

1. Stavimo X=broj pisama. Testiramo:

$$H_0 : X \sim \begin{pmatrix} 0 & 1 & 2 & 3 \\ \frac{1}{8} & \frac{3}{8} & \frac{3}{8} & \frac{1}{8} \end{pmatrix}$$
$$H_1 : \text{ ne } H_0$$

2. Neka je $X_1, \ldots, X_{250}$ slučajni uzorak za X. Ovdje je particija: $A_i = \{i\}, i = 0, 1, 2, 3$. Tada su opažene frekvencije N_i i teorijske frekvencije $n_i = n \cdot \mathbb{P}(X = i|H_0)$ dane u tablici:

i	N_i	n_i	$(N_i - n_i)^2/n_i$
0	24	$250 \cdot \frac{1}{8}$=31.25	1.682
1	108	$250 \cdot \frac{3}{8}$=93.75	2.166
2	95	$250 \cdot \frac{3}{8}$=93.75	0.0167
3	23	$250 \cdot \frac{1}{8}$=31.25	2.178
Σ	240		6.0427

Opažena vrijednost testne statistike je

$$h = \sum_{i=0}^{3} \frac{(N_i - n_i)^2}{n_i} = 6.0427.$$

3. Vrijedi $k = 4, d = 0, df = k - d - 1 = 3, \alpha = 0.05$ pa je $\chi^2_\alpha(df) = \chi^2_{0.05}(3) = 7.8147$. Dakle, kritično područje je

$$[\chi^2_\alpha(df), +\infty\rangle = [7.8147, +\infty\rangle.$$

Budući da je $h = 6.0427 < 7.8147$ zaključujemo da na temelju podataka ne možemo odbaciti H_0 na nivou značajnosti $\alpha = 0.05$.

U R-u:

```
> kocka<-c(24,108,95,23)
> pp<-c(1/8,3/8,3/8,1/8)
> chisq.test(kocka,p=pp)

Chi-squared test for given probabilities

data:  kocka
X-squared = 6.0427, df = 3, p-value = 0.1096
```

△

Zadatak 7.2 U biblioteci je slučajno odabrano 200 uzoraka po 5 knjiga. Dobivena je sljedeća empirijska distribucija broja oštećenih knjiga:

i	0	1	2	3	4	5
N_i	72	77	34	14	2	1

Uz značajnost od $\alpha = 0.05$ testirajte hipotezu da diskretna slučajna varijabla X, koja označava broj oštećenih knjiga u uzorku od 5 knjiga, ima binomnu razdiobu.

Rješenje: $n = \sum_{i=0}^{5} N_i = 200$

Stavimo X = broj oštećenih knjiga. Testiramo:

$$\begin{aligned} H_0 &: X \sim B(5, p) \\ H_1 &: \text{ ne } H_0 \end{aligned}$$

Procijenimo parametar p metodom maksimalne vjerodostojnosti. Definiramo $L\colon \langle 0, 1\rangle \to \mathbb{R}$ sa

$$L(p) = \prod_{i=0}^{5} \left[\binom{5}{i} p^i (1-p)^{5-i} \right]^{N_i} = c \cdot p^{\sum_{i=0}^{5} i \cdot N_i} (1-p)^{\sum_{i=0}^{5} (5-i) \cdot N_i},$$

gdje je $c > 0$ konstanta koja ne ovisi o p pa je dovoljno maksimizirati funkciju

$$l(p) = \left[\sum_{i=0}^{5} i \cdot N_i\right] \ln p + \left[\sum_{i=0}^{5} (5-i) \cdot N_i\right] \ln(1-p),$$

odakle deriviranjem dobijemo da se maksimum dostiže u $\hat{p} = \dfrac{\sum_{i=0}^{5} i \cdot N_i}{5 \cdot \sum_{i=0}^{5} N_i}$. Dakle,

$$\hat{p} = \frac{0 \cdot 72 + 1 \cdot 77 + 2 \cdot 34 + 3 \cdot 14 + 2 \cdot 4 + 1 \cdot 5}{5 \cdot 200} = \frac{200}{5 \cdot 200} = \frac{1}{5} = 0.2.$$

Neka je $X_1, \ldots, X_{200}$ slučajni uzorak za X. Ovdje je particija: $A_i = \{i\}, i = 0, \ldots, 5$. Tada su opažene frekvencije N_i i teorijske frekvencije $n_i = n \cdot \mathbb{P}(X = i | H_0)$ dane tablicom:

i	N_i		n_i		$(N_i - n_i)^2/n_i$
0	72		65.536		0.638
1	77		81.92		0.295
2	34		40.96		1.204
3	14		10.24		
4	2	17	1.28	11.584	2.532
5	1		0.064		
Σ	200				4.669

Teorijske frekvencije računamo rekurzivno koristeći sljedeću relaciju:

$$\frac{n_{i+1}}{n_i} = \frac{n \cdot \binom{5}{i+1} p^{i+1} (1-p)^{5-(i+1)}}{n \cdot \binom{5}{i} p^i (1-p)^{5-i}} = \frac{5-i}{i} \cdot \frac{p}{1-p}.$$

Opažena vrijednost testne statistike je

$$h = \sum_{i=0}^{5} \frac{(N_i - n_i)^2}{n_i} = 4.669.$$

Vrijedi $k = 4, d = 1, df = k - d - 1 = 2, \alpha = 0.05$ pa je $\chi^2_\alpha(df) = \chi^2_{0.05}(2) = 5.9915$. Dakle, kritično područje je

$$[\chi^2_\alpha(df), +\infty\rangle = [5.9915, +\infty\rangle.$$

Budući da je $h = 4.669 < 5.9915$ zaključujemo da podaci ne podržavaju odbacivanje H_0 na nivou značajnosti $\alpha = 0.05$.

$\triangle$

Zadatak 7.3 Uz nivo značajnosti od $\alpha = 0.05$ testirajte hipotezu da podaci iz tablice

i	1	2	3	4	5	≥ 6
N_i	71	28	5	2	2	1

dolaze iz geometrijske distribucije.

Rješenje: $n = \sum_{i=1}^{6} N_i = 109$

Testiramo:

$$H_0 : X \sim G(p), \text{ tj. } \mathbb{P}(X = i) = (1-p)^{i-1}p, i = 1, 2, \ldots$$
$$H_1 : \text{ ne } H_0$$

Procijenimo parametar p metodom maksimalne vjerodostojnosti. Definiramo $L\colon \langle 0, 1\rangle \to \mathbb{R}$ sa

$$\begin{aligned} L(p) &= \prod_{i=1}^{5} \mathbb{P}(X = i)^{N_i} \cdot \mathbb{P}(X \geq 6)^{N_6} \\ &= \prod_{i=1}^{5} [(1-p)^{i-1}p]^{N_i} \cdot [(1-p)^5]^{N_6} \\ &= (1-p)^{\sum_{i=1}^{6}(i-1)\cdot N_i} \cdot p^{\sum_{i=1}^{5} N_i}. \end{aligned}$$

Maksimiziramo funkciju

$$l(p) = \left[\sum_{i=1}^{6} (i-1) \cdot N_i\right] \ln(1-p) + \sum_{i=1}^{5} N_i \cdot \ln p,$$

odakle deriviranjem dobijemo da se maksimum dostiže u $\hat{p} = \dfrac{\sum_{i=1}^{5} N_i}{\sum_{i=1}^{5} i \cdot N_i + 5N_6}$. Dakle,

$$\hat{p} = \frac{108}{1 \cdot 71 + 2 \cdot 28 + 3 \cdot 5 + 4 \cdot 2 + 5 \cdot 2 + 5 \cdot 1} = \frac{109}{165} = 0.6545.$$

Neka je $X_1, \ldots, X_{109}$ slučajni uzorak za X. Ovdje je particija: $A_i = \{i\}, i = 1, \ldots, 5$ i $A_6 = \{6, 7, \ldots\}$ Tada su opažene frekvencije N_i i teorijske frekvencije $n_i = n \cdot \mathbb{P}(X \in A_i | H_0)$ dane tablicom:

<table>
<tr><th>i</th><th colspan="2">N_i</th><th colspan="2">n_i</th><th>$(N_i - n_i)^2/n_i$</th></tr>
<tr><td>1</td><td colspan="2">71</td><td colspan="2">71.3455</td><td>0.0016</td></tr>
<tr><td>2</td><td colspan="2">28</td><td colspan="2">24.6466</td><td>0.4562</td></tr>
<tr><td>3</td><td>5</td><td rowspan="4">10</td><td>8.5143</td><td rowspan="4">13.0079</td><td rowspan="4">0.6955</td></tr>
<tr><td>4</td><td>2</td><td>2.9413</td></tr>
<tr><td>5</td><td>2</td><td>1.0161</td></tr>
<tr><td>≥ 6</td><td>1</td><td>0.5363</td></tr>
<tr><td>Σ</td><td colspan="2">109</td><td colspan="2"></td><td>1.1535</td></tr>
</table>

Teorijske frekvencije računamo pomoću:

$$\begin{aligned} n_1 &= np \\ n_{i+1} &= (1-p)\, n_i,\ i = 1,2,3,4 \\ n_6 &= n - \sum_{i=1}^{5} n_i \end{aligned}$$

Opažena vrijednost testne statistike je

$$h = \sum_{i=1}^{6} \frac{(N_i - n_i)^2}{n_i} = 1.1535.$$

Vrijedi $k = 3, d = 1, df = k - d - 1 = 1, \alpha = 0.05$ pa je $\chi^2_\alpha(df) = \chi^2_{0.05}(1) = 3.8415$. Dakle, kritično područje je

$$[\chi^2_\alpha(df), +\infty\rangle = [3.8415, +\infty\rangle.$$

Budući da je $h = 1.1535 < 3.8415$ zaključujemo da podaci ne podržavaju odbacivanje H_0 na nivou značajnosti $\alpha = 0.05$.

$\triangle$

Zadatak 7.4 Mjeri se duljina života neke vrste žarulja. Testirajte hipotezu da je duljina života žarulja eksponencijalno distribuirana ako su dobiveni podaci:

sati	$[0, 1000\rangle$	$[1000, 2000\rangle$	$[2000, 3000\rangle$	$[3000, 4000\rangle$	$[4000, +\infty\rangle$
broj žarulja	124	51	18	10	3

Nivo značajnosti je $\alpha = 0.05$.

Rješenje: $n = 124 + 5 + 18 + 10 + 3 = 206$

Neka je X =duljina života žarulje. Testiramo

$$\begin{aligned} &H_0 : X \sim \text{Exp}(\lambda), \\ &H_1 : \text{ ne } H_0 \end{aligned}$$

Procijenimo parametar λ metodom maksimalne vjerodostojnosti. Prvo definiramo

$$\begin{aligned} p_1 &= \mathbb{P}(0 \le X < 1000 | H_0) = 1 - e^{-1000\lambda}, \\ p_2 &= \mathbb{P}(1000 \le X < 2000 | H_0) = e^{-1000\lambda}(1 - e^{-1000\lambda}), \\ p_3 &= \mathbb{P}(2000 \le X < 3000 | H_0) = e^{-2000\lambda}(1 - e^{-1000\lambda}), \\ p_4 &= \mathbb{P}(3000 \le X < 4000 | H_0) = e^{-3000\lambda}(1 - e^{-1000\lambda}), \\ p_5 &= \mathbb{P}(X \ge 4000 | H_0) = e^{-4000\lambda}. \end{aligned}$$

Funkcija vjerodostojnosti $L\colon \langle 0, +\infty\rangle \to \mathbb{R}$ je

$$L(\lambda) = \prod_{i=1}^{5} p_i^{N_i} = (1 - e^{-1000\lambda})^{203} e^{-129000\lambda}.$$

Maksimiziramo funkciju

$$l(\lambda) = \ln L(\lambda) = 203 \ln(1 - e^{-1000\lambda}) - 129000\lambda.$$

Deriviranjem dobivamo

$$\lambda = -\frac{\ln \frac{129}{332}}{1000} = 0.00094532.$$

Stoga su opažene frekvencije N_i i teorijske frekvencije $n_i = n \cdot p_i$ dane tablicom:

i	N_i		n_i		$(N_i - n_i)^2/n_i$
1	124		125.9577		0.0304
2	51		48.9413		0.0866
3	18		19.0163		0.0543
4	10	13	7.3888	12.0842	0.0694
5	3		4.6954		
Σ	206				0.2407

Realizacija testne statistike je

$$h = \sum_{i=1}^{5} \frac{(N_i - np_i)^2}{np_i} = 0.2407.$$

Ovdje je $k = 4, d = 1, df = k - d - 1 = 2$ pa je $\chi^2_{0.05}(2) = 5.9915 > 0.2407 = h$ pa ne odbacujemo H_0.

$\triangle$

7.2 Kolmogorov-Smirnovljev test

Kolmogorov-Smirnovljevim testom se testira pripadnost nekoj neprekidnoj razdiobi F_0:

$$\begin{aligned} \mathrm{H}_0 &: \ \mathrm{F} = \mathrm{F}_0 \\ \mathrm{H}_1 &: \ \text{ne } \mathrm{H}_0 \end{aligned}$$

Definiramo empirijsku funkciju distribucije:

$$\hat{F}_n(x) = \frac{1}{n} \sum_{k=1}^{n} 1_{\{X_i \le x\}}, \ x \in \mathbb{R}.$$

Realizaciju slučajnog uzorka $x_1, \ldots, x_n$ uredimo i dobijemo $x_{(1)}, \ldots, x_{(n)}$. Testna statistika:

$$D_n = \sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F_0(x)| = \max_{1\le i\le n}\{\max\{|\frac{i-1}{n} - F_0(x_{(i)})|, |\frac{i}{n} - F_0(x_{(i)})|\}\}$$

Kritično područje je

$$[d_\alpha(n), +\infty\rangle,$$

gdje $d_\alpha(n)$ čitamo iz tablice.

Zadatak 7.5 Nezavisnim mjerenjem neke veličine dobiveni su sljedeći podaci

$$0.07 \quad 0.30 \quad 0.51 \quad 0.54 \quad 0.95.$$

Dolaze li gornji podaci iz uniformne razdiobe na intervalu $[0, 1]$? ($\alpha = 0.05$)

Rješenje: $n = 5$

Testiramo:

$$\begin{aligned} &\mathrm{H}_0: \ \mathrm{F}=\mathrm{F}_0 \\ &\mathrm{H}_1: \ \text{ne } \mathrm{H}_0 \end{aligned}$$

gdje je

$$F_0(x) = \begin{cases} 0, & x < 0 \\ x, & 0 \le x \le 1 \\ 1, & x > 1 \end{cases}$$

i	$x_{(i)}$	$\frac{i-1}{5}$	$\frac{i}{5}$	$F_0(x_{(i)})$	$\lvert\frac{i-1}{5} - F_0(x_{(i)})\rvert$	$\lvert\frac{i}{5} - F_0(x_{(i)})\rvert$	max
1	0.07	0	0.2	0.07	0.07	0.13	0.13
2	0.30	0.2	0.4	0.30	0.10	0.10	0.10
3	0.51	0.4	0.6	0.51	0.11	0.09	0.11
4	0.54	0.6	0.8	0.54	0.06	0.26	0.26
5	0.95	0.8	1	0.95	0.15	0.05	0.15
							$\underline{d = 0.26}$

Iz tablice slijedi da je $d_{0.05}(5) = 0.5633$ pa iz $d = 0.26 < 0.5633 = d_{0.05}(5)$ zaključujemo da ne odbacujemo hipotezu H_0.

U R-u:

```
> podaci<-c(0.07,0.3,0.51,0.54,0.95)
> ks.test(podaci,"punif",0,1,alternative="two.sided")

One-sample Kolmogorov-Smirnov test

data:  podaci
D = 0.26, p-value = 0.8123
alternative hypothesis: two-sided
```

$\triangle$

Zadatak 7.6 Testirajte hipotezu da podaci dolaze iz jedinične normalne razdiobe uz razinu značajnosti $\alpha = 0.05$:

1.464	1.137	3.455	0.677	0.932	1.296	0.812	2.298
1.241	0.043	1.060	−1.526	0.469	−0.588	−0.190	−0.865

Rješenje: $n = 16$

Testiramo:

$$\begin{array}{ll} \mathrm{H}_0: & \mathrm{F}=\mathrm{F}_0 \\ \mathrm{H}_1: & \text{ne } \mathrm{H}_0 \end{array}$$

gdje je

$$F_0(x) = \Phi(x) \text{ (čitamo iz tablice jedinične normalne razdiobe)}$$

i	$x_{(i)}$	$\frac{i-1}{16}$	$\frac{i}{16}$	$F_0(x_{(i)})$	$\lvert\frac{i-1}{16} - F_0(x_{(i)})\rvert$	$\lvert\frac{i}{16} - F_0(x_{(i)})\rvert$	max
1	−1.526	0	0.0625	0.0635	0.0635	0.0010	0.0635
2	−0.865	0.0625	0.1250	0.1935	0.1310	0.0685	0.1310
3	−0.588	0.1250	0.1875	0.2783	0.1533	0.0908	0.1533
4	−0.190	0.1875	0.2500	0.4247	0.2372	0.1747	0.2372
5	0.043	0.2500	0.3125	0.5171	0.2671	0.2046	0.2671
6	0.469	0.3125	0.3750	0.6805	0.3680	0.3055	0.3680
7	0.677	0.3750	0.4375	0.7508	0.3758	0.3133	0.3758
8	0.812	0.4375	0.5000	0.7916	0.3541	0.2916	0.3541
9	0.932	0.5000	0.5625	0.8243	0.3243	0.2618	0.3243
10	1.060	0.5625	0.6250	0.8554	0.2929	0.2304	0.2929
11	1.137	0.6250	0.6875	0.8722	0.2472	0.1847	0.2472
12	1.241	0.6875	0.7500	0.8927	0.2052	0.1427	0.2052
13	1.296	0.7500	0.8125	0.9025	0.1525	0.0900	0.1525
14	1.464	0.8125	0.8750	0.9284	0.1159	0.0534	0.1159
15	2.298	0.8750	0.9375	0.9892	0.1142	0.0517	0.1142
16	3.455	0.9375	1.0000	0.9997	0.0622	0.0003	0.0622
							$\underline{d = 0.3758}$

Iz tablice slijedi da je $d_{0.05}(16) = 0.3273$ pa iz $d = 0.3758 > 0.3273 = d_{0.05}(5)$ zaključujemo da odbacujemo hipotezu H_0.

U R-u:

```
>podaci=c(1.464,1.137,3.455,0.677,0.932,1.296,0.812,2.298,1.241,0.043,1.060,
-1.526,0.469,-0.588,-0.190,-0.865)
```

```
> ks.test(podaci,"pnorm",0,1,alternative="two.sided")

One-sample Kolmogorov-Smirnov test

data:  podaci
D = 0.3758, p-value = 0.01543
alternative hypothesis: two-sided
```

$\triangle$

7.3 χ^2-test o nezavisnosti

Pretpostavimo da imamo slučajan uzorak $(X_1, Y_1), \ldots, (X_n, Y_n)$ iz dvodimenzionalnog diskretnog statističkog obilježja (X, Y) pri čemu slučajna varijabla X može poprimiti vrijednosti $\{a_1, \ldots, a_r\}$, a Y vrijednosti $\{b_1, \ldots, b_c\}$. Stavimo li

$$N_{ij} = \sum_{k=1}^{n} \mathbb{1}_{\{(X_k,Y_k)=(a_i,b_j)\}}, N_i = \sum_{j=1}^{c} N_{ij}, M_i = \sum_{i=1}^{r} N_{ij},$$

Kontingencijska tablica glasi:

$X \backslash Y$	b_1	b_2	$\cdots$	b_c	Σ
a_1	N_{11}	N_{12}	$\cdots$	N_{1c}	N_1
a_2	N_{21}	N_{22}	$\cdots$	N_{2c}	N_2
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
a_r	N_{r1}	N_{r2}	$\cdots$	N_{rc}	N_r
Σ	M_1	M_2	$\cdots$	M_c	n

Stavimo

$$\begin{aligned} p_{ij} &= \mathbb{P}((X,Y) = (a_i, b_j)), \\ p_i &= \mathbb{P}(X = a_i), \\ q_j &= \mathbb{P}(Y = b_j). \end{aligned}$$

Testiramo

$$\begin{aligned} &\mathrm{H}_0 : \text{X i Y su nezavisna obilježja tj. } p_{ij} = p_i q_j, \forall i, j \\ &\mathrm{H}_1 : \text{X i Y nisu nezavisna obilježja tj. } \exists i, j, p_{ij} \neq p_i q_j. \end{aligned}$$

Testna statistika:

$$H_n = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(N_{ij} - n\hat{p}_i\hat{q}_j)^2}{n\hat{p}_i\hat{q}_j} = n \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(\hat{p}_{ij} - \hat{p}_i\hat{q}_j)^2}{\hat{p}_i\hat{q}_j},$$

gdje je

$$\hat{p}_i = \frac{N_i}{n}, \hat{q}_j = \frac{M_j}{n}, \hat{p}_{ij} = \frac{N_{ij}}{n}.$$

Kritično područje: $[\chi^2_\alpha(df), +\infty\rangle$, gdje je $df = rc - (r-1) - (c-1) - 1 = (r-1)(c-1)$.

Zadatak 7.7 Neki fakultet ima četiri smjera: elektrotehnika, brodogradnja, strojarstvo i računarstvo. Odabran je slučajni uzorak od 500 studenata i dobiveni podaci su dani sljedećom tablicom:

	elektrotehnika	brodogradnja	strojarstvo	računarstvo	Σ
student	100	80	70	50	300
studentica	50	50	50	50	200
Σ	150	130	120	100	500

Ovisi li odabir smjera o spolu?($\alpha = 0.05$)

Rješenje: Testiramo nezavisnost obilježja X koje poprima vrijednosti: student, studentica i obilježja Y koje poprima vrijednosti: elektrotehnika, brodogradnja, strojarstvo, računarstvo. Očekivane frekvencije $n\hat{p}_i\hat{q}_j$ dane su tablicom:

	elektrotehnika	brodogradnja	strojarstvo	računarstvo	Σ
student	90	78	72	60	300
studentica	60	52	48	40	200
Σ	150	130	120	100	500

Tada je:

$$\begin{aligned} h = {} & \frac{(100-90)^2}{90} + \frac{(80-78)^2}{78} + \frac{(70-72)^2}{72} + \frac{(50-60)^2}{60} + \\ & + \frac{(50-60)^2}{60} + \frac{(50-52)^2}{52} + \frac{(50-48)^2}{48} + \frac{(50-40)^2}{40} = 7.21. \end{aligned}$$

Broj stupnjeva slobode je $df = (2-1)(4-1) = 3$ pa je $\chi_\alpha(df) = \chi_{0.05}(3) = 7.8147$. Zbog $h = 7.21 < 7.8147 = \chi_{0.05}(3)$ ne odbacujemo hipotezu H_0.

U R-u:

```
> studij<-matrix(c(100,80,70,50,50,50,50,50),nrow=2,byrow=T)
> studij
     [,1] [,2] [,3] [,4]
[1,]  100   80   70   50
[2,]   50   50   50   50
> colnames(studij)<-c("elektrotehnika","brodogradnja","strojarstvo","racunarstvo")
> rownames(studij)<-c("student","studentica")
> studij
```

```
           elektrotehnika brodogradnja strojarstvo racunarstvo
student               100           80          70          50
studentica             50           50          50          50
> chisq.test(studij)

Pearson's Chi-squared test

data:  studij
X-squared = 7.2115, df = 3, p-value = 0.06545
```

$\triangle$

7.4 χ^2-test o homogenosti

Pretpostavimo da imamo $m \geq 2$ populacija te nas zanima razdioba istog diskretnog statističkog obilježja X u tim populacijama. Neka su $X^{(i)}$ slučajne varijable koje predstavljaju X u i-toj populaciji.

Iz svake populacije nezavisno odabiremo slučajni uzorak:

$$\begin{aligned}
&X_1^{(1)}, \dots, X_{n_1}^{(1)} \\
&X_1^{(2)}, \dots, X_{n_2}^{(2)} \\
&\vdots \\
&X_1^{(m)}, \dots, X_{n_m}^{(m)}
\end{aligned}$$

Empirijske frekvencije uzoraka dane su tablicom:

populacija\\X	a_1	a_2	$\cdots$	a_k	Σ
1	N_{11}	N_{12}	$\cdots$	N_{1k}	n_1
2	N_{21}	N_{22}	$\cdots$	N_{2k}	n_2
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
m	N_{m1}	N_{m2}	$\cdots$	N_{mk}	n_m
Σ	M_1	M_2	$\cdots$	M_k	n

Neka je

$$X^{(i)} \sim \begin{pmatrix} a_1 & \cdots & a_k \\ p_{i1} & \cdots & p_{ik} \end{pmatrix}, i = 1, \dots, m,$$

$$p_j = \mathbb{P}(X = a_j), j = 1, \dots, k.$$

Testiramo

$$\mathrm{H}_0 : X^{(1)} \overset{D}{=} X^{(2)} \overset{D}{=} \cdots \overset{D}{=} X^{(m)} \text{ tj. } p_{ij} = p_j, j = 1, \ldots, k, i = 1, \ldots, m.$$

$$\mathrm{H}_1 : \exists i, j, \text{ t.d. } X^{(i)} \overset{D}{\neq} X^{(j)}$$

Testna statistika:

$$H_n = \sum_{i=1}^{m} \sum_{j=1}^{k} \frac{(N_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}.$$

gdje je

$$\hat{n}_{ij} = \frac{n_i M_j}{n}.$$

Kritično područje: $[\chi^2_\alpha(df), +\infty\rangle$, gdje je $df = (m-1)(k-1)$.

Zadatak 7.8 Za obradu određenog nastavnog gradiva primjenjene su dvije različite nastavne metode. Metoda M_1 primjenjena je u skupini A od 100 učenika, a metoda M_2 u skupini B od 200 učenika. Da bi se utvrdio učinak, svi su učenici ispitani i ocijenjeni odgovarajućom ocjenom od 1 do 5:

skupina \ ocjena	1	2	3	4	5	Σ
A	14	26	34	16	10	100
B	18	36	58	56	32	200
Σ	32	62	92	72	42	300

Jesu li obje metode jednako učinkovite ($\alpha = 0.05$)?

Rješenje: Testiramo:

H_0 : nizovi A i B potječu iz iste diskretne razdiobe s vrijednostima $\{1, 2, 3, 4, 5\}$.

Očekivane frekvencije su dane sljedećom tablicom:

skupina \ ocjena	1	2	3	4	5	Σ
A	32/3	62/3	92/3	24	14	100
B	64/3	124/3	184/3	48	28	200
Σ	32	62	92	72	42	300

Broj stupnjeva slobode je $df = (2-1)(5-1) = 4$ pa je $\chi_\alpha(df) = \chi_{0.05}(4) = 9.4877$. Zbog $h = 9.8847 > 9.4877 = \chi_{0.05}(4)$ odbacujemo hipotezu H_0.

U R-u:

```
> ocjene<-matrix(c(14,26,34,16,10,18,36,58,56,32),nrow=2,byrow=T)
> ocjene
```

```
     [,1] [,2] [,3] [,4] [,5]
[1,]   14   26   34   16   10
[2,]   18   36   58   56   32
> colnames(ocjene)<-c("1","2","3","4","5")
> rownames(ocjene)<-c("A","B")
> ocjene
   1  2  3  4  5
A 14 26 34 16 10
B 18 36 58 56 32
> chisq.test(ocjene)

Pearson's Chi-squared test

data:  ocjene
X-squared = 9.8848, df = 4, p-value = 0.04241
```

$\triangle$

7.5 Zadaci za vježbu

7.9 Potomstvo dobiveno križanjem dvije vrste biljaka se pojavljuje u tri genotipa označena s A, B i C. Teoretski model o genskom nasljeđivanju sugerira da bi se tipovi genotipa A, B i C trebali pojavljivati u omjeru 1:2:1. Izvedena je eksperimentalna provjera tog genetskog modela na 90 biljaka i dobiveni su sljedeći rezultati:

Genotip	A	B	C
Frekvencija	18	44	28

Potvrđuju li ovi podaci teoretski model ($\alpha = 0.05$)?

7.10 Broj knjiga posuđenih u nekoj knjižnici u jednom tjednu dan je sljedećom tablicom:

Dan	Pon	Uto	Sri	Čet	Pet
Broj posuđenih knjiga	135	108	120	114	146

Ovisi li broj posuđenih knjiga o danu u tjednu ($\alpha = 0.05$)?

7.11 Broj kvarova nekog stroja po smjeni je zabilježen za 180 smjena i dobiveni su sljedeći podaci:

Broj kvarova	0	1	2	3	4	5	6
Broj smjena	82	42	31	12	8	3	2

(a) Koji model bi bio dobar za ovaj tip podataka?

(b) Testirajte na razini značajnosti od $\alpha = 0.05$ da je vaš model dobar za dane podatke.

7.12 Izvršeno je 69 mjerenja neke slučajne pojave. Podaci su dani tablicom:

i	0	1	2	3	4	5	6	≥ 7
N_i	14	24	17	9	2	2	1	0

Uz nivo značajnosti od $\alpha = 0.05$ testirajte hipotezu da mjerena pojava X ima Poissonovu distribuciju.

7.13 Benfordov zakon tvrdi da prva znamenka slučajno odabrane prirodne, fizikalne konstante neće biti uniformno distribuirana, već da znamenka d dolazi s vjerojatnošću $\ln(1 + \frac{1}{d}) / \ln(10)$, $d = 1, 2, ..., 9$. Iz jedne tablice s 52 fizikalne konstante dobiveni su podaci

d	1	2	3	4	5	6	7	8	9
f_d	17	11	3	5	5	4	3	1	3

gdje je f_d broj konstanti čija je prva znamenka d. Testirajte Benfordov zakon na ovom uzorku na nivou značajnosti od 5%.

7.14 Mjeri se vijek trajanja neke vrste žarulja. Dobiveni su sljedeći podaci:

0.4	53.3	254.7	41.1	220.6	201.6
73.4	143.3	108.8	131.7	54.0	233.1

Na razini značajnosti od $\alpha = 0.05$ testirajte hipotezu da podaci dolaze iz Exp(0.005) razdiobe.

7.15 Mjeri vlačna čvrstoća čelične žice (MPa). Dobiveni su sljedeći podaci:

285	341	323	300	313
294	305	317	286	312

Na razini značajnosti od $\alpha = 0.05$ testirajte hipotezu da podaci dolaze iz N(300,289) razdiobe.

7.16 Lanac prehrambenih prozvoda ima u ponudi četiri vrste žitarica A, B, C i D. Lanac je bilježio vrstu prodane žitarice i dob kupca tijekom jednog tjedna. Kupci su podijeljeni u tri starosne kategorije: mlađi od 20 godina, između 20 i 40 godina i stariji od 40 godina. Dobiveni su sljedeći podaci:

	A	B	C	D	Σ
< 20	90	64	78	48	280
20-40	88	78	70	64	300
> 40	62	58	52	48	220
Σ	240	200	200	160	800

Testirajte na nivou značajnosti $\alpha = 0.10$ je li odabir žitarica ovisan o dobi.

7.17 U gradu je provedena anketa o gradnji novog parka u centru. Ponuđeni odgovori su bili da, ne i suzdržan. Uzet je uzorak od 1000 glasača podijeljen u tri starosne skupine. Dobiveni su sljedeći podaci:

	Da	Ne	Suzdržan	Σ
18-30	170	60	20	250
31-50	255	140	55	450
51-70	175	100	25	300
Σ	600	300	100	1000

Testirajte na nivou značajnosti od $\alpha = 0.10$ je li odgovor građana ovisan o dobi.

7.18 Provedena je anketa o referendumu u četiri županije i dobiveni su sljedeći podaci:

	Da	Ne	Suzdržan	Σ
županija 3	18	20	12	50
županija 8	26	16	8	50
županija 11	20	24	6	50
županija 16	28	12	10	50
Σ	92	72	36	200

Je li distribucija glasova u tim županijama jednaka ($\alpha = 0.10$)?

7.19 U New Yorku i Bostonu je proučavana distribucija punoljetnih osoba u tri kategorije: ispod 30 godina, između 30 i 60 godina i iznad 60 godina starosti. Dobiveni su sljedeći podaci:

	< 30	30-60	> 60
New York	51	77	22
Boston	29	63	8

Testirajte na nivou značajnosti $\alpha = 0.05$ utječe li grad na starosnu distribuciju punoljetnih osoba.

Rješeja: **7.9.** podaci potvrđuju teoretski model **7.10.** ne **7.11.** (a) Poissonov model, (b) model nije dobar **7.12.** podaci dolaze iz Poissonove distribucije **7.13.** podaci potvrđuju Benfordov zakon **7.14.** $d = 0.2799$ pa ne odbacujemo H_0 **7.15.** $d = 0.2599$ pa ne odbacujemo H_0 **7.16.** $h = 4.6978 < 10.6446$ pa ne odbacujemo H_0 **7.17.** $h = 11.9907 > 7.7794$ pa odbacujemo H_0 **7.18.** $h = 9.6232 < 10.6446$ pa ne odbacujemo H_0 **7.19.** $h = 4.1493 < 5.9915$ pa ne odbacujemo H_0