(BIO)STATISTIKA

seminari

smjer: **Prehrambena tehnologija i Biotehnologija**

pripremila:
dr.sc. Iva Franjić

Sadržaj

1 DESKRIPTIVNA STATISTIKA

Prilikom opažanja ili eksperimentiranja, pažnja istraživača redovito je usmjerena na jednu ili više veličina. Ako se promatra samo jedna veličina, označimo ju s X, onda je rezultat jednog mjerenja jedan realan broj x. Višestrukim ponavljanjem mjerenja veličine X dobiva se konačni niz brojeva $x_1, x_2, \ldots, x_n$ kao rezultat n ponovljenih mjerenja koji nazivamo **realizacija od** X. Veličina X obično se naziva **statističko obilježje**, a dobiveni niz brojeva $x_1, x_2, \ldots, x_n$ **statistički podaci** o promatranom statističkom obilježju X.

1.1 Grafički prikaz podataka

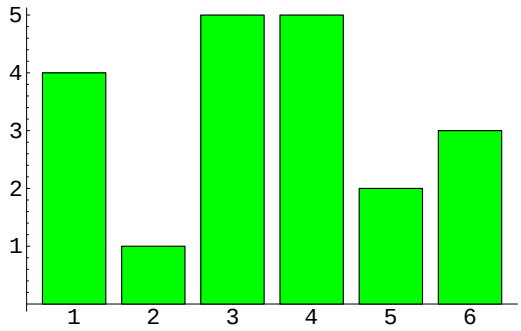
Primjer 1 *Neka X označava broj dobiven bacanjem igrače kocke. Kocku smo bacali 20 puta i dobiveni su sljedeći podaci:* $1, 3, 1, 6, 2, 6, 4, 6, 3, 3, 4, 3, 1, 4, 4, 1, 4, 5, 3, 5$.

- statističko obilježje $X =$ "broj na kocki"
- $\mathrm{Im}X = \{1, 2, 3, 4, 5, 6\} \Rightarrow$ skup svih vrijednosti koje X može poprimiti
- u našem primjeru, $\mathrm{Im}X$ je diskretan, tj. konačan skup, pa kažemo da je X **diskretno obilježje**
- obilježje može biti **numeričko** ili **nenumeričko**
- nenumeričko obilježje nazivamo i **kategorija**; npr. spol, boja i slično; možemo mu pridijeliti neku numeričku vrijednost, ali tada nema smisla računati npr. aritmetičku sredinu podataka!
- svakom elementu $a_i \in \mathrm{Im}X$ možemo pridružiti broj $f_i \Rightarrow$ **frekvencija (učestalost) pojavljivanja elementa** a_i u nizu podataka
- broj $f_{r_i} = \dfrac{f_i}{n}$: **relativna frekvencija** od a_i
  (n je broj ponavljanja pokusa, u ovom primjeru $n = 20$)

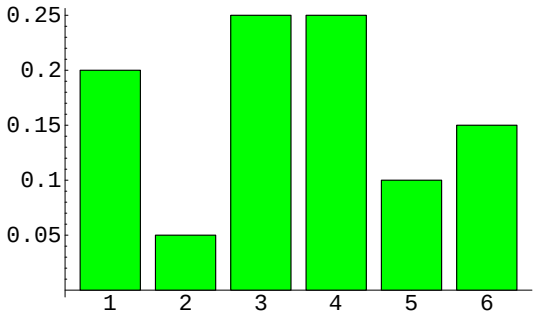
Prikažimo podatke u **TABLICI FREKVENCIJA**

a_i	f_i	f_{r_i}
1	4	$\frac{4}{20} = 0.2$
2	1	$\frac{1}{20} = 0.05$
3	5	$\frac{5}{20} = 0.25$
4	5	$\frac{5}{20} = 0.25$
5	2	$\frac{2}{20} = 0.1$
6	3	$\frac{3}{20} = 0.15$
Σ	20	1.00

GRAFIČKI PRIKAZ PODATAKA POMOĆU **STUPČASTOG DIJAGRAMA (BAR - CHART)**



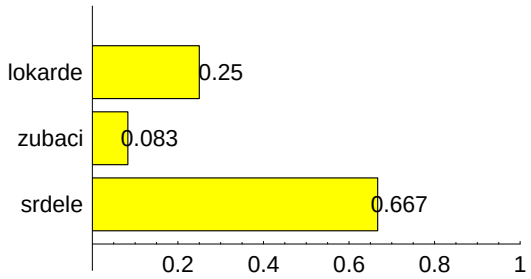
Stupčasti dijagram može se crtati i tako da ukupna površina stupića bude jednaka 1, što je bolje zbog usporedbe, npr. za različite n:



Primjer 2 *U uzorku od 144 ribe ulovljene u Bračkom kanalu, bilo je 36 lokardi, 12 zubataca i 96 srdela.*

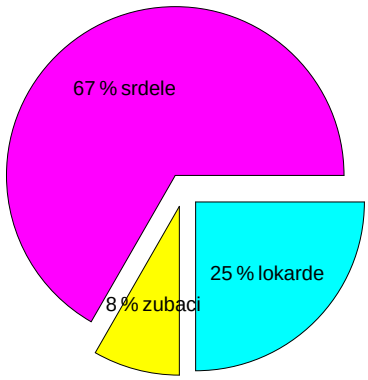
vrsta ribe	frekvencija	relativna frekvencija	
lokarde	36	$\frac{36}{144} = 0.25$	25%
zubaci	12	$\frac{12}{144} = 0.083$	8.3%
srdele	96	$\frac{96}{144} = 0.667$	66.7%
Σ	144	1.00	100%

HORIZONTALNI STUPČASTI DIJAGRAM



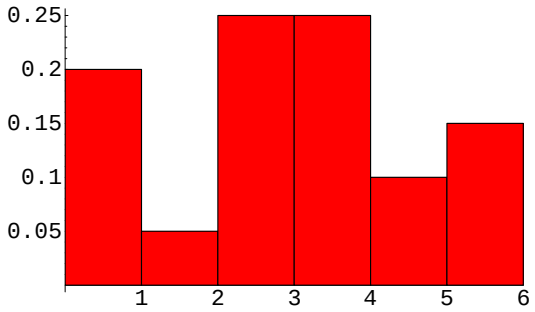
STRUKTURNI KRUG (PIE CHART)

-ako imamo relativno malo različitih vrijednosti koje statističko obilježje može poprimiti



Primjer 3 *Nacrtajmo* ***histogram*** *za podatke iz Primjera 1.*

- svaka 2 susjedna stupića se dodiruju i svaki ima težište u vrijednosti visina f_i ili f_{r_i}
- površina svakog stupića jednaka je relativnoj frekvenciji pa je površina ispod cijelog grafa jednaka je 1
- nema smisla za nenumeričke vrijednosti



Primjer 4 *Mjerena je visina (u metrima) 30 20-ogodišnjaka. Dobiveni su podaci: 1.85, 1.88, 1.78, 1.72, 1.80, 1.72, 1.75, 1.72, 1.79, 1.82, 1.69, 1.76, 1.60, 1.78, 1.76, 1.74, 1.70, 1.86, 1.72, 1.75, 1.69, 1.79, 1.83, 1.79, 1.65, 1.76, 1.59, 1.68, 1.74, 1.86.*

- statističko obilježje X = visina $\Rightarrow$ neprekidno statističko obilježje (poprima vrijednosti iz nekog intervala)
- podatke najprije moramo svrstati u razrede:
  1. odredimo x_{min} i x_{max} : $x_{min} = 1.59,\ x_{max} = 1.88$
  2. izaberemo adekvatan broj razreda (okvirno: $\sqrt{n}$) $\Rightarrow$ k = 6
  3. odredimo zajedničku širinu razreda:

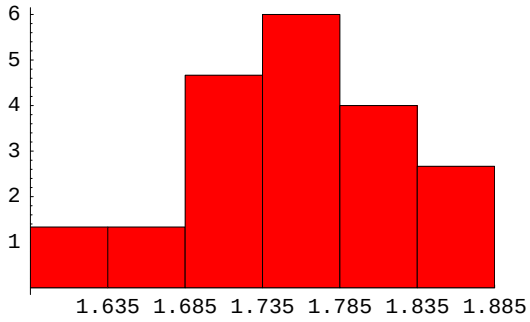
$$c = \frac{x_{max} - x_{min}}{k} = \frac{1.88 - 1.59}{6} = 0.0483 \ \Rightarrow \ \mathbf{c{=}0.05}$$

  (uvijek zaokružujemo na više!)

4. odredimo razrede (tj. lijevi prag razreda): $I_1, ..., I_k$
   pritom $I_1 \cup I_2 \cup ... \cup I_k$ mora obuhvaćati sve podatke
   $I_i = [a_{i,1}, a_{i,2}], \quad a_{i,2} = a_{i+1,1}$
   $I_{i+1} = [a_{i+1,1}, a_{i+1,2}]$

RAZREDI	f_i	$f_{r_i} = f_i/n$	f_{r_i}/c
$I_1 = [1.585, 1.635]$	2	0.067	1.34
$I_2 = [1.635, 1.685]$	2	0.067	1.34
$I_3 = [1.685, 1.735]$	7	0.233	4.66
$I_4 = [1.735, 1.785]$	9	0.3	6
$I_5 = [1.785, 1.835]$	6	0.2	4
$I_6 = [1.835, 1.885]$	4	0.133	2.66
Σ	30	1	20

Nacrtajmo histogram za ove podatke. Širina stupića više nije proizvoljna (sada je jednaka širini razreda, tj. c=0.05), pa da bi suma površina svih pravokutnika (odnosno površina ispod grafa) bila jednaka 1, na ordinatu ucrtavamo $\frac{f_{r_i}}{c}$ a ne f_{r_i}. Naime, $20 \cdot c = 20 \cdot 0.05 = 1$.



STEM AND LEAF DIJAGRAM

stem	leaf
1.5	9
1.6	90958
1.7	82252968640259964
1.8	5802636

stem	leaf
1.5	9
1.6_*	0
1.6^*	5899
1.7_*	2224024
1.7^*	8596865996
1.8_*	023
1.8^*	5866

1.2 Srednje vrijednosti uzorka

1.2.1 Aritmetička sredina uzorka

Aritmetička sredina uzorka je broj

$$\bar{x} := \frac{1}{n}(x_1 + x_2 + \ldots + x_n).$$

Ako je $\mathrm{Im}X = \{a_1, a_2, \ldots, a_k\}$ i pritom se a_i u uzroku ponavlja f_i puta, tada

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} f_i \cdot a_i, \quad n = \sum_{i=1}^{k} f_i.$$

- ima smisla samo za numeričke podatke

Primjer 5 *Izračunajte $\bar{x}$ za podatke iz Primjera 4.*

Rješenje:

$$\begin{aligned}\bar{x} &= \frac{1}{30}(1.59 + 1.60 + 1.65 + 1.68 + 2 \cdot 1.69 + 4 \cdot 1.72 + 1.70 + 2 \cdot 1.74 + 2 \cdot 1.75 \\ &\quad +3 \cdot 1.76 + 2 \cdot 1.78 + 3 \cdot 1.79 + 1.80 + 1.82 + 1.83 + 1.85 + 2 \cdot 1.86 + 1.88) \\ &= \frac{52.57}{30} \approx 1.75\end{aligned}$$

□

1.2.2 Medijan uzorka

- uredimo podatke (sortiramo ih po veličini): $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$
- ima smisla samo za numeričke podatke

Medijan uzorka je broj za koji vrijedi da je 50% svih podataka manje od ili jednako njemu i 50% svih podataka veće od ili jednako njemu.

Ako je broj podataka neparan, tj $n = 2k - 1,\ k \in \mathbf{N}$, tada je $m = x_{(k)}$. Za paran n $(n = 2k)$, vrijedi

$$m = \frac{x_{(k)} + x_{(k+1)}}{2}.$$

Općenito, $m = x_{(\frac{n+1}{2})}$. Vrijedi

$$x_{(\frac{p}{q})} = x_{(k+\frac{r}{q})}$$
$$x_{(\frac{p}{q})} := x_{(k)} + \frac{r}{q}\left(x_{(k+1)} - x_{(k)}\right)$$

Primjer 6 *Nađite medijan uzorka za podatke iz Primjera 1.*

Rješenje: Sortiramo podatke po veličini:

$$1 \leq 1 \leq 1 \leq 1 \leq 2 \leq 3 \leq 3 \leq 3 \leq 3 \leq \mathbf{3} \leq \mathbf{4} \leq 4 \leq 4 \leq 4 \leq 4 \leq 5 \leq 5 \leq 6 \leq 6 \leq 6$$
$$n = 20 = 2 \cdot 10$$
$$m = \frac{x_{(10)} + x_{(11)}}{2} = \frac{3+4}{2} = 3.5$$

□

1.2.3 Uzorački mod

Mod je ona vrijednost statističkog obilježja koja se u uzorku javlja s najvećom frekvencijom.

- koristan kod statističkih obilježja koja nisu numerička, pa nema aritmetičke sredine

- BIMODALNI UZORAK: uzorak u kojem postoje 2 vrijednosti s jednakom frekvencijom
- UNIMODALNI UZORAK: uzorak u kojem postoji samo jedan mod
- Ako svi podaci imaju istu frekvenciju pojavljivanja u uzorku, tada uzorak nema mod.

Primjer 7 *Nađite mod za podatke iz Primjera 1 i 4.*

Rješenje:

- u Primjeru 1: mod = 3 & mod=4 $\Rightarrow$ bimodalan uzorak
- u Primjeru 4: mod = 1.72

□

1.3 Mjere disperzije ili varijabiliteta

1.3.1 Raspon uzorka

Neka je $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ uređeni niz podataka. Broj

$$d = x_{(n)} - x_{(1)}$$

naziva se **raspon uzorka.**

Primjer 8 *Odredite raspon uzorka iz Primjera 4.*

Rješenje:

$$d = 1.88 - 1.59 = 0.29$$

□

1.3.2 Interkvartil

Donji kvartil q_L je ona vrijednost uzroka za koju vrijedi da je 25% svih podataka manje ili jednako od nje i 75% svih podataka veće ili jednako od nje.

$$q_L = x_{(\frac{n+1}{4})}$$

Gornji kvartil q_U je ona vrijednost uzroka za koju vrijedi da je 75% svih podataka manje ili jednako od nje i 25% svih podataka veće ili jednako od nje.

$$q_U = x_{(\frac{3(n+1)}{4})}$$

Interkvartil: $d_q = q_U - q_L$

Primjer 9 *Odredite interkvartil za podatke iz Primjera 4.*

Rješenje:

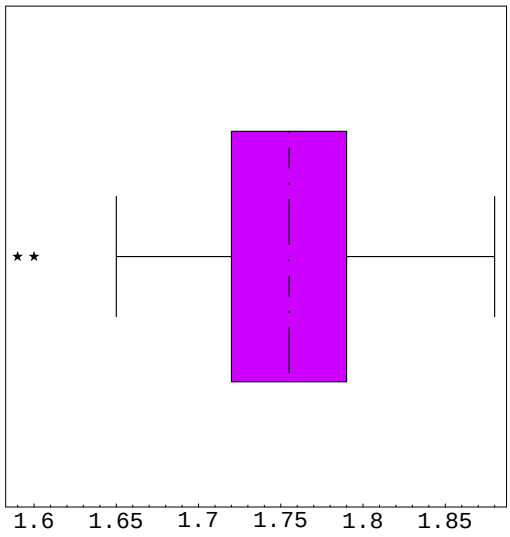
$$\begin{aligned}
q_L &= x_{(\frac{n+1}{4})} = x_{(\frac{30+1}{4})} = x_{(7+\frac{3}{4})} = x_{(7)} + \frac{3}{4}\left(x_{(8)} - x_{(7)}\right) \\
&= 1.70 + \frac{3}{4}(1.72 - 1.70) = 1.715 \approx 1.72 \\
q_U &= x_{(\frac{3(n+1)}{4})} = x_{(\frac{93}{4})} = x_{(23+\frac{1}{4})} = x_{(23)} + \frac{1}{4}\left(x_{(24)} - x_{(23)}\right) \\
&= 1.79 + \frac{1}{4}(1.80 - 1.79) = 1.7925 \approx 1.79 \\
d_q &= q_U - q_L = 1.79 - 1.72 = 0.07
\end{aligned}$$

□

Uređenu petorku $(x_{(1)}, q_L, m, q_U, x_{(n)})$ zovemo **karakteristična petorka uzorka.** Pomoću nje crtamo tzv. **"box and whisker" dijagram**, odnosno dijagram pravokutnika.

Primjer 10 *Nacrtajte box and whisker dijagram za podatke iz Primjera 4.*

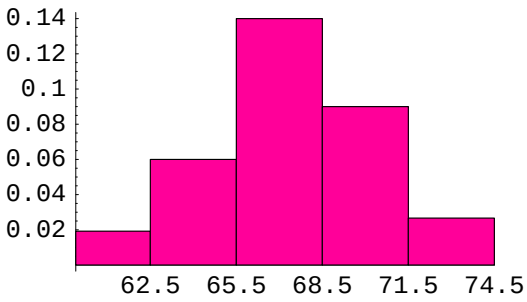
$x_{(1)} = 1.59,\ q_L = 1.72,\ m = 1.75,\ q_U = 1.79,\ x_{(30)} = 1.88,\ d_q = 0.07$



Zadatak 1 *U tablici su dane težine 100 studenata PBF-a. Nacrtajte histogram, nađite aritmetičku sredinu, medijan te interkvartil ovog uzorka.*

težina (kg)	*broj studenata*	*sredina razreda*	f_{r_i}	f_{r_i}/c
$60-62$	5	61	0.05	0.017
$63-65$	18	64	0.18	0.06
$66-68$	42	67	0.42	0.14
$69-71$	27	70	0.27	0.09
$72-74$	8	73	0.08	0.027
Σ	100		1	0.334

Rješenje:



Aritmetička sredina:

$$\bar{x} = \frac{1}{100}(61 \cdot 5 + 64 \cdot 18 + 67 \cdot 42 + 70 \cdot 27 + 73 \cdot 8) = 67.45$$

Medijan: U prva 2 razreda upada 5+18=23 podataka, a u prva 3 razreda 5+18+42=65 podataka, što znači da se medijan nalazi negdje unutar 3.razreda, tj. $65.5 \leq m \leq 68.5$. Medijan dobivamo interpolacijom:

$$m = 65.5 + \frac{27}{42}(68.5 - 65.5) = 65.5 + \frac{27}{42} \cdot 3 = 67.43$$

Vrijednost medijana može se očitati i sa histograma - medijan je apscisa koja odgovara liniji koja dijeli histogram na 2 dijela jednake površine.

Interkvartil: Najprije moramo odrediti donji i gornji kvartil. Postupak je sličan kao kod određivanja medijana - donji kvartil nalazi se negdje unutar 3.razreda tj. $65.5 \leq q_L \leq 68.5$, dok se gornji kvartil nalazi unutar 4.razreda (budući prva 3 razreda sadrže 65, a prva 4: 5+18+42+27=92 podatka), tj. $68.5 \leq q_U \leq 71.5$. Imamo:

$$\begin{aligned} q_L &= 65.5 + \frac{2}{42}(68.5 - 65.5) = 65.5 + \frac{2}{42} \cdot 3 = 65.643 \\ q_U &= 68.5 + \frac{10}{27}(71.5 - 68.5) = 68.5 + \frac{10}{27} \cdot 3 = 69.61 \\ d_q &= q_U - q_L = 69.61 - 65.643 = 3.967 \end{aligned}$$

□

Definirajmo još i **koeficijent kvartilne varijacije**:

$$v_q = \frac{d_q}{q_L + q_U}$$

-koristan kada varijabilitet želimo izraziti pomoću RELATIVNE veličine (neovisne o mjernim jedinicama)

1.3.3 Uzoračka varijanca i uzoračka standardna devijacija

Uzoračka varijanca:

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

Uzoračka standardna devijacija:

$$s = +\sqrt{s^2}$$

Vrijedi:

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

Ovaj oblik formule je puno praktičniji za računanje.

Ako se u uzroku $x_1, x_2, \ldots, x_n$ vrijednosti $a_1, a_2, \ldots, a_k$ pojavljuju s frekvencijom $f_1, f_2, \ldots, f_k$, onda vrijedi:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{k}(a_i-\bar{x})^2 \cdot f_i = \frac{1}{n-1}\left(\sum_{i=1}^{k}f_i \cdot a_i^2 - n\bar{x}^2\right)$$

Primjer 11 *Izračunajte uzoračku varijancu s^2 i uzoračku standardnu devijaciju s za podatke iz Primjera 4.*

Rješenje:

$$\begin{aligned} s^2 &= \frac{1}{n-1}\left(\sum_{i=1}^{k}f_i \cdot a_i^2 - n\bar{x}^2\right) = \frac{1}{29}\left[(1.59^2 + 1.60^2 + 1.65^2 + 1.68^2 + 2\cdot 1.69^2\right. \\ &+ 1.70^2 + 4\cdot 1.72^2 + 2\cdot 1.74^2 + 2\cdot 1.75^2 + 3\cdot 1.76^2 + 2\cdot 1.78^2 + 3\cdot 1.79^2 \\ &+ \left. 1.80^2 + 1.82^2 + 1.83^2 + 1.85^2 + 2\cdot 1.86^2 + 1.88^2) - 30\cdot 1.75^2\right] \approx 0.0051 \\ s &= +\sqrt{s^2} = 0.071 \end{aligned}$$

□

1.4 Mjere lokacije

Medijan, te gornji i donji kvartil spadaju u mjere lokacije. Tu su još i:

- **DECILI**: k-ti uzorački decil je broj

$$D_k = x_{\left(\frac{k(n+1)}{10}\right)}, \quad k = 1, 2, \ldots, 9$$

(k/10 podataka je manje ili jednako njemu)

- **PERCENTILI:** k-ti uzorački percentil je broj

$$P_k = x_{\left(\frac{k(n+1)}{100}\right)}, \quad k = 1, 2, \ldots, 99$$

(k% podataka je manje ili jednako njemu)

- decili su specijalni slučaj percentila: $D_1 = P_{10},\ D_2 = P_{20}, \ldots, D_9 = P_{90}$

Zadatak 2 *Izmjeren je kapacitet na 485 istovrsnih kondenzatora. Rezultati mjerenja su dani sljedećom tablicom frekvencija (podaci su u μF zaokruženi na dvije decimale)*

i	*razred*	f_i	$\bar{a}_i$	d_i	$f_i d_i$	$f_i d_i^2$	f_{r_i}	F_i
1	$19.58 - 19.62$	3	19.60	-6	-18	108	0.006	0.006
2	$19.63 - 19.67$	5	19.65	-5	-25	125	0.010	0.016
3	$19.68 - 19.72$	5	19.70	-4	-20	80	0.010	0.026
4	$19.73 - 19.77$	20	19.75	-3	-60	180	0.041	0.067
5	$19.78 - 19.82$	35	19.80	-2	-70	140	0.072	0.139
6	$19.83 - 19.87$	74	19.85	-1	-74	74	0.153	0.292
7	$19.88 - 19.92$	***92***	19.90	0	0	0	0.190	0.482
8	$19.93 - 19.97$	83	19.95	1	83	83	0.171	0.653
9	$19.98 - 20.02$	70	20.00	2	140	280	0.144	0.797
10	$20.03 - 20.07$	54	20.05	3	162	486	0.111	0.908
11	$20.08 - 20.12$	27	20.10	4	108	432	0.056	0.964
12	$20.13 - 20.17$	12	20.15	5	60	300	0.025	0.989
13	$20.18 - 20.22$	2	20.20	6	12	72	0.004	0.993
14	$20.23 - 20.27$	3	20.25	7	21	147	0.006	0.999
	Σ	485			319	2507		

(1) Nacrtajte histogram. (DZ)

(2) Kako bi procijenili aritmetičku sredinu i varijancu uzroka?

(3) Kako bi procijenili medijan te gornji i donji kvartil?

Rješenje:

$$n = \sum_{i=1}^{14} f_i = f_1 + \ldots + f_{14} = 485$$

Budući je $n = 485$ vrlo velik broj, $\frac{1}{n-1}$ u formuli za s^2 približno je jednak $\frac{1}{n}$. Dovoljno je, dakle, uzeti:

$$s^2 = \frac{1}{n}\sum_{i=1}^{k} f_i \cdot (\bar{a}_i - \bar{x})^2 \text{ gdje je } \bar{x} = \frac{1}{n}\sum_{i=1}^{k} f_i \cdot \bar{a}_i$$

Nadalje, širina razreda je $c = 0.05$. Definirajmo:

$$d_i := \frac{\bar{a}_i - \bar{a}_0}{c} \Leftrightarrow \bar{a}_i = \bar{a}_0 + c \cdot d_i,$$

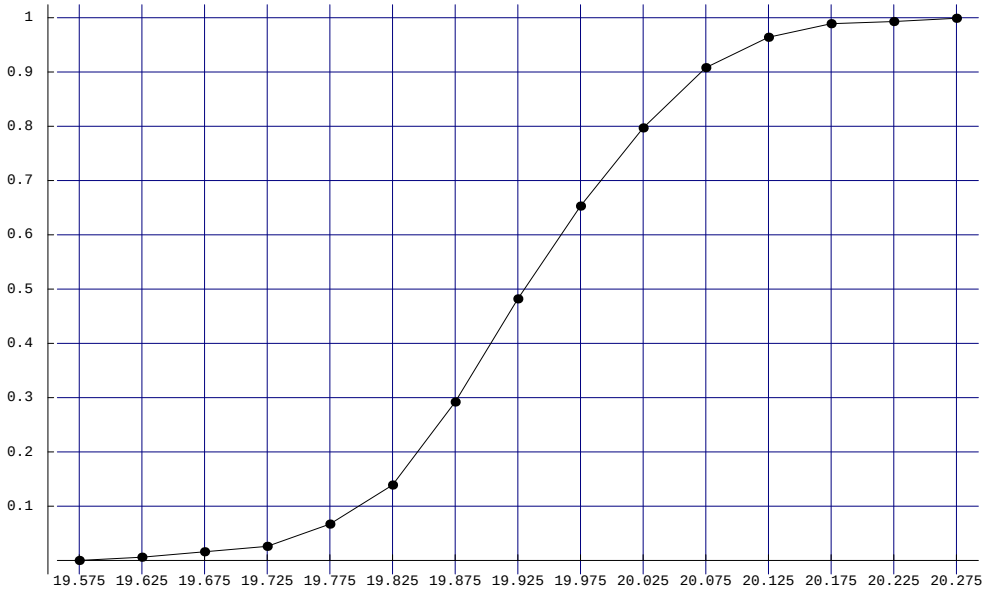
gdje je $\bar{a}_0$ referentna vrijednost aritmetičkog niza $\bar{a}_1, \ldots, \bar{a}_k$. Za $\bar{a}_0$ se obično uzima vrijednost s najvećom frekvencijom. Dakle, $\bar{a}_0$ je mod (ili jedan od). U ovom zadatku $\bar{a}_0 = 19.90$. Imamo:

$$\begin{aligned}
\bar{x} &= \frac{1}{n}\sum_{i=1}^{k} f_i \cdot \bar{a}_i = \frac{1}{n}\sum_{i=1}^{k} f_i(\bar{a}_0 + c \cdot d_i) = \frac{1}{n}\left(\bar{a}_0 \sum_{i=1}^{k} f_i + c\sum_{i=1}^{k} f_i \cdot d_i\right) \\
&= \bar{a}_0 + c \cdot \bar{d}, \text{ gdje je } \bar{d} = \frac{1}{n}\sum_{i=1}^{k} f_i \cdot d_i, \\
s^2 &= \frac{1}{n}\sum_{i=1}^{k} f_i \cdot (\bar{a}_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{k} f_i \cdot (\bar{a}_0 + c \cdot d_i - \bar{a}_0 - c \cdot \bar{d})^2 \\
&= \frac{1}{n}\sum_{i=1}^{k} f_i \cdot (c(d_i - \bar{d}))^2 = c^2 \cdot \frac{1}{n}\sum_{i=1}^{k} f_i \cdot (d_i - \bar{d})^2 = \ldots = c^2\left[\frac{1}{n}\sum_{i=1}^{k} f_i d_i^2 - \bar{d}^2\right]
\end{aligned}$$

Iz podataka dobivamo da je

$$\bar{d} = \frac{319}{485} = 0.658 \Rightarrow \bar{x} = 19.90 + 0.05 \cdot 0.658 = 19.93\ \mu F$$
$$s^2 = 0.05^2\left(\frac{1}{485} \cdot 2507 - 0.658^2\right) = 0.012 \Rightarrow s = 0.11\ \mu F$$

Kod određivanja medijana, te donjeg i gornjeg kvartila pomoći će nam **graf kumulativnih relativnih frekvencija** koji je prikazan na donjoj slici.



Za kumulativne relativne frekvencije F_j vrijedi:

$$F_j = \sum_{i=1}^{j} f_{r_i}, \quad j = 1, \ldots, k$$

Medijan m je x-koordinata točke $(m, 0.5)$ na grafu kumulativnih relativnih frekvencija. Ta točka leži na pravcu određenom točkama $(a_7, F_7) = (19.925, 0.482)$ i $(a_8, F_8) = (19.975, 0.653)$ pa medijan možemo izračunati linearnom interpolacijom:

$$\frac{1}{2} - F_7 = \frac{F_8 - F_7}{a_8 - a_7}(m - a_7)$$
$$\frac{1}{2} - 0.482 = \frac{0.653 - 0.482}{0.05}(m - 19.925) \;\Leftrightarrow\; m = 19.93\ \mu F$$

Slično se mogu izračunati donji q_L i gornji kvartil q_U. Njima su na grafu pridružene, redom, točke $(q_L, 0.25)$ i $(q_U, 0.75)$:

$$\frac{1}{4} - F_5 = \frac{F_6 - F_5}{a_6 - a_5}(q_L - a_5) \;\Leftrightarrow\; q_L = 19.86\ \mu F$$

$$\frac{3}{4} - F_8 = \frac{F_9 - F_8}{a_9 - a_8}(q_U - a_8) \Leftrightarrow q_U = 20.01\ \mu F$$

□

1.5 Mjere oblika

Slično kao što se definira uzoračka varijanca, može se definirati **uzorački k-ti centralni moment**, $k \in \mathbb{N}$:

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

Specijalno,

$$\mu_1 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}) = \frac{1}{n-1}\sum_{n=1}^{n} x_i - \frac{n\bar{x}}{n-1} = \frac{n\bar{x}}{n-1} - \frac{n\bar{x}}{n-1} = 0$$

$$\mu_2 = s^2$$

$$\mu_3 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^3$$

Primjer 12 *Promatrajmo uzorak:* $1, 2, 4, 5$. *Srednja vrijednost tog uzorka je* $\bar{x} = \frac{1}{4}(1+2+4+5) = 3$.
S druge strane, 3.centalni moment tog uzorka je

$$\mu_3 = \frac{1}{3}\left((1-3)^3 + (2-3)^3 + (4-3)^3 + (5-3)^3\right) = 0$$

Odavde možemo zaključiti da kada je uzorak simetričan s obzirom na aritmetičku sredinu, 3.centalni moment $\mu_3 = 0$.

Koeficijent asimetrije uzorka (skewness) definiran je s:

$$\alpha_3 = \frac{\mu_3}{s^3} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^3 = \frac{1}{n-1}\sum_{i=1}^{k} f_i \cdot \left(\frac{a_i - \bar{x}}{s}\right)^3$$

Vrijedi:

(i) $\alpha_3 = 0 \Rightarrow$ uzorak je SIMETRIČAN

(ii) $\alpha_3 > 0 \Rightarrow$ uzorak je POZITIVNO ASIMETRIČAN

(iii) $\alpha_3 < 0 \Rightarrow$ uzorak je NEGATIVNO ASIMETRIČAN

1.6 Linearna regresija

Imamo n parova podataka $(x_i, y_i),\ i = 1, \ldots, n$. Želimo odrediti vezu između nezavisne varijable x (nju možemo kontrolirati) i zavisne varijable y. Pretpostavimo da je veza **linearna**, tj. da je graf pripadajuće funkcije **pravac** $y = \alpha x + \beta$. Želimo odrediti procjenitelj za taj pravac oblika

$$y = \hat{\alpha}x + \hat{\beta}$$

Procjenitelji su:

$$\hat{\alpha} = \frac{s_{xy}}{s_x^2}, \quad \hat{\beta} = \bar{y} - \hat{\alpha}\bar{x},$$

pri čemu:

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)$$
$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)$$
$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right)$$

Zadatak 3 *U tablici su dani podaci o broju emitiranih reklama tijekom mjesec dana za neki proizvod i ostvarenoj zaradi na tom proizvodu. Procijenite pravac regresije za ove podatke.*

broj reklama	16	59	65	43	82	90	31	22
promet (u tisućama kuna)	18	63	28	71	85	98	20	25

Rješenje:

$$n = 8$$
$$\bar{x} = \frac{1}{8}\sum_{i=1}^{8} x_i = \frac{1}{8}(16 + 59 + 65 + 43 + 82 + 90 + 31 + 22) = 51$$
$$\bar{y} = \frac{1}{8}\sum_{i=1}^{8} y_i = \frac{1}{8}(18 + 63 + 28 + 71 + 85 + 98 + 20 + 25) = 51$$
$$\sum_{i=1}^{8} x_i^2 = 16^2 + 59^2 + 65^2 + 43^2 + 82^2 + 90^2 + 31^2 + 22^2 = 26080$$

$$\Rightarrow s_x^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) = \frac{1}{7}\left(26080 - 8\cdot 51^2\right) = 753.143$$

$$\sum_{i=1}^{8} x_i y_i = 16\cdot 18 + 59\cdot 63 + 65\cdot 28 + 43\cdot 71 + 82\cdot 85 + 90\cdot 98 + 31\cdot 20 + 22\cdot 25 = 25838$$

$$\Rightarrow s_{xy} = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right) = \frac{1}{7}\left(25838 - 8\cdot 51\cdot 51\right) = 718.571$$

$$\Longrightarrow \hat{\alpha} = \frac{s_{xy}}{s_x^2} = \frac{718.571}{753.143} = 0.954$$

$$\Longrightarrow \hat{\beta} = \bar{y} - \hat{\alpha}\bar{x} = 51 - 0.954\cdot 51 = 2.346$$

$$\Longrightarrow y = 0.954\cdot x + 2.346$$

□

(PEARSONOV) KOEFICIJENT KORELACIJE

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\cdot\frac{y_i - \bar{y}}{s_y}\right) = \frac{s_{xy}}{s_x\cdot s_y}, \quad -1 \le r \le 1$$

1. $r = 0$ nema korelacije
2. $r > 0$ pozitivna korelacija (ako x raste, i y u pravilu raste)
3. $r < 0$ negativna korelacija (ako x raste, y u pravilu pada)

Primjer 13 *Nađite koeficijent korelacije za podatke iz Zadatka 3.*

Rješenje:

$$s_{xy} = 718.571$$

$$s_x^2 = 753.143 \;\Rightarrow\; s_x = 27.4435$$

$$s_y^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right) = \frac{1}{7}\left(27972 - 8\cdot 51^2\right) = 1023.43 \;\Rightarrow\; s_y = 31.9911$$

$$\Longrightarrow r = \frac{s_{xy}}{s_x\cdot s_y} = \frac{718.571}{27.4435\cdot 31.9911} = 0.818467$$

što je razmjerno visoka korelacija. □

2 Slučajni uzorak i osnovne razdiobe

2.1 Slučajni uzorak

Neka je X statističko obilježje koje izučavamo. Cilj statističke analize je da se na osnovi uzorka iz populacije izvedu određeni zaključci o distribuciji obilježja X.

Recimo da želimo raditi ispitivanje o zečevima (npr.duljini njihovih ušiju) u nekoj šumi. Populacija iz koje izabiremo uzorak su svi zečevi koji žive u toj šumi. Uzorak zečeva biramo na slučajan način. Dvije su različite mogućnosti da to učinimo: nakon što ulovimo zeca i izmjerimo mu dužinu ušiju, možemo ga pustiti i tako omogućiti da ga još (bar) jednom ulovimo te da on uđe u uzorak (bar) dva puta. Druga mogućnost je da ga zadržimo dok ne izaberemo cijeli uzorak kako taj isti zec ne bi ušao u uzorak više od jednog puta.

Slučajni uzorak kojeg uzimamo tako da svaki član populacije može ući u uzorak više od jednog puta zovemo **jednostavni slučajni uzorak s ponavljanjem** (slučaj kada zečeve puštamo natrag u šumu), a ukoliko svaki član populacije može ući u uzorak točno jednom tada se radi o **jednostavnom slučajnom uzorku bez ponavljanja** (slučaj kada zečeve ne puštamo prije nego izaberemo ostatak uzorka).

Bitna razlika između ove dvije vrste biranja uzroka je u tome što je u jednom slučaju populacija konačna a u drugom beskonačna. Naime, ako uzimamo uzorak s vraćanjem, tada možemo uzeti uzorak proizvoljne veličine, veće čak i od ukupnog broja članova same populacije. To je dakako u slučaju uzimanja uzorka bez vraćanja, nemoguće. S druge strane, često se promatraju konačne populacije koje su dovoljno velike da ih s aspekta statističke analize možemo smatrati beskonačnim.

ZADAVANJE I RAČUNANJE VJEROJATNOSTI

Primjer 14 *Bacamo simetričnu kocku. Kolika je vjerojatnost da je pao paran broj?*

Rješenje: Označimo s Ω skup svih elementarnih događaja, tj. skup svih mogućih ishoda pokusa kojeg radimo. $|\Omega|$ je kardinalni broj skupa Ω (ukupan broj njegovih članova). Tada je:

$$\Omega = \{1, 2, 3, 4, 5, 6\}, \quad |\Omega| = 6$$

Označimo s A događaj čiju vjerojatnost računamo. Prebrojimo koliko elementarnih događaja je "povoljno" za događaj A. Dakle, zanimaju nas oni elementarni događaju, tj. ishodi, koji kad se dogode dogodi se i A. Imamo:

$$A = \{\text{na kocki je pao paran broj}\} = \{2, 4, 6\}, \quad |A| = 3$$

Vjerojatnost događaja A računamo kao kvocijent odgovarajućih kardinalnih brojeva:

$$\Rightarrow P(A) = \frac{|A|}{|\Omega|} = \frac{3}{6} = \frac{1}{2}$$

□

Primjer 15 *Bacamo 2 simetrične kocke. Kolika je vjerojatnost da zbroj na te 2 kocke bude jednak 7?*

Rješenje: Primjer se rješava slično kao prethodni - potrebno je odrediti i prebrojati skup Ω, te vidjeti koliko od njegovih elemenata je povoljno za događaj A (koji je u ovom slučaju događaj da je zbroj na 2 kocke jednak 7).

$$\begin{aligned}
\Omega &= \{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),(2,1),(2,2),(2,3),\ldots,(6,6)\} \\
&= \{(i,j) : 1 \leq i,j \leq 6\}, \qquad |\Omega| = 6 \cdot 6 = 36 \\
A &= \{\text{zbroj na 2 kocke je jednak 7}\} \\
&= \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}, \qquad |A| = 6 \\
\Longrightarrow P(A) &= \frac{|A|}{|\Omega|} = \frac{6}{36} = \frac{1}{6}
\end{aligned}$$

□

Laplaceov model vjerojatnosti: Neka je $\Omega = \{\omega_1, \ldots, \omega_m\}$, $m \in \mathbb{N}$. Pretpostavimo da su svi elementarni događaji `jednako vjerojatni`. Tada

je $P(\omega_i) = \frac{1}{m}$ pa je vjerojatnost događaja $A,\ A \subseteq \Omega$, jednaka:

$$\begin{aligned} P(A) &= \sum_{\omega_i \in A} P(\omega_i) = \sum_{\omega \in A} \frac{1}{m} = \frac{1}{m} \sum_{\omega \in A} 1 \\ &= \frac{|A|}{m} = \frac{|A|}{|\Omega|} = \frac{\text{broj povoljnih elementarnih događaja}}{\text{ukupan broj elementarnih događaja}} \end{aligned}$$

Primjer 16 *U kutiji se nalazi 20 mačića od kojih je 12 tigrastih i 8 crno-bijelih. Kolika je vjerojatnost da od 5 odabranih (na slučajan način izvučenih) mačića budu točno 3 tigrasta i 2 crno-bijela ako*
a) mačiće ne vraćamo
b) mačiće vraćamo?

Rješenje:
a) Pretpostavimo prvo da mačiće ne vraćamo. Događaj čiju vjerojatnost želimo izračunati je

$$A = \{\text{izvukli smo 3 tigrasta i 2 crno-bijela mačića}\}.$$

Broj načina na koji od 20 mačića možemo izabrati njih 5, a da nam pritom nije važno koliko je izvučeno tigrastih a koliko crno-cijelih, je $\binom{20}{5}$. Zapravo, imamo $|\Omega| = \binom{20}{5}$.

Broj načina na koji od ukupno 12 tigrastih mačića možemo izabrati njih 3 je $\binom{12}{3}$. Analogno, broj načina na koji od ukupno 8 crno-bijelih mačića možemo izabrati 2 je $\binom{8}{2}$. Ako istovremeno izvučemo 3 tigrasta i 2 crno-bijela mačića, dogodit će se događaj A. Imamo: $|A| = \binom{12}{3} \cdot \binom{8}{2}$

Vjerojatnost događaja A je:

$$P(A) = \frac{\binom{12}{3} \cdot \binom{8}{2}}{\binom{20}{5}} = \frac{\frac{12!}{3!\cdot 9!} \cdot \frac{8!}{2!\cdot 6!}}{\frac{20!}{5!\cdot 15!}} = \frac{\frac{12\cdot 11\cdot 10}{3\cdot 2} \cdot \frac{8\cdot 7}{2}}{\frac{20\cdot 19\cdot 18\cdot 17\cdot 16}{5\cdot 4\cdot 3\cdot 2}} = 0.39732$$

b) Pretpostavimo sada da mačiće vraćamo. Bitna razlika u odnosu na prethodni slučaj je što je sada *vjerojatnost* da izvučemo tigrastog mačića *u svakom izvlačenju ista* jer mačiće vraćamo pa svaki put izvlačimo iz istog skupa. U prethodnom slučaju, vjerojatnost da izvučemo tigrastog mačića se

smanjuje iz izvlačenja u izvlačenje budući u kutiji svaki put (kad izvučemo tigrastog mačića) ostaje sve manje tigrastih mačića.

Vjerojatnost da (u jednom izvlačenju) izvučemo jednog tigrastog mačića (označimo taj događaj s B) je:

$$P(B) = \frac{\binom{12}{1}}{\binom{20}{1}} = \frac{12}{20} = \frac{3}{5} = 0.6$$

Događaj da je izvučen crno-bijeli mačić (označimo taj događaj s C) je *suprotan* ili *komplementaran* događaju B - međusobno se isključuju a zajedno pokrivaju sve mogućnosti koje se mogu dogoditi (tj. $B \cup C$ je siguran događaj). Zbroj vjerojatnosti suprotnih događaja jednak je 1. Odatle lako izračunamo vjerojatnost od C:

$$P(C) = P(B^c) = 1 - P(B) = 1 - \frac{3}{5} = \frac{2}{5} = 0.4$$

Naravno, $P(C)$ možemo izračunati i direktno, slično kao što smo izračunali $P(B)$:

$$P(C) = \frac{\binom{8}{1}}{\binom{20}{1}} = \frac{8}{20} = \frac{2}{5} = 0.4$$

Ostalo je izračunati $P(A)$. Kako mačiće vraćamo, postoji *uređaj* pri njihovom izlačenju - zna se koji (i kakav) je bio prvi, koji drugi, koji treći itd. Tu nam se otvara mogućnost izbora: koji po redu je bio svaki od 3 izvučena tigrasta mačića? Prvi, treći i peti? Drugi, četvrti i peti? Drugi, treći i četvrti? Sve su to naime različiti elementarni događaji. Dakle, od 5 mjesta (u poretku izvlačenja) moramo izabrati 3 na kojima su bili tigrasti mačići (na preostala 2 su onda crno-bijeli mačići). To možemo učiniti na $\binom{5}{3}$ načina. Vjerojatnost da u jednom izvlačenju bude izvučen tigrasti mačić je, kao što znamo, $\frac{3}{5}$. Sljedeće izvlačenje je nezavisno od prethodnog, pa je vjerojatnost sa smo izvukli 2 tigrasta mačića jednaka $\frac{3}{5} \cdot \frac{3}{5} = \left(\frac{3}{5}\right)^2$ i analogno, vjerojatnost da smo ih izvukli 3 je $\left(\frac{3}{5}\right)^3$. U preostala 2 izvlačenja morao se dogoditi suprotan događaj, odnosno morao je biti izvučen crno-bijeli mačić, vjerojatnost čega je $\left(\frac{2}{5}\right)^2$. Uzmemo li sve do sad rečeno u obzir dobivamo:

$$P(A) = \binom{5}{3}\left(\frac{3}{5}\right)^3\left(\frac{2}{5}\right)^2 = 0.3456$$

Zadatak smo mogli riješiti i tako da razmatramo izvlačenje crno-bijelih mačića, odnosno da biramo mjesta na koja su došla 2 izvučena crno-bijela, što je moguće učiniti na $\binom{5}{2}$ načina. Sada bi komplementaran događaj bio izvlačenje tigrastog mačića i tako bi dobili:

$$P(A) = \binom{5}{2}\left(\frac{2}{5}\right)^2\left(\frac{3}{5}\right)^3$$

što zbog simetrije binomnih koeficijenata očito daje isti rezultat kao gore. □

Napomena 1 *Općenito, kad imamo N predmeta od kojih je M jedne vrste, a $N-M$ druge, vjerojatnost da - bez vraćanja - izvučemo točno k predmeta prve vrste ako izvlačimo ukupno n predmeta (dakle, $n-k$ predmeta druge vrste) je:*

$$p_X(k) = \frac{\binom{M}{k}\cdot\binom{N-M}{n-k}}{\binom{N}{n}}, \qquad \max(0, n-(N-M)) \leq k \leq \min(M,n) \quad (1)$$

Mora vrijediti $k \leq \min(M,n)$ jer niti možemo izvući više predmeta prve vrste nego što ih ukupno izvlačimo (zato $k \leq n$) niti možemo izvući više predmeta prve vrste nego što ih ukupno uopće ima (zato $k \leq M$). Uvjet $k \geq n-(N-M) \Leftrightarrow n-k \leq N-M$ osigurava pak da ne izvlačimo predmeta druge vrste ($n-k$) više nego što ih ukupno ima ($N-M$).

Napomena 2 *Ako je vjerojatnost realizacije događaja A u nekom pokusu ("uspjeh") jednaka p, tada je vjerojatnost da se događaj A dogodi točno k puta u n nezavisnih pokusa dana s*

$$p_X(k) = \binom{n}{k}p^k\cdot q^{n-k}, \qquad 0 \leq k \leq n, \quad (2)$$

gdje je $q = 1-p$ vjerojatnost suprotnog događaja, odnosno "neuspjeha".

Primjer 16 odgovara izboru slučajnog uzorka bez, odnosno s vraćanjem. S (1) je zadana **hipergeometrijska razdioba** koja opisuje biranje slučajnog uzorka bez vraćanja, a s (2) **binomna razdioba** koja opisuje biranje s vraćanjem.

2.2 Osnovne razdiobe

Slučajna varijabla je funkcija X koja elementarnim događajima pridružuje brojeve. Dakle,

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

2.2.1 Diskretne slučajne varijable

Označimo s ImX skup svih različitih vrijednosti koje slučajna varijabla X može poprimiti. Kažemo da je zadan **zakon razdiobe** ili **distribucija** slučajne varijable X ako je zadan skup Im$X = \{a_1, a_2, a_3, \ldots\}$, te niz brojeva $p_i \geq 0$ tako da

1) $p_i = P(X = a_i)$

2) $\sum_{i=1}^{\infty} p_i = 1$

Zakon razdiobe zapisujemo u obliku tablice:

$$X \sim \begin{pmatrix} a_1 & a_2 & a_3 & \ldots \\ p_1 & p_2 & p_3 & \ldots \end{pmatrix}$$

Budući je skup svih vrijednosti koje slučajna varijabla može poprimiti Im$X = \{a_1, a_2, a_3, \ldots\}$ diskretan (prebrojiv) skup, kažemo da je X **diskretna slučajna varijabla**.

Definicija 1 *Neka je $X : \Omega \to \mathbb{R}$ slučajna varijabla.* ***Funkcija gustoće vjerojatnosti*** *od X je funkcija $p_X :$ Im$X \to [0, 1]$ definirana s*

$$p_X(a_i) := P(X = a_i) = p_i$$

Definicija 2 ***Funkcija distribucije*** *slučajne varijable X je funkcija $F_X : \mathbb{R} \to [0, 1]$ definirana s*

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

Vrijedi

$$F_X(x) = \sum_{a_i \leq x} p_X(a_i).$$

Definicija 3 ***Matematičko očekivanje*** *diskretne slučajne varijable je broj* $\mathrm{E}[X]$ *definiran s*

$$\mathrm{E}[X] = \sum_{a_i \in \mathrm{Im}X} a_i \cdot p_X(a_i)$$

Vrijedi:

(i) $\mathrm{E}[g(X)] = \sum_{a_i \in \mathrm{Im}X} g(a_i) \cdot p_X(a_i), \quad g : \mathbb{R} \to \mathbb{R}$

(ii) $\mathrm{E}[X + Y] = \mathrm{E}[X] + \mathrm{E}[Y]$

(iii) $\mathrm{E}[cX] = c\,\mathrm{E}[X], \quad c \in \mathbb{R}$

(iv) $\mathrm{E}[c] = c, \quad c \in \mathbb{R}$

Iz svojstva (ii) vidimo da je očekivanje aditivno, iz svojstva (iii) da je homogeno. Ta dva svojstva zajedno daju svojstvo linearnosti.

Definicija 4 *Broj*

$$\mathrm{Var}[X] := \sum_{a_i \in \mathrm{Im}X} (a_i - \mathrm{E}\,X)^2 \cdot p_X(a_i)$$

zove se ***varijanca*** *diskretne slučajne varijable* X*.*
Standardna devijacija *slučajne varijable* X *je broj*

$$\sigma_X := +\sqrt{\mathrm{Var}[X]}$$

Vrijedi:

(i) $\mathrm{Var}[X] = \mathrm{E}[(X - \mathrm{E}\,X)^2]$
(primijenimo svojstvo (i) od očekivanja za $g(x) = (x - \mathrm{E}\,X)^2$)

(ii) $\mathrm{Var}[X] = \mathrm{E}[(X - \mathrm{E}\,X)^2] = \mathrm{E}[X^2 - 2X \cdot \mathrm{E}\,X + (\mathrm{E}\,X)^2]$
$= \mathrm{E}[X^2] - 2\,\mathrm{E}\,X \cdot \mathrm{E}\,X + (\mathrm{E}\,X)^2 = \mathrm{E}[X^2] - (\mathrm{E}\,X)^2$
$\Longrightarrow \mathrm{Var}[X] = \mathrm{E}[X^2] - (\mathrm{E}\,X)^2$
pritom: $\mathrm{E}[X^2] = \sum_{a_i \in \mathrm{Im}X} a_i^2 \cdot p_X(a_i)$

(iii) $\mathrm{Var}[aX + b] = \mathrm{E}[(aX + b - \mathrm{E}\,(aX + b))^2] = \mathrm{E}[(aX + b - a\,\mathrm{E}\,X - b)^2]$
$= \mathrm{E}[a^2(X - \mathrm{E}\,X)^2] = a^2\,\mathrm{E}[(X - \mathrm{E}\,X)^2] = a^2\,\mathrm{Var}\,X$

Nadalje, ako su X i Y *nezavisne* slučajne varijable, onda vrijedi:

$$\mathrm{E}[X \cdot Y] = \mathrm{E}[X] \cdot \mathrm{E}[Y]$$
$$\mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]$$

Općenito, ta 2 identiteta ne vrijede!

Zadatak 4 *Slučajna varijabla zadana je razdiobom*

$$X \sim \begin{pmatrix} -2 & -1 & 0 & 1 & 2 \\ 0.1 & 0.2 & 0.2 & 0.3 & 0.2 \end{pmatrix}$$

Odredite funkciju distribucije te slučajne varijable, te nacrtajte njen graf. Izračunaj vjerojatnost događaja $|X| \leq 1$*. Nadalje, izračunajte očekivanje i varijancu od* X*, te* $\mathrm{E}[3X]$*.*

Rješenje: Funkciju distribucije moramo promatrati po intervalima. Krenimo od $x \in \langle -\infty, -2\rangle$, tj. $x < -2$. U ovom slučaju:

$$F_X(x) = P(X \leq x) = 0$$

budući slučajna varijabla X ne može poprimiti vrijednost x strogo manju od -2.

Dalje, neka je $x \in [-2, -1\rangle$. Tada:

$$F_X(x) = P(X \leq x) = P(X = -2) = 0.1$$

budući je -2 jedina vrijednost unutar intervala $\langle -\infty, -1\rangle$ (drugim riječima: jedina vrijednost manja od x) koju X može poprimiti, a vjerojatnost da se to dogodi znamo jer je dan zakon razdiobe od X.

Neka je $x \in [-1, 0\rangle$. Tada:

$$F_X(x) = P(X \leq x) = P(X = -2) + P(X = -1) = 0.1 + 0.2 = 0.3$$

budući su -2 i -1 jedine vrijednosti unutar intervala $\langle -\infty, 0\rangle$ koje X može poprimiti, a vjerojatnost da se to dogodi očitavamo iz zakona razdiobe od X.

Dalje zaključujemo analogno

ako je $x \in [0,1\rangle$:

$$F_X(x) = P(X=-2) + P(X=-1) + P(X=0) = 0.1 + 0.2 + 0.2 = 0.5$$

ako je $x \in [1,2\rangle$:

$$\begin{aligned} F_X(x) &= P(X=-2) + P(X=-1) + P(X=0) + P(X=1) \\ &= 0.1 + 0.2 + 0.2 + 0.3 = 0.8 \end{aligned}$$

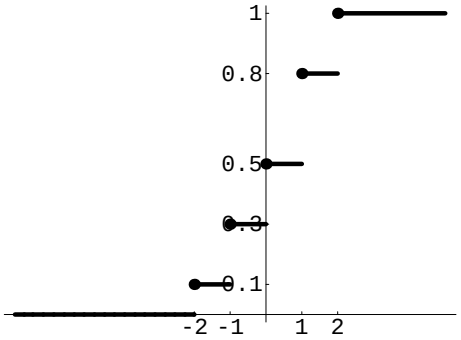
te konačno, ako je $x \in [2,+\infty\rangle$, tj. $x \geq 2$:

$$\begin{aligned} F_X(x) &= P(X=-2) + P(X=-1) + P(X=0) + P(X=1) + P(X=2) \\ &= 0.1 + 0.2 + 0.2 + 0.3 + 0.2 = 1 \end{aligned}$$

što je bilo prirodno za očekivati budući X `sigurno` poprima vrijednost manju ili jednaku 2 (tj. nikada ne poprima vrijednost veću od 2). Tako smo dobili:

$$F_X(x) = \begin{cases} 0, & x < -2 \\ 0.1, & -2 \leq x < -1 \\ 0.3, & -1 \leq x < 0 \\ 0.5, & 0 \leq x < 1 \\ 0.8, & 1 \leq x < 2 \\ 1, & x \geq 2 \end{cases}$$

Grafički prikaz funkcije distribucije



Nadalje, treba izračunati $P(|X| \leq 1)$. Vrijedi:

$$\begin{aligned} P(|X| \leq 1) &= P(-1 \leq X \leq 1) = P(X=-1) + P(X=0) + P(X=1) \\ &= 0.2 + 0.2 + 0.3 = 0.7 \end{aligned}$$

Uočimo pritom da

$$P(|X| < 1) = P(-1 < X < 1) = P(X = 0) = 0.2$$

Preostalo je izračunati $\mathrm{E}[X]$ i $\mathrm{Var}[X]$.

$$\begin{aligned} \mathrm{E}[X] &= \sum_{a_i \in ImX} a_i \cdot p_X(a_i) = \sum_{a_i \in ImX} a_i \cdot P(X = a_i) \\ &= -2 \cdot 0.1 + (-1) \cdot 0.2 + 0 \cdot 0.2 + 1 \cdot 0.3 + 2 \cdot 0.2 = 0.3 \end{aligned}$$

Varijancu računamo po formuli:

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

$\mathrm{E}[X]$ smo upravo izračunali, treba nam još $\mathrm{E}[X^2]$:

$$\begin{aligned} \mathrm{E}[X^2] &= \sum_{a_i \in ImX} a_i^2 \cdot p_X(a_i) = \sum_{a_i^2 \in ImX} a_i^2 \cdot P(X = a_i) \\ &= (-2)^2 \cdot 0.1 + (-1)^2 \cdot 0.2 + 0^2 \cdot 0.2 + 1^2 \cdot 0.3 + 2^2 \cdot 0.2 = 1.7 \end{aligned}$$

pa imamo:

$$\mathrm{Var}[X] = 1.7 - 0.3^2 = 1.61$$

Koliko je $\mathrm{E}[3X]$? To možemo izračunati na 2 načina. Jedan je - odrediti razdiobu slučajne varijable $Y = 3X$.

$$\mathrm{Im}X = \{-2, -1, 0, 1, 2\} \quad \Rightarrow \quad \mathrm{Im}Y = \mathrm{Im}\ 3X = \{-6, -3, 0, 3, 6\}$$

i pritom

$$\begin{aligned} P(Y = -6) &= P(3X = -6) = P(X = -2) = 0.1 \\ P(Y = -3) &= P(3X = -3) = P(X = -1) = 0.2 \\ P(Y = 0) &= P(3X = 0) = P(X = 0) = 0.2 \end{aligned}$$

i tako analogno dalje. Slijedi:

$$Y = 3X \sim \begin{pmatrix} -6 & -3 & 0 & 3 & 6 \\ 0.1 & 0.2 & 0.2 & 0.3 & 0.2 \end{pmatrix}$$

pa onda

$$\begin{aligned} \mathrm{E}[Y] &= \mathrm{E}[3X] = \sum_{a_i \in ImY} a_i \cdot p_Y(a_i) = \sum_{a_i \in ImY} a_i \cdot P(Y = a_i) \\ &= -6 \cdot 0.1 + (-3) \cdot 0.2 + 0 \cdot 0.2 + 3 \cdot 0.3 + 6 \cdot 0.2 = 0.9 \end{aligned}$$

Jednostavniji način rješavanja je iskoristiti svojstvo homogenosti očekivanja prema kojem je

$$\mathrm{E}[3X] = 3\,\mathrm{E}[X] = 3 \cdot 0.3 = 0.9$$

□

2.2.2 Binomna razdioba

Definicija 5 *Slučajna varijabla X ima **binomnu razdiobu** ili **distribuciju** s parametrima n i p ako je X poprima vrijednosti iz skupa $\{0, 1, 2, \ldots, n\}$ s vjerojatnostima*

$$p_X(k) = P(X = k) = \binom{n}{k} p^k \cdot q^{n-k}, \qquad 0 \leq k \leq n, \tag{3}$$

gdje je $q = 1 - p$. S (3) je zadana funkcija gustoće binomne razdiobe.

- Slučajnu varijablu X koja ima binomnu razdiobu označavamo s:

$$X \sim B(n, p)$$

- Očekivanje bionomne razdiobe: $\mathrm{E}[X] = np$
- Varijanca bionomne razdiobe: $\mathrm{Var}[X] = npq$
- Osnovna svojstva koja opisuju binomnu distribuciju:

  1. Pokus ponavljamo n puta.
  2. Postoje samo 2 mogućnosti ishoda u svakom pokusu. Jedan ishod ćemo zvati "uspjeh" a drugi "neuspjeh".
  3. Vjerojatnost "uspjeha" p jednaka je u svakom pokusu. Vjerojatnost "neuspjeha" je tada $q = 1 - p$.

4. Pokusi su međusobno nezavisni.

5. Binomna slučajna varijabla broji broj "uspjeha" k u tih n pokusa.

Zadatak 5 *Košarkaš gađa koš 5 puta i u svakom pokušaju pogađa s vjerojatnošću 3/4. Kolika je vjerojatnost da će košarkaš pogoditi koš:*
a) točno 3 puta
b) barem 3 puta
c) najviše 2 puta

Rješenje: Pokus je gađanje u koš; ponavljamo ga 5 puta. "Uspjeh" je pogodak u koš, "neuspjeh" je promašaj. Vjerojatnost "uspjeha" je zadana i jednaka je 3/4; vjerojatnost "neuspjeha" je tada 1/4 (=1-3/4). Definiramo slučajnu varijablu X koja broji pogotke. Ona ima binomnu distribuciju:

$$X \sim B\left(5,\ \frac{3}{4}\right)$$

Funkcija gustoće od X je:

$$p_X(k) = P(X = k) = \binom{5}{k}\left(\frac{3}{4}\right)^k\left(\frac{1}{4}\right)^{5-k}, \quad k = 0, 1, 2, 3, 4, 5 \qquad (4)$$

a) Želimo izračunati vjerojatnost da je košarkaš pogodio koš točno 3 puta. Zanima nas zapravo $P(X = 3)$. Uvrštavanjem $k = 3$ u (4) lako dobijemo rješenje:

$$P(X = 3) = \binom{5}{3}\left(\frac{3}{4}\right)^3\left(\frac{1}{4}\right)^2 = \binom{5}{3}\frac{3^3}{4^5} = 0.26$$

b) Kolika je vjerojatnost da košarkaš pogodi koš barem 3 puta? Taj događaj izražen pomoću slučajne varijable X je $X \geq 3$. Želimo dakle izračunati:

$$P(X \geq 3) = P(X = 3) + P(X = 4) + P(X = 5)$$

$P(X = 3)$ smo već izračunali pod a). $P(X = 4)$ i $P(X = 5)$ računamo na sličan način - uvrštavanjem $k = 4$ odnosno $k = 5$ u (4). Slijedi:

$$P(X = 4) = \binom{5}{4}\left(\frac{3}{4}\right)^4\left(\frac{1}{4}\right)^1 = 5 \cdot \frac{3^4}{4^5} = 0.395$$
$$P(X = 5) = \binom{5}{5}\left(\frac{3}{4}\right)^5\left(\frac{1}{4}\right)^0 = \left(\frac{3}{4}\right)^5 = 0.24$$
$$\Longrightarrow P(X \geq 3) = 0.26 + 0.395 + 0.24 = 0.895$$

c) Kolika je vjerojatnost da košarkaš pogodi koš najviše 2 puta? Taj događaj izražen pomoću slučajne varijable X je $X \leq 2$ a to je zapravo suprotan događaj događaju kojeg smo promatrali pod b) pa vrijedi:

$$P(X \leq 2) = 1 - P(X \geq 3) = 1 - 0.895 = 0.105$$

Naravno, vjerojatnost tog događaja mogla bi se računati i direktno:

$$P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2)$$

□

Zadatak 6 *Odredite očekivanje, varijancu i devijaciju slučajne varijable* $X \sim B\left(5,\ \frac{3}{4}\right)$. *Konstruirajte interval* $\mathrm{E}[X] \pm 2\sigma_X$ *te izračunajte* $P(\mathrm{E}[X] - 2\sigma_X < X < \mathrm{E}[X] + 2\sigma_X)$

Rješenje:

$$\begin{aligned}
&\mathrm{E}[X] = 5 \cdot \frac{3}{4} = 3.75 \\
&\mathrm{Var}[X] = 5 \cdot \frac{3}{4} \cdot \frac{1}{4} = 0.9375 \\
&\sigma_X = \sqrt{0.9375} = 0.968
\end{aligned}$$

Traženi interval je:

$$3.75 \pm 2 \cdot 0.968 = 3.75 \pm 1.936 \quad \Rightarrow \quad 1.814 < X < 5.686$$

Nadalje,

$$\begin{aligned}
&P(1.814 < X < 5.686) = P(2 \leq X \leq 5) \\
&= P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5)
\end{aligned}$$

Posljednje 3 vjerojatnosti već smo izračunali u prethodnom zadatku. Fali nam još:

$$P(X = 2) = \binom{5}{2} \left(\frac{3}{4}\right)^2 \left(\frac{1}{4}\right)^3 = 10 \cdot \frac{3^2}{4^5} = \frac{45}{512} = 0.088$$

Sada dobivamo:

$$P(\mathrm{E}[X] - 2\sigma_X < X < \mathrm{E}[X] + 2\sigma_X) = P(1.814 < X < 5.686)$$
$$= 0.088 + 0.26 + 0.395 + 0.24 = 0.983$$

Odavde možemo zaključiti da će ova slučajna varijabla u 98.3% slučajeva odstupati od svog očekivanja za najviše 2 devijacije. □

Zadatak 7 *Četiri prijatelja igraju neku igru s kartama. Prilikom podjele 52 karte jedan od igrača 3 puta zaredom nije dobio asa. Kolika je vjerojatnost da mu se to dogodi?*

Rješenje: Definirajmo slučajnu varijablu X koja broji koliko puta Igrač nije dobio asa. X ima binomnu razdiobu: $X \sim B(3, p)$. Potrebno je izračunati vrijednost parametra p što je vjerojatnost uspjeha, tj. vjerojatnost da u jednom izvlačenju igrač nije dobio asa. Ta vjerojatnost jednaka je omjeru broja svih ishoda u kojima Igrač nije dobio asa kroz broj svih mogućih ishoda dijeljenja karata. Oduzmemo li sve aseve, ostat će nam 48 karata. Stoga,

$$p = P(\text{igrač nije dobio niti jednog asa}) = \frac{\binom{48}{13}}{\binom{52}{13}} = 0.3038$$
$$\implies \quad X \sim B(3, 0.3038)$$

Vjerojatnost da Igrač 3 puta zaredom nije dobio asa jednaka je:

$$P(X = 3) = \binom{3}{3}(0.3038)^3(1 - 0.3038)^0 = 0.028.$$

□

2.2.3 Hipergeometrijska razdioba

Definicija 6 *Slučajna varijabla X ima **hipergeometrijsku razdiobu** ili **distribuciju** ako je funkcija gustoće te slučajne varijable zadana s:*

$$p_X(k) = P(X = k) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}, \quad \max(0, n - (N - M)) \leq k \leq \min(M, n) \tag{5}$$

- Očekivanje hipergeometrijske razdiobe: $\mathrm{E}[X] = \dfrac{nM}{N}$
- Varijanca hipergeometrijske razdiobe: $\mathrm{Var}[X] = \dfrac{nM(N-M)(N-n)}{N^2(N-1)}$
- Osnovna svojstva koja opisuju hipergeometrijsku distribuciju:

    1. Pokus se sastoji od slučajnog izvlačenja, bez vraćanja, n elemenata iz skupa od N elemenata, od kojih je njih M jedne vrste (izvlačenje takvog smatramo uspjehom) i $N-M$ neke druge vrste (izvlačenje takvog smatramo neuspjehom).
    2. Hipergeometrijska slučajna varijabla broji broj "uspjeha" (odnosno elemenata prve vrste) k u izvlačenju ukupno n elemenata.

Zadatak 8 *Kolika je vjerojatnost da se od 7 suglasnika i 5 samoglasnika napravi riječ koja se sastoji od 4 suglasnika i 3 samoglasnika? (riječ ne mora imati smisao)*

Rješenje: Od ukupno 7+5=12 slova želimo izabrati 4+3=7 slova. Neka je izvlačenje samoglasnika "uspjeh". Slučajna varijabla X koja broji "uspjehe" ima hipergeometrijsku razdiobu a funkcija gustoće joj je zadana s:

$$p_X(k) = P(X = k) = \frac{\binom{5}{k}\binom{7}{7-k}}{\binom{12}{7}}, \quad 0 \leq k \leq 5$$

Događaj da su izabrana 4 suglasnika i 3 samoglasnika pomoću slučajne varijable X možemo izraziti kao $X = 3$. Sada:

$$p_X(3) = P(X = 3) = \frac{\binom{5}{3}\binom{7}{4}}{\binom{12}{7}} = \frac{175}{396} = 0.442$$

□

Zadatak 9 *Iz vaze koja sadrži 4 crvene i 6 bijelih ruža izvlačimo 3 ruže. S X označimo slučajnu varijablu koja broji izvučene crvene ruže. Odredite njen zakon razdiobe, te prosječan broj izvučenih crvenih ruža.*

Rješenje: $X =$ broj crvenih ruža

Od ukupno 10=4+6 ruža izvlačimo 3, a izvlačenje crvene ruže smatramo uspjehom. Tada X ima hipergeometrijsku distribuciju a funkcija gustoće joj je:

$$p_X(k) = P(X = k) = \frac{\binom{4}{k}\binom{6}{3-k}}{\binom{10}{3}}, \quad 0 \leq k \leq 3$$

Dakle, Im$X = \{0, 1, 2, 3\}$. Trebaju nam pripadne vjerojatnosti. Imamo

$$p_X(0) = P(X = 0) = \frac{\binom{4}{0}\binom{6}{3}}{\binom{10}{3}} = \frac{6 \cdot 5 \cdot 4}{10 \cdot 9 \cdot 8} = \frac{1}{6}$$

$$p_X(1) = P(X = 1) = \frac{\binom{4}{1}\binom{6}{2}}{\binom{10}{3}} = \frac{4 \cdot \frac{6 \cdot 5}{2}}{\frac{10 \cdot 9 \cdot 8}{3 \cdot 2}} = \frac{1}{2}$$

$$p_X(2) = P(X = 2) = \frac{\binom{4}{2}\binom{6}{1}}{\binom{10}{3}} = \frac{\frac{4 \cdot 3}{2} \cdot 6}{\frac{10 \cdot 9 \cdot 8}{3 \cdot 2}} = \frac{3}{10}$$

$$p_X(3) = P(X = 3) = \frac{\binom{4}{3}\binom{6}{0}}{\binom{10}{3}} = \frac{4 \cdot 1}{\frac{10 \cdot 9 \cdot 8}{3 \cdot 2}} = \frac{1}{30}$$

$$\Rightarrow \; X \sim \begin{pmatrix} 0 & 1 & 2 & 3 \\ 1/6 & 1/2 & 3/10 & 1/30 \end{pmatrix}$$

Provjera da smo dobro računali:

$$\frac{1}{6} + \frac{1}{2} + \frac{3}{10} + \frac{1}{30} = 1$$

Izračunajmo sada $\mathrm{E}[X]$ (što je zapravo prosječan broj):

$$\mathrm{E}[X] = \sum_{k=0}^{3} a_k \cdot p_X(k) = \sum_{k=0}^{3} k \cdot p_X(k) = 0 \cdot \frac{1}{6} + 1 \cdot \frac{1}{2} + 2 \cdot \frac{3}{10} + 3 \cdot \frac{1}{30} = \frac{6}{5} = 1.2$$

Umjesto pomoću definicije, očekivanje smo mogli izračunati i pomoću gore navedene formule:

$$\mathrm{E}[X] = \frac{nM}{N} = \frac{3 \cdot 4}{4 + 6} = \frac{12}{10} = 1.2$$

□

2.2.4 Poissonova razdioba

Definicija 7 *Slučajna varijabla X ima **Poissonovu razdiobu** ili **distribuciju** s parametrom $\lambda > 0$ ako je funkcija gustoće te slučajne varijable zadana s:*

$$p_X(k) = P(X = k) = \frac{\lambda^k}{k!}\, e^{-\lambda}, \quad k = 0, 1, 2, 3, \ldots \tag{6}$$

Poissonova distribucija daje model vjerojatnosti "rijetkih" događaja (ponekad se naziva i "zakon rijetkih događaja") koji se događaju u jedinici vremena, površine, volumena i slično. Broj prometnih nesreća na određenoj dionici autoceste u jednom danu, telefonski pozivi na centrali u jednoj minuti, defekti po jedinici duljine bakrene žice, broj mjesečnih nesreća u tvornici, broj oboljelih stabala po aru šume te broj vidljivih grešaka na dijamantu su npr. varijable čije se relativne frekvencije mogu dobro aproksimirati Poissonovom distribucijom.

- Slučajnu varijablu X koja ima Poissonovu razdiobu označavamo s:

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

- Očekivanje Poissonove razdiobe: $\mathrm{E}[X] = \lambda$
- Varijanca Poissonove razdiobe: $\mathrm{Var}[X] = \lambda$
- Osnovna svojstva koja opisuju Poissonovu distribuciju:
    1. Pokus se sastoji od prebrojavanja koliko puta (k) se neki događaj dogodi u jedinici vremena, jedinici površine, volumena, težine, daljine ili bilo kojoj drugoj mjerenoj jedinici.
    2. Vjerojatnost da će se događaj kojeg promatramo dogoditi jednaka je za svaku mjernu jedinicu (za svaku sekundu, svaki metar, svaki karat i sl.).
    3. Broj događaja koji se dogode u pojedinoj jedinici vremena, površine ili volumena nezavisan je od broja događaja koji se dogodi u bilo kojoj drugoj jedinici.

4. Prosječni ili očekivani broj događaja u jednoj jedinici jednak je parametru λ, odnosno $\mathrm{E}[X] = \lambda$.

Zadatak 10 *Dokažite da je* $\mathrm{E}[X] = \lambda$.

Rješenje: Koristeći definiciju očekivanja, funkciju gustoće Poissonove razdiobe zadanu s (6) te svojstvo

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

dobivamo:

$$\begin{aligned} \mathrm{E}[X] &= \sum_{k=0}^{\infty} k \cdot p_X(k) = \sum_{k=1}^{\infty} k \cdot \frac{\lambda^k}{k!}\, e^{-\lambda} \\ &= \lambda \cdot \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!}\, e^{-\lambda} = \lambda \cdot \sum_{m=0}^{\infty} \frac{\lambda^m}{m!}\, e^{-\lambda} = \lambda \cdot 1 = \lambda \end{aligned}$$

□

Zadatak 11 *Slučajna varijabla X ima Poissonovu razdiobu. Ako vrijedi*

$$P(X = 1) = P(X = 2),$$

izračunajte očekivanje $\mathrm{E}[X]$ *i* $P(X \geq 4)$.

Rješenje:

$X \sim P(\lambda)$

$p_X(k) = P(X = k) = \dfrac{\lambda^k}{k!}\, e^{-\lambda}, \quad k = 0, 1, 2, 3, \ldots$

$P(X = 1) = P(X = 2) \;\Rightarrow\; \dfrac{\lambda^1}{1!}\, e^{-\lambda} = \dfrac{\lambda^2}{2!}\, e^{-\lambda} \;\Rightarrow\; \lambda = \dfrac{\lambda^2}{2} \;\Rightarrow\; \lambda^2 - 2\lambda = 0$

$\Rightarrow \lambda(\lambda - 2) = 0 \;\Rightarrow \lambda_1 = 0,\ \lambda_2 = 2$

Budući mora biti $\lambda > 0$, jedino rješenje je $\lambda = 2$.

$$\Rightarrow X \sim P(2), \quad P(X = k) = \frac{2^k}{k!}\, e^{-2}$$

$$
\begin{aligned}
&\mathrm{E}\,[X] = \lambda = 2 \\
&P(X \geq 4) = 1 - P(X < 4) = 1 - P(X \leq 3) \\
&\quad = 1 - P(X = 0) - P(X = 1) - P(X = 2) - P(X = 3) \\
&\quad = 1 - \frac{2^0}{0!}\, e^{-2} - \frac{2^1}{1!}\, e^{-2} - \frac{2^2}{2!}\, e^{-2} - \frac{2^3}{3!}\, e^{-2} \\
&\quad = 1 - e^{-2}\left(1 + 2 + 2 + \frac{4}{3}\right) = 1 - e^{-2} \cdot \frac{19}{3} = 0.143
\end{aligned}
$$

□

Zadatak 12 *Pretpostavimo da je 220 grešaka raspoređeno slučajno unutar knjige od 200 stranica. Odredite vjerojatnost da dana stranica knjige sadrži:*
a) niti jednu grešku
b) točno jednu grešku
c) barem dvije greške

Rješenje: Definirajmo slučajnu varijablu X koja broji greške na pojedinoj stranici. Ona ima Poissonovu distribuciju. Kako bi odredili njenu funkciju gustoće, potreban nam je parametar λ. Znamo da je taj parametar jednak očekivanom ili prosječnom broju događaja (= broj grešaka) koji se dogode u jednoj jedinici (= na jednoj stranici). Stoga

$$
\begin{aligned}
&\lambda = \frac{220}{200} = 1.1 \\
&p_X(k) = P(X = k) = \frac{\lambda^k}{k!}\, e^{-\lambda} = \frac{(1.1)^k}{k!}\, e^{-1.1}, \quad k = 0, 1, 2, \ldots \qquad (7)
\end{aligned}
$$

Pomoću ovako definirane slučajne varijable, događaj pod a) možemo zapisati kao $X = 0$, događaj pod b) kao $X = 1$, a događaj pod c) kao $X \geq 2$. Vjerojatnosti tih događaja računamo uvrštavanjem odgovarajućih k u (7). Dobivamo:

$$
\begin{aligned}
&a) \quad P(X = 0) = \frac{(1.1)^0}{0!}\, e^{-1.1} = e^{-1.1} = 0.333 \\
&b) \quad P(X = 1) = \frac{(1.1)^1}{1!}\, e^{-1.1} = 0.366 \\
&c) \quad P(X \geq 2) = 1 - P(X = 0) - P(X = 1) = 1 - 0.333 - 0.366 = 0.301
\end{aligned}
$$

□

Prethodni zadatak lijepo ilustrira zašto se Poissonova distribucija naziva i "zakon rijetkih događaja". Događaji da na stranici nema niti jedne greške ($X = 0$) i da je na stranici točno jedna greška ($X = 1$) - dakle "rijetki" događaji (u smislu malog broja grešaka) - imaju veću vjerojatnost nego događaj da su stranici 2 ili 3 ili 4 ili 5 ili ... ili n ili ... grešaka ($X \geq 2$).

2.2.5 Aproksimacija binomne razdiobe Poissonovom

• binomna razdioba $\mathbf{B(n, p)}$ može se **aproksimirati** Poissonovom razdiobom $\mathbf{P(np)}$. Aproksimacija je to bolja što je parametar n veći, a parametar p manji.

Zadatak 13 *Kolika je vjerojatnost da među 200 ljudi bude barem 4 ljevaka, ako ljevaka ima prosječno* 1%*?*

Rješenje: Definirajmo slučajnu varijabu X koja broji ljevake. Ona ima binomnu razdiobu s parametrima $n = 200$ (promatramo 200 ljudi, tj. 200 puta ponavljamo pokus) i $p = 1/100$ (što je vjerojatnost "uspjeha", odnosno vjerojatnost da je izabrani čovjek ljevak). Njena funkcija gustoće zadana je s:

$$P(X = k) = \binom{200}{k} \left(\frac{1}{100}\right)^k \left(\frac{99}{100}\right)^{200-k}, \quad 0 \leq k \leq 200.$$

Zanima nas kolika je $P(X \geq 4)$:

$$\begin{aligned} P(X \geq 4) &= 1 - P(X = 0) - P(X = 1) - P(X = 2) - P(X = 3) \\ &= 1 - \binom{200}{0} \left(\frac{1}{100}\right)^0 \left(\frac{99}{100}\right)^{200} - \binom{200}{1} \left(\frac{1}{100}\right)^1 \left(\frac{99}{100}\right)^{199} \\ &\quad - \binom{200}{2} \left(\frac{1}{100}\right)^2 \left(\frac{99}{100}\right)^{198} - \binom{200}{3} \left(\frac{1}{100}\right)^3 \left(\frac{99}{100}\right)^{197} = \dots \end{aligned}$$

Dobiveni izrazi nisu baš "praktični za računanje". Tu će nam pomoći aproksimacija Poissonovom razdiobom:

$$B(n, p) \sim P(np)$$

$$\lambda = n \cdot p = 200 \cdot \frac{1}{100} = 2$$
$$\Rightarrow\ P(X = k) = \frac{2^k}{k!} \cdot e^{-2}, \quad k = 0, 1, 2, \ldots$$

Sada dobivamo:

$$\begin{aligned} P(X \geq 4) &= 1 - P(X = 0) - P(X = 1) - P(X = 2) - P(X = 3) \\ &= 1 - \frac{2^0}{0!} \cdot e^{-2} - \frac{2^1}{1!} \cdot e^{-2} - \frac{2^2}{2!} \cdot e^{-2} - \frac{2^3}{3!} \cdot e^{-2} \\ &= 1 - \left(1 + 2 + 2 + \frac{4}{3}\right) e^{-2} = 0.143 \end{aligned}$$

□

Zadatak 14 *Stroj proizvodi* 99.8% *ispravnih i* 0.2% *neispravnih proizvoda. Kolika je vjerojatnost da u uzorku od 500 proizvoda više od 3 budu neispravna?*

Rješenje: Definirajmo slučajnu varijabu X koja broji neispravne proizvode. X ima binomnu razdiobu: $X \sim B(500, 0.002)$. Nas zanima

$$P(X > 3) = 1 - P(X = 0) - P(X = 1) - P(X = 2) - P(X = 3)$$

Direktno korištenje binomne razdiobe ponovo bi dovelo do nezgrapnih izraza. Iskoristimo stoga aproksimaciju Poissonovom razdiobom:

$$\lambda = n \cdot p = 500 \cdot 0.002 = 1$$
$$\Rightarrow\ P(X = k) = \frac{1^k}{k!} \cdot e^{-1} = \frac{1}{k! \cdot e}, \quad k = 0, 1, 2, \ldots$$

Slijedi:

$$P(X > 3) = 1 - \frac{1}{0! \cdot e} - \frac{1}{1! \cdot e} - \frac{1}{2! \cdot e} - \frac{1}{3! \cdot e} = 1 - \frac{8}{3e} = 0.019$$

□

2.3 Uvjetna vjerojatnost. Nezavisni događaji.

Pretpostavimo da `znamo` da se dogodio događaj B. Utječe li to na vjerojatnost događaja A?

`Vjerojatnost događaja` A `uz uvjet da se dogodio događaj B` zovemo **uvjetna vjerojatnost**, označavamo s $\mathbf{P(A|B)}$ i definiramo s:

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

Događaj B pritom na neki način postaje novi skup svih elementarnih događaja, tj. novi Ω.

Za događaje A i B kažemo da su **nezavisni** ako vrijedi:

$$P(A \cap B) = P(A) \cdot P(B),$$

gdje $A \cap B$ predstavlja događaj kada se `istovremeno dogode` A `i` B.

Pretpostavimo da su događaji A i B nezavisni. Tada

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

Dakle, ako su događaji nezavisni, onda uvjet da se dogodio jedan od njih ne utječe na vjerojatnost događanja onog drugog. Vrijedi i obrat - ukoliko vrijedi gornji identitet, tada su događaji nezavisni. Naime,

$$P(A|B) = P(A) \Leftrightarrow \frac{P(A \cap B)}{P(B)} = P(A) \Leftrightarrow P(A \cap B) = P(A) \cdot P(B)$$

$\Leftrightarrow$ događaji A i B su nezavisni

Primjer 17 *Promatramo obitelj s 3 djece. Pretpostavimo da je svih* $2^3 = 8$ *mogućnosti kombinacija djece (po spolu i po starosti) jednako vjerojatno. Skup svih elementarnih događaja* Ω *je:*

$$\Omega = \{MMM,\ MMŽ,\ MŽM,\ ŽMM,\ ŽŽM,\ ŽMŽ,\ MŽŽ,\ ŽŽŽ\}.$$

Promatramo događaje:

$$A = \{u\ obitelji\ su\ djeca\ oba\ spola\},$$

$$B = \{u \text{ obitelji nema više od 1 djevojčice}\}.$$

Jesu li ti događaji nezavisni? Da bismo odgovorili na to pitanje, potrebno je provjeriti vrijedi li definicija.

Izračunajmo najprije $P(A)$ i $P(B)$. Što su "povoljni" elementarni događaji za A? Svi oni, koji pripadaju Ω, i koji opisuju obitelji s bar jednom djevojčicom odnosno bar jednim dječakom. Dakle,

$$A = \Omega \backslash \{\ MMM,\ \check{Z}\check{Z}\check{Z}\ \}$$

pa je

$$P(A) = 1 - \frac{2}{8} = \frac{3}{4}.$$

Slično vidimo da je

$$B = \{\ MMM,\ MM\check{Z},\ M\check{Z}M,\ \check{Z}MM\ \}$$

pa je

$$P(B) = \frac{4}{8} = \frac{1 + \binom{3}{1}}{8} = \frac{1}{2}.$$

Događaj $A \cap B$ opisuje **istovremeno** *događanje događaja A i B, što znači da obitelj mora imati djecu oba spola i pritom najviše jednu djevojčicu - što znači zapravo točno jednu djevojčicu! - pa stoga*

$$A \cap B = \{MM\check{Z},\ M\check{Z}M,\ \check{Z}MM\} = B \backslash \{MMM\}$$

a odatle slijedi

$$P(A \cap B) = \frac{3}{8}.$$

Kako je

$$P(A) \cdot P(B) = \frac{3}{4} \cdot \frac{1}{2} = \frac{3}{8} = P(A \cap B)$$

time smo pokazali da su događaji A i B - u ovom slučaju - nezavisni.

No, vrijedi li to općenito, odnosno za obitelji s proizvoljnim brojem djece? Pokazuje se da za obitelji s 2 ili 4 djece ova 2 događaja nisu nezavisna!

Dokazat ćemo to za slučaj obitelji s 4 djece; samostalno to pokušajte učiniti za slučaj obitelji s 2 djece. Imamo:

$$P(A) = 1 - \frac{2}{2^4} = \frac{7}{8}, \qquad P(B) = \frac{1 + \binom{4}{1}}{2^4} = \frac{5}{16}$$
$$P(A \cap B) = \frac{\binom{4}{1}}{2^4} = \frac{1}{4}.$$

Konačno, kako je

$$P(A) \cdot P(B) = \frac{7}{8} \cdot \frac{5}{16} = \frac{35}{128} \neq \frac{1}{4} = P(A \cap B),$$

zaključujemo da događaji A i B nisu nezavisni!

Zadatak 15 *Bacamo 2 kocke. Koja je vjerojatnost da je na prvoj kocki pao broj 5, ako je zbroj na dvije kocke jednak 7? Izračunajte $P(\max(X, Y) \leq 2)$.*

Rješenje:

$\Omega = \{(i, j) : 1 \leq i, j \leq 6\}$
definiramo slučajne varijable $X, Y : \Omega \to \mathbb{R}$
tako da X pamti broj na prvoj kocki, a Y na drugoj
$\mathrm{ImX} = \mathrm{ImY} = \{1, 2, 3, 4, 5, 6\}$

Treba izračunati:

$$P(X = 5 \mid X + Y = 7) = \frac{P(X = 5,\ X + Y = 7)}{P(X + Y = 7)}$$

Imamo

$$\{X + Y = 7\} = \{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}$$
$$\{X = 5\} \cap \{X + Y = 7\} = \{(5, 2)\}$$

pa stoga

$$P(X = 5 \mid X + Y = 7) = \frac{\frac{1}{6^2}}{\frac{6}{6^2}} = \frac{1}{6}.$$

Nadalje, treba izračunati $P(\max(X,Y) \leq 2)$. Događaj $\{\max(X,Y) \leq 2\}$, realizirat će se ako na obje kocke ne padne broj veći od 2 - samo tako maksimum može biti ne veći od 2. Bacanje prve kocke nezavisno je od bacanja druge, odnosno realizacija na prvoj kocki ne utječe na realizaciju na drugoj, stoga možemo koristiti svojstva nezavisnih događaja iz njihove definicije. Slijedi:

$$\begin{aligned} P(\max(X,Y) \leq 2) &= P(X \leq 2,\ Y \leq 2) \\ &= P(X \leq 2) \cdot P(Y \leq 2) = \frac{2}{6} \cdot \frac{2}{6} = \frac{1}{9} \end{aligned}$$

□

2.4 Bayesova formula

Događaji $H_1, H_2, \ldots H_n$ čine **potpun sistem događaja** ako je:

$$\begin{aligned} &1)\ P(H_i) > 0 \ \text{ za } \ i = 1, 2, \ldots, n \\ &2)\ H_i \cap H_j = \emptyset \ \text{ za } \ i \neq j, \ \ i, j = 1, 2, \ldots, n \\ &3)\ \cup_{i=1}^{n} H_i = \Omega \end{aligned}$$

Elemente potpunog sistema događaja $H_1, H_2, \ldots, H_n$ nazivamo **hipoteze**. Važno! Hipoteze se uzajamno isključuju (svojstvo 2) i točno jedna od njih se mora dogoditi (svojstvo 3), u svakom izvođenju pokusa.

Formula potpune vjerojatnosti: Neka je $\{H_1, H_2, \ldots, H_n\}$ potpun sistem događaja i neka je A proizvoljan događaj. Tada vrijedi:

$$P(A) = \sum_{i=1}^{n} P(H_i) \cdot P(A|H_i)$$

Neka je zadan potpun sistem događaj $\{H_1, H_2, \ldots, H_n\}$. Pretpostavimo da je pokus izveden i da se kao njegov ishod **pojavio događaj A**. Vjerojatnosti $P(H_i)$ bile su poznate prije izvođenja pokusa. Koliku vjerojatnost imaju hipoteze H_i $(i = 1, \ldots, n)$ `nakon izvođenja pokusa`?

Bayesova formula: Neka je $\{H_1, H_2, \ldots, H_n\}$ potpun sistem događaja i neka je $A \subseteq \Omega$ događaj takav da je $P(A) > 0$. Tada za svaki $i = 1, 2, \ldots, n$

vrijedi

$$P(H_i|A) = \frac{P(H_i) \cdot P(A|H_i)}{\sum_{j=1}^{n} P(H_j) \cdot P(A|H_j)}$$

Dokaz. Primjenom definicije uvjetne vjerojatnosti slijedi

$$P(H_i|A) = \frac{P(H_i \cap A)}{P(A)}$$

i s druge strane

$$P(A|H_i) = \frac{P(A \cap H_i)}{P(H_i)} \quad \Rightarrow \quad P(A \cap H_i) = P(H_i) \cdot P(A|H_i).$$

Primjenimo ovo pa iz gornje jednakosti dobivamo

$$P(H_i|A) = \frac{P(H_i \cap A)}{P(A)} = \frac{P(H_i) \cdot P(A|H_i)}{\sum_{j=1}^{n} P(H_j) \cdot P(A|H_j)}$$

□

Spoznaja da se dogodio događaj A **mijenja** naše uvjerenje o mogućnosti pojavljivanja hipoteza $H_1, H_2, \ldots, H_n$. Vrijedi:

$$\begin{aligned}\sum_{i=1}^{n} P(H_i|A) &= \sum_{i=1}^{n} \frac{P(H_i) \cdot P(A|H_i)}{\sum_{j=1}^{n} P(H_j) \cdot P(A|H_j)} \\ &= \frac{1}{\sum_{j=1}^{n} P(H_j)P(A|H_j)} \cdot \sum_{i=1}^{n} P(H_i)P(A|H_i) = \frac{P(A)}{P(A)} = 1\end{aligned}$$

Primjer 18 *Pri obradi jednoga pacijenta sumnja se na 2 bolesti, H_1 i H_2. U danim uvjetima njihove su vjerojatnosti dane s $P(H_1) = 0.6$ i $P(H_2) = 0.4$. Radi preciziranja dijagnoze obavlja se određena pretraga na pacijentu, čiji su rezultati pozitivna ili negativna reakcija. U slučaju bolesti H_1 vjerojatnost pozitivne reakcije je 0.9, a negativne 0.1, a u slučaju bolesti H_2 i pozitivna i negativna reakcija imaju vjerojatnost 0.5. Pretraga je obavljena 2 puta i oba puta reakcija je bila negativna. Kolike su vjerojatnosti svake od bolesti poslije ovih pretraga? Koja hipoteza je* vjerodostojnija*?*

Rješenje: Skup $\{H_1, H_2\}$ je potpun sistem događaja - događaji H_1 i H_2 međusobno se isključuju a jedan se mora dogoditi. Definirajmo događaj A:

A = { pretraga je napravljena 2 puta i oba puta reakcija je bila negativna}

Želimo izračunati $P(H_1|A)$ = vjerojatnost da pacijent ima bolest H_1 ako znamo da se dogodio A, te $P(H_2|A)$ = vjerojatnost da pacijent ima bolest H_2 ako znamo da se dogodio A. To ćemo učiniti koristeći Bayesovu formulu. Treba nam $P(A|H_1)$ = vjerojatnost da se dogodio A ako pacijent ima bolest H_1 i $P(A|H_2)$ = vjerojatnost da se dogodio A ako pacijent ima bolest H_2. Razumno je pretpostaviti da su 2 napravljenje pretrage nezavisne jedna od druge pa imamo

$$P(A|H_1) = 0.1 \cdot 0.1 = 0.01$$
$$P(A|H_2) = 0.5 \cdot 0.5 = 0.25$$

Primjenom Bayesove formule dobivamo:

$$P(H_1|A) = \frac{P(H_1) \cdot P(A|H_1)}{\sum_{j=1}^{2} P(H_j) \cdot P(A|H_j)} = \frac{0.6 \cdot 0.01}{0.6 \cdot 0.01 + 0.4 \cdot 0.25} \approx 0.06$$
$$P(H_2|A) = \frac{P(H_2) \cdot P(A|H_2)}{\sum_{j=1}^{2} P(H_j) \cdot P(A|H_j)} = \frac{0.4 \cdot 0.25}{0.6 \cdot 0.01 + 0.4 \cdot 0.25} \approx 0.94$$

ili, jednostavnije,

$$P(H_2|A) = 1 - P(H_1|A) \approx 0.94$$

Zaključujemo da dobiveni rezultati pretraga daju "jak" razlog da se pretpostavi bolest H_2! Hipoteza H_2 je vjerodostojnija.

2.5 Neprekidne slučajne varijable

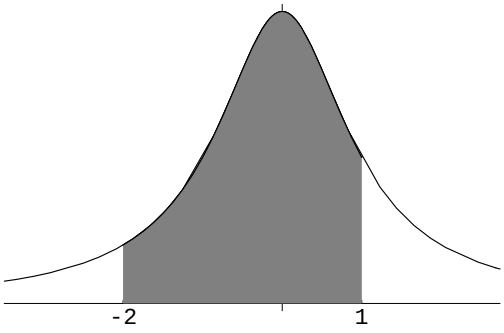
Za slučajnu varijablu X kažemo da je **neprekidna** ako vrijedi sljedeće:

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

(ii) postoji nenegativna funkcija $f_X : \mathbb{R} \to \mathbb{R}$ tako da za svaka dva broja a, b $(a < b)$ vrijedi

$$P(a \leq X \leq b) = \int_a^b f_X(t)dt$$

Funkciju f_X zovemo **funkcija gustoće** od X. Vjerojatnost da vrijednost slučajne varijable X upadne u interval $[a, b]$ jednaka je dakle površini ispod grafa funkcije gustoće na tom intervalu. Ako je na slici prikazana funkcija gustoće od X, tada je $P(-2 \leq X \leq 1)$ jednaka sljedećoj površini:



Funkcija distribucije F_X od X definirana je s:

$$F_X(x) := P(X \leq x) = \int_{-\infty}^{x} f_X(t)dt \tag{8}$$

Vrijedi:

$$P(a \leq X \leq b) = F_X(b) - F_X(a) \tag{9}$$

Navedimo još dva svojstva neprekidne slučajne varijable:

(1) Za svaki broj $a \in \mathbb{R}$ je

$$P(X = a) = \lim_{b \to a} P(a \leq X \leq b) = \lim_{b \to a} \int_a^b f_X(t)dt = \int_a^a f_X(t)dt = 0$$

(2)

$$\int_{-\infty}^{\infty} f_X(t)dt = P(-\infty < X < \infty) = 1$$

što znači da je ukupna površina ispod grafa funkcije gustoće jednaka 1.

Matematičko očekivanje od X definirano je s:

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} t \cdot f_X(t)dt, \tag{10}$$

a za **varijancu** vrijedi relacija kao i kod diskretnih slučajnih varijabli

$$\mathrm{Var}\,[X] = \mathrm{E}\,[X^2] - (\mathrm{E}\,[X])^2 \tag{11}$$

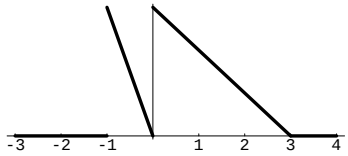
gdje je sada

$$\mathrm{E}\,[X^2] = \int_{-\infty}^{\infty} t^2 \cdot f_X(t)dt. \tag{12}$$

Općenito, za $g : \mathbb{R} \to \mathbb{R}$ vrijedi

$$\mathrm{E}\,[g(X)] = \int_{-\infty}^{\infty} g(t) \cdot f_X(t)dt.$$

Zadatak 16 *Funkcija gustoće neke slučajne varijable X dana je grafom. Odredite analitički prikaz od $f_X(x)$, $F_X(x)$, te izračunajte* $\mathrm{Var}[X]$ *i* $P(|X| \leq 1)$.



Rješenje: Da bi neka funkcija bila funkcija gustoće, mora zadovoljavati:

$$1^\circ) \quad f_X(t) \geq 0, \;\; t \in \mathbb{R}$$
$$2^\circ) \int_{-\infty}^{+\infty} f_X(t)dt = 1$$

Prvo svojstvo dana funkcija očito zadovoljava, a iz drugog svojstva slijedi da površina dva trokuta sa slike - što je površina ispod grafa zadane funkcije - mora biti jednaka 1. Označimo li nepoznatu visinu na y-osi s v, slijedi:

$$\frac{1 \cdot v}{2} + \frac{3 \cdot v}{2} = 1 \quad \Leftrightarrow \quad v = \frac{1}{2}$$

Točke $(-1, \frac{1}{2})$ i $(0, 0)$ jednoznačno određuju pravac $y = -\frac{x}{2}$, a točke $(0, \frac{1}{2})$ i $(3, 0)$ pravac $y = \frac{1}{2} - \frac{x}{6}$, pa smo tako dobili analitički prikaz funkcije gustoće:

$$f_X(x) = \begin{cases} 0, & x < -1 \\ -\frac{x}{2}, & -1 \leq x < 0 \\ \frac{1}{2} - \frac{x}{6}, & 0 \leq x \leq 3 \\ 0, & x \geq 3 \end{cases}$$

Sljedeći korak je odrediti funkciju distribucije $F_X(x)$. Prisjetimo se njene definicije (8). Imamo:

$$
\begin{aligned}
x \le -1: \qquad & F_X(x) = \int_{-\infty}^{x} f_X(t)dt = \int_{-\infty}^{x} 0dt = 0 \\
-1 \le x \le 0: \qquad & F_X(x) = \int_{-\infty}^{x} f_X(t)dt = \int_{-\infty}^{-1} 0dt + \int_{-1}^{x} \left(-\frac{t}{2}\right) dt = -\frac{1}{2}\cdot\frac{t^2}{2}\Big|_{-1}^{x} \\
& = -\frac{1}{4}(x^2-1) = \frac{1}{4}(1-x^2) \\
0 \le x \le 3: \qquad & F_X(x) = \int_{-\infty}^{x} f_X(t)dt = \int_{-\infty}^{-1} 0dt + \int_{-1}^{0} \left(-\frac{t}{2}\right) dt + \int_{0}^{x} \left(\frac{1}{2} - \frac{t}{6}\right) dt \\
& = -\frac{1}{2}\cdot\frac{t^2}{2}\Big|_{-1}^{0} + \frac{1}{2}\cdot t\Big|_{0}^{x} - \frac{1}{6}\cdot\frac{t^2}{2}\Big|_{0}^{x} = \frac{1}{4} + \frac{1}{2}x - \frac{1}{12}x^2 \\
x \ge 3: \qquad & F_X(x) = \int_{-\infty}^{x} f_X(t)dt = \int_{-\infty}^{-1} 0dt + \int_{-1}^{0} \left(-\frac{t}{2}\right) dt + \int_{0}^{3} \left(\frac{1}{2} - \frac{t}{6}\right) dt \\
& + \int_{3}^{x} 0dt = -\frac{1}{2}\cdot\frac{t^2}{2}\Big|_{-1}^{0} + \frac{1}{2}\cdot t\Big|_{0}^{3} - \frac{1}{6}\cdot\frac{t^2}{2}\Big|_{0}^{3} = \frac{1}{4} + \frac{3}{2} - \frac{9}{12} = 1
\end{aligned}
$$

pa slijedi:

$$
F_X(x) = \begin{cases} 0, & x \le -1 \\ \frac{1}{4}(1-x^2), & -1 \le x \le 0 \\ \frac{1}{12}(3+6x-x^2), & 0 \le x \le 3 \\ 1, & x \ge 3 \end{cases}
$$

Koliko je varijanca zadane slučajne varijable X? Izračunat ćemo je koristeći (11). Najprije izračunajmo očekivanje $\mathrm{E}[X]$ pomoću (10).

$$
\begin{aligned}
\mathrm{E}[X] &= \int_{-\infty}^{+\infty} t\cdot f_X(t)dt = \int_{-\infty}^{-1} t\cdot 0dt + \int_{-1}^{0} t\cdot\left(-\frac{t}{2}\right)dt + \int_{0}^{3} t\cdot\left(\frac{1}{2}-\frac{t}{6}\right)dt + \int_{3}^{x} t\cdot 0dt \\
&= -\frac{1}{2}\cdot\frac{t^3}{3}\Big|_{-1}^{0} + \frac{1}{2}\cdot\frac{t^2}{2}\Big|_{0}^{3} - \frac{1}{6}\cdot\frac{t^3}{3}\Big|_{0}^{3} = -\frac{1}{6} + \frac{9}{4} - \frac{27}{18} = \frac{7}{12}
\end{aligned}
$$

Nadalje, $\mathrm{E}[X^2]$ računamo pomoću (12):

$$
\begin{aligned}
\mathrm{E}[X^2] &= \int_{-\infty}^{+\infty} t^2\cdot f_X(t)dt = \int_{-\infty}^{-1} t^2\cdot 0dt + \int_{-1}^{0} t^2\cdot\left(-\frac{t}{2}\right)dt \\
&+ \int_{0}^{3} t^2\cdot\left(\frac{1}{2}-\frac{t}{6}\right)dt + \int_{3}^{x} t^2\cdot 0dt \\
&= -\frac{1}{2}\cdot\frac{t^4}{4}\Big|_{-1}^{0} + \frac{1}{2}\cdot\frac{t^3}{3}\Big|_{0}^{3} - \frac{1}{6}\cdot\frac{t^4}{4}\Big|_{0}^{3} = \frac{1}{8} + \frac{27}{6} - \frac{81}{24} = \frac{5}{4}
\end{aligned}
$$

Sada, prema (11), imamo:

$$\mathrm{Var}[X] = \frac{5}{4} - \left(\frac{7}{12}\right)^2 = \frac{131}{144}$$

Preostalo je još izračunati $P(|X| \leq 1) = P(-1 \leq X \leq 1)$. Primjenom (9) dobivamo:

$$\begin{aligned}
&P(|X| \leq 1) = P(-1 \leq X \leq 1) = F_X(1) - F_X(-1) \\
&F_X(1) = \frac{1}{12}(3 + 6 \cdot 1 - 1^2) = \frac{2}{3} \\
&F_X(-1) = 0 \quad \left(\text{ili} \quad F_X(-1) = \frac{1}{4}(1-1) = 0\right) \\
&\Rightarrow \quad P(|X| \leq 1) = F_X(1) - F_X(-1) = \frac{2}{3} - 0 = \frac{2}{3}
\end{aligned}$$

□

2.5.1 Normalna razdioba

Definicija 8 *Kažemo da neprekidna slučajna varijabla X ima* ***normalnu razdiobu*** *s parametrima μ i σ^2 ako joj je funkcija gustoće zadana s:*

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{13}$$

Oznaka:

$$X \sim N(\mu, \sigma^2)$$

Vrijedi:

1. $f_X(x) > 0, \ \forall x \in \mathbb{R} \ \Rightarrow \ \mathrm{Im}X = \mathbb{R}$
2. $\mathrm{E}[X] = \displaystyle\int_{-\infty}^{\infty} t \cdot \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \mu$
3. $\mathrm{Var}[X] = \mathrm{E}[(X-\mu)^2] = \displaystyle\int_{-\infty}^{\infty} (t-\mu)^2 \cdot \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \sigma^2$

Iz 2. i 3. vidimo da parametri μ i σ^2 zapravo predstavljaju očekivanje, odnosno varijancu od X.

Normalna razdioba je invarijantna na afine transformacije, tj. ako je

$$X \sim N(\mu, \sigma^2) \quad \text{i} \quad a, b \in \mathbb{R},\ a \neq 0$$

tada je

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

Zato svakoj normalno distribuiranoj slučajnoj varijabli $X \sim N(\mu, \sigma^2)$ možemo pridružiti **standardiziranu slučajnu varijablu**

$$X^* := \frac{X - \mathrm{E}[X]}{\sigma_X} = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

koja je također normalno distribuirana ali s parametrima 0 i 1.

Funkciju distribucije jedinične normalne razdiobe $N(0, 1)$ označavamo s $\Phi(x)$ i vrijedi:

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

Funkcija koju ćemo koristiti prilikom rješavanja zadataka i čije vrijednosti su tabelirane je

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

Veza među funkcijama Φ i Φ_0:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} e^{-\frac{t^2}{2}} dt + \frac{1}{\sqrt{2\pi}} \int_{0}^{x} e^{-\frac{t^2}{2}} dt = \frac{1}{2} + \Phi_0(x),$$

uz dogovor

$$\Phi_0(x) = -\Phi_0(-x) \quad \text{za} \quad x < 0.$$

Zadatak 17 *Neka je zadana slučajna varijabla $X \sim N(0, 1)$. Odredite vjerojatnosti događaja: a) $0 \leq X \leq 1$, b) $-1 \leq X \leq 2$, c) $X \leq 1$, d) $X \geq 1$*

Rješenje:

$$\begin{aligned} a) \qquad P(0 \leq X \leq 1) &= \Phi(1) - \Phi(0) = \frac{1}{2} + \Phi_0(1) - \frac{1}{2} - \Phi_0(0) \\ &= \Phi_0(1) - \Phi_0(0) = 0.3413 - 0 = 0.3413 \end{aligned}$$

$$b) \quad P(-1 \le X \le 2) = \Phi(2) - \Phi(-1) = \frac{1}{2} + \Phi_0(2) - \frac{1}{2} - \Phi_0(-1)$$
$$= \Phi_0(2) + \Phi_0(1) = 0.4772 + 0.3413 = 0.8185$$
$$c) \quad P(X \le 1) = \Phi(1) = \frac{1}{2} + \Phi_0(1) = 0.5 + 0.3413 = 0.8413$$
$$d) \quad P(X \ge 1) = 1 - P(X < 1) = 1 - P(X \le 1) = 1 - (1/2 + \Phi_0(1)) = 0.1587$$

□

Zadatak 18 *Neka je zadana slučajna varijabla $X \sim N(2, 4)$. Odredite vjerojatnosti događaja: a) $0 \le X \le 4$, b) $X \ge 4$*

Rješenje:

$$X \sim N(2,4) \quad \Rightarrow \quad X^* = \frac{X-\mu}{\sigma} = \frac{X-2}{2} \sim N(0,1)$$
$$a) \quad P(0 \le X \le 4) = F_X(4) - F_X(0)$$
$$= P\left(\frac{0-2}{2} \le \frac{X-2}{2} \le \frac{4-2}{2}\right) = P(-1 \le X^* \le 1)$$
$$= \Phi_0(1) - \Phi_0(-1) = 2\Phi_0(1) = 2 \cdot 0.3413 = 0.6826$$
$$b) \quad P(X \ge 4) = 1 - P(X < 4) = 1 - P(X \le 4) = 1 - F_X(4)$$
$$= 1 - P\left(X^* \le \frac{4-2}{2}\right) = 1 - \Phi(1) = 1 - \left(\frac{1}{2} + \Phi_0(1)\right)$$
$$= \frac{1}{2} - \Phi_0(1) = 0.5 - 0.3413 = 0.1587$$

□

Zadatak 19 *Slučajna varijabla X mjeri odstupanje aviona od sredine dozvoljenog koridora. Ona je normalno distribuirana, s očekivanjem 100m i standardnom devijacijom 200m. Ako je avion upravljen da leti sredinom koridora, nađite vjerojatnost da:*
a) avion leti kroz koridor širine 500m
b) iznad tog koridora.

Rješenje: Slučajna varijabla X mjeri odstupanje aviona od sredine koridora. Vrijedi: $X \sim N(100, 200^2)$
a) Ako želimo da avion leti sredinom koridora širine 500m, tada on od sredine

tog koridora može odstupati najviše 250m prema gore ili prema dole pa imamo:

$$\begin{aligned} P(-250 \leq X \leq 250) &= P\left(\frac{-250-100}{200} \leq X^* \leq \frac{250-100}{200}\right) \\ &= \Phi\left(\frac{3}{4}\right) - \Phi\left(-\frac{7}{4}\right) = \Phi_0\left(\frac{3}{4}\right) + \Phi_0\left(\frac{7}{4}\right) \\ &= 0.2734 + 0.4599 = 0.7333 \end{aligned}$$

b) Ako je avion iznad koridora, tada je $X \geq 250$.

$$\begin{aligned} P(X \geq 250) &= 1 - P(X < 250) = 1 - P(X \leq 250) = \\ &= 1 - \Phi\left(\frac{250-100}{200}\right) = 1 - \frac{1}{2} - \Phi_0\left(0.75\right) = 0.5 - 0.2734 = 0.2266 \end{aligned}$$

□

Zadatak 20 *Slučajna varijabla X ima normalnu razdiobu $N(2,4)$. Izračunajte uvjetnu vjerojatnost: $P(-1 \leq X \leq 1 \mid 0 < X < 3)$.*

Rješenje:

$$\begin{aligned} &P(A \mid B) = \frac{P(A \cap B)}{P(B)} \\ &P(A \cap B) = P(-1 \leq X \leq 1,\ 0 < X < 3) = P(0 < X \leq 1) \\ &P(-1 \leq X \leq 1 \mid 0 < X < 3) = \frac{P(0 < X \leq 1)}{P(0 < X < 3)} = \frac{P\left(\frac{0-2}{2} < X^* \leq \frac{1-2}{2}\right)}{P\left(\frac{0-2}{2} < X^* < \frac{3-2}{2}\right)} \\ &= \frac{\Phi_0(-0.5) - \Phi_0(-1)}{\Phi_0(0.5) - \Phi_0(-1)} = \frac{\Phi_0(1) - \Phi_0(0.5)}{\Phi_0(1) + \Phi_0(0.5)} = \frac{0.3413 - 0.1915}{0.3413 + 0.1915} = 0.2812 \end{aligned}$$

□

2.5.2 Aproksimacija binomne razdiobe normalnom

Neka je $X \sim B(n,p)$. Znamo da vrijedi $\mathrm{E}[X] = np$ i $\mathrm{Var}\,[X] = npq$.

Za velike n, vrijedi aproksimacija:

$$X^* = \frac{X - np}{\sqrt{npq}} \ \sim \ N(0,1)$$

Aproksimacija je to bolja što je vrijednost parametra p bliža $\frac{1}{2}$.

Vrijedi:

$$P(a \leq X \leq b) = \Phi\left(\frac{(b+0.5)-np}{\sqrt{npq}}\right) - \Phi\left(\frac{(a-0.5)-np}{\sqrt{npq}}\right)$$

Naime,

$$\begin{aligned} P(a \leq X \leq b) &= P\left(a - \frac{1}{2} < X < b + \frac{1}{2}\right) \\ &= P\left(\frac{a-\frac{1}{2}-np}{\sqrt{npq}} < X^* < \frac{b+\frac{1}{2}-np}{\sqrt{npq}}\right) = \Phi\left(\frac{b+\frac{1}{2}-np}{\sqrt{npq}}\right) - \Phi\left(\frac{a-\frac{1}{2}-np}{\sqrt{npq}}\right) \end{aligned}$$

Zadatak 21 *Neki stroj proizvodi* 60% *proizvoda prve kvalitete. Izračunajte vjerojatnost da u uzorku od 75 proizvoda bude barem 40 prve kvalitete.*

Rješenje: Slučajna varijabla X koja broji proizvode prve kvalitete ima razdiobu: $X \sim B(75, 0.6)$. Zanima nas $P(X \geq 40)$. Za realizaciju tog događaja "povoljni" su elementarni događaji $X = 40,\ X = 41,\ X = 42,\ \ldots, X = 50, \ldots, X = 60, \ldots, X = 75$. Račun direktnim korištenjem binomne razdobe bio bi stoga predug. Aproksimacija Poissonovom razdiobom, osim što bi izrazi koje moramo zbrojiti bili nešto jednostavniji, ne bi smanjila broj pribrojnika. No, aproksimacija normalnom razdiobom znatno će pojednostavniti stvar:

$$\begin{aligned} P(X \geq 40) &= 1 - P(X \leq 39) = 1 - P\left(X^* \leq \frac{39 + 0.5 - 75 \cdot 0.6}{\sqrt{75 \cdot 0.6 \cdot 0.4}}\right) \\ &= 1 - P\left(X^* \leq -1.296\right) = 1 - \Phi(-1.3) = 1 - \left(\frac{1}{2} + \Phi_0(-1.3)\right) \\ &= \frac{1}{2} + \Phi_0(1.3) = \frac{1}{2} + 0.4032 = 0.9032 \end{aligned}$$

□

2.5.3 Eksponencijalna razdioba

Definicija 9 *Neprekidna slučajna varijabla X ima* ***eksponencijalnu razdiobu*** *s parametrom λ ($\lambda > 0$) ako joj je funkcija gustoće zadana s:*

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

Oznaka:

$$X \sim Exp(\lambda)$$

Pogledajmo kako izgleda njena funkcija distribucije $F(x)$. Za $x \leq 0$, očito $F(x) = 0$, budući je tada

$$F(x) = \int_{-\infty}^{x} f(t)dt = \int_{-\infty}^{x} 0dt = 0.$$

Pretpostavimo sada da je $x > 0$. Tada:

$$F(x) = \int_{-\infty}^{x} f(t)dt = \int_{-\infty}^{0} 0dt + \int_{0}^{x} \lambda e^{-\lambda t}dt = 0 - e^{-\lambda t}\Big|_0^x = 1 - e^{-\lambda x}$$

Funkcija distribucije $F(x)$ slučajne varijable s eksponencijalnom razdiobom je dakle:

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

Očekivanje i varijanca:

$$\begin{aligned}
\mathrm{E}[X] &= \int_{-\infty}^{\infty} t \cdot f(t)dt = \int_{0}^{\infty} t \cdot \lambda e^{-\lambda t}dt = \lim_{M\to+\infty} \int_{0}^{M} t \cdot \lambda e^{-\lambda t}dt \\
&= \left| \begin{array}{ll} u = t & dv = \lambda e^{-\lambda t}dt \\ du = dt & v = -e^{-\lambda t} \end{array} \right| = \lim_{M\to+\infty} \left(-te^{-\lambda t}\Big|_0^M + \int_0^M e^{-\lambda t}dt \right) \\
&= \lim_{M\to+\infty} \left(-\frac{M}{e^{\lambda M}} - \frac{1}{\lambda} e^{-\lambda t}\Big|_0^M \right) = -\lim_{M\to+\infty} \frac{1}{\lambda \cdot e^{\lambda M}} - \frac{1}{\lambda} \lim_{M\to+\infty} (e^{-\lambda M} - 1) \\
&= 0 - \frac{1}{\lambda}(0 - 1) = \frac{1}{\lambda} \\
\mathrm{Var}[X] &= \mathrm{E}[X^2] - (\mathrm{E}[X])^2 = \ldots = \frac{1}{\lambda^2}
\end{aligned}$$

Zadatak 22 *Vrijeme ispravnog rada nekog uređaja je slučajna varijabla distribuirana po eksponencijalnom zakonu s očekivanjem 2 mjeseca. Kolika je vjerojatnost da će uređaj pokvariti u tijeku:*

a) prvog mjeseca

b) drugog mjeseca

c) drugog mjeseca, ako je poznato da u tijeku prvog mjeseca nije bio u kvaru.

Rješenje:

$$\mathrm{E}[X] = \frac{1}{\lambda} = 2 \;\; \Rightarrow \;\; \lambda = \frac{1}{2}$$

Slučajna varijabla X koja mjeri vrijeme ispravnog rada uređaja (izraženo u mjesecima) ima razdiobu $X \sim Exp\left[\frac{1}{2}\right]$. Njena funkcija distribucije je

$$F_X(x) = 1 - e^{-x/2}, \;\; x > 0$$

Događaj pod a) možemo izraziti kao $\{X \leq 1\}$, događaj pod b) kao $\{1 \leq X \leq 2\}$ a događaj pod c) kao $\{1 \leq X \leq 2 \mid X \geq 1\}$. Izračunajmo vjerojatnosti tih događaja:

$a)$ $\quad P(X \leq 1) = F_X(1) = 1 - e^{-1/2} = 0.393$

$b)$ $\quad P(1 \leq X \leq 2) = F_X(2) - F_X(1) = 1 - e^{-1} - (1 - e^{-1/2}) = 0.239$

$c)$ $\quad P(1 \leq X \leq 2 \mid X \geq 1) = \dfrac{P(1 \leq X \leq 2 \,,\; X \geq 1)}{P(X \geq 1)} = \dfrac{P(1 \leq X \leq 2)}{1 - P(X \leq 1)}$

$$= \frac{F_X(2) - F_X(1)}{1 - F_X(1)} = \frac{0.239}{1 - 0.393} = 0.393$$

□

Primjer slučajne varijable, odnosno distribucije, koja nema očekivanje: *Cauchyeva distribucija*

Definicija 10 *Kažemo da neprekidna slučajna varijabla X ima **Cauchyevu distribuciju** (s parametrima 1 i 0) ako joj je funkcija gustoće zadana s*

$$f_X(x) = \frac{1}{\pi(1 + x^2)}, \quad x \in \mathbb{R}$$

Provjerimo najprije da je razdioba dobro definirana. Očito je $f_X(x) \geq 0$ i nadalje:

$$\int_{-\infty}^{\infty} f_X(t)dt = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{1}{1+t^2}dt = \frac{2}{\pi}\int_0^{\infty} \frac{dt}{1+t^2} = \frac{2}{\pi}\lim_{M\to\infty}\int_0^M \frac{dt}{1+t^2}$$
$$= \frac{2}{\pi}\lim_{M\to\infty} \operatorname{arctg} t\Big|_0^M = \frac{2}{\pi}\left(\lim_{M\to\infty} \operatorname{arctg} M - \operatorname{arctg} 0\right) = \frac{2}{\pi}\left(\frac{\pi}{2} - 0\right) = 1$$

Svojstvo! $\mathrm{E}\,X$ postoji ako $\mathrm{E}\,|X| = \int_{-\infty}^{\infty} |t| f_X(t)dt < \infty$

$$\begin{aligned}
\mathrm{E}\,|X| &= \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{|t|dt}{1+t^2} = \frac{2}{\pi}\int_{0}^{\infty}\frac{|t|dt}{1+t^2} = \frac{2}{\pi}\lim_{M\to\infty}\int_{0}^{M}\frac{tdt}{1+t^2} \\
&= \{z = 1+t^2,\ dz = 2tdt\} = \frac{1}{\pi}\lim_{M\to\infty}\int_{1}^{M}\frac{dz}{z} \\
&= \frac{1}{\pi}\lim_{M\to\infty}\ln z\Big|_{1}^{M} = \frac{1}{\pi}\left(\lim_{M\to\infty}\ln M - \ln 1\right) = +\infty
\end{aligned}$$

pa zaključujemo da $\mathrm{E}\,X$ ne postoji.

3 Procjena parametara

Zanima nas statističko obilježje X neke promatrane populacije. Pretpostavimo da je X slučajna varijabla s konačnim očekivanjem $\mu = \mathrm{E}[X]$ i varijancom $\sigma^2 = \mathrm{Var}[X]$. μ i σ^2 su parametri razdiobe od X.

Promatramo **slučajne uzorke** koji se sastoje od n nezavisnih jednako distribuiranih slučajnih varijabli

$$X_1, X_2, \ldots, X_n$$

s distribucijom jednakom distribuciji mjerenog statističkog obilježja.

Nepristran i konzistentan procjenitelj za očekivanje μ je aritmetička sredina uzorka:

$$\bar{X}_n := \frac{1}{n}(X_1 + X_2 + \ldots + X_n)$$

Nepristran i konzistentan procjenitelj za varijancu σ^2 je uzoračka varijanca:

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

Definicija 11 *Neka su $L_n = l(X_1, \ldots, X_n)$ i $D_n = d(X_1, \ldots, X_n)$ slučajne varijable (statistike), funkcije slučajnog uzorka $X_1, \ldots, X_n$.*
*Kažemo da je $[L_n, D_n]$ $(1-\alpha)\cdot 100\%$ **pouzdan interval** za parametar τ ako vrijedi*

$$P(L_n \leq t \leq D_n) \geq 1 - \alpha, \quad \alpha \in \langle 0, 1\rangle$$

3.1 Pouzdani intervali za očekivanje normalne populacije

3.1.1 Varijanca poznata

Neka je X slučajna varijabla s nepoznatim očekivanjem μ i poznatom varijancom σ^2.

- imamo slučajni uzorak veličine n : $X_1, \ldots, X_n$

- $(1-\alpha)\cdot 100\%$ pouzdan interval za μ dobit ćemo ako promatramo uzoračku distribuciju statistike $\bar{X}_n$ (aritmetička sredina uzorka):

$$\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

(ona je normalno distribuirana ako je X normalno distribuirana i aproksimativno je normalna ako smo u uvjetima Centralnog graničnog teorema)

$$\Rightarrow\ \bar{X}_n^* = \frac{\bar{X}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}} = \frac{\bar{X}_n - \mu}{\sigma}\sqrt{n} \sim N(0,1)$$

Vrijedi:

$$\Phi_0(z_{\frac{\alpha}{2}}) = \frac{1-\alpha}{2}$$

i nadalje:

$$\begin{aligned}
&P(-z_{\frac{\alpha}{2}} \leq \bar{X}_n^* \leq z_{\frac{\alpha}{2}}) = 1-\alpha \\
\Leftrightarrow\ &P(-z_{\frac{\alpha}{2}} \leq \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{\frac{\alpha}{2}}) = 1-\alpha \\
\Leftrightarrow\ &P\left(\bar{X}_n - z_{\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + z_{\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha
\end{aligned}$$

$(1-\alpha)\cdot 100\%$ pouzdan interval za očekivanje normalne populacije (varijanca poznata)
$\bar{X}_n - z_{\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + z_{\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}}$

Zadatak 23 *Vrijeme trajanja neke vrste elektronskih cijevi je normalno distribuirana slučajna varijabla X s nepoznatim očekivanjem μ i standardnom devijacijom $\sigma = 40h$.*
a) Uzet je uzorak od 30 elektronskih cijevi za koji je dobiveno prosječno vrijeme trajanje od 780h. Nađite 99% *pouzdan interval za očekivanje μ vremena*

trajanja ove vrste elektronskih cijevi.
b) Koliki uzorak treba uzeti da bi se s vjerojatnošću 0.99, sredina uzorka $\bar{x}$ razlikovala od sredine μ manje od 10h?

Rješenje:

$X \sim N(\mu, 40^2)$

a) $n = 30, \quad \bar{x}_{30} = 780, \quad \alpha = 0.01$

$\Phi_0(z_{\frac{\alpha}{2}}) = \Phi_0(z_{0.005}) = \dfrac{1-\alpha}{2} = 0.495 \quad \Rightarrow \quad z_{0.005} = 2.58$

99% pouzdan interval za očekivanje:

$$\bar{x}_{30} \pm z_{0.005} \cdot \frac{\sigma}{\sqrt{n}} = 780 \pm 2.58 \cdot \frac{40}{\sqrt{30}} = 780 \pm 18.84$$
$$\Rightarrow \qquad 761.16 \leq \mu \leq 798.84$$

b) $P(|\bar{X}_n - \mu| < 10) = 0.99, \quad n = ?$

$$P(-10 < \bar{X}_n - \mu < 10) = P\left(-\frac{10}{\frac{\sigma}{\sqrt{n}}} < \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} < \frac{10}{\frac{\sigma}{\sqrt{n}}}\right)$$
$$= P\left(-\frac{10\sqrt{n}}{40} < \bar{X}_n^* < \frac{10\sqrt{n}}{40}\right) = 0.99$$
$$\Rightarrow \ \Phi_0\left(\frac{\sqrt{n}}{4}\right) - \Phi_0\left(-\frac{\sqrt{n}}{4}\right) = 0.99 \ \Leftrightarrow \ 2\Phi_0\left(\frac{\sqrt{n}}{4}\right) = 0.99 \ \Leftrightarrow \ \Phi_0\left(\frac{\sqrt{n}}{4}\right) = 0.495$$
$$\Rightarrow \ \frac{\sqrt{n}}{4} = 2.58 \ \Leftrightarrow \ \sqrt{n} = 10.32 \ \Rightarrow n = 106.5$$
$$\Longrightarrow \ n \geq 107, \quad \text{tj. treba uzeti uzorak duljine bar 107.}$$

Možemo razmišljati i ovako: $(1-\alpha)100\%$ pouzdani interval za očekivanje je:

$$\bar{X}_n - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

a odatle:

$$\Leftrightarrow \ -z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu - \bar{X}_n \leq z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \quad \Leftrightarrow \quad |\bar{X}_n - \mu| \leq z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

Kako nas zanima $P(|\bar{X}_n - \mu| < 10) = 0.99$, traženi n možemo odrediti iz uvjeta:

$$z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} = z_{0.005} \cdot \frac{\sigma}{\sqrt{n}} < 10.$$

Odatle dobivamo:

$$\sqrt{n} > \frac{z_{0.005} \cdot \sigma}{10} = \frac{2.58 \cdot 40}{10} = 10.32 \Rightarrow n > 106.5024 \quad \Rightarrow \quad n \geq 107$$

□

Zadatak 24 *(DZ) Neka mašina proizvodi kuglične ležajeve. Promjer kugličnog ležaja je normalna slučajna varijabla X s varijancom 1. Dužine 9 slučajno odabranih kugličnih ležajeva bile su*

$$20.1,\ 19.9,\ 20.0,\ 19.8,\ 19.7,\ 20.2,\ 20.1,\ 23.1,\ 22.8.$$

Odredite 95% *pouzdan interval za matematičko očekivanje slučajne varijable X.*

3.1.2 Varijanca nepoznata

Neka je $X \sim N(\mu, \sigma^2)$, μ i σ^2 nepoznati

- želimo naći $(1-\alpha) \cdot 100\%$ pouzdan interval za μ
- imamo slučajan uzorak veličine $n:\ X_1, \ldots, X_n$
- $\bar{X}_n$: aritmetička sredina uzorka
- varijancu σ^2 **procijenimo** pomoću S_n^2

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

(S_n^2 je nepristran i konzistentan procijenitelj za σ^2)

- standardiziranu varijablu

$$\bar{X}_n^* = \frac{\bar{X}_n - \mu}{\sigma} \sqrt{n}$$

aproksimativno zapisujemo pomoću procjene za σ^2:

$$T_n := \frac{\bar{X}_n - \mu}{S_n} \sqrt{n}$$

Statistika T_n ima **Studentovu ili t-distribuciju s** $(n-1)$ **stupnjeva slobode** : $T_n \sim t(n-1)$
Vrijedi:

$$P\left(-t_{\frac{\alpha}{2}}(n-1) \leq T_n \leq t_{\frac{\alpha}{2}}(n-1)\right) = 1-\alpha$$
$$\Leftrightarrow\ P\left(-t_{\frac{\alpha}{2}}(n-1) \leq \frac{\bar{X}_n - \mu}{\frac{S_n}{\sqrt{n}}} \leq t_{\frac{\alpha}{2}}(n-1)\right) = 1-\alpha$$
$$\Leftrightarrow\ P\left(\bar{X}_n - t_{\frac{\alpha}{2}}(n-1)\cdot\frac{S_n}{\sqrt{n}} \leq \mu \leq \bar{X}_n + t_{\frac{\alpha}{2}}(n-1)\cdot\frac{S_n}{\sqrt{n}}\right) = 1-\alpha$$

$(1-\alpha)\cdot 100\%$ pouzdan interval za očekivanje normalne populacije (varijanca nepoznata)

$$\bar{X}_n - t_{\frac{\alpha}{2}}(n-1)\cdot\frac{S_n}{\sqrt{n}} \leq \mu \leq \bar{X}_n + t_{\frac{\alpha}{2}}(n-1)\cdot\frac{S_n}{\sqrt{n}}$$

Napomena: Za $n \to \infty$, Studentova razdioba po distribuciji konvergira jediničnoj normalnoj razdiobi. Za broj stupnjeva slobode $n-1 \geq 30$ možemo aproksimativno uzeti da je $t(n-1) \approx N(0,1)$

Zadatak 25 *NASA testira komponente svojih raketa. Recimo da NASA želi procijeniti srednje vrijeme trajanja neke mehaničke komponente korištene u raketi "Columbia". Zbog ograničenja troškova, u simuliranim uvjetima svemira mogu testirati samo 10 komponenti. Dobiveni su podaci za vrijeme trajanja tih komponenti (u satima):* $\bar{x}_{10} = 1173.6, \quad s_{10} = 36.3$. *Procijenite očekivanje vijeka trajanja tih mehaničkih komponenti s* 95% *pouzdanim intervalom (pretpostavite da je vrijeme trajanja mehaničkih komponenti normalno distribuirano).*

Rješenje:

$$1-\alpha = 0.95 \ \Leftrightarrow\ \alpha = 0.05 \ \Leftrightarrow\ \frac{\alpha}{2} = 0.025$$
$$t_{0.025}(9) = 2.262$$
$$\bar{x}_{10} \pm t_{0.025}(9)\cdot\frac{s_{10}}{\sqrt{n}} = 1173.6 \pm 2.262\ \frac{36.3}{\sqrt{10}} = 1173.6 \pm 25.97$$
$$\Rightarrow\ 1147.63 \leq \mu \leq 1199.57$$

□

Zadatak 26 *(DZ) U svrhu istraživanja utjecaja toksičnih tvari koje luči jedna vrsta plijesni na kukuruz, biokemičar u 9 ekstrakata plijesni mjeri količinu toksičnih supstanci u mg. Dobiveni su rezultati: 1.2, 0.8, 0.6, 1.1, 1.2, 0.9, 1.5, 0.9, 1.0. Uz pretpostavku da su podaci iz normalne distribucije, procijenite* 98% *pouzdan interval za očekivanje te populacije.*

3.2 Pouzdani intervali za očekivanje populacije na osnovi velikih uzoraka

- pretpostavimo da je zadan slučajni uzorak $X_1, X_2, \ldots, X_n$ velike duljine $(n \to \infty)$ za X općenito nepoznate razdiobe, ali konačne varijance
- neka je μ parametar očekivanja i σ^2 varijanca
- želimo konstruirati aproksimativni pouzdani interval za μ
- prema Centralnom graničnom teoremu,

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

- nadalje, zbog konzistentnosti, $S_n \to \sigma$, pa

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

Dakle, za velike n $(n \to \infty)$ vrijedi

$$\frac{\bar{X}_n - \mu}{S_n}\sqrt{n} \overset{D}{\approx} N(0,1)$$

pa se $(1-\alpha)100\%$ pouzdani interval konstruira kao u slučaju normalne populacije s poznatom varijancom (za $\sigma \approx S_n$)

$(1-\alpha)\cdot 100\%$ pouzdan interval za očekivanje populacije na osnovi velikih uzoraka

$$\bar{X}_n - z_{\frac{\alpha}{2}} \cdot \frac{S_n}{\sqrt{n}} \leq \mu \leq \bar{X}_n + z_{\frac{\alpha}{2}} \cdot \frac{S_n}{\sqrt{n}}$$

Zadatak 27 *Zoolog želi procijeniti očekivanu količinu šećera u krvi određene životinjske vrste nastale nakon ubrizgavanja određene količine adrenalina. Dobivena srednja vrijednost uzorka od 55 životinja je 126.9 mg/100 ml uz standardnu devijaciju uzorka od 10.5 mg /100 ml. Odredite* 90% *pouzdan interval za očekivanje.*

Rješenje:

$$
\begin{aligned}
& n = 55, \quad \bar{x}_{55} = 126.9, \quad s_{55} = 10.5 \\
& 1 - \alpha = 0.9 \;\Leftrightarrow\; \alpha = 0.1 \;\Leftrightarrow\; \frac{\alpha}{2} = 0.05 \\
& \Phi_0(z_{0.05}) = \frac{0.9}{2} = 0.45 \;\Rightarrow\; z_{0.05} = 1.65 \\
& 126.9 \pm 1.65 \cdot \frac{10.5}{\sqrt{55}} = 126.9 \pm 2.34 \\
& \quad \Longrightarrow\; 124.56 \leq \mu \leq 129.24
\end{aligned}
$$

3.2.1 Pouzdan interval za parametar p binomne razdiobe

Tražimo $(1-\alpha)100\%$ pouzdani interval za proporciju p

$$X \sim B(n,p), \quad \mathrm{E}[X] = np, \quad \mathrm{Var}[X] = npq, \; q = 1-p$$

$\hat{P} = \frac{X}{n} = \bar{X}$ je nepristrani procjenitelj od p, tj. $E[\hat{P}] = p$

$$
\begin{aligned}
& \mathrm{E}[\hat{P}] = \mathrm{E}\left[\frac{X}{n}\right] = \frac{1}{n}\,\mathrm{E}[X] = \frac{1}{n} \cdot np = p \\
& \mathrm{Var}[\hat{P}] = \mathrm{Var}\left[\frac{X}{n}\right] = \frac{1}{n^2}\,\mathrm{Var}[X] = \frac{1}{n^2} \cdot npq = \frac{pq}{n} \\
& \Rightarrow\; \bar{X} = \hat{P} \sim N\left(p, \frac{pq}{n}\right) \\
& \Rightarrow\; \bar{X}^* = \frac{\bar{X} - p}{\sqrt{\frac{pq}{n}}} \sim N(0,1)
\end{aligned}
$$

Imamo:

$$
\begin{aligned}
& P(-z_{\frac{\alpha}{2}} \leq \bar{X}^* \leq z_{\frac{\alpha}{2}}) = 1 - \alpha \\
& \Leftrightarrow\; P(-z_{\frac{\alpha}{2}} \leq \frac{\bar{X} - p}{\sqrt{\frac{pq}{n}}} \leq z_{\frac{\alpha}{2}}) = 1 - \alpha \\
& \Leftrightarrow\; P\left(\bar{X} - z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{pq}{n}} \leq p \leq \bar{X} + z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{pq}{n}}\right) = 1 - \alpha
\end{aligned}
$$

Za veliki n, dobit ćemo dovoljno dobre rezultate ako zamijenimo p s $\bar{X} = \hat{p}$:

$$\boxed{\bar{X} - z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\bar{X}(1-\bar{X})}{n}} \leq p \leq \bar{X} + z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\bar{X}(1-\bar{X})}{n}}}$$

Zadatak 28 *Uzorak od 100 kućanstava nekog grada pokazao je da se u* 55% *kućanstava bar jedan član koristi Internetom.*
a) Nađite 95% *pouzdan interval za omjer kućanstava u tom gradu koja se služe Internetom.*
b) Koliko kućanstava treba uzeti da bi s vjerojatnošću od 0.95 mogli tvrditi da se najmanje 50% *kućanstava služi Internetom?*

Rješenje:

a) $\hat{p} = \bar{x} = 0.55, \quad n = 100$

$1 - \alpha = 0.95 \Leftrightarrow \alpha = 0.05 \Leftrightarrow \dfrac{\alpha}{2} = 0.025$

$z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$

$\hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}} = 0.55 \pm 1.96 \sqrt{\dfrac{0.55 \cdot 0.45}{100}} = 0.55 \pm 0.09751$

$\Longrightarrow 0.4525 \leq p \leq 0.6475$

b) $n = ?$

$$P\left(0.55 - 1.96\sqrt{\frac{0.55 \cdot 0.45}{n}} \leq p \leq 0.55 + 1.96\sqrt{\frac{0.55 \cdot 0.45}{n}}\right) = 0.95$$

želimo da vrijedi: $p \geq 0.5$ pa odatle

$0.55 - 1.96\sqrt{\dfrac{0.55 \cdot 0.45}{n}} \geq 0.5 \Leftrightarrow \dfrac{0.9751}{\sqrt{n}} \leq 0.05$

$\Leftrightarrow \sqrt{n} \geq 19.502 \Rightarrow n \geq 380.328$

$\Longrightarrow n \geq 381$

□

4 Testiranje statističkih hipoteza

Mnoge praktične situacije u vezi sa slučajnim pojavama zahtijevaju da se donesu odluke tipa DA ili NE. Npr. pri praćenju procesa proizvodnje nekog proizvoda treba, na temelju rezultata mjerenja $x_1, \ldots, x_n$ statističkog obilježja X, donijeti odluku o tome da li proces proizvodnje osigurava ili ne osigurava zahtjevanu kvalitetu. Pretpostavlja se, dakako, da obilježje X, koje karakterizira kvalitetu pojedinog proizvoda (količina određenog sastojka npr.) ima slučajni karakter.

Teorijski gledano, riječ je o tome da se na temelju n mjerenja slučajne varijable X, odnosno na temelju vrijednosti $(x_1, \ldots, x_n)$ slučajnog uzorka $(X_1, \ldots, X_n)$, donese odluka o prihvaćanju (DA) ili odbacivanju (NE) određene pretpostavke o svojstvima slučajne varijable X. Takva pretpostavka zove se *statistička hipoteza*, a postupak donošenja odluke o prihvaćanju ili odbacivanju statističke hipoteze zove se *testiranje*.

Primjer 19 *Želimo testirati da li je očekivanje trajanja neke vrste žarulja jednako npr. 1000h.*

Definiramo

$$H_0 : \mu = 1000h$$
$$H_1 : \mu \neq 1000h$$

H_0 je **nulta hipoteza**, a H_1 **alternativna hipoteza**. Budući iz alternativne hipoteze slijedi da može biti $\mu > 1000h$ ili $\mu < 1000h$, kažemo da je H_1 **dvostrana alternativna hipoteza**.

Ponekad je zgodnije imati **jednostranu alternativnu hipotezu**. Npr.

$$H_0 : \mu = 1000h$$
$$H_1 : \mu > 1000h$$

ili

$$H_1 : \mu < 1000h$$

Ukratko, nulta hipoteza u testu je na neki način "fiksna", dok je alternativna ona kod koje imamo mogućnost izbora.

Testiranje hipoteze (odnosno provjeru da li je ona istinita ili nije) provodimo na sljedeći način: uzmemo slučajni uzorak, izračunamo vrijednost odgovarajuće test-statistike, te na osnovu njene vrijednosti odlučujemo o istinitosti hipoteze.

Prilikom donošenja odluke o istinitosti hipoteze, postoji mogućnost pogreške, tj. krive odluke. Dvije su vrste mogućih pogrešaka:

→ *pogreška 1.vrste*: odbacili smo nultu hipotezu ako je ona istinita

→ *pogreška 2.vrste*: prihvatili smo nultu hipotezu ako je ona neistinita

	H_0 istinita	H_0 neistinita
prihvaćamo H_0	$\surd$	pogreška 2.vrste
odbacujemo H_0	pogreška 1.vrste	$\surd$

α = P(pogreška 1.vrste)= P(odbacujemo H_0 | H_0 istinita) ⇒ **nivo signifikantnosti** ili **razina značajnosti**

β = P(pogreška 2.vrste)= P(prihvaćamo H_0 | H_0 neistinita)

1-β=P(odbacujemo H_0 | H_0 neistinita) ⇒ **snaga testa**

Testiranja hipoteza (koja su ovdje obrađena) baziraju se na odgovarajućim *pouzdanim intervalima*. Ako izračunata vrijednost odgovarajuće test-statistike upadne u pouzdan interval tražene pouzdanosti, tada nultu hipotezu ne možemo odbaciti; ukoliko ona ne upadne u isti interval, nultu hipotezu odbacujemo!

4.1 Test o očekivanju normalno distribuirane populacije

4.1.1 Varijanca poznata

- neka je $X \sim N(\mu, \sigma^2)$, σ poznata

- imamo slučajni uzorak veličine n : $(X_1, \cdots, X_n)$
- želimo testirati da li je očekivanje μ jednako nekom unaprijed zadanom broju μ_0. Nulta hipoteza je $H_0 : \mu = \mu_0$. Za alternativnu možemo uzeti bilo koju od sljedeće tri:

$$H_1 : \mu \neq \mu_0 \quad \text{ili} \quad H_1 : \mu > \mu_0 \quad \text{ili} \quad H_1 : \mu < \mu_0$$

- u sva 3 slučaja koristimo istu test-statistiku:

$$Z = \frac{\bar{X}_n - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

  Ako je nulta hipoteza $H_0 : \mu = \mu_0$ istinita, tada je $E[\bar{X}] = \mu_0$, odnosno $Z \sim N(0, 1)$

Promotrimo redom slučajeve različitog izbora alternativne hipoteze:

$$\begin{aligned} 1. \qquad & H_0 : \mu = \mu_0 \\ & H_1 : \mu \neq \mu_0 \end{aligned}$$

Ako je $H_0 : \mu = \mu_0$ istinita, tada

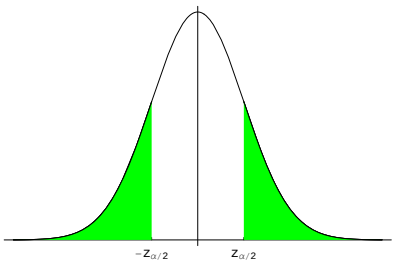
$$P(-z_{\frac{\alpha}{2}} \leq Z \leq z_{\frac{\alpha}{2}}) = 1 - \alpha$$

što je vjerojatnost da prihvatimo H_0 ako je ona istinita. S druge strane,

$$P((Z < -z_{\frac{\alpha}{2}}) \cup (Z > z_{\frac{\alpha}{2}})) = \alpha$$

je vjerojatnost da *ne* prihvatimo H_0 ako je one istinita.
Dakle,

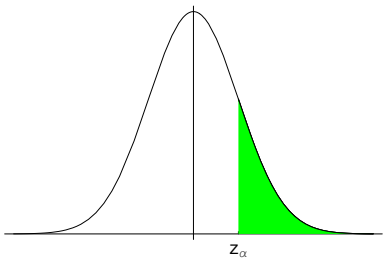
ako je $Z < -z_{\frac{\alpha}{2}}$ ili $Z > z_{\frac{\alpha}{2}}$	$\Rightarrow$ odbacujemo H_0
Ako je $-z_{\frac{\alpha}{2}} \leq Z \leq z_{\frac{\alpha}{2}}$	$\Rightarrow$ ne možemo odbaciti H_0



2. $H_0 : \mu = \mu_0$ H_0 odbacujemo ako je

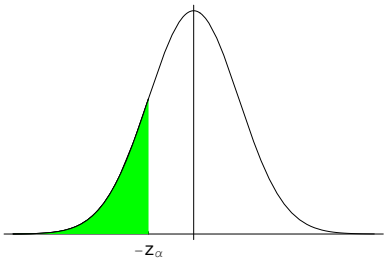
$H_1 : \mu > \mu_0$ $\boxed{Z > z_\alpha}$

(ne $z_{\frac{\alpha}{2}}$, nego z_α!!! Kritično područje površine α je svo na desnoj strani)



3. $H_0 : \mu = \mu_0$ H_0 odbacujemo ako je

$H_1 : \mu < \mu_0$ $\boxed{Z < -z_\alpha}$



Napomena: Treba paziti na terminologiju: ne kaže se "prihvaćamo hipotezu", nego "ne možemo ju odbaciti".

Zadatak 29 *Poznato je da napon u električnoj mreži od 220 volti ima normalnu distribuciju sa standarnom devijacijom od 6 volti. Ako je 16 nezavisnih mjerenja dalo rezultate:*

208, 216, 215, 228, 210, 224, 212, 213, 224, 218, 206, 209, 208, 218, 220, 206,

s razinom značajnosti 0.01 provjerite pretpostavku da je došlo do pada srednjeg napona u elektičnoj mreži.

Rješenje:

$$X \sim N(\mu, 6^2), \quad n = 16$$

Postavljamo hipoteze:

$$H_0 : \mu = 220$$
$$H_1 : \mu < 220$$

Nulta hipoteza je da je srednja vrijednost napona jednaka 220 (odnosno da je veća od te vrijednosti), dakle da *nije došlo* do pada napona, dok je alternativna da je srednja vrijednost napona manja od 220, odnosno da *je došlo* do pada napona, što je tvrdnja za koju želimo provjeriti da li vrijedi. Kad bismo kao alternativnu hipotezu uzeli $H_1 : \mu \neq 220$, u slučaju odbacivanja nulte hipoteze $H_0 : \mu = 220$, mogli bismo zaključiti samo da srednji napon *nije jednak* 220, no ne bismo znali je li on veći ili manji od te vrijednosti.

Računamo vrijednost test-statistike: $Z = \dfrac{\bar{X}_n - \mu_0}{\sigma}\sqrt{n}$

$$\mu_0 = 220, \qquad \bar{x}_{16} = 214.6875$$
$$\Rightarrow \; z = \frac{214.6875 - 220}{6}\sqrt{16} = -3.54167$$
$$z_\alpha = z_{0.01} = 2.325$$
$$\Longrightarrow \; z < -z_{0.01}$$

$\Longrightarrow$ odbacujemo nultu hipotezu H_0, tj. došlo je do pada napona! □

4.1.2 Varijanca nepoznata

- neka je $X \sim N(\mu, \sigma^2)$, $\quad \sigma$ nepoznata
- imamo njen slučajni uzorak veličine n : $(X_1, \cdots, X_n)$
- želimo testirati da li je očekivanje μ jednako nekom unaprijed zadanom broju μ_0
- koristimo test-statistiku:

$$T = \frac{\bar{X}_n - \mu_0}{S_n}\sqrt{n}$$

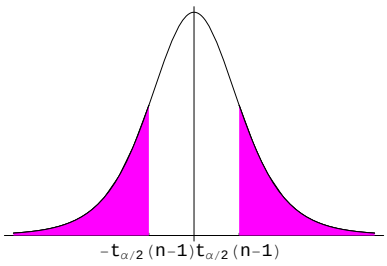
Ako je nulta hipoteza $H_0 : \mu = \mu_0$ istinita, tada je $T \sim t(n-1)$

1.

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

Nultu hipotezu H_0 odbacujemo ako je

$$T > t_{\frac{\alpha}{2}}(n-1) \quad \text{ili} \quad T < -t_{\frac{\alpha}{2}}(n-1)$$

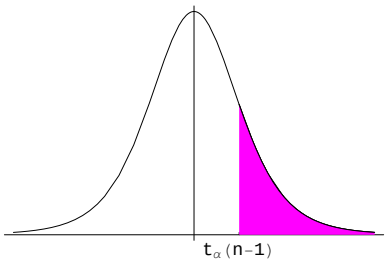


2.

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

H_0 odbacujemo ako je

$$T > t_\alpha(n-1)$$

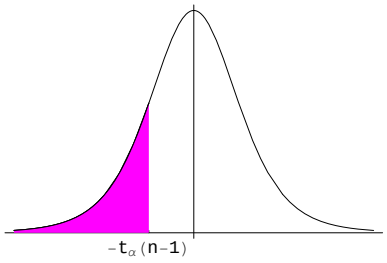


3.

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

H_0 odbacujemo ako je

$$T < -t_\alpha(n-1)$$



Zadatak 30 *Tvornica tvrdi da je prosječan vijek trajanja baterija iz te tvornice 21.5 sati. Na slučajnom uzorku od 6 baterija iz te tvornice laboratorijskim mjerenjima vijeka trajanja dobivene su vrijednosti od 19, 18, 22, 20, 16, 25 sati. S razinom značajosti* $\alpha = 0.05$*, testirajte da li dobiveni uzorak indicira kraći prosječan vijek trajanja baterija.*

Rješenje:

$$\mu_0 = 21.5, \quad n = 6, \quad \alpha = 0.05$$

$$H_0 : \mu = 21.5$$
$$H_1 : \mu < 21.5$$

Treba nam vrijednosti test-statistike: $T = \dfrac{\bar{X}_n - \mu_0}{S_n}\sqrt{n} \sim t(n-1)$

$$\bar{x}_6 = \frac{1}{6}(19 + 18 + 22 + 20 + 16 + 25) = 20$$

$$s_6^2 = \frac{1}{5}\sum_{i=1}^{6}(x_i - \bar{x}_6)^2 = \frac{1}{5}\left(\sum_{i=1}^{6} x_i^2 - 6 \cdot \bar{x}_6^2\right) = \frac{50}{5} = 10$$

$$\Rightarrow\ t = \frac{20 - 21.5}{\sqrt{10}}\sqrt{6} = -1.162$$

$$t_{0.05}(5) = 2.015$$

$$\Rightarrow\ t > -t_{0.05}(5)$$

Nultu hipotezu H_0 ne možemo odbaciti, tj. ne možemo zaključiti da uzorak indicira kraći prosječni vijek trajanja baterija. □

4.2 Testovi o očekivanju na osnovi velikih uzoraka

- NE pretpostavljamo da slučajni uzorak uzimamo iz normalno distribuirane populacije
- iz Centralnog graničnog teorema za $n \to \infty$ slijedi da test-statistika

$$Z = \frac{\bar{X}_n - \mu_0}{S_n}\sqrt{n} \overset{H_0}{\approx} N(0,1)$$

- osnovna hipoteza je ponovo oblika $H_0 : \mu = \mu_0$ za neki unaprijed zadani broj μ_0
- svodi se na testiranje očekivanja normalno distribuirane populacije uz $\sigma \approx S_n$ jer $S_n^2 \to \sigma^2$ kad $n \to \infty$

4.2.1 Test o proporciji

Pogledajmo kako izgleda test za očekivanje na osnovi velikih uzoraka u slučaju kada imamo binomno distribuiranu populaciju.

- promatramo statističko obilježje $X \sim B(n, p)$

- želimo testirati da li je proporcija p jednaka nekom unaprijed zadanom broju p_0. Nulta hipoteza je

$$H_0 : p = p_0.$$

  Za alternativnu možemo uzeti bilo koju od sljedeće tri:

$$H_1 : p \neq p_0 \quad \text{ili} \quad H_1 : p > p_0 \quad \text{ili} \quad H_1 : p < p_0$$

- u sva 3 slučaja koristimo istu test-statistiku:

$$Z = \frac{\bar{X} - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n} \sim N(0,1)$$

  gdje je $\bar{X} = \hat{P}$

Promotrimo redom slučajeve različitog izbora alternativne hipoteze:

1.

$$H_0 : p = p_0$$
$$H_1 : p \neq p_0$$

   Nultu hipotezu H_0 odbacujemo ako je $\quad Z > z_{\frac{\alpha}{2}} \quad$ ili $\quad Z < -z_{\frac{\alpha}{2}}$

2.

$$H_0 : p = p_0$$
$$H_1 : p > p_0$$

   H_0 odbacujemo ako je $\quad Z > z_\alpha$

3.

$$H_0 : p = p_0$$
$$H_1 : p < p_0$$

   H_0 odbacujemo ako je $\quad Z < -z_\alpha$

Zadatak 31 *Proizvođač tvrdi da njegove pošiljke sadrže najviše* 7% *defektnih proizvoda. Uzet je slučajni uzorak od 200 komada iz jedne pošiljke i bilo je 11 defektnih. Da li biste prihvatili tvrdnju proizvođača uz razinu značajnosti* 0.05*?*

Rješenje: Postavljamo hipoteze:

$$H_0 : p = 0.07$$
$$H_1 : p < 0.07$$

Kada bi za alternativnu hipotezu postavili $H_1 : p \neq 0.07$, u slučaju odbacivanja nulte hipoteze mogli bi zaključiti samo da proporcija defektnih nije 0.07, a to može značiti da je veća, ali i da je manja od te vrijednosti što je još bolje. Izračunajmo vrijednost odgovarajuće test-statistike:

$$\bar{x}_{200} = \hat{p} = \frac{11}{200} = 0.055 \implies z = \frac{0.055 - 0.07}{\sqrt{0.07 \cdot 0.93}}\sqrt{200} = -0.83$$
$$z_\alpha = z_{0.05} = 1.65$$
$$\Rightarrow \; z > -z_{0.05}$$

Nultu hipotezu H_0 ne možemo odbaciti, tj. ne možemo zaključiti da pošiljke sadrže manje od 7% defektnih proizvoda. □

4.3 Usporedba očekivanja dviju normalno distribuiranih populacija (t-test)

- pretpostavimo da mjerimo isto statističko obilježje X na dvije različite populacije
- pretpostavimo da je u obje populacije X normalno distribuirana slučajna varijabla **s jednakom varijancom** σ^2

$$X^{(1)} : \text{statističko obilježje } X \text{ za populaciju 1}, \quad X^{(1)} \sim N(\mu_1, \sigma^2)$$
$$X^{(2)} : \text{statističko obilježje } X \text{ za populaciju 2}, \quad X^{(2)} \sim N(\mu_2, \sigma^2)$$

- iz svake populacije uzimamo uzorak:

$$X_1^{(1)},\ X_2^{(1)},\ldots,X_{n_1}^{(1)} \text{ za } X^{(1)} \text{ duljine } n_1$$
$$X_1^{(2)},\ X_2^{(2)},\ldots,X_{n_2}^{(2)} \text{ za } X^{(2)} \text{ duljine } n_2$$

- želimo testirati sljedeću nultu hipotezu

$$H_0 : \mu_1 = \mu_2$$

u odnosu na neku od jednostranih alternativa

$$H_1 : \mu_1 < \mu_2 \quad \text{ili} \quad H_1 : \mu_1 > \mu_2$$

ili u odnosu na dvostranu alternativu

$$H_1 : \mu_1 \neq \mu_2$$

- u svim slučajevima koristimo istu test-statistiku

$$T = \frac{\bar{X}_1 - \bar{X}_2}{S} \cdot \frac{1}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

gdje su

$$\bar{X}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i^{(1)}, \qquad \bar{X}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} X_i^{(2)},$$

$$S^2 = \frac{1}{n_1 + n_2 - 2}\left((n_1 - 1)S_1^2 + (n_2 - 1)S_2^2\right)$$

za $S_1^2,\ S_2^2$ uzoračke varijance uzoraka 1 i 2. S^2 se interpretira kao **zajednička varijanca uzoraka** 1 i 2. Ako je H_0 istinita, tada je

$$T \sim t(n_1 + n_2 - 2)$$

1.

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

Nultu hipotezu H_0 odbacujemo ako

$$T > t_{\frac{\alpha}{2}}(n_1 + n_2 - 2) \quad \text{ili} \quad T < -t_{\frac{\alpha}{2}}(n_1 + n_2 - 2)$$

2.

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 > \mu_2$$

Nultu hipotezu H_0 odbacujemo ako

$$T > t_\alpha(n_1 + n_2 - 2)$$

3.

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 < \mu_2$$

Nultu hipotezu H_0 odbacujemo ako

$$T < -t_\alpha(n_1 + n_2 - 2)$$

Zadatak 32 *Ista vrsta jabuka uzgaja se u Slavoniji i u Zagorju. Na slučajan način izabrano je 7 slavonskih stabala te je izmjeren njihov prinos (u kg): 28, 26, 33, 29, 31, 27, 28; prinos sa 10 zagorskih stabala bio je: 36, 25, 21, 29, 30, 36, 27, 28, 30, 37. Uz razinu značajnosti 0.01, testirajte hipotezu da jabuke u Zagorju daju veći prinos, ako je poznato da je prinos normalna slučajna varijabla. Možemo li, uz istu razinu značajnosti, zaključiti da se prinosi jabuka u Slavoniji i Zagorju razlikuju?*

Rješenje:

$$n_1 = 7, \quad n_2 = 10$$

Postavljamo hipoteze

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 < \mu_2$$

Koristimo test-statistiku

$$T = \frac{\bar{X}_1 - \bar{X}_2}{S} \cdot \frac{1}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$$

$$\bar{x}_1 = \frac{1}{7}(28+26+33+29+31+27+28) = 28.857$$
$$\bar{x}_2 = \frac{1}{10}(36+25+21+29+30+36+27+28+30+37) = 29.9$$
$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)$$
$$\Rightarrow\ s_1^2 = \frac{1}{6}\cdot 34.855 = 5.81, \quad s_2^2 = \frac{1}{9}\cdot 240.9 = 26.767$$
$$s^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = \frac{6\cdot 5.81 + 9\cdot 26.767}{7+10-2} = 18.3842$$
$$\Rightarrow\ s = 4.2877$$
$$t = \frac{28.857 - 29.9}{4.2877\sqrt{\frac{1}{7}+\frac{1}{10}}} = -0.4936$$
$$t_\alpha(n_1+n_2-2) = t_{0.01}(15) = 2.602$$
$$\Rightarrow t > -t_{0.01}(15)$$

Ne možemo odbaciti H_0, tj. ne možemo zaključiti da jabuke u Zagorju daju veći prinos.
Ako želimo testirati da li su prinosi različiti, moramo postaviti hipoteze

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

Tada nam treba

$$t_{\frac{\alpha}{2}}(n_1+n_2-2) = t_{0.005}(15) = 2.949$$

Kako je

$$t > -t_{0.005}(15)$$

(i očito $t < t_{0.005}(15)$) ponovo ne možemo odbaciti nultu hipotezu, tj. ne možemo zaključiti da se prinosi jabuka razlikuju. □

4.4 Usporedba proporcija

- promatramo dvije populacije i neko njihovo Bernoullijevo statističko obilježje X

$X^{(1)}$: slučajna varijabla koja reprezentira X u populaciji 1
$X^{(2)}$: slučajna varijabla koja reprezentira X u populaciji 2

- pripadni parametri (vjerojatnosti uspjeha): $p_1,\ p_2$
- sa $\hat{p}_1$ i $\hat{p}_2$ označimo procjenitelje od p_1 i p_2 na bazi uzorka iz svake od populacija duljine n_1 i n_2 (uzorci su međusobno nezavisni), te sa

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

procjenu zajedničke vjerojatnosti uspjeha

- koristimo test-statistiku

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})}} \cdot \frac{1}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

- za velike uzorke, tj. kada $\min(n_1, n_2) \to +\infty$, vrijedi $Z \approx N(0,1)$

1.

$$H_0 : p_1 = p_2$$
$$H_1 : p_1 \neq p_2$$

Nultu hipotezu H_0 odbacujemo ako

$$Z > z_{\frac{\alpha}{2}} \quad \text{ili} \quad Z < -z_{\frac{\alpha}{2}}$$

2.

$$H_0 : p_1 = p_2$$
$$H_1 : p_1 > p_2$$

Nultu hipotezu H_0 odbacujemo ako

$$Z > z_\alpha$$

3.

$$H_0 : p_1 = p_2$$
$$H_1 : p_1 < p_2$$

Nultu hipotezu H_0 odbacujemo ako

$$Z < -z_\alpha$$

Zadatak 33 *Uzorci od 300 glasača iz županije A i 200 glasača iz županije B pokazali su da će* 56% *i* 48% *ljudi, redom, glasati za nekog određenog kandidata. S razinom značajnosti 0.05, testirajte hipotezu da*
a) postoji razlika među županijama
b) tog kandidata više "vole" u županiji A.

Rješenje:

$$n_1 = 300, \quad \hat{p}_1 = 0.56$$
$$n_2 = 200, \quad \hat{p}_2 = 0.48$$

a) $H_0 : p_1 = p_2$
$H_1 : p_1 \neq p_2$

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{300 \cdot 0.56 + 200 \cdot 0.48}{500} = 0.528$$

$$z = \frac{0.56 - 0.48}{\sqrt{0.528 \cdot 0.472}} \cdot \frac{1}{\sqrt{\frac{1}{300} + \frac{1}{200}}} = 1.75$$

$$z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$$
$$\Rightarrow \ z < z_{0.025}$$

$\Longrightarrow$ Ne možemo odbaciti nultu hipotezu, tj. ne možemo zaključiti da postoji razlika među županijama.

b) $H_0 : p_1 = p_2$

$$H_1 : p_1 > p_2$$

$$z_\alpha = z_{0.05} = 1.64 \quad \Rightarrow \; z > z_{0.05}$$

$\Longrightarrow$ Odbacujemo nultu hipotezu, tj. možemo zaključiti da kandidata više "vole" u županiji A. □

4.5 Usporedba varijanci dviju normalno distribuiranih populacija (F-test)

- neka je $X^{(1)} \sim N(\mu_1, \sigma_1^2), \quad X^{(2)} \sim N(\mu_2, \sigma_2^2)$
- imamo slučajne uzorke veličine n_i od $X_i,\ i = 1, 2$

$$\begin{array}{l} X_1^{(1)},\ X_2^{(1)}, \ldots, X_{n_1}^{(1)} \text{ za } X^{(1)} \text{ duljine } n_1 \\ X_1^{(2)},\ X_2^{(2)}, \ldots, X_{n_2}^{(2)} \text{ za } X^{(2)} \text{ duljine } n_2 \end{array}$$

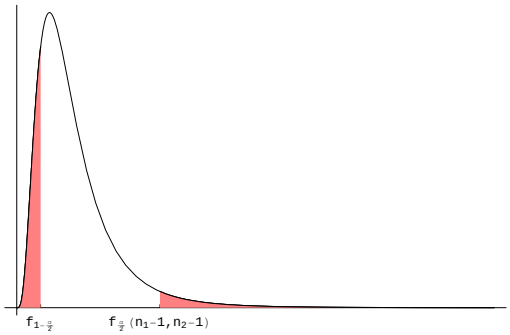
- test- statistika
$$F = \frac{S_1^2}{S_2^2} \;\sim\; F(n_1 - 1, n_2 - 1)$$
ima **Fisherovu** ili **F-distribuciju** sa parom stupnjeva slobode $(n_1 - 1, n_2 - 1)$.

- Vrijedi
$$f_{1-\frac{\alpha}{2}}(n_1, n_2) = \frac{1}{f_{\frac{\alpha}{2}}(n_2, n_1)}$$

1. $H_0 : \sigma_1^2 = \sigma_2^2$
   $H_1 : \sigma_1^2 \neq \sigma_2^2$

Nultu hipotezu H_0 odbacujemo ako

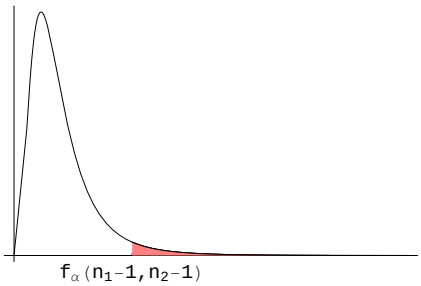
$$F > f_{\frac{\alpha}{2}}(n_1 - 1, n_2 - 1) \quad \text{ili} \quad F < f_{1-\frac{\alpha}{2}}(n_1 - 1, n_2 - 1)$$



2. $H_0 : \sigma_1^2 = \sigma_2^2$
   $H_1 : \sigma_1^2 > \sigma_2^2$

Nultu hipotezu H_0 odbacujemo ako

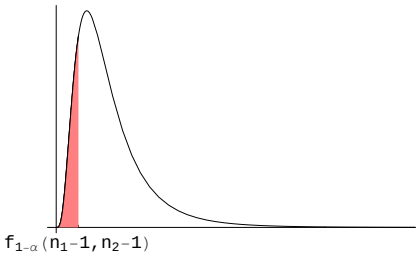
$$F > f_\alpha(n_1 - 1, n_2 - 1)$$



3. $H_0 : \sigma_1^2 = \sigma_2^2$
   $H_1 : \sigma_1^2 < \sigma_2^2$

Nultu hipotezu H_0 odbacujemo ako

$$F < f_{1-\alpha}(n_1 - 1, n_2 - 1)$$



Zadatak 34 *Iz dva 3.razreda neke srednje škole izabrano je, na slučajan način, po 10 učenika i izmjerena je njihova težina (zna se da je težina normalno distribuirana), a podaci su dani u tablici. Uz razinu značajnosti 0.02, testirajte hipotezu da su varijance jednake.*

3a:	57	60	63	59	62	60	58	56	54	62
3b:	58	62	60	56	63	58	61	57	53	61

Rješenje:

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

$$\bar{x}_1 = 59.1, \quad \bar{x}_2 = 58.9$$
$$s_1^2 = \frac{1}{9}\left(\sum_{i=1}^{10} x_i^2 - n\bar{x}^2\right) = 8.322, \quad s_2^2 = 9.433$$
$$\Rightarrow \; f = \frac{s_1^2}{s_2^2} = \frac{8.322}{9.433} = 0.8822$$
$$f_{\frac{\alpha}{2}}(n_1 - 1, n_2 - 1) = f_{0.01}(9, 9) = 5.35$$
$$f_{1-\frac{\alpha}{2}}(n_1 - 1, n_2 - 1) = f_{0.99}(9, 9) = \frac{1}{f_{\frac{\alpha}{2}}(n_2 - 1, n_1 - 1)} = \frac{1}{f_{0.01}(9, 9)} = 0.1869$$
$$\Rightarrow \; f_{0.99}(9, 9) < f < f_{0.01}(9, 9)$$

Ne možemo odbaciti nultu hipotezu, tj. ne možemo zaključiti da se varijance u ova dva uzorka razlikuju. □

4.6 χ^2 - test o prilagodbi modela podacima

- test-statistika je općenito

$$H = \sum_{i=1}^{k} \frac{(f_i - f_i')^2}{f_i'}$$

gdje su f_i eksperimentalne, a $f_i' = np_i$ teorijske frekvencije.

- ako vrijedi H_0, tada za velike n $(n \to \infty)$

$$H \sim \chi^2(k - r - 1)$$

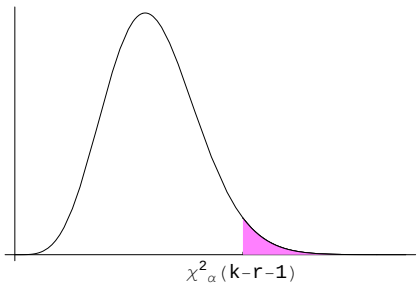
gdje $\chi^2(m)$ označava χ^2-**razdiobu s m stupnjeva slobode**.

- pritom je

$$k = \text{(konačan) broj razreda u tablici}$$
$$r = \text{broj nepoznatih parametara}$$

- nultu hipotezu da se radi o određenoj razdiobi odbacujemo ako

$$H \geq \chi^2_\alpha(k - r - 1)$$



Zadatak 35 *Proizvođač tvrdi da je* 5% *njegovih proizvoda prve klase,* 92% *druge i* 3% *treće klase. U slučajnom uzorku od 500 proizvoda nađeno je 40 proizvoda prve, 432 druge i 28 treće klase. Uz razinu značajnosti 0.05, testirajte hipotezu da je proizvođač u pravu.*

Rješenje: Proizvođač tvrdi da njegovi proizvodi imaju neku distribuciju, odnosno razdiobu. Govori li istinu, provjerit ćemo χ^2- testom. Duljina uzorka je $n = 500$. Kako bismo izračunali vrijednost odgovarajuće test-statistike trebaju nam teorijske frekvencije. Njih računamo po formuli $f_i' = np_i$ gdje je p_i odgovarajuća vjerojatnost, odnosno u ovom slučaju odgovarajuća proporcija. Tako je

$$p_1 = \frac{5}{100}, \quad p_2 = \frac{92}{100}, \quad p_3 = \frac{3}{100}.$$

Formirajmo tablicu:

i	f_i	f_i'	$\frac{(f_i - f_i')^2}{f_i'}$
1	40	$500 \cdot \frac{5}{100} = 25$	9
2	432	$500 \cdot \frac{92}{100} = 460$	1.7
3	28	$500 \cdot \frac{3}{100} = 15$	11.27
Σ	500	500	21.97

Suma poslijednjeg stupca u tablici daje nam vrijednost tražene test-statistike:

$$h = \sum_{i=1}^{3} \frac{(f_i - f_i')^2}{f_i'} = 21.97$$

Tablična vrijednost s kojom ju moramo usporediti kako bismo donijeli odluku o istinitosti nulte hipoteze je $\chi^2_\alpha(k - r - 1)$. α je zadana (=0.05), $k = 3$ (ukupan broj razreda), a $r = 0$ (nije bilo nijednog nepoznatog parametra pa ništa nije bilo potrebno procijenjivati). Dakle,

$$\chi^2_\alpha(k - r - 1) = \chi^2_{0.05}(2) = 6.0$$

Kako je

$$h > \chi^2_{0.05}(2),$$

što znači da je vrijednost test-statistike upala u kritično područje, moramo odbaciti nultu hipotezu. Drugim riječima, odbacujemo tvrdnju proizvođača, tj. on nije u pravu. □

Zadatak 36 *Pet novčića, s istom ali nepoznatom vjerojatnošću p da padne pismo, bacaju se 100 puta (rezultati su dani u tablici). Uz razinu značajnosti*

0.01, testirajte hipotezu da broj pisama koji se dobije u jednom bacanju predstavlja binomnu slučajnu varijablu.

broj pisama x_i	0	1	2	3	4	5
frekvencija f_i	3	16	36	32	11	2

Rješenje: Potrebno je provjeriti imaju li dani podaci binomnu distribuciju. Pokus koji izvodimo (ponavljamo ga 100 puta, dakle $n = 100$) je bacanje novčića 5 puta a "uspjeh" je "palo je pismo". Slučajna varijabla X broji pisma. Parametar n binomne distribucije je stoga jednak 5. Parametar p nije zadan te moramo ga procijeniti. Oprez! mali n sada označava i duljinu uzorka i parametar distribucije, no to su različite stvari i različite vrijednosti pa treba na to pripaziti.

Parametar p jednak je vjerojatnosti "uspjeha" u jednom bacanju novčića. Njegovu procjenu dobijemo tako da ukupan broj palih pisama podijelimo sa ukupnim brojem bacanja novčića. Novčić je ukupno bačen $5 \cdot 100 = 500$ puta (100 pokusa a svaki se sastoji od 5 bacanja). Ukupan broj pisama računamo pomoću dane tablice:

$$0 \cdot 3 + 1 \cdot 16 + 2 \cdot 36 + 3 \cdot 32 + 4 \cdot 11 + 5 \cdot 2 = 238.$$

Konačno,

$$\hat{p} = \frac{238}{500} = 0.476$$

Sljedeći korak je izračunati teorijske frekvencije $f_i' = np_i$. Funkcija gustoće slučajne varijable $X \sim B(5, 0.476)$ je

$$p_i := p_X(i) = P(X = i) = \binom{5}{i}(0.476)^i \cdot (0.524)^{5-i},$$

pa dobivamo

$$f_0' = 100 \cdot p_0 = 100 \cdot \binom{5}{0}(0.476)^0 \cdot (0.524)^5 = 3.95054$$

$$f_1' = 100 \cdot p_1 = 100 \cdot \binom{5}{1}(0.476)^1 \cdot (0.524)^4 = 17.9433$$

$$f_2' = 100 \cdot p_2 = 100 \cdot \binom{5}{2}(0.476)^2 \cdot (0.524)^3 = 32.6$$

$$f_3' = 100 \cdot p_3 = 100 \cdot \binom{5}{3}(0.476)^3 \cdot (0.524)^2 = 29.613$$
$$f_4' = 100 \cdot p_4 = 100 \cdot \binom{5}{4}(0.476)^4 \cdot (0.524)^1 = 13.45$$
$$f_5' = 100 \cdot p_5 = 100 \cdot \binom{5}{5}(0.476)^5 \cdot (0.524)^0 = 2.4436$$

Uočimo da je teorijska frekvencija prvog i poslijednjeg razreda < 5. Stoga ćemo te razrede spojiti s njima susjednim razredima. Ukoliko bi tako opet dobili razred čija je teorijska frekvencija strogo manja od 5, postupak bi ponovljali dok ne bi dobili razred s (ukupnom) teorijskom frekvencijom > 5. Sada formiramo tablicu:

i	f_i	f_i'	$\frac{(f_i - f_i')^2}{f_i'}$
1	$3 + 16 = \mathbf{19}$	$3.95054 + 17.9433 = \mathbf{21.89384}$	0.3825
2	36	32.6	0.3546
3	32	29.613	0.1924
4	$11 + 2 = \mathbf{13}$	$13.45 + 2.4436 = \mathbf{15.8936}$	0.5268
Σ	100	100	1.4563

Vrijednost test-statistike je dakle

$$h = 1.4563.$$

Konačan broj razreda $k = 4$, a broj procijenjenih parametara $r = 1$. Iz tablice očitavamo

$$\chi^2_\alpha(k - r - 1) = \chi^2_{0.01}(2) = 9.2$$

Kako je

$$h < \chi^2_{0.01}(2),$$

dakle vrijednost test-statistike nije ušla u kritično područje, ne možemo odbaciti nultu hipotezu, odnosno ne možemo zaključiti da se ne radi o binomnoj distribuciji. □

Zadatak 37 *Anketirano je 100 studenata i dobiven je prosječan broj njihovih odlazaka u kazalište tijekom godine. S nivoom signifikantnosti 0.05,*

testirajte hipotezu da se radi o uzorku iz populacije s normalnom distribucijom.

broj posjeta	$[0,2\rangle$	$[2,4\rangle$	$[4,6\rangle$	$[6,8\rangle$	$[8,10\rangle$	$[10,12\rangle$	$[12,14\rangle$
broj studenata	5	10	20	33	18	10	4

Rješenje: Normalna distribucija ima 2 parametra - očekivanje μ i varijancu σ^2. Kako nijedan od njih nije zadan, moramo ih procijeniti, pa odmah slijedi da je $r = 2$. Procjenitelj za očekivanje je $\hat{\mu} = \bar{x}$ a za varijancu $\hat{\sigma}^2 = s_n^2$.

U tablici su dani sortirani podaci. Vidimo da je 5 studenata išlo u kazalište 0 ili 1 put ali ne znamo koliko točno od tih 5 je išlo 0 a koliko 1 put. Treba nam "predstavnik" tog razreda - uzimamo sredinu razreda. Sada

$$\hat{\mu} = \bar{x} = \frac{1 \cdot 5 + 3 \cdot 10 + 5 \cdot 20 + 7 \cdot 33 + 9 \cdot 18 + 11 \cdot 10 + 13 \cdot 4}{100} = 6.9$$

$$\hat{\sigma}^2 = s_n^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) = \frac{1}{n-1}\left(\sum_{i=1}^{k} a_i^2 \cdot f_i - n\bar{x}^2\right)$$

no kako je $n = 100$ velik možemo umjesto s $n-1$ dijeliti s n:

$$\Rightarrow \hat{\sigma}^2 = \frac{1^2 \cdot 5 + 3^2 \cdot 10 + 5^2 \cdot 20 + 7^2 \cdot 33 + 9^2 \cdot 18 + 11^2 \cdot 10 + 13^2 \cdot 4}{100} - 6.9^2 = 7.95$$

Postavljamo (nultu) hipotezu da slučajna varijabla X koja broji odlaske u kazalište ima distribuciju

$$X \sim N(6.9, 7.95)$$

Sljedeći korak je odrediti teorijske frekvencije $f_i' = 100 \cdot p_i$. Imamo

$$\begin{aligned}
p_1 &= P(0 \le X < 2) = P\left(\frac{0-6.9}{\sqrt{7.95}} \le X^* < \frac{2-6.9}{\sqrt{7.95}}\right) \\
&= \Phi_0(-1.74) - \Phi_0(-2.45) = \Phi_0(2.45) - \Phi_0(1.74) \\
&= 0.4928572 - 0.4591 = 0.0338 \quad \Rightarrow \quad f_1' = 3.38 \\
p_2 &= P(2 \le X < 4) = P\left(\frac{2-6.9}{2.82} \le X^* < \frac{4-6.9}{2.82}\right) \\
&= \Phi_0(-1.03) - \Phi_0(-1.74) = \Phi_0(1.74) - \Phi_0(1.03)
\end{aligned}$$

$$
\begin{aligned}
&= 0.4591 - 0.3485 = 0.1106 \quad \Rightarrow \quad f_2' = 11.06 \\
p_3 &= P(4 \le X < 6) = P\left(-1.03 \le X^* < -0.32\right) \\
&= \Phi_0(-0.32) - \Phi_0(-1.03) = 0.223 \quad \Rightarrow \quad f_3' = 22.3 \\
p_4 &= P(6 \le X < 8) = P\left(-0.32 \le X^* < 0.39\right) \\
&= \Phi_0(0.39) - \Phi_0(-0.32) = 0.2772 \quad \Rightarrow \quad f_4' = 27.72 \\
p_5 &= P(8 \le X < 10) = P\left(0.39 \le X^* < 1.10\right) \\
&= \Phi_0(1.10) - \Phi_0(0.39) = 0.2126 \quad \Rightarrow \quad f_5' = 21.26 \\
p_6 &= P(10 \le X < 12) = P\left(1.1 \le X^* < 1.8\right) \\
&= \Phi_0(1.8) - \Phi_0(1.1) = 0.09974 \quad \Rightarrow \quad f_6' = 9.97 \\
p_7 &= P(12 \le X < 14) = P\left(1.8 \le X^* < 2.52\right) \\
&= \Phi_0(2.52) - \Phi_0(1.8) = 0.03006 \quad \Rightarrow \quad f_7' = 3
\end{aligned}
$$

Budući je $f_1' < 5$ i $f_7' < 5$, spojit ćemo prva dva i posljednja dva razreda, pa će tako ostati ukupno 5 razreda. Dakle, $k = 5$. Formiramo tablicu:

i	1	2	3	4	5	Σ
f_i	15	20	33	18	14	100
f_i'	14.44	22.3	27.72	21.26	12.97	
$\frac{(f_i - f_i')^2}{f_i'}$	0.022	0.237	1.006	0.499	0.082	1.846

Vrijednost test-statistike je prema tome

$$h = \sum_{i=1}^{5} \frac{(f_i - f_i')^2}{f_i'} = 1.846,$$

a

$$\chi_\alpha^2(k - r - 1) = \chi_{0.05}^2(2) = 6,$$

pa kako je $h < \chi_{0.05}^2(2)$, nultu hipotezu ne možemo odbaciti, odnosno ne možemo zaključiti da se ne radi o uzorku iz normalno distribuirane populacije. □

Zadatak 38 *(DZ) Bilježen je broj četvorki rođenih u nekoj županiji tijekom 70 godina. Podaci su dani u tablici. Uz razinu značajnosti 0.05, testirajte hipotezu da su podaci uzeti iz populacije s Poissonovom distribucijom.*

broj rođenih četvorki	0	1	2	3	4	5	6
broj godina	14	24	17	10	2	2	1

Napomena: $\hat{\lambda} = \bar{x}$

4.7 χ^2 - test nezavisnosti dviju varijabli

Neka je $(X_1, Y_1),\ (X_2, Y_2),\ \ldots\ (X_n, Y_n)$ slučajni uzorak za dvodimenzionalno diskretno statističko obilježje (X, Y) i neka je pritom:

$$\begin{aligned} &\mathrm{Im}X = \{a_1, \ldots, a_r\} \\ &\mathrm{Im}Y = \{b_1, \ldots, b_s\} \\ &\Rightarrow\ \mathrm{Im}(X,Y) = \{(a_i, b_j) : 1 \le i \le r,\ \ 1 \le j \le s\} \end{aligned}$$

Nadalje,

f_{ij} : frekvencija od (a_i, b_j) u uzorku

f_i : (marginalna) frekvencija od a_i u uzorku

g_j : (marginalna) frekvencija od b_j u uzorku

Vrijedi:

$$f_i = \sum_{j=1}^{s} f_{ij}, \quad g_j = \sum_{i=1}^{r} f_{ij}$$

Kontingencijska frekvencijska tablica:

$X \backslash Y$	b_1	b_2	$\ldots$	b_s	Σ
a_1	f_{11}	f_{12}	$\ldots$	f_{1s}	f_1
a_2	f_{21}	f_{22}	$\ldots$	f_{2s}	f_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a_r	f_{r1}	f_{r2}	$\ldots$	f_{rs}	f_r
Σ	g_1	g_2	$\ldots$	g_s	n

Označimo:

$$p_{ij} = P(X = a_i,\ Y = b_j)$$

$$p_i = P(X = a_i)$$
$$q_j = P(X = b_j)$$

Hipoteze su:

$$H_0 \ : \ p_{ij} = p_i \cdot q_j, \quad \forall\, i, j$$
tj. X i Y su nezavisne slučajne varijable
$$H_1 \ : \ \exists\, i, j \text{ takvi da } p_{ij} \neq p_i \cdot q_j$$

Uz H_0, procjene za p_i i q_j su:

$$\hat{p_i} = \frac{f_i}{n}, \quad \hat{q_j} = \frac{g_j}{n}$$

Očekivane vrijednosti f'_{ij} od f_{ij} uz H_0 su:

$$f'_{ij} = n\, \hat{p_i}\, \hat{q_j} = n \cdot \frac{f_i}{n} \cdot \frac{g_j}{n} = \frac{f_i \cdot g_j}{n}$$

Koristimo test-statistiku

$$H = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(f_{ij} - f'_{ij})^2}{f'_{ij}}$$

Ako je H_0 istinita, tada

$$H \sim \chi^2((r-1)(s-1))$$

Hipotezu o nezavisnosti odbacujemo ako

$$H \geq \chi^2_\alpha((r-1)(s-1))$$

Zadatak 39 *U cilju ispitivanja sklonosti potrošača proizvodu A uzet je uzorak na temelju kojeg su dobiveni podaci dani u tablici. Možete li na osnovu ovih podataka zaključiti da sklonost potrošača proizvodu A NE ovisi o njihovom dohotku, uz razinu značajnosti 0.05?*

mjesečni dohodak	*sklonost potrošnji*		
anketiranih kupaca u kn	*stalno kupuju*	*povremeno kupuju*	*ne kupuju*
-3000	70	17	21
$3000 - 5000$	165	56	28
$5000 - 7000$	195	85	26
$7000-$	170	42	25

Rješenje: Označimo s X slučajnu varijablu koja mjeri visinu dohotka, a s Y onu koja mjeri sklonost potrošnji. Postavljamo hipoteze:

H_0 : X i Y su nezavisne slučajne varijable

H_1 : X i Y su zavisne slučajne varijable

Provest ćemo χ^2-test o nezavisnosti dviju varijabli. Potrebno je izračunati teorijske frekvencije f'_{ij} za $i = 1, 2, 3, 4, \quad j = 1, 2, 3$, no pogledajmo najprije kolike su marginalne frekvencije f_i i g_j:

mjesečni dohodak	stalno kupuju	povremeno kupuju	ne kupuju	$\sum$
-3000	70	17	21	$f_1 = 108$
$3000 - 5000$	165	56	28	$f_2 = 249$
$5000 - 7000$	195	85	26	$f_3 = 306$
$7000-$	170	42	25	$f_4 = 237$
$\sum$	$g_1 = 600$	$g_2 = 200$	$g_3 = 100$	$n = 900$

Sada dobivamo:

$$f'_{11} = \frac{f_1 \cdot g_1}{n} = \frac{108 \cdot 600}{900} = 72 \qquad f'_{31} = \frac{f_3 \cdot g_1}{n} = \frac{306 \cdot 600}{900} = 204$$

$$f'_{12} = \frac{f_1 \cdot g_2}{n} = \frac{108 \cdot 200}{900} = 24 \qquad f'_{32} = \frac{f_3 \cdot g_2}{n} = \frac{306 \cdot 200}{900} = 68$$

$$f'_{13} = \frac{f_1 \cdot g_3}{n} = \frac{108 \cdot 100}{900} = 12 \qquad f'_{33} = \frac{f_3 \cdot g_3}{n} = \frac{306 \cdot 100}{900} = 34$$

$$f'_{21} = \frac{f_2 \cdot g_1}{n} = \frac{249 \cdot 600}{900} = 166 \qquad f'_{41} = \frac{f_4 \cdot g_1}{n} = \frac{237 \cdot 600}{900} = 158$$

$$f'_{22} = \frac{f_2 \cdot g_2}{n} = \frac{249 \cdot 200}{900} = 55.3 \qquad f'_{42} = \frac{f_4 \cdot g_2}{n} = \frac{237 \cdot 200}{900} = 52.67$$

$$f'_{23} = \frac{f_2 \cdot g_3}{n} = \frac{249 \cdot 100}{900} = 27.67 \qquad f'_{43} = \frac{f_4 \cdot g_3}{n} = \frac{237 \cdot 100}{900} = 26.3$$

Da bismo lakše izračunali vrijednost test-statistike, zgodno je, radi preglednosti, u tablici eksperimentalnim frekvencijama pridružiti odgovarajuće teorijske:

mjesečni dohodak	stalno kupuju	povremeno kupuju	ne kupuju
-3000	$70/\mathit{72}$	$17/\mathit{24}$	$21/\mathit{12}$
$3000-5000$	$165/\mathit{166}$	$56/\mathit{55.3}$	$28/\mathit{27.67}$
$5000-7000$	$195/\mathit{204}$	$85/\mathit{68}$	$26/\mathit{34}$
$7000-$	$170/\mathit{158}$	$42/\mathit{52.67}$	$25/\mathit{26.3}$

Preostalo je izračunati vrijednost test-statistike:

$$h = \sum_{i=1}^{4}\sum_{j=1}^{3}\frac{(f_{ij}-f'_{ij})^2}{f'_{ij}} = 18.532$$

Iz tablice očitavamo:

$$\chi^2_\alpha((r-1)(s-1)) = \chi^2_{0.05}((4-1)(3-1)) = \chi^2_{0.05}(6) = 12.6,$$

pa kako je

$$h > \chi^2_{0.05}(6)$$

vidimo da je vrijednost test-statistike upala u kritično područje. Nultu hipotezu o nezavisnosti stoga odbacujemo i zaključujemo da su visina mjesečnog dohotka (slučajna varijabla X) i sklonost potrošnji (slučajna varijabla Y) međusobno zavisne. □

4.8 χ^2 - test homogenosti populacija

- zanima nas razdioba istog diskretnog statističkog obilježja u raznim populacijama
- na osnovi nezavisnih uzoraka uzetih iz tih populacija, testiramo osnovnu hipotezu da su razdiobe od X u tim populacijama jednake, tj. da su populacije *homogene* obzirom na X

- m : broj populacija koje promatramo
  $X^{(i)}$: slučajna varijabla koja predstavlja X u i-toj populaciji ($i = 1, \ldots, m$); vrijedi

$$X^{(i)} \sim \begin{pmatrix} a_1 & a_2 & \ldots & a_k \\ p_1^{(i)} & p_2^{(i)} & \ldots & p_k^{(i)} \end{pmatrix}$$

- nulta hipoteza je da su sve $X^{(i)}$ jednake po distribuciji, a alternativna je da postoji bar jedna koja se po distribuciji razlikuje od ostalih, odnosno:

$$\begin{aligned} H_0 &: \; X^{(1)} \overset{D}{=} X^{(2)} \overset{D}{=} \ldots \overset{D}{=} X^{(m)} \\ H_1 &: \; \exists\, i, j \quad \text{tako da} \quad X^{(i)} \overset{D}{\neq} X^{(j)} \end{aligned}$$

- H_0 se može zapisati i ovako

$$H_0 \; : \; p_j^{(i)} = p_j, \quad j = 1, \ldots, k, \quad i = 1, \ldots, m$$

  gdje p_j predstavljaju zajedničke (tj. po populacijama jednake) vjerojatnosti od a_j

Frekvencijska tablica:

X	a_1	a_2	$\ldots$	a_k	$\sum$
populacija 1	f_{11}	f_{12}	$\ldots$	f_{1k}	n_1
populacija 2	f_{21}	f_{22}	$\ldots$	f_{2k}	n_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
populacija m	f_{m1}	f_{m2}	$\ldots$	f_{mk}	n_m
$\sum$	f_1	f_2	$\ldots$	f_k	n

- n_i : duljina uzroka iz i-te populacije,
  f_{ij} : frekvencija od a_j u uzorku iz i-te populacije
  $f_j = \sum_{i=1}^{m} f_{ij}$: frekvencija od a_j u svim uzorcima zajedno

- vrijedi: $n_i = \sum_{j=1}^{k} f_{ij}$
- procjena zajedničkih vrijednosti p_j ako vrijedi H_0:

$$\hat{p}_j = \frac{f_j}{n}, \quad j = 1, \ldots, k$$

- očekivane frekvencije (ako vrijedi H_0):

$$f'_{ij} = n_i \cdot \hat{p}_j = \frac{n_i \cdot f_j}{n}$$

- koristimo test-statistiku:

$$H = \sum_{i=1}^{m} \sum_{j=1}^{k} \frac{(f_{ij} - f'_{ij})^2}{f'_{ij}}$$

  Ako je H_0 istinita, tada

$$H \sim \chi^2((m-1)(k-1))$$

- hipotezu o homogenosti populacija odbacujemo ako

$$H \geq \chi^2_\alpha((m-1)(k-1))$$

Zadatak 40 *U tvorničkom pogonu proizvode se televizori. Svakog radnog dana u tjednu registrira se broj neispravnih televizora. Provedena su opažanja tijekom 750 dana i rezultati su prikazani u tablici. Može li se, uz razinu značajnosti 0.05, zaključiti da nema značajne razlike u pojavi neispravnih televizora tijekom tjedna?*

broj neispravnih televizora	*PON*	*UTO*	*SRI*	*ČET*	*PET*
$0-2$	60	63	62	68	51
$3-5$	70	61	60	52	70
$6->$	20	26	28	30	29

Rješenje: Neka je X broj neispravnih televizora po danu. Ako dane u tjednu shvatimo kao 5 različitih populacija (iz kojih su uzeti uzorci), tada je potrebno provjeriti ima li X jednaku distribuciju u svih tih 5 populacija, odnosno dana. To ćemo provjeriti χ^2-testom o homogenosti populacija. Hipoteze su dakle:

H_0 : podaci iz svih 5 populacija potječu iz iste vjerojatnosne razdiobe, tj. $X^{(1)} \stackrel{D}{=} X^{(2)} \stackrel{D}{=} \ldots \stackrel{D}{=} X^{(5)}$

H_1 : ne potječu iz iste razdiobe

Da bismo izračunali vrijednost odgovarajuće test-statistike, potrebne su nam procjene frekvencija f'_{ij}, pa najprije pogledajmo kolike su duljine uzoraka n_i iz svake od populacija ($i = 1, 2, 3, 4, 5$) i kumulativne frekvencije f_j svake od mogućih vrijednosti koje X poprima ($j = 1, 2, 3$):

broj neispr.tv	$0-2$	$3-5$	$6->$	$\sum$
PON	60	70	20	$n_1 = 150$
UTO	63	61	26	$n_2 = 150$
SRI	62	60	28	$n_3 = 150$
ČET	68	52	30	$n_4 = 150$
PET	51	70	29	$n_5 = 150$
$\sum$	$f_1 = 304$	$f_2 = 313$	$f_3 = 133$	$n = 750$

Sada:

$$f'_{11} = \frac{n_1 \cdot f_1}{n} = \frac{150 \cdot 304}{750} = 60.8$$

Kako je $n_1 = n_2 = n_3 = n_4 = n_5 = 150$, to je

$$f'_{21} = \frac{n_2 \cdot f_1}{n} = f'_{11}, \quad f'_{31} = \frac{n_3 \cdot f_1}{n} = f'_{11}, \quad f'_{41} = \frac{n_4 \cdot f_1}{n} = f'_{11}, \quad f'_{51} = \frac{n_5 \cdot f_1}{n} = f'_{11},$$

pa je

$$f'_{11} = f'_{21} = f'_{31} = f'_{41} = f'_{51} = 60.8$$

Slično,

$$f'_{12} = \frac{n_1 \cdot f_2}{n} = \frac{150 \cdot 313}{750} = 62.6 = f'_{22} = f'_{32} = f'_{42} = f'_{52}$$
$$f'_{13} = \frac{n_1 \cdot f_3}{n} = \frac{150 \cdot 133}{750} = 26.6 = f'_{23} = f'_{33} = f'_{43} = f'_{53}$$

Vrijednost test-statistike je:

$$H = \sum_{i=1}^{5} \sum_{j=1}^{3} \frac{(f_{ij} - f'_{ij})^2}{f'_{ij}} = 8.615$$

Iz tablice za χ^2-razdiobu očitavamo

$$\chi^2_\alpha((m-1)(k-1)) = \chi^2_{0.05}(4 \cdot 2) = \chi^2_{0.05}(8) = 15.5$$

Kako je

$$h < \chi^2_{0.05}(8),$$

vidimo da vrijednost test-statistike nije upala u kritično područje pa nultu hipotezu ne možemo odbaciti. Dakle, ne možemo zaključiti da su populacije nisu homogene što znači da promatrano statističko obilježje (= broj pokvarenih televizora) ima jednaku distribuciju u svim populacijama (= u svim danima). □

4.9 Usporedba očekivanja više normalno distribuiranih populacija (jednofaktorska analiza varijance ANOVA)

- ANOVA-u koristimo za usporedbu *više od dvije* normalno distribuirane populacije (za usporedbu *točno dvije* normalno distribuirane populacije koristimo `t-test`!)

- neka su

$$\begin{array}{lll} X_{11},\ X_{12},\ldots,X_{1n_1} & \text{za} & X^{(1)} \sim N(\mu_1, \sigma^2) \\ X_{21},\ X_{22},\ldots,X_{2n_2} & \text{za} & X^{(2)} \sim N(\mu_2, \sigma^2) \\ \vdots & & \vdots \\ X_{k1},\ X_{k2},\ldots,X_{kn_k} & \text{za} & X^{(k)} \sim N(\mu_k, \sigma^2) \end{array}$$

  k nezavisnih slučajnih uzoraka, svaki za normalno distribuirano obilježje X reprezentirano s $X^{(i)}$ za i-tu populaciju iz koje je uzet uzorak duljine n_i $(i = 1, 2, \ldots, k)$

- pretpostavljamo da su varijance od $X^{(i)}$ jednake (u svim populacijama)
- želimo testirati nultu hipotezu

$$H_0 \ : \ \mu_1 = \mu_2 = \ldots = \mu_k,$$

tj. hipotezu da *nema razlike* u očekivanjima među populacijama; alternativna hipoteza je onda naravno da razlika postoji, odnosno da se bar dvije populacije razlikuju po očekivanjima

- za test-statistiku treba nam sljedeće, za $i = 1, 2, \ldots k$:

$$\bar{X}_i = \frac{1}{n_i}(X_{i1} + \ldots + X_{in_i})$$
$$S_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2$$

- ukupna aritmetička sredina svih podataka:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij} = \frac{1}{n}\sum_{i=1}^{k} n_i\bar{X}_i, \quad n = \sum_{i=1}^{k} n_i$$

- suma kvadrata odstupanja srednjih vrijednosti uzoraka od ukupne sredine (= suma kvadrata u odnosu na tretman)

$$SST = \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2 = \sum_{i=1}^{k} n_i\bar{X}_i^2 - n\bar{X}^2$$

- suma kvadrata pogrešaka

$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2 = \sum_{i=1}^{k}(n_i - 1)S_i^2$$
$$= \sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}^2 - \sum_{i=1}^{k} n_i\bar{X}_i^2$$

- srednjekvadratno odstupanje među uzorcima (zbog razlike u tretmanima)

$$MST = \frac{SST}{k-1}$$

- srednjekvadratna pogreška

$$MSE = \frac{SSE}{n-k}$$

- konačno, test-statistika je

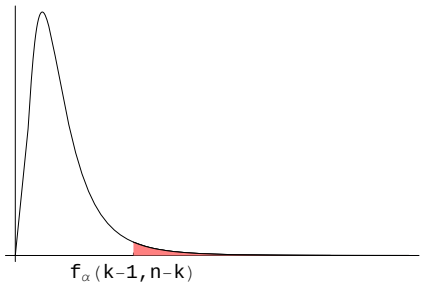
$$F = \frac{MST}{MSE}$$

  Ako je H_0 istinita, tada je

$$F \sim F(k-1, n-k)$$

- nultu hipotezu odbacujemo ako

$$F \geq f_\alpha(k-1, n-k)$$



ANOVA tablica:

izvor rasipanja	stupnjevi slobode	suma kvadrata	srednjekvadratno odstupanje	vrijednost test-statistike
zbog razlike među tretmanima	$k-1$	SST	MST	F
zbog greške	$n-k$	SSE	MSE	
$\sum$	$n-1$	SS		

pritom je

$$SS = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2$$

Zadatak 41 *Pivovara koristi 3 različite linije punjenja limenki piva. Sumnja se da se srednji neto sadržaj limenki razlikuje od linije do linije. Na slučajan način bira se 5 limenki sa svake linije i mjeri se njihov neto sadržaj. Testirajte postoji li značajna razlika između sredina neto sadržaja po linijama uz razinu značajnosti 0.05.*

linija	*sadržaj*	*u*	*dcl*		
1	3.633	3.651	3.66	3.645	3.654
2	3.615	3.627	3.636	3.63	3.624
3	3.645	3.63	3.627	3.63	3.633

Rješenje: Potrebno je provjeriti postoji li razlika između sredina neto sadržaja po linijama. Budući imamo 3 populacije (=linije), t-test nam ne može pomoći, već moramo provesti ANOVA-u. Krenimo redom:

$$k = 3, \quad n_1 = n_2 = n_3 = 5, \quad n = \sum_{i=1}^{3} n_i = 15$$

$$\bar{x}_1 = \frac{3.633 + 3.651 + 3.66 + 3.645 + 3.654}{5} = 3.6486$$

$$\bar{x}_2 = \frac{3.615 + 3.627 + 3.636 + 3.63 + 3.624}{5} = 3.6264$$

$$\bar{x}_3 = 3.633$$

$$\bar{x} = \frac{1}{15}\sum_{i=1}^{3}\sum_{j=1}^{5} x_{ij} = \frac{1}{15}\sum_{i=1}^{3} n_i \cdot \bar{x}_i = \frac{1}{3}\sum_{i=1}^{3} \bar{x}_i = 3.636$$

$$SST = \sum_{i=1}^{3} n_i \bar{X}_i^2 - n\bar{X}^2 = 5\sum_{i=1}^{3} \bar{x}_i^2 - 15\bar{x}^2 = 0.0013$$

$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}^2 - \sum_{i=1}^{k} n_i \bar{X}_i^2 = \sum_{i=1}^{3}\sum_{j=1}^{5} x_{ij}^2 - 5\sum_{i=1}^{3} \bar{x}_i^2 = 0.00086$$

$$MST = \frac{SST}{k-1} = \frac{0.0013}{2} = 0.00065$$

$$MSE = \frac{SSE}{n-k} = \frac{0.00086}{15-3} = 0.000072$$

i konačno dobivamo vrijednost test-statistike:

$$\Rightarrow \quad f = \frac{MST}{MSE} = \frac{0.00065}{0.000072} = 9.02778$$

Iz tablice za F-razdiobu potrebno je očitati:

$$f_\alpha(k-1, n-k) = f_{0.05}(2, 12) = 3.89$$

Kako je

$$f > f_{0.05}(2, 12)$$

vidimo da je vrijednost test-statistike upala u kritično područje što znači da nultu hipotezu o jednakosti očekivanja moramo odbaciti. Zaključujemo stoga da `postoji` značajna razlika među sredinama neto sadržaja po linijama.

ANOVA tablica:

izvor rasipanja	stupnjevi slobode	suma kvadrata	srednjekvadratno odstupanje	vrijednost test-statistike
zbog tretmana	2	0.0013	0.00065	
zbog greške	12	0.00086	0.000072	9.02778
$\sum$	14	0.00216		

□

4.10 Test koreliranosti dviju varijabli

- neka je

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

  slučajni uzorak za normalno distribuirani slučajni vektor (X, Y)

- $\bar{X},\ \bar{Y}$: aritmetičke sredine uzoraka

- $S_x^2,\ S_y^2$: uzoračke varijance

- kovarijanca od X i Y:

  $$S_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$$

  Vrijedi:

  $$\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^{n} X_i Y_i - \bar{X} \cdot \sum_{i=1}^{n} Y_i - \bar{Y} \cdot \sum_{i=1}^{n} X_i + n\bar{X}\bar{Y}$$

  $$= \sum_{i=1}^{n} X_i Y_i - \bar{X} \cdot n\bar{Y} - \bar{Y} \cdot n\bar{X} + n\bar{X}\bar{Y} = \sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}$$

pa onda

$$S_{xy} = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}\right)$$

- želimo testirati nultu hipotezu

$$H_0 \ : \ \rho = 0 \quad (= \text{nema korelacije})$$

u odnosu na jednostranu alternativu

$$H_1 \ : \ \rho > 0 \quad (= \text{korelacija postoji i pozitivna je})$$

ili

$$H_1 \ : \ \rho < 0 \quad (= \text{korelacija postoji i negativna je})$$

ili u odnosu na dvostranu alternativu

$$H_1 \ : \ \rho \neq 0 \quad (= \text{korelacija postoji})$$

- *Pearsonov koeficijent korelacije* je statistika

$$R = \frac{S_{xy}}{S_x \cdot S_y}$$

- test-statistika je:

$$Z = \frac{R}{\sqrt{1-R^2}} \cdot \sqrt{n-2}$$

Ako je H_0 istinita, tada

$$Z \sim t(n-2)$$

1.

$$H_0 : \rho = 0$$
$$H_1 : \rho \neq 0$$

Nultu hipotezu H_0 odbacujemo ako je

$$Z > t_{\frac{\alpha}{2}}(n-2) \qquad \text{ili} \qquad Z < -t_{\frac{\alpha}{2}}(n-2)$$

2.

$$H_0 : \rho = 0$$
$$H_1 : \rho > 0$$

H_0 odbacujemo ako je

$$Z > t_\alpha(n-2)$$

3.

$$H_0 : \rho = 0$$
$$H_1 : \rho < 0$$

H_0 odbacujemo ako je

$$Z < -t_\alpha(n-2)$$

Zadatak 42 *U jednom razredu od 30 učenika promatra se ocjena iz matematike (X) i ocjena iz fizike (Y). Uvidom u imenik dobiveni su ovi podaci:* $(1,3),(4,3),(2,2),(3,2),(1,2),(1,1),(2,2),(4,4),(2,2),(3,3),(4,4),(5,5),$ $(3,5),(2,1),(2,3),(2,2),(5,5),(3,3),(2,2),(2,2),(3,3),(3,2),(4,4),(2,2),$ $(3,3),(2,1),(3,2),(3,2),(3,2),(2,2).$
Uz razinu značajnosti 0.05, testirajte hipotezu da nema značajne korelacije između ocjena iz matematike i fizike.

Rješenje: Zanima nas postoji li korelacija između ocjena iz matematike i fizike. To ćemo ispitati pomoću testa o koreliranosti dviju varijabli - X (koja označava ocjene iz matematike) i Y (koja označava ocjene iz fizike). Budući nas zanima samo postoji li korelacije ili ne, a ne da li je (ako postoji) ona pozitivna ili negativna, dovoljno je za alternativnu hipotezu H_1 postaviti $\rho \neq 0$. Dakle,

$$\begin{aligned} H_0 &: \rho = 0 \\ H_1 &: \rho \neq 0 \end{aligned}$$

Izračunajmo sada vrijednost odgovarajuće test-statistike:

$$\bar{x} = \frac{1}{30}(1+4+2+3+1+1+2+4+2+3+\ldots+3+2) = 2.7$$
$$\bar{y} = \frac{1}{30}(3+3+2+2+1+2+4+2+3+4+\ldots+2+2) = 2.63$$
$$s_x^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) = \frac{1}{29}(251 - 30\cdot 2.7^2) = 1.114 \quad \Rightarrow \quad s_x = 1.056$$
$$s_y^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right) = \frac{1}{29}(245 - 30\cdot 2.63^2) = 1.293 \quad \Rightarrow \quad s_y = 1.137$$
$$s_{xy} = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right) = \frac{1}{29}(239 - 30\cdot 2.7\cdot 2.63) = 0.896$$
$$\Rightarrow \quad r = \frac{s_{xy}}{s_x \cdot s_y} = \frac{0.896}{1.056\cdot 1.137} = 0.746$$

Vrijednost test-statistike je

$$z = \frac{r}{\sqrt{1-r^2}} \cdot \sqrt{n-2} = \frac{0.746}{\sqrt{1-0.746^2}} \cdot \sqrt{28} = 5.927$$

Iz tablice očitavamo

$$t_{\frac{\alpha}{2}}(n-2) = t_{0.025}(28) = 2.048$$

Kako je

$$z > t_{0.025}(28)$$

vidimo da je vrijednost test-statistike upala u kritično područje, pa nultu hipotezu odbacujemo. Zaključujemo stoga da korelacija između ocjena iz matematike i fizike *postoji*, odnosno da su varijable X i Y korelirane. □

5 Linearni regresijski model

Imamo n parova podataka

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

koji su dobiveni mjerenjem (opažanjem) nekog dvodimenzionalnog numeričkog statističkog obilježja (X, Y) promatrane populacije. Nezavisna varijabla X interpretira se kao neslučajna a zavisna varijabla Y kao slučajna. Da bi se to naglasilo, X se najčešće zapisuje kao "malo" x. Želimo odrediti linearnu vezu između x i Y:

$$Y = \alpha\, x + \beta + \varepsilon,$$

pri čemu su $\alpha,\ \beta$ parametri, x je broj (neslučajna varijabla), a ε slučajna varijabla za koju vrijedi $\mathrm{E}[\varepsilon] = 0$ i koja se najčešće interpretira kao slučajna greška ili šum.

- procjenitelji od (α, β) dobiveni metodom najmanjih kvadrata:

$$\hat{\alpha} := \frac{S_{xy}}{S_x^2}$$
$$\hat{\beta} := \bar{y} - \hat{\alpha}\, \bar{x}$$

- procjenitelj za varijancu σ^2 je:

$$\hat{\sigma}^2 = \frac{SSE}{n-2}$$

  pri čemu je

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta} - \hat{\alpha}\, x_i)^2 = (n-1) \cdot (S_y^2 - \hat{\alpha}^2 S_x^2)$$

  i dakle vrijedi: $\quad \hat{Y}_i = \hat{\alpha}\, x_i + \hat{\beta}$

- $(1-\alpha) \cdot 100\%$ **pouzdan interval za** α:

$$\hat{\alpha} - t_{\frac{\alpha}{2}}(n-2) \cdot \frac{\hat{\sigma}}{\sqrt{(n-1)S_x^2}} \ \leq\ \alpha\ \leq\ \hat{\alpha} + t_{\frac{\alpha}{2}}(n-2) \cdot \frac{\hat{\sigma}}{\sqrt{(n-1)S_x^2}}$$

- $(1-\alpha)\cdot 100\%$ **pouzdan interval za** β:

$$\hat{\beta} - t_{\frac{\alpha}{2}}(n-2)\cdot\hat{\sigma}\,\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{(n-1)S_x^2}} \;\leq\; \beta \;\leq\; \hat{\beta} + t_{\frac{\alpha}{2}}(n-2)\cdot\hat{\sigma}\,\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{(n-1)S_x^2}}$$

- test-statistike za testiranje sljedećih nul-hipoteza:

  1. $H_0 \;:\; \alpha = \alpha_0 \;\; (\alpha_0 \in \mathbb{R})$ (u odnosu na razne alternative):

$$T_\alpha = \frac{\hat{\alpha}-\alpha_0}{\hat{\sigma}}\,\sqrt{(n-1)S_x^2}$$

  Ako je H_0 istinita tada je

$$T_\alpha \sim t(n-2)$$

  2. $H_0 \;:\; \beta = \beta_0 \;\; (\beta_0 \in \mathbb{R})$ (u odnosu na razne alternative):

$$T_\beta = \frac{\hat{\beta}-\beta_0}{\hat{\sigma}\cdot\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{(n-1)S_x^2}}}$$

  Ako je H_0 istinita tada je

$$T_\beta \sim t(n-2)$$

Zadatak 43 *Izabrano je 5 osoba starih 35, 45, 55, 65 i 75 godina (x), kojima je izmjeren krvni tlak (Y), pri čemu su dobiveni podaci: 114, 124, 143, 158, 166 redom. Odredite:*

a) procjenu pravca regresije za ove podatke

b) 95% *pouzdane intervale za* α *i* β

c) testirajte hipotezu da je koeficijent smjera tog pravca jednak 0, tj. da između x *i* Y *ne postoji linearna veza, uz razinu značajnosti 0.01.*

Rješenje:

a) izračunajmo procjenu parametara α i β : $\hat{\alpha} = \dfrac{S_{xy}}{S_x^2}, \quad \hat{\beta} = \bar{Y} - \hat{\alpha}\,\bar{x}$

$$\bar{x} = \frac{35+45+55+65+75}{5} = 55$$
$$\bar{y} = \frac{114+124+143+158+166}{5} = 141$$
$$s_x^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\cdot\bar{x}^2\right) = \frac{1}{4}\left(16125 - 5\cdot 55^2\right) = 250$$
$$s_{xy} = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\cdot\bar{x}\bar{y}\right) = \frac{1}{4}\left(40155 - 5\cdot 55\cdot 141\right) = 345$$
$$\Rightarrow \quad \hat{\alpha} = \frac{345}{250} = 1.38$$
$$\Rightarrow \quad \hat{\beta} = \bar{y} - \hat{\alpha}\,\bar{x} = 141 - 1.38\cdot 55 = 65.1$$
$$\Longrightarrow y = 1.38x + 65.1 \text{ je procjena pravca regresije za ove podatke}$$

b) Zanimaju nas pouzdani intervali za α i β. Najprije moramo izračunati $\hat{\sigma}^2$:

$$\hat{\sigma}^2 = \frac{SSE}{n-2}, \qquad SSE = \sum_{i=1}^{n}(Y_i - \hat{Y_i})^2$$

Znamo da je $\hat{Y_i} = \hat{\alpha}\,x_i + \hat{\beta}$ pa onda:

$$\hat{y_1} = \hat{\alpha}\,x_1 + \hat{\beta} = 1.38\cdot 35 + 65.1 = 113.4$$
$$\hat{y_2} = \hat{\alpha}\,x_2 + \hat{\beta} = 1.38\cdot 45 + 65.1 = 127.2$$
$$\hat{y_3} = \hat{\alpha}\,x_3 + \hat{\beta} = 1.38\cdot 55 + 65.1 = 141$$
$$\hat{y_4} = \hat{\alpha}\,x_4 + \hat{\beta} = 1.38\cdot 65 + 65.1 = 154.8$$
$$\hat{y_5} = \hat{\alpha}\,x_5 + \hat{\beta} = 1.38\cdot 75 + 65.1 = 168.6$$

Formirajmo tablicu:

i	1	2	3	4	5	$\sum$
x_i	35	45	55	65	75	
y_i	114	124	143	158	166	
$\hat{y}_i$	113.4	127.2	141	154.8	168.6	
$(y_i - \hat{y}_i)^2$	0.36	10.24	4	10.24	6.76	31.6

Dobili smo: $SSE = 31.6$ pa je onda

$$\hat{\sigma}^2 = \frac{31.6}{3} = 10.5\dot{3} \quad \Rightarrow \quad \sigma = 3.246$$

Pogledajmo sada kako izgleda 95% pouzdan interval za α, odnosno β:

$$\hat{\alpha} \pm t_{\frac{\alpha}{2}}(n-2) \cdot \frac{\hat{\sigma}}{\sqrt{(n-1)s_x^2}} = 1.38 \pm t_{0.025}(3) \cdot \frac{3.246}{\sqrt{4 \cdot 250}}$$
$$= 1.38 \pm 3.182 \cdot 0.103 = 1.38 \pm 0.33$$
$$\Longrightarrow \quad 1.05 \ \leq \ \alpha \ \leq \ 1.71$$

$$\hat{\beta} \ \pm \ t_{\frac{\alpha}{2}}(n-2) \cdot \hat{\sigma} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}} = 65.1 \ \pm \ t_{0.025}(3) \cdot 3.246 \sqrt{\frac{1}{5} + \frac{55^2}{1000}}$$
$$= 65.1 \pm 18.55$$
$$\Longrightarrow \quad 46.55 \ \leq \ \beta \ \leq \ 83.65$$

c) Želimo, uz razinu značajnosti 0.01, testirati hipotezu da ne postoji linearna veza između x i Y. Linearna veza ne postoji jedino ako je koeficijent smjera pravca regresije jednak 0. Ako je on različit od 0, bez obzira da li je pozitivan (tj. > 0) ili negativan (tj. < 0), linearna veza postoji. Postavljamo stoga hipoteze:

$$H_0 \ : \ \alpha = 0$$
$$H_1 \ : \ \alpha \neq 0$$

Sljedeći korak je izračunati vrijednost odgovarajuće test-statistike:

$$T_\alpha = \frac{\hat{\alpha} - \alpha_0}{\hat{\sigma}} \sqrt{(n-1)S_x^2} \ \sim \ t(n-2)$$

Imamo:

$$t_\alpha = \frac{1.38 - 0}{3.246} \sqrt{1000} = 13.44$$

Iz tablice za t-razdiobu očitavamo

$$t_{\frac{\alpha}{2}}(n-2) = t_{0.005}(3) = 5.841$$

Kako je

$$t_\alpha > t_{0.005}(3)$$

vrijednost test-statistike je upala u kritično područje, pa nultu hipotezu $H_0 : \alpha = 0$ moramo odbaciti. Zaključujemo stoga da koeficijent smjera pravca regresije nije jednak 0, pa onda linearna veza postoji. □

Linearni model najčešće se koristi u dvije svrhe:

1. za predviđanje (procjenu) vrijednosti *srednje tj. očekivane vrijednosti od* Y za neku danu vrijednost x_0 od x, tj. $\mathrm{E}[Y|x = x_0]$. U ovom slučaju, nastoji se procijeniti *srednja vrijednost* mjerenja *velikog broja pokusa* pri zadanoj vrijednosti od x.

   - procjenitelj od $\mathrm{E}[Y|x = x_0]$ je

   $$\widehat{\mathrm{E}[Y|x = x_0]} = \hat{\alpha}\, x_0 + \hat{\beta}$$

   - $(1-\alpha)$ 100% pouzdan interval za $\mathrm{E}[Y|x = x_0]$:

   $$\left[\widehat{\mathrm{E}[Y|x = x_0]} \;-\; t_{\frac{\alpha}{2}}(n-2) \cdot \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)S_x^2}} \;, \right. \tag{14}$$
   $$\left. \widehat{\mathrm{E}[Y|x = x_0]} \;+\; t_{\frac{\alpha}{2}}(n-2) \cdot \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)S_x^2}} \right]$$

2. za predviđanje (procjenu) *vrijednosti* Y za neku danu vrijednost x_0 od x. U ovom slučaju, nastoji se procijeniti *rezultat jednog pokusa* provedenog pri zadanoj vrijednosti od x, dakle rezultat nekog budućeg mjerenja.

   - procjenitelj od Y za $x = x_0$ je

   $$\hat{Y} = \hat{\alpha}\, x_0 + \hat{\beta}$$

   - $(1-\alpha)$ 100% pouzdan interval za Y u $x = x_0$:

   $$\left[\hat{Y} \;-\; t_{\frac{\alpha}{2}}(n-2) \cdot \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)S_x^2}} \;, \right. \tag{15}$$
   $$\left. \hat{Y} \;+\; t_{\frac{\alpha}{2}}(n-2) \cdot \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)S_x^2}} \right]$$

Uočimo da je pouzdani interval (15) za Y širi, odnosno manje precizan od pouzdanog intervala (14) za $\mathrm{E}[Y|x = x_0]$, što je bilo prirodno za očekivati.

Zadatak 44 *Nađite* 95% *pouzdan interval za* Y *u* $x = 55$*, te* 95% *pouzdan interval za* $\mathrm{E}[Y|x = 55]$ *za podatke iz Zadatka 43.*

Rješenje: Pouzdane intervale za Y u $x = 55$ i $\mathrm{E}[Y|x = 55]$ dobit ćemo uvrštavanjem odgovarajućih vrijednosti u (15) i (14), redom. Većina parametara već je izračunata, treba nam još samo:

$$\hat{Y} = \mathrm{E}[\widehat{Y|x = 55}] = \hat{\alpha} \cdot 55 + \hat{\beta} = 1.38 \cdot 55 + 65.1 = 141$$

Sada:

$$\begin{aligned}
&\mathrm{E}[\widehat{Y|x = x_0}] \ \pm \ t_{\frac{\alpha}{2}}(n-2) \cdot \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)S_x^2}} \\
&= \mathrm{E}[\widehat{Y|x = 55}] \ \pm \ t_{0.025}(3) \cdot 3.246 \sqrt{\frac{1}{5} + \frac{(55-55)^2}{4 \cdot 250}} = 141 \pm 4.62
\end{aligned}$$

pa slijedi da je 95% pouzdan interval za $\mathrm{E}[Y|x = 55]$:

$$136.38 \ \leq \ \mathrm{E}[Y|x = 55] \ \leq \ 145.62$$

Slično dobivamo:

$$\begin{aligned}
&\hat{Y} \ \pm \ t_{\frac{\alpha}{2}}(n-2) \cdot \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)S_x^2}} \\
&= 141 \ \pm \ t_{0.025}(3) \cdot 3.246 \sqrt{1 + \frac{1}{5} + \frac{(55-55)^2}{4 \cdot 250}} = 141 \pm 11.3146
\end{aligned}$$

pa je 95% pouzdan interval za procjenu (predviđanje) vrijednosti Y u $x = 55$:

$$129.685 \ \leq \ Y \ \leq \ 152.315$$

□

Pokazatelji da li je predloženi linearni model dobar (prihvatljiv) model za dane podatke:

- **koeficijent determinacije**

$$R^2 := \frac{(n-1)S_y^2 - SSE}{(n-1)S_y^2} = 1 - \frac{SSE}{(n-1)S_y^2} \ \in \ [0, 1]$$

- što je R^2 bliže vrijednosti 1, to je prilagodba linearnog modela podacima bolja

- koeficijent determinacije jednak je kvadratu koeficijenta korelacije

• **test značajnosti linearnog modela**

- svodi se na testiranje

$$H_0 \ : \ \alpha = 0$$
$$H_1 \ : \ \alpha \neq 0$$

Zadatak 45 *Izračunajte koeficijent determinacije za podatke iz Zadatka 43.*

Rješenje:

Znamo da je: $SSE = 31.6$

Treba nam još:

$$(n-1) \cdot s_y^2 = \sum_{i=1}^{n} y_i^2 - n \cdot \bar{y}^2 = 101341 - 5 \cdot 141^2 = 1936$$

$$\Rightarrow \ R^2 = 1 - \frac{SSE}{(n-1)S_y^2} = 1 - \frac{31.6}{1936} = 0.984$$

Linearni model je dakle za ove podatke jako dobar.

□