(BIO)STATISTIKA

skripta

studij: *Prehrambena tehnologija i Biotehnologija*

doc. dr. sc. Iva Franjić

2012.

Sadržaj

Poglavlje 1

DESKRIPTIVNA STATISTIKA

Deskriptivna ili opisna statistika je dio statistike koji obuhvaća jednostavne metode obrade podataka pomoću grafičkih prikaza te nekih jednostavnih numeričkih pokazatelja.

Kada nas zanima neka pojava ili veličina, izvodimo pokus čiji rezultati su mjerenja koja bi nam mogla biti od pomoći i dati uvid u vrijednosti te pojave, odnosno veličine. Veličinu koju promatramo nazivamo **(statističko) obilježje** i označavamo s X. Rezultat mjerenja statističkog obilježja X je (realan) broj x. Ponavljanjem pokusa n puta, dobivamo višestruka mjerenja istog statističkog obilježja $x_1, x_2, \ldots, x_n$, koja nazivamo **(statistički) podaci**.

1.1 Grafički prikaz podataka

Primjer 1. *Neka X označava broj dobiven bacanjem igrače kocke. Kocku smo bacali 20 puta, te dobili sljedeće podatke:*

$$1, 3, 1, 6, 2, 6, 4, 6, 3, 3, 4, 3, 1, 4, 4, 1, 4, 5, 3, 5.$$

- Statističko obilježje koje promatramo u ovom primjeru je *broj na kocki*. Označimo ga s X. Dakle, X="broj na kocki"
- Skup svih vrijednosti koje X može poprimiti je

  $$\mathrm{Im}X = \{1, 2, 3, 4, 5, 6\}.$$

  Budući je taj skup diskretan, odnosno konačan, kažemo da je X **diskretno obilježje**.
- Općenito, obilježje može biti **numeričko** ili **nenumeričko**. Nenumeričko obilježje nazivamo i **kategorija**. Klasični primjeri su npr. spol i boja. Takvim podacima možemo dodijeliti neku numeričku vrijednost (recimo muški spol = 0, ženski spol = 1), no tada naravno nema smisla računati npr. aritmetičku sredinu podataka.
- Primjetimo da se u našem primjeru pojedine vrijednosti pojavljuju više puta. Svakom elementu skupa vrijednosti koje diskretno obilježje X može poprimiti - označimo ih s a_i - možemo pridružiti njegovu **frekvenciju** ili **učestalost** pojavljivanja f_i.
- Radi lakše usporedbe više uzoraka različitih duljina, uvodi se i pojam **relativne frekvencije** od a_i koji se označava s f_{r_i}, a računa po formuli

  $$f_{r_i} = \frac{f_i}{n}$$

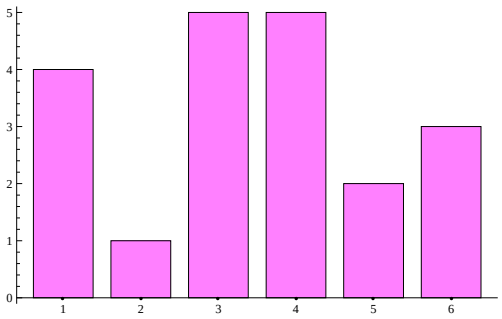
gdje je n duljina uzorka, odnosno broj ponavljanja pokusa (u ovom našem primjeru $n = 20$).

- Prikažimo sada podatke iz našeg primjera u **TABLICI FREKVENCIJA**

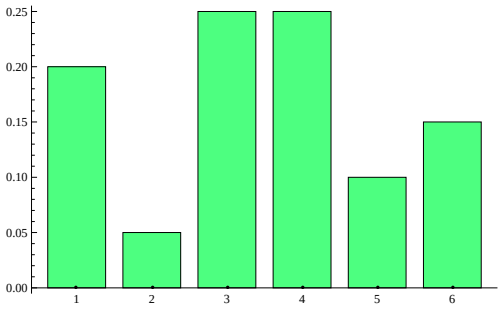
a_i	f_i	f_{r_i}	%
1	4	$\frac{4}{20} = 0.2$	20%
2	1	$\frac{1}{20} = 0.05$	5%
3	5	$\frac{5}{20} = 0.25$	25%
4	5	$\frac{5}{20} = 0.25$	25%
5	2	$\frac{2}{20} = 0.1$	10%
6	3	$\frac{3}{20} = 0.15$	15%
	$\Sigma = 20$	$\Sigma = 1.00$	$\Sigma = 100\%$

STUPČASTI DIJAGRAM (BAR - CHART)

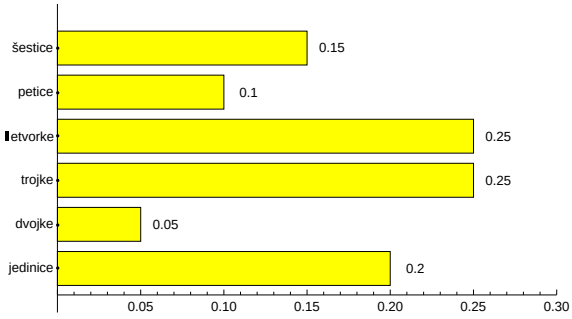
- Kada crtamo stupčasti dijagram diskretnog obilježja X, na x-os možemo nanijeti proizvoljnu širinu stupića; na y-os nanosimo frekvenciju f_i; stupića ima onoliko koliko ima različitih vrijednosti a_i koje obilježje X može poprimiti.



- Stupčasti dijagram diskretnog obilježja može se crtati i tako da se na y-os nanese relativna frekvencija f_{r_i}, što je bolje zbog usporedbe, npr. za različite duljine uzoraka n.



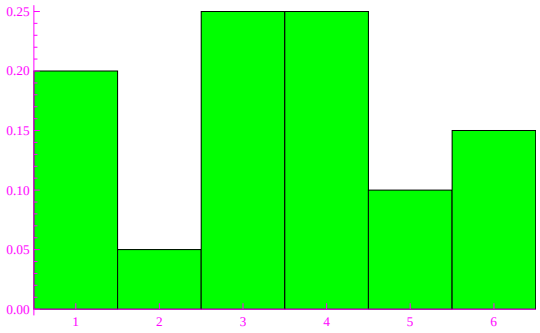
- **Horizontalni stupčasti dijagram** je varijacija stupčastog dijagrama: frekvencije, odnosno relativne frekvencije, nanosimo na x-os.



HISTOGRAM

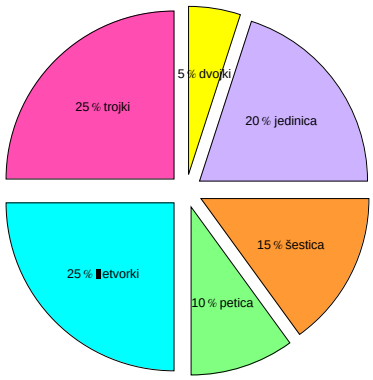
Histogram je još jedna (jako važna) varijacija stupčastog dijagrama. Glavna osobina mu je da je ukupna površina svih stupića koji ga čine jednaka 1, te da se svaka dva susjedna stupića međusobno dodiruju. Kako se (u slučaju diskretnog obilježja) širina baze stupića može izabrati proizvoljno, možemo ju izabrati tako da bude jednaka

1. Na y-os nanosimo relativne frekvencije. Tako postižemo da je površina svakog stupića jednaka upravo relativnoj frekvenciji, pa je stoga površina cijelog nastalog lika (što će reći ukupna površina ispod grafa) jednaka 1. Prikaz nema smisla za nenumeričke podatke!



STRUKTURNI KRUG (PIE CHART)

Ako imamo relativno malo različitih vrijednosti koje diskretno statističko obilježje može poprimiti, tada kao grafički prikaz možemo koristi strukturni krug.



Primjer 2. *Mjerena je visina (u metrima) 30 dvadesetogodišnjaka. 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 koje promatramo u ovom primjeru je *visina*. Kako ono može poprimiti (sve) vrijednosti iz određenog intervala, kojih ima beskonačno mnogo (za razliku od obilježja iz prethodnog primjera, gdje ih je bilo samo konačno mnogo), kažemo da je ono **neprekidno statističko obilježje**.

- Tablicu frekvencija sada je nemoguće napraviti kao u prvom primjeru. Ovdje je potrebno najprije *svrstati podatke u razrede (intervale)*. Postupak je sljedeći:

  1. Izaberemo odgovarajući broj razreda k - ne premalo i ne previše. Kao orijentacijsku pomoć pri tome možemo uzeti $k \approx \sqrt{n}$. U ovom primjeru $\sqrt{n} = \sqrt{30} \approx 5.477226$, što znači da možemo uzeti $k = 5$ ili $k = 6$. Uzet ćemo npr. $k = 6$. Sami pokušajte ponoviti ovaj postupak za $k = 5$.

  2. Odredimo širinu razreda c

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

  Važno! Uvijek zaokružujemo na više i na onaj broj decimala koliko imaju podaci! Da smo npr. dobili $c = 0.05023$, uzeli bismo $c = 0.06$.

  3. Formiramo razrede, odnosno odredimo njihove rubove. **Važno!** Rubove razreda uvijek uzimamo na jednu decimalu više

nego što ih imaju podaci! Pritom je zgodno (zbog jednostavnijeg računa i preglednosti) uzimati 5 kao zadnju decimalu. Lijevi rub prvog razreda biramo tako da bude malo manji od najmanjeg podatka (i tako da mu zadnja decimala bude 5). Kako je ovdje $x_{min} = 1.59$, za lijevi rub prvog razreda uzimamo 1.585. Desni rub prvog razreda dobivamo tako da lijevom rubu dodamo *širinu razreda* c. Dakle, desni rub prvog razreda je 1.585+0.05 = 1.635. Desni rub prvog razreda je ujedno i lijevi rub sljedećeg (drugog) razreda, pa ponavljanjem gornjeg postupka lako odredimo i rubove preostalih razreda.

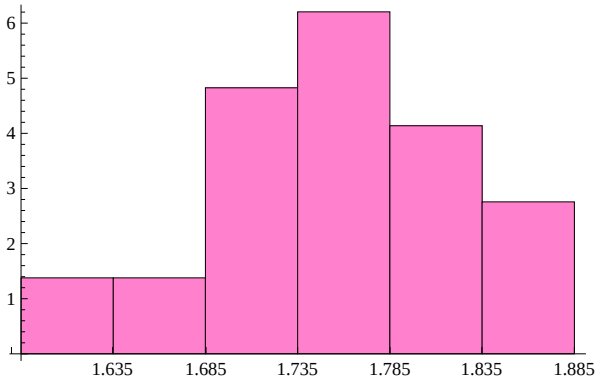
4. Prebrojimo koliko podataka je ušlo u pojedini razred, te formiramo donju tablicu.

RAZREDI	f_i
$[1.585, 1.635]$	2
$[1.635, 1.685]$	2
$[1.685, 1.735]$	7
$[1.735, 1.785]$	9
$[1.785, 1.835]$	6
$[1.835, 1.885]$	4
	$\Sigma = 30$

▶ Nacrtajmo sada **histogram** za ove podatke. Prisjetimo se najprije da je ovdje riječ o **neprekidnom obilježju**. To za posljedicu ima da *širinu stupića više ne možemo birati proizvoljno*, već je ona sada nužno *jednaka širini razreda*, tj. u ovom primjeru $c = 0.05$. **Važno!** Ukupna površina svih stupića koji čine his-

togram (odnosno površina ispod grafa) mora biti jednaka 1, pa na y-os ucrtavamo $\frac{f_{r_i}}{c}$, a ne f_{r_i}!

RAZREDI	f_i	f_{r_i}	f_{r_i}/c
$[1.585, 1.635]$	2	0.067	1.33
$[1.635, 1.685]$	2	0.067	1.33
$[1.685, 1.735]$	7	0.233	4.67
$[1.735, 1.785]$	9	0.3	6
$[1.785, 1.835]$	6	0.2	4
$[1.835, 1.885]$	4	0.133	2.67
	$\Sigma = 30$	$\Sigma = 1$	



STEM AND LEAF ("PETELJKA I LIST") DIJAGRAM

stem	leaf
1.5	9
1.6	05899
1.7	02222445566688999
1.8	0235668

stem	leaf
1.5	9
1.6_*	0
1.6^*	5899
1.7_*	0222244
1.7^*	5566688999
1.8_*	023
1.8^*	5668

1.2 Srednje vrijednosti uzorka

1.2.1 Aritmetička sredina uzorka

Aritmetička sredina uzorka je broj

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

Ima ju smisla računati samo za numeričke podatke. Ako je $\mathrm{Im}X = \{a_1, a_2, \ldots, a_k\}$ i pritom se a_i u uzorku pojavljuje f_i puta, tada

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

Primjer 3. *Izračunajte aritmetičku sredinu $\bar{x}$ za podatke iz Primjera 2.*

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

Medijan uzorka m je broj takav da je 50% svih podataka manje od ili jednako njemu i 50% svih podataka veće od ili jednako njemu. Ima smisla samo za numeričke podatke. Prilikom određivanja medijana podatke je najprije potrebno sortirati po veličini:

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

Pritom $x_{(j)}$ označava podatak koji je u promatranom uzorku j-ti po veličini. Općenito,

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

Ako je duljina uzorka neparan broj $n = 2k-1,\ k \in \mathbb{N}$, tada je $m = x_{(k)}$.
Ako je duljina uzorka paran broj $n = 2k,\ k \in \mathbb{N}$, tada je
$m = \dfrac{x_{(k)} + x_{(k+1)}}{2}$.

Primjer 4. *Nađite medijan uzorka za podatke iz Primjera 1.*

Rješenje: Sortirajmo podatke po veličini.

$$\begin{aligned} 1 \leq 1 \leq 1 \leq 1 \quad &\leq \quad 2 \leq 3 \leq 3 \leq 3 \leq 3 \leq \mathbf{3} \leq \mathbf{4} \leq 4 \leq 4 \leq 4 \leq 4 \\ &\leq \quad 5 \leq 5 \leq 6 \leq 6 \leq 6 \end{aligned}$$

$$n = 20 = 2 \cdot 10 \quad \Rightarrow \quad m = \frac{x_{(10)} + x_{(11)}}{2} = \frac{3+4}{2} = 3.5$$

■

1.2.3 Uzorački mod

Mod je vrijednost statističkog obilježja koja se u uzorku pojavljuje s najvećom frekvencijom. Koristan je kod statističkih obilježja koja nisu numerička, pa ne možemo izračunati aritmetičku sredinu i medijan.

Unimodalan uzorak je onaj u kojem postoji samo jedan mod. *Bimodalan uzorak* je onaj u kojem postoje dva moda - dvije različite vrijednosti s jednakom (najvećom) frekvencijom. Ukoliko svi podaci imaju jednaku frekvenciju, tada uzorak nema mod.

Primjer 5. *Nađite mod za podatke iz Primjera 1 i 2.*

Rješenje:

- u Primjeru 1: mod = 3 & mod=4 $\Rightarrow$ bimodalan uzorak
- u Primjeru 2: mod = 1.72

■

1.3 Mjere disperzije ili varijabiliteta

1.3.1 Uzoračka varijanca i uzoračka standardna devijacija

Uzoračka varijanca je kvadratno odstupanje podataka od aritmetičke sredine uzorka:

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

Uzoračka standardna devijacija je pozitivni korijen iz varijance:

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

Izvedimo sada formulu za varijancu koja će biti puno praktičnija za računanje. Vrijedi:

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \bar{x})^2 &= \sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2) = \sum_{i=1}^{n}x_i^2 - 2\bar{x}\sum_{i=1}^{n}x_i + \sum_{i=1}^{n}\bar{x}^2 \\ &= \sum_{i=1}^{n}x_i^2 - 2\bar{x}\cdot n\bar{x} + n\bar{x}^2 = \sum_{i=1}^{n}x_i^2 - n\bar{x}^2,\end{aligned}$$

pa smo tako dobili

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

Ako se u uzorku (međusobno različite) vrijednosti a_i pojavljuju s frekvencijom f_i $(i = 1, 2, \ldots, k)$, onda vrijedi:

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

Kao posljedica *Čebiševljeve nejednakosti* slijedi da se za $k > 1$ u intervalu

$$\langle \bar{x} - ks,\ \bar{x} + ks \rangle$$

nalazi barem $1 - \frac{1}{k^2}$ podataka. Specijalno, za $k = 2$, u intervalu $\langle \bar{x} - 2s,\ \bar{x} + 2s \rangle$ nalazi se barem $3/4$, odnosno 75% podataka. Za $k = 3$, u intervalu $\langle \bar{x} - 3s,\ \bar{x} + 3s \rangle$ nalazi se barem $8/9$, odnosno 89% podataka.

Primjer 6. *Izračunajte uzoračku varijancu s^2 i uzoračku standardnu devijaciju s za podatke iz Primjera 2.*

Rješenje:

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

■

1.3.2 Raspon uzorka

Raspon uzorka je razlika najvećeg i najmanjeg podatka u uzorku:

$$d = x_{max} - x_{min}.$$

Primjer 7. *Odredite raspon uzorka iz Primjera 2.*

Rješenje:

$$d = 1.88 - 1.59 = 0.29$$

■

1.3.3 Interkvartil

Donji kvartil q_L je broj takav da je 25% svih podataka manje od ili jednako njemu i 75% svih podataka veće od ili jednako njemu.

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

Gornji kvartil q_U je broj takav da je 75% svih podataka manje od ili jednako njemu i 25% svih podataka veće od ili jednako njemu.

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

Interkvartil je razlika gornjeg i donjeg kvartila:

$$d_q = q_U - q_L.$$

Prilikom određivanja kvartila trebat će nam sljedeća formula:

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

Primjer 8. *Odredite interkvartil za podatke iz Primjera 2.*

Rješenje:

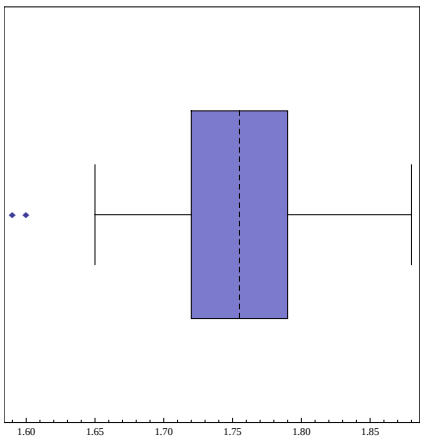
$$\begin{aligned} q_L &= x_{(\frac{n+1}{4})} = x_{(\frac{30+1}{4})} = x_{(7+\frac{3}{4})} = x_{(7)} + \frac{3}{4}\left(x_{(8)} - x_{(7)}\right) \\ &= 1.70 + \frac{3}{4}\left(1.72 - 1.70\right) = 1.715 \approx 1.72 \\ q_U &= x_{(\frac{3(n+1)}{4})} = x_{(\frac{93}{4})} = x_{(23+\frac{1}{4})} = x_{(23)}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

■

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

Primjer 9. *Nacrtajte box and whisker dijagram za podatke iz Primjera 2.*

$x_{(1)} = 1.59, \;\; q_L = 1.72, \;\; m = 1.755, \;\; q_U = 1.79, \;\; x_{(30)} = 1.88, \;\; d_q = 0.07$



1.4 Mjere lokacije

Medijan i kvartili spadaju u mjere lokacije jer se među podacima nalaze na specifičnoj lokaciji. Tu su još i:

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

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

  takav da je $k/10$ podataka manje od ili jednako njemu.

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

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

  takav da je $k/100$ ($k\%$) podataka manje od ili jednako njemu.

Decili su specijalan slučaj percentila: $D_1 = P_{10},\ D_2 = P_{20}, \ldots, D_9 = P_{90}$. Također, $m = D_5 = P_{50}$, te $q_L = P_{25}$ i $q_U = P_{75}$.

1.5 Mjere oblika

Uzorački k-ti centralni moment ($k \in \mathbb{N}$) definira se na sljedeći način:

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

Specijalno,

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

$\mu_2 = s^2$ uzoračka varijanca

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

Primjer 10. *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 μ_3 *je jednak* 0.

Koeficijent asimetrije uzorka (skewness) definiran je s:

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

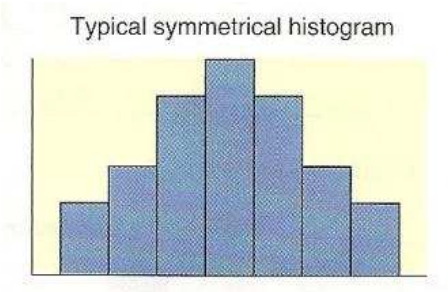
Vrijedi:

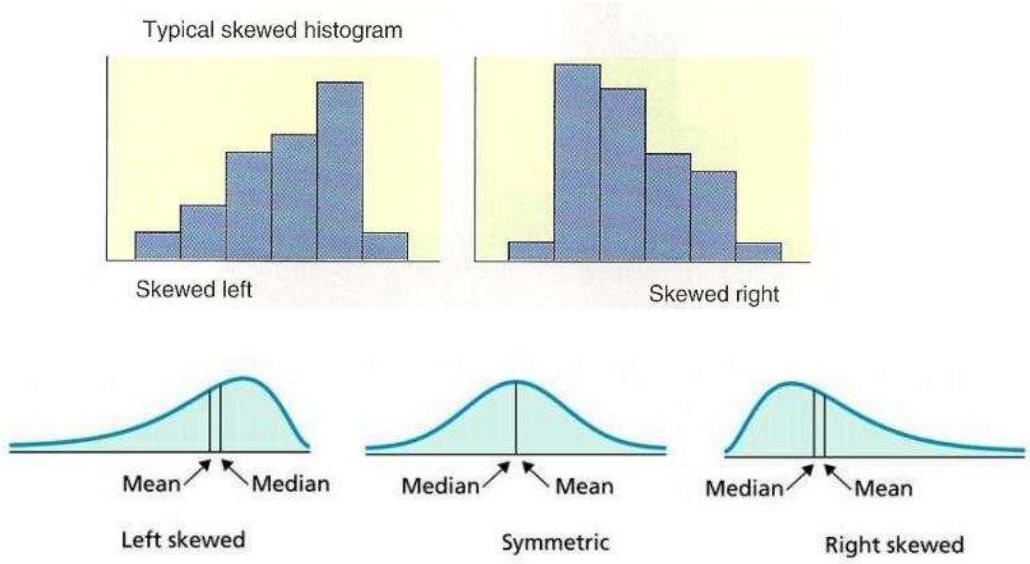
(i) $\alpha_3 = 0 \Rightarrow$ uzorak je SIMETRIČAN

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

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

Grubo govoreći, uzorak je **pozitivno asimetričan** (*positive skew, right-skewed*) ako se desni kraj pripadnog histograma produljuje i pretvara u "rep", dok je **negativno asimetričan** (*negative skew, left-skewed*) ako se isto događa na lijevom kraju. Uzorak je **simetričan** ako je pripadni histogram simetričan.





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

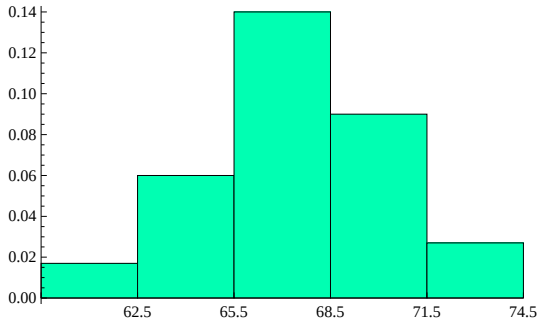
težina (kg)	*broj studenata*
$60-62$	5
$63-65$	18
$66-68$	42
$69-71$	27
$72-74$	8
	$\Sigma = 100$

Rješenje: Nacrtajmo najprije histogram (slika 1.1). Za razliku od Primjera 2, ovdje nam nisu zadani "goli" podaci, nego su oni već svrstani u razrede. Znamo da se kod histograma stupići moraju dodirivati, no ako pogledamo prikazane razrede vidjet ćemo da kod

njih postoje "rupe". Kako bi rupe "zakrpali", prvi razred shvaćat ćemo kao $[59.5, 62.5]$, drugi kao $[62.5, 65.5]$ i tako redom. Stvarna širina razreda je tako $c = 3$ (a ne $c = 2$!).

težina (kg)	$[59.5, 62.5]$	$[62.5, 65.5]$	$[65.5, 68.5]$	$[68.5, 71.5]$	$[71.5, 74.5]$
f_i	5	18	42	27	8
f_{r_i}	0.05	0.18	0.42	0.27	0.08
f_{r_i}/c	0.017	0.06	0.14	0.09	0.027

Slika 1.1: Histogram



Kada su podaci zadani, aritmetičku sredinu i uzoračku varijancu lako je izračunati. No, kao što smo već primijetili, ovdje to nije slučaj. Podatke stoga moramo *procijeniti*. Najmanje ćemo pogriješiti ako ih procijenimo *sredinama razreda* u kojima se nalaze. Sredine razreda označavat ćemo s $\bar{a}_i$.

težina (kg)	$[59.5, 62.5]$	$[62.5, 65.5]$	$[65.5, 68.5]$	$[68.5, 71.5]$	$[71.5, 74.5]$
f_i	5	18	42	27	8
$\bar{a}_i$	61	64	67	70	73

Sada imamo

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{5} f_i \cdot \bar{a}_i = \frac{1}{100}(5 \cdot 61 + 18 \cdot 64 + 42 \cdot 67 + 27 \cdot 70 + 8 \cdot 73) = 67.45$$

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{5} f_i \cdot \bar{a}_i^2 - n\bar{x}^2\right) = \frac{1}{99}\left(5 \cdot 61^2 + 18 \cdot 64^2 + 42 \cdot 67^2 + 27 \cdot 70^2 + 8 \cdot 73^2 - 100 \cdot 67.45^2\right) = 8.614$$

Medijan ćemo također morati procijeniti, kao i donji i gornji kvartil. Počnimo s medijanom. Odredimo najprije u kojem razredu se nalazi. U prva dva razreda sadržana su $5+18 = 23$ podatka, dok je u prva tri razreda sadržano njih $5+18+42 = 65$. Kako je medijan pedeseti podatak po veličini (budući je $n = 100$), to znači da se on nalazi negdje unutar trećeg razreda, odnosno $65.5 \leq m \leq 68.5$. Medijan ćemo dobiti interpolacijom, za što su nam potrebne **kumulativne relativne frekvencije**

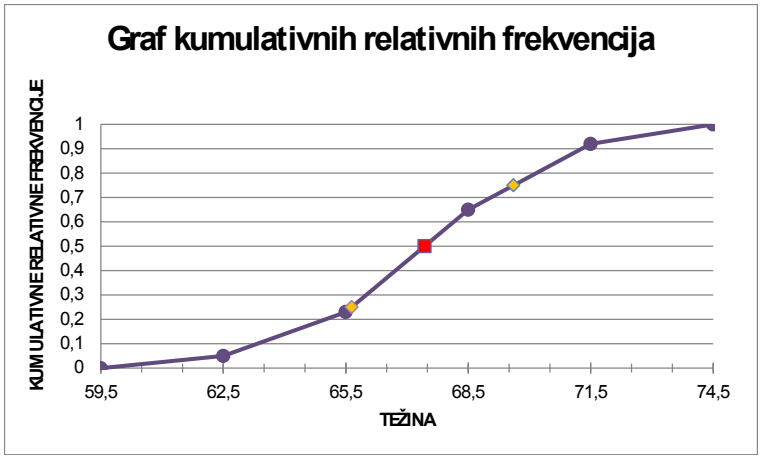
$$F_i = \sum_{j=1}^{i} f_{r_j}, \quad i = 1, \ldots, k.$$

težina (kg)	$[59.5, 62.5]$	$[62.5, 65.5]$	$[65.5, 68.5]$	$[68.5, 71.5]$	$[71.5, 74.5]$
f_{r_i}	0.05	0.18	0.42	0.27	0.08
F_i	0.05	0.23	0.65	0.92	1

Pomoću kumulativnih relativnih frekvencija možemo nacrtati **GRAF KUMULATIVNIH RELATIVNIH FREKVENCIJA**. Postupak je sljedeći: najprije ucrtamo točke u koordinatni sustav. Prva točka ima x-koordinatu jednaku *lijevom rubu prvog razreda*, a y-koordinata joj je jednaka 0. Sve ostale točke imaju x-koordinatu jednaku *desnom rubu odgovarajućeg razreda*, dok su im y-koordinate jednake *kumulativnoj relativnoj frekvenciji pripadajućeg razreda*. Dakle, točaka

uvijek imamo za jedan više od ukupnog broja razreda. Dvije po dvije susjedne točke zatim spojimo pravcem. Tako dobivamo graf po dijelovima linearne funkcije prikazan na slici 1.2.

Slika 1.2: Graf kumulativnih relativnih frekvencija



Na grafu kumulativnih relativnih frekvencija:

- medijan je x-koordinata točke čija je y-koordinata jednaka 0.5: $(m, 0.5)$
- donji kvartil je x-koordinata točke čija je y-koordinata jednaka 0.25: $(q_L, 0.25)$
- gornji kvartil je x-koordinata točke čija je y-koordinata jednaka 0.75: $(q_U, 0.75)$

Preciznije, točka $(m, 0.5)$ leži na pravcu koji je određen točkama $(65.5, 0.23)$ i $(68.5, 0.65)$. Sada imamo (jednadžba pravca kroz dvije točke):

$$0.5 - 0.23 = \frac{0.65 - 0.23}{68.5 - 65.5}(m - 65.5)$$
$$\Rightarrow \quad m = 65.5 + \frac{0.5 - 0.23}{0.65 - 0.23}(68.5 - 65.5) = 67.43$$

Primijetimo da se gornji račun može pojednostaviti

$$\begin{aligned} m &= 65.5 + \frac{0.5 - 0.23}{0.65 - 0.23}(68.5 - 65.5) = 65.5 + \frac{\frac{50}{100} - \frac{23}{100}}{\frac{65}{100} - \frac{23}{100}}(68.5 - 65.5) \\ &= 65.5 + \frac{27}{42} \cdot 3 = 67.43, \end{aligned}$$

pa sada vidimo da se medijan može odrediti na brz i jednostavan način budući je:

- **65.5**: lijevi rub razreda unutar kojeg smo odredili da se medijan nalazi
- **27**: broj podataka koje treba dodati na ona 23 podataka koliko ih ukupno ima u prva dva razreda da bi se došlo do podatka koji je na polovici po veličini - u ovom slučaju 50.-og podatka (23+27)
- **42**: broj podataka u razredu unutar kojeg smo odredili da se medijan nalazi
- **3**: širina razreda

Ostalo je još odrediti interkvartil. Kvartile ćemo odrediti koristeći ovaj kraći način (pokušajte ih sami naći koristeći graf kumulativnih relativnih frekvencija). Budući je duljina uzorka 100, donji kvartil je otprilike 25. podatak po veličini, a gornji kvartil 75. To znači da je donji kvartil drugi po veličini podatak u trećem razredu (23+2=25) i $65.5 \leq q_L \leq 68.5$. Kako bi primijenili gornju formulu treba nam: lijevi rub razreda u kojem se kvartil nalazi (65.5), broj podataka koji nam još treba do 25.-og podatka po veličini (2), broj podataka unutar razreda u kojem se traženi kvartil nalazi (42), te širina razreda (3). Sada imamo:

$$q_L = 65.5 + \frac{2}{42} \cdot 3 = 65.643$$

Znamo da se u prva tri razreda nalazi ukupno 65 podataka. U prva četiri se pak nalaze 92(=5+18+42+27) podatka. Dakle, gornji kvartil se nalazi unutar četvrtog razreda ($68.5 \leq q_U \leq 71.5$) i deseti je podatak po veličini unutar tog razreda (65+10=75). Slijedi:

$$q_U = 68.5 + \frac{10}{27} \cdot 3 = 69.61$$

i konačno

$$d_q = q_U - q_L = 69.61 - 65.643 = 3.967$$

■

Tablica "KONDENZATORI"

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		

Zadatak 2. *Izmjeren je kapacitet na 485 istovrsnih kondenzatora. Rezultati mjerenja dani su u gornjoj tablici "KONDENZATORI"; podaci su u μF zaokruženi na dvije decimale.*

(1) Nacrtajte histogram. (DZ)

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

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

Rješenje: Budući imamo uzorak veličine $n = 485$, $\frac{1}{n-1}$ 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:

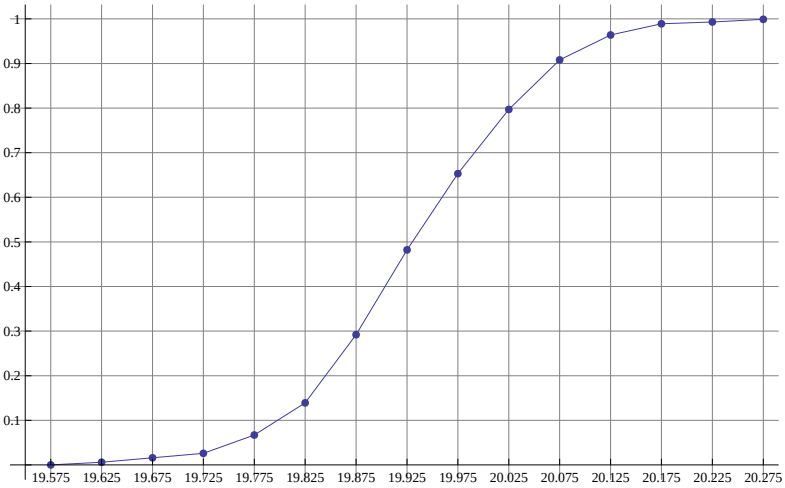
$$\begin{aligned}
\bar{x} &= \frac{1}{n}\sum_{i=1}^{k} f_i \cdot \bar{a}_i = \frac{1}{n}\sum_{i=1}^{k} f_i(\bar{a}_0 + c \cdot d_i) = \frac{1}{n}\left(\bar{a}_0 \sum_{i=1}^{k} f_i + c\sum_{i=1}^{k} f_i \cdot d_i\right) \\
&= \bar{a}_0 + c \cdot \bar{d}, \text{ gdje je } \bar{d} = \frac{1}{n}\sum_{i=1}^{k}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

Iz podataka dobivamo da je

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

Kod određivanja medijana, te donjeg i gornjeg kvartila pomoći će nam graf kumulativnih relativnih frekvencija (vidjeti Zadatak 2) koji je prikazan na slici 1.3.

Slika 1.3: Graf kumulativnih relativnih frekvencija



Točka $(m, 0.5)$ 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:

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

Slično se mogu izračunati donji q_L i gornji kvartil q_U.

$$\begin{aligned} & \frac{1}{4} - F_5 = \frac{F_6 - F_5}{a_6 - a_5}(q_L - a_5) \\ \Leftrightarrow \quad & \frac{1}{4} - 0.139 = \frac{0.292 - 0.139}{0.05}(q_L - 19.825) \;\Leftrightarrow\; q_L = 19.86\ \mu F \end{aligned}$$

$$\begin{aligned} & \frac{3}{4} - F_8 = \frac{F_9 - F_8}{a_9 - a_8}(q_U - a_8) \\ \Leftrightarrow \quad & \frac{3}{4} - 0.653 = \frac{0.797 - 0.653}{0.05}(q_U - 19.975) \;\Leftrightarrow\; q_U = 20.01\ \mu F \end{aligned}$$

■

Poglavlje 2

OSNOVE TEORIJE VJEROJATNOSTI

2.1 Osnovni pojmovi

- Slučajni pokus je pokus s više mogućih ishoda.
- Ishode slučajnog pokusa nazivamo **događaji**.
- Događaje koje ne možemo rastaviti na jednostavnije događaje nazivamo **elementarni događaji**.
- Događaje koji nisu elementarni nazivamo **složeni događaji**.
- **Skup** svih **elementarnih događaja** označavamo s Ω.
- Događaji (i elementarni i složeni) su *podskupovi* skupa elementarnih događaja.

- Kardinalni broj $|\Omega|$ skupa Ω jednak je ukupnom broju njegovih članova.

Primjer 11. *Bacamo simetričnu kocku. Kolika je vjerojatnost da je pao paran broj?*

Rješenje: Skup elementarnih događaja ovog slučajnog pokusa je:

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

Primijetimo da je kardinalni broj tog skupa $|\Omega| = 6$.

Označimo s A događaj čiju vjerojatnost želimo izračunati (događaje uvijek označavamo velikim štampanim slovom i pritom krećemo od početka abecede):

$$A = \{\text{na kocki je pao paran broj}\}.$$

Ako nije paran, kakav još broj može biti? Neparan, naravno. Označimo:

$$B = \{\text{na kocki je pao neparan broj}\}.$$

Budući 3 elementarna ishoda daju paran broj, tj. $A = \{2, 4, 6\}$ i isto tako 3 elementarna ishoda daju neparan broj, tj. $B = \{1, 3, 5\}$, jasno je da se događaji A i B pojavljuju s jednakom vjerojatnošću, odnosno vrijedi:

$$P(A) = P(B) = \frac{1}{2}.$$

Uočimo još nešto: događaji A i B su **komplementarni** ili **suprotni** - međusobno se isključuju ($A \cap B = \emptyset$), a zajedno pokrivaju sve moguće ishode ($A \cup B = \Omega$). Pišemo: $B = A^c$. ■

Neka je Ω skup elementarnih događaja i $\mathcal{P}(\Omega)$ skup svih podskupova od Ω. **Vjerojatnost** je funkcija $P : \mathcal{P}(\Omega) \to \mathbb{R}$ koja svakom događaju $A \in \mathcal{P}(\Omega)$ pridružuje broj $P(A)$ tako da vrijedi:

$(P1)$ $\quad P(A) \geq 0, \quad \forall A \in \mathcal{P}(\Omega)$

$(P2)$ $\quad P(\Omega) = 1$

$(P3)$ $\quad P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$, za sve $A_i \in \mathcal{P}(\Omega)$ takve da je $A_i \cap A_j = \emptyset,\ i \neq j$

Uređenu trojku $(\Omega, \mathcal{P}(\Omega), P)$ nazivamo **vjerojatnosni prostor**. Vjerojatnosni prostor je matematički model za promatrani slučajni pokus.

SVOJSTVA VJEROJATNOSTI

(1) $P(\emptyset) = 0$

(2) $A \subseteq B \quad \Rightarrow \quad P(A) \leq P(B)$

(3) $P(A^c) = 1 - P(A)$

(4) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

LAPLACEOV MODEL VJEROJATNOSTI

Neka je $\Omega = \{\omega_1, \ldots, \omega_m\}$, $m \in \mathbb{N}$. Pretpostavimo da su svi elementarni događaji *jednako vjerojatni*, tj. da je $P(\omega_i) = \frac{1}{m}$. Tada je vjerojatnost događaja A $(A \subseteq \Omega)$ jednaka:

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

Zadatak 3. *Bacamo 2 simetrične kocke. Kolika je vjerojatnost da zbroj na te 2 kocke bude jednak 7?*

Rješenje: Odredimo najprije prostor elementarnih događaja Ω.

$$\Omega = \{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),(2,1),(2,2),(2,3),\ldots,(6,6)\}$$
$$= \{(i,j) : 1 \leq i,j \leq 6\} \qquad \Rightarrow \quad |\Omega| = 6 \cdot 6 = 36$$

Zanima nam vjerojatnost događaja

$$A = \{\text{zbroj na 2 kocke jednak 7}\}.$$

Budući su svi elementarni događaji jednako vjerojatni (jer bacamo simetrične kocke), možemo primijeniti Laplaceov model računanja vjerojatnosti. Potrebno je prebrojati koliko je elementarnih događaja "povoljno" za događaj A. Dakle, zanimaju nas oni elementarni događaji, tj. ishodi, koji kad se dogode - dogodi se i A. Imamo:

$$A = \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\} \qquad \Rightarrow \quad |A| = 6$$

Vjerojatnost događaja A računamo kao kvocijent odgovarajućih kardinalnih brojeva:

$$\Longrightarrow P(A) = \frac{|A|}{|\Omega|} = \frac{6}{36} = \frac{1}{6}$$

■

2.2 Nezavisni događaji

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.

Zadatak 4. *Promatramo obitelj s 3 djece i događaje*

A = { *u obitelji su djeca oba spola* }
B = { *u obitelji nema više od 1 djevojčice* }.

Jesu li događaji A i B nezavisni?

Rješenje: Da bismo mogli odgovoriti na ovo pitanje, moramo provjeriti vrijedi li definicija.

Odredimo najprije skup svih elementarnih događaja:

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

Prirodno je pretpostaviti da je svih $2^3 = 8 = |\Omega|$ varijacija djece po spolu i starosti jednako vjerojatno.

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 ili bar jednim dječakom. Dakle,

$$A = \Omega \backslash \{\text{ MMM, ŽŽŽ }\}$$

pa je

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

Slično vidimo da je

$$B = \{\text{ MMM, MMŽ, MŽM, ŽMM }\}$$

pa je

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

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

$$A \cap B = \{\text{MMŽ, MŽM, ŽMM}\} = B \backslash \{\text{MMM}\}$$

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 ovi događaji nisu nezavisni! 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! ■

2.3 Uvjetna vjerojatnost

Pretpostavimo da znamo da se dogodio događaj B, te je $P(B) > 0$. Utječe li to na vjerojatnost događaja A?

Vjerojatnost događaja A uz uvjet da se dogodio događaj B nazivamo **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.

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

Zadatak 5. *Dvije prijateljice, Mirjana i Silvija, sretnu se u gradu. Mirjana kaže Silviji da ima 2 djece. Kolika je vjerojatnost, sa Silvijinog gledišta, da je drugo Mirjanino dijete kći ako Silvija otprije zna:*
a) da je jedno dijete sin,
b) da je sin starije dijete.

Rješenje: Definirajmo najprije prostor elementarnih događaja:

$$\Omega = \{\text{MM, MŽ, ŽM, ŽŽ}\}.$$

Zanima nas događaj

$$A = \{\text{jedno dijete je kći}\} = \{\text{MŽ, ŽM, ŽŽ}\}.$$

Točnije, zanima nas uvjetna vjerojatnost događaja A - uz uvjet da se dogodio

$$B_1 = \{\text{jedno dijete je sin}\} = \{\text{MM, MŽ, ŽM}\},$$

odnosno

$$B_2 = \{\text{starije dijete je sin}\} = \{\text{MM, MŽ}\}.$$

Moramo izračunati $P(A|B_1)$ i $P(A|B_2)$. Uočimo da je

$$A \cap B_1 = \{\text{MŽ, ŽM}\} \quad \text{i} \quad A \cap B_2 = \{\text{MŽ}\}.$$

Sada

$$P(A|B_1) = \frac{P(A \cap B_1)}{P(B_1)} = \frac{\frac{2}{4}}{\frac{3}{4}} = \frac{2}{3}, \qquad P(A|B_2) = \frac{P(A \cap B_2)}{P(B_2)} = \frac{\frac{1}{4}}{\frac{2}{4}} = \frac{1}{2}$$

■

2.4 Bayesova formula

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

$$1)\ P(H_i) > 0 \ \text{ za sve } \ i = 1, 2, \ldots, n$$

$$2)\ H_i \cap H_j = \emptyset \ \text{ za } \ i \neq j, \ \ i, j = 1, 2, \ldots, n$$

$$3)\ \bigcup_{i=1}^{n} H_i = \Omega$$

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

Formula potpune vjerojatnosti

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

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

Zadatak 6. *U jednoj kutiji šibica nalazi se 5 neupotrebljenih i 6 potrošenih šibica, a u drugoj 2 neupotrebljene i 9 potrošenih šibica. Na slučajan način iz svake kutije biramo po jednu šibicu i stavljamo u treću praznu kutiju. Zatim iz te treće kutije izvlačimo jednu šibicu. Kolika je vjerojatnost da ćemo njome moći zapaliti vatru?*

Rješenje: Zanima nas vjerojatnost događaja:

$$A = \{\text{izvučena je neupotrebljena šibica}\}.$$

Definiramo potpun sistem događaja:

H_1 = { u treću kutiju i iz prve i iz druge kutije
prebačena neupotrebljena šibica }
H_2 = { u treću kutiju iz prve kutije prebačena neupotrebljena,
a iz druge potrošena šibica }
H_3 = { u treću kutiju iz prve kutije prebačena potrošena,
a iz druge neupotrebljena šibica }
H_4 = { u treću kutiju i iz prve i iz druge kutije
prebačena potrošena šibica }

Vjerojatnost događaja A izračunat ćemo pomoću formule potpune vjerojatnosti. Izračunajmo najprije vjerojatnosti hipoteza.

$$P(H_1) = \frac{5}{11} \cdot \frac{2}{11} = \frac{10}{121}\,, \qquad P(H_2) = \frac{5}{11} \cdot \frac{9}{11} = \frac{45}{121}$$
$$P(H_3) = \frac{6}{11} \cdot \frac{2}{11} = \frac{12}{121}\,, \qquad P(H_4) = \frac{6}{11} \cdot \frac{9}{11} = \frac{54}{121}$$

Potrebne uvjetne vjerojatnosti su

$$P(A|H_1) = 1\,, \qquad P(A|H_2) = P(A|H_3) = \frac{1}{2}\,, \qquad P(A|H_4) = 0.$$

Slijedi:

$$\begin{aligned} P(A) &= \sum_{i=1}^{4} P(H_i)\cdot P(A|H_i) \\ &= \frac{10}{121}\cdot 1+\frac{45}{121}\cdot\frac{1}{2}+\frac{12}{121}\cdot\frac{1}{2}+\frac{54}{121}\cdot 0=\frac{7}{22}. \end{aligned}$$

■

Neka je zadan potpun sistem događaja $\{H_1, H_2, \ldots, H_n\}$. 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? Odgovor na to pitanje daje Bayesova formula.

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 li ovo, iz gornje jednakosti slijedi

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

□

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

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

Zadatak 7. *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 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 Diskretne slučajne varijable

Slučajna varijabla je funkcija $X : \Omega \to \mathbb{R}$ koja elementarnim događajima pridružuje brojeve.

Označimo s $\mathrm{Im}X$ skup svih različitih vrijednosti koje slučajna varijabla X može poprimiti. Ako je taj skup diskretan (prebrojiv), onda kažemo da je X **diskretna slučajna varijabla**.

Zakon razdiobe ili **distribucije** diskretne slučajne varijable X je zadan ako je zadan skup

$$\mathrm{Im}X = \{a_1, a_2, a_3, \ldots\},$$

te niz brojeva $p_i \geq 0$ takvih da je

$$p_i = P(X = a_i) \quad \text{i} \quad \sum_{i=1}^{\infty} p_i = 1.$$

Zakon razdiobe zapisujemo u obliku tablice:

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

Neka je $X : \Omega \to \mathbb{R}$ diskretna slučajna varijabla.

Funkcija gustoće (vjerojatnosti) od X je funkcija $p_X : \mathrm{Im}X \to [0,1]$ definirana s

$$p_X(a_i) = P(X = a_i) = p_i, \quad a_i \in \mathrm{Im}X.$$

Funkcija distribucije od 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

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

Zadatak 8. *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. Izračunajte vjerojatnost događaja $|X| \leq 1$.

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.

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

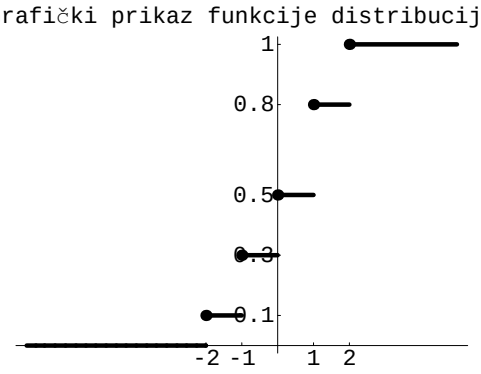
$$\begin{aligned} F_X(x) &= P(X = -2) + P(X = -1) + P(X = 0) + P(X = 1) \\ &= 0.1 + 0.2 + 0.2 + 0.3 = 0.8 \end{aligned}$$

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

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

Tako smo dobili:

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



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

$$P(|X| \leq 1) = P(-1 \leq X \leq 1) = P(X = -1) + P(X = 0) + P(X = 1)$$
$$= 0.2 + 0.2 + 0.3 = 0.7$$

Uočimo pritom da

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

■

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

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

Svojstva:

(E1) $\mathrm{E}[g(X)] = \sum_{a_i \in \mathrm{Im}X} g(a_i) \cdot p_X(a_i), \quad g : \mathbb{R} \to \mathbb{R}$

(E2) $\mathrm{E}[X + Y] = \mathrm{E}[X] + \mathrm{E}[Y]$ SVOJSTVO ADITIVNOSTI

(E3) $\mathrm{E}[cX] = c\,\mathrm{E}[X], \quad c \in \mathbb{R}$ SVOJSTVO HOMOGENOSTI

(E4) $\mathrm{E}[c] = c, \quad c \in \mathbb{R}$

ADITIVNOST + HOMOGENOST = LINEARNOST!

Varijanca diskretne slučajne varijable X je broj $\mathrm{Var}[X]$ definiran s

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

Standardna devijacija slučajne varijable X je broj

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

Svojstva:

(V1) $\mathrm{Var}[X] = \mathrm{E}[(X - \mathrm{E}[X])^2]$

(V2) $\mathrm{Var}[X] = \mathrm{E}[X^2] - (\mathrm{E}[X])^2$, gdje je $\mathrm{E}[X^2] = \sum_{a_i \in \mathrm{Im}X} a_i^2 \cdot p_X(a_i)$

(V3) $\mathrm{Var}[aX + b] = a^2 \mathrm{Var}[X]$

Svojstvo (V1) varijance slijedi odmah iz definicije nakon primjene svojstva (E1) očekivanja za $g(x) = (x - \mathrm{E}[X])^2$.

Dokaz svojstva (V2):

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

Dokaz svojstva (V3):

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

Neka su X i Y diskretne slučajne varijable i neka je $\mathrm{Im}X = \{a_1, a_2, \ldots\}$ i $\mathrm{Im}Y = \{b_1, b_2, \ldots\}$. Kažemo da su X i Y **nezavisne slučajne varijable** ako vrijedi

$$P(X = a_i, Y = b_j) = P(X = a_i) \cdot P(Y = b_j), \qquad \forall i, j$$

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

$$\begin{aligned} \mathrm{E}[X \cdot Y] &= \mathrm{E}[X] \cdot \mathrm{E}[Y] \\ \mathrm{Var}[X + Y] &= \mathrm{Var}[X] + \mathrm{Var}[Y] \end{aligned}$$

Općenito, ta dva identiteta ne vrijede!

Zadatak 9. *Neka je X diskretna slučajna varijabla iz Zadatka 8. Izračunajte njeno očekivanje $\mathrm{E}[X]$ i varijancu $\mathrm{Var}[X]$, te $\mathrm{E}[3X]$.*

Rješenje: Primjenom definicije dobivamo:

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

Varijancu računamo po formuli:

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

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

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

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

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

Jednostavniji način rješavanja je iskoristiti svojstvo homogenosti očekivanja prema kojem je

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

■

Zadatak 10. *Bacamo 2 kocke. Kolika 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: Skup elementarnih događaja je $\Omega = \{(i, j) : i, j = 1, 2, 3, 4, 5, 6\}$. Definirajmo slučajne varijable $X, Y : \Omega \to \mathbb{R}$ tako da X pamti broj na

prvoj kocki, a Y na drugoj. Imamo: $\mathrm{ImX} = \mathrm{ImY} = \{1,2,3,4,5,6\}$. 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 svojstvo nezavisnih događaja iz njihove definicije. Slijedi:

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

■

2.5.1 Binomna razdioba

Slučajna varijabla X ima **binomnu razdiobu s parametrima n i p** ako je funkcija gustoće te slučajne varijable zadana s:

$$p_X(k) = P(X = k) = \binom{n}{k} p^k q^{n-k}, \qquad k = 0, 1, 2, \ldots, n,$$

gdje je $q = 1 - p$.

- Oznaka: $X \sim B(n, p)$
- Očekivanje: $\mathrm{E}[X] = np$
- Varijanca: $\mathrm{Var}[X] = npq$

Osnovna svojstva koja opisuju binomnu razdiobu:

1. Pokus ponavljamo n puta.
2. Pokusi su međusobno nezavisni.
3. Postoje samo dva moguća ishoda u svakom pokusu. Jedan nazivamo "uspjeh", a drugi "neuspjeh".
4. Vjerojatnost "uspjeha" p jednaka je u svakom pokusu. Vjerojatnost "neuspjeha" je tada $q = 1 - p$.
5. Binomna slučajna varijabla poprima vrijednost k koja je jednaka broju "uspjeha" koji su se dogodili u tih n pokusa.
6. Binomna razdioba opisuje slučajno izvlačenje s vraćanjem.

Zadatak 11. *U kutiji se nalazi 20 kuglica od kojih je 12 crnih i 8 bijelih. Kolika je vjerojatnost da od 5 slučajno odabranih kuglica točno 3 budu crne i točno 2 budu bijele ako kuglice vraćamo?*

Rješenje: Događaj čiju vjerojatnost želimo izračunati je

$$A = \{\text{izvučene su 3 crne i 2 bijele kuglice}\}.$$

Kada biramo s vraćanjem, vjerojatnost da u jednom izvlačenju izvučemo crnu kuglicu jednaka je u svakom izvlačenju. Ako taj događaj označimo s C, imamo

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

Događaj da je izvučena bijela kuglica (označimo taj događaj s B) komplementaran je događaju C. Znamo da je zbroj vjerojatnosti komplementarnih događaja jednak 1, pa odatle lako izračunamo vjerojatnost od B:

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

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

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

Ostalo je izračunati $P(A)$. Kako kuglice vraćamo, postoji *uređaj* pri njihovom izvlačenju - zna se koja (i kakva) je bila prva, koja druga, koja treća itd. Tu nam se otvara mogućnost izbora: koja po redu je bila svaka od 3 izvučene crne kuglice? Prva, treća i peta? Druga, četvrta i peta? Druga, treća i četvrta? Sve su to naime različiti elementarni događaji. Dakle, od 5 mjesta (u poretku izvlačenja) moramo izabrati 3 na kojima su crne kuglice (na preostala 2 su onda automatski bijele). To možemo učiniti na $\binom{5}{3}$ načina. Vjerojatnost da u jednom izvlačenju bude izvučena crna kuglica je, kao što znamo, $\frac{3}{5}$. Sljedeće izvlačenje je nezavisno od prethodnog, pa je vjerojatnost da smo izvukli 2 crne kuglice 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 morala je biti izvučena bijela kuglica, 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 bijelih kuglica, odnosno da biramo mjesta na kojima su bijele kuglice, što je moguće učiniti na $\binom{5}{2}$ načina. Sada bi komplementaran događaj bio izvlačenje crne kuglice 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.

No, primijetimo da se zadatak može riješiti i na puno kraći način ako *uvedemo slučajnu varijablu* X koja *broji izvučene crne kuglice* (mogli bismo, naravno, definirati i varijablu koja broji bijele kuglice). Ta slučajna varijabla X očito ima binomnu razdiobu, s parametrima $n = 5$ (=broj ponavljanja pokusa = broj izvlačenja kuglica) i $p = 3/5$ (=vjerojatnost "uspjeha" = vjerojatnost izvlačenja crne kuglice u jednom izvlačenju). Dakle, $X \sim B(5,\ 3/5)$, a funkcija gustoće od X je oblika:

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

Nas zanima $P(X = 3)$, dakle slučaj $k = 3$, što je jednako

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

Ako bi definirali slučajnu varijablu Y koja broji izvučene bijele kuglice, tada bismo imali $Y \sim B(5,\ 2/5)$, budući se u ovom slučaju

mijenja ono što smatramo "uspjehom" (pa time i njegova vjerojatnost). Funkcija gustoće od Y je oblika

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

a nas zanima $P(Y = 2) = P(A)$ što je jednako

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

■

Zadatak 12. *Košarkaš gađa koš 4 puta i u svakom pokušaju pogađa s vjerojatnošću 0.82. 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 4 puta. "Uspjeh" je pogodak u koš, "neuspjeh" je promašaj. Vjerojatnost "uspjeha" je zadana i jednaka je 0.82; vjerojatnost "neuspjeha" je tada 0.18 (=1-0.82). Definirajmo slučajnu varijablu X koja broji pogotke. Ona ima binomnu razdiobu:

$$X \sim B\,(4,\ 0.82)\,,$$

a funkcija gustoće od X je:

$$p_X(k) = P(X = k) = \binom{4}{k} (0.82)^k\,(0.18)^{4-k}\,, \quad k = 0, 1, 2, 3, 4. \qquad (2.1)$$

a) Želimo izračunati vjerojatnost da je košarkaš pogodio koš točno 3 puta, što je zapravo $P(X = 3)$. Uvrštavanjem $k = 3$ u (2.1) lako dobijemo rješenje:

$$P(X = 3) = \binom{4}{3} (0.82)^3\,(0.18)^1 = 0.397.$$

b) Kolika je vjerojatnost da košarkaš pogodi koš barem 3 puta? Moramo izračunati

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

$P(X = 3)$ smo već izračunali pod a), a $P(X = 4)$ računamo analogno - uvrštavanjem $k = 4$ u (2.1).

$$P(X = 4) = \binom{4}{4} (0.82)^4 (0.18)^0 = 0.452$$
$$\Longrightarrow P(X \geq 3) = 0.397 + 0.452 = 0.849.$$

c) Kolika je vjerojatnost da košarkaš pogodi koš najviše 2 puta, odnosno $P(X \leq 2)$? Taj događaj je suprotan događaju kojeg smo promatrali pod b), te vrijedi:

$$P(X \leq 2) = 1 - P(X \geq 3) = 1 - 0.849 = 0.151.$$

Naravno, vjerojatnost tog događaja možemo izračunati i direktno:

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

■

Zadatak 13. *Odredite očekivanje, varijancu i devijaciju slučajne varijable* $X \sim B(4,\ 0.82)$. *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:

$$\begin{aligned} \mathrm{E}[X] &= np = 4 \cdot 0.82 = 3.28 \\ \mathrm{Var}[X] &= npq = 4 \cdot 0.82 \cdot 0.18 = 0.5904 \\ \sigma_X &= \sqrt{\mathrm{Var}[X]} = \sqrt{0.5904} = 0.768 \end{aligned}$$

Traženi interval je:

$$3.28 \pm 2 \cdot 0.768 = 3.28 \pm 1.536 \quad \Rightarrow \quad 1.744 < X < 4.816.$$

Nadalje,

$$\begin{aligned} P(1.744 < X < 4.816) &= P(2 \leq X \leq 4) \\ &= P(X = 2) + P(X = 3) + P(X = 4) \end{aligned}$$

$P(X = 3)$ i $P(X = 4)$ već smo izračunali u prethodnom zadatku. Nedostaje nam još:

$$P(X = 2) = \binom{4}{2} (0.82)^2 (0.18)^2 = 0.131,$$

pa sada imamo:

$$\begin{aligned} &P(\mathrm{E}[X] - 2\sigma_X < X < \mathrm{E}[X] + 2\sigma_X) = P(1.744 < X < 4.816) \\ &= 0.131 + 0.397 + 0.452 = 0.98. \end{aligned}$$

Odavde možemo zaključiti da će vrijednost ove slučajne varijable u 98% slučajeva odstupati od svog očekivanja za najviše 2 devijacije. ■

Zadatak 14. *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. Maknemo li sve aseve, ostat će nam 48 karata. Stoga,

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

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

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

■

2.5.2 Hipergeometrijska razdioba

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

$$p_X(k)=P(X=k)=\frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}},$$

$$\max\{0,n-(N-M)\}\le k\le\min\{n,M\}$$

- Očekivanje: $\mathrm{E}[X]=\dfrac{nM}{N}$
- Varijanca: $\mathrm{Var}[X]=\dfrac{nM(N-M)(N-n)}{N^2(N-1)}$

Osnovna svojstva koja opisuju hipergeometrijsku razdiobu:

1. Pokus se sastoji od slučajnog izvlačenja bez vraćanja n elemenata iz skupa od ukupno N elemenata, od kojih je M jedne vrste (izvlačenje kojih smatramo "uspjehom") i $N-M$ neke druge vrste (izvlačenje kojih smatramo "neuspjehom").

2. Hipergeometrijska slučajna varijabla poprima vrijednost k jednaku broju "uspjeha", odnosno broju izvučenih elemenata prve vrste od ukupno izvučenih n elemenata.

Zadatak 15. *U kutiji se nalazi 20 kuglica od kojih je 12 crnih i 8 bijelih. Kolika je vjerojatnost da od 5 slučajno odabranih kuglica točno 3 budu crne i 2 budu bijele ako kuglice ne vraćamo?*

Rješenje: Primijetimo najprije da je ovaj zadatak vrlo sličan Zadatku 11. Zanima nas vjerojatnost istog događaja A, no bitna razlika je u tome što sada kuglice *ne vraćamo*.

Broj načina na koji od 20 kuglica možemo izabrati njih 5 (bilo kojih 5) je $\binom{20}{5}$. Imamo $|\Omega| = \binom{20}{5}$ i pritom su svi elementarni događaji jednako vjerojatni.

Broj načina na koji od ukupno 12 crnih kuglica možemo izabrati njih 3 je $\binom{12}{3}$. Analogno, broj načina na koji od ukupno 8 bijelih kuglica možemo izabrati 2 je $\binom{8}{2}$. Ako istovremeno izvučemo 3 crne i 2 bijele kuglice, dogodit će se A. Imamo: $|A| = \binom{12}{3} \cdot \binom{8}{2}$

Vjerojatnost događaja A je:

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

Kao i Zadatak 11, i ovaj možemo riješiti tako da uvedemo *slučajnu varijablu* X koja *broji izvučene crne kuglice*. Ona ima hipergeometrijsku razdiobu, a funkcija gustoće joj je zadana s

$$p_X(k) = P(X = k) = \frac{\binom{12}{k}\binom{8}{5-k}}{\binom{20}{5}}, \quad k = 0, 1, 2, 3, 4, 5.$$

Sada

$$P(A) = P(X = 3) = \frac{\binom{12}{3}\binom{8}{2}}{\binom{20}{5}} = 0.39732.$$

Možemo uvesti i slučajnu varijablu Y koja *broji izvučene bijele kuglice*. Njena funkcija gustoće je

$$p_Y(k) = P(Y = k) = \frac{\binom{8}{k}\binom{12}{5-k}}{\binom{20}{5}}, \quad k = 0, 1, 2, 3, 4, 5,$$

pa odatle

$$P(A) = P(Y = 2) = \frac{\binom{8}{2}\binom{12}{3}}{\binom{20}{5}} = 0.39732.$$

■

Zadatak 16. *Vještica želi otrovati Snjeguljicu. Uzela je košaru s 10 jabuka i otrovala 2. Na putu je počela malo peći savjest pa je odlučila dati Snjeguljici da sama izabere 2 jabuke iz košare umjesto da joj odmah da one 2 otrovane. a) Kolika je vjerojatnost da Snjeguljica izabere barem 1 otrovanu jabuku? b) Koliki je očekivani broj izvučenih neotrovanih jabuka?*

Rješenje: Definirajmo slučajnu varijablu X koja broji izvučene otrovane jabuke. Ona ima hipergeometrijsku razdiobu (izvlačenje je očito bez vraćanja), predmeti prve vrste ("uspjesi") su otrovane jabuke, a funkcija gustoće od X je dana s:

$$p_X(k) = P(X = k) = \frac{\binom{2}{k}\binom{8}{2-k}}{\binom{10}{2}}, \quad k = 0, 1, 2.$$

Na pitanje pod a) moći ćemo odgovoriti nakon što izračunamo $P(X \geq 1)$. To možemo na dva načina:

$$P(X \geq 1) = P(X = 1) + P(X = 2)$$

ili $$P(X \geq 1) = 1 - P(X = 0).$$

Drugi put je očito kraći, pa računamo:

$$P(X \geq 1) = 1 - P(X = 0) = 1 - \frac{\binom{2}{0}\binom{8}{2}}{\binom{10}{2}} = 1 - \frac{28}{45} = \frac{17}{45} = 0.378.$$

Da bismo odgovorili na pitanje pod b), uvedimo novu slučajnu varijablu Y koja broji izvučene neotrovane jabuke; ona također ima hipergeometrijsku razdiobu, no sada su predmeti prve vrste ("uspjesi") neotrovane jabuke, odnosno $M = 8$. Uvrštavanje u formulu za očekivanje hipergeometrijske razdiobe daje:

$$\mathrm{E}[Y] = \frac{nM}{N} = \frac{2 \cdot 8}{10} = 1.6.$$

■

Zadatak 17. *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: Definirana je slučajna varijabla X = broj izvučenih crvenih ruža. Ona ima hipergeometrijsku razdiobu i pritom izvlačenje crvene ruže smatramo "uspjehom", pa je funkcija gustoće dana s:

$$p_X(k) = P(X = k) = \frac{\binom{4}{k}\binom{6}{3-k}}{\binom{10}{3}}, \quad k = 0, 1, 2, 3.$$

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

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

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

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

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

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

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

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

Umjesto primjenom definicije, očekivanje smo mogli izračunati i pomoću formule:

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

■

2.5.3 Poissonova razdioba

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

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

- Oznaka: $X \sim P(\lambda)$
- Očekivanje: $\mathrm{E}[X] = \lambda$
- Varijanca: $\mathrm{Var}[X] = \lambda$

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.

Osnovna svojstva koja opisuju Poissonovu razdiobu:

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 u bilo kojoj drugoj jedinici.

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

Zadatak 18. *Slučajna varijabla X ima Poissonovu razdiobu. Ako vrijedi $P(X = 1) = P(X = 2)$, izračunajte $\mathrm{E}[X]$ i $P(X \geq 3)$.*

Rješenje: Zadano je:

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

Iz danog uvjeta slijedi:

$$\begin{aligned} P(X = 1) = P(X = 2) \quad &\Rightarrow \quad \frac{\lambda^1}{1!}\, e^{-\lambda} = \frac{\lambda^2}{2!}\, e^{-\lambda} \;\Rightarrow\; \lambda = \frac{\lambda^2}{2} \;\Rightarrow\; \lambda^2 - 2\lambda = 0 \\ &\Rightarrow \quad \lambda(\lambda - 2) = 0 \;\Rightarrow \lambda_1 = 0,\; \lambda_2 = 2 \end{aligned}$$

Kako po definiciji mora biti $\lambda > 0$, jedino rješenje je $\lambda = 2$, pa imamo:

$$\begin{aligned} X \sim P(2), \qquad & P(X = k) = \frac{2^k}{k!}\, e^{-2}, \quad k = 0, 1, 2, 3, \ldots \\ \mathrm{E}[X] \;&=\; \lambda = 2 \\ P(X \geq 3) \;&=\; 1 - P(X < 3) = 1 - P(X \leq 2) \\ &=\; 1 - P(X = 0) - P(X = 1) - P(X = 2) \\ &=\; 1 - \frac{2^0}{0!}\, e^{-2} - \frac{2^1}{1!}\, e^{-2} - \frac{2^2}{2!}\, e^{-2} \\ &=\; 1 - e^{-2}\,(1 + 2 + 2) = 0.323 \end{aligned}$$

■

Zadatak 19. *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. Svako slovo na pojedinoj stranici može biti greška (tipfeler) ili ne. Ako je greška, smatramo ga uspjehom (budući naš X broji upravo greške). Bi li ovdje mogla pomoći binomna razdioba? Da vidimo - kolika je vjerojatnost "uspjeha" (odnosno da slovo bude greška)? To ne znamo. Koliko ima slova na stranici (broj ponavljanja pokusa)? Ni to ne znamo. Što znamo? Znamo *prosječan* broj grešaka na stranici; on je jednak $220/200 = 1.1$. Stoga, naš X ima Poissonovu razdiobu. Kako znamo da je parametar Poissonove razdiobe jednak očekivanom ili prosječnom broju događaja (= broj grešaka) koji se dogode u jednoj jedinici (= na jednoj stranici), imamo $\lambda = 1.1$. Slijedi

$$p_X(k) = P(X = k) = \frac{\lambda^k}{k!}\, e^{-\lambda} = \frac{(1.1)^k}{k!}\, e^{-1.1}, \quad k = 0, 1, 2, \ldots \tag{2.2}$$

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 (2.2). Dobivamo:

$a)$ $\quad P(X = 0) = \dfrac{(1.1)^0}{0!}\, e^{-1.1} = e^{-1.1} = 0.333$

$b)$ $\quad P(X = 1) = \dfrac{(1.1)^1}{1!}\, e^{-1.1} = 0.366$

$c)$ $\quad P(X \geq 2) = 1 - P(X = 0) - P(X = 1) = 1 - 0.333 - 0.366 = 0.301$

■

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

2.5.4 Aproksimacija binomne razdiobe Poissonovom

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

Zadatak 20. *Kolika je vjerojatnost da među 200 ljudi barem 4 budu ljevaci ako ljevaka prosječno ima* 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:

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

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

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 21. *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. Ona 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:

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

Slijedi:

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

■

2.6 Neprekidne slučajne varijable

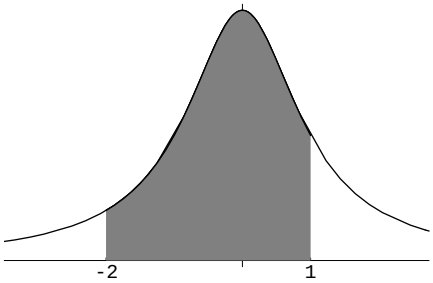
Za slučajnu varijablu X kažemo da je **neprekidna** ako je $\mathrm{Im}X$ interval u $\mathbb{R}$ i postoji nenegativna funkcija $f_X : \mathbb{R} \to \mathbb{R}$ tako da vrijedi

$$P(a \leq X \leq b) = \int_a^b f_X(t)dt, \quad \text{za sve } a, b \in \mathbb{R}\ (a < b).$$

Funkciju f_X zovemo **funkcija gustoće** od X.

Vjerojatnost da vrijednost slučajne varijable X bude u intervalu $[a, b]$ jednaka je površini ispod grafa funkcije gustoće na intervalu $[a, b]$. Ako je na slici 2.1 prikazan graf funkcije gustoće od X, tada je $P(-2 \leq X \leq 1)$ jednaka osjenčanoj površini.

Slika 2.1: Graf funkcije gustoće i vjerojatnost



Funkcija distribucije F_X neprekidne slučajne varijable X definirana je s:

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)dt \tag{2.3}$$

Vrijedi:

$$P(a \leq X \leq b) = F_X(b) - F_X(a) \tag{2.4}$$

Navedimo još dva svojstva neprekidne slučajne varijable:

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

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

(2) Ukupna površina ispod grafa funkcije gustoće jednaka je 1, odnosno vrijedi

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

Matematičko očekivanje neprekidne slučajne varijable X definirano je s

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} t \cdot f_X(t)dt, \tag{2.5}$$

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

$$\mathrm{Var}[X] = \mathrm{E}[X^2] - (\mathrm{E}[X])^2 \tag{2.6}$$

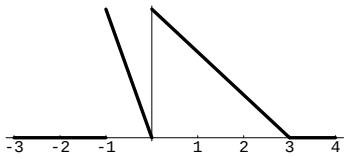
no sada je

$$\mathrm{E}[X^2] = \int_{-\infty}^{\infty} t^2 \cdot f_X(t)dt. \tag{2.7}$$

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

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

Zadatak 22. *Funkcija gustoće neke slučajne varijable X dana je grafom prikazanim na donjoj slici. 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:

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

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

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

Sljedeći korak je odrediti funkciju distribucije $F_X(x)$. Prisjetimo se njene definicije (2.3). Imamo:

$$x \leq -1: \quad F_X(x) = \int_{-\infty}^{x} f_X(t)dt = \int_{-\infty}^{x} 0dt = 0$$

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

pa slijedi:

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

Varijancu zadane neprekidne slučajne varijable X izračunat ćemo koristeći (2.6). Najprije izračunajmo očekivanje $\mathrm{E}[X]$ pomoću (2.5).

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

Nadalje, $\mathrm{E}[X^2]$ računamo pomoću (2.7):

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

Sada, prema (2.6), 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 (2.4) dobivamo:

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

■

2.6.1 Normalna razdioba

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

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

- Oznaka: $X \sim N(\mu, \sigma^2)$
- $f_X(x) > 0, \ \forall x \in \mathbb{R} \ \Rightarrow \ \mathrm{Im}X = \mathbb{R}$
- Očekivanje: $\mathrm{E}[X] = \mu$
- Varijanca: $\mathrm{Var}[X] = \sigma^2$

Normalna razdioba je invarijantna na linearne 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 $\Phi(x)$. Imamo:

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

i vrijedi

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

Zadatak 23. *Neka je zadana slučajna varijabla $X \sim N(0,1)$. Odredite vjerojatnosti događaja: $a)$ $X \leq 1$ $b)$ $X \geq 1$ $c)$ $0 \leq X \leq 1$ $d)$ $-1 \leq X \leq 2$.*

Rješenje:

$a)$ $P(X \leq 1) = \Phi(1) = 0.84134$

$b)$ $P(X \geq 1) = 1 - P(X < 1) = 1 - P(X \leq 1) = 1 - \Phi(1) = 0.15866$

$c)$ $P(0 \leq X \leq 1) = \Phi(1) - \Phi(0) = 0.84134 - 0.5 = 0.34134$

$d)$ $P(-1 \leq X \leq 2) = \Phi(2) - \Phi(-1) = \Phi(2) - (1 - \Phi(1))$

$= 0.97725 - 1 + 0.84134 = 0.81859$

■

Zadatak 24. *Neka je zadana slučajna varijabla $X \sim N(2,4)$. Odredite vjerojatnosti događaja: $a)$ $X \geq 4$, $b)$ $0 \leq X \leq 4$.*

Rješenje:

$$X \sim N(2,4) \quad \Rightarrow \quad X^* = \frac{X-\mu}{\sigma} = \frac{X-2}{2} \sim N(0,1)$$

$a)$ $$P(X \geq 4) = 1 - P(X < 4) = 1 - P(X \leq 4) = 1 - F_X(4)$$
$$= 1 - P\left(X^* \leq \frac{4-2}{2}\right) = 1 - \Phi(1) = 1 - 0.84134 = 0.15866$$

$b)$ $$P(0 \leq X \leq 4) = F_X(4) - F_X(0)$$
$$= P\left(\frac{0-2}{2} \leq \frac{X-2}{2} \leq \frac{4-2}{2}\right) = P(-1 \leq X^* \leq 1)$$
$$= \Phi(1) - \Phi(-1) = \Phi(1) - (1 - \Phi(1)) = 2 \cdot 0.84134 - 1 = 0.68268$$

■

Zadatak 25. *Slučajna varijabla X mjeri odstupanje aviona od sredine dozvoljenog koridora. Normalno je distribuirana; očekivanje joj je 100m, a standardna devijacija 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 i 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 dolje pa imamo:

$$P(-250 \leq X \leq 250) = P\left(\frac{-250-100}{200} \leq X^* \leq \frac{250-100}{200}\right)$$
$$= \Phi(0.75) - \Phi(-1.75) = \Phi(0.75) - 1 + \Phi(1.75)$$
$$= 0.77337 - 1 + 0.95994 = 0.73331$$

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

$$P(X \geq 250) = 1 - P(X < 250) = 1 - P(X \leq 250) =$$
$$= 1 - \Phi\left(\frac{250-100}{200}\right) = 1 - \Phi(0.75) = 1 - 0.77337 = 0.22663$$

■

Zadatak 26. *Slučajna varijabla X ima normalnu razdiobu $N(2,4)$. Izračunajte uvjetnu vjerojatnost: $P(-1 \leq X \leq 1 \mid 0 < X < 3)$.*

Rješenje:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$
$$P(A \cap B) = P(-1 \leq X \leq 1,\ 0 < X < 3) = P(0 < X \leq 1)$$
$$P(-1 \leq X \leq 1 \mid 0 < X < 3) = \frac{P(0 < X \leq 1)}{P(0 < X < 3)}$$
$$= \frac{P\left(\frac{0-2}{2} < X^* \leq \frac{1-2}{2}\right)}{P\left(\frac{0-2}{2} < X^* < \frac{3-2}{2}\right)} = \frac{\Phi(-0.5) - \Phi(-1)}{\Phi(0.5) - \Phi(-1)}$$
$$= \frac{\Phi(1) - \Phi(0.5)}{\Phi(0.5) - 1 + \Phi(1)} = \frac{0.84134 - 0.69146}{0.69146 - 1 + 0.84134} = 0.281306$$

■

2.6.2 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, prema Centralnom graničnom teoremu (CGT), vrijedi aproksimacija:

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

Aproksimacija je to bolja što je vrijednost parametra p bliža $1/2$.

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

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

■

2.6.3 Eksponencijalna razdioba

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

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

Zadatak 28. *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:

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

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

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

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:

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

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

$$c) \quad P(1 \leq X \leq 2 \mid X \geq 1) = \frac{P(1 \leq X \leq 2\ ,\ X \geq 1)}{P(X \geq 1)} = \frac{P(1 \leq X \leq 2)}{1 - P(X \leq 1)}$$
$$= \frac{F_X(2) - F_X(1)}{1 - F_X(1)} = \frac{0.239}{1 - 0.393} = 0.393$$

■

Poglavlje 3

TESTIRANJE STATISTIČKIH HIPOTEZA I POUZDANI INTERVALI

3.1 Procjena parametara

Cilj statističke analize je na osnovi uzorka uzetog iz populacije izvesti određene zaključke o distribuciji statističkog obilježja, odnosno slučajne varijable X, koju proučavamo.

Neka je X slučajna varijabla s konačnim očekivanjem $\mu = \mathrm{E}[X]$ i varijancom $\sigma^2 = \mathrm{Var}[X]$. Promatramo **slučajni uzorak** koji se sastoji od n *nezavisnih jednako distribuiranih slučajnih varijabli* $X_1, X_2, \ldots, X_n$, čija je distribucija jednaka distribuciji mjerenog statističkog obilježja. Želimo procijeniti μ i σ^2 na osnovi uzorka.

Pretpostavimo najprije da je $X \sim N(\mu, \sigma^2)$. To onda znači i da je

$X_i \sim N(\mu, \sigma^2),\ i = 1, 2, \ldots, n$. **Aritmetičku sredinu uzorka** definiramo s

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \ldots + X_n). \tag{3.1}$$

$\bar{X}$ je slučajna varijabla, pa pokušajmo odrediti njenu razdiobu. Ona će "naslijediti" normalnu razdiobu budući je dobivena kao zbroj nezavisnih normalno distribuiranih slučajnih varijabli, no s kojim parametrima? Imamo:

$$\begin{aligned} \mathrm{E}[\bar{X}] &= \mathrm{E}\left[\frac{1}{n}(X_1 + X_2 + \ldots + X_n)\right] = \frac{1}{n}\left(\mathrm{E}[X_1] + \mathrm{E}[X_2] + \ldots + \mathrm{E}[X_n]\right) \\ &= \frac{1}{n}\left(\mu + \ldots + \mu\right) = \frac{1}{n} \cdot n\mu = \mu \end{aligned} \tag{3.2}$$

budući su sve X_i jednako distribuirane, s očekivanjem μ. Nadalje, budući su X_i i međusobno nezavisne imamo:

$$\begin{aligned} \mathrm{Var}[\bar{X}] &= \mathrm{Var}\left[\frac{1}{n}(X_1 + X_2 + \ldots + X_n)\right] \\ &= \frac{1}{n^2}\left(\mathrm{Var}[X_1] + \mathrm{Var}[X_2] + \ldots + \mathrm{Var}[X_n]\right) = \frac{1}{n^2} \cdot n\sigma^2 = \frac{\sigma^2}{n} \end{aligned} \tag{3.3}$$

Konačno, imamo da je

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

Pokazuje se da svojstva (3.2) i (3.3), vrijede i ako X, odnosno X_i, imaju neku drugu razdiobu (dakle, ne nužno normalnu) s očekivanjem μ i (konačnom) varijancom σ^2.

Bi li aritmetička sredina uzorka $\bar{X}$ mogla biti dobar **procjenitelj** za očekivanje μ? Oboje su srednje vrijednosti, pa je to prirodna ideja. Odgovor je: da.

Svojstvo (3.2) je zapravo provjera da je $\bar{X}$ *nepristran procjenitelj* - očekivanje procjenitelja je jednako parametru kojeg procjenjuje. Iz (3.3) lagano slijedi da je

$$\lim_{n\to\infty} \mathrm{Var}[\bar{X}] = 0 \tag{3.4}$$

što je zapravo provjera da je $\bar{X}$ *konzistentan procjenitelj* (ima minimalnu varijancu). Zbog tih "lijepih" svojstava, $\bar{X}$ je dobar izbor za procjenitelja od μ. Kažemo da je $\bar{X}$ točkasti procjenitelj za μ.

Slično se može pokazati da je **uzoračka varijanca**

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

nepristran i konzistentan procjenitelj za varijancu σ^2. Kažemo da je S^2 točkasti procjenitelj za σ^2.

Problem točkastih procjenitelja je što je teško odrediti *pouzdanost* njihove procjene. Zato je ponekad zgodno i potrebno raditi **intervalne procjene** (procjene u obliku intervala).

Kažemo da je $[L, D]$ $(1-\alpha)\cdot 100\%$ **pouzdan interval za parametar** ξ, ako su L i D slučajne varijable ovisne o slučajnom uzorku $X_1,\ X_2, \ldots, X_n$, te ako vrijedi

$$P(L \leq \xi \leq D) = 1-\alpha, \quad \alpha \in (0,1).$$

3.2 Statistički testovi

Mnoge situacije povezane 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 osigurava li proces proizvodnje zahtjevanu kvalitetu ili ne. Pri tom se naravno pretpostavlja da obilježje X koje karakterizira kvalitetu pojedinog proizvoda (količina određenog sastojka npr.) ima slučajni karakter.

Riječ je zapravo 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. Želimo testirati je li 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**. Alternativna hipoteza je tvrdnja *suprotna* nultoj hipotezi; ona je njena negacija. Budući iz alternativne hipoteze $\mu \neq 1000h$ slijedi da može biti $\mu > 1000h$ ili $\mu < 1000h$, kažemo da je H_1 **dvostrana alternativna hipoteza**.

Ponekad je zgodnije imati **jednostranu alternativnu hipotezu**:

$$H_1 : \mu > 1000h \quad \text{ili} \quad H_1 : \mu < 1000h.$$

Kako znamo da je alternativna hipoteza tvrdnja suprotna nultoj hipotezi, tako onda uz alternativnu hipotezu $H_1 : \mu > 1000h$ stoji nulta u obliku $H_0 : \mu \leq 1000h$, a uz alternativnu $H_1 : \mu < 1000h$ nulta u obliku $H_0 : \mu \geq 1000h$. Tako ih zaista i shvaćamo i interpretiramo, no uvijek pišemo $H_0 : \mu = 1000h$.

Kada izaberemo dvostranu alternativnu hipotezu, kažemo da provodimo **dvostrani test**, a kada izaberemo jednostranu kažemo da provodimo **jednostrani test**.

Prilikom donošenja odluke o istinitosti hipoteze, postoji mogućnost pogreške, tj. donošenja krive odluke. To je jedan od razloga zašto se nikad ne kaže "prihvaćamo hipotezu", već "ne možemo ju odbaciti". Dvije su vrste mogućih pogrešaka:

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

★ **pogreška 2.vrste**: nismo odbacili nultu hipotezu ako je ona neistinita

Svi mogući ishodi testa prikazani su u donjoj tablici.

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

Vjerojatnost pogreške prve vrste α, uz koju provodimo test, nazivamo **nivo signifikantnosti** ili **razina značajnosti**:

α = P(pogreška 1.vrste)= P(odbacujemo $H_0 \mid H_0$ istinita)

Pomoću vjerojatnosti pogreške druge vrste β izražavamo **snagu testa** $1-\beta$.

β = P(pogreška 2.vrste)= P(ne odbacujemo $H_0 \mid H_0$ neistinita)

1-β=P(odbacujemo $H_0 \mid H_0$ neistinita)

Postupak provođenja statističkog testa

1. Najprije postavljamo HIPOTEZE (odnosno biramo odgovarajući test)

   H_0: nulta hipoteza
   H_1: alternativna hipoteza

   Nultu hipotezu smatramo istinitom dok ne uspijemo dokazati da je alternativna hipoteza istinita.

   Test završava zaključkom: ODBACUJEMO nultu hipotezu H_0 (tada zaključujemo da je istinita altenativna hipoteza H_1) **ili** NE MOŽEMO ODBACITI nultu hipotezu H_0 (tada ne možemo zaključiti da nulta hipoteza H_0 nije istinita, odnosno, uvjetno rečeno, možemo zaključiti da je istinita).

2. TEST STATISTIKA (TS) - svaki test ima test-statistiku čiju vrijednost treba izračunati na osnovi podataka.

3. KRITIČNO PODRUČJE (KP) - svaki test ima kritično područje. Kritično područje je interval (ili unija intervala).

4. PROVJERA je li vrijednost test-statistike ušla u kritično područje, tj. je li $TS \in KP$.

   Ako je $TS \in KP$, odnosno ako je vrijednost test-statistike ušla u kritično područje, tada nultu hipotezu H_0 odbacujemo.

   Ako je $TS \notin KP$, odnosno ako vrijednost test-statistike nije ušla u kritično područje, tada nultu hipotezu H_0 ne možemo odbaciti.

3.3 Test i pouzdani interval za očekivanje

U ovom poglavlju govorit ćemo o testu o očekivanju, te pouzdanom intervalu za očekivanje, uz različite početne pretpostavke.

$(1-\alpha)100\%$ pouzdan interval za očekivanje μ je takav interval unutar kojeg se prava vrijednost parametra μ nalazi s vjerojatnošću $1-\alpha$. Forme tog intervala razlikovat će se, ovisno o početnim pretpostavkama. Pouzdani interval vrlo je srodan dvostranom testu za očekivanje.

Nadalje, željet ćemo testirati je li očekivanje μ jednako nekom unaprijed zadanom broju $\mu_0 \in \mathbb{R}$. Nulta hipoteza je

$$H_0 : \mu = \mu_0,$$

dok za alternativnu hipotezu možemo izabrati bilo koju od sljedeće tri:

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

Zbog različitih početnih uvjeta, ono što će se od slučaja do slučaja razlikovati jest test-statistika.

3.3.1 Normalna populacija i poznata varijanca

Neka je X normalno distribuirana slučajna varijabla s nepoznatim očekivanjem μ i poznatom varijancom σ^2 i neka je $X_1, X_2, \ldots, X_n$ slučajni uzorak duljine n.

Znamo da ako je X normalno distribuirana, onda za aritmetičku sredinu uzorka $\bar{X}$ vrijedi

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

a odatle slijedi

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

TEST O OČEKIVANJU μ

Ako je nulta hipoteza $H_0 : \mu = \mu_0$ istinita, tada je $E[\bar{X}] = \mu = \mu_0$, odnosno

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

Ovako definiranu slučajnu varijablu $Z \sim N(0,1)$ koristit ćemo kao test-statistiku u ovom testu. Test provodimo uz razinu značajnosti α. Promotrimo redom slučajeve različitog izbora alternativne hipoteze.

(1)

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

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 *ne odbacimo* H_0 ako je ona istinita, a gdje je

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

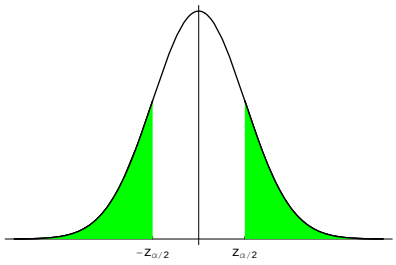
S druge strane,

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

je vjerojatnost da *odbacimo* H_0 ako je ona istinita (vidi sliku 3.1). Dakle,

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

Slika 3.1: Kritično područje dvostranog testa



Područje obojeno zelenom bojom na slici 3.1 naziva se **kritično područje**. Ukupna površina kritičnog područja uvijek je jednaka razini značajnosti α uz koju se test provodi.

(2)

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

Kritično područje površine α u ovom slučaju cijelo je na desnoj strani (slika 3.2).

Nultu hipotezu H_0 odbacujemo ako je $Z > z_\alpha$

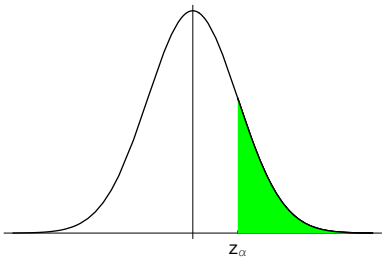
(3)

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

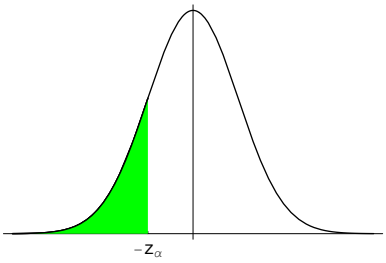
Kritično područje površine α sada je cijelo na lijevoj strani (slika 3.3).

Nultu hipotezu H_0 odbacujemo ako je $Z < -z_\alpha$

Slika 3.2: Kritično područje jednostranog (desno) testa



Slika 3.3: Kritično područje jednostranog (lijevo) testa



POUZDANI INTERVAL za očekivanje μ

Znamo da je u ovom slučaju $\bar{X}^* \sim N(0,1)$, pa stoga vrijedi (vidi dvostrani test o očekivanju):

$$P(-z_{\frac{\alpha}{2}} \leq \bar{X}^* \leq z_{\frac{\alpha}{2}}) = 1-\alpha.$$

Odatle lako slijedi:

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

pa konačno

$(1-\alpha)\cdot 100\%$ pouzdan interval za očekivanje

normalna populacija + varijanca poznata

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

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:*

$$208,\ 216,\ 215,\ 228,\ 210,\ 224,\ 212,\ 213,\ 224,\ 218,$$

$$206,\ 209,\ 208,\ 218,\ 220,\ 206,$$

možemo li uz razinu značajnosti 0.01 zaključiti da je došlo do pada srednjeg napona u elektičnoj mreži?

Rješenje: Zadano je:

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

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 vrijedi li. 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: $Z = \dfrac{\bar{X} - \mu_0}{\sigma}\sqrt{n}$

$$\mu_0 = 220, \qquad \bar{x} = 214.6875$$
$$\Rightarrow \; z = \frac{214.6875 - 220}{6}\sqrt{16} = -3.54167$$
$$z_\alpha = z_{0.01} = 2.326$$
$$\Longrightarrow \; z < -z_{0.01}$$

$\Longrightarrow$ odbacujemo nultu hipotezu H_0, odnosno možemo zaključiti da je došlo do pada napona! ■

Zadatak 30. *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 na osnovi kojeg 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 kako bi se artimetička sredina uzorka $\bar{X}$ razlikovala od sredine μ za manje od 10h s vjerojatnošću 0.99?

Rješenje: $X \sim N(\mu, 40^2)$

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

$$\Phi(z_{\frac{\alpha}{2}}) = \Phi(z_{0.005}) = 1 - \frac{\alpha}{2} = 0.995 \;\Rightarrow\; z_{0.005} = 2.58$$

99% pouzdan interval za očekivanje:

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

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

$$P(-10<\bar{X}-\mu<10)=P\left(-\frac{10}{\frac{\sigma}{\sqrt{n}}}<\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}<\frac{10}{\frac{\sigma}{\sqrt{n}}}\right)$$

$$=P\left(-\frac{10\sqrt{n}}{40}<\bar{X}^*<\frac{10\sqrt{n}}{40}\right)=0.99$$

$$\Rightarrow\ \Phi\left(\frac{\sqrt{n}}{4}\right)-\Phi\left(-\frac{\sqrt{n}}{4}\right)=0.99\ \Leftrightarrow\ 2\Phi\left(\frac{\sqrt{n}}{4}\right)=1.99\ \Leftrightarrow\ \Phi\left(\frac{\sqrt{n}}{4}\right)=0.995$$

$$\Rightarrow\ \frac{\sqrt{n}}{4}=2.58\ \Leftrightarrow\ \sqrt{n}=10.32\ \Rightarrow n=106.5$$

$\Longrightarrow\ n\geq 107,$ tj. treba uzeti uzorak duljine barem 107.

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

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

a odatle:

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

Kako nas zanima $P(|\bar{X}-\mu|<10)=0.99$, 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:

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

■

3.3.2 Normalna populacija i nepoznata varijanca

Neka je X normalno distribuirana slučajna varijabla s nepoznatim očekivanjem μ i nepoznatom varijancom σ^2 i neka je $X_1, X_2, \ldots, X_n$ slučajni uzorak duljine n.

Kao u prethodnom potpoglavlju, prirodno bi bilo krenuti od

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

No, ovdje se sada pojavljuje problem u tome što ne znamo σ, odnosno σ^2.

Kada nam je u statistici nešto nepoznato, onda to *procijenimo*. Varijancu σ^2 procjenjujemo pomoću njenog nepristranog i konzistentnog procjenitelja - uzoračke varijance:

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

Tako dobivamo *aproksimaciju* slučajne varijable $\bar{X}^*$ u obliku $\frac{\bar{X} - \mu}{S}\sqrt{n}$, koja više nema jediničnu normalnu razdiobu.

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

$$T = \frac{\bar{X} - \mu_0}{S}\sqrt{n} \sim t(n-1)$$

ima **Studentovu** ili **t-razdiobu s** $(n-1)$ **stupnjeva slobode**.

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 $t(n-1) \approx N(0,1)$

TEST O OČEKIVANJU μ

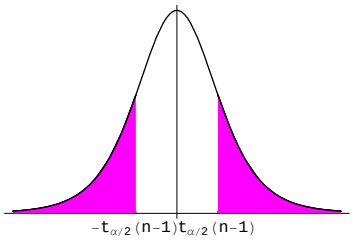
Gore definiranu slučajnu varijablu T koristit ćemo kao test-statistiku. Promotrimo različite izbore alternativnih hipoteza.

(1) $H_0 : \mu = \mu_0, \qquad H_1 : \mu \neq \mu_0$

Nultu hipotezu H_0 odbacujemo ako je (vidi sliku 3.4)

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

Slika 3.4:

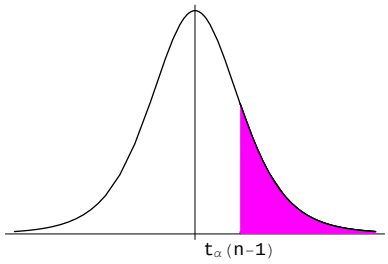


(2) $H_0 : \mu = \mu_0, \qquad H_1 : \mu > \mu_0$

Nultu hipotezu H_0 odbacujemo ako je (vidi sliku 3.5)

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

Slika 3.5:

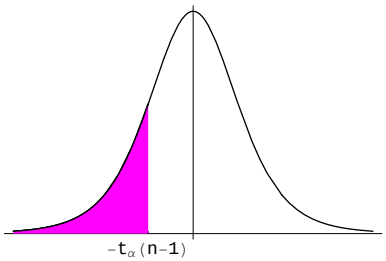


(3) $H_0 : \mu = \mu_0, \qquad H_1 : \mu < \mu_0$

Nultu hipotezu H_0 odbacujemo ako je (vidi sliku 3.6)

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

Slika 3.6:



POUZDANI INTERVAL za očekivanje μ

Vrijedi (vidi sliku 3.4):

$$P\left(-t_{\frac{\alpha}{2}}(n-1) \leq \frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}} \leq t_{\frac{\alpha}{2}}(n-1)\right) = 1-\alpha,$$

a odatle

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

čime dobivamo

$(1-\alpha)\cdot 100\%$ pouzdan interval za očekivanje

normalna populacija + varijanca nepoznata

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

Zadatak 31. *Tvornica tvrdi da je prosječan vijek trajanja njenih baterija 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. Može li se, uz razinu značajosti $\alpha = 0.05$, zaključiti da 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: $T = \dfrac{\bar{X} - \mu_0}{S}\sqrt{n} \sim t(n-1)$

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

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

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

$t_{0.05}(5) = 2.015$

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

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

Zadatak 32. *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):* $\bar{x} = 1173.6, \quad s = 36.3$. *Procijenite očekivanje vijeka trajanja tih mehaničkih komponenti* 95% *pouzdanim intervalom (uz pretpostavku da je vrijeme trajanja mehaničkih komponenti normalno distribuirano).*

Rješenje:

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

■

3.3.3 Veliki uzorak

Neka je X slučajna varijabla nepoznate razdiobe s nepoznatim očekivanjem μ i nepoznatom, ali konačnom varijancom σ^2, te neka je $X_1, X_2, \ldots, X_n$ slučajni uzorak velike duljine n ($n \to \infty$).

Situacija je slična kao u prethodnom poglavlju: standardna devijacija σ nam je nepoznata, pa ju procijenimo uzoračkom standardnom devijacijom S. No, bitna razlika je u tome što ovdje radimo s **velikim uzorkom**. Posljedica toga (Centralni granični teorem!) je da slučajna varijabla $\frac{\bar{X}-\mu}{S}\sqrt{n}$ ima (aproksimativno) jediničnu normalnu razdiobu (bez obzira na to koju razdiobu ima promatrano obilježje X!), odnosno

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

TEST O OČEKIVANJU μ

Ako je nulta hipoteza $H_0 : \mu = \mu_0$ ($\mu_0 \in \mathbb{R}$) istinita, tada vrijedi

$$Z = \frac{\bar{X}-\mu_0}{S}\,\sqrt{n} \;\sim\; N(0,1),$$

te tako definiranu slučajnu varijablu Z koristimo kao test-statistiku u ovom slučaju.

Kritična područja, odnosno kriteriji odbacivanja nulte hipoteze $H_0 : \mu = \mu_0$ isti su kao u prvom slučaju (kada pretpostavljamo da imamo normalno distribuirano obilježje i poznatu varijancu); vidi slike 3.1, 3.2 i 3.3. Razlog tome je što u oba ova slučaja (prvom i trećem) test-statistika ima *jediničnu normalnu razdiobu*, dok se u drugom slučaju (normalno distribuirano obilježje i nepoznata varijanca) pojavljuje *Studentova ili t-razdioba*. Kritično područje uvijek ovisi o razdiobi test-statistike.

POUZDANI INTERVAL za očekivanje μ

U ovom slučaju vrijedi:

$$P\left(-z_{\frac{\alpha}{2}} \leq \frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}} \leq z_{\frac{\alpha}{2}}\right) = 1-\alpha,$$
$$\Rightarrow P\left(\bar{X} - z_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}}\right) = 1-\alpha,$$

čime dobivamo

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

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

Zadatak 33. *Zoolog želi procijeniti očekivanu količinu šećera u krvi određene životinjske vrste nastale nakon ubrizgavanja određene količine adrenalina. Dobivena aritmetička sredina uzorka od 55 životinja je 126.9 mg uz standardnu devijaciju uzorka od 10.5 mg. a) Odredite 90% pouzdan interval za očekivanje. b) Može li se, uz razinu značajnosti $\alpha = 0.01$, zaključiti da je prosječna količina šećera u krvi ovih životinja bila veća od 128 mg?*

Rješenje: a) Imamo: $n = 55, \quad \bar{x} = 126.9, \quad s = 10.5$. Nadalje,

$$1 - \alpha = 0.9 \Leftrightarrow \alpha = 0.1 \Leftrightarrow \frac{\alpha}{2} = 0.05$$
$$\Phi(z_{0.05}) = 0.95 \Rightarrow z_{0.05} = 1.65,$$

pa konačno

$$126.9 \pm 1.65 \cdot \frac{10.5}{\sqrt{55}} = 126.9 \pm 2.34 \implies 124.56 \leq \mu \leq 129.24.$$

b) Želimo provjeriti je li prosječna količina šećera u krvi bila veća od 128, pa postavljamo hipoteze

$$H_0 : \mu = \mu_0 = 128$$
$$H_1 : \mu > 128$$

Test-statistika koju ćemo koristiti je:

$$Z = \frac{\bar{X} - \mu_0}{S} \sqrt{n} \sim N(0, 1),$$

a njena vrijednost na ovom uzorku je:

$$z = \frac{126.9 - 128}{10.5}\sqrt{55} = -0.777.$$

Kako je

$$z_\alpha = z_{0.01} = 2.326,$$

očito je

$$z = -0.777 < z_{0.01} = 2.326,$$

pa stoga nultu hipotezu ne možemo odbaciti, odnosno ne možemo zaključiti da je prosječna količina šećera u krvi bila veća od 128. Primijetimo da zapravo uopće nismo morali koristiti tablice: kritično područje je ovdje cijelo na desnoj strani (nad pozitivnim dijelom x-osi), a test-statistika je negativnog predznaka, pa očito nije moguće da je ušla u kritično područje. ■

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

Pretpostavimo da mjerimo isto statističko obilježje X u dvije različite populacije, te da je u obje te populacije X normalno distribuirana slučajna varijabla s jednakom varijancom σ^2.

Označimo s $X^{(1)}$ statističko obilježje X u populaciji 1, te s $X^{(2)}$ statističko obilježje X u populaciji 2. Imamo:

$$X^{(1)} \sim N(\mu_1, \sigma^2) \quad \text{i} \quad X^{(2)} \sim N(\mu_2, \sigma^2).$$

Po pretpostavci, varijanca σ^2 od $X^{(1)}$ i $X^{(2)}$ je jednaka; ono što nas zanima je jesu li možda i očekivanja μ_1 i μ_2 jednaka.

Iz obje populacije uzimamo uzorak; iz populacije 1 uzorak $X_1^{(1)}, X_2^{(1)}, \ldots, X_{n_1}^{(1)}$ za obilježje $X^{(1)}$ duljine n_1, te iz populacije 2 uzorak $X_1^{(2)}, X_2^{(2)}, \ldots, X_{n_2}^{(2)}$ za obilježje $X^{(2)}$ duljine n_2. Duljine n_1 i n_2 ne moraju biti jednake.

Želimo testirati sljedeću nultu hipotezu

$$H_0 : \mu_1 = \mu_2.$$

Za alternativnu hipotezu možemo izabrati jednu od sljedeće tri:

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

Koristimo sljedeću 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

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

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

i pritom $\bar{X}_1$ i S_1^2 predstavljaju aritmetičku sredinu i uzoračku varijancu prvog uzorka, a $\bar{X}_2$ i S_2^2 drugog. S^2 se interpretira kao **zajednička varijanca uzoraka** 1 i 2.

Ako je nulta hipoteza $H_0 : \mu_1 = \mu_2$ istinita, tada je

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

odnosno test-statistika T ima Studentovu ili t-razdiobu s $n_1 + n_2 - 2$ stupnjeva slobode.

Razmotrimo slučajeve različitih izbora alternativne hipoteze, te pripadne kriterije odbacivanja nulte hipoteze. Kako test-statistika T ima Studentovu razdiobu, kritična područja se formiraju slično kao u slučaju testa o očekivanju uz pretpostavku da je obilježje normalno distribuirano, a varijanca nepoznata.

(1) $H_0 : \mu_1 = \mu_2, \qquad H_1 : \mu_1 \neq \mu_2$

Nultu hipotezu H_0 odbacujemo ako (vidi sliku 3.4)

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

(2) $H_0 : \mu_1 = \mu_2, \qquad H_1 : \mu_1 > \mu_2$

Nultu hipotezu H_0 odbacujemo ako (vidi sliku 3.5)

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

(3) $H_0 : \mu_1 = \mu_2, \qquad H_1 : \mu_1 < \mu_2$

Nultu hipotezu H_0 odbacujemo ako (vidi sliku 3.6)

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

Zadatak 34. *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 s 10 zagorskih stabala bio je: 36, 25, 21, 29, 30, 36, 27, 28, 30, 37. Može li se, uz razinu značajnosti 0.01, zaključiti da jabuke u Zagorju daju veći prinos, ako je poznato da je prinos normalna slučajna varijabla. Može li se, 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

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

$$\bar{x}_1 = \frac{1}{7}(28 + 26 + 33 + 29 + 31 + 27 + 28) = 28.857$$
$$\bar{x}_2 = \frac{1}{10}(36 + 25 + 21 + 29 + 30 + 36 + 27 + 28 + 30 + 37) = 29.9$$
$$s_1^2 = \frac{1}{6} \cdot 34.855 = 5.81, \quad s_2^2 = \frac{1}{9} \cdot 240.9 = 26.767$$

$$s^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = \frac{6\cdot 5.81 + 9\cdot 26.767}{7+10-2} = 18.3842$$
$$\Rightarrow \; s = 4.2877$$

$$t = \frac{28.857 - 29.9}{4.2877\sqrt{\frac{1}{7}+\frac{1}{10}}} = -0.4936$$
$$t_\alpha(n_1+n_2-2) = t_{0.01}(15) = 2.602$$
$$\Rightarrow t > -t_{0.01}(15)$$

Odatle zaključujemo da ne možemo odbaciti H_0, tj. da ne možemo zaključiti da jabuke u Zagorju daju veći prinos.
Ako želimo testirati jesu li 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. ■

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

Pretpostavimo da mjerimo isto statističko obilježje X u dvije različite populacije, te da je u obje te populacije X normalno distribuirana slučajna varijabla.

Označimo s $X^{(1)}$ statističko obilježje X u populaciji 1, te s $X^{(2)}$ statističko obilježje X u populaciji 2. Imamo:

$$X^{(1)} \sim N(\mu_1, \sigma_1^2) \quad \text{i} \quad X^{(2)} \sim N(\mu_2, \sigma_2^2).$$

Iz populacije 1 uzimamo uzorak $X_1^{(1)}, X_2^{(1)}, \ldots, X_{n_1}^{(1)}$ duljine n_1, a iz populacije 2 uzorak $X_1^{(2)}, X_2^{(2)}, \ldots, X_{n_2}^{(2)}$ duljine n_2. Duljine n_1 i n_2 ne moraju biti jednake.

Želimo testirati sljedeću nultu hipotezu

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

Za alternativnu hipotezu možemo izabrati jednu od sljedeće tri:

$$H_1 : \sigma_1^2 \neq \sigma_2^2, \quad H_1 : \sigma_1^2 > \sigma_2^2 \quad \text{ili} \quad H_1 : \sigma_1^2 < \sigma_2^2.$$

Koristimo sljedeću test-statistiku

$$F = \frac{S_1^2}{S_2^2}.$$

Ako je nulta hipoteza $H_0 : \sigma_1^2 = \sigma_2^2$ istinita, tada

$$F \ \sim \ F(n_1 - 1,\ n_2 - 1)$$

odnosno test-statistika F ima **Fisherovu ili F-razdiobu s (uređenim) parom stupnjeva slobode** $(n_1 - 1,\ n_2 - 1)$.

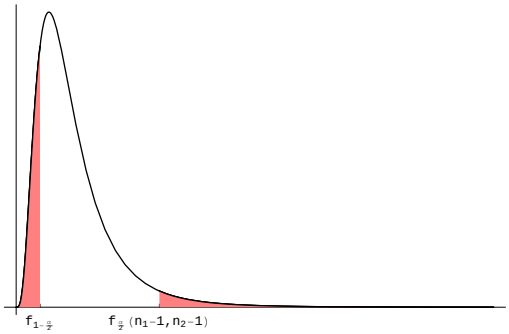
Razmotrimo sada različite slučajeve izbora alternativne hipoteze, te pripadne kriterije odbacivanja nulte hipoteze.

(1) $H_0 : \sigma_1^2 = \sigma_2^2, \qquad H_1 : \sigma_1^2 \neq \sigma_2^2$

Nultu hipotezu H_0 odbacujemo ako (vidi sliku 3.7)

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

Slika 3.7:

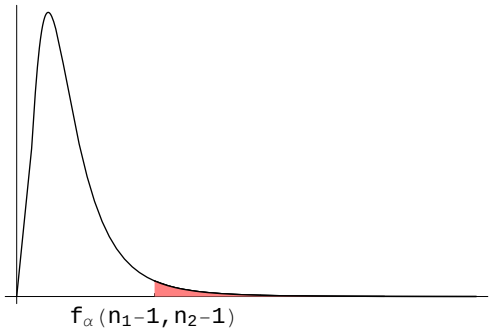


(2) $H_0 : \sigma_1^2 = \sigma_2^2, \qquad H_1 : \sigma_1^2 > \sigma_2^2$

Nultu hipotezu H_0 odbacujemo ako (vidi sliku 3.8)

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

Slika 3.8:

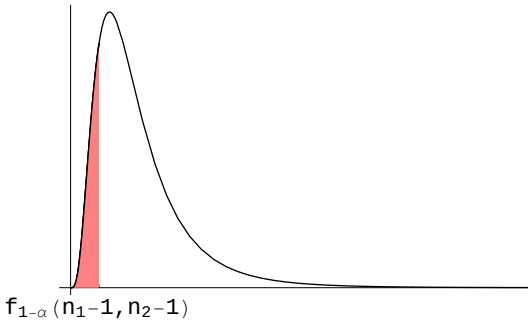


(3) $H_0 : \sigma_1^2 = \sigma_2^2, \qquad H_1 : \sigma_1^2 < \sigma_2^2$

Nultu hipotezu H_0 odbacujemo ako (vidi sliku 3.9)

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

Slika 3.9:



Vrijedi

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

Zadatak 35. *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. Može li se, uz razinu značajnosti 0.02, zaključiti da su varijance težina u ta dva razreda 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. ■

3.6 Usporedba očekivanja više normalno distribuiranih populacija (jednofaktorska analiza varijance ANOVA)

Jednofaktorska analiza varijance - skraćeno ANOVA (ANalysis Of VAriance) - je test koji koristimo za usporedbu očekivanja *barem dvije* normalno distribuirane populacije. Za usporedbu očekivanja *točno dvije* normalno distribuirane populacije koristimo `t-test` (iako, naravno, i u tom slučaju možemo primijeniti ANOVA-u).

Pretpostavimo da mjerimo isto statističko obilježje X u k $(k \geq 2)$ različitih populacija, te da je u svim promatranim populacijama X

normalno distribuirana slučajna varijabla s jednakom varijancom σ^2 (ista pretpostavka kao kod t-testa!).

Označimo s $X^{(i)}$ statističko obilježje X u populaciji i $(i = 1, 2, \ldots, k)$. Imamo:

$$X^{(i)} \sim N(\mu_i, \sigma^2), \quad i = 1, 2, \ldots, k.$$

Uzimamo k nezavisnih slučajnih uzoraka $X_{i1}, X_{i2}, \ldots, X_{in_i}$ $(i = 1, 2, \ldots, k)$, po jedan iz svake populacije; iz i-te populacije za obilježje $X^{(i)}$ duljine n_i $(i = 1, 2, \ldots, k)$.

Želimo testirati nultu hipotezu

$$H_0 \ : \ \mu_1 = \mu_2 = \ldots = \mu_k,$$

odnosno hipotezu da *nema razlike u očekivanjima* među populacijama. Alternativna hipoteza je tada naravno da *razlika postoji*, točnije, da se barem dvije populacije razlikuju po očekivanjima

$$H_1 \ : \ \exists\, i, j, \ \ i \neq j \ \text{ takvi da } \ \mu_i \neq \mu_j.$$

Primijetimo da je alternativna hipoteza ovdje po prvi put jedinstvena (nemamo mogućnosti izbora).

Za test-statistiku trebaju nam:

★ aritmetička sredina svakog od uzoraka $(i = 1, 2, \ldots k)$

$$\bar{X}_i = \frac{1}{n_i}(X_{i1} + \ldots + X_{in_i})$$

★ uzoračka varijanca svakog od uzoraka $(i = 1, 2, \ldots k)$

$$S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2$$

★ ukupna aritmetička sredina svih podataka

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij} = \frac{1}{n} \sum_{i=1}^{k} n_i \bar{X}_i, \qquad n = \sum_{i=1}^{k} n_i$$

★ suma kvadrata odstupanja srednjih vrijednosti uzoraka od ukupne sredine, odnosno suma kvadrata u odnosu na tretman ("**S**um of **S**quares due to **T**reatment")

$$SST = \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2 = \sum_{i=1}^{k} n_i\bar{X}_i^2 - n\bar{X}^2$$

★ suma kvadrata pogrešaka ("**S**um of **S**quares due to **E**rror")

$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2 = \sum_{i=1}^{k}(n_i - 1)S_i^2$$
$$= \sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}^2 - \sum_{i=1}^{k} n_i\bar{X}_i^2$$

★ srednjekvadratno odstupanje među uzorcima (zbog razlike u tretmanima)

$$MST = \frac{SST}{k-1}$$

★ srednjekvadratna pogreška

$$MSE = \frac{SSE}{n-k}$$

Konačno, test-statistika je oblika

$$F = \frac{MST}{MSE}.$$

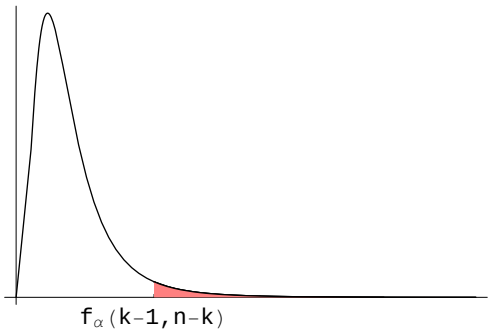
Ako je nulta hipoteza H_0 istinita, tada je

$$F \sim F(k-1,\ n-k),$$

odnosno test-statistika ima Fisherovu ili F-razdiobu s parom stupnjeva slobode $(k-1,\ n-k)$.

Nultu hipotezu H_0 odbacujemo ako

$$F > f_\alpha(k-1,\ n-k).$$



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 36. *Pivovara koristi 3 različite linije za punjenje limenki piva. Postoji sumnja da se srednji neto sadržaj limenki razlikuje od linije do linije. Na slučajan način izabrano je 5 limenki sa svake linije, te je izmjeren njihov neto sadržaj. Može li se, uz razinu značajnosti 0.05, zaključiti da postoji značajna razlika u prosječnim neto sadržajima limenki po linijama?*

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 prosječnih neto sadržaja limenki po linijama. Budući imamo 3 populacije (=linije), t-test nam ne može pomoći, već moramo provesti ANOVA-u. Krenimo redom:

$$k = 3, \quad n_1 = n_2 = n_3 = 5, \quad n = \sum_{i=1}^{3} n_i = 15$$

$$\begin{aligned}
\bar{x}_1 &= \frac{1}{5}(3.633 + 3.651 + 3.66 + 3.645 + 3.654) = 3.6486 \\
\bar{x}_2 &= \frac{1}{5}(3.615 + 3.627 + 3.636 + 3.63 + 3.624) = 3.6264 \\
\bar{x}_3 &= 3.633 \\
\bar{x} &= \frac{1}{15}\sum_{i=1}^{3}\sum_{j=1}^{5} x_{ij} = \frac{1}{15}\sum_{i=1}^{3} n_i \cdot \bar{x}_i = \frac{1}{3}\sum_{i=1}^{3} \bar{x}_i = 3.636 \\
SST &= \sum_{i=1}^{3} n_i \bar{X}_i^2 - n\bar{X}^2 = 5\sum_{i=1}^{3} \bar{x}_i^2 - 15\bar{x}^2 = 0.0013 \\
SSE &= \sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}^2 - \sum_{i=1}^{k} n_i \bar{X}_i^2 = \sum_{i=1}^{3}\sum_{j=1}^{5} x_{ij}^2 - 5\sum_{i=1}^{3} \bar{x}_i^2 = 0.00086 \\
MST &= \frac{SST}{k-1} = \frac{0.0013}{2} = 0.00065
\end{aligned}$$

$$MSE = \frac{SSE}{n-k} = \frac{0.00086}{15-3} = 0.000072$$

Konačno dobivamo vrijednost test-statistike:

$$\Rightarrow \quad f = \frac{MST}{MSE} = \frac{0.00065}{0.000072} = 9.02778.$$

Iz tablice za F-razdiobu očitamo:

$$f_\alpha(k-1,\ n-k) = f_{0.05}(2,\ 12) = 3.89$$

Kako je

$$f > f_{0.05}(2, 12)$$

vidimo da je vrijednost test-statistike ušla u kritično područje, što znači da nultu hipotezu o jednakosti očekivanja moramo odbaciti. Zaključujemo stoga da `postoji` značajna razlika među prosječnim neto sadržajima limenki 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		

■

3.7 Test i pouzdani interval za proporciju p binomne razdiobe

Test i pouzdani interval koji ćemo predstaviti u ovom poglavlju zapravo su specijalan slučaj testa i pouzdanog intervala za očekivanje

na osnovi velikih uzoraka, no zbog svoje važnosti zaslužuju posebno poglavlje.

Pretpostavimo da promatramo obilježje X koje je binomno distribuirano, odnosno neka je $X \sim B(n,p)$, te neka je $X_1, X_2, \ldots, X_n$ uzorak velike duljine n ($n \to \infty$). **Oprez!** Oznaka n koristi se i kao oznaka za duljinu uzorka i kao oznaka za parametar binomne razdiobe, što su općenito različite stvari i njihove vrijednosti nisu nužno uvijek jednake. Srećom, u kontekstu ovog testa one jesu jednake; pokus se sastoji od prebrojavanja predmeta/jedinki u uzorku koje zadovoljavaju određeno svojstvo. Vrijednost parametra n, dakle, poznata je iz konteksta.

Parametar koji je najčešće nepoznat, a zanima nas, jest parametar p, kojeg nazivamo i **proporcija**. Za *procjenitelja* parametra p uzet ćemo

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

Ovo je dobar izbor procjenitelja budući je to nepristran i konzistentan procjenitelj. Provjerimo! Znamo da je $\mathrm{E}[X] = np$ i $\mathrm{Var}[X] = npq,\ q = 1-p$. Imamo:

$$\begin{aligned} \mathrm{E}[\hat{p}] &= \mathrm{E}\left[\frac{X}{n}\right] = \frac{1}{n}\mathrm{E}[X] = \frac{1}{n}\cdot np = p \\ \mathrm{Var}[\hat{p}] &= \mathrm{Var}\left[\frac{X}{n}\right] = \frac{1}{n^2}\mathrm{Var}[X] = \frac{1}{n^2}\cdot npq = \frac{pq}{n} \end{aligned}$$

i očito $\lim_{n\to\infty} \mathrm{Var}[\hat{p}] = 0$.

Budući po pretpostavci radimo s velikim uzorkom ($n \to \infty$), te smo u uvjetima Centralnog graničnog teorema, vrijedi

$$\hat{p}^* = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} \sim N(0,1).$$

TEST O PROPORCIJI p

Želimo testirati je li proporcija p jednaka nekom unaprijed zadanom broju $p_0 \in (0,1)$. Nulta hipoteza je

$$H_0 : p = p_0,$$

dok za alternativnu možemo izabrati bilo koju od sljedeće tri:

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

Ako je nulta hipoteza $H_0 : p = p_0$ istinita, onda vrijedi

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n} \sim N(0,1).$$

Ovako definiranu slučajnu varijablu Z koristit ćemo kao test-statistiku. Kako test-statistika Z ima jediničnu normalnu razdiobu, kritična područja ista su kao kod testa o očekivanja na osnovi velikog uzorka (čega je ovo, ponovimo još jednom, specijalan slučaj), odnosno za normalno distribuirano obilježje s poznatom varijancom. Dajmo ipak i kratak pregled različitih mogućnosti izbora alternativne hipoteze:

(1) $H_0 : p = p_0, \qquad H_1 : p \neq p_0$

Nultu hipotezu H_0 odbacujemo ako je $Z > z_{\frac{\alpha}{2}}$ ili $Z < -z_{\frac{\alpha}{2}}$ (vidi sliku 3.1).

(2) $H_0 : p = p_0, \qquad H_1 : p > p_0$

Nultu hipotezu H_0 odbacujemo ako je $Z > z_{\alpha}$ (vidi sliku 3.2).

(3) $H_0 : p = p_0, \qquad H_1 : p < p_0$

Nultu hipotezu H_0 odbacujemo ako je $Z < -z_{\alpha}$ (vidi sliku 3.3).

POUZDANI INTERVAL za proporciju p

Vrijedi

$$P(-z_{\frac{\alpha}{2}} \leq \frac{\hat{p}-p}{\sqrt{\frac{pq}{n}}} \leq z_{\frac{\alpha}{2}}) = 1-\alpha,$$

pa odatle

$$P\left(\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{pq}{n}} \leq p \leq \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{pq}{n}}\right) = 1-\alpha.$$

Primijetimo da smo dobili pouzdan interval za p čiji rubovi uključuju vrijednost od p - koju ne znamo i želimo procijeniti. No, za velike n, dobit ćemo dovoljno dobre rezultate ako p zamijenimo s $\hat{p}$, pa tako dobivamo:

$(1-\alpha)\cdot 100\%$ pouzdan interval za proporciju p
(na osnovi velikog uzorka)

$$\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Zadatak 37. *Proizvođač tvrdi da njegove pošiljke sadrže najviše 7% defektnih proizvoda. Uzet je slučajan uzorak od 200 komada iz jedne pošiljke i u njemu je nađeno 11 defektnih proizvoda. Možemo li, uz razinu značajnosti 0.05, zaključiti da proizvođač govori istinu?*

Rješenje: Postavljamo hipoteze:

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

Kad 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 bi u danom kontekstu bilo još i bolje. Izračunajmo vrijednost odgovarajuće test-statistike:

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

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

Zadatak 38. *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 kako bi s vjerojatnošću 0.95 mogli tvrditi da se najmanje 50% kućanstava služi Internetom?
Rješenje:

$a)$ $\hat{p} = 0.55, \quad n = 100$
$$1 - \alpha = 0.95 \;\Leftrightarrow\; \alpha = 0.05 \;\Leftrightarrow\; \frac{\alpha}{2} = 0.025$$
$$z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$$
$$\hat{p} \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.55 \pm 1.96\sqrt{\frac{0.55 \cdot 0.45}{100}} = 0.55 \pm 0.09751$$
$$\implies 0.4525 \leq p \leq 0.6475$$

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

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

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

$$\Longrightarrow n \geq 381$$

■

3.8 Usporedba proporcija

Pretpostavimo da mjerimo isto statističko obilježje X u dvije različite populacije, te da je u obje te populacija X slučajna varijabla koja ima binomnu razdiobu. Označimo s $X^{(1)}$ slučajnu varijablu koja predstavlja X u populaciji 1, a s $X^{(2)}$ slučajnu varijablu koja predstavlja X u populaciji 2. Imamo:

$$X^{(1)} \sim B(n_1,\ p_1) \quad \text{i} \quad X^{(2)} \sim B(n_2,\ p_2).$$

Pritom je istovremeno n_1 duljina uzorka uzeta iz populacije 1, a n_2 duljina uzorka uzeta iz populacije 2. Pretpostavljamo da su uzorci velike duljine ($n_1 \to \infty,\ n_2 \to \infty$), te da su međusobno *nezavisni*.

Želimo testirati sljedeću nultu hipotezu

$H_0 : p_1 = p_2.$

Za alternativnu hipotezu možemo izabrati jednu od sljedeće tri:

$H_1 : p_1 \neq p_2, \quad H_1 : p_1 > p_2 \quad \text{ili} \quad H_1 : p_1 < p_2.$

Koristimo sljedeću test-statistiku

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

gdje su s $\hat{p}_1$ i $\hat{p}_2$ označeni procjenitelji parametara p_1 i p_2 (vjerojatnosti uspjeha), dok je

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

procjena zajedničke vjerojatnosti uspjeha.

Ako je nulta hipoteza $H_0 : p_1 = p_2$ istinita, tada je

$$Z \sim N(0,1).$$

Sada imamo:

(1) $H_0 : p_1 = p_2, \qquad H_1 : p_1 \neq p_2$

Nultu hipotezu H_0 odbacujemo ako je $Z > z_{\frac{\alpha}{2}}$ ili $Z < -z_{\frac{\alpha}{2}}$ (vidi sliku 3.1).

(2) $H_0 : p_1 = p_2, \qquad H_1 : p_1 > p_2$

Nultu hipotezu H_0 odbacujemo ako je $Z > z_\alpha$ (vidi sliku 3.2).

(3) $H_0 : p_1 = p_2, \qquad H_1 : p_1 < p_2$

Nultu hipotezu H_0 odbacujemo ako je $Z < -z_\alpha$ (vidi sliku 3.3).

Zadatak 39. *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. Može li se, uz razinu značajnosti 0.05, zaključiti da*

a) postoji razlika među županijama

b) tog kandidata više "vole" u županiji A.

Rješenje:

$$n_1 = 300, \quad \hat{p}_1 = 0.56$$

$$n_2 = 200, \quad \hat{p}_2 = 0.48$$

$a)$ $H_0 : p_1 = p_2$

$H_1 : p_1 \neq p_2$

$$\begin{aligned}
\hat{p} &= \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{300 \cdot 0.56 + 200 \cdot 0.48}{500} = 0.528 \\
z &= \frac{0.56 - 0.48}{\sqrt{0.528 \cdot 0.472}} \cdot \frac{1}{\sqrt{\frac{1}{300} + \frac{1}{200}}} = 1.75 \\
z_{\frac{\alpha}{2}} &= z_{0.025} = 1.96 \\
&\Longrightarrow \; z < z_{0.025}
\end{aligned}$$

$\Longrightarrow$ 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. ■

3.9 χ^2 - test o prilagodbi modela podacima

χ^2 - test o prilagodbi modela podacima je test koji primjenjujemo kada želimo provjeriti *imaju li podaci neku unaprijed pretpostavljenu razdiobu*. Pritom je izbor te moguće razdobe naravno šarolik.

Nulta hipoteza ovog testa je

H_0 : podaci odgovaraju pretpostavljenom modelu,

drugim riječima,

H_0 : podaci imaju pretpostavljenu razdiobu.

Alternativna hipoteza je time i ovdje jednoznačno određena

H_1 : podaci ne odgovaraju pretpostavljenom modelu.

Test-statistika je oblika

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

gdje su f_i eksperimentalne, a

$f_i' = np_i$

teorijske frekvencije.

Ako je nulta hipoteza H_0 istinita, tada za velike n $(n \to \infty)$, gdje je n duljina uzorka, vrijedi

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

gdje $\chi^2(k-r-1)$ označava χ^2–**razdiobu s k-r-1 stupnjeva slobode**.

Pritom je

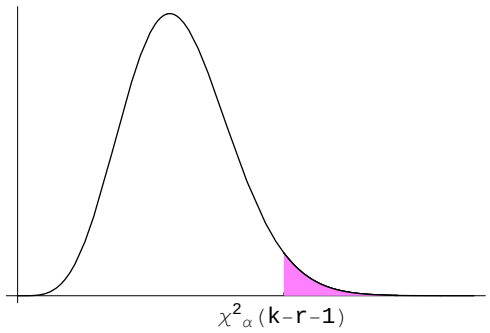
k = (konačan) broj razreda

r = broj nepoznatih (procijenjenih) parametara

Nultu hipotezu H_0 odbacujemo ako

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

Slika 3.10:



Zadatak 40. *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. Može li se, uz razinu značajnosti 0.05, zaključiti da je proizvođač u pravu?*

Rješenje: Proizvođač tvrdi da njegovi proizvodi imaju navedenu 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 $\chi^2_\alpha(k-r-1)$; pritom je α zadana razina značajnosti $(= 0.05)$, $k = 3$ (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 ušla u kritično područje, moramo odbaciti nultu hipotezu. Drugim riječima, odbacujemo tvrdnju proizvođača da se kvaliteta njegovih proizvoda podvrgava navedenoj razdiobi, tj. proizvođač nije u pravu. ■

Zadatak 41. *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, može li se zaključiti da broj pisama koji se dobije u jednom bacanju ima binomnu radiobu?*

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 razdiobu. Neka slučajna varijabla X broji koliko puta je palo pismo. Pokus koji izvodimo (ponavljamo ga 100 puta, dakle $n = 100$) je bacanje novčića 5 puta, a događaj koji tretiramo kao "uspjeh" je "palo

je pismo". Parametar n binomne razdiobe od X je stoga jednak 5. Parametar p nije zadan, te ga moramo procijeniti. Oprez! n sada označava i duljinu uzorka i parametar razdiobe, no to su različite stvari i različite vrijednosti, pa na to treba pripaziti.

Parametar p jednak je vjerojatnosti "uspjeha" u jednom bacanju novčića. Njegovu procjenu dobijemo tako da ukupan broj palih pisama podijelimo s 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 vjerojatnosti 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 bismo tako opet dobili razred čija je teorijska frekvencija strogo manja od 5, postupak bismo ponavljali dok ne bismo dobili razred s (ukupnom) teorijskom frekvencijom ≥ 5. Nakon toga formiramo novu tablicu:

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

Vrijednost test-statistike je dakle

$$h=\sum_{i=1}^{4}\frac{(f_i-f_i')^2}{f_i'}=1.4563.$$

Konačan broj razreda je $k=4$, a broj procijenjenih parametara je $r=1$ (procijenili smo parametar p; da je bilo rečeno da bacamo simetrične novčiće, tada bismo imali $p=1/2$, pa taj parametar ne bi bilo potrebno procjenjivati, odnosno bilo bi $r=0$). Iz tablice očitavamo

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

Kako je

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

što znači da vrijednost test-statistike nije ušla u kritično područje, ne možemo odbaciti nultu hipotezu, odnosno ne možemo zaključiti da se ne radi o binomnoj razdiobi. ■

Zadatak 42. *Anketirano je 100 studenata vezano za broj njihovih odlazaka u kazalište tijekom godine; podaci su u tablici. Može li se, uz razinu značajnosti 0.05, zaključiti da se radi o uzorku iz normalno distribuirane populacije?*

broj posjeta	$[-0.5, 1.5]$	$[1.5, 3.5]$	$[3.5, 5.5]$	$[5.5, 7.5]$	$[7.5, 9.5]$	$[9.5, 11.5]$
broj studenata	6	11	21	33	19	10

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

U tablici su prikazani podaci sortirani u razrede. Kao "predstavnike" podataka iz svakog pojedinog razreda uzimamo sredinu tog razreda (kao što smo radili u deskriptivnoj statistici; vidi Zadatak 1). Sada

$$\begin{aligned}\hat{\mu} &= \frac{1}{100}(0.5 \cdot 6 + 2.5 \cdot 11 + 4.5 \cdot 21 + 6.5 \cdot 33 + 8.5 \cdot 19 + 10.5 \cdot 10) = 6.06 \\ \hat{\sigma}^2 &= \frac{1}{99}(0.5^2 \cdot 6 + 2.5^2 \cdot 11 + 4.5^2 \cdot 21 + 6.5^2 \cdot 33 + 8.5^2 \cdot 19 + 10.5^2 \cdot 10 \\ &\quad -100 \cdot 6.06^2) = 6.996\end{aligned}$$

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

$$X \sim N(6.06,\ 6.996)$$

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

$$
\begin{aligned}
p_1 &= P(-0.5 \le X \le 1.5) = P\left(\frac{-0.5-6.06}{\sqrt{6.996}} \le X^* \le \frac{1.5-6.06}{\sqrt{6.996}}\right) \\
&= \Phi(-1.72) - \Phi(-2.48) = \Phi(2.48) - \Phi(1.72) \\
&= 0.99343 - 0.95728 = 0.03615 \\
&\Longrightarrow \quad f_1' = 3.615 \\
p_2 &= P(1.5 \le X \le 3.5) = P\left(\frac{1.5-6.06}{\sqrt{6.996}} \le X^* \le \frac{3.5-6.06}{\sqrt{6.996}}\right) \\
&= \Phi(-0.97) - \Phi(-1.72) = \Phi(1.72) - \Phi(0.97) \\
&= 0.95728 - 0.83397 = 0.12331 \\
&\Longrightarrow \quad f_2' = 12.331 \\
p_3 &= P(3.5 \le X \le 5.5) = P\left(-0.97 \le X^* \le -0.21\right) = \Phi(-0.21) - \Phi(-0.97) \\
&= \Phi(0.97) - \Phi(0.21) = 0.83397 - 0.58317 = 0.2508 \\
&\Longrightarrow \quad f_3' = 25.08 \\
p_4 &= P(5.5 \le X \le 7.5) = P\left(-0.21 \le X^* \le 0.54\right) = \Phi(0.54) - \Phi(-0.21) \\
&= 0.70540 - 1 + 0.58317 = 0.28857 \\
&\Longrightarrow \quad f_4' = 28.857 \\
p_5 &= P(7.5 \le X \le 9.5) = P\left(0.54 \le X^* \le 1.30\right) = \Phi(1.3) - \Phi(0.54) \\
&= 0.90320 - 0.70540 = 0.1978 \\
&\Longrightarrow \quad f_5' = 19.78 \\
p_6 &= P(9.5 \le X \le 11.5) = P\left(1.30 \le X^* \le 2.06\right) = \Phi(2.06) - \Phi(1.3) \\
&= 0.98030 - 0.90320 = 0.0771 \\
&\Longrightarrow \quad f_6' = 7.71
\end{aligned}
$$

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

i	1	2	3	4	5
f_i	**17**	21	33	19	10
f_i'	**15.946**	25.08	28.857	19.78	7.71
$\frac{(f_i-f_i')^2}{f_i'}$	0.07	0.66	0.6	0.03	0.68

Vrijednost test-statistike je prema tome

$$h = \sum_{i=1}^{5} \frac{(f_i - f_i')^2}{f_i'} = 2.04,$$

a

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

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

Zadatak 43. *(DZ) Bilježen je broj četvorki rođenih u nekoj županiji tijekom 70 godina. Podaci su dani u tablici. Može li se, uz razinu značajnosti 0.05, zaključiti da podaci dolaze iz populacije s Poissonovom razdiobom?*

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

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

3.10 χ^2 - test nezavisnosti dviju varijabli

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

$$\mathrm{Im}X = \{a_1, \ldots, a_r\}, \qquad \mathrm{Im}Y = \{b_1, \ldots, b_s\}$$
$$\Longrightarrow\ \mathrm{Im}(X, Y) = \{(a_i, b_j) : i = 1, 2, \ldots, r,\ \ j = 1, 2, \ldots, s\}$$

Nadalje, neka je:

- f_{ij} frekvencija (a_i, b_j) u uzorku
- $f_i = \sum_{j=1}^{s} f_{ij}$ (marginalna) frekvencija a_i u uzorku
- $g_j = \sum_{i=1}^{r} f_{ij}$ (marginalna) frekvencija b_j u uzorku

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

Konačno, uvedimo oznake:

$$\begin{aligned} p_{ij} &= P(X = a_i,\ Y = b_j) \qquad (i = 1, 2, \ldots, r, \quad j = 1, 2, \ldots, s) \\ p_i &= P(X = a_i) \qquad (i = 1, 2, \ldots, r) \\ q_j &= P(X = b_j) \qquad (j = 1, 2, \ldots, s). \end{aligned}$$

Želimo testirati nultu hipotezu

$H_0 \ : \ p_{ij} = p_i \cdot q_j, \quad \forall\, i, j$

odnosno

H_0 : X i Y su međusobno nezavisne slučajne varijable

Alternativna hipoteza je ponovo jednoznačno određena

$H_1 \ : \ \exists\, i, j, \ \ i \neq j$ takvi da $p_{ij} \neq p_i \cdot q_j$

odnosno

H_0 : X i Y *nisu* međusobno nezavisne slučajne varijable

Ako je nulta hipoteza H_0 istinita, kao procjene za p_i i q_j možemo uzeti:

$$\hat{p_i} = \frac{f_i}{n} \quad \text{i} \quad \hat{q_j} = \frac{g_j}{n}.$$

Očekivane vrijednosti f'_{ij} od f_{ij} (teorijske frekvencije) su tada:

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

Koristimo test-statistiku

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

Nultu hipotezu o nezavisnosti odbacujemo ako (vidi sliku 3.10)

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

Zadatak 44. *U cilju ispitivanja sklonosti potrošača nekom proizvodu, uzet je uzorak na temelju kojeg su dobiveni podaci dani u tablici. Može li se, uz razinu značajnosti 0.05, na osnovi ovih podataka zaključiti da sklonost potrošača tom proizvodu ne ovisi o njihovom dohotku?*

mjesečni dohodak anketiranih kupaca (u kn)	*sklonost potrošnji*		
	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 $i = 1, 2, 3, 4,\ j = 1, 2, 3$, pa u tu svrhu 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 imamo:

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

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

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

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

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

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

Kako bismo lakše izračunali vrijednost test-statistike, zgodno je, radi preglednosti, eksperimentalnim frekvencijama u tablici 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 ušla 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. ■

3.11 χ^2 - test homogenosti populacija

Pretpostavimo da nas zanima razdioba istog diskretnog statističkog obilježja X u raznim populacijama. Točnije, zanima nas je li razdioba od X jednaka u svim populacijama, odnosno jesu li populacije *homogene* s obzirom na X.

Promatramo m različitih populacija. Iz svake uzimamo uzorak duljine n_i $(i = 1, 2, \ldots, m)$; uzorci su međusobno nezavisni.

S $X^{(i)}$ označimo slučajnu varijablu koja predstavlja X u i-toj populaciji ($i = 1, 2, \ldots, m$). Nulta hipoteza je da su sve $X^{(i)}$ jednake po distribuciji, a alternativna je da postoji barem jedna koja se po distribuciji razlikuje od ostalih, odnosno

$$H_0 \ : \ X^{(1)} \stackrel{D}{=} X^{(2)} \stackrel{D}{=} \ldots \stackrel{D}{=} X^{(m)}$$
$$H_1 \ : \ \exists\, i, j, \ \ i \neq j \quad \text{tako da} \quad X^{(i)} \stackrel{D}{\neq} X^{(j)}$$

Kako su $X^{(i)}$ po pretpostavci diskretne slučajne varijable, njihov zakon razdiobe možemo prikazati sljedećom tablicom

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

pa sada H_0 možemo zapisati i ovako

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

gdje je $p_j = P(X = a_j)$, što će reći da p_j predstavljaju zajedničke (tj. po populacijama jednake) vjerojatnosti da X poprimi vrijednost a_j.

Uvedimo oznake:

- f_{ij} frekvencija a_j u uzorku iz i-te populacije
- $f_j = \sum_{i=1}^{m} f_{ij}$ frekvencija a_j u svim uzorcima zajedno
- $n_i = \sum_{j=1}^{k} f_{ij}$ duljina uzorka iz i-te populacije

Frekvencijska tablica:

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

Ako je nulta hipoteza H_0 istinita, kao procjenu zajedničkih vjerojatnosti p_j možemo uzeti:

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

i tada su očekivane (teorijske) frekvencije f'_{ij} od a_{ij}:

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

Koristimo test-statistiku:

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

Nultu hipotezu o homogenosti populacija odbacujemo ako (vidi sliku 3.10)

$$H > \chi^2_\alpha((m-1)(k-1))$$

Zadatak 45. *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), onda je potrebno provjeriti ima li promatrano obilježje X jednaku distribuciju u svih tih 5 populacija, tj. dana. To ćemo provjeriti χ^2-testom o homogenosti populacija. Hipoteze su:

H_0 : X ima istu vjerojatnosnu razdiobu u svih 5 populacija, tj. $X^{(1)} \stackrel{D}{=} X^{(2)} \stackrel{D}{=} X^{(3)} \stackrel{D}{=} X^{(4)} \stackrel{D}{=} X^{(5)}$

H_1 : X *nema* istu vjerojatnosnu razdiobu u svih 5 populacija

Kako bismo izračunali vrijednost odgovarajuće test-statistike, potrebne su nam teorijske frekvencije f'_{ij}, pa pogledajmo najprije kolike su duljine uzoraka n_i iz svake od populacija $(i = 1, 2, 3, 4, 5)$, te

marginalne frekvencije f_j svake od vrijednosti a_j koje X može poprimiti $(j = 1, 2, 3)$.

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

$$f'_{21} = \frac{n_2 \cdot f_1}{n} = f'_{31} = \frac{n_3 \cdot f_1}{n} = f'_{41} = \frac{n_4 \cdot f_1}{n} = f'_{51} = \frac{n_5 \cdot f_1}{n} = f'_{11},$$

odnosno

$$f'_{11} = f'_{21} = f'_{31} = f'_{41} = f'_{51} = 60.8.$$

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 ušla u kritično područje, pa nultu hipotezu ne možemo odbaciti. Dakle, ne možemo zaključiti da populacije nisu homogene, što znači da promatrano statističko obilježje (= broj pokvarenih televizora) ima jednaku distribuciju u svim populacijama (= u svim danima). ■

Poglavlje 4

LINEARNI REGRESIJSKI MODEL

4.1 Linearna regresija

Pretpostavimo da želimo odrediti postoji li veza između dvije varijable: x i y. Pritom je x nezavisna (neslučajna, kontrolirana) varijabla, dok je y zavisna (slučajna) varijabla. Kako bismo naglasili razliku između njih, nezavisnu varijablu označavamo malim slovom (x), a zavisnu velikim (Y). [1]

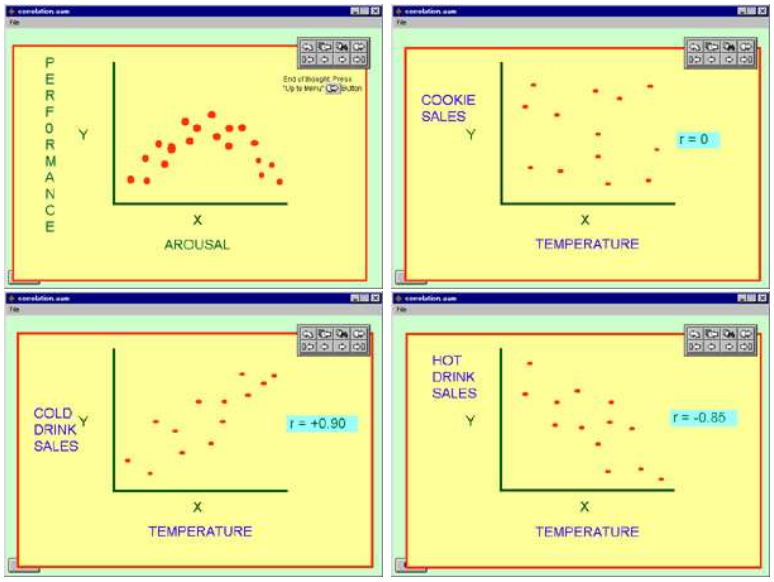
Funkcija koja opisuje vezu između x i Y može biti različitih ti-

[1] *Odakle naziv "**regresija**"?* Sir Francis Galton je u svom radu "Regression Toward Mediocrity in Hereditary Sature" iz 1885. istraživao povezanost visina očeva i sinova i došao je do zaključka da su sinovi vrlo visokih (niskih) očeva uglavnom viši (niži) od prosjeka, ali ipak niži (viši) od svojih očeva - visina sinova teži prema srednjoj vrijednosti, a ne prema još većim (manjim) vrijednostima $\Rightarrow$ regresija!

pova. Može biti npr. polinom, eksponencijalna funkcija, logaritamska funkcija, trigonometrijska funkcija, itd. Tip veze koji će nas ovdje zanimati je **linearna veza**.

Pretpostavimo da imamo n parova podataka $(x_i, Y_i),\ i = 1, 2, \ldots, n$. Dobivene podatke potrebno je najprije prikazati grafički (**dijagram raspršenja**), kako bi se sa slike vidjelo postoji li eventualno veza među varijablama, te kojeg je ona tipa; vidi sliku 4.1.

Slika 4.1:



Na prvoj gornjoj slici, veza među prikazanim varijablama je očito kvadratna funkcija - kroz točke je moguće provući parabolu koja će dosta dobro opisivati povezanost varijabli. Na drugoj gornjoj slici, veza ne postoji - točke su jako raspršene. Konačno, na donje dvije slike, veza među varijablama je očito **linearna**; graf pripadajuće funkcije je **pravac** $Y = ax + b$. Želimo naći

JEDNOSTAVNI LINEARNI REGRESIJSKI MODEL

$$Y_i = a\, x_i + b + \varepsilon_i, \quad i = 1, 2, \ldots, n,$$

gdje su a i b nepoznati parametri, x_i vrijednosti nezavisne, a Y_i vrijednosti zavisne varijable. Slučajne varijable ε_i su slučajne greške. Pretpostavljamo da su one međusobno nezavisne, te da vrijedi $\varepsilon_i \sim N(0, \sigma^2)$, $i = 1, ..., n$.

Sada treba naći "dobre" procjenitelje $\hat{a}$ i $\hat{b}$ za parametre a i b. Nepristrane i konzistentne procjenitelje za a i b dobivamo metodom najmanjih kvadrata i oni su oblika:

$$\hat{a} = \frac{S_{xy}}{S_x^2}, \qquad \hat{b} = \bar{Y} - \hat{a}\bar{x},$$

pri čemu je S_{xy} **kovarijanca**:

$$S_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y}) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i Y_i - n\bar{x}\bar{Y}\right)$$

Nepristran i konzistentan procjenitelj za varijancu σ^2 slučajnih grešaka ε_i je

$$\hat{\sigma}^2 = \frac{SSE}{n-2}$$

pri čemu je

$$\begin{aligned} SSE &= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}\left(Y_i - (\hat{a}x_i + \hat{b})\right)^2 \\ &= (n-1)\left(S_y^2 - \hat{a}^2 S_x^2\right) = (n-1)\left(S_y^2 - \hat{a}S_{xy}\right) \end{aligned}$$

Sada znamo kako "točkasto" procijeniti parametre a i b, no uvijek je korisno imati i intervale procjene, te mogućnost testiranja hipoteza vezanih uz vrijednosti parametara a i b.

$(1-\alpha)100\%$ pouzdan interval za parametar a

$$\hat{a} - t_{\frac{\alpha}{2}}(n-2) \cdot \frac{\hat{\sigma}}{\sqrt{(n-1)S_x^2}} \leq a \leq \hat{a} + t_{\frac{\alpha}{2}}(n-2) \cdot \frac{\hat{\sigma}}{\sqrt{(n-1)S_x^2}}$$

$(1-\alpha)100\%$ pouzdan interval za parametar b

$$\hat{b} - t_{\frac{\alpha}{2}}(n-2) \cdot \hat{\sigma} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1)S_x^2}} \leq b \leq \hat{b} + t_{\frac{\alpha}{2}}(n-2) \cdot \hat{\sigma} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1)S_x^2}}$$

Želimo li testirati nultu hipotezu:

$H_0 : a = a_0 \ (a_0 \in \mathbb{R})$

pri čemu za alternativnu hipotezu možemo izabrati bilo koju od:

$H_1 : a \neq a_0, \quad H_1 : a > a_0 \quad$ ili $\quad H_1 : a < a_0$

koristimo test-statistiku

$$T_a = \frac{\hat{a} - a_0}{\hat{\sigma}} \sqrt{(n-1)S_x^2} \overset{H_0}{\sim} t(n-2)$$

Želimo li testirati nultu hipotezu:

$H_0 : b = b_0 \ (b_0 \in \mathbb{R})$

pri čemu za alternativnu hipotezu možemo izabrati bilo koju od:

$H_1 : b \neq b_0, \quad H_1 : b > b_0 \quad$ ili $\quad H_1 : b < b_0$

koristimo test-statistiku

$$T_b = \frac{\hat{b} - b_0}{\hat{\sigma} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1)S_x^2}}} \overset{H_0}{\sim} t(n-2)$$

Kako u oba testa test-statistika ima Studentovu ili t-razdiobu, kritična područja, tj. kriteriji odbacivanja nulte hipoteze, formiraju se slično kao kod t-testa i testa o očekivanju za normalno distribuiranu populaciju s nepoznatom varijancom (vidi slike 3.4, 3.5 i 3.6). Razlika je jedino u broju stupnjeva slobode.

Procjena linearnog modela može se dobiti za svaki set podataka, no pitanje je *koliko je takav model dobar za te podatke*. **Pokazatelji je li nađeni linearni model dobar (prihvatljiv) model za dane podatke** su:

- **test značajnosti linearnog modela** - svodi se na testiranje

$$H_0 \ : \ a = 0, \quad H_1 \ : \ a \neq 0$$

- **koeficijent determinacije**

$$R^2 = 1 - \frac{SSE}{(n-1)S_y^2} \ \in \ [0,1]$$

  Što je vrijednost R^2 bliže 1, to je prilagodba linearnog modela podacima bolja. Koeficijent determinacije jednak je kvadratu koeficijenta korelacije (vidi sljedeće poglavlje).

Zadatak 46. *U tablici su dani podaci o broju emitiranih reklama tijekom 8 uzastopnih mjeseci za neki proizvod i ostvarenoj zaradi na tom proizvodu.*

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

a) Odredite procjenu pravca regresije za ove podatke.

b) Odredite 95% *pouzdane intervale za* a *i* b.

c) Uz razinu značajnosti 0.05, testirajte hipotezu da je koeficijent smjera tog pravca jednak 0, tj. da izmedu x i Y ne postoji linearna veza.

d) Je li predloženi linearni model dobar za dane podatke?

Rješenje: **a)** Procijenimo najprije parametre a i b:

$$\hat{a} = \frac{S_{xy}}{S_x^2}, \qquad \hat{b} = \bar{Y} - \hat{a}\bar{x}.$$

$$\bar{x} = \frac{1}{8}\sum_{i=1}^{8} x_i = \frac{1}{8}(16+59+65+43+82+90+31+22) = 51$$

$$\bar{y} = \frac{1}{8}\sum_{i=1}^{8} y_i = \frac{1}{8}(18+63+28+71+85+98+20+25) = 51$$

$$\sum_{i=1}^{8} x_i^2 = 16^2+59^2+65^2+43^2+82^2+90^2+31^2+22^2 = 26080$$

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

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

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

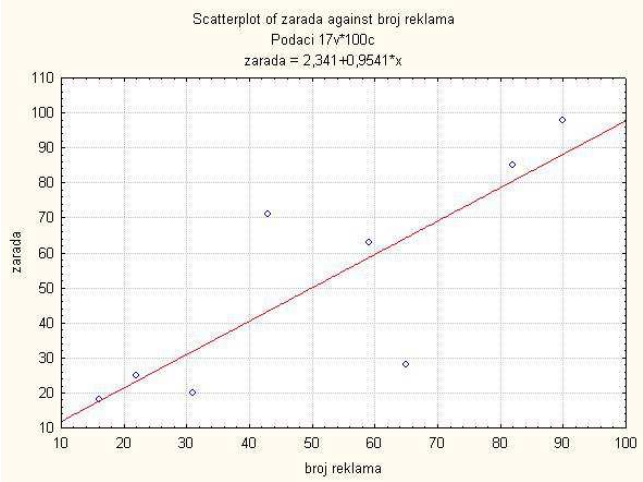
$$\Longrightarrow \hat{a} = \frac{s_{xy}}{s_x^2} = \frac{718.571}{753.143} = 0.954$$

$$\Longrightarrow \hat{b} = \bar{y} - \hat{a}\bar{x} = 51 - 0.954\cdot 51 = 2.346$$

$$\Longrightarrow y = 0.954\cdot x + 2.346$$ procjena pravca regresije za ove podatke

b) Moramo naći pouzdane intervale za parametre a i b. U tu svrhu moramo izračunati:

$$\hat{\sigma}^2 = \frac{SSE}{n-2}, \qquad SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$



Znamo da je $\hat{Y}_i = \hat{a}x_i + \hat{b}$, pa onda:

$$\hat{y}_1 = \hat{a}x_1 + \hat{b} = 0.954 \cdot 16 + 2.346 = 17.61$$

$$\hat{y}_2 = \hat{a}x_2 + \hat{b} = 0.954 \cdot 59 + 2.346 = 58.632$$

$$\hat{y}_3 = \hat{a}x_3 + \hat{b} = 0.954 \cdot 65 + 2.346 = 64.356$$

$$\hat{y}_4 = \hat{a}x_4 + \hat{b} = 0.954 \cdot 43 + 2.346 = 43.368$$

$$\hat{y}_5 = \hat{a}x_5 + \hat{b} = 0.954 \cdot 82 + 2.346 = 80.574$$

$$\hat{y}_6 = \hat{a}x_6 + \hat{b} = 0.954 \cdot 90 + 2.346 = 88.206$$

$$\hat{y}_7 = \hat{a}x_7 + \hat{b} = 0.954 \cdot 31 + 2.346 = 31.92$$

$$\hat{y}_8 = \hat{a}x_8 + \hat{b} = 0.954 \cdot 22 + 2.346 = 23.334$$

Formirajmo tablicu:

i	1	2	3	4	5	6	7	8
x_i	16	59	65	43	82	90	31	22
y_i	18	63	28	71	85	98	20	25
$\hat{y}_i$	17.61	58.632	64.356	43.368	80.574	88.206	31.92	23.334
$(y_i - \hat{y}_i)^2$	0.1521	19.08	1321.759	763.53	19.59	95.92	142.09	2.78

Odatle dobivamo:

$$SSE = \sum_{i=1}^{8} (y_i - \hat{y}_i)^2 = 2364.9 \ ,$$

pa onda

$$\hat{\sigma}^2 = \frac{2364.9}{6} = 394.15 \quad \Rightarrow \quad \hat{\sigma} = 19.85$$

SSE smo mogli izračunati i kraćim putem, primjenom formule:

$$SSE = (n-1)(S_y^2 - \hat{a}^2 S_x^2).$$

Tada

$$s_y^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - n \cdot \bar{y}^2\right) = \frac{1}{7}(27972 - 8 \cdot 51^2) = 1023.43$$
$$\Rightarrow \ SSE = 7(1023.43 - 0.954^2 \cdot 753.143) = 2365.88$$

Pogledajmo sada kako izgleda 95% pouzdan interval za a, odnosno b:

$$\hat{a} \pm t_{\frac{\alpha}{2}}(n-2) \cdot \frac{\hat{\sigma}}{\sqrt{(n-1)s_x^2}} = 0.954 \pm t_{0.025}(6) \cdot \frac{19.85}{\sqrt{7 \cdot 753.143}}$$
$$= 0.954 \pm 2.447 \cdot 0.273 = 0.954 \pm 0.668$$
$$\Longrightarrow \quad 0.286 \ \leq \ a \ \leq \ 1.622$$
$$\hat{b} \ \pm \ t_{\frac{\alpha}{2}}(n-2) \cdot \hat{\sigma} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}}$$
$$= 2.346 \ \pm \ t_{0.025}(6) \cdot 19.85 \sqrt{\frac{1}{8} + \frac{51^2}{7 \cdot 753.143}} = 2.346 \pm 38.196$$
$$\Longrightarrow \quad -35.85 \ \leq \ b \ \leq \ 40.542$$

c) Želimo, uz razinu značajnosti 0.05, testirati hipotezu da ne postoji linearna veza između x i Y. Linearna veza ne postoji ako je koeficijent smjera pravca regresije jednak 0. Ako je on različit od 0, bez obzira je li pozitivan (tj. > 0) ili negativan (tj. < 0), linearna veza postoji. Postavljamo stoga hipoteze:

$$H_0 \ : \ a = 0$$
$$H_1 \ : \ a \neq 0$$

Sljedeći korak je izračunati vrijednost odgovarajuće test-statistike:

$$T_a = \frac{\hat{a} - a_0}{\hat{\sigma}} \sqrt{(n-1)S_x^2} \ \sim \ t(n-2)$$

Imamo:

$$t_a = \frac{0.954 - 0}{19.85} \sqrt{7 \cdot 753.143} = 3.4896$$

Iz tablice za t-razdiobu očitavamo: $t_{\frac{\alpha}{2}}(n-2) = t_{0.025}(6) = 2.447$.

Kako je

$$t_a > t_{0.025}(6),$$

vidimo da je vrijednost test-statistike ušla u kritično područje, pa nultu hipotezu $H_0 : a = 0$ možemo odbaciti. Odatle zaključujemo, uz razinu značajnosti 0.05, da koeficijent smjera pravca regresije nije jednak 0, pa stoga linearna veza za ove podatke postoji.

d) Test o značajnosti linearnog modela proveli smo pod c). Ostaje još izračunati koeficijent determinacije za ove podatke. Već smo izračunali $SSE = 2364.9$ i $s_y^2 = 1023.43$.

$$R^2 = 1 - \frac{SSE}{(n-1)S_y^2} = 1 - \frac{2364.9}{7 \cdot 1023.43} = 0.67 \ ,$$

što ponovo ukazuje na to da je za ove podatke dobiveni linearni model relativno dobar. ■

Linearni model najčešće se koristi u dvije svrhe

1. za **predviđanje (procjenu) 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

$$\widehat{\mathrm{E}[Y|x=x_0]} = \hat{a}x_0 + \hat{b}$$

$(1-\alpha)100\%$ pouzdan interval za $\mathrm{E}[Y|x=x_0]$

$$\widehat{\mathrm{E}[Y|x=x_0]} \pm t_{\frac{\alpha}{2}}(n-2)\cdot\hat{\sigma}\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{(n-1)S_x^2}}$$

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

$$\hat{Y} = \hat{a}x_0 + \hat{b}$$

$(1-\alpha)100\%$ pouzdan interval za Y uz $x=x_0$

$$\hat{Y} \pm t_{\frac{\alpha}{2}}(n-2)\cdot\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{(n-1)S_x^2}}$$

Uočimo da je pouzdani interval za Y širi, odnosno manje precizan od pouzdanog intervala za $\mathrm{E}[Y|x=x_0]$.

Zadatak 47. *Prošli mjesec firma "Firma" platila je emitiranje 52 reklame za jedan mliječni proizvod koji proizvodi i prodaje. Na osnovi podataka iz Zadatka 46, nađite 95% pouzdan interval za zaradu te firme. Nađite i 95% pouzdan interval za prosječnu zaradu svih firmi koje prodaju isti proizvod i prošli mjesec su platile emitiranje 52 reklame za taj proizvod.*

Rješenje: Predviđanje zarade firme "Firma" odgovara predviđanju vrijednosti Y za zadani $x_0 = 52$. Imamo:

$$\hat{y} = \hat{a} \cdot 52 + \hat{b} = 0.954 \cdot 52 + 2.346 = 51.954$$

pa onda

$$\hat{Y} \pm t_{\frac{\alpha}{2}}(n-2) \cdot \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)S_x^2}}$$
$$= 51.954 \pm t_{0.025}(6) \cdot 19.85 \sqrt{1 + \frac{1}{8} + \frac{(52-51)^2}{7 \cdot 753.143}} = 51.954 \pm 51.5237$$

Traženi 95% pouzdan interval za Y je

$$0.4303 \leq Y \leq 103.478$$

Procjena prosječne zarade svih firmi koje prodaju isti proizvod odgovara procjeni vrijednosti $\mathrm{E}[Y|x = x_0]$ za zadani $x_0 = 52$. Imamo kao i gore:

$$\widehat{\mathrm{E}[Y|x = x_0]} = \hat{a} \cdot 52 + \hat{b} = 0.954 \cdot 52 + 2.346 = 51.954$$

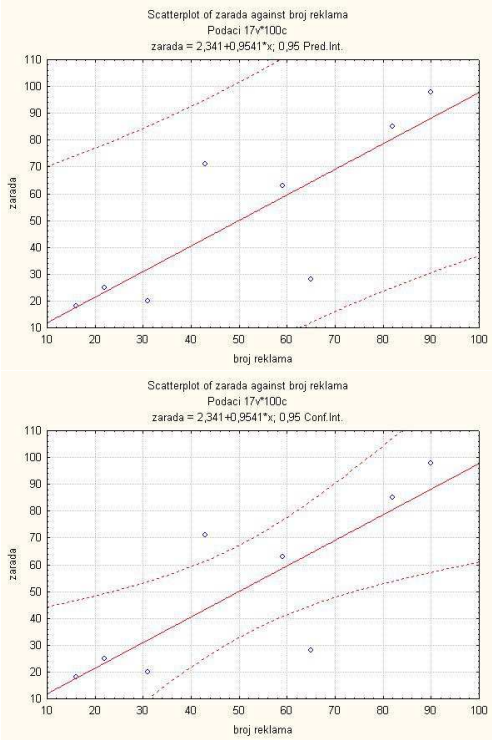
ali sada

$$\widehat{\mathrm{E}\left[Y|x = 52\right]} \pm t_{\frac{\alpha}{2}}(n-2) \cdot \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)S_x^2}}$$
$$= 51.954 \pm t_{0.025}(6) \cdot 19.85 \sqrt{\frac{1}{8} + \frac{(52-51)^2}{7 \cdot 753.143}} = 51.954 \pm 17.1862$$

pa slijedi da je 95% pouzdan interval za $\mathrm{E}\left[Y|x = 52\right]$:

$$34.7678 \leq \mathrm{E}\left[Y|x = 52\right] \leq 69.1402$$

■



4.2 Test koreliranosti dviju varijabli

(Pearsonov) koeficijent korelacije dviju varijabli X i Y definiran je s

$$R = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{S_x}\cdot\frac{Y_i - \bar{Y}}{S_y}\right) = \frac{S_{xy}}{S_x \cdot S_y}, \quad -1 \leq R \leq 1$$

Ukoliko korelacija među varijablama postoji, to ukazuje na moguće postojanje linearne veze među tim varijablama.

Ako je $R > 0$, to znači da je korelacija pozitivna, što će reći da ako X raste, i Y u pravilu raste, odnosno moguće je da postoji rastuća linearna veza.

Ako je $R < 0$, to znači da je korelacija negativna, što će reći da ako X raste, Y u pravilu pada, odnosno moguće je da postoji

padajuća linearna veza.

Ako je pak $R = 0$, to znači da korelacije nema, pa ne postoji ni linearna veza među varijablama.

Primjer 12. *Nađite koeficijent korelacije za podatke iz Zadatka 46.*

Rješenje:

$$s_{xy} = 718.571, \quad s_x^2 = 753.143 \Rightarrow s_x = 27.4435, \quad s_y^2 = 1023.43 \Rightarrow s_y = 31.9911$$
$$\Longrightarrow \; r = \frac{s_{xy}}{s_x \cdot s_y} = \frac{718.571}{27.4435 \cdot 31.9911} = 0.818467$$

što je razmjerno visoka korelacija. ■

Neka je $(X_1, Y_1),\ (X_2, Y_2), \ldots, (X_n, Y_n)$ slučajni uzorak za normalno distribuirano dvodimenzionalno statističko obilježje (X, Y). Želimo testirati nultu hipotezu

$H_0 \ : \ \rho = 0$ (= nema korelacije)

u odnosu na alternativnu

$H_1 \ : \ \rho \neq 0$ (= korelacija postoji)

ili

$H_1 \ : \ \rho > 0$ (= korelacija postoji i pozitivna je)

ili

$H_1 \ : \ \rho < 0$ (= korelacija postoji i negativna je)

Test-statistika koju koristimo je

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

Kritična područja formiraju se analogno kao kod svakog testa u kojem test-statistika ima Studentovu ili t-razdiobu (vidjeti npr. t-test).

Zadatak 48. *U jednom razredu od 30 učenika promatra se ocjena iz matematike (X) i ocjena iz fizike (Y). Uvidom u imenik dobiveni su ovi podaci:* $(1,3),(4,3),(2,2),(3,2),(1,2),(1,1),(2,2),(4,4),(2,2),(3,3),(4,4),(5,5),$ $(3,5),(2,1),(2,3),(2,2),(5,5),(3,3),(2,2),(2,2),(3,3),(3,2),(4,4),(2,2),$ $(3,3),(2,1),(3,2),(3,2),(3,2),(2,2).$

Može li se, uz razinu značajnosti 0.05, zaključiti da ne postoji korelacija između ocjena iz matematike i fizike?

Rješenje: Postoji li korelacija između ocjena iz matematike i fizike ispitat ćemo pomoću testa o koreliranosti dviju varijabli. Neka je X = ocjena iz matematike i Y = ocjena iz fizike. Budući nas zanima samo postoji li korelacije ili ne, a ne i je li ona (ako postoji) pozitivna ili negativna, dovoljno je za alternativnu hipotezu postaviti $H_1 : \rho \neq 0$. Dakle, hipoteze su

$$\begin{aligned} H_0 &: \rho = 0 \\ H_1 &: \rho \neq 0 \end{aligned}$$

Izračunajmo sada vrijednost odgovarajuće test-statistike:

$$\bar{x} = \frac{1}{30}(1+4+2+3+1+1+2+4+2+3+\ldots+3+2) = 2.7$$

$$\bar{y} = \frac{1}{30}(3+3+2+2+1+2+4+2+3+4+\ldots+2+2) = 2.63$$

$$s_x^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) = \frac{1}{29}(251 - 30\cdot 2.7^2) = 1.114 \quad \Rightarrow \quad s_x = 1.056$$

$$s_y^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right) = \frac{1}{29}(245 - 30\cdot 2.63^2) = 1.293 \quad \Rightarrow \quad s_y = 1.137$$

$$s_{xy} = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right) = \frac{1}{29}(239 - 30\cdot 2.7\cdot 2.63) = 0.896$$

$$\Longrightarrow \quad r = \frac{s_{xy}}{s_x \cdot s_y} = \frac{0.896}{1.056 \cdot 1.137} = 0.746$$

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 za Studentovu razdiobu 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 ušla u kritično područje, pa stoga nultu hipotezu odbacujemo. Zaključujemo stoga da korelacija između ocjena iz matematike i fizike *postoji*, odnosno da su varijable X i Y korelirane. ■