

6. Procjene parametara pomoću uzoraka

6.1. Uvod

Jedna od najčešćih matematičkih disciplina koja se koristi u svakodnevnom životu je statistika. Recimo ankete na temelju kojih se predviđaju pobjednici na izborima, ispitivanja kvalitete proizvoda, . . . , sve su to rezultati dobiveni statističkim metodama.

Ako zamislimo, recimo, ankete glasača na izborima, svakom je jasno da ne možemo anketirati **sve** glasače, jer je to preskupo i fizički neostvarivo. Umjesto anketiranja cijele grupe, koja se u statistici zove **populacija**, anketirat ćemo samo mali dio te grupe nazvan **uzorak**.

Populacija može biti konačna ili beskonačna. Na primjer, broj glasača na nekim izborima je primjer konačne populacije, dok bi recimo izvlačenje brojeva iz vrećice (s vraćanjem kad je neki broj izvučen) sve dok u vrećici ima brojeva, dalo beskonačnu populaciju.

Ako uzorak reprezentira populaciju, onda na temelju uzorka možemo zaključiti važne stvari o populaciji. Međutim, ne mora svaki uzorak dobro reprezentirati populaciju. Na primjer, želimo izračunati prosječnu visinu svih studenata nekog sveučilišta. Ako u uzorku izaberemo samo djevojke, svima je jasno da taj uzorak neće dobro reprezentirati sve studente sveučilišta, jer su djevojke, u prosjeku, niže od momaka. Grana statistike koja kaže kako treba birati uzorak da bi se dobila neka informacija o populaciji, zove se **induktivna statistika**. Budući da nikada ne možemo biti apsolutno sigurni da smo dobro obavili uzorkovanje, onda se u zaključcima o populaciji javlja jezik vjerojatnosti (s vjerojatnošću tom i tom je to i to).

Dio statistike koji se bavi samo opisom neke izdvojene grupe, bez da se iz toga izvlači ikakav zaključak o cjelokupnoj populaciji, zove se **opisna ili deduktivna statistika**.

6.2. Definicije osnovnih pojmova

Velikim slovima bez indeksa $X, Y \dots$ označavat ćemo slučajne varijable koje proučavamo. Njihovo očekivanje označavat ćemo s $\mu_X, \mu_Y \dots$, odnosno kad je jednoznačno jasno o kojoj se slučajnoj varijabli radi i samo s μ . Varijancu tih slučajnih varijabli označavat ćemo sa σ_X^2 , odnosno $\sigma_Y^2 \dots$, odnosno ako je jasno o kojoj se slučajnoj varijabli radi samo sa σ^2 .

Nakon neformalnog uvođenja pojma uzorak u uvodu, dajmo matematičku definiciju tog pojma.

Neka je zadana slučajna varijabla Y s funkcijom razdiobe F . Napravimo n njezinih nezavisnih identičnih kopija, koje nazovimo s Y_1, \dots, Y_n s identičnom razdiobom F . Tako dobivenu n -torku slučajnih varijabli nazivamo slučajni uzorak dimenzije n , a vrijednosti koje one poprimaju y_1, \dots, y_n nazivamo vrijednostima uzorka.

Slučajne varijable Y_i možemo interpretirati kao nezavisno ponavljanje istog pokusa u istim uvjetima.

Uzorak možemo uzimati iz konačne ili beskonačne populacije, što se, naravno, razlikuje. U slučaju da kod konačne populacije izvlačimo uzorak i ne vraćamo ga, populaciju možemo iscrpsti, tako da više ne možemo dobiti novi uzorak.

Ako u konačnoj populaciji vraćamo uzorak, tj. možemo ga više puta ponovno izabrati, onda je situacija slična beskonačnoj populaciji, jer ne možemo iscrpsti uzorke.

Na primjer, ako iz kutije od 100 loptica izvlačimo kao uzorak, koji ne vraćamo, njih 10, onda se radi o konačnoj populaciji. S druge strane, ako novčić bacamo 50 puta i brojimo broj palih glava, onda se radi o konačnoj populaciji s ponavljanjem (glave mogu pasti i više puta), što je slično situaciji kad imamo beskonačnu populaciju.

Prvo promatramo jednostavniju situaciju, kad nam je **populacija** iz koje “vaidimo” uzorak **poznata**. Tada možemo nešto reći i o uzorku.

Ako promotrimo sve moguće vrijednosti uzorka duljine n iz neke populacije (svejedno jesu li uzimani s ili bez ponavljanja), onda za svaku od tih vrijednosti uzorka možemo promatrati očekivanje i varijancu. Kako se vrijednosti uzorka mijenjaju, tako se mijenja očekivanje i varijanca svakog od njih.

Međutim, pitanje je kad gledamo očekivanja svih mogućih vrijednosti uzorka, što ćemo dobiti za aritmetičku sredinu očekivanja uzorka. Drugim riječima gledamo

što se događa s očekivanjem slučajne varijable

$$\bar{Y} = \frac{Y_1 + \cdots + Y_k}{k},$$

pri čemu je k broj svih mogućih uzoraka duljine n . Promotrimo to na jednom primjeru.

Primjer 6.2.1. *Populacija se sastoji od 5 brojeva i to: 2, 3, 6, 8 i 11. Iz te populacije izvlačimo sve moguće uzorke duljine 2*

- (a) *s vraćanjem (dakle dozvoljeno nam je izvući jednake brojeve),*
 (b) *bez vraćanja (jednaki brojevi nisu mogući).*

Odredite očekivanje populacije μ_Y i aritmetičku sredinu očekivanja vrijednosti uzoraka $\mu_{\bar{Y}}$.

Očekivanje populacije je

$$\mu_Y = \frac{2 + 3 + 6 + 8 + 11}{5} = 6.$$

- (a) *Ispišimo sad sve moguće vrijednosti uzoraka duljine 2 iz te populacije (s vraćanjem). Ima ih 25, jer prvu vrijednost uzorka možemo uzeti na 5 načina, a to možemo napraviti i za drugu vrijednost uzorka:*

(2, 2)	(2, 3)	(2, 6)	(2, 8)	(2, 11)
(3, 2)	(3, 3)	(3, 6)	(3, 8)	(3, 11)
(6, 2)	(6, 3)	(6, 6)	(6, 8)	(6, 11)
(8, 2)	(8, 3)	(8, 6)	(8, 8)	(8, 11)
(11, 2)	(11, 3)	(11, 6)	(11, 8)	(11, 11)

Nađimo sad očekivanja svake vrijednosti uzorka,

2.0	2.5	4.0	5.0	6.5
2.5	3.0	4.5	5.5	7.0
4.0	4.5	6.0	7.0	8.5
5.0	5.5	7.0	8.0	9.5
6.5	7.0	8.5	9.5	11.0

Srednja vrijednost tih očekivanja je zbroj svih vrijednosti iz prethodne tablice podijeljenih s brojem uzoraka, tj. s 25

$$\mu_{\bar{Y}} = \frac{2.0 + 2.5 + 4.0 + 5.0 + 6.5 + \cdots + 7.0 + 8.5 + 9.5 + 11.0}{25} = 6.$$

Je li to slučajno? Očito nije. Primijetite da se u vrijednostima uzoraka svaki broj pojavljuje točno $2n_p = 10$ puta, pri čemu je n_p veličina populacije. Kad

gledamo očekivanja svake vrijednosti uzorka, onda je to očekivanje na mjestu (i, j) u tablici

$$E_{i,j} = \frac{x_i + x_j}{2}.$$

Kad sad promatram $\mu_{\bar{Y}}$, onda je

$$\begin{aligned} \mu_{\bar{Y}} &= \frac{E_{11} + E_{12} + \cdots + E_{1n_p} + E_{21} + \cdots + E_{n_p n_p}}{n_p^2} \\ &= \frac{(x_1 + x_1) + (x_1 + x_2) + \cdots + (x_1 + x_{n_p}) + (x_2 + x_1) + \cdots + (x_{n_p} + x_{n_p})}{2n_p^2} \\ &= \frac{2n_p(x_1 + x_2 + \cdots + x_{n_p})}{2n_p^2} \\ &= \frac{x_1 + x_2 + \cdots + x_{n_p}}{n_p} = \mu_Y. \end{aligned}$$

(b) U ovoj situaciji, kad ne smijemo vraćati brojeve u populaciju, to odgovara situaciji da istovremeno izvlačimo 2 broja iz vrećice s 5 brojeva, pa nam poredak izvučениh brojeva nije bitan. Takvih uzoraka ima $\binom{n_p}{2}$, što je kod nas 10. Sve moguće vrijednosti uzorka su:

$$(2, 3), (2, 6), (2, 8), (2, 11), (3, 6), (3, 8), (3, 11), (6, 8), (6, 11), (8, 11).$$

Očekivanja svih vrijednosti uzoraka su:

$$2.5, 4.0, 5.0, 6.5, 4.5, 5.5, 7.0, 7.0, 8.5, 9.5,$$

i

$$\mu_{\bar{Y}} = \frac{2.5 + 4.0 + 5.0 + 6.5 + 4.5 + 5.5 + 7.0 + 7.0 + 8.5 + 9.5}{10} = 6.$$

Ponovno, radi li se o čudu? Odgovor je ne! Primijetimo da se svaki od brojeva nalazi točno jednak broj puta u svim mogućim vrijednostima uzoraka. Kako uzoraka ima $\binom{n}{2}$, onda brojeva u uzorcima ima $2\binom{n_p}{2}$, a svaki od brojeva pojavljuje se točno $\frac{2}{n_p}\binom{n_p}{2}$ puta. Sada je očekivanje

$$\mu_{\bar{Y}} = \frac{\frac{2}{n_p}\binom{n_p}{2} \frac{x_1 + \cdots + x_{n_p}}{2}}{\binom{n_p}{2}} = \frac{x_1 + x_2 + \cdots + x_{n_p}}{n_p} = \mu_Y.$$

Na sličan način dokazali bismo i za uzorke nekih drugih duljina n , (ne samo 2).

Bez obzira izvlačimo li uzorak duljine n iz konačne populacije s vraćanjem ili bez vraćanja, ili iz beskonačne populacije, uvijek je srednja vrijednost očekivanja uzorka jednaka očekivanju populacije

$$\mu_{\bar{Y}} = \mu_Y.$$

Sada pogledajmo standardnu devijaciju očekivanja svih mogućih vrijednosti uzoraka. Promotrimo to ponovno na jednom primjeru.

Primjer 6.2.2. *Populacija se sastoji od 5 brojeva i to: 2, 3, 6, 8 i 11. Iz te populacije izvlačimo sve moguće uzorke duljine 2*

- (a) *s vraćanjem (dakle dozvoljeno nam je izvući jednake brojeve),*
 (b) *bez vraćanja (jednaki brojevi nisu mogući).*

Odredite standardnu devijaciju populacije σ_Y i standardnu devijaciju svih očekivanja vrijednosti uzoraka $\sigma_{\bar{Y}}$.

U Primjeru 6.2.1. pokazali smo da je očekivanje populacije $\mu_Y = 6$, pa za varijancu populacije vrijedi

$$\begin{aligned}\sigma_Y^2 &= \frac{(2-6)^2 + (3-6)^2 + (6-6)^2 + (8-6)^2 + (11-6)^2}{5} \\ &= \frac{16 + 9 + 0 + 4 + 25}{5} = 10.8,\end{aligned}$$

odnosno

$$\sigma_Y = \sqrt{10.8} \approx 3.28634.$$

- (a) *Da bismo dobili varijancu uzorka svakom očekivanju vrijednosti uzorka (već izračunato)*

2.0	2.5	4.0	5.0	6.5
2.5	3.0	4.5	5.5	7.0
4.0	4.5	6.0	7.0	8.5
5.0	5.5	7.0	8.0	9.5
6.5	7.0	8.5	9.5	11.0

treba oduzeti očekivanje uzorka, kvadrirati, zbrojiti i podijeliti s 25. Dobivamo

$$\sigma_{\bar{Y}}^2 = \frac{(2.0-6)^2 + (2.5-6)^2 + (4.0-6)^2 + \dots + (11.0-6)^2}{25} = \frac{135}{25} = 5.4,$$

a za standardnu devijaciju vrijedi

$$\sigma_{\bar{Y}} = \sqrt{5.4} \approx 2.32379.$$

Primijetimo da je

$$\sigma_{\bar{Y}}^2 = \frac{\sigma_Y^2}{2},$$

pri čemu je duljina uzorka baš 2. I općenito vrijedi

$$\sigma_{\bar{Y}}^2 = \frac{\sigma_Y^2}{n},$$

pri čemu je n duljina uzorka.

(b) U ovoj situaciji, kad ne smijemo vraćati brojeve u populaciju, očekivanja svih vrijednosti uzoraka su:

$$2.5, 4.0, 5.0, 6.5, 4.5, 5.5, 7.0, 7.0, 8.5, 9.5,$$

pa je

$$\sigma_{\bar{Y}}^2 = \frac{(2.5 - 6)^2 + (4.0 - 6)^2 + \dots + (8.5 - 6)^2 + (9.5 - 6)^2}{10} = 4.05.$$

Primijetite da se to razlikuje od situacije kad uzorak možemo vraćati. Može se pokazati da ako uzorak duljine n bez vraćanja biramo iz populacije od n_p , onda je

$$\sigma_{\bar{Y}}^2 = \frac{\sigma_Y^2}{n} \cdot \frac{n_p - n}{n_p - 1}.$$

Ako uzorak duljine n izvlačimo iz konačne populacije bez vraćanja, onda je varijanca svih mogućih očekivanja uzoraka

$$\sigma_{\bar{Y}}^2 = \frac{\sigma_Y^2}{n} \cdot \frac{n_p - n}{n_p - 1}.$$

Ako uzorak duljine n izvlačimo iz beskonačne populacije ili iz konačne populacije s vraćanjem, onda je varijanca svih mogućih očekivanja uzoraka

$$\sigma_{\bar{Y}}^2 = \frac{\sigma_Y^2}{n}.$$

Nadalje, ako uzimamo **velike** uzorke $n \geq 30$, iz populacije koja je baram dvostruko veća od veličine uzorka, onda je za praktične potrebe \bar{Y} normalna slučajna varijabla s očekivanjem $\mu_{\bar{Y}}$ i standardnom devijacijom $\sigma_{\bar{Y}}$ bez obzira ponaša li se populacija kao normalna slučajna varijabla.

Ako je populacija normalno distribuirana, onda prethodni zaključak vrijedi i za male uzorke ($n < 30$).

Primjer 6.2.3. *Populacija od 3000 dojenčadi ima raspodjelu visina po normalnoj slučajnoj varijabli s očekivanjem $\mu = 68$ cm i standardnom devijacijom $\sigma = 3$ cm. Ako uzmemo 80 uzoraka po 25-ero dojenčadi, koliko je $\mu_{\bar{Y}}$ i $\sigma_{\bar{Y}}$, ako smo uzorke birali*

- (a) *s ponavljanjem,*
 (b) *bez ponavljanja.*

Budući da smo izabrali dosta uzoraka (ali ne sve moguće), onda će $\mu_{\bar{Y}}$ i $\sigma_{\bar{Y}}$ biti blizu onima koje dobivamo ako uzmemo sve moguće uzorke, ali neće biti jednaki. Dakle, i u slučaju (a) i u slučaju (b), očekivanje će biti blizu teoretskom očekivanju

$$\mu_{\bar{Y}} \approx \mu_Y = 68 \text{ cm},$$

a za varijancu imamo

- (a) *Standardna devijacija za uzorak ponavljanjem bit će približno jednaka*

$$\sigma_{\bar{Y}} \approx \frac{\sigma_Y}{\sqrt{25}} = \frac{3}{5} = 0.6 \text{ cm}.$$

- (b) *Standardna devijacija za uzorak bez ponavljanja bit će približno jednaka*

$$\sigma_{\bar{Y}} \approx \frac{\sigma_Y}{\sqrt{25}} \cdot \sqrt{\frac{3000 - 25}{3000 - 1}} = \frac{3}{5} \cdot \sqrt{\frac{2975}{2999}} = 0.59520 \text{ cm},$$

što se vrlo malo razlikuje od slučaja kad ponavljamo uzorak.

Primjer 6.2.4. *Kolika je vjerojatnost da će \bar{Y} iz primjera 6.2.3. za jedan uzorak biti između 66.8 i 68.3 cm? U koliko će se to od 80 uzoraka dogoditi?*

Ovdje nam je očekivanje uzorka približno jednako očekivanju populacije,

$$\mu_{\bar{Y}} = \mu_Y = 68 \text{ cm},$$

pa za traženu vjerojatnost imamo

$$\begin{aligned} \Pr(66.8 \leq \bar{Y} \leq 68.3) &= \Pr\left(\frac{66.8 - 68}{0.6} \leq \frac{\bar{Y} - 68}{0.6} \leq \frac{68.3 - 68}{0.6}\right) \\ &= \Pr(-2 \leq Z \leq 0.5) = F(0.5) - F(-2) \\ &= 0.69146 - (1 - F(2)) = 0.69146 - 0.02275 = 0.66871. \end{aligned}$$

Budući da smo uzeli 80 uzoraka, a vjerojatnost da će se dogoditi traženi događaj je 0.66871, onda se očekuje da je to u

$$80 \cdot 0.66871 = 53.49 \approx 53 \text{ uzoraka}.$$

6.3. Procjena parametara populacije

U zemlji Nedođiji živi milijun stanovnika. Neki od njih su plavi, a neki crveni. Međutim, ako nemamo vremena, novaca, niti volje ispitivati karakteristike svih milijun Nedođijaca, pitanje je kako procijeniti parametre populacije (koliko ih je crvenih i nižih od 130 cm) uzevši relativno mali slučajni uzorak Nedođijaca.

Dakle ideje je na temelju parametara uzorka procijeniti parametre populacije. Takva i slična pitanja mnogo su interesantnija statističarima nego, ako znamo parametre populacije, promatranje uzoraka.

Sjetite se, rekli smo da je uzorak kopija n identičnih nezavisnih slučajnih varijabli $Y_i = Y$.

Neka je g funkcija $g : \mathbb{R}^n \rightarrow \mathbb{R}$. Slučajnu varijablu

$$\Theta = g(Y_1, Y_2, \dots, Y_n)$$

zovemo statistika, ako funkcija g ne ovisi o nepoznatim parametrima.

Recimo slučajna varijabla

$$\bar{Y} = \frac{Y_1 + \dots + Y_n}{n}.$$

je statistika i zove se sredina uzorka. I slučajna varijabla

$$S^2 = \frac{(Y_1 - \bar{Y})^2 + \dots + (Y_n - \bar{Y})^2}{n - 1}$$

(nije greška ovaj $n - 1$ u nazivniku) je jedna statistika koja se zove varijanca uzorka.

Objašnjenje zašto $n - 1$, a ne n slijedi malo kasnije.

Ako u populaciji Y imamo nepoznati parametar ϑ koji treba odrediti, onda za statistiku

$$\Theta = g(Y_1, Y_2, \dots, Y_n)$$

kažemo da je procjenitelj parametra ϑ , a vrijednost te statistike (kad uvrstimo konkretne vrijednosti za slučajne varijable),

$$\hat{\vartheta} = g(y_1, y_2, \dots, y_n)$$

procjena parametra ϑ .

Za tu sredinu uzorka vrijedi:

$$E(\bar{Y}) = E\left(\frac{Y_1 + \dots + Y_n}{n}\right) = \frac{nE(Y)}{n} = E(Y),$$

$$\text{Var}(\bar{Y}) = \text{Var}\left(\frac{Y_1 + \dots + Y_n}{n}\right) = \frac{n \text{Var}(Y)}{n^2} = \frac{\text{Var}(Y)}{n}.$$

Očekivanje slučajne varijable \hat{S}^2

$$\hat{S}^2 = \frac{(Y_1 - \bar{Y})^2 + \dots + (Y_n - \bar{Y})^2}{n}$$

je (koristiti da je $E(X^2) = (E(X))^2 + \text{Var}(X)$)

$$\begin{aligned} E(\hat{S}^2) &= \frac{1}{n} \left(E((Y_1 - \bar{Y})^2) + \dots + E((Y_n - \bar{Y})^2) \right) \\ &= \frac{1}{n} \left((E(Y_1 - \bar{Y}))^2 + \text{Var}(Y_1 - \bar{Y}) + \dots + (E(Y_n - \bar{Y}))^2 + \text{Var}(Y_n - \bar{Y}) \right). \end{aligned}$$

Budući da je

$$E(Y_i - \bar{Y}) = \mu - \mu = 0,$$

onda je

$$\begin{aligned}
 E(\widehat{S}^2) &= \frac{1}{n} \left(\text{Var}(Y_1 - \bar{Y}) + \dots + \text{Var}(Y_n - \bar{Y}) \right) \\
 &= \frac{1}{n} \left(\text{Var} \left(Y_1 - \frac{1}{n} \sum_{i=1}^n Y_i \right) + \dots + \text{Var} \left(Y_n - \frac{1}{n} \sum_{i=1}^n Y_i \right) \right) \\
 &= \frac{1}{n} \left(\text{Var} \left(\frac{n-1}{n} Y_1 - \frac{1}{n} \sum_{i=2}^n Y_i \right) + \dots + \text{Var} \left(\frac{n-1}{n} Y_n - \frac{1}{n} \sum_{i=1}^{n-1} Y_i \right) \right) \\
 &= \frac{1}{n} \left(\left(\frac{n-1}{n} \right)^2 \text{Var}(Y_1) + \frac{1}{n^2} \sum_{i=2}^n \text{Var}(Y_i) + \dots + \left(\frac{n-1}{n} \right)^2 \text{Var}(Y_n) + \frac{1}{n^2} \sum_{i=1}^{n-1} \text{Var}(Y_i) \right) \\
 &= \frac{1}{n} \left(\left(\frac{n-1}{n} \right)^2 n\sigma^2 + \frac{1}{n^2} (n-1)\sigma^2 \right) \\
 &= \frac{n-1}{n} \sigma^2.
 \end{aligned}$$

Međutim, vrlo često se želi da je očekivanje statistike jednako samom parametru kojeg procjenjujemo.

Ako je

$$E(\Theta) = \vartheta,$$

onda se takav procjenitelj zove nepristrani procjenitelj.

Nepristrani procjenitelji mogu imati bolja svojstva od drugih procjenitelja, recimo njihova varijanca može biti manja.

Za procjenitelj \widehat{S}^2 pokazali smo da je

$$E(\widehat{S}^2) = \frac{n-1}{n} \sigma^2,$$

pa on nije nepristrani procjenitelj. Nepristrani procjenitelj dobit ćemo ako pomnožimo \widehat{S}^2 s $n/(n-1)$. Množenjem izlazi da je

$$S^2 = \frac{n}{n-1} \widehat{S}^2 = \frac{(Y_1 - \bar{Y})^2 + \dots + (Y_n - \bar{Y})^2}{n-1}$$

i da je to nepristrani procjenitelj, tj. da je

$$E(S^2) = \sigma^2.$$

Na temelju toga koje sve podatke imamo poznate o populaciji, iz parametara uzorka možemo procijeniti i parametre populacije.

6.3.1. Procjena očekivanja

Nepoznato očekivanje μ populacije Y procjenjujemo sredinom uzorka:

$$\bar{Y} = \frac{Y_1 + \cdots + Y_n}{n}.$$

6.3.2. Procjena varijance uz poznato očekivanje

Ako je poznato očekivanje populacije μ , onda njezinu varijancu procjenjujemo korištenjem procjenitelja \hat{S}^2 .

Ako je očekivanje poznato, varijancu $\text{Var}(Y)$ procjenjujemo na sljedeći način:

$$\hat{S}^2 = \frac{(Y_1 - \mu)^2 + \cdots + (Y_n - \mu)^2}{n},$$

pri čemu je μ poznato očekivanje slučajne varijable.

6.3.3. Procjena i očekivanja i varijance

Nepoznato očekivanje μ populacije Y procjenjujemo sredinom uzorka:

$$\bar{Y} = \frac{Y_1 + \cdots + Y_n}{n},$$

a varijancu procjeniteljem S^2

$$S^2 = \frac{(Y_1 - \bar{Y})^2 + \cdots + (Y_n - \bar{Y})^2}{n - 1}.$$

Primjer 6.3.1. *Da bi se utvrdila preciznost uređaja za mjerenje debljine dasaka, debljina daske mjerena je pet puta i izmjerene su sljedeće vrijednosti (u centimetrima) 2.64, 2.68, 2.71, 2.73, 2.65.*

(a) *Odredite procjenu varijance, ako je poznato da je stvarna debljina daske 2.7 cm.*

(b) *Odredite procjenu očekivanja i varijance ako nije poznata debljina daske.*

Pretpostavljamo da se izmjereni rezultati debljine daske ponašaju kao normalna slučajna varijabla.

(a) *U ovom slučaju je procjena za varijancu (uz poznato očekivanje)*

$$\begin{aligned} s^2 &= \frac{(y_1 - \mu)^2 + \cdots + (y_n - \mu)^2}{n} \\ &= \frac{1}{5} \left((2.64 - 2.7)^2 + (2.68 - 2.7)^2 \right. \\ &\quad \left. + (2.71 - 2.7)^2 + (2.73 - 2.7)^2 + (2.65 - 2.7)^2 \right) \\ &= 0.0015. \end{aligned}$$

(b) *U ovom slučaju izabiremo procjenu za nepoznato očekivanje i varijancu*

$$\begin{aligned} \bar{y} &= \frac{2.64 + 2.68 + 2.71 + 2.73 + 2.65}{5} = 2.682, \\ s^2 &= \frac{(y_1 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2}{n - 1} \\ &= \frac{1}{4} \left((2.64 - 2.682)^2 + (2.68 - 2.682)^2 \right. \\ &\quad \left. + (2.71 - 2.682)^2 + (2.73 - 2.682)^2 + (2.65 - 2.682)^2 \right) \\ &= 0.00147. \end{aligned}$$

6.4. Intervali pouzdanosti

6.4.1. Intervali pouzdanosti za parametre populacije

Neka su μ_S i σ_S očekivanje i standardna devijacija uzorka neke statistike S . Ako smo uzeli dovoljno velike uzorke $n \geq 30$ (koji se mogu ponavljati ili iz beskonačne populacije), mnoge statistike S ponašaju se kao normalne slučajne varijable s očekivanjem $\mu_S \approx \mu$ i standardnom devijacijom $\sigma_S \approx \frac{\sigma}{\sqrt{n}}$. Ovdje ne treba paziti imamo li u nazivniku \sqrt{n} ili $\sqrt{n-1}$, jer je za dovoljno velike n to za sve praktične potrebe jednako. Dakle,

$$S \approx N \left(\mu, \frac{\sigma}{\sqrt{n}} \right).$$

Drugim riječima, sa slučajnom varijablom S ponašamo se kao s normalnom slučajnom varijablom. Ako za statistiku uzmemo sredinu uzorka \bar{Y} , onda smo već

zaključili (vidjeti prethodno predavanje) da je

$$\Pr \left(|\bar{y} - \mu| < k \frac{\sigma}{\sqrt{n}} \right) = 2(F(k) - F(0)),$$

pri čemu je \bar{y} vrijednost slučajne varijable \bar{Y} .

Za neke k -ove već smo izračunali kolika je desna strana. Faktor s desne strane $2(F(k) - F(0))$ obično se zove pouzdanost i piše se u obliku postotka. Neke tipične vrijednosti su (vidjeti tablicu za funkciju distribucije normalne razdiobe):

k	3	2.58	2.33	2.05	2	1.96	1.645	1.28	1	0.6745
pouzdanost u %	99.73	99	98	96	95.45	95	90	80	68.27	50

Drugim riječima, kad kažemo da je

$$\Pr \left(|\bar{y} - \mu| < 2 \frac{\sigma}{\sqrt{n}} \right) = 95.45\%,$$

to znači da se parametar μ može kretati od

$$\bar{y} - 2 \frac{\sigma}{\sqrt{n}} < \mu < \bar{y} + 2 \frac{\sigma}{\sqrt{n}}$$

s pouzdanošću 95.45%. To se obično zapisuje da je

$$\mu = \bar{y} \pm 2 \frac{\sigma}{\sqrt{n}}.$$

Primjer 6.4.1. Mjeranja dijametra slučajnog uzorka od 200 kugličnih ležajeva pokazala su da je očekivanje uzorka 0.824 mm, a standardna devijacija 0.042 mm. Nađite s

- (a) 95%,
(b) 99%,

pouzdanošću očekivanje dijametra svih ležajeva.

Budući da smo uzali veliki uzorak $200 > 30$, onda se slučajna varijabla ponaša kao normalna slučajna varijabla s očekivanjem 0.824 i standardnom devijacijom

$$\frac{0.042}{\sqrt{200}} \approx 0.00296984.$$

- (a) Za 95% pouzdanost je $k = 1.96$, pa je

$$\mu = 0.824 \pm 1.96 \cdot 0.00296984 = 0.824 \pm 0.0058209 \text{ mm.}$$

- (a) Za 99% pouzdanost je $k = 2.58$, pa je

$$\mu = 0.824 \pm 2.58 \cdot 0.00296984 = 0.824 \pm 0.0076622 \text{ mm.}$$

6.4.2. Intervali pouzdanosti za proporcije

Ako statistika S daje proporciju uspjeha u uzorku od n , $n \geq 30$, izvučenih iz binomne slučajne varijable (s ponavljanjem), s vjerojatnošću uspjeha p , pitanje je kako ćemo naći interval pouzdanosti za p .

Kad smo počeli razmatrati binomnu slučajnu varijablu, onda smo prvo uveli jednostavnu indikatorsku varijablu

X	Pr
1	p
0	q

za koju je

$$\mu = p, \quad \sigma = \sqrt{pq}.$$

Ako uzmemo dovoljno uzoraka varijable X (s ponavljanjem), onda vrijede intervali pouzdanosti za $\mu = p$ za normalnu slučajnu varijablu iz prethodnog odjeljka, s time da za \bar{y} treba staviti izračunati \bar{p} koji dobijemo iz uzorka, a za $\sigma = \sqrt{pq} = \sqrt{p(1-p)}$.

Uočimo da je

$$\sigma \leq \frac{1}{2},$$

što se lako dokazuje. Naime, dovoljno je dokazati da je

$$p(1-p) \leq \frac{1}{4}.$$

Budući da je $p(1-p) = 0$ parabola okrenuta nadolje, onda je njeno tjeme na polovištu nultočaka, tj. u $\frac{1}{2}$ i u njemu poprima vrijednost baš $\frac{1}{4}$.

Tu činjenicu možemo iskoristiti na sljedeći način. Ako želimo interval pouzdanosti 95.45%, onda je interval pouzdanosti za p

$$p = \bar{p} \pm 2 \frac{\sqrt{p(1-p)}}{\sqrt{n}} \approx \bar{p} \pm \frac{1}{\sqrt{n}},$$

što se često koristi u jednostavnim procjenama.

Ako je uzorak duljine n izvučen iz binomne slučajne varijable kojoj je vjerojatnost uspjeha p , onda je interval pouzdanosti za p

$$p = \bar{p} \pm k \sqrt{\frac{p(1-p)}{n}}$$

ako je uzorak izvučen iz beskonačne populacije ili iz konačne populacije s vraćanjem, a

$$p = \bar{p} \pm k \sqrt{\frac{p(1-p)}{n}} \sqrt{\frac{n_p - n}{n_p - 1}},$$

ako je izvučen iz konačne populacije veličine n_p .

Primjer 6.4.2. Na uzorku od $n = 1000$ pojedinaca našli smo da je $\bar{p} = 0.43 = 43\%$, koliki je interval pouzdanosti od 95.45% za p ?

Iz gore navedenog vrijedi

$$p = 0.43 \pm \frac{1}{\sqrt{1000}} = 0.43 \pm 0.032 = 43\% \pm 3.2\%.$$

Primjer 6.4.3. Želimo odrediti postotak crvenih Nedođijaca u granicama od $\pm 1\%$ s pouzdanošću 95.45%. Koliki uzorak moramo uzeti.

Iz jednostavne intervalne procjene za p imamo

$$p = \bar{p} \pm \frac{1}{\sqrt{n}}$$

s pouzdanošću 95.45%. Mi želimo da je

$$\frac{1}{\sqrt{n}} = 0.01 = 1\%.$$

Odatle odmah izlazi da je

$$\sqrt{n} = \frac{1}{0.01} = 100,$$

pa je $n = 10000$.

Ako je uzorak duljine n , $n \geq 30$ uzet iz binomne slučajne varijable kojoj ne znamo p , onda standardnu devijaciju σ možemo zamijeniti sa standardnom devijacijom uzorka, tj.

$$\sigma = \sqrt{p(1-p)} \approx s = \sqrt{\bar{p}(1-\bar{p})}.$$

Primjer 6.4.4. U 40 bacanja (nesimetričnog) novčića dobili smo 24 glave. Nađite s pouzdanošću 99.73% granice za p ako novčić bacamo neograničen broj puta.

Imamo

$$\bar{p} = \frac{24}{40} = 0.6,$$

a σ aproksimiramo sa

$$s = \sqrt{0.6 \cdot 0.4} = \sqrt{0.24} = 0.4898979,$$

pa je

$$p = 0.6 \pm 3 \cdot \frac{0.4898979}{\sqrt{40}} = 0.6 \pm 0.2324.$$

Primjer 6.4.5. Na uzorku od 100 pojedinaca izmjerili smo da je visina tih pojedinaca $\bar{y} = 172$ cm i $s = 7$ cm. Nađite procjenu za μ s pouzdanošću 99.73%.

Zbog veličine uzorka možemo uzeti da je

$$\sigma \approx s,$$

pa je

$$\mu = 172 \pm 3 \frac{7}{\sqrt{100}} = 172 \pm 2.1 \text{ cm.}$$

Primjer 6.4.6. Želimo odrediti prosječnu visinu muškaraca u Hrvatskoj u granicama ± 1 cm i pouzdanošću 99.73%, uz pretpostavku da je $\sigma \approx 10$ cm. Koliki uzorak trebamo uzeti za procjenu?

$$\mu = \bar{y} \pm 3 \frac{10}{\sqrt{n}}.$$

Mi zahtijevamo da je

$$\frac{30}{\sqrt{n}} = 1 \text{ cm.}$$

Odatle slijedi da je $n = 900$.