5 Procjena parametara

5.1 Slučajni uzorak

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

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

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

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

5.2 Procjena populacijskog očekivanja i populacijske varijance

Zanima nas statističko obilježje X neke promatrane populacije. Pretpostavimo da je X slučajna varijabla s konačnim očekivanjem $\mu = \mathrm{E}[X]$ i varijancom $\sigma^2 = \mathrm{Var}[X]$. μ i σ^2 su parametri razdiobe od X. U ovom poglavlju zanima nas procjena vrijednosti od μ i σ^2. Razlikujemo točkovnu i intervalnu procjenu parametara.

Cilj **točkovnog procjenjivanja** je izvođenje jedinstvenog broja iz slučajnog uzorka za kojeg je razumno smatrati da je blizak nepoznatoj vrijednosti parametra populacije.

Označimo sa Θ određeni parametar populacije. Promatramo **slučajne uzorke** koji se sastoje od n nezavisnih jednako distribuiranih slučajnih varijabli

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

s distribucijom jednakom distribuciji mjerenog statističkog obilježja.

Želimo formirati funkcije slučajnog uzorka (**statistike**)

$$f(X_1, X_2, \ldots, X_n)$$

tako da kad izračunavamo vrijednost te funkcije na podacima iz uzorka (vrijednost statistike, realizacija statistike) dobijemo broj blizak određenom parametru populacije što je najviše moguće.

Funkcije slučajnog uzorka nazivat ćemo statistikama (to su slučajne varijable).

Statistika koja ima za cilj procijeniti određeni parametar Θ nazivat ćemo **točkovnim procjeniteljem** ili samo procjeniteljem parametra Θ.

Kako procijeniti parametre μ i σ^2 iz slučajnog uzorka

$$X_1, X_2, \ldots, X_n.?$$

Označimo sa

$$\overline{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$$

statistiku koju zovemo aritmetičkom sredinom uzorka (uzoračkom aritmetičkom sredinom).

Svojstva:

1.

$$\begin{aligned} E[\overline{X}_n] &= \text{lineranost} = \frac{1}{n}(E[X_1] + \ldots + E[X_n]) \\ &= \text{jednaka distribucija} = \frac{1}{n} n\mu = \mu, \end{aligned}$$

2.

$$\mathrm{Var}[\overline{X}_n] = \frac{1}{n^2}\mathrm{Var}(X_1 + \ldots + X_n) = \text{nezavisnost}$$

$$= \frac{1}{n^2}(\mathrm{Var}X_1 + \ldots + \mathrm{Var}X_n) = \text{jednaka distribucija} = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}.$$

Slučajnu varijablu $\overline{X}_n$ uzimamo za (točkovnog) procjenitelja partametra μ. Prethodna svojstva (1) i (2) govore da je $\overline{X}_n$ "dobar" procjenitelj za μ.

Za procjenitelj $T_n = f(X_1, \ldots, X_n)$ **kažemo da je nepristrani procjenitelj** za parametar Θ ako je

$$E[T_n] = \Theta.$$

Za nepristrani procjenitelj $T_n = f(X_1, \ldots, X_n)$ **parametra** Θ **kažemo da je konzistentan procjenitelj** ako je

$$\lim_{n\to\infty} \mathrm{Var}[T_n] = 0.$$

Iz prethodnog očito slijedi da $\overline{X}_n$ nepristran i konzistentan procjenitelj parametra μ.

Neka je sada definirana statistika

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2.$$

Očito je

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - n\overline{X}_n^2\right].$$

Prisjetimo se još da je $E[X^2] = \text{Var}[X] + \mu^2$. Imamo:

$$E[S_n^2] = \frac{1}{n-1}\sum_{i=1}^{n} E[X_i^2] - \frac{n}{n-1}E[\overline{X}_n^2]$$

$$= \frac{1}{n-1}n(\sigma^2+\mu^2) - \frac{n}{n-1}\left(\frac{\sigma^2}{n}+\mu^2\right) = \sigma^2.$$

Zaključujemo da je statistika S_n^2 (koju nazivamo uzoračka varijanca) nepristrani procjenitelj varijance σ^2. Može se pokazati da je S_n^2 i konzistentan procjenitelj.

5.3 Pouzdani intervali za očekivanje normalne populacije

Definicija 18 *Neka su $L_n = l(X_1,\dots,X_n)$ i $D_n = d(X_1,\dots,X_n)$ slučajne varijable (statistike), funkcije slučajnog uzorka $X_1,\dots,X_n$.*
Kažemo da je $[L_n, D_n]$ $(1-\alpha)\cdot 100\%$ ***pouzdan interval*** *za parametar τ ako vrijedi*

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

5.3.1 Varijanca poznata

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

- imamo slučajni uzorak veličine $n: \; X_1,\dots,X_n$
- $(1-\alpha)\cdot 100\%$ pouzdan interval za μ dobit ćemo ako promatramo uzoračku distribuciju statistike $\bar{X}_n$ (aritmetička sredina uzorka):

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

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

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

Vrijedi:

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

i nadalje:

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

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

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

Zadatak 16 *Vrijeme trajanja neke vrste elektronskih cijevi je normalno distribuirana slučajna varijabla X s nepoznatim očekivanjem μ i standardnom devijacijom $\sigma = 40h$.*
a) Uzet je uzorak od 30 elektronskih cijevi za koji je dobiveno prosječno vrijeme trajanje od 780h. Nađite 99% *pouzdan interval za očekivanje μ vremena trajanja ove vrste elektronskih cijevi.*
b) Koliki uzorak treba uzeti da bi se s vjerojatnošću 0.99, sredina uzorka $\bar{x}$ razlikovala od sredine μ manje od 10h?

Rješenje:

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

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

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

99% pouzdan interval za očekivanje:

$$\bar{x}_{30} \pm z_{0.005} \cdot \frac{\sigma}{\sqrt{n}} = 780 \pm 2.58 \cdot \frac{40}{\sqrt{30}} = 780 \pm 18.84$$

$$\Rightarrow \qquad 761.16 \leq \mu \leq 798.84$$

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

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

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

$$\Rightarrow \Phi_0\left(\frac{\sqrt{n}}{4}\right) - \Phi_0\left(-\frac{\sqrt{n}}{4}\right) = 0.99 \Leftrightarrow 2\Phi_0\left(\frac{\sqrt{n}}{4}\right) = 0.99 \Leftrightarrow \Phi_0\left(\frac{\sqrt{n}}{4}\right) = 0.495$$

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

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

$\square$

Zadatak 17 *(DZ) Neka mašina proizvodi kuglične ležajeve. Promjer kugličnog ležaja je normalna slučajna varijabla X s varijancom 1. Dužine 9 slučajno odabranih kugličnih ležajeva bile su*

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

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

5.3.2 Varijanca nepoznata

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

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

- varijancu σ^2 **procijenimo** pomoću S_n^2

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

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

- standardiziranu varijablu

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

aproksimativno zapisujemo pomoću procjene za σ^2:

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

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

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

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

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

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

Zadatak 18 *NASA testira komponente svojih raketa. Recimo da NASA želi procijeniti srednje vrijeme trajanja neke mehaničke komponente korištene* *u raketi "Columbia". Zbog ograničenja troškova, u simuliranim uvjetima svemira mogu testirati samo 10 komponenti. Dobiveni su podaci za vrijeme trajanja tih komponenti (u satima):* $\bar{x}_{10} = 1173.6, \quad s_{10} = 36.3$. *Procijenite očekivanje vijeka trajanja tih mehaničkih komponenti s* 95% *pouzdanim intervalom (pretpostavite da je vrijeme trajanja mehaničkih komponenti normalno distribuirano).*

Rješenje:

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

□

Zadatak 19 *(DZ) U svrhu istraživanja utjecaja toksičnih tvari koje luči jedna vrsta plijesni na kukuruz, biokemičar u 9 ekstrakata plijesni mjeri količinu toksičnih supstanci u mg. Dobiveni su rezultati: 1.2, 0.8, 0.6, 1.1, 1.2, 0.9, 1.5, 0.9, 1.0. Uz pretpostavku da su podaci iz normalne distribucije, procijenite* 98% *pouzdan interval za očekivanje te populacije.*

5.4 Pouzdani intervali za očekivanje populacije na osnovi velikih uzoraka

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

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

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

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

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

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

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

$$\boxed{\begin{array}{c}(1-\alpha)\cdot 100\% \text{ pouzdan interval za očekivanje populacije}\\ \text{na osnovi velikih uzoraka}\\ \bar{X}_n - z_{\frac{\alpha}{2}} \cdot \frac{S_n}{\sqrt{n}} \leq \mu \leq \bar{X}_n + z_{\frac{\alpha}{2}} \cdot \frac{S_n}{\sqrt{n}}\end{array}}$$

Zadatak 20 *Zoolog želi procijeniti očekivanu količinu šećera u krvi određene životinjske vrste nastale nakon ubrizgavanja određene količine adrenalina. Dobivena srednja vrijednost uzorka od 55 životinja je 126.9 mg/100 ml uz standardnu devijaciju uzorka od 10.5 mg /100 ml. Odredite* 90% *pouzdan interval za očekivanje.*

Rješenje:

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

5.4.1 Pouzdan interval za parametar p binomne razdiobe

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

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

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

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

Imamo:

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

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

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

Zadatak 21 *Uzorak od 100 kućanstava nekog grada pokazao je da se u* 55% *kućanstava bar jedan član koristi Internetom.*
a) Nađite 95% *pouzdan interval za omjer kućanstava u tom gradu koja se služe Internetom.*
b) Koliko kućanstava treba uzeti da bi s vjerojatnošću od 0.95 mogli tvrditi da se najmanje 50% *kućanstava služi Internetom?*

Rješenje:

a)
$$
\begin{aligned}
&\hat{p} = \bar{x} = 0.55, \quad n = 100 \\
&1 - \alpha = 0.95\ \Leftrightarrow\ \alpha = 0.05\ \Leftrightarrow\ \frac{\alpha}{2} = 0.025 \\
&z_{\frac{\alpha}{2}} = z_{0.025} = 1.96
\end{aligned}
$$

$$\hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.55 \pm 1.96 \sqrt{\frac{0.55 \cdot 0.45}{100}} = 0.55 \pm 0.09751$$
$$\Longrightarrow 0.4525 \leq p \leq 0.6475$$

b) $n = ?$

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

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

$$0.55 - 1.96\sqrt{\frac{0.55 \cdot 0.45}{n}} \geq 0.5 \;\Leftrightarrow\; \frac{0.9751}{\sqrt{n}} \leq 0.05$$
$$\Leftrightarrow\; \sqrt{n} \geq 19.502 \;\Rightarrow\; n \geq 380.328$$
$$\Longrightarrow n \geq 381$$

$\square$