10. Matematička statistika

Sadržaj poglavlja

Novčić je bačen stotinu puta, pri čemu se pismo pojavilo 40 puta. Je li način bacanja bio korektan?

Ako se pri 100 bacanja kocke šestica pojavi 10 puta, je li ta kocka ispravna?

Pri ponavljanju identičnog pokusa slučajna varijabla poprimila je vrijednosti 6.1, 6.3, 6.4, 6.8, 7.2. Ako je njezina razdioba normalna, a parametri te razdiobe nepoznati, koji ćemo broj uzeti za očekivanje a koji za disperziju te varijable?

Slučajna varijabla poprima vrijednosti unutar intervala $[0,1]$. Bilježenje rezultata dalo je podatke: 0.11, 0.14, 0.28, 0.44, 0.48, 0.68, 0.76. O razdiobi ove varijable nemamo nikakvu informaciju. S kojom sigurnošću možemo tvrditi da je riječ o jednolikoj razdiobi?

Na ova, i slična pitanja, odgovor daje *matematička statistika*.

10.1. Točkaste procjene parametara

10.1.1. Uvod

Predmet svakog statističkog proučavanja jest neki (masovni) skup, kojeg nazivamo **populacija** ili **generalni skup**. Populaciju mogu činiti na primjer

- stanovnici države, općine, mjesta;
- privredni potencijali države, regije, grada;
- proizvodnja neke tvornice u jednom danu, mjesecu ili godini

i slično. Podatak koji proučavamo u danoj populaciji nazivamo **obilježje**. Kod iste populacije možemo promatrati više obilježja. Npr. ako je u pitanju stanovništvo, možemo se zanimati za, recimo

- promjenu brojčanog stanja stanovništva tijekom godina;
- broj stanovnika prema starosnoj dobi;
- zaposlenost po vrstama zanimanja;
- školsku spremu

i za stotinjak drugih podataka. Promatramo li proizvodnju, obilježja mogu biti

- broj (količina) proizvedenih dobara u nekom vremenu;
- proizvodnja po vrstama proizvoda;
- profit;
- broj (postotak) škartnih proizvoda u ukupnoj proizvodnji.

Statistički se mogu pratiti i mnoge druge pojave. Tako na primjer, analiziraju se

- meteroološke prilike,
- učestalost i vrsta bolesti,
- ispitivanje kupovne moći, tržišta i slično.

U modelu matematičke statistike, populacija čini skup Ω. Obilježje je opisano vrijednošću slučajne varijable X. Osnovni problem matematičke statistike je u određivanju razdiobe varijable X, ili pak nekih njezinih numeričkih karakteristika.

Statistika se može baviti proučavanjem podataka koji **točno** opisuju stanje u svakoj populaciji. Ti se podaci dobivaju uglavnom popisom, redovitim evidencijama i praćenjima. Tako na primjer, svake desete godine se organiziraju popisi cjelokupnog stanovništva države. Analiziranjem i prikazivanjem takvih podataka bavi se tzv. **deskriptivna statistika**.

Vrlo često je nemoguće (i nepotrebno!) statistički obraditi čitavu populaciju. Djelom zbog toga što je ona prevelika da bi se taj postupak mogao sprovesti ili da bi bio isplativ. Drugi mogući razlog jest što se u nekim postupcima ispitivanja (recimo u kontroli kvalitete proizvodnje) u samom postupku ispitivanja uništava taj element populacije. Zamislimo na primjer ispitivanje duljine života žarulje!

U tom slučaju se proučava samo jedan mali dio populacije koji nazivamo **uzorak**. Na osnovu tog uzorka, donosimo potom sud o čitavoj populaciji. Predmet **matematičke statistike** jest statistička obrada uzorka: način odabira uzorka (da bi on dobro predstavljao čitavu populaciju) analiza obilježja u uzorku i procjena u kojoj su mjeri ti rezultati vjerodostojni za čitavu populaciju. Kako zaključci u ovom slučaju ne mogu nikad biti apsolutno sigurni (oni se donose uvijek s nekim stupnjem vjerojatnosti), matematička statistika se izražava i koristi metodama teorije vjerojatnosti.

Primjer 10.1.

Rezultati općih izbora postaju poznati (i službeni) kad se zna za glas svakog birača, tj. tek nakon što se obradi čitava populacija. Međutim, mnogo prije toga se rezultati mogu predvidjeti (s velikom dozom sigurnosti) na osnovu glasanja nekog dobro izabranog uzorka, koji može biti po veličini i 10 000 puta manji od čitave populacije!

Primjer 10.2.

Jedna proizvodna traka proizvodi otpornike. Dozvoljena granica škarta je 2%. Kako ćemo kontrolirati je li proizvodnja ispravna, t.j. je li postotak škarta unutar tih granica?

Bilo bi nerazumno, i skoro nemoguće za ovakav tip proizvoda, kontrolirati čitavu proizvodnju. Umjesto toga, uzimamo relativno maleni uzorak, odabran na pogodan način: recimo, svaki stoti proizvod. Ako je broj škartnih proizvoda u tom uzorku veći od određene granice, uz veliku dozu sigurnosti možemo zaključiti da je broj škartova u čitavoj populaciji veći od 2%, tj. da je došlo do grešaka u proizvodnji koje treba ispraviti. Kolika je ta dozvoljena granica škartnih proizvoda u uzorku i kolika je sigurnost u našem zaključku, to je predmet izučavanja metematičke statistike.

10.1.2. Populacija. Uzorak

Upoznajmo se s oznakama i temeljnim pojmovima matematičke statistike. Sa X ćemo označiti slučajnu varijablu koja će biti predmet proučavanja. Nju ćemo zvati **populacija**. Njezinu funkciju distribucije označavat ćemo sa F, funkciju gustoće (ako postoji) sa f, očekivanje s a i disperziju sa σ^2.

U ovisnosti o problemu koji promatramo, neki parametri ϑ_1, $\vartheta_2, \ldots$ u ovoj razdiobi mogu biti nepoznati. Najčešći zadatak matematičke statistike jest dati odgovarajuću procjenu za te parametre. Ta se procjena postiže na temelju poznatih realizacija $x_1, x_2, \ldots, x_n$ slučajne varijable X.

Informacije o nepoznatoj razdiobi populacije X dobivamo samo na temelju realizacija te slučajne varijable.

Uzorak

Neka je X slučajna varijabla s razdiobom F. Za slučajne varijable $X_1, \ldots, X_n$ kažemo da su **nezavisne kopije** slučajne varijable X, ako one imaju svojstva:

1. međusobno su nezavisne,

2. imaju razdiobu identičnu razdiobi slučajne varijable X.

Tako dobivenu n-torku slučajnih varijabli $(X_1, \ldots, X_n)$ nazivamo **uzorak**. Ako je x_1 je realizacija varijable X_1, x_2 realizacija varijable X_2 i t.d., tada se $(x_1, \ldots, x_n)$ naziva **vrijednost** ili **realizacija** uzorka $(X_1, \ldots, X_n)$.

Broj n označava **veličinu** (**dimenziju** ili **volumen**) uzorka.

Možemo zamisliti da varijable $X_1, \ldots, X_n$ opisuju ponašanje slučajne varijable X pri ponavljanju stohastičkog eksperimenta u nepromijenjenim uvjetima.

Radi jednostavnosti, pretpostavimo za sada da je u razdiobi varijable X nepoznat jedan parametar ϑ. Funkciju gustoće varijable X označavat ćemo s $f_\vartheta(x)$ ili s $f(\vartheta, x)$, jer ona ovisi o tom nepoznatom parametru ϑ.

Vrijednost parametra ϑ trebamo procijeniti na temelju realizacija $x_1, x_2, \ldots, x_n$ varijable X. Bit će definirana funkcija

$$\widehat{\vartheta} = g(x_1, x_2, \ldots, x_n)$$

koja će dati procjenu $\widehat{\vartheta}$ parametra ϑ. Ta procjena ovisi, dakle, o realizacijama $x_1, x_2, \ldots, x_n$. Realizacije su slučajne, pa je prirodno da će se pri ponavljanju pokusa pojaviti neka druga n-torka, a onda i druga vrijednost za procjenu $\widehat{\vartheta}$. Zato je normalna situacija da procjena $\widehat{\vartheta}$ nije jednaka nepoznatom parametru ϑ. (Jedan od zadataka matematičke statistike jest da pruži *mjeru sigurnosti* za točnost ove procjene.)

Budući da su $x_1, x_2, \ldots, x_n$ realizacije slučajnih varijabli $X_1, X_2, \ldots, X_n$, onda će i $\widehat{\vartheta}$ biti realizacija slučajne varijable

$$\Theta := g(X_1, X_2, \ldots, X_n).$$

Statistika, procjenitelj i procjena

Slučajna varijabla

$$\Theta := g(X_1, X_2, \ldots, X_n). \tag{1}$$

naziva se **statistika**. Statistikom nazivamo *svaku* funkciju koja ovisi o uzorku $X_1, X_2, \ldots, X_n$, a ne ovisi (eksplicitno) o nepoznatom parametru.

Neka je ϑ nepoznati parametar u populaciji X. Za statistiku (1) kažemo da je **procjenitelj** parametra ϑ. Vrijednost te statistike

$$\widehat{\vartheta} = g(x_1, x_2, \ldots, x_n) \tag{2}$$

nazivamo **procjenom** parametra ϑ.

Prema tome, procjenitelj je slučajna varijabla. Nakon realizacije uzorka, vrijednost procjenitelja daje nam procjenu nepoznatog parametra.

10.1.3. Statistika za procjenu očekivanja

Želimo procijeniti nepoznato očekivanje a populacije X. Prirodno je onda odabrati statistiku

$$\overline{X} := \frac{X_1 + X_2 + \ldots + X_n}{n}.$$

Ona se naziva **sredina** uzorka.

Označimo nepoznato očekivanje i disperziju populacije X:

$$\begin{cases} \boldsymbol{E}(X) = a, \\ \boldsymbol{D}(X) = \sigma^2. \end{cases}$$

Varijabla $\overline{X}$ je slučajna. Izračunajmo njezino očekivanje i disperziju! Prema svojstvima očekivanja, vrijedi

$$\begin{aligned} \boldsymbol{E}(\overline{X}) &= \boldsymbol{E}\left(\frac{X_1 + X_2 + \ldots + X_n}{n}\right) \\ &= \frac{1}{n}\Big[\boldsymbol{E}(X_1) + \boldsymbol{E}(X_2) + \ldots + \boldsymbol{E}(X_n)\Big] = a. \end{aligned}$$

Varijable $X_1, \ldots, X_n$ su nezavisne, pa je

$$\begin{aligned} \boldsymbol{D}(\overline{X}) &= \boldsymbol{D}\left(\frac{X_1 + X_2 + \ldots + X_n}{n}\right) \\ &= \frac{1}{n^2}\Big[\boldsymbol{D}(X_1) + \boldsymbol{D}(X_2) + \ldots + \boldsymbol{D}(X_n)\Big] = \frac{\sigma^2}{n}. \end{aligned}$$

Procjena očekivanja

Nepoznato očekivanje a populacije X procjenjujemo pomoću **sredine** uzorka:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{3}$$

Za tu slučajnu varijablu vrijedi

$$\boldsymbol{E}(\overline{X}) = a, \qquad \boldsymbol{D}(\overline{X}) = \frac{\sigma^2}{n}, \tag{4}$$

gdje je σ^2 **varijanca** (**disperzija**) populacije.

Primjećujemo da je disperzija statistike $\overline{X}$ obrnuto proporcionalna veličini uzorka. Ako je uzorak dovoljno velik, vrijednosti varijable $\overline{X}$ bit će koncentrirane oko srednje vrijednosti $\boldsymbol{E}(\overline{X}) = a$. Zato je jasno da će $\overline{X}$ biti dobra procjena za a. O kvaliteti te procjene bit će više riječi u nastavku. ◀

10.1.4. Nepristrani procjenitelji

Među svim statistikama želimo odabrati one koje su, po nekim kriterijima, bolje od drugih. Zato ćemo izdvojiti neka poželjna svojstva statistika te dati kriterij za usporedbu različitih statistika.

U prethodnom primjeru, statistika $\overline{X}$ za parametar a imala je svojstvo:

$$\boldsymbol{E}(\overline{X}) = a.$$

Dakle, *očekivanje statistike podudara se s vrijednošću parametra*. Statistike koje posjeduju to poželjno svojstvo nazvat ćemo posebnim imenom.

Nepristrani procjenitelji

Za statistiku Θ kažemo da je **nepristrani procjenitelj** ili **nepristrana statistika** parametra ϑ, ukoliko vrijedi

$$\boldsymbol{E}(\Theta) = \vartheta.$$

Kriterij nepristranosti svakako je poželjan, ali nije jedini odlučujući za odabir statistike. Upoznat ćemo primjere kod kojih pristrani procjenitelji mogu imati bolja svojstva od nepristranih. (Na primjer, njihova disperzija može biti manja.)

10.1.5. Usporedba statistika

Usporedba statistika

Neka je $(X_1, \ldots, X_n)$ uzorak, ϑ nepoznati parametar te $\Theta_1(X_1, \ldots, X_n)$, $\Theta_2(X_1, \ldots, X_n)$ dvije nepristrane statistike za ϑ. Kažemo da je Θ_1 **bolja** (**efikasnija**) od Θ_2 ako je $\boldsymbol{D}(\Theta_1) < \boldsymbol{D}(\Theta_2)$.

Još je jedno poželjno svojstvo koje bi dobra statistika trebala imati: povećanjem uzorka statistika mora davati sve bolju aproksimaciju nepoznatog parametra.

Valjane statistike

Statistiku $\Theta_n = \Theta(X_1, X_2, \ldots, X_n)$ nazivamo **valjanom** procjenom parametra ϑ ako za svaki $\varepsilon > 0$ slučajna varijabla Θ_n konvergira prema ϑ *po vjerojatnosti*:

$$\lim_{n\to\infty} \boldsymbol{P}(|\Theta_n - \vartheta| < \varepsilon) = 1.$$

Teorem 10.1.

Da bi nepristrana statistika bila valjana, dovoljno je da joj disperzija teži u nulu (kad n teži u beskonačnost).

DOKAZ. Ta tvrdnja slijedi iz Čebiševljeve nejednakosti:

$$\boldsymbol{P}(|\Theta_n - \vartheta| < \varepsilon) \geqslant 1 - \frac{\boldsymbol{E}[(\Theta_n - \vartheta)^2]}{\varepsilon^2} = 1 - \frac{\boldsymbol{D}(\Theta_n)}{\varepsilon^2} \to 1.$$

10.1.6. Procjena disperzije, uz poznato očekivanje

Pretpostavimo sad da nam je očekivanje populacije poznato, a disperzija σ^2 nije. Za procjenu disperzije biramo statistiku

$$D^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - a)^2.$$

Očekivanje ove statistike je:

$$\boldsymbol{E}(D^2) = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{E}(X_i - a)^2 = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{D}(X_i) = \boldsymbol{D}(X) = \sigma^2.$$

Dakle, ova je statistika nepristrana.

O kvaliteti procjene odlučivat će disperzija statistike. Zbog nezavisnosti i jednake distribuiranosti slučajnih varijabli X_i bit će:

$$\begin{aligned}\boldsymbol{D}(D^2) &= \frac{1}{n^2}\sum_{i=1}^{n}\boldsymbol{D}\Big[(X_i - a)^2\Big] = \frac{1}{n}\boldsymbol{D}\Big[(X - a)^2\Big] \\ &= \frac{1}{n}\Big(\boldsymbol{E}\Big[(X-a)^4\Big] - \Big[\boldsymbol{E}(X-a)^2\Big]^2\Big) = \frac{1}{n}\Big(\mu_4 - \sigma^4\Big) \qquad (5)\end{aligned}$$

Ovdje je

$$\mu_4 = \boldsymbol{E}\Big[(X-a)^4\Big]$$

četvrti centralni moment populacije X.

Vidimo da disperzija statistike D^2 opada obrnuto proporcionalno veličini uzorka. Prema Teoremu 10.1, ova je statistika valjana.

10.1.7. Procjena disperzije, uz nepoznato očekivanje

Ako je očekivanje poznato, tada je statistika

$$D^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - a)^2$$

nepristrani procjenitelj za disperziju. Koju ćemo statistiku koristiti ako je i očekivanje a nepoznato? Prirodno je zamijeniti ga u ovoj formuli s $\overline{X}$. Tako dobivamo statistiku

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

Provjerimo je li ona nepristrana. Njezino očekivanje je

$$\boldsymbol{E}(\Theta) = \boldsymbol{E}\Big(\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\Big) = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{E}(X_i - \overline{X})^2.$$

Vrijedi $\boldsymbol{E}(X_i - \overline{X}) = a - a = 0$, pa je $\boldsymbol{E}(X_i - \overline{X})^2 = \boldsymbol{D}(X_i - \overline{X})$. Sada je, zbog nezavisnosti varijabli $X_1, X_2, \ldots, X_n$,

$$\begin{aligned}\boldsymbol{E}(\Theta) &= \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{D}(X_i - \overline{X}) \\ &= \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{D}\Big[X_i - \frac{1}{n}\sum_{j=1}^{n}X_j\Big] \\ &= \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{D}\Big[\frac{n-1}{n}X_i - \frac{1}{n}\sum_{j\neq i}X_j\Big] \\ &= \frac{1}{n}\sum_{i=1}^{n}\Big[\Big(\frac{n-1}{n}\Big)^2\boldsymbol{D}(X_i) + \frac{1}{n^2}\sum_{j\neq i}\boldsymbol{D}(X_j)\Big] \\ &= \frac{1}{n}\cdot n\Big[\Big(\frac{n-1}{n}\Big)^2\sigma^2 + \frac{1}{n^2}\cdot(n-1)\sigma^2\Big] \\ &= \frac{n-1}{n}\sigma^2.\end{aligned}$$

Prema tome, očekivanje statistike Θ ne podudara se s parametrom σ^2. Ovaj procjenitelj nije nepristran. Primjetimo ipak da se razlika očekivanja procjenitelja i parametra smanjuje povećavanjem veličine uzorka n.

Međutim, množenjem s konstantnim faktorom $\dfrac{n}{n-1}$ ovaj se procjenitelj može učiniti nepristranim:

Procjene disperzije

Ako je očekivanje a populacije X poznato, nepristrana procjena nepoznate disperzije σ^2 računa se formulom

$$D^2 := \frac{1}{n}\sum_{i=1}^{n}(X_i - a)^2. \qquad (6)$$

Ako su očekivanje a i disperzija σ^2 populacije X nepoznati, onda se nepristrani procjenitelj za disperziju računa formulom

$$S^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2. \qquad (7)$$

Statistika S^2 je valjana, jer joj disperzija teži k nuli. Vrijedi naime

$$\begin{aligned}\boldsymbol{D}(S^2) &= \boldsymbol{E}[(S^2 - \sigma^2)^2] \\ &= \boldsymbol{E}(S^4) - 2\sigma^2\boldsymbol{E}(S^2) + \sigma^4 = \boldsymbol{E}(S^4) - \sigma^4.\end{aligned}$$

Iz prikaza

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - a)^2 - \frac{1}{n(n-1)}\sum_{\substack{i,j=1\\ i\neq j}}^{n}(X_i - a)(X_j - a)$$

nakon kvadriranja ovog izraza i računanja očekivanja svakog člana, dobivamo izraz sličan (5):

$$\boldsymbol{D}(S^2) = \frac{1}{n}\Big(\mu_4 - \frac{n-3}{n-1}\sigma^4\Big). \qquad (8)$$

Primjer 10.3.

Da bi se utvrdila preciznost mjernog geodetskog uređaja koji nema sistematske pogreške, načinjeno je šest mjerenja. Dobiveni su rezultati (u metrima): 3540, 3582, 3555, 3578, 3564, 3548. Odredi nepristranu procjenu za varijancu, u slučajevima (a) ako je poznato da iznos mjerene veličine iznosi 3560 m, (b) ako nije poznat iznos mjerene veličine.

► (a) U ovom je slučaju poznato očekivanje slučajne varijable, jer ono mora biti jednako mjerenoj vrijednosti (zbog odsustva sistematske pogreške): $a = 3560$. Zato procjenu za varijancu računamo ovako:

$$\widehat{d}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - a)^2 = 232.17\ \mathrm{m}^2.$$

(b) Očekivanje je nepoznato, pa ga računamo iz uzorka:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n}x_i = 3561.17\ \mathrm{m}.$$

Nepristranu procjenu varijance računamo ovako:

$$\widehat{s}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = 276.97\ \mathrm{m}^2. \ ◄$$

10.1.8. Uporaba džepnog računala

Formula

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

nije najprikladnija za račun džepnim računalom. Ona zahtjeva izvođenje točno $5n+1$ operacija (za računala s inverznom notacijom, inače je broj neznatno veći). Pod operacijom se smatra svako unošenje podataka ili njihov poziv iz memorije, te svaka funkcijska ili aritmetička operacija.

Transformirajmo ovaj izraz na sljedeći način:

$$\widehat{s}^2 = \frac{1}{n-1}\Big(\sum_{i=1}^{n}x_i^2 - 2\overline{x}\sum_{i=1}^{n}x_i + n\overline{x}^2\Big) = \frac{1}{n-1}\Big(\sum_{i=1}^{n}x_i^2 - n\overline{x}^2\Big)$$

Sad je nužno napraviti točno $3n+5$ operacija. Zato ćemo procjenu disperzije računati ovom formulom.

Praktički svi džepni kalkulatori imaju ugrađene elementarne statističke funkcije. Neki među njima specijalizirani su upravo za rješavanje statističkih zadataka. Na različitim računalima mogu postojati različiti načini korištenja tih funkcija, ali zajednički principi mogu se opisati ovako.

Niz podataka $x_1, x_2, \ldots, x_n$ unosi se posebnom tipkom, obično označenom s $\boxed{\Sigma}$. Na koncu unosa, u posebnim registrima spremljeni su sljedeći podatci:

- volumen uzorka n;
- zbroj elemenata uzorka, $\sum x_i$;
- zbroj kvadrata elemenata uzorka, $\sum x_i^2$.

U posebnim su registrima također spremljene izračunate statističke funkcije. Na boljim računalima, pozivi tih registara nalaze se na posebnim tipkama označenim s $\boxed{\overline{x}}$, $\boxed{s}$ i $\boxed{\sigma}$.

10.1.9. Računanje s grupiranim podatcima

Podatci dani u uzorku vrlo su često grupirani u *razrede*. Uzorak tada ima ovakav oblik

x_1	n_1
x_2	n_2
$\vdots$	$\vdots$
x_r	n_r

Ovdje je $n = n_1 + \ldots + n_r$ volumen uzorka. Sredina i disperzija uzorka računaju se tada formulama

$$\begin{aligned}\overline{x} &= \frac{1}{n}\sum_{i=1}^{r}n_i x_i, \\ \widehat{s} &= \frac{1}{n-1}\sum_{i=1}^{r}n_i(x_i - \overline{x})^2 = \frac{1}{n-1}\Big(\sum_{i=1}^{r}n_i x_i^2 - n\overline{x}^2\Big).\end{aligned}$$

Primjer 10.4.

Odredimo procjenu za očekivanje i disperziju na temelju uzorka normalne populacije:

x_i	2560	2600	2620	2650	2700
n_i	2	3	10	4	1

► Volumen uzorka je

$$n = n_1 + \ldots + n_5 = 20.$$

Računanje očekivanja i disperzije olakšano je ako vrijednosti slučajne varijable translatiramo za isti iznos C. Tu je C po volji odabrani broj. Pri tom vrijedi

$$\boldsymbol{E}(X) = C + \boldsymbol{E}(X - C), \qquad \boldsymbol{D}(X) = \boldsymbol{D}(X - C)$$

(U računu koji slijedi koristit ćemo samo prvo svojstvo.) Za pogodnu konstantu u ovom primjeru možemo uzeti $C = 2620$:

$$\begin{aligned}\overline{x} &= 2620 + \frac{1}{n}\sum_{i=1}^{5}n_i(x_i - 2620) \\ &= 2620 + \frac{1}{20}\Big[-60\cdot 2 + (-20)\cdot 3 + 30\cdot 4 + 80\Big] = 2620 + 1 = 2621. \\ \widehat{s}^2 &= \frac{1}{n-1}\Big(\sum_{i=1}^{5}n_i x_i^2 - n\overline{x}^2\Big) = 967.4 \ ◄\end{aligned}$$

Primjer 10.5. ■ **Određivanje varijance na temelju zadanog uzorka varijanci**

Pri kontroli kvalitete nekog proizvoda, ispituje se varijanca na kontrolnim uzorcima tijekom svakog dana. Dobivene su vrijednosti $s_1^2, s_2^2, \ldots, s_k^2$, na temelju uzoraka veličina $n_1, n_2, \ldots, n_k$. Kako ćemo odrediti procjenu varijance ove populacije?

► Trebamo odrediti nepristrani procjenitelj za nepoznatu varijancu σ^2. Izabrat ćemo statistiku

$$\widehat{\Theta} = \frac{a_1 S_1^2 + a_2 S_2^2 + \ldots + a_k S_k^2}{A}$$

gdje su $a_1, a_2, \ldots, a_k$ i A konstante koje treba odrediti.

Očekivanje ove statistike je

$$\boldsymbol{E}(\widehat{\Theta}) = \frac{a_1\boldsymbol{E}(S_1^2) + \ldots + a_k\boldsymbol{E}(S_k^2)}{A} = \frac{a_1 + \ldots + a_k}{A}\cdot\sigma^2$$

Statistika će biti nepristrana ako je $A = a_1 + \ldots + a_k$. Konstante $a_1, \ldots, a_k$ možemo birati po volji, ali je prirodno da one odgovaraju veličinama pojedinih uzoraka. Tako će dnevne procjene temeljene na većem uzorku imati veću težinu u konačnoj procjeni. Prema tome, tražena procjena je

$$\widehat{\vartheta} = \frac{n_1 s_1^2 + n_2 s_2^2 + \ldots + n_k s_k^2}{n_1 + n_2 + \ldots + n_k}. \ ◄$$

10.1.10. * Nepristrana procjena standardnog odstupanja

Pokazali smo da je

$$D^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - a)^2$$

nepristrana procjena disperzije σ^2. **Standardno odstupanje** (**devijacija**) definira se kao korijen disperzije, $\sigma = \sqrt{\sigma^2}$.

Logično je postaviti pitanje: je li veličina

$$D = \sqrt{D^2} = \Big(\frac{1}{n}\sum_{i=1}^{n}(X_i - a)^2\Big)^{1/2}$$

nepristrana procjena za standardno odstupanje?

Može izgledati neobično, ali odgovor je negativan. Razlog tome je što funkcija drugog korijena "jače skuplja" velike brojeve od malih. Za bilo koju nedegeneriranu pozitivnu slučajnu varijablu Y općenito vrijedi

$$\boldsymbol{E}(Y) < \sqrt{\boldsymbol{E}(Y^2)}.$$

(Ova nejednakost slijedi iz Cauchy-Schwarz-Bunjakovskijeve nejednakosti.) Zato je

$$\boldsymbol{E}(D) < \sqrt{\boldsymbol{E}(D^2)} = \sqrt{\sigma^2} = \sigma.$$

Nepristranu procjenu za standardno odstupanje praktički je nemoguće utvrditi u općem slučaju, za bilo koju distribuciju populacije X. Ako populacija ima normalnu razdiobu, onda se može dokazati da nepristrana procjena glasi

$$\tilde{D} = k_{n+1}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - a)^2}, \qquad (9)$$

pri čemu se koeficijent k_n računa formulom

$$k_n = \sqrt{\frac{n-1}{2}}\cdot\frac{\Gamma\Big(\dfrac{n-1}{2}\Big)}{\Gamma\Big(\dfrac{n}{2}\Big)}. \qquad (10)$$

Na isti se način dobiva nepristrana procjena standardnog odstupanja ukoliko očekivanje nije poznato:

$$\tilde{S} = k_n\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2},$$

Za velike vrijednosti od n može se koristiti jednostavnija formula:

$$\tilde{S} = \sqrt{\frac{1}{n-1.45}\sum_{i=1}^{n}(X_i - \overline{X})^2},$$

Za male vrijednosti od n može biti korisna sljedeća tablica:

n	k_n	n	k_n	n	k_n
3	1.1284	10	1.0280	30	1.0087
4	1.0853	12	1.0230	35	1.0072
5	1.0640	15	1.0181	40	1.0064
6	1.0506	20	1.0134	45	1.0056
7	1.0423	25	1.0104	50	1.0051

Međuvrijednosti se mogu utvrditi interpolacijom.

U većini zadataka i teoriji koja slijedi, kao procjenu za odstupanje ćemo ipak, jednostavnosti radi, koristiti korijen varijance. Izuzetak će biti zadatci u kojima se eksplicitno traži nepristrana procjena standardnog odstupanja.

10.2. Kriterij najveće izglednosti

Pretpostavimo da nam je u razdiobi slučajne varijable X nepoznata vrijednost jednog parametra ϑ. Označimo sa $f(\vartheta, x)$ zakon razdiobe te slučajne varijable,

$$f(\vartheta, x) = \boldsymbol{P}_\vartheta(\{X = x\}), \qquad x \in S,$$

ako je X diskretnog tipa, odnosno, neka je

$$f(\vartheta, x) = f_\vartheta(x),$$

funkcija gustoće, ako je X neprekinuta slučajna varijabla. Indeks ϑ označava da se vjerojatnost događaja $\{X = x\}$ i vrijednost funkcije gustoće $f(x)$ varijable X računaju uz pretpostavku da je nepoznata vrijednost parametra, o kojem ovise te vrijednosti, jednaka ϑ. Tu nepoznatu vrijednost ćemo pokušati procijeniti iz vrijednosti uzorka $(x_1, \ldots, x_n)$.

Kriterij najveće izglednosti

Neka je $x_1, x_2, \ldots, x_n$ realizacija uzorka populacije X, čija funkcija gustoće $f(\vartheta, x)$ ovisi o nepoznatom parametru ϑ. **Funkcija izglednosti** definira se kao umnožak

$$L(\vartheta, x_1, \ldots, x_n) := f(\vartheta, x_1) f(\vartheta, x_2) \cdots f(\vartheta, x_n). \tag{11}$$

Za procjenu parametra ϑ uzimamo onu vrijednost $\widehat{\vartheta}$ za koju funkcija izglednosti poprima globalni maksimum.

Zašto se ova funkcija naziva *funkcija izglednosti*? Za zadani x, vrijednost $f(\vartheta, x)$ opisuje vjerojatnost da slučajna varijabla poprimi vjerojatnost u okolišu broja x. Zato umnožak (11) predstavlja vjerojatnost da uzorak $(X_1, X_2, \ldots, X_n)$ poprimi vrijednost u okolišu od $(x_1, x_2, \ldots, x_n)$. Postavlja se pitanje: za koju će vrijednost parametra ϑ ta vjerojatnost biti najveća? Na taj način dobivamo *kriterij za odabir procjene parametra* ϑ. Za ϑ ćemo odabrati onaj parametar koji maksimizira funkciju izglednosti. Na taj način *maksimiziramo vjerojatnost pojavljivanja* uzorka koji se ostvario!

Opravdanje ovog uvjeta je intuitivna pretpostavka da onom uzorku koji se stvarno realizira trebamo dati prednost u odnosu na one koji se nisu ostvarili. Nakon što nam je poznata vrijednost uzorka $(x_1, \ldots, x_n)$, uzimamo onu vrijednost parametra ϑ za koju ta realizacija ima najveću vjerojatnost pojavljivanja, veću nego bilo koja druga realizacija.

Primjer 10.6. Procjena parametra eksponencijalne razdiobe

Vrijeme X ispravnog rada uređaja kojem se karakteristike ne mijenjaju vremenom, dobro je opisano eksponencijalnom razdiobom, s gustoćom

$$f(x) = \lambda e^{-\lambda x}, \qquad x > 0.$$

Ovdje je λ nepoznati parametar razdiobe. Bilježeni su rezultati na probnom uzorku i dobiven niz $x_1, x_2, \ldots, x_n$. Na temelju tih rezultata, koriseći se kriterijem najveće izglednosti, treba procijeniti očekivanje varijable X.

▶ U ovom je primjeru

$$\begin{aligned} L(\lambda, x_1, \ldots, x_n) &= f(\lambda, x_1) f(\lambda, x_2) \cdots f(\lambda, x_n) \\ &= \lambda e^{-\lambda x_1} \cdot \lambda e^{-\lambda x_2} \ldots \lambda e^{-\lambda x_n} \\ &= \lambda^n e^{-\lambda z}. \end{aligned}$$

Tu smo označili $z = \sum_{i=1}^n x_i$. Izračunajmo maksimum ove funkcije:

$$\frac{\partial L}{\partial \lambda} = \lambda^{n-1} e^{-\lambda z}(-\lambda z + n).$$

Ova se derivacija poništava kad je $\lambda = 0$ ili kad je $-\lambda z + n = 0$. Prva je mogućnost besmislena, a iz druge slijedi

$$\widehat{\lambda} = \frac{n}{z} = \frac{n}{x_1 + x_2 + \ldots + x_n} = \frac{1}{\overline{x}}.$$

Za eksponencijalnu funkciju je poznato da vrijedi $\boldsymbol{E}(X) = 1/\lambda$. zato je procjena za očekivanje

$$\overline{x} = \frac{1}{\widehat{\lambda}} = \frac{x_1 + x_2 + \ldots + x_n}{n}.$$

Ovaj je rezultat u skladu sa standardnom procjenom za očekivanje slučajne varijable. ◀

— ◆ —

Korisno je primjetiti da je funkcija izglednosti L uvijek pozitivna, pa je stoga definirana i funkcija $\ln L$. S obzirom da vrijedi

$$(\ln L)' = \frac{L'}{L},$$

ova funkcija poprima maksimum u istim točkama kao i L. Vrlo je često nju praktičnije derivirati nego funkciju L. U prethodnom primjeru je

$$\ln L = n \ln \lambda - \lambda \sum_{i=1}^n x_i,$$

$$\frac{\partial \ln L}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^n x_i = 0,$$

pa slijedi isti zaključak kao prije.

10.2.1. Procjena vjerojatnosti događaja

Događaj A ima nepoznatu vjerojatnost realiziranja p. Kako ćemo procijeniti tu vjerojatnost?

Više je mogućih odgovora. Navedimo dva najjednostavnija.

Primjer 10.7. Procjena vjerojatnosti korištenjem relativne frekvencije

Pokus u kojem se događaj A može ostvariti ponavljamo n puta, pri nepromijenjenim uvjetima. U svakom ponavljanju bilježimo je li se događaj zbio ili nije. Na taj je način dobiven uzorak $x_1, x_2, \ldots, x_n$, pri čemu je $x_k = 0$ ako se A nije ostvario, a $x_k = 1$ ako se A ostvario. Na temelju tog uzorka treba procijeniti vjerojatnost p.

▶ Rezultat pokusa prate indikatorske slučajne varijable koje su nezavisne kopije slučajne varijable

$$X \sim \begin{pmatrix} 0 & 1 \\ q & p \end{pmatrix}.$$

Ova je razdioba zadana vjerojatnostima (ovisnim o nepoznatom parametru p):

$$f(p, x) = p^x (1-p)^{1-x}, \qquad x = 0 \text{ ili } 1.$$

Funkcija izglednosti je

$$L(p, x_1, \ldots, x_n) = \prod_{k=1}^n p^{x_k} (1-p)^{1-x_k}.$$

Zato je

$$\ln L = \sum_{k=1}^n \Big[x_k \ln p + (1 - x_k) \ln(1-p) \Big],$$

$$\frac{\partial \ln L}{\partial p} = \sum_{k=1}^n \left(\frac{x_k}{p} - \frac{1 - x_k}{1-p} \right) = \frac{m}{p} - \frac{n-m}{1-p} = 0. \tag{12}$$

Tu smo s m označili

$$m = x_1 + \ldots + x_n,$$

a taj je zbroj jednak broju pojavljivanja događaja A u n pokusa.

Sređivanjem jednakosti (12) dobivamo $p = \dfrac{m}{n}$. ◀

Primjer 10.8. Procjena vjerojatnosti korištenjem geometrijske razdiobe

Ponavljamo pri nepromijenjenim uvjetima pokus u kojem se događaj A može ostvariti i bilježimo broj pokusa kad se to dogodilo. Čitav se postupak ponavlja n puta. Na taj je način dobiven uzorak $x_1, x_2, \ldots, x_n$. Na temelju tog uzorka treba procijeniti vjerojatnost p.

▶ Slučajna varijabla koja opisuje pojavljivanje događaja A ima geometrijsku razdiobu s parametrom p. Ona je zadana vjerojatnostima

$$f(p, x) = p(1-p)^{x-1}, \qquad x = 1, 2, 3, \ldots.$$

Funkcija izglednosti je (nepoznati parametar i dalje označavamo s p):

$$L(p, x_1, \ldots, x_n) = p^n (1-p)^{x_1 - 1} \cdots (1-p)^{x_n - 1}.$$

Sada imamo

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

$$\frac{\partial \ln L}{\partial p} = \frac{n}{p} - \frac{\sum x_i - n}{1-p} = 0.$$

Iz posljednje jednakosti, nakon sređivanja, dobivamo

$$\widehat{p} = \frac{n}{x_1 + x_2 + \ldots + x_n} = \frac{1}{\overline{x}}. \quad ◀$$

Primjer 10.9. Procjena vjerojatnosti

Postotak bijelih kuglica u kutiji je nepoznat. Zagrabili smo n kuglica i pobrojali m bijelih. Kolika je procjena za postotak bijelih kuglica?

▶ Taj postotak jednak je vjerojatnosti da izvučena kuglica iz kutije bude bijela. Neka je p ta vjerojatnost. Slučajna varijabla X koju promatramo je broj bijelih kuglica u uzorku veličine n. Njezina je razdioba $X \sim B(n, p)$. Zato vrijedi

$$f(p, x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \ldots, n.$$

Odavde je

$$\begin{aligned} \frac{\partial}{\partial p} \ln f(p, x) &= \frac{\partial}{\partial p} [x \ln p + (n - x) \ln(1-p)] \\ &= \frac{x}{p} - \frac{n-x}{1-p} = 0 \implies \widehat{p} = \frac{x}{n} \end{aligned}$$

Ako se u uzorku pojavilo m bijelih kuglica, onda je najbolja procjena $\widehat{p} = \dfrac{m}{n}$. ◀

Primjer 10.10. Procjena parametra Poissonove razdiobe

Neka X ima Poissonovu razdiobu, $X \sim \mathscr{P}(\lambda)$, λ nepoznat. Procijenimo vrijednost od λ.

▶ Sada je

$$f(\lambda, x) = \boldsymbol{P}(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}.$$

Funkcija izglednosti je

$$L(\lambda, x_1, \ldots, x_n) = \frac{\lambda^{x_1 + \ldots + x_n}}{x_1! \cdots x_n!} e^{-n\lambda},$$

$$\ln L(\lambda, x_1, \ldots, x_n) = -n\lambda + (x_1 + \ldots + x_n) \ln \lambda - \sum \ln(x_i!),$$

$$\frac{\partial \ln L}{\partial \lambda} = -n + \frac{x_1 + \ldots + x_n}{\lambda} = 0,$$

$$\widehat{\lambda} = \frac{x_1 + \ldots + x_n}{n} = \overline{x}. \quad ◀$$

— ◆ —

Kriterijom najveće izglednosti možemo odrediti i više od jednog nepoznatog parametra. Ako funkcija izglednosti ima oblik

$$L(\vartheta_1, \ldots, \vartheta_s, x_1, \ldots, x_n) = f(\vartheta_1, \ldots, \vartheta_s, x_1) \cdots f(\vartheta_1, \ldots, \vartheta_s, x_n)$$

onda nepoznate parametre $\vartheta_1, \ldots, \vartheta_s$ dobivamo iz uvjeta

$$\frac{\partial L(\vartheta_1, \ldots, \vartheta_s, x_1, \ldots, x_n)}{\partial \vartheta_i} = 0, \qquad i = 1, 2, \ldots, s. \tag{13}$$

Primjer 10.11. Procjena parametara normalne razdiobe

Slučajna varijabla X ima normalnu razdiobu $\mathscr{N}(a, \sigma^2)$ s nepoznatim i očekivanjem a i disperzijom σ^2. Odredimo procjenu tih parametara koristeći kriterij najveće izglednosti.

▶ Pripadna funkcija izglednosti je

$$L(a, \sigma) = L(a, \sigma, x_1, \ldots, x_n) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\Big\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - a)^2 \Big\}.$$

Logaritam ove funkcije je

$$\ln L(a, \sigma) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - a)^2.$$

Ekstrem dobivamo ako vrijedi:

$$\frac{\partial}{\partial a} \ln L(a, \sigma) = \frac{1}{2\sigma^2} \cdot 2 \sum_{i=1}^n (x_i - a) = 0$$

i odavde

$$\widehat{a} = \frac{1}{n} \sum_{i=1}^n x_i,$$

$$\frac{\partial}{\partial \sigma} \ln L(a, \sigma) = -\frac{n}{2} \cdot \frac{2}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^n (x_i - a)^2 = 0$$

i odavde

$$\widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \widehat{a})^2.$$

Primjer 10.12. Odabir statistike

Zadana je jednolika razdioba na intervalu $[0, c]$, c nepoznat. Vrijednost od c trebali bismo procijeniti na osnovu uzorka: nekoliko na sreću odabranih brojeva iz tog intervala. Koju statistiku je prikladno upotrebiti?

▶ Recimo, zbog jednostavnijeg razmatranja, da je uzorak dao sljedeće vrijednosti

$$3,\ 5,\ 8,\ 1,\ 3,\ 2,\ 6,\ 2.$$

Neka je X slučajna varijabla: vrijednost na sreću odabranog broja iz intervala $[0, c]$. Njezina je funkcija gustoće

$$f(c, x) = \begin{cases} 1/c, & 0 \leqslant x \leqslant c, \\ 0, & \text{inače.} \end{cases}$$

Pri tom vrijedi $\boldsymbol{E}(X) = \dfrac{c}{2}$, $\boldsymbol{D}(X) = \dfrac{c^2}{12}$.

Odaberimo statistiku pomoću koje možemo odrediti c. Budući je $\boldsymbol{E}(X) = c/2$, a znamo statistiku pomoću koje određujemo $\boldsymbol{E}(X)$, onda možemo izabrati statistiku

$$\Theta_1 = 2\overline{X} = \frac{2}{n} \sum_{k=1}^n X_k.$$

Vrijedi

$$\boldsymbol{E}(\Theta_1) = \frac{2}{n} \sum_{k=1}^n \boldsymbol{E}(X_k) = \frac{2}{n} \cdot \frac{c}{2} \cdot n = c$$

pa je ova statistika nepristrana. Međutim, ona nije baš najsretnije odabrana. Ukoliko uzorak poprimi opisanu vrijednost, onda dobivamo sljedeću procjenu za c

$$\widehat{c} = \frac{2}{8}(3 + 5 + 8 + 1 + 3 + 2 + 6 + 2) = 7.5$$

što je apsurd, budući se pojavila realizacija 8.

Pokušajmo odrediti prikladniju statistiku. Pogledajmo što će dati kriterij najveće izglednosti. Imamo

$$L(c, x_1, \ldots, x_n) = \begin{cases} \left(\dfrac{1}{c} \right)^n, & x_k \leqslant c, \ \forall k, \\ 0, & \text{inače.} \end{cases}$$

Maksimum se postiže ako je c najmanji moguć, a to je za $\widehat{c} = \max_{1 \leqslant k \leqslant n} x_k$.

Time se nameće statistika

$$Y = \max\{X_1, \ldots, X_n\}.$$

U zadanom uzorku, dobili bismo procjenu $\widehat{c} = 8$. Očevidno je da niti ona nije posve zadovoljavajuća: ako je *zaista* $c = 8$, nevjerovatno je da se baš ta maksimalna vrijednost i izabere.

Pogledajmo je li ova statistika nepristrana. Odredimo njezinu razdiobu:

$$\begin{aligned} F_Y(y) &= \boldsymbol{P}(Y < x) = \boldsymbol{P}(X_1 < x, \ldots, X_n < x) \\ &= \boldsymbol{P}(X < x)^n = \left(\frac{x}{c} \right)^n, \quad 0 < x \leqslant c. \end{aligned}$$

Funkcija gustoće je (u ovisnosti o parametru c)

$$f(c, x) = \frac{n}{c^n} x^{n-1}, \quad 0 < x < c.$$

Odavde dobivamo očekivanje

$$\boldsymbol{E}(Y) = \frac{n}{c^n} \int_0^c x \cdot x^{n-1} \,\mathrm{d}x = \frac{n}{n+1} c.$$

Statistika nije nepristrana. Stoga činimo korekciju i promatramo drugu statistiku

$$\Theta_2 = \frac{n+1}{n} Y = \frac{n+1}{n} \max\{X_1, \ldots, X_n\}$$

koja daje nepristranu procjenu za c. ◀

— ◆ —

Usporedimo statistike Θ_1 i Θ_2 iz ovog primjera. Vrijedi

$$\boldsymbol{D}(\Theta_1) = \boldsymbol{D}(2\overline{X}) = 4\boldsymbol{D}\left(\frac{X_1 + \ldots + X_n}{n} \right) = 4 \cdot \frac{\boldsymbol{D}(X)}{n} = \frac{c^2}{3n}.$$

Izračunajmo disperziju statistike Y:

$$\boldsymbol{E}(Y^2) = \frac{n}{c^n} \int_0^c x^2 \cdot x^{n-1} \,\mathrm{d}x = \frac{n}{n+2} c^2$$

i odavde

$$\begin{aligned} \boldsymbol{D}(Y) = \boldsymbol{E}(Y^2) - \boldsymbol{E}(Y)^2 &= \left(\frac{n}{n+2} - \frac{n^2}{(n+1)^2} \right) c^2 \\ &= \frac{n}{(n+2)(n+1)^2} c^2. \end{aligned}$$

Sada je

$$\boldsymbol{D}(\Theta_2) = \boldsymbol{D}\Big(\frac{n+1}{n} Y \Big) = \frac{(n+1)^2}{n^2} \cdot \frac{n}{(n+2)(n+1)^2} c^2 = \frac{c^2}{n(n+2)}.$$

Čim je $n \geqslant 2$, statistika Θ_2 je efikasnija od statistike Θ_1.

Primjer 10.13. Nepristranost i valjanost statistike

Kao procjenu za nepoznato očekivanje $a = \boldsymbol{E}(X)$ možemo uzeti statistike

$$\begin{aligned} \Theta_1 &= \overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}, \\ \Theta_2 &= X_1. \end{aligned}$$

Provjerimo da su obje procjene nepristrane. Koja među njima je valjana?

▶ Za statistiku Θ_1 znamo da je nepristrana. Isto vrijedi i za Θ_2:

$$\boldsymbol{E}(\Theta_2) = \boldsymbol{E}(X_1) = a.$$

Provjerimo da je Θ_1 valjana. Iz poznatog svojstva

$$\boldsymbol{D}(\Theta_1) = \frac{1}{n^2} \sum_{i=1}^n \boldsymbol{D}(X_i) = \frac{\sigma^2}{n}$$

s pomoću Čebiševljeve nejednakosti dobivamo

$$\boldsymbol{P}(|\Theta_1 - a| > \varepsilon) \leqslant \frac{\boldsymbol{D}(\Theta_1)}{\varepsilon^2} = \frac{\sigma^2}{\varepsilon^2 n} \to 0$$

kad $n \to \infty$. Dakle, $\Theta_1 = \overline{X}$ je valjana procjena za očekivanje.

Za statistiku Θ_2 vrijedi pak

$$\boldsymbol{P}(|\Theta_2 - \vartheta| > \varepsilon) = \boldsymbol{P}(|X_1 - \vartheta| > \varepsilon) > 0$$

čim je razdioba od X_1 netrivijalna. Zato Θ_2 nije valjana statistika. ◀