Korisne metode opisa skupova podataka

Balog, Karla

Sveučilište J. J. Strossmayera u Osijeku
Odjel za matematiku
Sveučilišni preddiplomski studij matematike

Karla Balog

Korisne metode opisa skupova podataka

Završni rad

Osijek, 2020.

Sveučilište J. J. Strossmayera u Osijeku
Odjel za matematiku
Sveučilišni preddiplomski studij matematike

Karla Balog

Korisne metode opisa skupova podataka

Završni rad

Mentor: prof. dr. sc. Mirta Benšić

Osijek, 2020.

Sažetak

U ovome radu navodimo što su uzorak i slučajan uzorak, definiramo parametarski statistički model te neke granične rezultate. Zatim opisujemo, kroz primjere, statistički niz podataka, prvo grafički, a zatim i numerički pomoću parametara lokacija i parametara raspršenja. Na poslijetku, radimo točkovnu procjenu parametara parametarskog statističkog modela pomoću danog uzorka.

Ključne riječi: parametarski statistički model, parametri lokacije, parametri raspršenja, grafički prikaz, točkovna procjena

Abstract

In this paper, the definition of a sample and a random sample will be listed as well as the definition of the parametric statistical model and some important limit theorems. We then describe, through examples, the statistical data sets, first graphically and then numerically using measures of location and measures of dispersion. Finally, estimation of the parameters for a hypothesized parametric statistical model will be made using a sample of data.

Key words: parametric statistical model, measures of location, measures of dispersion, graphical respresentation, point estimation

Sadržaj

1 Statistički model

1.1 Populacija i uzorak

Kao matematička disciplina, statistika se bavi prikupljanjem podataka, njihovim opisivanjem te različitim metodama zaključivanja o prikupljenim podacima.

Zaključivanje se vrši na populaciji, no da bismo mogli zaključivati, prvo je potrebno prikupiti podatke. S obzirom da je često nemoguće skupiti potrebne podatke iz cijele populacije, podaci skupljaju i analiziraju na jednom njezinom dijelu - *uzorku* - koji je *slučajan* ukoliko svaka jedinka populacije ima jednaku šansu biti odabrana.

Prikupljene podatke organiziramo u obliku tablice koju nazivamo *baza podataka*. Svi podaci koje prikupimo prilikom jednog istraživanja su slučajni jer kada bismo ponovno proveli istraživanje na nekom drugom uzorku, dobili bismo drugačije rezultate. Zbog toga se svaki podatak koji prikupimo smatra realizacijom jedne slučajne varijable. Svaki slučajni vektor $X = (X_1, \ldots, X_n) \in \mathbb{R}^n$ nazivamo *slučajni uzorak*, a svaku njegovu realizaciju $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ nazivamo *uzorak*.

Ono što nas zanima jest donošenje zaključaka o distribuciji tog slučajnog vektora, X, koju najčešće opisujemo funkcijom distribucije F:

$$F(x_1, \ldots, x_n) = P(X_1 \leq x_1, \ldots, X_n \leq x_n). \tag{1}$$

Sve što nam je poznato o funkciji distribucije slučajnog vektora opisujemo **statističkim modelom**, familijom funkcija distribucije slučajnog vektora. Najčešće se bavimo statističkim modelom **jednostavnog slučajnog uzorka** (j.s.u.) u kojemu su slučajne varijable $X_1, \ldots, X_n$ međusobno nezavisne i jednako distribuirane s funkcijom distribucije F.

1.2 Parametarski statistički model

Često funkcija distribucije F ovisi o nekom k-dimenzionalnom parametaru θ, pa ju označavamo s F_θ i kažemo da se radi o **parametarskom statističkom modelu**

$$\mathcal{P} = \{F_\theta : \theta \in \Theta\},$$

gdje je $\Theta \subseteq \mathbb{R}^k$ **prostor parametara**. U suprotnom je model neparametarski.

Primjer 1.1. *Želimo zaključivati o broju valjanih kišobrana koje jedna tvornica proizvede prilikom jedne ture proizvodnje. Najbolji način za to jest uzeti nasumičan uzorak od n kišobrana i prebrojati među njima ispravne i neispravne kišobrane što možemo označiti s:*

$$0 - \textit{neispravan kišobran}$$
$$1 - \textit{ispravan kišobran.}$$

Dobivamo uzorak $x_1, \ldots, x_n$, gdje je $x_i \in \{0, 1\}, \forall i$, koji je realizacija slučajnog uzorka $X_1, \ldots, X_n$. Distribuciju od X_i možemo zapisati tablicom distribucije:

$$\begin{pmatrix} 0 & 1 \\ 1-\theta & \theta \end{pmatrix}$$

gdje je

$$P(X_i = 1) = \theta, \quad P(X_i = 0) = 1 - \theta, \quad i \in \{1, \ldots, n\}.$$

Statistički model je model jednostavnog slučajnog uzorka iz Bernoullijeve distribucije:

$$\mathcal{P} = \{F_\theta : \theta \in \langle 0, 1\rangle\},$$

pri čemu je

$$\begin{aligned} F_\theta(t_1, \ldots, t_n) &= P(X_1 \leq t_1, \ldots, X_n \leq t_n) \\ &= \prod_{i=1}^{n} P(X_i \leq t_i) = \prod_{i=1}^{n} F_\theta(t_i) \\ &= \prod_{i=1}^{n} \theta_i^x (1-\theta)^{1-x_i}. \end{aligned}$$

Kako je očekivanje Bernoullijeve slučajne varijable jednako vjerojatnosti θ, procjenu nepoznatog parametra vjerojatnosti možemo dobiti procjenom očekivanja (više u potpoglavlju 3.1, str. 21):

$$\overline{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Primjer 1.2. *Još neki primjeri parametarskih statističkih modela:*

a) statistički model jednostavnog slučajnog uzorka iz Poissonove distribucije:

$$\mathcal{P} = \left\{ p_\lambda(x) = \frac{\lambda^j}{j!} e^{-\lambda} : j \in \mathbb{N} \right\}$$

b) statistički model jednostavnog slučajnog uzorka iz normalne distribucije:

$$\mathcal{P} = \left\{ p_{(\mu,\sigma^2)}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} : (\mu, \sigma^2) \in \mathbb{R} \times \langle 0, \infty \rangle \right\}$$

c) statistički model jednostavnog slučajnog uzorka iz binomne distribucije:

$$\mathcal{P} = \left\{ p_{(n,\theta)}(x) = \binom{n}{i} \theta^i (1-\theta)^{n-i}, i = 0, \ldots, n : (n, \theta) \in \mathbb{N}_0 \times \langle 0, 1 \rangle \right\}$$

Prije nego što kažemo nešto više o procjeni parametara, bilo bi dobro prisjetiti se dva osnovna granična rezultata koji će nam biti važni prilikom promatranja velikih uzoraka, odnosno kada $n \to \infty$, iako smatramo da su oni približno točni i za konačne slučajne uzorke kada je n dovoljno velik.

1.3 Slabi zakon velikih brojeva i centralni granični teorem

Definicija 1.1. *Za niz slučajnih varijabli $(X_n, n \in \mathbb{N})$ kažemo da je* ***niz nezavisnih jednako distribuiranih slučajnih varijabli*** *(niz n.j.d.) ako sve slučajne varijable $X_1, X_2, \ldots$ imaju jednaku distribuciju i ako su $\forall n \in \mathbb{N}$ slučajne varijable $X_1, X_2, \ldots$ nezavisne.*

Teorem 1.1. ***Slabi zakon velikih brojeva***
Neka je $(X_n, n \in \mathbb{N})$ niz n.j.d. s varijancom μ i očekivanjem $\sigma^2 < \infty$ te $\overline{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$. Tada $\forall \varepsilon > 0$ vrijedi

$$\lim_{n \to \infty} P\Big(\big| \overline{X}_n - \mu \big| > \varepsilon \Big) = 0.$$

(prosjek konvergira po vjerojatnosti prema svom očekivanju).

Za slučajne varijable, osim po vjerojatnosti, također možemo promatrati i konvergenciju po distribuciji.

Definicija 1.2. *Neka je* $(X_n, n \in \mathbb{N})$ *niz slučajnih varijabli s pripadnim nizom funkcija distribucije* $(F_n, n \in \mathbb{N})$ *i* X *slučajna varijabla s funkcijom distribucije* F. *Kažemo da niz* $(X_n, n \in \mathbb{N})$ ***konvergira po distribuciji*** *prema* X, *u oznaci*

$$X_n \xrightarrow{D} X$$

ako u svakoj točki u kojoj je F *neprekidna vrijedi:*

$$\lim_{n\to\infty} F_n(x) = F(x).$$

Ova definicija, odnosno konvergencija po distribuciji znači da vjerojatnost $PX_n \leq x$ možemo aproksimirati s $F(x)$ kada $n \to \infty$.

Teorem 1.2. ***Centralni granični teorem (CGT)***
Neka je $(X_n, n \in \mathbb{N})$ *niz n.j.d. slučajnih varijabli s očekivanjem* μ *i varijancom* $\sigma^2 < \infty$ *te* $S_n = \sum_{i=1}^{n} X_i$. *Tada*

$$\frac{S_n - ES_n}{\sqrt{VarS_n}} \xrightarrow{D} Z \sim \mathcal{N}(0,1), \quad n \to \infty.$$

CGT kaže da niz standardiziranih parcijalnih suma niza n.j.d. slučajnih varijabli konvergira po distribuciji prema standardnoj normalnoj slučajnoj varijabli.

Ako uz očekivanje $EX_1 = \mu$ i varijancu s $VarX_1 = \sigma^2$, označimo prosjek s $\overline{X}_n = \frac{1}{2}\sum_{i=1}^{n} X_i = \frac{1}{n}S_n$, zbog pretpostavke da je $X_1, X_2, \ldots$ niz n.j.d. vrijedi:

$$ES_n = E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} EX_i = \sum_{i=1}^{n} \mu = n\mu$$

i

$$VarS_n = Var\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} VarX_i = \sum_{i=1}^{n} \sigma^2 = n\sigma^2$$

pa iz teorema slijedi

$$\frac{S_n - ES_n}{\sqrt{VarS_n}} = \frac{S_n - n\mu}{\sqrt{n\sigma^2}} = \frac{n\left(\frac{S_n}{n} - \mu\right)}{\sigma\sqrt{n}} = \frac{\overline{X}_n - \mu}{\sigma}\sqrt{n} \xrightarrow{D} Z \sim \mathcal{N}(0,1), \quad n \to \infty.$$

Odnosno, prosjek $\overline{X}_n$ asimptotski ima normalnu distribuciju $\mathcal{N}(\mu, \frac{\sigma^2}{n})$.

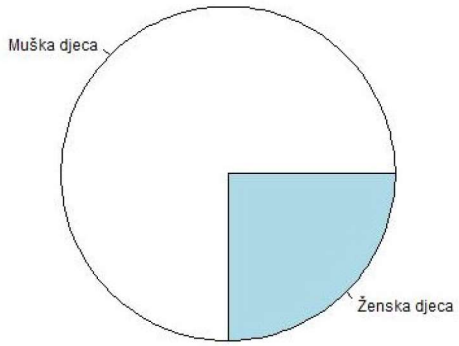
2 Parametri i grafički prikazi niza podataka

U većini slučajeva, kada promatramo danu bazu podataka, ne možemo odmah reći nešto općenito o populaciji iz koje podaci dolaze. Stoga, kako bismo dobili detaljniji uvid u podatke, tj. slučajnu varijablu iz čije distribucije podaci dolaze, u ovom poglavlju opisat ćemo grafičke prikaze te parametre niza podataka (brojčane vrijednosti koje se definiraju pomoću danog niza podataka $x_1, \ldots, x_n$ o slučajnoj varijabli X).

2.1 Grafički prikazi

2.1.1 Kružni i stupčasti dijagram

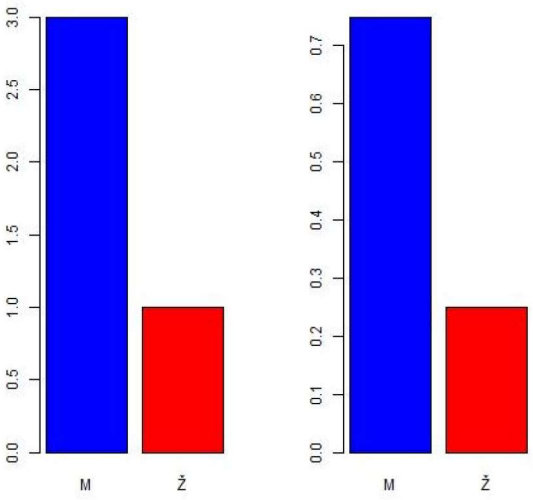
Za vizualni prikaz diskretne slučajne varijable s konačnom slikom, kao i za kvalitativne podatke, koristimo kružni dijagram (eng. *pie chart*). Recimo, ako je u nekoj obitelji ukupan broj djece 4, od kojih su 3 dječaka i 1 djevojčica, što su frekvencije tih podataka u uzorku, onda su odgovarajuće relativne frekvencije 3/4, odnosno 1/4 (dječaci čine 75% djece u obitelji, a djevojčice 15%), pa je odgovarajući kružni dijagram idući:



Slika 1: Primjer kružnog dijagrama

Kružni dijagram koristan je jedino u slučajevima kada raspolažemo sa malim brojem kategorija, npr. 4 ili 5. Što je broj kategorija veći, dijagram je nepregledniji i na osnovu njega ne možemo puno toga zaključiti. U potonjem slučaju bolje je koristiti **stupčasti dijagram** (eng. *bar plot*), u kojem svaki stupac predstavlja jednu kategoriju, pri čemu je visina frekvencija ili relativna frekvencija kategorije, a širine stupaca su jednake. Stupčasti dijagram se koristi kao grafički prikaz neovisno o broju kategorija, pa na idućoj slici

možemo vidjeti jedan za prethodne podatke o djeci u obitelji (M - muško dijete, Ž - žensko dijete).



Slika 2: Primjer stupčastog dijagrama

2.1.2 Histogram i QQ graf

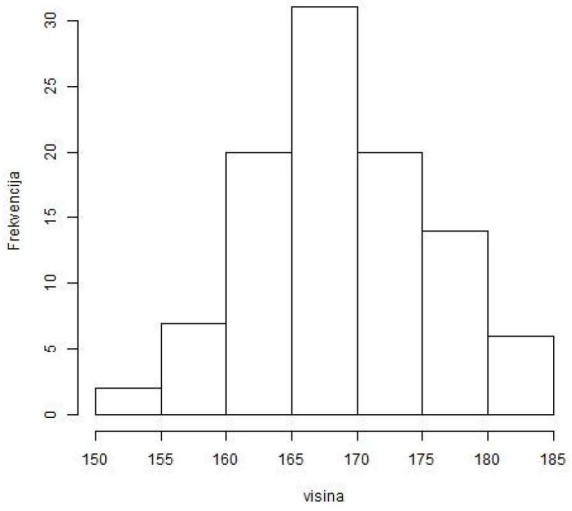
Kada skup podataka koji promatramo dolazi iz neprekidne distribucije, npr. prosjek ocjena, tada skup svih podataka grupiramo u razrede (najčešće disjunktne intervale). Za grafički prikaz ovako kategoriziranih podataka koristimo poseban stupčasti dijagram kojeg nazivamo **histogram**, pri čemu je bitno da broj kategorija bude dobro odabran kako ne bismo imali previše ili premalo stupaca, pa samim time i nepregledan graf. I u ovom slučaju visine stupaca predstavljaju (relativnu) frekvenciju podataka iz intervala, ali sada širine stupaca nisu međusobno jednake već predstavljaju duljinu odgovarajućih intervala.

Primjer 2.1. *Neka je dana tablica u kojoj su zabilježene visine 100 učenika iz jedne srednje škole.*

visine učenika u cm							
176.76	171.78	180.63	167.27	173.40	171.78	176.55	170.03
182.75	169.68	184.96	169.10	161.00	165.82	172.85	159.90
170.63	155.25	164.70	168.55	169.85	169.39	178.32	164.12
174.91	153.15	168.66	182.67	165.45	171.41	165.45	173.88
167.75	169.70	169.77	182.18	163.77	156.92	168.68	161.90
160.84	166.28	168.89	178.78	171.07	167.46	175.29	176.56
167.20	166.12	175.04	173.74	162.66	156.67	169.98	164.69
164.01	165.68	167.66	160.84	166.87	171.05	177.72	164.54
157.96	168.86	168.54	179.30	173.89	160.90	162.10	173.12
177.92	176.46	162.07	152.15	166.46	161.48	164.90	165.20
174.64	181.06	178.84	177.43	168.70	168.12	177.38	171.08
169.43	163.19	174.99	160.40	166.65	158.77	157.50	171.87
174.70	170.41	163.05	160.95				

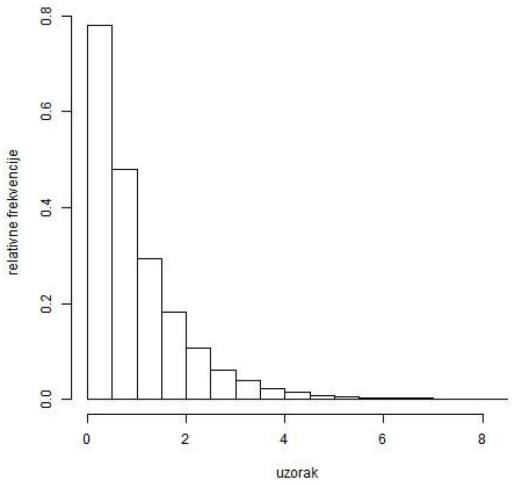
Tablica 1: Visine učenika u cm

Iz ovako dane tablice ne možemo ništa zaključiti o populaciji osim pojednih visina u cm, no iz histograma vidimo da je najveći broj osoba visine $165-170cm$*, najviše osobe su visoke* $180-185cm$*, dok najnižih osoba (*$150-155cm$*) ima najmanje. Također pomoću histograma možemo naslutiti da ovaj uzorak dolazi iz normalne distribucije jer oblik histograma prati oblik grafa funkcije gustoće normalne distribucije.*



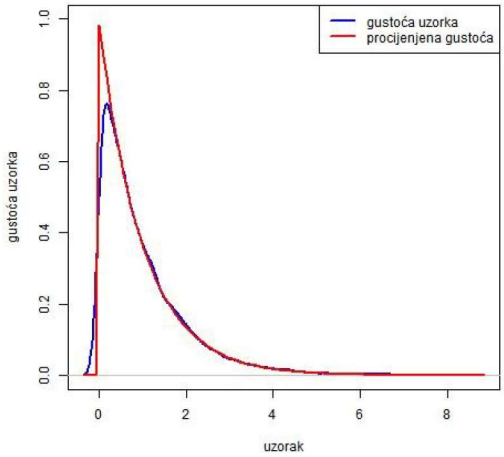
Slika 3: Primjer histograma

Primjer 2.2. *Uzmimo slučajan uzorak od 1000 brojeva*[1] *iz eksponencijalne distribucije s parametrom* $\lambda = 1$ *i pretpostavimo da ne znamo iz koje distribucije dolazi uzorak. Želimo li na osnovu danog niza podataka procijeniti o kojoj se distribuciji radi, možemo pogledati histogram tih podataka kako bismo dobili vizualan uvid, tj. kako bismo mogli naslutiti distribuciju uzorka.*



Slika 4: Histogram relativnih frekvencija uzorka

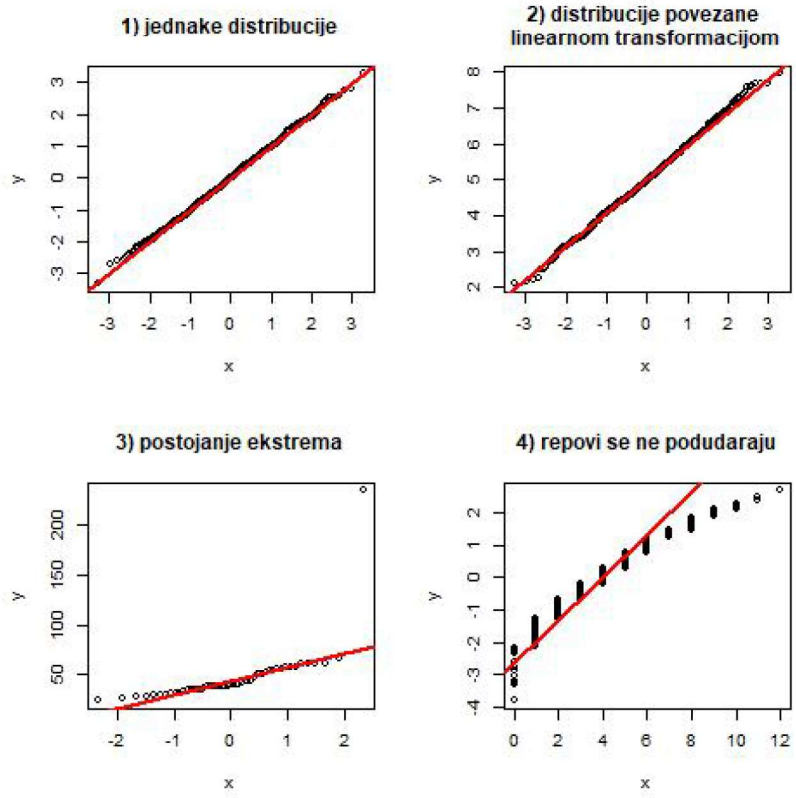
Iz grafa možemo isčitati grubu procjenu funkcije gustoće i pretpostaviti da se radi o eksponencijalnoj distribuciji. Iduća slika koja prikazuje grafove teorijske funkcije gustoće i funkcije gustoće uzorka potvrđuje pretpostavku.



Slika 5: Grafovi funkcija gustoće

[1]podaci simulirani u programskom paketu R

Osim histograma, za usporedbu distribucije uzorka sa nekom pretpostavljenom najčešće se koristi **QQ graf** (eng.*quantile-quantile plot*). Na jednoj osi grafa (x koordinata) nalaze se kvantili teorijske razdiobe, na drugoj kvantili stvarne razdiobe (y koordinata). U slučaju da je pretpostavljena distribucija jednaka stvarnoj, uređeni parovi trebali bi ležati, barem približno, na pravcu $y = x$. Ako točke leže na nekom pravcu $y = ax + b, a, b \in \mathbb{R}$, znači da je jedna distribucija linearna transformacija druge, a ako ne leže na istom pravcu, može biti nekoliko različitih uzroka. Npr. ako nekoliko točaka ne leži na pravcu, a ostale leže, znači da postoje ekstremi. Ako se točke na krajevima nalaze ispod/iznad pravca, tada repovi distribucija nisu jednaki. Zakrivljenost grafa upućuje na asimetričnost pretpostavljene distribucije u odnosu na teorijsku. U slučaju da je graf stepenast, distribucija uzorka je diskretna. Na idućim grafovima možemo vidjeti neke od tih slučajeva:



Slika 6: Nekoliko primjera QQ grafova

Histogram nam, kao što smo vidjeli, pruža uvid u distribuciju podataka, no često te iste podatke želimo opisati pomoću različitih numeričkih karakteristika pa ćemo u idućem dijelu promotriti parametre lokacije, kao i parametre raspršenosti niza podataka uz još neke grafičke prikaze istih.

2.2 Parametri lokacije

2.2.1 Aritmetička sredina

Jedan od najvažnijih lokacijskih parametara je **aritmetička sredina**, definirana s:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2}$$

Propozicija 2.1. *(Važna svojstva aritmetičke sredine)*

i) $\displaystyle\sum_{i=1}^{n}(x_i - \overline{x}) = 0$

ii) $\displaystyle\sum_{i=1}^{n}(x_i - c)^2 \geq \sum_{i=1}^{n}(x_i - \overline{x})^2, \quad c \in \mathbb{R}$

Dokaz. Dokažimo prvo tvrdnju *i)*:

$$\sum_{i=1}^{n}(x_i - \overline{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n}\overline{x} = n\overline{x} - n\overline{x} = 0.$$

Za dokaz tvrdnje *ii)* pretpostavljamo da je $c \in \mathbb{R}$. Ako raspišemo lijevu stranu nejednakosti, imamo:

$$\begin{aligned}\sum_{i=1}^{n}(x_i - c)^2 &= \sum_{i=1}^{n}\left[(x_i - \overline{x}) + (\overline{x} - c)\right]^2 \\ &= \sum_{i=1}^{n}\left[(x_i - \overline{x})^2 + 2(x_i - \overline{x})(\overline{x} - c) + (\overline{x} - c)^2\right] \\ &= \sum_{i=1}^{n}(x_i - \overline{x})^2 + 2(\overline{x} - c)\sum_{i=1}^{n}(x_i - \overline{x}) + \sum_{i=1}^{n}(\overline{x} - c)^2\end{aligned}$$

Nadalje, zbog svojstva i), imamo:

$$\sum_{i=1}^{n}(x_i - c)^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2 + n(\overline{x} - c)^2 \geq \sum_{i=1}^{n}(x_i - \overline{x})^2,$$

iz čega slijedi tvrdnja. □

Svojstvo i) prethodne propozicije kaže da aritmetička sredina ima svojstvo da je suma odstupanja od $\overline{x}$ jednaka 0, a svojstvo ii) da je suma kvadratnog odstupanja od aritmetičke sredine najmanja.

Primjer 2.3. *Aritmetička sredina niza podataka iz primjera 2.1 iznosi* $\overline{x} = 168.9143$. *Na histogramu tih podataka vidimo da se* $\overline{x}$ *nalazi unutar središnjeg intervala, pa je ona zato mjera centralne tendencije.*

Aritmetička sredina koristan je predstavnik danog niza podataka jer pokazuje njihovu lokaciju na osi apscisa, no nekada može dati krivu informaciju zbog tzv. **stršećih vrijednosti**, vrijednosti koji su znatno manje ili znatno veće u odnosu na ostatak podataka.

Primjer 2.4. *Provedeno je istraživanje u jednoj zgradi vezano uz mase pojedinih članova obitelji. Jedan uzorak daje nam sljedeće informacije:*

	otac	majka	sin	kćer
prva obitelj	131	63	44	58
druga obitelj	71	63	50	59
treća obitelj	79	68	15	65
četvrta obitelj	87	62	97	32
peta obitelj	76	67	32	45
šesta obitelj	111	63	32	51
sedma obitelj	85	55	41	49
osma obitelj	89	57	45	18
deveta obitelj	99	59	39	44
deseta obitelj	84	66	37	88

Tablica 2: Mase pojedinih članova obitelji u kg

Pretpostavimo da nas zanima prosječna težina oca po obitelji. Ona iznosi 91.2kg. S obzirom na stršeće vrijednosti (u ovome nizu podataka najviše iz niza odskaču vrijednosti 131 i 111), ponekada se koristi ***skraćena srednja vrijednost****, kod koje se iz niza podataka izbaci određeni postotak najvećih i*

najmanjih vrijednosti kako bi se dobila bolja procjena. Kada bismo iz ovoga niza podataka izbacili 20% gornjih i donjih vrijednosti, za prosječnu masu oca bismo dobili 87.17kg.

U ovom primjeru cijelu populaciju čine sve četveročlane obitelji u zgradi, a uzorak je 10 četveročlanih obitelji. Dobiveni uzorak masa očeva smatramo realizacijom slučajnog vektora $X = (X_1, \ldots, X_{10})$. *Obitelji su slučajno birane, pa možemo pretpostaviti da su slučajne varijable* $X_1, \ldots, X_{10}$ *nezavisne i jednako distribuirane (svaka masa ovisi o pojedinoj osobi, nijedan podatak ne utječe na drugi), pa se radi o j.s.u.*

S obzirom da je masa normalno distribuirano obilježje, svaka slučajna varijabla X_i, $i = 1, \ldots, n$ *ima* $\mathcal{N}(\mu, \sigma^2)$ *distribuciju, pri čemu su očekivanje* μ *i varijanca* σ^2 *nepoznati parametri. Nepoznati parametar je dvodimenzionalan, pa imamo parametarski statistički model:*

$$\mathcal{P} = \left\{ F_{(\mu,\sigma^2)} : (\mu, \sigma^2) \in \mathbb{R} \times \langle 0, \infty \rangle \right\}$$

pri čemu je

$$F_{(\mu,\sigma^2)}(t_1, \ldots, t_n) = \prod_{i=1}^{n} F_{(\mu,\sigma^2)}(t_i),$$

a $F_{(\mu,\sigma^2)}$ *funkcija distribucije iz* $\mathcal{N}(\mu, \sigma^2)$ *distribucije. Kada ne bismo imali pretpostavku o normalnoj distribuiranosti obilježja, tada bi imali neparametarski model jer je distribucija iz koje dolazi uzorak potpuno nepoznata.*

Aritmetička sredina slučajne varijable njezina je očekivana vrijednost, odnosno očekivanje μ *slučajne varijable, pa je u ovom primjeru jedna procjena očekivane mase oca u cijeloj populaciji na osnovu uzorka veličine 10 jednaka 91.2kg.*

Bilo bi korisno na primjeru ovih podataka spomenuti stabljika-list ili stablo-list dijagram (eng. *stem-and-leaf plot*), koji služi za prikaz numeričkih varijabli i prikazuje oblik distribucije podataka, te iz njega možemo isčitati najmanju i najveću vrijednost iz niza podataka. Kako sa pravi takav dijagram?

Primjer 2.5. *Prikupljene podatke poredamo u rastućem poretku i kategoriziramo tako da one podatke koji one koji su manji od 100 grupiramo po prvoj znamenci, a one koji su veći od 100 po prve dvije: 1, 3, 4, 5, 6, 7, 8, 9, 11,*

13. Zatim za listove kod odgovarajuće znamenke prepišemo ostatak podatka. Za tablicu podataka iz prethodnog primjera stabljika-list dijagram je idući:

1	|	58
3	|	22279
4	|	144559
5	|	0157899
6	|	23335678
7	|	169
8	|	45789
9	|	79
11	|	1
13	|	1

Npr., ako uzmemo 7 | 169*, znači da se u niz nalaze brojevi 71, 76, 79. Prema tome, vidimo da je najmanja masa 15kg, najveća 131kg, da većina ljudi ima masu manju od 70kg, da je frekvencija broja 63 jednaka 3, itd.*

2.2.2 Medijan

Medijan, u oznaci m, se definira kao središnja vrijednost niza podataka poredanih u rastućem poretku. U slučaju kada je ukupan broj vrijednosti neparan, medijan je točno ona vrijednost u sredini niza, a ako je ukupan broj vrijednosti paran, za medijan se uzima aritmetička sredina dvaju središnjih vrijednosti. U oba slučaja je 50% podataka manje ili jednako i 50% podataka veće ili jednako od medijana:

$$m(x_i, \ldots, x_n) = \begin{cases} x_{\frac{n+1}{2}}, & \text{n neparan} \\ \dfrac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}, & \text{n paran} \end{cases} \tag{3}$$

Također možemo promatrati i medijan apsolutnih odstupanja slučajne varijable od medijana podataka, odnosno ako niz podataka $x_1, \ldots, x_n$ ima medijan m, onda se za medijan apsolutnih odstupanja uzima medijan brojeva $|x_1 - m|, \ldots, |x_n - m|$.

Postavlja se pitanje: kada je korisnije koristiti aritmetičku sredinu, a kada medijan kao parametar centra? Odabir parametra centra ovisi o tome

kako su podaci za koje ih računamo distribuirani.

Ako podaci dolaze iz distribucije koja je simetrična, možemo koristiti i medijan i aritmetičku sredinu za mjeru centra jer će vrijednosti biti približno jednake. Na primjer, medijan podataka iz primjera 2.1 je $m = 168.78$, što je gotovo jednako kao i $\overline{x} = 168.9143$. No, u ovakvim slučajevima pretežito se koristi aritmetička sredina jer ju možemo direktno izračunati bez da prvo sortiramo niz.

U slučaju kada u uzorku koji promatramo postoje stršeće vrijednosti ili je raspodjela podataka asimetrična, bolje je koristiti medijan jer po samoj definiciji možemo vidjeti da na njega ne utječu svi podaci, nego samo srednje vrijednosti iz niza, za razliku od aritmetičke sredine na koju te stršeće vrijednosti uvelike utječu. Pogledajmo primjer.

Primjer 2.6. *Prosječna neto plaća po osobi u rujnu 2019. godine u Hrvatskoj iznosila je 6 148 kuna, dok je medijalna neto plaća za isti mjesec iznosila 5 539 kuna. Najviša prosječna neto plaća bila je 11 448kn, a najniža 4 233kn. Ovdje možemo primijetiti koliko neke visoke, odnosno niske vrijednosti u nizu podataka mogu utjecati na razliku između prosjeka i medijana. Medijan je u ovom slučaju korisniji kao mjera centra nego aritmetička sredina jer daje realniju sliku o plaćama: 50% stanovništva imalo je neto plaću manju od 5 539kn, dok je najviša prosječna neto plaća bila otprilike 2.0668 puta veća.* [2]

2.3 Parametri raspršenosti

Parametri, odnosno mjere lokacije nam daju informacije o srednjoj vrijednosti niza podataka, no to nije dovoljno da se o podacima da potpuna slika.

Primjer 2.7. *Neka su dana dva niza podataka:*

$$x_1 = 1, x_2 = 2, x_3 = 3, x_4 = 4, x_5 = 5$$

i

$$y_1 = y_2 = y_3 = y_4 = y_5 = 3.$$

Oba niza imaju jednake prosjeke i medijane ($\overline{x} = \overline{y} = m_x = m_y = 3$), no vidimo da nizovi nisu jednaki. Drugi niz je stacionaran (sve vrijednosti su u

[2] podaci preuzeti sa stranice Državnog zavoda za statistiku Republike Hrvatske

istoj točki), a u prvom nizu je raspon podataka od 1 do 5. Stoga, kako bismo u potpunosti opisali dane podatke pomoću nekoliko parametara, osim mjera centra, također koristimo i mjere raspršenosti.

Pogledajmo najčešće korištene mjere raspršenja - mjere koje nam daju uvid u raspon danog niza podataka, odnosno koliko se u prosjeku podaci razlikuju od srednje vrijednosti niza.

Prvo definiramo najjednostavniju mjeru: **raspon** niza podataka $x_1, \ldots, x_n$, u oznaci d, kao razliku

$$d = x_{min} - x_{max}, \tag{4}$$

gdje je x_{min} najmanja (**minimum**), a x_{max} najveća vrijednost u danom nizu (**maksimum**). No, možemo vidjeti da je raspon osjetljiv na ekstremne vrijednosti. Na primjer, raspon masa očeva u primjeru 2.4 je $131 - 71 = 60$, ali kada bismo izbacili najmanju i najveću vrijednost, dobili bismo $111 - 76 = 35$. Zbog ove osjetljivosti raspona, bolje je koristiti **interkvartilni raspon** koji se definira kao razlika između gornjeg i donjeg kvartila:

$$d' = q_{0.75} - q_{0.25} \tag{5}$$

Kako bismo objasnili što je gornji (donji) kvartil, prvo ćemo definirati kvantile.

2.3.1 Kvantili

Neka je X slučajna varijabla. Kvantil ili postotna vrijednost q_p za neki $p \in \langle 0, 1\rangle$ je realan broj takav da je slučajna varijabla X manja od tog broja s vjerojatnošću barem p i veća od njega s vjerojatnošću barem $1 - p$:

$$P(X \leq q_p) \geq p \quad \text{i} \quad P(X \geq q_p) \geq 1 - p.$$

Najčešće korišteni p% kvantili:

1) medijan (50% kvantil, 0.5-kvantil)

2) **donji kvartil** (0.25-kvantil) - barem 25% podataka je manje ili jednako i barem 75% podataka je veće ili jednako $q_{0.25}$

3) **gornji kvartil** (0.75-kvantil) - barem 75% podataka je manje ili jednako i barem 25% podataka je veće ili jednako $q_{0.75}$

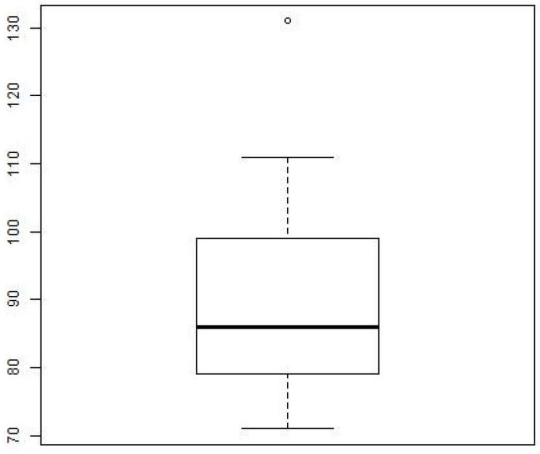
Primjer 2.8. *Minimum, donji kvartil, medijan, gornji kvartil i maksimum čine tzv.* ***karakterističnu petorku****. Ako ponovno pogledamo primjer 2.4, odnosno dobiveni uzorak masa očeva, karakteristična petorka je:*

$$\begin{aligned} x_{min} &= 71, \\ q_{0.25} &= 79, \\ m &= 86, \\ q_{0.75} &= 99, \\ x_max &= 131, \end{aligned}$$

a interkvartilni raspon $d' = 99 - 79 = 20$.

Podatke iz prethodnog primjera lakše je povezati sa samim uzorkom ako pogledamo grafički prikaz: **kutijasti dijagram** ili brkatu kutiju (eng. *box and whisker plot*). Osim što prikazuje karakterističnu petorku, na njemu vidimo raspršenje (raspon i interkvartilni raspon), asimetričnost podataka te eventualne stršeće vrijednosti. Na samom dijagramu horizontalne linije predstavljaju, redom, od dolje prema gore: minimum, donji kvartil, medijan, gornji kvartil i maksimum. Minimum i maksimum nalaze se na udaljenosti 1.5 interkvartilnog raspona, a ekstremne vrijednosti prikazane su točkama ili zvjezdicama.

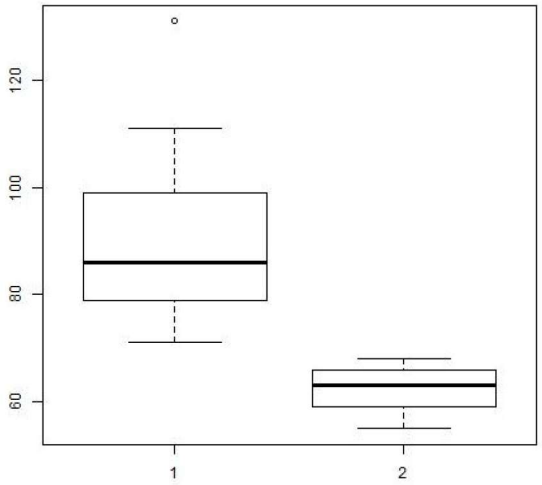
Primjer 2.9. *Pogledajmo kutijasti dijagram svih masa iz primjera 2.4 o masama pojedinih članova obitelji.*



Slika 7: Primjer kutijastog dijagrama

Na dijagramu vidimo da postoji jedna ekstremna vrijednost te da je distribucija iz koje dolazi uzorak asimetrična (u odnosu na medijan). Kutijasti

dijagrami posebno su korisni kada želimo usporediti dva skupa podataka. Usporedimo kutijaste dijagrame mase očeva (lijevo) i mase majki (desno).



Slika 8: Kutijasti dijagrami masa očeva i masa majki

Ovakvim vizualnim prikazom možemo, recimo, o populaciji zaključiti da je najveća masa među ženskim roditeljima manja od najmanje mase među muškim roditeljima, odnosno da je općenito masa majki manja nego masa očeva.

Kao što smo vidjeli na prethodnim primjerima, interkvartilni raspon vrlo je koristan prikaz mjere raspršenosti podataka s obzirom da nije osjetljiv na ekstremne vrijednosti. No, kao i kod medijana, da bismo ga mogli izračunati, prvo podatke moramo poredati u rastućem poretku, što za velike uzorke može oduzeti puno vremena. Zato ćemo definirati još neke mjere za opisivanje raspršenosti koje su lakše za izračunati: varijanca i standardna devijacija.

2.3.2 Varijanca i standardna devijacija

Varijanca slučajne varijable X definira se kao srednje kvadratno odstupanje niza podataka od njihove aritmetičke sredine:

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \tag{6}$$

Standardna devijacija kvadratni je korijen varijance:

$$s_n = \sqrt{s_n^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{7}$$

U idućem izvodu možemo vidjeti koje je značenje standardne devijacije za neki niz podataka. Iz definicije varijance (6) kada jednakost pomnožimo s n, za proizvoljan $k > 0$ slijedi:

$$\begin{aligned} ns_n^2 &= \sum_{i=1}^{n}(x_i - \overline{x})^2 \\ &= \sum_{x_i<\overline{x}-ks_n}(x_i-\overline{x})^2 + \sum_{\overline{x}-ks_n\le x_i\le\overline{x}+ks_n}(x_i-\overline{x})^2 + \sum_{x_i>\overline{x}+ks_n}(x_i-\overline{x})^2 \\ &\ge \sum_{\overline{x}-ks_n\le x_i\le\overline{x}+ks_n}(x_i-\overline{x})^2 + \sum_{x_i>\overline{x}+ks_n}(x_i-\overline{x})^2 = \sum_{|x_i-\overline{x}|>ks_n}(x_i-\overline{x})^2 \end{aligned}$$

Sada kvadriranjem nejednakosti $|x_i - \overline{x}| \ge ks_n$ slijedi da je $(x_i - \overline{x})^2 \ge k^2 s_n^2$, pa uvrštavanjem u gornji niz imamo:

$$\sum_{|x_i-\overline{x}|\ge ks_n}(x_i-\overline{x})^2 \ge \sum_{|x_i-\overline{x}|\ge ks_n} k^2 s_n^2.$$

Zatim ako još dodatno pretpostavimo da ima $l \le n$ podataka u nizu za koje je $|x_i - \overline{x}| > ks_n$ te da je $s_n > 0$, dobijemo:

$$ns_n^2 \ge lk^2 s_n^2 \quad \Rightarrow \quad l \le \frac{1}{k^2}n.$$

Možemo promotriti neke posebne slučajeve:

k=2 $\Rightarrow l \le \frac{1}{4}n \Rightarrow$ oko 25% podataka se nalazi izvan, odnosno barem 75% podataka se nalazi unutar intervala $[\overline{x} - 2s_n, \overline{x} + 2s_n]$

k=3 $\Rightarrow l \le \frac{1}{9}n \Rightarrow$ oko 11% podataka se nalazi izvan, odnosno barem 89% podataka se nalazi unutar intervala $[\overline{x} - 3s_n, \overline{x} + 3s_n]$

Primijetimo da prethodni izvod odgovara Čebiševljevoj nejednakosti: Ako je X slučajna varijabla s konačnim očekivanjem μ i konačnom varijancom σ^2, tada za $k > 0$ vrijedi

$$P(|X - \mu| \ge k\sigma) \le \frac{1}{k^2}$$

iz čega slijedi

$$P(|X - \mu| < k\sigma) = 1 - P(|X - \mu| \ge k\sigma) \ge 1 - \frac{1}{k^2},$$

odnosno da je vjerojatnost da slučajna varijabla odstupa od svog očekivanja za manje od k standardnih devijacija veća ili jednaka $1 - \frac{1}{k^2}$.

Zbog Čebiševljeve nejednakosti vrijedi slabi zakon velikih brojeva (teorem 1.1) koji kaže da vjerojatnost da se realizacija prosjeka slučajnih varijabli $X_1, \ldots, X_n$ razlikuje od svog očekivanja za više od po volji odabranog malog broja možemo učiniti proizvoljno malom s povećanjem broja slučajnih varijabli u izračunu:

$$\lim_{n\to\infty} P\Big(\big|\overline{X}_n - \mu\big| > \varepsilon\Big) \leq \frac{Var(\overline{X_n})}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \to 0 \text{ kada } n \to \infty.$$

Primjer 2.10. *Pogledajmo ponovno primjer 1.1 i dodatno pretpostavimo da je vjerojatnost da odabrani kišobran bude ispravan jednaka vjerojatnosti da bude neispravan. Biramo kišobran 10,000 puta i ispravan kišobran izbrojimo 5067 puta. U ovom slučaju, svaka slučajna varijabla X_i ima Bernoullijevu distribuciju:*

$$\begin{pmatrix} 0 & 1 \\ 1/2 & 1/2 \end{pmatrix}.$$

Pokus ponavljamo 10,000 puta i za $E(X_i) = 1/2, Var(X_i) = 1/4$ dobijemo:

$$Var(\overline{X}_{10,000}) = \frac{\frac{1}{4}}{10,000} = 2.5 \times 10^{-5},$$

pa je standardna devijacija $\sigma = \sqrt{2.5 \times 10^{-5}} = 0.005$, što znači da je udio ispravnih kišobrana koji smo mi izbrojali, 0.5067, za malo više od jedne standardne devijacije veći od očekivane vrijednosti 0.5. Ovo se poklapa sa Čebiševljevom nejednakosti: za $k = 2$ dobivamo da je:

$$P\left(\left|\overline{X}_n - \frac{1}{2}\right| > 0.01\right) \leq \frac{1}{4},$$

odnosno vrijedi da je $\overline{x}_{10,000} = 0.5067 \in [0.49, 0.51]$.

3 Procjena parametara

U potpoglavlju 1.2 definirali smo parametarski statistički model u kojem distribucija F slučajnog vektora X, odnosno slučajnog uzorka $(X_1, \ldots, X_n)$ ovisi o nekom k-dimenzionalnom parametru θ. U prethodnom poglavlju (2) naveli naveli smo o kojim sve parametrima lokacije te raspršenja F može ovisiti. U ovom poglavlju ćemo definirati procjenitelja i procjenu tih parametara.

Prilikom određivanja vrijednosti parametra, nikada to nećemo sa 100%-tnom vjerojatnošću moći napraviti, jedino ga možemo aproksimirati iz realizacije $(x_1, \ldots, x_n)$ jer svakim ponovljenim istraživanjem dobivamo novu, drugačiju procjenu. Definirajmo sada neke osnovne pojmove.

Definicija 3.1. *Neka je $(X_1, \ldots, X_n)$ slučajan uzorak iz parametarskog statističkog modela.* ***Procjenitelj*** *nepoznatog parametra θ je slučajna varijabla $T = t(X_1, \ldots, X_n)$, pri čemu $t : \mathbb{R}^n \to \Theta$. Realizaciju procjenitelja nazivamo* ***procjena****.*

Važno je naglasiti da je procjenitelj **slučajna varijabla**, a procjena **broj**, odnosno da je procjena određenog parametra također realizacija slučajne varijable, kao i uzorak iz kojeg računamo tu procjenu. S obzirom na to da se parametar procjenjuje jednom vrijednošću, ovakvu procjenu još nazivamo točkastom.

Definicija 3.2. *Neka je $(X_1, \ldots, X_n)$ slučajan uzorak.* ***Statistika*** *je slučajna varijabla*

$$T = t(X_1, \ldots, X_n), \tag{8}$$

Drugim riječima, statistika je svaka funkcija koja ovisi o slučajnom uzorku $(X_1, \ldots, X_n)$, a ne ovisi eksplicitno o nepoznatom parametru θ.

Definicija 3.3. *Neka je $\mathcal{P} = \{F_\theta : \theta \in \Theta \subseteq \mathbb{R}^k\}$ parametarski statistički model, a θ k-dimenzionalan parametar. Kažemo da je procijenitelj T* ***nepristran*** *ukoliko vrijedi*

$$E_\theta[T] = \theta, \quad \forall \theta \in \Theta,$$

pri čemu je θ stvarna vrijednost parametra distribucije F_θ.

Definicija 3.4. *Neka je $\mathcal{P} = \{F_\theta : \theta \in \Theta \subseteq \mathbb{R}^k\}$ parametarski statistički model i θ jednodimenzionalan parametar. Kažemo da je procjenitelj T* ***konzistentan*** *ako za pripadni niz procjenitelja $(T_n, n \in \mathbb{N})$, $\forall \varepsilon > 0$ vrijedi:*

$$\lim_{n \to \infty} P_\theta\big(|T_n - \theta| > \varepsilon \big) = 0, \quad \forall \theta \in \Theta.$$

Ako međusobno uspoređujemo dva procjenitelja, s obzirom da procjenitelj slučajna varijabla, bolji je, odnosno efikasniji, onaj procjenitelj koji ima manju varijancu:

Definicija 3.5. *Neka je $(X_1, \ldots, X_n)$ slučajan uzorak iz parametarskog statističkog modela te $T_1 = t_1(X_1, \ldots, X_n)$ i $T_2 = t_2(X_1, \ldots, X_n)$ dva nepristrana procjenitelja nepoznatog jednodimenzionalnog parametra θ. Kažemo da je procjenitelj T_1* ***bolji*** *ili* ***efikasniji*** *od T_2 ako je $Var(T_1) < Var(T_2)$.*

3.1 Točkasta procjena očekivanja

Vrlo često u praktičnim situacijama trebamo procijeniti srednju vrijednost neke veličine na osnovu n mjerenja. Na primjer, mjerimo masu nekog predmeta n puta i dobijemo n različitih mjerenja, tj. statistički niz $x_1, \ldots, x_n$. Stvarnu težinu (označimo ju s μ) možemo shvatiti kao matematičko očekivanje $E(X)$ slučajne varijable X, pa problem utvrđivanja stvarne mase predmeta svodimo na problem procjene parametra μ slučajne varijable X na temelju n nezavisnih mjerenja. Prema tome, intuitivno procjenitelja očekivanja μ definiramo na sljedeći način:

Definicija 3.6. *Neka je $(X_1, \ldots, X_n)$ j.s.u.*[3] *iz distribucije slučajne varijable X pri čemu je nepoznati parametar očekivanje od X, $EX = EX_i = \mu \in \mathbb{R}$, $i = 1, \ldots, n$. Statistika*

$$\overline{X}_n = t_n(X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{9}$$

je procjenitelj za očekivanje slučajne varijable X i nazivamo ju ***uzoračka aritmetička sredina****.*
Za realizaciju $(x_1, \ldots, x_n)$ aritmetička sredina

$$\overline{x}_n = t_n(x_1, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{10}$$

je procjena očekivanja slučajne varijable X.

[3] jednostavan slučajan uzorak (1.1, str. 1)

Pretpostavimo sada da je $VarX < \infty$, odnosno da je $Var(X_i) = \sigma^2$, $\forall i = 1, \ldots, n$ i izračunajmo očekivanje i varijancu uzoračke aritmetičke sredine. Prema linearnosti očekivanja vrijedi:

$$E\big[\overline{X}_n\big] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n}\cdot n\mu = \mu,$$

Iz ovog rezultata vidimo da je $\overline{X}_n$ nepristran procjenitelj jer je

$$E_\mu\big[\overline{X}_n\big] = \mu, \quad \forall \mu \in \mathbb{R}.$$

Prisjetimo se da za varijancu slučajne varijable X vrijedi:

$$Var(aX + b) = a^2 Var(X), \quad a, b \in \mathbb{R}.$$

Prema tome,

$$\begin{aligned} Var\big(\overline{X}_n\big) &= Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{1}{n^2}\cdot n\sigma^2 = \frac{\sigma^2}{n}. \end{aligned}$$

Pokažimo još da je $\overline{X}_n$ konzistentan procjenitelj za očekivanje. Prema definiciji, to će vrijediti ako

$$(\forall \mu \in \mathbb{R})(\forall \varepsilon > 0) \lim_{n\to\infty} P_\mu\big(\,\big|\overline{X}_n - \mu\big|\,\big) = 0.$$

To direktno slijedi iz slabog zakona velikih brojeva (teorem 1.1, str. 3).

Definicija 3.7. *Kažemo da je niz procjenitelja* $(T_n, n \in \mathbb{N})$ *parametra* θ ***asimptotski normalan procjenitelj*** *ako vrijedi*

$$\frac{T_n - E_\theta[T_n]}{\sqrt{Var(T_n)}} \xrightarrow{D} Z \sim \mathcal{N}(0,1).$$

Iz ove definicije i CGT-a (teorem 1.2, str.4) slijedi da je uzoračka aritmetička sredina asimptotski normalan procjenitelj očekivanja slučajne varijable X.

Primjer 3.1. *Pretpostavimo da želimo izračunati očekivanu ocjenu studenata nekog fakulteta na jednom kolokviju. Uzmemo li od svih studenata uzorak od 80 ljudi, ocjenu modeliramo neprekidnom slučajnom varijablom X koja ima funkciju distribucije F_μ, gdje je očekivanje μ nepoznati parametar koji želimo procijeniti. U idućoj tablici zadane su frekvencije ocjena:*

ocjena	1	2	3	4	5
frekvencija	21	25	15	12	7

Tablica 3: Tablica frekvencija ocjena

Uređena 80-orka frekvencija ocjena predstavlja jednu realizaciju jednostavnog slučajnog uzorka $(X_1, \ldots, X_{80})$ iz distribucije F_μ, pri čemu je procjena očekivanja slučajne varijable X jednaka:

$$\overline{x}_{80} = \frac{1 \cdot 21 + 2 \cdot 25 + 3 \cdot 15 + 4 \cdot 12 + 5 \cdot 7}{80} \approx 2.49.$$

3.2 Točkasta procjena varijance

Definicija 3.8. *Neka je $(X_1, \ldots, X_n)$ j.s.u. iz distribucije slučajne varijable X, pri čemu je $\sigma^2 > 0$ nepoznati parametar te $\overline{X}_n = \sum_{i=1}^{n} X_i$. Statistika*

$$\overline{S}_n^2 = \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \overline{X}_n\right)^2 \tag{11}$$

procjenitelj je varijance slučajne varijable X i nazivamo ju ***uzoračka varijanca****.*

Problem sa ovako definiranom statistikom jest što ona nije nepristrana. Pokažimo to.

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

Ako sada označimo $E[X] = \mu$, $\overline{S}_n^2$ možemo drugačije zapisati:

$$\begin{aligned}
\overline{S}_n^2 &= \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \mu - \left(\overline{X}_n - \mu\right)\right)^2 \\
&= \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \mu\right)^2 - \frac{2}{n} \sum_{i=1}^{n} \left(X_i - \mu\right)\left(\overline{X}_n - \mu\right) + \frac{1}{n} \sum_{i=1}^{n} \left(\overline{X}_n - \mu\right)^2
\end{aligned}$$

Kako je

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

i

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

slijedi da je

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

Sada za očekivanje vrijedi:

$$E\left[\overline{S}_n^2\right]=E\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\mu\right)^2-\left(\overline{X}_n-\mu\right)^2\right]$$
$$=\frac{1}{n}\sum_{i=1}^{n}E\left[\left(X_i-\mu\right)^2\right]-E\left[\left(\overline{X}_n-\mu\right)^2\right]$$
$$=\frac{1}{n}\sum_{i=1}^{n}VarX_i-Var\overline{X}_n$$
$$=\frac{1}{n}\cdot n\cdot\sigma^2-\frac{\sigma^2}{n}$$
$$=\frac{n-1}{n}\sigma^2\neq\sigma^2$$

Pa prema tome, uzoračka varijanca nije nepristran procjenitelj, no množenjem sa $\frac{n}{n-1}$, dobivamo nepristrani procjenitelj

$$S_n^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}_n\right)^2 \tag{12}$$

kojeg nazivamo **korigirana uzoračka varijanca**. Provjerimo nepristranost.

$$\begin{aligned} E_{\sigma^2}\left[S_n^2\right] &= E_{\sigma^2}\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2\right] \\ &= E_{\sigma^2}\left[\frac{n}{n-1}\overline{S}_n^2\right] = \frac{n}{n-1}E_{\sigma^2}\left[\overline{S}_n^2\right] \\ &= \frac{n}{n-1}\cdot\frac{n-1}{n}\cdot\sigma^2 = \sigma^2 \end{aligned}$$

Nepristranu procjenu varijance računamo s:

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}_n\right)^2,$$

a procjenu standardne devijacije, s_n, kao kvadratni korijen istog izraza. Valjalo bi napomenuti da ovako definirana procjena standardne devijacije nije nepristrana i gotovo je nemoguće napraviti takvu koja će vrijediti za bilo koju distribuciju.

Primjer 3.2. *Pogledajmo ponovno bazu podataka iz primjera 3.1 iz prethodnog podnaslova u kojem smo izračunali procjenu očekivanja,* $\overline{x}_{80} \approx 2.49$. *S obzirom da u tablici imamo grupirane podatke (dane su frekvencije ocjena), procjenu korigirane uzoračke varijance računamo po formuli:*

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^{k} n_i(x_i - \overline{x}_n)^2,$$

gdje je k *broj grupa* ($n = n_1 + n_2 + \ldots + n_k$), *a* n_i *odgovarajuća frekvencija podatka* x_i. *Prema tome, imamo:*

$$\begin{aligned} s_{80}^2 &= \frac{1}{80-1}\sum_{i=1}^{5} n_i(x_i - 2.49)^2 \\ &= \frac{1}{79}\big[21\cdot(1-2.49)^2 + 25\cdot(2-2.49)^2 + 15\cdot(3-2.49)^2 + \\ &\qquad 12\cdot(4-2.49)^2 + 7\cdot(5-2.49)^2\big] \approx 1.62, \end{aligned}$$

pa procjena standardne devijacije iznosi:

$$s_{80} = \sqrt{\frac{1}{80-1}\sum_{i=1}^{5} n_i(x_i - 2.49)^2} \approx 1.27.$$

3.3 Procjena funkcije distribucije i funkcije gustoće

Procijeniti funkciju distribucije F znači procijeniti $F(x), \forall x \in \mathbb{R}$. Po definiciji, $F(x) = P(X \leq x)$, što znači da se problem procjene distribucije svodi na procjenu vjerojatnosti. S obzirom da je F u potpunosti nepoznata, kažemo da se radi o **neparametarskom modelu**.

Definicija 3.9. *Neka je F f-ja distribucije slučajne varijable X i $(X_1, \ldots, X_n)$ j.s.u. iz distribucije F. Kao jednog procjenitelja od F definiramo* ***empirijsku funkciju distribucije*** *s:*

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{\{X_i \leq x\}}, \quad x \in \mathbb{R}, \tag{13}$$

gdje je

$$I_{\{X_i \leq x\}} = \begin{cases} 1, & \text{ako je } X_i \leq x \\ 0, & \text{inače} \end{cases} \tag{14}$$

$\Rightarrow F_n$ je proporcija (relativna frekvencija) vrijednosti manjih ili jednakih od x, tj. prosjek uzorka $I_{\{X_1 \leq x\}}, \ldots, I_{\{X_n \leq x\}}$.

Kako se procjena vjerojatnosti svodi na procjenu očekivanja, slijedi da je $F_n(x)$ nepristran i konzistentan procjenitelj za $F(x), \forall x \in \mathbb{R}$. Ako ponovno promotrimo slučajnu varijablu (14), vidimo da ona ima Bernoullijevu distribuciju u kojoj je vjerojatnost realizacije uspjeha jednaka $F(x)$, pa za očekivanje i varijancu procjenitelja vrijedi:

$$E[F_n(x)] = F(x)$$

$$Var[F_n(x)] = \frac{1}{n} F(x)[1 - F(x)]$$

S obzirom da su slučajne varijable $X_1, \ldots, X_n$ nezavisne i da je

$$I_{\{X_i \leq x\}} = \begin{pmatrix} 0 & 1 \\ 1 - F(x) & F(x) \end{pmatrix},$$

imamo $nF_n(x) \sim \mathcal{B}(n, F(x)), \forall x \in \mathbb{R}$, pa iz Borelovog jakog zakona velikih brojeva[4] i fundamentalnog teorema statistike[5] slijedi da empirijska funkcija distribucije $F_n(x)$ uniformno konvergira po x prema $F(x), \forall x \in \mathbb{R}$.

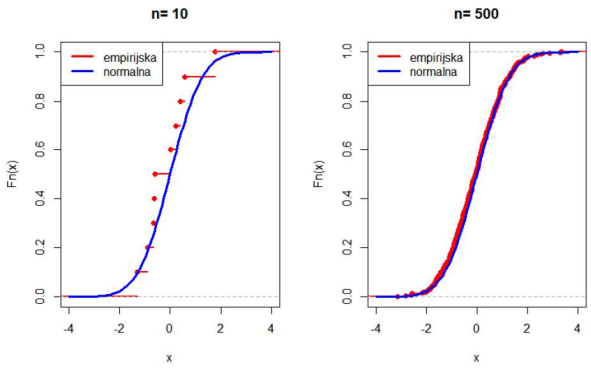
[4]Neka je $Z_n \sim \mathcal{B}(n, p), n \in \mathbb{N}$. Tada je $(g.s.) \lim_{n \to \infty} \frac{Z_n}{n} = p.$ (g.s. - gotovo sigurno)

[5]Neka je $X_1, \ldots, X_n$ niz n.j.d. slučajnih varijabli s funkcijom distribucije F i F_n empirijska funkcija distribucije bazirana na uzorku $X_1, \ldots, X_n$. Tada je $P\{ \lim_{n \to \infty} [sup_{x \in \mathbb{R}} |F_n(x) - F(X)|] = 0 \} = 1$

Primjer 3.3. *Za zapis distribucije neprekidne slučajne varijable koristimo funkciju gustoće, f, za koju vrijedi*

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

pa je vjerojatnost da se X realizira određenom vrijednošću iz nekog intervala jednaka površini ispod grafa funkcije f.
Stoga, ako uzmemo dva slučajna uzorka od, na primjer, 10 i 500 brojeva iz standardne normalne distribucije[6] *i procijenimo tu distribuciju sa empirijskom, na grafovima gustoća empirijske funkcije distribucije i teorijske distribucije za oba uzorka vidimo da empirijska funkcija distribucije uniformno konvergira prema teorijskoj (standardnoj normalnoj distribuciji).*



Slika 9: Grafovi teorijske distribucije i distribucije uzorka

Primjer 3.4. *Pomoću online ankete ispitano je 100 ljudi za njihovo mišljenje o novoizašloj seriji, gde 1 označava "ne sviđa mi se", 2 označava "sviđa mi se", 3 označava "niti mi se sviđa niti mi se ne sviđa" i 4 "nisam ju pogledao/la". Prikupljeni podatci prikazani su u potonjoj tablici:*

x_i	1	2	3	4
n_i	33	28	3	36

Tablica 4: Tablica frekvencija odgovora ispitanika

S obzirom da se radi o diskretnoj slučajnoj varijabli X kojom modeliramo odgovore, njezinu distribuciju možemo procijeniti ***empirijskom tablicom***

[6]simulacija napravljena u programskom paketu R

distribucije

$$\begin{pmatrix} x_1 & \dots & x_n \\ \hat{p_1} & \dots & \hat{p_n} \end{pmatrix}$$

pri čemu je $\hat{p}_i$ *nepoznat za svaki* i*. Svaku od tih vrijednosti procjenjujemo odgovarajućom relativnom frekvencijom* $\hat{p}_i = \dfrac{f_i}{n}$*, gdje je* f_i *frekvencija od* x_i*. Tada je procjena distribucije od* X *jednaka*

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 0.33 & 0.28 & 0.03 & 0.36 \end{pmatrix}.$$

Kao jednog procjenitelja funkcije gustoće neprekidne slučajne varijable koristimo procjenitelja jezgrom:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} k\left(\frac{x - X_i}{h}\right),$$

gdje je k funkcija jezgre koja zadovoljava uvjet nenegativnosti

$$\int_{-\infty}^{\infty} k(x) = 1,$$

n veličina uzorka, a h parametar zaglađivanja povezan uz širinu stupca histograma.

Literatura

[1] M. Benšić, N. Šuvak, *Primijenjena statistika*, Sveučilište J.J. Strossmayera, Odjel za matematiku, Osijek, 2013.

[2] F. Daly, D.J. Hand, M.C. Jones, A.D. Lunn, K.J. McConway, *Elements of Statistics*, Addison Wesley, 1995.

[3] Ž. Pauše, *Uvod u matematičku statistiku*, Školska knjiga, Zagreb, 1993.

[4] John A. Rice, *Mathematical Statistics and Data Analysis*, Treće izdanje, University of California, Berkeley, 2007.