Statistike

Statistike su slučajne promenljive $Y = f(X_1, X_2, ..., X_n)$ koje se formiraju na osnovu prostog slučajnog uzorka $X_1, X_2, ..., X_n$.

Na osnovu definicije prostog slučajnog uzorka, osobina funkcije f i raspodele obeležja, mogu se odrediti karakteristike sp Y.

Polazeći od realizovanog uzorka $(x_1, x_2, ..., x_n)$ računamo realizovanu vrednost $y = f(x_1, x_2, ..., x_n)$ sp Y.

Uzoračka sredina

- **Defnicija. Neka je $X_1, X_2, ..., X_n$ prost slučajan uzorak obima n za posmatrano obeležje X. Uzoračka sredina je statistika**

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

- Na osnovu definicije prostog slučajanog uzorka i osobina matematičkog očekivanja

$$E(\overline{X}_n) = E(X) = m \qquad D(\overline{X}_n) = \frac{\sigma^2}{n}$$

Izračunavanje uzoračke sredine

- Ako su podaci u uzorku dati kao niz vrednosti $x_1, \ldots, x_n$ bez sređivanja, tada je realizovana vrednost uzoračke sredine

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

- Ako je uzorak dat u obliku tabele, tada je realizovana vrednost statistike $\bar{X}_n$

Tabela 1.

Vrednost obeležja	x_1	x_2	...	x_k
frekvencija	n_1	n_2	...	n_k

$$\bar{x}_n = \frac{1}{n}(n_1 x_1 + \ldots + n_k x_k)$$

Izračunavanje uzoračke sredine

- Ako je uzorak dat u obliku Tabele 2, prvo se odrede predstavnici intervala $[a_j, a_{j+1})$ – najčešće su to njihove sredine x_j'. Tada je realizovana vrednost statistike $\overline{X}_n$

Tabela 2.

Vrednost obeležja	$[a_1, a_2)$	$[a_2, a_3)$	...	$[a_j, a_{j+1}]$
frekvencija	n_1	n_2	...	n_k

$$\overline{x}_n = \frac{1}{n}(n_1 x'_1 + \ldots + n_k x'_k)$$

Uzoračka disperzija

- Uzoračka sredina je blisko povezana sa matematičkim očekivanjem obeležja, daje nam podatak o prosečnoj vrednosti obeležja na uzorku.
- Uzoračka disperzija daje odstupanje vrednosti obeležja od prosečne vrednosti.
- **Neka je $X_1, X_2, ..., X_n$ prost slučajan uzorak obima n za posmatrano obeležje X. Ukoliko se smatra da je za obeležje X poznato matematičko očekivanje $E(X)=m$, tada je uzoračka disperzija statistika**

$$\widetilde{S}_n^2 = \frac{1}{n}((X_1 - m)^2 + ... + (X_n - m)^2)$$

Uzoračka disperzija, nastavak

- Ako matematičko očekivanje nije poznato, tada je **uzoračka disperzija**

$$\overline{S}_n^2 = \frac{1}{n}((X_1 - \overline{X}_n)^2 + \ldots + (X_n - \overline{X}_n)^2)$$

- **Korigovana uzoračka disperzija** je

$$\hat{S}_n^2 = \frac{1}{n-1}((X_1 - \overline{X}_n)^2 + \ldots + (X_n - \overline{X}_n)^2)$$

- **Uzoračka disperzija** se može računati i po formuli:

$$\overline{S}_n^2 = \frac{1}{n}(X_1 + \ldots + X_n)^2 - \left(\overline{X}_n\right)^2$$

Uzoračka disperzija i mat. očekivanje

- Veza korigovane i uzoračke disperzije je:

$$\hat{S}_n^2 = \frac{n}{n-1}\overline{S}_n^2$$

- Na osnovu definicije uzoračke disperzije i osobina matematičkog očekivanja dobija se:

$$E(\widetilde{S}_n^2) = E\left(\frac{1}{n}((X_1 - m)^2 + \ldots + (X_n - m)^2)\right) = D(X) = \sigma^2$$

$$E(\overline{S}_n^2) = E\left(\frac{1}{n}(X_1 + \ldots + X_n)^2 - \left(\overline{X}_n\right)^2\right) = \frac{n-1}{n}D(X)$$

$$E(\hat{S}_n^2) = E\left(\frac{n}{n-1}\overline{S}_n^2\right) = D(X)$$

Uzorački moment drugog reda

- Ako su podaci dati po intervalima, javlja se razlika između vrednosti uzoračke disperzije dobijene na osnovu podataka i uzoračke disperzije koja se dobija na osnovu podataka sređenih intervalno.
- Neka je $\overline{x}_n^2$ uzorački moment drugog reda računat na osnovu Tabele 1

$$\overline{x}_n^2 = \frac{1}{n}\sum_{j=1}^{k} n_j x_j^2$$

a uzorački moment drugog reda računat za iste podatke predstavljene u Tabeli 2 je $\overline{x}_n^{2^*}$ (intervali su dužine d). Tada

$$\overline{x}_n^2 = \overline{x}_n^{2^*} - \frac{d^2}{12}$$

Šepardova korekcija

Uzorački mod

- **Mod uzorka je** (u slučaju Tabele 1) **svaka vrednost x_j obeležja za čiju odgovarajuću frekvenciju n_j važi:**

$$n_j > n_{j-1} \text{ i } n_j > n_{j+1} .$$

- U slučaju Tabele 2, ako su dužine intervala jednake c, a mod se nalazi u intervalu $[a_j, a_{j+1})$, tada je mod uzorka

$$m_0 = a_j + \frac{\delta c}{\delta + \Delta}$$

$$\delta = n_j - n_{j-1} \qquad \Delta = n_j - n_{j+1}$$

- Ako postoji samo jedan mod, raspodela je ***unimodalna***, ako ima dva moda, ***bimodalna***, a ako ima više modova, raspodela je ***polimodalna***.

Medijana uzorka

- **Medijana uzorka** se u slučaju Tabele 1 dobija tako što se prvo napiše varijacioni niz

$$y_1 \le y_2 \le \ldots \le y_n$$

pa je medijana uzorka

$$m_e = \begin{cases} y_{k+1}, & n = 2k+1 \\ \frac{1}{2}(y_k + y_{k+1}), & n = 2k \end{cases}$$

- Ako su podaci dati u obliku Tabele 2, a medijana se nalazi u intervalu $[a_j, a_{j+1})$, tada je uzoračka medijana:

$$m_e = a_j + [\frac{n}{2} - \sum_{k=1}^{j+1} n_j] \frac{c}{n_j}$$

Uzorački moment reda k

- Neka je $x_1, \ldots, x_n$ realizovani prost slučajan uzorak obima n za obeležje X. **Običnи uzorački moment reda k** je

$$\frac{1}{n}\sum_{i=1}^{n} x_i^k$$

- **Centralni uzorački moment reda k** je

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

- Uzoračka sredina je uzorački moment prvog reda, a uzoračka disperzija je uzorački moment drugog reda.

Uzorački koeficijenti

- **Uzorački koeficijent varijacije** je

$$c_V = \frac{\bar{s}_n}{\bar{x}_n}$$

$$\bar{s}_n = \sqrt{\frac{1}{n}\sum_{j=1}^{k} n_j (x_j - \bar{x}_n)^2}$$

- **Uzorački koeficijent asimetrije** je

$$\frac{c_3}{(\bar{s}_n)^3}$$

$$c_3 = \frac{1}{n}\sum_{j=1}^{k} n_j (x_j - \bar{x}_n)^3$$

- **Uzorački koeficijent spljoštenosti** je

$$\frac{c_4}{(\bar{s}_n)^4} - 3$$

$$c_4 = \frac{1}{n}\sum_{j=1}^{k} n_j (x_j - \bar{x}_n)^4$$

Računanje realizovanih statistika

- Za sledeći primer, odrediti uzoračku sredinu, disperziju, mod, medijanu, koeficijent varijacije, koeficijent asimetrije i koeficijent spljoštenosti.

Broj četvorki	0	1	2	3	4	5	6
Broj godina	12	21	14	8	2	2	1

- Uzoračka sredina je $\bar{x}_{60} = \frac{1}{60}(0 \cdot 12 + \ldots + 6 \cdot 1) = 1{,}6167$
- Uzoračka disperzija je $\bar{s}_{60}^2 = \frac{1}{60}\left((0 - \bar{x}_{60})^2 \cdot 12 + \ldots + (6 - \bar{x}_{60})^2 \cdot 1\right) = 1{,}8364$
- Uzorački mod je $m_0 = 1$, uzoračka medijana je 3.
- Uzorački koeficijent varijacije je $c_V = \frac{\bar{s}_{60}}{\bar{x}_{60}} = 0{,}8382$
- Uzorački koeficijent asimetrije je $\frac{1}{60(\bar{s}_{60})^3}\left((0 - \bar{x}_{60})^3 \cdot 12 + \ldots + (6 - \bar{x}_{60})^3 \cdot 1\right) = 1{,}7839$
- Uzorački koeficijent spljoštenosti je $\frac{1}{60(\bar{s}_{60})^4}\left((0 - \bar{x}_{60})^4 \cdot 12 + \ldots + (6 - \bar{x}_{60})^4 \cdot 1\right) - 3 = 1{,}0048$

Statistike kao slučajne promenljive

- Neka je dat prost slučajan uzorak $X_1, X_2, ..., X_N$. Uz standardnu oznaku $\overline{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j$
- ***Uzoračka koeficijent varijacije*** je statistika

$$C_V = \frac{\sqrt{\frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X}_n)^2}}{\overline{X}_n}$$

- ***Uzoračka koeficijent asimetrije*** je statistika

$$\pi_1 = \frac{\frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X}_n)^3}{\left(\sqrt{\frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X}_n)^2}\right)^3}$$

Statistike poretka

- **Statistika poretka prvog ranga** je

$$Y_1 = \min_{1 \le j \le n} X_j$$

- **Statistika poretka n-tog ranga** je

$$Y_n = \max_{1 \le j \le n} X_j$$

- **Statistike poretka** prvog, drugog, …, n-tog ranga su redom prvi, drugi, …, n-ti element varijacionog niza.
- **Raspon uzorka** je razlika statistike poretka n-tog i prvog ranga

$$R = Y_n - Y_1$$

Statistike kao slučajne promenljive

- ***Uzorački koeficijent spljoštenosti*** je statistika

$$\pi_3 = \frac{\frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X}_n)^4}{\left(\sqrt{\frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X}_n)^2}\right)^4} - 3$$

- ***Uzoračka medijana*** je statistika

$$m_e = \begin{cases} Y_{k+1}, & n = 2k+1 \\ \frac{1}{2}(Y_k + Y_{k+1}), & n = 2k \end{cases}$$

Zadatak

- Neka je $X_1, X_2, ..., X_n$ prost slučajan uzorak za obeležje X koje ima normalnu raspodelu $\mathcal{N}(m, \sigma^2)$. Dokazati da je raspodela uzoračke sredine $\overline{X}_n$ raspodela $\mathcal{N}(m, \sigma^2/n)$, a raspodela za

$$\frac{\widetilde{S}_n^2}{\sigma^2} \quad \text{je} \quad \chi_n^2 \quad \text{raspodela.}$$

$$\widetilde{S}_n^2 = \frac{1}{n}((X_1 - m)^2 + ... + (X_n - m)^2)$$

Uzorački kvantili

- ***Uzorački l-procentni kvantil*** je broj koji je veći od l% vrednosti iz uzorka. Ako je u pitanju 25% elemenata iz uzorka, kvantil se zove ***prvi kvartil*** i označava se sa q_1.
- Ako je u pitanju 50% elemenata iz uzorka odgovarajući kvantil se poklapa sa ***medijanom***.
- Ako je u pitanju 75% elemenata – ***treći kvartil*** - q_3.

Box-plot dijagram

- Pravougaoni dijagram je jedan način grafičkog prikazivanja podataka. Na izabranoj osi se odrede tačke koje odgovaraju uzoračkoj medijani i kvantilima q_1 i q_3.
- Zatim se računaju unutrašnje f_1 i f_3 i spoljašnje F_1 i F_3 granice dijagrama

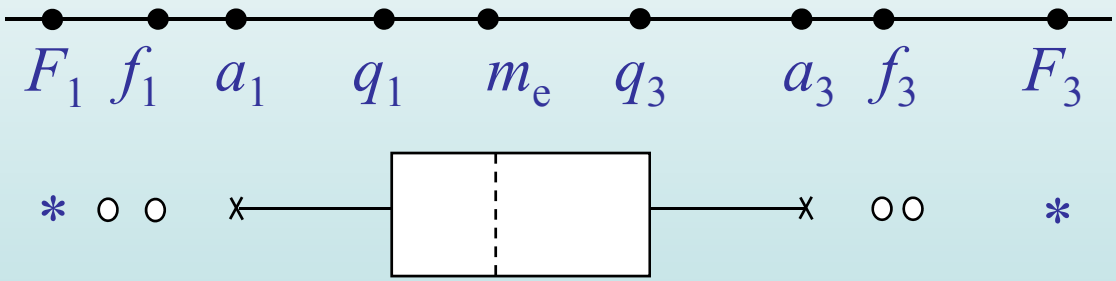
$$f_1 = q_1 - 1{,}5(q_3 - q_1) \qquad f_3 = q_1 + 1{,}5(q_3 - q_1)$$

$$F_1 = q_1 - 3(q_3 - q_1) \qquad F_3 = q_1 + 3(q_3 - q_1)$$

- Zatime se određuju a_1–najmanji među elementima uzorka koji su veći od f_1 i a_3–najveći među elementima uzorka koji su manji od f_3.

Box-plot dijagram

- Dijagram se sastoji od pravougaonika čija je jedna strana paralelna izabranoj osi i jednaka odsečku (q_1, q_3).
- U pravougaonik se ucrta linija koja odgovara uzoračkoj medijani m_e.
- Ako je linija blizu sredine pravougaonika, raspodela bi mogla biti simetrična.



- Kružićem o su označeni svi elementi uzorka koji su u intervalima $[F_1, f_1]$ i $[f_3, F_3]$, a zvezdicom * svi elementi uzorka manji od F_1 ili veći od F_3 (*extreme outliers*).

Korelaciona tabela

- Imamo prost slučajan uzorak obima n i posmatramo dva obeležja X i Y na elementima uzorka. Podaci iz uzorka mogu biti dati u obliku tabele koja se naziva ***korelaciona tabela***.

$X \setminus Y$	y_1	y_2	…	y_s	zbir
x_1	n_{11}	n_{12}	…	n_{1s}	$n(x_1)$
x_2	n_{21}	n_{22}	…	n_{2s}	$n(x_2)$
…	…	…	…	…	
x_m	n_{m1}	n_{m2}	…	n_{ms}	$n(x_m)$
zbir	$n(y_1)$	$n(y_2)$		$n(y_s)$	$\boldsymbol{n}$

- Broj n_{ij} označava da se par (x_i, y_j) pojavio n_{ij} puta u uzorku.

$$i = 1,\ldots,m \qquad j = 1,\ldots,s \qquad \sum_i \sum_j n_{ij} = n$$

obim uzorka

Uzorački koeficijent korelacije

- Kada su podaci dati u obliku korelacione tabele, uzoračke sredine obeležja X i Y su:

$$\bar{x}_n = \frac{1}{n}\sum_{j=1}^{m} n(x_j)x_j \qquad \bar{y}_n = \frac{1}{n}\sum_{j=1}^{s} n(y_j)y_j$$

- Uzoračke disperzije obeležja X i Y su:

$$\bar{S}_X^2 = \frac{1}{n}\sum_{j=1}^{m} n(x_j)(x_j - \bar{x}_n)^2 \qquad \bar{S}_Y^2 = \frac{1}{n}\sum_{j=1}^{s} n(y_j)(y_j - \bar{y}_n)^2$$

- **Uzoračke koeficijent korelacije** meri međusobnu zavisnost obeležja X i Y:

$$r = \frac{\frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{s} n_{ij}x_i y_j - \bar{x}_n\bar{y}_n}{\sqrt{\bar{S}_X^2\bar{S}_Y^2}}$$

Uzorački koeficijent korelacije

- Kada su podaci dati u obliku tabele

Vrednosti za X	x_1	x_2	…	x_n
Vrednosti za Y	y_1	y_2	…	y_n

- Uzoračke sredine su:

$$\bar{x}_n = \frac{1}{n}\sum_{j=1}^{n} x_j \qquad \bar{y}_n = \frac{1}{n}\sum_{j=1}^{n} y_j$$

- Uzoračke disperzije su:

$$\bar{S}_{nx}^2 = \frac{1}{n}\sum_{j=1}^{n} (x_j - \bar{x}_n)^2 \qquad \bar{S}_{ny}^2 = \frac{1}{n}\sum_{j=1}^{n} (y_j - \bar{y}_n)^2$$

- Uzorački koeficijent korelacije je:

$$r = \frac{\frac{1}{n}\sum_{j=1}^{n} (x_j y_j) - \bar{x}_n \bar{y}_n}{\sqrt{\bar{S}_X^2 \bar{S}_Y^2}}$$

Primer

- Neka je obeležje X broj dobijenih šestica u jednom bacanju, a obeležje Y broj dobijenih parnih brojeva.

X	2	1	0	2	0	1	1	1	1	2
Y	2	3	1	3	2	2	1	2	2	2

- Korelaciona tabela je

X/Y	1	2	3	
0	1	1	0	**2**
1	1	3	1	**5**
2	0	2	1	**3**
	2	**6**	**2**	**10**

- Uzoračke sredine su

$$\bar{x}_n = 11/10 = 1{,}1 \qquad \bar{y}_n = 20/10 = 2.$$

- Uzoračke disperzije su

$$\bar{S}^2_{nx} = 0{,}51 \qquad \bar{S}^2_{ny} = 0{,}4$$

- Uzorački koeficijent korelacije je

$$r = 0{,}4428$$