Deskriptivna statistika

Deskriptivna ili opisna statistika je grana statistike koja se bavi predočavanjem i opisivanjem glavnih karakteristika sakupljenih podataka (tablice, histogrami, srednje vrijednosti,…).

a) FREKVENCIJSKA TABLICA

- koristi se za pregledniji prikaz podataka

i	a_i	f_i	$r_i = \frac{f_i}{n}$
$\sum$		n	1

a_i =različite vrijednosti koje imamo u pokusu

f_i =frekvencije (broj pojavljivanja) tih različitih vrijednosti

r_i =relativne frekvencije (udjeli vrijednosti) - ne računamo ih ako nam ne trebaju kasnije u zadatku

n =ukupan broj podataka (broj ponavljanja pokusa)

b) HISTOGRAM RELATIVNIH FREKVENCIJA

- svaka vrijednost a_i se crta u posebnom stupcu (oznaka vrijednosti u sredini stupca)

- stupci su spojeni

- širina stupca je c (udaljenost između vrijednosti)

- visina stupca je $\frac{r_i}{c}$

- ukupna površina stupaca je 1 (uspoređujemo s funkcijom gustoće dane razdiobe)

- ne moramo koristiti iste skale na osima

- ukoliko se radi o diskretnom slučaju, možemo crtati i stupčasti dijagram (stupci nisu povezani, visina je uvijek jednaka ili frekvencijama ili relativnim frekvencijama)

c) ARITMETIČKA SREDINA UZORKA (PROSJEK)

$x_1, x_2, \dots, x_n$= uzorak, tj. dani podaci

$a_1, a_2, \dots, a_k$= različite vrijednosti (k=broj različitih vrijednosti)

$$\overline{\boldsymbol{x}} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\boldsymbol{a_1 f_1 + a_2 f_2 + \cdots + a_k f_k}}{\boldsymbol{n}}$$

d) VARIJANCA I STANDARDNA DEVIJACIJA UZORKA

One mjere koliko podaci odstupaju od aritmetičke sredine $\bar{x}$ (uzimaju se kvadrati kako bismo dobili pozitivne elemente u sumi).

VARIJANCA

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1} = \frac{{x_1}^2 + {x_2}^2 + \cdots + {x_n}^2 - n \cdot \bar{x}^2}{n-1}$$
$$= \frac{\boldsymbol{{a_1}^2 f_1 + {a_2}^2 f_2 + \cdots + {a_k}^2 f_k - n \cdot \bar{x}^2}}{\boldsymbol{n-1}}$$

- pojavljuje se i oznaka Var

STANDARDNA DEVIJACIJA

- ima poželjno svojstvo da mjeri varijabilnost u originalnim mjernim jedinicama

- pozitivni korijen varijance uzorka

$$\boldsymbol{s = \sqrt{s^2}}$$

e) MOD UZORKA

- oznaka Mod

- vrijednost s najvećom frekvencijom (tj. najčešća vrijednost)

f) RASPON UZORKA

- razlika između maksimuma i minimuma podataka

$x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ - uzlazno sortirani podaci

$$\boldsymbol{d = x_{(n)} - x_{(1)}}$$

g) MEDIJAN UZORKA

- medijan *m* je takav broj za koji vrijedi da je 50% svih podataka manje ili jednako njemu i 50% svih podataka veće ili jednako od njega

1.način – za uzlazno sortirane (uređene) nizove s neparnim brojem podataka, to je srednji član niza – ukoliko je broj podataka paran, *m* je aritmetička sredina srednja dva člana tog uređenog niza

2. način

$$\boldsymbol{m = x_{\left(\frac{n+1}{2}\right)}; \qquad x_{\left(\frac{p}{q}\right)} = x_{\left(k+\frac{r}{q}\right)} = x_{(k)} + \frac{r}{q}\left(x_{(k+1)} - x_{(k)}\right); \qquad k \in \mathbb{N}, \frac{r}{q} \in \langle 0, 1 \rangle}$$

h) KVARTILI i INTERKVARTIL

- **donji kvartil** q_L je takav broj da je 25% podataka niza manje od njega, a 75% veće od njega

- **gornji kvartil** q_U je takav broj da je 75% podataka niza manje od njega, a 25% veće od njega

1. način – donji kvartil je medijan prvog podniza dobivenog dijeljenjem početnog niza medijanom na dva podniza, dok je gornji kvartil medijan drugog takvog podniza (tj. možemo ih računati na 1. način kao medijan)

2. način

$$\boldsymbol{q_L = x_{\left(\frac{n+1}{4}\right)}}, \qquad \boldsymbol{q_U = x_{\left(\frac{3(n+1)}{4}\right)}}$$

INTERKVARTIL – razlika gornjeg i donjeg kvartila, oznaka d_q: $\boldsymbol{d_q = q_U - q_L}$

i) DIJAGRAM PRAVOKUTNIKA (box-and-whisker plot)

- crtamo pomoću **karakteristične petorke ($\boldsymbol{x_{(1)}, q_L, m, q_U, x_{(n)}}$)**

- rubovi pravokutnika (box) su donji i gornji kvartil, medijan se ucrtava kao zadebljana linija unutar pravokutnika

- ako su maksimum i minimum (i eventualno još neke druge vrijednosti uz njih) udaljeni od kvartila za **više od** $\boldsymbol{\frac{3}{2} d_q}$, nazivamo ih OUTLIERS (pripisuju se najčešće pogrešci tijekom mjerenja, uobičajeni za velike uzorke) i posebno se označavaju npr. kružićem

- brkovi (whiskers) su najmanja i najveća vrijednost koje nisu outlieri

- ako nema outliera, brkovi su maksimum i minimum

j) UZORAČKI CENTRALNI MOMENT REDA k, $k \in \mathbb{N}$

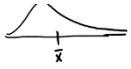
$\mu_k = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^k$; uvijek je $\mu_1 = 0,\ \mu_2 = s^2$

centralni moment reda 3: $\boldsymbol{\mu_3 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^3 = \frac{1}{n-1}\sum_{i=1}^{k} f_i(a_i - \bar{x})^3}$

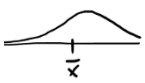
KOEFICIJENT ASIMETRIJE UZORKA: $\boldsymbol{\alpha_3 = \frac{\mu_3}{s^3}}$

Ako je $\mu_3 = 0$, tj. ako je $\alpha_3 = 0$, uzorak je simetričan s obzirom na aritmetičku sredinu $\bar{x}$.

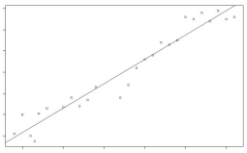
Ako je $\alpha_3 > 0$, uzorak je pozitivno asimetričan (right tail; positive skew).



Ako je $\alpha_3 < 0$, uzorak je negativno asimetričan (left tail; negative skew).



Linearna regresija



Tražimo pravac $y = ax + b$ koji je (u određenom smislu) najbliži danim točkama, tj. parovima podataka.

Dane su točke (x_1, y_1), (x_2, y_2), …, (x_n, y_n). Pa je n **- broj parova podataka.**

Prvo, računamo prosjek x-eva, tj. $\overline{x} = \frac{\sum x_i}{n}$.

Računamo prosjek y-a, tj. $\overline{y} = \frac{\sum y_i}{n}$.

Tada,

$$s_{xx} = \frac{1}{n-1}\left(x_1^2 + x_2^2 + \cdots + x_n^2 - n \cdot \overline{x}^2\right)$$

$$s_{xy} = \frac{1}{n-1}(x_1y_1 + x_2y_2 + \cdots + x_ny_n - n \cdot \overline{x} \cdot \overline{y}).$$

Pa dobivamo

$$a = \frac{s_{xy}}{s_{xx}} \qquad i \qquad b = \overline{y} - a \cdot \overline{x}.$$

Na kraju, uvrstimo a i b u formulu za pravac, tj. uvrstimo ih u

$$y = ax + b.$$

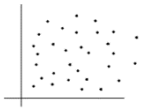
Pearsonov koeficijent korelacije

Računamo i $s_{yy} = \frac{1}{n-1}\left(y_1^2 + y_2^2 + \cdots + y_n^2 - n \cdot \overline{y}^2\right)$.

Pearsonov koeficijent korelacije iznosi

$$r = \frac{s_{xy}}{\sqrt{s_{xx} \cdot s_{yy}}}.$$

Uvijek je $-1 \leq r \leq 1$.



Ako je $r = 0$, tada nema korelacije između x-eva i y-a.

Ako je $r > 0$, tada kažemo da je korelacija pozitivna. To znači da ako x raste, tada u pravilu y također raste.

Ako je $r < 0$, tada kažemo da je korelacija negativna. To znači da ako x raste, tada u pravilu y pada.