

# Statistika – sažetak i popis formula

## 1. Deskriptivna statistika

**Aritmetička sredina** brojeva  $x_1, x_2, \dots, x_n$  :

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

Na primjer, aritmetička sredina brojeva 1,2,3,4,5 je broj  $\frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$ .

**Frekvencija** nekog podatka je broj pojavljivanja tog podatka. Na primjer, za podatke 1,1,2,2,2,3,4 broj 1 ima frekvenciju 2, broj 2 frekvenciju 3, a brojevi 3 i 4 po frekvenciju 1.

Ako podatke grupiramo u razrede, onda slično definiramo **frekvencije razreda**.

**Relativna frekvencija** (podatka ili razreda), po definiciji je kvocijent obične frekvencije i ukupnog broja podataka. Zato je zbroj relativnih frekvencija jednak 1.

**Medijan** skupa podataka je srednji podatak ako je broj podataka neparan, a aritmetička sredina dvaju srednjih ako je broj podataka paran.

Na primjer, za podake 1,2, 4, 11, 13 medijan je 4 (srednji podatak),

a za podatke 1,2,4,7,11,13 medijan je  $\frac{4+7}{2} = 5.5$  (aritmetička sredina 3. i 4. podatka)

**Raspon** podataka  $x_1, x_2, \dots, x_n$  poredanih prema veličini je razlika  $x_n - x_1$  najvećeg i najmanjeg podatka.

Na primjer, raspon podataka 1,1,2,2,3,11,64 je  $64-1=63$

**Kvartili** dijele podatke u četiri jednakobrojne skupine.

**Prvi ili donji kvartil** je broj od kojega je 25% podataka manje ili je njemu jednako.

**Drugi** je kvartil medijan.

**Treći ili gornji kvartil** je broj od kojega je 75% podataka manje ili je njemu jednako.

**Mjere rasipanja** (disperzije) podataka.

### 1. Suma apsolutnih vrijednosti odstupanja podataka od aritmetičke sredine:

$$SAO := |x_1 - \bar{x}| + |x_2 - \bar{x}| + \dots + |x_n - \bar{x}|.$$

### 2. Prosječno apsolutno odstupanje od aritmetičke sredine:

$$PAO := \frac{|x_1 - \bar{x}| + |x_2 - \bar{x}| + \dots + |x_n - \bar{x}|}{n}$$

**3. Varijanca uzorka**  $(s')^2$  definira se kao **prosječno kvadratno odstupanje od prosjeka**:

$$(s')^2 := \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n}$$

**4. Standardna devijacija uzorka**  $s'$  je drugi korijen iz varijance uzorka:

$$s' := \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n}}$$

**5. Korigirana varijanca** (nepristrana procjena varijance populacije)

$$s^2 := \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n-1}$$

**(razlikuje se po tome što u nazivniku, umjesto n ima n-1, a u oznaci što nema crtice).**

**6. korigirana standardna devijacija uzorka**  $s$ , kojom se procjenjuje standardna devijacija populacije:

$$s := \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n-1}}$$

Dosadašnje pojmove ilustriramo Primjerom 9. iz lekcije: *Deskriptivna statistika*.

**Primjer 9.** Mjerenjem vremena između dviju uzastopnih poruka pristiglih na neku adresu dobiveni su sljedeći podatci (u sekundama):

12, 8, 1, 7, 24, 4, 4, 6, 20, 10, 3, 2, 22, 23, 8, 6, 5, 25, 16, 3, 1, 14, 15, 18, 2, 6, 27, 19, 12, 4, 20, 14, 3, 13, 8, 15, 30, 5, 7, 16.

(I) Prebrojimo podatke. Vidimo da ih ima 40, dakle  $n = 40$ .

(II) Poredajmo podatke prema veličini (od manjeg prema većem):

1, 1, 2, 2, 3, 3, 3, 4, 4, 4, 5, 5, 6, 6, 6, 7, 7, 8, 8, 8, 10, 12, 12, 13, 14, 14, 15, 15, 16, 16, 18, 19, 20, 20, 22, 23, 25, 27, 30.

(III) Napravimo tablicu frekvencija:

1	2	3	4	5	6	7	8	10	12	13	14	15	16	18	19	20	22	23	24	25	27	30
2	2	3	3	2	3	2	3	1	2	1	2	2	2	1	1	2	1	1	1	1	1	1

Vidimo da frekvencije variraju iako imaju i opći trend prema opadanju. To bi još izrazitije bilo da smo stavili frekvencije 0 za brojeve od 1 do 30 koji se ne pojavljuju.

(IV) Grupirajmo podatke u razrede duljine 5:

0.5 - 5.5    5.5 - 10.5    10.5 - 15.5    15.5 - 20.5    20.5 - 25.5    25.5 - 30.5

11            9            7            6            4            2

Vidimo da, nakon ovakvog grupiranja, frekvencije razreda opadaju, što se dobro vidi i iz histograma. To je jedan od najvažnijih razloga grupiranja.

(V) Odredimo, najmanji podatak, najveći podatak i raspon:

min = 1

max = 30

raspon = max – min = 30-1 = 29.

(VI) Odredimo medijan i aritmetičku sredinu i unaprijed procijenimo njihov odnos.

Odredimo kvartile.

S obzirom da su podatci više grupirani na početak, medijan je manji od aritmetičke sredine.

Kako je n = 40, medijan je aritmetička sredina 20-og i 21-og podatka. Dakle:

$$\text{Medijan} = \frac{8+10}{2} = 9$$

**Aritmetička sredina**,  $\bar{x} = \frac{458}{40} = 11.45$  (zaista je medijan manji).

**Prvi kvartil**:  $q_1 = 4.5$

**Drugi kvartil** (medijan):  $q_2 = 9$

**Treći kvartil**:  $q_3 = 17$

(VII) Odredimo varijancu i standardnu devijaciju te korigiranu varijancu i korigiranu standardnu devijaciju uzorka.

**Varijanca**:  $(s')^2 = 63.1975$

**Standardna devijacija**:  $s' = 7.9497$  (na 4 decimale)

**Korigirana varijanca**:  $s^2 = 64.8179$  (na 4 decimale)

**Korigirana standardna devijacija**:  $s = 8.0510$  (na 4 decimale).

### **Empirijsko pravilo za zvonolike distribucije frekvencija.**

Kažemo da podatci imaju **zvonoliku distribuciju** ako za histogram frekvencija (ili relativnih frekvencija, svejedno) vrijedi:

(N1) Površina je koncentrirana oko aritmetičke sredine.

(N2) Površina je približno simetrično raspoređena lijevo i desno od aritmetičke sredine

(N3) Površine rastu odprilike do aritmetičke sredine, potom padaju.

Uz ove uvjete histogram (odnosno pripadna krivulja) ima **zvonolik oblik**. Praksa pokazuje da takav oblik imaju histogrami distribucija kod **velikih uzoraka**, pri mjerenju mnogih statističkih fenomena (**statističkih obilježja**), poput mase, visine, postotka elementa koji se može nekom tehnološkom metodom izdvojiti iz neke rudače, grješaka pri mjerenju, kvocijenta inteligencije itd. Za takva statistička obilježja **uočeno je** sljedeće **empirijsko pravilo**:

U intervalu  $\langle \bar{x} - s', \bar{x} + s' \rangle$  ima oko 68% podataka, tj. oko 2/3 podataka (površine histograma)

U intervalu  $\langle \bar{x} - 2 \cdot s', \bar{x} + 2 \cdot s' \rangle$  ima oko 95% podataka (površine histograma)

U intervalu  $\langle \bar{x} - 3 \cdot s', \bar{x} + 3 \cdot s' \rangle$  su gotovo svi podatci (gotovo čitava površina).

## 2. Procjenjivanje.

Neka je  $X$  slučajna varijabla.

Očekivanje  $E(X)$  procjenjujemo aritmetičkom sredinom podataka

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

Varijancu  $V(X)$  procjenjujemo izrazom

$$s^2 = \frac{(x_1 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n-1}, \quad (\text{u nazivniku je } n-1, \text{ a ne } n)$$

Standardnu devijaciju  $s(X)$  procjenjujemo izrazom  $s = \sqrt{\frac{(x_1 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n-1}}$ .

## 2. Interval pouzdanosti za očekivanje – prava vrijednost mjerene veličine.

Označimo  $E(X) = \mu$  i  $V(X) = \sigma^2$ , bez obzira je li  $X$  normalno distribuirana.

Očekivanje procjenjujemo aritmetičkom sredinom podataka, ali aritmetička sredina ne mora biti (i u pravilu nije) jednaka (nepoznatom) očekivanju. Zato nas zanima **interval** oko  $\bar{x}$  unutar kojega će, uz određenu sigurnost, biti očekivanje  $\mu$ . To je **interval pouzdanosti**.

### Postupak određivanja intervala pouzdanosti.

1. **Ako je  $X$  normalno distribuirana i ako je poznata standardna devijacija  $\sigma$ .**

Tada je, uz 95% vjerojatnost, interval pouzdanosti (odprilike)

$$\left\langle \bar{x} - 2 \frac{\sigma}{\sqrt{n}}, \bar{x} + 2 \frac{\sigma}{\sqrt{n}} \right\rangle$$

Smisao intervala pouzdanosti nije da se očekivanje  $\mu$  u njemu nalazi s vjerojatnošću 0.95 (naime  $\mu$  nije slučajna veličina i nalazi se ili ne nalazi u tom intervalu). Taj se smisao može interpretirati na primjer tako da bi se odprilike u 95 od 100 ponavljanja ovih  $n$  mjerenja, aritmetička sredina  $\bar{x}$  našla u intervalu

$$\left\langle \mu - 2 \frac{\sigma}{\sqrt{n}}, \mu + 2 \frac{\sigma}{\sqrt{n}} \right\rangle \quad (\text{što bismo mogli provjeriti da znamo } \mu \text{ i } \sigma),$$

a to je isto kao da kažemo da bi se odprilike u 95 od 100 ponavljanja, očekivanje  $\mu$  našlo u intervalu  $\left\langle \bar{x} - 2 \frac{\sigma}{\sqrt{n}}, \bar{x} + 2 \frac{\sigma}{\sqrt{n}} \right\rangle$  (što bismo opet mogli provjeriti da znamo  $\mu$  i  $\sigma$ ).

Umjesto broja 2, za vjerojatnost 0.95, mogli bismo u tablici jedinične normalne razdiobe  $T$  (ili odgovarajućoj proceduri u Excelu ili Mathematici) naći precizniji podatak: 1.96. Naime,  $P(|T| < 1.96) = 0.95$

Slično bismo mogli odrediti simetrične intervale oko aritmetičke sredine za druge vjerojatnosti, a ne samo za 0.95.

Općenito je interval pouzdanosti za vjerojatnost  $1-2p$ , jednak

$$\left\langle \bar{x} - z_p \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + z_p \cdot \frac{\sigma}{\sqrt{n}} \right\rangle$$

gdje je  $z_p$  takav realni broj, za kojega vrijedi  $P(T > z_p) = p$ , zj. broj iza kojega je površina ispod grafa funkcije gustoće jedinične normalne razdiobe jednaka  $p$ .

Veličina  $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$  koja se tu pojavljuje zove se **standardna grješka**, gdje je  $n$  broj mjerenja (duljina uzorka).

2. **Ako je  $n$  velik** (obično se uzima ako je  $n > 30$ ), **i ako je poznata standardna devijacija  $\sigma$** , a  **$X$  ne mora biti normalno distribuirana.**

Tada možemo postupiti kao u 1.

Treba napomenuti da je pretpostavka da znamo  $\sigma$  (a da  $\mu$  procijenjujemo iz  $n$  mjerenja) nerealna, iako nije nemoguća. U praksi smo gotovo uvijek prisiljeni procijeniti  $\sigma$  pomoću  $s$ . Tada se situacija usložnjava, međutim za parametre normalne razdiobe, tj. ako pretpostavimo da je  $X$  normalno distribuirana, problem se može riješiti.

3.  **$n < 30$ ,  $X$  je normalno distribuirana, a  $\sigma$  nepoznat – procjenjujemo ga pomoću  $s$**  (postupak korektan za sve  $n$ )

Tada je interval pouzdanosti, uz vjerojatnost  $1-2p$ :

$$\left\langle \bar{x} - t_p(k) \frac{s}{\sqrt{n}}, \bar{x} + t_p(k) \frac{s}{\sqrt{n}} \right\rangle.$$

gdje je  $t(n-1)$  Studentova razdioba s  $k=n-1$  stupnjeva slobode, a značenje broja  $t_p(k)$  je sljedeće:

$$P(|t(k)| > t_p(k)) = 2p, \quad \text{tj.} \quad P(t(k) > t_p(k)) = p$$

**Ako je  $n$  dovoljno velik**, recimo oko 30, onda je  $t(n-1)$  praktično jednaka jediničnoj normalnoj razdiobi, pa možemo umjesto Studentove razdiobe koristiti jediničnu normalnu. Naravno, ako se služimo određenim statističkim paketom, to je nepotrebno. Također, tada interval pouzdanosti dobijemo izravno.

## Testiranje hipoteze $\mu = \mu_0$ (t-test)

Predpostavimo da je  $X$  normalno distribuirana slučajna veličina s očekivanjem  $\mu$  i varijancom  $\sigma^2$ .

Neka smo na osnovi  $n$  mjerenja dobili procjene:

$\bar{x}$  za njeno očekivanje  $\mu$ ,

$s^2$  za njenu varijancu  $\sigma^2$ .

Testiramo hipotezu:

$H_0: \mu = \mu_0$ ,

gdje je  $\mu_0$  neka deklarirana vrijednost.

Napominjemo da bismo prije toga trebali provjeriti hipotezu o bliskosti varijanca (koju treba formulirati), a nakon što testiranje varijanaca pozitivno prođe, možemo pristupiti testiranju očekivanja.

Testiranje se zasniva na činjenici da broj  $\frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$  možemo interpretirati kao slučajnu

vrijednost slučajne varijable  $t(n-1)$  (ta se razdioba zove **test-statistika**).

Postupak opisujemo uz **kontrahipotezu**  $\mu \neq \mu_0$ , dakle imamo:

**(I)**

$H_0: \mu = \mu_0$

$H_a: \mu \neq \mu_0$

1. Računamo  $t_{\text{exp}} = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$ .

2. Biramo **nivo signifikantnosti** (razinu značajnosti)  $\alpha$  što je obično 0.05. Značenje nivoa signifikantnosti je  $\alpha = P(H_0 \text{ odbacujemo} | H_0 \text{ je istinita})$ .

Taj se broj zove i **pogrješka prve vrste**.

3. U tablici t-razdiobe određujemo kritičnu vrijednost  $t_0$  (ovisno o broju stupnjeva slobode  $k=n-1$ , i kontrahipotezi koja je, ako drukčije ne specificiramo  $\mu \neq \mu_0$ ).

Značenje kritične vrijednosti:  $t_0 = t_{\frac{\alpha}{2}}(k)$ , tj.  $P(|t(k)| > t_0) = \alpha$ .

4. Ako je  $|t_{\text{exp}}| < t_0$  hipotezu prihvaćamo, inače je odbacujemo. Područje između kritične vrijednosti i njoj suprotne  $<-t_0, t_0>$  zovemo **područjem prihvaćanja (kritično područje)**, ostatak je **područje odbacivanja**. Smisao je u tome, što hipotezu prihvaćamo ako  $t_{\text{exp}}$  upadne u područje prihvaćanja, inače je odbacujemo.

Ovaj test zovemo **dvostrukim**, naziv možemo tumačiti tako što se područje odbacivanja od dvaju simetričnih dijelova. Naime, tu područje odbacivanja ima dva simetrična dijela, svaki površine  $\frac{\alpha}{2}$ , gdje je  $\alpha$  nivo signifikantnosti. To je zato što je kontrahipoteza oblika  $\mu \neq \mu_0$ ,

pa se dopuštaju otkloni na obje strane. Dakle, u slučaju  $\alpha=0.05$ , broj  $t_0$ , označava broj iza kojega je ispod grafa t-razdiobe površina jednaka 0.025.

Kontrahipotezu  $\mu \neq \mu_0$  koristimo u pravilu onda ako su neki podatci iz uzorka manji, a neki veći od deklarirane vrijednosti  $\mu_0$ .

(II).

$$H_0: \mu = \mu_0$$

$$H_a: \mu > \mu_0$$

Tu hipotezu koristimo u pravilu onda ako ako su svi podatci iz uzorka (ili većina od njih) veći od  $\mu_0$ .

1. korak je kao i u (I).

2. Tu je  $t_0 = t_{\alpha}(k)$ ,  $P(t(k) > t_0) = \alpha$  (a ne  $\frac{\alpha}{2}$  kao u (I)):

3. Ako je  $t_{\text{exp}} < t_0$ , hipotezu prihvaćamo, inače je odbacujemo.

Dakle, područje prihvatanja je  $<-\infty, t_0>$ , a odbacivanja  $<t_0, +\infty>$ .

Ovo je primjer **jednostrukog** testa (područje odbacivanja je od jednoga dijela).

(III).

$$H_0: \mu = \mu_0$$

$$H_a: \mu < \mu_0$$

Tu hipotezu koristimo u pravilu onda ako ako su svi podatci iz uzorka (ili većina od njih) manji od  $\mu_0$ .

Postupak je sličan onome iz (II), samo što je područje prihvatanja  $<-\infty, t_0>$ .

## Testiranje hipoteze $\mu_1 = \mu_2$ (t-test).

Tom testu u pravilu predhodi F-test. Nakon što taj prođe nastavlja se s t-testom (testiranju očekivanja), tj. s testiranjem hipoteze:

$$H_0: \mu_1 = \mu_2 \text{ (nulta hipoteza)}$$

Hipoteza se, primjenom t-testa, provodi se slično kao kod  $\mu = \mu_0$  (razlika je samo u prvom koraku).

1. Izračuna se:

$$t_{\text{exp}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1 - 1) s_1^2 + (n_2 - 1) s_2^2}{n_1 + n_2 - 2}} \sqrt{\frac{n_1 + n_2}{n_1 n_2}}}$$

gdje obično označavamo:  $s_d = \sqrt{\frac{(n_1 - 1) s_1^2 + (n_2 - 1) s_2^2}{n_1 + n_2 - 2}} \sqrt{\frac{n_1 + n_2}{n_1 n_2}}$

2. Odredi se broj stupnjeva slobode  $k=n_1+n_2-2$ .

3. Prihvati se neki nivo signifikantnosti  $\alpha$  (obično  $\alpha=0.05$ , ali može i  $\alpha=0.01$  ili  $\alpha=0.1$ )  
Smisao nivoa signifikantnosti u testiranju je, kao i inače, sljedeći:  
 $P(\text{Postavljena se hipoteza odbacuje} | \text{postavljena je hipoteza istinita}) = \alpha$ .
4. Iz tablica t-razdiobe izračuna se kritična vrijednost pomoću koje određujemo upada li izračunata vrijednost  $t_{exp}$  u kritično područje. Kritična vrijednost ovisi o nivou signifikantnosti  $\alpha$ , o broju stupnjeva slobode (dakle o broju mjerenja), ali i o našoj kontrahipotezi koja može biti:
  - a)  $\mu_1 \neq \mu_2$  (kad testiramo jesu li te dvije veličine jednake ili različite). Tada kritična vrijednost  $t_0$  ima značenje:  $P(|t|>t_0) = \alpha$ , gdje  $t$  označava Studentovu (t-razdiobu). Hipotezu prihvaćamo ako je  $|t_{exp}|<t_0$  (inače je odbacujemo).  
Ako izričito drukčije ne kažemo uvijek smatramo da je kontrahipoteza takva.
  - b)  $\mu_1 > \mu_2$  (koja ima smisla samo ako je  $\bar{x}_1 > \bar{x}_2$ , iako se može provoditi i inače). Tada kritična vrijednost  $t_0$  ima značenje:  $P(t>t_0) = \alpha$  ( $t_0$  je drukčiji od onog iz a)). Hipotezu prihvaćamo ako je  $t_{exp}<t_0$ , inače je odbacujemo.
  - c)  $\mu_1 < \mu_2$  (koja ima smisla samo ako je  $\bar{x}_1 < \bar{x}_2$ , iako se može provoditi i inače). Tada kritična vrijednost  $t_0$  također ima značenje:  $P(t>t_0) = \alpha$ . Hipotezu prihvaćamo ako je  $t_{exp} > -t_0$ , inače je odbacujemo.

## $\chi^2$ - test.

Rezultate mjerenja slučajne varijable zapišemo u tablicu tako da u gornji redak stavljamo postignute rezultate podijeljene u  $L$  razreda: nulti, prvi,...,(L-1)-ti, a u donji frekvencije  $f_i$  tih razreda.

Iz pretpostavke o teoretskoj distribuciji izračunaju se pripadne teoretske frekvencije (u lekciji je to pokazano za Poissonovu distribuciju).

Hipoteza je da se podatci ravnaju prema teoretskoj distribuciji.

Postupak se provodi ovako:

1. Računanje broja *hikvadrat eksperimentalno* koji je **mjera udaljenosti** eksperimentalnih i teoretskih frekvencija.

$$\chi_{exp}^2 := \frac{(f_0 - f_{t0})^2}{f_{t0}} + \frac{(f_1 - f_{t1})^2}{f_{t1}} + \dots + \frac{(f_{L-1} - f_{t,L-1})^2}{f_{t,L-1}}$$

2. Određivanje broja stupnjeva slobode:  $k=L-1-l$

gdje je  $l$  broj parametara teoretske razdiobe (za Poissonovu i eksponencijalnu  $l=1$ , za normalnu i binomnu  $l=2$ ), i nivoa signifikantnosti  $\alpha$  (u pravilu  $\alpha=0.05$ ).

3. Određivanje kritične vrijednosti  $\chi_{\alpha}^2(k)$  koja ima značenje

$$P(\chi^2(k) > \chi_{\alpha}^2(k)) = \alpha,$$

gdje je  $\chi^2(k)$  *hikvadrat razdioba* s  $k$  stupnjeva slobode (to je **test-statistika**).

4. Hipotezu prihvaćamo ako je  $\chi_{exp}^2 < \chi_{\alpha}^2(k)$

(tada smatramo da udaljenost između eksperimentalnih i teoretskih podataka nije prevelika), inače je odbacujemo.

Dakle područje prihvatanja (kritično područje) je  $\langle 0, \chi^2_\alpha(k) \rangle$ , a područje odbacivanja  $\langle \chi^2_\alpha(k), +\infty \rangle$ .

**Općenito kod testiranja imamo ove nazive:**

**Pogrješka prve vrste:**  $\alpha := P(\text{Hipotezu odbacujemo} \mid \text{Hipoteza je istinita})$ .

**Pogrješke druge vrste:**  $\beta := P(\text{Hipotezu prihvaćamo} \mid \text{Hipoteza je lažna})$ .

**Jakost testa:**  $1 - \beta$ .

## Metoda najmanjih kvadrata i koeficijent regresije

Ako smo mjerenjem dviju zavisnih veličina, za prvu od njih – veličinu  $x$ , dobili podatke

$x_1, x_2, \dots, x_n$ ,

a za drugu, veličinu  $y$ , korespondirajuće podatke

$y_1, y_2, \dots, y_n$ ,

onda te podatke možemo shvatiti kao  $n$  uređenih parova:

$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$

koje geometrijski možemo predočiti kao  $n$  točaka ravnine.

Tada među svim pravcima s jednadžbom  $y = ax + b$ ,

najbolje ovim podacima odgovara onaj s parametrima

$$a = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \cdot \sum x_i^2 - (\sum x_i)^2}, \quad b = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n \cdot \sum x_i^2 - (\sum x_i)^2}.$$

Dobiveni pravac s jednadžbom  $y = ax + b$  zove se **regresijski pravac**.

Geometrijski to znači da regresijski pravac *najmanje odstupa* od početnih točaka. Ti su se parametri dobili metodom najmanjih kvadrata koja se zasniva na načelu da

*suma kvadrata razlika eksperimentalnih i teoretskih podataka bude minimalna.*

Više o tome ima u lekciji.

Ako su točke  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  grupirane oko regresijskog pravca, onda govorimo da su podatci **korelirani (linearno korelirani)**. Na osnovi toga govori se da su pripadne veličine  $x, y$  korelirane. Razina koreliranosti mjeri se **koeficijentom korelacije**

$$r := \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \cdot \sqrt{n \sum y_i^2 - (\sum y_i)^2}}$$

Taj je broj između -1 i 1. Ako je  $r$  blizu 1, to je visoka pozitivna, a ako je blizu -1 to je visoka negativna koreliranost. Ako je, pak,  $r$  blizu nule koreliranost je vrlo niska.