



Osnovne statistične metode in statistična pismenost

Žan Železnik
SURS

Osnovni pojmi v statistiki

- **Populacija** – množica vseh proučevanih elementov; pomembna je natančna opredelitev: stvarna (koga ali kaj preučujemo), prostorska (kje?) in časovna (kdaj?)
 - Primer: vsa podjetja v gradbeni dejavnosti registrirana v Sloveniji
- **Enota** – posamezni proučevani element
 - Primer: gradbeno podjetje registrirano v Sloveniji
- **Vzorec** – podmnožica oz. del populacije, na osnovi katere sklepamo o lastnosti cele populacije; **slučajni oz. naključni vzorec** – izbira enot v vzorec je naključna.
 - Primer: naključni vzorec 300 podjetij
- **Spremenljivka** – lastnost enote;
 - Primer: število zaposlenih, prihodek, standardna klasifikacija dejavnosti (SKD)
- Vsaka spremenljivka ima določeno **vrednost**, glede na to to ločimo:
 - **Opisne (atributivne, kvalitativne)** spremenljivke – vrednosti se izražajo z opisi (besede, znaki, števili)
 - **Številске (numerične, kvantitativne)** spremenljivke – vrednosti se izražajo s števili, s katerimi lahko računamo
 - **Diskretne / nezvezne** spremenljivke: zavzame le določene vrednosti iz nekega intervala, navadno cela števila
 - **Zvezne** spremenljivke: zavzame vsako vrednost na nekem intervalu

Merske lestvice (vrste spremenljivk)

Merske lestvice opredelimo glede na urejenost vrednostne spremenljivke:

- **imenska (nominalna) spremenljivka:** vrednosti spremenljivke nimajo nobene urejenosti, lahko jih le razlikujemo med seboj
 - Primer: kraj rojstva, poklic, barva oči
 - Posebna vrsta so **dvojiške (binarne, dihotomne) spremenljivke**, ki imajo samo po dve vrednosti. Primer: spol (moški, ženski), zaposlen (da, ne)
- **urejenostna (ordinalna, vrstna) spremenljivka:** vrednosti spremenljivke lahko uredimo od najmanjše do največje
 - Primer: stopnja izobrazbe, strinjanje z vrednostmi (sploh se ne strinjam, ne strinjam se, niti niti, strinjam se, zelo se strinjam)
- **razmična (intervalna) spremenljivka:** izraža se številsko, vrednosti pa lahko odštevamo, ne pa tudi delimo.
 - Primer: temperatura zraka, leto rojstva
- **razmernostna spremenljivka:** številske vrednosti, razmerje poljubnih dveh njenih vrednosti je smiselno; lahko povemo, kolikokrat večja/manjša je vrednost neke enote od druge enote
 - Primer: bruto plača, letna količina padavin

Opisna statistika

- **Opisna statistika je prvi korak analize podatkov**, ko so bili ti zbrani in urejeni.
- **Skupina statističnih metod**, ki se ukvarjajo s povzemanjem pridobljenih podatkov.
- Temeljne opisne statistike sestavlja **grafični opis**, **tabelarni opis** in **statistični povzetki**, ki na osnovi nekaterih računov predstavijo bistvene podatke (npr. srednje vrednosti, mere razpršenosti ...).

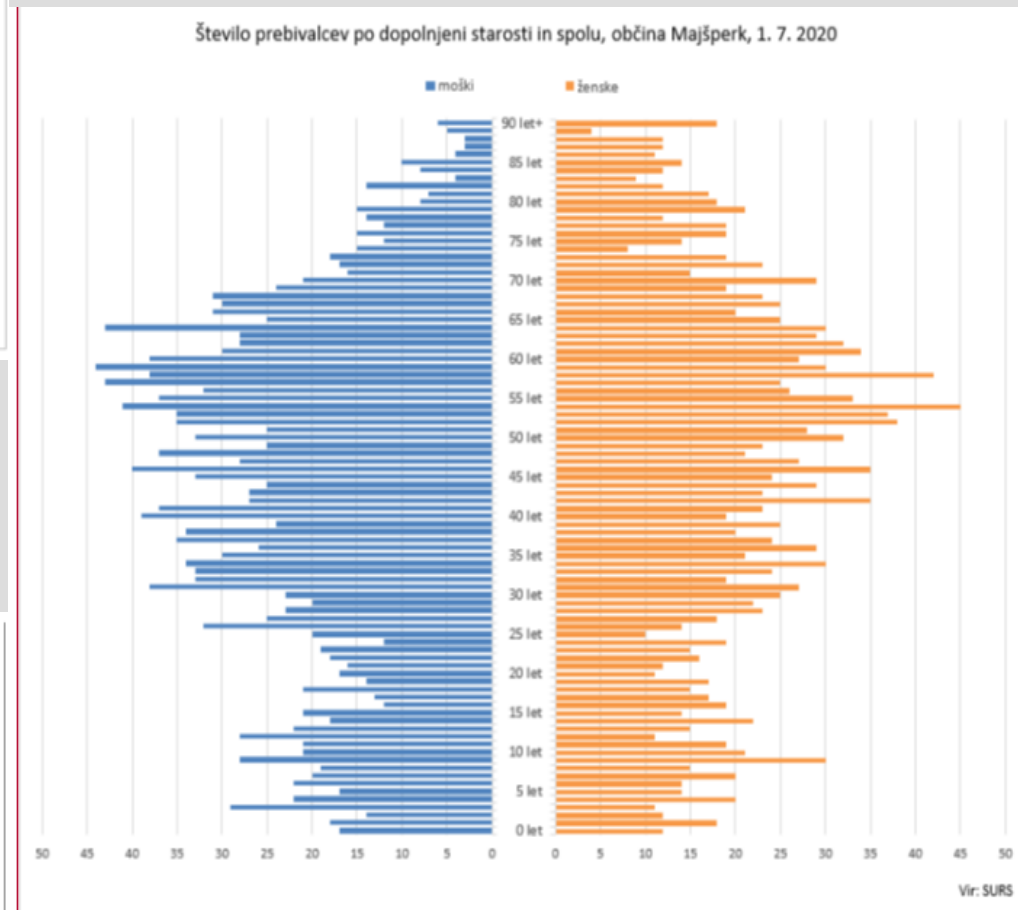
Primerjava podatkov

- Podatki postanejo zanimivi takrat, ko jih med seboj **primerjamo**
- Števila lahko primerjamo **absolutno**, na primer z razliko, ali pa **relativno**, na primer z razmerjem
- **Relativna števila** so tista, ki primerjajo dva podatka z njunim razmerjem. Glede na to, v kakšni **vsebinski povezavi** sta podatka, imamo tri vrste relativnih števil:
 - **Strukture**
 - **Koeficienti**
 - **Indeksi**

Absolutna števila

Frekvence, vsote

- Histogram
- Starostna piramida
- Tabelarični prikaz

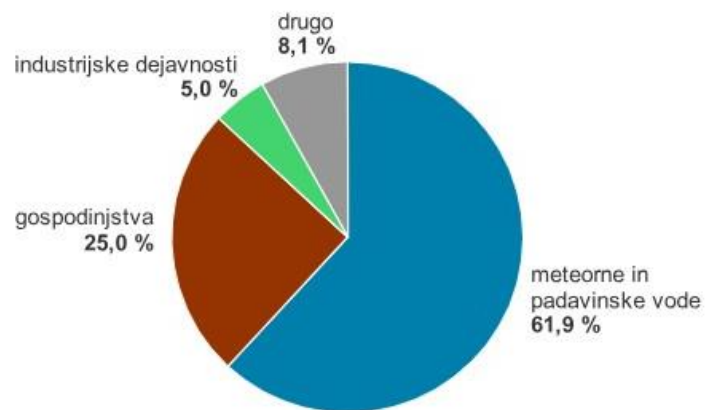


		V času absolventa bom šel na potovanje		
		Da	Ne	SKUPAJ
Sem absolvent	Da	90	29	119
	Ne	14	17	31
	SKUPAJ	104	46	150

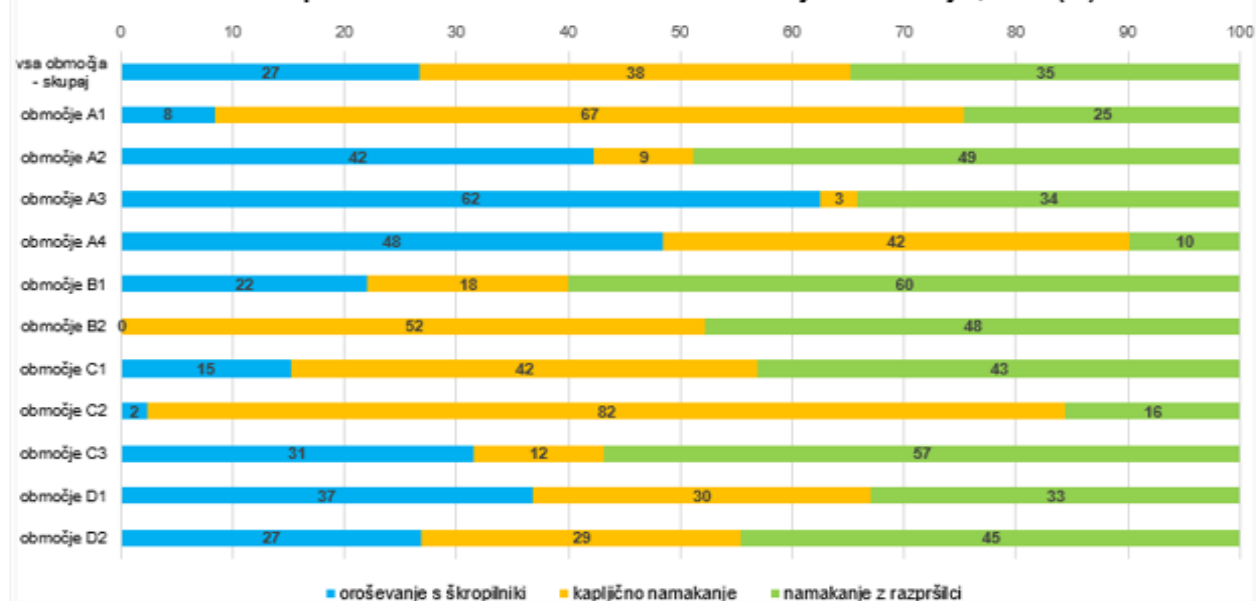
Relativna števila Strukture

- Deleži (0 – 1) ali odstotki (0 % - 100 %)
- Govorimo o *strukturi* po spolu, *strukturi* po smeri študija ...
- Strukturni krog ali stolpčni diagram

Odpadne vode po viru onesnaževanja, Slovenija, 2018



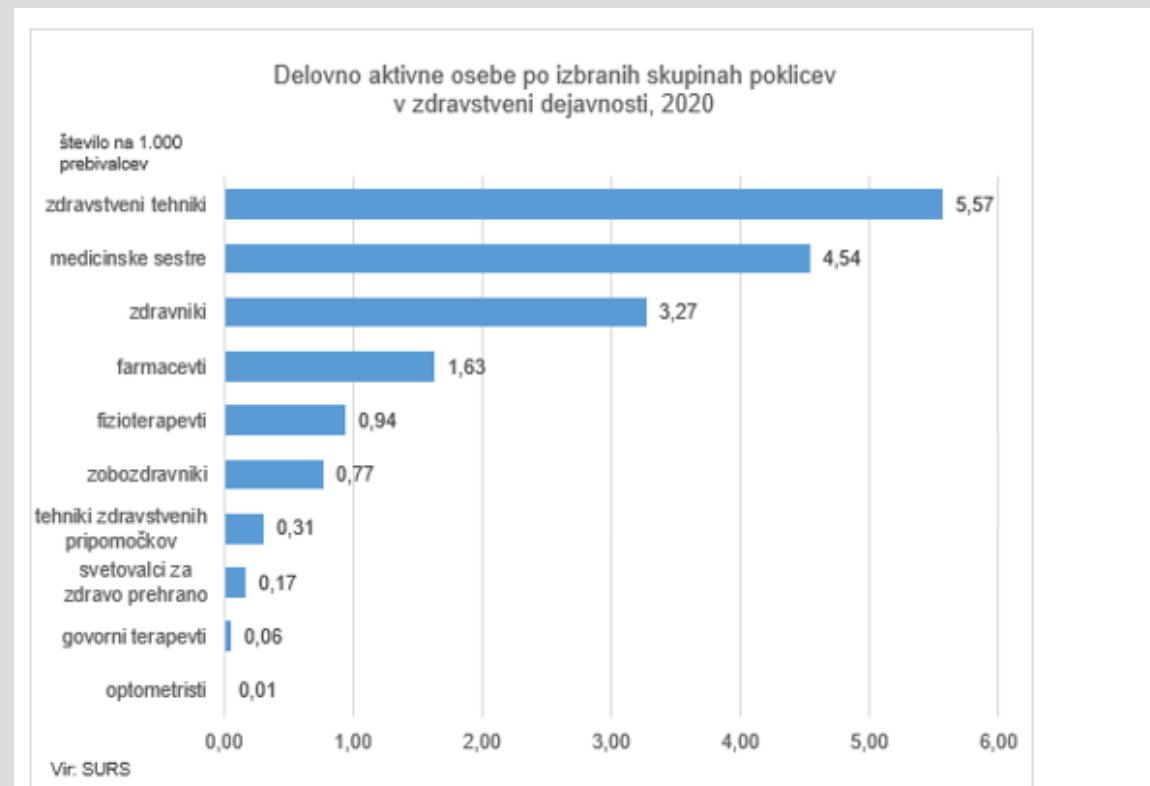
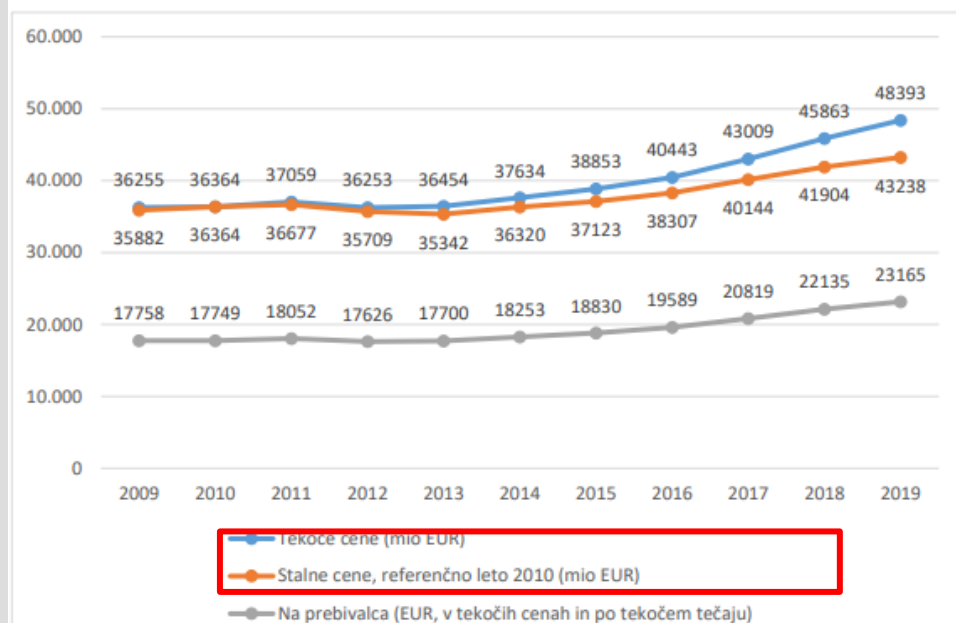
Tipi namakalnih sistemov na izbranih kmetijskih območjih, 2015 (%)



Relativna števila Koeficienti

- Razmerje dveh podatkov, ki sta vsebinsko povezana: BDP *na* prebivalca, število bolnikov *na* zdravnika ...)
- Posebne vrste koeficienti so stopnje (demografija, medicina). Govorimo o *stopnji* rodnosti:
 - število živorojenih v vsem letu / št. prebivalcev (v 1000)

Spodnji grafikon prikazuje gibanje bruto domačega proizvoda (BDP) v Sloveniji od leta 2009 do leta 2019.



Relativna števila

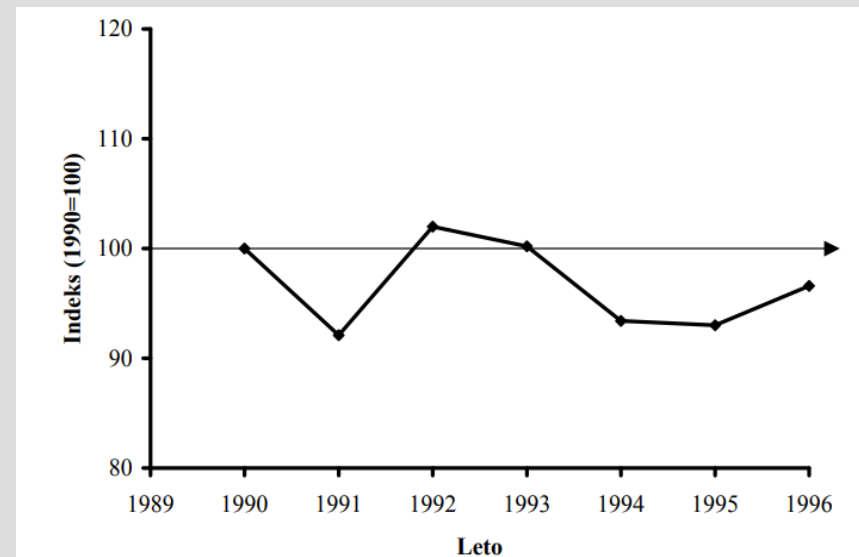
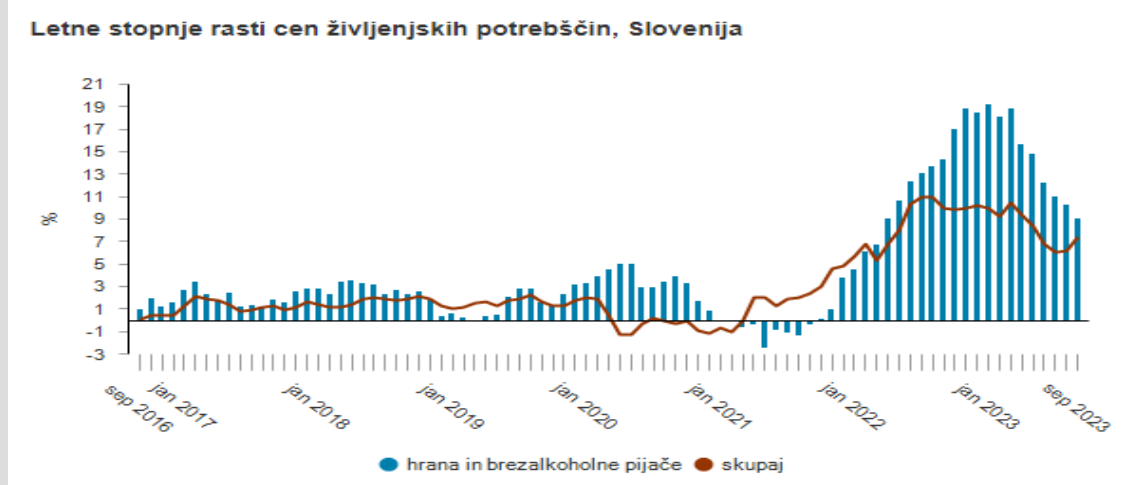
Indeksi

- Statistična vrsta (krajevna, stvarna, časovna): $x_1, x_2, \dots, x_n$; x_0 je osnova
- Indeks je primerjava podatkov statistične vrste z osnovo (%):

$$I_{i/0} = \frac{x_i}{x_0} * 100, \quad i = 1, 2, \dots, n$$

$$I_{i/i-1} = \frac{x_i}{x_{i-1}} * 100, \quad i = 1, 2, \dots, n$$

- Indeks 120,5 pomeni, da je podatek za 20,5% večji od osnove, indeks 85,0 pomeni, da je indeks 15% manjši od osnove
- Časovne vrste
 - Primer: število brezposelnih po letih
 - Indeksi s stalno osnovo (bazni indeks)
 - Indeksi s premično osnovo (verižni indeks)



Slika 2-9: Indeksi z osnovo 1990 za število študentov v Sloveniji po letih v obdobju 1990-1996

Opisne statistike

Kvantili

- X je številska spremenljivka, njene vrednosti so $x_1, x_2, \dots, x_n$.
- Vrednosti **uredimo od najmanjše do največje vrednosti** → **ranžirna vrsta**.
- **Rang R** je zaporedno mesto v ranžirni vrsti → $R = 1, 2, \dots, n$
- Relativni rang **P** izražamo v deležih ($0 < P < 1$):

$$P = \frac{R-0,5}{n} \rightarrow R = P * n + 0,5$$

- **Kvantil** razdeli ranžirno vrsto na enake dele
- Glede na to, na kolikšne dele razdeli kvantil ranžirno vrsto, ločimo: **mediano** (2 dela), **kvartile** (4 dele), **decile** (10 delov), **percentile** (100 delov)

Opisne statistike

Kvantili, minimum, maksimum

- **Mediana (Me)** – razdeli ranžirno vrsto na polovico

$$P(Me) = 0,5$$

$$R(Me) = 0,5 * n + 0,5$$

- Če je število zapisov sodo, je mediana povprečje dveh vrednosti na sredini ranžirne vrste.

- **Kvartili (Q_1, Q_2, Q_3)** – razdeli ranžirno vrsto na 4 enake dele

$$P(Q_1) = 0,25$$

$$R(Q_1) = 0,25 * n + 0,5$$

$$P(Q_2) = 0,5$$

$$R(Q_2) = 0,5 * n + 0,5$$

$$P(Q_3) = 0,75$$

$$R(Q_3) = 0,75 * n + 0,5$$

Mediana

- **Kvartilni razmik Q** je razlika med 1. in 3. kvartilom (vsebuje 50 % podatkov).

- **Minimum (min)** - najmanjša vrednost v r. vrsti

- **Maksimum (max)** - najvišja vrednost v r. vrsti

- **Variacijski razmik VR** = $max - min$

Skupina planincev se je odpravila na daljšo turo. Pred odhodom so stehali pripravljene nahrbtnike. Dva sta tehtala po 15 kg, dva po 10 kg in dva po 12 kg, pet po 14 kg, trije po 13 kg, eden 16 kg in eden 11 kg. Kolikšna je mediana?

Rešitev:

Ranžirna vrsta nahrbtnikov, urejenih od najlažjega do najtežjega:

Rang (R)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Teža (kg)	10	10	11	12	12	13	13	13	14	14	14	14	14	15	15	16

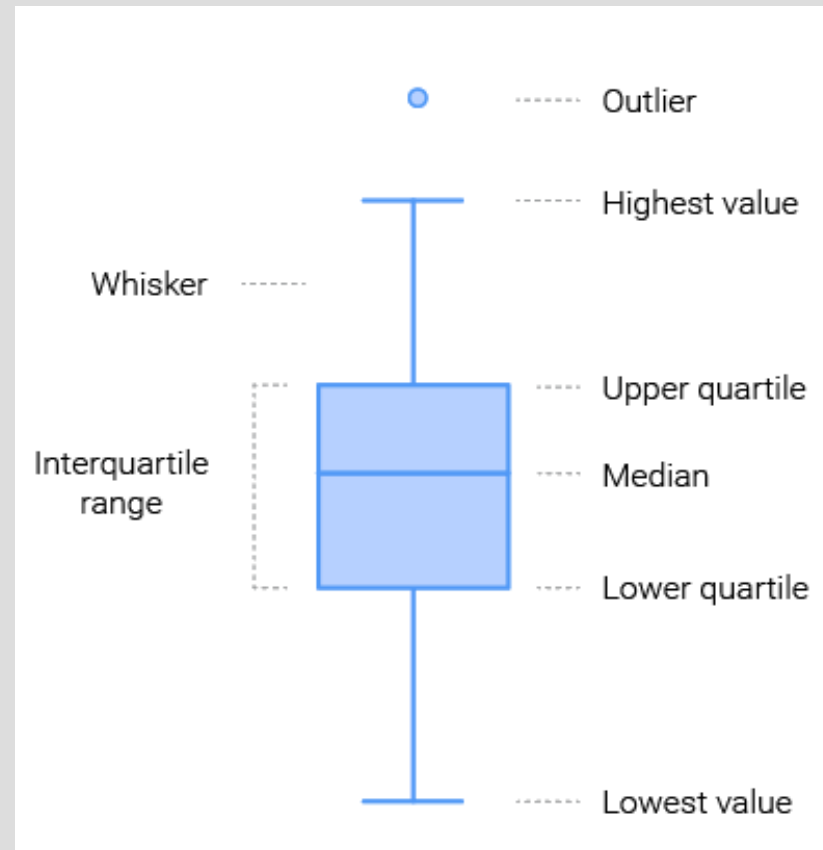
$$R = 0,5 * 16 + 0,5 = 8,5$$

$$Me = 0,5 * (13 + 14) = 13,5$$

Opisne statistike

Okvir z ročaji (škatla z brki, diagram kvartilov)

- Zelo nazoren grafični prikaz podatkov številske spremenljivke
- Potrebujemo podatke: $Q_1, Q_2, Q_3, \min^*, \max^*, Q$
- **Osamelec** je vrednost, ki **bistveno odstopa** od večine ostalih vrednosti. To so vrednosti, ki nastopajo izven intervala $[Q_1 - 1,5 * Q, Q_3 + 1,5 * Q]$. V škatli z brki jih označimo s krožcem (°).
 - **Ekstremni osamelec** je vrednost, ki nastopa izven intervala $[Q_1 - 3 * Q, Q_3 + 3 * Q]$. V škatli z brki jih označimo z zvezdico (*).
 - **Pogojni $\min^*/\max^*$** je **najmanjša / največja** vrednost, ki **še ni** osamelec.



Opisne statistike

Srednje vrednosti

- **Modus** (*Mo*) je najpogostejša vrednost spremenljivke, torej vrednost z največjo frekvenco.
- **Povprečje (aritmetična sredina)** $\bar{x}$ številske spremenljivke z vrednostmi $x_1, x_2, \dots, x_n$ je

$$\bar{x} = \frac{1}{n} * \sum_{i=1}^n x_i$$

- **Utežena (tehtana) aritmetična sredina** se izračuna za frekvenčne porazdelitve oziroma ko imamo utežene podatke:

$$\bar{x} = \frac{1}{n} * \sum_{i=1}^n f_i * x_i \quad \text{oz.} \quad \bar{x} = \frac{\sum_{i=1}^n w_i * x_i}{\sum_{i=1}^n w_i}$$

kjer je x_i sredina posameznega razreda frekvenčne porazdelitve in f_i frekvenca posamezne vrednosti, w_i pa utež posamezne vrednosti

- **Mediana (središčnica)**

Izračunajmo povprečno število točk pri izpitu iz statistike za primer 200 študentov:

Tabela 2-29: Študenti po številu točk pri izpitu

Število točk	Število študentov	Sp. meja	Zg. meja	Sredina	Širina
51-60	34	50,5	60,5	55,5	10
61-70	105	60,5	70,5	65,5	10
71-80	35	70,5	80,5	75,5	10
81-90	25	80,5	90,5	85,5	10
91-100	1	90,5	100,5	95,5	10
Skupaj	200				

$$\bar{x} = \frac{1}{200} (34 \cdot 55,5 + 105 \cdot 65,5 + \dots + 1 \cdot 95,5) = 68,2$$

Opisne statistike

Mere variabilnosti

- **Varianca** je povprečen kvadriran odklon vrednosti od aritmetične sredine:

$$\sigma^2 = \frac{1}{n} * \sum_{i=1}^n (x_i - \bar{x})^2$$

- **Varianca frekvenčne porazdelitve** je

$$\sigma^2 = \frac{1}{n} * \sum_{i=1}^n f_i * (x_i - \bar{x})^2,$$

kjer je x_i sredina posameznega razreda frekvenčne porazdelitve in f_i frekvenca posamezne vrednosti.

- **Standardni odklon:** $\sigma = \sqrt{\sigma^2}$
- **Koeficient variacije** meri, kakšen odstotek aritmetične sredine predstavlja standardni odklon:

$$\text{KV}\% = \frac{\sigma}{\bar{x}} * 100$$

Porazdelitvena funkcija

- Porazdelitvena funkcija:

$$F(x) = P(X \leq x)$$

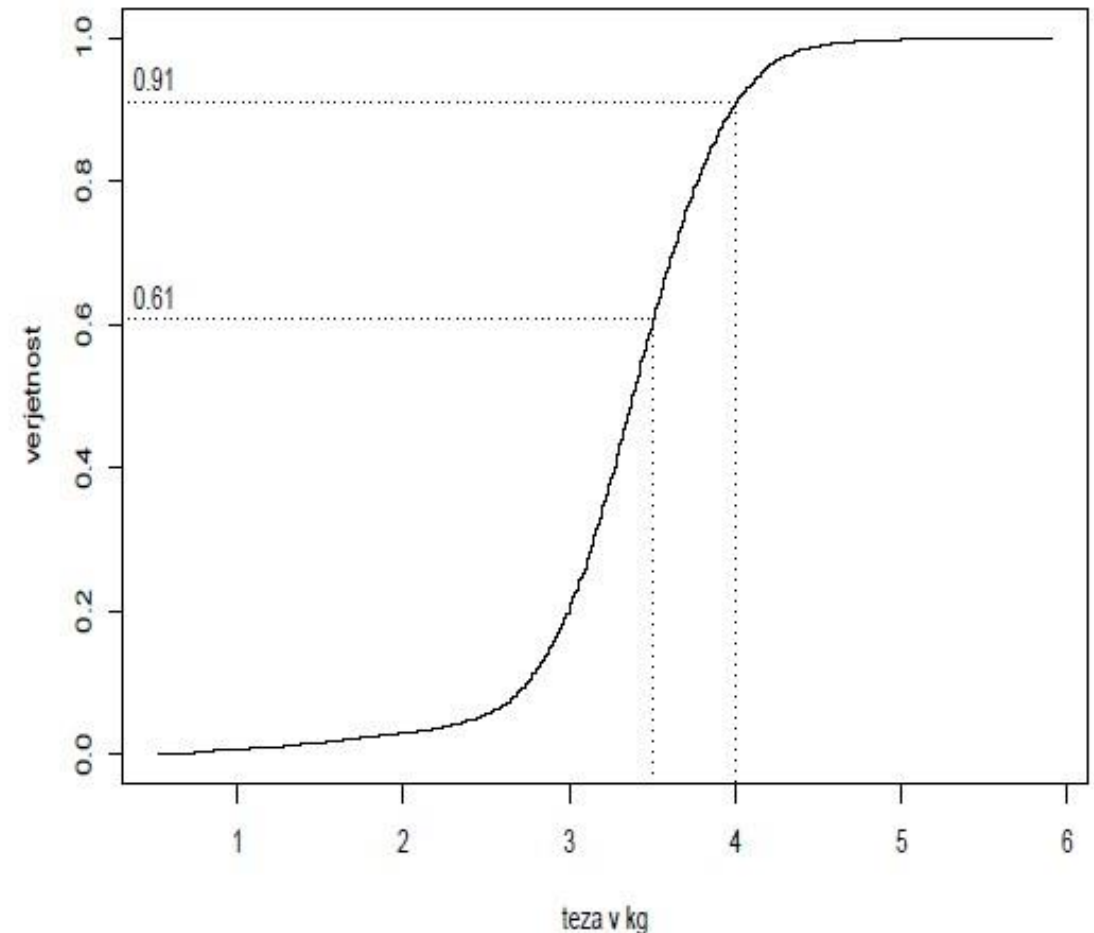
Pri dani vrednosti x je $F(x)$ torej **verjetnost, da so vrednosti slučajne spremenljivke X manjše od x** . Z drugimi besedami, $F(x)$ je delež vrednosti X , ki so manjše od x .

Primer: Porazdelitvena funkcija porodne teže novorojenčkov.

Iz nje lahko razberemo, da je 61% novorojenčkov lažjih od 3,5 kg in kar 91% lažjih od 4 kg. Med 3,5 in 4 kg pa je

$$P(\text{teža} \leq 4 \text{ kg}) - P(\text{teža} \leq 3,5 \text{ kg}) = F(4) - F(3,5) = 0,91 - 0,61 = 0,3,$$

torej 30% novorojenčkov.



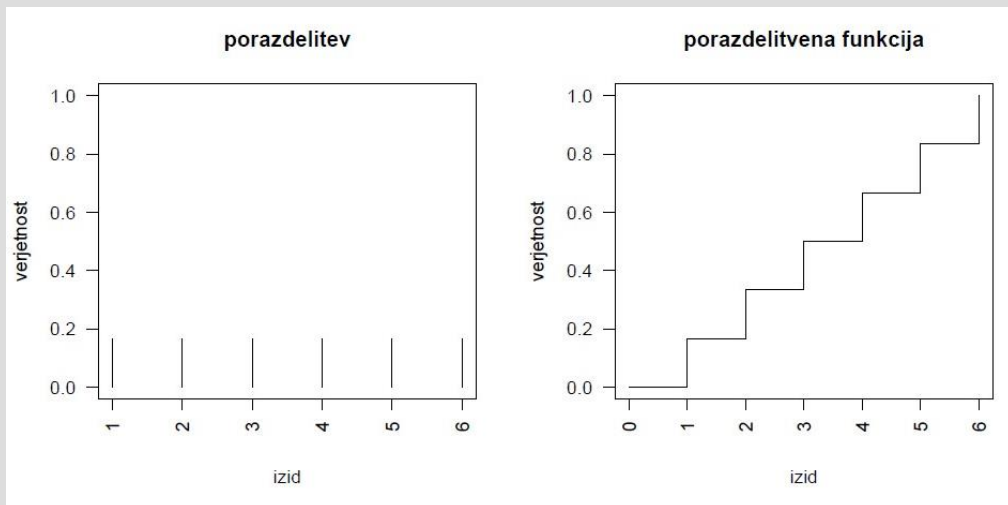
Verjetnostna porazdelitev in gostota verjetnosti

Diskretne porazdelitve

- **Diskretna slučajna spremenljivka** -> zaloga vrednosti je neka števna množica
- Po navadi lahko navedemo verjetnost pojava posameznih vrednosti -> **verjetnostna porazdelitev**
- Primer: Met kocke

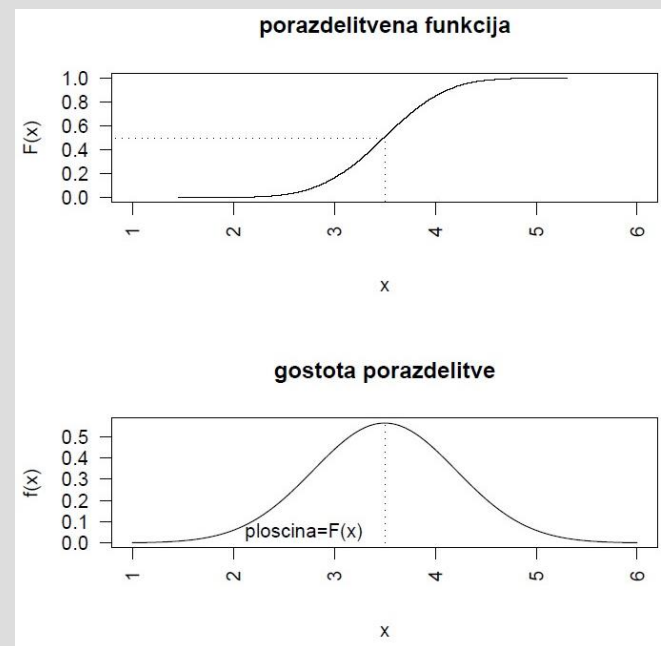
$$X \sim \begin{pmatrix} x_1 & x_2 & \cdots & x_m & \cdots \\ p_1 & p_2 & \cdots & p_m & \cdots \end{pmatrix}$$

$$X \sim \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \end{pmatrix}$$



Zvezne porazdelitve

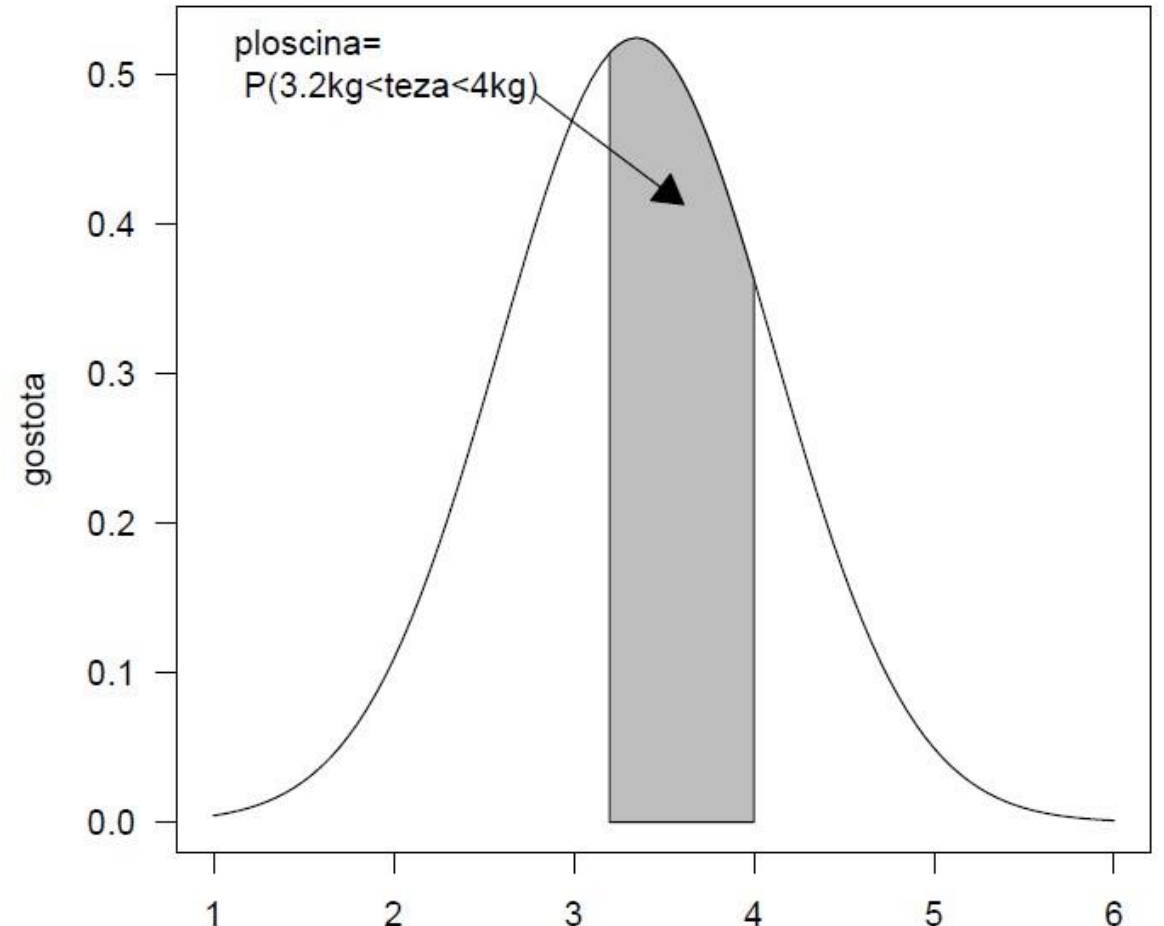
- **Zvezna slučajna spremenljivka** -> zavzema vsako realno število znotraj določenega intervala
- Ne moremo govoriti o verjetnostih posameznih vrednosti, ampak o gostoti okrog danega x -> **gostota porazdelitve**
- Primer: Porodna teža novorojencev



Zvezna porazdelitev – gostota verjetnosti

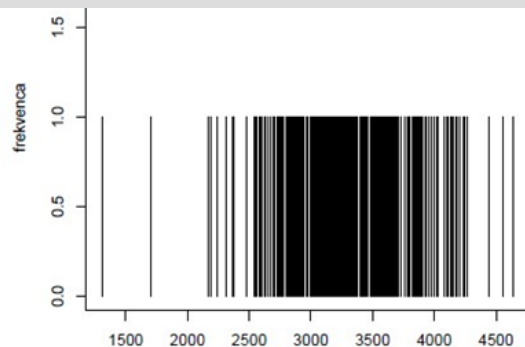
- Če je **porazdelitev zvezna**, je verjetnost, da je vrednost spremenljivke na nekem intervalu (a, b) enaka ustrezni **ploščini pod krivuljo**, ki predstavlja **gostoto porazdelitve**
- Primer: Porodna teža novorojencev

Na sliki je ilustrirana verjetnost, da je porodna teža otroka med 3,2 in 4 kg

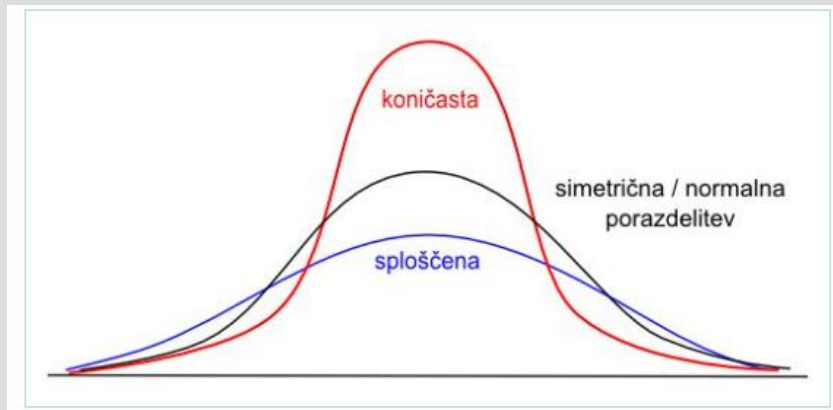
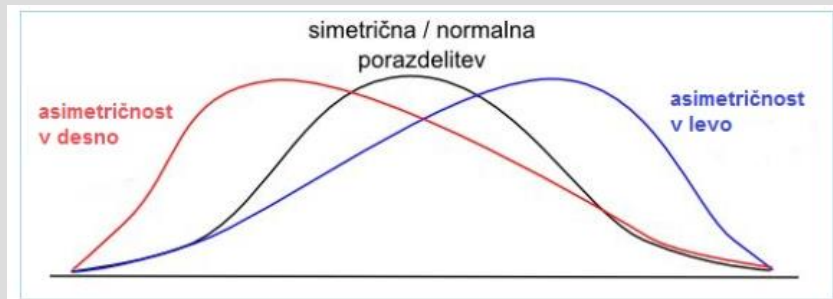
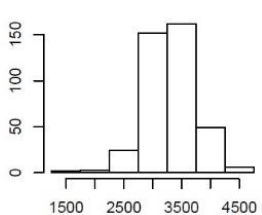
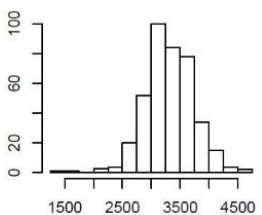
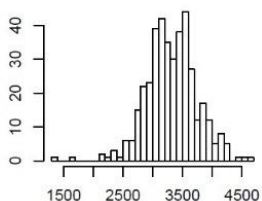
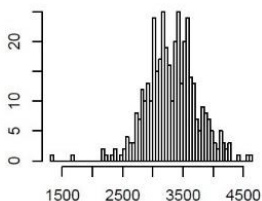


Porazdelitve slučajnih spremenljivk

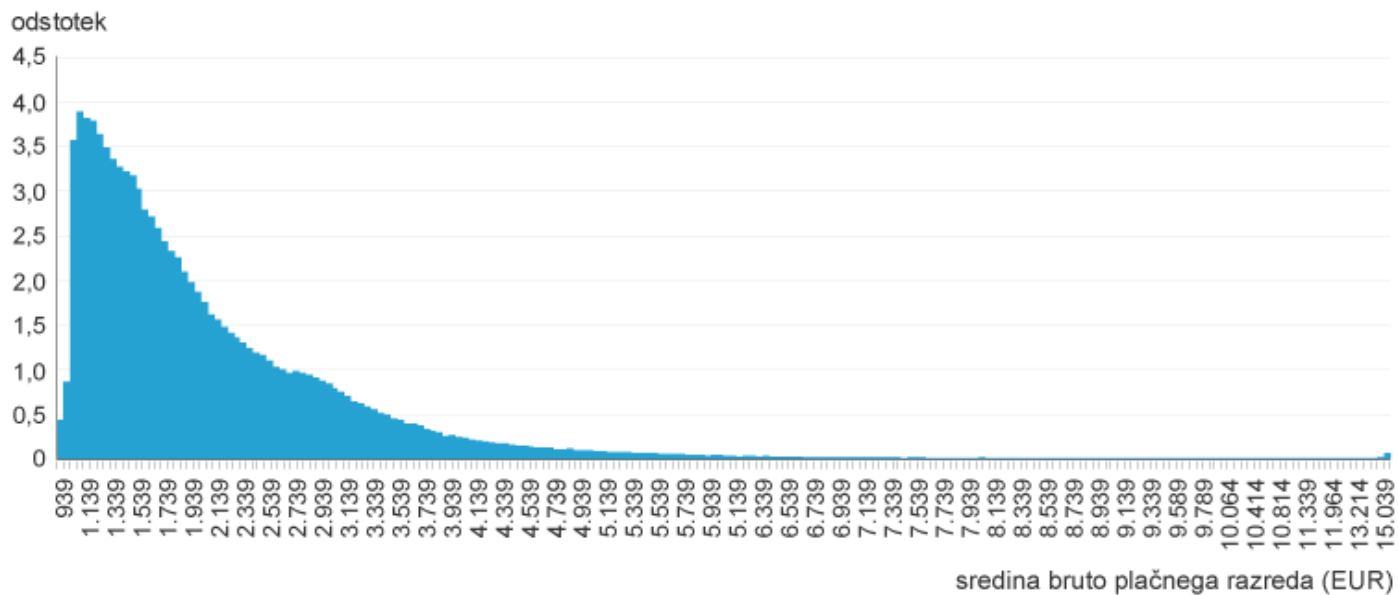
- Mere asimetrije in sploščenosti
- Unimodalna porazdelitev = 1 vrh, multimodalna = več vrhov



porodna teža

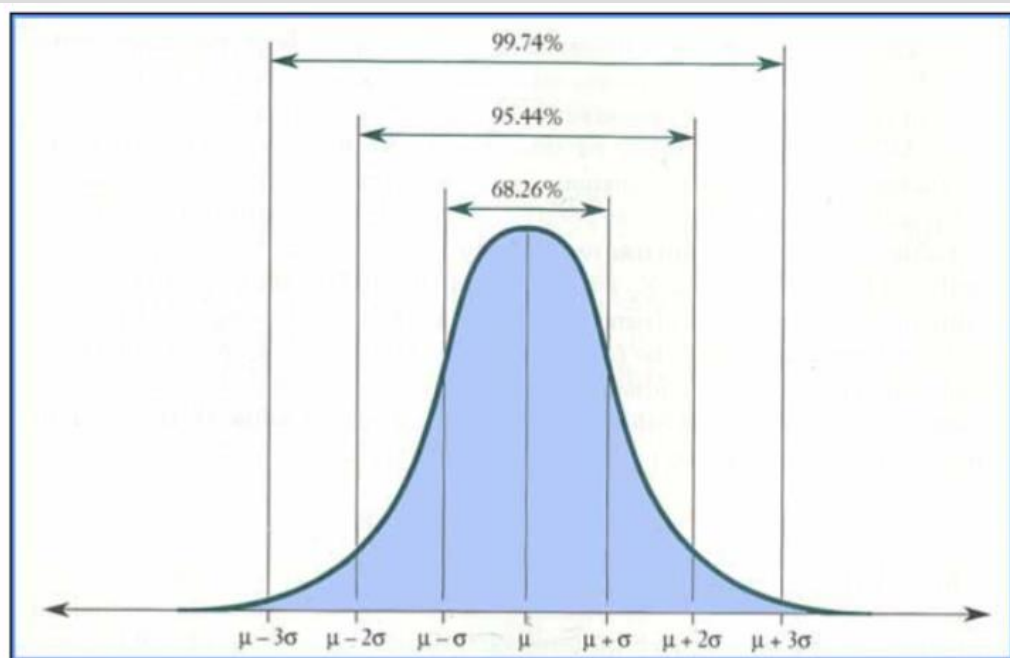
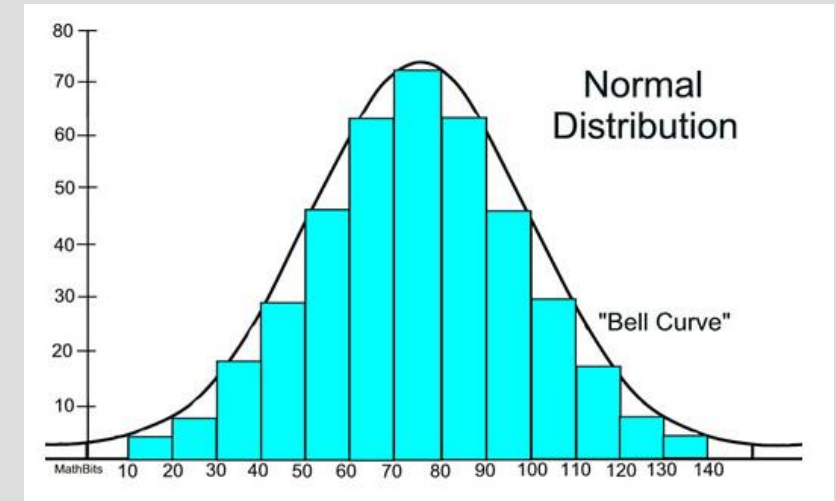


Porazdelitev povprečne mesečne bruto plače (EUR) po plačnih razredih širine 50 EUR, Slovenija, 2021¹⁾

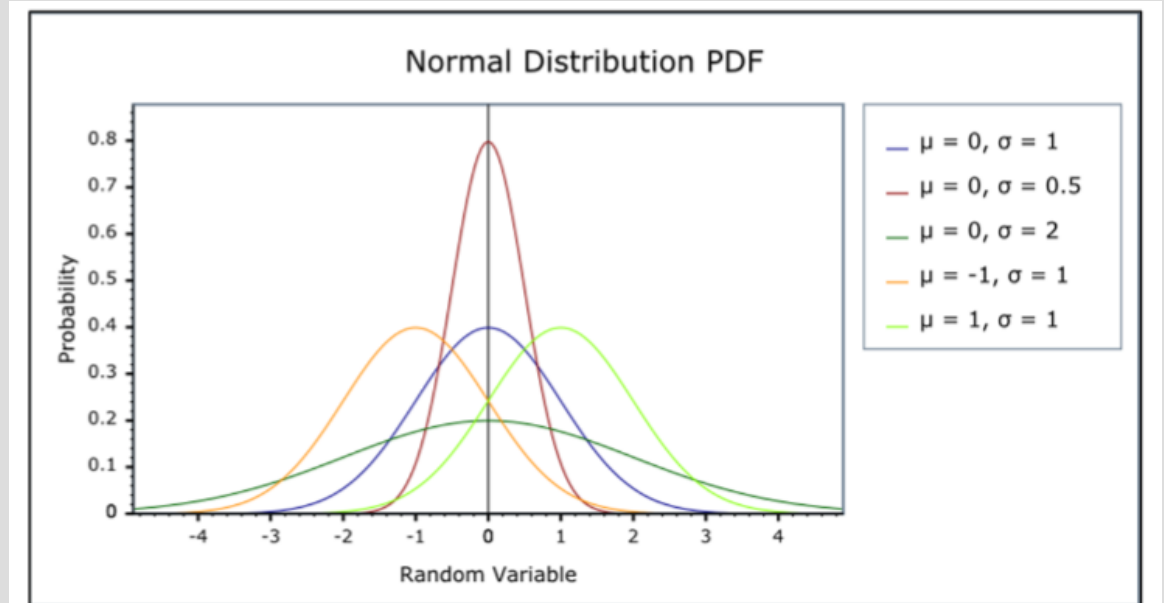


Normalna (Gaussova) porazdelitev $N(\mu, \sigma)$

- Najpomembnejša zvezna porazdelitev
- Funkcija gostote verjetnosti je krivulja, ki opiše podatke $X \sim N(\mu, \sigma)$, $\mu \in R, \sigma > 0$
- μ je **aritmetična sredina** X , σ je **standardni odklon** X
- **Simetrična** porazdelitev



Gostota verjetnosti za normalno porazdelitev



Porazdelitve slučajne spremenljivke

Standardizirana normalna porazdelitev $N(0,1)$

- Vsako normalno porazdelitev lahko prevedemo v $N(0, 1)$

Če je $X \sim N(\mu, \sigma)$, je spremenljivka

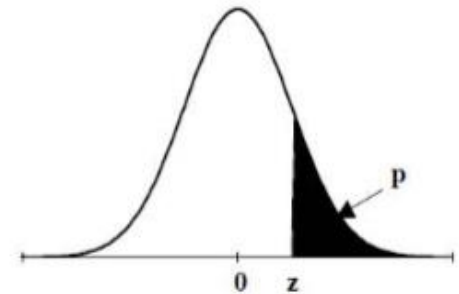
$Z = \frac{X - \mu}{\sigma}$ porazdeljena po stand. norm. porazdelitvi $N(0, 1)$, torej velja $Z \sim N(0, 1)$.

- Iz tabele za porazdelitev $N(0,1)$ preberemo verjetnost (delež vrednosti) za posamezen odklon.

TABELA 1: STANDARDIZIRANA NORMALNA PORAZDELITEV

V tabeli je za vrednost standardiziranega odklona z navedena verjetnost $p = P(Z \geq z)$.

Primer: $z = 0,93$
 $p = 0,1762$



z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641
0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
0,7	0,2420	0,2389	0,2358	0,2327	0,2296	0,2266	0,2236	0,2206	0,2177	0,2148
0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
1,0	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0869	0,0853	0,0838	0,0823
1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0721	0,0708	0,0694	0,0681

Vzorčenje

Vzorčenje je postopek, kjer izberemo vnaprej določeno število enot opazovanja iz večje populacije.

Razlogi za vzorčenje:

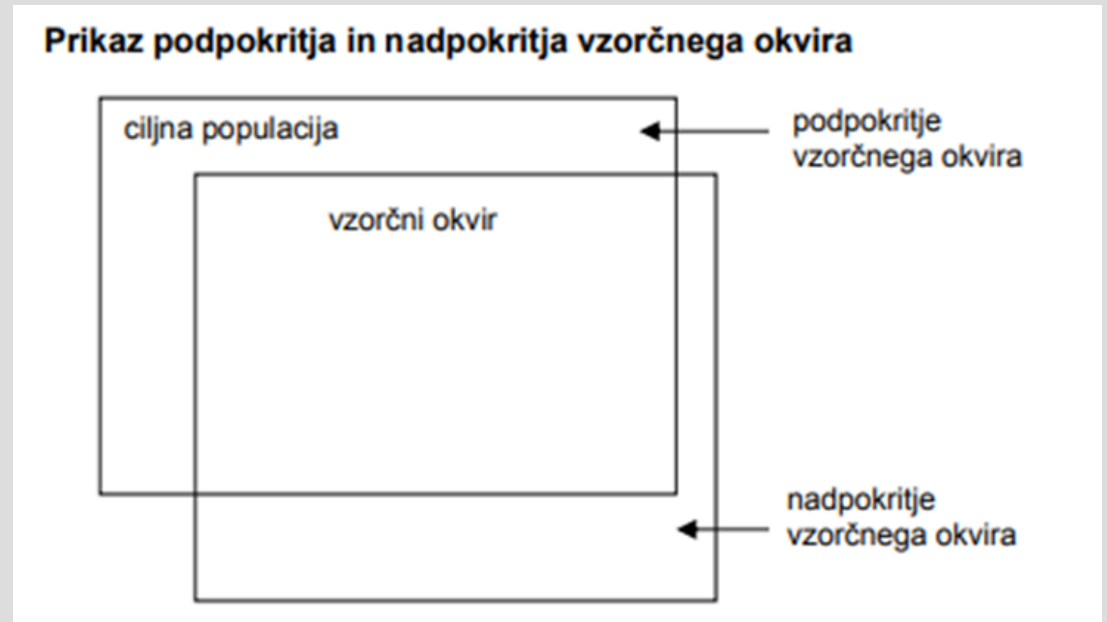
- nižji stroški
- manjša obremenitev poročevalskih enot in oseb, ki izvajajo raziskovanje
- krajše časovno obdobje izvedbe raziskovanja

Izbor enot opazovanja



Vzorčni okvir raziskovanja

- Iz splošnega vzorčnega okvir se pripravi okvir vzorčni okvir za ciljno (opazovano) populacijo.
- Vzorčni okvir sestavlja seznam enot populacije, ki jo opazujemo.
- Cilj je pripraviti vzorčni okvir z čim boljšim pokritjem.



- Vzorčni okvir raziskovanja vsebuje tudi dodatne informacije, ki so potrebne pri vzorčenju in izračunu vzorčnih ocen.

Izbor vzorca

Velikost vzorca

Alokacija (razporeditev velikosti vzorca med stratumne)

Načini izbora vzorca: verjetnostni (različne metode- SRS, sistematično, koordinirano), zajem s pragom, popolni zajem

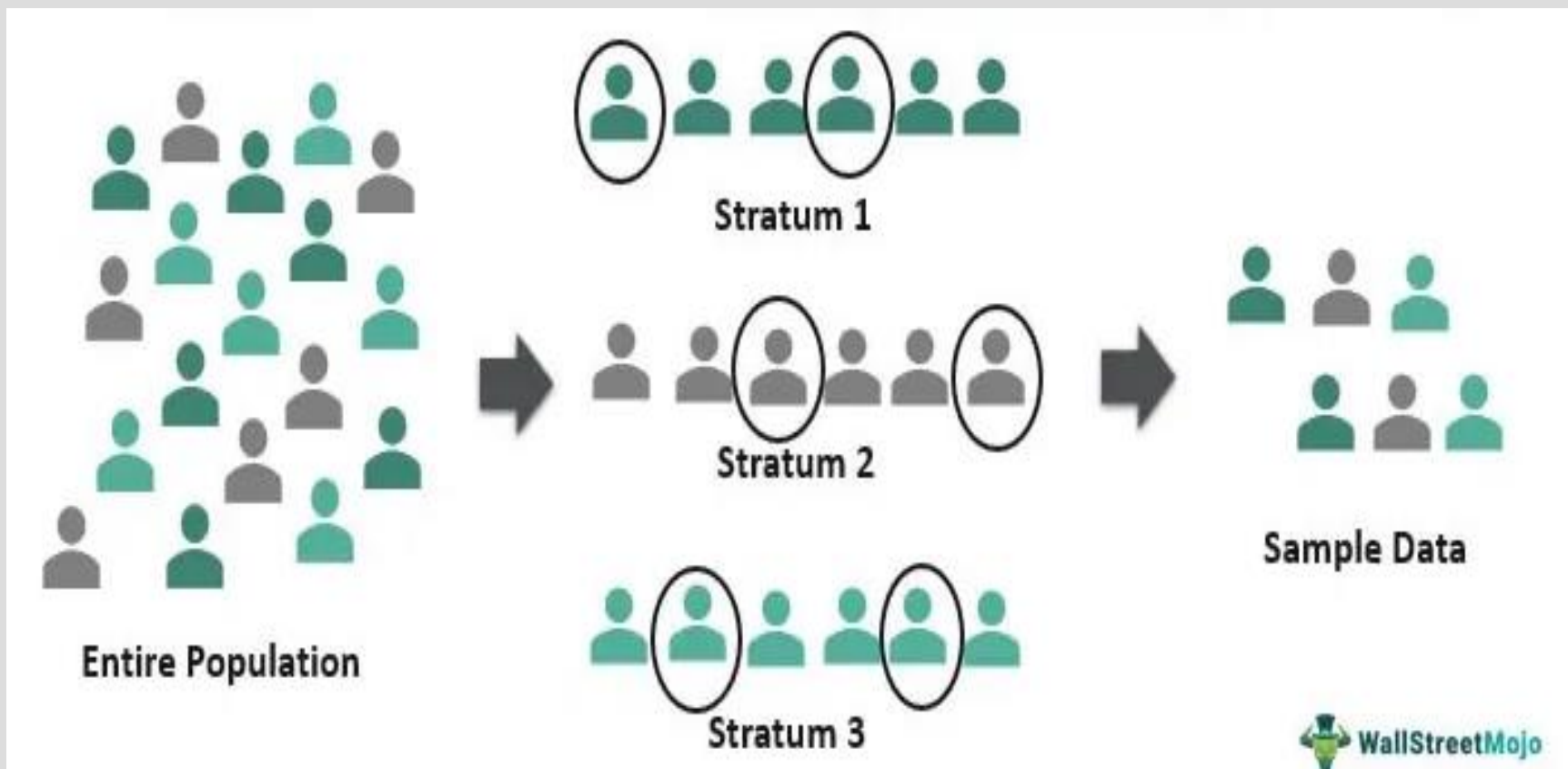
Običajno se pri poslovnih raziskovanjih uporablja **stratificiran** (enostopenjski) vzorčni načrt.

Stratumi so paroma neprekrivajoče množice podpopulacij opazovane populacije.

Primer statumskih spremenljivki:

- statistična dejavnost (npr. dvomestna statistična dejavnost podjetij) in
- velikostni razred (le ta se določi na podlagi števila zaposlenih in/ali prihodka)

Stratificiran vzorčni načrt



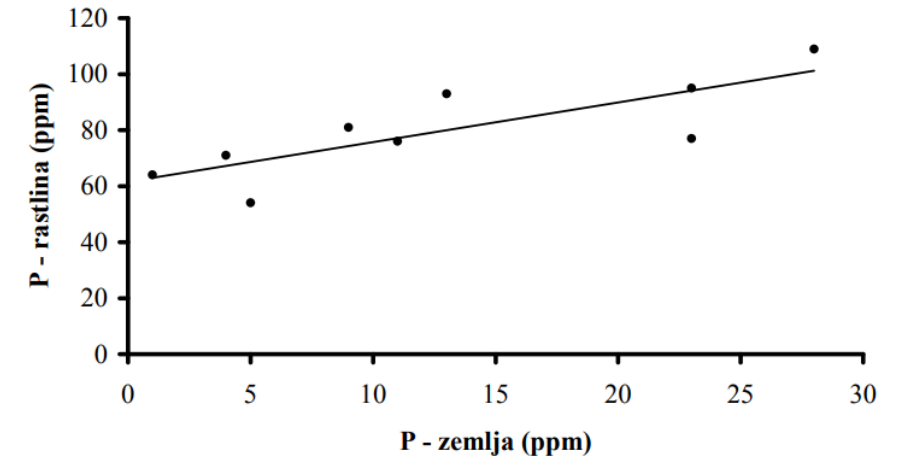
Uteževanje

- Utež izbora
- Utež zaradi neodgovora
- Razmernostna cenilka, post-stratifikacija

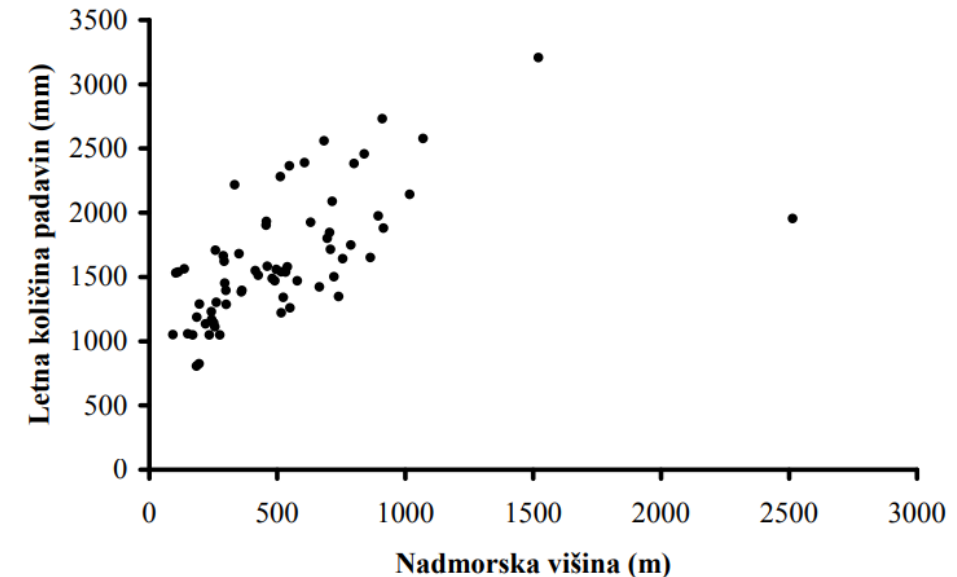
Linearna regresija

Imamo 2 spremenljivki x in y :

- y_i ($i = 1 \dots n$) so vrednosti **odvisne (odzivne) spremenljivke**
- x_i ($i = 1 \dots n$) so vrednosti **neodvisne (napovedne) spremenljivke**
- Proučujemo **vpliv neodvisne na odvisno spremenljivko**
 - Iščemo premico, ki se najbolj prilega podatkom
- Če je v modelu samo ena neodvisna spremenljivka x , ki je številska, ta model imenujemo **enostavna linearna regresija**
- **Razsevni grafikon**: na ordinati je odvisna spremenljivka, na abscisi pa neodvisna spremenljivka, vsaka enota je predstavljena z eno točko (na vsaki i -ti enoti imamo par vrednosti)
- Primeri:
 - višina in teža otroka sta **soodvisni** (povezani) spremenljivki;
 - količina padavin je odvisna od nadmorske višine, vendar nadmorska višina ni odvisna od količine padavin



Slika 6-3: Vsebnost fosforja v rastlini v odvisnosti od vsebnosti fosforja v zemlji za 9 poskusnih lončkov in linearni regresijski model, dobljen na osnovi teh podatkov



Linearna regresija

- Regresija je prilagajanje ustrezne matematične funkcije empiričnim podatkom. To funkcijo imenujemo **regresijska funkcija**. Zanima nas linearna odvisnost povprečne vrednosti odvisne spremenljivke od neodvisne spremenljivke. Ocena regresijske premice:

$$\hat{Y} = k * X + n$$

- Kako dobiti oceni parametrov k in n ? Izhajamo iz kriterija, da se premica čim bolje prilega podatkom. Najbolj pogosto uporabljena metoda je **metoda najmanjših kvadratov**.

$$k = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$n = \bar{y} - k * \bar{x}$$

- Vrednost n je presečišče premice z ordinatno osjo. Vrednost k pa predstavlja naklon premice in pomeni, če se X poveča za eno enoto, se povprečje za Y spremeni (poveča, zmanjša) za k enot.

Korelacijski koeficient in koeficient determinacije

- $X = (x_1, x_2, \dots, x_n)$; $Y = (y_1, y_2, \dots, y_n)$ sta številske spremenljivke. Ali sta povezani?

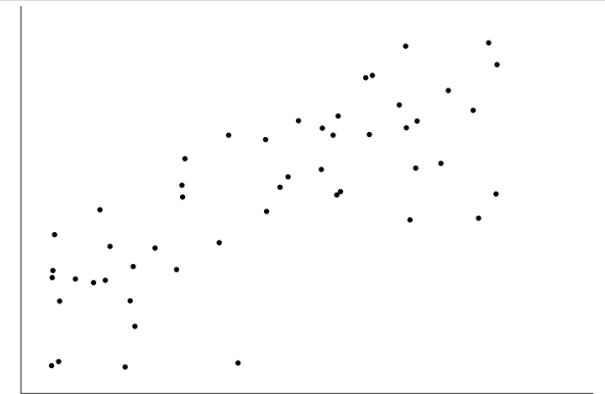
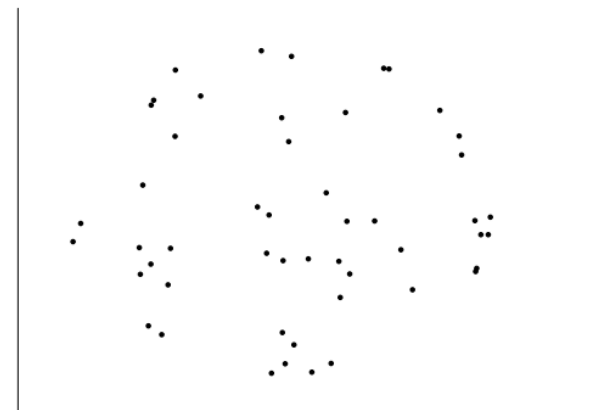
- Mera linearne povezanosti -> **korelacijski koeficient**:

$$-1 \leq r \leq 1$$

- **Pearsonov koeficient korelacije r** :

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

- **Koeficient determinacije r^2** oceni kakovost regresijskega modela
 - $0 \leq r^2 \leq 1$
 - izračuna kakšen delež variabilnosti odvisne spremenljivke je pojasnjen z modelom
 - Razmerje med pojasnjeno varianco in skupno varianco
 - izračuna se kot **kvadrat koeficienta korelacije** = r^2

Slika 6-10: Pozitivna korelacija $r=0,76$ Slika 6-11: Močna negativna korelacija, $r=-0,97$ Slika 6-12: Korelacija nič, $r=0,02$



Hvala za pozornost!

Vprašanja in komentarji na
zan.zeleznik@gov.si.