2011

Opisna statistika, predavanja prof. Podlesek

Mojca Svetek

Predavanja prof. Podlesek

Uvod v statistiko: Osnovni pojmi

Opredelitev vede

•Statistika je veda, ki se ukvarja z masovnimi pojavi.

→ velike količine kvantitativnih podatkov

•Veda o zbiranju, urejanju, opisovanju in interpretaciji podatkov.

•Matematična statistika: teoretične osnove statistike

•Uporabna statistika: aplikacije na različnih področjih, npr. v psihologiji

Populacija in vzorec

•Populacija in vzorec
- Populacijo prestavljajo vsi posamezniki, ki nas zanimajo in jih želimo preučevati
- Vzorec je del populacije, ki populacijo predstavlja v določeni raziskavi

•Parameter in statistika, napaka vzorčenja
- Parameter – populacija vseh študentov psi; vrednost, ki opiše celotno populacijo
- Statistika – vzorec študentov psi; vrednost, ki opiše vzorec

V populaciji težko preučimo (izmerimo) vse enote.

↓

Načrtovanje vzorčenja in meritev

Vzorčenje, priprava na meritve: Zbiranje podatkov

Opisna statistika: Urejanje in opisovanje (povzemanje) empiričnih podatkov

Inferenčna statistika: Posploševanje z vzorca na populacijo*

*Inferenčna statistika se ukvarja z ugotavljanjem, koliko rezultati vzorčne skupine odstopajo

Spremenljivke

1. Spremenljivka = znak, ki predstavlja različne izraženosti nekega pojava (telesna višina ima lahko različne izraženosti, če ima samo eno je konstanta)
2. Podatek = (izmerjena, zabeležena) vrednost neke spremenljivke pri nekem objektu merjenja

Oseba	Višina (cm)	Spol
A	168	Ž
B	172	Ž
C	175	Ž
Č	165	Ž
D	172	M
E	178	M
F	180	M
G	192	M
H	185	M
I	174	M

–kategorialni podatki

- dihotomni, trihotomni, politomni
- načela klasifikacije: enakost objektov v isti kategoriji, izključnost kategorij, izčrpnost kategorij

–numerični podatki

Podatek je stopnja izraženosti neke lastnosti spremenljivke, ena višina je podatek, vse višine so spremenljivka

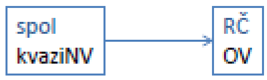
- Kategorialni podataki - kvalitativno
    - Dihotomna – dve vrednosti/skrajnosti
    - Trihotomna
    - Politomna
- Numerični podatki - kvantitativno
    - Številke prestavljajo stopnjo izraženosti pojava

Vrste raziskovalnih načrtov

- **korelacijske metode**

–Opazujemo dve ali več naravno obstoječih spremenljivk. Ali med njimi obstaja odnos? Kakšna je povezanost med dvema pojavoma? (nimamo neodvisne oz. odvisne spremenljivke)
–Korelacijska metoda nam ne da informacije o vzročno-posledičnih povezavah med dvema spremenljivkama

- **eksperimentalne metode**



–Eksperimentator spreminja neodvisno spremenljivko in opazuje spremembe v odvisni. Ostale spremenljivke ohranja konstantne. Ali neodvisna spremenljivka vpliva na odvisno? Ali med obema obstaja vzročno-posledični odnos? Kdo na koga vpliva, vzročno-posledični odnos, neodvisna spremenljivka:odvisna spremenljivka
–Cilj eksperimenta je vzpostaviti vzročno-posledično zvezo med dvema spremenljivkama
–Raziskovalec manipulira z eno spremenljivko (neodvisno spremenljivko) in pri tem opazuje drugo, da bi ugotovil, kako manipulirana spremenljivka vpliva na drugo (odvisno spremenljivko)
–Raziskovalec mora nadzorovati situacijo, saj se ne sme zgoditi, da bi na spremembo vplivala kakšna druga spremenljivka

- **kvazieksperimentalne metode**

–Opazujemo razlike v odvisni spremenljivki glede na naravno obstoječo kvazineodvisno spremenljivko; pojavljajo se sovplivajoče spremenljivke. Ali kvazineodvisna spremenljivka vpliva na odvisno?
–Uporabljamo podobne statistične metode kot pri eksperimentalnih metodah
–Imamo napol neodvisno spremenljivko in napol odvisno (spola ne moreš zamenjati pri isti osebi, taka spremenljivka je kvazineodvisna, saj v resnici eksperimentator ne more z njimi manipulirati; tudi s časom ne more manipulirati, možno je, da pacienti po 3 mesecih terapije napredujejo zaradi nje same ali pa je to zgolj posledica časa, za to ne more suvereno sklepati na učinek terapije same)

Število spremenljivk v analizi

1. **Univariatna analiza** = analiza posamezne spremenljivke: njenega povprečja, razpršenosti, porazdelitve
2. **Bivariatna analiza** = analiza odnosa med dvema spremenljivkama (korelacija), napovedovanje vrednosti Y na podlagi X (regresija, napovedovanje?)
3. **Multivariatna analiza** = analiza odnosa med več spremenljivkami

Merske ravni numeričnih spremenljivk (NOIR + A)

1. Nominalna → kategorialne, kvalitativne mere
- EMŠO
- 1 - moški, 2 - ženske (1 ni več od 2)
2. Ordinalna
- mesto na tekmovanju (razlika med 1. in 2. in 2. in 3. ni nujno enaka)
- pogostost: 1 - nikoli, 2 - občasno, 3 - pogosto, 4 – vedno
- starostne kategorije: 15-20 let, 20-30 let, 31-40 let, več kot 40 let
3. Intervalna
- temperatura (25°C, 77°F) (lahko razporedimo po velikosti, vmes so enake razdalje)
- IQ
4. Razmernostna
- telesna višina (176 cm)
- najpogostejše v naravoslovnih znanostih
5. Absolutna

Najnižja raven številčnih podatkov je **nominalna raven** (1 – moški, 0 – ženski) [nom=ime] to ni prava številčna vrednost, ampak oznaka kategorije.

Druga merska raven – **ordinalna** – že pove nekaj o urejenosti števil (1. mesto na tekmovanju, 2., 3., 4.) – ne vemo pa koliko je drugi slabši od prvega, razdalje med temi številkami lahko niso enake in jih ne moremo kar tako seštevati

Tretja raven – **intervalna** – nastopi, ko je razdalja med posameznimi vrednosti (npr. temperatura, dolžina, IQ …) enaka. Med posameznimi vrednostmi so natančno enake razdalje, zato je z njimi mogoče tudi računati.

- ➥ **0** je **arbitrarna vrednost**, nekdo ki ima IQ 100 ni 2x pametnejši od nekoga z IQ 50 – zato ne moremo tvorit pravih razmerij (ni absolutne ničle)

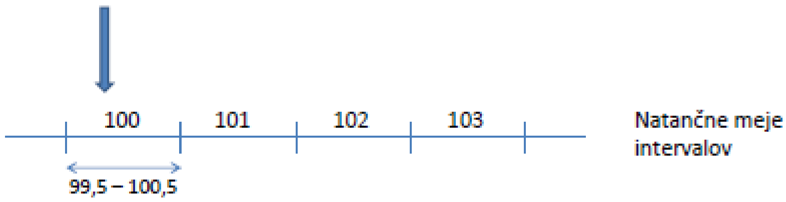
Razmernostna raven ima vse lastnosti intervalne, a poleg tega še **absolutno ničlo**. To pomeni, da vrednost 0 pomeni odsotnost merjene spremenljivke. Prednost te ravni je, da lahko merimo prava razmerja med spremenljivkami.

- ➥ 0 cm pomeni odsotnost dolžine, medtem ko 40 cm pomeni dvakrat manj od 80 cm

Absolutna raven – pri štetju, dva elementa sta 2x več kot 1, je hkrati tudi razmernostna

Zveznost spremenljivk

- Diskretne spremenljivke - npr. število otrok ali pa številka parkirnega mesta
- Zvezne spremenljivke - med dvema številoma je neskončno drugih števil (npr. anksioznost, sposobnost, inteligentnost … gre za eno kontinuus izraženosti)
    - v praksi jih ne moremo neskončno fino izmeriti (odvisno od instrumenta, ki mora biti zelo natančno izbran; nekdo lahko med dvema odgovoroma DA NE koleba, mi pa ju ocenimo bodisi z 1 bodisi z 2)
    - zaokroževanje (IQ 100 pomeni, da se oseba nahaja med 99,5 in 100,5, oseba se nahaja na intervalu, ki ima natančno zgornjo in spodnjo mejo)



Pomembna mesta
Št. preb. mesta: 22.000; države: 22.000.000

Pomembna mesta

Načela določanja pomembnih mest

- Vse neničelne števke so pomembne (tudi 0, če je med dvema neničelnima); 105,43 (5 pomembnih mest)
- Vodilne ničle niso pomembne; 0,0023 (dve pomembni mesti)
- Decimalne ničle so pomembne; 120 (3 pomembna mesta) vs. 120,00 (5 pomembnih mest)
- Dvoumno pri ničlah pred decimalno vejico → 2000 (znanstvenik podrčrta zadnje pomembno mesto – v tem primeru gre na stotice natančen zapis), 2000 (2 s.f.), 2000 ± 5 % (5 % od 2000 je 100)
- Znanstveni zapis:
    - 0,00012 = 1,2 x 10^{-4}
    - 0,001200 = 1,200 x 10^{-3}
    - 1300 (4 s.f.) = 1,300 x 10^{3} ‽
    - 1300 (2 s.f.) = 1,3 x 10^{3} ‽
- Kadar imamo števila zaokrožena na celo mesto, pri računanju ne moremo navajati več pomembnih mest, edino če jih je nad 10 lahko na eno

Pomembna mesta

Zaokroževanje na pomembna mesta

- 1,2459 (3 s.f.) → 1,25
- 1,25 (2 s.f.) → 1,3 ali 1,2 (po dogovoru)

Pomembna mesta

Aritmetika

- Množenje in deljenje: Rezultat naj ima toliko pomembnih mest, kot je najmanjše število pomembnih mest (upoštevamo izmerjena števila). 25,2 x 4,5 = 1$\underline{1}$0
    - Interval ena: (25˙15, 25˙25)
    - Interval dva: (4˙45, 4˙55)
        - Oseba se lahko nahaja na čisto spodnji meji obeh intervali ali pa na zgornjih (dejanske vrednosti v danem primeru so lahko med 111,92 in 114,89, sredina je 113,4, ampak bi bil ta rezultat lahko zavajajoč zato se odločimo za 1$\underline{1}$0)
- Seštevanje in odštevanje: Rezultat naj ima toliko decimalnih mest, kot jih ima izmerjeno število z najmanjšim številom decimalnih mest. 25,22 + 4,5 = 29,7
- Pri računanju uporabljamo čim več mest, rezultat zaokrožimo na koncu.

Notacija

X_i (npr. X_1, X_2 … X_n) *

Y_{ij} *

$\sum$ vsota

$\prod$ zmnožek

N numerus (veliki *N* uporabimo, ko vključimo vse osebe, če bi bil samo njihov del, npr. ena skupina n_1, n_2 … pa z malo)

*X je neodvisna spremenljivka, Y je odvisna spremenljivka Drugi indeks po navadi pomeni skupino (Y_{22} pomeni drugo osebo v drugi skupini)

X_1 predstavlja vrednost spremenljivke pri 1. osebi

X_n predstavlja vrednost spremenljivke pri osebi *n*

$$\sum_{i=1}^{N} x_i$$

x_1=25 x_2=30 x_3=32 x_4=18

$$\sum_{i=1}^{4} x_i = x_1 + x_2 + x_3 + x_4$$

$$\prod_{i=1}^{4} x_i = x_1 x_2 x_3 x_4$$

Vrstni red operacij

1. Operacije v oklepajih
2. Kvadriranje (oz. izračun drugih potenc)

3. Množenje oz. deljenje (od leve proti desni)
4. Seštevanje znotraj Σ
5. Preostala seštevanja oz. odštevanja

Sumacijski znak

Sumacijski znak

$$\sum_{i=1}^{N} X_i = X_1 + X_2 + X_3 + \ldots + X_N$$

$$\sum_{i=1}^{N} cX_i = c\sum_{i=1}^{N} X_i$$

$$\sum_{i=1}^{N} c = Nc$$

$$\sum_{i=1}^{N} (X_i + Y_i + Z_i) = \sum_{i=1}^{N} X_i + \sum_{i=1}^{N} Y_i + \sum_{i=1}^{N} Z_i$$

$$\sum_{i=1}^{N} X_i Y_i = X_1 Y_1 + X_2 Y_2 + \ldots + X_N Y_N$$

Zgled:

$$\sum_{i=1}^{N} c = c + c + c + \cdots + c_n = N \cdot c$$

$$\sum_{i=1}^{N} (x_i + y_i + z_i)$$

Primer:

spremenljivka / oseba	**X**	**Y**	**Z**
A	1	1	1
B	1	2	1
C	3	1	4
seštevek	**5**	**4**	**6**
	$\sum_{i=1}^{N} x_i$	$\sum_{i=1}^{N} y_i$	$\sum_{i=1}^{N} z_i$

5 + 4 + 6 = $\sum_{i=1}^{N} x_i + \sum_{i=1}^{N} y_i + \sum_{i=1}^{N} z_i$ = 15

Sumacijski znak

Je pravilno ali napačno?

$$\sum_{i=1}^{N} X_i \sum_{i=1}^{N} Y_i = \sum_{i=1}^{N} X_i Y_i$$

$$\left(\sum_{i=1}^{N} X_i \right)^2 = \sum_{i=1}^{N} X_i^{\,2}$$

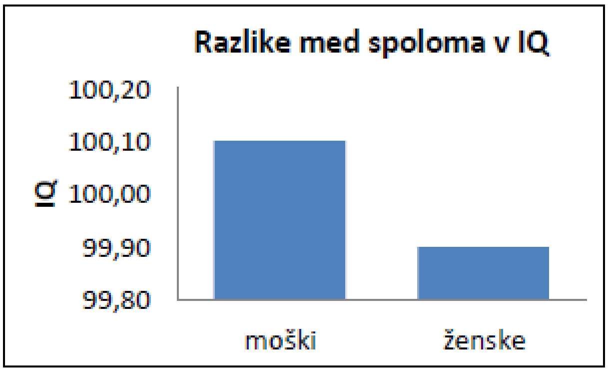
$$\sum_{i=1}^{N} (X_i + c)(X_i - c) = \sum_{i=1}^{N} X_i^{\,2} - Nc^2$$

$$\sum_{i=1}^{N} (X_i + Y_i)^2 = \sum_{i=1}^{N} X_i^2 + \sum_{i=1}^{N} Y_i^2 + 2\sum_{i=1}^{N} X_i Y_i$$

1. N
2. N
3. P
4. P

Zakaj je dobro biti statistično pismen?

- »There are three kinds of lies: lies, damned lies, and statistics« (M. Twain)
- Otroci z večjimi stopali bolje črkujejo.
- Povprečna mesečna neto plača za april 2011 je znašala 975,98 EUR.



Zajetje
V statistično raziskovanje Mesečno poročilo o izplačanih plačah pri pravnih osebah so vključene zaposlene osebe s pogodbami o zaposlitvi (če delajo po pogodbah o delu ali avtorskih pogodbah, jih ne upoštevamo). Upoštevane so vse zaposlene osebe, ne glede na to, ali delajo za določen ali nedoločen čas, polni delovni čas ali delovni čas, krajši od polnega.
Vključeni niso samostojni podjetniki posamezniki in pri njih zaposlene osebe, osebe, ki opravljajo poklicno dejavnost, osebe, ki opravljajo javna dela, napotene zaposlene osebe (detaširani delavci) in kmetje.

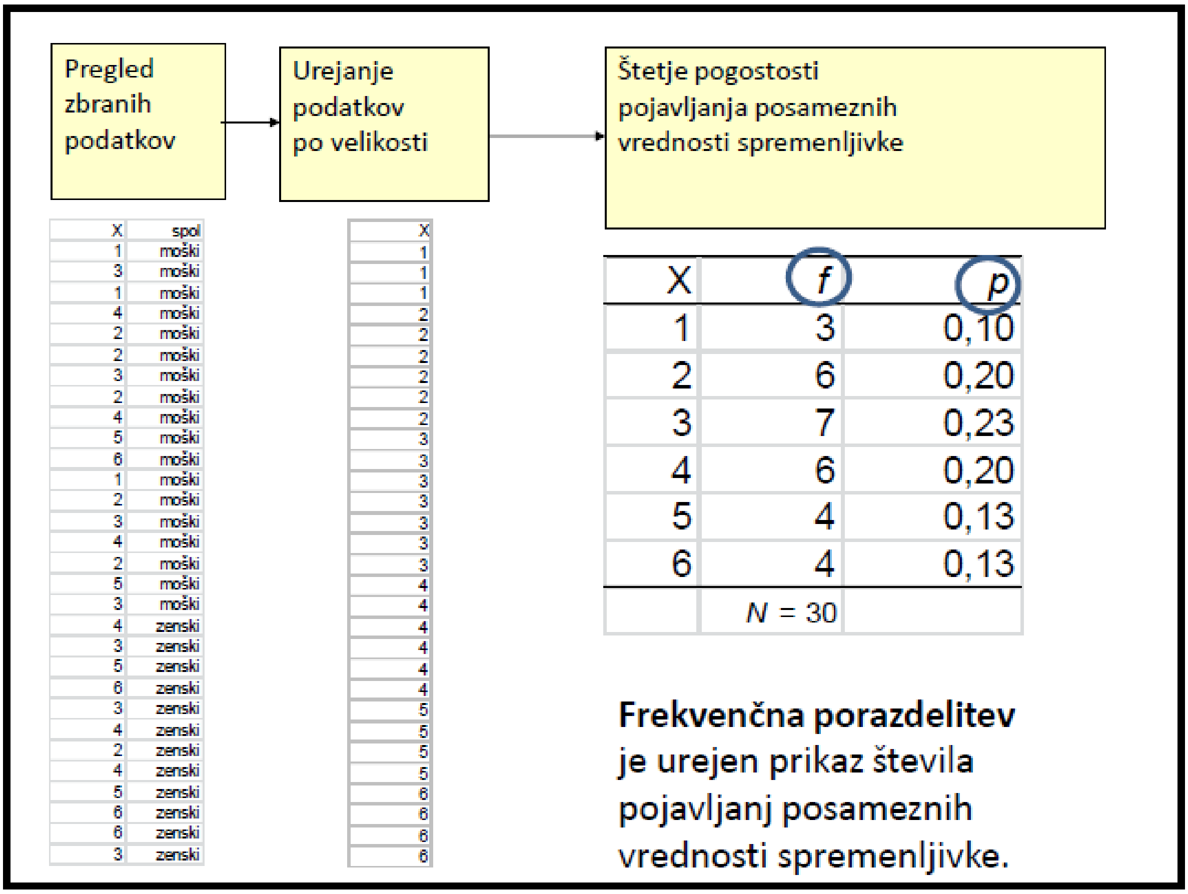
Povprečna mesečna plača je povprečni znesek, ki ga kot plačilo za mesec dela prejmejo zaposlene osebe pri pravni osebi.

Urejanje podatkov

Definicija

Urejanje podatkov in deskriptivna statistika
= opis glavnih značilnosti zbranih podatkov
= opis preučevanih značilnosti vzorca (navadno)

Intervalne spremenljivke

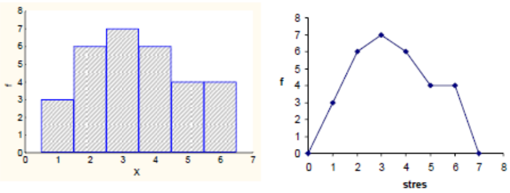


Najpogostejše spremenljivke so intervalne.
Intervalna lestvica od 1 do 6 (npr. stopnja stresa), med posameznimi številkami so enake razlike. Gre za diskretne vrednosti (ne zvezne).

X	*f (preštete frekvence)*	*p(relativna oblika, oblika deleža)* $p_i = f_i / N$	%
1	3	0,10	10
2	6	0,20	20
3	7	0,23	23
4	6	0,20	20
5	4	0,13	13
6	4	0,13	13
	N=30 (velikost vzorca)	1,00	

Grafični prikazi frekvenčne porazdelitve

Histogram   Frekvenčni poligon



Grafični prikazi zelo dobro prikazujejo podatke. Na navpični osi (ordinati) predstavimo frekvenco, na vodoravni (abcisi) pa spremenljivko. Histogram (njegova širina) pri vrednosti 1 dejansko zajema vrednosti od 0,5 do 1,5 (*I*(0˙5, 1˙5)) itd. Razmerje ordinata:abcisa je pribl. 2:3 (zlati rez). Pri frekvenčnem poligonu sta dve vrednosti še teoretično dodani (ena višja od najvišje in ena nižja od najnižje). Ploščini histograma in poligona sta enaki.

Grafični prikazi frekvenčne porazdelitve

- *Stem-and-leaf plot*

```
Y Stem-and-Leaf Plot

Frequency   Stem & Leaf

  3.00       1 . 000
  6.00       2 . 000000
  7.00       3 . 0000000
  6.00       4 . 000000
  4.00       5 . 0000
  4.00       6 . 0000

Stem width:  1
Each leaf:     1 case(s)
```

Ničle na nek način prikazujejo histogram, ničel je toliko kot je oseb, ki predstavljajo določeno vrednost. Vidimo lahko ne samo koliko oseb je pri določeni vrednosti, ampak tudi o podrobnostih posameznikih. V takem primeru morda sploh ne bi bilo nobene ničle ampak bi zapis za npr. frekvenco 3 izgledalo 142, kar bi pomenilo, da so vrednosti za prvo osebo npr. 1˙1, drugo 1˙4, tretjo pa 1˙2.

Urejanje v ranžirno vrsto

13	21	11	7	22	30	27	27	12
16	21	23	17	18	13	15	20	29
24	21	25	23	15	18	17	21	

Ranžirna vrsta
7, 11, 12, 13, 13, 15, 15, 16, 17, 17, 18, 18, 20, 21, 21, 21, 21, 22, 23, 23, 24, 25, 27, 27, 29, 30

Frekvenčna porazdelitev

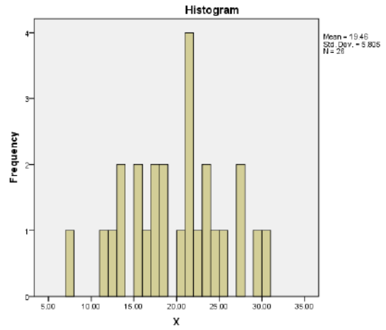
x

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	7.00	1	3.8	3.8	3.8
	11.00	1	3.8	3.8	7.7
	12.00	1	3.8	3.8	11.5
	13.00	2	7.7	7.7	19.2
	15.00	2	7.7	7.7	26.9
	16.00	1	3.8	3.8	30.8
	17.00	2	7.7	7.7	38.5
	18.00	2	7.7	7.7	46.2
	20.00	1	3.8	3.8	50.0
	21.00	4	15.4	15.4	65.4
	22.00	1	3.8	3.8	69.2
	23.00	2	7.7	7.7	76.9
	24.00	1	3.8	3.8	80.8
	25.00	1	3.8	3.8	84.6
	27.00	2	7.7	7.7	92.3
	29.00	1	3.8	3.8	96.2
	30.00	1	3.8	3.8	100.0
	Total	26	100.0	100.0	

Vrsto uredimo po velikosti in preštejemo kolikokrat se posamezne vrednosti pojavljajo.

💻 Izkoriščamo Excelovo funkcijo (f_x) »COUNTIF«(countif (A1: A27; C2) prenašamo (A2: A28; C3) oz. fiksiramo (A$1: A$27; Cx))

Frekvenčna porazdelitev



Združevanje v razrede

- arbitrarno
- izbor širine razredov in števila razredov
  - 6 do 8 razredov
  - minimalno število razredov = 3,3 log N + 1
- RI = (X_{max} – X_{min} + 1) / št.razredov
- Pojmi: SM, ZM, natančna SM, natančna ZM, sredina razreda, zaprt/odprt razredni interval
- določitev spodnje meje prvega razreda:
  - večkratnik širine razreda,
  - spodnja meja prvega razreda = min X
  - sredina prvega razreda = min X

Vrednost 7 se pojavlja enkrat, vrednosti 8 pa nimamo, zato ni stolpca pri vrednosti 8. Podatke grupiramo v skupine. Ni pravil za tvorjenje razredov. Odločimo se za širino razreda, pri grafu frekvenčne razdelitve je vsaka vrednost svoj razred, lahko pa naredimo širše razrede (npr. vsak razred je širok 3 vrednosti). Razpon razreda ni x_{min}-x_{max}, natančen razpon je še za 1 enoto večij, torej je $\boxed{x_{min}-x_{max}+1}$ (razred od 7 do 30 je v resnici natančno I(6˙5, 3˙5)). Glede na numerus lahko ocenimo, kakšna je prib. priporočljiva št. razredov: $\boxed{3{,}3 \log N + 1}$. Natančne meje naj bodo tako določene, da točno vemo, kam sodi podatek (natančne meje naj bodo za eno mesto bolj natančne od vrednosti, s katerimi se ukvarjamo – npr. če bi bila natančna ZM intervala (10, 13) 13, in SM intervala (13, 15) 13, ne bi vedeli, kam spada število 13). Natančna ZM spodnjega razreda je namreč enaka natančni SM zgornjega razreda.

Razred (širina: 3)	Natančna SM	Natančna ZM	Xj (X') (sredina razreda)	*f*	*p*
07 - 09	6,5	9,5	8,0	1	0,039
10 - 12	9,5	12,5	11,0	2	0,077
13 - 15	12,5	15,5	14,0	3	0,154
16 - 18	15,5	18,5	17,0	5	
19 - 21	18,5	21,5	20,0	5	
22 - 24	21,5	24,5	23,0	4	
25 - 27	24,5	27,5	26,0	3	
28 - 30	27,5	30,5	29,0	2	

Združevanje v razrede

Ranžima vrsta
7, 11, 12, 13, 13, 15, 15, 16, 17, 17, 18, 18, 20, 21, 21, 21, 21, 22, 23, 23, 24, 25, 27, 27, 29, 30

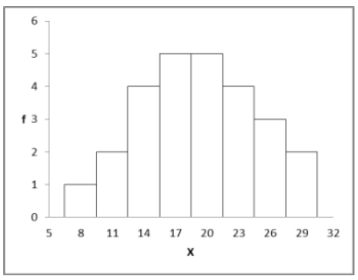
Frekvenčna tabela

razred	X'	*f*	%
7 - 9	8	1	3.9
10 - 12	11	2	7.7
13 - 15	14	4	15.4
16 - 18	17	5	19.2
19 - 21	20	5	19.2
22 - 24	23	4	15.4
25 - 27	26	3	11.5
28 - 30	29	2	7.7
		26	100.0

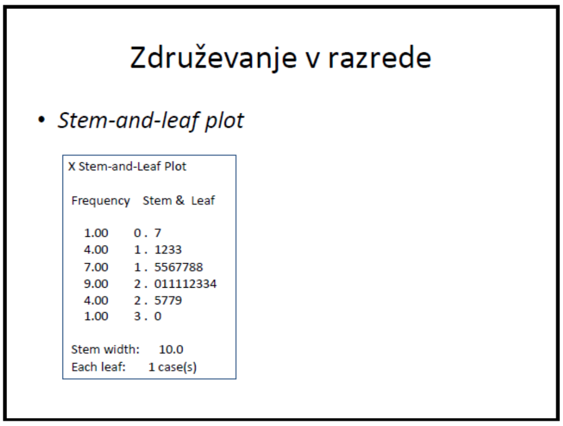
Združevanje v razrede

Histogram
↓
Frekvenčna porazdelitev

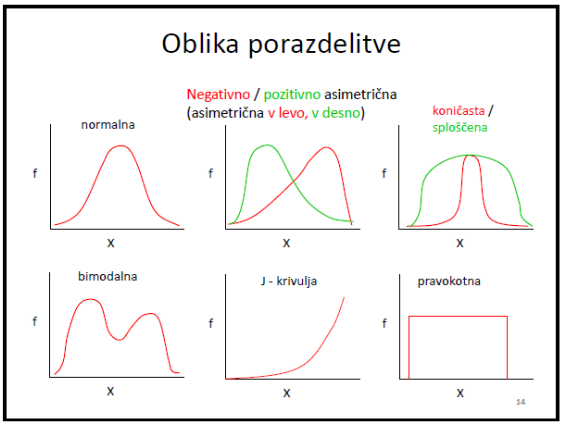
Risanje slik v skladu s standardi APA



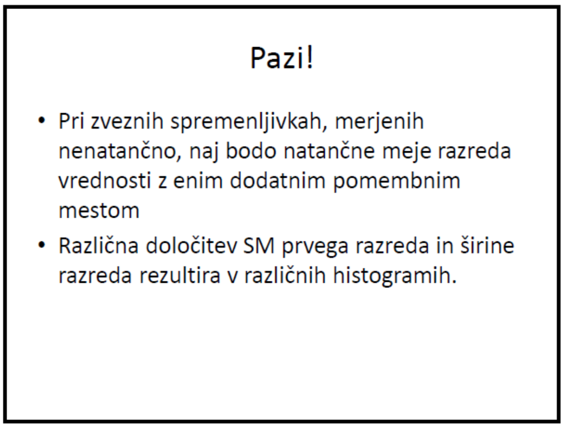
Na abcisno os lahko nanašamo natančne meje razredov ali pa sredine razredov.

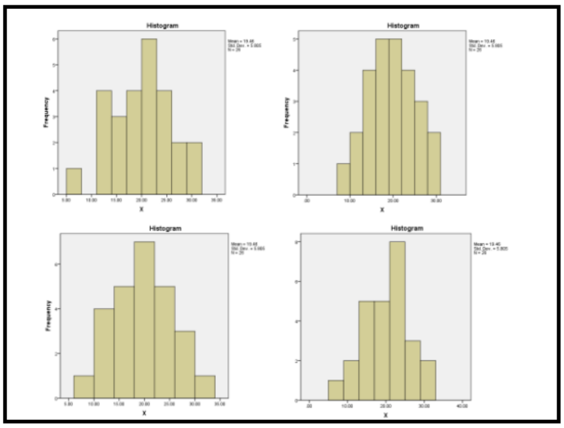


Stem (=deblo, sreblo) & *leaf* (=list) *display* je prvi predstavil J.W. Tukey. Vsako vrednost je razdelil na dva dela, npr. med stebla desetice in med liste enice. Vrednost X=15 je torej razdelil na steblo 1 in listek 5, vrednost X=20 pa na steblo 2 in listek 0. Na tak način je vrednosti pogrupiral, v opisanem primeru bi steblo 1 pomenilo interval $I_1(10, 19)$, steblo 2 pa interval $I_2(20, 29)$ itn. Na tak način lahko ustvarimo zasukan histogram, a ne izgubimo posameznih vrednosti znotraj intervalov. Lahko ustvarimo še več razredov, kot prikazuje slika, pri tem višji vključuje vrednosti 0-4, nižji pa 5-9. Ta način prikazovanja podatkov nudi seznam točnih vrednosti znotraj skupin in sliko histograma.



V psihologiji je najbolj pogosta oblika porazdelitve t.i. normalna (simetrična). Pogosta je takrat, ko je v igri ogromno različnih spremenljivk, ki jih ne moremo kontrolirati (IQ je odvisen od socialnega statusa, izobrazbe mame, prehrane v otroštvu, spola …). Pri koničasti porazdelitvi je več poszameznikov s povprečno vrednostjo (je višji!) in več tistih z ekstremnimi vrednostmi (izgleda kot da je za kakšno enoto bolj dvignen glede na graf, ki prikazuje normalno porazdelitev). Pri sploščenem pa je ravno nasprotno. Modus je najpogostejša vrednost – bimodalna porazdelitev ima 2 modusa, trimodalna ima 3 modusa (vrhove), polimodalna več ...

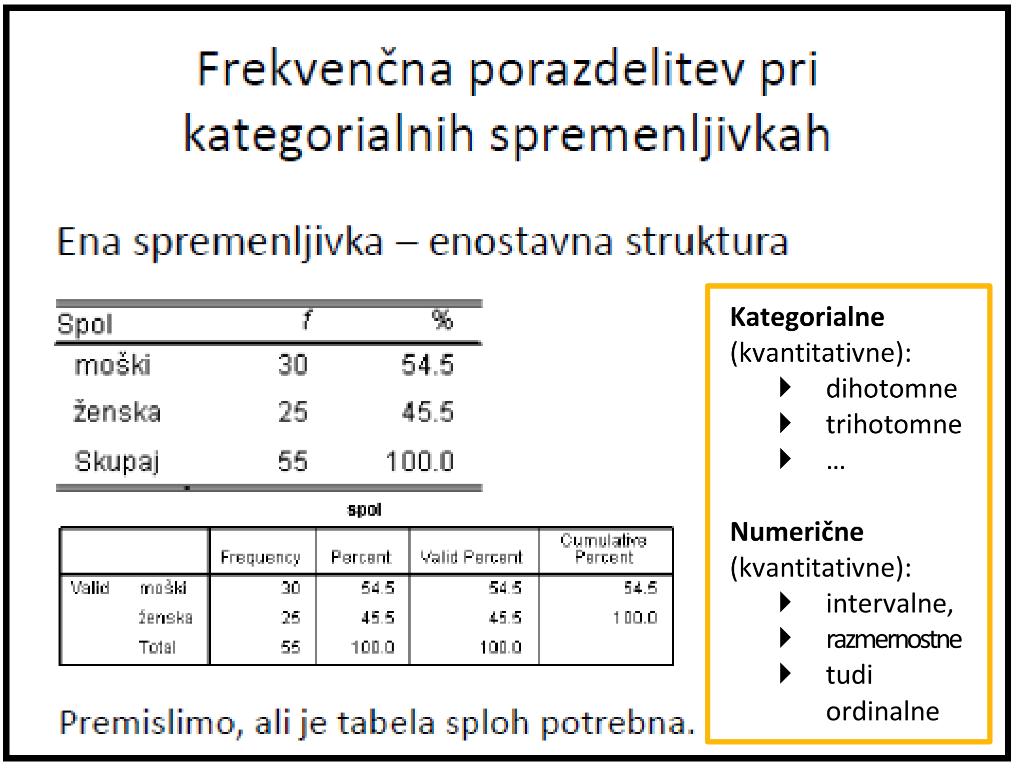




Če izberemo drugačno mejo spodnjega razreda in drugačno število in širino razredov, bodo naši histogrami drugačni. Oblika histograma pri grupiranih podatkih je okvirna. Z grupiranjem lahko izgubljamo podatke (znotraj skupine I(7,9) ne vemo, koliko je sedmic, osmic, devetic), kar se odraža v obliki. Vsi 4 histogrami prikazujejo iste podatke.

Pri štetju frekvenc se v *Excelu* pomagamo z funkcijo »Freaquency« (izberemo zgornje meje razredov in vrednosti spremenljivke), pri tem potrdimo izbiro s CTRL, SHIFT in ENTER, da se izpišejo vrednosti za vse razrede. Excel riše stolpčne diagrame, da dobimo histogram s klikom na stolpec v meniju izbrišemo razmik med njimi, saj vrednosti intervalnih in razmernostnih ravni prikazujemo s histogrami, nominalnih in ordinalnih pa ne, saj posamezne »vrednosti« niso »povezane« med seboj.
V *SPSS* uporabljamo Analyze → Descriptive Statistic → Frequency. SPSS riše histograme »po svoje«, s klikom nanje lahko prilagodimo število in širino razredov. Izriše tudi tabelo, v kateri prešteva frekvence posameznih vrednosti spremenljivke.

Kategorialne spremenljivke



Frekvenčna porazdelitev pri kategorialnih spremenljivkah

Ena spremenljivka – enostavna struktura

Spol	*f*	%
moški	30	54.5
ženska	25	45.5
Skupaj	55	100.0

spol

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	moški	30	54.5	54.5	54.5
	ženska	25	45.5	45.5	100.0
	Total	55	100.0	100.0	

Premislimo, ali je tabela sploh potrebna.

Kategorialne (kvantitativne):
- dihotomne
- trihotomne
- …

Numerične (kvantitativne):
- intervalne,
- razmernostne
- tudi ordinalne

To so tiste spremenljivke, ki so nominalne. Njihove vrednosti so ime za kategorijo (spol: ženski – 1, moški – 2). Frekvenčna analiza je skoraj vse, kar lahko pri kategorialnih spremenljivkah naredimo (npr. srednje vrednosti spola ne moremo računati).

Če npr. 3 osebe ne bi želele povedati spola, bi dodali še spremenljivko »Missing« s frekvenco 3, vrednost »Total« bi bila torej 58. *Manjkajoč komentar – tiste pomanjkljive namreč ne obdelujemo vedno in povsod.*

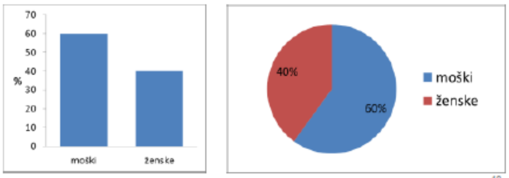
»*Cumulative Percent*« (glej spodnjo tabelo zgoraj) pomeni odstotek posameznikov, ki so v raziskavi dosegli dotičen rezultat ali nižjega. $\boxed{c\% = {}^{cf}/_{N}\,(100\%)}$ »*Cumulative frequence*« (*cf*) pa predstavlja število posameznikov, ki so se uvrstili na ali pod določeno vrednost.

Frekvenčna porazdelitev pri kategorialnih spremenljivkah

Ena spremenljivka – enostavna struktura

$f_{st} = 3.6° f_{\%}$

strukturni stolpci — strukturni krogi



Risanje tortnega diagrama (pite) na roko – vsak odstotek »pite« predstavlja 3,6° kroga iz centra.

Frekvenčna porazdelitev pri kategorialnih spremenljivkah

Več spremenljivk - večstopenjska struktura
→ kontingenčna (kombinacijska) tabela

spol * kajenje Crosstabulation

Count

		kajenje		
		kadi	ne kadi	Total
spol	moški	13	17	30
	ženska	10	15	25
Total		23	32	55

Večstopenjska struktura – več spremenljivk, ki jih križamo (npr. križamo spol s tem, ali oseba ne kadi ali oseba kadi). Podatke razdelimo. Gre za kontingenčno tabelo.

Frekvenčna porazdelitev pri kategorialnih spremenljivkah

Več spremenljivk - večstopenjska struktura
→ tabela s kotnim sto

spol * kajenje Crosstabulation

			kajenje		
			kadi	ne kadi	Total
spol	moški	Count	13	17	30
		% of Total	23.6%	30.9%	54.5%
	ženska	Count	10	15	25
		% of Total	18.2%	27.3%	45.5%
Total		Count	23	32	55
		% of Total	41.8%	58.2%	100.0%

Frekvenčna porazdelitev pri kategorialnih spremenljivkah

Več spremenljivk - večstopenjska struktura
→ tabela s stolpičnim sto

spol * kajenje Crosstabulation

			kajenje		Total
			kadi	ne kadi	
spol	moški	Count	13	17	30
		% within kajenje	56.5%	53.1%	54.5%
	ženska	Count	10	15	25
		% within kajenje	43.5%	46.9%	45.5%
Total		Count	23	32	55
		% within kajenje	100.0%	100.0%	100.0%

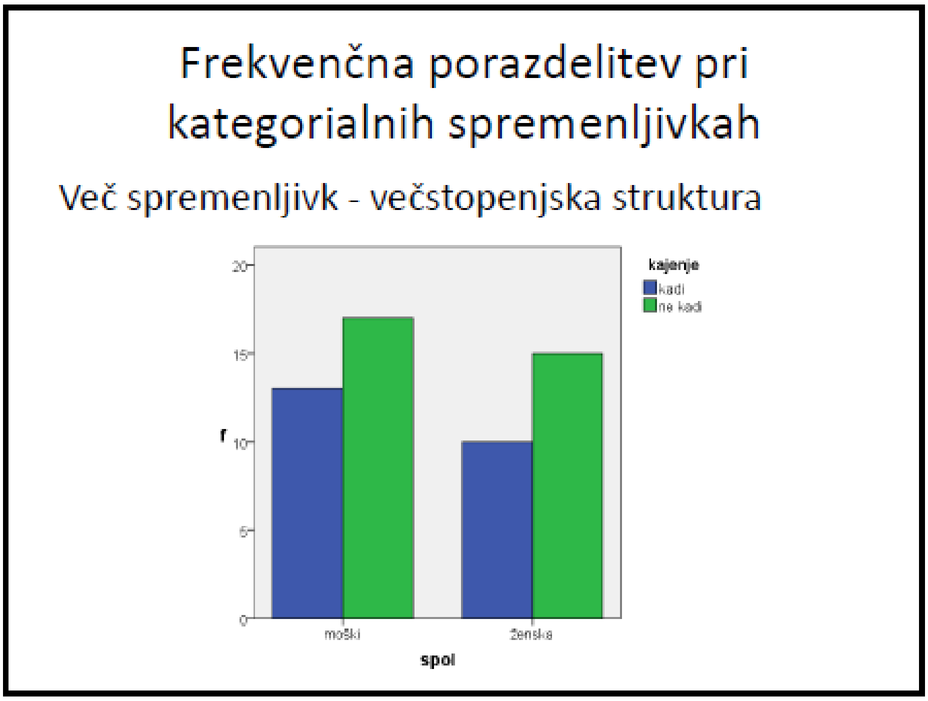
Poudarek pri posamezni tabeli je drugje. Pri stolpčnem ugotovimo, da je poudarjen delež moških kadilcev. Pri vrstičnim pa pojasnjujemo koliko jih kadi ali ne kadi pri moških in ženskih. Kako to ugotovimo? Beremo po stolpcih. Ljudje lažje primerjajo podatke po stolpcih, zato jih primerno predstavimo.

Frekvenčna porazdelitev pri kategorialnih spremenljivkah

Več spremenljivk - večstopenjska struktura
→ tabela z vrstičnim sto

spol * kajenje Crosstabulation

			kajenje		Total
			kadi	ne kadi	
spol	moški	Count	13	17	30
		% within spol	43.3%	56.7%	100.0%
	ženska	Count	10	15	25
		% within spol	40.0%	60.0%	100.0%
Total		Count	23	32	55
		% within spol	41.8%	58.2%	100.0%



Predstavitev rezultatov raziskave

Predstavitev rezultatov

- povzetek lastnosti zbranih podatkov in statistične obdelave
- vključimo vse relevantne rezultate, tudi tiste, ki ne potrjujejo naše hipoteze! (včasih avtorji navajajo le statistično pomembne rezultate)
  - ↳ metaanalize – raziskovalci poskušajo zbrati podatke iz različnih raziskav, metaanaliza ne da prave slike stanja, če bi drugi raziskovalci navajali samo raziskave, ki kažejo na npr. velike razlike med moškimi in ženskami, ne pa tistih, ki ne kažejo na razlike
- individualnih podatkov ne vključujemo, razen kot primere ali pri raziskavah na eni osebi
  - ↳ nikoli ne navajamo imen, priimkov, vedno kažemo povprečje, standardno deviacijo, deleže, ne pa izpisane dosežke posameznikov

Predstavitev statistik

- bralec je statistično izobražen
  - ni mu treba pojasnjevati, kaj je npr. mediana
- po branju bi moral bralec znati ponoviti analizo, torej predstavimo vse potrebno (in samo to)
- deskriptivne statistike (če predstavimo mero centralne tendence, predstavimo tudi mero razpršenosti) – najprej predstavimo deskriptivne statistike, nato inferenčne
- inferenčne statistike: t (39) = 7,12, p < ,01

Predstavitev statistik

Statistike v besedilu:

Aritmetične sredine (in standardne deviacije v oklepajih) so za poskuse od 1 do 4 znašale 2,34 (0,50), 2,59 (1,21), 2,68 (0,39) in 2,86 (0,12).

poskus	*M*	*SD*
1	2,34	0,50
2	2,59	1,21
3	2,68	0,39
4	2,86	0,12

Kot smo predvidevali, je skupina oseb, ki so zelo dojemljive za hipnozo (M = 8,19, SD = 7,12), bolj pogosto kot skupina preostalih oseb (M = 5,26, SD = 4,25) poročala, da so zaznali premik nepremične lučke, $t(60)$ = 1,99, p = ,05. (poseben način zapisovanja rezultatov statističnih testov)

Tabel in slik naj ne bi bilo več kot 4 na poročilo. Enostavne tabele ne vključujemo, če lahko podatke jasno in enostavno vključimo v besedilo.

Ker so vejice včasih v slovenščini nepregledne, lahko v dilemi uporabimo podpičje med pomembnimi členi.

»t« pomeni statistiko (lahko bi jo označevale tudi druge črke), ki ima pripadajočo prostostno stopnjo (60 [velikost vzorca – št. ocenjevalnih parametrov]), rezultat statistike (1,99) in raven statistične pomembnosti (p=,05).

Izbor tabel in slik

- jasnost, ekonomičnost (pišemo v besedilu, dokler je še pregledno)
- prikaz natančnih kvantitativnih vrednosti
  - ↪ tabele
- prikaz splošnega vzorca v rezultatih, prikaz interakcije, primerjave, strukture
  - ↪ slike
- brez podvajanja!

Tabele

- za prikaz velike količine podatkov, predvsem kvantitativnih
- nekaj namigov:
  - zaokroževanje (enako št. dec. mest, vejice na enakih mestih)
  - navpično je lažje primerjati kot vodoravno (bralec lažje primerja vrednosti po stolpcu navzdol)
  - robna povprečja

Enostavna tabela

Tabela 1

Delež napak na preizkusu A pri različnih ravneh težavnosti

glava

čelo

Raven težavnosti	*M*	*SD*
nizka	,15	,08
srednja	,25	,07
visoka	,40	,10
ekstremno visoka	,89	,06

jedro

✍ Vsaka tabela potrebuje naslov. »Tabela« in zaporedno številko pišemo v pokončnem tisku nad tabelo, pod oznako pa naslov v poševnem tisku. Med spremenljivkami ni tabelčnih črt, ta razmejuje le glavo tabele. Zelo velikokrat se ničle spuščajo, ko gre število od -1 do 1, ali pa pri deležih (kjer ne moremo imeti večjega deleža od 1), če pa vrednosti lahko presegajo vrednost 1, moramo ničle obvezno pisati.

Večstopenjska tabela

Tabela 2

Povprečno število pravilnih odgovorov pri otrocih z različno količino vaje

		Ocena		
Skupina	*N*[a]	3	4	5
11-letni				
brez vaje	14	200	205	213
7 dni vaje	15	240	251	260
14 dni vaje	17	250	263	288
21 dni vaje	19	281	290	306
28 dni vaje	20	290	290	307
12-letni				
brez vaje	15	221	234	221
7 dni vaje	17	243	250	270
14 dni vaje	18	261	274	290
21 dni vaje	19	290	296	299
28 dni vaje	20	301	305	308

Opombe: Maksimalni dosežek = 320.

[a] Število otrok (od 20), ki so rešili vse teste.

Imamo dve spremenljivki – starost in dnevi vaje + še spremenljivko ocene pri matematiki, torej 3 spremenljivke. Kakor pišemo naslov tabele poševno, pišemo tudi »*Opombe:*«/»*Opombe.*«. Oznaka »[a]« označuje opombe za stolpec N[a].

Večstopenjska tabela

Tabela 3

Povprečno število pravilnih odgovorov pri otrocih z vajo in brez nje

		Ocena		
Skupina	*N*[a]	3	4	5
	Verbalni testi			
deklice				
z vajo	18	280	297	301
brez vaje	19	240	251	260
dečki				
z vajo	19	281	290	306
brez vaje	20	232	264	221
	Matematični testi			
deklice				
z vajo	20	201	214	221
brez vaje	17	189	194	216[b]
dečki				
z vajo	19	210	236	239
brez vaje	18	199	210	213

Opombe: Maksimalni dosežek = 320.
[a] Število otrok (od 20), ki so rešili vse teste.
[b] Ena deklica v tej skupini je pravilno rešila le dve nalogi.

Naslovi tabel

- preveč telegrafski : Odnos med smerjo in ocenami
- preveč natančen : Povprečna ocena za predmete pedagoškega paketa, statistiko in informatiko pri študentih psihologije, pedagogike in ekonomije
- ravno pravšnji : Povprečne ocene treh predmetov pri študentih različnih študijskih smeri

Jedro tabele

- vsa polja enako število decimalnih mest
- prazna polja – v tem primeru nujno vključimo opombe, zakaj podatka ni
  - nemogoč pogoj: [prazno polje]
  - nismo zbrali podatkov ali nismo računali: / (/ pojasnimo v opombah)
- oblikovanje (širina stolpcev, prazne vrstice, poravnava jedra desno [glave pa levo])
- posebne vrste tabel (ANOVA, regresijska analiza)

		Poskus 1	Poskus 2	Poskus 3	Poskus 4
Sk. 1					
	M	2,34	2,59	2,65	/
	SD	0,50	1,21	0,39	/

Sk. 2

	M	2,60	2,50	2,60	2,30
	SD	0,50	1,25	0,38	2,25

Opombe. Znak / pomeni, da …

Opombe pod tabelo

- hierarhija opomb po pomembnosti:
  - **splošne opombe**: pojasnjevanje celotne tabele, kratic, simbolov
    - *Opombe.* Vse statistično nepomembne trismerne interakcije smo izpustili. U = ujemanje, N = neujemanje.
  - **specifične opombe**: pojasnjevanje dela tabele
    - [a]N = 25.
  - **verjetnostne opombe**
    - **p* < ,05. ***p* < ,01. ****p* < ,001.
- opombe se končajo s pikami: [a] Število otrok, ki so rešili vse teste. [b] Ena deklica iz skupine je rešila le 2 nalogi. [c] Dva dečka iz zadnje skupine sta dosegla vse točke. (sledijo si po novih APA standardih, ne več ena pod drugo)

Journal of Experimental Psychology: Learning, Memory, and Cognition
2006, Vol. 32, No. 3, 316–332

Copyright 2006 by the American Psychological Association
0278-7393/06/$12.00 DOI: 10.1037/0278-7393.32.3.316

Recency Effects as a Window to Generalization: Separating Decisional and Perceptual Sequential Effects in Category Learning

Matt Jones, Bradley C. Love, and W. Todd Maddox
University of Texas at Austin

Table 1
Comparison of Predictions Regarding Category Contrast Effect

		Hypothesis	
Feedback	Category	Decisional	Perceptual
Modal	Same	−	−
Modal	Different	+	+
Amodal	Same	+	−
Amodal	Different	−	+

Note. All predictions are based on responses to intermediate stimuli following extreme stimuli. Cases where the response is likely to be optimal (i.e., matching the modal category) are indicated by "+"; cases where the nonoptimal (amodal) response is predicted are indicated by "−." The "Category" column indicates whether the two stimuli are from the same half of the stimulus range. The "Feedback" column refers to feedback given for the extreme stimulus. Deterministic tasks only involve the modal case (rows 1 and 2), where decisional and perceptual effects are in concert, whereas probabilistic tasks also include the amodal case, where the hypotheses make opposite predictions.

Roškar, S., Podlesek, A., Kuzmanić, M., Omejc Demšar, L., Zaletel, M., & Marušič, A. (2011). Suicide risk and its relationship to change in marital status. *Crisis*, *32*(1), 24-30.

Table 1. Sociodemographic characteristics of suicide victims and community controls and incidence of different types of marital status change within the two groups in the last 5 years

	Suicide victims	Matched controls
N	1614	4617
Mean age (*SD*) in years	51.6 (17.1)	51.3 (17.0)
n male (%)	1243 (77.0)	3546 (76.8)
n female (%)	371 (23.0)	1071 (23.2)
Change in marital status	172 (10.7)[a]	257 (5.6)[b]
Got married (%)	31 (1.9)	62 (1.3)
Got widowed (%)	87 (5.4)	115 (2.5)
Got divorced (%)	54 (3.3)	80 (1.7)

Note. [a]Another 6 participants ([b]Another 5 participants) had changed their marital status for the second or third time in the last 5 years or had a wrong entry in the database. They are not included in these numbers and in subsequent analyses.

✍ Neoptimalna tabela.

Roškar, S., Podlesek, A., Kuzmanić, M., Omejc Demšar, L., Zaletel, M., & Marušič, A. (2011). Suicide risk and its relationship to change in marital status. *Crisis*, *32*(1), 24-30.

Table 2. Frequencies (%) of suicide victims and controls with a certain type of change in marital status during the last 5 years

	Years prior to the change in marital status				
Type of change	1	2	3	4	5
Getting married					
Controls	10 (16)	11 (18)	13 (21)	14 (23)	14 (23)
Suicide victims	19 (61)	* (7)	* (3)	* (7)	7 (23)
Becoming widowed					
Controls	20 (17)	27 (24)	31 (27)	17 (15)	20 (17)
Suicide victims	40 (46)	13 (15)	12 (14)	14 (16)	8 (9)
Getting divorced					
Controls	18 (23)	15 (19)	13 (16)	20 (25)	14 (18)
Suicide victims	20 (37)	12 (22)	10 (19)	6 (11)	6 (11)

Notes. Year 1 denotes that the change has happened in the last year prior to suicide or census; year 2 denotes that the change has happened more than 1 and less than 2 years before the event, etc. *For data protection purposes, frequencies were not included in the table as they were lower than 5.

✍ Tabela ima preširoke stolpce. Večji razmik je lahko le med glavo in jedrom. Z bližino lahko pokažemo tudi to, kaj sodi bolj skupaj.

Psychological Review
2005, Vol. 112, No. 2, 383–416

Copyright 2005 by the American Psychological Association
0033-295X/05/$12.00 DOI: 10.1037/0033-295X.112.2.383

The Dynamics of Scaling: A Memory-Based Anchor Model of Category Rating and Absolute Identification

Alexander A. Petrov
University of California, Irvine

John R. Anderson
Carnegie Mellon University

Table 4
Performance of a Hierarchy of Models on a Battery of Measures of Various Phenomena

Phenomenon and statistic	Model							
	$\mathcal{M}_{P}$	$\mathcal{M}_{PS}$	$\mathcal{M}_{PSC}$	$\mathcal{M}_{PSA}$	$\mathcal{M}_{PSAC}$	$\mathcal{M}_{PSCR}$	$\mathcal{M}_{PSACR}$	Empirical
Absolute identification								
Transmitted information (T)	1.79	1.53	1.57	1.50	1.66	1.57	1.57	1.68
Nonuniform response distribution (s)	2.59	2.55	2.52	2.55	2.50	2.50	2.50	2.40
Edge effect (*bow*)	−.42	−.47	−.30	−.38	−.29	−.27	−.31	+.14
Repetition effect (*rep*)	.01	.01	.01	.15	.05	.01	.04	.11
Assimilative context effect (a)	−.00	+.00	+.00	+.14	+.11	+.00	+.11	+.14
Practice effect (*pr*)	.00	.00	.00	.06	.01	.01	.02	.06
Bayesian information criterion (BIC)	504	496	491	522	493	504	504	—
Category rating								
Overall accuracy (R^2)	.82	.68	.80	n. a.	.82	.73	.77	.77
Nonuniform response distribution (s)	2.08	2.06	2.05	n. a.	1.91	2.56	1.93	1.77
Nonstationary distribution (Δs)	.00	.00	.00	n. a.	−.01	.12	.21	.55
Sequential effect (r)	.00	.00	.00	n. a.	.04	.08	.17	.34
Gradual trend (ΔARL)	.00	.00	.00	n. a.	−.01	.08	.27	.49
Compensatory context effect (d)	+.00	+.01	+.01	n. a.	+.26	−.62	−.53	−.21
Bayesian information criterion (BIC)	615	595	614	n. a.	537	552	472	—

Note. Each cell reports the mean of 9,600 model runs. Model $\mathcal{M}_{P}$ has perceptual variability only. Models $\mathcal{M}_{PS}$ and $\mathcal{M}_{PSC}$ add anchor selection and response correction. Models $\mathcal{M}_{PSA}$, $\mathcal{M}_{PSAC}$, and $\mathcal{M}_{PSCR}$ introduce activation learning and running averaging of anchor locations. Model $\mathcal{M}_{PSACR}$ is synonymous with the full ANCHOR model. Compare with the empirical data in Tables 1 and 3. Dashes indicate cells in which BIC is not defined. n. a. = not applicable; ARL = average response level.

Perception & Psychophysics
2004, 66 (4), 665-678

The influence of category membership of stimuli on sequential effects in magnitude judgment

PETER PETZOLD and GERT HAUBENSAK
Justus Liebig University of Giessen, Giessen, Germany

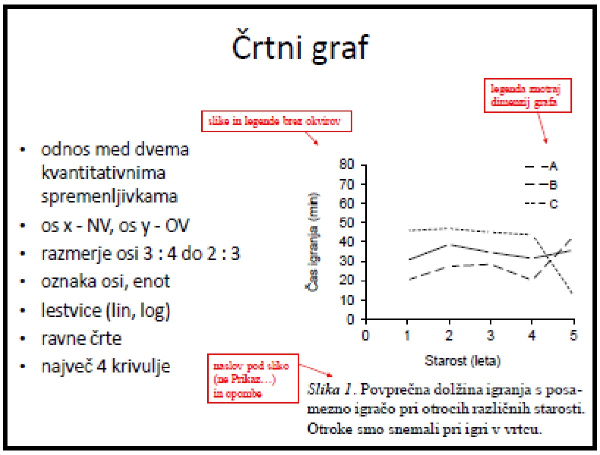
Table 2
Results of Linear Regression Analyses for Two Instructions.

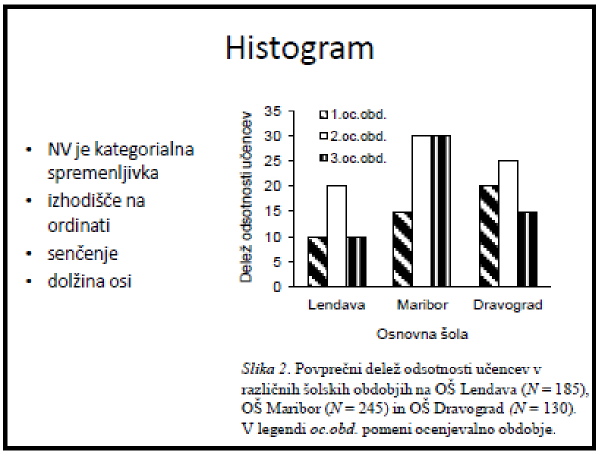
Subject	a_0^*	a_1^*	b_1	a_2^*	b_2	c	u_1^*	v_1^*	u_2^*	v_2^*	Δ_1	Δ_2
Separation Instruction												
1	.493	−.189	.514	−.065	.111	.805	.174	−.438	.024	.027	.064	−.010
2	.363	−.217	.638	−.079	.222	.700	.219	−.584	.081	−.168	.015	.002
3	.488	−.177	.449	−.147	.251	.589	.164	−.412	.258	−.510	.042	−.025
4	.464	−.199	.508	−.053	.163	.818	.165	−.527	.074	−.164	.037	.023
5	.395	−.264	.609	−.033	.183	.394	.217	−.588	.093	−.203	−.023	.039
6	.460	−.210	.641	−.089	.153	.809	.200	−.573	.023	−.094	.085	−.019
M	.444	−.209	.560	−.078	.181	.686	.190	−.520	.092	−.185	.037	.002
Integration Instruction												
7	.042	−.032	.350	.019	.166	.004	−.018	.026	−.054	.088	−.017	.026
8	.316	−.184	.544	−.077	.163	.202	.066	−.160	−.015	.020	−.012	−.025
9	.304	−.182	.594	−.133	.218	.145	.027	−.075	.038	−.118	−.001	−.067
10	.186	−.192	.722	.020	−.048	−.203	.203	−.416	−.055	.130	−.058	.011
11	.316	−.034	.392	−.020	.108	.428	−.099	.223	.057	−.032	.090	.014
12	.295	−.141	.644	−.062	.174	.037	.011	−.010	.015	−.065	.049	−.011
M	.240	−.128	.541	−.042	.130	.101	.032	−.069	.002	.004	.008	−.009

💻 Minus (–) dobimo s: Ctrl+Num-, da dobimo pravi minus. Zgornja tabela je primerna.

Slike

- tipi:
  - grafi: črtni, histogrami, pite, razpršitveni diagrami
  - sheme, skice, fotografije …
- slika naj kaže le bistvene podatke
- preglednost in razumljivost





Črtne grafe uporabljamo za intervalne spremenljivke. Na abcisi navajamo neodvisno spremenljivko, na ordinato pa odvisno. Ni nujno, da je lestvica linearna (1, 2, 3, 4, 5 …), lahko je tudi npr. logaritemska (1, 10, 100 …).

💻 Posebej pozorni moramo biti v Excelu, ki bo kategorije postavil eno zraven druge (5, 6, 7, 9, 11, 12) – med vsako vrednostjo bo enaka razlika, v tem primeru uporabljamo »scatter plot« - peščični diagram.

✍ Pri črtnem grafu po navadi ne uporabljamo barv, ampak različen stil črtkanja. Legenda je po navadi pozicionirana pod sliko grafa, znak ob znaku ne pa v stolpcih kot na vzorcnem primeru.

Histograme uporabljamo, ko imamo opravka z eno nominalno in eno intervalno spremenljivko.

✍ Stolpcev ne barvamo – tudi v sivinah ne, ampak jih šrafiramo. Grafe označujemo z »Slika« in zaporedno številko, naslov je lociran pod grafom v poševnem tisku, sledi pika in naslov v pokončnem tisku.

Oblikovanje

- širina slike za A4 format:
  1 stolpec: 5,0 - 8,45 cm
  2 stolpca: 10,6 - 17,5 cm
- velikost nabora črk in številk
  > 8, < 14 pt, razlika največ 4 pt
- nabori sans serif, ne serif
- prekinitev osi —//—
- enotna oblika za vse slike

Roškar, S., Podlesek, A., Kuzmanić, M., Omerje Demšar, L., Zaletel, M., & Marušić, A. (2011). Suicide risk and its relationship to change in marital status. *Crisis, 32*(1), 24-30.

Figure 1. Percentage of marital status change in the suicide victims and the control group according to years prior to suicide.

Naslove pišemo jasno, jedernato: Odnos med A in B, Odstotek tega in tega – ne uporabljamo izrazov, kot so »prikaz« ipd. Širina naslova je enaka širini grafa.

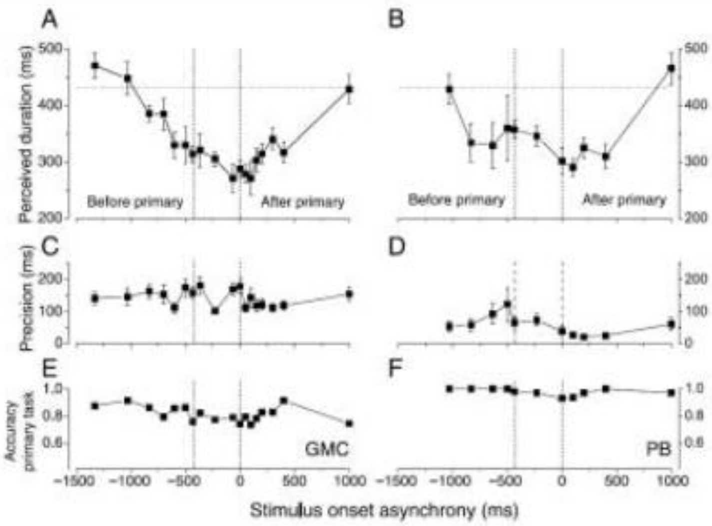
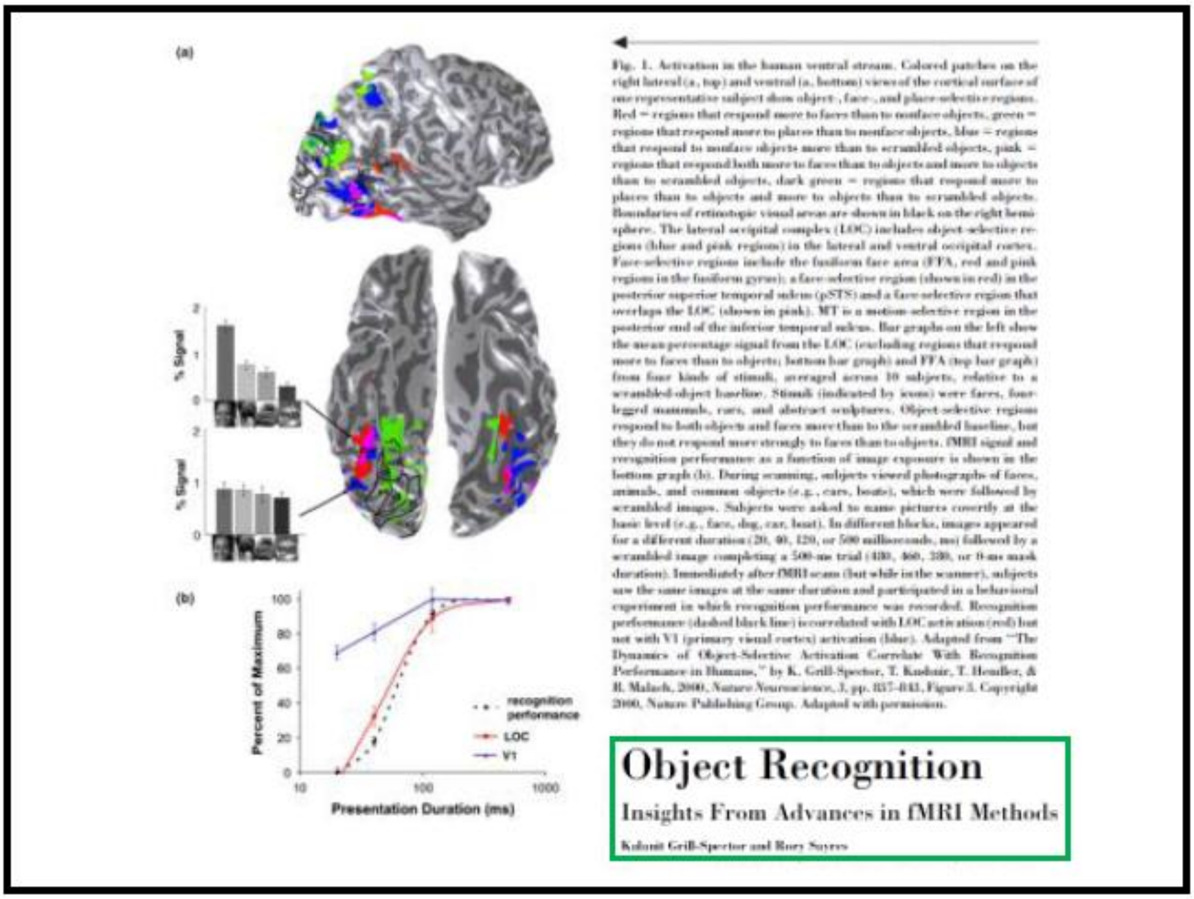
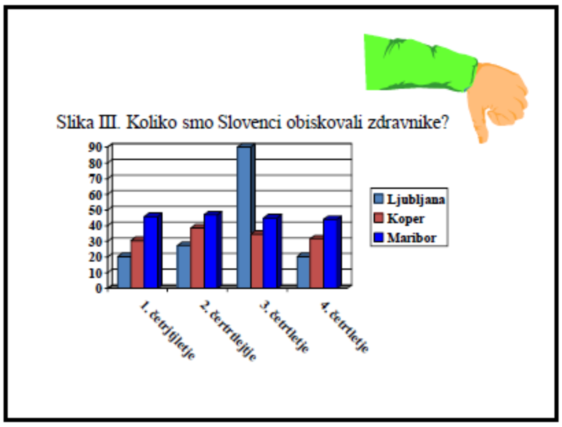


Figure 2. Data for two representative subjects for primary task stimuli displayed along the horizontal meridian. Top row (panels A and B): perceived duration of a 433 ms interval as function of the Stimulus Onset Asynchrony between the primary task and secondary task stimuli. The horizontal line indicates the physical target duration of 433 ms. Positive SOA values indicate intervals starting after the appearance of the primary task stimuli. Negative values indicate intervals where the first bar is flashed before the appearance of the primary task stimuli. The two vertical bars, placed at −433 ms and 0 ms indicate respectively, intervals which terminate with the presentation of the primary task stimulus and intervals which start with primary task stimulus. Error bars are *SEM* calculated via bootstrap procedure (see Methods). Middle row (panels C and D): precision of the duration judgement, measured as difference between the 50th percentile and the 75th percentile of the psychometric function. Error bars are *SEM*. Bottom row (panels E and F): accuracy of the primary task as function of Stimulus Onset Asynchrony between the primary task and secondary task stimuli. Chance level is 50%.

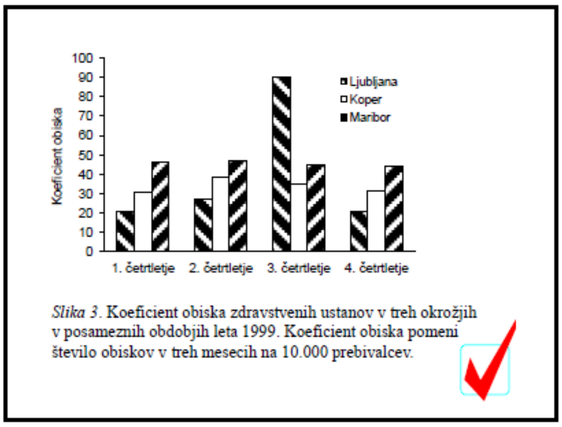
Citation: Cicchini, G. M., & Morrone, M. C. (2009). Shifts in spatial attention affect the perceived duration of events. *Journal of Vision, 9*(1):9, 1–13, http://journalofvision.org/9/1/9/, doi:10.1167/9.1.9.



Pojasnimo, kaj je v posameznem delu (delu (a), delu (b) …) navedeno.



Napake: 3-D se izogibamo, nikoli ne obračamo ime kategorij poševno (pazimo pri Excelu!), legenda bi morala biti brez okvirčka npr. v grafu ali spodaj, »Slika« bi morala biti zapisana poševno, njena zaporedna številka ne sme biti v rimskih številkah, barv ne uporabljamo, manjka oznaka osi y, naslov je neprimeren …



Rešitev.

Tabele, slike in besedilo

- Tabele in slike dopolnjujejo besedilo, ga ne podvajajo.
- V besedilu se nanje sklicujemo:
  - V tabeli 2 vidimo, da ...
  - … (glej sliko 7) …manjka oznaka y
- V besedilu pojasnimo, kaj je zapisano v tabeli, kako naj bralec bere tabelo in pregleda sliko.
- Vsaka tabela in slika mora biti samostojna, razumljiva brez besedila.

Položaj podatka v porazdelitvi: Percentili in percentilni rangi

9 oseb smo vprašali, koliko zelo dobrih prijateljev imajo. Rezultati so:

X	1	2	2	3	3	3	4	4	5

- *Koliko podatkov je manjših od 4?* 6 podatkov oz. ${}^{6}/_{9} = {}^{2}/_{3}$ podatkov. Gre za grobe ocene, saj ne moremo vedeti, če ima druga oseba, ki ima rezultat 4, res višji rezultat od te.
- *Koliko podatkov je manjših ali enakih 4?* 8 podatkov oz. ${}^{8}/_{9}$ podatkov.

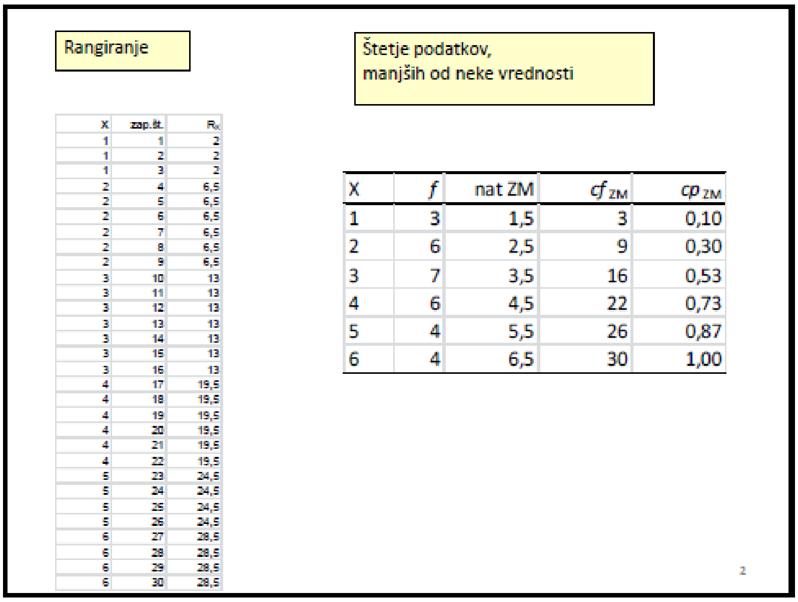
Percentilni rang dosežka nam pove, kolikšen odstotek podatkov je enakih ali nižjih od tega dosežka.

- Diskretne spremenljivke - npr. število otrok ali pa številka parkirnega mesta
- Zvezne spremenljivke - med dvema številoma je neskončno drugih števil (npr. anksioznost, sposobnost, inteligentnost … gre za eno kontinuus izraženosti)
  - v praksi jih ne moremo neskončno natančno izmeriti (odvisno od instrumenta, ki mora biti zelo dobro izbran; nekdo lahko med dvema odgovoroma DA NE koleba, mi pa ju ocenimo bodisi z 1 bodisi z 2)
  - zaokroževanje (IQ 100 pomeni, da se oseba nahaja med 99,5 in 100,5, oseba se nahaja na intervalu, ki ima natančno zgornjo in spodnjo mejo)

---

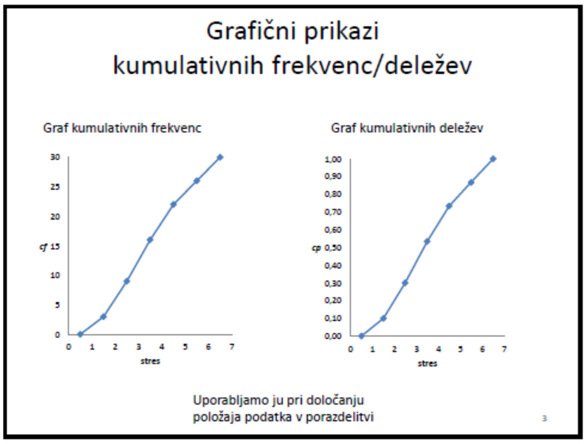
1. korak: Razvrstimo (rangiramo) podatke po vrsti.
2. korak: Absolutni rangi so enaki zaporedni številki osebe. Vendar ko se srečamo z enakimi podatki (npr.: X=1, *f*=3), upoštevamo povprečni rang (npr. $R_{X/A}$=2). Za pomoč si lahko napišemo zaporedne številke.
3. korak: Da bi si lahko predstavljali umeščenost v skupino, potrebujemo podatek o numerusu (za spodnjo tabelo je *N*=26)
4. korak: Pozorni moramo biti, ker govorimo po navadi o zveznih spremenljivkah, katere meje intervalov uporabljamo – sredine, zgornje ali spodnje meje? Do kje kumuliramo?
5. korak: kumulativno frekvenco *cf* izračunamo po enačbi: $cf = R_X - \frac{1}{2}$.
6. korak: glede na numerus in kumulativno frekvenco izračunamo kumulativni %: $cp = cf/N$

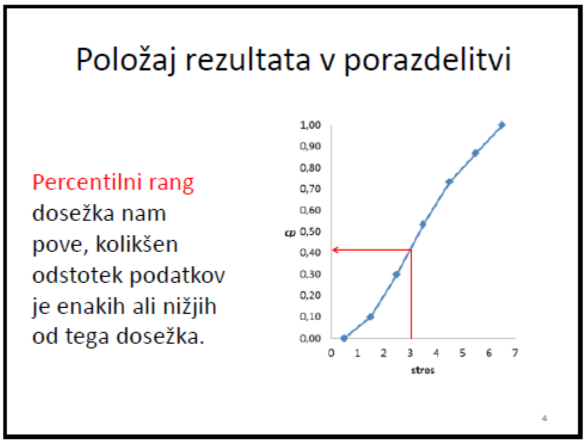
X	R_A/R_X	*cf*	*cp*
7	1,0	0,5	0,02
11	2,0	1,5	0,06
12	3,0	2,5	0,10
13	4,5	4,0	0,15
13	4,5	.	.
15	6,5	6,0	0,23
15	6,5	.	.
16	8,0	7,5	0,29
17	9,5	9,0	0,35
17	9,5	.	.
18	11,5	11,0	0,42
18	11,5	.	.
20	13,0	12,5	0,48
21	15,5	15,0	0,58
21	15,5	.	.
21	15,5	.	.
21	15,5	.	.
22	18,0	17,5	0,67
23	19,5	19,0	0,73
23	19,5	.	.
24	21,0	20,5	0,79
25	22,0	21,5	0,83
27	23,5	23,0	0,88
27	23,5	.	.
29	25,0	24,5	0,94
30	26,0	25,5	0,98



X	zap.št.	R_X
1	1	2
1	2	2
1	3	2
2	4	6,5
2	5	6,5
2	6	6,5
2	7	6,5
2	8	6,5
2	9	6,5
3	10	13
3	11	13
3	12	13
3	13	13
3	14	13
3	15	13
3	16	13
4	17	19,5
4	18	19,5
4	19	19,5
4	20	19,5
4	21	19,5
4	22	19,5
5	23	24,5
5	24	24,5
5	25	24,5
5	26	24,5
6	27	28,5
6	28	28,5
6	29	28,5
6	30	28,5

X	*f*	nat ZM	cf_{ZM}	cp_{ZM}
1	3	1,5	3	0,10
2	6	2,5	9	0,30
3	7	3,5	16	0,53
4	6	4,5	22	0,73
5	4	5,5	26	0,87
6	4	6,5	30	1,00

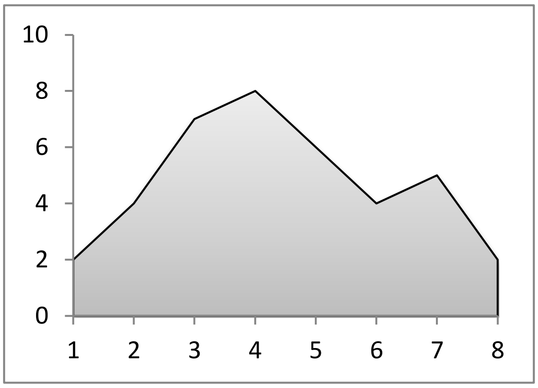


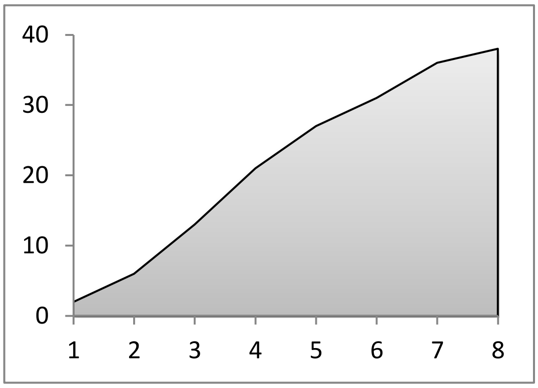


Pri risanju grafa za kumulativne frekvence se upoštevajo **zgornje meje**, pikice so na 1˙5, 2˙5 itn.

Grafi kumulativnih frekvenc (graf desno) so po ploščini enaki poligonom (graf levo) pod pogojem, da so vsi izpostavljeni npr. merilu zlatega reza. Ker sicer na prvi pogled (in razmislek) ne izgleda tako, se lahko prepričamo s sliko. Torej: *P*(poligon) = *P*(graf *cf*), seveda pod pogojem, da morata biti sliki enako visoki in široki, sicer pa ploščini zagotovo nista enaki.

f	*cf*
2	2
4	6
7	13
8	21
6	27
4	31
5	36
2	38





Določanje percentilnega ranga podatka v ranžirni vrsti

X_i	$R(X_i)$	$P(X_i)$
7	1	.02
11	2	.06
12	3	.10
13	4.5	.15
13	4.5	.15
15	6.5	.23
15	6.5	.23
16	8	.29
17	9.5	.35
17	9.5	.35
18	11.5	.42
18	11.5	.42
20	13	.48
21	15.5	.58
21	15.5	.58
21	15.5	.58
21	15.5	.58
22	18	.67
23	19.5	.73
23	19.5	.73
24	21	.79
25	22	.83
27	23.5	.88
27	23.5	.88
29	25	.94
30	26	.98

Absolutni rang R(X)

- Nevezani rangi
- Vezani rangi

Percentilni rang P(X)

Za zvezne spremenljivke, merjene nenatančno

$$P(X) = \frac{R(X) - 0{,}5}{N}$$

R(16,0) = 8
R(20,0) = 13
R(21,0) = 15,5

P(16,0) = ,29
P(20,0) = ,48
P(21,0) = ,58

Kolikšen je percentilni rang za X=21?

$P(21) = ?\quad P(21) = \frac{R(21) - 0{,}5}{N} = \frac{15{,}5 - 0{,}5}{26} = 0{,}58$

Kolikšen je percentilni rang za X=13?

$P(13) = ?\quad P(13) = \frac{4{,}5 - 0{,}5}{26} = 0{,}15$

Absolutni rangi so primerni le za primerjave znotraj skupine. Da bi primerjali položaje v različno velikih skupinah, potrebujemo relativne (percentilne) range.

Kolikšen je percentilni rang za X=10? Ne vemo, ampak bomo izvedeli, če *interpoliramo*.

Linearna interpolacija

Določanje percentilnega ranga vrednosti, ki se ne pojavlja v ranžirni vrsti

x

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	7.00	1	3.8	3.8	3.8
	11.00	1	3.8	3.8	7.7
	12.00	1	3.8	3.8	11.5
	13.00	2	7.7	7.7	19.2
	15.00	2	7.7	7.7	26.9
	16.00	1	3.8	3.8	30.8
	17.00	2	7.7	7.7	38.5
	18.00	2	7.7	7.7	46.2
	20.00	1	3.8	3.8	50.0
	21.00	4	15.4	15.4	65.4
	22.00	1	3.8	3.8	69.2
	23.00	2	7.7	7.7	76.9
	24.00	1	3.8	3.8	80.8
	25.00	1	3.8	3.8	84.6
	27.00	2	7.7	7.7	92.3
	29.00	1	3.8	3.8	96.2
	30.00	1	3.8	3.8	100.0
	Total	26	100.0	100.0	

Pri zveznih spremenljivkah je to kumulativni % do zgornje meje intervala.

X	*cp*
7,5	0,038
11,5	0,077
12,5	0,115
…	

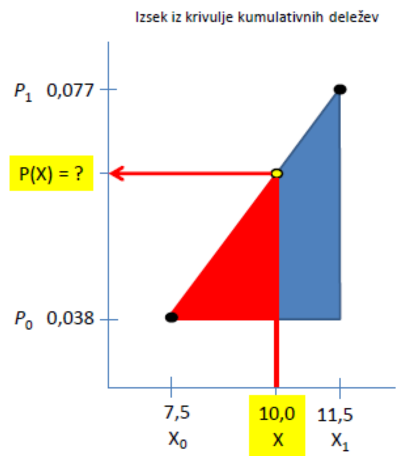
→ Kolikšen je *cp* do 10,0, tj. *P*(10,0)?

Določanje percentilnega ranga vrednosti, ki se ne pojavlja v ranžirni vrsti

$$P(X)=\frac{R(X)-0{,}5}{N}$$

$$\frac{P(X)-P_0}{P_1-P_0}=\frac{X-X_0}{X_1-X_0}$$

$$P(X)=P_0+\frac{X-X_0}{X_1-X_0}(P_1-P_0)$$



Interpolacije se lahko lotimo s pravokotnim trikotnikom. Če se premikamo od vrednosti X_0 do X_1, se X nahaja na določenemu deležu razdalje. Na enakem deležu razdalje se mora nahajati percentilni rang, ko potujemo od P_0 proti P_1.

$$\frac{P(X)-0{,}02}{0{,}06-0{,}02}=\frac{10-7}{11-7}$$

$$P(X)=0{,}05$$

Drugi način: Če se 10 nahaja na $^3/_4$ poti od 7 do 11, potem mora biti iskani percentilni rang prav tako na $^3/_4$ od 0,02 do 0,06.

Interpolacija nam omogoča, da najdemo pribl. vmesno vrednost med dvema danima številoma.

Primer 1: Zanima nas percentilni rang za X=7,0. Upoštevamo natančne intervalne meje.

X	*f*	*cf*	*c%*
10	2	25	100%
9	8	23	92%
8	4	15	60%
7	6	11	44%
6	4	5	20%
5	1	1	4%

7 se nahaja na intervalu s pravimi mejami 6,5 in 7,5. Kumulativni procenti, ki pripadajo spodnji in zgornji meji so 20% in 44%.

X	%
7,5	44%
7,0	**?**
6,5	20%

Razlika med X=7,5 in X=6,5 je 1. Razlika med 44% in 20% pa 24. Naša iskana vrednost je locirana 0,5 pod zgornjo oz. nad spodnjo mejo, torej na sredini. Temu pripada posledično sredinska vrednost med 44 in 20 (24:2=12). 12% pod 44% je 32% (44%-12%=32%), torej P(7,0)=0,32 oz. 32%.

Primer 2: Dopolni tabelo in poišči percentilni rang za X=6,5 in za X=9,5. Upoštevamo natančne zgornje meje.

X	*f*	*cf*	*c%*
10	2		
9	3		
8	5		
7	6		
6	2		
5	2		

X	*f*	*cf*	*c%*
10	2	*20*	*100%*
9	3	*18*	*90%*
8	5	*15*	*75%*
7	6	*10*	*50%*
6	2	*4*	*20%*
5	2	*2*	*10%*

Ker upoštevamo zgornje meje, lahko preberemo, da je P(6,5)=0,20. Tisti za 9,5 pa 0,90.

Primer 3: Iščemo percentilni rang za X=5. (Opomba: V tabeli niso zajeti vsi podatki, zgolj izsek.) Upoštevamo natančne zgornje meje.

X	*f*	*cf*	*c%*
6	2	20	100%
5	2	18	90%
4	4	16	80%

Zgornja in spodnja meja za X=5 sta 4,5 in 5,5. P(4,5)=0˙80, P(5,5)=0˙90. 5 se nahaja na polovici med 4,5 in 5,5 → 0˙10/2=0˙05. Rezultat je 85% oz. 0,85. (Nepreverjeno!)

Primer 4: Izračunaj P(15,2) in P(26,0).

X	*f*	R_A/R_X	*cf*	*cp*
7	1	1,0	0,5	0,02
11	1	2,0	1,5	0,06
12	1	3,0	2,5	0,10
13	2	4,5	4,0	0,15
15	2	6,5	6,0	0,23
16	1	8,0	7,5	0,29
17	2	9,5	9,0	0,35
18	2	11,5	11,0	0,42
20	1	13,0	12,5	0,48
21	4	15,5	15,0	0,58
22	1	18,0	17,5	0,67
23	2	19,5	19,0	0,73
24	1	21,0	20,5	0,79
25	1	22,0	21,5	0,83
27	2	23,5	23,0	0,88
29	1	25,0	24,5	0,94
30	1	26,0	25,5	0,98

P(15,2)=?

$$P = \frac{R - 0{,}5}{N} \rightarrow R_x = R_0 + \frac{x - x_o}{x_1 - x_0}$$

$$R_x = 6{,}5 + \frac{15{,}2 - 15}{16 - 15} = 6{,}5 + 0{,}2 = 6{,}7$$

$$P = \frac{6{,}7 - 0{,}5}{26} = \frac{6{,}2}{26} = 0{,}24$$

P(26,0)=?

$$R_x = 22 + \frac{26 + 25}{27 - 25} = 22 + 0{,}5 = 22{,}5$$

$$P = \frac{22{,}5 - 0{,}5}{26} = \frac{22}{26} = 0{,}85$$

Rezultata: P(15,2)=0,24 in P(26,0)=0,85.

Določanje percentilov

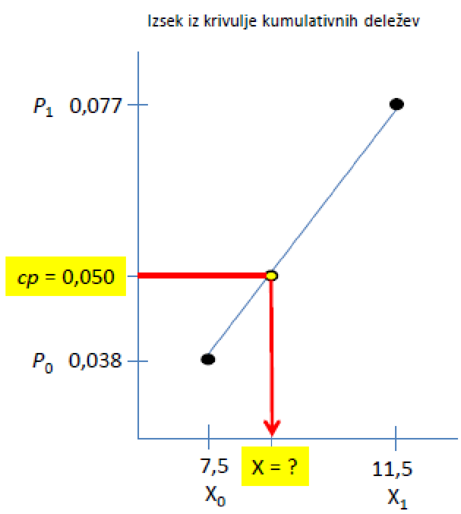
k-ti percentil je vrednost spremenljivke, od katere je imelo *k* % posameznikov nižji dosežek.

$$\frac{P(X) - P_0}{P_1 - P_0} = \frac{X_P - X_0}{X_1 - X_0}$$

$$R(X) = N \cdot P(X) + 0{,}5$$

$$\frac{R(X) - R_0}{R_1 - R_0} = \frac{X_P - X_0}{X_1 - X_0}$$

$$X_P = X_0 + \frac{(X_1 - X_0) \cdot (R(X) - R_0)}{R_1 - R_0}$$



Gre za obratno operacijo iskanju rangov. Percentili predstavljajo namreč posameznike, percentilni rangi pa deleže, odstotke. Označujemo jih s $\mathbf{P_x}$ (percentilne range pa s P(X)).

Primer 1: Iščemo 25. percentil (P_{25}) in 75. percentil (P_{75}).

X	*f*	R_A/R_X	*cf*	*cp*
7	1	1,0	0,5	0,02
11	1	2,0	1,5	0,06
12	1	3,0	2,5	0,10
13	2	4,5	4,0	0,15
15	2	6,5	6,0	0,23
16	1	8,0	7,5	0,29
17	2	9,5	9,0	0,35
18	2	11,5	11,0	0,42
20	1	13,0	12,5	0,48
21	4	15,5	15,0	0,58
22	1	18,0	17,5	0,67
23	2	19,5	19,0	0,73
24	1	21,0	20,5	0,79
25	1	22,0	21,5	0,83
27	2	23,5	23,0	0,88
29	1	25,0	24,5	0,94
30	1	26,0	25,5	0,98

Takoj izpeljemo, da je percentilni rang P(X)=0,25. Gre namreč za isto stvar, le da enkrat govorimo o deležih, drugič pa posameznih posameznikih.

P_{25} mora ležati nekje med vrednostjo X=15 in vrednostjo X=16 (oz. med P(0,23) in P(0,29)).

$$\frac{0{,}25-0{,}23}{0{,}29-0{,}23}=\frac{P_{25}-15{,}0}{16{,}0-15{,}0}\rightarrow P_{25}=15{,}3$$

P_{75} mora ležati med vrednostima X=23 in X=24, torej med P(0,73) in P(0,79).

$$\frac{0{,}75-0{,}73}{0{,}79-0{,}73}=\frac{P_{75}-23}{24-23}\rightarrow P_{75}=23{,}3 \text{ ali } x=23+\frac{(24-23)(20-19{,}5)}{21-19{,}5}=23+\frac{0{,}5}{1{,}5}=23{,}3^*$$

Primer 2. Poiščite absolutni rang od P_{75}. Upoštevajte podatke iz zgornje tabele.

$$R_X(P_{75})=0{,}75\cdot 26+0{,}5=20^*$$

Primer 3. Iščemo vrednost x, kateri odgovarja absolutni rang 20.

$$20=19{,}5+\frac{x-23}{24-23}\rightarrow 0{,}5=x-23\rightarrow 23{,}5=x$$

Dosedanje enačbe

$P = \dfrac{R - 0.5}{N}$	[1] določanje percentilnega ranga iz absolutnega ranga podatka v ranžirni vrsti (za grupirane podatke brez -0.5)
$\dfrac{R_x - R_0}{R_1 - R_0} = \dfrac{x - x_0}{x_1 - x_0}$	[2] spremembe rangov so sorazmerne spremembam vrednosti
$R_x = R_0 + \dfrac{x - x_0}{x_1 - x_0}$	[3] določanje absolutnega ranga v ranžirni vrsti; R_x – absolutni rang vrednosti x, R_0 – predhodni absolutni rang, x_0 - predhodna vrednost v vrsti, x_1 – naslednja vrednost v vrsti

Pomembnejši percentili

- kvartili ($P_{25} = Q_1$, $P_{50} = Q_2$, $P_{75} = Q_3$)
- decili ($P_{10} = D_1$, $P_{20} = D_2$ …)
- mediana ($P_{50} = Q_2 = D_5 = Mdn$)

GRUPIRANI PODATKI

Določanje percentilnih rangov in percentilov pri grupiranih podatkih

$$\frac{cf(X) - cf(\mathrm{SM_j})}{cf(\mathrm{ZM_j}) - cf(\mathrm{SM_j})} = \frac{X - \mathrm{SM_j}}{\mathrm{ZM_j} - \mathrm{SM_j}}$$

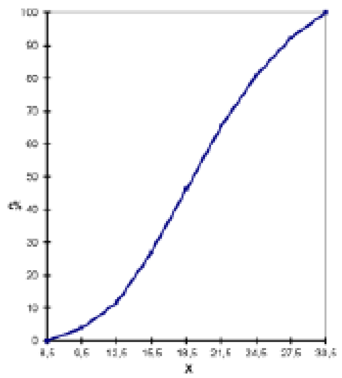
$$cf(X) = cf(\mathrm{SM_j}) + \frac{X - \mathrm{SM_j}}{\mathrm{ZM_j} - \mathrm{SM_j}}\left(cf(\mathrm{ZM_j}) - cf(\mathrm{SM_j})\right)$$

$$cf(X) = cf(\mathrm{SM_j}) + \frac{f_j(X - \mathrm{SM_j})}{i}$$

$$cf_X \rightarrow cp_X$$

$$X_P = \mathrm{SM_j} + \frac{i \cdot (R(X) - cf_{\mathrm{SM_j}})}{f_j}$$

Razred j	X_j'	f_j	%	cf_{sp}	cf_{zg}	$c\%_{sp}$	$c\%_{zg}$
7 - 9	8	1	3.9	0	1	0	3.9
10 - 12	11	2	7.7	1	3	3.9	11.6
13 - 15	14	4	15.4	3	7	11.6	27.0
16 - 18	17	5	19.2	7	12	27.0	46.2
19 - 21	20	5	19.2	12	17	46.2	65.4
22 - 24	23	4	15.4	17	21	65.4	80.8
25 - 27	26	3	11.5	21	24	80.8	92.3
28 - 30	29	2	7.7	24	26	92.3	100.0
		26	100.0	26		100.0	



Podatki:

SM	ZM	SM(nat)	ZM(nat)	X'	f	cf(sm)	cf(zg)	cp(sm)	cp(zm)
7,00	9,00	6,50	9,50	8	1	0,00	1,00	0	0,038
10,00	12,00	9,50	12,50	11	2	1,00	3,00	0,038	0,115
13,00	15,00	12,50	15,50	14	4	3,00	7,00	0,115	0,269

16,00	18,00	15,50	18,50	17	5	7,00	12,00	0,269	0,461
19,00	21,00	18,50	21,50	20	5	12,00	17,00	0,461	0,653
22,00	24,00	21,50	24,50	23	4	17,00	21,00	0,653	0,807
25,00	27,00	24,50	27,50	26	3	21,00	24,00	0,807	0,923
28,00	30,00	27,50	30,50	29	2	24,00	26,00	0,923	1,000

$$cf_{x'} = cf_{zm} - {}^{f}/_{2} = cf_{sm} + {}^{f}/_{2} = \frac{cf_{zm} + cf_{sm}}{2}$$

Zgled: Izračunaj P_{25} iz grupiranih podatkov.

$$x_p = x_0 + \frac{i(R_p - F_0)}{f_o} \rightarrow R_p = NP$$
$$R_p = 26 * 0{,}25 = 6{,}5$$
$$x_p = 12{,}5 + \frac{3*(6{,}5-3)}{4} = 12{,}5 + 2{,}625 = 15{,}125$$

P_{75} = 23.4 (R = 19.5)
D_1 = 11.9 (R = 2.6)
$D_5 = P_{50}$ = 19.1 (R = 13)
D_9 = 26.9 (R = 23.4)

Enačbe za grupirane podatke:

$\frac{R_x - F_0}{f_0} = \frac{x - x_0}{i}$	[4] izpeljava [2] za grupirane podatke; F_0 – kumulativna absolutna frekvenca do spodnje meje razreda, f_0 – frekvenca v razredu, x_0 – spodnja meja razreda, i – širina razreda
$R_x = F_0 + \frac{f_0(x - x_0)}{i}$	[5] določanje absolutnega ranga pri grupiranih podatkih
$R_p = NP + 0.5$	[6] določanje absolutnega ranga percentila za podatke v ranžirni vrsti (za grupirane podatke brez +0.5)
$x_p = x_0 + \frac{(x_1 - x_0)(R_p - R_0)}{R_1 - R_0}$	[7] določanje percentila iz ranžirne vrste
$x_p = x_0 + \frac{i(R_p - F_0)}{f_0}$	[8] določanje percentila pri grupiranih podatkih

natančna SM kvantilnega razreda … x_0 **$R(x_0) = F_0$**
natančna ZM kvantilnega razreda … x_1 **$R(x_1) = F_1$**

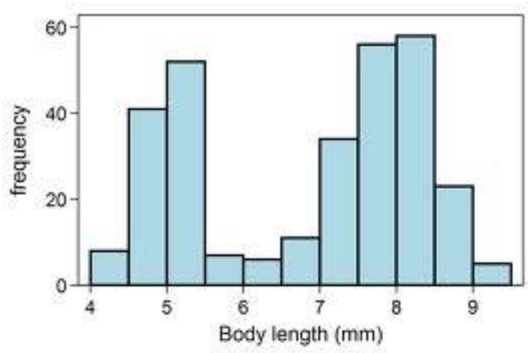
Arbitrarnost percentilnih rangov in percentilov

- Dogovor, na katero vrednost kumuliramo (ZM, X') → percentilni rangi so lahko malenkost različni
- Pri grupiranih podatkih različna določitev SM prvega razreda in širine razreda rezultira v različnih histogramih in tudi krivuljah kumulativnih frekvenc → percentilni rangi so lahko malenkost različni

Opis povprečja

Modus

- v frekvenčni distribuciji je modus tista vrednost, ki ima največjo frekvenco (se največkrat pojavlja)
- bimodalna distribucija – dva vrha, ki sta oba modusa (ni nujno, da sta čisto enaka po vrednosti, en je lahko večji in drug manjši z manjšo razliko)
- multimodalna distribucija
- 💻 Excel: =MODE



Primer 1.

PRILJUBLJENE RESTAVRAVIJE	f
Emonska klet	10
Ajdovo zrno	12
Hollyfood	6
Joe Pena's	10
La Storia	9

Modus je Ajdovo zrno in ne frekvenca 12!

GRUPIRANI PODATKI

- pri grupiranih podatkih je modus sredina modalnega razreda
  - modus se nahaja na meji, sredini dveh srednjih sosednjih razredov.

$$Mo = SM_{Mo} + \frac{f_{Mo} - f_n}{2f_{Mo} - f_n - f_v} = 2{,}5 + \frac{7-6}{2 \cdot 7 - 6 - 5} = 2{,}5 + \frac{1}{3} = 2{,}8$$

$$\boxed{Mo = SM_{Mo} + \frac{f_{Mo} - f_n}{2f_{Mo} - f_n - f_v} i}$$

$SM_{Mo} \ldots spodnja\ meja\ modalnega\ razreda$

$f_{Mo} \ldots število\ rezultatov\ v\ modalnem\ razredu$

$f_n \ldots število\ rezultatov\ v\ razredu\ pod\ modalnim\ razredom$

$f_v \ldots število\ rezultatov\ v\ razredu\ nad\ modalnim\ razredom$

$i \ldots širina\ razreda$

Mediana

- Mdn (Me)
- Mdn=P_{50}=D_5=Q_2
- Je dosežek, ki skupino razdeli na dva dela z enakim številom elementov. Točno 50 % dosežkov je višjih oz. nižjih od Mdn
- Excel: =MEDIAN

Aritmetična sredina

- Artimetična sredina (*M*) je težišče frekvenčne porazdelitve.
- M grupiranih podatkov
- Aritmetična sredina vzorca … *M*
- Aritmetična sredina populacije … μ
- Excel: =AVERAGE

Aritmetična sredina

- *M* je težišče frekvenčne porazdelitve.

$$M = \frac{\sum_{i=1}^{N} X_i}{N}$$

- M grupiranih podatkov

$$M = \frac{\sum_{j=1}^{k} f_j X_j'}{N}$$

- Tehtana *M* $\xrightarrow{\text{več skupin z različnimi } n_k}$

$$M = \frac{\sum w_k X_k}{\sum w_k}$$

μ … aritmetična sredina populacije, *M* … aritmetična sredina vzorca

Aritmetična sredina grupiranih podatkov

X	X_i'	f_i
7 – 9	8	1
10 – 12	11	2
13 – 15	14	4
16 – 18	17	5
19 – 21	20	5
22 – 24	23	4
25 – 27	26	3
28 – 30	29	2

Opomba. Pri grupiranih podatkih ne vemo točno, kaj so vmesne vrednosti.

$$M = \frac{8 + 2 \cdot 11 + 4 \cdot 14 + 5 \cdot 17 + 5 \cdot 20 + 4 \cdot 23 + 3 \cdot 26 + 2 \cdot 29}{26} = \frac{499}{26} = 19{,}19$$

$$M = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{\sum f_i X_i'}{N}$$

Tehtana aritmetična sredina

Primer 1. Merimo povprečni dosežek na kratkem preizkusu znanja.

1. N=83	M=80,2
2. N=72	M=75,4
3. N=57	M=62,8

$$M_{tot} = \frac{83 \cdot 80{,}2 + 72 \cdot 75{,}4 + 57 \cdot 62{,}8}{83 + 72 + 57} = \frac{15665}{212} = 73{,}9$$

Gre za tehtane sredine - je sredina, ki je obtežena, saj velikost vzorca (w_k) deluje kot neka obtežitev. Aritmetične sredine ne moremo kar povprečiti, ampak jih moramo obtežiti.

$$M = \frac{\sum w_k X_k}{\sum w_k}$$

Primer 2. Iščemo demostratorja (1., 2. ali 3.), osebo z najboljšimi ocenami. Ni vseeno, kako težimo, ker predmeti niso enako pomembni.

	w	1.	2.	3.
OS	3	8	10	8
SZ	5	8	9	9
MPR	1	9	6	8

$$M_1 = \frac{3 \cdot 8 + 5 \cdot 8 + 1 \cdot 9}{3 + 5 + 1} = 8{,}1 \qquad M_2 = 9{,}0 \qquad M_3 = 8{,}6$$

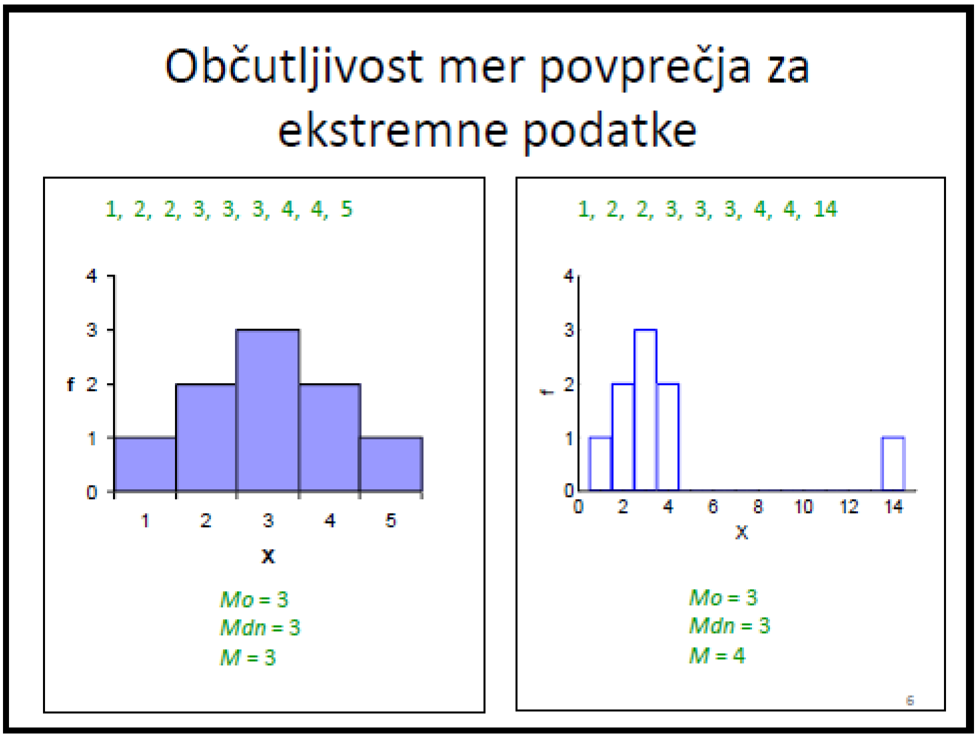
Izbrali bi drugo osebo.

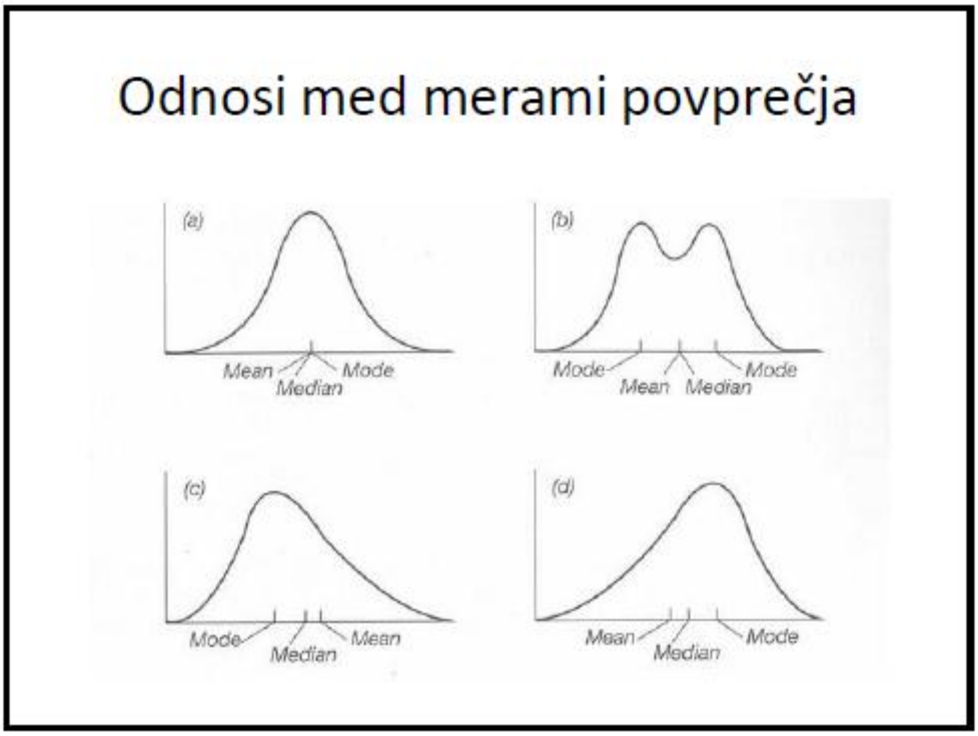
Lastnosti M:

- Primerna je samo za intervalne in razmernostne spremenljivke (ne pa za ordinalne – pri podatkih te vrste uporabljamo mediano, ker natančne razdalje niso pomembne)
- $\sum(X - M) = 0$; odkloni sredine se morajo vedno sešteti do nič, drugače to ni aritmetična sredina
- Ob transformaciji (premiku) lestvice:
  - Če vsakemu podatki prištejemo konstanto *c*, se *M* zveča za to konstanto
  - Če vsak podatek pomnožimo s konstanto *c*, se *M* poveča za c-krat
- Točka zloma pri aritmetični sredini je 0 %, M oz. µ se spremeni za vsakega posameznika

Občutljivost mer povprečja za ekstremne podatke

Aritmetična sredina je zelo občutljiva za ekstremne podatke, modus manj, na mediano pa nimajo vpliva, zato raje uporabljamo mediano, saj posamezni primeri lahko preveč izkrivijo aritmetično sredino.

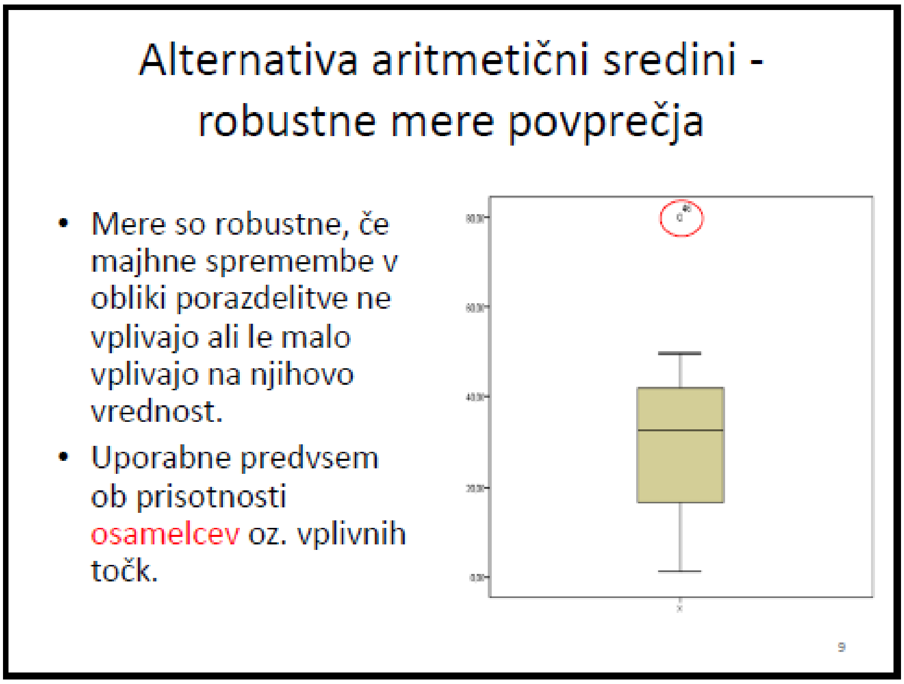




- Indikator normalnosti (normalne porazdelitve): Mdn, Mo in M so enaki
- Pri desno asimetričnih: kjer je vrh je Mo, M je skrajno proti ekstremu, Mdn je med njima
- Pri levo asimetričnih: M nizke ekstremne vrednosti vlečejo na levo, Mo je premaknjena pod vrh, med njima je Mdn

Kdaj uporabimo katero od osnovnih mer povprečja?

- Mo
    - Pri nominalnih spremenljivkah
    - Pogost je tudi pri diskretnih spremenljivkah
    - Tudi pri ordinalnih spremenljivkah, ga lahko uporabljamo
        - Ni precej natančna mera
- Mdn
    - Pri ordinalnih spremenljivkah
    - Pri intervalnih spremenljivkah
        - Če je nekaj podatkov ekstremnih oz. je porazdelitev asimetrična
        - Če nekateri dosežki nimajo določljivih vrednosti
        - Če je porazdelitev na enem repu odprta (npr. razred od 28 naprej, aritmetične sredine ne bi mogli računati, uporabljamo Mdn)
- M
    - Pri intervalnih (in razmernostnih) spremenljivkah v preostalih primerih (skupaj s SD … standardno deviacijo)



Robustne mere povprečja:

- Mediana
- Prirezana aritmetična sredina (trimmed mean) – odrežemo določen % najvišjih in najnižjih vrednosti (5-20 %; npr. 2,5 % na levi in 2,5 % na desni strani), in aritmetično sredino izračunamo na ostalih npr. 95 % primerov
- Winsorizirane ocene – ekstremne podatke (10-25 %) zamenjamo s sosednjimi, a manj ekstremnimi vrednostmi, npr. pri 90 % winsoniziranju 5 % spodnjim vrednostim pripišemo vrednost 5. percentila in 5 % zgornjega 95. Percentila
    - Primer Mw: $\frac{x_2+x_2+x_3+x_4+x_5+x_6+x_7+x_7}{8}$

Poročanje:

- »Eksperimentalna skupina je pri reševanju nalog naredila manj napak (M=2,54) kot kontrolna (M=7,63)
- »Modialna kategorija so bili otroci z enim sorojencem (takih je bilo 35 od 50).«

Geometrična sredina

$$G = \sqrt[n]{X_1 X_2 X_3 \dots X_n}$$

- Primerna je za intervalne in razmernostne spremenljivke
- G pove, katero število nam da enako vrednost v n-kratnem množenju
- Uporabna za računanje povprečne hitrosti sprememb (relativne rasti)
  - Skupni verižni indeks
  - Povprečni verižni indeks
- Excel: =GEOMEAN za geometrično sredino in povprečni verižni indeks

Primer 1.

1. merjenje: 10 mišk (× 2/3)
2. merjenje: 15 mišk (× 2)
3. merjenje: 30 mišk (× 2,33)
4. merjenje: 70 mišk

Kakšna je povprečna hitrost spremembe? Indeksi nam to pokažejo.

$$I_{2/1} = \frac{15}{10} \cdot 100 = 150$$

Sprememba od 1. do 2. merjenja je bila 150.
Sprememba od 2. do 3. merjenja je bila 200.
Sprememba od 3. do 4. merjenja je bila 233.
➥ Gre za 3 verižne indekse.

$$I_{4/1} = \frac{70}{10} \cdot 100 = 700$$

Skupni verižni indeks, indeks spremembe od 1. do 4. obdobja (zgoraj).

$$70 = 2{,}33 \cdot 30 = 2{,}33 \cdot (2 \cdot 15) = 2{,}33 \cdot 2 \cdot (1{,}5 \cdot 10) = 2{,}33 \cdot 2 \cdot 1{,}5 \cdot 10$$

$\Delta_{4/1} = 2{,}33 \cdot 2 \cdot 1{,}5 = 7$ … *ker je 10 mišk ×7 enako 70 mišk*

$I_{sk} = \frac{150}{100} \cdot \frac{200}{100} \cdot \frac{233}{100} \cdot 100 = 700$ … *skupni verižni indeks*

Povprečni verižni indeksi:

$\Delta_{2/1} = 1{,}5$

$\Delta_{3/2} = 2$

$\Delta_{4/3}= 2{,}33$

$\overline{\Delta_{4/1}}= \sqrt[3]{1{,}5 \cdot 2 \cdot 2{,}33} = 1{,}91$ … *ker je táko povprečno povečanje ob vsakem merjenju*

Po domače: $\bar{I} = \sqrt[št.indeksov]{zmnožek\ vseh\ posamičnih\ indeksov}$

$\overline{I_{4/1}} = \sqrt[3]{\frac{700}{100}} \cdot 100 = 191$ … *povprečni verižni indeks*

1. merjenje: 10 mišk (× 2/3)	10 (×1,91)
2. merjenje: 15 mišk (× 2)	19,1 (×1,91)
3. merjenje: 30 mišk (× 2,33)	36,5 (×1,91)
4. merjenje: 70 mišk	70

Primer 2. S kakšno hitrostjo se je večalo št. listov zapiskov?

	Strani zapiskov	I	$I_{50/10}$
10. dan	10		
20. dan	22		
30. dan	28		
40. dan		$I_{40/30}$=152	
50. dan		$I_{50/40}$=92	

a. Na 40. dan je 42,56 strani zapiskov.

$28 \cdot 1{,}52 = 42{,}56$

b. Na 50. dan pa 39, 16 strani zapiskov.

$42{,}56 \cdot 0{,}92 = 39{,}16$

$\Delta_{sk}= 3{,}916 \rightarrow I_{sk} = 391{,}6$

$\bar{\Delta}= \sqrt[4]{3{,}916} = 1{,}41$ $\overline{I_{50/10}} = 141$

$\bar{I} = 1{,}41 \rightarrow 141$

Primer 3.

Leta 1990 je bilo na 1.200.000 delovno sposobnih ljudi 40.000 brezposelnih, leta 1991 pa na 1.210.000 delovno sposobnih 65.000 brezposelnih. Kakšen je indeks rasti stopnje brezposelnosti?

delovno sposobni	brezposelni	delež
1200000	40000	0,033333
1210000	65000	0,053719

$$I = \frac{0{,}053719}{0{,}033333} \cdot 100 = 161$$

Indeks rasti stopnje brezposelnosti je 161.

Geometrična sredina

$$G = \sqrt[n]{X_1 X_2 X_3 \ldots X_n}$$

- Primerna za intervalne in razmernostne spremenljivke.
- Uporabna za računanje povprečne hitrosti sprememb (relativne rasti)
  - Skupni verižni indeks $I_{sk} = \prod_{i=1}^{k} \frac{I_{(i+1)/i}}{100} \times 100$
  - Povprečni verižni indeks $\overline{I}_{(k+1)/i} = \sqrt[k]{\frac{I_{sk}}{100}} \cdot 100$

Harmonična sredina

$$H = \frac{N}{\sum_{i=1}^{N} \frac{1}{X_i}}$$

Povprečevanje koeficientov (hitrosti):

- *H*: kadar je konstanten števec koeficienta
- *M*: kadar je konstanten imenovalec

Kvadratna sredina

- RMS (angl. *root mean square*)

$$RMS = \sqrt{\frac{\sum X^2}{N}} = \sqrt{\overline{X^2}}$$

Harmonična sredina

Primer 1.

Študent se vsak dan vozi v Ljubljano in nazaj, od nje živi stran 1 uro. Če se v Ljubljano vozi s 90 km/h, nazaj pa s 60 km/h. Kolikšna je njegova povprečna hitrost.

$$H = \frac{2}{\frac{1h}{90km} + \frac{1h}{60km}} = 72\frac{km}{h}$$

Pazi: 90 km od LJ štartamo in se vozimo 1 h, nato nadaljujemo vožnjo 60 km/h in se peljemo 1 h.

$$M = \frac{90 + 60}{2} = \frac{150}{2} = 75\frac{km}{h}$$

Če je razdalja konstantna razdalja, uporabimo harmonično sredino; če je pa konstanten čas, je pa primerna aritmetična sredina.

Primer 2.

Imamo dva lista besed. Na vsakem je 40 besed. Otroci berejo besede od začetka do konca in od konca do začetka. Hitrost branja naprej je 20 bes/min, nazaj pa 8 bes/min.

Naprej berejo 2 min, nazaj pa 5 min. Cel list preberejo v 7 min.

Primer 3.

V neki slaščičarni so štiri leta zapored peki tort namenili 1.500 €. V prvem letu so za eno torto porabili 3 €, drugo leto 5 €, tretje leto 5,5 € in četrto leto 7,5 €. Koliko tolarjev so v povprečju v štirih letih porabili za eno torto?

leto	eur/torta
1	3
2	5
3	5,5
4	7,5

$$H = \frac{4}{\frac{1}{3} + \frac{1}{5} + \frac{1}{5{,}5} + \frac{1}{7{,}5}} = 4{,}7$$

Opis razpršenosti

Mere razpršenosti

- variacijski razmik

  VR = X_{max} - X_{min} + 1

- povprečni absolutni odklon od povprečja

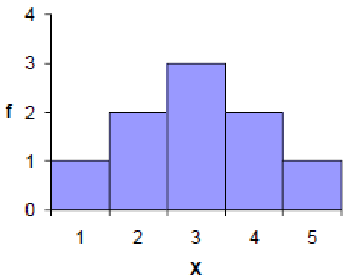
  $$\mathrm{PO_M} = \frac{\sum_{i=1}^{N} |X_i - M|}{N}$$

- varianca

  $$SD^2 = \frac{\sum_{i=1}^{N} (X_i - M)^2}{N}$$

- standardna deviacija

  $$SD = \sqrt{\frac{\sum_{i=1}^{N} (X_i - M)^2}{N}}$$



X	X-M	abs(X-M)	$(X-M)^2$
1	-2	2	4
2	-1	1	1
2	-1	1	1
3	0	0	0
3	0	0	0
3	0	0	0
4	1	1	1
4	1	1	1
5	2	2	4
vsota	0	8	12

N = 9

M = 3

VR = 5

$\mathrm{PO_M}$ = 8/9

var = 12/9

$$SD = \sqrt{\frac{12}{9}} = 1{,}15$$

2

Variacijski razmik

- Pove o razponu vrednosti
- Natančne vrednosti: VR = Xmax – Xmin = 5,5–0,5 = 5
- Nenatančne vrednosti: VR = Xmax – Xmin + 1 = 5–1+1 = 5
- Gre za enostavno mero razpršenosti, ki pove, kako razpršeni so podatki
- Primerno za ordinalne, intervalne in razmernostne spremenljivke

Variacijski razmik je zelo občutljiv na posamezne vrednosti, zato je včasih bolj primeren povprečni absolutni odklon povprečja. Povprečni absolutni odklon je povprečna oddaljenost rezultatov. Odklone lahko seštevamo le v absolutni obliki! Drugače se odštejejo v 0.

X	|X-M|	X-M	$(X\text{-}M)^2$
2	1,2	-1,2	1,44
4	0,8	+0,8	0,64
5	1,8	+1,8	3,24
3	0,2	-0,2	0,04
2	1,2	-1,2	1,44
M=3,2	∑=5,2	∑=0	

Povprečni absolutni odklon od povprečja

$$PO_M = \frac{\sum|X-M|}{N} = \frac{5,2}{5} = 1,04$$

Povprečnega absolutnega odklona se poslužujemo le pri **intervalnih** in **razmernostnih** spremenljivkah. Vsebuje namreč aritmerično sredino (M), ki jo ne moremo uporabljati na ordinalnih ali nominalnih spremenljivkah. To pa ne velja za prej omenjeni variacijski razmik ali v nadaljevanju omenjen interkvartilni razmik.

Standardna deviacija in varianca

SD pove, ali so vrednosti na splošno blizu ali daleč od aritmetične sredine. Deviacija je razdalja vrednosti od sredine ($X - M$) in je lahko pozitivna ali negativna (odvisno ali je manjša ali večja od M. Vrednosti deviacije se med sabo seštejejo (in odštejejo) v nič ($\sum(X-M) = 0$) zato ne moremo na tak način priti do informacije o deviaciji (razpršenosti). Zato vrednosti deviacije kvadriramo in se izognemo negativnemu predznaku [$(X-M)^2$]. Če kvadrirane deviacije seštejemo in delimo z številom vrednosti (N) dobimo varianco [$var = \frac{\sum(X-M)^2}{N}$]. Standardna deviacija je koren iz variance [$SD = \sqrt{var} = \sqrt{\frac{\sum(X-M)^2}{N}}$].

$$var = \frac{\sum(X-M)^2}{N} = \frac{6,8}{5} = 1,36$$

$$var = SD^2 \rightarrow SD = \sqrt{var} = \sqrt{\frac{\sum(X-M)^2}{N}} \rightarrow SD = \sqrt{1,36} = 1,17$$

Mere razpršenosti

- variacijski razmik

  $VR = X_{max} - X_{min} + 1$

- povprečni absolutni odklon od povprečja

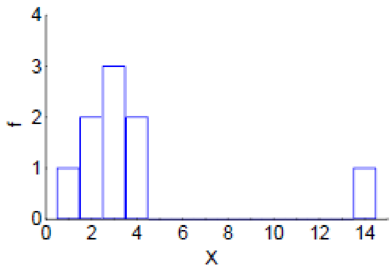
  $$PO_M = \frac{\sum_{i=1}^{N} |X_i - M|}{N}$$

- varianca

  $$SD^2 = \frac{\sum_{i=1}^{N} (X_i - M)^2}{N}$$

- standardna deviacija

  $$SD = \sqrt{\frac{\sum_{i=1}^{N} (X_i - M)^2}{N}}$$



X	X-M	abs(X-M)	$(X-M)^2$
1	-3	3	9
2	-2	2	4
2	-2	2	4
3	-1	1	1
3	-1	1	1
3	-1	1	1
4	0	0	0
4	0	0	0
14	10	10	100
vsota	0	20	120

$N = 9$

$M = 4$

VR = 14

$PO_M = 20/9$

$var = 120/9$

$$SD = \sqrt{\frac{120}{9}} = 3{,}65$$

3

Petico smo zamenjali s 14. Lahko opazimo, kako to vpliva na mere razpršenosti. Vse te mere so zelo občutljive na posamezne odstopajoče skupine. Sredino in standardno deviacijo uporabljamo, kadar imamo pribl. normalno porazdelitev, saj bi drugače lahko en osamljen ekstremen podatek popačil sliko.

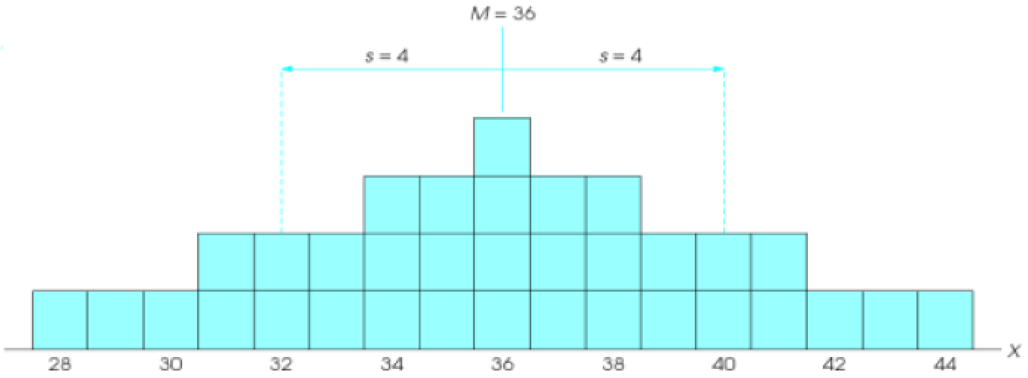


Figure 4-9 (p. 125)

A sketch of a distribution for a sample with a mean of M = 36 and a standard deviation of s = 4. Notice that the distribution is centered around 36 and that most of the scores are within a distance of 4 points from the mean, although some scores are farther away.

Statistics for the Behavioral Sciences, Sixth Edition by Frederick J. Gravetter and Larry B. Wallnau
Copyright © 2004 by Wadsworth Publishing, a division of Thomson Learning. All rights reserved.

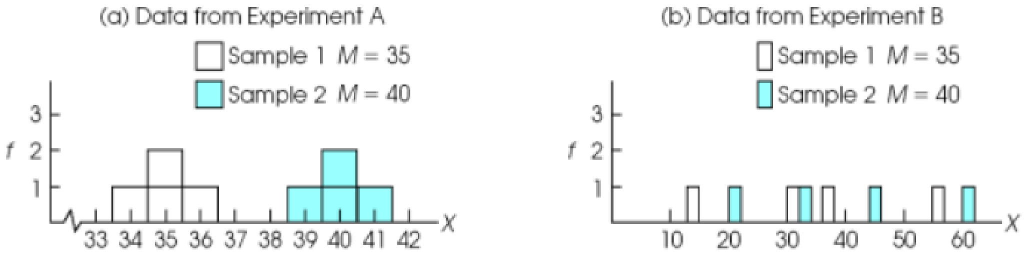


Figure 4-10 (p. 127)

Graphs showing the results from two experiments. In Experiment A, the variability within samples is small and it is easy to see the 5-point mean difference between the two samples. In Experiment B, however, the 5-point mean difference between samples is obscured by the large variability within samples.

Statistics for the Behavioral Sciences, Sixth Edition by Frederick J. Gravetter and Larry B. Wallnau
Copyright © 2004 by Wadsworth Publishing, a division of Thomson Learning. All rights reserved.

Lastnosti SD

- Ob transformacijah lestvice:
  - če vsakemu dosežku prištejemo konstanto, se SD ne spremeni
  - če vsak dosežek pomnožimo s konstanto c, se SD poveča za c-krat, varianca pa za c^2-krat
- Občutljiva na ekstremne vrednosti
- Neizračunljiva pri porazdelitvah z odprto zadnjo kategorijo.

Primer A:

X	|X-M|	X-M	$(X-M)^2$
2	1,2	-1,2	1,44
4	0,8	+0,8	0,64
5	1,8	+1,8	3,24
3	0,2	-0,2	0,04
2	1,2	-1,2	1,44
M=3,2	∑=5,2	∑=0	∑=6,8

$$SD_A = \sqrt{\frac{\sum(X-M)^2}{N}} = \sqrt{\frac{1{,}44+0{,}64+3{,}24+0{,}04+1{,}44}{5}} = \sqrt{\frac{6{,}8}{5}} = 1{,}17$$

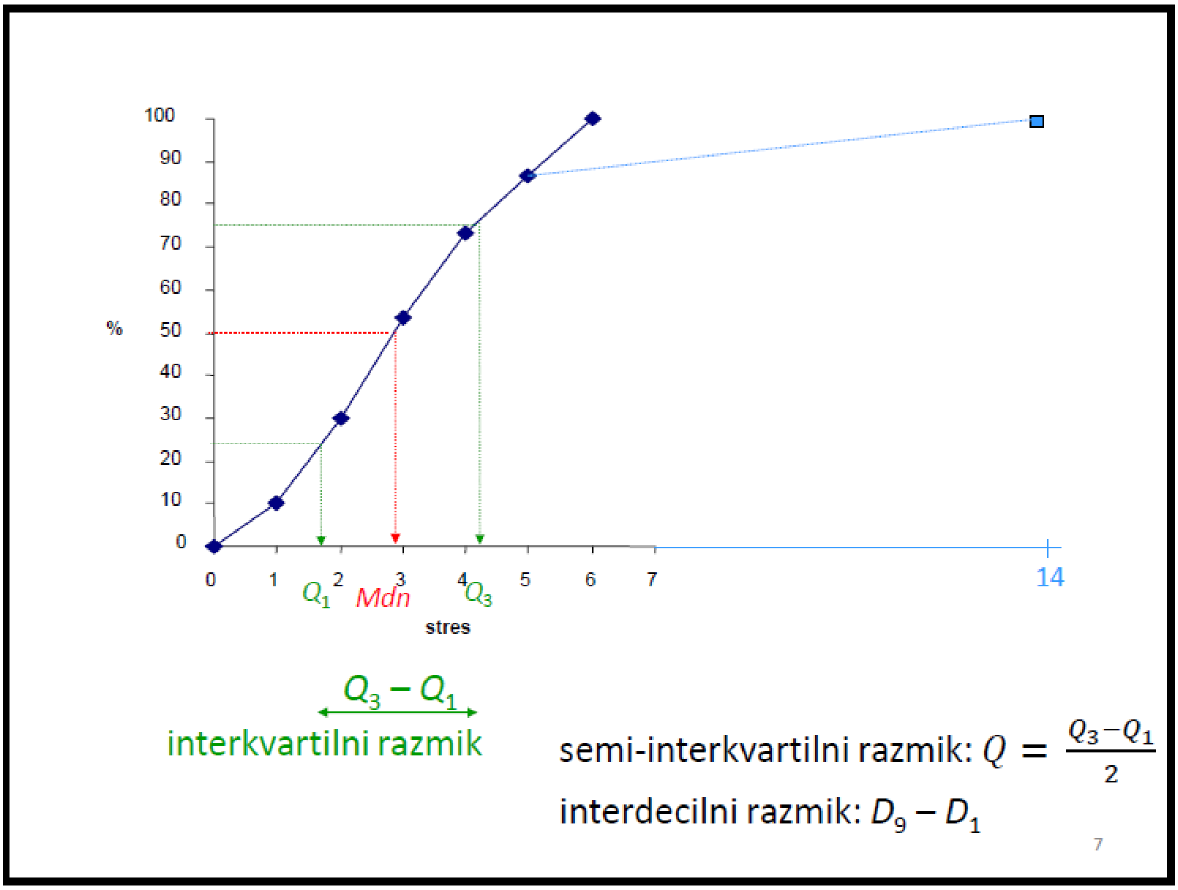
Primer B: $\boxed{c=2}$

2*X	|X-M|	X-M	$(X-M)^2$
4			5,76
8			2,56
10			12,96
6			0,16
4			5,76
M=6,4			∑=27,2

$$SD_B = \sqrt{\frac{27{,}2}{5}} = 2{,}33$$

$$\boxed{SD_B = 2 \cdot SD_A}$$

Interkvartilni in interdecilni razmik



Robustne mere razpršenosti, ki jih najpogosteje uporabljamo so:

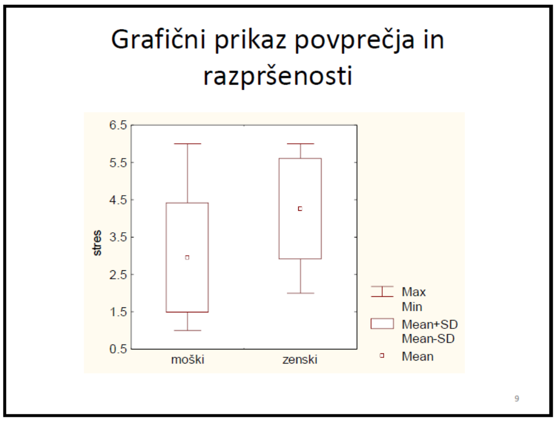
- interdecilni razmik – razlika med 90. percentilom in 10. percentilom
- interkvartilni razmik – razlika med tretjim in prvim kvartilom (Q_3–Q_1)
- semi-interkvartilni razmik – polovica razdalje interkvartilnega razmika

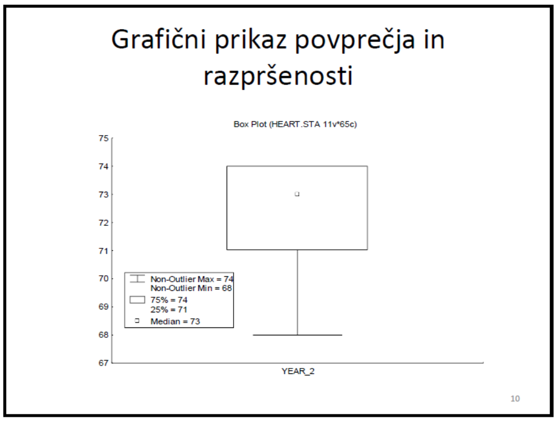
Interkvartilni razmik predstavlja odklon od mediane, ni pa nujno, da je mediana točno na sredini med prvim in tretjim kvartilom. Odklon od mediane opisuje približno (glej sliko).

Gre za mere, ki so neobčutljive na ekstreme. Predvsem interkvartilni razmik ne, intertercilni je nekoliko bolj.

Poročanje

- »Otroci staršev z nizko izobrazbo so uporabljali več enostavčnih povedi (M=4,12, SD=0,87) kot otroci staršev z visoko izobrazbo (M=2,42, SD=0,75)« - če bi obstajali ekstremni primeri, navajamo mediano in semi-inetrkavrtilni razmik (Mdn=, Q=)
  - M in SD
  - Mdn in semi-interkvartilni razmik (Q)





Zaboji z ročaji: zaboji prikazujejo standardno deviacijo, ročaji pa minimalne in maksimalne vrednosti.

Grupirani podatki

- Povprečni absolutni odklon

$$\mathrm{PO}_M = \frac{\sum f_j |X_j - M|}{N}$$

- Varianca

$$SD^2 = \frac{\sum f_j (X_j - M)^2}{N}$$

→ Sheppardov popravek za varianco: $SD_c^2 = SD^2 - \frac{i^2}{12}$

Tehtana varianca

- kadar je M enaka v vseh skupinah

$$SD^2 = \frac{\sum N_k SD_k^2}{N}$$

- kadar M ni enaka

$$var_{tot} = M_{var} + var_M = \frac{\sum N_k var_k}{N} + \frac{\sum N_k (M_k - M_{tot})^2}{N}$$

M_{tot} ... tehtana (totalna) aritmetična sredina

X	X_i'	f_i	|X−M|	$(X-M)^2$	$f \times (X-M)^2$
7 – 9	8	1	11,19	125,22	125,23
10 – 12	11	2	8,19	67,11	134,23
13 – 15	14	4	5,19	26, 96	107,84
16 – 18	17	5	2,19	4,80	24,03
19 – 21	20	5	0,81	0,65	3,26
22 – 24	23	4	3,81	14,50	57,99
25 – 27	26	3	6,81	46,34	139,03
28 – 30	29	2	9,81	96,19	192,38
		N=26			
		M=19,19			

$$PO_M = \frac{|8-19{,}19| + 2\cdot|11-19{,}19| + 4\cdot|14-19{,}19| + \cdots + 2\cdot|29-19{,}19|}{26} = 4{,}56$$

$$var = \frac{\sum f \cdot (X-M)^2}{N} = \frac{125{,}23 + 134{,}23 + \cdots + 190{,}38}{26} = 30{,}15$$

$$SD = \sqrt{var} = \sqrt{30{,}15} = 5{,}49$$

SD moramo popraviti, če so podatki normalno razporejeni, mi jih pa grupiramo. Varianco korigiramo tako: $var_c = 30{,}15 - 3^2/12 = 29{,}40$.

Tehtana varianca

Kadar so aritmetične sredine enake:

$N_1=8$	$N_2=5$
$M_1=3$	$M_2=3$
$var_1=1,5$	$var_2=2$

$$var_{teh} = \frac{\sum N_k \cdot var_k}{\sum N_k} = \frac{8 \cdot 1,5 + 2 \cdot 5}{8+5} = 1,69$$

Kadar aritmetične sredine niso enake:

$N_1=8$	$N_2=5$
$M_1=3$	$M_2=4$
$var_1=1,5$	$var_2=2$

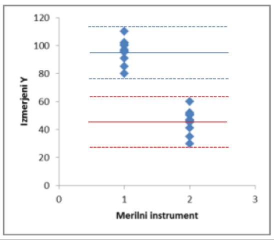
$$M_{tot} = \frac{\sum(N_k * M_k)}{\sum N_k}$$

$$M_{tot} = \frac{8*3+4*5}{8+5} = 3,38$$

$$var_{tot} = \frac{\sum N_k * var_k}{N} + \frac{\sum N_k (M_k - M_{tot})^2}{N} = \frac{8*1,5+5*2}{13} + \frac{8*(3-3,38)^2 + 5*(4-3,38)^2}{13}$$
$$= 1,69 + 0,24 = 1,93$$

Koeficient variacije

$$KV = \frac{SD}{M}(100\,\%)$$



smiseln le pri razmernostnih spremenljivkah
→ absolutna ničla

Standardizirani odklon

$$z = \frac{X_i - M}{SD}$$

Pove nam,

- za koliko standardnih deviacij je nek rezultat drugačen od aritmetične sredine,
- ali je rezultat nad- ali podpovprečen.

→ predstava položaja podatka v skupini

Koeficient variacije pove kakšno je relativno variiranje glede na povprečje podatkov. Uporablja se izključno za primerjave, nikoli kot osnovna mera razpršenosti.

Imeli smo 6-letne dečke. V povprečju so bili visoki 120,1 cm, težki pa 25,3 kg. Standardna deviacija višine dečkov je bila 6,05 cm, standardna deviacija teže pa 3,22 kg. Pri deklicah, starih 6 let smo izračunali enako. Povprečna višina je znašala 121,2 cm, teža 26,5 kg, standardna deviacija 6,51 cm in 4,82 kg. Kateri bolj variirajo v višini in kateri v teži?

$$KV_{v(f)} = \frac{6,05\ cm}{120,1\ cm} = 5,0\ \%\ ;\ KV_{t(f)} = \frac{3,22\ kg}{25,3\ kg} = 12,7\ \%\ ;\ KV_{v(p)} = 5,4\ \%\ ;\ KV_{t(p)} = 18,2\ \%$$

In v višini in v teži deklice bolj variirajo. → Taka razmerja lahko opazujemo samo, kadar imamo **absolutno ničlo**. Ne npr. pri IQ!!

Standardiziran odklon nekega podatka pove, za koliko se ta podatek odklanja od neke aritmetične sredine.

X	Z
2	−1,03
4	+0,68
5	+1,54
3	−0,17
2	−1,03
M=3,2	
SD=1,17	

$$Z = \frac{2 - 3{,}2}{1{,}17} = \frac{-1{,}2}{1{,}17} = -1{,}03$$

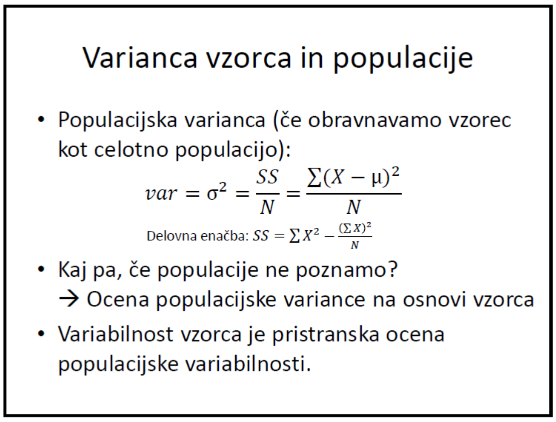
$$Z = \frac{4 - 3{,}2}{1{,}17} = 0{,}68$$

Povprečen podatek ima Z vrednost Z=0. Z=−1 pomeni da je podatek eno standardno deviacijo pod povprečjem. Ugoden je za primerjavo podatkov v različnih vzorcih in skupinah podatkov.

Pišemo test, rezultati so:

M	SD	X
10	2	8
15	3	12

Na obeh testih enko dobro!! Na obeh 1 SD od povprečja.



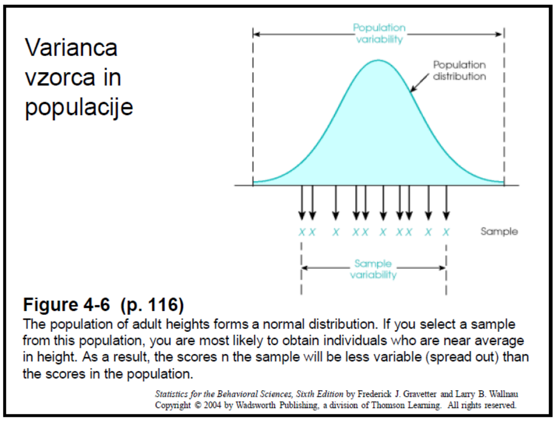


Figure 4-6 (p. 116)
The population of adult heights forms a normal distribution. If you select a sample from this population, you are most likely to obtain individuals who are near average in height. As a result, the scores n the sample will be less variable (spread out) than the scores in the population.

Statistics for the Behavioral Sciences, Sixth Edition by Frederick J. Gravetter and Larry B. Wallnau
Copyright © 2004 by Wadsworth Publishing, a division of Thomson Learning. All rights reserved.

$$SS = sum\ of\ squared\ deviations = \sum(X - M)^2$$

varianca je povprečna SS → $var = mean\ squared\ deviation = \frac{SS}{N}$

Varianca vzorca in populacije

- Ocena populacijske variance na osnovi vzorca (angl. *sample variance*):

$$var' = \sigma'^2 = s^2 = \frac{SS}{n-1} = \frac{\sum(X-M)^2}{n-1}$$

- S tem ko delimo SS z *n*-1, se vrednost zveča → boljša ocena (nepristranska ocena) populacijske variance.

	X
	1
	1
	5
	5
M	3
var	4
σ	2

vzorci (n = 2)		*M*	*var'*	σ'
1	1	1	0	0,00
1	5	3	8	2,83
5	1	3	8	2,83
5	5	5	0	0,00
povprečje		3	4	1,41

Na ta način pridemo do pravilnejše ocene populacijske variance. Dobimo boljšo nepristransko oceno variance populacije.

Opis oblike frekvenčne porazdelitve

Momenti

$$m_r = \overline{(X - A)^r}$$ ...*r*-ti moment okrog vrednosti A

$$A = 0 \Rightarrow \overline{X_i^r} = \frac{X_1^r + X_2^r + \ldots + X_N^r}{N} = \frac{\sum X_i^r}{N}$$ ... *r*-ti začetni moment (*r*-ti moment okrog ničle)

$$A = \overline{X} \Rightarrow m_r = \frac{\sum (X_i - \overline{X})^r}{N} = \overline{(X_i - \overline{X})^r}$$ ... *r*-ti centralni moment (*r*-ti moment okrog aritmetične sredine) ($m_1 = 0$, $m_2 = var$)

$$m_r = \frac{\sum f_j (X_j - \overline{X})^r}{N}$$ ... *r*-ti centralni moment grupiranih podatkov

$$a_r = \frac{m_r}{SD^r} = \frac{m_r}{(\sqrt{m_2})^r} = \frac{m_r}{\sqrt{m_2^r}} = \frac{\sum (X_i - \bar{X})^r}{N \cdot SD^r}$$ ... *r*-ti moment v brezdimenzionalni obliki ($a_1 = 0$, $a_2 = 1$)

$$SD = \sqrt{m_2}$$

Dokazi:

$$\sum (X - M) = 0$$

$$\boldsymbol{m_1} = \frac{0}{N} = 0$$

$$\boldsymbol{a_1} = \frac{0}{N \cdot SD} = 0$$

$$\boldsymbol{m_2} = \frac{\sum (X - M)^2}{N} = var = SD^2$$

$$\boldsymbol{a_2} = \frac{\sum (X - M)^2}{N \cdot SD^2} = \frac{\sum (X - M)^2 \cdot N}{N \cdot (X - M)^2} = \frac{SD^2}{SD^2} = 1$$

- prvi začetni moment okrog ničle = **aritmetična sredina** (glej drugo enačbo)
- drugi moment = kvadratna vrednost, tretji = kubična itn.
- drugi centralni moment okrog M = **varianca**
- tretji centralni moment v brezdimenzionalni obliki = **asimetričnost porazdelitve**
- četrti centralni moment v brezdimenzionalni obliki = **sploščenost porazdelitve**

Asimetričnost porazdelitve

$(-\infty, +\infty)$

$$\gamma_1 = a_3 = \frac{m_3}{\sigma^3} = \frac{m_3}{(\sqrt{m_2})^3} = \frac{\sum (X - \mu)^3}{N \cdot \sigma^3} = \frac{\sum z^3}{N}$$ ...As je tretji centralni moment v brezdimenzionalni obliki (Fisherjeva as.)

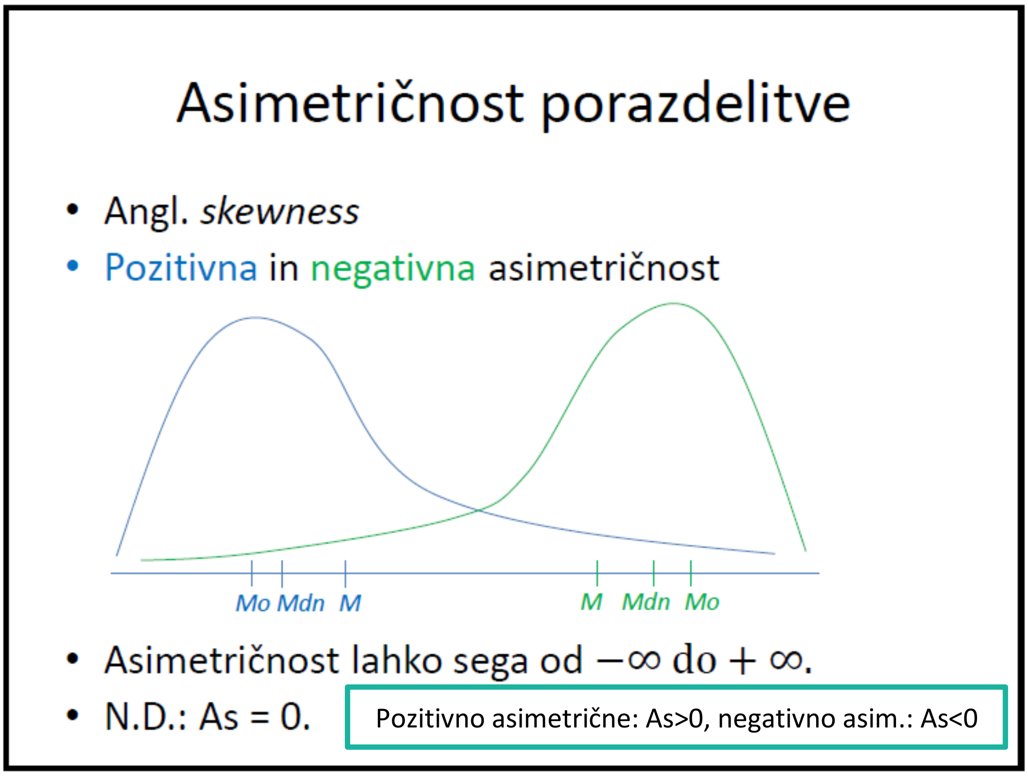
$$g_1 = \frac{N \sum z^3}{(N-1)(N-2)}$$ ... ocena Fisherjeve as. na osnovi vzorčnih podatkov

$$As = \frac{\mu - Mo}{\sigma}$$ ... Pearsonov 1. koeficient asimetrije

$$As' = \frac{3(M - Mdn)}{s}$$ ... Pearsonov 2. koeficient asimetrije

$$As = \frac{Q_3 - 2Q_2 + Q_1}{Q_3 - Q_1}$$ ... kvartilni koeficient asimetrije

$$As = \frac{P_{90} - 2P_{50} + P_{10}}{P_{90} - P_{10}}$$ ... 10-90 percentilni koeficient asimetrije



Porazdelitev je asimetrična, če je en rep krivulje daljši od drugega. Pozitivna asimetričnost je tista, pri kateri je daljši rep orientiran v pozitivno smer in obratno. Če je porazdelitev simetrična nima nobenega repa daljšega in ni pozitivna oz. negativna.

Asimetričnost porazdelitve izračunamo tako: $skew = \frac{\sum(X-M)^3}{N \cdot SD^3}$.

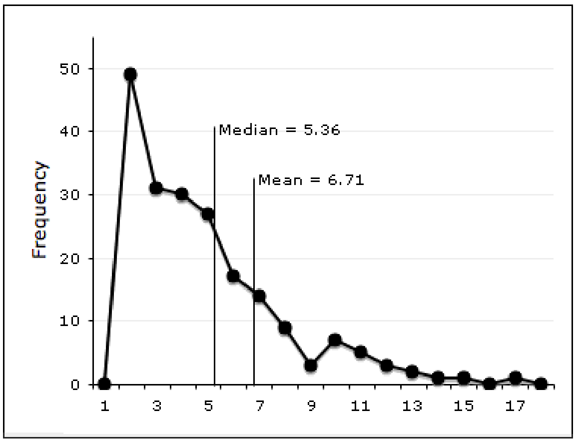
Pri normalni porazdelitvi je vrednost $\frac{\sum(X-M)^3}{N \cdot SD^3} = 0$.

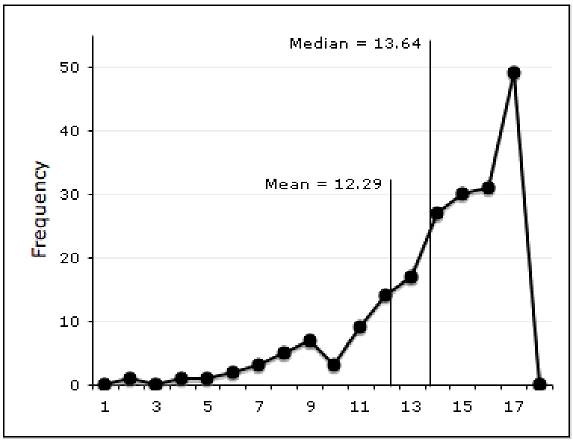
Da bi lahko zaključili, da neka porazdelitev ni normalna, mora biti As ≥ 0,5.

Oznake: $\mu = M\ (aritmetična\ sredina); \sigma = SD\ (standardna\ deviacija); As = skew$

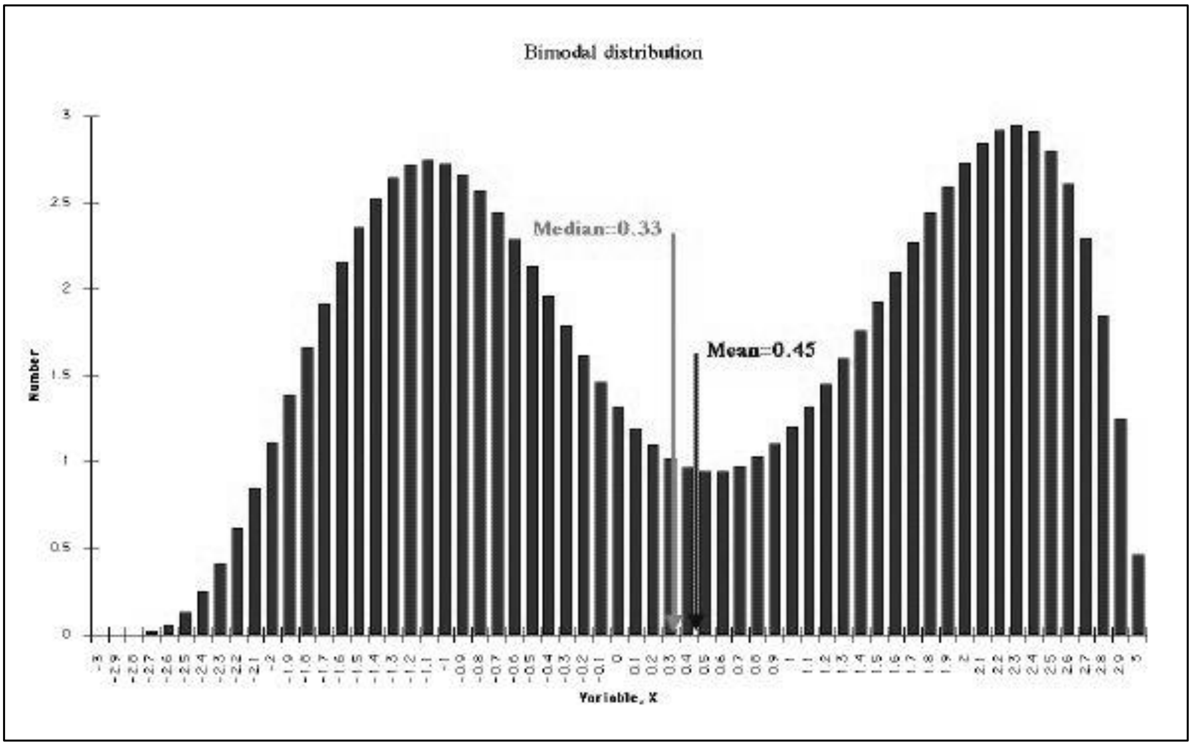
Če je aritmetična sredina večja od mediane gre za pozitivno (tudi desno) asimetričnost in obratno:

$\boxed{M > Mdn \ldots pozitivna\ asimetričnost}$





<u>Nasprotno</u> velja npr. za mnoge pozitivno/negativno asimetrične bimodalne distribucije (bimodalne, ker imajo dva modusa).



[−2, +∞)

Sploščenost porazdelitve

$$Spl = a_4 - 3 = \frac{m_4}{\sigma^4} - 3 = \frac{\sum (X-\mu)^4}{N \cdot \sigma^4} - 3 = \frac{\sum z^4}{N} - 3$$

$$g_2 = \frac{N(N+1)\sum z^4}{(N-1)(N-2)(N-3)} - \frac{3(N-1)^2}{(N-2)(N-3)}$$

... Spl je četrti centralni moment v brezdimenzionalni obliki
$a_4 = \beta_2$ ... Pearsonova sploščenost
γ_2 ... Fisherjeva sploščenost ($\beta_2 - 3$)
g_2 ... nepristranska ocena game 2

$$Spl_K = \frac{Q}{P_{90} - P_{10}}$$

... percentilni koeficient sploščenosti

Sploščenost porazdelitve (ang. kurtosis) je odvisna od velikosti porazdelitvenih repov. Porazdelitev, ki ima razmeroma dolge repe se imenuje leptokurtična, tista s kratkimi pa platikurtična. Normalna distribucija/porazdelitev pa je mezokurtična.

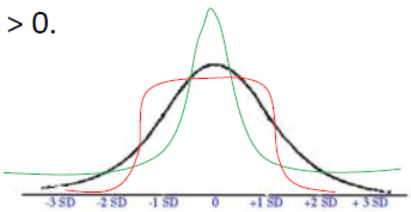
Sploščenost računamo po tej formuli: $kurtosis = \frac{\sum(X-M)^4}{N \cdot SD^4} - 3$. Pri normalni porazdelitvi je vrednost $\frac{\sum(X-M)^4}{N \cdot SD^4} - 3 = 0$.

Oznake: $\mu = M$ $(aritmetična\ sredina)$; $\sigma = SD$ $(standardna\ deviacija)$; $Spl = kurtosis$

Dve različni porazdelitvi imata lahko enako varianco in približno enako asimetričnost porazdelitve, ampak se pomembno razlikujeta v sploščenosti.

Sploščenost porazdelitve

- Sploščenost:
  - Normalna porazdelitev je mezokurtična: Spl = 0.
  - V primerjavi z normalno distribucijo ima platikurtična porazdelitev ravnejši vrh in krajše repe; Spl < 0,
  - leptokurtična porazdelitev pa ima ožji vrh in širše, gostejše repe; Spl > 0.
  - Možne vrednosti: -2 do +∞



Sploščenost porazdelitve

- Asimetrične porazdelitve so leptokurtične.
- Pri simetričnih porazdelitvah:
  - Večji Spl nakazuje unimodalnost krivulje.
  - Če je *var* približno konstantna, višanje sploščenosti nakazuje večje število oz. bolj ekstremne vrednosti.

Table 2.
Kurtosis for Seven Simple Distributions Not Differing in Variance

X	Freq. A	Freq. B	Freq. C	Freq. D	Freq. E	Freq. F	Freq. G
−6.6	0	0	0	0	0	0	1
−0.4	0	0	0	0	0	3	0
1.3	0	0	0	0	5	0	0
2.9	0	0	0	10	0	0	0
3.9	0	0	20	0	0	0	0
4.4	0	20	0	0	0	0	0
5	20	0	0	0	0	0	0
10	0	10	20	20	20	20	20
15	20	0	0	0	0	0	0
15.6	0	20	0	0	0	0	0
16.1	0	0	20	0	0	0	0
17.1	0	0	0	10	0	0	0
18.7	0	0	0	0	5	0	0
20.4	0	0	0	0	0	3	0
26.6	0	0	0	0	0	0	1
Kurtosis	−2.0	−1.75	−1.5	−1.0	0.0	1.33	8.0
Variance	25	25.1	24.8	25.2	25.2	25.0	25.1

Iz: Wuensch, K. L. (2007). Skewness, Kurtosis, and the Normal Curve. Sneto s spletne strani: http://core.ecu.edu/psyc/wuenschk/StatsLessons.htm.

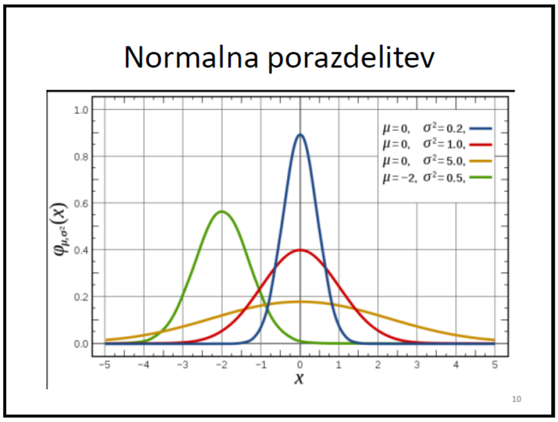
EXCEL: **Spl**: =KURT (populacijska); **As**: =SKEW (populacijska);

EXCEL (grupirani podatki): **As**: =SUM(f*(X'-M)^3)/(N*SD^3); **Spl**: =SUM(f*(X'-M)^4)/(N*SD^4)-3

Normalna porazdelitev

- zvezna verjetnostna porazdelitev
- Gaussova porazdelitev
- $f(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(X-\mu)^2}{\sigma^2}}$
- Standardizirana $N(0,1)$: $f(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{-z^2}$
- Dobro opisuje pojave, na katere delujejo slučajni vplivi.

9



Normalna porazdelitev je družina porazdelitev z različnimi, a podobnimi oblikami (graf desno). Gre za simetrične porazdelitve z vrednostmi skoncentriranimi pretežno okoli sredine. Definirajo jih aritmetična sredina (M oz. μ) in standardna deviacija (SD oz. σ).

Pri psihologiji imamo največkrat opravka z t.i. približki normalni porazdelitvi.

Formula po kateri izračunamo višino normalne distribucije se glasi: $f(X) = \frac{1}{\sqrt{2\pi \cdot SD^2}} \cdot e^{\frac{-(X-M)^2}{SD^2}}$

- $e \dots naravni\ logaritem\ [2{,}718282 \dots]$
- X lahko zavzema vrednosti od – neskončno do + neskončno $[-\infty, +\infty]$

f(X) je blizu 0, če je X več kot 3 standardne deviacije oddaljen od aritmetične sredine $[-3, +3]$

Normalna distribucija je lahko preoblikovana v standardizirano normalno distribucijo s sledečo formulo: $z = \frac{X-M}{SD}$. Imenujemo jo tudi z-distribucija. »z« predstavlja število standardnih deviacij nad in pod aritmetično sredino za določeno X vrednost.

Primer: M=50; SD=10; X=70.

$$z = \frac{70-50}{10} = 2$$

Vrednost X=70 je od aritmetične sredine oddaljena 2 standardni deviaciji.

$$f(X) = \frac{1}{\sqrt{2\pi \cdot SD^2}} \cdot e^{-z^2}$$

Standardna napaka (ang. standard error) je standardna deviacija od vzorčne distribucije. Na splošno velja, da večji kot je vzorec, manjša je standardna napaka. Standardna napaka aritmetične sredine je označena z *σM*.

Normalna porazdelitev

Table 7.2 Normal Distribution

Mean	μ
Variance	σ^2
Standard deviation	σ
Moment coefficient of skewness	$\alpha_3 = 0$
Moment coefficient of kurtosis	$\alpha_4 = 3$
Mean deviation	$\sigma\sqrt{2/\pi} = 0.7979\sigma$

Preverjanje odstopanja frekvenčne porazdelitve od normalne porazdelitve na osnovi deskriptivne statistike

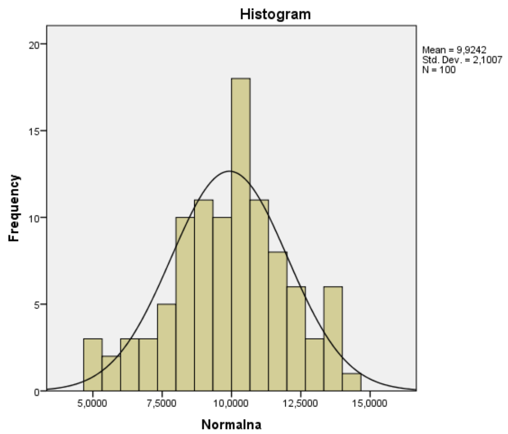
- Primerjava *M* in *Mdn*
- Primerjamo koeficient asimetričnosti (3. standardizirani centralni moment) z njegovo standardno napako: $SE(As) \cong \sqrt{6/n}$
- Primerjamo koeficient sploščenosti z njegovo standardno napako: $SE(Spl) \cong \sqrt{24/n}$
- Koeficient, ki je več kot 1,96-krat večji od *SE*, kaže preveliko odstopanje od normalnosti (gledamo absolutno vrednost koeficienta).

12

Normalnost porazdelitve lahko testiramo na več načinov: s sliko histograma, z prikazom Q-Q, s prikazom P-P, z statističnimi testi normalnosti (v SPSS imamo 2 pod Explore – Plots), lahko pa tudi preko asimetričnosti in sploščenosti, pri čemer velja: $As/SE(As)\ ali\ Spl/SE(Spl) < |1{,}96|$.

Normalna porazdelitev

N	Valid	100
	Missing	0
Mean		9,924156
Std. Deviation		2,1006583
Skewness		-,292
Std. Error of Skewness		,241
Kurtosis		-,069
Std. Error of Kurtosis		,478

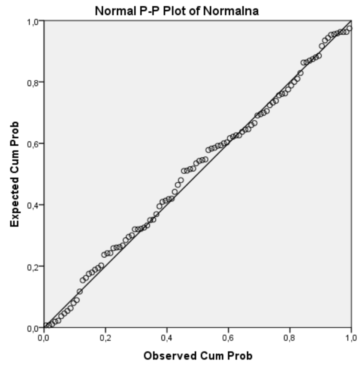


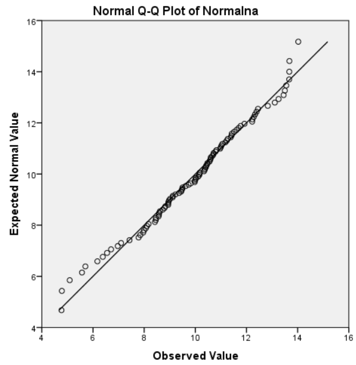
$\frac{As}{SE(As)} = \frac{-0{,}292}{0{,}241} = -1{,}21$ in $\frac{Spl}{SE(Spl)} = \frac{-0{,}069}{0{,}478} = -0{,}14$, ker je količnik v obeh primerih < |1,96|.

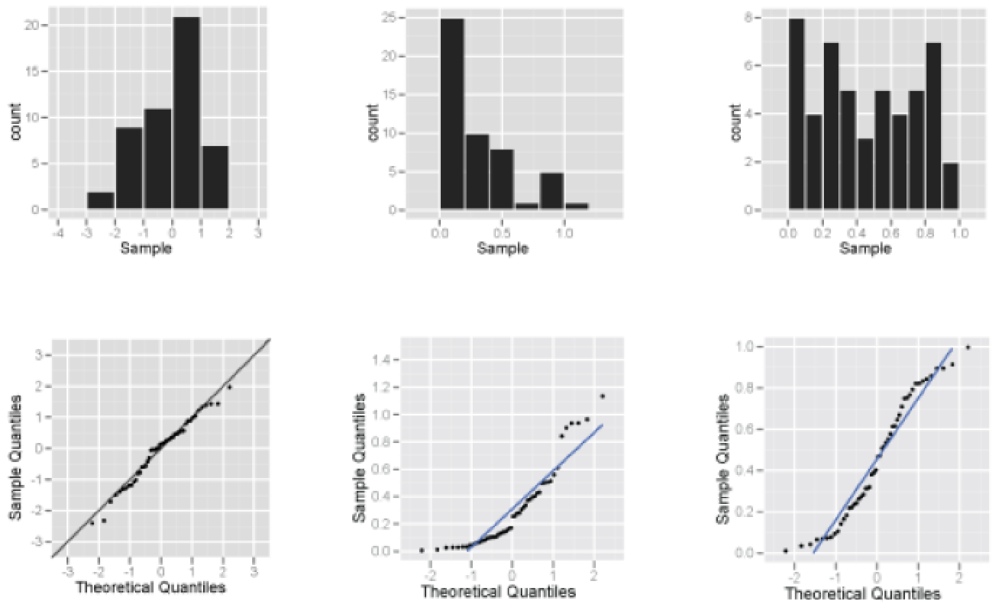
Grafične metode primerjanja z normalno porazdelitvijo

- Primerjava histograma z normalno krivuljo
- Normalni verjetnostni grafi
  - Prikaz Q-Q
    - Primerjamo empirične kvantile s teoretičnimi (ki bi jih dobili pri standardni normalni porazdelitvi)
    - Prileganje premici (*r*)
  - Prikaz P-P
    - Primerjamo kumulativno porazdelitev (cp_e) z normalno kumulativno porazdelitvijo (cp_t).

14

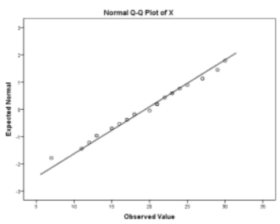


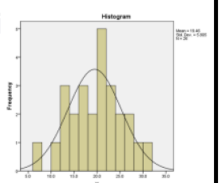


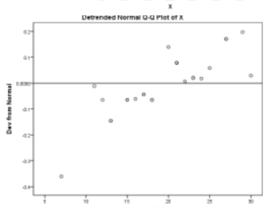


17

Odstopanje porazdelitve od normalnosti







Opis porazdelitve: povzetek

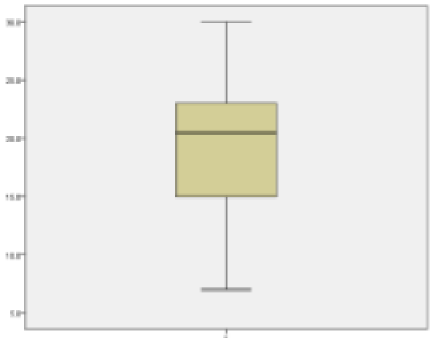
- frekvenčna porazdelitev:
  - N, O, I, R
- povprečje:
  - O: *Mo, Mdn*
  - I, R: *Mo, Mdn, M*
- razpršenost:
  - O: *VR, Q*
  - I, R: *VR, Q,* PO_M*, var, SD*
- oblika porazdelitve:
  - I: As, Spl

20

Odstopanje porazdelitve od normalnosti

Descriptives

			Statistic	Std. Error
X	Mean		19.462	1.1385
	95% Confidence Interval for Mean	Lower Bound	17.117	
		Upper Bound	21.806	
	5% Trimmed Mean		19.530	
	Median		20.500	
	Variance		33.698	
	Std. Deviation		5.8050	
	Minimum		7.0	
	Maximum		30.0	
	Range		23.0	
	Interquartile Range		8.3	
	Skewness		-.105	.456
	Kurtosis		-.455	.887



Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
X	.105	26	.200*	.985	26	.961

a. Lilliefors Significance Correction
*. This is a lower bound of the true significance.

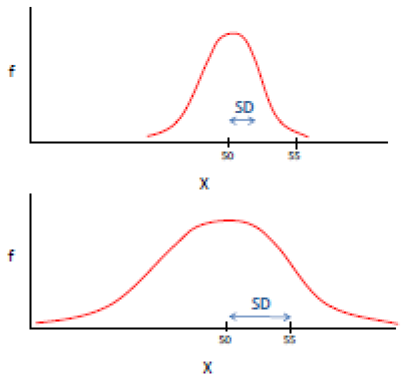
Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Normalna	,060	100	,200*	,982	100	,185

Za naš primer oba testa na *Sig*. kažeta **več od 0,05**, saj gre za normalno distribucijo.

Položaj podatka v normalni porazdelitvi

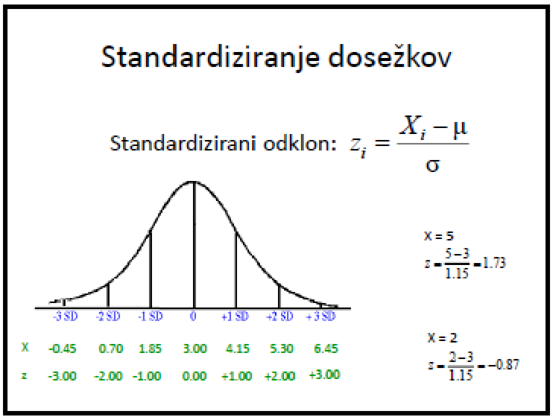
Standardiziranje dosežkov

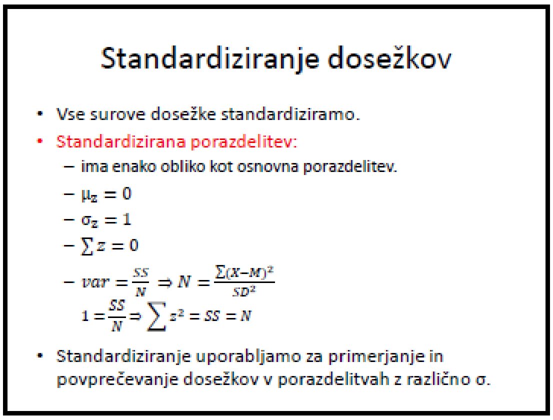


standardizirani odklon:

$$z_i = \frac{X_i - \mu}{\sigma}$$

Primerjava rezultatov na različnih testih





Primer A.

- ekstravertnost: M=12, SD=2, X_E=12
- nevroticizem: M=20, SD=4, X_N=12

Izračunaj povprečen položaj, posameznikov povprečen dosežek.

$$z_E = \frac{12-12}{2} = 0$$

$$z_N = \frac{12-20}{4} = -2{,}0$$

$$\bar{z} = \frac{0-2{,}0}{2} = -1{,}0$$

Primer B.

- SD=1,5
- z=–1,50
- M=3

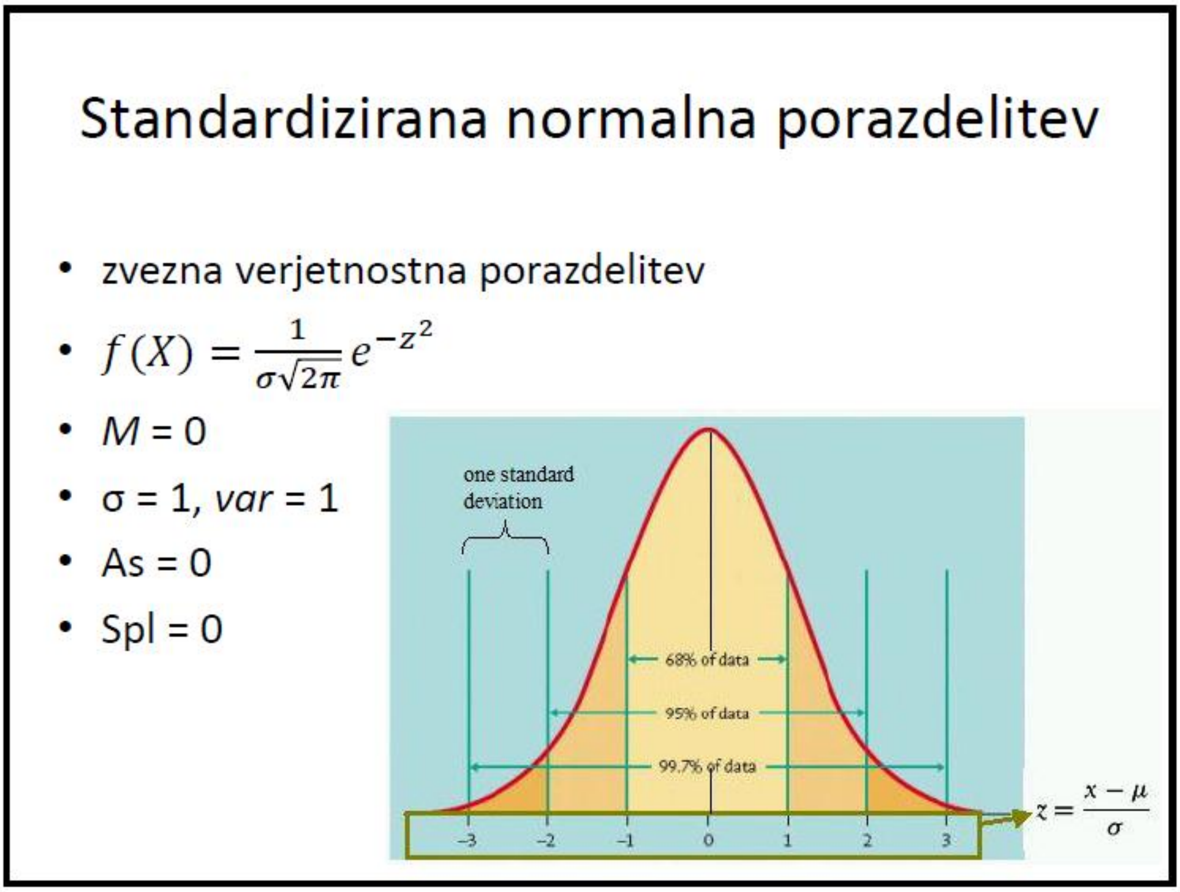
x=?

$$\frac{x-3}{1{,}15} = -1{,}50 \ \rightarrow x = -1{,}50 \cdot 1{,}15 + 3 = 1{,}28$$

Primer C.

- z=3
- SD=1,15
- M=3

$$x = 3 + 1{,}15 \cdot 3 = 6{,}45$$

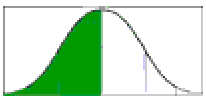


Standardizirana normalna porazdelitev

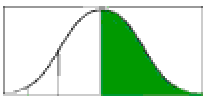
- le pri zveznih spremenljivkah
- M=0
- σ=1, var=1
- As=0
- Spl=0

- $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-z^2}$
- med -1 in 1 je 2/3 podatkov (68 %)
- med -2 in 2 je 95 % podatkov
- med -3 in 3 je 99,7 % podatkov

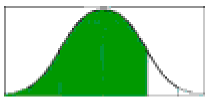
Skladno z ugotovitvami o osnovnih lastnostih standardizirane normalne distribucije lahko ugotavljamo, kolikšen delež podatkov leži pod ali nad določeno vrednostjo Z:



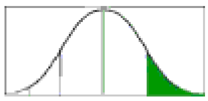
z=0; % P pod = 50 %



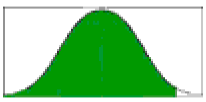
z=0; % P nad = 50 %



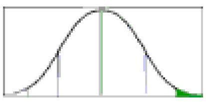
z=1; % P pod = 84,1 %



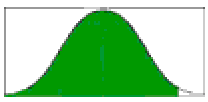
z=1; % P nad = 15,9 %



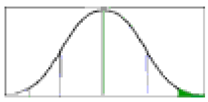
z=2; % P pod = 97,7 %



z=2; % P nad = 2,3 %



z=3; % P pod = 99,9 %

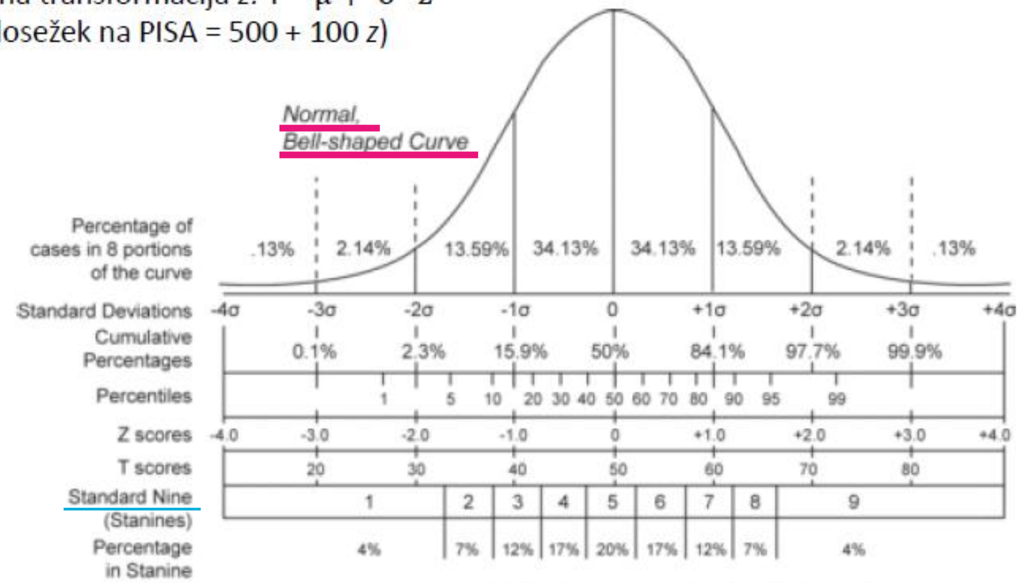


z=3; % P nad = 0,1 %

Površina pod Z je, kot bomo ugotovili, enaka **percentilnemu rangu** vrednosti Z oz. NORMSDIST(z) v Excelu (glej str. 63).

Druge standardizirane porazdelitve

Linearna transformacija z: $Y = \mu + \sigma \cdot z$
(npr. dosežek na PISA = 500 + 100 z)



Z-scores	< -2.0	-2 to -1.5	-1.5 to -1.0	-1.0 to -.5	-.5 to 0	0 to +.5	+.5 to +1.0	+1.0 to +1.5	+1.5 to +2.0	> +2.0
Percent	2.3%	4.4%	9.2%	15.0%	19.2%	19.2%	15.0%	9.2%	4.4%	2.3%
Sten	1	2	3	4	5	6	7	8	9	10

- *osnovna formula*: Y=M+SD*z
- T=50+10*z (M je P_{50}, 10 kategorij)
- STANINE=5+2*z (dosežke razdelimo na 9 kategorij)
- STEN=5,5+2*z (dosežke razdelimo na 10 kategorij)
- IQ=100+15*z
- velja le za normalno porazdelitev

Druge standardizirane porazdelitve

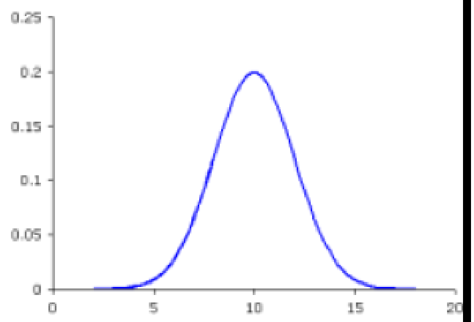
Figure 5.3 Standard score systems

Percentile score for normal distribution	z score Mean 0 SD 1	T score Mean 50 SD 10	Deviation IQ Mean 100 SD 15	Stanine Mean 5 SD 2	Sten Mean 5.5 SD 2
	-4 SD	10			
1	-3 SD	20	55		
2.5	-2 SD	30	70	1	1.5
16	-1 SD	40	85	3	3.5
50	0	50	100	5	5.5
84	+1 SD	60	115	7	7.5
97.5	+2 SD	70	130	9	9.5
99	+3 SD	80	145		
	+4 SD	90			

Iz: Jackson, C. (1996). *Understanding psychological testing*. Leicester: BPS Books.

Verjetnost v normalni porazdelitvi

- N. D. je zvezna verjetnostna porazdelitev → Zvezna naključna spremenljivka lahko zasede zvezno množico vrednosti.
- Limita poligona relativnih frekvenc je zvezna porazdelitev.
- Ne govorimo o verjetnosti pojavljanja posameznih vrednosti, ampak intervala vrednosti. → (verjetnostna) gostotna funkcija (angl. *probability density function*). Verjetnost, da bo imela naključna spremenljivka vrednost v intervalu A, je integral gostotne funkcije nad intervalom A.



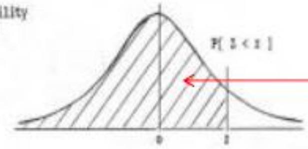
8

STANDARD STATISTICAL TABLES

1. Areas under the Normal Distribution

The table gives the cumulative probability up to the standardised normal value z
i.e.

$$P[Z < z] = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp(-\tfrac{1}{2}Z^2)\, dZ$$



p levo od vrednosti *z* nam pove, kolikšen je percentilni rang vrednosti *z*.

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	0.5040	0.5080	0.5120	0.5159	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7854
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9773	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9865	0.9868	0.9871	0.9874	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9924	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9980	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986

z	3.00	3.10	3.20	3.30	3.40	3.50	3.60	3.70	3.80	3.90
P	0.9986	0.9990	0.9993	0.9995	0.9997	0.9998	0.9998	0.9999	0.9999	1.0000

Excelove funkcije:

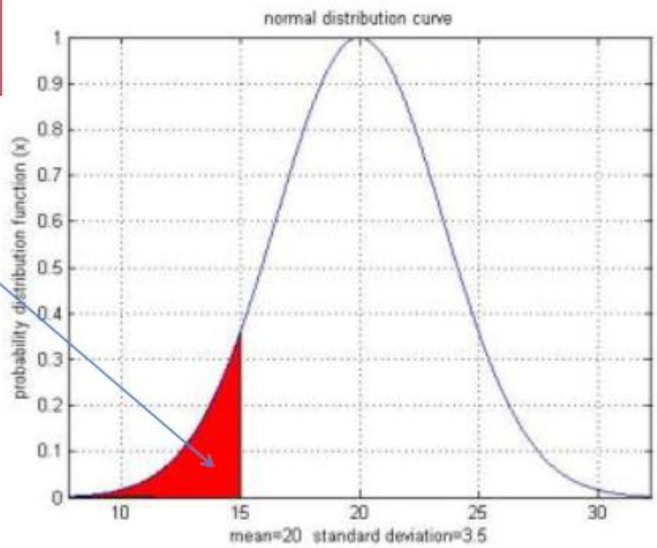
NORMSDIST(z) → *p* (levo od z)
NORM.S.DIST(z;TRUE)

NORMSINV(p) → *z*, od katerega levo je površina pod krivuljo enaka *p*
NORM.S.INV(p)

NORM.S.DIST(z;FALSE) → višina krivulje/ordinata

9

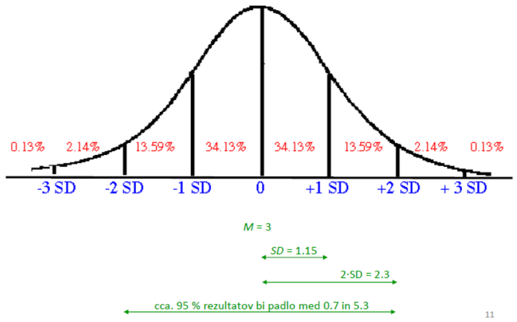
p levo od vrednosti X nam pove, kolikšen je percentilni rang vrednosti X.



Excelove funkcije
NORMDIST(X;M;SD;TRUE) – NORM.DIST(X;M;SD;TRUE)
NORMINV(p;M;SD) – NORM.INV(p;M;SD)

10

Normalna porazdelitev kot model frekvenčne porazdelitve spremenljivke



11

Binomska porazdelitev kot model frekvenčne porazdelitve spremenljivke

- Če ima spremenljivka dve vrednosti (kategoriji), so podatki binomski:
  - naravna dihotomnost: Moški/ženska, cifra/glava, da/ne ...
  - umetna dihotomnost (podatke združimo v dve kategoriji): visoko-nizko anksiozen, pade-ne pade šestica
  - → Diskretne spremenljivke

Binomska porazdelitev

Primeri vprašanj:

- Če 4-krat vržemo kovanec, kakšna je verjetnost, da bomo dobili 0/1/2/3/4 glave?
- Če 5-krat potegnemo karto (z vračanjem), kakšna je verjetnost, da bomo dobili natančno 0/1/2/3/4 ase?
- Če kocko vržemo 20-krat, kakšna je verjetnost posameznega števila šestic?
- Če ima test 20 nalog z dvema odgovornima alternativama, kakšna je verjetnost, da ob ugibanju pravilno rešimo 1 nalogo, 2 nalogi, 3 naloge, 4 naloge, 5 nalog, ... vseh 20 nalog?

13

Osnove teorije verjetnosti

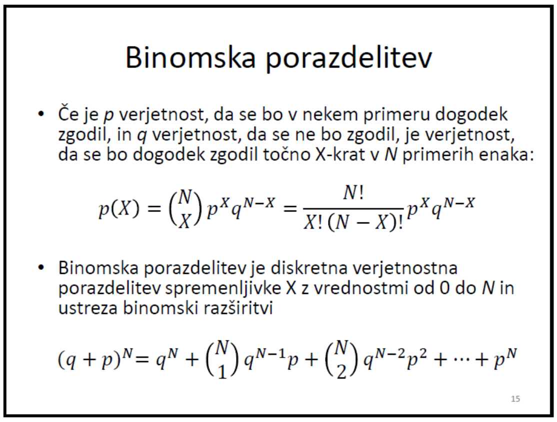
- Verjetnost = število, ki nam pove, kolikšna je možnost, da se zgodi nek izid A ... P(A)

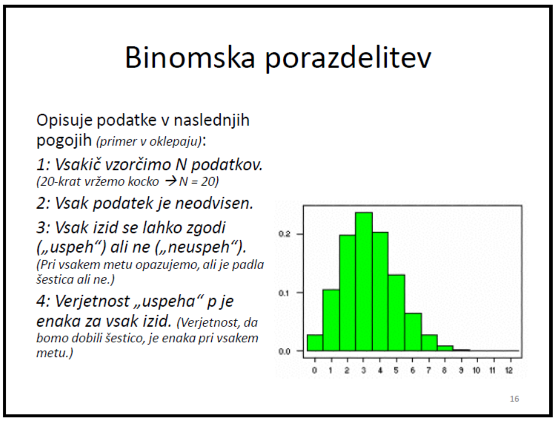
$$P(A) = \frac{m}{n} = \frac{\text{število ugodnih izidov}}{\text{število vseh izidov}}$$

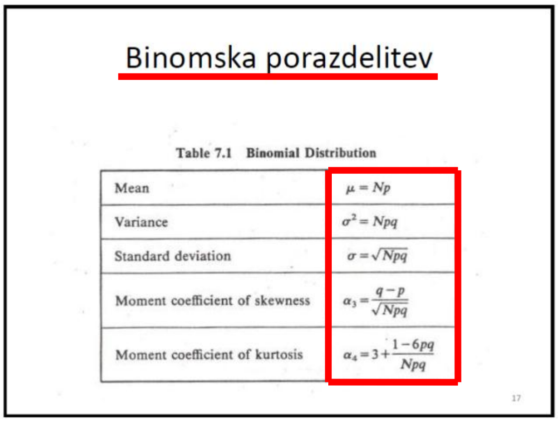
$$P(A) = \frac{822}{6850} = \frac{3}{25} = 0.12 = 12\%$$

- Ta definicija velja samo za naključne vzorce.
  - Vsak element populacije mora imeti enako možnost izbora v vzorec.
  - Če izberemo več kot en element, morajo imeti vsi enako možnost izbora (*p* je konstantna; vzorčenje z vračanjem).

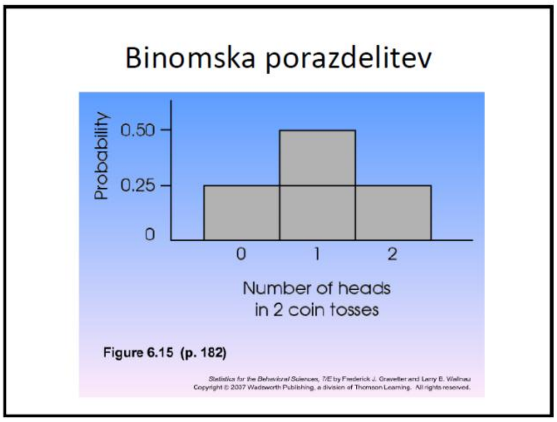
14







Mean	$\mu = Np$
Variance	$\sigma^2 = Npq$
Standard deviation	$\sigma = \sqrt{Npq}$
Moment coefficient of skewness	$\alpha_3 = \frac{q-p}{\sqrt{Npq}}$
Moment coefficient of kurtosis	$\alpha_4 = 3 + \frac{1-6pq}{Npq}$

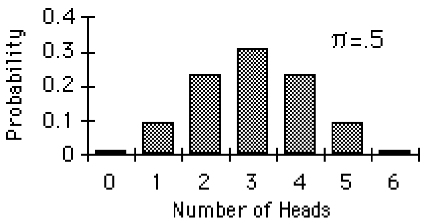


Binomska porazdelitev je podobna normalni – se ji dobro približuje.

Primer A. Mečemo kovanec. Kakšna je verjetnost, da trikrat dobimo grb, če mečemo 6-krat?

N=6; X=3; P= $\frac{1}{2}$

$$P(3) = \binom{6}{3} \cdot \left(\frac{1}{2}\right)^3 \cdot \left(\frac{1}{2}\right)^{6-3} = \frac{6!}{3! \cdot (6-3)!} \cdot \frac{1}{8} \cdot \frac{1}{8}$$
$$= 20 \cdot \frac{1}{64} = 0{,}3125$$



Primer B. 20-krat vržemo kocko. Kakšna je verjetnost da vsak met posebej pade 6? Kakšna je verjetnost, da pade 0-krat? Kakšna je verjetnost, da pade 1-krat?

Verjetnost, da pade 6 je 1/6.

$$p(0) = \binom{20}{0} \cdot \left(\frac{1}{6}\right)^0 \cdot \left(\frac{5}{6}\right)^{20} = \frac{20!}{0! \cdot 20!} \cdot 1 \cdot \left(\frac{5}{6}\right)^{20} = 0{,}026$$

$$p(1) = \binom{20}{1} \cdot \left(\frac{1}{5}\right)^1 \cdot \left(\frac{5}{6}\right)^{19} = \frac{20!}{1! \cdot 19!} \cdot \frac{1}{6} \cdot \left(\frac{5}{6}\right)^{19} = 0{,}104$$

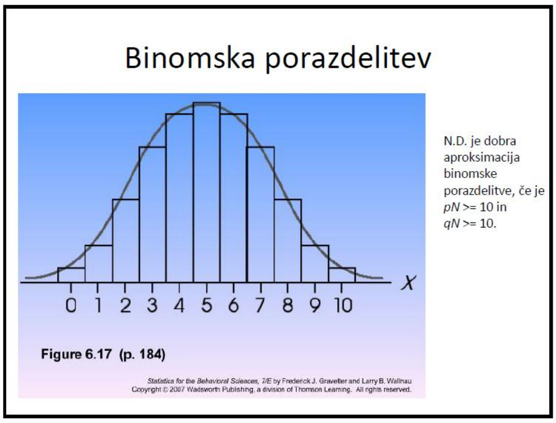
$$M = N \cdot p = 20 \cdot \frac{1}{6} = 3{,}33$$

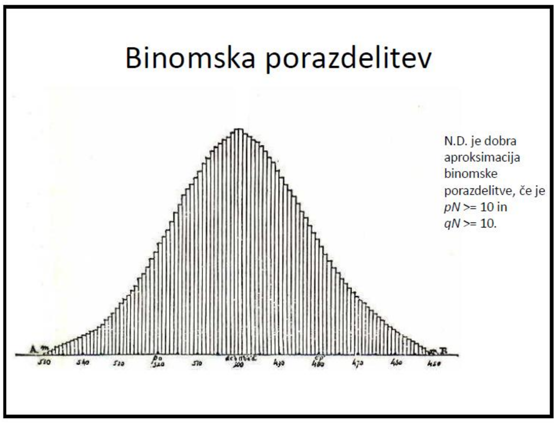
$$var = N \cdot p \cdot q = 20 \cdot \frac{1}{6} \cdot \frac{5}{6} = 2{,}78$$

$$SD = \sqrt{var} = \sqrt{2{,}76} = 1{,}67$$

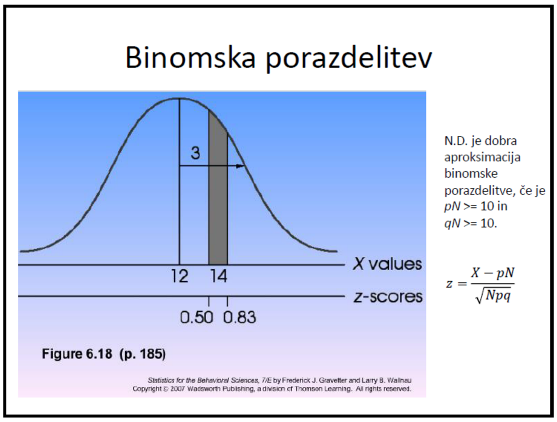
$$As = \frac{q - p}{\sqrt{Npq}} = \frac{\frac{5}{6} - \frac{1}{6}}{1{,}67} = 0{,}40$$

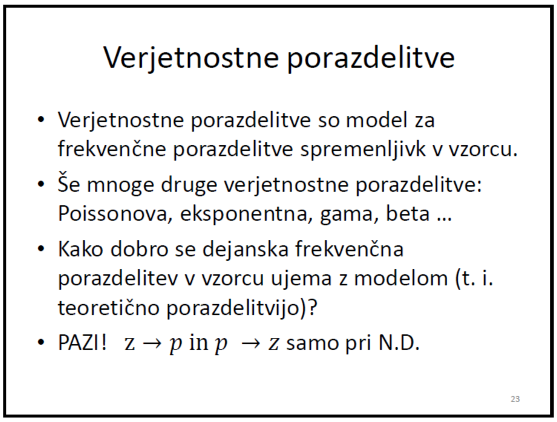
$$Spl = \frac{1 - 6 \cdot pq}{Npq} = \frac{1 - 6 \cdot \frac{1}{6} \cdot \frac{5}{6}}{20 \cdot \frac{1}{6} \cdot \frac{5}{6}} = 0{,}06$$





Recimo, da mečemo kovance – večkrat kot ga vržemo, bolj je krivulja podobna normalni – zvezni. Z naraščajočim št. poskusov postaja binomska porazdelitev podobnejša normalni z enakima M in SD.





Primer C. 48 nalog = N; a, b, c, d tipi nalog; dosežek je 14 = X. Kakšna je verjetnost tega rezultata?

$$p = \frac{1}{4}, \qquad M = N \cdot p = 48 \cdot \frac{1}{4} = 12$$

$$SD = \sqrt{Npq} = \sqrt{48 \cdot \frac{1}{4} \cdot \frac{3}{4}} = 3$$

a. binomska: $p(X = 14) = \binom{48}{14} \cdot \left(\frac{1}{4}\right)^{14} \cdot \left(\frac{3}{4}\right)^{34}$

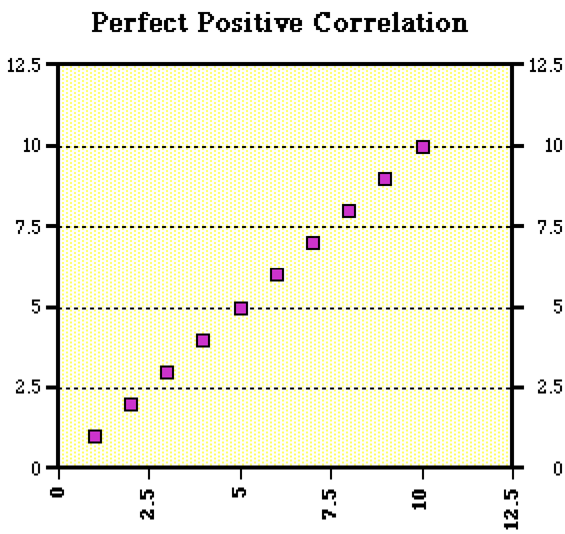
b. normalna: $x_{sp} = 13{,}5 \to z_{sp};\ x_{zg} = 14{,}5 \to z_{zp}$ – med z vrednostnima določimo p med krivuljama (=aproksimacija z uporabo normalne porazdelitve)

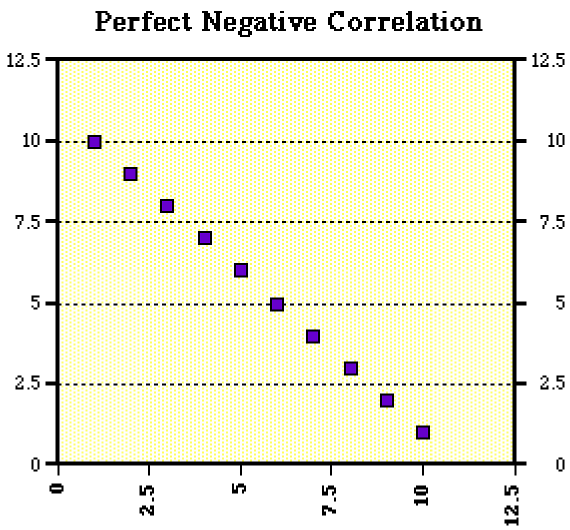
Korelacija

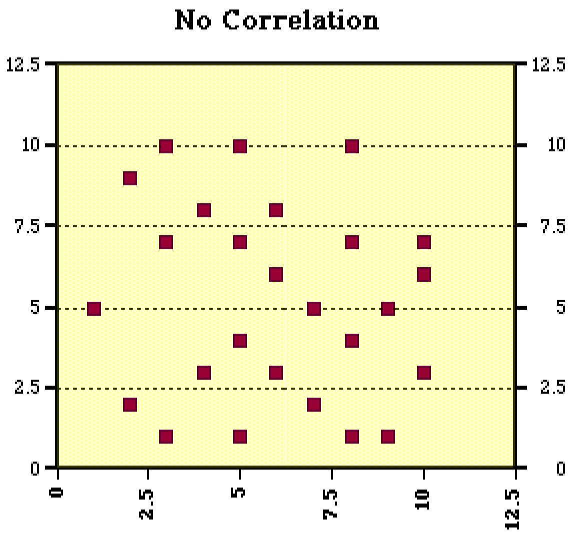
Vrste odnosov med spremenljivkami

Korelacija (povezanost, asociacija, sovariiranje)

- **Smer**: nična korelacija, pozitivna (tisti, ki ima višji dosežek na X osi ima višji dosežek n Y osi), negativna korelacija (nižje vrednosti na X in višje na Y osi oz. obratno)
- **Oblika odnosa**: linearna, krivuljčni odnos (obrnjena U krivulja) in drugi nelinearni odnosi
- **Stopnja korelacije**: 0 (ničelna), +1 (popolna pozitivna povezanost), -1 (popolna negativna povezaost)
    - **korelacijski koeficienti**: [-1, +1]





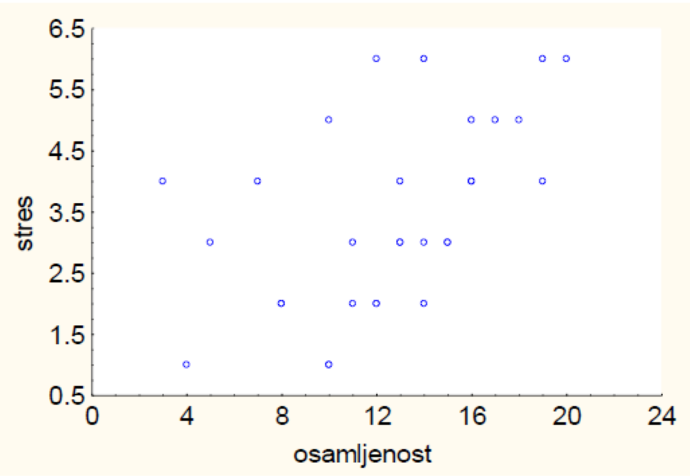


Cela vrsta posledic, zakaj sta dve spremenljivki povezani, v korelacijah: spremenljivki sta povezani, a ne moremo govoriti o tem, da ena vpliva na drugo:

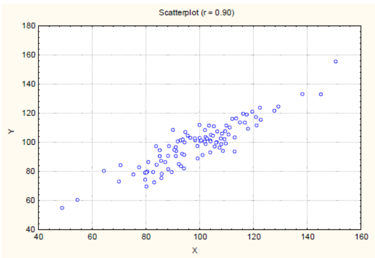
- **Napovedovanje**: naravno dane spremenljivke, spontane (ni neodvisnih in odvisnih, čeprav se včasih uporabljajo ti izrazi, saj iz ene napovedujemo drugo; previdnost pri interpretaciji!), anksioznost in agresivnost nista taki spremenljivki, da bi bila ena predhodna/posledična drugi, ampak se sočasno pojavljajta
- **Odvisnost**: odvisno od tega, kakšen je šolski uspeh v srednji šoli, bo uspeh na fakulteti, vendar to ni pogoj (predhodna spremenljivka), med njima ni direkten vzročno-posledični odnos
- **Vzročnost (kavzalnost)**: če povečamo število izbir, se bo normalno povečal reakcijski čas

- razpršitveni/razsevni diagram (scatterplot), vsaka točka nosi koordinate (x, y) → korelacijski oblak – kadar so točke blizu, ležijo blizu korelacijske premice, pomni, da je odnos med X in Y precej jasen

stres	osamlj
1	10
3	15
1	4
4	16
2	12
2	8
3	15
2	14
4	19
5	10
6	20
1	10
2	8
3	11
4	16
2	12
5	17
3	5
4	3
3	13
5	16
6	12
3	13
4	13
2	11
4	7
5	18
6	14
6	19
3	14

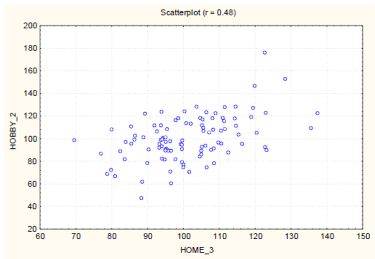


Vrste povezanosti

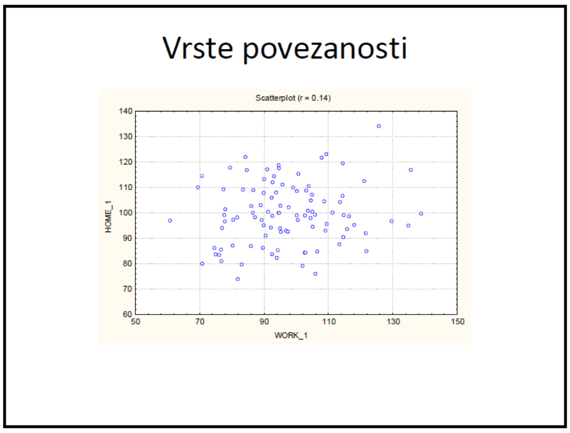


Tesneje ko so točke zbrane okoli korelacijske premice, če bi vse točke ležale na premici, bi bila korelacija 1, če je ozek, ampak niso vse točke na premici, je visok (Pearsonov diagram=opisovanje linearnih odnosov, v tem primeru je korelacija 0,90). Strmina krivulje s samo korelacijo nima veze. Strmina je odvisna od vrste lestvic, na katerih merimo npr. X [4,16]

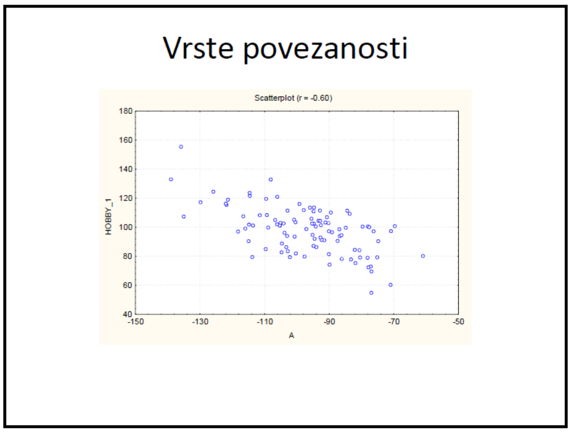
Vrste povezanosti



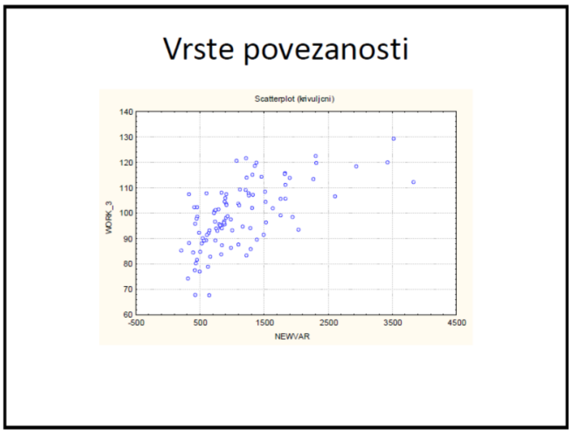
Gre za šibkejšo pozitivno korelacijo. Korelacija *r*=0,48 je v psihologiji visoka korelacija, saj smo ljudje zelo različni in je razpršenost velika.



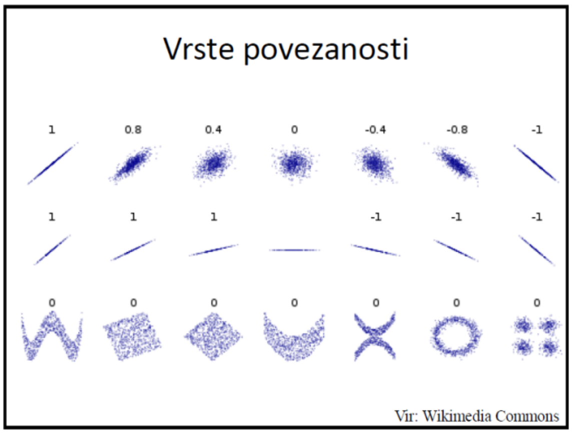
Obstaja pozitivna povezanost (r=0,14), vendar je šibka, gre za majhno, nejasno povezanost.



Negativna korelacija.



Pearsonov korelacijski koeficient (r) je namenjen ocenjevanju linearnih odnosov, ne krivuljčnih. To pomeni, da Pearsonov korelacijski koeficient lahko uporabimo, ko sta obe spremenljivki intervalni in je zveza med njima linearna. Med vsemi korelacijskimi koeficienti ta najbolje odraža povezanost.



Vidimo, da se odnos dobro vidi iz korelacijskega oblaka, Pearsonov koeficient ni primeren, saj je med spremenljivkami v nelinearnih odnosih lahko 0.

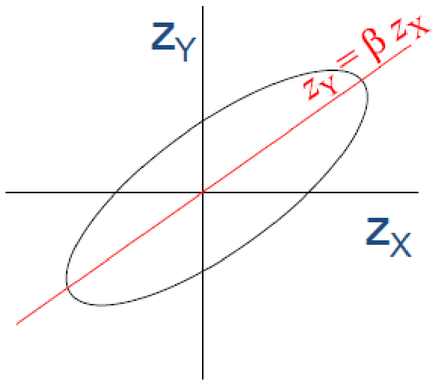
Osnovni obrazec za Pearsonov korelacijski koeficient:

$$r = \frac{kovaricana}{standardni\ odklon\ na\ x * stadradni\ odklon\ na\ y} = \frac{cov}{\sigma_x \sigma_y} = \frac{\sum (x - M_x)(y - M_y)}{N \cdot \sigma_x \sigma_y}$$

‘produkt-moment’ pristop k *r*

Ali je visok dosežek na eni povezan z visokim dosežkom na drugi spremenljivki? Ali je odklon dosežka od sredine pri prvi spremenljivki povezan z odklonom pri drugi spremenljivki?

r ... indeks tendence po skupnem variiranju obeh spremenljivk

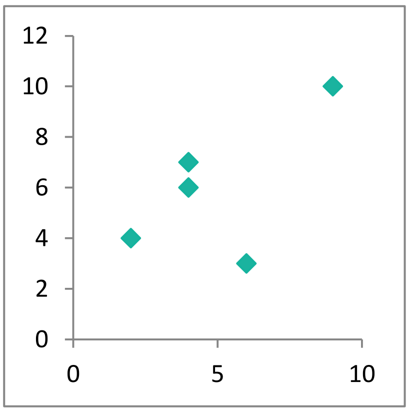


$$r = \frac{\sum_{i=1}^{N} z_{X_i} z_{Y_i}}{N}$$

$$r = \frac{\sum (X_i - M_X)(Y_i - M_Y)}{N \cdot SD_X \cdot SD_Y}$$

Razpršitvene diagrame rišemo v razmerju 1:1 (kvadrat), da lažje opazujemo odnose.

oseba	X	Y	X-Mx	Y-My	(X-Mx)(Y-My)	Zx	Zy	Zx*Zy
1	6	3	1	-3	-3	0,42	-1,22	-0,52
2	9	10	4	4	16	1,69	1,63	2,76
3	4	7	-1	1	-1	-0,42	0,41	-0,17
4	2	4	-3	-2	6	-1,27	-0,82	1,03
5	4	6	-1	0	0	-0,42	0,00	0,00
M	5	6						∑=3,1
SD	2,37	2,45						

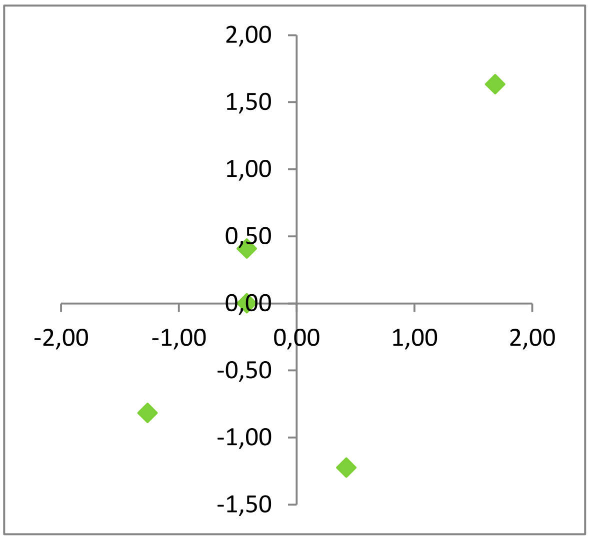


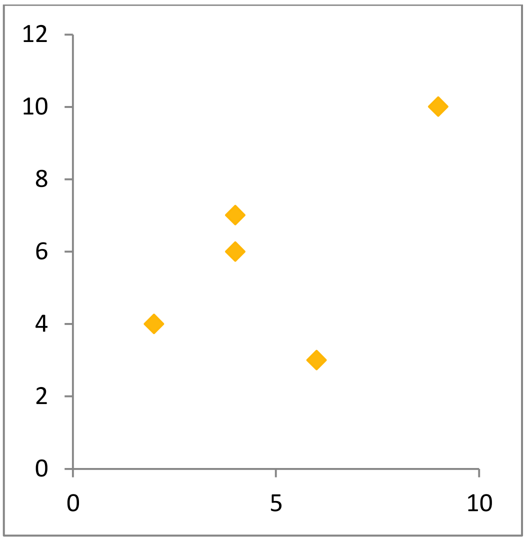
$$cov_{xy} = \frac{\sum(X - M_x)(Y - M_y)}{N} = \frac{18}{5} = 3{,}6$$

Kovarianca nam pove nekaj o odnosu med spremeljivkami. Če bi X povečali za 5× in Y za 10× bi bila kovarianca 3,6×5×10.

$$r = \frac{cov}{SD_x \cdot SD_y} = \frac{3{,}6}{2{,}37 \cdot 2{,}45} = 0{,}62$$

$$z = X - M/_{SD} \rightarrow r = \frac{\sum(z_x \cdot z_y)}{N} = \frac{3{,}1}{6} = 0{,}62$$





Oba prikaza sta enaka – tisti standardiziranimi in tisti z surovimi vrednostmi.

'produkt-moment' pristop k *r*

Ali je visok dosežek na eni povezan z visokim dosežkom na drugi spremenljivki? Ali je odklon dosežka od sredine pri prvi spremenljivki povezan z odklonom pri drugi spremenljivki?

r ... indeks tendence po skupnem variiranju obeh spremenljivk

$$r = \frac{stopnja\ sovariiranja\ X\ in\ Y}{stopnja\ neodvisnega\ variiranja\ X\ in\ Y}$$

$$r = \frac{\mathrm{cov}_{XY}}{\sigma_X \sigma_Y}$$

$$r = \frac{SP}{\sqrt{SS_X SS_Y}} \Rightarrow r = \frac{\sum (X_i - M_X)(Y_i - M_Y)}{N \cdot SD_X \cdot SD_Y}$$

$$SP = \sum (X - M_x)(Y - M_y)$$

$$SS_x = \sum (X - M_x)^2; \quad SS_y = \sum (Y - M_y)^2$$

	X	Y	X-M(X)	Y-M(Y)	(X-M(X))(Y-M(Y))
	6	3	1	-3	-3
	9	10	4	4	16
	4	7	-1	1	-1
	2	4	-3	-2	6
	4	6	-1	0	0
M	5	6	0	0	3,6
SD	2,366432	2,44949	2,366432	2,44949	kovarianca

	5*X	10*Y	X-M(X)	Y-M(Y)	(X-M(X))(Y-M(Y))
	30	30	5	-30	-150
	45	100	20	40	800
	20	70	-5	10	-50
	10	40	-15	-20	300
	20	60	-5	0	0
M	25	60	0	0	180
SD	11,83216	24,4949	11,83216	24,4949	kovarianca

kovarianca = 3.6*5*10

Vedno enako vrednost bi dala enačba $\frac{\mathrm{cov}_{XY}}{SD_X SD_Y}$ ... = enačba za *r*

Kovarianca je merilo koliko dve spremenljivki variirata skupaj (za razliko od variance, ki meri koliko variira ena sama spremenljivka). Obrazec: $C_{xy} = cov_{xy} = \frac{\sum(x-M_x)(y-M_y)}{N}$

Če dve spremenljivki težita k skupni variaciji (ko je ena nad pričakovano vrednostjo teži tudi druga nad svojo pričakovano vrednost), je kovarianca med njima pozitivna.

Poročanje

- „Prepričanje, da se bodo šli naslednje leto cepit proti gripi, je s starostjo naraščalo, *r* = 0,65.“
- Korelacijske matrike:

Table 3. *Correlations among items used to assess forgiveness dimensions for husbands (above diagonal) and wives (below diagonal)*

Items	1	2	3	4	5	6
1. “… want to see hurt and miserable”	1.00	.47	.39	−.37	−.36	−.30
2. “… think of how to even the score”	.57	1.00	.31	−.35	−.34	−.35
3. “… think of ways to make them regret”	.59	.53	1.00	−.22	−.30	−.23
4. “… just accept partner’s humanness, flaws and failures”	−.54	−.29	−.53	1.00	.52	.51
5. “… let bygones be bygones”	−.44	−.26	−.29	.56	1.00	.56
6. “… quick to forgive”	−.53	−.31	−.42	.53	.55	1.00

Pri poročanju vedno navajamo *r* in to v ***poševnem tisku***.

Zapisovanje korelacij v tabeli. Včasih so zapisane samo vrednosti zgornjega trikotnika, oz. vrednosti spodnjega trikotnika (npr. sivih cifer ni izpisanih, saj so razvidne).

	X	Y	Z
X	1,00	0,50	0,25
Y	0,50	1,00	0,40
Z	0,25	0,40	1,00

Velikost učinka

0,10 - nizek *r*
0,30 - srednje visok *r*
0,50 - visok *r*

Odstotek pojasnjene variance

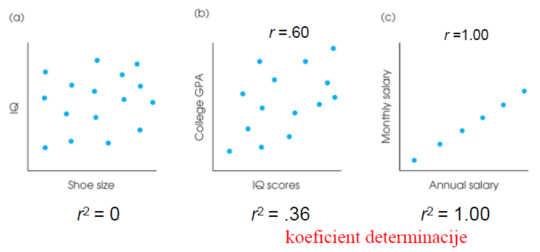


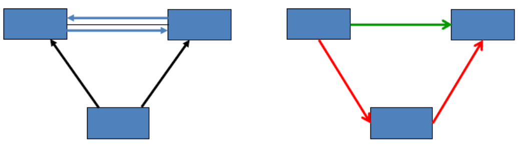
Figure 15-8 (p. 425)
Three sets of data showing three different degrees of linear relationship.

Essentials of Statistics for the Behavioral Sciences, 5/E by Frederick J Gravetter and Larry B. Wallnau
Copyright © 2005 by Wadsworth Publishing, a division of Thomson Learning

Če koeficient korelacije kvadriramo, dobimo koeficient determinacije. 36 % odstotkov variabilnosti znamo pojasniti z odnosom dosežka na testu in IQ-ju. Popolnoma povezani sta leta in mesečna plača, ker lahko letno plačo v celoti pojasnimo z mesečno plačo (r^2=1,00).

Težave pri interpretaciji *r*

korelacija in vzročnost



Vpliv omejenosti obsega

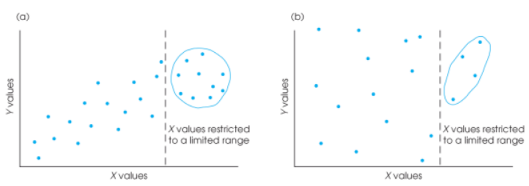


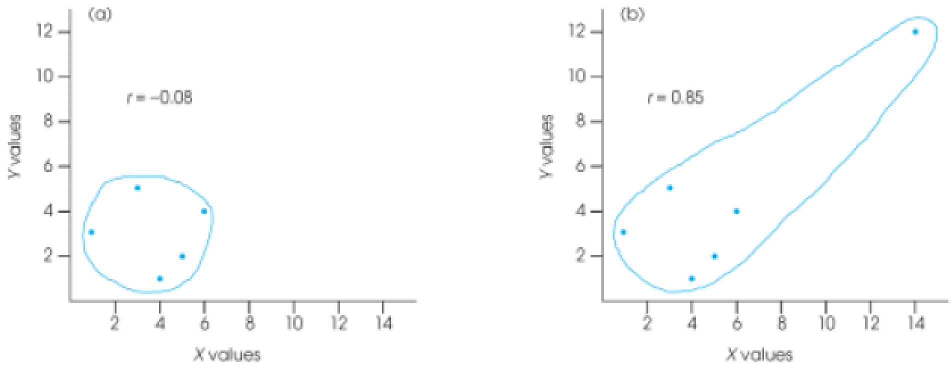
Figure 15-6 (p. 423)
(a) In this example, the full range of *X* and *Y* values shows a strong, positive correlation, but the restricted range of scores produces a correlation near zero.
(b) Here the full range of *X* and *Y* values shows a correlation near zero but the scores in the restricted range produce a strong positive correlation.

Essentials of Statistics for the Behavioral Sciences, 5/E by Frederick J Gravetter and Larry B. Wallnau
Copyright © 2005 by Wadsworth Publishing, a division of Thomson Learning

Če je korelacija med štorkljami in dojenčki visoka, ne moremo sklepati, da štorklje nosijo dojenčke ali da dojenčki kličejo štorklje. Neka tretja spremenljivka zelo pogosto vpliva na povezanost dveh spremenljivk (npr. poseljenost in hrana). Če dve spremenljivki korelirata, ni nujno da vplivajo ena na drugo (desna slikica) – npr. reakcijski čas in število izbir – to ne pomeni, da je odnos direkten.

Ko govorimo o povezanosti spremenljivk, je pomembno upoštevati zajeti vzorec. X os je matura, Y os pa IQ. Na vzorcu študentov psihologije ne bi izmerili povezave med IQ in rezultatom na maturi, medtem ko bi na splošnem vzorcu prišli do višje korelacije (študentje psihologije bi torej imeli v povprečju višji IQ).

Vpliv izstopajočih vrednosti



Original Data

Subject	X	Y
A	1	3
B	3	5
C	6	4
D	4	1
E	5	2

Data with Outlier Included

Subject	X	Y
A	1	3
B	3	5
C	6	4
D	4	1
E	5	2
F	14	12

Figure 15-7 (p. 424)
A demonstration of how one extreme data point (an outlier) can influence the value of a correlation.

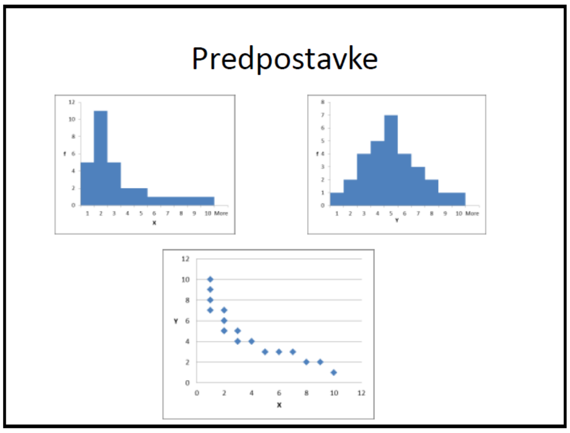
Essentials of Statistics for the Behavioral Sciences, 5/E by Frederick J Gravetter and Larry B. Wallnau
Copyright © 2005 by Wadsworth Publishing, a division of Thomson Learning

Ekstremne vrednosti imajo lahko velik vpliv, zato je vedno smiselno pogledati celoten razpršitveni diagram. Vpliv ekstremnega primera pokažemo tako, da odrežemo ekstremno vrednost in pokažemo korelacijo za skupino z in brez ekstremnega posameznika.

Ekstremne vrednosti:
- Ali jih imamo?
- Kako vplivajo?

Predpostavke

- Merske lestvice za Pearsonov *r*:
  - Y mora biti intervalna (zvezna, merjena) spremenljivka
  - X je načeloma lahko na katerikoli ravni (glej npr. točkovno biserialni koeficient)
- Odnos med X in Y je linearen → pregledati razpršitveni diagram
  - problem nenormalni distribuciji
  - če ni, raje uporabimo Spearmanov ρ



Korelacija spremenljivk,
ki niso merjene na intervalni merski ravni

druga od sprem. \ ena od sprem.	nominalna	ordinalna
nominalna	n.d.-n.d.: Φ koeficient u.d.-u.d.: Tetrahorični koef. Več kategorij: koeficient kontingence *C*, Cramerjev *V*	Kendallov τ
ordinalna	Kendallov τ	Spearmanov ρ Kendallov τ
intervalna	n.d.- Točkovno biserialni koef. u.d.- Biserialni koef.	Spearmanov ρ Kendallov τ

n.d.: naravno dihotomna
u.d.: umetno dihotomizirana

Kendallov koeficient konkordance *W* - skladnost med ocenjevalci

Linearna regresija

REGRESIJA

↓

opis odnosov, napovedovanje

KORELACIJA

↓

opis velikosti povezanosti

Napovedovanje

Predikcija, ocena Y na osnovi vrednosti X:

- Y … **kriterijska** spremenljivka (odvisna)
- X … **prediktorska** spremenljivka (neodvisna)

Regresijska analiza:

- enostavna (bivariatna: iz ene prediktorske napovedujemo eno kriterijsko spremenljivko) in multipla regresija
- linearna in nelinearna regresija

Uvod v regresijo

- Pearsonova korelacija (koeficient *r*) je tehnika, ki opisuje in meri linearno zvezo (linearen odnos) med dvema spremenljivkama
  - vsaka točka predstavlja odnos med neko vrednostjo na X in neko vrednostjo na Y
- čez korelacijski oblak rišemo premico
  - tako lažje razberemo odnos med X in Y (kriterijsko in prediktorsko spremenljivko)
  - premica označuje centralno tendenco (sredino) odnosa (tako kot npr. aritmetična sredina označuje centralno tendenco nekega seta rezultatov)

  - gre za poenostavljen opis odnosa med X in Y
- naš cilj je vedeti, kakšna je ta premica (oz. njena enačba)

Pogojne aritmetične sredine

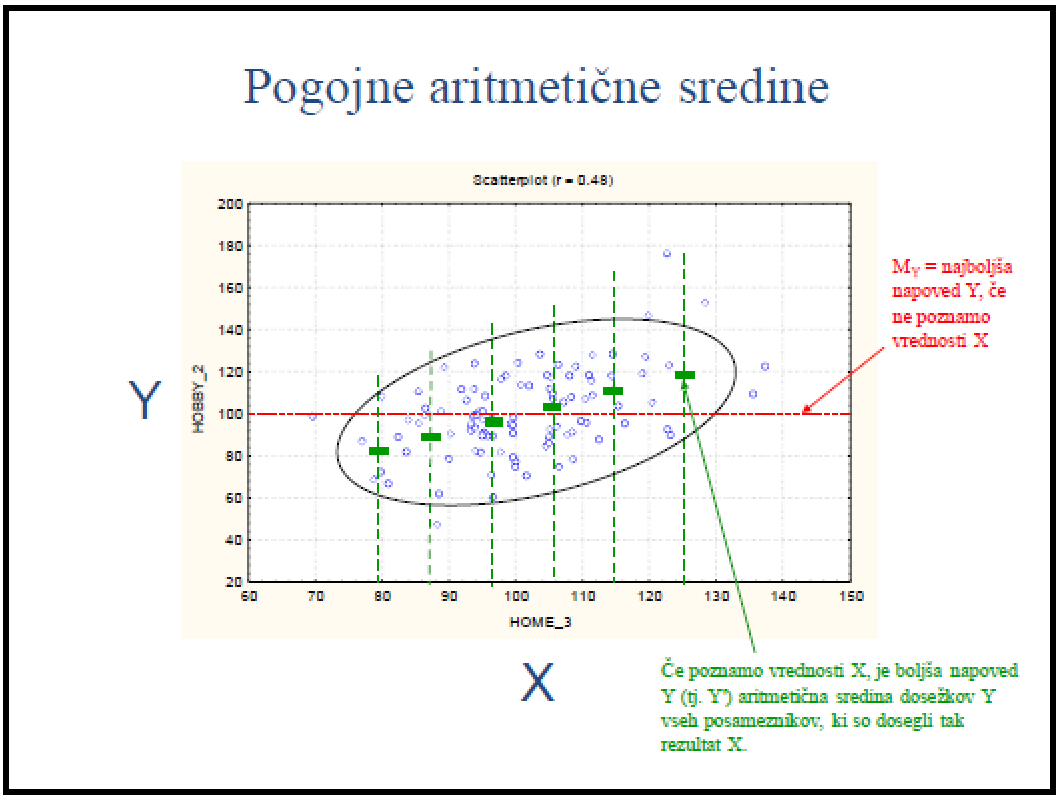
Gre za osnovni koncept linearne regresije.

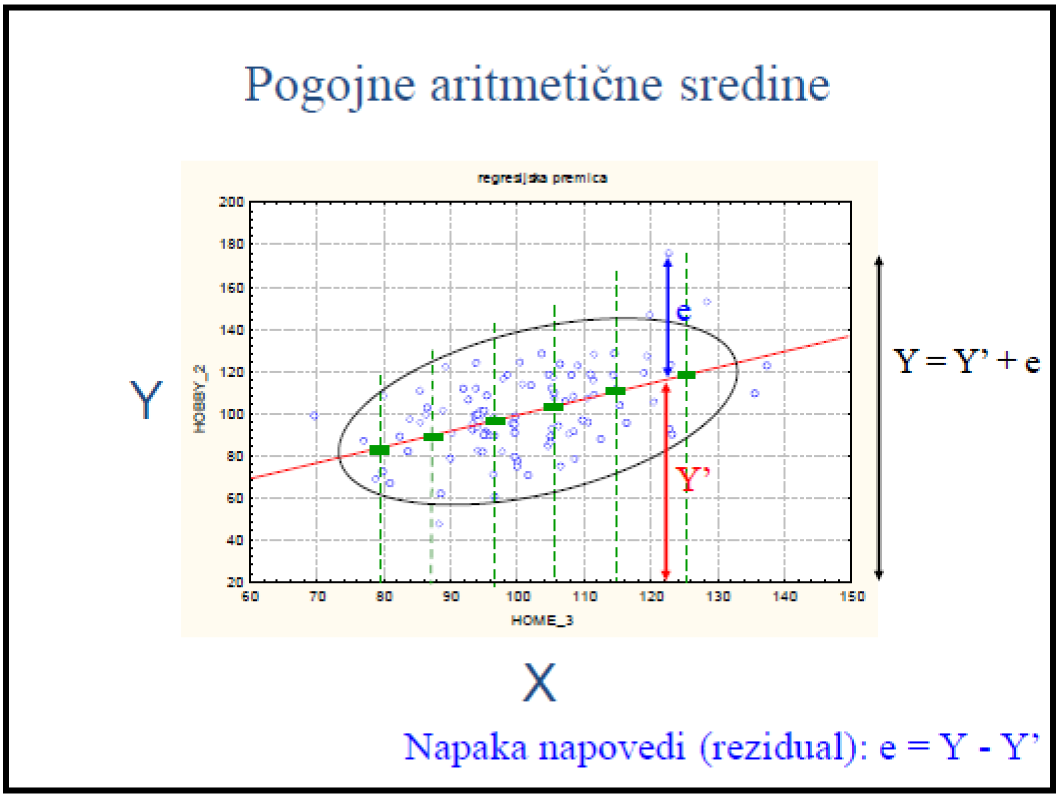
Kadar med spremenljivkama ni sistematičnega odnosa (korelacija je ničelna) je najbolje napovedati, da bo imel posameznik povprečno vrednost na drugi spremenljivke.

Pogojne aritmetične sredine označujemo z Y'.

Če je odnos med spremenljivkama linearen, lahko potegnemo premico, na kateri ležijo pogojne sredine.

Rezidual (napaka napovedi): različni posamezniki se razlikujejo v vrednosti kriterijske spremenljivke. Nekateri posamezniki odstopajo od pogoje sredine. Rezidual je razlika med dejansko vrednostjo posameznikovega kriterija in napovedano vrednostjo: e=Y-Y'.





Načelo najmanjših kvadratov

- da bi izvedeli, kako dobro se premica prilega podatkom, moramo definirati razdaljo med premico in vsako točko na grafu
- za vsak podatek X definiramo predvideno vrednost na Y (Y')
  - razdalja med predvideno vrednostjo in resnično je: $Y - Y'$ (ta razdalja pomeni napako med nekim točno določenim Y in predvidenim Y)
- Y' napovedujemo na osnovi pogojnih sredin
- linearna regresija: Y' ležijo na premici
- korelacijska premica potuje najbolj optimalno skozi korelacijski oblak
- regresijska premica je tista premica za katero je vsota razlike odklonov (rezidualov) minimalna: $\sum(Y - Y') = 0$
- vsota kvadratov odklonov (rezidualov) je minimalna: $\sum(Y - Y')^2$
  - kvadriramo zato, ker so nekatere razdalje pozitivne, druge pa negativne, tega pa nočemo, saj se odštejejo
  - vsoto uporabljamo zato, ker želimo dobiti splošen podatek za vse vrednosti
- standardna napaka napovedi = razpršitev dejanskih okoli napovedanih vrednosti

Napovedovanje

- za napovedovanje potrebujemo le enačbo premice: $Y' = a + bX$
- če poznamo načelo najmanjših kvadratov lahko določimo premico, ki se najbolje prilega (ki ima najmanjšo napako kvadratov
- ta premica pove, za vsako vrednost X najboljši Y → Y', ki najbolje napove Y

- a … vrednost Y, ko je X=0 (presečišče z ordinatno osjo, pove kakšna je vrednost kriterija, če je vrednost prediktorja 0)
- b … povečanj Y', ko X naraste za 1 enoto $b = {}^{\Delta Y}/_{\Delta X}$ (za koliko enot se spremeni kriterij, če se prediktor spremeni za eno enoto)

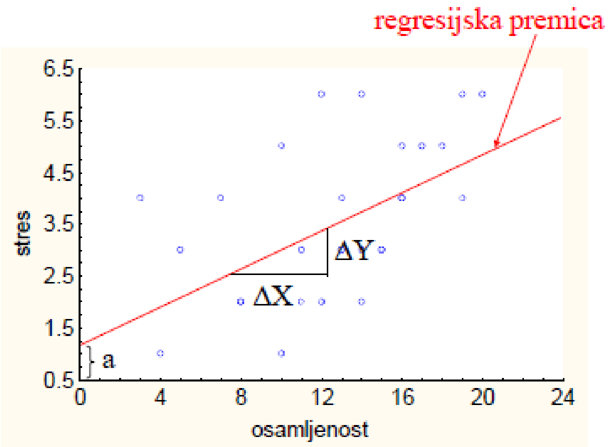
Za napovedovanje potrebujemo le enačbo premice:

$Y' = a + bX$

a:
vrednost Y,
ko je X = 0

b:
povečanje Y',
ko X naraste
za 1 enoto

$$b = \frac{\Delta Y}{\Delta X}$$



$$b_{YX} = \frac{N\sum XY - (\sum X)(\sum Y)}{N\sum X^2 - (\sum X)^2}$$

$$b_{YX} = \frac{\sum xy}{\sum x^2} = \frac{\mathrm{cov}_{XY}}{\sigma_X^2} = r_{YX}\frac{\sigma_Y}{\sigma_X} \qquad b = \frac{SP}{SS}$$

$$a_{YX} = \overline{Y} - b_{YX}\overline{X}$$

$$b_{XY} = \frac{N\sum XY - (\sum X)(\sum Y)}{N\sum Y^2 - (\sum Y)^2}$$

$$a_{XY} = \overline{X} - b_{XY}\overline{Y}$$

$SP \ldots sum\ of\ products$
$SS_x \ldots sum\ of\ squares\ (za\ X)$

opazimo, da je za izračun a in b potrebno imeti iste podatke kot za Pearsonov r:

$$r = \frac{SP}{\sqrt{SS_x SS_y}}$$

Primer 1. Napovedujemo vrednosti Y.

X	Y	X-Mx	Y-My	(X-Mx)(Y-My)	(X-Mx)^2
7	11	2	5	10	25
4	3	-1	-3	3	9

6	5	1	-1	-1	1
3	4	-2	-2	4	4
5	7	0	1	0	1
M=5	M=6			SP=16	SSx=10

$$SP = \sum (X - M_x)(Y - M_y) = 16$$

$$SS_x = \sum (X - M_x)^2 = 10$$

$$b = \frac{SP}{SS_x} = \frac{10}{16} = 1{,}6$$

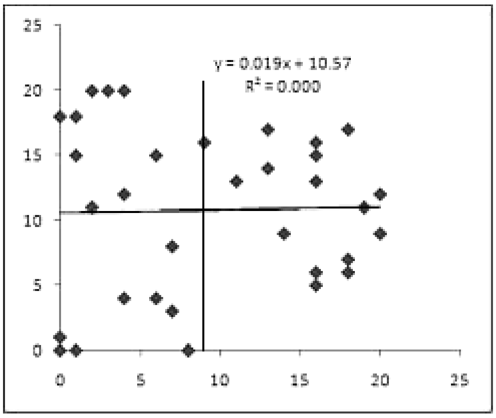
$$a = M_y - b * M_x = 6 - 1{,}6 * 5 = -2$$

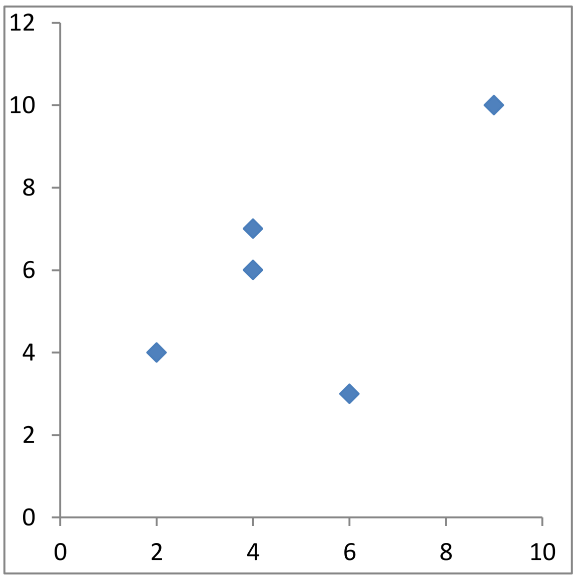
$$Y' = 1{,}6X - 2$$

$$\boxed{M_x = \bar{X}\ in\ M_y = \bar{Y}}$$

Kadar korelacija ni pomembna je zelo pomembno, ali napovedujemo X iz Y ali Y iz X!

Če je korelacija 0, je regresijska premica vodoravna, z vsakim X bi napovedali enak Y, enako, če bi napovedovali X iz Y, je pa premica popolnoma drugačna! Ti premici sta pravokotni.





Primer 2. Napovedovanje iz X in iz Y.

X	Y	XY	X^2
6	3	18	36
9	10	90	81
4	7	28	16
2	4	8	4
4	6	24	16
∑=25	∑=30	∑=168	∑=153
M=5	M=6		

$b_{yx} = \frac{5*168-25*30}{5*153-25^2} = 0{,}64$

ali: $b_{yx} = \frac{\sum(X-M_x)(Y-M_y)}{\sum(X-M_x)^2} = \frac{18}{28} = 0{,}64$

$a_{yx} = 6 - 0{,}64 * 5 = 2{,}80$

ter: $a_{yx} = M_y - b * M_x = 6 - 0{,}64 * 5 = 2{,}80$

$Y' = a_{yx} + b_{yx} * X = 2{,}80 + 0{,}64 * X$

$X' = a_{xy} + b_{xy} * Y$

$$b_{xy} = \frac{5 * 168 - 25 * 30}{5 * 216 - 30^2} = 0{,}60$$

ali: $b_{xy} = \frac{\sum(X-M_x)(Y-M_y)}{\sum(Y-M_y)^2} = \frac{18}{30} = 0{,}60$

$a_{xy} = 5 - 0{,}60 * 6 = 1{,}40$

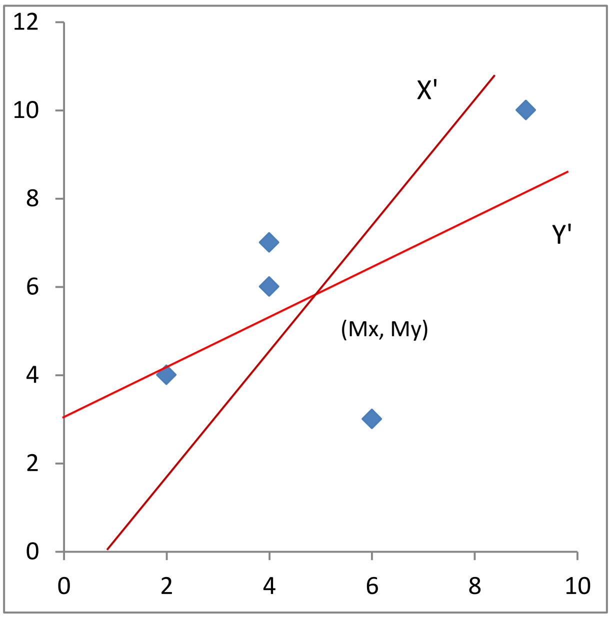
ter: $a_{xy} = M_x - b * M_y = 5 - 0{,}60 * 6 = 1{,}40$

$X' = 1{,}4 + 0{,}6 * Y$

$X' = a_{xy} + b_{xy} * Y' = 1{,}4 + 0{,}6 * (2{,}80 * 0{,}64 * X') = 1{,}4 + 1{,}68 + 0{,}384 * X' \rightarrow 5 = 5$

tudi: $b_{yx} = \frac{cov_{xy}}{\sigma_x^2} = \frac{3{,}6}{2{,}37^2} = 0{,}64$

(X-Mx)(Y-My)	(X-Mx)^2	(Y-My)^2
-3	1	9
16	16	16
-1	1	1
6	9	4
0	1	0
18	28	30



X	Y	XY	X^2	Y^2	Y'	Y-Y'	(Y-Y')^2	|Y-Y'|	(Y'-My)
6	3	18	36	9	6,64	-3,64	13,25	3,64	0,64
9	10	90	81	100	8,56	1,44	2,07	1,44	3,56
4	7	28	16	49	5,36	1,64	2,69	1,64	-0,64
2	4	8	4	16	4,08	-0,08	0,01	0,08	1,92
4	6	24	16	36	5,36	0,64	0,41	0,64	-0,64
25	30	168	153	210		0	18,43		

$$var = \frac{\sum e^2}{N} = 3{,}68$$

$\sigma = \sqrt{3{,}68} = 1{,}92$

$\sigma_y^2 = 6{,}0$

$\sigma_{y'}^2$=2,29 … standardna deviacija napovedi

$\sigma_e^2 = 3{,}36$

Standardizirane vrednosti

Kadar imamo opravka s standardiziranimi vrednostmi, je a=0, torej regresijska premica teče skozi koordinatno izhodišče.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-2,488	,695		-3,580	,001
	Normalna	1,128	,068	,859	16,599	,000

a. Dependent Variable: Spremenljivka

Correlations

		Spremenljivka	Normalna
Pearson Correlation	Spremenljivka	1,000	,859
	Normalna	,859	1,000
Sig. (1-tailed)	Spremenljivka	.	,000
	Normalna	,000	.
N	Spremenljivka	100	100
	Normalna	100	100

Nestandardizirane vrednosti: a=–2,49, b=1,13;

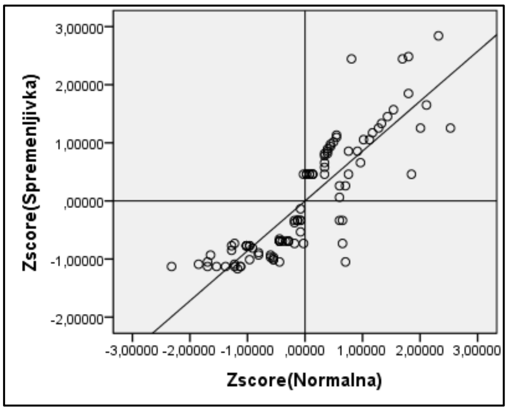
Pearsonova korelacija: r=,859

Standardiziran β koeficient: β=r=,859

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)		,051		,000	1,000
	Zscore(Normalna)	,859	,052	,859	16,599	,000

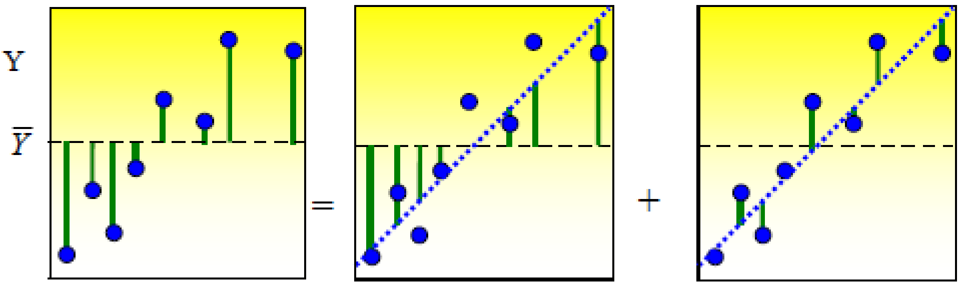
a. Dependent Variable: Zscore(Spremenljivka)



Če vrednosti obeh spremenljivk (Spremenljivke in Normalne) standardiziramo, vidimo, da postane naš b enak β oz. Pearsonovi korelaciji. Konstante a pa seveda ni (oz. je nesmiselna), saj premica poteka skozi koordinatno izhodišče.

Grafa sta enaka (tisti s standardiziranimi in tisti z empiričnimi), le da je koordinatno izhodišče »premaknjeno« in regresijska premica potuje čezenj. (Glej še str. 72!)

Testiranje hipotez I: Razstavljanje totalne vsote kvadratov



Totalna SS = Modelna (pojasnjena) SS + Nepojasnjena SS (napaka)

$$\sum_{i=1}^{N}(Y_i-\bar{Y})^2 \qquad \sum_{i=1}^{N}(\hat{Y}_i-\bar{Y})^2 \qquad \sum_{i=1}^{N}(Y_i-\hat{Y}_i)^2$$

$$\sigma_Y^{\,2}=\frac{\sum(Y-\bar{Y})^2}{N} \qquad \sigma_{\hat{Y}}^{\,2}=\frac{\sum(\hat{Y}-\bar{Y})^2}{N} \qquad \sigma_{Y.X}^2=\frac{\sum(Y-\hat{Y})^2}{N}$$

totalna *var* = pojasnjena *var* + nepojasnjena *var*

Standardna napaka nam da neko mero standardne razdalje med regresijsko premico in resničnimi vrednostmi (točkami) – zelo podobna je standardni deviaciji, saj obe povesta mero standardne razdalje, tudi njun izračun je podoben: $SD=\sqrt{var}=\sqrt{\frac{SS}{N}}=\sqrt{\frac{\sum(X-M)^2}{N}}$.

Računamo pojasnjeno varianco in pojasnjeno vsoto kvadratov:

Najprej varianca: $var=\frac{SS}{N}$

Potem vsota kvadratov (razdalje/<u>napake</u>):

$$SS_{(error)}=\sum(Y-Y')^2$$

Najdemo standardno napako napovedi (<u>nepojasnjeno</u>):

$$\sigma=\sqrt{\frac{SS_e}{N}}=\sqrt{\frac{\sum(Y-Y')^2}{N}}$$

Standardna napaka napovedi

Standardna napaka napovedi za populacijo

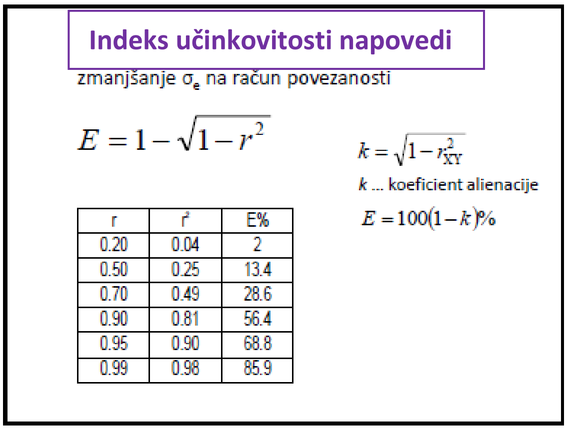
$SS_e=(1-r^2)\,SS_Y$ → večja kot je korelacija, manjša je SS_e

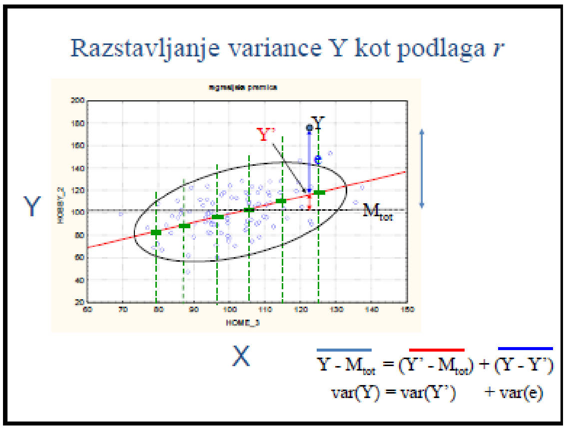
$$\sigma_{Y.X}=\sqrt{\frac{\sum(Y-\hat{Y})^2}{N}}=\sigma_Y\sqrt{1-r_{XY}^2}$$

Nepristranska ocena standardne napake napovedi na osnovi vzorčnih podatkov

$$\hat{\sigma}_{Y.X}=\sqrt{\frac{\sum(Y-\hat{Y})^2}{N-2}}=\sigma_Y\sqrt{1-r_{XY}^2}\sqrt{\frac{N}{N-2}}$$

💻 EXCEL: totalna varianca: =var(n); pojasnjena varianca: =var(Y'-M_Y)

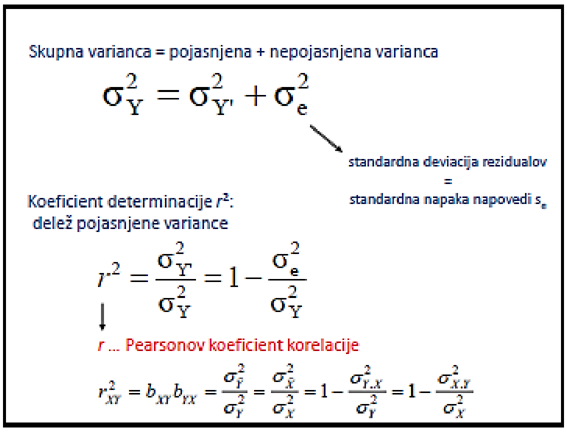


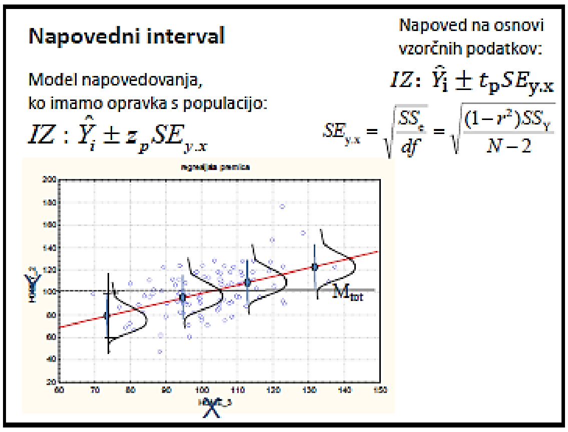


r^2 je t.i. **koeficient determinacije**, saj meri delež variabilnosti ene spremenljivke, ki je lahko določena z odnosom z drugo spremenljivko. Če je korelacija $r = 0{,}80\ (ali - 0{,}80)$, je $r^2 = 0{,}64\ (ali\ 64\ \%)$, kar pomeni, da lahko 64 % variabilnosti na Y napovemo iz X.

Koeficient determinacije pomeni razmerje pojasnjene variance in celotne variance. Zavzame vrednosti med 0 in 1. Če ga pomnožimo s 100, rezultat imenujemo **odstotek pojasnjene variance**. Koeficient determinacije je indikator linearnosti! **Koeficient alienacije** pa kaže pomanjkanje linearnosti med dvema spremenljivkama (tudi koeficient nedeterminacije).

Praktično koristnost korelacije si lažje predstavljamo, če izračunamo **indeks učinkovitosti napovedi**, ki nam pove, za koliko se zaradi korelacije zmanjša standardna napaka napovedi.



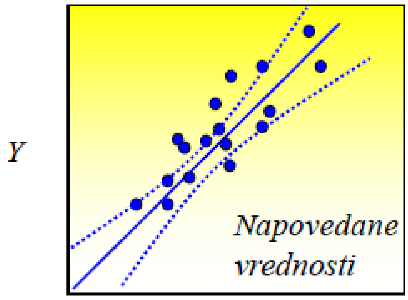


Koeficient determinacije: vrednost 0 pomeni, da med x in y ni nikakršne korelacije (pojasnjene variance sploh ni in torej x ne vpliva na y). Vrednost 1 pomeni, da je korelacija najmočnejša (nepojasnjene variance sploh ni, kar pomeni, da samo x vpliva na y). Pojasnjena varianca je posledica delovanja neodvisne spremenljivke (x), nepojasnjena varianca pa posledica delovanja ostalih vplivov.

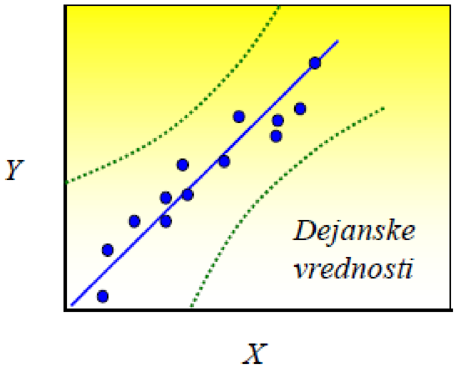
Napovedni interval: območje, v katerem se z določeno verjetnostjo giblje dejanska vrednost odvisne spremenljivke. Lahko jih imenujemo tudi intervali zaupanja, saj nam dajo neko približno oceno – večji kot je interval, večje je zaupanje, a hkrati manjša natančnost.

+ upoštevati vzorčne variacije v regresijskih koeficientih

95 % interval zaupanja za napovedane vrednosti
= interval, v katerem se bi pri 95 % vzorcev nahajala napovedana vrednost Y' pri nekem X



95 % napovedni interval za Y (včasih tudi interval zaupanja za dejanske [angl. *observed*] vrednosti)
= interval okrog napovedane vrednosti, v katerem se nahaja srednjih 95 % dejanskih vrednosti Y pri posameznikih, ki imajo določeno vrednost X



- Napovedni interval je večji kot interval zaupanja za napovedane vrednosti.
- Širina intervala zaupanja za dejanske in za napovedane vrednosti narašča z naraščanjem razdalje med M_X in X_i.

💻 EXCEL:

- standardna napaka napovedi: =STEYX(y-; x-i)
- spodnja in zgornja meja 95 % intervala zaupanja:
  - =NORMSINV(0,975) oz. =NORMSINV (0,025) – vseeno, saj računamo z dvema nasprotnima enačbama, ki kompenzirata za to, da izračunamo le eno Z vrednost, lahko pa bi izračunali obe in uporabili le eno formulo z obema ☺
  - Y'+z*SE_{yx} oz. Y'+z**STEYX*(y-i, x-i)
  - Y'–z*SE_{yx} oz. Y'–z**STEYX*(y-i, x-i)

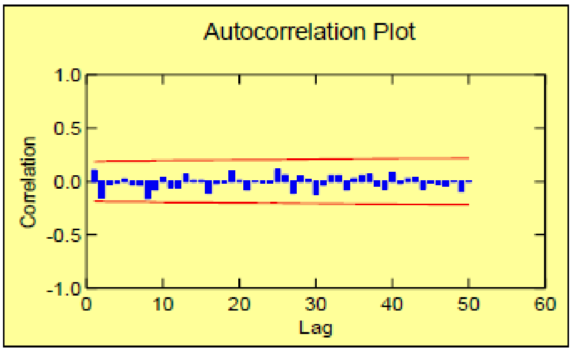
Predpostavke v regresijski analizi → analiza rezidualov + druga regresijska diagnostika

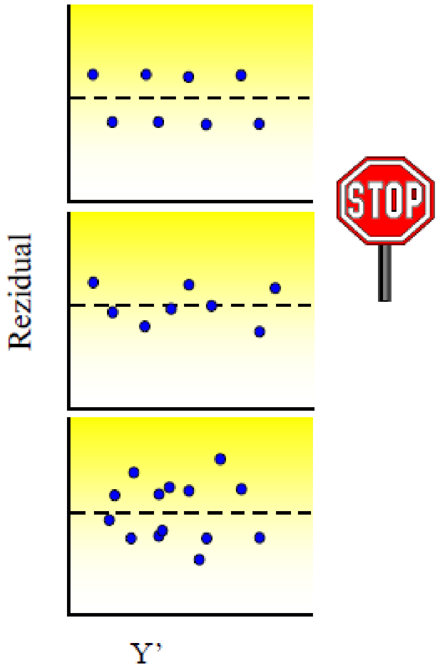
1. **naključno vzorčenje**: napake napovedi ne smejo biti korelirane (avtokolerirane)
2. **linearnost odnosa**: kakšni so grafi rezidualov, ali opazimo specifične vzorce?
3. **homoscedastičnost**: ali je standardna napaka napovedi enaka na celotnem razponu?
4. **normalnost porazdelitve rezidualov**: kakšen je odnos med reziduali in Y'?

Rezidual pomeni odklon oz. oddaljenost dejanske vrednosti od teoretične napovedane/ocenjene vrednosti.

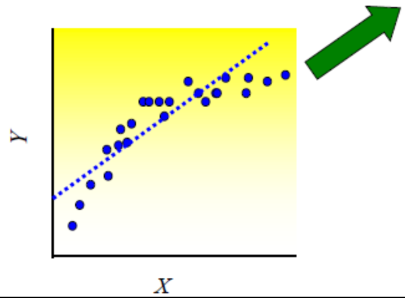
Naključno vzorčenje oz. neodvisnost podatkov

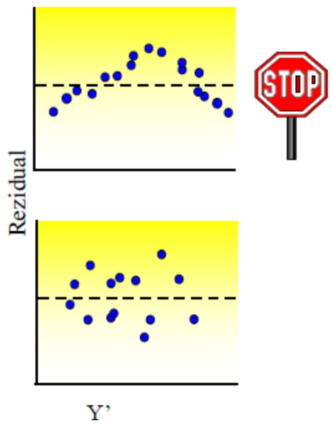
1. Napake napovedi ne smejo biti korelirane (problem: pri časovnih vrstah).
2. Mera t. i. serialne korelacije: Durbin-Watson
   - Zaseda lahko vrednosti od 0 do 4.
   - 2 = ni avtokorelacije,
     0 = pozitivna avtokorelacija, 4 = negativna avtokorelacija
3. Narišemo odnos med reziduali in napovedanimi vrednostmi. Pregledamo, če obstajajo kakšni vzorci. Naredimo lahko tudi graf ACF → analize časovnih vrst



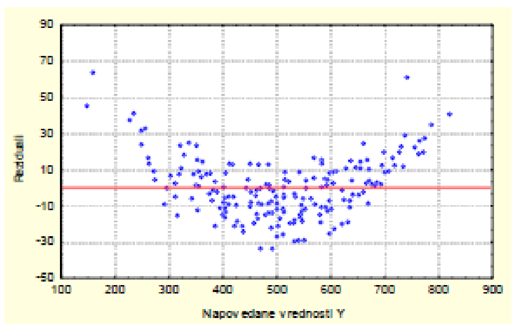


Za preverjanje linearnosti odnosa:
- Narišemo odnos med reziduali in Y'.
- Iščemo morebitne vzorce.



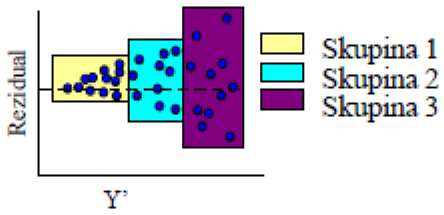


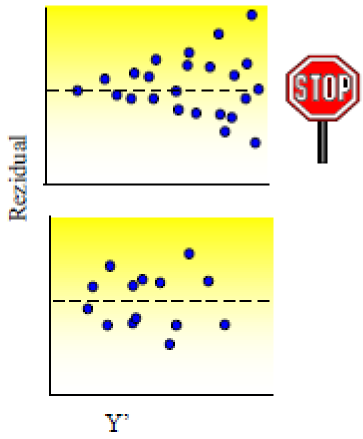
Primer rezidualnega grafa
pri nelinearni povezanosti



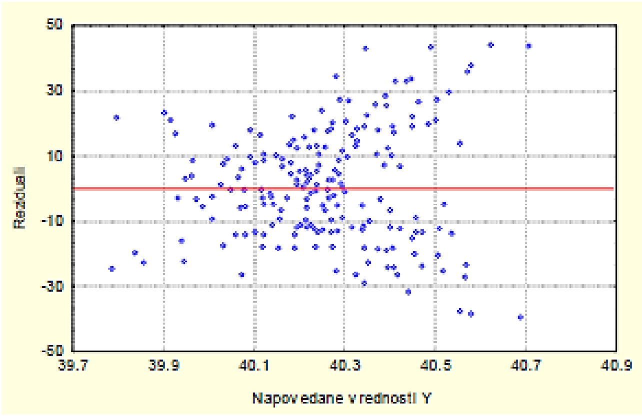
Homoscedastičnost =
Standardna napaka napovedi je enaka na celotnem razponu X.

- Narišemo odnos med reziduali in Y'. Iščemo morebitne vzorce.
- Y združimo v nekaj razredov in z Levenovim testom preverimo enakost varianc.

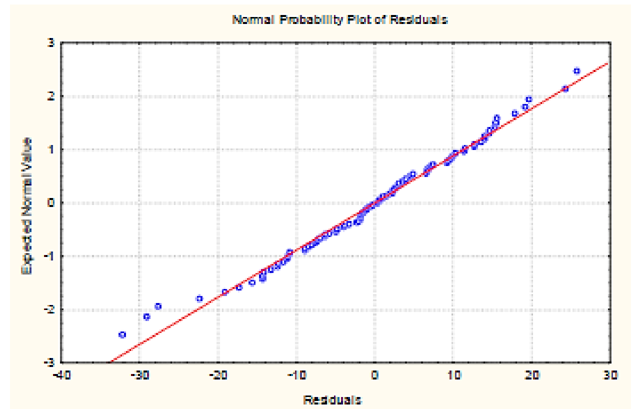


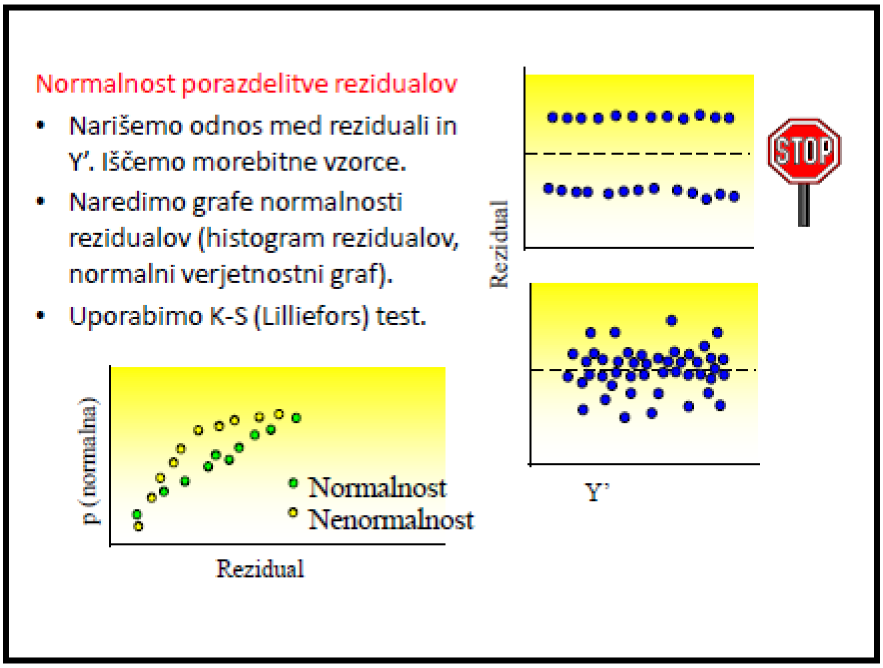


Primer rezidualnega grafa
pri heteroscedastičnosti



Primer rezidualnega grafa
pri normalni porazdeljenosti rezidualov



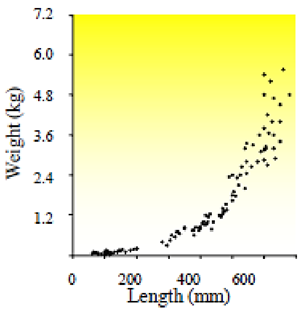


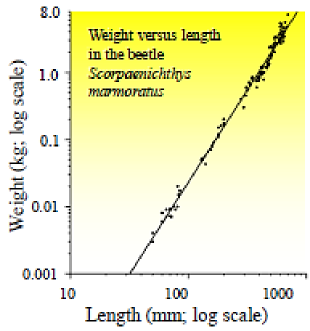
Robustnost regresije pri kršenju predpostavk

Predpostavka	Na kaj vpliva kršenje?	Robustnost	Opomba
Normalnost	Na inferenčne teste, intervale zaupanja	Visoka	Le če je *N* dovolj velik (>10)
Neodvisnost	Na inferenčne teste	Nizka	Odvisno od višine avtokorelacije
Homoscedastičnost	Na intervale zaupanja za regresijske koeficiente in intervale napovedi, na inferenčne teste	Nizka	Posebej pri majhnih vzorcih
Linearnost	Na interpretacijo *r* in regresijskih koeficientov	Nizka	Prepričaj se!

Kaj lahko naredimo, če so predpostavke kršene?

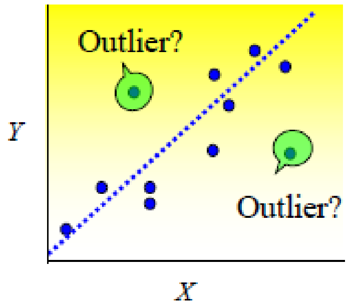
- Lahko poskusimo transformirati podatke, toda:
  - pri nekaterih podatkih ne bo uspešna nobena transformacija;
  - včasih dobro transformacijo težko najdemo.
- Uporabimo nelinearno regresijo.





Izstopajoče in vplivne točke

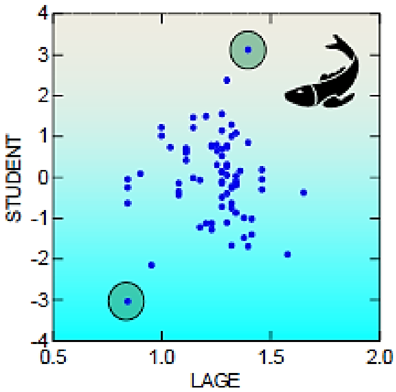
- točke, ki so močno oddaljene od regresijske premice
- *Problem 1*: Ali so to res izstopajoče (posebne, drugačne) vrednosti (angl. *outliers*)?
  - Pregled rezidualov (surovih, standardiziranih, studentiziranih)
- *Problem 2*: Ali pomembno vplivajo na statistične zaključke (angl. *influential points*)?
  - Pregled različnih mer (DFFit, DFBeta, Cookova razdalja, ročica, Mahalanobisova razdalja)

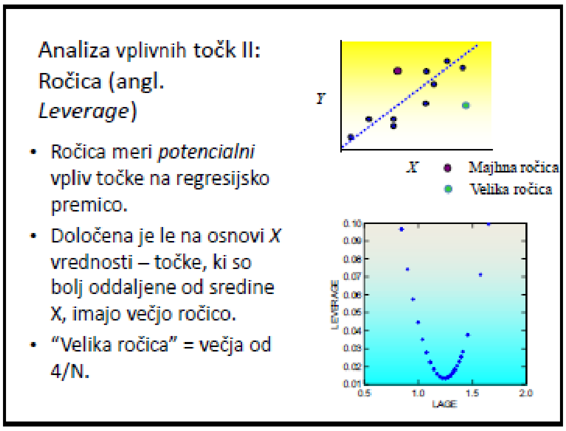


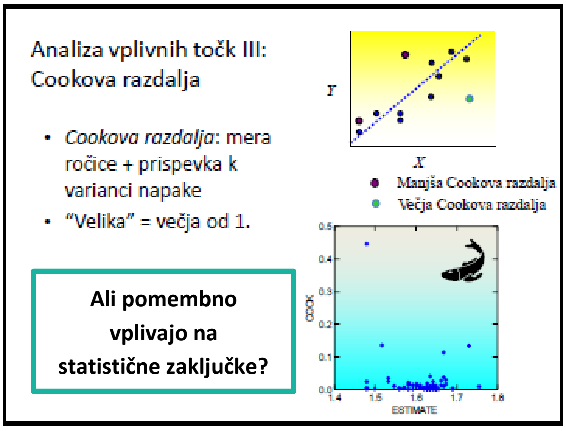
Analiza vplivnih točk I: Studentizirani reziduali

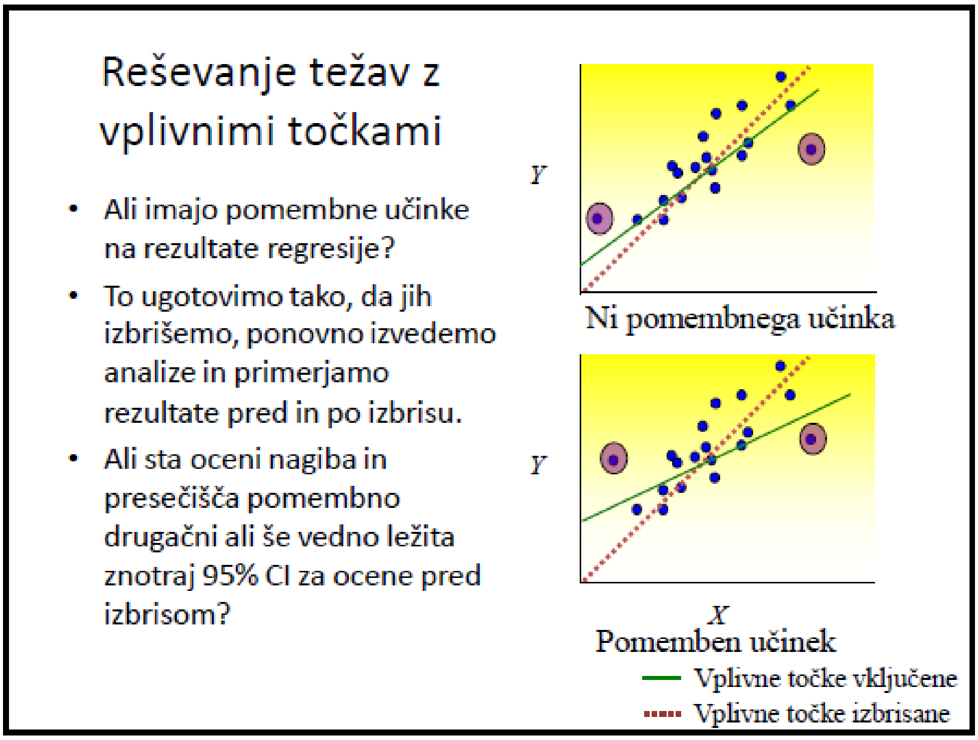
Ali so to res izstopajoče vrednosti?

- Narišemo odnos med *Studentiziranimi reziduali* in napovedanimi vrednostmi.
- "Veliki" reziduali so tisti, ki imajo studentizirano vrednost > 3,0.
- Taki primeri močno prispevajo k nepojasnjeni varianci (varianci napake).









Učinek brisanja vplivnih točk

Ponovimo: Posamezni primeri lahko nesorazmerno vplivajo na velikost korelacijskih in regresijskih koeficientov.

Vplivnost je odvisna od oddaljenosti točke od:

- aritmetične sredine X in
- regresijske premice

Kaj se lahko zgodi po brisanju vplivnih točk?

- Zmanjša se velikost vzorca (*N*) in s tem tudi moč analize.
- Zmanjša se MS_e, s tem se zmanjša s_b in poveča moč.
- Če je *N* majhen, je prvi učinek najbrž večji od drugega, razen če so vplivne točke zelo ekstremne.

Razni korelacijski koeficienti

Korelacija spremenljivk,ki niso merjene na intervalni merski ravni

ena od sprem. / druga od sprem.	nominalna	ordinalna
nominalna	n.d.-n.d.: Fi koeficient u.d.-u.d.: Tetrahorični koef. Več kategorij: koeficient kontingence *C*, Cramerjev *V*	Kendallov tau
ordinalna	Kendallov tau	Spearmanov r Kendallov tau
intervalna	n.d.- Točkovno biserialni koef. u.d.- Biserialni koef.	Spearmanov r Kendallov tau

Opombe: n.d.: naravno dihotomna; u.d.: umetno dihotomizirana

☞ Kendallov koeficient konkordance W - skladnost med ocenjevalci

Spearmanov koeficient korelacije rangov

- ρ ali r_S
- Pri **ordinalnih spremenljivkah**
- Rangiramo podatke, izračunamo Pearsonov koeficient korelacije na rangih
- Meri konsistentnost smeri povezave (t. i. monotonost povezave)
- Uporaba pri **nelinearnih odnosih**
- $d = R_y - R_x$ ... razlika med rangoma X in Y

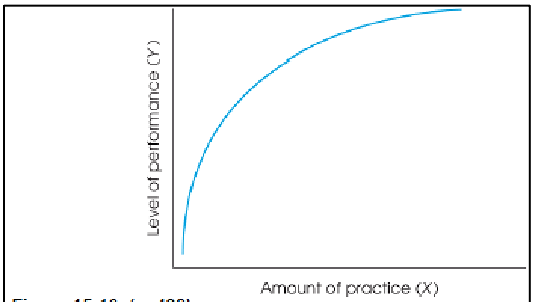


Figure 15-10 (p. 432)
Hypothetical data showing the relationship between practice and performance. Although this relationship is not linear, there is a consistent positive relationship. An increase in performance tends to accompany an increase in practice.

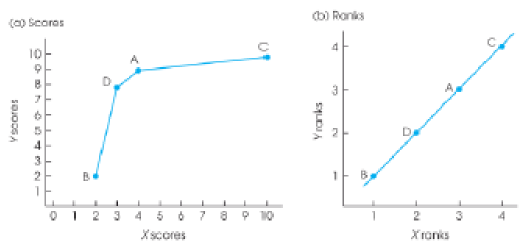
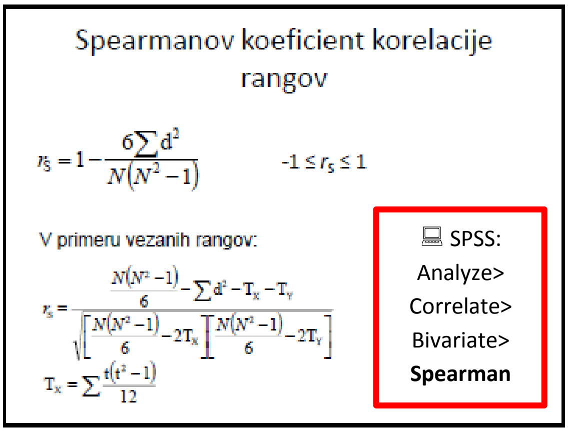


Figure 15-11 (p. 433)
Scatterplots showing (a) the scores and (b) the ranks for the data in Example 15.8. Notice that there is a consistent, positive relationship between the X and Y scores, although it is not a linear relationship. Also notice that the scatterplot of the ranks shows a perfect linear relationship.
Essentials of Statistics for the Behavioral Sciences, 5/E by Frederick J Gravetter and Larry B. Wallnau
Copyright © 2005 by Wadsworth Publishing, a division of Thomson Learning



Ko imamo opravek z rangi se moramo zavedati, da le-ti niso intervalna spremenljivka (ampak ordinalna), zato Pearsonov korelacijski koeficient ne ustreza. Korelacija med rangi je pozitivna, kadar velja, da višje kot so rangirane enote na eni spremenljivki, višje so enote na drugi spremenljivki. Če je obratno pa je negativna. Jakost korelacije nam pove, kako pogosto se rangi za enote ujemajo. Za nevezane range je značilno, da ima vsaka oseba drugačen rang, pri vezanih pa imata dve osebi lahko isti rang – to so vezani rangi.

Spearmanov koeficient korelacije rangov

Oseba	R(X)	R(Y)	d	d²
A	1	3	-2	4
B	4	5	-1	1
C	3	4	-1	1
Č	2	1	1	1
D	5	2	3	9

$d = R(Y) - R(X)$

$$r_S = 1 - \frac{6\sum d^2}{N(N^2-1)} = 1 - \frac{6 \cdot 16}{5 \cdot 24} = 0{,}20$$

Spearmanov koeficient korelacije rangov

Oseba	R(X)	R(Y)	d	d²
A	1	3	-2	4
B	3	5	-2	4
C	3	4	-1	1
Č	3	1,5	1,5	2,25
D	5	1,5	3,5	12,25

$$T_X = \sum \frac{t(t^2-1)}{12} = \frac{3 \cdot (3^2-1)}{12} = 2$$

$$T_Y = \sum \frac{t(t^2-1)}{12} = \frac{2 \cdot (2^2-1)}{12} = 0{,}5$$

$$r_S = \frac{\frac{N(N^2-1)}{6} - \sum d^2 - T_X - T_Y}{\sqrt{\left[\frac{N(N^2-1)}{6} - 2T_X\right]\left[\frac{N(N^2-1)}{6} - 2T_Y\right]}} = \frac{\frac{5(5^2-1)}{6} - 23{,}5 - 2 - 0{,}5}{\sqrt{\left[\frac{5(5^2-1)}{6} - 2 \cdot 2\right]\left[\frac{5(5^2-1)}{6} - 2 \cdot 0{,}5\right]}} = -0{,}34$$

http://faculty.vassar.edu/lowry/corr_rank.html

Kendallov τ

$$\tau = \frac{S}{S_{max}} = \frac{2S}{N(N-1)}$$

S = število ujemanj – število zamenjav
$-1 \le \tau \le 1$

V primeru vezanih rangov:

$$\tau = \frac{S}{\sqrt{\left[\frac{1}{2}N(N-1) - T_X\right]\left[\frac{1}{2}N(N-1) - T_Y\right]}}$$

$$T_X = \frac{\sum t(t-1)}{2}$$

t ... št. vezanih vrednosti

SPSS:
Analyze>
Correlate>
Bivariate>
Kendall's tau-b

Kendallov τ – primer

Oseba	R(X)	R(Y)
A	1	3
B	4	5
C	3	4
Č	2	1
D	5	2

Pari oseb – Razvrstitev rangov na X in Y
AB – ujemanje v razvrstitvi oseb
AC – ujemanje
AČ – neujemanje
AD – neujemanje
BC – ujemanje
BČ – ujemanje
BD – neujemanje
CČ – ujemanje
CD – neujemanje
ČD – ujemanje

Skupaj: 6 ujemanj, 4 neujemanja

$$\tau = \frac{S}{S_{max}} = \frac{6-4}{10} = 0{,}20$$

$$\tau = \frac{2S}{N(N-1)} = \frac{2(6-4)}{5(5-1)} = 0{,}20$$

http://www.wessa.net/rwasp_kendall.wasp

Kendallov τ – primer z vezanimi rangi

Oseba	R(X)	R(Y)
A	1	3
B	3	5
C	3	4
Č	3	1
D	5	2

$$T_X = \frac{\sum 3(3-1)}{2} = 3 \qquad T_Y = 0$$

Pari oseb – Razvrstitev rangov na X in Y
AB – ujemanje v razvrstitvi oseb
AC – ujemanje
AČ – neujemanje
AD – neujemanje
BC – neodločeno
BČ – neodločeno
BD – neujemanje
CČ – neodločeno
CD – neujemanje
ČD – ujemanje

Skupaj: 3 ujemanja, 4 neujemanja

$$\tau = \frac{S}{\sqrt{\left[\frac{1}{2}N(N-1) - T_X\right]\left[\frac{1}{2}N(N-1) - T_Y\right]}} = \frac{3-4}{\sqrt{\left[\frac{1}{2}5(5-1) - 3\right]\left[\frac{1}{2}5(5-1) - 0\right]}} = -0{,}12$$

Kendallov τ

- Primeren tudi za opisovanje odnosa nominalne in ordinalne spremenljivke

Oseba	R(X)	R(Y)
A	1	3
B	1	5
C	1	4
Č	0	1
D	0	2

Pari oseb – Razvrstitev rangov na X in Y
AB – neodločeno
AC – neodločeno
AČ – ujemanje (arbitrarno)
AD – ujemanje
BC – neodločeno
BČ – ujemanje
BD – ujemanje
CČ – ujemanje
CD – ujemanje
ČD – neodločeno

Skupaj: 6 ujemanj, 0 neujemanj

$$\tau = \frac{S}{\sqrt{\left[\frac{1}{2}N(N-1) - T_X\right]\left[\frac{1}{2}N(N-1) - T_Y\right]}} = \frac{6-0}{\sqrt{\left[\frac{1}{2}5(5-1) - 4\right]\left[\frac{1}{2}5(5-1) - 0\right]}} = 0{,}77$$

Kendallov tau je neparametričen test uporabljen pri **ordinalnih spremenljivkah** (druga spremenljivka je lahko na kateri koli merski ravni). Meri korelacijo med **rangi**, torej podobnost v porazdelitvi, ko so podatki razvrščeni po vrsti. Ko ugotavljamo ujemanje in neujemanje med posameznimi rangi, gledamo (X_1, Y_1) oz. ($R(X_1)$, $R(Y_1)$), (X_2, Y_2) oz. ($R(X_2)$, $R(Y_2)$) itn. Če gre za ujemanje velja $R(X_1) > R(Y_1)$ in $R(X_2) > R(Y_2)$ ali obratno.

Kendallov koeficient konkordance *W*

- Skladnost ocenjevalcev

Os./objekt	A	B	C	D
A. A.	1	2	3	4
B. B.	2	1	3	4
C. C.	2	3	4	1
Č. Č.	3	1	4	2
D. D.	1	2	4	3
Rj	9	9	18	14
$(R_j - \frac{\sum R_j}{N})^2$	12,25	12,25	30,25	2,25

$\sum R_j = 50 \rightarrow povprečje\ 12{,}5$

$\sum = 57$

$$S = \sum_{j=1}^{N}\left(R_j - \frac{\sum R_j}{N}\right)^2$$

$$W = \frac{S}{S_{max}} = \frac{12S}{m^2(N^3 - N)}$$

R_j ... vsota rangov objekta j
m ... število ocenjevalcev
N ... število objektov

$0 \le W \le 1$

$$W = \frac{S}{S_{max}} = \frac{12S}{m^2(N^3 - N)} = \frac{12 \cdot 57}{5^2(4^3 - 4)} = 0{,}46$$

http://www.stattools.net/KendallW_Pgm.php

Točkovno biserialni koeficient korelacije

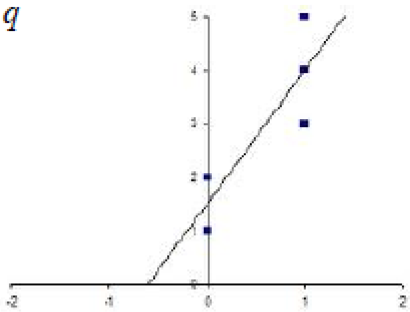
- Intervalna spremenljivka – nominalna (naravno dihotomna) spremenljivka.

$$r_{pb} = \frac{M_p - M_q}{\sigma_{tot}}\sqrt{pq} = \frac{M_p - M_{tot}}{\sigma_{tot}}\sqrt{\frac{p}{q}}$$

$$r_{pb} = \frac{N\sum Y_p - N_p \sum Y}{\sqrt{N_p N_q\left[N\sum Y^2 - \left(\sum Y\right)^2\right]}}$$

$-1 \le r_{pb} \le 1$

SPSS: Analyze>Correlate>Bivariate>**Pearson**



Točkovni biserialni koeficient uporabljamo, kadar je **neodvisna spremenljivka nominalna**. Same točkovne biserialne koeficiente med seboj lahko primerjamo, izogibamo pa se primerjavi z drugimi! Zavzema vrednosti med -1 in 1 Po vrednosti teh sredin interpretiramo smer. Točkovni biserialni koeficient označujemo kot r_{pb}. EXCEL: =CORREL(matrika1;matrika2)

M (1)	26
M (0)	18
SD (tot)	5,83
p	0,5
q	0,5
r_{pb}	0,69

Data for the independent-measures t
The data consist of two separate samples with $n = 10$ scores in each sample.

Data (number of words recalled)			
Group 1 (images)		Group 2 (no images)	
19	32	23	12
20	30	22	16
24	27	15	19
30	22	16	14
31	25	18	25
$n = 10$		$n = 10$	
$M = 26$		$M = 18$	
$SS = 200$		$SS = 160$	

Data for the point-biserial correlation
The data consist of two scores (X and Y) for each of the 20 participants.

Participant	Memory score	Treatment condition
A	19	1
B	20	1
C	24	1
D	30	1
E	31	1
F	32	1
G	30	1
H	27	1
I	22	1
J	25	1
K	23	0
L	22	0
M	15	0
N	16	0
O	18	0
P	12	0
Q	16	0
R	19	0
S	14	0
T	25	0

Table 15-1 (p. 430)
The same data are organized in two different formats. On the left-hand side, the data appear as two separate samples (Group 1 and Group 2), appropriate for an independent-measures t hypothesis test. On the right-hand side, the same data are shown as a single sample, with two scores for each individual: the original memory score (X) and a dichotomous score (Y) that designates the treatment condition (images or no-images) in which the participant is located. The data on the right are appropriate for a point-biserial correlation.

Essentials of Statistics for the Behavioral Sciences, 5/E by Frederick J Gravetter and Larry B. Wallnau
Copyright © 2005 by Wadsworth Publishing, a division of Thomson Learning

Biserialni koeficient korelacije

- Intervalna spremenljivka – umetno dihotomizirana spremenljivka (s kodiranjem 0 in 1 aproksimiramo zvezno spremenljivko, ki se normalno porazdeljuje; 0 … X < *Mdn*, 1 … X > *Mdn*).

$$r_b = \frac{M_p - M_q}{\sigma_{tot}} \cdot \frac{pq}{y} = \frac{M_p - M_{tot}}{\sigma_{tot}} \cdot \frac{p}{y} = r_{pb} \frac{\sqrt{pq}}{y}$$

$-1 \leq r_b \leq 1$

y … višina ordinate v standardizirani normalni krivulji na točki, ki deli ploščino pod krivuljo na *p* in *q*

Biserialni koeficient korelacije uporabimo kadar proučujemo **vpliv <u>umetno</u> dihotomne spremenljivke**, ki ima le **dve stopnji** (kategoriji). Odvisna spremenljivka mora biti vsaj **intervalna**, njene vrednosti pa morajo biti porazdeljene približno normalno (včasih tudi ordinalne, npr. dosežek na testu, obravnavamo kot intervalne). Neodvisna spremenljivka ne sme imeti več kot dve stopnji. Tudi pojav, ki ga ta izraža, mora biti normalen, pa tudi zvezen. Poleg tega mora biti korelacija med spremenljivkama linearna. Simbol je r_b. Korelacija pri biserialnem je lahko tudi negativna.

Biserialni koeficient korelacije

X	Y		
1	0	*M* (0)	4,25
3	0	*M* (1)	4,75
5	0	*SD* (tot)	2,29
8	0	*p*	0,5
2	1	*q*	0,5
4	1	*z*	0
6	1	*y*	0,3989
7	1	r_b	0,14

X	Y		
5	0	*M* (0)	4
2	0	*M* (1)	6,75
5	0	*SD* (tot)	1,73
4	0	*p*	0,5
6	1	*q*	0,5
6	1	*z*	0
7	1	*y*	0,3989
8	1	r_b	1,00

X	Y		
1	0	*M* (0)	2,75
2	0	*M* (1)	6,25
3	0	*SD* (tot)	2,29
5	0	*p*	0,5
4	1	*q*	0,5
6	1	*z*	0
7	1	*y*	0,3989
8	1	r_b	0,96

X	Y		
1	0	*M* (0)	2,5
2	0	*M* (1)	6,5
3	0	*SD* (tot)	2,29
4	0	*p*	0,5
5	1	*q*	0,5
6	1	*z*	0
7	1	*y*	0,3989
8	1	r_b	1,09

$$r_b = \frac{M_p - M_q}{\sigma_{tot}} \cdot \frac{pq}{y} = \frac{M_p - M_{tot}}{\sigma_{tot}} \cdot \frac{p}{y} = r_{pb}\frac{\sqrt{pq}}{y}$$

$-1 \leq r_b \leq 1$

EXCEL: r_b: =r_{pb}*SQRT(p*q)/Y; **Y**: =NORM.DIST(x;0;1;FALSE); **x**: =NORMSINV(p)

Φ koeficient

- Dve naravno dihotomni spremenljivki

$$\Phi = \frac{AD - BC}{\sqrt{(A+B)(C+D)(A+C)(B+D)}} = \frac{p_{XY} - p_X p_Y}{\sqrt{p_X q_X p_Y q_Y}}$$

$-\phi_{max} \leq \phi \leq \phi_{max}$

Interpretacija ϕ

$$\Phi_{max} = \sqrt{\frac{p_m q_v}{p_v q_m}} \qquad p_v > p_m$$

pomen predznaka

$$\chi^2 = \sum \frac{(f_o - f_t)^2}{f_t}$$

$\chi^2 = N\phi^2$

	0	1	
0	15 (A)	30 (B)	45
1	5 (C)	25 (D)	30
	20	55	75

$$\Phi = \frac{AD - BC}{\sqrt{(A+B)(C+D)(A+C)(B+D)}} = \frac{15 \cdot 25 - 30 \cdot 5}{\sqrt{45 \cdot 30 \cdot 20 \cdot 55}} = 0{,}18 \qquad \Phi_{max} = \sqrt{\frac{p_X q_Y}{p_Y q_X}} = 0{,}49$$

Gre za posebno obliko Cramérjevega koeficienta za tabele 2×2. Pogosto ga uporabljamo s Cramerjevim hkrati. S Pearsonom je slabo primerljiv.

Tetrahorični koeficient korelacije

- Dve umetno dihotomizirani spremenljivki

$$r_{tet} \approx \cos\left(\frac{\pi}{1+\sqrt{\frac{AD}{BC}}}\right)$$

A, D: ujemanja; B, C: neujemanja; enačba velja za radiane, če računamo s stopinjami, je v števcu 180°

	0	1	
0	5	30	35
1	45	30	75
	50	60	110

$r_b = -0{,}71$

Tetrahorični koeficient korelacije uporabljamo, kadar sta obe spremenljivki **umetno dihotomni**. Obe spremenljivki sta **ordinalni** (glej še tabelo, kjer piše, da sta nominalni) in imata po dve stopnji (kategoriji). Veljajo tudi vsi ostali pogoji kot pri biserialnem korelacijskem koeficientu. Tetrakorični koeficient je primerljiv s Pearsonovim (kot tudi biserialni v nekaterih primerih). Podatke o dveh takih spremenljivkah urejamo v tabele 2×2. Kategoriji A in D morata biti frekvenci ujemajočih se skupin (pri obeh skupinah boljši in pri obeh slabši).

Primer. Mentor fotografskega krožka je učence razvrstil po prizadevnosti (X) na bolj prizadevne (BP) in manj prizadevne (MP) ter po uspešnosti (Y) na bolj uspešne (BU) in manj uspešne (MU).

x \ y	MU	BU	skupaj
BP	4	10	14
MP	12	5	17
skupaj	16	15	31

Boljših po obeh spremenljivkah je 10 in slabših po obeh 12. To sta frekvenci za števec pod korenom! Za imenovalec posledično ostaneta 4 in 5.

$$r_{tet} = \cos\left(\frac{\pi}{1+\sqrt{\frac{10*12}{4*5}}}\right) = 0{,}61$$

Koeficient kontingence *C*

- Dve kategoriji ali več

$$C = \sqrt{\frac{\chi^2}{N + \chi^2}} \qquad \chi^2 = \sum \frac{(f_e - f_t)^2}{f_t}$$

$0 \leq C \leq C_{max}$

kvadratne tabele:

$$C_{\max} = \sqrt{\frac{m - 1}{m}}$$

$$C_{\text{kor}} = \frac{C}{C_{\max}}$$

	psih.	soc.	filoz.	
kolo	20	35	30	85
avtobus	40	15	30	85
	60	50	60	170

$\chi^2 = 14{,}67$

$C = 0{,}28$

Za ugotavljanje povezanosti **nominalnih spremenljivk** lahko uporabimo koeficient kontingence. Če posamezna opisna spremenljivka nima le dveh možnih vrednosti, ampak več, pri analizi povezanosti med spremenljivkama govorimo o kontingenci. Vrednosti obeh spremenljivk uredimo hkrati v eni preglednici, ki jo imenujemo kontingenčna preglednica.

Vrednosti tega koeficienta so med 0 in 1. Na splošno pri opisnih spremenljivkah dobimo nižje stopnje povezanosti kor pri številskih. Zato že nižje koeficiente interpretiramo kot pomembno povezane. Poleg tega nimajo (kontingenčni koeficienti na splošno) predznaka. Pri nominalnih namreč ni ne pozitivne ne negativne smeri, zato je to ok. Smer presojamo glede na tabelčne vrednosti.

Analitično bomo v takem primeru povezanost med spremenljivkama obravnavali z:

- vrednostjo χ2 (hi kvadrat)
- koeficientom kontingence
- popravljenim koeficientom kontingence

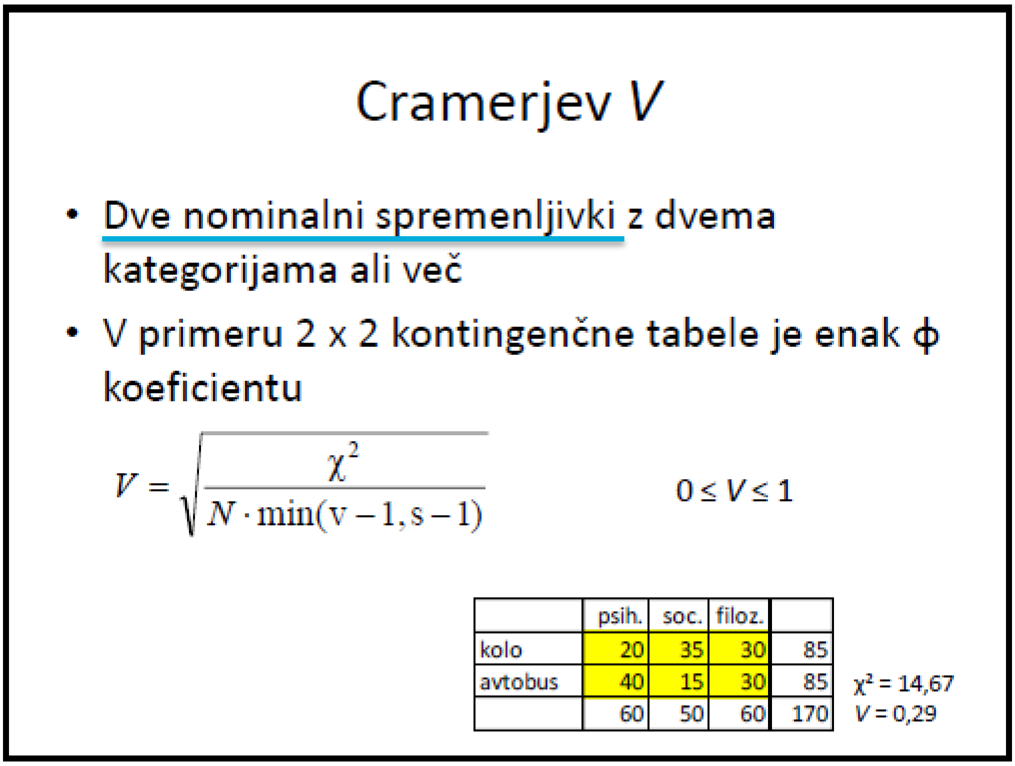
χ2-preizkus – preizkus hipoteze neodvisnosti

Preiskujemo povezanost med dvema opisnima spremenljivkama. Vprašanje: Ali sta v osnovni množici povezani? Postavimo hipotezo neodvisnosti: Spremenljivki v osnovni množici sta neodvisni. Neodvisna spremenljivka ima le dve kategoriji. Potrebujemo frekvenčno tabelo s stvarnimi/ empiričnimi frekvencami (f_e). Vse frekvence preračunamo v procente. Če sta spremenljivki neodvisni pričakujemo, da so procenti po stolpcih enaki (kot »skupaj«) – to so teoretične frekvence (f_t). Če so frekvence zelo različne, sklepamo na odvisnost in obratno.

Odstotne frekvence					*Pričakovane frekvence*				
	za	vseeno	proti	skupaj		za	vseeno	proti	skupaj
moški	26,8	12,2	61,0	100	**moški**	45,0	17,5	37,5	100
ženske	54,4	20,3	25,3	100	**ženske**	45,0	17,5	37,5	100
skupaj	45,0	17,5	37,5	100	**skupaj**	45,0	17,5	37,5	100

$$\chi 2 = \sum \frac{(f_E - f_T)^2}{f_T}$$

SPSS: Analyze>**Descriptive Statistics**>**Crostabs**>Statistics>Contingency Coefficient



Uporabljamo ga, ko računamo korelacijo po tabelah, ki imajo več kot 2×2 vrstice in stolpce. Uporablja se po tem, ko **chi²** (c^2; »chi-square test«) pokaže pomembnost, da določimo moč povezave.

SPSS: Analyze>**Descriptive Statistics**>**Crostabs**>Statistics>Phi and Cramer's V

SPSS: Analyze>**Descriptive Statistics**>**Crostabs**>Statistics>Chi-square