

Opisna statistika

Inštitut za biostatistiko in medicinsko informatiko
Medicinska fakulteta, Univerza v Ljubljani

Primer – pulz pred in po obremenitvi

- Populacija: študenti Univerze v Queenslandu
- Vzorec: študenti pri predmetu Osnove statistike
- Spremenljivke:
 - Višina
 - Teža
 - Starost
 - Spol (1 – moški, 2 – ženska)
 - Kajenje (0 – ne, 1 – da)
 - Alkohol (0 – ne, 1 – da)
 - Rekreacija (1 – malo, 2 – povprečno, 3 – veliko)
 - Obremenitev (0 – ne, 1 – da)
 - Pulz1 (pred obremenitvijo, udarci/min)
 - Pulz2 (po obremenitvi)
 - Leto (1993 do 1998)

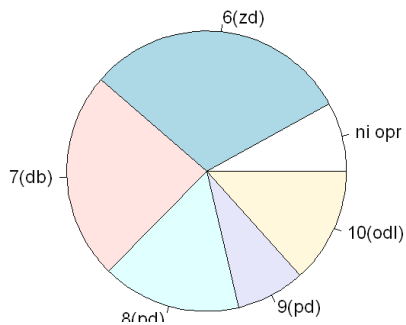
Spremenljivke

Opisne

Številске

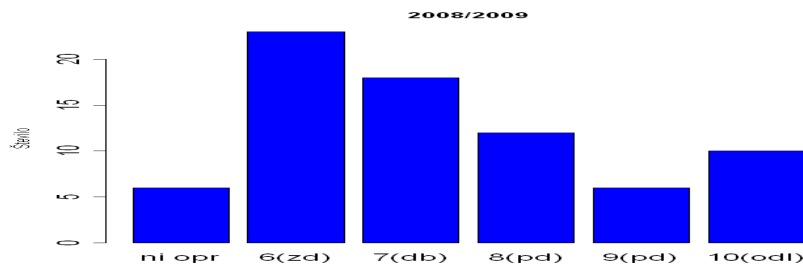
Imenske

Spol, Kajenje (Da/Ne),
Barva las



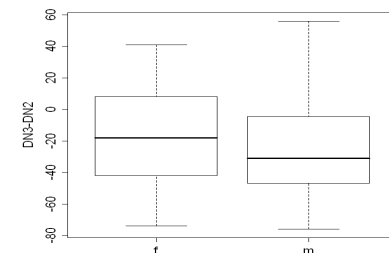
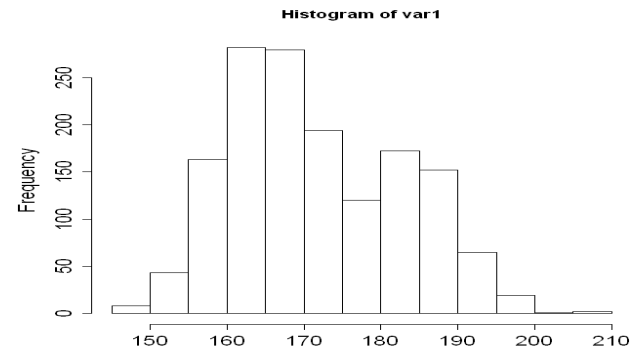
Urejenostne

Stadij bolezni, Kajenje
(Ne/Občasno/Redno),
Ocena, Izobrazba



Razmične

Temperatura v C
ali F



Razmernostne

Temperatura v
Kelvinih, teža

Kako lahko predstavimo podatke?

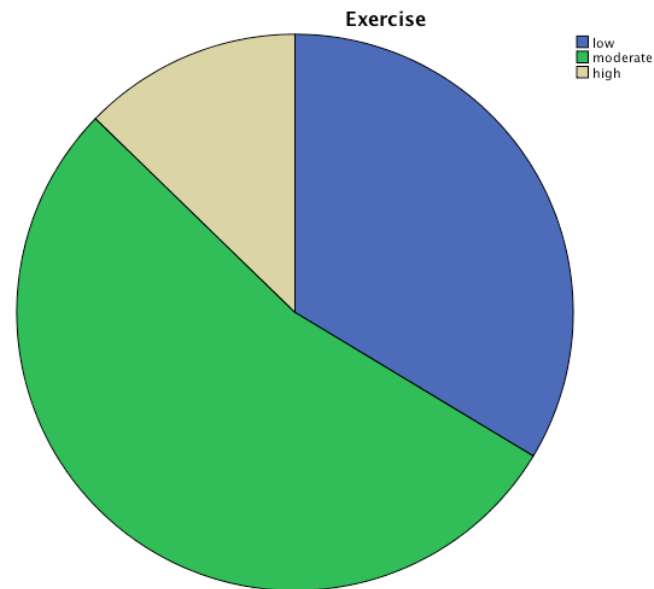
- Izbira primerne metode je odvisna od **vrste podatkov**
- Z **opisno statistiko** povzamemo lastnosti podatkov
 - **Mere središčnosti** (centralne tendence)
 - **Mere razpršenosti** (variabilnosti)
- Grafične predstavitve podatkov
- Povezanost med dvema ali več spremenljivkami

Rekreacija?

Frekvenčna tabela (frequency table)

Kolač, strukturni krog, pita
(pie chart)

	Frekvenca	Relativna frekvenca	Kumulativna relativna frekvenca
malo	37	0,33	0,33
srednje	59	0,54	0,87
veliko	14	0,13	1
Skupaj	110	1	



Možne vrednosti

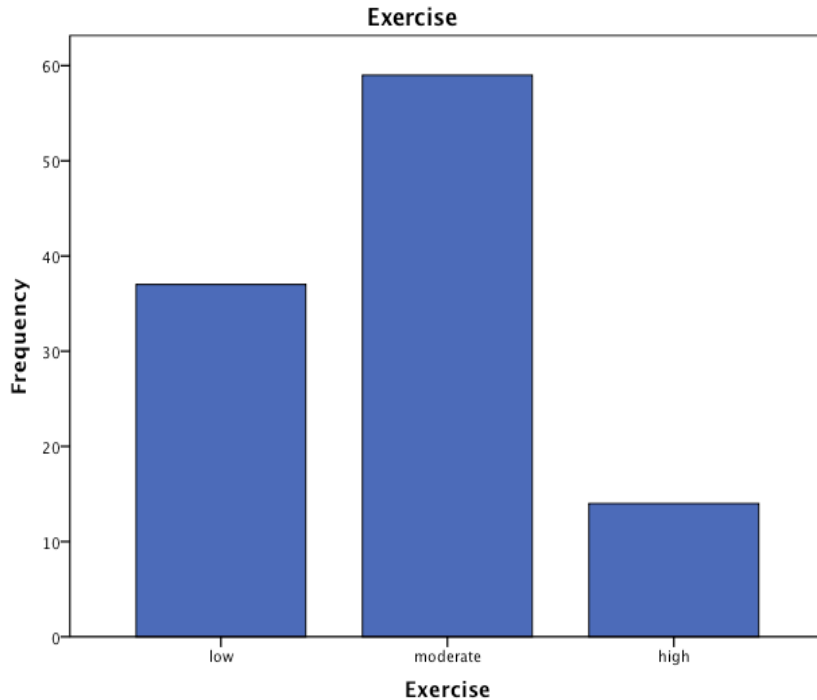
Pogostost

Relativna pogostost

$$0,33 + 0,54 = 0,87$$

$$\frac{14}{110} = 0,13$$

Stolpični diagram (bar plot)



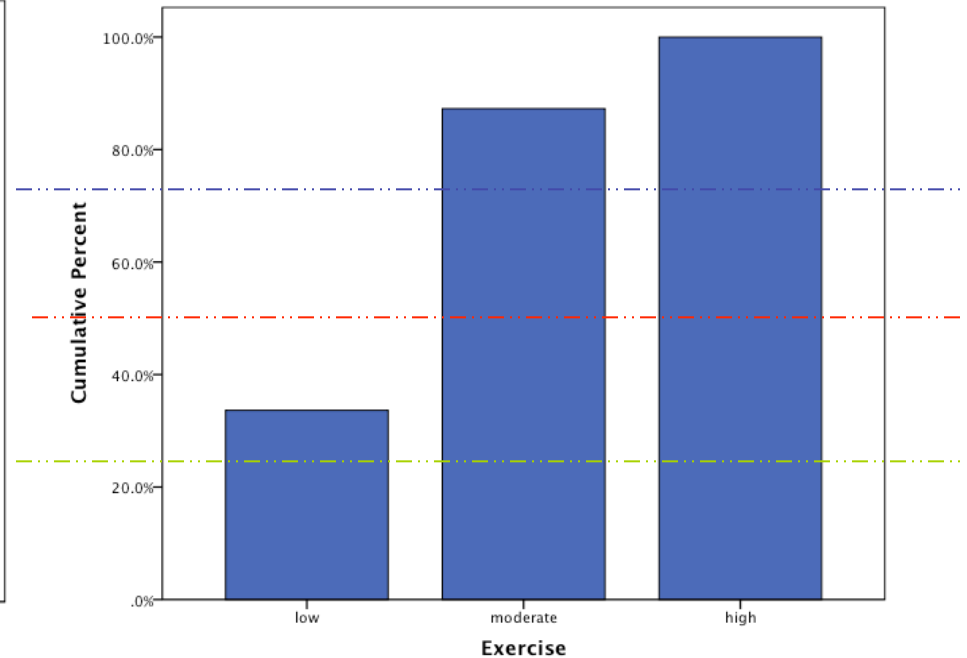
Katera je najpogostejša vrednost?

MODUS (mode)

Ali lahko izračunamo povprečno oceno?

POVPREČJE

Empirična porazdelitvena funkcija



Katera je “srednja” vrednost?

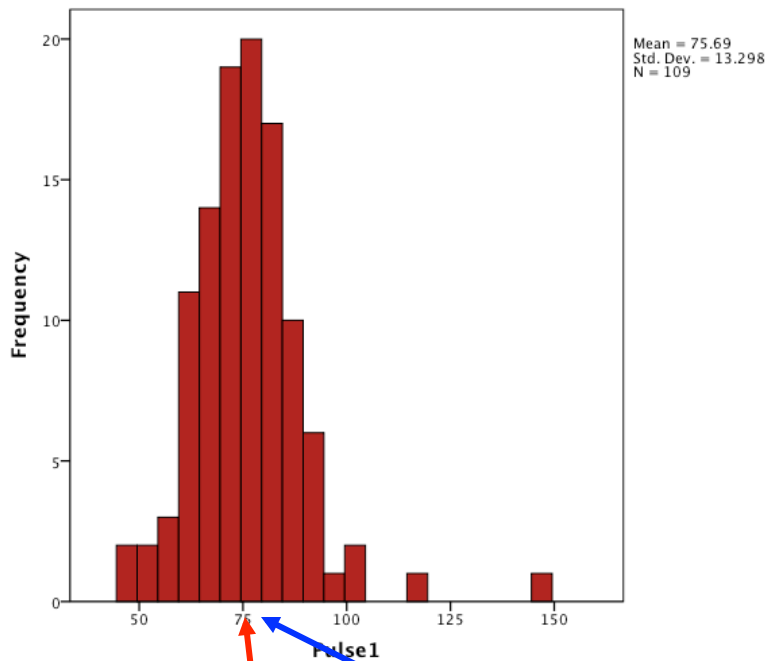
25. percentil (1. kvartil)?

MEDIANA (ali 50. percentil ali 2.kvartil)

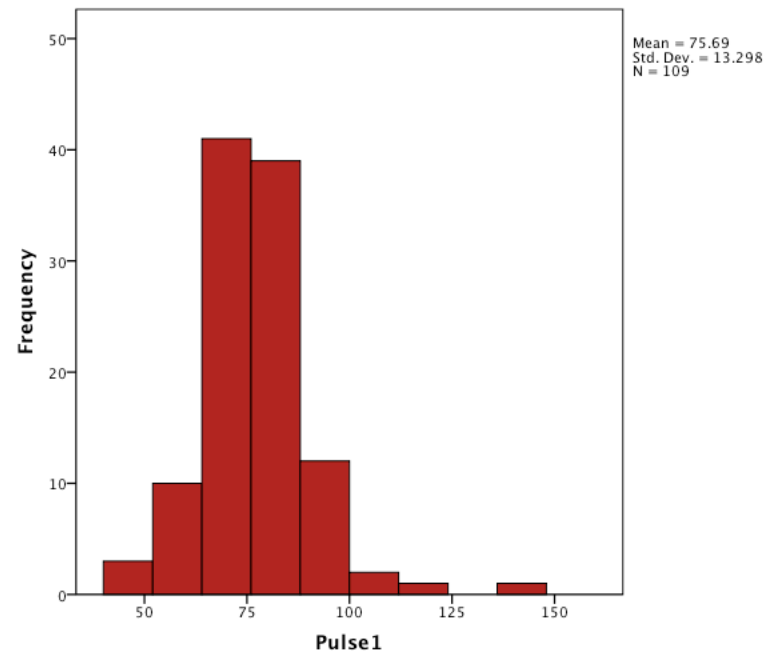
75. percentil (3. kvartil)?

Poglejmo še pulz v mirovanju

Histogram



Druga izbira velikosti intervala



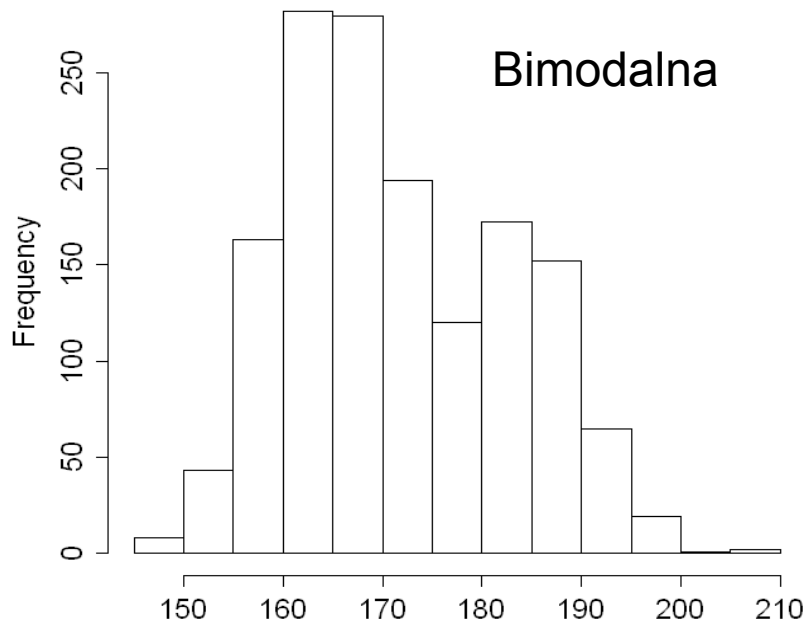
Kaj lahko rečemo o:

- Razpršenosti
- asimetričnosti?

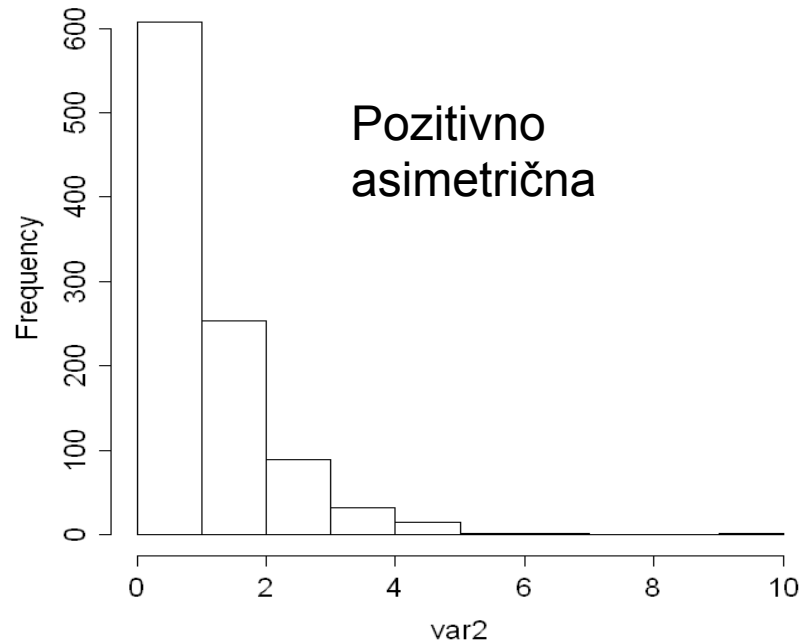
Koliko študentov ima pulz več kot 100?

Kolikšno je **povprečje**? Kolikšna je **mediana**?

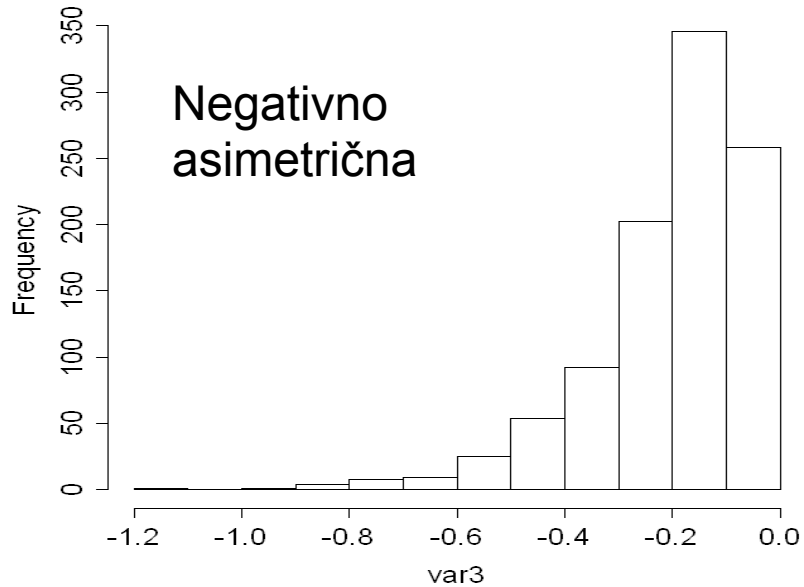
Histogram of var1



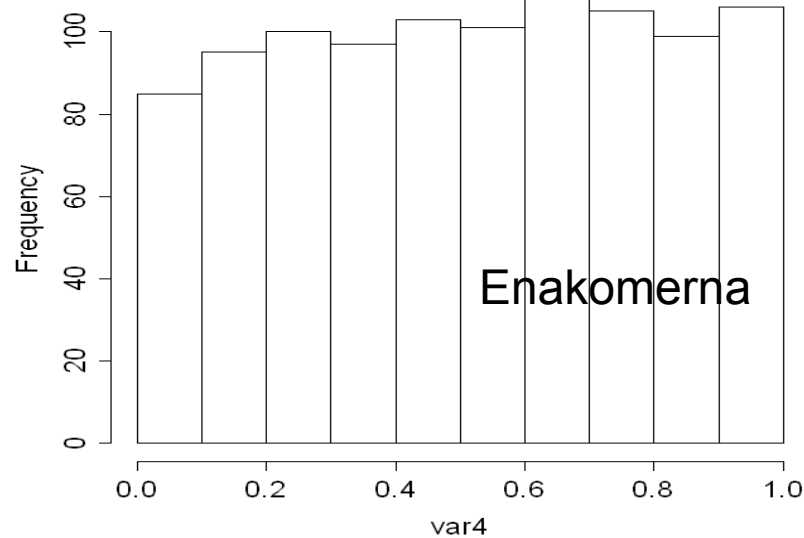
Histogram of var2



Histogram of var3

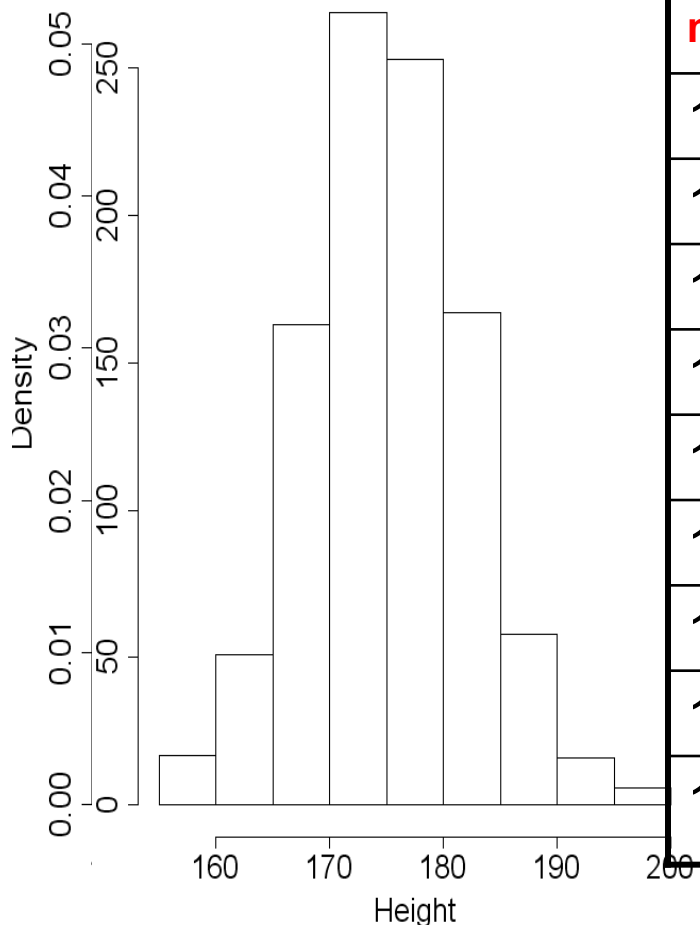


Histogram of var4



Kako pridemo do histograma?

Distribution of Height



vključena izključena

n=1000, velikost vzorca

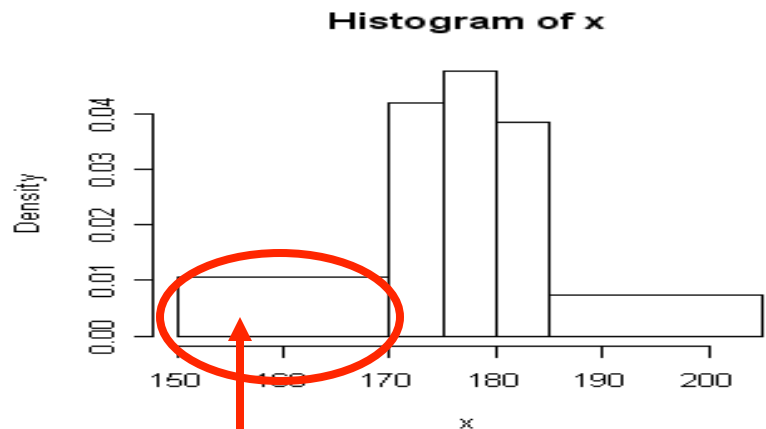
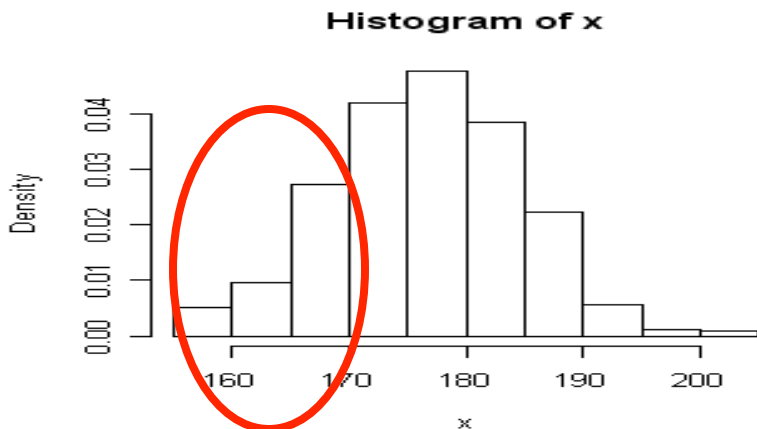
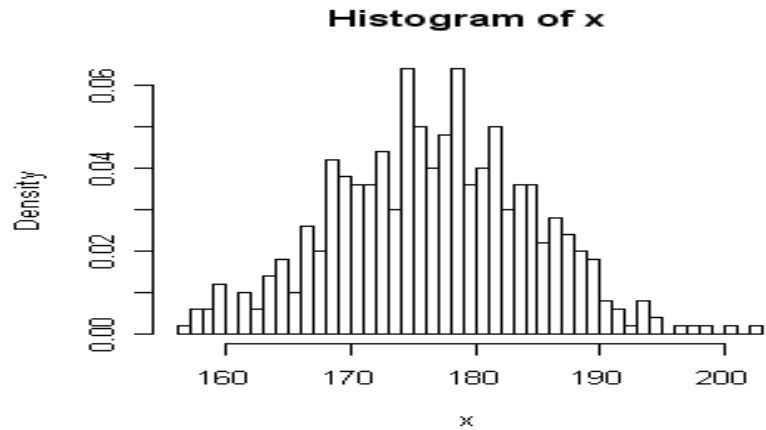
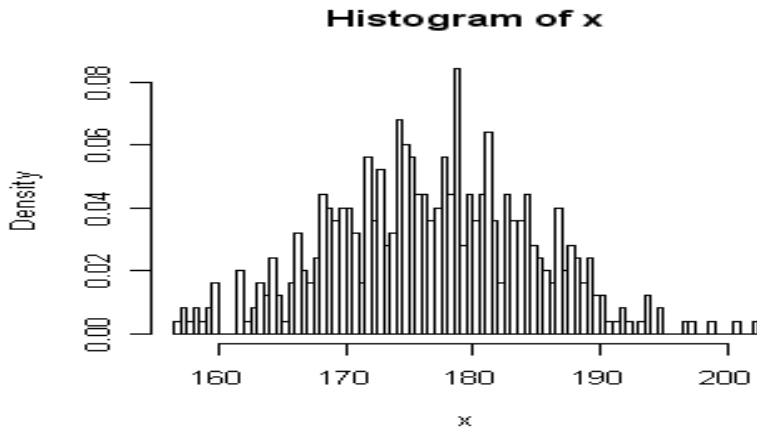
Spodnja meja	Zgornja meja	Frekvenca razreda	Relativna frekvenca	Gostota
155	160	17	.0017	0.003
160	165	51	0.051	0.010
165	170	163	0.163	0.033
170	175	269	0.269	0.054
175	180	253	0.253	0.051
180	185	167	0.167	0.033
185	190	58	0.058	0.012
190	195	16	0.016	0.003
195	200	6	0.006	0.001

Histogram

Gostota (Density)=Relativna frekvenca/širina razreda

Isti podatki, različni prikazi

Širina intervalov vpliva na videz



Katera predstavitev je najboljša?

Kako se razlikujeta Hist3 in Hist4?

Ploščina (in ne višina!) je sorazmerna relativni frekvenci

Mere središčnosti (measures of central tendency)

- **Povprečje** (mean, arithmetic mean)

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

- **Mediana** (median)

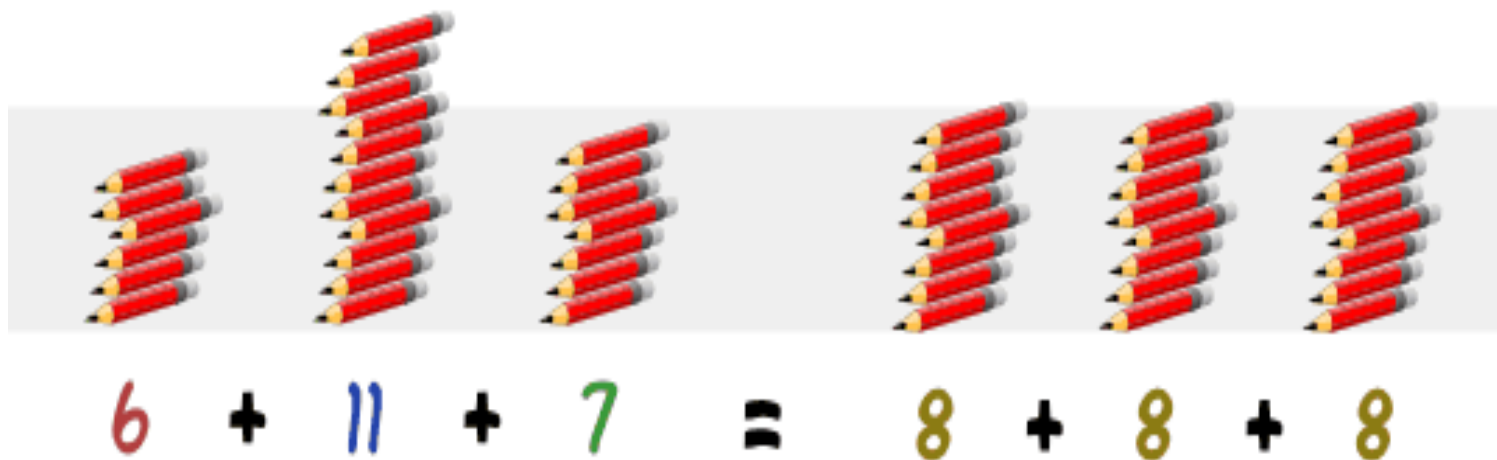
- Srednja vrednost porazdelitve (polovica od opazovanj ima vrednost, ki je manjša ali enaka mediani, polovica pa vrednost, ki je večja ali enaka mediani).

- *50. percentil, 2. kvartil*

- **Modus** (mode)

- Najpogostejša vrednost

Zakaj potrebujemo mere razpršenosti?



Povprečje=8

SD=2.65

razpon=5

Povprečje=8

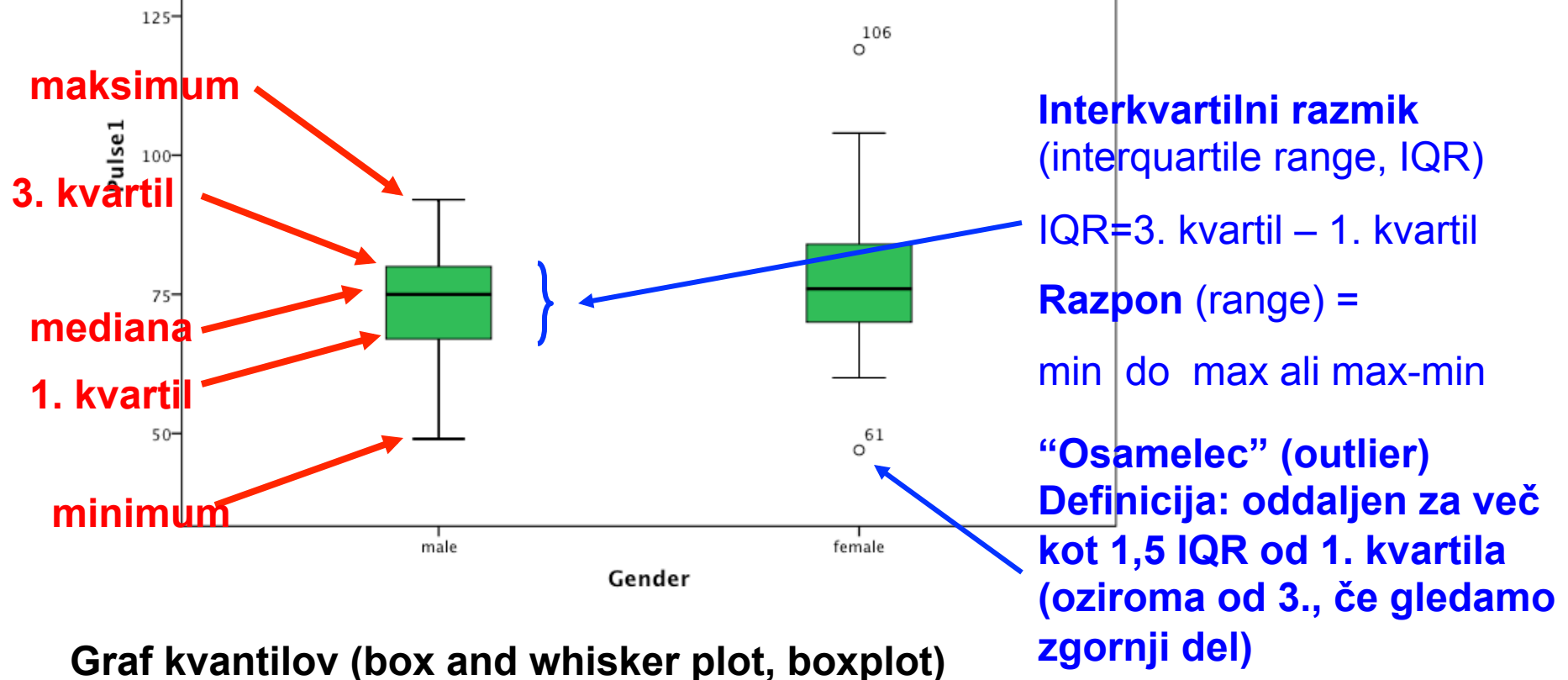
SD=0

razpon=0

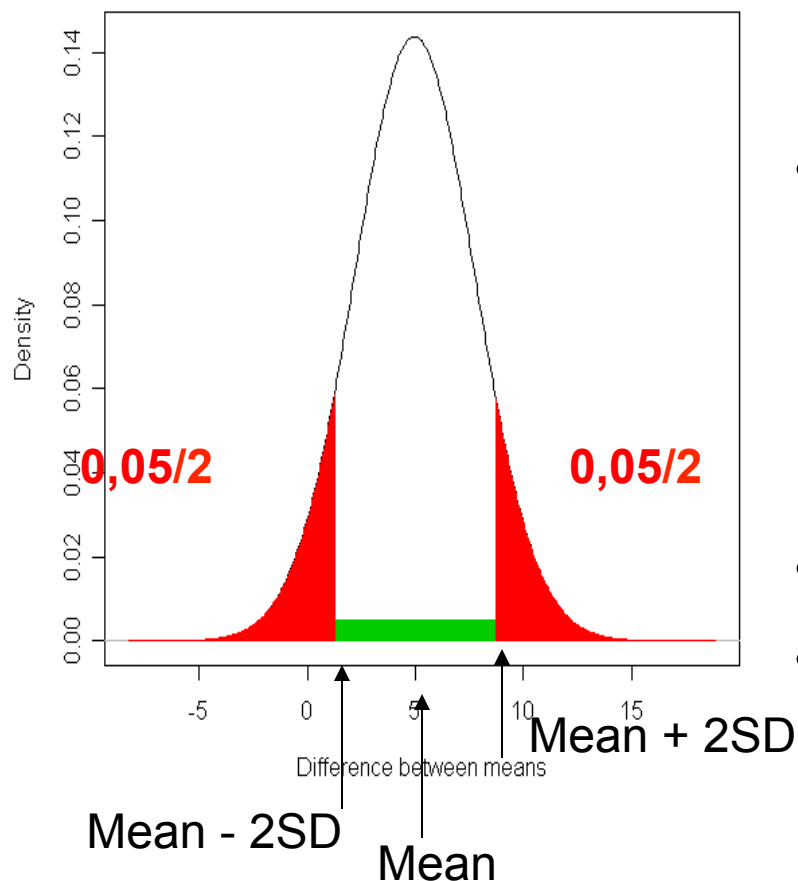
Ali imajo študentke višji pulz kot študenti?

Povprečje:	74,1 (m)	77,5 (f)
Standardni odklon:	13,8 (m)	12,6 (f)

Ali je ta razlika dovolj velika, da lahko sklepamo, da so pulzi deklet značilno višji od pulzov fantov?



Kdaj lahko interpretiramo standardni odklon?



- Če je porazdelitev podatkov *zmerno simetrična*, večina opazovanj (95%) bo znotraj 2 standardnih odklonov od povprečja
- ... in 68% 1 SD od povprečja
- ... in 99.7% 3 SD od povprečja

Mere razpršenosti (measures of dispersion)

- **Varianca**

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

- (**standardni odklon** (SD, standard deviation) = $\sqrt{\text{variance}}$)

- **Razpon** (range)

- Najvišja vrednost – najnižja vrednost

- **Interkvartilni razmik** (interquartile range)

- 3. kvartil – 1. kvartil

- 1. kvartil = 25. percentil

V katerih enotah je izražena varianca?

Mere središčnosti: ali jih lahko izračunamo za vsako spremenljivko?

- Število mladičev: 3, 1, 7, 2, 2
 - Povprečje = ?
 - Mediana = ?
 - Modus = ?
- Stadij raka dojke: I, IV, III, II, I
 - Povprečje = ?
 - Mediana = ?
 - Modus = ?
- Spol: ženski, moški, moški, moški, ženski
 - Povprečje = ?
 - Mediana = ?
 - Modus = ?

Mere razpršenosti: ali jih lahko izračunamo za vsako spremenljivko?

- Število mladičev: 3, 1, 7, 2, 2
 - Varianca = ?
 - Interkvartilni razmik = ?
 - Razpon = ?
- Stadij raka dojke: I, IV, III, II, I
 - Varianca = ?
 - Interkvartilni razmik = ?
 - Razpon = ?
- Spol: ženski, moški, moški, moški, ženski
 - Varianca = ?
 - Interkvartilni razmik = ?
 - Razpon = ?

Spremenljivke

Opisne

Številске

Imenske

Urejenostne

Razmične

Razmernostne

Modus



Mediana



Povprečje



Razpon



IQR



Varianca/ standardni odklon



$X + \text{ali} - Y$



X / Y



Kaj je opisna statistika (descriptive statistics)?

“je skupina statističnih metod, ki se ukvarjajo s **povzemanjem pridobljenih podatkov**. Te metode iščejo opisne (meta) podatke o populaciji in njenih sestavnih delih, da bi ustvarile pregledni opis.“ (Wikipedia, Oktober2009)

KAKO?

- grafikoni
- tabele
- “statistični povzetki“