

Statistika



Statistična obdelava podatkov

FRI – 2008

Aleksandar Jurišić



Pogled od zunaj



Be good & you will be lonesome.
Mark Twain

Števila so me pogosto begala, še posebej, če sem se imel pred seboj neko njihovo razvrstitev, tako da je tem primeru obveljala misel, ki so jo pripisali Diaraeliju, z vso pravico in močjo: **"Obstajajo tri vrste laži: laži, preklete laži in statistika."**

Autobiography of Mark Twain

Figures often beguile me, particularly when I have the arranging of them myself; in which case the remark attributed to Disraeli would often apply with justice and force: **"There are three kinds of lies: lies, damned lies and statistics."**

- *Autobiography of Mark Twain*



Načrt

- **Opisna statistika**
 - ena spremenljivka
 - Mere centralne tendence
 - Mere razpršenosti
 - Mere oblike
 - dve spremenljivki
 - Mere asociacije
- **Inferenčna (analitična) statistika**
 - točkovno in intervalno ocenjevanje
 - ena- in dva- vzorčno testiranje hipotez
 - kontingenčne tabele
 - regresija

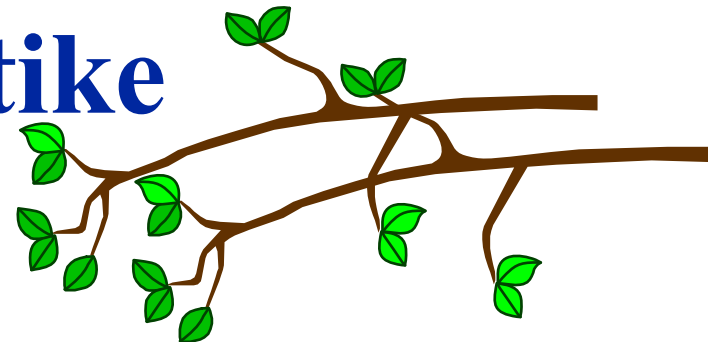


Statistika

**preučuje podatke, jih
zbira,
klasificira,
povzema,
organizira,
analizira in
interpretira.**



Dve glavni veji statistike



Opisna statistika se ukvarja z
organiziranjem, povzemanjem in
opisovanjem zbirk podatkov
(reduciranje podatkov na povzetke)

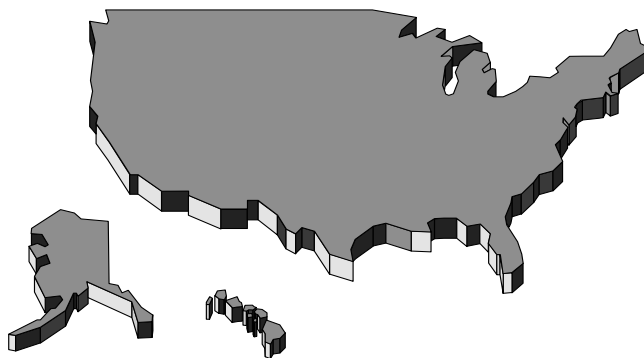
Analitična statistika jemlje vzorce podatkov
in na osnovi njih naredi zaključke
(inferenčnost) o populaciji
(ekstrapolacija).



Tipi podatkovnih množic

- **Populacija**

- vsi objekti, ki jih opazujemo



- Primer: vsi registrirani glasovalci

- **Vzorec**

- podmnožica populacije



- Primer: 100 registriranih glasovalcev

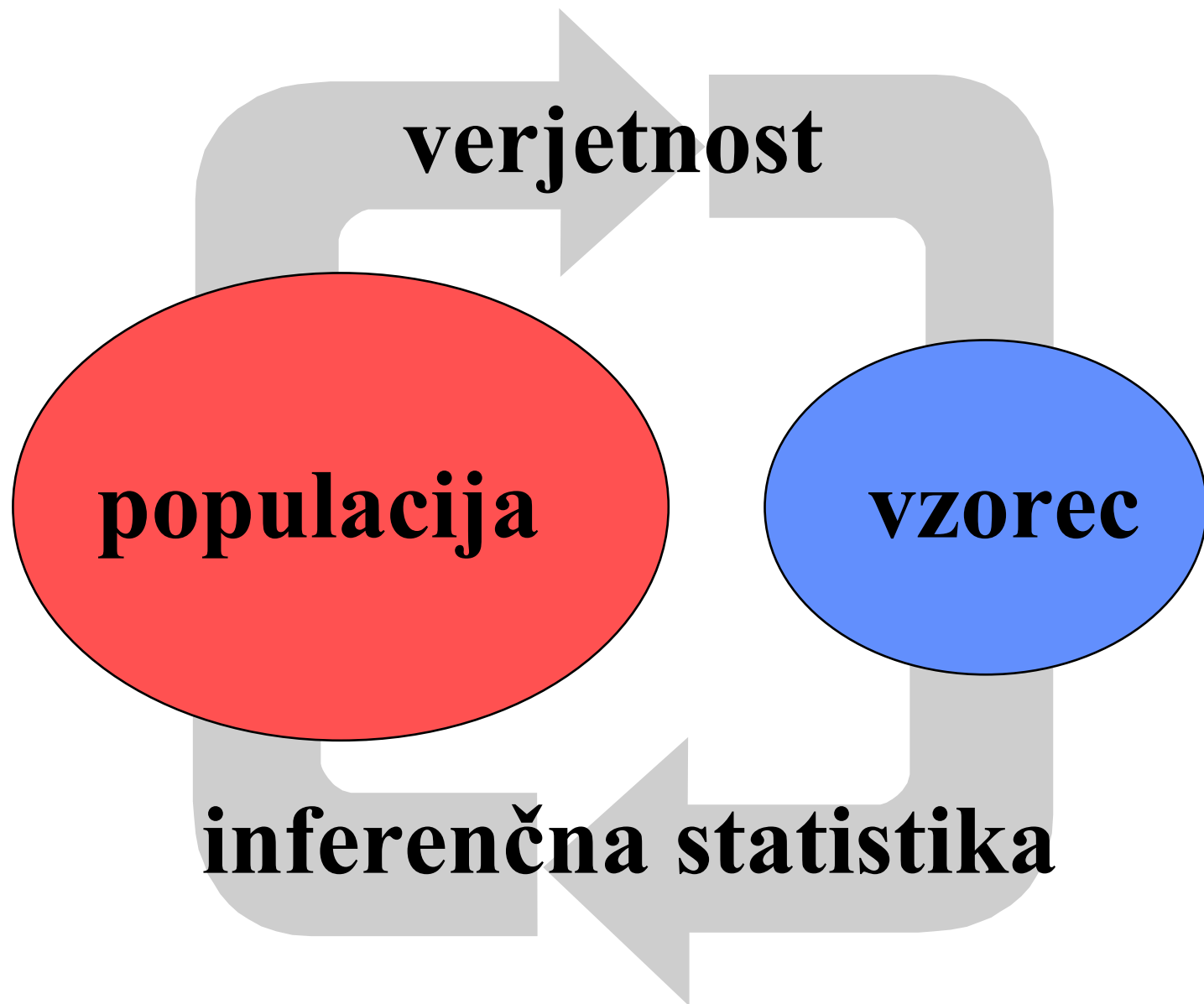


Populacija je podatkovna množica, ki ji je namenjena naša pozornost.

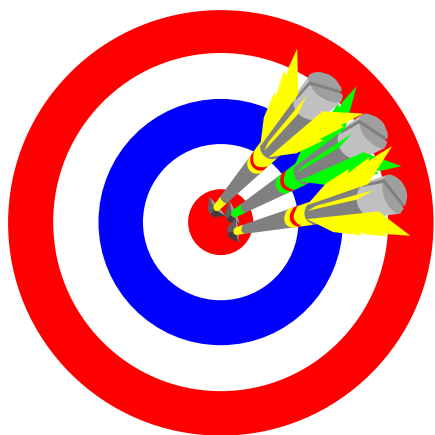


Vzorec je podmnožica podatkov,
ki so izbrani iz
Populacije
(po velikosti
bistveno manjši
od populacije).





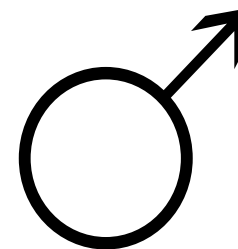
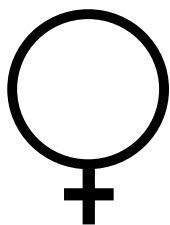
Tipi podatkov



- **kvantitativni**
(numerični)
predstavljajo kvantiteto
ali količino nečesa.



- **kvalitativni** (kategorije)
ni kvantitativnih interpretacij.



Kvantitativni (numerični)

- **interval**
 - poljubna ničla
 - enaki intervali predstavljajo enake količine
- **razmerje**
 - smiselna točka nič
 - operacije seštevanje, odštevanje, množenje in deljenje so smiselne



Kvalitativni (kategorični)

- **nominalni**
 - kategorije brez odgovarjajočega vrstnega reda – urejenosti
- **ordinalni/številski**
 - kategorije z urejenostjo



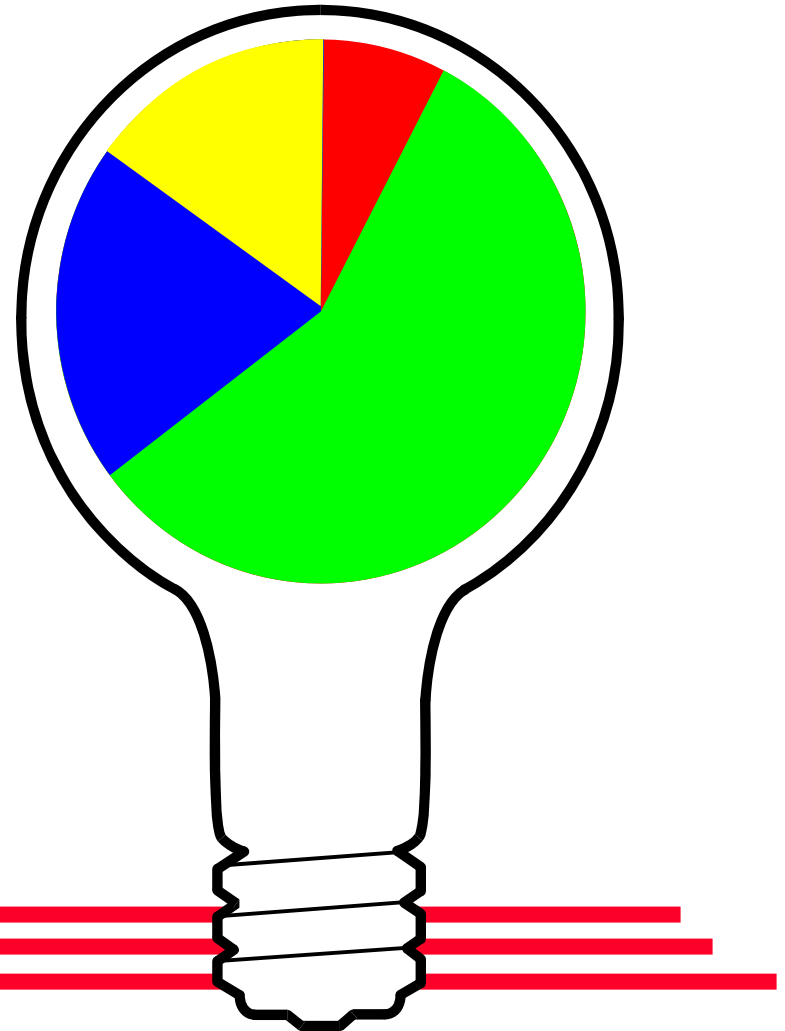
Oddelek sistemskih inženirjev

kategorija	frekvenca	relativna frekvenca
vrsta zaposlenih	število zaposlenih	delež
učitelji	16	0,8421
skupne službe	3	0,1579
skupaj	19	1,0000



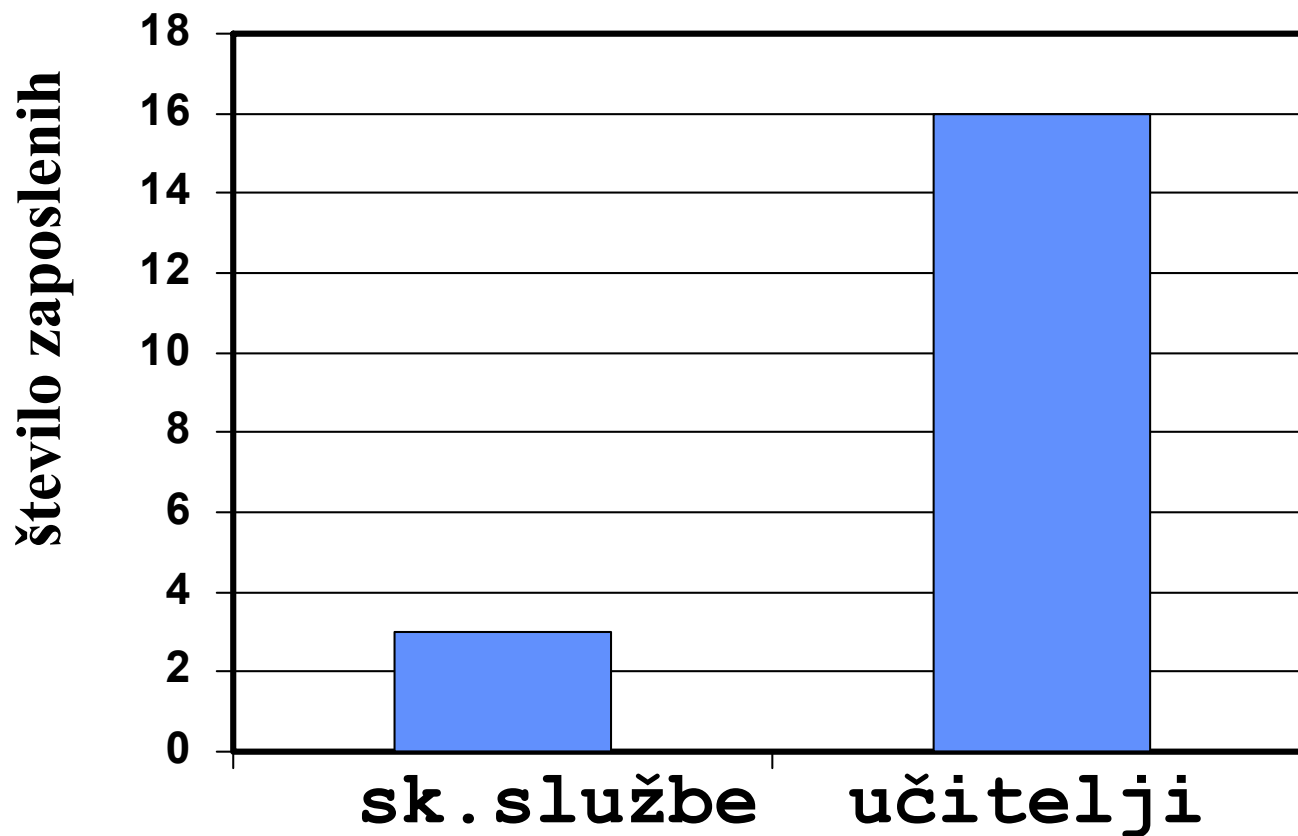
Grafična predstavitev kvalitativnih podatkov

- stolpčni graf,
poligonski diagram
- strukturni krog
pogača, kolač



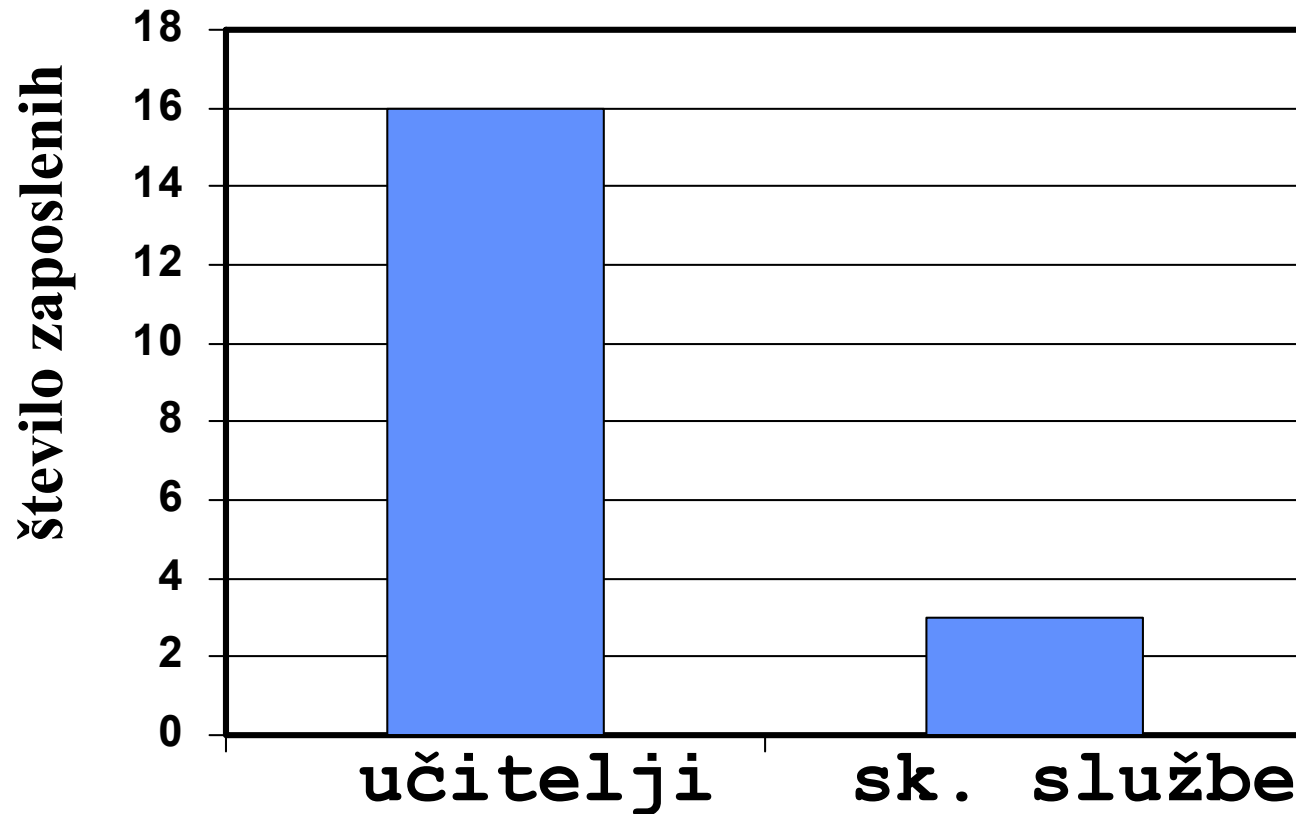
Stolpčni graf

oddelek sistemskih inženirjev



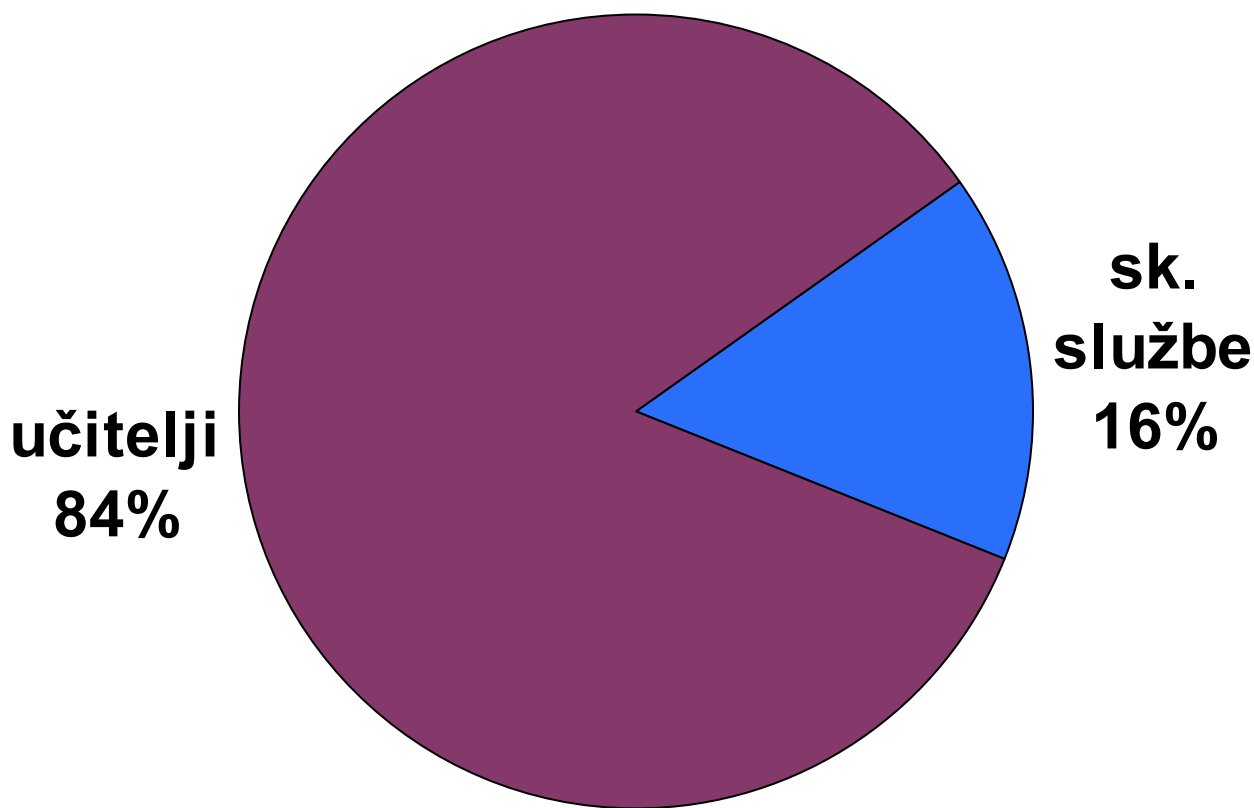
Pareto diagram (po italijanskem ekonomistu)

oddelek sistemskih inženirjev



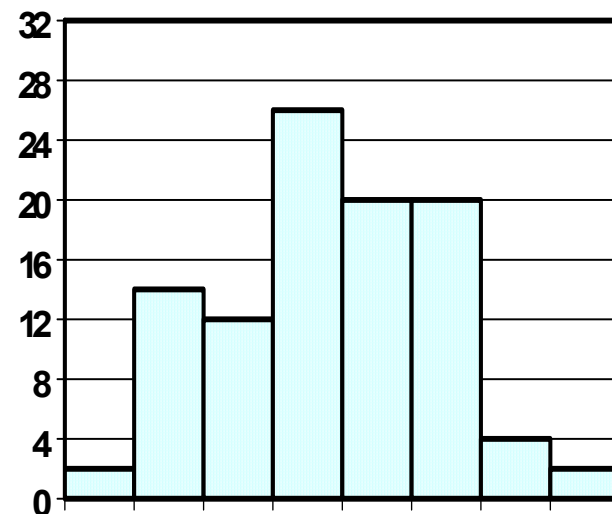
Strukturni krog (pogača, kolač)

oddelek sistemskih inženirjev

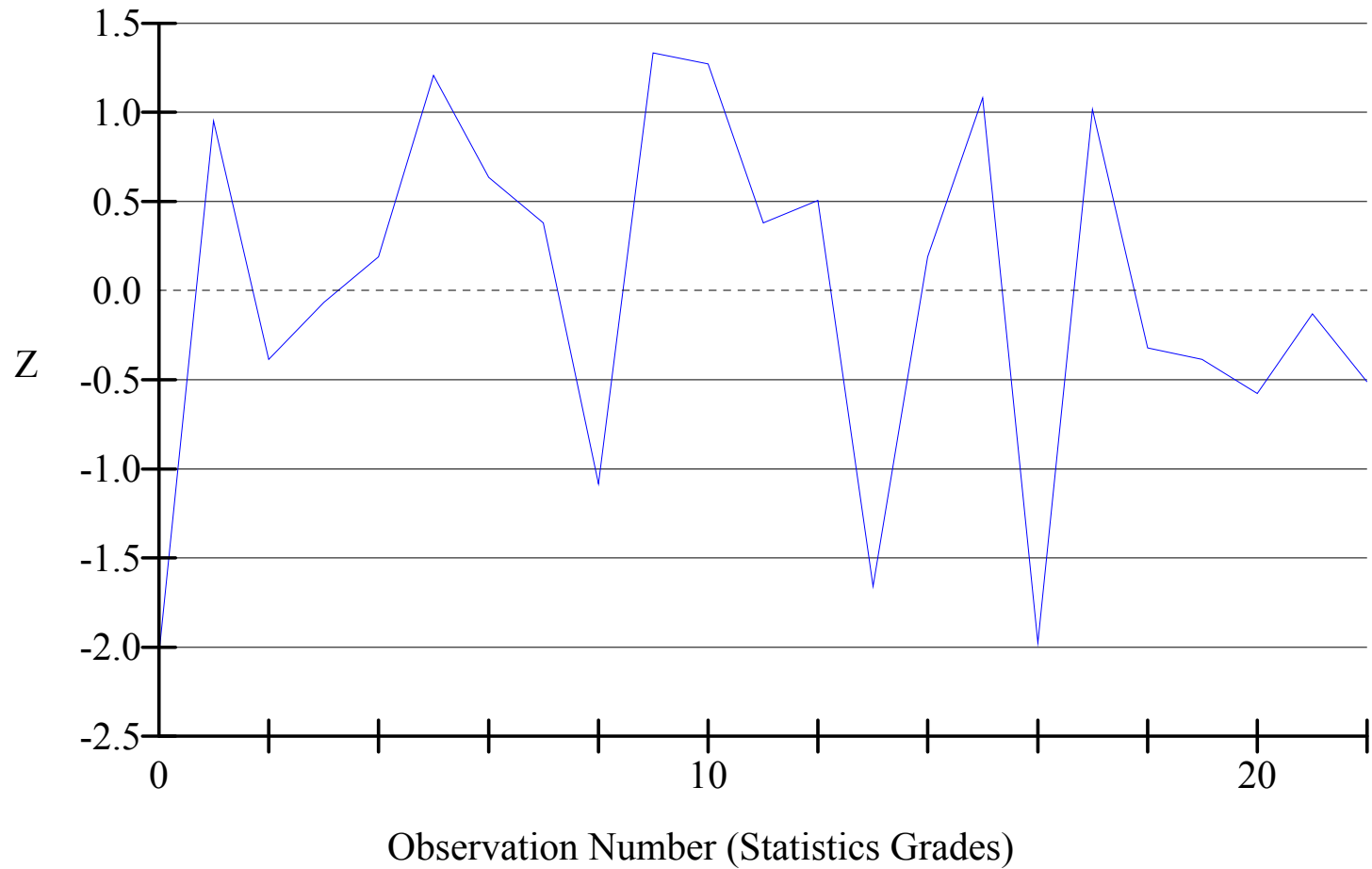


Grafična predstavitev kvantitativnih podatkov

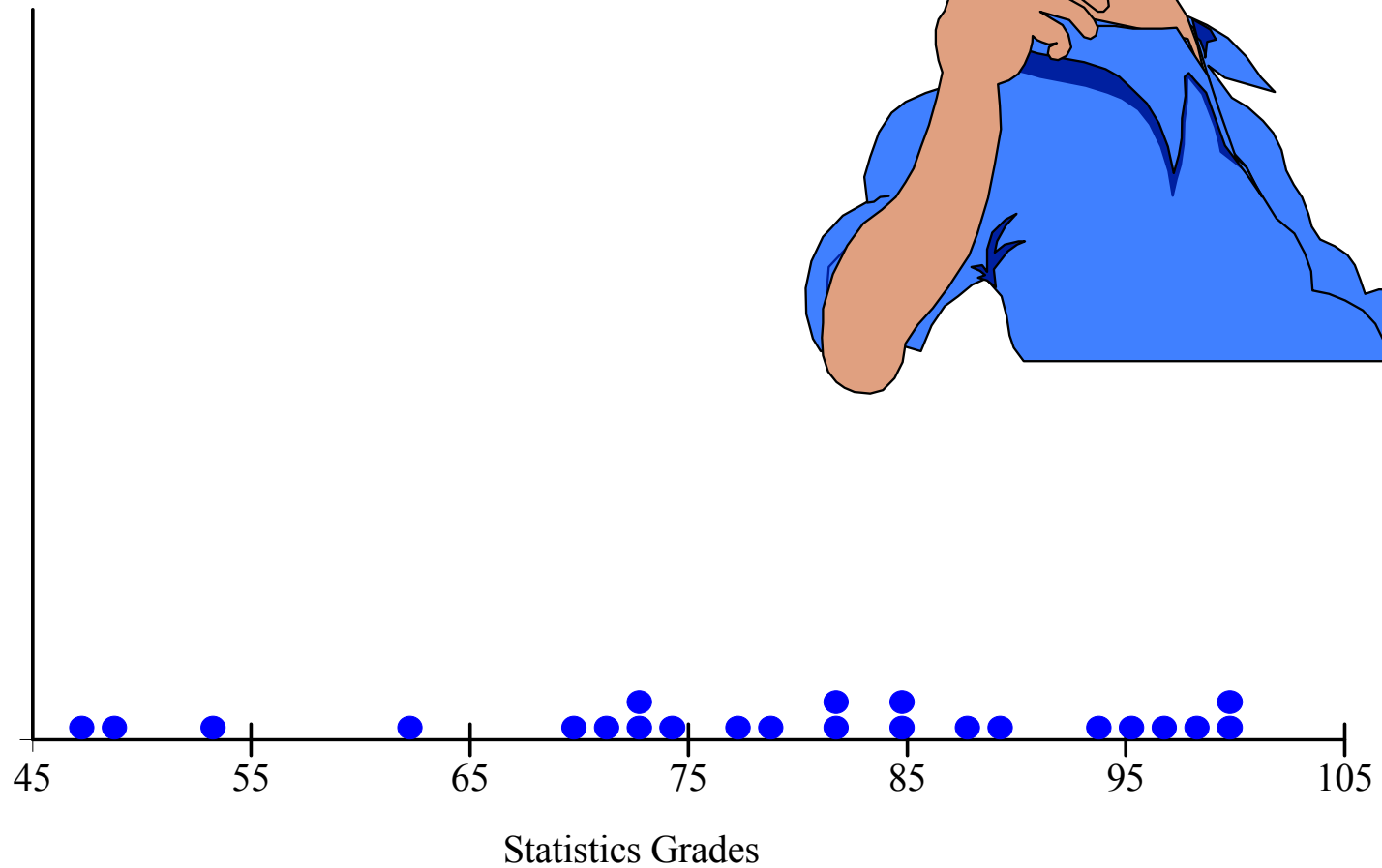
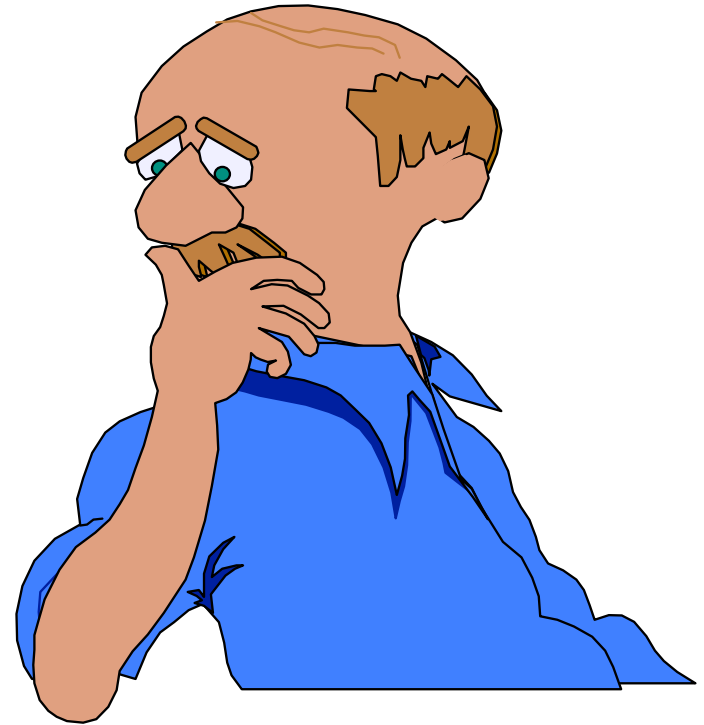
- runs plot (X, Y plot)
- zaporedje (dot plot)
- steblo-list predstavitev
(angl. stem-and-leaf)
- histogrami
- škatla z brki (box plot)



Runs Chart

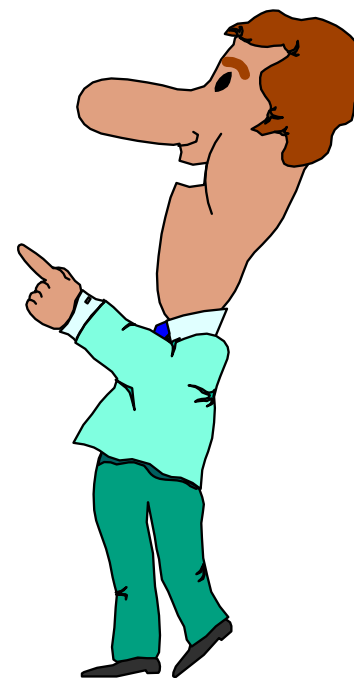


Dot Plot



Urejeno zaporedje/ranžirana vrsta

Urejeno zaporedje je zapis
podatkov v vrsto po
njihovi numerični
velikosti (ustreznemu
mestu pravimo **rang**).



Primer zaporedja podatkov (nal. 2.48, str.64)

a. Konstruiraj urejeno zaporedje.	88	103	113	122	132
	92	108	114	124	133
	95	109	116	124	133
	97	109	116	124	135
b. Nariši steblo-list diagram.	97	111	117	128	136
	97	111	118	128	138
	98	112	119	128	138
	98	112	120	131	142
c. Naredi histogram.	100	112	120	131	146
	100	113	122	131	150



Koraki za konstrukcijo steblo-list predstavitve

1. Razdeli vsako opazovanje-podatke na dva dela, **stebila** (angl. stem) in **listi** (angl. leaf).
2. Naštej stebila po vrsti v stolpec, tako da začneš pri najmanjšem in končaš pri največjem.



Koraki za konstrukcijo steblo-list predstavitve

- 3. Upoštevaj vse podatke in postavi liste za vsak dogodek/meritev v ustrezno vrstico/steblo.**
- 4. Naštej frekvence za vsako steblo.**



Steblo-list diagram

steblo/listi

frekvenca relativna
frekvenca

steblo	listi	frekvenca	relativna frekvenca
08	8	1	2%
09	2 5 7 7 7 8 8	7	14%
10	0 0 3 8 9 9	6	12%
11	1 1 2 2 2 3 3 4 6 6 7 8 9	13	26%
12	0 0 2 2 4 4 4 8 8 8	10	20%
13	1 1 1 2 3 3 5 6 8 8	10	20%
14	2 6	2	4%
15	0	1	2%
		50	100%



Histogrami

- **kako zgradimo histogram**
- **število razredov**
- **frekvenca**
- **procenti**



Kako zgradimo histogram

1. Izračunaj **razpon** podatkov.
2. Razdeli razpon na 5 do 20 **razredov** enake širine.
3. Za vsak razred preštej število vzorcev, ki spadajo v ta razred.
To število imenujemo **frekvenca razred**.
4. Izračunaj vse **relativne frekvence razredov**.



Pravilo za določanje števila razredov v histogramu

število vzorcev
v množici podatkov

število
razredov

manj kot 25

5 ali 6

25 – 50

7 – 14

več kot 50

15 – 20

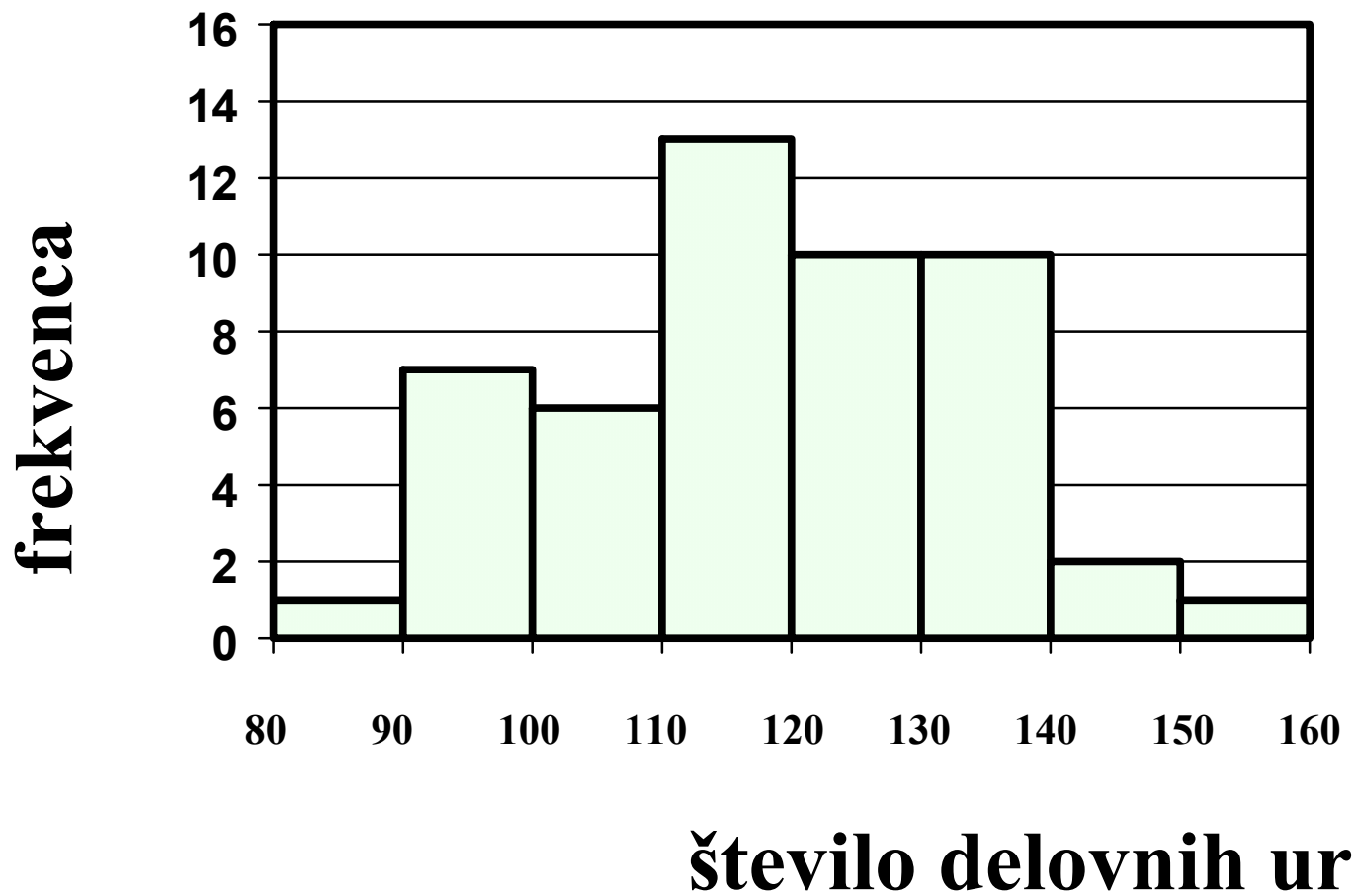


Frekvenčna porazdelitev

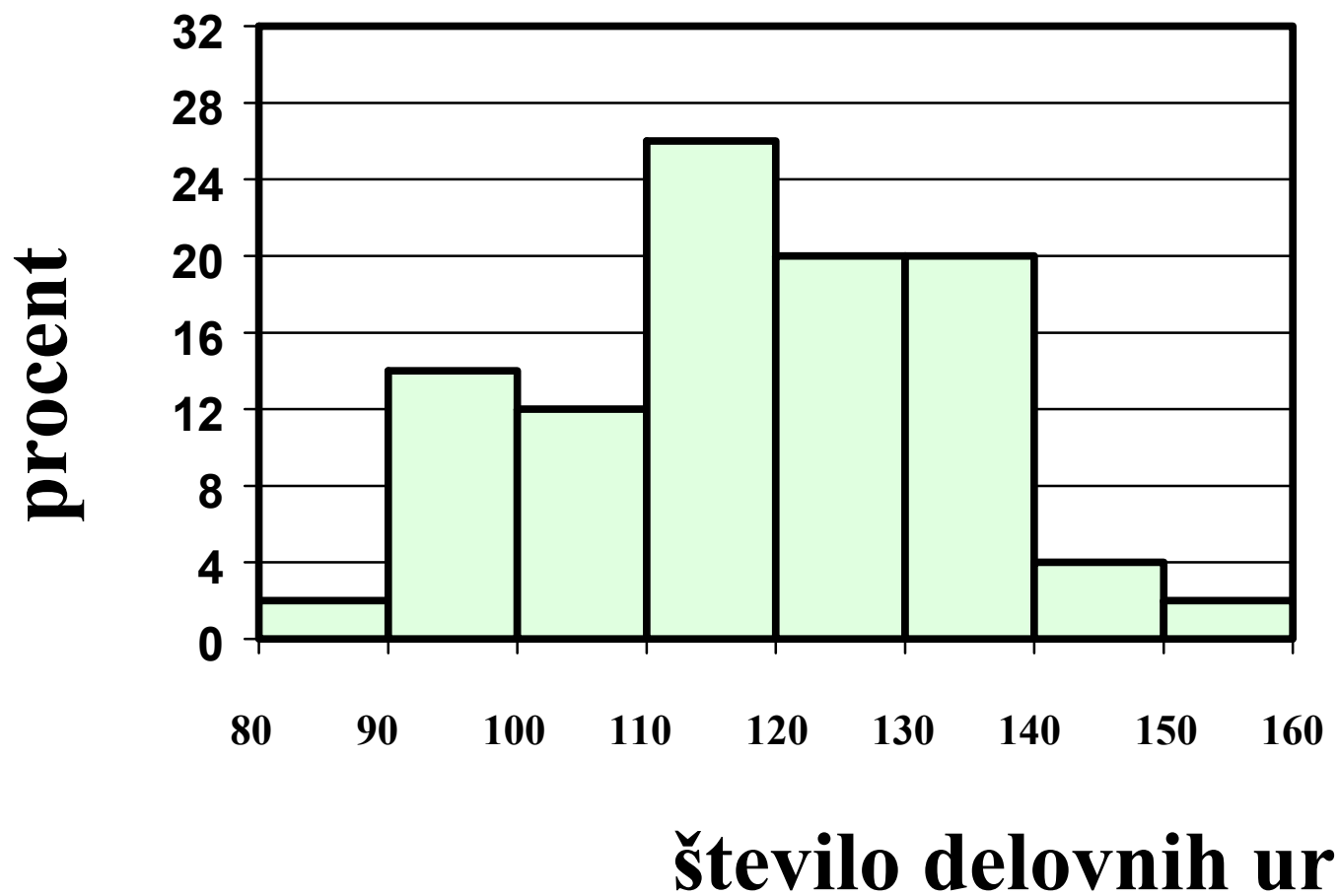
razred	interval razreda	frekvenca	relativna frekvenca
1	80 – 90	1	2%
2	90 – 100	7	14%
3	100 – 110	6	12%
4	110 – 120	13	26%
5	120 – 130	10	20%
6	130 – 140	10	20%
7	140 – 150	2	4%
8	150 – 160	1	2%
		50	100%



Frekvenčni histogram



Procentni histogram





Mere za lokacijo in razpršenost

- **srednje vrednosti**
- **razpon (min./max)**
- **centili, kvartili**
- **varianca**
- **standardni odklon**
- **Z-vrednosti**



Modus (M_o)

**Modus množice podatkov
je tista vrednost, ki se
pojavi z največjo
frekvenco.**



Mediana (M_e)



Da bi prišli do mediane za neko množico podatkov, naredimo naslednje:

- 1. podatke uredimo po velikosti v naraščujočem vrstnem redu,**
- 2. če je število podatkov liho, potem je mediana podatek na sredini,**
- 3. če je število podatkov sodo, je mediana enaka povprečju dveh podatkov na sredini.**



Mediana

populacije: μ



vzorca: m



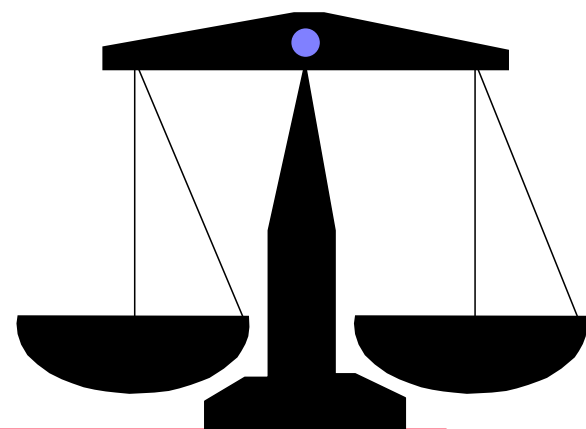
Povprečje

populacije:

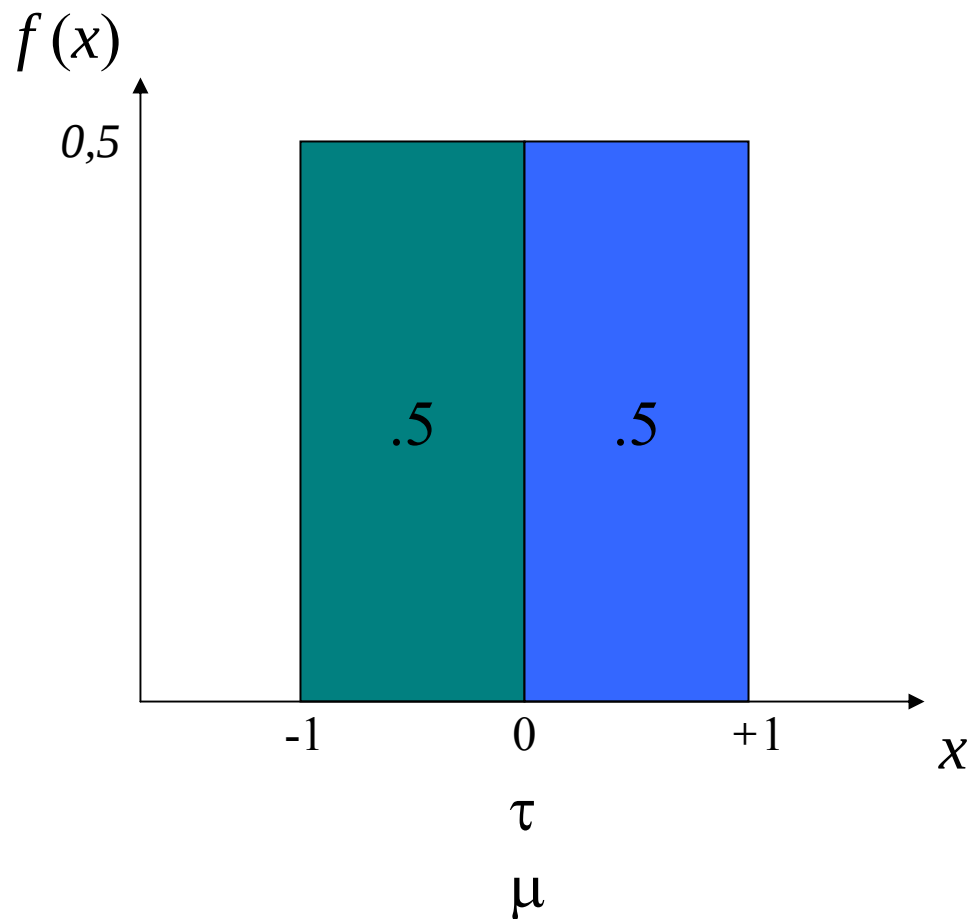
$$\mu = \frac{\sum_{i=1}^n y_i}{n}$$

vzorca:

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n}$$



Povprečje in mediana



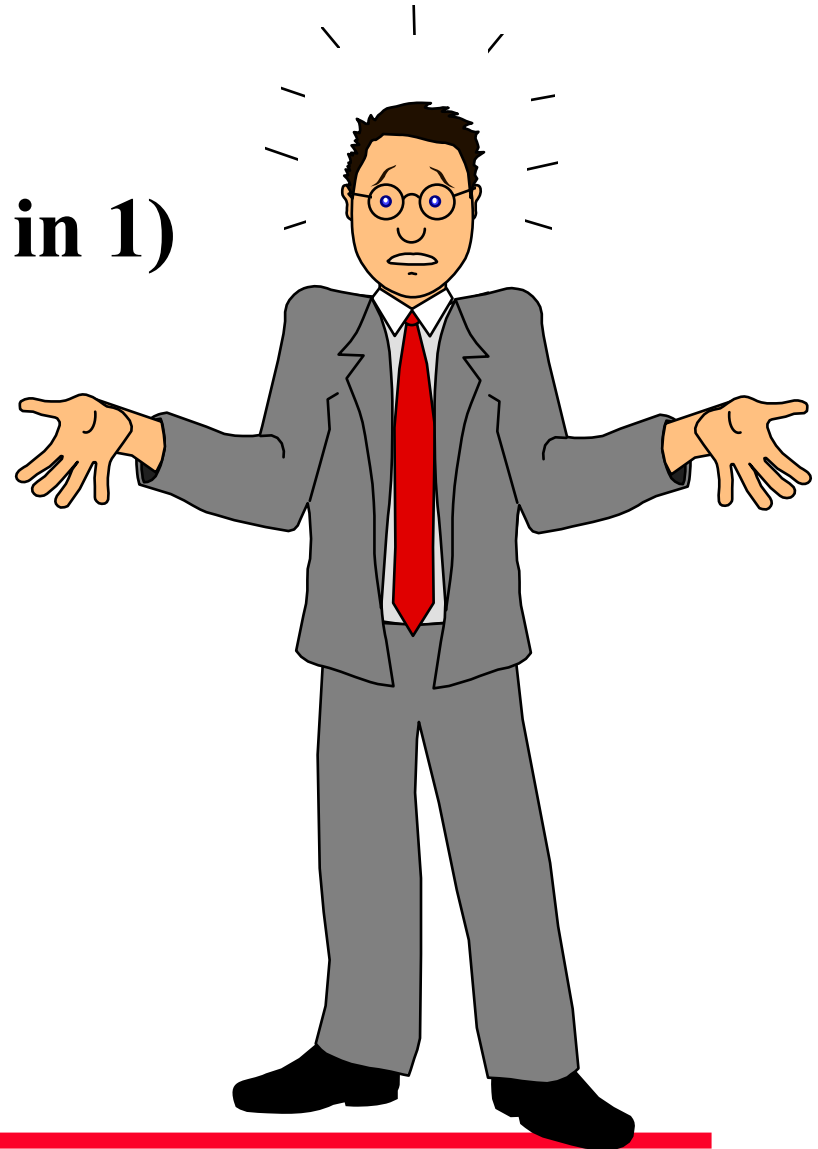
Razpon ali variacijski razmik

Razpon je razlika med največjo in najmanjšo meritvijo v množici podatkov.



Centili

$100p$ -ti centil (p je med 0 in 1)
je definiran kot število,
od katerega ima $100p$
procentov meritev
manjšo ali enako
numerično vrednost.



Določanje 100*p*-tega centila

Izračunaj vrednost $i = p(n+1)$

in jo zaokroži na najbližje celo število.

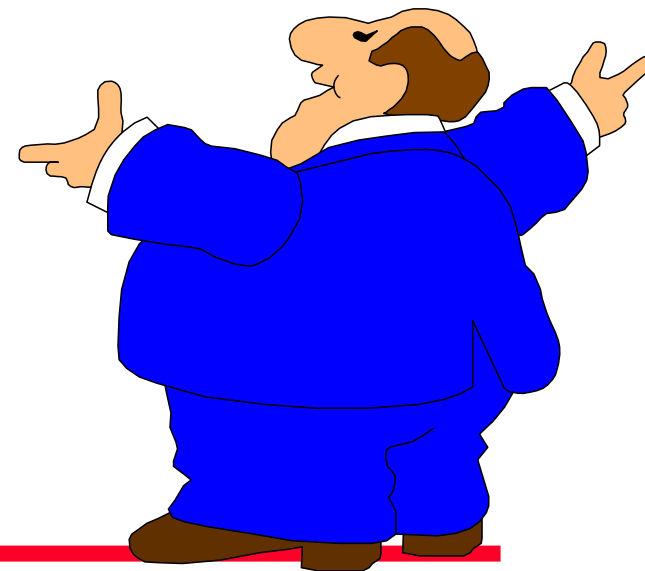
To število je enako i .

Izmerjena vrednost z i -tim rangom

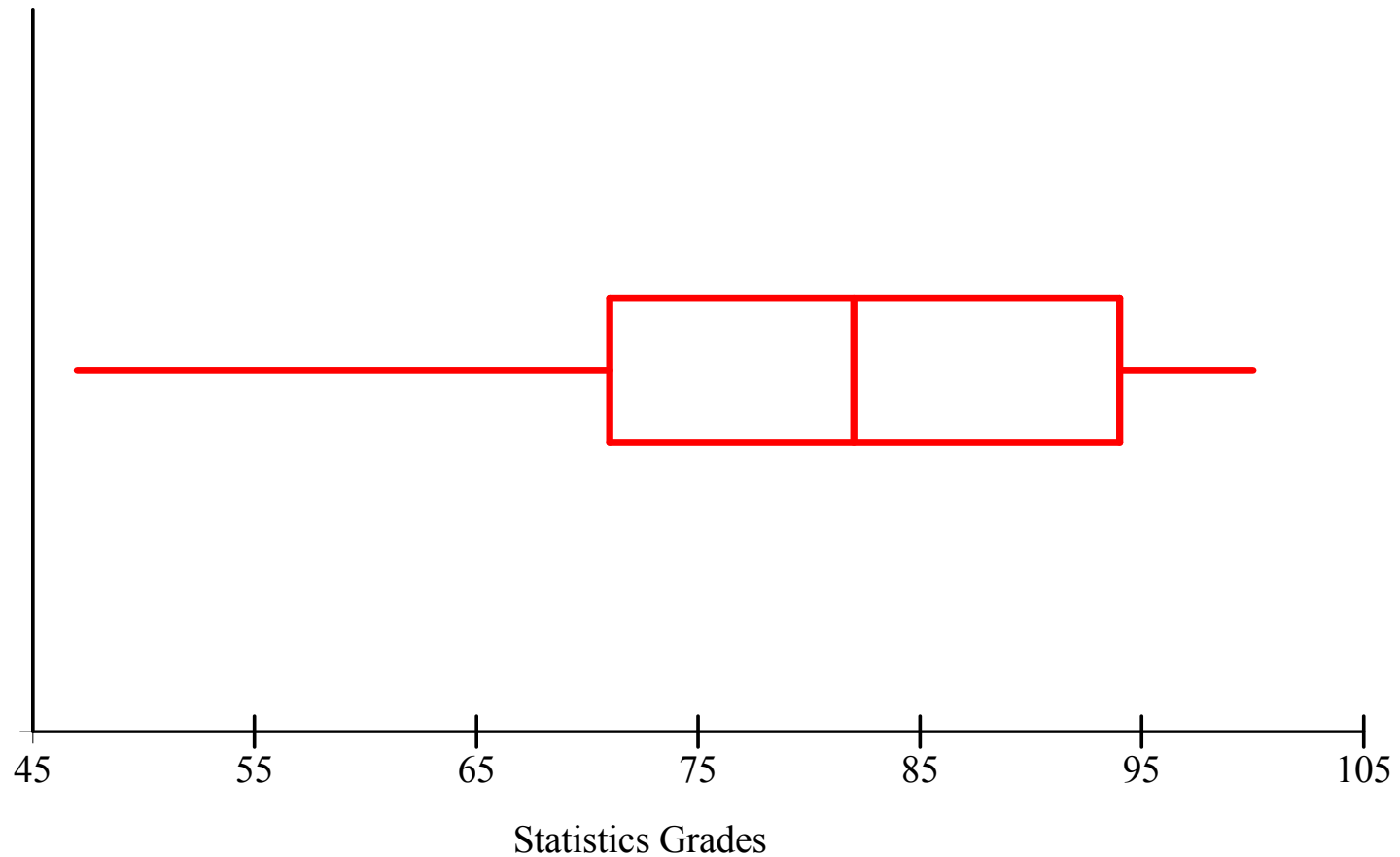
je **100*p*-ti centil**.



- 25. centil se imenuje tudi **1. kvartil**.
- 50. centil se imenuje **2. kvartil** ali **mediana**.
- 75. centil se imenuje tudi **3. kvartil**.



Škatla z brki (angl. box plot)



Mere razpršenosti

- **varianca**
 - kvadrat pričakovanega odklona (populacije)
 - vsota kvadratov odklonov deljena s stopnjo prostosti (vzorec)
- **standardni odklon (deviacija)**
 - pozitivni kvadratni koren variance
- **koeficient variacije**
 - standardni odklon deljen s povprečjem



Mere razpršenosti

	populacija	vzorec
varianca	σ^2	S^2, s^2
standardni odklon	σ	S, s



Za vzorec smo vzeli osebe na FRI.

Zabeležili smo naslednje

število otrok:

1

2

2

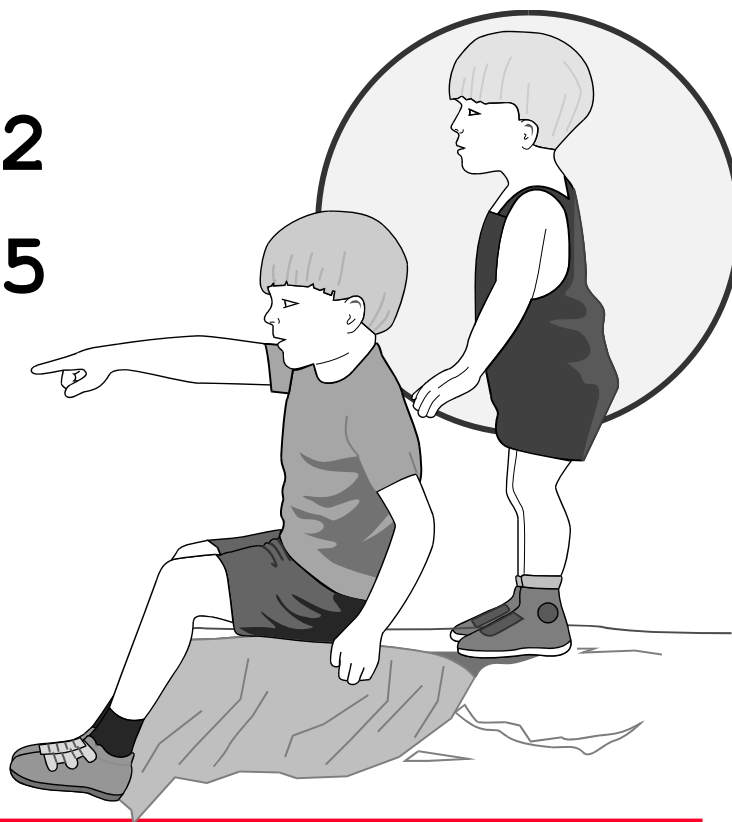
1

2

5

1

2



Varianca

populacije:

$$\sigma^2 = \frac{\sum_{i=1}^n (y_i - \mu)^2}{n}$$

**(končne populacije z
 n meritvami).**



Varianca

vzorca:

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

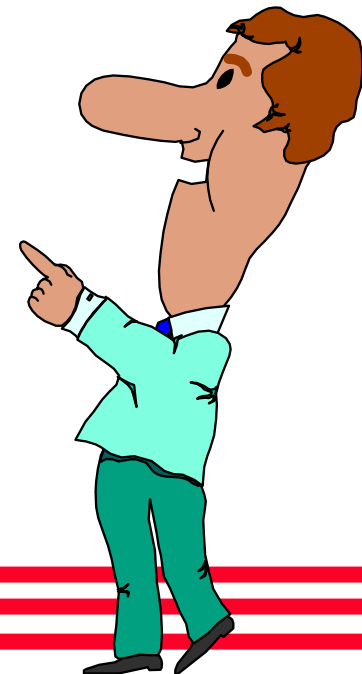
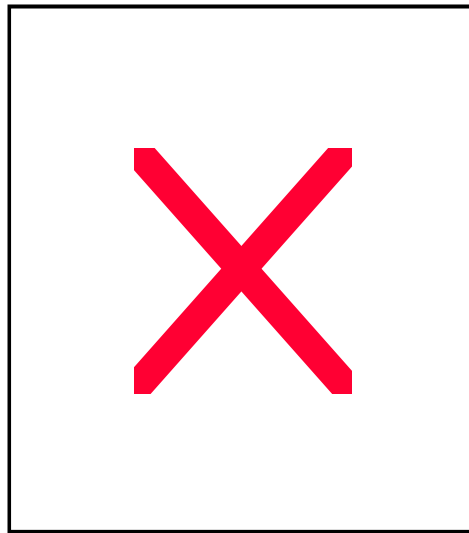
$$= \frac{\sum_{i=1}^n y_i^2 - \frac{\left(\sum_{i=1}^n y_i\right)^2}{n}}{n-1}$$

(z n meritvami).



Standardni odklon

**Standardni odklon je pozitivno
predznačen kvadratni koren variance.**



Empirična pravila

Če ima podatkovna množica porazdelitev približno zvonaste oblike (unimodalna oblika – ima en sam vrh), potem veljajo naslednja pravila (angl. rule of thumb), ki jih lahko uporabimo za opis podatkovne množice:

1. Približno **68,3%** vseh meritev leži na razdalji **1 x standardnega odklona** od njihovega povprečja.



Empirična pravila

2. Približno **95,4%** meritev leži na razdalji do **2 x standardnega odklona** od njihovega povprečja.
3. **Skoraj vse** meritve (**99,7%**) ležijo na razdalji **3 x standardnega odklona** od njihovega povprečja.



Mere oblike

Če je spremenljivka približno normalno porazdeljena, potem jo statistični karakteristiki povprečje in standardni odklon zelo dobro opisujeta.

V primeru unimodalne porazdelitve spremenljivke, ki pa je bolj asimetrična in bolj ali manj sploščena (koničasta), pa je potrebno izračunati še stopnjo **asimetrije** in **sploščenosti** (koničavosti).





Centralni momenti

***l*-ti centralni moment je**

$$m_l = \frac{\sum_{i=1}^n (y_i - \mu)^l}{n}$$

$$m_1 = 0, \quad m_2 = \sigma^2$$





Mere asimetrije

Razlike med srednjimi vrednostimi so tem večje, čim bolj je porazdelitev asimetrična:

$$KA_{M_o} = (\mu - M_o)/\sigma$$

$$KA_{M_e} = 3(\mu - M_e)/\sigma$$

Koeficient asimetrije (s centralnimi momenti):

$$g_1 = m_3/m_2^{3/2}$$





Mera sploščenosti (kurtosis)

**Koeficient sploščenosti
(s centralnimi momenti)**

$$***K = g_2 = m_4/m_2^2 - 3***$$



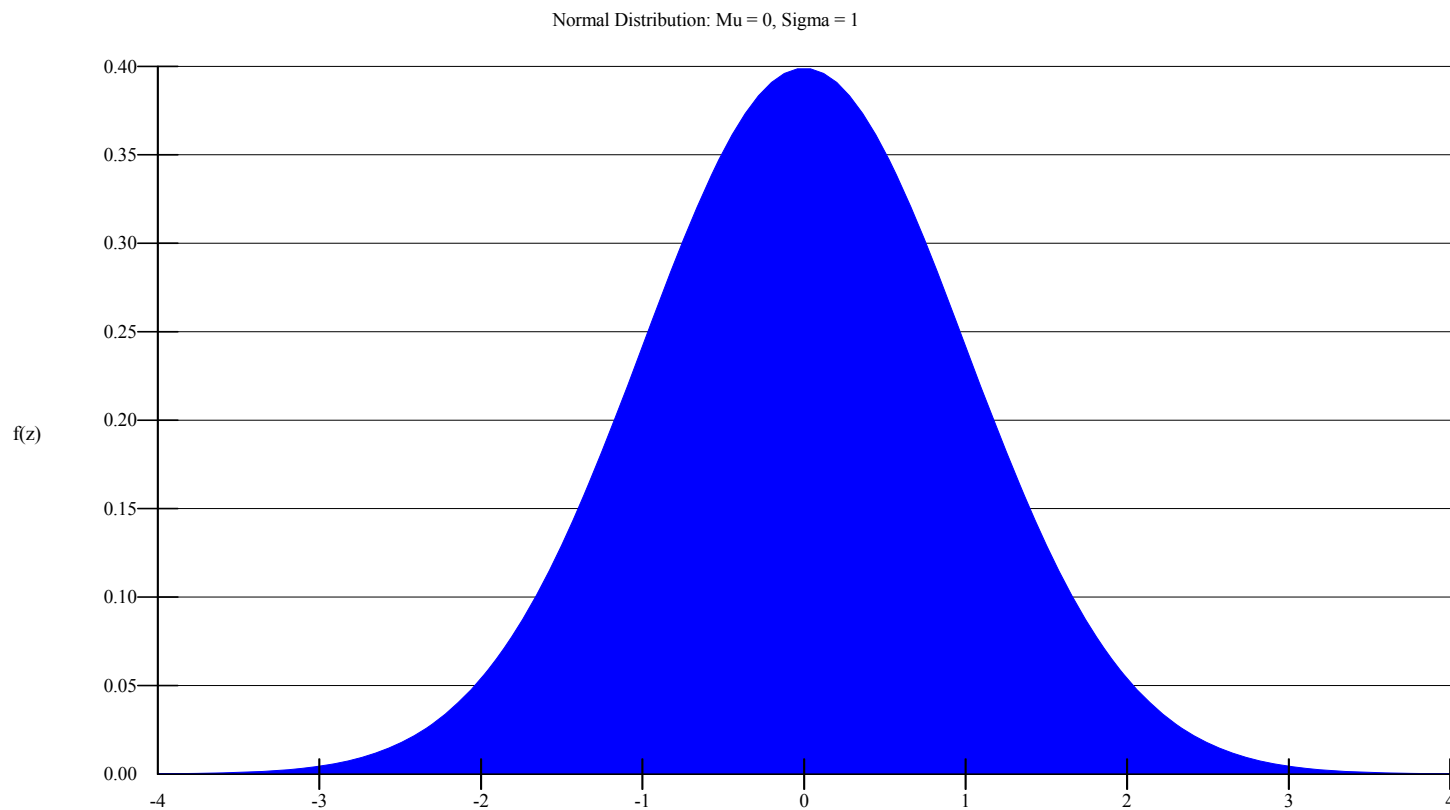


Mera sploščenosti (kurtosis)

- **$K = 3$ (ali 0)**
 - normalna porazdelitev zvonaste-oblike (mesokurtic)
- **$K < 3$ (ali negativna)**
 - bolj kopasta kot normalna porazdelitev, s krajšimi repi (platykurtic)
- **$K > 3$ (ali pozitivna)**
 - bolj špičasta kot normalna porazdelitev, z daljšimi repi (leptokurtic)

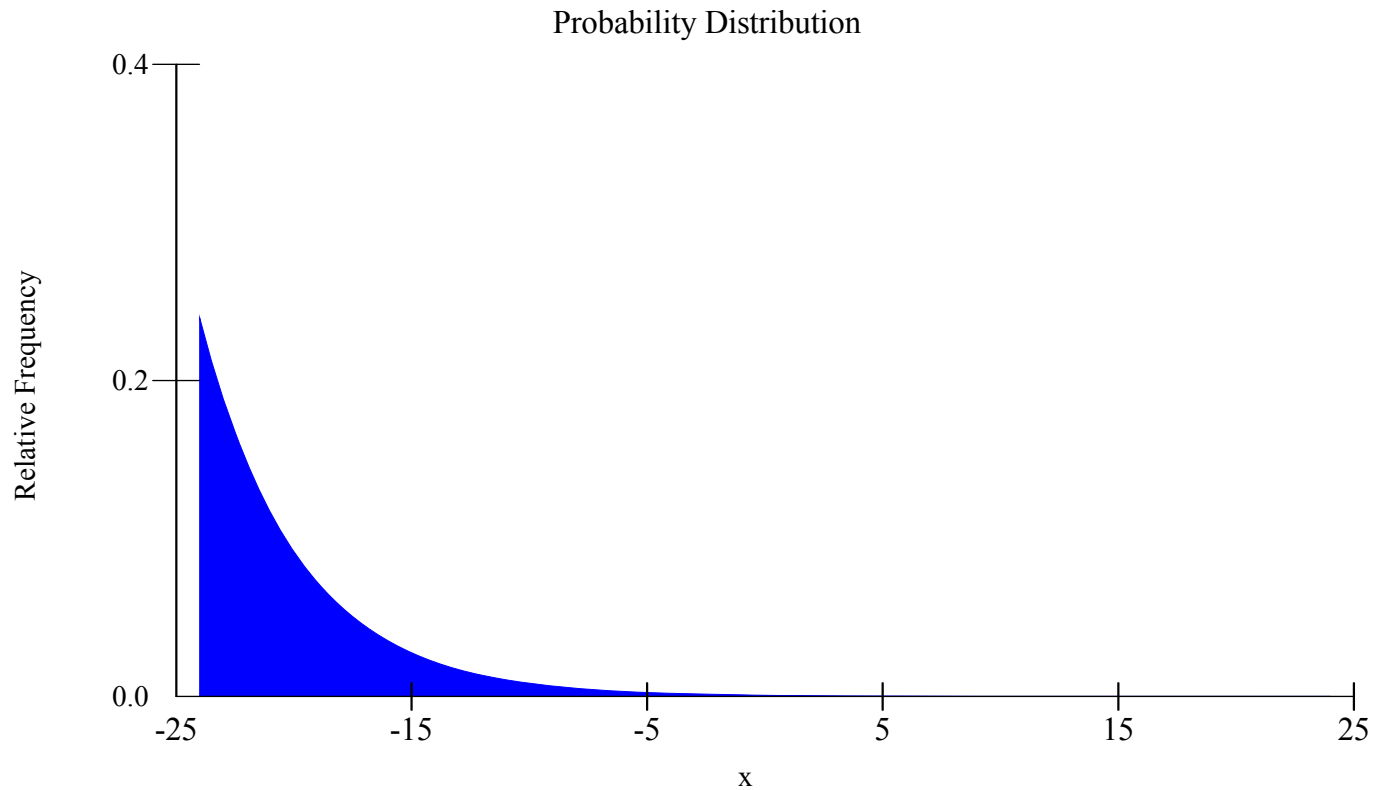


Normalna porazdelitev



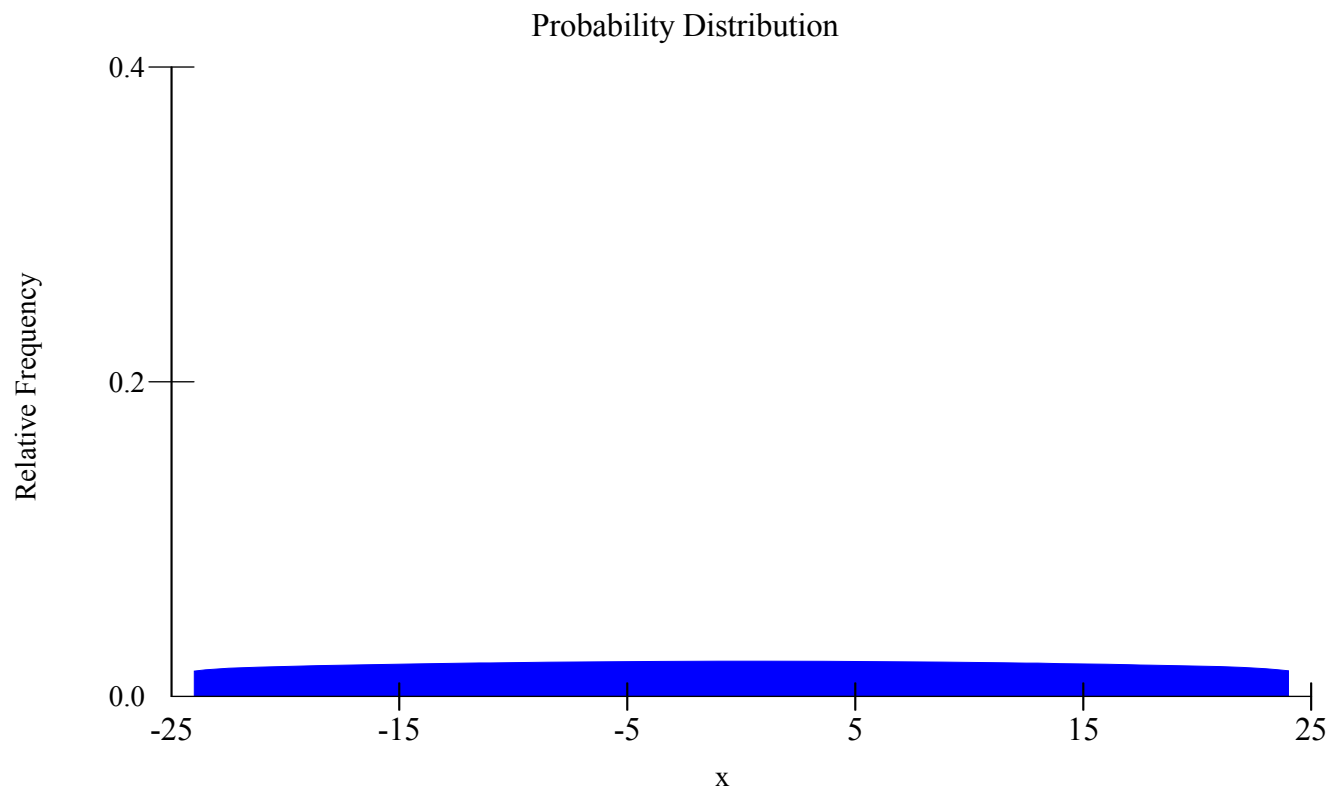
asimetričnost = 0, sploščenost = 3 (mesokurtic)

Asimetrična v desno



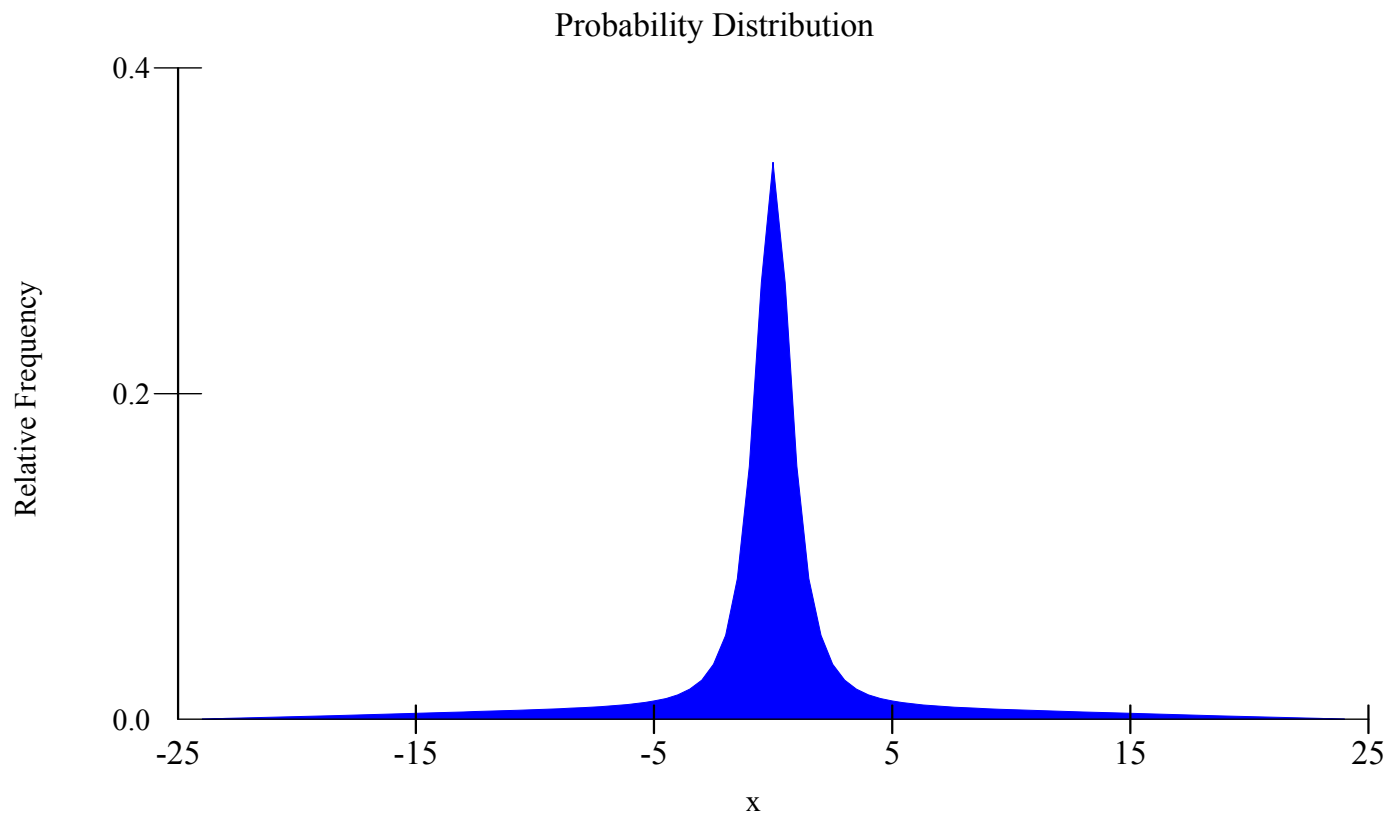
asimetričnost = 1,99, sploščenost = 8,85 (leptokurtic)

Kopasta porazdelitev



asimetričnost = 0, sploščenost = 1,86 (platykurtic)

Špičasta porazdelitev



asimetričnost = $-1,99$, sploščenost = $8,85$ (leptokurtic)



Standardizacija

Vsaki vrednosti x_i spremenljivke X odštejemo njeno povprečje μ in delimo z njenim standardnim odklonom σ :

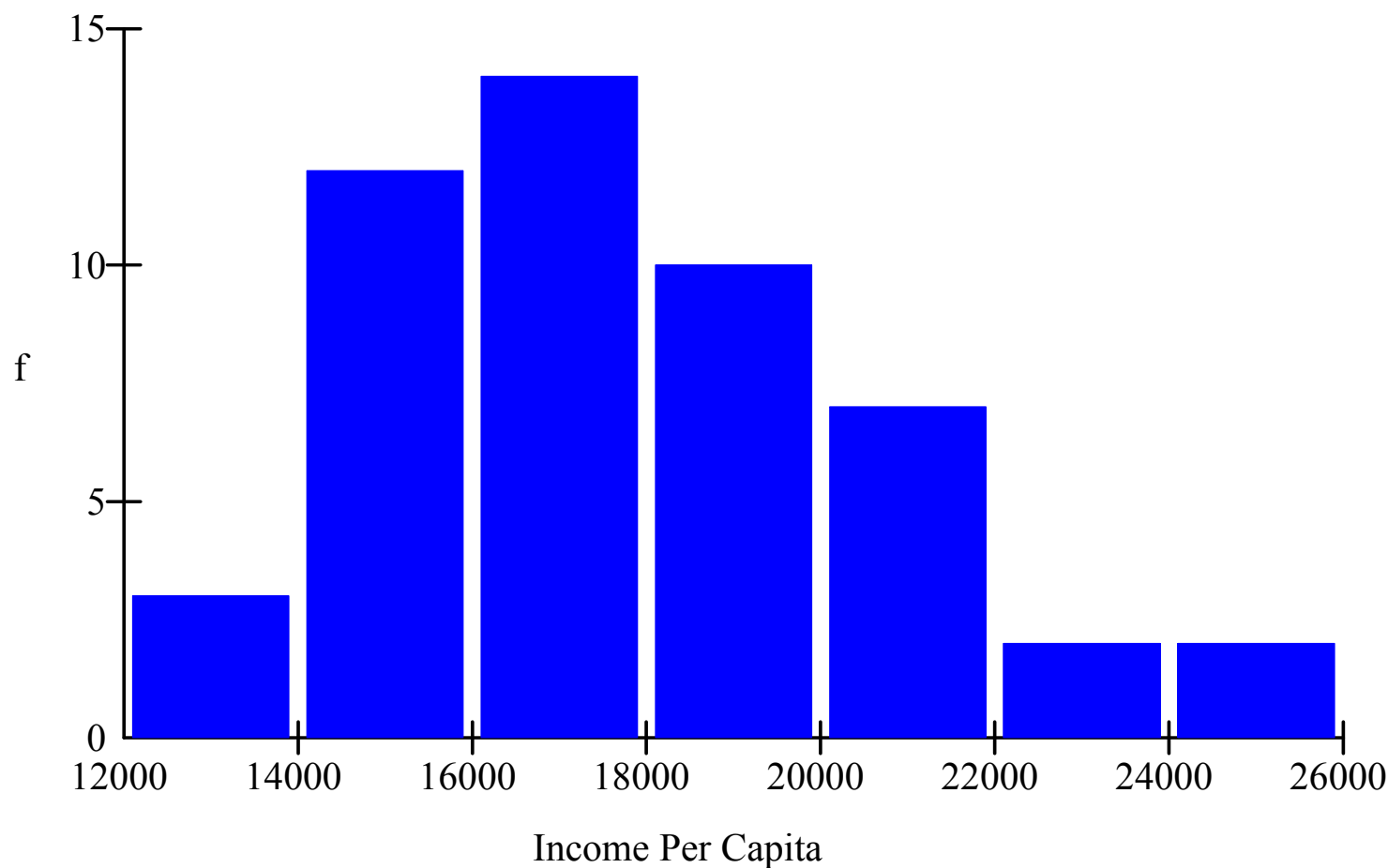
$$z_i = (x_i - \mu) / \sigma$$

Z imenujemo standardizirana spremenljivka, z_i pa standardizirana vrednost.

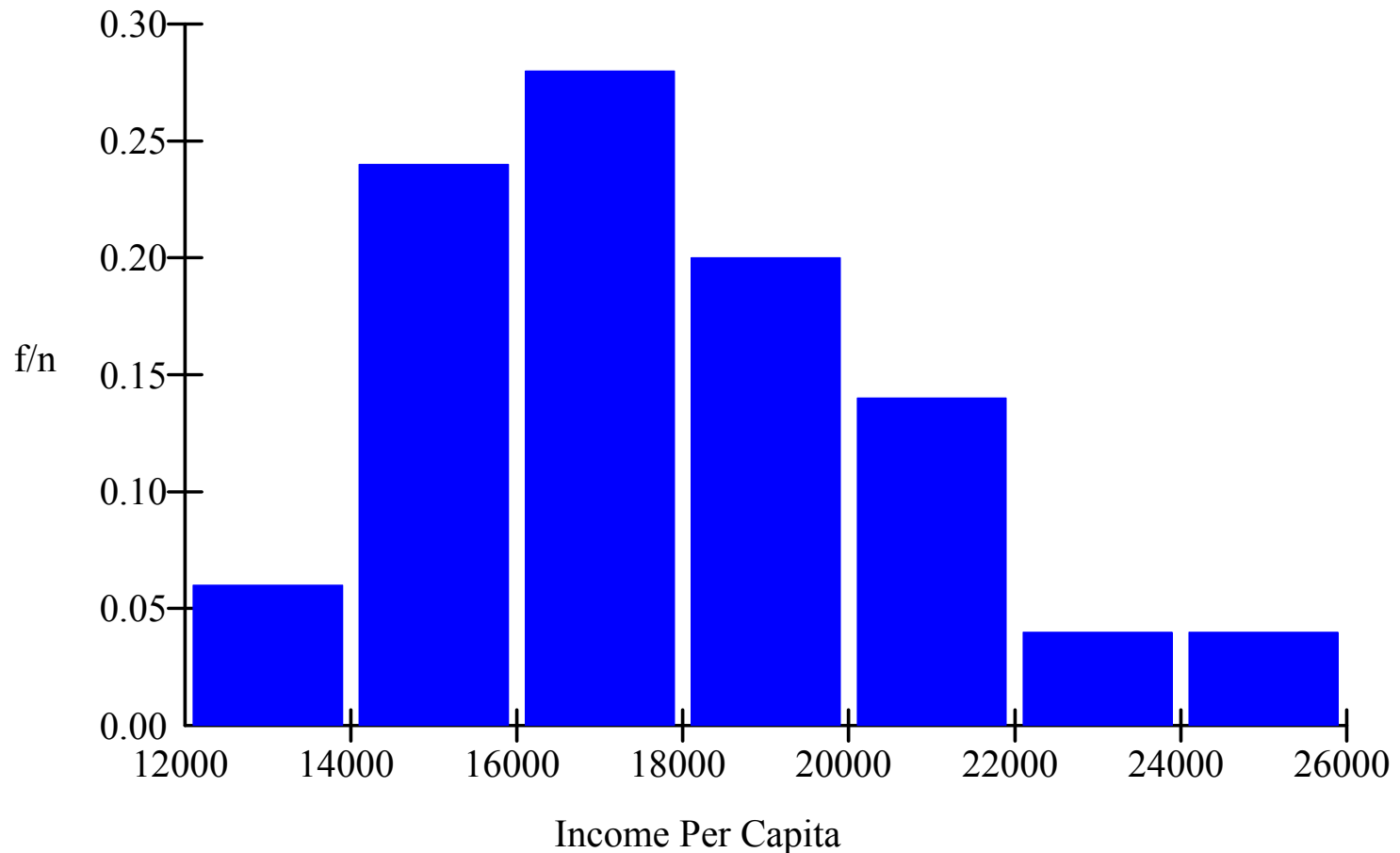
Potem je $\mu(Z)=0$ in $\sigma(Z)=1$.



Frekvenční histogram



Relativni frekvenčni histogram



Histogram standardiziranih Z-vrednosti

