Statistike

Milena Kovač

12. oktober 2012

Naključni vzorec

- merili smo lastnost - spremenljivko x_i

- opravili smo n meritev X_1, X_2, ... X_n
  (npr. na 1000 na naključno izbranih kravah)

- dobimo niz podatkov Z

$$Z = (x_1, x_2, ...x_n)$$

- za prikazovanje podatke preuredimo v statistike

Podatki: mlečnost - krave :-(

23.5, 17.6, 10.8, 15.2, 22.0, 14.4, 9.5, 11.3, 28.9, 36.0, 21.5, 16.2, 13.5, 12.1, 17.7, 13.2, 33.6, 24.5, 8.3, 19.4, 20.5, 11.6, 12.3, 17.6, 10.8, 15.2, 23.7, 22.0, 18.4, 9.5, 11.3, 18.9, 26.0, 24.5, 16.2, 13.5, 12.1, 17.7, 13.2, 33.6, 24.5, 8.3, 19.4, 20.8, 11.6, 12.3, 17.6, 10.8, 15.2, 22.0, 14.4, 9.5, 11.3, 28.2, 34.7, 21.5, 16.2, 13.5, 12.1, 17.4, 13.2, 13.6, 24.2, 28.3, 19.4, 20.1, 11.6, 14.3, 15.6, 20.8, 15.3, 22.1, 14.9, 9.5, 11.3, 28.3, 26.1, 21.2, 16.6, 23.5, 12.6, 17.7, 23.2, 30.3, 24.5, 8.3, 19.4, 20.9, 11.6, 12.3, 15.2, 22.0, 14.4, 9.5, 11.3, 28.2, 34.7, 21.5, 16.2, 13.5, 15.1, 17.4, 13.2, 13.6, 24.2, 28.3, 19.4, 20.8, 11.6, 14.3, 15.6, 20.8, 15.3, 22.1, 14.9, 9.5, 11.3, 28.3, 26.1, 21.2, 16.6, 23.5, 12.5, 17.7, 23.2, 30.3, 24.5, 8.3, 19.4, 20.8, 11.6, 15.2, 22.0, 14.4, 9.5, 14.3, 28.2, 34.7, 21.5, 16.2, 13.5, 12.1, 17.4, 13.2, 13.6, 24.2, 28.3, 19.4, 20.1, 11.6, 14.3, 15.6, 20.3, 15.7, 22.1, 14.9, 9.5, 11.3, 28.3, 26.1, 21.2, 16.6, 23.5, 12.1, 17.7, 23.2, 30.3, 24.5, 8.3, 19.4, 20.8, 14.9, 13.2, 13.6, 24.2, 28.3, 19.4, 20.1, 11.6, 14.3, 15.6, 20.8, 15.3, 22.1, 14.9, 9.5, 11.3, 28.3, 26.1, 21.2, 16.6, 23.5, 13.1, 17.7, 23.2, 30.3, 24.5, 8.3, 19.4, 20.8, 11.6, 12.3, 15.2, 22.0, 14.4, 9.5, 11.3, 28.2, 34.7, 21.5, 16.2, 13.5, 18.1, 17.4, 13.2, 13.6, 24.2, 28.3, 19.4, 20.1, 11.6, 14.3, 15.6, 23.8, 15.3, 22.1, 14.9, 9.5, 11.3, 28.3, 12.6, 17.7, 23.2, 30.3, 24.5, 8.3, 19.4, 20.8, 11.6, 12.3, 15.2, 22.0, 14.4, 9.5, 11.3, 28.2, 34.7, 21.5, 16.2, 13.5, 12.1, 17.4, 13.2, 13.6, 24.2, 28.3, 19.4, 21.9, 11.6, 14.3, 15.6, 25.8, 15.3, 22.1, 14.9, 9.5, 11.6, 28.3, 26.1, 21.2, 16.4, 23.5, 18.2, 37.8, 5.7, 17.3, 12.4, 21.7, 24.4, 25.8, 10.9, 41.5, 15.1, 27.4, ...

Podatki: mlečnost - krave :-)

Krava	#lakt	Dt. kont.	DKM (kg)	Mašč. (%)	Belj. (%)	Rejec	Pasma
76811	2	20.2.08	23.5	4.07	3.34	Lojtrca	ČB
65231	3	20.2.08	17.6	3.99	3.18	Lojtrca	ČB
70023	1	20.2.08	10.8	4.23	3.60	Lojtrca	ČB
54617	1	24.2.08	15.2	4.15	3.47	Vošca	RJ
43213	7	24.2.08	22.0	4.03	3.26	Vošca	RJ
34572	5	24.2.08	14.4	3.84	3.19	Ortnik	ČB
56801	4	2.3.08	9.5	4.27	3.32	Ortnik	CIKA
54521	2	2.3.08	11.3	3.95	3.04	Ortnik	RJ
...							

Statistike

1. Število in frekvenca

2. Srednje vrednosti

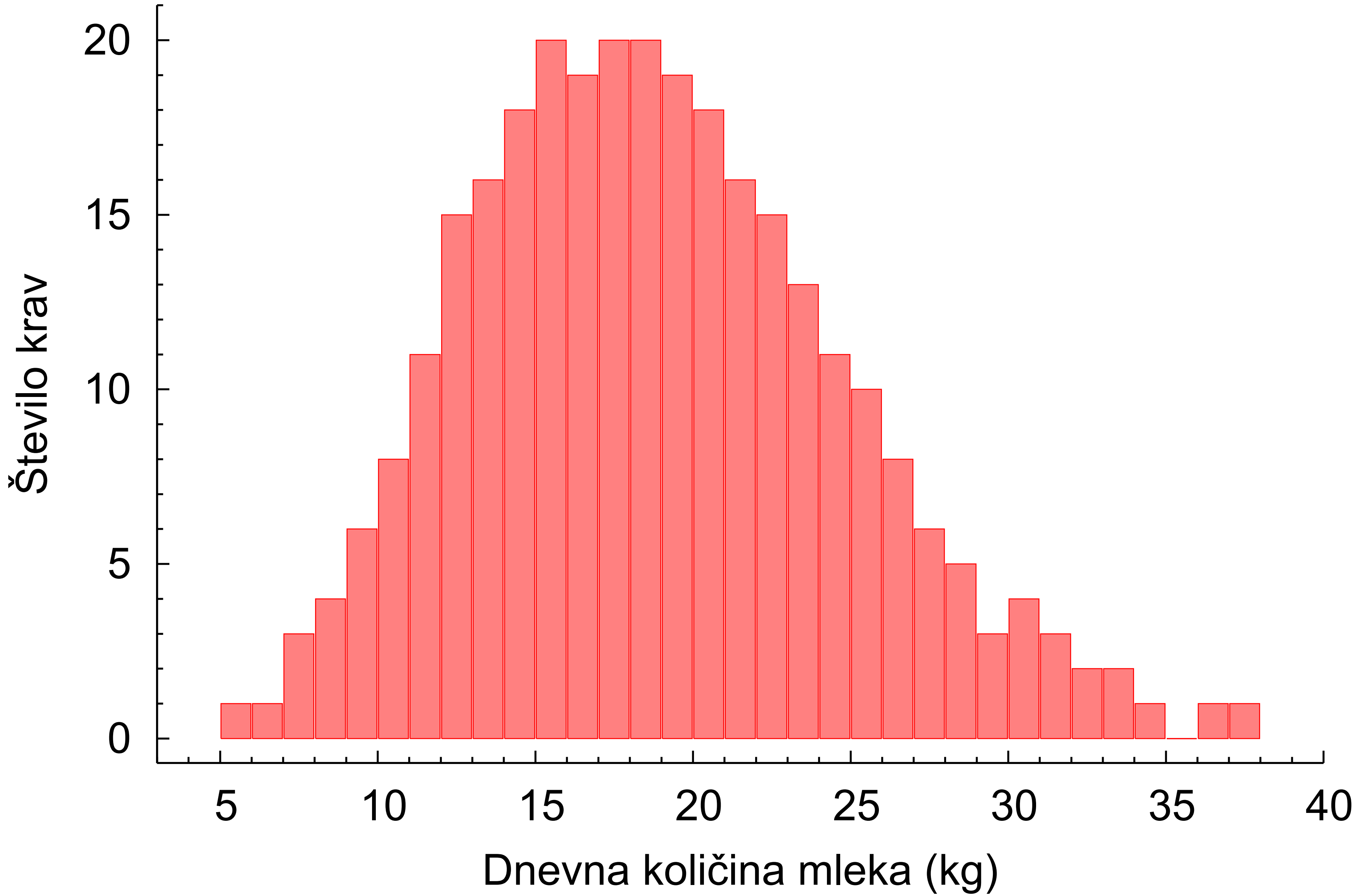
3. Mere razpršenosti

4. Mere podobnosti

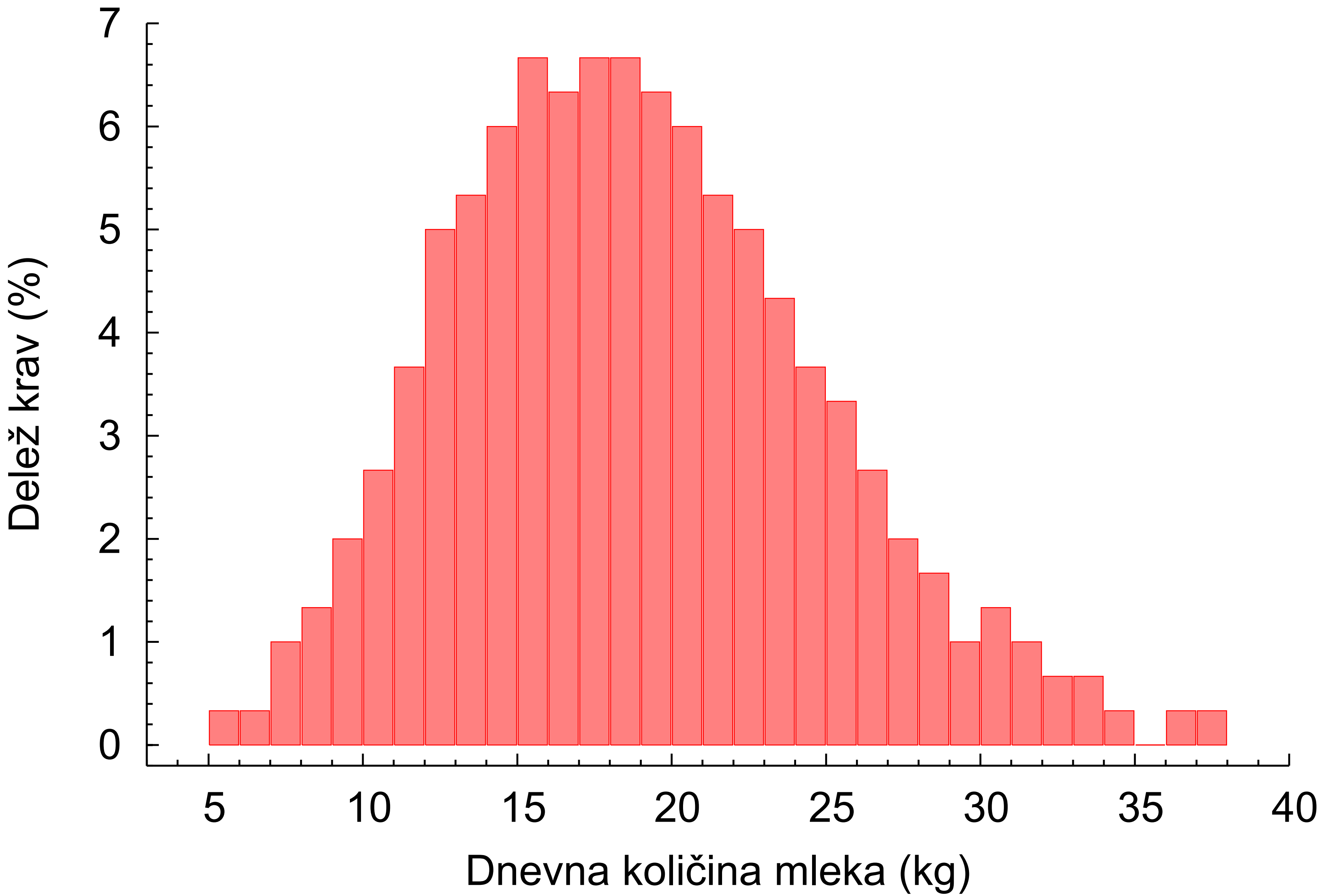
Grupiranje v razrede

DKM	Frekvenca		DKM	Frekvenca		DKM	Frekvenca	
5	I	1	16	𝍸 𝍸 𝍸 IIII	19	27	𝍸 I	6
6	I	1	17	𝍸 𝍸 𝍸 𝍸	20	28	𝍸	5
7	III	3	18	𝍸 𝍸 𝍸 𝍸	20	29	III	3
8	IIII	4	19	𝍸 𝍸 𝍸 IIII	19	30	IIII	4
9	𝍸 I	6	20	𝍸 𝍸 𝍸 III	18	31	III	3
10	𝍸 III	8	21	𝍸 𝍸 𝍸 I	16	32	II	2
11	𝍸 𝍸 I	11	22	𝍸 𝍸 𝍸	15	33	II	2
12	𝍸 𝍸 𝍸	15	23	𝍸 𝍸 III	13	34	I	1
13	𝍸 𝍸 𝍸 I	16	24	𝍸 𝍸 I	11	35		0
14	𝍸 𝍸 𝍸 III	18	25	𝍸 𝍸	10	36	I	1
15	𝍸 𝍸 𝍸 𝍸	20	26	𝍸 III	8	37	I	1

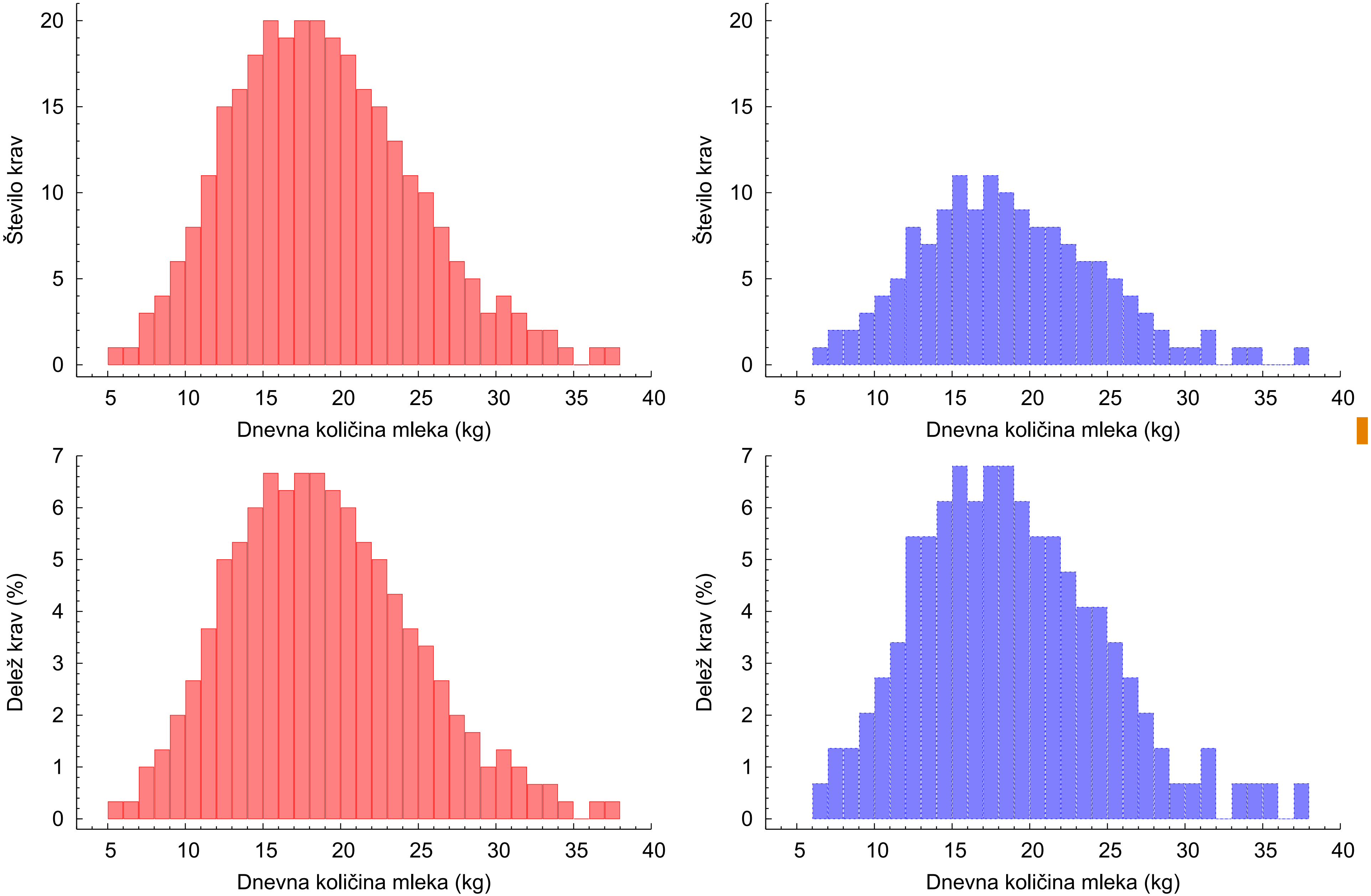
Absolutna frekvenca



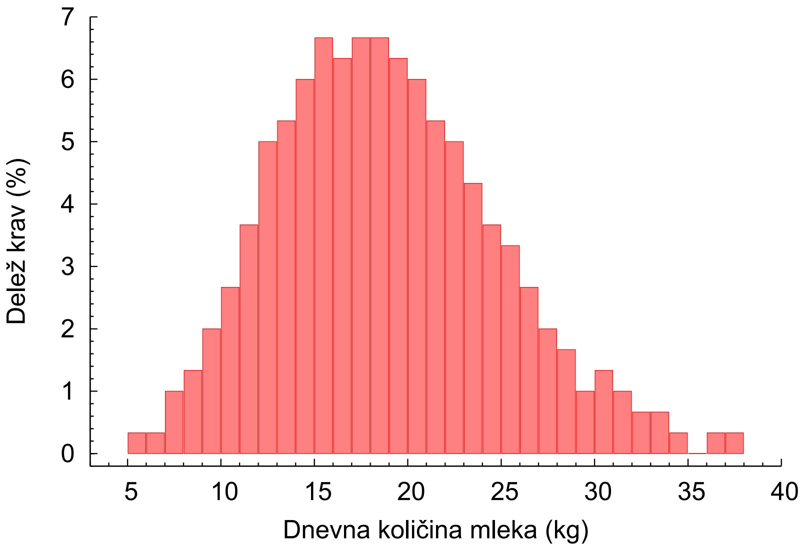
Relativna frekvenca



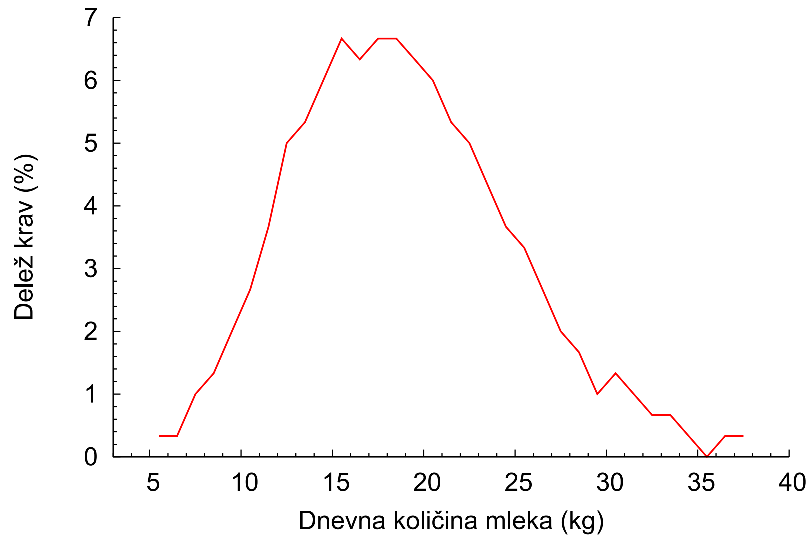
Absolutna ali relativna frekvenca?



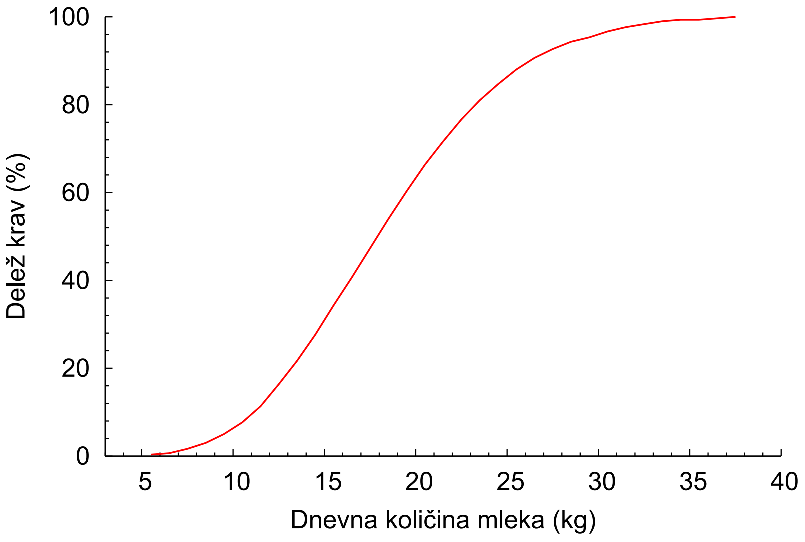
Grafikoni za prikaz frekvenc



Histogram



Poligon



Kumulativni prikaz

Srednje vrednosti

1. Aritmetična sredina

2. Geometrijska sredina

3. Harmonična sredina

4. Mediana

5. Modus

Aritmetična sredina

- **Sinonimi**: povprečje, srednja vrednost, matematično upanje ...
- predstavlja lokacijo porazdelitve $\rightarrow$ lokacijski parameter
- uporaben parameter pri normalni, Poissonovi porazdelitvi, a izgubimo
  - razpršitev, povezave med meritvami,
  - spreminjanje meritev s časom ...
- **ni vedno zadostna statistika & primerna statistika**
- leži med minimalno in maksimalno vrednostjo

(Navadna) aritmetična sredina

- vsoto meritev (x_i) delimo s številom opazovanj (n)

$$\overline{X} = \frac{1}{n} * \sum_{i=1}^{n} x_i$$

- vsota odklonov $\left(x_i - \overline{X}\right)$ je enaka 0

$$\sum_{i=1}^{n} \left(x_i - \overline{X}\right) = 0$$

- vsota kvadratnih odklonov $\left(x_i - \overline{X}\right)^2$ je minimalna

$$\sum_{i=1}^{n} \left(x_i - \overline{X}\right)^2 = min$$

Povprečna masa rogovja

V dveh revirjih so tehtali rogovje jelenov. Našli so 16 rogovij, podatki so navedeni spodaj.

Revir	Jelen	Masa (kg)	Revir	Jelen	Masa (kg)
1	1	13.0	2	9	10.6
1	2	10.3	2	10	10.2
1	3	12.0	2	11	13.7
1	4	13.3	2	12	12.7
1	5	14.0	2	13	14.1
1	6	12.4	2	14	14.2
1	7	14.2	2	15	12.0
1	8	13.5	2	16	10.4

Izračunajmo!

Skupina	Revir	Število	Povprečje
A	1		
B	2		
C	1+2		

Povprečna masa rogovja

V dveh revirjih so tehtali rogovje jelenov. Našli so 16 rogovij, podatki so navedeni spodaj.

Revir	Jelen	Masa (kg)	Revir	Jelen	Masa (kg)
1	1	13.0	2	9	10.6
1	2	10.3	2	10	10.2
1	3	12.0	2	11	13.7
1	4	13.3	2	12	12.7
1	5	14.0	2	13	14.1
1	6	12.4	2	14	14.2
1	7	14.2	2	15	12.0
1	8	13.5	2	16	10.4

Izračunajmo!

Skupina	Revir	Število	Povprečje
A	1	8	12.8
B	2	8	12.2
C	1+2	16	12.5

$$\overline{X}_C = \frac{12.8 + 12.2}{2} = 12.5$$

Povprečna masa rogovja

Tudi jeleni, a malo drugače ...

Revir	Jelen	Masa (kg)	Revir	Jelen	Masa (kg)
1	1	13.0	2	9	10.6
1	2	14.1	2	10	10.1
1	3	12.0	2	11	13.7
1	4	13.2	2	12	12.7
1	5	14.0	2	13	10.3
1	6	12.4	2	14	10.4
1	7	14.2	2	15	12.0
1	8	13.5			
1	16	14.2			

Izračunajmo!

Skupina	Revir	Število	Povprečje
A	1	9	13.4
B	2	7	11.4
C	1+2	16	12.5

$$\overline{X}_C = \frac{13.4 + 11.4}{2} = 12.4$$

Primerjajmo!

	Poskus 1		Poskus 2	
Revir	Število	Povprečje	Število	Povprečje
1	8	12.8	9	13.4
2	8	12.2	7	11.4
3	16	12.5	16	12.5
$\overline{X}_C$		12.5		12.4
Poskus je	uravnotežen		neuravnotežen	

Pri neuravnoteženih poskusih ne smemo uporabljati navadne aritmetične sredine.

Tehtana aritmetična sredina

Frekvenca opazovanj f_i v razredih $i = 1, 2, \dots r$

$$\overline{X} = \sum_{i=1}^{r} f_i \bar{x}_i$$

Število opazovanj n_i v razredih $i = 1, 2, \dots r$

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{r} n_i \bar{x}_i$$

$$\overline{X}_C = \frac{9 * 13.4 + 7 * 11.4}{9 + 7} = 12.5$$

Poskus: ovce

Kmet je dobil analizo s številom gnezd in povprečji za velikost gnezda za vsako pasmo, zanima pa ga povprečje za celotno čredo.

Pasma	Število (n_i)	Delež (f_i)	Število jagnjet/gnezdo($\bar{x}_i$)
A	8	0.08	1.25
B	90	0.90	2.30
C	2	0.02	4.00
Skupaj	100	1.00	

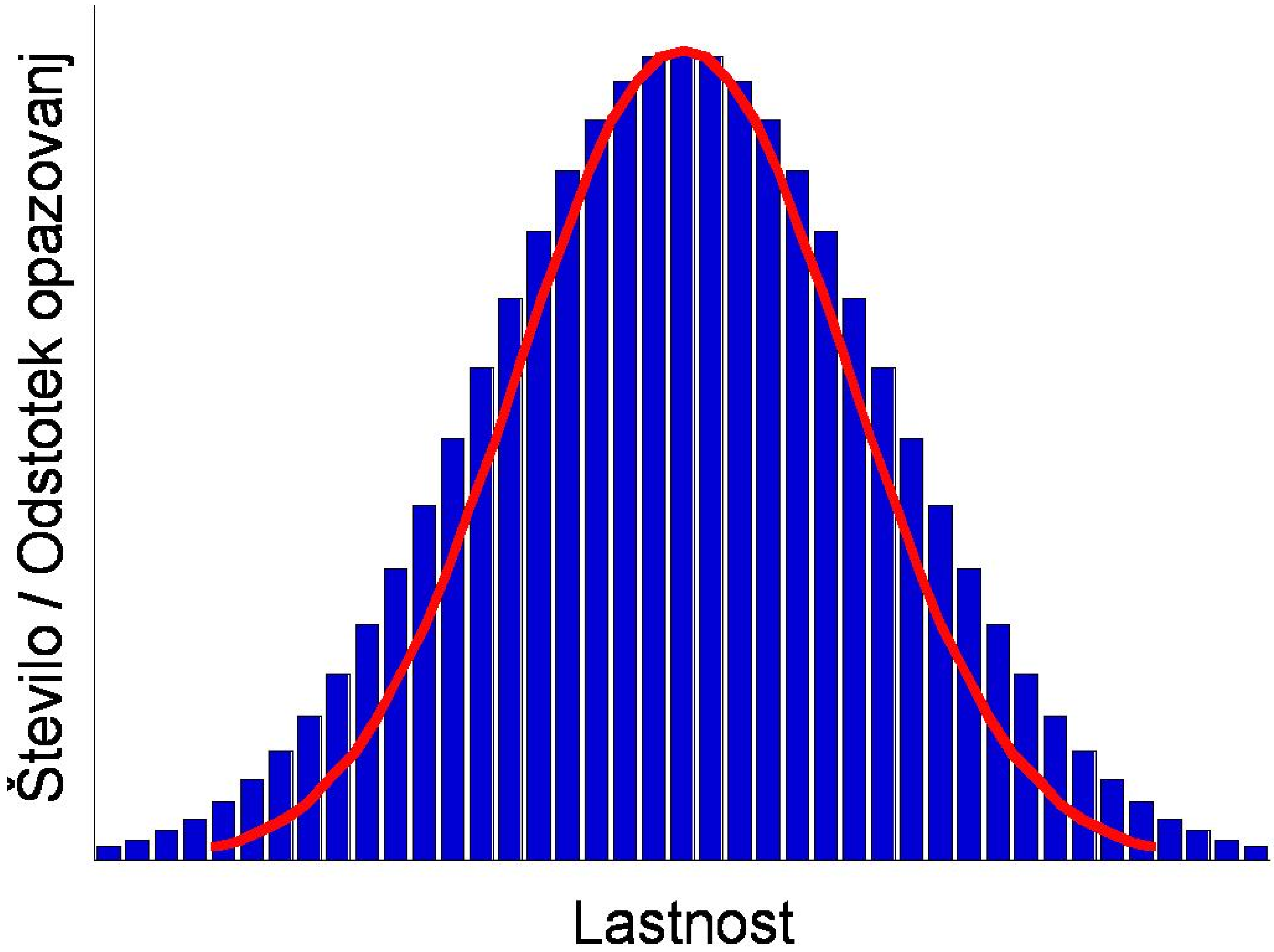
$$\overline{X}_S = (8 * 1.25 + 90 * 2.30 + 2 * 4.00)/(8 + 90 + 2) = 2.25$$

$$\overline{X}_S = 0.08 * 1.25 + 0.90 * 2.30 + 0.02 * 4.00 = 2.25$$

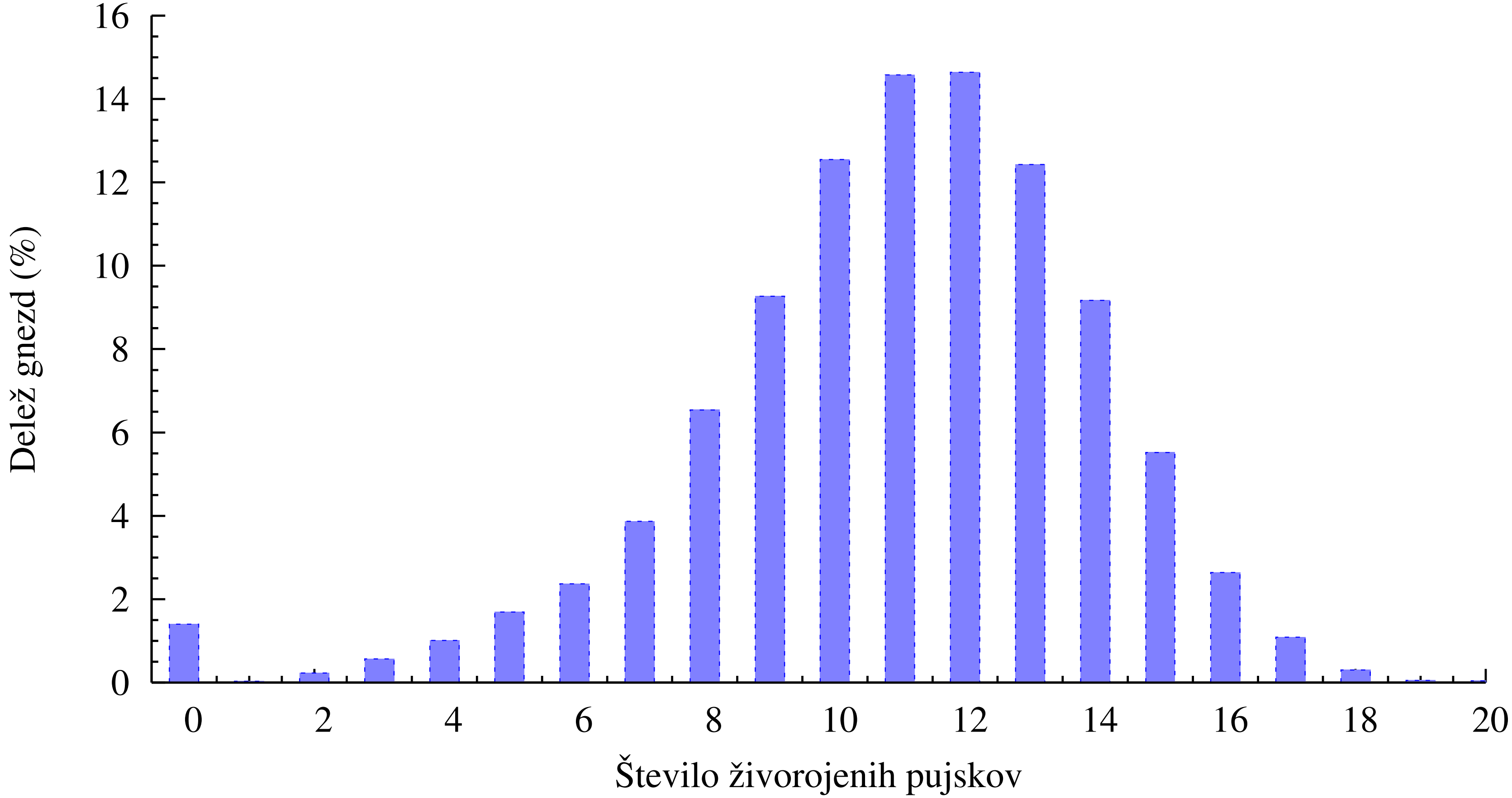
Uporaba aritmetične sredine

- pri normalni ali Gaussovi porazdelitvi
- pri Poissonovi porazdelitvi
- v populacijah: μ
- za vzorec: $\bar{x}$ ali $\overline{X}$
- **ni primerno**: neenovite populacije, asimetrične porazdelitve, porazdelitve z več vrhovi ...

Normalna porazdelitev



Diskretna porazdelitev



Geometrijska sredina

- n-ti koren iz produkta posameznih vrednosti spremenljivke x_i

$$\overline{X}_g = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

- uporablja se samo pri pozitivnih vrednostih spremenljivke x_i

- imenujejo jo tudi srednja proporcionalna vrednost

Geometrijska sredina (nadalj.)

- $min. < \overline{X}_g < max.$

- uporabna pri

  - verižnih indeksih,
  - koeficientih rasti,
  - stopnjah rasti,
  - število somatskih celic (obdelujemo na logaritemski skali)...

Preverimo!

- antilogaritem aritmetične sredine logaritmiranih vrednosti spremenljivke x_i

$$log\left(\sqrt[n]{\prod_{i=1}^{n} x_i}\right) = log\left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}} =$$

$$= \frac{1}{n} * \sum_{i=1}^{n} log(x_i) = \overline{X_{log}}$$

- geometrijska sredina

$$\overline{X_g} = alog(\overline{X_{log}}) = exp(\overline{X_{log}})$$

Harmonična sredina

$$\overline{X_h} = n * \left(\sum \frac{1}{x_i} \right)^{-1}$$

$$= \frac{n}{1/x_1 + 1/x_2 + \cdots + 1/x_n}$$

- inverzna vrednost povprečja inverznih vrednosti spremenljivke x_i
- povprečna efektivna velikost populacije po generacijah
- dnevni prirast - hitrost rasti (!)

Mediana

Sinonim: centralna vrednost

- vrednost spremenljivke, ki razdeli meritve na dve polovici
- število meritev večjih in manjših od nje je enako
- neobčutljiva na posamezne vrednosti spremenljivk, dokler spremenjena vrednost ostane na isti strani mediane
- pove o podatkih manj kot povprečje
- primerna, če porazdelitev ni simetrična

Določanje mediane

- razvrstimo meritve po vrednosti
  - od minimuma k maksimumu
  - ali obratno

- **liho število opazovanj**

  $\Rightarrow$ mediana dobi vrednost srednje enote

- **sodo število opazovanj**

  $\Rightarrow$ mediana je povprečje srednjega para meritev

Primer: trajanje zdravljenja

Sedem živali smo zdravili. Zabeležili smo trajanje bolezni. Večina živali se je pozdravila po približno enem tednu, dve pa sta se pozdravili šele po mesecu dni.

Žival	1	2	3	4	5	6	7
Trajanje bolezni (dni)	6	6	7	7	8	29	35

Izračunajte:

1. povprečje = 14 dni
   (velik vpliv dveh dolgo bolnih živali)

2. mediana = 7 dni (bolje opiše trajanje zdravljenja)

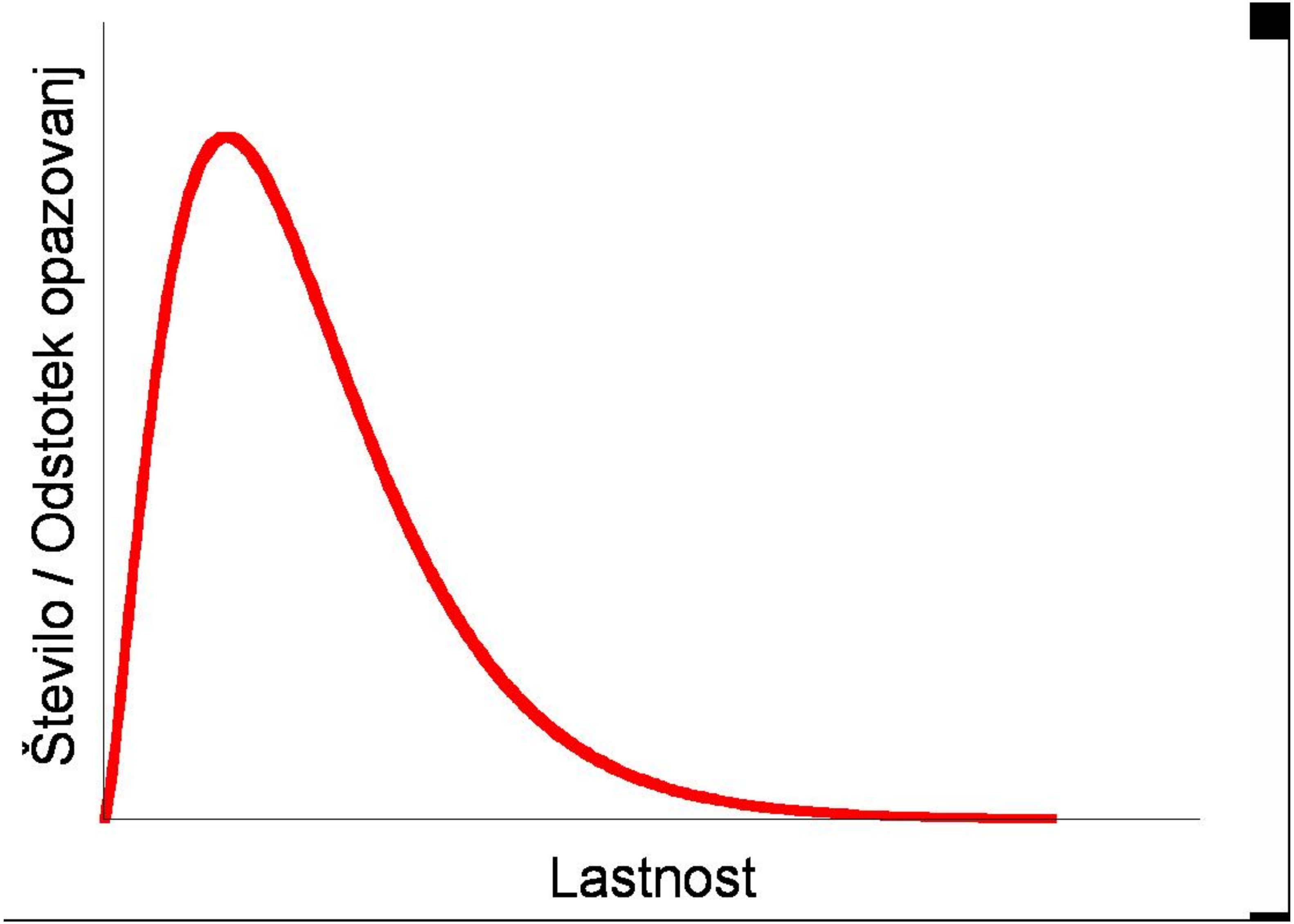
Modus

- **najpogostejša vrednost**
- unimodalne, bimodalne in polimodalne porazdelitve
- pri manjših vzorcih nezanesljivo
- primeri:
  - pojav estrusa po odstavitvi pri svinjah
  - dnevni ritmi za lastnosti obnašanja

Uporaba modusa

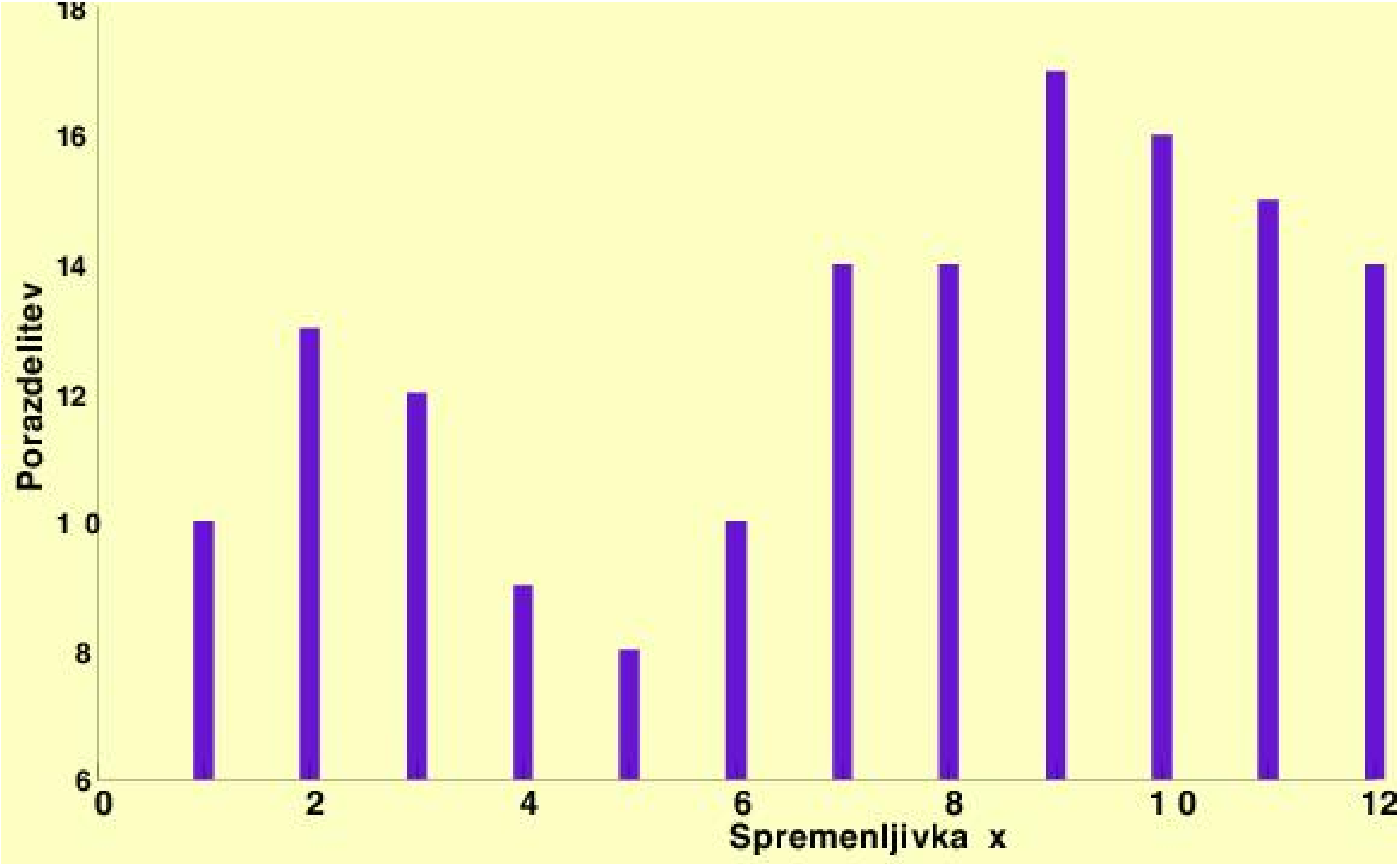
- pri selekcioniranih vzorcih
  - pri lastnostih, blizu (biološke) limite
  - pri stopenjskih odbirah
- pri heterogenih in asimetričnih porazdelitvah
  - lastnosti, omejene navzdol ali navzgor
- **ali je mogoče lastnost transformirati?**

Primer: Asimetrična zvezna porazdelitev



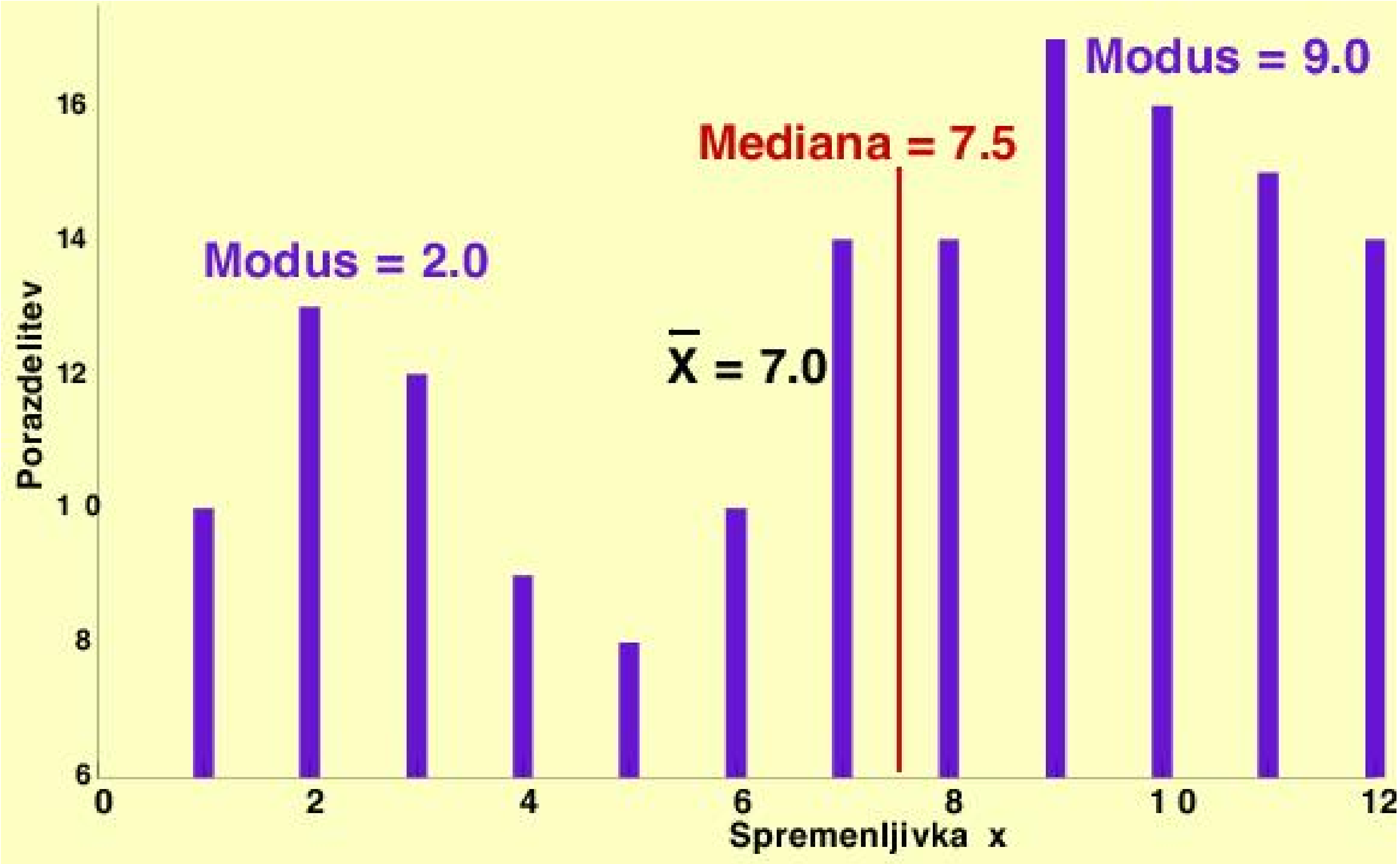
Določite srednjo vrednost, mediano in modus!

Primer: diskretna porazdelitev z dvema vrhovoma



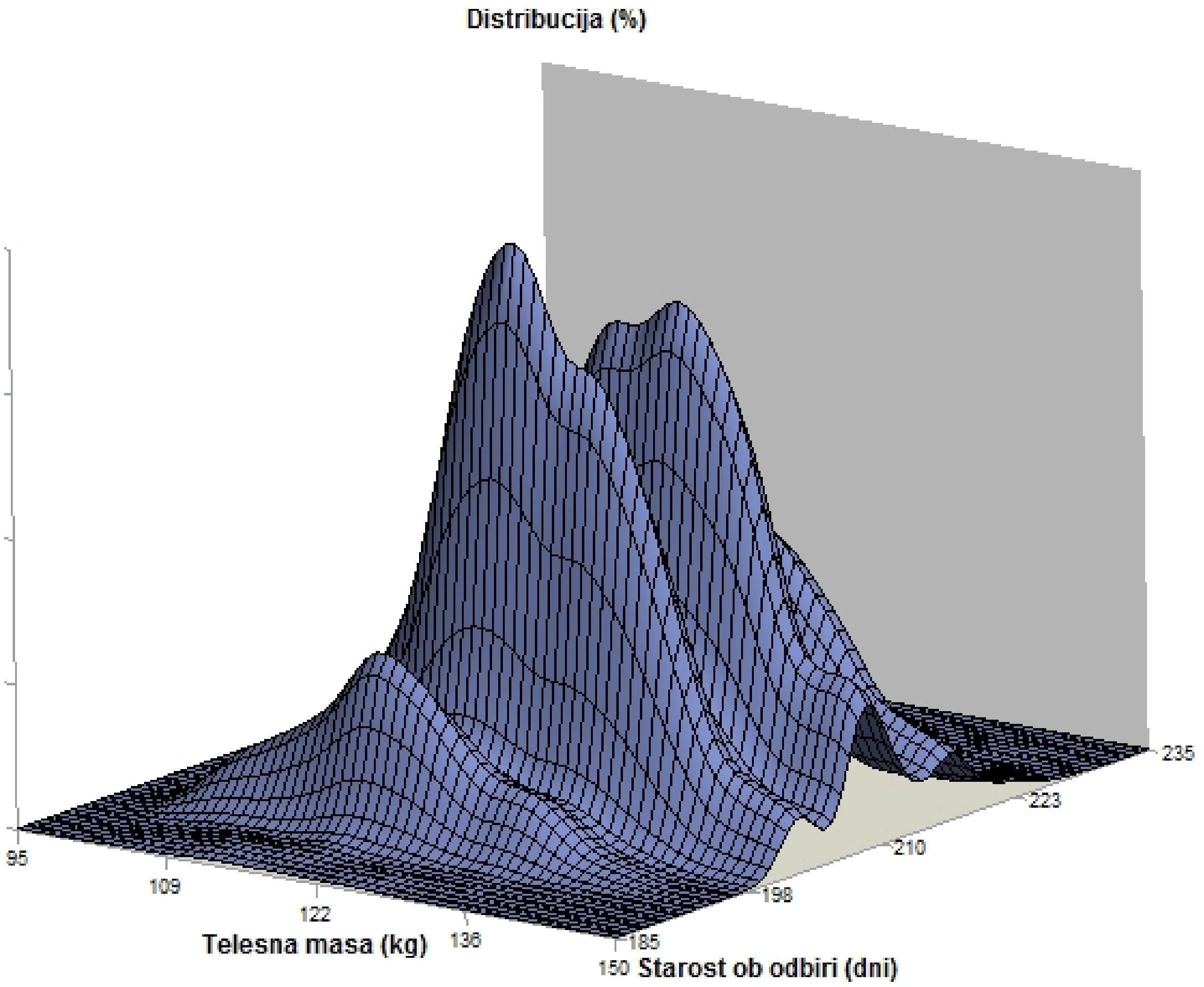
Določite srednjo vrednost, mediano in modus!

Primer: diskretna porazdelitev z dvema vrhovoma

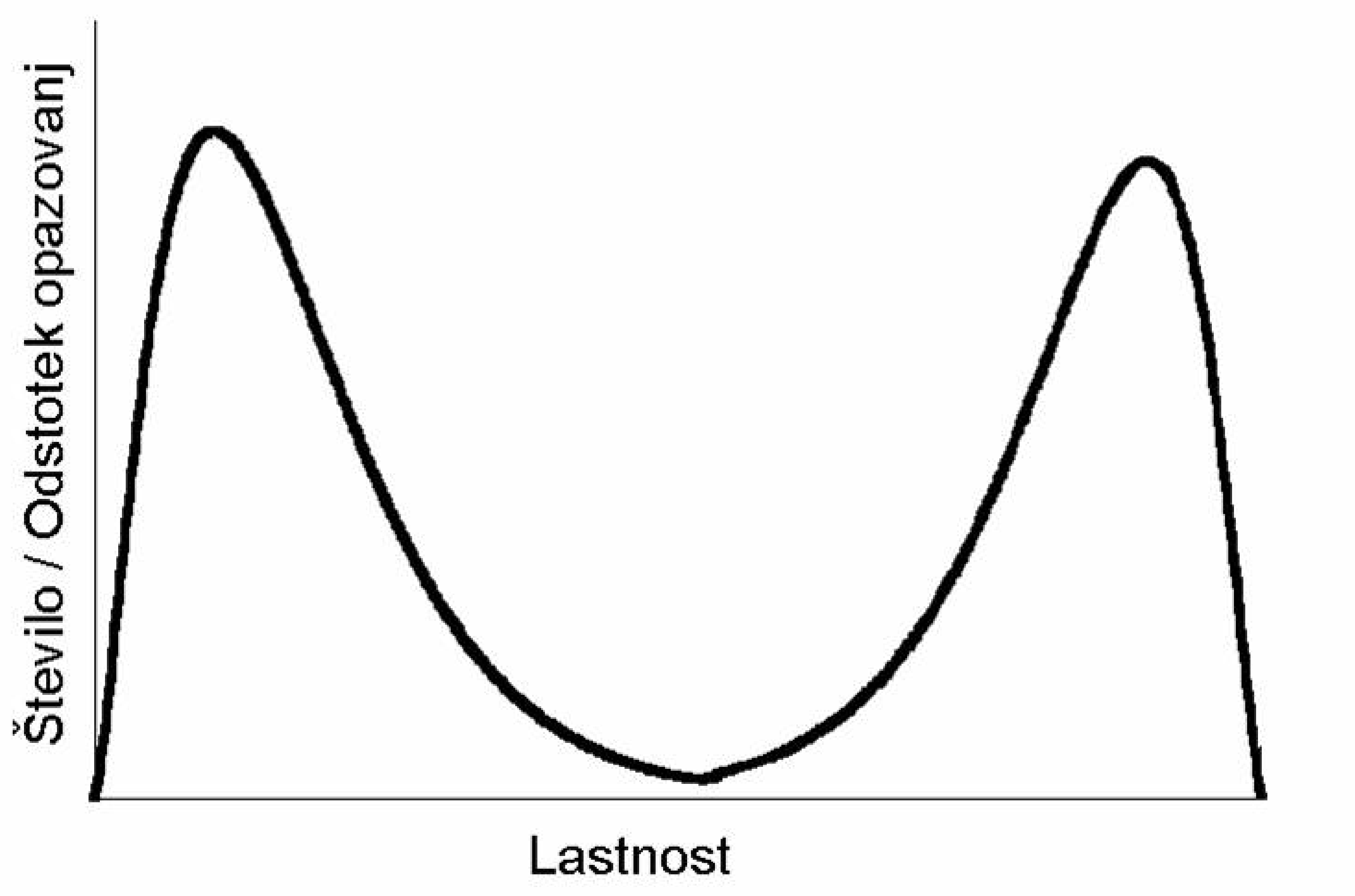


Ali je srednja vrednost primerna statistika?

Primer: večmodalna porazdelitev



Primer: zvezna porazdelitev z dvema vrhovoma



Določite srednjo vrednost, mediano in modus!

Ali je srednja vrednost primerna statistika?