OSNOVE STATISTIKE

FKKT-kemijski tehnologi 1.letnik

2011

Miran Černe

- Statistika je način, kako iz množice podatkov izluščiti ustrezne informacije.

- Izraz izhaja iz latinskih besed

  STATUS = stanje

  STATO = država

- Danes pod besedo statistika razumemo:

  številske podatke,

  zbiranje podatkov,

  statistične urade (SURS)

- Statistika je znanstvena veda, ki proučuje množične pojave

OSNOVNI POJMI STATISTIKE

- **Opisna ali deskriptivna statistika** – analiza zbranih podatkov brez težnje, da bi iz teh podatkov posploševali čez njihov obseg. Uporablja metode organiziranja, sumiranja in predstavitve podatkov na primeren informativen način.

  Zgled: Iz podatkov o osebnih dohodkih v Krki lahko izračunamo njihov povprečni dohodek, ne moremo pa sklepati na povprečni dohodek v RS.

- **Sklepno ali inferenčno statistiko**-le ta uporablja statistično sklepanje iz vzorca (dela populacije) na celotno populacijo

  Zgled: Ankete o referendumu, ankete o gledanosti programov, ugotavljanje kakovosti izdelkov iz množične proizvodnje

POJMI V STATISTIKI

1. **Statistična množica ali populacija**- to je množica **statističnih enot**, ki jo statistično proučujemo. Običajno je to velika množica (lahko tudi neskončna).

Zgledi statističnih populacij:

- Državljani RS v letu 2011
- Študentje UL v študijskem letu 2010/11
- Študentje FKKT UL v študijskem letu 2010/11
- Avtomobili, ki so jih izdelali v Revozu v letu 2010.

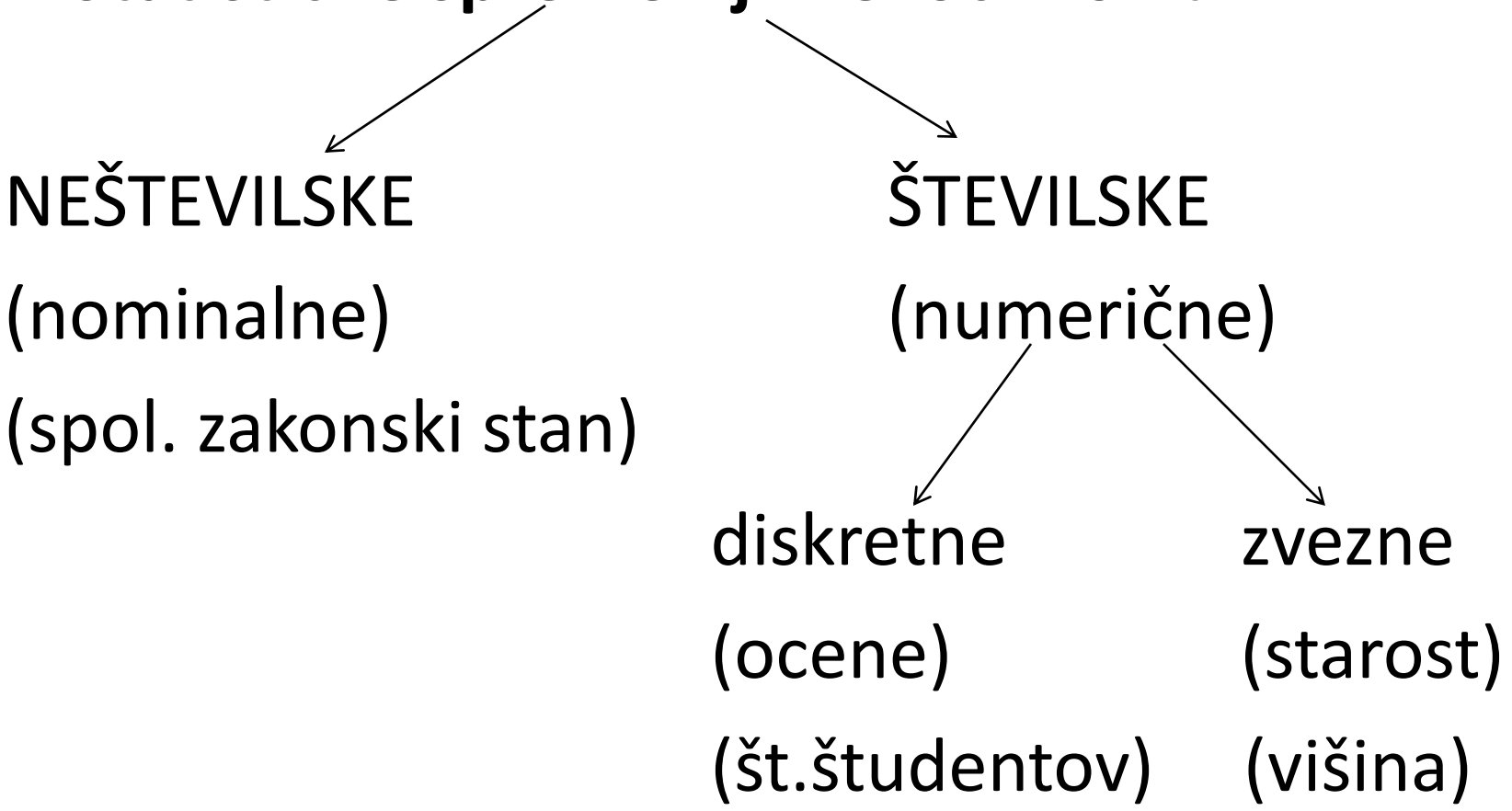
Zgledi statistični enot:

- Državljan RS v letu 2011
- Študent UL v študijskem letu 2010/11
- Študent FKKT UL v študijskem letu 2010/11
- Avtomobil, izdelan v Revozu v letu 2010

2. Opisne mere statistične populacije se imenujejo **statistični parametri** – to so značilnosti statistične populacije kot celote. Najpogostejši statistični parameter je **velikost populacije (oznaka N).** Poleg tega so še drugi – **povprečna vrednost** podatkov (če se da izračunati), **razpršenost ali disperzija** podatkov....

3. **Vzorec** – primerno, ponavadi naključno, izbrana končna podmnožica statistične populacije. Opisne mere vzorca so **cenilke** ali statistike. Le te uporabljamo pri statističnem sklepanju glede parametrov celotne populacije (intervali zaupanja, testiranje hipotez)
4. **Spremenljivke** – statistične enote opisujemo s statističnimi spremenljivkami, npr. starost, spol, ocena izdelka, višina osebnega dohodka....

- **Statistične spremenljivke ločimo** na:



Glede na **mersko lestvico** pa ločimo naslednje statistične spremenljivke:

- **Nominalne** (imenske) – spol, ime (nimajo nobene naravne urejenosti)
- **Ordinalne** (urejenostne) – izobrazba, kvalifikacija (imajo urejenost, vendar z njimi ne moremo računati)
- **Intervalne** – temperatura, letnica rojstva (so opisane s števili, vendar nimajo naravne ničelne točke,lahko jih odštevamo).
- **Razmernostne** – količina, velikost (z njimi lahko “poljubno” računamo, tudi razmerja)

- **Zgled:** Struktura prebivalce RS po spolu ob popisu leta 2002:

Spol	Št. Prebivalcev
Ženske	1005460
Moški	958576
Skupaj	**1964036**

- Populacija: množica prebivalcev RS ob popisu leta 2002
- Statistična enota: prebivalec RS ob popisu leta 2002
- Spremenljivka: spol (neštevilska, nominalna)
- Parametri populacije: velikost populacije N=1964036

Zgled: Po koncu tečaja angleškega jezika je profesor v namen ocene svojega dela pripravil vprašalnik, s katerim so tečajniki ocenili njegovo delo z ocenami od 1 do 5. Od 30 tečajnikov jih je na zadnji uri bilo 20 in 15 jih je oddalo izpolnjen vprašalnik.

Ocena	Število odgovorov
1	1
2	2
3	4
4	5
5	3
Skupaj	**15**

- Statistična populacija: skupina 30 tečajnikov
- Statistična enota: posamezen tečajnik
- Statistična spremenljivka: ocena tečaja (številska, diskretna, intervalna)
- Ali je vzorec slučajen: NE!
- Lahko izračunamo povprečno oceno:

$$\frac{(1x1 + 2x2 + 4x3 + 5x4 + 3x5)}{15}$$

in dobimo približno 3,47.

Zgled: Pri anketi o branju neke revije so o bralcih te revije pridobili naslednje tipe podatkov:

- Starost
- Spol
- Zakonski stan
- Letni dohodek
- Ocena revije

Statistična populacija: bralci revije

Statistična enota: en bralec/bralka revije

Statistične spremenljivke:

- Starost-številska, zvezna, razmernostna
- Spol-neštevilska, nominalna
- Zakonski stan-neštevilska, nominalna
- Letni dohodek-številska, zvezna, razmernostna
- Ocena revije-številska, diskretna, intervalna

PREDSTAVITEV PODATKOV

Če je število vrednosti statistične spremenljivke veliko (t.j. imamo veliko podatkov), te vrednosti običajno razdelimo v **frekvenčne razrede**.

Absolutna frekvenca (oznaka f_k za k-ti razred) je število vrednosti statistične spremenljivke v k-tem razredu.

Relativna frekvenca (oznaka f_k') pa je delež absolutne frekvence f_k glede na celoto. Če je N število enot v populaciji ali morda vzorcu, je

$$f_k' = \frac{f_k}{N}$$

Zgled: V podjetju so med 30 zaposlenimi izvedli anketo o zadovoljstvu z delovnimi pogoji. Vsak delavec je z oceno od 1 do 5 ocenil svoje delovne pogoje. Rezultati ankete so bili:

1,2,4,3,2,3,2,4,4,5,3,4,1,3,4,3,2,1,1,4,3,4,
2,3,4,1,5,2,4,3

Rezultate ankete lahko predstavimo v tabeli frekvenc in relativnih frekvenc.

Anketa v podjetju

Ocena	Frekvenca	Relativna frekvenca
1	5	5/30 oziroma 16,7%
2	6	6/30 oziroma 20%
3	8	8/30 oziroma 26,7%
4	9	9/30 oziroma 30%
5	2	2/30 oziroma 6,7%
Skupaj	**30**	**1 oziroma 100%**

Zgled: V Kamniško-Savinjski Alpah so planinske postojanke na naslednjih nadmorskih višinah

(podatki PZS) : 1086, 1793, 1526, 1453, 1375, 837, 1123, 600, 1534, 1396, 1864, 1808, 1478, 1460, 1208, 1471, 1534, 1548, 444, 1700, 434, 1356, 961, 725, 1491, 1534 (v metrih).

Razvrstite podatke v frekvenčne razrede širine 200m (npr. (1200,1400]). Postojanke pod 1000m postavi v en frekvenčni razred.

Planinske postojanke

Višina	Frekvenca
0-1000	6
1000-1200	2
1200-1400	4
1400-1600	10
1600-1800	2
1800-2000	2
Skupaj	**26**

Zgled: Podatki o telefonskih računih za 200 novih naročnikov:

42.19 38.45 29.23 89.35 118.04 110.46 **0** 72.88 83.05 95.73 103.15
94.52 26.84 93.93 90.26 72.78 101.36 104.8 74.01 56.01 39.21 48.54
93.31 104.88 30.61 22.57 63.7 104.84 6.45 16.47 89.5 13.36 44.16
92.97 99.56 92.62 78.89 87.71 93.57 0 75.71 88.62 99.5 85 0 8.41
70.48 92.88 3.2 115.5 2.42 1.08 76.69 13.62 88.51 55.99 12.24
119.63 23.31 11.05 8.37 7.18 11.07 1.47 26.4 13.26 21.13 95.03
29.04 5.42 77.21 72.47 0 5.64 6.48 6.95 19.6 8.11 9.01 84.77 1.62
91.1 10.88 30.62 100.05 26.97 15.43 29.25 1.88 16.44 109.08 2.45
21.97 17.12 19.7 6.93 10.05 99.03 29.24 15.21 28.77 9.12 118.75 0
13.95 14.34 79.52 2.72 9.63 21.34 104.4 2.88 65.9 20.55 3.43 10.44
21.36 24.42 95.52 6.72 35.32 117.69 106.84 8.4 90.04 3.85 91.56
10.13 5.72 33.69 115.78 0.98 19.45 0 27.21 89.27 14.49 92.17 21
106.59 13.9 9.22 109.94 10.7 0 11.27 72.02 7.74 5.04 33.4 6.95 6.48
11.64 83.26 15.42 24.49 89.13 111.14 92.64 53.9 114.67 27.57 64.78
45.81 56.04 20.39 31.77 94.67 44.32 3.69 19.34 13.54 18.89 1.57 0
5.2 2.8 5.1 3.03 9.16 15.3 75.49 68.69 35 9.12 18.49 84.12 13.68
20.84 100.04 112.94 20.12 53.21 15.3 49.24 9.44 2.67 4.69 41.38
45.77

Sedaj je podatkov bistveno več in so tudi veliko bolj nepregledni. Razdelimo jih v frekvenčne razrede.

Na **koliko razredov** naj bi razdelili te podatke?

Približno pravilo je naslednje:

Število podatkov	Število razredov
Manj kot 50	5-7
50-200	7-9
20-500	9-10
500-1000	10-11
1000-5000	11-13
5000-50000	13-17
Več kot 50000	17-20

Lahko pa se uporabi Sturgesova formula:

Število razredov = $1 + 3.3 \log_{10} N$

Odločimo se za 8 razredov (po formuli dobimo približno 8,59). Koliko bo **širina razredov**? Le to izračunamo po formuli:

Širina razredov = (Max. – Min.)/Število razredov

V našem primeru dobimo

Širina = (119,63 – 0)/8 ≈ 14,95

Odločimo se za širino 15. Tako dobimo naslednje razrede: [0,15], (15,30],(30,45],...(105,120]

in tabelo:

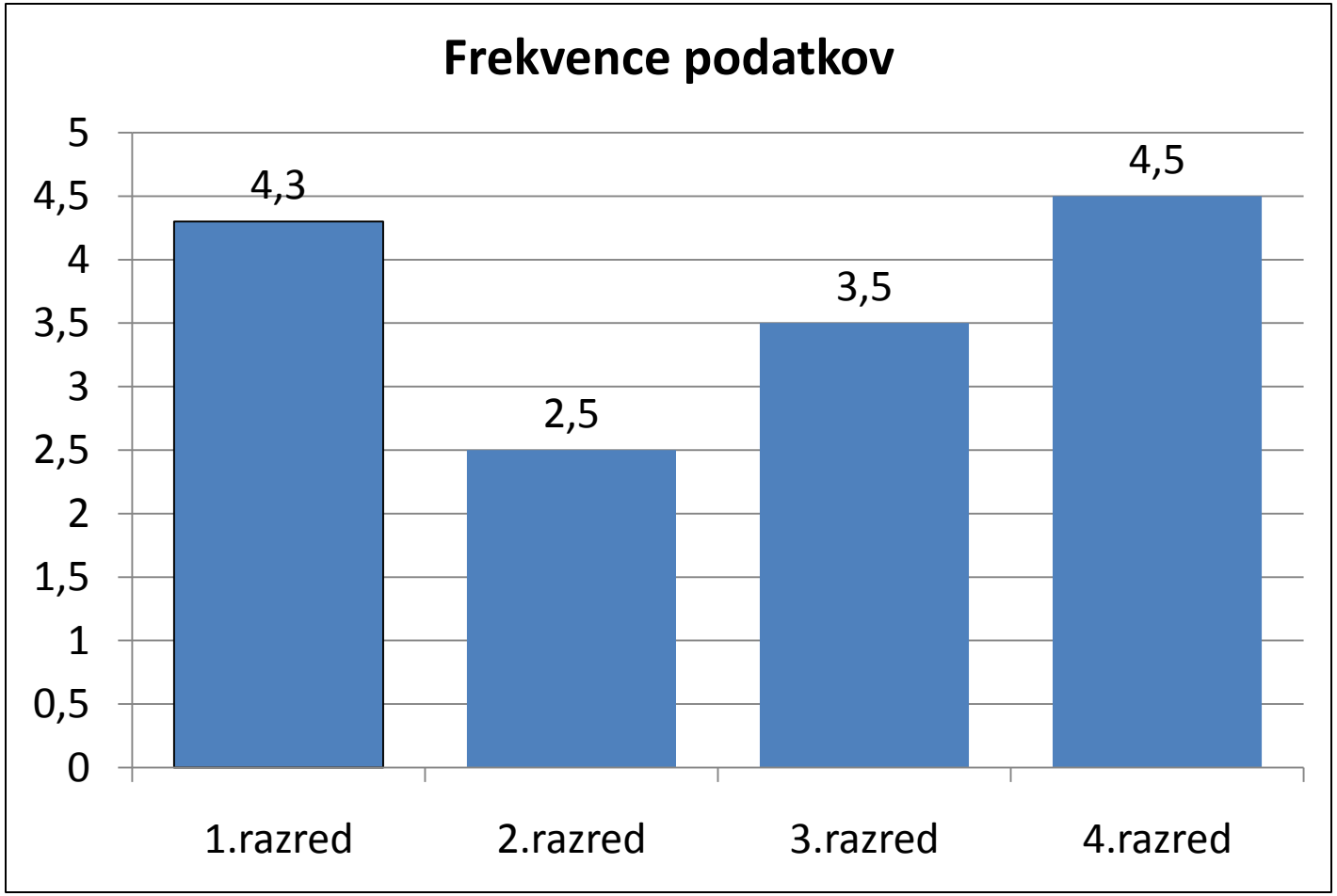
Frekvence računov

Razred računov	Frekvenca
0-15	71
15-30	37
30-45	13
45-60	9
60-75	10
75-90	18
90-105	28
105-120	14
Skupaj	**200**

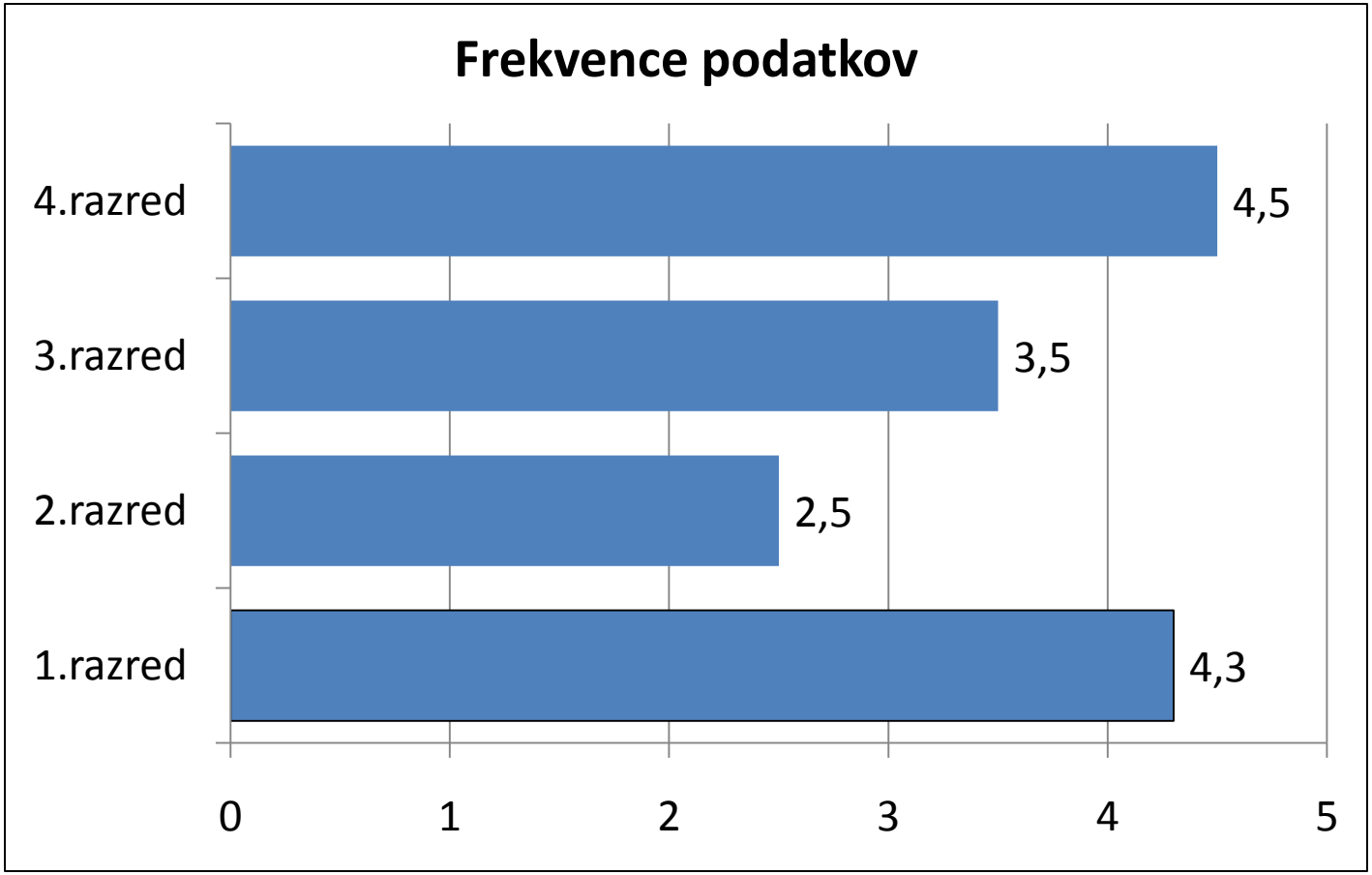
GRAFIČNE PREDSTAVITVE PODATKOV

Več kot frekvenčne tabele pa nam povedo grafične predstavitve podatkov, t.j. prikazovanje podatkov grafično. Osnovni načini so:

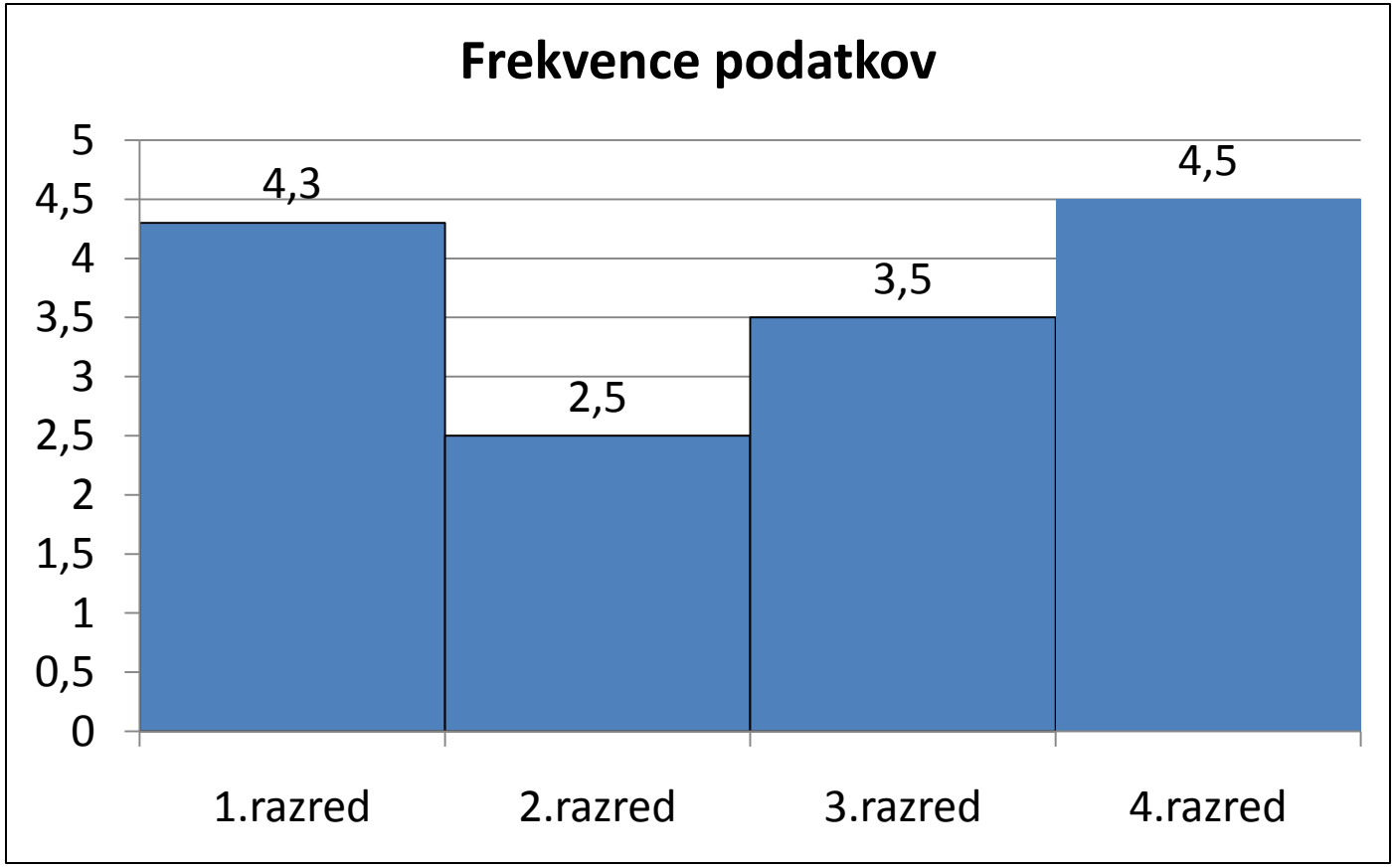
- Stolpčni diagram – pokončni (nominalna ali diskretna spr.)



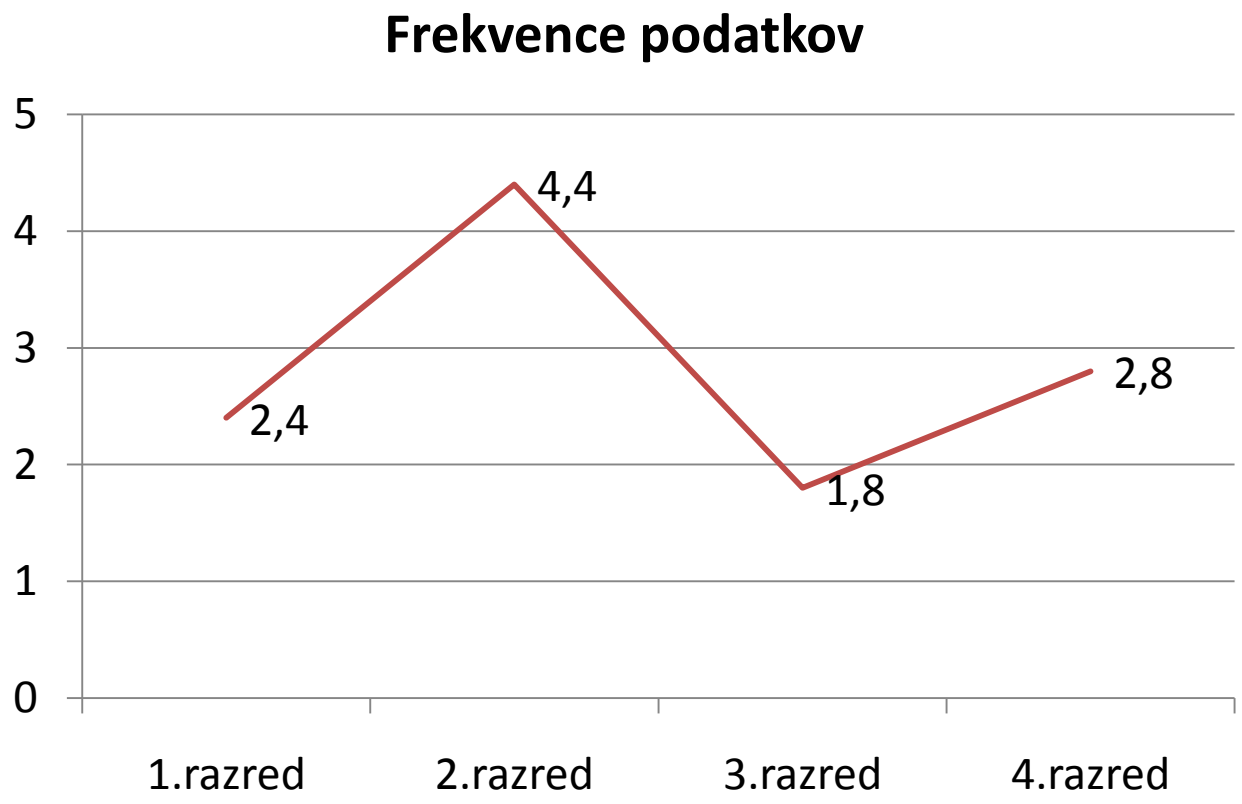
- Stolpčni diagram-ležeči (nominalna ali diskretna spremenljivka)



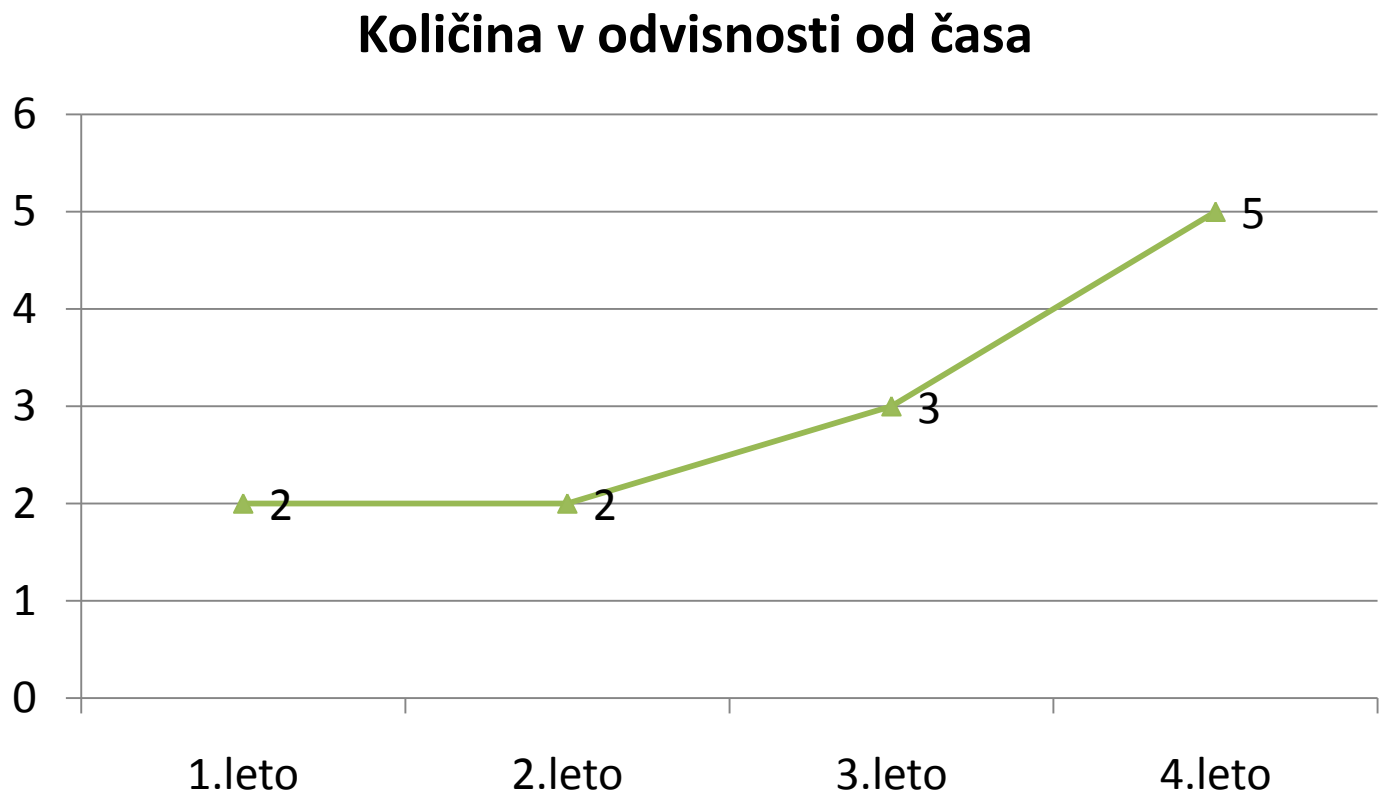
- Histogram (številska intervalna spr.)



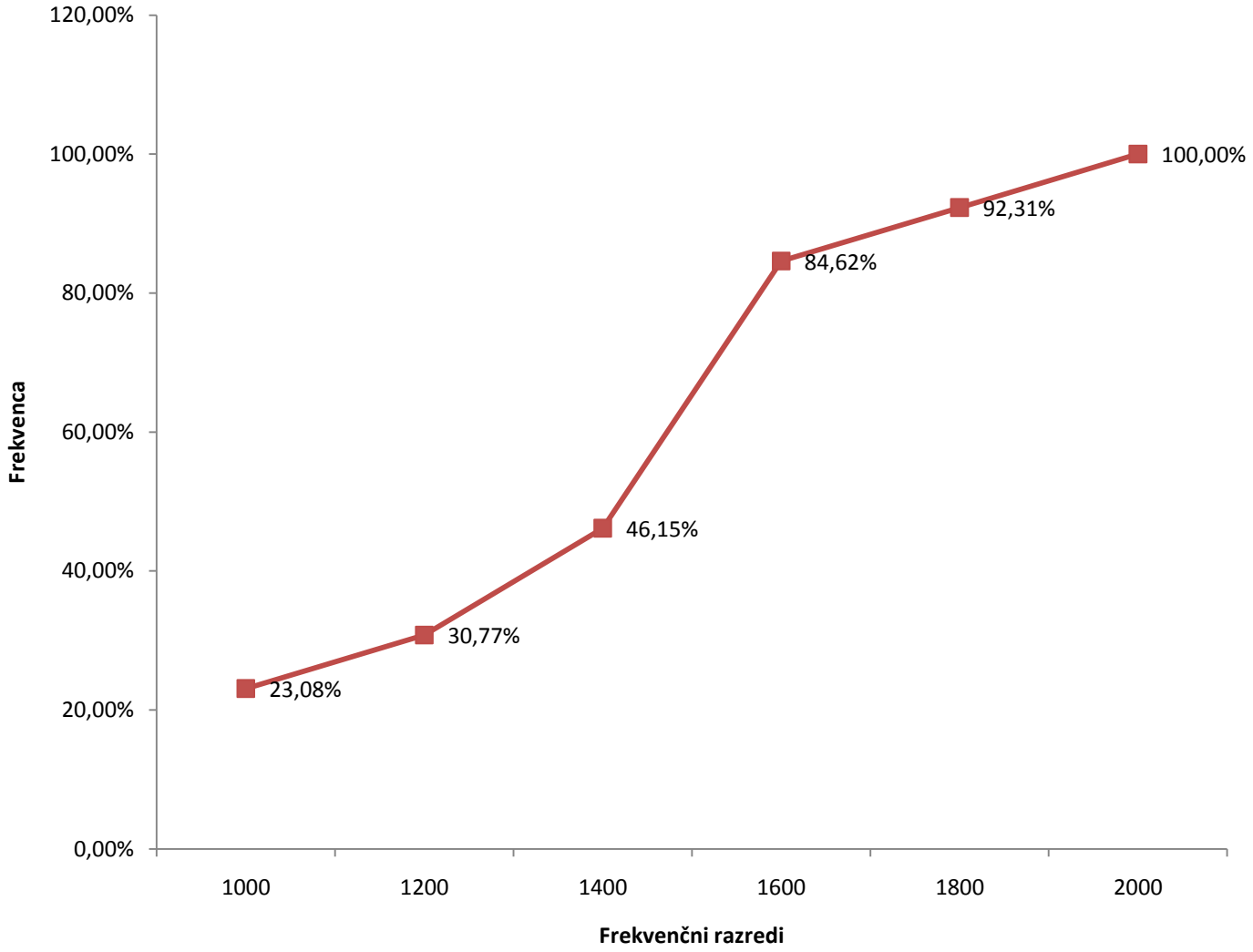
- Linijski diagram (npr. frekvenčni poligon):



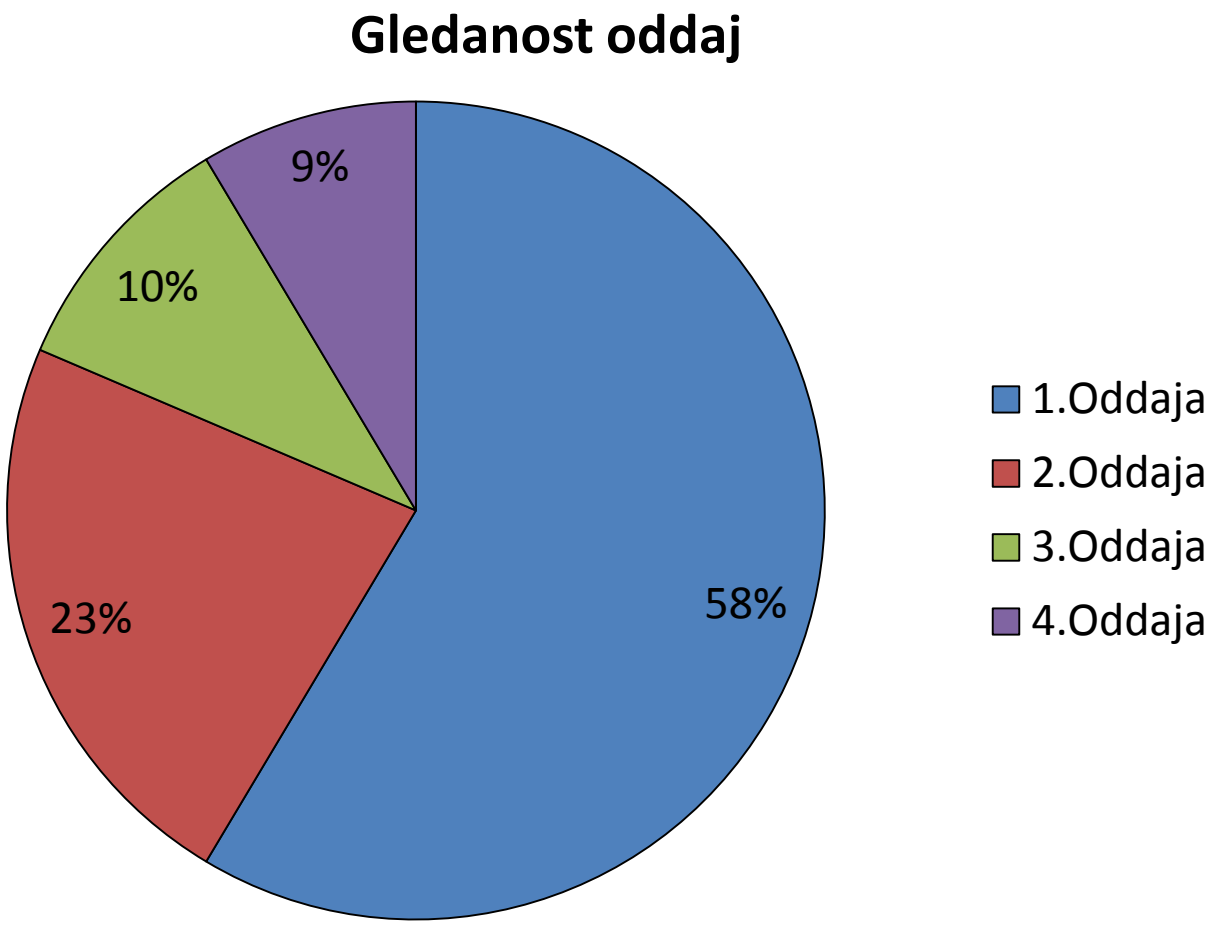
- Linijski diagram (časovne vrste – odvisnost vrednosti spremenljivke od časa):



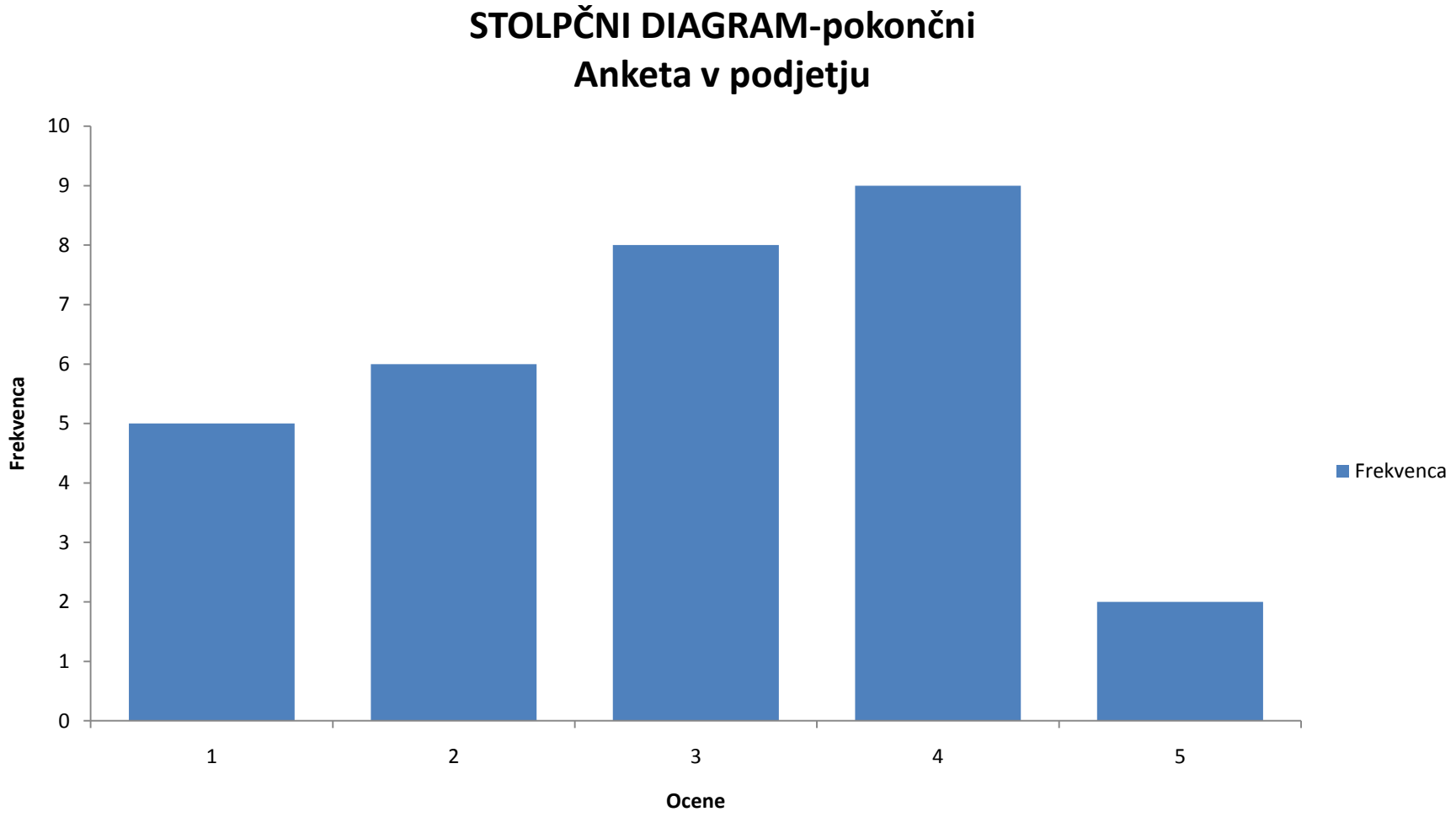
- Ogive (kumulativna relativna frekvenca):

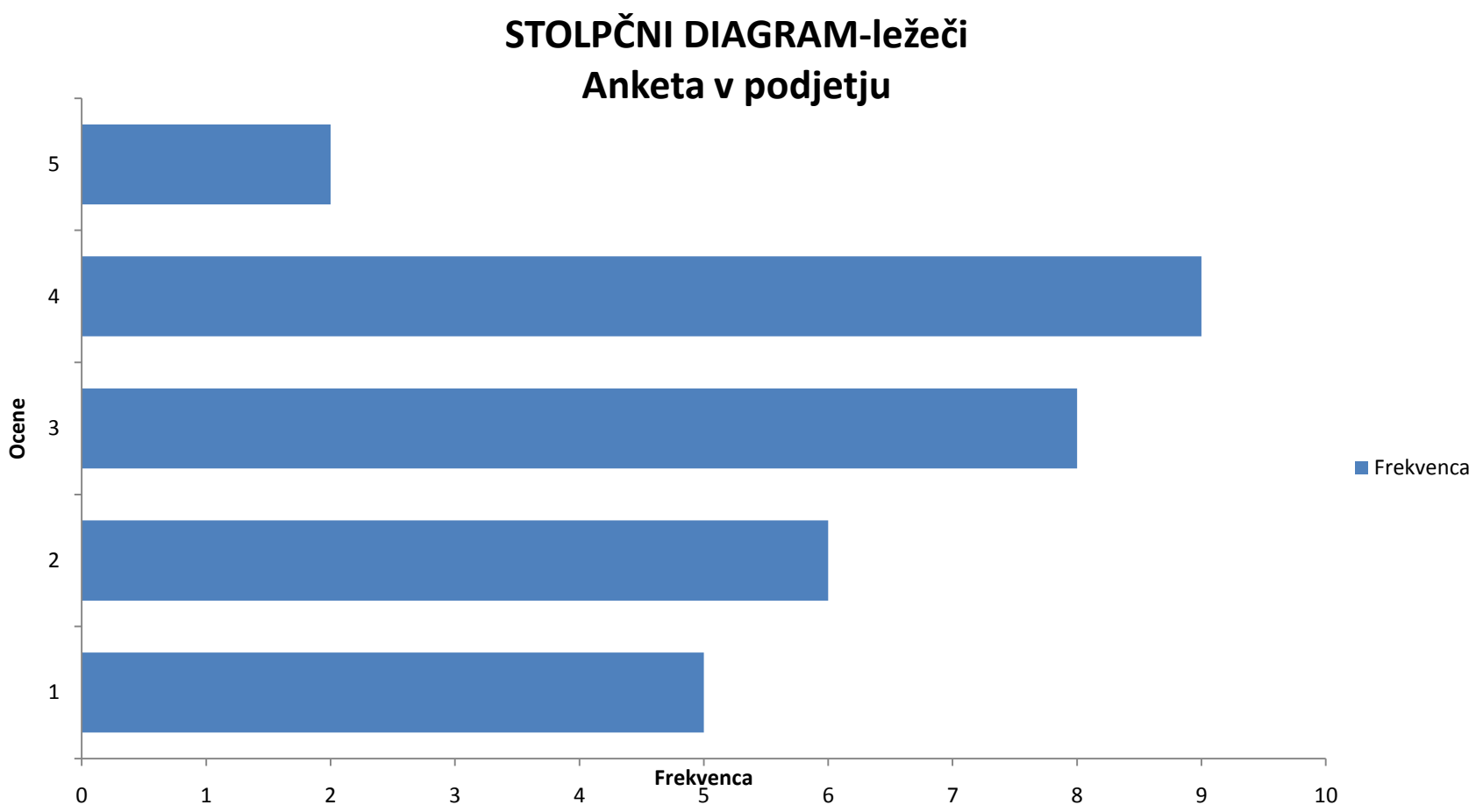


- Strukturni krog:

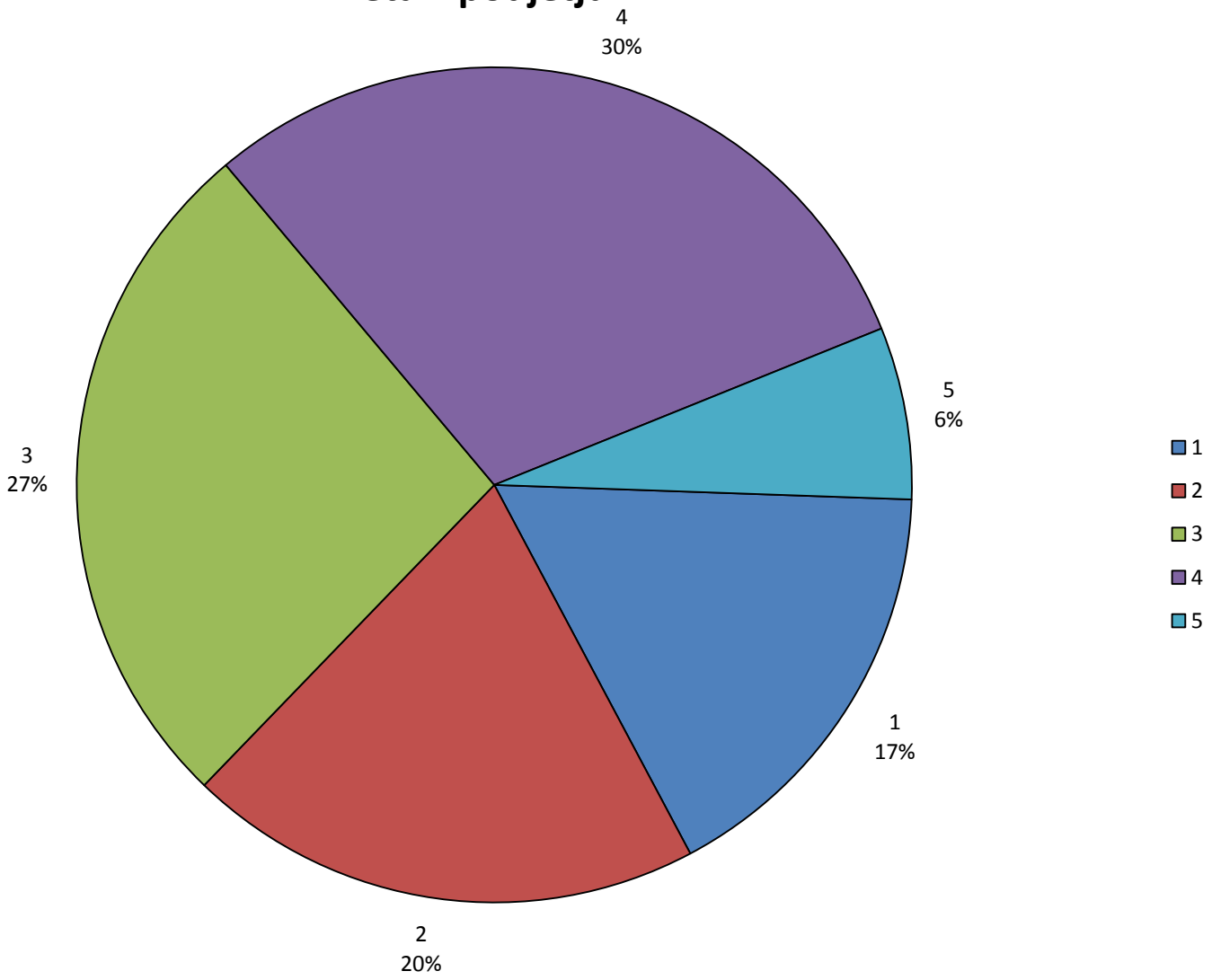


- Zgled: Ocene o zadovoljstvu v podjetju

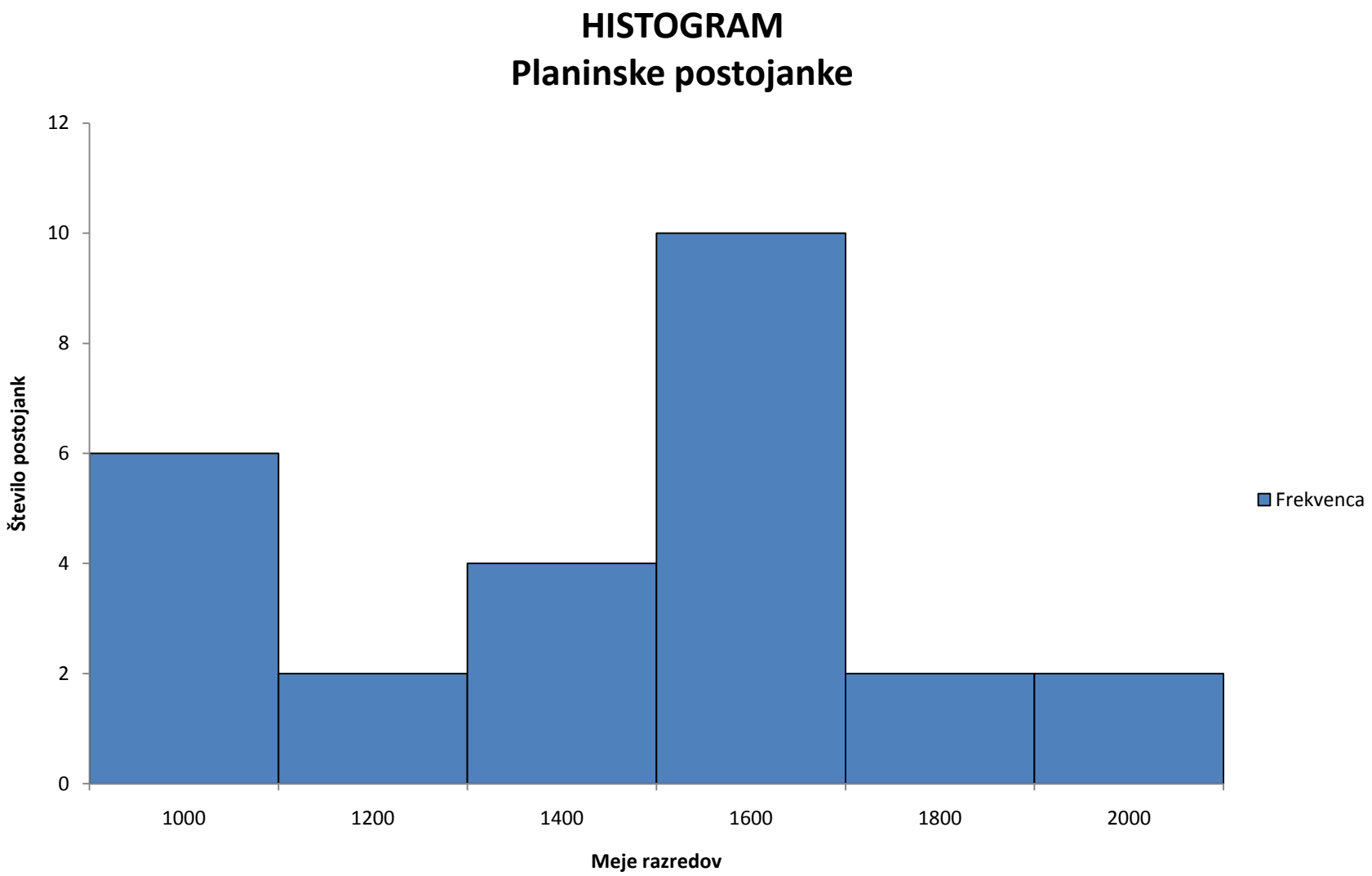



STOLPČNI DIAGRAM-ležeči
Anketa v podjetju
Ocene
5
4
3
2
1
0
1
2
3
4
5
6
7
8
9
10
Frekvenca
Frekvenca

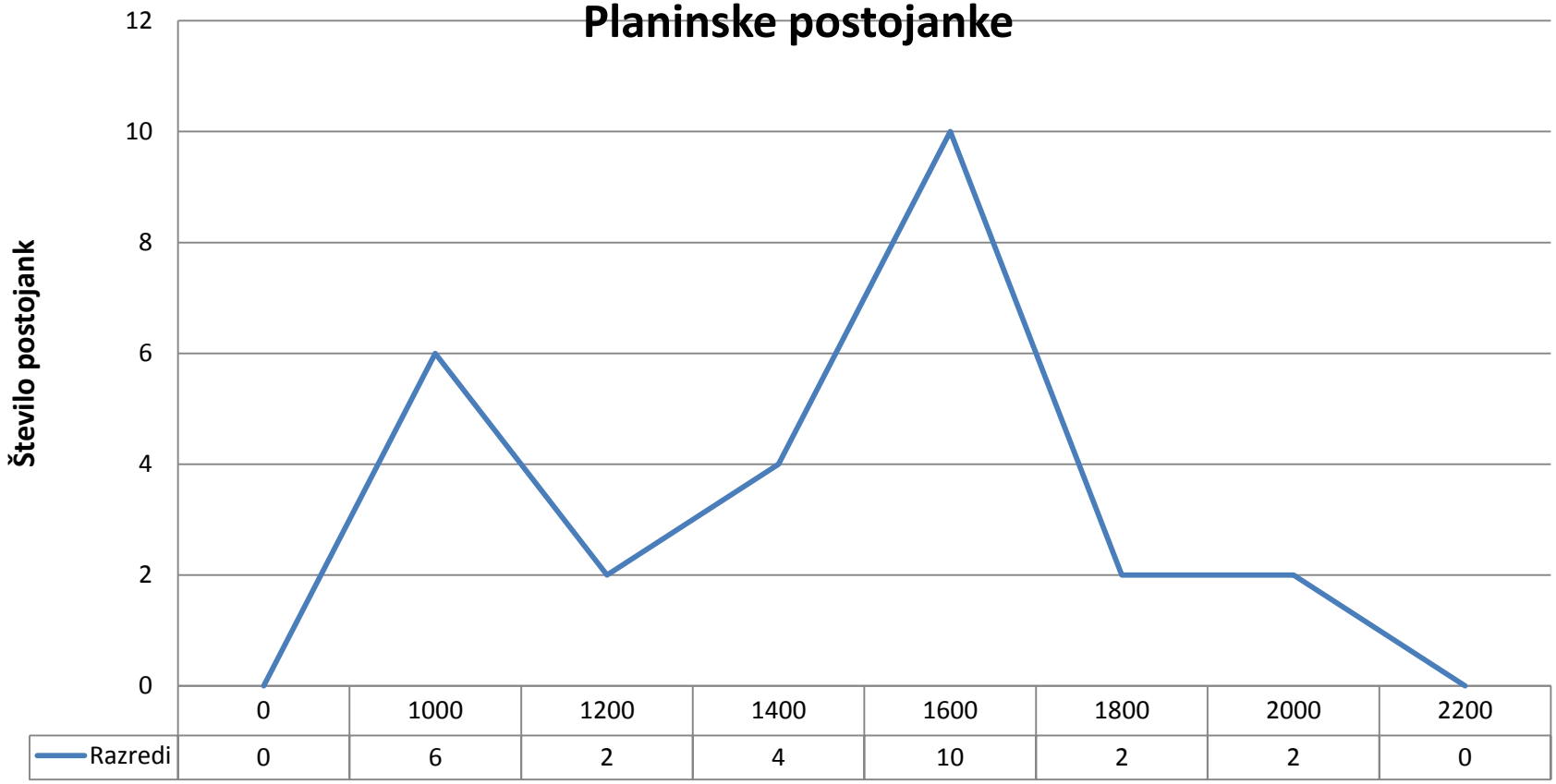
STRUKTURNI KROG
Anketa v podjetju



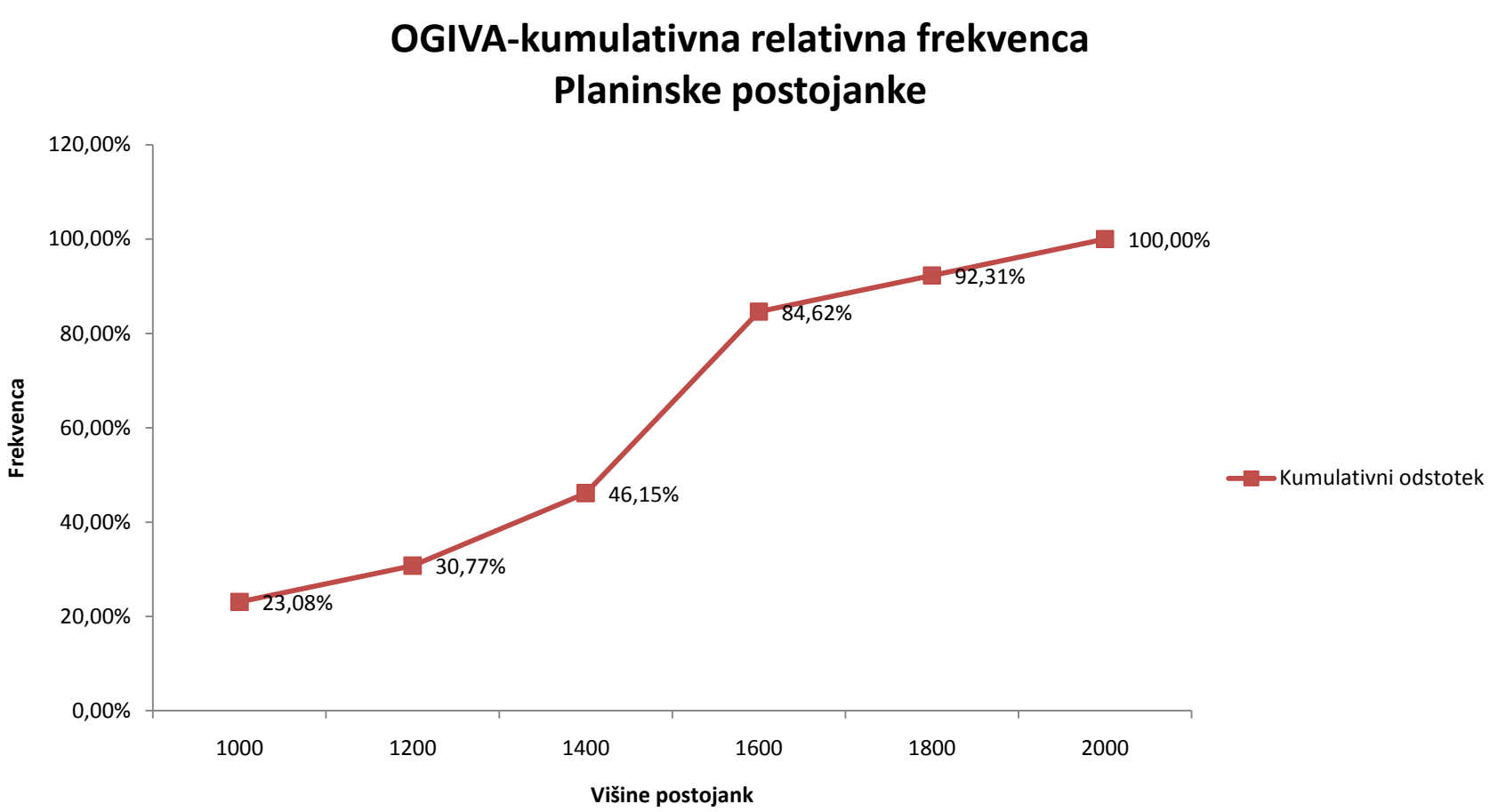
- Zgled: Planinske postojanke



LINIJSKI DIAGRAM
Planinske postojanke

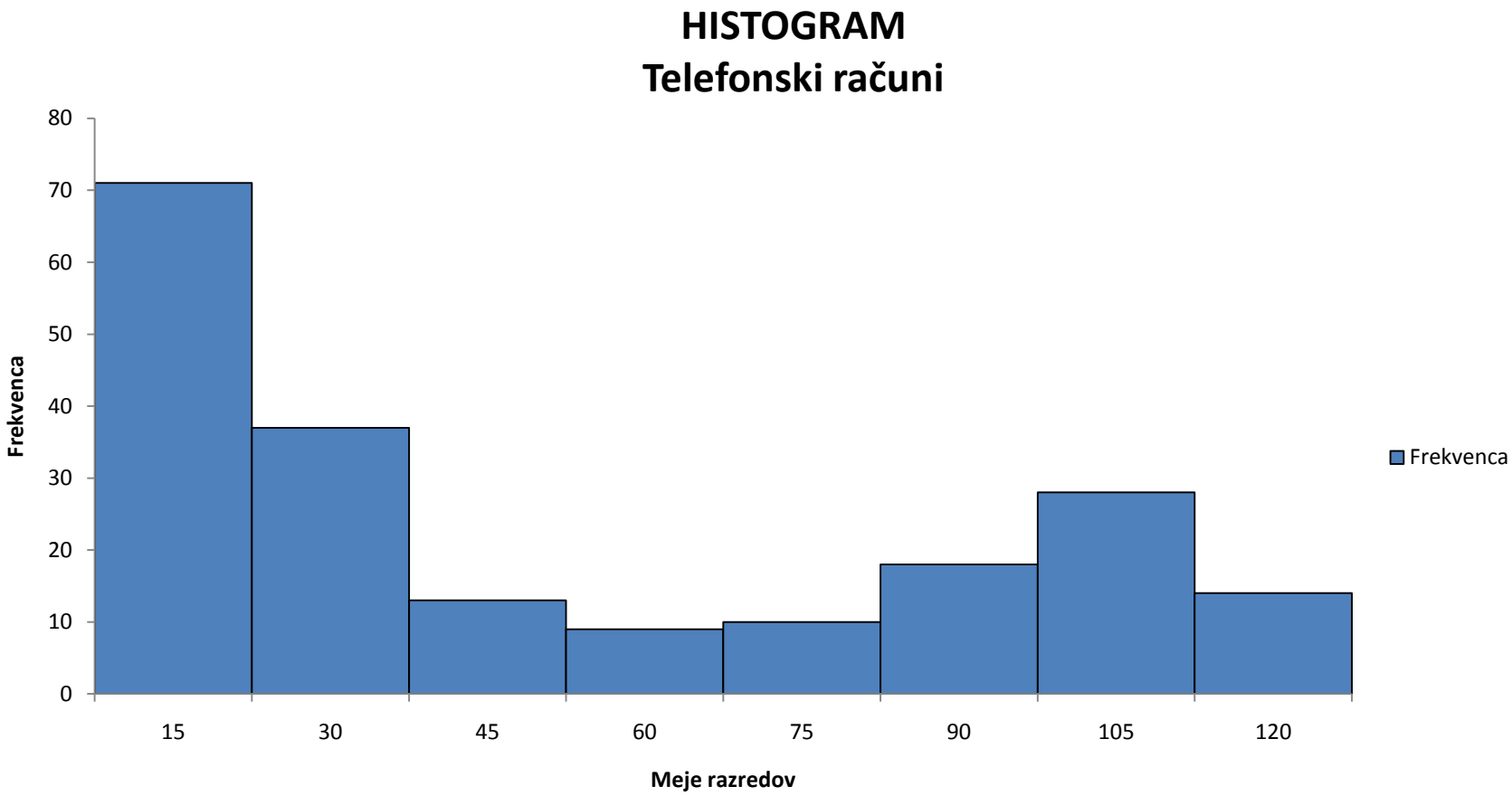


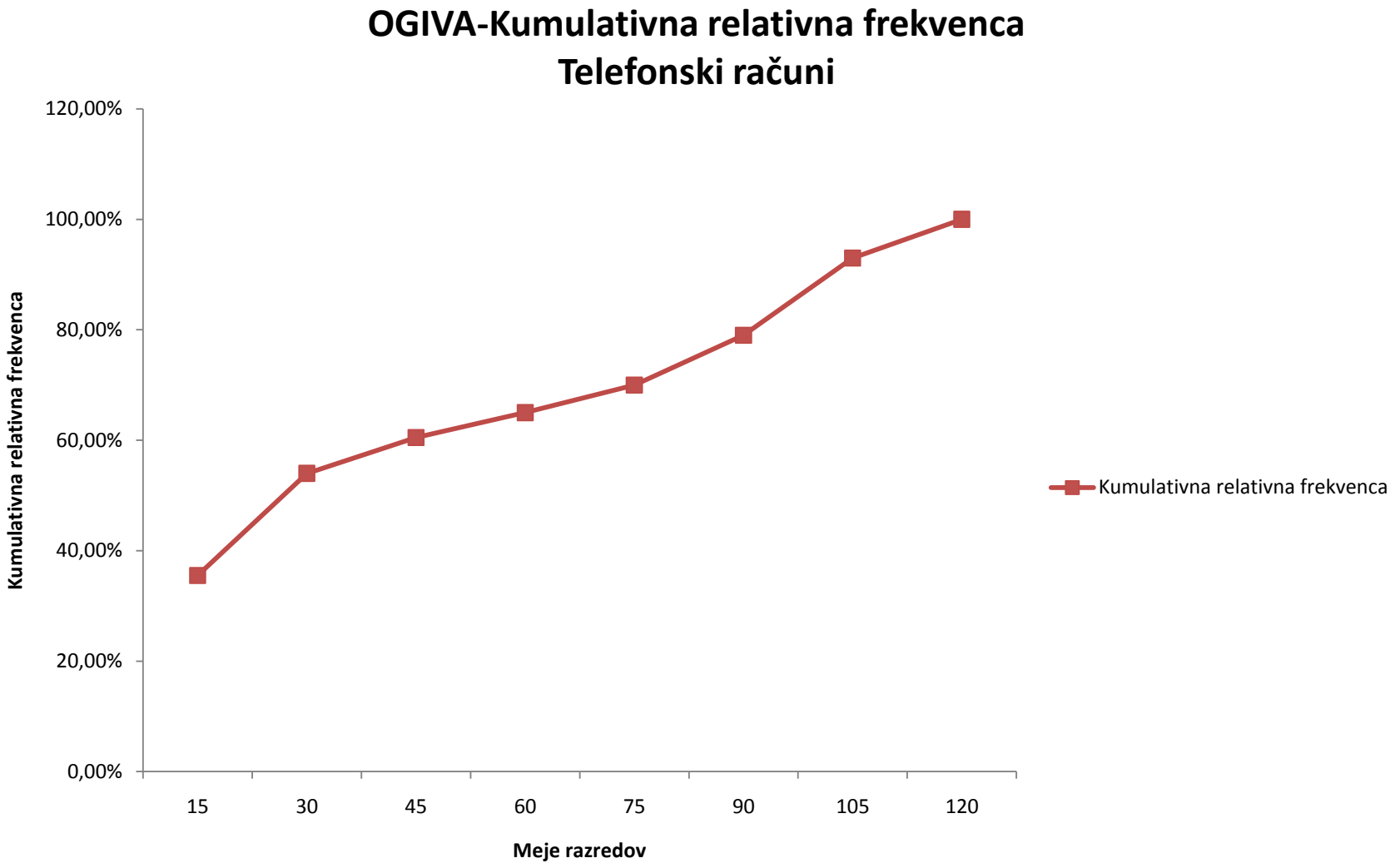
	0	1000	1200	1400	1600	1800	2000	2200
Razredi	0	6	2	4	10	2	2	0


OGIVA-kumulativna relativna frekvenca
Planinske postojanke
120,00%
100,00%
80,00%
60,00%
40,00%
20,00%
0,00%
Frekvenca
23,08%
30,77%
46,15%
84,62%
92,31%
100,00%
1000
1200
1400
1600
1800
2000
Višine postojank
Kumulativni odstotek

- Zgled: Podatki o telefonskih računih

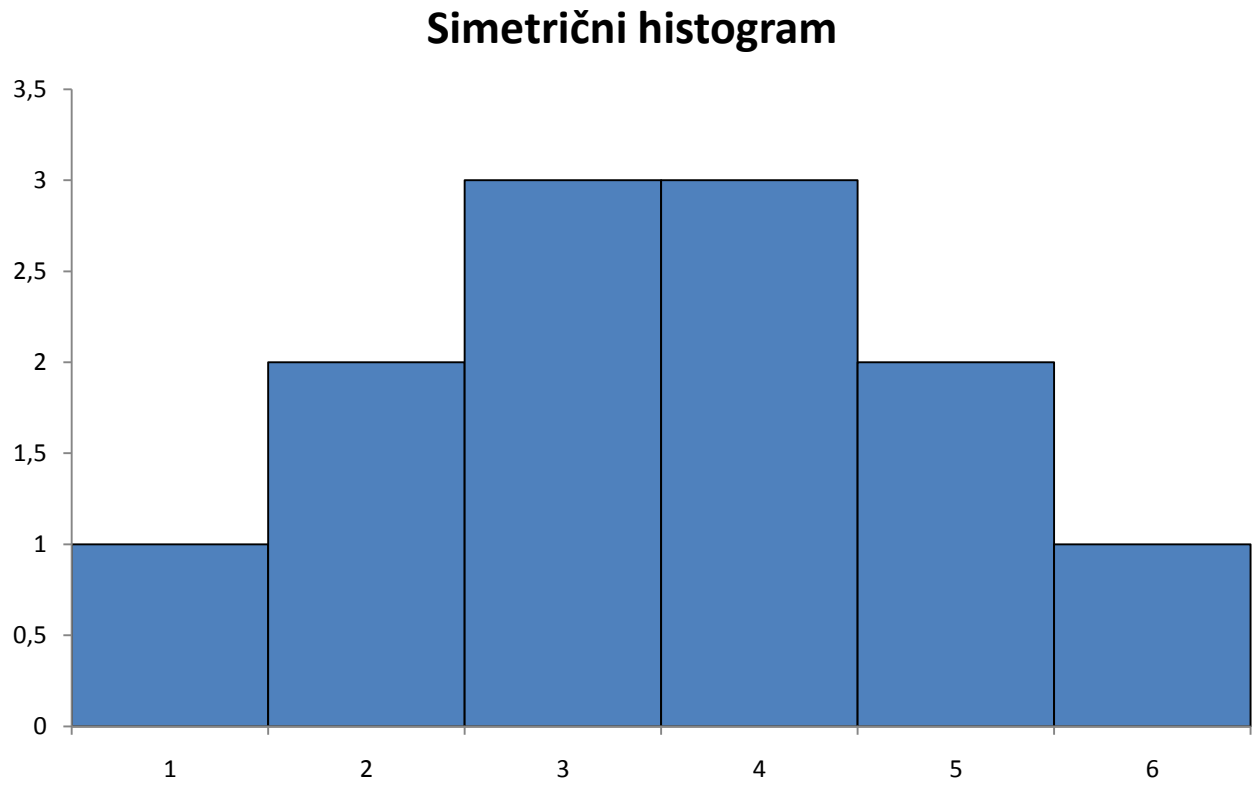
Meje razredov	*Frekvenca*	*Kumulativni %*
15	71	35,50%
30	37	54,00%
45	13	60,50%
60	9	65,00%
75	10	70,00%
90	18	79,00%
105	28	93,00%
120	14	100,00%


HISTOGRAM
Telefonski računi
Frekvenca
80
70
60
50
40
30
20
10
0
15
30
45
60
75
90
105
120
Meje razredov
Frekvenca

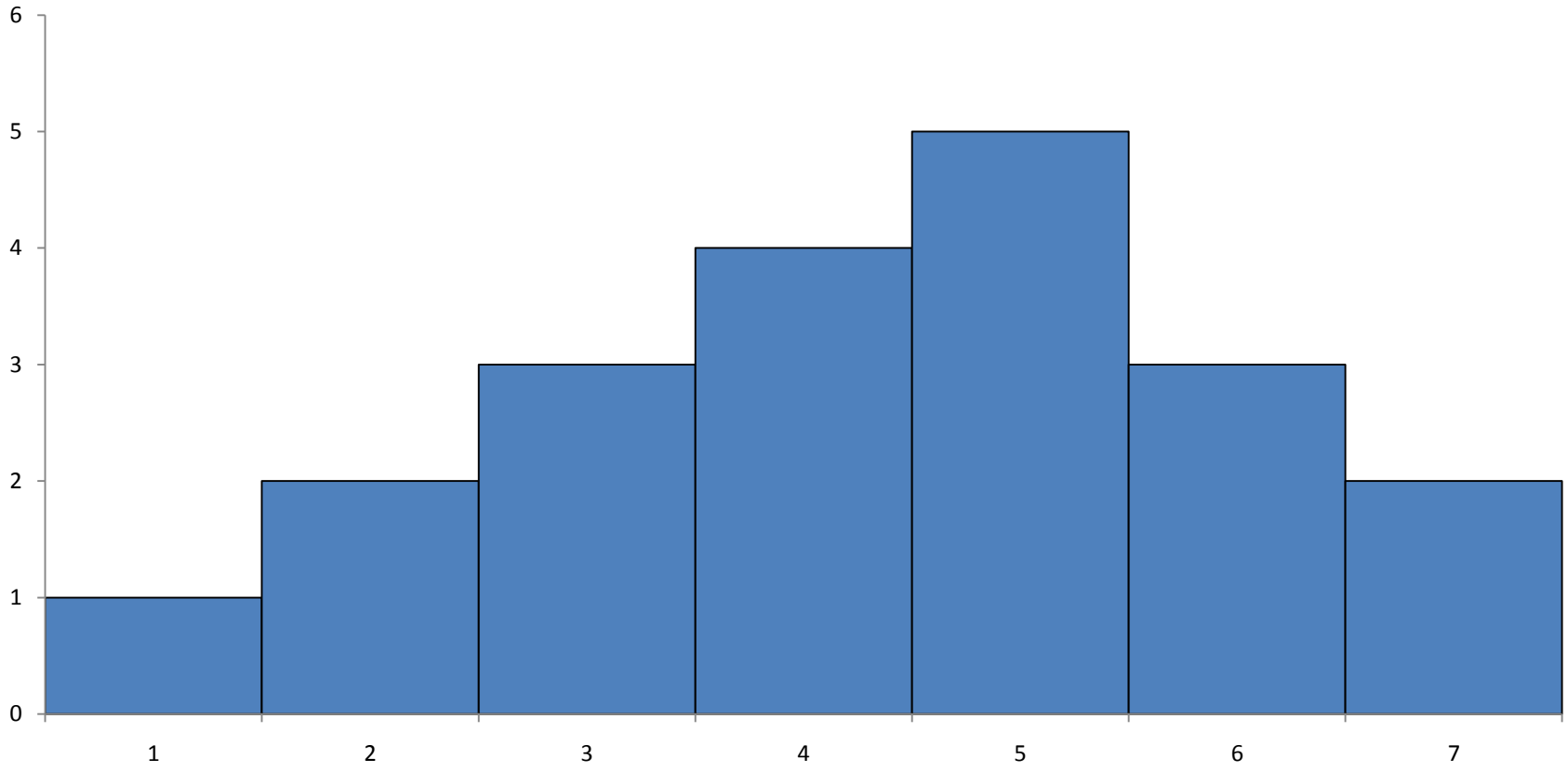

OGIVA-Kumulativna relativna frekvenca
Telefonski računi
120,00%
100,00%
80,00%
60,00%
40,00%
20,00%
0,00%
Kumulativna relativna frekvenca
15
30
45
60
75
90
105
120
Meje razredov
Kumulativna relativna frekvenca

Oblike histogramov

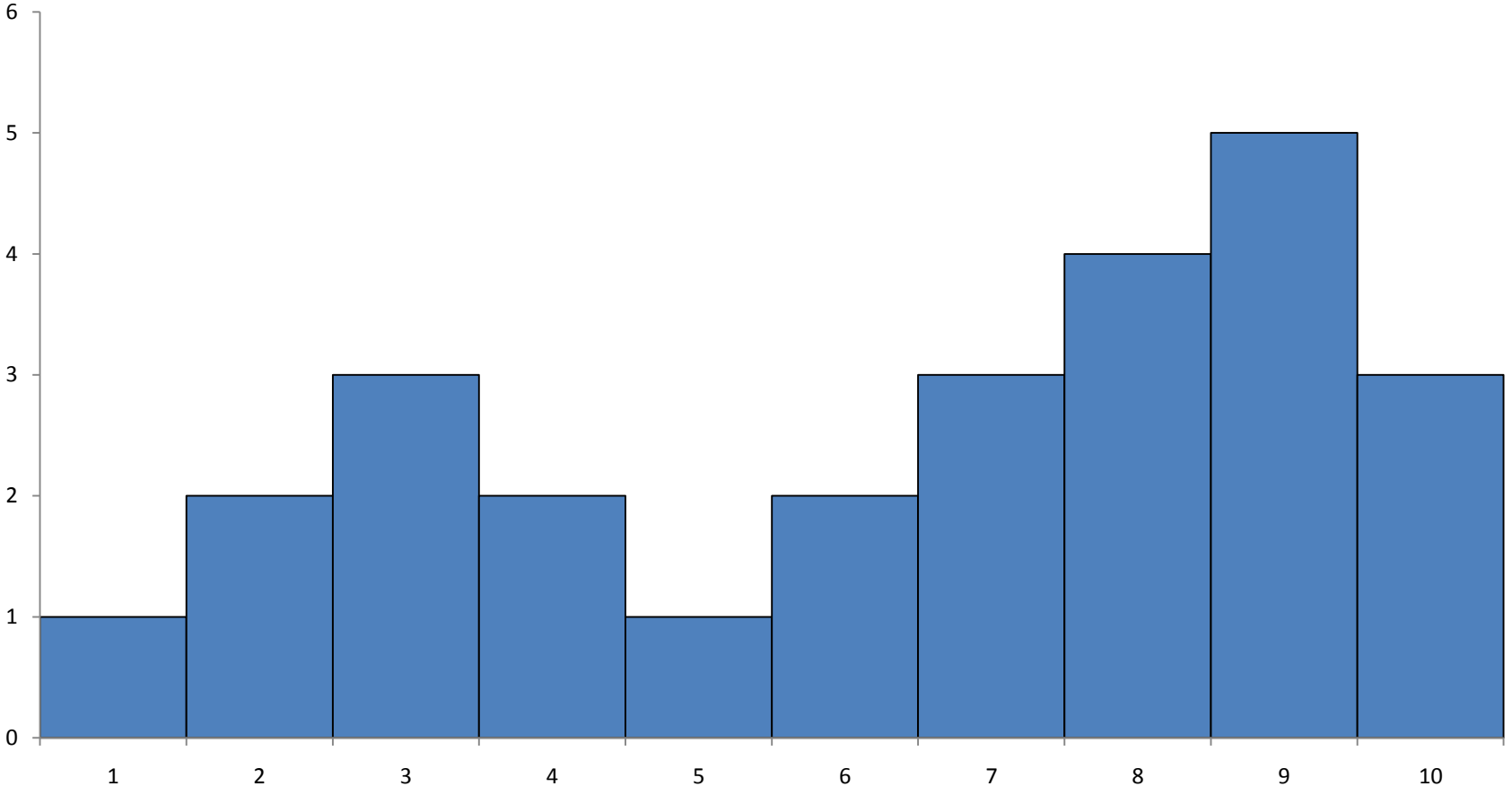
Nekatere značilne oblike histogramov imajo svoja imena:



Unimodalni histogram

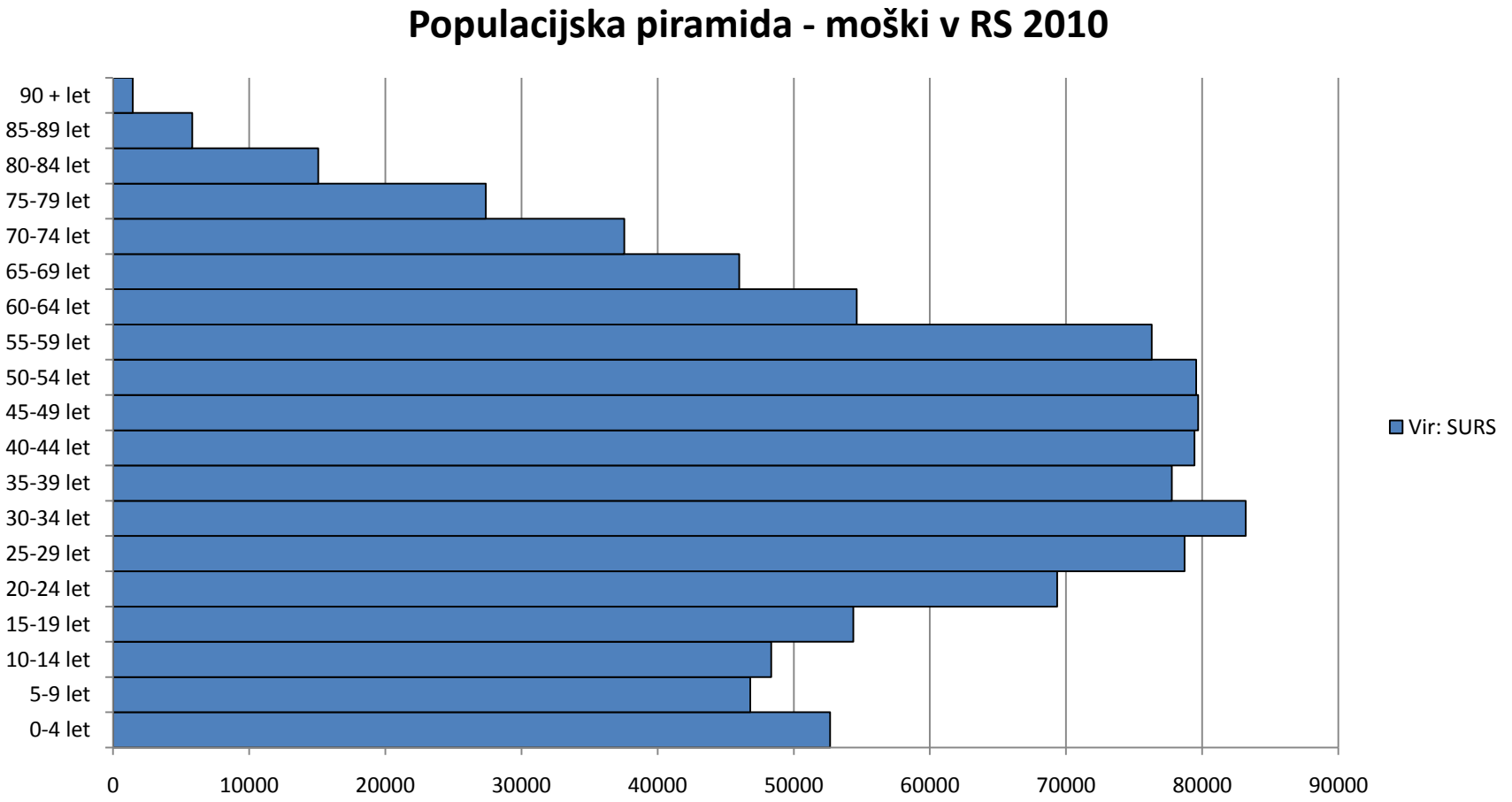


Bimodalni histogram



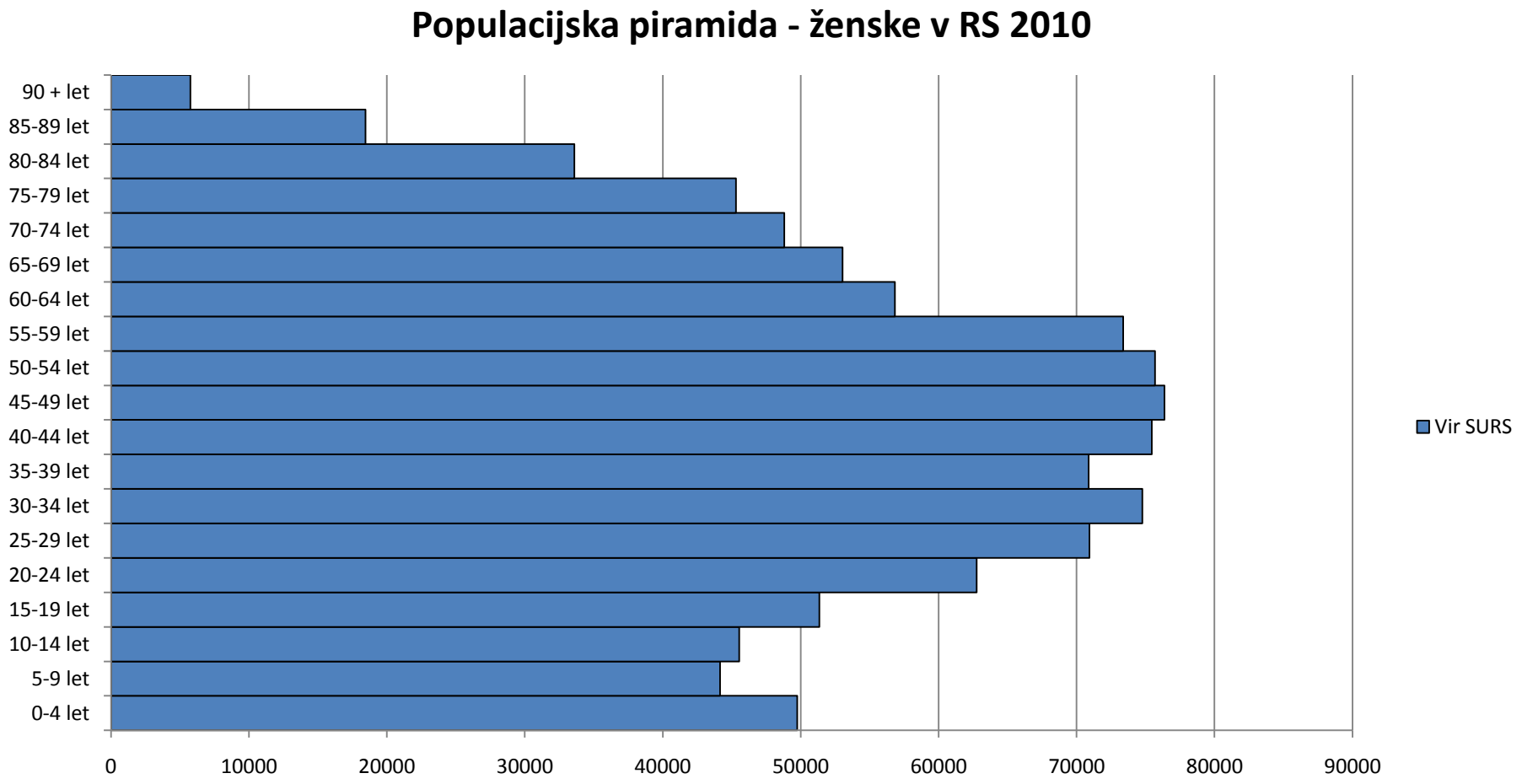
- Zgled: Populacija moških v RS 2010 - vir SURS

0-4 let	5-9 let	10-14 let	15-19 let	20-24 let	25-29 let	30-34 let	35-39 let	40-44 let	45-49 let	50-54 let	55-59 let	60-64 let	65-69 let	70-74 let	75-79 let	80-84 let	85-89 let	90 + let
52666	46802	48340	54376	69356	78708	83189	77763	79426	79697	79546	76299	54632	45999	37556	27388	15077	5828	1459



- Zgled: Populacija žensk v RS 2010 – vir SURS

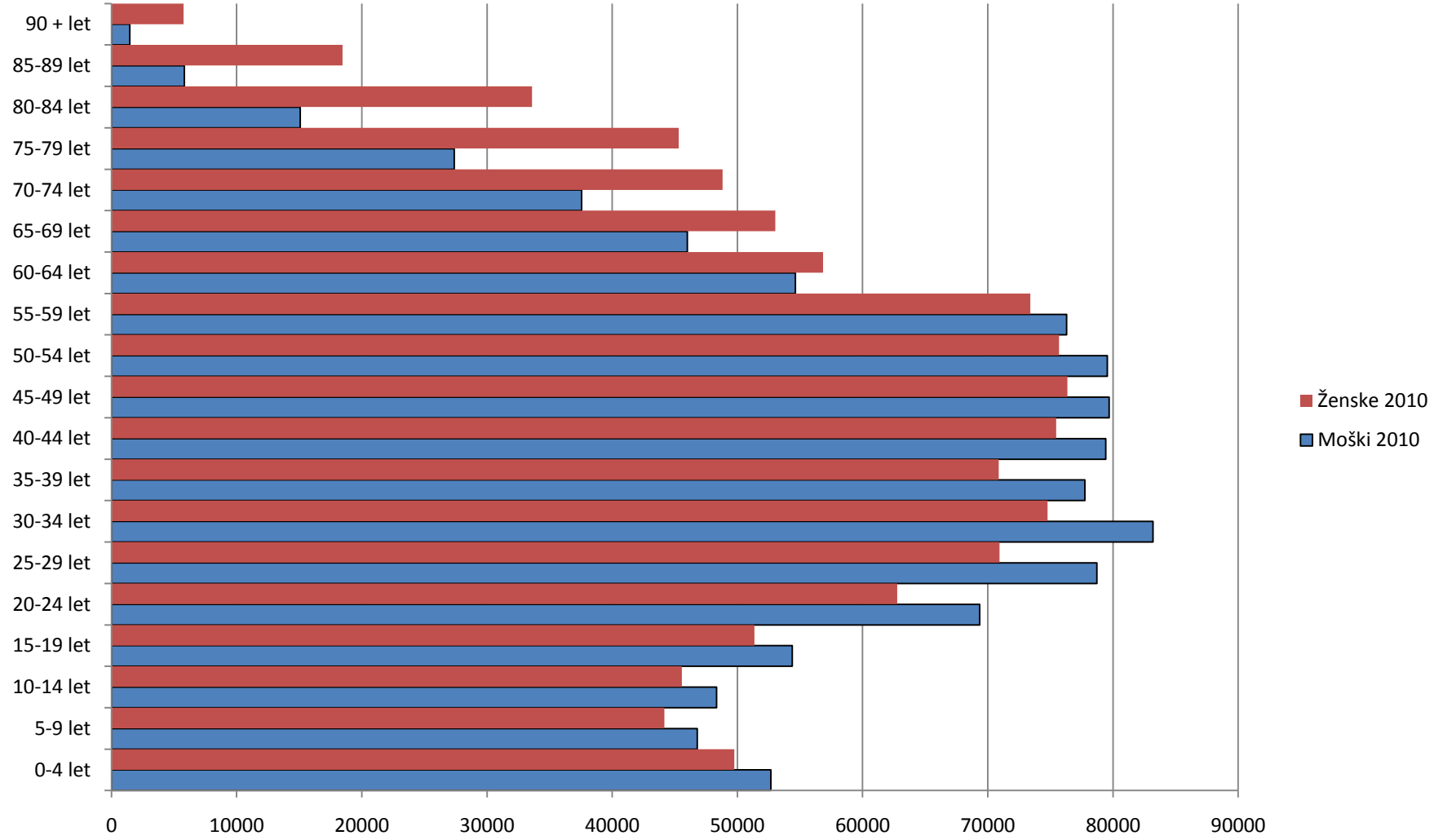
0-4 let	5-9 let	10-14 let	15-19 let	20-24 let	25-29 let	30-34 let	35-39 let	40-44 let	45-49 let	50-54 let	55-59 let	60-64 let	65-69 let	70-74 let	75-79 let	80-84 let	85-89 let	90 + let
49752	44162	45553	51359	62763	70931	74777	70876	75461	76365	75689	73389	56834	53033	48812	45307	33591	18457	5758



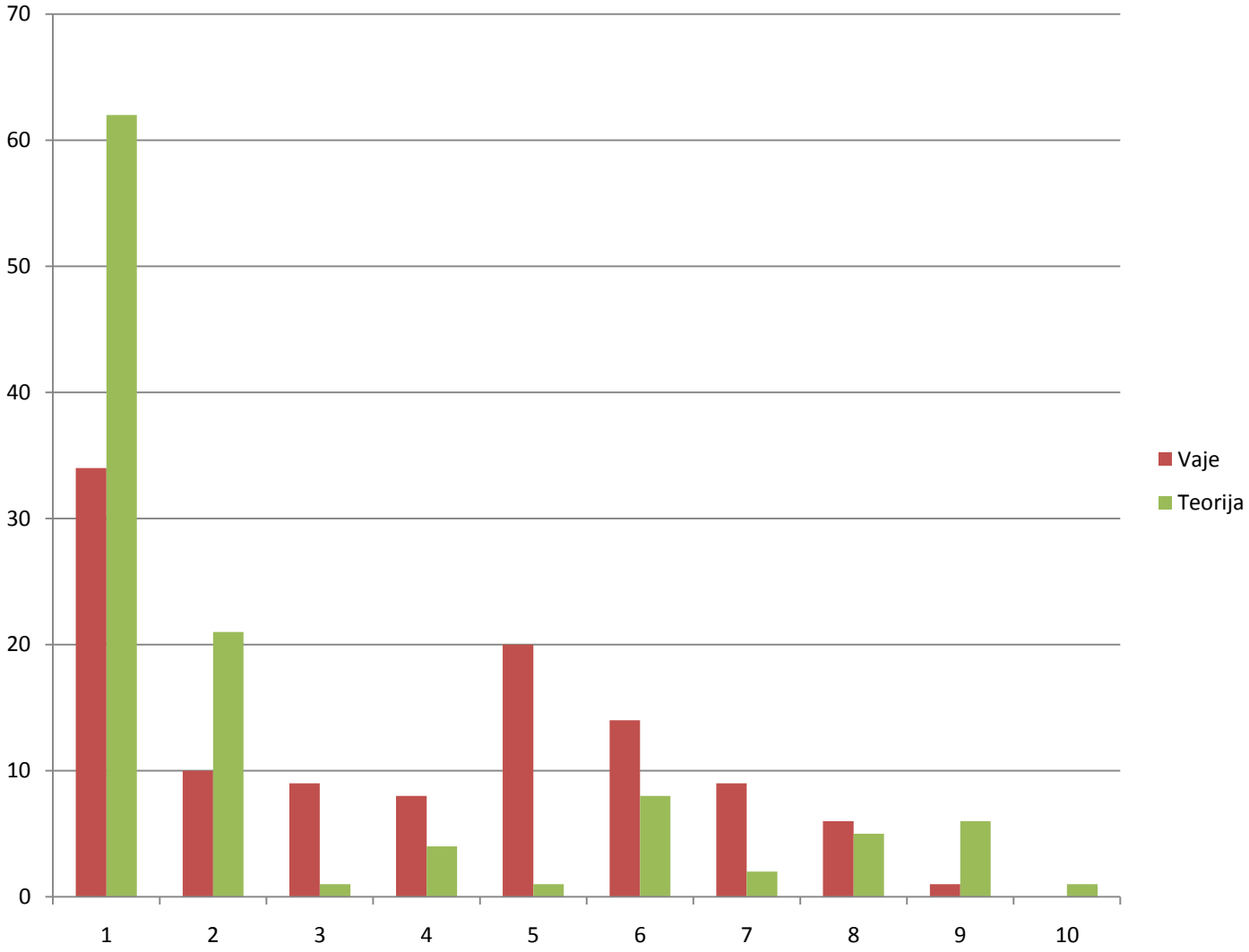
Če pa hočemo na istem grafikonu prikazati porazdelitev dveh ali celo več spremenljivk z enakimi vrednostmi, vendar uporabljenih na različnih statističnih populacijah, lahko to naredimo tako, da **združimo** njihove grafikone.

V naslednjem zgledu združimo obe populacijski piramidi v enem grafikonu. Vsaka je predstavljena s svojo barvo.

- Oba diagrama združimo za lažjo primerjavo.



- Zgled: Primerjava rezultatov kolokvija in teorije za 3.kolokvij MS2



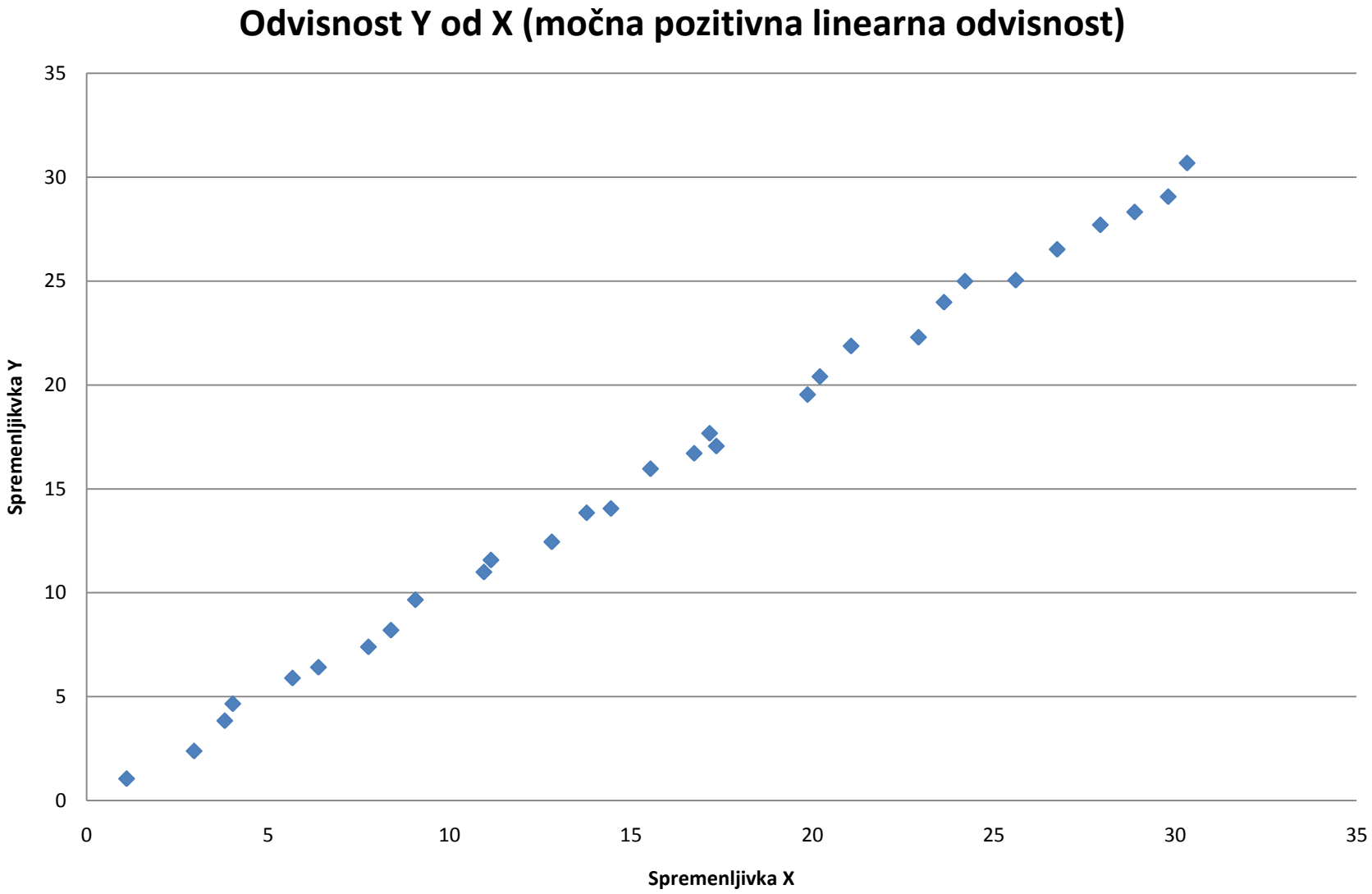
Če imamo dve numerični intervalni spremenljivki lahko njuno odvisnost predstavimo v t.i. **RAZSEVNEM GRAFIKONU:**

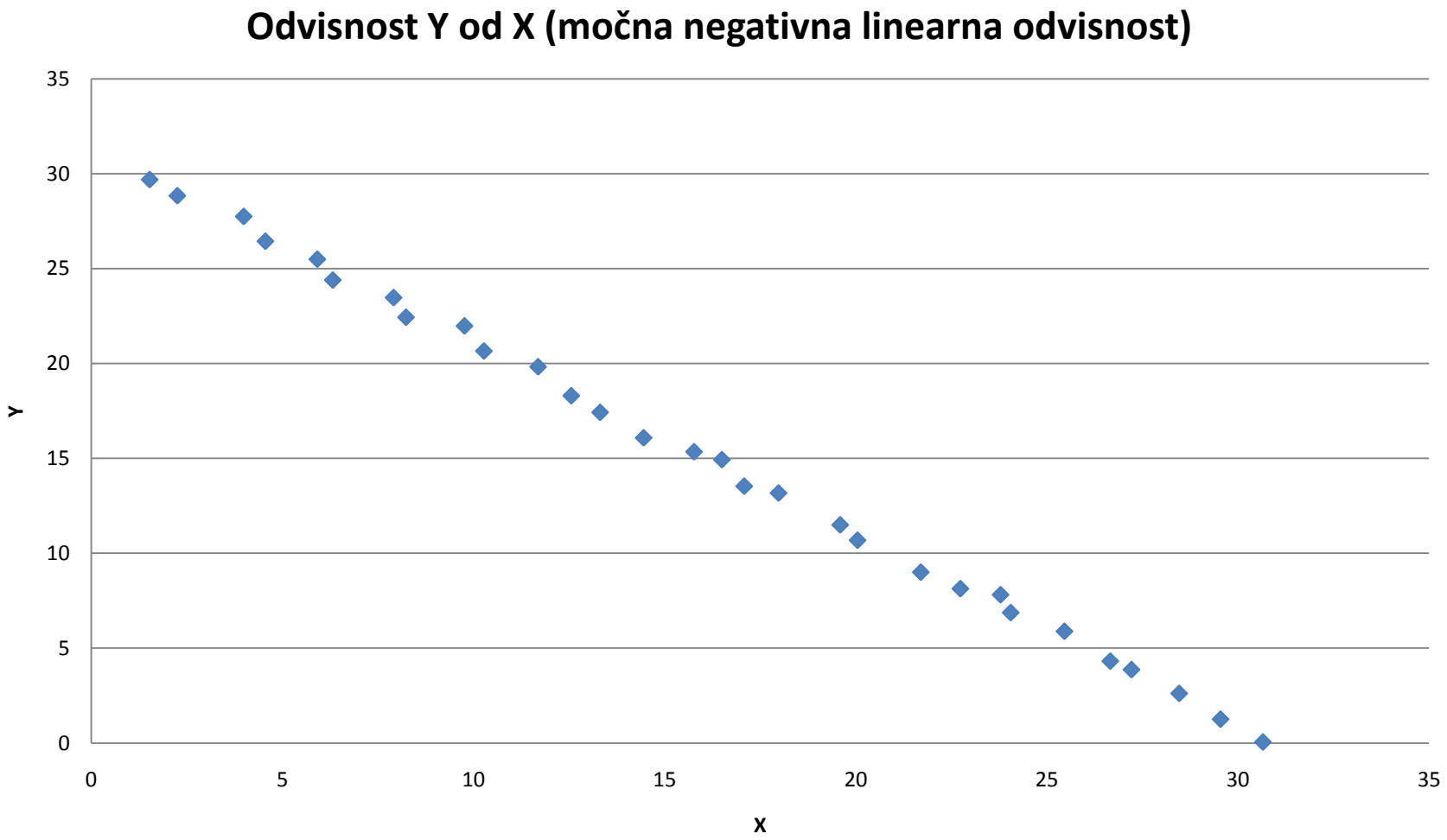
Imamo dve množici podatkov:

X)1,524505 2,694677 3,709311 4,795746 5,50784 6,185665 7,757768 8,103441 9,161323 10,96609 11,64939 12,12109 13,78198 14,51035 15,91111 16,13804 17,07927 17,39785 19,99723 20,27036 21,66084 22,20269 23,765 24,24774 25,01835 26,80562 27,68878 28,63467 29,01699 30,73425

Y) 1,664431 2,480836 3,820409 4,686738 5,095184 6,357992 7,763891 8,94842 9,367478 10,7598 11,64767 12,94306 13,97939 14,72763 15,52571 16,99587 17,88907 17,10049 19,97464 20,98283 21,24549 22,88656 23,35022 24,64358 25,77929 26,96972 27,04937 28,22255 29,33158 30,23772

Iz tega ne vidimo nobene povezame med X in Y. Če pa ju predstavimo grafično, dobimo

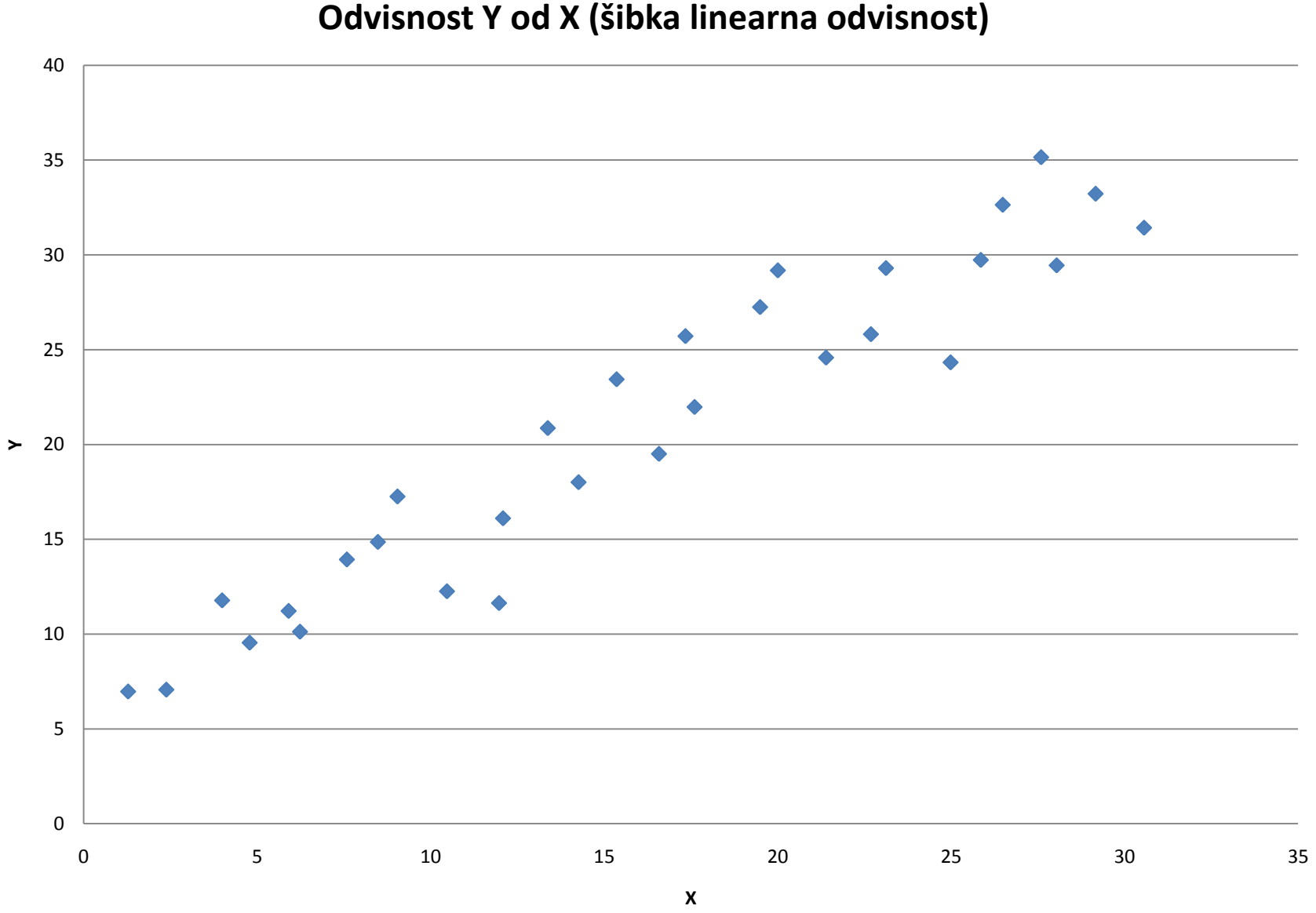

Odvisnost Y od X (močna pozitivna linearna odvisnost)
Spremenljikvka Y
Spremenljivka X
0
5
10
15
20
25
30
35

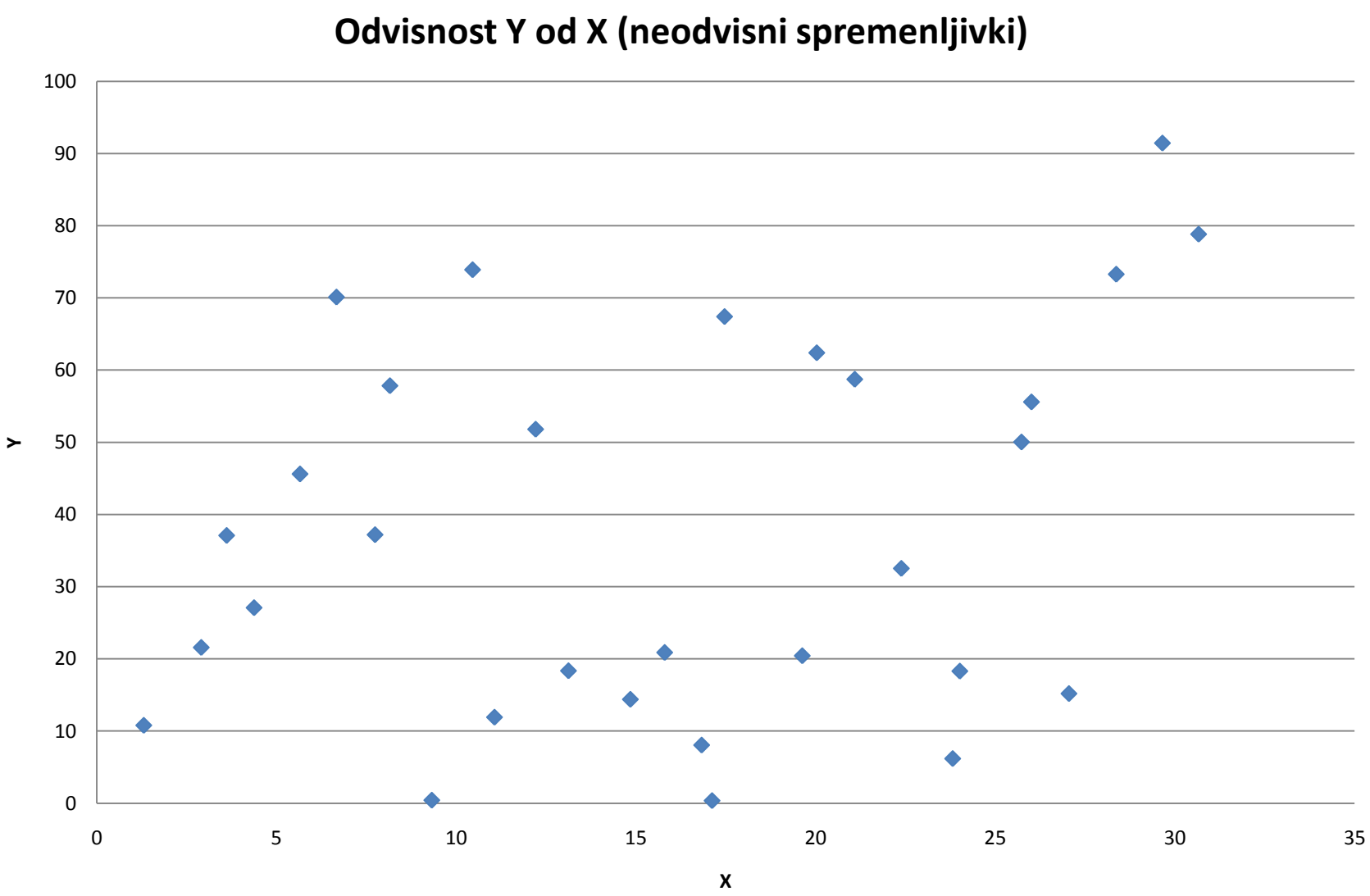

Odvisnost Y od X (močna negativna linearna odvisnost)
35
30
25
20
15
10
5
0
Y
0
5
10
15
20
25
30
35
X

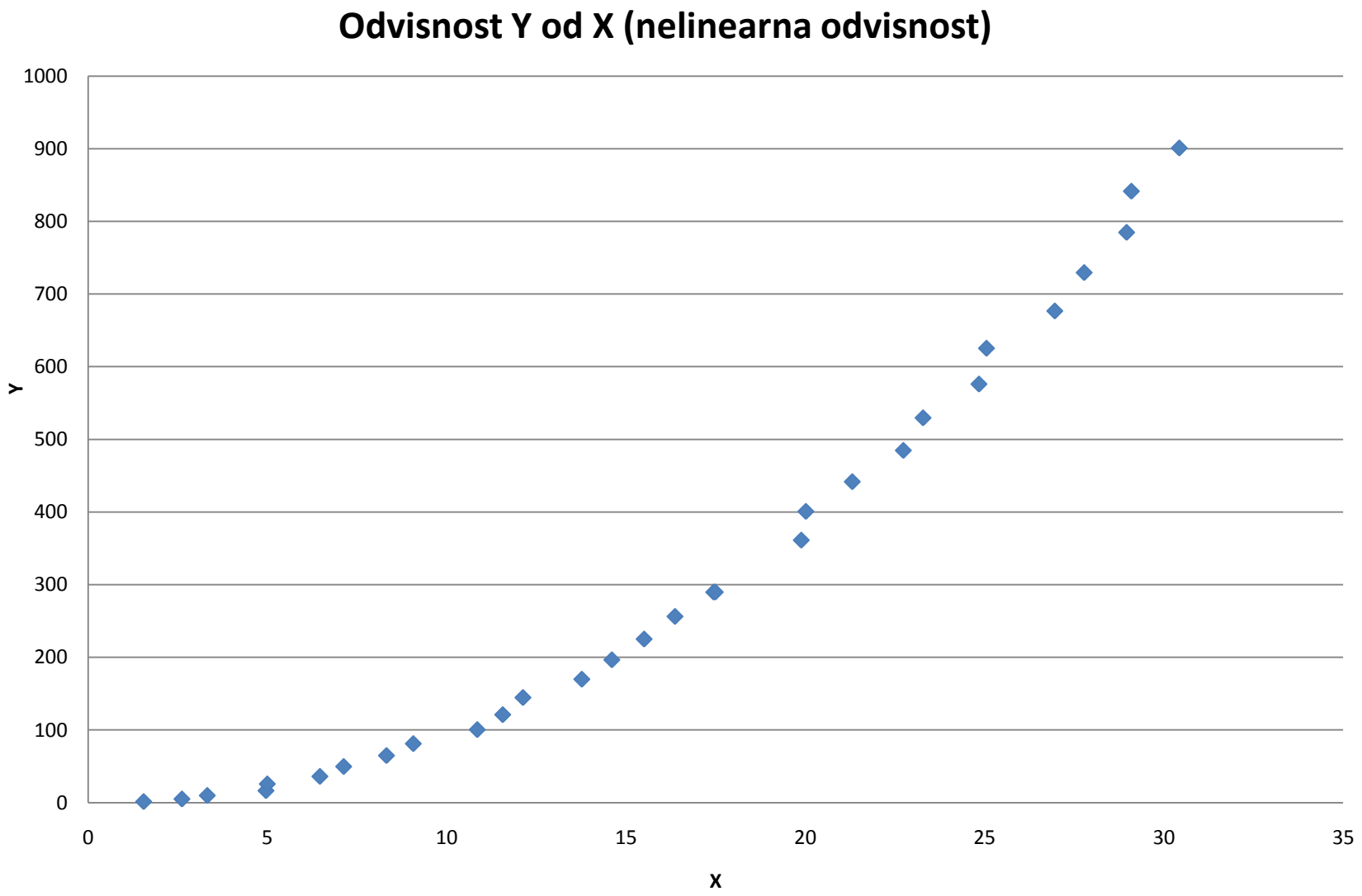
V grafičnem prikazu za **vsak par podatkov (X,Y) zapišemo točko v ravnini**. Tako dobimo RAZSEVNI GRAFIKON za spremenljiki X in Y.

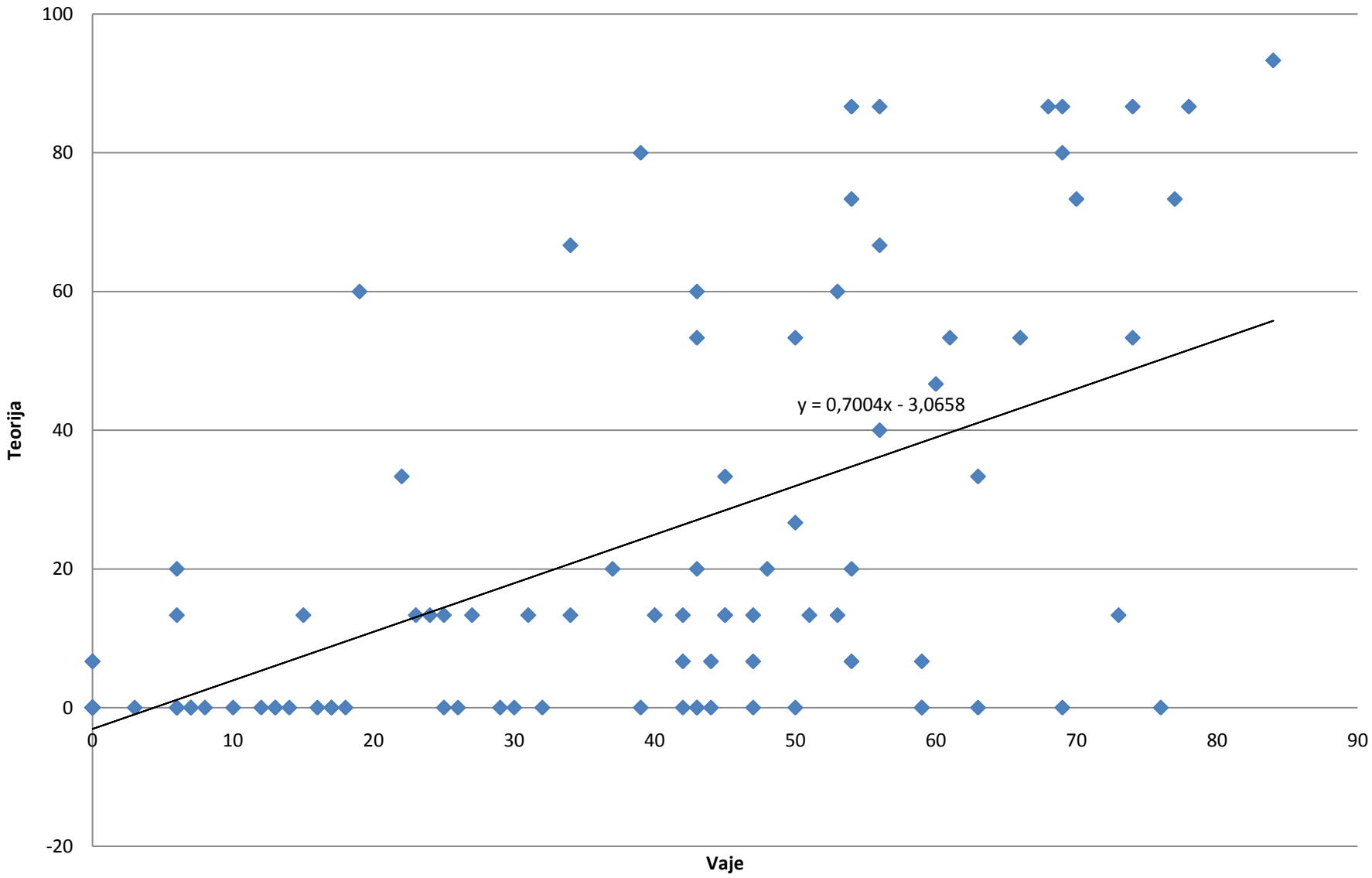
V tem primeru vidimo, da sta X in Y močno **linearno povezani.** To bomo kasneje skušali bolje opisati.

Narišimo še nekaj razsevnih grafikonov.


Odvisnost Y od X (šibka linearna odvisnost)
40
35
30
25
20
15
10
5
0
y
0
5
10
15
20
25
30
35
x


Odvisnost Y od X (neodvisni spremenljivki)
100
90
80
70
60
50
40
30
20
10
0
Y
0
5
10
15
20
25
30
35
X


Odvisnost Y od X (nelinearna odvisnost)
1000
900
800
700
600
500
400
300
200
100
0
Y
0
5
10
15
20
25
30
35
X


Prikaz zveze rezultatov med vajami in teorijo 3.kolovij MS2
Teorija
Vaje
y = 0,7004x - 3,0658
100
80
60
40
20
0
-20
0
10
20
30
40
50
60
70
80
90

STATISTIČNI PARAMETRI
za intervalne spremenljivke

SREDNJE VREDNOSTI

- **ARITMETIČNA SREDINA ali POVPREČNA VREDNOST:** če statistična spremenljivka na populaciji z N elementi zavzame vrednosti x_1, x_2,...,x_N (vsako le enkrat), je njena srednja vrednost enaka

$$\overline{x} = (x_1 + x_2 + \ldots + x_N)/N$$

- Če statistična spremenljivka zavzame vrednosti $x_1, x_2,\ldots,x_k$ s frekvencami $f_1, \ldots f_k$, pri čemer je

$$f_1 + f_2 + \ldots + f_k = N$$

Tedaj dobimo srednjo vrednost kot **tehtano srednjo vrednost:**

$$\overline{x} = (f_1x_1 + f_2 x_2 + \ldots + f_k x_k)/N$$

- **Zgled:** Tečajniki angleškega jezika

Ocena	Število odgovorov
1	1
2	2
3	4
4	5
5	3
Skupaj	15

Dobimo povprečno oceno kot:

$$\frac{(1x1 + 2x2 + 4x3 + 5x4 + 3x5)}{15} \approx 3,47$$

- Zgled: Planinske postojanke.

(podatki PZS) : 1086, 1793, 1526, 1453, 1375, 837, 1123, 600, 1534, 1396, 1864, 1808, 1478, 1460, 1208, 1471, 1534, 1548, 444, 1700, 434, 1356, 961, 725, 1491, 1534 (v metrih).

Povprečna višina= (1086+...+1534)/26 ≈ 1297,65

- Zgled: Telefonski računi. Aritmetična sredina je 43,59.

- **MEDIANA ali SREDIŠČNICA:** mediano izračunamo tako, da vse vrednosti spremenljivke razvrstimo po velikosti (naraščajoče ali padajoče). Vrednost, ki pade v sredino teh podatkov, je mediana. Če je podatkov sodo mnogo, za mediano vzamemo povprečje srednjih dveh vrednosti. Prednost mediane pred aritmetično sredino je v tem, da nanjo podatki, ki močno odstopajo ne vplivajo močno.
- **Zgled:** Plače v podjetju- 20 plač po 1000 Eur in ena plača po 10000 Eur. Aritmetična sredina je 1428,57 Eur. Mediana je 1000 Eur.

- **Zgled:** Tečajniki angleškega jezika. Povprečna ocena je 3,47. Mediana je: 4

  1,2,2,3,3,3,3,4,4,4,4,4,5,5,5

- **Zgled:** Planinske postojanke. Povprečna višina je 1297,65m. Mediana: 1456,5 m

  434 444 600 725 837 961 1086 1123 1208 1356 1375 1396 1453 1460 1471 1478 1491 1526 1534 1534 1534 1548 1700 1793 1808 1864

- **Zgled:** Telefonski računi. Povprečen račun je 43,59. Mediana: 26,91

- **MODUS ali GOSTIŠČNICA** je definirana kot tista vrednost spremenljivke, ki se pojavi z največjo frekvenco. Pri velikih populacijah je običajno bolje gledati modusni razred, t.j., razred vrednosti, ki se pojavi z največjo frekvenco. Če našo spremenljikvo predstavljamo s histogramom oz. razredi, ki imajo enake širine, se modus oz. modusni razred pojavi tam, kjer je histogram **najvišji**. Prav tako modus ni nujno enolično določen (denimo pri bimodalnih histogramih).

- **Zgled:** Tečajniki angleškega jezika. Povprečna ocena je 3,47. Mediana je 4. Modus je 4.

  1,2,2,3,3,3,3,4,4,4,4,4,5,5,5
- **Zgled:** Planinske postojanke. Povprečna višina je 1297,65m. Mediana je 1456,5 m. Modus je 1534m, ki se pojavi trikrat. Modusni razred je od 1400-1600m.

  434 444 600 725 837 961 1086 1123 1208 1356 1375 1396 1453 1460 1471 1478 1491 1526 1534 1534 1534 1548 1700 1793 1808 1864

MERE VARIABILNOSTI

- **VARIANCA ali DISPERZIJA ali RAZPRŠENOST**:

Tako imenovane srednje vrednosti podatkov nam ne povedo veliko o podatkih, ki so močno razpršeni. To razpršenost podatkov merimo z varianco ali disprezijo in pa tudi s **STANDARDNIM ODKLONOM ali STANDARDNO DEVIACIJO.**

OZNAKE: varianca : σ^2

standardna deviacija : $\sigma = \sqrt{\sigma^2}$

- Varianco ali disperzijo dobimo s formulo:

$\sigma^2 = (f_1 (x_1 - \overline{x})^2 + f_2 (x_2 - \overline{x})^2 + \ldots + f_k (x_k - \overline{x})^2)/N$

Pri tem je vsota frekvenc $f_1 + \ldots + f_k = N$.

Število $\overline{x}$ je aritmetična sredina vrednosti spremenljivke.

- **Zgled:** Tečajniki angleškega jezika:

Povprečje: 3,47

Vrednosti: 1,2,2,3,3,3,3,4,4,4,4,4,5,5,5

Torej je

$\sigma^2 = ((1-3,47)^2 + 2(2-3,47)^2 + 4(3-3,47)^2 + 5(4-4,47)^2 + 3(5-3,47)^2)/15 \approx 1,32$

In zato je standardni odklod $\sigma \approx 1,15$

- **Zgled:** Planinske postojanke. Varianca 159173,1 m^2 standardni odklon 398,97 m.

- **Zgled:** Telefonski računi. Varianca je 1511,045 in standardni odklon je 38,87.

INTERPRETACIJA STANDARDNEGA ODKLONA:

Za večino vsakodnevnih spremenljivk (**normalno porazdeljenih**) veljajo naslednja empirična pravila:

- Približno 68% vseh vrednosti spremenljikve je od povprečja oddaljeno za manj kot 1 standardni odklon.
- Približno 95% vseh vrednosti spremenljikve je od povprečja oddaljeno za manj kot 2 standardna odklona.
- Približno 99,7% vseh vrednosti spremenljikve je od povprečja oddaljeno za manj kot 3 standardne odklone.

MERE LINEARNE ODVISNOSTI

Za intervalni statistični spremenljivki bomo vpeljali mere njune linearne odvisnosti.

- **KOVARIANCA:** Naj bosta X in Y intervalni statistični spremenljivki na isti statistični množici z N elementi. Spremenljika X naj ima vrednosti $x_1,x_2,...,x_N$ in spremenljivka Y naj ima vrednosti $y_1,y_2,...,y_N$. Naj bosta $\bar{x}$ in $\bar{y}$ njuni povprečni vrednosti.

Tedaj se njuna **kovarianca** glasi:

$$\mathrm{COV}(X,Y) = \frac{\sum_{1}^{N}(x_j - \bar{x})(y_j - \bar{y})}{N}$$

Zakaj je to mera linearne odvisnosti statističnih spremenljivk X in Y ?

Če sta X in Y močno linearno odvisni spremenljivki, t.j. $Y \approx a\,X + b$, potem za njuno kovarianco dobimo vrednost

$$a\,\sigma^2(X)$$

- Če sta X in Y močno **pozitivno linearno odvisni** (a>0), je njuna kovarianca **“močno” pozitivna**.

-ko je $x_j - \bar{x}$ pozitivno število, je tudi $y_j - \bar{y}$ pozitivno število in vsota “veliko” pozitivnih števil je “močno” pozitivna.

- Če sta X in Y močno **negativno linearno odvisni** (a<0), je njuna kovarianca **“močno” negativna**.

-ko je $x_j - \bar{x}$ pozitivno število , je tudi $y_j - \bar{y}$ negativno število in vsota “veliko” negativnih števil je “močno” negativna.

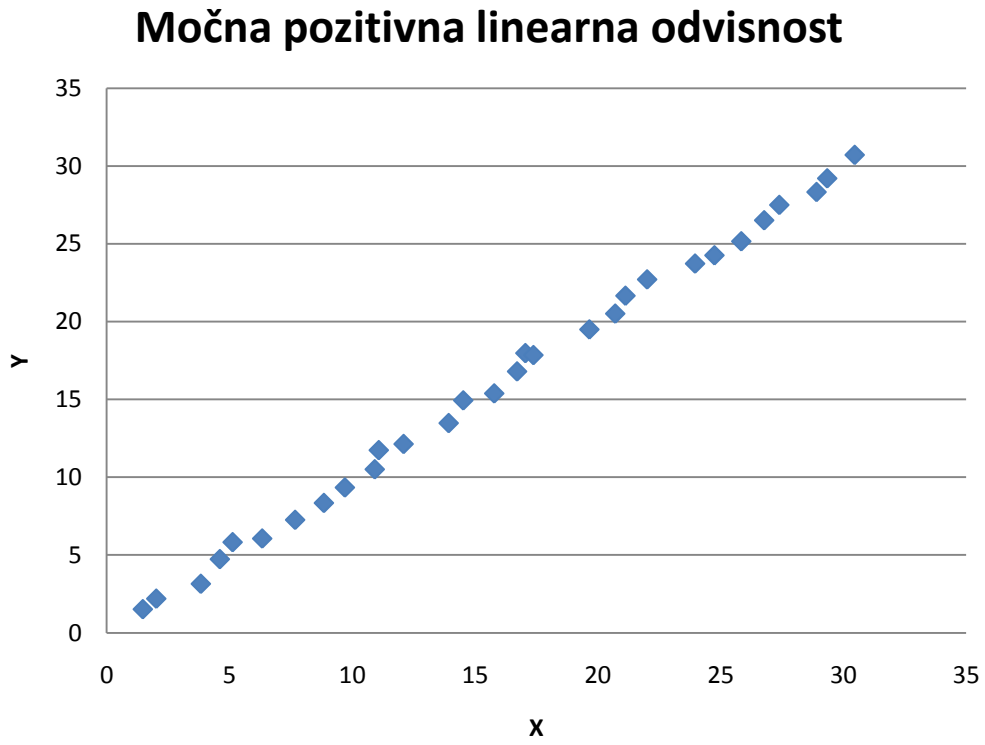
- Če pa X in Y nista linearno odvisni, se zdi, da se členi v vsoti za kovarianco odštevajo in vsota je “bližje” 0, kot v prejšnjih primerih.

Zgled: Izračunajmo kovarianco za množici podatkov:

X) 1,475616 2,023828 3,838456 4,611971 5,131717 6,335841 7,677096 8,848525 9,702592 10,91907 11,08242 12,09369 13,92795 14,5224 15,78299 16,71286 17,05027 17,37885 19,65718 20,70894 21,12557 22,00823 23,95943 24,74859 25,84091 26,77276 27,38948 28,90513 29,33939 30,4585

Y) 1,517662 2,195559 3,150433 4,735985 5,82588 6,061236 7,260428 8,346499 9,340551 10,5098 11,74202 12,13735 13,48112 14,93904 15,38828 16,8019 17,97917 17,85161 19,50056 20,51712 21,66606 22,71492 23,73351 24,26439 25,16746 26,51828 27,50525 28,33544 29,21302 30,72098

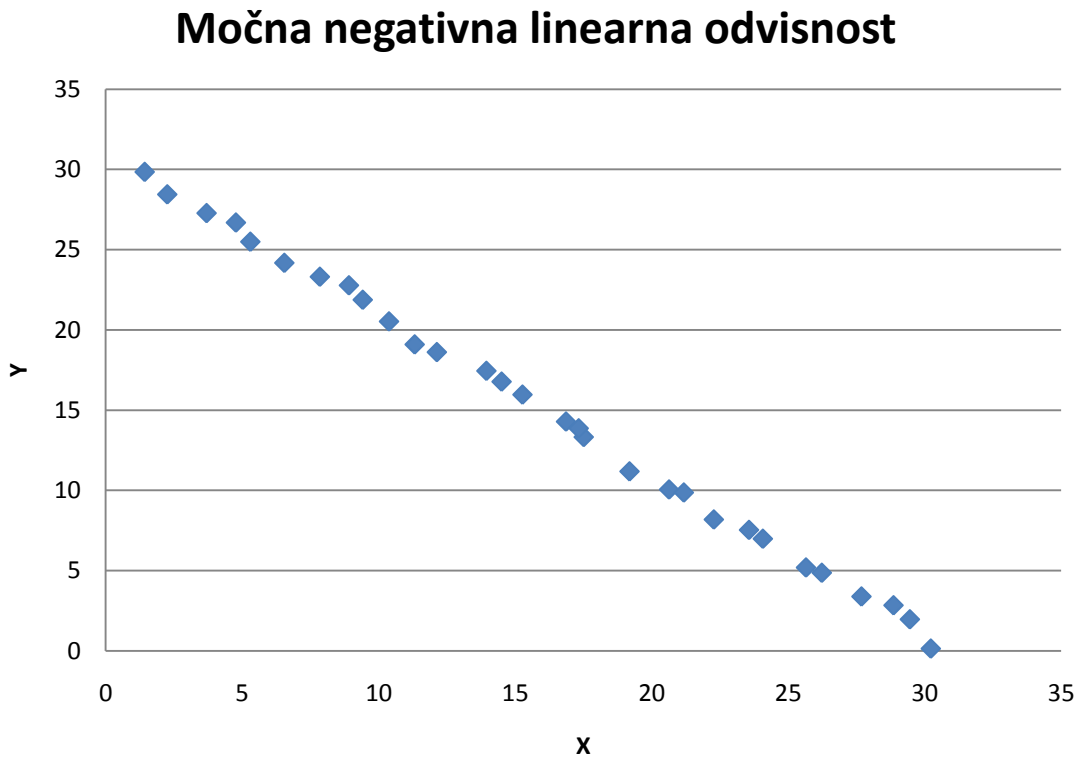
- V tem primeru dobimo naslednji razsevni grafikon



In kovarianco:

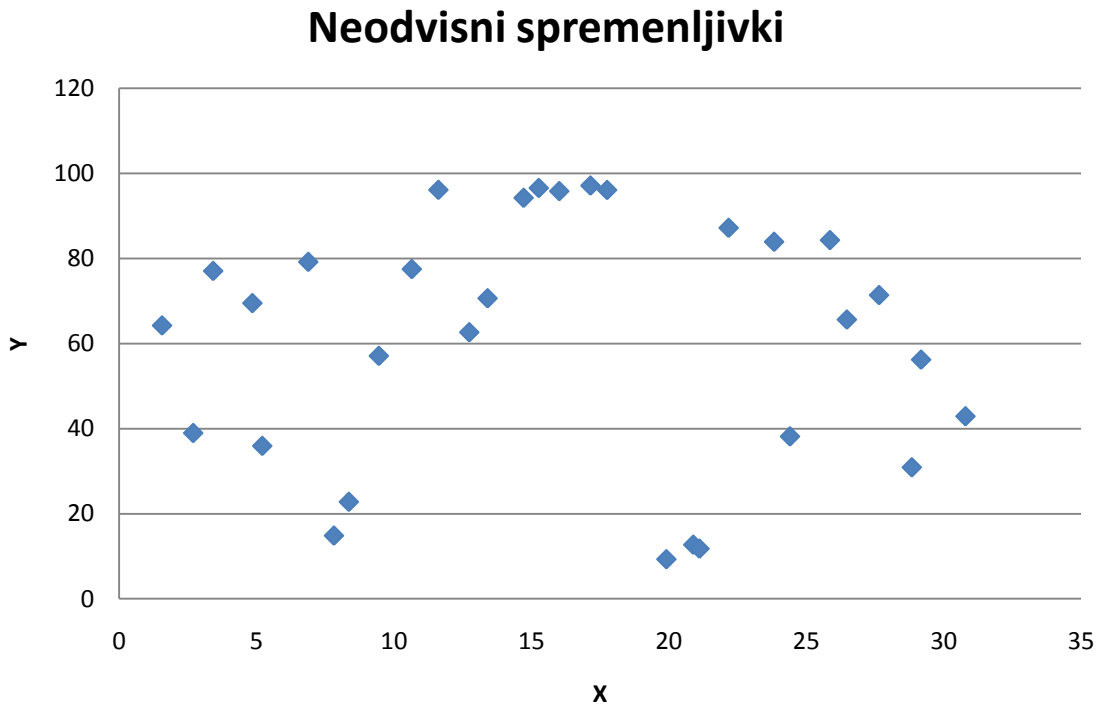
COV(X,Y) = 75,33588 (pozitivna!)

Če pa vzamemo podatke, ki imajo močno negativno linearno odvisnost – razsevni grafikon oblike:



Dobimo kovarinaco: COV(X,Y)= -74,3718 (negativna!)

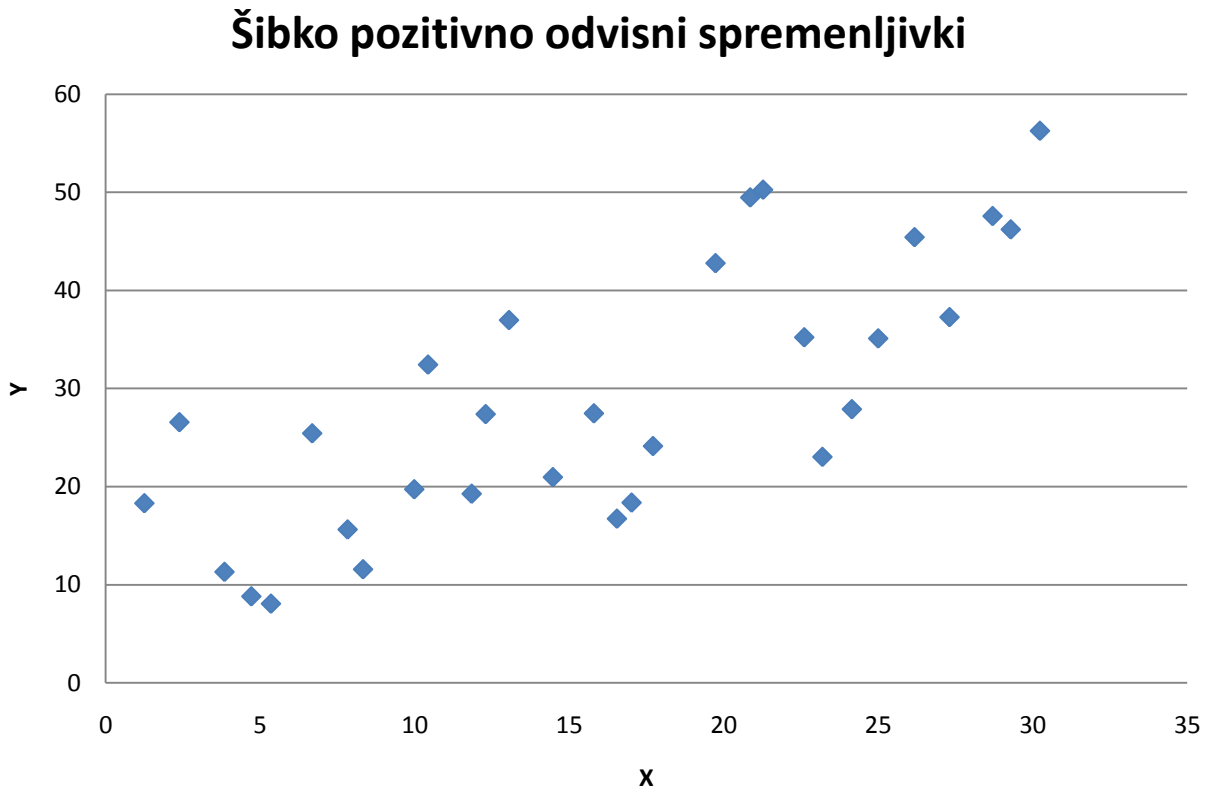
- Pri podatkih, kjer linearne odvisnosti ne pričakujemo:



Kovarianca je COV(X,Y) = -11,5698

Kljub temu, da kovarianca nosi informacijo o linearni odvisnosti dveh statističnih spremenljivk pa ta informacija še zdaleč ni popolna. Namreč, večja vrednost kovariance še ne pomeni, da sta spremenljivki močneje linearno povezani. Kot zgled si lahko ogledamo dve šibko pozitivno linearno odvisni spremenljivki.

- **Zgled:**



Kovarianca je COV(X,Y)= 86,6734

Taki primeri povedo, da moramo spremenljivki ustrezno normalizirazi. Če imata spremenljivki sami veliko varianco, le ta vpliva na velikost njune kovariance. Zato vpeljemo

KORELACIJSKI KOEFICIENT

ali tudi

PEARSONOV KOEFICIENT KORELACIJE

po britanskem statistiku Karlu Pearsonu. Definiran je kot kvocient

$$\rho = COV(X,Y)/\sigma_X\sigma_Y$$

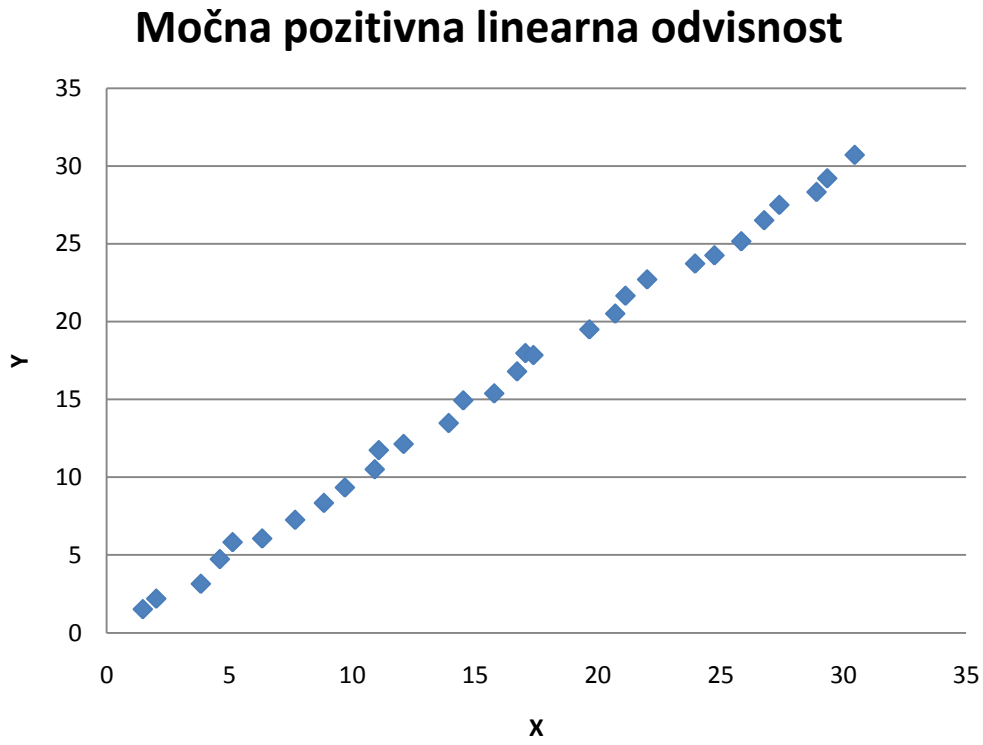
- **Vrednosti** korelacijskega koeficienta so vedno **med -1 in 1!**
- Če je vrednost korelacijskega koeficienta **-1**, imata spremenljivki **negativno** linearno odvisnost.
- Če je vrednost korelacijskega koeficienta **1**, imata spremenljivki **pozitivno** linearno odvisnost.
- Če je vrednost korelacijskega koeficienta **0**, med spremenljivkama **ni linearne odvisnosti**.

- **Zgled:** Za podatke, ki smo jih že srečali

X) 1,475616 2,023828 3,838456 4,611971 5,131717 6,335841 7,677096 8,848525 9,702592 10,91907 11,08242 12,09369 13,92795 14,5224 15,78299 16,71286 17,05027 17,37885 19,65718 20,70894 21,12557 22,00823 23,95943 24,74859 25,84091 26,77276 27,38948 28,90513 29,33939 30,4585

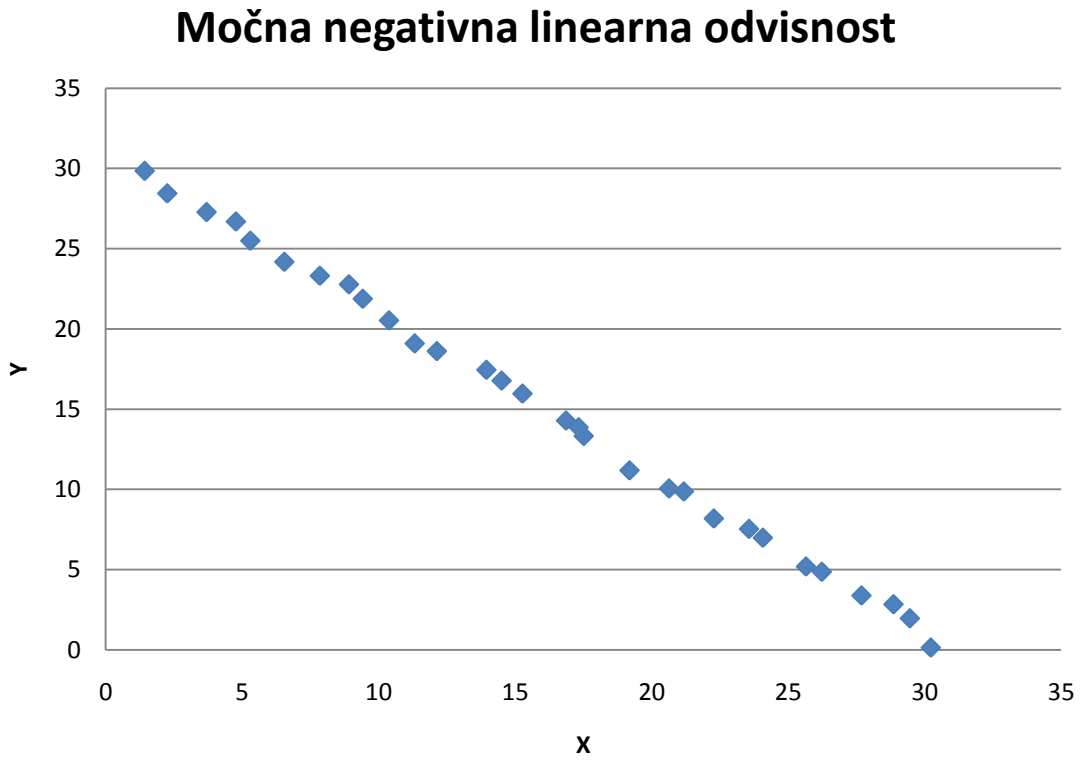
Y) 1,517662 2,195559 3,150433 4,735985 5,82588 6,061236 7,260428 8,346499 9,340551 10,5098 11,74202 12,13735 13,48112 14,93904 15,38828 16,8019 17,97917 17,85161 19,50056 20,51712 21,66606 22,71492 23,73351 24,26439 25,16746 26,51828 27,50525 28,33544 29,21302 30,72098

z razsevnim grafikonom:



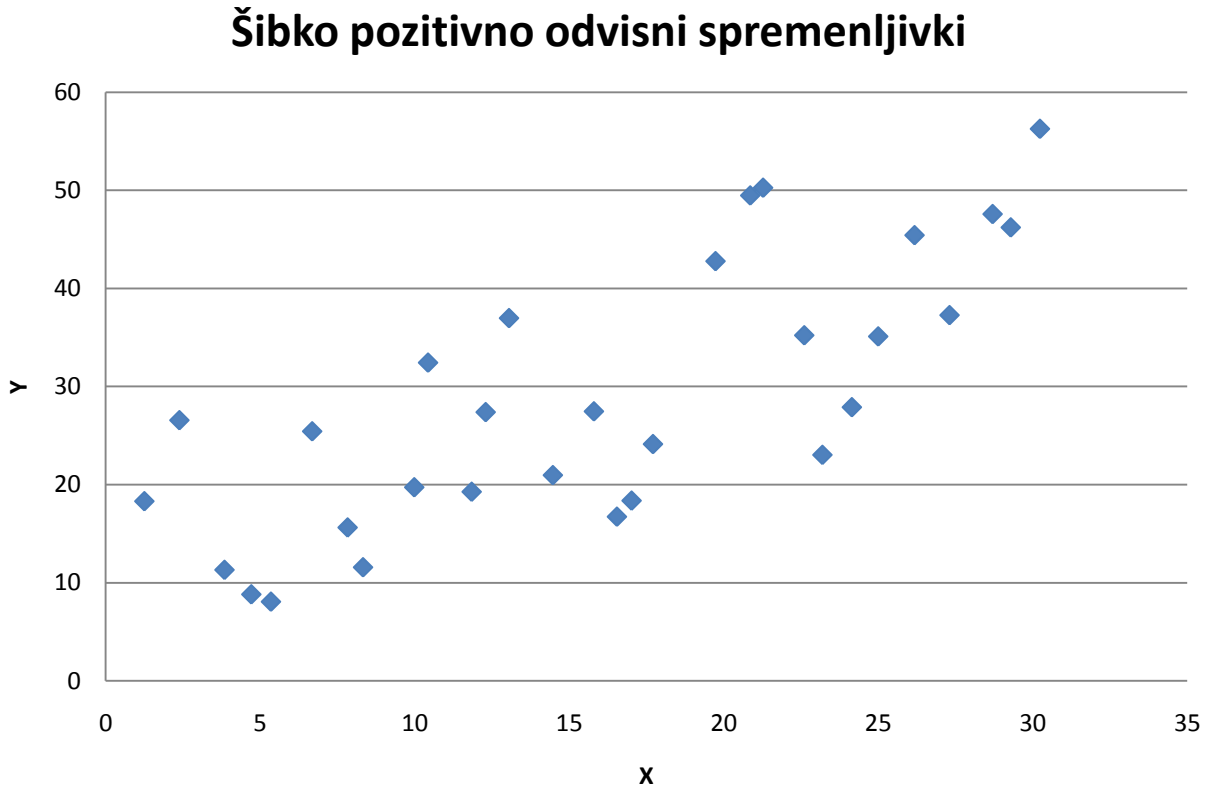
dobimo $\rho = 0{,}998694$ (blizu 1!)

- Zgled: za podatke z razsevnim grafikonom



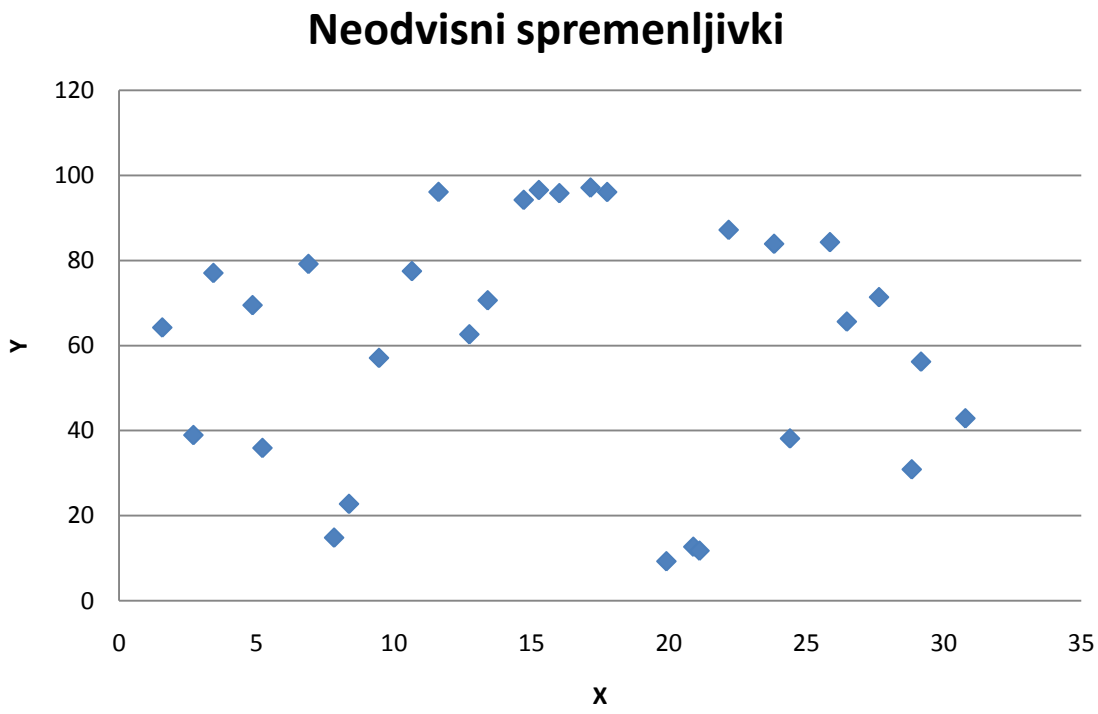
dobimo $\rho = -0{,}99916$ (blizu -1!)

- **Zgled:** Za šibko linearno odvisne podatke dobimo:



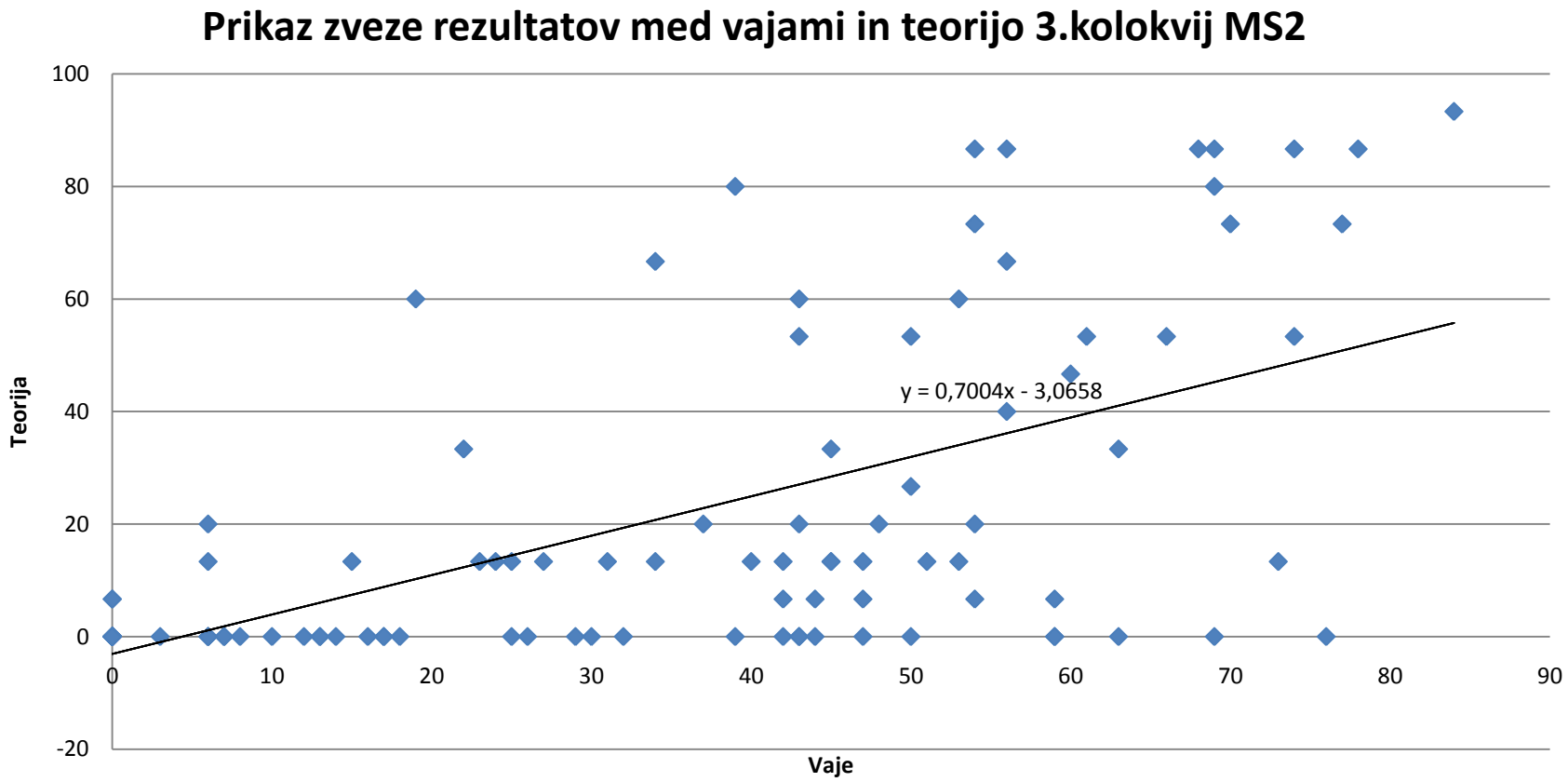
- Korelacijski koeficient je $\rho = 0{,}762955$

- **Zgled:** Dve neodvisni spremenljivki



- Korelacijski koeficient je ρ = -0,04712 (blizu 0!)

- **Zgled:**



- **Tokrat je korelacijski koeficient ρ =** 0,627728

- Zelo močna pozitivna linearna korelacija: $\rho > 0,9$
- Močna pozitivna linearna korelacija: $0,7 < \rho < 0,9$
- Srednja pozitivna linearna korelacija: $0,5 < \rho < 0,7$
- Nizka pozitivna linearna korelacija: $0,3 < \rho < 0,5$
- Ni korelacije: $-0,3 < \rho < 0,3$
- Nizka negativna linearna korelacija: $-0,5 < \rho < -0,3$
- Srednja negativna linearna korelacija: $-0,7 < \rho < -0,5$
- Močna negativna linearna korelacija: $-0,9 < \rho < -0,7$
- Zelo močna negativna linearna korelacija: $\rho < -0,9$

REGRESIJSKA PREMICA (LINEARNA REGRESIJA)

- Če za dve intervalni statistični spremenljivki X in Y slutimo oz. celo vemo, da sta nekoliko linearno odvisni, skušamo najti premico, ki dane podatke “najboljše” aproksimira. Odločimo se za premico

$$y = \alpha + \beta x,$$

ki podatke najboljše aproksimira v smislu **metode najmanjših kvadratov.**

- Iščemo torej premico **y = α + β x**, oziroma iščemo koeficienta **α** in **β** za katero velja, da je **vsota kvadratov napak**

$$\sum_{1}^{N} (y_i - \alpha - \beta x_i)^2$$

minimalna! Pri tem je β REGRESIJSKI KOEFICIENT in pove za koliko merskih enot se v povprečju spremeni vrednost Y-a, če se vrednost X-a poveča za eno mersko enoto.

- Kako pridemo do koeficientov **α** in **β**? Ko iščemo ekstreme funkcije, le te dobimo v točkah, kjer so odvodi enaki 0! Zato odvajamo izraz

$$\sum_{1}^{N} (y_i - \alpha - \beta x_i)^2$$

po spremenljikah **α** in **β** ter dobljene izraze enačimo z 0. Dobimo:

- po odvajanju na **α**:

$$2\sum_{1}^{N}(y_i - \alpha - \beta x_i) = 0$$

- ter po odvajanju na **β**

$$2\sum_{1}^{N}(y_i - \alpha - \beta x_i)x_i = 0$$

Od tod dobimo:

$$\bar{y} - \alpha - \beta\bar{x} = 0$$

$$(\sum_{1}^{N} y_i x_i) - \alpha(\sum_{1}^{N} x_i) - \beta(\sum_{1}^{N} x_i^2) = 0$$

- Zadnjo enačbo lahko preoblikujemo. Še prej pa moramo opaziti naslednji povezavi:

$$\sigma_X^2 = \frac{\sum_1^N (x_i - \bar{x})^2}{N} = \frac{\sum_1^N x_i^2}{N} - \bar{x}^2$$

ter

$$COV(X,Y) = \frac{\sum_1^N (x_i - \bar{x})(y_i - \bar{y})}{N} = \frac{\sum_1^N x_i y_i}{N} - \bar{x}\bar{y}$$

- Od tod dobimo enačbi:

$$\bar{y} = \alpha + \beta \bar{x}$$

$$\frac{\sum_{1}^{N} x_i y_i}{N} = \alpha \bar{x} + \beta \frac{\sum_{1}^{N} x_i^2}{N}$$

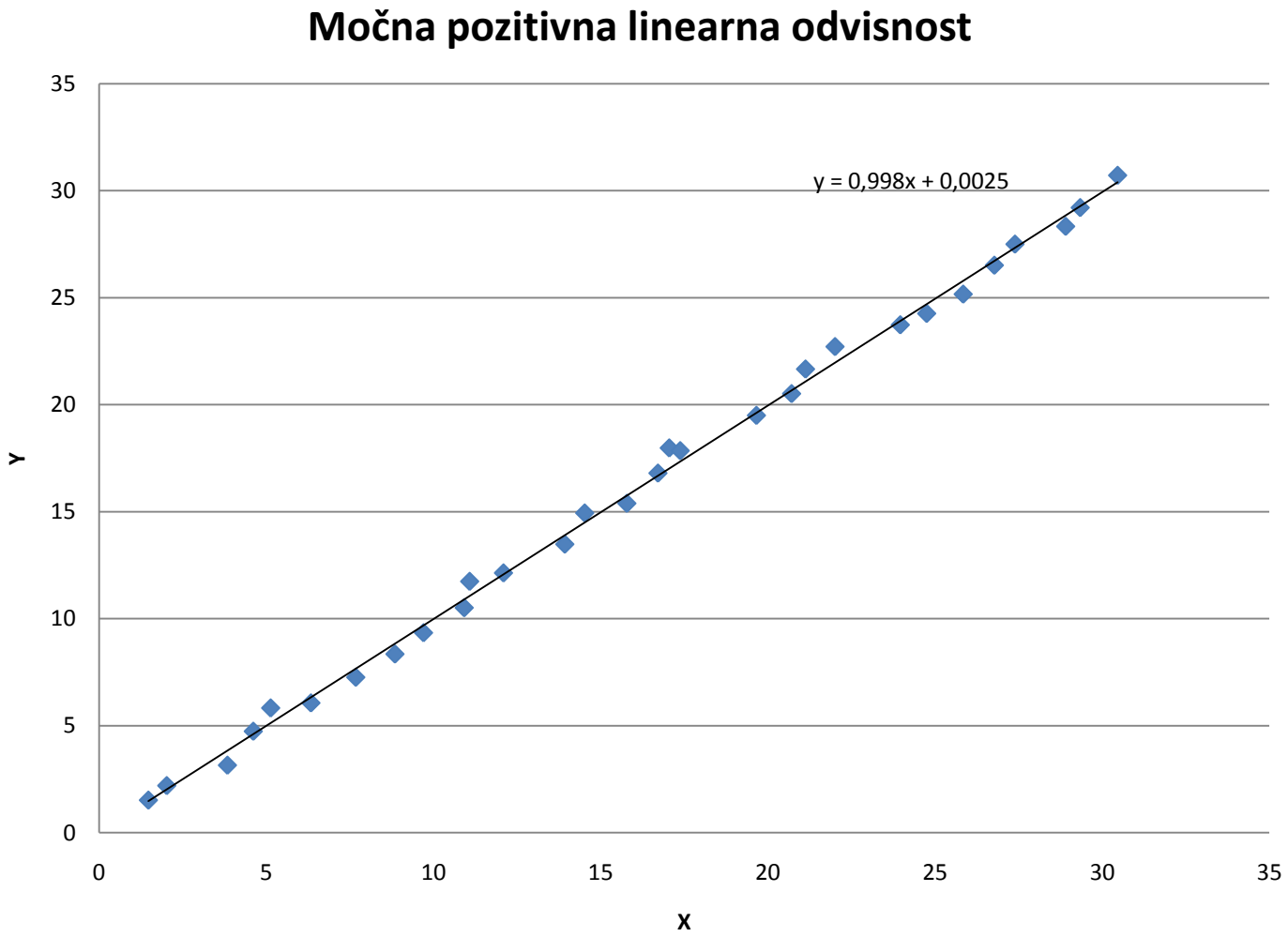
in nato

$$\beta = \frac{COV(X,Y)}{\sigma_X^2}$$

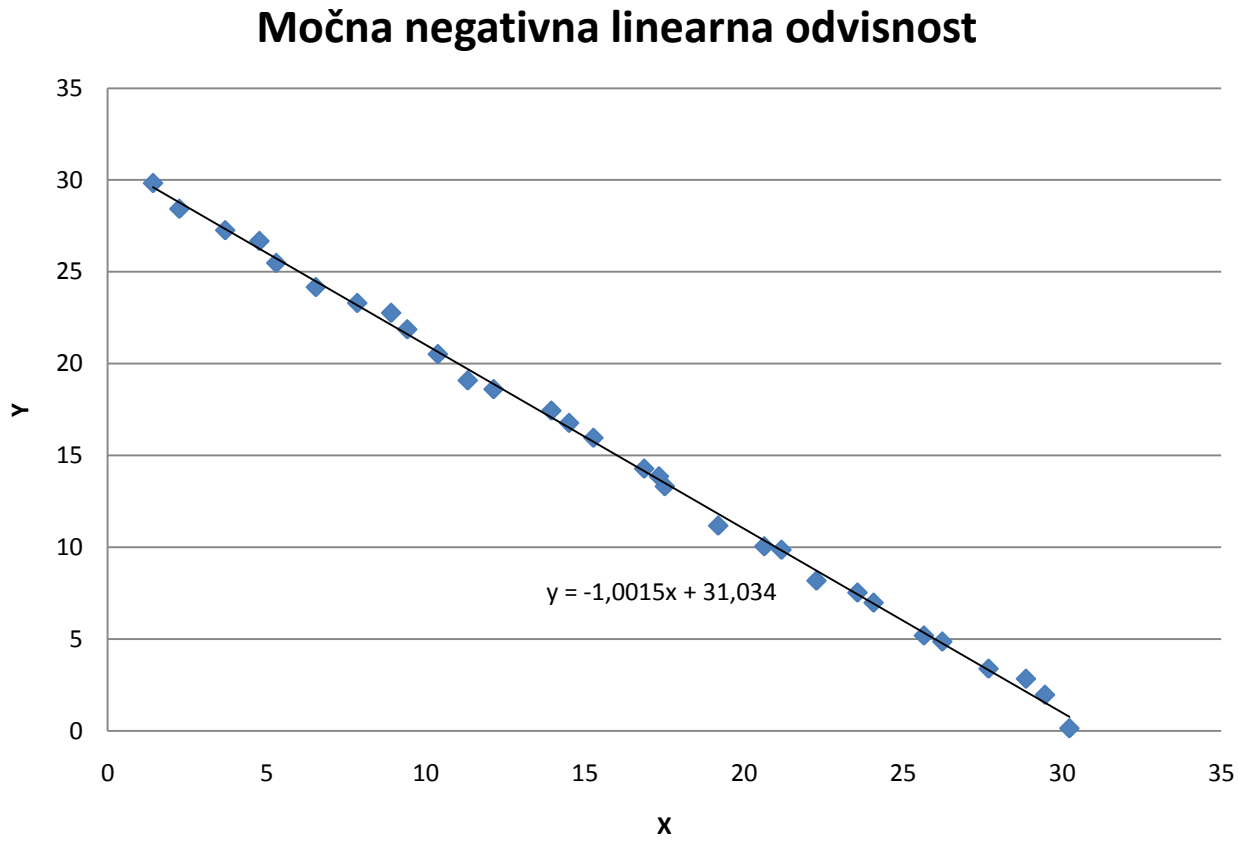
ter

$$\alpha = \bar{y} - \beta \bar{x}$$

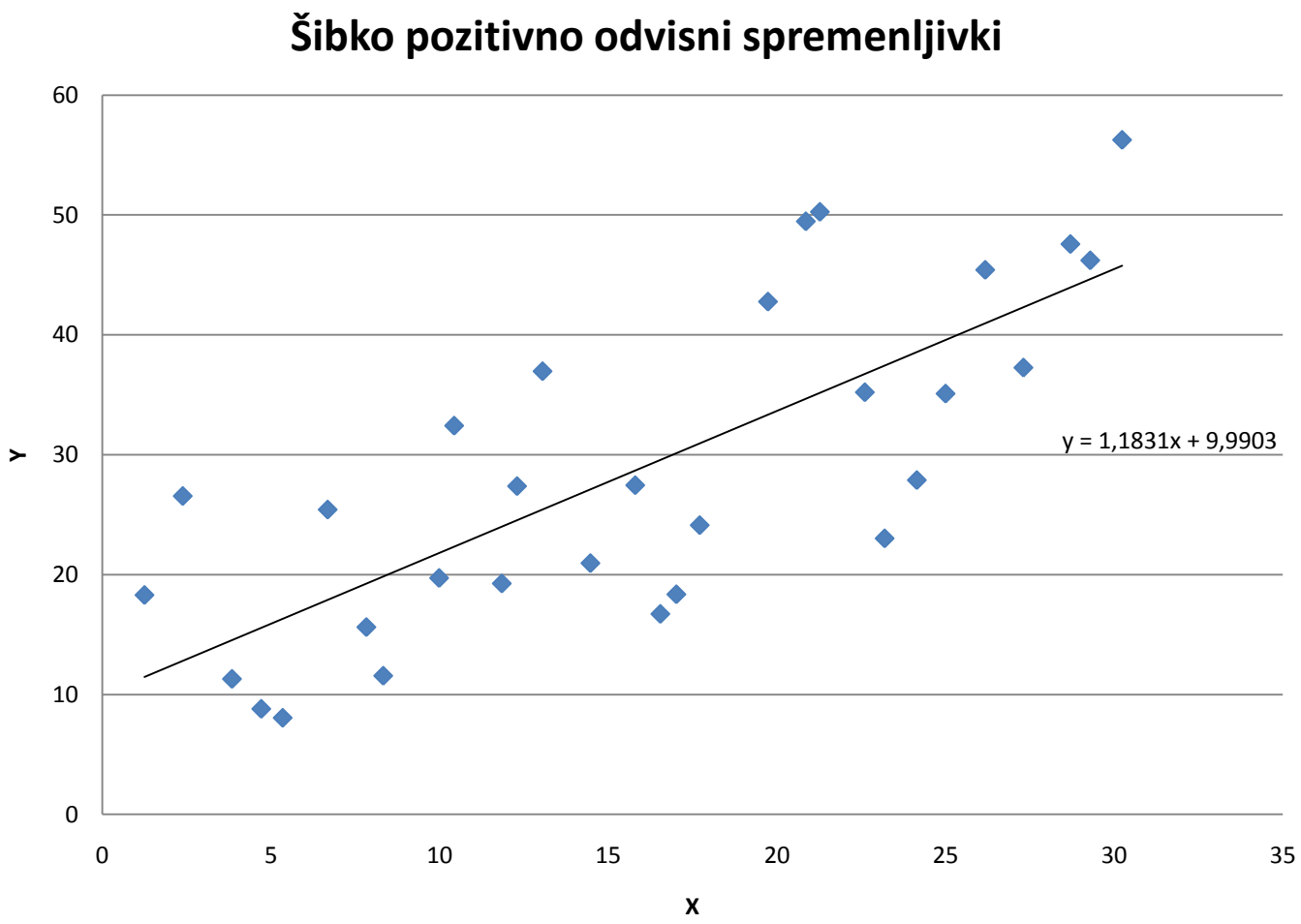
- **Zgled:**


Močna pozitivna linearna odvisnost
y = 0,998x + 0,0025
y
x
0
5
10
15
20
25
30
35

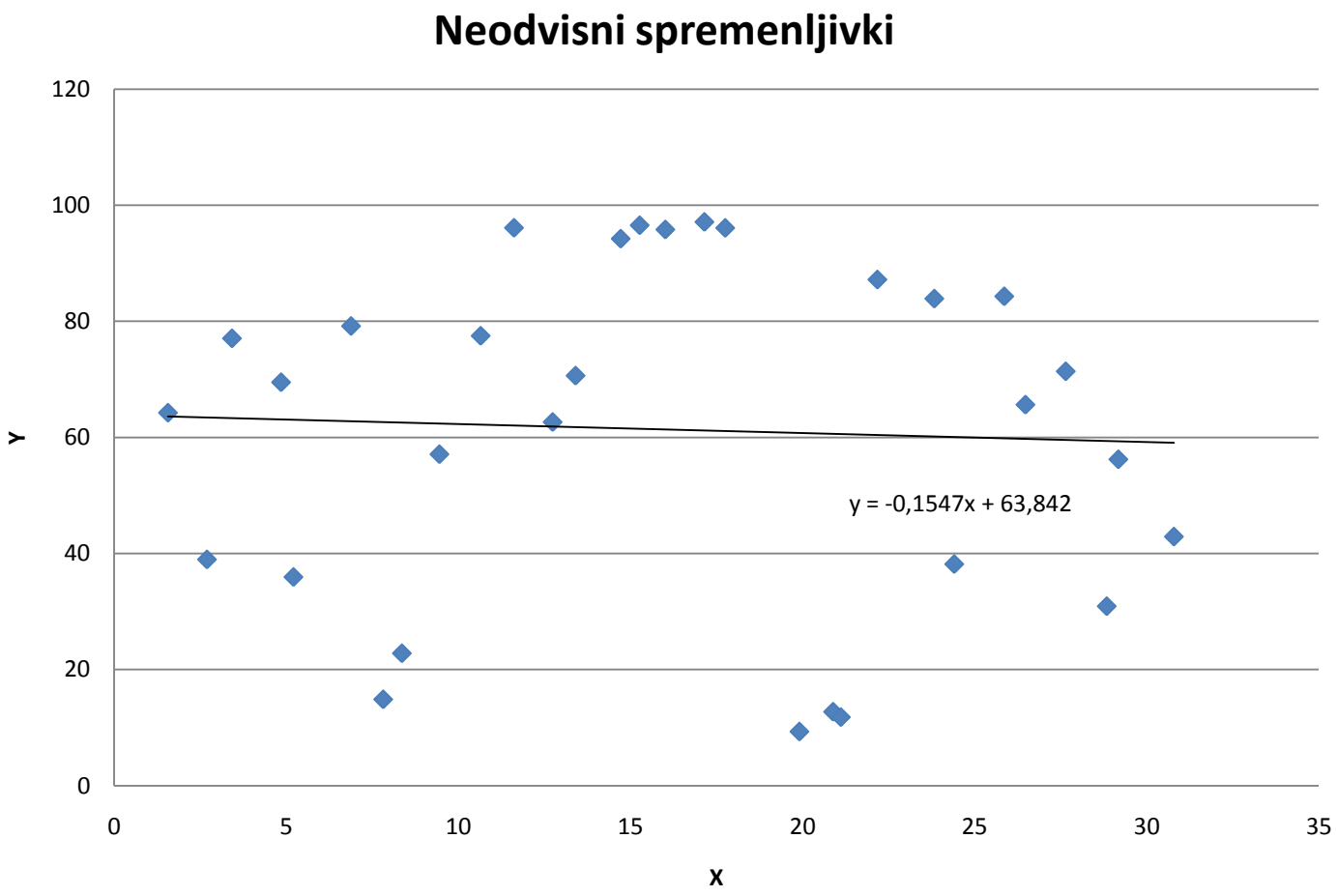
- **Zgled:**


Močna negativna linearna odvisnost
y = -1,0015x + 31,034
y
x
0
5
10
15
20
25
30
35

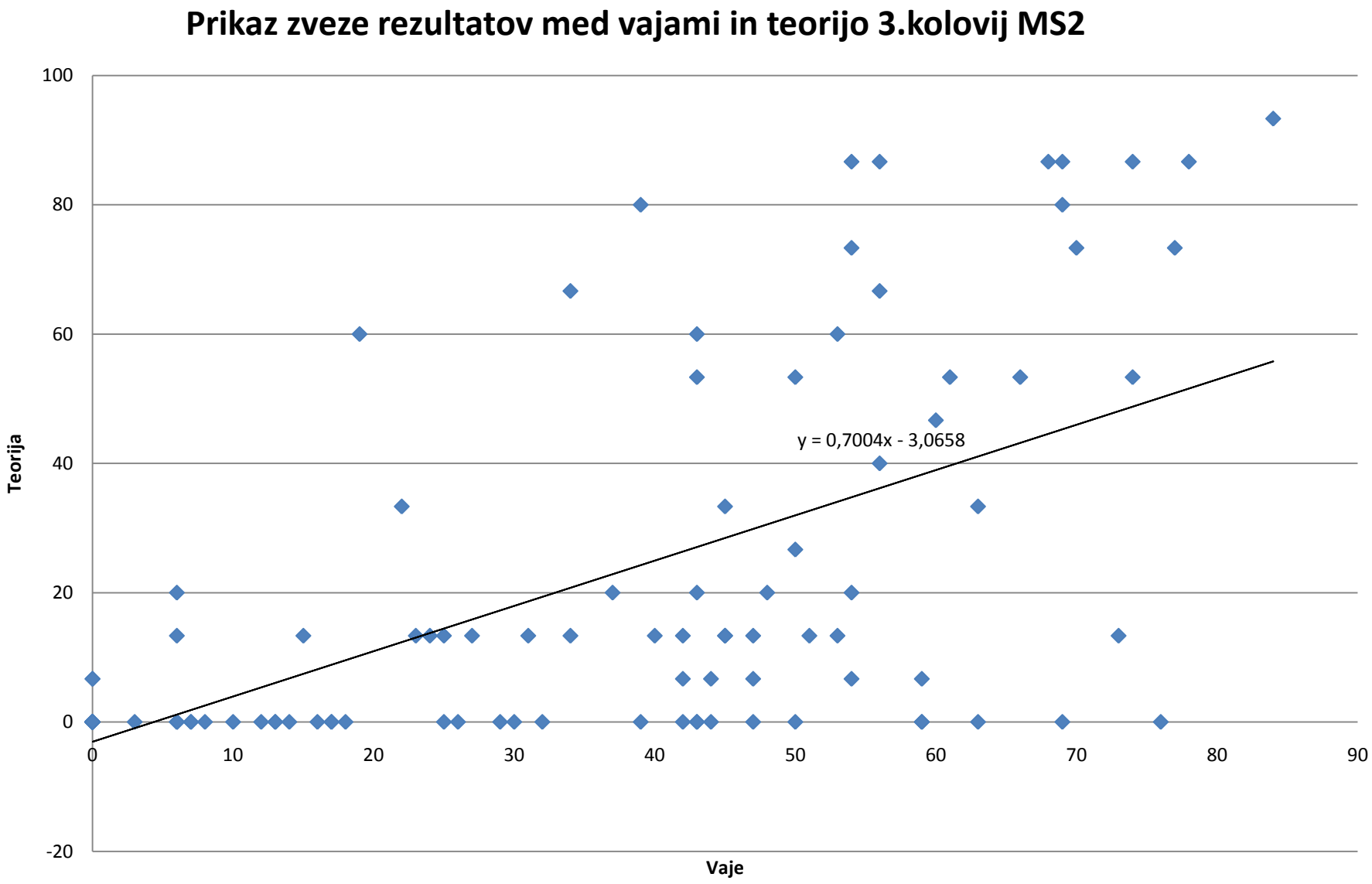
- **Zgled:**


Šibko pozitivno odvisni spremenljivki
y = 1,1831x + 9,9903
y
x
0
10
20
30
40
50
60
0
5
10
15
20
25
30
35

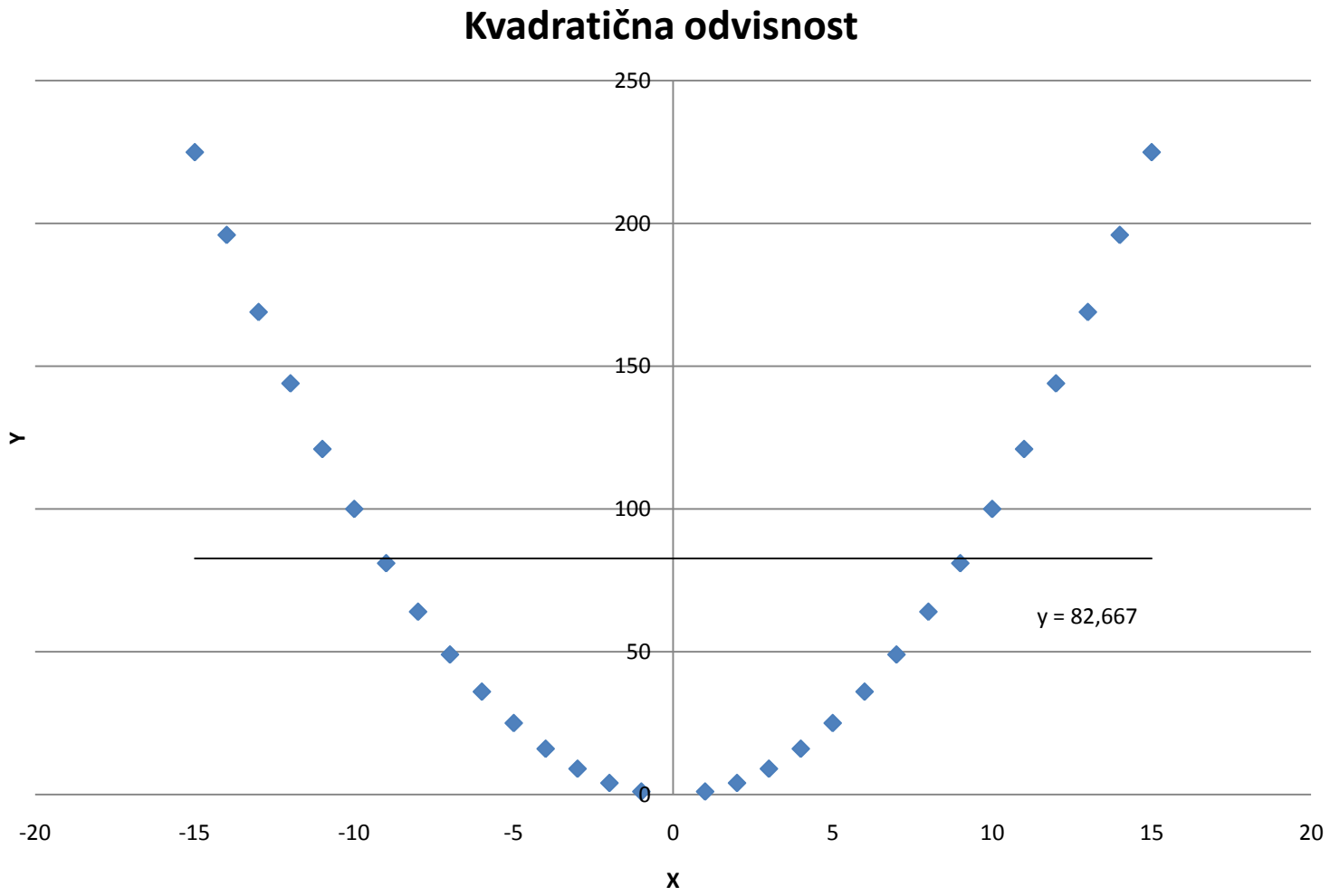
- **Zgled:**


Neodvisni spremenljivki
120
100
80
60
40
20
0
y
0
5
10
15
20
25
30
35
x
y = -0,1547x + 63,842

- **Zgled:**


Prikaz zveze rezultatov med vajami in teorijo 3.kolovij MS2
Teorija
Vaje
y = 0,7004x - 3,0658
100
80
60
40
20
0
-20
0
10
20
30
40
50
60
70
80
90

- **Zgled:** Korelaciijski koeficient je 0.



Oglejmo si še zadnjega od glavnih statističnih koeficientov, povezanih z linearno odvisnostjo dveh statističnih spremenljivk oz. linearno regresijo. To je

DETERMINACIJSKI KOEFICIENT ρ^2, katerega vrednost izračunamo kot kvadrat korelacijskega koeficienta ρ.

Pomen determinacijskega koeficienta je najbolje razviden iz naslednje povezave:

$$\rho^2 = \text{Varianca}(\hat{Y}=\alpha+\beta X) / \text{Varianca}(Y)$$

Vrednost determinacijskega koeficienta je torej razmerje med varianco, ki bi jo imela spremenljivka Y , ki bi bila s spremenljivko X linearno povezana s premico, ki jo dobimo z metodo najmanjših kvadratov, in varianco spremenljivke Y. **Determinacijski koeficient torej pove kolikšen delež celotne variance spremenljivke Y je pojasnjen z regresijsko premico.** Ker je kvadrat korelacijskega koeficienta, nam, seveda, tudi nekaj pove o linearni odvisnosti spremenljivk X in Y.

Izpeljimo povezavo med determinacijskim koeficientom in variancami Y in Y:

1. Najprej izračunajmo aritmetično sredino spremenljivke Y . To je

$$\bar{y} = \frac{\sum_{1}^{N} y_i}{N} = \frac{\sum_{1}^{N} (\alpha + \beta x_i)}{N} = \alpha + \beta \bar{x} = \bar{y}$$

Pri zadnjem enačaju smo uporabili eno od dveh enačb, ki določata koeficienta α in β.

Povprečji sta torej enaki!

2. Sedaj izračunamo varianco Y . Vemo:

$$\sigma_Y^2 = \frac{\sum_1^N y_i^2}{N} - \bar{y}^2 = \frac{\sum_1^N (\alpha + \beta x_i)^2}{N} - \bar{y}^2 =$$

$$= \alpha^2 + 2\alpha\beta\bar{x} + \beta^2 \frac{\sum_1^N x_i^2}{N} - \bar{y}^2 =$$

Upoštevajmo še $\bar{y} = \alpha + \beta\bar{x}$ in dobimo

$$= \beta^2 (\frac{\sum_1^N x_i^2}{N} - \bar{x}^2) = \beta^2 \sigma_X^2 = \frac{COV^2(X,Y)}{\sigma_X^2}$$

Pri tem smo upoštevali še drugo formulo, ki nam podaja koeficienta α in β:

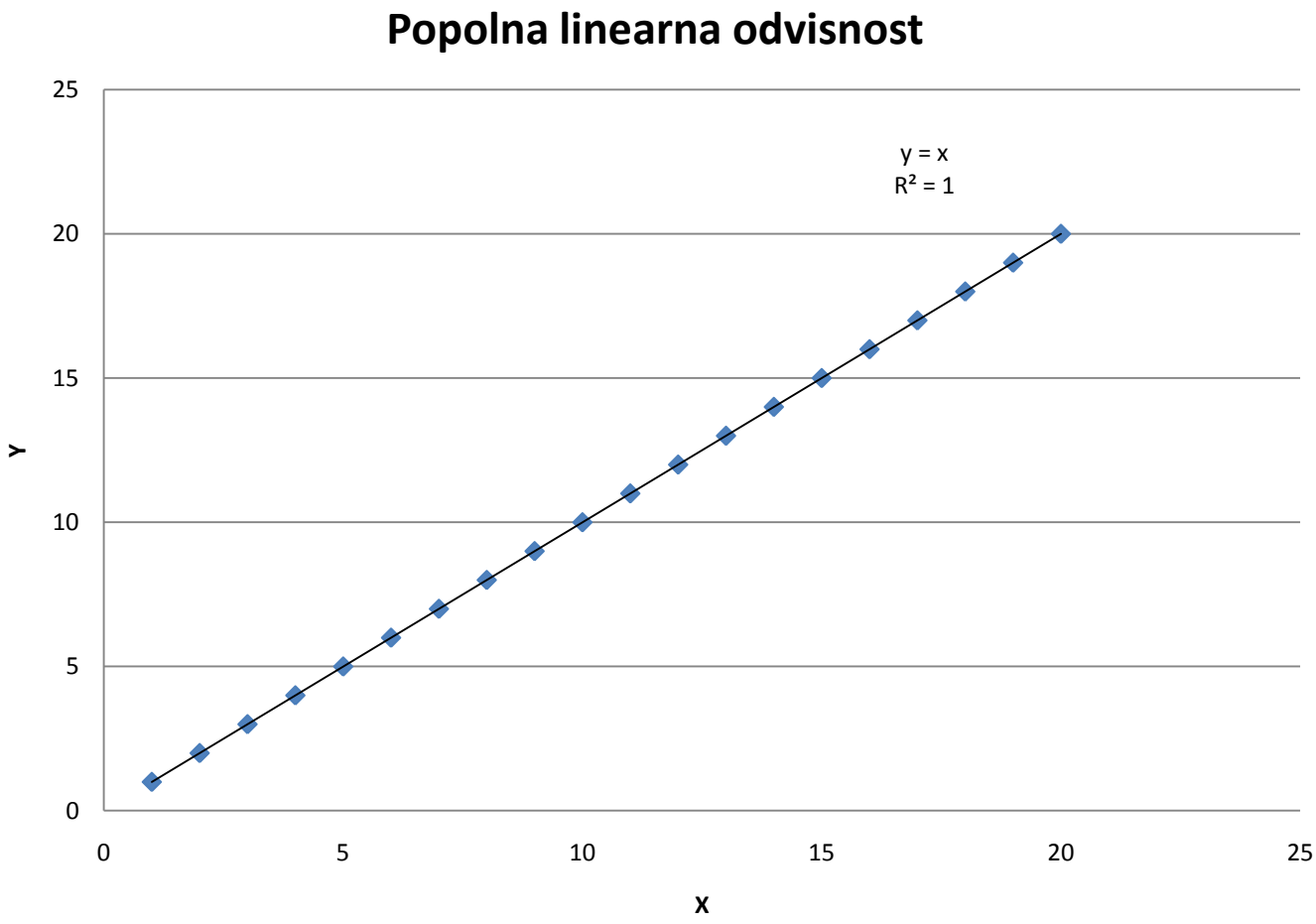
$$\beta = \frac{COV(X,Y)}{\sigma_X^2}$$

Torej imamo: $$\rho^2 = \frac{COV^2(X,Y)}{\sigma_X^2 \sigma_Y^2} = \frac{\sigma_Y^2}{\sigma_Y^2}$$

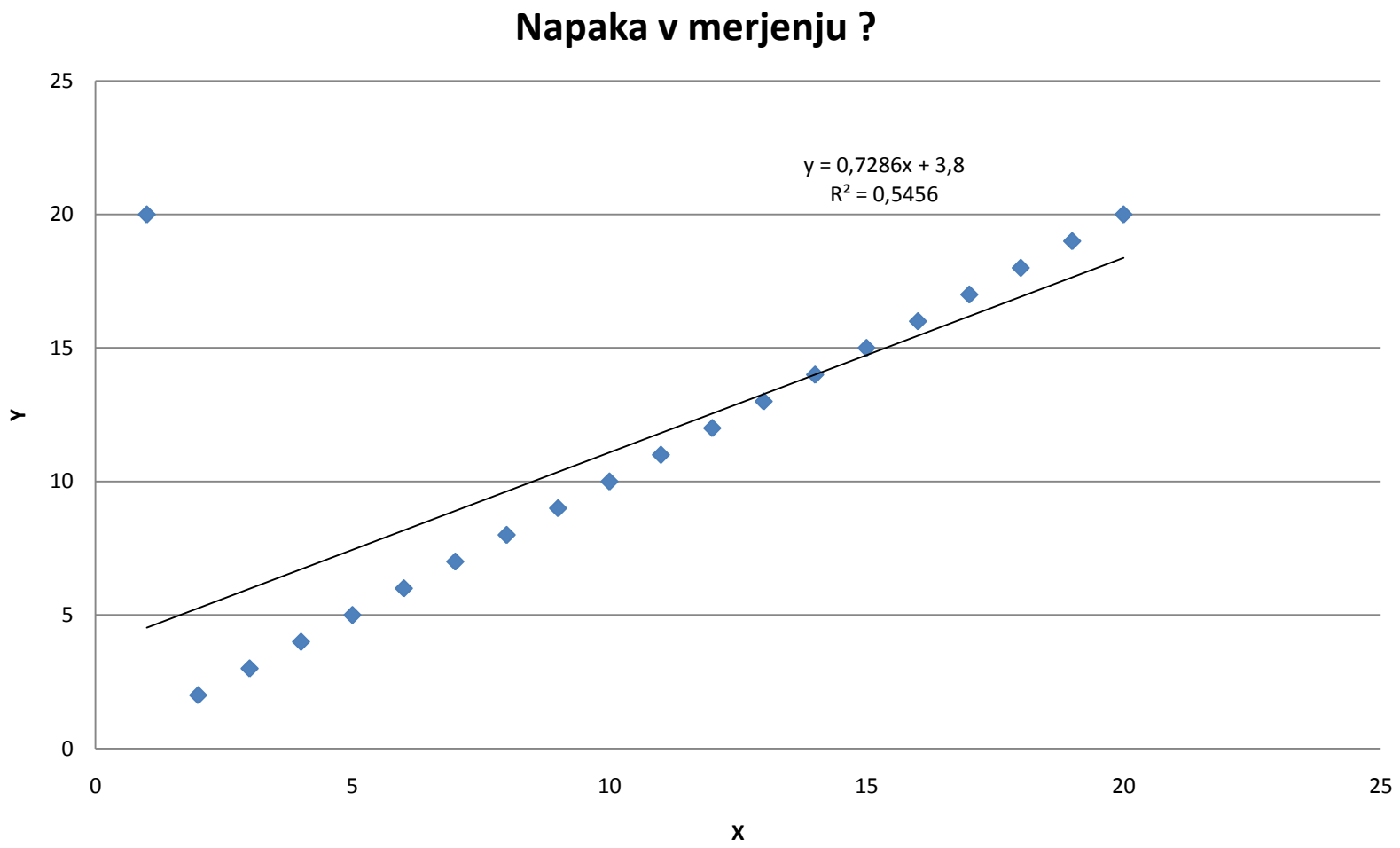
In to smo želeli pokazati.

Za konec si oglejmo še nekaj zgledov.

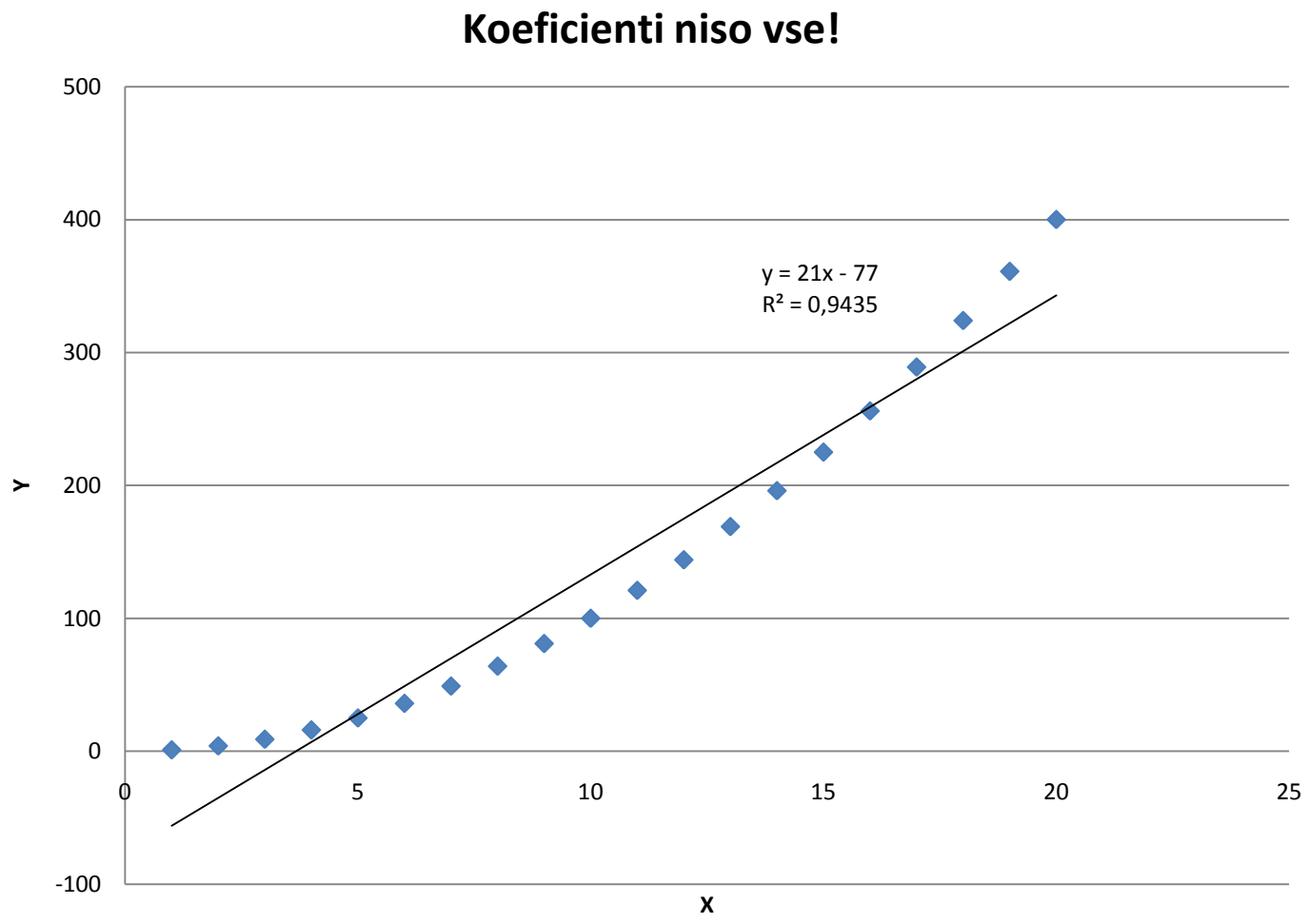
Zgled:



- **Zgled:**


Napaka v merjenju ?
y = 0,7286x + 3,8
R² = 0,5456
25
20
15
10
5
0
0
5
10
15
20
25
y
x

- **Zgled:**


Koeficienti niso vse!
y = 21x - 77
R² = 0,9435
500
400
300
200
100
0
-100
0
5
10
15
20
25
Y
X