Statistika

Vanja Wagner

Link na video predavanje

Predavanje možete pogledati na sljedećem **linku**.

1. Deskriptivna statistika

Ponašanje mjera pri linearnoj transformaciji varijabli

Primjer: Ako varijabli X pribrojimo konstantu (broj), što se događa sa srednjom vrijednosti i varijancom?

Ponašanje mjera pri linearnoj transformaciji varijabli

Primjer: Ako varijabli X pribrojimo konstantu (broj), što se događa sa srednjom vrijednosti i varijancom?

	x_i	$x_i + 10$	$x_i + 20$
	62.30	72.30	82.30
	62.28	72.28	82.28
	69.77	79.77	89.77
	68.95	78.95	88.95
	69.88	79.88	89.88
srednja vrijednost	66.64	76.64	86.64
varijanca uzorka	12.695	12.695	12.695

Ponašanje mjera pri linearnoj transformaciji varijabli

Neka je varijabla Y jednaka varijabli X pomnoženoj s kontantom a i uvećanom za b, odnosno promatramo linearnu transformaciju podataka $x_1, \ldots, x_N$:

$$y_i = ax_i + b.$$

Ponašanje mjera pri linearnoj transformaciji varijabli

Neka je varijabla Y jednaka varijabli X pomnoženoj s kontantom a i uvećanom za b, odnosno promatramo linearnu transformaciju podataka $x_1, \ldots, x_N$:

$$y_i = ax_i + b.$$

Tada je

$$\begin{aligned} \overline{y} &= a\overline{x} + b && (1) \\ \sigma_y^2 &= a^2\sigma_x^2 && (2) \\ \sigma_y &= |a|\sigma_x \end{aligned}$$

gdje je σ_x^2 varijanca uzorka $x_1, \ldots, x_N$, a σ_y^2 varijanca uzorka $y_1, \ldots, y_N$.

Srednja vrijednost i varijanca zbroja dvaju uzoraka

Skakačica	Bodovi 1. skok (X)	Bodovi 2. skok (Y)	Bodovi Ukupno (Z)
TAKANASHI Sara	122.7	120.7	243.4
ITO Yuki	117.0	122.0	239.0
IRASCHKO-STOLZ Daniela	113.3	110.3	223.6
SEIFRIEDSBERGER Jacqueline	114.6	106.6	221.2
RUPPRECHT Anna	107.7	111.5	219.2
PINKELNIG Eva	109.6	108.8	218.4
ALTHAUS Katharina	112.8	102.1	214.9
ROGELJ Spela	105.6	108.6	214.2
CLAIR Julia	108.3	105.9	214.2
AVVAKUMOVA Irina	105.0	108.4	213.4
Srednja vrijednost ($\overline{x}$)	111.66	110.49	222.15
Varijanca (σ^2)	27.53	35.52	101.68
Standardna devijacija (σ)	5.247	5.960	10.084

Srednja vrijednost i varijanca zbroja dvaju uzoraka

Skakačica	Bodovi 1. skok (X)	Bodovi 2. skok (Y)	Bodovi Ukupno (Z)
TAKANASHI Sara	122.7	120.7	243.4
ITO Yuki	117.0	122.0	239.0
IRASCHKO-STOLZ Daniela	113.3	110.3	223.6
SEIFRIEDSBERGER Jacqueline	114.6	106.6	221.2
RUPPRECHT Anna	107.7	111.5	219.2
PINKELNIG Eva	109.6	108.8	218.4
ALTHAUS Katharina	112.8	102.1	214.9
ROGELJ Spela	105.6	108.6	214.2
CLAIR Julia	108.3	105.9	214.2
AVVAKUMOVA Irina	105.0	108.4	213.4
Srednja vrijednost ($\overline{x}$)	111.66	110.49	222.15
Varijanca (σ^2)	27.53	35.52	101.68
Standardna devijacija (σ)	5.247	5.960	10.084

Iz primjera vidimo da se srednje vrijednosti zbrajaju dok to ne vrijedi za varijancu i standardnu devijaciju. Općenito vrijedi:

Ako je

$$z_i = x_i + y_i$$

tada je

$$\overline{z} = \overline{x} + \overline{y}$$

ali

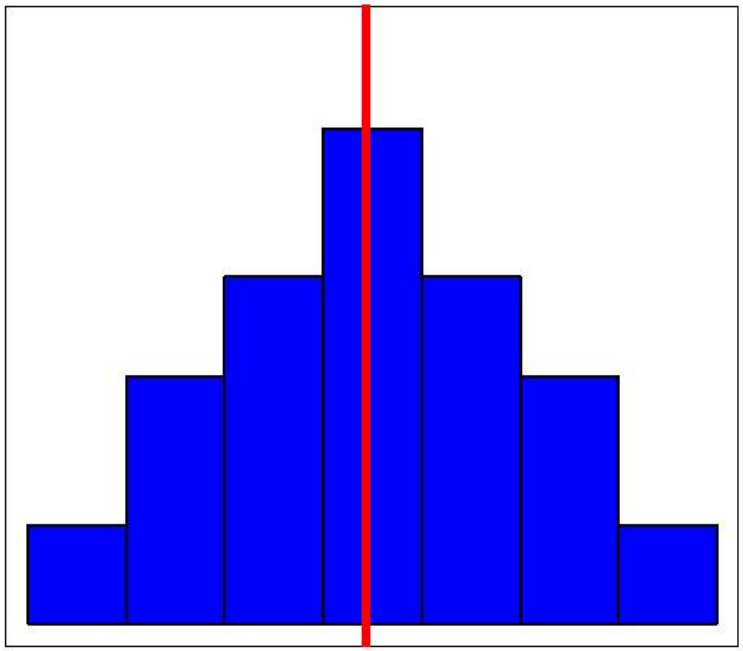
$$\sigma_z^2 \neq \sigma_x^2 + \sigma_y^2.$$

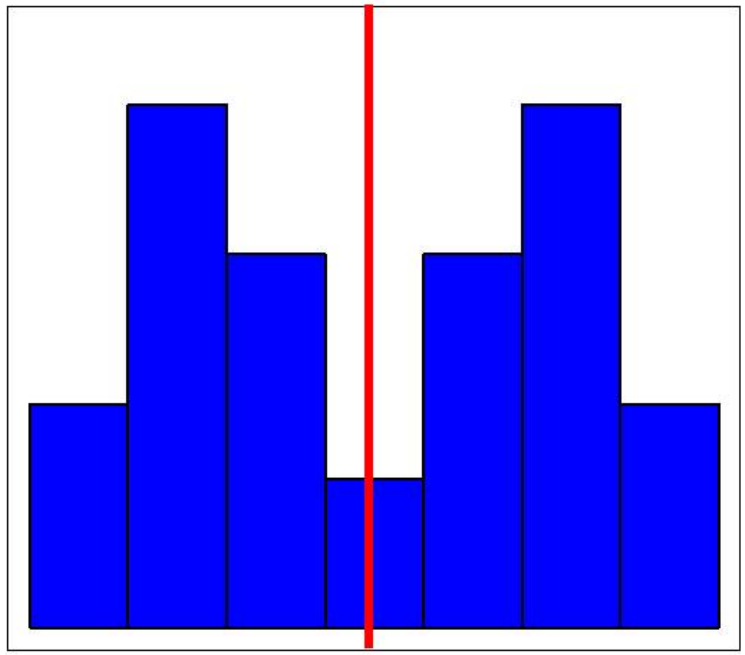
Mjere asimetrije

Često će nas zanimati kojeg je *oblika* razdioba uzorka. Jedno od osnovnih pitanja koje si tada postavljamo je - Je li razdioba simetrična?

Mjere asimetrije

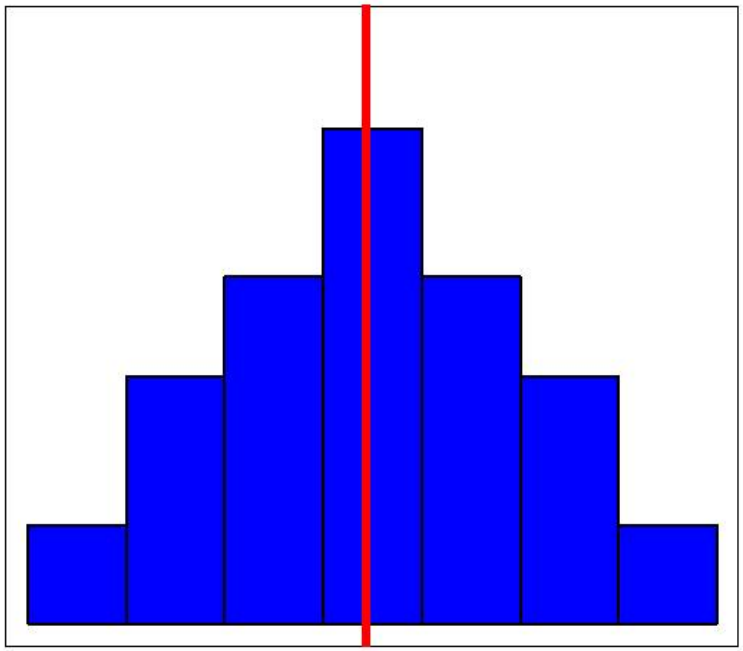
Često će nas zanimati kojeg je *oblika* razdioba uzorka. Jedno od osnovnih pitanja koje si tada postavljamo je - Je li razdioba simetrična?
U tu svrhu možemo promotriti histogram uzorka:

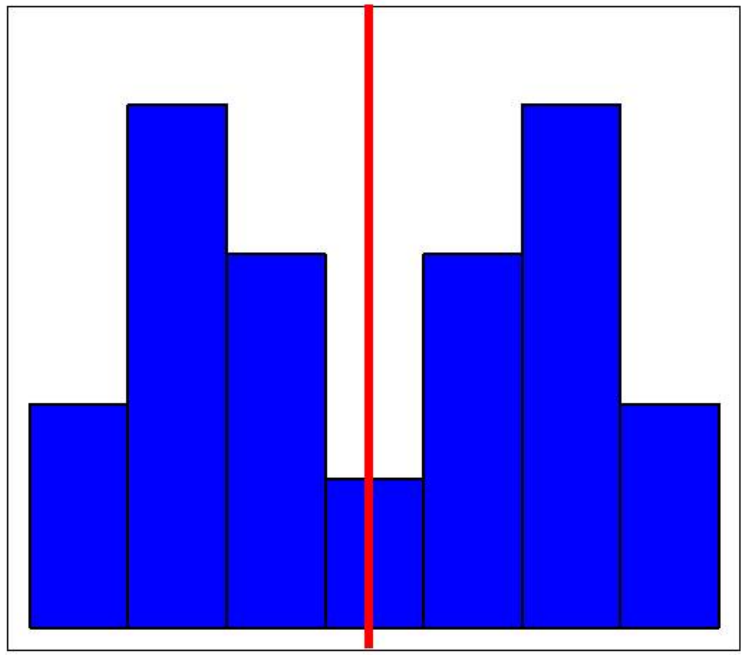




Mjere asimetrije

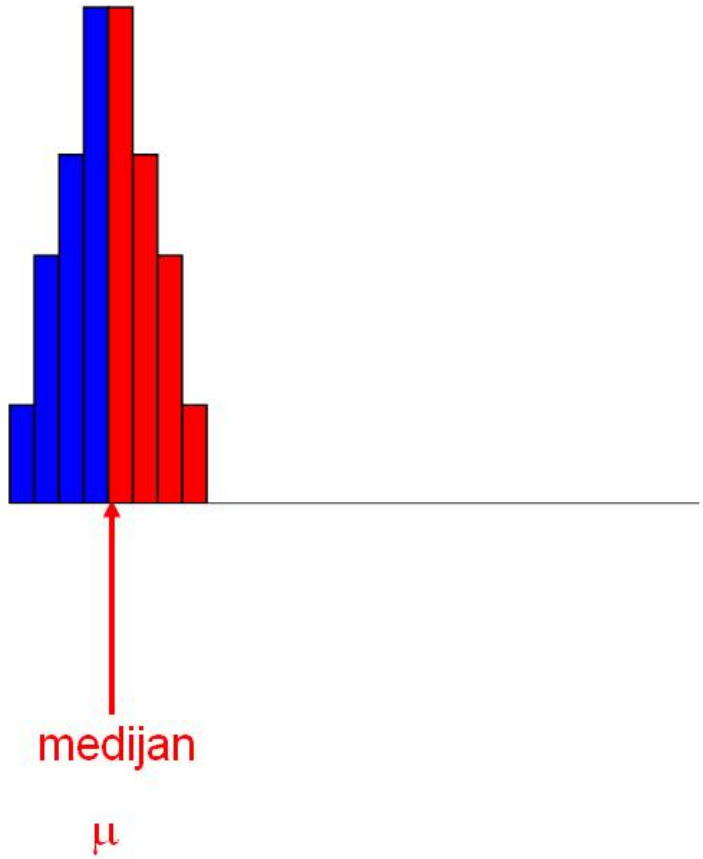
Često će nas zanimati kojeg je *oblika* razdioba uzorka. Jedno od osnovnih pitanja koje si tada postavljamo je - Je li razdioba simetrična?
U tu svrhu možemo promotriti histogram uzorka:



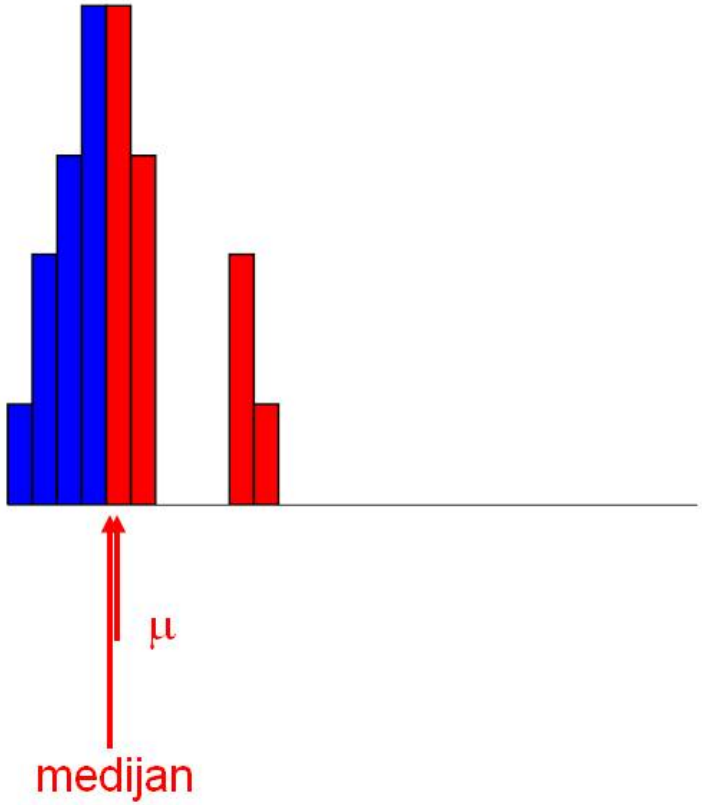


Simetrična razdioba!

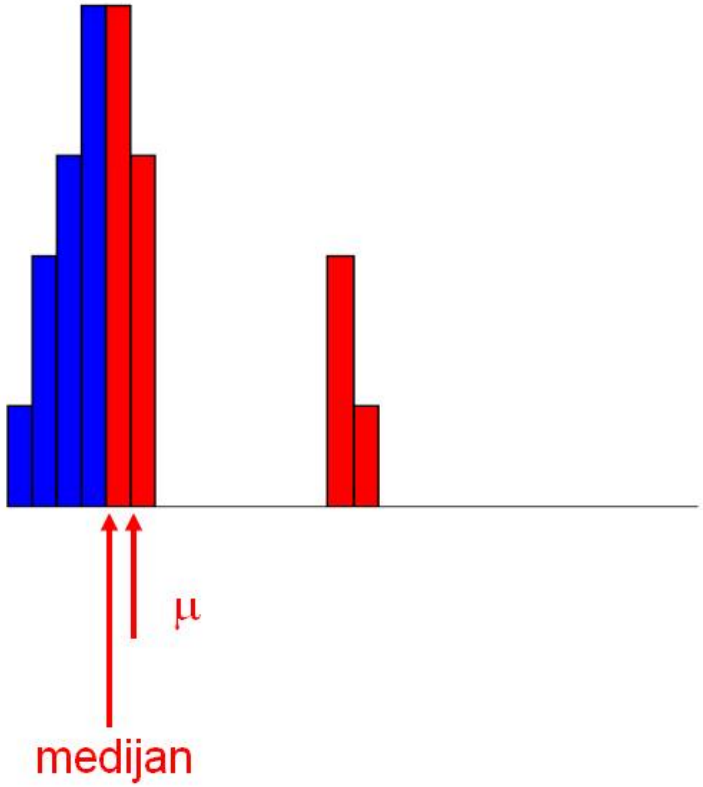
Što se događa sa srednjom vrijednošću i medijanom kada razdioba nije simetrična?



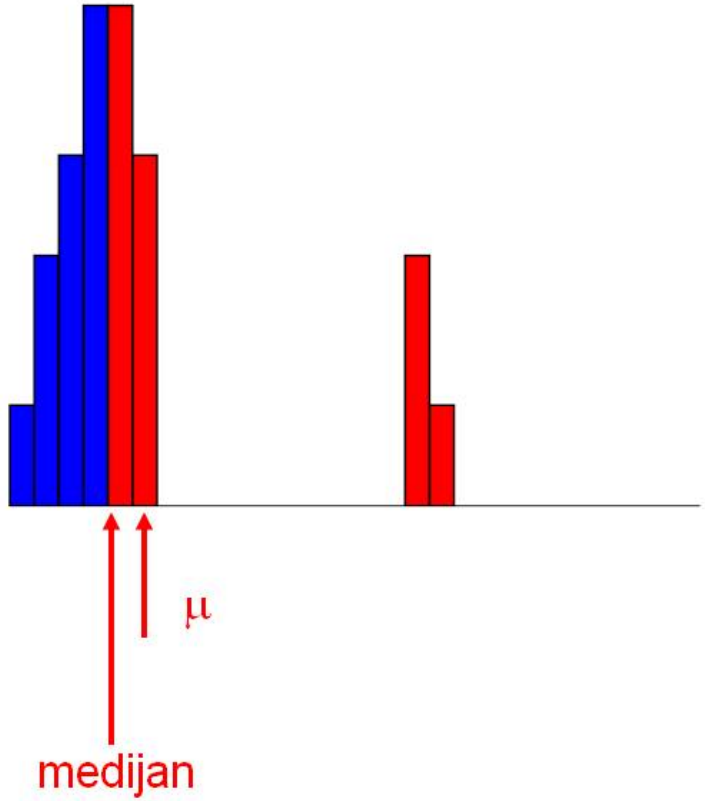
Što se događa sa srednjom vrijednošću i medijanom kada razdioba nije simetrična?



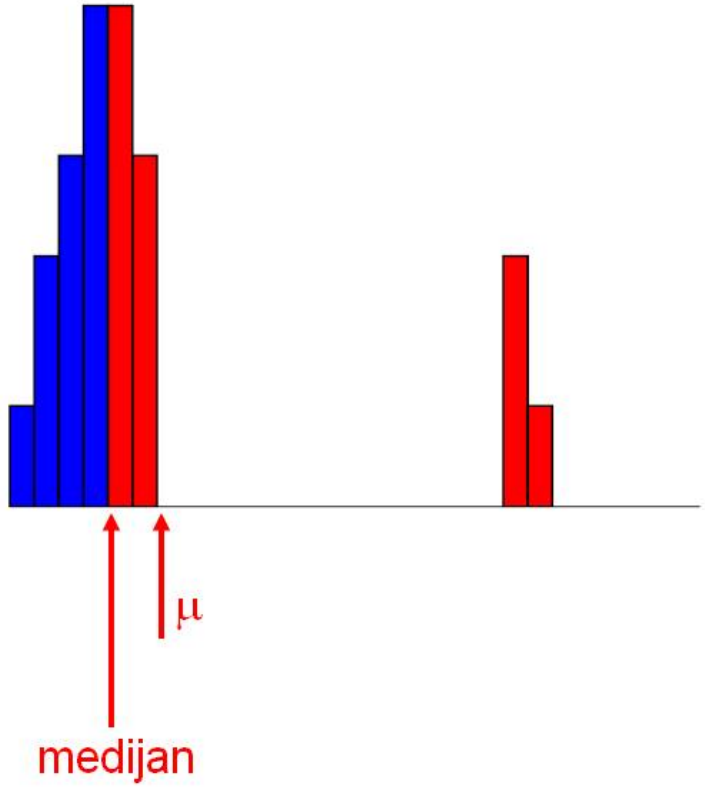
Što se događa sa srednjom vrijednošću i medijanom kada razdioba nije simetrična?



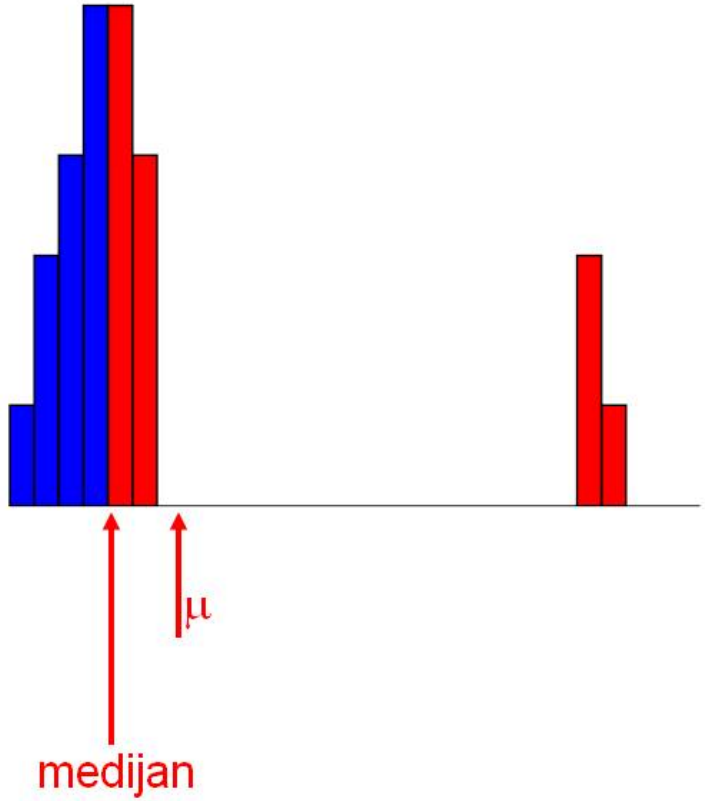
Što se događa sa srednjom vrijednošću i medijanom kada razdioba nije simetrična?



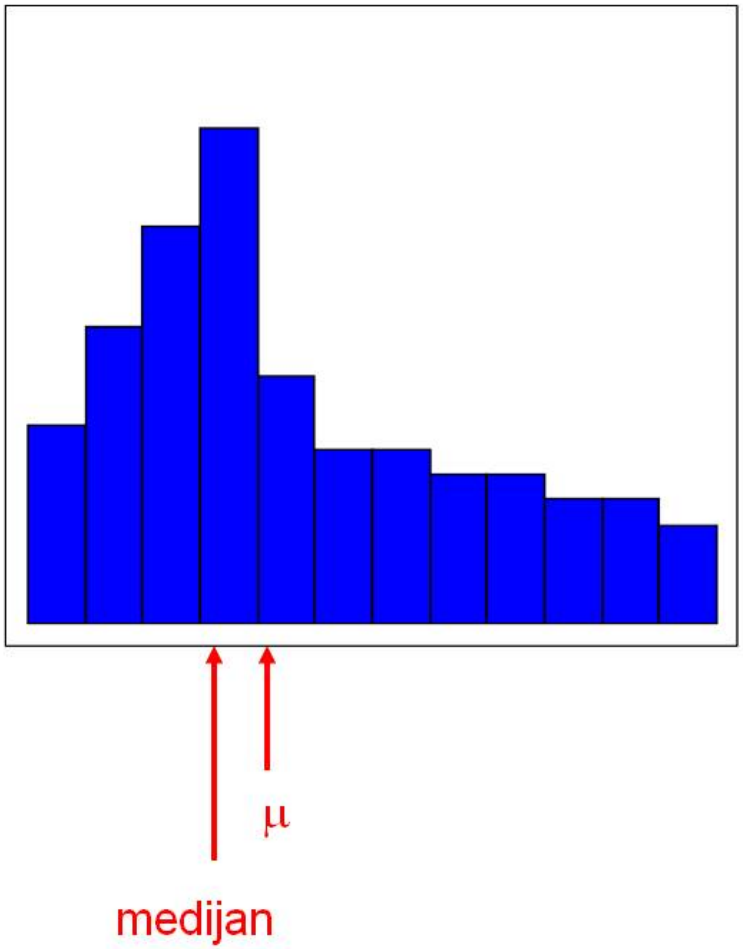
Što se događa sa srednjom vrijednošću i medijanom kada razdioba nije simetrična?

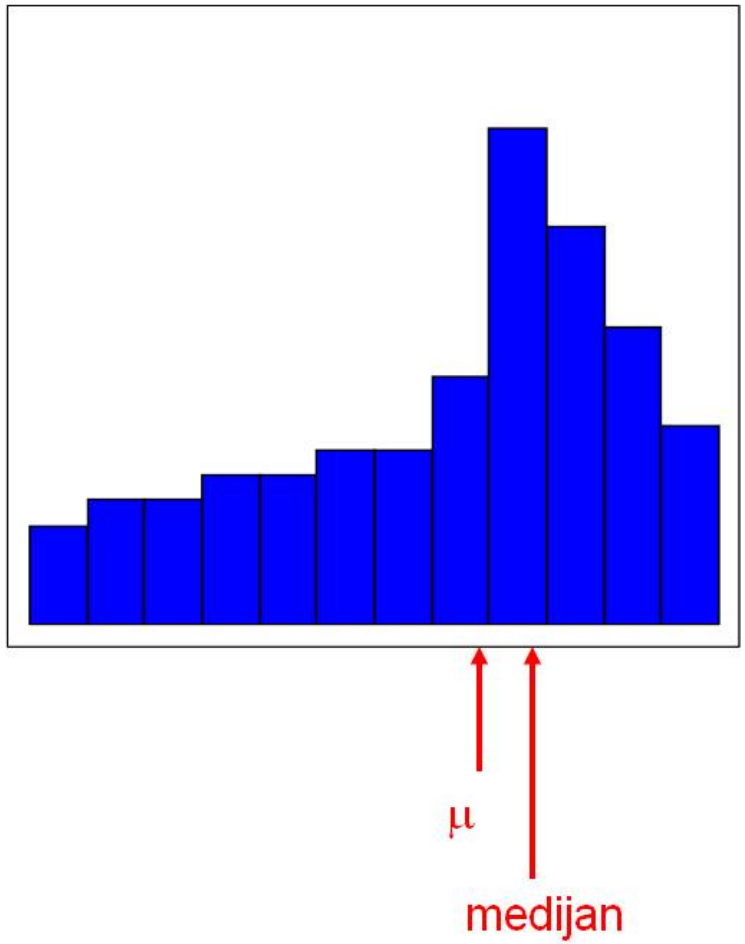


Što se događa sa srednjom vrijednošću i medijanom kada razdioba nije simetrična?

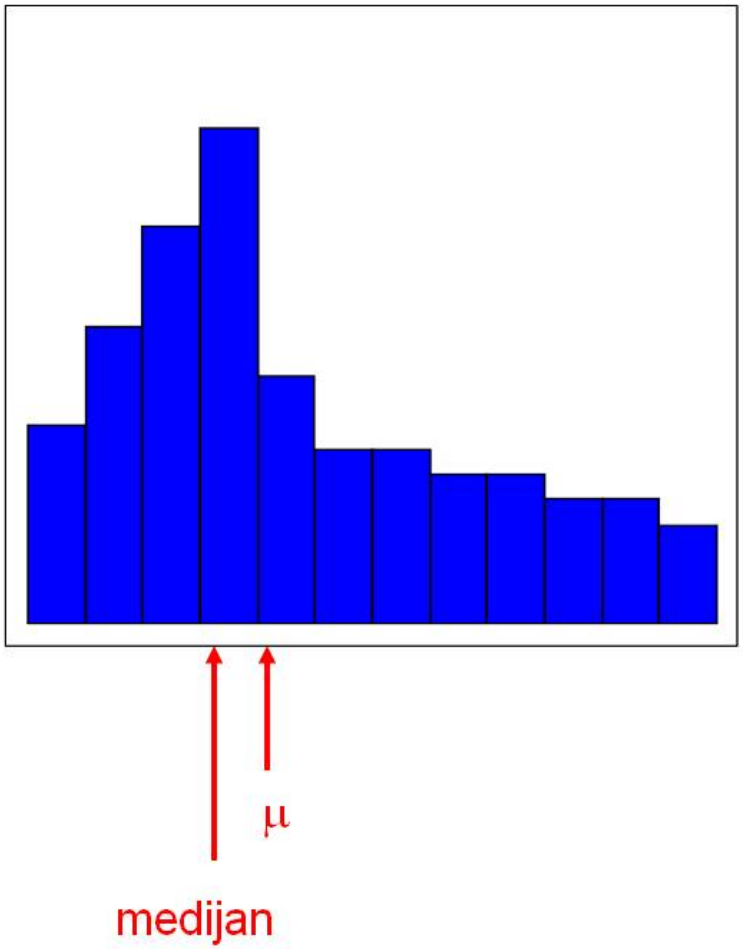


Asimetrične razdiobe:

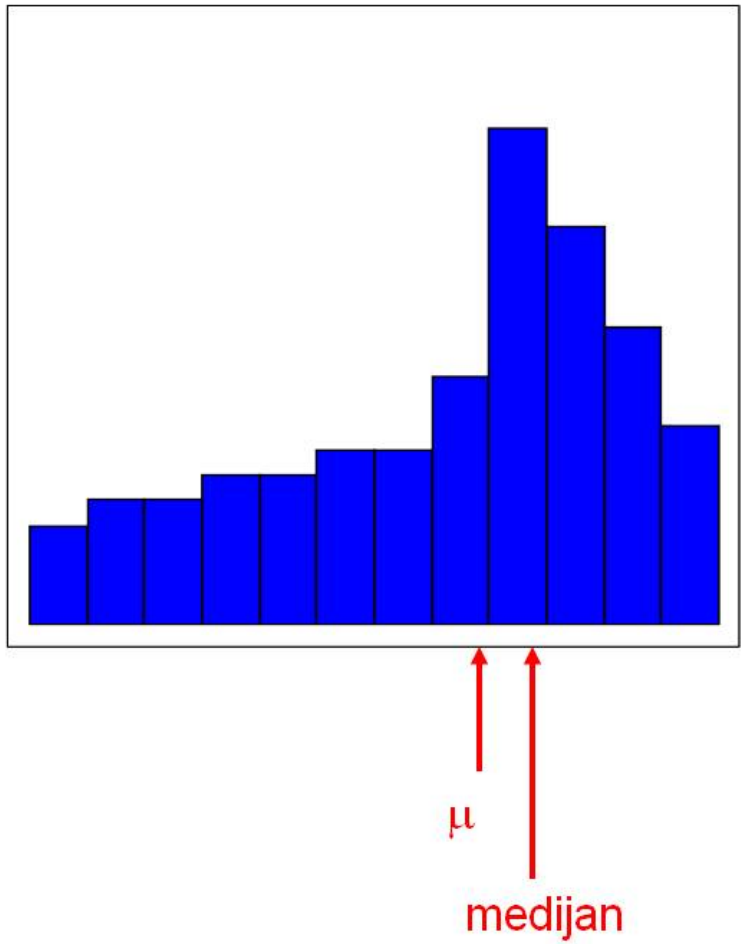




Asimetrične razdiobe:

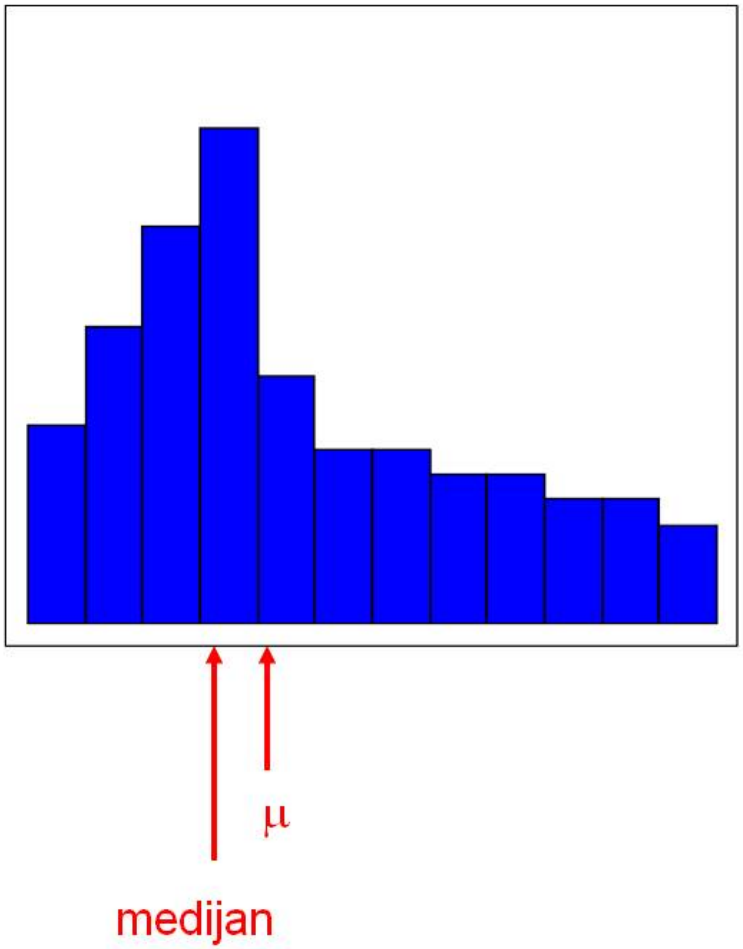


Pozitivno zakošena

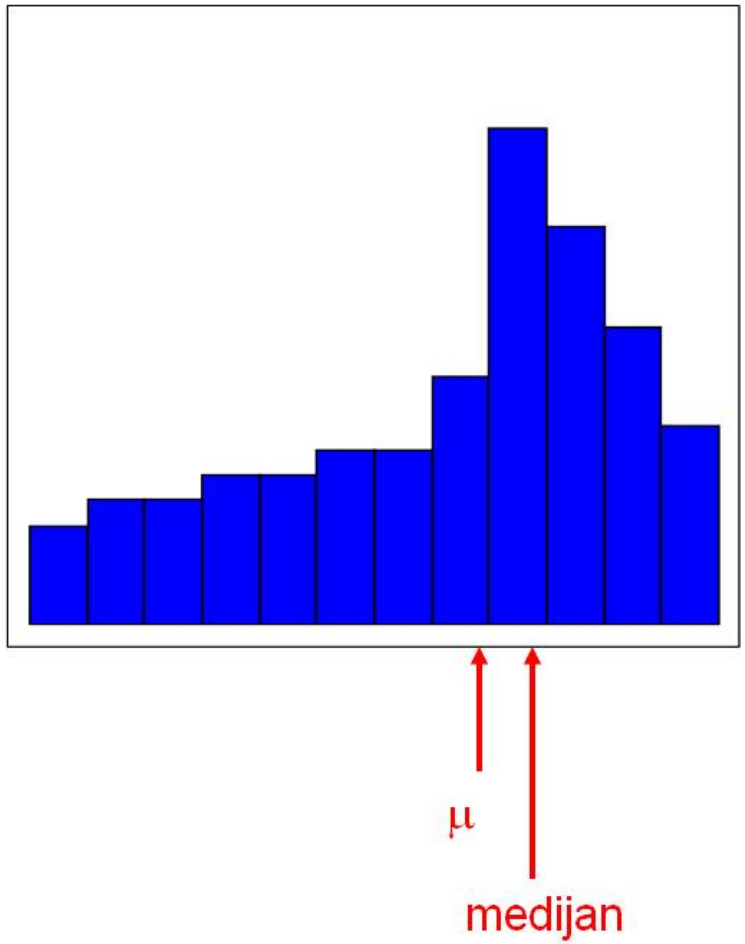


Negativno zakošena

Asimetrične razdiobe:



Pozitivno zakošena



Negativno zakošena

- Asimetrija ili nakošenost (*skewness*)

Mjere asimetrije

Promatrat ćemo dvije mjere asimetrije, a prva od njih dvije uzet će u obzir odnos srednje vrijednosti i medijana koji smo promotrili u prethodnom primjeru:

[1] O standardiziranom uzorku ćemo govoriti detaljnije nešto kasnije tijekom ovog predavanja.

Mjere asimetrije

Promatrat ćemo dvije mjere asimetrije, a prva od njih dvije uzet će u obzir odnos srednje vrijednosti i medijana koji smo promotrili u prethodnom primjeru:
Pearsonov koeficijent asimetrije je vrijednost:

$$Sk_P = 3\frac{\overline{x} - m}{\sigma},$$

gdje je $\overline{x}$ srednja vrijednost, m medijan i σ standardna devijacija promatranih podataka.

[1] O standardiziranom uzorku ćemo govoriti detaljnije nešto kasnije tijekom ovog predavanja.

Mjere asimetrije

Promatrat ćemo dvije mjere asimetrije, a prva od njih dvije uzet će u obzir odnos srednje vrijednosti i medijana koji smo promotrili u prethodnom primjeru:
Pearsonov koeficijent asimetrije je vrijednost:

$$Sk_P = 3\frac{\overline{x} - m}{\sigma},$$

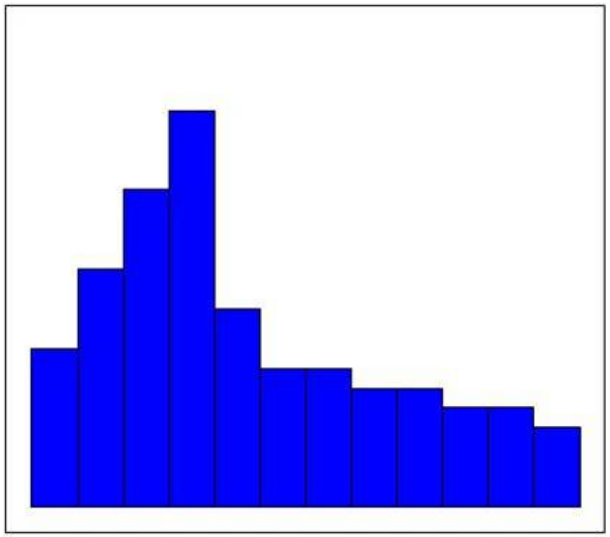
gdje je $\overline{x}$ srednja vrijednost, m medijan i σ standardna devijacija promatranih podataka.

Druga mjera koju ćemo jednostavno zvati **koeficijent asimetrije** uzorka (*skewness*) zapravo predstavlja 3. moment standardiziranog uzorka[1]:

$$skew(X) = \frac{1}{N}\sum_i \left(\frac{x_i - \overline{x}}{\sigma}\right)^3.$$

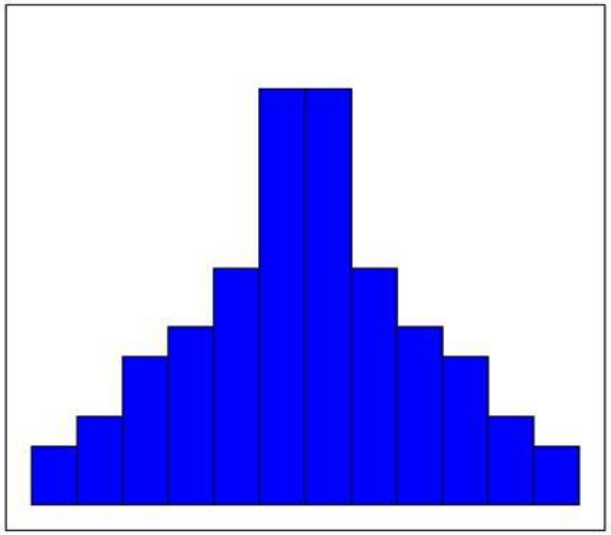
[1] O standardiziranom uzorku ćemo govoriti detaljnije nešto kasnije tijekom ovog predavanja.

Primjer.



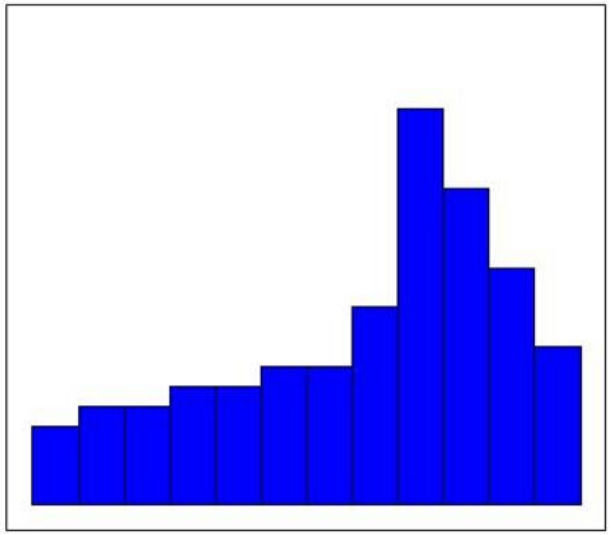
Zakošenost = 0.64

$Sk_P = 1.22$



Zakošenost = 0

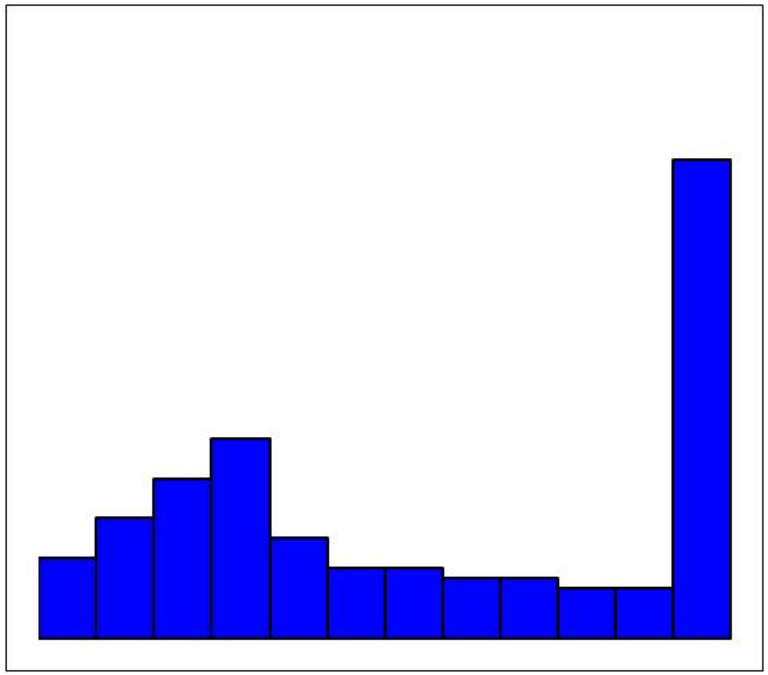
$Sk_P = 0$



Zakošenost = -0.64

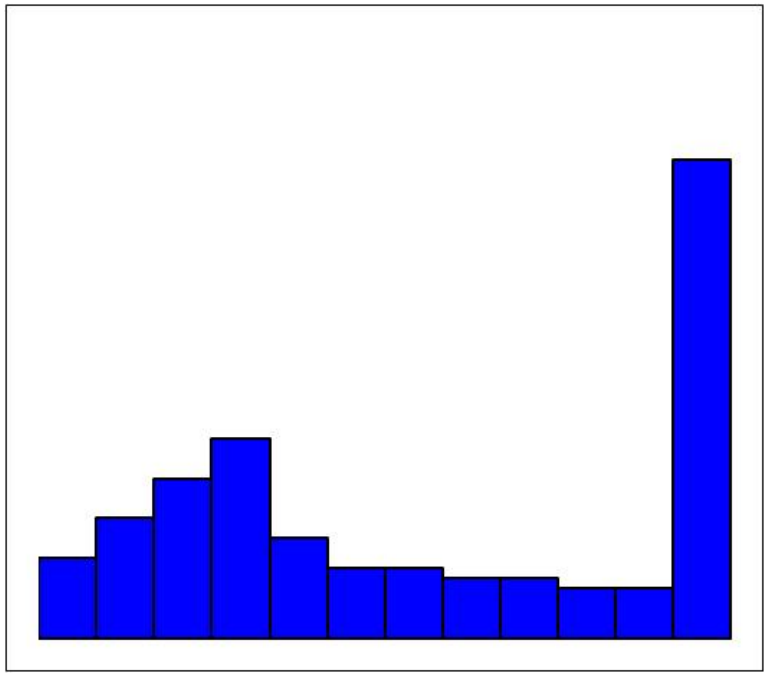
$Sk_P = -1.22$

Oprez!



Zakošenost $= 0$

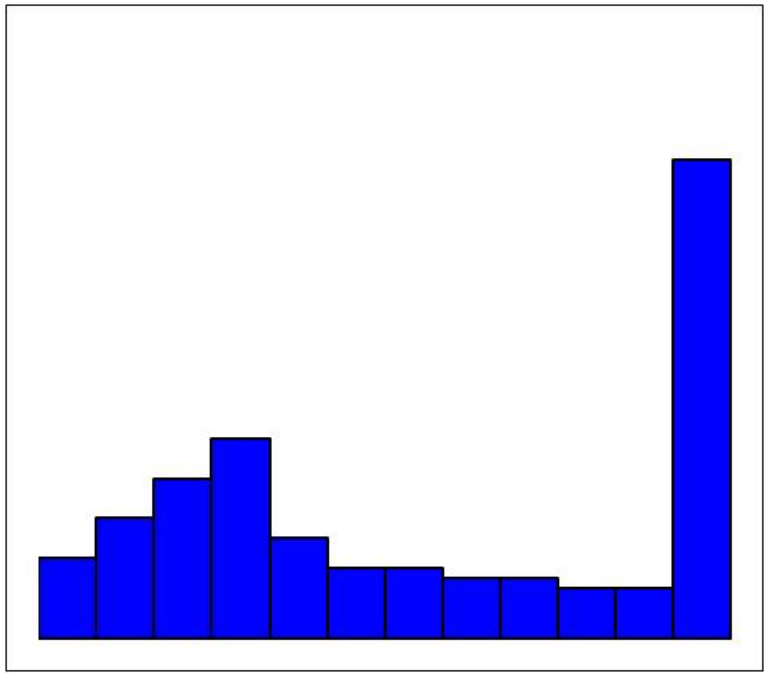
Oprez!



Zakošenost $= 0$

Bimodalna distribucija!

Oprez!

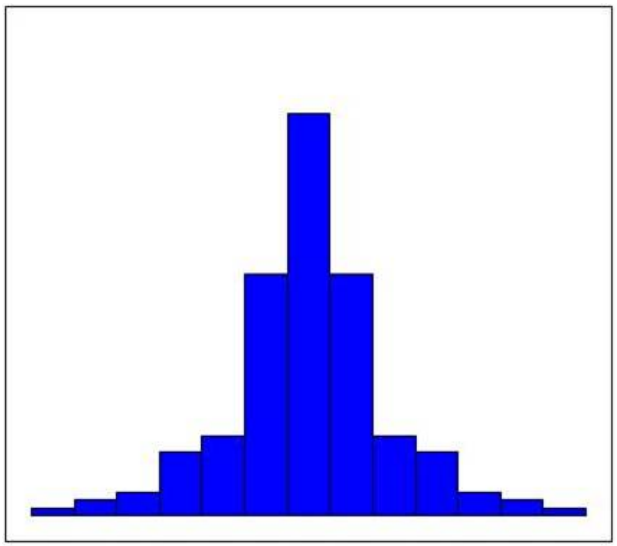


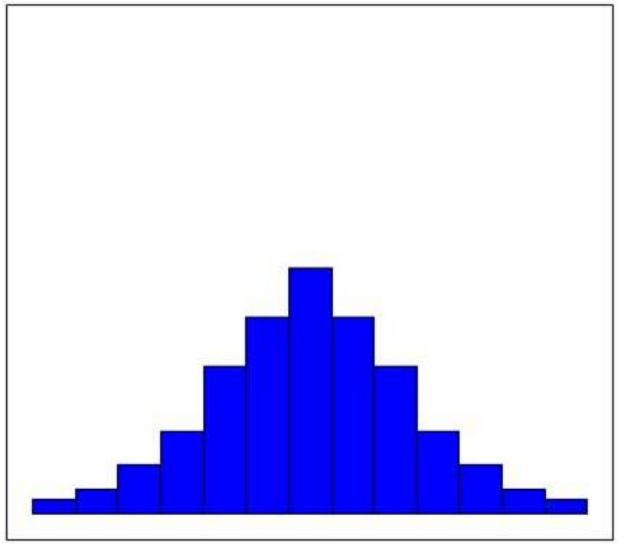
Zakošenost $= 0$

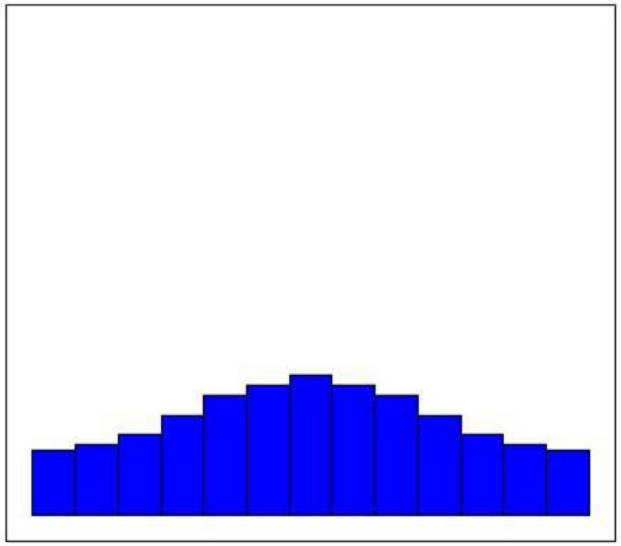
Bimodalna distribucija!

Zakošenost je dobar pokazatelj asimetrije za unimodalne distribucije.

Spljoštenost







Koeficijent spljoštenosti

Definiramo koeficijent spljoštenosti

$$kurt(X) = \frac{1}{N} \sum_i \left(\frac{x_i - \overline{x}}{\sigma} \right)^4,$$

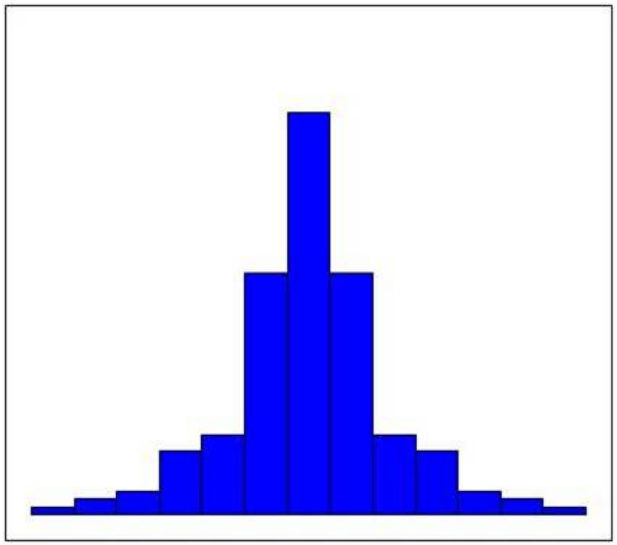
gdje je $\overline{x}$ srednja vrijednost iσ standardna devijacija naših podataka $x_1, \ldots, x_N$.

Koeficijent spljoštenosti

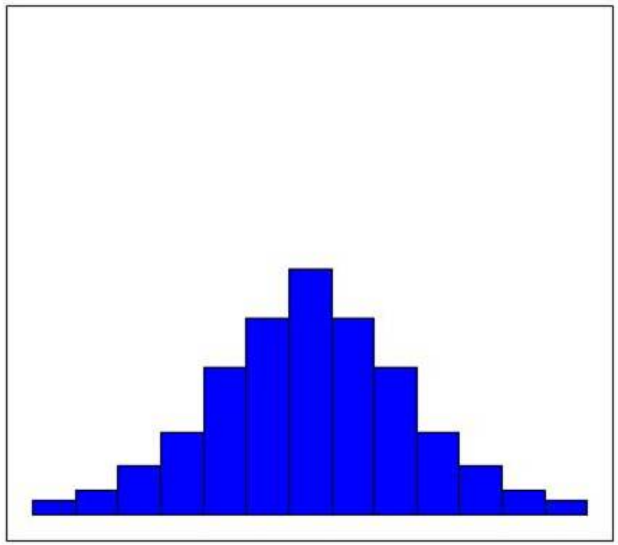
Definiramo koeficijent spljoštenosti

$$kurt(X) = \frac{1}{N}\sum_i \left(\frac{x_i - \overline{x}}{\sigma}\right)^4,$$

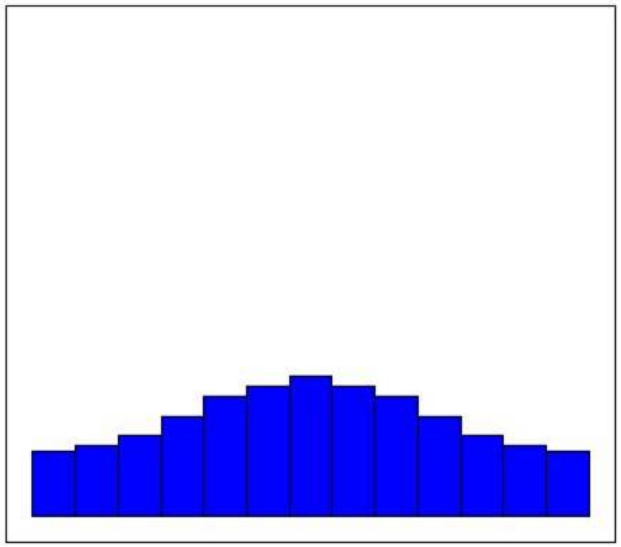
gdje je $\overline{x}$ srednja vrijednost iσ standardna devijacija naših podataka $x_1, \ldots, x_N$.



$kurt(X) = 4.31$



$kurt(X) = 3$



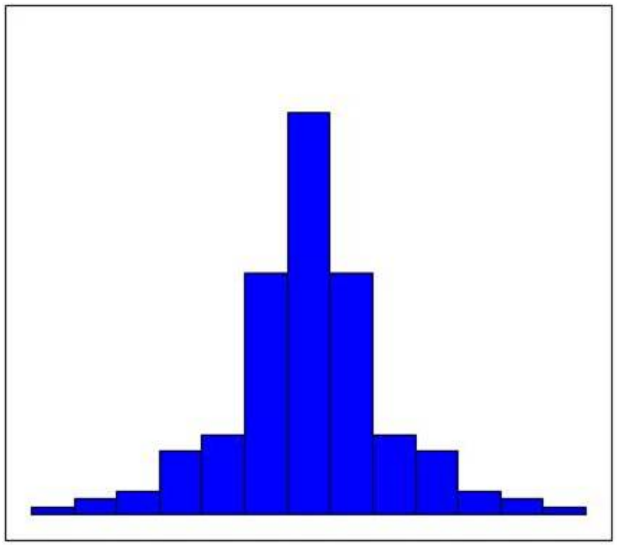
$kurt(X) = 2.13$

Koeficijent spljoštenosti

Definiramo koeficijent spljoštenosti

$$kurt(X) = \frac{1}{N}\sum_i \left(\frac{x_i - \overline{x}}{\sigma}\right)^4,$$

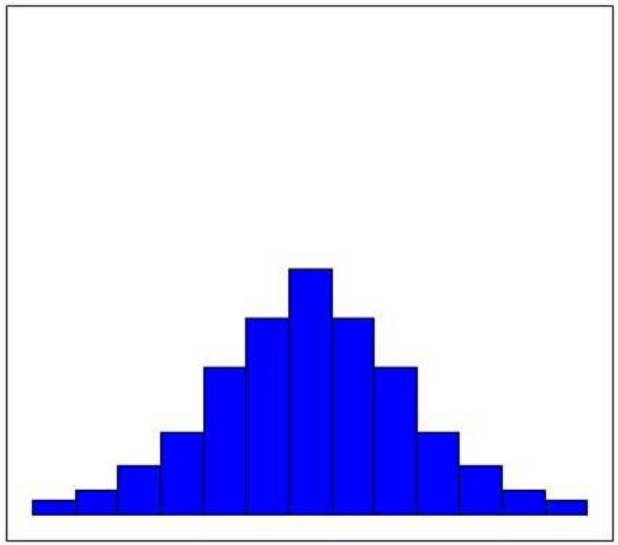
gdje je $\overline{x}$ srednja vrijednost iσ standardna devijacija naših podataka $x_1, \ldots, x_N$.



$kurt(X) = 4.31$

$kurt(X) > 3$

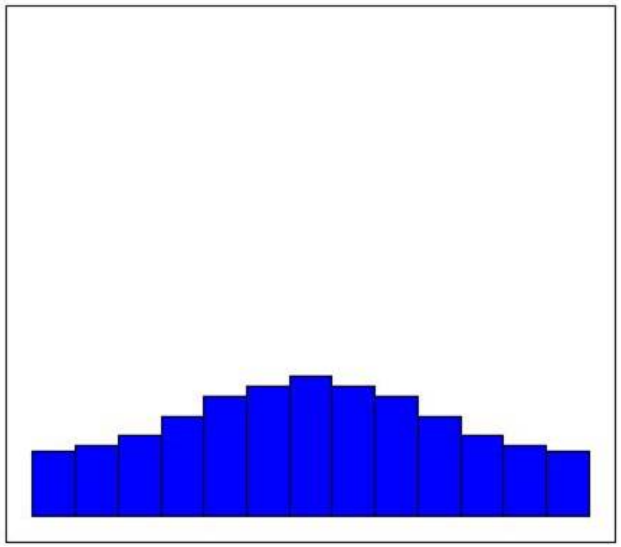
Leptokurtična (izbočena)



$kurt(X) = 3$

$kurt(X) = 3$

Mezokurtična distribucija



$kurt(X) = 2.13$

$kurt(X) < 3$

Platikurtična (spljoštena)

Položaj podataka - rang

Rang je položaj (redni broj) podatka u nizu podataka.

Ukoliko dva ili više podataka imaju istu vrijednost, njihov rang je aritmetička sredina rangova koje bi imali da su vrijednosti različite.

Položaj podataka - rang

Rang je položaj (redni broj) podatka u nizu podataka.

Ukoliko dva ili više podataka imaju istu vrijednost, njihov rang je aritmetička sredina rangova koje bi imali da su vrijednosti različite.

Primjer. Odredite rangove za zadani skup podataka:

15 2 21 13 8 17 15 8 15.

Položaj podataka - rang

Rang je položaj (redni broj) podatka u nizu podataka.

Ukoliko dva ili više podataka imaju istu vrijednost, njihov rang je aritmetička sredina rangova koje bi imali da su vrijednosti različite.

Primjer. Odredite rangove za zadani skup podataka:

$$15 \quad 2 \quad 21 \quad 13 \quad 8 \quad 17 \quad 15 \quad 8 \quad 15.$$

Rješenje. Uredimo niz podataka od najmanjeg do najvećeg:

$$2 \quad 8 \quad 8 \quad 13 \quad 15 \quad 15 \quad 15 \quad 17 \quad 21,$$

Položaj podataka - rang

Rang je položaj (redni broj) podatka u nizu podataka.

Ukoliko dva ili više podataka imaju istu vrijednost, njihov rang je aritmetička sredina rangova koje bi imali da su vrijednosti različite.

Primjer. Odredite rangove za zadani skup podataka:

$$15 \quad 2 \quad 21 \quad 13 \quad 8 \quad 17 \quad 15 \quad 8 \quad 15.$$

Rješenje. Uredimo niz podataka od najmanjeg do najvećeg:

$$2 \quad 8 \quad 8 \quad 13 \quad 15 \quad 15 \quad 15 \quad 17 \quad 21,$$

a zatim im pridružimo rang:

Položaj podataka - rang

Rang je položaj (redni broj) podatka u nizu podataka.

Ukoliko dva ili više podataka imaju istu vrijednost, njihov rang je aritmetička sredina rangova koje bi imali da su vrijednosti različite.

Primjer. Odredite rangove za zadani skup podataka:

$$15 \quad 2 \quad 21 \quad 13 \quad 8 \quad 17 \quad 15 \quad 8 \quad 15.$$

Rješenje. Uredimo niz podataka od najmanjeg do najvećeg:

$$2 \quad 8 \quad 8 \quad 13 \quad 15 \quad 15 \quad 15 \quad 17 \quad 21,$$

a zatim im pridružimo rang:

Vrijednost	2	8	8	13	15	15	15	17	21
	1	2	3	4	5	6	7	8	9

Položaj podataka - rang

Rang je položaj (redni broj) podatka u nizu podataka.

Ukoliko dva ili više podataka imaju istu vrijednost, njihov rang je aritmetička sredina rangova koje bi imali da su vrijednosti različite.

Primjer. Odredite rangove za zadani skup podataka:

$$15 \quad 2 \quad 21 \quad 13 \quad 8 \quad 17 \quad 15 \quad 8 \quad 15.$$

Rješenje. Uredimo niz podataka od najmanjeg do najvećeg:

$$2 \quad 8 \quad 8 \quad 13 \quad 15 \quad 15 \quad 15 \quad 17 \quad 21,$$

a zatim im pridružimo rang:

Vrijednost	2	8	8	13	15	15	15	17	21
	1	2	3	4	5	6	7	8	9
Rang	1	2.5	2.5	4	6	6	6	8	9

Položaj podataka - rang

Rang je položaj (redni broj) podatka u nizu podataka.

Ukoliko dva ili više podataka imaju istu vrijednost, njihov rang je aritmetička sredina rangova koje bi imali da su vrijednosti različite.

Primjer. Odredite rangove za zadani skup podataka:

$$15 \quad 2 \quad 21 \quad 13 \quad 8 \quad 17 \quad 15 \quad 8 \quad 15.$$

Rješenje. Uredimo niz podataka od najmanjeg do najvećeg:

$$2 \quad 8 \quad 8 \quad 13 \quad 15 \quad 15 \quad 15 \quad 17 \quad 21,$$

a zatim im pridružimo rang:

Vrijednost	2	8	8	13	15	15	15	17	21
	1	2	3	4	5	6	7	8	9
Rang	1	2.5	2.5	4	6	6	6	8	9

Npr. rang za vrijednost 8 je $(2+3)/2 = 2.5$ a za 15: $(5+6+7)/3=6$.

Standardizirani uzorak

Za uzorak $x_1, \ldots, x_N$ čija je srednja vrijednost $\overline{x}$ i standardna devijacija σ, **standardizirani uzorak** (z-vrijednost) $z_1, \ldots, z_N$ je

$$z_i = \frac{x_i - \overline{x}}{\sigma}.$$

Standardizirana varijabla pokazuje koliko je standardnih devijacija vrijednost podatka udaljena od srednje vrijednosti uzorka.

Standardizirani uzorak

Za uzorak $x_1, \ldots, x_N$ čija je srednja vrijednost $\overline{x}$ i standardna devijacija σ, **standardizirani uzorak** (z-vrijednost) $z_1, \ldots, z_N$ je

$$z_i = \frac{x_i - \overline{x}}{\sigma}.$$

Standardizirana varijabla pokazuje koliko je standardnih devijacija vrijednost podatka udaljena od srednje vrijednosti uzorka.
Uočimo da standardizirani uzorak zadovoljava:

$$\overline{z} = 0 \quad \text{i} \quad \sigma_z = 1.$$

Standardizirani uzorak

Za uzorak $x_1, \ldots, x_N$ čija je srednja vrijednost $\overline{x}$ i standardna devijacija σ, **standardizirani uzorak** (z-vrijednost) $z_1, \ldots, z_N$ je

$$z_i = \frac{x_i - \overline{x}}{\sigma}.$$

Standardizirana varijabla pokazuje koliko je standardnih devijacija vrijednost podatka udaljena od srednje vrijednosti uzorka.
Uočimo da standardizirani uzorak zadovoljava:

$$\overline{z} = 0 \quad \text{i} \quad \sigma_z = 1.$$

Naime, standardiziranjem zapravo linearno transformiramo podatke:

$$\overline{z_i} = \frac{1}{\sigma} x_i - \frac{\overline{x}}{\sigma},$$

pa primjenom formula (1) i (2) za $a = \frac{1}{\sigma}$ i $b = -\frac{\overline{x}}{\sigma}$ dobivamo

Standardizirani uzorak

Za uzorak $x_1, \ldots, x_N$ čija je srednja vrijednost $\overline{x}$ i standardna devijacija σ, **standardizirani uzorak** (z-vrijednost) $z_1, \ldots, z_N$ je

$$z_i = \frac{x_i - \overline{x}}{\sigma}.$$

Standardizirana varijabla pokazuje koliko je standardnih devijacija vrijednost podatka udaljena od srednje vrijednosti uzorka.
Uočimo da standardizirani uzorak zadovoljava:

$$\overline{z} = 0 \quad \text{i} \quad \sigma_z = 1.$$

Naime, standardiziranjem zapravo linearno transformiramo podatke:

$$\overline{z_i} = \frac{1}{\sigma} x_i - \frac{\overline{x}}{\sigma},$$

pa primjenom formula (1) i (2) za $a = \frac{1}{\sigma}$ i $b = -\frac{\overline{x}}{\sigma}$ dobivamo

$$\overline{z} = a\overline{x} + b = \frac{1}{\sigma}\overline{x} - \frac{\overline{x}}{\sigma} = 0$$

Standardizirani uzorak

Za uzorak $x_1, \ldots, x_N$ čija je srednja vrijednost $\overline{x}$ i standardna devijacija σ, **standardizirani uzorak** (z-vrijednost) $z_1, \ldots, z_N$ je

$$z_i = \frac{x_i - \overline{x}}{\sigma}.$$

Standardizirana varijabla pokazuje koliko je standardnih devijacija vrijednost podatka udaljena od srednje vrijednosti uzorka.
Uočimo da standardizirani uzorak zadovoljava:

$$\overline{z} = 0 \quad \text{i} \quad \sigma_z = 1.$$

Naime, standardiziranjem zapravo linearno transformiramo podatke:

$$\overline{z_i} = \frac{1}{\sigma} x_i - \frac{\overline{x}}{\sigma},$$

pa primjenom formula (1) i (2) za $a = \frac{1}{\sigma}$ i $b = -\frac{\overline{x}}{\sigma}$ dobivamo

$$\overline{z} = a\overline{x} + b = \frac{1}{\sigma}\overline{x} - \frac{\overline{x}}{\sigma} = 0$$

i

$$\sigma_Z = |a|\sigma = \frac{1}{\sigma}\sigma = 1.$$

Čebiševljeva nejednakost

Promotrimo uzorak $x_1, \ldots, x_N$. Možemo se zapitati:

Koliko elemenata uzorka očekujemo da se nađe u intervalu $[\overline{x} - k\sigma, \overline{x} + k\sigma]$, za $k > 1$?

Čebiševljeva nejednakost

Promotrimo uzorak $x_1, \ldots, x_N$. Možemo se zapitati:

Koliko elemenata uzorka očekujemo da se nađe u intervalu $[\overline{x} - k\sigma, \overline{x} + k\sigma]$, za $k > 1$?

To pitanje je ekvivalentno pitanju:

Koliko elemenata pripadnog standardiziranog uzorka $z_1, \ldots, z_N$ očekikujemo da bude u $[-k, k]$?

Čebiševljeva nejednakost

Promotrimo uzorak $x_1, \ldots, x_N$. Možemo se zapitati:

Koliko elemenata uzorka očekujemo da se nađe u intervalu $[\overline{x} - k\sigma, \overline{x} + k\sigma]$, za $k > 1$?

To pitanje je ekvivalentno pitanju:

Koliko elemenata pripadnog standardiziranog uzorka $z_1, \ldots, z_N$ očekikujemo da bude u $[-k, k]$?

Ta vrijednost se može aproksimirati korištenjem nekih vjerojatnosnih principa. Korištenjem **Čebiševljeve nejednakosti** može se pokazati da se između $\overline{x} - k\sigma$ i $\overline{x} + k\sigma$ približno nalazi najmanje

$$\left(1 - \frac{1}{k^2}\right) \cdot 100\%$$

članova uzorka $x_1, \ldots, x_N$.

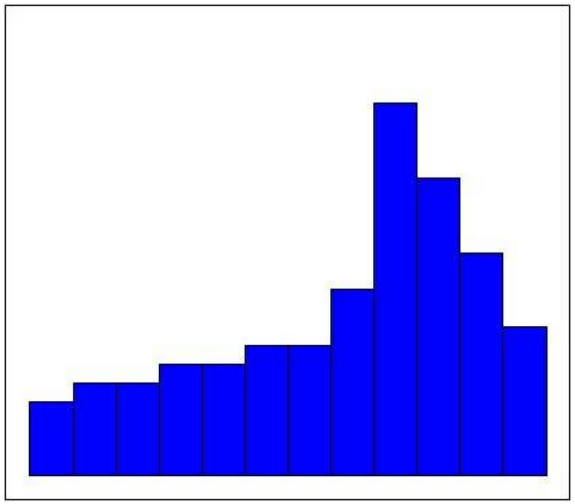
k	Min % unutar $k\sigma$ od $\overline{x}$	Max % izvan $k\sigma$ od $\overline{x}$
1	0.00	100.00
$\sqrt{2}$	50.00	50.00
1.5	55.56	44.44
2	75.00	25.00
3	88.89	11.11
4	93.75	6.25
5	96.00	4.00
6	97.22	2.78
7	97.96	2.04
8	98.44	1.56
9	98.77	1.23
10	99.00	1.00

Vysochanskij-Petuninova nejednakost

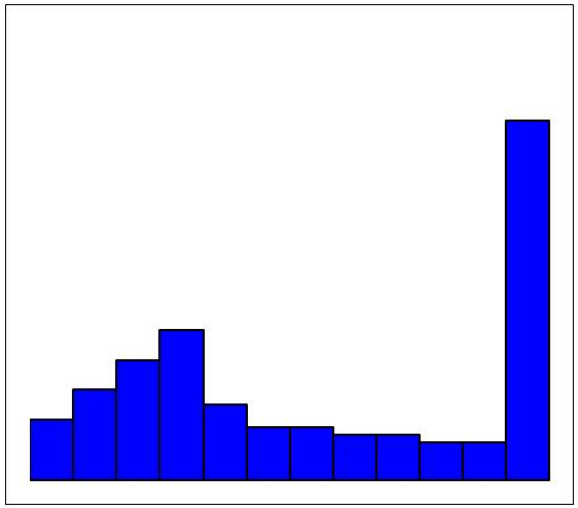
Na prethodno pitanje možemo dati i precizniji odgovor, ako je promatran uzorak iz *unimodalne* distribucije:

Vysochanskij-Petuninova nejednakost

Na prethodno pitanje možemo dati i precizniji odgovor, ako je promatran uzorak iz *unimodalne* distribucije:



Unimodalna distribucija

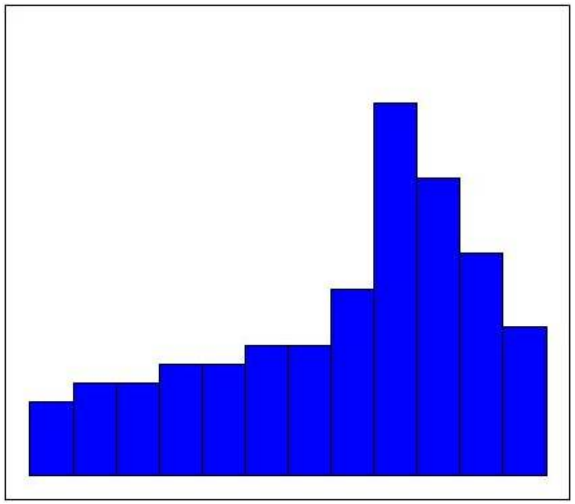


Bimodalna distribucija

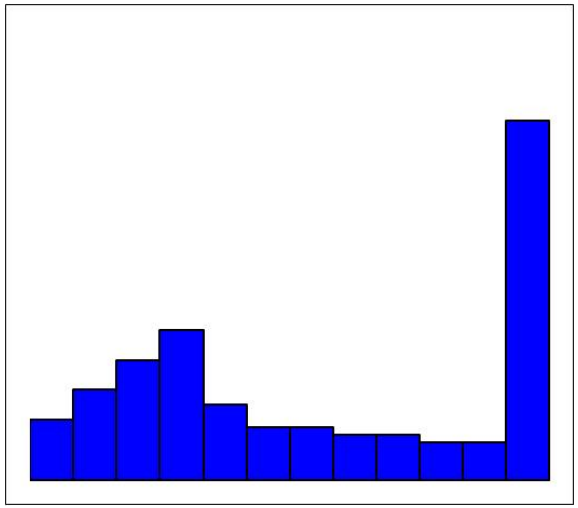
(unimodalna distribucija - samo jedan vrh, bimodalna distribucija - dva vrha)

Vysochanskij-Petuninova nejednakost

Na prethodno pitanje možemo dati i precizniji odgovor, ako je promatran uzorak iz *unimodalne* distribucije:



Unimodalna distribucija

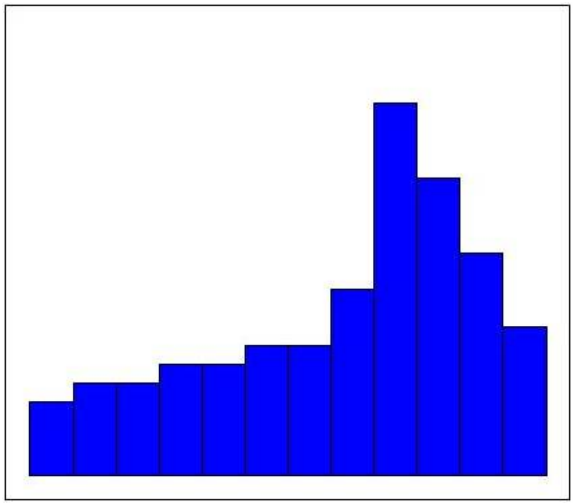


Bimodalna distribucija

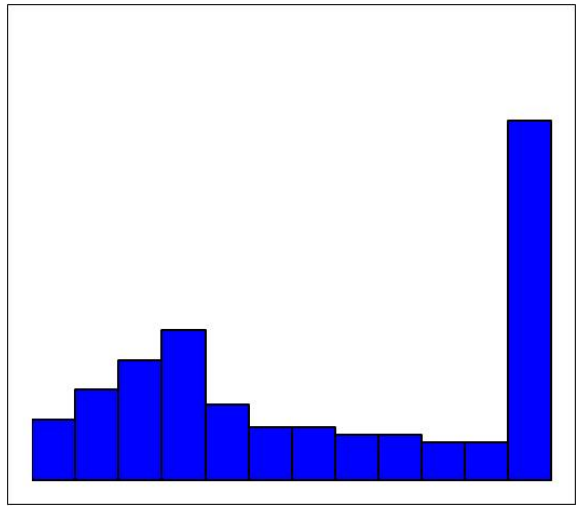
(unimodalna distribucija - samo jedan vrh, bimodalna distribucija - dva vrha)

Vysochanskij-Petuninova nejednakost

Na prethodno pitanje možemo dati i precizniji odgovor, ako je promatran uzorak iz *unimodalne* distribucije:



Unimodalna distribucija



Bimodalna distribucija

(unimodalna distribucija - samo jedan vrh, bimodalna distribucija - dva vrha)

Ukoliko je distribucija varijable unimodalna, tada između $\overline{x} - k\sigma$ i $\overline{x} + k\sigma$, $(k > 1)$, očekujemo da se nalazi najmanje

$$\left(1 - \frac{4}{9k^2}\right) \cdot 100\%$$

k	Čebiševljeva nejednakost	Vysochanskij-Petuninova nejednakost
1	0.00	55.56
$\sqrt{2}$	50.00	77.78
1.5	55.56	80.25
2	75.00	88.89
3	88.89	95.06
4	93.75	97.22
5	96.00	98.22
6	97.22	98.77
7	97.96	99.09
8	98.44	99.31
9	98.77	99.45
10	99.00	99.56