



Deskriptivna statistika

Teme

- ▶ **Mjere centralne tendencije**
 - ▶ Aritmetička sredina
 - ▶ Medijana
 - ▶ Modus
- ▶ *Odabir mjere centralne tendencije*
- ▶ *Mjere centralne tendencije i oblik distribucije*
- ▶ **Mjere varijacije**
 - ▶ Raspon
 - ▶ Varijansa
 - ▶ Standardna devijacija

Mjere centralne tendencije

- ▶ **Svrha** deskriptivne statistike: organizovanje i sumiranje seta vrijednosti
- ▶ Cilje je pronalaženje **jedne** vrijednosti koja najbolje predstavlja cijeli set vrijednosti
- ▶ **Definicija:** *Centralna tendencija je statistička mjera koja pokušava odrediti jednu vrijednost koja je najtipičnija i najreprezentativnija za određenu grupu podataka.*
- ▶ Korisne su za pravljenje **komparacija**
- ▶ Ne postoji standardizovana **procedura** za određivanje adekvatne mjere
- ▶ Najčešće korišćenje mjere: **aritmetička sredina, medijana i modus**

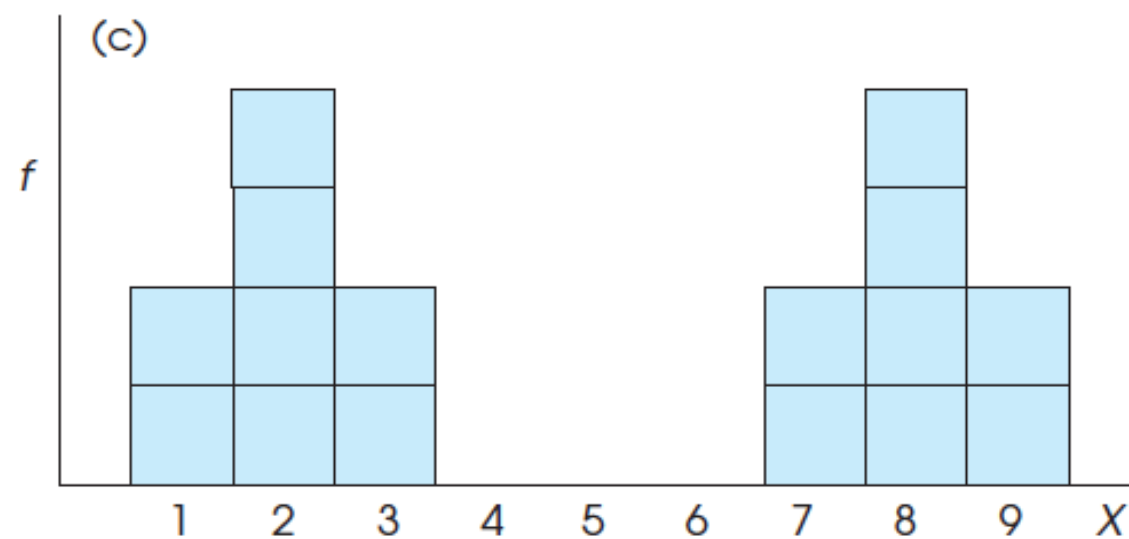
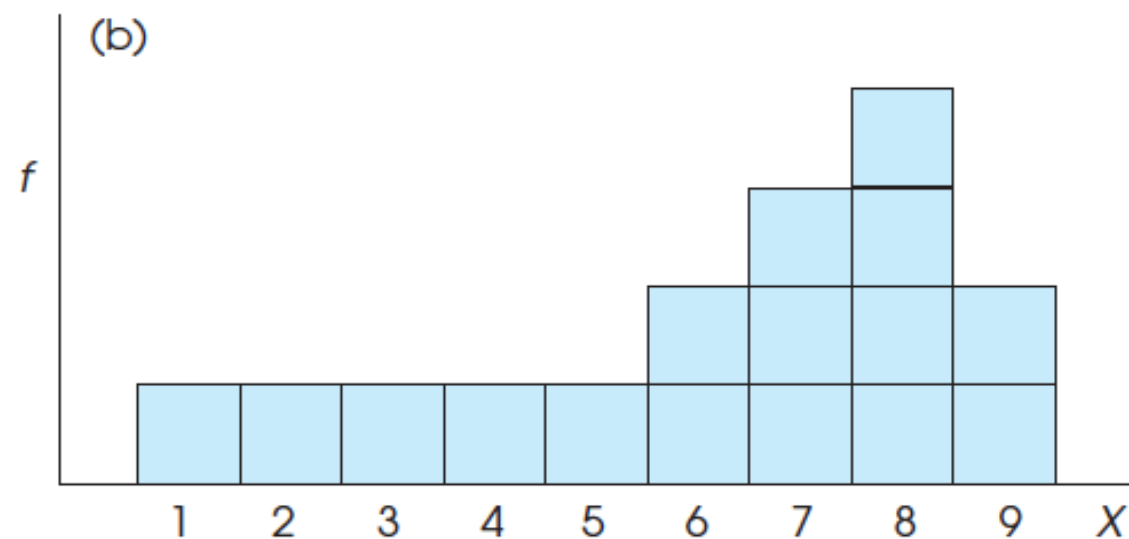
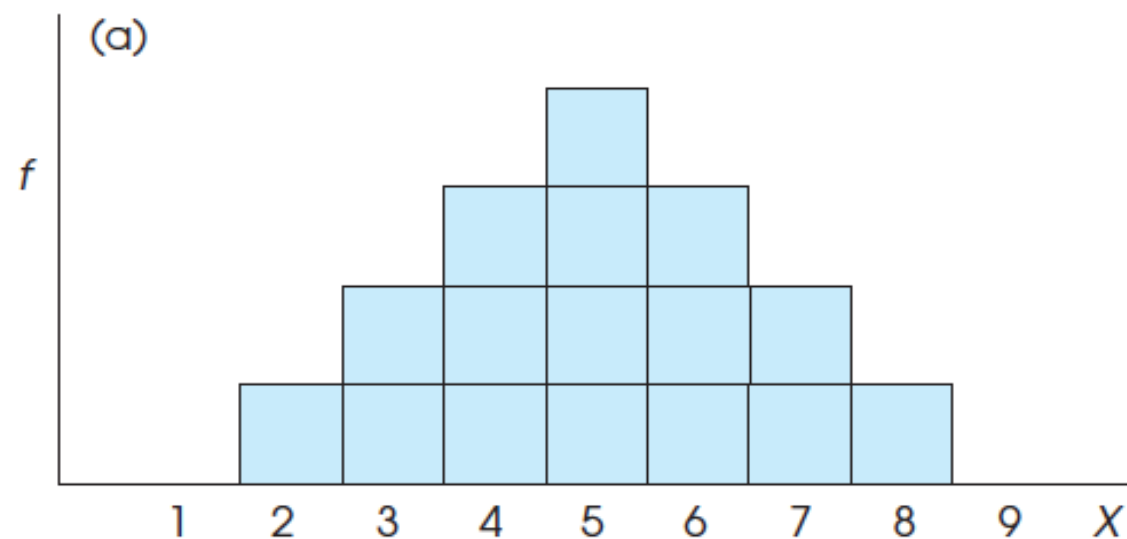


FIGURE 3.1

Three distributions demonstrating the difficulty of defining central tendency. In each case, try to locate the “center” of the distribution.

Aritmetička sredina (prosjeak)

- ▶ Najčešće korišćena mjera centralne tendencije
- ▶ Računanje: **sabrati sve vrijednost i podijeliti sa brojem vrijednosti u setu**
- ▶ Formalna notacija:

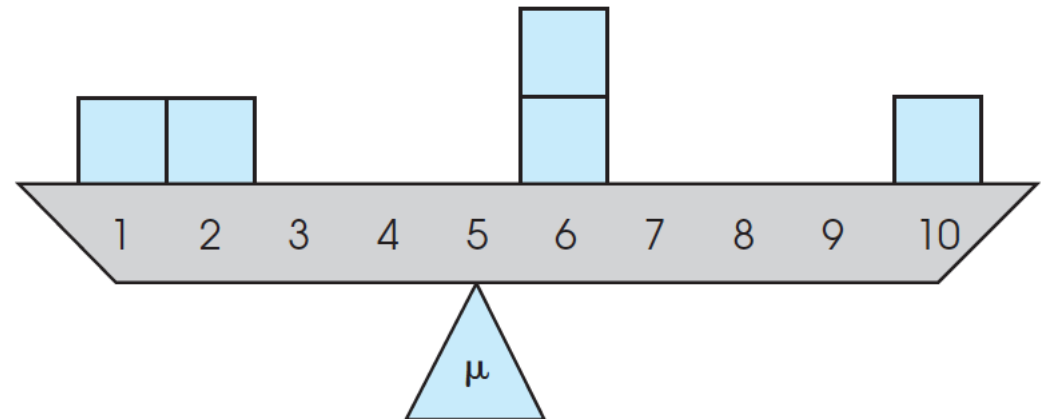
$$\bar{X} = \frac{\sum_{i=1}^n x_i}{n} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

- ▶ Drugačije obilježena shodno tome da li se odnosi na prosek u uzorku ili populaciji

Alternativna definicija

- ▶ Prosjek je tačka balansa određene distribucije
- ▶ Zamislimo populaciju koja sadrži sljedeće vrijednosti: **1, 2, 6, 6, 10**.
- ▶ Prosjek ovoj populaciji je **5**, i ako je prosjek dobra mjera ova vrijednost bi trebala kvalitetno balansirati između dvije strane
- ▶ Ukoliko saberemo pojedinačne udaljenosti od prosjeka (ispod i iznad) dobićemo isti broj

<i>Score</i>	<i>Distance from the Mean</i>
$X = 1$	4 points below the mean
$X = 2$	3 points below the mean
$X = 6$	1 point above the mean
$X = 6$	1 point above the mean
$X = 10$	5 points above the mean



Računanje prosjeka iz tebele frekvencije

- ▶ Najčešće vrijednosti organizovane u tabele frekvencije
- ▶ Uzimamo u obzir dvije informacije: **1) konkretnu vrijednost** i **2) koliko često se ponavlja**
- ▶ **Obavezno množiti!**

Računanje prosjeka iz tebele frekvencije

- ▶ Najčešće vrijednosti organizovane u tabele frekvencije
- ▶ Uzimamo u obzir dvije informacije: **1) konkretnu vrijednost** i **2) koliko često se ponavlja**
- ▶ Obavezno množiti!

Quiz Score (X)	f	fX
10	1	10
9	2	18
8	4	32
7	0	0
6	1	6

Računanje prosjeka iz tebele frekvencije

- ▶ Najčešće vrijednosti organizovane u tabele frekvencije
- ▶ Uzimamo u obzir dvije informacije: **1) konkretnu vrijednost** i **2) koliko često se ponavlja**

Quiz Score (X)	f	fX
10	1	10
9	2	18
8	4	32
7	0	0
6	1	6

$$\Sigma X = 10 + 9 + 9 + 8 + 8 + 8 + 8 + 6 = 66$$

- ▶ **Obavezno množiti!**

Računanje prosjeka iz tebele frekvencije

- ▶ Najčešće vrijednosti organizovane u tabele frekvencije
- ▶ Uzimamo u obzir dvije informacije: **1) konkretnu vrijednost** i **2) koliko često se ponavlja**

Quiz Score (X)	f	fX
10	1	10
9	2	18
8	4	32
7	0	0
6	1	6

$$\Sigma X = 10 + 9 + 9 + 8 + 8 + 8 + 8 + 6 = 66$$

- ▶ Obavezno množiti!

$$M = \frac{\Sigma X}{n} = \frac{66}{8} = 8.25$$

Karakteristike prosjeka

- ▶ Promjena vrijednosti bilo kojeg člana skupa mijenja prosjek
- ▶ Dodavanje nove, ili isključivanje već postojeće vrijednosti mijenja vrijednost prosjeka (osim u slučaju da je dodata vrijednost identična prosjeku)
- ▶ Ukoliko se svim vrijednostima doda ili oduzme konstantna, vrijednost prosjeka će se promijeniti za tu istu konstantu
- ▶ Množenje ili dijeljenje sa konstantom mijenja vrijednost prosjeka na identičan način

Medijana

- ▶ Medijana je locirana tačno na **sredini** distribucije
- ▶ Ne postoji formalna notacija
- ▶ **Definicija:** Ukoliko su vrijednosti poredane po veličini, **od najmanje do najveće**, medijana je tačka koja listu dijeli na pola
- ▶ Definisanje medijane znači podijeliti distribuciju na **dva jednaka dijela**
- ▶ Primjer: **5, 9, 7, 2, 4**
- ▶ Ukoliko skup ima **neparan** broj članova umjesto središnje vrijednosti uzimamo prosjek između središnje dvije
- ▶ Primjer: **5, 9, 7, 2, 4, 6**

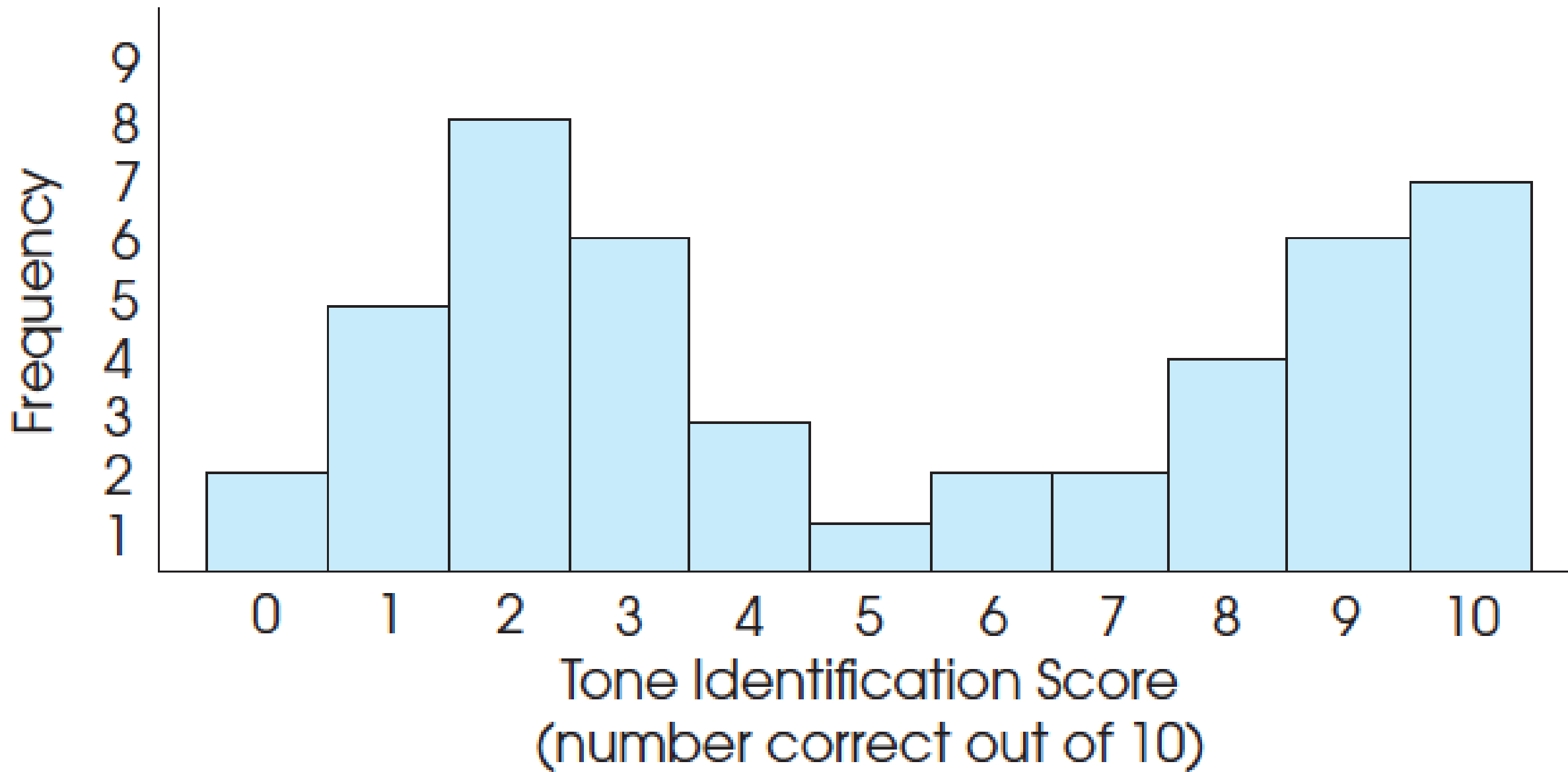
Modus

- ▶ **Definicija:** Najčešće primijećena vrijednost u jednom skupu
- ▶ Ne postoji formalna notacija
- ▶ Korsina za opservacije na kojima se ne mogu sprovesti matematičke operacije
- ▶ Distribucija ne mora imati modus, ali ih može imati više

Modus

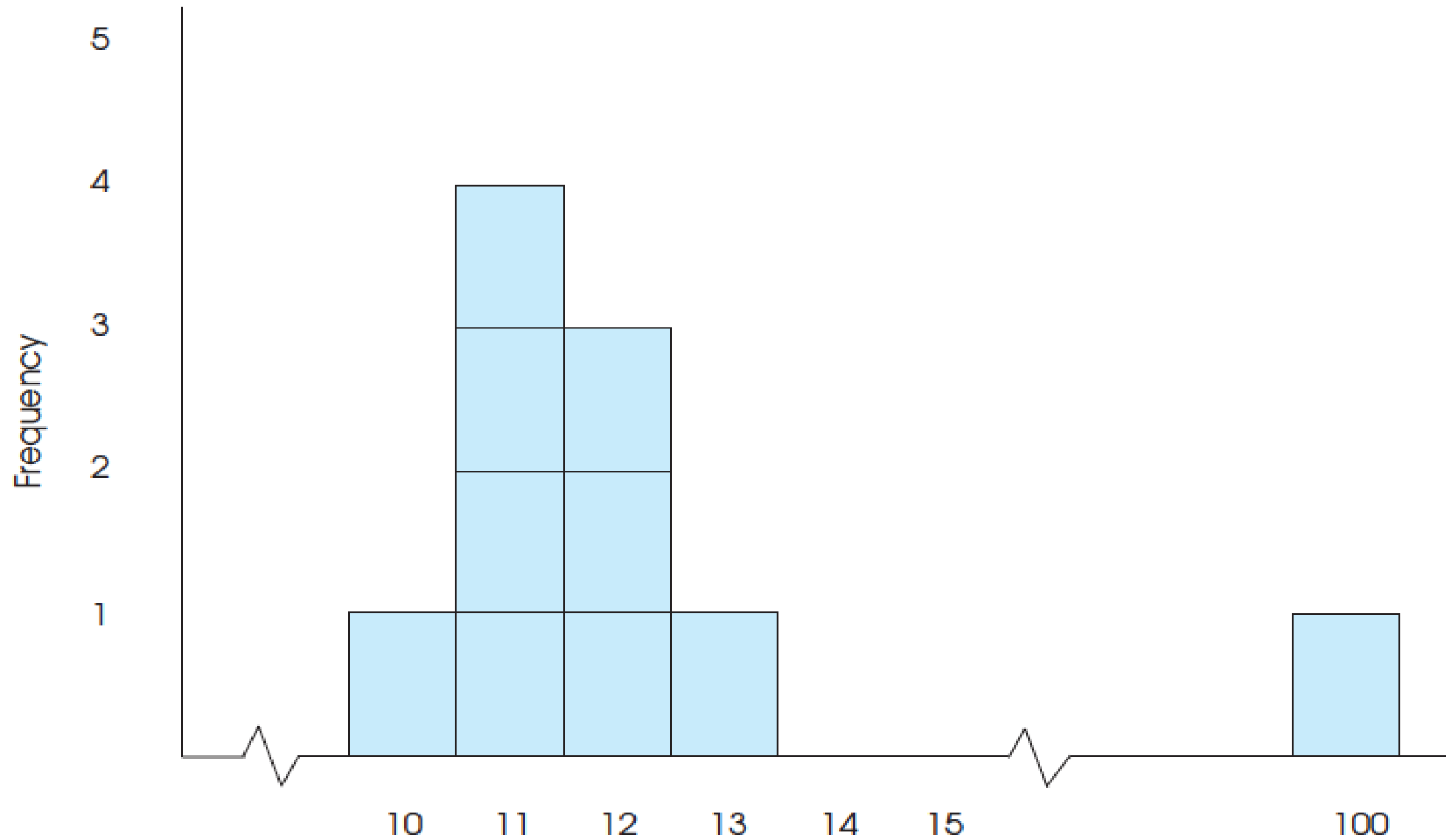
- ▶ **Definicija:** Najčešće primijećena vrijednost u jednom skupu
- ▶ Ne postoji formalna notacija
- ▶ Korsina za opservacije na kojima se ne mogu sprovesti matematičke operacije
- ▶ Distribucija ne mora imati modus, ali ih može imati više

<i>Restaurant</i>	<i>f</i>
College Grill	5
George & Harry's	16
Luigi's	42
Oasis Diner	18
Roxbury Inn	7
Sutter's Mill	12



Odabir mjere centralne tendencije

- ▶ Prosjek je najčešće korišćena mjera (najinformativnija)
- ▶ Nije nužno dobar reprezent svih vrijednosti
- ▶ Koristiti **medijanu** umjesto prosjeka u sljedećim situacijama:
 1. *Postoje ekstremne vrijednosti (distribucija je iskrivljena)*
 2. *Nedostajuće vrijednosti*
 3. *Otvorene distribucije (distribucije bez donje ili gornje granice)*
 4. *Kada baratamo ordinalnim skalama*



Odabir mjere centralne tendencije

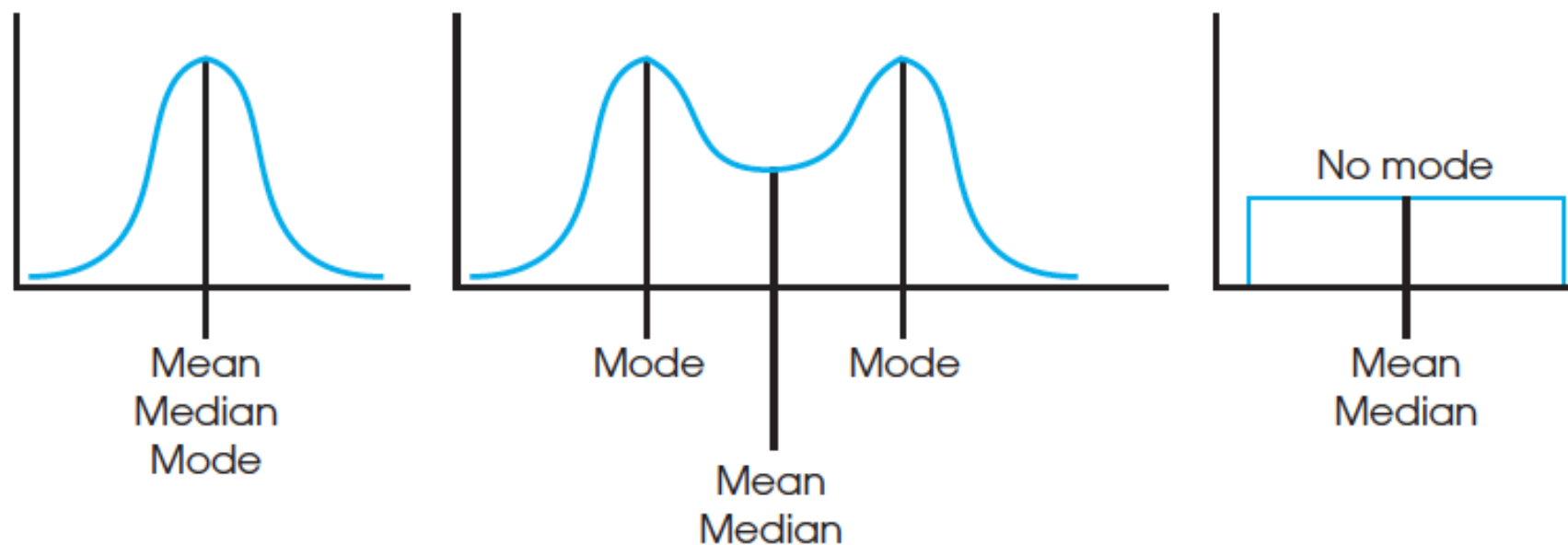
► Koristiti **modus** umjesto prosjeka ili medijane u sljedećim situacijama:

1. *Kada baratamo nominalnim skalama*
2. *Diskretne varijable*
3. *Opisivanje neobičnih oblika distribucije (nrp. ekstremno bimodalne)*

Oblici distribucija: *simetrične*

FIGURE 3.10

Measures of central tendency for three symmetrical distributions: normal, bimodal, and rectangular.



Oblici distribucija: *asimetrične*

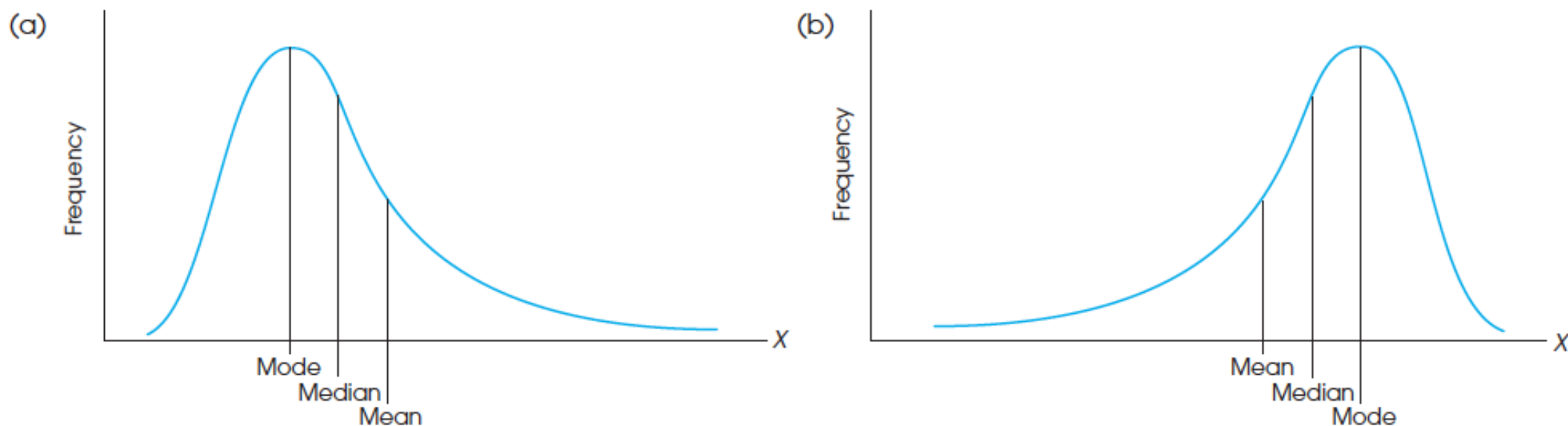


FIGURE 3.11

Measures of central tendency for skewed distributions.

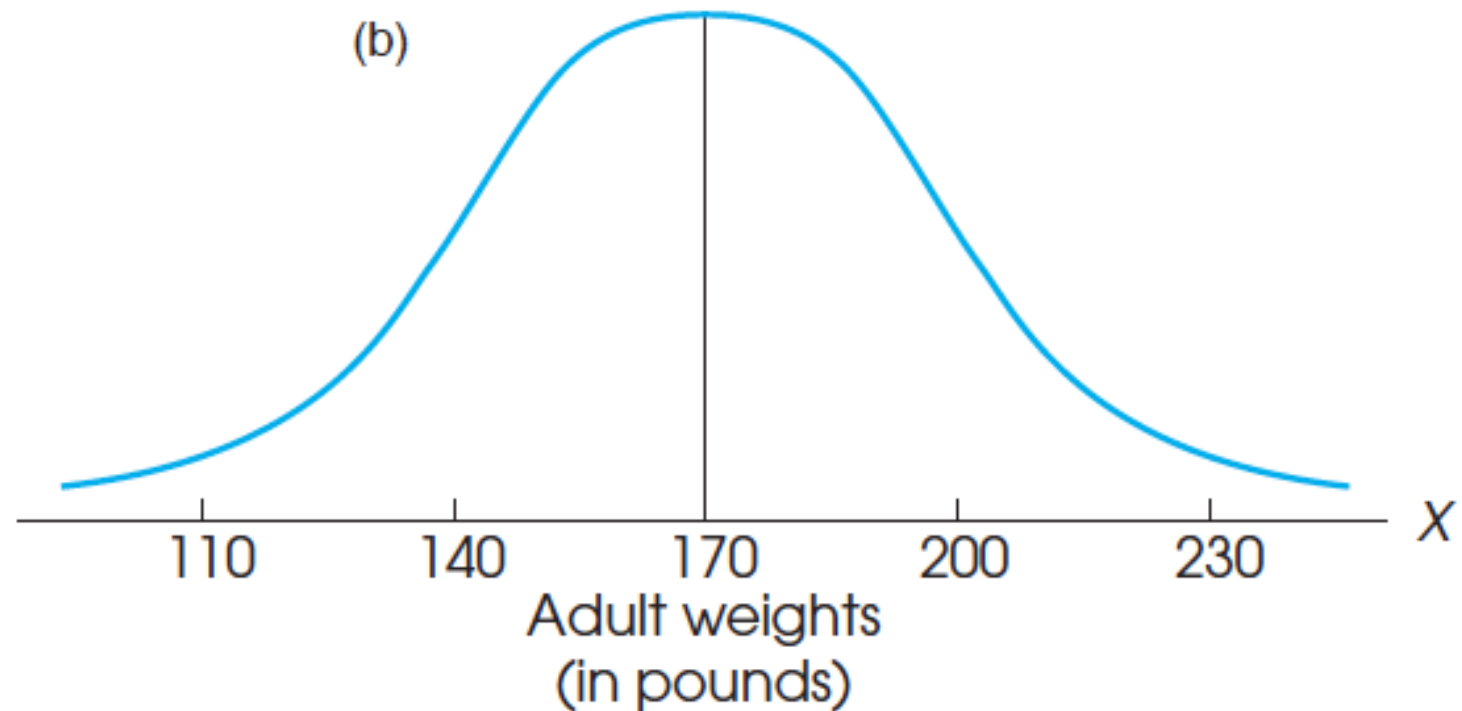
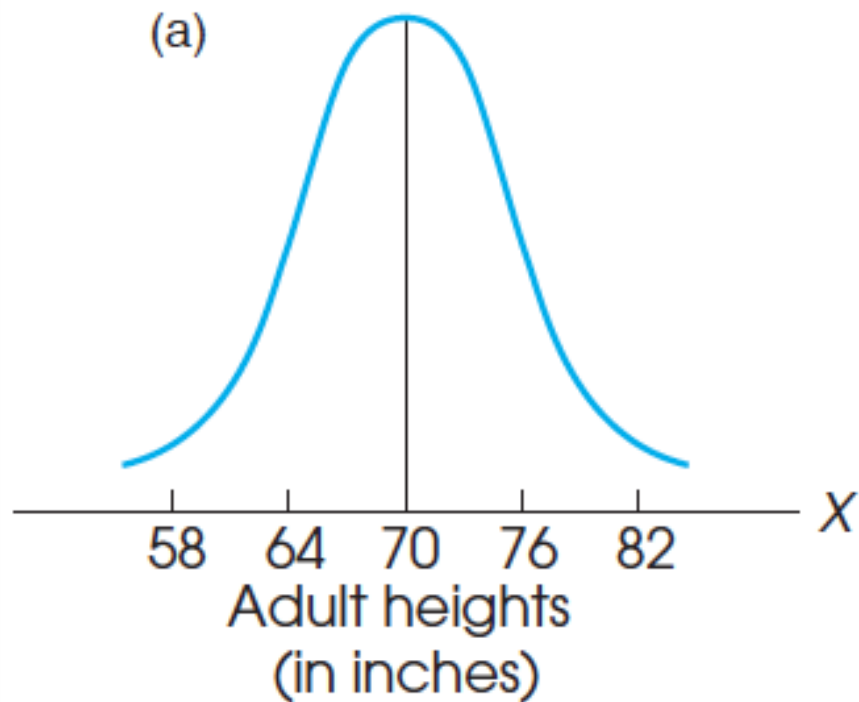


Mjere varijacije

METODOLOGIJA POLITIČKIH NAUKAME

Mjere varijacije

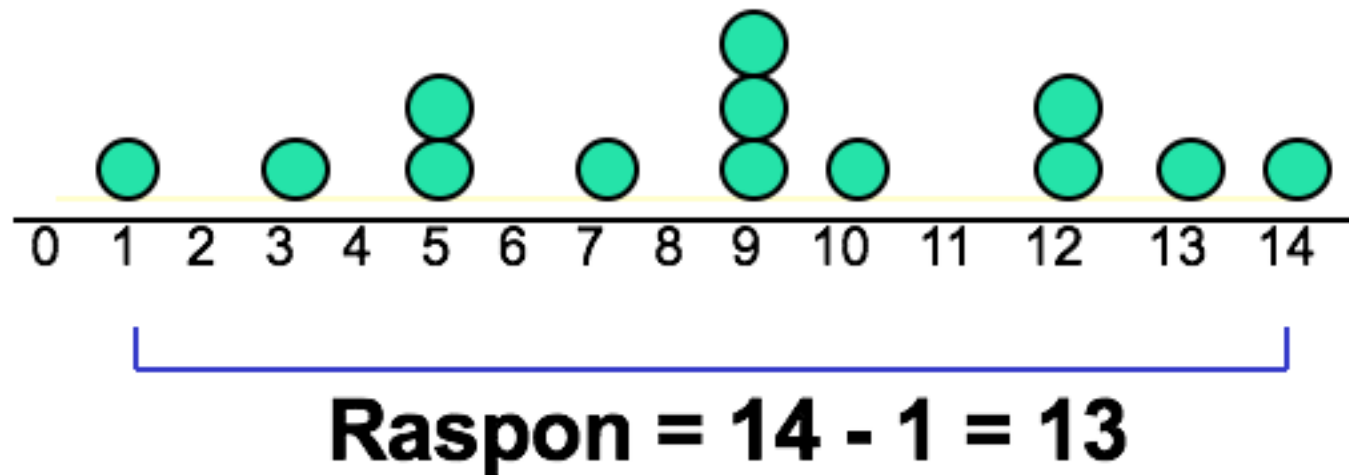
- ▶ Varijacija znači da opservacije koje istražujemo nijesu identične
- ▶ Varijacija je ključna za nauku, jer identične pojave nijesu interesantne za istraživanje
- ▶ **Definicija:** *Varijabilnost je kvantitativna mjera razlika među opservacijama, koja opisuje stepen u kojem su vrijednosti raspršene ili grupisane zajedno.*
- ▶ Najčešće korišćene mjere: **raspon, varijansa i standardna devijacija**



Varijabilnost je definisana u terminima **distance**. Govori nam koliko distance trebamo očekivati između **pojedinačnih skorova** i **prosjeaka**.

Raspon

- ▶ **Definicija:** distanca koja je pokrivena distribucijom, opseg između najveće i najmanje vrijednosti
- ▶ **Računanje:** $X_{max} - X_{min}$



Varijansa/Standardna devijacija

- ▶ **Definicija:** *prosječno kvadratno odstupanje pojedinačnih vrijednosti od prosjeka*
- ▶ Razlikuje se od od uzorka do populacije
- ▶ Standardna devijacija – najčešće korišćena mjera varijacije
- ▶ SD je samo korijenovana varijansa

Varijansa/Standardna devijacija

- ▶ **Definicija:** *prosječno kvadratno odstupanje pojedinačnih vrijednosti od prosjeka*
- ▶ Razlikuje se od od uzorka do populacije
- ▶ Standardna devijacija – najčešće korišćena mjera varijacije
- ▶ SD je samo korijenovana varijansa

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

\bar{x} = aritmetička sredina

n = veličina uzorka

x_i = i^{ta} vrijednost varijable X

Računanje

- ▶ **Korak 1:** Pronađi **aritmetičku sredinu**
- ▶ **Korak 2:** Od svake pojedinačne vrijednosti oduzmi aritmetičku sredinu (**odredi distancu**)
- ▶ **Korak 3:** **Kvadrirati** dobijenu distancu
- ▶ **Korak 4:** **Saber**i sve kvadrirane distance
- ▶ **Korak 5:** **Podijeli** sa brojem opservacija (ili $N - 1$, ukoliko je uzorak) **(VARIJANSA)**
- ▶ **Korak 6:** Izračunaj **kvadratni korijen** iz dobijene vrijednosti **(STAND. DEV)**

Zašto kvadriramo distance?

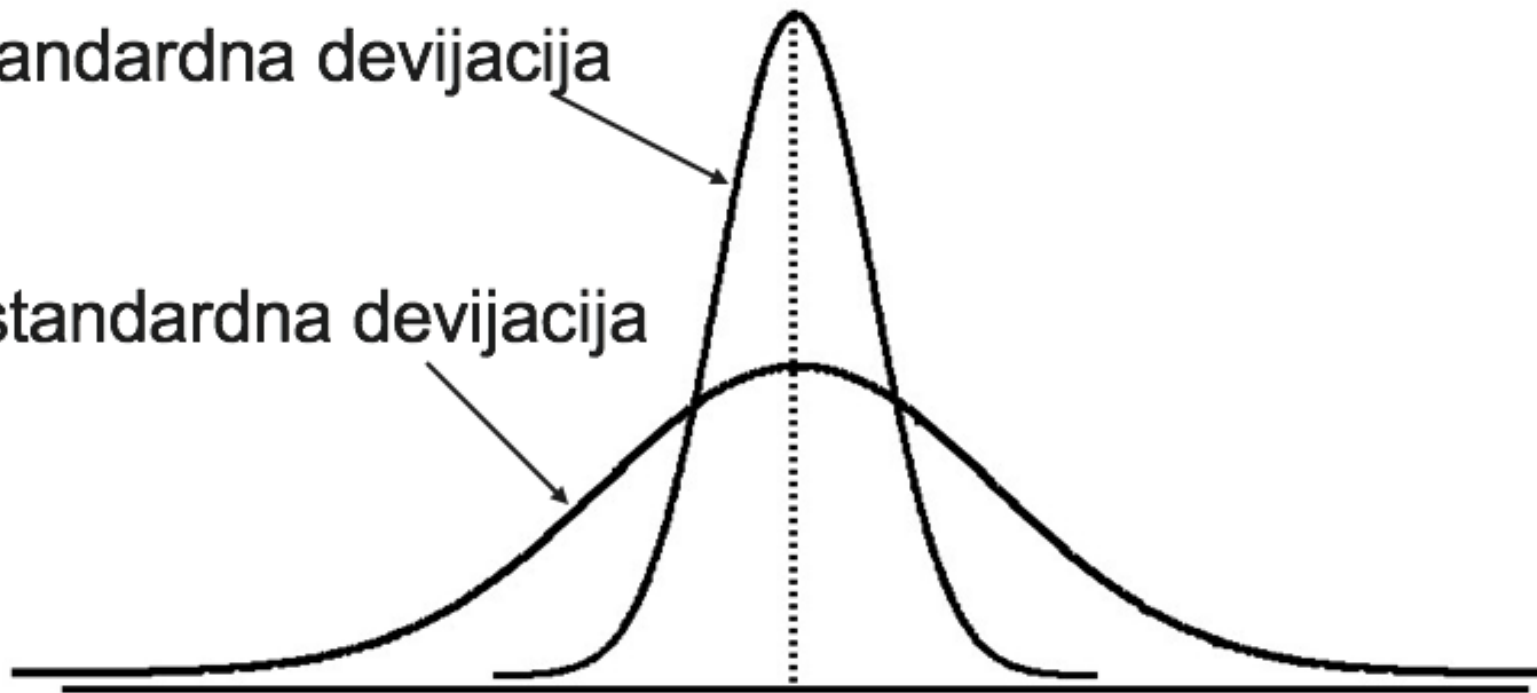
<i>Score X</i>	<i>Deviation $X - \mu$</i>	<i>Squared Deviation $(X - \mu)^2$</i>
1	-5	25
9	3	9
5	-1	1
8	2	4
7	1	1

40 = the sum of the squared deviations

U praksi...

Mala standardna devijacija

Velika standardna devijacija



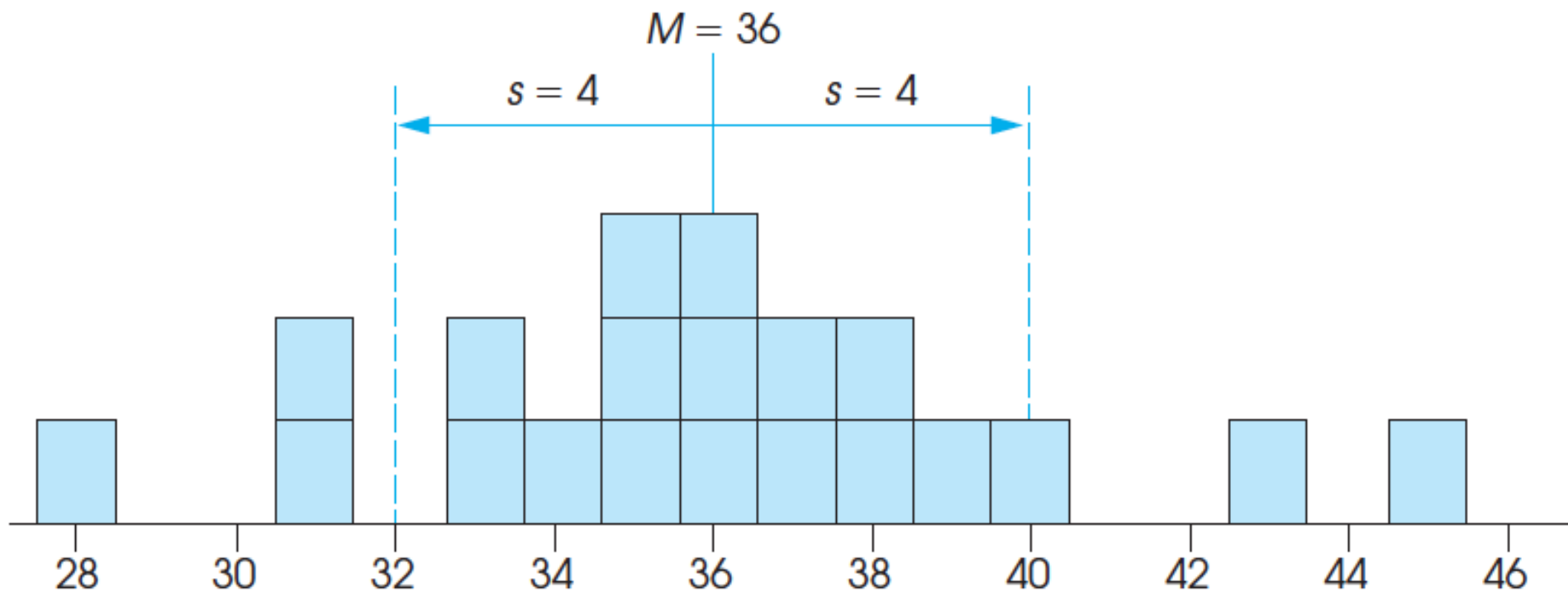


FIGURE 4.7

A sample of $n = 20$ scores with a mean of $M = 36$ and a standard deviation of $s = 4$.

Empirijsko pravilo: $\pm 1SD$ (68%)
 $\pm 2SD$ (95%)
 $\pm 3SD$ (99.7%)

Zašto marimo za empirijsko pravilo?

METODOLOGIJA POLITIČKIH NAUKAME