PREDAVANJE 2

DESKRIPTIVNA STATISTIKA

Deskriptivna statistika

- početna faza statističke obrade podataka
- prikazuje i opisuje podatke
- Obuhvata sledeće grafičke i numeričke metode:
  1) tabele i grafikone
  2) mere centralne tendencije
  3) mere oblika
  4) mere varijacije
  5) mere pozicije

Odabir metoda i načina izračunavanja zavisi od vrste uzorka

Vrste uzoraka

Prost uzorak

- Pojedinačno navedeni svi elementi uzorka
- nesortiran:

  2,3,2,1,5,4,4
- sortiran:

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

Grupisan uzorak

- navedene su klase elemenata i broj elemenata uzorka koji pripadaji datoj klasi

klase	Broj elemenata
1	1
2	2
3	1
4	2
5	1

Grupisan uzorak

- Klase kod grupisanog uzorka mogu sadržati jednu ili više vrednosti (intervali).

klase	Broj elemenata
1	1
2	2
3	1
4	2
5	1

klase	Br elemenata
1 - 3	4
4 - 5	3

Iz intervalnog uzorka ne možemo rekonstruisati prost, deo informacije se gubi!

Tabele i grafikoni

Raspodela frekvencija

Raspodela frekvencija je tabela koja prikazuje klase ili intervale podataka, zajedno sa brojem elemenata uzorka koji pripadaju svakoj klasi (frekvencija, f).

Intervali

I	f
[1, 5)	4
[5, 9)	5
[9, 13)	3
[13, 17)	4
[17, 20]	2

frekvencije

[označava da interval sadrži rubnu tačku, a (da je ne sadrži.

Npr, tačka 5 pripada intervalu [5, 9) a ne pripada [1, 5).

Širina i opseg

Širina intervala je razlika izmedju gornje i donje granice.

Intervali mogu imati istu ili različitu širinu.

5 – 1 = 4

20 – 17 = 3

I	f
[1, 5)	4
[5, 9)	5
[9, 13)	3
[13, 17)	4
[17, 20]	2

Opseg je razlika izmedju gornje granice poslednjeg i donje granice prvog intervala (20 – 1 = 19).

Pravljenje raspodele frekvencija

Primer:

Podaci predstavljaju starost studenata. Napraviti tabelu raspodele frekvencija sa 5 klasa.

Starost studenata

18	20	21	27	29	20
19	30	32	19	34	19
24	29	18	37	38	22
30	39	32	44	33	46
54	49	18	51	21	21

Primer (nastavak):

1. broj klasa treba da bude 5.

2. Minimalni podatak je 18, maksimalni je 54, pa je opseg 54 – 18 = 36. Opseg podeljen brojem klasa daje širinu klase.

$$\text{Širina} = \frac{36}{5} = 7.2$$ Zaokružićemo na 8.

Primer (nastavak):

- Ako su podaci diskretni (celi brojevi), možemo koristiti jednostavniji zapis intervala npr. 18 – 25 umesto [18, 26)

starost studenata

I		f
18 – 25		13
26 – 33		8
34 – 41		4
42 – 49		3
50 – 57		2
		$\Sigma f = 30\ (= n)$

Zbir frekvencija odgovara obimu uzorka!

Sredina intervala

$$\text{Sredina } (x) = \frac{(\text{gornja granica}) + (\text{donja granica})}{2}$$

I	f	x
[1, 5)	4	3

$$x = \frac{1+5}{2} = 3$$

Relativna frekvencija

Relativna frekvencija neke klase je udeo (proporcija) elemenata koji upadaju u tu klasu u odnosu na ukupan broj elemenata.

$$\mathrm{p} = \frac{f}{n}$$

I	f	p
[1, 5)	4	0.222

$\sum f = 18$

$$\mathrm{p} = \frac{f}{n} = \frac{4}{18} \approx 0.222$$

Primer:

Odrediti relativne frekvencije za starost studenata.

I	f	p
$18-25$	13	$\frac{f}{n}$ 0.433
$26-33$	8	0.267
$34-41$	4	0.133
$42-49$	3	0.1
$50-57$	2	0.067
	$\sum f = 30$	$\sum \frac{f}{n} = 1$

$= \frac{13}{30}$

≈ 0.433

Zbir relativnih frekvencija je 1

Kumulativne frekvencije

Kumulativna frekvencija neke klase je suma frekvencija te klase i svih prethodnih.

starost studenata

I	*f*	kumulativna f
18 – 25	13	13
26 – 33	+ 8	21
34 – 41	+ 4	25
42 – 49	+ 3	28
50 – 57	+ 2	30
	$\sum f = 30$	

poslednja kumulativna f odgovara obimu

Korigovana frekvencija

Korigovana frekvencija nekog intervala je frekvencija po jedinici širine intervala (gustina).

Koristi se kao dopunska informacija kada intervali nisu jednake širine.

$$f' = \frac{f}{h}$$

Primer:

I	f	h	f'
[0, 5)	10	5	2
[5, 15)	10	10	1

f sugeriše da intervali imaju jednaku zastupljenost,

ali f ' nam pokazuje da je prvi interval gušći.

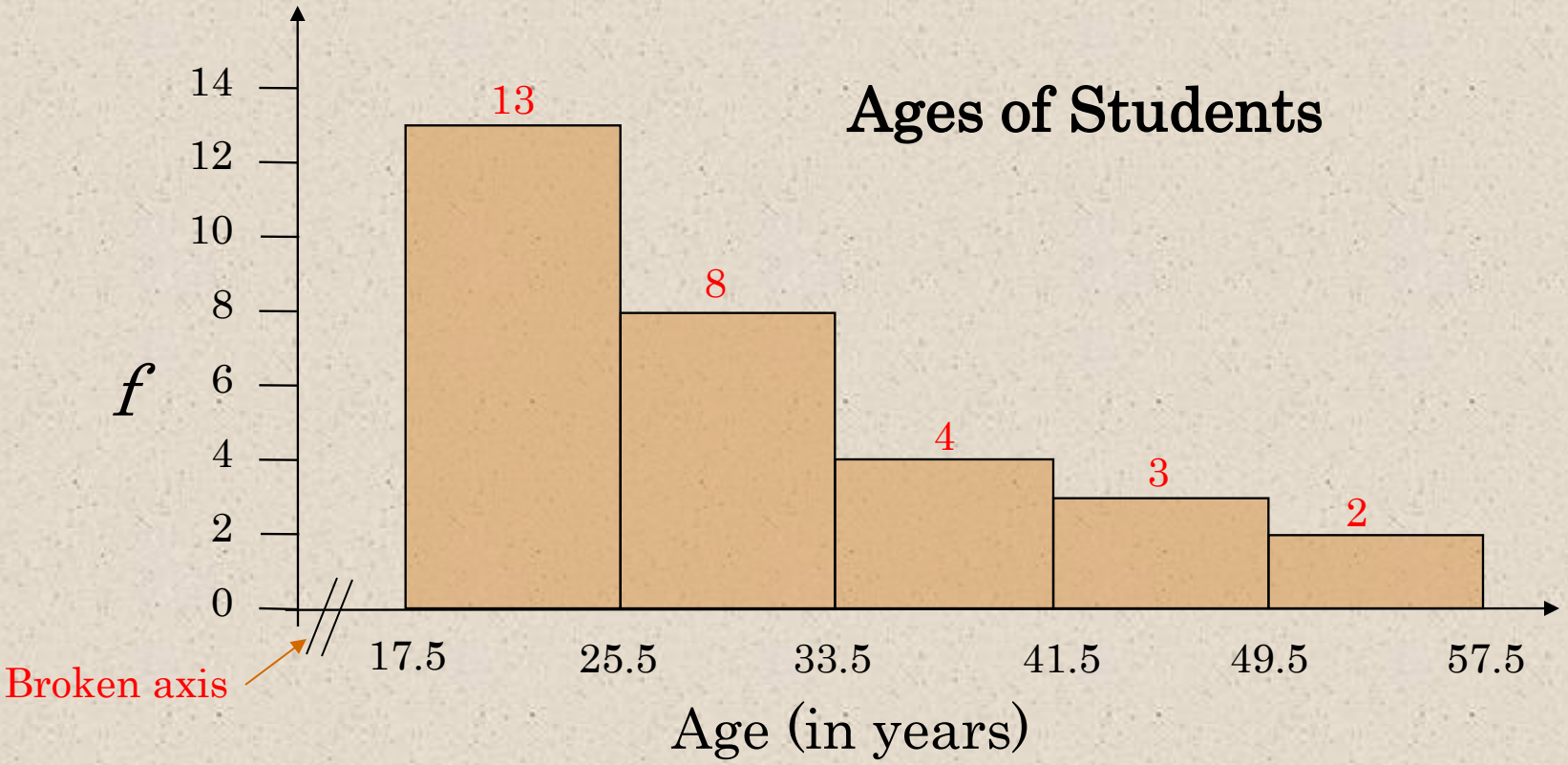
Rezime

- Uzorak može biti prost, grupisan u klase od jednog elementa, i intervalni (klase sa više elemenata).
- Tabela raspodele frekvencija, obim uzorka
- Opseg, širina, sredina intervala
- Relativne, kumulativne i korigovane frekvencije.

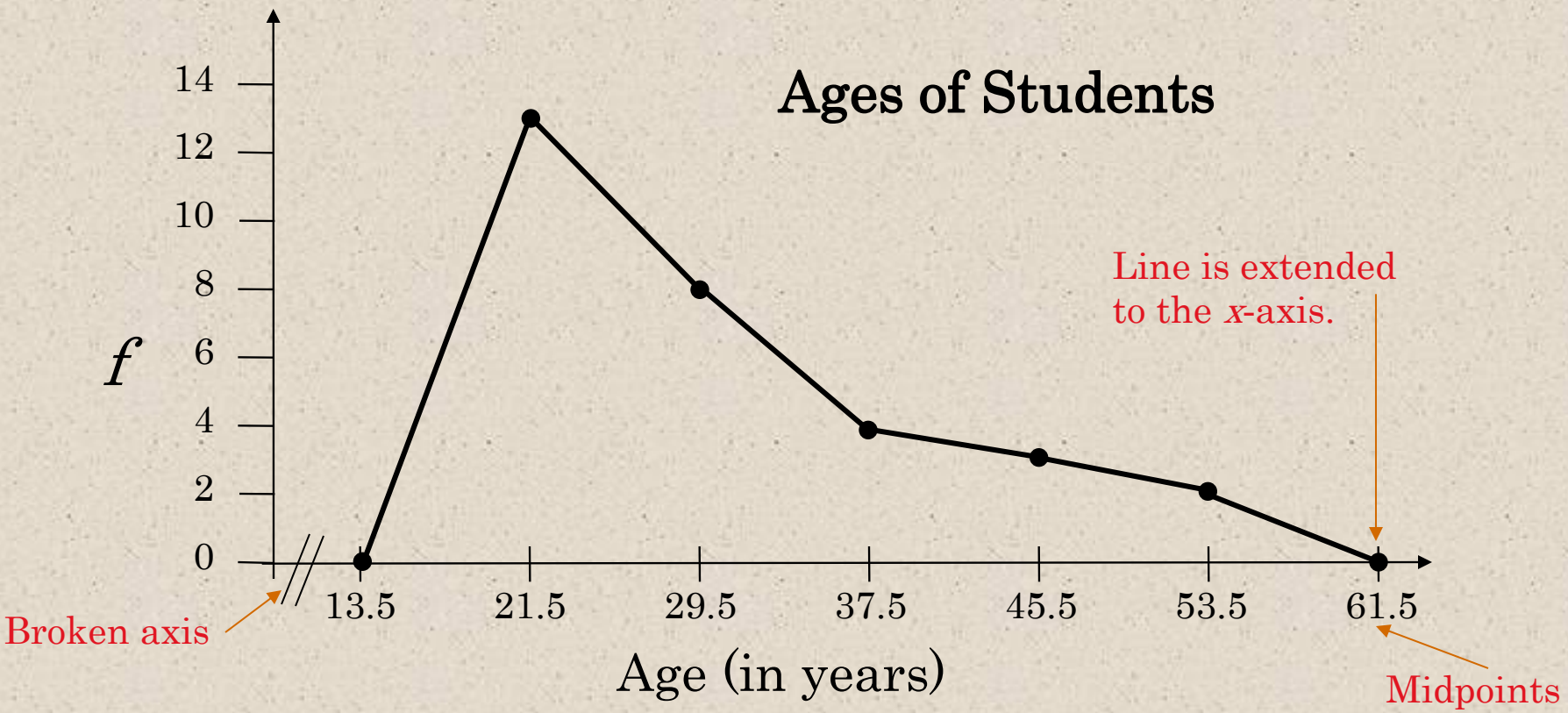
Primer: Tabela prikazuje uzrasnu strukturu stanara jedne zgrade. Napraviti detaljni tabelarni prikaz uzorka (sve iz rezimea)

Starost (god)	0-15	16-25	26-45	46-65	66-90
Broj stanara	10	5	28	35	12

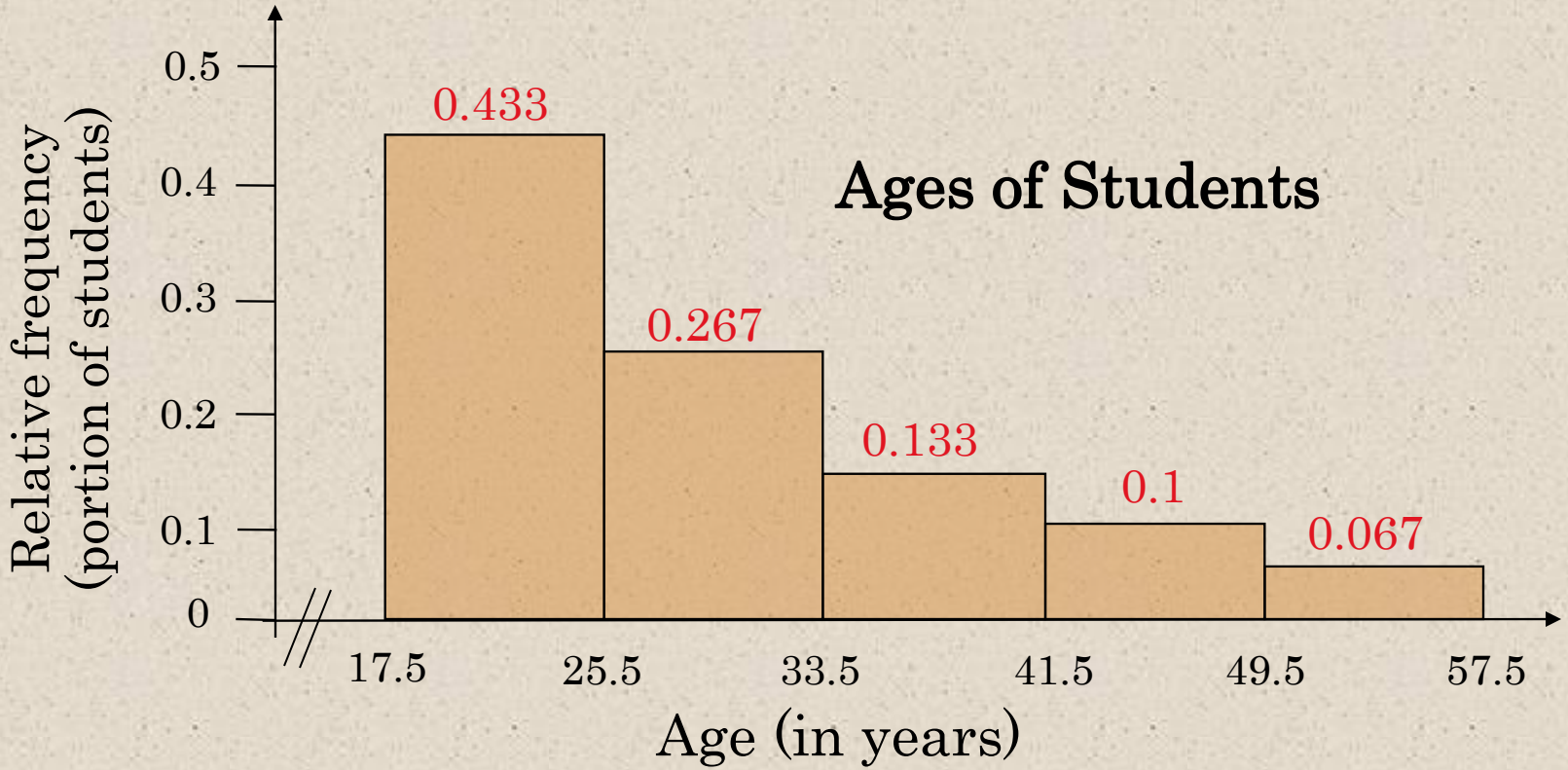
Histogram frekvencija



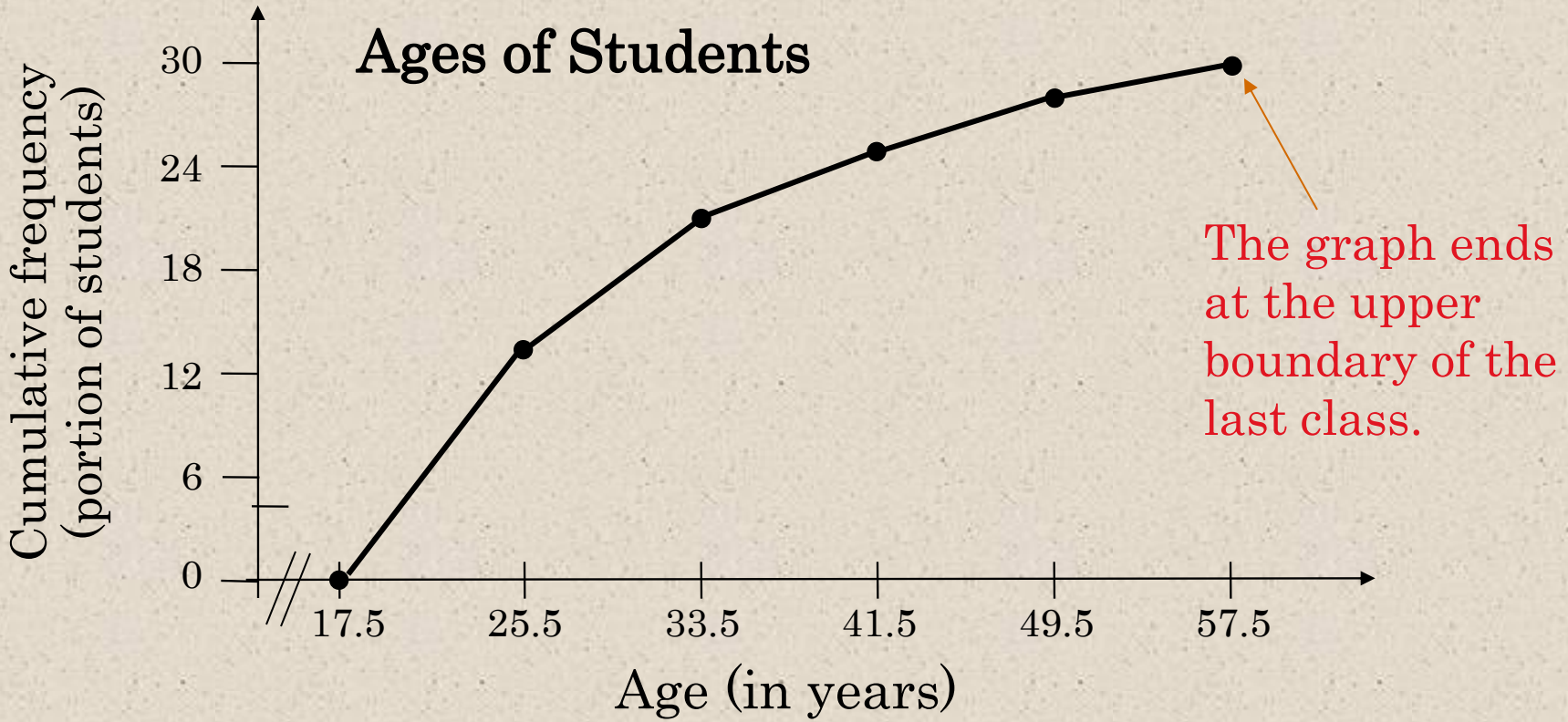
Poligon frekvencija



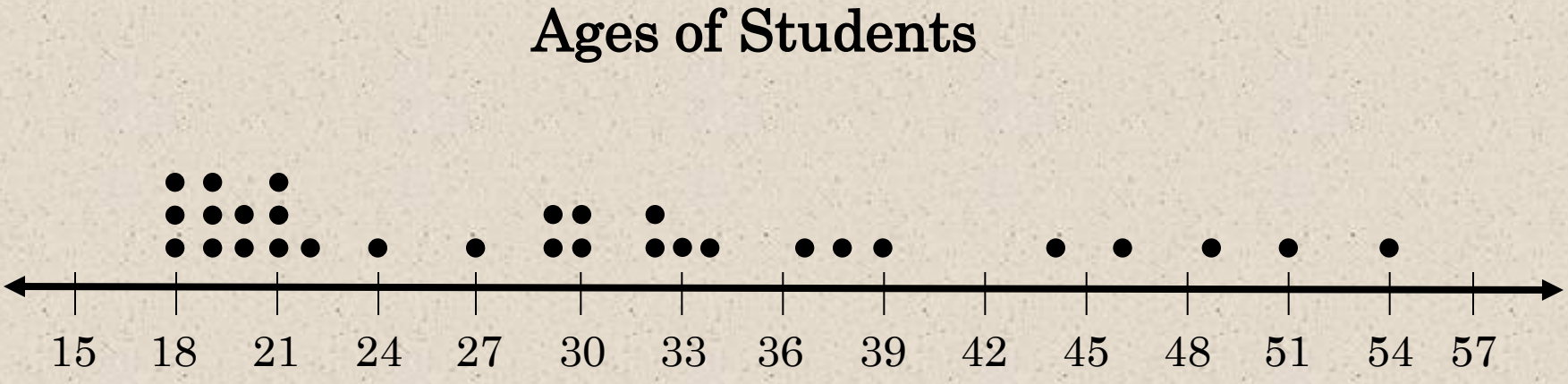
Histogram relativnih frekvencija



Ogiva (graf kumulativnih f)



Tačkasti dijagram (dot plot)



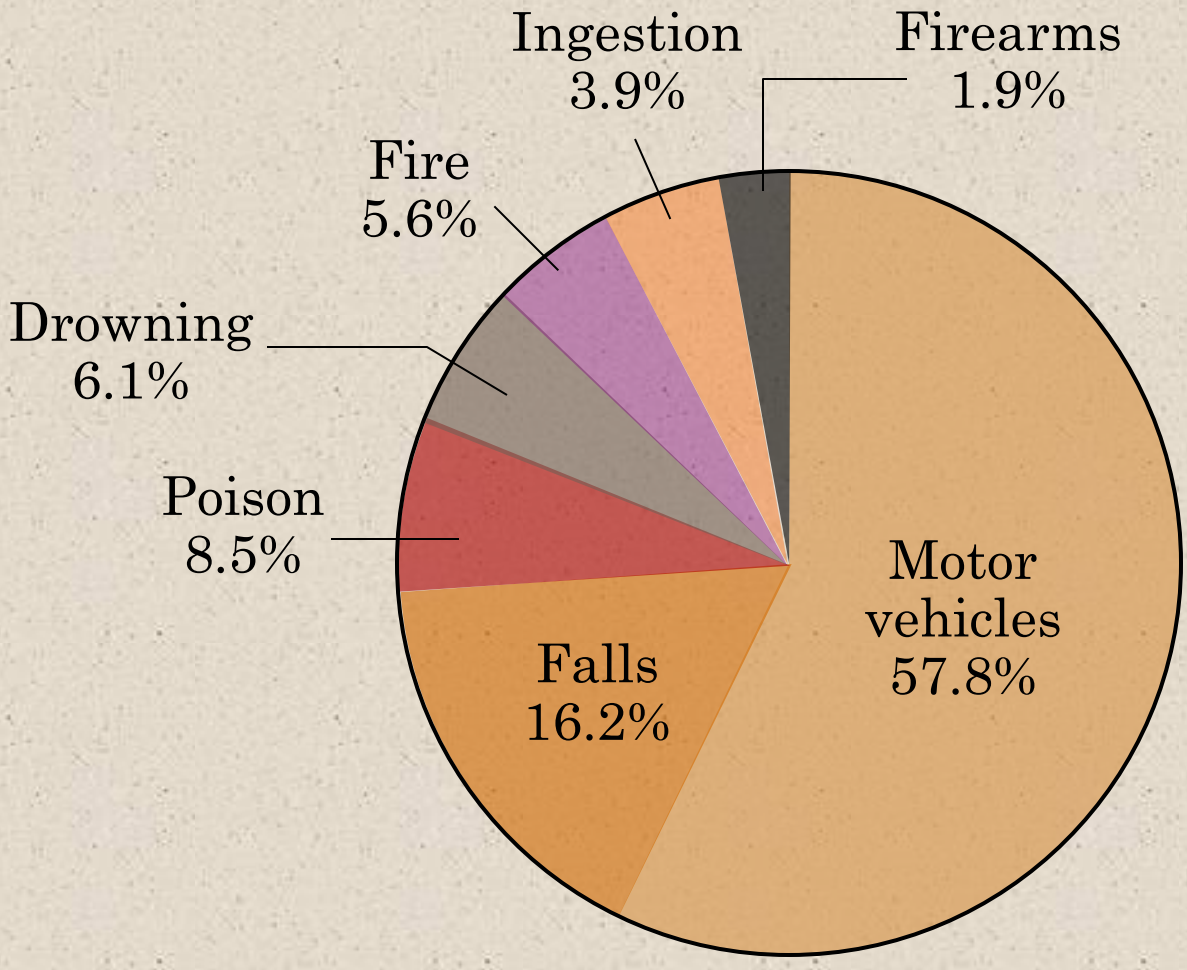
Pita graf (pie chart)

- koristi se najčešće kod nominalnih podataka
- Primer: podaci predstavljaju uzroke nesreća sa smrtnim ishodom u USA 2002.

I	f	p
Motor Vehicle	43 500	0.578
Falls	12 200	0.162
Poison	6 400	0.085
Drowning	4 600	0.061
Fire	4 200	0.056
Ingestion of Food/Object	2 900	0.039
Firearms	1 400	0.019

$n = 75\ 200$

Pita graf


Ingestion
3.9%
Firearms
1.9%
Fire
5.6%
Drowning
6.1%
Poison
8.5%
Motor
vehicles
57.8%
Falls
16.2%

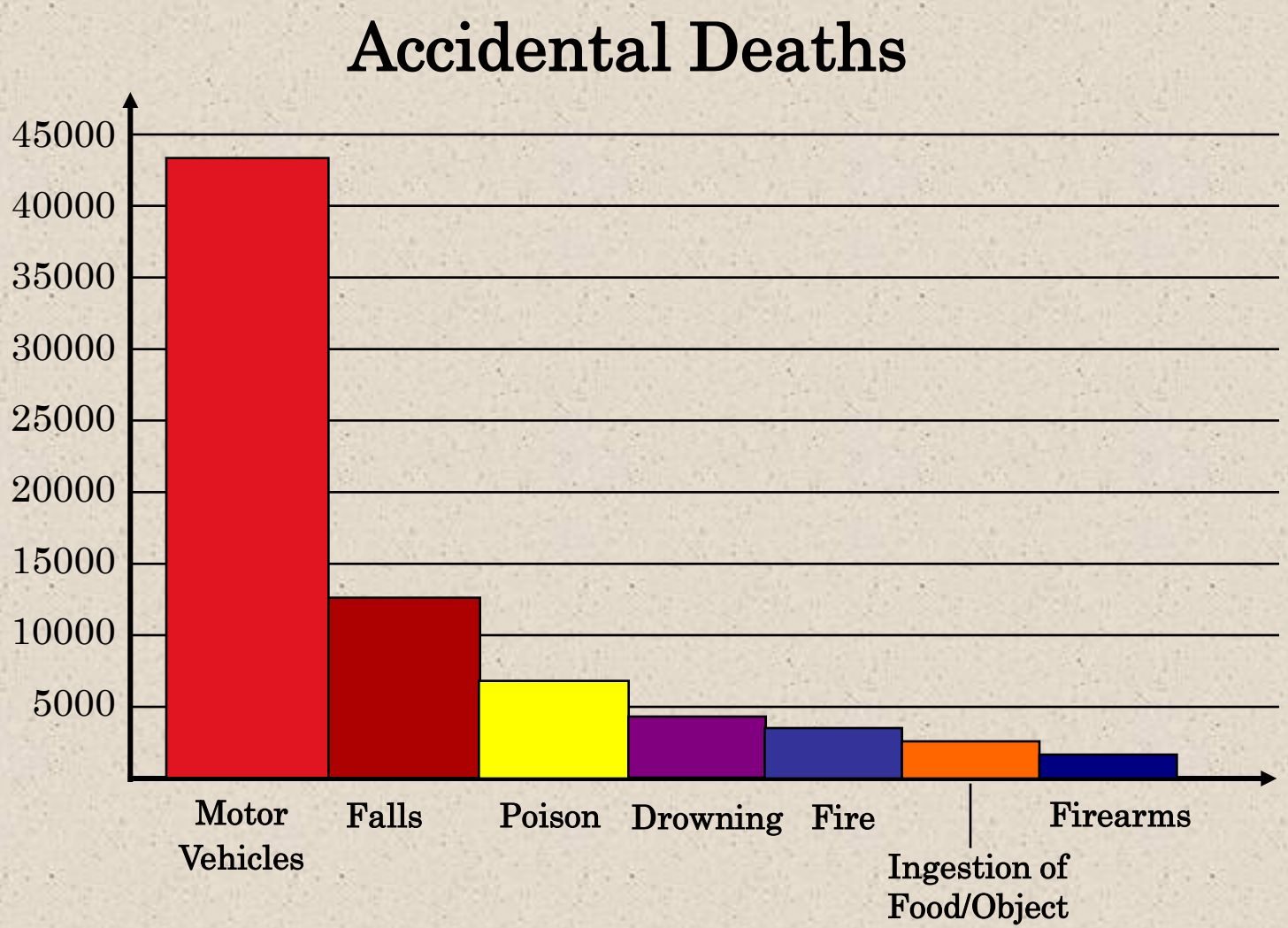
Pareto graf

- Pareto graf je stubičasti grafikon za nominalne podatke, koristi frekvencije, i klase sortirane prema frekvencijama.

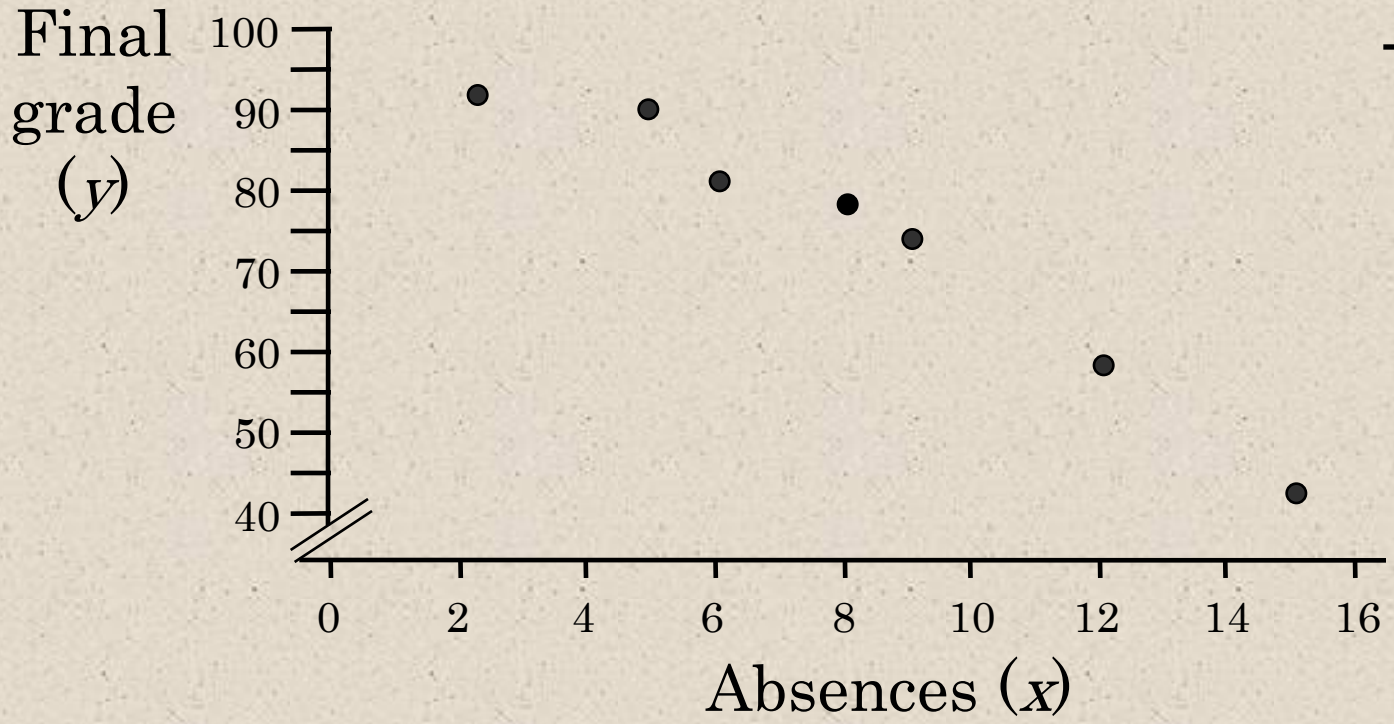
Accidental Deaths in the USA in 2002

Type	Frequency
Motor Vehicle	43,500
Falls	12,200
Poison	6,400
Drowning	4,600
Fire	4,200
Ingestion of Food/Object	2,900
Firearms	1,400

Pareto graf


Accidental Deaths
45000
40000
35000
30000
25000
20000
15000
10000
5000
Motor Vehicles
Falls
Poison
Drowning
Fire
Ingestion of Food/Object
Firearms

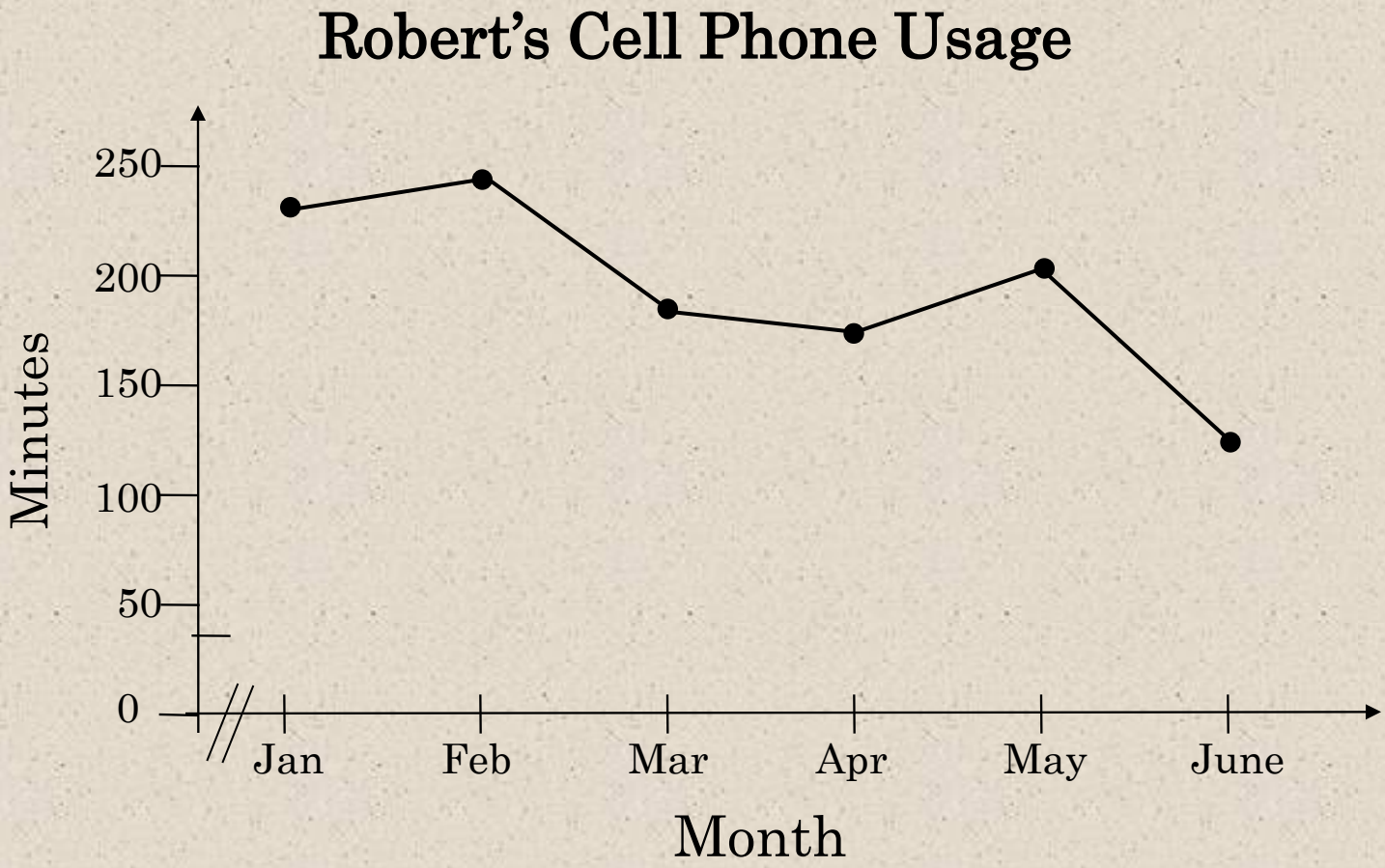
Dijagram rasipanja (Scatter Plot)



Absences x	Grade y
8	78
2	92
5	90
12	58
15	43
9	74
6	81

- Koristi se za parove podataka (X,Y);
- polazna tačka u regresionoj analizi.

Graf vremenskih serija



II DEO

Mere centralne tendencije

Aritmetička sredina (Mean)

Mere centralne tendencije su tipične, srednje, reprezentativne vrednosti obeležja na uzorku ili populaciji.

Tri najčešće korišćene su: aritmetička sredina, modus i medijana.

Aritmetička sredina:

A.S. populacije:

$$\mu = \frac{\sum x}{N}$$

A.S. prostog uzorka:

$$\bar{x} = \frac{\Sigma x}{n}$$

A.S. grupisanog uzorka:

$$\bar{x} = \frac{\Sigma f x}{n}$$

- **Primer:**
- Navedene su godine starosti svih 7 zaposlenih jednog malog preduzeća. Izračunati aritmetičku sredinu.

50 32 61 57 39 47 57

U pitanju je A.S. populacije:

$$\mu = \frac{\sum x}{N} = \frac{50+32+\cdots+57}{7} = 49$$

Prosečna starost zaposlenih je 49 godina.

Medijana

Medijana je srednja vrednost sortiranog uzorka

- ukoliko je obim uzorka **neparan** broj, imamo jednu srednju vrednost.
- ukoliko je obim **paran**, imamo dve srednje vrednosti i medijana je njihova aritmetička sredina.

Primer:

Izračunati medijanu starosti 7 zaposlenih.

50 32 61 57 39 47 57

Prvo sortiramo uzorak, a onda odredimo srednji element.

32 39 47 50 57 57 61

Medijana iznosi 50 godina.

Modus

Modus uzorka je vrednost sa najvećom frekvencijom.

Uzorak može imati jedan ili više modusa (bimodalni, multimodalni uzorak), a može i da nema modusa (ako se sve vrednosti javljaju sa istom frekvencijom).

Primer:

Odrediti modus starosti 7 zaposlenih

50 32 61 **57** 39 47 **57**

Modus je 57 godina, ta vrednost se najčešće javlja.

Autlajer (oulier) je vrednost koja je najviše udaljena od ostalih vrednosti uzorka.

Upoređivanje tri mere CT

Primer:

- jedan 22-ogodišnjak se zaposlio u posmatranoj kompaniji.

50 32 61 57 39 47 57 22

Izračunaj ponovo AS, medijanu i modus. Koja mera se najviše promenila?

A.S. = 45.6

Medijana = 48.5

Modus = 57

AS je najpodložnija dejstvu autlajera, jer uzima u obzir sve vrednosti.

Medijana je umereno pod uticajem aulajera, a modus uopšte nije.

Modalni i medijalni interval

- Kod intervalnog uzorka, modus i medijana su pozicionirani u modalnom tj medijalnom intervalu.
- Modalni interval je onaj koji ima najveću korigovanu frekvenciju.
- Medijalni interval je onaj u kome kumulativna frekvencija prvi put prelazi polovinu obima.

	I	f	h	Kumulativna f	Korigovana f
Modalni int. →	[0,2)	10	2	10	5
Medijalni int. →	[2,5)	12	3	22	4
	[5,8)	10	3	32	3.33
	[8,10)	8	2	40	4

obim n = 40, > prva kum. f veća od 40/2

Težinska aritmetička sredina

Koristi se kada elementi uzorka imaju različit stepen važnost, tj. različite težine.

Težinska A.S. se računa po formuli:

$$\bar{x} = \frac{\sum(x \cdot w)}{\sum w}$$

gde je w težina pridružena elementu x.

Primer:

Različiti predmeti nose različit broj ECTS bodova, koji se mogu smatrati težinama. Ukoliko želimo da izračunamo prosečnu ocenu koja uzima u obzir ECTS bodove, računaćemo težinsku aritmetičku sredinu.

Primer (nastavak):

Tabela sa predmetima, ocenama i ECTS bodovima:

predmet	Ocena x	ECTS w
A	8	6
B	10	9
C	6	4

$$\bar{x} = \frac{\sum(x \cdot w)}{\sum w} = \frac{48+90+24}{6+9+4} = 8.53$$

Težinska AS je 8.53, a obična (8+10+6)/3 = 8.00

Oblik raspodele

Grafikonom (histogramom) predstavljena raspodela frekvencija može biti:

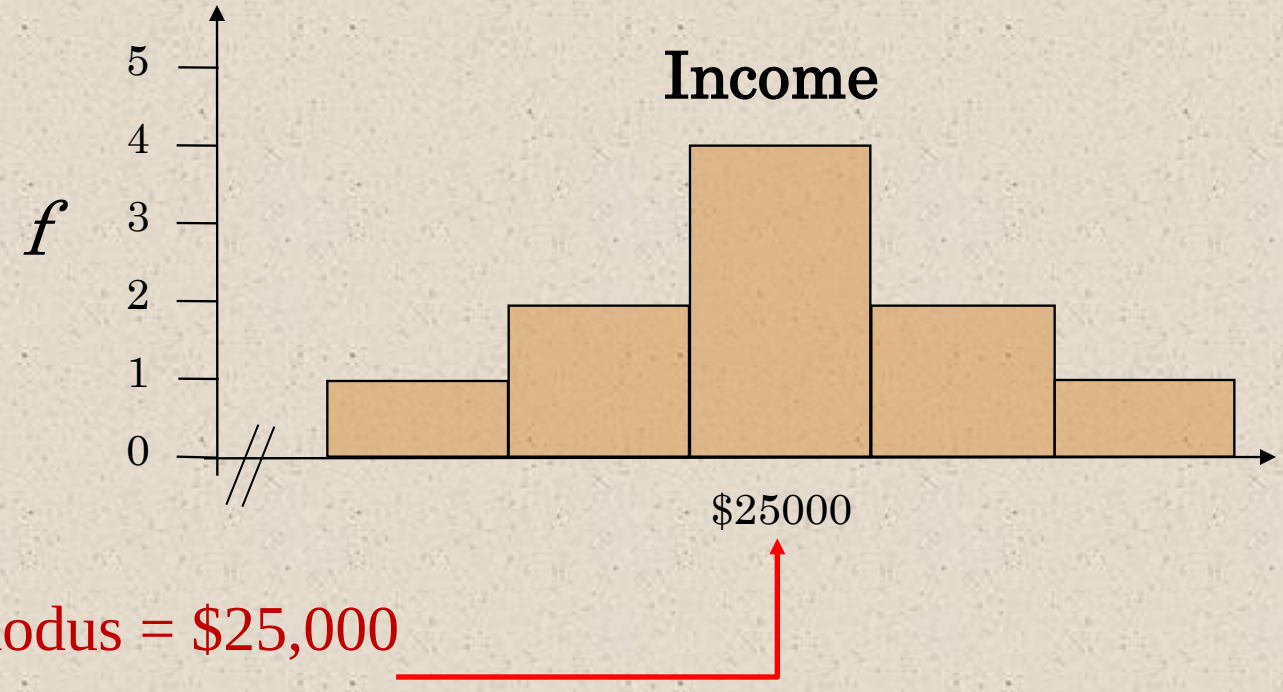
- simetrična
- asimetrična

U simetrične raspodele spadaju normalna i uniformna,

a asimetrične mogu biti iskrivljene u levo ili u desno.

Primer normalne raspodele

10 Annual Incomes
15,000
20,000
22,000
24,000
25,000
25,000
26,000
28,000
30,000
35,000

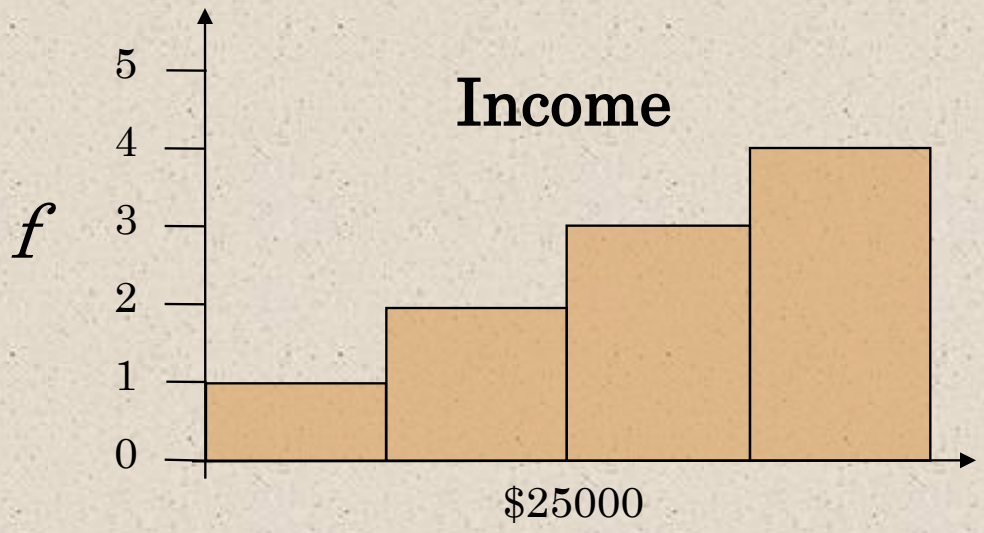


AS = medijana = modus = $25,000

Raspodela iskrivljena u levo

10 Annual Incomes
0
20,000
22,000
24,000
25,000
25,000
26,000
28,000
30,000
35,000

„rep“ se nalazi sa leve strane



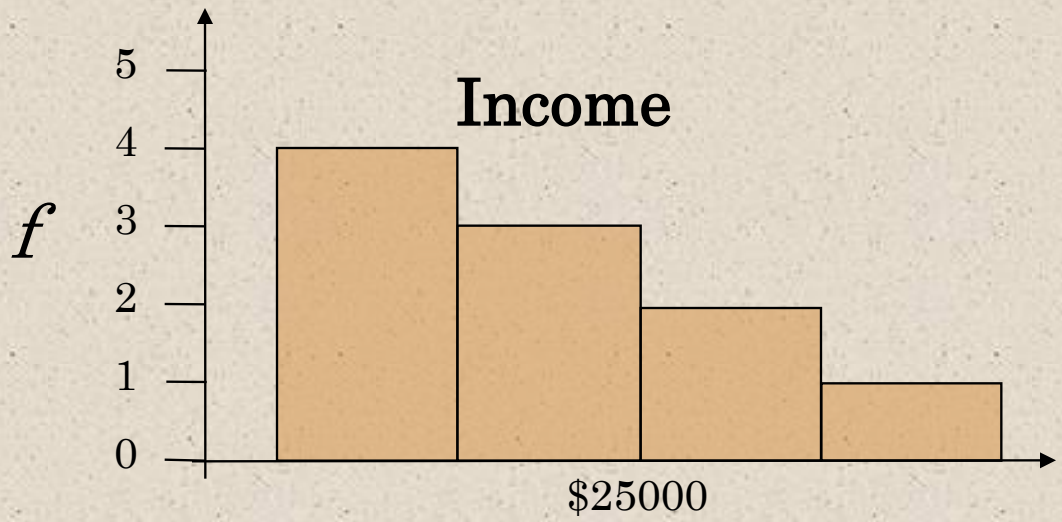
AS = $23,500
medijana = modus = $25,000

AS < Medijana

Raspodela iskrivljena u desno

10 Annual Incomes
15,000
20,000
22,000
24,000
25,000
25,000
26,000
28,000
30,000
1,000,000

„rep“ se nalazi sa desne strane

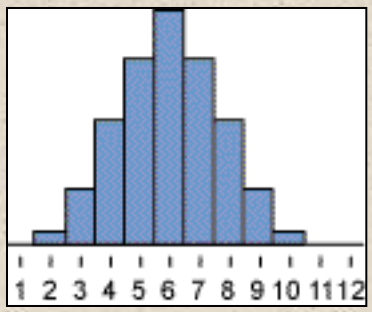


AS = $121,500
medijana = modus = $25,000

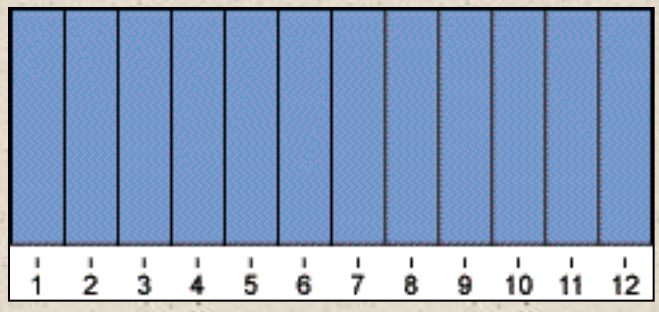
AS > Medijana

Pregled oblika raspodele

Normalna

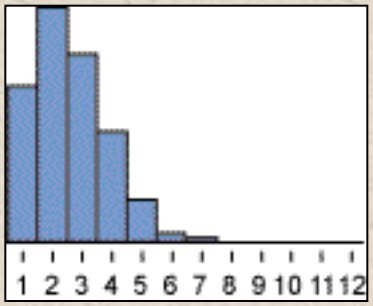


Uniformna



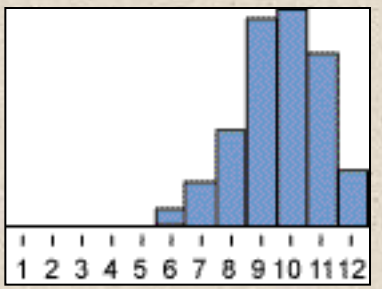
AS = Medijana

Desno iskrivljena



AS > Medijana

Levo iskrivljena



AS < Medijana

Mere CT - rezime

- Mere CT: aritmetička sredina, modus, medijana.

Primer 1: 14 studenata 2. godine nekog smera je anketirano o broju položenih ispita. Realizovani uzorak je:

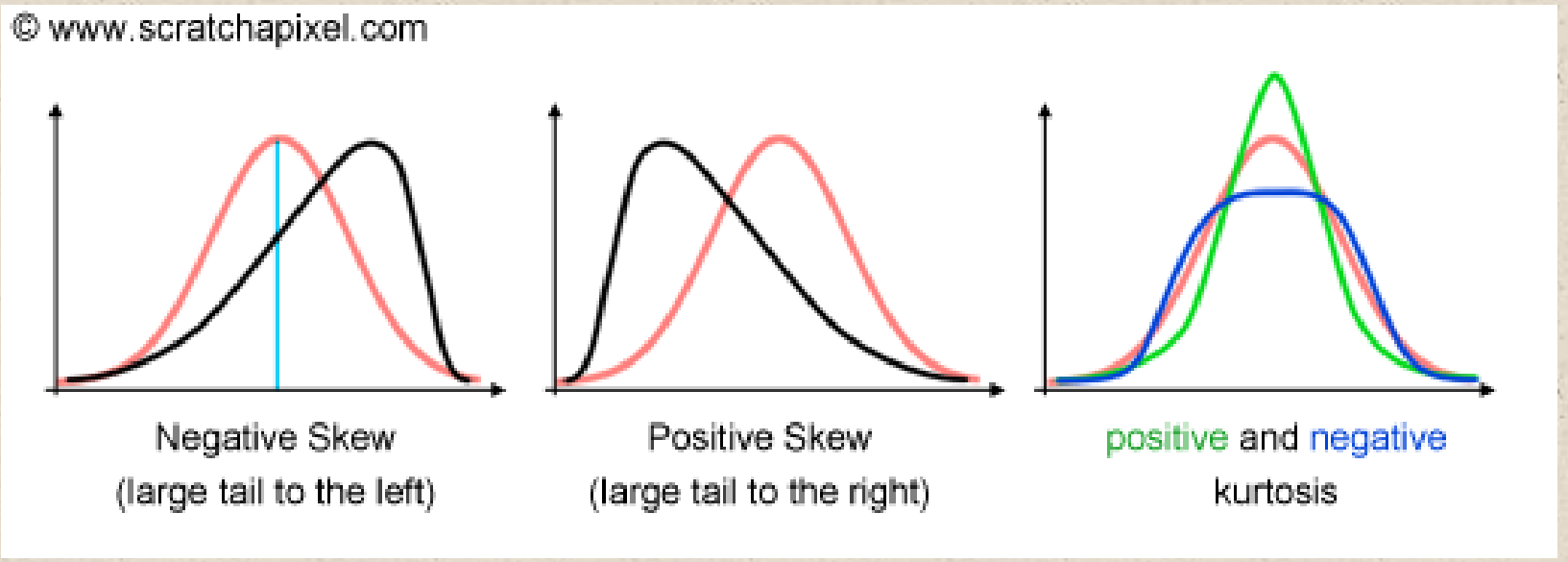
10,13,15,11,12,8,9,10,12,12,9,14,8,10

Izračunati ar. sredinu, modus i medijanu.

Primer 2: Iz tabele starosti stanara odrediti ar. sredinu, modalni i medijalni interval.

Mere oblika

- Iskrivljenost raspodele (skewness) ukazuje na stepen (a)simetričnosti.
- Spljoštenost raspodele (kurtosis) poredi vertikalni opseg grafikona sa opsegom normalne raspodele.



III DEO

Mere varijacije

Opseg (range)

Opseg podataka je razlika najveće i najmanje vrednosti uzorka, tj. maksimuma i minimuma.

opseg = (max vrednost) – (min vrednost)

Primer:

Odredi minimum, maksimum i opseg datog uzorka.

uzorak	58	56	67	56	67	63	63	67	63	57

Minimum je 56, maksimum je 67, opseg je 67 – 56 = 11.

Varijansa i standardna devijacija

Varijansa populacije

$$\sigma^2 = \frac{\sum(x-\mu)^2}{N}.$$

"sigma na kvadrat"

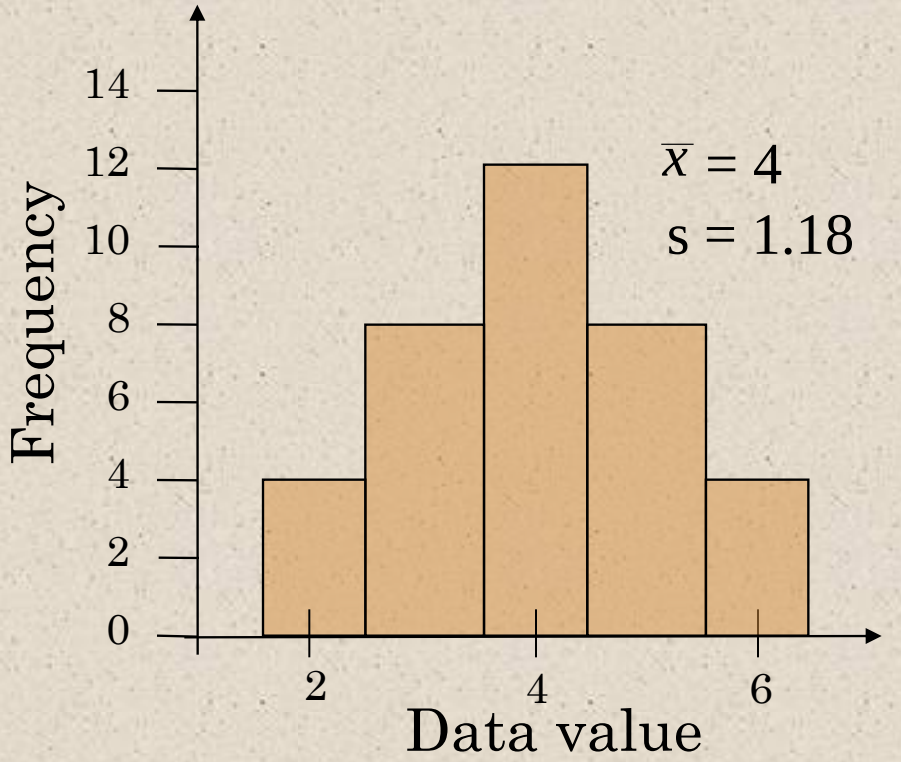
Standardna devijacija populacije

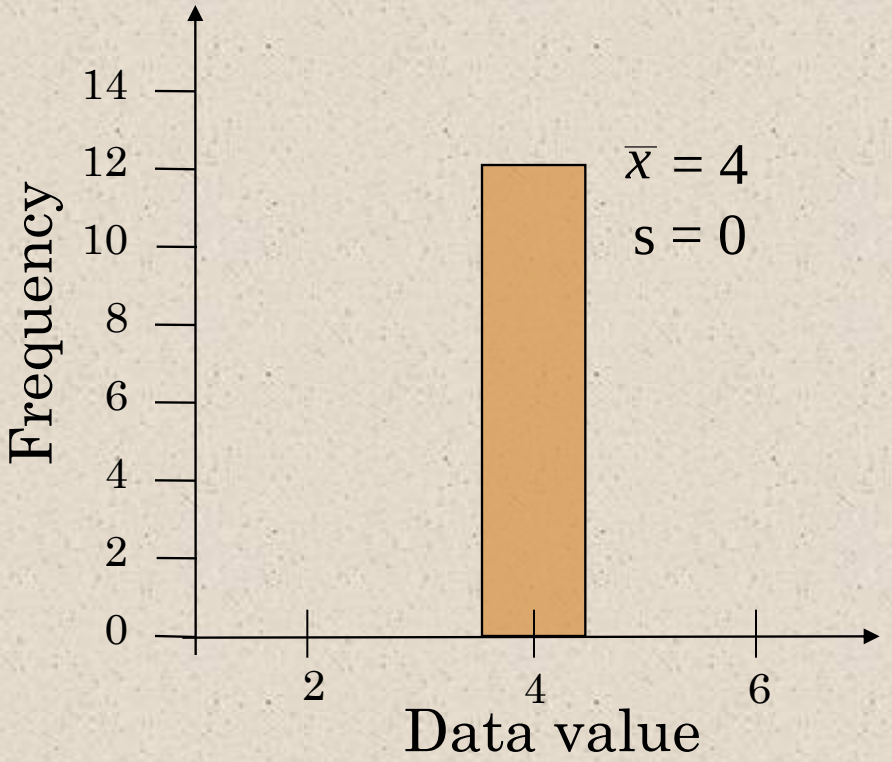
$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum(x-\mu)^2}{N}}.$$

"sigma"

Interpretacija standardne devijacije

Standardna devijacija i varijansa mere odstupanje vrednosti uzorka od sredine, što su veće to je veće rasipanje (disperzija) uzorka.



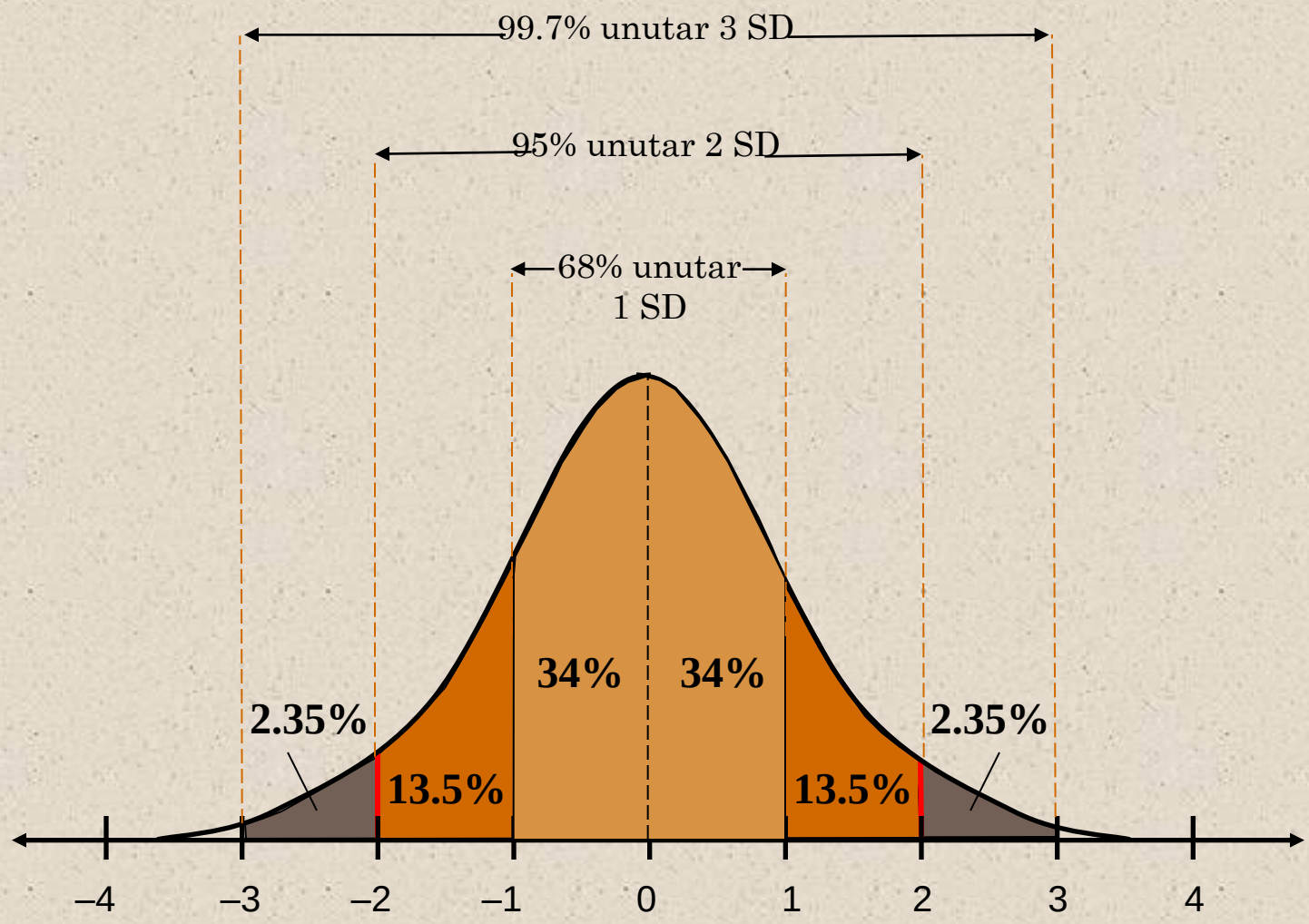


Empirijsko pravilo (68-95-99.7%)

- **Empirijsko pravilo**
- Za podatke sa (simetričnom) normalnom raspodelom, standardna devijacija SD ima sledeće karakteristike

1. Oko 68% populacije se nalazi unutar jedne SD od sredine populacije.
2. Oko 95% populacije se nalazi unutar dve SD od sredine populacije.
3. Oko 99.7% populacije se nalazi unutar tri SD od sredine populacije.

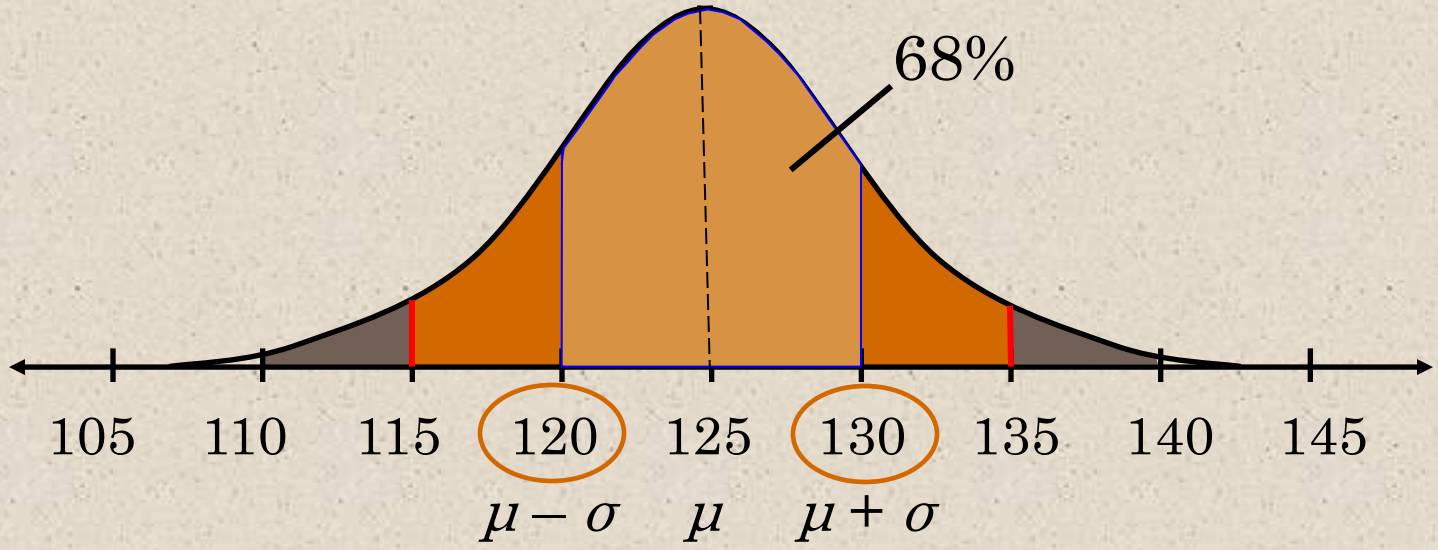
Empirijsko pravilo (68-95-99.7%)



Primena empirijskog pravila

Primer:

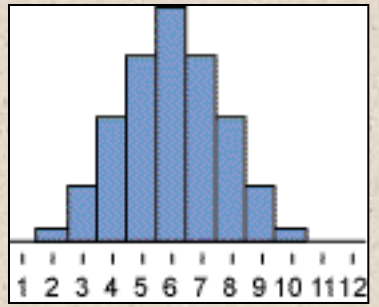
- Neka je srednja vrednost cene nekog proizvoda \$125, a standardna devijacija \$5. Poznato je da cena ima normalnu raspodelu. Proceni koji procenat proizvoda košta između \$120 and \$130.



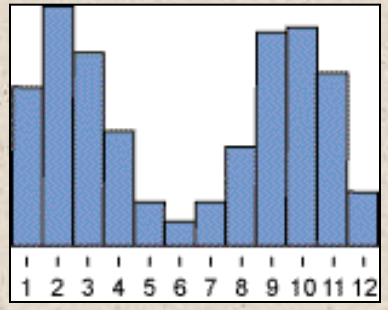
Približno 68% posmatranog proizvoda košta između \$120 i \$130.

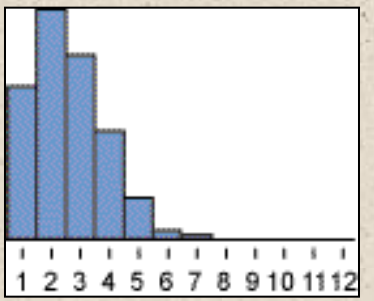
Čebiševljeva teorema

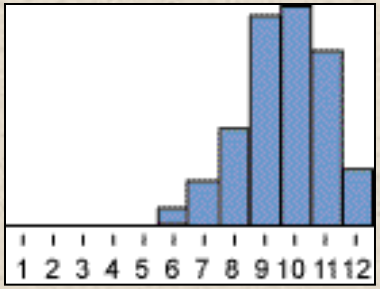
Empirijsko pravilo se može koristiti samo kod normalno raspoređenih obeležja.



Čebiševljeva teorema se može koristiti uvek, nezavisno od oblika raspodele.







Čebiševljeva teorema

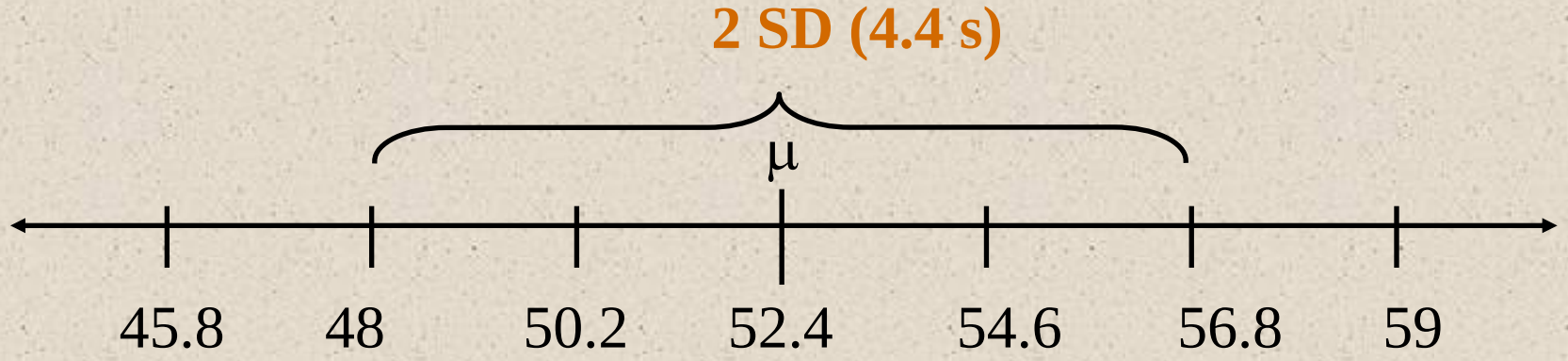
- Udeo populacije koji se nalazi unutar *k* standardnih devijacija (*k* > 1) od srednje vrednosti je bar $1-\frac{1}{k^2}$.

- Za *k* = 2: najmanje $1-\frac{1}{2^2}=1-\frac{1}{4}=\frac{3}{4}$, tj 75% se nalazi unutar 2 SD od srednje vrednosti.

- Za *k* = 3: najmanje $1-\frac{1}{3^2}=1-\frac{1}{9}=\frac{8}{9}$, tj 88.9% se nalazi unutar 3 SD od srednje vrednosti.

Primer:

Srednje vreme u trci na 400m za žene je 52.4 s sa standardnom devijacijom od 2.2 s. U kom intervalu će trčati najmanje 75% žena trkača?



Najmanje75% žena će imati rezultat između 48 i 56.8 sekundi.

Uzoračka disperzija i SD

Varijansa (S^2) i standardna devijacija (S) za grupisan uzorak

$$S^2=\sum \frac{f(x-\bar{x})^2}{n} = (\frac{1}{n}\sum f x^2)\text{-}\bar{x}^2 \qquad S=\sqrt{S^2}$$

gde je:

- n = Σf obim uzorka,
- f su frekvencije,
- x sredine intervala i
- $\bar{x}$ aritmetička sredina

Varijansa uzorka se zove i **uzoračka disperzija**

Koeficijent varijacije

- Varijansa i standardna devijacija su apsolutne mere varijacije
- Koeficijent varijacije (*CV*) je relativna mera varijacije, što znači da rezultate ne interpretiramo u fizičkim jedinicama već u procentima.

$$CV = \frac{\sigma}{\mu} * 100\% \quad \text{(za populaciju)}$$

$$CV = \frac{s}{\bar{x}} * 100\% \quad \text{(za uzorak)}$$

Obeležja kod kojih je *CV* > 100% se mogu smatrati visoko varijabilnim.

Rezime – mere varijacije

- Varijansa (uzoračka disperzija), standardna devijacija i koeficijent varijacije.
- Korektna sumirana deskripcija uzorka uključuje obim, meru CT i meru varijacije (najčešći oblik: n, AS±SD)
- Na primer: „Na uzorku od 254 ispitanika, utvrđena je prosečna neto zarada u iznosu 56.730±16.855 dinara“

Primer: Iz tabele starosti stanara izračunati SD i CV, i sumirano prikazati ispitivano obeležje.

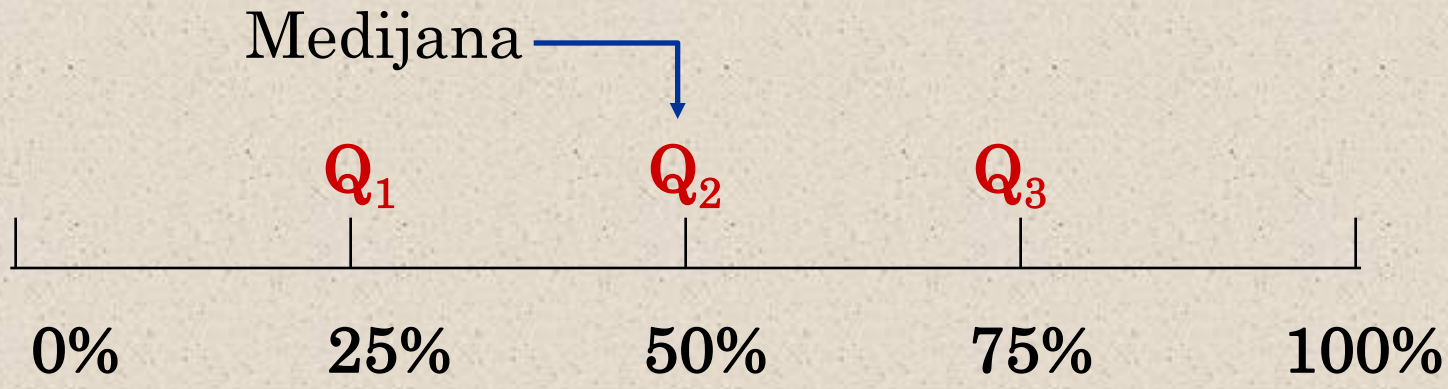
Rešenje: Ispitivanjem populacije od 90 stanara, utvrđena je prosečna starost 44.56±20.15 godina.

IV DEO

Mere pozicije

Kvartili

Tri **kvartila Q1, Q2 i Q3** približno dele sortirani uzorak na četiri jednaka dela.



Q1 je donji kvartil, Q2 je medijana, Q3 je gornji kvartil

Određivanje kvartila

Primer:

15 studenata je na testu osvojilo sledeće bodove. Odredi i protumači kvartile.

28 43 48 51 43 30 55 44 48 33 45 37 37 42 38

Potrebno je sortirati uzorak:

donja polovina gornja polovina

28 30 33 37 37 38 42 43 43 44 45 48 48 51 55

Q_1 Q_2 Q_3

Interpretacija: oko 25% studenata je osvojilo 37 bodova ili manje; oko pola studenata 43 ili manje; a oko 75% studenata je osvojilo 48 ili manje bodova.

Interkvartilni opseg

Interkvartilni opseg (IQR) je mera varijacije izražena preko kvartila:

$$\text{IQR} = Q_3 - Q_1.$$

Primer:

Odrediti i protumačiti IQR za podatke iz prethodnog primera.

$Q_1 = 37$ $Q_2 = 43$ $Q_3 = 48$

$$\begin{aligned}(\text{IQR}) &= Q_3 - Q_1 \\ &= 48 - 37 \\ &= 11\end{aligned}$$

Bodovi srednje (centralne) polovine studenata se nalaze unutar 11 bodova.

„Box and Whisker“ grafikon

Box-and-whisker graf predstavlja alat za eksploratornu analizu podataka koji ističe 5 važnih tačaka u uzorku.

Koristi sledeće vrednosti:

- Minimum
- Q_1
- Q_2 (medijana)
- Q_3
- Maksimum

Primer:

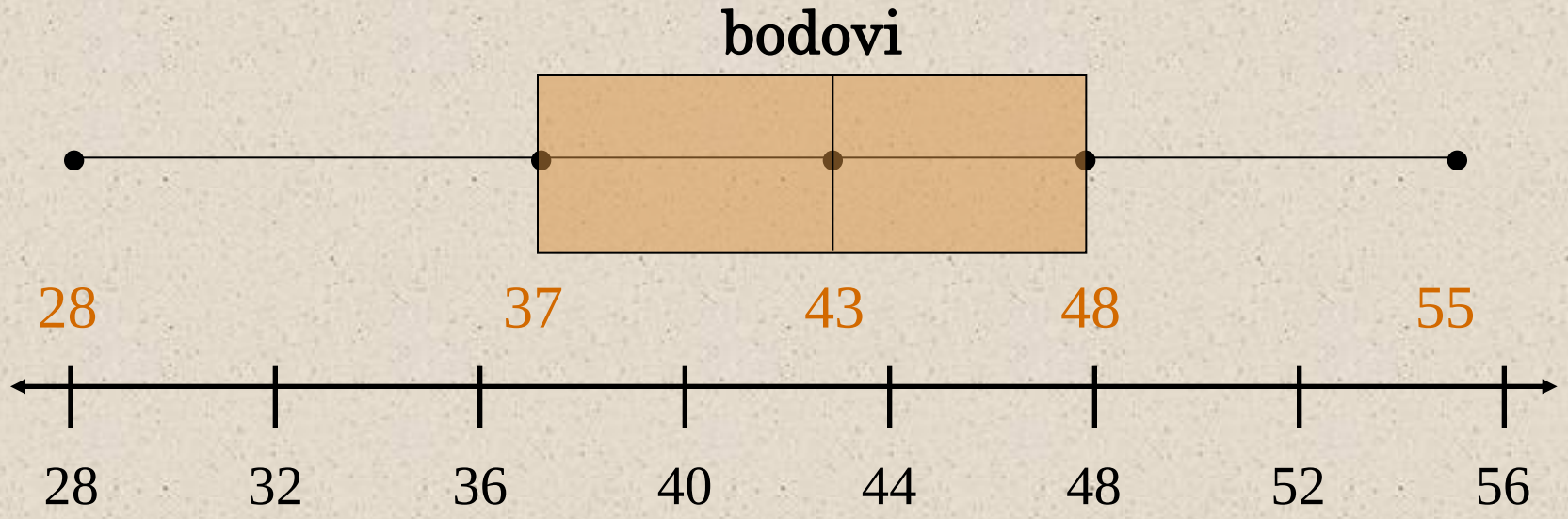
Nacrtati box-and-whiskers graf za bodove 15 studenata.

28 30 33 37 37 38 42 43 43 44 45 48 48 51 55

Box and Whisker Plot

5 tačaka:

- Minimum 28
- Q_1 37
- Q_2 43
- Q_3 48
- Maksimum 55



Percentili i decili

Percentili dele sortiran skup na 100 približno jednakih delova. Ima 99 percentila: P_1, P_2, P_3…P_{99}.

Decili dele sortiran skup na 10 približno jednakih delova. Ima 9 decila: D_1, D_2, D_3…D_9.

Ako je nečija težina na 80om percentilu (P_{80}), to znači da je on teži od 80% populacije, a da je preostalih 20% teže od njega.