

Statistikë e aplikuar



Statistika deskriptive

Faton Berisha

Kapitulli 2

Statistika deskriptive

Descriptive Statistics

- 2.1 Përshkrimi i formës së një shpërndarjeje
- 2.2 Përshkrimi i tendencës qendrore
- 2.3 Masat e variacionit
- 2.4 Persentilet, kuartilet dhe afishimet kuti
- 2.5 Përshkrimi i të dhënave kualitative
- 2.6 Përdorimi i grafikëve për të studiuar relacionin ndërmjet variablave
- 2.8 Mesataret me peshë (të ponderuara) dhe të dhënat e grupuara

Forma e një shpërndarjeje

- ❖ Për të ditur se si duket popullimi, gjendet “forma” e shpërndarjes së tij
- ❖ Paraqitet grafikisht shpërndarja me anë të cilësdo nga metodat vijuese:
 - ❖ Paraqitja trung-gjethe
 - ❖ Shpërndarja e frekuencave
 - ❖ Histogrami
 - ❖ Grafiku pikësor

Paraqitja trung-gjethe

- ❖ Qëllimi i paraqitjes trung-gjethe është të kuptohet shablloni i përgjithshëm i të dhënave, duke i grupuar të dhënat në klasa
 - ❖ Për të parë:
 - ❖ variacionin nga klasa në klasë
 - ❖ sasinë e të dhënave në secilën klasë
 - ❖ shpërndarjen e të dhënëve përbrenda secilës klasë
- ❖ Më e mira për shpërndarje të dhënash madhësish të vogla deri të mesme

Shembulli i kilometrazheve të veturave

- ❖ Shembull 2.1, Rasti i kilometrazheve të veturave
 - ❖ Të dhënat në tabelën 2.1
- ❖ Paraqitja trung-gjethe e kilometrazheve të veturave:

29	8	$29 + 0.8 = 29.8$
30	13445666889	
31	00123344455566777889	
32	0001122344556788	
33	3	$33 + 0.3 = 33.3$

2-6

Shembulli i kilometrazheve të veturave. (Vazhdim)

- ❖ Një paraqitje tjetër e të dhënave të njëjta duke përdorur më tepër klasa
 - ❖ Klasat me yllëz (*) shtrihen nga 0.0 deri 0.4
 - ❖ P.sh., rreshti me ballinën 30* përmban të dhënat nga 30.0 deri 30.4 mpg
 - ❖ Klasat pa yllëz shtrihen nga 0.5 deri 0.9
 - ❖ P.sh., rreshti me ballinën 30 përmban të dhënat nga 30.5 deri 30.9 mpg

29	8
30*	1344
30	5666889
31*	001233444
31	55566777889
32*	0001122344
32	556788
33*	3

Kilometrazhet e veturave: Rezultatet

- ❖ Nga paraqitja e fundit trung-gjethe, shpërndarja duket pothuajse “simetrike”
 - ❖ Pjesa e sipërme e paraqitjes...
 - ❖ Trungjet 29, 30*, 30, dhe 31*
 - ❖ ... është pothuajse imazh i pasqyruar i pjesës së poshtme të paraqitjes
 - ❖ Trungjet 31, 32*, 32, dhe 33*
- ❖ Por jo saktësisht refleksion pasqyre
 - ❖ Ndoshta, diçka më tepër të dhëna në pjesën e poshtme sesa në pjesën e sipërme
 - ❖ Më vonë, këtë do ta quajmë shpërndarje “të shtrembëruar me bisht nga e majta”

Shpërndarja e frekuencave

- ❖ *Shpërndarje frekuencash* është një listë klasash me numrin ose “*frekuencën*” e vlerave që i përkasin secilës klasë
 - ❖ “Klasifiko dhe numëro”
 - ❖ Shpërndarja e frekuencave është tabelë
- ❖ Shpërndarja e frekuencave grafikisht paraqitet në *histogram*
 - ❖ Histogrami është vizatim i shpërndarjes së frekuencave
- ❖ Shembujt 2.2 dhe 2.4, Rasti i kohës së pagesës

Konstruktimi i një shpërndarje frekuencash

- ❖ Hapat për konstruktimin e një shpërndarje frekuencash:
 1. Përcakto numrin K të klasave
 2. Përcakto gjerësinë e klasës
 3. Përcakto vlerën fillestare të klasave, d.m.th., kufirin e poshtëm të shpërndarjes
 4. Llogarit kufijt e klasave
 5. Vendos të gjitha klasat
- ❖ Shpërndaj të dhënat në K klasat dhe regjistro frekuencat

Numri i klasave

- ❖ Grupohen të gjitha n të dhënat në K klasa
- ❖ K është numri më i vogël i plotë për të cilin

$$2^K \geq n$$

- ❖ Në shembujt 2.2 dhe 2.4, $n = 65$
 - ❖ Për $K = 6$, $2^6 = 64, < n$
 - ❖ Për $K = 7$, $2^7 = 128, > n$
 - ❖ Prandaj përdoren $K = 7$ klasa

Gjatësia e klasës

- ❖ Gjatësia L e klasës është madhësia e hapit nga njëra te tjetra

$$L = \frac{\text{Vlera më e madhe} - \text{Vlera më e vogël}}{K}$$

- ❖ Në shembujt 2.2 dhe 2.4, vlera më e madhe është 29 ditë dhe vlera më e vogël është 10 ditë:

$$L = \frac{29 - 10 \text{ ditë}}{7 \text{ klasa}} = \frac{19 \text{ ditë}}{7 \text{ klasa}} = 2.7143 \text{ ditë/klasë}$$

- ❖ Rrumbullakohet, sipas dëshirës, gjatësia e klasës nga sipës në 3 ditë/klasë

Fillimi i klasave

- ❖ Klasat fillojnë në vlerën më të vogël të të dhënëve
 - ❖ Kjo është kufiri i poshtëm i klasës së parë
- ❖ Kufiri i sipërm i klasës së parë është
Vlera më e vogël + L
 - ❖ Në shembujt, klasa e parë fillon nga 10 ditë dhe shkon deri në 13
- ❖ Klasa e dytë fillon nga kufiri i sipërm i klasës së parë dhe vazhdon për L më tepër
 - ❖ Klasa e dytë fillon nga 13 ditë dhe shkon deri në 15 ditë
- ❖ E kështu me rradhë...

Shpërndarja dhe frekuencat:

Shembulli 2.4

Klasat (ditë)	Frekuencat
10 to 13	3
13 to 16	14
16 to 19	23
19 to 22	12
22 to 25	8
25 to 28	4
28 to 31	<u>1</u>
	65

Prova: Shuma e të gjitha frekuencave duhet të jetë n

Frekuenca relative

- ❖ *Frekuenca relative* e një klase është proporcioni (thyesa) i numrit të të dhënave që përmbahen në klasën me numrin n të të gjitha të dhënave
 - ❖ Llogaritet duke pjesëtuar frekuencën e klasës me numrin total n të vlerave të të dhënave
 - ❖ Frekuenca relative mund të shprehet si numër thyesor ose si përqindje
 - ❖ *Shpërndarje e frekuencave relative* është një listë e të gjitha klasave të të dhënave dhe frekuencave relative përkatëse të tyre

Frekuenca relative: Shembulli 2.4

Klasa (ditë)	Frekuenca	Frekuenca relative
10 to 13	3	$3/65 = 0.0462$
13 to 16	14	$14/65 = 0.2154$
16 to 19	23	0.3538
19 to 22	12	0.1846
22 to 25	8	0.1231
25 to 28	4	0.0615
28 to 31	<u>1</u>	<u>0.0154</u>
	65	1.0000

Prova: Shuma e të gjitha frekuencave relative duhet të jetë 1

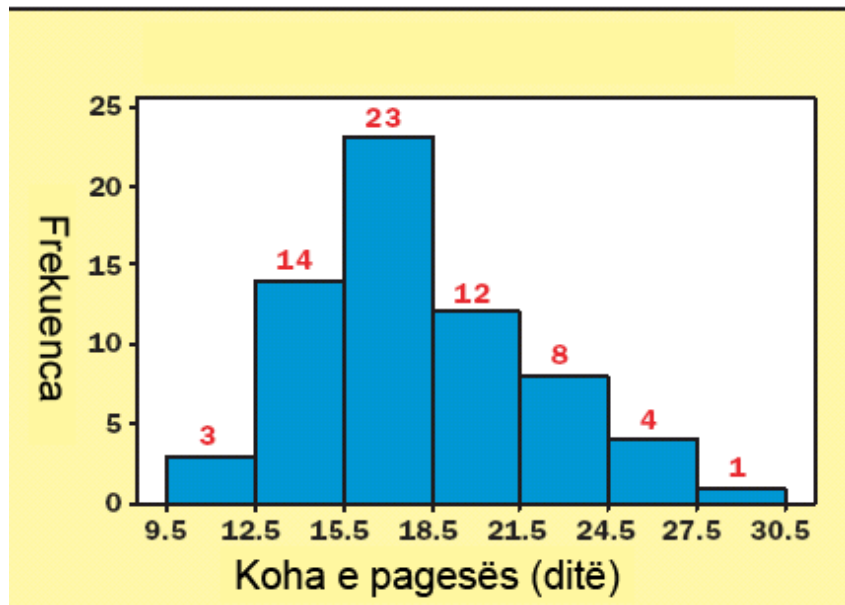
Histogrami

- ❖ Grafik në të cilin drejtkëndëshat paraqesin klasat
- ❖ Baza e drejtkëndëshit paraqet gjatësinë e klasës
- ❖ Lartësia e drejtkëndëshit paraqet
 - ❖ frekuencën në një histogram frekuencash
 - ❖ frekuencën relative në një histogram frekuencash relative

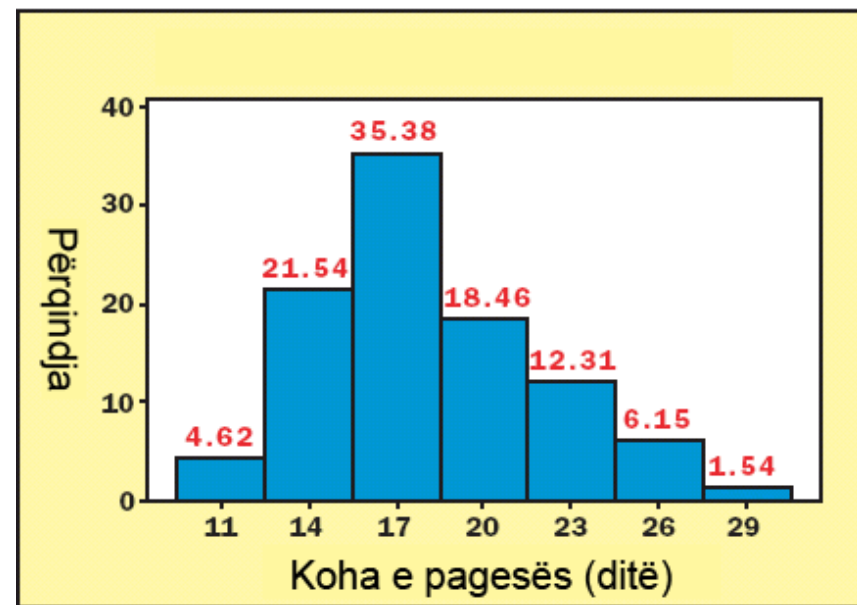
Histogramet

Shembulli 2.2: Rasti i kohëve të pagesës

Histogrami i frekuencave

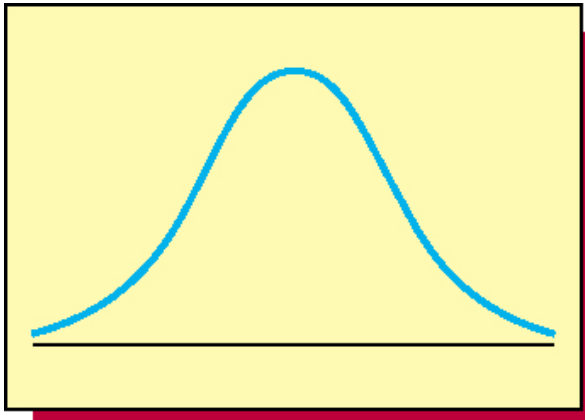


Histogrami i frekuencave relative



Sikur me rastin e paraqitjes trung-gjethë, bishti në të djathtë duket të jetë më i gjatë sesa bishti në të majtë.

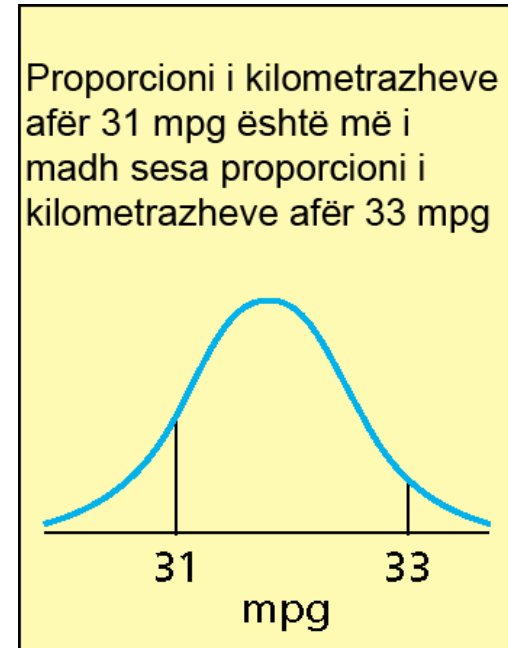
Lakorja normale



Lakore simetrike me formë zileje
për popullim me shpërndarje normale

Lartësia e normalës mbi cilëndo pikë
paraqet proporcionin relativ
të vlerave afër asaj pike

Shembulli 2.1. Rasti i kilometrazheve
të veturave

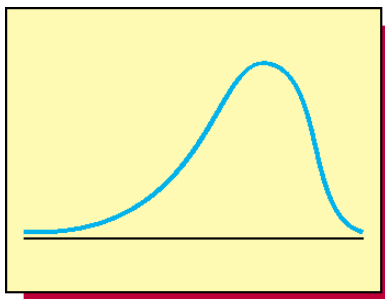


Shpërndarjet e shtrembëruara

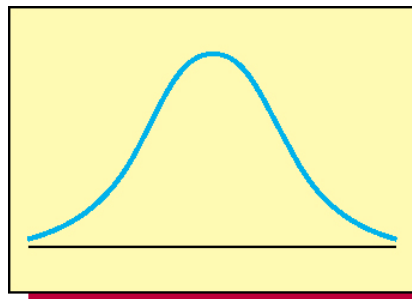
Shpërndarjet e shtrembëruara nuk janë simetrike ndaj qendrës së tyre. Kanë bisht më të gjatë nga njëra anë.

- Një popullim ka shpërndarje sipas lakores së frekuencave relative të veta.
- E shtrembëruar është ana me bisht më të gjatë

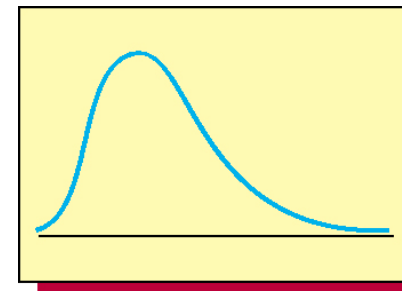
E shtrembëruar në majtë



Simetrike



E shtrembëruar në djathtë

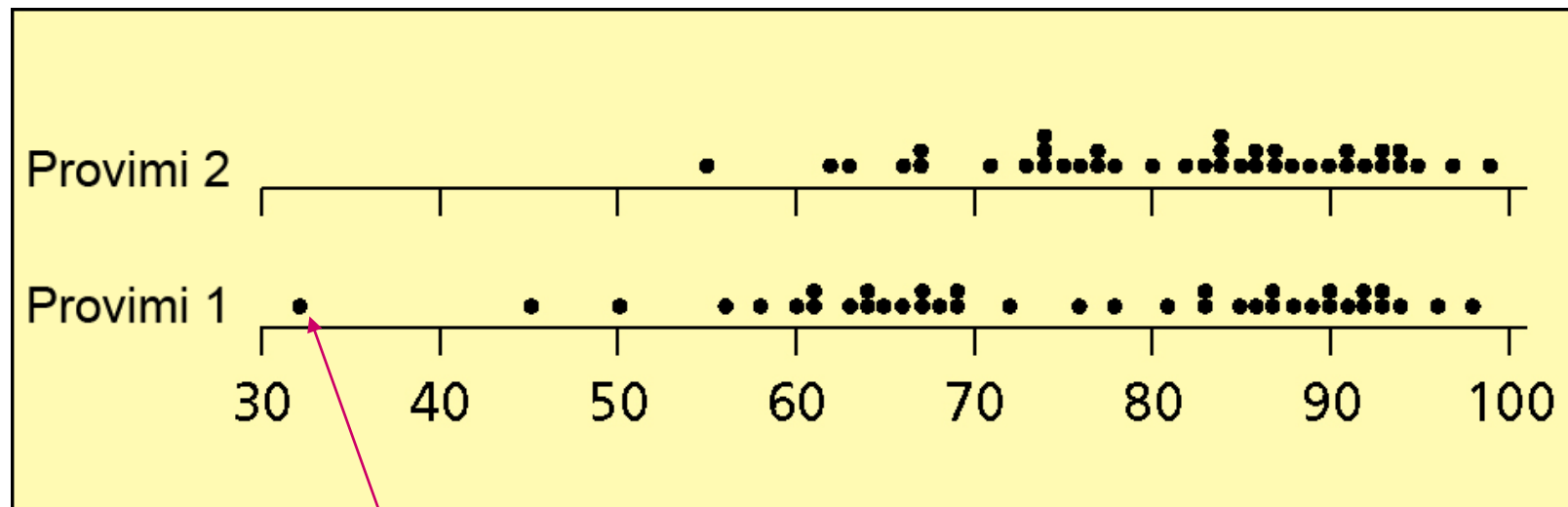


Lartësia e lakores së frekuencës relative mbi cilëndo pikë paraqet proporcionin relativ të vlerave afër asaj pike

Grafikët pikësorë

Në drejtëz numerike, secila vlerë e të dhënave paraqitet me një pikë të vendosur mbi vlerën përkatëse të shkallës

Rezultatet nga provimet 1 dhe 2



Rezultat jashtëzakonisht i ulët, prandaj “i jashtashtrirë”

Parametrat e një popullimi

Parametër popullimi është një numër i llogaritur nga të gjitha masat e popullimit që përshkruan ndonjë aspekt të popullimit.

Mesatarja e popullimit, shënohet me μ , është parametër popullimi dhe është mesi aritmetik i masave të popullimit.

Vlerësimet pikësore dhe statistikat e mostrës

Vlerësim pikësor është një vlerësim njënumërsht i vlerës së një parametri të popullimit.

Statistikë mostre është një numër i llogaritur duke përdorur masat e mostrës që përshkruan ndonjë aspekt të mostrës.

❖ Statistikat e mostrës përdoren si vlerësime pikësore të parametrave të popullimit. —

Mesatarja e mostrës, shënohet me \bar{X} , është statistikë mostre dhe është mesi aritmetik i masave të mostrës.

❖ Mesatarja e mostrës është vlerësim pikësor i mesatarës së popullimit.

Masat e tendencës qendrore

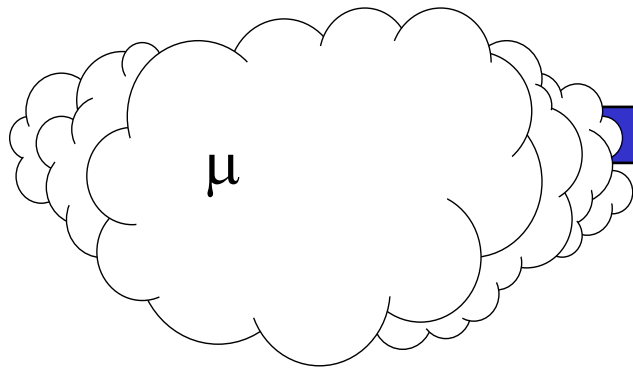
Mesatarja, μ Vlera e mesit aritmetik, ose e pritur.

Mediana, M_d Vlera e pikës së mesme të masave të renditura

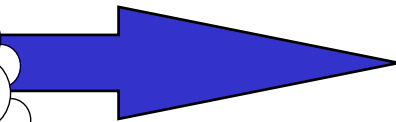
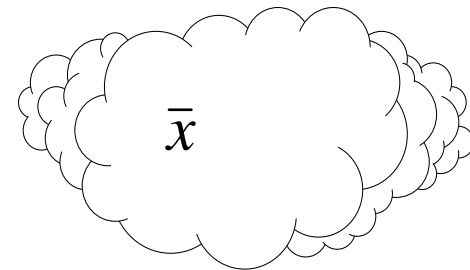
Moda, M_o Vlera më e frekuentuar
(me frekuencën më të madhe)

Mesatarja

Popullimi X_1, X_2, \dots, X_N



Mostra x_1, x_2, \dots, x_n



Mesatarja e popullimit

$$\mu = \frac{\sum_{i=1}^N X_i}{N}$$

Mesatarja e mostrës

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

Mesatarja e mostrës

Për një mostër madhësie n , *mesatarja e mostrës* përkufizohet me

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

dhe është vlerësim pikësor i mesatares së popullimit μ

- ❖ Është vlera për t'u pritur,
në të mesmen aritmetike dhe në kohë të gjatë.

Shembull: Rasti i kilometrazheve të veturave

Shembulli 2.5: Mesatarja e mostrës për pesë kilometrazhet e para të veturave nga tabela 2.1

30.8, 31.7, 30.1, 31.6, 32.1

$$\bar{x} = \frac{\sum_{i=1}^5 x_i}{5} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5}$$

$$\bar{x} = \frac{30.8 + 31.7 + 30.1 + 31.6 + 32.1}{5} = \frac{156.3}{5} = 31.26$$

Mediana

Mediana M_d e një popullimi ose mostre është vlerë e tillë që 50% të të gjitha masave, pasi të jenë renditur në renditje numerike, shtrihet mbi (ose nën) të.

Mediana M_d gjendet si vijon:

1. Në qoftë se numri i masave është tek, mediana është masa e ndërmjeme në vlerat e renditura
2. Në qoftë se numri i masave është çift, mediana është mesi aritmetik i dz y vlerave të ndërmjeme në vlerat e renditura

Shembull: Mediana e mostrës

Shembull 2.6: Pagat vjetore të internistëve (x\$1000)

127 132 138 141 144 146 **152** 154 165 171 177 192 241

(Vëreni se vlerat janë në renditje numerike rritëse nga e majta në të djathtë)

Meqë $n = 13$ (tek), është vlera e ndërmjeme ose e 7^{ta} në të dhënat e renditura, prandaj

$$M_d = 152$$

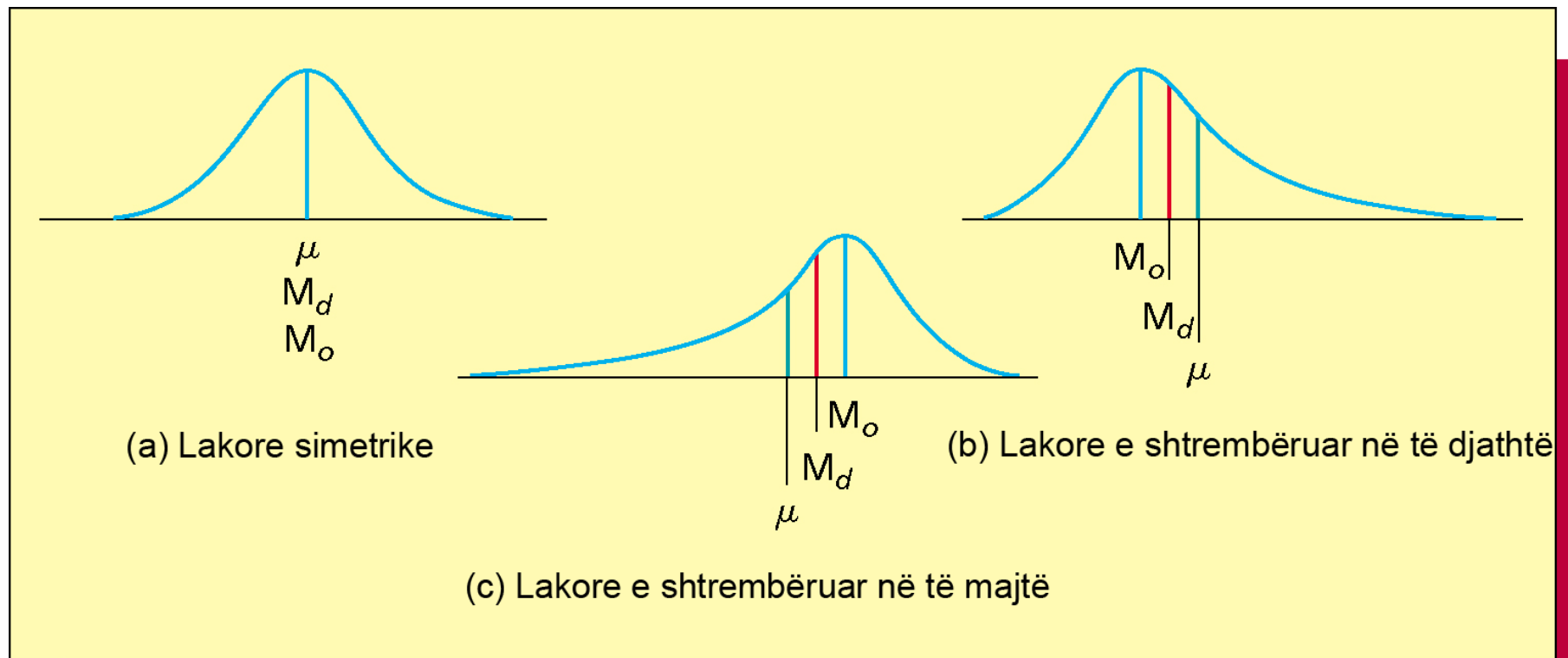
- ❖ Një pagë vjetore prej \$180,000 është në krahun e sipërm, goxha mbi pagën median e prej \$152,000
- ❖ Në fakt, \$180,000 është një page shumë e lartë dhe kompetitive

Moda

Moda M_o e një popullimi ose mostre masash është masa e cila paraqitet më së shpeshti.

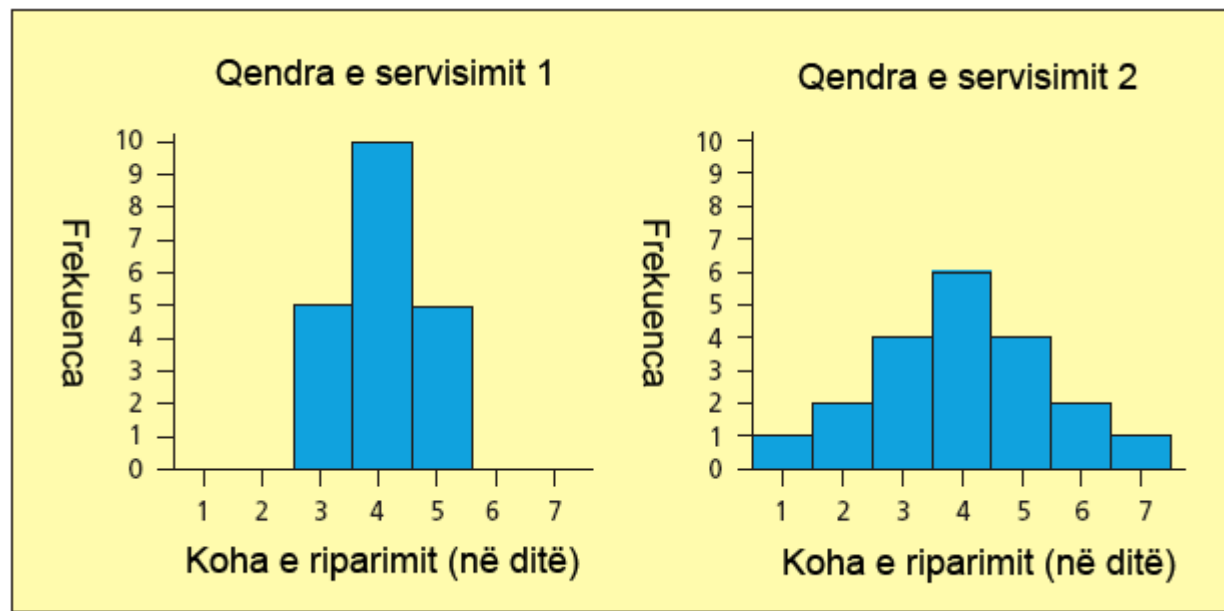
- ❖ Modat janë vlerat “më tipike” të observuara.
- ❖ Ndonjëherë frekuencat më të larta në dy ose më tepër vlera
 - ❖ Në qoftë se ka dy moda, të dhënat janë bimodale
 - ❖ Në qoftë se ka më tepër se dy moda, të dhënat janë multimodale.
- ❖ Kur të dhënat janë në klasa, klasa me frekuencë më të madhe është *klasa modale*.
 - ❖ Drejtkëndëshi më i lartë në histogram

Relacionet ndërmjet mesatares, medianës dhe mode



Tendenca qendrore nuk mjafton

- ❖ Njohja e masave të tendencës qendrore nuk mjafton.
- ❖ Të dy shpërndarje e mëposhtme kanë masa identike të tendencës qendrore.



Masat e variacionit

Rangu

Masa më e madhe (maksimale) minus
masa më e vogël (minimale)

Variance

Mesatarja e katrorëve të devijimeve
të të gjithamasave të popullimit nga mesatarja
e popullimit

Standard Deviation

Rrënja katrore e variancës

Rangu

Rangu = Vlera më e madhe – Vlera më e vogël

Rangu mat intervalin e shtrirjes së të gjitha të dhënave.

Shembull:

Pagat e internistëve (në mijë dollarë)

127 132 138 141 144 146 152 154 165 171 177 192 241

$\text{Rangu} = 241 - 127 = 114$ (\$114,000)

Varianca

Për popullim të madhësisë N , *varianca e popullimit* σ^2 përkufizohet me

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N} = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_N - \mu)^2}{N}$$

Për mostër të madhësisë n , *varianca e popullimit* s^2 përkufizohet me

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1}$$

dhe është vlerësim pikësor për σ^2

Devijimi standard

Devijimi standard i popullimit, σ : $\sigma = \sqrt{\sigma^2}$

Devijimi standard i mostrës, s : $s = \sqrt{s^2}$

Shembull: Varianca dhe devijimi standard i popullimit

Popullimi i margjinave të profitit për pesë kompani të mëdha:
8%, 10%, 15%, 12%, 5%

$$\mu = \frac{8 + 10 + 15 + 12 + 5}{5} = \frac{50}{5} = 10 \%$$

$$\begin{aligned}\sigma^2 &= \frac{(8 - 10)^2 + (10 - 10)^2 + (15 - 10)^2 + (12 - 10)^2 + (5 - 10)^2}{5} \\ &= \frac{(-2)^2 + 0^2 + 5^2 + 2^2 + (-5)^2}{5} \\ &= \frac{4 + 0 + 25 + 4 + 25}{5} = \frac{58}{5} = 11.6\end{aligned}$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{11.6} = 3.406 \%$$

Shembull: Varianca dhe devijimi standard i popullimit

Shembulli 2.12: Varianca dhe devijimi standard i mostrës për pesë kilometrazhet e para të veturave nga tabela 2.1

$$s^2 = \frac{\sum_{i=1}^5 (x_i - \bar{x})^2}{5 - 1} \quad 30.8, 31.7, 30.1, 31.6, 32.1 \text{ so } Z = 31.26$$
$$= \frac{(30.8 - 31.26)^2 + (31.7 - 31.26)^2 + (30.1 - 31.26)^2 + (31.6 - 31.26)^2 + (32.1 - 31.26)^2}{4}$$

$$s^2 = 2.572 / 4 = 0.643$$

$$s = \sqrt{s^2} = \sqrt{.643} = 0.8019$$

Rregulla empirike për popullime normale

Në qoftë se një popullim ka mesatare μ dhe devijim standard σ dhe përshkruhet nga një lakore normale, atëherë

68.26% e masave të popullimit shtrihen brenda një devijimi standard nga mesatarja: $[\mu - \sigma, \mu + \sigma]$

95.44% e masave të popullimit shtrihen brenda dy devijimesh standarde nga mesatarja: $[\mu - 2\sigma, \mu + 2\sigma]$

99.73% e masave të popullimit shtrihen brenda tri devijimesh standarde nga mesatarja: $[\mu - 3\sigma, \mu + 3\sigma]$

Teorema e Chebyshev-it

Në qoftë se μ dhe σ janë mesatarja dhe devijimi standard i një popullimi, atëherë për çdo vlerë $k > 1$

Së paku $100(1 - 1/k^2)\%$ e masave të popullimit shtrihen në intervalin:

$$[\mu - k\sigma, \mu + k\sigma]$$

Shfrytëzohet vetëm për shpërndarjet jonormale:

- ❖ Të shtrembëruar
 - ❖ Por jo ekstremisht të shtrembëruar
- ❖ Bimodale

Z-Vlerat

- ❖ Për çfarëdo x në një popullim ose mostër, z-vlera përkatëse është

$$z = \frac{x - \text{Mesatarja}}{\text{Devijimi standard}}$$

- ❖ z-Vlera është numri i devijimeve standarde për të cilat x shtrihet nga mesatarja.
 - ❖ z-Vlerë pozitive është për x mbi (më të madhe se) mesatare.
 - ❖ z-Vlerë negative është për x nën (më të vogël se) mesatare.

Koeficienti i variacionit

- ❖ Matë madhësinë e devijimit standard relativisht me madhësinë e mesatares.
- ❖ Koeficienti i variacionit = $\frac{\text{Devijimi standard}}{\text{Mesatarja}} \cdot 100\%$
- ❖ Përdoret për të:
 - ❖ Krahasuar ndryshueshmërinë relative të vlerave rreth mesatares.
 - ❖ Krahasuar ndryshueshmërinë relative të popullimeve ose mostrave me mesatare të ndryshme dhe devijime standarde të ndryshme.
 - ❖ Matur rrezikun.