



Seminarski rad



KVALITET PODATAKA

Predmet: Prikupljanje i predobrada
podataka za mašinsko učenje

Mentor:

Aleksandar Stanimirović

Kandidat:
Filip Trajković 1574

SADRŽAJ

[1] UVOD

[2]
Kvalitet
podataka

[3]
Raspodela
podataka

[4]
Korelacija

[5]
Varijansa

[6] ZAKLJUČAK

Uvod

U toku preprocesiranja podataka neophodno je obaviti niz koraka kako bi se podaci pripremili za proces obučavanja modela mašinskog učenja.



Kvalitet podataka

Kvalitet podataka se odnosi na razvoj i implementaciju aktivnosti koje primenjuju tehnike upravljanja kvalitetom na podatke kako bi se osiguralo da podaci odgovaraju specifičnim potrebama organizacije u određenom kontekstu.

Podaci za koje se smatra da odgovaraju njihovoj nameni smatraju se podacima visokog kvaliteta. Tako dobijeni podaci visokog kvaliteta predstavljaju adekvatnu grupu uzoraka nad kojima se kasnije mogu vršiti analize uspešnosti algoritama, problema u toku modelovanja i dr.



Pojam kvaliteta podataka

Pojam kvaliteta podataka predstavlja **kvalitativnu meru dobijenog uzorka čija se vrednost u procesu procene kvaliteta proverava u odnosu na prethodno stečeno znanje iz datog domena.**

Vrednost kvaliteta svakog pojedinačnog uzorka se uz pomoć mera kvaliteta proverava na osnovu čega se donosi konačni zaključak o **kvalitativnim svojstvima posmatranog uzorka.**



Mere kvaliteta podataka



Osnovne mere kvaliteta su:

- Tačnost
- Jedinstvenost
- Konzistentnost
- Potpunost
- Relevantnost
- Pravovremenost



Tačnost

Predstavlja meru kvaliteta podataka koja definiše vrednost odstupanja podataka od stvarne ili ispravne vrednosti originalnog podatka.

		True Class	
		Positive	Negative
Predicted Class	Positive	TP	FP
	Negative	FN	TN

Tačnost je validan izbor evaluacije za probleme klasifikacije podataka koji su dobro izbalansirani i bez velikog broja "outlier"-a.

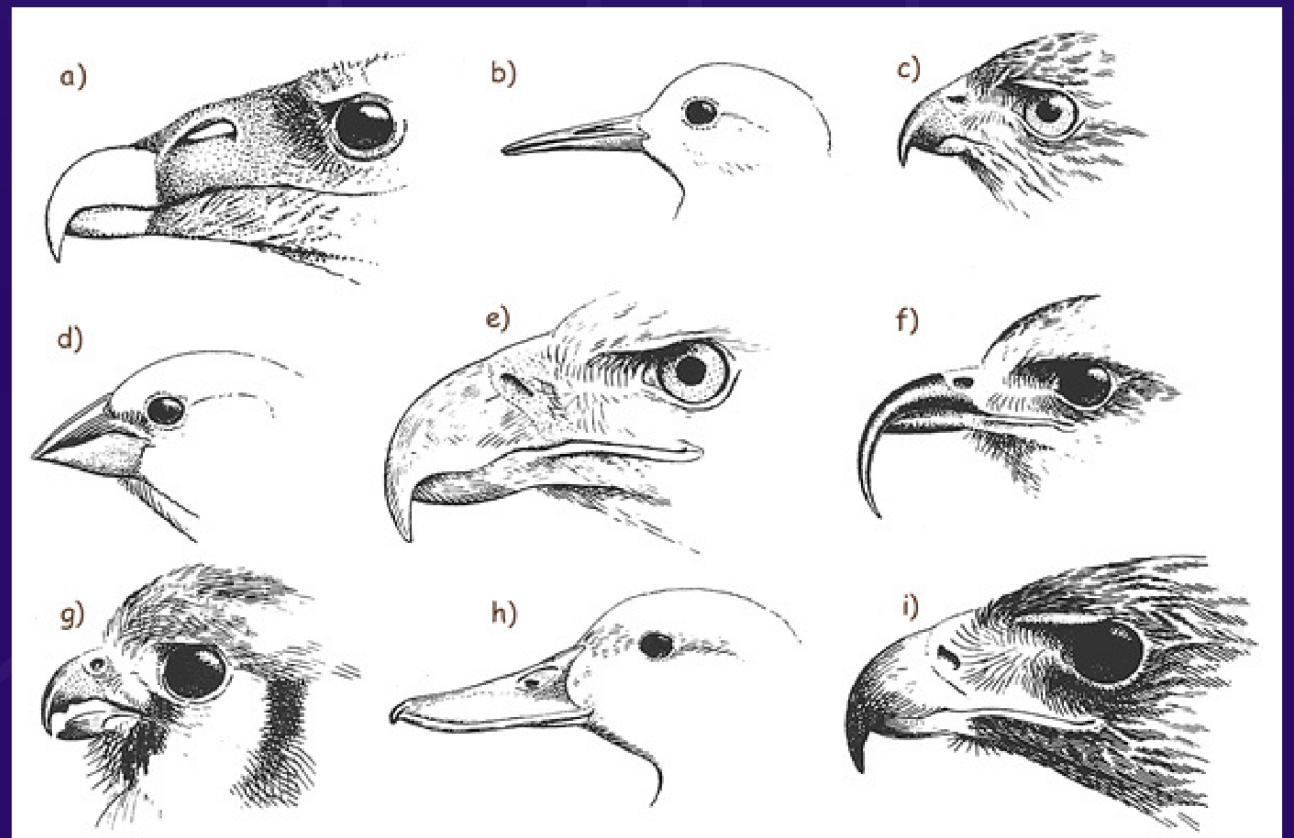
$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$



Jedinstvenost

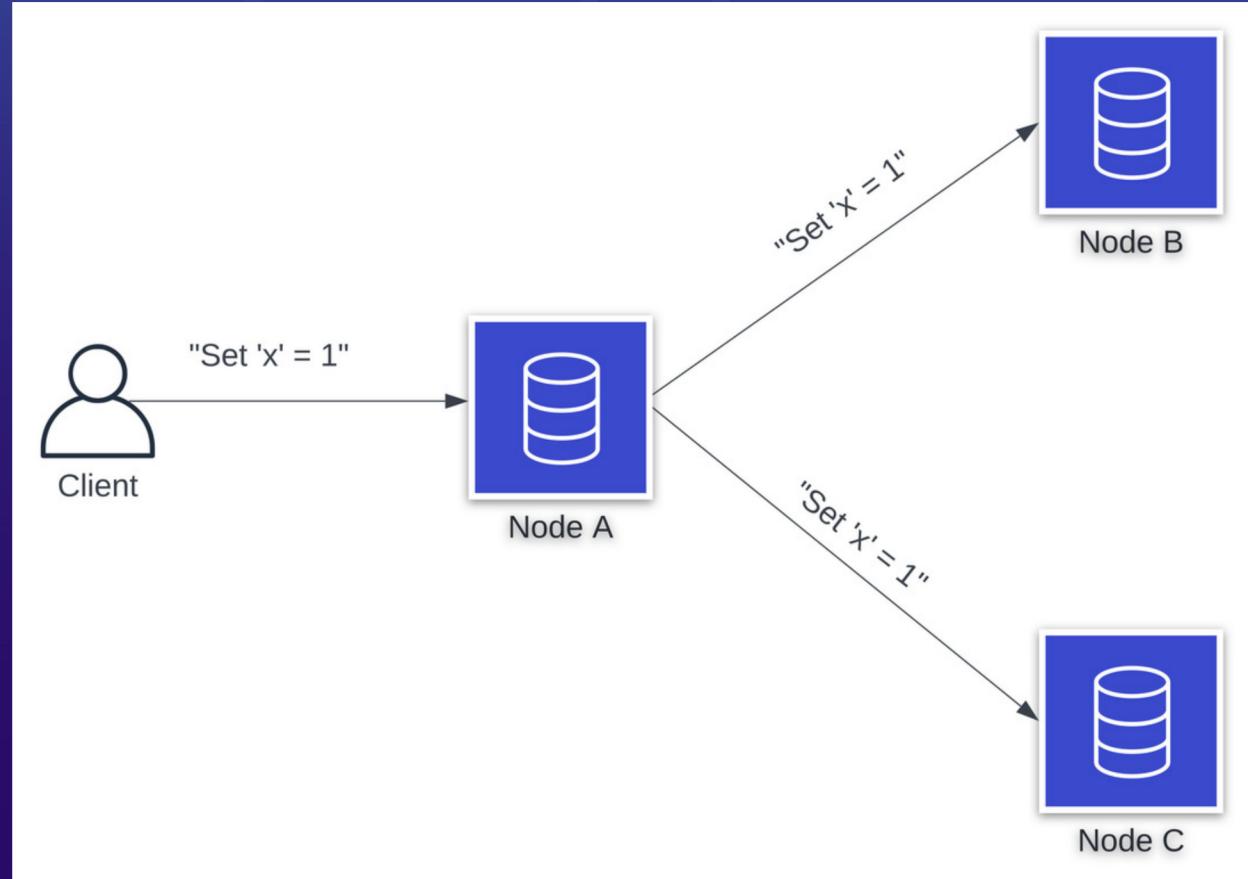
Jedinstvenost predstavlja osobinu podataka koja se odnosi na svaku pojedinačnu stavku u podacima gde se **većim kvalitetom podrazumeva i veća količina jedinstvenih podataka.**

Jedinstvenost jeste **suprotnost multiplikativnosti podataka** u tabeli podataka. Multiplikativnost dovodi do povećanja obima skupa podataka bez unošenja varijabilnosti.



Konzistentnost

Konzistentnost podataka se odnosi na uniformnost podataka dok se kreću kroz mreže i aplikacije.



Konzistentnost proverava da li su vrednosti podataka uskladištene za isti zapis u različitim izvorima bez kontradiktornosti i da li su potpuno iste – u smislu značenja, kao i strukture i formata.



Potpunost



Potpunost podataka se odnosi na broj popunjениh vrednosti unutar skupa podataka što doprinosi celovitosti ili sveobuhvatnosti skupa podataka.

Kako bi skup podataka bio potpun neophodno je da ne postoje praznine ili nedostajuće informacije koje na taj način skup podataka čine težim za obradu i izvođenje daljih zaključaka iz podataka.

Name	Gender	Age	No. of Children
Mary Jane	F	33	3
Jason Charles	M	56	
Isaac Brown	M	22	

Relevantnost



Relevantnost podataka se odnosi na **stepen značaja informacija** koje se nalaze u određenom skupu podataka.

Relevantnost podataka pokazuje **korisnost** tipa podataka koji se pribavlja, njegovu **kvalitativnu vrednost** i **stepen iskorišćenosti** tog podatka u procesu obrade i analize skupa podataka.

Pravovremenost

Pravovremenost podataka predstavlja meru kvaliteta podataka koja se odnosi na **dostupnost i ažuriranost** podataka u određenom vremenskom trenutku.



Raspodela podataka

Raspodela podataka predstavlja meru koja opisuje **način distribucije podataka** u odnosu na celokupan skup podataka prema **učestalosti ponavljanja podataka** u rasponu od minimalne do maksimalne vrednosti.

Raspodela podataka se predstavlja grafikom vrednosti pri čemu su na **X-osi predstavljene sve vrednosti** datog atributa, dok su na **Y-osi predstavljene učestalosti tih vrednosti** u skupu podataka.

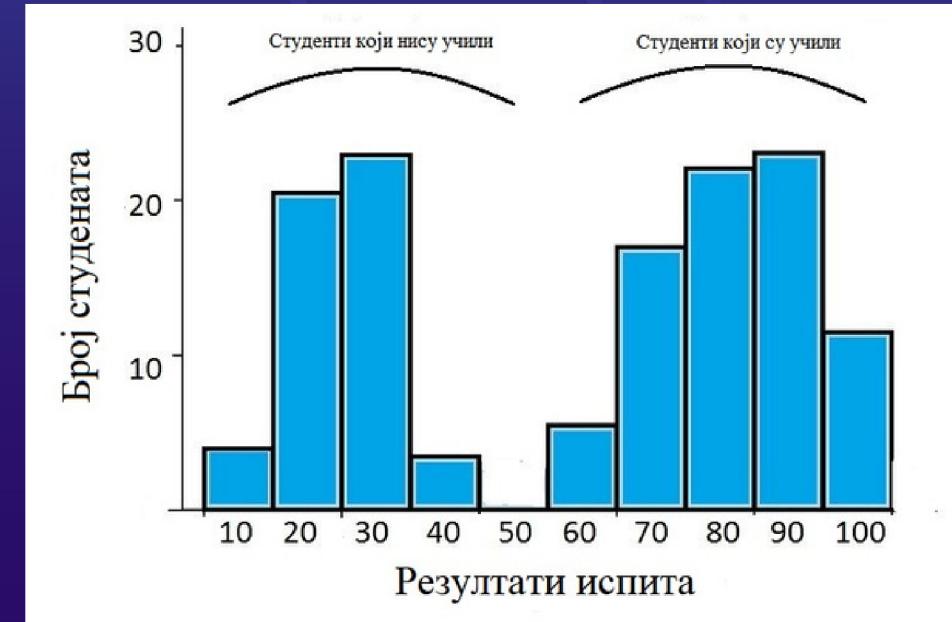


Tipovi raspodele podataka

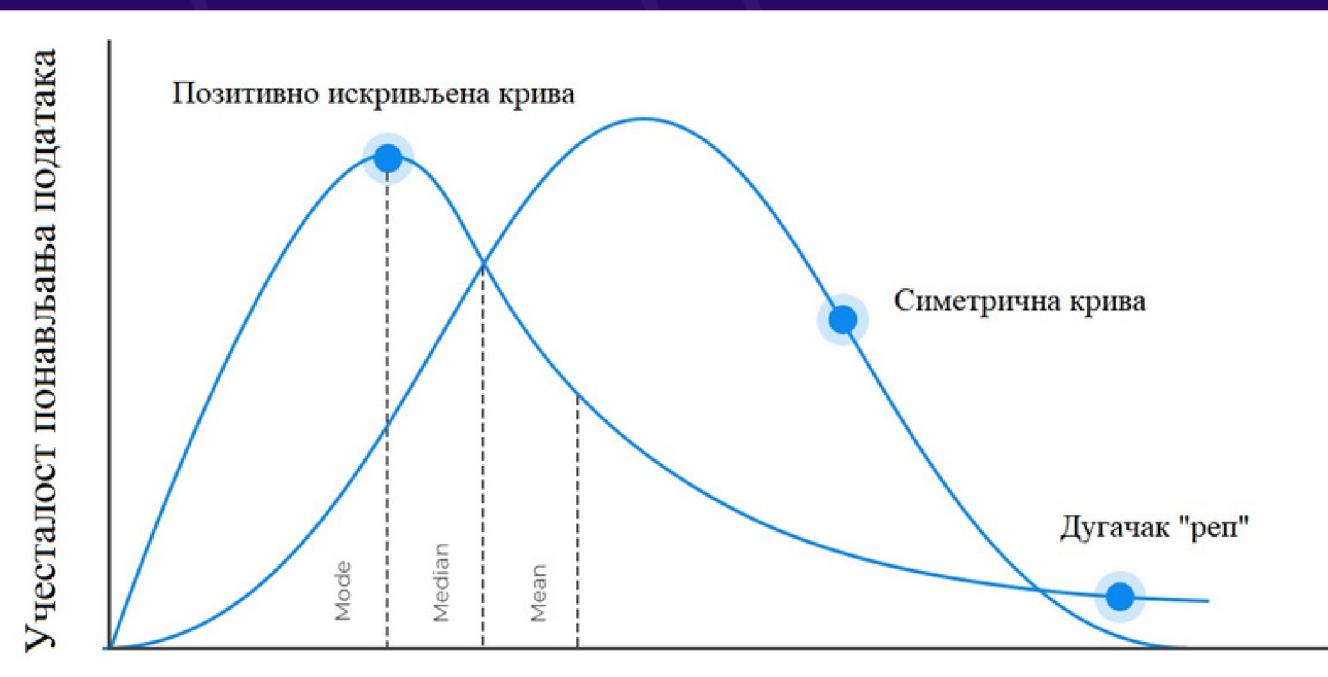
- Simetrična raspodela



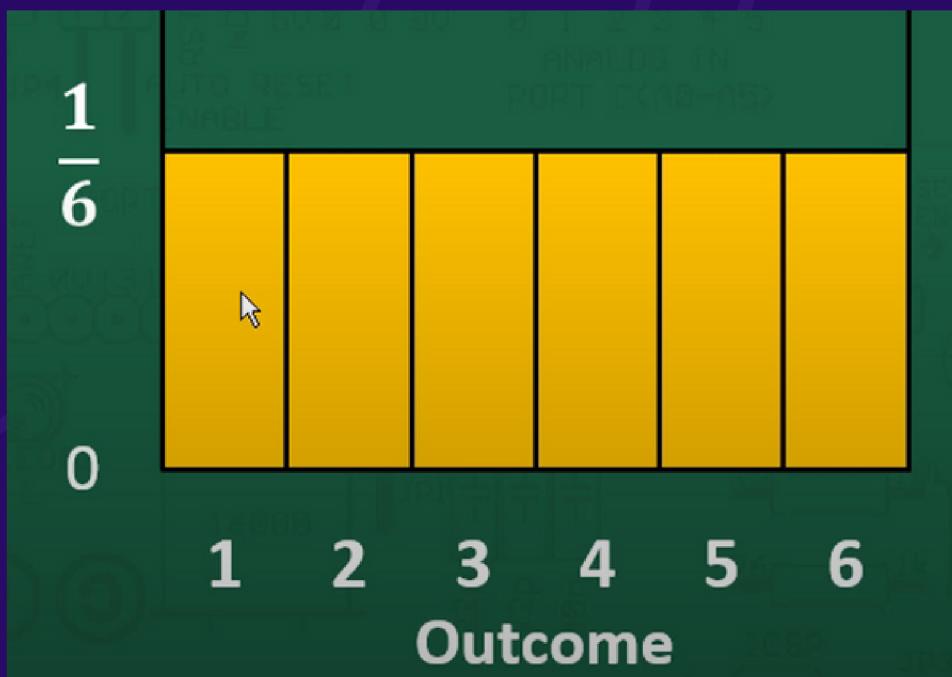
- Bimodalna raspodela



- Asimetrična raspodela



- Uniformna raspodela



Mere centralne tendencije

Mere centralne tendencije predstavljaju familiju vrednosti koje imaju za cilj da celokupan skup podataka opišu uz pomoć jedne jedinstvene vrednosti. Ovakvim predstavljanjem, teži se da se skup podataka predstavi nekim zajedničkim svojstvom.

Najbitniji elementi ove familije vrednosti su:

- Aritmetička srednja vrednost
- Modus
- Medijana



Aritmetička srednja vrednost

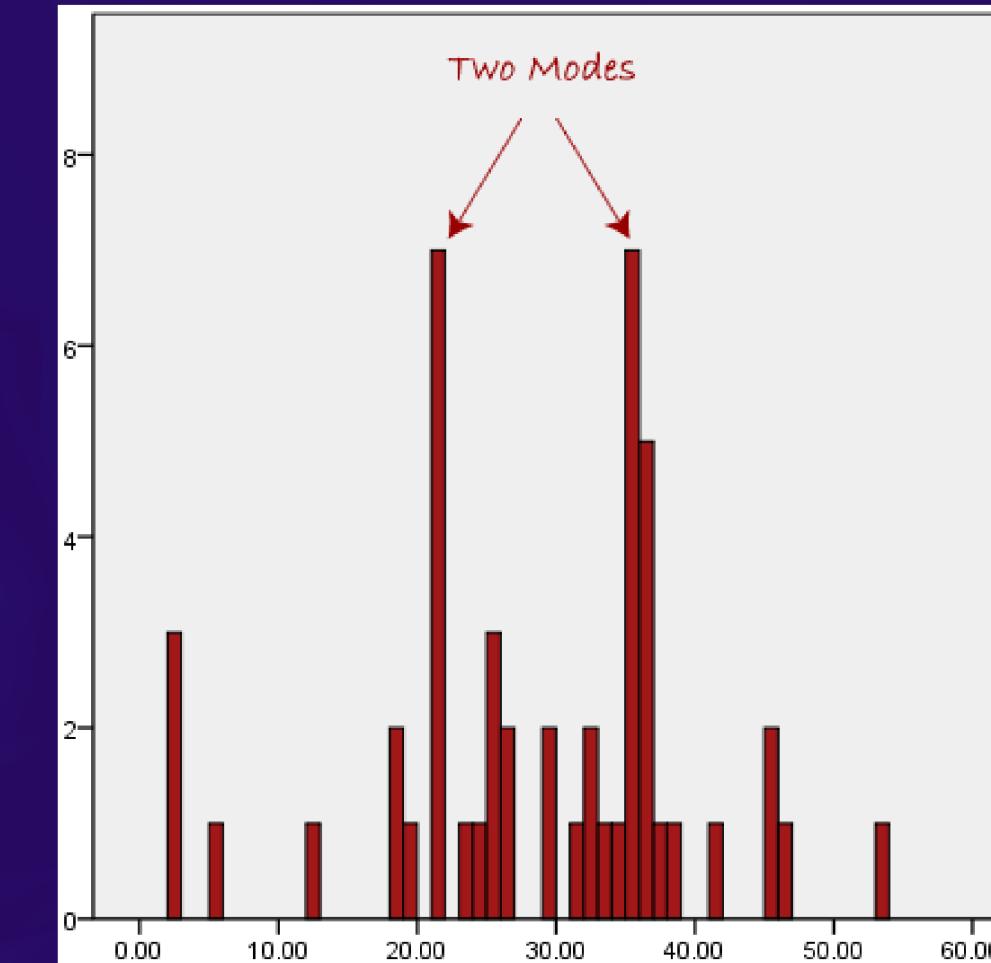
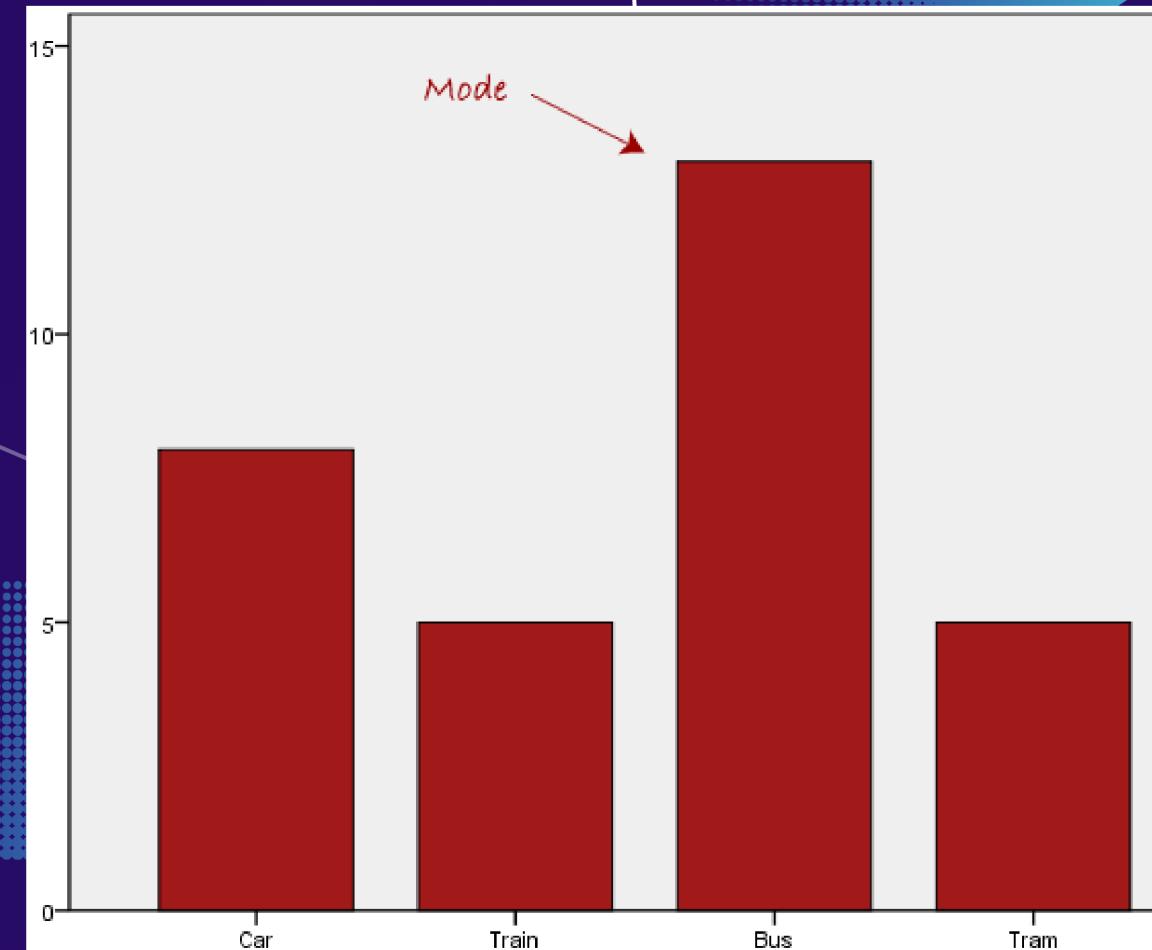
Ova mera centralne tendencije ima visoku upotrebnu vrednost kod skupova podataka koji imaju Gausovu (normalnu) raspodelu podataka. Ukoliko je raspodela podataka asimetrična tj. iskošena neophodno je koristiti neki drugi pristup zato što tada srednja vrednost ne prikazuje realnu vrednost oko koje se podaci gomilaju.

Formula izračunavanja:

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

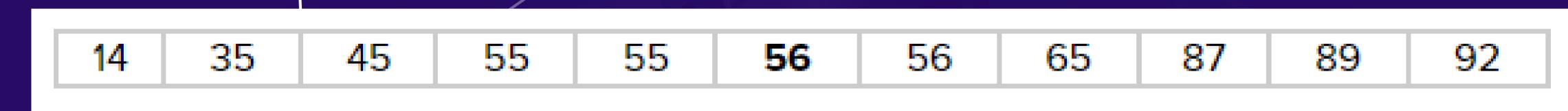
Modus

Modus vrednost predstavlja vrednost uzorka koji ima najveću vrednost učestalosti u skupu podataka. Ukoliko postoji više uzoraka koji imaju identičnu vrednost frekvencije u skupu podataka, svi uzorci se uzimaju kao validne modus vrednosti.



Medijana

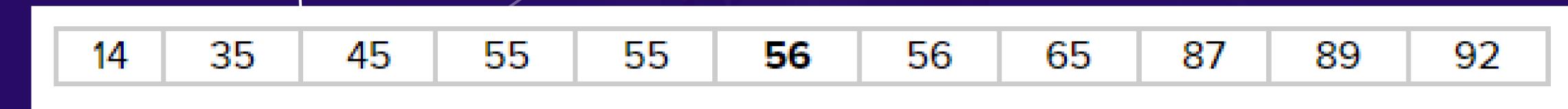
Medijana predstavlja meru centralne tendencije koja sa višim stepenom verodostojnosti određuje srednju vrednost u odnosu na aritmetičku srednju vrednost. U procesu izračunavanja medijne vrednosti se eliminišu vrednosti koje mogu negativno da utiču na realno stanje raspodele podataka.



$$\text{Median} = \frac{56 + 55}{2} = 55.5$$

Medijana

Medijana predstavlja meru centralne tendencije koja sa višim stepenom verodostojnosti određuje srednju vrednost u odnosu na aritmetičku srednju vrednost. U procesu izračunavanja medijne vrednosti se eliminišu vrednosti koje mogu negativno da utiču na realno stanje raspodele podataka.



$$\text{Median} = \frac{56 + 55}{2} = 55.5$$

Vrste raspodele podataka



U zavisnosti od tipova podataka koji se obrađuju, raspodelu podataka je moguće podeliti u dve reprezentativne vrste:

Diskretna raspodela:

- Bernulijeva
- Binomialna
- Poasonova

Kontinualna raspodela:

- Normalna/Gausova
- "Log-Normal"

Diskretna raspodela

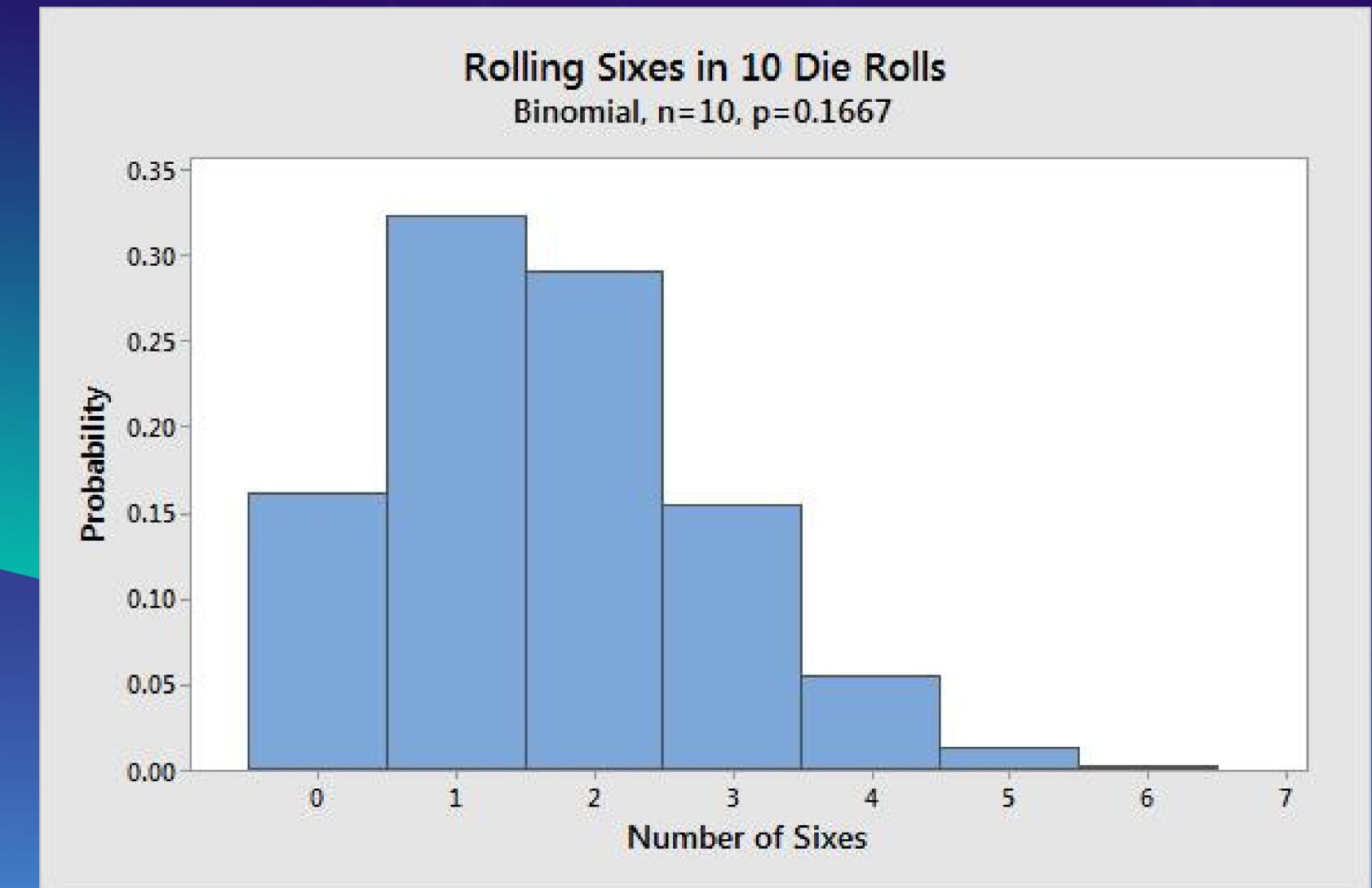
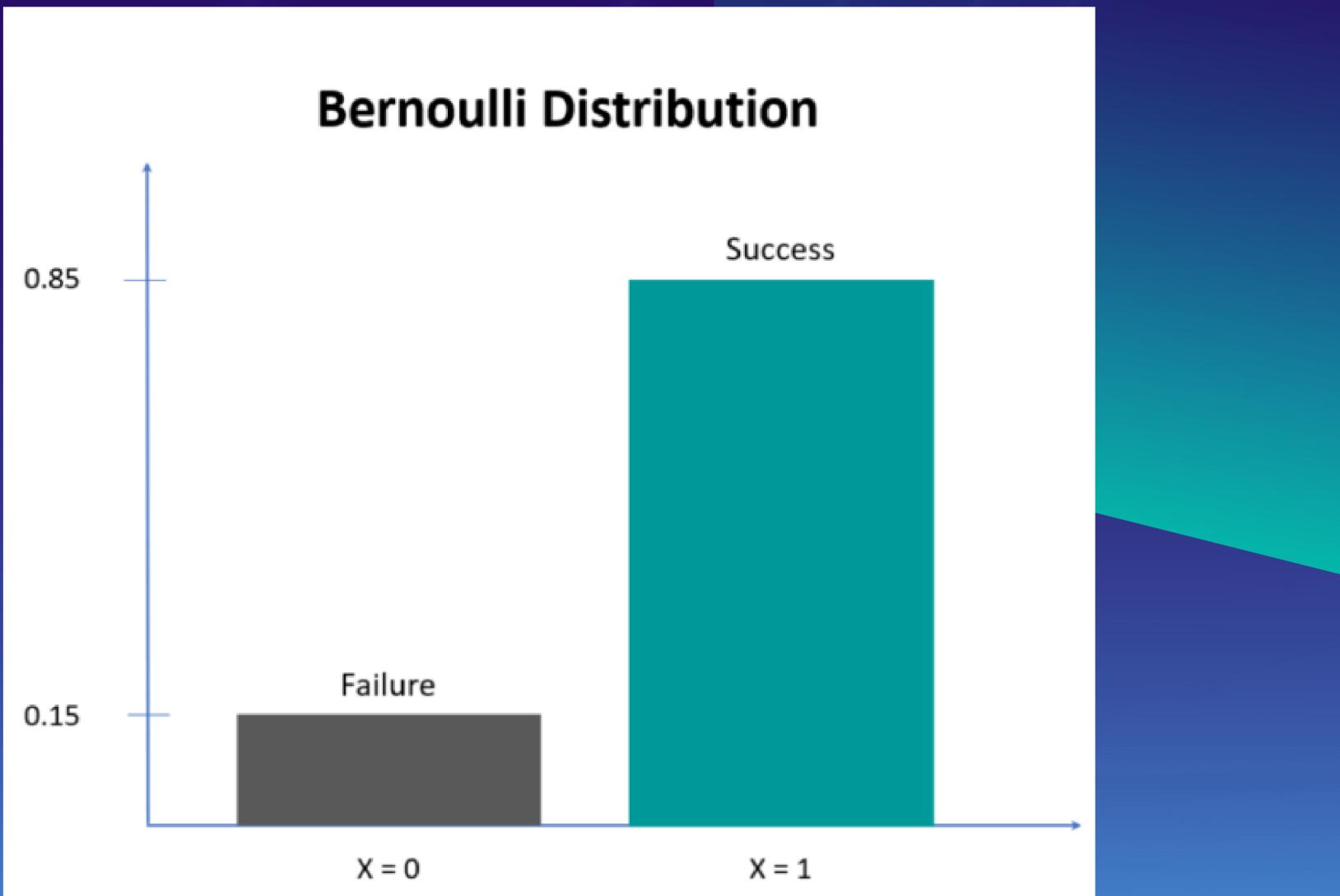


Bernulijeva

$$P(X = x) = p^x(1 - p)^{1-x}; P(x) = \begin{cases} 1 - p, & \text{for } x = 0 \\ p, & \text{for } x = 1 \end{cases}$$

Binomialna

$$P(X = x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{(n-x)}$$



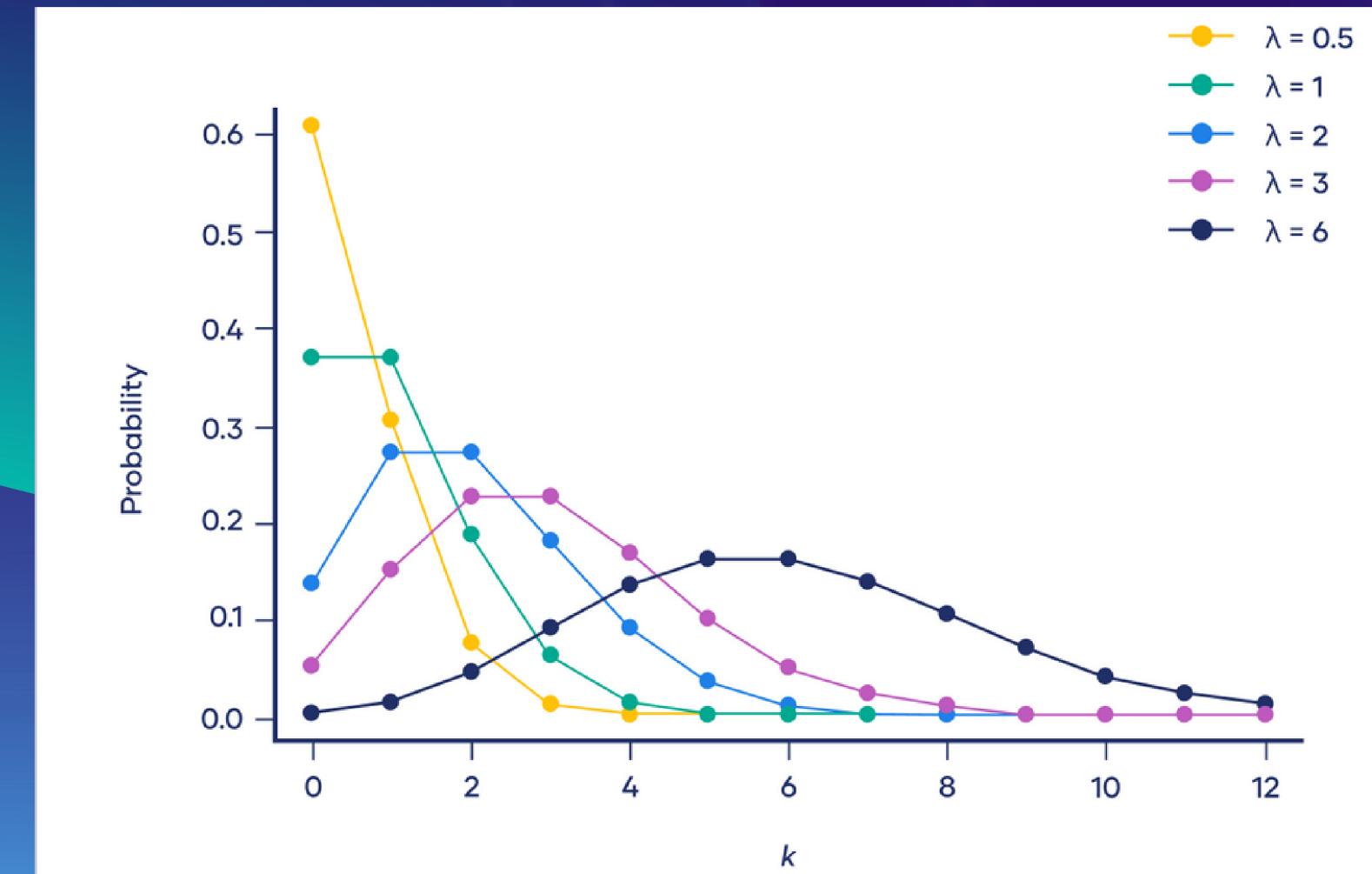
Diskretna raspodela



Poasonova raspodela

Poasonova raspodela predstavlja diskretnu raspodelu podataka koja se bavi verovatnoćom da li će se ili neće dogoditi određeni događaj u određenom intervalu koji se posmatra.

$$f(x) = \frac{\lambda^x}{x!} e^{-\lambda}$$

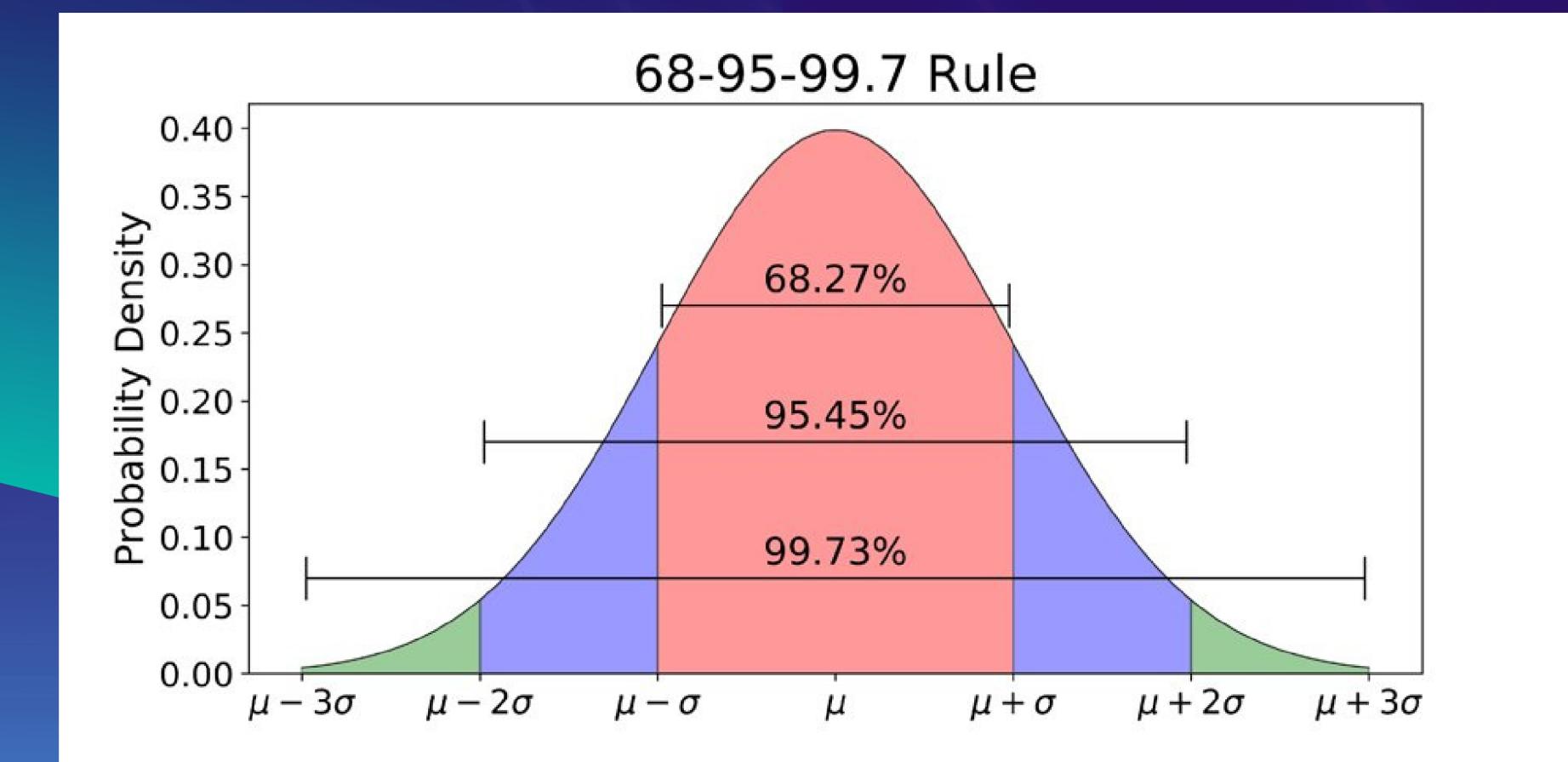
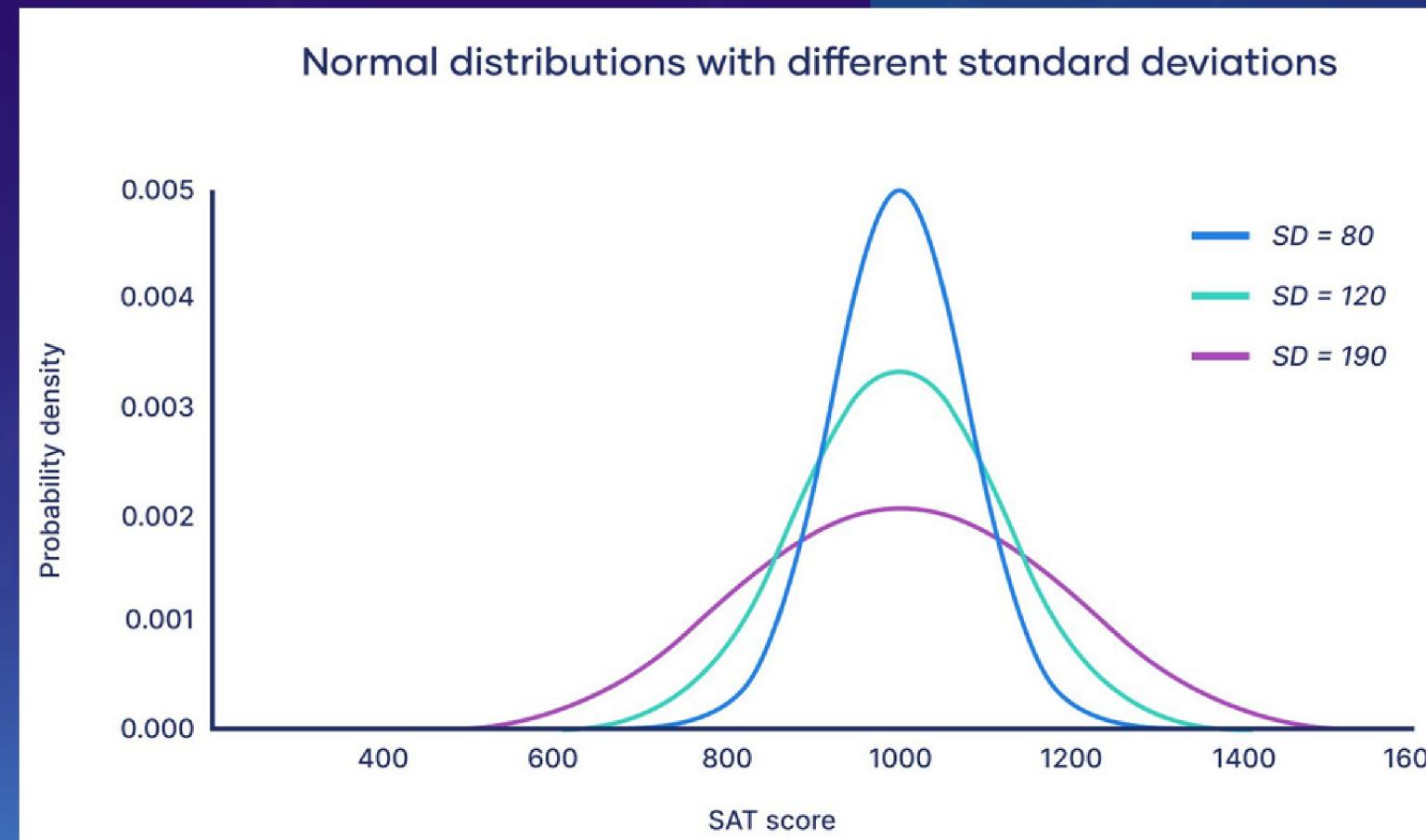


Kontinualna raspodela



Normalna/Gausova raspodela

Glavna karakteristika podataka koji su predstavljeni ovom raspodelom jeste da su mere centralne tendencije srednja vrednost, medijana i modus jednake.



Kontinualna raspodela

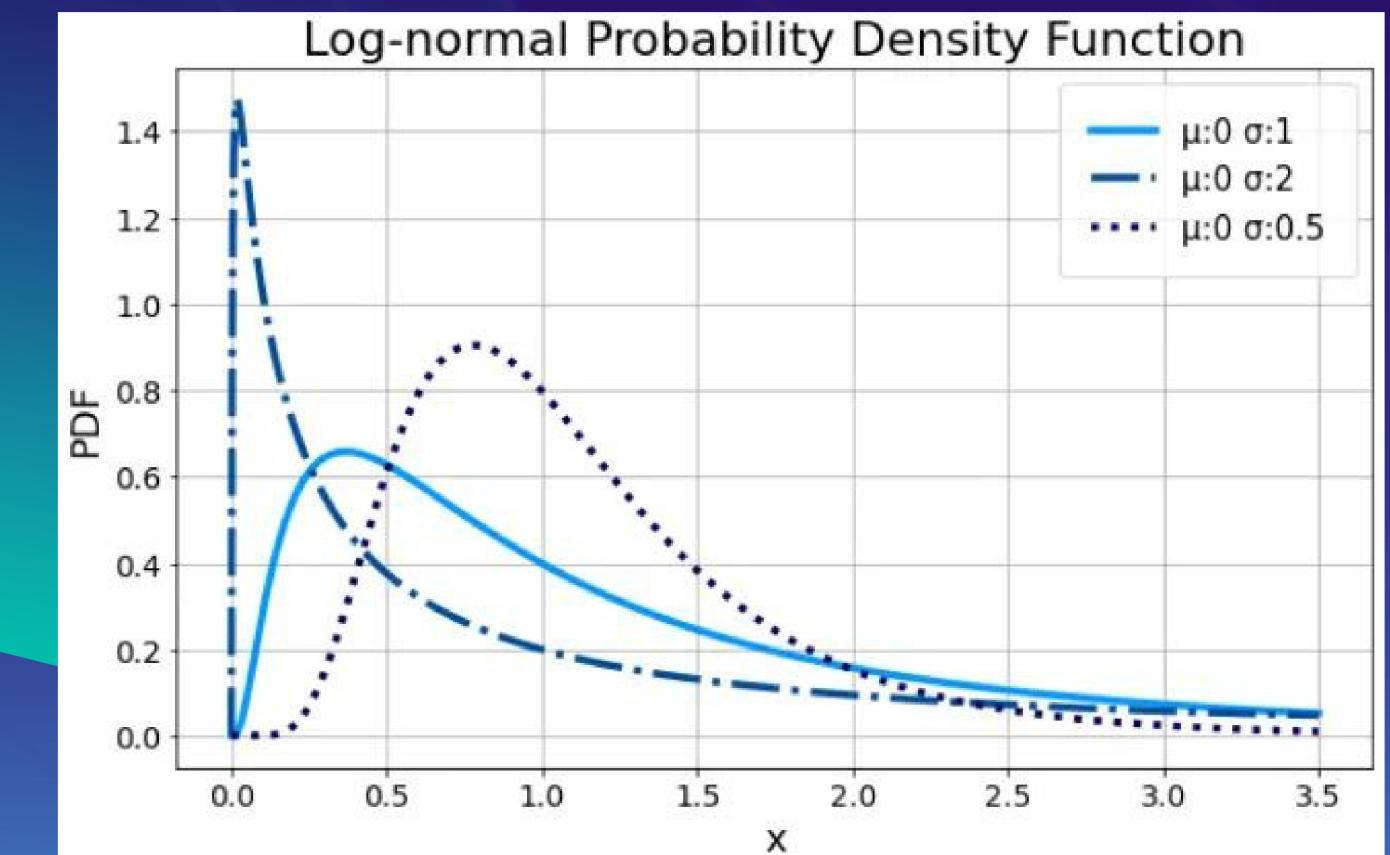


"Log-Normal" raspodela

"Log-Normal" raspodela podataka predstavlja desno iskrivljenu raspodelu podataka. Vrednost funkcije počinje iz nulte vrednosti i strmo raste do maksimalne vrednosti funkcije, nakon čega vrednost funkcije postepeno opada.

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\ln(x)-\mu}{\sigma}\right)^2}$$

$$\hat{\mu} = \frac{\sum_k \ln x_k}{n} \text{ and } \hat{\sigma}^2 = \frac{\sum_k (\ln x_k - \hat{\mu})^2}{n}$$



μ - lokacioni parametar, σ – parametar skaliranja distribucije

Korelacija

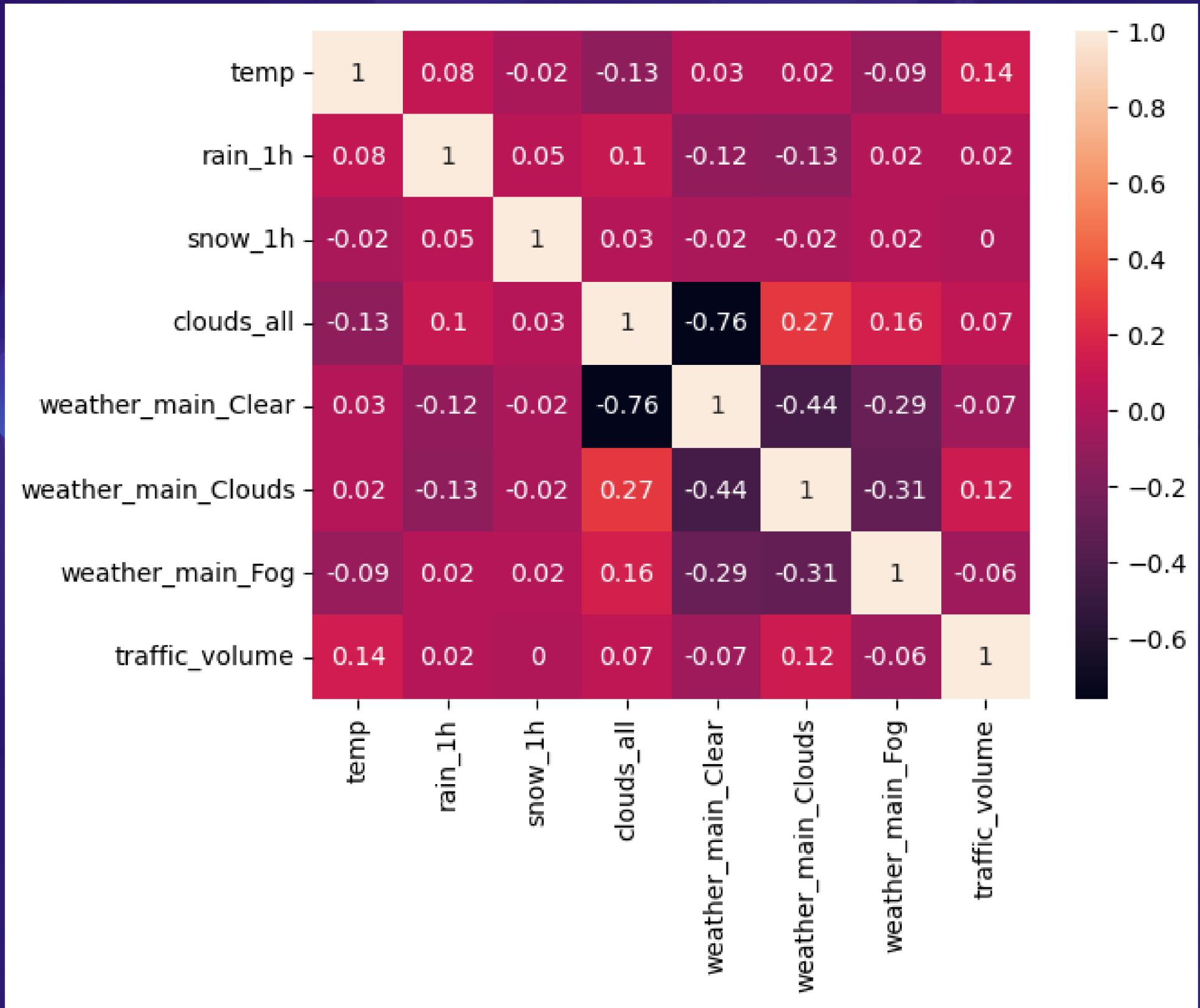
Korelacija predstavlja numeričku vrednost linearne zavisnosti između pojedinačnih fičera unutar posmatranog skupa podataka.

Vrednost korelacije između dva fičera se računa kao količnik vrednosti kovarijanse između dva fičera i proizvoda pojedinačnih vrednosti standardnih devijacija.

$$cov(x, y) = \frac{1}{n} \times \sum_{i=1}^n (x_i - mean(x)) \times (y_i - mean(y))$$

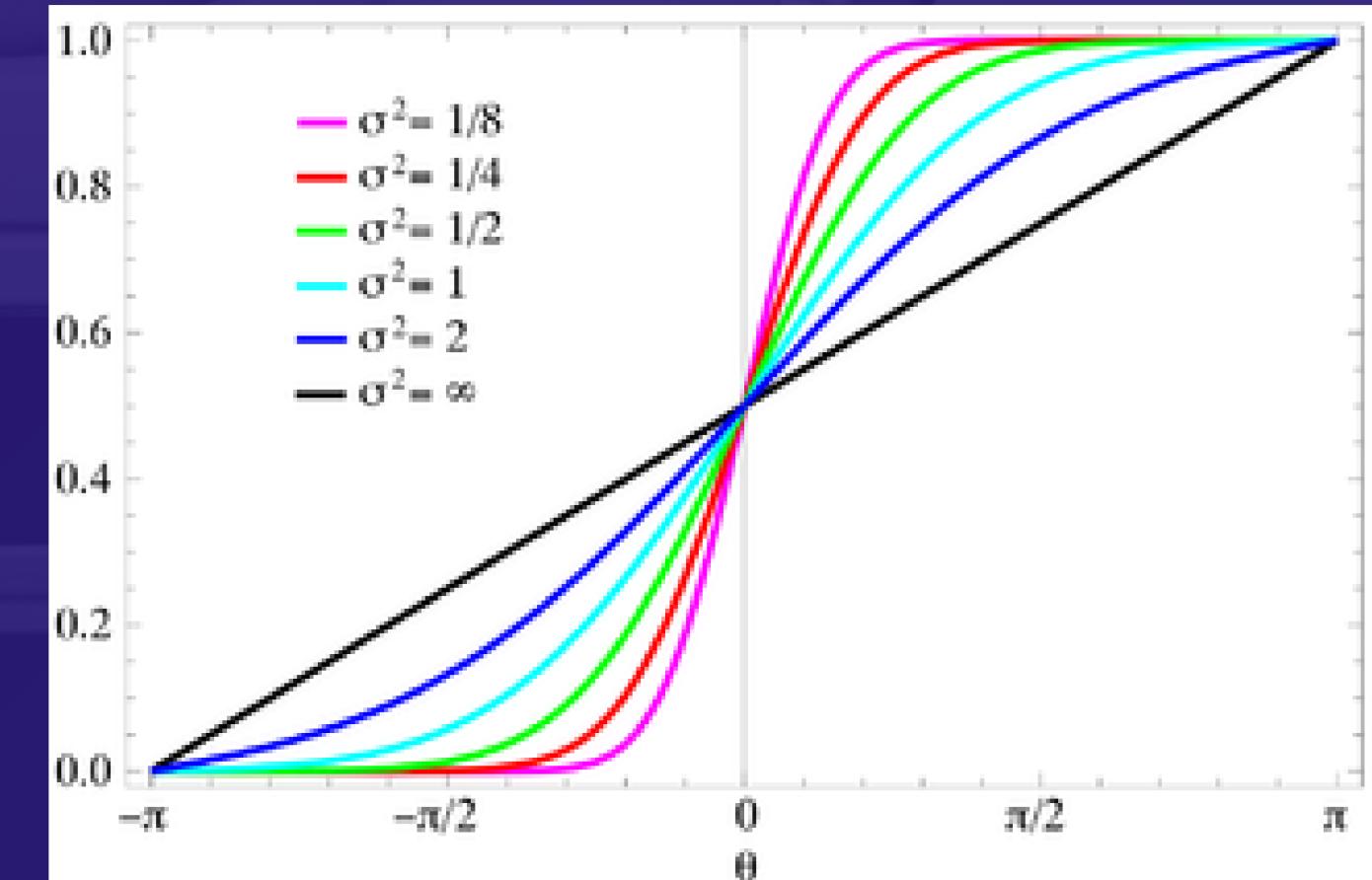
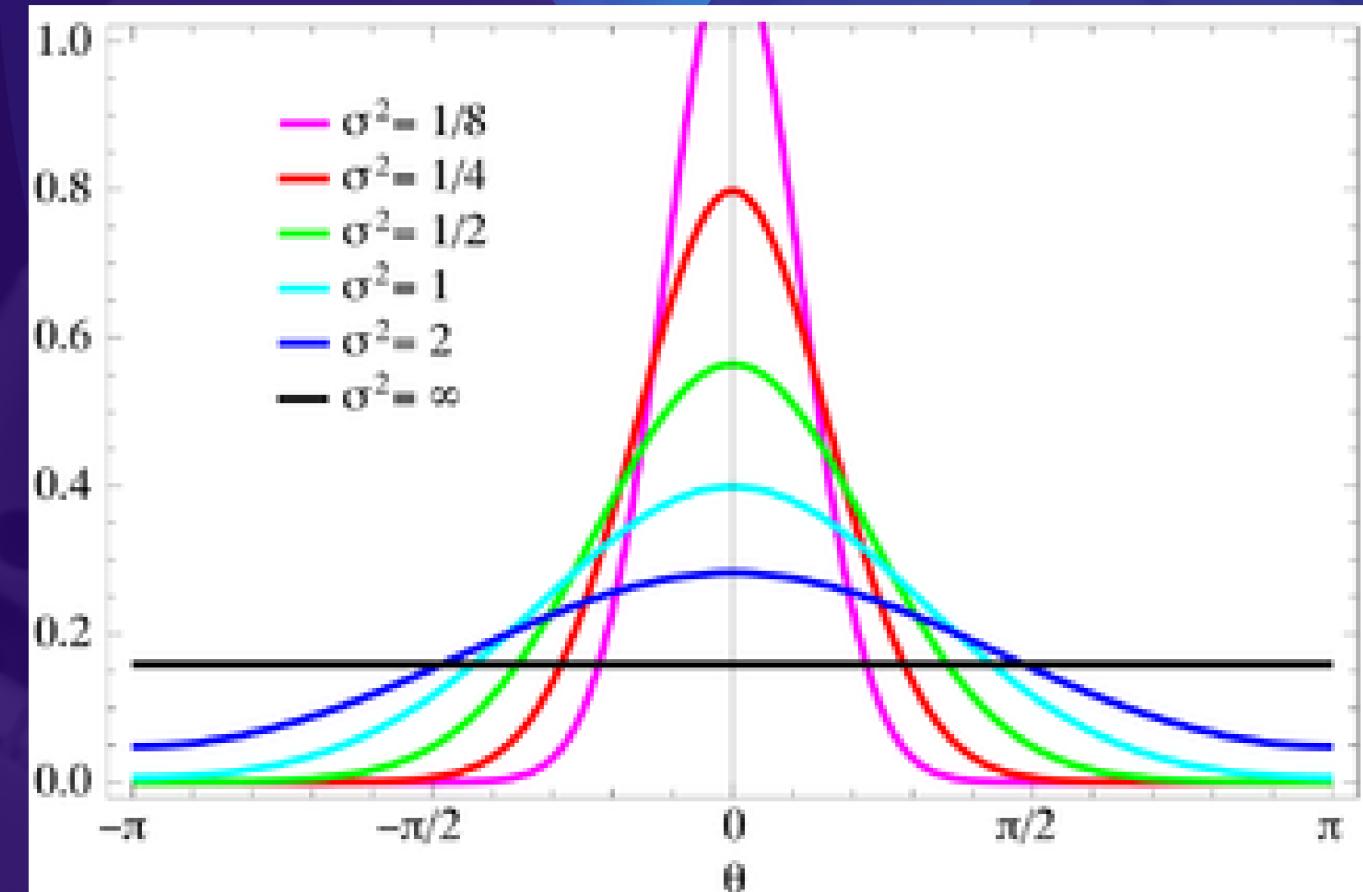
$$\text{Pearson's correlation coefficient} = \frac{cov(x, y)}{stdev(x) \times stdev(y)}$$

Korelacija



Varijansa

Varijansa predstavlja mjeru koja pokazuje koliko prosečna vrednost uzorka odstupa ili varira od mere centralne tendencije – srednje vrednosti. Ova mera detaljnije opisuje način širenja vrednosti u celokupnom skupu podataka. Varijansa može imati nisku ili visoku vrednost što utiče na to koliko će podaci biti grupisani oko srednje vrednosti.





PRAKTIČNI DEO SEMINARSKOG RADA



**HVALA
NA
PAŽNJI**



ficax.al@elfak.rs

