Statistics-2

Agenda:

1. Percentiles and Quartiles
2. 5 number summary
3. Box Plot
4. Covariance & correlation
5. Probability Distribution Function
6. Different types of distribution.

1 Percentiles & Quartiles

Percentage: 1, 2, 3, 4, 5, 6

% of numbers that are odd $= \frac{3}{6} = \frac{\text{No. of odd numbers}}{\text{Total no. of numbers}}$

$= \frac{1}{2} = 50\%$

Percentiles:

A Percentile is a value below which a certain Percentage of data points lie.

n = total no. of values.

$x = \{2, 3, 3, 4, 6, 6, 6, 7, 8, 8, 9, 9, 10, 11, 12\}$

$$\text{Percentile Rank of } 10 = \frac{\#\text{ of values below } 10}{n} \times 100$$

$$= \frac{\cancel{12}^{4}}{\cancel{15}_{\cancel{5}\,1}} \times \cancel{100}^{20} = 80 \text{ Percentile.}$$

80 Percentile = 80% of the distribution fall below the value 10.

(*) What value exists at 25 Percentile?

$$\boxed{\text{Value} = \frac{\text{Percentile}}{100} \times (n+1)}$$

$$= \frac{\cancel{25}^{1}}{\cancel{100}_{4}} \times \cancel{16}^{4}$$

$$\boxed{= 4^{th} \text{ element}}$$

$X = \{ 2, 3, 3, 4, 6, 6, 6, 7, 8, 8, 9, 9, 10, 11, 12 \}$

* If we get value as decimal like 4.5 we can take average of 4th and 5th value.

Quartiles

$Q_1 \rightarrow 25$ Percentile

$Q_2 \rightarrow$ median $\rightarrow 50$ Percentile

$Q_3 \rightarrow 75$ Percentile.

② 5 Number summary

i, minimum

ii, First quartile (25 Percentile) (Q_1)

iii, Median (Q_2)

iv, Third quartile (75 percentile) (Q_3)

v, maximum

Remove the outliers

$X = \{ 1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 29 \}$

[lower Fence ⟷ upper Fence]

Lower Fence $= Q_1 - 1.5(IQR)$

Inter Quartile Range $= \boxed{Q_3 - Q_1}$

Higher Fence $= Q_3 + 1.5(IQR)$

$X = \{1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 29\}$

$Q_1 = 25 \text{ Percentile} = \frac{25}{100} \times (20) = 5^{th} \text{ value} = 3$

$Q_3 = 75 \text{ Percentile} = \frac{75}{100} \times 20 = 15^{th} \text{ value} = 7$

$IQR = Q_3 - Q_1$

$= 7 - 3$

$= 4$

Lower Fence $= Q_1 - 1.5(IQR)$

$= 3 - 1.5(4)$

$= 3 - 6$

$= -3$

Higher Fence $= Q_3 + 1.5(IQR)$

$= 7 + 1.5(4)$

$= 7 + 6$

$= 13$

$[-3, 13]$

Hence, we can consider 29 as an outlier for the above data.

③ Box Plot :- [To visualize outliers]

i, minimum value = 1

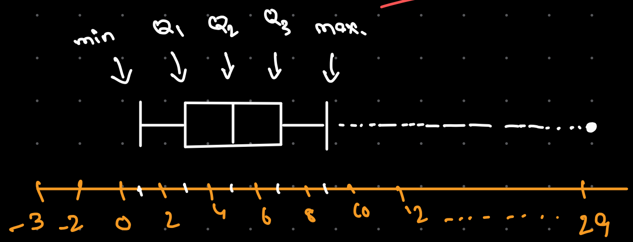
ii, $Q_1 = 3$

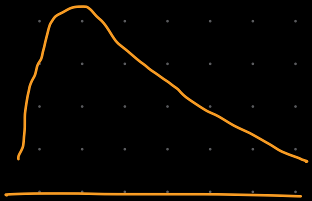
iii, median $= Q_2 = 5$

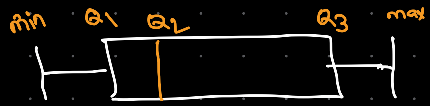
iv, $Q_3 = 7$

v, maximum value = 9

Box Plot

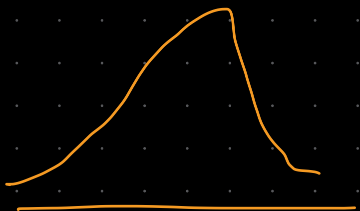






$Q_3 - Q_2 > Q_2 - Q_1$

mean > median > mode



$Q_3 - Q_2 < Q_2 - Q_1$

mode > median > mean

Assignment:

$y = \{-13, -12, -6, -5, 3, 4, 5, 6, 7, 7, 8, 10, 10, 11, 24, 55\}$

$Q_1\,(25) =$

$= \frac{25}{100} \times (16+1)$

$= \frac{\cancel{25}^{1}}{\cancel{100}_{4}} \times 17$

$= 4.25$ element

$\boxed{Q_1 = -1}$

$Q_3 = \frac{\cancel{75}^{3}}{\cancel{100}_{4}} \times (17)$

$= 12.75$ element

$\boxed{Q_3 = 10}$

$Q_2\,(50)$

$\frac{6+7}{2} = \frac{13}{2}$

$= 6.5$

$\boxed{Q_2 = 6.5}$

$IQR = Q_3 - Q_1$

$= 10 - (-1) = 11$

lower Fence

$= Q_1 - 1.5(IQR)$

$= -1 - 1.5(11)$

$= -1 - 16.5$

Higher Fence =

$Q_3 + 1.5(IQR)$

$= 10 + 16.5$

$\boxed{26.5}$

$= \boxed{-17.5}$

$[-17.5,\ 26.5]$

Boxplot

① min = −13 ② $Q_1 = -1$ ③ $Q_2 = 6.5$

④ $Q_3 = 10$ ⑤ max = 24

$\boxed{\text{Outlier} = 55}$

② $Z = \{1, 2, 4, 6, 7, 12, 18, 34, 77, 66, 108, 99, 14\}$

Sorted $= \{1, 2, 4, 6, 7, 12, 14, 18, 34, 66, 77, 99, 108\}$

$Q_1 = \frac{25}{100} \times 14 = 3.5$ element

$\boxed{Q_1 = 5}$

$\boxed{Q_2 = 14}$

$Q_3 = \frac{\cancel{75}^{3}}{\cancel{100}_{4\,2}} \times \cancel{14}^{7} = 10.5$ element

$IQR = Q_3 - Q_1$

$\boxed{= 66.5}$

$\boxed{Q_3 = 71.5}$

lower fence $= Q_1 - 1.5(IQR)$

$= 5 - 1.5(66.5)$

$= -94.75$

High fence $= Q_3 + 1.5(IQR)$

$= 71.5 + 99.75$

$= 171.25$

$[-94.75,\ 171.25]$

③ Covariance and Correlation

[Relationship between X and Y]

X	Y
2	3
4	5
6	7
8	9

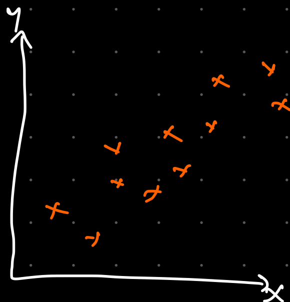
$X \uparrow \quad Y \uparrow$ ✓

$X \uparrow \quad Y \downarrow$

Size of house → Price

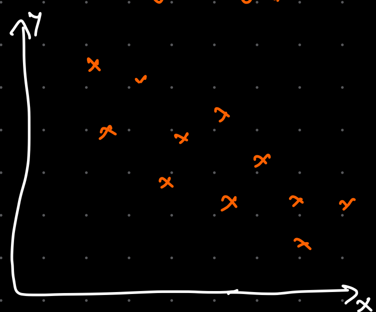
$X \downarrow \quad Y \uparrow$

$X \downarrow \quad Y \downarrow$ ✓



$x \uparrow \quad y \uparrow$

$x \downarrow \quad y \downarrow$



$x \uparrow \quad y \downarrow$

$x \downarrow \quad y \uparrow$

Covariance

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

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

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

$$\Downarrow$$

$$\boxed{\text{var}(x) \Leftarrow \text{cov}(x,x)}$$

cov (x, y)

$\boxed{\begin{matrix} x\uparrow & y\uparrow \\ x\downarrow & y\downarrow \end{matrix}}$ → +ve covariance

$\boxed{\begin{matrix} x\downarrow & y\uparrow \\ x\uparrow & y\downarrow \end{matrix}}$ → −ve covariance

Ex:-

x	y
2	3
4	5
6	7
$\bar{x} = 4$	$\bar{y} = 5$

X and Y are having a positive covariance

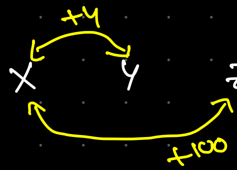
$$\text{cov}(x,y) = \sum_{i=1}^{n} \frac{(x_i-\bar{x})(y_i-\bar{y})}{n-1}$$

$$= \frac{[(2-4)(3-5) + (4-4)(5-5) + (6-4)(7-5)]}{2}$$

$$= \frac{4+0+4}{2} = 4 \quad \text{+ve covariance}$$

Advantages

① Relationship between X and Y



Disadvantages

1) Covariance does not have a specific limit value

$-\infty \longleftrightarrow +\infty$

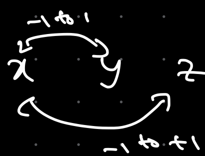
(It will not gives the strength of the relation between variables)
(It only gives the direction)

* Pearson Correlation coefficient [-1 to +1]

$$\rho_{x,y} = \frac{cov(x,y)}{\sigma_x \cdot \sigma_y}$$

(row of x and y) ← $\rho_{x,y}$

$\sigma_x \cdot \sigma_y$ → (std deviation of x and y)



* The more the value towards +1, the more +ve correlated it is.
* The more the value towards -1, the more -ve correlated it is.

X Y 0.4

X Z 0.7 → strong

Disadv: only linear relationship will be captured by Pearsons correlation coefficient.

* Spearman Rank correlation

$$r_s = \frac{cov(R(x), R(y))}{\sigma_{R(x)} * \sigma_{R(y)}}$$

X	Y	R(x)	R(y)
5	6	3	1
7	4	2	2
8	3	1	3
1	1	5	4
2	2	4	5

Feature selection

+ve	+ve	+ve	no impact	-ve	O/P :
size of house	no. of rooms	Location	No. of People staying	is Haunted	Price.

---

* Probability Distribution Function
* Probability Density Function
* Probability Mass Function.

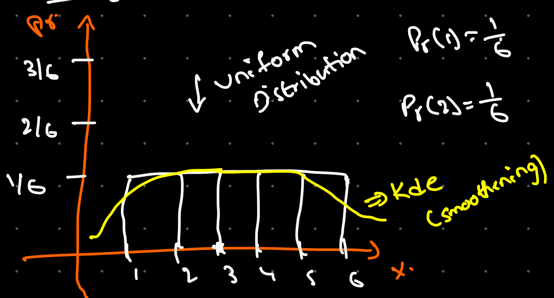
→ **Probability Distribution Function**

i) Probability Density Fn
ii) Probability Mass Fn
iii) Cumulative Distributive Fn.

① **Probability Mass Function (PMF)**

* when dataset has Discrete Random Variable

Ex: Rolling a Dice {1,2,3,4,5,6}



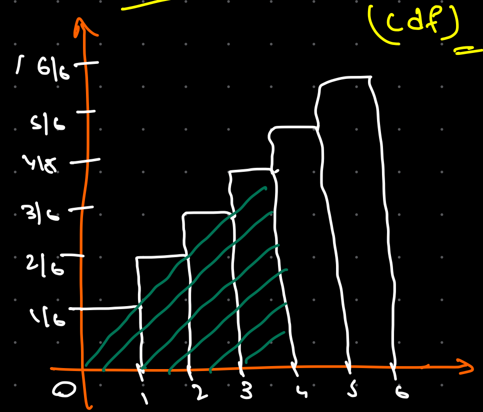
$Pr(1) = \frac{1}{6}$

$Pr(2) = \frac{1}{6}$

$Pr(1 \text{ or } 2) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6}$

$= \frac{1}{3}$

Cumulative Distribution Function (cdf)



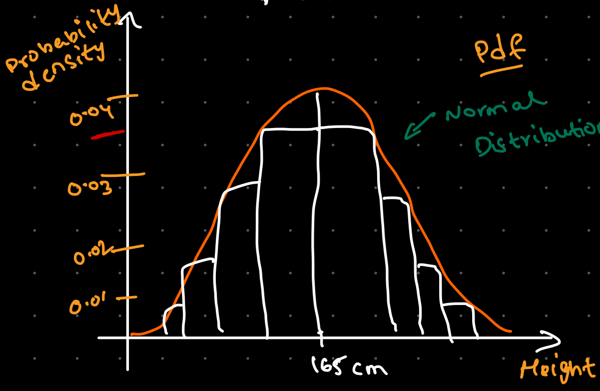
$Pr(X \le 4) = Pr(X=1) + Pr(X=2)$

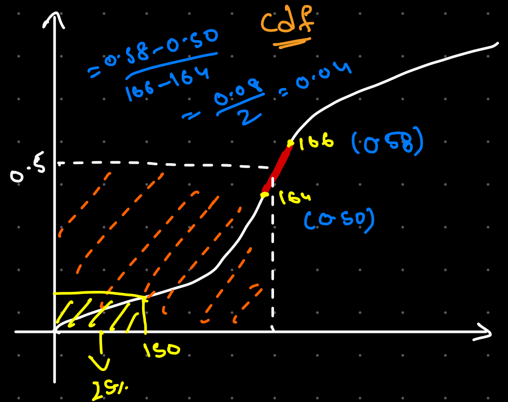
$+ Pr(X=3) + Pr(X=4)$

$= 4/6.$

② Probability Density Function

* Distribution of continous Random Variable.

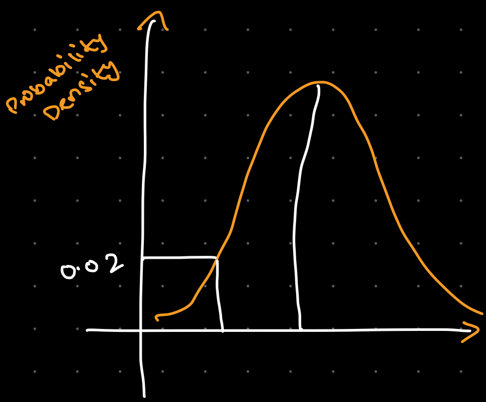


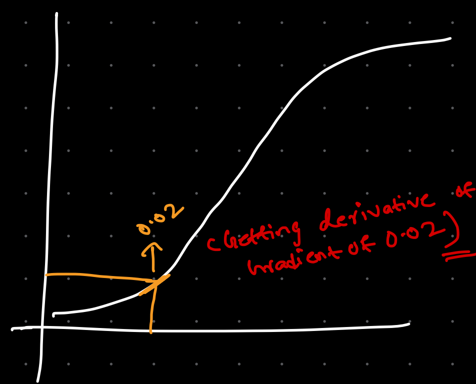


* During the reduce in the size of histogram bins, the slope of the curve gets reduced

* For every probability density that we are calculating in the PDF is coming from the gradient of the CDF. *

* Probability Density ⇒ Gradient of cumulative curve.



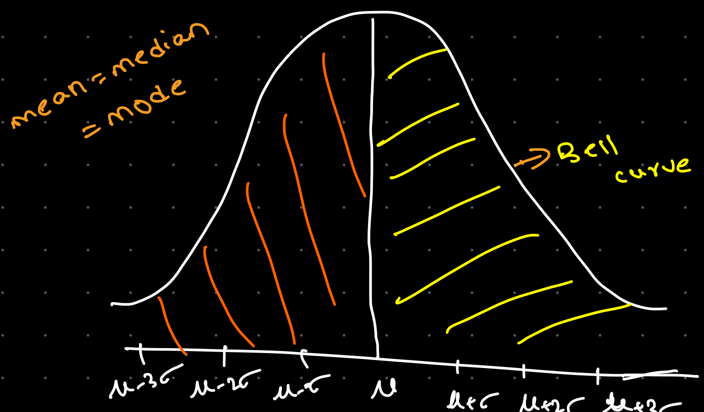


Different types of Distribution

1. Normal Distribution → PDF
2. Standard Normal Distribution → PDF
3. log Normal Distribution → PDF
4. Power Law Distribution → PDF
5. Bernoulli Distribution → PMF
6. Binomial Distribution → PMF
7. Poisson Distribution → PMF
8. Uniform Distribution
   - Discrete → PDF
   - continous → PMF
9. Exponential Distribution → PDF
10. CHI SQUARE Distribution → PDF
11. F- Distribution → PDF

1. Normal / Gaussian Distribution

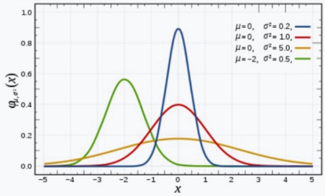
X : continous Random variable



Ex: Height, weight, age .. (IRIS Dataset)

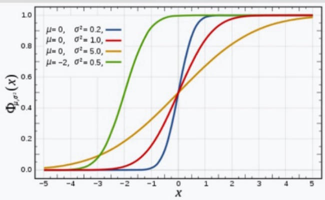
Normal distribution

Probability density function



The red curve is the *standard normal distribution*

Cumulative distribution function



Notation	$\mathcal{N}(\mu, \sigma^2)$

$$X \approx N(\mu, \sigma^2)$$

Support Parameters μ = mean

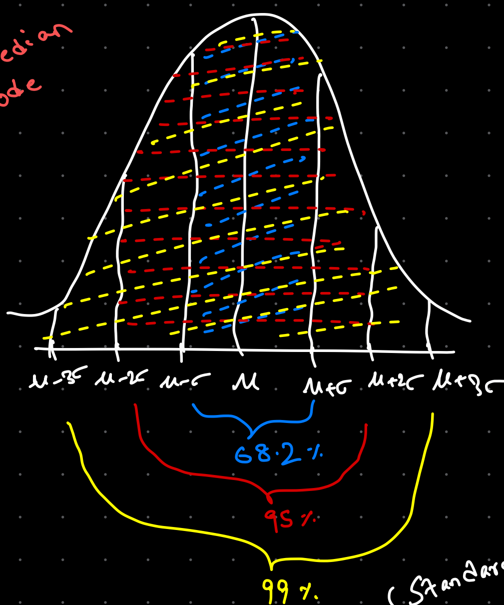
σ^2 = variance

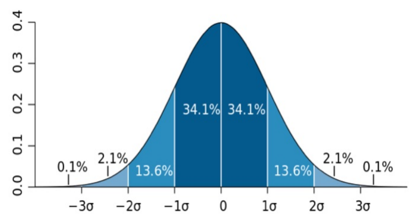
$$\text{PDF} = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Empirical Rule (68-95-99.7% Rule)

(3-Sigma Rule)

mean = median = mode





For the normal distribution, the values less than one standard deviation away from the mean account for 68.27% of the set; while two standard deviations from the mean account for 95.45%; and three standard deviations account for 99.73%.

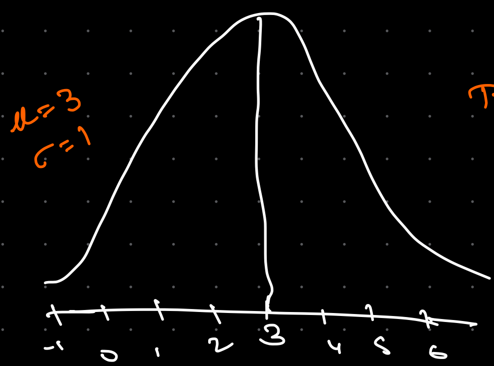
(Standard Deviation and coverage)

② Standard Normal Distribution

The simplest case of normal distribution is known as standard normal distribution or unit normal distribution

This is special case when $\mu = 0$ and $\sigma = 1$

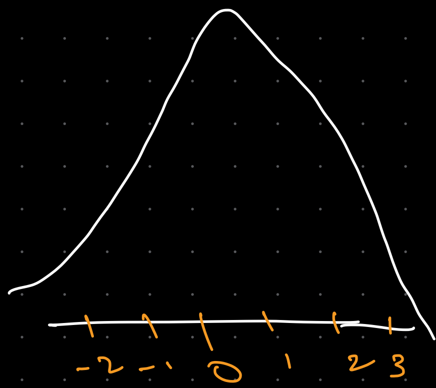
X $\quad \mu = 3 \quad \sigma = 1$



Transformation

$\Rightarrow$

$\mu = 0, \ \sigma = 1$



(standard normal distribution)

$\downarrow$

$$Z\text{-score} = \frac{x_i - \mu}{\sigma}$$

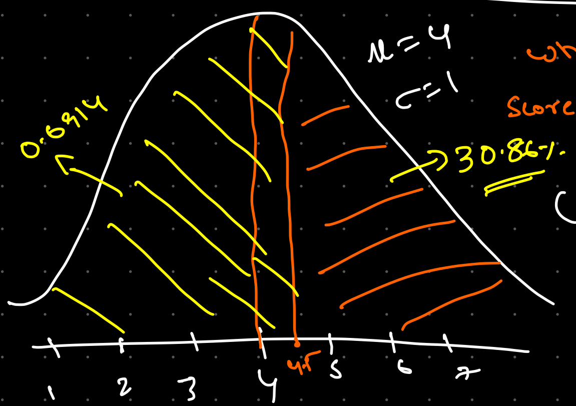
$$\rightarrow \frac{1-3}{1} = -2$$

$$\frac{3-3}{1} = 0 \qquad \frac{2-3}{1} = -1$$

$$\boxed{4.5} \rightarrow \frac{4.5-3}{1} = 1.5$$

Z-score tells you about a value how many standard deviation away from mean

X



$\mu = 4$

$\sigma = 1$

What is the percentage of scores lies above 4.5?

(we need to use Z-table to calculate the area).

$$z\text{-score} = \frac{4.5-4}{1}$$

$$= 0.5$$

Area under curve $= 1 - 0.6914$

$= 0.3086 \Rightarrow 30.86\%$