

Agenda

- ① Percentiles and Quartiles ✓
- ② 5 Number Summary {Outliers} ✓
- ③ Box plot ✓
- ④ Covariance And Correlation }.
- ⑤ Probability distribution function
- ⑥ Different types of distribution

① Percentiles And Quartiles [GATE, CAT]

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

$$\% \text{ of numbers that are odd} = \frac{3}{6} = \frac{\text{No. of odd numbers}}{\text{No. of total no.}} \\ = \frac{1}{2} = 50\%$$

Percentiles :

Defn : A percentile is a value below which a certain percentage of data points lie.

$n = 15$

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

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

$$= \frac{7 + 8}{15} \times 100 = 80 \text{ percentile}$$

80 percentile = 80% of the distribution fall below the value of 10 //

② What value exists at 25 percentile?

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

$$= \frac{18}{100} * 164 = \boxed{4} \text{ Element } = 4$$

20%
Q1

$$X: \{ \begin{matrix} & \downarrow & \downarrow & \downarrow & \downarrow \\ 2, 3, 3, 4, & 6, 6, 6, 7, 8, 8, 9, 9, & \boxed{10}, & 11, 12 \end{matrix} \}$$

$\uparrow \quad \uparrow \quad \uparrow \quad \uparrow$

$$\frac{4+6}{2} = \underline{\underline{5}}$$

4.5

Quartiles

① Q1 → 25 percentile

Q2 → Median → 50 percentile

Q3 → 75 percentile

5 Number Summary

(1) Minimum

(2) First Quartile (25 percentile) (Q1)

(3) Median (Q2)

(4) Third Quartile (75 percentile) (Q3)

(5) Maximum

Remove the Outliers

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

[Lower Fence \longleftrightarrow Higher Fence]

$$\text{Lower Fence} = Q_1 - 1.5(IQR)$$

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

$$\text{Higher Fence} = Q_3 + 1.5(IQR)$$

$X = \{1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, \underline{\boxed{12, 9}}\}$

$$Q_1 = 25^{\text{percentile}} = \frac{25}{100} * 20 = 5^{\text{th value}} = 3 \quad n+1=19+1=20$$

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

$$IQR = 7 - 3 = 4$$

$$\text{Lower Fence} = Q_1 - 1.5(IQR)$$

$$= 3 - 1.5(4)$$

$$= 3 - 6$$

$$= -3$$

$$\text{Higher Fence} = Q_3 + 1.5(IQR)$$

$$= 7 + 1.5(4)$$

$$= 7 + 6$$

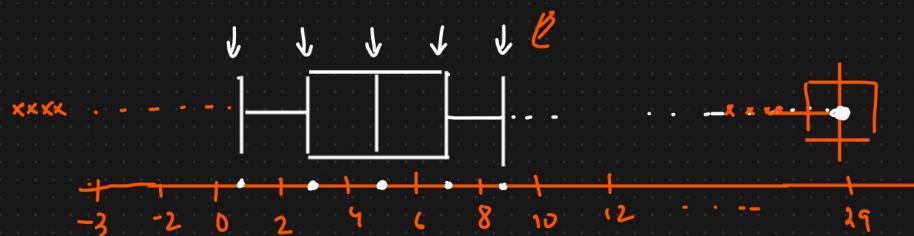
$$= 13$$

$$[-3, 13]$$

$X = \{1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, \underline{\boxed{12, 9}}\}$

Box plot [To visualize Outliers]

Box plot



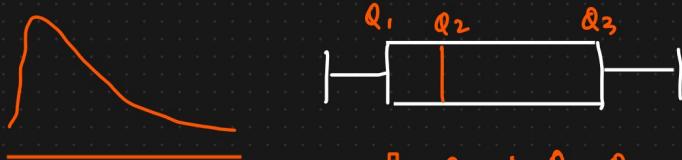
$$\textcircled{1} \text{ Minimum value} = 1$$

$$\textcircled{2} \text{ } Q_1 = 3$$

$$\textcircled{3} \text{ Median } Q_2 = 5$$

$$\textcircled{4} \text{ } Q_3 = 7$$

$$\textcircled{5} \text{ Maximum} = 9$$



mean > median > mode

$$Q_3 - Q_1 > Q_2 - Q_1$$

↳ box plot of this graph will be like this



Internal Assignment

$$-5+3 = 1 \quad \uparrow \overline{2}$$

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

[lower fence \longleftrightarrow higher fence]

$$Q_1 = \frac{21}{4} \times 17^4 = 4.25$$

$$\text{lower fence} = -1 - 1.5(10 + 1)$$

$$Q_3 = \frac{75}{100} \times 17 = 12.75$$

$$= -1 - 1.5(11)$$

$$= -17.5$$

$$[-17.5, 26.5].$$

$$\text{higher fence} = 10 + 1.5(10 + 1)$$

↓
55 is an outlier

$$= 10 + 16.5$$

$$= 26.5$$

\equiv

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

$$Q_1 = 5$$

$$Q_3 = 71.5$$

\equiv

$$[-94.75, 171.25]$$

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

+53

$$Q_3 = \frac{75}{100} \times 14 = \frac{42}{4} 2 + 10.5$$

$$Q_3 = \frac{66 + 77}{2} = 71.5$$

Covariance And Correlation

[Relationship between X and Y]

X	Y
2	3
4	5
6	7

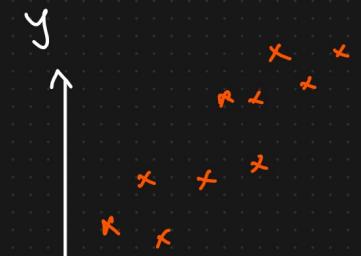
\rightarrow	X↑	Y↑
\rightarrow	X↓	Y↑
\rightarrow	X↑	Y↓

IS
Household

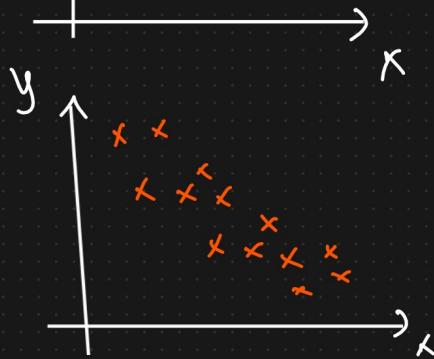
Size of
house

Price

$$8 \quad 9 \quad \rightarrow \boxed{x \downarrow \quad y \downarrow}$$



$$\boxed{\begin{matrix} x \uparrow & y \uparrow \\ x \downarrow & y \downarrow \end{matrix}}$$



$$\boxed{\begin{matrix} x \uparrow & y \downarrow \\ x \downarrow & y \uparrow \end{matrix}}$$

① Covariance

$$\text{Cov}(x, y) = \sum_{i=1}^n \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1} \quad \text{Var}(x) = \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{n-1}$$

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

$$\text{Cov}(x, y) \rightarrow \boxed{\begin{matrix} x \uparrow & y \uparrow \\ x \downarrow & y \downarrow \end{matrix}} \Rightarrow \text{pos covariance}$$

$$\boxed{\begin{matrix} x \uparrow & y \downarrow \\ x \downarrow & y \uparrow \end{matrix}} \Rightarrow \text{neg covariance.}$$

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

$$= \left[(2-4)(3-5) + (4-4)(5-5) + (6-4)(7-5) \right]$$

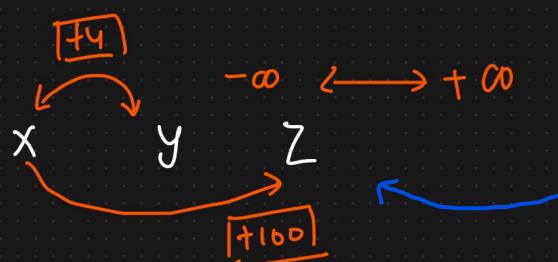
$$\begin{array}{cc} x & y \\ \rightarrow 2 & 3 \\ \rightarrow 4 & 5 \\ \underline{6} & \underline{7} \\ \bar{x} = 4 & \bar{y} = 5 \end{array}$$

$$= \frac{4+0+4}{2} = \frac{8}{2} = 4 \quad \text{pos covariance}$$

X and Y are having a positive covariance

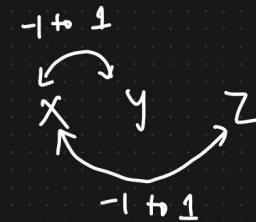
Advantages

- ① Relationship between X & Y



- ① Covariance does not have a specific limit value

i.e. If the covariance value is higher that doesn't mean highly correlated.



- ② Pearson Correlation Coefficient

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

$[-1 \text{ to } 1]$

- ④ The more the value towards +1 the more the correlated it is
 ② The more the " " " " -1 " " " " -ve "

Pearson's Correlation: Measures the linear relationship between two continuous variables. It assumes that both variables are normally distributed and that the relationship between them is linear.

$$X \quad Y \quad 0.6$$

Spearman's Rank Correlation: Measures the monotonic relationship between two variables. It can be used with ordinal data or when the assumptions of Pearson's correlation (like normality and linearity) are not met. It works by ranking the data points and then applying Pearson's correlation formula to the ranks.

$$X \quad Z \quad 0.7$$

If still having doubt, search on internet

- ① Spearman Rank Correlation

$$\rho_s = \frac{\text{Cov}(R(x), R(y))}{\sqrt{R(x) \cdot R(y)}}$$

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

Feature Selection

+ve Size of house	+ve No. of rooms	+ve location	≈ 0 No. of people Staying	-ve bounced	$\frac{0/p}{P_{\text{Price}}} \uparrow$
----------------------	---------------------	-----------------	---	----------------	---

Probability Distribution Function And Probability Density Function

Probability Mass function

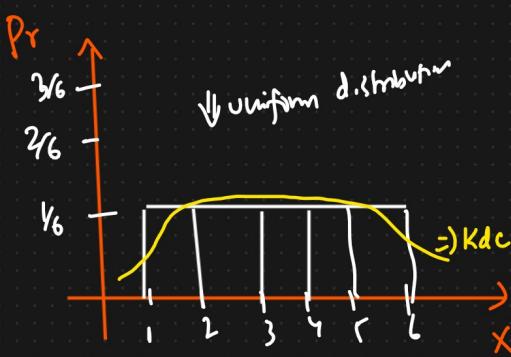
Probability Distribution Function

- ① Probability density fn
 - ② Probability mass fn ✓
 - ③ Cumulative distributn fn.
- $\left. \right\} \text{Important}$

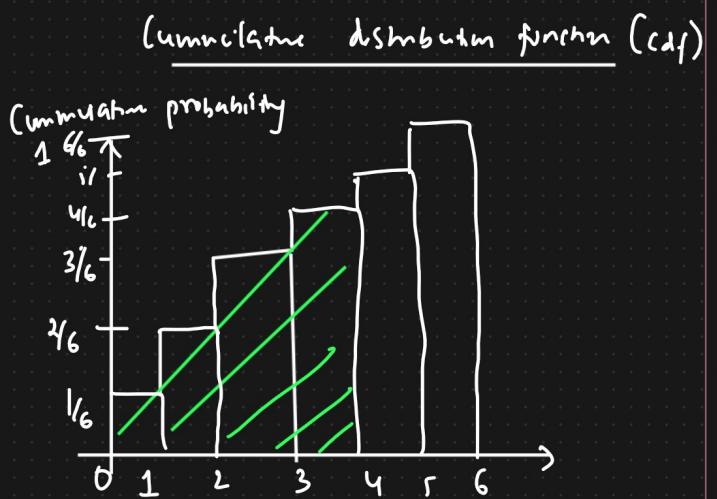
① PMF ↴

① Discrete Random Variable ⇩

Eg: Rolling a dice {1, 2, 3, 4, 5, 6}



$$\begin{aligned} \Pr(1) &= \frac{1}{6} \\ \Pr(2) &= \frac{1}{6} \end{aligned}$$



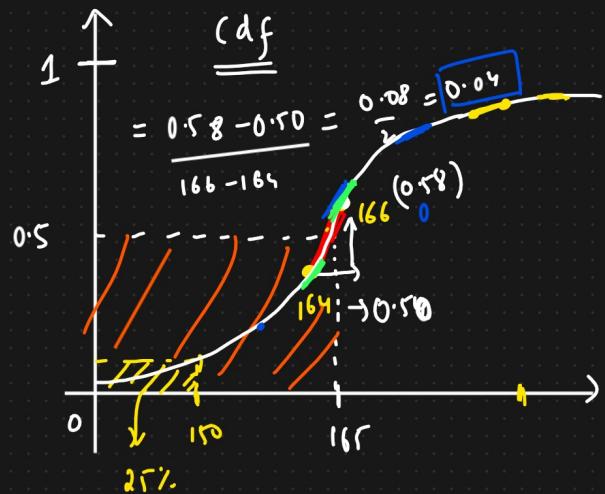
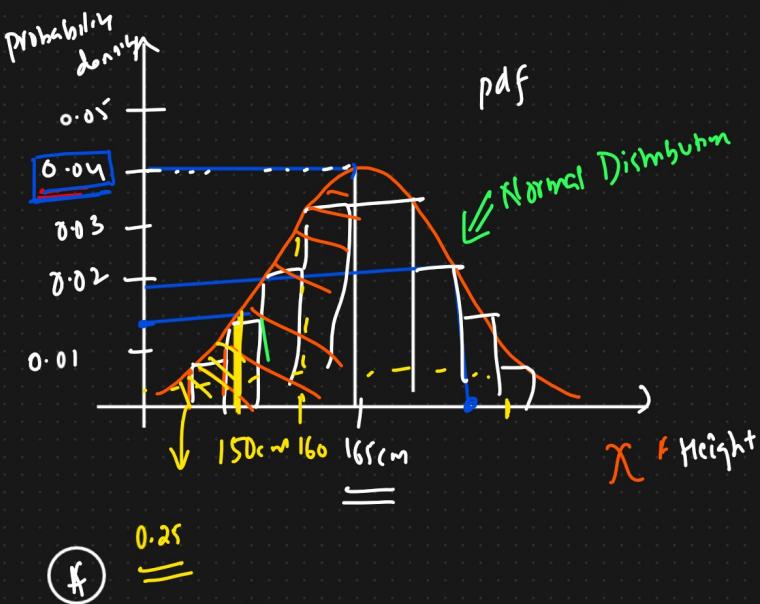
$$\Pr(1 \text{ or } 2) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}$$

$$\begin{aligned} \Pr(X \leq 4) &= \Pr(X=1) + \Pr(X=2) + \Pr(X=3) \\ &\quad + \Pr(X=4) \end{aligned}$$

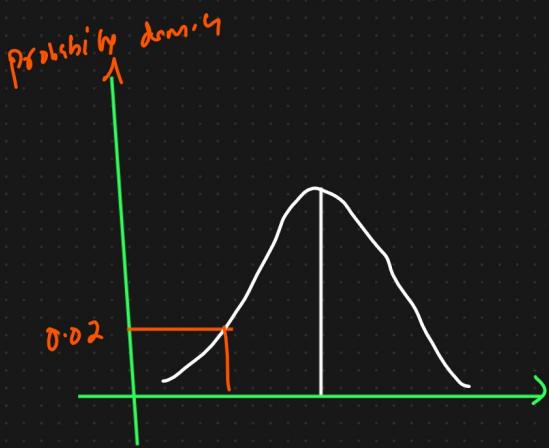
= 0/p

② Probability Density Function

① Distribution of continuous Random Variable



Probability Density \Rightarrow Gradient of Cumulative Curve



Different types of Distribution

① Normal / Gaussian Distribution \rightarrow pdf

② Standard Normal Distribution \rightarrow pdf

③ Log Normal Distribution \rightarrow pdf

④ Power law Distribution \rightarrow pdf

⑤ Bernoulli Distribution \rightarrow pmf

⑥ Binomial Distribution \rightarrow pmf

⑦ Poisson Distribution \rightarrow pmf

agr hm cdf ko smooth krde toh hme milta hai
S curve. or agr hm usme hr ek point pei gradient
nikaale toh hme milta hai PDF

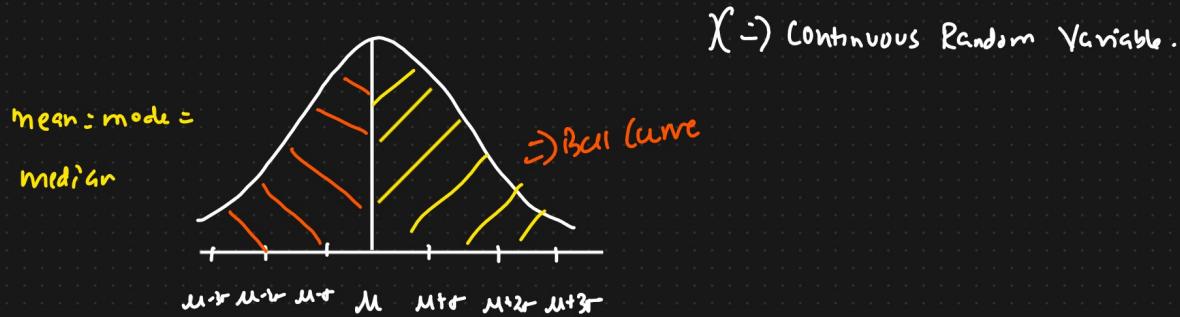
⑧ Uniform Distribution \rightarrow Discrete \rightarrow pmf
 \rightarrow Continuous \rightarrow pdf

⑨ Exponential Distribution. \rightarrow pdf

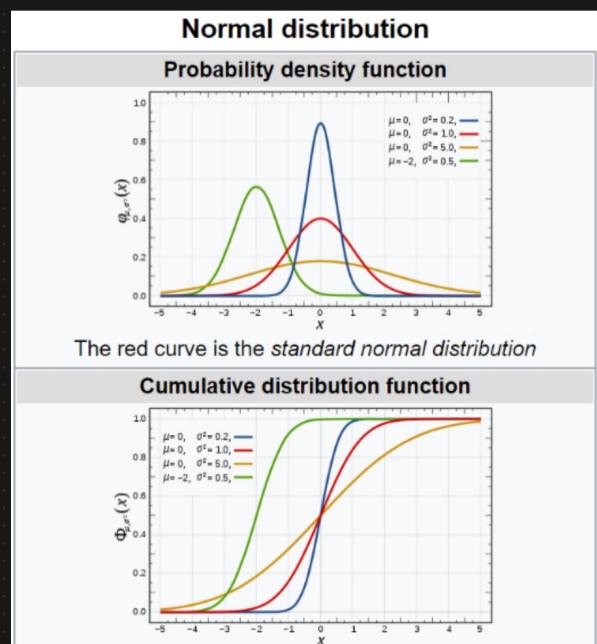
⑩ CHI SQUARE Distribution \rightarrow pdf

⑪ F Distribution \rightarrow pdf.

⑫ Normal/Gaussian Distribution



Eg:- Height, weight, age, IRIS dataset



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

Support parameters $\mu = \text{mean}$
 $\sigma^2 = \text{variance}$

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

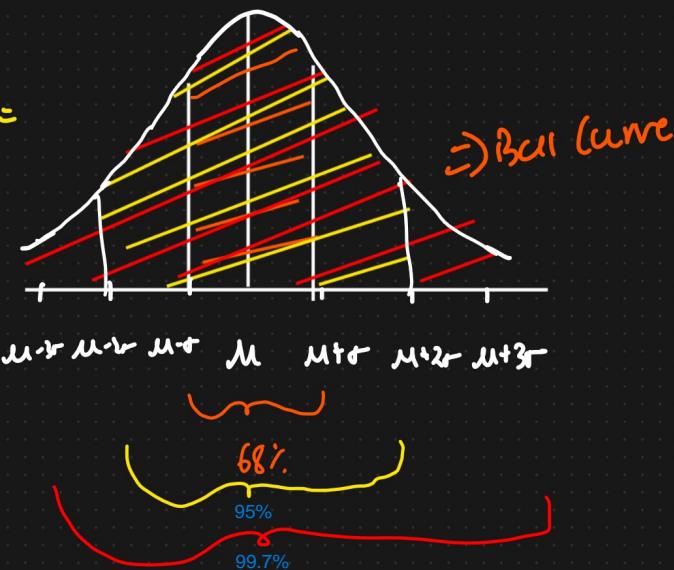
Empirical Rule $68 - 95 - 99.7\%$ Rule

100 datapoint

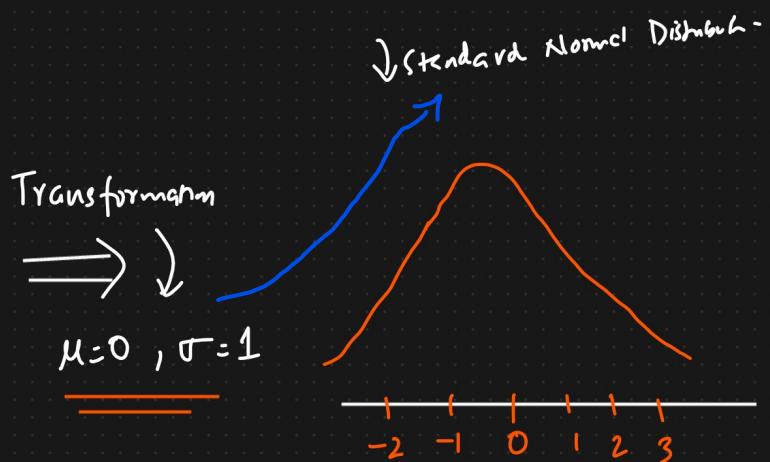
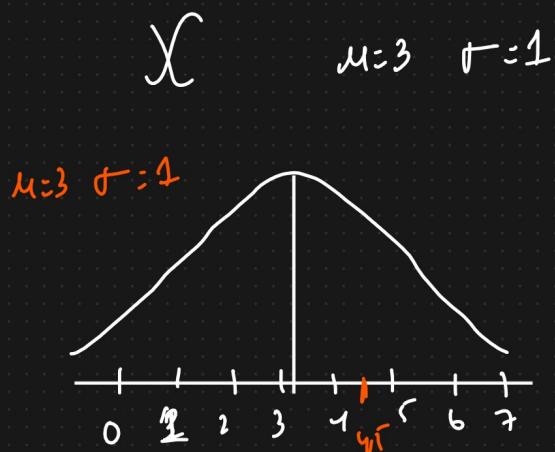
$$X = \{ \quad \}$$

mean = mode =

median



Standard Normal Distribution



by applying this formula, we can find standard normal distribution

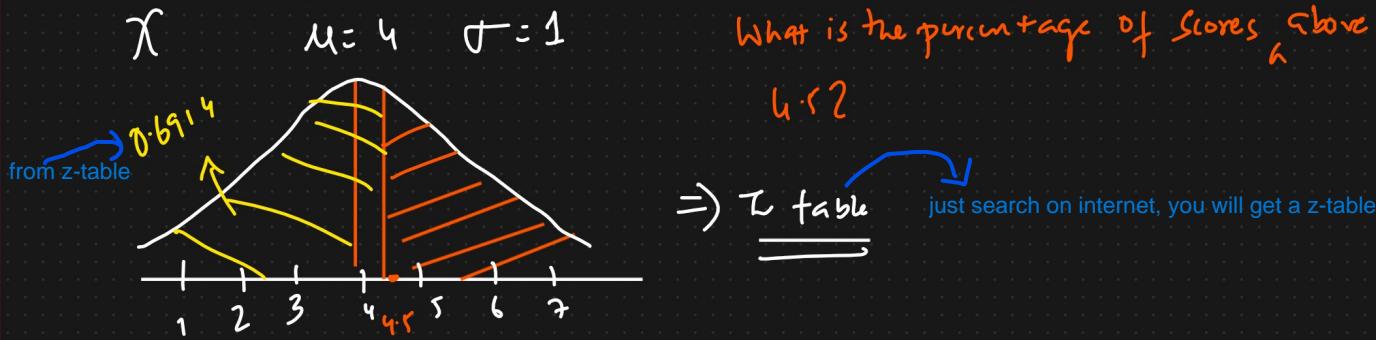
$$\text{Z-score} = \frac{x_i - \mu}{\sigma}$$

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

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

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

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



$$z\text{-score} = \frac{4.5 - 4}{1} = 0.5 \Leftrightarrow$$

$$\begin{aligned} \text{Area under curve} &= 1 - 0.6914 \\ &= 0.3086 // = 30.86\% \end{aligned}$$