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 distributions

## 1. Percentiles And Quartiles [GATE, CAT]

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

% 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.

$n = 15$

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

Percentile Rank of (value) 10 = $\frac{\#\text{ of values below } 10}{n} * 100$

$= \frac{12}{15} \times 100 = 80$ percentile

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

Here are the steps on how to calculate the percentile of a value:

Collect the data set.
Arrange the data set in ascending order.
Determine the total number of observations.
Identify the data value for which you are interested to find the percentile.
Count the number of data values that are less than the above value.
Divide the number from Step 5 by the number from Step 3 to find the percentile of the given data value.

② What value exists at 25 percentile ?

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

$$= \frac{25}{100} * 16 = 4^{th} \text{ Element} = 4$$

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

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

4.5

If value will come in decimal, then take avg of that no. before decimal and the next no.

then that'th no. will be the value

## Quartiles

① Q1 → 25 percentile

Q2 → Median → 50 percentile

Q3 → 75 percentile

## ② 5 Number Summary

(1) Minimum
(2) First Quartile (25 percentile) (Q1)
(3) Median (Q2)
(4) Third Quartile (75 percentile) (Q3)
(5) Maximum

After removing the outlier

## Remove the Outliers

Meaning of taking this lower fence and higher fence is that below lower fence everything will be outlier and above higher fence everything will be outliers.

These fence only use for checking outliers.

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

[Lower Fence ⟷ Higher Fence]

Lower Fence = Q1 − 1.5(IQR)

Inter Quartile Range = Q3 − Q1

Higher Fence = Q3 + 1.5(IQR)

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

$Q1 = 25^{th}$ percentile $= \frac{25}{100} * 20 = 5^{th}$ value $= 3$

$Q3 = 75^{th}$ percentile $= \frac{75}{100} * 20 = 15^{th}$ value $= 7$

$IQR = 7 - 3 = 4$

Lower Fence $= Q_1 - 1.5(IQR)$
$= 3 - 1.5(4)$
$= 3 - 6$
$= -3$

Higher Fence $= Q3 + 1.5(IQR)$
$= 7 + 1.5(4)$
$= 7 + 6$
$= 13$

$[-3, 13]$.

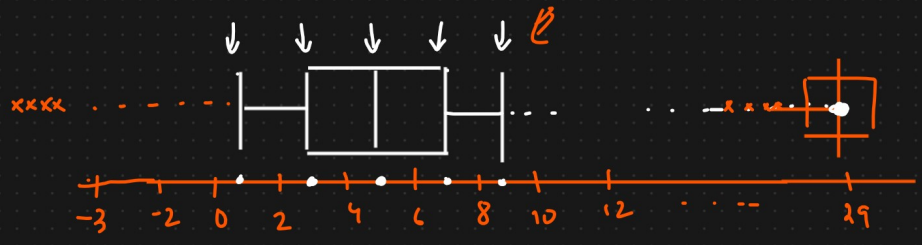
$X = \{1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 29\}$ (29 is outlier)

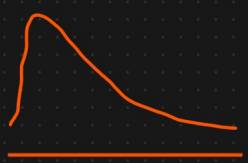
## Box plot [To visualize Outliers]

5 number summary is used for Box plot

1. Minimum value = 1
2. Q1 = 3
3. Median Q2 = 5
4. Q3 = 7
5. Maximum = 9

Box plot

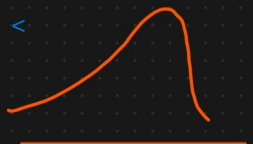


mean > median > mode


$Q3 - Q2 > Q2 - Q1$

Here Q3 - Q2 < Q2 - Q1



Q2 is left side as more data is towards right side. So Q3 - Q2 should be higher that Q2 - Q1.

## Internal Assignment

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

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

[Lower Fence ⟷ Higher Fence]

$Q1 = \frac{25}{100} \times 17 = 4.25$

$Q3 = \frac{75}{100} \times 17 = 12.75$

Lower Fence $= -1 - 1.5(10+1)$

$= -1 - 1.5(11)$

$= -17.5$

Higher Fence $= 10 + 1.5(10+1)$

$= 10 + 16.5$

$= 26.5$

$[-17.5, 26.5]$.

⇓

55 is an outlier

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

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

$Q1 = 5$

$Q3 = 71.5$

$[-94.75, 171.25]$

$Q3 = \frac{75}{100} \times 14 = \frac{42}{4} \; 2 + 10.5$

$Q3 = \frac{66+77}{2} = 71.5$

## Covariance And Correlation

[Relationship between X and Y]

| X | Y |
|---|---|
| 2 | 3 |
| 4 | 5 |
| 6 | 7 |

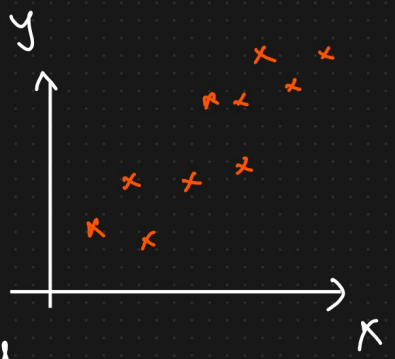
| | | |
|---|---|---|
| → | X↑ | Y↑ |
| → | X↓ | Y↑ |
| → | X↑ | Y↓ |

IS Haunted

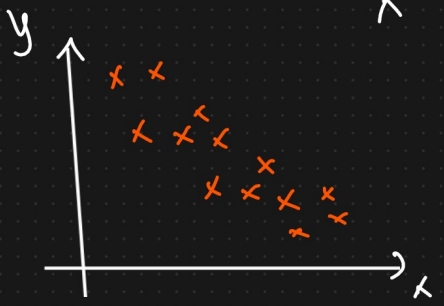
Size of House → Price

8 9

→ | x↓ y↓ |




| x↑ y↑ |
| --- |
| x↓ y↓ |




| x↑ y↓ |
| --- |
| x↓ y↑ |

## ① Covariance

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

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

After applying above formula if ---

Cov(x,y)

| x↑ y↑ |
| --- |
| x↓ y↓ |

⇒ +ve Covariance

| x↑ y↓ |
| --- |
| x↓ y↑ |

⇒ -ve Covariance.

⇓

| Var(x) ⇐ Cov(x,x) |
| --- |

Variance of x is covariance of x, x means relationship of x with x itself is covariance of (x,x)

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

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

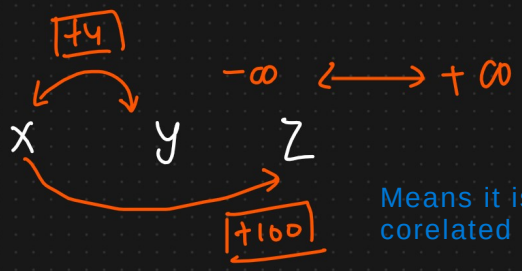
$$= \frac{[(2-4)(3-5) + (4-4)(5-5) + (6-4)*(7-5)]}{2}$$

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

X and y are having a positive covariance

## Advantages

① Relation ship between X & y



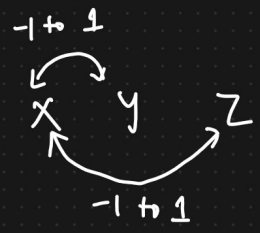

## Disadvantage

① Covariance does not have a specific limit value

It can be any +ve value or any -ve value but can't be any limit or strength that is why there is no limit in covariance

Means it is not saying if x is highly corelated with Y or Z.

① Pearson Correlation Coefficient Range [-1 to 1]




For linear dataset

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

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

② The more the " " -1 " " -ve " " "

| | | |
|---|---|---|
| X | y | 0.6 |
| X | Z | 0.7 |

So from here we can say that X is high correlated with Z than Y with the help of Pearson correlation function.

② Spearman Rank correlation

For non linear dataset will also be capture

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

R is rank

| 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 |

Created rank wise table which are bigger no.

## Feature Selection

How these value correlated with price

| +ve | +ve | +ve | ≈0 | -ve | O/P |
|---|---|---|---|---|---|
| Size of house | No. of Rooms | Location | ~~No. of people Staying~~ | Haunted | Price ↑ |

## Probability Distribution Function And Probability Density Function

### Probability Mass function

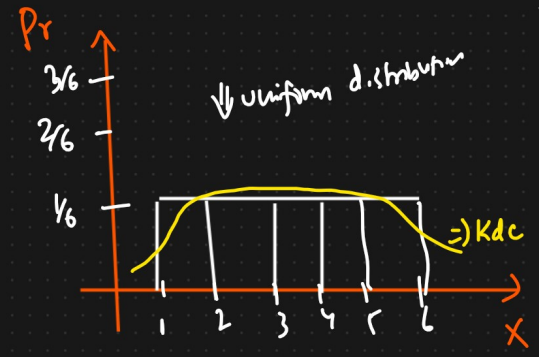
### Probability Distribution Function

① Probability density fn
② Probability mass fn ✓
③ Cummulative distribute fn.

① PMF

① Discrete Random Variable

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



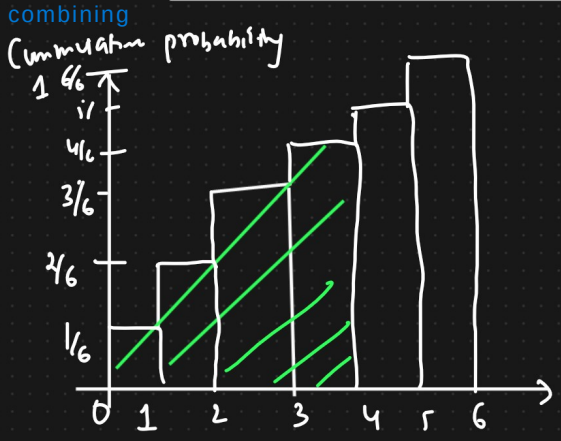

$Pr(1) = 1/6$

$Pr(2) = 1/6$

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

### Cummulative distribution function (cdf)

combining

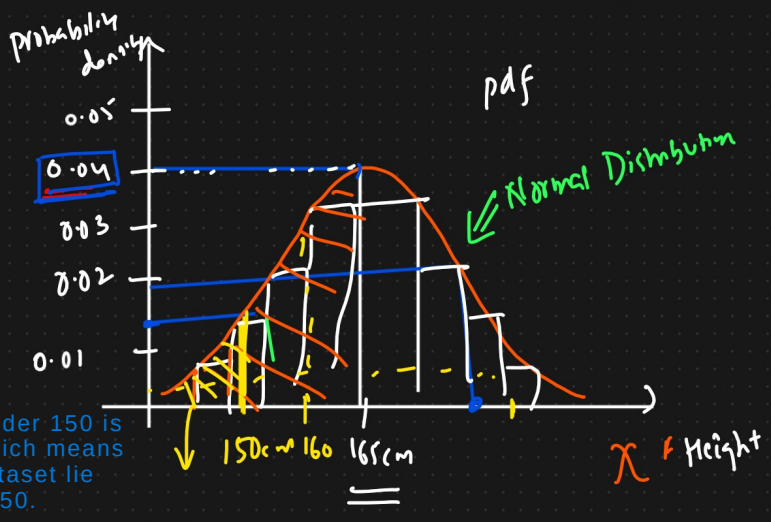



$Pr(X \leq 4) = Pr(X=1) + Pr(X=2) + Pr(X=3) + Pr(X=4)$

$= O/P$

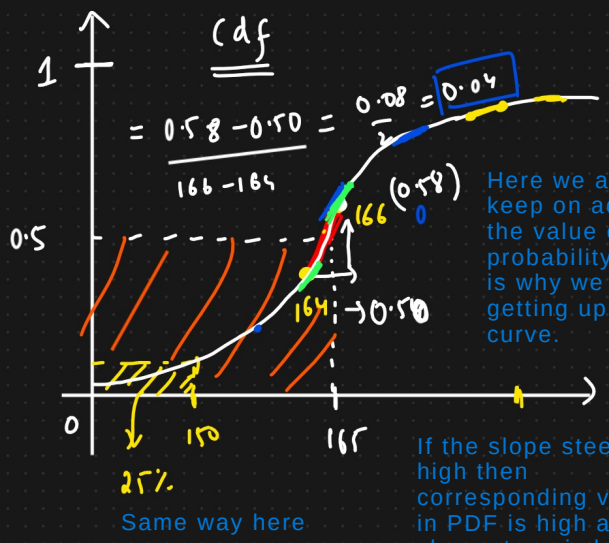
## ② Probability Density Function

### ① Distribution of continuous Random Variable




Area under 150 is 0.25 which means 25% dataset lie under 150.

(A) 0.25

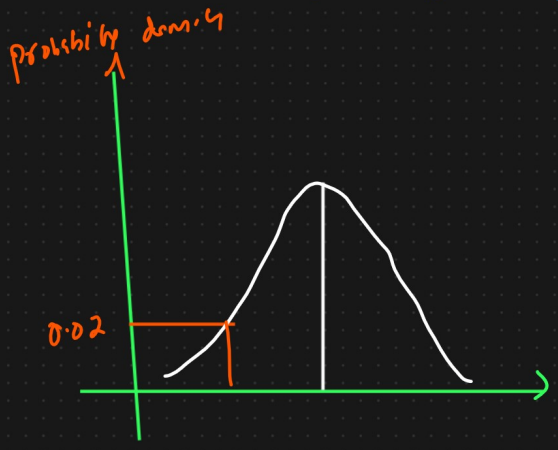


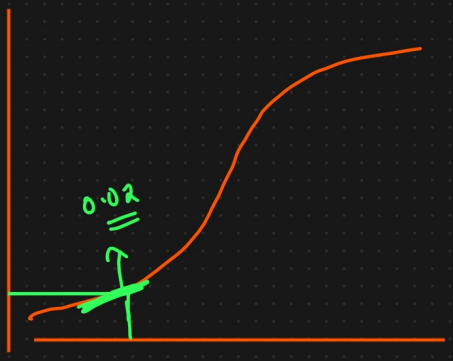

Same way here

Here we are keep on adding the value of probability that is why we are getting upper curve.

If the slope steep is high then corresponding value in PDF is high and if slope steep is less the value at PDF is very less that is why after max value second half is going down in PDF

Probability Density ⇒ Gradient or Slope of Cumulative Curve




## Different types of Distribution

① Normal / Gaussian 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 → pmf
→ Continuous → pdf

⑨ Exponential Distribution. → pdf

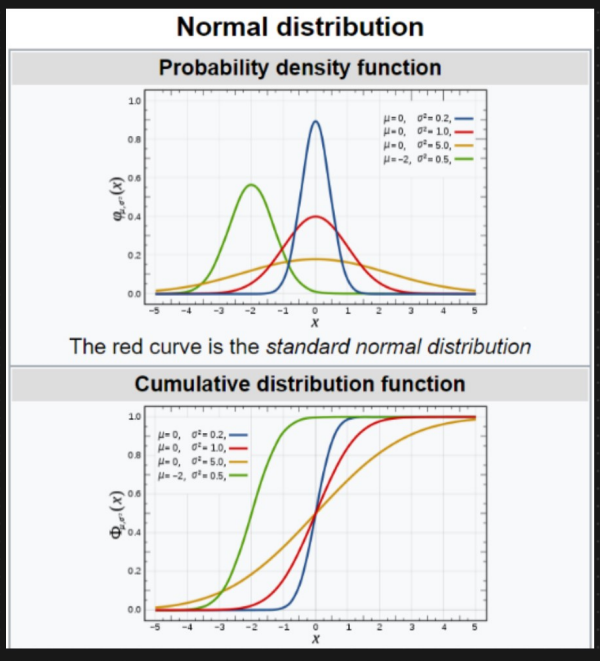
⑩ CHI SQUARE Distribution → pdf

⑪ F Distribution → pdf.

## ① Normal/Gaussian Distribution

X :) Continuous Random Variable.


Eg: Height, weight, age, IRIS dataset




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

Support parameters $\mu$ = mean

$\sigma^2$ = variance

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

The empirical rule, also known as the 68-95-99.7 rule or the three-sigma rule, is a statistical rule that applies to normally distributed data. It states that approximately 68% of the data will fall within one standard deviation of the mean, 95% of the data will fall within two standard deviations of the mean, and 99.7% of the data will fall within three standard deviations of the mean.

The empirical rule can be used to make inferences about the probability of a particular value occurring in a normally distributed population. For example, if we know that the average height of a population of adults is 170 centimeters and the standard deviation is 5 centimeters, we can use the empirical rule to say that 68% of the adults in the population will have heights between 165 and 175 centimeters.

## Empirical Rule — 68 - 95 - 99.7% Rule

100 datapoint

$X = \{ \quad \}$

mean : mode : median

$\mu-3\sigma \quad \mu-2\sigma \quad \mu-\sigma \quad \mu \quad \mu+\sigma \quad \mu+2\sigma \quad \mu+3\sigma$

⇒ Bell Curve

First Standard deviation range-- 68% data lie

68%

Second SD range -- 95% data lie

Third SD range -- 99.7 % data lie

## Standard Normal Distribution

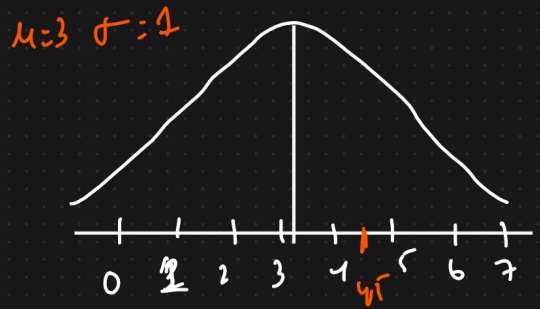
$X \qquad \mu = 3 \quad \sigma = 1$

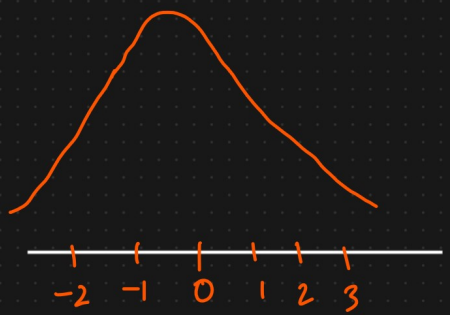
↓ Standard Normal Distribution

Then we get

Applied some Transformation ⟹

$\mu = 0, \sigma = 1$




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

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

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

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

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

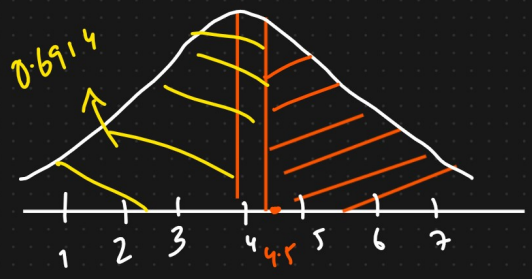
Z-score tells you about a value how many standard deviation away from the mean

In other words, if we want to find for any value that it is how many STandard deviation Away from the mean. Then Z - score is used.

$X$ $\mu = 4$ $\sigma = 1$

What is the percentage of scores lies above 4.5?

Simply mean what is the area above 4.5 in the graph




$\Rightarrow$ $z$ table

$$z\text{-score} = \frac{4.5 - 4}{1} = 0.5$$ ⇐ Means 0.5 SD Away from the mean

Area under curve $= 1 - 0.6914$

Check from z table from google PDF and check for +ve SD 0.5 and it will area under curve left side that is why doing minus for are above 4.5 score here.

$= 0.3086 = 30.86\%$