What Is Statistics?

Statistics is a branch of mathematics that deals with collecting, analyzing, interpreting, and visualizing empirical data. Descriptive statistics and inferential statistics are the two major areas of statistics. Descriptive statistics are for describing the properties of sample and population data (what has happened). Inferential statistics use those properties to test hypotheses, reach conclusions, and make predictions (what can you expect).

• Use of Statistics in Machine Learning

1. Asking questions about the data
2. Cleaning and preprocessing the data
3. Selecting the right features
4. Model evaluation
5. Model prediction

• Population and Sample

• Population: (N)

In statistics, the population comprises all observations (data points) about the subject under study.

An example of a population is studying the voters in an election. In the 2019 Lok Sabha elections, nearly 900 million voters were eligible to vote in 543 constituencies.

• Sample: (n)

In statistics, a sample is a subset of the population. It is a small portion of the total observed population.

An example of a sample is analyzing the first-time voters for an opinion poll.

- # Measures of Central Tendency

Measures of central tendency are the measures that are used to describe the distribution of data using a single value. Mean, Median and Mode are the three measures of central tendency

Mean:

The arithmetic mean is the average of all the data points.

If there are n number of observations and xi is the ith observation, then mean is:

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

Example :->

Consider the data frame below that has the names of seven employees and their salaries.

	Name	Salary
0	Jane	50000
1	Michael	54000
2	Willian	50000
3	Rosy	189000
4	Hana	55000
5	Ferdie	40000
6	Graeme	59000

To find the mean or the average salary of the employees, you can use the mean() functions in Python.

```
print(df['Salary'].mean())
```

```
71000.0
```

Solution :=>

```
50000 + 54000 + 50000 + 189000 + 55000 + 40000 + 59000
= 497000
```

There are 7 employees.

$$\text{Mean Salary} = \frac{497000}{7} = 71000$$

Ans:

Median:

Median is the middle value that divides the data into two equal parts once it sorts the data in ascending order.

If the total number of data points (n) is odd, the median is the value at position (n+1)/2.

When the total number of observations (n) is even, the median is the average value of observations at n/2 and (n+2)/2 positions.

The median() function in Python can help you find the median value of a column. From the above data frame, you can find the median salary as:

```
print(df['Salary'].median())
```

```
54000.0
```

Question :=>

Consider the data frame below that has the names of seven employees and their salaries.

	Name	Salary
0	Jane	50000
1	Michael	54000
2	Willian	50000
3	Rosy	189000
4	Hana	55000
5	Ferdie	40000
6	Graeme	59000

Solution

Step 1: Sort salaries

[40000, 50000, 50000, 54000, 55000, 59000, 189000]

Step 2: Find the middle value (since there are 7 values)

The 4th value (middle one) = 54000

Median Salary = 54000

$\left(\frac{7+1}{2}\right)$

$\Rightarrow 4^{th}$ value

Mode:

The mode is the observation (value) that occurs most frequently in the data set. There can be over one mode in a dataset.

The mode salary from the data frame can be calculated as:

```
print(df['Salary'].mode())
```

```
0    50000
dtype: int64
```

Question

Given below are the heights of students (in cm) in a class:

155, 157, 160, 159, 162, 160, 161, 165, 160, 158

Mode = 160 cm.

Outliers

Outliers are values "**outside**" the other values:

```
99,86,87,88,111,86,103,87,94,78,300,85,86
```

Outliers can change the mean a lot. Sometimes we don't use them (they might be an error), or we use the median or the mode instead.

Descriptive v/s Inferential Statistics

Descriptive Statistics

Descriptive Statistics summarizes (describes) observations from a set of data.

Since we register every newborn baby, we can tell that 51 out of 100 are boys.

From these collected numbers, we can predict a 51% chance that a new baby will be a boy.

It is a mystery that the ratio is not 50%, like basic biology would predict. We only know that we have had this tilted sex ratio since the 17th century.

Note

Raw observations are only data. They are not real knowledge.

You use **Descriptive Statistics** to transform raw observations into data that you can understand.

Descriptive Statistics Measurements

Descriptive statistics are broken down into different measures:

Tendency (Measures of the Center)

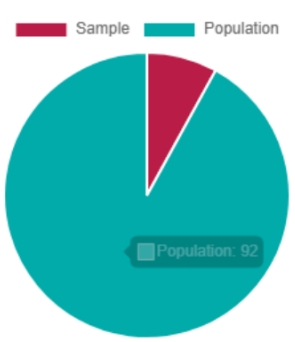
- The Mean (the average value)value
- The Median (the mid point value)
- The Mode (the most common value)

Spread (Measures of Variability)

- Min and Max
- Standard Deviation
- Variance
- Skewness
- Kurtosis

Inferential Statistics

Inferential statistics are methods for quantifying properties of a population from a small **Sample**:



You take data from a sample and make a prediction about the whole population.

For example, you can stand in a shop and ask a **sample of 100 people** if they like chocolate.

From your research, using inferential statistics, you could predict that 91% of **all shoppers** like chocolate.

Incredible Chocolate Facts

Nine out of ten people love chocolate.

50% of the US population cannot live without chocolate every day.

You use **Inferential Statistics** to predict whole domains from small samples of data.

Descriptive Statistics is broken down into **Tendency** and **Variability**.

Tendency is about **Center Measures**:

- The **Mean** (the average value)
- The **Median** (the mid point value)
- The **Mode** (the most common value)

Variability uses these measures:

- Min and Max
- Variance
- Deviation
- Distribution
- Skewness
- Kurtosis

The Variance

In statistics, the **Variance** is the average of the squared differences from the **Mean Value**.

In other words, the variance describes how far a set of numbers is **Spread Out** from the mean (average) value.

Mean value is described in the previous chapter.

Question

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n}$$

This table contains 11 values:

7	8	8	9	9	9	10	11	14	14	15

Calculate the Variance:

```
// Calculate the Mean (m)
let m = (7+8+8+9+9+9+10+11+14+14+15)/11;

// Calculate the Sum of Squares (ss)
let ss = (7-m)**2 + (8-m)**2 + (8-m)**2 + (9-m)**2 + (9-m)**2 + (9-m)**2 + (9-m)**2 + (10-m)**2 + (11-m)**2 + (14-m)**2 + (15-m)**2;

// Calculate the Variance
let variance = ss / 11;
```

To calculate the variance of the Grade, use the following:

```
print(df['Grade'].var())
685.6190476190476
```

The **variance** is a statistical measure that describes how spread out the values in a dataset are. There are **two types of variance** depending on the context:

✅ 1. Population Variance (σ^2)

Used when you have data for the **entire population**.

Formula:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

- σ^2: population variance
- N: total number of data points
- x_i: each individual data point
- μ: population mean

✅ 2. Sample Variance (s^2)

Used when you have a **sample** (subset) of the population.

Formula:

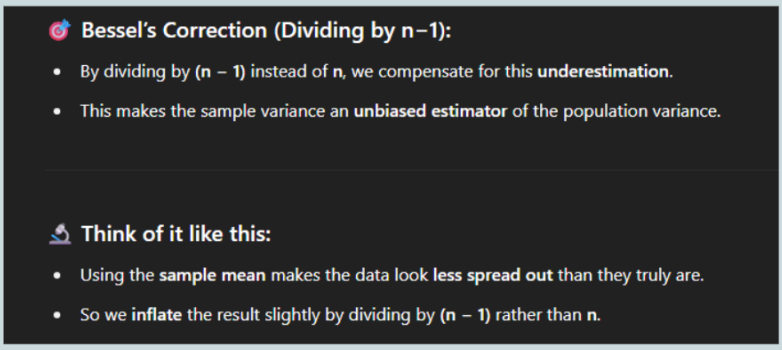
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- s^2: sample variance
- n: number of sample data points
- x_i: each sample data point
- $\bar{x}$: sample mean

> 🔍 **Why divide by (n – 1) instead of n?**
> This is called **Bessel's correction**. It corrects the bias in estimating the population variance from a sample.

Metric	Formula	Denominator	Used For
Population Variance	$\frac{1}{N}\sum(x_i - \mu)^2$	N	Entire population
Sample Variance	$\frac{1}{n-1}\sum(x_i - \bar{x})^2$	n – 1	Subset of population

Why divide by (n − 1) instead of n? In sample variance



Standard Deviation

Standard Deviation is a measure of how spread out numbers are.
The symbol is **σ** (Greek letter sigma).
The formula is the **√** variance (the square root of the variance).

The Standard Deviation is (in JavaScript):

```
// Calculate the Mean (m)
let m = (7+8+8+9+9+9+10+11+14+15)/11;

// Calculate the Sum of Squares (ss)
let ss = (7-m)**2 + (8-m)**2 + (8-m)**2 + (9-m)**2 + (9-m)**2 + (9-m)**2 + (9-m)**2 + (10-m)**2 + (11-m)**2 + (14-m)**2 + (15-m)**2;

// Calculate the Variance
let variance = ss / 11;

// Calculate the Standard Deviation
let std = Math.sqrt(variance);
```

Deviation is a measure of **Distance**.

How far (on average), all values are from **the Mean** (the Middle).

You can find the standard deviation using the std() function in Python.

```
print(df['Grade'].std())
```

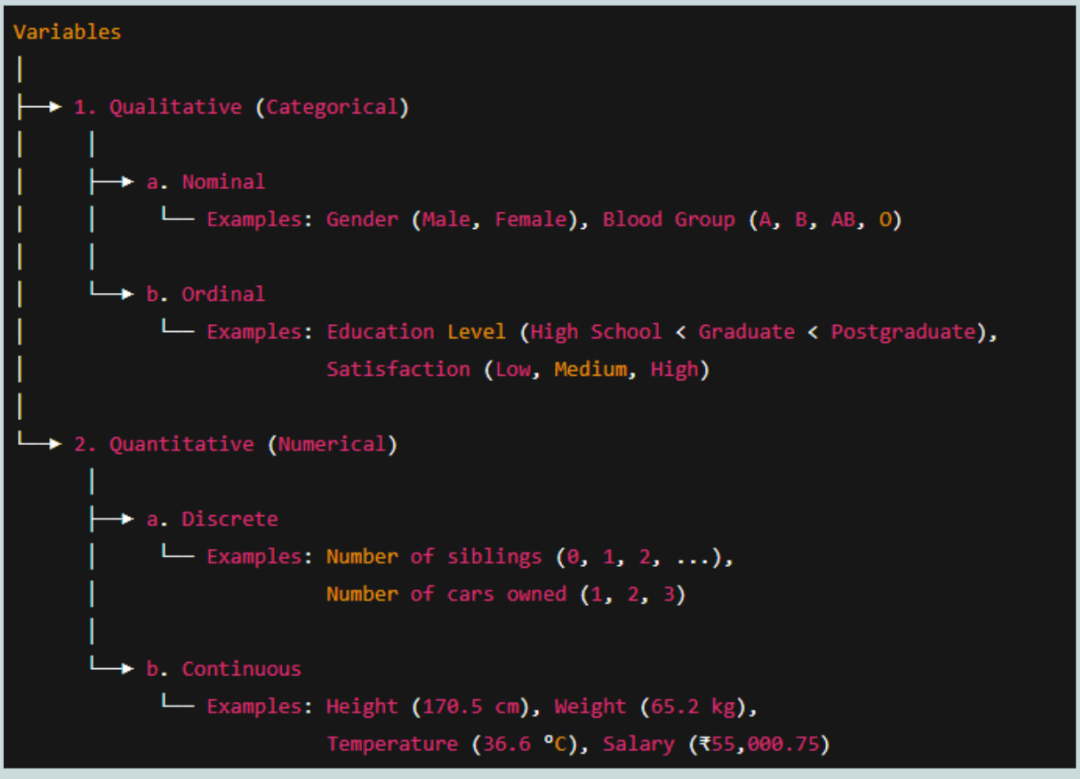
```
26.184328282754315
```

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n}}$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{\text{variance}}$$

Variables

Any characteristic that can be measured or categorized.



1. Qualitative Variables (Categorical)

Represent categories or labels; **non-numeric**.

a. Nominal

Categories with **no natural order**.

Example: Gender (Male, Female), Blood Type (A, B, O)

b. Ordinal

Categories with a **meaningful order**, but **no fixed difference** between levels.

Example: Satisfaction (Low, Medium, High), Education Level (UG < PG < PhD)

2. Quantitative Variables (Numerical)

Represent **numeric values**; can be **measured** or **counted**.

a. Discrete

Countable values with gaps; usually whole numbers.
Example: Number of siblings (0, 1, 2), Number of books

b. Continuous

Can take **any value within a range**, including decimals.
Example: Height (165.3 cm), Temperature (36.6°C), Salary (₹50,000.75)

- Random variables

A **random variable** is a variable that takes on **numerical values** determined by the **outcome of a random experiment**.

In simple terms:
A random variable maps the outcomes of a random process (like rolling a die or flipping a coin) to numbers.

Types of Random Variables

1. Discrete Random Variable

Takes a **finite or countable** number of values.
Example:
Let X be the number shown on a rolled die.
Possible values: $X=\{1,2,3,4,5,6\}$

2. Continuous Random Variable

Can take **any value within a range** (infinite possibilities).
Example:
Let Y be the height of a person in cm.
Possible values: $Y \in [140,200]$ (like 170.3, 165.8 etc.)

Example:

Experiment: Toss a coin twice

- Sample space: {HH, HT, TH, TT}

- Define a random variable X: number of heads

Then:

- HH → X = 2
- HT → X = 1
- TH → X = 1
- TT → X = 0

So, X∈{0,1,2} → **Discrete Random Variable**

Histogram

A **histogram** is a graphical representation used to visualize the **distribution of a numerical dataset**. It shows how **frequently** data falls within **specific ranges** or **intervals** (called **bins**).

It looks like a **bar chart**, but it shows the **distribution** (not categories).

Key Features of a Histogram:

- X-axis: **Intervals (bins)** of data values
- Y-axis: **Frequency** (how many data points fall in each bin)
- Bars **touch** each other (unlike bar charts)

Example:

Suppose we have test scores of 15 students:

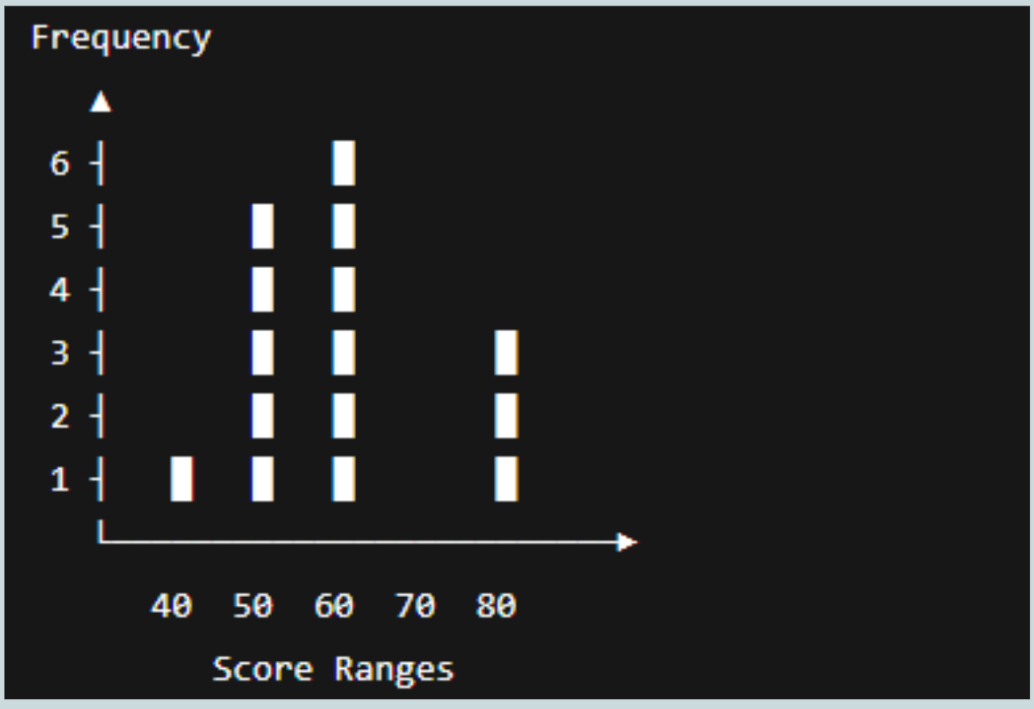
makefile Copy Edit

```
Scores: 45, 50, 52, 53, 55, 57, 59, 61, 62, 65, 67, 68, 70, 72, 75
```

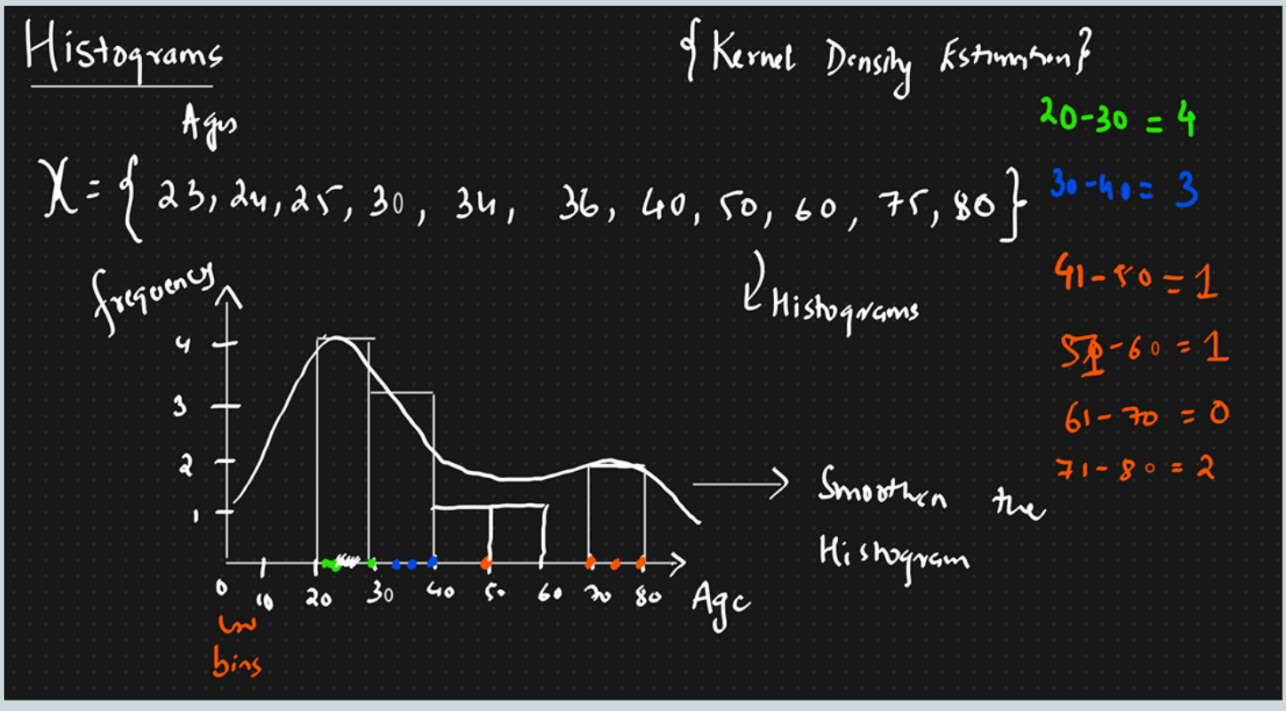
We can create bins:

Bin (Score Range)	Frequency (Count)
40–49	1
50–59	5
60–69	6
70–79	3

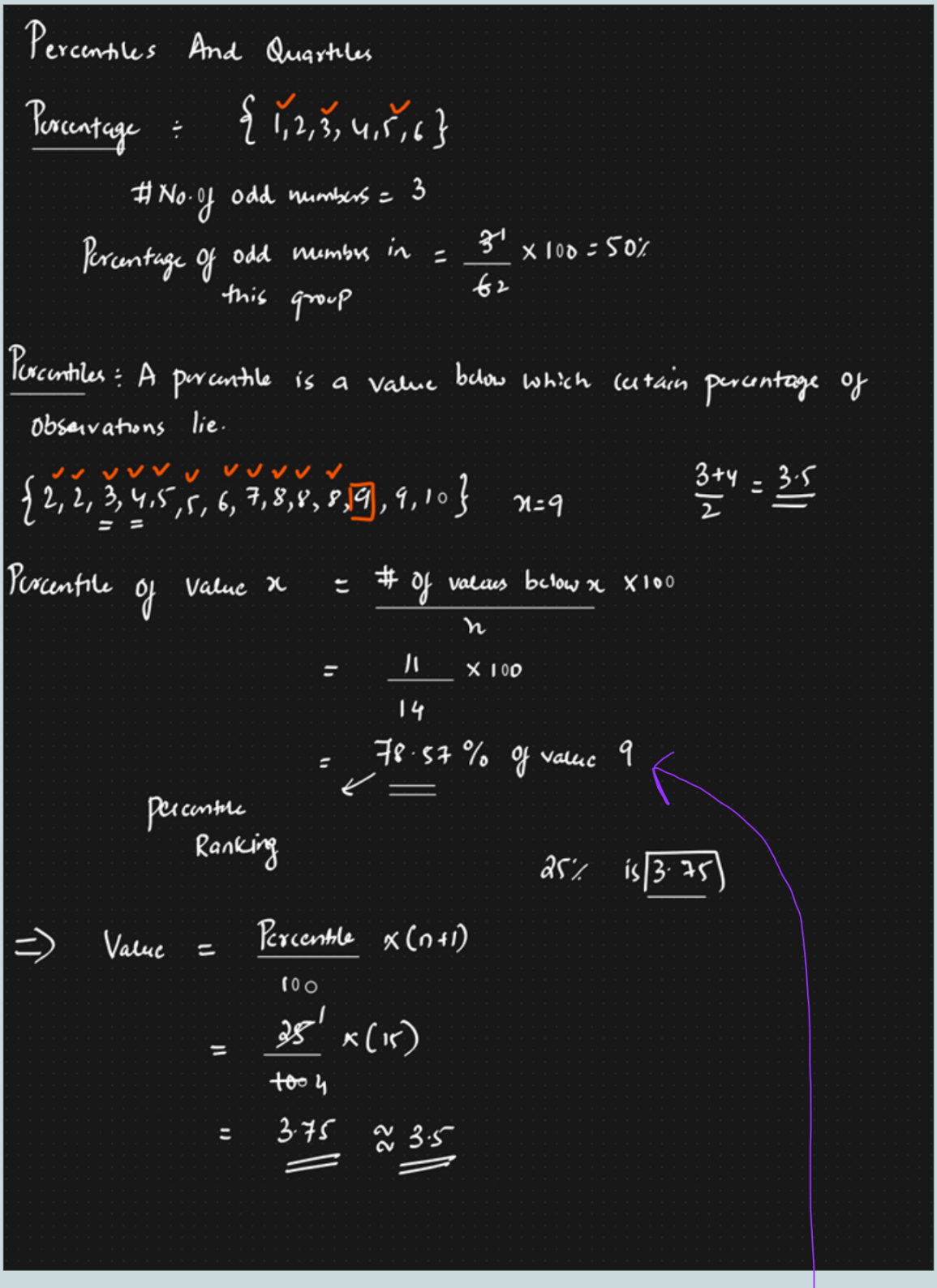
Solution :→



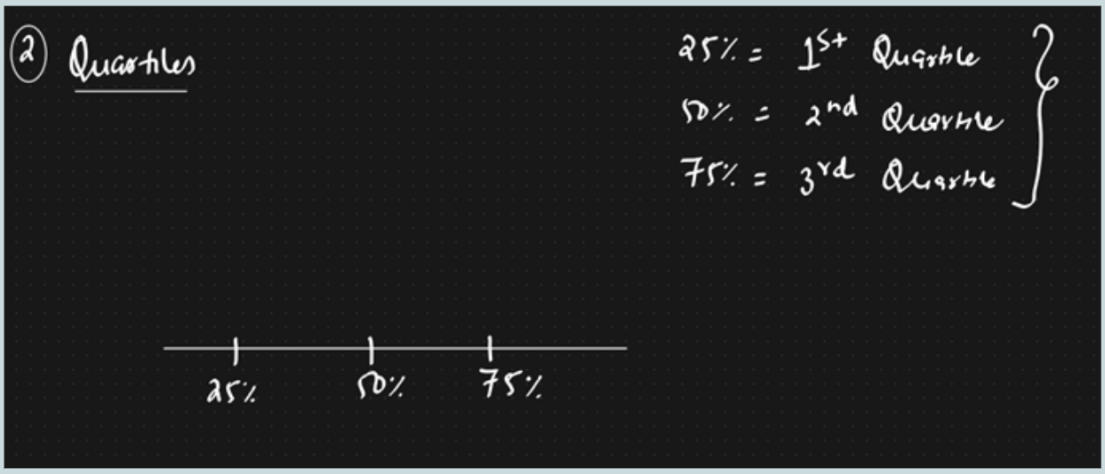
Ques :→



Percentile and Quartiles



78.57 percentage of entire distribution is less than 9



- ## 5 Number Summary

1). Minimum
2). 1st Quartile (25%) (Q1)
3). Median
4). 3rd Quartile (75%) (Q3)
5). Maximum

$$Q1/Q3 = \frac{\text{Percentile}}{100} \times (n+1)$$

Eg:- 1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 27 ← Outlier

[Lower fence ⟷ Higher fence]

Anything outside this is considered as outliers

Lower fence = $Q1 - 1.5(IQR)$

Higher fence = $Q3 + 1.5(IQR)$

$$IQR = Q3 - Q1$$

IQR → Inter Quartile Range

(25%)

$$Q1 = \frac{\text{Percentile}}{100} \times (n+1) = \frac{25}{100} \times (19+1) = 5^{th} \text{ position}$$

value ⇒ 3

75%.

$Q3 = \frac{75}{100} \times 20 = 15^{th}$ position

value $\Rightarrow 7$

$\Rightarrow IQR = Q3 - Q1$

$= 7 - 3 = 4$

$\Rightarrow$ Lower fence $= Q1 - 1.5(IQR)$

$= 3 - 1.5(4)$

$= 3 - 6 = -3$

$\Rightarrow$ Higher fence $= Q3 + 1.5(IQR)$

$= 7 + 1.5 \times 4$

$= 13$

$[-3, 13]$

So, anything below -3 or greater than 13 is considered as outlier

So, 27 is outlier

So, After removing outlier

Eg :- 1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9

So, Minimum = 1

1st Quartile = 3

Median = 5

3rd Quartile = 7

Maximum = 9

Box - Plot ⇒ This plot is used to visualise the outliers

Covariance And Correlation

Covariance and correlation are two statistical measures used to determine the relationship between two variables. Both are used to understand how changes in one variable are associated with changes in another variable.

Covariance

Definition: Covariance is a measure of how much two random variables change together. If the variables tend to increase and decrease together, the covariance is positive. If one tends to increase when the other decreases, the covariance is negative.

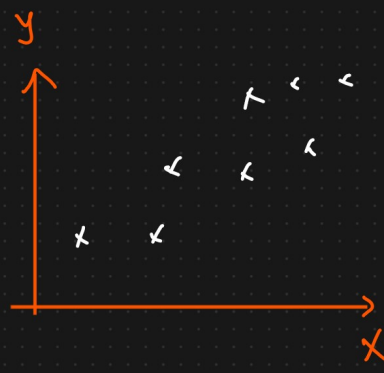
[Quantify the Relationship between X and Y]

X	Y
2	3
4	5
6	7
8	9

X↑	Y↑
X↓	Y↑
X↑	Y↓
X↓	Y↓

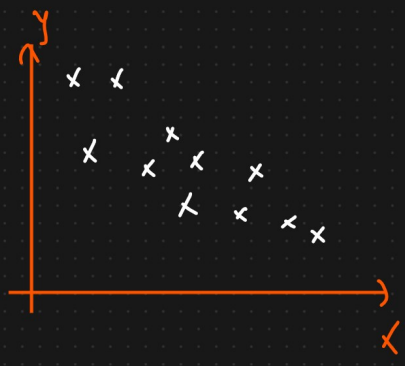
Dataset

↓ ↑ Size of House	Price ↑ ↓
1200	45 lakhs
1300	50 lakh
1500	75 lakh



X↑	Y↑
X↓	Y↓

⇒ +ve Covariance ⇒ +ve value



X↓	Y↑
X↑	Y↓

X	Y
7	10
6	12
5	14
4	16

⇒ -ve Covariance ⇒ -ve value

Covariance

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

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

$$\boxed{Cov(x,x) = Var(x)}$$

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

$x_i \rightarrow$ Datapoint of random variable x

$\bar{x} \rightarrow$ Sample mean of x

$y_i \rightarrow$ Datapoints of random variable y

$\bar{y} \rightarrow$ Sample mean of y

Students

Hour studied (X)	Exam Score (Y)
2	50
3	60
4	70
5	80
6	90

$x\uparrow \; y\uparrow$ $x\downarrow \; y\downarrow$ $\Rightarrow$ +ve covariance

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

① $\bar{x} = \frac{2+3+4+5+6}{5} = 4$ //

② $\bar{y} = \frac{50+60+70+80+90}{5} = 70$ //

$$Cov(x,y) = \frac{(2-4)(50-70)+(3-4)(60-70)+(4-4)(70-70)+(5-4)(80-70)+(6-4)(90-70)}{4}$$

$Cov(x,y) = 20$ //.

$\Rightarrow$ The positive covariance indicates the no. of hours studied increased the exam score also.

X	Y
7	10
6	12
5	14

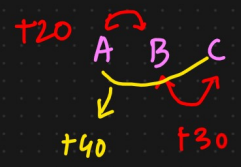
$\Rightarrow$ -Ve //.

$x\uparrow \; y\downarrow$ $x\downarrow \; y\uparrow$

0.96 Cov(A, B) 0.88 Cov(B, C)

-200 +100 20 = Cov(A, B)

-300 +300 30 = Cov(B, C)



Advantages

-1 to 1

Disadvantage

① Quantify the Relationship between X and Y

① Covariance does not have a specific limit value.

$$Cov(X,Y) \Rightarrow -\infty \text{ to } \infty$$

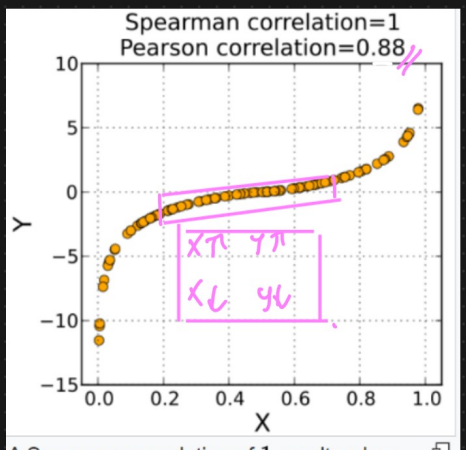
② Correlation
- → Pearson Correlation Coefficient
- → Spearman Rank Correlation

① Pearson Correlation Coefficient $\Rightarrow [-1 \text{ to } 1]$

$$\rho_{x,y} = \frac{Cov(x,Y)}{\sigma_x \cdot \sigma_y} = \frac{20}{\sigma_x \cdot \sigma_y} \Rightarrow 0 \text{ to } 1$$

(Y) The more the value towards +1 the more +ve correlated X & Y is.

(K) The more the value towards -1 the more -ve correlated it is (X,Y)

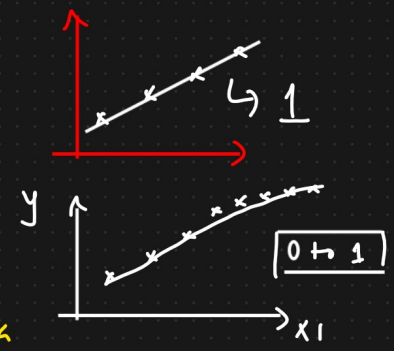


A Spearman correlation of 1 results when the two variables being compared are monotonically related, even if their relationship is not linear. This means that all data points with greater x values than that of a given data point will have greater y values as well. In contrast, this does not give a perfect Pearson

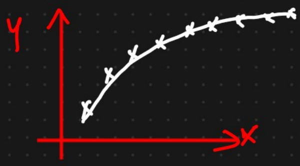
⇒ X↑ Y↑
⇒ X↓ Y↓

Pearson Correlation = 0.88

② Spearman Rank Correlation



Pearson Correlation for non linear data



x	y	R(x)	R(y)
1	2	2	1
3	4	3	2
5	6	4	3
7	8	5	5
0	7	1	4

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

Feature Selection

+ve Correlated

Size of House ↑ | No. of Room ↑ | Location ↑ | ~~No. of People staying in the house~~ (≈0) → NO → Price ↑↑ ↓↓

Haunted ↔ Price: -ve Correlation

o/p: Price