AI/ML Course

Statistics

- **Definition:**

Statistics is the field that deals with collection, organization, analysis, interpretation and the presentation of data

↓ to perform

Decision making

↓ to increase

Profitability in Business

- **Applications:**
  - Machine Learning
  - Data Analyst
  - Risk Analyst
  - Business Analyst
- **Example:**

We are running shop where customers of different age do shopping. So "Age" is field we can apply statistics like mean,

median, mode, or different distributions to provide discount to age group.

⇒ Type of Statistics:

1. Descriptive:

It consists of organizing and summarizing of data.

e.g

(i) Measure of Central Tendency (Mean, Median, Mode)

(ii) Measure of Dispersion (Variance, Standard Deviation)

2. Inferential:

It consists of collecting data called sample data, we perform some experiments and we derive conclusions/ inferences for population

data.

⇛ **Example:**

There is a college A of 1000 students

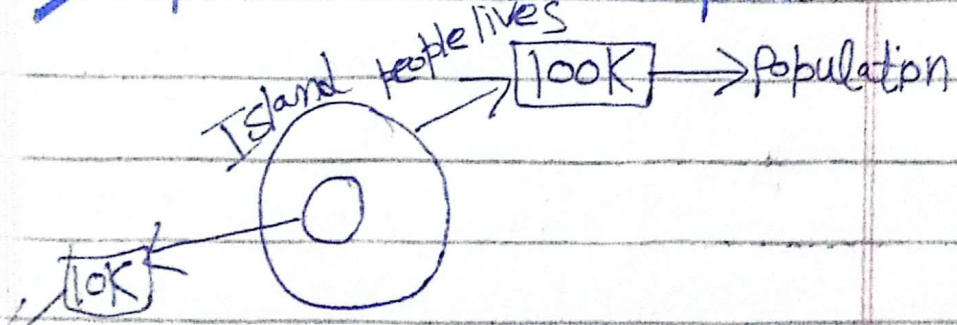
class Statistics

Sample: {180cm, 170cm, 162cm, 150cm, 160cm}

Descriptive: Take mean to find average height in sample.

Inferential: Using this sample inferena the average height of population i.e all 1000 students through experiments.

⇛ **Population and Sample**



mple • **Task:** Collect all the weights of all the people which is very difficult.

So we just pick up a sample from population and perform inferential statistics

Notations

Sample	n
Population	N

- **Usecase:**

Exit poles: Predict the result of elections. So news channel pick up some sample data and try to inference which party will win.

⇒ Measures of Central Tendency:

① Mean:

- **population size(N):**

$$M = \sum_{i=1}^{N} \frac{x_i}{N}$$

- **Sample (n)**

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

- **Example:**

Let's take a field

Ages = {1, 3, 4, 5} → Distribution

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

$$\boxed{\bar{x} = 3.25}$$

- We call it measure of central tendency because it is showing you the mean part of this distribution.

2). Median:

Let Ages = {1, 3, 4, 5, 100}

$$\mu(\text{mean}) = \frac{1+3+4+5+100}{5} = \frac{113}{5}$$

$$\mu(\text{mean}) = 22.6$$

I have just added the outlier (100) a big number, the average mean became so big.

⇒ To overcome the impact of large number we use median.

Let

Ages = {4, 3, 1, 5, 100}

↳ Sort the numbers {1, 3, 4, 5, 100}

↳ Median = 4 (The central element so it overcome outlier)

- **If numbers are odd:**

$$\text{Median} = \left(\left\lfloor \frac{n}{2} \right\rfloor + 1\right)\text{th element}$$

This is floor

- **If numbers are even:**

There will be two middle elements and we calculate their average.

$$\text{Median} = \frac{\left\lfloor \frac{n}{2} \right\rfloor \text{th} + \left(\left\lfloor \frac{n}{2} \right\rfloor + 1\right)\text{th}}{2}$$

- **eg**

Ages = {1, 3, 4, 5, 100, 200}

$$\text{Median} = \frac{\frac{6}{2}\text{th} + \left(\frac{6}{2}+1\right)\text{th}}{2}$$

$$\text{Median} = \frac{4+5}{2} = \boxed{4.5}$$

③ Mode:

It is also used to overcome outlier.

⇛ In mode, we select the element which has maximum frequency (occuring maximum times in data).

Let ages = {4, 3, 2, 1, 1, 4, 4, 5, 2, 100}

mode = 4 (occuring maximum 3 times)

⇛ Measure of Dispersion:

① Variance

② Standard Deviation

- Example:

Let ages = {2, 2, 4, 4} → Distribution1

Ages2 = {1, 1, 5, 5} → Distribution2

μ(mean of ages) = $\frac{2+2+4+4}{4}$ = 3

μ(mean of ages2) = $\frac{1+1+5+5}{4}$ = 3

⇛ Both the distribution have same average (mean) but we can see that both have different elements and data points, so how we differentiate them.

⇛ See the data points of distribution 1 i.e 2 and 4 are close to the average i.e 3 means **less spread**

⇛ The data points of distribution 2 i.e 1 and 5 are far to the average i.e 3 means **more spread**

⇛ So we will find difference b/w two populations through variance and standard deviation.

① **Variance:**

Population (N):

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

Let Ages1 = {2, 2, 4, 4}

Ages2 = {1, 1, 5, 5}

μ(mean of ages 1) $= \frac{2+2+4+4}{4} = \boxed{3}$

μ(mean of ages 2) $= \frac{1+1+5+5}{4} = \boxed{3}$

For ages 1

x_i	μ	$(x_i-\mu)^2$
2	3	1
2	3	1
4	3	1
4	3	1

$\sum (x_i-\mu)^2 = 4$

And $N=4$

So,

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

σ^2(variance) $= \frac{4}{4}$

$\boxed{\sigma^2 = 1}$

So less spread dispersed

For ages 2

x_i	μ	$(x_i-\mu)^2$
1	3	4
1	3	4
5	3	4
5	3	4

$\sum (x_i-\mu)^2 = 16$

And $N=4$

So,

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

σ^2(variance) $= \frac{16}{4}$

$\boxed{\sigma^2 = 4}$

More spread dispersed.

- For sample data(n)

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

n = Sample data

$\bar{x}$ = Sample mean

- Important Interview Question:

In Sample variance, why we devide with (n-1)?

Actually when we pick random sample data from population the

$\bar{x}$ (Sample mean) $\approx$ μ (population mean)

And S^2 (Sample variance) $\approx$ σ^2 (population variance)

But when we pick sample from one side of population the sample (mean) becomes too low or high and dispersion (sample variance) becomes too high which underestimates population variance as:

$\bar{x} \lll \mu$

$s^2 \lll \sigma^2$

So, to avoid under estimation of s^2 we use $(n-1)$ in formula which decreases s^2 by a factor. This technique is also called [redacted] Bessel's Correction

2) Standard Deviation

- **Population Standard Deviation:**

$\sigma = \sqrt{\sigma^2}$ → Population variance

σ → Population Standard Deviation

⇒ Variance gives you the spread/dispersion

⇒ Standard Deviation tells us how far the data point is away from the mean

- **Sample Standard Deviation:**

$s = \sqrt{s^2}$

$s = \sqrt{\text{Sample variance}}$

↓

Sample Standard Deviation.

⇒ Variable:

Def:

Variable is a property that can take up any value

- **e.g**

Age = 25 → Variable can have single value at a time

Gender = Male

Height = 7.2

Different types of variable:

1. Quantitative Variable:

Quantitative Variable has further two types:

(i) Discrete Quantitative variable:

- In discrete quantitative You have to take a whole number. Fraction is not allowed.

- **e.g**

A = 2 (A has 2 children, we can't say A has 2.5 children)

(ii) Continuous Quantitative variable:

In countinuous quantitative variable, we can take fractions. Decimal values are allowed.

e.g

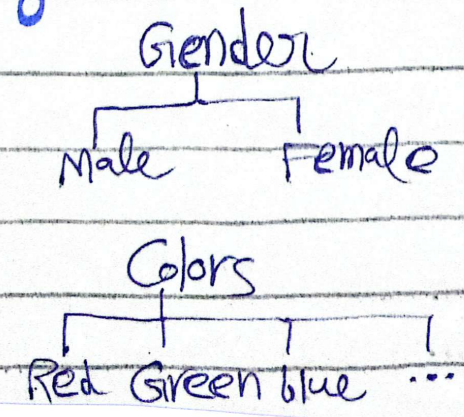
height = 175cm

Weight = 62.32 Kgs

2. Qualitative / Categorical Variable:

In Qualitative/Categorical variable, we don't have numerical values but we have different categories.

e.g



⇒ Random Variables

- Denoted by capital letter i.e X
- Random variable is a function whose values can be derived from some processes or experiments.

e.g

(i) Tossing a Coin

$$X = \begin{cases} 0 & H \\ 1 & T \end{cases}$$

(ii) Rolling a Fair Dice

$$X = \begin{cases} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{cases}$$

- Different Types of Random variables:

1. Discrete Random variables:

e.g Tossing a Coin
Rolling a dice

means only number values no fraction (just whole numbers)

2. **Continuous Random Variable**

e.g

(i) Tommorow how many inches it is going to rain

⇒ inches = 3.1 $X = \{3.1, 2, 4, 0, \ldots\}$

Fractions are allowed.

(ii) Height of the people attending the event tommorow

e.g 160.5cm

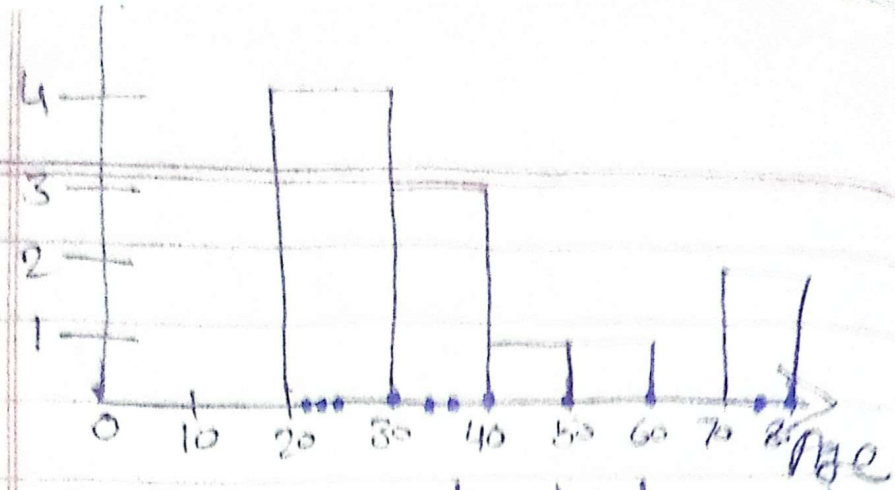
height = $\{150, 160.5, 161, 162.57, \ldots\}$

⇒ **Histograms:**

Let's consider a random variable of age

$X = \{23, 24, 25, 30, 34, 36, 40, 50, 60, 75, 80\}$

Let's create it's histogram



So I have make points according to data. Now let's count

	Points	
20-30	4	
31-40	3	
41-50	1	51-60 = 1
61-70	0	
71-80	2	

⇒ So, histogram actually helps you to find out frequency in the bins which you can change used in distribution.

⇒ We can also derive **Probability Distribution Function or Probability Density Function (PDF)** by smoothening the carve of histogram.

⇒ For smoothening we use Kernel Density Estimation (KDE) which will be discussed later:

⇒ Percentiles and Quartiles:

- Percentage:

{1, 2, 3, 4, 5, 6}

No. of odd numbers: 3

percentage of odd numbers in this group: $\frac{3}{6} \times 100\% = 50\%$

$$\text{Percentage} = \left(\frac{\text{obtained}}{\text{total}} \times 100\right)\%$$

1. Percentiles:

Def:

A percentile is a value below which a certain percentage of ~~observations~~ distributions lie.

- e.g

{2, 2, 3, 4, 5, 5, 6, 7, 8, 8, 8, 9, 9, 10}

percentile of value 9: $\frac{\text{No. of values below 9}}{n}$

n → Size of Data

Percentile of value 9 = $\frac{11}{14} \times 100\%$

Percentile of value 9 = 78.57% of value 9

78.57% → Percentile Ranking

This tells that 78.57% of whole data is distribution below 9.

Generic Formula

Percentile of value x = $\frac{\text{No. of values below } x}{n} \times$

- Now do vice versa get 25% percentile of given distribution i.e value

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

$$\text{value} = \frac{25}{100} \times (14+1)$$

value = 3.75

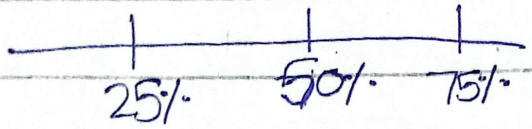
But there is no value 3.75 in distribution so I will pick 3d and 4th and take their averge which is 3.5 so

value = 3.5

Means 25% of distribution is less than 3.5

2) Quartiles:

Let suppose I have calculated percentiles



So,

25% = 1st Quartile

50% = 2nd Quartile

75% = 3rd Quartile

⇒ 5 Number Summary

1) Minimum
2) 1st Quartile (25 percentile)
3) Median
4) 3rd Quartile (75 percentile)
5) Maximum

• e.g:

1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 27

We used 5 Number Summary to detect and Remove outliers from data

• Step1: Specify Lower Fence and higher fence

⇒ Anything out of these fences will be removed as outlier

Lower Fence = Q1 − 1.5(IQR)

Higher Fence = Q3 + 1.5(IQR)

where

IQR (Inter Quartile Range) = Q3 − Q1

$$Q = \left(\frac{\text{Percentile}}{100} \times (n+1)\right)^{th}$$

$$Q3 = \frac{75}{100} \times (20) = 15\text{th position}$$

Q3 = 7

$$Q1 = \frac{25}{100} \times (20) = 5\text{th position}$$

Q1 = 3

$$IQR = Q3 - Q1 = 7 - 3$$

IQR = 4

$$\text{Lower Fence} = Q1 - 1.5(IQR)$$
$$= 3 - 1.5(4)$$

Lower Fence = −3

$$\text{Higher Fence} = Q3 - 1.5(IQR)$$
$$= 7 + 1.5(4)$$

Higher Fence = 13

Remove 27 as it is out of Higher fence so outlier Now

1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9

5-Number Summary

Minimum = 1
1st Quartile = 3
Median = 5
3rd Quartile = 7
Maximum = 9

On the basis of this 5-Number Summary we can draw Box Plot and it indicates 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:

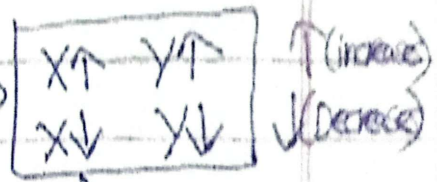
Covariance is measure of how much two random

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 covariance is negative.

- **Example:**

Let we have two random variables in our dataset

X	Y
2	3
4	5
6	7
8	9



Here I will get Positive

Quantify the reationship Covariance b/w X and Y

Now take a real example of house Dataset

Size of house	price
1200	45 Lakhs
1300	50 Lakhs
1500	75 Lakhs

Again

X	Y
7	10
6	12
5	14
4	16

→ $X\uparrow\ Y\downarrow$ / $X\downarrow\ Y\uparrow$

You will get Negative Covariance here

Covariance Formula:

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

If we take cov of a variable withitself it will be it's own variance

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

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

$$Cov(x,x) = \text{variance}(x)$$

Now come to original Formula

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

Where

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

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

n = Sample size

e.g Students

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

X↑ Y↑ and X↓ Y↓

So covariance will be positive

① $\bar{X} = \frac{2+3+1+4+5+6}{5} = \boxed{4}$

② $\bar{Y} = \frac{50+60+70+80+90}{5} = \boxed{70}$

X_i	Y_i	$X_i-\bar{X}$	$Y_i-\bar{Y}$	$(X_i-\bar{X})(Y_i-\bar{Y})$
2	50	−2	−20	40
3	60	−1	−10	10
4	70	0	0	0
5	80	1	10	10
6	90	2	20	40

$\Sigma = 100$

So,

$$\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y}) = 100$$

and $n = 5$

So,

$$Cov(X,Y) = \sum_{i=1}^{n}\frac{(X_i-\bar{X})(Y_i-\bar{Y})}{n-1}$$

$$Cov(X,Y) = \frac{100}{5-1} = \frac{100}{4}$$

$$\boxed{Cov(X,Y) = 25}$$

⇒ The positive covariance indicates the no. of hours studied increase the exam

score also.

- Advantages of Covariance.

① Quantify the relationship between X and Y

- Disadvantage of covariance

Covariance doesn't have a specific limit value

$Cov(x, y) \Rightarrow -\infty$ to ∞

For example A, B and C are fields

$Cov(A, B) = 200$.

$Cov(B, C) = 300$

but we can't say the relation ship between B and C is stronger

For limiting we introduced correlation

② Correlation:

There are two types

(i) Pearson Correlation Coefficient

⇛ Limit [-1 to 1]

- Formula:

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

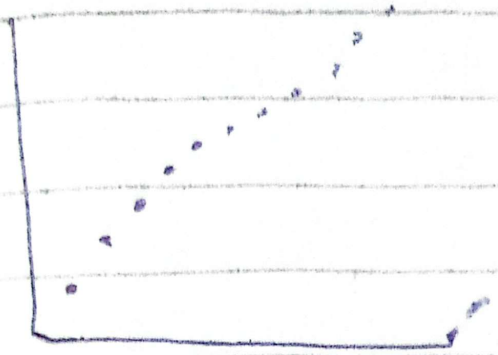
σ ← Standard Deviation

⇛ The more the value towards +1 the more positively correlated X and Y is

⇛ The more the value towards -1 the more negatively correlated x and y is

(ii) Spearman Rank Correlation:

Pearson Correlation is not able to capture the whole information in non-linear data like



Like this where not straight line

so we introduced Spearman Rank Correlation

- **Spearman Rank Correlation Formula**

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

(Rank)

Rank means sort the number and assign Rank minimum $\Rightarrow$ rank = 0

Let

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

$\Rightarrow$ Used for capturing Non-Linear things

How we use Covariance and Correlation in real word Scenarios

⇒ Feature Selection in Feature Engineering