Statistics!

→ Statistics is the science of collecting, organizing and analysing data.

Data Analyst → Report → visualizations

→ Data Analyst will build the reports and present the visualizations to the stake holders. They will take the meaningful decisions.

Data:- facts or pieces of information

Eg:- weights of students in the class

{60kg, 50kg, 45kg, 30kg, ...}

IQ of students in class.

{100, 90, 95, 99, ...}

House price dataset:-

City	Area	no. of Rooms	Price
Bng/R	1000	2	45L
New York	1250	2.5	50L
Mumbai	800	3	60L

After analysing this data. A Data scientist need to build a model which predicts a price for new houses, based on past data.

Types of statistics:-

① Descriptive statistics:-

→ This involves methods for summarizing & organizing data to make it understandable. This type of statistics helps to describe the basic features of the data in a study.

(i) Measure of Central Tendency
[mean, median, mode]

(ii) Measure of dispersion
[variance, standard deviation]

(iii) Data Distribution
- → Histograms
- → Box Plot
- → pie chart
- → PDF, PMF..

(iv) Summary statistics
- → five number summary
- → Q1, Q2, Q3

(2) Inferential statistics

→ This involves methods for making predictions or inferences about a population based on a sample of data. It allows for hypothesis testing estimation and drawing conclusions.

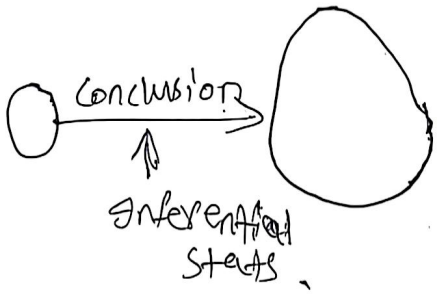
(i) Hypothesis Testing
(ii) P-value
(iii) Confidence Interval
(iv) Statistical Analysis Test
- → Z-test
- → t-Test
- → ANOVA → F-Test
- → CHI square

Eg:- Let say there are 20 statistics class in your college, and you have collected the height of students in the class.

Heights {175cm, 180 cm, 140 cm, 135, 160 cm, 120 cm}

Descriptive question:
"what is avg height of entire classroom"
↑
Measure of central Tendency.

Inferential question:
"Are the height of the sample students in classroom similar to what you expect in the entire college"?



Population

→ It is the entire set of individuals.
→ It includes all members of a defined group.

Characteristics:
(1) Complete set: Contains all the observation of interest.
(2) Parameter: A numerical value summarizing the entire population.
- Population Mean (μ)
- Population Variance (σ^2)

Ex:-
(1) Population in a school study
* All students enroled in a school.
* Determine avg height of student, population mean.

Sample Data:

→ It is a subset of the Population that is used to represent the entire group.

characteristics:

(1) Subset: Represents a portion of the Population.

(2) Statistic: A numerical value summarizing the sample data.
[Sample mean, sample variance]

(3) Random Sampling:
Samples should be randomly selected to avoid bias.

Types of Sampling Techniques:

1. Probability Sampling
2. Non Probability "

(1) Probability Sampling:

(a) Simple Random Sampling:
Every member of the population has an equal chance of being selected.
Eg:- selecting people randomly.

(b) Systematic Sampling:
Select every nth member of the population after a random starting point.

Eg:- Airport → credit card
↳ 5th person, 10th person, 15th person ...
→ Feedback survey → select every 11th member

(c) Stratified Sampling:
Divide the population into strata (groups) based on specific characteristics & then randomly sampling from each strata.

Eg:- Divide employees by department and then randomly select a proportional number from each department to form a survey sample.

Eg:- Age → <12 12-18 >18
[Politics]

① Cluster Sampling:-

→ Divide the population into clusters, randomly selecting clusters, then sampling all the members from the selected clusters.

Eg:- Randomly selecting several schools from a district, & surveying all teachers within those schools.

② Non Probability Sampling:-

Select individuals who are easiest to reach.

Eg:- Surveying people at mall.

ⓐ Convenience Sampling:-

→ selecting individual who are easiest to reach.

ⓑ Judgmental (purposive Sampling)

→ Select individual based on the researcher's Judgement

Eg:- choose experts in a field to participate in Data science.

ⓒ Snowball Sampling:-

Existing study subjects recruit future subjects from among their acquaintances.

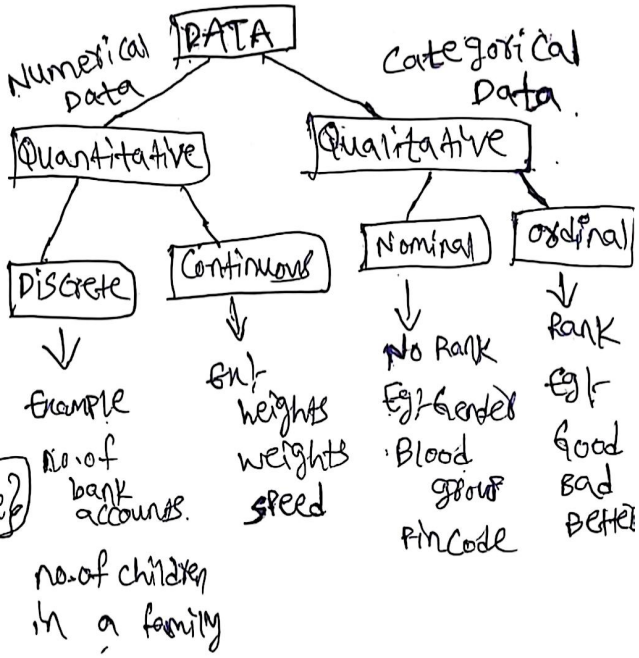
Eg:- survey members of a rare disease.

ⓓ Quota Sampling:-

Age, group, gender, caste.

* Selecting the sampling technique depends on use cases.

Types of DATA:-



Scales of Measurement of Data:

→ This describes the nature of information within the values assigned to variables.

4 Primary scales of measurement:

1. Nominal Scale
2. Ordinal Scale
3. Interval "
4. Ratio "

① Nominal scale:-

→ This scale classifies data into distinct categories that do not have an intrinsic order.

→ Qualitative/Categorical data.

Characteristics:

(i) Data is categorized based on labels, names or qualities.

(ii) Categories are mutually exclusive.

Ex:- Gender
male
female.

(iii) No logical order among categories. [No Rank]

② Ordinal Scale:

→ This scale classifies the data into categories that can be ranked or ordered.

Characteristics:

(i) Data is categorized & Ranked in a specific order.

(ii) The interval b/w ranks are not necessarily equal.

Examples:

Education level	⇒ Feature Engineering ⇒ Ranks
High School	1
Bachelor	2
Masters	3
Doctorate	4

Customer feedback	
Satisfied	1
Very Satisfied	2
Not Satisfied	0

Socio Economic Status	
Low	0
medium	1
High	2

③ Interval Scale:-

→ The interval scale not only categorizes & orders but also specify the exact difference between intervals. It lacks a true zero point.

Characteristics:

(i) Data is ordered with consistent interval b/w values.

(ii) Allows for meaningful comparison of differences.

(iii) No true zero point.

④ Ratio scale:

→ The order matters.

→ Differences are measurable.

→ Contains a 0 starting point.

→ Ratio can be measured.

Eg:- marks of students

0, 30, 40, 90.

5

0

0

1

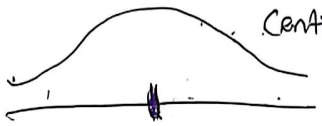
-5

Measure of Central Tendency:

→ Measures of central tendency are statistical metrics that describe the center point or typical value of a dataset. They provide a single value that summarizes a set of data by identifying the central position within the dataset.

Ex:- Age = [24, 32, 12, 48, 16, 20]

Central Point
Central Position.



① Mean:

→ Mean is the sum of all values divided by the no. of values.

Population Mean (μ)

Population (N)

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

Here x is a random variable

$x = \{5, 8, 12, 15, 20\}$

$N = 5$

$$\mu = \frac{5+8+12+15+20}{5} = 12$$

Sample mean ($\bar{x}$)

Sample (n)

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

n → sample size.

Characteristics :

→ mean is affected by extreme outliers

→ Used for interval and Ratio Data

② Median :

→ The median is the middle value in a dataset after arranging the values in a ascending or descending order.

Ex:- $X = \{1, 2, 3, 4, 5\}$

no. of elements = 5

5 is odd

median = 3

$X = \{3, 4, 1, 5, 2, 100\}$

↳ $\{1, 2, 3, 4, 5, 100\}$

no. of elements = 6

6 is even

$\text{Median} = \frac{3+4}{2} = 3.5$

Characteristics :

* Not affected by outliers.
* Used for ordinal, interval and ratio data.

③ Mode :

→ Most frequently appeared value in a dataset.

Ex:- 2, 4, 4, 6, 7, 7, 7, 9

mode = 7 (most frequent value)

3, 5, 5, 6, 6, 8

Mode = 5, 6 (bimodal)

Characteristics.

* This is not affected by extreme values.
* Used for Nominal, ordinal, interval and ratio data.

Choosing the Appropriate Measure

1. Mean : Best used when data is symmetrically distributed without outliers.
2. Median : Best used when data is skewed or contains outliers. Provides the middle value, which better represents the center of a skewed dataset.
3. Mode : Best used for categorical data to identify the most common category. Also useful for identifying the most frequent value in ordinal, interval or ratio data.

Measures of Dispersion:

→ This describe the spread or variability of a dataset. They indicate how much the values in a dataset differ from the central tendency.

Common measures of Dispersion

(1) Range
(2) Variance
(3) standard deviation
(4) Inter quartile Range (IQR)

(1) Range:

→ It is the difference b/w the maximum & minimum value in a dataset.

Range = Maximum value − minimum value

Ex: Ages {14, 13, 10, 20, 25, 75, 15}

Range = 75 − 10 = 65

Characteristics:

→ simple to calculate
→ sensitive to outliers
→ Rough measure of dispersion.

(2) Variance:

↔ It measures the average squared deviation of each value from the mean. It provides a sense of how much the values in a dataset vary.

Population Variance

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

Sample variance:

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

Ex:- size of flower petals

{5, 8, 12, 15, 20} ⇒ variance of this distribution.

$N = 5$

$$\mu = \frac{5+8+12+15+20}{5} = 12$$

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

$$\text{Variance} = \frac{(5-12)^2 + (8-12)^2 + \ldots + (20-12)^2}{5}$$

Variance = 27.6

Characteristics:

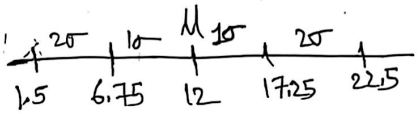
→ Provide a precise measure of variability.
→ units are squared of the original data units.
→ More sensitive to outliers than the Range.

(3) Standard Deviation:

→ It is the square root of the variance.

$$\sigma = \sqrt{27.6} \approx 5.25$$

→ By using standard deviation we can say how far the data point is from mean



{5, 8, 12, 15, 20}

Where 20 will locate?

→ It is 2σ (two standard deviations) to the right mea

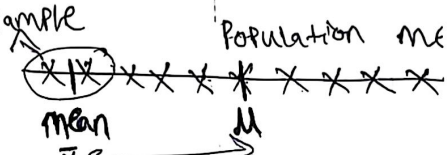
Characteristics:

→ Provides a clear measure spread in the same uni as the data.
→ sensitive to outliers.

Why Sample Variance is divided by n−1?

Ex:-

Ages = {- - - - - -



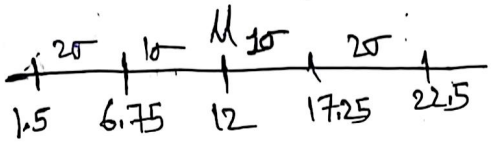
→ When you take Sa the mean of Sam

(3) Standard Deviation :

→ It is the square root of the variance.

$$\sigma = \sqrt{27.6} \approx 5.25$$

→ By using standard deviation we can say how far the data point is from mean.



{5, 8, 12, 15, 20}

Where 20 will locate?

→ It is 2σ (two standard deviations) to the right of mean.

Characteristics :

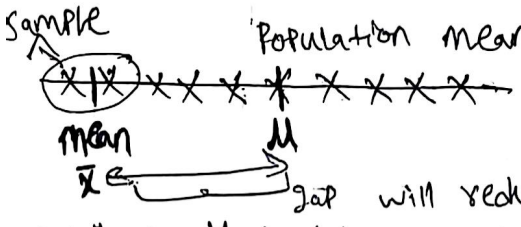
→ Provides a clear measure of spread in the same units as the data.

→ Sensitive to outliers.

Why Sample Variance is divided by n-1 ?

Ex :-

Ages = {- - - - - - - - - - }



→ When you take, sample, the mean of sample is at more distance compared to population mean.

$$\bar{x} < < < M$$

$$s^2 < < < \sigma^2$$

→ If we divide sample mean with 'n', there is so much difference b/w sample mean & population mean.

→ To do correction & to decrease difference some what, bessel's correction is introduced.

'n-1'

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

↓

understanding the true population mean.

Degrees of freedom = n-1

Random Variables :

X → function → Values

⇓

Process or Experiments

$y = 5x + 2$.

x = 1, 2, 3 ... are random numbers variables

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

$X = \begin{cases} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{cases}$ Rolling a fair dice

Random Variables

Discrete Random
Eg:- Tossing a coin
Rolling a dice

Continuous Random variable
Eg:- How many inches of Rain
$[0, 1.1, 5.5 \ldots]$

**

→ Random variable is a function which defines its values through some process or experiments.

Percentiles and Quartiles

Percentage: $\{1, 2, 3, 4, 5, 6\}$

no. of odd numbers = 3

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

Percentiles:-

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

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

$$\text{Percentile of value } x = \frac{\#\text{ of values below } x}{n} \times 100$$

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

$= 78.57\%$ of value 9

Percentile Ranking.

→ This says that 78.57% of entire distribution is less than 9.

→ If value want to find?

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

$= \frac{25}{100} \times (15)$

Value $= 3.75$

→ 25% percentile values are below $3.5 = \frac{3+4}{2}$

→ The values that are less than 3.5 are fall in 25 percentile.

② Quartiles

25% = 1st quartile
50% = 2nd quartile
75% = 3rd quartile

Q_3 = 75 percentile $= \frac{75}{100} * (19+1)$

$= 15^{th}$ value

$Q_3 = 7$

$IQR = Q_3 - Q_1 = 7 - 3 = 4$

lower fence $= Q_1 - 1.5(IQR)$

$= 3 - 6 = -3$

upper fence $= Q_3 + 1.5(IQR)$

$= 7 + 1.5(4)$

$= 7 + 6$

$= 13$

$[-3, 13]$

outlier = 29 in the above data

5 Number Summary and Boxplot

[To find outliers]

(i) ~~Maximum~~ minimum
(ii) First quartile (25 percentile) Q1
(iii) median
(iv) Third quartile (75 percentile) Q3
(v) Maximum.

Removing the outlier?

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

Q_1 → 7 ← Q_3 ; 29 → outlier.

[Lower fence ⟷ Higher fence]

→ Any value that is lower than lower fence & any value that is higher than higher fence is determined as outlier.

Lower fence $= Q_1 - 1.5(IQR)$

Higher fence $= Q_3 + 1.5(IQR)$

$IQR = Q_3 - Q_1$ = Inter quartile Range

$Q_1 = 25 \text{ percentile} = \frac{25}{100} * (19+1)$

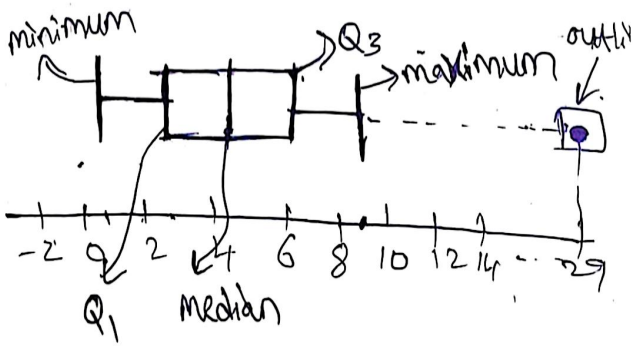
$= 5^{th}$ value

$Q_1 = 3$

5 number summary:

Minimum = 1
$Q_1 = 3$
median = 5
$Q_3 = 7$
maximum = 9

Box plot:



Histogram and Skewness:

→ A histogram is a graphical representation of the distribution of numerical data.

→ It is an estimate of the probability distribution of a continuous variable and is used to visualize the shape, central tendency, and variability of a dataset.

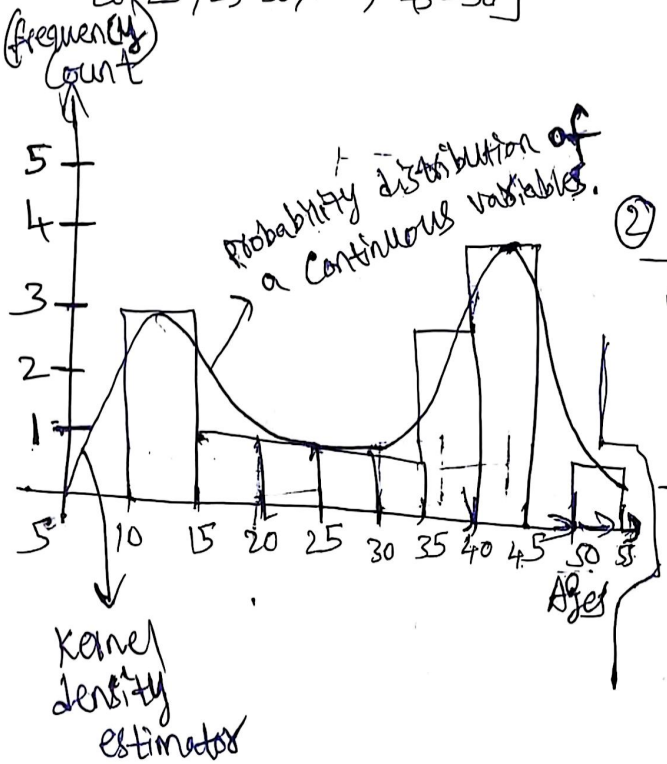
Ages = {11, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 41, 42, 43, 50}

0 – 50 ↓ distribution of numerical data

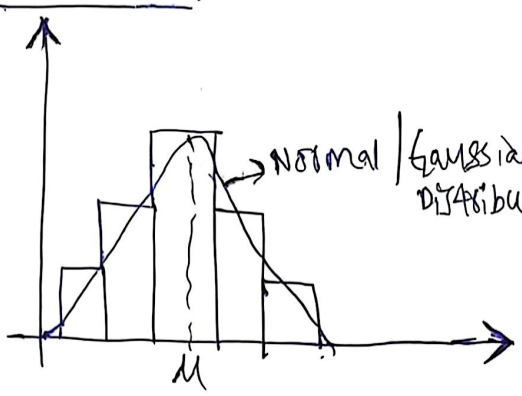
① No. of bins = 10

$\frac{50}{10} = 5 \Rightarrow$ bin size

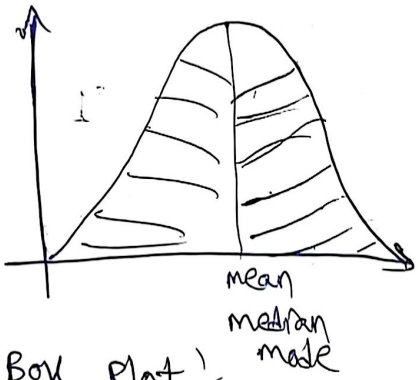
Bins → [0–5, 5–10, 10–15, 15–20, 20–25, 25–30, --> 45–50]



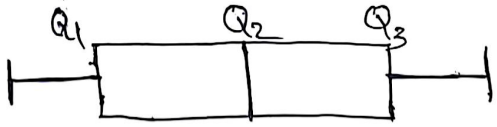
Skewness:



⇒ Symmetric Distribution ⇒ No skewness



Box Plot:

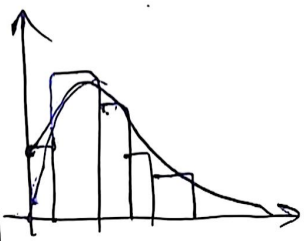


$Q_3 - Q_2 \approx Q_2 - Q_1$

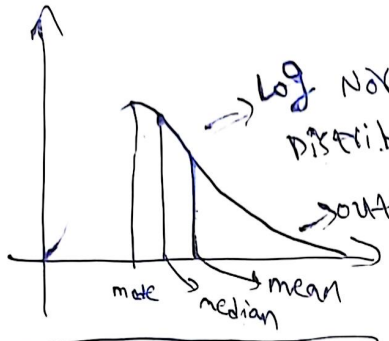
→ The mean, median and mode are all perfectly at the center.

Mean = median = mode

② Right Skewed:

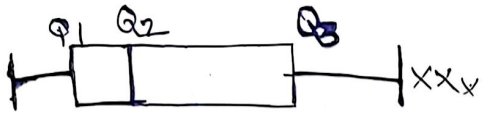


Positive skewed



mean ≥ median ≥ mode

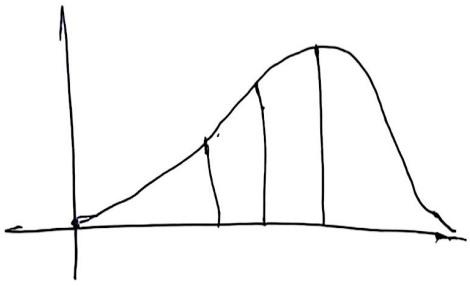
Box plot :-



→ mean will shift to left.

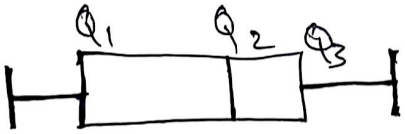
median $Q_3 - Q_2 \geq Q_2 - Q_1$

(3) Left skewed distribution :



mean ≤ median ≤ mode

Negative skewed.



$Q_2 - Q_1 \geq Q_3 - Q_2$

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 :

→ It is a measure of how much two random variables change together.

→ If the variables tend to increase & decrease together, the covariance is positive. If one tends to increase when the other decreases, the covariance is negative.

X	Y
2	3
4	5
6	7
8	9

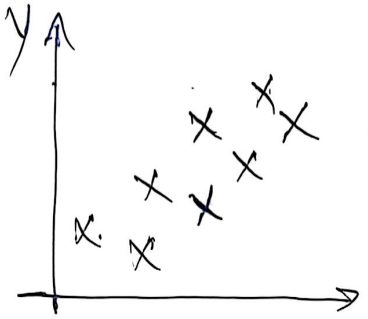
Quantify the Relationship b/w X & Y

If

X↑	Y↑
X↓	Y↑
X↑	Y↓
X↓	Y↓

Dataset

Size of House	Price
1200	45 lakhs
1300	50 lakhs
1500	75 lakhs



In this case

$x\uparrow\ y\uparrow$

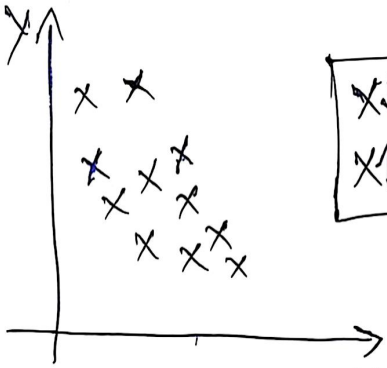
$x\downarrow\ y\downarrow$

$\Downarrow$

+ve Covariance

+ve value

Ex 2:-



$x\downarrow y\uparrow$

$x\uparrow y\downarrow$

$\Downarrow$

-ve covariance

-ve value.

X	Y
7	10
6	12
5	14
4	16

① Covariance

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

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

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

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

$x_i \rightarrow$ Data Points of random variable X

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

$y_i \rightarrow$ Data Points of random variable Y

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

Example:-

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

Here $x\uparrow y\uparrow$, $x\downarrow y\downarrow$ $\Rightarrow$ +ve covariance

we will see with formula.

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

① $\bar{x} = \frac{2+3+4+5+6}{5} = 4$

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

$$Cov(X,Y) = \frac{(2-4)(50-70)+(3-4)(60-70)+\cdots+(6-7)(90-70)}{5-1}$$

$Cov(X,Y) \simeq 20$ // → +'ve value

+'ve covariance.

→ The positive covariance indicates the no. of hours studied increased the exam score also..

Advantages!

→ Quantify the relationship between X & Y.

Disadvantage!

→ Covariance does not have a specific limit value.

2) Correlation → Pearson Correlation Coefficient
→ Spearman Rank Correlation.

1) Pearson Correlation Coefficient!

→ value in b/w [-1, 1]

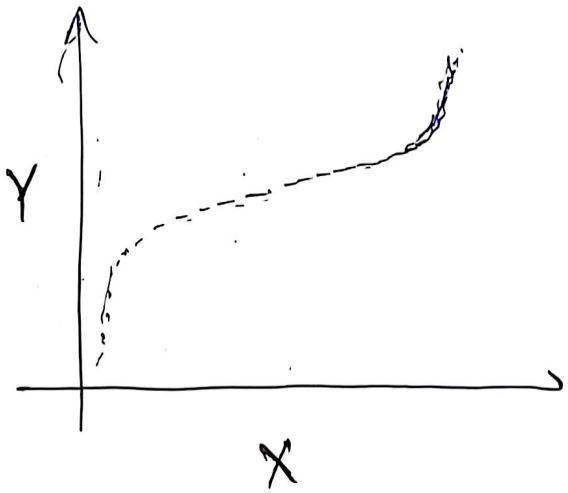
$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

* The more the value towards +1, the more the correlated X & Y is.

* The more the value towards -1, the more negative correlated X & Y is.

ii) Spearman Rank Correlation.

→ This is use for non-linear data.



Spearman correlation = 1

Pearson correlation = 0.88

→ Here, a spearman correlation of 1 results when the two variables being compared are monotonically related, even if their relationship is not linear.

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

Distributions:

Probability Distribution functions

→ It describes how the probabilities are distributed over the values of a random variable.

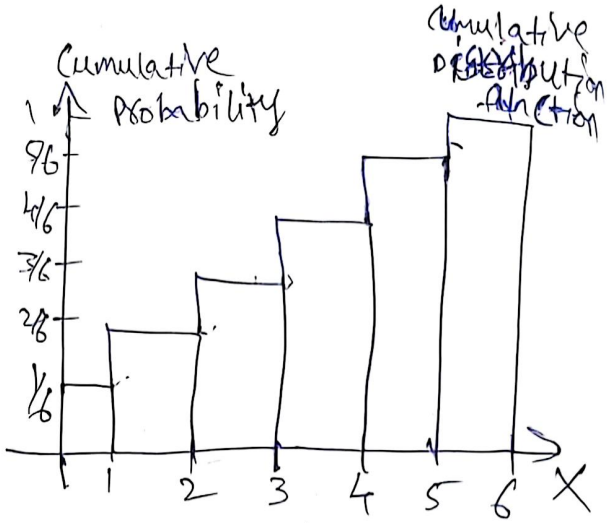
Probability Mass function (PMF)

→ used for discrete random variable.

Probability Density function (PDF)

→ used for continuous random variable.

Converting PMF to CDF



Cumulative Distribution function (cdf)

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

$= 1/6 + 1/6 = 2/6 = 1/3$

$Pr(X \leq 6) = Pr(X=1) + Pr(X=2) + \dots + Pr(X=6)$

$Pr(X \leq 6) = 1$

PMF

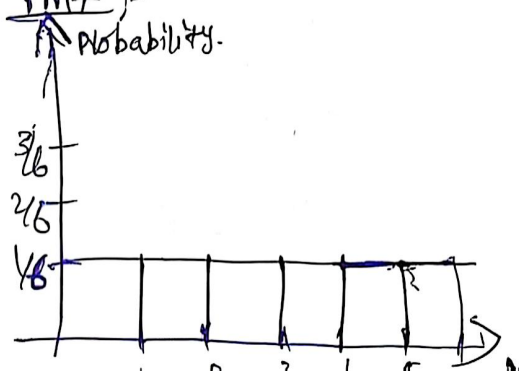
Ex 1:

Rolling a dice.

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

$Pr(1) = Pr(2) = \dots = Pr(6) = 1/6$

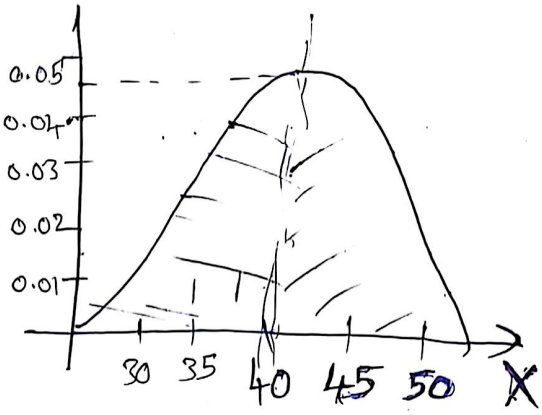
PMF :-



PDF

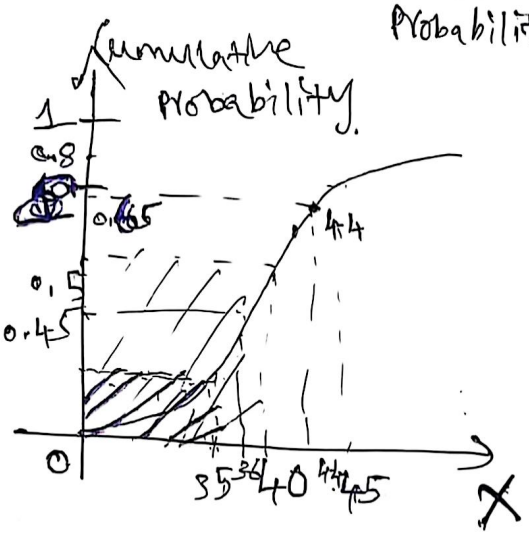
Ex :-

Ages = {- - -}



Converting PDF to CDF

Cumulative Probability



$Pr(X \leq 40) = ?$

→ Area under the curve

$Pr(X \leq 35) = ?$

→ For getting probability density from cumulative density function we use **slope**.

$$\text{Slope} = \left\{\frac{0.65 - 0.45}{44 - 38}\right\} = 0.05$$

↓ Probability Density.

$$\text{Slope} = \left\{\frac{x_2 - x_1}{y_2 - y_1}\right\} \Rightarrow \text{Gradient.}$$

→ Probability density is the Gradient of cumulative Density function.

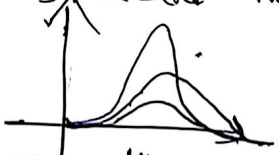
1. Normal Distribution (Gaussian Distribution)

→ Symmetrical, bell-shaped curve characterised by its mean (center) & standard deviation (spread)

→ Emprical rule (68-95-99.7)

→ If a normal distribution has a mean of 0 & a standard deviation 1, it is referred to as a standard normal distribution.

→ Z-score can be used to standardize values from any normal distribution to the standard normal distribution



(2) Bernoulli Distribution

→ It is a simple & fundamental discrete probability distribution that models a random experiment with two possible outcomes

"success" → 1

"failure" → 0

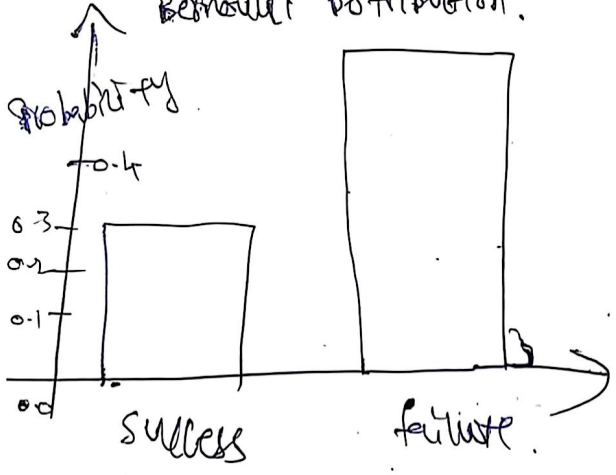
Ex:- flipping a coin (heads or tails)

Ext o/p of a single question
Yes / no

mean (expected value) $E[X] = p$

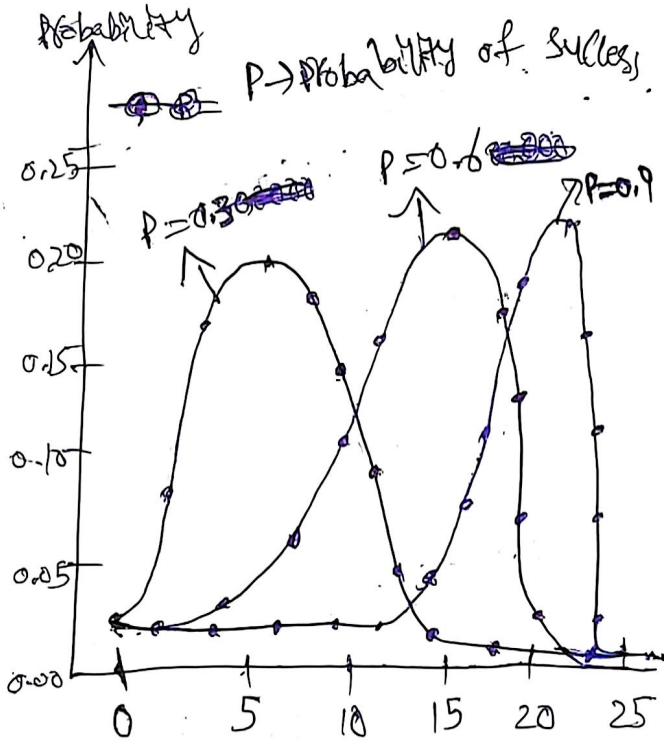
Variance $Var[X] = p(1-p)$

Bernoulli distribution.



Prob of getting success $= 0.3$
& in failure $= 0.7$

Ex:-
→ no. of heads in a fixed no. of coin flips.
→ defective / non-defective items in a production process.



Binomial Distribution:

→ It is a discrete probability distribution.

→ It describes a no. of successes in a fixed no. of independent and identical Bernoulli trials.

mean $\Rightarrow E[X] = np$

$Var[X] = np(1-p)$

Poisson distribution:

→ It is a discrete probability distribution.

→ It models the no. of events occurring in a fixed interval of time or space.

→ It is particularly useful for situations where events are rare & random.

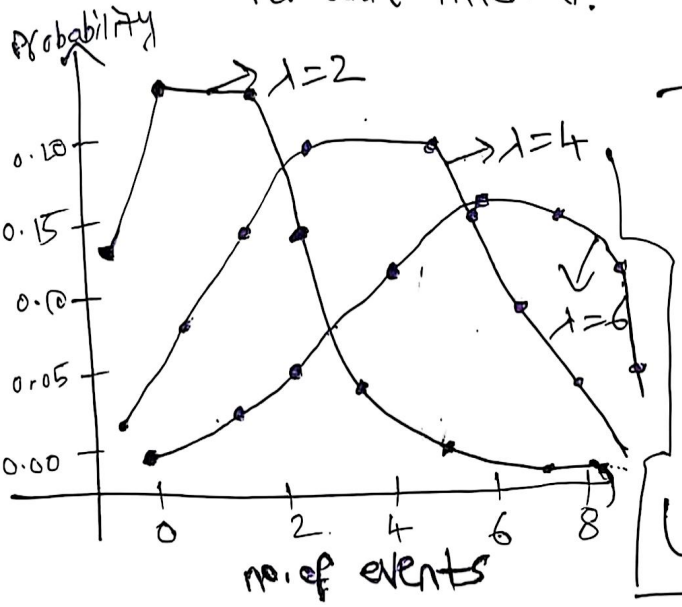
Ex:- no. of phone calls at a call center in a given minute.

→ no. of arrivals at a service point.

$E[X] = \lambda$

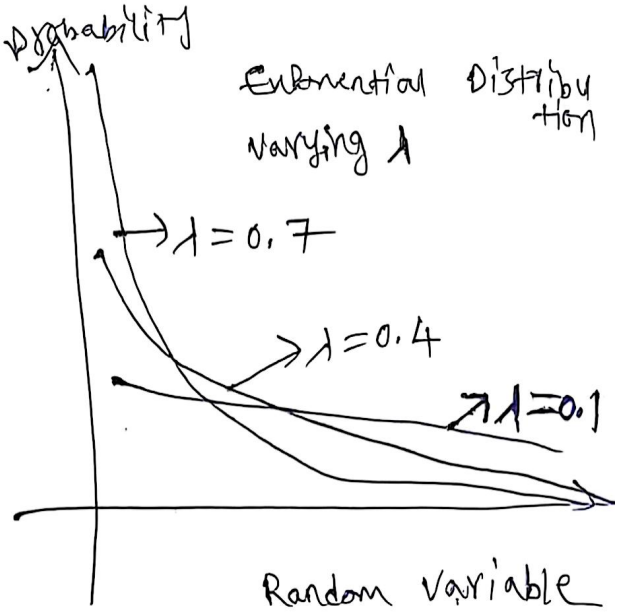
$Var[X] = \lambda$

λ → Avg rate of events per unit interval.



Poisson distribution varying λ.

$Var[T] = 1/\lambda^2$



Exponential Distribution varying λ

Exponential Distribution:

→ It is a continuous probability distribution that models the time b/w events in a Poisson process.

Ex:- time b/w arrivals at a service point.

Ex:- → Time until a system failure.

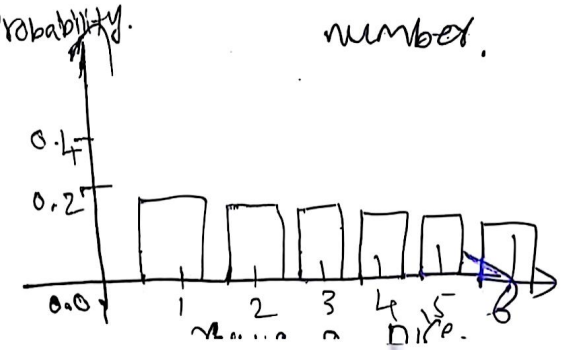
mean $E[T] = 1/\lambda$

Uniform Distribution:-

→ All values in distribution have an equal probability of occurring.

→ There are continuous & discrete versions of the uniform Distribution.

Ex:- generating a random number.

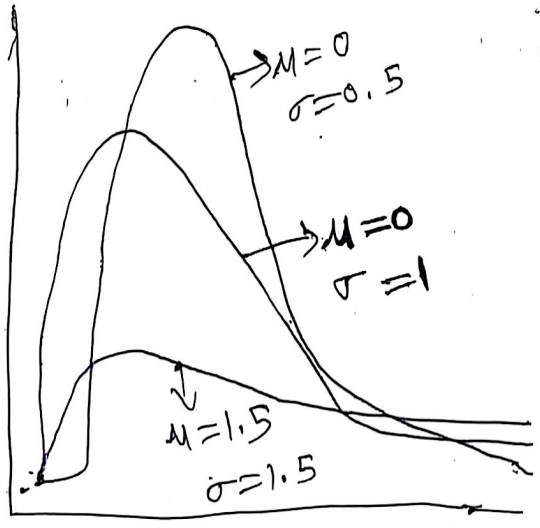


Log-normal Distribution :

→ It is a continuous probability distribution of a random variable whose logarithm is normally distributed.

→ ~~If you take a logarithm for a Gaussian variable~~

→ By taking gaussian variable and exponentiate it, you get the log normal distribution.



Ex:- The length of Comments Posted in Internet discussion forums follows a log normal distribution.

Power-Law Distribution :

→ It describes a functional relationship between two variables, where one variable is proportional to a power of the other.

$$f(x) = Kx^{\alpha}$$

x, y → variables of interest
'α' → law exponent
'K' → constant

Ex:- If side of a square increased by two units, the area increases by four units.

Pareto Distribution :

→ A specific type of distribution that follows power law is called Pareto Distribution.

→ Pareto principle states that 80% of the

effects come from 20% of the cause.

→ for Ex:- 80% of the world's wealth is earned by 20% of the people.

→ The idea sometimes expressed as "80-20" rule".

→ The CLT says that the sampling distribution of mean will always be normally distributed, as long as the sample size is large enough.

→ Regardless of whether the population has a normal, Poisson, binomial or any other distribution, the sampling distribution of mean will be normal.

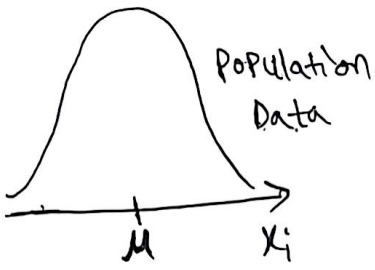
Note!-

Need to study formulas (derivations) from Krish naik course.

Central Limit Theorem! (CLT)

→ The Central limit theorem relies on the concept of a sampling distribution, which is the probability distribution of a statistic for a large no. of samples taken from a population.

① $X \approx N(\mu, \sigma)$



n = sample size ⇒ any value

$S_1 = \{x_1, x_2, \ldots, x_n\} = \bar{x}_1$

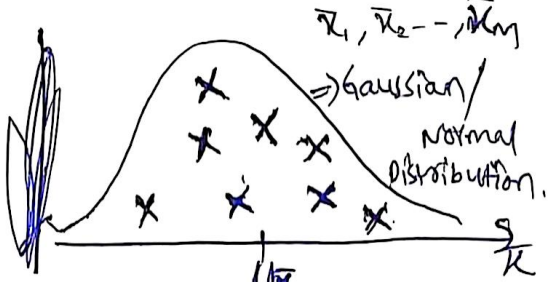
$S_2 = \{x_2, x_3 \ldots, x_n\} = \bar{x}_2$

$S_3 = \bar{x}_3$

$S_m = \bar{x}_m$

Sampling distribution of means

$\bar{x}_1, \bar{x}_2 \ldots, \bar{x}_m$



② $X \not\sim N(\mu, \sigma)$ → not a normal distribution.

↓ random variable

Here sample size must be $n \geq 30$

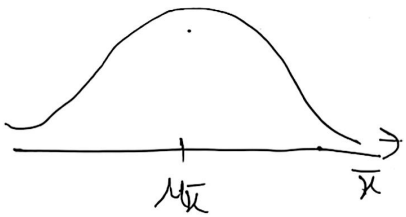
$S_1 = \{x_1, \ldots, x_{30}\} = \bar{x}_1$

$S_2 = \{x_1, \ldots, x_{30}\} = \bar{x}_2$

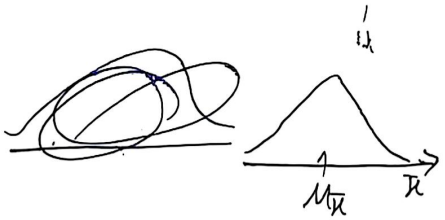
$S_3 = \ldots = \bar{x}_3$

$\vdots$

$S_m = \ldots = \bar{x}_m$



Sampling distribution of mean



σ = Population std Deviation

μ = Population mean

n = sample size.

$$\bar{X} \approx N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

After distribution

✳✳ → for large sample sizes, the sampling distribution of means will approximate to normal distribution even if the population distribution is not normal.

A Hypothesis and Hypothesis testing Mechanism:

① Null Hypothesis (H_0)
↳ It is the assumption we are beginning with

② Alternate Hypothesis (H_1)
↳ opposite of Null Hypothesis.

③ Experiments → statistical Analysis {p-value, significance value}

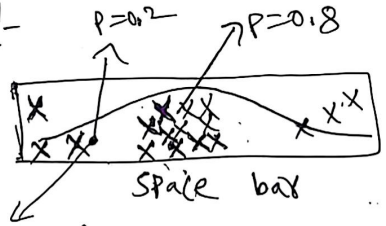
④ Accept the Null Hypothesis
or Reject " " "

P Value:

→ The p value is a number, calculated from a statistical test, that describes how likely you a to have found a particula set of observations if th null hypothesis were true.

→ P values are used in hypothesis testing to hel decide whether to reject the null hypothesis.

Ex:-



out of 100 touches, we around 20 times in this

Hypothesis Testing:

Eg: Coin is fair or
{100 tim

① Null Hypothesis:

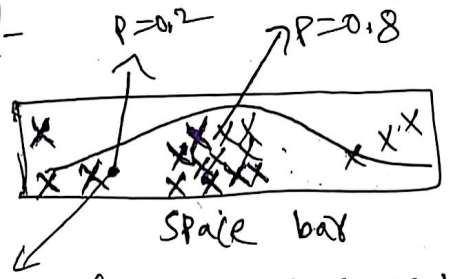
H_0 → Coin is fair

P Value :

→ The p value is a number, calculated from a statistical test, that describes how likely you are to have found a particular set of observations if the null hypothesis were true.

→ P values are used in hypothesis testing to help decide whether to reject the null hypothesis.

Ex :-



out of 100 touches, we touch around 20 times in this region.

Hypothesis Testing :

Eg: Coin is fair or not {100 times}

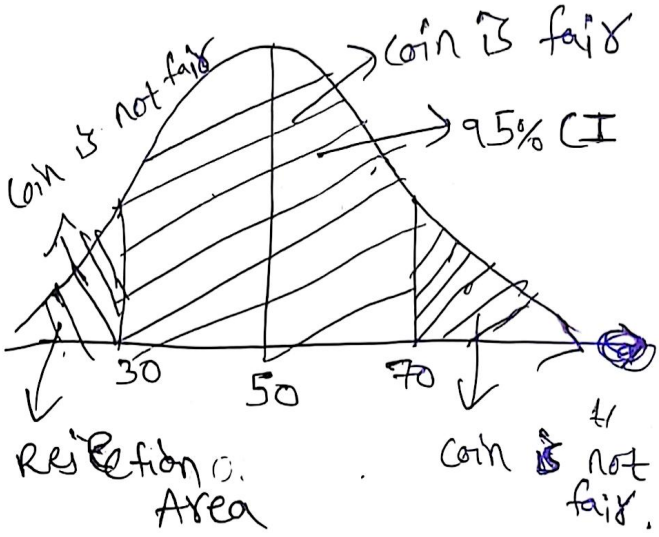
① Null Hypothesis :

H_0 → Coin is fair

② Alternate Hypothesis

H_1 → Coin is not fair

③ Experiment :- {100 times}



④ significance value :

$\alpha = 0.05$

$CI = 1 - 0.05 = 0.95$

⑤ Conclusion

p < significance value

Reject the null Hypothesis.

else

fail to reject the null Hypothesis.

Hypothesis Testing and Statistical Analysis:

① Z Test
② t Test
} Average (mean) ⇒ Z-table / t-table ⇒ Z score & P value

③ CHI SQUARE ⇒ Categorical Data
④ ANNOVA ⇒ Variance.

① Z-Test

→ we Z-Test when you know (i) population standard deviation (ii) $n \geq 30$

→ The average heights of all residents in a city is 168cm. with a $\sigma = 3.9$. A doctor believes the mean to be different. He measured the height of 36 individuals and found the average height to be 169.5 cm.

(a) State null & Alternate Hypothesis
(b) At a 95% Confidence level, is there enough evidence to reject the null hypothesis.

Ans)

$\mu = 168cm \quad \sigma = 3.9$

$n = 36 \quad \bar{x} = 169.5$

$CI = 0.95 \quad \alpha = 1 - 0.95$

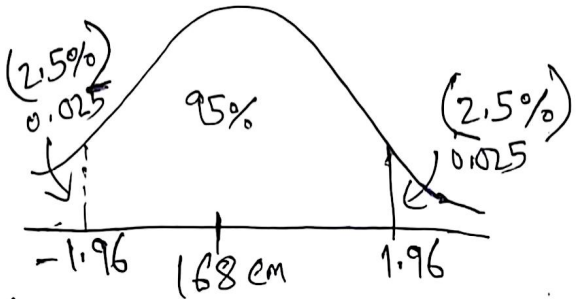
(i) Null Hypothesis

$H_0 = \mu = 168cms$

(ii) Alternate Hypothesis

$H_1 = \mu \neq 168cms$

(iii) Based on C.I. we will draw decision boundary.



* This is 2 Tail test, because we are trying to find, the avg height is greater than or less than 168cm we are checking in both the directions. so it is 2-Tail.

Area under curve

$1 - 0.025 = 0.9750$

check Z-table & Compute Z-score.

for 0.9750 the Z-score is 1.96

if Z is less than -1.96 or greater than +1.96, Reject the null Hypothesis.

Z-test

$$Z_d = \frac{\bar{x} - \mu}{3.9/\sqrt{n}}$$

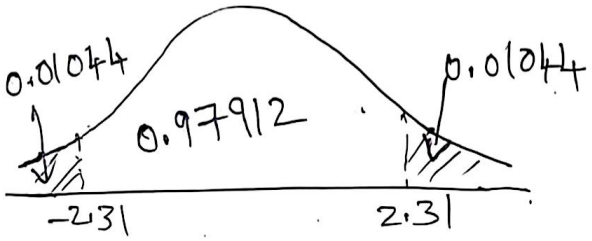
$$= \frac{169.5 - 168}{3.9/\sqrt{36}}$$

$$\boxed{Z_d = 2.31}$$

Conclusion :-

$2.31 > 1.96$, Reject null Hypothesis

Based on calculation.



from Z-table

The area corresponds to '2.31' is 0.98956

Getting the tail Area

$1 - 0.98956$

$= 0.01044$

$P\text{value} = 0.01044 + 0.01044$

$= 0.02088$

$P < 0.05$

$0.02088 < 0.05$

$\Rightarrow$ Reject the null Hypothesis.

Final Conclusion :-

Avg Height $\neq$ 168cm's

$\rightarrow$ The avg height seems to be increasing & based on sample heights.

(2) A factory manufactures bulbs with a avg warranty of 5 years with standard deviation of 0.50. A worker believes that the bulb will malfunction is less than 5 years. He tests a sample of 40 bulbs and find the average time to be 4.8 years

(a) State null and Alternate Hypothesis

(b) At a 2% significance level, is there enough evidence to support that the warranty should be revised?

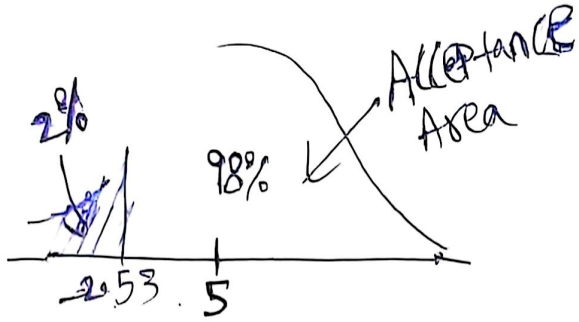
Ans:- $\mu = 5$ $\sigma = 0.5$ $n = 40$ $\bar{x} = 4.8$

(a) Null Hypothesis $H_0 = \mu = 5$

Alternate " $H_1 = \mu < 5$

{1 Tail Test}

(b) Decision Boundary



Z-Test

$$Z_d = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{4.8 - 5}{0.5/\sqrt{40}}$$

$$= -2.53$$

From Z-table

Area under curve with Z-score -2.53 is 0.0570

P-value = 0.0570

Compare P-Value with α

$0.0570 < 0.02 \Rightarrow$ False.

we accept the null hypothesis

we fail to reject the null hypothesis.

Student t distribution

→ In Z-stats when we perform any analysis using Z-score we require σ (Population standard deviation)

↳ is already known.

→ How do we perform any analysis when we don't know the Population standard deviation?

⇓

for this scenario we use Student's t distribution.

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

s = sample standard deviation,

T-stats : T-test

↳ One Sample t-test

(1) In the Population the average IQ is 100. A team of researchers want to test a new medication to see if it has either a positive or negative effect on intelligence, or no effect at all. A sample of 30 participants who have taken the medication has

a mean of 140 with a standard deviation of 20. Did the medication affect intelligence.

Ans)

$\mu = 100 \quad n = 30$

$\bar{x} = 140 \quad S = 20$

$CI = 95\%$

$\alpha = 0.05$

(i) null Hypothesis $H_0: \mu = 100$

Alternate $H_1: \mu \neq 100$

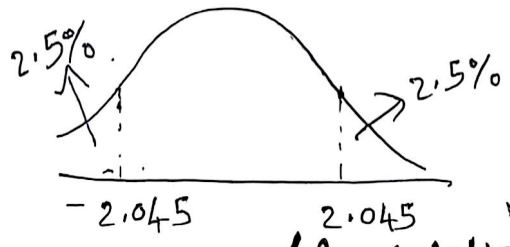
{Two tail test}

(ii) $\alpha = 0.05$

(iii) Degree of freedom

$df = n - 1 = 30 - 1 = 29$

(iv) Decision Rule)



(from t-table)

→ if t-test is less than -2.045 or greater than 2.045, reject the null hypothesis.

(v) Calculate Test statistic)

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{140 - 100}{20/\sqrt{30}} = 10.96$$

$t = 10.96$

$10.96 > 2.045$

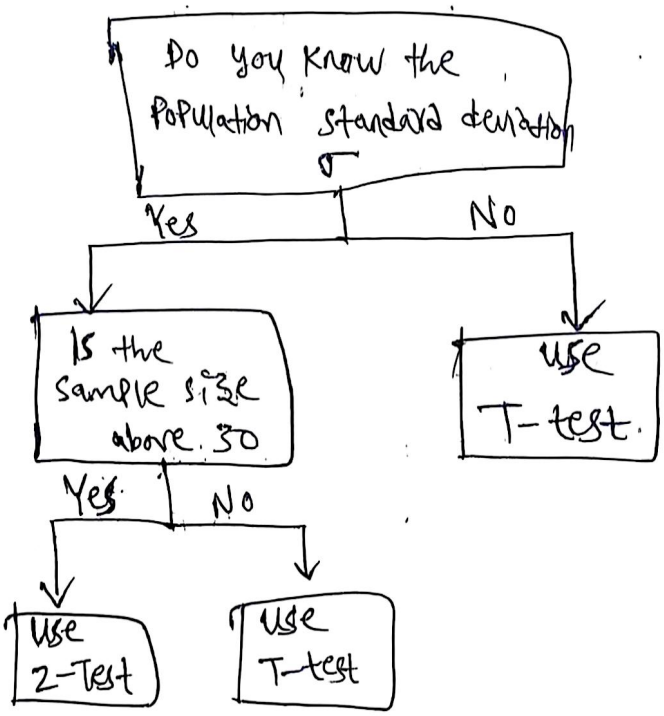
Reject null Hypothesis

Conclusion

* Medication used has affected the intelligence

medication has increased the intelligence

When to use T-tests vs Z-tests)



Type 1 and Type 2 Errors / Confidence Interval & Margin of Error

Reality : Null Hypothesis is True
or " " " false

Decision : Null Hypothesis is True
or " " " False

Outcome 1 : we reject the null Hypothesis when in reality it is false → Good.

Outcome 2 :-
we reject the null Hypothesis when in reality it is True
↳ Type 1 Error.

Outcome 3 :-
we retain the null Hypothesis, when in reality it is False
↳ Type 2 Error.

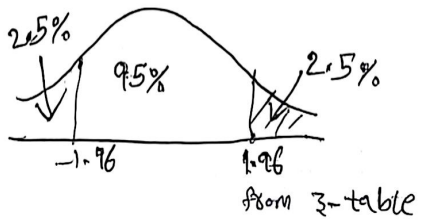
Outcome 4 :-
we retain the null Hypothesis when in reality it is TRUE → Good.

Point estimate ± Margin of Error

If it is z-Test
$\Rightarrow \bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

Eg: On the verbal section of CAT exam, the standard deviation is known to be 100. A sample of 30 test takers has a mean of 520. Construct 95% CI about the mean.

$\alpha = 0.05$ $\quad \bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$



Lower C.I $= 520 - (1.96)\frac{100}{\sqrt{25}} = 480.8$

Higher C.I $= 520 + (1.96)\frac{100}{\sqrt{25}} = 559.2$

Conclusion :-
I am 95% confident about the mean CAT score is between 480.8 and 559.2

CHI SQUARE TEST

→ The Chi square Test for Goodness of fit test claims about Population Proportions.

→ It is a non-Parametric test that is performed on categorical [ordinal & nominal] data.

There is a population of male who likes different color bikes.

		Theory	Sample
Yellow	Bike	1/3	22
Red	"	1/3	17
Orange	"	1/3	59

↙ Theory categorical distribution

↓ observed. Categorical distribution.

Goodness of fit test:

→ In a science class of 75 students, 11 are left Handed. Does this class fit the theory that 12% of people are left handed

	O	E
left handed	11	9 → $\frac{12}{100} \times 75$
Right handed	64	66

Ex:-

In 2010 Census of the city, the weight of the individuals in a small city were found to be the following.

<50kg	50-75	>75
20%	30%	50%

In 2020, weight of $n=500$ individuals were sampled. Below are the results.

<50	50-75	>75
140	160	200

Using $\alpha = 0.05$, would you Conclude the Population difference of weights has changed in the last 10 years?

Ans:-

2010 Expected

<50kg	50-75	>75
20%	30%	50%

2020 n=500 observed

<50	50-75	>75
140	160	200

Expected

<50	50-75	>75
0.2X500 =100	0.3X500 =150	0.5X500 =250

(i) Null Hypothesis : H_0

The data meets the expectation.

Alternate Hypothesis : H_1

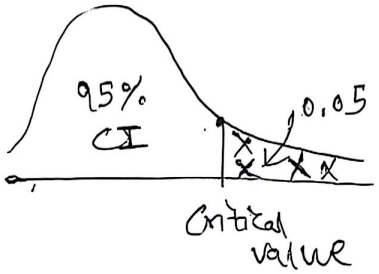
The data doesn't meet the expectation.

(ii) $\alpha = 0.05$ C.I = 95%

(iii) Dof

$3 - 1 = 2$

(iv) Decision Boundary



dof = 2

significance = 0.05

see chi SQUARE Table for value

5.991 (from Table)

If χ^2 is greater than 5.99,

Reject H_0

else

we fail to reject the null Hypothesis

(v) Calculate CHI SQUARE Test statistic

$$\chi^2 = \Sigma \frac{(O-E)^2}{E}$$

$$\chi^2 = \frac{(140-100)^2}{100} + \frac{(160-150)^2}{150} + \frac{(200-250)^2}{250}$$

$$\chi^2 = 26.66$$

$\chi^2 > 5.991$

Reject Null Hypothesis.

Conclusion :

The weights of 2020 population are different than those expected in the 2010.

Analysis of Variance (ANOVA)

→ ANOVA is a Statistical method used to compare the means of three or more groups.

ANOVA :

① Factors (variables)

② Levels

Eg: medicine (Factor)

Dosage: 5mg, 10mg, 15mg → levels

mode of payment (factor)

GPAY, PHNPE, IMPS, NEFT

(levels)

Assumptions in ANOVA:

1. Normality of Sampling Distribution of Mean.
   → The distribution of sample mean is normally distributed.

2. Absense of outliers.
   → outliers need to be removed from the dataset.

3. Homogenity of variance
   → Population variance in different levels of each independent variable are equal.

$$[\sigma_1^2 = \sigma_2^2 = \sigma_3^2]$$

4. Samples are independent & Random.

~~Mathematics And Statistics for data science by Code Basics~~

Complete Statistics for Data Science By Krish Naik !

① Descriptive statistics

↓

(i) Measure of central Tendency

(ii) " " Dispersion

summarizing the data

Histograms, Pdf, Cdf, Probability

- Permutation, Mean, median, Mode, Variance, standard deviation.

→ Gaussian Distribution

→ LogNormal Distribution

→ Binomial "

→ Bernoulli's "

→ Power law distribution

→ Standard Normal distribution

→ Transformation and standardization

→ Q-Q Plot

② Inferential statistics

→ Z-test

→ t-test

→ ANOVA (also called F-Test)

→ CHISQUARE

→ HYPOTHESIS TESTING { P value

Confidence Intervals

What is statistics?

→ It is the science of collecting, organizing and analyzing data.

→ statistics helps in taking better decision making.

What is Data?

→ facts or pieces of information that can be measured.

Ex: The IQ of a class
{98, 97, 60, 55, 75, 65} → Data

Age of students of a class
{30, 25, 24, 23, 27, 28} → Data

Types of statistics:

① DESCRIPTIVE STATS

→ It consists of organizing and summarizing data

② INFERENTIAL STATS

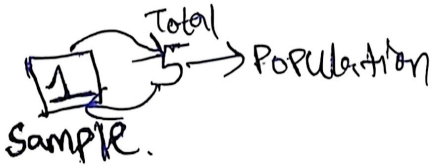
→ Techniques where in we used the data that we have measured to form conclusions.

Ex:
Classroom of Maths student(20)
Marks of the 1st sem
84, 86, 78, 72, 75, 65, 80, 81, 92, ——

Eg:- Descriptive stats
① What is the average marks of the students in the class?

Eg:- Inferential stats question be like
→ Are the marks of the students of this classroom similar to the marks of the maths classroom in the College?

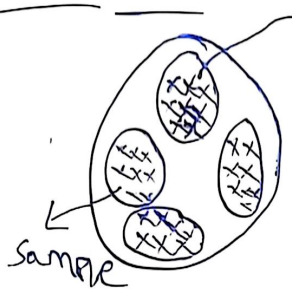
→ Population
→ If there are 5 classrooms, then I take only 1 classroom to analyse. ✓ sample



Population and Sample :

Elections → Goa, UP

Exit Poll



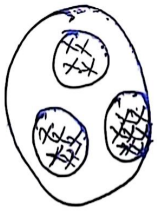
Based on Maximum no. of people voting the winner will be predicted

Population (N) Sample (n)

Sampling Techniques!

① Simple Random Sampling!

→ When performing simple Random sampling, Every member of the population (N) has an equal chance of being selected for your sample (n)



② Stratified Sampling :-

→ In this the population (N) is split into non-overlapping groups (strata)

Eg:- Gender → Male, Female

③ Systematic Sampling

from population (N) → nth individual can be examined

Eg:- Mall → survey (Covid)
↳ for every 8th person you do survey
↙ It's not fixed, you can do survey to every 1st person or 3rd person, whatever person you need you can do survey to the same nth person in a systematic way

④ Convenience Sampling

→ For example we are doing survey, we can take only expertise persons in that domain.

Ex:- Data science

→ If we are doing survey on data science, we can consider persons who have knowledge on datascience only.

Variables !

→ A variable is a property that can take on any value.

Eg:- Height = {192, 178, 168, 150, 160, 170}
weight = {78, 99, 100, 60, 50}

Eg:- survey
Age groups (1-10), (11-20), (21-30) ? → In this there is no chance of overlapping

Two Kinds of Variables

① Quantitative Variable → Measured Numerically, {Add, Substract, multiply divide}

② Qualitative / Categorical Variables

↳ Eg :- Gender → Male, Female

{Based on some characteristics we can derive categorical variables}

Eg: IQ

0–10 ⇓ less IQ

10–50 ⇓ medium IQ

50–100 ⇓ Good IQ

→ It is two types ~~Blood Group~~

Quantitative ~~Quantitative~~

Discrete Variables

Eg :- Whole number

① no. of Bank Accounts
2, 3, 4 ..

② Total no. of child in family
2, 3, 4

Continuous Variables

Eg:

Height = {172.5, 162, 163.7 ...}

weight = {100, 99.5, 99.75 ...}

Rainfall = {1, 1.25, 1.3 ...}

Examples :

① What kind of variable Gender is ? Categorical

② " " " Marital status? Categorical

③ " " " River length ? Continuous

④ " " " Population of state is? Discrete

⑤ " " " Song length ? Continuous.

Blood pressure ? Continuous

PIN CODE ? {Discrete or categorical}

↙

As classes are going we will know about it

Variable Measurement Scales

4 types of measured variable

① Nominal ② ordinal

③ Interval ④ Ratio

① Nominal data

→ This is categorical / qualitative

→ This comes from names. data

Eg:- Colors, Gender, Type of flower

Eg:- Married or not married, Yes or No, blood type (A, B, AB, O) etc.

② Ordinal

→ order of data matters, value does not

→ Ordinal means order. order means ranking

Eg:- Students (marks)	Rank
100	1
96	2
57	4
85	3
42	5

→ Ordinal scales both name and order.

Eg:- finish a race (first, second, third...)

Cancer stage (stage I, stage II, stage III)

③ Interval data

→ In this order matters, values also matters but natural zero is not present

Eg:- Temperature

Fahrenheit

70-80, 80-90, 90-100, 0

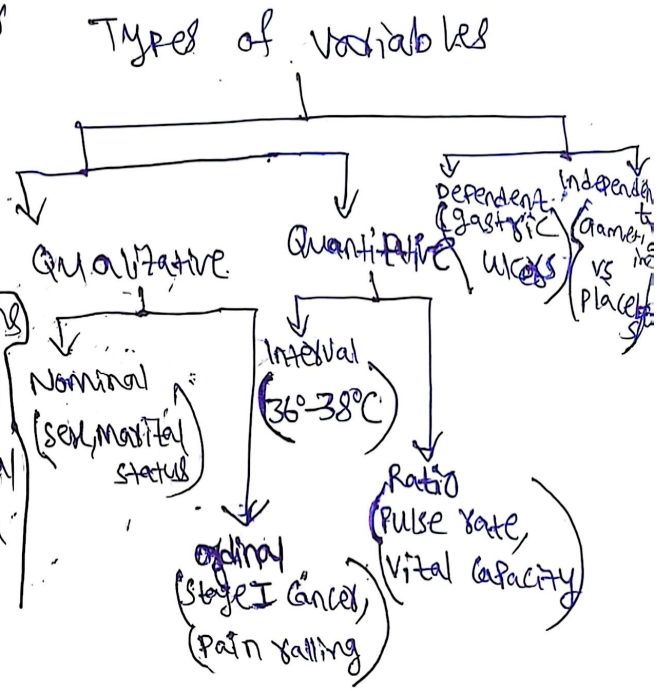
This is not present in this. That's why natural

④ Ratio data

→ This have all the properties of interval variables plus a real absolute zero.

→ value of zero represents the total absence of the variable.

Ex:- 100 feet is twice as long as 50 feet.



Frequency Distribution

Sample data: Rose, lilly, sunflower, Rose, lilly, sunflower, Rose, lilly, lilly

Flower	Frequency	cumulative frequency
Rose	3	3
Lilly	4	7
sunflower	2	9

① BAR GRAPH:

} Bar charts

BAR VS HISTOGRAM :-

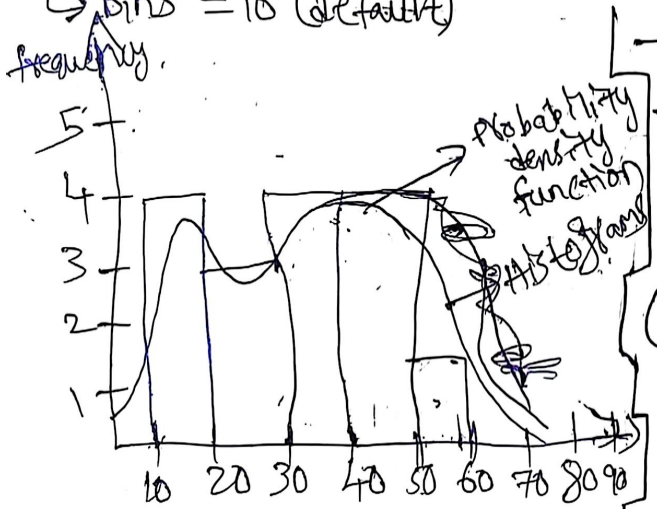
→ Bar is used for discrete values

→ Histogram is used for continuous values.

② Histograms: continuous

Ages = {10, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 41, 42, 43, 50, 51}

↳ Bins = 10 (default)



Basics to Intermediate stats:

→ measure of central tendency

→ measure of dispersion

→ Gaussian distribution

→ Z-score

→ standard normal distribution

① Arithmetic mean for Population & Sample:-

Mean (Average)

Population (N)	Sample (n)

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

$x = \{1, 1, 2, 2, 3, 3, 4, 5, 5, 6\}$

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

$$= \frac{1+1+2+2+3+3+4+5+5+6}{10}$$

$$= \frac{32}{10} = 3.2$$

Probability density function:-

→ Smoothening of histograms.

How the smoothening is done?

Ans:- Based on {kernel density estimator}

we will learn about it in upcoming classes.

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

$$= 3.2$$

Measure of Central Tendency!

What is Central Tendency?

→ Refers to the measure, used to determine the centre of the distribution of data.

① Mean ② Median ③ Mode

② Median

{1,1,2,2,3,3,4,5,5,6,100}

$$\text{Mean} = \frac{1+1+\cdots+100}{11}$$

$$= \frac{132}{11} = 12$$

→ outliers can effect the "mean".

→ outliers make adverse effects on the distribution (mean)

median! {1,1,2,2,3,3,4,5,5,6,100}

→ first sort the numbers

odd number = 11

$$\text{median} = \frac{11}{2} = 5.5$$

= 6th element is median

even number = 12

{1,1,2,2,3,3,4,5,5,6,100,112}

even number = 12

$$\text{median} = \frac{12}{2} = 6$$

= Avg of 6th & 7th element

$$= \frac{3+4}{2} = \frac{7}{2} = 3.5$$

median = 3.5

→ median works well even if outliers present.

Mode!

{1,2,2,3,4,5,6,6,6,7,8,100,100,100,100}

mode = {most frequent Element}

mode = 3

→ mode works well with "categorical" data!!

Where mode is used?

Type of flower	Petal length	Petal width	DATASET 10% missing
Rose			
Lily			
Sunflower			

{missing value → fixed with most recurring element

= missing data

mode works well with Categorical variable.

Measure of Dispersion :-

1. Variance
2. Standard Deviation

↓ It means how well the data is spread

$\{1, 1, 2, 2, 4\} \rightarrow \mu = \frac{80}{5} = 2$

$\{2, 2, 2, 2, 2\} \rightarrow \mu = \frac{10}{5} = 2$

How these two distributions is different?

→ So, to say difference we use variance and standard deviation.

① Variance : (σ^2)

→ It is nothing but spread.

Population variance

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

(Variance)

Sample variance → sample mean

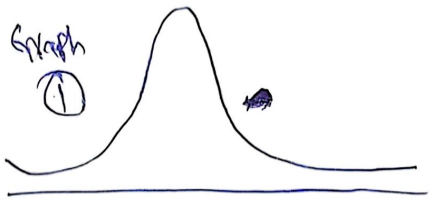
$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

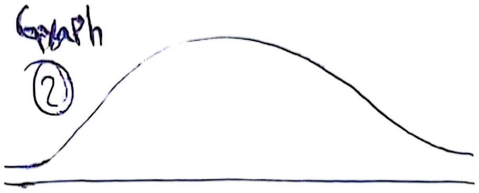
Ex:-

x	μ	$x-\mu$	$(x-\mu)^2$
1	2.83	−1.83	3.34
2	2.83	−0.83	0.6889
2	2.83	+0.83	0.6889
3	2.83	0.17	0.03
4	2.83	1.17	1.37
5	2.83	2.17	4.71
			10.84

$\mu = 2.83$

$\sigma^2 = \frac{10.84}{6} = 1.81$

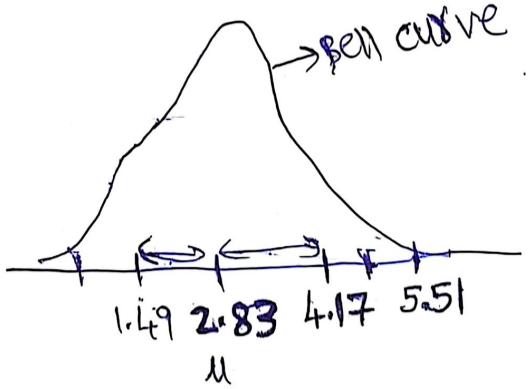




→ The variance is more in graph 2. Because the spread is more in Graph 2 compared to Graph 1

$\sigma = \sqrt{\text{Variance}} = \sqrt{1.81} = 1.345$

"Standard deviation"



$\mu = 2.83$

$\sigma = 1.34$

2.83 + 1.34 = 4.17 → first standard deviation towards right

2.83 − 1.34 = 1.49 → first standard deviation towards left

4.17 + 1.34 = 5.51 → second standard deviation towards right

If we want to find '5'.

→ It falls between 4.17 & 5.51

So, it is 1.5σ from the M

How many are below '11'?

$= \frac{17}{20} \times \frac{5}{100} = 85\%$

Percentiles and Quartiles:

→ These are used to find the outliers.

Percentage:

1, 2, 3, 4, 5

% of the numbers that are odd?

$\% = \frac{\#\text{ of numbers that are odd}}{\text{Total numbers}}$

$= \frac{3}{5} = 0.6 = 60\%$

Percentile:

→ A percentile is a value below which a certain percentage of observation lie.

Ex:

Dataset: 2, 2, 3, 4, 5, 5, 5, 6, 7, 8, 8, 8, 9, 9, 10, 11, 11, 12

What is the percentile ranking of 10?

$x = 10$

$\text{Percentile rank of } x = \frac{\#\text{ of values below } x}{n} \times 100$

$= \frac{16}{20} \times \frac{20}{100} = 80\%$

(Q2) what value exists at percentile ranking of 25%?

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

$= \frac{25}{100} \times (21) = 5.25$

↓ Index position

So take 5th & 6th element and do average.

In the example the 25th percentile value is '5'.

Quartiles

Five number Summary

① Minimum ② first quartile (Q_1)

③ median ④ Third " (Q_3)

⑤ Maximum

Removing the outliers

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

Lower fence $= Q_1 - 1.5(IQR)$

Upper fence $= Q_3 + 1.5(IQR)$

Q_1 – 25% percentile

Q_3 – 75 percentile

As per example

$Q_1 = \frac{25}{100} \times (19+1) = 5$ → index

$= 3$

$Q_1 = 3$

$Q_3 = \frac{75}{100} \times (19+1) = 15$ → index

$Q_3 = 7$

Interquartile Range (IQR)

$IQR = Q_3 - Q_1$

$= 7 - 3$

$= 4$

Lower fence $= Q_1 - 1.5(IQR)$

$= 3 - 1.5(4)$

$= -3$

Higher fence $= Q_3 + 1.5(IQR)$

$= 7 + 1.5(4)$

$= 7 + 6 = 13$

→ so, anything lesser than lower fence and anything higher than upper fence are outliers here in example less than −3 & higher than 13 are outliers.

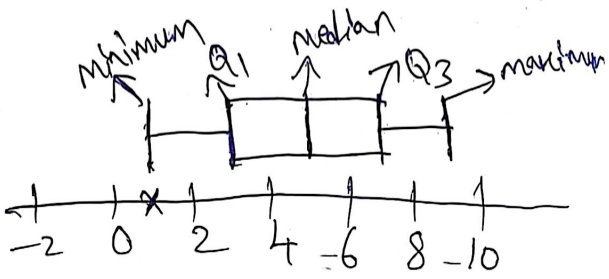
→ After Removing outliers the data is

minimum = 1
$Q_1 = 3$
Median = 5
$Q_3 = 7$
Max = 9

→ 5 number summary

Let us draw Box Plot by using that data.

Box Plot



→ To find outlier we use Box Plot.

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

→ Bessels correction
degree of freedom

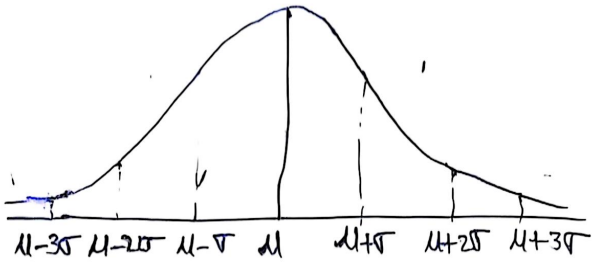
Why sample variance is $n-1$?

And the answer to this question.

Advance statistics!

Distributions

- Normal distribution
- standard Normal Distribution
- Z-score Distribution
- Log Normal Distribution
- Bernoulli's Distribution
- Binomial Distribution



Empirical formula:

→ 68% - 95% - 99.7%

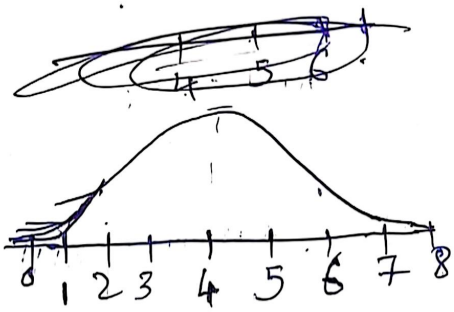
Practical Programming with Python

1. mean, median, mode
2. variance, standard deviation
3. Histogram, pdf, Bar plot, violen plot
4. IQR
5. Log Normal Distribution.

Examples of Gaussian Distribution

Eg: ① Height → Normal Distribution
(Domain expert will say this)
② weight ③ Iris Dataset

Eg: $\mu = 4$ $\sigma = 1$



Where 4.5 → standard deviation is we can say easily, it 0.5σ to the right

But when 4.75 or something like that comes, we can use Z-score

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

$$= \frac{4.75 - 4}{1} = 0.75$$

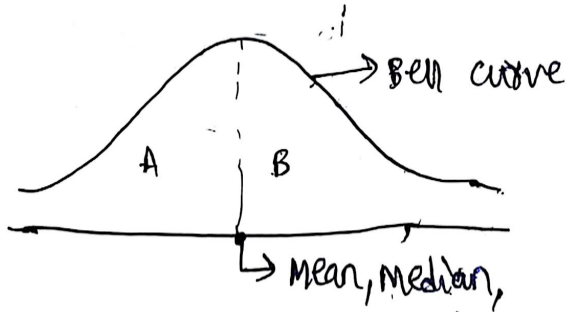
→ Then we can say it is 0.75σ to the right

① Distributions

Ages = {24, 26, 27, 28, 30, ...}

→ Distributions are used to know the data how it is distributed.

(i) Gaussian / Normal Distribution



→ Both A & B are symmetrical

→ If you want to find how many standard deviation to the right or how many to the left you can use 'Z-score'

Ex:-

$\mu = 4 \quad \sigma = 1$

{1, 2, 3, 4, 5, 6, 7} → Normal distribution

⇓ After applying Z-score

{−3, −2, −1, 0, 1, 2, 3} → Standard Normal distribution

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

→ If the condition is satisfied $(\mu = 0, \sigma = 1)$, then it is the "standard Normal distribution".

Practical Application

Dataset

Age (years)	Salary (units)	weight (kg)
24	40k	70
25	80k	80
26	60k	55
27	70k	45

→ These all are having different units, to bring them into one range you have to apply Z-score $(\mu = 0, \sigma = 1)$

→ This process is called standardization.

Normalization

→ To bring data into (0 to 1 range) you can apply Normalization. For this you can use MinMax Scaler (0 to 1)

Practical example on Z-score

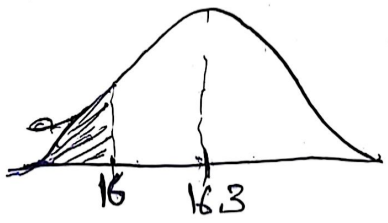
Ex:- your local Pizza shop's claims their large is atleast 16 inches or its free. Their Pizza is normally distributed with $\mu = 16.3$ inches & $\sigma = 0.2$ inches.

(i) what's the probability of getting free Pizza

(ii) what's the probability of getting lucky with Pizza over 16.5 inches?

(iii) what's the probability of getting Pizza between 15.95 & 16.63 inches?

(i) Sol:-



$x_i = 16$

first step is to standardize the distribution

$$Z = \frac{x_i - \mu}{\sigma} = \frac{16 - 16.3}{0.2} = -1.5$$

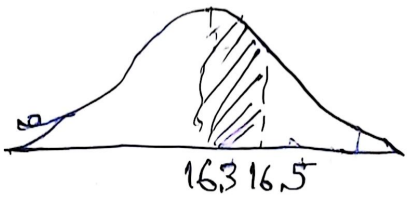
$$P(X < 16) \Rightarrow P(Z < -1.5)$$

Look Left Z-table

$P(Z < -1.5) = 0.0668$

→ we have 6.68% of chances of getting free Pizza.

(ii) $P(X>16.5)$



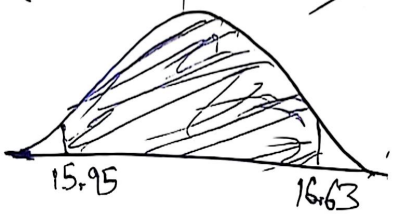
$$Z = \frac{x_i - \mu}{\sigma} = \frac{16.5 - 16.3}{0.2} = 1$$

$P(X>16.5) \Rightarrow P(Z>1)$ = 0.1587

see right z-table

Probability of getting lucky with pizza over 16.5 inches is 15.87%

(iii) $P(15.95 < X < 16.63)$



$$Z = \frac{x_i - \mu}{\sigma} = \frac{15.95 - 16.3}{0.2} = -1.75$$

$$= \frac{16.63 - 16.3}{0.2} = 1.65$$

$P(-1.75 < Z < 1.65)$

$= 0.9505 - 0.0401$

$= 0.9104.$

91.04% people will get pizza of length in between 15.95 & 16.63 inches

Day 4 - statistics

1. IQR - Python
2. Probability
3. Permutation & Combination
4. Confidence Intervals
5. P-value
6. Hypothesis Testing.

(1) Inter quartile range (IQR)

→ This is used to detect outliers.

→ use sns.boxplot to see outliers.

→ find q_1 & q_3 quartiles using

$q_1, q_3 = np.percentile(dataset, [25, 75])$

$IQR = q_3 - q_1$

lower_fence $= q_1 - (1.5 * IQR)$

higher_fence $= q_3 + (1.5 * IQR)$

(2) Probability

→ Probability is a measure of the likelihood of an event

eg: rolling a dice {1, 2, 3, 4, 5, 6}

$P(\text{getting } 6) = \frac{1}{6}$

(i) Addition Rule (Probability, 'or')

Mutual Exclusive Event

If two events are mutually exclusive, they cannot occur at the same time.

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

either 1 or 2 or ... or 6, only one digit can happen at once.

Non mutual Exclusive:

multiple events can occur at the same time.

Eg:- Deck of cards {Q}

↓

It can be heart, diamond, spade etc.

Exclusive Example:

Eg:- If I toss a coin, what is the probability of the coin landing on heads or tails?

Ans:- mutual exclusive

Addition Rule

$$P(A \text{ or } B) = P(A) + P(B)$$
$$= \frac{1}{2} + \frac{1}{2}$$
$$= 1$$
$$P(A \text{ or } B) = 1$$

Rolling a dice

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

Non-mutual exclusive Example:

Eg:- If you are picking a card randomly from deck. What is the probability of choosing a card that is queen or a heart?

Ans:- Non-mutual exclusive

$$P(Q) = \frac{4}{52} \qquad P(\heartsuit) = \frac{13}{52}$$
$$P(Q \text{ and } \heartsuit) = \frac{1}{52}$$

Addition Rule for non-mutual exclusive events.

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$
$$P(Q \text{ or } \heartsuit) = P(Q) + P(\heartsuit) - P(Q \text{ and } \heartsuit)$$
$$= \frac{4}{52} + \frac{13}{52} - \frac{1}{52}$$
$$= \frac{16}{52} = \frac{4}{13}$$

Note:

→ Addition rule for mutual exclusive events

$$\boxed{P(A \text{ or } B) = P(A) + P(B)}$$

→ Addition rule for non-mutual exclusive events.

$$\boxed{P(A \text{ or } B) = P(A) + P(B) - P(A \cap B)}$$

(ii) Multiplication Rule:

Independent Events

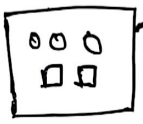
Eg:- Rolling a dice {1,2,3,4,5,6}

one event cannot depend on another. Each & every event are independent.

When you rolled a dice, coming of 1 or anything, cannot depend on another

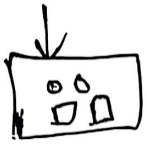
Dependent Event :-

Ex:-



$\rightarrow P(\text{circle}) = \frac{3}{5}$

If we take out one circle



$P(\text{square}) = \frac{2}{4}$

→ Naive Bayes algorithm is based on dependent Event.

Example of Independent Events :

But what is the probability of rolling a "5" and then a "4" in dice?

Ans:- Independent Event

multiplication Rule

$P(A \text{ and } B) = P(A) * P(B)$

$P(4 \text{ and } 5) = \frac{1}{6} * \frac{1}{6} = \frac{1}{36}$

Dependent Event example

What is the probability of drawing a queen & then a Aces from a deck of cards?

Ans:- Dependent

$P(A \text{ and } B) = P(A) * P(B/A)$ → conditional probability

$P(Q \text{ and Aces}) = P(Q) * P(\text{Aces}/Q)$

$= \frac{4}{52} * \frac{4}{51}$

③ Permutation & Combination :-

Permutation :-

Ex:-

School trip (Chocolate factory)

Dairy, 5star, Milkybar, Eclairs, Gem, Kitkat.

Student {Assignment}

When he visit how many 3 types of chocolates he can see.

$\underline{6} \times \underline{5} \times \underline{4} = 120$

formula

$$nP_r = \frac{n!}{(n-r)!} = \frac{6!}{(6-3)!}$$

$= \frac{6 \times 5 \times 4 \times 3!}{3!}$

$= 120.$

Combinations

$$nC_r = \frac{n!}{r!(n-r)!}$$

$6C_3 = \frac{6!}{3!(6-3)!} = \frac{6 \times 5 \times 4 \times 3!}{3! \times 3 \times 2}$

$= 20.$

'20' unique combinations