STATISTICS

M T W T F S S

Statistics is the science of collecting, Siganizing and analyzing data.

Data: "facts & pieces of information"

Ex:) Height of students in a classroom = {175 cm, 150 cm, 140 cm, 155 cm}

2) Intelligence Quotient (IQ) of 5 randomly selected individuals.
= (109, 89, 129, 101, 105)

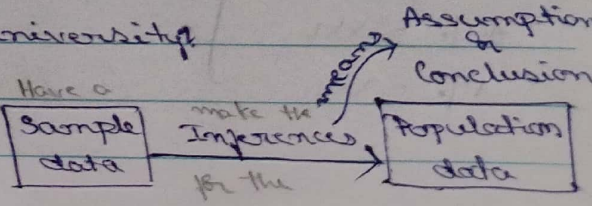
Two types of statistics: Descriptive Stats
Inferential stats

Descriptive Statistics	Inferential Statistics
• It consists of Saganizing and summarizing of data	• It consists of using data that you've measured to form conclusions.
• Ex: Probability Density Function (PDF), Histograms, Box plot, Bar chart, Pie chart	• Ex: Hypothesis Testing, P-value, z-test, t-test, Anova test, chi-square test

Example: Lets say there are 20 maths classes at your university and you've collected the ages of students in one class

Ages = {21, 20, 18, 34, 17, 22, 24, 25, 26, 23, 22, 20}.

• What is the average (& mode & median & min) age of student in your maths class?	• Are the ages of the students in this maths class similar to what you would expect in a normal maths class at this university?

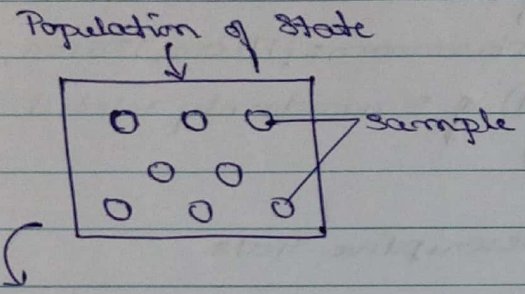


M T W T F S S

Population and Sample data

- working with Population and sample data is very much important for statistics inferential
- There are different ways of sampling too.

Elections:



News channels can not ask each and every person, as to whom they have cast their votes to... for prediction

So, news channel pick up some sample of people to know to whom they have voted

Let total no. of ppl selected for sample is 1000.

Parties for election { BJP, Congress, Independent
60% 30% 20%

⇒ Exit Poll Results

Example: Have to take up the task of ordering the Jacket for the employees of a company (considering 4 branches at different locations) 60k - 70k employees approx.

- It is not possible to ask each and every employee the size.
- Let's take sample data & make some assumptions.,

40% - L
30% - XL
20% - S
5% - XS
4% - XXL
1% - XXL

- with this only 1-2% of jackets were wasted (like not of correct size for ppl)

M T W T F S S

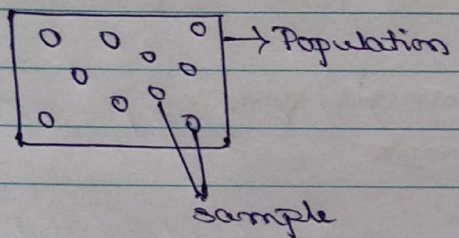
Population (N)

Sample (n)

Sampling Techniques

1) Simple Random Sampling

Every member of the population (N) has an equal chance of being selected for your sample (n)



2) Stratified Sampling

Strata (means) ⟶ layers
↓ also
Non overlapping groups
↓ or
clusters

Ex:
- Gender → Male, Female
- Age groups → 0-18, 18-35, 35-60
- Blood groups
- Courses
- Education Qualification

3) Systematic Sampling

In a population of (N) ⟶ Select every n^{th} individual → 7^{th}

Ex: Survey → Mall (SBI credit card)

Airport → Customs (checking luggage)

M T W T F S S

4) Convenience Sampling

Only those people who are intrested will only be participating

Ex:- Data Science Survey → sending only to ppl who knows about it & Artifical Intelligence → anyone intrested in it

- Youtube survey
- Healthcare Disease
  Blind people survey → Only blind ppl participate in it
- Economics - RBI
  House Hold Survey → Info collected from Female

Exit Poll : Stratified + Random Sampling Sampling

Variable:

A variable is a property that can take on any value

Ex: Height = 182

a = 150

b = 145

xyz = 160

A = {7, 9, 10}.

There are 2 kinds of variable → Quantitative variable, Qualitative variable

1) Quantitative variable:

Can be measured numerically {& can also perform mayb operation like Add, Sub, *, ÷}

M T W T F S S

2) Qualitative Variable: Based on some characteristics we can derive categorical variables

Ex: Gender → Male / Female

- Can we convert a Quantitative variable → Qualitative variable?
Yes it is possible.

Ex: IQ

0-10	10-50	50-100	
Low IQ	Medium IQ	Good IQ	→ 3 categories (Qualitative)

Quantitative Variable

- Discrete variable
- Continuous variable

Ex:
- Whole number.
- Number of Bank Accounts = {2, 3, 4, 7}
- Total no. of children in a family = {2, 3, 7, 10}
- Total no. of employees in a company = {10k, 30k, 50k}

Ex:
- Height = {172.5, 165.5, 130.78} cm
- Rainfall = 1.35, 1.25, 1.75 cm
- Weight of a person
- Temperature
- Measurement of ingredients

Questions:

1) What kind of variable does it belong to:

[a] Marital status — Categorical (Qualitative)
[b] Nile River length — Continuous quantitative variable
[c] Movie duration — Continuous quantitative variable
[d] IQ — Continuous quantitative variable

M T W T F S S

Frequency Distribution

Consider

Sample dataset : Green, Red, Yellow, Green, Red, Yellow, Green, Red

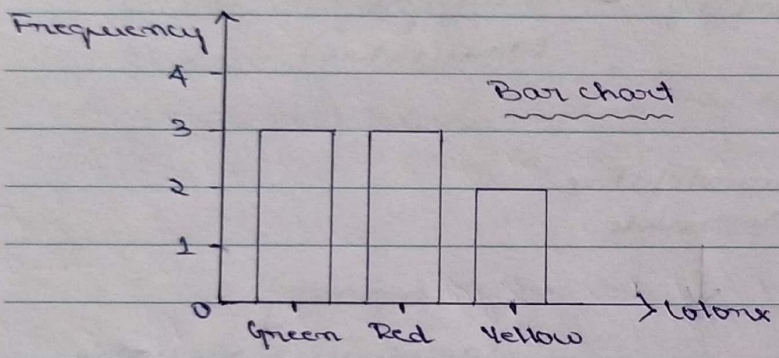
WRT colors

Color	Frequency
Green	3
Red	3
Yellow	2

↳ Categorical variable.

- Bar graph - categorical variable
  Histogram - continous variable

BAR GRAPH



Variable Measurement Scales

- 4 types of Measured variables.

1) Nominal data : {categorical data}
   ↳ Ranking is not that important
   Ex: Colors, Gender, Flowers types

2) Ordinal data : {categorical data}
   ↳ Rank can be assigned to it

Ex: [a]

Student (Marks)	Rank
100	1
96	2
85	3
57	4
45	5

ordinal data (percentiles)

Ranking will definitely help us to understand things in a better way.

M T W T F S S

[b]

Degree	Salary
PHD	(1) ...
B.E	(3) ...
Masters	(2) ...
BCA	(4) ...
12th	(5) ...

3) Interval data

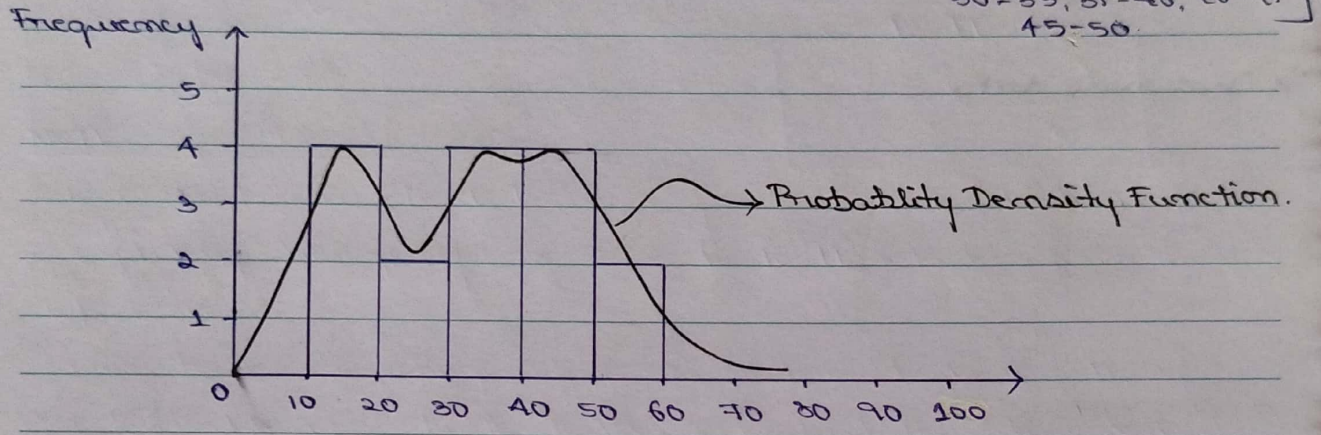
Binx: The number of parts an interval/range is divided into.

HISTOGRAMS [Continuous variables]

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

Binx = 10 (Default) ⇒ x-axis
↳ group/range size

[Ex: 0-50 —bins→ 0-5, 5-10, 10-15, 15-20, 20-25, 25-30, 30-35, 35-40, 40-45, 45-50]



Ages	Frequency
0-10	0
10-20	4
20-30	2
30-40	4
40-50	4
50-60	2

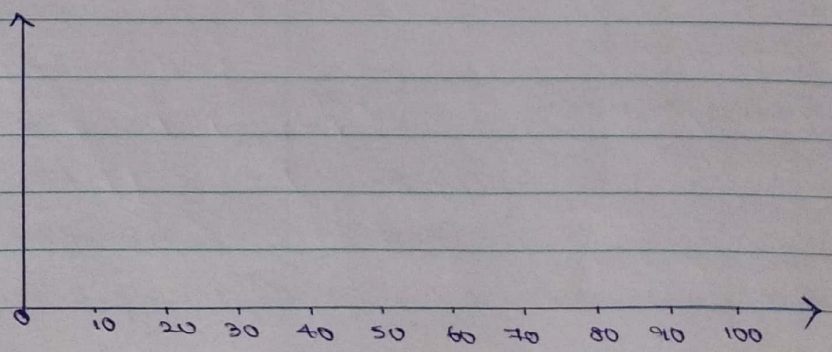
Ages ⇒ a-b (excluding b)

- **Probability Density Function (PDF)**
  It is the smoothened version of Histogram
- **Kernel Density Estimator**
  Used to smoothen histogram.

Example: 10, 13, 18, 22, 27, 32, 38, 40, 45, 51, 56, 57, 88, 90, 92, 94, 99

Create a histogram with total bins as 10.

- 0-10, 10-20, 20-30, 30-40, 40-50, 50-60, 60-70, 70-80, 80-90, 90-100



M T W T F S S

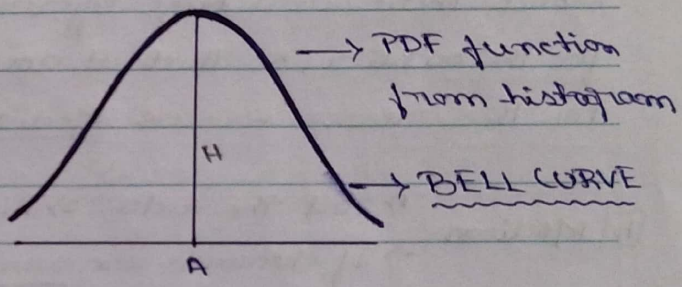
Intermediate Statx

1) Measure of Central Tendency
2) Measure of Dispersion
3) Gaussian Distribution
4) Z-Score
5) Standard Normal Distribution
6) Central Limit Theorem

1) Measure of Central Tendency

↳ Basically identifies the central position of the dataset

[a] Mean
[b] Median
[c] Mode



→ PDF function from histogram

→ BELL CURVE

- Most of the elementx are present in the central region only near A.

- A can be mean/median/mode

H determines in what region the curve is at the top

- Let'x consider:

Population (N)

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

$$\text{Population mean } (\mu) = \sum_{i=1}^{N} \frac{x_i}{N}$$

where, N = size of dataset

x_i = elementx in dataset

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

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

Sample (n)

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

Outlier:
It is an element that looks completely different than that of the entire distribution

[a] Mean

Let $x = \{1, 2, 2, 3, 4, 5\}$
↳ sample

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

$$\bar{x} = \frac{1+2+2+3+4+5}{6} = \frac{17}{6}$$

$$\bar{x} = 2.83$$

- adding outlier to x

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

$$\bar{x} = \frac{1+2+2+3+4+5+100}{7} = \frac{117}{7}$$

$$\bar{x} = 16.71$$

NOTE:

- We can observe that because of an outlier addition, mean completely changes from 2.83 to 16.71, which gives a huge very difference
- Since there is a huge difference, so have to do something for the outlier, so that it does not affect the entire distribution. For this reason we use Median

[b] Median 1) Sort the data. 2) Pick up the central element from data. odd
3) If elements are even taken 2 elements & ÷ by 2 $\left[\frac{a+b}{2}\right]$

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

Median $= \frac{2+3}{2} = \frac{5}{2} = 2.5 \approx 2.83$ (Mean)

- $x = \{1, 2, 2, 3, 4, 5, 100\}$ (outlier)

Median $= 3 \approx 2.83$ (Mean)

- Median definitely works in the case of outliers

[c] Mode - Element with highest frequency is chosen

$x = \{1, 2, 2, 3, 3, 3, 4, 5, 6, 6, 7\}$

Mode = 3

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

Mode = [2, 3, 4] is not in python

If such scenario occurs, use Median

Example:

A girl went to a market and picked up some flowers.

⇒ Lilly, Sunflower, Rose, Lilly, Lotus, NAN, NAN

- NAN is replaced by frequently picked up flower, Lilly here.

M T W T F S S

Usage of Mode:

In EDA → Feature Engineering

[If it has [in the dataset]]

NAN values (NULL)

If the data set has

- Categorical variable → then NAN is replaced using → Mode
- Continuous values → then we can replace NAN values with → Mean
- Continuous values + outliers → then the technique used to replace NAN values → Median

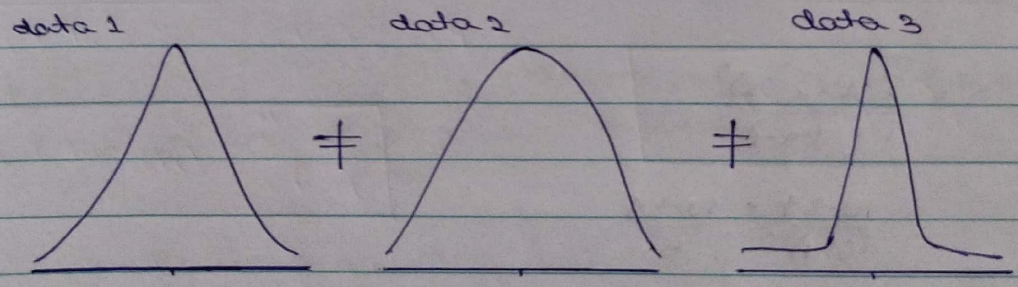
2) Measure of Dispersion

↳ Spread {How the data is spread}

[a] Variance

[b] Standard Deviation

- Consider a specific data (here data 1, data 2, data 3) giving different spread for the datas respectively, which is measured using [a] & [b]



with the help of these different spreadness, we can understand something about data

M T W T F S S

[a] Variance

It basically says how well the data is spread

We defined variance based on 2 things / Population Variance
\ Sample Variance

Population Variance (σ^2) → sigma square

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

where,
μ = population mean

Sample variance (s^2)

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

where,
$\bar{x}$ = sample mean

Q) Why sample variance is divided by $(n-1)$?

Example: x → sample $= \{1, 2, 2, 3, 4, 5\} \Rightarrow \bar{x} = 2.83$
$n = 6$

x	$\bar{x}$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
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

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

$= \frac{10.84}{(6-1)} = \frac{10.84}{5}$

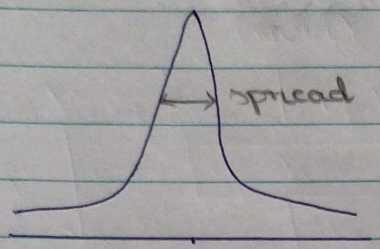
$\therefore \boxed{s^2 = 2.168}$

M T W T F S S

Variance & spread

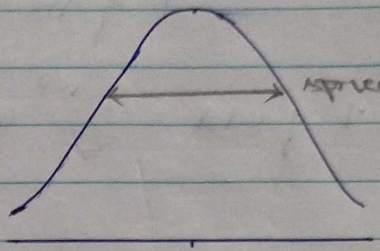
Now, Lets consider, variance (defines the spread)

- From previous example
  Variance = 2.168



Mumbai (less space)
Vertical growth (Tall buildings)

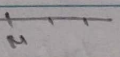
- Random consideration
  Variance = 6.42



Bangalore — villages to cities
Horizontal growth
(Outskirts of bangalore considered & now coming under bangalore)

[b] Standard Deviation (σ)

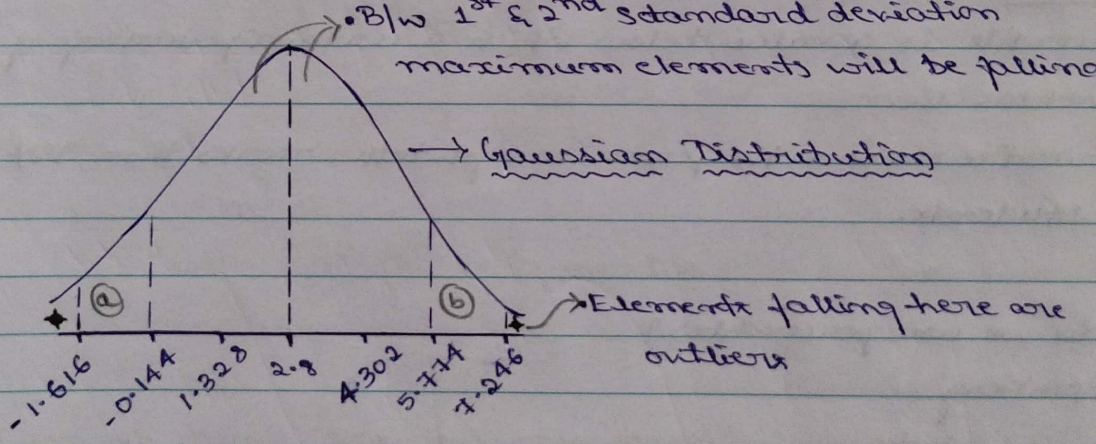
↳ Granular knowledge



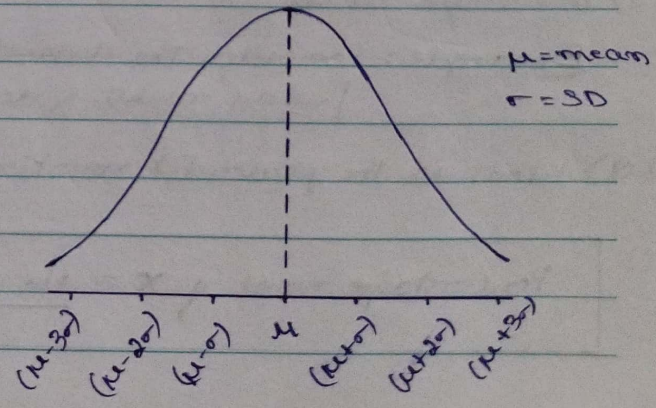
$$\sigma = \sqrt{\text{variance}}$$

Example: $x = \{1, 2, 2, 3, 4, 5\}$, $s^2 = 2.168$

$\sigma = \sqrt{2.168} = 1.472$



- Regions (a) & (b) minimum elements will be falling



M T W T F S S

Concepts:

PERCENTILES AND QUARTILES

Percentages:

Example: 1, 2, 3, 4, 5.

% of the numbers that are odd?

$$\% \text{ of odd} = \frac{3}{5} \times 100 = 60\%$$

Percentiles:

Exams like CAT, GATE, SAT here the scoring happens in terms of percentile

99 percentile $\neq$ 99%.

↓ it means

Out of all the people that have written that exam, his/her rank is somewhere near this particular (99) value.

Definition:

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

99 percentile means the person has got better marks than 99% of the students.

Why do we use percentiles?

1) For Ranking

2) Used in Stats

Example: Consider the dataset {2, 2, 3, 4, 5, 5, 5, 6, 7, 8, 8, 8, 8, 8, 9, 9, 10, 11, 11, 12}
[where n = 20, & data already sorted]

1Q) What is the percentile ranking of 10

$$\text{Percentile rank of } x = \frac{\text{No. of values below } x}{n} \times 100$$

M T W T F S S

Percentile Rank of 10 = $\frac{16}{20} \times 100 = 80$ percentile

Indicates that 10 is greater than 80% of entire dataset distribution

2) What is the percentile rank of 11?

Percentile Rank of 11 = $\frac{17}{20} \times 100 = 85$ percentile.

3) What value exists at percentile ranking of 25%?

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

$$= \frac{25}{100_4} \times (20+1)$$

$= 5.25 \rightarrow$ index not value.

1 2 3 4 5.25 6.

∴ dataset: 2, 2, 3, 4, [5, 5] 5, 6, ..., 12

↳ combine them & find the average as index 5.25 is not a whole no.

Average = $\frac{5+5}{2} = \frac{10}{2} = 5$

Value = 5 for 25 percentile

- These all things are for understanding purpose only no need of calculation as code will automatically do it.
- Formulas derivation are not necessary.

Quartiles

Quartiles basically mean 25 percentile

M T W T F S S

FIVE NUMBER SUMMARY

1) Minimum Value
2) First Quartile Value (25 percentile) Q_1
3) Median (~~50 percentile~~) (Measure of central tendency)
4) Third Quartile Value (75 percentile) Q_3
5) Maximum Value

Removing the outliers

Dataset : {1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 27}

- How do we get to know some element is a outlier?

We have to define the fences for that i.e.,

Lower Fence ⟷ Higher Fence.

Any element that does not belong to the fence is considered as outliers

Lower Fence = Q1 - 1.5(IQR)
Higher Fence = Q3 + 1.5(IQR)
IQR = Inter Quartile Range
IQR = Q3 - Q1
$Q_1 = \frac{25}{100} \times (n+1)$
$Q_3 = \frac{75}{100} \times (n+1)$

For given dataset : $n = 19$

$Q_1 = \frac{25}{100} \times (20) = 5^{th}$ index

$\boxed{Q_1 = 3}$

$Q_3 = \frac{75}{100} \times (20) = 15^{th}$ index

$\boxed{Q_3 = 7}$

IQR = 7 - 3

$\boxed{IQR = 4}$

Lower Fence = 3 - 1.5(4) = -3

Higher Fence = 7 + 1.5(4) = 13

-3 ⟷ 13

∴ 27 does not come into this Fence, hence it is a outlier.

→ Found after excluding outlier

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

M T W T F S S

- Five number summary
(For the dataset)

Box plot

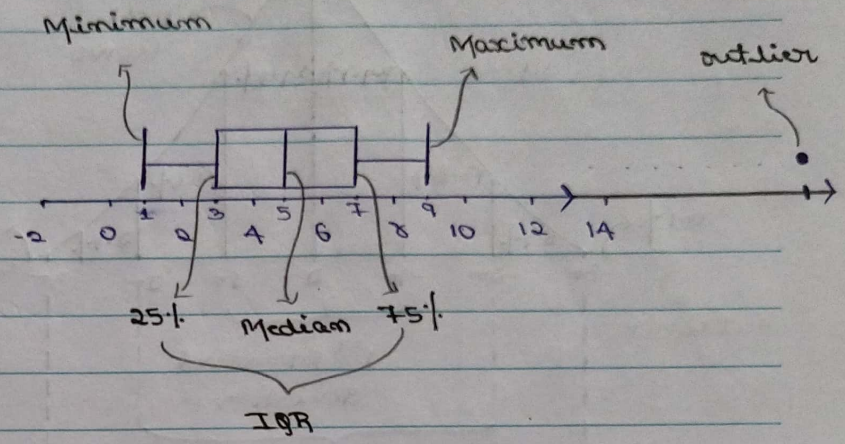
Minimum = 1

$Q_1 = 3$

Median = 5

$Q_3 = 7$

Maximum = 9



Use of Box Plot

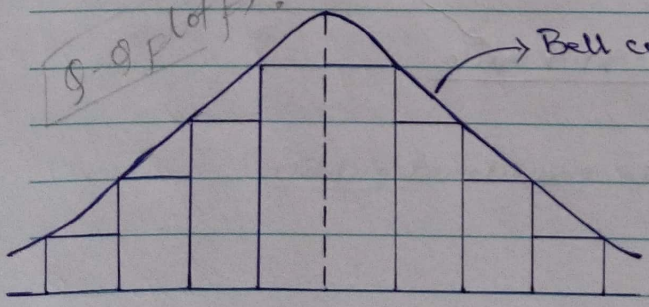
- It is a visualization technique to see the outlier.

3) Distributions

[a] Normal / Gaussian Distribution
[b] Standard Normal Distribution
[c] I - Sobre
[d] Log Normal Distribution
[e] Bernoulli's Distribution } No need of formulas to remember.
[f] Binomial Distribution }
[g] Power law

[a] Normal / Gaussian Distribution

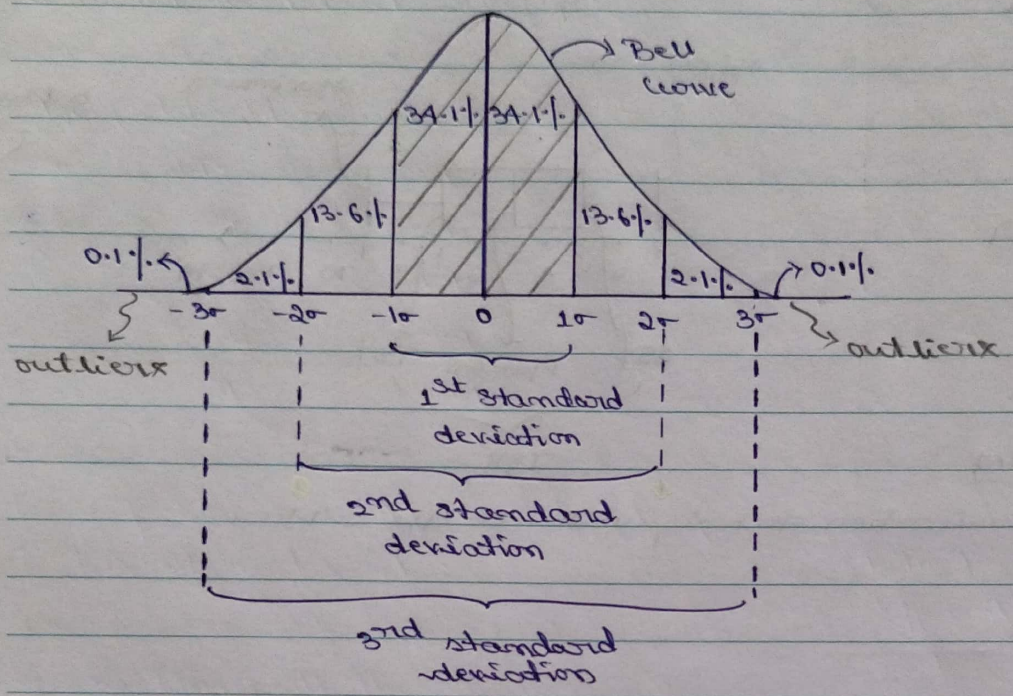
It is nothing but the simple bell curve (as seen before) constructed based on mean/median/mode and the standard deviation.



M T W T F S S

Properties:

- Emperical Rule of Gaussian Distribution



⇝ Area b/w 1st standard deviation (to the right & left)

$34.1\% + 34.1\% = 68.2\%$

Most of the elements are present over here i.e., around 68.2%.

⇝ Area b/w 2nd standard deviation (to the right & left)

$68.2\% + 13.6\% + 13.6\% = 95.4\%$

Here around 95.4% of the entire dataset will be falling in 2nd SD

⇝ Area b/w 3rd standard deviation (to the right & left)

$95.4\% + 2.1\% + 2.1\% = 99.6\%$

Here around 99.6% of data will be found

⁂ Emperical Rule.

This Rule is known as 68-95-99.7 rule

⇝ Anything falling after 3rd SD is considered outlier.

M T W T F S S

Which all kind of dataset will fall into (Gaussian distribution / bell curve)?

1) IRIS Dataset
2) Weight of human being
3) Height

} Domain expertise have conformed them

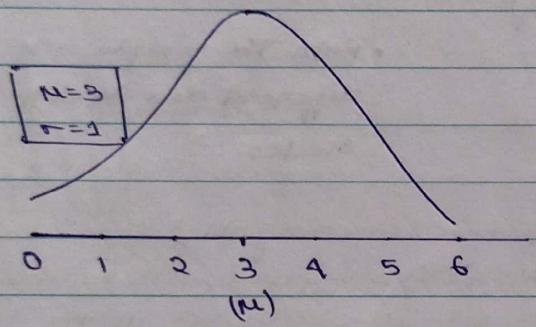
[b] Standard Normal Distribution

Dataset : {1, 2, 3, 4, 5}

(μ) Mean = $3\left(\frac{15}{5}\right)$

(SD) Standard deviation = 1.414

↳ Let's gracefully consider (σ) SD ≈ 1 (to make the calculation easier)



• Can we convert this entire dataset with μ = 3 & SD = 1/5/10 in such a way that μ = 0 & SD = 1?

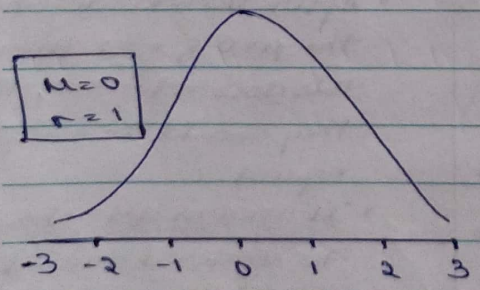
⇒ Yes, it is possible using z-score.

[c] Z-score

{It helps us understand, how far / how many SD(σ) that particular no. falls away from the mean}

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

Why?



elements: {1, 2, 3, 4, 5}.

⇒ Z-score

- $\frac{1-3}{1} = -2$
- $\frac{2-3}{1} = -1$
- $\frac{3-3}{1} = 0$
- $\frac{4-3}{1} = 1$
- $\frac{5-3}{1} = 2$

M T W T F S S

Concept

Standardization vs Normalization

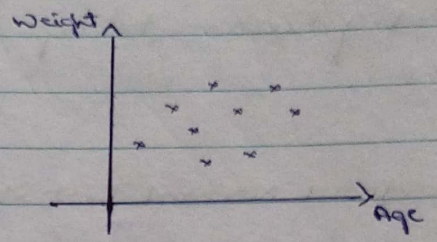
Standardization:

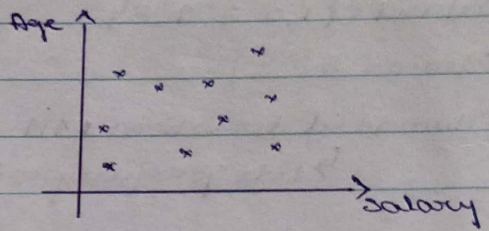
Consider the dataset, with

Age(yr)	Weight(kg)	Salary(INR)
25	75	25k
26	80	30k
28	85	40k
30	60	80k
32	72	85k

$\mu = 28.2$

Scatter plot





- Each feature here (Age, Weight, Salary) has different unit.
- Both the graphs are different due to different scales

- Can we bring all these features into the same unit scale? [For Machine learning Algorithm what ever Maths gets applied, will be able to calculate it quickly as all the values are in the same scale]

Yes, using Standardization applying Z-Score wrt each & every feature.

using Z-score ↓

Consider,

Age	μ	$x_i - \mu$	$(x_i - \mu)^2$	New Age (Standardized)
25	28.2	-3.2	10.24	-1.25
26	28.2	-2.2	4.84	-0.977
28	28.2	-0.2	0.04	-0.078
30	28.2	1.8	3.24	0.703
32	28.2	3.8	14.44	1.484

$\mu = 28.2$ total = 32.8

$\left(\frac{141}{5}\right)$

- After standardization, the $\mu = 0$ & $\sigma = 1$. However, this does not mean that they should be exactly equal.
- It means, data should be centered around zero & the spread of the data should be similar for all features.

$$\text{variance} = \frac{\text{total}}{n} = \frac{32.8}{5} = 6.56$$

$$\text{SD}(\sigma) = \sqrt{\text{variance}} = \sqrt{2.56} = 2.56$$

M T W T F S S

Normalization:

there we specific the min & max value, so that values get entirely converted b/w them.

[Min Max Scaler]

↓ ↓

0 to 1

0 to 1

- It is done using the formula:

Min Max Scaler

$$x_{Nor} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

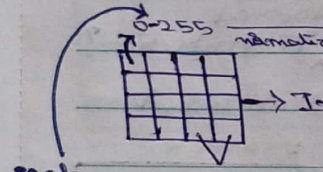
- When should we use standardization & normalization?

⇒ Most of the ML use case - Standardization

⇒ Convolution Neural Network (CNN) - Normalization

↓ In images we have pixels

Pixels are normalized b/w 0-1



Consider:

f1	f1'
2	0.14
5	0.571
6	0.71
8	1
1	0

After using Min Max Scaler

Calculation

$x_{Nor} = \frac{2-1}{8-1} = \frac{1}{7} = 0.14$

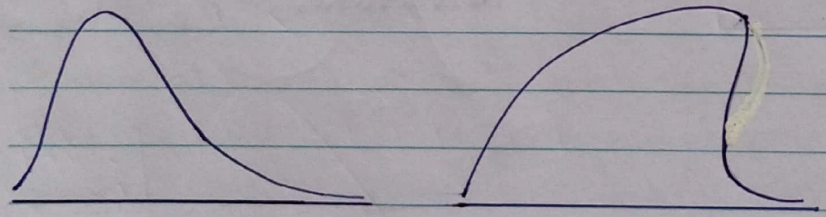
Now these values are b/w 0 to 1

$x_{min} = 1$

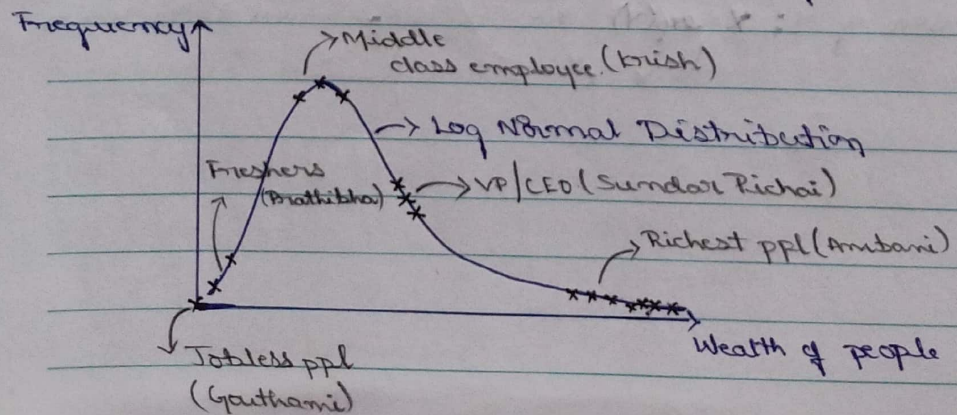
$x_{max} = 8$

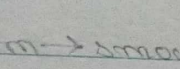
[d] Log Normal Distribution

Skewed Curves → They have lot of outliers



Log Normal Distribution is a type of skewed curves Distribution.



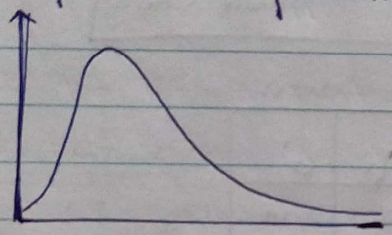
data → Histogram → smoothen →  → log normal distribution.

Basically log Normal Distribution is saying that..

- Very less no. of people will be present in one sufficient area (Richest)
- There will be max. no. of people falling in one area (Middle class)
- There will be v. less no. of people in another area (Jobless, Freshers)

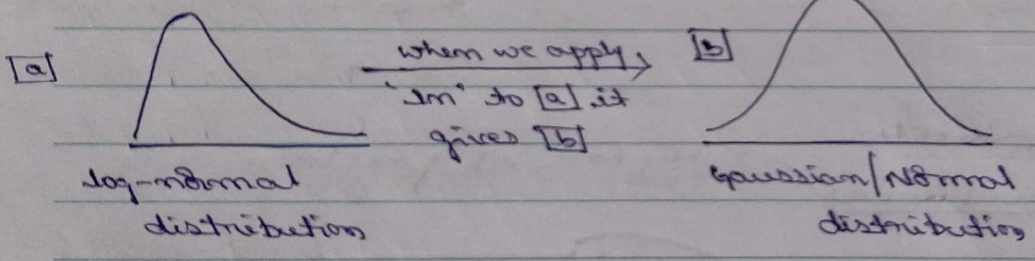
Another Example:

- People writing comments (length) in Youtube channel (krish).



Q) Can we convert log Normal Distribution to Gaussian & Normal Distribution?

→ Yes, if the random variable X is log-normally distributed, then $Y = \ln(X)$ has a normal distribution.



Example.

X	$Y = \ln(X)$
20	- - - -
25	- - - -
30	- - - -
40	- - - -
45	- - - -
↓ [a]	↓ [b]

NOTE:

$\log_{10}$ ⇒ Base 10

$\ln$ ⇒ Base e

→ Equivalently, if Y has a normal distribution, then the exponential function of Y; $X = \exp(Y)$, has a log-normal distribution

↳ $X = e^Y$

Look at wikipedia for more examples.

M T W T F S S

[6] Bernoulli's Distribution

- Categorical variable - with binary outcome.

  Example: Tossing a coin

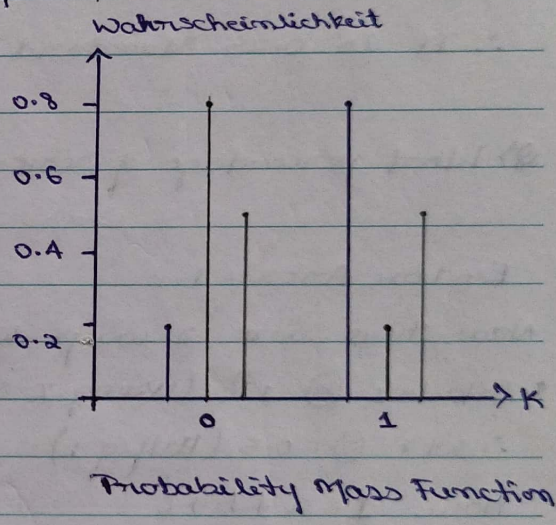
  Probability of passing exam

- Bernoulli's Distribution is the discrete probability distribution of a random variable which takes the value 1 with probability p and the value 0 with probability $q = 1 - p$.

- Probability Mass Function (PMF)

  Whenever we are trying to plot for a categorical variable, at that point of time we say it as PMF

$$\text{PMF} = \begin{cases} q = 1-p & \text{if } k = 0 \\ p & \text{if } k = 1 \end{cases} \quad p^k (1-p)^{1-k}$$



Probability Mass Function

Examples of Bernoulli distribution:

- $P(x=0) = 0.2$ and $P(x=1) = 0.8$
- $P(x=0) = 0.8$ and $P(x=1) = 0.2$
- $P(x=0) = 0.5$ and $P(x=1) = 0.5$

[7] Binomial Distribution (Extention of Bernoulli's Distribution)

Here we combine multiple Bernoulli Distribution

2 o/p Denoted as $B(n, p)$

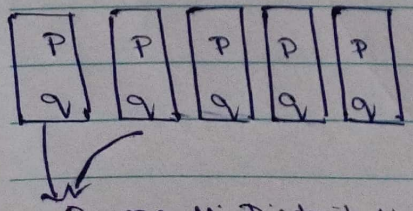
n = no. of experiments/trials, $n \in \{0, 1, 2, \ldots\}$

p = probability (success propability for each trial) $p \in [0, 1]$

$q = 1 - p$; $k \in \{0, 1, \ldots, n\}$ - no. of successes

Example: Tossing a coin 5 times

$$\text{PMF} = \binom{n}{k} p^k q^{n-k}$$



combining multiple Bernoulli is Binomial

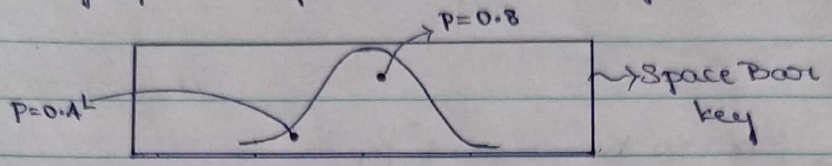
Bernoulli Distribution

(Power law | 80-20 Rule)

M T W T F S S

P value (Probability value)

Pressing/operating a space bar key, maximum clicks at the middle region.



- P=0.8, means., Out of 100 touches, the no. of touches at P=0.8 is 80
- P=0.4, means., Out of all 100 touches, the no. of times we are touching at P=0.4 is 40

Hypothesis Testing

Consider, Coin experiment ⟶ Test whether the coin is fair coin & not by tossing the coin 100 times

Probability of tossing { Head(50 times) ⇒ P(H) = 0.5 ; Tail(50 times) ⇒ P(T) = 0.5 } ⇒ Fair coin

[It is a coin that has an equal chance of landing heads & tails when it is flipped]

- Now how do we test a coin is fair & not using Hypothesis Testing?

Steps in Hypothesis Testing:

1) Null Hypothesis - (Coin is fair) [H_0]

By default whichever statement is true, that is taken as Null Hypothesis.

Example: Criminal (has made some crime) $\xrightarrow[\text{to the}]{\text{goes}}$ court

By default the court will say that criminal is not guilty, until proved.

2) Alternate Hypothesis - (Coin is not fair) - [H_1]

It should be the opposite of Null Hypothesis

3) Experiment - (z-test, t-test, chi-square test, ANOVA test)

It is done to prove the coin is fair & not

M T W T F S S

4) Reject & Accept the Null Hypothesis.

Based on the experiments we perform this.

⇒ Now, lets test whether the coin is fair & not, 100 tosses.

Step 1:

Possibilities of tossing 100 times ⇒

1) 50 times head, 50 times tail — coin is fair

2) 60 times head, 40 times tail — coin is fair

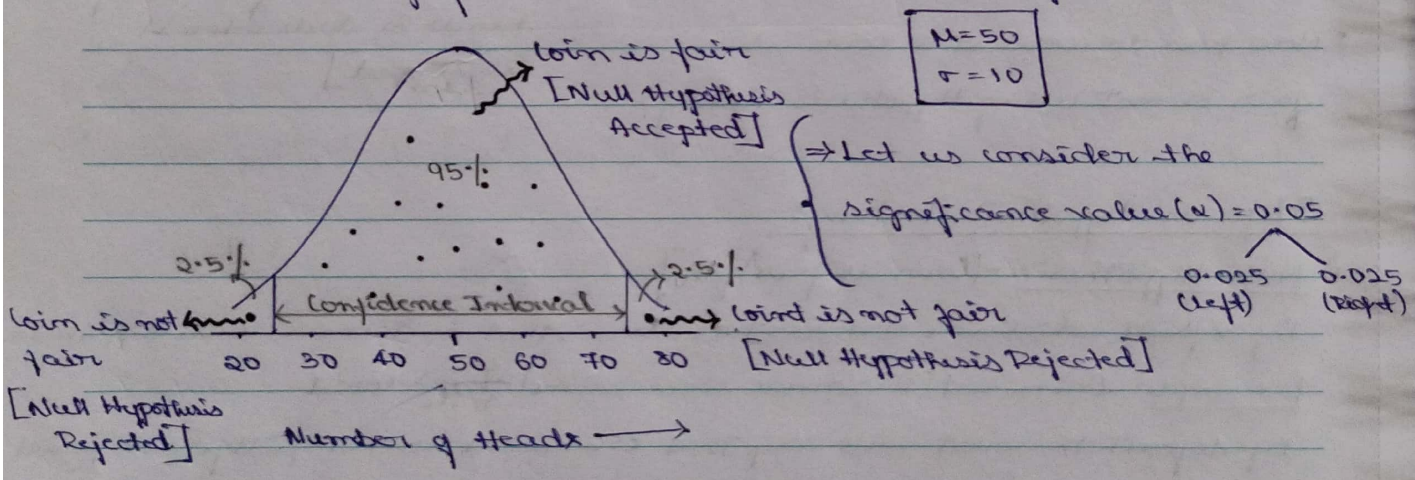
3) 30 times head, 70 times tail — Can't decide fair & not

↓ so to determine this we use

[Confidence interval, Significance value]

Confidence Interval, Significance value.

Based on the significance value we derive our confidence Interval (CI)



⇒ Let us consider the significance value (α) = 0.05

0.025 (Left) 0.025 (Right)

Confidence Interval: (CI)

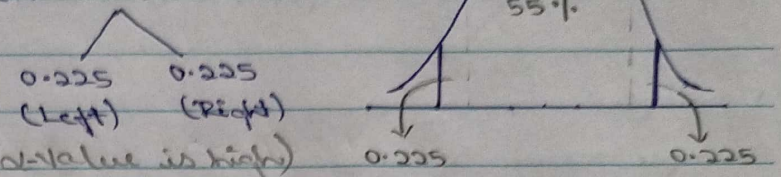
It basically specifies a specific range based on significance interval, if the value falls within that range. Then it is considered the coin is fair Out of that range coin is not fair.

$CI = 1 - \alpha = 1 - 0.05 = 0.95$

$\therefore CI = 95\%$.

M T W T F S S

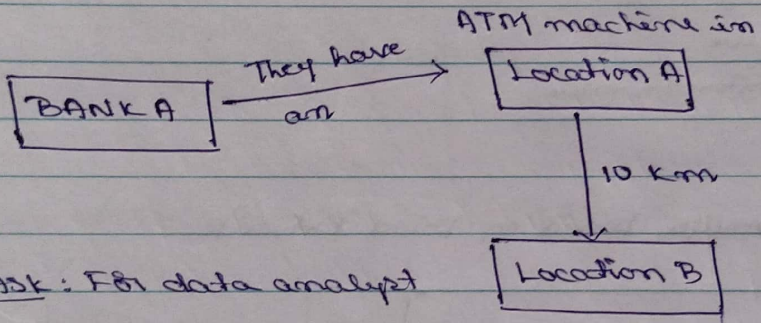
- Consider, Health care domain where Covid Vaccine Test is conducted.
  Error here should be very less.
  Let significance value (α) = 0.15.



  $CI = 1 - 0.15 = 0.55$
  CI = 55% (Restricted as α-value is high)

- Can we derive p value from significance value?
  Yes, possible

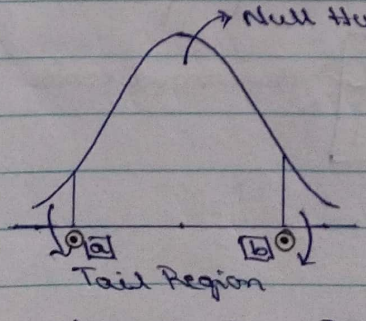
- Real world project: (Question Interview).



TASK: For data analyst
Based on the data that they have WRT location A.
They have to find out whether an ATM should be opened in Location B & not.

This is done with the help of hypothesis testing

Confidence Interval:
There will be some range, if our values are within this range Null Hypothesis will be accepted.
If value comes in tail region Null Hypthesis will be rejected.



- How do we find the values at point [a] & [b] for marking range?
  ⇒ ✗

M T W T F S S

Point Estimate :

The value of any statistics that estimates the value of a parameter is called point estimate.

Example:

Sample mean $\bar{x}$
Population mean μ
A perfect example of point estimate.

$\bar{x}$ is a point estimate of μ

Which means, if $\mu = 3$ then $\bar{x}$ can be 2.8 | 2.9 | 3.5 | 3.5

↓

Keeping sample mean value we can estimate population mean value. (conclude)

Confidence Interval formula based on Point Estimate:

CI = Point Estimate $\pm$ Margin of Error → This in total gives the estimate of population.

Point Estimate ($\bar{x}$) (Sample data)

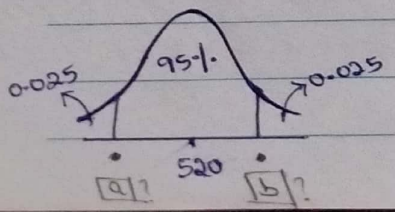
Q) On the quant (quantitative) test of CAT Exam, the population Standard Deviation is known to be 100. A sample of 25 test takers has a mean of 520. Construct a 95% CI about the mean.

Given:

$SD = 100 = \sigma$ (population)
$n = 25$
$\bar{x} = 520$
$CI = 95\%$
$\alpha = 0.05$ $(1 - 0.95)$

- Population SD is given → so we use → Z-score → Z-table
- CI = Point Estimate $\pm$ Margin of Error

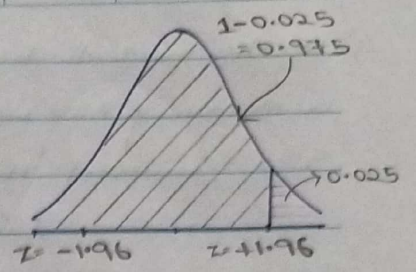
$$CI = \bar{x} \pm Z_{\frac{\alpha}{2}} \left[\frac{\sigma}{\sqrt{n}}\right]$$ → standard error



M T W T F S S

[a] • Lower Fence CI = $\bar{x} - Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$

[b] • Higher Fence CI = $\bar{x} + Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$



• Lower Fence CI = $520 - (1.96)\frac{100}{\sqrt{25}}$

$= 520 - (1.96)20$

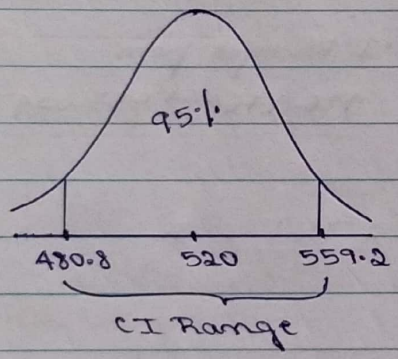
$= 480.8$

$Z_{\frac{\alpha}{2}} = Z_{\frac{0.05}{2}} = Z_{0.025} = 1.96$

Z-table
for value
0.9750
$Z = 1.96$

• Higher Fence CI = $520 + (1.96)\frac{100}{\sqrt{25}}$

$= 559.2$



⇒ For the above problem compute C.I Range values for SD = 80.

• Lower Fence = $520 - (1.96)\frac{80}{5} = 488.64$

• Higher Fence = $520 + (1.96)\frac{80}{5} = 551.36$

M T W T F S S

Q) On the quant test of CAT Exam, a sample of 25 test takers has a mean of 520 with a sample standard deviation of 80. Construct 95% CI about the mean.

Given

$n = 25$

$\bar{x} = 520$

$CI = 95\%$

$\alpha = 0.05$

S.D (Sample) = 80 = s

∴ NOTE:

When ever population Standard Deviation is not given, we need to apply t-test (t-table)

$$C.I = \bar{x} \pm t_{\alpha/2} \left[\frac{s}{\sqrt{n}}\right]$$ → Standard Error.

For t-test we need to calculate

- Degree of freedom = $n - 1$

$25 - 1 = 24$

$$t_{\frac{\alpha}{2}} = t_{\frac{0.05}{2}} = t_{0.025} = 2.064$$

From t-table

$$C.I = 520 \pm 2.064\left(\frac{80}{5}\right)$$

C.I Ranges from

486.976 ⟷ 553.024

Type 1 and Type 2 Error.

Consider the coin tossing as Example,

H_0 = coin is fair

H_1 = coin is not fair

Reality check:

1) The Null Hypothesis is True & Null Hypothesis is False

Decision: [Based on Experiments]

1) Null Hypothesis is True & False

H_0 - Criminal is not guilty

H_1 - Criminal is guilty

M T W T F S S

Outcome 1:

We reject the Null Hypothesis in reality if it is False - Yes

Example: Criminal is guilty, making them a prisoner - Right decision

Outcome 2:

We reject the Null Hypothesis when in reality if it is True - No (Error)

Example: Criminal not guilty, but making them a Prisoner - Wrong decision

This kind of Error is Type 1 Error

Outcome 3:

We accept the Null Hypothesis, when in reality it is False - No (Error)

Example: Criminal is guilty, but not prisoned / set free - wrong decision

This kind of Error is Type 2 Error

Outcome 4:

We accept the Null Hypothesis, when in reality it is True - Yes

Example: Criminal is not guilty, so not prisoned / set free - Right decision

Confusion matrix:

	H_0 (Not Guilty)	H_1 (Guilty)
Not Prisoned (or set free)	Correct	Type II Error
Prisoned	Type I Error	Correct

Example for Errors:

1) Person having cancer → ML model predicts not having cancer.
2) Predicting stock market is going to crash → but it did not crash next day

M T W T F S S

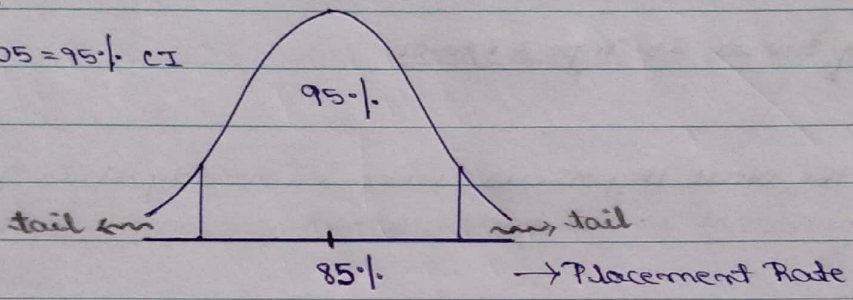
1 Tail and 2 Tail Test

Example:

College in Karnataka has an 85% placement rate. A new college was recently opened and it was found that a sample of 150 students had a placement rate of 88%, with a standard deviation of 4%. Does this college have a different placement rate?

Let $\alpha = 0.05$

then $1 - 0.05 = 95\%$ CI

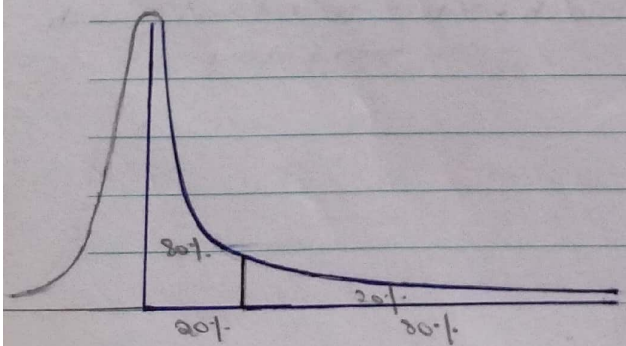


[a] Does this college have different placement rate?
(placement rate)
It can either increase & decrease – 2 Tail Test
(i.e., it can either be greater than 85% or lesser than 85%)

[b] Does this college have placement rate greater than 85%?
Placement rate focused only on one region – 1 Tail Test
(Right side)

[c] Does this college have placement rate lesser than 85%?
Placement rate focused only on one region – 1 Tail Test
(left side)

[g] Power Law (Distribution) not for 2018



- Pareto Distribution
  [Non-Gaussian Distribution]
- 80–20% rule
  80% of data acquired with 20% of area & remaining 20% of data acquired with 80% of area
- This graph is extension of Log Normal Distribution

M T W T F S S

Examples:

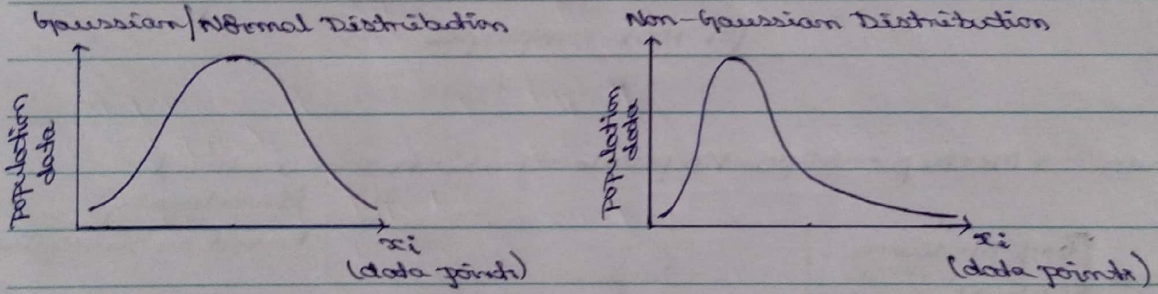
1) x% of team is responsible for winning 80% of match
2) 80% of score was scored by 20% of the team
3) 20% of entire nation hold 80% of the oils
4) 80% of wealth is held by 20% of the people
5) 80% of amazon revenue comes from x% of the products

Day 3

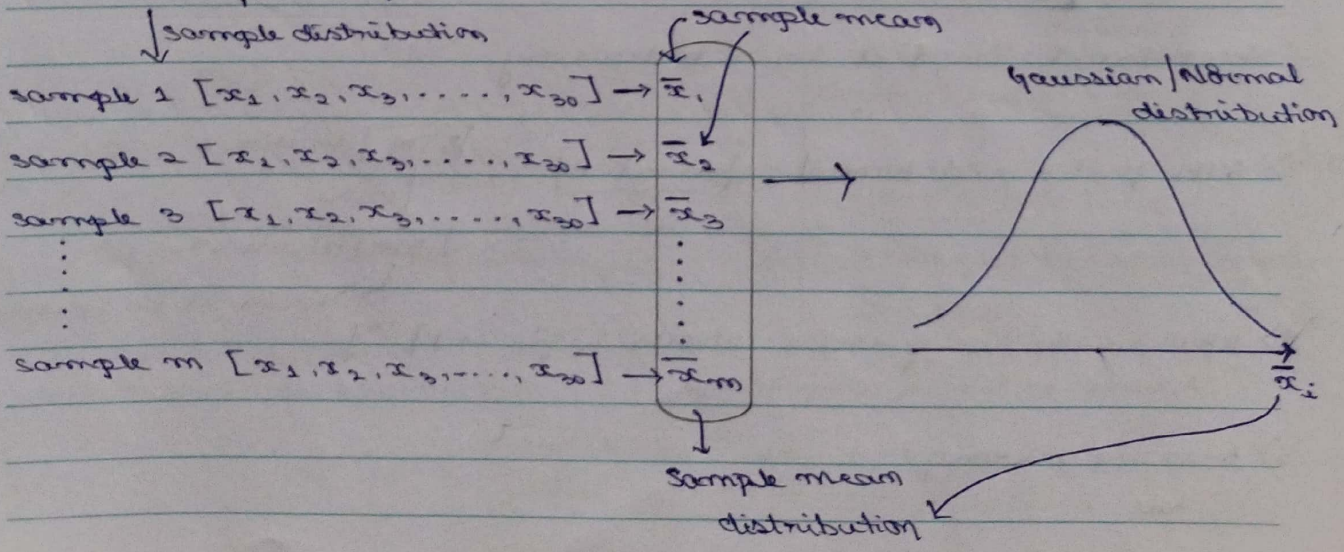
Central Limit Theorem

Let's say we have a distribution, it might be a Gaussian/Normal Distribution & Non-Gaussian Distribution.

If we keep on taking multiple samples from distribution where sample size is greater than & equal to 30, i.e., $n \geq 30$, then sample mean distribution that is obtained is plotted with the help of histogram we get Gaussian Distribution.



- From distribution pick up n elements for sample ($n \geq 30$)



M T W T F S S

- Is it compulsory to take sample size (n) ≥ 30?

[a] Yes, its compulsory for Non-Gaussian Distribution. If $n \geq 30$ is not followed we can not convert it into Gaussian Distribution

[b] If we consider $n \geq 30$ for Gaussian Distribution, we will get much better distribution from the sample.

Inferential Statistics (Important for Data Analyst & Data Scientist)

- z test (z table)
- t test (t table)
- z test proption population

Example:

1) Company with 100k employees → Order T-shirts for employees → ~~Drop a mail to each employee for T-shirt size~~

Collect Sample Data for T-shirt size

Based on this

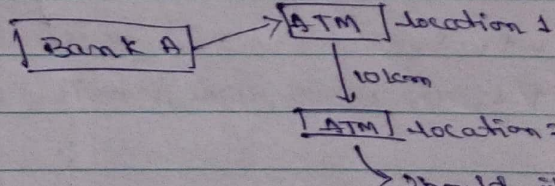
Conclude the sizes for 100k employees

2) iNeuron → Meetup → 300-400 people → T-shirts → ordered (Randomly) Based on Sampling

500 t-shirts
- 60% Medium
- 10% XL
- 10% XXL
- 20% L

Next Event: Plan it in a better way

Interview question
3) ATM problem statement



Interview question
4) Measure the size of entire shorts & give CI [ ]
Assume μ & σ & compute C.I

5) Amazon delivery [percentile, Quartile]
late delivery tracking.

M T W T F S S

Hypothesis Testing

[a] Z-test

Q) A factory has a machine that fills 80 ml of baby medicine in a bottle. An employee believes the average amount of baby medicine is not 80 ml. Using 40 samples, he measures the average amount dispensed by the machine to be 78 ml with a standard deviation of 2.5.

[a] State Null and Alternate Hypothesis

[b] At a 95% C.I, is there enough evidence to support the machine is not working properly? 2-Tail Test (can be <80 & >80)
(2 Rejection region)

Step 1:

Null Hypothesis (H_0) $\Rightarrow \mu = 80$

Alternate Hypothesis (H_1) $\Rightarrow \mu \neq 80$

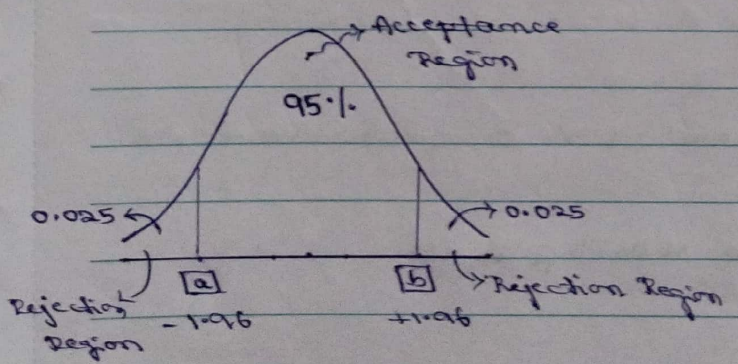
$n = 40$
$\bar{x} = 78$
$s = 2.5$
} sample

Step 2:

C.I = 95%

$\alpha = 0.05$

Step 3: Create Decision Boundary



z-score $\Rightarrow (1 - 0.025) = 0.9750$

↳ look for this value in Z-table $\Rightarrow$ 1.96

Step 4: Calculate Test Statistics

~~FSD~~ sample standard deviation

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

$s/\sqrt{n}$ → standard Error

(SD for z-test can be s & σ depending on the question)

$$z = \frac{78 - 80}{2.5/\sqrt{40}} = \frac{-2}{0.395} = -5.05$$

$z = -5.05$

wkt, for $n = 1$

$$z = \frac{\bar{x}_i - \mu}{\sigma/\sqrt{n}}$$

NOTE:

- why z-test?
  1) $n \gg 30$ & $n < 30$ standard
  2) Population deviation & sample standard deviation

- why t-test?
  1) sample standard deviation
  2) $n < 30$

M T W T F S S

Step 5: State the results

Decision Rule: If $z = -505$ is less than -1.96 & greater than 1.96, then reject the null hypothesis with 95% C.I

Result: Reject Null Hypothesis (H_0)

{There is some fault in the machine with 95% C.I}

Q) In the population the average IQ is 100 with a standard deviation of 15. A team of scientists wants to test a new medication to see if it has a +ve & -ve effect, & no effect at all. A sample of 30 participants who have taken the medication has a mean of 140. Did the medication affect Intelligence? 95% C.I (2 Tail Test)

Step 1:

$H_0: \mu = 100$

$H_1: \mu \neq 100$

$n = 30$

$\sigma = 15$

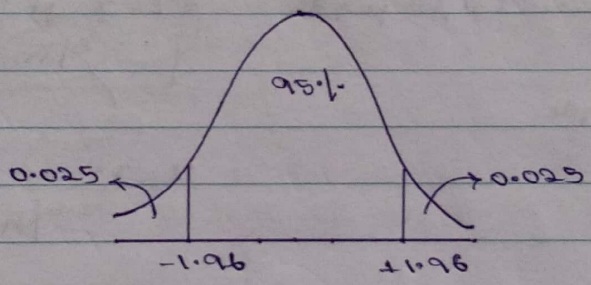
$\bar{x} = 140$

Step 2:

C.I = 95%

$\alpha = 0.05$

Step 3:



• use z-score table

$(1 - 0.025) = 0.9750$

↓

1.96

Step 4: z-test

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{140 - 100}{15 / \sqrt{30}} = \frac{40 \times \sqrt{30}}{15} = 14.60$$

$z = 14.60$

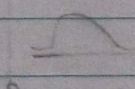
Step 5:

Decision Rule: If z is less than -1.96 & greater than 1.96 Reject the H_0

Result: $14.60 > 1.96$ Reject the Null Hypothesis

↓

+ve impact increasing IQ



M T W T F S S

Q) A complain was registered, the boys in the Municipal Primary School are underfed. Average weight of boys of age 10 is 32 kgs with standard deviation is 9 kgs. A sample of 25 boys was selected from the municipal school and the average weight was found to be 29.5 kgs? With C.I = 95%, Check whether it is True or False? (1 Tail Test)

Step 1:

$H_0: \mu = 32$

$H_1: \mu < 32$

$\sigma = 9$ kg

$n = 25$

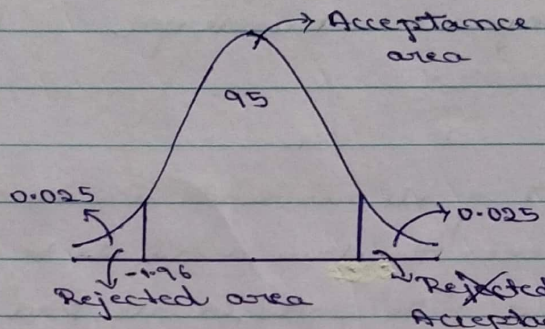
$\bar{x} = 29.5$

Step 2:

C.I = 95%

$\alpha = 0.05$

Step 3:



Z-table gives 0.9750 → 1.96

=> Since $\mu < 32$ (H_1) (1 Tail test)

Step 4:

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{29.5 - 32}{9/\sqrt{25}} = \frac{-2.5 \times 5}{9} = -1.38$$

Step 5:

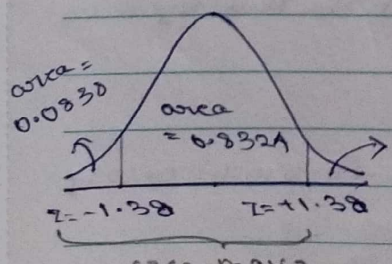
Decision Rule: If z is less than -1.96 (for 1 Tail Test) ~~or greater than 1.96~~, Reject Null Hypothesis.

Result: $-1.38 > -1.96$, Therefore we accept the Null Hypothesis

∴ The boys are not underfed.

⇝ We all know Significance value α, but what about p-value?

$\boxed{\alpha = 0.05}$ ⇒ p value = ?



∴ $1 - 0.9162 = 0.0838$

p-value = 0.0838 + 0.0838

$\boxed{\text{p-value} = 0.1676}$

For z = 1.38, from z-table gives value 0.9162 (It is the area to the left of +1.38)

the process of find

Is p-value same for both 1 tail & 2 tail test?

For all the problems, we found out H_0 is accepted or not using z value & value obtained from z_{new} - test value

We can also find H_0 is accepted or not using p-value & significance value.

p-value calculated here is WRT 2 tail

$p = 0.1676$, $\alpha = 0.05$ — Here $0.1676 > 0.05$

i.e., if $p \geq \alpha$ we accept the H_0

∴ The Boys are not underfed.

NOTE:
p-value is calculated based on new z-value

Q) The average weight of all residents in town XYZ is 168 lbs. A nutritionist believes the true mean to be different. She measured the weight of 36 individuals and found the mean to be 169.5 lbs with a standard deviation of 3.9.

[a] At 95% C.I is there enough evidence to discard the Null Hypothesis? (2 Tail Test)

Step 1:

$H_0: \mu = 168$

$H_1: \mu \neq 168$

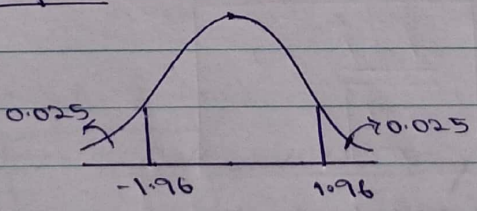
$n = 36$, $\bar{x} = 169.5$, $s = 3.9$ — sample.

Step 2:

$\alpha = 0.05$

C.I = 95%

Step 3:



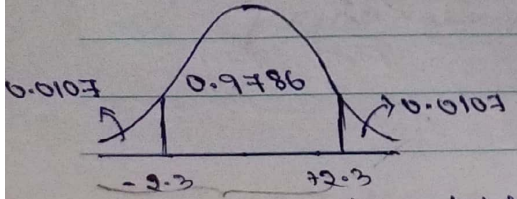
Step 4:

$$z = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{169.5 - 168}{3.9/\sqrt{36}} = \frac{1.5 \times 6}{3.9} = 2.3$$

Step 5:

As $2.3 > 1.96$, Reject the Null Hypothesis

Step 6: WRT p value.



↳ z-table = 0.9893

p-value = 0.0107 + 0.0107 = 0.0214

$0.0214 < 0.05$

∴ Reject the Null Hypothesis

[b] t-test

M T W T F S S

Q) A company manufactures bike batteries with an average life span of 2 or more years. A engineer believes this value to be less. Using 10 samples, he measures the average life span to be 1.8 years, with a standard deviation of 0.15.

[a] State the Null & Alternate hypothesis

[b] At a 99% C.I, is there enough evidence to discard the H_0?
(1 Tail Test)

Step 1:

$H_0: \mu \geq 2$

$H_1: \mu < 2$

$n = 10$

$\bar{x} = 1.8$ } sample

$s = 0.15$

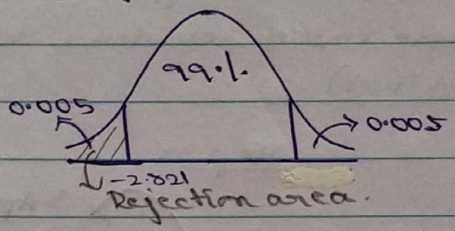
Step 2:

C.I = 99%

$\alpha = 0.01$

0.005 0.005

Step 3:



$1 - 0.0005 = 0.9995$

[T-test as $n < 30$ & σ is given]

T-test

- Degree of freedom $\Rightarrow n - 1 = 10 - 1 = 9$.

$\alpha = 0.01$

Value in t-table = 2.821

Step 4: Calculate t-test statistics

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

$$t = \frac{1.8 - 2}{0.15/\sqrt{10}} = \frac{-0.2 \times \sqrt{10}}{0.15} = -4.216$$

Step 5:

$-4.216 < -2.821$, Reject the Null hypothesis.

P-value can be calculate for this too.
This value is high & not within 3rd standard deviation.
So well not be able to see in the table

M T W T F S S

[C] Z-test with proportions (Proportions means ratio)

(No standard deviation required here)

Q) A tech company believes that the percentage of residents in town XYZ, that owns a cell phone is 70%. A marketing manager believes that this value to be different. He conducts a survey of 200 individuals and found that 130 responded yes to owning a cell phone.

[a] State the Null and Alternate hypothesis.

[b] At a 95% C.I, is there enough evidence to reject the H_0?
(2 Tail Test)

Step 1: > percentage

$H_0: p_0 = 0.70$

$H_1: p_0 \neq 0.70$

$n = 200$
$x = 130$
] calculate proportion $(\hat{p})$ WRT this.

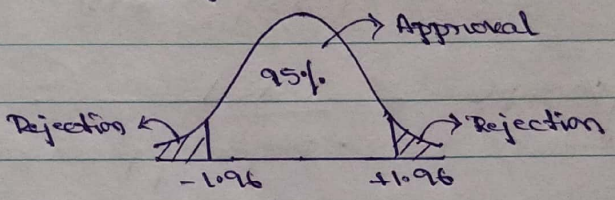
$\hat{p} = \frac{x}{n} = \frac{130}{200} = 0.65 \Rightarrow 65\%$

Step 2:

C.I = 95%

$\alpha = 0.05$

Step 3:



Step 4: Calculate Z-test, with proportion value

$$Z_{test} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$$

where, $q_0 = 1 - p_0$

$$Z = \frac{0.65 - 0.70}{\sqrt{\frac{0.70 \times 0.3}{200}}}$$

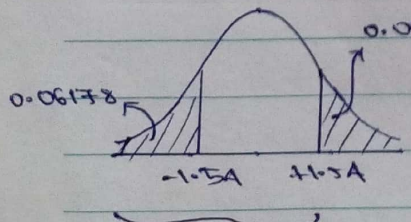
$$= \frac{-0.05}{0.0324} = -1.54$$

$\therefore Z = -1.54$

Step 5:

At 95% C.I there is $-1.54 > -1.96$,
so we Accept the Null hypothesis.

Step 6: p-value



Z table $\Rightarrow 1.54 \rightarrow 0.93822$

$1 - 0.93822 = 0.06178$

$\Rightarrow$ P-value $= 0.06178 + 0.06178 = 0.12356$

$\therefore p > \alpha \Rightarrow$ Accept H_0

M T W T F S S

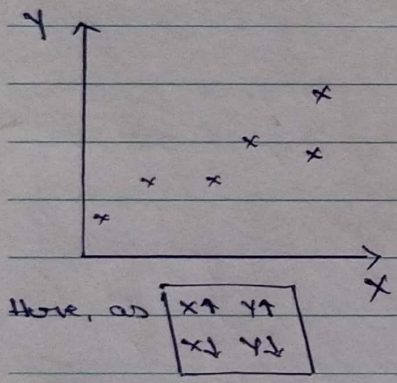
Day 1

Covariance (of numerical variables)

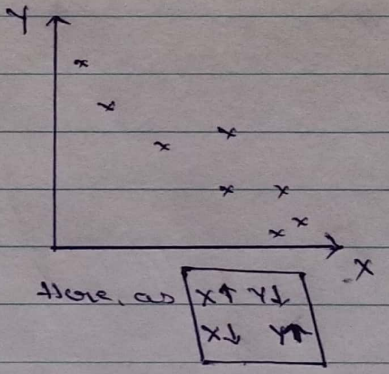
Consider 2 features X & Y

X	Y
..	..
...	...
....	....
.....	.....
.....	.....

- We need to quantify the relationship b/w X & Y (Meaning what is the exact relationship b/w X & Y)
- When X↑ is Y↓
  X↑ is Y↑
  X↓ is Y↓
  X↓ is Y↑
  & vice versa



Here, as X↑ Y↑ / X↓ Y↓



Here, as X↑ Y↓ / X↓ Y↑

$$\text{Cov}_{x,y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

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

$\bar{y}$ = mean of y

NOTE $\text{Var}_x = \frac{\sum (x_i - \bar{x})^2}{N-1}$

Relation b/w Cov(x,y) and variance(x) is:

$$\text{Cov}(x,x) = \frac{\sum (x_i - \bar{x})^2}{n-1} = \text{Var}(x)$$

M T W T F S S

- Covariance gives relationship b/w 2 variables

- If $x\uparrow$ $y\uparrow$ ⇒ Positively correlated

Example:

x	y
2	3
4	5
6	7
$\bar{x}=4$	$\bar{y}=5$
$(\frac{12}{3})$	$(\frac{15}{3})$

$$cov(x,y) = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{N-1}$$

$$= \frac{(2-4)*(3-5)+(4-4)(5-5)+(6-4)(7-5)}{(3-1)}$$

$$= \frac{(-2)(-2)+0+(2)(2)}{2} = \frac{8}{2} = 4$$

$cov(x,y) = 4$ ⇒ positive value

∴ Positively correlated

Why $x\uparrow$ $y\uparrow$ ⇒ positive value

AND

$x\uparrow$ $y\downarrow$ } $cov(x,y)$ Negative value

$x\downarrow$ $y\uparrow$ } ⇓

Negatively correlated

Conclusion:

Case 1: $cov(x,y) = +ve$

Positive correlation

Case 2: $cov(x,y) = -ve$

Negative correlation

Disadvantages of covariance:

- $cov(x,y)$ ⇒ Value will be either

+ve value or -ve value ⇒ there is no limit to the value like, (500 / -400 / -300 / 1000 ...) so it may go to infinity

⇒ Aim is to restrict wrt the covariance, with some specific value

Lets say -1 to +1

⇓

This is done by another correlation called as Pearson correlation co-efficient

M T W T F S S

Pearson Correlation Co-efficient (For Numerical Values)

→ Ecta symbol called rho

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

→ because of this the value gets restricted b/w [-1 to 1]

σ_x = Standard deviation of x

σ_y = Standard deviation of y

- More the value towards 1 more +ve it is correlated.

- Why do we use Pearson correlation ?

Let say we have a dataset with 1000 features

features.

x y z A B C o/p

↳ we will not use all the features in ML models

↓

It is not possible due to

'curse of dimensionality' (As features ↑ cause problems, detailed in ML).

which all feature should be selected to predict the output (o/p) feature

↓

(Dependent feature)

x, y, z, A, B, C

↓

(Independent features)

→ We are obviously supposed to select the most important feature that will definitely help us to predict the output

→ To find the relationship b/w x & o/p, y & o/p, ...

⇓

If they are highly correlated (either +ve or -ve)

keep those features

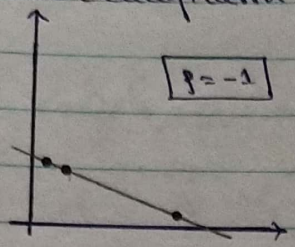
→ Now correlation b/w x & y if it's 99%,

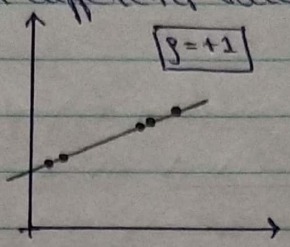
then definitely delete one of the feature x or y as they are similar.

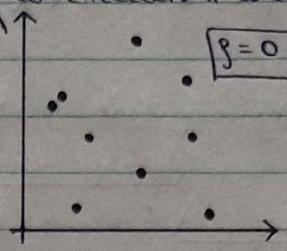
→ If $\rho = 0$ (not at all correlated to output) remove that feature

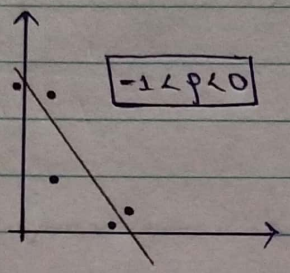
M T W T F S S

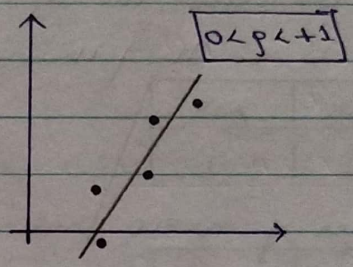
Scatter diagrams with different values of correlation coefficient (ρ)











Spearman Rank Correlation (For Numerical Values)

(graph... disadvantage of Pearson correlation Rank kaha)

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

R(x) = Rank of x

R(y) = Rank of y

Consider,

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

Rank = 1, for highest values

Rank = same, for repeated values

Spearman graph

M T W T F S S

d) Chi square Test (χ^2)

- The chi square test claims about population proportions.
- It is a non parametric test that is performed on categorical (numerical & ordinal) data.

Q) In the 2000 U.S census, the ages of individuals in a small town were found to be the following.

<18	18-35	>35
20%	30%	50%

In 2010, ages of $n=500$ individuals were sampled. Below are the results.

<18	18-35	>35
121	288	91

Using $\alpha = 0.05$, would you conclude the population distribution of ages has changed in the last 10 years?

$n=500$

	<18	18-35	>35
Expected (As it is calculated WRT $n=500$)	20%	30%	50%
	100	150	250
Observed	121	288	91

- $\frac{20}{100} \times 500 = 100$
- $\frac{30}{100} \times 500 = 150$
- $\frac{50}{100} \times 500 = 250$

Step 1:
H_0: The data meets the expected distribution
H_1: The data does not meet the expected distribution.

Step 2:
C.I = 95%
$\alpha = 0.05$

Step 3: Calculate the degree of freedom = (3 categories)
$df = n - 1 = 3 - 1 = 2$

In Chi-square table look for value using df and α.
Now, $df = 2$, $\alpha = 0.05$, value = 5.991

Step 4: Decision
If χ^2 is greater than value (5.991) then, Reject H_0.
Otherwise accept H_0

M T W T F S S

Step 5:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

$$\chi^2 = \frac{(121-100)^2}{100} + \frac{(288-150)^2}{150} + \frac{(91-250)^2}{250}$$

$$\chi^2 = \frac{441}{100} + \frac{19044}{150} + \frac{25281}{250}$$

where f_o = observed value

f_e = Expected value

$$\chi^2 = 4.41 + 126.96 + 101.124$$

$$\chi^2 = 232.494$$

Step 6:

As $232.494 > 5.991$, Reject the Null Hypothesis.

Q) A school principal would like to know which days of the week students are most likely to be absent. The principal expect the students will be absent equally during the 5-day school week. The principal selects a random sample of 100 teachers asking them which day of the week they had the higher number of student absences. The observed and expected results are shown in the table below. Based on the results, do the days for the highest number of absentes occur with equal frequencies (C.I = 95%.)

	Monday	Tuesday	Wednesday	Thursday	Friday
Observed	23	16	14	19	28
Expected	20	20	20	20	20

$n = 100$ (sample)

Step 1:

H_0 : The data meets the expected distribution

H_1 : The data does not meet the expected distribution

Step 2:

C.I = 95%

$\alpha = 0.05$

Step 3: n(categories)

$df = n - 1 = 5 - 1 = 4$

Chi-square table $df = 4$, $\alpha = 0.05$, value = 9.488

Step 4:

If χ^2 greater than 9.488, Reject H_0 ; if χ^2 less than 9.488 accept H_0

M T W T F S S

Step 5:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = \frac{(23-20)^2}{20} + \frac{(16-20)^2}{20} + \frac{(14-20)^2}{20} + \frac{(19-20)^2}{20} + \frac{(28-20)^2}{20}$$

$$= \frac{9}{20} + \frac{16}{20} + \frac{36}{20} + \frac{1}{20} + \frac{64}{20}$$

$$= \frac{126}{20} = 6.3$$

$\therefore \chi^2 = 6.3$.

Step 6:

As $6.3 < 9.488$, Accept the Null Hypothesis.

[e] ANOVA (F-Test) [Inferential Statistics]

ANOVA basically means Analysis Of Variance.

- ANOVA is a statistical method used to compare the means of 2 or more group.

- Two important things in ANOVA

[a] Factors (also called variables)

[b] Levels (are subcategories)

Example: 1) Consider Medicine for reducing Anxiety with different dosages.

0 mg	50 mg	100 mg
9	7	4
8	6	3
7	6	2
8	7	3
6	8	6

Here,

Factor: Dosage

Levels: 0 mg, 50 mg, 100 mg

Example: 2) Factor: Gender

Levels: Male, Female

M T W T F S S

- Types of ANOVA:

1) One way ANOVA:

One factor with atleast 2 levels, levels are independent.

Example: 1) Medicine...

2) Repeated Measures ANOVA:

One factor with atleast 2 levels, but levels are dependent.

Example: a) Running & covering certain distance each day in kms by people (consider 10 ppl). / b) Study hours of certain 10 ppl on each day.

c) No. of hours spent each day in gym by 10 ppl.

	Day 1	Day 2	Day 3	Day 4
1	6	4	2	8
2	5	7	2	4
⋮	⋮	⋮	⋮	⋮
10	7	5	3	2

Here Day 1, Day 2, Day 3, Day 4 are dependent
(As a person will not have same energy every day)

3) Factorial ANOVA:

Two or more factors, with atleast 2 levels with each factor, levels can be either independent or dependent or both (mixed).

Example:

	Day 1	Day 2	Day 3
Men	9	7	4
	8	6	3
	7	5	3
Women	8	7	3
	8	8	4
	9	9	3

Gender	Days
Men	Day 1
Women	Day 2
	Day 3

[a] Levels: Men, Women
[b] Day 1, Day 2, Day 3 — Here dependent days

Factors: Gender, Days

M T W T F S S

One Way ANOVA (F-test).

Q) Researchers want to test a new anxiety medication. They split participants into 3 conditions (0 mg, 50 mg, 100 mg), then ask them to rate their anxiety level on the scale of 1-10. Are there any differences b/w the 3 conditions using $\alpha = 0.05$?

0 mg	50 mg	100 mg
9	7	4
8	6	3
7	6	2
8	7	3
8	8	4
9	7	3
8	6	2

Step 1:

H_0: $\mu_{0\,mg} = \mu_{50\,mg} = \mu_{100\,mg}$

H_1: Not all means are equal

Step 2:

$\alpha = 0.05$

C.I = 95%.

Step 3: Calculate the degree of freedom.

$df_{Between} = a - 1$

$df_{within} = N - a$

$df_{Total} = df_{Between} + df_{within}$

$= N - 1$

where,

- N = Population
  Combining all the datapoints in a dataset
- n = Sample
  No. of datapoints in a sample in each category/level
- a = no. of levels.

M T W T F S S

Here, $N = 7 \times 3 = 21$; $n = 7$; $a = 3$

$df_{Between} = a - 1 = 3 - 1 = 2$

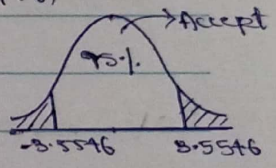
$df_{within} = N - a = 21 - 3 = 18$

$df_{Total} = N - 1 = 21 - 1 = 20$ (this is nothing but $2 + 18$)

Step 4: State Decision Rule

$df_{Between} = df_1$
$df_{within} = df_2$
α value
} Required for F-table

$df_1 = 2$
$df_2 = 18$
$\alpha = 0.05$
] Decision Rule range (2, 18)



F-table value = 3.5546

If F test is greater than F-table/~~test~~ value, Reject Null Hypothesis, if lesser than F-~~table~~ value, Accept Null Hypothesis.

Step 5: Calculate F-Test Statistics.

	SS	df	MS	F Test
Between	98.67	2	49.34	86.56
Within	10.29	18	0.57	
Total	108.96	20		

[a] Sum of squares: SS

- $SS_{Between} = \frac{\sum(\sum a_i)^2}{n} - \frac{T^2}{N}$ where, $T = \left[\sum(\sum a_i)\right]^2$

- $SS_{within} = \sum y^2 - \frac{\sum(\sum a_i)^2}{n}$ where, y^2 = Each & every value square

M T W T F S S

[b] Degree of Freedom (df)

[c] Mean square (MS)

$$MS = \frac{SS}{df}$$

[d] F Test

$$F = \frac{MS_{Between}}{MS_{Within}}$$

Calculations:

[a] SS

- $SS_{Between} = \frac{\Sigma(\Sigma a_i)^2}{n} - \frac{T^2}{N}$

$$= \frac{57^2+47^2+21^2}{7} - \frac{(57+47+21)^2}{21}$$

$SS_{Between} = 98.67$

$\Sigma(\Sigma a_i)^2 = (9+8+7+8+8+9+8)^2 + (7+6+6+7+8+7+6)^2 + (4+3+2+3+4+3+2)^2$

$\Sigma(\Sigma a_i)^2 = 57^2+47^2+21^2$

- $SS_{Within} = \Sigma y^2 - \frac{\Sigma(\Sigma a_i)^2}{n}$

$= 853 - 842.7$

$SS_{Within} = 10.29$

$\Sigma y^2 = 9^2+8^2+7^2+8^2+8^2+9^2+8^2+$
$7^2+6^2+6^2+7^2+8^2+7^2+6^2$
$4^2+3^2+2^2+3^2+4^2+3^2+2^2$

$\Sigma y^2 = 853$

Step 6:

As $86.56 > 3.5516$, We Reject the Null Hypothesis.