Statics

Statistics

Statistics is the Science of collecting, Organizing and analyzing data.

Data:- "fact or pieces of information"

Types of Statistics:

① Descriptive statistics

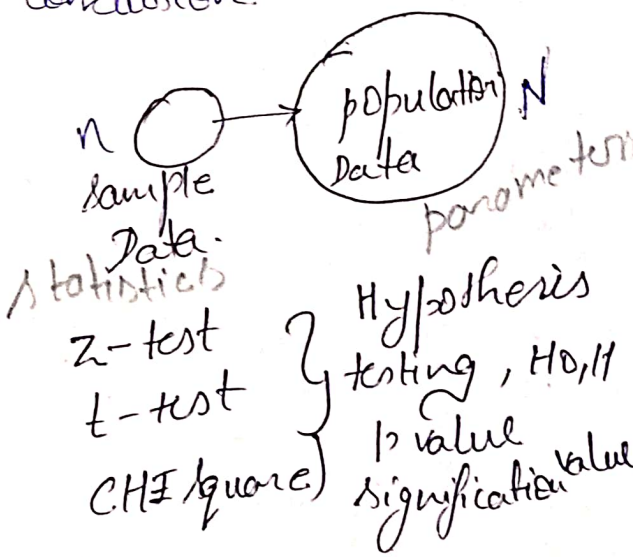
⇒ It consists of Organizing and Summarizing data.

Sub Types:-

* Measure of Central Tendency [mean, median, Mode]
* Measure of Dispersion [Variance, std]
* Different type of Distribution of data. [Histogram, pdf, pmf]

② Inferential Statistics

→ It consists of Using data you have measured to form conclusion.

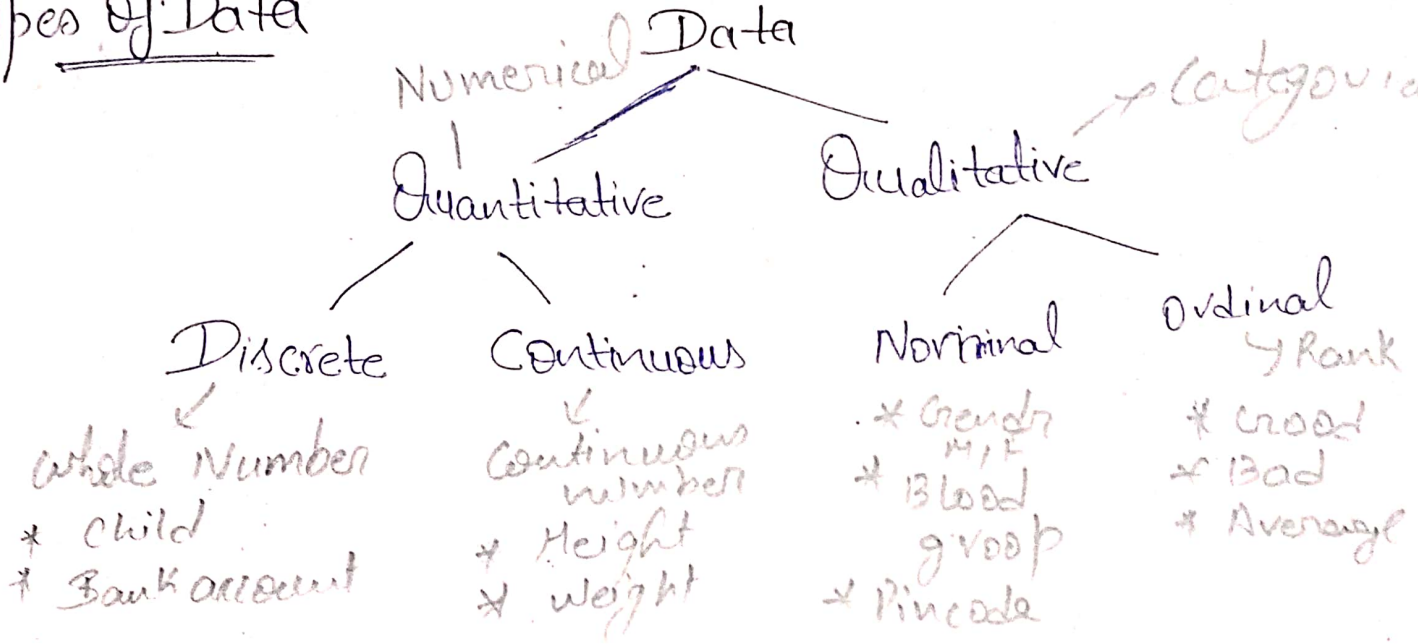


Population and Sample Data.

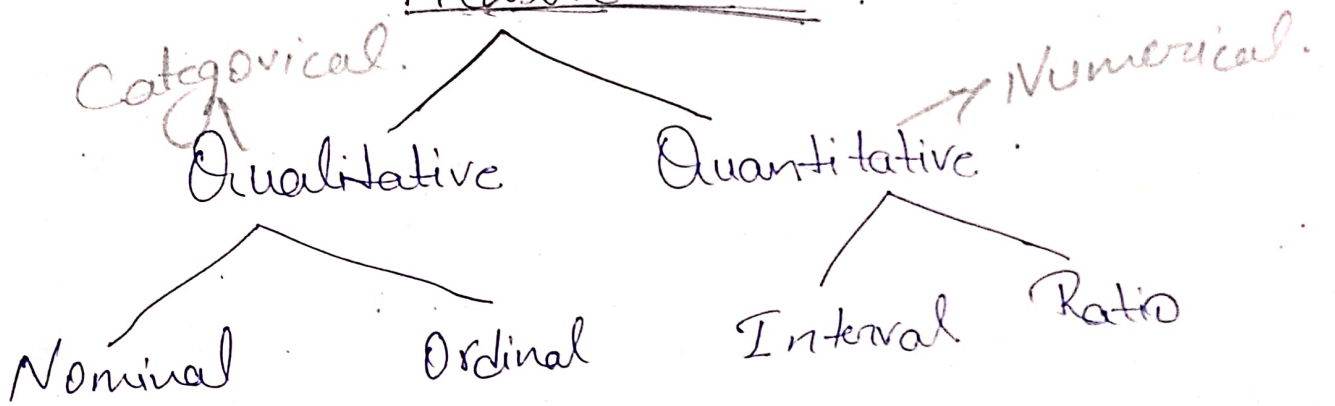
Population:- the group you are intrested in studing

Sample: a subset of population.

Types of Data



Measurent level



* **Nominal Scale Data**
  * Order does not matter
  * Gender, colors, labels

eg: favorite color
Red → 5
Blue → 3
Orage → 2
50%
30%
20%

* **Ordinal Scale Data**
  * Ranking is important
  * Order matter
  * Difference cannot be measured

eg: 1 → Best
2 → Good
3 → Bad
Differ
Differ

* **Interval Data** (context)
  * The order matter
  * Difference can be measured
  * Ratio cannot be measure
  * No true "0" starting point

exception
0K = -273°d

eg: Temperature
30F
60F
120F
30 di
60 di
3:1 ×
−∞ ∞+

* Ratio Scale Data (contact)

* The order matter
* Difference are measurable (including ratio)
* Contains a "0" starting point → 0 mark, starting no negative

0°K -> -273°C

eg: Student marking class

1st mark = 90, 60, 30, 10, 0

ratio = 90 : 30 = (3 : 1)

Measure of Central Tendency: (Descriptive statistics)

* Mean, * Median, * Mode

(i) Mean:

Population (N)

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

Sample (n)

$$\text{Sample mean} = \sum_{i=1}^{n} \frac{x_i}{n}$$

(ii) Median:

* Sort the Random variable.
* No. of element. if count is even then (mid + mid+1)/2 else mid
* It is used to find the Central Tendency when Outliers is present.

(iii) Mode:

* frequency maximum
* for categorical [1, 1, 1, 2, 3, 3, 4, 5, 5] → 1

2] Measure of Dispersion [Spread of the Data]

* Variance * Standard deviation

i) Variance

population Variance

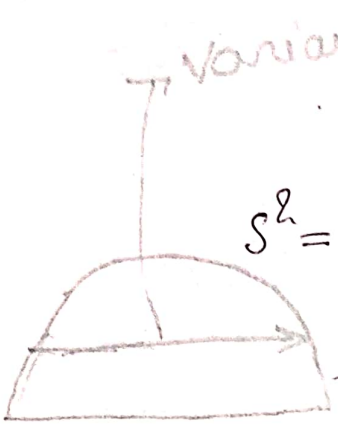
$$\sigma^2 = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N}$$

x_i = Data points

μ = population mean

N = Population size

Sample Variance.



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

x_i = Data Points

$\bar{x}$ = Sample mean

n = Sample size.

$(n-1)$ → Un biased estimator to create.

Let:

$s^2 = 2.5$

Variance spread.

$s^2 = 7.5$

2) Standard Deviation:

Population Standard Deviation

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

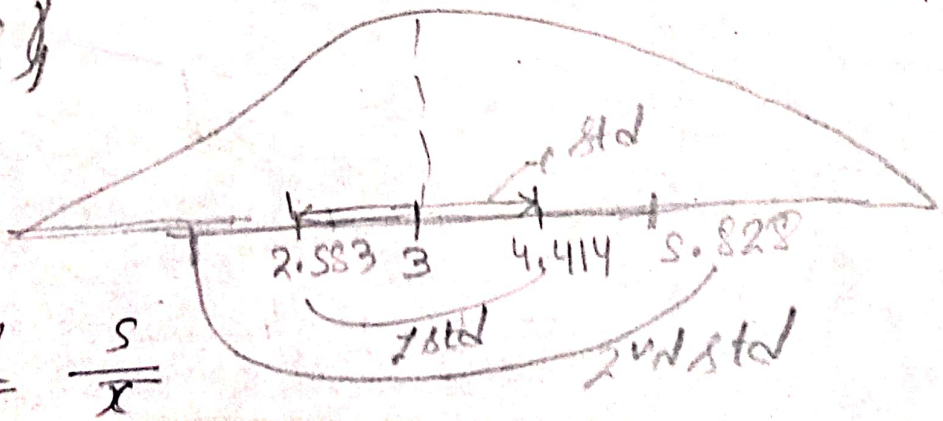
Let: {1, 2, 3, 4, 5}

$\bar{x} = 3$

$\sigma = 1.414$

Sample std

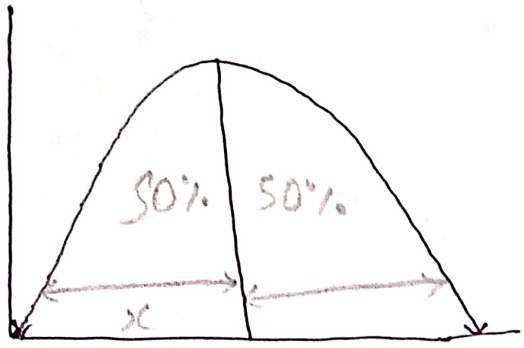
$$\text{std} = \sqrt{s^2}$$ s^2 = sample variance



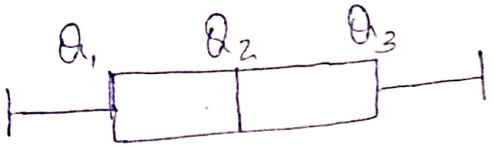
Sample ~~Standard~~ deviation coefficient of variation = $\frac{s}{\bar{x}}$

Skewness: measure of the asymmetry

* Normal / Gaussian / Symmetrical Distribution

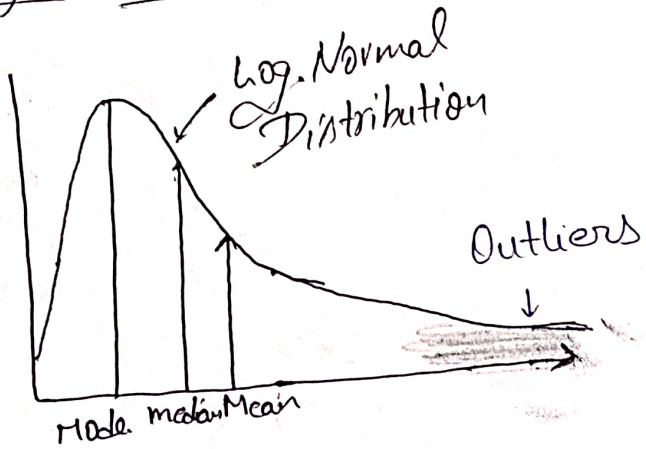


* No Skewness
* Mean = Median = Mode

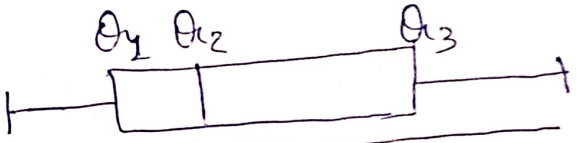


$Q_3 - Q_2 \cong Q_2 - Q_1$

* Right Skewed

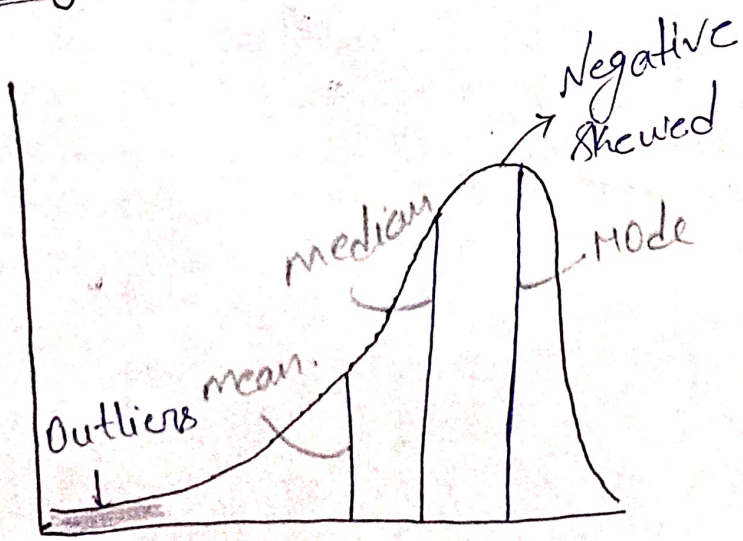


* Positive Skewed
* mean ≥ median ≥ mode
* Outliers on right side

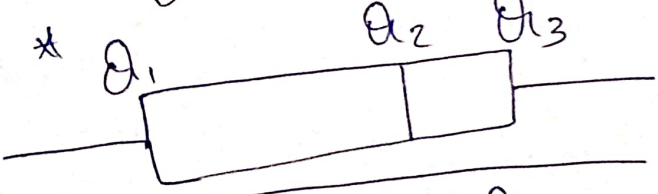


$Q_3 - Q_2 \geq Q_2 - Q_1$

* Left Skewed Distribution



* mean ≤ median ≤ mode
* Outliers on left side
* Negative Skewed



$Q_2 - Q_1 \geq Q_3 - Q_2$

Covariance (Direction only)

Covariance measure the direction of the linear relationship between two variable.

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

$$S_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1} \rightarrow \text{Sample formula}$$

* Covariance gives a sense of direction
  * (>0) greaterthan the two variables move together
  * (<0) lessthen the two variables move in opposite directions
  * (=0) the two variables are independent.
* Covariance is not standardized, so it value depends on the scale of the variable eg:. ($-\infty$ to ∞)

Correlection: coefficient Indicates direction & strength.

* Standardized (unit-free)
* -1 to $+1$
* Indicates direction & strength
* use to compare relationships across datasets with different scales.

Correlation range [-1 to 1]
* 1 perfect positive corr
* -1 perfect negative
* 0 No linear relation

$$\text{Corr}(x,y) = \frac{\text{Cov}(x,y)}{\sigma_x, \sigma_y}$$

σ_x, σ_y are the standard deviation of x and y

Types: * Pearson's → same
* Spearman's Rank → $\frac{\text{cov}(x,y)}{\sigma(R(x)) \cdot \sigma(R(y))}$

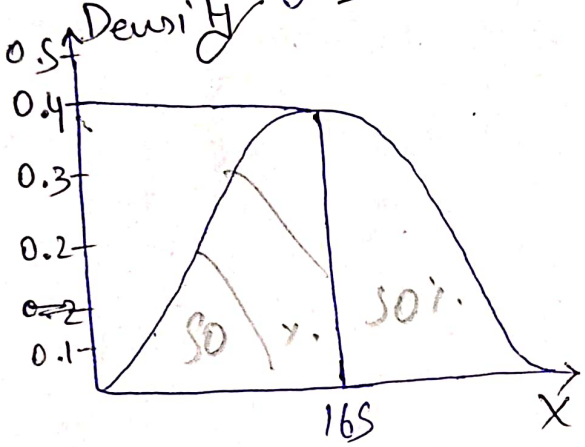
Advanced

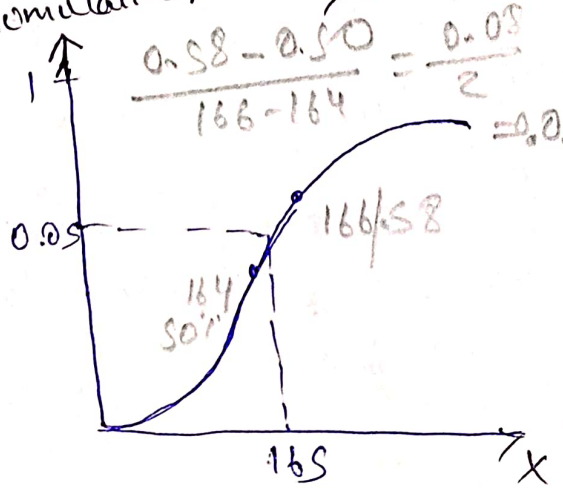
Probability Distribution Function / Density function

Pdf / Pmf / CDF (Cumulative Distribution function)

* Probability Density function (Pdf) / (CDF)

* Distribution of Continuous Random variable

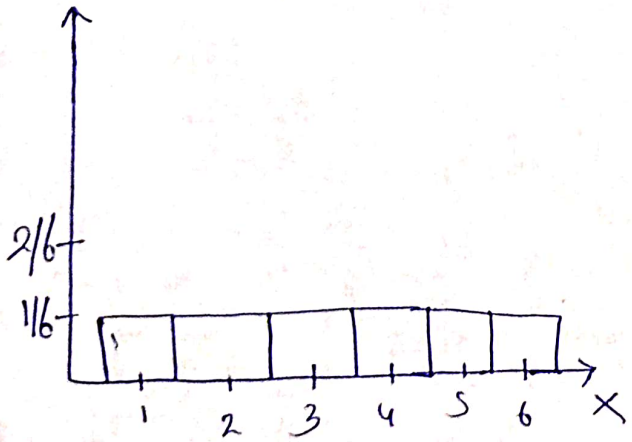


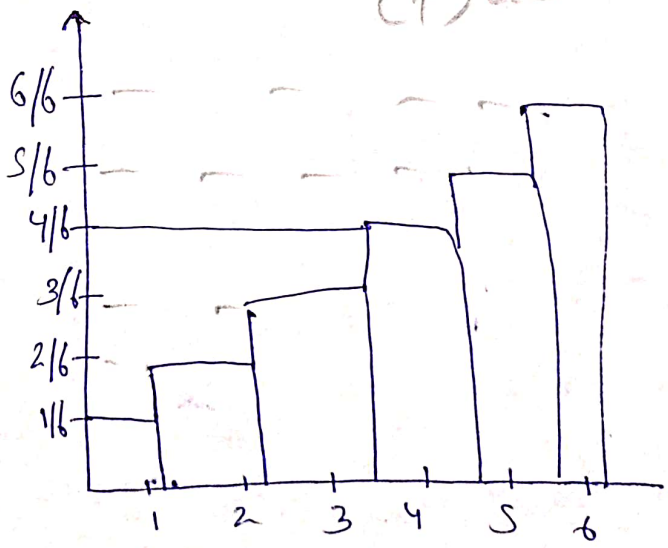


* Probability Mass function:

* Discrete Random Variable.

Eg: Rolling Dice





Types of Probability Distribution:

(i) Normal / Gaussian Distribution (pdf) → Continuous
(ii) Bernoulli Distribution (pmf) → Discrete / binary
(iii) Uniform Distribution (pmf)
(iv) Poission Distribution (pmf)
(v) Log Normal Distribution (pdf)
vi) Binomial Distribution (pmf)

* Bernoulli Distribution:

* Discrete Random Variable (pmf)
* Outcomes are Binary (0, 1)
* $k = \{0, 1\}$

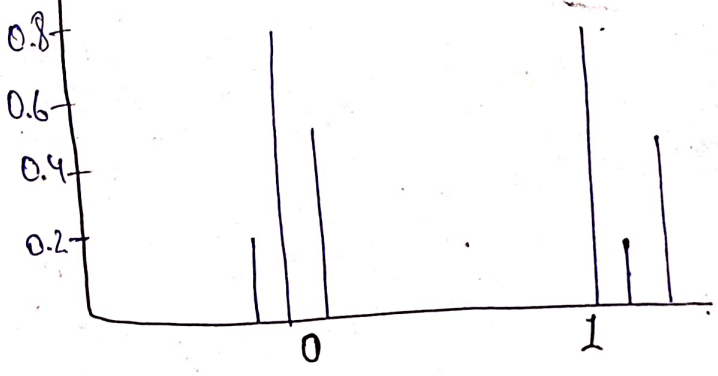
eg: Tossing a coin (H, T)

$Pr(H) = 0.5 = p$
$Pr(T) = 0.5 = 1 - p = q$

Let whether the person will pass/fail

$Pr(Pass) = 0.7 = p$
$Pr(fail) = 1 - 0.7 = 0.3$
$0.3 = q$
$1 - p = q$

* Variance = pq
* Std = $\sqrt{pq}$
* mean = p
* Median = $\begin{cases} 0 & \text{if } p < 1/2 \\ 0,1 & p = 1/2 \\ 1 & p > 1/2 \end{cases}$



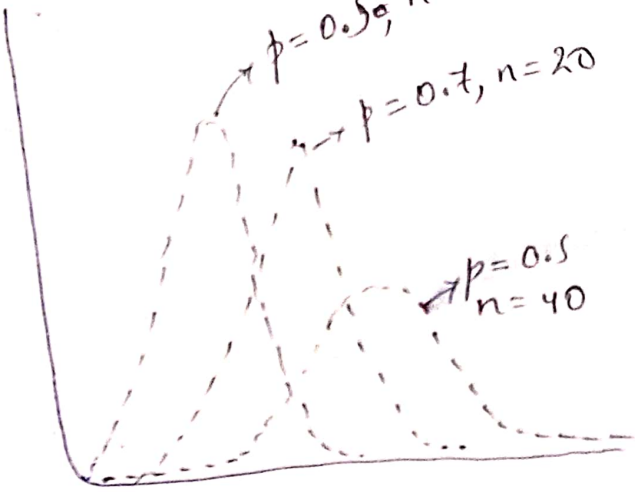
Three examples of Bernoulli Distribution

$P(X=0) = 0.2$ and $P(X=1) = 0.8$
$P(X=0) = 0.8$ and $P(X=1) = 0.2$
$P(X=0) = 0.5$ and $P(X=1) = 0.5$

$PMF = p^k * (1-p)^{1-k} \rightarrow (k = 0, 1)$

* Binomial Distribution: (pmf), $\{B(n,p)\ n=\{1,\dots n\}$

* A single sucess/failure experiment is also called a Bernoulli trial Distribution and sequence of outcomes is said to Binomial
* Sequence of independent Bernoulli trail can collectively Binomial distribution.



* Discerete Random Variable
* Every experiment outcome is binary
* These experiment/trail is performed for n times

eg: Tossing a coin 10 times.

Support: $K \in \{0,1,2,3\dots n\}$ → Number of sucess

PMF: $Pr(K,n,p) = {}^{n}C_{K}\, p^{K}(1-p)^{n-K}$

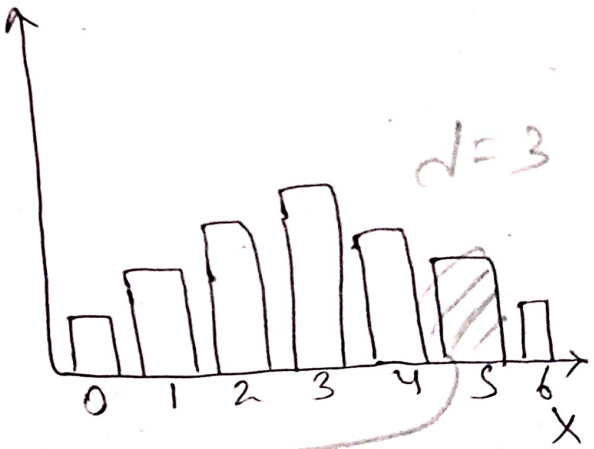
Mean → np
Variance → npq
std → $\sqrt{npq}$

---

* Poisson Distribution: (pmf)

* Discrete Random Variable
* Describe the number of event occuring in a fixed time interval.

eg: No. of people visiting hospitial every hour.



* λ → Expected No. of events to occur at every time interval.
* T = Time interval

$$PMF = P(x=-) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

$$P(x=5) = \frac{e^{-3}3^{5}}{5!} \quad (\lambda=3)$$

$$= 0.101$$

$$= 10.1\%$$

mean = $\lambda * t$

variance = $\lambda * T$
$= \lambda * t$

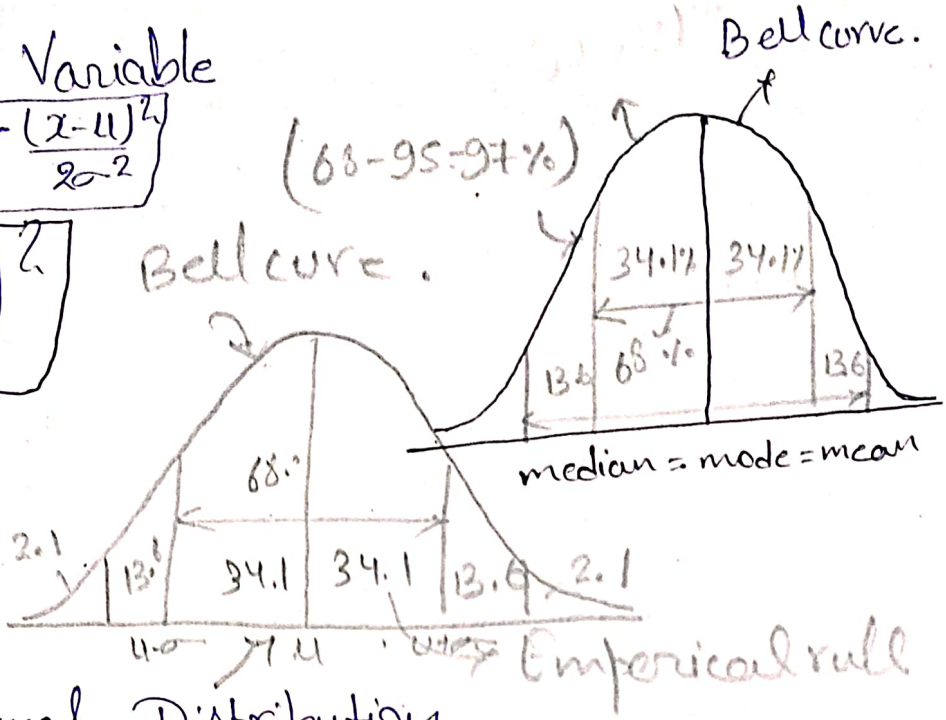
* Normal / Gaussian distribution (pdf),

* Continuous Random Variable

P.D.F = $\frac{1}{\sqrt{2\pi\sigma^2}} \times e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

$\frac{1}{\sigma\sqrt{2\pi}} e^{-1/2\left(\frac{x_i-\mu}{\sigma}\right)^2}$

* mean = μ = Average
* Variance = σ^2
* Std = σ



Emperical rule of Normal Distribution
⇒ (68-95-97)%.

eg: Weight of the Student in the class
Height " " " "

---

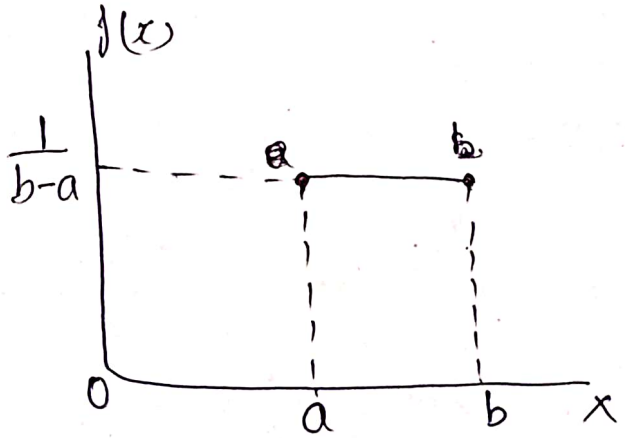
*Uniform Distribution

* Continuous Uniform Distribution (pdf)
* Discrete Uniform Distribution (pmf)

Continuous Uniform Distribution (pdf)

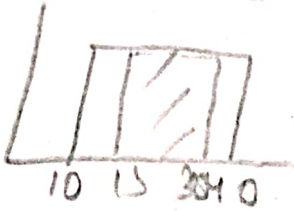
* Continuous random Variable
* minimum & maximum values (a,b)
* $-\infty < a < b < \infty$

$$pdf = \begin{cases} \frac{1}{b-a} & \text{for } x \in [a,b] \\ 0 & \text{otherwise} \end{cases}$$

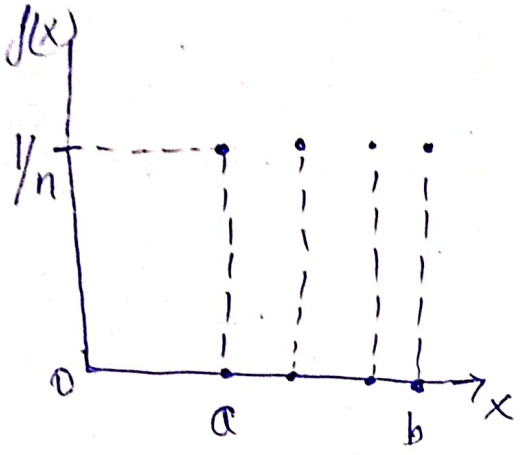


Mean = $\frac{1}{2}(a+b)$
Median = $\frac{1}{2}(a+b)$
Variance = $\frac{1}{2}(b-a)^2$

Q: max 40 & min 10
(i) fall between 15 & 30
$P(15 \le x \le 30) = (x_2 - x_1) * \frac{1}{b-a}$
$= 15 \times \frac{1}{30} \Rightarrow 50\%$



* Discrete Uniform Distribution



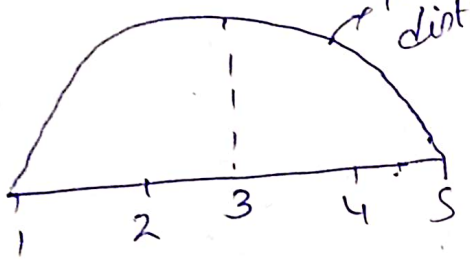
* finite number of outcomes equally likely to happen.
* Discrete Uniform Distribution
eg: Rolling a dice → {1/6}
Notation → U(a,b)
PMF ⇒ 1/n

mean, median = $\frac{a+b}{2}$

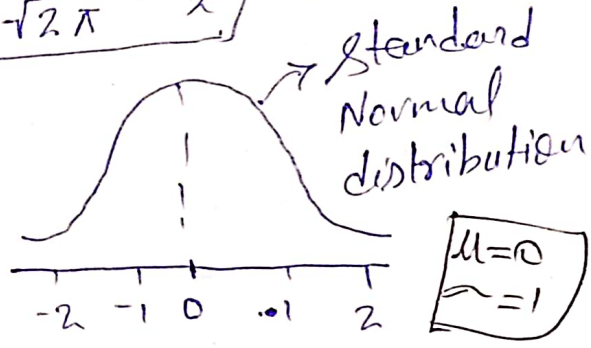
Standard Normal Distribution and Z-Score

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

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

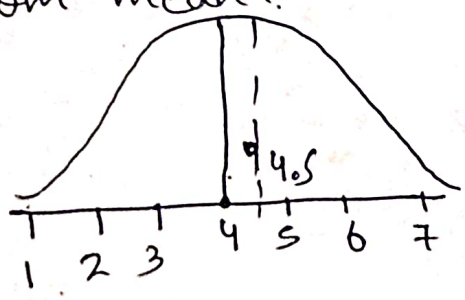


$\mu = 0$
$\Rightarrow$
$\sigma = 1$



* To convert Normal distribution to Standard Normal distribution we use Z-Score. $= \frac{x_i - \mu}{\sigma}$

Q1: How many Standard deviation 4.5 is away from mean?

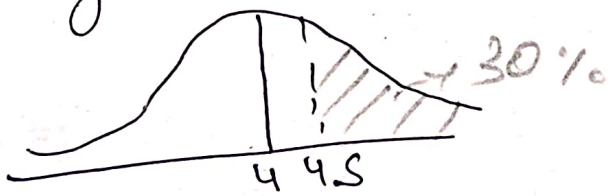


$x_1 = 4.5$
$\mu = 4$
$\sigma = 1$

$So, = \frac{4.5 - 4}{1}$
$= 0.5$

Using Z-Score table

Q. what percentage of data is falling above 4.5

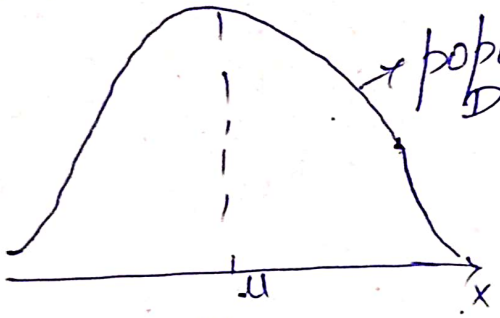


So, Z-Score $= \frac{4.5 - 4}{1} = 0.5$

using Z-Score table
$= 1 - 0.69146$
$= .3085$
$= 30\%$

Central Limit Theorem (CLT) + Always Normal distribut

① $X \simeq N(\mu, \sigma)$



taking sample (n ≥ 30)

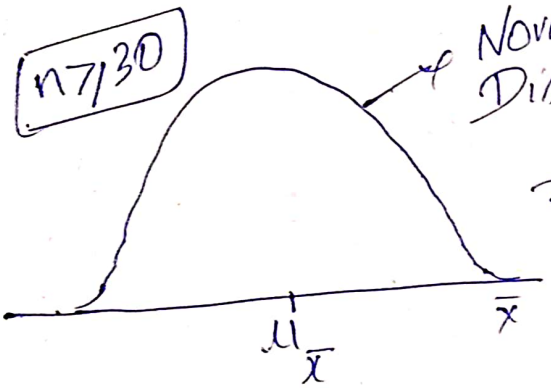
$S_1 = \{ x_1, x_2, x_3 \ldots \} = \bar{x}_1$

$S_2 = \{ x_1, x_3 \ldots \} = \bar{x}_2$

S_3

$S_{30} = \{ \ldots \} = \bar{x}_n$

So, By plotting that mean we get Normal distribution



$\bar{X} \simeq N(\mu, \frac{\sigma}{\sqrt{n}})$ ← sample error of the mean

Note (even, possion, binomial any type)

* Population if Normal distribute then N.of. sample can be any
* If population is not normal distributed then N.of. sample $n \geq 30$

Estimate:

It is an observed numerical value used to estimate an unknown population parameter.

① Point Estimate

Single numerical value used to estimate the unknown parameters

* Sample mean is a point estimate of a population mean.

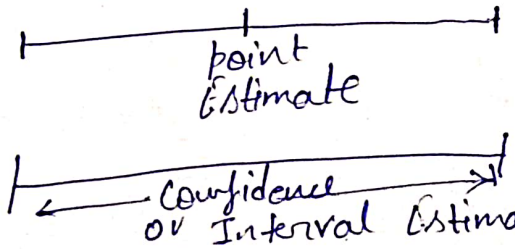
$C.I = \bar{x} + Z^{*}_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$ → population

sample mean; critical value; sample size.

② Interval Estimate.

Range of values used to estimate the unknown population parameter

* Interval estimate of population parameters are called Confidence Intervals



Pvalue

The pvalue 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. Pvalues are used in hypothesis testing to help decide whether to reject the null hypothesis

* low pvalue (<0.05): strong evidence against the null hypothesis → Reject the null hypothesis
* High p-value (>0.05) weak evidence ag--- hy---
→ fail to reject the null hypothesis

Significance value:

Significance value is also know as alpha (α) (0.05, 0.01, 0.10) 5%, 1%, 10%

Relationship b/w p-value & S.V

* If p value $\leq \alpha$, reject the null hypothesis
* If pvalue $> \alpha$, fail to reject null hypothesis

Confidence Level:

The probability that the Confidence Interval contains the true population parameter.

$C.L = (1-\alpha)$

if: Significance value (α) = 0.05 then C.I = 1 − 0.05 = 0.95

Error

Type 1 Error: We reject the Null Hypothesis when in reality it is true

Type 2 Error: We retain the null hypothesis when in reality it is false.

Hypothesis Testing and Statistical Analysis

z table → z score & p value

1. z test → } Average
2. t test →
3. Chi Square → categorical data
4. ANNOVA → variance

Z-Test (z-test is a statistical test used to determine whether there is a significant difference between sample and population mean or between two sample means)

(* population std & * n ≥ 30 ↑↑)
* data normally distributed * Sample should be independent

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

μ → population mean
$\bar{x}$ → sample mean
n → sample size

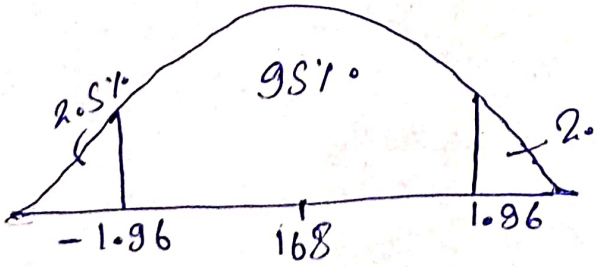
Q1

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 heights to be 169.5 cm.

(a) state null and alternate hypothesis
(b) At a 95% confidence level, is there enough evidence to reject the null hypothesis

Sol: $\mu = 168$ $\sigma = 3.9$, $\bar{x} = 169.5$, $n = 36$

a) Null Hypothesis H_0 $\mu = 168$cm
b) Alternate hypothesis H_1 $\mu \neq 168$cm (2 Tail Test)
c) C.I = 0.95 $\alpha = 1 - 0.95 = 0.05$



Note: always
* z test
* pvalue

area under the curve = 1 - 0.25 = 0.9750
Using z-score table we find (0.9750)
z-score = 1.96

d) Statistical Analysis / formula apply

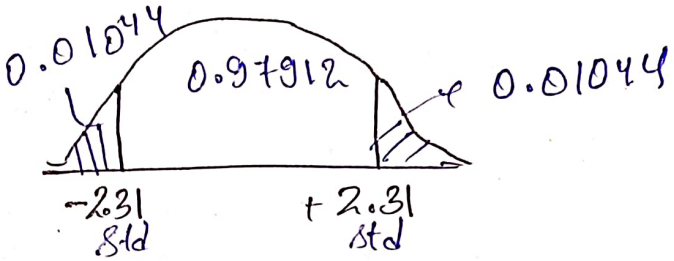
$$\text{Z-score} = \frac{x_i - \mu}{\sigma} \rightarrow \text{population}$$

$$\text{Z-test} = \frac{x_i - \mu}{\sigma/\sqrt{n}} \Rightarrow \frac{169.5 - 168}{3.9/\sqrt{36}} \Rightarrow 2.31$$

If Z-test value is less then -1.96 or greater $+1.96$ then, we Reject the Null hypothesis
else, we Accept Null Hypothesis

$2.31 > +1.96$ & we Reject the Null Hypothesis §

(2) P-value (By using)



$$\text{Z-test} = \frac{169.5 - 168}{3.9/\sqrt{36}} \Rightarrow 2.31$$

$= 1 -$ area under the curve of $+2.31 \Rightarrow 1 - 0.98956$
$\Rightarrow 0.01044$

$$\text{pvalue} = 0.01044 + 0.01044 = \underline{0.02088}$$

if p-value < Significance
$0.02088 < 0.05$ & we Reject the Null Hypothesis §

(2) A factory manufactors bulbs with a average warranty of 5 years with Standard deviation of 0.50. A worker believes that the bulb will malfunctions in less than 5 years. He tests a sample of 40 bulbs and find the average time to be 4.8 years.

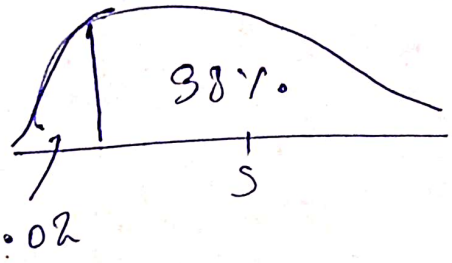
a) State null and alternate hypothesis
b) At a 2% significance level, is there enough evidence to support idea that the warranty should be revised?

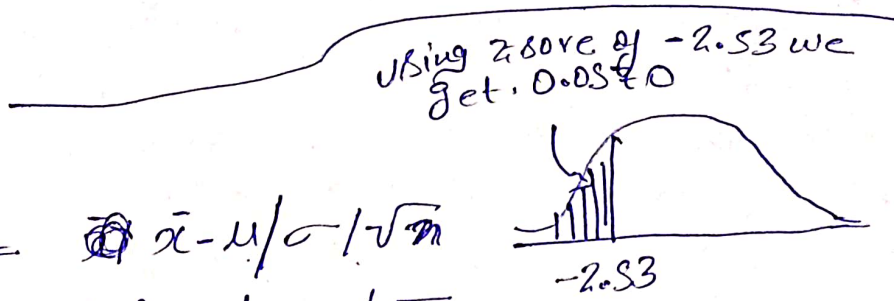
Sol: $\mu = 5$, $\sigma = 0.50$, $n = 40$, $\bar{x} = 4.8$

1) Null Hypothesis → $H_0 : \mu = 5$
2) Alternative Hypothesis $H_1 : \mu < 5$ & 1 tail Test
3) Desicision Boundary

C.I = 0.98
$\alpha = 1 - 0.98 = 0.02$

Or
C.I = 1 − 0.02 = 0.98





4) Pvalue → z. Test = $\bar{x} - \mu / \sigma / \sqrt{n}$

$= 4.8 - 5 / 0.50 / \sqrt{40}$

$= -2.53$

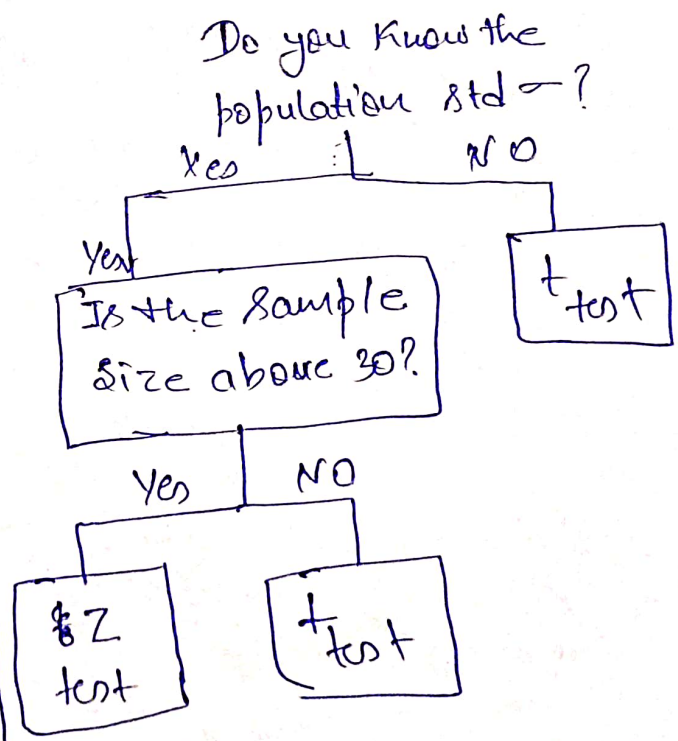
Area under the curve of -2.53, z-value is = 0.0570

p-value → 0.0570

∴ if p-value < Significance

0.0570 < 0.02 → false ∴ we accept the Null Hypothesis

T-test	z-test
Small ($n \leq 30$)	Larger ($n > 30$)
population variance unknown	Know
Distribution used Student's t-distribution	Normal distribution



T-test

A t-test is a statistical test used to compare the mean of one or more groups when the population standard deviation is unknown and the sample size is small ($n < 30$).

* **Assumptions:**
  * The data should be numerical, not categorical.
  * The sample must be randomly selected :- Random sampling
  * The data should follow an approximately normal distribution

formula

(a) one-sample

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

s = sample std
μ = population mean
$\bar{x}$ = sample mean
n = population size

(b) Two sample

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$\bar{x}_1, \bar{x}_2$ = sample mean
s_1^2, s_2^2 = sample variance
n_1, n_2 = sample size

(c) paired T-Test

$$t = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}}$$

$\bar{d}$ = Mean of difference b/w paired values
s_d = std of difference
n = Number of pairs

Degrees of freedom (df)

The Degree of freedom (df) represent the number of independent values in a statistical calculation that are free to vary.

(a) one sample

$df = n - 1$
n = sample size

(b) Two sample

$df = n_1 + n_2 - 2$
n_1, n_2 = sample sizes of two group

(c) paired T-test

$df = n - 1$
n = Number of paired observations.

* Df determine the shape of t-distribution
  * smaller d.f. mean wider distribution.
  * larger closer to normal

Q. t. Test

In the population the average IQ is 100. A team of researcher want to test a new medication to see if it has either a positive or negative effective 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? C.I = 95%.

Sol: $\mu = 100$, $n = 30$, $s = 20$, $\bar{x} = 140$, CI = 95% = .95

$\alpha = 0.05$

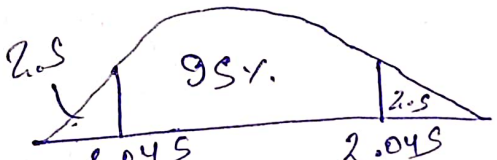
① Null Hypothesis $H_0 : \mu = 100$

② Alternate Hypothesis $H_1 : \mu \neq 100$ (2 tail Test)

③ $\alpha = 0.05$

④ Degree of freedom = $n - 1 = 30 - 1 = 29$

⑤ Decision Rule



0.05

By using t-table and the d.f is 29 for (2 tail) = 2.045

t-table 29 and (2 Tail) — 0.05

= 2.045

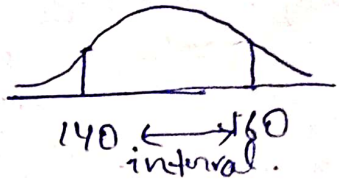
If t test is less than −2.045 and greater than 2.045, then Reject the Null Hypothesis

⑤ calculate t test stat

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

6) conclusion we reject the null hypothesis | 2.04 < 10.96 |

Confidence Interval and Margin of Error



We construct a confidence interval to help estimate what the actual value of the unknown population mean is

Point Estimate ± Margin of error

$$\text{Z test} = \bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

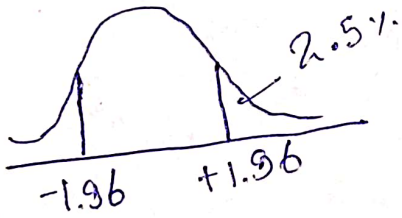
Q. On the verbal section of CAT exam the standard deviation is known to be 100. A Sample of 25 test taken has a mean of 520. Construct a 95% C.I about the mean.

Sol $\bar{x} = 520$, $\sigma = 100$, $n = 25$, C.I = 0.95, $\alpha = 0.05$

$= 1 - 0.025$

$= 0.9750$ (using z-table)

$= 1.96$



$$\bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

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

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

So range between 480.8 to 559

So, I am 95% confident that the mean CAT score lies between 480.8 to 559.2

Chi-Square Test.

The Chi-Square Test help us check if two things are related or if the difference we see are just luck.

eg: you ask 100 people if they like pizza or burgers, and you check if men and women have different preferences the Chi-Square Test you if gender really affects their choice or if the difference happened by chance.

Types of Chi-Square Test:

(i) Chi square Goddness-of-fit Test

* 1 categorical variable
* Tests if a sample follow a known distribution
* Use: you compaired observed vs expected frequencies

(ii) Chi-Square Test of independence

* 2 categorical variable.
* Tests if two categorical variable are related or independent
* you check if one ~~catory~~ category affects the others.

Formula: $x^2 = \sum \frac{(O-E)^2}{E}$

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

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

In 2020 ages of n=500 individuals were sampled. Below are result.

<50	50-75	75
140	160	200

Using $\alpha = 0.05$, would you conclude the population difference of weights has charged in the last 10 years.

Sol:

<50	50-75	775
20%	30%	50%

Expected

<50	50-75	775
140	160	200

Observed

(1)

Expected	<50	50-75	775
	$\frac{20 \times 500}{100}$	$\frac{30 \times 500}{100}$	$\frac{50 \times 500}{100}$
	100	150	250

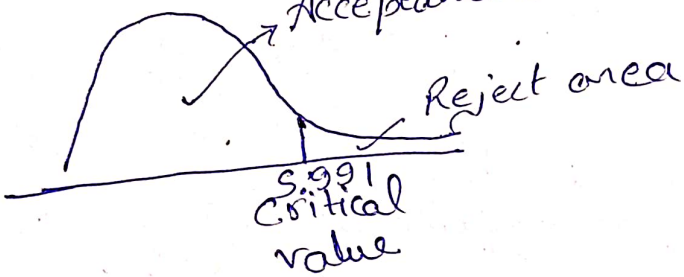
(1) Null Hypothesis: H_0: The data meets the exception

(2) Alternate Hypothesis: H_1: The data does not meet the exception.

(3) $\alpha = 0.05$ CI = 95%.

(4) Degree of freedom $df = k-1 = 3-1 = 2$

(5) Decision Boundary



By using Chi-table
$df = 2$ & $\alpha = 0.05$
we get 5.991

(5) calculate chi square test statistics

$$x^2 = \sum \frac{(O-E)^2}{E} \Rightarrow \frac{(40)^2}{100} + \frac{(10)^2}{150} + \frac{(50)^2}{250} = 26.66$$

(6) Con"

If x^2 is greater than 5.99; Reject H_0

else we fail to reject the Null Hypothesis.

So, $x^2 = 26.66 > 5.99$ {Reject the Null Hypothesis}

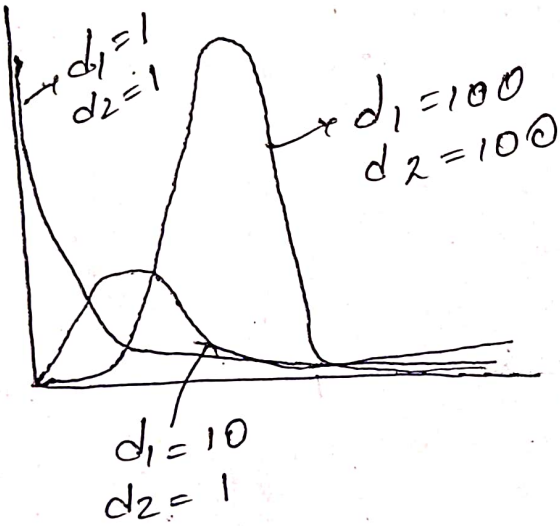
F-Distribution

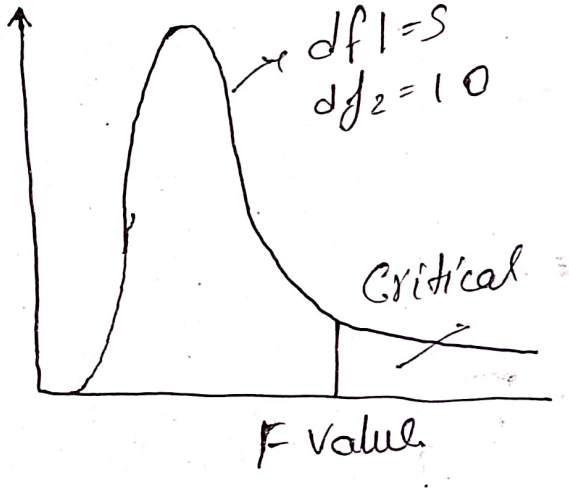
The F-Distribution is a probability distribution used mainly in ANOVA (Analysis of Variance) and F-Test to compaire of two datasets. It help determine if the mean (variance) the difference in data are real or just due to Chance
for anova

Key features of F-Distribution

* Always positive (values are never negative)
* Right-Skewed (not symmetric like a normal distribution)
* Used in hypothesis testing when comparing two variance
* Defined by two degree of freedom (df_1, df_2) > 0 always

Probability density function





Formula

$$Pdf = f(x; d_2, d_2) = \frac{\sqrt{\frac{(d_1 x)^{d_1} \, d_2^{2}}{(d_1 x + d_2)^{d_1 + d_2}}}}{x \, B\left(\frac{d_1}{2}, \frac{d_2}{2}\right)}$$

→ Beta function or $B(m,n)$

$$B(m,n) = \frac{(m-1)!\,(n-1)!}{(m+n-1)!} = \frac{m+n}{mn} \Big/ \frac{(m+n)}{m}$$

$$F = \frac{S_1^2}{S_2^2}$$ ∴ S_1^2, S_2^2 variance of first and second sample.

F-Test

The F-Test is a statistical test used to compare the variance of two datasets. It helps check if the spread (variability) of data in two group is significantly different.

Q The following data shows the no. of bulbs produced daily for some days by 2 workers A and B

A	40	30	38	41	38	35	
B	39	38	41	33	32	39	40	34

Can we consider based on the data workers B is more stable and efficient. $\alpha = 0.05$ C.I = 95% = 0.95

Sol: ① Null Hypothesis $H_0: \sigma_1^2 = \sigma_2^2$

② Alternate Hypothesis $H_1: \sigma_1^2 \neq \sigma_2^2$

③ calculate of variance

A x_1	$\bar{x}_1$	$(x_1 - \bar{x}_1)^2$
40	37	9
30	37	49
38	37	1
41	37	16
38	37	1
35	37	4
	$\bar{x}_1 = 37$	$\Sigma (x - \bar{x})^2 = 80$

B x_2	$\bar{x}_2$	$(x_2 - \bar{x}_2)^2$
39	37	4
38	37	1
41	37	36
33	37	16
32	37	25
39	37	4
40	37	9
34	37	9
	$\bar{x}_2 = 37$	$\Sigma (x_2 - \bar{x}_2)^2 = 84$

$$s_1^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x}_1)^2}{n-1}$$

$$= \frac{80}{6-1}$$

$$= 13$$

$$s_2^2 = \frac{84}{8-1}$$

$$= \frac{84}{7}$$

$$= 12$$

④ Calculate of variance ratio of F-test

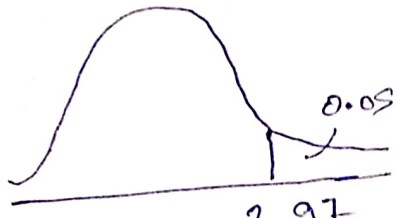
$$F = \frac{s_1^2}{s_2^2} = \frac{13}{12} = 1.33$$

⑤ Decision Rule

$df_1 = 6 - 1 = 5$ and $df_2 = 8 - 1 = 7$

By using F-table and $df_1 = 5$ and $df_2 = 7$

⑥



⑦ So, if F test is greater than 3.97 ; Rejected the Null Hypothesis

$1.33 < 3.97$

We fail to reject Null

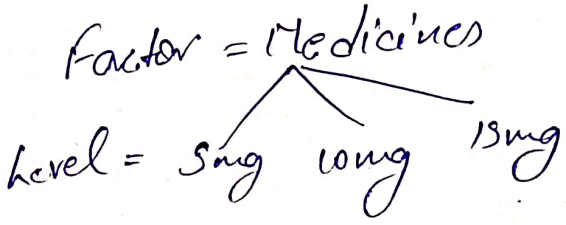
ANOVA (Analysis of Variance)

ANOVA is a statistical method used to compare the mean of 2 or more groups

Anova

① Factors (Variables)
② Level

eg:



Assumption in ANOVA

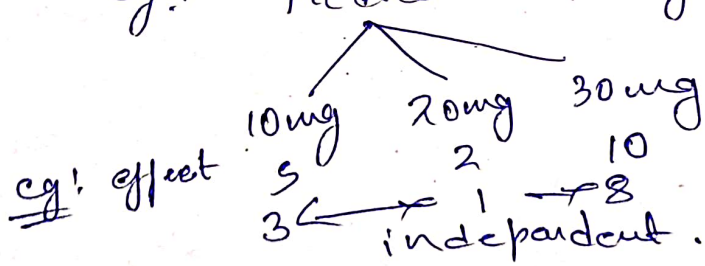
(i) Data is numerical and normally distributed
(ii) Homogeneity of Variance
Each one of the population has same variance
$[\sigma_1^2 = \sigma_2^2 = \sigma_3^2]$
population variance in different levels of each independent variable are equal.
(iii) Samples are independent and random
(iv) Absense of outliers.

Types of ANOVA

formula → $F = \frac{\text{Variance between groups}}{\text{Variance within groups}}$

(i) One way ANOVA: 1 factor with atleast 2 levels these level are independent.

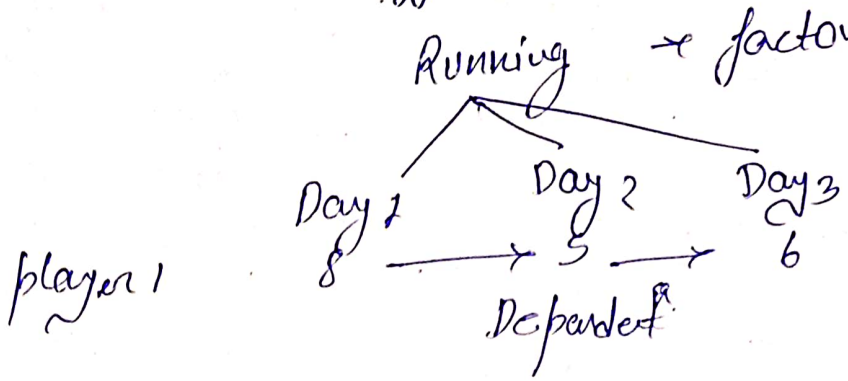
eg: Medication → factor



(2) Repeated Measures ANOVA:

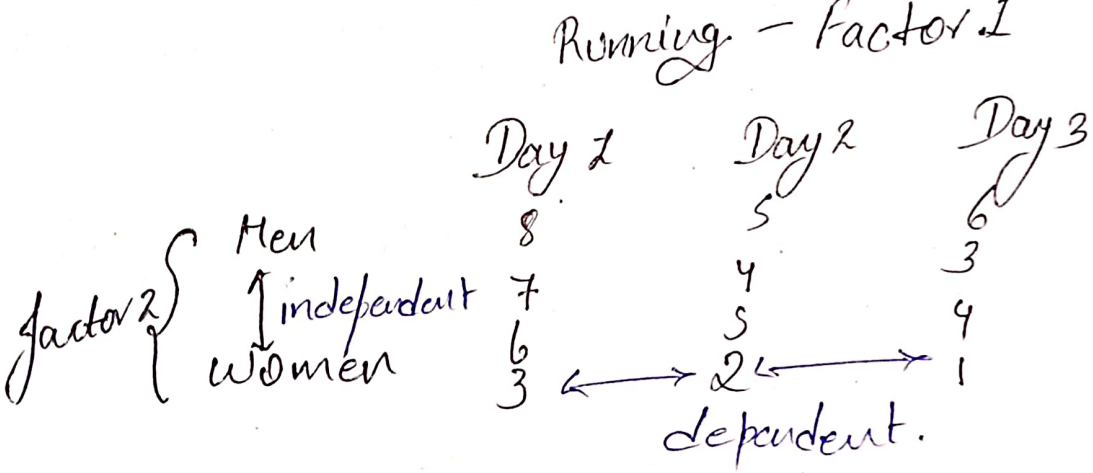
One factor with atleast 2 levels, and level are dependents.

* It used when you measure the same subjects multiples times under different conditions or over time.



(3) Factorial Anova / Two factor

Two factor or more factor each of which with atleast 2 level, level can be either independent and dependant



Hypothesis Testing in Anova

Null Hypothesis H_0: $\mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$

Alternate hypothesis H_1: Atleast one of the mean is not equal.

$$\mu_1 \neq \mu_2 = \mu_3 \dots \mu_k$$

Q.1 Doctors want to test a new medication with reduce headache. They splits the participant into 3 condition [15mg, 30mg, 45mg]. Later on the doctor ask the patient to rate the headache between [1-10] Are they any differences between the 3 condition using alpha 0.05?

15mg	30mg	45mg
9	7	4
8	6	3
7	6	2
8	7	3
8	8	4
9	7	3
8	6	2

So:

① Null Hypothesis: $H_0 = \mu_{15} = \mu_{30} = \mu_{45}$

② Alternate Hypothesis H_1: not all μ_i are equal.

$\alpha = 0.05$, C.I = 0.95

③ calculate Degree of freedom

$N = 21$ $a = 3$ $n = 7$

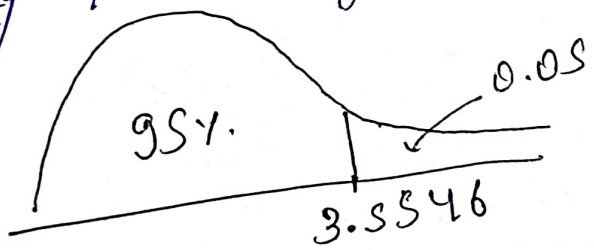
$df_{between} = a - 1 = 3 - 1 = 2$

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

$df_{total} = N - 1 = 21 - 1 = 20$ ($df_{between} + df_{within}$)

Using F-table of (2, 18) we get = 3.5546

Critical = 3.5546



④ State Decision Rule

if F is greater than 3.5546, reject the Null Hypothesis

⑤ Calculate test Statistics

	SS	df	MS	F
Between	98.67	2	98.67/2 = 49.34	49.34 / 0.54
within	10.29	18	10.29/18 = 0.54	= 86.56
Total	108.95	20	49.88	

$SS_{between}$, SS_{within}, SS_{total}

15mg.	30mg	45mg
9	7	4
8	6	3
7	6	2
8	7	3
8	8	4
9	7	3
8	6	2

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

15mg = 9+8+7+8+8+9+8 = 57

30mg = 7+6+6+7+8+7+6 = 47

45mg = 4+3+2+3+4+3+2 = 21

$$SS_{between} = \frac{57^2+47^2+21^2}{7} - \frac{[57+47+21]}{21} = 98.67$$

$$SS_{within} = \Sigma y^2 - \frac{\Sigma(\Sigma ai)^2}{n}$$

$$= \Sigma y^2 - \left[\frac{57^2+47^2+21^2}{7}\right]$$

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

$= 853$

$$= 853 - \frac{[57^2+47^2+21^2]}{7}$$

$SS_{within} = 10.29$

(5) $SS_{total} = \Sigma y^2 - \frac{T^2}{N}$

$= 853 - \frac{125^2}{21} = 108.95$ $(SS_{between} + SS_{within})$

(6) $F = \frac{MS\ between}{MS\ within}$

$= \frac{49.34}{0.54} = 86.56$

So, Con"

If F is greater than 3.5546, reject the Null Hypothesis

$86.56 > 3.5546$ Reject the Null Hypothesis