Hand Wrtten / Easy explained Notes

STATISTICS FOR THE DATA SCIENCE

Part - 1

- WHY STATISTICS IN DATA SCIENCE?
- TYPES OF STATISTICS: DESCRIPTIVE VS. INFERENTIAL
- POPULATION VS. SAMPLE DATA
- SCALE OF MEASUREMENT
- MEASURES OF CENTER TENDENCY
- MEASURES OF DISPERSION
- SETS IN STATISTICS

#Value_freeContent

@Krishan kumar

Statistics

Defination Statistics in the science of collecting, organizing and analyzing data

Data facts or pieces of imformation → It can be mejor and collect

ex → ① Height in the classroom of the student [175, 180, 160] cm

② IQ [100, 90, 80, 60]

Why statistics ?

To innovate any product, to bring value to any product, the role of data is very important. Statistic give a lots of information about data because statistic provide such tools, we can get many infraction and concussion from the providing data.

Types of Statistics :-

1. Descriptive statistic →

↳ It consits of organizing and summarizing data.

(A) Measure of centre Tendency →

[Mean, Median, Mode]

(B) Measure of Dispersion →

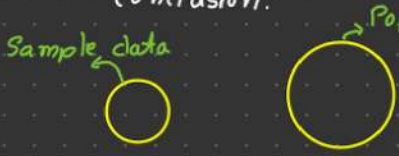
[Variance, Std (Standard deviation)]

(C) Different type of Distribution of data →

tools :- [Histogram, PMF (Probability Function), PDF (" Density ")]

2. Infernfion statistics →

↳ It consists of using data you have measured to from conclusion.



Problem lets say there are 20 Statistic class at your college and you have collected the height of the student in the class. heights are recorded [175, 180, 175, 180, 176, 160, 135, 190] cm.

* **Descriptive Question** → What are the average height of the entire classroom.

Sol ⇒ Mean → Average → this is the part of descriptive stats.

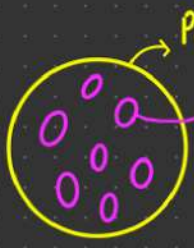
↳ $\frac{175+180+175+180+176+160+135+190}{8}$ → Average height

Infernition Question → Are the height of the student (Sample data) in the classroom smilliar what you expect in the entire college. (population data)

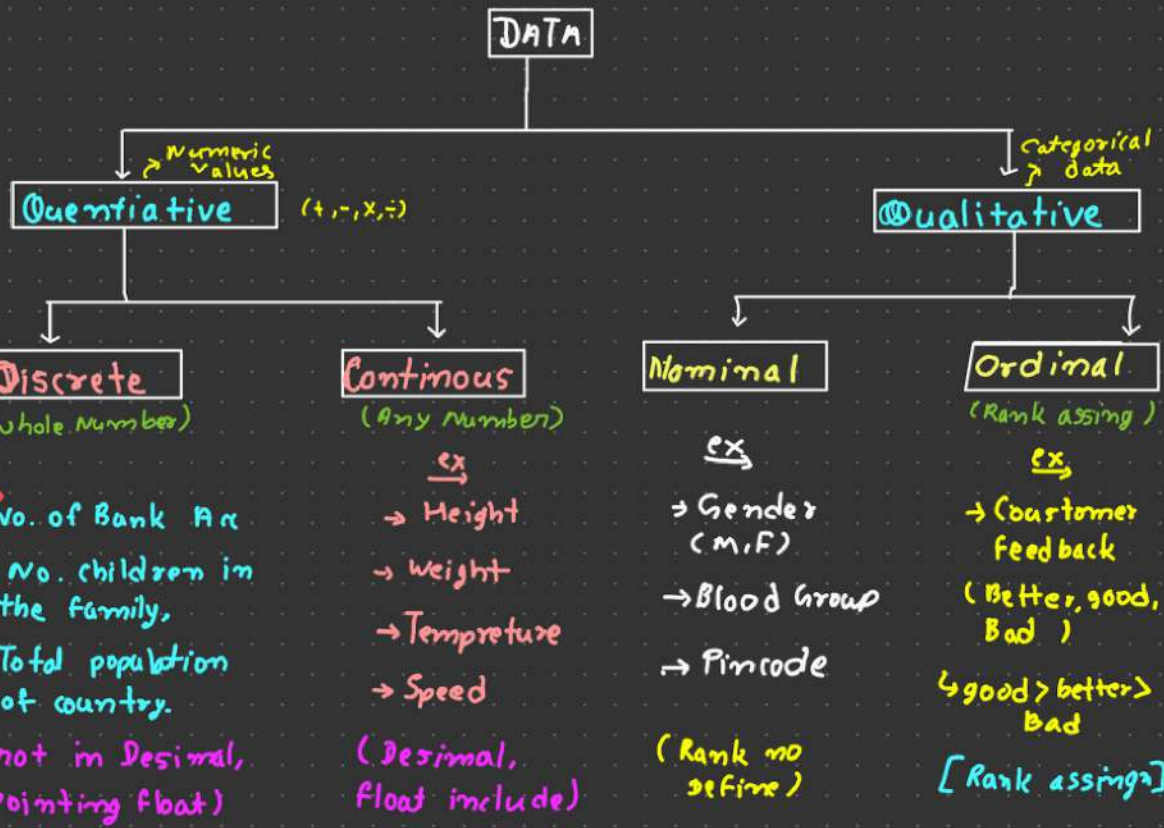
Population and sample data

* Population Data ⇒ The group you are intrested in studying
* Sample Data ⇒ A Subset of population data

Ex) Exit pole of election



Types of Data



Scale of Measurement

- Nominal scale data.
- Ordinal scale data
- Internal scale data
- Ratio scale data.

Nominal scale Data :-

→ Qualitative OR categorical data
→ Order does not matter
↳ ex → Gender / color / habit

ex Fevorite color.

Red → 5 → 50%
Black → 3 → 30%
Blue → 2 → 20%
10

there are No any Specific Rank

↳ by nominal data we find (called → cumulative freq.

Ordinal scale data

* Ranking is imp.
* Order is matter
* Difference can't be m measured

ex → 1 → Best
2 → Good
3 → Bad

ex Working Profession Productivity

Hours		Day
1	-	Mon.
2	-	Tue.
3	-	Wed.
4	-	Th
5	-	Fri

Best work on → friday

ex → Race →
1st
2nd
3rd
→ In ordinal Difference can't be measure.

Internal scale data

* The order matter
* Difference can't be measured.
* Ratio can't measured
* No '0' starting fixed

It can be -ve OR +ve

Ex, Temp Variable →

60F, 90F → Ratio → 60:90 → 2:3

→ Diff. can be measure

120 F, 180F → 90 - 30 → 60 F

$\frac{180}{120} \rightarrow \frac{60}{40} \rightarrow 3/2$

Ratio does not mean that Temp. increase OR ↓↓ 3/2 ↑.

Ratio scale data

* The order matter
* Diff. are measurable (including ratio)
* Contain a '0' Starting point.

Ex, Student marks in a class.

0, 90, 60, 30, 75, 80, 85, 50

Assending order → 0, 30, 50, 60, 75, 80, 85, 90

Diff. → 60 - 30 = 10

80 - 50 = 30

Ratio → $\frac{30}{90}$ = 3:1 This person gain 3X marks from that one (30 mark)

Ratio measured in Grading.

Question not in Temp. Type Question

Measure of centre tendency

- Mean OR Average
- Median
- Mode

Mean OR Average :-

Population (N)

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

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

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

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

$N = 3.2$

Sample mean $= \bar{x}$

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

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

$\bar{x} = 3.2$

Median :

$x = [4,5,2,3,2,1]$

Steps

- Sort the random variable → [1,2,2,3,4,5]
- No. of element , count → 6
- if count == even

  [1,2,2,3,4,5]

  ↳ Median → $\frac{2+3}{2}$ → 2.5

- if count == odd

  [1,2,2,3,3,4,5]

  Median = 3

Why Median ?

↳ Median is used to find centre of tendency when outlier is present.

$x = [1, 2, 3, 4, 5]$

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

$\bar{x} = 3$

Diff. in mean

$x = [1, 2, 3, 4, 5, 100]$ → outlier

$\bar{x} = \frac{1+2+3+4+5+100}{6} = \frac{115}{6}$

$\bar{x} = 19.16$

In Median →

$x = [1, 2, 3, 4, 5, 100]$

Median → $\frac{3+4}{2} \rightarrow \frac{7}{2}$

Median = 3.5

So the result is, Mean with outlier → 3 (Diff) 19.16

Median with outlier

3 → 3.5 (very close)

Note

Whenever we find out centre of tendency we should do median.

Mode :-

↳ Frequency maximum

↳ Maximum repeating Number

$x = [1, 2, 3, 4, 5, 8, 9, 1, 2, 3, 2, 8, 9, 1, 2, 3, 4]$

Mode = 2

Where and why we use mean, median, mode,

We use these all in EDA and feature engineering.

for ex →

Age	weight	Salary	Gender	Degree
	Numeric fe.		Mode	Mode
24	70	40K	M	BE
25	80	50K	F	-
27	95	70K	M	-
24	-	35K	-	BE
32	-	70K	M	PHD
-	60	40K	-	Master
-	65	-	F	BSc
40	72	-	M	BE

→ Data is missing

Note

Missing data filling

* if data in → Categorical → Fill by Mode (best result)
* if data in → Numerical → Fill by Mean (best result)
* Data with Outlier → Fill by Median (best result)

Measure of Disperson :- (How much our data spread)

Types → (i) Variance
(ii) Standard deviation

To calculate the this spreaded data we use → Measure of disperson

(i) Variance :-

(A) Population variance

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

x_i → Data point
μ → Population mean
N → Population size

(B) Sample variance

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

x_i → Data point
$\bar{x}$ → Sample mean
n → Sample size

Q Why we devide Sample variance by n-1 ?

Sol The sample variance devided by n-1 so that we can create an unbiased estimater of the population variance
↳ This scenerio → called → Bencle correction

Ex → [1,2,3,4,5] , s^2 = ?

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

⇒ $s^2 = \frac{10}{5}$ ⇒ 2

Sample variance → 2

x	$\bar{x}$	$(x_i - \bar{x})$
1	3	4
2	3	1
3	3	0
4	3	1
5	3	4
means 15/5 = 3		10

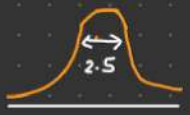
* What difference in these two sample variance ? X, y

↳ $X = 2.5$ $\quad y = 7.5$

$s^2 = 2.5$ $\quad s^2 = 7.5$

Ans → s^2 = denote → Dispersion of spread

$s^2 = 2.5$



→ When spreadness = ↑↑ (increase)

Height becames = ↓↓ (decrease)

Standard deviation :- (std) (σ)

(A) Population Standard deviation

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

(B) Sample std.

Sample std = $\sqrt{s^2}$

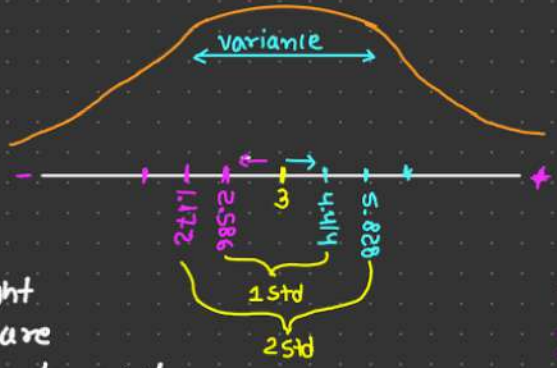
s^2 → sample variance

ex →

$X = [1, 2, 3, 4, 5]$

Sample mean = $\bar{x} = 3$

$\sigma = \sqrt{3} = 1.414$



When one step towards the right and one step towards the left are combained then we called them 'one std'

Q. Where the 2 element fold in the variance.

Ans → one step left to the mean.

3.000
-1.414
2.586
-1.414
1.172

3.000
+1.414
4.414
+1.414
5.828

Random variable

Random variable is a process of maping the output of a random process or experiments to a number there are not any fix value.

ex,

Tossing a coin

$X = \begin{cases} 0, \text{ if Head} \\ 1, \text{ if Tell} \end{cases}$

Rolling a dice

$Y = \{\text{Sum of the rolling a dice 7 times}\}$

↳ $P(Y \geq 15)$ OR $P(Y < 10)$

Probability of

Sets

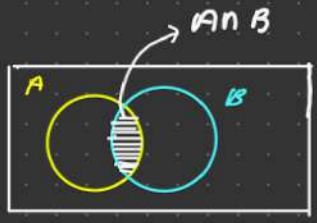
$A = \{1, 2, 3, 4, 5, 6, 7, 8\}$

$B = \{3, 4, 5, 6, 7\}$

Intersection ($A \cap B$)

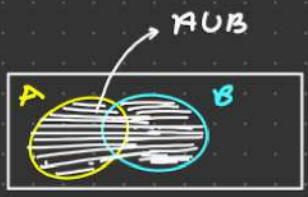
(common value)

$A \cap B = \{3, 4, 5, 6, 7\}$



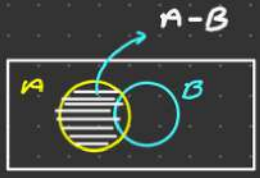
Union

$A \cup B = \{1,2,3,4,5,6,7,8\}$



Difference

$A - B = \{1,2,8\}$



Subset (Som)

$A \rightarrow B$ = False (∵ all element of B present in A)

$B \rightarrow A$ = True (∵ All element of A present in A)

Superset

$B \rightarrow A$ = False

$A \rightarrow B$ = True

STATISTICS FOR THE DATA SCIENCE

Part - 2

- HISTOGRAME AND SKEWNESS
- COVARIANCE AND CORRELATION
- ADVANTAGE VS DISADVANTAGE OF VARIANCE
- PEARSON CORRELATION COFFICIENT
- PDF / PMF / CDF
- BERNOULLY DISTRIBUTION
- BINOMIAL DISTRIBUTION

#Value_freeContent

@Krishan kumar

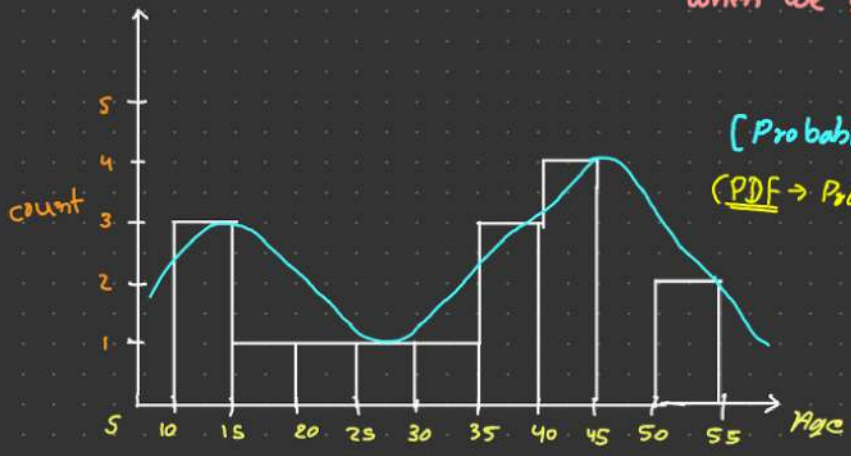
Histograms And Skewness

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

0 to 50 value

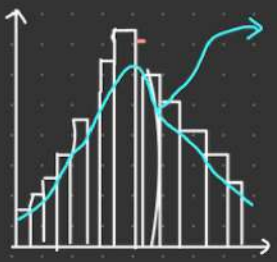
↳ $\frac{50}{10}$ = (5) bin size (No. of bins → 10)

↳ $\frac{50}{20}$ = (2.5) bin size (No. of bins → 20)



when we smothen our histogram
↓
Technique
↓
[Probability distribution fun.]
(PDF → Probability density function)

Skewness



Normal / Gaussian destribution
↘
These types of destribution is called → Symmetrical destribution.

* Symmetrical Data

No Skewness
(No curve)
Equal from both side

→ Boxplot :- To finding outliers

$$\Rightarrow Q_3 - Q_2 \cong Q_2 - Q_1$$

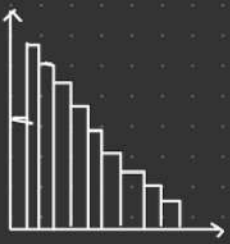
The mean, median, mode are all perfectly at the centre.

Mean = Median = Mode

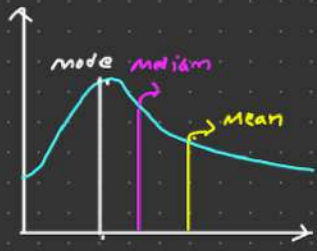
There are no skewness in the symmetrical distribution data and just because there are no any maxi value and minimam value in Right OR Left corner of data.

② Right Skewed data

Means longitute in Right side (Mejority data on right side present)



⇒ Positive / Right skewed → Distribution in PDF →

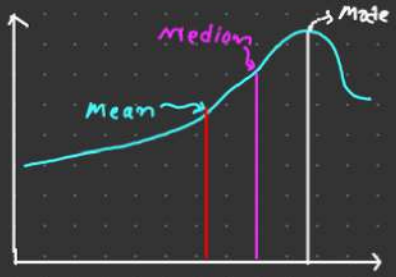


Relation btw Mean, median, mode ⇒ **Mean > Median > Mode**

* Box plot for Right Skewed data

$\Rightarrow \quad Q_3 - Q_2 \geq Q_2 - Q_1$

* Left Skewed data



* Box. plot

$\Rightarrow \quad Q_2 - Q_1 \geq Q_3 - Q_2$

Relation

Mean < Median < Mode

Covariance and Correlation

[To reactify Relation btw X and Y]

X	y
2	3
4	5
6	7
8	9

In this scenario →

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

↳ For example →

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

→ Plotted like →

Covariance →

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

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

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

$$\Rightarrow Cov(x,x) = Var(x)$$

x_i → Data point of x

$\bar{x}$ → Sample mean of x

y_i → Data point of y

$\bar{y}$ → Sample mean of y

Q What are the difference between covariance and variance?

Ans Variance of x $var(x)$ is nothing but $cov(x,x)$, whenever we talk about variance(x) it is specifically told about spread of the data

⇒ $Cov(x,y)$

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

→ Then the output will be +ve covariance

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

→ Then output will be -ve covariance

X	y
2	3
4	5
6	7
$\bar{x} \to 4$	$\bar{y} \to 5$

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

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

$$\Rightarrow \frac{4+0+4}{2} = 4 \to \text{+ve value}$$

positive covariance = +ve value

So, X and Y are having a positive covariance.

Advantage of covariance

* Relation between X and Y +ve OR -ve value of covariance

* Disadvantage of covariance

* We don't conclude weather which value is covariance with any other value Because ↓

Covariance does not have a specific limit value.

ex 

[there is no limit of Difference]

Note: To fix the disadvantage of covariance we use :-

↳ Pearson correlation coefficient

Pearson correlation cofficient

limition → +1 to -1

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

* The more value towards +1 the more +ve correlated it is. (x,y)

* The more value towards -1 the more -ve correlated it is (x,y)

Spearman Rank correlation

Rang → -1 to +1

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

Rank → Highest element in the values makes → 1st Rank

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

* Feature selection

+ve	+ve	+ve	≈ 0	-ve
Size of House	No. of Rooms	Location	No. of people staying	Hounte ⇒ Price ↑↑

PDF / PMF

* PDF = Probability Density function
* PMF = Probability mass function.

↳ PDF is denote a distribution of data. That helps to understand that how our data will distributed.

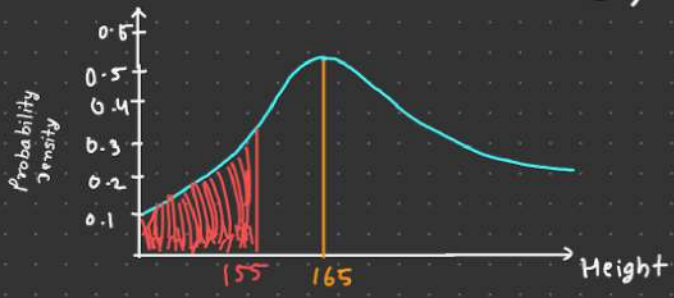
Types of Probability Distribution function

1. Pdf
2. Pmf
3. cdf

Probability density function (Pdf)

↳ Continous Random variables, ex → Height of students in classroom

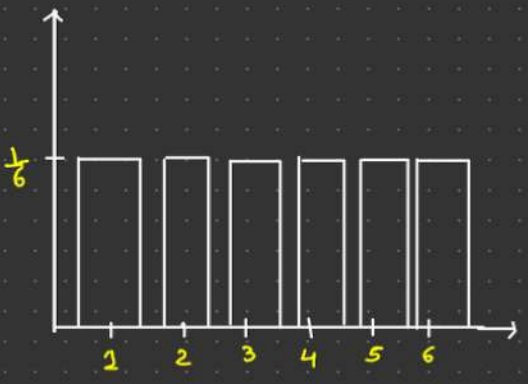
↳ $P(X \leq 155)$



Probability mass function (Pmf)

↳ Variable → Discrete Random variable

ex → Rolling a dice { 1,2,3,4,5,6 }



$P(X \leq 4)$

$\Downarrow$

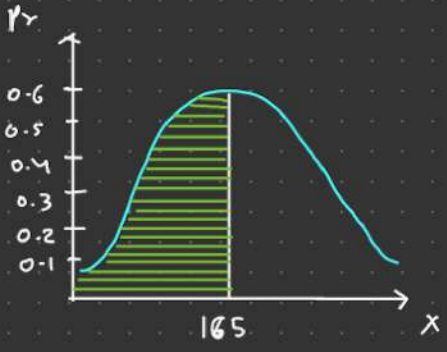
$P(x=1) + P(x=2) + P(x=3) + P(x=4)$

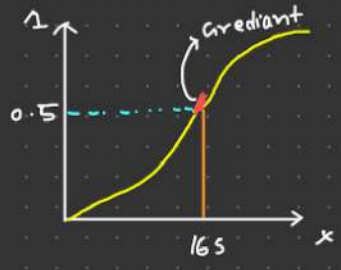
$\Rightarrow \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} \rightarrow \frac{4}{6} = \frac{2}{3}$

$P(1) \rightarrow \frac{1}{6}$

$P(2) \rightarrow \frac{1}{6}$

Cumulative Distribution fun. (cdf)





PDF Vs PMF Vs CDF

Relation and difference

① Pmf

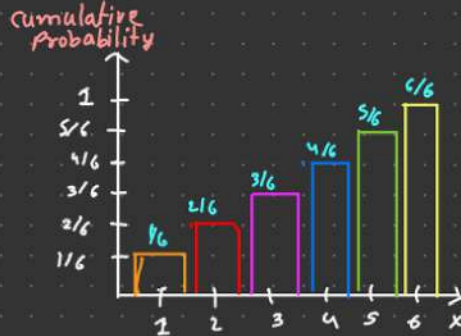
* Descrete Random variable.

Ex → Rolling a dice ⇒ [1,2,3,4,5,6]

P(1) → 1/6
⋮
P(6) → 1/6

→ This called → Uniform distribution

CDF



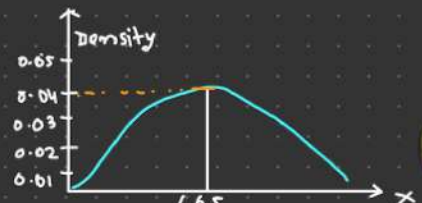
Ex → cdf for P(X ≤ 2)

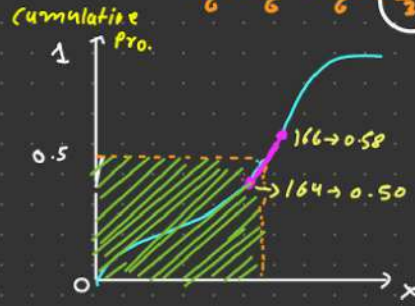
↳ cdf = P(X ≤ 2) = P(X=1) + P(X=2)

$= \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}$

② PDF

↳ Distribution of continous Random variable





→ slope ⇒ $\frac{0.53 - 0.5}{166 - 164} \rightarrow \frac{0.03}{2}$

⇒ 0.04

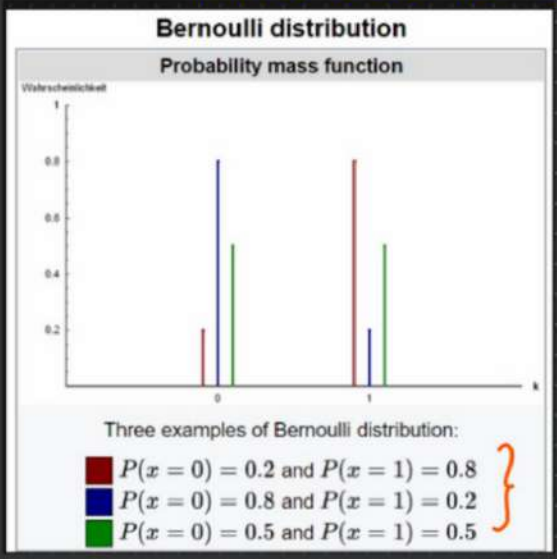
Types of Probability Distribution

- Normal / Gaussian distribution (Pdf)
- Bernoully distribution (Pmf)
- Uniform distribution (Pmf)
- Poisson distribution (Pmf)
- Log normal distribution (Pdf)
- Binomial distribution (Pmf)

Bernoully distribution

In the discrete probability distribution of a random variable which takes the value 1 with probability P and the value 0 with Probability $q = 1-p$ Less formally it can be though us a model for the set of possible outcomes of any single expriement that asks a yes-no question.

Pmf → To see the distribution



Points

① Discrete random variable (PmF)

② Out comes are binary

ex: 1. Tossing a coin (H, T)

$P(H) \rightarrow 0.5 \rightarrow p$ (Pro. of Succuss)

$P(T) \rightarrow 0.5 \rightarrow 1-p \Rightarrow q$

(q = Pr. of failure)

③ When there person will fail / Pass

↳ $P(\text{Pass}) \rightarrow \boxed{0.7 \rightarrow p}$ → Succus value

$P(\text{fail}) \rightarrow 1-0.7 \Rightarrow \boxed{0.3 = q}$ → Failare value

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

if $k=1$

Pmf → $P(k=1) = p^1 * (1-p)^{1-1}$

$= p$

if $k=0$

$Pmf = P(k=0) = p^0 * (1-p)^{1-0}$

$= (1-p) = q$

$\Rightarrow q$

Simplified

$$Pmf = \begin{cases} q=1-p & , \text{if } k=0 \\ p & , \text{if } k=1 \end{cases}$$

⇒ Mean of bernaully distribution,

$$\langle \varepsilon(k) = \sum_{j=1}^{k} k.P(k)$$

, $P(k=1) = 0.6 \Rightarrow p$

, $P(k=0) = 0.4 \Rightarrow 1-p \Rightarrow q$

→ $[0 \times 0.4 + 1 \times 0.6] \Rightarrow 0.6 = p$

Whenever we abstract mean in bernaully distribution, then we get P-Value.

⇒ Median of Bernoully dis,

$$\text{Median} = \begin{cases} 0 & , \text{if } p < \frac{1}{2} \\ [0,1] & , \text{if } p = 1/2 \\ 1 & , \text{if } p > 1/2 \end{cases}$$

⇒ Variance of bernoully dis.,

$Var = p(1-p)$

$\boxed{var = pq}$

⇒ Std of bernoully dis,

$\boxed{Sdtd = \sqrt{pq}}$

Binomial Distribution

In Statistic the binomial dis. with parameter n and P is the discrete Pro. Dis of the number of success in a squence of n independent experiments. Each asking Yes or No. question. and each its own boolen-value outcome. success (with Pro. P) or failare ($q = 1-P$).

A single success / failare experiement is also called a bernoulli trial OR Bernoully experiment and a sequence of outcome is called a bernoully process.

For a single trial ex⇒ $n=1$ the binomial dis. is a bernoully dis. the binomial dis. is the basis for the popular binomial test of statistial singnificant.

$B(n,P)$

Notation: $B(n,P)$

* Every experiment outcome is binary
* This experiment is Perfoms for n trial
* Group of bernoully dis → Binomial Distribution
* Discrete random variable.

Parametters

* $n \in \{0,1,2,\dots\}$ → No. of trial
* $P \in [0,1]$ → Success Pro. of each trial
* $q = 1-P$
* Ex⇒ Tossing a coin 10 time

Support:- $k \in \{0,1,2,3,\dots n\}$
→ Number of success

* Pmf :-

$$P(k, n, p) = {}^{n}C_{k}\, p^{k} (1-p)^{n-k}$$

for $k = 0, 1, 2, 3, \ldots, n$ where

$$^{n}C_{k} = \frac{\lfloor n}{\lfloor k \, \lfloor n-k}$$

⇒ **Mean :-**

$$\text{Mean} = np$$

⇒ **Variance :-**

$$\text{Var} = npq$$

⇒ **Std**

$$\text{Std} = \sqrt{npq}$$

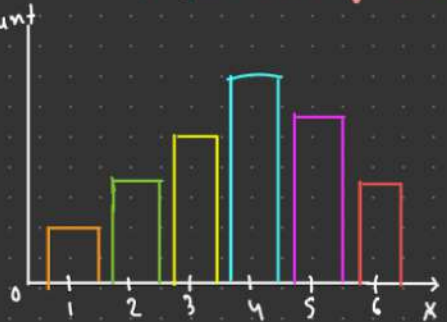
Poisson Distribution

* Discrete random variable. (Pmf)
* Discribe the number of event occuring in fixed time interval.

ex) ① No. of people visiting hospital every hour.

② No. of people visiting bank every hour.

Pmf



$\lambda = 3$, Expected no. of event occure at every time interval.

Means → How many people come at every hour.

STATISTICS FOR THE DATA SCIENCE

Part - 3

- POISSON DISTRIBUTION
- NORMAL / GAUSSIAN DISTRIBUTION
- UNIFORM DISTRIBUTION
- Z - SCORE
- CENTRAL LIMIT THEORAM
- ESTIMATOR
- HYPOTHESIS AND TESTING MECHANISIM
- P -VALE
- Z - TEST

#Value_freeContent

@Krishan kumar

Q What is the probability of a person to come at 5th hour.

Ans

$$Pmf = P(X=5) = \frac{e^{\lambda} - \lambda^{x}}{\lfloor x}, \quad \text{if } \lambda = 3 \text{ (औसत)}$$

$$= \frac{e^{3} - 3^{5}}{\lfloor 5} \Rightarrow 0.101 \rightarrow 10.1\%$$

there is 10% posibilities that at 5th hour 3 person come.

Q What is the probability to visit at 5th hour OR 4th hour.

Ans

Pmf $P(X=5) + P(X=4)$

* Mean of poission distribution

Mean $= \Sigma(x) = \mu$

$$\boxed{\mu = \lambda * t}$$

λ = expected no. of event occur at every time interval

t = time interval

Variance of poission = $\lambda * t$

Normal / Gaussian Distribution (PDF)

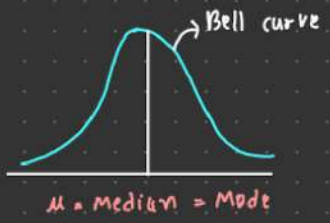
Notation :- $N(\mu, \sigma^2)$

Parameter :-

$\mu \in R$ (Mean) — Real number

$\sigma^2 \in R > 0$ = Variance

$x \in R$ = Data points



$$PDF = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma}}$$

⇒ **Mean of normal distribution**

Mean = μ = Average

⇒ **Variance**

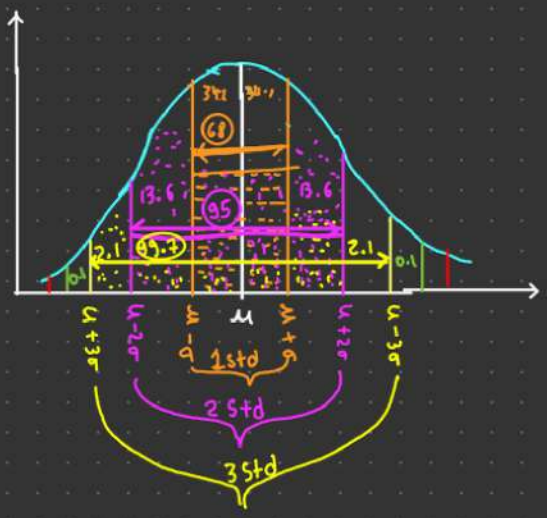
Var = σ^2

⇒ **Std**

$\sigma = \sqrt{Var}$

This role follow

68 - 95 - 99.7 % Rule



$P(\mu-\sigma \leq x \leq \mu+\sigma) \simeq 68\%$

$P(\mu-2\sigma \leq x \leq \mu+2\sigma) \simeq 95\%$

$P(\mu-3\sigma \leq x \leq \mu+3\sigma) \simeq 99.7\%$

Ex → ① Weight of the students in the class.
② Height of " " " " ".
③ IRIS Dataset [Sepal widht]

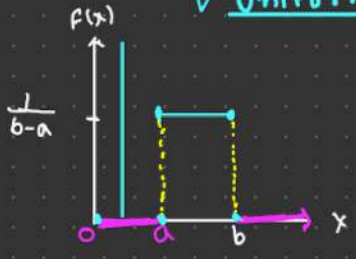
Uniform Distribution →

(A) Continous uniform distribution (PdF)
(B) Discrete Uniform distribution (Pmf)

(A) Continous Uniform Distribution →

In statistic, the continous uniform dis. or Rectangular dis. is a family of symmetric probability dis. The dis discribe an experiment where there is an arbitrary outcome that lie between certain bounds.
the bounds are defined by the parametter a and b, which are the minimum and maximum va

V Uniform (PdF)

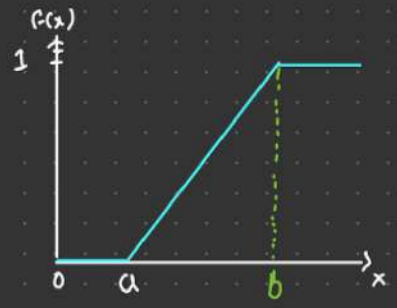


Notation → V(a,b)

Parametter →

$-\infty < a < b < \infty$

- **CDF**



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

$$cdf \rightarrow \begin{cases} 0 & \text{for} \quad x < a \\ \frac{x-a}{b-a} & \text{for} \quad x \in [a,b] \\ 1 & \text{for} \quad x > b \end{cases}$$

Mean $= \frac{1}{2}(a+b)$

Variance $\rightarrow \frac{1}{12}(b-a)^2$

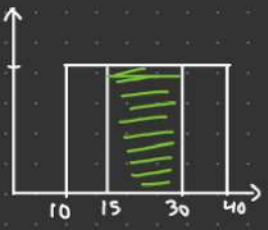
Median $\rightarrow \frac{1}{2}(a+b)$

Ex → The no. of candie sold daily at a shop is uniformally dis. with a maxi at 40 and mini. of 10.

① Probability of daily Sales to falls between 15 and 30.

Sol ⇒

$\frac{1}{b-a} \Rightarrow \frac{1}{40-10} \Rightarrow \frac{1}{30}$



$x_1 = 15$

$x_2 = 30$

$$P(15 \leq x \leq 30) = (x_2 - x_1) \times \frac{1}{b-a}$$

$$\Rightarrow 15 \times \frac{1}{30} \Rightarrow \frac{1}{2} \Rightarrow 0.5$$

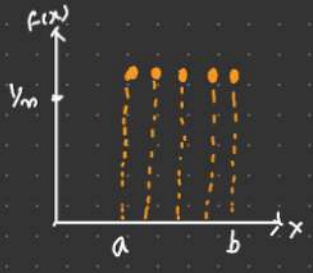
② $P(x \geq 20)$

Sol. $(40-20) \times \frac{1}{30}$

$\Rightarrow 0.66 \rightarrow 66\%$

(13) Discrete uniform dis. (Pmf)

In statistic the discrete uniform dis. is symmetric probability dis. wherin a finite number of value are equally likely to be observed every on of n value has equally probability 1/n. Another way of saying that "discrete uniform Dis" would be a known, finite number of outcome equally likely to happen.



Ex Rolling a dice

[1,2,3,4,5,6]

↳ P(1) → 1/6

⋮

P(6) → 1/6

a → 1

b → 6

$\frac{1}{n} = n = b - a + 1$

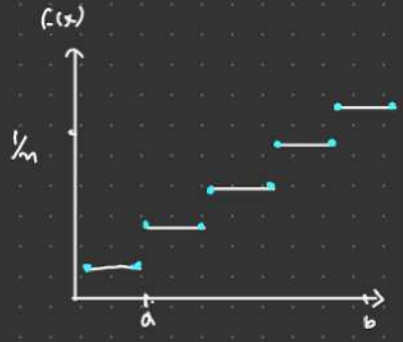
Natation → u (a,b)

Parametter → a,b with $b \geq a$

Pmf = $\frac{1}{n}$

Mean = $\frac{a+b}{2}$

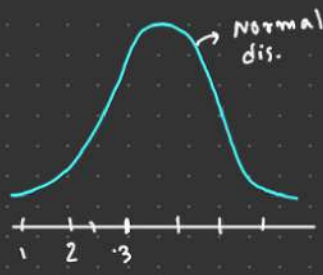
Medion ↗



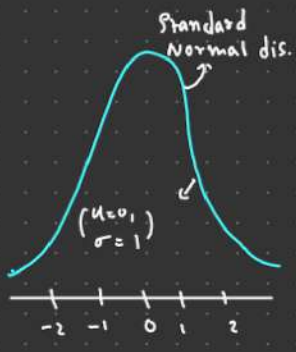
Standard Normal Distribution and Z-score

$X \in [1,2,3,4,5]$, $\mu = 3$

$\sigma = 1.414 \simeq 1$



Transition Technique → When $\mu = 0$, $\sigma = 1$



⇒ Z-Score

Z-score is used for the help for Normal dis. to Transform in standard normal distribution.

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

$x_i = 1$, $\frac{1-3}{1} \Rightarrow -2$

$x_i = 2$, $\frac{2-3}{1} \Rightarrow -1$

$x_i = 3$, $\frac{3-3}{1} \Rightarrow 0$

$x_i \to 4 \Rightarrow \frac{4-3}{1} \to 1$

$x_i \to 5 \Rightarrow \frac{5-3}{1} \to 2$

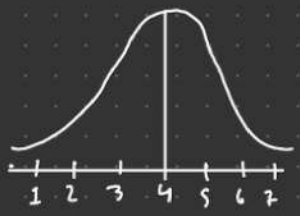
Q Away from the mean for a specific mean how much Std fall for the 4.

↳ by Z-score → $x_i = 4 \Rightarrow \frac{4-3}{1} \to \boxed{1}$

Note :-

One std to the right → When value is +ve

One std to the left → When value is -ve



$\{\mu = 4, \sigma = 1\}$ (Normal dis. fun.)

Q How many standard deviation 4.5 is away from the mean?

Ans → $x_1 = 4.5$

Z-Score → $\frac{4.5 - 4}{1}$ → 0.5

0.5 Std from the right.

Q What percentage of data is falling above 4.5

$\mu \to 4$

$\sigma \to 1$

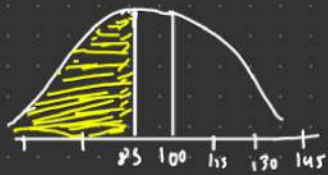
Z-Score → $\frac{2.5 - 4}{1} \Rightarrow 2.5$

Area under the curve (≤ 2.5) ⇒ 1 − 0.6681 → 6.6%

Problem → In india average IQ is 100, with a std of 15 what is the percentage of the population where you expect to have an IQ lower then 85.

$\mu \to 100$, $\sigma = 15$

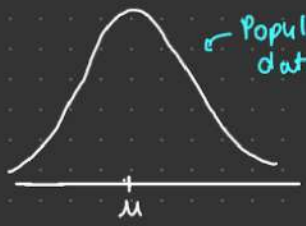
Z-Score → $\frac{85 - 100}{15} \to \frac{-15}{15} \to -1$



Area under the curve = (≤ 85) = 0.15866 ⇒ 15.8%

Central limit theorem

type → ① $X \sim N(\mu, \sigma)$



$n = 20$

Sample distribution

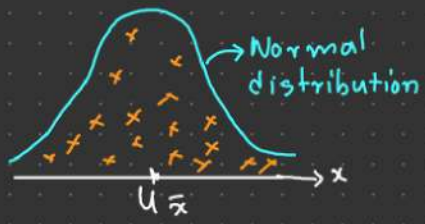
$S_1 \rightarrow [x_1, x_2, \ldots, x_{20}] = \bar{x}_1$

$S_2 \rightarrow [x_2, x_3, \ldots, x_{20}] = \bar{x}_2$

$\vdots$

$S_n \rightarrow \bar{x}_n \Rightarrow$ Sample mean.

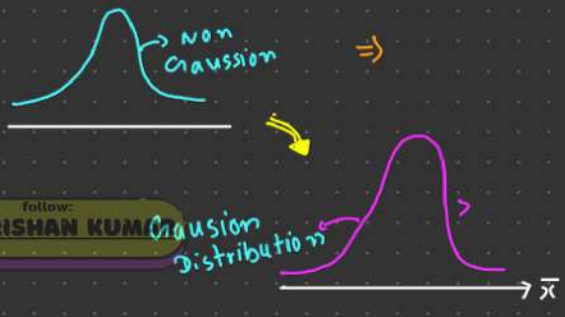
$\bar{x} = \{\bar{x}_1 + \bar{x}_2 + \ldots + \bar{x}_n\}$



If we have population data who is normally distributed. On this- we applied Sampling distri. and after that calculating sample mean.

↳ when we ploted all those mean → We got normal distribution

② $X \not\sim N(\mu, \sigma)$



$n \geq 30$

$S_1 \rightarrow \{x_1, x_2, \ldots x_{30}\} \rightarrow \bar{x}_1$

$S_2 \rightarrow \{x_2, x_3, \ldots, x_{30}\} \rightarrow \bar{x}_2$

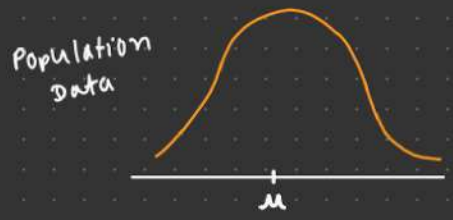
$\vdots$

$S_n \rightarrow \bar{x}_n$

⇒ CLT Properties

the central limit theoram says that the sampling dis. of the mean will be always be normally distributed, as long as the sample size is large enough. Regardless of wheather the population has a normal, poisson, binomial or any other distr. the sampling distribution of the mean will be normal.

Important for the interview



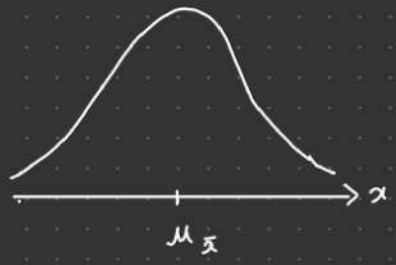
$x \simeq N(\mu, \sigma)$

$\sigma \Rightarrow$ population std

$\mu \Rightarrow$ Population mean

$n \Rightarrow$ Sample size.

Sampling distribution of mean (CLT)



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

n can be any value

Q What is the Std in sample distribution ?

↳ Std $\Rightarrow \frac{\sigma}{\sqrt{n}}$

Infrential Statistics

Estimate :- It is an observe numerical value used to estimate an unknown population parametter.

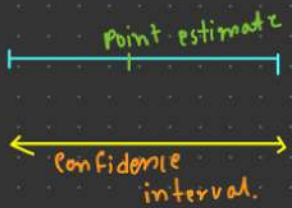
① Point estimate → Single numerical value used to estimate the unknown population perumetter.

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

There is a huge gape, to counter those gape we use Interval estimate.

② Interval estimate → Range of value used to estimate the unknown population parametter.

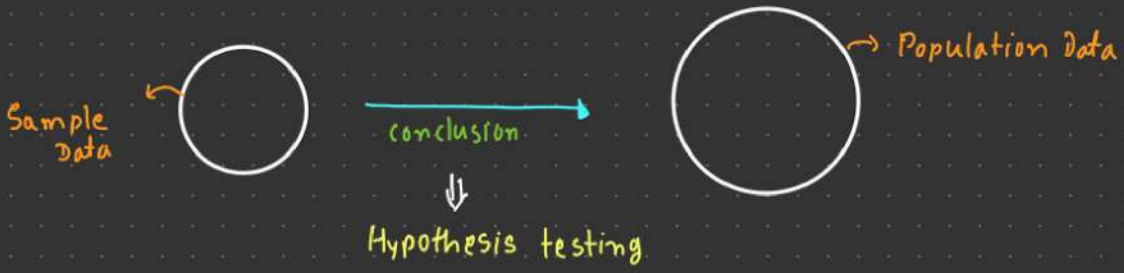
* Interval estimate of population parameter are called confidence interval.



55 - 65

A Hypothesis and Hypothesis testing mechanism :-

↳ In inferntial Stats. used for → Conclusion or Infernces.



Hypothesis testing Mechanism →

ex → Person → done Crime

(1) Null Hypothesis (H_0) → The person is not Guilty

↳ The assumption that you are begining with.

(2) Alternate Hypothesis (H_1) → The person is Guilty

↳ Opposite of Null Hypothesis

(3) Experiments → profe collect (by help of → Statistical Analysis) (P-value)

↳ {DNA, Tests, finge prints}

(4) Accept the null hypothesis OR Reject the null hypothesis

for example,

College at district a stats says its average passed percentage of statistic are 85%. A New college opened in the district and it was found that a sample of 100 Students have a pass % of 90 with a Standard deviation of 40%. Does this school have a different Passed %.

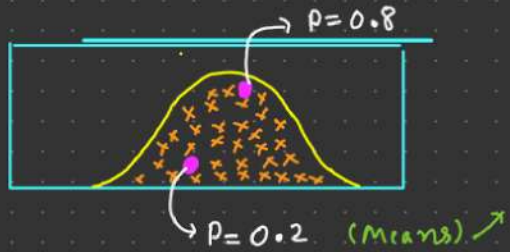
Ans → Null Hypothesis → (H_0) → $\mu = 85\%$

Alternate hypothesis → (H_0) → $\mu \neq 85\%$

P-Value

The p-value is a number, calculated from, a Statistical test, that discribe how likely you are to have found a particular set of observations if the null hypothe. were true. P-values are used in hypothesis testing to help decide wheather to reject the null hypothesis.

Ex → Using keyboards space tap.



Out of 100 touches in this key the probability of touching in this regin ≥0

↳ Hypothesis testing

Ex → Wheather coin is far or Not. {100 Tosses}

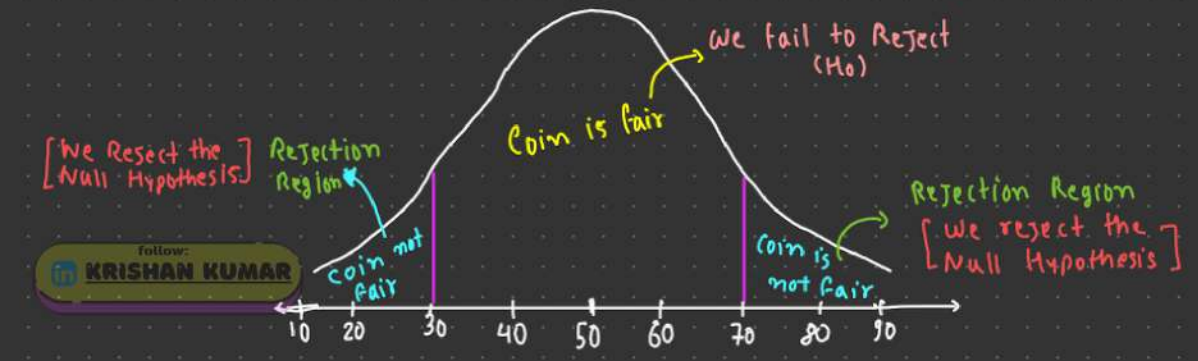
Ans →

① Null Hypothesis (H_0): Coin is fair

② Alternate hypothesis (H_1): Coin is not fair

③ Experiment :

$$\begin{cases} P(H) \to 0.5,\ P(T) \to 0.5 \\ P(H) \to 0.6,\ P(T) \to 0.4 \\ P(H) \to 0.7,\ P(T) \to 0.3 \end{cases}$$



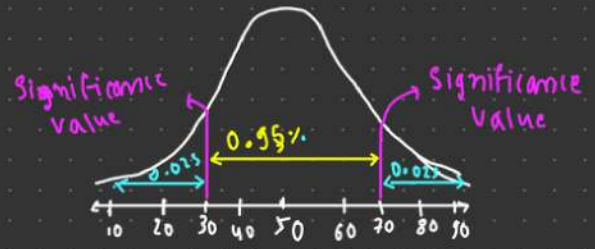
Significance value (α)

Let, $\alpha = 0.05$

Confidence interval (C.I.)

C.I. $\Rightarrow 1 - 0.05$

$\Rightarrow 0.95$



* Conclusion

Let, $P = 0.01\%$

$\rightarrow$ $P <$ Significance

$\therefore$ We reject the null hypothesis

else:

We fail to reject null hypothesis.

Hypothesis testing and statistical Analysis :-

(A) Z - Test
(B) T - Test } Average
(C) CHI SQUARE } categorical data
(D) ANNOVA } varianie

(A) Z - test :-

Condition : Z-score only apply on where 2
(i) Population std
(ii) $n \geq 30$

Problem The average height of all the residents in a city is 168cm. with a $\sigma = 3.9$, A Doctor belive that mean to be different. He measure the height of 36 individuals and found the average height to be 169.5 cm

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

Method 1 ⟹ Z - test

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

(a) Null hypothesis (H_0) = $\mu = 168cm$

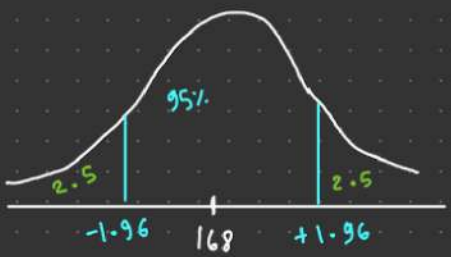
Alternate hypothesis (H_1) = $\mu \neq 168$ {2 Tail test}

It can be greater OR lesser then also.

(b) C.I. = 0.95 , $\alpha = 1 - 0.95$

$= 0.05$

Decision boundry



* Area under the curve

$1 - 0.95 \Rightarrow 0.05$

0.25 0.25

$\Rightarrow 1 - 0.25 \Rightarrow \boxed{0.9750}$

Area under the curve

Z-Score $\rightarrow 0.9750 \rightarrow +1.96$

Z-Score for population data

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

Z-Score for sample data

$$\text{Z-Score} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

* Statistical Analysis

$$\text{Z-Score} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \Rightarrow \frac{169.5 - 168}{3.9/\sqrt{36}} \Rightarrow \boxed{2.31}$$

If Z-test value is less then -19.6 OR Greater then +1.96 we

Reject the Null hypothesis

else

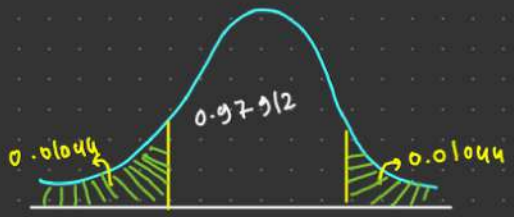
we Accept the Null hypothesis

$2.31 > +1.96$ { So, we Reject the null hypothesis }

Method 2 → p-value

* Statistical Analysis →

$$2\text{-Score} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{169.5 - 168}{3.9/\sqrt{36}} \Rightarrow 2.31$$



In Z-table $2.31 \Rightarrow 0.98956$

1 - Aera under the curve

↳ $1 - 0.98956$

$\Rightarrow$ **0.01044**

* How to calculate p-value →

$\Rightarrow$ $0.01044 + 0.01044$

$\Rightarrow$ 0.02088

if P value < Significance

$0.02088 < 0.05$ {We Rejecting the null hypothesis}

else-

we accept the null hypothesis.

↳ Here we accepting the null hypothesis.

Problem: 2 A factory manufacture bulbs with a average warenty of 5 Yr. with standard deviation of 0.50 Yr. A worker belive that the bulb will manufacture in less then 5 Yr. He test a sample of 40 bulbs and find the average time to be 4.8 Years.

(a) Stats the null hypothesis and alternate

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

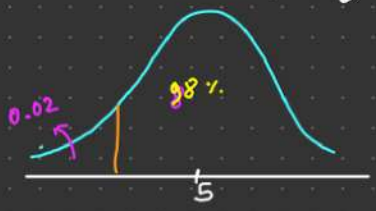
Ans → $\mu = 5$, $\sigma = 0.50$, $\bar{x} = 4.8$, $n = 40$

↳ (A) Null hypothesis (Ho) ⇒ $\mu = 5$

↳ Alternate hypothesis (H_1) ⇒ $\mu < 5$ {2 Tail test} (only checking one side)

* Method (P-Value)

Decision Boundry



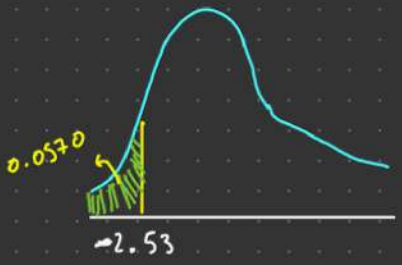
C.I. → 0.98

α = 1 − 0.98

⇒ 0.02

→ p-value →

z-test → $\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ → $\frac{4.8 - 5}{\frac{0.05}{\sqrt{40}}}$ → $\frac{-0.2}{0.079}$ → −2.53



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

P-value = 0.0570

if

Pvalue < Significance

0.0570 < 0.02 → False

* Conclusion → The warranty needs to be revised. So, we accept the Null hypothesis.

STATISTICS FOR THE DATA SCIENCE

Part - 4

- STUDENT T-DISTRIBUTION
- BAYES THEOREM
- CHI - SQUARE TEST
- F - DISTRIBUTION
- ANOVA AND ITS TYPES

#Value_freeContent

@Krishan kumar

(B) Student t-distribution → Statistical analysis using Z-score we need population standard deviation (σ).

* How do we perform an analysis when we don't know the population standard deviation.

Ans → Student t-distribution

in Z-score →

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

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

S = Sample std.

* Degree of freedom →

$$dof = n - 1$$

$n \Rightarrow$ Sample size

We 3 people

[X] [ ] [ ]

➤ We can find aera under the curve by using t-value OR dof. for t-distribution where population std (not given) only Sample std (given)

Problem → In the population the average IQ is 100. A team of researchers want to test a new medication to see if it has eighter a +ve OR -ve effects on intelligence, or no effect at all. A Sample of 30 participants who have taken the medicine has a mean of 140 with a std of 20. did the medication affect intelligence? C.I. ⇒ 95%.

Ans → $\mu = 100$, $\sigma = 20$, $n = 30$, $\bar{x} = 140$, C.I ⇒ 95%., $\alpha = 0.05$

① Null hypothesis (H_0) → $\mu = 100$

② Alternate '' (H_1) → $\mu \neq 100$ { 2 Tail test }

↓

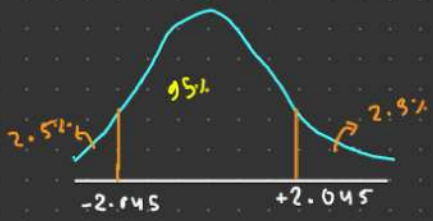
· only apply when the value of a mean is eighter ↑↑ OR ↓↓

③ $\alpha = 0.05$

④ Degree of freedom :-

$df \rightarrow 30 - 1 \rightarrow \underline{29}$

⑤ Decision boundry :-



by using t-table,

df = 29

$\alpha = 0.05$

we got ⇒ +2.045

if t-test is less then -2.045 and greater then 2.045, we reject the null hypothesis.

⑥ Calculate t-test statistic →

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \Rightarrow \frac{140 - 100}{20/\sqrt{30}} \Rightarrow \frac{40}{3.65} \Rightarrow \underline{10.96}$$

⑦ Conclusion

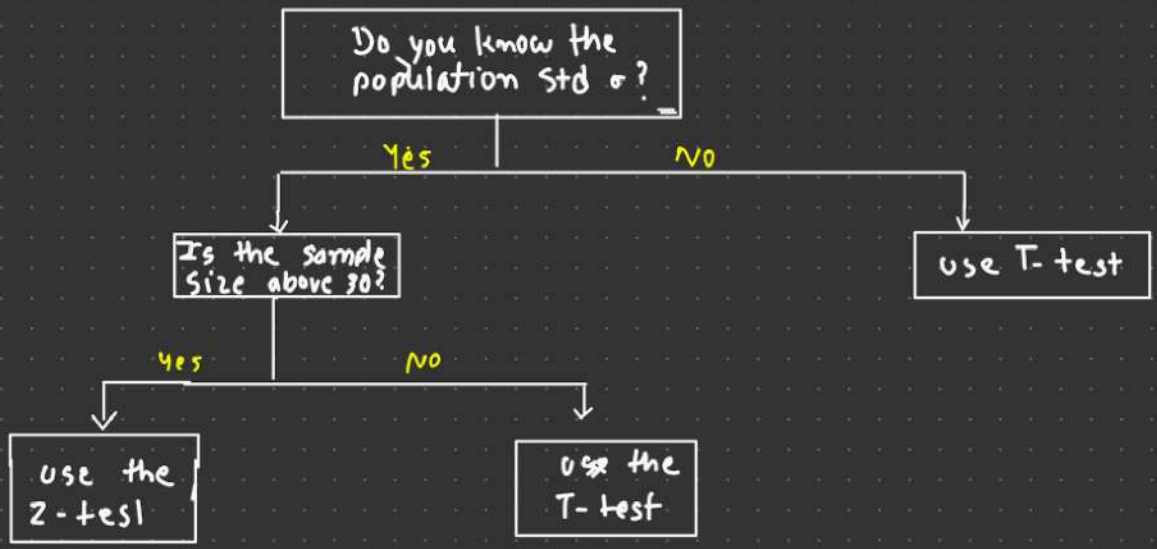
Decision Rule :- if t is less then -2.045 and greater then 2.045 we reject the null hypothesis.

$t = 10.96 > 2.045$ {we Reject the null hypothesis}

Final → Medication has increased the intelligence.

⭐ Interview Questions

when to use t-test Vs Z-test



* Types 1 and Type 2 Errors

Reality : Null hypothesis is true OR H_0 is false

Decision : H_0 is True OR H_0 is False.

Conclusion

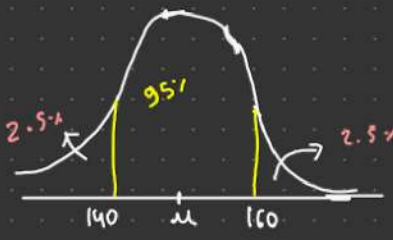
* Outcome : 1 — we reject the Ho (Null hypothesis) when in reaily it false → Good

* outcome : 2 — we reject the null hypothesis when in reaility it is True → True → Type 1 Error

* outcome : 3 — we retain (accept) the Ho, when in reality it is false → Type 2 Error

* outcome 4 : we retain the Ho, when in reality it is True → Good

Confidence Interval and Margin of Error :



$\mu \rightarrow 160$

C.I. → 95%

140 ⟷ 160

Point estimate : A value of any statistics that estimate the value of an unknown population parametter is called "point estimat".

$\bar{x} \rightarrow \mu$

2.95 , $\mu = 3$

Confidence interval: We construct a C.I. to help the estimate what the actual value of the unknown population mean is :-

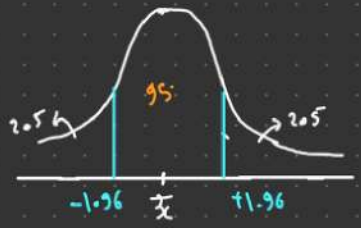
Point estimate ± Margin Error

Z-test

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

{ α = Significant value who decided How much C.I. is

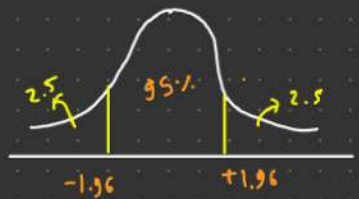
$\alpha = 0.05$



$\Rightarrow Z_{\frac{0.05}{2}} \Rightarrow Z_{0.025}$

Problem: On a 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.

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



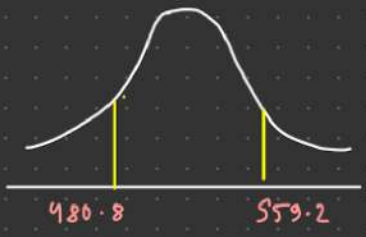
$\Rightarrow 1 - 0.25$

$\Rightarrow 0.9750$

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

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

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



→ Conclusion : I am 95% sure (confident) that the mean CAT Exam score lies between 480.8 and 559.2

Bayes theoram

Probability → Independent Event
Probability → Dependent Event.

* Independent Event

ex. Rolling a dice
{1,2,3,4,5,6}

P(1) → 1/6
⋮
P(6) → 1/6

Ex. Tossing a coin
P(H) → 1/2
P(T) → 1/2

* Dependent Event

ex

→ P(R) → 2/5
→ P(B) → 3/4

Here one event effect another events. also.

Ex.

$$P(R \text{ and } B) = P(R) \times P(B/R)$$

$2/5 \times 3/4 \rightarrow 6/20$

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

$P(A) \times P(B/A) = P(B) \times P(A/B)$

$$P(B/A) = \frac{P(B) \times P(A/B)}{P(A)}$$

Bayes theoram

$P(A/B) \rightarrow$ Conditional Probability

OR

$$P(A/B) = \frac{P(A) \times P(B/A)}{P(B)}$$

A, B → Events

$P(A/B)$ → Probability of A given B is already happ

$P(B/A)$ → Pro. of B given A is true.

$P(A)$, $P(B)$ → The independent Pro. of A and B

* Uses of Bayes theoram in ML : (ex)

Dataset / Pradict the price → (Y = ?)

Size of the House	No. of Room	Location	Price
x_1	x_2	x_3	y

By Bayes theoram →

$$P\left(\frac{y}{(x_1, x_2, x_3)}\right) = \frac{P(y) \times P\left(\frac{x_1, x_2, x_3}{y}\right)}{P(x_1, x_2, x_3)}$$

Analysis by using bayes theoram → Bayesian Statistics

(1) CHI SQUARE-TEST :-

It is non parametric test that performed categorical {ordinal, nominal} data.

→ CHI SQUARE TEST for Goodness of fit test claims about population propertions [categorical variance].

ex → (1)

	Theory	Sample
yellow bike	1/3	22
orange bike	1/3	17
Red bike	1/3	59

Applyed Goodness fit

↓ Theory categorical distribution

↳ Observed categorical distribution

ex → (2) Goodness of fit test : In a student's class of 100 student where 30 are right hand. Does this class fit the theory 12% of people are right handed.

↳

	O	E
Right handed	30	12
left handed	70 / 100	88 / 100

O → Observed value

E → Expected value

$\frac{12}{100} \times 100$

↓ Sample info

↳ Theory categorical distribution

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

≤ 50kg	50 - 75	>75
20%	30%	50%

In 2020, age of $n = 500$ individuals were sampled. Below are the Result.

<50	50-75	>75
140	160	200

using $\alpha = 0.05$ would you conclude the population difference of weights has changed in the last 10 Years.

Ans:

Expected Value

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

(2010)

$n = 500$

observed Value

<50	50-75	>75
140	160	200

(2020)

<50	50-75	>75
0.2×500 $\Rightarrow 100$	0.3×500 $\Rightarrow 150$	0.5×500 $\Rightarrow 250$

20% of 500 $\rightarrow \frac{20 \times 500}{100}$

30% of 500 $\rightarrow 0.3 \times 500$

50% of 500 $\rightarrow 0.5 \times 500$

① Null hypothesis: H_0: The Data meet the expectation

Alternate Hypothesis: H_1: The data does not meet the expectation.

② $\alpha = 0.05$, C.I. → 95%.

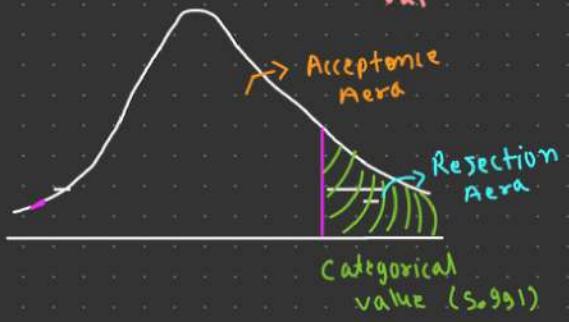
③ Degree of freedom

$$dof = k-1$$

$k \rightarrow$ No. of category

$3-1 \rightarrow 2$

④ Decision Boundry : We used CHI-Square table for the categorical value.



denote :

$\chi^2 \Rightarrow$ CHI Square (Sumbal)

If χ^2 is greater than 5.991, Reject H_0

else:

We fail to Reject the H_0

⑤ Calculate CHI-SQUARE Test-Statistic :

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

$O \rightarrow$ observed, $E \rightarrow$ expected value

$$\Rightarrow \left(\frac{40}{100}\right)^2 + \left(\frac{10}{150}\right)^2 + \left(\frac{-50}{250}\right)^2$$

$$\Rightarrow \frac{1600}{100} + \frac{100}{150} + \frac{2500}{250}$$

$\Rightarrow 16 + 0.66 + 10 \Rightarrow 26.66$

E value

<50	50-75	>75
140	160	200

O value

<50	50-75	>75
100	150	250

Final: If x^2 greater than 5.99 Reject the H_0, else fail to reject the H_0.

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

Cons: The weights of 2020 population are different than those expected on the 2010 population.

F-distribution :-

In probability theory and statistic the F-distribution or F-ratio also known ass snedecor's F-dis. or the Fisher-snedecor distribution is continous probubility distribution that aries frequently as the null dis. of a test test statistic, most notebly in the analysis of variance (ANOVA) and other F-test.

* F-test used to campare variance of mean between two group. F-test also called 'variance ratio test'

parametter: $d_1, d_2 > 0$ (degree of freedom)

Support : $x \in (0, +\infty)$

F-distribution with d_1 and d_2 degree of freedom is the dis. of.

$$\chi = \frac{S_1/d_1}{S_2/d_2}$$

- $S_1 \rightarrow$ Independent random variable { CHI Square distribution }
- $S_2 \rightarrow$ " " "
- $d_1 \rightarrow$ Degree of freedom
- $d_2 \rightarrow$ Degree of freedom

⇒ F-test:

Problem: The following data show the no. of bulbs produced daily for same day by 2 worker A and B

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

Can be consider based on the data worker B is more stable and effiecient

$\alpha = 0.05$

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

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

② Calculation of variance:

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

A

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

$S_1^2 \Rightarrow \frac{(x_1 - \bar{x})^2}{n-1}$

$S_1^2 \Rightarrow \frac{80}{6-1} \Rightarrow \frac{80}{5} \Rightarrow 13$

B

x_2	$\bar{x}$	$(x_2 - \bar{x})^2$
39	37	4
38	37	1
41	'	36
33	'	16
32	'	25
39	'	4
40	'	9
34	'	9
$\Sigma = 37$		$\Sigma = (x_1 - \bar{x})^2$ ⇒ 84

$S_2^2 = \frac{(x_2 - \bar{x})^2}{n-1}$

$S_2^2 = \frac{84}{8-1} \Rightarrow \frac{84}{7} \Rightarrow 12$

→ Calculation of variance Ratio (F-test):

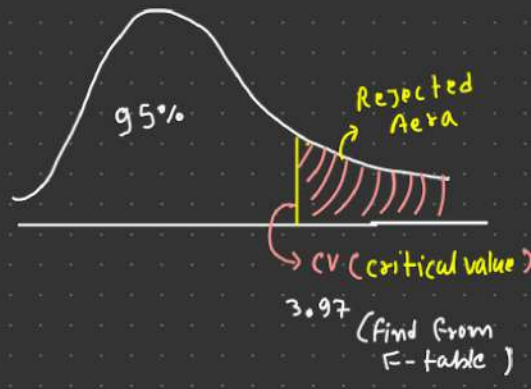
$$F = \frac{S_1^2}{S_2^2} \Rightarrow \frac{13}{12} \Rightarrow 1.33$$

(3) Decision Rule:

$df_1 \rightarrow 6-1 \Rightarrow 5$

$df_2 \rightarrow 8-1 \rightarrow 7$

$\alpha \rightarrow 0.05$



If F-test greater than 3.97, Reject the NULL Hypothesis

$1.33 < 3.97$, {so, we Reject the H_0}

Final conclusion :- Worker B is not efficient when compare to worker A.

(ANOVA) Analysis of variance :

Anova is a statistical method to compared the mean of 2 or more group.

ANOVA :

① factors (variable)

② levels

Ex ① Factors = Medicine

level = 5mg, 10 mg, 15mg [Dosage]

Ex ① Modle of Payment [Factors]

G-Pay Phone-Pe IMPS NEFT [levels]

⇒ Assumptions in ANOVA:

① Normality of Sampling Distribution of mean The dis. of Sample mean is normaly distributed.

② Absense of outlier.

Outlier score need to be Remove from dataset.

③ Homogenity 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 independe variable are equal.

④ Sample are independent and Random.

⇒ Types of ANOVA:

① One way ANOVA
② Repeated Measured ANOVA
③ Factorial ANOVA

① One way ANOVA: One factor with at least 2 levels, there levels are independent.

② **Repeated Measured anova:** One factor with at least 2 levels, levels are dependent.

③ **Factorial ANOVA:** Two or more factor (each of which with after 2 level), levels can be eighter dependent or independent.

⇒ Hypothesis testing in ANOVA:

Null Hypothesis: H_0: $\mu_1 = \mu_2 = \mu_3, \ldots \mu_n$

H_1: At least one of the mean is not equal

$\times$ $\boxed{\mu_1 = \mu_2 = \mu_3 = \ldots \mu_n}$

Test statistic:

$$F = \frac{\text{Variation between sample}}{\text{Variation within sample}}$$

X_1	X_2	X_3
1	6	5
2	7	6
4	3	3
5	2	2
3	1	4
$\Sigma X_1 = 15$	$\Sigma X_2 = 14$	$\Sigma X_3 = 20$
$\bar{X} = 15/3 \Rightarrow 3$	$\bar{X_2} = 14/5$	$\bar{X_3} = \frac{20}{5} \Rightarrow 4$

H_0: $\bar{X_1} = \bar{X_2} = \bar{X_3}$

H_1: at least one sample mean is not equal

* One way anova :

One factor at least 2 levels, levels are independent

Problem : Doctor want to test a new medication which reduce headach. They split the participant into 3 condition [15mg, 30mg, 45mg]. later on the doctor ask the patient to rate the headach between [1 → 10] are there any difference between 3 condition using Alpha (α) → 0.05 ?

Ans:

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

① Define null Hypothesis :

H_0 : $\mu_{15} = \mu_{30} = \mu_{45}$

H_1 : not all μ's are equal

② State significant value :

$\alpha = 0.05$

C.I. → 0.95

③ Calculate degree of freedom :

$N = 21$, $a = 3$, $n = 7$ (a → Number of Sample)

$df_{between}$: $a - 1 \rightarrow 3 - 1 \Rightarrow 2$

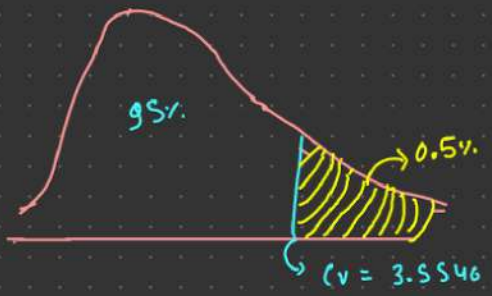
df_{within} : $N - a \rightarrow 21 - 3 \Rightarrow 18$

df_{total} : $N - 1 \rightarrow 21 - 1 \Rightarrow 20$

(2, 18)

↳ super useful to finding critical value in table

F - Table



Critical Value → 3.5546

④ **State decision Rule :**

If F is greater than 3.5546, we Reject the H_0.

⑤ **Calculate test Statistic :**

SS = Sum of Square

$SS_{between}$ 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

⑥ $$SS_{between} = \frac{\Sigma (\Sigma a_i)^2}{n} - \frac{T^2}{N}$$

15 mg → 9+8+7+8+8+9+8

30 mg → 7+6+6+7+8+7+6

45 mg → 4+3+2+3+4+3+2

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

$$\Rightarrow 98.67$$

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

{y^2 ⇒ Given table value में हर element का square करना}

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

$\Sigma y^2 \Rightarrow 9^2+8^2+7^2+8^2+\ldots\ldots+3^2+2^2 \Rightarrow 853$

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

$\Rightarrow 10.29$

(3) $SS_{Total}: \Sigma y^2 - \frac{T^2}{N}$

$\Rightarrow 853 - \frac{125^2}{21}$

$\Rightarrow 108.95$

	SS	df	Ms	F
Between	98.67	2	49.34	
within	10.29	18	0.54	
Total	108.95	20	49.88	

{$MS = SS/df$}

F ⇒ $\frac{Ms\ between}{Ms\ Within}$ is equal to ⇒ $F = \frac{\text{Variation between sample}}{\text{variation within sample}}$

$\Rightarrow \frac{49.34}{0.54} \Rightarrow 86.56 \Rightarrow$ **F = 86.56** (1)

Final:

If F is greater than 3.5546, we Reject the Ho

86.57 > 3.5546, So we Reject H_0.