Introduction to stats

Descriptive stats

* Measure of central tendency
* Measure of Dispersion
* Summarizing the data
  Histograms, Pdf, cdf,
  Probability, Permutations
  Mean, Median, Mode, Variance
  and Standard deviation
* Gaussian Distribution
* Lognormal Distribution
* Biomial distribution
* Standard Normal distribution
* Transformation & Standardization
* Q-Q plot

Inferential stats

* z test
* t-test
* ANOVA
* CHI square
* Hypothesis testing
* Confidence Intervals
* z-table, t-table

What is statistics?

Statistics is the science of collecting, organizing and analyzing the data. { Better decision making }

Data :- fact or piece of information that can be measured

Eg:- The IQ of a class

{ 10, 12, 16, 19, 23, 25 }

Types of statistics:

(i) Descriptive stats :- It consists of organizing and Summarizing.

Eg:- What is the avg of the student in the class.

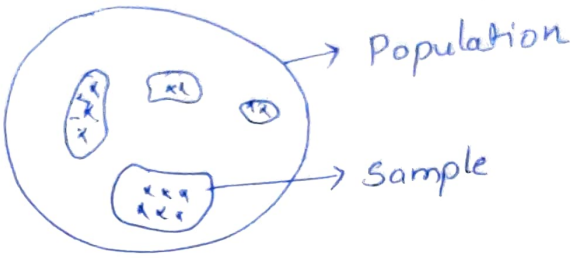
(ii) Inferential stats :- Technique where in we used the data we have measured to form Conclusion.

Eg:- Are the marks of the student of the classroom similar to the age of the math classroom in the College.

Population & Sample :—

Elections → Goa, UP

Population (N) Sample (n)



(i) Simple Sample Sampling :— Every member of the population(N) has equal chance of being selected for your Sample(n).

(ii) Stratified Sampling :— Where the population(N) is split into non-overlapping groups.

Eg :— Gender — Male, Female

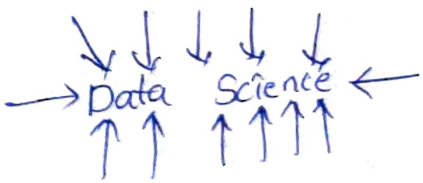
Age Grops :
(10-20) (20-30) (30-40)

(iii) Systematic Sampling :— Consider every n^{th} indivisual from Population.

N → n^{th} indivisual

Eg :— Shopping Mall → Covid Survey
↳ 8^{th} person enter mall.

(iv) Convenience Sampling :—
Sample from the only had knowledge on the domain.
(Domain expert knowledge) → only those people



Variable :- A variable is a property that can take on any value.

Eg:- {128, 173, 142, 162, 168}

Height, weight, age, marks

Two kinds of variable :

(i) Quantitative Variable :- Measured Numerically. Where we can Perform arthematic operations (+, -, *, ÷)

(ii) Qualitative variable / categorical :- Based on some characterstic we can derive categorical variables.

Eg:- Gender — Male, Female.

Quantitative

Discrete

Eg:- Whole number

(i) No. of Bank Account

(ii) No. of children family

Continous

Eg:- non-finite numbers

Height = { 121.6, 137.01, 198.06 }

Weight = { 98.6, 77.7, 88.2 }

Rain fall = { 0.2, 0.8, 1.1 }

Variable Measurements :-

→ 4 types of Measured variable

① Nominal :- Categorial data classes, color, Gender, type of flower

② Ordinal :- Order of the data matter, value does not

③ Interval :- Order matters, value also matters, natural zero is not present.

④ Ratio :-

Eg:-

Student (marks)	Rank
100	1
96	2
57	3
85	4

ordinal data

Temperature in farenheit

70-80 , 80-90, 90-100 →

Frequency Distribution:-

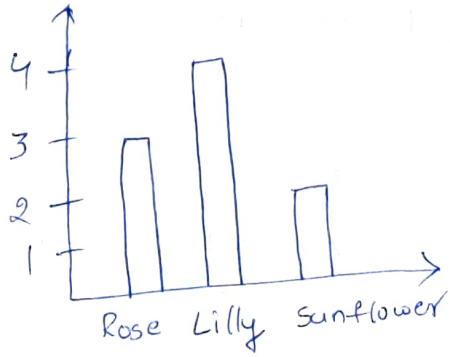
Sample Dataset :-

Rose, Lilly, sunflower, Rose, Lilly, Sunflower, Rose, Lilly, Lilly

Flower	frequency	Cummulative frequency
Rose	3	3
Lilly	4	7
Sunflower	2	9

Bar graph:-

→ Bar graph is used on discrete data



Histogram :—

→ Histogram is used on Continous data

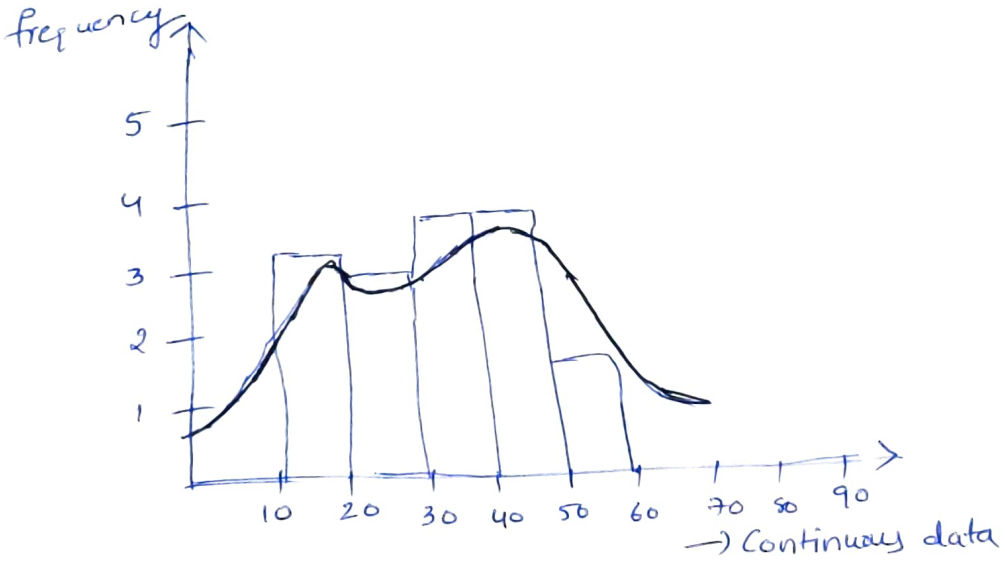
Age :- {10, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 41, 42, 43, 50, 51}

Assume, Bins = 10

Pdf : probability distribution function

↳ used to smoothing of histogrm

(Kernal density ~~function~~) Estimator)



Bar ↓ Discrete Vs Histogram ↓ Continuous

① Measure of central tendency
② Measure of Dispersion
③ Gaussian Distribution
④ Z-score
⑤ Standard Normal distribution

(i) Arithmetic mean for population & Sample :-

Mean (Average)

Population (N) Sample (n)

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

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

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

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

Central tendency :- Refers to the measure used to determine the centre of the distribution of data.

① Mean ② Median ③ Mode

(ii) Median :-

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

Suppose add 100 to the data

$x = \{1, 1, 2, 2, 3, 3, 4, 5, 5, 6, 100\}$

Mean $= \frac{132}{11} = 12$, Mean becomes 12, it is very vary from before (3.2)

→ 100 is outlier → In such cases use Median

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

steps:-

(i) Sort the numbers (ASC/DSC)

(ii) Consider middle value (if n = odd)

or

take avg of two middle values (if n = even)

Median = 3

if x = {1, 1, 2, 2, 3, 3, 4, 5, 5, 6, 100, 112}

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

→ Median works well with outlier

* outlier : The value completely different from distribution.

Mode :-

→ Most frequent element in data

x = {1, 2, 2, 3, 4, 5, 6, 6, 6, 7, 8, 100, 200}

Mode = 6

Where specifically used :-

Mode works both numerical and categorical data. But it suite most to categorial.

Eg:-

Type of flower	Petal length	Sepal length
Rose		
Lilly		
Sunflower		
—		
—		
—		

{ Missing value → Most frequent value (Mode) }

→ Specifically works with Categorical.

Measure of Dispersion:-

↳ Dispersion means spread.

(i) Variance
(ii) Standard deviation

* How 2 distributions are distributed, at that point of time, we use variance.

Population Variance

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

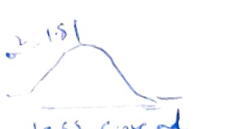
Sample Variance

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

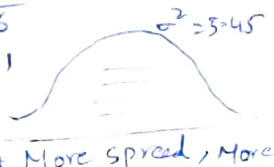
x	μ	$x-\mu$	$(x-\mu)^2$
1	2.83	-1.83	3.34
2	2.83	-0.83	0.6889
2	2.83	0.83	0.6889
3	2.83	0.17	0.03
4	2.83	1.17	1.37
5	2.83	2.17	4.71

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

$$= 1.81$$

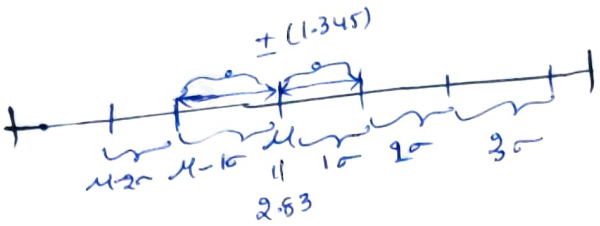


less spread



→ More spread, More elements

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



Variance = Spread

S.D = What range of value falls.

⊛ Percentiles & Quartiles ⇒ [Find outliers]

Percentage : 1, 2, 3, 4, 5

% of the numbers that are odd ?

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

$$= 3/5 = 0.6 = 60\%.$$

Percentile :-

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

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

(Q) what is the percentile ranking of 10 ?

$$\text{percentile Rank of } 10 = \frac{\text{\# of values below 10}}{n} \times 100$$

$$= 16/20 \times 100$$

$$= 80\%.$$

* This means 80 % of the data lie below 10.

(Q) percentile rank of 11 $= \frac{17}{20} \times 100$

$= 85\%$ [85% elements are below 11]

(Q) What is value exists at percentile ranking of 25%?

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

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

$= 5.25 \Rightarrow$ [Index position]

Value at index 5 is {5}

Q) What is value at percentile of 75%?

$$\text{Value} = \frac{75}{100} \times (20+1)$$

$= 15.75$

Value at 16th position is {9}.

Five number Summary :—

(i) Minimum
(ii) First Quartile (Q_1)
(iii) Median
(iv) Third Quartile (Q_3)
(v) Maximum

Removing the outlier :—

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

less than values are outliers ⟸ [Lower fence, Higher fence] ⟹ greater than values are outlier

$$\text{Lower fence} = Q_1 - 1.5(IQR)$$

$$\text{Higher fence} = Q_3 + 1.5(IQR)$$

Interquartile Range (IQR):

$$IQR = Q_3 - Q_1$$

$Q_3 = 75\%$.
$Q_1 = 25\%$.

value at 75% percentile (Q_3) $= \frac{75}{100} \times (20)$

$Q_3 = 15 \Rightarrow$ index $\Rightarrow \{7\}$

value at 25% percentile (Q_1) $= \frac{25}{100} \times (20)$

$Q_3 = 5$ index $\Rightarrow \{3\}$

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

Lower fence $= Q_1 - 1.5 \times (IQR)$

$= 3 - 1.5(4)$

$= -3$

Higher fence $= Q_3 + 1.5(IQR)$

$= 7 + 1.5(4)$

$= 7 + 6$

$= 13$

outlier $< [-3, 13] >$ outlier [27 is outlier]

Remaining data $= \{1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9\}$

Minimum $= 1$

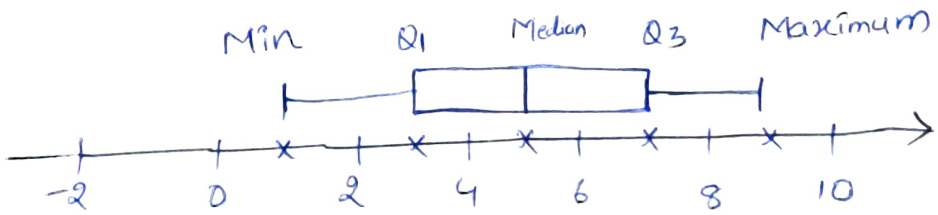
$Q_1 = 3$

Median $= 5$

$Q_3 = 7$

Maximum $= 9$

Box plot :-



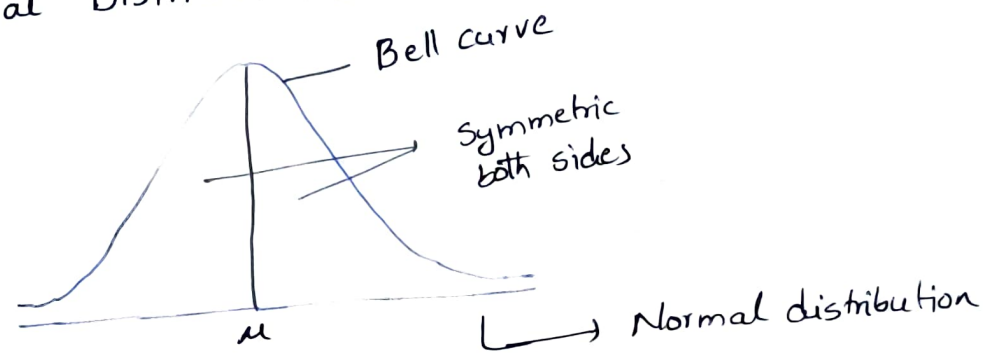
Distribution

(i) Normal Distribution
(ii) Standard Normal Distribution
(iii) Z-score
(iv) Log Normal Distribution
(v) Bernoulis Distribution
(vi) Binomial Distribution

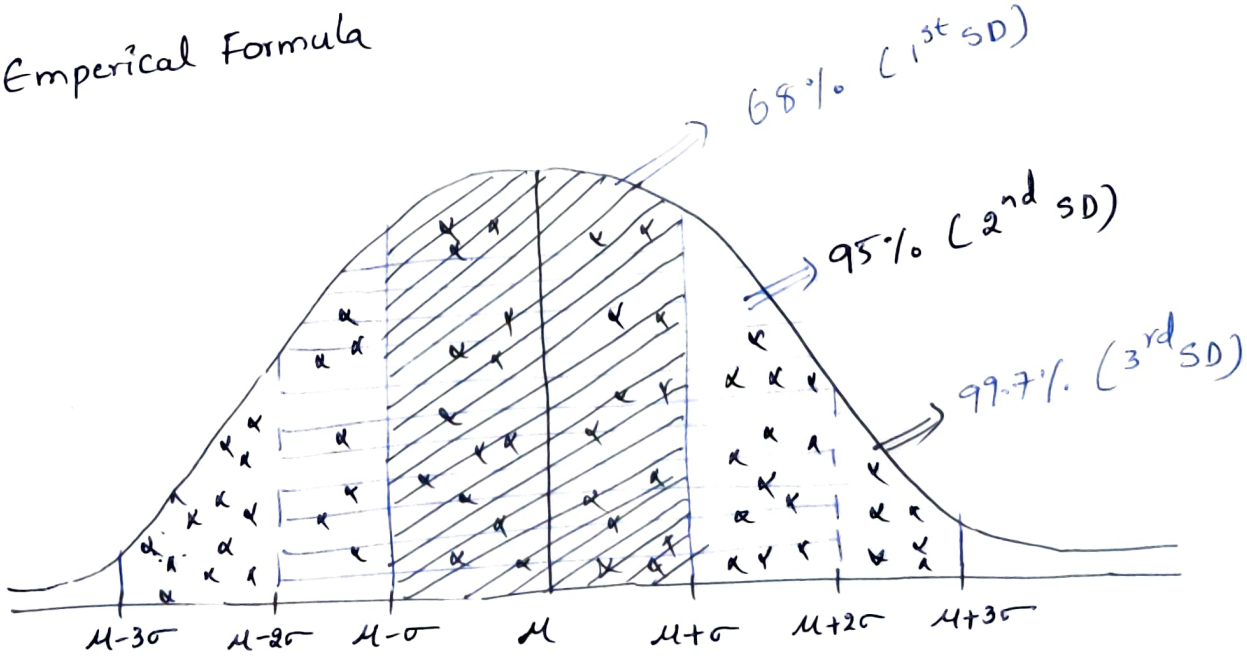
Practicals

(i) Mean, Median, Mode
(ii) Variance, SD
(iii) Histogram, Pdf, Bar plot, Violon plot
(iv) IQR
(v) Log Normal Distribution

Gaussian / Normal Distribution :—



Emperical Formula



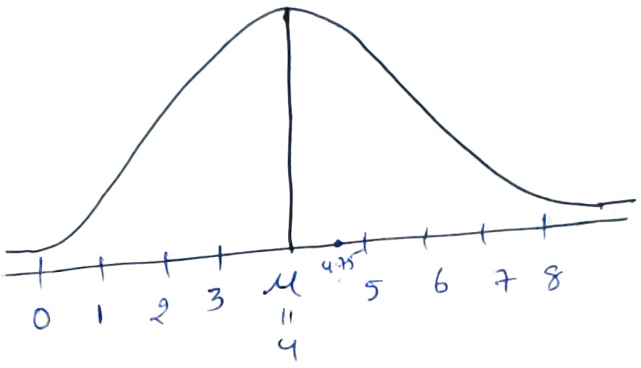
formal => 68 - 95 - 99.7%

Eg:

Height → Normally distributed
weight → " "
IRIS dataset → Normally distributed

Example :-

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



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

if $x = 4.75$

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

$z = 0.75$ sd to right

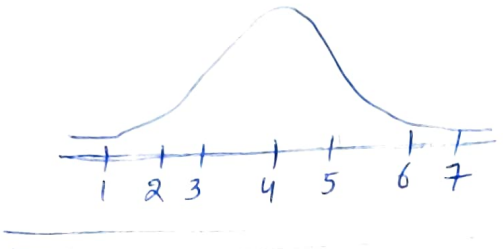
if $x = 3.75$

$$Z_{score} = \frac{3.75 - 4}{1}$$

$= -0.25$ [S.D to left]

Example :—

data = { 1, 2, 3, 4, 5, 6, 7 }



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

⇓

Z-score

⇓

{ -3, -2, -1, 0, 1, 2, 3 }

↳ Standard Normal distribution

which satisfy ⇒ $\mu = 0$, $\sigma = 1$

↳ A Random variable y specifically satisfy ⇒ { $\mu = 0$, $\sigma = 1$ }

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

$z_1 = \frac{1-4}{1} = -3$

$z_2 = \frac{2-4}{1} = -2$

$z_3 = \frac{3-4}{1} = -1$

$z_4 = \frac{4-4}{1} = 0$

$z_5 = \frac{5-4}{1} = 1$

$z_6 = \frac{6-4}{1} = 2$

$z_7 = \frac{7-4}{1} = -3$

{ -3, -2, -1, 0, 1, 2, 3 }

Dataset :—

Age (years)	Salary (Rs) (units)	Weight (kg)
21	40k	41 kg
26	30k	25 kg
31	10k	33 kg
26	55k	56 kg
11	60k	69 kg

⇒ Apply standard deviation

⇓

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

⇓

This process is called standard deviation

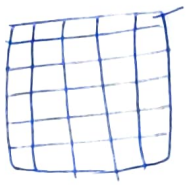
Normalization :- (Range => (0,1))

↳ change all values between range (0,1) or Interval.

MinMax scalar used to perform Normalization

↳ It changes all changes between (0,1)

CNN → Image classification



0, to 255

⇓

divide by 255

↓

It becomes [0,1]

Example 6

(Q) ODI Series

Series average in 2021 = 250	Series average score 2020 = 260
S.D of score = 10	S.D of score = 12
Rishabpant avg score = 240	Rishabpant avg score = 245

Q1) Compare both series in which year Rishab pant score way better?

2021

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

$$= \frac{270-250}{10}$$

$$= -1$$

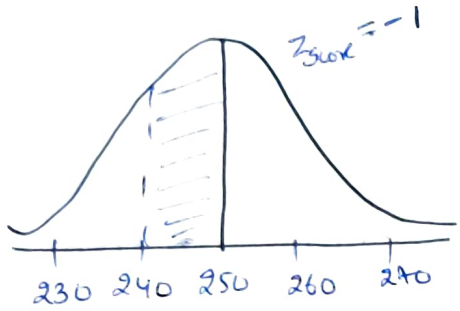
2020

$$Z = \frac{245-260}{12}$$

$$= -1.25$$

In 2021

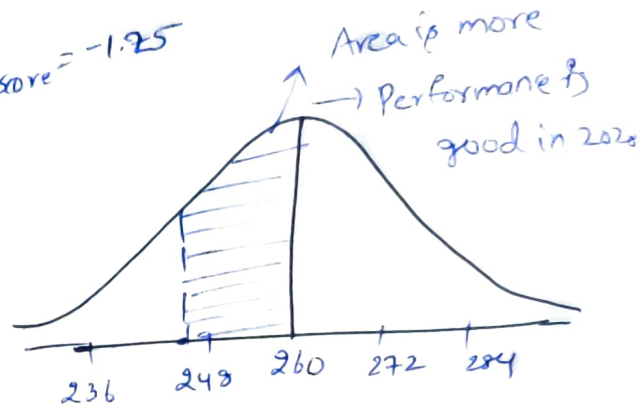
$\mu = 250$, $x_i = 240$, $\sigma = 10$



In 2020

$\mu = 260$, $x_i = 245$, $\sigma = 12$

$Z_{score} = -1.25$



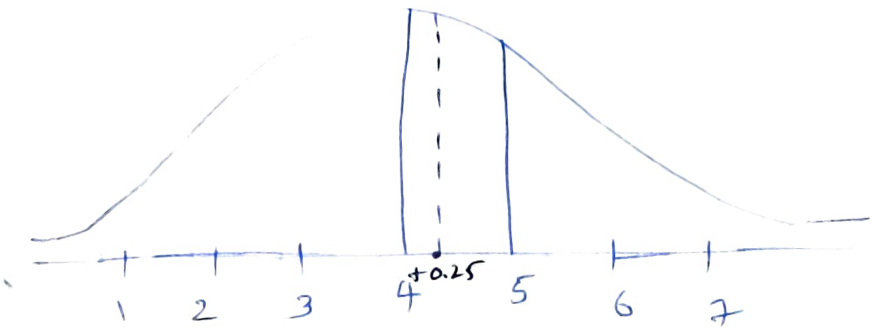
Stats Interview Questions

(Q) What is percentage of score fall above 4.25 ?

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

$$Z_{score} = \frac{x_i - \mu}{\sigma} = \frac{4.25 - 4}{1}$$

$$= 0.25 \quad [\text{S.D to left}]$$



(*) look value at 0.25 in z-table

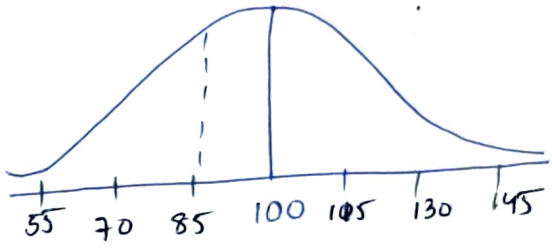
(Q) In india, the avg IQ is 100, with standard deviation of 15, what percentage of the population would you except to have an IQ lower than 85.?

Here given

$\mu = 100$

$x_i = 85$

$\sigma = 15$



$$Z_{score} = \frac{x_i - \mu}{\sigma}$$

$$= \frac{85 - 100}{15}$$

$$= -1 \Rightarrow \text{look at z-table}$$

Value at -1 in z-table $= 0.84$

$= 84\ \%$

Day 4 :-

(i) IQR python
(ii) Probability
(iii) Permutation & Combinations
(iv) Confidence Interval
(v) P_value
(vi) Hyphothesis testing

$$Z_{score} = \frac{x_i - \mu}{\sigma}$$

(*) using z-score and Emperical formala, Any value outside 3rd standard deviations are outliers.

Probability :-

Probability is the measure of the likelihood of an event.

Roll a dice = {1, 2, 3, 4, 5, 6}

$$pr(x) = \frac{\text{\# of way an event can occur}}{\text{Total no. of outcomes}}$$

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

Toss a coin = {H, T}

$$P(H) = 1/2$$

Addition Rule :- (or)

Mutual Exclusive Events : Two events are mutual exclusive if they cannot occur at the same time

Eg :- Roll a dice
Toss a coin

Non-Mutual Exclusive Events :- Multiple event can occur at the same time.

Eg :- Deck of cards

(Q) If I toss a coin, what is the probability of the coin landing on head or tail?

* They are mutual exclusive

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

Roll a dice $= \{1, 2, 3, 4, 5, 6\}$

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

(Q) you are picking a card randomly from a deck. what is the probability of choosing a card that is queen or heart ?

* This Non-mutual exclusive events

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

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

Multiplication Rule :-

Independent Events :-

Each event is independent

Roll a dice $= \{1, 2, 3, 4, 5, 6\}$
↳ each is independent

Dependent Events :—

=) Events are dependent on next occuring events.

Box

3 Red
2 green

$\Rightarrow P(R) = 3/5$

Total = 3+2
= 5

After drawing red ball

Remaining

2 Red
2 green

$\Rightarrow P(g) = 2/4$

Total = 4

Example for Independent Event

(Q) What is the probability of rolling '5' and then "4" in dice ?

* This is Independent event

use multiplication Rule here

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

$P(5 \text{ and } 4) = P(5) * P(4)$

$= 1/6 * 1/6$

$= 1/36.$

Example for Dependent Event :—

(Q) What is the probability of drawing a 'queen' and then 'ace' from a deck of cards ?

→ These are dependent event → Conditional probability

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

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

$= 4/52 * 4/51$

Permutations & Combinations :—

Permutations :—

Chocolate factory = { Dairy, 5 star, Milky bar, Eclas, Silk }

Total = 5

Student need fill three places saw

1st place — 4 chance — 3 - chance

hay

5 chance

$= 5 \times 4 \times 3$

Permutations $= {}^{n}P_{r} = \frac{n!}{(n-r)!} = \frac{5!}{2!} = \frac{5 \times 4 \times 3 \times 2 \times 1}{2 \times 1}$

$= 60$

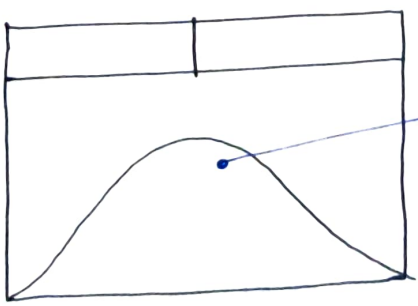
Combinations :—

$${}^{n}C_{r} = \frac{n!}{(n-r)!\, r!}$$

Eg :— ${}^{5}C_{3} = \frac{5!}{(2!)(3!)} = \frac{5 \times 4 \times 3 \times 2 \times 1}{3 \times 2 \times 1 \quad 2 \times 1}$

$= 10$

P_value :—



Every 100 times we touch 80% here.

i.e $p = 80\%$

Mouse pad

Hypothesis testing, Confidene Interval And Significance Value ?

(Q)
Coin → Test whether this coin a fair or not by perform 100 tosses?

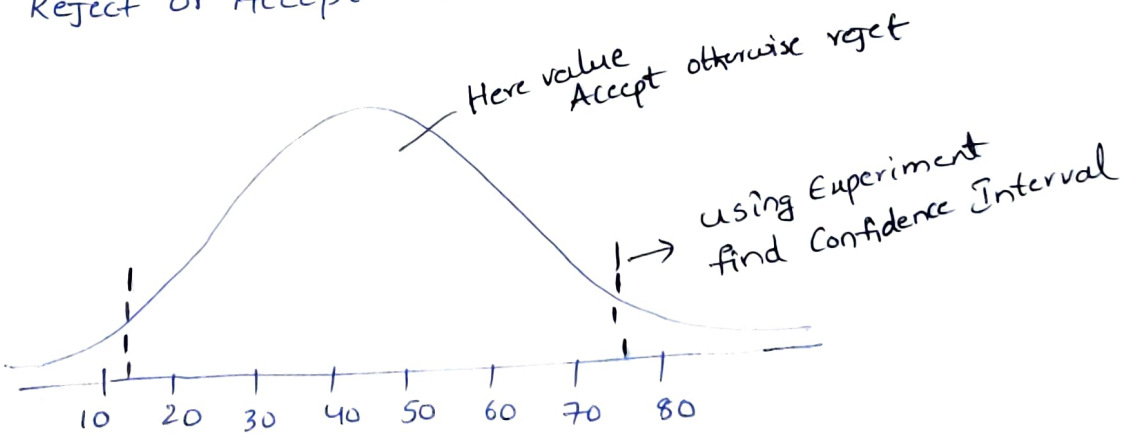
Hyphothesis testing :—

(i) Null Hyphothesis : Coin is fair

(ii) Alternative Hyphothesis : Coin is unfair

(iii) Experiment [z-test / t-test / f-test ----]

(iv) Reject or Accept the Null hyphothesis



Significance value = 0.05
↳ decided by Domain expert

Day-5 :—

① Type I or Type II Error

② One tail and 2-tail test

③ Confidence Interval

④ z-test, t-test, chi-square test

(1) Type I or Type II Errors :-

Null hyphothesis (H_0) = Coin is fair

Alternative hyphothesis (H_1) = Coin is unfair

Reality check :-

Null hyphothesis is True or Null hyphothesis is false

Decision :-

Null hyphothesis is True or Null hyphothesis is false.

outcome 1 :-

*) We ~~reject~~ reject the null hyphothesis, when in reality it is false
↳ good.

outcome 2 :-

* We reject the null hyphothesis, when in reality it is True.
⇓
[Type I Error]

outcome 3 :-

* we retain or accept Null hypothesis, when in reality it is false.
⇓
[Type II Error]

outcome 4 :-

* we accept Null hypothesis, when in reality it is True.
↳ good.

Confusion Matrix :—

	P	N
T	TP	TN
F	FP	FN

FP → Type I Error

FN → Type II Error

1 tail and 2-tail test :—

Eg: Colleges in karnataka have an 85% placements rate. A new college was recently opened and it was found that a sample of 150 students had a placements rate of 88 % with a S.D 4 % . Does this college has a different placement rate ?

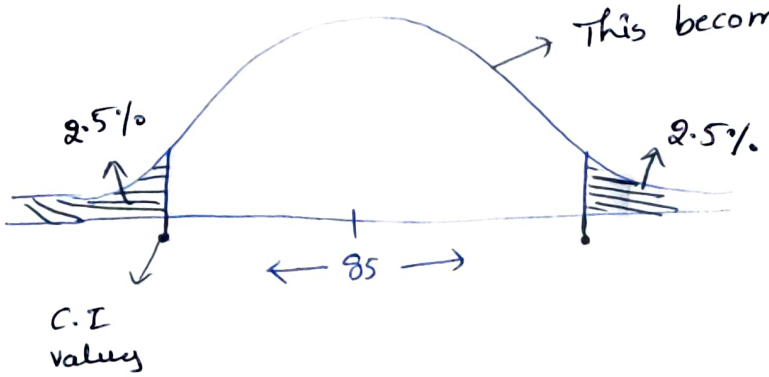
Suppose = $\alpha = 0.05$

Confidence Interval (CI) $= 1 - \alpha$

$= 1 - 0.05$

$= 0.95$

$= 95\%$

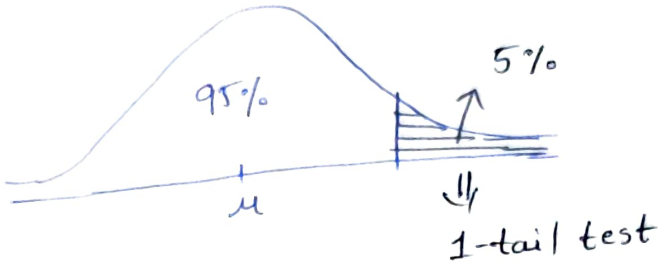


Let say

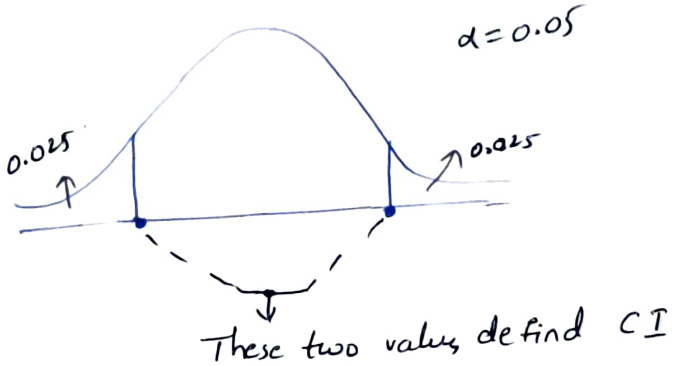
Does this college have a placements greater than 95%?

$\alpha = 0.05$

$CI = 95\%$



Confidence Interval :-



Point Estimate :- The value of any statistics that estimates the value of a parameter.

Inferential stats :-

Sample		Population
mean	⟶	Mean
$\bar{x}$	estimate approximately ⟶	μ

Eg :-

$\bar{x} = 2.9 \Longrightarrow \mu = 3$

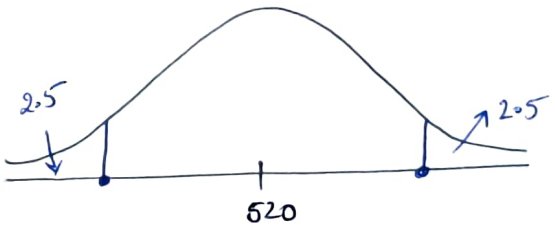
Confidence Interval = point estimate ± Margin of error

(Q) On the quant test of CAT exam, the population standard deviation is known to be 100. A sample of 25 test takers has a mean of 520 score. Construct 95% CI about the mean?

Sol:

$\sigma = 100 \quad n = 25 \quad \bar{x} = 520$

$CI = 95\% \implies \alpha = 0.05$



(i) population SD is given
(ii) $n \geq 30$
} [Z - test]

Point Estimate ± Margin of Error

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

$\because \left[\frac{\sigma}{\sqrt{n}} \Rightarrow \text{Standard Error.}\right]$

Upper Bound $= \bar{x} + Z_{\frac{0.05}{2}} \frac{\sigma}{\sqrt{n}}$

$= 520 + (1.96)\left(\frac{100}{\sqrt{25}}\right)$

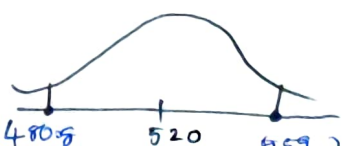
$= 559.2$

Lower Bound $= \bar{x} - Z_{\frac{0.05}{2}} \frac{\sigma}{\sqrt{n}}$

$= 520 - (1.96)\left(\frac{100}{\sqrt{25}}\right)$

$= 480.8$

Range = [480.8, 559.2]



(Q) find the avg size of the shark throughout the world?

Assume

$\alpha = 0.05$ $\quad \bar{x} = 1050$ $\quad n = 50$

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 1050 + z_{0.05/2} \frac{1050}{\sqrt{50}}$$

$$= 1050 + 291.045$$

$$= 1341.045$$

⊗ if population SD is not given then use t-test.

(Q) on the quant test of CAT exam, a sample of 25 test takers has a mean of 520, with S.D of 80. Construct 95% CI about the mean?

Sol:-

Here population SD is not given. This situation use t-test.

$n = 25$ $\quad \bar{x} = 520$ $\quad s = 80$

$\alpha = 0.05$

formula:-

Point Estimate ± Margin of Error

$$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$

$$\text{Upper Bound} = \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}} \qquad \text{Lower Bound} = \bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}}$$

Degree of freedom $= n-1$

$= 25-1$

$= 24$

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

$UB = 520 + 2.064\left(\frac{80}{5}\right)$

$= 553.024$

$LB = 520 - 2.064\left(\frac{80}{5}\right)$

$= 486.97$

$\Rightarrow [486.97, 553.024]$

One sample z-test:—

(Q) In the population, the avg IQ is 100 with S.D of 15. Researcher wants to test a new medication to see if there is positive or negtive effect on intelligence, or no effect at all of Sample 30 participants who have taken the medication has a mean IQ of 140. Did the medication effect the intelligence?

Sol:

(i) Define Null hyphothesis

$\mu = H_0 = 100$

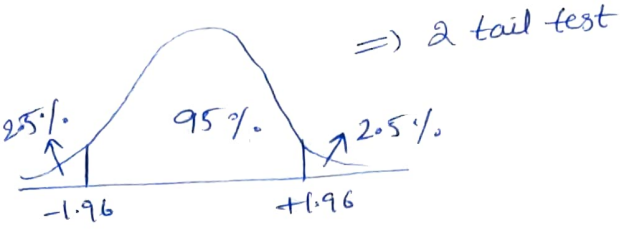
ii) Alternative Null hyphothesis

$H_1 = \mu \neq 100$

iii) State α value

$\alpha = 0.05$

iv) State decision rule



$\Rightarrow$ 2 tail test

$$Z_{\alpha/2} = \frac{Z_{0.05}}{2} = Z_{0.025} = Z_{1-0.025} = Z_{0.975}$$

$$= 1.96$$

(v) Calculate z-statistics

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

for one sample $\sqrt{n} = 1$

$$= \frac{140 - 100}{\frac{15}{\sqrt{30}}}$$

$$= \frac{40}{15} \times \sqrt{30}$$

$$= 14.60$$

vi) State out decision

$$14 > 6.0 > 1.96$$

If z value is less than -1.96 or greater than 1.96 $\Rightarrow$ reject Null hyphothesis.

Medication improved Intelligence ? $\Rightarrow$ yes improved.

one Sample t-test

Population avg IQ = 100

$n = 30$, $\bar{x} = 140$, $s = 20$

Did the medication effect intelligance?

$\alpha = 0.05$

(i) Null hyphothesis (H_0) = $\mu = 100$

(ii) Alternative hyphothesis (H_1) = $\mu \neq 100$

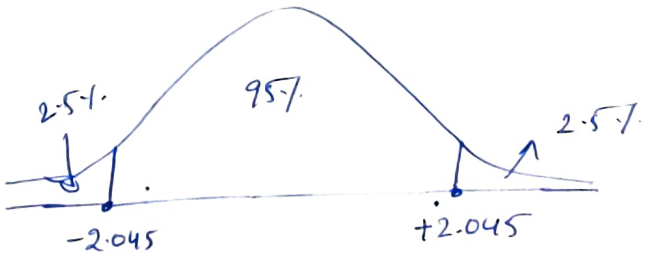
(iii) Calculate degree of freedom

$= n - 1$

$= 30 - 1$

$= 29$

(iv) State Decision rule



(v) Calculate t-statistics

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

$$= 2 \times \sqrt{30}$$

$$= 10.95$$

$10.95 > 2.045 \Rightarrow$ Reject Null hyphothesis

$\Rightarrow$ Improved intelligence.

Day-6 ✓

① CHI square
② Covariance
③ Pearson Corelation Coefficient
④ Spearsman Rank Corelation
⑤ practical Implementation
z-test, t-test, chi square test
⑥ f-test (Anova)

CHI square test :—

* Chi square test claims about population proportions

→ It is a non parametric test that is performed on categorical (nomial or ordinal) data.

(Q) In the 2000 Indian census, the ages of the indivial in a small town were to be the following :

Less than 18	18-35	> 35
20%	30%	50%.

Sol :—

In 2010, Ages of sample $n = 500$ indiviuals were sampled. Below are the results

<15	18-35	>35
121	288	91

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

Solb

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

<18	18-35	>35	
121	288	91	→ observed
500x0.2 = 100	500x0.3 = 150	500x0.5 = 250	→ This is expected.

$\Rightarrow n = 500$ sample

① H_0 = The data meets the distribution 2000 censey

H_1 = The data does not meet the distribution 2000 censey

② $\alpha = 0.005 \Rightarrow CI = 95\%$

③ Degree of freedom $= n-1$

$= 3-1$

$= 2$

④ Decision Boundary → check chisquare table

$\alpha^2_{0.005} = 5.991$

⑤ Calculate test statistics

$$\alpha^2 = \frac{\Sigma (f_o - f_e)^2}{f_e}$$

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

$$= 232.94.$$

Covariance :-

[X] Weight	[Y] Height
50	160
60	170
70	180
75	181

$\Rightarrow$ Here relation ip

[X] ↑ [Y] ↑

No. of hour study	play
2	6
3	4
4	3

[X] ↑ [Y] ↓

$$\text{Covariance} = \boxed{\text{Cov}(x,y) = \sum_{i=1}^{N} \frac{(x_i - \bar{x})(y_i - \bar{y})}{N}} \Rightarrow \text{population}$$

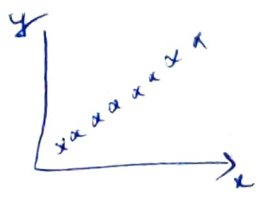
$$\boxed{\text{Cov}(x,y) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1}} \Rightarrow \text{Sample data}$$

$= \{+ve\}$ or $\{-ve\}$ or $\{zero\}$

+ve

X ↑ Y ↑

X ↓ Y ↓

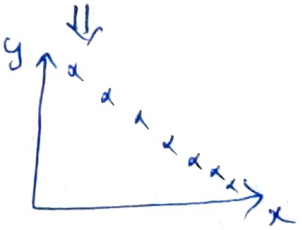


Positive Co-relation

-ve

X ↑ Y ↓

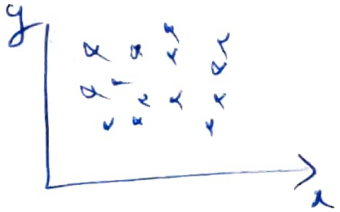
X ↓ Y ↑



Negative Co-relation

0

⇓ No relation



Disadvantage of Co-variance :-

(1) There is no fix range of magnitude.

Pearson Coefficient :-

→ because of disadvantage of Co-variance, we will use pearson Coefficient.

it ranges {-1 to +1}

* The more towards +1 is more positively Co-related
* The more towards -1 is more negatively Co-related.

$$\rho_{(x,y)} = \frac{Cov(x,y)}{\sigma_x \sigma_y} \quad \because \{-1, 1\}$$

(*) Pearson Coefficent not works well with non-linear properties.

Spearman Co-relation Coefficient :-

(*) Here Non-linear properties also works very well.

$$Spea(x,y) = \frac{Cov(R(x), R(y))}{R\sigma_x \; R\sigma_y}$$

X	Y	R(x)	R(Y)
170	75	2	2
160	62	3	3
150	60	4	4
145	55	5	5
180	85	1	1

{$p \leq \alpha \Rightarrow$ Reject Null}

Here p means probability.

Day 7 :-

1. P-value and Significance value
2. Distribution
3. Central limit theorem
4. Bernoulis distribution
5. Bionomial distribution
6. Pacto's distribution
7. Log Normal distribution
8. Poison distribution
9. Power law.

P-value & Significance value :-

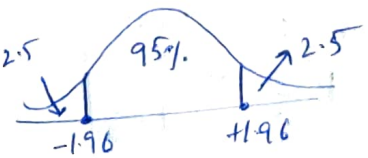
(Q) The avg weight of all residents in bangolore city is 168 pound with standard deviation 3.9. we take sample 36 indiviual and mean ip 169.5 pounds. CI = 95%.

$\mu = 168$, $\sigma = 3.9$ $\bar{x} = 169.5$ $n = 36$ $\alpha = 0.05$

(i) $H_0 = \mu = 168$

(ii) $H_1 = \mu \neq 168$

(iii) Decision boundary



(iv) Calculate Z- stats

$$Z = \frac{x_i - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{169.5 - 168}{\frac{3.9}{\sqrt{36}}}$$

$$= 2.307$$

(v) Decision Rule

$2.307 > 1.96$, so we Reject Null hypothesis.

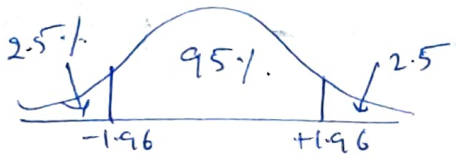
(Q) Average age of a college is 24 years with a std 1.5 we take sample 36 students and mean is 25 and with $\alpha = 0.05$.

$\mu = 24$, $\bar{x} = 25$, $\sigma = 1.5$ $\alpha = 0.05$, $n = 36$

(i) $H_0 = \mu = 24$

(ii) $H_1 = \mu \neq 24$

(iii) Decision boundary



(iv) Calculate z- Statistics

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{25 - 24}{\frac{1.5}{\sqrt{36}}}$$

$$= \frac{1 \times 6}{1.5}$$

$$= 4$$

(v) Decision Rule

$4 > 1.96$ the we Reject the null hypothesis.

Log Normal distribution :—

→ This is log Normal distribution

→ Apply log(y) to become Normal distribution.

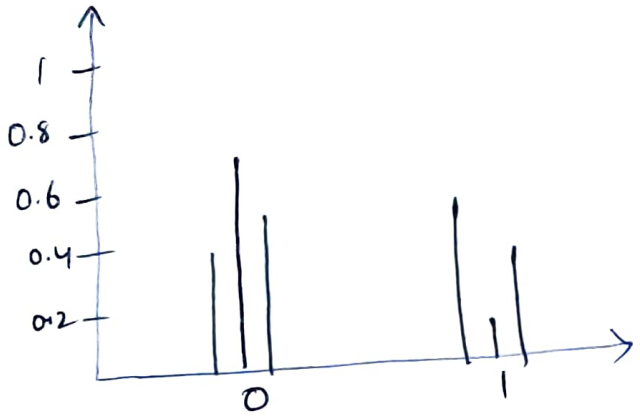
Bernoulis Distribution :—

Here 2 outcomes either {0 or 1}

Assume { Tossing a coin
$P(H) = 0.5$

$p = 0.5$

$q = 1 - p = 0.5$



$\Rightarrow$ $p^k (1-q)^k$

→ probability mass function

⇓

Specifically for categorical variables.

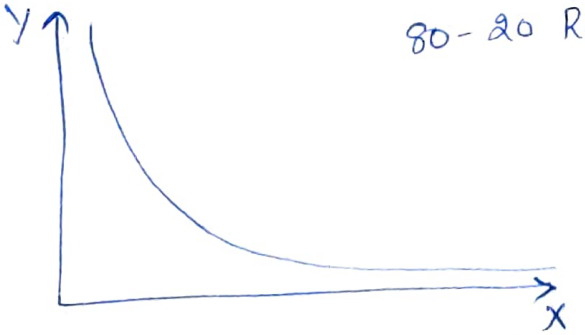
Bionomial Distribution :—

With multiple traits

$$\binom{n}{k} p^k q^{n-k}$$

Pareto distribution :—

It is non-Gaussian distribution



80-20 Rule $\Rightarrow$ Power law distribution.

Eg : 80% of the health is distributed of 20% of people

80% of the Company project by 20% of people in the team.

Central limit theoream

if sample data ≥ 30

$n \geq 30$

* Start picking multiple samples.

$S_1 \longrightarrow \bar{x}_1$

$S_2 \longrightarrow \bar{x}_2$

$S_3 \longrightarrow \bar{x}_3$

$\vdots$

$S_n \longrightarrow \bar{x}_n$

Sample mean