Introduction To Statistics

Total 47

① Basics To Advance

(i) Descriptive stats

{ ① Measure of central Tendency
② Measure of Dispersion }

Summarizing the data.

Histograms, Pdf, Cdf, Probability,
Permutation, mean, Median, Mode,
Variance, Standard deviation

① Gaussian Distribution
② Lognormal Distribution
③ Binomial Distribution
④ Bernoulli's Distribution
⑤ Pareto Distribution (Power law)
⑥ Standard Normal Distribution → Python
⑦ Transformation and Standardization
⑧ Q-Q Plot

(ii) Inferential stats

Z test → Python
t test → Python
ANOVA → F test
CHI SQUARE
Hypothesis Testing
{P values}
Confidence Intervals
Z table, t table

CamScanner

What is Statistics?

Statistics is the science of collecting, organizing and analyzing data. {Better Decision Making}

Def. data??

Data: Facts or pieces of information that can be measured.

Eg: The IQ of a class

{98, 97, 60, 75, 55, 65}

Ages of students of a class

{30, 25, 24, 23, 27, 28} → Data

Types of Statistics

① Descriptive stats: It consists of organizing and summarizing data.

② Inferential stats: Techniques where in we used the data that we have measured to form conclusions.

① classroom of Maths student (20)

84, 86, 78, 72, 75, 65, 80, 81, 92, 95, 96, 97, ...

eg: Descriptive stats.

① what is the average marks of the students in the class.

CamScanner

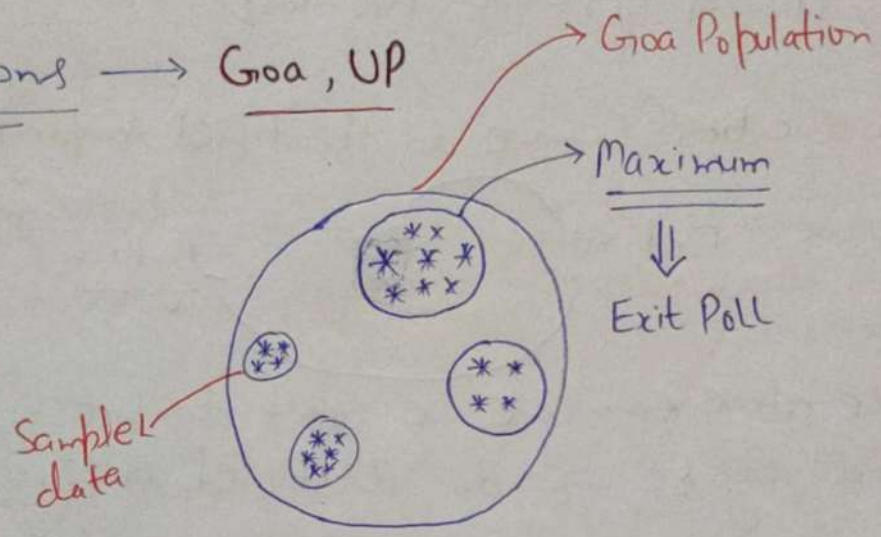
Inferential Stats

Sample $\boxed{1}$ ← 5 → Population

eg. Are the marks of the students of this classroom similar to the age of the Maths classroom in the college?

Population And Sample

Elections ⟶ Goa, UP

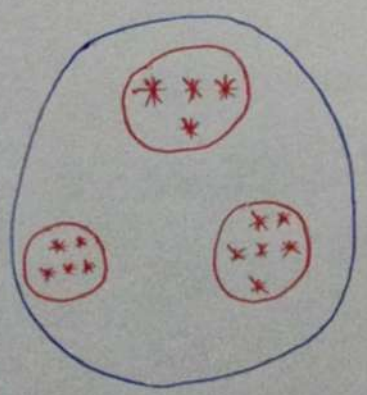


Population (N) Sample (n)

Interview

Sampling Techniques

① Simple Random Sampling: Every member of the population (N) has an equal chance of being selected for your sample (n).



② Stratified Sampling: where the Population (N) is split into non-overlapping groups. (strata)

eg: Gender → Male ✓ → Female ✓ Survey

Age-group

(0-10) (10-20) (20-40) (40-100)

Interview Question: Can we do stratified sampling based on profession?

Profession: Dot net PHP Python

In the above ~~can~~ case we can't do but by applying some condition it may be we do stratified sampling.

Profession: Doctors, Engineers

In that case we do stratified sampling.

③ Systematic Sampling:

(N) ⟶ n th individual

eg: Mall → Survey

→ 8th person → Survey

④ Convenience Sampling:

→ Survey ←

→ Data Science

Only those people

use case

eg: Exit Poll

{Random Sampling}

RBI → Household Survey

⇓

Survey ⇒ Women

eg: Drug → Tested ⇒ Stratified + Convenience
⇓
What kind of Sample?

Variables: A variable is a property that can take any value.

{182, 178, 168, 150, 160, 170}

eg: Height
weight = {78, 99, 100, 60, 50}

Two kinds of variables:

① Quantitative variable → Measured numerically, {add, subtract, multiply, divide}
eg: Age, Weight, height

② Qualitative / Categorical variables

↳ eg: Gender → H, → F { Based on some characterstics we can derive Categorical variables. }

eg: IQ

0–10	10–50	50–100
⇓	⇓	⇓
Less IQ	Medium IQ	Good IQ

Blood group	T-shirt size
A +ve	L
	XL
	M
	S

Quantitative
- Discrete variable
- Continuous Variables

eg: Whole number
No. of Bank Accounts

eg: 2, 3, 4, 5, 6, 7,

Total children in a family
eg: 2, 3, 4, 5

eg: Height = 172.5, 162 cm, 163.5 cm.
weight = 100kg, 99.5, 99.75

Rainfall = 1.1, 1.25, 1.35 ---

eg: ① What kind of variable is Gender ?? Categorical

② What " " " Marital status ? "

③ " " " " River length ? Continuous

④ " " " " Population of the state is ? Discrete

⑤ " " " " Song length ? Continuous

⑥ " " " Blood pressure ? Continuous

⑦ PIN Code ? {Discrete or Categorical}

⊛ Variable Measurement Scales

4 types of measured variable

① Nominal data – {Categorical data} → classes

② Ordinal – Order of the data matters, value does not.

③ Interval – Order matters, value also matters, natural zero is not present. eg (0 k temp.)

④ Ratio data
⇓
the measurement have a true natural zero.

Interview question: Difference b/w ordinal and Nominal data

eg: Student marks – 0, 10, 20, 30, 70

eg:

Student (Marks)	Rank
100	1
96	2
57	4
85	3
44	5

→ Ordinal Data

Frequency Distribution:

Sample dataset :- Rose, Lilly, Sun flower, Rose, Lilly, Sunflower, Rose, Lilly, Lilly.

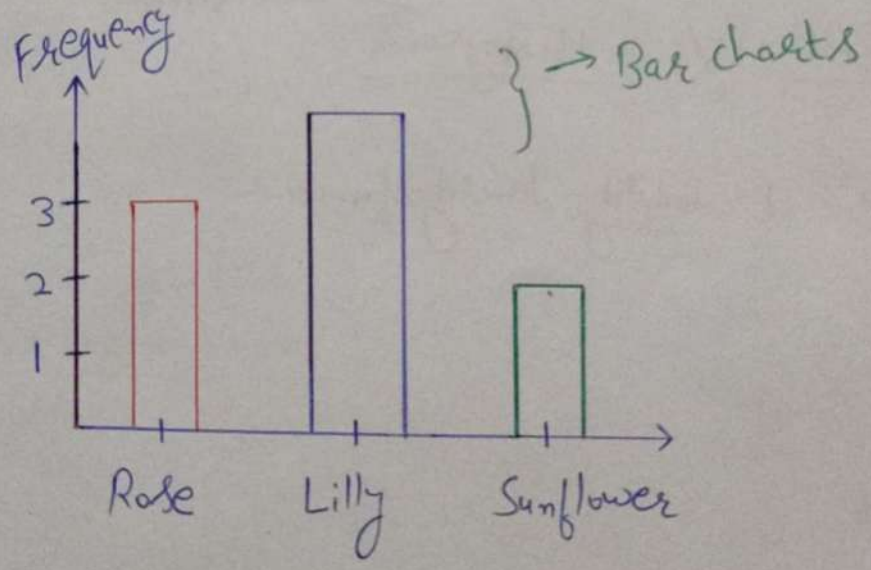
Flower	Frequency	Commulative Frequency
Rose	3	3
Lilly	4	7
Sunflower	2	9

Interview

When there is discrete variables :- we draw a bar graphs

when " " continuous variables :- we draw a histogram.

① Bar Graph



For Summarizing the data we plot a Bar Graph

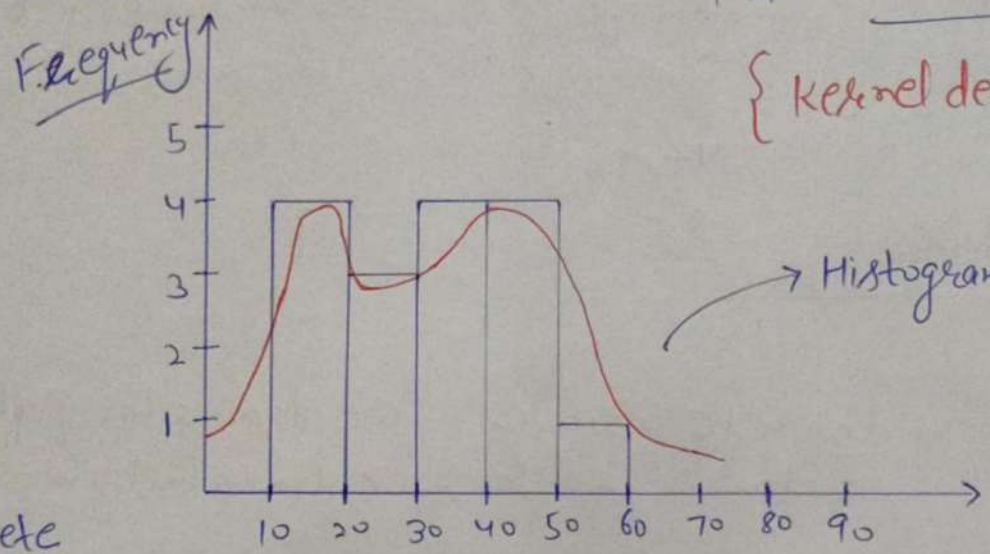
② Histograms :- Continuous

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

↳ Bins = 10

Pdf :- Smoothing of histograms

{ Kernel density function }



→ Histograms

Discrete

→ Continuous

BAR Vs Histograms

Pdf : Probability density function

Basics To Intermediate Stats

1. Measure of Central Tendency
2. Measure of dispersion
3. Gaussian Distribution
4. Z score
5. Standard normal Distribution

1. Arithmetic Mean for population & Sample

Mean (Average)

Population (N)

Population data

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

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

$\Downarrow$

Population mean $= \dfrac{1+1+2+2+3+3+4+5+5+6}{10}$

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

Sample (n)

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

$= \underline{\underline{3.2}}$

Interview

(*) **Central Tendency**: Central tendency refers to the measure used to measure the "centre" of the distribution of data.

1. Mean
2. Median
3. Mode

CamScanner

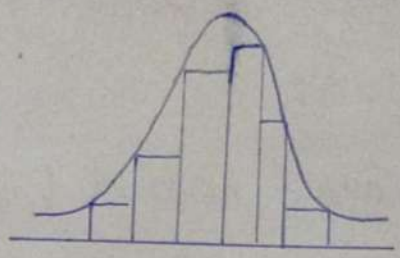
① **Mean**: $\{1, 1, 2, 2, 3, 3, 4, 5, 5, 6, \boxed{100}\}$ ✗ → outliers

$$\text{Mean} = \frac{32 + 100}{11}$$

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

$\mu = 3.2$
$+$
100
$\mu = 12$

Distribution



② **Median**: $\{1, 1, 2, 2, \boxed{3}, 3, 4, 5, 5, 6, \boxed{100}, 112\}$

Steps: ① Sort all the numbers ✓

② Find the central element → odd length / → even length

$\boxed{\mu = 12}$ ✓

Median = 3

Median = 3.5

$\boxed{\text{Avg} = \frac{3+4}{2} = 3.5}$ find the average of two middle elements.

$\boxed{\mu = 3.2}$

$\boxed{\mu = 12}$

$\boxed{\text{Median} = 3}$

$\boxed{\text{Median} = 3.5}$

* {Median works well with outliers}

CamScanner

③ **Mode**: { most frequent elements }

$\{1, \underbrace{2, 2}_{2}, 3, 4, 5, \underbrace{6, 6, 6}_{3}, 7, 8, 100, \underline{100, 100, 100}\}$

$\boxed{\text{mode} = 6}$ → Measure of central tendency

when to use:

① Outliers → median

② Categorical feature → mode

eg: Type of flower | Petal length | Petal width | Data set → 10% missing data

Rose
Lilly
Sunflower
—
—
—

{ Missing Value → Most frequent occuring element }

⇓

Categorical variable

Ages of students

age
25
26
—
—
—
—
32
34
38

⇒ Mean? ✓
Median?
Mode??

(*) Measure of Dispersion

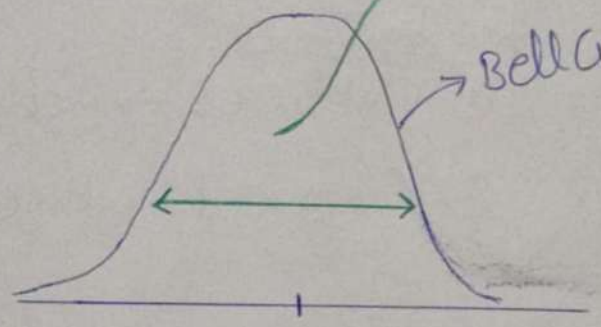
1. Variance
2. Standard Deviation

{Dispersion} $\overset{\mu}{=}$ $\{1,1,2,2,4\}$ $\frac{10}{5}=2$

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

⇓

Spread



Variance, Standard Deviation

Bell Curve

Normal Distribution

Mean, Median, Mode

⇓

Measure of central Tendency

Population (N)	Sample (n)
$\mu = \sum_{i=1}^{N} \frac{(x_i)}{N}$	$\bar{x} = \sum_{i=1}^{n} \frac{(x_i)}{n}$
$\sigma^2 = \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{N}$	$S^2 = \sum_{i=1}^{n} \frac{(x_i-\bar{x})^2}{\boxed{n-1}}$

Bessels Correction

Degree of freedom

① Variance

Population Variance

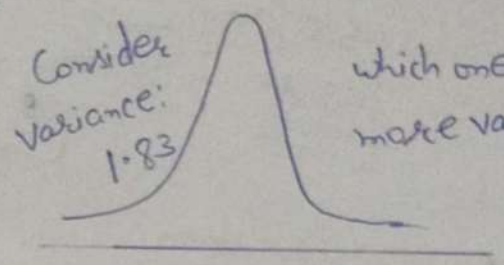
$$\sigma^2 = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N} = \frac{10.84}{6} = 1.81$$

Sample variance

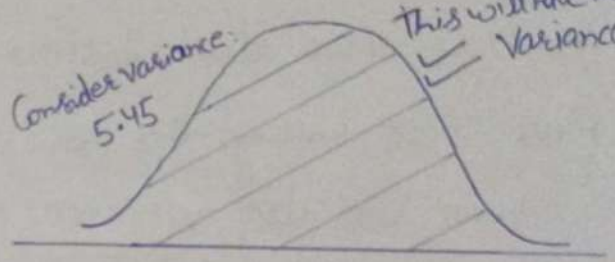
$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$$

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
			10.84

$\mu = 2.83$



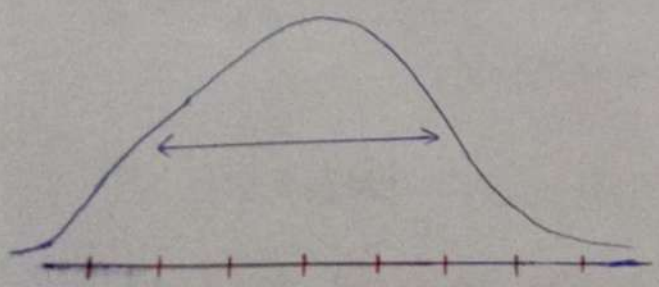
which one have more variance?



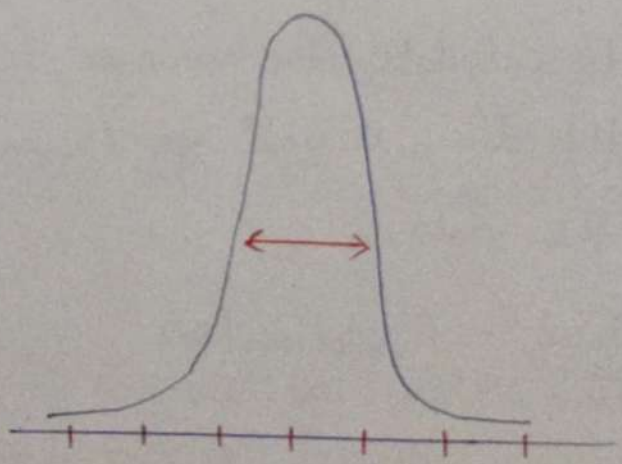
* more variance value: more spread.

less variance value = less spread but more peak

more variance = more spread



graph variance → Big Number
S.D. → Big Number
⇓
means more spread



graph variance → Small number
S.D. → Small number
⇓
means less spread.

② Standard Deviation:

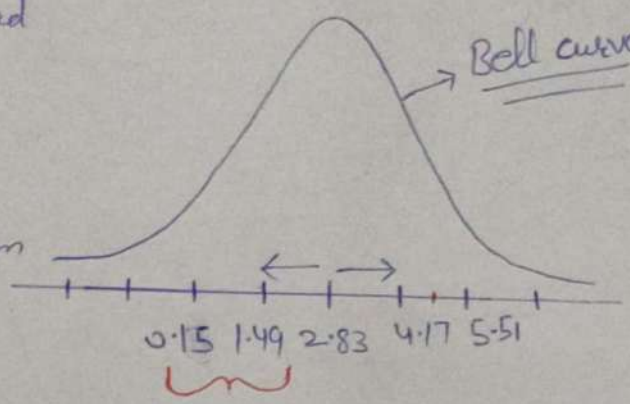
$$S.D = \sigma = \sqrt{\text{variance}} = \sqrt{1.81} = 1.345$$

variance = spread

Let's Consider

S.D. = $\boxed{5}$

↳ 1.5σ from the μ



$$\begin{array}{r} 2.83 \\ +\ 1.34 \\ \hline 4.17 \end{array} \qquad \begin{array}{r} 2.83 \\ -1.34 \\ \hline 1.49 \end{array}$$

$$\begin{array}{r} 4.17 \\ +\ 1.34 \\ \hline 5.51 \end{array} \qquad \begin{array}{r} 1.49 \\ 1.34 \\ \hline 0.15 \end{array}$$

Interview Question

① How two distributions are different?

Ans For knowing the how two distributions are different at that time we use variance and standard deviation.

Most important

② Why sample variance is divided by n−1?

Ans The reason dividing by n−1 corrects the bias is because we are using the sample mean, instead of the population mean, to calculate the variance. Since the sample mean is based on the data, it will get drawn toward the centre of mass for the data.

Proof: Population (N)

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

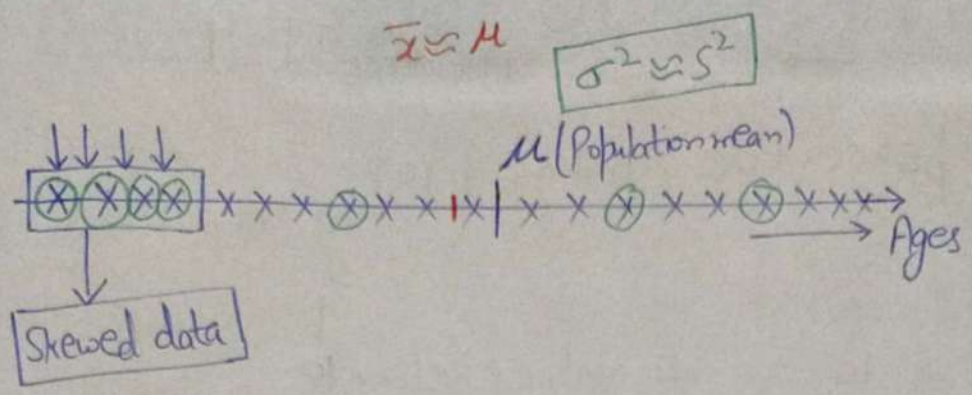
Sample (n)

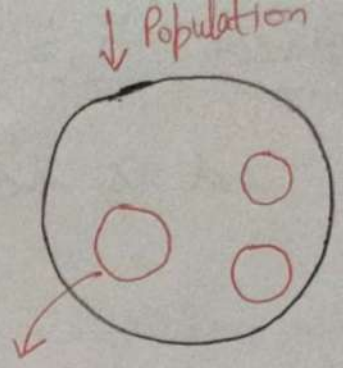
$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$$

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

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

CamScanner





$n-1, n-2, n-3, n-4, \ldots$

(*) Percentiles and Quartiles

Percentage: 1, 2, 3, 4, 5

% of the numbers that are odd?

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

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

Percentiles: { GATE, CAT, SAT }

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

95 percentile means that the person has got better marks than 95% of the entire student.

Dataset : 2, 2, 3, 4, 5, 5, 5, 6, 7, 8, 8, 8, 8, 9, 9, $\boxed{10}$, $\boxed{11}$, 11, 12

$\downarrow 5+5/2 = 5$

① What is the percentile ranking of 10?

$n = 20$

$$\text{Percentile rank of } \overset{x}{10} = \frac{\#\text{ values below } \overset{x}{10}}{n} \times 100$$

$$= \frac{16}{\cancel{20}} \times \overset{5}{\cancel{100}} = \underline{\underline{80 \text{ percentile}}}$$

What is the percentile ranking of 11?

$$\text{Percentile rank of } \overset{x}{11} = \frac{\#\text{ values below } \overset{x}{11}}{n} \times 100$$

$$= \frac{17}{\cancel{20}} \times \overset{5}{\cancel{100}} = \underline{\underline{85 \text{ percentile}}}$$

② What value exists at percentile rank of 25?

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

$$= \frac{\cancel{25}}{\cancel{100}_{4}} \times 21 = 5.25 \Rightarrow \underline{\underline{\text{index position}}}$$

$$\Downarrow$$

$\underline{\underline{5}}$ $\underline{\underline{\text{Answer}}}$

(*) Five Number Summary And Box Plot

1. Minimum
2. First Quartile (25%) Q1
3. Median
4. Third Quartile (75%) Q3
5. Maximum

$\Rightarrow$ Box Plot $\rightarrow$ Outlier detect

Removing the Outliers

$\{1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, \boxed{7}, 8, 8, 9, \boxed{27}\}$ $\rightarrow$ Outlier $\downarrow$ BoxPlot

[Lower Fence $\longleftrightarrow$ Higher Fence]

$$Q1 = \frac{25}{100} \times (\overset{n}{19}+1) = \frac{\cancel{25}^{5}}{\cancel{100}_{\cancel{5}}} \times \cancel{20}^{4} = 5 \Rightarrow \text{index}$$

$\boxed{(25\%)\, Q1 = 3}$

$$\boxed{IQR = Q3 - Q1}$$

$$\boxed{\begin{aligned} \text{Lower Fence} &= Q1 - 1.5(IQR) \\ \text{Higher Fence} &= Q3 + 1.5(IQR) \end{aligned}}$$

$$Q3 = \frac{75}{100}(19+1)$$

$$= \frac{\cancel{75}^{15}}{\cancel{100}_{\cancel{5}}} \times \cancel{20}^{4}$$

$$= 15 \Rightarrow \text{index}$$

$\boxed{(75\%)\, Q3 = 7}$

$$IQR = Q3 - Q1$$

$$= 7 - 3$$

$$\boxed{IQR = 4}$$

$$\boxed{\begin{aligned} L.F &= 3 - 1.5(4) = 3 - 6 = -3 \\ H.F &= 7 + 1.5(4) = 7 + 6 = 13 \end{aligned}}$$

$[-3 \leftrightarrow 13]$

CamScanner

Box Plot

① Minimum = 1

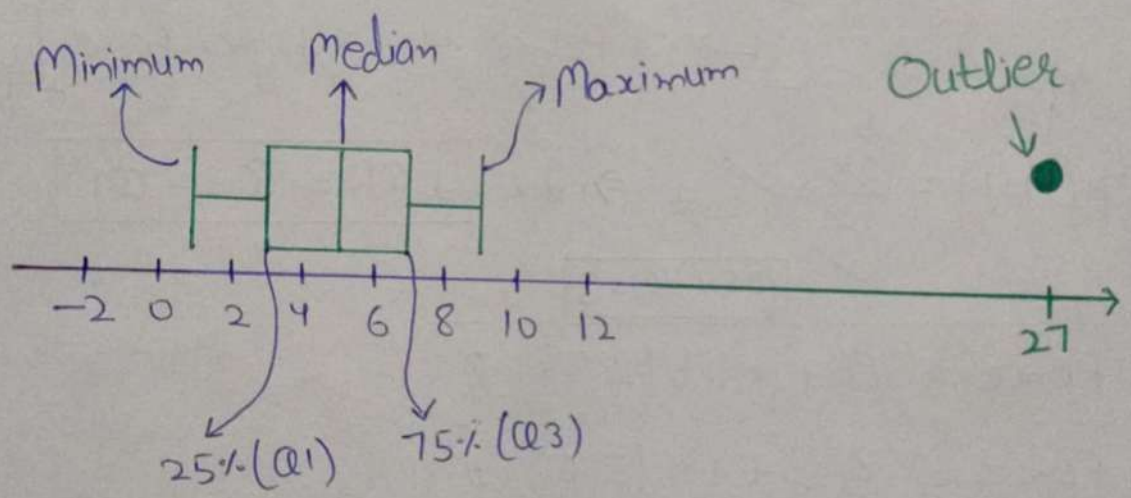
② First Quartile (25%) Q1 = 3

③ Median = 5

④ Third Quartile = 7

⑤ Maximum = 9

outlier
(27)



Interview Question: What is Box Plot and applications?

Ans: A box-plot displays the five-number summary of a set of data. The five-number summary is the minimum, first quartile, median, third quartile and maximum.
We draw a box from the first quartile to the third quartile. A vertical line goes through the box at the median

Applications: ① It is use to remove the outliers.

② It provides us visualization way where the outlier is present.

CamScanner

Advance Statistics

① Distributions (Gaussian Distribution)

- ↳ Normal Distribution
- ↳ Standard Normal Distribution
- ↳ Z score
- ↳ Log Normal Distribution
- ↳ Bernoulli's Distribution
- ↳ Binomial Distribution

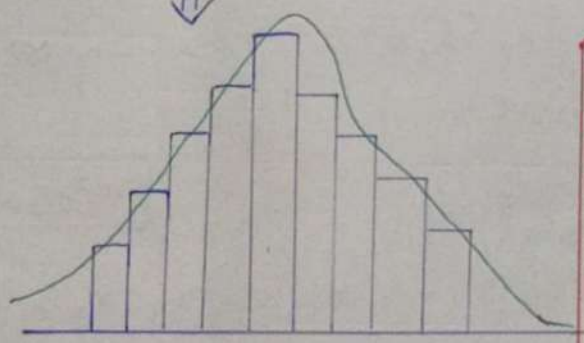
Practical

1. Mean, Median, Mode
2. Variance, Standard Deviation
3. Histogram, Pdf, Bar Plot, violan Plot
4. IQR
5. Log Normal Distribution

① Distribution

Ages = {24, 26, 27, 25, 30, 32 - - - - - - - - -}

⇓

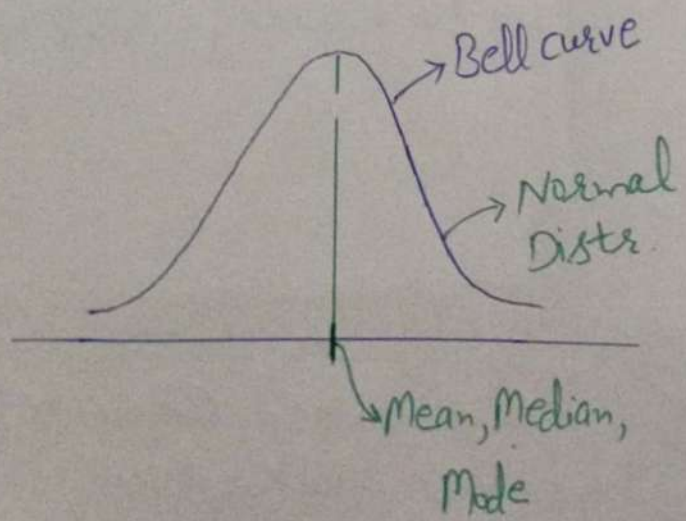


Interview

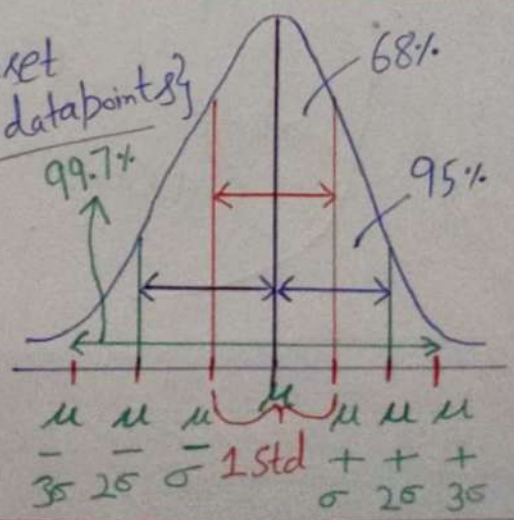
[3-Sigma Rule]
or
[Emperical Formula]
68-95-99.7% Rule

Interview

Normal / Gaussian Distribution



Dataset {100 datapoints}

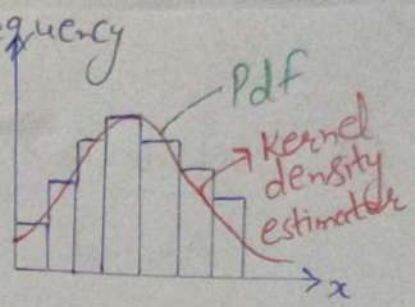


Symmetrical: Left area of the curve = Right area of the curve

eg: ① Height → Normally Distributed

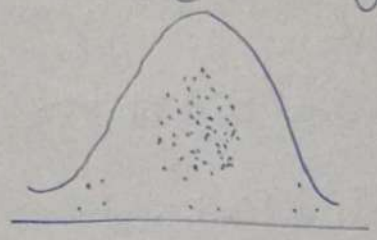
↓

Domain Expert → {Doctor}



② Weight ③ IRIS DATASET ⇒ {sepal length, petal length}

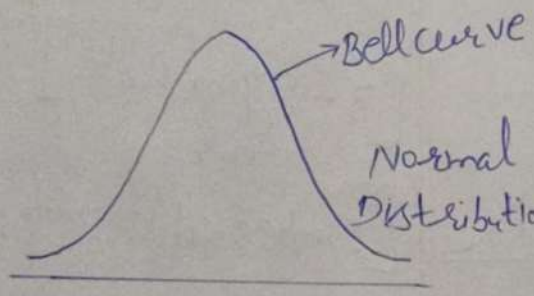
④ Laptop Touchpad →



⑤ ineuron Team meetup ⇒ T-shirt ⇒ L, M, XL, XXL

Interview Question: Central limit Theorem

Prob: $x = \{65, 72, 83, 55, 64, 67, \ldots\} \approx N(\mu, \sigma)$



$\boxed{n \geq 30}$ ⇒ Sample size

→ $\{x_1, x_2, x_3, x_4 \ldots\} \rightarrow \bar{x}_1$

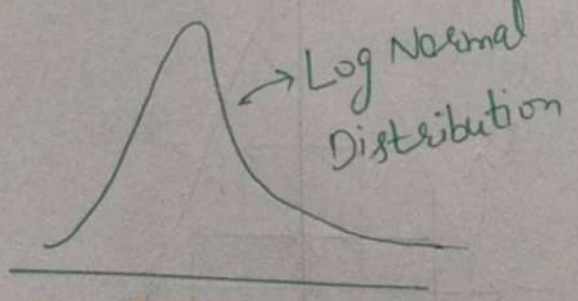
→ $\{x_1, x_3, x_7, x_5 \ldots\} \rightarrow \bar{x}_2$

→ $\{x_1, x_3, x_4 \ldots\} \rightarrow \bar{x}_3$

⋮

→ $\bar{x}_n$

$y = \{\ldots\}$ weather



* Interview

log normal Distribution: The curve which is skewed either left or right hand side is called log normal distribution.

In above example, it is skewed towards the right hand side

Interview Can we convert log Normal Distribution into a Normal Distribution - Yes (11)

if $x \simeq$ Log Normal Distribution

then

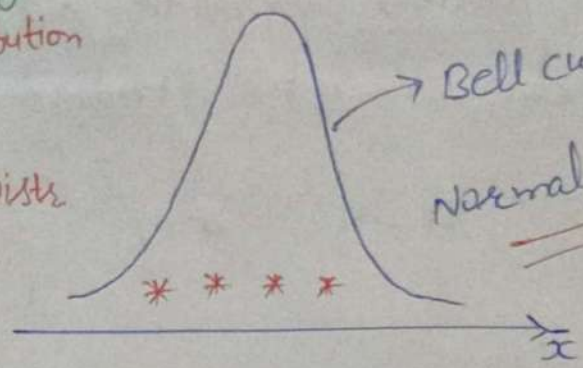
$y \simeq \log(x) \Rightarrow$ Normal Distr

Natural log $\boxed{\log_e}$

$\Downarrow$

$x \simeq \exp(y)$

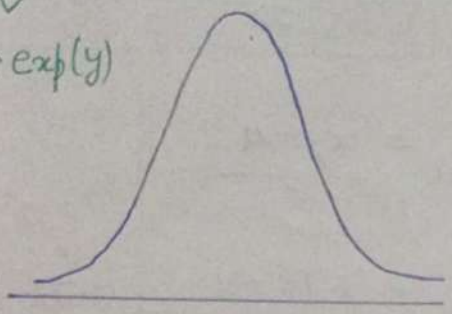
by applying natural log - $y \simeq \log(x)$



eg: of log Normal Distribution:

① Wealth Distribution

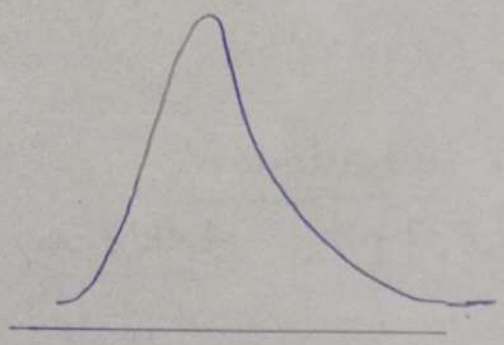
② People writing big Comments



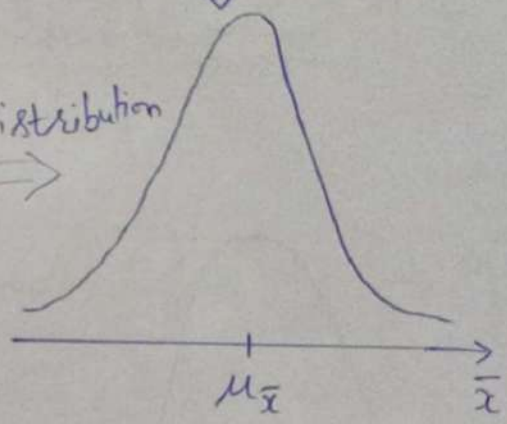
$n \geq 30$

Sample Distribution

$\Downarrow$



Normal Distribution $\Longrightarrow$



Interview

Definition of Central Limit Theorem:

It says whether your distribution is Gaussian distribution or non-Gaussian distribution, if we take the sample where $n \geq 30$ and we take all the sample mean and we populated we get a normal distribution.

CamScanner

eg: Example of Normal Distribution and using the Z-score convert the dataset into Standard Normal Distribution.

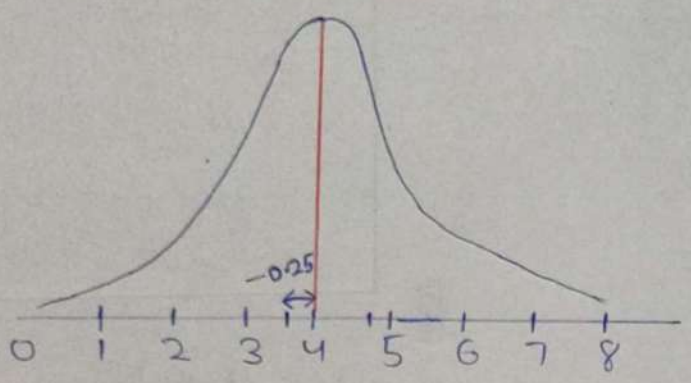
Interview How many standard deviation 4.5 fall from the mean?

Let Consider $\mu = 4$ $\sigma = 1$

$\boxed{4.5} \rightarrow$ standard deviation

$+0.5$ s.d

$\boxed{4.75}$??

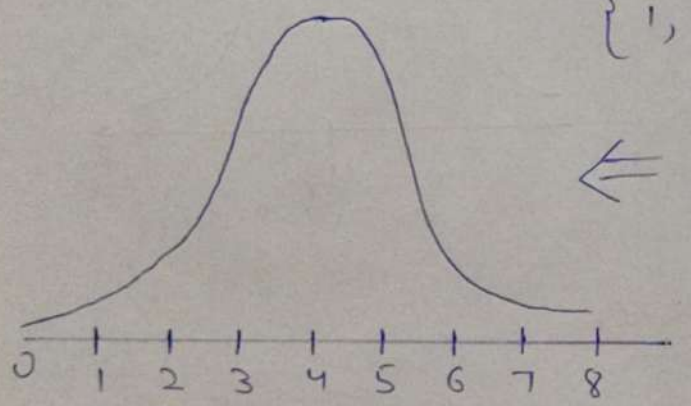


$$Z_{score} = \frac{x_i - \mu}{\sigma}$$

$$= \frac{4.75 - 4}{1} = 0.75 \text{ sd}$$ +ve

3.75

$$Z_{score} = \frac{3.75 - 4}{1} = -0.25$$



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

Let's say $\mu = 4$, $\sigma = 1$

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

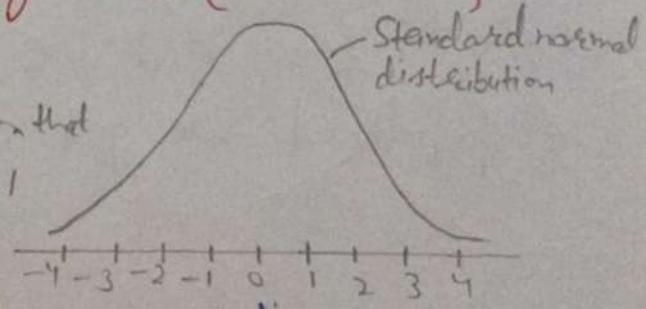
$\{-3, -2, -1, 0, 1, 2, 3\}$

$y \sim SND(\mu = 0, \sigma = 1)$

$Z(1) = \frac{1-4}{1} = -3$ $Z(3) = \frac{3-4}{1} = -1$

$Z(2) = \frac{2-4}{1} = -2$

Any distribution that have $\mu = 0$ and $\sigma = 1$ that distribution is called standard Normal distribution.



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

$\Downarrow$

Z score

$\Downarrow$

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

Standard Normal Distribution $(\mu = 0, \sigma = 1)$

Satisfying this property

Practical Application

DATASET

units

years (Rs) (kg)

Age	Salary	Weight
24	40k	70
25	80k	80
26	60k	55
27	70k	45

Z score

$\mu = 0, \sigma = 1$

$\Downarrow$

Standard Normal Distribution

Standadization ← Z score

⇒ Model ⇒ Mathematical equation

Feature Scaling

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

Normalization

(−1 to 1)

(0 to 1)

By using Min Max Scaler formula we are able to convert this into (0 to 1).

Feature Scaling → Model → O/P Normal Data

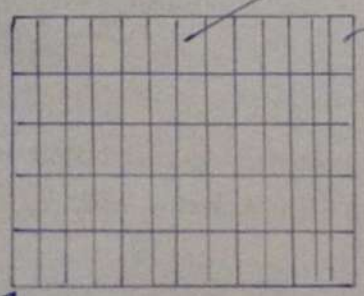
Same unit

Rescale unit

Where we use Normalization?

In Deep Learning,

CNN → Image Classification



Divide by 255

0-255

↓↓

Minmax Scaler

0-1

Pixels

Normalization

Interview Question: Difference between Normalization and Standardization?

(0,1) ⇒ used in Deep Learning

Ans Normalization (min-max Normalization): In this approach we will scale down the values of the features between 0 to 1 by using min-max scaler formula.

Feature Transformation

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Feature Scaling ⇓ used in Feature Engineering

Standardization (Z-Score Normalization): Here all the features will be transformed in such a way that it will have the properties of standard normal distribution with mean (μ) = 0 and Standard deviation (σ) = 1 using the Z-score formula.

Feature Transformation technique

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

---

* **Practical Example**: {India Vs SA}

① ODI Series (2021)

Series Average Score in 2021 = 250

Standard Deviation = 10

Series < Team final Score = 240

In 2020

Series Average Score in 2020 = 260

Standard Deviation = 12

Team final Score = 245

Compare both the scores in which year Team final Score was better?

Ans In 2021,

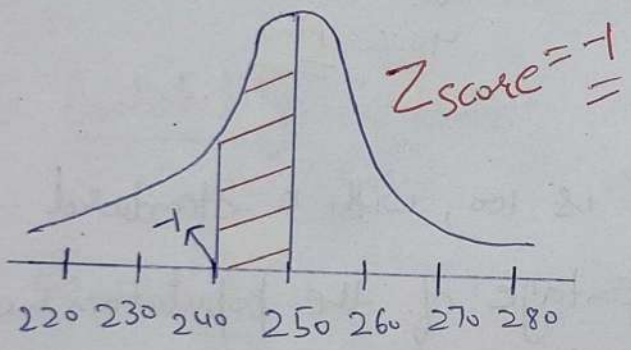
$$Z_{score} = \frac{x_i - \mu}{\sigma} = \frac{240-250}{10} = \frac{-\cancel{10}}{\cancel{10}} = -1$$

In 2020,

$$Z_{score} = \frac{x_i - \mu}{\sigma} = \frac{245-260}{12} = \frac{-15}{12} = -1.25$$

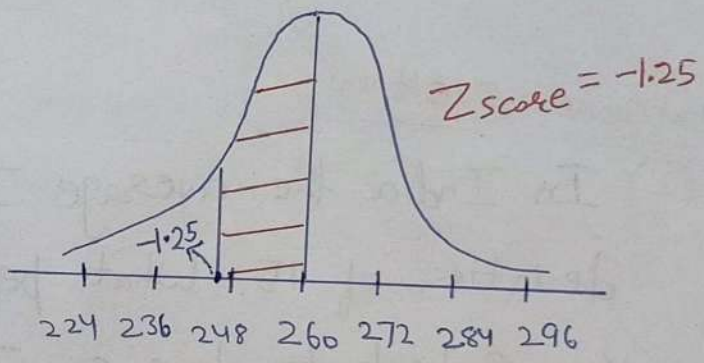
In 2021 ✓

$\boxed{\mu = 250 \quad x_i = 240 \quad \sigma = 10}$



In 2020 ✓

$\boxed{\mu = 260 \quad x_i = 245 \quad \sigma = 12}$

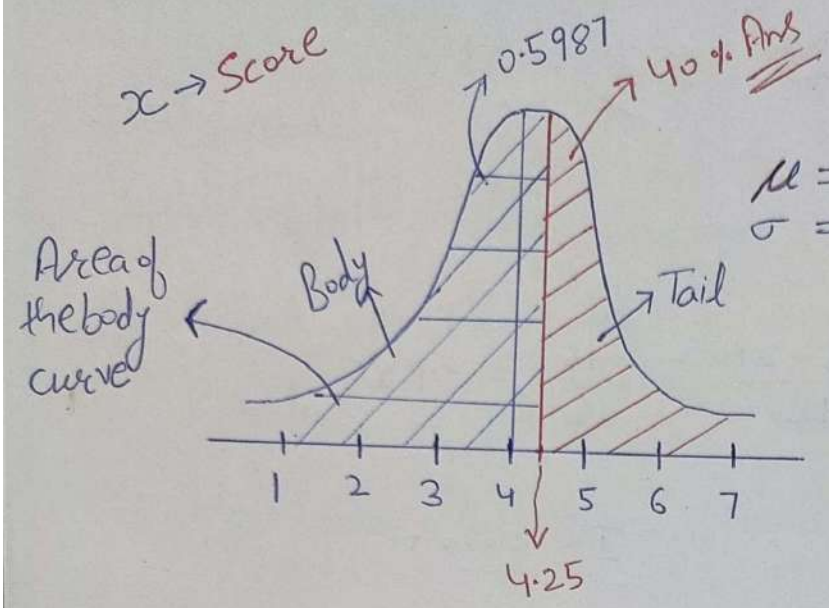


Standard deviation is greater in 2020 as compare to in 2021, So in 2020 teams final score is better than the 2021.

Important

Interview Question: ① What percentage of scores falls above 4.25?

$x \rightarrow$ Score



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

$\boxed{1 - \text{Left Area}}$

use Z-table

$1 - 0.5987$

$= \underline{0.4013}$

$\Downarrow$

40% Ans

$$Z = \frac{x_i - \mu}{\sigma} = \frac{4.25 - 4}{1} = 0.25$$

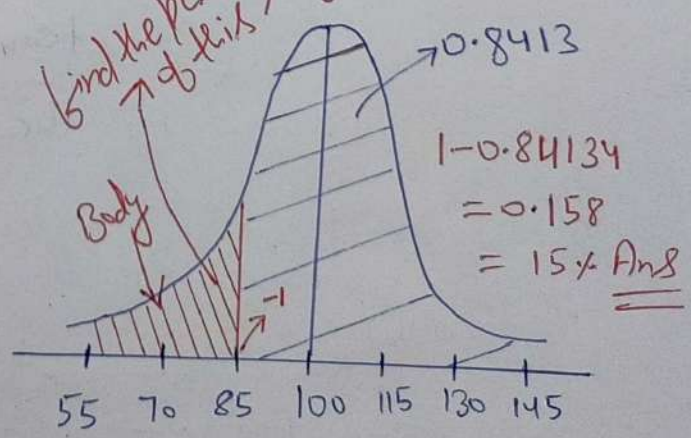
Interview Question

② In India the average IQ is 100, with a standard deviation of 15. What percentage of the population would you expect to have an IQ lower than 85?

Ans

$$Z = \frac{85 - 100}{15} = \frac{-15}{15} = -1$$

use Z-table



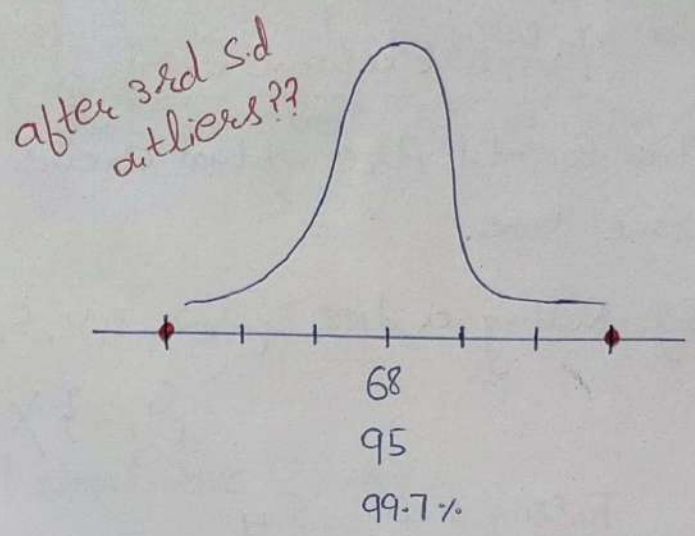
$1 - 0.84134$
$= 0.158$
$= 15\%$ Ans

Also come in interview

IQ $\Downarrow$ 90 to 120

— STATS Today's Agenda

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



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

(*) **Probability**: Probability is a measure of the likelihood of an Event.

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

$$Pr(6) = \frac{\#\text{ of way an even can occur}}{\#\text{ of possible outcome}}$$

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

Toss a Coin $\{H, T\}$

$$\boxed{Pr(H) = \frac{1}{2}}$$

CamScanner

② Addition Rule: (Probability, "or")

Mutual Exclusive Event

Two Events Are mutual exclusive if they cannot occur at the same time.

eg: Rolling a dice $\{1, 2, 3, 4, 5, 6\}$

$\{1, 2\}$ ✗

Tossing a coin $\{H$
$T\}$

Non Mutual Exclusive

Multiple events can occur at the same.

eg: Deck of Cards $\{Q, ♡\}$

Example for Mutual Exclusive

① If I Toss a Coin, what is the probability of the coin landing on heads or tails?

Ans Mutual Exclusive Addition Rule

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

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

$$\boxed{Pr(A \text{ or } B) = 1}$$

⇒ Roll a Dice

$$Pr(1 \text{ or } 3 \text{ or } 6) = Pr(1) + Pr(3) + Pr(6)$$

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

$$= \frac{\cancel{3}}{\cancel{6}_2} = \frac{\cancel{1}}{\cancel{2}} = 0.5$$

CamScanner

Example for Non Mutual Exclusive

You are Picking a card randomly from a deck. What is the Probability of choosing a card that is queen or a heart? $\rightarrow (52)$

Ans Non Mutual Exclusive

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

Addition Rule for non mutual exclusive Events

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

$$P(Q \text{ or } \heartsuit) = P(Q) + P(\heartsuit) - P(Q \text{ and } \heartsuit)$$

$$= \frac{4}{52} + \frac{13}{52} - \frac{1}{52}$$

$$= \frac{\cancel{16}\,4}{\cancel{52}\,13} = \frac{4}{13}$$

③ Multiplication Rule

{Independent Events}

eg: Rolling a Dice $\{1, 2, 3, 4, 5, 6\}$

1, 1, 2, Each and Every are independent

Dependent Events:

Red marble •

$Pr(Red) = \frac{3}{5} \rightarrow Pr(Green) = \frac{2}{4}$

$\Downarrow$

Naive Baye's {Conditional Probability}

Also called (Baye's Theorem)

Independent Events

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

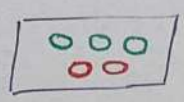
Ans Independent Event

Multiplication Rule

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

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

(*) What is the probability of drawing a Queen and then a Aces from a deck of cards?



3/5 $\boxed{\frac{2}{4}}$

Ans Dependent

$P(G \text{ and } R) = P(G) * P(R/G)$

↳ event occured

$P(A \text{ and } B) = P(A) + P(B/A)$ Conditional Probability

This is very important in Baye's Theorem

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

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

⊛ Permutation and Combination

Permutation

School trip {choclate factory} → Dairy, 5 star, Milky bar, Eclairs, Gem, Silk

Student {Assignment}

Student →

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

Dairy, Gem, Milky

Milky, Gem, Dairy

Permutation

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

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

$$= \underline{\underline{120}}$$

⊛ Combination

Dairy	Gem	Eclair
—	—	—

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

$$= \frac{^{2}\cancel{6} \times 5 \times \cancel{4}^{2} \times \cancel{3!}}{\cancel{3} \times \cancel{2} \times 1 \times \cancel{3!}} = \underline{\underline{20}}$$

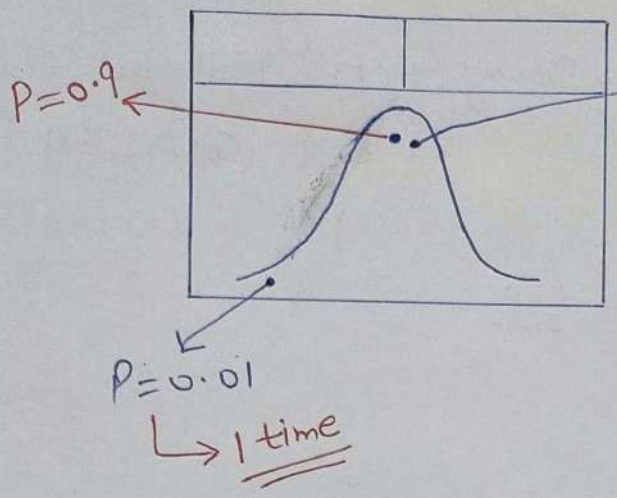
CamScanner

Interview

① **P value** {Many people get's confused but its very easy}

(Significance value)

Every 100 time I touch the mouse pad 80 times I touch this specific region.



Significance value derived from C.I.

NOTE: Significance value $\neq$ P value

P value is also called significance value.

* P value means you are just saying probability.
* Significance says what should be within your C.I.

Definition of P value (Significance value): It is the probability for the Null Hypothesis to be true.

Interview

② **Confidence Interval**: The range of values that we observe in our sample and for which we expect to find the value that accurately reflects the population.

For finding the confidence interval its neccessary to understand the concept of point Estimate.

Point Estimate: The value of any statistic that estimates the value of a parameter is called Point Estimate.

Point Estimate

$\Uparrow$

$\boxed{\bar{x}} \longrightarrow \mu$

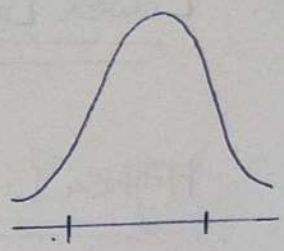
$$\boxed{\begin{array}{l} \bar{x} \geq \mu \\ \bar{x} \leq \mu \end{array}}$$

Statistic $\quad$ Parameter

$\boxed{\bar{x}} \longrightarrow \boxed{\mu} \longleftarrow$ Population mean

Sample mean

$\boxed{\text{Point Estimate}} \pm \boxed{\text{Margin of Error}} = \boxed{\text{Parameter} \Rightarrow \text{Population mean}}$



Lower C.I = Point Estimate − Margin of Error

Higher C.I = Point Estimate + Margin of Error

α = Significance value

$$\text{Margin of Error} = Z_{\alpha/2} \boxed{\frac{\sigma}{\sqrt{n}}} \Rightarrow \text{Standard Error}$$

$\sigma \rightarrow$ Population Sd

Interview Question: Average size of sharks in the sea with C.I. 95%

For Solving this question you allowed to assume anything.

Ans There are various ways of solving this question:

① By using Central limit Theorem

② By using Confidence interval with Population S.D and without Population S.D

Now I am going to solve this question with the help of Confidence interval where Population S.D. is given.

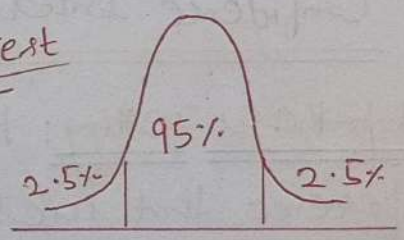
Let consider, Population Standard Deviation is:

$\sigma = 100$

Sample $(n) = 30$

Sample mean $(\bar{x}) = 500$

2 Tail Test



C.I = Point Estimate ± Margin Error

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

This is the formula when your population S.D. is given

$$= 500 \pm Z_{0.05/2} \frac{100}{\sqrt{n}}$$

$Z_{0.025} = 1 - 0.025$
$= 0.976$
$\rightarrow 1.96$

$$\text{Lower Limit} = 500 - 1.96 \times \frac{100}{\sqrt{30}} = 386$$

$$\text{Higher Limit} = 500 + 1.96 \times \frac{100}{\sqrt{30}} = 613$$

Conclusion, If we have 95% C.I., my population mean will be falling between 386 to 613.

* Most Important

Note : ① Z-test is usually applied when the Population Standard deviation is given. or $n \geq 30$.

② When the Population Standard deviation is not given then, we apply t-test. and $n < 30$ and Sample S.D. is given.

Interview

③ Hypothesis Testing And Relationship between Hypothesis testing, Confidence Interval, Significance Value, - - - - -

Hypothesis Testing: Hypothesis testing is a form of statistical inference that uses data from a sample to draw conclusions about a population parameter or a population probability distribution.

$$\boxed{\text{Sample data}} \xrightarrow{\text{Conclusion}} \boxed{\text{Population Data}}$$

$\Uparrow$
Hypothesis

⇒ Inside Hypothesis Testing there are different different experiments (Testings) are available like: Z test, t test, Anova test, Chi square test etc.

Relationship between Hypothesis Testing, C.I. and Significance value:

Hypothesis Testing :-

Coin → Test whether this Coin is a fair Coin or not by performing 100 tosses.

Sholay Coin $P(H) = 100\%$

$$\boxed{P(H) = 0.5 \qquad P(T) = 0.5}$$

100

50 times Head (The coin is fair)

Hypothesis Testing steps:

① Null Hypothesis (H_0) = Coin is fair.

② Alternate Hypothesis (H_1) = Coin is not fair or Coin is unfair.

③ Experiment :- Toss a Coin for 100 times.

④ Reject or Accept the Null Hypothesis.

CamScanner

Let's Consider,

you are getting 50 times head,

Assume, mean is, $\mu = 50$

S.D. $\sigma = 10$

Assume,

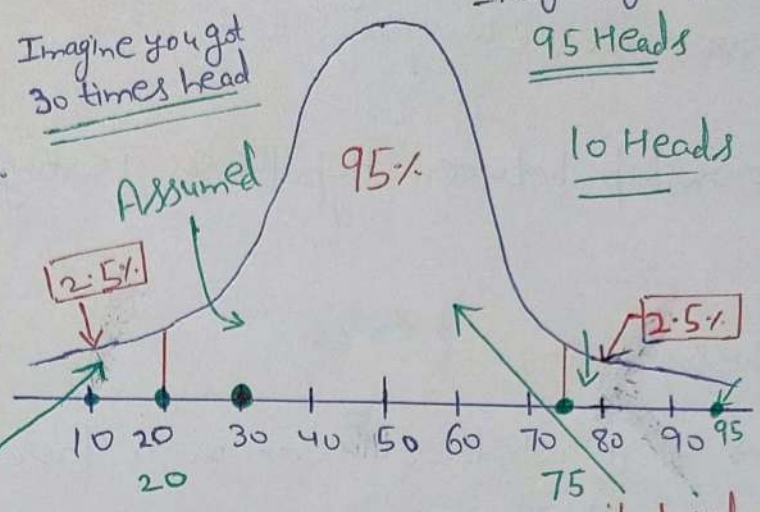
C.I. between 20 to 75

100% − 5%

⇓

95%

C.I.



if Pvalue < 0.05 its falling Here

if p value > 0.05 its falling Here

Significance value

$\alpha = 0.05$ { given by Domain Expert }

NOTE:

$\boxed{\text{Significance value} \neq \text{P value}}$

* P value means you are just saying probability

* Significance says what should be within your C.I.

Today Topics

1. Type 1 and Type 2 Error
2. One(1) Tailed and Two(2) Tailed Test
3. Confidence Interval
4. Z-test, t-test, Chi-square test, ANOVA Test (F test)

1. Type 1 and Type 2 Error

Null Hypothesis (H_0) = Coin is fair
Alternate Hypothesis (H_1) = Coin is not fair

Reality check

Null Hypothesis is True or Null Hypothesis is False

Decision

Null Hypothesis is True or Null Hypothesis is False.

Outcome 1:

We reject the Null Hypothesis, when in reality it is false
↓
Yes

Outcome 2:

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

eg: Movies

Person – Death Sentence

Outcome 3:

We Accept the Null Hypothesis when in reality it is false. → Type 2 Error

Outcome 4:

We Accept the Null Hypothesis when in reality it is true. → Good

	P	N
T	TP	TN
F	FP	FN

FN → Type 2

FP → Type 1

② 1 Tail and 2 tail Test

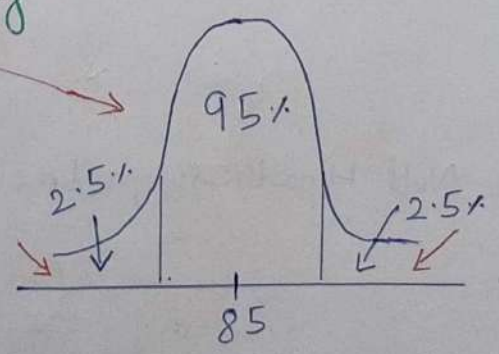
Example: Colleges in Karnataka have an 85% placement rate. A new college was recently opened and it was found that a sample of 150 students had a placement rate of 88%. With a standard deviation 4%. Does this college has a different placement rate?

↳ 85%

$\alpha = 0.05$ (Assume)

85 is our mean.

2 tailed Test → because here two possibilities occur Placement rate may be greater than 85% or may be less than 85%.

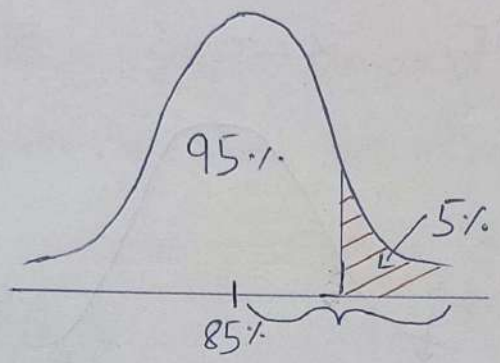


Just change the question – Does this college have a placement rate greater than 85% ?

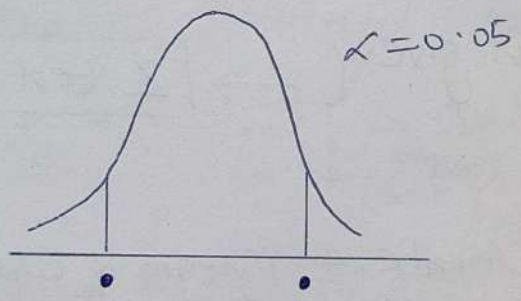
$\alpha = 0.05$ (Assume)

One tailed test
because it have only one condition or say possibility which is greater.

1 Tail Test



③ Confidence Interval



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

Point Estimate : The value of any statistic that estimates the value of a parameter. or Single numerical value used to estimate the unknown population parameter.

Inferential stats

Sample → Population

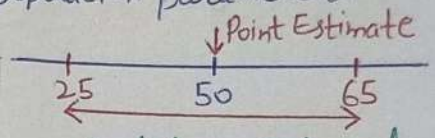
mean → mean

$\boxed{\bar{x}} \longrightarrow \mu$

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

Confidence Intervals

Interval Estimate: Range of values used to estimate the unknown population parameter.



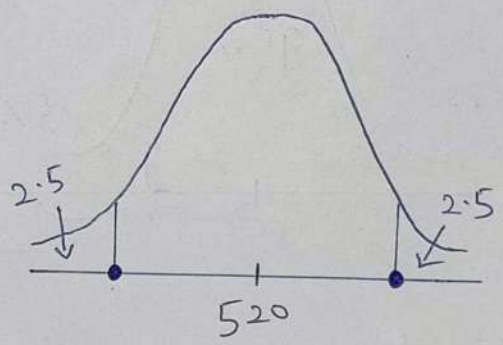
Confidence Interval

Point Estimate ± Margin of Error

Q On the quant test of CAT Exam, the standard deviation is known to be 100. A sample of 25 test takers has a mean of 520 score. Construct a 95% C.I about the mean?

Ans $\sigma = 100$ $\quad n = 25$ $\quad \alpha = 0.05$ $\quad \bar{x} = 520$

$\alpha = 1 - 0.95 = 0.05$



① Population Std is given
② $n \geq 30$

→ $\boxed{\text{Z test}}$

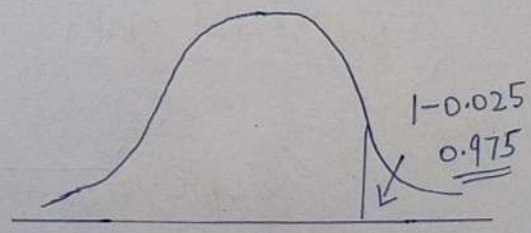
usually sample is always greater or equal 30, here just for example sample is taken 25.

Point Estimate ± Margin of Error

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

→ Standard Error

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



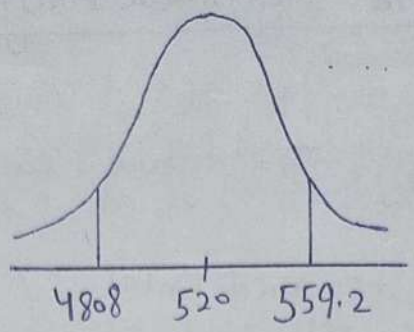
$\boxed{1.96}$

Upper bound $= \bar{x} + Z_{\frac{0.05}{2}} \frac{100}{\sqrt{25}}$

Lower bound $= \bar{x} - Z_{\frac{0.05}{2}} \frac{100}{\sqrt{25}}$

Upper $= 520 + 1.96(20) = 559.2$
Lower $= 520 - 1.96(20) = 480.8$

Ans



Q On the quant test of CAT exam, a sample of 25 test takers has a mean of 520 with a standard deviation of 80. Construct 95% Confidence interval about the mean?

Ans Condition

$n = 25$ $\bar{x} = 520$ $S = 80$

$\alpha = 0.05$

Here, Population S.D. is not given $\longrightarrow$ t - test

Point Estimate $\pm$ Margin of Error

$$\bar{x} \pm t_{\alpha/2}\left(\frac{S}{\sqrt{n}}\right) \longrightarrow \text{Standard Error}$$

$$\text{upper bound} = \bar{x} + t_{0.05/2}\left(\frac{S}{\sqrt{n}}\right) \qquad t_{\frac{0.05}{2}} = 2.064$$

Degree of freedom $= n-1 = 25-1 = 24$

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

$$= 553.024$$

$$\text{Lower bound} = \bar{x} - t_{0.05/2}\left(\frac{S}{\sqrt{n}}\right)$$

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

$$= 486.97$$

$$[486.97 \longleftrightarrow 553.024]$$

CamScanner

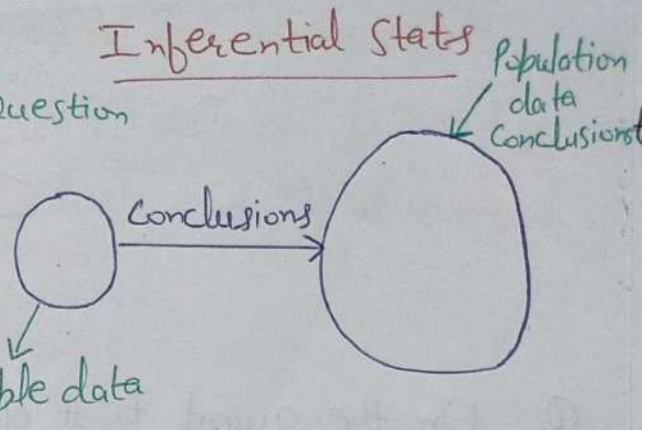
Hypothesis Testing And Statistical Analysis

① Z test
② t test } ⇒ Average ⇒ Interview Question

③ CHI SQUARE ⇒ Categorical Data

④ ANOVA (F test) ⇒ Variance

Inferential Stats

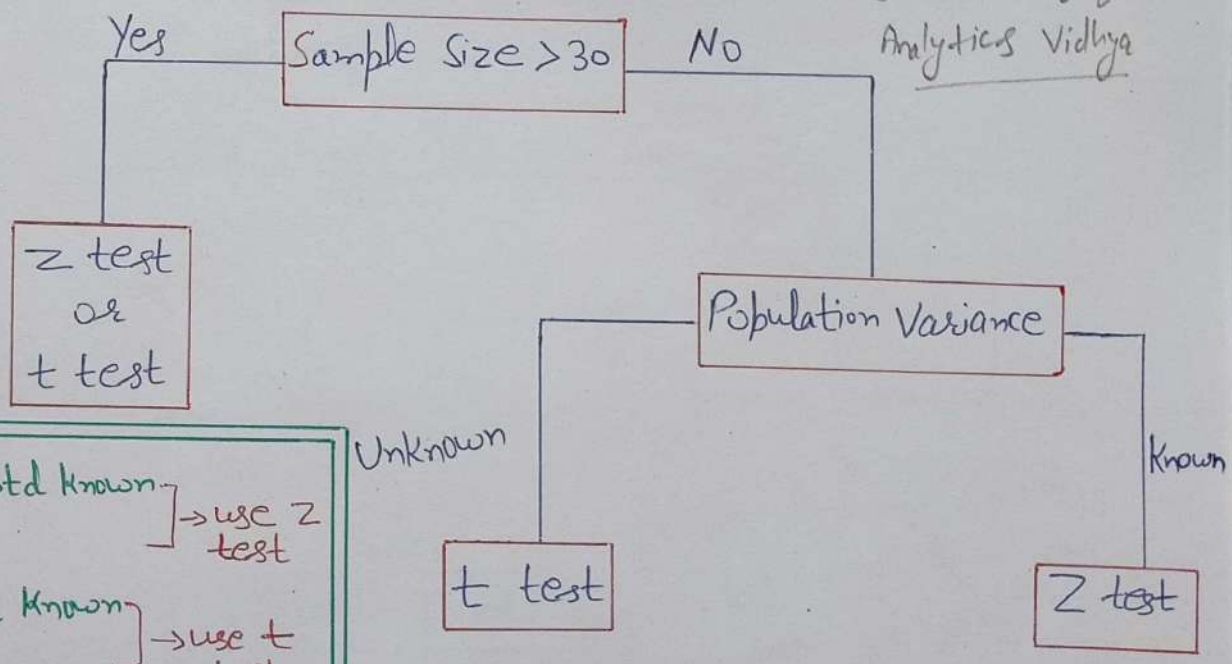


Z test ⇒ Z table

Interview Question

When to use T-tests Vs Z-test

Search on google Analytics Vidhya



**

if Population Std known and n = 30 → use Z test

if Sample std known and n = 30 → use t test

① Population Std known

OR

② Population Std unknown and Sample size > 30

⇒ Z-test

① Population Std unknown
② Sample size ≤ 30
③ Sample Std is given

⇒ t-test

t-test works well with small sample size ($n \leq 30$).

Also works if n = 31, 32, 33 ...
∵ 31, 32, 33 just a minor change

Z-test

① The average heights of all residents in a city is 168 c.m. with a population std $\sigma = 3.9$. A doctor believe the mean to be different. He measured the height of 36 individuals and found the average to 169.5 c.m.

(a) State Null And Alternate Hypothesis

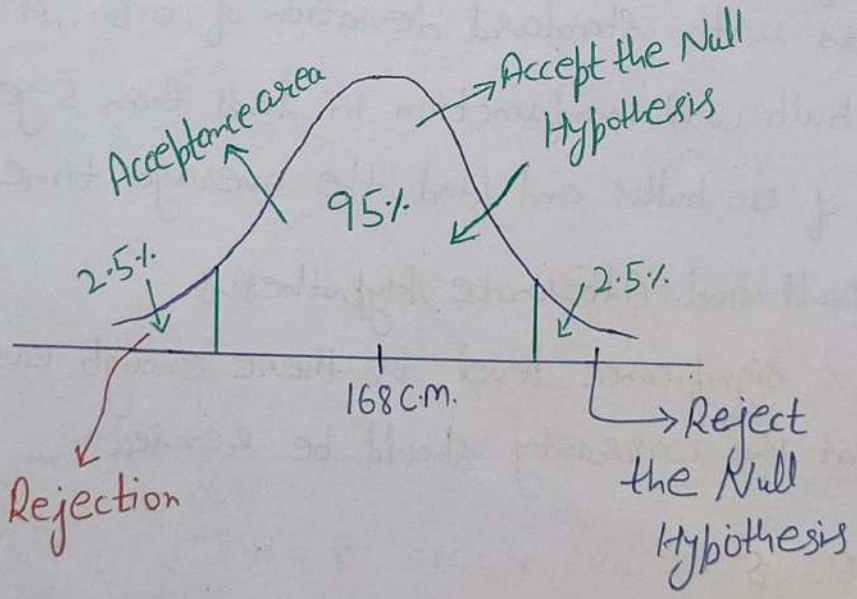
(b) At 95% C.I., is there enough evidence to Reject the Null Hypothesis.

Ans $\mu = 168$ c.m. $\sigma = 3.9$ $n = 36$ $\bar{x} = 169.5$ c.m.

(a) Null Hypothesis H_0 $\mu = 168$ c.m.

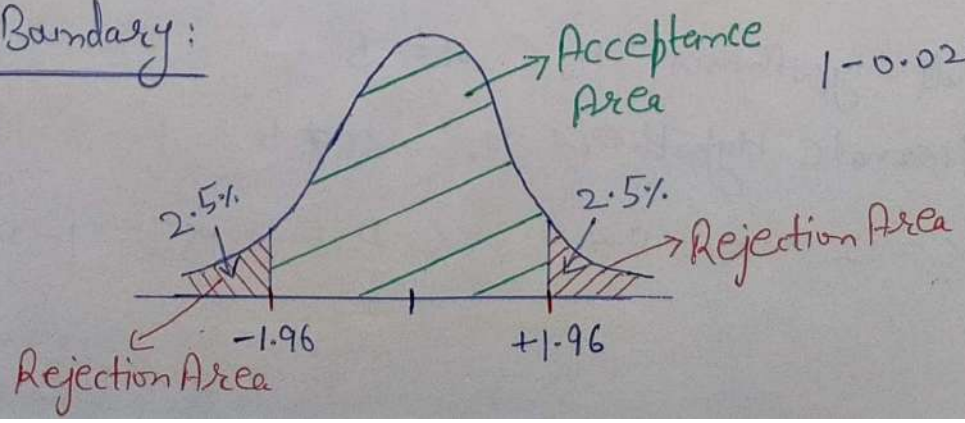
(b) Alternate Hypothesis H_1 $\mu \neq 168$ c.m. {2 Tail Test}

(c) C.I. $= 0.95 \Rightarrow 95\%$ $\alpha = 1 - C.I. = 1 - 0.95 = 0.05$



(d) Decision Boundary:

$1 - 0.025 = \boxed{0.9750}$



If Z-test value falls between -1.96 to $+1.96$ then we we fail to reject the Null Hypothesis.

(e) Statistical Analysis

$$Z\ score = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \qquad \boxed{n=1}$$

$$Z_{test} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \qquad \boxed{\sigma/\sqrt{n}} \Rightarrow \text{Standard error}$$

$$= \frac{169.5 - 168}{3.9/\sqrt{36}} = \boxed{2.31}$$

Conclusion: If Z-test is less than -1.96 or greater than $+1.96$ then Reject the Null Hypothesis.

Here, $2.31 > 1.96$ {We Reject the Null Hypothesis}

② A factory manufactures bulbs with a average warranty of of 5 years with standard deviation of 0.50. A worker believes that the bulb will manfunction 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 the idea that the warranty should be revised.

1 Tail Test

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

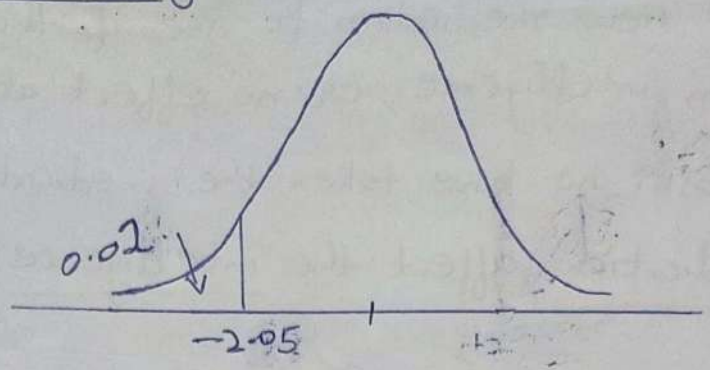
(a) Null Hypothesis H_0 $\mu = 5$

(b) Alternate Hypothesis H_1 $\mu \neq 5$

(c) $\alpha = 2\% \Rightarrow 0.02$ $\quad C.I = 1 - \alpha = 1 - 0.02 = 0.98$

CamScanner

(d) Decision Boundary:

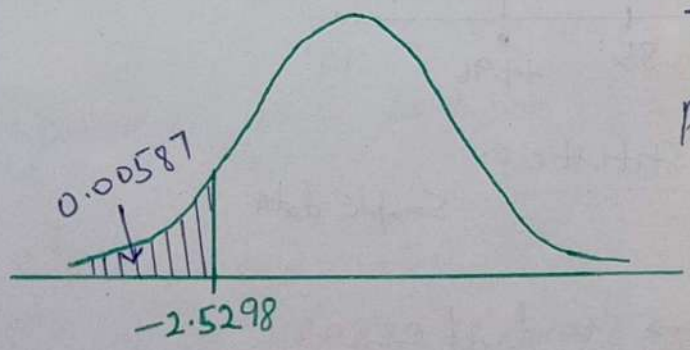


(e) Calculate test statistics (Z-Test)

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{4.8 - 5}{0.50/\sqrt{40}} = -2.5298$$

Conclusion: Here, $-2.52 < -2.05$. So we reject the Null Hypothesis. Warranty needs to be revised.

P value



$\alpha = 0.02$

p-value $< \alpha$ [yes]

Reject the Null Hypothesis

Q In the Population, the average IQ is 100 with a SD of 15. Researchers wants to test a new medication to see if there is positive or negative effect on intelligence, or no effect at all. A sample of 30 participants who have taken the medication has a mean of 140. Did the medication affect the intelligence.

$\alpha = 0.05$ C.I. = 95%

$\sigma = 15$, $n = 30$, $\bar{x} = 140$

$\mu = 100$

Ans ① Define Null Hypothesis

$H_0 = \mu = 100$

② Alternate Hypothesis H_1 $\mu \neq 100$

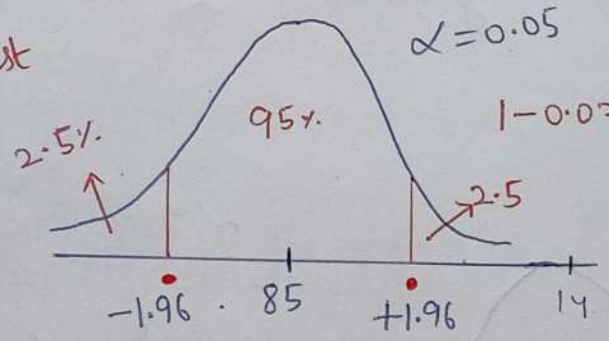
③ State Alpha

$\alpha = 0.05$

if it's rejected we definitely say $\boxed{P \leq 0.05}$

④ State Decision Rule

2 Tail Test



$\alpha = 0.05$

Z table

$1 - 0.025 = 0.975$

⑤ Calculate Z-Test Statistics:

Sample data

$$Z = \frac{\bar{x} - \mu}{\left\{\frac{\sigma}{\sqrt{n}}\right\}} \rightarrow \text{Standard error}$$

$$= \frac{140 - 100}{\frac{15}{\sqrt{30}}} = \frac{40}{15} \times \sqrt{30} = 14.30$$

⑥ State our Decision

$\{Z = -0.2\}$

$14.60 > 1.96$

$Z = 14.60$

If Z is less than −1.96 or greater than 1.96, reject the null hypothesis.

Medication improve the intelligence or decrease?

Improve the intelligence.

Q In the population the average IQ is 100. A team of researchers want to test a new medication to see if it has either a positive or negative effect on intelligence, or no effect at all. A sample of 30 participants who have taken the medication has a mean of 140 with a standard deviation of 20. Did the medication affect intelligence? C.I. = 95%

Ans $\mu = 100$, $n = 30$ $S = 20$ C.I = 95% $\alpha = 1 - 0.95 = 0.05$

① Null Hypothesis $H_0 : \mu = 100$

Alternate Hypothesis $H_1 : \mu \neq 100$ {2 Tail Test}

② Significance value $\alpha = 0.05$

③ Degree of freedom $= n-1$

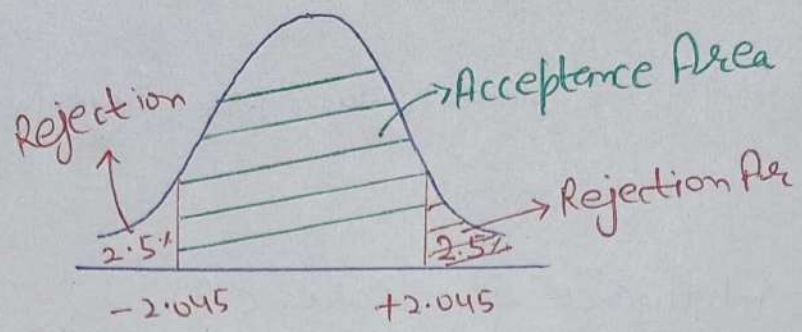
$= 30-1$

$= 29$

CamScanner

④ Decision boundary

dof = 29



If t test is less than -2.045 or greater than 2.045, then we reject the Null Hypothesis.

⑤ Calculate the t test statistics

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

Conclusion:

Decision Rule: If t is less than -2.0452 or greater than 2.0452, Reject the Null Hypothesis.

$t = 10.96 > 2.0452 \Rightarrow$ Reject the Null Hypothesis

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

(a.) State Null and Alternate Hypothesis

(b) At 95% C.I., is there enough evidence to support Machine is working properly or not.

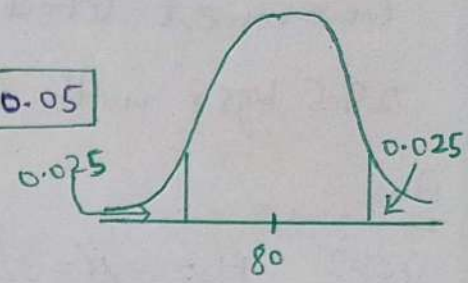
Ans Step I

$\mu = 80ml$ $n = 40$ $\bar{x} = 78$ $S = 2.5$

Null Hypothesis $\mu = 80$

Alternate Hypothesis $\mu \neq 80$

Step 2 C.I = 0.95 S.V(α) = 1 − 0.95 = $\boxed{0.05}$



Step 3

$n = 40$

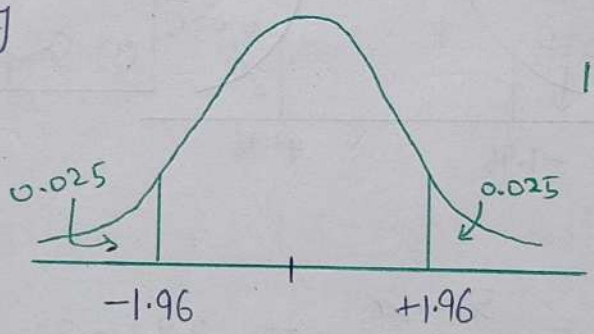
$S = 2.5$

Z test

Let's perform the experiment

Decision Boundary

$1 - 0.025 = 0.975$



Step 4 Calculate test statistics (Z-test)

$$Z = \frac{\bar{x} - \mu}{\boxed{S/\sqrt{n}}} = \frac{78 - 80}{\frac{2.5}{\sqrt{40}}} = -5.05$$

$\boxed{S/\sqrt{n}} \rightarrow$ Standard error

Conclusions:

Decision Rule: If $Z = -5.05$ is less than -1.96 or greater $+1.96$, Reject the Null Hypothesis with 95% C.I.

Reject the Null Hypothesis { There is some fault in the machine.

Q A Complain was registered, the boys in a Government Scholl are underfed. Average weight of boys of age 10 is 32 kgs. with S.D. = 9 kgs. A sample of 25 boys were selected from the Government School and the average weight was found to be 29.5 kgs? with C.I. = 95%. Check it is True or False.

Population mean

Population S.D

Ans

Step I $H_0: \mu = 32$

$H_1: \mu \neq 32$

1 Tail Test

Conditions for Z-test

① We know the Population Sd. OR

② We do not know the Population Sd but our Sample is large ≥ 30

② C.I. = 0.95

$\alpha = 1 - 0.95 = 0.05$

$\bar{x} = 29.5, \sigma = 9, n = 25$

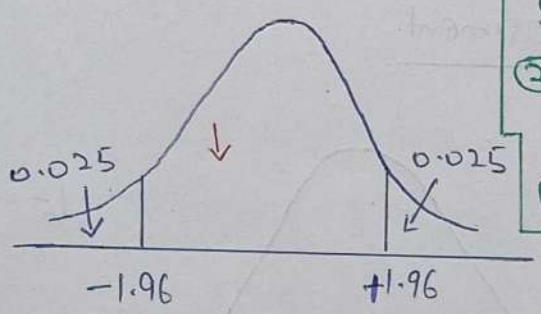
③ Z-test

Conditions for T-test

① We do not know the Population Sd.

② Our Sample size is Small $n \leq 30$

③ Sample Sd is given.



$$Z\text{-score} = \frac{\bar{x} - \mu}{\boxed{\sigma/\sqrt{n}}} = \frac{29.5 - 32}{9/\sqrt{25}} = -1.39$$

S
σ

Conclusion: $-1.39 > -1.96$. So, we Accept the Null Hypothesis

95% C.I. We fail to Reject Null Hypothesis

The boys are fed well.

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

(a) Null {Alternate Hypothesis}

(b) 95%. Is the enough evidence to discard the null Hypothesis?

Ans $\bar{x} = 169.5$ $S = 3.9$ $n = 36$ $\mu = 168$ $C.I. = 0.95$

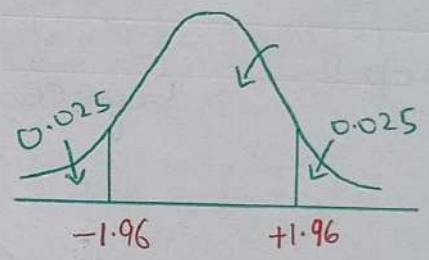
Step I

$H_0: \mu = 168$

$H_1: \mu \neq 168$

Step 2: $C.I = 0.95$ $\alpha = 0.05$

Step 3:



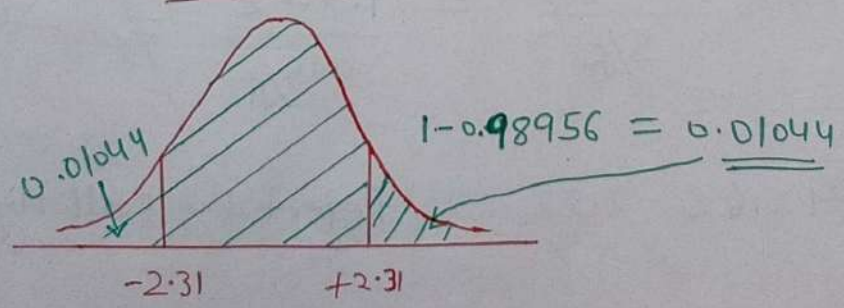
Step 4:

$$Z\text{-Score} = \frac{169.5 - 168}{3.9/\sqrt{36}} = \simeq \boxed{2.31} \Rightarrow 2.3076$$

$2.31 > 1.96$ {Reject the Null Hypothesis}

P-value

2 Tail Test



P-value $= 0.01044 + 0.01044 = 0.02088$

$0.02088 < \boxed{\alpha}$ 0.05 {Reject the Null Hypothesis}

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

(a) State the Null and Alternate Hypothesis

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

Ans $\mu = 2$ $n = 10$ $\bar{x} = 1.8$ $S = 0.15$ $C.I. = 0.99$ $\alpha = 0.01$

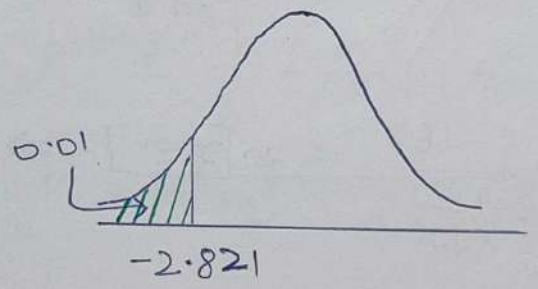
Step I $H_0: \mu \geq 2$

$H_1: \mu < 2$

1 Tail Test

Step II Degree of freedom $= n-1$

$= 10-1 = 9$



Step III Calculate t test statistics

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}} = \frac{1.8 - 2}{0.15/\sqrt{10}} = -4.216$$

Step IV $-4.216 < -2.82$ {Reject the Null Hypothesis}

Conclusion: The average life of the battery is less than 2 years.

(*) Z test with Proportions

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

(a) State Null and Alternate Hypothesis — 2 Tail Test

(b) At a 95% C.I., is there enough evidence to reject the Null Hypothesis?

Ans Step 1

Null Hypothesis: $P_0 = 0.70$

Alternate Hypothesis: $P_0 \neq 0.70$

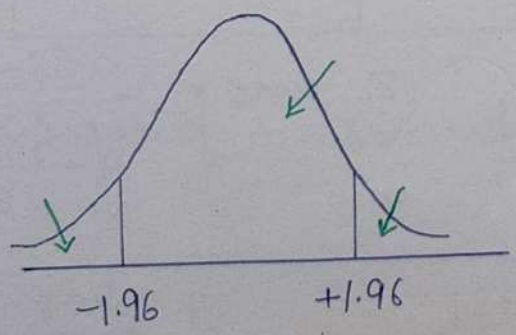
$n = 200$ $\boxed{x = 130}$

$\hat{P} = \frac{130}{200} = \boxed{0.65}$

$q_0 = 1 - P_0 = \boxed{0.30}$

Step 2 $C.I = 0.95$ $\alpha = 0.05$

Step 3:



Step 4: Z-test with Proportion

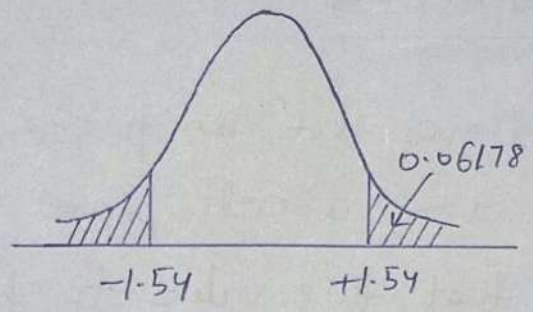
$$Z\text{-test} = \frac{\hat{P} - P_0}{\sqrt{\frac{P_0 q_0}{n}}}$$

$$= \frac{0.65 - 0.70}{\sqrt{\frac{0.7 \times 0.3}{200}}} \approx -1.54$$

Conclusion: $-1.54 > -1.96$ Fail to Reject the Null Hypothesis.

CamScanner

Ratio



P-value = 0.06178 + 0.06178

= 0.12356

P-value > Significance value → Fail to Reject the Null Hypothesis.

Q A car company believes that the percentage of residents in city ABC that owns a vehicle is 60% or less. A sales manager disagrees with this. He conducts a Hypothesis testing surveying 250 residents and found that 170 responded yes to owning a vehicle.

(a) State the Null and Alternate Hypothesis

(b) At a 10% significance level, is there enough evidence to support the idea that vehicle ownership in city ABC is 60% or less?

Ans $H_0: P_0 \leq 0.60$

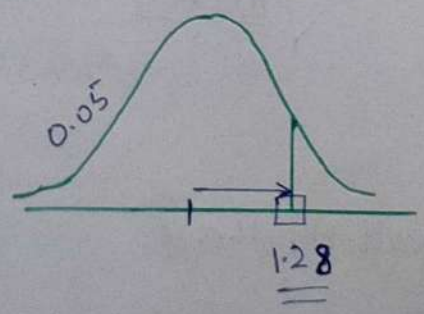
$H_1: P_0 > 0.60$

$q_0 = 1 - P_0 = 0.40$

$\hat{p} = \frac{170}{250} = 0.68$

$$Z\text{-score} = \frac{0.68 - 0.60}{\sqrt{\frac{0.6 \times 0.4}{250}}}$$

$$= \frac{0.08}{0.0309} = 2.588$$



Reject the Null Hypothesis

Agenda

1. CHI SQUARE
2. Covariance
3. Pearson Corelation Coefficient
4. Spearman Rank Correlation
5. Practical Implementation
   Z-test, t-test, Chi-Square test
6. F test (ANOVA)

Why CHI SQUARE TEST Performed?

Interview

1. CHI Square Test claims about population proportions.
   It is a non parametric test that is performed on categori-cal (nominal or ordinal) data.

Q In the 2000 Indian Census, the age of the individual in a small town were found to be the following:

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

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

<18	18-35	>35
121	288	91

CamScanner

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

Ans

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

{Population} 2000

Expected

<18	18-35	>35
121	288	91
500×0.2	500×0.3	500×0.5
100	150	250

$\boxed{n = 500}$

observed

Expected

(Chi Square Table)

<18	18-35	>35
121	288	91
100	150	250

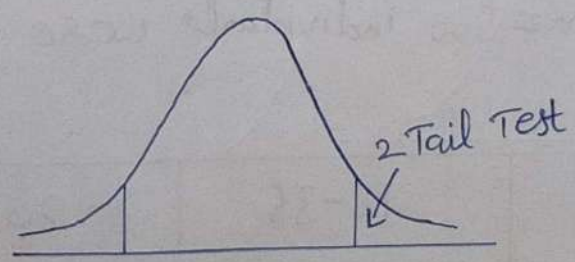
① H_0 = The data meets the distribution 2000 census

H_1 = The data does not meet " " "

② $\alpha = 0.05$ (95% C.I.) $df = 2$, $\alpha = 0.05$

③ Degree of freedom $= n-1 = 3-1 = 2$

④ Decision Boundary



If χ^2 is greater than 5.99, then Reject the Null Hypothesis H_0.

⑤ Calculate Test Statistics

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

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

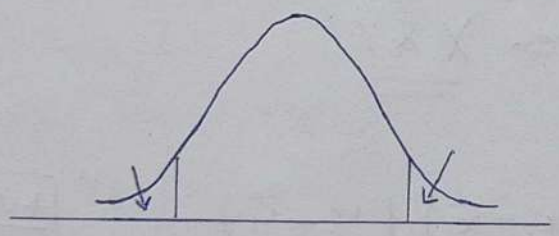
$$= 232.494$$

$x^2 = 232.494 > 5.99$ { Reject the Null Hypothesis }

$\alpha = 0.05$ 0.01, 0.10 } Domain

$0.11 > 0.05$
⇓
Accept the Null Hypothesis

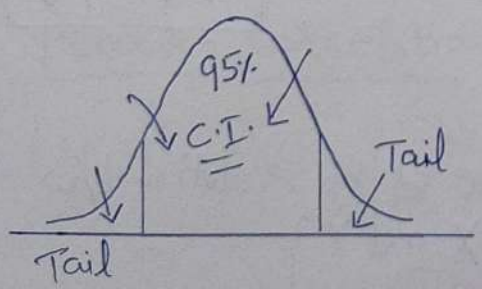
$0.002 < 0.05$



{ P-value < Significance value
⇓
Reject the Null Hypothesis }

OR

{ Accept the Null Hypothesis }



P $0.11 > 0.05$
⇓
Accept

$P = \boxed{0.002} < 0.05$
⇓
Reject the Null Hypothesis

② Covariance:

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

$\left\{ \begin{array}{l} \boxed{X\uparrow \quad Y\uparrow} \\ \boxed{X\downarrow \quad Y\downarrow} \end{array} \right\}$

No. of hour Study	Play
2	6
3	4
4	3

$\boxed{\begin{array}{ll} X\uparrow & Y\downarrow \\ X\downarrow & Y\uparrow \end{array}}$

Quantity relationship between X & Y

Covariance

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

$\boxed{+ve}$ Positive Correlation

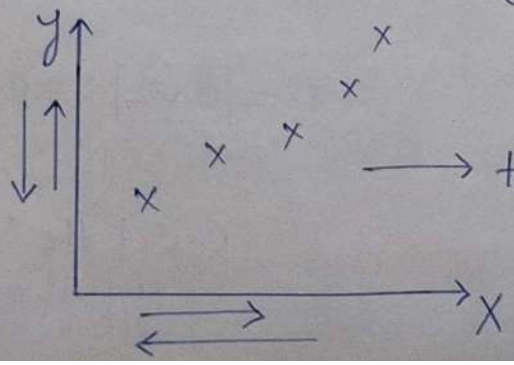
$\boxed{\begin{array}{ll} X\uparrow & Y\uparrow \\ X\downarrow & Y\downarrow \end{array}}$

$\boxed{-ve}$ negative Correlation

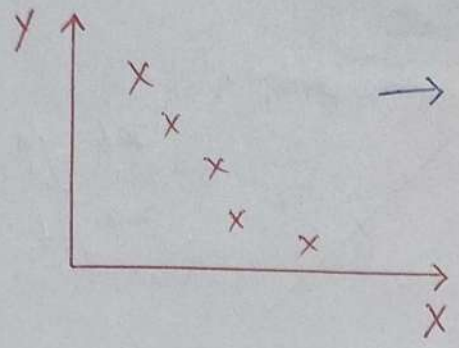
$\boxed{\begin{array}{ll} X\downarrow & Y\uparrow \\ X\uparrow & Y\downarrow \end{array}}$

= +ve or -ve

0 {There is no relation b/w X and Y}



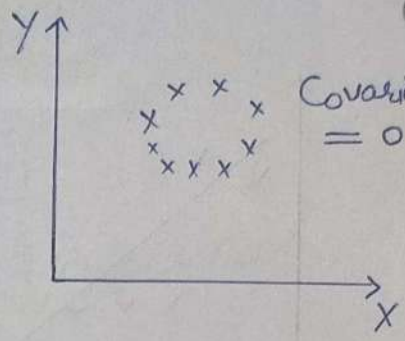
→ +ve Correlation



→ -ve Correlation

⇓

-ve Covariance



Covariance = 0

Disadvantage of Covariance

① Positive or Negative ✓

+100 +1000 −200 −2000

{Direction}

X	Y
10	4
8	6
7	8
6	10
7.75	7

$Cov(x,y) = -ve$

$= [(10-7.75)(4-7)+(8-7.75)(6-7)+(7-7.75)(8-7)$
$+ (6-7.75)(10-7)$

$= -3.25$

X↑	Y↓
X↓	Y↑

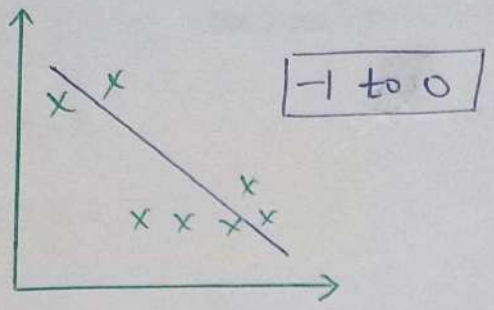
③ Pearson Correlation Coefficient:

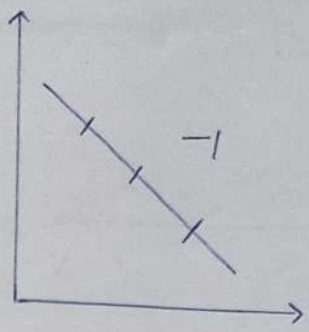
(−1 to 1)

The more towards +1 more positively correlated.

The more towards −1 more negatively correlated.

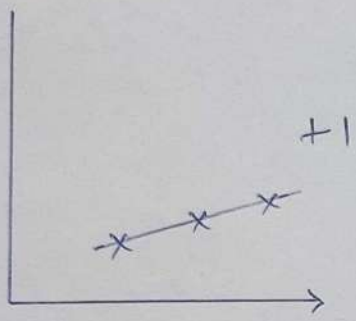
$$\rho(x,y) = \frac{Cov(X,Y)}{\sigma_x \sigma_y}$$ {−1 to 1}

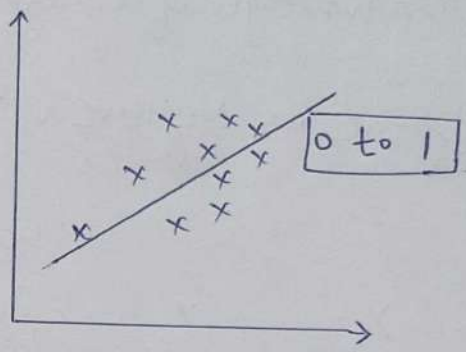




$x\downarrow \; y\uparrow$

$x\uparrow \; y\downarrow$





Interview

Why use Pearson Correlation?

For testing whether there is significant relationship b/w two variables.

(4) Spearman Rank Correlation

$$Spear(x,y) = \frac{Cov(R(x), R(y))}{R\sigma_x * R\sigma_y}$$

x Height	y Weight
170	75
160	62
150	60
145	55
180	85

$R(x)$	$R(y)$
2	2
3	3
4	4
5	5
1	1

Interview

Why we use Spearman Rank Correlation?

It Captures Non Linear Properties

1. P value & significance value
2. Distribution
3. Bernoulli's Distribution
4. Binomial Distribution
5. Pareto's Distribution {Power law}
6. F Test (ANOVA)

Log Normal Distribution

1. P value & Significance value

↳ Derive the P value

Q The average weight of all residents in Bangalore city is 168 pounds with a standard deviation 3.9. We take a sample of 36 individuals and the mean is 169.5 pounds. C.I. = 95%

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

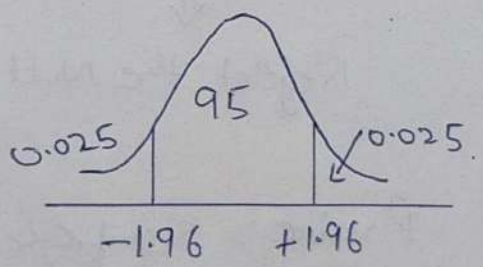
1. $H_0 = \mu = 168$

   $H_1 : \mu \neq 168$

2. $\alpha = 0.05$

3. Decision Boundary

$1 - 0.025 = 0.975$



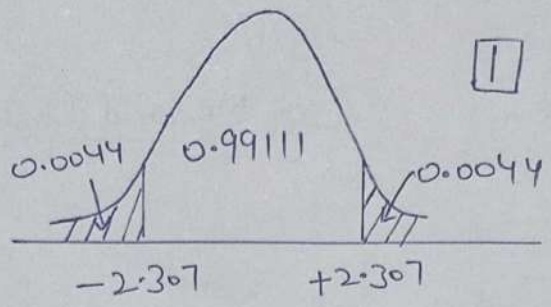
4. Z test

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

$$= \frac{1.5}{3.9} \times 6 = 2.307$$

$Z = 2.307 > 1.96$ Reject the Null Hypothesis

P value



$1 - 0.99111 = 0.00889$

$$\boxed{\begin{aligned} P \text{ value} &= 0.0044 + 0.0044 \\ &= \underline{\underline{0.0088}} \end{aligned}}$$

P value < 0.05

$\boxed{0.0088 < 0.05} \rightarrow$ Reject the Null Hypothesis

NOTE:

P value $\leq$ Significance value

$\Downarrow$

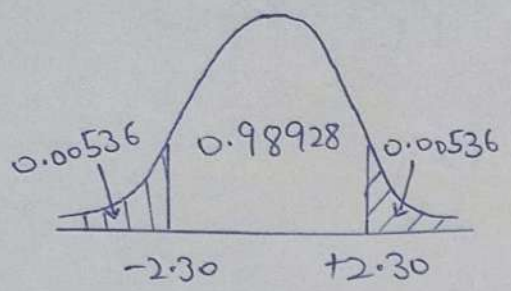
Reject the Null Hypothesis

P value $>$ Significance value

$\Downarrow$

Fail to Reject the Null Hypothesis

CamScanner



$Z = 2.30$

$1 - 0.98928 = 0.01072$

$2.30 > 1.96$ {Reject the Null Hypothesis}

P value $= 0.00536 + 0.00536$

$= 0.01072 \leq \alpha \Rightarrow$ Reject the Hypothesis

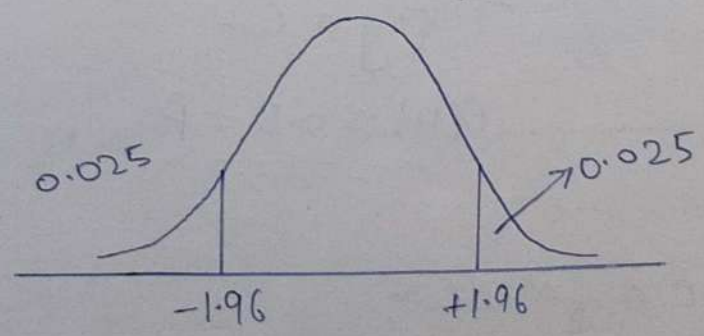
Q Average age of a College is 24 years with a standard deviation 1.5. Sample of 36 students mean is 25 years. with $\alpha = 0.05$, C.I = 95%, does the age vary?

Ans ① $H_0: \mu = 24$ $\sigma = 1.5$ $n = 36$ $\bar{x} = 25$ $\alpha = 0.05$

$H_1: \mu \neq 24$

② $\alpha = 0.05$

③ Decision Boundary

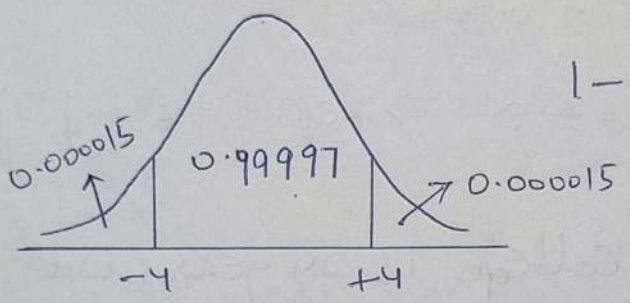


(4) $$Z\text{-Score} = \frac{x-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{25-24}{1.5} \times 6$$

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

$$= 4$$

$4 > 1.96$ {Reject Null Hypothesis}



$1 - 0.99997 = 0.00003$

P value $= 0.000015 + 0.000015$

$= 0.00003$

P value $\leq$ Significance { Reject the Null Hypothesis}

(3) Bernoulli's Distribution:

2 outcomes $\boxed{0 \text{ or } 1}$ Single Trial

Tossing a Coin

$P = 0.5$

$q = 1 - P = 0.5$

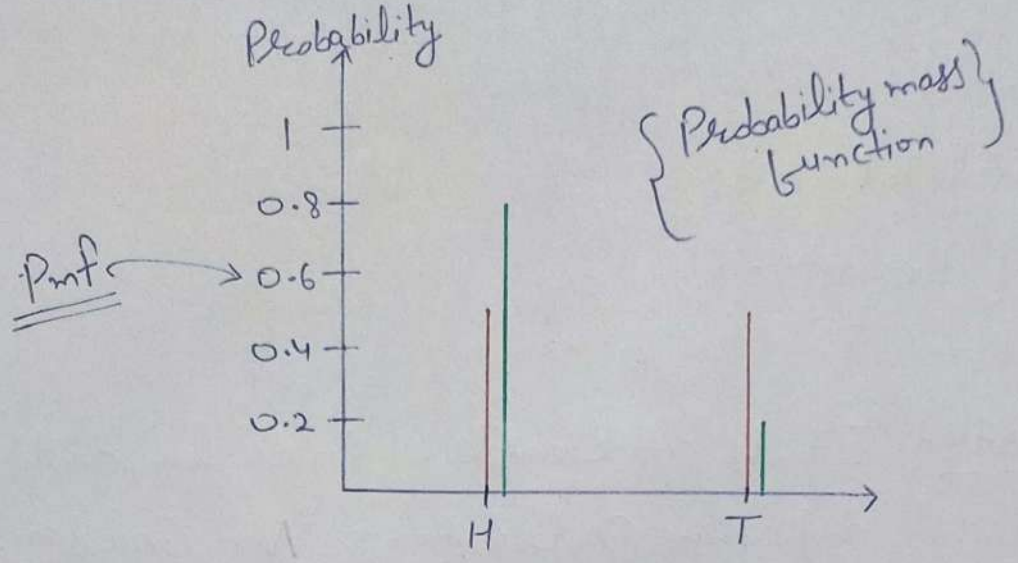
$P(H) = 0.5 = P$

Do not have a fair Coin

$P(H) = 0.3 = P$

$P(T) = q = 1 - P = 1 - 0.3 = 0.7$

CamScanner



* Probability mass function (Pmf): used for Categorical variables.

Probability density function (Pdf): used for Continuous variables.

④ Binomial Distribution:

Every Trial → Bernoulli's distribution

Multiple Trial

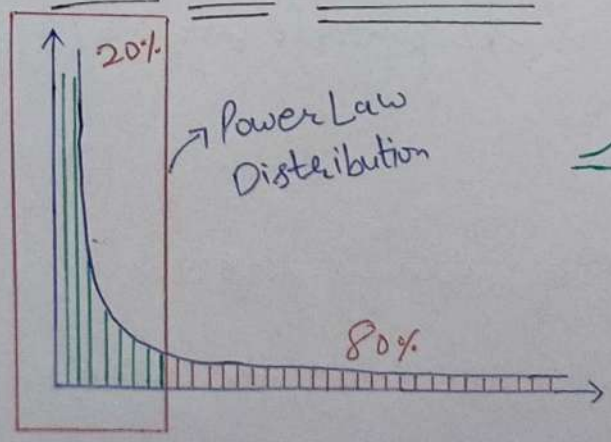
↓

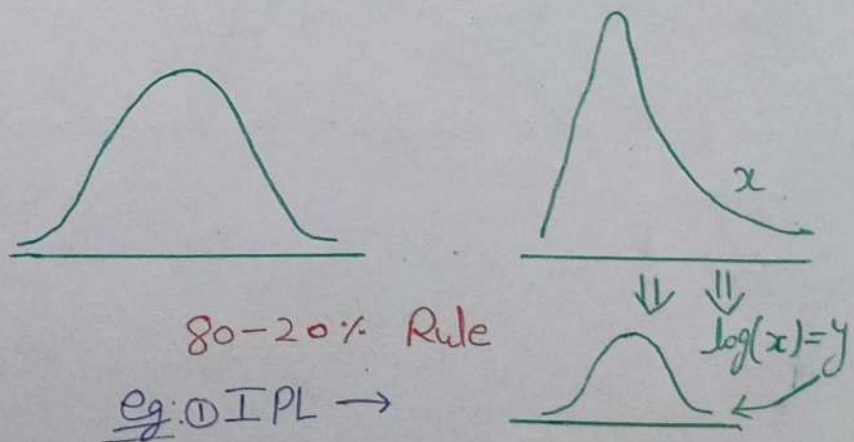
$P(H) = 0.5$ $\quad$ $P(H) = 0.6$ $\quad$ - - - - -

$P(T) = 0.5$ $\quad$ $P(T) = 0.4$

Interview

⑤ Power Law Distribution





80-20% Rule

eg: ① IPL →

20% of Team is responsible for winning 80% of match

② 80% of wealth is distribution 20% of the total Population.

③ Oil richnation

{ 80% of the total oil is with 20% of the nation }

Interview

Pareto Distribution: It is an example of Power law distributio and this Distribution is Non Gaussian.

Interview: Can we Convert a Pareto Distribution into a Normal Distribution? Interview Question

Ans Yes, By using Box Cox transformation we convert a Pareto distribution into Normal Distribution.

Importance of Normal Distribution: If we have any Normal Distributed data, then by using the Machine Learning algorithms like Linear Regression, Logistic Regression we are easily able to make a Efficient Model.

Example of Pareto Distribution:

① 20% of the Product in Amazon is responsible for 80% of sales.

② 80 - 20%

③ 20% of defects solves the 80% of upcoming defects.

④ 20% of Team is responsible in delivering the 80% of the projects.

CamScanner

Interview: Relationship between Log Normal and Pareto Distribution?

Done ANOVA (F Test)

Done Poisson Distribution

Box Cox Transformation

→ Homework

Done Q-Q Plot

↓

it is done in FE

Tutorial-16

Hindi Playlist

Probability density function and Probability distribution fun.

⇓

Done

Q Why we divide sample variance by $n-1$?

Ans The sample variance is divide by $n-1$ so that we can create an unbiased estimator of the population variance.

Random Variables: Random variable is a process of mapping the output of a random process or experiment to a number.

$\Downarrow$

Experiment

eg:- Tossing a Coin

$$x = \begin{cases} 0 & \text{if Head} \\ 1 & \text{if Tail} \end{cases}$$

Rolling a dice 6 times

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

$y = \{\text{Sum of Rolling a dice 7 times}\}$

$Pr(y \geq 15)$ $\quad$ $Pr(y < 10)$

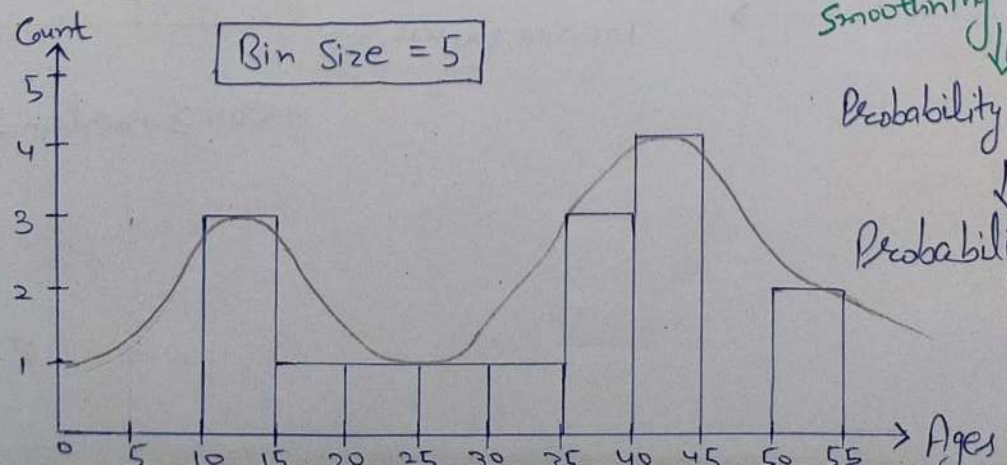
Histograms And Skewness $\rightarrow$ [Frequency] $\Rightarrow$ Distribution

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

Bins, Bin Size

$\frac{50}{5} = \boxed{10} \Rightarrow$ No. of bins

$\boxed{\text{Bin Size} = 5}$



[Kde]

Smoothning the Histogram

$\Downarrow$

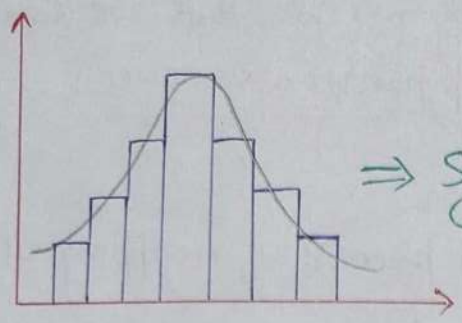
Probability distribution function

$\Downarrow$

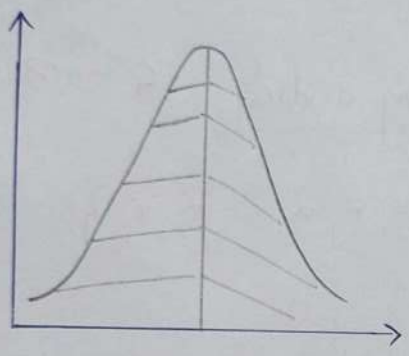
Probability density function

Skewness

Normal / Gaussian Distribution



⇒ Symmetrical Distribution

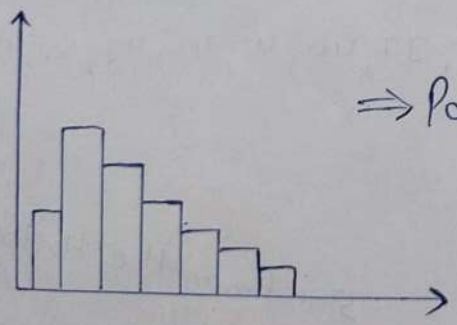


① No Skewness

Mean, Median, Mode

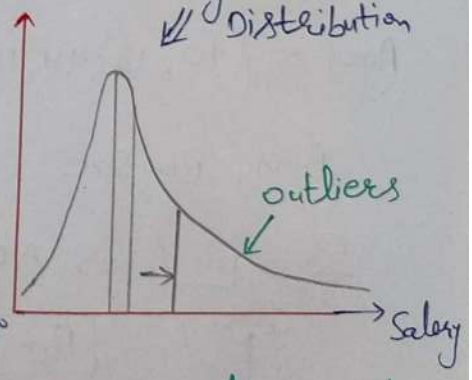
$\boxed{\text{Mean} = \text{Median} = \text{Mode}}$

Right Skewed



⇒ Positive skewed ⇒

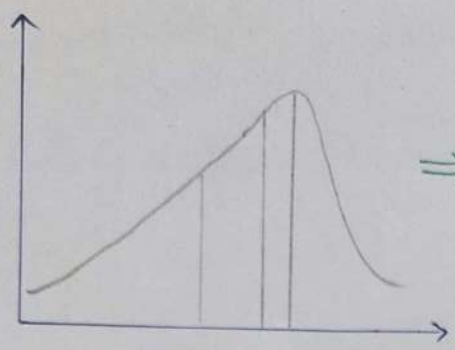
Log Normal Distribution



100, 200, 300, 400, 100000 →

mean ≥ median ≥ mode

Left skewed



$\Rightarrow$ Negative skewed

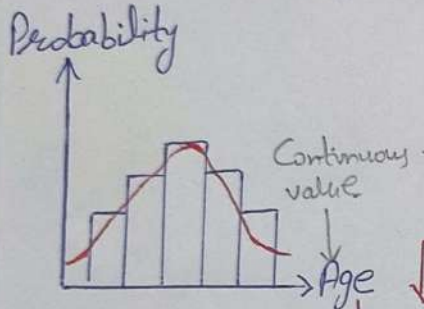
Relationship: mean $\leq$ median $\leq$ mode

CamScanner

Agenda

1. Probability Distribution function
2. Probability Density function (Pdf)
3. Probability mass function (Pmf)
4. Different types of distribution
5. CDF {Commulative distribution function}
6. Log Normal Distribution

$x = \{\text{Continuous}\}$ or Discrete values

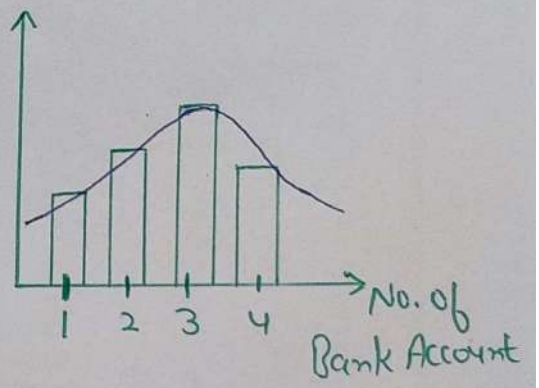


{Distribution of data}

Probability Distribution function

Discrete value → No. of Bank accounts

- Probability density function
- Probability mass function



1. Probability Density function

1. Continuous Random variable

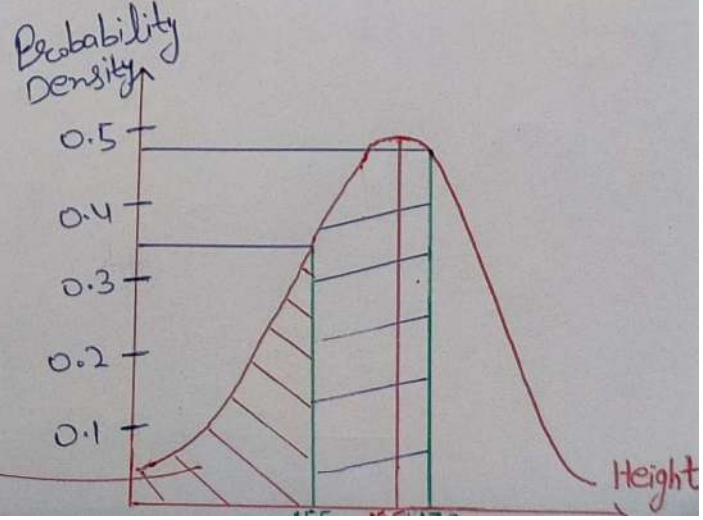
eg: Height of students in classroom

Formula

$\boxed{Pr(H \leq 155)} \Rightarrow$ Formula

$Pr(H \geq 155 \text{ and } H \leq 170)$

Area under this curve



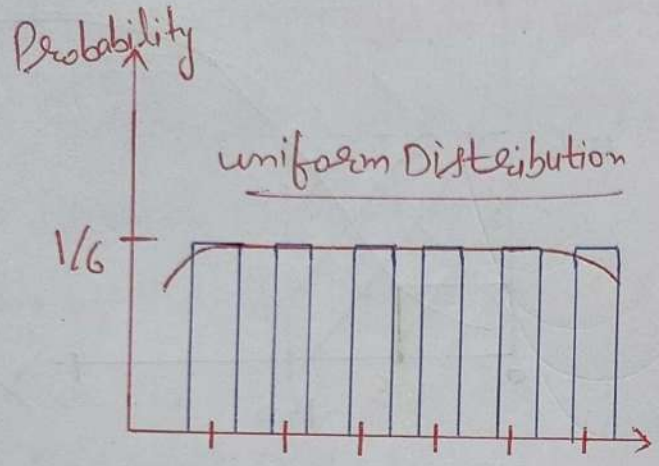
② Probability mass Function (Pmf)

① Discrete Random variable

eg:- Rolling a Dice $\{1, 2, 3, 4, 5, 6\}$

$Pr(1) = \frac{1}{6}$ $\quad Pr(3) = \frac{1}{6}$

$Pr(2) = \frac{1}{6}$ $\quad Pr(4) = \frac{1}{6}$



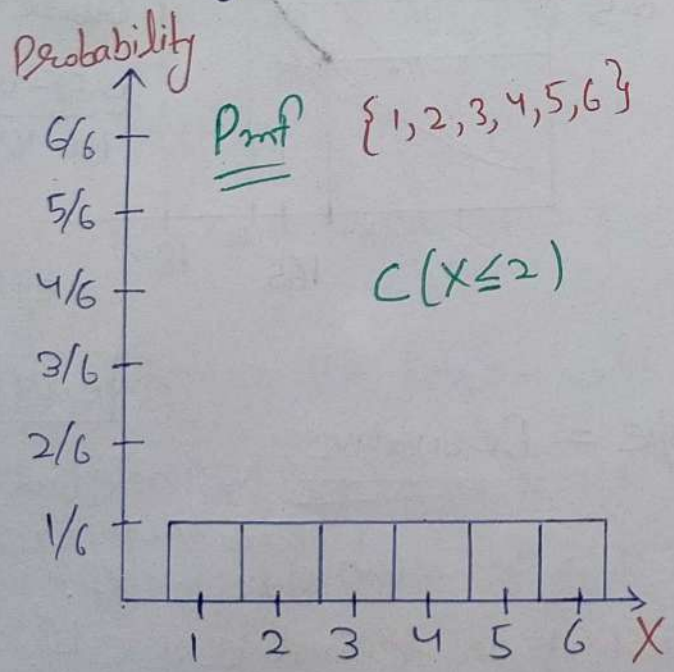
$$Pr(X \leq 4) = Pr(X=1) + Pr(X=2) + Pr(X=3) + Pr(X=4)$$

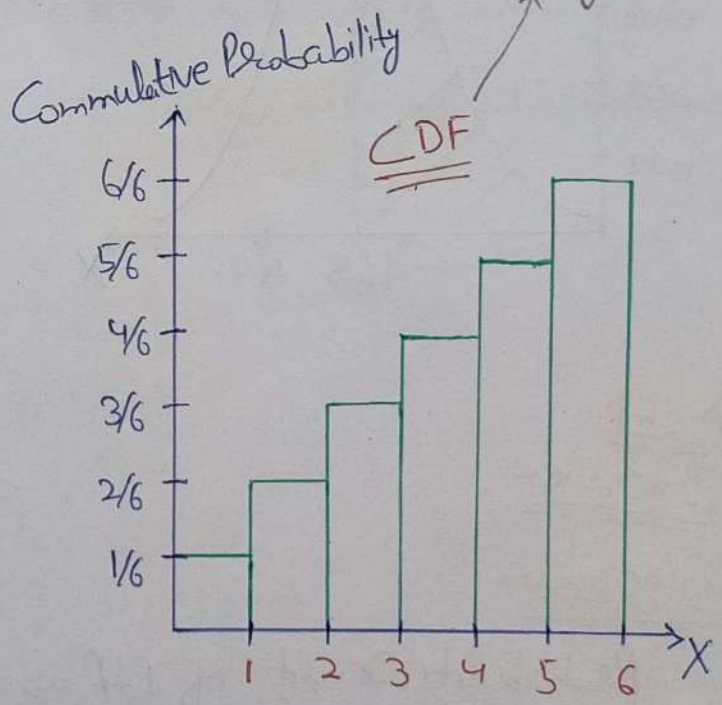
$$= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{\cancel{4}^{2}}{\cancel{6}_{3}} = \frac{2}{3}$$

③ Commulative Distribution Function (Relationship b/w Pmf and CDF)

Commulative Density function

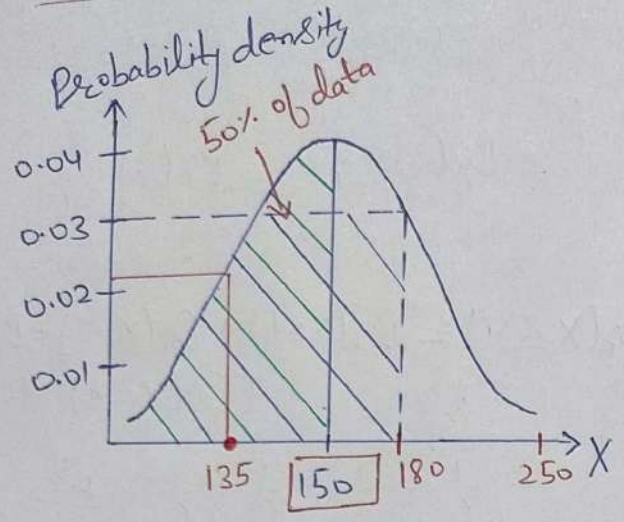
eg: Rolling a Dice

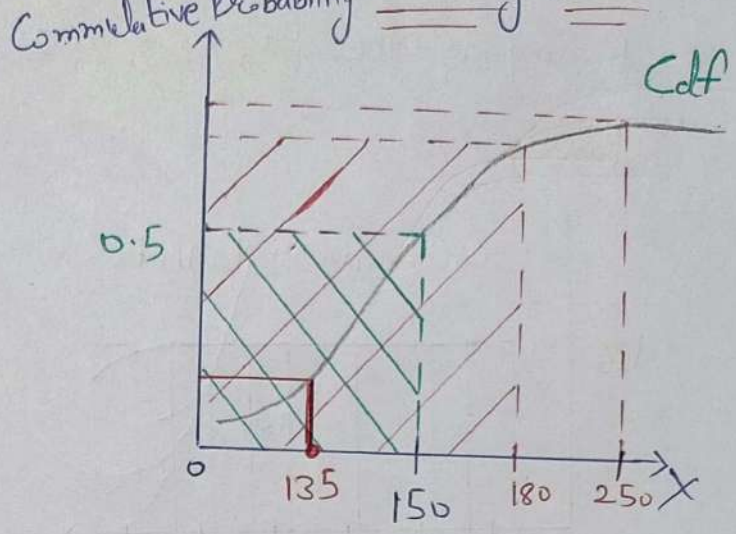


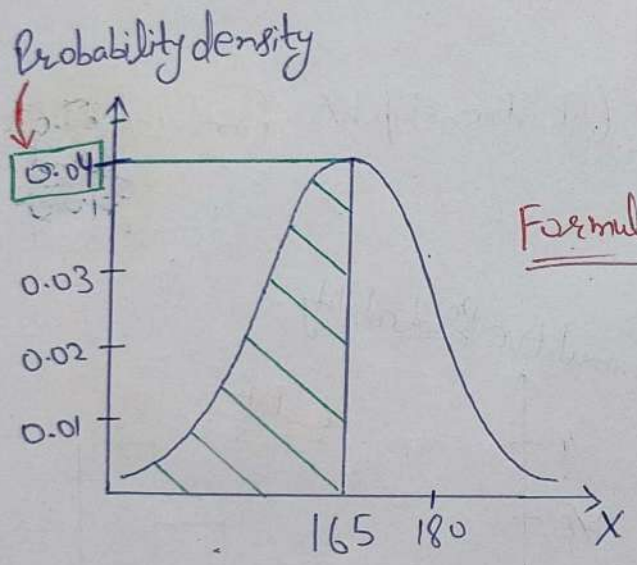


M.Imp Interview

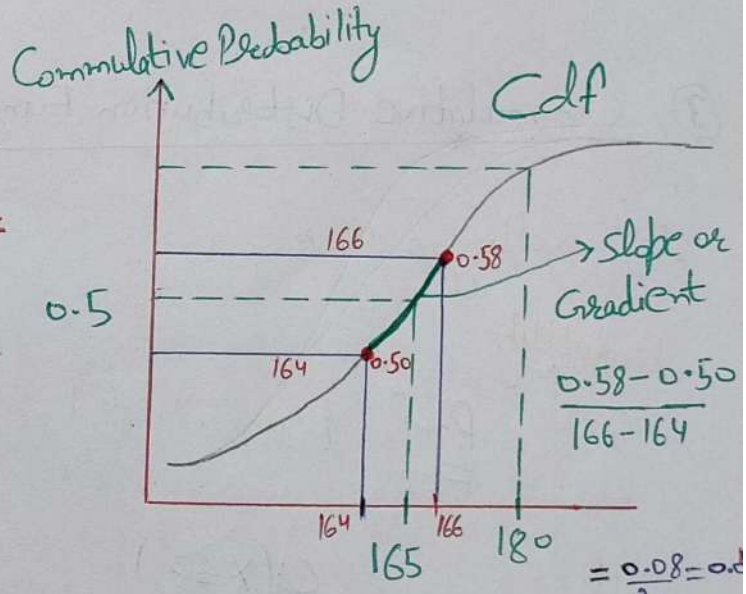
Relationship b/w Pdf (Probability density function) and Cdf (Commulative density function)







Formula ⇐



$= \frac{0.08}{2} = 0.$

$= \boxed{0.04}$

M.Imp Interview

Slope = Derivative

we are find the Prob. density by Calculat the slope in cdf.

Conclusion:

Probability Density of Pdf ⇒ Gradient or derivative of Cdf.

Probability density function ← Pdf is the derivative of cdf.

Commulative density function ← cdf is the integration of Pdf.

Types of Probability Distribution

1. Normal / Gaussian Distribution (Pdf)
2. Bernoulli Distribution (Pmf) → outcomes are binary only.
3. Uniform Distribution (Pmf)
4. Poisson Distribution (Pmf)
5. Binomial Distribution (Pmf)
6. Log Normal Distribution (Pdf)

1. Bernoulli Distribution:

1. Discrete Random variable {Pmf}
2. Outcomes are Binary

$P \Rightarrow$ Probability of one outcome

$q \Rightarrow$ Probability of other outcome

eg: Tossing a coin {H, T}

$Pr(H) = 0.5 = P$

$Pr(T) = 0.5 = 1-P = q$

eg: Whether the person will Pass / Fail

$Pr(Pass) = 0.7 = P$

$Pr(Fail) = 1-P = 0.3 = q$

Parameters

$0 \le P \le 1$

$q = 1-P$

$K = \{0, 1\} \Rightarrow$ outcomes

PMF

$$\begin{cases} q = 1-P & \text{if } K=0 \\ P & \text{if } K=1 \end{cases}$$

$\Downarrow$

Formula

$$\boxed{PMF = P^K * (1-P)^{1-K}}$$

if $K=1$, $PMF = p^{(1)} * (1-p)^{1-1}$

$$\boxed{PMF = p}$$

Heads and Tail

$K = 0, 1$

p	q
$Pr(H)$	$Pr(T)$

if $K=0$, $K \in \{0,1\}$

$$Pmf = p^0 * (1-p)^1$$

$$= (1-q) = q$$

② Mean of Bernoulli Distribution

$$E(K) = \sum_{i=0}^{K} K \cdot P(K)$$

$Pr(K=1) = 0.6 \Rightarrow P$

$Pr(K=0) = 0.4 \Rightarrow 1-P$

$$= [0 * 0.4 + 1 * 0.6]$$

$$\boxed{= 0.6 = p}$$

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

Interview

* Bernoulli Distribution

Note → Refer Wikipedia

① Discrete Random variable.

② Every Experiment outcome is Binary.

③ These experiment is performed for n trials.

** Bernoulli Distribution → for n number of trials.

eg: Tossing a Coin 10 times

Notation: $B(n, p)$

$Pr(H) = 0.5 = p$

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

Parameters:- $n \in \{0, 1, 2, 3, \ldots\} \Rightarrow$ no. of trials

$p \in \{0, 1\} \rightarrow$ Success probability for each trial.

$q = 1 - p$

Support = $k \in \{0, 1, 2, 3, \ldots, n\} \Rightarrow$ Number of Successes.

PMF

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

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

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

Interview

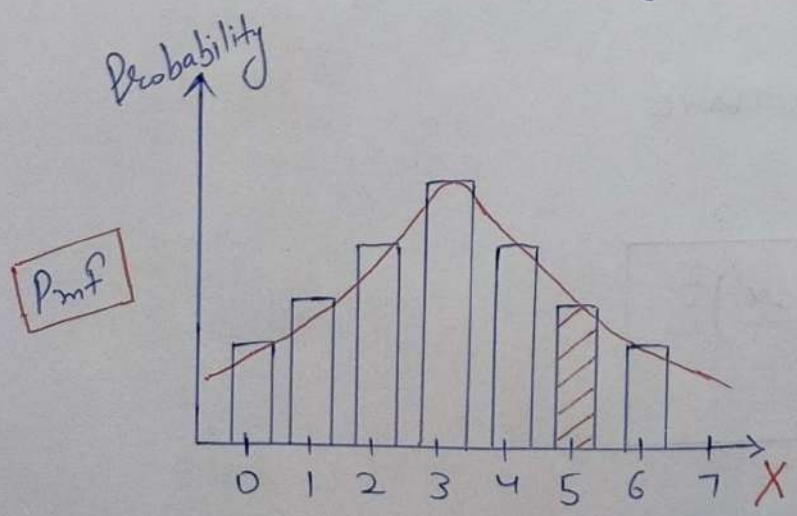
* Poisson Distribution

① Discrete Random variable (Pmf)

② Describe the Number of event occuring in a fixed time interval.

eg: No. of people visiting hospital every hour.
No. of people visiting banks at 11 am.

$\boxed{\lambda = 3} \Rightarrow$ Expected no. of event occur at every time interval. (Threshold)



$x = 5$

$\boxed{\lambda}$

$\lambda = 3 \Rightarrow$ Let us consider $\lambda = 3$

$$Pr(X=5) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

$$= \frac{e^{-3}3^{5}}{5!} = 0.101 = \underline{\underline{10.1\%}}$$

4Pm and 5Pm

$$Pr(X=4 \text{ and } X=5) = Pr(X=4) + Pr(X=5)$$

* Normal / Gaussian Distribution (Pdf)

(*) Continuous Random variable

Ages, Height, weight

$x = \{ - - - - - - - - - \}$

$\Downarrow$

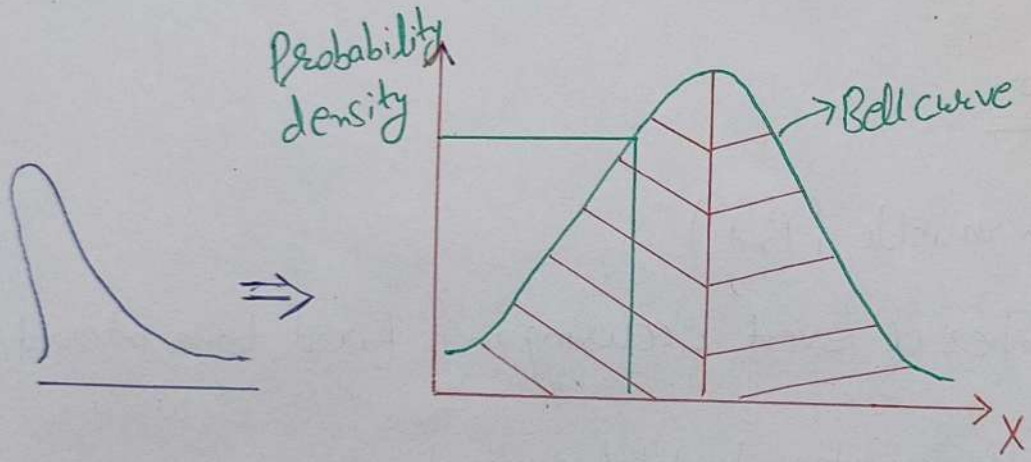
50%

IRIS

Domain Experts

$\Downarrow$

Doctors



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

Parameters: $\mu \in R$

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

$x \in R$

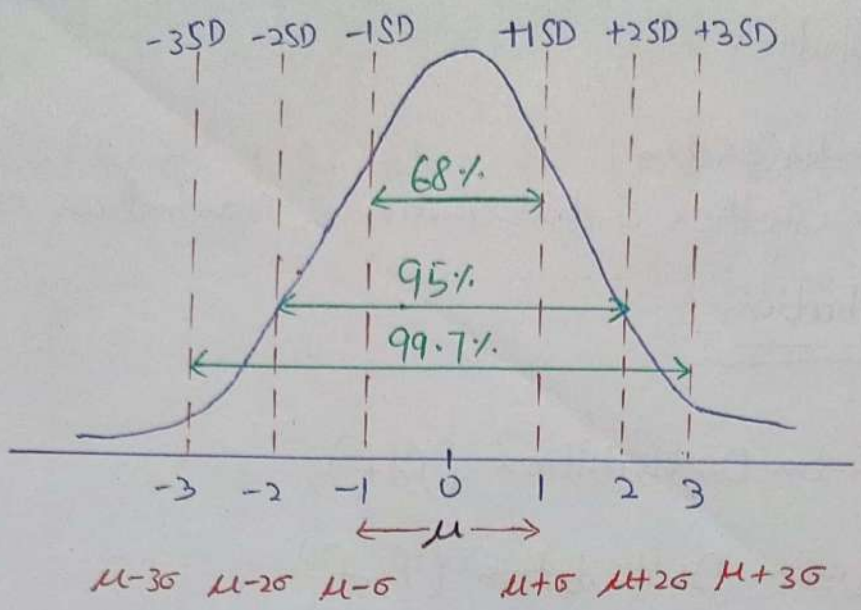
$$\boxed{Pdf = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}}$$

Mean $= \mu$ variance $= \sigma^2$ Std $= \sqrt{\text{variance}}$

Empirical Rule of Normal Distribution

Interview

Standard Normal Distribution



$$\mu-\sigma \le x \le \mu+\sigma \approx 68\%$$

$$\mu-2\sigma \le x \le \mu+2\sigma \approx 95\%$$

$$\mu-3\sigma \le x \le \mu+3\sigma \approx 99.7\%$$

Interview: How to convert Poisson distribution into normal distribution?

Agenda

① Uniform Distribution

② Central Limit Theorem

③ Q-Q Plot (Check whether a distribution is Normal or Not)

① Uniform Distribution

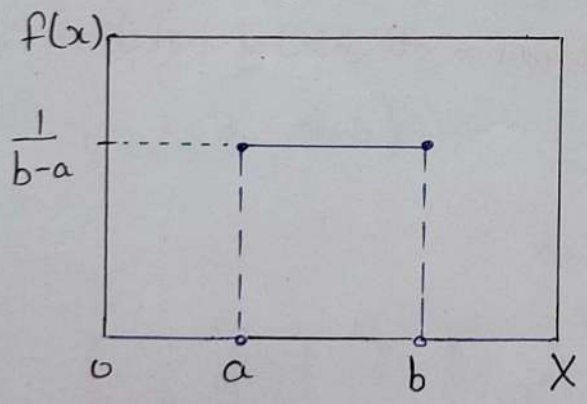
① Continuous Uniform Distribution (Pdf)

② Discrete Uniform Distribution (Pmf)

① Continuous uniform Distribution [Continuous Random variables]

Continuous uniform distribution
with parameters a and b

Probability density function



Notation = $U(a,b)$

Parameters = $-\infty < a < b < \infty$

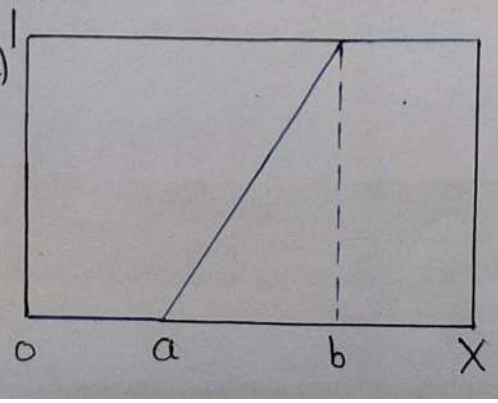
$$\text{Pdf} = \begin{cases} \frac{1}{b-a} & \text{for } x \in [a,b] \\ 0 & \text{otherwise} \end{cases}$$

Cummulative distribution function

mean and median is same because it is symmetrical.

$$\boxed{\text{mean} = \frac{a+b}{2}}$$

$$\boxed{\text{median} = \frac{a+b}{2}}$$



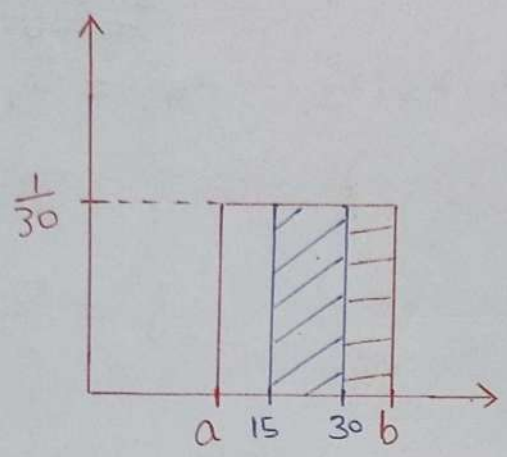
$$\text{Cdf} = \begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{x-b} & \text{for } x[a,b] \\ 1 & \text{for } x > b \end{cases}$$

CamScanner

eg: The number of candies sold daily at a shop is uniformly distributed with a maximum of 40 and a minimum of 10.

(i) Probability of daily sales to fall between 15 and 30?

Ans



$\boxed{\frac{1}{b-a}}$

$\boxed{a=10}$ $\boxed{b=40}$

$x_1 = 15$

$x_2 = 30$

$Pr(15 \le x \le 30) = (x_2 - x_1) * \frac{1}{b-a}$

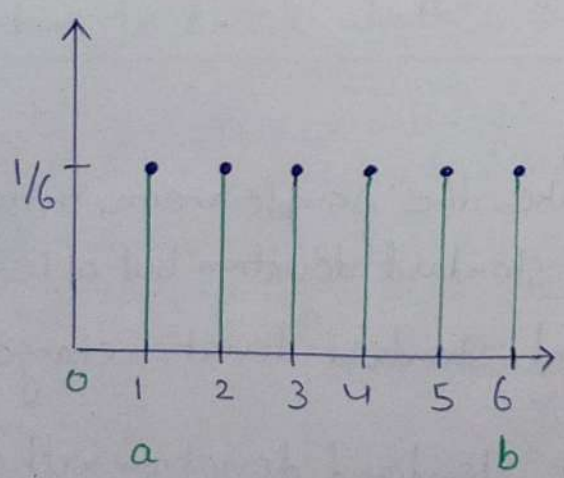
$= \cancel{15}^{1} * \frac{1}{\cancel{30}_{2}} = 0.5$

$Pr(X \ge 30) = (40-30) \times \frac{1}{30} = \cancel{10} \times \frac{1}{\cancel{30}_{3}} = \frac{1}{3} = 0.33 = \boxed{33\%}$

② Discrete Uniform Distribution (Pmf)

eg: Rolling a Dice

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



$n = 6$

$n = b - a + 1$

$= 6 - 1 + 1$

$n = \underline{\underline{6}}$

$Pr(1) = \frac{1}{n} = \frac{1}{6}$

Notation $u(a,b)$

Parameter a, b $\boxed{b \ge a}$

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

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

$Median = \frac{a+b}{b}$

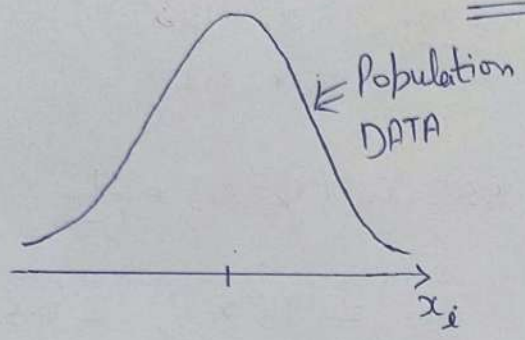
Symmetrical

Interview M.Imp

Central Limit Theorem: It says whether your distribution is Gaussian or non-Gaussian distribution, if we take the sample where $n \geq 30$ and we take all the sample mean and we populated we get a normal distribution.

$$\boxed{n \geq 30}$$

① $X \simeq N(\mu, \sigma)$



n sample

Sampling mean

$S_1 = \{x_1, x_2, x_3 \text{ ---- } x_n\} = \bar{x}_1$

$S_2 = \{ \qquad \} = \bar{x}_2$

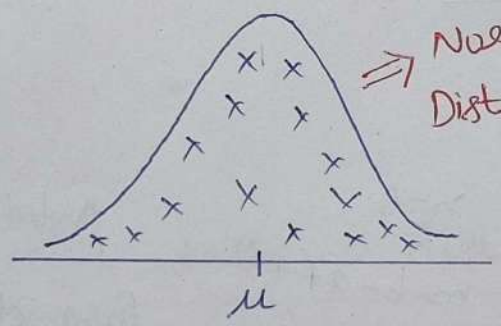
$S_3 = \{ \qquad \} = \bar{x}_3$

$\vdots$

$S_n = \{ \qquad \} = \bar{x}_n$

$\bar{X} = \{\bar{x}_1, \bar{x}_2, \bar{x}_3, \text{ ------ } \bar{x}_n\} \Rightarrow$ Sampling distribution mean

$\Downarrow$



$\Rightarrow$ Normally Distributed

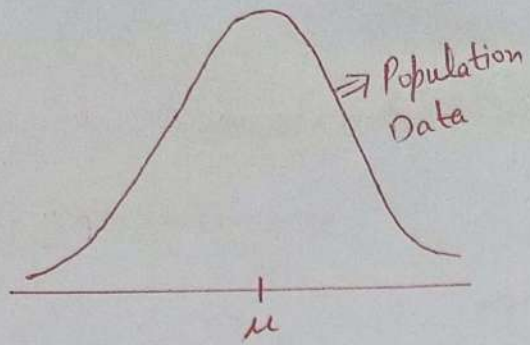
Most Important for Interview: If I take the sample mean, initially my population data had some mean and standard deviation but after taking the sampling mean will the mean and standard deviation change?

Ans mean will be the same but standard deviation will change.

$$\bar{X} \simeq N(\mu, \sigma/\sqrt{n})$$

Here, n can be any value.

CamScanner



$X \approx N(\mu, \sigma)$

σ = Population Std

μ = Population mean

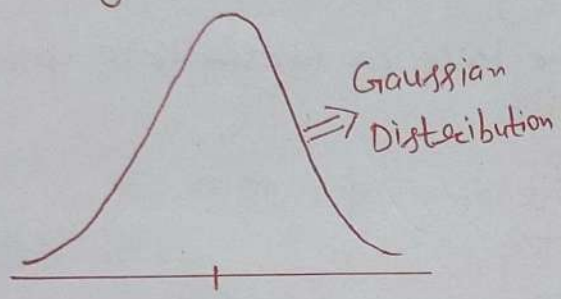
n = Sample size

$\Downarrow$

Sampling distribution of mean

$\bar{X} \approx N(\mu, \sigma/\sqrt{n})$

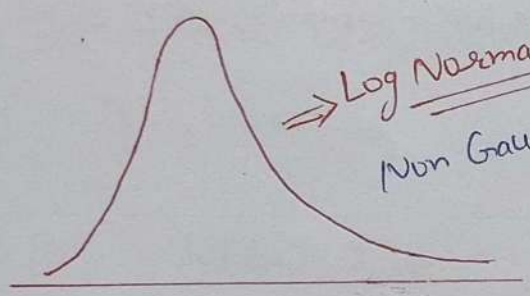
$\boxed{\frac{\sigma}{\sqrt{n}}}$ $\Downarrow$ Standard Error



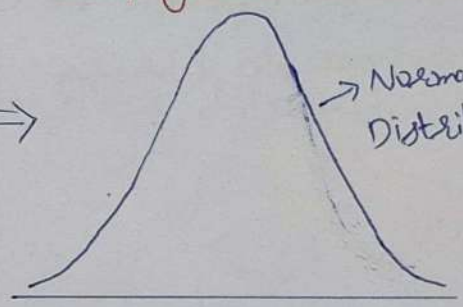
$\boxed{n \text{ Can be any value}}$

* $\boxed{n \geq 30}$

Sampling Distribution Mean



$\Rightarrow$



(*) How to check whether a distribution is Normal or Not (Q-Q Plot)

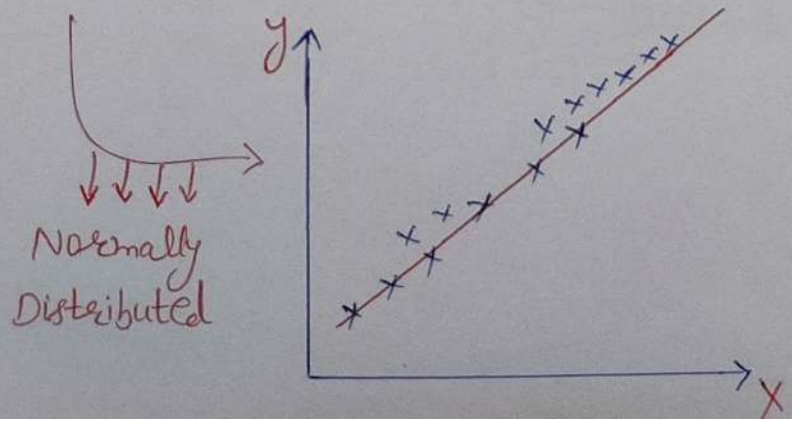
Q-Q Plot

This we will check it is normal distribution or not.

Normal Distribution

$X = \{ - - - - - - - \}$

$y = \{ - - - - - - - - - - \}$



when data is following as a straight line, then we say that the data is normally distributed.

Interview

① Which libraries you use for Statistics implementation?

Pandas - used for organize the data

numpy - used for mathematical calculation

Matplotlib and Seaborn - used for plotting

Statsmodel -

Scipy

Statistics - used for performing some basic or intermediate operations.

② How you implement the normal test?

③ How you know data is normal or not?

④ How you implement the Central limit Theorem?

⑤ What are the libraries for Probability distribution function etc.?

CamScanner

Agenda

① Chi Square Test

② ANNOVA Test

① Chi Square Test:

The chi square test for Goodness of fit claims about Population proportions [Categorical variables].

or

It is a non parametric test that is performed on categorical data. [nominal, ordinal].

eg: There is a population of male who likes different colors of bikes.

	Theory	Sample
Yellow Bike	1/3	22
Orange Bike	1/3	17
Red Bike	1/3	59
	⇓	↳ observed
	Theory Categorical Distribution	Categorical Distribution

Goodness of fit

(*) In a student class of 100 students, 30 are Right handed. Does this class fit the theory 12% of people are right handed.

	O	E
Right handed	30	12
Left handed	70	88

$\Rightarrow$ Theory Categorical Distribution

$\Downarrow$

Observed Categorical Distribution

Chi Square for Goodness of fit

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

< 50	50 – 75	> 75
140	160	200

C.I. = 95%

using $\alpha = 0.05$, would you conclude the population differences of weights has changed in the last 10 years?

Ans

2010 Expected

< 50 kg	50 – 75 kg	> 75
20%	30%	50%

observed n = 500 2020

< 50	50 – 75	> 75
140	160	200

Expected

< 50	50 – 75	> 75
100	150	250

$\frac{20}{100} \times 500$

$\frac{30}{100} \times 500$

$\frac{50}{100} \times 500$

① Null Hypothesis: H_0 : The data meets the expectation.

② Alternate Hypothesis: H_1 : The data does not meet the expectation.

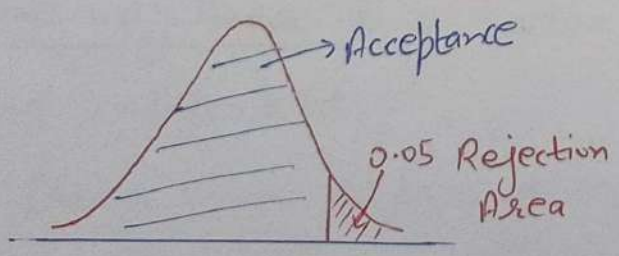
③ $\alpha = 0.05$ C.I. = 95%

④ Degree of freedom

$df = K - 1 = 3 - 1 = 2$

CHI Square Table is a 1 Tail test only

⑤ Decision Boundary → CHI SQUARE Table



CV = 5.991
(critical value)

w.r.t chi square test it is not a normal distribution it is a right skewed distribution.
before Z-test and t-test these all are normal distribution so you have 1 tail and 2 tail whenever you have right skewed it is just 1 Tail test

If χ^2 is greater than 5.991, Reject H_0
else
we fail to Reject H_0

⑤ Calculate Chi Square Test Statistics

$$\chi^2 = \sum_{i=1}^{k} \frac{(O-E)^2}{E}$$

observed $n=500$

2020

<50	50–75	>75
140	160	200

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

Expected

<50	50–75	>75
100	150	250

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

$$= 16 + 0.66 + 10$$

$$\boxed{\chi^2 = 26.66}$$

Assignment
⇓
$\boxed{\alpha = 0.25}$

$\boxed{26.66 > 5.991, \text{ Reject } H_0}$

CamScanner

② Analysis of Variance (ANOVA) (F-test)

Definition: ANOVA is a Statistical method used to compare the mean of 2 or more groups.

why? → if we have more than 2 groups at that time we use ANOVA test to compare the mean of 2 or more groups

ANOVA

① Factors (variable)

② Levels

eg: Factor = Medicines

Levels ⇒ 5mg, 10mg, 20mg [Dosage]

Mode of Payment → Factor

GPay, PhonePE, IMPS, NEFT [Levels]

Types of ANOVA:

① **One way Anova:** One factor with at least 2 levels, these levels are independent.

eg: Doctor want to test a new medication to decrease headache. They split the participants into 3 Conditions with respect to dosage [10mg, 20mg, 30mg]. Doctors ask the participant to rate the headache [1-10].

CamScanner

Medication → Factor

10 mg	20 mg	30 mg → Levels
5	7	2
9	8	7
–	–	–
–	–	–

② <u>Repeated Measures Anova</u>: One factor with atleast 2 levels, levels are dependent.

Running → Factor

Day 1	Day 2	Day 3 → Levels
8 km	5 km	6 km

③ <u>Factorial Anova</u>: Two or more factors (each of which with atleast 2 levels) levels can be either independent or dependent

Running → Factor

Independent ↓ | Day 1 | Day 2 | Day 3 ← Dependent

Gender		Day 1	Day 2	Day 3
	Male	8	5	6
		7	4	3
	Female	6	5	4
		3	2	1

CamScanner

Hypothesis Testing in ANOVA

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

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

F Test Statistics

$$F = \frac{\text{variation between Samples (variance)}}{\text{variation within Samples (variance)}}$$

variance between Sample

	X_1	X_2	X_3
variance within Sample	1	6	5
	2	7	6
	4	3	3
	5	2	2
	3	1	4
	$\bar{X}_1 = 3$	$\bar{X}_2 = 19/5$	$\bar{X}_3 = 4$

$H_0 = \bar{X}_1 = \bar{X}_2 = \bar{X}_3$

H_1: Atleast one Sample is not equal

Q One way ANOVA

One factor with atleast 2 levels, levels are independent.

① Doctors want to test a new medication which reduces 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 there any differences between the 3 Conditions using alpha = 0.05 ?

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

Ans

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

② Alternate " " H_1: Atleast one mean is not equal.

③ State Significance Value

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

④ Calculate Degree of freedom

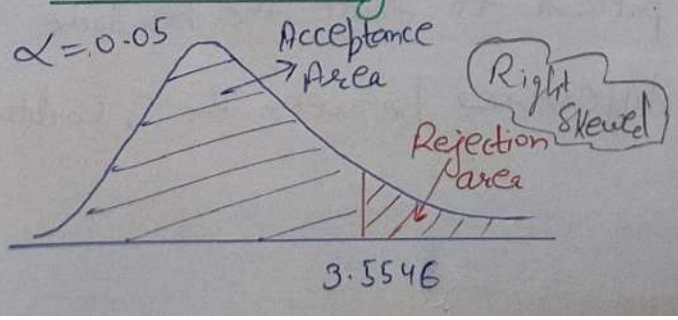
$N = 21 \quad a = 3 \quad n = 7$

$(2, 18) \Rightarrow$ F test table

$$\begin{cases} df\ between = a-1 = 2 \\ df\ within = N-a = 21-3 = 18 \end{cases}$$

$df\ total = N-1 = 21-1 = 20$

Decision Boundary



If F test is greater than 3.556, Reject the H_0.

⑤ Calculate F test statistics

	Sum of Square SS	df	mean Square MS	F
Between	98.67	2	49.34 (98.67/2)	86.56 (49.34/0.54)
Within	10.29	18	0.54 (10.29/18)	
Total	108.95	20		

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

① $SS_{between} = \frac{\Sigma(\Sigma a_i)^2}{n} - \frac{T^2}{N}$

$$\left.\begin{aligned} 15mg &= 9+8+7+8+8+9+8 = 57 \\ 30mg &= 7+6+6+7+8+7+6 = 47 \\ 45mg &= 4+3+2+3+4+3+2 = 21 \end{aligned}\right\} \Rightarrow T^2$$

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

$$= \boxed{98.67}$$

$\Sigma Y^2 = 9^2+8^2+7^2+8^2+8^2+9^2$
$8^2+7^2+6^2+6^2+7^2+8^2$
$+7^2+6^2+\ldots$
$= 853$

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

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

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

$$= \boxed{10.29}$$

CamScanner

③ $SS_{total} = \sum y^2 - \frac{T}{N}$

$$= 853 - \frac{125^2}{21} = \boxed{108.95}$$

$$F = \frac{\text{Variation between Samples}}{\text{variation within Samples}} = \frac{\text{MS between}}{\text{MS Within}}$$

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

$86.56 > 3.556$, Reject the Null Hypothesis.

Interview

① What is F test used for?

② Why F test is used? where (ANOVA)

③ Why chi square is used? What are the practical applications? where it is exactly used in data science.

(Chi square test also used in feature selection)

CamScanner