STATS

Statistics and its applications –

Statistics is a field that deals with collection, organization analysis, interpretation and presentation of the data.

Applications –

- Machine learning
- Data analyst
- BA
- risk analyst
- house wife
- Each & every domain

Types of statistics –

Statistics

Descriptive

It consists of organizing & summarizing of data

1. Measure of central tendency
   (i) Mean
   (ii) Median
   (iii) Mode
2. Measure of Dispersion
   (i) Variance
   (ii) Standard deviation

Inferential

Collect data → sample data

↓ {using some experiment}

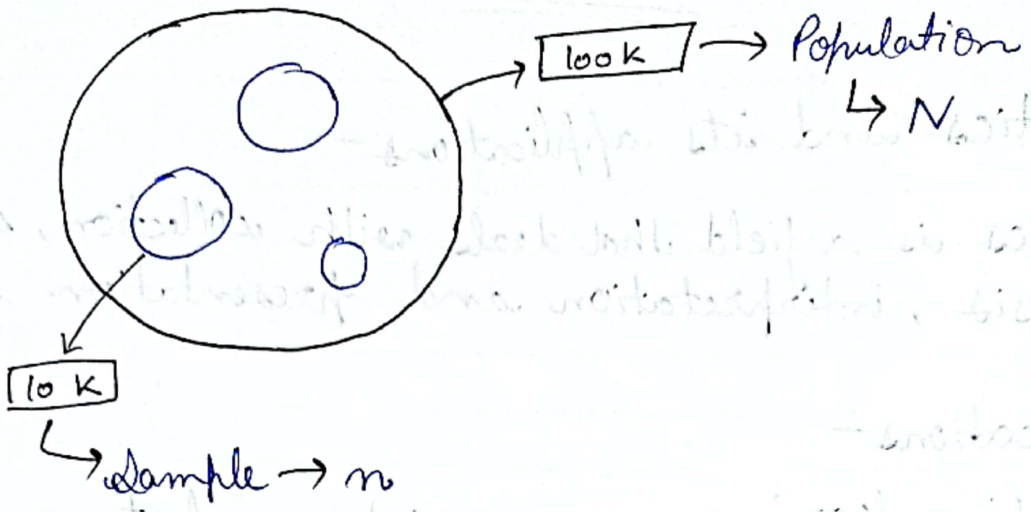
conclusions or Inferences → Ztest, Ttest, ..

↓

other data

↓

population data



Measure of central tendency -

① Mean
② Median
③ Mode

Mean -

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

Ages = {1, 3, 4, 5}

$$\mu = \frac{1+3+4+5}{4} = 3.25$$

Median -

Ages = {1, 3, 4, 5, 100} → outlier

$$\mu = \frac{113}{5} = 22.6$$

Ages = {4, 3, 1, 5, 100}

→ sort the numbers

Ages = {1, 3, 4, 5, 100}

* Instead of finding total averge pick up the central element.

Median - 4

Ages = {1, 3, 4, 5, 100, 200}

↳ Medians → $\frac{4+5}{2} = 4.5$

Mode - {4, 3, 2, 1, 1, 4, 4, 5, 2, 100}

* Select the element having maximum frequency.

Mode = 4

Measure of Dispersion

1. Variance
2. Standard deviation

Ages 1 = {2, 2, 4, 4}

$\mu = \frac{2+2+4+4}{4}$

$= 3$

Ages 2 = {1, 1, 5, 5} → {Spread is more or dispersion is more}

$\mu = \frac{1+1+5+5}{4}$

$= 3$

Variance -

For population data {N}

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

For sample data {n}

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

Eg- Using Ages 1 & Ages 2

x_i	μ	$(x_i-\mu)^2$
2	3	1
2	3	1
4	3	1
4	3	1

$4/4 = 1$

x_i	μ	$(x_i-\mu)^2$
1	3	4
1	3	4
5	3	4
5	3	4

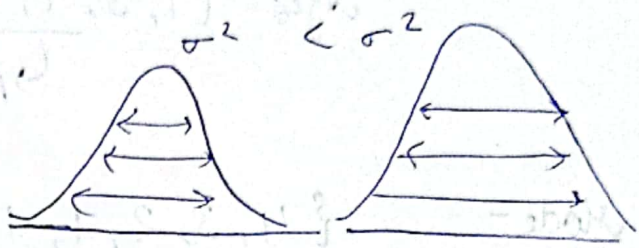
more deviation → $16/4 = 4$

Standard deviation -

variance ↑↑ spread ↑↑

$\sqrt{\sigma^2}$ → how far a data pt. is away from mean

population standard deviation



Variable - a variable is a property that can take up any value.

Age = 25 → variable

Ages = [11, 20, 25, 30] → not a variable

Variable

- Quantitative variable
  - Discrete
    - whole no.
    - Eg - Age
  - Continuous
    - can be decimal
    - Eg - Weight
- Qualitative / categorical
  - Eg - Gender < male / female
  - colors
  - [not is numerical values]

Random variables -

$y = 5x + 2$ → here x is a random variable which can take diff values.

↳ x = 1
x = 2
x = 3

X → is a → function → values ← are derived from ← process & experiments

Tossing a coin

$$X = \begin{cases} 0 & - \text{ if head} \\ 1 & - \text{ if tail} \end{cases}$$

Random variables

- Discrete
  - Eg - tossing a coin, rolling a dice
- Continuous
  - tomorrow how many inches it is going to rain

Percentiles & Quartiles -

① Percentage -

How many % of number are odd - {2, 4, 7, 5, 8, 9, 10}

$$\text{Percentage} = \frac{\#\text{ No.'s are odd}}{\#\text{ of obv.}} \times 100 = \frac{3}{7} \times 100$$

② Percentile - A percentile is a value below which a certain percentage of obv.'s lie.

Eg - 2, 2, 3, 4, 5, 5, 5, 6, 7, 8, 8, 8, 8, 9, 9, 10, 11, 11, 12

* What is the percentile of ranking '10'.

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

$$= \frac{16}{20} \times 100 = 80\%$$

* What value exists at the percentile ranking of 25% ??

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

$$= \frac{25}{100} \times (20+1) = 5.25$$

↳ Index

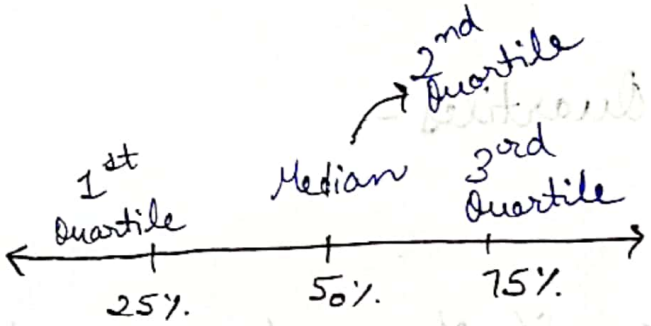
5.25 index indication –

$$\text{Avg}(5+6) = \frac{5+5}{2} = 5$$

↳ 25%.

↳ indexes

(3) Quartiles –



Five number summary –

1. Minimum
2. First Quartile
3. Median
4. Third Quartile
5. Maximum

Dataset -

{Remove the outlier}

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

[Lower fence (-3) → Higher fence (13)]

Lower fence = Q1 - 1.5 (IQR)

Higher fence = Q3 + 1.5 (IQR)

Inter Quartile range (IQR) = Q3 - Q1 = 7 - 3 = 4

$$Q1\ (25\%) = \frac{\text{percentile}}{100} \times (n+1) = \frac{25}{100}(20)$$

$$= 5^{th} \text{ index} = 3$$

$$Q3\ (75\%) = \frac{75}{100} \times 20 = 15^{th} \text{ index} = 7$$

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

Higher fence = 7 + 1.5(4) = 13

∴ 27 is an outlier

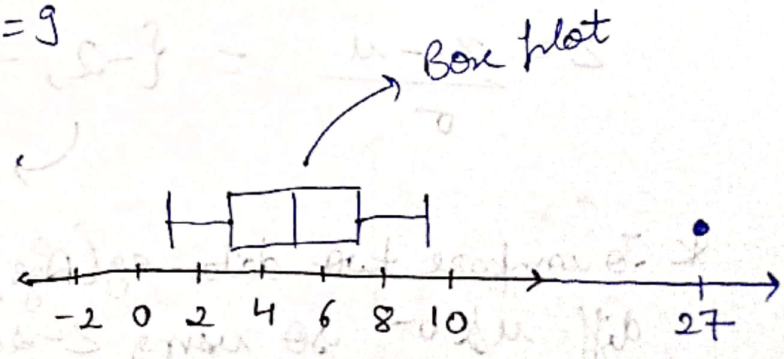
① Minimum = 1

② Q1 = 3

③ Median = 5

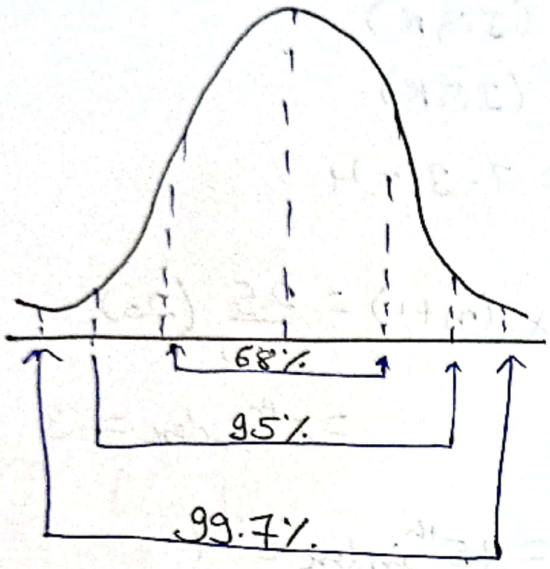
④ Q3 = 7

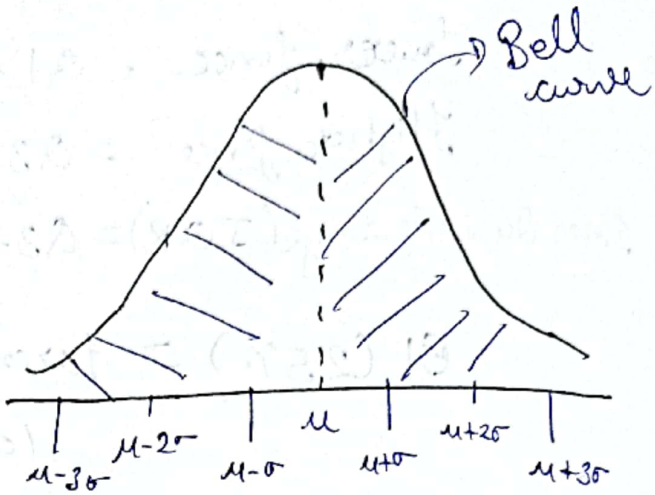
⑤ Maximum = 9



Normal Distribution / Gaussian distribution

Emperical formula
[68-95-99.7]





Standard normal distribution

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

$$\Downarrow$$

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

Z-score

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

∵ actually its: $\frac{\sigma}{\sqrt{n}}$
i.e. standard error
but as we apply this formula to each value so $n = 1 \Rightarrow$ it becomes only "σ".

Eg - 1, 2, 3, 4, 5

$\mu = 3, \sigma = 1$

$$Z = \frac{x_i - \mu}{\sigma} = \{-2, -1, 0, 1, 2\}$$

$$\hookrightarrow SND(\mu = 0, \sigma = 1)$$

* To compare two data eg(Age, Weight) both will have diff. μ & σ so using Z-score we find their SND.

Application of Z-score -

Z-score → these are standardized values that can be used to compare scores in diff. distribution.

Eg → Cricket - 2020 2021

India VS Australia test match

2020 (Stats - 3 test)

Test average = 181

standard avg. = 12

Rishabh finalscore = 187

2021 (Stats)

Rishabh test avg. = 182

standard deviation = 5

final score = 185

Compared to the rest of the test scores, in which year was Rishabh's score in final game better?

Ans: $Z = \frac{x-\mu}{\sigma}$

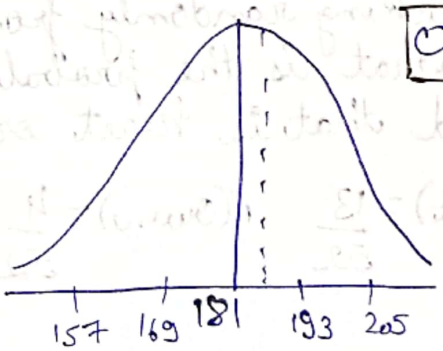
$= \frac{187-181}{12} = 0.5$

$Z = \frac{185-182}{5}$

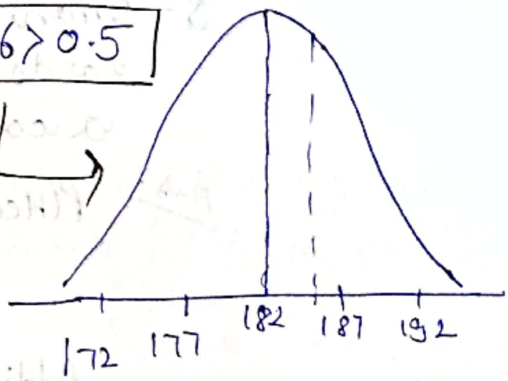
$= 0.6$

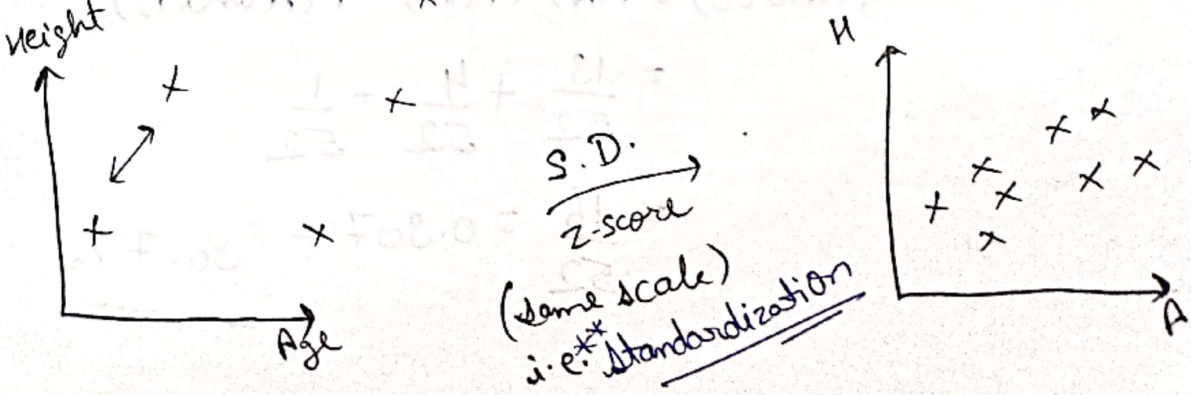
0.6 > 0.5

2020



2021





Log normal distribution -

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

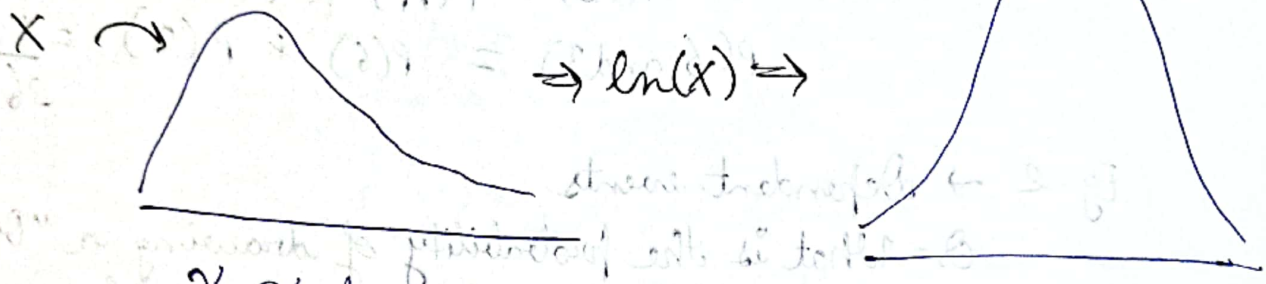
$\Downarrow$

$X \approx$ log normal distribution

$\Downarrow$

Then $y = \ln(X)$ have a normal distribution

$\ln = \log_e$



$X \approx$ log normal distribution

$y = \ln(X)$

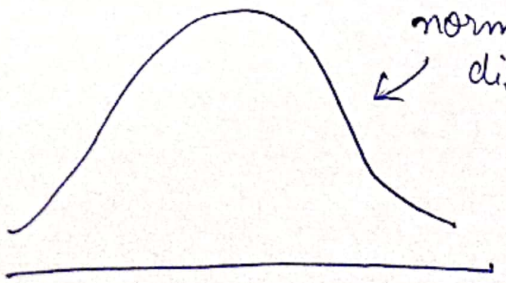
$X = \exp(Y)$

Central limit theorem -

The CLT states that regardless of the shape of population distributions the distribution of sample mean will be approximately normal.

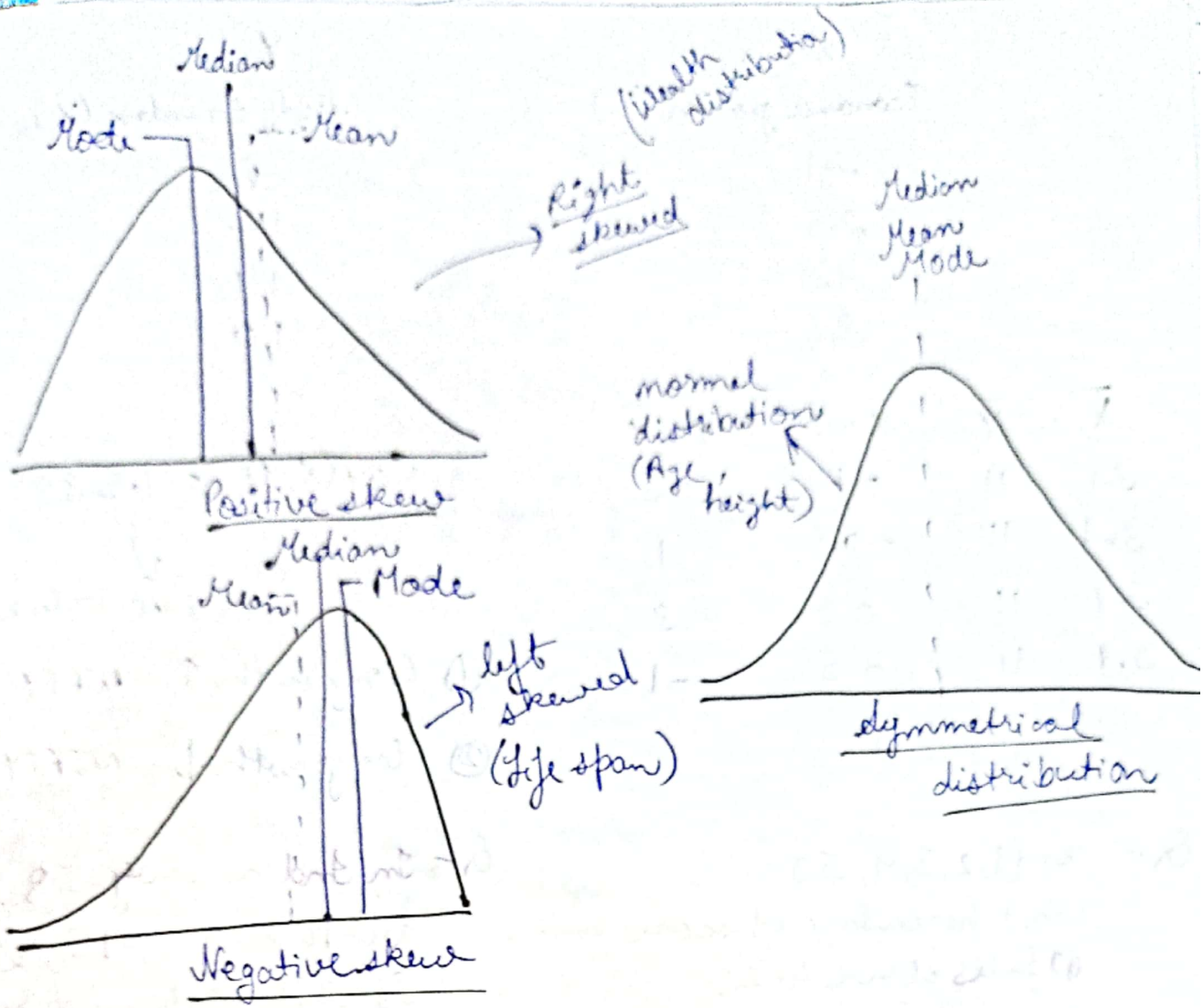
$\bar{S_1}, \bar{S_2}, \bar{S_3}, - - - , \bar{S_n}$

↳ All these will be ↘ normally distributed ↙



① The distribution of sample means will become more normal as its sample size increases.

② *RULE - sample distri. will be approximately normal if their sample size is $n \geq 30$



COVARIANCE - Quantify the relationship between X & Y.

+ve → { X↑ Y↑ , X↓ Y↓ } { X↑ Y↓ , X↓ Y↑ } ← -ve

population -

$$Cov(X,Y) = \sum_{i=1}^{N} \frac{(x-\bar{x})(y-\bar{y})}{N}$$

sample -

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

$$\boxed{Cov(x,x) = Var(x)}$$

X Economic growth (%)	Y Nifty 50 index (%)
2.1	8
2.5	12
4.0	14
3.6	10

$\bar{X}$	$\bar{Y}$	$x-\bar{x}$	$y-\bar{y}$
3.1	11	-1	-3
3.1	11	-0.6	1
3.1	11	0.9	3
3.1	11	0.5	-1

$Cov(X,Y) = 1.533$

↓

+ve value

① Eco. growth ↑ NIFTY ↑

② Eco growth ↓ NIFTY ↓

Q- $x = \{1, 2, 3, 4, 5\}$
What percentage of score
① falls above 4.25
② below 3.75
given $\mu = 4$, $\sigma = 1$

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

$= \frac{4.25 - 4}{1}$

$= 0.25$

② $Z_{score} = \frac{3.075 - 4}{1}$

$= -0.25$

$= 40\%$

Q- In India avg. IQ is 100 with a $\sigma = 15$, what is the % of population would you expect to have an IQ- ① lower than 85
② higher than 85
③ blw 85 & 100

Normalization – [lower scale ⟷ higher scale] → [0-1]

① Min-Max scaler [0-1]

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

x	→	y
1		0
2		0.25
3		0.5
4		0.75
5		1

we apply this in –
Deep learning & image processing
i.e. convert pixels (0-255) to (0-1)

* Normalization & standardization are a part of Feature scaling

mostly used in Deep learning

mostly used in Machine learning

Sampling techniques –

① Simple random sampling –
Every number of the population (N) has an equal chance of being selected for your sample (n)

② Stratified sampling –
Data is choosed according to the categories.

Gender → Male, Female

Education degree → High school, Masters, Phd

③ Systematic sampling –
Select every n^{th} individual out from the population (N)

↳ Convenience sampling –
Only those who are interested in the survey will only participate.

⊛ Pearson correlation coefficient - (-1 to 1)

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

→ most common way of measuring a linear correlation.

Between 0 to 1 → +ve correlation - when one variable changes, the other variable changes in the 'same direction'.

0 → No correlation - there is 'no relationship' b/w the variables.

Between 0 & -1 → -ve correlation - when one variable changes, the other variable changes in the 'opposite direction'.

⊛ Spearman Rank correlation -

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

→ when the variables are normally distributed or the relationship b/w the variables is not linear.

x	y
10	4
8	6
7	8
6	10

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

→ Ascending Order

PROBABILITY

$$\text{Probability} = \frac{\text{No. of ways can event can occur}}{\text{No. of possible outcomes}}$$

Mutually exclusive - if events can not occur at same time.

Non-mutually exclusive - if events can occur at same time.

Eg-1 → Mutually exclusive -

Q- Probability of getting heads or tails on flipping a coin?

Ans- $P(H) = \frac{1}{2} = 0.5$ $P(T) = \frac{1}{2} = 0.5$

Addition rule for mutually exclusive

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

i.e. $P(H \text{ or } T) = 0.5 + 0.5$

$= 1$

Eg-2 → Non-mutually exclusive event -

Q- When picking randomly from a deck of cards, what is the probability of choosing a card that is heart or a Queen?

Ans: $P(\text{Heart}) = \frac{13}{52}$ $P(\text{Queen}) = \frac{4}{52}$ $P(Q \text{ or } H) = \frac{1}{52}$

Addition rule for non-mutually exclusive

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

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

$$= \frac{16}{52} = 0.307 = 30.7\%$$

Independent events – two events are independent if they do not affect one another.

Dependent events – two events are dependent if they affect one another.

Eg-1 → Independent events

Q – What is the probability of rolling 6 & then a 3, with a normal 6 sided dice.

Ans – Multiplication rule for independent events

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

$$P(6 \text{ and } 3) = P(6) * P(3) = \frac{1}{36}$$

Eg-2 → Dependent events

Q – What is the probability of drawing a "Queen" and then a "King" from a deck of cards?

Ans – Multiplication rule for dependent events

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

$$P(Q \text{ and } K) = P(Q) * P(K|Q)$$

$$= \frac{4}{52} * \frac{4}{51} = .006 = 0.6\%$$

PERMUTATION –

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

n = total # of obj.
r = # of obj. you are picking

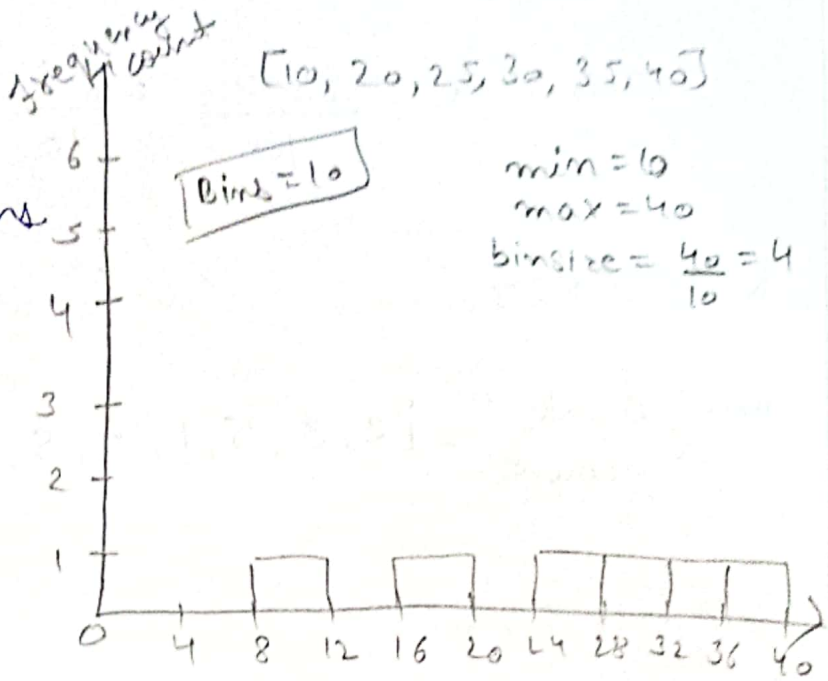
COMBINATION –

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

Histogram –

Ages = {0, 10, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 41, 42, 43, 50, 51, 65, 68, 78, 90, 95, 100}

1. Sort the numbers
2. Bins → no. of groups
3. Bin size → size of bins

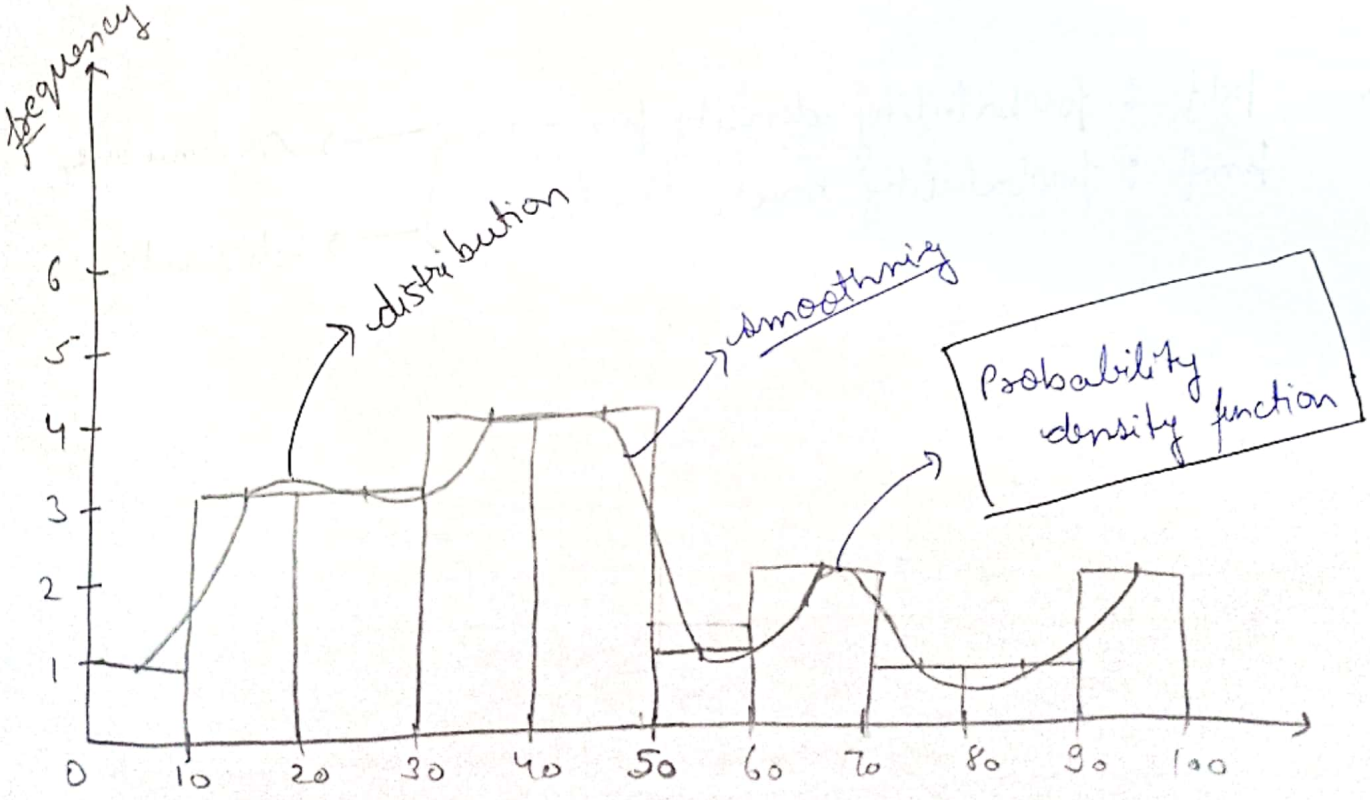


$$\text{Binsize} = \frac{\text{Max} - 0}{\text{bins}}$$

For ages-dataset –

bins = 20 $\quad$ binsize = $\frac{100}{20} = 10$

* CONTINUOUS



Weight = [30, 35, 38, 42, 46, 58, 59, 62, 63, 68, 75, 77
80, 90, 95]

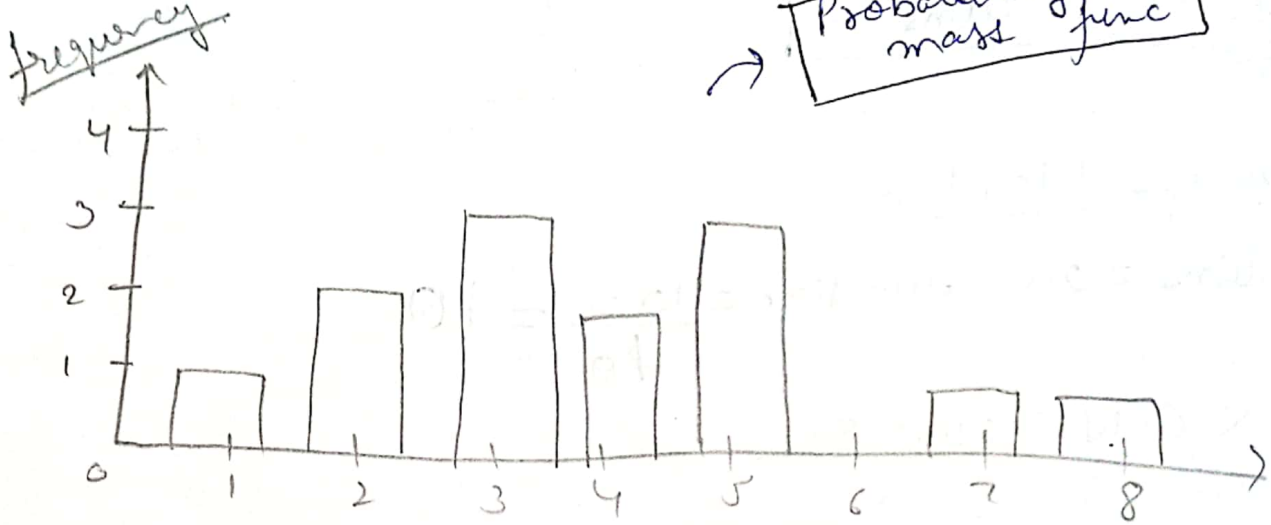
bins = 10

∴ bin size = $\frac{95-30}{10} = \frac{65}{10} = 6.5$ → continuous value

* DISCRETE —

No. of Bank accounts = [2, 3, 5, 1, 4, 5, 3, 7, 8, 3, 2, 4, 5]

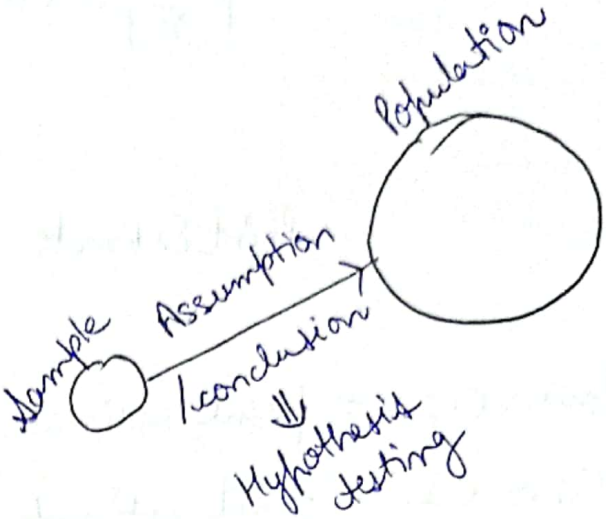
→ Probability mass func~



* pdf : probability density function → continuous
pmf : probability mass function → discrete

Inferential statistics -

- Hypothesis testing
- p-value
- Confidence interval
- Significance value



* Hypothesis testing

Steps of hypothesis testing -

① Null hypothesis (H_0)

② Alternate hypothesis (H_1)

③ Perform experiments

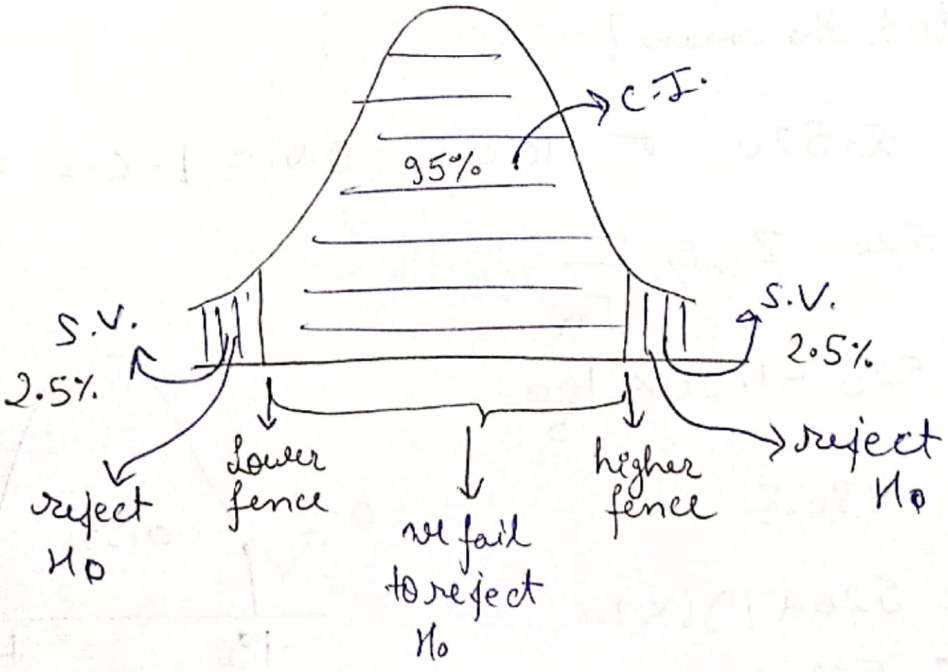
↳ we fail to reject the null hypothesis [within C.I.]

↳ we reject the null hypothesis [outside C.I.]

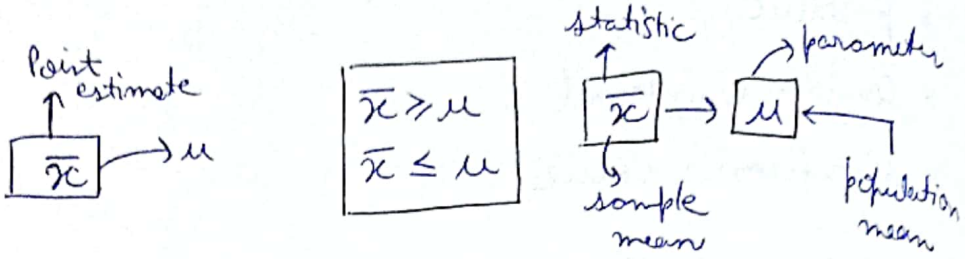
C.I.
[confidence interval]

S.V.
[Significance value]

S.V. = 1 − C.I.



Point estimate – the value of any statistic that estimates the value of a parameter is called point estimate.



Point Estimate ± Margin of Error = Parameter → Population mean

Lower C.I. = point estimate – Margin of Error

Higher C.I. = point estimate – Margin of Error

Margin of error = $Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ → σ = population Sd; $\sqrt{n}$ ⟹ standard error

α = significance value

Q- On the quant test of CAT Exam, a sample of 25 test takers has a mean of 520 with a population S.D. of 100. Construct a 95% C.I. about the mean?

$n = 25 \quad \bar{x} = 520 \quad \sigma = 100 \quad$ S.V. = 1 – C.I. = 0.05

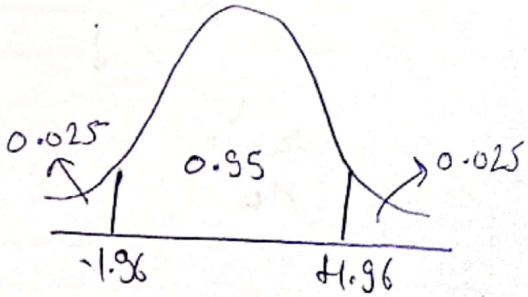
Lower C.I. $= 520 - Z_{0.05/2} \frac{\sigma}{\sqrt{n}}$

$= 520 - 1.96 \times \frac{100}{5}$

$= 480.8$

Higher C.I. $= 520 + 1.96 \times 20$

$= 559.2$



* Population S.D – Z test + $n \geq 30$
Sample S.D – T test + $n < 30$

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

$\bar{x} = 520$ $S = 80$ $n = 25$ C.I = 95% S.V. = 0.05

Lower C.I. $= 520 - t_{\frac{0.05}{2}} \left(\frac{80}{\sqrt{25}}\right)$

$= 520 - 2.064 \times 16$

$= 486.976$

Higher C.I. $= 553.024$

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

Q- A factory has a machine that fills 80 ml of baby medicines in a bottle. An employee the average amount of baby medicine is not 80 ml. Using 40 samples, he measures the average amount dispersed by the machine to be 79 ml with a standard deviation of 2.5.

a) State Null & Alternate hypothesis

b) At 95% CI, is there enough evidences to support machine is working properly or not.

Ans. Step 1 -

Null Hypothesis $\mu = 80$

Alternate hypothesis $\mu \neq 80$

Step 2 -

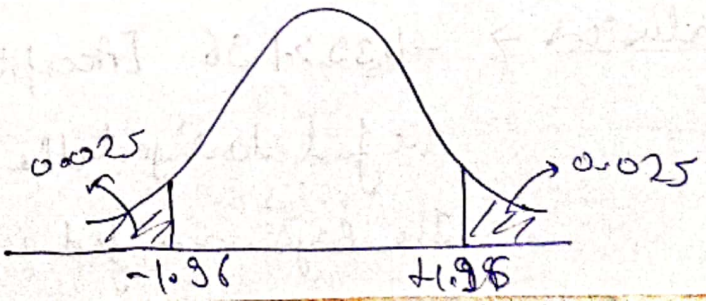
C.I. = 0.95 S.V. $(\alpha) = 1 - 0.95 = 0.05$

Step 3 -

$n = 40$ $S = 2.5$

Z test

Decision boundary



Calculate test statistics -

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

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.

8- A complain was registered, the boys in a Govt. school are underfed. Avg. weight of the boys of age 10 is 32 kgs with S.D. = 9 kgs. A sample of 25 boys were selected from the govt. school and the avg. weight was fund to be 29.5 kgs? with CI = 95%, check it is true or false.

A-8: $n = 25 \quad \mu = 32 \quad \sigma = 9 \quad \bar{x} = 29.5$

Step-1 -

$H_0 : \mu = 32$

$H_1 : \mu \neq 32$

Step 2 - C.I. $= 0.95 \quad \alpha = 1 - 0.95 = 0.05$

Step 3 -

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

Conclusion → $-1.39 > -1.96$ [Accept the Null Hypothesis 95% C.I.]

We fail to reject the Null Hypothesis

The Boys are fed well.

Q- A factory manufactures cars with a warranty of 5 years or more on the engine and transmission. An engineer believes that the engine or transmission will malfunction in less than 5 years. He tests a sample of 40 cars and finds the average time to be 4.8 years with a standard deviation of 0.50.

① State the null & alternate hypothesis.

② At a 2% significance level, is there enough evidence to support the idea that warranty should be revised?

Ans $n = 40$ $\bar{x} = 4.8$ years $S = 0.50$

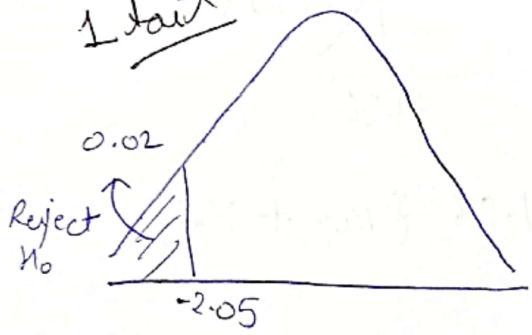
Step 1 -

$H_0 : \mu \geq 5$ {Null hypothesis}

$H_1 : \mu < 5$ {Alternate hypothesis}

Step 2 - $\alpha = 0.02$ C.I. $= 1 - 0.02 = 0.98 = 98\%$

Step 3 - 1 tail



Step-4 - $Z_{score} = \frac{\bar{x} - \mu}{S/\sqrt{n}}$

$Z_{score} = \frac{4.8 - 5}{\frac{0.50}{\sqrt{40}}}$

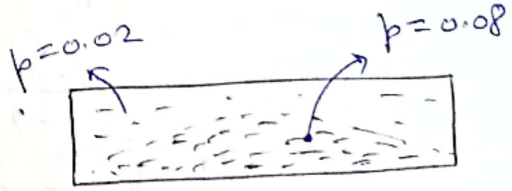
$Z_{score} = -2.52$

Step-5 - Conclusion : $-2.52 < -2.05$ – Reject H_0

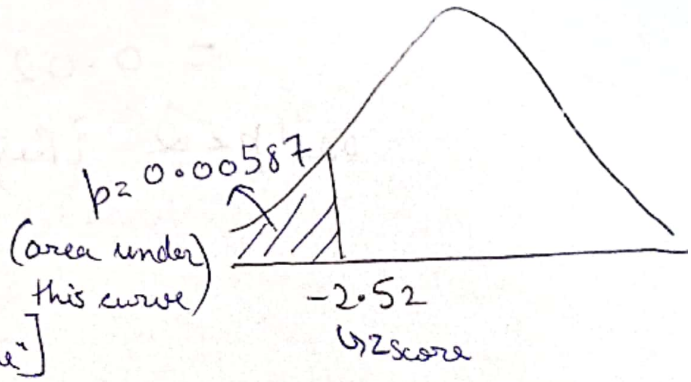
⇒ Warranty needs to be revised.

P-value

p-value ≠ significance value
↳ derived by C.I.
↳ should always be less than or equal to s.v.



As pvalue < α (0.02)
so we reject H_0
[the area under the curve we get through Zscore is "pvalue"]



Q- The avg. weight of all residents in a town xyz is 168 pounds. A nutritionist believes that 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 there enough evidence to discard the H_0?

A.S

$\bar{x} = 169.5 \quad s = 3.9 \quad n = 36 \quad \mu = 168$

C.I. = 0.95

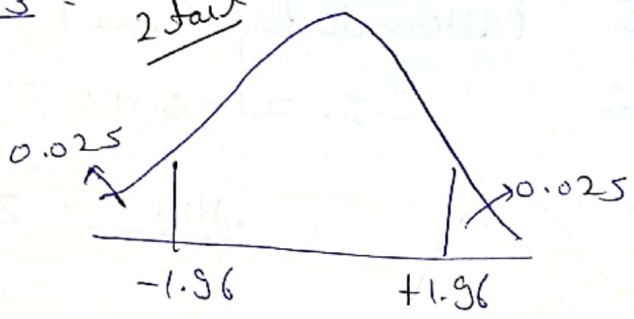
Step-1 - $H_0 \quad \mu = 168$

$H_1 \quad \mu \neq 168$

Step-2 - C.I. = 0.95

$\alpha = 0.05$

Step-3 - 2 tail

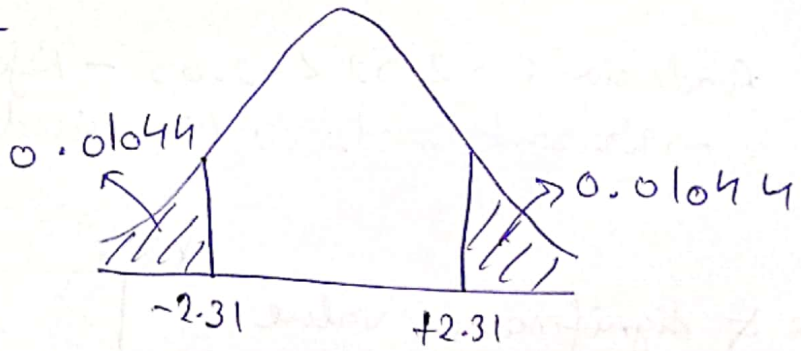


Step-4

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

Step-5 - $2.31 > 1.96$ {Reject H_0}

P-value -



pvalue = 0.01044 + 0.01044

= 0.02088

as $p < \alpha$ {Reject H_0}

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

a) State the Null and alternate hypothesis

b) At a 99% of CI, is there enough evidence to discard the H_0?

Ans: ① $H_0 : \mu \geq 2$

$H_1 : \mu < 2$

② C.I = 0.99 $\quad n = 10$

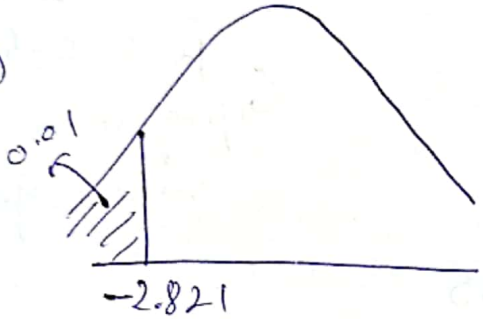
S.V. = 0.01

as $n < 30$

degree of freedom = $n - 1$

$= 10 - 1 = 9$

③



④ $t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{1.8 - 2}{0.15/\sqrt{10}}$

$= -4.216$

⑤ $-4.216 < -2.82$

{Reject H_0}

⇒ The avg. life span of battery is less than 2 years

* Z test with proportions -

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

p_0 – population proportion

$\hat{p}$ – sample proportion

$q_0 = 1 - p_0$

$\hat{p} = \frac{x}{n}$ ↳ proportion

pvalue > α {Fail to reject H_0}

else {accept H_0}

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 diff. He conducts a survey of 200 individuals and found that 130 responded Yes. Owning a cell phone?

① State H_0 & H_1.

② At a 95% CI, is there enough evidence to reject H_0?

Ans: Step-1

Null hypothesis $\Rightarrow P_0 = 0.70$

Alternate hypothesis $\Rightarrow P_0 \neq 0.70$

$q_0 = 1 - P_0 = 0.30$

$n = 200 \quad x = 130$

$\hat{p} = \frac{130}{200}$

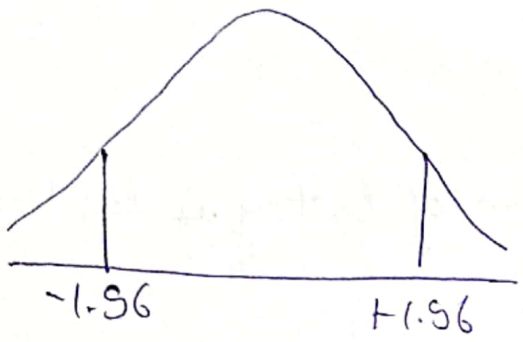
$\hat{p} = 0.65$

↳ proportion of saying "yes"

Step-2

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

Step-3



Step-4

$$Z_{test} = \frac{0.65 - .70}{\sqrt{\frac{0.7 \times 0.3}{200}}}$$

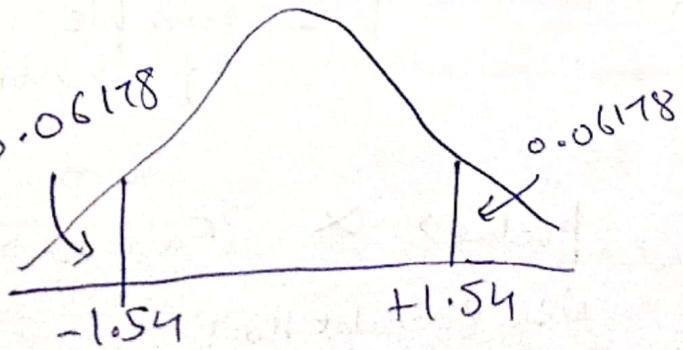
$\simeq -1.54$

Step-5

Conclusion: $-1.54 > -1.96$

{Fail to reject the null hypothesis}

p-Value



pvalue = 0.12356

Q- A car company believes that the % 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 & alternate hypothesis

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

As:

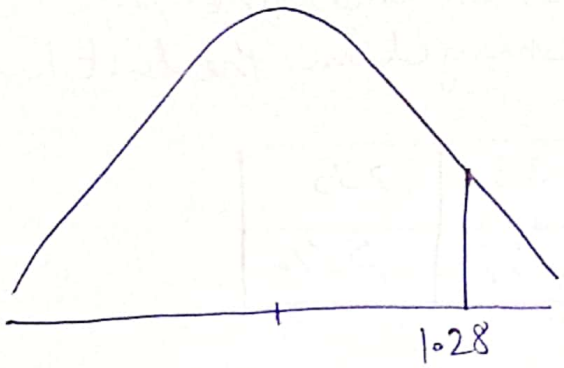
$H_0 : P_0 \leq 0.60$

$H_1 : P_0 > 0.60$

$q_0 = 0.40$

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

$$Z_{score} = \frac{0.68 - 0.60}{\sqrt{\frac{0.6 \times 0.4}{0.250}}} = \frac{0.08}{0.0309} = 2.588$$



Reject the Null hypothesis

* Chi-square test -

→ Chi-square test claims about population proportion. It is a non-parametric that test that is performed on categorical data.

↳ ordinal data / nominal data

$$\chi^2 = \sum \frac{(fo-fe)^2}{fe}$$

Q - In the 2000 U.S. census the age of individuals in a small town found to be following -

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

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

<18	18-35	>35
121	288	91

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

Ans.

	<18	18-35	>35
Expected →	20%	30%	50%

$n=500$

	<18	18-35	>35
Obs. →	121	288	91
Exp. →	100	150	250

Step-1 → Null hypothesis - H_0 : The data meets the exp. distribution

H_1 : The data does not meet the expected distribution

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

Step-3 : degree of freedom {categories}

$df = c - 1 = 3 - 1 = 2$

Step-4 : decision boundary = 5.991

{chi-square-table}

Step-5 : Test statistics -

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

$$\chi^2 = 232.494$$

Step-6 : Conclusion -

$\chi^2 > 5.99$ {Reject H_0}

* Anova test - (f-test)

→ The t-test works well when dealing with two groups, but sometime we want to compare more than two groups at the same time, so here we use f-test.

→ Extention of t-test & z-test.

$$F = \frac{\text{measure of effect (MSE effect)}}{\text{measure of error (MSE error)}}$$

$F > 1$, variation due to effect > due to error

$F = 1$, variation due to effect = due to error

$F < 1$, variation due to effect < due to error

* Bernoulli Distribution -

→ Used when we want to model the outcome of a single trail of an event.

Eg - Tossing of coin (1 time) - $S = \{H, T\}$

p - probability of success q - probability of failure

For getting head -

$P(H) = p = 0.5$

For not head -

$P(\text{not } H) = q = 0.5$

$\therefore$ $p = 1 - q$

$q = 1 - p$

* Binomial Distribution -

→ Used when we want to model the outcome of multiple trails of an event.

Eg - Tossing a coin (twice)

$S = \{HH, HT, TH, TT\}$

For X = no. of heads

$P(X=0) = 1/4 = 0.25$

$P(X=1) = P(HT) = 1/4 + 1/4 = 0.50$

$P(X=2) = P(HH) = 1/4 = 0.25$

* Similarly if tossed 3-times

$P(X=0) = 1/8$

$P(X=1) = 3/8$

$P(X=2) = 3/8$

$P(X=3) = 1/8$

$$P(x) = \frac{n!}{(n-x)!\,x!}\, p^x\, q^{n-x}$$

n = no. of trials

x = no. of success

p = prob. of success

$q = 1 - p$ = prob. of failure

* POWER LAW / PARETO PRINCIPLE

→ b/w two quantities

→ 80-20 rule

→ i.e. 80% of consequences come from 20% of causes