Math.

What is statistic?

Statistic is the science of collection, organizing and analyzing data.

Data: facts of information that can be measured.

Type of Stats:

1> DESCRIPTIVE STATS: It consists of organizing and summarizing data.

2> Infential Stats:
~~other~~ Technian where we used the data that we have measured to form Conclusion.

Population = N Sample = n

Sampling Techn.

1> Simple Random Sampling: ~~Each~~ Every member of Population (N) has an equal chance of being selected to your Sample (n).

2> Stratified Sampling: where the population (N) is split into non-overlapping Groups (Strats)

Gender → male
→ Female

Server

3> Systematic Sampling:

from Population N we take only n^{th} individuals

Eg: If I am out side of the mall and try to survey covid

I took 8th / 7th person of the mall.

4> Convenience Sampling:

Survey related to specific topic.

Eg: A survey on Govt Job people.

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

1> Quantitative Variable: Age, weight, Height, measured Numerically (can +, -, x, ...)

~~2> Quantitative~~

2> Qualitative / Categorical variable:

↳ Gender → M, F — { Based on some characteristic we can derive categorical variable }

Quantitative

↓ Discrete Variable — whole number No. of children in a family. Eg: 1, 2, 3, ...

↓ Continuous Variable — Height = 172.5, 162. Weight = 100k

What kind of variable.

Gender ⇒ Catagorical Marital status ⇒ Catagorical

River length ⇒ Continous Population of state ⇒ Discrete

Song length ⇒ "

Blood Pressure ⇒ "

Variable Measurement Scale :

① Nominal : Catagorical data.

② Ordinal : order of the data matter but value doesn't.

Student marks ⇒	100	90	80	10
Rank ⇒	1st	2nd	3rd	4th

③ Interval : Order matters, value matters, Zero is not fix.

Age line ⇒ (1 - 10) (10 - 90) (90 - 100)

④ Ratio : Naturally numerical, order data with absolute zero., zero matters

length, ~~width~~, weight, Temp

Frequency :-

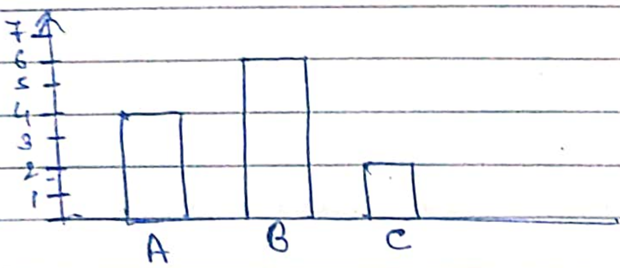
Data Set : A, B, C, A, A, B, C, A, B, B, B, B

Alphabet	Frequency Distribution	Commulative Frequency
A	4	4
B	6	10
C	2	12

Frequency Distribution → Data base-এ কাজে লাগে Some real কাজে

Commulative Frequency → আগেরটার সাথে যোগ হবে 4,

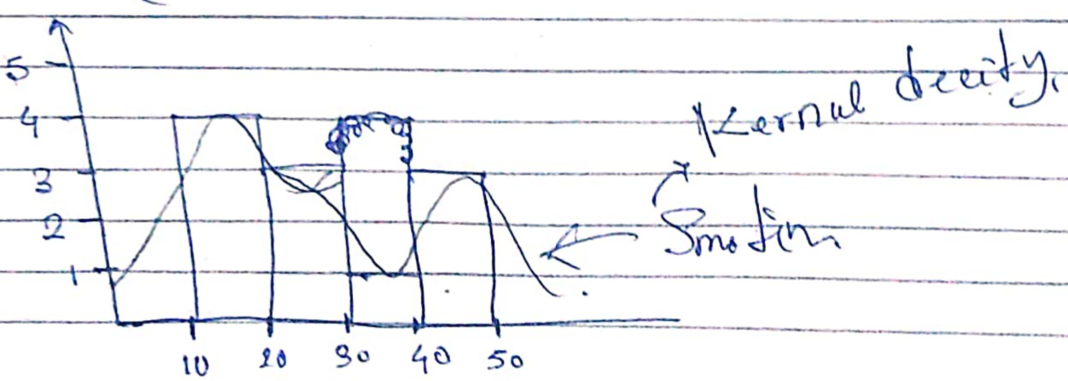
① BAR Graph



② Histogram:

Age = {10, 12, 13, 20, 25, 26, 31, 40, 44, 45, 50,}

Bins = 10 (এটা যা খুশি হতে পারে)



Arithmetic Mean for population & Sample.

MEAN: (Average) μ

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

Population (N)

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

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

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

$$= 3.2$$

Sample (n)

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

$$= 3.2$$

▷ Central Tendency:

Refers to the mesure used to determine the centre of the distribution of data

(i) Mean (ii) Median (iii) Mode.

If $X = \{1, 2, 2, 2, 3, 4, 5, 5, 5, 100\}$

$$\text{mean}(\mu) = \sum_{i=1}^{N} \frac{x_i}{N} = \frac{1+1+2+2+3+3+4+5+5+6+100}{11}$$

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

Median:

i) Short the number = {1, 1, 2, 2, 3, 3, 4, 5, 5, 6, 100} → 11 no

ii) find out the median number: 3

but calculation → {1, 1, 2, 2, 3, 3, 4, 5, 5, 6, 100, 111}

$$\text{median} \left(\frac{3+4}{2}\right) = 3.5$$

Mode : most frequent element

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

2 3

Mode = 6

* MEASURE AND DISPERSION

(i) Variance

(ii) Standard Deviation

~~Variance~~ :

Only we use this

If {1, 1, 2, 3, 4} $\mu = \frac{10}{5} = 2$

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

both mean is same. But distribution are diff, so we use those

(i) Variance :

Population Variance (σ)	Sample variance (s)
$\sigma^2 = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N}$	$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$

Variance মান নির্নয়ে

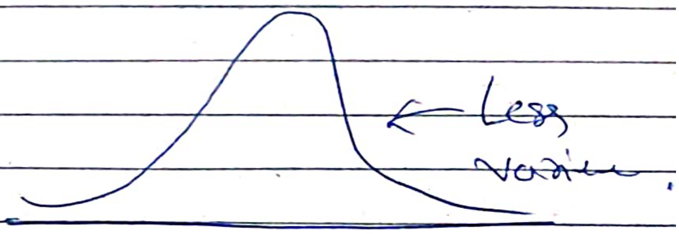
কতো দূরত্ব, কতো spread.

Ex-

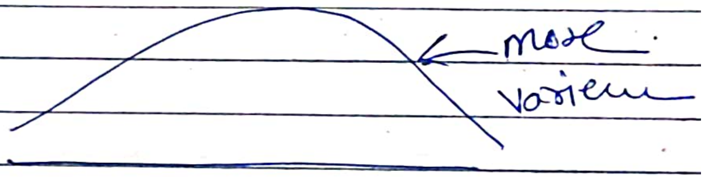
x	μ	x-μ	(x-μ)²
1		-1.83	3.34
2		-0.83	0.6889
3	2.83	+0.83	0.6889
3		0.17	0.03
4		1.17	1.37
5		2.17	4.71

μ = 2.83   10.84

$\sigma^2 = \frac{10.84}{6} = 1.81$



*Standard daviation



$= \sqrt{\text{varience}}$

$= \sqrt{\sigma^2}$ → with the help of this

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

what is the range.

Percentiles & Quartiles.

With the help of those we find the Out lier.

Percentiles: It is the value below which a certain percentage of observation lie.

Percentile Rank of x = $\frac{\# \text{ of value below } x}{n} \times 100$

=> number of

Q Data: 2, 2, 3, 4, 5, 5, 5, 6, 7, 8, 8, 8, 8, 8, 9, 9, 10, 11, 11

what is the Percentile ranking of 10?

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

$n = 20$ ← number of values

Percentile Ranking of 10 = $\frac{16}{20} \times 100$

$= 80\%$.

Q what value would be ~~Best~~ Percentile Ranking of 25%?

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

Index Position = $\frac{25}{100} \times 21 = 5.25$

5th index = 5

6th " = 5

$\frac{6\text{th} + 5\text{th}}{2} = \frac{5+5}{2} = 5$

↑ Value

• FIVE NUMBER Summary

* Removing the outliers

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

outlier

n = 19

Value are used to remove outliers we need [lowerfence ⟷ higherfence].

lowerfence = $Q_1 - 1.5(IQR)$

upperfence = $Q_1 + 1.5(IQR)$

IQR - Interquartile Range = $Q_3 - Q_1$

Q_3 = 75% (Percentile)

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

$= 15$ thindex

$= 7$

Q_1 = 25% (Percentile)

$\frac{25}{100} \times (19+1) = 5$ index

$= 3$

IQR = 7 - 3 = 4

lowerfence = 3 − 1.5 × 4 = −3

upperfence = 7 + 1.5 × 4 = 13

[−3 ⟷ 13]

any thing is greater than 13 is outliers

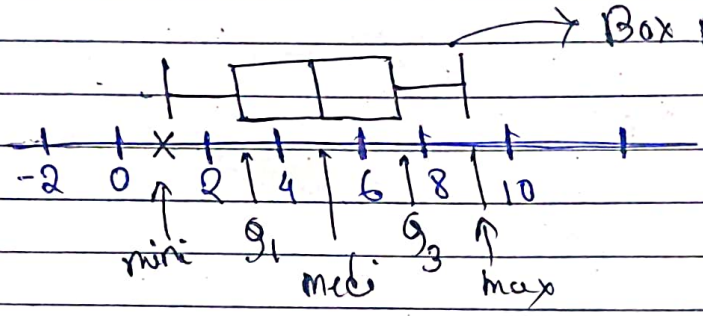
" " lesser " −3 " "

After Removing outlier

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

minimum = 1 $Q_1 = 3$

median = 5 $Q_3 = 7$ max = 9



for this outliers Remove

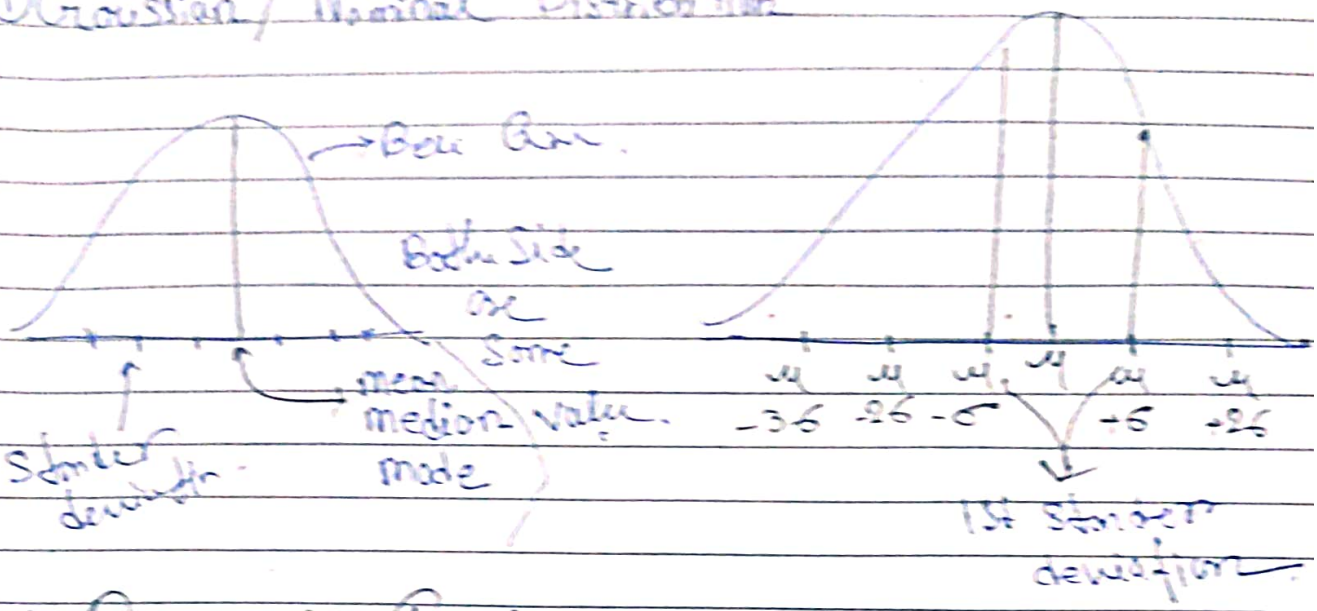
Q Why Sample variance is n-1 ?

DISTRIBUTIONS:

- Normal Distributions
- Stander normal Distributions
- Z Score
- Log normal Distrib
- Bernoulli Dis
- Binomial Dist

Ex- Ages = {24, 26, 27, 28, 30, 32 ——— }

① Gaussian / Normal Distribution



* Empirical Formula:

68-95-99.7% Rule

68% distribution fall in the reli.
⇒ 1st Standard deviation = ±σ → -σ to +σ

95% distribution = 2nd Standard dev. ⇒ -2σ to +2σ

99.7% " = 3rd " " ⇒ -3σ to +3σ

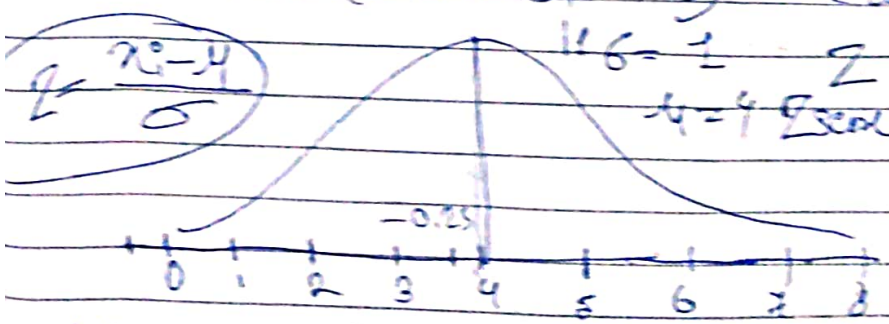
* Z score :

When we want to find out the Gaussian standard deviation (like 0.25, 0.75) we use this

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

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

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



After we apply Zscore in all value ⇒ {-3, -2, -1, 0, 1, 2, +3}

* Stande normal distribut. ($\mu=0, \sigma=1$)

{1, 2, 3, 4, 5, 6, 7} → Normal Distributim ($\mu=4, \sigma=1$)

⇓

Apply Zscore

⇓

{-3, -2, -1, 0, 1, 2, 3} → Standen Normal distribut ($\mu=0, \sigma=1$)

=>.

Practical applica

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

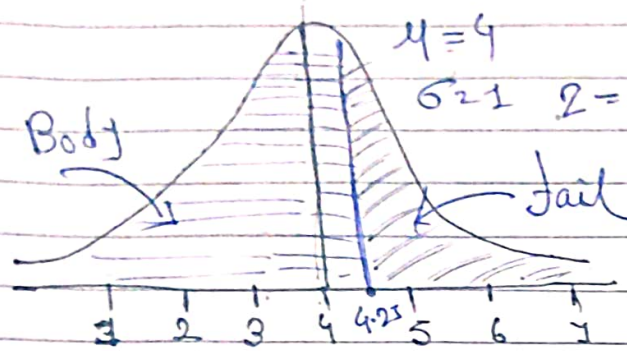
Normalizatin:

When we want to convert all value between 0 to 1 thus called Normalizatin.

6.

* Stat Interview qus...

what Percentage of Score fall above 4.25?



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

↑ Standard deviation.

this curve is symetrical so we can thought that whole one of this curve is = 1

So tail Area = 1 − left Area

= 1 − 0.5987

= 0.4013

= 40%

→ Standard deviation (0.25) → 0 (2500 left → 250 right)

→ we get the value from Z table (google)

Q In India the average IQ is 100, with a Standard deviation of 15. What Percentage of the ~~for~~ population would you expect is have an IQ lower then 85?

~~σ=15~~ σ=15 μ=100

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

Probability

Probability is a mesur of the likelihood of an Event

Roll dice {1, 2, 3, 4, 5, 6} Probab of getting 6

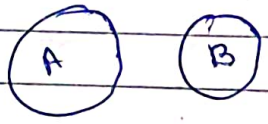
$$Pr(6) = \frac{\text{\# of ways an event can occur}}{\text{\# of possible outcome}}$$

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

* Addition Rule

▷ Mutual Exclusive Event : Two event are mutual exclusive if they can't occur at the same time

Eg: Rolling die {1, 2, 3, 4, 5, 6} it put 1 value at one time.



▷ Non Mutual Exclusive : multiple events can occur at the same time.

Q If I Toss ~~the~~ a coin, what is the Probability of the coin landing on ~~hon~~ to heads or tails?

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

OR → 6 + 1

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

$$= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2} = 0.5$$

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

total card = 52

one type card = 13 Cards

total type = 4

only Queen $P(Q) = \frac{4}{52}$ $P(Q \text{ and } B) = \frac{1}{52}$

" B $P(B) = \frac{13}{52}$

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

$= \frac{4}{52} + \frac{13}{52} - \frac{1}{52} = \frac{16}{52}$

* Multiplication Rule:

① Independent Event:

② Dependent Event:

* Independent:
Q what is the probability of rolling a '5' and then a '4' in dice?

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

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

$= \frac{1}{6} * \frac{1}{6} = \frac{1}{36}$

* Dependent

What is the probability of drawing a Queen and then Ace from a deck?

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

$$= \frac{4}{53} * \frac{4}{51}$$

* Permutation & Combination :-

Arrangement or formation (Permutation): $^nP_r = \frac{n!}{(n-r)!}$

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

* P Value : is a number describe how likely it is that your data would have occurred under the null hypothesis of your statistical test.

P ~~How many time~~ এ পাতি কথায় যে P value টা অবচেয়ে বেশি আসছে

Hypothesis Testing:

Let's a Quest to bose on Coin

There is some steps

1) Null Hypothesis : Coin is fair

2) Alternative Hypothesis: Coin is unfair

3) Experiment

4) Reject or Accept the null Hypothesis

How we find out the Coin is fair or unfair?

with the help of Significant value (α)

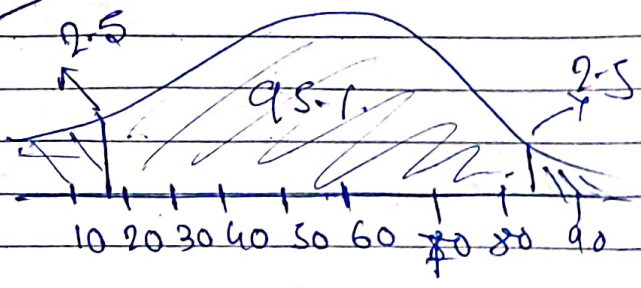
if $\alpha = 0.05 = 5\%$

$\therefore$ CI = Confidence Interval = $100\% - \alpha$

$= 100\% - 5\%$

$= 95\%$

If value fall in the CI, that is fair value.



$2.5 + 2.5 = 5\%$

(-) (+)

* ERROR.

Type 1 and Type 2 error

Null Hypothes (H_0) = Coin is fair

Alternative Hypothes (H_1) = Coin is unfair

Two type of desi—

Reality Check: Null Hypothesis is True or Null Hypothi, Fake

Desision:

Null Hypothes is True or Null Hypothes is Fake.

Outcome 1:

We reject the null Hypothisis, when in reality it is false → good Desi.

Reality

Outcome 2:

We reject the null Hypothesis, when in reality it is true → Good desi (Type 1 error)

Outcome 3:

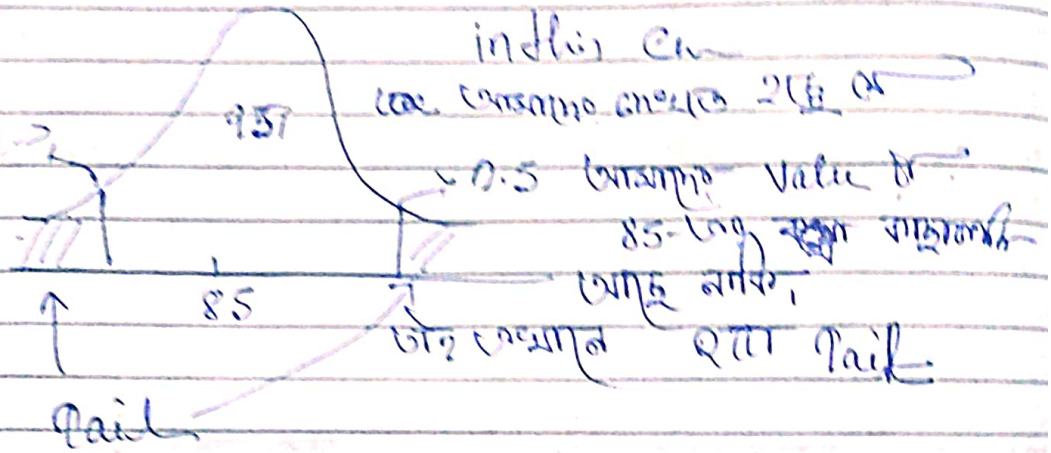
we accept / retain the Null Hypothesis when in reality it is false → Not Good desi (Type 2 error)

Outcome 4:

we accept the Null Hypothesis when in reality it is true → Good.

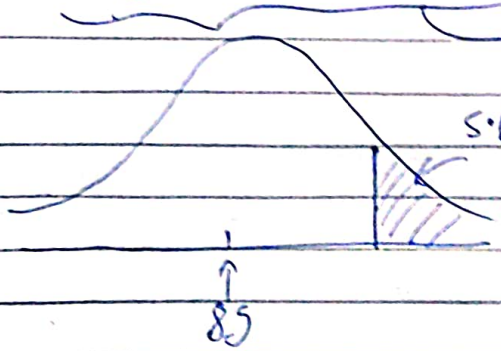
* 1 Tail and 2 Tail Test:

Q) Colleges in Karnataka have an 85% placement rate. A new College was recently opened and it was found that a sample of 100 Students had a placement rate of 88% with a Standard deviation 4%. Does this College has different placement rate? let $\alpha = 0.05$



∴ So this is two tail test.

But If Q) Does this College have placement rate greater than 85%?



for this one tail test.

* Confidence Interval (CI)

Point Estimate : The value of any statistic that estimate the value of a parameter.

$\bar{x} \longrightarrow \mu$ ↷ Estimate the mean

$\bar{x} = 2.4$ 3.

CI : Point Estimate ± margin of Error.

Q) on the quant test of CAT Exam, Standard deviation is known to be 100. A Sample of 25 test takers has a mean of 520 score. Construct a 95% CI about the mean?

Ans

$\sigma = 100, n = 25$ ~~$\alpha = 0.05$~~, $\bar{x} = 520$

$CI = 1 - \alpha \Rightarrow \alpha = 1 - CI = 1 - 95\%$

$= 1 - 0.95 = 0.05$

① Population Std given.

② $n > 30$

CI = Point Est ± Margin of error

$$\bar{x} \pm Z\frac{\alpha}{2}\left(\frac{\sigma}{\sqrt{n}}\right)$$ ↷ Standard error

Upper $= \bar{x} + Z\frac{0.05}{2} * \frac{100}{\sqrt{25}} = 559.2$ or $Z\frac{0.05}{2} = Z_{0.025}$

Lower $= \bar{x} - Z\frac{0.05}{2} * \frac{100}{\sqrt{25}} = 480.2$ for Z table $= 0.975$

One Sample Z-Test:

(i) Population Std is given
(ii) Sample size n ≥ 30

In the population, the avg IQ is 100 with a Sd of 15. Researchers want to test a new medication to see if there is position +ve or -ve effect on intelligence or no effect at all. A Sample of 30 patients who have taken the medication has mean of 140. Did the medicate affect intelligence? $\alpha = 0.05$

Ans

CI = 95%

1) Define Null Hypothesis

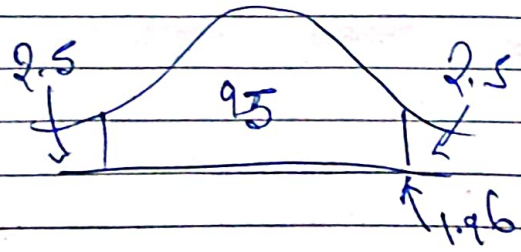
$H_0 = \mu = 100$

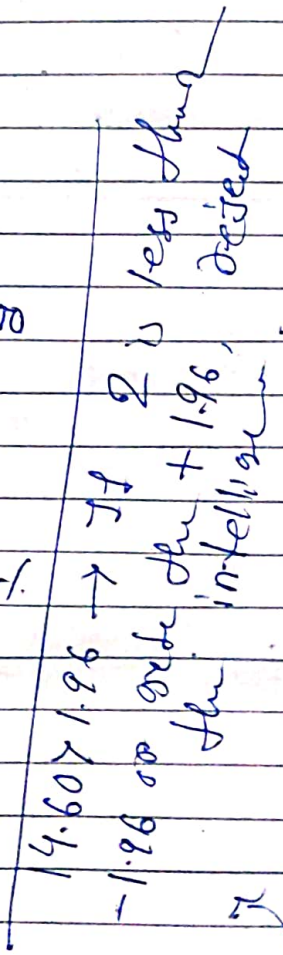
2) Arithmetic Hypothesis H_1 $\mu \neq 100$

3) Stat Alpha $\alpha = 0.05$

4) Stat decision Rule

CI = 95%





5) Calculate Z Stat

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{140 - 100}{\frac{15}{\sqrt{30}}} = 14.60$$

$\frac{\sigma}{\sqrt{n}}$ → Standard Error

* One Sample t-test.

We use Z-test when Population S.t d.

we " t-test " unknow Population

Q Population the avera IQ = 100
n = 30, $\bar{x}$ = 140, S = 20

α = 0.05 ?

Ans

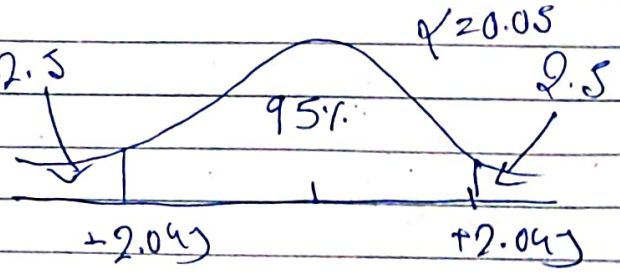
(i) $H_0 = \mu = 100$

(ii) $H_1 \Rightarrow \mu \neq 100$

(iii) Calculate degree of freedom

$n - 1 = 30 - 1 = 29$

(iv) State Decision Rule



(v) T-Test

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{10.96}{}$$

*CHI SQUARE TEST (काई स्क्वायर)

Chi Square test claims about Population Proportions.

It is a non Parametric test that is performed on Categorical (nominal or ordinal) data.

In ~~200~~ the 2000 Indian Census, the age of Individual in a Small town were found to be following:

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

In 2010, age of n=5000 individual were the Sample

<18	18-35	>35
121	288	91

Using $\alpha = 0.05$, would you conclude the Population distribution of ages has change in last 10 years

Ans:

<18	18-35	>35	2000
20%	30%	30%	← Expected

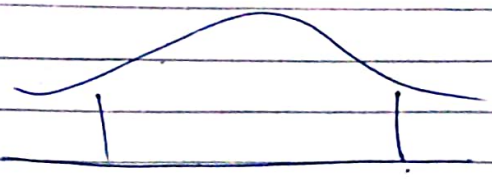
<18	18-35	>35	n = 500
121	288	91	← Observed
Exped: $500 \times \frac{20}{100} = 100$	$500 \times \frac{30}{100} = 150$	$500 \times \frac{50}{100} = 250$	

<18	18-35	>35	
121	288	91	→ Observed (f_o)
100	150	250	→ Exped (f_e)

1) H_0 = The data meets the distribution 2000 cars
H_1 = " " does not meet " " "

2) $\alpha = 0.05$ (95% CI)

4) Decision Boundary



5) χ^2 = chi square

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

$$= 232.94$$

Covariance : (there is no fixed value)

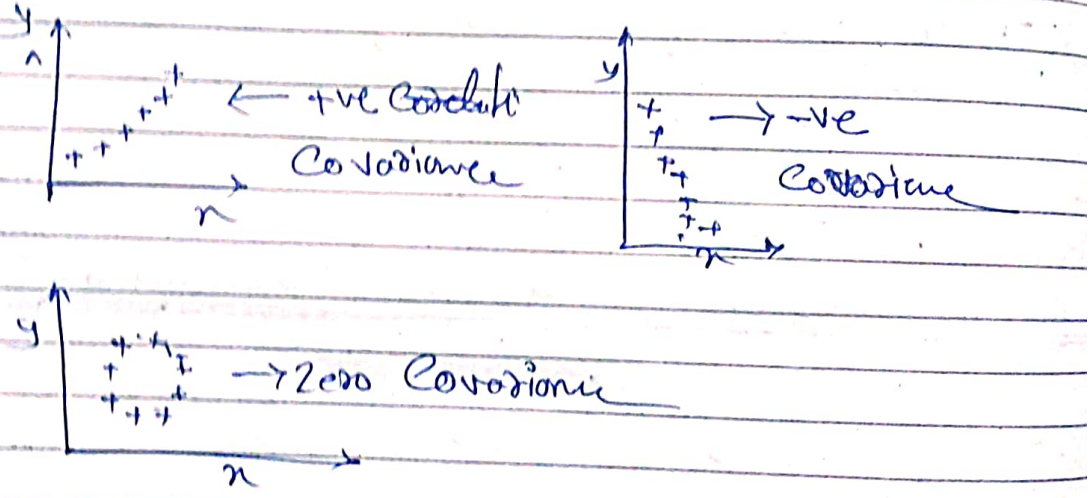
Quantify relationship between x & y

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

= +ve or -ve or 0

+ve → [x ↑ y ↑ or x ↓ y ↓]

-ve → [x ↓ y ↑ or x ↑ y ↓]



* Pearson Correlation Coefficient.

It value between (-1 to +1)

The more towards +1 more positively Correlated

The more towards -1 more negatively Correlated

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

* Spearman's Rank Correlation.

It Capture the non linear records

$$Spearm(x,y) = \frac{Cov(R_x, R_y)}{R_{Gx} * R_{Gy}} = \frac{Cov(R(x), R(y))}{R_{Gx} * R_{Gy}}$$

R → Rank of

→ It give the Ranks of value

X	Y	R(x)	R(y)
170	75	2	2
160	68	3	3
150	60	4	4
145	55	5	5
180	85	1	1

(Pvalue < Significance) → Reject Null Hypothesis

* P value & Significance. (Pvalue < Significance)

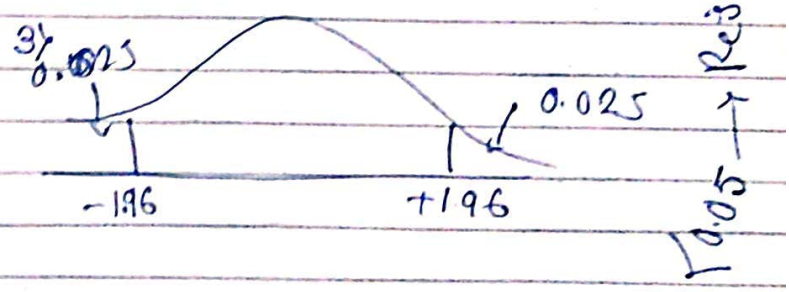
Q The avg weight of all residents in Bangalore city is 168 pound with a standard deviation 3.9. we take a sample of 36 individuals and the mean is 169.5 pounds. C.I = 95% ?

Ans:

$\mu = 168 \quad \sigma = 3.9 \quad \bar{X} = 169.5 \quad n = 36 \quad \alpha = 0.05$

1) $H_0 = \mu = 168$

$H_1: \mu \neq 168$



2) $\alpha = 0.05$

4) Z-test

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = 2.307 > 1.96$$ Reject the null Hypothesis

5) pvalue