"Statistics is the
Grammer of
Data Science / Data Analytics"

Hand Written Notes :

"Statistics Required for Data Analysis"

Ravindranath Kumbhar.

* Statistics

Statistics is the science of collecting, organizing and analyzing data.

Data :- "facts or pieces of information"

e.g. Age of students in classroom

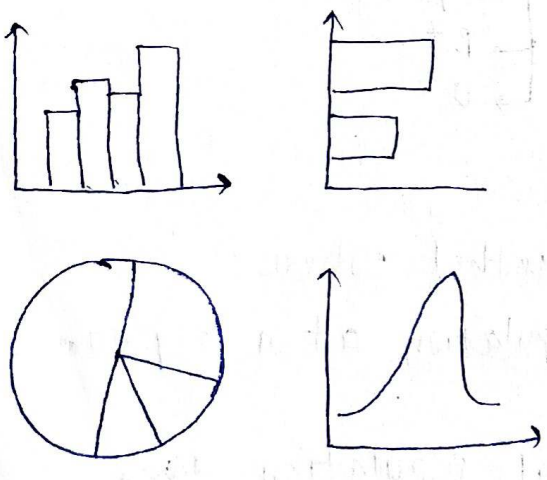
{24, 23, 22, 30, 32}

statistics is divided into two types.

Statistics

Descriptive stats

- It consist of organizing and summarizing of the data in the form of histogram, Barchart, Box plot, pdf, Pie i.e.



- It contains to find mean, median, mode, SD,

Inferential stats

- It consist of technique to form some conclusions.

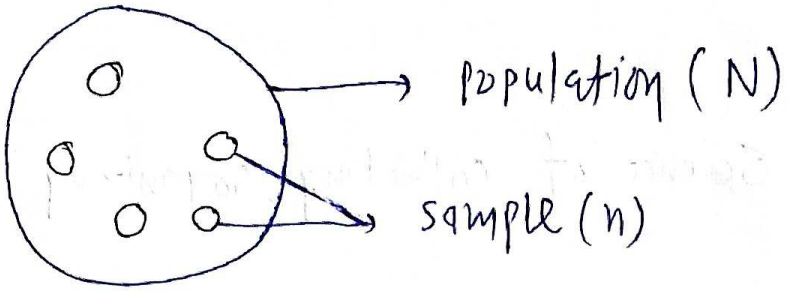
e.g. Suppose there are 1000 students in a college and we have to measure a average height of the students.

then we take some samples from this students and find the ~~avera~~ average height.

i.e.

using the sample size we can find the conclusion it belongs to inferential statistics.

* Population and sample :-



* Sampling techniques :-

① Simple random sampling :- Every member of a population (N) has an equal chance of being selected for your sample (n).

e.g.
. when we are tossing a coin then chances of probability of getting head and tail is equal i.e. $\frac{1}{2}, \frac{1}{2}$

② Stratified Sampling :-

Strata ⟹ Layers

It consists of different types of layers and this layers are non overlapping with each other.

e.g. ① Gender → male, female

② Blood group → A+, B+, O+

③ Systematic sampling :-

It is the probability of sampling method where researchers select members of the population at a regular interval.

e.g. suppose we have a list of total population then in systematic sampling we select every 5th person in the list or every 7th person in the list or so on.

④ Convinience sampling :-

Only those people who are interested will only be participating.

It is used in marketing.

e.g.

Suppose a company lonches a new product they they send a notification only who are interested.

* Variables :-

A variable is a property that can take an any value.

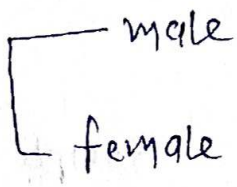
e.g. Age = 24

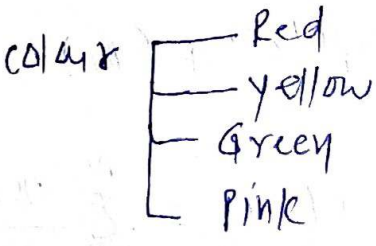
Height = 7.2

There are two kinds of variables.

① quantitative variable :- measured numirically

② qualitative variable :- categorical variable





(Based on some characteristics we can group categorical variables)

Quantitative Variable

Discrete variable

They are whole numbers

e.g. ① No. of bank accounts

② No. of children in family

- It does not contain fraction values

continous Variable

It is numeric

e.g. Height = 120.5 cm

weight = 50.5 kg

Rainfall, temperature.

* Intermediate Stats :-

* Measure of central tendency :-

A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position.

① Mean (Average) :-

Population (N)

$$\text{Popln mean } (\mu) = \sum_{i=1}^{N} \frac{x_i}{N}$$

Sample (n)

$$\text{Sample mean } (\bar{x}) = \sum_{i=1}^{n} \frac{x_i}{n}$$

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

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

$$\mu = 3$$

② Median :-

let e.g. {1, 2, 3, 4, 5, 100}

If there is any outlier is present in the set then steps to find the median :-

i] Sort the numbers

ii] find the central number.

① If no. of elements are even we find the average of central elements

② If no. of elements are odd we find the central element.

* Outlier :-

Outlier is the number that is completely different than the entire distribution.

E.g. let consider the above e.g.

{1, 2, [3, 4] 5, 100}

central

∴ Average = $\frac{3+4}{2}$ = 3.5

∴ median = 3.5

③ Mode :-

It is the most frequent occuring element

e.g. {4, 3, 1, 4, 4, 5, 6, 4, 8}

we select the element which has maximum frequency

i.e. 4

* Measure of Dispersion :-

① Variance (σ^2) :-

It is a measure of how far a set of numbers is spread out from their average value.

for the population data (N)

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

e.g. let consider the set $\{2, 2, 4, 4\}$

$\therefore$ mean (μ) $= \frac{12}{4} = 3$

$\therefore$

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

$\therefore \sigma^2 = \frac{4}{4} = 1$

for the sample data (n)

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

* $n-1$ is less than n.

⑤ Standard deviation ;-

$$\sigma = \sqrt{\sigma^2}$$

It is nothing but root of variance.

The Standard deviation is a measure of the amount of variation ~~of~~ dispersion of a set of values. A low Standard deviation indicates that the value tends to be close to the mean of the set.

* Percentile and Quartiles ;-

① Percentile ;-

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

Suppose,

a person A got 99 percentile it means that the person has got better marks than 99 percentage of the entire students.

e.g.

Dataset = {2, 2, 3, 4, (5), 5, 5, 6, 7, 8, 8, 8, 8, 8, 9, 9, [10], 11, 11, 12}

i] what is the percentile of ranking 10 ?

$$\therefore \text{Percentile of rank of } x = \frac{\text{no. of value below } x}{n} \times 100$$

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

$$= 80 \text{ percentile}$$

ii] What is the value that exist at 25 percentile ?

$$\text{value} = \frac{\text{Percentile}}{100} \times n$$

$$= \frac{25}{100} \times 20$$

$$= 5 \quad = \text{Index}$$

* 5 Number Summary :-

1. Minimum
2. first quintial (25%) Q_1
3. Median
4. Third quintile (75%) Q_3
5. Maximum

It is used for removing outliers.

Q. How to remove outlier ?

let consider a example

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

for removing the outliers we have to create lowerfance and higherfance.

[lowerfance ⟷ higherfance]

lowerfance = $Q_1 - 1.5(IQR)$ $\quad Q_1 = 25\%$.

higherfance = $Q_3 + 1.5(IQR)$ $\quad Q_3 = 75\%$.

$IQR = Q_3 - Q_1$ = difference betⁿ two quartile.

$Q_1 = 25\% = \frac{\text{Percentile}}{100} \times (n+1) = \frac{25}{100}(19+1) = 5^{th}$ index
$= 3$

$Q_3 = 75\% = \frac{75}{100} \times (19+1) = 15^{th}$ index
$= 7$

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

$\therefore$ lower fance $= Q_1 - 1.5(IQR)$
$= 3 - 1.5(4)$
$= -3$

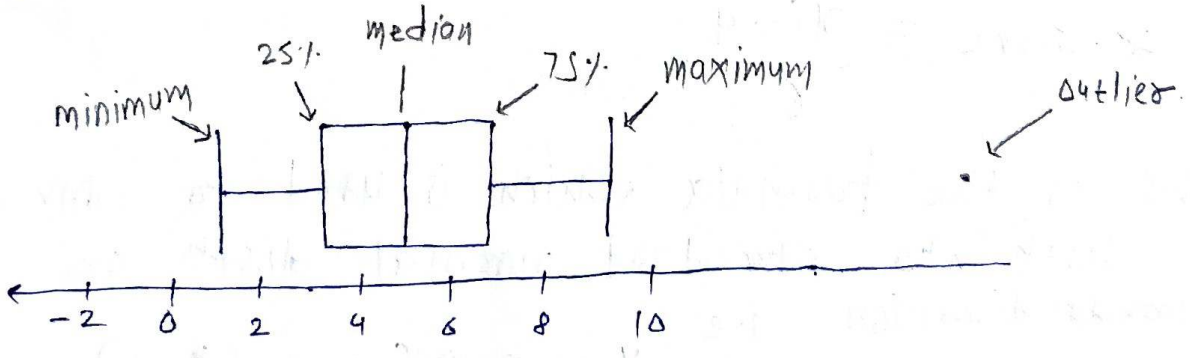
Higher fance $= Q_3 + 1.5(IQR)$
$= 7 + 1.5(4)$
$= 13$

$\therefore [-3, 13]$

$\therefore$ All the numbers are lies beth $[-3, 13]$
27 is the greater than 13 so we can remove it.
$\{1, 2, 2, 2, 3, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9\}$

$\therefore$ ① minimum = 1
② Q_1 = 3
③ median = 5
④ Q_3 = 7
⑤ maximum = 9

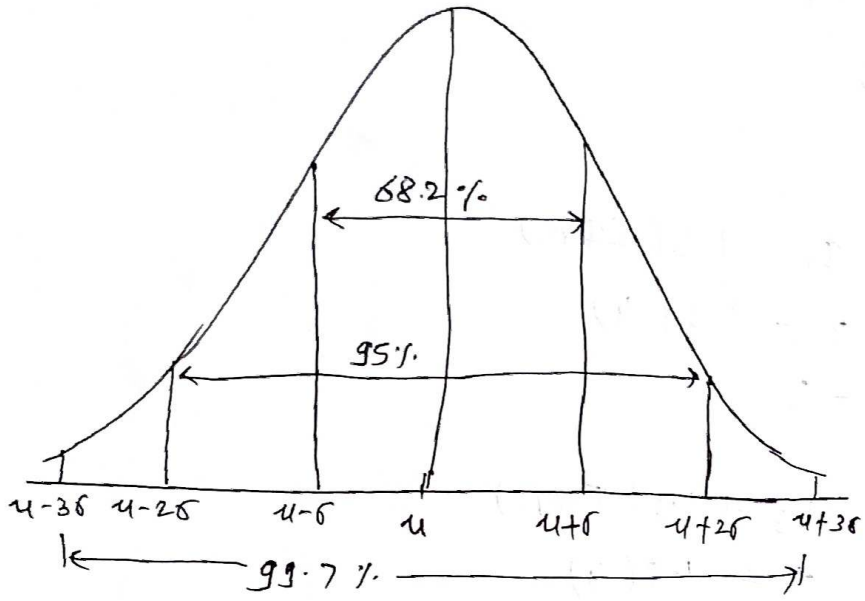
Box plot :-



* Normal distribution / Gaussion dist^n ;-

It is symmetric and it has a bell shap curve.

In the normal distribution mean, median and mode are as same.



e.g.

$\chi = \{$ 100 numbers in this set $\}$

then

68.2 % of the data are present in the 1st standard deviation.

95 % data are present in the 2nd S.D.

99.7 % data are present in the 3rd S.D.

* Empirical formula

[68 - 95 - 99.7]

* Standard Normal distribution ;-

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

This is the formula which is used to convert the dist^n into standard normal dist^n or standardization. i.e.

$$Y \approx SND(\mu = 0, \sigma^2 = 1)$$

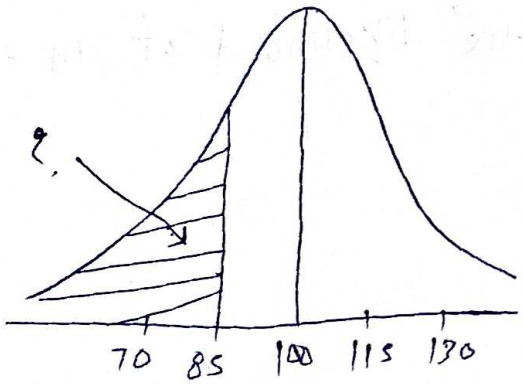
* why Standard normal distn ;-

Suppose in the dataset there are three diff. features and each having three different units. If we have to do some prediction then it is complicated with this data.

So, we can scale down of this data in some scale. The process is known as standardization.

e.g. Que.

In india the average IQ is 100 with a standard deviation of 15. what is the % of population would you expect to have an IQ lower than 85 ?



$\mu = 100$

$\sigma = 15$

$\therefore$ Z-score $= \frac{85 - 100}{15}$

$= -1$

now check the value of -1 in Z-table.

The value of -1 is 0.1587.

$\therefore$ IQ lower than 85 = 0.1587.

* Central limit theorem :-

If you take a sufficiently large sample size from a population with a finite level of variance σ^2, the mean of all samples from that population will be roughly equal to the population mean.

e.g.

Suppose we have a non-normal distn then we take a samples from this distribution and find the sample mean ($\bar{x}$) If we plot this sample sample mean then it follows normal distribution.

* Probability :-

Probability is the measure of the likelihood of an even

e.g. Tossing a coin

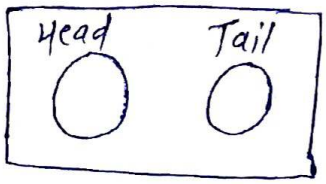
$P(H) = 0.5$

$P(T) = 0.5$

① Mutual exclusive events :-

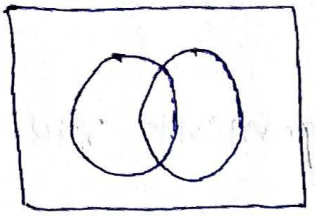
Two events are mutually exclusive events if they cannot occure at the same time.

e.g. Tossing a coin, Rolling a dice.



Non mutual exclesive events :-

They occures at the same time.



Q. what is the prob. of coin landing on heads or tails ?

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

* Independent event :-

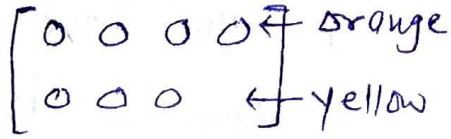
Two events are independent if they do not affect one another.

e.g. Tossing a coin

* Dependent events :-

Two events are dependent if they affects from one another.

e.g. Bag of colour marbles



$\therefore P(\text{orange}) = \frac{4}{7}$ and $P(\text{yellow}) = \frac{3}{6}$

* Permutation :-

It is the number of ways things can be ordered or arranged.

The order of the arrangement matters.

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

* combination :-

order of the selection does not matters.

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

* covariance :-

It is a systematic relationship between two random variables in which a change in the other reflects a change in one variable.

* If you have a two ~~variables~~ distribution and you have to find the relation beth them then use the covariance.

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

* $V(X) = cov(X,Y)$

While calculate the covariance if we get a more positive value or less negative value then we use pearson corln coeff.

i.e. +ve $\longrightarrow \infty$ -ve $\longrightarrow \infty$.

Pearson correlation works with linear data.

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

Spearman rank correlation works with non-linear data. and linear.

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

* P-value :-

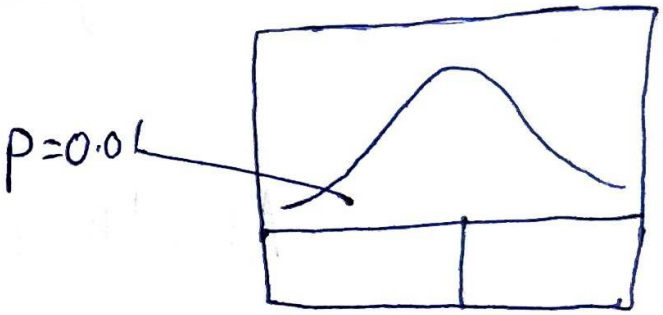
It is the prob. for the Null hypothesis to be true.

also called as significant value.

P=0.01 means that

out of 100 touches

1 time you touch there.



* Confidence Interval :-

Confidence interval is a range of values that is used to estimate population parameter such as mean, proportion and other parameters.

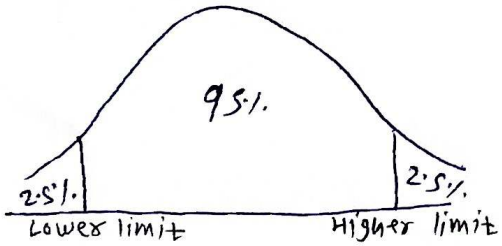
$$C.I = \text{point estimate} \pm \text{margin error}$$

* **Point estimate :-** Suppose in some experiment you are not able to find population mean in that case you have to find sample mean ($\bar{x}$) and using this sample mean you are able to estimate population mean. so the sample mean $\bar{x}$ is point estimate.

example :- what is the average size of the shark?

⇒ using the 95% confidence interval we can find it.

let consider $\sigma = 100$, $n = 30$, $\bar{x} = 500$



α = Significant value

$\alpha = 2.5 + 2.5 = 5\% = 0.05$

$Z_{0.025} = 1 - 0.025 = 0.976$

$= 1.96$

* check the $Z_{0.976}$ in the z-table

$$C.I = \text{point estimate} \pm \text{margin error}$$

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

$$= 500 \pm Z_{\frac{0.05}{2}} \frac{100}{\sqrt{30}}$$

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

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

∴ At the 95% C.I the average ~~value~~ size of shark lies between the range 386-613.

* Hypothesis testing :-

Hypothesis testing is a statistical method used to make inferences about the population parameters based on the sample data.

This process involves formulating two hypothesis

1] Null Hypothesis (H_0) :- This is the assumption or claim that there is no significant effect, relationship or difference.

2] Alternative Hypothesis (H_1) :-
This is the assumption that there is a significant effect, relationship or difference.

for example,

Peoples in ↳ particular city

Suppose we have to measure average height of the Then with the help of sampling technique we can take a sample data and perform some experiment ~~on the~~ i.e. statistical testing. Then we come up with some conclusion about the population. This entire process is called as Hypothesis testing.

Use case :-

① Hypothesis testing provides structured and objective framework for making decisions based on data.

② Researchers use hypothesis testing to determine whether there is enough evidence to support or reject a specific research hypothesis.

* Chi- Square test :-

It is used to check the dependency of two variables.

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

i.e. f_o = Observed values

f_e = Expected values

$$f_e = \frac{f_c f_r}{f_n}$$

e.g. 500 elementary school boys and girls are asked which is their favourite colour : Blue, green, or pink Result are shown below.

	Blue	Green	Pink	
Boy	100	150	20	270
Girl	20	30	180	230
	120	180	200	

using $\alpha = 0.05$ would you conclude that there is a relationship between gender and favourite colour.

=> Null Hypothesis :-

H_0 : Gender and colour are related

H_1 : Gender and colour are not related.

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

Degree of freedom :-

= (row - 1) (column - 1)

= (2 - 1) (3 - 1)

= 2

check the value of d.f 2 and $\alpha = 0.05$ in the chi-square table.

Test stetistics = 5.991

chi-square test :-

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} \quad , \quad f_e = \frac{f_e f_r}{f_n}$$

Given table

Observed	Blue	Green	Pink	
Boy	100	150	20	270
Girl	20	30	180	230
	120	180	200	500

(Boy , Blue) = (120 X 270)/500 = 64.8

(Boy , Green) = (180 X 270) /500 = 97.2

(Boy , Pink) = (200 X 270)/500 = 108

(Girl, Blue) = (120 X 230)/500 = 55.2

(Girl, Green) = (180 X 230)/500 = 82.8

(Girl, Pink) = (200 X 230) /500 = 92

Expected	Blue	Green	Pink
Boy	64·8	97·2	108
Girl	55·2	82·8	92

$$\chi^2 = \frac{(100-64\cdot8)^2}{64\cdot8} + \frac{(150-97\cdot2)^2}{97\cdot2} + \frac{(20-108)^2}{108} +$$

$$\frac{(20-55\cdot2)^2}{55\cdot2} + \frac{(30-82\cdot8)^2}{82\cdot8} + \frac{(180-92)^2}{92}$$

$$= 19\cdot12 + 28\cdot68 + 71\cdot70 + 22\cdot44 + 33\cdot66 + 84\cdot17$$

$$\chi^2 = 259\cdot79$$

$\chi^2 = 259\cdot79 > 5\cdot99$

Reject the H_0.

i.e. They are not related.

* Statistical Testing :-

① z test - (comparision of mean) - sample S.D is given (s)

② t test - (comparision of mean) - popln S.D is given (σ)

③ ANOVA - (Analysis of variance)

④ f test - (comparision of variance)

⑤ chi-square test (comparision of categorical variable)

* ~~One Sample~~ z test :-

z test is a statistical Hypothesis test which is used to determine ~~the~~ whether the means of two datasets are different. This test is useful when you have a large sample size and know the population standard deviation.

There are ~~two~~ main types of z test

① One sample z test

② Two sample z test

① One sample z test :-

This test is used whether the sample mean is significantly different (greater than, less than or not equal) than a population mean when the population standard deviation is known.

formula - $$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

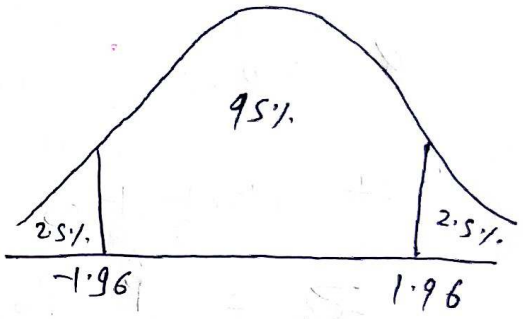
e.g.

In a popl^n the average IQ ($\mu = 100$) and $\sigma = 15$ then the doctor tested new medication to find out whether it increases the IQ or decreases IQ. After medication, after one month sample of 30 participents are taken ($n = 30$) and this 30 had a mean ($\bar{x} = 140$) so, did medication affect intelegents ? $\alpha = 0.05$

$\Rightarrow$ ① Null hypothesis $H_0 : \mu = 100$

Alternate hypothesis $H_1 : \mu \neq 100$

It is a two tail test



② State alpha $\alpha = 0.05$

③ State decision rule

$= 1 - 0.025$
$= 0.975$

If the value lies between the -1.96 to 1.96 then accept the null hypothesis. and the value $\begin{bmatrix} < -1.96 \\ > 1.96 \end{bmatrix}$ then reject the null hypothesis.

④ Test statistics

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

⑤ Result

$14.60 > 1.96$

$\therefore$ reject the null hypothesis and accept the alternate hypothesis.

* One sample Z-test with proportion :-

e.g.

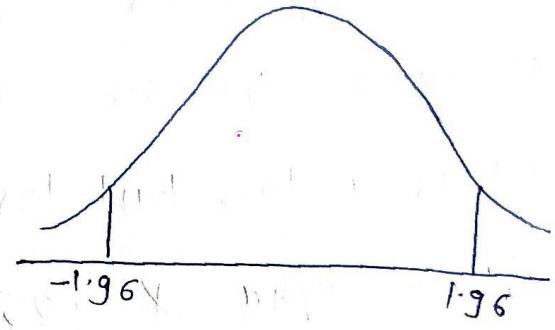
A survy claims that 9 out of 10 doctors recommend aspirin for their patients with headache. To test this claim a random sample of 100 doctors is taken. Out of this 100 doctors, 82 indicates that they recommend Asparin. Is this claim is accurate ? $\alpha = 0.05$

$\Rightarrow$

$H_0 : P = 0.90$

$H_1 : P \neq 0.90$

$\alpha = 0.05$



test statistics

$$Z_0 = \frac{\hat{P} - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}}$$

$\hat{P} = 0.80$

$P_0 = 0.90$

$n = 100$

$$= \frac{0.82 - 0.90}{\sqrt{\frac{0.90(0.10)}{100}}}$$

$$= -2.667$$

$\therefore \quad -2.667 \notin [-1.96, 1.96]$

$\therefore$ Reject the null hypothesis and accept the alternate hypothesis

$\therefore$ Claim is Inaccurate.

② Two Sample Z-test :-

This test is used to compare the means of two independent samples to determine if they are different or not.

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

e.g.

Suppose you want to determine if there is a significant difference in the mean strength of two diff. types of materials (material A and B) that you use in your products. you have collected samples of each material and conducted strength test on them. you want to know if the mean strength of material A is significantly different from the mean strength of material B. $\alpha = 0.05$

Material A :

$n_1 = 50$

$\bar{x}_1 = 75$

$s_1 = 10$

material B :

$n_2 = 60$

$\bar{x}_2 = 80$

$s_2 = 12$

n = sample size

$\bar{x}$ = sample mean

s = sample standard deviation.

$\Rightarrow$

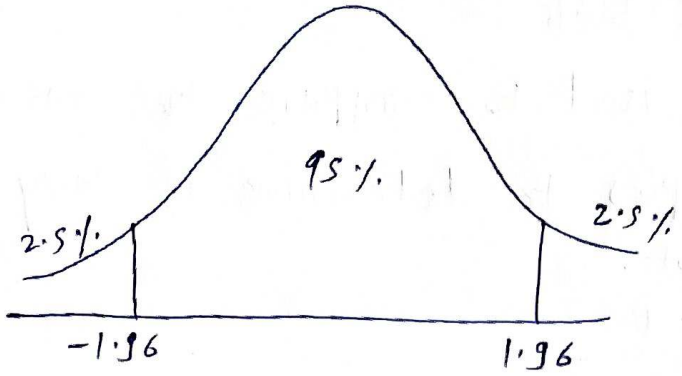
Null Hypothesis : $H_0 : \mu_1 = \mu_2$

Alternate Hypothesis : $H_1 : \mu_1 \neq \mu_2$

significant value $\alpha = 0.05$

Test statistics

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{75 - 80}{\sqrt{\frac{10^2}{50} + \frac{12^2}{60}}} = -2.369$$



$= 1 - 0.025$

$= 0.975$

Result :-

$-2.369 \notin [-1.96, 1.96]$

∴ we Reject the Null hypothesis.

* t-test :-

A t test is a statistical test that is used to compare the means of two groups. To determine if there is a significant difference between the means of two groups.

* In this the population S.D is not given and n should be less than 30.

e.g.

In the population the average IQ is 100, a team of scientist want to test a new medication to see if it has a +ve or -ve effect on intelegence or no effect at all. A sample of ~~30~~ 25 participents have taken the medication and has a mean IQ of 140 with standard deviation 20.

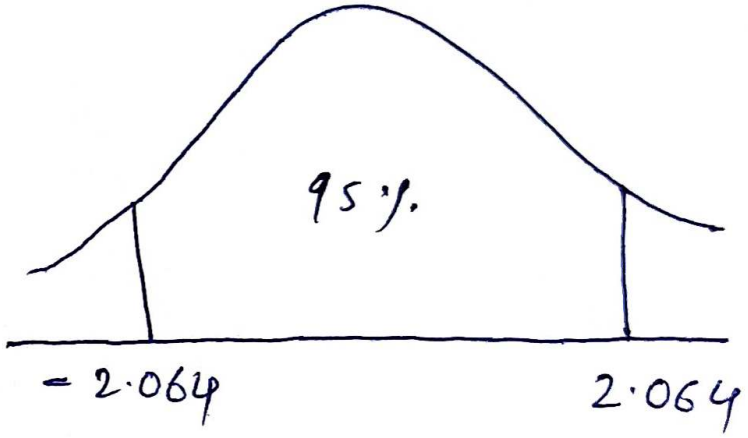
did medication affect the intelegence. $\alpha = 0.05$

› ① $H_0 : \mu = 100$

$H_1 : \mu \neq 100$

② Degrees of freedom

$n - 1 = 24$



$$t \text{ test} = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{140 - 100}{20/\sqrt{25}} = 10$$

$\therefore 10 > 2.064$

$\therefore$ Reject the Null Hypothesis.

$\therefore$ medication has increased the intelegence.