Data Analytics.

Definition: Techniques used for business analysis to get meaningful conclusions. The techniques are divided into three types.

1. Descriptive Analytics: This deals with the things that had already been done
   Ex: Visualizing the data, summary statistics etc
2. predictive analytics: This says what would happen if done something with the data
   Ex: classification, regression etc
3. Prescriptive analytics: This shows what should be done
   Ex:- Automated cars. Based on situation, it has to analyse the data and choose a path.

Data types:

Structured Data: In these, there are defined rows & columns with names defined
Ex: Relational DB, Matrix etc.

Unstructured data: These are not defined.
Ex: facebook chats, twitter etc.

Based on measurement scale: This depicts the types of operation that can be done.

Nominal: The order of data is not mentioned.
Ex: studentname, color etc.

Ordinal: The order is mentioned.
Ex: good, better, Excellent.

Interval: The form that cannot be measured in a fixed way.

Ex: dates, temperature

Ratio: the form that can be represented in the fixed way.

Ex: height, weight

- Nominal, ordinal are categorical
- Interval, Ratio are Numerical.

How to classify.?	Nominal	ordinal	Interval	Ratio
Order	x	✓	✓	✓
difference	x	x	✓	✓
Abs Zero	x	x	x	✓
	mode	mode, median	mean, median, mode	median, mean, Mode

Measure of central tendency:

Mean: Mean is an average. $x_1, x_2, x_3 \ldots \bar{n} = \sum_{i=1}^{n} \frac{x_i}{n}$

- Advantage is that it considers each and every value.
- Disadvantage is the outlier, the value.

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0.$$

- If frequency is taken, $\bar{x} = \sum_{i=1}^{n} \frac{f_i x_i}{f_i}$

Median: Median is the mid value.

- sort all the elements, and pick the middle value.
- Advantage is that it doesn't have any impact.

Mode: Most repeated value (or) frequent value.

→ To identify the position in the dataset.

Percentile (P_x) is the value of data at which x percentage of data (i.e) below that value.

position corresponds to $P_x = \frac{x(n+1)}{100}$.

$P_{25} = \frac{25(12)}{100}$.

Measure of variance: Use to understand the variability in data.

- Range = Max value - min value
- Interquartile range: $Q_3 - Q_1$
- Standard deviation.
- Variance: It is a measure of variance in the data from the mean.

variation of population (σ^2)

$$\sigma^2 = \sum_{i=1}^{n} \frac{(x_i - u)^2}{n}.$$

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

Standard deviation (σ):

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

Random variable: A function which maps a real number for each sample point in the sample space S.

Ex: Tossing the coin 3 times.

$S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$

To find no. of tails:

$X = f(\text{no. of tails})$

$\Rightarrow \{0, 1, 1, 2, 2 \ldots\}$

Types of Random Variable:

Discrete random variable: which can take only finite no. of values on finite observation interval.

Continuous random variable: A random variable which can take infinite no. of values in a finite observation interval.

Ex: Time taken to fill water tank. Each time you perform the o/p is different.

Probability Mass function: Discrete Random Variable.

probability that the discrete random variable X takes on the values given by $P(x_i) \Rightarrow f(x) = P(X = x_i)$

It should satisfy: i) $f(x) >= 0$ for $\forall x_i$ in X ii) $\sum f(x) = 1$

Ex: flip a coin 2 times.

X = no. of heads.

$X = \{0, 1, 2\}$

$f(0) = P(X=0) = P(t,t) = 1/4 = 0.25$

$f(1) = 2/4 = 0.25$

$f(2) = 1/4 = 0.25$

X	0	1	2
f(x)	0.25	0.5	0.25

$f(x_i) >= 0$

$\sum f(x_i) = 1$.

31st July 19.

probability density function: If X is a continuous random variable, the function f(x) is probability density function where $f(x) = P(a \le x \le b) = \int_a^b f(x)\,dx$. and i) $f(x) \ge 0$

ii) $\int_{-\infty}^{\infty} f(x)\,dx = 1$.

Example: $f(x) = \begin{cases} x & 0 < x < 1 \\ 2-x & 1 < x < 2 \\ 0 & \text{otherwise} \end{cases}$

$$\Rightarrow \int_{1/2}^{3/2} f(x)\,dx = \int_{1/2}^{1} f(x)\,dx + \int_{1}^{3/2} f(x)\,dx$$

$$\Rightarrow \int_{1/2}^{1} x\,dx + \int_{1}^{3/2} 2-x\,dx$$

$$\Rightarrow \left[\frac{x^2}{2}\right]_{1/2}^{1} + 2[x]_{1}^{3/2} - \left[\frac{x^2}{2}\right]_{1/2}^{3/2}$$

$$= \left[\frac{1}{2} - \frac{\frac{1}{4}}{2}\right] + 2\left[\frac{3}{2} - 1\right] - \left[\frac{1}{2} - \frac{1/2}{2}\right]$$

$$= \left[\frac{1}{2} - \frac{1}{8}\right] + \left[2\left[\frac{3-2}{1}\right]\right] - \left[\frac{1}{2} - \frac{1}{8}\right]$$

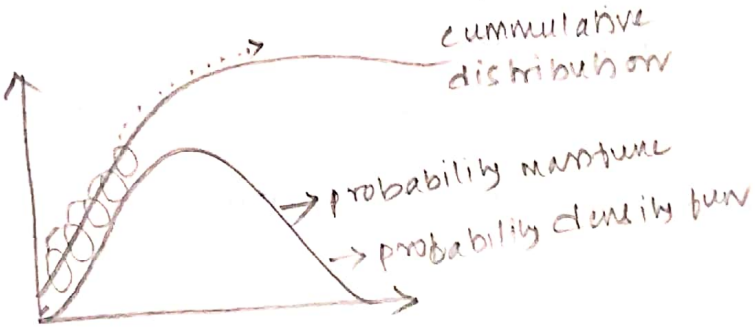
Cummulative distribution function cdf:

Cummulative distribution of random variable is defined as probability that the random variable 'x' takes a value "less than or equal to x".

→ Applicable for discrete and continuous random variable.

$$CDF = f(x) = P(X \leq x)$$

$$\Rightarrow f_x(a) = P(X \leq a) = \sum_{x \leq a} f(x)$$



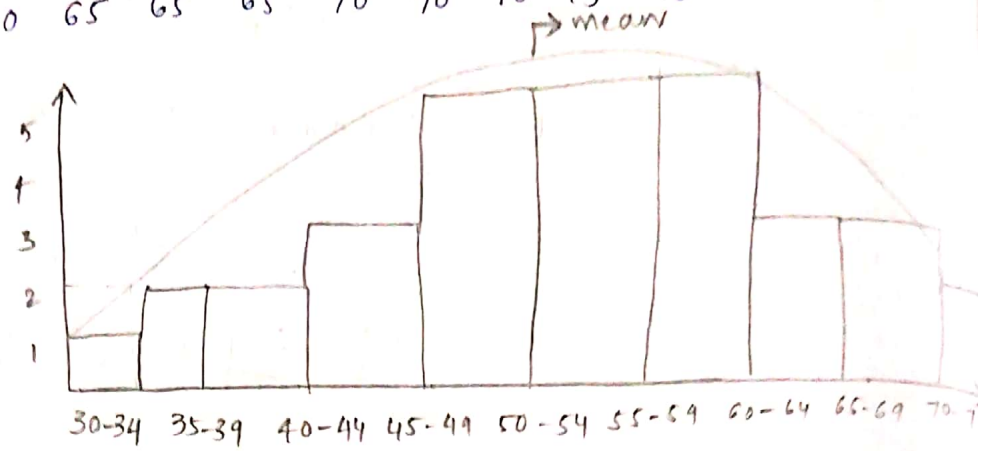
→ since the cummulative keeps on increasing, the graph for cummulative is always increasing.

Normal distribution: ($M = 0, \sigma = 1$)

→ Almost 90% of data analytics depends on normal distribution. (Most real world data shows up this distribution)

→ The frequency distribution centered around mean which follows a bell shape curve is normal distribution.

Ex: 35 35 35 40 40 45 45 45 45 50 50
50 50 50 55 55 55 55 55 60 60 60
60 60 65 65 65 70 70 70 75 75 80.



Bell shaped curve

- In normal distribution, mean, median, mode are same.

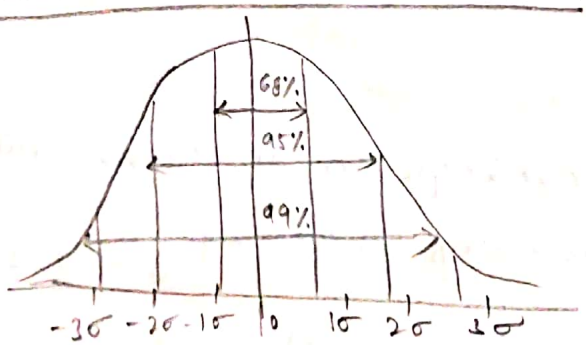
why normal distribution?

→ If the data follows dnormal distribution, several parametric tests can be applied and can draw meaningful conclusions by analysing the data

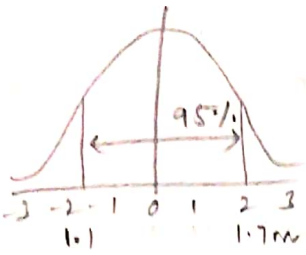
→ parametric tests statistics have lots of tools and it is well developed system.

→ parametric statistics use the mean, variance etc.

Standard Normalization distribution.



Ex: 95% students at school are in between 1.1m and 1.7m tall
Calculate mean and standard deviation.



since 95%, the 2σ to -2σ are filled, hence mean = $\frac{1.1+1.7}{2}$ = 1.4m

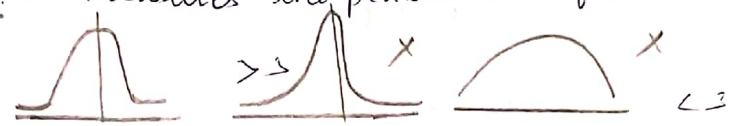
Standard deviation = 1.7 - 1.1 => 0.6

0.6/4 = 1.5m

* If variation from normal distribution is sufficiently large, all the statistical test are invalid.

To find variation: (balance of distribution)

1. Skewness: Measure of symmetry. (Mean to 50% left to mean to 50% right)
2. Kurtosis: It measures the peakedness of the distribution. (flatness)



Kurtosis = 0 — Peakedness — Kurtosis

[Mesokurtic] [leptokurtic] [platykurtic]

→ Mesokurtic is accepted whereas leptokurtic and platykurtic are not accepted

→ variation is accepted if Kurtosis $\leq$ 3 (standard error)

$$\text{Standard error} = \frac{\text{Standard deviation}}{\sqrt{n}}$$

If Kurtosis $\leq$ 3(SE), then distribution can be assumed as normal, else the distribution is not normal and the statistical tests are invalid.

Skewness: If company employee's salaries are taken, then CEO has high salaries other than employee. It gets affected by outlier.

Skew < 3 (standard error)

positive skew — negative skew

1st Aug'19.

Reason for peakedness is because of outliers: and is same for skew.

→ Removing outliers may cause a normal distribution.

Sampling and Estimation:

Sampling is a process of selecting subgroup from population to make inference about population parameters such as mean, standard deviation, proportion etc.

→ Its difficult to study the entire population, hence studying the sample infers studying some part of population.

Benefits of sampling:

i) Reduced cost

ii) speed.

Example: Literary Digest.

London 55%. 37%.

Roosvelt 41%. 61%.

(Expect) (original)

Steps used in sampling process:

1. Identification of correct target population.
2. Decide the sampling frame.
3. Determine the sample size.
4. Sampling method. (Probabilistic Sampling, Non-probabilistic Sampling)

Probabilistic Sampling:

i) select the sampling based on probability distribution

a) Random sampling

b) Stratified sampling

c) clustered sampling.

d) Bagging / bootstrap aggregration

stratified

Random sampling:

If homogenous, then random could be correct. If it is not homogenous, then dividing the heterogenous into class, where each class is a homogenous.

clustered sampling:

In stratified, from each group equal proportion is to be picked, but in cluster if the group is negligible, and does not affect in calculation are not selected. only the major ones that affect are collected.

Bootstrap aggregation: It checks whether the groups are similar or not.

Non probabilistic sampling:

i) convenience sampling: By convenience, the samples are collected without any cost, permission etc

ii) voluntary sampling: It is not mandatory to collect samples from particular group. Anyone who tends to provide the sample can give.

Sampling distribution: The probability distribution of statistic such as sample mean, standard deviation computed from random samples of population.

ex: population - 10 observations

1	2	3	4	5	6	7	8	9	10
5	10	15	20	25	30	35	40	45	50

Sample size 2:

5, 5	10, 15	20, 25	10, 30	40, 45 50	30, 40
mean: 5	12.5	22.5	20	45	35

10, 25	5, 20	40, 45	50, 20
17.5	12.5	42.5	35

→ As sample increases, it goes to normal distribution. If it applies normal distributions, statistical test can be applied.

5th Aug '19.

Standard Normal Distribution:

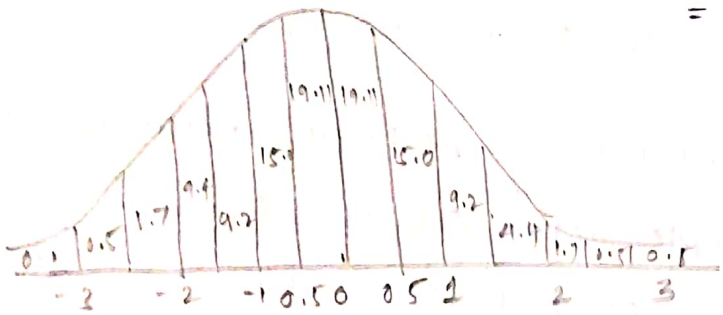
All the real world data is standardized to normal distribution by applying the below

$$z = \frac{x - \mu}{\sigma}$$ μ = mean, σ = standard deviation.

Let $\mu = 86$, $\sigma = 14$.

$P(x < z) \Rightarrow P(x < \frac{x-86}{14}) \Rightarrow P(x < -1) = P(\frac{72-86}{14})$

$= 15.87\%$.



Z score cummulative percentage

- $< = -3$: 0.1%.
- $< = -2$: 15.9%. 2.3%.
- $< = -1$: 15.9%.
- $= 0$: 50%.
- $< = 1$: 84%.
- $< = 2$: 97.7%.
- $< = 3$: 99.9%.

Confidence Interval:

Confidence interval is the range in which the value of population parameter is likely to lie with certain probability.

- Confidence interval is the sample mean or proportion plus or minus margin of error. (90% 95% 99%)

$= \boxed{\bar{x} \pm z \frac{\sigma}{\sqrt{n}}}$ if population standard deviation is known.

$\Rightarrow \bar{x} \pm \frac{s}{\sqrt{n}} \cdot z$ if population standard deviation is unknown.

ex: We measure the height of 40 randomly chosen men and get a mean height 175 cms, we also know that standard deviation of men's height is 20 cms.

Confidence Interval $= 175 \pm z \cdot \frac{20}{\sqrt{40}}$

since, most of the margin of error is 95%.

we take approximately z as 2 (97.7%).

$= 175 \pm 2 \cdot \frac{20}{\sqrt{40}}$

$= 175 + \frac{40}{\sqrt{40}} \; ; \; 175 - \frac{40}{\sqrt{40}}$

$\Rightarrow 175 + 6.32 \; ; \; 175 - 6.32$

$\Rightarrow 181.32 \quad 168.68$

z should be taken as 1.96 for margin of 95% error.

→ The population parameter cannot be random, then why are we guessing?

According to central limit theorem, if the distribution is plotted as frequency distribution the sample mean is μ and sample standard deviation = $\sigma/\sqrt{n}$.

* The sample mean we are trying to find is the estimation of population mean.

Central limit theorem — (CLT)

for a large sample drawn from a population with mean (μ) and standard deviation (σ), the sampling distribution of mean, $\bar{x}$ follows an approximate normal distribution with mean (μ) and standard deviation (Standard error) $\sigma/\sqrt{n}$ irrespective of the distribution of the population

* Central limit theorem is the basis for hypothesis test such as z-test and t-test. In many cases we will have access to only a sample and the inference about the population has to be made based on sample statistic
* assumption of CLT is that random variable have to be independent and identically distributed

Sample size estimation: —

1st Aug '19. Hypothesis testing:

Hypothesis: An assumption [for] certain characteristics about of population

Ex:- Statistics of different people based on teaching by HR.

25, 60, 45, 56, 32, 43, 47, 59, 39, 41.

Sample mean = 44.5

Standard deviation = 11.41

The minimum acceptable score for good employee is 45 as per research.

→ we cannot say that sample mean is not upto the minimum because, there may be sample bias. Considering the sample bias and the population, one should make the ranges.

- what is the probability or chance, we are still able to say that my sample has a mean, which is above 45 or 45.

H_0 (Null hypothesis), H_1 (Alternate hypothesis) are the types of hypothesis statistically.

H_0: There is no difference b/w population mean & hypothesized mean.

H_1: There is significant difference.

$H_0: \mu \geq 45$

$H_1: \mu < 45$

⇒ considering marginal error, it could be accepted within the range 37 to 46. But there can be a person at 36. belonging to other class.

Rejecting H_0, accepting H_1, which originally belongs to H_0 is known as Alpha (α) & Type I error.

→ Rejecting H_1, and accepting H_0 which originally comes from H_1 is known as Type 2 error termed as Bela (β)

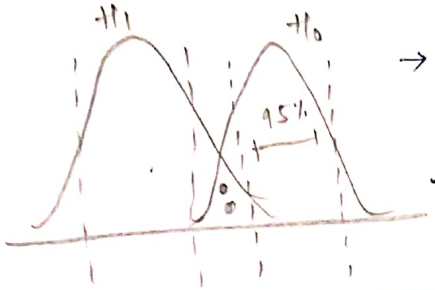
→ Type 1 error is produces problem, type 2 error is consumer problem.

Type 1 → is correct, but we are depicting as wrong

Type 2 → is wrong, but we are depicting as correct.

→ Always, null hypothesis is tested.

8 Aug '19.



→ False rejection of null hypothesis is type 1 error

→ False acceptance of null hypothesis is type 2 error.

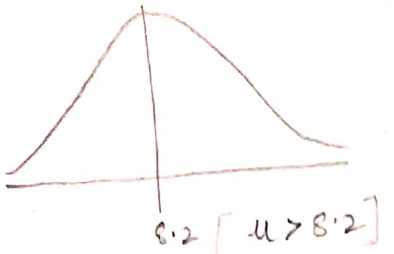
False acceptance = $1-\beta$. is called power of test, the lesser the value, more is accuracy.

Formulating hypothesis;

1. A researcher thinks that if knee surgery patients go to physical therapy, twice a week (instead of 3 times) their recovery period will be longer. Average recovery time for knee surgery patients is 8.2 weeks.

Soln:- $H_0: \mu \leq 8.2$

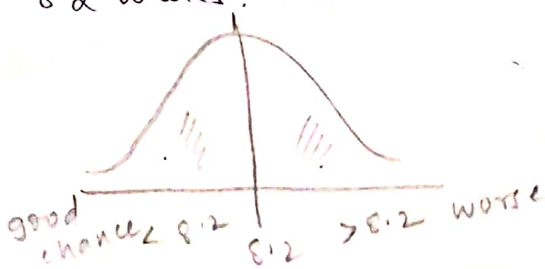
$H_1: \mu > 8.2$ ✓



2. A researcher is studying the effect of radical exercise program on knee surgery patients. There is good chance that the therapy will improve recovery time, but there is also possibility it will make it worse, the problem average recovery time is 8.2 weeks.

$H_0: \mu = 8.2$

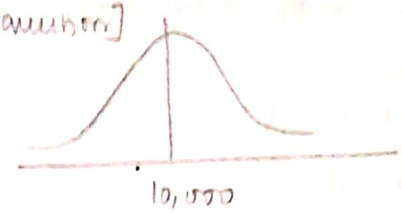
$H_1: \mu \neq 8.2$



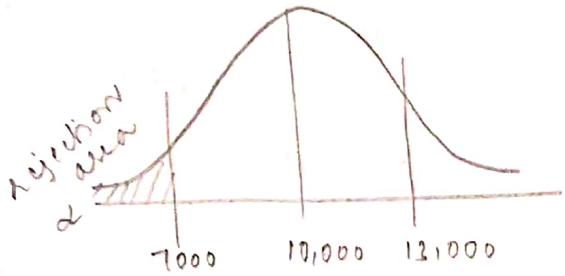
3. The salary of machine learning expert on an average is atleast 10,000 $. [hypothesis question]

$H_0 : \mu \geq 10,000$

$H_1 : \mu < 10,000$



If it's not researcher based question, it should be taken as hypothetical question, and the question itself becomes needed hypothesis alternate



- Its not a problem if the mean results greater than 10,000, but if the mean occurs less than 7000, the hypothesis is rejected because of type1 error.

This is called one tail test/ left tail test. $\alpha = 1.64$

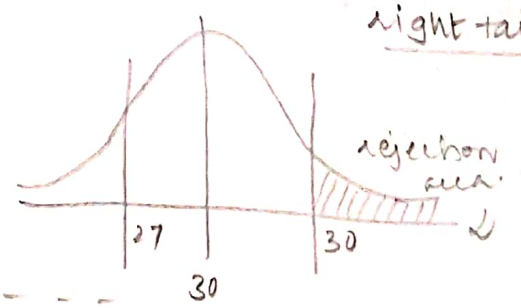
4. Average waiting time at London airport security check is less than 30 mins

$H_0 : \mu \geq 30$.

$H_1 : \mu < 30$

right tail test

$\alpha = 1.64$



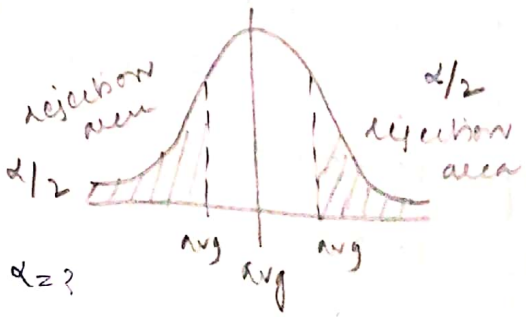
one sample test

Two sample test.

5. Average students of male & female at MBA graduation is different

$H_0 : \mu_M = \mu_F$

$H_1 : \mu_M \neq \mu_F$



7th Aug'19 Hypothesis testing:

Hypothesis: An assumption [for] about certain characteristics of population

Ex:- Statistics of different people based on teaching by HR.

25, 60, 45, 56, 32, 43, 47, 59, 39, 41.

Sample mean = 44.5

Standard deviation = 11.41

The minimum acceptable score for good employee is 45 as per research.

→ we cannot say that sample mean is not upto the minimum because, there may be sample bias. Considering the sample bias and the population, one should make the ranges.

- what is the probability or chance, we are still able to say that my sample has a mean, which is above 45 or 45.

H_0 (Null hypothesis), H_1 (Alternate hypothesis) are the types of hypothesis statistically.

H_0: There is no difference b/w population mean & hypothesized mean.

H_1: There is significant difference.

$H_0: \mu \geq 45$

$H_1: \mu < 45$

⇒ considering marginal error, it could be accepted within the range 37 to 46. But there can be a person at 36. belonging to other class.

Rejecting H_0, accepting H_1, which originally belongs to H_0 is known as Alpha (α) & Type I error.

→ Rejecting H_1, and accepting H_0 which originally comes from H_1 is known as Type 2 error termed as Beta (β)

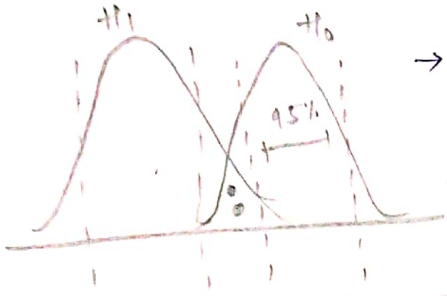
→ Type 1 error is producer problem, type 2 error is consumer problem.

Type 1 → its correct, but we are depicting as wrong

Type 2 → its wrong, but we are depicting as correct.

→ Always, null hypothesis is tested.

8th Aug '19.



→ False rejection of null hypothesis is type 1 error

→ False acceptance of null hypothesis is type 2 error.

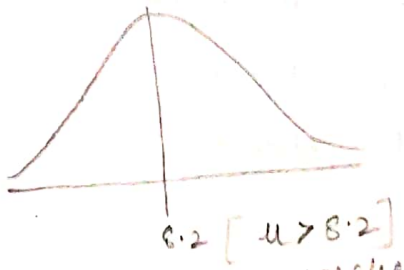
False acceptance = $1-\beta$. is called power of test, the lesser the value, more is accuracy.

Formulating hypothesis:

1. A researcher thinks that if knee surgery patients go to physical therapy, twice a week (instead of 3 times) their recovery period will be longer. Average recovery time for knee surgery patients is 8.2 weeks.

Soln:- $H_0: \mu \leq 8.2$

$H_1: \mu > 8.2$

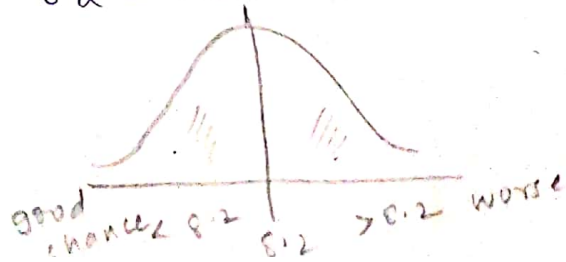


2. A researcher is studying the effect of radical exercise program on knee surgery patients there is good chance that the therapy will improve recovery time, but there is also possibility it will make it worse, the average recovery time is 8.2 weeks.

problem

$H_0: \mu = 8.2$

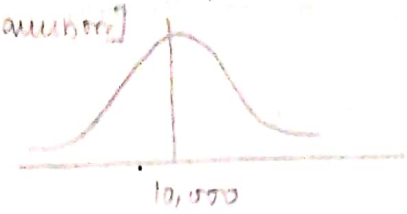
$H_1: \mu \neq 8.2$



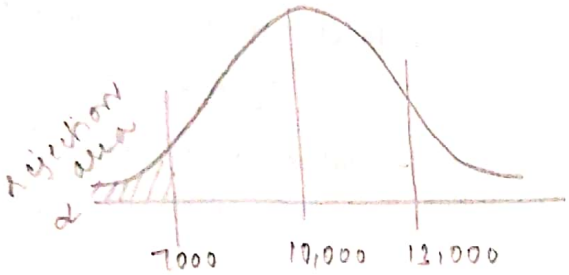
3. The salary of machine learning expert on an average is atleast 10,000 $. [hypothesis question]

$H_0 : \mu \geq 10,000$

$H_1 : \mu < 10,000$



If it not researchee based question, it should be taken as hypothetical question, and the question itself becomes null hypothesis alternate



- Its not a problem if the mean results greater than 10,000, but if the mean occurs less than 7000, the hypothesis is rejected because of type1 error.

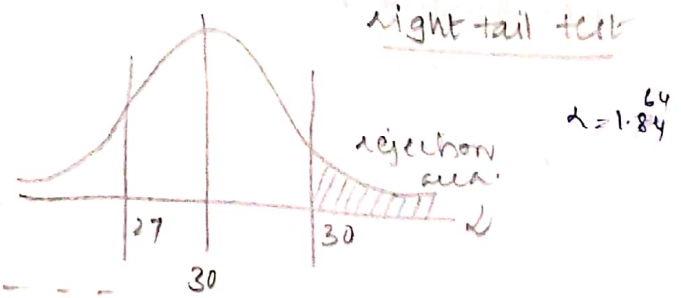
This is called one tail test/ left tail test. $\alpha = 1.84$ 1.64

4. Average waiting time at London airport security check is less than 30 mins

$H_0 : \mu \geq 30$.

$H_1 : \mu < 30$

one sample test

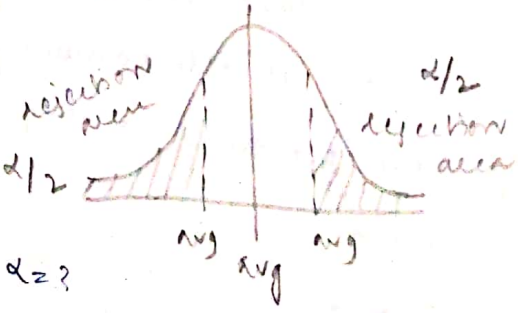


Two sample test.

5. Average students of male & female at graduation is different MBA

$H_0 : \mu_M = \mu_F$

$H_1 : \mu_M \neq \mu_F$

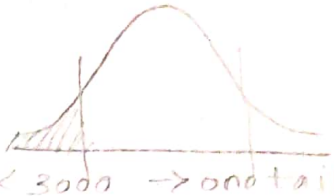


16 Aug 19.

A manufacturer purchased bulbs that are supposed to burn for a mean life of atleast 3000 hrs with a standard deviation of 500 hrs. If a sample of 100 bulbs is taken, with mean $\bar{x} = 2800$ hrs

Soln:- $H_0 : \mu \geq 3000$ $\quad \alpha = 5\%$.

$H_1 : \mu < 3000$



for one tailed test, $\alpha = 5\%$.

$z = -1.645$

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{2800 - 3000}{500/\sqrt{100}} = \frac{2800 - 3000}{50}$$

$$= -4$$

Reject Null hypothesis

2. A company claims that its weight reducing drug will cause a weight loss of atleast 10 pounds within 1 month, a random sample of 64 subjects is taken and the average weight loss is 7 pounds which Standard deviation is 4 pounds. Test $\alpha = 0.01$

Soln:- $H_0 : \mu \geq 10$ $\quad n = 64, \bar{x} = 7, S = 4, \alpha = 0.01,$

$H_1 : \mu < 10$

$$z = \frac{7 - 10}{4/\sqrt{64}} = \frac{-3}{4/8} = \boxed{-6}.$$

$-6 \geq 10$

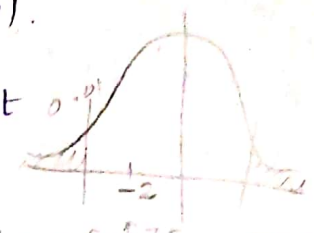
Reject Null hypothesis.

(3.) A researcher claims that 10 year old watch 6.6 hrs of TV daily. You try to verify this with following sample data.

$n = 100$, $\bar{x} = 6.1$ hrs, $S = 2.5$ hrs, $\alpha = 0.01$.

H_0: $\mu = 6.6$ $\Rightarrow$ Two tailed test

H_1: $\mu \neq 6.6$.



$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{6.1 - 6.6}{2.5/\sqrt{100}}$$

$$= \frac{6.1 - 6.6}{2.5/10} = \frac{-0.5}{3.9 \times 10 / 2.5} = \frac{-0.5}{0.39/2.5} = -2$$

// $z = -2$ since it lies in region of acceptance, null hypothesis should not be rejected.

confidence interval: $\bar{x} \pm \frac{\sigma}{\sqrt{n}} \times z$. $\Rightarrow 6.1 \pm \frac{2.5}{\sqrt{100}} \times 2.575$

$\Rightarrow 6.1 \pm \frac{2.5}{10} \times 2.575$

$5.46 \longleftrightarrow 6.744$

Since $\bar{x}$ (μ) is 6.6 and is in between the confidence interval, it can be accepted

4. A manufacturer produces gold with a thickness of exactly 1 inch. A customer takes a random sample of 100 golds and finds $\bar{x} = 1.2$ inches and $S = 0.40$ inches should the manufacturer's claim that the bar is exactly 1 inch be rejected? $\alpha = 0.01$

H_0: $\mu = 1$

H_1: $\mu \neq 1$

using confidence interval:-

$$\bar{x} \pm \frac{\sigma}{\sqrt{n}} \times z = 1.2 \pm \frac{0.40}{\sqrt{100}} \times 2.575$$

$$= 1.097 \longleftrightarrow 1.303$$

$\Rightarrow \bar{x} = 1.2$ inches, $\mu = 1.0$

$\therefore$ accept / rejected

using Z-score:

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{1.2 - 1.0}{0.40/\sqrt{100}} = \frac{0.2}{0.4/10} = \frac{0.2 \times 10}{0.4} = 5$$

5 is above the value of 2.575, hence rejected.

5. A company claims that its batteries have a life of atleast 100hrs and sample data is:

$n = 121$, $\bar{x} = 97$ hrs, $s = 3$ hrs, $\alpha = 0.05$

$Z = -1.645$

Soln:- $H_0 : \mu \geq 100$ hrs

$H_1 : \mu < 100$.

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{97 - 100}{3\ 121/\sqrt{121}} = \frac{97 - 100}{\frac{3}{11}} = \frac{-3 \times 11}{3} = -11$$

Null hypothesis should be rejected.

For Z-test to be applied:

→ standard deviation should be known (population)

→ sample size should be large (> 30)

→ should follow standard normal distribution.

Research case1: No verifications
Take as H_1

case 2: verify the research
Take as H_0

manufacturer company => Take as H_0

Hypothesis => Take as H_1

0.01 99%

TTL -2.575 2.575

OTL 2.33 / -2.33

0.5 95%

TTL = -1.96 1.96

OTL = ± 1.645

H_0: $\mu \leq 2000$

H_1: $\mu > 2000$

29th Aug '19

Two sample test:

A hypothesis test that is used to compare two sample groups to determine if they have originated from same population or from different populations.

→ Two sample z test requires:

i) Standard deviation to be known. or the sample size to be larger than 30.

ii) population should follow normal distribution.

iii) Assumptions:

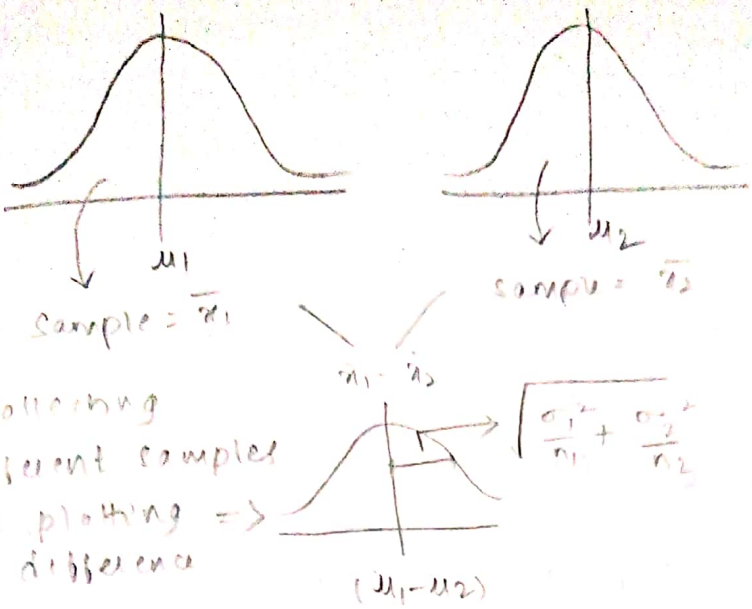
i) The sample size (say n_1 & n_2) of two samples (when) drawn from two populations are large (atleast 30) and the corresponding standard deviations σ_1 & σ_2 are known.

ii) The samples are drawn from two normally distributed populations with standard deviation, σ_1 & σ_2 known.

→ Assume that μ_1 & μ_2 are population means. Our interest is to check a hypothesis on difference between μ_1 and μ_2 that $(\mu_1 - \mu_2)$

→ If $\bar{x}_1$ and $\bar{x}_2$ are estimated means from two samples drawn from two populations, the statistic $(\bar{x}_1 - \bar{x}_2)$ follows standard normal distribution with mean $(\mu_1 - \mu_2)$ and standard deviation $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ where n_1 and n_2 are the sample sizes of two samples

→ corresponding z statistics is $$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$



By collecting different samples and plotting => the difference

1. The marketing specialization student earns atleast 5000 more per month than operation management student.

specialization	sample size	salary	(Pop) σ_s	[Research indirect]
marketing	120	67500	72500	
operations	45	58950	4600	

i. $\mu_1 - \mu_2 \leq 5000$
ii. $\mu_1 - \mu_2 > 5000$

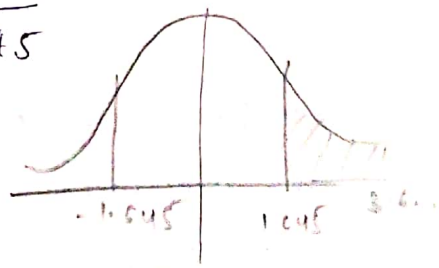
sol:- H_0 : $\mu_1 - \mu_2 \leq 5000$. [No difference]

H_1 : $\mu_1 - \mu_2 > 5000$ [Yes, difference]

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$= \frac{(67500 - 58950) - 5000}{\sqrt{\frac{(72800)^2}{120} + \frac{(4600)^2}{45}}}$$

$$= 38.787.$$



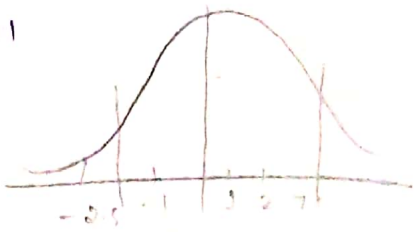
$\alpha = 0.05$ => 1.645 to 1.645

=> Right tailed test

=> Rejected

2. Typing speed on a PC, typing speed of men & women are not equal.

	Men	Women	
$\bar{x}$	65 wpm	68 wpm	$\alpha = 0.01$
s	10 wpm	14 wpm	
n	50 wpm	60	



$H_0: (\mu_1 - \mu) = 0$.

$H_1: (\mu_1 - \mu_2) \neq 0$.

$$Z = \frac{(65-68)-(0)}{\sqrt{\frac{10^2}{50} + \frac{(14)^2}{60}}} = \frac{-3}{\sqrt{\frac{100}{50} + \frac{196}{60}}} = \frac{-3}{1.633} \Rightarrow -1.837$$

$$2.29$$

$\Rightarrow -2.575$ to $+2.575$

$\Rightarrow$ Accepted (i.e), Null hypothesis is not rejected

3. who earns more, married or unmarried?

	Married	unmarried	
$\bar{x}$	639.60	658.20	$\alpha = 0.04$
s	60	90	
n	40	60	

4.

	Married	unmarried	who lives more?
$\bar{x}$	75.5	77.0	$\alpha = 0.01$
s	14	16	
n	140	160.	

5. Are the machine tools manufactured by company X & Y different with regards to how long they last?

	company x	company y	
$\bar{x}$	16.2 weeks	15.9 weeks	$\alpha = 0.08$
s	0.2 weeks	0.2 weeks	
n	40	40.	

3. $H_0: \mu_1 - \mu_2 = 0$

$H_1: \mu_1 - \mu_2 \neq 0$

$$Z = \frac{(63960 - 658 \cdot 20) - (60 - 40)}{\sqrt{\frac{(60)^2}{40} + \frac{(90)^2}{60}}}$$

4. $H_0: \mu_1 - \mu_2 = 0$

$H_1: \mu_1 - \mu_2 \neq 0$

5. $H_0: \mu_1 - \mu = 0$

$H_1: \mu_1 - \mu_2 \neq 0$

7th Sept'19:

Pair sample t-test:

In many cases, we would like to analyse whether an intervention & or treatment) such as training program, treatment for specific illness may have significantly changed the population parameter such as mean, or proportion before and after intervention

→ In paired t-test, the data related to the parameter is captured twice (from the) once before intervention, once after intervention.

Ex1: Body weight before and after attending yoga training.

Ex2: Cholestral levels before and after attending meditation training

* Assume that the mean difference in the estimated parameter before and after treatment is d 'D'.

before	After
x_1	x_2

$$x_1 - x_2 = D$$

and the corresponding standard deviation difference is S_d. [$S_1 - S_2 = S_d$] Let μ_d be the hypothesized mean difference, then statistic

$$t = \frac{D - \mu_d}{S_d/\sqrt{n}} \Rightarrow \frac{D - \mu_d}{S_d/\sqrt{n}}$$

Assume, p will follow normal distribution.

→ statistic 't' follows t-distribution with (n-1) degree of freedom.

P1: A researcher believes that people drink more coffee on monday than other days of week. Based on a sample of 50 coffee drinkers the mean difference was estimated as 14 ml, and the corresponding standard deviation difference is 8.5 ml. Conduct hypothesis test at $\alpha = 0.01$ to claim that people drink on an average 10ml more coffee on monday.

Sol: $n = 50$; $D = 14$; $S_d = 8.5$ ml $\mu_d = 10$ml.

H_0: $\mu \leq 10$.

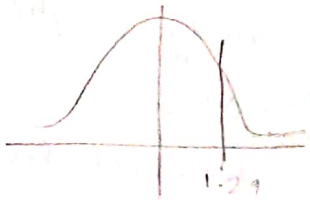
H_1: $\mu > 10$

$$t = \frac{14 - 10}{\frac{8.5}{\sqrt{50}}} \Rightarrow \frac{4}{1.20} = 3.33$$

[α: 0.01 ⇒ 2.33]

Degree of freedom: 49 ↔ 0.1

⇒ 1.2990.



The t-statistic value greater than t-critical value, we can reject null hypothesis.

P2: Difference of average weekly consumption of alcohol (in ml) before & after breakup is 11.5 ml and corresponding sample standard deviation difference is 95.67 ml for 20 candidates. Conduct a hypothesis test to check whether alcohol consumption is more after breakup. (i.e $\mu_d > 0$) at 95% C.I?

Sol:- $\bar{D} = 11.5$ ml, $S_d = 95.67$, $n = 20$, $\mu_d = 0$.

H_0: $\mu_d \leq 0$.

H_1: $\mu_d > \mu_1 - \mu_2 > 0$; $\mu_d > 0$

$$t = \frac{11.5 - 0}{95.67/\sqrt{20}} = \frac{11.5}{95.67/4.4721} = \frac{11.5}{21.39} = 0.537.$$

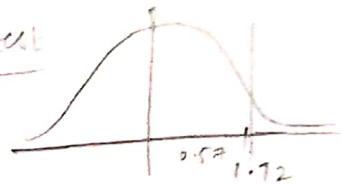
$\alpha = 0.05$

$n = 19$

0.05

19 — ▭ — 1.7291

+ table one tail test



Should not reject null hypothesis.

The t-stats_val < t-critical_val, we cannot reject H_0

P3: Two sample t-test: (with equal variance)

1) Difference in two population means when population standard deviations are unknown and believed to be equal.

→ If you want to estimate difference in two population means when the standard deviations of the populations are unknown

→ Hence, we need to estimate them from sample drawn from these two populations.

→ An additional assumption we make here is the standard deviation of two populations are equal (however unknown)

The sampling distribution of difference in estimated means $(\mu_1 - \mu_2)$ $(\bar{x}_1 - \bar{x}_2)$ follows a t-distribution with degree of freedom $= (n_1 + n_2 - 2)$ with mean $(\mu_1 - \mu_2)$ and standard deviation $= \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$, here s_p is pooled standard deviation

$$s_p^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 1}$$

The corresponding t-statistic is

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 1}} \times \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

P1: A company makes a claim that children who drink their health drink will grow taller than the children who do not drink that health drink.

	sample size	height	S.D.
health drink	80	7.6 cm	1.1 cm
No health drink	80	6.3 cm	1.3 cm

At $\alpha = 0.05$, test whether the increase in height for children who drink that health drink is atleast 1.2 cms.

Sol:- H_0: $\mu_1 - \mu_2 \geq 1.2$ $\quad \mu_1 - \mu_2 \leq 1.2$

H_1: $\mu_1 - \mu_2 < 1.2$ $\quad \mu_1 - \mu_2 > 1.2$

$$s_p = \frac{79 \times (1.1)^2 + 79 \times (1.3)^2}{(80 + 80 - 1)} = 1.45$$

$$t = \frac{(7.6 - 6.3) - 1.2}{\sqrt{1.45\left(\frac{1}{80} + \frac{1}{80}\right)}} \Rightarrow 0.5252$$

t-critical value for one tail test where $\alpha = 0.05$ and degree of freedom = 158 is 1.6546.

→ since t-stat < t-crit, we should not reject null hypothesis.

10th 9 Sept '19.

Sampling distribution:- (1st Aug 19)

The sample follows the normal distribution irrespective of the population.

ex:- Rolling of Dice $\{1,2,3,4,5,6\} = \{\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}\}$
This is called uniform distribution.

ex2:- Rolling 2 dice P(sum = 5) => $\{(1,4)(2,3)(4,1)\}$

* If σ is known and $n > 30$, then z test
* If σ is unknown and $n > 30$, then z test.
* If σ is unknown and $n < 30$, then t-test, because the curve has some errors

10th 11 Sept '19: Two sample t-test with unequal variance:

b) Difference in two population means when population standard deviations are unknown and not equal.

→ If we want to estimate difference in two population means when standard deviations of the two populations are unknown and unequal.

→ we need to estimate standard deviations from the samples drawn from these populations.

→ Then sampling distribution of the difference in estimated means $(\bar{x}_1 - \bar{x}_2)$ follows t distribution with mean $(\mu_1 - \mu_2)$ and standard deviation,

$$s_{\bar{d}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

→ The corresponding degree of freedom is given by

$$df = \frac{s_u^4}{s_d}$$

$$df = \left\lfloor \frac{Su^4}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1-1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2-1}} \right\rfloor$$

Here the symbol $\lfloor \rfloor$ implies rounding down to nearest integer $[25.5 \rightarrow 25]$.

→ The t-statistic for testing two populations with unequal variance is given by

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

P1: A researcher is interested in finding the average duration of marriage based on educational qualification of couples. Two groups were considered for study. Group1 consisted of couples with no bachelor's degree and group2 consisted of couples with bachelor degree or higher.

Group	sample size	duration of marriage	S.D
No degree	120	10.1 yrs	2.4 yrs
degree	100	9.5 yrs	3.1 yrs

At $\alpha = 0.05$, test whether the average duration of marriage is more for couples with no bachelor's degree as compared to couples with bachelor's degree

H_0: $\mu_1 \leq \mu_2 \Rightarrow \mu_1 - \mu_2 \leq 0$.

H_1: $\mu_1 > \mu_2 \Rightarrow \mu_1 - \mu_2 > 0$

$$t = \frac{10.1 - 9.5}{\sqrt{\frac{(2.4)^2}{120} + \frac{(3.1)^2}{100}}} = \frac{0.6}{\sqrt{0.048 + 0.0961}} = \frac{0.6}{0.379} = 1.58$$

$(0.379)^2 \Rightarrow 0.0207$

$$\alpha = 0.05 \; ; \; df = \frac{\left(\frac{(2.4)^2}{120}\right)^2 + \left(\frac{(3.1)^2}{100}\right)^2}{\frac{120-1}{} \qquad \frac{100-1}{}} = [184.33]$$

The critical value of t for $\alpha = 0.05$ & $df = 184$ is 1.6531

→ Null hypothesis should not be rejected

∴ since t statistic is less than critical value of t i.e the difference in duration of marriage b/w two groups is less than equal to 0.

12th Sept' 19:

Non parametric tests

→ when population is not known to be following distributions, non parametric tests can be applied.

→ No summary statistics are required.

→ Chi square test (χ^2).

→ If the distribution follows χ^2 distribution, then chi square test is applied.

→ We said, population distribution is unknown, the sample which we are testing if it follow chi square, then test can be applied.

parametric tests

z-test, t-test, F-test.

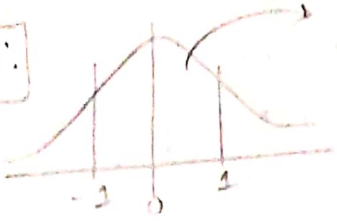
→ The parameter used here was taking sample mean & estimating population mean

→ Even proportion is taken and estimation of population mean can be done

→ common assumption: population is coming from normal distribution But sample may/ maynot if < 30 ⇒ T-distribution if > 30 ⇒ normal

Chisquare (χ^2) test: Let say the sample follows

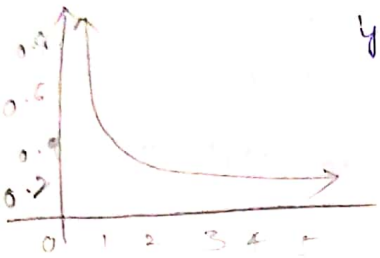
a) Distribution: $N(0,1)$



If we pick a point from 0 to 1, the probability for choosing the point near to 0 is high

Ex: $0.5 \Rightarrow x_1$

$x_1^2 = (0.5)^2 = 0.25 = Q_1$



If degree of freedom = 1, i.e. if sample size is less, it follow the curve.

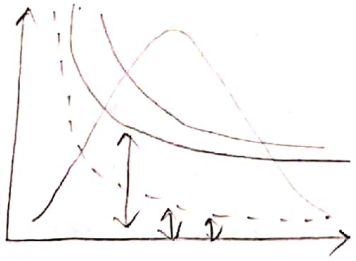
$Q_2 = 1.25$

→ If x_1, x_2 are taken to be $0.5 \& 1 \Rightarrow 0.25 \& 1$.

If 10 values are taken, more than 5 values results within the same range and some of them differ largely as quoted.

$$(0.1)^2 + (2)^2 + 0.1)^2 + (-3)^2$$

→ As sample size increases, the graph changes and the probability of 4, 5.. increases.



* As degree of freedom increases, the curve tends to be changed.

b) Test: Goodness fit test:-

Generally- If die is rolled 90 times, then each number should be 15 times. It could be accepted if 12, 13, 14, But if 67 times 'Number '6'' is dropped, it should not be accepted To test such kind of problems, goodness test fit is applied. An application of χ^2

Ex1:-

	M	T	W	Th	F	S
(i) Expected	10	10	15	20	30	15
(No) observed	30	14	34	45	57	20 [200]
(No) expected	20	20	30	40	60	30
ob- exp	$(10)^2$	$(6)^2$	$(4)^2$	$(5)^2$	$(3)^2$	$(10)^2$
	100	36	16	25	3	100

$$\chi^2 = x_1^2 + x_2^2 + x_3^2 + x_4^2 + x_5^2 + x_6^2$$

* χ^2 follows chi² distribution with degree of freedom = 5

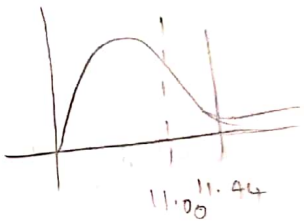
χ^2 is the error so if the error range is acceptable, it is said to be a good fit.

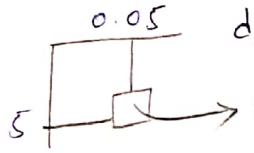
$$\chi^2 = \frac{(\text{Observed value} - \text{Expected val})^2}{\text{Expected val}}$$

$$= \frac{100}{20} + \frac{36}{20} + \frac{16}{30} + \frac{25}{40} + \frac{9}{60} + \frac{100}{30}$$

$$= 5 + \ldots + 3.33$$

$$= 11.4 \quad \text{// The error.}$$





$d_f = 5$; $\alpha = 0.05$ on χ^2 table:

check the value, its 11.00.

Since the error is more than critical value, the distribution is not correct

16th Sept'19:

Q1:- Hanuman Airlines operated daily flights to several Indian cities. One of the problems Hanuman airline faces is the food preference of the passengers Captain cook, the operations manager of Hanuman Airlines believes that 35% of passengers prefer vegetarian food 40% prefer non veg food, 20% low calorie food 5% requested for diabetic food. A sample of 500

passangers were taken to analyse the food preferences and the data is shown below.

Foodtype	veg	non veg	low cal	diabetic
no. of pass.	190	185	90	35 (O)
original	35%	40%	20%	5%

conduct a χ^2 test, to check whether captain cook's belief is true? at $\alpha = 0.05$

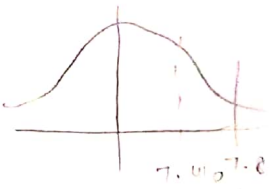
values:	175	200	100	25 (E)

H_0: cook's belief is true.

H_1: cook's belief is not true.

$$\chi^2 = \frac{(obs - exp)^2}{exp} \Rightarrow 1.22 + 1.25 - 1 + 4$$

$$= 7.410.$$



$\alpha = 0.05$; $d_f = 3 \Rightarrow 7.814$

Null hypothesis should not be rejected.

18th Sept' 19.

Central limit theorem for proportionate:

Sampling distribution of proportions $\bar{p}$ for a large sample follows an approximate normal distribution with mean P (population proportion) and standard deviation $\sqrt{\frac{P(1-p)}{n}}$

Hypothesis testing for proportion of population

Z test proportion:

According to central limit theorem of proportions the sampling distribution $\bar{p}$ for a large sample

follows approximate normal distribution. Then z-statistic is $z = \frac{\bar{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$. To calculate standard deviation $\sqrt{\frac{p(1-p)}{n}}$, we need the knowledge of p. However we can use the value of $\bar{p}$ estimated from large samples. One of the thumb rules used is the value of $n \times \bar{p} \times (1-\bar{p}) \geq 10$ to use z-stat i.e $z = \frac{\bar{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$.

1. According to a study exactly 12% of gift cards purchased from e-commerce portals are never used. One manager of ecommerce company wanted to test whether this claim is true. She collected data of 250 gift card purchases and found that 22 gift cards were not still used till its expiry date.

i) Conduct appropriate hypothesis test to claim that 5%. to check whether claim that exactly 12% gift cards are never used is true/not?

ii) calculate 95% confidence intervals for proportion of gift cards that are not used

i) $n = 250$. $H_0 : P = 0.12$

$H_1 : P \neq 0.12$

$\bar{p} = \frac{22}{250} = 0.088$

$P = 12\% = 0.12$

$z = \frac{\bar{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{0.088 - 0.12}{\sqrt{\frac{0.12(1-0.12)}{250}}} = -1.557$

$\alpha = 0.05 \Rightarrow \pm 1.96 \quad -1.96 < -1.57 < 1.96$.

No rejection of null.

ii) $\bar{p} \pm z\sqrt{\frac{p(1-p)}{n}}$

$= 0.088 \pm 1.96\sqrt{\frac{0.12(1-0.12)}{250}} = 0.0528, 0.123$

$0.0528 < 0.12 < 0.1231$

No rejection of null.

Hypothesis test for difference in population proport

Two sample z test for proportion:

When proportions are estimated from large samples then the sampling distribution of proportio follows a normal distribution according to CLT.

Let $\bar{p}_1$ & $\bar{p}_2$ be estimated values of proportions from large samples.

The difference $P_1 - P_2$ is the hypothesized diff in population proportion. When null hypothesis is $p_1 = p_2$ (i.e) $H_0 : P_1 = P_2$, then test statistic is

$$z = \frac{(\bar{P}_1 - \bar{P}_2) - 0}{\sqrt{\bar{P}(1-\bar{P})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where $\bar{p} = \frac{n_1\bar{P}_1 + n_2\bar{P}_2}{n_1 + n_2}$ is the pooled estimate for proportion.

23rd Sept'19.

F- distribution: (Fisher's distribution) Ronald Fisher.

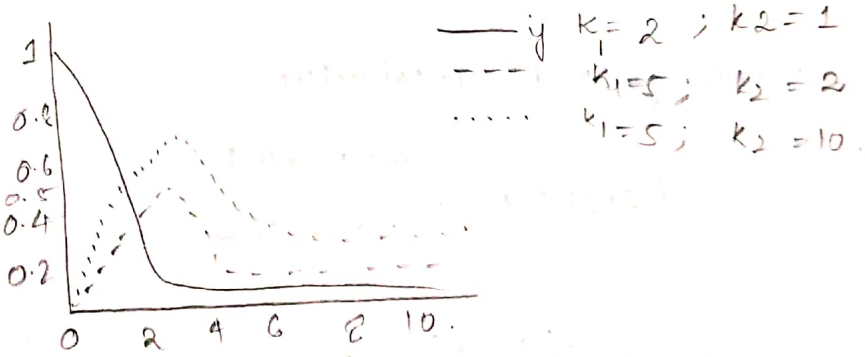
$F = \frac{Q_1}{Q_2}$ (The ratio of chi² is F)

It is the ratio of two χ^2 distribution. Let Q_1 and Q_2 be two independent χ^2 distributions with k_1 and k_2 degree of freedom respectively, then random variable X defined as $\boxed{X = \frac{Y_1/k_1}{Y_2/k_2}}$ is an F-distribution.

probability density function of F-distribution:

$$f(x) = \frac{\Phi\left(\frac{k_1+k_2}{2}\right)\left(\frac{k_1}{k_2}\right)^{k_1/2}}{\Psi\left(\frac{k_1}{2}\right)\Psi\left(\frac{k_2}{2}\right)} \times \frac{x^{\frac{k_1}{2}-1}}{\left(1+\frac{k_1 x}{k_2}\right)^{\frac{k_1+k_2}{2}}}$$

Using the above formula, the critical value is found, but we do not integrate the complex functions instead use the table values.



F distribution is used in analysis of variance test.

Properties of F distribution:

1. Mean of F distribution is $\frac{k_2}{(k_2-2)}$ for $k_2 > 2$
2. Standard deviation of $\sqrt{\frac{2k_2^2(k_1+k_2-2)}{k_1(k_2-2)^2(k_2-4)}}$ for $k_2 > 4$

3. F distribution is non symmetrical and the shape of distribution depends upon the values of k_1 & k_2.

4. F distribution is used in analysis of variance to test mean values of multiple groups.

P1. Conduct a F test on following samples.

Sample 1	variance = 109.63 n_1 = 41
Sample 2	variance = 65.99 n_2 = 21

With a two tailed F test, we just want to know if the variances are not equal to eachother.

1. H_0: $\sigma_1^2 = \sigma_2^2$

   H_1: $\sigma_1^2 \neq \sigma_2^2$.

2. calculate your F critical value.

$$F_{statistic} = \frac{\text{variance 1}}{\text{variance 2}}$$

$\text{Variance} = \frac{\Sigma(X - \bar{x})^2}{n-1}$. Since this seems to be the same as x^2, the ratio of x^2 is F.

$$\Rightarrow \frac{109.63}{65.99} = 1.66.$$

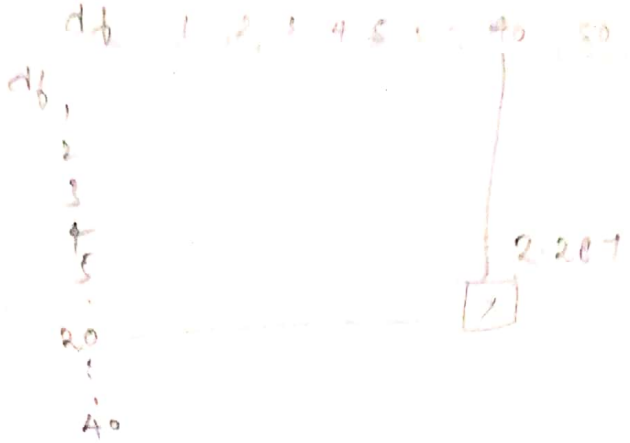
3. calculate degrees of freedom:

Sample 1 = 41 − 1 = 40.

Sample 2 = 21 − 1 = 20.

Let $\alpha = 0.05$ => Two tailed test = 0.025 · 0.025

4. find the critical value using F-table.



At $\alpha = 0.025$; $F(40,20) = 2.287$

· Since, F stat is 1.66 less than F critical 2.287, so we cannot reject null hypothesis.

25 sept'19. Analysis of variance (ANOVA).

On many situations, we need to compare conduct hypothesis to compare mean values for more than two groups (samples) created using a factor

→ for example, a marketer may like to understand the impact of 3 different discount value such as 10%, 0%, 20% on the average sales. when we to compare the impact of factor on mean or more than two groups, simultaneously

such as two sample t-test

Not ideal approach since they can result in type 1 and type 2 error.

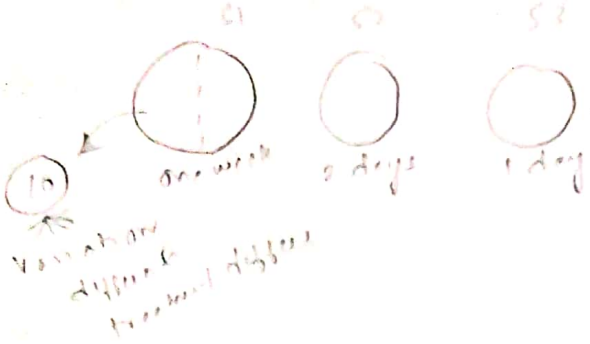
sample 1 sample 2 sample 3

μ_1 μ_2 μ_3

$\{\mu_1 = \mu_2 \quad \mu_2 = \mu_3 \quad \mu_3 = \mu_1\}$ => Type 2 error.

The reason behind the error is,

* On ANOVA, one objective is to verify whether the variation due to treatment is different from the variation due to randomness.



Multiple t-tests for comparing several means:

Null Hypo: A: $\mu_0 = \mu_{10}$ $H_1: \mu_0 \neq \mu_{10}$

B: $\mu_{10} = \mu_{20}$ $\mu_{10} \neq \mu_{20}$

C: $\mu_{20} = \mu_0$ $\mu_{20} \neq \mu_0$.

Anova H_0: $\mu_0 = \mu_{10} = \mu_{20}$.

P(A): Retain H_0 / in test A

P(B): Retain H_0 / in test B

P(C): Retain H_0 / in test C

$\therefore P(A) = P(B) = P(C) = 1 - \alpha$.

α is the error 0.05

$P(A) = P(B) = P(C) = 0.95$

$P(A \cap B \cap C) = 0.8573$

↳ individual probability $= 0.95$ whereas the combined probability $= 0.85$.

Because of this incorrectness, α & β errors, thus multiple t-tests cannot be performed.

One way Analysis of Variance:

1-way ANOVA is appropriate under following:

1. We would like to study the impact of a single

treatment (factor) at different levels on a continuous response variable. For example: The variable price discount is the factor and 0%, 10% & 20% price discounts are different levels (3 levels). Different levels of discount are likely are have varying impact on sales of the product; where sales is the outcome variable.

2. In each group, population response variable follows a normal distribution and the sample chosen using random sampling.

3. The population variances for different groups are assumed to be same. (i.e) variability in response variable values within different groups is same.

Setting up analysis of variance:

Assume that we would to study the impact of factor (discount) with K levels on continuous variables (quantity) then the null & alternate hypothesis for one way ANOVA

$H_0: \mu_1 = \mu_2 = \mu_3 \cdots \mu_K$

H_1: Not all μ's are equal.

P1. Using the following data, perform oneway 'anova' using $\alpha = 0.05$

group1	group2	group3
61	23	56
45	43	76
33	21	74
45	43	87
67	45	56

$\mu_1 = 48.2$ $\mu_2 = 35.4$ $\mu_3 = 69.8$

$$\text{Variance} = \sum \frac{(X - \bar{X})^2}{n-1}$$

$$val1 = \frac{(51-48.2)^2 + (45-48.2)^2 + (33-48.2)^2 + (45-48.2)^2 + (67-48.2)^2}{4}$$

$$= \frac{612.8}{4} \Rightarrow 153.2$$

Variance within group

$$val2 = \frac{(23-35.4)^2 \ldots)}{4} = \frac{515.2}{4} = 128.8$$

$$val3 = \frac{(36-69.8)^2 + \ldots}{4} = 183.2$$

$$MSD_g = \frac{153.2 + 128.8 + 183.2}{3} = 155.07$$ (mean squared deviation within group)

→ Variances across groups:

$$M_g \rightarrow \begin{matrix} 48.2 \\ 35.4 \\ 69.8 \end{matrix} \Rightarrow \frac{M_1 + M_2 + M_3}{3} = \frac{48.2 + 35.4 + 69.8}{3} = 51.13$$

$$\Rightarrow (48.2-51.13)^2 + (35.4-51.13)^2 + (69.8-51.13)^2$$

$$\Rightarrow 8.58 + 247.42 + 348.57$$

$$SS_{mean}: \frac{8.58 + 247.42 + 348.57}{2}$$

$$\Rightarrow 302.29.$$

$$MSB = 302.29 \times 5 = 1511.45.$$

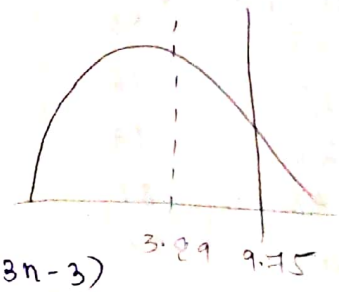
Variance across / variance within

$$F_{stat} = \frac{MSB}{MSW} = \frac{1511.45}{155.07} = 9.75$$

→ MSB: degree of freedom
(3 means => 2)

→ MSW: degree of freedom
15 values 3 groups
15 − 3 => 12 (3n − 3)

F(2, 12)



→ At $\alpha = 0.05$ & F(2, 12); value = 3.89

Reject Null hypo

26th Sept'19

Using one way ANOVA, perform $\alpha = 0.01$ for the data

n	mean	sd
30	50.76	10.45
30	45.32	12.76
50	53.67	11.46.

H_0: $\mu_1 = \mu_2 = \mu_3$

H_1: $\mu_1 \neq \mu_2 \neq \mu_3$

Variance within group :-

$V = (sd)^2 = (10.45)^2 + (12.76)^2 + (11.46)^2$

$= 109.2 + 162.82 + 131.56$

$= 403.40.$

Total variance $= \frac{403.4}{3} = 134.53 \Rightarrow MSW.$

variance across groups:

grand mean $= \frac{50.76 + 45.32 + 53.67}{3} = 49.75.$

variance : $\frac{(50.76 - 49.75)^2 + (45.36 - 45.32)^2 + (53.67 - 49.75)^2}{2}$

$= \frac{35.26}{2} \Rightarrow 17.62.$

$MSB = 17.62 \times 30$

$= 528.75$

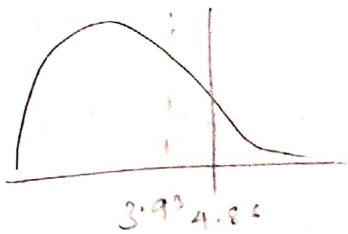
$F_{stat} = \frac{528.75}{134.53} \Rightarrow 3.93.$

Degree of freedom:

MSB $\Rightarrow 3 - 1 = 2$

msw $\Rightarrow 90 - 3 \Rightarrow 87$

$(2, 87) \Rightarrow 4.86.$



* how is the mean verification related to the calculations above ?

→ Even confidence interval is the variation only because, there were interval

→ Even now, we are performing the group variation.

3. A clinical psychologist has ran a between subjects experiment, comparing two treatments for depression (cognitive behavioural therapy) CBT and client centered therapy - CCT against a control condition. subjects were randomly assigned to the experimental conditions. After 12 weeks the subjects depression scores were measured using CESD depression scale. The data are summarized as

	n	mean	std
control	40	21.4	4.5
CBT	40	16.9	5.5
CCT	40	19.1	5.8

H_0: $\mu_1 = \mu_2 = \mu_3$

H_1: Not equal.

MSB 9: $(4.5)^2 + (5.5)^2 + (5.8)^2$

$= 20.25 + 30.25 + 33.64$

$= 84.14.$

Average $= \frac{84.14}{3} = 28.04$

Across group: $\frac{21.4 + 16.9 + 19.1}{3}$

grand mean $= 19.133$

variance $\frac{(21.4 - 19.33)^2 + (16.9 - 19.33)^2 + (19.1 - 19.33)^2}{2}$

$= \frac{4.28 + 4.86 + 0.0529}{2} = \frac{10.239}{2} = 5.116$

5.90

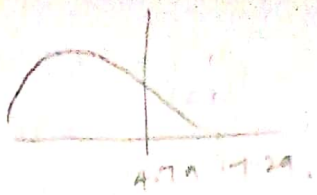
$MSB = 5.116 \times 40$

$= 204.65$

$F_{stat} = \frac{204.65}{28.04} = 7.29$

degree of freedom: (2, 117) => 4.79

// Reject null hypothesis.



1. An education researchee is comparing 4 diffcent algebra curicula. 8th grade students are randomly assigned to one of the four groups. Their state achievement test scores are compared at the end of the year. use the appropriate statistical procedure to determine whether the curicula differ with respect to math achievements.

$\alpha = 0.05$

	n	mean	sd
1	50	170.5	14.5
2	50	168.3	12.8
3	50	169.6	17.7
4	50	172.8	16.8

variance within group:

$$\Rightarrow (14.5)^2 + (12.8)^2 + (17.7)^2 + (16.8)^2$$

$$\Rightarrow 210.25 + 163.84 + 313.29 + 282.24$$

$$\Rightarrow 969.24$$

$$AvV \Rightarrow \frac{969.24}{4} = 242.405$$

variance across group:

grand mean = 170.5 + 168.3 + 169.6 + 172.8

= 170/3/4 = 42.575

$$\text{dvariance}: \frac{(170.5 - 170.3)^2 + (168.3 - 170.3)^2 + (169.6 - 170.3)^2 + (172.8 - 170.3)^2}{3}$$

(42.75, 42.572, 42.752, 42.572)

$$= \frac{0.04 + 4 + 0.49 + 6.25}{3} = 2.08$$

$$msb \Rightarrow 2.08 \times 50 \Rightarrow$$

$$= 104.16.$$

An-4

$f_{stat} = \frac{104.16}{242.4} \Rightarrow 0.42$ 200-4

degree of freedom: F(3, 196)

3rd Def'n: Correlation Analysis:

→ The relation between two variables.

→ correlation and covariance:

Covariance indicates the linear relationship between the variables.

correlation measures the strength and direction of linear relationship. To measure the correlation, we have a coefficient called pearson coefficient denoted by r.

Pearson coefficient (r) => pearson product moment correlation.

Pearson coefficient: It is used for measuring the strength and direction of linear relationship between two continuous random variables X and Y.

Ex: Consider two variables, the average call duration (y) and the age (x), we may like to know whether avg call duration is related to age of callers.

Let (x_i) be different values of X & y_i be different values of Y.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$$Z_x = \left(\frac{x_i - \bar{x}}{\sigma_x}\right) \qquad Z_y = \left(\frac{y_i - \bar{y}}{\sigma_y}\right)$$

$$r = \frac{\sum_{i=1}^{n} Z_x Z_y}{n}$$

$$\Rightarrow \frac{\sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sigma_x}\right)\left(\frac{y_i - \bar{y}}{\sigma_y}\right)}{n}$$

If taken for a sample:

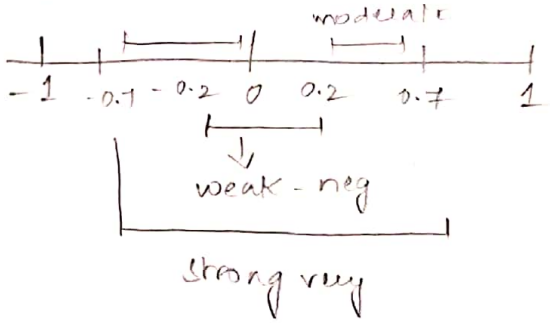
$$\frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(n-1) s_x \cdot s_y} \Rightarrow \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(n-1)\sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}} \cdot \sqrt{\sum_{i=1}^{n} \frac{(y_i - \bar{y})^2}{n-1}}}$$

$$\Rightarrow \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$ Pearson.

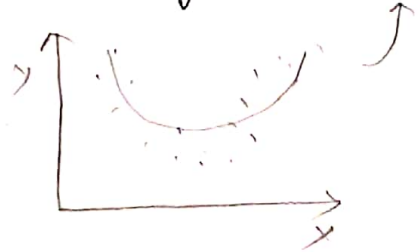
$$r = \frac{cov(x, y)}{\sigma_x, \sigma_y}$$

properties of pearson correlation coefficient:

1. The value of 'r' lies b/w -1 and +1.
2. High absolute value indicates strong relation.



3. $r^2 = R^2$.
4. Pearson correlation coefficient value may be 0, even there is strong non linear relationship b/w x & y.



P1: The average share price of two companies over past 12 months are shown below. Calculate 'r'

P	X	Y	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
161.02	274.58	219.50	-9.7	-16.6	94.08	275.96
25.09	287.96	242.92	3.68	6.82	13.5	46.5
51.4	290.35	245.90	6.07	9.80	36.84	96.04
=245.6	$\bar{x}$ = 284.286	$\bar{y}$ = 236.10.			144.42	418.1

$$\Rightarrow r = \frac{245.6}{\sqrt{144.42} \cdot \sqrt{418.1}} = \frac{245.6}{12.01 \times 20.44}$$

$$= \frac{245.6}{245.57} = 1.00$$

$$\frac{245.6}{244.8}$$

22nd oct'19.

Regression:

perform linear regression on striking rate and no. of sixes.

2nd
23 oct 19:

Simple Linear regression:

Simple linear regression is a statistical technique for finding the existence of an association relationship between dependent variable and independent variable.

→ SLR implies that there is only one independent variable in the model.

Functional form of SLR:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

for a dataset with n observations (X_i, Y_i) where $i = 1, 2, \ldots n$

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad \varepsilon_i = Y_i - \beta_0 - \beta_1 X_i$$

where Y_i is the value of.

Estimation of parameters using ordinary least squares:

$$\sum_{i=1}^{n} \varepsilon_i = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)$$

In ordinary least squares, the objective is find the optimal value of β_0 & β_1 that will minimize the sum of squares error (SSE)

$$SSE = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

To find the optimal values of β_0 & β_1 that will minimize SSE, we have to equate the partial derivate with respect to β_0 & β_1 to 0.

$$\frac{\partial SSE}{\partial \beta_0} = 0 \Rightarrow \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

$$\Rightarrow \sum_{i=1}^{n} 2(-1)(Y_i - \beta_0 - \beta_1 X)$$

$$= -\sum_{i=1}^{n} 2(Y_i - \beta_0 - \beta_1 X) = 0$$

$$\Rightarrow 2\left(n\beta_0 + \beta_1 \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} Y_i\right) = 0$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x} \quad \text{---(1)}$$

$$\frac{\partial SSE}{\partial \beta_1} = \sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 x_i) \cdot x_i = 0$$

$$= -2 \sum_{i=1}^{n} (y_i x_i - \beta_0 x_i - \beta_1 x_i^2) = 0 \quad \text{---(2)}$$

substitute β_0 in (2)

$$= -2 \sum_{i=1}^{n} (y_i x_i - (\bar{y} - \beta_1 \bar{x}) x_i - \beta_1 x_i^2) = 0$$

$$= -2 \sum_{i=1}^{n} (x_i y_i - x_i \bar{y} + \beta_1 \bar{x} x_i - \beta_1 x_i^2) = 0$$

$$= \sum_{i=1}^{n} (x_i y_i - x_i \bar{y}) - \beta_1 \sum_{i=1}^{n} (x_i^2 - \bar{x} \cdot x_i) = 0$$

$$\beta_1 \Rightarrow \frac{\sum_{i=1}^{n} (x_i y_i - x_i \bar{y})}{\sum_{i=1}^{n} (x_i^2 - x_i \bar{x})} \Rightarrow \sum_{i=1}^{n} \frac{x_i (y_i - \bar{y})}{x_i (x_i - \bar{x})}$$

Since, $\sum_{i=1}^{n} (\bar{x}\bar{y} - y_i \bar{x}) = 0$ & $\sum_{i=1}^{n} (\bar{x}^2 - x_i \bar{x}) = 0$

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

$$\beta_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{cov(x,y)}{Var(x)}$$

→ linear relationship - Pearson coefficient nonzero.

→ The above should be true to perform linear regression.

e) Does bodyweight affect cost of treatment from Dataset?

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$.

Residual : Original - predicted

Work in.

29 Oct '19.

Validation of simple linear Regression model :-

1. coefficient of determinant (R^2)
2. Hypothesis test for regression coefficient β_1
3. ANOVA (MLR)
4. Residual Analysis
5. Outlier Analysis.

1. coefficient of determinance :

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

variation in Y_i = explained variance + not explained.

$$R^2 = \frac{\text{explained variance}}{\text{total variance}}$$

$$\sum_{i=1}^{n} (x_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y} - \bar{y})^2 + (y_i - \hat{y})^2 \sum_{i=1}^{n}$$

Total variation = explained + not explained

Sum of squares of total variation (SST) = Sum of squares of variation explained by Regression + Sum of squares of errors

$$R^2 = \frac{SSR}{SST}$$

$$R^2 = \frac{SST - SSE}{SST} = \frac{SST}{SST} - \frac{SSE}{SST}$$

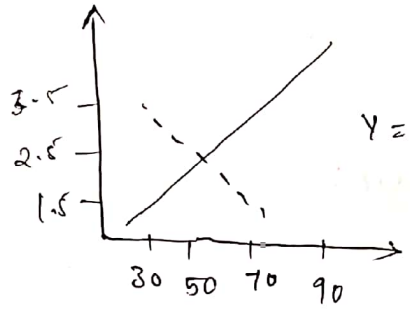
$$R^2 = 1 - \frac{SSE}{SST}$$

R^2 is the proportion of variation in response variable y explained by regression model.

Properties:-

- R^2 lies between 0 & 1
- R^2 higher value of R^2 implies better fit.
- $R^2 = r^2$ where r = pearson coefficient.

Hypothesis test on β_1:



$y = mx + c$

for same x & y array, 'm' is the one which changes the line.

Hence β_1 is the changing factor

Case1: $H_0 : \beta_1 = 0$ There is no correlation.

β_1 can't be always 0.

Ex: If for every 10 runs, 5000 increase, then

$\beta_1 = 5000$.

$$t = \frac{\bar{x} - u}{\sigma/\sqrt{n}}$$

$\Rightarrow$ errors follow normal distribution

$$Y_i = \beta_0 + \beta_1 X + \varepsilon_i$$

If ε_i follows normal distribution, then Y_i also follows normal distribution.

since β_1 also follows normal distribution, their hypothesis test can be performed.

$$t = \frac{\beta_1 - u}{\sigma/\sqrt{n}} \Rightarrow \frac{\beta_1 - 0}{se}$$

$$se = \frac{\sqrt{\sum_{i=1}^{n} (y_i - \hat{y})^2 / n-2}}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

t-value corresponding p-value is found, and is compared with α.

---

* Not a part of SLR.

21st oct 19: Z test:

A manufacturer produces thickness of 1 inch. A random sample mean is 1.2 & SD 0.4. Manufacturer claims that bolt exactly 1 inch be rejected?

Earlier $Z_{stat} = \frac{\bar{x} - u}{\sigma/\sqrt{n}}$

In python, from statsmodel.stats import weightstats as stest

Stests.zstat_generic(xbar, value2=0, std_diff=se, alternative='two sided', diff=mu).

=> 'larger, smaller'

=> If 'pvalue > α', null hypothesis should not be rejected.

Multiple Linear Regression:

→ Independent variables are more than 1.

Assumptions:-

→ All the X_1, X_2, X_3 should be linearly dependent on Y

→ X_1 & X_2 should not have co-relation. If it has, it's a multicollinearity problem.

→ If there are many no. of x_1, x_2, x_3, then some of them should be removed to avoid overfitting.

→ Those methods are forward selection and backward elimination.

On backward elimination:

According to p value of each independent variable is checked.

→ If Pvalue of attribute $< \alpha$, then the feature is relevant

→ If pvalue of attribute $> \alpha$, then the feature is not contributing.

$$Y = \beta_0 + \beta_1 X + \beta_2 X_2 \ldots \beta_n X_n.$$

After finding the OLS, printing the summary gives the values for each feature like pvalue, t etc

→ If the pval is greater than α, remove that feature and again fit the model. This process is repeated.

29 Oct '19

Avoiding overfitting:

To avoid overfitting some of the parameters should be removed. To know which among them is to be removed MALLOW's C_p is used.

$$C_p = \left(\frac{SSE_p}{MSE_{full}} \right) - (N - 2p)$$

where

N = no. of observations

p = parameters in model (including constant)

for each p value, some constant is obtained, for example if P=1 : 101, P=2 : 80 P=3 : 60 P=4 : 20

P=5 : 60 P=6 : 80.

From P=5; it starts overfitting. The point at which the p value minimizes the set of factors are to be considered. (P-CP) difference minimum should be there.

P1. Transformations :-

While finding the correlation between X & Y, if there's weak correlation, the X can be transformed into $\log X$, $\sqrt{X}$, $\frac{1}{X}$ etc. Then the correlation would be better and hence regression can be done

Advantages :-

1. Low R^2 can be overcomed
2. Residuals do not follow normal distribution. In this case, transformation can be done.
3. Residuals are not patternised.
4. If there's non-linear relationship, transformation helps.

MLP problems :-

1. Sold price increases by atleast $1000 for every unit in baking striking rate

$H_0 : \beta_1 \leq 1000$

$H_1 : \beta_1 > 1000$

$$t = \frac{2086 - 1000}{983.64}$$ [according to projected values]

$$= 1.104$$

corresponding P_value = 0.1357.

Let $\alpha = 0.05$ $0.1357 > 0.05$

P_val > α; null hypothesis not rejected.

1st Nov 19

T-test :

statsmodels.stats.weightstats.ttest_ind(x1, x2=None, value=0, alternative='two-sided', usevar='pooled', weights=(None, None))

- The statsmodels has the built in library for performing t-test

Shapiro test for normality :

To perform parametric tests we should know whether the sample/pop follows normal distribution.

→ population may be normal distribution, but sample may or may not variate

→ If sample has 3 * SE, then acceptable/not

→ But by just observation, we cannot assess the distribution exactly. Hence the test is Shapiro test.

A module of shapiro(s), gives the (0.9, 0.593) (p)

If $P > 0.5$ (α), then normal distribution (H0).

Bartlett's test :-

Bartlett's test for homogenity of variances is used to test that variances are equal for all samples.

ex: stats.bartlett([1,2,3,4,5], [3,4,6])

pvalue = 0.95 [variance is equal in all samples]

ex2: stats.bartlett([1,2,3,4,5], [3,40,80])

pvalue = 0.000006 [variance is not equal]

1 Nov'19.

Validation of MLR:

1. coefficient of multiple determination (R^2) and adjusted R square
2. t test b/w response variable and individual explanatory variable at given significance level.
3. F test to check the statistical significance of overall model
4. Conduct residual analysis
5. Multicollinearity.

1. coefficient of multiple determination:

If no. of variables are added/ increased,

i) If relevant = R^2 increases

ii) If irrelevant = R^2 const or decreases.

Hence, if irrelevant then the model has unnecessary variable.

Therefore instead of R^2, we go for adjusted R^2 score.

$$\text{adjusted } R^2 = 1 - \frac{SSE}{SST} \Rightarrow 1 - \frac{SSE/(N-k-1)}{SST/(N-1)}$$

-> The higher the adjusted R^2, the better the model is.

2. test b/w responsive variable....

(hypothesis test for $B_1 \ldots B_N$)

H_0: There is no relationship b/w X_1 & Y

H_1: There is a relationship b/w X & Y.

$H_0: B_1 = 0$

$H_1: B_1 \neq 0$.

The t statistic value of CTRP and P are 10.25 & 10.34 and corresponding pval = 0.000.

Hence at $\alpha = 0.05$, reject null hypothesis.

3. Validation of overall regression model F-test:

$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$

$H_1 : \beta_1 \neq \beta_2 \neq \beta_3 = 0$.

Instead of mean as in ANOVA, here we check for coefficients ($\beta_1 .. \beta_n$)

For an example, F-statistic is $\frac{86.62}{\alpha}$ ($\alpha + 14\, e^{-14}$), hence is almost 0. F-statistic p-value is $2.14\, e^{-14}$ and is 0.

→ For $P < \alpha$, reject null hypothesis.

4. Residual Analysis for Normality, homoscedasticity. (equally distributed overall the plot)

→ shapiro test can also be performed for checking normality with residuals as input.

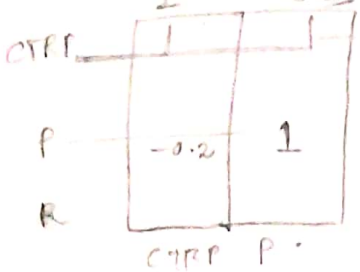
→ probplot (p-plot): The probality expected line is given and the predicted values are projected

5. check for the presence of multicollinearity:

→ IVF value is used.

→ Heat maps can also be used to check multicollinearity

1. colinearity coefficient pearson should be found for all individual variables pairwise.



→ weakly correlated
→ Should be weakly to avoid multicollinearity.

6th Nov 19:-

Chi² test in python:-

H0: Gender and voting preferences are independent

→ In pandas, crosstab function gives the summary.
→ In scipy, contingency, chisquare are used.
→ ravel() serializes the value from 2D-to 1D

ANOVA:- If you have more than 2 t-test

Anova doesnot tell which group mean is different.

Tukey hsd test is used to tell the group:

It compares each of the group.

1	2	diff	false
1	3		
1	4		
2	3	0.36	
2	4	no diff	True
3	4		

It does individual t-test again.

Post-hoc test: Fishers, least square