Two sample Z-test

- This test is used to check whether the two samples taken from the same population or not
- two samples are same or not

- Two sample Z-test assumptions

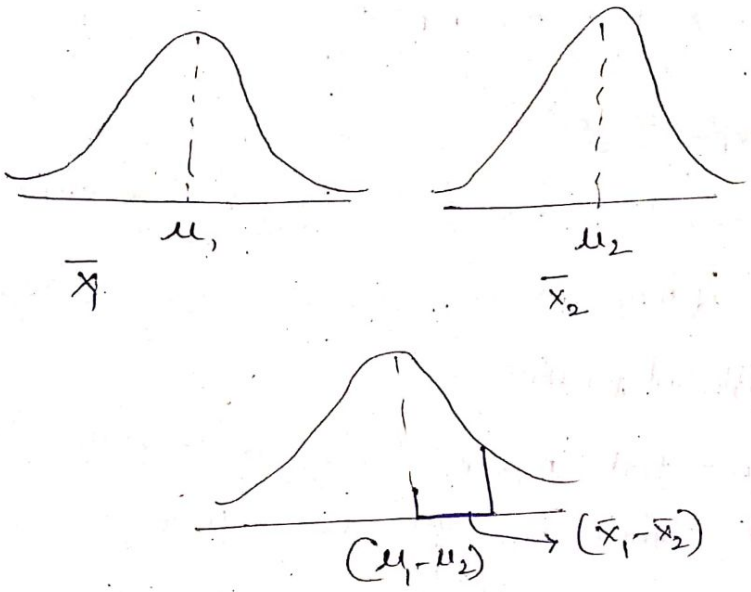
29/08/19 Two sample Z-test:

- A hypothesis test that is used to compare two sample groups to determine if they have originated from same population or from two different population.
- two Sample Z-test requires the standard deviation to be known & the sample size to be larger than 30 and should Population should follow normal distribution.

Two Sample Z-test assumptions:

- The sample size (say n_1 & n_2) of two samples drawn from two populations are large (say atleast 30) and corresponding standard deviations σ_1 & σ_2 are known.
- Samples are drawn from 2 normally distributed populations with standard deviations with σ_1 & σ_2 known.
- Assume that μ_1 & μ_2 are population means. our interest is to check a hypothesis on difference b/n μ_1 & μ_2

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



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

Pro The marketing specialization students earns atleast 5000 Rupees more per month than operation management students

Specialization	Samplesize	salary	standard deviation
Marketing	120	67500	7200
Operations	45	58950	7600

$\bar{x}_1 = 67500$ $\bar{x}_2 = 58950$

$n_1 = 120$ $n_2 = 45$

$s_1 = 7500$ $s_2 = 4,600$

$H_0 : \mu_1 - \mu_2 \leq 5000$

$H_1 : \mu_1 - \mu_2 > 5000$

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

$\because \mu_1 - \mu_2 = 5000$

$$\Rightarrow \frac{(67500 - 58950) - (5000)}{\sqrt{\frac{(7500)^2}{120} + \frac{(4,600)^2}{45}}}$$

$$\Rightarrow \frac{8550 - 5000}{\sqrt{\frac{5184000}{120} + \frac{211600}{45}}}$$

$$\Rightarrow \frac{3550}{4320 +}$$

$= 3.737$

Typing

Prob Typing speed on a PC

Men & women typing speeds are different.

consider $\alpha = 1\%$.

	men	women
$\bar{x}$	65 wpm	68 wpm
s	10 wpm	14 wpm
n	50	60

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$

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

$$\Rightarrow \frac{(\bar{x}_1 - \bar{x}_2) - (0)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$\Rightarrow \frac{(65-68)}{\sqrt{\frac{(10)^2}{50} + \frac{(14)^2}{60}}} \Rightarrow \frac{-3}{\sqrt{2+3.26}} \Rightarrow \frac{-3}{\sqrt{5.26}}$$

$$\Rightarrow \frac{-3}{2.29}$$

$$\Rightarrow -1.30$$

3
196
60
7
14 × 14
60 30 15
15) 49 (3.26
45
40
30
100

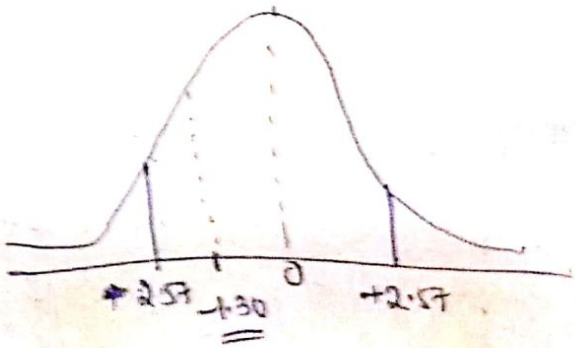
confidence fol 99%.

→ two tail test z-score = 2.57

-1.30 is in accepted Region.

So, H_0 will be accepted

ie, $\mu_1 = \mu_2$



Prob Are the mach
w.r.t how
sample co
$\bar{x}$
s
n
$\alpha = 0.08$

Sol

$z = (\bar{x}_1$

⇒

⇒

Ans Are the machine tools manufactured by X, & Y. different w.r.t how long they last.

sample	company 'x'	company 'y'
$\bar{x}$	16.2 weeks	15.9 weeks
s	0.2 weeks	0.2 weeks
n	40	40

$\alpha = 0.08 \Rightarrow 8\%$.

Sol

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

$$\Rightarrow \frac{(16.2 - 15.9)^2 - (0)}{\sqrt{\frac{(0.2)^2}{40} + \frac{(0.2)^2}{40}}}$$

$$\Rightarrow \frac{(0.3)^2}{\sqrt{\frac{0.04 + 0.04}{40}}}$$

$$\Rightarrow \frac{0.09}{\sqrt{\left(\frac{0.08}{40}\right)}} \Rightarrow \frac{0.09}{\sqrt{\left(\frac{2}{1000}\right)}} = \frac{0.09}{\sqrt{\frac{1}{500}}} \Rightarrow 0.09 \times \sqrt{5} \cdot 10$$

$$\Rightarrow 0.9 \times \sqrt{5}$$

$$\Rightarrow$$

P30 Who earns more married (M) unmarried

	married	unmarried
$\bar{x}$	$639.60	$658.20
s	$60	$90
n	40	60

Consider $\alpha = 0.05$

Sol^n

H_0 : earnings of married & unmarried are same

H_1 : earnings of married & unmarried are not same

z - statistic

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

$$\Rightarrow \frac{(639.60 - 658.20) - (0)}{}$$

12.4
65
100
120
20
$\Rightarrow 1.24$

125

Q2) who lives longer married & unmarried.

	married	unmarried
$\bar{x}$	78.5 years	77.0 years
s	14.0	16.0
n	140	150

$\alpha = 0.01$

Sol)

H_0

H_1

Z-statistic

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

$$\Rightarrow \frac{(78.5 - 77.0) - (0)}{\sqrt{\frac{(14)^2}{140} + \frac{(16)^2}{160}}}$$

$$\Rightarrow \frac{(1.5)}{\sqrt{\frac{196}{140} + \frac{256}{160}}}$$

$$\Rightarrow \frac{1.5}{\sqrt{3}}$$

$\Rightarrow$

$\frac{14 \times 14}{140} = \frac{16}{10}$

$\Rightarrow 1.4 + 1.6$

$\Rightarrow$ 1.4 + 1.6 = 3.0

$\frac{3}{2} \Rightarrow 1.5$

$\left(\frac{3^2}{5}\right) = \frac{9}{4} \Rightarrow 2.25$

3/9/19

Paired sample t-test

- applied on a single sample, but at different conditions
  (or) two different situation

Ex growth of a student before taking drug and after taking drug

→ In many cases we would like to analyse whether an intervention (or treatment) such as training programs, treatment for specific illness may have significantly change population Parameter. Such as mean (or) proportion. before & after intervention

- In pair t-test, the data related to parameter is captured twice. once before intervention and once after intervention

- Ex: 1) Body weight before & after attending Yoga training
  2) cholestrol level before & after attending meditation training.

Assume that the mean difference in the estimated parameter before & after treatment is 'd' and the corresponding standard deviation difference is S_d. Let μ_d be the hypothised mean difference. then statistic 't'

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

Assume 'd' follows normal distribution than statistic 't' follows 't' distribution with (n-1) degrees of

Prob A Researcher believes that people drink more coffee on monday than other days of week. Based on sample of 50 coffee drinkers, the mean difference was estimated as 14 ml and corresponding standard deviation 8.5 ml. Conduct a hypothesis test at $\alpha = 0.1$ to claim that people drink on average 10ml more coffee on monday compared to other days of week

Soln $n = 50$

$D = 14 \text{ ml}$

$S_d = 8.5 \text{ ml}$

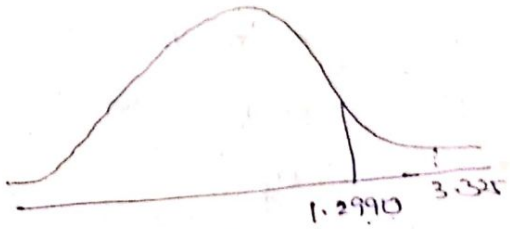
$\mu_d = 10 \text{ ml}$

$\alpha = 0.1$

$\mu_1 \quad \mu_2 + 10$

$\mu_d = 10$

$\mu_d - 10$



$H_0 : \mu_d \leq 10$

$H_1 : \mu_d > 10$

$$t = \frac{D - \mu_d}{S_d/\sqrt{n}} = \frac{14 - 10}{8.5/\sqrt{50}} = 3.3275$$

H_0 it have to be rejected

So, People will have more coffee on moday than other days.

P30

The difference in avg weekly consumption in alcoholic units before & after breakup is 11.5 and the corresponding standard deviation difference is 95.67 for 20 candidates,

Conduct a hypothesis test to check whether alcohol consumption is more after breakup ie., ($\mu_d > 0$) at 95% confidence

Soln

$\mu_d \neq 11.5$

$D = 11.5$

$S_d = 95.67$

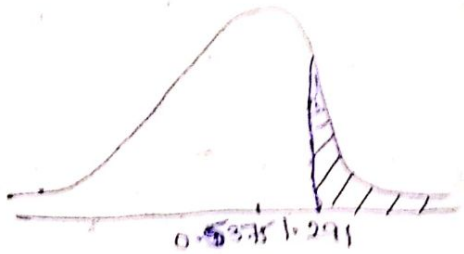
$n = 20$

$\alpha = 0.05\%$

confidence level = 95%

$H_0: \mu_d \leq 0$

$H_1: \mu_d > 0$



$$t = \frac{D - \mu_d}{S_d/\sqrt{n}} \Rightarrow \frac{11.5 - 0}{95.67/\sqrt{20}} = 0.5375$$

$t_{critical}$ value at degree of freedom = 19

$\alpha = 0.05$

So, 0.5375 is in acceptance region

So, H_0 is accepted ie., there is no difference.

syllabus

Two sample t-test

(a) Differencing Two population means when population standard deviations are unknown, and believe to be equal.

If we want to estimate difference in two population means when standard deviation of the population are unknown

hence, we need to estimate them from the samples drawn from two populations.

An additional anumption is that the standard deviati -on of two populations are equal. (Otherwise unknown).

Then sampling dishibution of the difference in estimated means $(\bar{x}_1 - \bar{x}_2)$ follows t-dishibution with degree of freedom $(n_1 + n_2 - 2)$ with mean $(\mu_1 - \mu_2)$ with standard deviation $\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$.

here s_p^2 is the pooled stann variance.

s_p^2 = pooled variance

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

Prob A Company makes a claim that children who drink their health drink will grow taller than the children who donot drink that health drink.

Group	Sample size	Increase in height	standard deviation
Drink health drink	80	7.6 cm	1.1 cm
donot drink health drink	80	6.3 cm	1.3 cm

at $\alpha = 0.05$, test whether the increase in height of the children who drink the health drink is atleast 1.2 cm

Soln given. $n_1 = 80$ $n_2 = 80$

$\bar{x}_1 = 7.6$ $\bar{x}_2 = 6.3$

$s_1 = 1.1$ $s_2 = 1.3$

$H_0 : \mu_0 (\mu_1 - \mu_2) \leq 1.2$

$H_1 : (\mu_1 - \mu_2) > 1.2$

$$t = \frac{(x_1 - x_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

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

$$\Rightarrow \frac{79 \cdot (1.1)^2 + 79 (1.3)^2}{160 - 1}$$

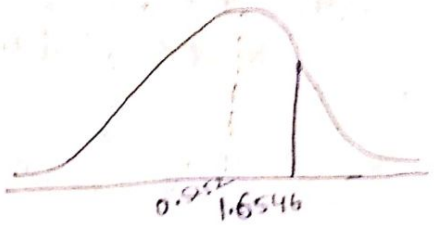
$$= \frac{(1.3) - 1.2}{\sqrt{1.45 \left(\frac{1}{80} + \frac{1}{80}\right)}}$$

$$\Rightarrow \frac{79 \cdot (1.21) + 79 (1.69)}{159}$$

$$\Rightarrow \frac{0.1}{\sqrt{1.45 \times \frac{1}{40}}} \Rightarrow 0.5252 \Rightarrow 1.45$$

t statistic = 0.5252

t critical value for 1-tail test where $\alpha = 0.05$ & degree of freedom $(n_1+n_2-2) = 158$ is 1.6546



$\alpha = 0.5$

So, t statistic is not in the rejection region.

So, accept H_0.

i.e., the difference is ≤ 1.2.

test!

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

two sample t-test with unequal variance

- 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 deviation from the samples drawn from these two populations.

Then the sampling distribution of the difference in estimated means $(\bar{x}_1 - \bar{x}_2)$ follows a t-distribution with mean $(\mu_1 - \mu_2)$ and standard deviation $S_u = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$, the corresponding degree of freedom is given by $df = \left\lfloor \frac{S_u^4}{\frac{(s_1^2/n_1)^2}{n_1+1} + \frac{(s_2^2/n_1)^2}{n_2+1}} \right\rfloor$.

- Then t :

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

9/9/22

Eg: A Researcher is interested in finding the avg duration of marriage based on educational qualifications of couples. Two groups were considered for study. Group1 consisted of couple with no batchelor degree (both partners) and Group2 consists of couples who both have batchelor's degree or higher.

Group	Sample size	duration of marriage	standard deviation of sample
no degree	120	10.1 years	2.4 years
with degree	100	9.5 years	3.1 years.

at $\alpha = 0.05$ test whether the avg. duration of marriage is more for couples with no batchelor's degree as compared to couples with batchelor's degree

Soln $n_1 = 120 \quad n_2 = 100$

$\bar{x}_1 = 10.1$ y. $\quad \bar{x}_2 = 9.5$ y.

$s_1 = 2.4$ y. $\quad s_2 = 3.1$ y.

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

$H_1: \mu_1 - \mu_2 > 0$

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

$$\Rightarrow \frac{(10.1 - 9.5) - 0}{\sqrt{\frac{(2.4)^2}{120} + \frac{(3.1)^2}{100}}}$$

$$\Rightarrow \frac{0.6}{\sqrt{\frac{5.76}{120} + \frac{9.61}{100}}}$$

$$\Rightarrow \frac{0.6}{\sqrt{\frac{0.48}{10} + \frac{9.61}{100}}}$$

$$= 1.5605$$

$$df = S_u = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$df = \left\lfloor \frac{(S_u)^4}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_1)^2}{n_2 - 1}} \right\rfloor = \frac{0.0207}{0.000113}$$

$$\approx \lfloor 184.33 \rfloor = 184$$

$\alpha = 0.05$ & $df = \lfloor 184.33 \rfloor \approx 184$

The critical value of 't' for $\alpha = 0.05$ & degree of freedom (df) = 184. is 1.6531

$\frac{24 \times 24}{96}$

$\frac{48}{5.76}$

$\frac{3.1 \times 31}{31}$

$\frac{93}{961}$

$12 \overline{)5.76}\ \overline{)6.48 \times 12}$

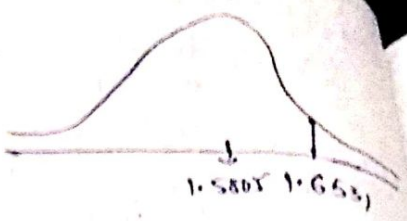
$\frac{48}{96}$ $\quad$ 5.76

t-critical value = 1.6531

t-statistic value = 1.5605

accept it ie, H_0 is accepted.

ie, $\mu_1 - \mu_2 \leq 0$.



Since the t-statistic is less than critical value of 't', we retain null hypothesis ie, the difference in duration of marriage between two groups is less than & equal to zero

11/9/19

Non parametric test

· good fit test

Parametric tests

Z-tests } mean
t-tests }
F-test — proportion

Day	M	T	W	Th	F	S	
Expected %	10	10	15	20	30	15	
Observed	30	14	34	45	57	20	= 200
Expected value	20	20	30	40	60	30	
(Expected-Observed) (E) (O)	10	-6	4	5	-3	-10	
$(E-O)^2$	10^2	$(-6)^2$	4^2	5^2	$(-3)^2$	$(-10)^2$	

or 1.653),

t' we retain

rriage

n

rtion

12/4/19

Non - parametric test

In previous sections we discussed z-test, t-test are called parametric tests. since the objective is to infer about population parameters such as mean and population parameters

To conduct z-test, t-test we need summary statistics (mean & standard deviation) not necessarily the entire distribution.

In parametric test, we assume that the population follows Normal distribution, Non-parametric tests are distribution free tests. since they do not have assumptions about distribution population.

The major difference between parametric & Non parametric tests.

- parametric tests needs only values of parameters and knowledge about distribution
- In non-parametric test we need entire distribution of the data.
- Importantly data may not follow any parametric distribution such as Normal distribution
- Also test is not about the population parameter, but about characteric of the entire distribution.
- Non - tests parameters

>> When do popular non-parametric test is used

- The test is not about the population parameter such as mean & standard deviation
- The method / test doesn't require assumption about Population distribution.

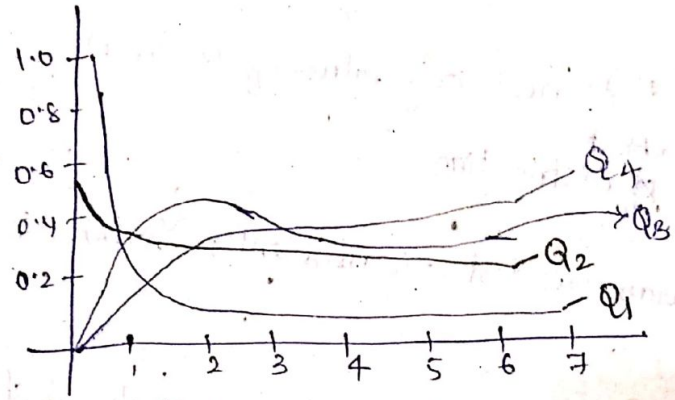
χ^2 - distribution chi-square (χ^2)

- let us take a random variable 'x' if 'x' is coming from standard Normal distribution

$$X \sim N(0,1)$$

Q is another random variable

$Q = x^2$, the distribution of Q random variable is χ^2-distribution with degree of freedom 1.



- If we take 2-samples x_1, x_2,

$Q_2 = x_1^2 + x_2^2$, which will follows χ^2-distribution with df = 2

$Q_3 = x_1^2 + x_2^2 + x_3^2$ — follows χ^2-distribution with df=

A Restaurant owner believes that his restaurant's walk incoming customers expected distribution would be

day	M	T	W	Th	F	S
Expected %	10	10	15	20	30	15

But on one week observed values are

Observed values	30	14	34	45	57	20

respectively

At $\alpha = 0.05$, check whether owners expected customer's distribution correct or not. with appropriate hypothesis test.

Sol^n H_0 : owners distribution is correct

H_1 : owners distribution is not correct

$\frac{10}{100} \times 200$

χ^2 - statistic

Expected %	10	10	15	20	30	15	
observed	30	14	34	45	57	20	= 200
Expected value	20	20	30	40	60	30	

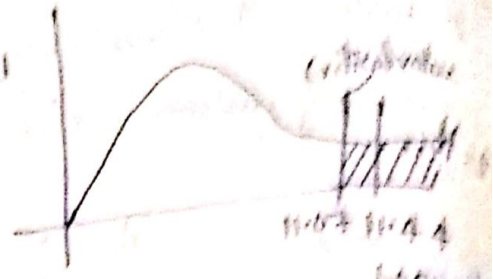
$$\chi^2\text{-statistic} = \frac{(30-20)^2}{20} + \frac{(14-20)^2}{20} + \frac{(34-30)^2}{30} + \frac{(45-40)^2}{40} + \frac{(57-60)^2}{60} + \frac{(20-30)^2}{30}$$

$$\Rightarrow \frac{10^2}{20} + \frac{6^2}{20} + \frac{4^2}{30} + \frac{5^2}{40} + \frac{3^2}{60} + \frac{10^2}{30}$$

$$\Rightarrow \frac{100}{20} + \frac{36}{20} + \frac{16}{30} + \frac{25}{40} + \frac{9}{60} + \frac{100}{30} = 11.44$$

χ^2 - distribution with $\alpha = 0.05$ is 11.07

$\because \chi^2$ - statistic value > χ^2 - critical value



We reject null hypothesis

- owner's distribution is not a good fit based on significance level 0.05

16/07/19

Hanuman Airlines operated indian airlines to several indian cities. one of the problems hanuman airlines faces is food Preferences given by the panengers. Captain cook the operations manager of hanuman airlines believes that 35% of panengers Prefer vegitarian food, 40% panengers prefer non-vegetarian food 20% Low calory food & 5% requested for diabetic food A sample of 500 panengers were chosen to analyse the food Preferences

food type	veg	non-veg	low calorie	diabetic
No of panengers	190	185	90	35

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

Sol^n H_0 : Captain cook's belief is true

H_1 : Captain cook's belief is not true

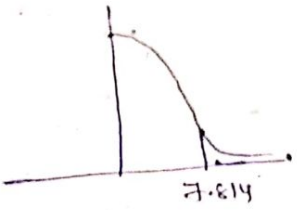
~~X~~ χ^2 - statistic

Food type	Veg	non-veg	Low calorie	Diabetic	$\Sigma E = 500$
Observed	190	185	90	35	
Expected %	35	45	20	5	
Expected value	175	225	100	25	

~~degree~~

degree of freedom = 3

χ^2 - critical value with d.f = 3, at $\alpha = 0.05$ is 7.814.



α	0 ...	0.05
1		
2		
3	----	7.814

χ^2 - statistic $\left(\frac{O-E}{E}\right)$

$$\Rightarrow \left(\frac{190-175}{175}\right) + \frac{(190-175)^2}{175} + \frac{(225-185)^2}{225} + \frac{(100-90)^2}{100} + \frac{(35-25)^2}{25}$$

$$\Rightarrow \frac{225}{175} + \frac{(40)^2}{225} + \frac{(10)^2}{100} + \frac{(10)^2}{25}$$

$$\left(\frac{45}{175}\right)^2 = \frac{1}{7}$$

$$\Rightarrow 1.2857 + 7.111 + 1.4$$

$$\Rightarrow 13.3967$$

7)9(1.285
7
20
14
60
56
40
35
50