Advance Statistics - 05

Topics:

(I) Chi Square Test

(II) Chi Square For Goodness of Fit (Problem)

(III) Chi Square Test with Python

(I) Chi Square Test

It claims about population proportion.

It is a non parametric test that is performed on categorical data to find out the relationship between them.

Example: There is a population of Male who likes different colors of bikes.

Likes Yellow bike → 1/3 rd of population

Likes Orange bike → 1/3 rd of population

Likes Red bike → 1/3 rd of population

find out the statement is true or not

Color	Theory	Sample
Yellow	1/3	22
Orange	1/3	17
Red	1/3	59 → Observed categorical distribution

1/3 → Theory categorical distribution

CHI Square Fitness of good:

In 2010, Census of the city, the weight of the individuals in a small city. were found to be the following.

<50kg	50-75	>75
20%	30%	50%

In 2020, ages of $n = 500$ individuals were sampled. Below are the results.

<50	50-75	>75
140	160	200

Using $\alpha = 0.05$, would you conclude the population differences of weights has changed in the last 10 year?

Ans:

Expected

<50	50-75	>75
~~140~~	~~160~~	~~200~~
20%	30%	50%

$n = 500$

Observed

<50	50-75	>75
140	160	200

New expected

<50	50 - 75	75
0·2×500 = 100	0·30×500 = 150	0·5 × 500 = 250

① **Null hypothesis (h_0)**: The data meets the expectation

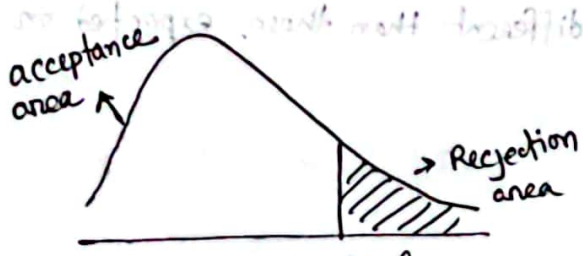
Alternate hypothesis (h_1): The data does not meet the expectation.

② $\alpha = 0.05$, C·I = 95%

③ **Degree of freedom:** $df = K - 1 = 3 - 1 = 2$

[K = number of categories]

④ **Decision Boundary:**



Critical Value → To find this critical value we use α, dof and chi square table

↓

value is 5·991

if chi square (χ^2_ϕ) is greater than 5·991, we reject the null hypothesis else, we failed to reject the null hypothesis.

⑤ **Chi square test statistics:**

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

O = Observed
E = Expected

$$= \frac{(40)^2}{100} + \frac{(10)^2}{150} + \frac{(-50)^2}{200}$$

$$= \frac{1600}{100} + \frac{100}{150} + \frac{2500}{200}$$

$$= 16 + 0.66 + 10$$

$$= 26.66$$

We can see that, $\chi^2 = 26.66 > 5.99$. So we reject the null hypothesis.

The weights of 2020 population are different than those expected on the 2010 population.

Advance Statistics 06

Topics:

(i) F Distribution

(ii) Variance Ratio Test (F Test)

(i) F Distribution: (Right skewed)

It is a continous probability distribution that arises frequently as the null distribution of a test statistic, most notably in the analysis of variance (ANOVA) and other F-tests.

Parameters:

$d_1, d_2 > 0$ (degree of freedom)

Support: $x \in (0, +\infty)$ [x = random variable]

~~Pdf~~ =

$$PdF = \frac{\sqrt{\frac{(d_1 x)^{d_1}\, d_2^{d_2}}{(d_1 x + d_2)^{d_1 + d_2}}}}{x\, B\left(\frac{d_1}{2}, \frac{d_2}{2}\right)}$$

$$B(m,n) = \frac{(m-1)!\,(n-1)!}{(m+n-1)!}$$

The F distribution with d1 and d2 degree of freedom is the distribution of $X = \frac{S_1/d_1}{S_2/d_2}$

S_1, S_2 = Independent random variables { follows Chi square Distribution }

Variance Ratio Test (F Test):

The following data shows the number of bubbs produced daily for some days by 2 workers ~~are~~ A and B

A	B
40	39
30	38
38	41
41	33
38	32
35	39
	40
	34

Can we consider based on the data that Worker B is more stable and efficient?

$\alpha = 0.05$

Ans:

① Null Hypothesis (H_0): $\sigma_1^2 = \sigma_2^2$

Alternate Hypothesis (H_1): $\sigma_1^2 \neq \sigma_2^2$

Sample Variance

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

② Calculation of variance:

for A, $\bar{x}_1 = 37$

x_1	$\bar{x}_1$	$(x_1 - \bar{x}_1)^2$
40	37	9
30	37	49
38	37	1
41	37	16
38	37	1
35	37	4
		$\Sigma (x_1 - \bar{x}_1)^2 = 80$

For B,

X_2	$\bar{X}_2$	$(X_2 - \bar{X})^2$
39	37	4
38	37	1
41	37	36
33	37	16
32	37	25
39	37	4
40	37	9
$\bar{X}_2 = 37$		$\Sigma (X_2 - \bar{X}_2)^2 = 84$

$$S_1^2 = \frac{80}{n_1 - 1}$$

$$= \frac{80}{6-1}$$

$$= 16$$

$$S_2^2 = \frac{84}{n_2 - 1}$$

$$= \frac{84}{8-1}$$

$$= 12$$

③ **Calculation of variation Ration (F Test):**

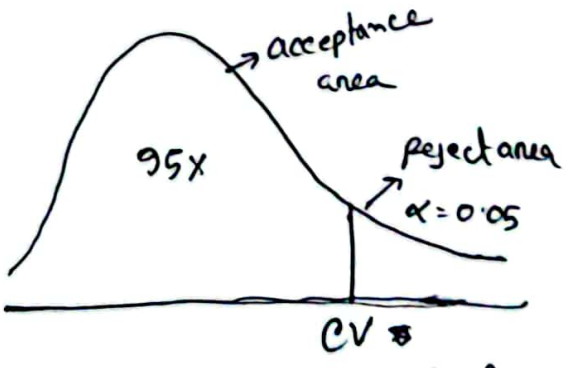
$$F = \frac{S_1^2}{S_2^2} = \frac{16}{12} = 1.33$$

④ **Decision Rule:**

degree of freedom: $df1 = 6 - 1 = 5$

$df2 = 8 - 1 = 7$

$\alpha = 0.05$



Check F table to find the value

cv found = 3.97

If F test is greater than 3.97, Reject the Null hypothesis

Else, we fail to reject the null hypothesis.

Conclusion: Worker B is not efficient when worked to worker A.