Data Science, 2022
Tut 5: Evaluation and Measurement- Hypothesis Testing

Make Assumptions about values when it is necessary in consistent manner. Refer necessary table from following link when necessary.

https://www.sheffield.ac.uk/polopoly_fs/1.43999!/file/tutorial-10-reading-tables.pdf

Testing a Proportion of small samples

1. H_0: $p = p_0$

2. One of the alternatives H\: $p < p_n$, $p > p_o$, or $p \neq p_o$

3. Choose a level of significance equal to a.

4. Test statistic: Binomial variable X with $p = p_0$.

5. Computations: Find x, the number of successes, and compute the appropriate P-value.

6. Decision: Draw appropriate conclusions based on the P-value

<u>Ex. 1</u>

A builder claims that air-conditions are installed in 70% of all homes being constructed today in the city of Mumbai. Would you agree with this claim
if a random survey of new homes in this city shows that 8 out of 15 had air-conditions installed? Use a 0.10 level of significance

1 $H_0: p = 0.7$ $\quad H_1: p \neq 0.7$ $\quad$ level of significance $= 0.10$

Binomial variable x with $p = 0.7$ and $n = 15$

$p = 2P(x \leq 8 \text{ when } p = 0.7)$
$= 2\, b(x, 15, 0.7)$
$= 2 * 0.1312$
$= 0.2624$

$\because$ $p > 0.1$ so we not reject H_0

<u>Ex.2</u>

A commonly prescribed drug for relieving nervous tension is believed to be only 60% effective. Experimental results with a new drug administered to a random sample of 100 adults who were suffering from nervous tension show that 70 received relief. Is this sufficient evidence to conclude that the new drug is superior to the one commonly prescribed? Use a 0.05 level of significance.

② $H_0: p = 0.6$ $\quad H_1: P > 0.6$ $\quad$ level of significance $= 0.05$.

$$z = \frac{\overset{x}{70} - \overset{n}{100} \times \overset{p}{0.6}}{\sqrt{\underset{n}{100} \times \underset{p}{0.6} \times 0.4}}$$

$1 - P_0 = 1 - 0.6 = 0.4$

$= 2.04$

$\therefore p = 0.0207$.

Since $P < 0.05$ we reject H_0

Ex.3

A vote is to be taken among the residents of a Mumbai and the surrounding area to determine whether a proposed Nuclear plant should be constructed. The construction site is within the Mumbai limits, and for this reason many voters in the surrounding area feel that the proposal will pass because of the large proportion of Mumbai voters who favor the construction. To determine if there is a significant difference in the proportion of Mumbai voters and surrounding area voters favoring the proposal, a poll is taken. If 120 of 200 Mumbai voters favor the proposal and 240 of 500 surrounding area residents favor it, would you agree that the proportion of Mumbai voters favoring the proposal is higher than the proportion of surrounding area voters? Use an a = 0.05 level of significance.

③ P_0 is voter favors Mumbai P_1 Surrounding area favors

$\therefore P_0 = \frac{120}{200} = 0.6$ $\quad P_1 = \frac{240}{500} = 0.48$ $\quad$ level of significance $= 0.05$

$P = \frac{120 + 240}{200 + 500} = 0.514$

(3) $H_0 : P_0 \le P_1$ $H_1 : P_0 > P_1$

$$z = \frac{P_0 - P_1}{\sqrt{p(1-P_0)\left(\frac{1}{n_1} - \frac{1}{n_2}\right)}}$$

$$= \frac{0.6 - 0.48}{\sqrt{0.514(0.486)(7\times10^{-3})}}$$

$z = 2.87$

$\therefore P = 0.004$

$\because P < 0.05$ so we reject H_0

Ex.4

State the null and alternative hypotheses to be used in testing the following claims, and determine generally where the critical region is located:

(a) At most, 20% of next year's wheat, crop will be exported to the Russia..

(b) On the average, Indian homemakers drink 3 cups of tea per day.

(c) The proportion of graduates in engineering this year majoring in the computer sciences is at least. 0.15.

(d) The average donation to the Indian Autism Association is no more than 500 INR.

(e) Residents in suburban Mumbai commute, on the average, 15 kilometers to their place of employment.

(4)

i) $H_0: p = 0.20$ $H_1 = p > 0.2$

Critical region is right tail

ii) $H_0: c = 3$ $H_1: c \neq 3$

Critical region is on both tail since 2 tail test.

iii) $H_0: p = 0.15$ $H_1 = p < 0.15$

Critical region is on left tail.

iv) $H_0 = \mu = 500$ $H_1: \mu > 500$

Critical region right tail it is

v) $H_0: p = 15$ $H_1: p \neq 15$

Critical region is in both tails

Ex.5

In a study conducted by the Department of computer Engineering and analyzed by the Statistics Consulting Center at SPIT the laptops supplied by two different companies were compared. Ten sample laptops were made out of the Intel chips supplied by each company and the "robustness" was studied. The data are as follows:

Company A: 9.3 8.8 6.8, 8.7 8.5 6.7 8.0 6.5 9.2 7.0

Company B: 11.0 9.8 9.9 10.2,10.1 9.7 11.0 11.1 10.2 9.6

Can you conclude that there is virtually no difference in means between the laptops supplied by the two companies? Use a P-value to reach your conclusion. Should variances be pooled here?

Q.5 μ_1 : mean popⁿ of laptops by company A.
μ_2 : mean popⁿ of laptops by company B.

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2 \qquad \alpha = 0.05$

$$\bar{x}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} x_{1i} = \frac{9.3+8.8+6.8+8.7+8.5+6.7+8.0+6.5+9.2+7.0}{10}$$

$$\bar{x}_1 = 7.95$$

$\bar{x}_2 = 10.26$

$$S_1^2 = \frac{1}{n_1-1}\left[\sum_{i=1}^{n_1} x_{1i}^2 - n_1\bar{x}_1^2\right] \quad \therefore S_1 = \frac{10.865}{9} = 1.207$$

$$S_2^2 = \frac{1}{n_2-1}\left[\sum_{i=1}^{n_2} x_{2i}^2 - n_2\bar{x}_2^2\right] = 0.325$$

Sample variance are diff. so we will use unpooled t-test.

$$v = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{s_2^2}{n_2}\right)^2} = 10.30$$

The test statistics

$$T = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

$T = -5.90$

$|t| = 5.90$ p value

~~pvalue~~ p-value $= 2P(T \geq |t|)$

$= 2P(T \geq 5.90)$

$t_{0.0005}(10) = 4.587$

p-value < 0.001

As $p < \alpha$, we can reject null hypothesis in favor of the alternative hypothesis and conclude that mean robustness of laptops is not the same for two companies.