Inferential statistics

Testing of Hypothesis

Z test

It is important parametric test based upon the normality assumption. Traditionally Z test is used, when the samples are selected from population of known parameter with sample size more than 30. We consider that if sample size is more than 30 then sample selected from non normal population is also approximately normal distributed.

Z test is defined as the ratio of difference between t and E(t) to the S.E.(t)

$$Z = \frac{t-E(t)}{SE.(t)} \sim N(0, 1),$$

where t = statistic, E(t) = Expected value of statistic and S.E.(t) = Standard error of the statistic.

Z test is used to test

- Significance of single mean.

- Significance of difference between two means.

- Significance of single proportion.

Significance of difference between two proportions

- Significance of difference between sample correlation and population correlation.

- Significance of difference between independent sample correlations

Test of significance of a single mean

Let us consider sample of size n (n > 30) has been drawn from the normal population N (μ, σ^2) then the sample mean $\bar{x} \sim$ N (μ, σ^2).

Different steps in the test are;

Problem to test

Ho:$\mu = \mu_0$ (sample is drawn from population with mean μ_0)

H1: $\mu \neq \mu_0$ (Two tailed test)

or H1: $\mu > \mu_0$ (One tailed right)

or H1: $\mu < \mu_0$ (One tailed left)

Test statistic

For the sample selected from the population of unknown size

$$Z = \frac{\bar{X} - E(\bar{X})}{SE.(\bar{X})} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$ for known variance

$\frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$ for unknown variance(for large sample size($\hat{\sigma} = s$)

For the sample selected from the population of known size

$$Z = \frac{\bar{X} - E(\bar{X})}{SE.(\bar{X})} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}}$$ for known variance

$$\frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}}$$ for unknown variance

Where $\bar{X}$ = sample mean, μ = population mean, σ = population s.d. s = sample s.d.,

N = population size, n = sample size

Level of significance

Let α be the level of significance. Usually we take α = 0.05 unless we are given.

Critical value

Critical or tabulated value of Z is obtained from table according to the level of significance and alternative hypothesis.

Decision

Reject Ho at α level of significance if I z I > $Z_{\text{tabulated}}$, accept otherwise.

Example

A sample of 400 students is found to have mean height of 170 cm. Can it be reasonably regarded as a sample from a large population with mean height 169.5 cm and standard deviation 3.5 cm?

Solution

Here,

Sample size (n) = 400

Sample mean ($\bar{X}$) = 170

Population mean (μ) = 169.5

Population S.D. (σ) = 3.5

Problem to test

Ho : Mean height of students is 169.5 cm (μ = 169.5)

H_1 = Mean height of student is not 169.5 cm ($\mu \neq$ 169.5) (Two tailed)

Test statistic

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{170 - 169.5}{\frac{3.5}{\sqrt{400}}} \quad \frac{0.5 \times 20}{3.5} = 2.857$$

Critical value

Let 5% be the level of significance then critical value is Ztab = $Z\alpha/2$ = 1.96 Decision Here Z= 2.857 > Z_{tab} = 1.96, reject Ho at 5% level of significance.

Conclusion

The sample of 400 students cannot be regarded as sample from large population with mean height 169.5 cm and standard deviation 3.5 cm.

Test of significance difference between two means

Let us consider two independent samples of size n_1 and n_2 be drawn from population having means means μ_1 and μ_2 and variances σ_1^2 and σ_2^2 respectively. Let $\bar{X}_1$ and $\bar{X}_2$ be the sample means.

For large n_1 and n_2.

$$\bar{X}_1 \sim N(\mu_1, \frac{\sigma_1^2}{n_1}$$

$$\bar{X}_2 \sim N(\mu_2, \frac{\sigma_2^2}{n_2}$$

$$\bar{X}_1 - \bar{X}_2 \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$$

Different steps in the test are

Problem to test

Ho: $\mu_1 = \mu_2$ There is no significant difference between two population mean.

H1: $\mu_1 \neq \mu_2$ (two tailed)

or H1 : $\mu_1 < \mu_2$ (one tailed left)

or H1 : $\mu_1 > \mu_2$ (one tailed right)

Test statistic

$$\boldsymbol{Z} = \frac{\bar{X}_1 - \bar{X}_2 - E(\bar{X}_1 - \bar{X}_2)}{\boldsymbol{S.E.}(\bar{X}_1 - \bar{X}_2)}$$

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

When population means and variances are known

$$\boldsymbol{Z} = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

When population variances are known

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

When population variances are unknown

for large sample size$\widehat{\sigma_1^2} = S_1^2$ and $\widehat{\sigma_2^2} = S_2^2$

$\bar{X}_1$ = sample mean of size n_1,

$\bar{X}_2$ = sample mean of size n_2

σ_1^2 = population variance of first population

σ_2^2 = population variance of second Population

S_1^2 = sample variance of first sample

S_2^2 = sample variance of second sample.

Level of significance

Let a α be the level of significance. Usually we take α = .05 unless we are given.

Critical value

Critical or tabulated value of Z is obtained from table according to the level of significance and alternative hypothesis.

Decision

Reject Ho at a level of significance if I z I > $_{\text{Ztabulated}}$, accept otherwise.

Example

In a random sample of 500 the mean is found to be 20. In another independent sample of 400 the mean is 15. Could the samples have been drawn from the same population with S.D. 4? Solution Here,

Sample size of first sample (n_1) = 500

Sample mean of first sample ($\bar{X}_1$) = 20

Sample size of second sample (n_2) = 400

Sample mean of second sample ($\bar{X}_2$) = 15

Population S.D. of first (σ_1) = 4

Population S.D. of second (σ_2) = 4

Problem to test

Ho : μ_1 = μ_2 (both the populations are same)

H1: $\mu_1 \neq \mu_2$ (population are different)

Test statistic

$$\boldsymbol{Z} = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$\boldsymbol{Z} = \frac{(20-15)}{\sqrt{\frac{16}{500}+\frac{16}{400}}} \mathbf{= 18.51}$$

Critical value

Let α = 5 % be the level of significance the critical value is $_{\text{Ztabulated}}$ = Z α /2 = 1.96. .

Decision

Z =18.51 > Ztabulated = 1.96, reject Ho at 5% level of significance.

Conclusion

We cannot conclude that the samples have been drawn from the same population.

Test of significance difference between two proportions:

Let P_1 and P_2 be the two population proportions possessing a certain characteristic. Let two independent samples of sizes n_1 and n_2 be drawn from the two population. Also p_1 and p_2 be the proportion of units possessing certain characteristic in the two samples.

For large sample size

$p_1 \sim N(p_1\ ,\frac{P_1Q_1}{n_1})$

$p_2 \sim N(p_2\ ,\frac{P_2Q_2}{n_2})$

Then

$p_1 - p_2 \sim (p_1\ \ - p_2\, , \frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2})$

Different steps in the test are;

Problem to test

$H_{0\,:}$ $P_1 = P_2$

$H_{1\,:}$ $P_1 \neq P_2$ (Two tail test)

$H_{1\,:}$ $P_1\ > P_2$ (One tail right)

H_{1} : $P_1 < P_2$ (One tail left)

Test statistic

$$\boldsymbol{Z} = \frac{(p_1 \ - p_2\,) - E\ ((p_1 \ - p_2\,)}{\boldsymbol{S.E.}((p_1 \ - p_2\,)}$$

$$= \frac{(p_1 \ - p_2 - (\mathrm{P1} - \mathrm{P2})}{\sqrt{\frac{\boldsymbol{P_1Q_1}}{\boldsymbol{n_1}} + \frac{\boldsymbol{P_2Q_2}}{\boldsymbol{n_2}}}}$$

$$= \frac{(\mathrm{P1} - \mathrm{P2})}{\sqrt{\frac{\boldsymbol{P_1Q_1}}{\boldsymbol{n_1}} + \frac{\boldsymbol{P_2Q_2}}{\boldsymbol{n_2}}}}$$

If population proportion are given

$$= \frac{(p_1 \ - p_2}{\sqrt{\boldsymbol{PQ}(\frac{\boldsymbol{1}}{\boldsymbol{n_1}} + \frac{\boldsymbol{1}}{\boldsymbol{n_2}})}}$$

If population proportion are not given

Where P_1 = population proportion of first population

P_2 = population proportion of second population

p_1= sample proportion of first sample of size n_1

p_2 = sample proportion of second sample of size n_2

Level of significance

Let α be the level of significance. Usually we take α = .05 unless we are given.

Critical value Critical or tabulated value of Z is obtained from table according to the level of significance and alternative hypothesis.

Decision

Reject Ho at a level of significance if I z I > $_{\text{Ztabulated}}$, accept otherwise.

Example

A machine puts out 21 defective articles in a sample of 500 articles. Another machine gives 3 defective articles in a sample of 100 are the two machines significantly different in their performance? Use p value method at 1% level of significance.

Solution

Here Defective articles by a machine (x_1) = 21

Number of articles by a machine (n_1) = 500

Defective articles by another machine (x_2) = 3

Number of articles by another machine (n_2) = 100

Sample proportion of defective article by a machine (p_1) = $\frac{x1}{n1} = \frac{21}{500} = 0.042$

Sample proportion of defective article by another machine (p_2) = $\frac{x2}{n2} = \frac{3}{100} = 0.03$

Let

P_1 = Population proportion of defective from a machine

P2 = Population proportion of defective from another machine

$$P = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{500 \times 0.042 + 100 \times 0.03}{500 + 100} = \frac{24}{100} = 0.04$$

Level of significance (α) = 1%

Problem to test

Ho : $P_1 = P_2$ (There is no significance difference in performance of machines)

H_1 : $P_1 \neq P_2$ (There is significance difference in performance of machines)

Test statistic

$$Z = \frac{(p_1 \ - p_2}{\sqrt{\boldsymbol{PQ}(\frac{\boldsymbol{1}}{\boldsymbol{n_1}} + \frac{\boldsymbol{1}}{\boldsymbol{n_2}})}}$$

$$\frac{0,042 - 0.03}{\sqrt{0.04 \times 0.96(\frac{1}{500} + \frac{1}{100})}} = 0.571$$

Now prob(Z$\geq$ Zcalculated$) = \text{prob}(Z \geq 0.571) = 0.5 - \text{prob}(0 \leq Z \leq 0.571)$

$$= 0.5 - 0.2175 = 0.284$$

For two tailed test, p value = 2 Prob (Z $\geq$ $_{\text{Zcalculated}}$) = 2 x 0.284 = 0.568

Here α = 1% = 0.01

Decision P value = 0.568 > α = 0.01, accept H_0 at 1% level of significance.

Conclusion

There is no significant difference in performance of two machines.

t test

When the sample size is small (traditionally it is assumed less than or equal to 30), then the sampling distribution of the sample mean is assumed to follow student's t distribution. The t distribution is also similar to normal distribution having shape as in normal distribution but little bit flatter. As the sample size increases the shape of t distribution is more likely to normal curve. Whatever be the sample size the statistical software uses the t test for all sample size instead of Z test, since it can compute the tail area of the curve (p value) or to compare with the pre-assigned value of a.

t test is based upon the assumption that

- Sample size small
- Sample is selected from normal population.
- Population standard deviation is not known.
- Samples are independent.

It is used to test

- **Significance of single mean**.
- **Significance of difference between means.**
- **Significance of correlation coefficient.**
- **Significance of regression coefficient**.

Some other test

i)Chi- square test

ii) ANOVA (Analysis of Variance)

iii) Run test

iv) Sign test

v) Mann Whitney U test

vi) The Kruskal- Wallis test