L- 7: Inferential statistics & Predictive Analytics

Agenda

- Central limit theorem
- Type I, Type II Errors
- Testing of Hypothesis – continuation from previous session
- Covariance
- Correlation
- Introduction to regression

Central Limit Theorem

If $\bar{x}$ is the mean of a sample of size n taken from a population having the mean μ and variance σ^2, then

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$ is a random variable

whose distribution function approaches that of the standard normal distribution as $n \longrightarrow \infty$.

Central Limit theorem

→ It does not matter what the distribution of x_i's is

→ In many real applications, the random variable is a sum of independent random variables. In all such cases, CLT helps to use normal distribution.

Examples

→ random noise in Comm. Systems

→ Errors in Lab measurements

→ Errors in regression analysis etc

Errors:

$p(\ ) = \alpha$

	H_0 is true	H_0 is False
Reject H_0	Type I Error (false positive)	Correct Decision
Accept H_0	Correct Decision	Type II Error (false negative)

$p(\ ) = \beta$

Testing of Hypothesis

or

Hypothesis testing

Mean

One mean

($H_0: \mu = \mu_0 / \mu \geq \mu_0 / \mu \leq \mu_0$)

Large sample

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

Small sample

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

for $(n-1)$ d.o.f

Two means

($H_0: \mu_1 - \mu_2 = \delta$)

Large

Small

(see next slide)

Mean

one mean

Two means
$(H_0: \mu_1 - \mu_2 = \delta)$

Large sample

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

Small sample

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

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

Testing of Hypothesis

Example - 1

Example :- 1

Can it be concluded that the average life span of Indians is more than 70 yrs. If a random sample of 100 indians has average life span of 71.8 years with a S.D of 8.9 years.

Example:- (contd)

Can it be concluded that the average life span of Indians is more than 70yrs.

If a random sample of 100 indians has average life span of 71.8 years with a S.D of 8.9 years.

Sample

$100 = n$

$\bar{x} = 71.8$

$s = 8.9$

Validation

$H_0 : \mu > 70$

population

Example:- (contd)

Can it be concluded that the average life span of Indians is more than 70 yrs. If a random sample of 100 Indians has average life span of 71.8 years with a S.D of 8.9 years.

→ One mean problem

→ $n = 100$: Large sample, so Z-test

$$\therefore Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \text{ or } \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

$H_0 : \mu > 70$

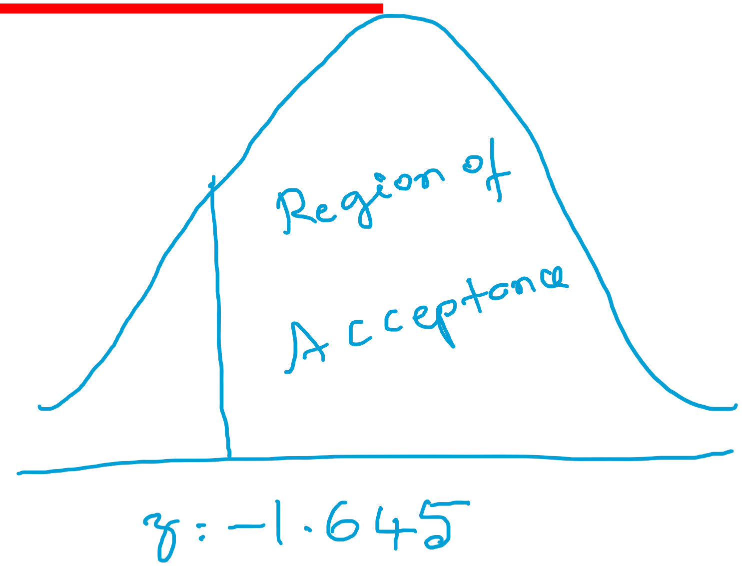
$H_1 : \mu \leq 70 \longrightarrow$ Left tailed test

$\alpha = 5\%$ (Let)

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{71.8 - 70}{8.9 / \sqrt{100}}$$

$$= 2.022$$

Lies in the region of acceptance



$\therefore$ H_0 is accepted

i.e Avg life is more than 70 years

Testing of Hypothesis

Example – 2.

Example - 2

A machine which produces Mica insulating washers for use in electronic devices said to have a thickness of 10mm. A sample of 10 washers has an average thickness of 9.52 mm with a S.D of 0.6mm. whether the sample is drawn from the given population

(use 5% Level of significance)

Example - 2 Small sample

Mean

A machine which produces Mica insulating washers for use in electronic devices said to have a thickness of 10mm. A sample of 10 washers has an average thickness of 9.52 mm with a S.D of 0.6mm. whether the sample is drawn from the given population

(use 5% Level of significance)

$\bar{x}$ S

$H_0 : \mu = 10$

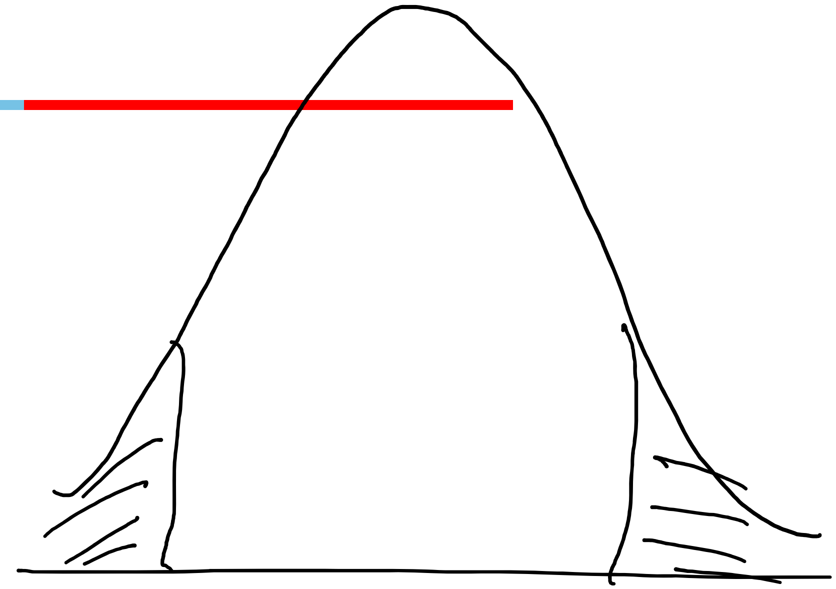
$H_1 : \mu \neq 10$

$\alpha = 0.05$

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

$$= \frac{9.52 - 10}{0.6/\sqrt{10}}$$

$$= -2.52 \longrightarrow \text{Reject } H_0$$



-2.26 $\qquad$ $+2.26$

Testing of Hypothesis

Example – 3

Example (3)

A random sample of 40 items produced by a company A have a mean life time of 647 hours with S.D 27 hours. While a sample of 40 items by company B has a mean life time of 638 hours with S.D of 31 hours.

Does this substantiate the claim of the company A that their items are superior to those produced by company B.

Example 3 — Solution ?

A random sample of 40 items produced by a company A have a mean life time of 647 hours with S.D 27 hours. While a sample of 40 items by company B has a mean life time of 638 hours with S.D of 31 hours.

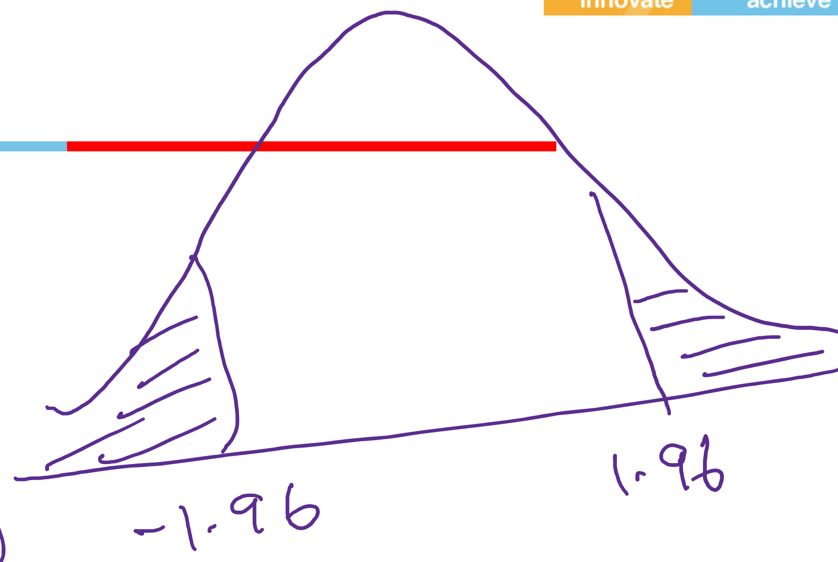
Does this substantiate the claim of the company A that their items are superior to those produced by company B.

Solution:-

$H_0: \mu_1 - \mu_2 = 0$

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

$\alpha = 0.05$



$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(647 - 638) - 0}{\sqrt{\frac{(27)^2}{40} + \frac{(31)^2}{40}}}$$

$$= 3.73$$

Reject H_0!

Testing of Hypothesis

Example – 4

Example- 4 :-

A Company believes that the advertisement A is more effective than advt. B. To test this sampling is done.

In a random sample of 60 customers who saw advertisement A, 18 tried the product. In a random sample of 100 customers, who saw advt B, 22 tried the product.

Does this indicate that advt A is more effective than advt. B.

Example- 4 :-

A Company believes that the advertisement A is more effective than advt. B. To test this sampling is done.

In a random sample of 60 customers who saw advertisement A, 18 tried the product. In a random sample of 100 customers, who saw advt B, 22 tried the product.

Does this indicate that advt A is more effective than advt. B.

Sample A : 18 out of 60
Sample B : 22 out of 100

Advt (A) > Advt (B)
????

Testing of Hypothesis

Example – 5

Example:-

consider the following data

Travel time	Stress: High	Moderate	Low	Total
< 20 min	9	5	18	32
20-50 min	17	8	28	53
> 50 min	18	6	7	31
Total	44	19	53	116

Based on this data, Can we conclude that stress levels depends on travel time

???

Thanks