STATISTICS FOR DATA SCIENCE
HYPOTHESIS and INFERENCE

Dr. Deepa Nair

Department of Science and Humanities

STATISTICS FOR DATA SCIENCE

UNIT-4 HYPOTHESIS and INFERENCE

Session-7

Large - Sample tests for Difference between two means

Dr. Deepa Nair
Department of Science and Humanities

STATISTICS FOR DATA SCIENCE

Large - Sample tests for Difference between two means

- We now investigate examples in which we wish to determine whether the means of two populations are equal. The data will consist of two samples, one from each population. The basic idea is quite simple. We will compute the difference of the sample means. If the difference is far from 0, we will conclude that the population means are different. If the difference is close to 0, we will conclude that the population means might be the same
- As an example, suppose that a production manager for a manufacturer of industrial machinery is concerned that ball bearings produced in environments with low ambient temperatures may have smaller diameters than those produced under higher temperatures. To investigate this concern, she samples 120 ball bearings that were manufactured early in the morning, before the shop was fully heated, and finds their mean diameter to be 5.068 mm and

Source: Text Book Chapter6

their standard deviation to be 0.011 mm. She independently samples 65 ball bearings manufactured during the afternoon and finds their mean diameter to be 5.072 mm and their standard deviation to be 0.007 mm. Can she conclude that ball bearings manufactured in the morning have smaller diameters, on average, than ball bearings manufactured in the afternoon?

Source: Text Book Chapter6

STATISTICS FOR DATA SCIENCE

Large - Sample tests for Difference between two means

- So let us see another example of a study on surgery verses physiotherapy which has been carried out for 6 months. From the study it is observed that mean % of the people carried by surgery is 72 with a standard deviation of 8 for a sample of 32 patients.
- At the same time the mean 90 of the people caused by physiotherapy is 75 with a standard deviation of 7 and the study has been denote for a sample of 36 patients.
- Now the questions is can we conduct a hypothesis test to test surgery is better physiotherapy .So here we need to found whether the means of two populations are equal the two samples are taken from two different population

STATISTICS FOR DATA SCIENCE

Large - Sample tests for Difference between two means

Example

Tennis Elbow

To operate or not to operate?

Mean % of the people cured by Surgery

$$\mu_X = 72, \sigma_X = 8, n_X = 32$$

Mean % of the people cured by Physiotherapy

$$\mu_Y = 75, \sigma_Y = 6, n_Y = 36$$

STATISTICS FOR DATA SCIENCE

Large - Sample tests for Difference between two means

Example:

Can we conduct a Hypothesis test ?

Can we say that Surgery is better than Physiotherapy ?

STATISTICS FOR DATA SCIENCE

Large - Sample tests for Difference between two means

- To determine whether the means of two populations are equal.
- The data will consist of two samples, one from each population.
- Let $X_1, \ldots X_{n_X}$ and $Y_1, \ldots . Y_{n_Y} (n_X > 30, n_Y > 30)$ be large sample from a population with means $\mu_X and\ \mu_Y$, and standard deviations $\sigma_X\ and\ \sigma_Y$.

Source: Text Book Chapter6

STATISTICS FOR DATA SCIENCE

Large - Sample tests for Difference between two means

- We will compute the difference of the sample means.

$$H_0: \mu_X - \mu_Y = \Delta_0,$$

$$H_0: \mu_X - \mu_Y > \Delta_0,$$

$$H_0: \mu_X - \mu_Y < \Delta_0$$

Source: Text Book Chapter6

STATISTICS FOR DATA SCIENCE

Large - Sample tests for Difference between two means

- Compute the z-score,

$$z = \frac{(\bar{X}-\bar{Y})-\Delta_0}{\sqrt{{\sigma^2}_X/_{n_X}+{\sigma^2}_Y/_{n_Y}}}$$

Type equation here.

- $\bar{X}-\bar{Y} \sim N(\Delta_0, {\sigma^2}_X/_{n_X} + {\sigma^2}_Y/_{n_Y})$

- If σ_X and σ_Y are unknown they may be approximated with s_X and s_Y respectively.

Source: Text Book Chapter6

STATISTICS FOR DATA SCIENCE

Large - Sample tests for Difference between two means

- Compute the P-value
- The P-value is an area under the normal curve, which depends on the alternate hypothesis as shown in the table:

Source: Text Book Chapter6

STATISTICS FOR DATA SCIENCE

Large - Sample tests for Difference between two means

Alternate Hypothesis	*P*-value
$H_1: \mu_X - \mu_Y > \Delta_0$	$Area\ to\ the\ right\ of\ z$
$H_1: \mu_X - \mu_Y < \Delta_0$	$Area\ to\ the\ left\ of\ z$
$H_1: \mu_X - \mu_Y \neq \Delta_0$	$Sum\ of\ the\ areas\ in\ the\ tails\ cut\ off\ by\ z\ and\ -z$

STATISTICS FOR DATA SCIENCE

Large - Sample tests for Difference between two means

Example:

A study has been carried out for a period of 6 months for the treatment of Tennis elbow and the results are as follows:

Mean % of the people cured by Surgery

$$\bar{X} = 72, S.D = 8, n_X = 32$$

Mean % of the people cured by Physiotherapy

$$\bar{Y} = 75, S.D = 6, n_Y = 36$$

Can we conclude that surgery is better than Physiotherapy?

STATISTICS FOR DATA SCIENCE

Large - Sample tests for Difference between two means

Solution:

$$H_0: \mu_X - \mu_Y \leq 0$$

$$H_1: \mu_X - \mu_Y > 0$$

$$\bar{X} = 72, \sigma_X \rightarrow s_X = 8, n_X = 32$$

$$\bar{Y} = 75, \sigma_Y \rightarrow s_Y = 6, n_Y = 36$$

STATISTICS FOR DATA SCIENCE

Large - Sample tests for Difference between two means

Solution:

$$z = \frac{(\bar{X} - \bar{Y}) - \Delta_0}{\sqrt{{\sigma^2}_{\bar{X}}/n_X + {\sigma^2}_Y/n_{\bar{Y}}}}$$

$$= \frac{75 - 72}{\sqrt{\frac{36}{36} + \frac{64}{32}}} = \frac{3}{\sqrt{3}} = \sqrt{3} = 1.73 > 1.645 at\ 5\%$$

So we can conclude that surgery is better than Physiotherapy.

STATISTICS FOR DATA SCIENCE

Large - Sample tests for Difference between two means

Solution:

- The P –Value is
- At 5% significance level

Solution:

- So we need to reject the null hypothesis and conclude that surgery is better than Physiotherapy.

STATISTICS FOR DATA SCIENCE

Large - Sample tests for Difference between two means

Example:

The article "Effect of Welding Procedure on Flux Cored Steel Wire Deposits" (N. Ramini de Rissone, I. de S. Bott, et al., *Science and Technology of Welding and Joining*, 2003:113–122) compares properties of welds made using carbon dioxide as a shielding gas with those of welds made using a mixture of argon and carbon dioxide.

Source: Text Book Chapter6

STATISTICS FOR DATA SCIENCE

Large - Sample tests for Difference between two means

Example:

- One property studied was the diameter of inclusions, which are particles embedded in the weld.
- A sample of 544 inclusions in welds made using argon shielding averaged 0.37 μm in diameter, with a standard deviation of 0.25 μm.
- A sample of 581 inclusions in welds made using carbon dioxide shielding averaged 0.40 μm in diameter, with a standard deviation of 0.26 μm. (Standard deviations were estimated from a graph.)
- Can you conclude that the mean diameters of inclusions differ between the two shielding gases?

STATISTICS FOR DATA SCIENCE

Large - Sample tests for Difference between two means

Solution:

$$H_0: \mu_X - \mu_Y = 0$$

$$H_1: \mu_X - \mu_Y \neq 0$$

$$\overline{X} = 0.37 \ , \overline{Y} = 0.40$$

$$s_X = 0.25, S_Y = 0.26, n_X = 544, n_Y = 581$$

$$\overline{X} - \overline{Y} \sim N(0.0.01521^2)$$

Large - Sample tests for Difference between two means

Solution:

$$z = \frac{(\bar{X} - \bar{Y}) - \Delta_0}{\sqrt{{\sigma^2}_X/n_X + {\sigma^2}_Y/n_Y}} = -1.97$$

Solution:

- This is a two-tailed test, and the P-value is 0.0488.
- A follower of the 5% rule would reject the null hypothesis.
- It is certainly reasonable to be skeptical about the truth of H_0

STATISTICS FOR DATA SCIENCE

Large - Sample tests for Difference between two means

Example:

- In previous example Can you conclude that the mean diameter for carbon dioxide welds (μ_Y) exceeds that for argon welds (μ_X) by more $than\ 0.015\ \mu m$?

STATISTICS FOR DATA SCIENCE

Large - Sample tests for Difference between two means

Solution:

$$H_0: \mu_X - \mu_Y \geq -0.015$$

$$H_1: \mu_X - \mu_Y < -0.015$$

$$\overline{X} = 0.37 \ , \overline{Y} = 0.40$$

$$s_X = 0.25, S_Y = 0.26, n_X = 544, n_Y = 581$$

$$\overline{X} - \overline{Y} \sim N(-0.015. 0.01521^2)$$

STATISTICS FOR DATA SCIENCE

Large - Sample tests for Difference between two means

Solution:

$$z = \frac{(\bar{X} - \bar{Y}) - \Delta_0}{\sqrt{{\sigma^2}_X/n_X + {\sigma^2}_Y/n_{\bar{Y}}}}$$

$$z = \frac{-0.03 - (-0.015)}{0.01521} = -0.99$$

Solution:

- This is a one-tailed test.
- The P −value is 0.1611.
- We cannot conclude that the mean diameter of inclusions from carbon dioxide welds exceeds that of argon welds by more than $0.015\ \mu m$.

STATISTICS FOR DATA SCIENCE

Large - Sample tests for Difference between two means

Example:

- In a random sample of 100 tube lights produced by company A, the mean lifetime (mlt) of tube light is 1190 hours with standard deviation of 90 hours.

- Also in a random sample of 75 tube lights from company B the mean lifetime is 1230 hours with standard B the mean lifetime is 1230 hours with standard deviation of 120 hours.

- Is there a difference between the mean lifetime of the two brands of tube lights at a significance level of (a) 0.05 (b) 0.01?

STATISTICS FOR DATA SCIENCE

Large - Sample tests for Difference between two means

Solution:

- Let X_A , X_B denote the lifetime(in hours) of tube lights produced by company A and B respectively.

- It is given that the mean lifetime of tube lights of company A is $\overline{X_A}$=1190, standard deviation for tube lights of A is $s_A = 90$.

- Similarly $\overline{X_B}$ =1230, s_B= 120, $n_A =$ sample size of tube lights from A= 100, $n_B =$ sample size from B = 75

STATISTICS FOR DATA SCIENCE

Large - Sample tests for Difference between two means

Solution:

$H_0: \mu_A - \mu_B = 0$ i.e., no difference.

Alternate hypothesis: $H_1: \mu_A - \mu_B \neq 0$ i.e., there is difference.

$$\sigma_{\bar{X}_A - \bar{X}_B} = \sqrt{{\sigma_{\bar{X}_A}}^2 + {\sigma_{\bar{X}_B}}^2} \quad \overset{\mu_{\bar{X}_A - \bar{X}_B} = \Delta_0 = 0}{=} \quad \sqrt{\frac{{S_A}^2}{n_A} + \frac{{S_B}^2}{n_B}}$$

$$= \sqrt{\frac{(90)^2}{100} + \frac{(120)^2}{75}} = 16.5227$$

STATISTICS FOR DATA SCIENCE

Large - Sample tests for Difference between two means

Solution:

Test statistic:

$$z = \frac{(\bar{X}_A - \bar{X}_B) - \Delta_0}{\sqrt{\frac{{S_A}^2}{n_A} + \frac{{S_B}^2}{n_B}}} = \frac{1190 - 1230}{16.5227} = -2.421$$

For $\alpha = 0.05$.

Reject N.H. since $\mathrm{P} - \mathrm{Value} 0.0156 < 0.05$

i.e., yes, there is difference between the mean lifetimes of the tube lights produced by

A and B.

Solution:

For $\alpha = 0.01$

Accept N.H. since Reject N.H. since $\mathbf{P - Value} 0.0156 > 0.05$

So there is no difference between $\overline{X}_A$ and $\overline{X}_B$.

STATISTICS FOR DATA SCIENCE

Large - Sample tests for Difference between two means

Example:

- To test the effects a new pesticide on rice production, a farm land was divided into 60 units of equal areas, all portions having identical qualities as to soil, exposure to sunlight etc.
- The new pesticide is applied to 30 units while old pesticide to the remaining 30.
- Is there reason to believe that the new pesticide is better than the old pesticide if the mean number of kg. of rice harvested/unit using new pesticide (N.P)is 496.31with s.d of 17.18 kgs while for old pesticide (O.P) is 485.41kgs and 14.73kgs.Test at a level of significance (a) $\alpha = 0.005$ (b) 0.01.

STATISTICS FOR DATA SCIENCE

Large - Sample tests for Difference between two means

Example:

$$H_0: \mu_X - \mu_Y \leq 0$$

$H_1: \mu_X - \mu_Y > 0$ i.e., new pesticide is superior to (better than) old pesticide.

$$\bar{X} = 496.31, \bar{Y} = 485.41, s_X = 17.18, s_Y = 14.73, n_X = 30, n_Y = 30$$

Test statistic is

$$z = \frac{(\bar{X}-\bar{Y})-\Delta_0}{\sqrt{{\sigma^2}_X/n_X + {\sigma^2}_Y/n_{\bar{Y}}}} = \frac{(496.31-485.41)-0}{\sqrt{\frac{(17.18)^2}{30}+\frac{(14.73)^2}{30}}} = 2.63814$$

Example:

Decision

$\alpha = 0.05$

Reject N.H. since $\mathrm{P-Value} 0.041 < 0.05$ i.e., accept A.H. or new pesticide is superior to old pesticide.

Dr. Deepa Nair

Department of Science and Humanities

deepanair@pes.edu