STATISTICS FOR DATA SCIENCE

Confidence Intervals

Prof. Uma D
Prof. Suganthi S
Prof. Silviya Nancy J

Department of Computer Science and Engineering

STATISTICS FOR DATA SCIENCE

Confidence Intervals for Difference Between two means

Prof. Uma D
Prof. Suganthi S
Prof. Silviya Nancy J

Topics to be covered...

- Sum/ Difference of two independent normally distributed random variables
- A Confidence Interval for the Difference Between Two Means
- Confidence Intervals Estimate for Paired data

Sum/ Difference of two independent normally distributed random variables is normal

If $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$ are independent random variables that are normally distributed, then their sum/difference is also normally distributed.

If,

$$X \sim N(\mu_X, \sigma_X^2)$$
$$Y \sim N(\mu_Y, \sigma_Y^2)$$

Then,

$$X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$$

$$X - Y \sim N(\mu_X - \mu_Y, \sigma_X^2 + \sigma_Y^2)$$

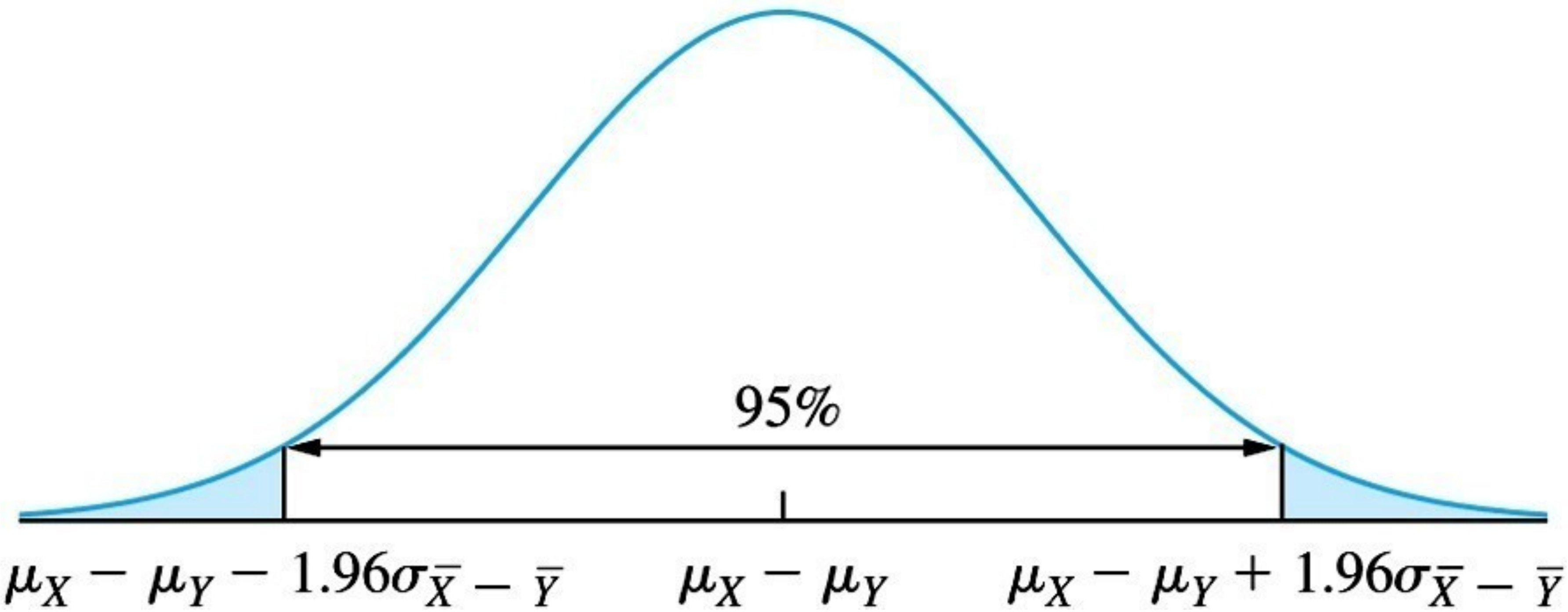
A Confidence Interval for the Difference Between Two Means

Let $X_1, \ldots, X_{n_X}$ be a *large* random sample of size n_X from a population with mean μ_X and standard deviation σ_X, and let $Y_1, \ldots, Y_{n_Y}$ be a *large* random sample of size n_Y from a population with mean μ_Y and standard deviation σ_Y. If the two samples are independent, then a level $100(1-\alpha)\%$ confidence interval for $\mu_X - \mu_Y$ is

$$\overline{X} - \overline{Y} \pm z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}} \qquad (5.16)$$

When the values of σ_X and σ_Y are unknown, they can be replaced with the sample standard deviations s_X and s_Y.

A Confidence Interval for the Difference Between Two Means



Example

A group of 75 people enrolled in a weight loss program that involved adhering to a special diet and to a daily exercise program. After 6 months, their mean weight loss was 25 pounds, with a sample standard deviation of 9 pounds.

A second group of 43 people went on the diet but didn't exercise. After 6 months, their mean weight loss was 14 pounds, with a sample standard deviation of 7 pounds.

Find a 95% confidence interval for the mean difference between the weight losses.

Solution

X_bar ~ N(25, 9/sqrt(75))

Y_bar ~ N(14, 7/sqrt(43))

since both the samples are independent,

a 95% Confidence Interval for $\mu_X - \mu_Y$ is given by

$$(\bar{X} - \bar{Y}) \pm z_{a/2} * \text{sqrt}\left((\sigma_X{}^2/n_1) + (\sigma_Y{}^2/n_2) \right)$$

$= (25 - 14) \pm 1.96 * \text{sqrt}\left((9^2/75) + (7^2/43) \right)$

$= 11 \pm 1.96 * \text{sqrt}(2.2195)$

$= 11 \pm 2.92$

$= (8.08, 13.92)$

Confidence Intervals with Paired Data

The data is described as paired when it arises from the same observational unit.

An example of paired data would be a before-after drug test.

The data is described as unpaired or independent when the sets of data arise from separate observational unit.

For example one clinical trial might involve measuring the blood pressure from one group of patients who were given a medicine and the blood pressure from another group not given it.

Constructing confidence Intervals with Paired Data

For large samples,

If the population of differences is approximately normal, then a $(1-\alpha)$

100% Confidence Interval for μ_D is given by:

$$D \pm z_{\alpha/2}\sigma_D.$$

In practice, σ_D is approximated with s_D/n .

Constructing confidence Intervals with Paired Data

For small samples (n < 30),

If the population of differences is approximately normal, then a $(1-\alpha)$

100% Confidence Interval for μ_D is given by:

$$D \pm t_{n-1,\alpha/2} \frac{s_D}{n} .$$

Example

Breathing rates, in breaths per minute were measured for a group of 10 people at rest and then during moderate exercise. The results are as follows:

N	Exercise	Rest
1	30	15
2	37	16
3	39	21
4	37	17
5	40	18
6	39	15
7	34	19
8	40	21
9	38	18
10	34	14

Find a 95% confidence interval for the increase in breathing rate due to exercise.

Solution

N	Exercise(X)	Rest (Y)	Difference (D = X – Y)
1	30	15	15
2	37	16	21
3	39	21	18
4	37	17	20
5	40	18	22
6	39	15	24
7	34	19	15
8	40	21	19
9	38	18	20
10	34	14	20

Solution

D_bar = mean of differences = 19.4

s_D = standard deviation of differences

s_D = 2.836273 , n = 10 , alpha = 0.05

$t_{10-1,.025}$ = 2.262

The 95% confidence interval is $19.4 \pm 2.262(2.836273/\sqrt{10})$, or (17.3712, 21.4288).

THANK YOU

Prof D.Uma
Prof S. Suganthi
Prof J. Silviya Nancy

Computer Science and Engineering

umaprabha@pes.edu
+91 99 7251 5335