STATISTICS FOR DATA SCIENCE

Confidence Intervals

Prof. Uma D

Department of Computer Science and Engineering

STATISTICS FOR DATA SCIENCE

Confidence Intervals for Difference Between two means

Prof. Uma D

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

Topics to be covered...

CI for Difference b/w Two (True) Means

$\boxed{\mu_x - \mu_y}$

single sample

$\bar{x}$ [ ] → $\mu = ?$

$\hat{p}$ [ ] → $P = ?$

Population

$(\bar{x} - \bar{y}) \pm MoE \Rightarrow \mu_x - \mu_y$

[ ]

All plausible values for $\mu_x - \mu_y$

Difference b/w Two Means

Sample 1 $\bar{x}$ → Population 1 μ_x

Sample 2 $\bar{y}$ → Population 2 μ_y

$(\bar{x}-\bar{y}) - MoE$ $(\bar{x}-\bar{y})$ $(\bar{x}-\bar{y}) + MoE$

LB   UB

A Confidence Interval for the Difference Between Two Means

Group 1

independent groups

Group 2

There are many situations where it is of interest to compare two groups with respect to their mean scores on a continuous outcome.

For example, we might be interested in comparing mean systolic blood pressure in men and women, or perhaps compare body mass index (BMI) in smokers and non-smokers.

Both of these situations involve comparisons between two **independent groups**, meaning that there are different people in the groups being compared.

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,

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

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

$\bar{X}$

$\bar{Y}$

μ_X

μ_Y

σ_X^2/n

SD

$\bar{X} \sim N(\mu_X, (\sigma_X/\sqrt{n})^2) \rightarrow$ variance

$\bar{Y} \sim N(\mu_Y, \sigma_Y/\sqrt{n}) \rightarrow$ SD

PES UNIVERSITY ONLINE

A Confidence Interval for the Difference Between Two Means

CI for Large Sample

CI for $\mu_x - \mu_y$.

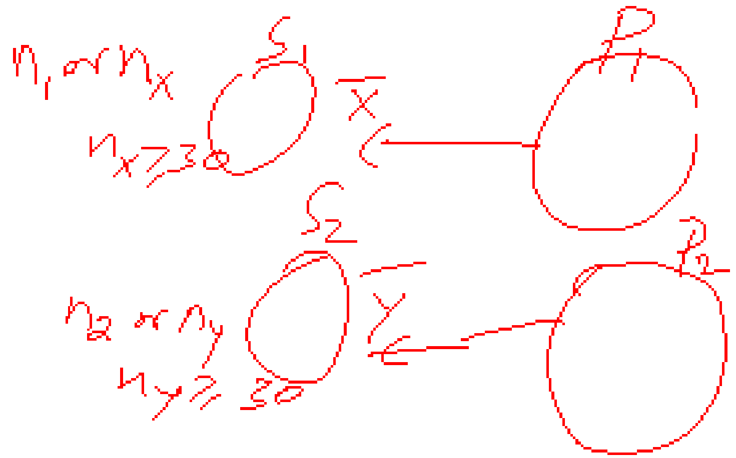
$(1-\alpha)*100\%$ CI is given by

$$(\bar{X} - \bar{Y}) \pm MoE$$

$$\boxed{(\bar{X} - \bar{Y}) \pm Z_{\alpha/2} * \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}}$$

$$\left(-\infty, \ (\bar{X} - \bar{Y}) + Z_{\alpha} * \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}\right)$$

$$\left((\bar{X} - \bar{Y}) - Z_{\alpha} * \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}, \ +\infty\right)$$



$$\bar{X} \sim N\left(\mu_x, \frac{\sigma_x^2}{n_x}\right)$$

$$\bar{Y} \sim N\left(\mu_y, \frac{\sigma_y^2}{n_y}\right)$$

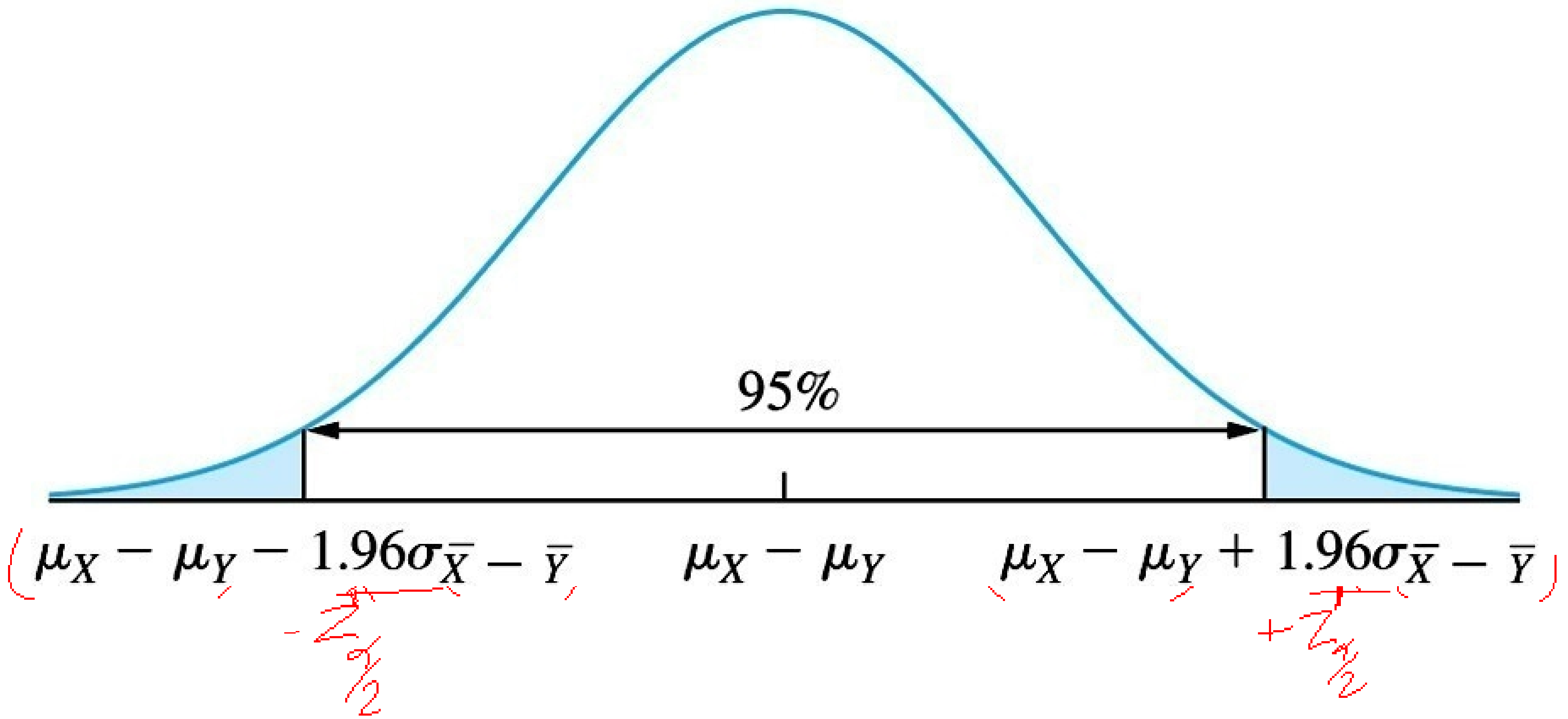
A Confidence Interval for the Difference Between Two Means

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.

n_x or $n_1 = 75$

$\bar{X} = 25$

$s_x = 9$

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.

n_y or $n_2 = 43$

$\bar{Y} = 14$

$s_y = 7$

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

$CL = 95\%$

Solution

$\bar{X} \sim N\left(25, \frac{9}{\sqrt{75}}\right)$ SD

$\bar{Y} \sim N\left(14, \frac{7}{\sqrt{43}}\right)$

$X \sim N(\mu_x, \sigma_x)$

$\bar{X} \sim N\left(\mu_x, \frac{s_x}{\sqrt{n}}\right)$

$\bar{Y} \sim N\left(\mu_y, \frac{s_y}{\sqrt{n}}\right)$

95% CI is given by

$\alpha = 0.05$

$$(\bar{X} - \bar{Y}) \pm z_{\alpha/2} * \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}$$

$$(25 - 14) \pm 1.96 * \sqrt{\frac{9^2}{75} + \frac{7^2}{43}}$$

$$11 \pm 1.96 \sqrt{2.2195}$$

$$11 \pm 2.92$$

95% CI for $\mu_x - \mu_y$ is $(8.08, 13.92)$

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

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

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

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

$= 11 \pm 2.92$

$= (8.08, 13.92)$

Solution

	Men			Women		
Characteristic	n		s	n		s
Systolic Blood Pressure	1,623	128.2	17.5	1,911	126.5	20.1
Diastolic Blood Pressure	1,622	75.6	9.8	1,910	72.6	9.7
Total Serum Cholesterol	1,544	192.4	35.2	1,766	207.1	36.7
Weight	1,612	194.0	33.8	1,894	157.7	34.6
Height	1,545	68.9	2.7	1,781	63.4	2.5
Body Mass Index	1,545	28.8	4.6	1,781	27.6	5.9

Source: sphweb.bumc.bu.edu

Population 1

Men

μ_x

$\mu_x - \mu_y$

$(\bar{X} - \bar{Y}) \pm z_{\alpha/2} \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}$

Women

μ_y

Population 2

STATISTICS FOR DATA SCIENCE

Solution

Group1 Group2

	Men	Women	Difference
Characteristic	Mean (s)	Mean (s)	95% CI
Systolic Blood Pressure	128.2 (17.5)	126.5 (20.1)	(0.44, 2.96)
Diastolic Blood Pressure	75.6 (9.8)	72.6 (9.7)	(2.38, 3.67)
Total Serum Cholesterol	192.4 (35.2)	207.1 (36.7)	(-17.16, -12.24)
Weight	194.0 (33.8)	157.7 (34.6)	(33.98, 38.53)
Height	68.9 (2.7)	63.4 (2.5)	(5.31, 5.66)
Body Mass Index	28.8 (4.6)	27.6 (5.9)	(0.76, 1.48)

G1 = M

G2 = W

Source: sphweb.bumc.bu.edu

Interpretation

Interpretation: With 95% confidence the difference in mean systolic blood pressures between men and women is between **0.44 and 2.96** units.

Our best estimate of the difference, the point estimate, is **1.7 units.**

The standard error of the difference is 0.641, and the margin of error is 1.26 units.

$\sqrt{\frac{S_x^2}{n_x}+\frac{S_y^2}{n_y}}$ $Z_{\alpha/2}$ * Standard error

$Z_{\alpha/2} * \frac{S_D}{\sqrt{n}}$

When comparing two independent samples in this fashion the confidence interval provides a range of values for the ***difference***.

high (+ve values)

In this example, we estimate that the difference in mean systolic blood pressures is between 0.44 and 2.96 units with **men having the higher values**. In this example, we arbitrarily designated the men as group 1 and women as group 2.

M W

$G_1 - G_2 > 0$

W − M

Had we designated the groups the other way (i.e., women as group 1 and men as group 2), the confidence interval would have been **-2.96 to -0.44,** suggesting that **women have lower systolic blood pressures** (**anywhere from 0.44 to 2.96 units lower than men**

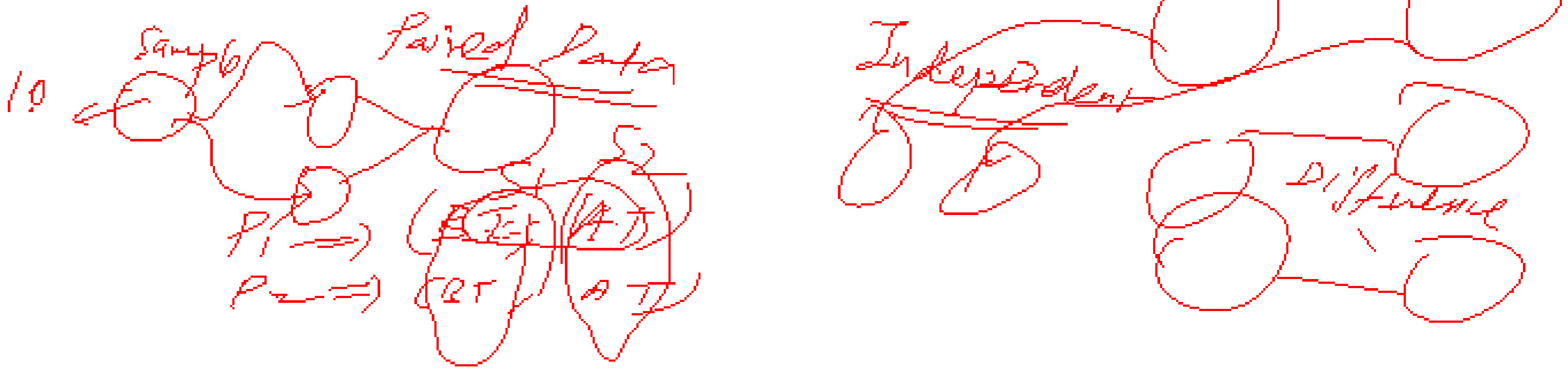
Interpretation

Notice that the 95% confidence interval for the **difference in mean total cholesterol levels** between men and women is **-17.16 to -12.24.**

Men have lower mean total cholesterol levels than women; anywhere from 12.24 to 17.16 units lower.

Conclusion

The men have higher mean values on each of the other characteristics considered (indicated by the positive confidence intervals).



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:

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

In practice, σ_D is approximated with $s_D/\text{sqrt}(n)$.

Paired

(Large)

Large

For $n \geq 30$

$\bar{D} \pm Z_{\alpha/2} * \sigma_D$

$\bar{D} \pm Z_{\alpha/2} * \frac{s_D}{\sqrt{n}}$

$n = 10$

Small Sample

$n < 30$

Small Sample $n < 30$

BE	AF	Differ (D)
x_1	y_1	$x_1 - y_1$
x_2	y_2	$x_2 - y_1$
$\bar{X}$	$\bar{Y}$	

$\to s_D$

Mean Difference

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

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:

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

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:

$n < 30$

Paired Data

Difference (D)

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.

$> +ve$

$\bar{D} \rightarrow$ Mean Difference

s_D

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

The 95% confidence interval is 19.4 ± 2.262(2.836273/ √10), or (17.3712, 21.4288).

$\bar{D} = 19.4$

$s_D = 2.836273$

CL = 95%

n = 10

$1-\alpha = .95$

$\alpha = .05$

$\alpha/2 = .025$

95% CI is given by

$\bar{D} \pm t_{n-1,\alpha/2} * \frac{s}{\sqrt{n}}$

$19.4 \pm t_{9,0.025} * \frac{2.836273}{\sqrt{10}}$

$19.4 \pm 2.262 *$

19.4 ± MoE

(17.3712, 21.4288)

∴ 95% CI is (17.3712, 21.42888)

THANK YOU

D. Uma
Computer Science and Engineering

umaprabha@pes.edu

+91 99 7251 5335