MATHEMATICS FOR COMPUTER SCIENCE ENGINEERS

Unit 2: Confidence Intervals

Mamatha.H.R

Department of Computer Science and Engineering

Mathematics for Computer Science and Engineering

Confidence intervals for Difference between Two Means

Dr. Mamatha. H.R
Professor
Department of Computer Science and Engineering
PES University, Bangalore

Topics to be covered

- 5.4- Confidence Intervals for the Difference between Two Means

Textbook1: Chapter 5, section 5.4

Estimating the Difference between Two Means

- Sometimes we are interested in comparing the means of two populations.

- ❑ **The average growth of plants fed using two different nutrients.**
- ❑ **The average scores for students taught with two different teaching methods.**

- To make this comparison,

A random sample of size n_1 drawn from population 1 with mean μ_1 and variance σ_1^2.

A random sample of size n_2 drawn from population 2 with mean μ_2 and variance σ_2^2.

Notations - Comparing Two Means

	Mean	Variance	Standard Deviation
Population 1	μ_1	$\sigma_1{}^2$	σ_1
Population 2	μ_2	$\sigma_2{}^2$	σ_2

	Sample size	Mean	Variance	Standard Deviation
Sample from Population 1	n_1	$\bar{x}_1$	$s_1{}^2$	s_1
Sample from Population 2	n_2	$\bar{x}_2$	$s_2{}^2$	s_2

Estimating the Difference between Two Means

- We compare the two averages by making inferences about $\mu_1 - \mu_2$, the difference in the two population averages.
- If the two population averages are the same, then $\mu_1 - \mu_2 = 0$.
- The best estimate of $\mu_1 - \mu_2$ is the difference in the two sample means,

$$\bar{x}_1 - \bar{x}_2$$

Mathematics for Computer Science and Engineering

Estimating the Difference between Two Means

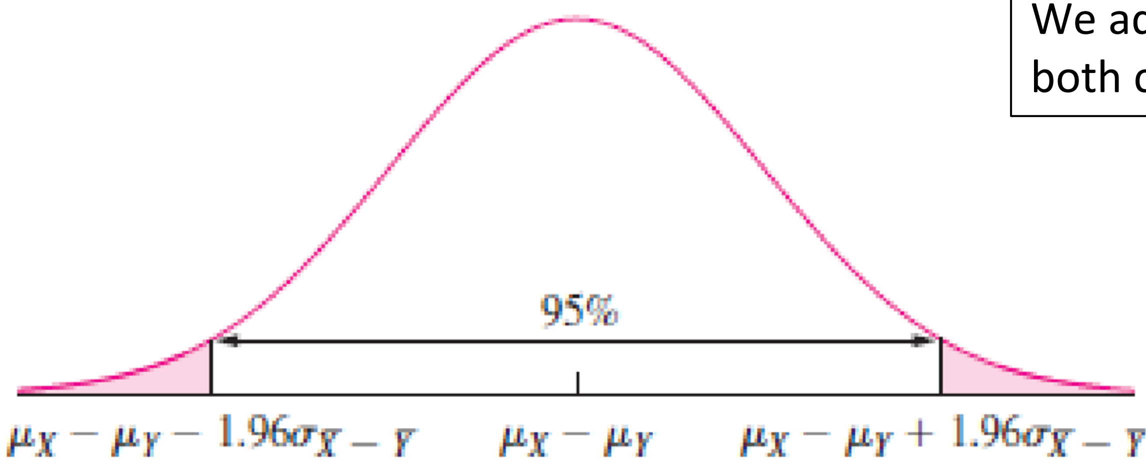
Let X and Y be independent, with $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$. Then

$$X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2) \quad (5.14)$$

$$X - Y \sim N(\mu_X - \mu_Y, \sigma_X^2 + \sigma_Y^2) \quad (5.15)$$

Note:

We add the variance for both cases.



Textbook: Page 354 onwards

Estimating the Difference between Two Means

Summary

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

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

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.

Textbook: Page 356

Estimating μ_1-μ_2

For large samples, point estimates and their margin of error as well as confidence intervals are based on the standard normal (z) distribution.

Point estimate for $\mu_1 - \mu_2$: $\{\bar{x}_1 - \bar{x}_2\} \pm MoE$; where

$$\text{Margin of Error}: \pm 1.96\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Note:
Assumption :
Both $n_1 \geq 30$ and $n_2 \geq 30$

Confidence interval for $\mu_1 - \mu_2$:

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Example 1

Compare the average daily intake of dairy products of men and women using a 95% confidence interval.

Avg Daily Intakes	Men	Women
Sample size	50	50
Sample mean	756	762
Sample Std Dev	35	30

Example Solution

Compare the average daily intake of dairy products of men and women using a 95% confidence interval.

Avg Daily Intakes	Men	Women
Sample size	50	50
Sample mean	756	762
Sample Std Dev	35	30

$$(\bar{x}_1 - \bar{x}_2) \pm 1.96\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad \Rightarrow (756 - 762) \pm 1.96\sqrt{\frac{35^2}{50} + \frac{30^2}{50}}$$

$$\Rightarrow \quad -6 \pm 12.78$$

$$\text{or } -18.78 < \mu_1 - \mu_2 < 6.78.$$

Example Solution, continued

- Could you conclude, based on this confidence interval, that there is a difference in the average daily intake of dairy products for men and women?

$$-18.78 < \mu_1 - \mu_2 < 6.78$$

- The confidence interval **contains the value μ_1-μ_2= 0.** . Therefore, it is possible that $\mu_1 = \mu_2$.

- **You would not want to conclude that there is a difference in average** daily intake of dairy products for men and women.

Example 2

The chemical composition of soil varies with depth. The article "Sampling Soil Water in Sandy Soils: Comparative Analysis of SomeCommon Methods" (M. Ahmed, M. Sharma, et al., *Communications in Soil Science and Plant Analysis*, 2001: 1677–1686) describes chemical analyses of soil taken from a farm in Western Australia. Fifty specimens were each taken at depths 50 and 250 cm. At a depth of 50 cm, the average NO3 concentration (in mg/L) was 88.5 with a standard deviation of 49.4. At a depth of 250 cm, the average concentration was 110.6 with a standard deviation of 51.5. Find a 95% confidence for the difference between the NO3 concentrations at the two depths.

Example Solution

Solution

Let $X_1, \ldots, X_{50}$ represent the concentrations of the 50 specimens taken at 50 cm, and let $Y_1, \ldots, Y_{50}$ represent the concentrations of the 50 specimens taken at 250 cm. Then $\overline{X} = 88.5$, $\overline{Y} = 110.6$, $s_X = 49.4$, and $s_Y = 51.5$. The sample sizes are $n_X = n_Y = 50$. Both samples are large, so we can use expression (5.16). Since we want a 95% confidence interval, $z_{\alpha/2} = 1.96$. The 95% confidence interval for the difference $\mu_Y - \mu_X$ is $110.6 - 88.5 \pm 1.96\sqrt{49.4^2/50 + 51.5^2/50}$, or 22.1 ± 19.8.

A 95% confidence interval for the difference in mean voltage output between two types of batteries is reported as (1.2, 2.8). A student interprets this result by stating:

“There is a 95% probability that the true difference in mean voltage is between 1.2 and 2.8 volts.”

Which of the following best evaluates the student’s statement?

A.The statement is correct because the confidence level directly reflects the probability the parameter lies within the computed interval.
B. The statement is incorrect because the confidence interval applies to sample data, not the population.
C. The statement is incorrect; the correct interpretation is that if we repeated the experiment many times, 95% of such intervals would contain the true difference in means.
D. The statement is correct only if the sample size exceeds 30, due to the Central Limit Theorem.

Solution: C

References

"Statistics for Engineers and Scientists", William Navidi, McGraw Hill Education, India, 6th Edition, 2024.

Dr. Mamatha. H.R
Professor
Department of Computer Science and Engineering
PES University
Bangalore