# Chapter 8: Confidence Interval Estimation: Further Topics

Department of Mathematics
Izmir University of Economics

Week 11
2014-2015

# Introduction

In this chapter we will focus on

- inferential statements concerning estimation of estimation of **two** populations and
- the comparison of two means from normally distributed populations (mainly) and two population proportions.

# Confidence Intervals for the Difference Between TWO Normally Distributed Population Means

Our analysis is based on the sample results. We present two sampling schemes for analyzing means:

- The first sampling scheme is for **dependent samples** and
- the second scheme is for **independent samples**.

For independent samples we will discuss the theory (similar to Chapter 7)

- when population variance is **known** and
- when the population variance is **unknown**.

When the population variance is unknown we consider the following cases:

- assuming that the population variances are **equal** or
- **NOT** assuming that they are **equal**.

## Confidence Intervals for the Difference Between TWO Normal Population Means: DEPENDENT Samples

In this section, we first deal with the sampling schemes for **DEPENDENT** samples, that is, the values in one sample are influenced by the values in the other sample. This procedure can be also called as **matched pairs**.

**Assumptions:**

- Let $x_1, x_2, \ldots, x_n$ be the values of the observations from the population with mean $\mu_x$,
- $y_1, y_2, \ldots, y_n$ be the **matched sampled values** from the population with mean $\mu_y$,
- $d_i = x_i - y_i$ be the $n$ differences,
- $\bar{d} = \frac{\sum d_i}{n}$ be the sample mean for the $n$ differences, and
- $s_d = \sqrt{\frac{\sum (d_i - \bar{d})^2}{n-1}}$ be the sample standard deviation for the $n$ differences.
- The population distribution of the differences is normal.

The $100(1-\alpha)\%$ confidence interval for the difference between means $\mu_d = \mu_x - \mu_y$ is

$$\bar{d} - t_{n-1,\frac{\alpha}{2}} \frac{s_d}{\sqrt{n}} < \mu_d < \bar{d} + t_{n-1,\frac{\alpha}{2}} \frac{s_d}{\sqrt{n}}.$$

The confidence interval can be also written as

$$\bar{d} \pm \mathrm{ME}.$$

**Example:** A medical study was conducted to compare the difference in effectiveness of two particular drugs in lowering cholesterol levels. The research team used a paired sample approach to control variation in reduction that might be due to factors other than the drug itself. Each member of a pair was matched by age, weight, lifestyle, and other pertinent factors. Drug $X$ was given to one person randomly selected in each pair and drug $Y$ was given to the other individual in the pair. After a specified amount of time each persons cholesterol level was measured again. Suppose that a random sample of nine pairs of patients with known cholesterol problems is selected from the large populations of participants. Estimate with a 99% confidence level the mean difference in the effectiveness of the two drugs $X$ and $Y$ to lower cholesterol where

| Pair | Drug $X$ | Drug $Y$ | Differences ($d_i = x_i - y_i$) |
|---|---|---|---|
| 1 | 29 | 26 | 3 |
| 2 | 32 | 27 | 5 |
| 3 | 31 | 28 | 3 |
| 4 | 32 | 27 | 5 |
| 5 | 30 | | |
| 6 | 32 | 30 | 2 |
| 7 | 29 | 26 | 3 |
| 8 | 31 | 33 | -2 |
| 9 | 30 | 36 | -6 |

# Confidence Intervals for the Difference Between TWO Normal Population Means: INDEPENDENT Samples with KNOWN population variances

**Assumptions:**
Suppose we have

- two **independent random samples** of sizes $n_x$ and $n_y$ observations
- from normally distributed populations with means $\mu_x$ and $\mu_y$, and
- variances $\sigma_x^2$ and $\sigma_y^2$.
- The sample means are $\bar{x}$ and $\bar{y}$.

The $100(1-\alpha)\%$ confidence interval for $\mu_x - \mu_y$ is

$$\bar{x} - \bar{y} - z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}} < \mu_x - \mu_y < \bar{x} - \bar{y} + z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}.$$

The confidence interval can be also written as

$$\bar{x} - \bar{y} \pm \mathrm{ME}.$$

**Example:** From a very large university, independent random samples of 120 students majoring in marketing and 90 students majoring in finance were selected. The mean GPA for the random sample of marketing majors was found to be 3.08, and the mean GPA for the random sample of finance majors was 2.88. From similar past studies the population standard deviation for the marketing majors is assumed to be 0.42; similarly, the population standard deviation for the finance majors is 0.64. Denoting the population mean for marketing majors by $\mu_x$ and the population mean for finance majors by $\mu_y$, find a 95 % confidence interval for $\mu_x - \mu_y$.

# Confidence Intervals for the Difference Between TWO Normal Population Means: INDEPENDENT Samples with UNKNOWN population variances

In this section we consider the case where the population variances are unknown. At this point, one should ask whether the unknown population variances are assumed to be equal or not.

For this purpose, we present two situations:
**A.** Population variances are assumed to be **EQUAL**.
**B.** Population variances are **NOT** assumed to be **EQUAL**.

# Confidence Intervals for the Difference Between TWO Normal Population Means: INDEPENDENT Samples with UNKNOWN population variances (Case A: Population variances are assumed to be EQUAL)

**Assumptions:**
Suppose we have

- two **independent random samples** of sizes $n_x$ and $n_y$ observations
- from normally distributed populations with means $\mu_x$ and $\mu_y$, and
- an **unknown** but **common** variance.
- The observed sample means are $\bar{x}$ and $\bar{y}$ and
- the observed sample variances are $s_x^2$ and $s_y^2$.

The $100(1-\alpha)\%$ confidence interval for $\mu_x - \mu_y$ is

$$\bar{x} - \bar{y} - t_{n_x+n_y-2,\frac{\alpha}{2}}\sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}} < \mu_x - \mu_y < \bar{x} - \bar{y} + t_{n_x+n_y-2,\frac{\alpha}{2}}\sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}},$$

where

$$s_p^2 = \frac{(n_x-1)s_x^2 + (n_y-1)s_y^2}{n_x + n_y - 2}$$

is called the **pooled sample variance**.

The confidence interval can be also written as $\bar{x} - \bar{y} \pm \mathrm{ME}$.

**Example:** The residents of St.Paul, Minnesota, complain that traffic speeding fines given in their city are higher than the traffic speeding fines that are given in nearby Minneapolis. Independent random samples of the amounts paid by residents for speeding tickets in each of the two cities over the last 3 months were obtained. These amounts were

| St.Paul | Minneapolis |
|---|---|
| 100 | 95 |
| 125 | 87 |
| 135 | 100 |
| 128 | 75 |
| 140 | 110 |
| 142 | 105 |
| 128 | 85 |
| 137 | 95 |
| 156 | |
| 142 | |

Assuming equal population variances, find a 95% confidence interval for the difference in the mean costs of speeding tickets in these two cities.

# Confidence Intervals for the Difference Between TWO Normal Population Means: INDEPENDENT Samples with UNKNOWN population variances (Case B: Population variances are NOT assumed to be EQUAL)

**Assumptions:**
Suppose we have

- two **independent random samples** of sizes $n_x$ and $n_y$ observations
- from normally distributed populations with means $\mu_x$ and $\mu_y$, and
- **unknown** and **NOT EQUAL** variances.
- The observed sample means are $\bar{x}$ and $\bar{y}$ and
- the observed sample variances are $s_x^2$ and $s_y^2$.

The $100(1-\alpha)\%$ confidence interval for $\mu_x - \mu_y$ is

$$\bar{x} - \bar{y} - t_{v,\frac{\alpha}{2}} \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}} < \mu_x - \mu_y < \bar{x} - \bar{y} + t_{v,\frac{\alpha}{2}} \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}},$$

where the degrees of freedom is

$$v = \frac{\left[\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}\right]^2}{\frac{\left(\frac{s_x^2}{n_x}\right)^2}{n_x - 1} + \frac{\left(\frac{s_y^2}{n_y}\right)^2}{n_y - 1}}.$$

The confidence interval can be also written as

$$\bar{x} - \bar{y} \pm \mathrm{ME}.$$

**Example:** An accounting firm conducted a random sample of the accounts payable for the east and the west offices of one of its clients. Among these two independent samples they wanted to estimate the difference between the population mean values of the payables. The sample statistics obtained were as follows:

| | East office(population $x$) | West office (population $y$) |
|---|---|---|
| Sample mean | 290 | 250 |
| Sample size | 16 | 11 |
| Sample standard deviation | 15 | 50 |

We do not assume that the unknown population variances are equal. Estimate the difference between the mean values of the payables for the two offices. Use a 95% confidence level.

## Confidence Interval for the Difference Between TWO Population Proportions (Large Samples)

Often a comparison of two population proportions is of much interest. In this section, we consider confidence intervals for the difference between two population proportions with independent large samples taken from these two populations.

**Assumptions:**
Suppose we have

- two **independent random samples** of sizes $n_x$ and $n_y$ with proportions of "successes" $p_x$ and $p_y$, respectively.
- **Sample proportions** are denoted as $\hat{p}_x$ and $\hat{p}_y$.

Note that, we examine the random variable $\hat{p}_x - \hat{p}_y$.

If the sample sizes are large (generally at least 40 observations in each sample), a $100(1-\alpha)\%$ confidence interval for the difference between two population proportions, $(p_x - p_y)$, is

$$\hat{p}_x - \hat{p}_y - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x} + \frac{\hat{p}_y(1-\hat{p}_y)}{n_y}} < p_x - p_y < \hat{p}_x - \hat{p}_y - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x} + \frac{\hat{p}_y(1-\hat{p}_y)}{n_y}}.$$

The confidence interval can be also written as

$$(\hat{p}_x - \hat{p}_y) \pm ME.$$

**Example:** During a presidential election year, many forecasts are made to determine how voters perceive a particular candidate. In a random sample of 120 registered voters in precinct $X$, 107 indicated that they supported the candidate in question. In an independent random sample of 141 registered voters in precinct $Y$, only 73 indicated support for the same candidate. The respective population proportions are denoted $p_x$ and $p_y$. Find a 95 % confidence interval for the population difference $(p_x - p_y)$.

**Example:** In a random sample of $n_x = 120$ large retailers, 85 used regression as a method of forecasting. In an independent random sample of $n_y = 163$ small retailers, 78 used regression as a method of forecasting. Find a 98% confidence interval for the difference between the two population proportions.