

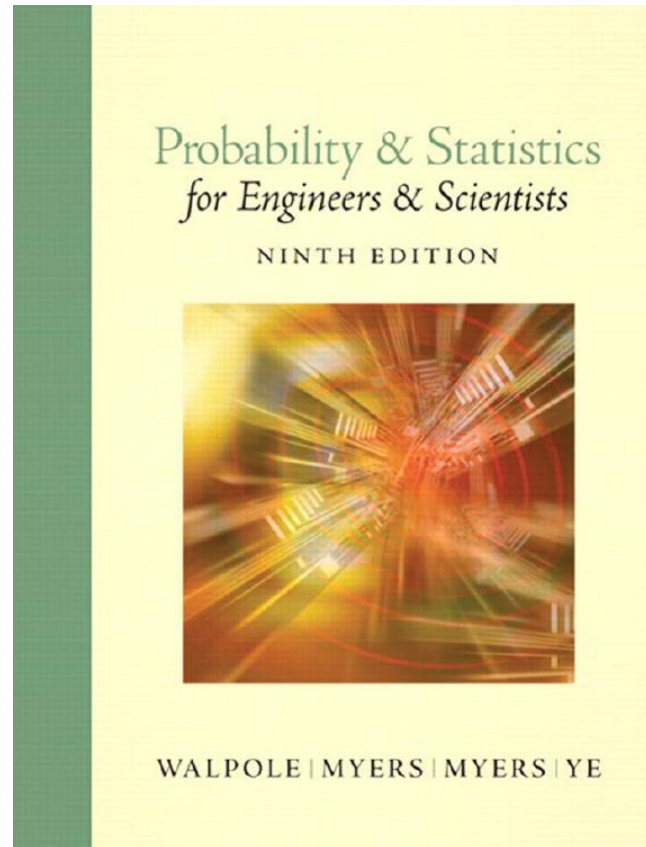
# Statistical Analysis

---

Lecture 05

# Books

---



# PowerPoint

<http://www.bu.edu.eg/staff/ahmedaboalatah14-courses/14767>

The screenshot shows a web interface for Benha University. At the top, there is a blue header with the university logo, the name 'Benha University', and a welcome message for 'Ahmed Hassan Ahmed Abu El Atta' with a 'Log out' link. Below the header, a navigation menu on the left lists various university services. The main content area displays course details for 'Automata and Formal Languages' by 'Ass. Lect. Ahmed Hassan Ahmed Abu El Atta'. The details are presented in a table with blue headers and white content. A 'Course password' section is also visible. On the right side, there are social media icons and a vertical toolbar with icons for Google, a book, RG, LinkedIn, Facebook, Twitter, Google+, YouTube, WordPress, a camera, a globe, a question mark, and an edit icon.

Benha University

Staff Search: **Welcome: Ahmed Hassan Ahmed Abu El Atta (Log out)**

You are in: [Home](#) / [Courses](#) / [Automata and Formal Languages](#) [Back To Courses](#)

Ass. Lect. Ahmed Hassan Ahmed Abu El Atta :: Course Details:  
Automata And Formal Languages [add course](#) | [edit course](#)

Course name	Automata and Formal Languages
Level	Undergraduate
Last year taught	2018
Course description	Not Uploaded
Course password	
Course files	<a href="#">add files</a>
Course URLs	<a href="#">add URLs</a>
Course assignments	<a href="#">add assignments</a>
Course Exams & Model Answers	<a href="#">add exams</a>

(edit)

# One- and Two- Sample Estimation Problems

---

CHAPTER 9

# 9.8 Two Samples: Estimating the Difference between Two Means

---

# Confidence Interval for $\mu_1 - \mu_2$ $\sigma_1^2$ and $\sigma_2^2$ Known

---

If we have two populations with means  $\mu_1$  and  $\mu_2$  and variances  $\sigma_1^2$  and  $\sigma_2^2$ , respectively, a point estimator of the difference between  $\mu_1$  and  $\mu_2$  is given by the statistic  $\bar{X}_1 - \bar{X}_2$ . Therefore, to obtain a point estimate of  $\mu_1 - \mu_2$ , we shall select two independent random samples, one from each population, of sizes  $n_1$  and  $n_2$ , and compute  $\bar{x}_1 - \bar{x}_2$ , the difference of the sample means. Clearly, we must consider the sampling distribution of  $\bar{X}_1 - \bar{X}_2$ .

According to Theorem 8.3, we can expect the sampling distribution of  $\bar{X}_1 - \bar{X}_2$  to be approximately normally distributed with mean  $\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2$  and standard deviation  $\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$ . Therefore, we can assert with a probability of  $1 - \alpha$  that the standard normal variable

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

# Confidence Interval for $\mu_1 - \mu_2$ $\sigma_1^2$ and $\sigma_2^2$ Known

---

will fall between  $-z_{\alpha/2}$  and  $z_{\alpha/2}$ . Referring once again to Figure 9.2, we write

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha.$$

Substituting for  $Z$ , we state equivalently that

$$P\left(-z_{\alpha/2} < \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} < z_{\alpha/2}\right) = 1 - \alpha,$$

which leads to the following  $100(1 - \alpha)\%$  confidence interval for  $\mu_1 - \mu_2$ .

# Confidence Interval for $\mu_1 - \mu_2$ $\sigma^2_1$ and $\sigma^2_2$ Known

---

If  $\bar{x}_1$  and  $\bar{x}_2$  are means of independent random samples of sizes  $n_1$  and  $n_2$  from populations with known variances  $\sigma_1^2$  and  $\sigma_2^2$ , respectively, a  $100(1 - \alpha)\%$  confidence interval for  $\mu_1 - \mu_2$  is given by

$$(\bar{x}_1 - \bar{x}_2) - z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}},$$

where  $z_{\alpha/2}$  is the  $z$ -value leaving an area of  $\alpha/2$  to the right.



# Example 9.10:

---

A study was conducted in which two types of engines,  $A$  and  $B$ , were compared. Gas mileage, in miles per gallon, was measured. Fifty experiments were conducted using engine type  $A$  and 75 experiments were done with engine type  $B$ . The gasoline used and other conditions were held constant. The average gas mileage was 36 miles per gallon for engine  $A$  and 42 miles per gallon for engine  $B$ . Find a 96% confidence interval on  $\mu_B - \mu_A$ , where  $\mu_A$  and  $\mu_B$  are population mean gas mileages for engines  $A$  and  $B$ , respectively. Assume that the population standard deviations are 6 and 8 for engines  $A$  and  $B$ , respectively.

# Example 9.10 Solution :

---

The point estimate of  $\mu_B - \mu_A$  is  $\bar{x}_B - \bar{x}_A = 42 - 36 = 6$ . Using  $\alpha = 0.04$ , we find  $z_{0.02} = 2.05$  from Table A.3. Hence, with substitution in the formula above, the 96% confidence interval is

$$6 - 2.05\sqrt{\frac{64}{75} + \frac{36}{50}} < \mu_B - \mu_A < 6 + 2.05\sqrt{\frac{64}{75} + \frac{36}{50}},$$

or simply  $3.43 < \mu_B - \mu_A < 8.57$ . 

# The Cases of $\sigma_1$ and $\sigma_2$ are Unknown

---

# Variances Unknown but Equal

$$\sigma^2_1 = \sigma^2_2$$

---

Consider the case where  $\sigma^2_1$  and  $\sigma^2_2$  are unknown. If  $\sigma^2_1 = \sigma^2_2 = \sigma^2$ , we obtain a standard normal variable of the form

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma^2[(1/n_1) + (1/n_2)]}}.$$

According to Theorem 8.4, the two random variables

$$\frac{(n_1 - 1)S_1^2}{\sigma^2} \quad \text{and} \quad \frac{(n_2 - 1)S_2^2}{\sigma^2}$$

have chi-squared distributions with  $n_1 - 1$  and  $n_2 - 1$  degrees of freedom, respectively. Furthermore, they are independent chi-squared variables, since the random samples were selected independently. Consequently, their sum

$$V = \frac{(n_1 - 1)S_1^2}{\sigma^2} + \frac{(n_2 - 1)S_2^2}{\sigma^2} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{\sigma^2}$$

has a chi-squared distribution with  $v = n_1 + n_2 - 2$  degrees of freedom.

# Variances Unknown but Equal

$$\sigma^2_1 = \sigma^2_2$$

---

Since the preceding expressions for  $Z$  and  $V$  can be shown to be independent, it follows from Theorem 8.5 that the statistic

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma^2[(1/n_1) + (1/n_2)]}} \bigg/ \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{\sigma^2(n_1 + n_2 - 2)}}$$

has the  $t$ -distribution with  $v = n_1 + n_2 - 2$  degrees of freedom.

# Pooled Estimate of Variance $S_p$

---

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

# Confidence Interval for $\mu_1 - \mu_2$ , $\sigma^2_1 = \sigma^2_2$ but Both Unknown

---

Substituting  $S_p^2$  in the  $T$  statistic, we obtain the less cumbersome form

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p \sqrt{(1/n_1) + (1/n_2)}}.$$

Using the  $T$  statistic, we have

$$P(-t_{\alpha/2} < T < t_{\alpha/2}) = 1 - \alpha,$$

where  $t_{\alpha/2}$  is the  $t$ -value with  $n_1 + n_2 - 2$  degrees of freedom, above which we find an area of  $\alpha/2$ . Substituting for  $T$  in the inequality, we write

$$P \left[ -t_{\alpha/2} < \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p \sqrt{(1/n_1) + (1/n_2)}} < t_{\alpha/2} \right] = 1 - \alpha.$$

After the usual mathematical manipulations, the difference of the sample means  $\bar{x}_1 - \bar{x}_2$  and the pooled variance are computed and then the following  $100(1 - \alpha)\%$  confidence interval for  $\mu_1 - \mu_2$  is obtained.

# Confidence Interval for $\mu_1 - \mu_2$ , $\sigma^2_1 = \sigma^2_2$ but Both Unknown

---

If  $\bar{x}_1$  and  $\bar{x}_2$  are the means of independent random samples of sizes  $n_1$  and  $n_2$ , respectively, from approximately normal populations with unknown but equal variances, a  $100(1 - \alpha)\%$  confidence interval for  $\mu_1 - \mu_2$  is given by

$$(\bar{x}_1 - \bar{x}_2) - t_{\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + t_{\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

where  $s_p$  is the pooled estimate of the population standard deviation and  $t_{\alpha/2}$  is the  $t$ -value with  $v = n_1 + n_2 - 2$  degrees of freedom, leaving an area of  $\alpha/2$  to the right.



# Example 9.11:

---

The article “Macroinvertebrate Community Structure as an Indicator of Acid Mine Pollution,” published in the *Journal of Environmental Pollution*, reports on an investigation undertaken in Cane Creek, Alabama, to determine the relationship between selected physiochemical parameters and different measures of macroinvertebrate community structure. One facet of the investigation was an evaluation of the effectiveness of a numerical species diversity index to indicate aquatic degradation due to acid mine drainage. Conceptually, a high index of macroinvertebrate species diversity should indicate an unstressed aquatic system, while a low diversity index should indicate a stressed aquatic system.

Two independent sampling stations were chosen for this study, one located downstream from the acid mine discharge point and the other located upstream. For 12 monthly samples collected at the downstream station, the species diversity index had a mean value  $\bar{x}_1 = 3.11$  and a standard deviation  $s_1 = 0.771$ , while 10 monthly samples collected at the upstream station had a mean index value  $\bar{x}_2 = 2.04$  and a standard deviation  $s_2 = 0.448$ . Find a 90% confidence interval for the difference between the population means for the two locations, assuming that the populations are approximately normally distributed with equal variances.

# Example 9.11 Solution :

---

Let  $\mu_1$  and  $\mu_2$  represent the population means, respectively, for the species diversity indices at the downstream and upstream stations. We wish to find a 90% confidence interval for  $\mu_1 - \mu_2$ . Our point estimate of  $\mu_1 - \mu_2$  is


$$\bar{x}_1 - \bar{x}_2 = 3.11 - 2.04 = 1.07.$$

The pooled estimate,  $s_p^2$ , of the common variance,  $\sigma^2$ , is

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(11)(0.771^2) + (9)(0.448^2)}{12 + 10 - 2} = 0.417.$$

Taking the square root, we obtain  $s_p = 0.646$ . Using  $\alpha = 0.1$ , we find in Table A.4 that  $t_{0.05} = 1.725$  for  $v = n_1 + n_2 - 2 = 20$  degrees of freedom. Therefore, the 90% confidence interval for  $\mu_1 - \mu_2$  is

$$1.07 - (1.725)(0.646)\sqrt{\frac{1}{12} + \frac{1}{10}} < \mu_1 - \mu_2 < 1.07 + (1.725)(0.646)\sqrt{\frac{1}{12} + \frac{1}{10}},$$

which simplifies to  $0.593 < \mu_1 - \mu_2 < 1.547$ . 

# Unknown and Unequal Variances $\sigma^2_1 \neq \sigma^2_2$

---

Let us now consider the problem of finding an interval estimate of  $\mu_1 - \mu_2$  when the unknown population variances are not likely to be equal. The statistic most often used in this case is

$$T' = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{(S_1^2/n_1) + (S_2^2/n_2)}},$$

which has approximately a  $t$ -distribution with  $v$  degrees of freedom, where

$$v = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{[(s_1^2/n_1)^2/(n_1 - 1)] + [(s_2^2/n_2)^2/(n_2 - 1)]}.$$

Since  $v$  is seldom an integer, we *round it down* to the nearest whole number. The above estimate of the degrees of freedom is called the Satterthwaite approximation (Satterthwaite, 1946, in the Bibliography).

# Confidence Interval for $\mu_1 - \mu_2$ , $\sigma_1^2 \neq \sigma_2^2$ but Both Unknown

---

Using the statistic  $T'$ , we write

$$P(-t_{\alpha/2} < T' < t_{\alpha/2}) \approx 1 - \alpha,$$

# Confidence Interval for $\mu_1 - \mu_2$ , $\sigma^2_1 \neq \sigma^2_2$ but Both Unknown

---

If  $\bar{x}_1$  and  $s_1^2$  and  $\bar{x}_2$  and  $s_2^2$  are the means and variances of independent random samples of sizes  $n_1$  and  $n_2$ , respectively, from approximately normal populations with unknown and unequal variances, an approximate  $100(1 - \alpha)\%$  confidence interval for  $\mu_1 - \mu_2$  is given by

$$(\bar{x}_1 - \bar{x}_2) - t_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + t_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},$$

where  $t_{\alpha/2}$  is the  $t$ -value with

$$v = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{[(s_1^2/n_1)^2/(n_1 - 1)] + [(s_2^2/n_2)^2/(n_2 - 1)]}$$

degrees of freedom, leaving an area of  $\alpha/2$  to the right.

# Example 9.12:

---

A study was conducted by the Department of Zoology at the Virginia Tech to estimate the difference in the amounts of the chemical orthophosphorus measured at two different stations on the James River. Orthophosphorus was measured in milligrams per liter. Fifteen samples were collected from station 1, and 12 samples were obtained from station 2. The 15 samples from station 1 had an average orthophosphorus content of 3.84 milligrams per liter and a standard deviation of 3.07 milligrams per liter, while the 12 samples from station 2 had an average content of 1.49 milligrams per liter and a standard deviation of 0.80 milligram per liter. Find a 95% confidence interval for the difference in the true average orthophosphorus contents at these two stations, assuming that the observations came from normal populations with different variances.

# Example 9.12 Solution :

---

For station 1, we have  $\bar{x}_1 = 3.84$ ,  $s_1 = 3.07$ , and  $n_1 = 15$ . For station 2,  $\bar{x}_2 = 1.49$ ,  $s_2 = 0.80$ , and  $n_2 = 12$ . We wish to find a 95% confidence interval for  $\mu_1 - \mu_2$ .

Since the population variances are assumed to be unequal, we can only find an approximate 95% confidence interval based on the  $t$ -distribution with  $v$  degrees of freedom, where

$$v = \frac{(3.07^2/15 + 0.80^2/12)^2}{[(3.07^2/15)^2/14] + [(0.80^2/12)^2/11]} = 16.3 \approx 16.$$

Our point estimate of  $\mu_1 - \mu_2$  is

$$\bar{x}_1 - \bar{x}_2 = 3.84 - 1.49 = 2.35.$$

Using  $\alpha = 0.05$ , we find in Table A.4 that  $t_{0.025} = 2.120$  for  $v = 16$  degrees of freedom. Therefore, the 95% confidence interval for  $\mu_1 - \mu_2$  is

$$2.35 - 2.120\sqrt{\frac{3.07^2}{15} + \frac{0.80^2}{12}} < \mu_1 - \mu_2 < 2.35 + 2.120\sqrt{\frac{3.07^2}{15} + \frac{0.80^2}{12}},$$

which simplifies to  $0.60 < \mu_1 - \mu_2 < 4.10$ . Hence, we are 95% confident that the interval from 0.60 to 4.10 milligrams per liter contains the difference of the true average orthophosphorus contents for these two locations. 