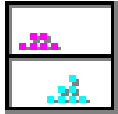


CHAPTER 10



Visualizing Two-Sample Hypothesis Tests

CONCEPTS

- Null Hypothesis, Alternative Hypothesis, Two-Sample Tests, One-Tail Test, Two-Tail Test, Confidence Interval, Decision Rule, Level of Significance, Sampling Distribution, Critical Value, Test Statistic, Type I Error, Type II Error, Power

OBJECTIVES

- Become familiar with the sampling distributions used in two-sample parametric tests for means and variances
- Understand the relationship between a confidence interval and the corresponding two-sample hypothesis test
- Be able to explain the meaning of Type I error, Type II error, and power
- Know the assumptions underlying two-sample hypothesis tests and recognize the effects of violating each of them

Overview of Concepts

A one-sample test compares a sample statistic with an assumed parameter value (usually a historic benchmark). In contrast, a **two-sample test** compares two sample statistics with each other. Usually, the main issue is simply whether or not the two samples differ (e.g., in terms of means or variances). The first step is to create two mutually exclusive hypotheses that exhaust the possibilities. For example, in a laptop PC battery test, does brand A have a different mean rundown time than brand B? The hypotheses are

$H_0: \mu_A = \mu_B$ (mean battery rundown time is the same for both brands)

$H_1: \mu_A \neq \mu_B$ (mean battery rundown time differs for the two brands)

These hypotheses are a special case of a more general test for a specified difference of means:

$H_0: \mu_A - \mu_B = D$ (mean difference in battery rundown time is D)

$H_1: \mu_A - \mu_B \neq D$ (mean difference in battery rundown time is not D)

However, since $D = 0$ is the usual choice, only the first form will be discussed. The **null hypothesis** (H_0) is the statement or belief that we try to reject. The **alternative hypothesis** (H_1) is the exact converse of the null hypothesis. Evidence is gathered by testing randomly chosen batteries (n_A of brand A and n_B of brand B). Samples will be as large as possible, given the constraints of time and budget for the test. Equal sample sizes are logical, but are not necessary (and sometimes impossible). A **two-tail test** is commonly used when there is no prior belief about directionality. A **one-tail test** (e.g., $H_0: \mu_A \leq \mu_B$ and $H_1: \mu_A > \mu_B$) might be relevant if there are prior reasons (such as a manufacturer's claim) to believe that one brand lasts longer than the other.

If the **confidence interval** around the estimated difference of means includes zero ($D = 0$) we would reject H_0 . But the usual way to test the hypothesis is to compute a **test statistic** and compare it with a **critical value** using our knowledge of the **sampling distribution** of the test statistic. The sampling distribution of the test statistic for the difference of two means is standard normal if the population variances are known or Student's t if the population variances are unknown. If the variances are unknown but are assumed equal, the t -test uses the sum of the sample degrees of freedom $(n_A - 1) + (n_B - 1)$. If the variances are unknown but are assumed unequal, a more complex approach is required (see Behrens-Fisher problem in the Glossary). The **decision rule** states the criterion for rejecting H_0 . If H_0 is not rejected, it is tentatively accepted (many statisticians prefer to say that H_0 is "not rejected").

If H_0 is true and we mistakenly reject it, we have committed **Type I error**, while if a false H_0 is mistakenly accepted we have committed **Type II error**. Otherwise, no error has been committed. The probability of Type I error is denoted α . The probability of Type II error is denoted β . The probability of correctly rejecting a false null hypothesis is called **power** and is equal to $1 - \beta$. The value of α (called the **level of significance**) is chosen in advance (typically 0.10, 0.05, or 0.01) and is embodied in the decision rule. The value of β cannot be chosen because it depends on the true parameter value, which is usually unknown. However, power may be found for any possible true parameter value. This module allows you to estimate Type I error or power empirically by sampling a known population many times.

Illustration of Concepts

Do the weights of Bartlett pears have the same variance as the weights of D'Anjou pears? The **null hypothesis** (H_0) is that the variances are equal, while the **alternative hypothesis** (H_1) is that they are unequal. A **two-tail** test is appropriate when it does not matter which variance is greater (otherwise a **one-tail test** could be used).

$$\begin{aligned} H_0: \quad \sigma_1^2 &= \sigma_2^2 \quad \text{or} \quad \sigma_1^2 / \sigma_2^2 = 1 && \text{(population variances are the same)} \\ H_1: \quad \sigma_1^2 &\neq \sigma_2^2 \quad \text{or} \quad \sigma_1^2 / \sigma_2^2 \neq 1 && \text{(population variances are different)} \end{aligned}$$

Early in the 20th century, the statistician R. A. Fisher (1890 – 1935) showed that these hypotheses may be tested using the F **sampling distribution** (named after its discoverer).

To test these hypotheses, two samples are taken. A sample of 18 Bartlett pears shows a mean weight of 242.4 grams with a standard deviation of 21.26 grams, while a sample of 13 D'Anjou pears shows a mean weight of 236.5 grams with a standard deviation of 9.563 grams. The dot plot (Figure 1) suggests that the D'Anjou pears (large dots) have less dispersion than the Bartlett pears (small dots). Also, the confidence intervals for each variance (top of Figure 2) do not overlap. More formally, the two-sided 95% **confidence interval** for the ratio of variances (bottom of Figure 2) does not include 1, suggesting that H_0 should be rejected. If H_0 is true (and if the population is normal) the ratio of the two sample variances will follow an F distribution with $df_1 = 18 - 1$ and $df_2 = 13 - 1$. The **test statistic** is $F = s_1^2 / s_2^2 = (21.26)^2 / (9.563)^2 = 4.940$. At the 0.05 **level of significance** the lower **critical value** is 0.354 and the upper critical value is 3.129, as illustrated in the **decision rule** (Figure 3). Since the test statistic lies in the right-tail rejection region, we reject H_0 and conclude that the population variances are unequal. The F-test must agree with the confidence interval approach.

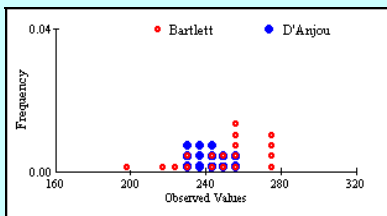


Figure 1: Dot Plot

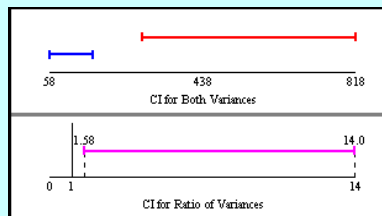


Figure 2: Confidence Intervals

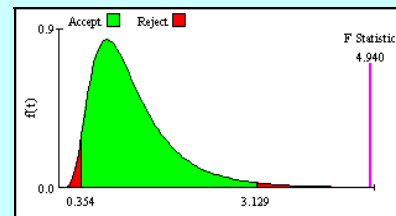


Figure 3: Decision Rule

Could different samples lead to different results? Yes. This sampling experiment was repeated four more times. The results of all five samples are shown in the table below. Samples 1, 2, 3, and 5 lead to rejection (and possible **Type I error**) while sample 4 leads to acceptance (and possible **Type II error**). Which decision is correct? We don't know. The **power** of the test is unknown since the true population variances are unknown. Uncertainty is inherent in all statistical decisions. Since only one sample usually is taken, we must accept the possibility of Type I or Type II error in any decision we make.

Sample	s_1^2	s_2^2	F Statistic	Decision
1	451.8	91.46	4.940	Reject
2	549.6	89.91	6.113	Reject
3	577.2	110.9	5.204	Reject
4	318.8	124.9	2.552	Accept
5	611.9	64.99	9.415	Reject

Orientation to Basic Features

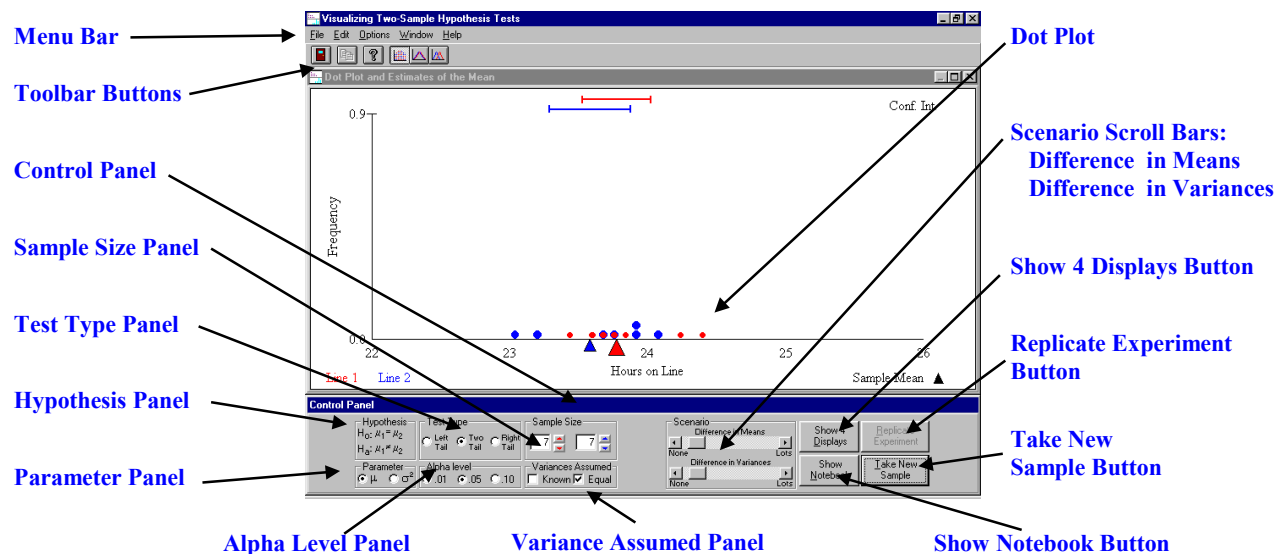
This module helps you learn how to test hypotheses for two means (with known or unknown population variances) and for two variances. It permits one-tail and two-tail tests and shows the relationship between confidence intervals and hypothesis tests using appropriate scenarios. You may also specify two populations and set their means and standard deviations. Using replication, you can study the meaning of Type I error, Type II error, and power.

1. Select a Scenario

Start the module by clicking on the module's icon, title, or chapter number in the *Visual Statistics* menu and pressing the **Run Module** button. When the module is loaded, you will be on the introduction page of the Notebook. Read the questions and then click the **Concepts** tab to see the concepts that you will learn. Click the **Scenarios** tab. Click on **Unknown Variances Assumed Equal**. Select a scenario, read it, and press **OK**.

2. Main Display

The main display opens with a large display (Dot Plot and Estimates of the Mean) and a Control Panel. Click **Take New Sample** and observe the changes in the sampled items (red and blue dots) and confidence intervals (horizontal red and blue line segments). The sample means (red and blue fulcrums) will change to reflect each new sample. The **Replicate** button will remain inactive until you press **Show 4 Displays** (but do not do it yet).

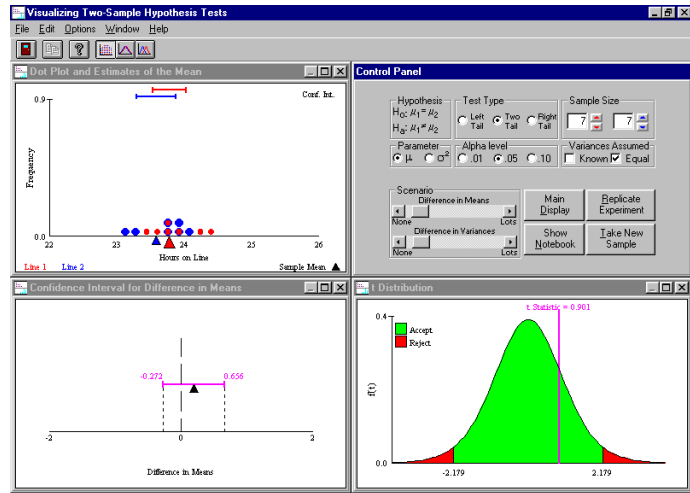


3. Toolbar Buttons

Place the cursor over the fifth toolbar button located above the dot plot. A label will appear (**Show Populations Sampled**). Click the button and see an overlay of the population to be sampled. This distribution does not change when you take samples, but will reflect any changes you make in the true or hypothesized parameters. Click the sixth toolbar button (**Show Sampling Distributions**) and observe the narrower sampling distributions. The sampling distributions will not change when you take samples, but will reflect any changes you make in the sample sizes or in the true and hypothesized parameters. Click the fourth toolbar button (**Show Dot Plot Only**) to restore the simple dot plot.

4. **Show 4 Displays**

Click **Show 4 Displays**. Read the Hint that appears and click **OK**. The Control Panel will move to the upper right and its controls will be rearranged. The dot plot will become smaller and will be in the upper left. Two new quadrant displays (Confidence Interval for Difference in Means and t Distribution) will be on the bottom. The dot plot and confidence interval will change as you take new samples, while the t Distribution display remains the same except for the t-statistic. Right-click on the Confidence Interval for Difference in Means display and select **Statistics and Parameters** to see a list of sample statistics and parameter values. Right-click on the Statistics and Parameters display and select **Analysis of Experiment** to see a verbal description of the results of your experiment. Right-click the Analysis of Experiment display and select **Confidence Interval** to see the original displays.

5. **Using the Control Panel**

In the Test Type panel, click each option (**Right-Tail**, **Left-Tail**, **Two-Tail**). The t distribution will match the test type you have selected. Click each option (**.10**, **.05**, **.01**) in the Alpha Level panel and verify that the t Distribution tail area and the confidence intervals change to reflect your choices. Check **Known** in the Variances Assumed panel and observe that the t distribution is replaced by a normal distribution. Uncheck **Known** to return to the t distribution. Uncheck **Equal** in the Variances Assumed panel. The t distribution and the confidence intervals will change. Click σ^2 in the Parameter panel. The t distribution is replaced by the F distribution and the confidence intervals now show variances instead of means. Click μ in the Parameter panel, and the displays again show a mean. Click the **Sample Size** spin button to alter the sample size (2 to 100). Click the fifth toolbar button to show an overlay of the population to be sampled. Change the position of any scroll bar in the Scenario panel (**Difference in Means**, **Difference in Variances**) and press **Take New Sample**. The population does not change as you take samples.

6. **Copying a Display**

Click on the display you wish to copy. Its window title will be highlighted. Select **Copy** from the **Edit** menu (on the menu bar) or Ctrl-C. It can then be pasted into other applications.

7. **Help**

Click on **Help** on the menu bar at the top of the screen. **Search for Help** lets you search an index for this module, **Contents** shows a table of contents for this module, **Using Help** gives instructions on how to use Help, and **About** gives licensing and copyright information.

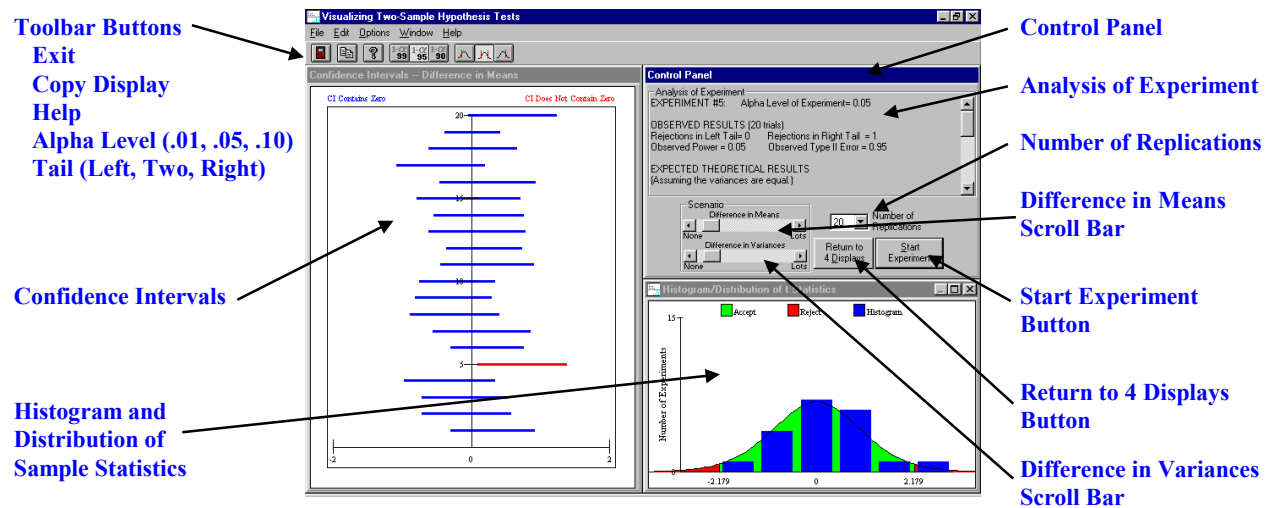
8. **Exit**

Close the module by selecting **Exit** in the **File** menu (or click  in the upper right-hand corner of the window). You will be returned to the *Visual Statistics* main menu.

Orientation to Additional Features

1. Replication

Click the **Replicate Experiment** button on the Control Panel. At the top are nine toolbar buttons. The new ones control the confidence level (**99%, 95%, 90%**) and the test type (**Left-tail, Two-Tail, Right-Tail**). On the left is the Confidence Intervals – Difference in Means display. On the right is a Control Panel containing the Analysis of Experiment window and scroll bar controls (**Difference in Means, Difference in Variances**) and a Histogram/Distribution display. Click the **Number of Replications** list box to vary the number of replications (20, 50, 100, 200, 500). If you click the **Start Experiment** button, it becomes **Finish Experiment** and **Return to 4 Displays** becomes **Pause Experiment**. If you click **Pause Experiment**, the displays are frozen in their current state and **Pause Experiment** becomes **Continue Experiment**. If you click **Finish Experiment**, the confidence interval and histogram displays go to their final state without showing any intermediate steps. Click **Return to 4 Displays** to exit replication.



2. Do-It-Yourself Controls

Click **Show Notebook**, choose the **Do-It-Yourself** tab, and click **OK**. The Parameters panel lets you change the means and standard deviations with scroll bars or by entering values into the edit boxes (not limited to the 1 to 10 scroll bar range). Choose any of six populations to sample (**Normal, Uniform, Skewed Left, Skewed Right, Very Skewed Left, Very Skewed Right**) by pressing its button. When you click **OK**, the Control Panel's scenario scroll bars are replaced by the Do-It-Yourself Summary panel shown at the right. Press **Change DIY Parameters** to use the Parameters panel again. The replication control panel also shows the Do-It-Yourself panel instead of scenario scroll bars.

Sample	μ	σ	Distribution
1	10	1	Normal
2	10	1	Normal

Change DIY Parameters

3. Options

Click **Options** on the menu bar. You may change terminology from **Accept Null Hypothesis** to **Do Not Reject Null Hypothesis**. This option may be preferred by some, since we do not actually prove a null hypothesis.

Basic Learning Exercises

Name _____

Press the **Show Notebook** button, select the **Scenarios** tab, and click on **Known Variances**. Select the **Day and Night Students** scenario, read it, and press **OK**. Make sure all four displays are showing.

Tests for Two Means with Known Variances

1. Express the hypotheses in words. Do you believe a two-tail test is appropriate? Is the null hypothesis true or false? If there is a difference in means, do you expect it will be easy to detect using sampling? **Hint:** Look at the scenario scroll bars.
2. If you are using quadrant displays, click **Main Screen**. Click the **Show Population Sampled** toolbar button. Describe the distribution of the populations being sampled. Are the two population variances equal? How do you know?
3. Click **Take New Sample** a few times. Assess the overlap of sampled items and the relative positions of the sample means (red and blue fulcrums). Then click the **Show Sampling Distribution** toolbar button. Do these displays tend to confirm or disconfirm your answer to exercise 1 regarding the anticipated degree of difficulty of detecting the actual difference in means through random sampling? Why aren't the variances of the sampling distributions identical, given that the population variances are the same?

4. Click **Show 4 Displays**. a) Why does the decision rule show a normal distribution? b) Is this normal distribution justified on the basis of sample size? Explain. c) When you press **Take New Sample**, do its critical values change? d) Does the test statistic change? Explain.

5. Click **Take New Sample** 10 times, observing how often the null hypothesis is accepted (z statistic falls in the green acceptance region or the confidence interval for the difference of means includes zero). a) Did the z-test always agree with the confidence interval approach? b) Based on sampling results alone, what decision would probably be made about H_0 ? c) If you make this decision, would you commit Type I error? Type II error? d) What do you conclude about the power of the test in this example? Should this conclusion about power be generalized?

6. Right-click on the Dot Plot and Estimates of Means display (in the upper left quadrant) and select **Statistics and Parameters** from the menu. a) What are the true population means? b) How large is the true difference in population means? c) Does the table's CI for Difference always agree with the visual display called Confidence Interval for Difference in Means? Click **Take New Sample** several times, watching these two displays.

7. Right-click on the Statistics and Parameters display (in the upper left quadrant) and select Analysis of Experiment from the menu. What does the analysis tell you?

Intermediate Learning Exercises

Name _____

Test for Means with Known Variances

Press the **Show Notebook** button, select the **Scenarios** tab, and click on **Known Variance**. Select the **Day and Night Students** scenario, read it, and press **OK**. If the dot plot display is not visible, right-click the upper left display and select **Dot Plot** from the menu.

8. Press **Replicate**. Set the number of replications to 100. Press **Start Experiment** and then **Finish Experiment** (to avoid having to watch the graphs build). A confidence interval that fails to include zero (shown in red) indicates rejection of the null hypothesis ($H_0: \mu_1 = \mu_2$). Record the number of rejections from the Analysis of Experiment window. Repeat this experiment twice. Find the total number of rejections and compute the average power. What is the advantage of doing the experiment three times? Compare the histogram of test statistics with the hypothesized normal distribution and comment on its appearance.

Experiment:	1	2	3	Total	Avg. Power
Rejections:	_____	_____	_____	_____	_____

9. Press **Return to 4 Displays**. Move the **Difference in Means** scroll bar to the middle (halfway between None and Lots). Examine the sampling distributions (by clicking the **Show Sampling Distribution** toolbar button) and make a general prediction about the power of the test. Press **Replicate**, do 100 replications three times, and record the results. Find the total number of rejections and compute the average power. Compare the histogram of test statistics to the normal distribution. Compare these results with exercise 8.

Experiment:	1	2	3	Total	Avg. Power
Rejections:	_____	_____	_____	_____	_____

10. Press **Return to 4 Displays**. Double both sample sizes (to 44 and 52). Describe the sampling distributions. Press **Replicate**. Do 100 replications three times and record the results below. Find the total number of rejections and the average power. Did the larger sample sizes improve the power? What does the histogram of test statistics tell you? Explain.

Experiment:	1	2	3	Total	Avg. Power
Rejections:	_____	_____	_____	_____	_____

Tests for Two Means with Unknown Variances

11. Press **Return to 4 Displays**. Press the **Show Notebook** button, select the **Scenarios** tab, and click on **Unknown Variances Assumed Equal**. Select the **Making Cars** scenario, read it, and click **OK**. Set the **Difference in Means** and **Difference in Variances** scroll bars to None and press **Take New Sample**. Why is the t distribution displayed instead of the normal distribution?
12. Right-click on the Confidence Interval for Difference in Means window and select **Statistics and Parameters**. Press **Take New Sample** several times. Do the degrees of freedom (df) and critical values change? Explain why or why not.
13. Set Variance Assumed to **Unequal** and press **Take New Sample** several times, watching what happens to degrees of freedom. Explain what you observe. **Hint:** Click **Help** or examine the formula for Welch's correction in the Glossary.
14. Set Variances Assumed to **Equal** and press **Replicate**. Choose 100 replications. What is the empirical Type I error? Repeat a few times. What does the histogram of test statistics tell you? Interpret these results.
15. Set the **Difference in Variances** scroll bar to Lots and again perform 100 replications. What is the empirical Type I error? Does it appear that Type I error is compromised if unknown, unequal variances are assumed equal? What does the histogram of test statistics look like?

16. Click **Return to 4 Displays**. Set Variances Assumed to **Unequal** and press **Replicate**. Choose 100 replications. What is the empirical Type I error? Repeat a few times. What does the histogram of test statistics tell you? Interpret these results.

17. Set the **Difference in Means** scroll bar to its midpoint. Write down the essential aspects of your sampling experiment, including sample size (use **Return to 4 Displays** if you need to refresh your memory). Then perform 100 replications. Estimate the empirical power. What does the histogram of test statistics look like?

18. Click **Return to 4 Displays**. Increase both sample sizes to 30, click **Replicate Experiment**, and perform 100 replications. Estimate the empirical power. Did the larger sample sizes have a strong effect on power? Compare these results with your answer in exercise 10.

Tests for Two Variances

19. Click **Return to 4 Displays**. Press the **Show Notebook** button, select the **Do-It-Yourself** tab, and press **OK**. Create unequal variances by setting $\sigma_1 = 1.5$ and $\sigma_2 = 1$ (the means are irrelevant). Make sure both populations are normal. Click **OK**. Choose *unknown* variances by unchecking **Known** in the Variances Assumed **panel**. Change Parameter from μ to σ^2 . Set both sample sizes to 10. Be sure Test Type is **Two-Tail** and Alpha Level is **0.05**. State the hypotheses in two different ways and explain their meaning. What is the true ratio of variances? Using sampling, how hard do you think it will be to detect the false null hypothesis you created?

20. Click **Take New Sample**. Why is the F distribution the relevant distribution? What are its degrees of freedom (df)? Explain. What are its critical values? **Hint:** Look at the Statistics and Parameters window or right-click on the lower left display to show the Analysis of Experiment window if you need further information.

Numerator df _____
 Denominator df _____

Left-tail critical value _____
 Right-tail critical value _____

21. Click **Take New Sample** 10 times, and count how many times the F test statistic falls in the red rejection region in either tail (refer to the F distribution that is displayed). Would you say the power of the test is high? Explain your reasoning.

22. Press **Replicate**. Set **Number of Replications** to 100 and press **Start Experiment**. Record the number of rejections. Repeat two more times. Find the total number of rejections and compute the average power. Why is Type I error not relevant?

Experiment:	1	2	3	Total	Power
Rejections:	_____	_____	_____	_____	_____

23. Describe the histogram of test statistics. Why doesn't it resemble the F distribution?

Advanced Learning Exercises

Name _____

Tests of Two Means: Skewness and Type I Error

Click **Return to 4 Displays**. Press the **Show Notebook** button, select the **Do-It-Yourself** tab, and click **OK**. Set $\mu_1 = 5$ and $\mu_2 = 5$ (equal means) and set $\sigma_1 = 1$ and $\sigma_2 = 1$ (equal variances). Make sure *both* populations are normal and click **OK**. Choose *unknown* variances by unchecking **Known** in the Variances Assumed panel. Set Parameter to μ . Set both sample sizes to 10. Set Test Type to **Two-Tail** and Alpha Level to **0.05**. Check **Equal** in the Variances Assumed panel. Make sure the four displays are showing.

24. Press **Replicate** and do 100 replications. Record the empirical Type I error (from the Analysis of Experiment window). Repeat twice and take the average. Compare the histogram of test statistics with its hypothesized Student's t sampling distribution. Use the **Change DIY Parameters** button to do the same experiment with both populations **Skewed Right**. Does skewness affect the shape of the histogram and Type I error? How much do your results vary? In terms of Type I error, is the test for two means robust to skewness?

<u>Distribution</u>	<u>Experiment 1</u>	<u>Experiment 2</u>	<u>Experiment 3</u>	<u>Avg Type I Error</u>
Normal	_____	_____	_____	_____
Skewed Right	_____	_____	_____	_____

Tests of Two Variances: Skewness and Type I Error

25. Click **Return to 4 Displays** and change Parameter to σ^2 . Increase both sample sizes to 50. Click **Change DIY Parameters** and set both populations to **Normal**. Do 100 replications and record the Type I error (from the Analysis of Experiment window). Repeat two times and take the average. Compare the histogram of test statistics with its hypothesized F sampling distribution. Use the **Change Distribution** button to do the same experiment with both populations **Skewed Right**. Describe the effects of skewness on the shape of the histogram and on Type I error. In terms of Type I error, is the test for two variances robust to skewness? Did the large sample size help? Are skewed populations encountered very often in reality?

<u>Distribution</u>	<u>Experiment 1</u>	<u>Experiment 2</u>	<u>Experiment 3</u>	<u>Avg Type I Error</u>
Normal	_____	_____	_____	_____
Skewed Right	_____	_____	_____	_____

Tests of Two Means: Skewness and Power

26. Press **Return to 4 Displays** and click **Change DIY Parameters**. To create a false null hypothesis about the means, set $\mu_1 = 6$ and $\mu_2 = 5$ (unequal means) but leave the variances unchanged at $\sigma_1 = 1.0$ and $\sigma_2 = 1.0$ (equal variances). Make both populations normal and click **OK**. Be sure the variances are still assumed unknown. Set both sample sizes to 10. Set Test Type to **Two-Tail**, Alpha Level to **0.05**, and Variances Assumed to **Equal**. Click the **Show Sampling Distributions** toolbar button. From the appearance of these distributions, will this test of difference of means have high power? Is Type I error relevant here? Explain.
27. Press **Replicate** and do 100 replications. Record the empirical power (from the Analysis of Experiment window). Repeat twice and take the average. Compare the histogram of test statistics with the hypothesized Student's t sampling distribution. Use the **Change DIY Parameters** button to do the same experiment with both populations skewed right. How much do your results vary? Does skewness adversely affect power in a test for two means?

<u>Distribution</u>	<u>Experiment 1</u>	<u>Experiment 2</u>	<u>Experiment 3</u>	<u>Avg Power</u>
Normal	_____	_____	_____	_____
Skewed Right	_____	_____	_____	_____

Tests of Two Variances: Skewness and Power

28. Press **Return to 4 Displays** and click **Change DIY Parameters**. Create a false null hypothesis about variances with $\mu_1 = 5$ and $\mu_2 = 5$ (equal means) and $\sigma_1 = 1.5$ and $\sigma_2 = 1.0$ (unequal variances). Make sure both populations are *normal* and click **OK**. Change **Parameter** to σ^2 . Do 100 replications three times, record the empirical power, and take the average. Compare the histogram of test statistics with its hypothesized F distribution. Do the same experiment when both populations are skewed right. Describe the effects of skewness on the histogram and on power. Compare this result with exercise 25. How does this illustrate the importance of the normality assumption in a test for variances?

<u>Distribution</u>	<u>Experiment 1</u>	<u>Experiment 2</u>	<u>Experiment 3</u>	<u>Avg Power</u>
Normal	_____	_____	_____	_____
Skewed Right	_____	_____	_____	_____

Individual Learning Projects

Write a report on one of the three topics listed below. Use the cut-and-paste facilities of the module to place the appropriate graphs in your report. Include in your report a copy of graphs and/or tables you feel are relevant for each different experimental setup.

1. Investigate the effects of the variance assumption and unbalanced sample size on Type I error in a test for two means. Using the Do-It-Yourself controls, set $\mu_1 = 5.0$ and $\mu_2 = 5.0$ (equal means), set $\sigma_1 = 3.0$ and $\sigma_2 = 1.5$ (unequal variances), make both populations normal, and let variances be assumed unknown. Specify a two-tailed test with $\alpha = 0.05$ and assume *equal* variances. Start with balanced samples $n_1 = 20$ and $n_2 = 20$ (combined sample size of 40). Do 100 replications several times, find the average Type I error of the test (check the Analysis of Experiment window), and compare the histogram of test statistics with the hypothesized t distribution. Repeat the replication for unbalanced samples $n_1 = 30$ and $n_2 = 10$ (combined sample size of 40) and again for unbalanced samples with $n_1 = 10$ and $n_2 = 30$ (combined sample size of 40). Then repeat the entire series of replication experiments with variances assumed *unequal*. Describe the effect on Type I error of the correctness of the assumption about variances with balanced or unbalanced sample sizes.
2. Construct two power curves for a two-tailed test of two means. With the Do-It-Yourself controls, set $\mu_1 = 5.0$ and $\mu_2 = 3.5$ (unequal means), $\sigma_1 = 1.6$ and $\sigma_2 = 1.6$ (equal variances), make both populations normal, and let variances be assumed *unknown* and equal. Choose $\alpha = 0.05$. Choose a small sample size (under 10) and make both samples the same size. Record the true difference of means, do 100 replications three times, and find the average power of the test. Repeat the experiment as you decrease μ_1 in steps of 0.5 until it equals μ_2 (true null hypothesis). Make a power curve by plotting the average power (on the Y-axis) against the true difference in means (on the X-axis). Then make a power curve for a sample size that is four times as large. Explain what the power curves tell you.
3. Investigate the effects of sample size on power in a test of two variances. Create a moderately false hypothesis using the Do-It-Yourself controls to set $\sigma_1 = 5.1$ and $\sigma_2 = 3.3$ (the means are irrelevant). Be sure the populations are normal. Choose $\alpha = 0.05$ and a two-tailed test. Start with small but equal sample sizes (under 6). Use 100 replications several times and average the power of the test (from the Analysis of Experiment window). Progressively double the sample sizes (n , $2n$, $4n$, $8n$, $16n$) and repeat the experiment. Describe the effect on power of progressively doubling the sample sizes. How large a sample is needed so the power of the test is at least 50%? At least 90%?

Team Learning Projects

Select one of the three projects listed below. In each case, produce a team project that is suitable for an oral presentation. Use presentation software or large poster boards to display your results. Graphs should be large enough for your audience to see. Each team member should be responsible for producing some of the graphs. Include in your report a copy of all graphs and statistics that you evaluated. Ask your instructor if a written report is also expected.

1. This is a project for a team of two. Investigate unbalanced samples on power in a test of two variances, and the possible mitigating effects of sample size. Create a moderately false hypothesis about variances using the Do-It-Yourself controls, setting $\sigma_1 = 6.0$ and $\sigma_2 = 4.0$ (the means are irrelevant). Be sure the populations are normal. Choose a two-tailed test for σ^2 with $\alpha = 0.05$. The first team member should do 100 replications several times and average the power of the test (from the Analysis of Experiment window) for sample sizes $n_1 = 6, 12, 24, 36, 42$ and $n_2 = 48 - n_1$ (so the combined sample size always remains at 48). The second team member will do a similar experiment for sample sizes $n_1 = 12, 24, 48, 72, 84$ and $n_2 = 96 - n_1$ (so the combined sample size always remains at 96). Are unbalanced sample sizes injurious to power? To what extent do larger sample sizes enhance power and/or mitigate any adverse effects that may exist?
2. This is a project for a team of three. Investigate the effects of population non-normality and sample size on Type I error. With the Do-It-Yourself controls set $\mu_1 = 6.0$ and $\mu_2 = 6.0$ (equal means) and $\sigma_1 = 2.0$ and $\sigma_2 = 2.0$ (equal variances). Be sure the variances are assumed *unknown*. Each team member should choose a different population shape (normal, uniform, skewed) and should use sample sizes of 6, 24, and 96, first in a test for equal means (a true hypothesis) assuming equal variances, and then in a test for equal variances (also a true hypothesis). Use 100 replications three times, find the average Type I error of the test for each sample size, and examine the histogram of test statistics. Summarize your findings about Type I error. Does larger sample size mitigate potential adverse effects of non-normality for means and variances?
3. This is a project for a team of three or more. Construct power curves for a two-tailed test of two means. With the Do-It-Yourself controls, set $\mu_1 = 6.0$ and $\mu_2 = 5.0$ (unequal means), set $\sigma_1 = 1.5$ and $\sigma_2 = 1.5$ (equal variances), make both populations normal, and let variances be assumed *unknown*. Assume equal variances and choose $\alpha = 0.05$. Each team member should choose a different sample size so that the range from 5 to 100 is covered (keep both samples the same size). Record the true difference of means, do 100 replications three times, and find the average power of the test. Repeat the experiment as you increase μ_2 in steps of 0.2 until μ_2 equals μ_1 (true null hypothesis). Present the combined team results on one graph with a power curve for each sample size. For each power curve, plot the average power (on the Y-axis) against the true mean (on the X-axis). Discuss what the power curves tell you.

Self-Evaluation Quiz

1. Type II error is
 - a. the probability of correctly rejecting the null hypothesis.
 - b. the probability of correctly accepting the null hypothesis.
 - c. the probability of incorrectly rejecting the null hypothesis.
 - d. the probability of incorrectly accepting the null hypothesis.
 - e. none of the above.
2. To compare two sample means with known variances, the statistician would use the
 - a. Student's t distribution.
 - b. chi-square distribution.
 - c. normal distribution.
 - d. F distribution.
 - e. binomial distribution.
3. Which procedure might be used to compare two sample means?
 - a. Draw dot plots for the samples and compare them visually.
 - b. Calculate a test statistic and compare it with a critical value.
 - c. Calculate a confidence interval and see if it contains zero.
 - d. Calculate a p-value for a test statistic and compare it with an α level.
 - e. All of the above are useful procedures.
4. To compare two sample means with unknown variances, the statistician would use the
 - a. Student's t distribution.
 - b. chi-square distribution.
 - c. normal distribution.
 - d. F distribution.
 - e. binomial distribution.
5. The one-sided confidence interval for a difference of means
 - a. always extends to infinity on one side.
 - b. is similar to a one-sided hypothesis test.
 - c. would signal rejection of the hypothesis of equal means if it does not include zero.
 - d. puts the entire α risk in one tail.
 - e. has all of the above characteristics.
6. In a test of two means with samples of size n_1 and n_2 , if the population variances are unknown and assumed unequal, which statement is *incorrect*?
 - a. We face a problem known as the Behrens-Fisher problem.
 - b. The sample variances are biased estimators.
 - c. Degrees of freedom will generally be less than $n_1 + n_2 - 2$
 - d. Degrees of freedom will generally be greater than the minimum of $n_1 - 1$ or $n_2 - 1$
 - e. It is appropriate to use Welch's correction for degrees of freedom.

7. In a two-sample test from unknown populations with sample sizes $n_A = 10$ and $n_B = 15$, which statement is correct?
 - a. To compare variances we would use F with $df_1 = 10$ and $df_2 = 15$.
 - b. To compare means we would use Student's t with $df = 23$, assuming unequal variances.
 - c. To compare means we would use Student's t with $df = 23$, assuming equal variances.
 - d. To deal with unequal variances, we might rely on Chebychev's inequality.
 - e. To find the confidence interval for the difference of means, we would use $F_{9,14}$.
8. With unknown but equal variances, the test statistic for equality of two means
 - a. will increase when the level of significance increases.
 - b. will decrease when the level of significance increases.
 - c. could either increase or decrease when the level of significance increases.
 - d. is unchanged when the level of significance increases.
 - e. is affected both by sample size *and* the level of significance.
9. To compare two sample variances, the statistician would use which distribution?
 - a. Normal.
 - b. Student's t.
 - c. Chi-square.
 - d. F.
 - e. None of the above.
10. The $1 - \alpha$ confidence interval for the ratio of two unknown variances
 - a. will enclose unity if the true population variances are equal.
 - b. will enclose unity with probability $1 - \alpha$ if the true population variances are equal.
 - c. will enclose unity if the hypothesis of equal variances is rejected at α .
 - d. will be reliable even if the populations are skewed (i.e., non-normal).
 - e. is based on the Student's t distribution if the populations are normal.
11. If either sample size is increased, the critical value for a right-tail test for equality of two variances would be likely to
 - a. increase.
 - b. decrease.
 - c. stay the same.
 - d. either increase or decrease.
 - e. depend on the test statistic.
12. Type I error in a two-sample test for variances is quite sensitive to population skewness, while the Type I error in a two-sample test for means is fairly robust to population skewness.
 - a. True.
 - b. False.

Glossary of Terms

Acceptance region Portion of the hypothesized distribution that is bounded by the critical value(s). Its area is $1 - \alpha$ where α is the chosen level of significance. Since a given sample can disprove (but cannot prove) the null hypothesis, it is more accurate to call it the *non-rejection region*. See **Type I error**.

Alternative hypothesis Denoted H_1 , it is the converse of the null hypothesis (e.g., $H_0: \mu_1 = \mu_2$ and $H_1: \mu_1 \neq \mu_2$). If the sample evidence contradicts H_0 , we reject H_0 in favor of the alternative hypothesis H_1 . A hypothesis test may be motivated by suspicion that H_0 is false.

Behrens-Fisher problem When population variances are unknown and assumed unequal, the distribution of the test statistic for a difference of two means is uncertain. Various approaches to this problem are used, but the simplest is to adjust the degrees of freedom. See **Welch's correction**.

Chi-square distribution If the population is normal, the sample variance has a chi-square distribution with $n - 1$ degrees of freedom.

Confidence interval Range of values that would enclose a true (generally unknown) population parameter (such as a difference of means or a ratio of variances) a given percentage of the time.

Difference of means with known variances:

$$\bar{X}_1 - \bar{X}_2 - z \sigma_{\bar{X}_1 - \bar{X}_2} < \mu_1 - \mu_2 < \bar{X}_1 - \bar{X}_2 + z \sigma_{\bar{X}_1 - \bar{X}_2}$$

Difference of means with unknown variances:

$$\bar{X}_1 - \bar{X}_2 - t s_{\bar{X}_1 - \bar{X}_2} < \mu_1 - \mu_2 < \bar{X}_1 - \bar{X}_2 + t s_{\bar{X}_1 - \bar{X}_2}$$

Ratio of two population variances:

$$F_{\text{Lower}}\left(\frac{s_1^2}{s_2^2}\right) < \frac{\sigma_1^2}{\sigma_2^2} < F_{\text{Upper}}\left(\frac{s_1^2}{s_2^2}\right)$$

Confidence level Desired probability of enclosing an unknown population parameter when creating a confidence interval from sample data. The confidence level, denoted $1 - \alpha$, is chosen by the researcher and is usually expressed as a percent (typically 90%, 95%, or 99%). The higher the confidence level, the wider the confidence interval. See **Level of significance**.

Critical value Value on the X-axis that defines the rejection region for a hypothesis. In a one-tail test, the critical value defines a right-tail or left-tail area. In a two-tail test, there are critical values defining rejection regions in each tail. The critical value is determined by the level of significance. A test statistic beyond the critical value(s) is unlikely if the H_0 is true.

Decision rule Diagram that illustrates the criterion for rejection of the null hypothesis. It shows the hypothesized sampling distribution with its critical value(s) labeled and a shaded rejection region for the specified level of significance.

Degrees of freedom For a Student's t-test for the difference of two means with unknown population variances that are assumed equal, degrees of freedom will be $n_1 + n_2 - 2$. If the unknown variances are assumed unequal, degrees of freedom may be modified using Welch's correction. For an F test of the ratio of two variances, the numerator and denominator degrees of freedom will be $df_1 = n_1 - 1$ and $df_2 = n_2 - 1$.

Equal variances When the population variances are unknown, the formula for the standard error of the difference of two means depends on whether the variances are assumed equal or unequal. The hypothesis of equal variances may be tested using an F test based on the sample variances.

F distribution If the null hypothesis of equal variances is true and the populations being sampled are normal, the sampling distribution of the ratio of two sample variances s_1^2 / s_2^2 is F. See **Degrees of freedom**.

Known variances If the variances in a test for two sample means are known, the normal distribution may be used for the sampling distribution of the test statistic. If the variances are unknown, the test statistic follows the Student's t distribution. If the sample sizes are large, the difference between the normal and Student's t distributions is slight.

Level of significance The desired probability of Type I error. It is set by the researcher (typical values are 0.10, 0.05, and 0.01) and is denoted α . Other things equal, the power of a hypothesis test increases as the level of significance increases. See **Confidence level**.

Normality assumption In tests of two means or two variances, the theoretical sampling distributions are predicated upon the assumption that the populations being sampled are normal. If the populations are not normal, the Type I error and power of the test for two variances may be compromised. However, because of the Central Limit Theorem, the test for two means is fairly insensitive to violations of the normality assumption. See **Robustness**.

Null hypothesis Denoted H_0 , it is a statement that we try to reject (for example, $H_0: \sigma_1^2 = \sigma_2^2$). The null hypothesis is not necessarily chosen because we believe it to be true, but rather as a reference point. If the sample evidence contradicts H_0 , the null hypothesis is rejected. Otherwise, it awaits further testing and could be rejected at a later time.

One-tail test In a one-tail test, the alternative hypothesis always contains $>$ (for a right-tail test) or $<$ (for a left-tail test). For example, the hypotheses $H_0: \mu_1 \geq \mu_2$ and $H_1: \mu_1 < \mu_2$ would require a left-tail test.

Power If the null hypothesis is false, *theoretical* power is the probability of rejecting the null hypothesis H_0 when it is false (or $1 - \beta$ where β is the probability of Type II error). For example, in a test comparing two means, power will be lower if the true means are nearly equal, but will be higher if the true means differ substantially. The ideal power is near 1. *Empirical* power is the ratio of the number of rejections to the number of times the test is performed.

P-value Probability that a result as extreme as (or more extreme than) the observed sample statistic would arise by chance if the null hypothesis were true.

Rejection region Area under the hypothesized sampling distribution that lies beyond the critical value(s). It is determined by α , the probability of Type I error. If the test statistic falls within this region, we will reject the null hypothesis. See **Level of significance**.

Robustness Quality of being unaffected by a violation of an assumption. For example, the Student's t test for two means is robust to non-normality in the population if the sample sizes are moderately large.

Sampling distribution Theoretical distribution of the test statistic assuming the null hypothesis is true. For example, if $H_0: \sigma_1^2 = \sigma_2^2$ is true, the sampling distribution of s_1^2 / s_2^2 is F.

Standard error Another name for the standard deviation of a sampling distribution. The formula for standard error is derived from statistical theory, but the standard error is usually estimated from samples.

Difference of two means with known variance:

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Difference of two means with unknown variances assumed equal:

$$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_1 - 1)s_1^2}{n_1 + n_2 - 2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Difference of two means with unknown variances assumed unequal:

$$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Student's t distribution If the populations being sampled are normal but with unknown variances, the test statistic for the difference of means follows the Student's t distribution. This distribution is also used to construct confidence intervals for one mean or for the difference of two means. See **Welch's correction**.

Test statistic Calculated value resulting from the comparison of a sample statistic and a hypothesized population parameter. Its distribution depends on the null hypothesis. The test statistic is compared with a critical value to see whether the null hypothesis should be rejected.

Difference of two means:

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{\bar{X}_1 - \bar{X}_2}} \text{ (known variances)} \quad \text{or} \quad t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{X}_1 - \bar{X}_2}} \text{ (unknown variances)}$$

Ratio of two variances:

$$F = \frac{s_1^2}{s_2^2}$$

Two-sample test Any statistical test that utilizes two samples to test a hypothesis about two populations. For example, two sample means may be used to test whether the population means are equal. In contrast, a one-sample test compares a sample with a fixed point of reference, such as a hypothesized value of a true population parameter, to test a hypothesis about the population.

Two-tail test In a two-tail test, the alternative hypothesis always contains \neq . For example, the hypotheses $H_0: \mu_1 = \mu_2$ and $H_1: \mu_1 \neq \mu_2$ imply a two-tail test.

Type I error The error of rejecting a true null hypothesis that is true. The probability of Type I error is denoted α . It is the area under the hypothesized sampling distribution that is beyond the critical value(s). See **Level of significance**.

Type II error The error of accepting a null hypothesis that is false. The probability of Type II error (denoted β) is the area under the true sampling distribution that is not in the rejection region. See **Power**.

Welch's correction When population variances are unknown and assumed unequal, the t test for a difference of two means may be used with modified degrees of freedom. The degrees of freedom will be between $\text{Min}(n_1, n_2)$ and $n_1 + n_2 - 2$. See **Behrens-Fisher problem**.

Welch's correction:

$$\text{d.f.} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{\left(s_1^2/n_1 \right)^2}{n_1 - 1} + \frac{\left(s_2^2/n_2 \right)^2}{n_2 - 1}}$$

Solutions to Self-Evaluation Quiz

1. d Read the Overview of Concepts. Consult the Glossary.
2. c Do Exercises 1–7. Read the Overview of Concepts. Consult the Glossary.
3. e Do Exercises 1–5. Read the Illustration of Concepts.
4. a Do Exercises 11–18. Read the Overview of Concepts. Consult the Glossary.
5. e Consult the Glossary. Read Overview of Concepts.
6. b Do Exercises 12 and 13. Consult the Glossary.
7. c Do Exercises 12, 13, and 20. Consult the Glossary.
8. d Read the Overview of Concepts. Consult the Glossary.
9. d Do Exercises 19–23. Read the Illustration of Concepts. Consult the Glossary.
10. b Do Exercises 19–23. Read the Illustration of Concepts.
11. b Consult the Glossary. Do Team Learning Project 1.
12. a Do Exercises 25–28. Do Team Learning Project 2.