Distributions of Differences Between Sample Means

INTRODUCTION

In this lesson, we will introduce methods for comparing means from independent samples.

Independent samples do not influence each other in any way. For example, we could compare the mean height of women to the mean height of men. The samples would be independent in this case, because a randomly selected women's height does not affect the height of a randomly selected man.

Suppose you want to estimate the difference between the mean commute time of students at your college and the mean commute time of faculty at your college. To carry out this study, you gather two random samples. One sample is from the students and one sample is from the faculty. Your samples give you the following sample statistics (commute times are in minutes):

Table 1: Student and Faculty Commute Times

	Sample 1: Students	**Sample 2: Faculty**
Sample Mean: $\bar{x}$	37.5 min.	45.2 min.
Sample Standard Deviation: s	10.60 min.	8.60 min.
Sample Size: n	45	35

1 Are the samples independent? If yes, what makes the samples independent? If not, why not?

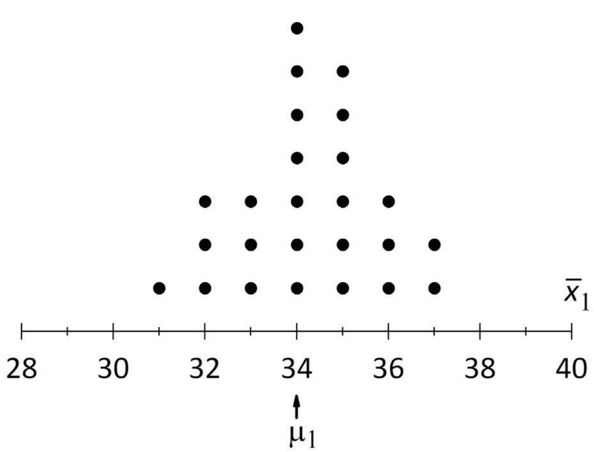
2 Look at the sample table. Which sample has the longer average commute?

3 Now calculate the difference in the sample means ($\bar{x}_1 - \bar{x}_2$). Let $\bar{x}_1$ be the sample mean commute time for the students (Sample 1) and $\bar{x}_2$ be the sample mean commute time for the faculty (Sample 2). The value $\bar{x}_1 - \bar{x}_2$ is called a *difference of sample means*. Explain how the sign of the difference (positive or negative) tells us which sample mean is greater.

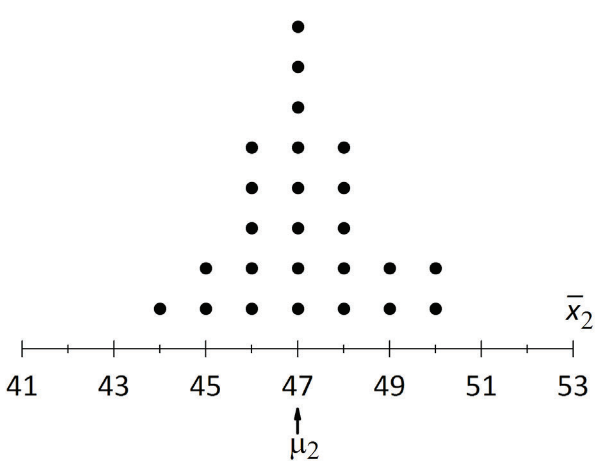
The mean commute times above are sample means. The population means will be different. If we want to know whether the difference ($\bar{x}_1 - \bar{x}_2$) above is unusual, then we need to use the sampling distribution of all possible differences between sample means.

After extensive research, the college determines that the population mean commute time for students is μ_1 = 34 minutes with a standard deviation of σ_1 = 9 minutes, and that the population mean for faculty is μ_2 = 47 minutes with a standard deviation of σ_2 = 8 minutes.

With these parameters, we can approximate sampling distributions of sample means for student and faculty commute times. To such simulations are represented below. For the students, 25 samples were randomly generated, each with 45 students. Sample means were computed from each sample. For the faculty 25 samples were generated as well each with a 35 faculty, and sampling means were computed. The sampling distributions are represented below first for students, then for faculty.

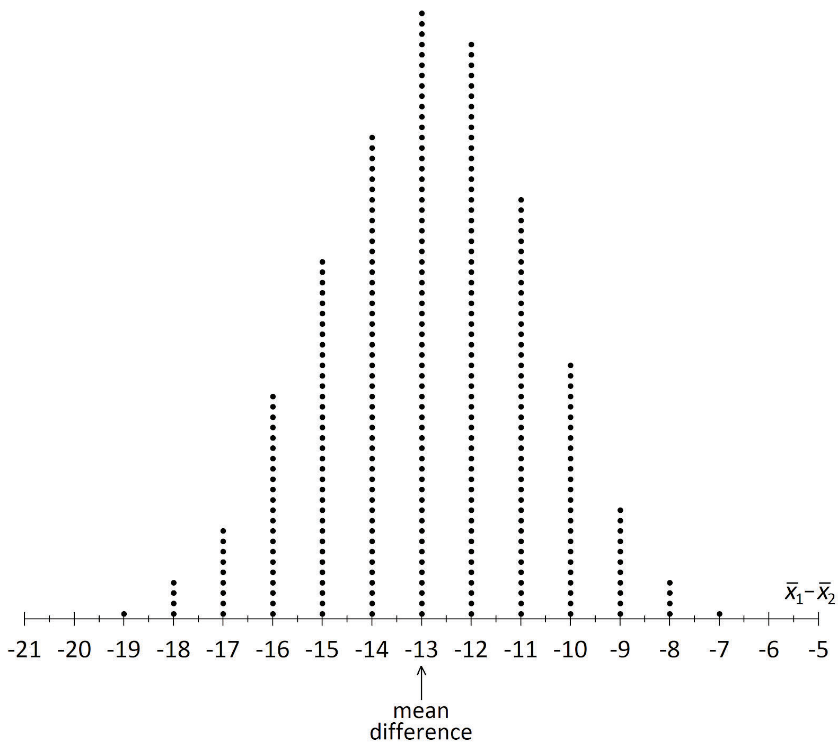


Twenty-five randomly generated sample means of student commute times, from samples of size 45. The simulations assume that the population mean is 34.



Twenty-five randomly generated sample means of faculty commute times, from samples of size 35. The simulations assume that the population mean is 47.

If we subtract each of the 25 faculty means from each of the 25 student means, we get 25·25 = 625 differences between sample means, $\bar{x}_1 - \bar{x}_2$. These are represented in the dotplot below (each dot represents *two* data points).



4 We have learned that the difference between two approximately normal variables will also be approximately normal.

A) Do both of the individual sampling distributions (of $\bar{x}_1$ and $\bar{x}_2$) above satisfy the criteria for approximate normality? Explain.

B) Does the distribution of differences appear approximately normal? Is this unexpected?

5 We have learned that the mean of all differences between sample means is the difference of the population means. Does that hold true for this example?

A Mean of differences between sample means (from the image above) = ________

B Difference between population means, $\mu_1 - \mu_2$ = _______

C Is the mean difference equal the difference of the means?

6 Compute the standard errors for the mean commute times for faculty and students.

Students: $\frac{\sigma_1}{\sqrt{n_1}} =$ Faculty: $\frac{\sigma_2}{\sqrt{n_2}} =$

7 We've learned that the variance of all differences between two variables is *sum* of their variances. Remember, variance is the *square* of standard deviation (or in this case, standard error).

A What is the variance of all differences ($\bar{x}_1 - \bar{x}_2$)?

B Standard deviation is the square root of variance. What is the standard deviation (i.e. standard error) of all differences, ($\bar{x}_1 - \bar{x}_2$)?

The problems you just finished illustrate properties that we have already established. The properties are:

- When two independent random variables are approximately normal in their distribution then differences between those variables will be approximately normal as well.
- The mean of differences between two random variables is the difference their means respectively.
- The variance of differences between two independent random variables is the *sum* of their variances, respectively. The standard deviation (i.e. standard error) is the square root of that variance.

The Central Limit Theorem for Differences Between Sample Means

- When two sampling distributions of sample means satisfy the conditions for approximate normality (where $n_1 > 30$ and $n_2 > 30$, or the respective populations are normal) then the distribution of all differences between sample means, $\bar{x}_1 - \bar{x}_2$, will be approximately normal as well.

- The mean of differences between sample means, $\bar{x}_1 - \bar{x}_2$, is the difference between the respective population means, μ_1 - μ_2.
- Recall that the standard error of sample means is $\frac{\sigma}{\sqrt{n}}$, and the variance is the square of this, $\frac{\sigma^2}{n}$. The variance of differences between *two* sample means is the sum of the respective variances, $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$. The standard deviation (i.e. standard error) is the square root of the variance, $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.

To standardize the difference between two sample means, we could take a *Z* score by subtracting the mean difference and then dividing by the standard error.

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

In practice, this statistic is rarely used because population standard deviations are typically unknown. Sample standard deviations are used instead, but with them comes added variability resulting in a statistic that is better approximated by a *T* distribution instead of *Z*.

Estimated standard error = $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

T statistic: $T = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$.

It doesn't take much imagination to see that, with two sample sizes, the degrees of freedom will not be straightforward for this statistic. Researchers have found that the statistic above varies, approximately, according to a *T* distribution if the following degrees of freedom are used:

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

This is a tedious formula, so we will be providing the degrees of freedom on each problem where they are needed.

Table 1: Student and Faculty Commute Times

	Sample 1: Students	**Sample 2: Faculty**
Sample Mean: $\bar{x}$	37.5 min.	45.2 min.
Sample Standard Deviation: s	10.60 min.	8.60 min.
Sample Size: n	45	35

8 Compute the *estimated* standard error (using the sample standard deviations this time) of differences between sample means using the information from Table 1 above.

9 Using the sample mean difference, $\bar{x}_1 - \bar{x}_2$, the population mean difference, μ_1 - μ_2, and the estimated standard error (you have already computed each of these), compute the *T* statistic.

$$T = \frac{(\bar{x}_1 - \bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} =$$

10 The *T* distribution for the statistic above has 77 degrees of freedom. Using this, find a *P*-value and state whether the difference between sample means is unusually high.

11 Use the *P*- value to make a conclusion in context about the relationship between the sample means.

CONCLUSION

In this lesson, we learned about the sampling distribution of differences between sample means. We learned that distribution is approximately normal under certain conditions. We learned that the mean of this distribution is the difference between the respective population means. We also learned a standard error of differences between sample means.

This knowledge will allow us to test claims about differences between population means, and construct confidence intervals about differences between population means.

Testing for Differences Between Population Means

INTRODUCTION

In this Section we use what we know about distributions of differences between sample means, $\bar{x}_1 - \bar{x}_2$, to test claims that compare two population means, μ_1 and μ_2.

Imagine that a community college math department uses the same final exam for all college algebra students. The department chair is concerned that students in online courses do not do as well as students in on-site courses (that is, courses in which the students are in the classroom). The fundamental or basic research question here is, *do students who take online courses perform poorly when compared to students in on-site courses*?

The instructor gathers random samples of 40 on-site students and 35 online students. She finds the final exam score for each student in the random sample. The mean score for the on-site students is 75, with a standard deviation of 12. The mean score for the online students is 72, with a standard deviation of 15.

We will conduct a test at the 5% significance level to determine if the mean score for all on-site students is higher than the mean score for all online students.

Hypothesis Tests for the Difference of Population Means

To answer this research question, let's see how the hypothesis testing process works in the context of differences between means, using independent samples.

Step 1: Determine the Hypotheses

The null hypothesis states that the *population means* are equal. This is the same as saying that the difference between the two population means is $\mu_1 - \mu_2 = 0$.

$H_0: \mu_1 = \mu_2$.

The alternative hypothesis will be one of the following inequalities:

- H_a: $\mu_1 < \mu_2$, for a left-tailed test.
- H_a: $\mu_1 > \mu_2$, for a right-tailed test.
- H_a: $\mu_1 \neq \mu_2$, for a two-tailed test.

TRY THESE

Step 1: Determine the Hypotheses

1 The on-site students are Population 1 and the online students are Population 2. What are the null and alternative hypotheses?

2 Is this a left-tailed, right-tailed, or two-tailed test?

Step 2: Collect the Data

We must verify that the criteria for the approximate normality of the sampling distributions are met.
The criteria are:

- Each population must be normal or the sample sizes must be greater than 30.
- The samples must be independent. (The values in the first sample cannot have any influence on the outcome of the values in the second.)
- The samples must be randomly selected from the populations.

Once the samples have been collected, compute

- the sample mean of each sample,
- the standard deviation of each sample, and
- the difference in sample means.

TRY THESE

3 We saw earlier in the Section that the sampling distribution of differences in sample means is approximately normal when the individual sampling distributions of sample means are approximately normal.

A Does each sampling distribution meet the approximate normality criteria? Explain.

B Are the samples independent? How do you know?

4 Calculate the difference of the sample means, $\bar{x}_1 - \bar{x}_2$.

5 Look at your alternative hypothesis. Are the sample means consistent with your alternative hypothesis?

Step 3: Assess the Evidence

If the null hypothesis is true, the sampling distribution of differences between sample means will have the following parameters:

- The mean of all differences between sample means is the difference between the population means. The null hypothesis says these means are equal, so the difference will be zero.

Mean difference: $\mu_1 - \mu_2 = 0$ (since H_0 assumes that $\mu_1 = \mu_2$)

- The variance of all differences between sample means is the sum of their individual variances. Standard deviation (and standard error) is the square root of variance. If we use sample standard deviations in the computation, it becomes an *estimated* standard error.

Estimated standard error: $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

Standardizing the difference between sample means involves subtracting the mean difference and dividing by the estimated standard error. Using sample standard deviations in the estimated standard error forces this statistic to be distributed, approximately, according to a *T*-distribution. Notice that $\mu_1 - \mu_2$ is replaced with zero in the simplest version of the formula below.

$$T = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{\bar{x}_1 - \bar{x}_2 - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Recall that *you will be given the degrees of freedom* for these problems because their computation is rather involved.

We will use the *T*-statistic and technology or tables to find *P*-values for this test. When the null hypothesis is true, the *P*-value is the probability of observing a random difference of sample means that is at least as extreme as the one observed.

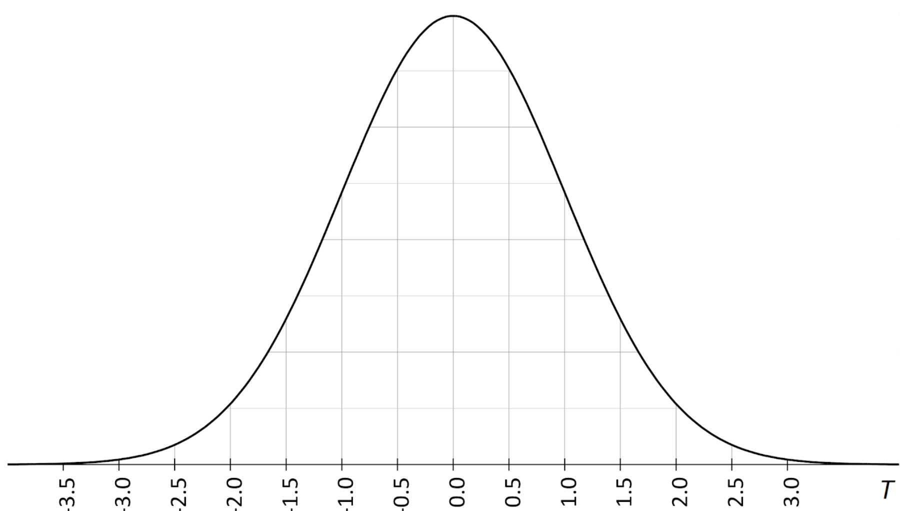
To find the *P*-value:

- For a left-tailed test, the *P*-value is the area to the left of the test statistic.
- For a right-tailed test, the *P*-value is the area to the right of the test statistic.
- For a two-tailed test, the *P*-value is *twice* the area to the right of a positive test statistic or *twice* the area to the left of a negative test statistic.

TRY THESE

6 Calculate the test statistic for your observed difference of sample means.

7 Identify the position of the observed test statistic on the *T*-distribution below. Shade the area that represents the *P*-value.



8 Find the *P*-value. In this example there are 64 degrees of freedom for the *T*-distribution.

Step 4: State a Conclusion

Finish the hypothesis test by presenting a complete conclusion with the following items:

- State whether you reject or fail to reject the null hypothesis.
- State whether you support or do not support the alternative hypothesis.
- Specify the *P*-value or level of significance.
- State a conclusion in context.

TRY THESE

9 How does the *P*-value compare to the significance level? Should we reject or fail to reject the null hypothesis?

10 What can we conclude about the alternative hypothesis?

11 State a conclusion in the context of the math department's research question.

In this Section we learned to test claims that compare two population means using independent samples. To this point, the methods used for testing hypothesis have been similar, regardless of the application. In upcoming modules, we will continue to follow the steps outlined here, but the test statistics will evolve to allow for more complex situations. These will include comparisons of three or more population means or proportions.