Section 3C – P-value and Significance Levels

Vocabulary

Population: The collection of all people or objects to be studied.

Sample: Collecting data from a small subgroup of the population.

Statistic: A number calculated from sample data in order to understand the characteristics of the data. For example, a sample mean average, a sample standard deviation, or a sample percentage.

Parameter: A number that describes the characteristics of a population like a population mean or a population percentage. Can be calculated from an unbiased census, but is often just a guess about the population.

Hypothesis Test: A procedure for testing a claim about a population.

Null Hypothesis (H_0): A statement about the population that involves equality. It is often a statement about "no change", "no relationship" or "no effect".

Sampling Variability: Also called "random chance". The principle that random samples from the same population will usually be different and give very different statistics. The random samples will usually be different than the population parameter.

P-value: The probability of getting the sample data or more extreme because of sampling variability (by random chance) if the null hypothesis is true. If the P-value is close to zero (lower than the significance level) is unlikely to have happened because of sampling variability. If the P-value is too large (higher than the significance level, then the sample could have occurred because of sampling variability.

Significance Level (α): Also called the Alpha Level. This is the probability of making a type 1 error. The P-value is compared to this number to determine significance and if sampling variability is likely to be involved in the hypothesis test.

Randomized Simulation: A technique for visualizing sampling variability in a hypothesis test. The computer assumes the null hypothesis is true, and then generates random samples. If the sample data or test statistic falls in the tail, then the sample data significantly disagrees with the null hypothesis. This technique can also calculate the P-value and standard error without a formula.

Test Statistic: A number calculated in order to determine if the sample data significantly disagrees with the null hypothesis. There are a variety of different test statistics depending on the type of data.

Critical Value: We compare a test statistic to this number to determine if the sample data significantly disagrees with the null hypothesis. If the absolute value of the test statistic is higher than the absolute value of the critical value, then the sample data significantly disagrees with the null hypothesis.

Introduction: There is a dilemma in hypothesis testing. In a hypothesis test, we want to determine if the sample data disagrees with the null hypothesis, but there is a problem. The principle of sampling variability (random chance) tells us that random samples will almost always be different than population parameters in the null hypothesis. So even if the population parameter in the null hypothesis is correct, my random sample data will still disagree with it. So how can we use random sample data to ever decide about the accuracy of a population parameter? Random samples almost always disagree with the null hypothesis.

The real question is why does the random sample data disagree? There are only two possible answers to that question and these are at the heart of the problem.

1. The random sample disagrees because the null hypothesis is wrong.

OR

2. The null hypothesis is correct and the random sample data disagrees because of sampling variability (random chance).

How do we know which option is correct in a situation? Does my random sample data disagree because all random samples disagree, or does my random sample data disagree because the null hypothesis is wrong?

To answer this question, statisticians invented the P-value.

P-value

P-value: The probability of getting the sample data or more extreme because of sampling variability (by random chance) if the null hypothesis is true.

We see from the definition, we see several important ideas addressed.

- The P-value is a conditional probability based on the null hypothesis being true. The P-value can only be calculated by assuming the null hypothesis was true.

- The P-value is a probability that our random sample data occurs. If the null hypothesis really is correct, then what is the probability of our random sample data occurring by random chance?

- The P-value helps us understand why our random sample data disagrees with the null hypothesis. Does it disagree because of sampling variability (random chance) or not? If it is not sampling variability, then the only other alternative is that the null hypothesis is wrong.

- The P-value not only takes into account the probability of the sample data occurring, but also any other samples that disagrees even more with the null hypothesis than our random sample data. This is what is meant by "or more extreme".

Reading your P-value

P-value can help us with the dilemma discussed above, but only if you know how to interpret it correctly. Remember, the P-value is the probability of your random sample data occurring because of sampling variability (by random chance). Does my random sample data disagree with the null hypothesis just because of sampling variability? If so, then the population parameter in the null hypothesis might be correct. If my sample data does not disagree because of sampling variability, then the only other alternative is that the null hypothesis must be wrong. In that case we will say that we "reject the null hypothesis".

Low P-value

Scientists like the P-value to be very close to zero. The lower the P-value, the better. Remember, the P-value is measuring the probability that the random sample data or more extreme occurred because of sampling variability. If the P-value is zero (or really close to zero), then the data probably did not occur because of sampling variability.

Think of sampling variability as a confounding variable that we need to control or at least make sure it is unlikely to be the reason the sample data disagrees. If the P-value is zero, then it is unlikely to be sampling variability (random chance).

Suppose you P-value is 0.013 (1.3%). If your car has only a 1.3% probability of starting, do you think your car will start or is it unlikely to start? If your car only has a 1.3% chance of starting, it is unlikely to start. That is a good way to think about P-value. If there is only a 1.3% probability of our random sample data disagreeing by random chance, it is probably not random chance! It is unlikely to be sampling variability.

Remember our dilemma about the two options in a hypothesis test.

1. The random sample disagrees because the null hypothesis is wrong.

OR

2. The null hypothesis is correct and the random sample data disagrees because of sampling variability (random chance).

Low P-value Key Idea: If the P-value is really close to zero, it is ruling out option 2. At least we can say that it is very unlikely to be sampling variability (option 2). In that case, the only other alternative is option 1. The random sample data disagrees with the null hypothesis because those population parameters in the null hypothesis are wrong. In that case, we can reject the null hypothesis.

P-value close to zero → Unlikely to be sampling variability → Reject H_0

High P-value

Remember, the goal is to totally rule out sampling variability as the reason our random sample data disagrees. We need the P-value to be zero or at least as close to zero as possible. It doesn't take much for a P-value to be too high. For example, suppose our P-value was 0.15 (15%). Don't events with a 15% probability sometimes happen? While 15% may be a low probability, is it really low enough to totally rule out that the event will not happen? For this reason, we need the P-value to be extremely low and extremely close to zero.

So how can we know if the P-value is too high?

The answer to this is to compare the P-value to the significance level.

In the last section, we saw that scientists pick a significance level at the beginning of the hypothesis test. We also saw that the significance level is connected to the critical value and determining if the test statistic falls in the tail. The proportion in the tail is the significance level. The significance level is also called the alpha level (α) and can be thought of as the complement of the confidence levels $(1-\alpha)$. We saw in the last section, that the most common significance level chosen is 5% $(\alpha = 0.05)$.

Confidence Levels $(1-\alpha)$	Significance Levels (α)
90% (0.90)	10% (0.10)
95% (0.95)	5% (0.05)
99% (0.99)	1% (0.01)

So the P-value must be less than or equal to the significance level, to rule out sampling variability (or at least to ensure it is very unlikely to be sampling variability). If the P-value is higher than the significance level, then the sample data could have occurred because of sampling variability.

P-value less than or equal to the significance level → Unlikely to be sampling variability → Reject H_0

P-value higher than the significance level → Could be sampling variability → Fail to reject H_0

Let's talk about these rules some. Let's look again at the P-value of 15% (0.15). If we are using a 5% significance level then the rule would indicate that the random sample data could have occurred because of sampling variability. This implies that the null hypothesis could be correct, and my sample data might disagree because all samples disagree. Does this guarantee that the null hypothesis is correct? Absolutely not. Let's go back to the car starting analogy. If my car only has a 15% probability of starting, is it guaranteed to start? No. It still has a low probability of starting, but it might start. That is the point. A high P-value does not tell us that the null hypothesis is correct for sure. It tells us that it might be correct.

So what about our dilemma? What does a high P-value tell us about our two options?

1. The random sample disagrees because the null hypothesis is wrong.

OR

2. The null hypothesis is correct and the random sample data disagrees because of sampling variability (random chance).

High P-value Key Idea: If the P-value is too high, then we cannot rule out option 2 (sampling variability). The high P-value does not guarantee it is sampling variability though. It just might be. When the P-value is large, we will not be able tell which option is correct. The null hypothesis might be wrong. Our sample data disagrees with it after all. On the other hand, the null hypothesis might be correct and our sample data disagrees because of sampling variability. In a sense, we cannot tell which option is correct. That is why we say "Fail to reject the null hypothesis". This means that we do not have a low enough P-value to rule out sampling variability, so we cannot say for sure that the null hypothesis is wrong. It might be correct.

P-value less than or equal to the significance level → Unlikely to be sampling variability → Reject H_0

P-value higher the significance level → Could be sampling variability → Fail to reject H_0

For this reason, high P-values are generally not preferred by data scientists. A low P-value rules out sampling variability (rules out random chance) and allows us to reject the null hypothesis and support the alternative hypothesis. A low P-value is also considered evidence. Scientific reports often require a low P-value as evidence to support their findings. When a scientist gets a high P-value, they do not have evidence. Sampling variability is involved and they cannot really say anything definitively. This does not mean that a high P-value has no value. A low P-value gives us evidence that the alternative hypothesis is probably correct. A high P-value indicates that the null hypothesis might be correct, but we do not have evidence.

Low P-value (Less than or equal to the significance level)

- Unlikely to be sampling variability
- Reject H_0
- H_A is probably correct
- We have significant evidence.

High P-value (Higher the significance level)

- Could be sampling variability
- Fail to reject H_0
- H_0 is probably correct
- We do not have evidence.

Example 1 (Interpreting P-values)

Suppose we have a 5% significance level and a P-value = 0.0278. Convert the P-value into a percentage and write a sentence to explain the P-value. Compare the P-value to the significance level. Is this a low P-value or a high P-value. Could this be sampling variability or is it unlikely to be sampling variability? Explain your answer. Does the sample data significantly disagree with the null hypothesis or not? Explain your answer. Should we reject the null hypothesis or fail to reject the null hypothesis? Explain your answer.

P-value = 0.0278 = 2.78%

P-value Sentence: If the null hypothesis is true, there is a 2.78% probability of getting the sample data or more extreme by random chance (because of sampling variability).

P-value (2.78%) is lower than our significance level (5%), so this is a low P-value close to zero.

Since the P-value is very low (2.78%), it is unlikely to be sampling variability (unlikely to be random chance).

A low P-value means the sample data fell in the tail and has a large test statistic. That means that the sample data does significantly disagree with the null hypothesis.

Reject the null hypothesis, since the P-value is lower than the significance level and is unlikely to be sampling variability.

Example 2 (Interpreting P-values)

Suppose we have a 10% significance level and a P-value = 0.414. Convert the P-value into a percentage and write a sentence to explain the P-value. Compare the P-value to the significance level. Is this a low P-value or a high P-value. Could this be sampling variability or is it unlikely to be sampling variability? Explain your answer. Does the sample data significantly disagree with the null hypothesis or not? Explain your answer. Should we reject the null hypothesis or fail to reject the null hypothesis? Explain your answer.

P-value = 0.414 = 41.4%

P-value Sentence: If the null hypothesis is true, there is a 41.4% probability of getting the sample data or more extreme by random chance (because of sampling variability).

P-value (41.4%) is higher than our significance level (10%), so this is a high P-value.

Since the P-value is very high (41.4%), the sample data could have occurred because of sampling variability (could be random chance).

A high P-value means the sample data did not fall in the tail and has a small test statistic. That means that the sample data does NOT significantly disagree with the null hypothesis.

Fail to reject the null hypothesis, since the P-value is higher than the significance level and could be sampling variability.

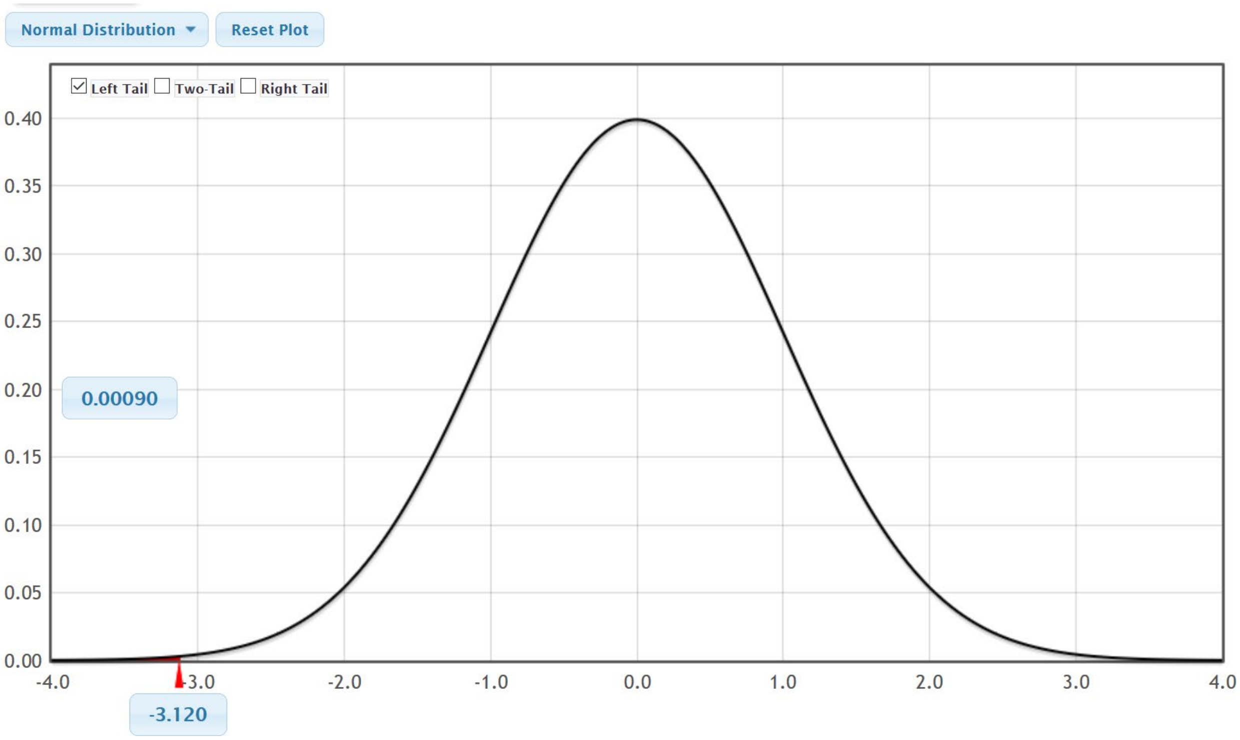
Calculating P-values

Method 1: Traditional Approach (Using the test statistic and a theoretical curve.)

One way to calculate P-value is with the test statistic and the theoretical curve that represents the sampling distribution. Remember, the P-value is the probability of getting the sample data or more extreme if the null hypothesis is true. Think of the test statistic as representing the sample data if the null hypothesis was true. "The probability of getting the sample data or more extreme" would be the proportion in the tail or tails using the test statistic as your cutoff and taking into account the type of test you are doing.

Example 1: Suppose we are doing a left tailed hypothesis test that uses the Z-test statistic for proportions. Our test statistic compares the sample data to the null hypothesis. In this example, our Z-test statistic was calculated to be Z = −3.12. What would the estimated P-value be?

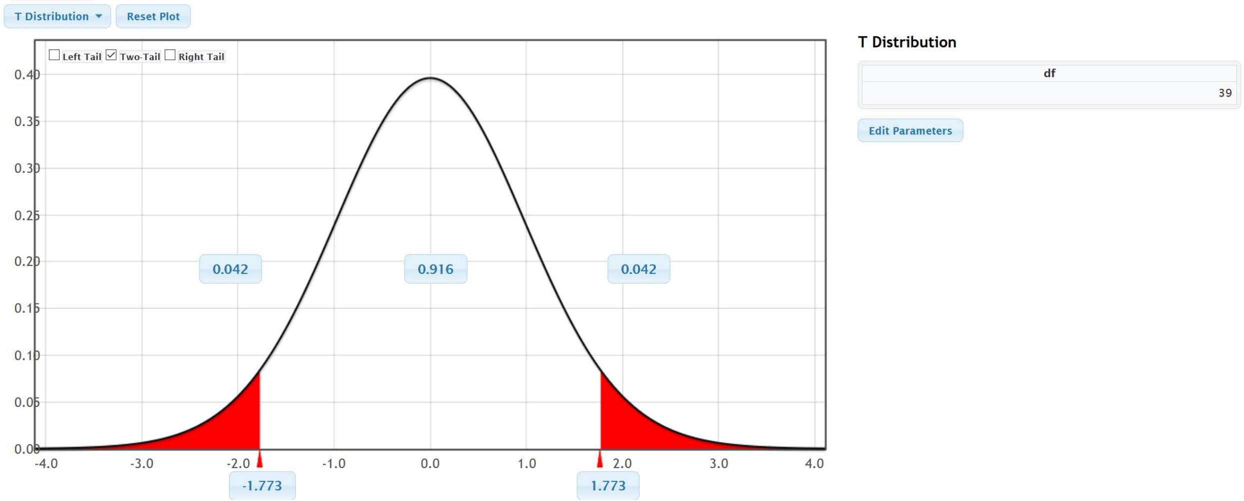
Go to the "theoretical distributions" menu in StatKey at www.lock5stat.com and click on "normal". Click on "Left Tail". In the bottom box in the left tail, put in the test statistic of −3.12. The proportion calculated above is the estimated P-value.



We see that the proportion in the left tail that corresponds to the test statistic cutoff is 0.0009 or 0.09%. This is the estimated P-value, the probability of getting the sample data or more extreme by random chance if the null hypothesis was true.

Example 2: Suppose we are doing a two-tailed hypothesis test that uses the T-test statistic. Remember, our test statistic compares the sample data to the null hypothesis. In this example, our T-test statistic was calculated to be T=1.773 and our degrees of freedom was 39. What would the estimated P-value be?

Go to the "theoretical distributions" menu in StatKey at www.lock5stat.com and click on "*t*". Under degrees of freedom put in 39 and then click on "Two Tail". A two-tailed P-value calculation takes a little thought. Notice that there are now two bottom boxes. One in the left tail and one in the right tail. If you T-test statistic is close to the left tail (negative) put in the bottom box in the left tail. If you T-test statistic is close to the right tail (positive) put in the bottom box in the right tail. Since our test statistic is closer to the right tail (positive), we will type in the test statistic of 1.773 into the right bottom box. You do not need to type the test statistic in both boxes. The left tail will automatically adjust to the number you typed in the right tail box. In a two-tailed hypothesis test, "or more extreme" could be any sample data that higher or lower than the parameter in the null hypothesis. So we need to include both of the proportions in the left and right tail. Add the two proportions calculated above the left and right tail. This is the estimated P-value.

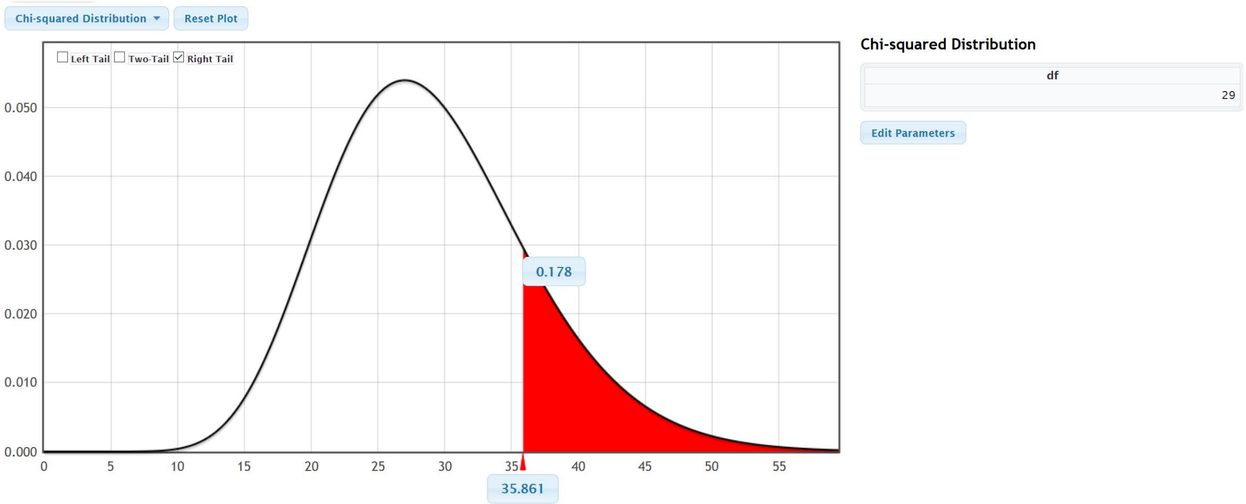


df
39

We see that the proportion in the right tail that corresponds to the T-test statistic is 0.042. We also see that the computer has calculated the left tail probability as well. Notice it is also 0.042. So the estimated P-value for this two-tailed hypothesis test is 0.042 + 0.042 = 0.084 or 8.4%. This is the estimated P-value, the probability of getting the sample data or more extreme by random chance if the null hypothesis was true. Note that a two-tailed P-value is twice as large as a one-tailed P-value from the same data. A common formula that is often used is to take the proportion in the tail corresponding to the test statistic and multiply by two $(0.042 \times 2 = 0.084)$.

Example 3: Suppose we are doing a right tailed hypothesis test that uses the chi-squared test statistic. Our test statistic compares the sample data to the null hypothesis. In this example, our chi-squared test statistic was calculated to be χ^2 = 35.861 and our degrees of freedom was 29. What would the estimated P-value be?

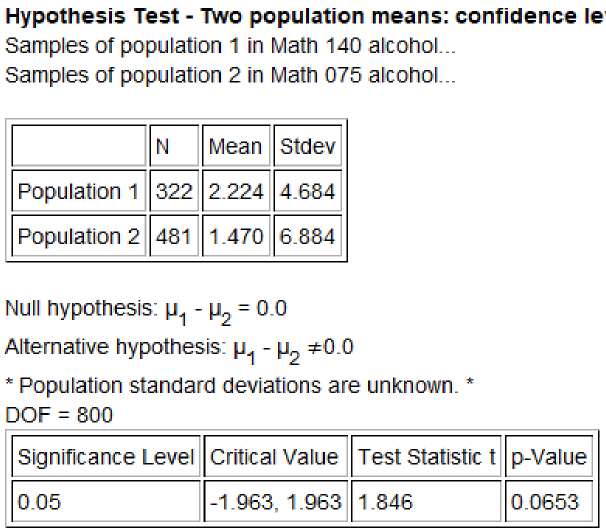
Go to the "theoretical distributions" menu in StatKey at www.lock5stat.com and click on "χ^2". Under degrees of freedom put in 29 and then click on right tail. In the bottom box, put in the test statistic of 35.861. The proportion calculated above is the estimated P-value.



df
29

We see that the proportion in the right tail that corresponds to the test statistic cutoff is 0.178 or 17.8%. This is the estimated P-value, the probability of getting the sample data or more extreme by random chance if the null hypothesis was true.

Example 4: Most traditional statistics programs calculate the P-value with this approach. In the previous section on test statistics, we compared the number of alcoholic beverages per week that Math 140 statistics students drink and the number of alcoholic beverages per week that Math 075 pre-statistics students drink. Statcato is using the test statistic, degrees of freedom, and theoretical T-curve to estimate the P-value. Notice that this is a two-tailed hypothesis test with a test statistic of T=1.846 and degrees of freedom of 800. We used these numbers with StatKey and got about the same result. In the StatKey printout, we added the tails 0.033 + 0.033 = 0.066 to get our estimated P-value.



Hypothesis Test - Two population means: confidence level = 0.95

Samples of population 1 in Math 140 alcohol...

Samples of population 2 in Math 075 alcohol...

	N	Mean	Stdev
Population 1	322	2.224	4.684
Population 2	481	1.470	6.884

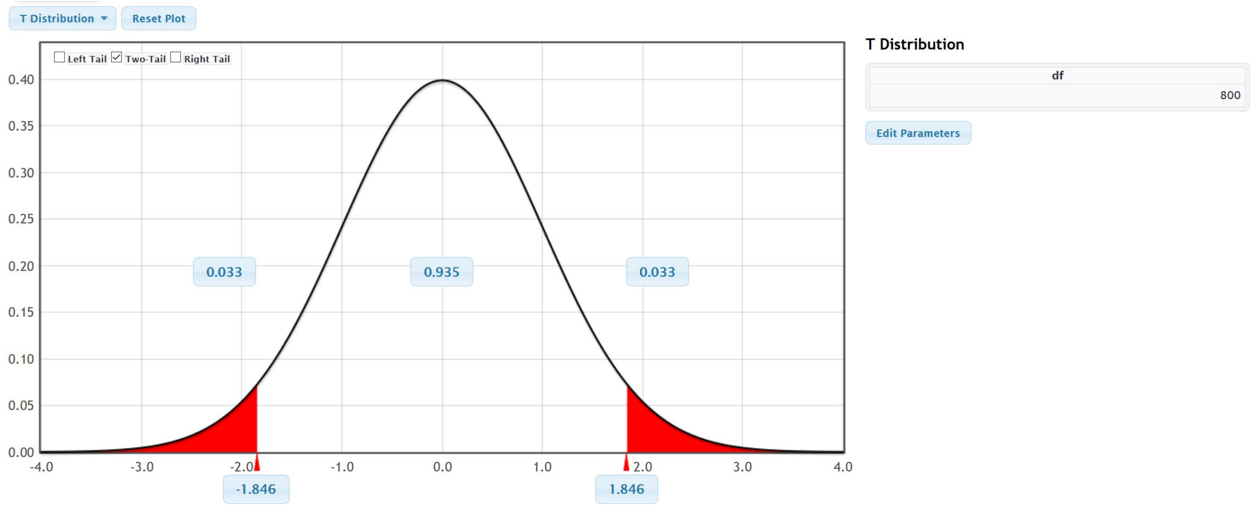
Null hypothesis: $\mu_1 - \mu_2 = 0.0$

Alternative hypothesis: $\mu_1 - \mu_2 \neq 0.0$

* Population standard deviations are unknown. *

DOF = 800

Significance Level	Critical Value	Test Statistic t	p-Value
0.05	-1.963, 1.963	1.846	0.0653



Method 2: Randomized Simulation (Randomization)

In the previous method, we saw that if we know the test statistic, we can estimate the P-value using a theoretical curve. There are many questions about the accuracy of calculating P-values in this way. For one, we need the data to meet certain assumptions to ensure that the curve is a good approximation of the sampling distribution.

Another approach that is sometimes used to calculate P-value is called "randomized simulation" or "randomization". This is a more direct way of calculating the P-value. Let's examine the P-value definition again.

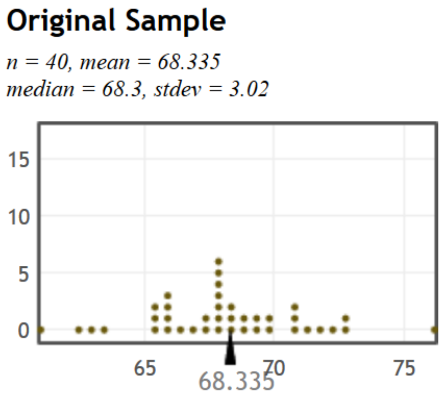
P-value: The probability of getting the sample data or more extreme because of sampling variability (by random chance) if the null hypothesis is true.

What would sampling variability look like if the null hypothesis was true? The idea behind randomized simulation is to address this question directly. Computers can create a simulated sampling distribution under the premise that the null hypothesis is true. Once this is accomplished, we can calculate the probability of getting the sample data or more extreme directly. One and two-population hypothesis tests do not require the test statistic calculation and can calculate the P-value directly from either the sample statistic or the difference between the two sample statistics. This technique is also not tied to the accuracy of a theoretical curve, so it has other advantages as well.

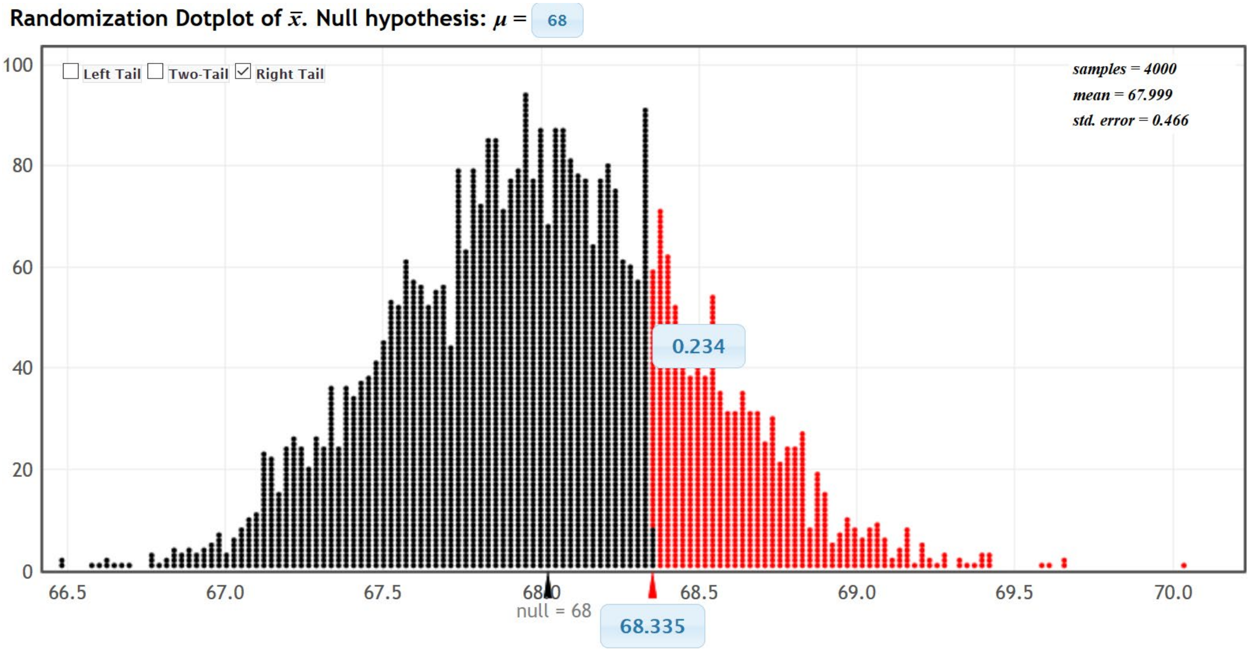
Example 1 (Simulation): Open the health data at www.matt-teachout.org and open the men's height data. Some believe that the population mean average height of all men is 68 inches. We want to test the claim that the population mean average height of men is now greater than 68 inches. Here is the null and alternative hypothesis.

H_0 : μ = 68
H_A : μ > 68 (claim)

Notice this is a right tailed test. Go to the "Randomization Hypothesis Tests" menu in StatKey at www.lock5stat.com. Click on "Test for Single Mean". Click on the "Edit Data" button and copy and paste the men's height data into StatKey. Uncheck "Identifier" and check "Header Row". An identifier is a word next to every number that explains something about that value. A header row is a title. Push "Ok". Change the null hypothesis to "μ = 68". Now click the "Generate thousand samples" button a bunch of times. We are simulating what sampling variability looks like if the null hypothesis is true. In simulation, it is important to not confuse the simulated samples with the actual original random sample data. The sample mean for the original data is 68.335, so click on "Right Tail" and then put in 68.335 in the bottom box. The proportion above the sample mean is the
P-value. Notice we have calculated the probability of getting the sample data or more extreme by sampling variability if the null hypothesis was true. We also did not require the test statistic to calculate it. This is called randomized simulation or randomization.


Original Sample

n = 40, mean = 68.335
median = 68.3, stdev = 3.02



We see that the sample mean of 68.335 inches is not in the tail. We also see that the estimated P-value is 0.234 or 23.4%. Both of these tell us that this sample data does not significantly disagree with the null hypothesis and could have occurred because of sampling variability. We will fail to reject the null hypothesis.

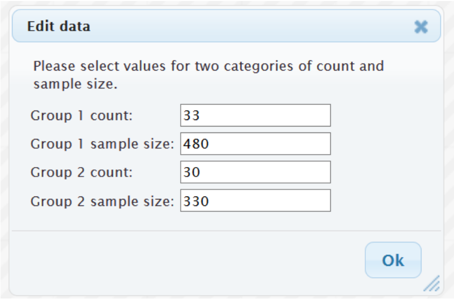
Example 2 (Simulation): Suppose we want to compare the percentage (proportion) of math 075 (pre-statistics) students that smoke cigarettes and the percentage of math 140 students that smoke cigarettes. Our claim is that the population proportions are the same. Here is the representative sample data from the Fall 2015 semester at COC and the null and alternative hypothesis.

Math 075 (pre-stat) students: 480 total students, 33 smoke cigarettes
Math 140 (statistics) students: 330 total students, 30 smoke cigarettes

π_1 : Population proportion of pre-stat students that smoke cigarettes at COC.
π_2 : Population proportion of statistics students that smoke cigarettes at COC.

H_0: $\pi_1 = \pi_2$ (claim)
H_A: $\pi_1 \neq \pi_2$

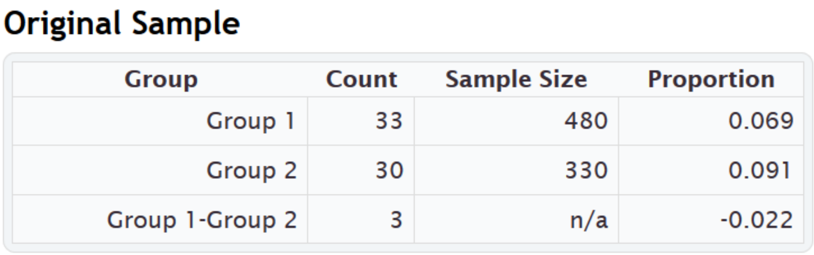
Notice this is a two-tailed two-population proportion hypothesis test. To use randomized simulation, go to the "Randomization Hypothesis Test" menu and click on "Test for Difference in Proportions". Under the "Edit Data" menu, put in the sample count and total sample size as follows. Since we designated the math 075 pre-stat students as group 1 and math 140 statistics students in group 2, we need to enter the data in that order. Now push "Ok".


Edit data

Please select values for two categories of count and sample size.

Group 1 count: 33
Group 1 sample size: 480
Group 2 count: 30
Group 2 sample size: 330

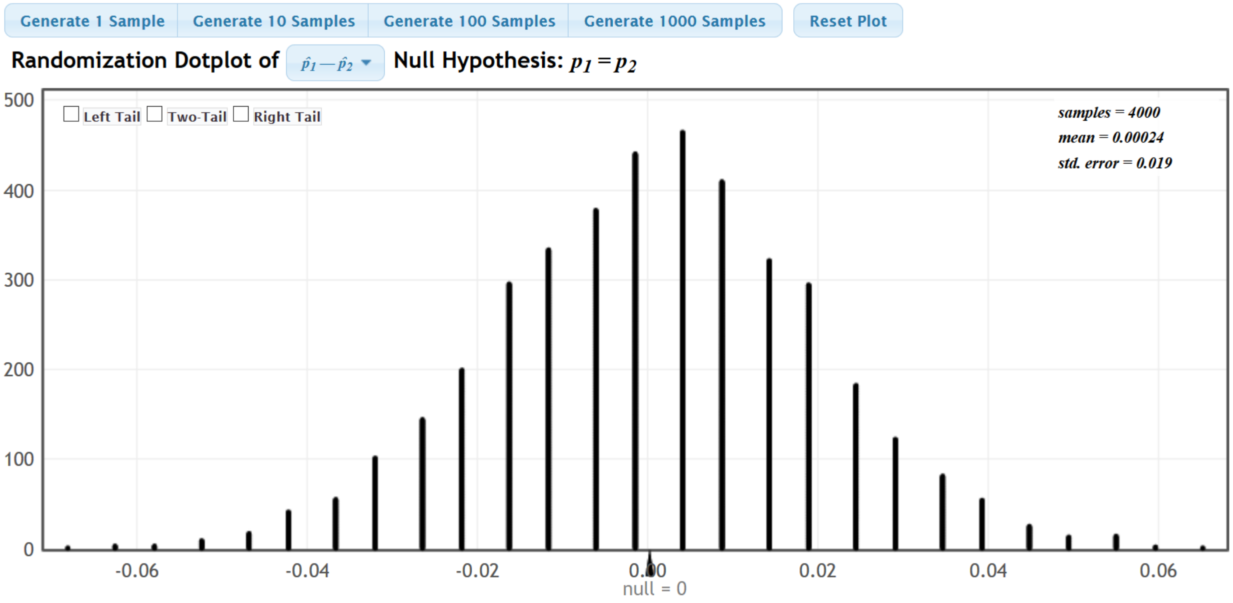
In simulation, it is important not to confuse the original real sample data with all of the simulated samples. Here is the original sample proportions. Notice that the sample proportion for the pre-stat students $(\hat{p}_1)$ was 0.069 and the sample proportion for the stat students $(\hat{p}_2)$ is 0.091. In two-population simulation we will be using the difference between the sample statistics $(\hat{p}_1 - \hat{p}_2)$ = −0.022 to calculate the P-value.



Original Sample

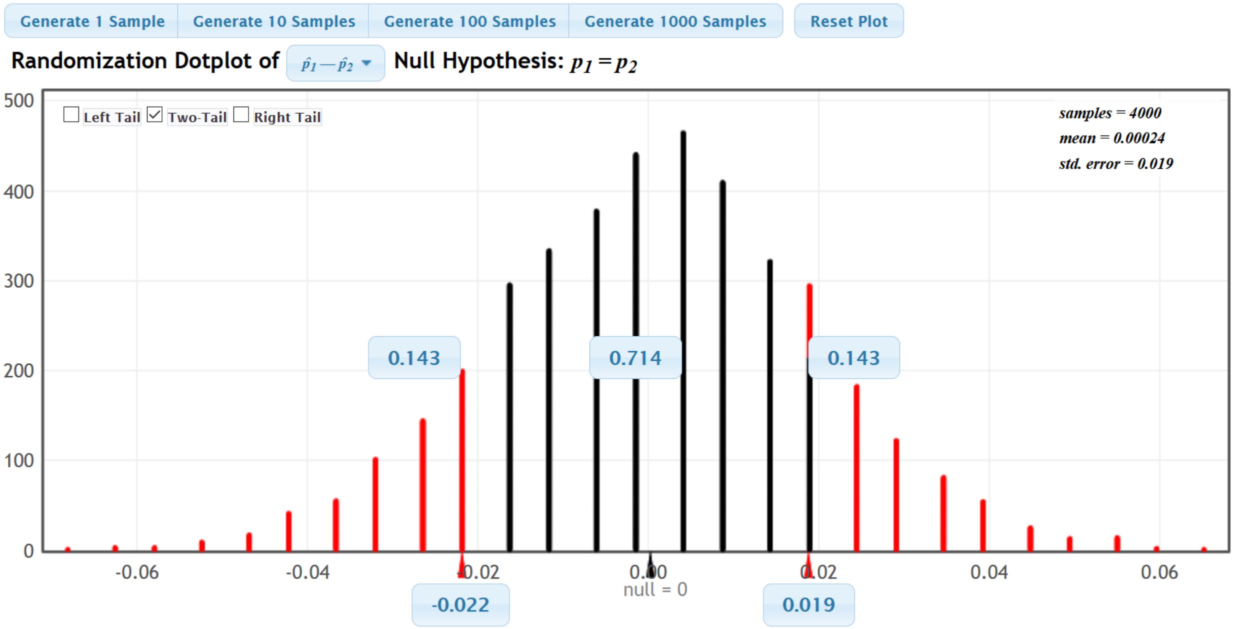
Group	Count	Sample Size	Proportion
Group 1	33	480	0.069
Group 2	30	330	0.091
Group 1-Group 2	3	n/a	-0.022

Let's simulate the null hypothesis. We are creating thousands of random samples from the premise that the populations are equal.



samples = 4000
mean = 0.00024
std. error = 0.019

Click on "Two-Tail". Since the difference between the proportions was negative and in the left tail, we will put the difference −0.022 in the left tail. The right tail will automatically adjust. Remember, in a two-tailed hypothesis test, we will need to add the proportions in the top boxes of the two tails to get the estimated P-value. Notice the estimated P-value = 0.143 + 0.143 = 0.286 or 28.6%.



Randomization Dotplot of $\hat{p}_1 - \hat{p}_2$ Null Hypothesis: $p_1 = p_2$

□ Left Tail ☑ Two-Tail □ Right Tail

samples = 4000
mean = 0.00024
std. error = 0.019