Chapter 4: Categorical and Quantitative Relationship Tests

Vocabulary

Categorical Data: Another word for qualitative data. Data that is generally in the form of labels that tell us something about the people or objects in the data set. For example, the country a person lives in, the person's occupation, type of pet, or smoking status.

Quantitative Data: Numerical measurement data. The data is made up of numbers that measure or count something and have units. Also taking an average of the data should make sense.

Random Sample: Collecting data from a population in such a way that every person in the population has an approximately equal chance of being chosen. This technique tends to give us data with less sampling bias.

Random Assignment: Take a group of people or objects and randomly put them into two or more groups. This is a technique used in experiments to create similar groups. Similar groups help to control confounding variables so that the scientist can prove cause and effect.

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

Random Chance: Another word for sampling variability. The principle that random samples from the same population will usually be different and give very different statistics.

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 and the number of populations.

Critical Value: If the test statistic is higher than this number, then the sample data significantly disagrees with the null hypothesis. The z or t score critical values are also used to calculate margin of error in confidence intervals.

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

Significance Level (α): Also called the Alpha Level. If the P-value is lower than this number, then the sample data significantly disagrees with the null hypothesis and is unlikely to have happened by random chance. This is also the probability of making a type 1 error.

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 without a formula.

Introduction: In the last chapter, we introduced the idea of a hypothesis test. This is a procedure for checking a claim about a population. People make claims all the time about populations. In the last chapter, we introduced the one-population hypothesis test to check a claim about a specific population. This last chapter will continue the discussion of hypothesis testing. A very common hypothesis test is determine if population variables may be related or not. The type of variable is very important though. We cannot analyze a categorical/categorical relationship the same way we analyze a quantitative/quantitative relationship.

There is a common thread in all of these relationship hypothesis tests that is very important to understand from the outset. If we find that a population parameter is the same in various groups (populations), then it does not seem to matter what group we are in, we get about the same thing. This would indicate that the variable that decides the groups is not related to the parameter we are studying. Alternatively, if the population parameter is significantly different in various groups (populations), then it does matter what group we are in. This would indicate that the variable that decides the groups is related to the parameter we are studying. Therefore, the null hypothesis will usually be "not related" or "independent" because we will need to show parameters are equal in various populations.

The alternative hypothesis will be "related" or "associated" because this corresponds to parameters being different or not equal. Remember equal is always the null hypothesis.

H_0: The variables are NOT related (not associated, independent) – *parameters from various populations are equal*
H_A: The variables are related (associated, dependent) – *parameters from various populations are not equal*

Note about cause and effect: Remember just because you prove two variables are related does not imply that one causes the other. In chapter 1, we learned that to prove cause and effect we need to control confounding variables with experimental design. When a scientist needs to prove cause and effect, they will often use random assignment instead of a random sample to control the confounding variables.

Section 4A – Categorical/Quantitative Relationships: Two Population Mean Hypothesis Test

Suppose we want to determine if categorical variables are related to a quantitative variable or not. A common technique would be to examine the population means from the various groups determined by the categorical variable. If the population means from the quantitative data are equal in the groups, then that would indicate that the categorical variable that determines the groups is not related to the quantitative variable. It did not matter what group we are in, since the means are about the same. If the mean averages for the groups are significantly different, then it does matter what group we are in. This would indicate that they are related. For this section, we will be focusing on categorical data with only two options. This leads to a two-population mean average hypothesis test. If the categorical data has three or more variables, then that would lead to an ANOVA test. We will cover that test in our next section.

Important note: Just because variables are related does not imply cause and effect. To prover cause and effect, we need to use experimental design.

Null and Alternative Hypotheses

Here are common null and alternative hypotheses for the two-population mean average hypothesis test. Notice equal (not related) is the null hypothesis and not equal (related) is the alternative hypothesis.

Let us suppose that the groups are independent of each other. There are a couple different ways of writing the null and alternative hypothesis. Notice that saying that the population means are equal is the same as saying the difference is zero. A not equal alternative hypothesis would be a two-tailed test.

μ_1: Mean Average of Population 1
μ_2: Mean Average of Population 2

(Two-tailed, two-population mean from independent groups)

$H_0 : \mu_1 = \mu_2$ (categorical variables are not related to the quantitative variable)
$H_A : \mu_1 \neq \mu_2$ (categorical variables are related to the quantitative variable)

OR

$H_0 : \mu_1 - \mu_2 = 0$ (categorical variables are not related to the quantitative variable)
$H_A : \mu_1 - \mu_2 \neq 0$ (categorical variables are related to the quantitative variable)

We can also specify that the population mean of population 1 is higher or lower than population 2. Notice that still indicates that the categorical and quantitative variables are related. If the alternative hypothesis is less than, then it is a left tailed test. Less than points to the left. If the alternative hypothesis is greater than, then it is a right tailed test.

Greater than points to the right. While some people prefer to use "≤" or "≥" symbol for the null hypothesis, I generally do not. Mainly because of the relationship idea. The null hypothesis is not related which must be equal to.

(Right-tailed, two-population mean from independent groups)

$H_0 : \mu_1 = \mu_2$ (categorical variables are not related to the quantitative variable)
$H_A : \mu_1 > \mu_2$ (categorical variables are related to the quantitative variable)

$H_0 : \mu_1 - \mu_2 = 0$ (categorical variables are not related to the quantitative variable)
$H_A : \mu_1 - \mu_2 > 0$ (categorical variables are related to the quantitative variable)

(Left-tailed, two-population mean from independent groups)

$H_0 : \mu_1 = \mu_2$ (categorical variables are not related to the quantitative variable)
$H_A : \mu_1 < \mu_2$ (categorical variables are related to the quantitative variable)

$H_0 : \mu_1 - \mu_2 = 0$ (categorical variables are not related to the quantitative variable)
$H_A : \mu_1 - \mu_2 < 0$ (categorical variables are related to the quantitative variable)

Sometimes we may have the same people measured twice or some one-to-one pairing between the groups. When this happens, we call this "matched pairs". If you recall from our discussion of matched pair confidence intervals, we subtract the ordered pairs. This creates the difference column of data. We then calculate the mean and standard deviation of the differences.

μ_d: Mean Average of Differences between the populations

(Two-tailed, two-population mean from matched pair data)

$H_0 : \mu_d = 0$ (categorical variables are not related to the quantitative variable)
$H_A : \mu_d \neq 0$ (categorical variables are related to the quantitative variable)

(Right-tailed, two-population mean from matched pair data)

$H_0 : \mu_d = 0$ (categorical variables are not related to the quantitative variable)
$H_A : \mu_d > 0$ (categorical variables are related to the quantitative variable)

(Left-tailed, two-population mean from matched pair data)

$H_0 : \mu_d = 0$ (categorical variables are not related to the quantitative variable)
$H_A : \mu_d < 0$ (categorical variables are related to the quantitative variable)

Two-population Mean Hypothesis Test Assumptions

The assumptions for two-population hypothesis tests are the same as for two-population confidence intervals that we discussed in previous chapters. The assumptions are slightly different depending on if the groups are matched pair or independent. Two-population hypothesis tests are also used in experimental design. In that case, we need the groups to be randomly assigned in order to control confounding variables. Another way to control confounding variables is to measure the same group of people twice (matched pair).

Two-population Mean Assumptions (Matched Pair)

- The quantitative ordered pair sample data should be collected randomly or be representative of the population.
- Data values within the sample should be independent of each other.
- There should be at least thirty ordered pairs or the differences should have a nearly normal shape.

Two-population Mean Assumptions (Not Matched Pair)

- The two quantitative samples should be collected randomly or be representative of the population.
- Data values within the sample should be independent of each other.
- Data values between the two samples should be independent of each other.
- The sample sizes should be at least 30 or have a nearly normal shape.

Two-population mean T-test statistic

The two population mean T-test statistic is very similar to the two-population proportion Z-test statistic. It just compares two sample means instead of two sample proportions. The two-population mean T-test statistic is used to determine if two sample means are significantly different. It can also be thought of as determining if the difference between the sample means is significantly different from zero or some other difference in the null hypothesis.

It is important not to confuse one and two-population test statistics. Recall that the one-population mean T-test statistic counts the number of standard errors that the sample mean $(\bar{x})$ is above or below the population mean (μ) in the null hypothesis. If the T-test statistic is positive, then the sample mean $(\bar{x})$ is a certain number of standard errors "above" the population mean (μ). If the T-test statistic is negative, then the sample mean $(\bar{x})$ is a certain number of standard errors "below" the population mean (μ).

The two-population mean T-test statistic will count how many standard errors that the sample mean for group 1 $(\bar{x}_1)$ is above or below the sample mean for group 2 $(\bar{x}_1)$. If the T-test statistic is positive, it is "above". If the T-test statistic is negative, it is "below". The two-population T-test statistic can also be thought of as the number of standard errors that the difference between the means is from zero or some other claimed difference.

Here are a couple of different formulas used by computer programs. If you recall in our discussions of confidence intervals two-population mean comparisons can come from data that is independent groups (like men and women) or matched pairs (like the same people measured twice). Again, it is not important for you to calculate these by hand with a calculator. Computers do the heavy lifting. Focus on being able to explain the test statistic and using it to determine significance.

These formulas are much easier to calculate if you already know the standard error. For independent groups, "n_1" is the sample size of group 1 and "n_2" is the sample size of group 2. The standard deviation for group 1 is "s_1" and the standard deviation for group 2 is "s_2".For matched pairs, the sample sizes of both groups are the same (n). The mean of the differences between the matched pairs is " $\bar{d}$ " and the standard deviation of the differences is " s_d ".

(Independent Groups) Two-population mean T-test statistic

$$T = \frac{(Sample\ Mean\ for\ group\ 1\ (\bar{x}_1) - Sample\ Mean\ for\ group\ 2\ (\bar{x}_2))}{Standard\ Error} \quad \text{OR} \quad \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\left[\left(\frac{s_1^2}{n_1}\right) + \left(\frac{s_2^2}{n_2}\right)\right]}}$$

(Matched Pairs) Two-population mean T-test statistic

$$T = \frac{(Mean\ of\ Differences\ between\ matched\ pairs\ (\bar{d}))}{Standard\ Error} \quad \text{OR} \quad \frac{(\bar{d})}{\left(\frac{s_d}{\sqrt{n}}\right)}$$

Example

Suppose we want to test the claim that the level of statistics student at COC is not related to the amount of alcohol they drink. If they are not related, then the population mean average amount of alcoholic beverages per week between COC pre-stat students should be the same as the mean average amount of alcoholic beverages per week between COC statistics students. In this case, the claim is the null hypothesis. Notice these are independent groups. Population 1 is COC statistics students and population 2 is COC pre-stat students. Here is the null and alternative hypothesis. Notice this will be a two-tailed hypothesis test. Use a 5% significance level.

$H_0 : \mu_1 = \mu_2$ (The level of stat student is not related to the amount of alcohol beverages per week) (Claim)
$H_A : \mu_1 \neq \mu_2$ (The level of stat student is related to the amount of alcohol beverages per week)

OR

$H_0 : \mu_1 - \mu_2 = 0$ (The level of stat student is not related to the amount of alcohol beverages per week) (Claim)
$H_A : \mu_1 - \mu_2 \neq 0$ (The level of stat student is related to the amount of alcohol beverages per week)

The data can be found on www.matt-teachout.org. We will be using the "COC Statistics Survey Data Fall 2015". We will be comparing the number of alcoholic drinks for Math 140 (statistics) students to the number of alcoholic drinks for Math 075 (pre-stat) students.

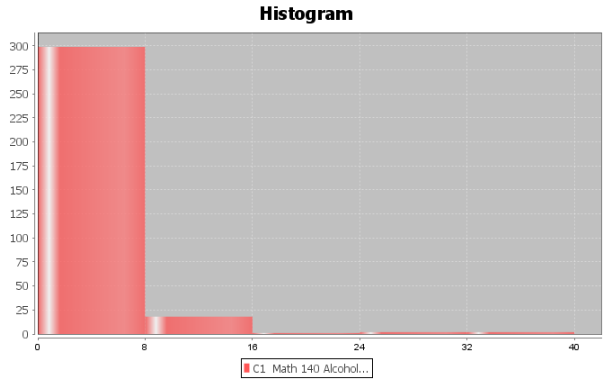
Let us start by checking the assumptions. The two data sets are not matched pair so we will check the assumptions for independent groups.

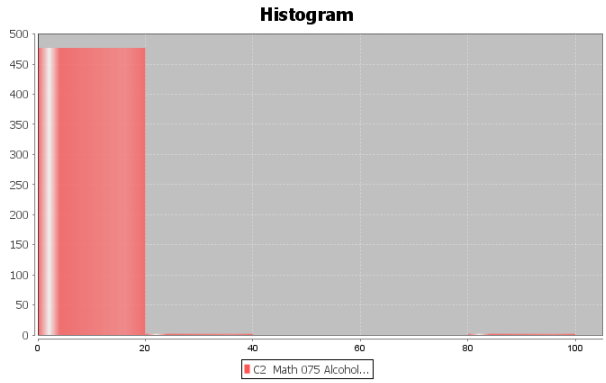
Assumptions for Two-Population Mean (Independent Groups)

- Sample data should be collected randomly or represent the population. If it is an experiment, then the groups should be randomly assigned. Yes. This sample data was not random, but it was a census of the fall 2015 semester, so is likely to be representative of COC students.
- The sample sizes for both groups should be at least 30 or nearly normal. Yes. The sample sizes are 322 and 481, which are both over 30. Even though both data sets are skewed right, it still passes the at least 30 or normal requirement.

Descriptive Statistics

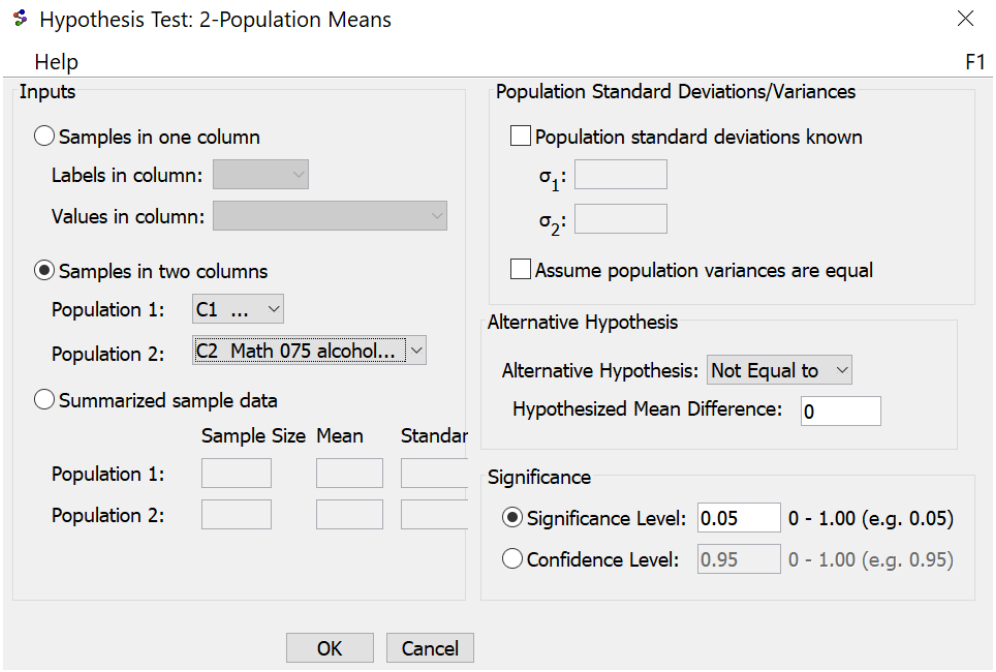
Variable	N total
C1 Math 140 Alcoholic Beverages per Week	322
C2 Math 075 Alcoholic Beverages per Week	481





- Data values within the samples and between the samples should be independent of each other. No. Some of the Math 140 students and Math 075 students may have come from the same class. Groups of friends may have similar alcohol consumption.

We want to use Statcato to calculate the test statistic, critical value and degrees of freedom. Since these data sets are over 300, we will need to add few rows before copy and pasting the data into Statcato. These data sets have a sample size of 322 and 481, so we will add about 200 rows to Statcato. Go to the "Edit" menu in Statcato and click on "Add multiple rows/columns". Put in 200 next to "rows" and push OK. We will get our sample data sets from the "Math 075 Survey Data Fall 2015" and the "Math 140 Survey Data Fall 2015" at www.matt-teachout.org under the "statistics" menu and "data sets". Copy and paste the alcohol beverages per week data for both groups into two columns of Statcato. Since these are independent groups, we will go to the "Statistics" menu in Statcato, click on "Hypothesis Tests", and then click on "2-population means". Since our raw data is in two columns, click on "Samples in two columns". Type in the column for math 140 alcoholic beverages under population 1 and math 075 as population 2. Notice saying that the groups are equal is the same as saying the difference is zero. Therefore, the "hypothesized mean difference" should be zero. In addition, the alternative hypothesis is "Not Equal To" and significance level is 0.05. Push OK.



Here is the Statcato printout.

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

Let us write a sentence to explain the T-test statistic. Remember, in this case group one is Math 140 statistics students and group 2 is Math 075 pre-statistics students. The sample mean number of alcoholic beverages per week for group 1 ($\bar{x}_1$) is 2.224 and the sample mean number of alcoholic beverages per week for group 2 ($\bar{x}_2$) is 1.47. Also, note that the test statistic is positive, indicating the group 1 is above group 2.

Sentence to explain the T-test statistic: The sample mean average number of alcoholic beverages per week for Math 140 statistics students is 1.846 standard errors above the sample mean average number of alcoholic beverages per week for Math 075 pre-statistics students.

Is it significant?

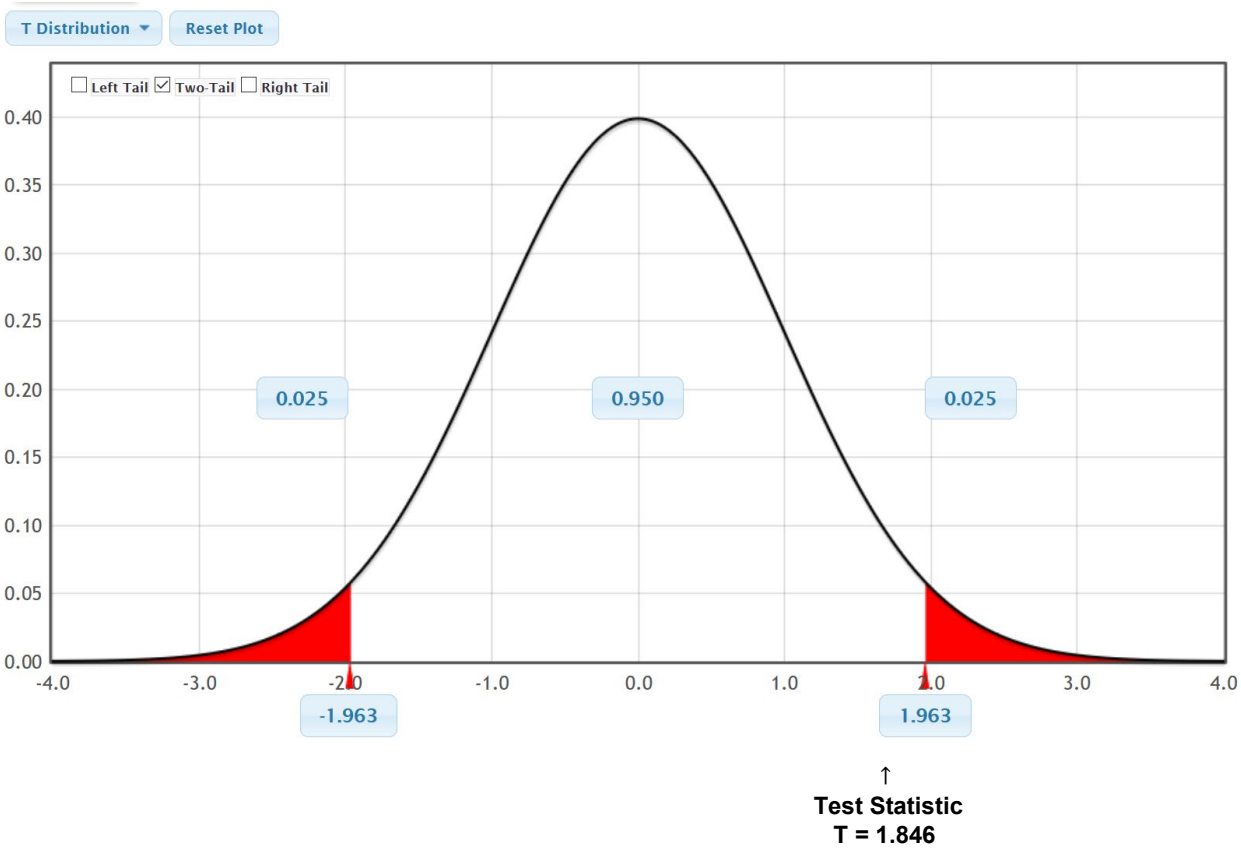
Notice that the test statistic did not fall in one of the tails determined by the critical values. This indicates that the sample means are not significantly different. This also indicates that the sample data does not significantly disagree with the null hypothesis.

Notice that the degrees of freedom is 800 in the Statcato printout. How did Statcato calculate this? The formula for two-population mean degrees of freedom from independent groups is given below. You will need the sample sizes and standard deviations for both groups. Again, never calculate this by hand. Statcato calculated it for us. If you do not have Statcato, there are many two-population mean degrees of freedom calculators for independent groups. Here is one I like to use. (http://web.utk.edu/~cwiek/TwoSampleDoF). You will want to round the degrees of freedom to the ones place. In this example, the app above gave "800.7819" which is close to what Statcato gave. We usually round the degrees of freedom down in order to account for more variability. An easier formula for two-population mean degrees of freedom for independent groups is to use the smaller of n_1-1 or n_2-1.

(Independent groups) Two-population mean degrees of freedom = $$\frac{\left[\left(\frac{s_1^2}{n_1}\right)+\left(\frac{s_2^2}{n_2}\right)\right]^2}{\left[\left(\frac{\left[s_1^2/n_1\right]^2}{n_1-1}\right)+\left(\frac{\left[s_2^2/n_2\right]^2}{n_2-1}\right)\right]}$$

(Matched Pair) Two-population mean degrees of freedom = n – 1

It is enough to know now that we can also use StatKey to calculate and visually see the critical values. Go to the "theoretical distributions" menu in StatKey at www.lock5stat.com and click on "t". Put in 800 degrees of freedom and click "two-tail". Since we are using a 5% significance level in two tails, each tail should have a proportion of 0.025 (2.5%). Notice StatKey gives the same critical values that Statcato gave.



↑
Test Statistic
T = 1.846

P-value and Conclusions

Let us see if we can finish this test about the level of statistics student and the amount of alcoholic beverages consumed per week. The Statcato printout indicated that the P-value is 0.0653. Since the P-value is higher than the 5% significance level, we will fail to reject the null hypothesis. The claim was the null hypothesis so our conclusion should be that we do not have significant evidence to reject the claim.

Conclusion: There is not significant evidence to reject the claim that the level of stat student is not related to the amount of alcohol beverages per week.

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

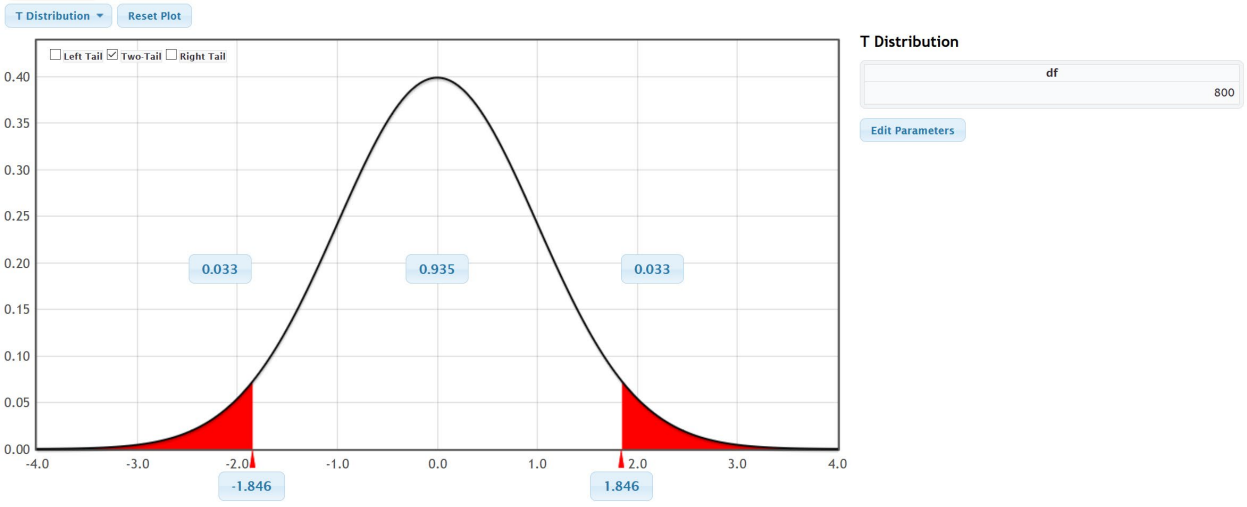
Significance Level	Critical Value	Test Statistic t	p-Value
0.05	-1.963, 1.963	1.846	0.0653

How was the P-value calculated?

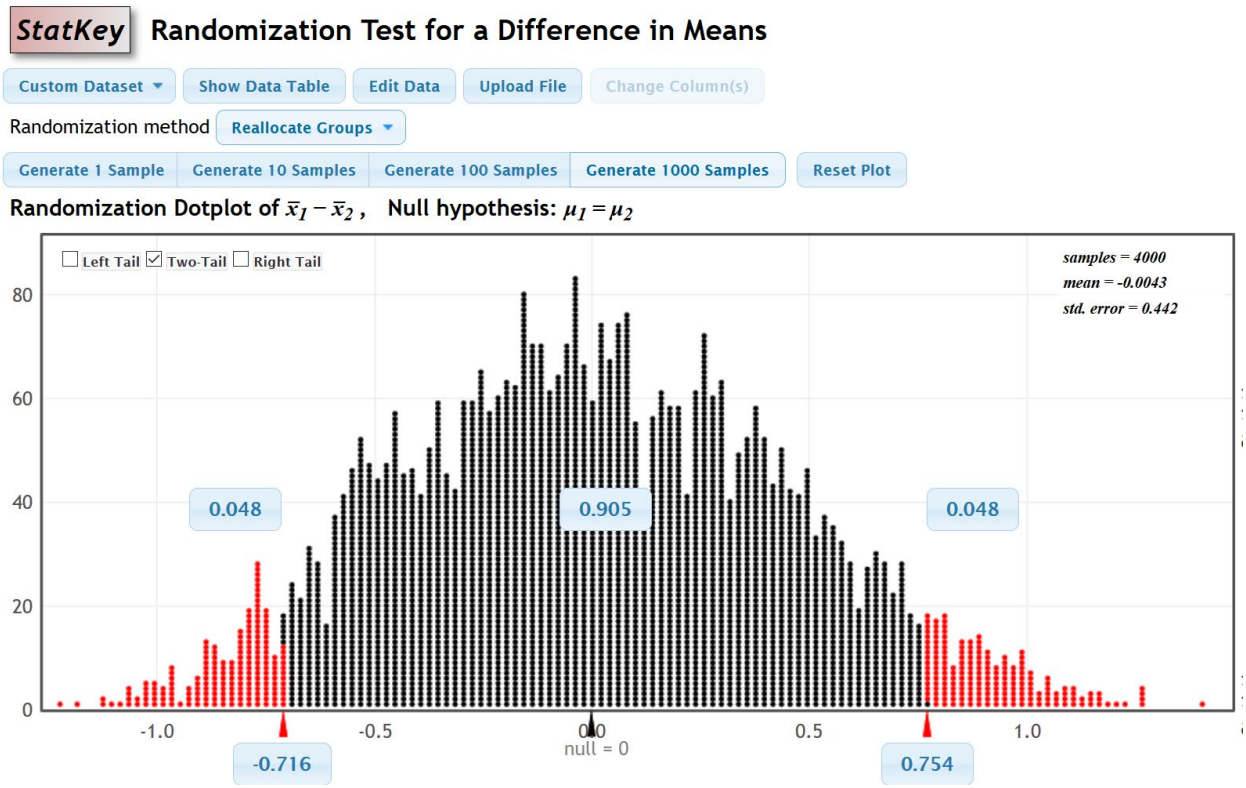
P-values again can be calculated in different ways. A traditional approach would be to calculate the proportion in the tail or tails corresponding to the test statistic. Recall the degrees of freedom was 800. Using the theoretical distribution T calculator in StatKey, we can calculate the proportion in the tails. We entered the degrees of freedom and clicked "Two-Tail". We then entered our test statistic of +1.846 in the right tail since it was positive. The left tail automatically adapted. This was a two-tailed test, so we will need to add the proportions in the tails to get the P-value.

P-value = 0.033 + 0.033 = 0.066

P-value sentence: If the null hypothesis is true and the level of statistics student is not related to alcohol, then there is a 6.6% probability of getting this sample data or more extreme because of sampling variability.

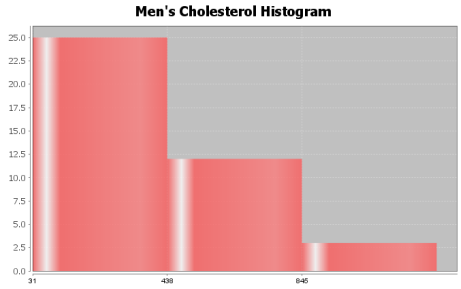


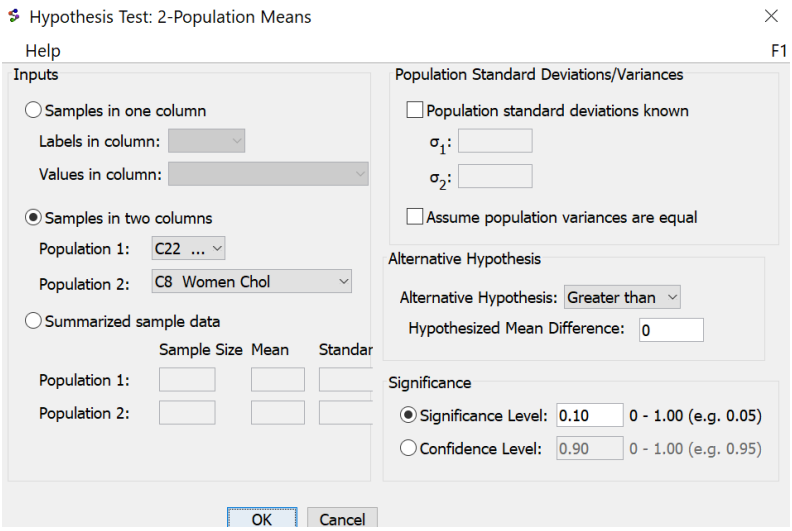
We learned in the last chapter that we could also use randomized simulation to estimate the P-value. Open StatKey at www.lock5stat.com. When computing a two-population mean hypothesis test with StatKey, we will need the raw categorical and quantitative data. Open the "Math 075 140 combined survey Data Fall 2015" at www.matt-teachout.org. Copy the student level data and the alcoholic beverages data next to each other in a fresh excel spreadsheet. Under the "Randomization Hypothesis Tests" menu click on "Test for Difference in Means". Click on "Edit Data" and copy and paste both columns into StatKey. Click "Generate 1000 Samples" a few times. Remember this was a two-tailed test. The sample difference between the population 1 and population 2 was 0.754. Enter the sample difference of 0.754 into the right tail. Add the proportions in the tails to get the approximate P-value of 0.048 + 0.048 = 0.096 or 9.6%.



Example 2 (Two-population mean average T-test with Independent groups)

Some people believe that the population mean average cholesterol for men and women is the same, while others think that men's cholesterol is higher. Use the randomly collected health data at www.matt-teachout.org to test the claim that the mean average cholesterol for men (μ_1) is higher than the mean average cholesterol for women (μ_2). This claim would indicate that gender is related to cholesterol. Use the following Statcato printout, graphs and a 10% significance level.





Hypothesis Test - Two population means: confidence level = 0.90

Samples of population 1 in C22 Men Chol

Samples of population 2 in C8 Women Chol

	N	Mean	Stdev
Population 1	40	395.225	292.412
Population 2	40	240.875	185.982

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

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

* Population standard deviations are unknown. *

DOF = 66

Significance Level	Critical Value	Test Statistic t	p-Value
0.10	1.295	2.817	0.0032

Null and alternative hypothesis

H_0 : $\mu_1 = \mu_2$ (Gender and Cholesterol are NOT related)

H_A : $\mu_1 > \mu_2$ (Gender and Cholesterol ARE related) (Claim)

Type of hypothesis test? Two-population Mean T-test (right tail with independent groups)

Assumptions? The data did pass all of the assumptions, so we can proceed with the hypothesis test.

- Both samples were collected randomly.
- Both samples pass the at least 30 or normal requirement. Even though both data sets were skewed right, they both had a sample size of 40.

- Data values within the samples were likely to be independent. These are simple random samples out of large populations. It is unlikely that the individual men in the data will be related. It is also unlikely that individual women will be related.
- Data values between the samples were likely to be independent. The groups were not matched pairs. They were not the same people measured twice or some other one to one pairing. Since the men and women were collected randomly out of a large population, it is unlikely they will be related.

T-test statistic = 2.817

Test Stat Sentence: The sample mean cholesterol for the men (395.225 mg/dL) is 2.817 standard errors above the sample mean cholesterol for the women (240.875 mg/dL).

- The right tail starts at the critical value of 1.295, so the test statistic definitely falls in the right tail and is significant.
- This tells us that the sample mean cholesterol for the men is significantly higher than for the women.
- The sample data significantly disagrees with the null hypothesis.

P-value = 0.0032 = 0.32%

P-Value Sentence: If Ho is true, and men and women have the same population mean average cholesterol, then we had a 0.32% probability of getting the sample data or more extreme because of sampling variability.

- This is a low P-value. (The P-value of 0.32% is much smaller than the 10% significance level.)
- If H_0 is true, this tells us that the sample data was unlikely to have happened by random chance (sampling variability).
- A low P-value also indicates significance. This tells us that the sample mean cholesterol for the men is significantly higher than for the women.
- There is a significant disagreement between the sample data and the null hypothesis.

Reject H_0 or Fail to reject H_0? Reject H_0 since the P-value (0.32%) is smaller than the 10% significance level.

Conclusion?

There is significant evidence to support the claim that the population mean average cholesterol for men is higher than the population mean average cholesterol for women. This also gives evidence that a gender is related to cholesterol.

(The random sample data significantly agrees with the claim that the population mean average cholesterol for men is higher than the population mean average cholesterol for women. We have a low P-value as evidence.)
