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 : μ_d = 0 (categorical variables are not related to the quantitative variable)
H_A : μ_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

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

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

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

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.

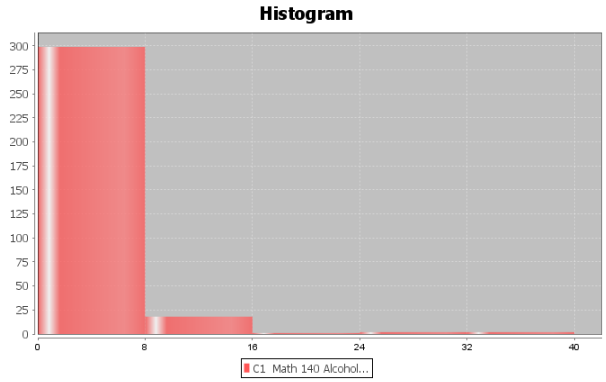
Let us start by checking the assumptions. The two data sets are not matched pair so we will check the assumptions for independent groups.

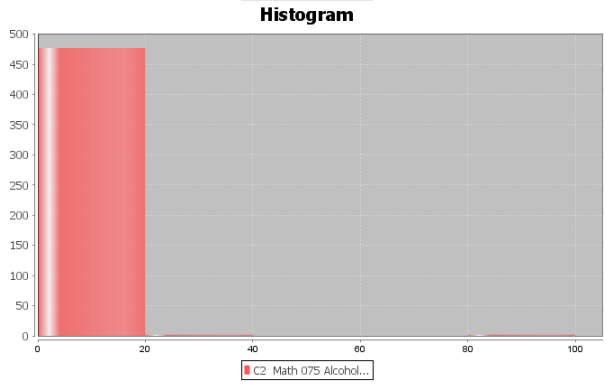
Assumptions for Two-Population Mean (Independent Groups)

- 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

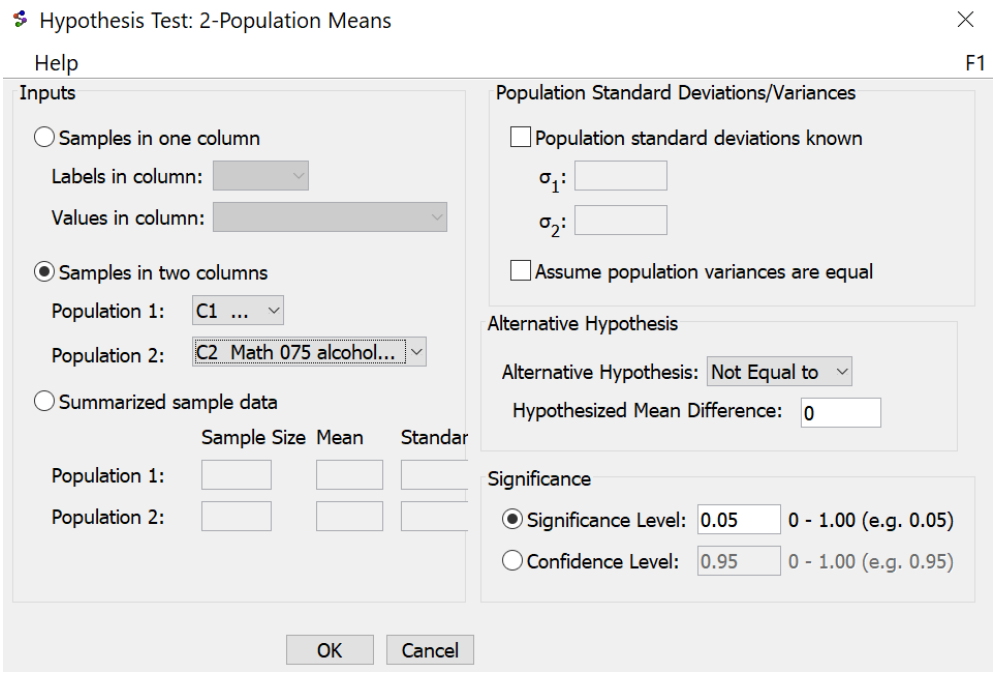
Variable	N total
C1 Math 140 Alcoholic Beverages per Week	322
C2 Math 075 Alcoholic Beverages per Week	481





- 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?

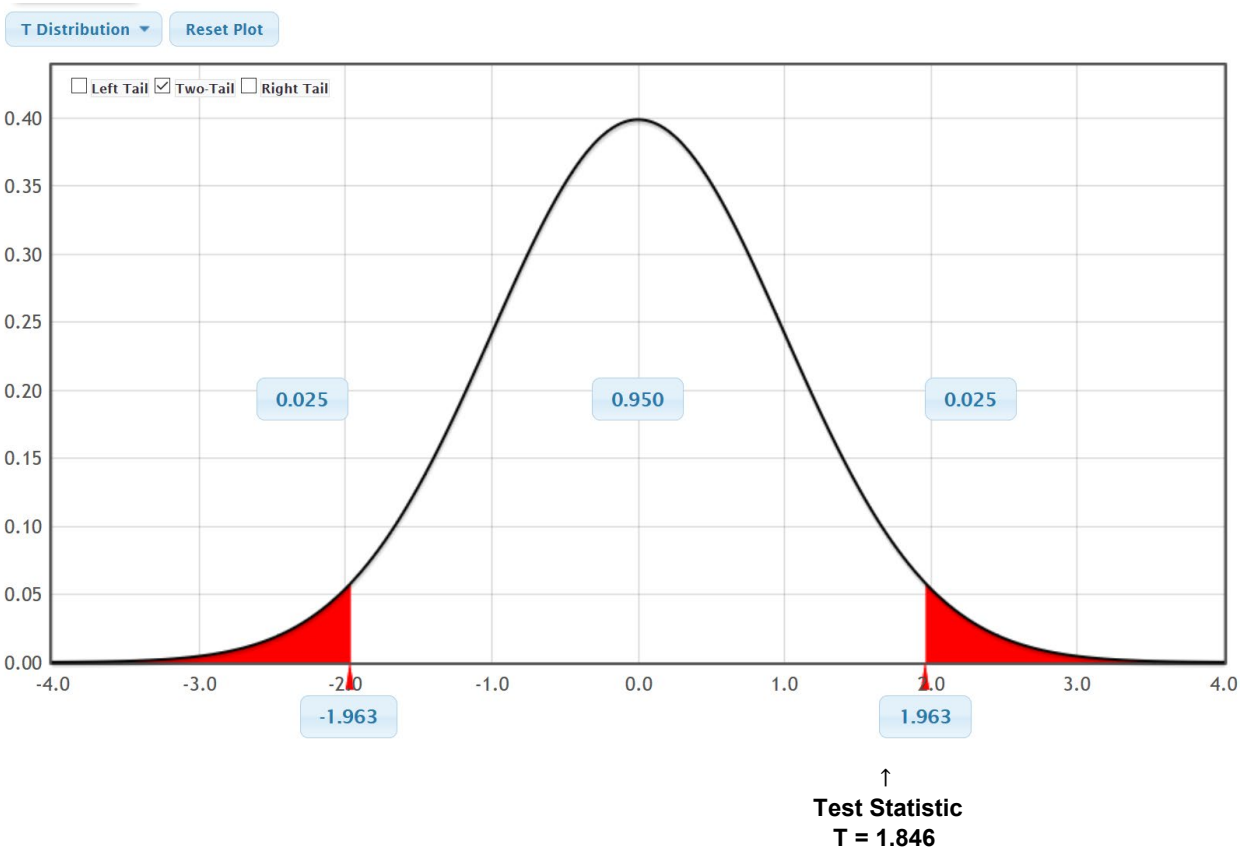
Notice that the test statistic did not fall in one of the tails determined by the critical values. This indicates that the sample means are not significantly different. This also indicates that the sample data does not significantly disagree with the null hypothesis.

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.



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

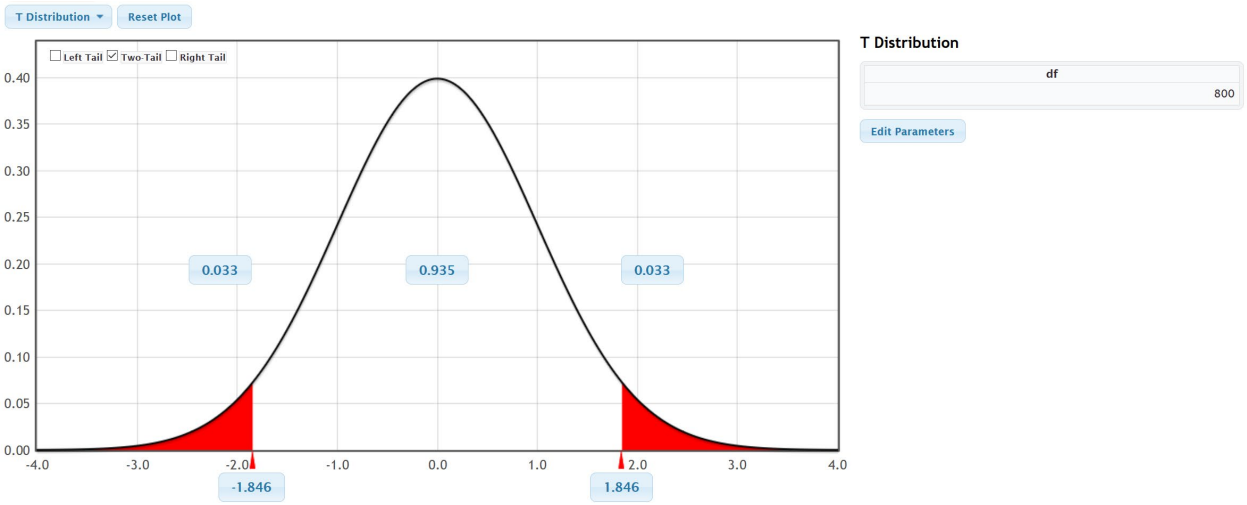
Significance Level	Critical Value	Test Statistic t	p-Value
0.05	-1.963, 1.963	1.846	0.0653

How was the P-value calculated?

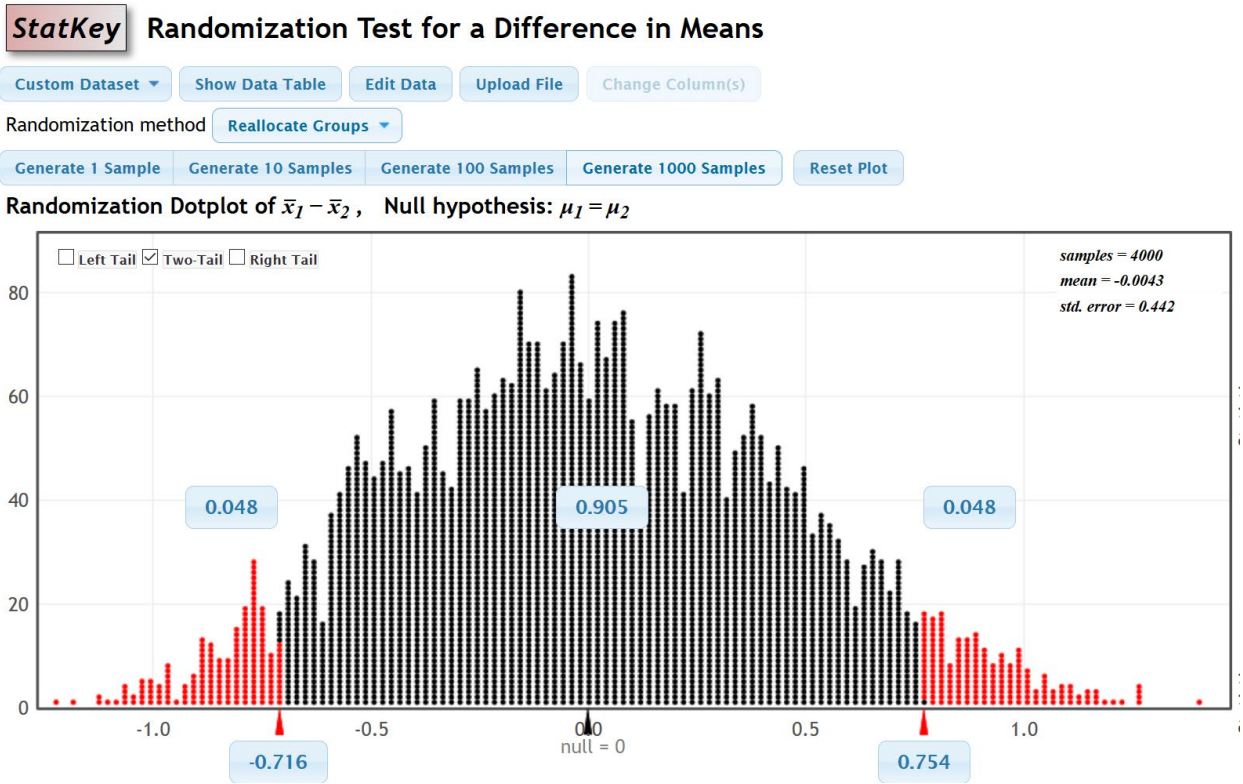
P-values again can be calculated in different ways. A traditional approach would be to calculate the proportion in the tail or tails corresponding to the test statistic. Recall the degrees of freedom was 800. Using the theoretical distribution T calculator in StatKey, we can calculate the proportion 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.

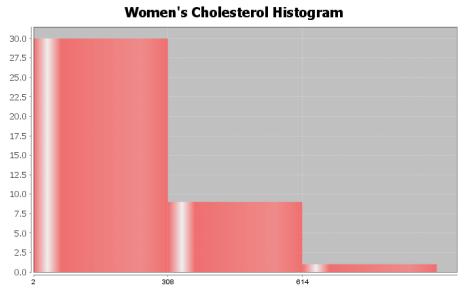


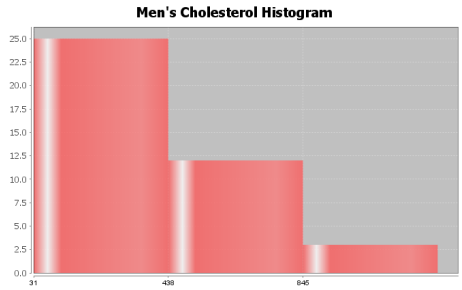
We learned in the last chapter that we could also use randomized simulation to estimate the P-value. Open StatKey at www.lock5stat.com. When computing a two-population mean hypothesis test with StatKey, we will need the raw categorical and quantitative data. Open the "Math 075 140 combined survey Data 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%.

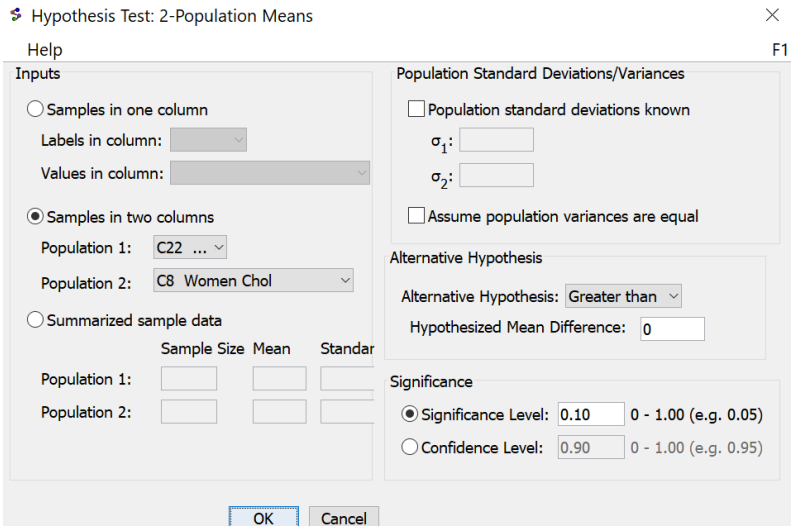


Example 2 (Two-population mean average T-test with Independent groups)

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.)

Practice Problems Section 4A

(#1-10) Use each of the following two-population mean T-test statistics and the corresponding critical values to fill out the table.

	Type of Test	T-test stat	Sentence to explain T-test statistic.	Critical Value	Does the T-test statistic fall in a tail determined by a critical value? (Yes or No)	Are the sample means from the two groups significantly different or not? Explain.	Does sample data significantly disagree with H_0? Explain.
1.	Right Tailed	+1.383		+2.447			
2.	Left Tailed	−2.851		−1.773			
3.	Two Tailed	−1.501		±2.006			
4.	Right Tailed	+3.561		+1.692			
5.	Two Tailed	+0.887		±1.943			
6.	Left Tailed	−1.003		−2.759			
7.	Two Tailed	−4.416		±1.994			
8.	Right Tailed	+0.275		+1.839			
9.	Left Tailed	−1.461		−1.674			
10.	Two Tailed	+2.330		±2.138			

(#11-20) Use each of the following P-values and corresponding significance levels to fill out the table.

	P-value Proportion	P-value %	Sentence to explain the P-value	Significance Level %	Significance level Proportion	If H_0 is true, could the sample data occur by sampling variability or is it unlikely?	Reject H_0 or Fail to reject H_0?
11.	0.0007			10%			
12.	0.421			1%			
13.	8.71×10^{-5}			5%			
14.	0.339			1%			
15.	0.076			5%			
16.	0			10%			
17.	0.528			5%			
18.	0.0277			10%			
19.	3.04×10^{-6}			1%			
20.	0.178			5%			

21. Explain the difference between matched pair data and independent groups.

22. Explain the difference between random samples and random assignment.

23. List the assumptions that we need to check for a two-population mean hypothesis test from independent groups.

24. List the assumptions that we need to check for a two-population mean hypothesis test from matched pairs.

25. List the assumptions that we need to check for a two-population mean hypothesis test that is using experimental design.

26. Explain how to use a two-population mean hypothesis test to show that categorical and quantitative data are related.

27. Explain how to use a two-population mean hypothesis test to show there is a cause and effect between categorical and quantitative data.

(#28-33) Directions:

a) Determine if the following two-population mean tests are matched pair or independent groups

b) Write the null and alternative hypothesis. Include relationship implications.

c) Check all of the assumptions for a two-population mean T-test. Explain your answers. Does the problem meets all the assumptions?

d) Write a sentence to explain the T-test statistic.

e) Use the test statistics and the critical value to determine if the sample data significantly disagrees with the null hypothesis. Explain your answer.

f) Write a sentence to explain the P-value.

g) Use the P-value and significance level to determine if the sample data could have occurred by random chance (sampling variability) or is it unlikely to random chance? Explain your answer.

h) Should we reject the null hypothesis or fail to reject the null hypothesis? Explain your answer.

i) Write a conclusion for the hypothesis test. Explain your conclusion in plain language.

j) Is the categorical variable related to the quantitative variable? Explain your answer.

28. The ACT exam is used by many colleges to test the readiness of high school students for college. Many high school students are now taking ACT prep classes. A local high school offers an ACT prep class, but wants to know if it really helps. Twenty students were randomly selected. They took the ACT exam before and after taking the ACT prep class. For each student the difference between the after and before scores were measured (d = after – before). Population 1 was the after prep class scores and population 2 was the before prep class scores. The mean of the differences was 1.5 ACT points with a standard deviation of 2.3 ACT points. A histogram of the differences yielded a bell shaped normal distribution. Use a 5% significance level to test the claim that the after prep class scores are higher than the before prep class scores. What does this data indicate about the relationship between taking a prep class or not and ACT scores.

N	Sample Mean	Stdev	Significance Level	Critical Value	Test Statistic t	p-Value
20	1.5	2.3	0.05	1.729	2.917	0.0044

29. A random sample of 20 male German Shepherds found that their average weight was 112 pounds with a standard deviation of 28 pounds. A random sample of 14 male Dobermans found that their average weight is 107 pounds with a standard deviation of 24 pounds. Assume that weights are normally distributed. Use the Statcato printout below and a 5% significance level to test the claim that the population mean average weight of male German Shepherds (population 1) is more than the population mean average weight of male Doberman Pinchers (population 2). What does this data indicate about the relationship between the weight and the type of dog?

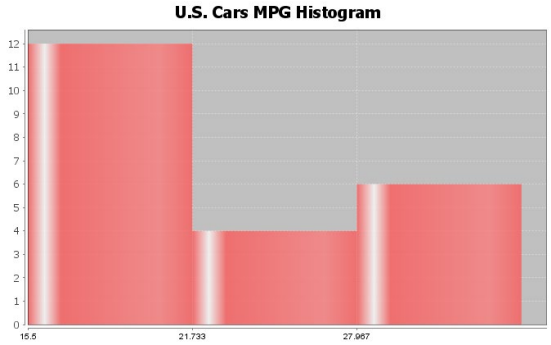
	N	Mean	Stdev
German Shep Sample 1	20	112.0	28.0
Doberman Sample 2	14	107.0	24.0

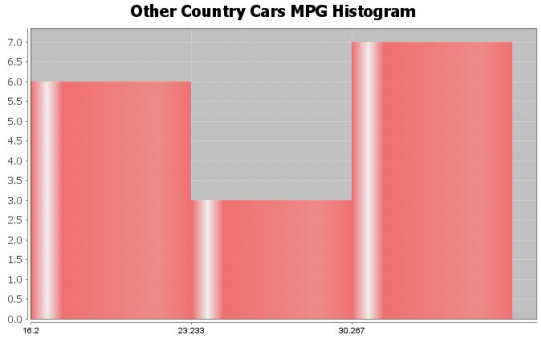
Significance Level	Critical Value	Test Statistic t	p-Value
0.05	1.697	0.558	0.2906

30. Cotinine is an alkaloid found in tobacco and is used as a biomarker for exposure to cigarette smoke. It is especially useful in examining a person's exposure to second hand smoke. A random sample of 32 non-smoking American adults was collected. These adults were not smokers and did not live with any smokers. The average cotinine level for this sample was 7.2 ng/mL with a standard deviation of 5.8 ng/mL. A second random sample of 35 non-smoking American adults was then collected. These adults did not smoke themselves, but did live with one or more smokers. The mean average cotinine level for this sample was 28.5 ng/mL and had a standard deviation of 11.4 ng/mL. Use a 1% significance level to test the claim that people that do not live with smokers have a lower cotinine level than those people that do live with smokers. What does this data indicate about the relationship between cotinine levels and living with a smoker or not.

Significance Level	Critical Value	Test Statistic t	p-Value
0.01	-2.402	-9.758	$1.4751 \cdot 10^{-13}$

31. We want to see if the country a car is made in is related to its gas mileage in miles per gallon. Specifically we wanted to see if cars made in the U.S. have a lower population mean average mpg than those made outside the U.S. We used the random car data at www.matt-teachout.org and a 5% significance level to create the following graphs and statistics with Statcato. Check the assumptions and perform the hypothesis test.



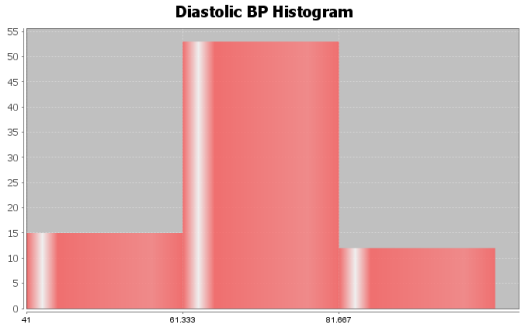


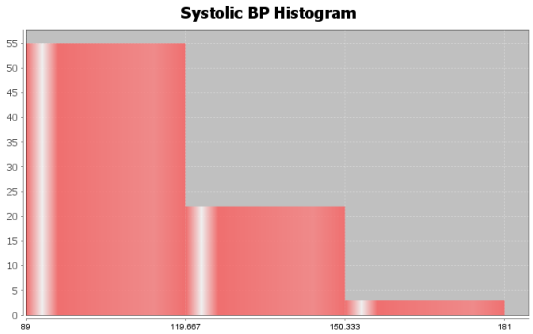
	N	Mean	Stdev
Population 1 USA mpg	22	22.995	6.054
Population 2 Other Country mpg	16	27.188	6.601

* Population standard deviations are unknown. *
DOF = 30

Significance Level	Critical Value	Test Statistic t	p-Value
0.05	-1.697	-2.001	0.0273

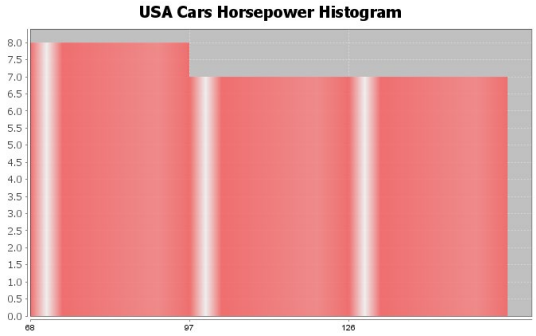
32. We want to test the claim that the diastolic blood pressure of a person is less than the systolic blood pressure of a person. We used the random health data at www.matt-teachout.org, Statcato, and a 1% significance level to create the following graphs and statistics. Check the assumptions and perform the hypothesis test. Notice that since the diastolic and systolic blood pressures came from the same randomly selected adults, we used a matched pair calculation.

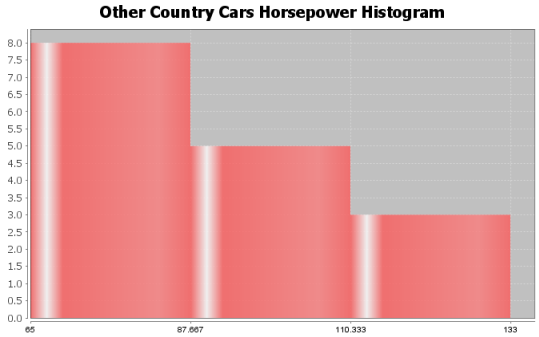




N	Sample Mean	Stdev	Significance Level	Critical Value	Test Statistic t	p-Value
80	-44.525	10.077	0.01	-2.375	-39.521	$4.188 \cdot 10^{-54}$

33. We want to see if the country a car is made in is related to the horsepower of the car. Specifically we wanted to see if cars made in the U.S. have a higher population mean average horsepower than those made outside the U.S. We used the random car data at www.matt-teachout.org and a 10% significance level to create the following graphs and statistics with Statcato. Check the assumptions and perform the hypothesis test.





	N	Mean	Stdev
Population 1 USA car horsepower	22	110.182	26.383
Population 2 Other Country car horsepower	16	90.125	22.408

Significance Level	Critical Value	Test Statistic t	p-Value
0.10	1.306	2.526	0.0081

(#34-37) Directions: For each problem, answer the following questions.

a) Determine if the following two-population mean tests are matched pair or independent groups

b) Write the null and alternative hypothesis. Include relationship implications.

c) Use randomized simulation to calculate the P-value. Write a sentence to explain the P-value.

d) Use the P-value and significance level to determine if the sample data could have occurred by random chance (sampling variability) or is it unlikely to random chance? Explain your answer.

e) Use the sample mean difference and the standard error in the simulation to calculate the T-test statistic.

$$T = \frac{Sample\ Mean\ Difference}{Standard\ Error}$$

e) Should we reject the null hypothesis or fail to reject the null hypothesis? Explain your answer.

f) Write a conclusion for the hypothesis test. Explain your conclusion in plain language.

g) Is the categorical variable related to the quantitative variable? Explain your answer.

34. Go to StatKey at www.lock5stat.com. Under the "Randomization Hypothesis Tests" menu, click on "Test for Difference for Difference in Means". Under the data sets menu on the top left, click on "Commute Time (Atlanta vs St Louis)". This took a random sample of people from Atlanta (population 1) and a random sample of people from St. Louis (population 2). Use randomized simulation and a 5% significance level to test the claim that the mean average commute time for people in Atlanta is greater than the mean average commute time for people from St. Louis. What does this data indicate about the relationship between the city and the commute time?

35. Go to StatKey at www.lock5stat.com. Under the "Randomization Hypothesis Tests" menu, click on "Test for Single Mean". Under the data sets menu on the top left, click on "Pulse Rate Difference (Quiz – Lecture)". An experiment was done on college students to determine if heart rate is related to taking a quiz or not. The heart rates of students were measured on a day they were taking a quiz (population 1) and again on a day when there was just lecture (population 2). The same students were measured twice. Use randomized simulation and a 1% significance level to test the claim that the mean average heart rate difference between the quiz and lecture days is greater than zero. This will indicate that the heart rates on quiz days tend to be higher than lecture days. What does this data indicate about the relationship between the heart rate and taking a quiz or not?

36. Go to StatKey at www.lock5stat.com. Under the "Randomization Hypothesis Tests" menu, click on "Test for Difference for Difference in Means". Under the data sets menu on the top left, click on "Exercise Hours (Male vs Female)". This took a random sample of male adults (population 1) and a random sample of female adults (population 2). Use randomized simulation and a 10% significance level to test the claim that the mean average amount of time that males and females exercise is the same. What does this data indicate about the relationship between exercise hours and gender?

37. Use StatKey and the random health data to test the claim that the population mean average pulse rate for women is higher than for men. Go to www.matt-teachout.org and click on the statistics tab and then the data sets tab. Open the health data. Copy and paste the gender data and pulse data columns next to each other in a fresh excel spreadsheet. Now copy the two columns. Go to StatKey at www.lock5stat.com. Under the "Randomization Hypothesis Tests" menu, click on "Test for Difference for Difference in Means". Under "Edit Data", paste the gender and pulse rate columns into StatKey. Click on "Generate 1000 Samples" a few times. Click on "Right Tail" and put in the sample mean difference of 6.9 beats per minute in the bottom box in order to estimate the P-value. Now answer the questions above.

Section 4B – Categorical/Quantitative Relationships: ANOVA

Introduction

In the last section, we saw that we could use a two-population mean average hypothesis test to determine if categorical and quantitative variables are related or not. If the mean averages were the same in two groups that would indicate that the categorical variable that determines the groups is not related to the quantitative variable mean average.

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

Many times, categorical data has more than just two options. This would mean that we would need to compare three or more population means.

H_0 : $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \ldots = \mu_k$ (categorical variable is not related to the quantitative variable)
H_A : at least one population mean is ≠ (categorical variable is related to the quantitative variable)

Unfortunately, a T test statistic can only compare two things at a time and cannot handle a hypothesis test involving three or more groups. To compare three or more population means, we will need to use ANOVA.

ANOVA

ANOVA stands for "Analysis of Variance". If you remember, variance is the square of the standard deviation. Variance measures the variability from the mean. There are two specific variances that are compared in an ANOVA test, the variance between the groups and the variance within the groups. The variance between the groups is a measure of how different the groups are. It measures how much variability each of the sample means are from the mean of all the groups combined. The variance within the groups measures the amount of variability that data values in each group are from their own sample mean. In ANOVA, we compare the variance between the groups to the variance within the groups.

ANOVA tests use the F-test statistic. In ANOVA, the F-test statistic divides the variance between the groups to the variance within the groups.

F Test Statistic Sentence: The ratio of the variance between the groups to the variance within the groups.

Calculating and Interpreting the F-test Statistic

As with all difficult calculations in statistics, use a computer program to calculate the F-test statistic. Never calculate it by hand. Always focus more on interpretation than on calculation. Let us see if we can better understand how the F-test statistic works.

Variance divides the sum of squares of the differences by the degrees of freedom.

$$Variance = \frac{Sum\ of\ Squares}{Degrees\ of\ Freedom}$$

To calculate the variance between the groups, the computer calculates the sum of squares between the groups and then divides by the degrees of freedom. The sum of squares between the groups subtracts the mean of all the groups combined ($\bar{\bar{x}}$)from the sample means ($\bar{x}_i$) for each group. It squares the differences and adds them. Since the variance between calculations is based on the number of groups, the degrees of freedom between is the number of groups – 1 or "k – 1".

$$Variance = \frac{Sum\ of\ Squares\ Between}{Degrees\ of\ Freedom\ Between} = \frac{\sum(\bar{x}_i - \bar{\bar{x}})^2}{k-1}$$

To calculate the variance within the groups, we divide the sum of squares within each group divided by the sum of squares within. The "sum of squares within" subtracts each data value minus its own sample mean, squares the differences and adds them up. If we look at the degrees of freedom for each data set $(n_i - 1)$and add them up for each group, we will get the "degrees of freedom within".

$$Variance = \frac{Sum\ of\ Squares\ Within}{Degrees\ of\ Freedom\ Within} = \frac{\sum(x-\bar{x}_i)^2}{\sum(n_i-1)}$$

There is a beauty in the mathematics behind the F test statistic. The total number of data values for all of the groups combined minus one is often called the total degrees of freedom. There is also a total sum of squares.

Sum of Squares Between + Sum of Squares Within = Total Sum of Squares

Degrees of Freedom Between + Degrees of Freedom Within = Total Degrees of Freedom

Variance Between + Variance Within = Total Variance

As we said, the F test statistic divides the Variance between the groups by the Variance within the groups. We often say that the F-test statistic is the ratio of two variances. In ANOVA, it is the ratio of the variance between the groups to the variance within the groups.

$$\text{F test statistic} = \frac{Variance\ Between\ the\ Groups}{Variance\ Within\ the\ Groups} = \frac{\left(\frac{Sum\ of\ Squares\ Between}{Degrees\ of\ Freedom\ Between}\right)}{\left(\frac{Sum\ of\ Squares\ Within}{Degrees\ of\ Freedom\ Within}\right)}$$

Computer Programs will often give you sum of squares, degrees of freedom, and variances for the F test statistic. Look at the following printout. This test used a 5% significance level.

Source of Variation	DOF	SS	MS	Test statistic F	Critical value F	p-Value
Treatment (Between Groups)	4	10484529.98982	2621132.49746	7.92175	2.4248	$7.03917 \cdot 10^{-6}$
Error (Within Groups)	170	56249274.83547	330878.08727			
Total	174	66733804.82529				

Let us see if we understand what we are seeing. Notice the MS (mean sum of squares) is the sum of squares (SS) divided by degrees of freedom (df).

"MS Treatment (Between Groups) is the variance between the groups 2621124.8 which was calculated by dividing the sum of squares (SS Between) 10484529.98982 by the degrees of freedom (DOF Between) 4.

"MS Error (Within Groups) is the variance within the groups 330878.08727 which was calculated by dividing the sum of squares (SS Within) 56249274.83547 by the degrees of freedom (DOF Within) 170.

So the F-test statistic is calculated by dividing the variances (MS).

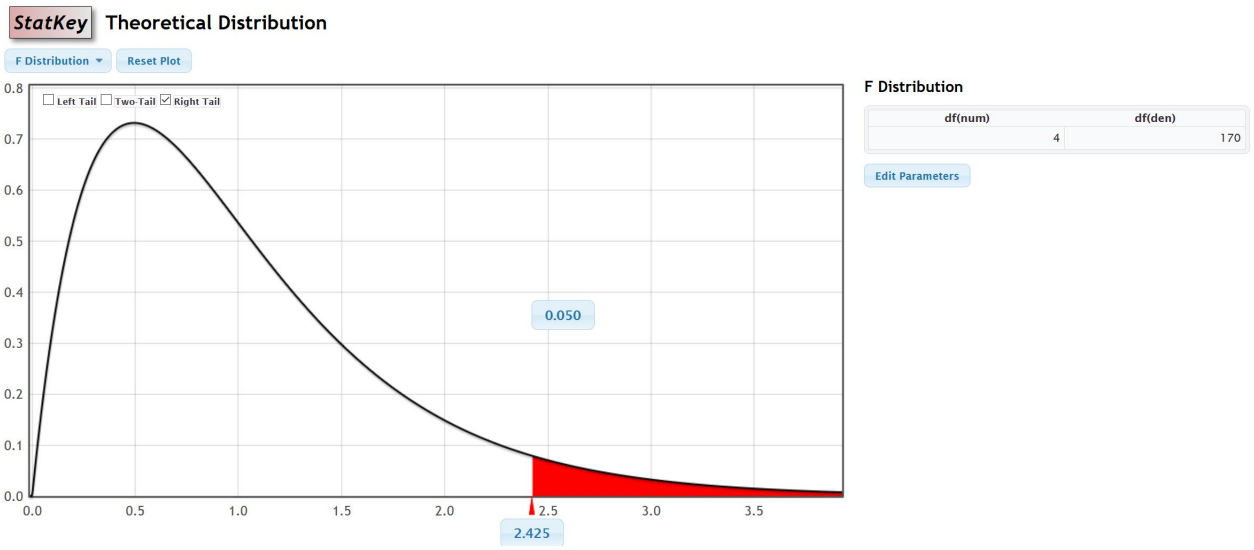
$$\text{F test statistic} = \frac{Variance\ Between\ the\ Groups}{Variance\ Within\ the\ Groups} = \frac{2621124.8}{330878.08727} = 7.92175$$

So the variance between the groups is almost 8 times greater than the variance within the groups.

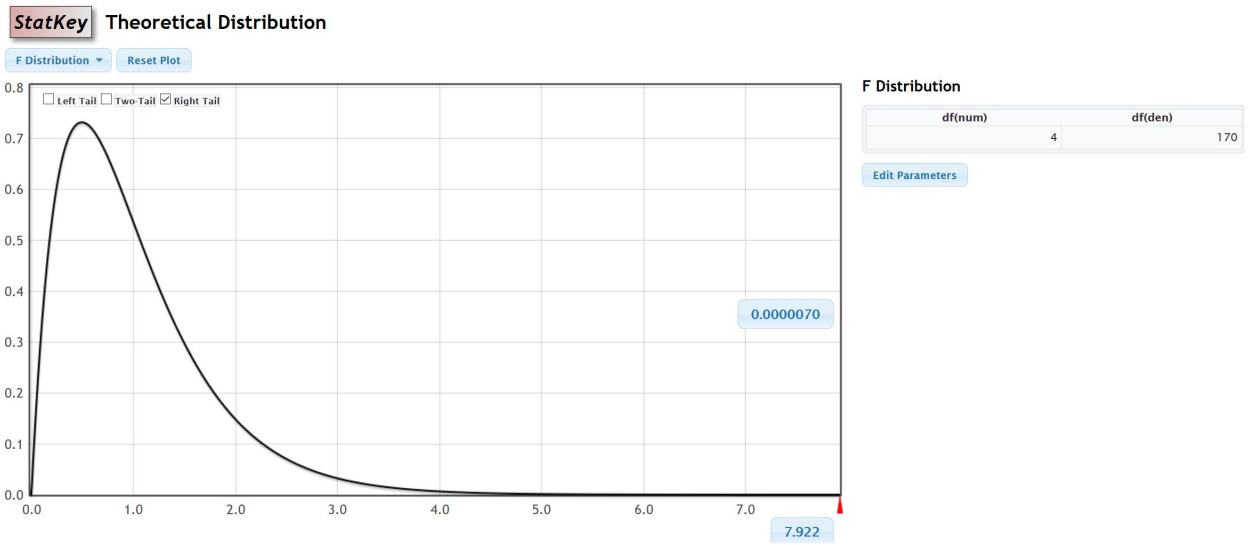
Is this F test statistic significant?

Notice Statcato has calculated a critical value to compare the test statistic to. ANOVA is always a right tailed test. Remember the test statistic needs to be in the tail determined by the critical value in order to be significant. Statcato thinks that the F test statistic has to be 2.4248 or higher to be significant. Our test statistic is 7.9217, which is definitely larger than the critical value 2.4248 and falls in the right tail. So our F test statistic is significant. Therefore, the F test statistic is significantly large and the variance between the groups is significantly greater than the variance within the groups. As with all test statistics, this also tells us that the sample data significantly disagrees with the null hypothesis.

We could have looked up the critical value with StatKey as well. You will need to know the degrees of freedom between (numerator degrees of freedom) which was four and the degrees of freedom within (denominator degrees of freedom). Go to www.lock5stat.com and open StatKey. Under "Theoretical Distributions" click on "F". Put in the numerator degrees of freedom as 4 and the denominator degrees of freedom as 170. Since ANOVA is always a right tailed test and the significance level is 5%, simply click on "Right Tail" and enter 0.05 for the right tail proportion. Notice the number on the bottom is the critical value. It is about the same as what Statcato gave.



Notice we can also use the same F theoretical curve to calculate the P-value with StatKey. Remember the numerator degrees of freedom is 4 and the denominator degrees of freedom is 170. Now just put the F test statistic in the bottom box in the right tail. The proportion is the P-value. Notice the P-value calculated by StatKey is very close to the P-value calculated by Statcato.



Notes about the F-test statistic

- In a fraction, when the numerator is significantly larger than the denominator, the overall fraction is large. If the variance between the groups is much larger than the variance within the groups, this will give a large F-test statistic (and a small P-value) and indicates that the sample means are significantly different. A small P-value indicates that the sample data is unlikely to happen by random chance. We will be reject the null hypothesis that the population means are the same. We are also rejecting that the categorical and quantitative variables are not related and supporting the alternative hypothesis that the variables are related.

- In a fraction, when the numerator is the same of smaller than the denominator, the overall fraction is small. So if the variance between the groups is much smaller than the variance within the groups, this will give a small F-test statistic (and a large P-value) and indicates that the sample means are not significantly different. A large P-value indicates that the sample data could have happened by random chance. We will fail to reject the null hypothesis that the population means are the same. In other words, the population means might be the same and the categorical and quantitative variables are probably not related.

- The F-test statistic can also be used in a two-population variance or two-population standard deviation hypothesis test. In that case, it compares the variance from two populations.

Here is the summary table from last chapter to remind you of the key decisions in a hypothesis test.

		Significant Test Statistic		Test Statistic NOT Significant
		(Test Statistic falls in tail determined by the critical value or values)		*(Test Statistic does NOT fall in tail determined by the critical value or values)*
		OR		**OR**
		Small P-value		**Large P-value**
		(P-value ≤ significance level)		*(P-value > significance level)*
		OR		**OR**
		Sample Data in Tail		**Sample Data NOT in Tail**
		(when simulating the Null Hypothesis)		*(when simulating the Null Hypothesis)*
Is the sample data significantly		Yes. Significantly different		Not Significantly different
different than H_0?				
Could the sample data happen		Unlikely		Could happen
by random chance				
(sampling variability) if H_0 is true?				
Reject H_0 or		Reject H_0		Fail to Reject H_0
Fail to Reject H_0?				
Is there significant		Yes. Is evidence		No evidence
Evidence?				

Assumptions for an ANOVA hypothesis test

- The 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 samples should be independent of each other.
- The sample sizes should be at least 30 or have a nearly normal shape
- No standard deviation for any sample is more than twice as large as any other sample.

Notice that we must have a random or representative sample. As with all mean average hypothesis tests, we require the sample size to be at least 30 or have a normal shape. Data values within the samples and between the samples should be independent of each other. This again is a difficult assumption to assess. If we have a small simple random sample out of a very large population, then the data values are unlikely to be related. ANOVA is based on variance, so variability in the samples is very important. If one sample has a lot more variability than the others do, this can be a problem. Therefore, we want all of our sample standard deviations to be close. An often-used rule is that no sample standard deviation can be more than twice as large as any other can. Notice that to check these assumptions, we need to look at the sample sizes, sample means and sample standard deviations for each of our groups. We should also look at the shape of the samples with histograms or dot plots.

ANOVA Example 1: Mean Average Salaries for people living in five states in Australia.

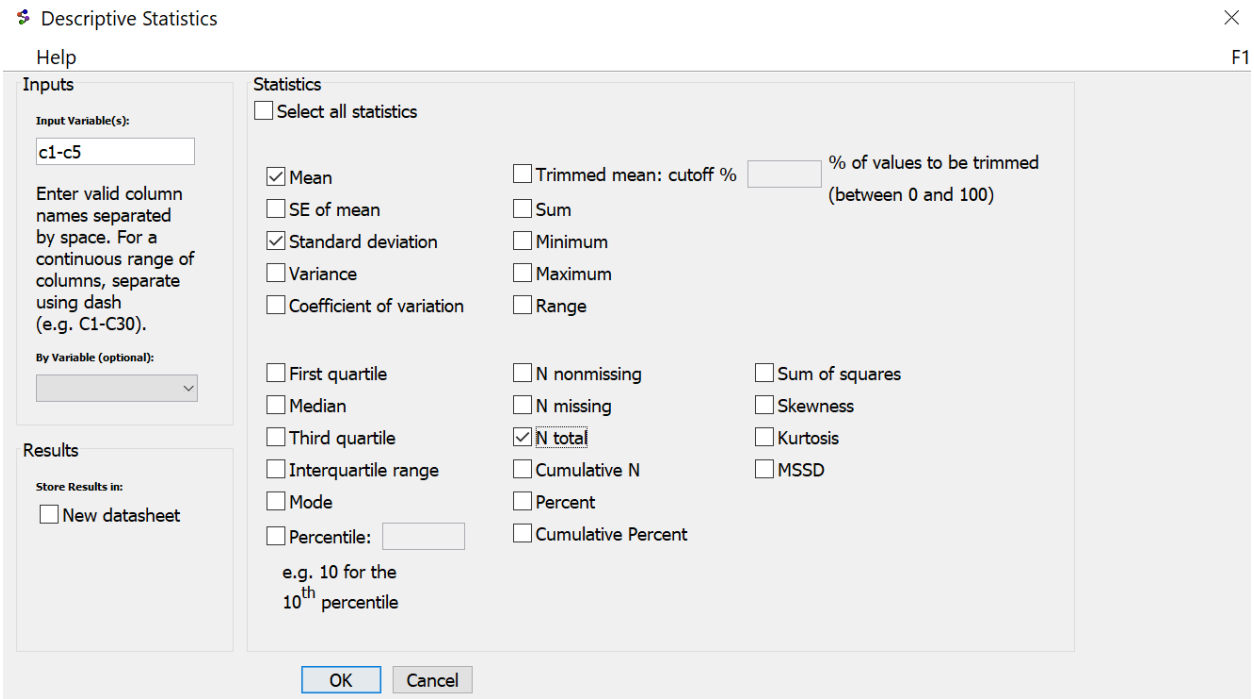
Suppose we want to compare the mean average weekly salary for people living in five states in Australia. The states are Northern Territory, New South Wales, Queensland, Victoria, and Tasmania. We claim that the mean average salary of people is related to where they live. To support this claim, we will need to show that the mean average salaries are different in these states. As with all multiple population hypothesis tests, you should label the populations. To perform this test, adults were randomly selected from each of the five states. We will be using a 5% significance level.

μ_1 : Northern Territory
μ_2 : New South Wales
μ_3 : Queensland
μ_4 : Victoria
μ_5 : Tasmania

Here is the null and alternative hypothesis for the ANOVA test. Remember an ANOVA is a multiple μ test for three or more groups. Notice that if the population mean average salary is the same, then it does not matter which state the person lives in. This implies that the state (categorical variable) is not related to the salary (quantitative variable). If at least one population mean average salary is different, then it does matter which state the person lives in. This implies that the state (categorical variable) is related to the salary (quantitative variable). Again, we see that "not related" is the null hypothesis and "related" is the alternative.

Ho: μ1 = μ2 = μ3 = μ4 = μ5 *(states in Australia are not related to salary)*
Ha: at least one is ≠ (CLAIM) *(states in Australia are related to salary)*

When doing an ANOVA test, it is good to find the sample size (n), the sample mean of each group, and the standard deviation for each group. We also will need to create histograms to check the shape of our samples. Go to www.matt-teachout.org and click on the "statistics" tab and then "data sets". You can either open the Australia Salary data Statcato file in Statcato or open the excel file and copy and paste it into Statcato. The adults in this sample data were randomly selected. To calculate the sample sizes, means and standard deviations, go to the "statistics" menu in Statcato, then click on "basic statistics" then "descriptive statistics". To create histograms, go the "graph" menu and click on histogram. In small data sets like these, I prefer three bins (bars). If makes it easier to see the shape. In addition, if you click on "Show Legend" the computer will also make a title for the graph. Here is the sample statistics and graphs from Statcato.



Descriptive Statistics

Variable	Mean	Standard Deviation
C1 North Territory	1534.540	701.525
C2 New South Wales	1536.823	677.140
C3 Queensland	1368.291	536.319
C4 Victoria	1149.050	516.553
C5 Tasmania	898.695	386.354

Variable	N total
C1 North Territory	35
C2 New South Wales	35
C3 Queensland	35
C4 Victoria	35
C5 Tasmania	35

Histogram

Help

Graph Variables

Graph Variables:

Ctrl-click to select multiple variables

C1 North T
C2 New Sc

Grouped By Categories in: [optional]

Heights of bars represent:

(•) Frequency

( ) Relative Frequency

X-Axis

X-axis (horizontal)

(•) Provide the number of classes, minimum, and maximum

Class width = (maximum - minimum) / classes

Number of bins (classes): 3

Minimum: [ ] Maximum: [ ] [automatic if left blank]

( ) Provide the class width and the minimum

Class width: [ ]

Minimum: [ ] [automatic if left blank]

Label: [ ]

Position of tick marks: ( ) Center of bar (•) Between bars

Other Options

Plot

Title: Histogram

[✓] Show Legend

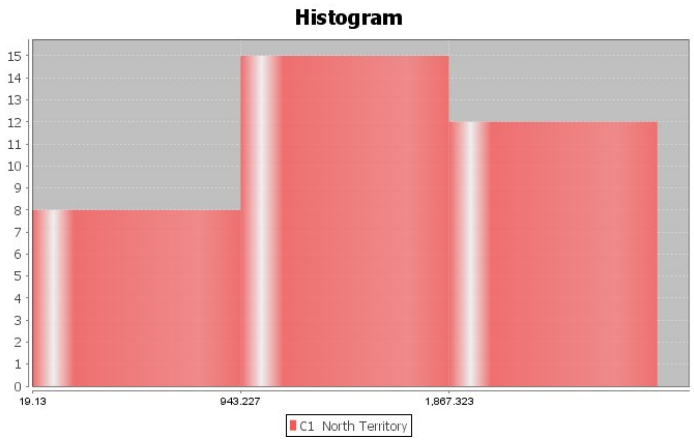
Y-Axis

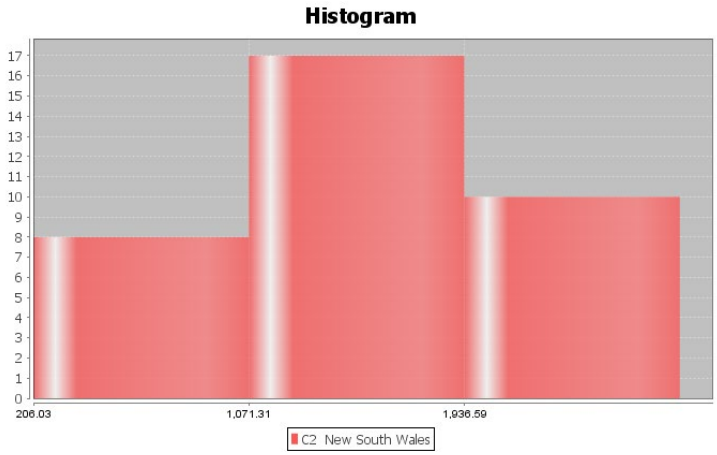
Y-axis (vertical)

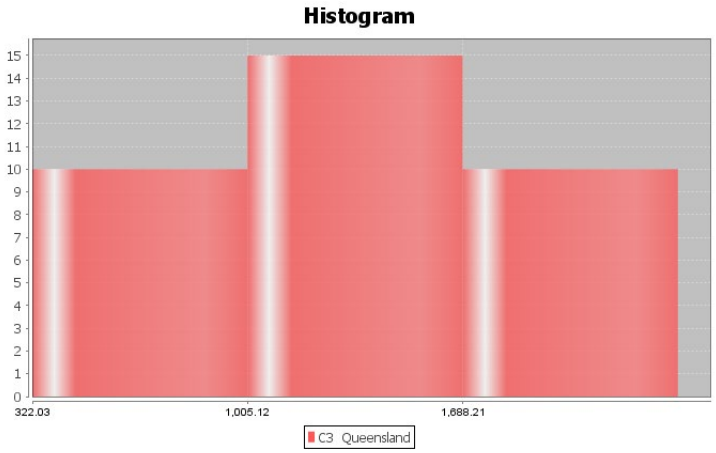
Label: [ ]

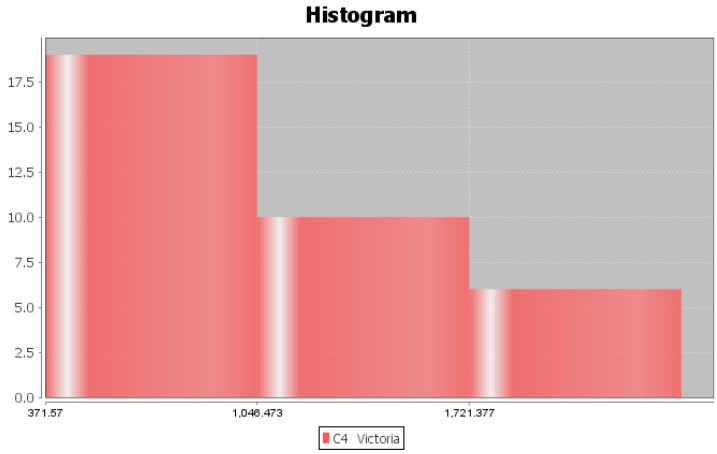
Tick mark units: [ ] automatic if left blank

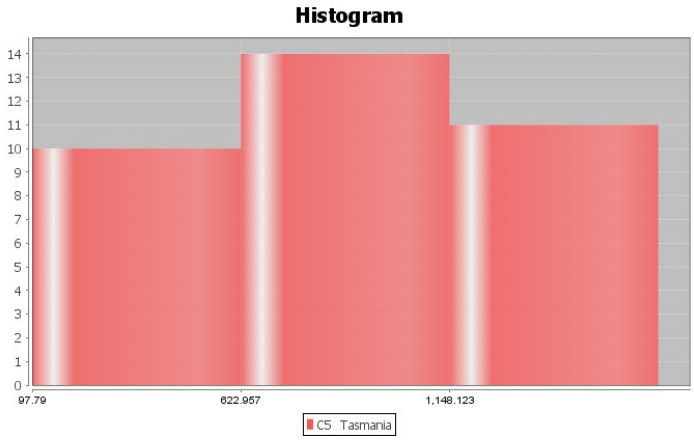
OK Cancel











Assumptions: Notice that this data passes all of the assumptions for the ANOVA hypothesis test.

1. The sample data should be random or representative of the population. Yes. The sample data sets were collected randomly.

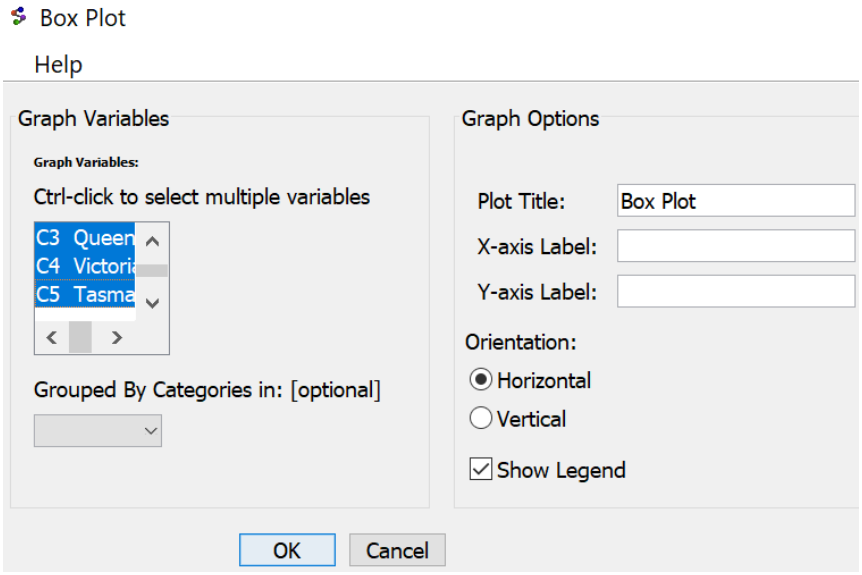
2. Each sample should have a sample size of at least 30 or be nearly normal. Yes. All of the samples had a nearly normal shape except for the data from Victoria, which was skewed right. The sample size for all of the samples was 35. So even though data from Victoria was skewed right, its sample size was still over 30. All of the other samples sizes were over 30 and normal.

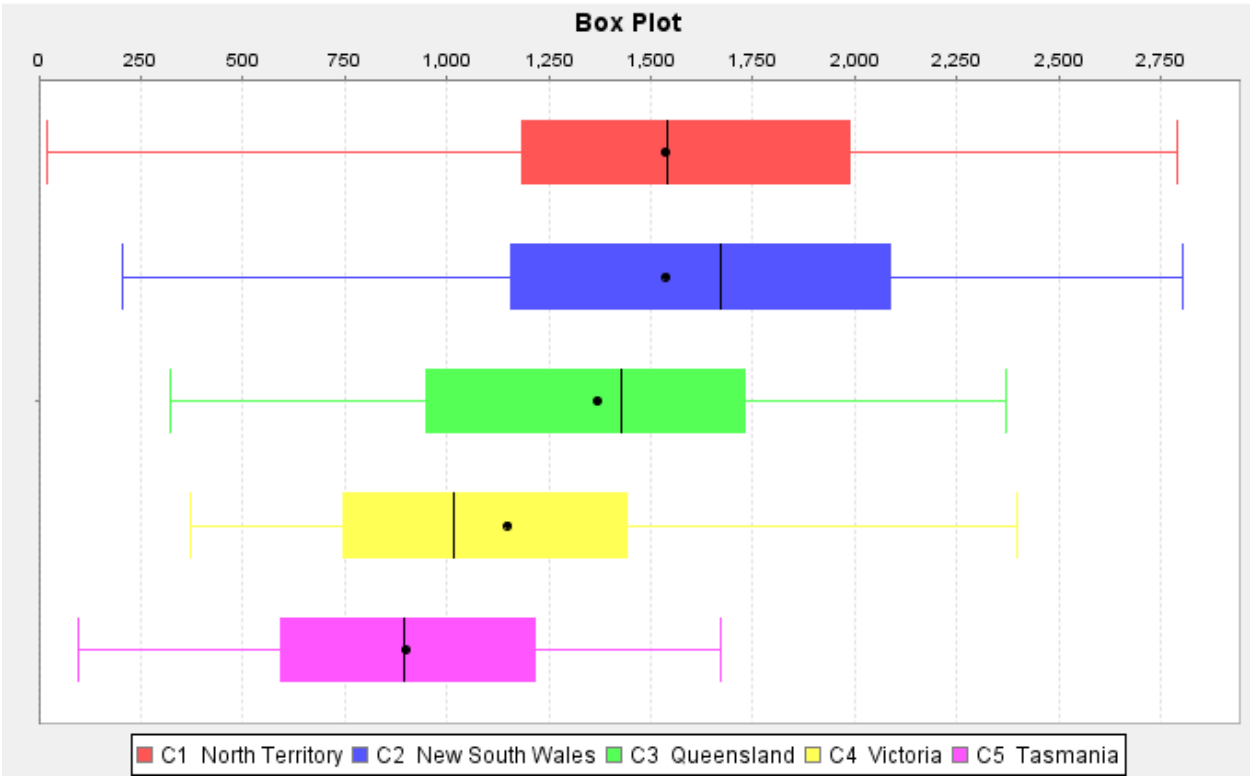
3. Data values within the samples and between the samples should be independent of each other. Yes. Since we are dealing with small random samples out of millions in the populations, it is unlikely that these data values are related.

4. The sample standard deviations for the groups should be close. No standard deviation should be more than twice as large as any other should. Yes. The sample standard deviations are close. No sample standard deviation is more than twice as large as any other sample standard deviation. Notice that the smallest standard deviation was 386.3 and the largest was 701.5 and all of the others are in between.

Some data scientists like to create a side-by-side boxplot when performing an ANOVA test. This is surprising since the ANOVA test looks at means and standard deviations for center and spread, yet box plots look at the median and interquartile range (IQR). The boxplot can still show us general tendencies about shape, center and spread.

In Statcato, go to the "Graph" menu and click on "Box Plot". Hold the control key down and highlight all five of the data sets. You can create a vertical or horizontal box plot. "Show Legend" will create a title. Do NOT click on the "Group By" button. Here is the box plot from Statcato.





This graph tells a lot. We can see that the centers are quite different. The length of the box is a measure of spread (IQR). The lengths of the boxes are all pretty similar. If one box was more than twice as long as another was, this might indicate that one group has a lot more variability than another does. We can also get a sense of the shapes of these data sets, though separate histograms are better. So this side-by-side boxplot shows us that the variability is similar in the groups but the centers are quite different. Only the data from Victoria looks skewed right.

The key question: Are these sample means different because of sampling variability (random chance) OR are they different because at least one of the populations really is different?

To answer this, we need the F test statistic and a P-value.

How to do an ANOVA test with Statcato

Copy and paste your raw quantitative data from each group into some columns in Statcato.

To calculate the F-test statistic and P-value, go to the "statistics" menu, then "Analysis of Variance", then "One-Way ANOVA".

Statistics ➔ Analysis of Variance ➔ One-Way ANOVA

Hold the control key down to select the columns where your data is and push "add to list". Select your significance level and push "OK". Here is the printout we got. Notice this is the same printout we were looking at before.

One-way ANOVA: Significance level = 0.05
Selected column variables: C1 North Territory C2 New South Wales C3 Queensland C4 Victoria C5 Tasmania

Source of Variation	DOF	SS	MS	Test statistic F	Critical value F	p-Value
Treatment (Between Groups)	4	10484529.98982	2621132.49746	7.92175	2.4248	$7.03917 \cdot 10^{-6}$
Error (Within Groups)	170	56249274.83547	330878.08727			
Total	174	66733804.82529				

Let us see if we understand what we are seeing. Notice the MS (variance) is the sum of squares (SS) divided by degrees of freedom (DOF).

MS (Treatment) is the variance between the groups (2621124.8)

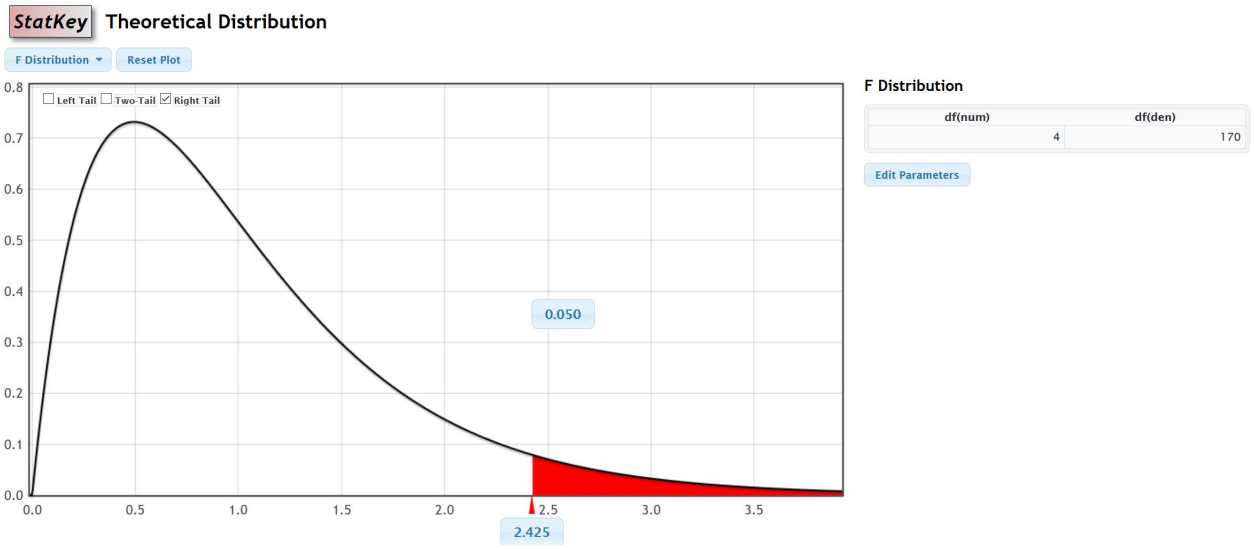
MS (Error) is the variance within the groups (330878.43)

So the F-test statistic is calculated by the formula

F test statistic = $\frac{Variance\ Between\ the\ Groups}{Variance\ Within\ the\ Groups} = \frac{2621124.8}{330878.08727} = 7.92175$

So the variance between the groups is almost 8 times greater than the variance within the groups. Is this significantly large for an F?

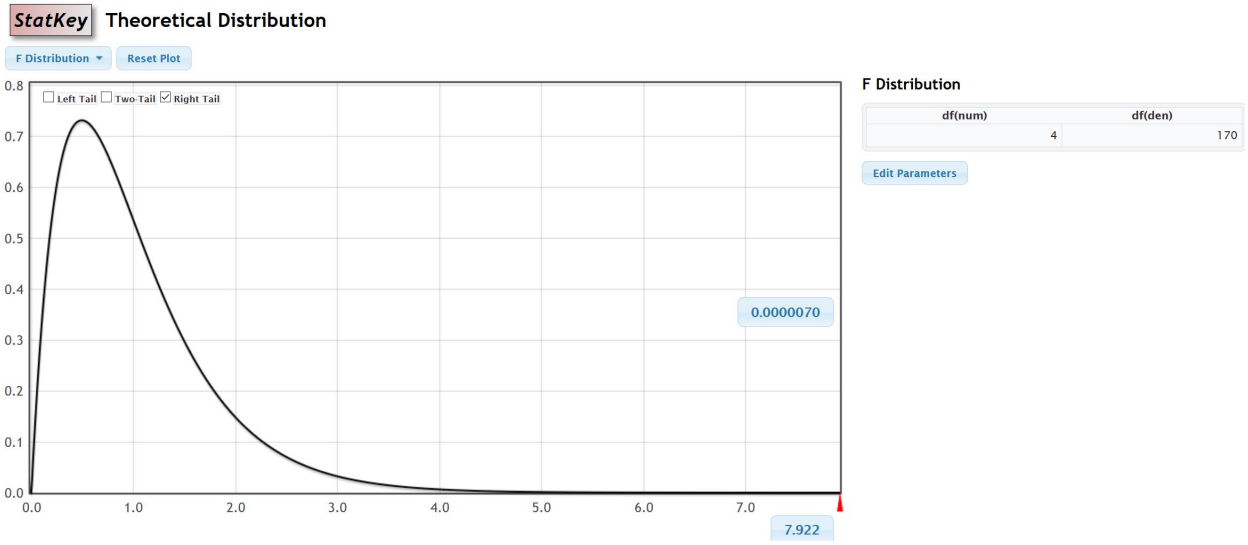
Notice Statcato has calculated a critical value to compare the test statistic to. Remember the test statistic needs to fall in the tail determined by the critical value to be significant. ANOVA is always a right tailed test. Look at the following picture created by StatKey. We see that our test statistic is 7.9217 falls in the right tail. So the F test statistic is significant. This also tells us that the sample data significantly disagrees with the null hypothesis and that the variance between the groups is significantly greater than the variance within the groups. Otherwise, the F test statistic would not have fallen in the right tail.



↑

F = 7.92175

Notice we can also use the same F theoretical curve to calculate the P-value with StatKey. Remember the numerator degrees of freedom is 4 and the denominator degrees of freedom is 170. Now just put the F test statistic in the bottom box in the right tail. The proportion is the P-value. Notice the P-value calculated by StatKey is very close to the P-value calculated by Statcato.



The test statistic fell in the tail determined by the critical value, so the sample data does significantly disagree with the null hypothesis.

Notice that in our printout from Statcato, we got the following P-value: “7.039 x 10^-6”. This is scientific notation. Move the decimal six places to the left to get the P-value as a decimal.

P-value = 0.00000704

The actual P-value is very close to zero and much lower than a 5% significance level. From our study of P-values, we know this is very significant and unlikely to happen by random chance (sampling variability).

Since the P-value is less than our significance level, we should reject the null hypothesis.

Conclusion

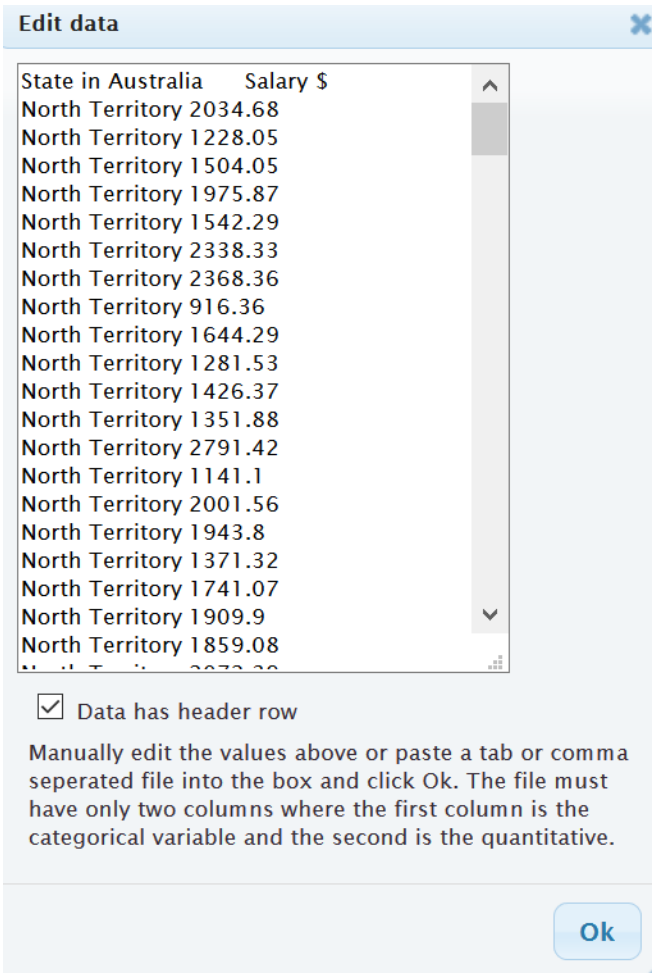
Recall the claim was the alternative hypothesis that where a person lives is related to the salary. To show this we needed to have evidence to support that at least one state was different from the others (alternative hypothesis). Since we rejected the null, we support this claim. Our P-value is very small and our F test statistic very large, so we have significant evidence.

Conclusion: There is significant evidence to support the claim that the mean average salaries of people in Northern Territory, New South Wales, Queensland, Victoria, and Tasmania are different and that the state a person lives in is related to the salary.

Note: Remember "relationship" does not mean "causation" though. This was not an experiment and did not control confounding variables. There are many reasons why a persons' salary is high or low. It would be wrong to say that the place a person lives causes their salary to be low or high.

Simulation

Remember we can also estimate the P-value and determine significance with randomized simulation. Go to www.lock5stat.com and open StatKey. We will need to go to www.matt-teachout.org and open the "Australia Salary Data" in Excel. In Statcato, we needed the quantitative data separated by group, but in StatKey, we need the raw categorical and quantitative data. StatKey will separate the data. In the excel spreadsheet you will see the column that says, "State in Australia" and "Salary". Copy these two data sets together. Under the "More Advanced Randomization Tests" menu click on "ANOVA for Difference in Means". Click on "Edit Data" and paste in the state and salary columns.



Notice under “Original Sample”, StatKey has calculated the F-test statistic for you along with the sample means, sample sizes and sample standard deviations. If you wish to see the variance between and the variance within calculations click on “ANOVA Table”. It looks similar to the Statcato printout.

Original Sample ANOVA Table

$n = 175, F = 7.922$

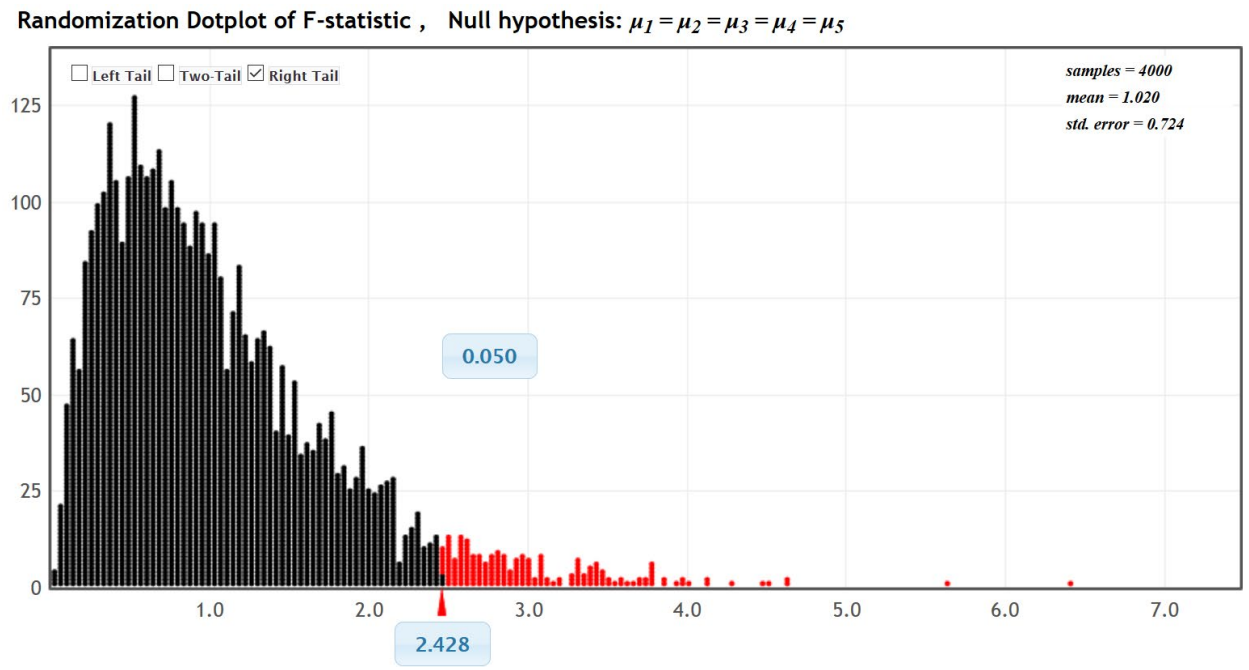
Statistics	North Territory	New South Wales	Queensland	Victoria	Tasmania	Overall
Sample Size	35	35	35	35	35	175
Mean	1534.5	1536.8	1368.3	1149.1	898.7	1297.5
Standard Deviation	701.5	677.1	536.3	516.6	386.4	619.3

ANOVA Table

	df	SS	MS	F
Groups	4	10484530.0	2621132.5	7.922
Error	170	56249274.8	330878.1	
Total	174	66733804.8		

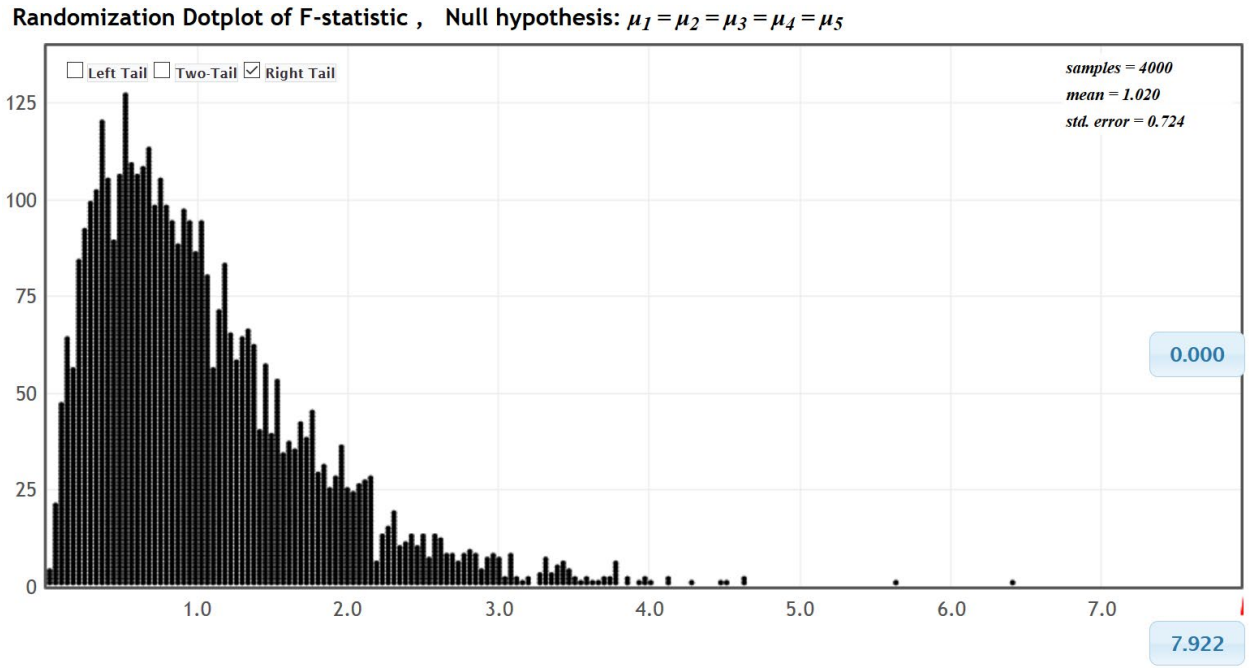
Notice the null hypothesis is that all five population means are equal. To simulate the null hypothesis click "Generate 1000 Samples" a few times. In simulations with only one or two groups, we usually use the sample mean, sample mean difference, sample proportion, or sample proportion difference. In tests with more than two groups, we cannot use that approach. When a test involves three or more groups, we will resort to using the test statistic to summarize the sample data.

In this simulation, the computer has randomly collected thousands of samples and calculated thousands of F test statistics. Remember the real test statistic can be found under "Original Sample". ANOVA is a right tailed test so we will click on "Right-Tail". If we put in the 5% significance level in the proportion box in the right tail, we will have the critical value. Because of sampling variability, you will get slightly different answers, but this simulation gave a critical value of 2.428, which is not far from the theoretical critical value calculated by Statcato earlier. We can now use this graph to determine if the test statistic falls in the tail. Notice our F-test statistic of 7.922 does fall in the tail, so our sample data significantly disagrees with the null hypothesis and our variance between the groups is significantly higher than the variance within the groups.



↑
F =7.922

We can also calculate the P-value by putting the test statistic in the bottom box of the simulation. Notice our P-value came out to be about zero. So this sample data is unlikely to occur because of sampling variability if the null hypothesis was true. We would reject the null hypothesis and get the same conclusion as we did with the traditional approach.



Problems Section 4B

(#1-10) Use each of the following ANOVA F-test statistics and the corresponding critical values to fill out the table.

	F-test stat	Sentence to explain F-test statistic.	Critical Value	Does the F-test statistic fall in a tail determined by the critical value? (Yes or No)	Does sample data significantly disagree with H_0?
1.	+5.573		+2.886		
2.	+1.192		+3.113		
3.	+0.664		+2.949		
4.	+4.415		+3.125		
5.	+3.718		+4.117		
6.	+0.991		+2.009		
7.	+2.652		+1.875		
8.	+1.585		+3.225		
9.	+2.447		+2.798		
10.	+8.133		+2.891		

(#11-20) Use each of the following P-values and corresponding significance levels to fill out the table.

	P-value Proportion	P-value %	Sentence to explain the P-value	Significance Level %	Significance level Proportion	If H_0 is true, could the sample data occur by random chance or is it unlikely?	Reject H_0 or Fail to reject H_0?
11.	0.186			10%			
12.	0.0042			1%			
13.	2.59×10^{-4}			5%			
14.	0.006			1%			
15.	0.353			5%			
16.	0			10%			
17.	0.041			5%			
18.	0.274			10%			
19.	1.04×10^{-8}			1%			
20.	0.067			5%			

21. The F-test statistic compares the variance between the groups to the variance within the groups. Explain how the variance between the groups is calculated and what it tells us. Explain how the variance within the groups is calculated and what it tells us. How can we use the variance between and the variance within to calculate the F-test statistic?

22. If the variance between the groups were significantly larger than the variance within, would the F-test statistic be large or small? Explain why.

23. If the variance between the groups were about the same as the variance within, would the F-test statistic be large or small? Explain why.

24. The ANOVA printout involves the degrees of freedom within the groups, the degrees of freedom between the groups and the total degrees of freedom. How are the different degrees of freedom calculated?

(#25-28) Directions: Use the following Statcato statistics, graphs and ANOVA printout to test the population claims. For each of the following problems answer the following.

a) Give the null and alternative hypothesis.
b) Check the assumptions for a One-Way ANOVA test.
c) Write a sentence to explain the F test statistic.
d) Use the F test statistic and Critical Value to determine if the sample data significantly disagrees with the null hypothesis. Explain your answer.
e) Use the P-value and Significance Level to answer the following: Could the sample data or more extreme have occurred because of sampling variability or is it unlikely that the sample data occurred because of sampling variability? Explain your answer.
f) Should we reject the null hypothesis or fail to reject the null hypothesis? Explain your answer.
g) Write a conclusion for the hypothesis test addressing evidence and the claim.
h) What is the variance between the groups? What is the variance within the groups? Was the variance between significantly higher than the variance within? Explain how you know.
i) Was the categorical and quantitative variables related or not. Explain your answer.

25. A random sample of black bears were weighed at various times of the year. Some of the bears were weighed in the spring, some in the summer and some in the fall. The bears were tagged so that the same bear was not measured more than once. Use a 1% significance level and the following Statcato statistics, graphs and ANOVA printout to test the population claim that the time of year (season) is related to the weight of the bears.

One-way ANOVA: Significance level = 0.01

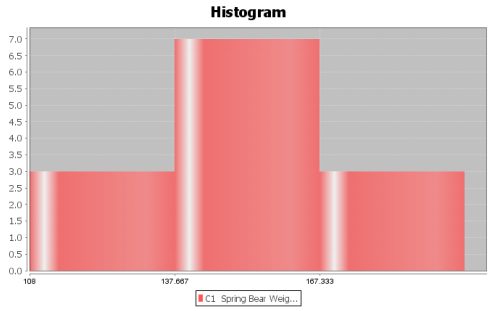
Selected column variables: C1 Spring Bear Weig... C2 Summer Bear Weig... C3 Fall Bear Weight...

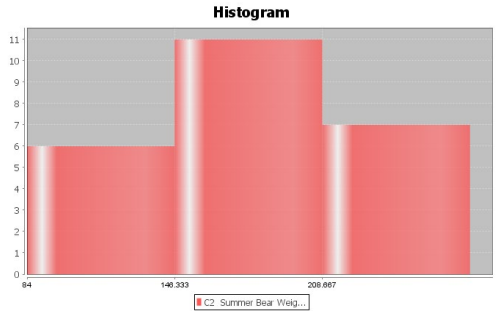
Source of Variation	DOF	SS	MS	Test statistic F	Critical value F	p-Value
Treatment (Between Groups)	2	45539.29263	22769.64632	13.55345	5.0472	0.00002
Error (Within Groups)	51	85679.46663	1679.98954			
Total	53	131218.75926				

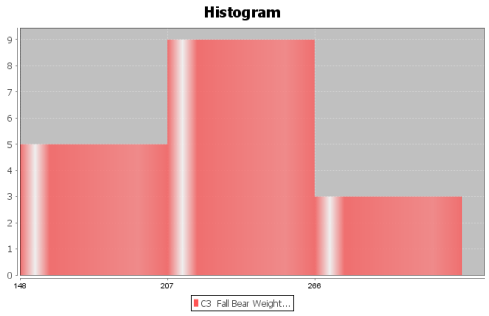
Descriptive Statistics

Variable	Mean	Standard Deviation
C1 Spring Bear Weights in Pounds	151.385	22.463
C2 Summer Bear Weights in Pounds	182.125	48.017
C3 Fall Bear Weights in Pounds	228.118	40.769

Variable	N total
C1 Spring Bear Weights in Pounds	13
C2 Summer Bear Weights in Pounds	24
C3 Fall Bear Weights in Pounds	17







26. A census of Math 075 pre-stat students was taken in the fall 2015 semester. The students were separated into three sleep groups: low amount of sleep, moderate amount of sleep, high amount of sleep. They were also asked how many total units they have completed at the college. Though the data was not random, you can assume it was representative of Math 075 students at COC. Use a 10% significance level and the following Statcato statistics, graphs and ANOVA printout to test the claim that sleep is not related the total number of units completed.

One-way ANOVA: Significance level = 0.1

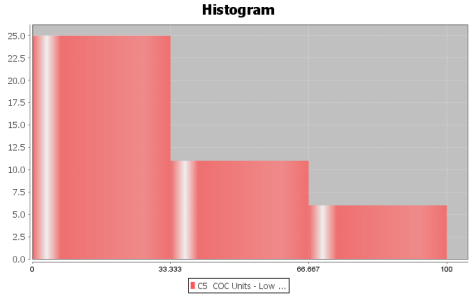
Selected column variables: C5 COC Units - Low ... C6 COC Units - Medi... C7 COC Units - High...

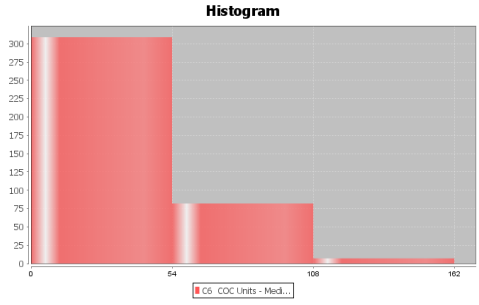
Source of Variation	DOF	SS	MS	Test statistic F	Critical value F	p-Value
Treatment (Between Groups)	2	2822.35625	1411.17813	1.83387	2.3133	0.16087
Error (Within Groups)	497	382446.38503	769.50983			
Total	499	385268.74128				

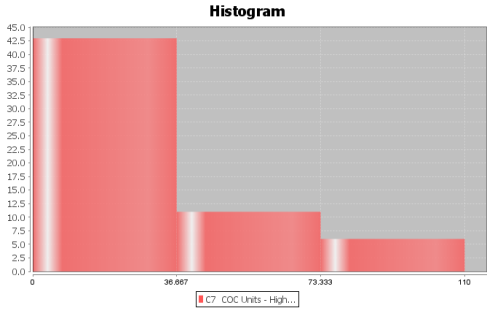
Descriptive Statistics

Variable	Mean	Standard Deviation
C5 COC Units - Low Sleep Group	32.952	28.586
C6 COC Units - Medium Sleep Group	32.990	27.585
C7 COC Units - High Sleep Group	25.675	28.178

Variable	N total
C5 COC Units - Low Sleep Group	42
C6 COC Units - Medium Sleep Group	398
C7 COC Units - High Sleep Group	60







27. A census of Math 075 pre-stat students was taken in the fall 2015 semester. The students were separated into four political parties: democratic, republican, independent party, and other political party. They were also asked number of alcoholic beverages they consume per week. Though the data was not random, you can assume it was representative of Math 075 students at COC. Use a 5% significance level and the following Statcato statistics, graphs and ANOVA printout to test the claim that political party is not related to the number of alcoholic beverages.

One-way ANOVA: Significance level = 0.05

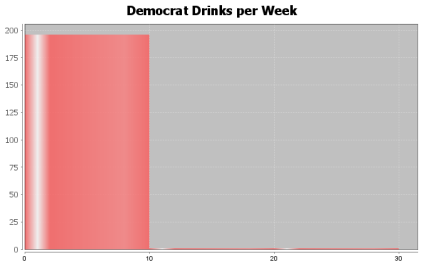
Selected column variables: C9 # Drinks per Wee... C10 # Drinks per Wee... C11 # Drinks per Wee... C12 # Drinks per Wee...

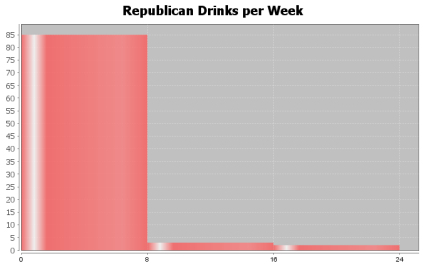
Source of Variation	DOF	SS	MS	Test statistic F	Critical value F	p-Value
Treatment (Between Groups)	3	25.44137	8.48046	0.89597	2.6228	0.44306
Error (Within Groups)	497	4704.16342	9.46512			
Total	500	4729.60479				

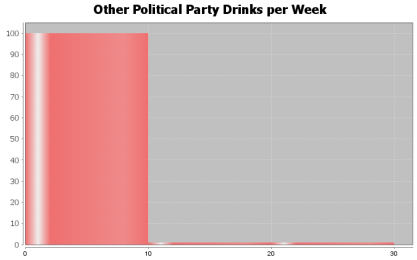
Descriptive Statistics

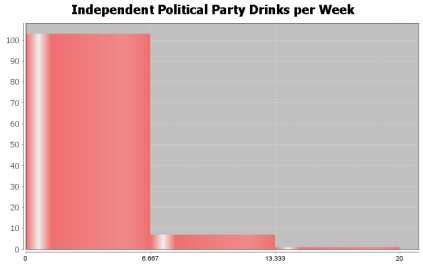
Variable	Mean	Standard Deviation
C9 # Drinks per Week - Democrats	0.914	2.566
C10 # Drinks per Week - Independent Political Party	1.342	2.943
C11 # Drinks per Week - Other Political Party	1.373	3.447
C12 # Drinks per Week - Republicans	1.411	3.753

Variable	N total
C9 # Drinks per Week - Democrats	198
C10 # Drinks per Week - Independent Political Party	111
C11 # Drinks per Week - Other Political Party	102
C12 # Drinks per Week - Republicans	90









28. A census of Math 075 pre-stat students was taken in the fall 2015 semester. The students were asked what their favorite social media is: Facebook, Instagram, Snapchat, or Twitter. They were also asked number minutes per day spent on social media. Though the data was not random, you can assume it was representative of Math 075 students at COC. Use a 5% significance level and the following Statcato statistics, graphs and ANOVA printout to test the claim that the type of social media is related to the number of minutes per day spent on social media.

One-way ANOVA: Significance level = 0.05

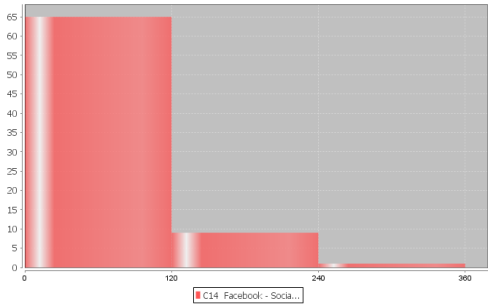
Selected column variables: C14 Facebook - Socia... C15 Instagram - Soci... C16 Snapchat - Socia... C17 Twitter - Social...

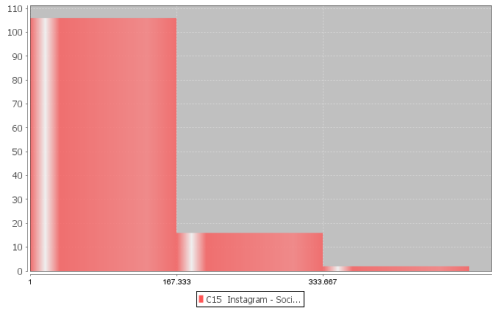
Source of Variation	DOF	SS	MS	Test statistic F	Critical value F	p-Value
Treatment (Between Groups)	3	169375.54058	56458.51353	8.20214	2.6354	0.00003
Error (Within Groups)	293	2016833.12272	6883.38950			
Total	296	2186208.66330				

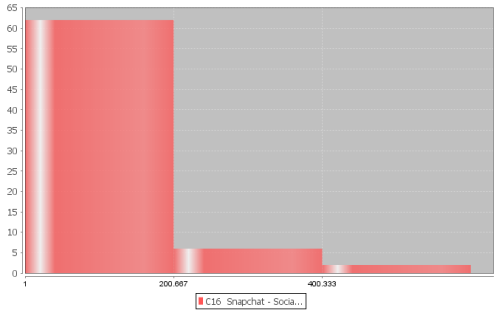
Descriptive Statistics

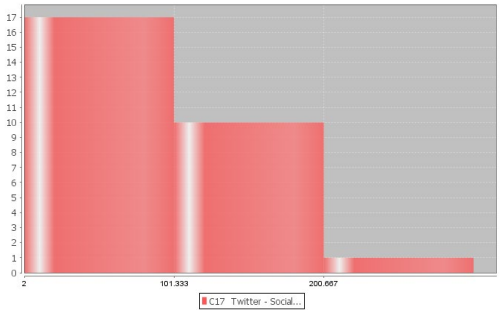
Variable	Mean	Standard Deviation
C14 Facebook - Social Media Minutes per day	43.867	58.103
C15 Instagram - Social Media Minutes per day	83.206	82.817
C16 Snapchat - Social Media Minutes per day	110.914	109.552
C17 Twitter - Social Media Minutes per day	90.964	59.408

Variable	N total
C14 Facebook - Social Media Minutes per day	75
C15 Instagram - Social Media Minutes per day	124
C16 Snapchat - Social Media Minutes per day	70
C17 Twitter - Social Media Minutes per day	28









(#29-33) Directions: Go to www.lock5stat.com and click on the StatKey button. Under the "More advanced randomization tests" menu click on "ANOVA for Difference in Means". For each of the following problems, use a randomized simulation to answer the following. Assume the data met the assumptions for an ANOVA hypothesis test. For each problem, answer the following questions.

a) *Give the null and alternative hypothesis.*
b) *The F-test statistic is given under "Original Sample". Write a sentence to explain the F test statistic.*
c) *Simulate the null hypothesis and put the significance level in the right tail to calculate the critical value. What was the critical value? (Answers will vary.)*
d) *Use the F test statistic and Critical Value to determine if the sample data significantly disagrees with the null hypothesis. Explain your answer.*
f) *Put in the test statistic into the right tail to calculate the P-value. What was the P-value? (Answers will vary.)*
g) *Use the P-value and Significance Level to answer the following: Could the sample data or more extreme have occurred because of sampling variability or is it unlikely that the sample data occurred because of sampling variability? Explain your answer.*
f) *Should we reject the null hypothesis or fail to reject the null hypothesis? Explain your answer.*
g) *Write a conclusion for the hypothesis test addressing evidence and the claim.*
h) *What is the variance between the groups? What is the variance within the groups? Was the variance between significantly higher than the variance within? Explain how you know.*
i) *Was the categorical and quantitative variables related or not. Explain your answer.*

29. Use the random car data and a 1% significance level to test the claim that the country a car is from is related to its gas mileage. Go to www.matt-teachout.org and open the random car data. Copy and paste the country and the miles per gallon columns next to each other in a new excel spreadsheet. The country should be on the left and the miles per gallon should be on the right. Then copy both columns together. Go to www.lock5stat.com and click on the StatKey button. Under the "More advanced randomization tests" menu click on "ANOVA for Difference in Means". Click on the "Edit Data" button and paste the country and mpg columns into StatKey. Click on "Generate 1000 Samples" a few times and then "Right-Tail". Put in the original sample F-test statistic in the bottom box to estimate the P-value. Complete the questions above.

30. Under the "ANOVA for Difference in Means" menu in StatKey, click on the button at the top left of the page and click on "Sandwich Ants". We are studying the number of ants that are drawn to different kinds of food. In this data, we are looking at the mean average number of ants that come to three different types of sandwiches left out to spoil. Use a 5% significance level to test the claim that the number of ants is not related to the type of sandwich.

31. Use the random car data and a 10% significance level to test the claim that the country a car is from is not related to its horsepower. Go to www.matt-teachout.org and open the random car data. Copy and paste the country and the horsepower columns next to each other in a new excel spreadsheet. The country should be on the left and the horsepower should be on the right. Then copy both columns together. Go to www.lock5stat.com and click on the StatKey button. Under the "More advanced randomization tests" menu click on "ANOVA for Difference in Means". Click on the "Edit Data" button and paste the country and horsepower columns into StatKey. Click on "Generate 1000 Samples" a few times and then "Right-Tail". Put in the original sample F-test statistic in the bottom box to estimate the P-value. Complete the questions above.

32. Under the "ANOVA for Difference in Means" menu in StatKey, click on the pulse rate and award data. This data looks at the average pulse rates of those people that have won Olympic, Academy and Nobel awards. Use a 1% significance level to test the claim that the population mean average pulse rate is related to the type of award the person won.

33. Under the "ANOVA for Difference in Means" menu in StatKey, click on the Homes for Sale (price by state) data. This data looks at the average selling price of homes in four different states. Use a 10% significance level to test the claim that the population mean average home price is related to the state the home is sold in.

Section 4C – Proportion Relationships: Two-population Proportion Test

Sometimes we wish to determine if a specific percentage from categorical data is related to various groups (populations). If we only have two populations, we can use a two-population proportion hypothesis test with a Z-score test statistic. If we have three or more populations, we will need to use a more advanced test statistic called the chi-squared test statistic. This is sometimes called a "Goodness of Fit" test. The key idea is to ask the question if the population percentage is the same in the various groups or is it significantly different.

Two-Population Proportion Test for Proportion Relationships

There are different ways of writing the null and alternative hypothesis. A population proportion can be described with the Greek letter pi (π) or with a "p". Remember equal proportions goes with the null hypothesis of "not related" while any difference between the proportions indicates a relationship.

H_0: $p_1 = p_2$ The population % is NOT related to a categorical variable (% is not related to the groups)
H_A: $p_1 \neq p_2$ The population % is related to a categorical variable (% is related to the groups)

As we learned in the last chapter, the alternative hypothesis determines the type of test. If the alternative hypothesis is greater than (>) it is a right-tailed test. If the alternative hypothesis is less than (<) it is a left-tailed test. If the alternative hypothesis is not equal ($\neq$) it is a two-tailed test. While some prefer to use ≥ or ≤ for the null hypothesis, I prefer not to because of relationship implications.

Two-Tailed Null and Alternative Hypothesis

H_0: $p_1 = p_2$ The population % is NOT related to a categorical variable (% is not related to the groups)
H_A: $p_1 \neq p_2$ The population % is related to a categorical variable (% is related to the groups)

OR

H_0: $\pi_1 = \pi_2$ The population % is NOT related to a categorical variable (% is not related to the groups)
H_A: $\pi_1 \neq \pi_2$ The population % is related to a categorical variable (% is related to the groups)

Right-Tailed Null and Alternative Hypothesis

H_0: $p_1 = p_2$ The population % is NOT related to a categorical variable (% is not related to the groups)
H_A: $p_1 > p_2$ The population % is related to a categorical variable (% is related to the groups)

OR

H_0: $\pi_1 = \pi_2$ The population % is NOT related to a categorical variable (% is not related to the groups)
H_A: $\pi_1 > \pi_2$ The population % is related to a categorical variable (% is related to the groups)

Left-Tailed Null and Alternative Hypothesis

H_0: $p_1 = p_2$ The population % is NOT related to a categorical variable (% is not related to the groups)
H_A: $p_1 < p_2$ The population % is related to a categorical variable (% is related to the groups)

OR

H_0: $\pi_1 = \pi_2$ The population % is NOT related to a categorical variable (% is not related to the groups)
H_A: $\pi_1 < \pi_2$ The population % is related to a categorical variable (% is related to the groups)

Assumptions

It is very important to always check the assumptions for a hypothesis test in order to make sure that our sample data is as unbiased as possible. Remember that biased data may lead to a wrong conclusion (type 1 or type 2 error). Since we are now using this test to determine relationships, we may also need to prove cause and effect. If that is the case, we will need to use random assignment instead of a random sample.

Assumptions for a Two-population Proportion Test for Relationship

1. Random: The sample categorical data either should be a random sample *(if proving there is relationship)* or have used random assignment *(if proving cause and effect)*.

2. Large sample size: The sample categorical data should have at least ten success $(x \geq 10)$ and at least ten failures $(n - x \geq 10)$. *For example, there should be at least 10 people with congestive heart failure (CHF) in the sample from the U.S. and at least 10 people without CHF in the sample from the U.S. There should also be at least 10 people with CHF in the sample from the Australia and at least 10 people without CHF in the sample from Australia.*

3. Data values within each sample and between the samples should be independent of each other. If the data was collected from one sample then the assumption is just that data values within the sample should be independent. If the data was collected from more than one sample, then the data values between the samples should also be checked for independence. *For example, we should not have people in our samples that are family members or the same people measured twice. The sample from the U.S. should not be connected to the sample from Australia. For example, the congestive heart failure (CHF) data should not come from a company that has hospitals in both countries. In an experiment, we should not control confounding variables by using the same group of people measured multiple times. This would fail the independent individuals' assumption. Random assignment is a better option for controlling confounding variables.*

Note: Some statisticians like to use the chi-squared test statistic even if there are only two populations of interest. If that is the case, use the assumptions for the goodness of fit test.

Z-test statistic for two-population proportion tests

The Z-test statistic measures the number of standard errors that the sample proportion from group 1 $(\hat{p}_1)$ is above or below the sample proportion from group 2 $(\hat{p}_2)$. It is "above" when the Z-test statistic is positive and "below" when the Z-test statistic is negative. It can also be thought of as the number of standard errors that the difference between the sample proportions $(\hat{p}_1 - \hat{p}_2)$ is from zero or some other claimed difference. Z-scores usually are significant around two standard errors, but it is always good to refer to the critical value or P-value when judging significance. The formula below seems daunting to calculate. Remember, no one in data science calculates this by hand with a calculator. Always use a computer program like R or Statcato. In the two-population proportion Z test, we often use pooling ($\bar{p}$). Pooling the proportions is combines the two data sets together before calculating the proportion. We need to assume that the population proportions are equal in the null hypothesis in order to pool. For this reason, pooling is usually used for a two-population proportion hypothesis test and is not used in a two-population proportion confidence interval.

p-pooled ($\bar{p}$) = $\frac{(x_1+x_2)}{(n_1+n_2)}$

Z-test statistic for two population proportion *(pooled)* = $\frac{(sample\ 1 - sample\ 2)}{standard\ error} = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\left(\frac{\bar{p}(1-\bar{p})}{n_1} + \frac{\bar{p}(1-\bar{p})}{n_2}\right)}}$

While this formula looks daunting, it is only counting how many standard errors the sample proportion for group 1 is above or below group 2. The most important thing is not calculating. It is interpreting and explaining the test statistic.

Z-test statistic for two-population sentence: The sample proportion for group 1 is # of standard errors (above or below) the sample proportion for group 2.

Look at the following two-population proportion printout.

Significance Level	Critical Value	Test Statistic Z	p-Value
0.05	-1.96, 1.96	-0.412	0.6800

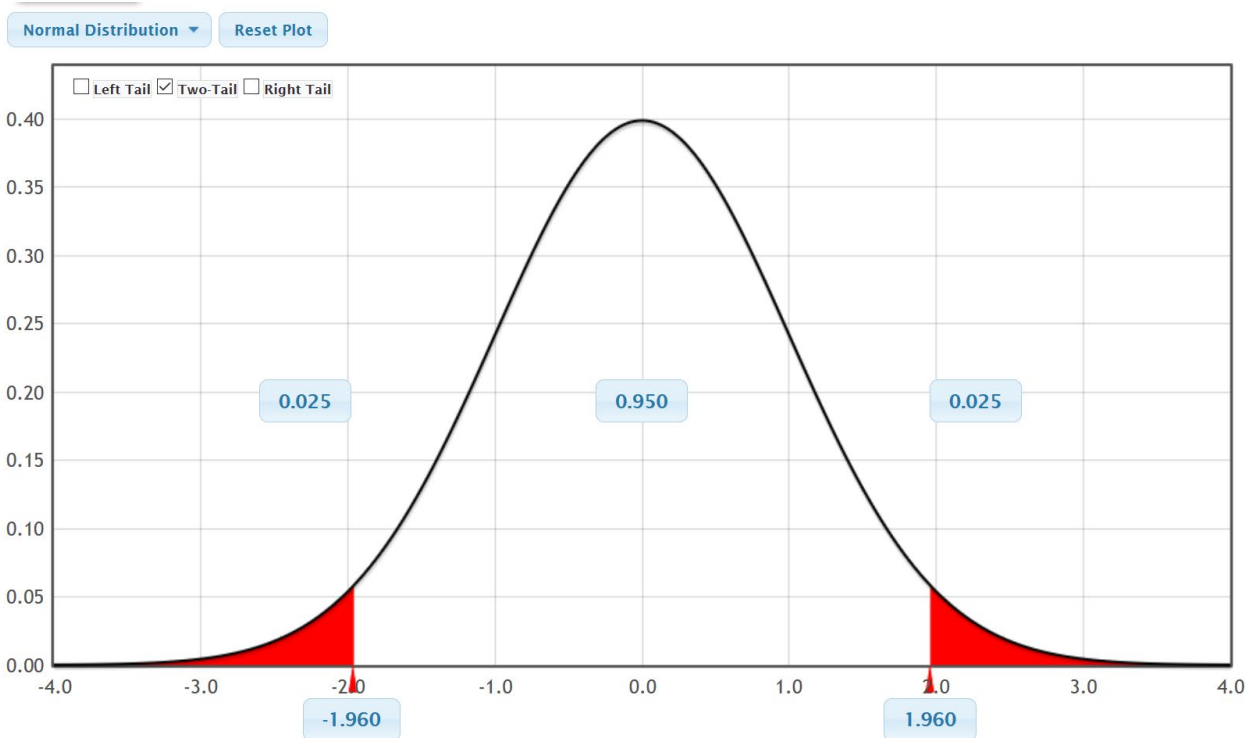
The Z-score test statistic is *−0.412*. Since it is negative, we know that the sample proportion for group 1 is lower than the sample proportion for group 2.

Z-test statistic sentence: The sample proportion for group 1 is 0.412 standard errors below the sample proportion for group 2. *(Notice we did <u>not</u> say "−0.412 standard errors". A Z-score of −0.412 means "0.412 standard errors below".)*

Test statistics tell us if the sample data significantly disagrees with the null hypothesis. Remember the following rules.

- If the test statistic falls in one of the tails determined by the critical value or values, then the sample data significantly disagrees with the null hypothesis.
- If the test statistic falls does NOT fall in one of the tails determined by the critical value or values, then the sample data does NOT significantly disagree with the null hypothesis.

In the Statcato printout, the critical values are ± 1.96. So the Z-test statistic does not fall in either tail. The sample data does not significantly disagree with the null hypothesis.



↑
Z = −0.412

P-value

We also learned in the last chapter, that it is vital to know if the sample data occurred because of sampling variability (random chance). Remember, sample data always disagrees with the null hypothesis to some extent. The key is to determine why it is different. There are two reasons why the sample data disagrees with the null hypothesis. Maybe the null hypothesis is correct and the sample data disagrees because of sample data is always different (random chance). Another option is that the sample data disagrees because the null hypothesis is wrong. The key is that if you determine that it was not random chance, the only other option is that the null hypothesis is wrong. P-value is the key to making this decision about whether the data occurred by random chance. If the P-value is low (close to zero) then it is unlikely to be random chance. If the P-value is high, there is a possibility of the sample data occurring because of random chance.

- If P-value ≤ significance level (α), then the sample data is unlikely to have occured by random chance. Since sampling variability is ruled out, the null hypothesis must be wrong. So we "reject the null hypothesis". A low P-value also indicates that the sample data significantly disagrees with the null hypothesis.
- If P-value > significance level (α), then the sample data is could have occurred by random chance. Since we do not know if sampling variability is involved or not, we also do not know if the null hypothesis is right or wrong. So we say we "fail to reject the null hypothesis" in this case. A high P-value also indicates that the sample data does not significantly disagree with the null hypothesis.

Randomized Simulation

In the last chapter, we saw that P-value could be calculated with randomized simulation or a randomization technique. This is a fabulous way for us to visualize what sampling variability (random chance) looks like if the null hypothesis is true. We have a computer create thousands of random samples under the premise that the null hypothesis is true. These simulated samples have the same sample sizes as the original sample data. If the real original sample data falls in the tail of the simulation it indicates that it is significant. The more in the tail the data is, the smaller the P-value. For a one-population proportion test, we see if the sample proportion is in the tail. For a two-population proportion test, we will see if the difference between the two sample proportions falls in the tail.

- If sample statistic falls is in a tail of the simulation, then the sample data is significant and significantly disagrees with the null hypothesis.
- If the sample statistic does not fall in a tail of the simulation, then the sample data is not significant and does not significantly disagree with the null hypothesis.

Here is a chart from chapter four that summarizes test statistics, P-value and simulation.

		Significant Test Statistic		Test Statistic NOT Significant
		(Test Statistic falls in tail determined by the critical value or values)		*(Test Statistic does NOT fall in tail determined by the critical value or values)*
		OR		**OR**
		Small P-value		**Large P-value**
		(P-value ≤ significance level)		*(P-value > significance level)*
		OR		**OR**
		Sample Data in Tail		**Sample Data NOT in Tail**
		(when simulating the Null Hypothesis)		*(when simulating the Null Hypothesis)*
Is the sample data significantly different than H_0?		Yes. Significantly different		Not Significantly different
Could the sample data happen by random chance (sampling variability) if H_0 is true?		Unlikely		Could happen
Reject H_0 or Fail to Reject H_0?		Reject H_0		Fail to Reject H_0
Is there significant Evidence?		Yes. Is evidence		No evidence

Example (Two-Population Proportion Categorical Relationship Test)

Many high school and college students love to listen to music when they study. Some like to listen to their favorite music, while others just like the background noise. Use a 5% significance level to test the claim that liking the music is related to being able to memorize a large amount of information. A randomized experiment was done to test this claim. A group of college students were randomly assigned into two groups. Both groups had to memorize the same amount of information. The number of students that were able to memorize a significant amount of the information were classified as "high retention". One group listened to their favorite music and the other group had to listen to a type of music they hated. Confounding variables like the room environment and music volume were the same in both groups.

Label your variables.

p_1: The percentage of college students that listen to liked music and can memorize a significant amount of information (high retention).

p_2: The percentage of college students that listen to hated music and can memorize a significant amount of information (high retention).

Here is the sample data.

Liked Music: 25 total people, 10 high retention, $\hat{p}_1 \approx 0.4$
Hated Music: 24 total people, 11 high retention, $\hat{p}_2 \approx 0.458$

Sample Difference: $\hat{p}_1 - \hat{p}_2 \approx 0.4 - 0.458 = -0.058$

H_0: $p_1 = p_2$ (The population % for high retention is NOT related to liking the music)
H_A: $p_1 \neq p_2$ (The population % for high retention is related to liking the music) CLAIM

(Notice this is a two-tailed test.)

Let us check the assumptions.

Is the sample data random or representative? Yes. The data was not a random sample of the population, but it was randomly assigned. So the sample data will not apply to all college students, but it has the capacity to prove cause and effect.

Is there at least 10 success and 10 failures in the sample data? Yes. In the liked music group, there were 10 high retention and 14 not high retention. In the hated music group, there were 11 high retention and 14 not high retention.

Are data values independent? It is difficult to know this without a detailed look at the people in the experiment. We did not have the same people measured twice, but instead used random assignment. We also should not have family members. If some of the students are friends and know each other, they may have similar taste in music. We will assume this data passes this assumption, but it might need further study.

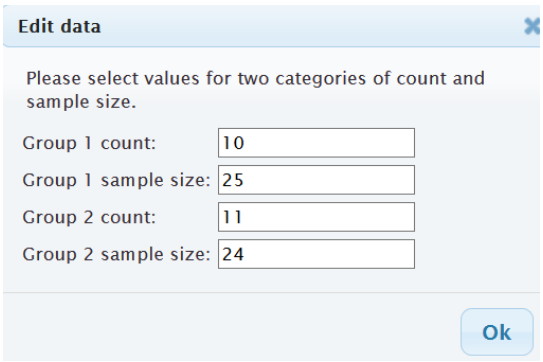
Simulation Approach

Let us use StatKey to simulate the null hypothesis. Remember, the null hypothesis is equivalent to the difference being zero, so the simulation should be centered close to zero.

StatKey Directions for Two Population Proportions (percentages)

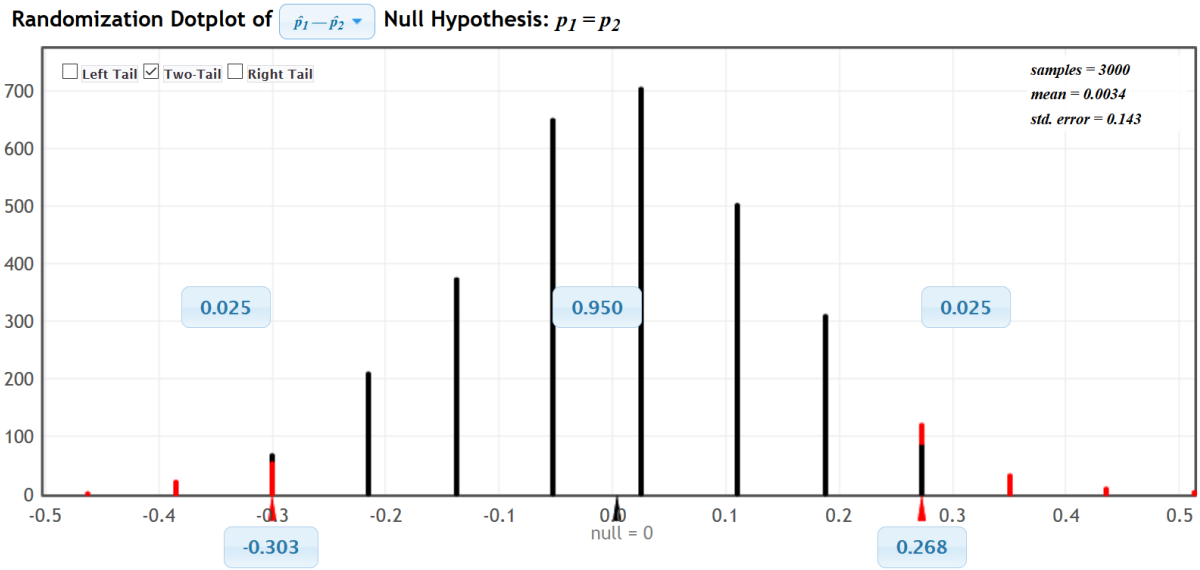
Randomization Hypothesis Tests → Test for Difference in Proportions → Under "edit data", put in summary counts → click "generate 1000 samples" multiple times → click on tail determined by the alternative hypothesis → Enter sample proportion difference in bottom box. (If the difference is negative, put it in the left box. If the difference is positive, put it in the right box. P-value will be automatically calculated above the sample difference in the tail.)

We put the data into StatKey, simulated the null hypothesis, and then clicked on "two-tail".



Original Sample

Group	Count	Sample Size	Proportion
Group 1	10	25	0.400
Group 2	11	24	0.458
Group 1-Group 2	-1	n/a	-0.058



↑
Sample Proportion Difference = −0.058

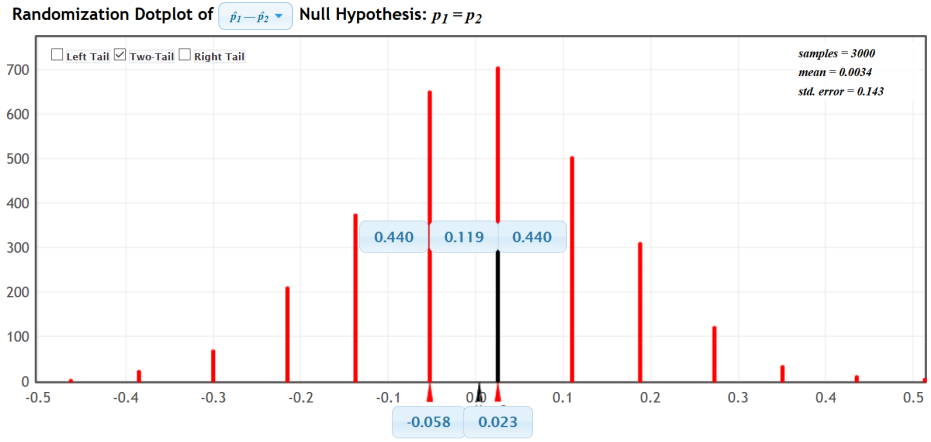
Notice that in simulation, it is important to identify the tail. With a 5% significance level and a two-tailed test, there is 2.5% in each tail. We see from the simulation that sample differences of approximately −0.303 or less are significant. Also sample differences of approximately +0.268 or higher are significant. Our real sample difference −0.058 was not in either of the tails. The sample data does not significantly disagree with the null hypothesis. It also tells us that the sample proportions are not significantly different.

StatKey does not calculate the Z-score test statistic, but we do have the approximate standard error from the simulation of about 0.143. Using the test statistic formula, we get the following.

$$Z = \frac{(sample\ proportion\ 1 - sample\ proportion\ 2)}{standard\ error} = \frac{-0.058}{0.143} \approx -0.41$$

So the sample proportion of high retention for the liked music group was only 0.41 standard errors below the sample proportion of high retention for the hated music group. This is not significant. We do not have a critical value, yet we saw that the sample difference was not in the tail of the simulation.

Now let us use the simulation, to calculate the P-value and check whether this data could have happened because of sampling variability (random chance). If we enter the original sample difference of −0.058 in the left bottom box, we get the following.



Notice in a two-tailed test, you need to add the proportions in both tails (upper boxes) to get the P-value.

P-value ≈ 0.440 + 0.440 = 0.880 = 88.0%

P-value Sentence: If the null hypothesis is true, there is an 88.0% probability of getting this sample data or more extreme because of sampling variability.

Interpret the P-value: This is a very large P-value and is much larger than the 5% significance level. This indicates that the population proportions may be equal and the sample data could have happened because of sampling variability. Since sampling variability is involved, we must fail to reject the null hypothesis.

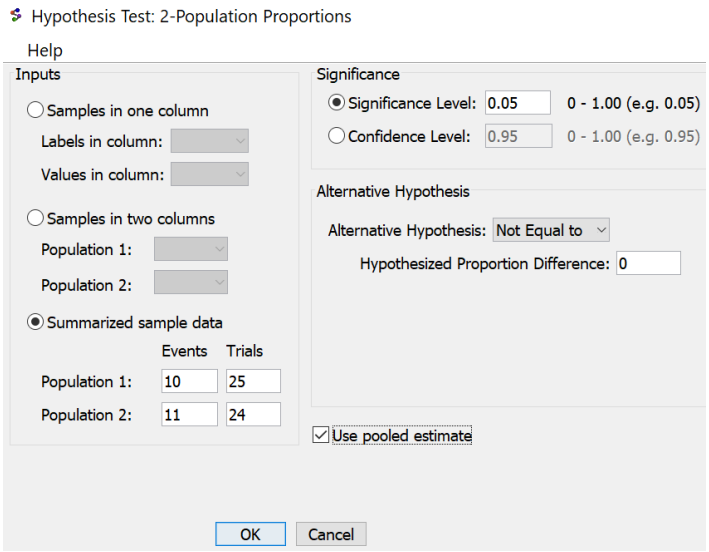
Conclusion: There is not significant evidence to support the claim that liking music is related to high retention. Notice that the alternative hypothesis (related) was the claim and we have a high P-value. Data seems to indicate they are not related, though we do not have significant evidence. This was an experiment with random assignment, so we may say the data indicates that liking the music does not cause a significant difference in the high retention percentage.

We could also use Statcato to calculate the test statistic, critical values and P-value.

Statcato Directions for Two Population Proportions (percentages)

Statistics → Hypothesis Tests → 2-Population Proportions → Samples in one column, samples in two columns or summarized sample data → put in alternative hypothesis sign (usually ≠ for relationships)
→ Hypothesized proportion difference: 0 → check "use pooled estimate" → put in significance level
→ push "OK".

Here is the Statcato printouts for the same problem.



Hypothesis Test - Two population proportions: confidence level = 0.95

	Number of Events	Number of trials	Proportion
Sample 1	10	25	0.4
Sample 2	11	24	0.458

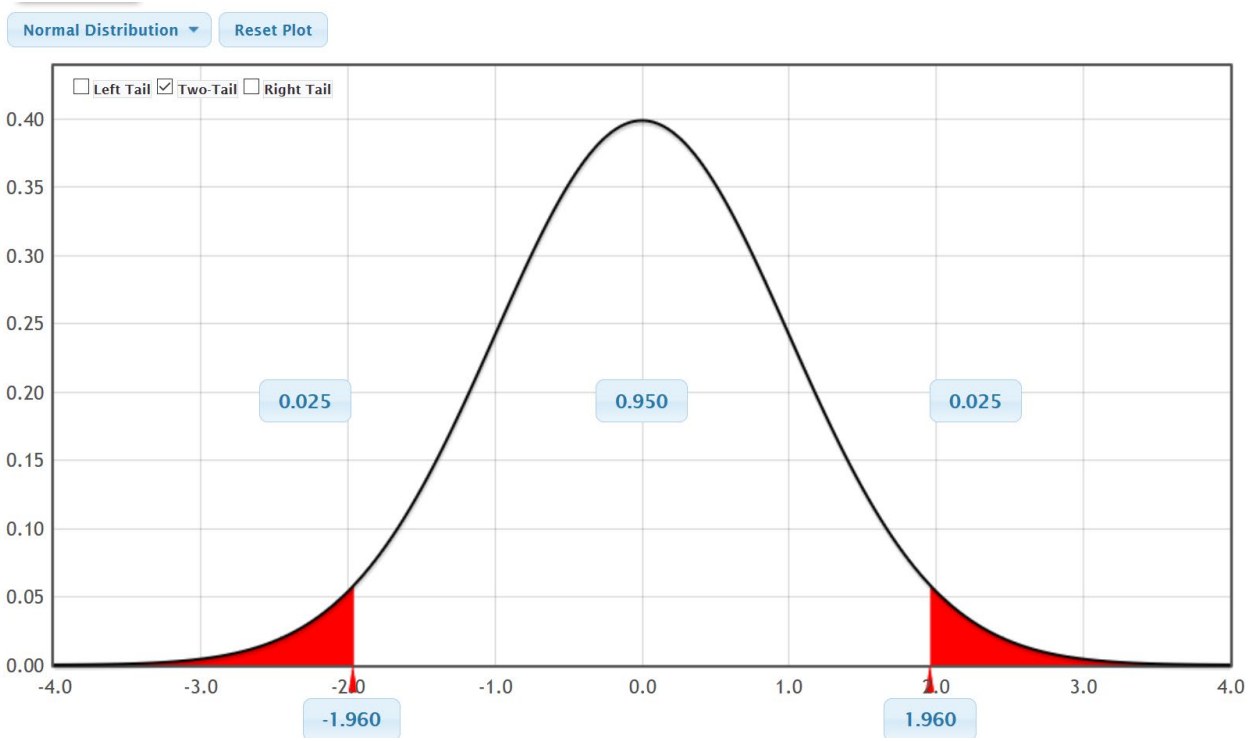
Null hypothesis: $p_1 - p_2 = 0.0$

Alternative hypothesis: $p_1 - p_2 \neq 0.0$

Significance Level	Critical Value	Test Statistic Z	p-Value
0.05	-1.96, 1.96	-0.412	0.6800

Notice the Z-test statistic is similar to what we got with StatKey, though the P-value is lower. Notice Statcato gave us critical values to compare the Z-test statistic to. The test statistic does not fall in the tail determined by the critical values. The P-value is still extremely large and indicates that if the null hypothesis was true, this data or more extreme could have happened because of sampling variability (random chance).

Also, notice the way Statcato wrote the null and alternative hypothesis. Saying two parameters are equal is the same as saying the difference is zero. You may see the null and alternative hypothesis written in different ways.



↑

Z = −0.412

Problems Section 4C

(#1-10) Use each of the following two-population proportion Z-test statistics and the corresponding critical values to fill out the table.

	Z-test stat	Sentence to explain Z-test statistic.	Critical Value	Does the Z-test statistic fall in a tail determined by a critical value? (Yes or No)	Does sample data significantly disagree with H_0?
1.	−1.835		±1.645		
2.	+0.974		+2.576		
3.	−1.226		−1.96		
4.	−3.177		±1.96		
5.	+2.244		+1.645		
6.	+1.448		±2.576		
7.	−0.883		−2.576		
8.	+1.117		+1.96		
9.	+2.139		±2.576		
10.	−0.199		−1.645		

(#11-20) Use each of the following P-values and corresponding significance levels to fill out the table.

	P-value Proportion	P-value %	Sentence to explain the P-value	Significance Level %	Significance level Proportion	If H_0 is true, could the sample data occur by random chance or is it unlikely?	Reject H_0 or Fail to reject H_0?
11.	0.728			10%			
12.	0.0421			1%			
13.	2.11×10^{-4}			5%			
14.	0.0033			1%			
15.	0.176			5%			
16.	0			10%			
17.	0.0628			5%			
18.	0.277			10%			
19.	3.04×10^{-6}			1%			
20.	0			5%			

21. Explain the difference between random samples and random assignment.

22. List the assumptions that we need to check for a two-population proportion hypothesis test.

23. List the assumptions that we need to check for a two-population proportion hypothesis test that is using experimental design.

24. Explain how to use a two-population proportion hypothesis test to show that two categorical variables are related.

25. Explain how to use a two-population proportion hypothesis test to show there is a cause and effect between two categorical variables.

(#26-30) Directions: Use the following Statcato printouts to answer the following questions.

a) Write the null and alternative hypothesis. Include relationship implications. Is this a left-tailed, right-tailed, or two-tailed test?

b) Check all of the assumptions for a two-population proportion Z-test. Explain your answers. Does the problem meets all the assumptions?

d) Write a sentence to explain the Z-test statistic in context.

e) Use the test statistics and the critical value to determine if the sample data significantly disagrees with the null hypothesis. Explain your answer.

f) Write a sentence to explain the P-value.

g) Use the P-value and significance level to determine if the sample data could have occurred by random chance (sampling variability) or is it unlikely to random chance? Explain your answer.

h) Should we reject the null hypothesis or fail to reject the null hypothesis? Explain your answer.

i) Write a conclusion for the hypothesis test. Explain your conclusion in plain language.

j) Is the population proportion related to the categorical variable or not? Explain your answer.

26. The United States has the highest teen pregnancy rate in the industrialized world. In 2008, a random sample of 1014 teenage girls found that 326 of them were pregnant before the age of 20. In 2012, a random sample of 1025 teenage girls was taken and 334 were found to be pregnant before the age of 20. Let population proportion 1 represent 2008 and population proportion 2 represent 2012. Use a 10% significance level and the following Statcato printout to test the claim that the population percentage of teen pregnancies in the U.S. is lower in 2008 than it is in 2012. This claim would indicate that the population percentage of U.S. teen pregnancies is related to the year.

	Number of Events	Number of trials	Proportion
Sample 1	334	1025	0.326
Sample 2	326	1014	0.321

Significance Level	Critical Value	Test Statistic Z	p-Value
0.05	-1.645	-0.210	0.4168

27. While many Americans favor the legalization of marijuana, opponents of legalization argue that marijuana may

be a gateway drug. They believe that if a person uses marijuana, then they are more likely to use other more dangerous illegal drugs. Use the table of random sample data given below and a 5% significance level to test the claim that marijuana users have a higher percentage of other drug use than non-marijuana users. This claim also would indicate that using Marijuana is related to using other drugs.

	Uses Other Drugs	Total
Uses Marijuana	87	213
Does not use Marijuana	26	219

	Number of Events	Number of trials	Proportion
Sample 1	87	213	0.408
Sample 2	26	219	0.119

Significance Level	Critical Value	Test Statistic Z	p-Value
0.05	1.645	6.850	$3.6839 \cdot 10^{-12}$

28. Use a 1% significance level and the following Statcato printout to test this claim that gender is not related to abstaining from drinking alcohol. If this is the case, then the percentage of men and women that do not drink alcohol should be the same. We took a random sample of 190 men and found that 66 of them never drink alcohol. We took a random sample of 250 women and found that that 137 of them never drink alcohol. We designated the proportion of men that never drink alcohol as population 1 and the women as population 2.

	Number of Events	Number of trials	Proportion
Sample 1	66	190	0.347
Sample 2	137	250	0.548

Significance Level	Critical Value	Test Statistic Z	p-Value
0.01	-2.576, 2.576	-4.182	$2.8937 \cdot 10^{-5}$

29. A health magazine claims that marriage status is one of the most telling factors for a person's happiness. Use a 10% significance level and the Statcato printout below to test the claim that the percent of married people that are unhappy is lower than the percent of single or divorced people that are unhappy. If this is the case, then perhaps being married, single or divorced is related to being unhappy. The following sample data was collected randomly. Population 1 represented married adults and population 2 represented single or divorced adults.

	Unhappy	Total
Married	74	200
Single or Divorced	97	200

	Number of Events	Number of trials	Proportion
Sample 1	74	200	0.37
Sample 2	97	200	0.485

Significance Level	Critical Value	Test Statistic Z	p-Value
0.10	-1.282	-2.325	0.0100

30. A tattoo magazine claimed that the percent of men that have at least one tattoo is greater than the percent of women with at least one tattoo. If this were true, then gender would be related to having a tattoo. Use a 5% significance level and the following Statcato printout to test this claim. A random sample of 857 men found that 146 of them had at least one tattoo. A random sample of 794 women found that 137 of them had at least one tattoo. Population 1 was the proportion of men with at least one tattoo and population 2 was the proportion of women with at least one tattoo.

	Number of Events	Number of trials	Proportion
Sample 1	146	857	0.170
Sample 2	137	794	0.173

Significance Level	Critical Value	Test Statistic Z	p-Value
0.05	1.645	-0.118	0.5468

(#31-34) Directions: go to www.lock5stat.com and click on StatKey. Then under the "Randomization Hypothesis Test" menu click on "Test for Difference in Proportions". Create a randomized simulation of the null hypothesis to answer the following questions.

a) Write the null and alternative hypothesis. Include relationship implications. Is this a left-tailed, right-tailed, or two-tailed test?

b) What is the difference between the sample proportions? Adjust the tails of your simulation to reflect the significance level. Did your sample proportion difference fall in the tail?

c) Does the sample data significantly disagree with the null hypothesis? Explain your answer.

d) Put the sample proportion difference into the bottom box in the appropriate tail of your simulation in order to calculate the P-value. What was the P-value? (Answers will vary.) Write a sentence to explain the P-value.

e) Use the P-value and significance level to determine if the sample data could have occurred by random chance (sampling variability) or is it unlikely to random chance? Explain your answer.

f) Should we reject the null hypothesis or fail to reject the null hypothesis? Explain your answer.

g) Write a conclusion for the hypothesis test. Explain your conclusion in plain language.

h) Is the population proportion related to the categorical variable or not? Explain your answer.

i) Use the following formula to calculate the Z-test statistic. Write a sentence to explain the Z-test statistic in context. (Answers will vary.)

$$Z\ test\ stat = \frac{Sample\ Proportion\ Difference}{Standard\ Error}$$

31. A body mass index of 20-25 indicates that a person is of normal weight for their height and body type. A random sample of 760 women found that 198 of the women had a normal BMI. A random sample of 745 men found that 273 of them had a normal BMI. A fitness magazine claims that the percent of women with a normal BMI is lower than the percent of men with a normal BMI. This would imply that gender is related to having a normal BMI. Let population 1 be the proportion of women with a normal BMI and population 2 be the proportion of men with a normal BMI. Use a 10% significance level and a randomized simulation in StatKey.

32. A new medicine has been developed that treats high cholesterol. An experiment was conducted and adults were randomly assigned into two groups. The groups had similar gender, ages, exercise patterns and diet. Of the 420 adults in the placebo group, 38 of them showed a decrease in cholesterol. Of the 410 adults in the treatment group, 49 of them showed a decrease in cholesterol. The FDA claims that the medicine is not effective in lowering cholesterol since the proportion for the placebo group and the treatment groups are about the same. Use a randomized simulation in StatKey, and a 1% significance level to test this claim.

33. A study was done to see if there is a relationship between smoking and being able to get pregnant. Two random samples of women trying to get pregnant were compared. A random sample of 135 women that smoke (population 1) found that 38 were able to get pregnant in the allotted amount of time. A random sample of 543 women that do not smoke (population 2) found that 206 were able to get pregnant in the allotted amount of time. Test the claim that the population percent of smoking women that were able to get pregnant is lower than the population percent of non-smoking women. This claim also implies that smoking is related to getting pregnant. Use a randomized simulation in StatKey, and a 5% significance level to test this claim.

34. A study was done to see if there is a relationship between the age of a person (teen or adult) and using text messages to communicate. A random sample of 800 teens (population 1) found that 696 of them use text messages regularly to communicate. A random sample of 2252 adults (population 2) found that 1621 of them use text messages regularly to communicate. Test the claim that population percentages are equal for the two groups implying that age is not related to using text messages. Use a randomized simulation in StatKey, and a 10% significance level to test this claim.

Section 4D – Proportion Relationships: Goodness of Fit Test

While the Z-score test statistic works well for two population proportion tests, it cannot handle proportions from three or more groups. For this case, we will introduce a new test statistic called "Chi-squared" (χ^2). This test statistic is usually used for more complicated categorical relationship analysis. The Goodness of Fit test works a lot like the two-population proportion relationship test except that there are now three or more groups. The opposite of three or more parameters being equal is not all of them being not equal. If even one is significantly different, we should reject the null hypothesis. For this reason, many statisticians prefer to use the phrase "at least one is not equal" or "the distribution is different than the null hypothesis". I prefer the former.

Remember, if the population proportion or percentage is the same for all the groups, then it does not matter what group we are in. That would tell us that the population percentage is not related to the categorical variable that determines the groups. If the population proportion or percentage is different in at least one of the groups, then it does matter what group we are in. That would tell us that the population percentage is related to the categorical variable that determines the groups.

Null and Alternative Hypothesis for the Goodness of Fit Test

H_0: $p_1 = p_2 = p_3 = p_4 = p_5$ The population % is NOT related to a categorical variable (% is not related to the groups)
H_A: *At least one* $\neq$ The population % is related to a categorical variable (% is related to the groups)

Expected Counts and Observed Counts

All hypothesis tests need to find some way of comparing the sample data to the null hypothesis. That is very difficult when you have three or more groups. The Goodness of Fit test compares the observed counts from the sample data to the expected counts from the null hypothesis. To calculate the Chi-Squared test statistic for a Goodness of Fit test, we will subtract the observed sample counts (number of successes) from each group to what we expect to happen if the null hypothesis was true (expected counts). Think of the observed counts as what really happened in the sample data and the expected counts as a theoretical count based on the null hypothesis being true. In this way, we can determine if the sample data significantly disagrees with the null hypothesis even if we have twenty groups.

Observed Counts: The counts from the sample data. Also called the number of successes or number of events.

Expected Counts: Theoretical counts based on the premise that the null hypothesis is true.

Calculating the Chi-squared test statistic (χ^2) for the Goodness of Fit Test

The Chi-squared test statistic works like a variance calculation. In fact, we have seen previously that the Chi-squared distribution is often used in one-population variance confidence intervals and hypothesis tests. Instead of calculating a sum of squares to measure the difference between data values and the mean, we will be calculating a sum of squares the measures the difference between the observed and expected counts. We need an average of the squares so we divide by the expected count for each group.

Chi-Squared Test Statistic formula: $\chi^2 = \sum \frac{(O-E)^2}{E}$

The more groups you have in your data, the more difficult this formula is to calculate. While we will show an example of how the Chi-squared test statistic is calculated, it is always better to use a computer program to calculate it for you. It is more important to be able to explain the test statistic and be able to use it to determine if the sample data significantly disagrees with the null hypothesis.

Chi-Squared Test Statistic (χ^2) Sentence: The sum of the averages of the squares of the differences between the observed sample counts and the expected counts from the null hypothesis.

Degrees of Freedom for Goodness of Fit Test = k – 1

Interpreting the Chi-squared test statistic for a Goodness of Fit Test

The first thing to know about a Goodness of Fit test is that it is always a right-tailed test. It is never left-tailed or two-tailed. You may be comparing the proportions of twenty groups, but the Goodness of Fit test condenses it into one right-tailed test.

Degrees of Freedom

If we have ten groups, we will have ten expected counts and ten observed counts. So the degrees of freedom for our calculation will be the number of groups (k) minus one. For ten groups, the degrees of freedom will be 10 – 1 = 9. This is important when looking up critical values.

Determining Significance

As with all hypothesis tests, if the test statistic falls in the tail determined by the critical value, then the sample data significantly disagrees with the null hypothesis. If the test statistic does not fall in the tail, then the sample data does not significantly disagree with the null hypothesis. The Goodness of Fit Test is a right-tailed test so the test statistic must fall in the right tail to be considered significant.

Assumptions for the Goodness of Fit Test for Categorical Relationships

1. Random: The sample categorical data should be either a random sample or representative *(if proving there is relationship)* or have used random assignment *(if proving cause and effect)*.

2. Large sample size: The expected counts should be at least five. *In the Chi-squared test statistic calculation, we calculate theoretical counts based on the null hypothesis being true. These counts are called the expected counts (expected frequencies or expected values). In the Goodness of Fit test we want all of the expected counts to be five or greater. An expected count below five indicates the sample size was not large enough for a Goodness of Fit test.*

3. Data values within each sample and between the samples should be independent of each other. If the data was collected from one sample then the assumption is just that individuals should be independent. If the data was collected from more than one sample, then the samples and the individuals should be checked for independence. *As with the two population proportion assumptions, if we are doing an experiment, we should not control confounding variables by using the same group of people measured multiple times. This would fail the independent individuals' assumption. Random assignment is a better option for controlling confounding variables.*

Example 1 (Goodness of Fit Categorical Relationship Test) *Case 1: Equal proportions but data collected from different samples with unequal sample sizes.*

In the previous example, we looked at data comparing two groups, those that listened to a music they liked and those that listened to a music they hated. From this data, we were able to see if liking a music or not is related memorizing information (high retention).

The scientists in this experiment also had a third group that did not listen to any music. If you recall from our discussion about experimental design, this is called the control group.

Here is the sample data.

Liked Music: 25 total people, 10 high retention, $\hat{p}_1 \approx 0.4$
Hated Music: 24 total people, 11 high retention, $\hat{p}_2 \approx 0.458$
No Music: 26 total people, 19 high retention, $\hat{p}_3 \approx 0.731$

Let us use a 5% significance level to test the claim the having music or not is related to high retention.

H_0: $p_1 = p_2 = p_3$ (High retention is NOT related to having music or not.)
H_A: *At least one* $\neq$ (High retention is related to having music or not.) CLAIM

What are the expected counts?

To calculate the expected counts, we have to think about what we would expect to happen if the null hypothesis was true. Remember if there is no relationship between the variables, then the music should not matter when it comes to

memorizing information. The percentage of high retention should be the same. So each of the three groups should have the same percentage and the same expected count. If we disregard music, then there was a total of 75 adults and 40 tested into high retention. So if the null hypothesis is true and music is not related to memorizing information, then all the music groups should have a proportion of 40/75 ≈ 0.533. In our study of categorical data analysis, we saw that to estimate an amount you multiple the proportion times the total number of people or objects in that group.

Expected Count for each Group = (proportion for group if null hypothesis was true) x (sample size of that group)

Expected Count for Liked Music Group = 0.533 x 25 ≈ 13.325
Expected Count for Hated Music Group = 0.533 x 24 ≈ 12.792
Expected Count for No Music Group = 0.533 x 26 ≈ 13.858

Let us check the assumptions for this test. Notice that this is an experiment, so will require random assignment instead of random samples. Since the data was collected from multiple samples, we will need the groups and individuals to be independent.

1. Was the sample data collected randomly? Yes. The groups were randomly assigned in the experiment. This will account for confounding variables in the cause and effect study.

2. Are all of the expected counts at least five? Yes. The expected counts were 13.325, 12.792, and 13.858. All of them are greater than five.

3. Are the data values independent? Yes. It is always difficult to judge independence. Since the groups were randomly assigned and not the same people measured three times, the groups are probably independent of each other. It is difficult to judge whether the individuals are independent in an experiment. They are often volunteers and may have some relationship like maybe they all came from the same college. We will assume it passes this assumption for now, but may need to check with the people running the experiment.

Note: Z-test statistics can only compare two proportions at a time and cannot compare three or more proportions. Hence, we will need to use the chi-squared test statistic (χ^2).

Chi-Squared Test Statistic (χ^2)

The goal of any test statistic is to see if the sample data significantly disagrees with the null hypothesis. To do this, we compare the actual sample data "observed" counts to the theoretical "expected" counts if the null hypothesis was true.

We need to know if the observed counts for high retention (10, 11 and 19) are significantly different from the expected counts in the null hypothesis.

Observed Sample Count for Liked Music Group: 10 high retention
Observed Sample Count for Hated Music Group: 11 high retention
Observed Sample Count for No Music Group: 19 high retention

Expected Count for Liked Music Group = 0.533 x 25 ≈ 13.325
Expected Count for Hated Music Group = 0.533 x 24 ≈ 12.792
Expected Count for No Music Group = 0.533 x 26 ≈ 13.858

Calculating Chi-Squared

We want to know if the expected counts were significantly different from the observed counts. This will tell us if the sample data significantly disagrees with the null hypothesis. The chi-squared test statistic will tell us.

Chi-Squared Test Statistic (χ^2): The sum of the averages of the squares of the differences between the observed sample counts and the expected counts if the null hypothesis was true.

When calculating the Chi-squared test statistic for the Goodness of Fit test, it is important that you pair the observed count with the correct expected count from that same group. For this reason, the observed and expected counts are labeled to reflect what group they came from.

Observed Sample Count for Liked Music Group: O_1 = 10
Observed Sample Count for Hated Music Group: O_2 = 11
Observed Sample Count for No Music Group: O_3 = 19

Expected Count from H_0 for Liked Music Group: E_1 ≈ 13.325
Expected Count from H_0 for Hated Music Group: E_2 ≈ 12.792
Expected Count from H_0 for No Music Group: E_3 ≈ 13.858

Chi-Squared Test Statistic formula = $\sum \frac{(O-E)^2}{E}$

$\chi^2 = \frac{(O-E)^2}{E} = \frac{(10-13.325)^2}{13.325} + \frac{(11-12.792)^2}{12.792} + \frac{(19-13.858)^2}{13.858} \approx 0.830 + 0.251 + 1.908 = 2.989$

Interpreting Chi-Squared

Is the test statistic of χ^2 = 2.989 significant? To judge if this is significant, we will need a critical value, P-value, or to see if it is in the tail of a simulation.

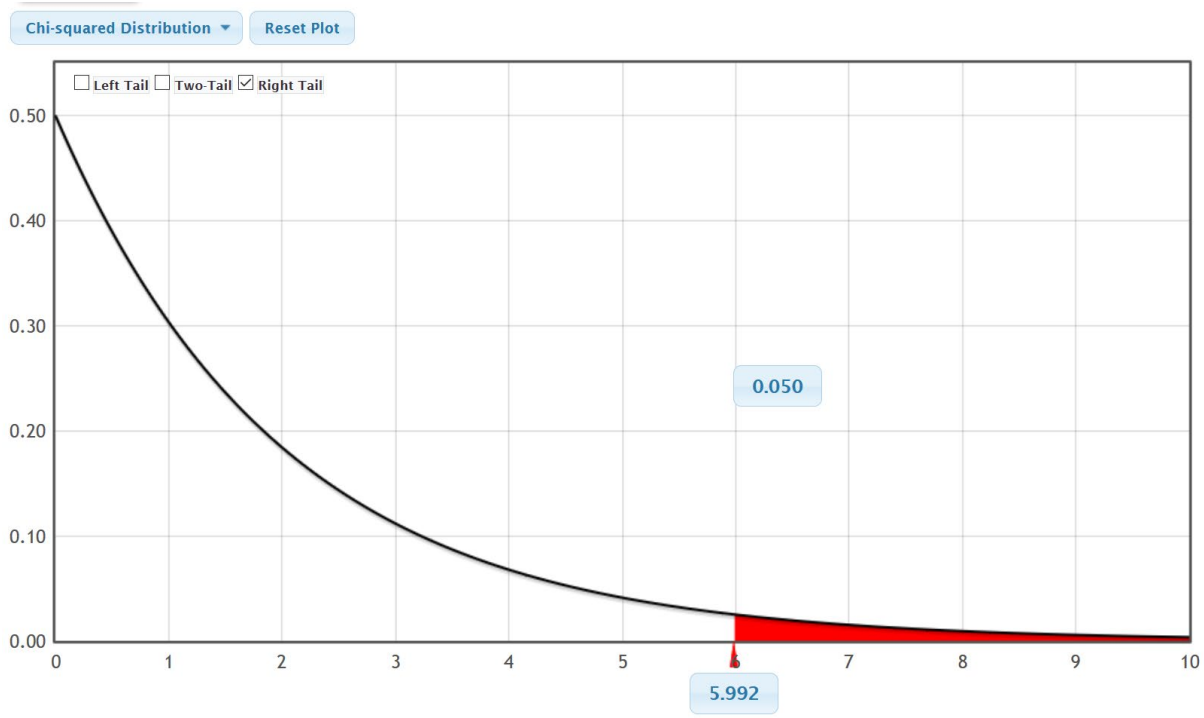
Note: Chi-squared Goodness of Fit test is always right tailed. Remember if the null hypothesis were true, then the observed and expected counts would be about the same. If that is the case then when you subtract them you should get about zero. So the center of the Chi-squared distribution should be close to zero. If you square numbers and add them up, it would be impossible for them to ever be negative.

Degrees of Freedom: The chi-squared test statistic is based on the counts for the number of groups (k) so the degrees of freedom for a goodness of fit test is k-1. In this problem, there were three groups so the degrees of freedom is 3-1 = 2.

Since we have already calculated the test statistic, we can look up the critical value and P-value with the StatKey theoretical chi-squared function.

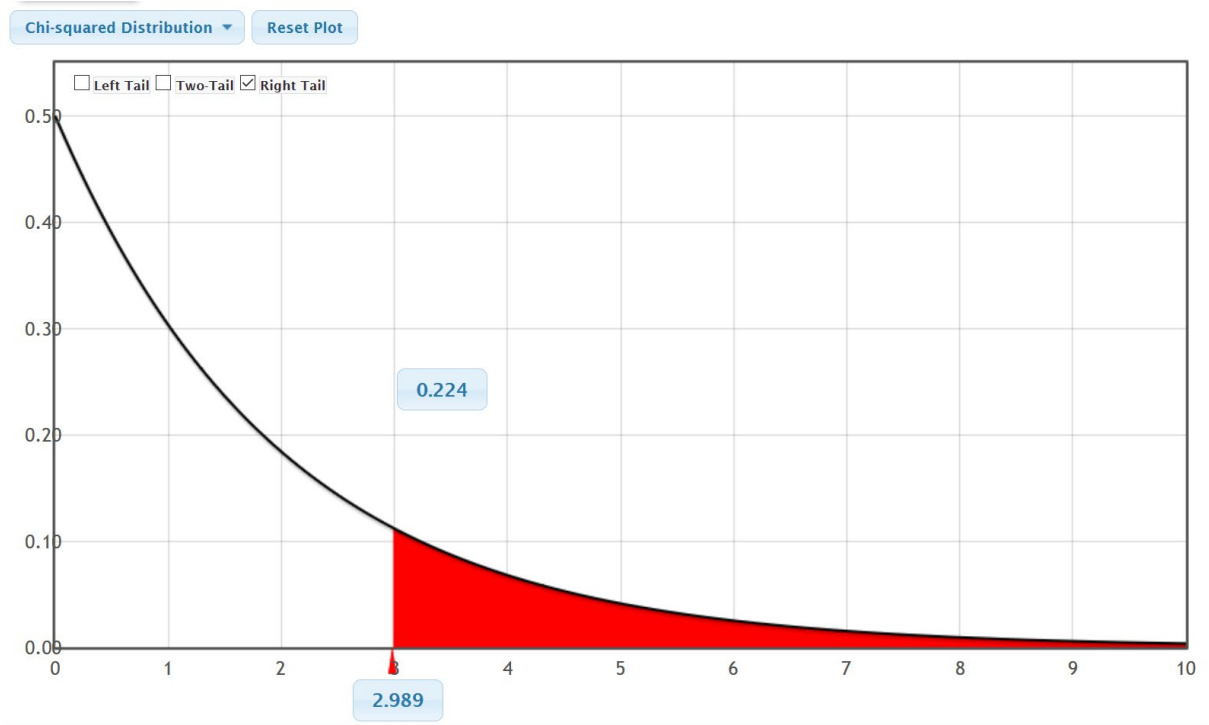
StatKey (Theoretical Chi-Squared)

Theoretical Distributions → χ^2 → degrees of freedom: 2 → Click "right tail". (Remember chi-squared is always a right tailed test.) → To calculate the critical value, put in the significance level into the upper box. (The critical value will be below it.) → For the P-value, put the test statistic into the lower box. (The P-value will be above it.)



↑
$\chi^2 = 2.989$

Notice that the critical value for a 5% significance level is 5.992. This means that the test statistic should be greater than 5.992 to be considered significant. Notice this implies that our chi-squared test statistic of 2.989 is not in the tail and not significant. So our sample data does not significantly disagree with the null hypothesis.



Notice that when we plugged in the test statistic of 2.989 into the theoretical Chi-squared curve, the estimated P-value is about 0.224 = 22.4%. This is a rather large P-value and is much larger than the 5% significance level. If the null hypothesis was correct, then this sample data or more extreme could have happened because of sampling variability (random chance).

Since we cannot rule out sampling variability, we should fail to reject the null hypothesis.

Conclusion: Since the P-value is high and the claim is the alternative hypothesis, our conclusion should be that we do not have significant evidence to support the claim that listening to music is related to high retention. The sample data indicates that listening to music is not related to high retention, though we do not have evidence.

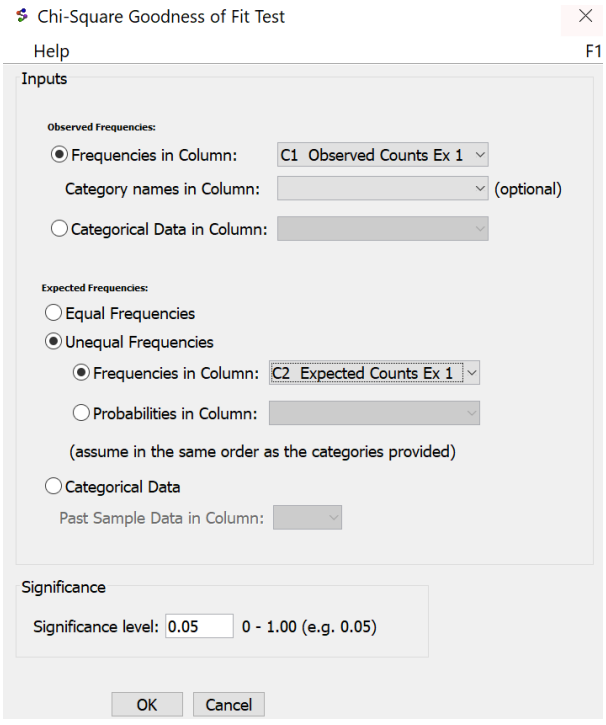
We could also have calculated the test statistic, critical value and P-value with Statcato.

Statcato Directions for Goodness of Fit

First type in the observed counts in one column of Statcato and the expected counts into a second column. Take note of whether your expected counts are equal or not. In this case, they are not equal. The proportions were assumed equal in the null hypothesis, but since the sample sizes of the groups are different, the expected counts will be different.

	C1	C2
Var	Observed Counts Ex 1	Expected Counts Ex 1
1	10	13.325
2	11	12.792
3	19	13.858

Statistics → Multinomial Experiment → Chi-Square Goodness-of-Fit → Under (observed) Frequencies in Column: C1 (or whatever column has your observed sample counts) → Under Expected Frequencies: Click "Unequal Frequencies"→ Under "Frequencies in column put in the column where you typed your expected counts → Put in the significance level → push OK.



Chi-Square Goodness-of-Fit Test:
Input: C1 Observed Counts Ex 1
Expected frequencies in C2 Expected Counts Ex 1

Category	Observed Frequency	Expected Frequency	Contribution to X^2
0	10.0	13.325	0.8297
1	11.0	12.792	0.2510
2	19.0	13.858	1.9079

N	Number of Categories	DOF	Significance	Critical Value	Test statistics	p-Value
40.0	3	2	0.05	5.9915	2.9887	0.2244

Notice that our test statistic, critical value and P-value is about the same as the theoretical distribution in StatKey.

Example 2 (Goodness of Fit Categorical Relationship Test) *Case 2: Equal proportions from one sample.*

In the fall 2015 semester at COC, we asked the Math 140 statistics students what their favorite social media is. Here is the sample data. Use a 1% significance level to test the claim that the population proportions for each social media are not the same. This would indicate that the population percentage for social medias are related the type of social media. Notice the data came from one sample and has five types of social media. This means our sample size will be the same for each social media.

H_0: $p_1 = p_2 = p_3 = p_4 = p_5$ (The population proportion of COC statistics students that prefer a social media is not related to the type of social media.)
H_A: *At least one* ≠ (The population proportion of COC statistics students that prefer a social media is related to the type of social media.) CLAIM

AB
Which social media do you use the most?
Snapchat
Other
Facebook
Instagram
Facebook
Instagram
Other
Facebook
Facebook
Facebook
Snapchat
Twitter

	Count
Facebook	75
Instagram	124
Other	27
Snapchat	71
Twitter	31

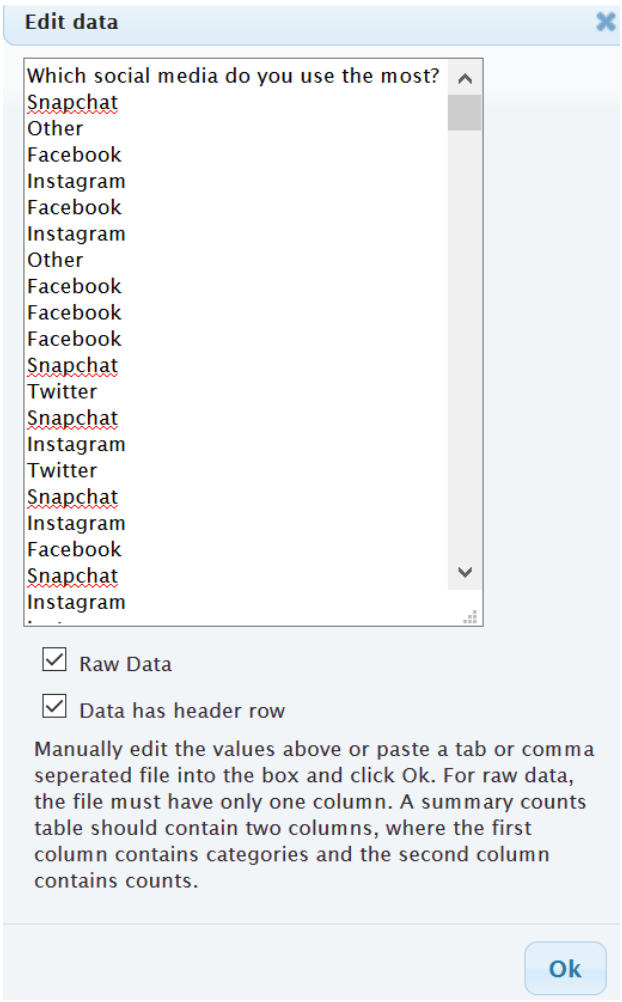
Using Randomized Simulation

We can calculate the test statistic, critical value and P-value with a randomized simulation in StatKey. Like ANOVA, since there are three or more groups involved, we will not be able to put in the sample proportions directly into the simulation. Instead, the computer will use the Chi-squared test statistic to summarize the sample data.

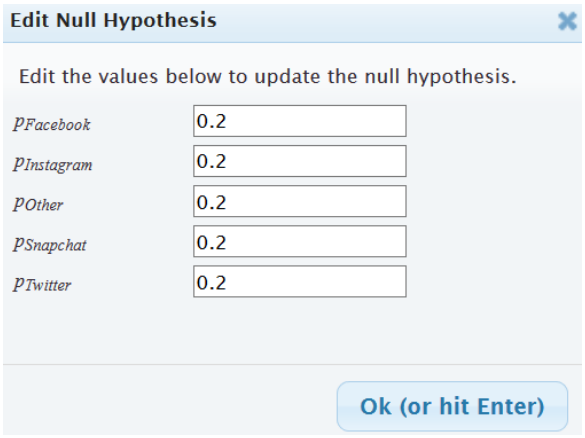
Go to www.lock5stat.com and click on StatKey. Under the "More Advanced Randomization Tests" menu click on "χ^2 Goodness of Fit". Under "Edit Data", type in the following. Do not forget to put a space after the comma. You can also use raw categorical data if you have it. Then push OK.

Choice, Count
Facebook, 75
Instagram, 124
Other, 27
Snapchat, 71
Twitter, 31

OR



It is good to look at the null hypothesis and make sure it is correct. Notice that saying that the proportions are equal is the same as saying each is 20% since we are dealing with one sample.



Notice that StatKey calculated the Chi-squared test statistic as χ^2 = 94.744 under "Original Sample".

Original Sample Show Details

$n = 328, \chi^2 = 94.744$

	Count
Facebook	75
Instagram	124
Other	27
Snapchat	71
Twitter	31

Let us check the assumptions for the Goodness of Fit Test. Under the "Original Sample" menu, click on "Show Details" to see the expected counts. Notice all of the expected counts are equal to 65.6. We can also see that the groups with the largest "Contribution to χ^2" have the most disagreement with the null hypothesis. Notice the largest "Contribution to χ^2" was 51.99, which came from the Instagram group.

Detailed Sample Table

	Count
Facebook	75 65.6 1.347
Instagram	124 65.6 51.99
Other	27 65.6 22.713
Snapchat	71 65.6 0.445
Twitter	31 65.6 18.249

Observed, Expected, Contribution to χ^2

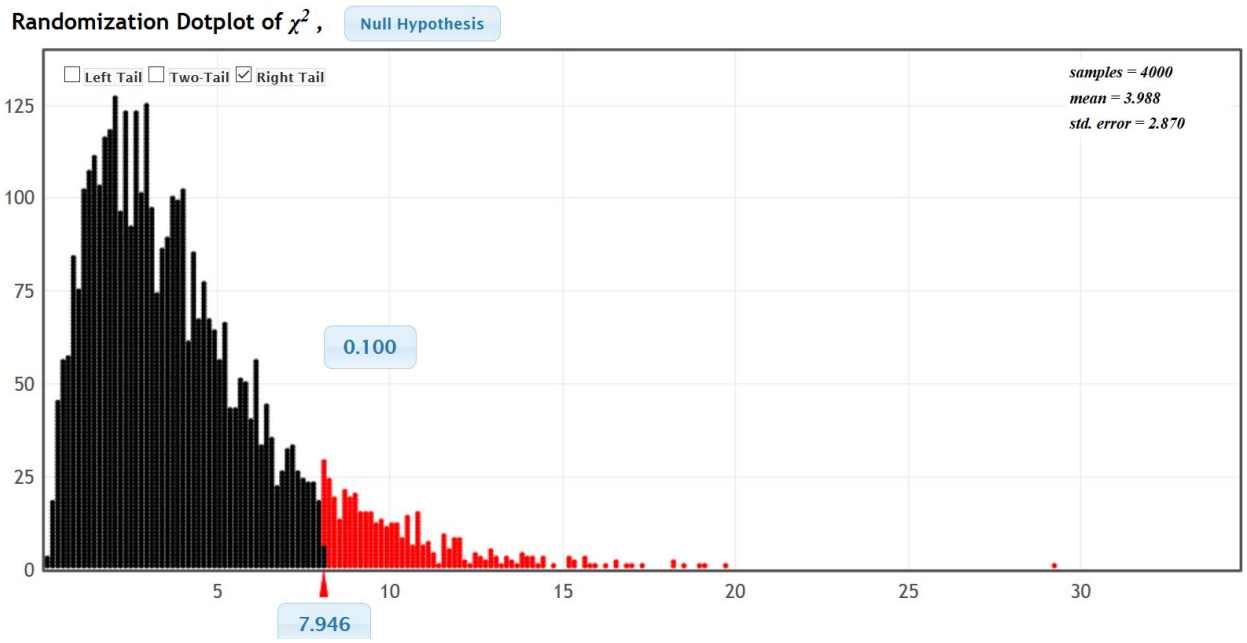
Checking the Assumptions

1. Is the sample data random or representative? Yes. Since the data was a census of all of the stat students in the fall 2015 semester, it is probably representative of all stat students at COC even though it is not a random sample.

2. Are the expected counts at least five? Yes. All of the expected counts were 65.6, which is greater than five.

3. Are the data values independent? No. Since this data came from one sample, we do not have to check that the samples are independent. It is difficult to judge if the individual stat students are independent or not. There are probably groups of friends or siblings in the data. In that case, they may have similar views about social media.

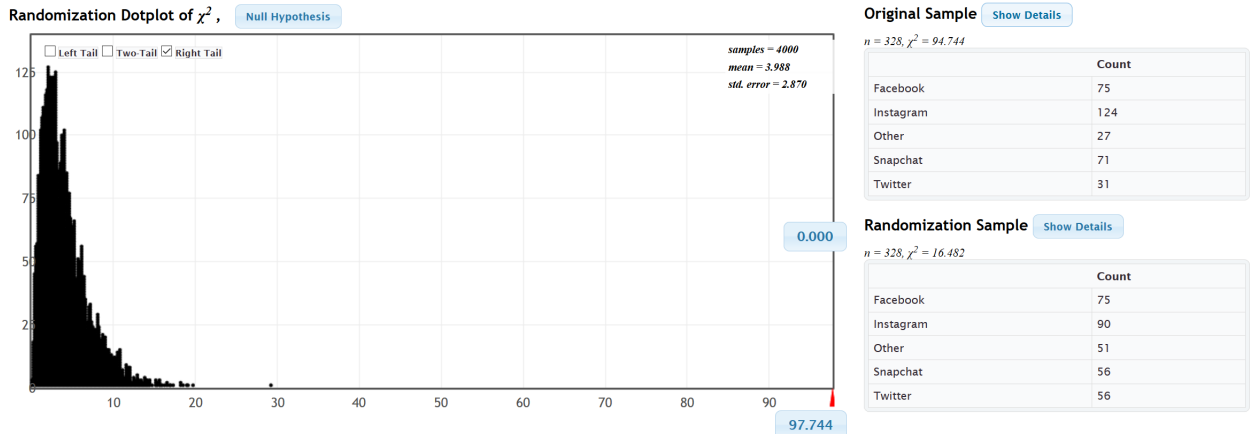
We can simulate the null hypothesis by clicking "Generate 1000 Samples" a few times. Notice the simulated distribution looks very skewed to the right. Remember the Goodness of Fit test is a right tailed test, so to calculate the Critical Value, click on "Right Tail" and put in the 10% significance level (0.10) in the tail proportion. The critical value came out to be approximately 7.946 in this simulation, but remember answers can vary due to sampling variability. In any case, our test statistic of 94.744 is way in the right tail.



↑

χ^2 = 94.744

We can also calculate the P-value by plugging in the test statistic of 94.744. Remember not to confuse the simulated chi-squared values with the actual original sample test statistic. We have calculated 4000 Chi-squared values, but only the one under "Original Sample" is the real one based on the data. Our estimated P-value is zero.



Since our test statistic fell in the tail of the simulation, we know the sample data significantly disagrees with the null hypothesis. Since our P-value was zero, we know it is highly unlikely that this sample data or more extreme occurred due to sampling variability if the null hypothesis was true.

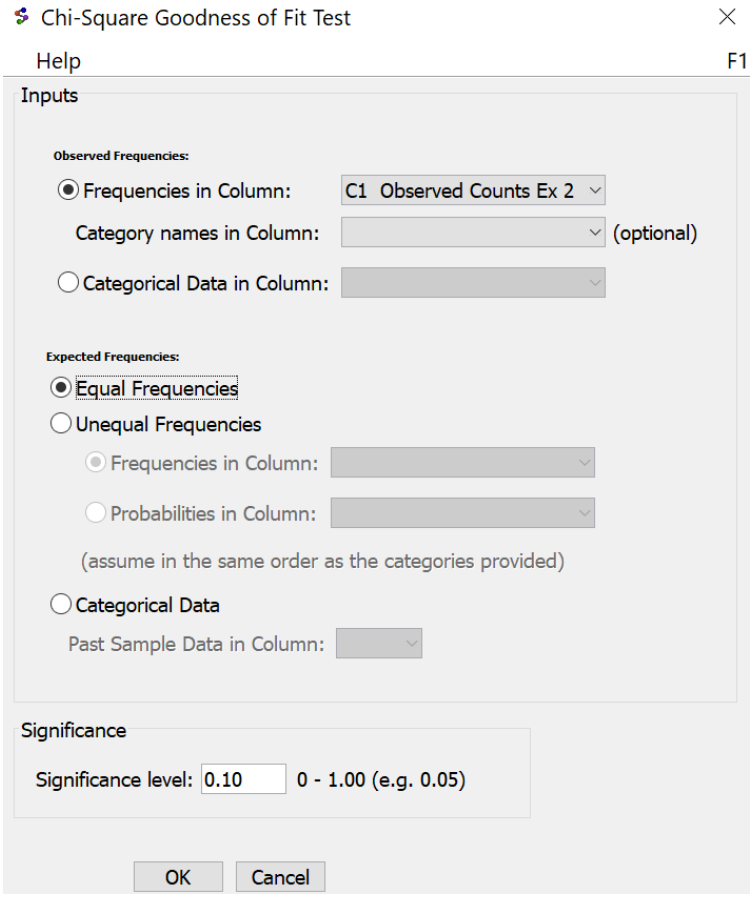
Since our P-value was low, we will reject the null hypothesis.

Since our P-value was low and our claim was the alternative hypothesis, our conclusion should be that there is significant evidence to support the claim that the population proportions are related to the type of social media. However, remember that we may have failed one of the assumptions regarding independence.

We can also calculate the test statistic, critical value and P-value with Statcato.

In Statcato, we will go to the “Statistics” menu, and then click on “Multinomial Experiments” and “Goodness of Fit”. Since we are dealing with one sample and equal proportions, the expected counts will be equal. In that case, we can click on the equal (expected) frequencies button. We will still need to type in the observed counts or we can copy and paste the raw data. Put in our 10% significance level and push OK. Notice that the critical value, test statistic, and P-value are virtually the same as the simulation in StatKey.

	C1
Var	Observed Counts Ex 2
1	75
2	124
3	27
4	71
5	31
6	



Chi-Square Goodness-of-Fit Test:
Input: C1 Observed Counts Ex 2
Expected frequency = 65.6

Category	Observed Frequency	Expected Frequency	Contribution to X^2
0	75.0	65.6	1.3470
1	124.0	65.6	51.9902
2	27.0	65.6	22.7128
3	71.0	65.6	0.4445
4	31.0	65.6	18.2494

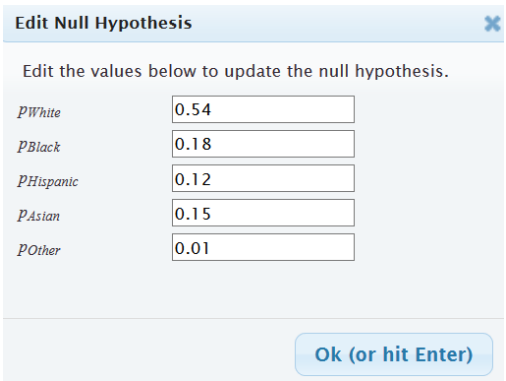
N	Number of Categories	DOF	Significance	Critical Value	Test statistics	p-Value
328.0	5	4	0.10	7.7794	94.7439	0

Another Type of Goodness of Fit Test

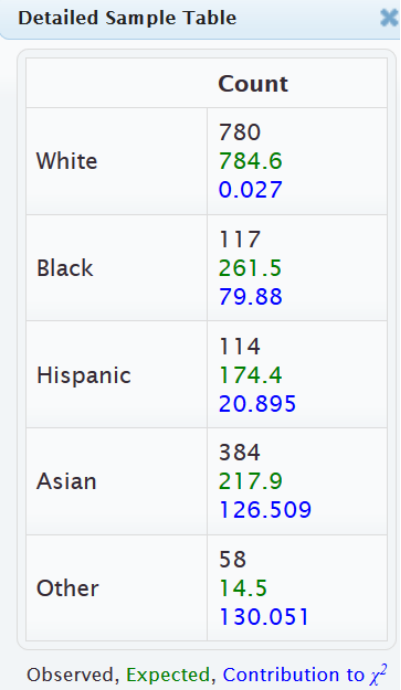
To determine a proportion relationship with a Goodness of Fit test, the null hypothesis will be that the population proportions are equal. However, Goodness of Fit tests can also be used to determine if sample data fits a specific distribution of proportions that are not necessarily all equal. When the proportions are not equal in the null hypothesis, the expected counts will also not be equal.

Example 3: Goodness of Fit Test: Unequal proportions in the null hypothesis.

A famous example of using a Goodness of Fit test in this way occurred in the case of juries in Alameda County, USA. Juries are required to represent the racial demographic of their county, yet Alameda county juries were way out of compliance. Here is the racial demographic of Alameda county at the time of the scandal. This is our null hypothesis. We will use a 1% significance level and a Goodness of Fit test to test the claim that the juries were out of compliance with these proportion.



H_0 : p_1 = 0.54 , p_2 = 0.18 , p_3 = 0.12 , p_4 = 0.15 , p_5 = 0.01 (Juries represent the Alameda racial demographic)
H_A : at least one is ≠ (CLAIM) (Juries do NOT represent the Alameda racial demographic)



Detailed Sample Table

	Count
White	780 784.6 0.027
Black	117 261.5 79.88
Hispanic	114 174.4 20.895
Asian	384 217.9 126.509
Other	58 14.5 130.051

Observed, Expected, Contribution to χ^2

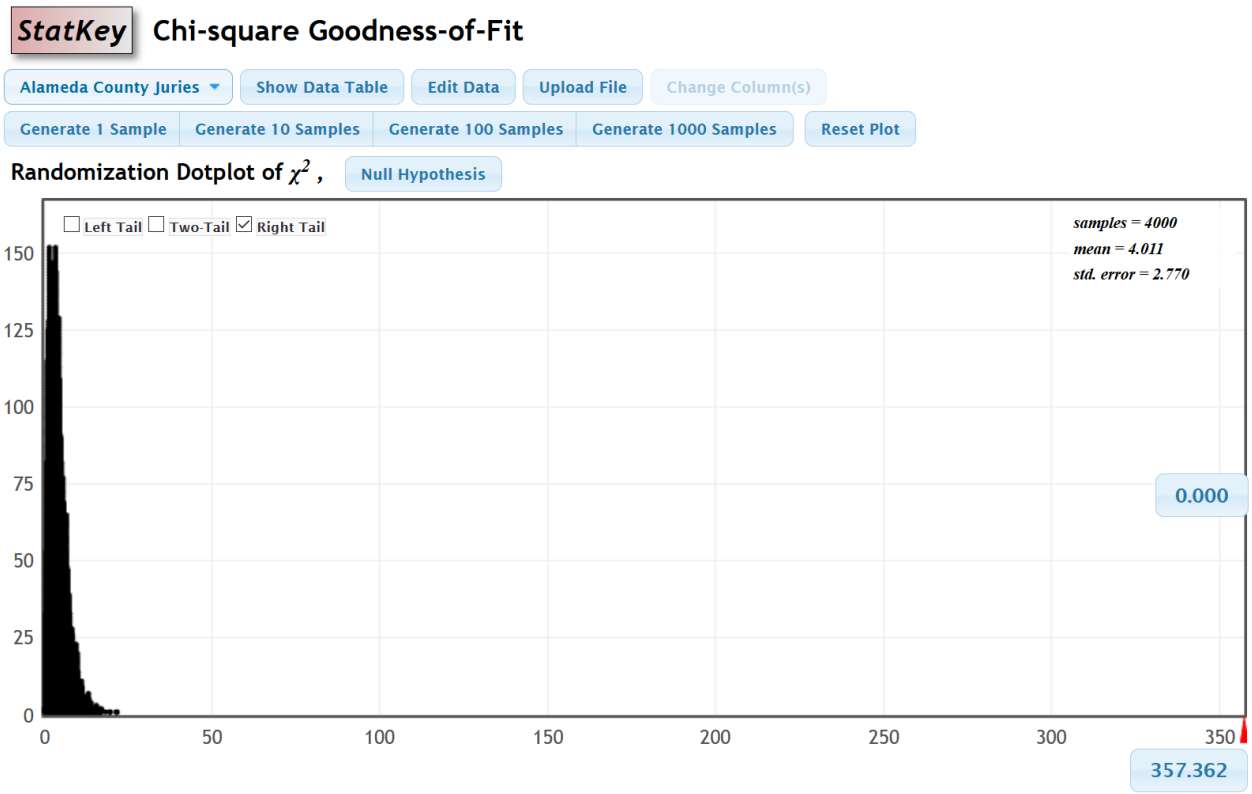
Original Sample Show Details

$n = 1453, \chi^2 = 357.362$

	Count
White	780
Black	117
Hispanic	114
Asian	384
Other	58

Here is the sample data and Chi-squared test statistic. Notice the observed and expected counts are very different for African American, Hispanic American and Asian American.

Using a randomized simulation in StatKey, we see that the test statistic was in the tail and the P-value was zero.

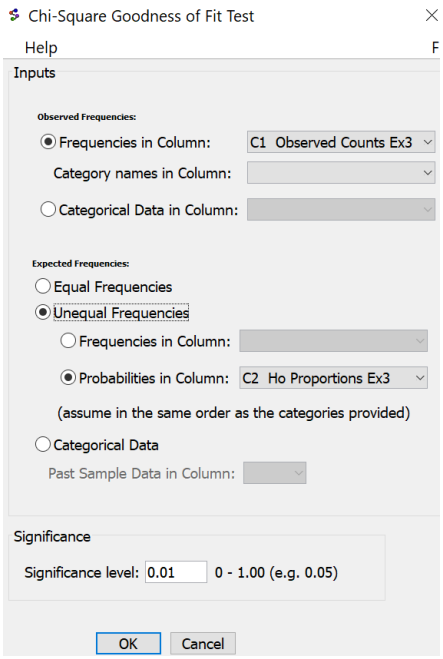


Hence we will reject the null hypothesis that the juries are in compliance and support the claim that the racial demographic of the juries in Alameda county were significantly out of compliance with the racial demographic of the county.

We can also calculate the test statistic and P-value with Statcato. Since the null hypothesis has specific proportions, we will need to type them in a column of Statcato. We will also need to type in the observed sample counts.

	C1	C2
Var	Observed Counts Ex3	Ho Proportions Ex3
1	780	0.54
2	117	0.18
3	114	0.12
4	384	0.15
5	58	0.01

Now got to the “Statistics” menu in Statcato, click on “Multinomial Experiments” and then “Chi-Square Goodness of Fit”. We will need to enter the observed counts column under “Observed Frequencies”. Under “Expected Frequencies” click on “Unequal Frequencies” and then “Probabilities in Column”. Enter the column that has the proportions for the null hypothesis.



Chi-Square Goodness-of-Fit Test:

Input: C1 Observed Counts Ex3

Expected probabilities in C2 Ho Proportions Ex3

Category	Observed Frequency	Expected Frequency	Contribution to X^2
0	780.0	784.62	0.0272
1	117.0	261.5400	79.8800
2	114.0	174.3600	20.8954
3	384.0	217.95	126.5088
4	58.0	14.5300	130.0510

N	Number of Categories	DOF	Significance	Critical Value	Test statistics	p-Value
1453.0	5	4	0.01	13.2767	357.3625	0

Notice the test statistic and P-value are the same as we calculated with StatKey.

Problems Section 4D

(#1-10) Use each of the following Goodness of Fit χ^2-test statistics and the corresponding critical values to fill out the table.

	χ^2-test stat	Sentence to explain χ^2-test statistic.	Critical Value	Does the χ^2-test statistic fall in a tail determined by the critical value? (Yes or No)	Does sample data significantly disagree with H_0?
1.	+28.573		+9.117		
2.	+1.226		+7.113		
3.	+2.137		+5.521		
4.	+14.415		+6.114		
5.	+3.718		+7.182		
6.	+0.891		+3.994		
7.	+51.652		+14.881		
8.	+1.185		+4.181		
9.	+2.442		+8.619		
10.	+14.133		+10.336		

(#11-20) Use each of the following P-values and corresponding significance levels to fill out the table.

	P-value Proportion	P-value %	Sentence to explain the P-value	Significance Level %	Significance level Proportion	If H_0 is true, could the sample data occur by random chance or is it unlikely?	Reject H_0 or Fail to reject H_0?
11.	0.0006			10%			
12.	0.042			1%			
13.	9.16×10^{-7}			5%			
14.	0.739			1%			
15.	0.0035			5%			
16.	0			10%			
17.	0.419			5%			
18.	0.0274			10%			
19.	3.77×10^{-5}			1%			
20.	0.067			5%			

21. How is the degrees of freedom calculated in a Goodness of Fit test?

22. The $\boldsymbol{\chi^2}$**-**test statistic compares the observed sample counts to the expected counts from $\boldsymbol{H_0}$. Explain how the expected counts are calculated.

23. Explain how the $\boldsymbol{\chi^2}$**-**test statistic is calculated from the observed and expected counts.

24. If the observed sample counts were significantly different from the expected counts, would the χ^2-test statistic be large or small? Explain why.
25. If the observed sample counts were close to the expected counts, would the χ^2-test statistic be large or small? Explain why.

(#26-29) Directions: Use StatKey at www.lock5stat.com to simulate the following Chi-squared Goodness of Fit tests. Go to "more advanced randomization tests" at the bottom of the StatKey page. Click on the button that says "χ^2 Goodness of Fit". Under "Edit Data", type in the given sample data. Create a randomized simulation of the null hypothesis to answer the following questions.

a) Write the null and alternative hypothesis. Include relationship implications.

b) What is the degrees of freedom?

c) What is the Chi-squared test statistic? Write a sentence to explain the test statistic.

d) Adjust the right tail of your simulation to reflect the significance level. Did the Chi-squared test statistic fall in the tail?

e) Does the sample data significantly disagrees with the null hypothesis? Explain your answer.

f) Are the observed counts in the sample data significantly different from the expected counts from the null hypothesis? Explain your answer.

g) Put the Chi-squared test statistic into the bottom box in the right tail of your simulation in order to calculate the P-value. What was the P-value? (Answers will vary.) Write a sentence to explain the P-value.

h) Use the P-value and significance level to determine if the sample data could have occurred by random chance (sampling variability) or is it unlikely to random chance? Explain your answer.

i) Should we reject the null hypothesis or fail to reject the null hypothesis? Explain your answer.

j) Write a conclusion for the hypothesis test. Explain your conclusion in plain language.

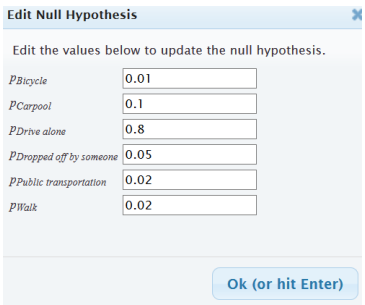
k) Is the population proportion related to the categorical variable or not? Explain your answer.

26. It is a big job to write and grade the AP-statistics exam for high school students each year. It is a difficult multiple-choice exam. All questions have five possible answers A-E. Use a 5% significance level and the following sample data to test the claim that percent of A answers is the same as the percent of B answers which is the same as C, D and E. This would indicate that the letter of the answer is not related to the percentage of times it happens. You can assume that the sample data meets the assumptions. Type the following sample data under the "Edit Data" menu of StatKey.

Choice, Count
A, 85
B, 90
C, 79
D, 78
E, 68

27. We collected data from all of the math 140 statistics students in the fall 2015 semester. A person that works at COC thinks that 80% of COC students drive alone, 10% carpool, 5% are dropped off by someone, 2% walk, 1% bike, and 2% use public transportation. Use a 5% significance level and the following sample data to test the claim that these percentages are wrong. You can assume that the data meets the assumptions for inference. Type in the proportions under "Null Hypothesis" in StatKey. Under "Edit Data" type in the following sample data from the fall 2015 survey data.

Choice, Count
Bicycle, 1
Carpool, 30
Drive Alone, 267
Dropped Off, 18
Public Transportation, 6
Walk, 10



28. We collected data from all of the math 140 statistics students in the fall 2015 semester. Use a randomized simulation in StatKey, a 5% significance level, and the following sample data to test the claim that the population percentages for the different political parties are different. This would indicate that the political party is related to the population percentages. You can assume that the data meets the assumptions for inference. Under "Edit Data", type in the following sample data from the fall 2015 survey data.

Choice, Count
Democratic, 110
Republican, 63
Independent, 65
Other, 90

29. Juries are required to meet the racial demographic of the county they represent. Here is the racial demographic for Alameda county: 54% Caucasian, 18% African American, 12% Hispanic American, 15% Asian American, and 1% other. We are worried that the juries in Alameda County may not be representing these percentages. Use randomized simulation, a 1% significance level, and the following observed sample counts to test the claim that the juries do not represent the demographic of the county. Under the "Edit Data" menu in StatKey, type in the following sample counts.

Jury Sample Data Observed Counts

Race, Count

Caucasian, 780

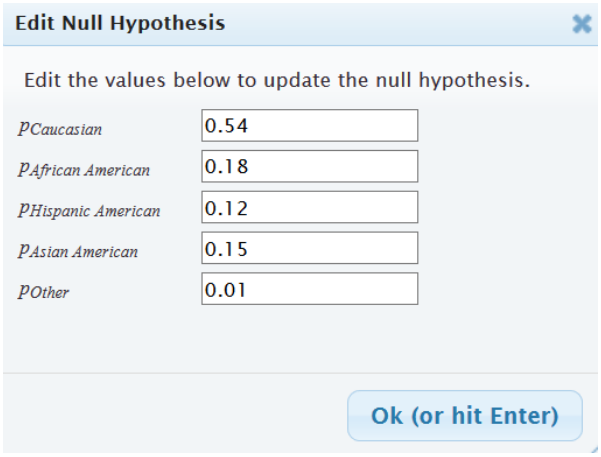
African American, 117

Hispanic American, 114

Asian American, 384

Other, 58

Under the "Null Hypothesis" menu, type in the following.



(#30-32) Directions: Use the following Statcato printouts to answer the following questions.

a) Write the null and alternative hypothesis. Include relationship implications.

b) Check the assumptions for a Goodness of Fit test.

c) What is the Chi-squared test statistic? Write a sentence to explain the test statistic.

d) Did the Chi-squared test statistic fall in the tail determined by the critical value?

e) Does the sample data significantly disagrees with the null hypothesis? Explain your answer.

f) Are the observed counts in the sample data significantly different from the expected counts from the null hypothesis? Explain your answer.

g) What was the P-value? Write a sentence to explain the P-value.

h) Use the P-value and significance level to determine if the sample data could have occurred by random chance (sampling variability) or is it unlikely to random chance? Explain your answer.

i) Should we reject the null hypothesis or fail to reject the null hypothesis? Explain your answer.

j) Write a conclusion for the hypothesis test. Explain your conclusion in plain language.

k) Is the population proportion related to the categorical variable or not? Explain your answer.

30. An online sports magazine wrote an article about the favorite sports in America. It said that 43% of Americans prefer Football, 23% of Americans prefer Baseball, 20% of Americans prefer Basketball, 8% of Americans prefer Hockey, and 6% of Americans prefer Soccer. When 130 randomly selected adults were asked their favorite sport, we found the following: 44 said Football, 26 said Baseball, 29 said Basketball, 13 said Hockey, and 18 said Soccer. Use a 5% significance level to test the claim that the proportions match the distribution claimed in the magazine article.

Chi-Square Goodness-of-Fit Test:

Input: C1 Observed Counts#4

Expected probabilities in C2 Null Hypothesis

Category	Observed Frequency	Expected Frequency	Contribution to X^2
0	44.0	55.9	2.5333
1	26.0	29.9000	0.5087
2	29.0	26.0	0.3462
3	13.0	10.4	0.6500
4	18.0	7.8	13.3385

N	Number of Categories	DOF	Significance	Critical Value	Test statistics	p-Value
130.0	5	4	0.05	9.4878	17.3766	0.0016

31. Thousands of people die from car accidents across the U.S. every year, but is the day of the week related to the probability of having a fatal car accident? To test this claim, use a 1% significance level and a Goodness of Fit test to determine if the probabilities of a fatal car accident are significantly different. The following random sample data summary gives the observed number of the number of deaths from car accidents in the U.S. for each day of a randomly selected week. The total number of deaths for the week was 805.

Day	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Number of Fatal Car Accidents	106	104	103	113	130	132	117

Chi-Square Goodness-of-Fit Test:
nput: C5 Observed #5
Expected frequency = 115.0

Category	Observed Frequency	Expected Frequency	Contribution to X^2
0	106.0	115.0	0.7043
1	104.0	115.0	1.0522
2	103.0	115.0	1.2522
3	113.0	115.0	0.0348
4	130.0	115.0	1.9565
5	132.0	115.0	2.5130
6	117.0	115.0	0.0348

N	Number of Categories	DOF	Significance	Critical Value	Test statistics	p-Value
805.0	7	6	0.01	16.8119	7.5478	0.2731

32. The National Highway Traffic Safety Administration (NHTSA) publishes reports about motorcycle fatalities and helmet use. The following distribution shows the proportion of fatalities by location of injury for motorcycle accidents.

Location of Injury	Multiple Locations	Head	Neck	Thorax	Abdomen/Spine
Proportion	0.57	0.31	0.03	0.06	0.03

The random sample data below shows the distribution of 2068 randomly selected fatalities from riders that were not wearing a helmet. Use a 0.01 significance level to test the claim that the distribution for the sample does not match the proportions given by the NHTSA. Where is the largest discrepancy between the observed and expected value? What does this tell us about the importance of wearing helmets?

Location of Injury	Multiple Locations	Head	Neck	Thorax	Abdomen/Spine
Number of Deaths	1036	864	38	83	47

Chi-Square Goodness-of-Fit Test:
Input: C3 observed#6
Expected probabilities in C4 Ho#6

Category	Observed Frequency	Expected Frequency	Contribution to X^2
0	1036.0	1178.76	17.2897
1	864.0	641.08	77.5150
2	38.0	62.04	9.3153
3	83.0	124.08	13.6006
4	47.0	62.04	3.6461

N	Number of Categories	DOF	Significance	Critical Value	Test statistics	p-Value
2068.0	5	4	0.01	13.2767	121.3667	0

Section 4E – Categorical Relationships: Contingency Tables

Vocabulary

Categorical data: Also called qualitative data. Data in the form of labels that tell us something about the people or objects in the data set. For example, the country they live in, occupation, or type of pet.

Contingency Table: Also called a two-way table. This table summarizes the counts when comparing two different categorical data sets each with two or more variables.

Marginal Percentage (Marginal Proportion): A single percentage or proportion without any conditions. In a contingency table, this can found with numbers in the margins.

Conditional Percentage (Conditional Proportion): The percentage or proportion calculated from a particular group or if a particular condition was true. These are the very important when studying categorical relationships.

Joint Percentage (Joint Proportion): A percentage or proportion involving two variables being true about the person or object, but does not have a condition. There are generally two types (AND, OR).

Introduction

An important field of exploration when analyzing data is the study of relationships between variables. A lot of thought has been put into determining which variables have relationships and the scope of that relationship. Is a person's diet related to having high blood pressure? Is the city a person lives in related to whether or not they have tuberculosis? Is being in a car accident related to texting while driving? These are all important questions that statisticians, data analysts and data scientists explore.

Relationships can be categorical ⇔ categorical, categorical ⇔ quantitative, and quantitative ⇔ quantitative. In this chapter, we will begin to explore the relationships between two categorical variables.

Remember, statistics is a deep well of mathematics and knowledge learned by years of study. There are much more advanced techniques for studying relationships, but we will be focusing on a basic introduction to the topic. You will find that a good understanding of this chapter will help tremendously when you go on to the more advanced techniques later on. For example, I find my students have many problems understanding the Chi-Squared distribution because they lack the foundational understanding of contingency (two-way) tables and analyzing differences between categories.

Note on Terminology: *When studying relationships between variables you will hear different words used to describe the relationship. The most common are "relationship", "association", or "correlation". "Correlation" is often used for describe a relationship between two quantitative variables (quantitative ⇔ quantitative), while "relationship" and "association" are used for two categorical variables (categorical ⇔ categorical) or for a categorical - quantitative relationship study (categorical ⇔ quantitative).*

In this chapter, we will be using the terms "relationship" or "association".

Note on Causation: *One of the most famous statements in statistics is that "correlation is not causation". Proving that one thing causes another is a much more complex kind of study and involves controlling confounding variables and experimental design. Remember that just because there is a relationship, that does not prove causation. There may be many other factors involved.*

To analyze categorical data we need to know the counts (frequencies) for each categorical variable. This particularly important when you are studying categorical relationships. No data scientist or statistician finds the frequencies by hand. They use computer programs to make a contingency table (or two-way table).

Creating a contingency table with raw data and StatKey

Let us look at an example. Go to www.matt-teachout.org and click on the math 140 survey data fall 2015. We want to explore the relationship between the campus a person goes to and their political party.

First, we will need to check the data. When exploring relationships between two data sets, the data needs to be ordered pair. This usually means the data came from the same people. We also need to be careful of blanks. This means a person did not answer one or both of the questions. Start by copy and pasting the campus data and political party data into a fresh excel spreadsheet. A good rule of thumb is never mess up an original data set. Always copy and paste into a new excel file if you want to change things. The two columns of categorical data need to be in next to each other in the new Excel sheet. Otherwise, StatKey will not accept it. Go through the data and make sure there are no blanks. If there is a blank, delete that entire row. If you remember from chapter 1, this is called non-response bias. This process of deleting out missing cells is sometimes called "cleaning the data".

To make a contingency table with StatKey, go to www.lock5stat.com and click the "StatKey" button. Now click on "Two Categorical Variables" under the "Descriptive Statistics and Graphs" menu. Then click on the "edit data" button. Copy both columns together in your excel spreadsheet and paste them into StatKey. Check the "raw data" box and the "data has header row" box and push "OK".

Counts Table Switch Variables

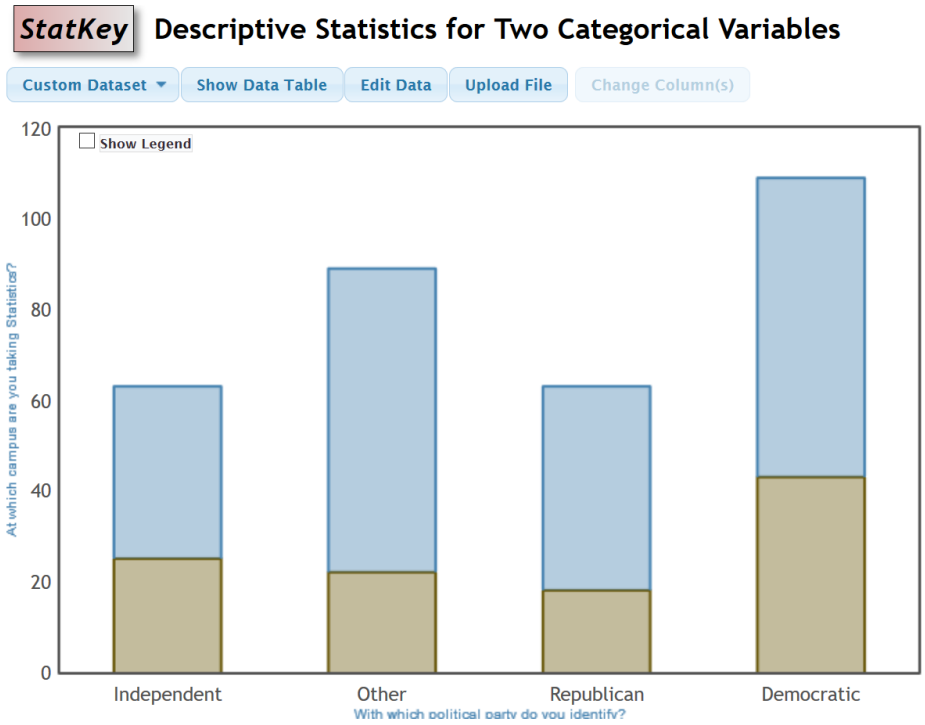
At which campus are you taking Statistics? \ With which political party do you identify?	Independent	Other	Republican	Democratic	Total
Canyon Country Campus	25	22	18	43	108
Valencia Campus	38	67	45	66	216
Total	63	89	63	109	324

Proportions Row Column Overall

This is called a contingency table. Notice that we a lot of frequency information. Notice 66 is in the Democratic column and Valencia row, so 66 Math 140 students are both Democrat and go to the Valencia campus. Similarly, 18 math 140 students are both Republican and go to the Canyon Country campus.

The size of a contingency table is the number of rows by the number of columns. Totals are not included. This table has two rows (CCC and Valencia) and four columns (Independent, Other, Republican, and Democratic), so this is a "2 by 4" or "2×4" contingency table.

StatKey has several cool features with the contingency table. Notice it has created a stacked bar chart. This graph gives a visual representation of a contingency table. Notice if you place your curser on any section of the graph the corresponding count lights up in the contingency table.



The "proportion" buttons are particularly useful. If we click on the "overall" proportion button. The computer calculates the intersection (AND) percentages for the entire data set. If we click on the "row" proportion button it gives conditional percentages for the rows. If we click on the "column" proportion button it gives the conditional percentages for the columns. We will discuss these more later, but these are very useful.

Proportions Row | Column | Overall

At which campus are you taking Statistics? \ With which political party do you identify?	Independent	Other	Republican	Democratic	Total
Canyon Country Campus	0.077	0.068	0.056	0.133	0.333
Valencia Campus	0.117	0.207	0.139	0.204	0.667
Total	0.194	0.275	0.194	0.336	1

Proportions Row | Column | Overall

At which campus are you taking Statistics? \ With which political party do you identify?	Independent	Other	Republican	Democratic	Total
Canyon Country Campus	0.231	0.204	0.167	0.398	1
Valencia Campus	0.176	0.31	0.208	0.306	1
Total	0.194	0.275	0.194	0.336	1

Proportions Row | Column | Overall

At which campus are you taking Statistics? \ With which political party do you identify?	Independent	Other	Republican	Democratic	Total
Canyon Country Campus	0.397	0.247	0.286	0.394	0.333
Valencia Campus	0.603	0.753	0.714	0.606	0.667
Total	1	1	1	1	1

Another feature is the "switch variables" button. Clicking on this button will switch the rows and columns.

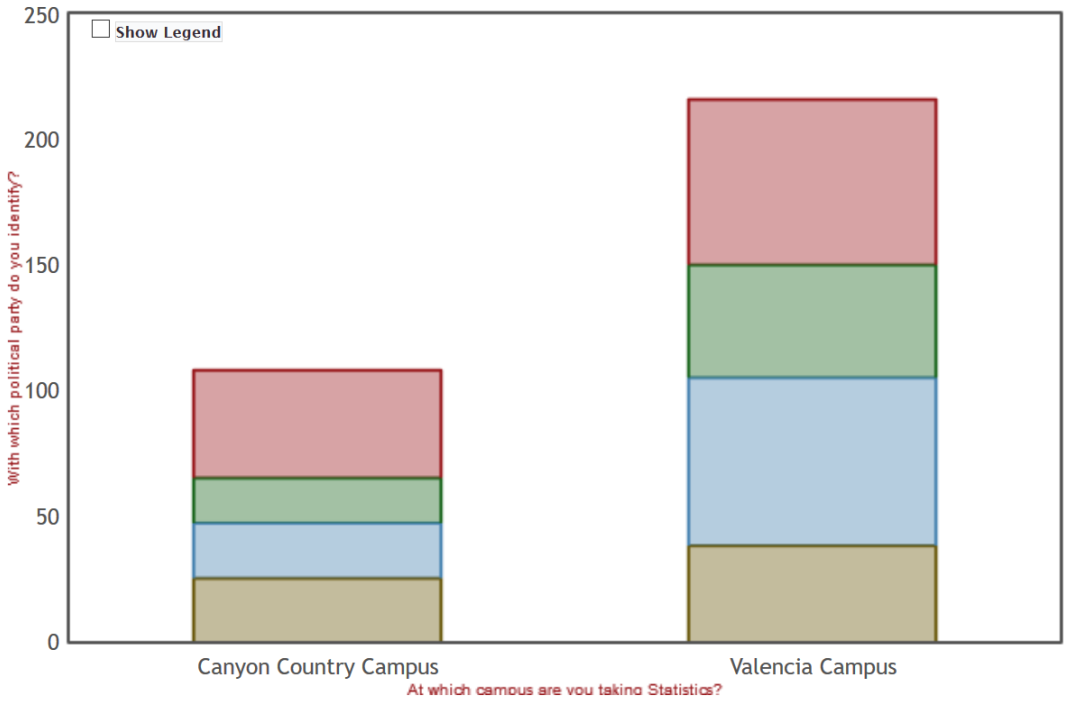
Counts Table Switch Variables

With which political party do you identify? \ At which campus are you taking Statistics?	Canyon Country Campus	Valencia Campus	Total
Independent	25	38	63
Other	22	67	89
Republican	18	45	63
Democratic	43	66	109
Total	108	216	324

Proportions Row | Column | Overall

StatKey **Descriptive Statistics for Two Categorical Variables**

Custom Dataset | Show Data Table | Edit Data | Upload File | Change Column(s)



Notice that you can click on any section in the graph and it will highlight the count it came from in the contingency table. In addition, you can click on the proportion buttons to calculate and compare various proportions.

Creating a contingency table with summary counts and StatKey

Let us look at the same example again. As we said in the last section, categorical data is often not given in raw form. Sometimes a person may give you the summary counts (frequencies). In that case, you already have the contingency table, yet it is good to be able to put that into StatKey to create the stacked bar chart and use the switch variable and proportion features. To put in a contingency table into StatKey, go to www.lock5stat.com and click the "StatKey" button. Now click on "Two Categorical Variables" under the "Descriptive Statistics and Graphs" menu. Then click on the "edit data" button. Type in the table as seen below. Note that there should be a space after every comma and the totals are not included. There should also be a "[blank]" in the upper left corner. Uncheck the "raw data" box and check the "data has header row" box and push "OK". Notice this gives us the exact same table and graphs as if we had used the raw data.

[blank], Independent, Other, Republican, Democratic
Canyon Country Campus, 25, 22, 18, 43
Valencia Campus, 38, 67, 45, 66

Creating a contingency table with raw data and Statcato

You can also create a contingency table with Statcato. Copy and paste the ordered pair categorical data into a fresh excel spreadsheet. Make sure to clean the data and delete out any rows with missing values. Since this data set is over 300 values, go to the "edit" menu, "add multiple rows" and add another 100 rows. When that is done, copy and paste the two columns one at a time into Statcato. Statcato does not copy and paste multiple columns at the same time very well. It is best to copy and paste one at a time. Now go to the "Statistics" menu and click on "Multinomial Experiments". Now click on "Cross Tabulation and Chi-Square". Pick one column of data to be the row and the other column of data as the column. Uncheck the box that says, "Perform chi-squared test". That is a more advanced analysis. Also, do not click on anything under the "frequency (optional)" menu. Now push "OK".

Statistics => Multinomial Experiments => Cross Tabulation => OK

If we use the campus and political party data from the previous example, we get the following from Statcato. Notice it gives the counts (frequencies), totals (All), and the intersection percentages (AND).

Cross Tabulation and Chi-Square

rows in C1 At which campus ..., columns in C2 With which polit...

	Democratic	Independent	Other	Republican	All
Canyon Country Campus	43.0 (13.27%)	25.0 (7.72%)	22.0 (6.79%)	18.0 (5.56%)	108.0 (33.33%)
Valencia Campus	66.0 (20.37%)	38.0 (11.73%)	67.0 (20.68%)	45.0 (13.89%)	216.0 (66.67%)
All	109.0 (33.64%)	63.0 (19.44%)	89.0 (27.47%)	63.0 (19.44%)	324.0 (100.00%)

Calculating Marginal Percentages

Marginal Percentages are percentages that involve only one variable and do not have a condition. They get their name because the amount and total are found in the margins (totals). Let us look at a couple examples. Remember a proportion and percentage can be found from the amount (frequency) and the total.

Proportion = $\frac{Amount\ (Frequency)}{Total}$

Percentage = $\frac{Amount\ (Frequency)}{Total} \times 100\%$

Example: Find the proportion and percentage of the math 140 students are democrat. Notice we need to find the amount of democrats and the total number of students. The amount of democrats will be in total part of the democrat row or column. The total number of students is often called the grand total and is found in the bottom right of the table.

Proportion = $\frac{Amount\ (Frequency)}{Total} = \frac{109}{324} \approx 0.336$

Percentage = proportion × 100% = 0.336 × 100% = 33.6%

It is always better to use technology when you can instead of calculating something by hand. We could had found the proportion with StatKey by clicking on the "overall" proportion button. Statcato had this percentage already calculated as well. Notice the democrat data is summarized as a column. In both StatKey and Statcato, we need to look at the total in the democratic column to get the proportion. We can then convert the answer into a percentage or proportion as needed.

Proportions Row | Column | Overall

At which campus are you taking Statistics? \ With which political party do you identify?	Independent	Other	Republican	Democratic	Total
Canyon Country Campus	0.077	0.068	0.056	0.133	0.333
Valencia Campus	0.117	0.207	0.139	0.204	0.667
Total	0.194	0.275	0.194	0.336	1

Cross Tabulation and Chi-Square

rows in C1 At which campus ..., columns in C2 With which polit...

	Democratic	Independent	Other	Republican	All
Canyon Country Campus	43.0 (13.27%)	25.0 (7.72%)	22.0 (6.79%)	18.0 (5.56%)	108.0 (33.33%)
Valencia Campus	66.0 (20.37%)	38.0 (11.73%)	67.0 (20.68%)	45.0 (13.89%)	216.0 (66.67%)
All	109.0 (33.64%)	63.0 (19.44%)	89.0 (27.47%)	63.0 (19.44%)	324.0 (100.00%)

Example: Use the tables above to give the proportion and percentage of the Math 140 students that attended the Canyon Country campus. Look for the Canyon Country campus data. Notice it is in the first row. So the number we are looking for is at the end of the first row under "total" or "All".

Proportion of Math 140 students at the Canyon Country campus ≈ 0.333
Percentage of Math 140 students at the Canyon Country campus ≈ 33.3%

Calculating Joint Percentages

There are two types of joint percentages. The first type is the percentage of the total that has two things true about the person. We often call this the intersecting percentage or "AND". The second type is the proportion or percentage of the total that has either one of two things true about the person. This is sometimes called the union percentage or "OR". Intersecting percentages means that both things must be true about the person or object. Let us look at a few examples. Remember a proportion and percentage can be found from the amount (frequency) and the total.

Example: Find the proportion and percentage of the math 140 students that are both democrat AND attend the Valencia campus. Both things must be true about the person. In an "AND" (intersection) proportion, the amount can be found in the cell where the column and row meet. We will still use the "grand total" in the lower right corner as the total, since we need to include everyone in the data set. Look at the where the democratic column meets the Valencia row. There are 66 students that have both characteristics. This is the amount we need. The grand total is still 324 so here is the proportion and percentage calculation. Round your answer to three significant figures.

"AND" Proportion = $\frac{Frequency\ in\ intersection\ cell}{Grand\ Total} = \frac{66}{324} \approx 0.2037037 \approx 0.204$

"AND" Percentage = proportion × 100% = 0.204 × 100% = 20.4%

Again, we could have used technology to get that answer. We could had found the proportion with StatKey by clicking on the "overall" proportion button. Statcato had this percentage already calculated as well. Both times, we need to look at the cell where the Democratic column meets the Valencia row.

Proportions Row | Column | Overall

At which campus are you taking Statistics? \ With which political party do you identify?	Independent	Other	Republican	Democratic	Total
Canyon Country Campus	0.077	0.068	0.056	0.133	0.333
Valencia Campus	0.117	0.207	0.139	0.204	0.667
Total	0.194	0.275	0.194	0.336	1

Cross Tabulation and Chi-Square

rows in C1 At which campus ..., columns in C2 With which polit...

	Democratic	Independent	Other	Republican	All
Canyon Country Campus	43.0 (13.27%)	25.0 (7.72%)	22.0 (6.79%)	18.0 (5.56%)	108.0 (33.33%)
Valencia Campus	66.0 (20.37%)	38.0 (11.73%)	67.0 (20.68%)	45.0 (13.89%)	216.0 (66.67%)
All	109.0 (33.64%)	63.0 (19.44%)	89.0 (27.47%)	63.0 (19.44%)	324.0 (100.00%)

Example: Use the tables above to give the proportion and percentage of the Math 140 students that both attend the Canyon Country campus AND are Republican. Look for where the Canyon Country campus row meets the Republican Column.

Proportion of Math 140 students at the Canyon Country campus AND Republican ≈ 0.056
Percentage of Math 140 students at the Canyon Country campus AND Republican ≈ 5.6%

Example: Now calculate the proportion and percentage of Math 140 students that either are at the Valencia campus OR are Democratic. This means we need to include anyone that was Democrat regardless of campus and include anyone at the Valencia campus regardless of the political affiliation. This is a more difficult calculation. Here is a couple common formulas for "OR" (union) percentages.

"OR" (Union) Proportion = $\frac{(Row\ Total + Column\ Total - Intersection\ Cell)}{Grand\ Total} = \frac{(216 + 109 - 66)}{324} = \frac{(259)}{324} \approx 0.79938 \approx 0.799$

It is better to use technology if we can. StatKey and Statcato printouts can help us calculate the "OR" (union) proportion or percentage.

"OR" (Union) Proportion = Row Total Proportion + Column Total Proportion − Intersection Cell Proportion

Notice these proportions are given in the StatKey table we can use them to calculate the "OR" proportion.

Proportions Row | Column | Overall

At which campus are you taking Statistics? \ With which political party do you identify?	Independent	Other	Republican	Democratic	Total
Canyon Country Campus	0.077	0.068	0.056	0.133	0.333
Valencia Campus	0.117	0.207	0.139	0.204	0.667
Total	0.194	0.275	0.194	0.336	1

"OR" (Union) Proportion = Row Total Proportion (Valencia) + Column Total Proportion (Democratic) − Intersection Cell Proportion (where Valencia and Democratic meet)

= 0.667 + 0.336 – 0.204 = 0.799

(We can convert this proportion to a percentage if needed. Percent = Proportion × 100% ≈ 0.799 × 100% ≈ 77.9%)

"OR" (Union) Percentage = Row Total % + Column Total % − Intersection Cell %

Notice these percentages are given in the Statcato table we can use them to calculate the "OR" percentage.

Cross Tabulation and Chi-Square

rows in C1 At which campus ..., columns in C2 With which polit...

	Democratic	Independent	Other	Republican	All
Canyon Country Campus	43.0 (13.27%)	25.0 (7.72%)	22.0 (6.79%)	18.0 (5.56%)	108.0 (33.33%)
Valencia Campus	66.0 (20.37%)	38.0 (11.73%)	67.0 (20.68%)	45.0 (13.89%)	216.0 (66.67%)
All	109.0 (33.64%)	63.0 (19.44%)	89.0 (27.47%)	63.0 (19.44%)	324.0 (100.00%)

"OR" (Union) Percentage = Row Total % (Valencia) + Column Total % (Democratic) − Intersection Cell % (where Valencia and Democratic meet) = 66.7% + 33.6% − 20.4% = 79.9%

Conditional Proportions and Percentages

Conditional proportions and percentages are the key to understanding categorical relationships. A condition is thought of as prior knowledge about the person or situation that may change the percentage. Let us say that the Los Angeles Lakers have a 75% chance of beating the Phoenix Suns. If the Lakers best player LeBron James does not play, will that change the percentage? Of course. Knowing that LeBron James will not play is called a condition.

In contingency tables, a condition involves restricting to one particular group before you calculate the percentage.

Example: What percentage of the Canyon Country campus Math 140 students are Democrat?

First notice that this is not a joint proportion. It does NOT ask for the percentage of all students that are both Democrat and go to the Canyon Country campus.

The key is to identify which group we are restricting ourselves to. In other words, what is the condition? Look for words that say "if" or "given this is true" or "out of". This designates the condition. In this example, notice that the problem said "of the Canyon Country students". That means that we are supposed to only look at the Canyon Country students when we find our amount (frequency) and total. A commonly used method for calculating conditional percentages from a contingency table is to circle the row or column that has your condition (Canyon Country). Then only use numbers in that row or column.

Counts Table Switch Variables

At which campus are you taking Statistics? \ With which political party do you identify?	Independent	Other	Republican	Democratic	Total
Canyon Country Campus	25	22	18	43	108
Valencia Campus	38	67	45	66	216
Total	63	89	63	109	324

Proportions Row Column Overall

Notice that the Canyon Country Campus counts are in the first row. So we should only use numbers in the first row. We should not use the grand total anymore. We need the total number of students that attend the Canyon Country campus. In other words, the total from our condition. The amount will be the number of democrats in the Canyon Country row. In other words the intersection cell frequency.

Conditional Proportion = $\frac{Amount\ in\ Intersection\ Cell\ (Canyon\ Country\ meets\ Democratic)}{Row\ or\ Column\ Total\ (Row\ total\ Canyon\ Country)} = \frac{43}{108} \approx 0.398148 \approx 0.398$

We can use the "row" and "column" proportion buttons in StatKey to find this conditional proportion. Since the condition is a row, we should click the "row" proportion button.

Proportions Row Column Overall

At which campus are you taking Statistics? \ With which political party do you identify?	Independent	Other	Republican	Democratic	Total
Canyon Country Campus	0.231	0.204	0.167	0.398	1
Valencia Campus	0.176	0.31	0.208	0.306	1
Total	0.194	0.275	0.194	0.336	1

Notice the answer we are looking for is given in the intersecting cell. If we restrict ourselves to considering only the Canyon Country students, 0.398 or 39.8% of them are democrat.

Example: What proportion of the republican math 140 students attend the Valencia campus? To answer this we need to recognize that we are no longer considering all the students. We are restricting our proportion to considering only the republican students ("out of"). Since the condition is being republican, we should only use numbers in the republican column. The total will now be the total number of republicans and the amount will be the amount of republicans that attend the Valencia campus.

Counts Table Switch Variables

At which campus are you taking Statistics? \ With which political party do you identify?	Independent	Other	Republican	Democratic	Total
Canyon Country Campus	25	22	18	43	108
Valencia Campus	38	67	45	66	216
Total	63	89	63	109	324

Proportions Row Column Overall

Conditional Proportion = $\frac{Amount\ in\ Intersection\ Cell\ (Republican\ meets\ Valencia)}{Row\ or\ Column\ Total\ (column\ total\ Republican)} = \frac{45}{63} \approx 0.7142857 \approx 0.714$

We can also use StatKey to find what proportion of Republican Math 140 students attend the Valencia campus. Notice our condition is now republican ("out of"). This is a column so I will click the "column" proportion button in StatKey.

Proportions Row | Column | Overall

At which campus are you taking Statistics? \ With which political party do you identify?	Independent	Other	Republican	Democratic	Total
Canyon Country Campus	0.397	0.247	0.286	0.394	0.333
Valencia Campus	0.603	0.753	0.714	0.606	0.667
Total	1	1	1	1	1

Notice now we want to restrict ourselves to the Republican column. The conditional proportion we are looking for is 0.714 or 71.4%.

Relationship Principle

Let us go back to the LeBron James example. The key to understanding categorical relationships is to judge how close or far apart conditional percentages are.

Chances of Lakers winning if LeBron James plays ≈ 75%
Chances of Lakers winning if LeBron James does not play ≈ 40%

These percentages are significantly different, so it tells us that the condition of LeBron James playing in the game is related to the Lakers winning.

Let us look at another example using the Lakers chances of beating the Phoenix Suns.

Chances of Lakers winning if it snows in Nebraska ≈ 75%
Chances of Lakers winning if it does not snow in Nebraska ≈ 75%

These percentages are not significantly different, so it tells us that the condition of snowing in Nebraska is not related to the Lakers winning. The condition does not matter.

Relationship Principle:

Close Conditional Percentages = Condition is NOT related to the categorical variable

Significantly Different Conditional Percentages = Condition IS related to the categorical variable

Note: You cannot compare any conditional percentages you want. They must be the same variable for the percentage and from different groups (different condition). You cannot compare the percentage of republicans from the Canyon Country campus to the percentage of democrats from the Valencia campus. They are not the same thing and will likely have very different percentages regardless of the relationship. Compare the percentage of republicans from the Canyon Country campus to the percentage of republicans from the Valencia campus. That will give us information about the relationship. Conditional percentage analysis is the basis behind the Chi-Squared test statistic we will learn in chapter 5.

Practice Problems Section 4E

1. If the proportions for a categorical variable from one group are significantly different from another group, what does that indicate about the relationship between that variable and the groups?

2. If the proportions for a categorical variable from one group are almost the same as another group, what does that indicate about the relationship between that variable and the groups?

(#3-10) Open the math 140 fall 2015 survey data at www.matt-teachout.org. Copy and paste the smoking status column and the type of transportation column next to each other in a new excel spread sheet. Then copy both columns together. Open StatKey at www.lock5stat.com. Under the "Descriptive Statistics and Graphs" menu, click on "Two Categorical Variables". Paste the two columns into StatKey. Be sure to check the boxes for "Raw Data" and "Header Row" and push "OK". Use StatKey to create a contingency table for smoking status and transportation. Use the table to answer the following questions.

3. What percent of the math 140 students smoke?

4. What proportion of the math 140 students drive alone to school?

5. What percent of the math 140 students both carpool and do not smoke?

6. What proportion of the math 140 students both smoke and drive alone to school?

7. What percent of the math 140 students either do not smoke or are dropped off by someone?

8. What proportion of the math 140 students either walk to school or smoke?

9. What percent of the smoking math 140 students carpool? What percent of the non-smoking math 140 students carpool? Do the proportions appear to be close or significantly different? What does this tell us about the relationship between smoking and carpooling to school?

10. What proportion of the drive alone math 140 students smoke? What proportion of the dropped off math 140 students smoke? Do the proportions appear to be close or significantly different? What does this tell us about the relationship between smoking and the type of transportation?

(#11-18) Open the math 140 fall 2015 survey data at www.matt-teachout.org. Copy and paste the texting and driving column and the car accident column next to each other in a new excel spreadsheet. Then copy both columns together. Open StatKey at www.lock5stat.com. Under the "Descriptive Statistics and Graphs" menu, click on "Two Categorical Variables". Paste the two columns into StatKey. Be sure to check the boxes for "Raw Data" and "Header Row" and push "OK". Use StatKey to create a contingency table for texting and driving and car accidents. Use the table to answer the following questions.

11. What percent of the math 140 students text and drive?

12. What proportion of the math 140 students have been in a car accident?

13. What percent of the math 140 students both text and drive and have been in a car accident?

14. What proportion of the math 140 students do not text and drive and have not been in a car accident?

15. What percent of the math 140 students either text and drive or have not been in a car accident?

16. What proportion of the math 140 students either do not text and drive or have been in a car accident?

17. What percent of the text and drive math 140 students have been in a car accident? What percent of the not text and drive math 140 students have been in a car accident? Do the proportions appear to be close or significantly different? What does this tell us about the relationship between texting and driving and car accidents?

18. What proportion of the car accident math 140 students text and drive? What proportion of the no car accident math 140 students text and drive? Do the proportions appear to be close or significantly different? What does this tell us about the relationship between texting and driving and car accidents?

(#19-26) Open the math 140 fall 2015 survey data at www.matt-teachout.org. Copy and paste the tattoos column and the favorite social media column next to each other in a new excel spread sheet. Then copy both columns together. Open StatKey at www.lock5stat.com. Under the "Descriptive Statistics and Graphs" menu, click on "Two Categorical Variables". Paste the two columns into StatKey. Be sure to check the boxes for "Raw Data" and "Header Row" and push "OK". Use StatKey to create a contingency table for tattoos and favorite social media. Use the table to answer the following questions.

19. What percent of the math 140 students have a tattoo?

20. What proportion of the math 140 students prefer snapchat?

21. What percent of the math 140 students both prefer Facebook and do not have a tattoo?

22. What proportion of the math 140 students both have a tattoo and prefer twitter?

23. What percent of the math 140 students either prefer Instagram or have a tattoo?

24. What proportion of the math 140 students either prefer twitter or do not have a tattoo?

25. What percent of the tattoo math 140 students prefer twitter? What percent of the no tattoo math 140 students prefer twitter? Do the proportions appear to be close or significantly different? What does this tell us about the relationship between liking twitter and having a tattoo?

26. What proportion of the Instagram math 140 students have a tattoo? What proportion of the Facebook math 140 students have a tattoo? Do the proportions appear to be close or significantly different? What does this tell us about the relationship between social media and having a tattoo?

(#27-34) Open the car data at www.matt-teachout.org. Copy and paste the country column and the cylinders column next to each other in a new excel spread sheet. Then copy both columns together. Open StatKey at www.lock5stat.com. Under the "Descriptive Statistics and Graphs" menu, click on "Two Categorical Variables". Paste the two columns into StatKey. Be sure to check the boxes for "Raw Data" and "Header Row" and push "OK". Use StatKey to create a contingency table for the country and cylinders. Use the table to answer the following questions.

27. What percent of the cars were made in Germany?

28. What proportion of the cars have six cylinders?

29. What percent of the cars have four cylinders and are made in Japan?

30. What proportion of the cars have eight cylinders and are made in the U.S.?

31. What percent of the cars either have six cylinders or are made in Germany?

32. What proportion of the cars either have six cylinders or are made in the Japan?

33. What proportion of the cars made in Japan have four cylinders? What proportion of cars made in Germany have four cylinders? Are the proportions appear to be close or significantly different? What does this tell us about the relationship between the country and the number of cylinders?

34. What proportion of the cars with six cylinders were made in the U.S.A? What proportion of the cars with eight cylinders were made in the U.S.A? Are the proportions appear to be close or significantly different? What does this tell us about the relationship between cars made in the U.S.A and the number of cylinders?

Section 4F – Categorical Relationships: Categorical Association Test

In this chapter, we looked at the Goodness of Fit test. The Goodness of Fit Test determines if a single proportion is related to some other categorical variable. In this section, we will look at situations where more than one proportion is involved. When we have multiple different proportions in multiple groups, we call this the Categorical Association Test.

In the Categorical Association Test, we will be determining if categorical variables are related or not. Many students confuse the Goodness of Fit test and the Categorical Association Test because they are both categorical relationship tests. Look at your sample counts. If you have a single observed count for each group, you are doing a Goodness of Fit test since we are only looking at one proportion in the groups. If your observed counts are summarized in a contingency table, then more than one proportion is involved in your groups. That makes it a Categorical Association Test.

For example, suppose we wanted to see if the amount of education a person has is related to their health. Notice the amount of education has multiple options and the health status has multiple options. This cannot be a Goodness of Fit test. The observed counts are summarized in a contingency (two-way) table so we will be using the Categorical Association Test.

	Excellent Health	Good Health	Fair Health	Poor Health
Less than High School	72	202	199	62
High School Diploma	465	877	358	108
Some College/Associates Degree	80	138	49	11
Bachelor's Degree	229	276	64	12
Graduate Degree	130	147	32	2

A Goodness of Fit Test would only look at one of these. For example, suppose we want to see if the proportion for excellent health is related to education. In that case, the data would look like this.

	Excellent Health
Less than High School	72
High School Diploma	465
Some College/Associates Degree	80
Bachelor's Degree	229
Graduate Degree	130

The Categorical Association Test

So let us look at the categorical association test. This test determines if categories are related or not. The categories can have multiple options. The sample data for this test is either two raw categorical data sets or summary counts summarized in a contingency (two-way) table.

Null and Alternative Hypothesis

In the last section, we examined conditional proportions. We saw that we need to compare conditional proportions from the same variable. If the conditional proportions are equal or close in our groups, it indicates that the categorical variables are not related. If the conditional proportions are significantly different, then the categorical variables are related.

For example, we will want to compare the proportions for excellent health in all of our education groups. We will also want to compare the proportions for poor health in all of our education groups and so on. We will not want to compare the proportion of excellent health to poor health since they are not the same variable.

	Excellent Health	Good Health	Fair Health	Poor Health
Less than High School	72	202	199	62
High School Diploma	465	877	358	108
Some College/Associates Degree	80	138	49	11
Bachelor's Degree	229	276	64	12
Graduate Degree	130	147	32	2

If there is no relationship between the categories, we expect the conditional proportions for each variable to be equal in all the groups. If there is a relationship between the categories, we expect at least one or more of the conditional proportions for each variable to be different. It is difficult to specify all of the conditional proportions and groups, so to summarize, we often say that the "distribution of conditional proportions are the same" or the "distribution of conditional proportions are different".

Note

- Saying that the categories are "related" can also be described as "associated" or "dependent".
- Saying that the categories are "not related" can also be described as "not associated" or "independent".

Categorical Association Test Null and Alternative Hypothesis

H_0 : The categories are not related (distribution of conditional proportions are equal)
H_A : The categories are related (distribution of conditional proportion are different)

The Chi-Squared Test Statistic (χ^2)

Since multiple proportions in multiple groups are involved, we will be using the Chi-Squared test statistic (χ^2) again. In our previous study of using the Chi-squared test statistic, we saw that this test statistic compares the observed sample counts to the expected counts based on the null hypothesis.

Chi-Squared Test Statistic $(\chi^2) = \sum \frac{(O-E)^2}{E}$

The expected counts are what we expect to happen if the null hypothesis is true. If the null hypothesis is true and there is no relationship between the categories, we expect the conditional proportions to be equal in the various groups. If we multiply the equal proportions by the size of the group, we can get our expected counts. Here is a formula that computer programs use to calculate the expected counts.

Expected Counts (for contingency table) = $\frac{Row\ Total \times Column\ Total}{Grand\ Total}$

How does this formula give us the expected counts and account for equal conditional proportions? To answer this, we will look at an example.

Example 1

A sample of 75 people were paid to participate in an experiment. The goal of the experiment was to determine if listening to music is related to a person's ability to memorize information. The people were randomly assigned into three groups. One group tried to memorize some information while listening to their favorite music. Another group tried to memorize some information while listening to music they hated. The third group tried to memorize some information in a silent room. All of the people attempted to memorize the same information and took a test to determine how much of the information they remembered. Confounding variables were controlled. For example, the volume of music was the same in the music groups. We will use a 10% significance level.

	Liked Music	**Disliked Music**	**No Music**	*Total*
High Retention	10	11	18	39
Low Retention	14	15	7	36
Total	24	26	25	*Grand Total = 75*

H_0 : Listening to music and memorizing information are not related (not associated, independent).
H_A : Listening to music and memorizing information are related (associated, dependent). CLAIM

Remember conditional proportions are important to explore when analyzing relationships between categorical variables. Here are two very important principles.

1. When conditional proportions were close or equal, it indicated that the variables were not related to each other.

2. When conditional proportions were significantly different, it indicated that the variables were related to each other.

Let us look at some conditional proportions. Remember we need to compare the same variable proportion in different groups.

Let us calculate the proportion of people in the liked music group that were able to memorize a lot of the information? Let us compare that to the proportion of people in the hated music group that were able to memorize a lot of the information.

P(high retention | liked music) = 10/24 = 0.417 or 41.7%
P(high retention | disliked music) = 11/26 = 0.423 or 42.3%

These two conditional probabilities are close and so indicate that the music and high retention are not related or independent.

Here lies the fundamental problem. We are not really taking the entire contingency (two-way) table and all of the conditional probabilities into account. If we look at another conditional probability, we may come to a different conclusion. Look at these two.

Let us calculate the proportion of people in the liked music group that were able to memorize a lot of the information? Let us compare that to the proportion of people in the no music group that were able to memorize a lot of the information.

P(high retention | liked music) = 10/24 = 0.417 or 41.7%
P(high retention | no music) = 18/25 = 72%

These two probabilities are significantly different and so indicate that the music and high retention are related.

So it is difficult to determine if categorical variables are related or not by just looking at two conditional proportions. We need a better way to do this.

Calculating the Chi-Squared Test Statistic

A much better way to determine if the categories are related or not is by using the chi-squared test statistic. It takes into account all of the conditional probabilities possible instead of relying on only two. Remember to calculate the chi-squared test statistic, we need to compare the expected counts (expected frequencies) to the observed counts (observed frequencies).

Expected Counts

The "expected counts" or "expected frequencies" are what we expect to happen if the null hypothesis is true. For the Categorical Association Test, the null hypothesis is that the categories are not related (independent). This would imply that the distribution of conditional proportions are equal.

Let us work this out for the music and retention problem.

	Liked Music	**Disliked Music**	**No Music**	*Total*
High Retention	10	11	18	39
Low Retention	14	15	7	36
Total	24	26	25	*Grand Total = 75*

If the null hypothesis is true, we expect the proportion for high retention to be the same regardless of the music choice. If we disregard music, then the proportion of high retention would be the amount of high retention (39) divided by the grand total (75). So if the null hypothesis is true and music is not related to retention, then 52% of every group should memorize a significant about of the information.

P (high retention) = 39/75 = 0.52

Remember the expected values are found by multiplying the proportion times the total number of people or objects in that group.

$E = n \times p$

Only the n is not the grand total, it is the total for each column (each music group).

If the null hypothesis is true, we expect the p for high retention to always be 0.52 and the expected values will be 0.52 x total students for each music choice.

$E_{liked\ music\ high\ retention} = n \times p = 24 \times 0.52 = 12.48$

$E_{hated\ music\ high\ retention} = n \times p = 26 \times 0.52 = 13.52$

$E_{no\ music\ high\ retention} = n \times p = 25 \times 0.52 = 13.0$

Similarly, we expect the proportion for low retention to be the same in all of the music groups. If we disregard music, then the proportion of low retention would be the amount of low retention (36) divided by the grand total (75). So if the null hypothesis is true and music is not related to retention, then 48% of every group should not be able to memorize much of the information.

P (low retention) = 36/75 = 0.48

So if the null is true we expect the p for low retention to always be 0.48 and the expected counts will be 0.48 x total people for each music group.

$E_{liked\ music\ low\ retention} = n \times p = 24 \times 0.48 = 11.52$

$E_{hated\ music\ low\ retention} = n \times p = 26 \times 0.48 = 12.48$

$E_{no\ music\ low\ retention} = n \times p = 25 \times 0.48 = 12.0$

Earlier we saw that computer programs often use this formula to calculate the expected counts.

Expected Counts (for contingency table) = $\frac{Row\ Total \times Column\ Total}{Grand\ Total}$

Notice that the column total is the total number of people in each music group. The row total divided by the grand total is the proportion that must be equal for all of the groups.

$E = n \times p$ = Column Total $\times \left(\frac{Row\ Total}{Grand\ Total}\right)$

Now let us calculate the Chi-Squared Test Statistic

We learned that the Chi-Squared test statistic is a comparison of the observed sample values and the expected values from the null hypothesis. Here is the formula again.

Chi-Squared Test Statistic $(\chi^2) = \sum \frac{(O-E)^2}{E}$

So Chi-Squared subtracts the observed and expected values to find the difference. Since some differences are negative, it squares the differences. It also divides by E to make it a kind of average of squares and finally it adds up these values for every variable.

Here is the sentence to explain Chi-Squared again:

"The sum of the averages of the squares of the differences between the observed sample data and the expected values if the null hypothesis were true."

	Liked Music	**Disliked Music**	**No Music**	*Total*
High Retention	10	11	18	39
Low Retention	14	15	7	36
Total	24	26	25	*Grand Total = 75*

In this example, the numbers in the two-way table are the observed counts. Note: The observed counts do not include the totals! This two-way table has two rows and three columns (not counting totals). This is often called a "two by three" (2x3) table. So we have six observed counts and six expected counts.

	Liked Music	**Disliked Music**	**No Music**
High Retention	10	11	18
Low Retention	14	15	7

Let us calculate the Chi-Squared test statistic for this problem. Here are the expected counts again. It is good to label so that you subtract the correct expected count from the correct observed count.

$E_{liked\ music\ high\ retention} = n \times p = 24 \times 0.52 = 12.48$

$E_{hated\ music\ high\ retention} = n \times p = 26 \times 0.52 = 13.52$

$E_{no\ music\ high\ retention} = n \times p = 25 \times 0.52 = 13.0$

$E_{liked\ music\ low\ retention} = n \times p = 24 \times 0.48 = 11.52$

$E_{hated\ music\ low\ retention} = n \times p = 26 \times 0.48 = 12.48$

$E_{no\ music\ low\ retention} = n \times p = 25 \times 0.48 = 12.0$

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(10-12.48)^2}{12.48} + \frac{(11-13.52)^2}{13.52} + \frac{(18-13)^2}{13} + \frac{(14-11.52)^2}{11.52} + \frac{(15-12.48)^2}{12.48} + \frac{(7-12)^2}{12}$$

$\approx 0.49282 + 0.46970 + 1.92308 + 0.53388 + 0.50885 + 2.08333 \approx 6.012$

The numbers that were added to get the Chi-Squared test statistic are called the "Contributions to Chi-Squared". Notice that the largest contributions to Chi-squared were 1.92308 and 2.08333. These calculations came from the low and high retention from the no music group.

Is this Chi-squared test statistic significant?

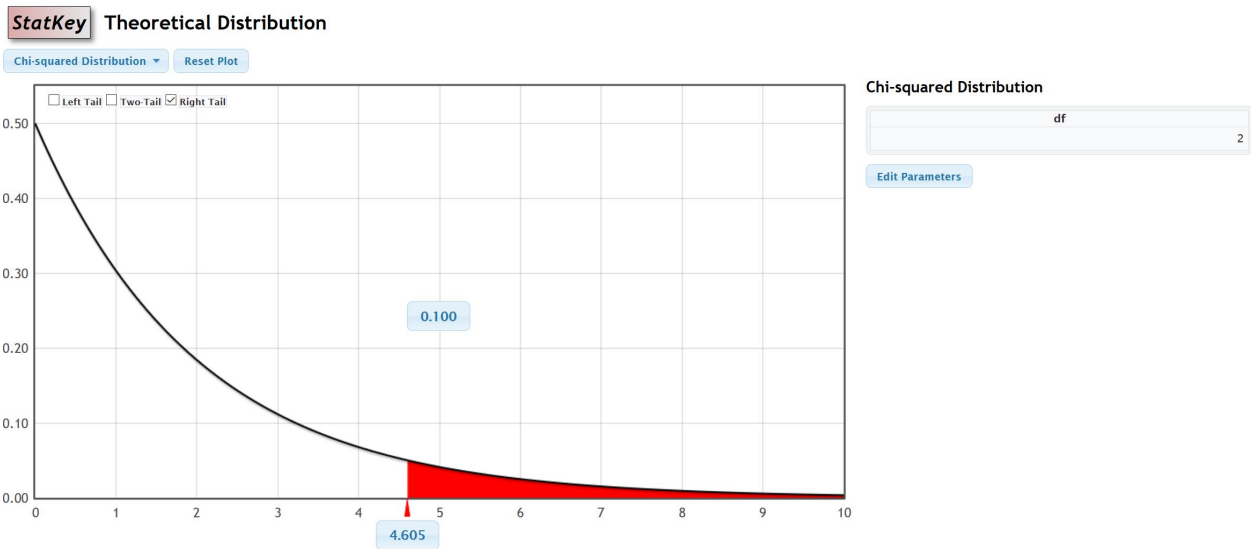
In a contingency table the degrees of freedom is the number of rows minus one times the number of columns minus one. Note: Do not include the totals when you count the number of rows and columns.

Degrees of Freedom (for Contingency Table Data) = $(r-1)\times(c-1)$
where *"r"* is the number of rows and *"c"* is the number of columns.

In the music and retention data, there were two rows and three columns so the degrees of freedom will be two.

Degrees of Freedom (for Contingency Table Data) = $(r-1)\times(c-1) = (2-1)\times(3-1) = 1\times 2 = 2$

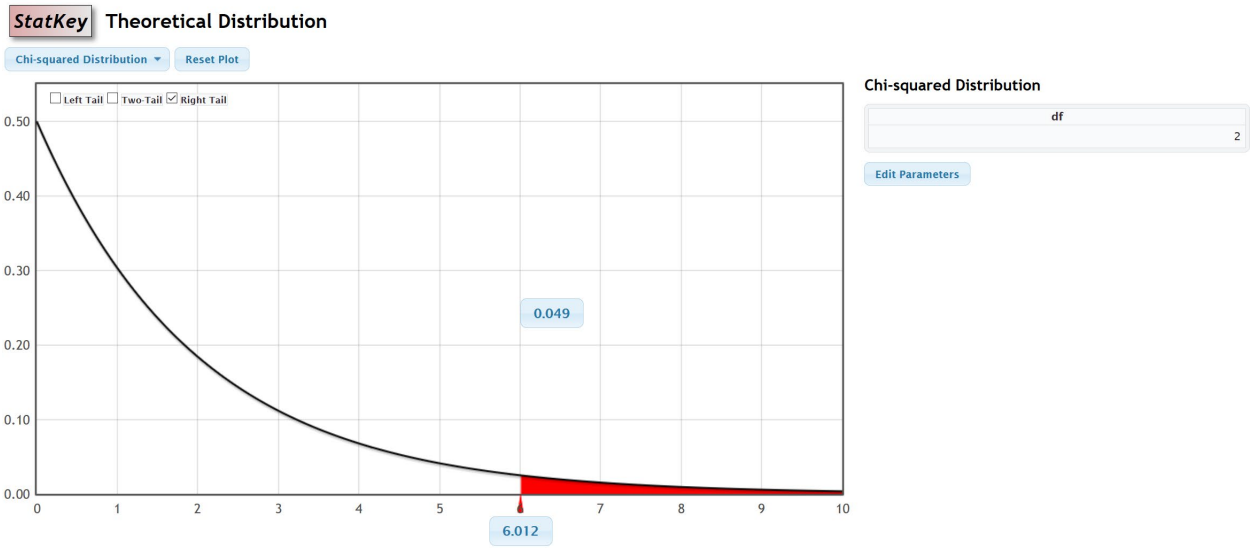
The Categorical Association Test is a right tailed test. We can use the degrees of freedom to look up the critical value using the Chi-Squared theoretical distribution calculator in StatKey. Go to www.lock5stat.com and click on StatKey. Under the "Theoretical Distributions" menu, click on "χ^2". Put in the degrees of freedom and click "right-tail". Since we are using a 10% significance level, we will enter 0.10 in the proportion for the right tail.



↑
χ^2 = 6.012

Notice that the critical value was 4.605. So our test statistic must be 4.605 or greater to be considered significant. Our Chi-squared test statistic was 6.012, which is in the right tail. This tells us that the sample data significantly disagrees with the null hypothesis. It also tells us that our observed counts are significantly different from our expected counts.

We can also use the theoretical Chi-squared curve to calculate the P-value. Just put the test statistic of 6.012 in the bottom box. Notice the computer calculated a P-value of 0.049 or 4.9%. This is less than our 10% significance level.



We can also use a randomized simulation in StatKey. Under the "More Advanced Randomization Tests" menu, click on "χ^2 Test for Association". We will need to type in our observed counts into StatKey. Do not include the totals. The computer will calculate the totals automatically.

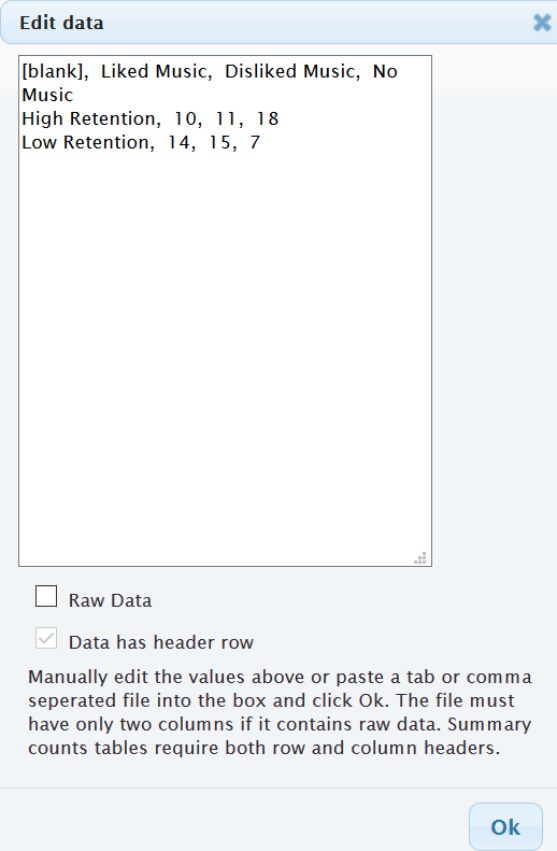
	Liked Music	Disliked Music	No Music
High Retention	10	11	18
Low Retention	14	15	7

Under the "Edit Data" menu, type in the contingency table with commas. Notice that "[blank]" must be in the top left corner.

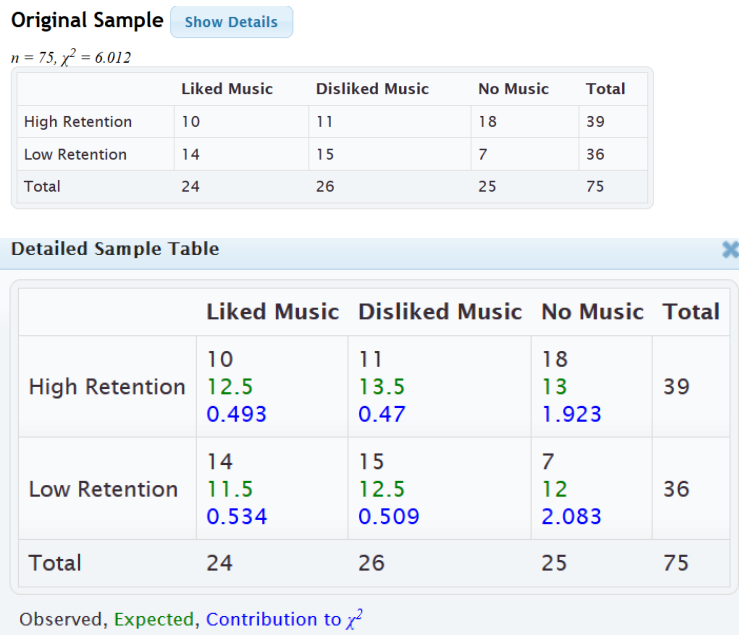
[blank], Liked Music, Disliked Music, No Music

High Retention, 10, 11, 18

Low Retention, 14, 15, 7



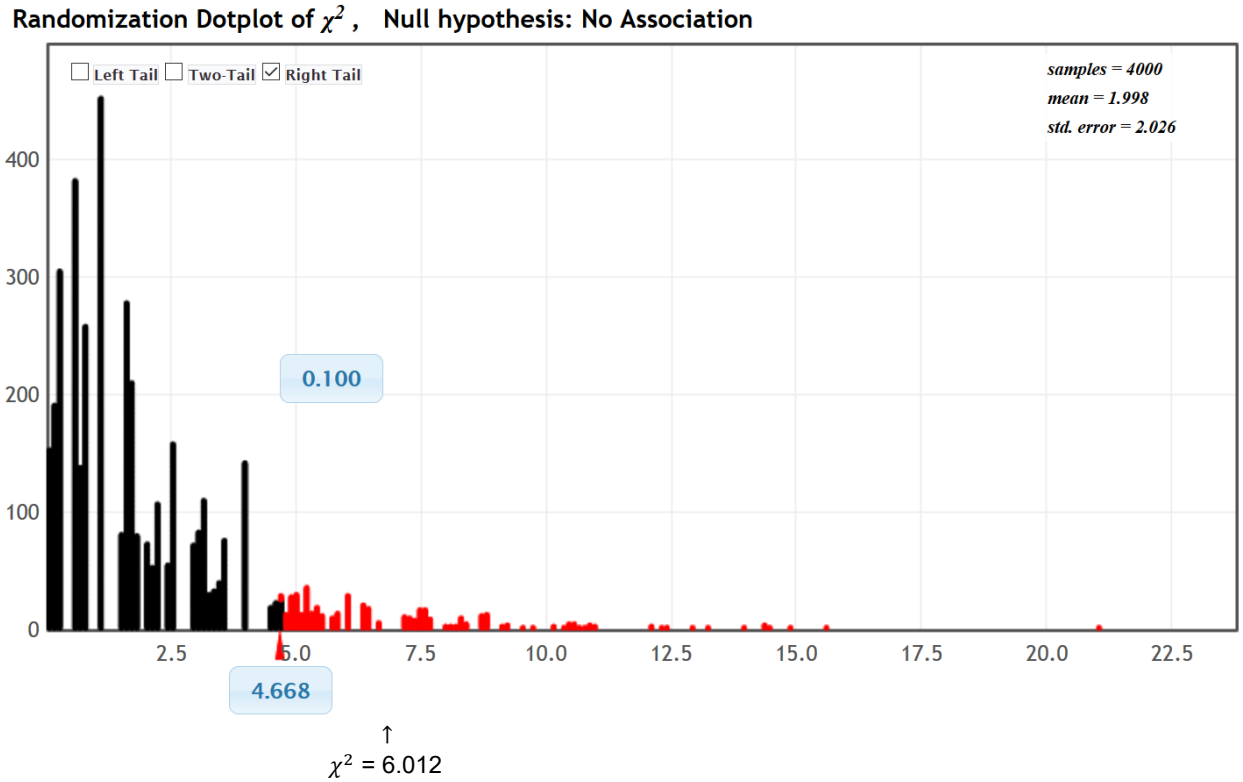
Under "Original Sample", we see that StatKey has calculated the Chi-Squared test statistic of 6.012. If you click on "Show Details", you can see the expected counts and the contributions to chi-squared.



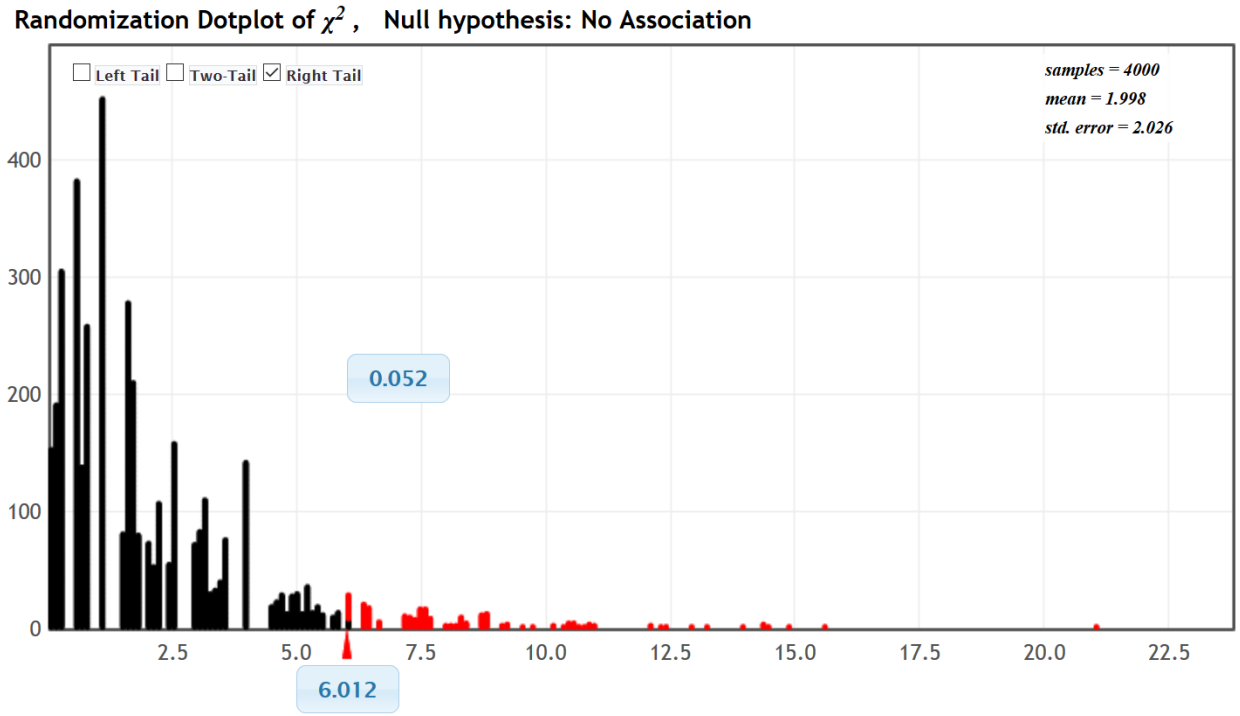
	Liked Music	Disliked Music	No Music	Total
High Retention	10	11	18	39
Low Retention	14	15	7	36
Total	24	26	25	75

	Liked Music	Disliked Music	No Music	Total
High Retention	10 12.5 0.493	11 13.5 0.47	18 13 1.923	39
Low Retention	14 11.5 0.534	15 12.5 0.509	7 12 2.083	36
Total	24	26	25	75

If we click “Generate 1000 Samples” a few times, we get our randomized simulation. Notice the null hypothesis is “No Association” (not related). Putting our 10% significance level in the tail proportion, gives us an approximate critical value of 4.668. This is close to what we got with the theoretical curve. Notice that our Chi-squared test statistic of 6.012 does fall in the tail of the simulation.



We can also use the simulation to calculate the approximate P-value. Just put the test statistic of 6.012 into the bottom box. Notice the P-value came out to be 0.052. This is almost the same P-value as we calculated with the theoretical Chi-squared distribution.



Calculating the Chi-Squared test statist and P-value with Statcato

- First type in the contingency (two-way) table exactly as you see it into Statcato. The column titles will be in the grey cells (VAR) at the top, but the row titles will be in the regular cells. Do not type the totals. Statcato will automatically calculate them.

	C1	C2	C3	C4
Var		Liked Music	Disliked Music	No Music
1	High Retention	10	11	18
2	Low Retention	14	15	7

- Go to the "statistics" menu and click on "multinomial experiment" then "chi-square contingency table". Hold the control key down and click on the columns that have your observed frequencies (C2, C3 & C4). Do not include the row labels (C1). They can be added later.

Chi-Square Test: Contingency Table:

	C2 Liked Music	C3 Disliked Music	C4 No Music	Total
High Retention	10.0 (12.48) [0.49]	11.0 (13.52) [0.47]	18.0 (13.0) [1.92]	39.0
Low Retention	14.0 (11.52) [0.53]	15.0 (12.48) [0.51]	7.0 (12.0) [2.08]	36.0
Total	24.0	26.0	25.0	75.0

(expected frequency), [test statistic contribution]

Significance Level	DOF	χ^2	Critical value	p-Value
0.05	2	6.0117	5.9915	0.0495

Notice the test statistic, critical value and P-value in Statcato are about the same as we got in the simulation.

The expected counts (expected frequencies) are in parenthesis in the Statcato printout and the contributions to chi-squared are in brackets.

Assumptions

The assumptions for the Categorical Association Test are as follows:

Categorical Association Test Assumptions

- The categorical sample or samples should be collected randomly or be representative of the population.
- Data values within each sample should be independent of each other.
- If multiple samples are collected, then the data values between the samples should be independent of each other.
- The expected counts from the null hypothesis should be at least five.

Did the music and retention data meet the assumptions?

Random? Yes. This was an experiment and the groups were randomly assigned.

Were the expected counts at least 5? Yes. The expected counts were 12.4, 13.52, 13.0, 11.52, 12.48, and 12.0. All the expected frequencies were at least five.

Independence? The individuals in the experiment were randomly assigned, so we can probably assume the data met the independence requirement.

Writing the conclusion

Should we reject the null hypothesis or fail to reject the null hypothesis? Since our P-value is lower than our significance level, we reject the null hypothesis.

Remember a conclusion must address the evidence and the claim. Since our P-value is low, we have significant evidence. We rejected the null hypothesis, but the claim was the alternative hypothesis. We will therefore support the claim.

H_0 : Listening to music and memorizing information are not related (not associated, independent).
H_A : Listening to music and memorizing information are related (associated, dependent). CLAIM

Conclusion: We have significant sample evidence to support the claim that listening to music and retaining information are related.

In fact, since confounding variables were controlled, the experiment proves cause and effect. The “no music” group did significantly better than either of the music groups. This experiment appears to indicate that when a person listens to either music they like or music they hate, they have a harder time memorizing information.

Note about “Independence” and “Homogeneity”

Sometimes we obtain multiple categorical data from one random sample of people or objects. We ask multiple categorical questions from the same group of people. Statisticians sometimes refer to that situation as an “independence” test.

If we collect the same categorical data from multiple random samples, then that is sometimes referred to as a “homogeneity” test. We may ask the same categorical question from different groups of people.

Whether you collected the data from one sample or multiple samples, the data still summarizes into a contingency table. Look at the following sample data.

	Business	English	History	Music	Biology	Math
Female	89	71	62	48	56	9
Male	112	58	59	53	62	13

Suppose we took one random sample of college students and asked them two categorical questions. What gender do you most identify with? What is your major? This would be referred to as an "independence" test.

We could collect this data in another way. Suppose we took a random sample of female college students and asked them what their major was. Later we took a random sample of male college students and asked them the same question. This would now be referred to as a "homogeneity" test.

Notes about the Categorical Association Test

- The Categorical Association Test is used to determine if categories with multiple options are related or not.
- The categorical association test is always a right-tailed test.
- The degrees of freedom for the categorical association test is the number of rows minus one times the number of columns minus one. $df = (r-1) \times (c-1)$
- The categorical association test uses the Chi-squared test statistic (χ^2), which compares the observed sample counts to the expected counts if the null hypothesis was true.
- Always use a computer to calculate the test statistic. Focus on being able to interpret and judge significance.

Practice Problems Section 4F

(#1-10) Use each of the following categorical association χ^2-test statistics and the corresponding critical values to fill out the table.

	χ^2-test stat	Sentence to explain χ^2-test statistic.	Critical Value	Does the χ^2-test statistic fall in a tail determined by the critical value? (Yes or No)	Does sample data significantly disagree with H_0?
1.	+1.573		+4.117		
2.	+6.226		+5.118		
3.	+2.144		+4.121		
4.	+3.415		+5.091		
5.	+13.718		+7.189		
6.	+0.972		+4.812		
7.	+31.652		+12.557		
8.	+11.185		+5.181		
9.	+25.443		+7.008		
10.	+1.133		+8.336		

(#11-20) Use each of the following P-values and corresponding significance levels to fill out the table.

	P-value Proportion	P-value %	Sentence to explain the P-value	Significance Level %	Significance level Proportion	If H_0 is true, could the sample data occur by random chance or is it unlikely?	Reject H_0 or Fail to reject H_0?
11.	0.263			10%			
12.	0.0042			1%			
13.	5.22×10^{-4}			5%			
14.	0.0639			1%			
15.	0			5%			
16.	0.539			10%			
17.	0.0419			5%			
18.	0.0027			10%			
19.	7.73×10^{-8}			1%			
20.	0.674			5%			

21. If we have two raw categorical data sets, what must we click on in Statcato to perform a categorical association test?

22. If we have summary counts organized in a contingency table, what must we click on in Statcato to perform a categorical association test?

23. What are the assumptions for a categorical association test if the data was collected from on random sample?

24. What are the assumptions for a categorical association test if the data was collected from multiple random samples?

25. How are the expected counts calculated in a categorical association test?

26. If the expected counts from the null hypothesis are significantly different from the observed sample counts, describe the effect on the Chi-Squared test statistic.

27. If the expected counts from the null hypothesis are close to the observed sample counts, describe the effect on the Chi-Squared test statistic.

(#28-31) Directions: For each of the following problems, use the Statcato printout provided to answer the following questions.

a) Write the null and alternative hypothesis. Make sure to label which one is the claim.

b) Check the assumptions for the categorical association test.

c) What is the Chi-squared test statistic? Write a sentence to explain the test statistic.

d) Does the test statistic fall in the tail determined by the critical value?

e) Does the sample data significantly disagree with the null hypothesis? Explain your answer.

f) Are the observed counts significantly different from the expected counts? Explain your answer.

g) What is the P-value? Write a sentence to explain the P-value.

h) Compare the P-value to the significance level. Should we reject the null hypothesis or fail to reject the null hypothesis? Explain your answer.

i) If the null hypothesis was true, could the sample data or more extreme have occurred by sampling variability or is it unlikely to be sampling variability? Explain your answer.

j) Write a conclusion for the test addressing evidence and the claim. Explain your conclusion in non-technical language.

k) Are the categories related or not? Explain your answer.

28. A random sample of male college students were asked their major. Later, a random sample of female college students were asked their major. The goal of the study was to show that gender is not related to major. Use a 5% significance level and the Statcato printout below to answer the questions given above.

Chi-Square Test: Contingency Table:

	Business	English	History	Music	Biology	Math	Total
Female	89.0 (97.30) [0.71]	71.0 (62.45) [1.17]	62.0 (58.58) [0.20]	48.0 (48.89) [0.02]	56.0 (57.12) [0.02]	9.0 (10.65) [0.26]	335.0
Male	112.0 (103.70) [0.67]	58.0 (66.55) [1.10]	59.0 (62.42) [0.19]	53.0 (52.11) [0.02]	62.0 (60.88) [0.02]	13.0 (11.35) [0.24]	357.0
Total	201.0	129.0	121.0	101.0	118.0	22.0	692.0

(expected frequency), [test statistic contribution]

Significance Level	DOF	χ^2	Critical value	p-Value
0.05	5	4.6014	11.0705	0.4664

29. A random sample of adults were asked their blood type and Rh status. (Blood tests were provided for those that did not know their blood type and Rh status.) The goal of the study was to show that blood type is related to Rh status (dependent). Use a 10% significance level and the Statcato printout below to answer the questions given above.

Chi-Square Test: Contingency Table:

	Type A	Type B	Type AB	Type O	Total
Rh+	35.0 (36.03) [0.03]	24.0 (23.0) [0.04]	11.0 (16.1) [1.62]	91.0 (85.87) [0.31]	161.0
Rh-	12.0 (10.97) [0.10]	6.0 (7.0) [0.14]	10.0 (4.9) [5.31]	21.0 (26.13) [1.01]	49.0
Total	47.0	30.0	21.0	112.0	210.0

(expected frequency), [test statistic contribution]

Significance Level	DOF	χ^2	Critical value	p-Value
0.10	3	8.5522	6.2514	0.0359

30. A hospital wanted to determine if the age of a patient is not related to what part of the hospital they were in. They took a random sample of patients that have visited their hospital and determined both their age and the part of the hospital. The ages were broken up into age groups. Use a 1% significance level and the Statcato printout below to answer the questions given above.

Chi-Square Test: Contingency Table:

	Med/Surg	ICU	SDS	ER	Total
18-35 years old	19.0 (19.12) [$7.98 \cdot 10^{-4}$]	4.0 (11.47) [4.87]	25.0 (17.85) [2.87]	16.0 (15.55) [0.01]	64.0
36-49 years old	27.0 (19.42) [2.96]	7.0 (11.65) [1.86]	22.0 (18.13) [0.83]	9.0 (15.80) [2.92]	65.0
50-64 years old	17.0 (18.53) [0.13]	13.0 (11.12) [0.32]	15.0 (17.29) [0.30]	17.0 (15.07) [0.25]	62.0
65+ years old	12.0 (17.93) [1.96]	21.0 (10.76) [9.75]	8.0 (16.73) [4.56]	19.0 (14.58) [1.34]	60.0
Total	75.0	45.0	70.0	61.0	251.0

(expected frequency), [test statistic contribution]

Significance Level	DOF	χ^2	Critical value	p-Value
0.01	9	34.9208	21.666	$6.153 \cdot 10^{-5}$

31. A random sample of American adults was taken and their health and education status obtained. Test to test the claim that health and education are related. Use a 5% significance level and the Statcato printout below to answer the questions given above.

Chi-Square Test: Contingency Table:

	Excellent Health	Good Health	Fair Health	Poor Health	Total
Less Than High School	72.0 (148.64) [39.51]	202.0 (249.76) [9.13]	199.0 (106.91) [79.33]	62.0 (29.70) [35.14]	535.0
High School Diploma	465.0 (502.31) [2.77]	877.0 (844.04) [1.29]	358.0 (361.29) [0.03]	108.0 (100.36) [0.58]	1808.0
Some College / Associates Degree	80.0 (77.24) [0.10]	138.0 (129.78) [0.52]	49.0 (55.55) [0.77]	11.0 (15.43) [1.27]	278.0
Bachelor's Degree	229.0 (161.42) [28.30]	276.0 (271.23) [0.08]	64.0 (116.10) [23.38]	12.0 (32.25) [12.72]	581.0
Graduate Degree	130.0 (86.40) [22.00]	147.0 (145.19) [0.02]	32.0 (62.15) [14.62]	2.0 (17.26) [13.49]	311.0
Total	976.0	1640.0	702.0	195.0	3513.0

(expected frequency), [test statistic contribution]

Significance Level	DOF	χ^2	Critical value	p-Value
0.05	12	285.0610	21.0261	0

(#32-35) Directions: Use StatKey at www.lock5stat.com to simulate the following chi-squared categorical association tests. Go to the "More Advanced Randomization Tests" menu at the bottom of the StatKey page. Click on the button that says, "χ^2 Test for Association". Click on "Edit Data" and type in the contingency table provided. Click on "Generate 1000 Samples" a few times to create the simulated sampling distribution and answer the following questions.

a) Write the null and alternative hypothesis. Make sure to label which one is the claim.

b) Check the assumptions for the categorical association test. Assume the data was collected randomly. Under "Original Sample", click on "Show Details" to see the expected counts.

c) Use the formula df = (r – 1)(c – 1) to calculate the degrees of freedom. "r" is the number of rows and "c" is the number of columns not counting the totals.

d) What is the Chi-squared test statistic? Write a sentence to explain the test statistic.

e) Put the significance level proportion in the right tail proportion to calculate the critical value. What is the critical value? (Answers will vary slightly.) Does the original sample χ^2 test statistic fall in the tail determined by the critical value?

f) Does the sample data significantly disagree with the null hypothesis? Explain your answer.

g) Are the observed counts significantly different from the expected counts? Explain your answer.

h) Put the original sample test χ^2 test statistic in the bottom box in the simulation to calculate the P-value. What is the P-value? (Answers will vary slightly.) Write a sentence to explain the P-value.

i) Compare the P-value to the significance level. Should we reject the null hypothesis or fail to reject the null hypothesis? Explain your answer.

j) If the null hypothesis was true, could the sample data or more extreme have occurred by sampling variability or is it unlikely to be sampling variability? Explain your answer.

k) Write a conclusion for the test addressing evidence and the claim. Explain your conclusion in non-technical language.

l) Are the categories related or not? Explain your answer.

32. We want to know if the state a home is built in is related to the size of the home. A random sample of homes in the U.S was taken. Click on "Edit Data" in StatKey and type in the following contingency table. Do not forget to include a space after the commas. Use a 5% significance level and randomized simulation to test the claim that the state is not related to size of the home.

[blank], CA, NJ, NY, PA

Large, 7, 6, 7, 3

Small, 23, 24, 23, 27

33. Open the "Car Data" at www.matt-teachout.org. Copy and paste the "Country" and "Cylinders" columns next to each other in a new Excel spreadsheet. Then copy the two columns together. Click on "Edit Data" in StatKey and paste the two columns into StatKey. Use a 1% significance level to test the claim that the country a car is made in is related to the cylinders. Answer the questions above.

34. We want to show that gender is related to getting an award. A random sample of people that won famous awards in the Olympic, Academia, and Nobel was taken and their gender was noted. Click on "Edit Data" in StatKey and type in the following contingency table. Do not forget to include a space after the commas. Use a 10% significance level and randomized simulation to test the claim that awards are related to gender.

[blank], Olympic, Academy, Nobel

Male, 109, 11, 73

Female, 73, 20, 76

35. Open the "Math 140 Fall 2015 Survey Data" at www.matt-teachout.org. Copy and paste the "Tattoo" and "Favorite Social Media" columns next to each other in a new Excel spreadsheet. Then copy the two columns together. Click on "Edit Data" in StatKey and paste the two columns into StatKey. Use a 5% significance level to test the claim that having a tattoo or not is not related to social media. Answer the questions above.

Section 4G – Quantitative Relationships: Correlation and Regression

Vocabulary

Correlation: Statistical analysis that determines if there is a relationship between two different quantitative variables.

Regression: Statistical analysis that involves finding the line or model that best fits a quantitative relationship, using the model to make predictions, and analyzing error in those predictions.

Explanatory Variable (x): Another name for the x-variable or independent variable in a correlation study.

Response Variable (y): Another name for the y-variable or dependent variable in a correlation study.

Correlation Coefficient (r): A statistic between −1 and +1 that measures the strength and direction of linear relationships between two quantitative variables.

R-squared (r^2): Also called the coefficient of determination. This statistic measures the percent of variability in the y-variable that can be explained by the linear relationship with the x-variable.

Residual $(y - \hat{y})$: The vertical distance between the regression line and a point in the scatterplot.

Standard Deviation of the Residual Errors (s_e): A statistic that measures how far points in a scatterplot are from the regression line on average and measures the average amount of prediction error.

Slope (b_1): The amount of increase or decrease in the y-variable for every one-unit increase in the x-variable.

Y-Intercept (b_0): The predicted y-value when the x-value is zero.

Regression Line $(\hat{y} = b_0 + b_1 x)$: Also called the line of best fit or the line of least squares. This line minimizes the vertical distances between it and all the points in the scatterplot.

Scatterplot: A graph for visualizing the relationship between two quantitative ordered pair variables. The ordered pairs (x, y) are plotted on the rectangular coordinate system.

Residual Plot: A graph that pairs the residuals with the x values. This graph should be evenly spread out and not fan shaped.

Histogram of the Residuals: A graph showing the shape of the residuals. This graph should be nearly normal and centered close to zero.

Introduction

Sometimes we want to know if two different quantitative variables are related to each other. This kind of relationship study is difficult because the units are different. We cannot directly compare the height of man in inches to his weight in pounds. Inches and pounds are completely different. Statisticians and mathematicians developed a type of analysis for this situation called “correlation and regression”. The idea is to let one variable be X and the other variable be Y. Then use ordered pair data to create a graph called a scatterplot and look for patterns. The most common is a linear pattern (correlation). If we see a linear pattern, we can also calculate the line that best fits the data and use this line to make predictions (regression).

Choosing your variables

It is important to determine which variable will be X and which variable will be Y. In statistics, we call the X-variable the “explanatory variable” or the “independent variable”. We call the Y-variable the “response variable” or “dependent variable”. How do we chose? Here are a couple key questions to ask yourself.

- Does one variable respond more than the other does?
- Which variable is the focus of the study and the variable I might want to make predictions about?

Let us look at some examples.

Example: Year (time) and unemployment rates in U.S.

Ask yourself the following question. Does one of the variables responds more than the other? Does time fluctuate in response to the unemployment rate? That does not sound right. Time seems to go on no matter what happens with unemployment. Do you think unemployment might fluctuate in response to time? That seems more likely. So we should let the explanatory variable X be time (years) and let the response variable y be unemployment rate. Unemployment responds to time, but not the other way around.

Example: The unemployment rate in U.S. and the national debt in the U.S.

These variables respond to each other, so either variable could be the response variable Y. In this case, pick the response variable (Y) to be the one you are most interested in (focus of the study) or the variable you may want to make predictions about. If there is a relationship, then the Y-variable will be the variable you can make predictions about.

Suppose the focus of your study and the variable you want to predict is the national debt. Unemployment may just be one factor that may be related to the national debt. If that is the case, you should make the national debt your response variable Y. By default, that means that unemployment rate would be explanatory variable X.

Correlation Graphs and Statistics with StatKey

To study the relationship between two different quantitative variable, you will need ordered pair data. For example, we will need the height and weight of the same men, or the unemployment rate and national debt of the same countries. Decide which variable should be X and which variable should be Y. The computer will then make ordered pairs from your data (X , Y) and plot all the points on the rectangular coordinate system. This graph of all the ordered pairs is called a scatterplot.

Example

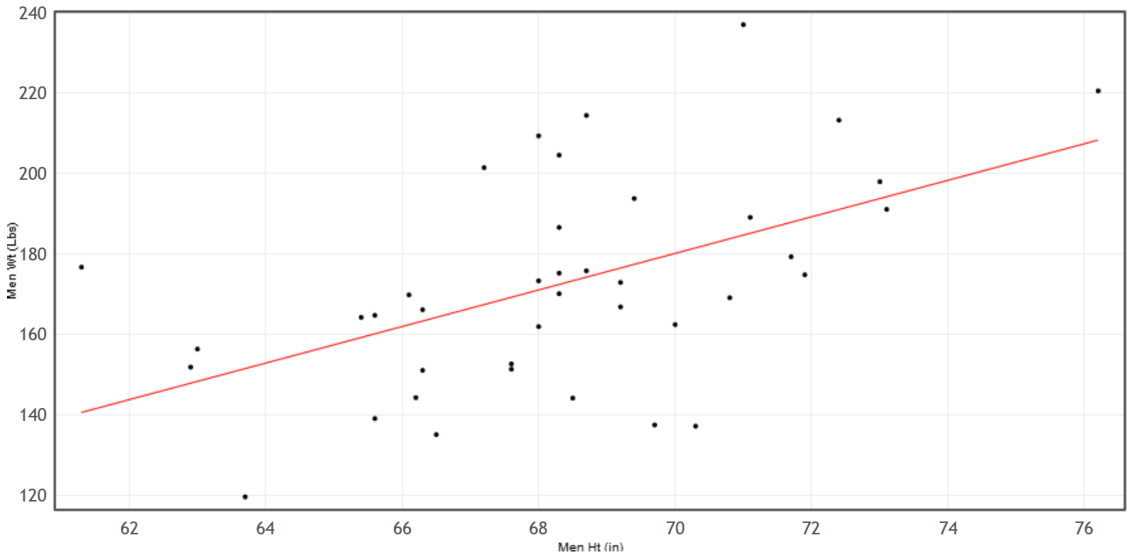
Suppose we want to study if the weights in pounds of the men in the health data is related to their heights in inches. I am most interested in predicting the weights of men from their heights so I will let the weight be the response variable Y and height be the explanatory variable X. Notice these are ordered pairs, since the heights and weights came from the same 40 men.

To put the data into StatKey, you will want to open a fresh excel spreadsheet and place the two data sets side by side. These two data sets are already next to each other in the health data, but in general, the data sets may not be. Copy the two columns of data together.

Go to www.lock5stat.com and click on “StatKey”. Under the “Descriptive Statistics and Graphs” menu, click on “two quantitative variables”. Under the “edit data” tab, paste the height and weight data into StatKey. The graph you see is the scatterplot. Notice StatKey has placed the heights on the horizontal x-axis and the weights on the vertical y-axis. If it is backward, simply click the “switch variables” button. It is also nice to check the “show regression line” box. The regression line is the line that best fits the points in the scatterplot. StatKey has also given us some statistics to help understand the relationship.

StatKey **Descriptive Statistics for Two Quantitative Variables**

Custom Dataset | Show Data Table | Edit Data | Upload File | Change Column(s)



Analyzing scatterplots is an important skill. In this graph, we see that the points seem to follow the linear pattern reasonably well and are reasonably close to the line. Shorter men on the left tend to have lower weights than taller men on the right. The line goes up from left to right. We call this a "positive linear relationship", or a "positive correlation". If the line goes down from left to right, we would call that a "negative linear relationship", or a "negative correlation".

Summary Statistics Switch Variables

Statistic	Men Ht (in)	Men Wt (Lbs)
Mean	68.335	172.550
Standard Deviation	3.020	26.327
Sample Size	40	
Correlation	0.522	
Slope	4.553	
Intercept	-138.607	

Scatterplot Controls

☑ Show Regression Line

We see that StatKey has given us the mean and standard deviation of each data set (heights and weights). It has also given us the sample size (n) of 40. There were 40 ordered pairs (40 heights and 40 weights from the same 40 men). The number next to the word "Correlation" is 0.522. This is called the "correlation coefficient" (r) and is an important statistic in measuring the direction and strength of the linear relationship. Here are some general guidelines for understanding the correlation coefficient "r".

Correlation Coefficient (r)

The correlation coefficient (r) is a number between -1 and +1 that measures the strength and direction of correlation. The correlation coefficient is an extremely difficult calculation that is very time consuming. Like most statistics, it is better to use a computer program like StatKey or Statcato to calculate it.

If the r is negative, the regression line will go down from left to right. If you remember from algebra classes, this means the line has a negative slope. If the r is positive, the regression line will go up from left to right. This means the line has a positive slope. The closer r is to +1 or −1, the stronger the relationship. This means the points are very close to the line. The closer r is to zero, the weaker the relationship. The points are very far from the line. It is important to always look at the scatterplot with the r-value. Do not just look at an r-value without looking at the scatterplot. These are not strict rules, but general guidelines. A scatterplot with many points and a 0.7 r-value can mean something different from a scatterplot with only a few points and a 0.7 r-value.

- If r is close to +1 (like r = +0.893) ➔ Strong, Positive Correlation (line going up from left to right (positive slope) and the points in scatterplot are close to line) ,
(r ≈ +0.6, +0.7, +0.8 , +0.9 usually indicate pretty strong positive correlation)

- If r is close to −1 (like r = −0.916) ➔ Strong Negative Correlation (line going down from left to right (negative slope) and the points in the scatterplot are close to the line)
(r ≈ −0.6, −0.7, −0.8 , −0.9 usually indicate pretty strong negative correlation)

- If r close to zero (like +0.037 or −0.009) ➔ No linear correlation. Points in the scatterplot do not follow any linear pattern. There still could be a nonlinear curved pattern though.
(r ≈ ±0.1, ±0.0 usually indicate no linear correlation)

- If r ≈ ±0.2, ±0.3 usually indicate very weak linear correlation. There is some linear pattern but the points are very far from the regression line.

- If r ≈ ±0.4, ±0.5 usually indicate moderate linear correlation. There is a linear pattern and points are only moderately close to the regression line.

In the men's height and weight example, the r-value was +0.522. This tells us that there is a moderate positive linear relationship (or moderate positive correlation) between the height and weight of these men.

Important Note: Remember relationships or associations do not imply causation. Just because there is a positive linear relationship between the height and weight of these men, it does not give me the right to say that the height causes a man to have a certain weight. There are many confounding variables involved.

Correlation ≠ Causation

Coefficient of Determination (r^2)

If you square the r-value, you get the coefficient of determination. This statistic tells us the percentage of variability in the response variable (Y) that can be explained by the explanatory variable (X). In general, the higher the r^2 percentage, the stronger the relationship.

StatKey does not calculate r^2 for us, but it is not a difficult calculation. If we square the r-value, we get the following.

$r^2 = (0.522)^2 = 0.522 \times 0.522 \approx 0.272$ or 27.2%

So about 27.2% of the variability in the men's weights can be explained by the relationship with their heights.

Slope

The slope of the regression line is an important statistic in correlation and regression. It is a difficult calculation. If you are wondering how it is calculated, here is the formula the computer used.

Slope of the Regression Line = $\frac{Correlation\ Coefficient \times Standard\ Deviation\ of\ Y}{Standard\ Deviation\ of\ X}$

The slope is the amount of increase of decrease in Y for every 1-unit increase in X (per unit of X). If the slope is negative, than it is a "decrease" in Y and if the slope is positive, it is an "increase" in Y.

In this problem, StatKey gave us the slope as 4.553. Notice this is a positive slope so is indicating an increase in Y. The slope tells us that the weights of the men in the data set are increasing 4.553 pounds on average for every 1 inch taller they get. Another way to say that is that the weights are increasing on average 4.552 pounds per inch.

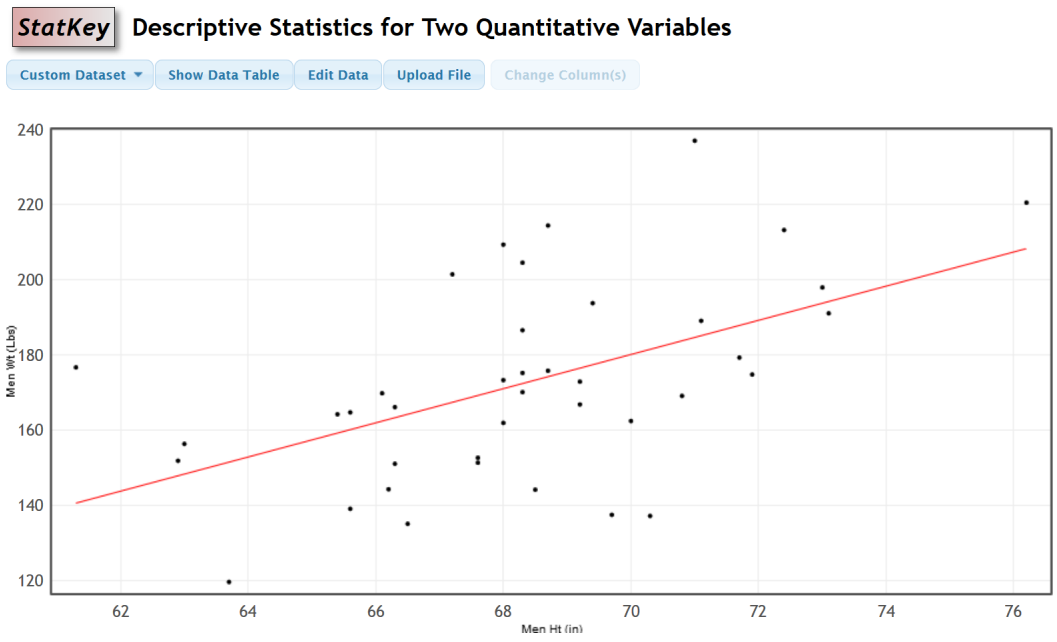
Y-intercept

The Y-intercept is another difficult calculation. In case you are wondering, here is the formula the computer used to calculate the Y-intercept. You must calculate the slope first, before you can find the Y-intercept.

Y-intercept of Regression Line = Mean of Y values − $(Slope\ \times Mean\ of\ X\ values)$

Y-intercepts can be difficult because they do not always make sense in context. The definition of a Y-intercept is the predicted Y-value when X is zero. StatKey calculated the Y-intercept for the height and weight data as −138.607. So by definition, the predicted average height of men that are zero inches tall is negative 138.607 pounds. That does not make sense.

In many situations (like heights of men), it is impossible for the X to be zero. Look at the scatterplot again for the height and weight data.



Notice that the points in the scatterplot have X values between about 61 inches and about 76 inches. This is called the scope of the X-values. The accuracy of this regression line is based on X values between about 61 and 76 inches. If we use this data to predict a man's weight from his height, we should only use heights in the scope (between 61 and 76). Going outside the scope is called extrapolation and can result in bad errors. So let us get back to the Y-intercept. The Y intercept is plugging in zero for X. Notice zero is not in the scope of the X values, so is an extrapolation. That means we will not expect the Y-intercept to make sense in this context. The number is correct and is important for the regression line accuracy, but a man cannot have a height of zero.

Some Y-intercepts do make sense in context. Suppose we are looking at the number of months a company has been in business (X) and their monthly revenue in thousands of dollars (Y). The Y-intercept may represent their starting capital at month zero or the amount of money the company had when they started their business.

Regression Line and Predictions

The regression line is also called the "line of best fit" or the "line of least squares". It minimizes the vertical distances between the points in the scatterplot and regression line itself. If there is correlation between the variables, then the regression line is also a prediction formula. If you plug in an X value into the equation for X, you can solve for Y and get a predicted Y value. The regression line is represented by the following formula.

$\hat{Y}$ = (Y-intercept) + (Slope) X

Plugging in our Y intercept (−138.607) and our slope (4.553), we get the following equation.

Regression Line for Heights and Weights of mean in the health data: $\hat{Y} = -138.607 + 4.553\ X$

The $\hat{Y}$ refers to the "predicted Y value" which can be very different from the actual Y values in the data set. You may also see computer programs put in the variable names for X and $\hat{Y}$.

Weights in pounds = −138.607 + 4.553 (Heights in inches)

We said already that there was a moderate correlation between the heights and weights of these men. So we should be able to use the formula to make a prediction.

Use the regression line equation to predict the average weights of men that are 73 inches tall. Remember Y represents weight and X represents height. Simply plug in 73 for X and solve for Y. Remember to follow the order of operations. Multiply the X value by the slope first, before you add it to the Y-intercept. Also, be aware of negative Y-intercepts and negative slopes.

$\hat{Y} = -138.607 + 4.553\ X$

$\hat{Y} = -138.607 + 4.553\ (73)$

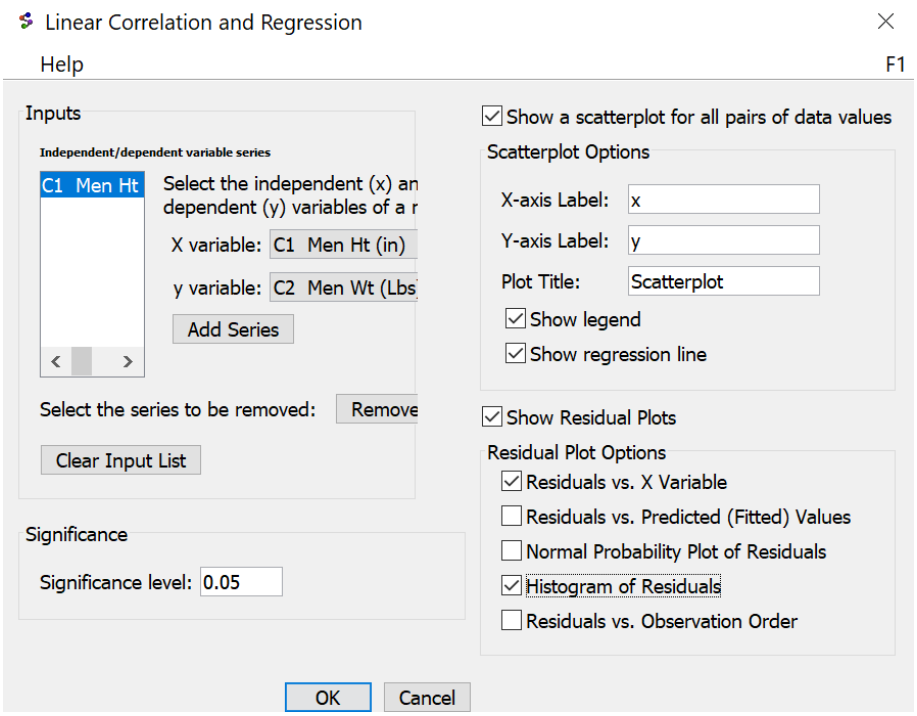
$\hat{Y} \approx -138.607 + 332.369$

$\hat{Y} \approx +193.762$

Therefore, we predict that the average weight of men that are 73 inches tall is about 193.8 pounds. Be careful of applying this prediction to all men. This data came from sample data and may not reflect the heights of all men on earth.

Calculating Correlation Graphs and Statistics with Statcato

We can also make scatterplots and calculate correlation statistics with Statcato. Copy and paste the men's height and weight data into two columns of Statcato. Go to the "statistics" menu, click on "correlation and regression" and then click on "linear". Click on the height to be the X-variable and the weight to be the Y-variable and then push "add series". Check the box that says "show scatterplot" and the box that says "show regression line". Statcato also has the capability of making residual plots. These are more advanced kinds of graphs that are studied in regression analysis. Check the box that says, "Show residual plots", the box that says "residuals vs x-variable", and the box that says "histogram of the residuals". Now push "OK".



Correlation and Regression: Significance level = 0.05

Series: C1 Men Ht (in),C2 Men Wt (Lbs)
x = C1 Men Ht (in)
y = C2 Men Wt (Lbs)
Sample size n = 40
Degrees of freedom = 38

Correlation:
H_0: ρ = 0 (no linear correlation)
H_1: ρ ≠ 0 (linear correlation)

	Test Statistic	Critical Value
r	0.5222	±0.3120
t	3.7750	±2.0244

p-Value = 0.0005

Regression:
Regression equation $Y = b_0 + b_1x$
b_0 = -138.6070
b_1 = 4.5534

Variation:
Explained variation = 7372.6464
Unexplained variation = 19659.0136
Total variation = 27031.66
Coefficient of determination r^2 = 0.2727
Standard error of estimate = 22.7452

Some of the information in this printout refers to the correlation hypothesis test that we will study in chapter five. Notice Statcato gave us the correlation coefficient r of 0.522 and the coefficient of determination r^2 = 0.2727 (27.27%). The slope is given as b_1 = 4.5534 and the Y-intercept is given as b_0 = −138.6070. Notice these are the same numbers as StatKey.

There is one statistic on the Statcato printout that was not on the StatKey printout that is important.

Standard error of estimate = 22.7452

This statistic is called the standard deviation of the residual errors (s_e). It measures the average vertical distance that points in the scatterplot are from the regression line. It also tells us the average prediction error for predictions made in the scope of the X-values. The units of the standard deviation of the residual errors is the same as the Y-variable (pounds). This statistic tells us the following.

The points in the scatterplot are 22.7452 pounds on average from the regression line.

If we use the regression line and the height of a man to predict the weight, our prediction could have an average error of 22.7452 pounds.

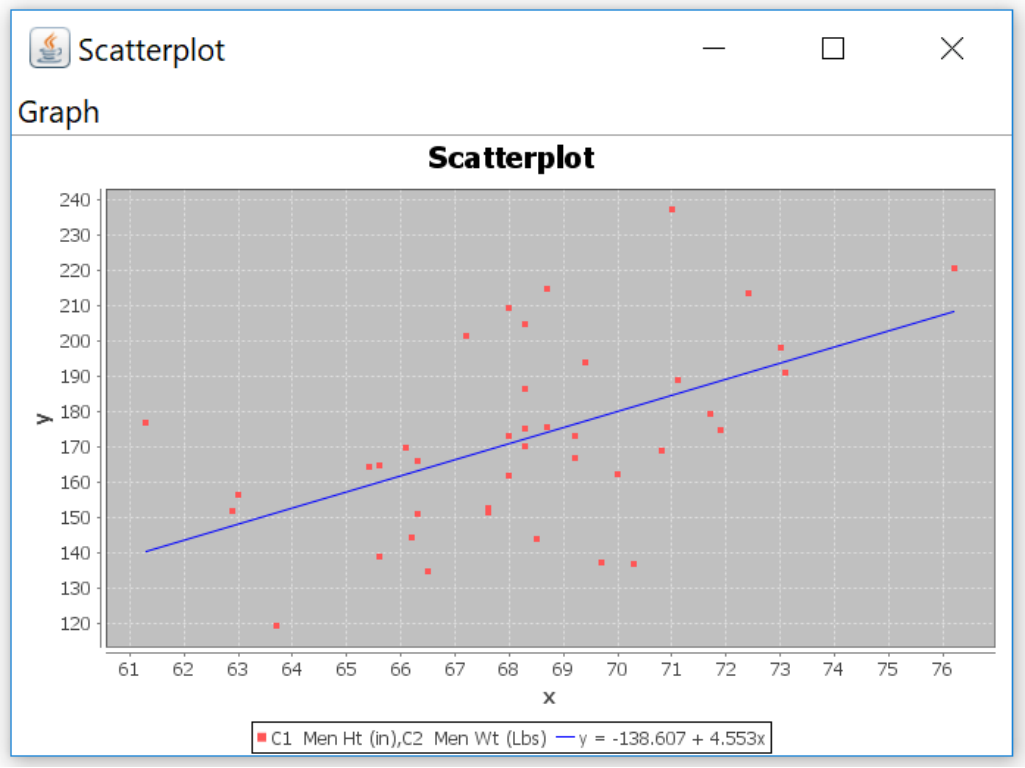
Remember the prediction we made earlier. We predicted that the average weight of men that are 73 inches tall is about 193.8 pounds. Well that prediction could be off by 22.7452 pounds on average.

A "residual" is the vertical distance that each point is from the regression line. Suppose a point has an ordered pair (X , Y). The point on the regression line with the same X value would have an ordered pair (X , $\hat{Y}$). To calculate a residual the computer subtracts the predicted $\hat{Y}$ value from the actual Y value of the point in the scatterplot. This gives the vertical distance that point is from the regression line.

Residual = $Y - \hat{Y}$

The standard deviation of the residual errors is an average of the residuals. The actual formula is shown below. Notice that we divide by n – 2 instead of n – 1 because there were two data sets. This again is called the degrees of freedom and will be discuss more in later chapters.

$$s_e = \sqrt{\frac{\sum(y - \hat{y})^2}{n-2}}$$



Notice Statcato also gave us a scatterplot of the data with the regression line drawn. The regression line formula is at the bottom of the graph.

Practice Problems Section 4G

1. How can tell which variable should be the explanatory variable and which variable should be the response variable?

2. How can we use the correlation coefficient (r) to determine if there is strong positive correlation? How can we use the correlation coefficient (r) to determine if there is strong negative correlation? How can we use the correlation coefficient (r) to determine if there is no correlation?

3. What is the definition of the coefficient of determination (r^2)?

4. What are the two definitions for the standard deviation of the residual errors (s_e)?

5. What is the definition of the slope of the regression line?

6. What is the definition of the y-intercept of the regression line?

7. What is extrapolation? Why should we avoid extrapolation?

(#8-16) Directions: Go to www.matt-teachout.org, click on "Statistics" and then "Data Sets". Open the indicated data. Copy and paste the two indicated columns of quantitative data next to each other on a new Excel spreadsheet. Then copy the two columns together. Now go to www.lock5stat.com and click on StatKey. Under the "Descriptive Statistics and Graphs" menu click on "Two Quantitative Variables". Click on "Edit Data" and paste the two columns together into StatKey. Then answer indicated questions.

8. Open the cigarette data. Let the explanatory variable (X) represent the amount of nicotine (milligrams) and the response variable (Y) represent the amount of tar (milligrams).

   a) Look at the scatterplot and the correlation coefficient (r). Describe the strength and direction of the linear relationship.
   b) Square the correlation coefficient in StatKey to calculate r^2. This is also called the coefficient of determination. Write a sentence to explain r^2.
   c) Find the slope of the regression line. Write a sentence to explain the slope.
   d) Find the y-intercept. Write a sentence to explain the y-intercept. Does the y-intercept make sense in the context of this data?
   e) The standard deviation of the residual errors was 1.3 mg. Explain the two meanings of this statistic.
   f) Use the regression line formula to predict the amount of tar if a cigarette contains 1.2 mg of nicotine. How much error could there be in this prediction.

9. Open the cigarette data. Let the explanatory variable (X) represent the amount of nicotine (mg) and the response variable (Y) represent the amount of carbon monoxide in parts per million (PPM).

   a) Look at the scatterplot and the correlation coefficient (r). Describe the strength and direction of the linear relationship.
   b) Square the correlation coefficient in StatKey to calculate r^2. This is also called the coefficient of determination. Write a sentence to explain r^2.
   c) Find the slope of the regression line. Write a sentence to explain the slope.
   d) Find the y-intercept. Write a sentence to explain the y-intercept. Does the y-intercept make sense in the context of this data?
   e) The standard deviation of the residual errors was 2.3 PPM. Explain the two meanings of this statistic.
   f) Use the regression line formula to predict the amount of carbon monoxide if a cigarette contains 1.2 mg of nicotine. How much error could there be in this prediction.

10. Open the health data. Let the explanatory variable (X) represent the systolic blood pressure (mm of Hg) and the response variable (Y) represent the diastolic blood pressure (mm of Hg). Use the combined columns with 80 randomly selected adults. Do not separate by gender.

a) Look at the scatterplot and the correlation coefficient (r). Describe the strength and direction of the linear relationship.
b) Square the correlation coefficient in StatKey to calculate r^2. This is also called the coefficient of determination. Write a sentence to explain r^2.

c) Find the slope of the regression line. Write a sentence to explain the slope.
d) Find the y-intercept. Write a sentence to explain the y-intercept. Does the y-intercept make sense in the context of this data?
e) The standard deviation of the residual errors was 7.4579 mm of Hg. Explain the two meanings of this statistic.
f) Use the regression line formula to predict the diastolic blood pressure of a person who has a systolic blood pressure of 130. How much error might there be in that prediction?

11. Open the health data. Let the explanatory variable (X) represent the waist size in centimeters and the response variable (Y) represent the weight in pounds. Use the combined columns with 80 randomly selected adults. Do not separate by gender.

a) Look at the scatterplot and the correlation coefficient (r). Describe the strength and direction of the linear relationship.
b) Square the correlation coefficient in StatKey to calculate r^2. This is also called the coefficient of determination. Write a sentence to explain r^2.
c) Find the slope of the regression line. Write a sentence to explain the slope.
d) Find the y-intercept. Write a sentence to explain the y-intercept. Does the y-intercept make sense in the context of this data?
e) The standard deviation of the residual errors was 14.6809 pounds. Explain the two meanings of this statistic.
f) Use the regression line formula to predict the weight of a person who has a waist size of 100 cm. How much error might there be in that prediction?

12. Open the health data. Let the explanatory variable (X) represent the age in years and the response variable (Y) represent the cholesterol in milligrams per deciliter (mg/dL). Use the combined columns with 80 randomly selected adults. Do not separate by gender.

a) Look at the scatterplot and the correlation coefficient (r). Describe the strength and direction of the linear relationship.
b) Square the correlation coefficient in StatKey to calculate r^2. This is also called the coefficient of determination. Write a sentence to explain r^2.
c) Find the slope of the regression line. Write a sentence to explain the slope.
d) Find the y-intercept. Write a sentence to explain the y-intercept. Does the y-intercept make sense in the context of this data?
e) The standard deviation of the residual errors was 255.3625 mg/dL. Explain the two meanings of this statistic.
f) Use the regression line formula to predict the cholesterol of a person that is 40 years old. How much error might there be in that prediction?

13. Open the bear data. Let the explanatory variable represent the age of the bear in months and the response variable represent the length of the bear in inches.

a) Look at the scatterplot and the correlation coefficient (r). Describe the strength and direction of the linear relationship.
b) Square the correlation coefficient in StatKey to calculate r^2. This is also called the coefficient of determination. Write a sentence to explain r^2.
c) Find the slope of the regression line. Write a sentence to explain the slope.
d) Find the y-intercept. Write a sentence to explain the y-intercept. Does the y-intercept make sense in the context of this data?
e) The standard deviation of the residual errors was 7.51 inches. Explain the two meanings of this statistic.
f) Use the regression line formula to predict the length of a bear that is 24 months old. How much error might there be in that prediction?

14. Open the bear data. Let the explanatory variable represent the neck circumference of the bear and the response variable represent the weight of the bear in pounds.

a) Look at the scatterplot and the correlation coefficient (r). Describe the strength and direction of the linear relationship.
b) Square the correlation coefficient in StatKey to calculate r^2. This is also called the coefficient of determination. Write a sentence to explain r^2.
c) Find the slope of the regression line. Write a sentence to explain the slope.
d) Find the y-intercept. Write a sentence to explain the y-intercept. Does the y-intercept make sense in the context of this data?
e) The standard deviation of the residual errors was 43.9 pounds. Explain the two meanings of this statistic.
f) Use the regression line formula to predict the weight of a bear that has a neck circumference of 24 inches. How much error might there be in that prediction?

15. Open the car data. Let the explanatory variable (X) represent the weight of the car in tons and the response variable (Y) represent the gas mileage in miles per gallon.

a) Look at the scatterplot and the correlation coefficient (r). Describe the strength and direction of the linear relationship.
b) Square the correlation coefficient in StatKey to calculate r^2. This is also called the coefficient of determination. Write a sentence to explain r^2.
c) Find the slope of the regression line. Write a sentence to explain the slope.
d) Find the y-intercept. Write a sentence to explain the y-intercept. Does the y-intercept make sense in the context of this data?
e) The standard deviation of the residual errors was 2.8516 mpg. Explain the two meanings of this statistic.
f) Use the regression line formula to predict the mpg for a car that weighs 3 tons. How much error might there be in that prediction?

16. Open the car data. Displacement is the amount of liquid in cubic centimeters forced out by the piston. Let the explanatory variable (X) represent the horsepower of the car and the response variable (Y) represent the displacement of the car (cc).

a) Look at the scatterplot and the correlation coefficient (r). Describe the strength and direction of the linear relationship.
b) Square the correlation coefficient in StatKey to calculate r^2. This is also called the coefficient of determination. Write a sentence to explain r^2.
c) Find the slope of the regression line. Write a sentence to explain the slope.
d) Find the y-intercept. Write a sentence to explain the y-intercept. Does the y-intercept make sense in the context of this data?
e) The standard deviation of the residual errors was 44.138 cubic centimeters. Explain the two meanings of this statistic.
f) Use the regression line formula to predict the number of cc's of displacement for a car with 120 horsepower. How much error might there be in that prediction?

Section 4H – Quantitative Relationships: The Correlation Test

We saw in the last section, that two quantitative samples are related if their correlation coefficient (r) is close to 1 or −1. When the correlation coefficient (r) is close to zero, the two quantitative samples are not related. How does this apply to populations? What if we want to determine if there is a relationship between two quantitative variables in a population? For this, we will need to look at the correlation hypothesis test.

The Correlation Hypothesis Test

If the sample correlation coefficient (r) is zero, tells us that the two quantitative samples are not related. For populations, we need to look at the population correlation coefficient "rho" (ρ). While this looks like a "p", it is not. It is the Greek letter "rho" and represents the population correlation coefficient.

If you recall from the last section, the correlation coefficient is related to the slope of the regression line.

Sample Slope $b_1 = \frac{(correlation\ coefficient\ times\ standard\ deviation\ of\ the\ y\ values\ )}{standard\ deviation\ of\ the\ x\ values} = \frac{(r \times s_y)}{s_x}$

So if there is no correlation between variables the correlation coefficient and the slope both go to zero. This principle applies to populations as well. As the population correlation coefficient "rho" (ρ) goes to zero, the population slope "Beta 1" (β_1) also goes to zero.

Correlation Test Null and Alternative Hypothesis

There are several ways of writing the null and alternative hypothesis for a correlation hypothesis test. We can use the population correlation coefficient "rho" (ρ) or the population slope "Beta 1" (β_1). We can also specify positive or negative correlation. Remember the correlation coefficient and the slope always have the same sign.

To show positive correlation the correlation coefficient should be close to +1 (greater than zero) and the slope should also be significantly positive (greater than zero). A positive relationship is also called a "direct" relationship. As the X variable increases, the Y variable also tends to increase. As the X variable decreases, the Y variable also tends to decrease.

To show negative correlation the correlation coefficient should be close to −1 (less than zero) and the slope should also be significantly negative (less than zero). A negative relationship is also called an "indirect" or "inverse" relationship. As the X variable increases, the Y variable also tends to decrease. As the X variable decreases, the Y variable also tends to increase.

Note about Statcato: Statcato only has the option for the two-tailed correlation test and cannot specify positive correlation (right-tailed) or negative correlation (left-tailed) hypothesis tests.

Two-Tailed Correlation Test: *For determining if variables are related or not. Does not specify if the direction is positive or negative.*

H_0 : ρ = 0 *(The two quantitative variables in the population are not related.)*
H_A : ρ ≠ 0 *(The two quantitative variables in the population are related.)*

OR

H_0 : β_1 = 0 *(The two quantitative variables in the population are not related.)*
H_A : β_1 ≠ 0 *(The two quantitative variables in the population are related.)*

Right-Tailed Correlation Test: *For determining if variables have a positive (or direct) relationship or not. Notice the alternative hypothesis symbol ">" points to the right.*

H_0 : ρ = 0 *(The two quantitative variables in the population are not related.)*
H_A : ρ > 0 *(The two quantitative variables in the population have a positive (direct) relationship.)*

OR

H_0 : β_1 = 0 *(The two quantitative variables in the population are not related.)*
H_A : β_1 > 0 *(The two quantitative variables in the population have a positive (direct) relationship.)*

Left-Tailed Correlation Test: *For determining if variables have a negative (or inverse) relationship or not. Notice the alternative hypothesis symbol "<" points to the left.*

H_0 : ρ = 0 *(The two quantitative variables in the population are not related.)*
H_A : ρ < 0 *(The two quantitative variables in the population have a negative (inverse) relationship.)*

OR

H_0 : β_1 = 0 *(The two quantitative variables in the population are not related.)*
H_A : β_1 < 0 *(The two quantitative variables in the population have a negative (inverse) relationship.)*

T-test statistic

The relationship between correlation and the slope of the regression line is highlighted in the test statistic. For a correlation test, you can use either the correlation coefficient "r" or a T-test statistic. I prefer the T-test statistic. The null hypothesis is that there is not a relationship between the quantitative variables. This would indicate that that the correlation coefficient and the slope would be close to zero. So the T-test statistic counts how many standard error the slope is from zero. If the T-test statistic is positive, the slope will be above zero and if the T-test statistic is negative, the slope will be below zero.

T-test statistics for correlation = $\frac{(slope - 0)}{standard\ error}$

T-test statistics sentence for Correlation: The number of standard errors that the slope of the regression line is above or below zero.

As with all test statistics, we will want to see if the T-test statistic falls in a tail determined by the critical value or values. If so, the sample data significantly disagrees with the null hypothesis and the slope is significantly different from zero. If the T-test statistic does not fall in a tail determined by the critical value or values, then the sample data does not significantly disagree with the null hypothesis and the slope is not significantly different from zero.

Residual Errors

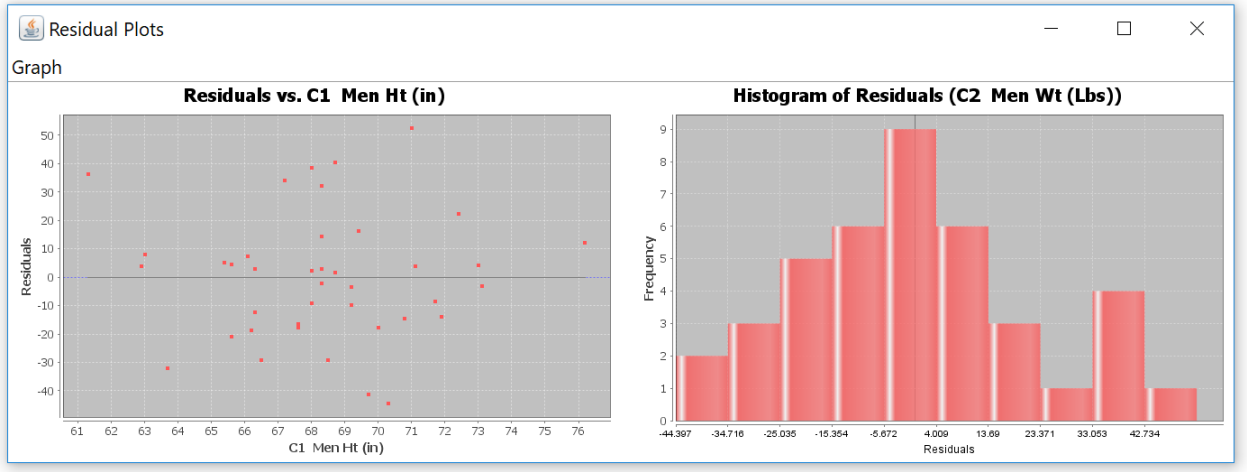
The correlation test has many assumptions. Some of the assumptions are centered on the understanding of "residuals" or "residual errors". We learned in the last section that a residual is the vertical distance between each point in the scatterplot and the regression line. To calculate a residual, the computer subtracts the actual y coordinate of the point minus the predicted ŷ value on the regression line. We also saw that the average of all the residuals is called the standard deviation of the residual errors (S_e). This tells us the average vertical distance that the data is from the regression line and the average prediction error.

Residual = $y - \hat{y}$

$$S_e = \sqrt{\frac{\Sigma(y-\hat{y})^2}{n-2}}$$

Besides the standard deviation of the residual errors, there are also residual graphs that statisticians often like to examine when doing a correlation test. We will only look at two. They are the "histogram of the residuals errors" and the "residual plot verses the x-values".

Here is an example. These graphs were created with Statcato. The explanatory variable (X) was the height of men and the response variable (Y) was the weight of men.



Residual Plot

The graph on the left is called the "residual plot verses the x-variable". This graph shows the vertical distances that each point is from the regression line. A point that is 40 above the line will have a residual of +40. A point that is 19 below the line will have a residual of −19. The zero line represents the regression line since points on the regression line have a residual of zero. We want the residual plot to be evenly spread out. When the points are evenly spread out, our standard deviation of the residuals is a consistent measure of spread. When the residual plot is not evenly spread out, you will see parts of the x-axis where all the points are very close and other parts of the x-axis where the points are very far away. This is an uneven spread (or fan shaped). We want the standard deviation to be a consistent measure of spread for all x value in the scope. If the points are close for some x values, then the standard deviation will be an overestimate of the variability for those x values. Similarly, if the points are far away for other x values, then the standard deviation will be an underestimate of the variability for those x values. Residual plots can be very difficult to read. I tell my intro students to put all the points on the left side of the graph between your fingers. Now put all the points on the right side of the graph between your fingers. If your fingers are about the same width on both the left and right side, you are probably ok. The data is evenly spread out and the standard deviation is a consistent measure of spread (variability). If your fingers are much closer on one side than the other, that may indicate a fan shape or uneven spread. In that case, the standard deviation is not a consistent measure of variability. Notice that points with an x-value greater than 72 are much closer to the regression line than those below 72. This could indicate an uneven spread (fan shape). This also could indicate that the regression line predictions are more accurate for taller men in the data (over 72 inches) and less accurate for shorter men in the data.

Histogram of the Residuals

The graph on the right is called the "histogram of the residuals". Remember that the calculation of the regression line uses the mean and standard deviation. If you remember from previous chapters, the mean and standard deviations are only accurate for normal data. We could check the shape of each data set separately, but instead we prefer to check the shape of the residuals. The histogram of the residuals should be normal (bell shaped). It should also be centered close to zero. Statcato gives a dark vertical line at zero for this purpose. This line should be close to the highest bar in the histogram. This histogram above passes both criteria.

Let us look at the assumptions for a correlation test.

Correlation Test Assumptions

1. The quantitative ordered pair data should be collected randomly or be representative of the population. *(The two samples usually have different units, but must have a one-to-one pairing).*
2. Data values within the sample should be independent of each other. *(The two samples are not independent since they are ordered pair. The individual data values within each sample should be independent. If you have small simple random sample from a large population, then the data values are probably not related.)*
3. The sample size should be at least 30. *(There should be 30 or more ordered pairs.)*
4. The scatterplot and correlation coefficient (r) should show some linear pattern. *(The correlation coefficient (r) should not be close to zero.)*
5. There should be no influential outliers in the scatterplot. *(If your correlation coefficient is close to 1 or −1, then you probably have no influential outliers. Remember to look for outliers on the scatterplot. A residual plot magnifies the distances, so everything looks like an outlier in a residual plot.)*
6. The histogram of the residuals should be nearly normal.
7. The histogram of the residuals should be centered close to zero. *(The zero line should be touching the highest bar in the histogram, or at least very close to the highest bar.)*
8. The residual plot verses the x variables should be evenly spread out with no fan shape or sideways "V" pattern. *(Put all the points in the residual plot between your fingers on the left side of the graph. Now put all the points in between your fingers on the right side. If your fingers are about the same width apart on the left and right side, the graph is close to evenly spread out.)*

Correlation Test Example 1

Let us use Statcato and the random "Health" data at www.matt-teachout.org to test the claim that there is no relationship between the age of a man and his cholesterol. In the Health data, we have the ages and cholesterol of forty randomly selected men. We will designate the age to be the explanatory variable (X) and the cholesterol to be the response variable (Y). Let us use a 5% significance level.

We can write the null and alternative hypothesis in one of two ways. We can use the population correlation coefficient "rho" (ρ) or the population slope "beta 1" (β_1). Remember our claim is "not related" so that must be the null hypothesis. Since positive or negative relationship was not mentioned, we will assume this is the general two-tailed test.

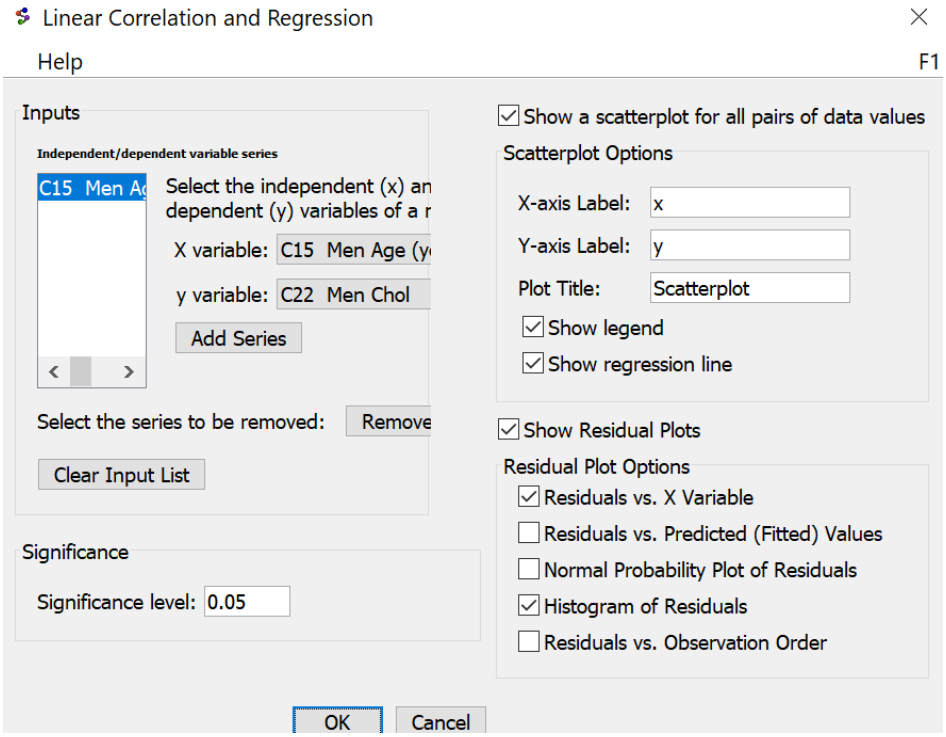
H_0 : ρ = 0 *(The age and cholesterol of men are not related.) CLAIM*
H_0 : ρ ≠ 0 *(The age and cholesterol of men are related.)*

OR

H_0 : β_1 = 0 *(The age and cholesterol of men are not related.) CLAIM*
H_0 : β_1 ≠ 0 *(The age and cholesterol of men are related.)*

Copy and paste the men's age and cholesterol data into two columns of Statcato. Go to the "statistics" menu, click on "correlation and regression" and then click on "linear". Click on the men's age to be the X-variable and the men's cholesterol to be the Y-variable and then push "add series". Check the box that says "show scatterplot" and the box that says "show regression line". Statcato also has the capability of making residual plots. Check the box that says, "Show residual plots", the box that says "residuals vs x-variable", and the box that says "histogram of the residuals". Now push "OK". Here is the Statcato printout, with the test statistic, P-value, correlation coefficient and all of the graphs.

Note: Some versions of Statcato do not have residual plots.



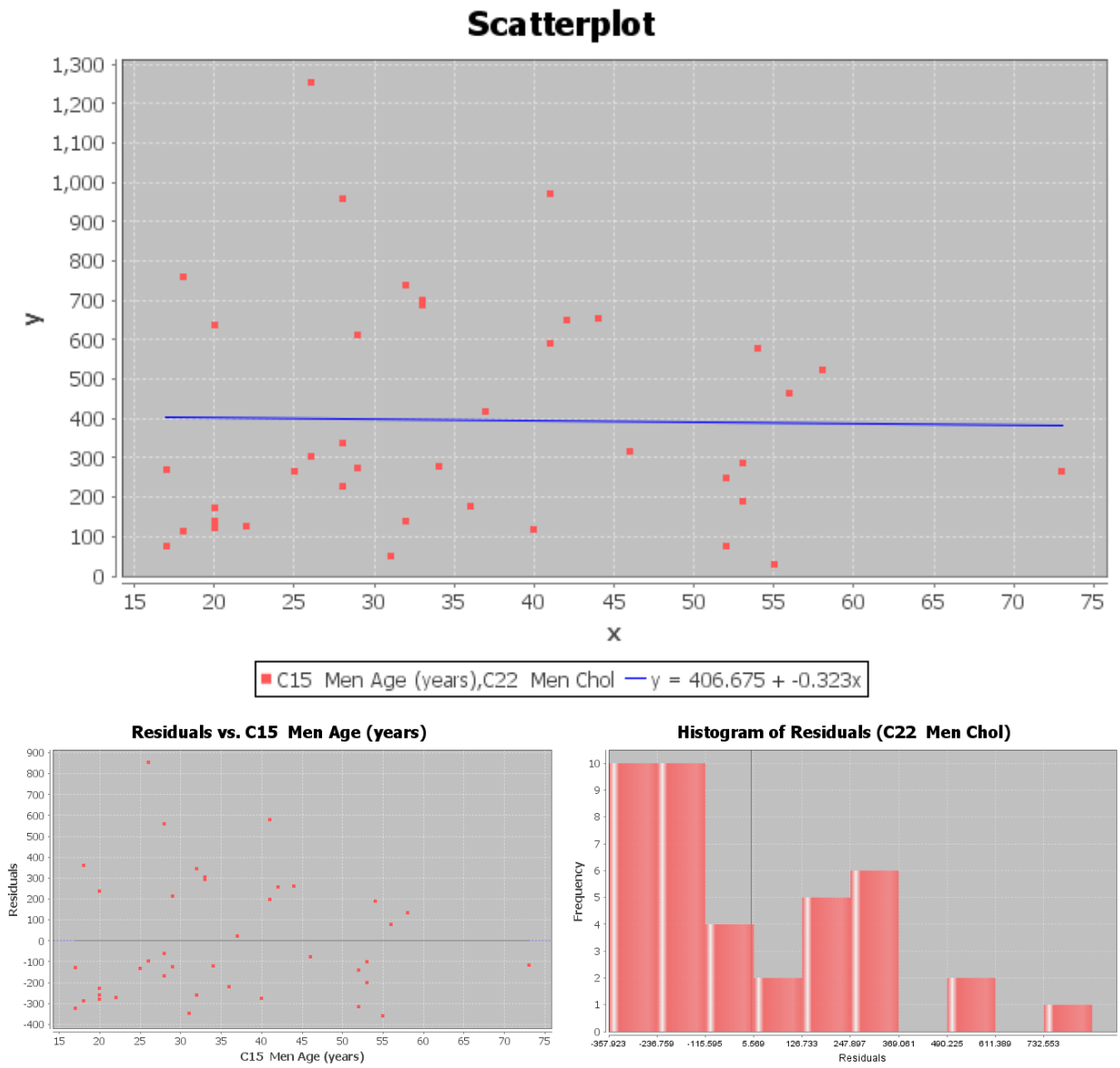
Correlation and Regression: Significance level = 0.05

Series: C15 Men Age (years),C22 Men Chol
x = C15 Men Age (years)
y = C22 Men Chol
Sample size n = 40
Degrees of freedom = 38

Correlation:
H_0: ρ = 0 (no linear correlation)
H_1: ρ ≠ 0 (linear correlation)

	Test Statistic	Critical Value
r	-0.0154	±0.3120
t	-0.0948	±2.0244

p-Value = 0.9250



Let us start by checking the assumptions for the men's age and cholesterol problem. Notice that this data fails many of the assumptions. That means our hypothesis test is compromised. We should also not use this regression line to make predictions about men's cholesterol.

1. Two quantitative ordered pair random samples. Yes. Age and cholesterol are both quantitative. The data had randomly selected men with the age and cholesterol of each man. It is ordered pair data.

2. Data values within each sample should be independent of each other. Yes. Since there is only forty randomly selected men out of millions of men in the population, the men are not likely to be related.

3. The sample size should be at least 30. Yes. There was forty men in the data. This is greater than thirty.

4. The scatterplot and correlation coefficient (r) should show some linear pattern. No. The regression line does not seem to fit the points in the scatterplot at all and the correlation coefficient r is very close to zero.

5. There should be no influential outliers in the scatterplot. No. There seem to be many influential outliers in the scatterplot and the correlation coefficient r is very close to zero.

6. The histogram of the residuals should be nearly normal. No. The histogram of the residuals is skewed right and not normal.

7. The histogram of the residuals should be centered close to zero. No. The histogram of the residuals seems to be centered to the left of zero. The zero line is not touching the highest bar in the histogram.

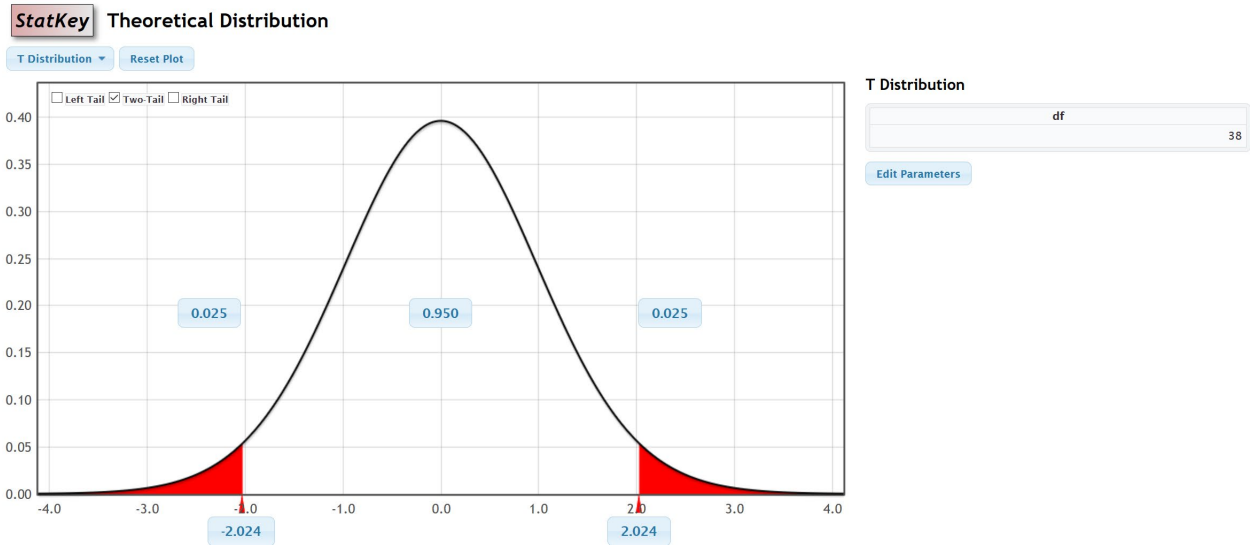
8. The residual plot verses the x variables should be evenly spread out. No. The residual plot seems to show a distinct fan shape and is not evenly spread out. The points on the left side of the graph seem to have a very wide spread while the points on the right side of the graph seem to be very close.

Test Statistic: T = −0.0948

Sentence: The slope of the regression line is 0.0948 standard errors below zero.

Our T-test statistic is −0.0948 and does not fall in either of the tails determined by the critical value. Our random sample data does not significantly disagree with the null hypothesis. This also indicates the slope is not significantly different from zero.

We put the degrees of freedom 38 into the theoretical T distribution calculator in StatKey to get the following picture.



↑
T = −0.0948

P-value = 0.9250

Sentence: If the null hypothesis is true and there is no relationship between the age and cholesterol for men, then there is a 92.5% probability of getting the sample data or more extreme because of sampling variability.

Notice the P-value is greater than our 5% significance level. This indicates that the sample data or more extreme could have occurred because of sampling variability if the null hypothesis was true. Since sampling variability cannot be ruled out, we must fail to reject the null hypothesis.

Fail to reject the Null Hypothesis.

We have a high P-value and the null hypothesis is the claim. The sample data did not pass all of the assumptions for the correlation test.

Conclusion: There is not significant evidence to reject the claim that the age and cholesterol of men is not related.

Age and cholesterol of men are probably not related. This sample data did not provide evidence since the P-value was high and it failed many of the assumptions for the correlation test.

Example 2

We can also use randomized simulation on StatKey to determine significance and calculate the P-value. StatKey can calculate the scatterplot and the correlation coefficient and slope, but does not calculate any of the residual graphs.

We are going to be using the "mpg weight horsepower" data on www.matt-teachout.org to test the claim that there is a negative (inverse) relationship between the weight of a car and the miles per gallon of gas (mpg). We will be using a 5% significance level and assume the data met all of the assumptions.

H_0 : ρ = 0 *(The weight and mpg of a car are not related.)*
H_A : ρ < 0 *(The weight and mpg of a car have a negative (inverse) relationship.) CLAIM*

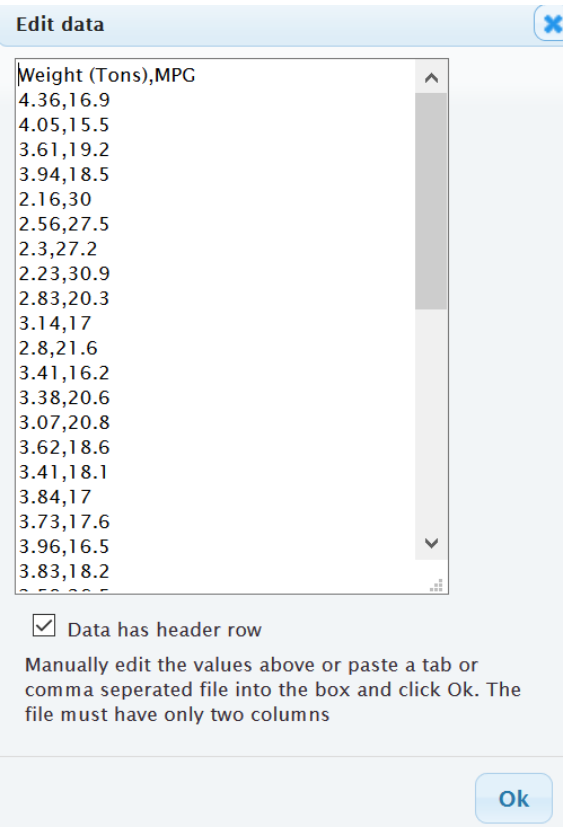
OR

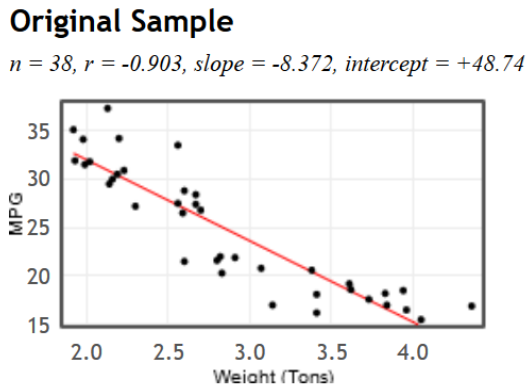
H_0 : β_1 = 0 *(The weight and mpg of a car are not related.)*
H_A : β_1 < 0 *(The weight and mpg of a car have a negative (inverse) relationship.) CLAIM*

We will designate the weight of the car as the explanatory variable (X) and the miles per gallon as the response variable (Y). Copy and paste the weight of the cars and mpg into a fresh excel spreadsheet. Put the weight data on the left and the mpg on the right. Now copy both columns together.

Weight (Tons)	MPG
4.36	16.9
4.05	15.5
3.61	19.2
3.94	18.5
2.16	30
2.56	27.5
2.3	27.2

Go to www.lock5stat.com and open StatKey. Under the "Randomized Hypothesis Tests" menu, click on "Test for Slope, Correlation". Under the "Edit Data" menu, paste in the weight and mpg columns. Since the data sets have titles, check the box that says, "Header has header row" and push OK. Under "Original Sample" we see the scatterplot, correlation coefficient (r) and the sample slope (b_1).





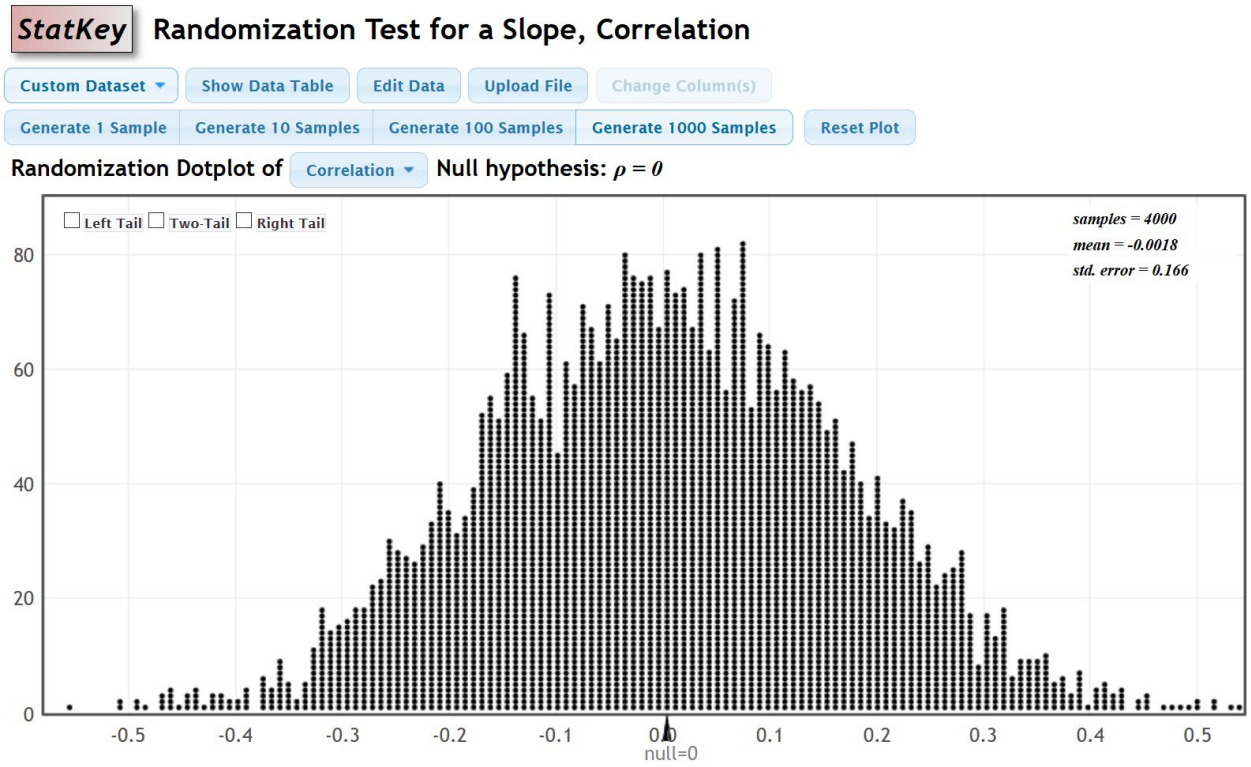
Let us give a quick analysis of the sample data as we did in the last section. We see that the scatterplot and the correlation coefficient (r) show a strong negative relationship between the samples. Notice that r = −0.903 and is close to −1. The points in the scatterplot seem to be close to the regression line and there does not appear to be any influential outliers.

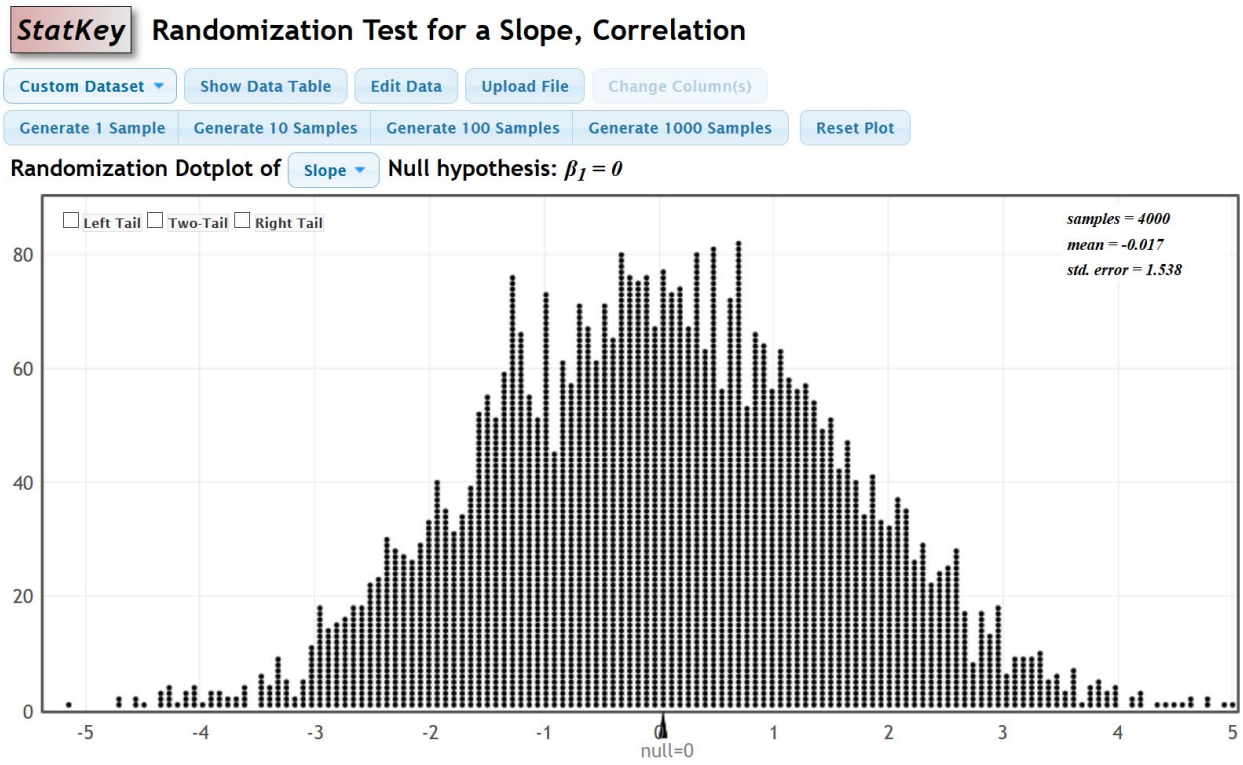
The slope is −8.372. In our last section, we saw that the slope is the amount of increase or decrease in the Y variable per unit of X. Since the slope was negative, it is a decrease. In addition, the X variable is the weight of the car in tons and the Y variable is the gas mileage in miles per gallon.

Sample Slope Sentence: For every 1 ton heavier the car, the average miles per gallon of the cars in the samples are decreasing 8.372 mpg.

Randomized Simulation

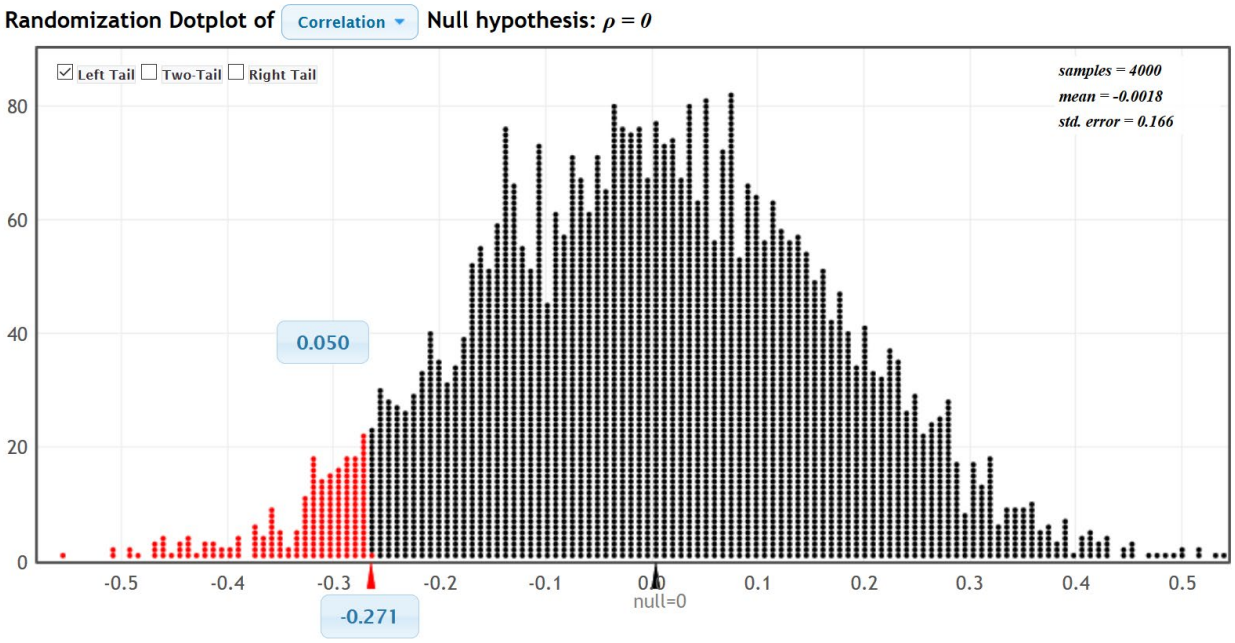
There are two ways to do the randomized simulation. We can have the computer create thousands of random samples and calculate the correlation coefficient for each. Another way is to have the computer create thousands of random samples and calculate the slope for each. At the top of the distribution, you will see we can change the setting to "correlation" or "slope". Notice how the null hypothesis changes to reflect the setting. Click on "Generate 1000 Samples" a few times.





Simulating with the Correlation Coefficient

Let us start with looking at the correlation coefficient simulation. These are thousands of correlation coefficients. When the setting is on "Correlation", we will need to use the "Original Sample" correlation coefficient (r) to determine significance and calculate the P-value. Since the alternative hypothesis was less than "<", this was a left-tailed test. Click on left tail. Since we are using a 5% significance level, we will put in 0.05 in the left tail proportion. Notice the simulation indicates that our "Original Sample" correlation coefficient (r) needs to be −0.271 or less to be significant. Our "Original Sample" correlation coefficient (r) is −0.903 and definitely falls in the left tail.

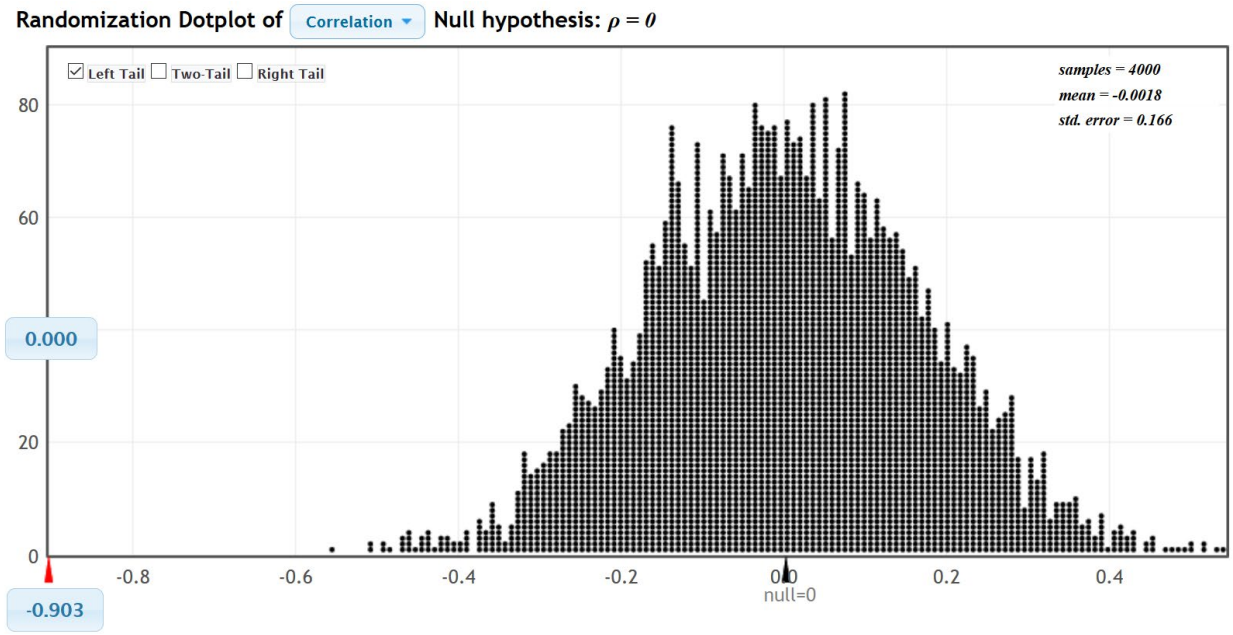


↑
r = −0.903

Since our correlation coefficient r falls in the left tail of the simulation, the sample data significantly disagrees with the null hypothesis.

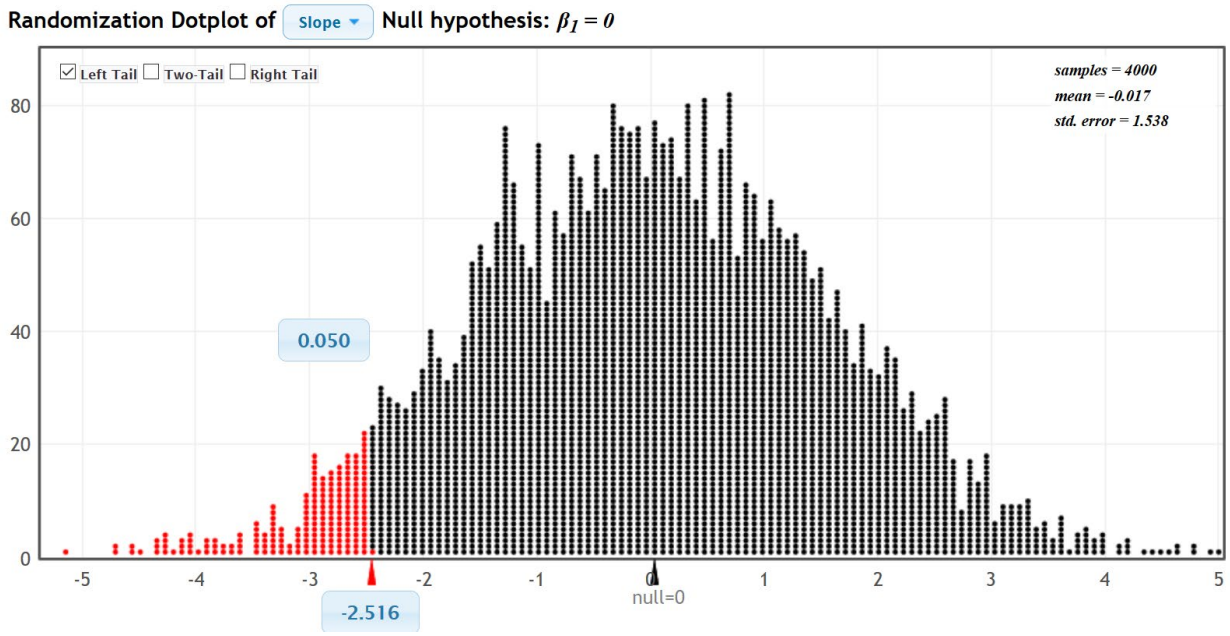
Calculating the P-value

The P-value is the probability of getting the sample data or more extreme by sampling variability if the null hypothesis is true. This simulated distribution is a view of sampling variability if the null is true. We just need to figure out the probability of the sample data or more extreme. Since this simulation created thousands of correlation coefficients, we will enter the real "original sample" correlation coefficient (r = −0.903) in the bottom box of the simulation. The left tail probability will give us the probability we are looking for. In this case, the P-value was approximately zero.



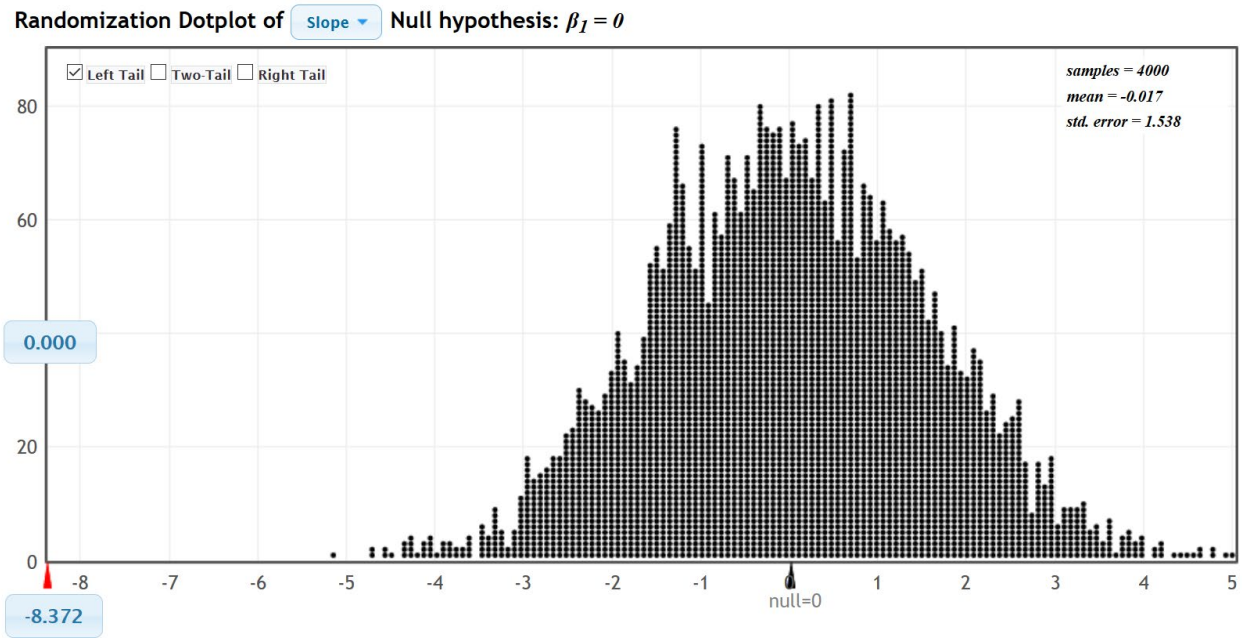
Simulating with the Slope

We can simulate with either the correlation coefficient or the slope. Here is the randomized simulation of thousands of sample slopes. Putting the 5% significance level in the tail, shows us that the real "original sample" slope needs to be −2.516 or less to be in the left tail. So if the real "original sample" slope is less than −2.516, the sample data will significantly disagree with the null hypothesis. The real "original sample" slope is −8.372 so it does fall in the left tail.



↑
Slope = −8.372

Now let us calculate the P-value with the slope. Enter the real "original sample" slope in the bottom box in the left tail. We see that the P-value is zero. Notice this is the same P-value as we got when we simulated with the correlation coefficient.



Notice that the original sample slope or correlation coefficient fell in the tail. So the sample data significantly disagrees with the null hypothesis. The slope is significantly different from zero.

What is the T-test statistic? Remember in a simulation, you do necessarily have to use the test statistic to judge significance. We used the sample correlation coefficient and slope to judge significance. We can calculate the T-test statistic though using the formula. Notice in the slope simulation, the approximate standard error for this simulation is 1.538. The standard error will vary between simulations though.

T-test statistic (for the correlation test) = $\frac{(Slope - Zero)}{Standard\ Error} = \frac{(-8.372 - 0)}{1.538} \approx -5.443$

T-test statistics Sentence: The slope of the regression line is 5.443 standard errors below zero.

P-value ≈ 0

P-value sentence: If the null hypothesis is true and there is no relationship between the weight of a car and the miles per gallon, then there is zero probability of getting this sample data or more extreme by sampling variability.

The P-value also tells us that it is extremely unlikely for this sample data to occur because of sampling variability.

The P-value is less than our 5% significance level, so we will reject the null hypothesis.

Conclusion: There is significant evidence to support the claim that there is a negative (inverse) relationship between the weight of a car and the miles per gallon. This does not imply that a heavy car causes the car to have

Notes

- Remember, if you simulate with the correlation coefficient, then you have to use the real "original sample" correlation coefficient when you calculate the approximate P-value. If you simulate with the slope, then you have to use the real "original sample" slope when you calculate the approximate P-value.
- You do not have to simulate with both the correlation coefficient and the slope. The point is that either simulation gives you approximately the same P-value.
- In all randomized simulations, there is sampling variability. Answers will vary slightly in different simulations.

Practice Problems Section 4H

(#1-10) Use either the correlation coefficient or the T-test statistic and the corresponding critical values to fill out the table.

	T-test statistic or Correlation Coefficient (r)	Sentence to explain T-test statistic or Correlation Coefficient (r)	Critical Value (T or r)	Does the T-test statistic or r-value fall in a tail determined by a critical value? (Yes or No)	Does sample data significantly disagree with H_0?
1.	T = −2.441		±1.775		
2.	r = 0.183		0.316		
3.	T = +1.166		+2.003		
4.	r = −0.799		±0.286		
5.	T = +3.118		+2.714		
6.	r = 0.921		0.339		
7.	T = −0.852		±2.322		
8.	r = −0.026		−0.279		
9.	T = +1.339		±1.997		
10.	r = 0.483		+0.303		

(#11-20) Use each of the following P-values and corresponding significance levels to fill out the table.

	P-value Proportion	P-value %	Sentence to explain the P-value	Significance Level %	Significance level Proportion	If H_0 is true, could the sample data occur by random chance or is it unlikely?	Reject H_0 or Fail to reject H_0?
11.	0.521			10%			
12.	0.0426			1%			
13.	3.41×10^{-5}			5%			
14.	0.0033			1%			
15.	0.768			5%			
16.	0			10%			
17.	0.0428			5%			
18.	0.277			10%			
19.	6.04×10^{-6}			1%			
20.	0.0178			5%			

21. List the assumptions that we need to check when performing a correlation hypothesis test.

22. How can we use the scatterplot and the correlation coefficient (r) to determine if the sample data follows a linear pattern?

23. Points in the scatterplot that are far from the regression line are considered outliers, but it is difficult to know if the outliers are influential or not. How can we use the scatterplot and the correlation coefficient (r) to determine if potential outliers are influential or not?

24. Explain the two assumptions that we check by using the histogram of the residuals.

25. Explain how to determine if the residual plot is evenly spread out or not.

(#26-29) Directions: For each of the following problems, use the Statcato printouts provided to answer the following questions.

a) Write the null and alternative hypothesis for the correlation test. Address the quantitative relationship and label which is the claim.

b) Write a sentence to explain the strength and direction based on the correlation coefficient (r).

c) Write a sentence to explain the sample slope (b_1).

d) Check all of the assumptions for the correlation test. Explain your answers.

e) Write a sentence to explain the T-test statistic.

f) Compare the T-test statistic to the critical value. Does the test statistic fall in a tail determined by the critical value?

g) Does the sample data significantly disagree with the null hypothesis? Explain your answer.

h) Is the sample slope significantly different from zero? Explain your answer.

i) Write a sentence explaining the P-value.

j) Compare the P-value to the significance level. Could the sample data or more extreme occur by sampling variability if the null hypothesis was true or is it unlikely? Explain your answer.

k) Should we reject the null hypothesis or fail to reject the null hypothesis? Explain your answer.

l) Write a conclusion for the test addressing evidence and the claim.

26. Use a 5% significance level and the Statcato printout below to test the claim that there is a linear relationship between the height (X) of a man and his weight (Y). This printout came from the random health data at www.matt-teachout.org.

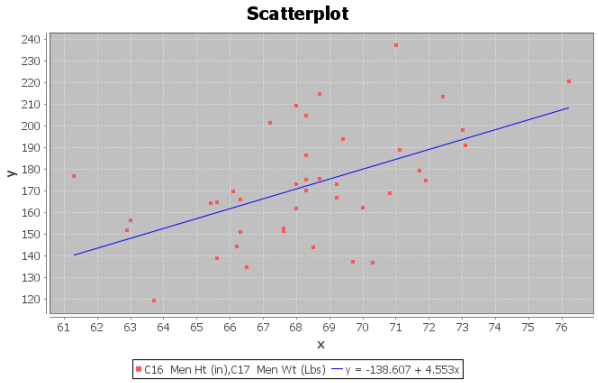
Correlation and Regression: Significance level = 0.05

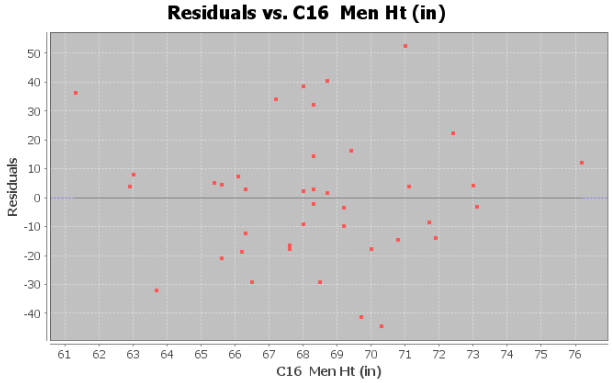
x = C16 Men Ht (in)
y = C17 Men Wt (Lbs)
Sample size n = 40
Degrees of freedom = 38

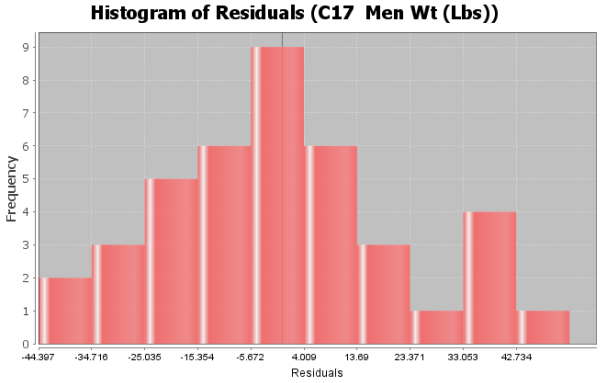
	Test Statistic	Critical Value
r	0.5222	±0.3120
t	3.7750	±2.0244

p-Value = 0.0005

Regression:
Regression equation $Y = b_0 + b_1x$
$b_0 = -138.6070$
$b_1 = 4.5534$







27. Use a 5% significance level and the Statcato printout below to test the claim that there is NO linear relationship between the systolic blood pressure (X) of a woman and her diastolic blood pressure (Y). This printout came from the random health data at www.matt-teachout.org.

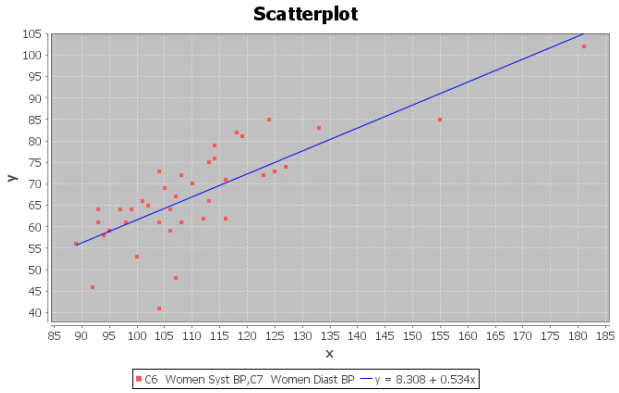
Correlation and Regression: Significance level = 0.05

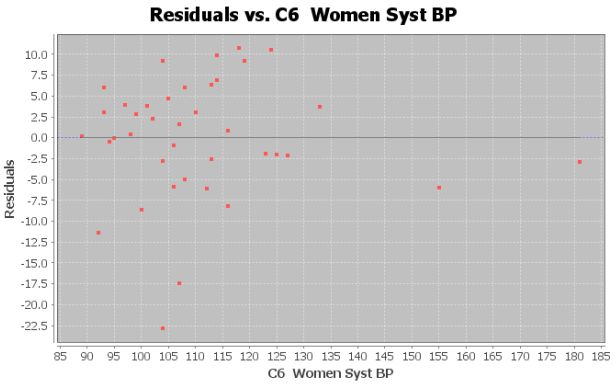
x = C6 Women Syst BP
y = C7 Women Diast BP
Sample size n = 40
Degrees of freedom = 38

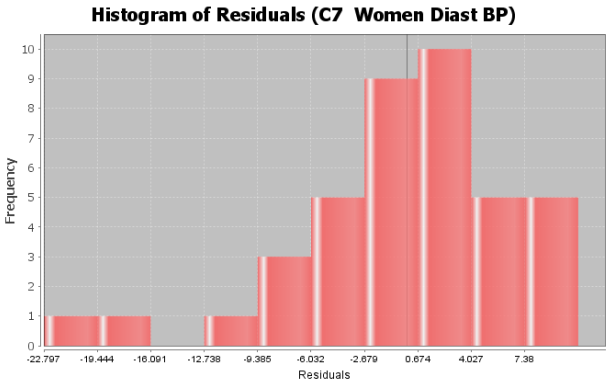
	Test Statistic	Critical Value
r	0.7854	±0.3120
t	7.8209	±2.0244

p-Value = $1.9615 \cdot 10^{-9}$

Regression:
Regression equation $Y = b_0 + b_1x$
$b_0 = 8.3079$
$b_1 = 0.5335$







28. Use a 5% significance level and the Statcato printout below to test the claim that there is a relationship between the head width (X) of a bear and its chest size (Y). This printout came from the random bear data at www.matt-teachout.org.

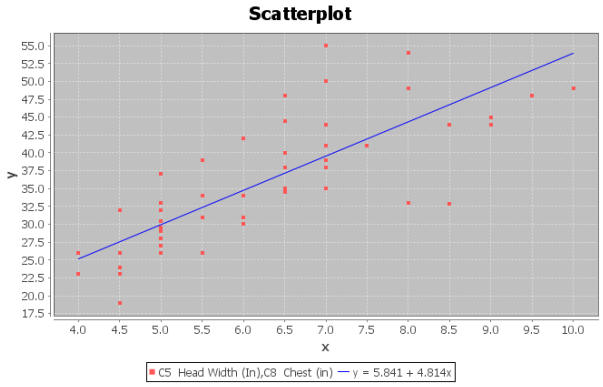
Correlation and Regression: Significance level = 0.05

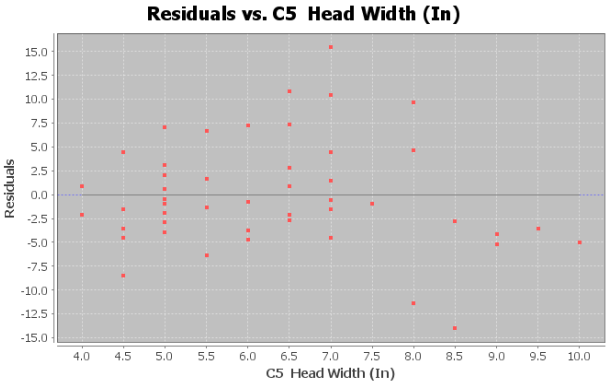
x = C5 Head Width (In)
y = C8 Chest (in)
Sample size n = 54
Degrees of freedom = 52

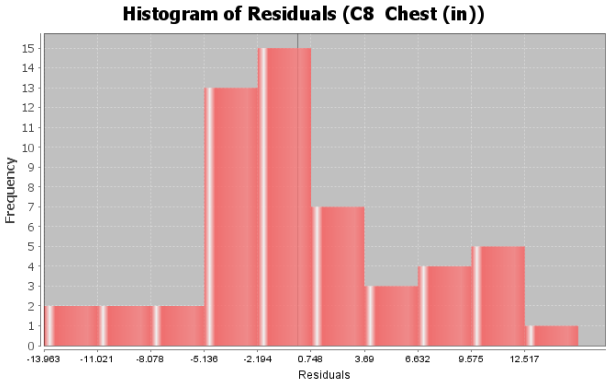
	Test Statistic	Critical Value
r	0.7785	±0.2681
t	8.9451	±2.0067

p-Value = $4.2208 \cdot 10^{-12}$

Regression:
Regression equation $Y = b_0 + b_1x$
$b_0 = 5.8408$
$b_1 = 4.8143$







29. Use a 5% significance level and the Statcato printout below to test the claim that there is NO relationship between the neck circumference (X) of a bear and its weight (Y). This printout came from the random bear data at www.matt-teachout.org.

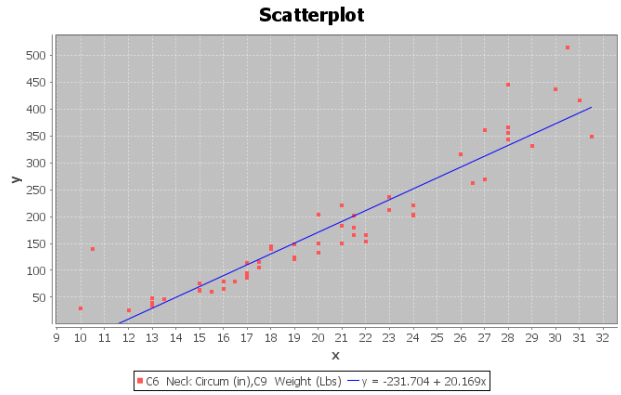
Correlation and Regression: Significance level = 0.05

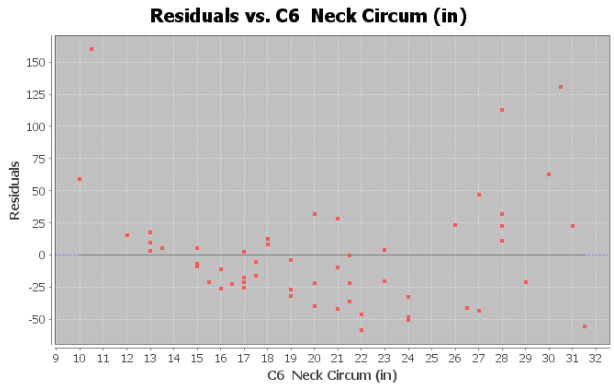
x = C6 Neck Circum (in)
y = C9 Weight (Lbs)
Sample size n = 54
Degrees of freedom = 52

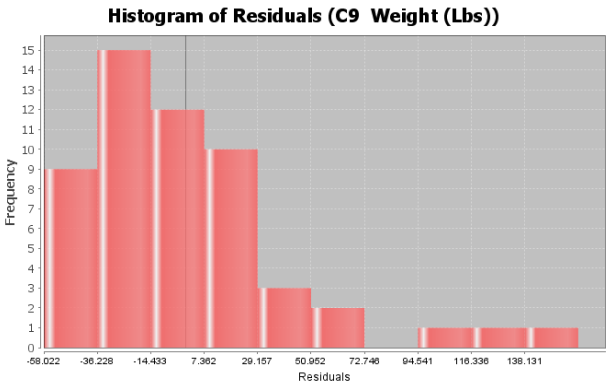
	Test Statistic	Critical Value
r	0.9341	±0.2681
t	18.8612	±2.0067

p-Value = 0

Regression:
Regression equation $Y = b_0 + b_1x$
$b_0 = -231.7044$
$b_1 = 20.1694$







(#30-32) Directions: Go to www.lock5stat.com and click on the "StatKey" button. Under "Randomization Hypothesis Tests", click the one that says, "Test for Slope, Correlation". Click "Generate 1000 Samples" a few times. Remember there are two ways of getting the P-value. If the top of the graph says "Randomization Dot plot of Correlation", then the null hypothesis is $\rho = 0$. *Remember rho looks like a "* ρ *" but it is not a "P". If the top of the graph says "Randomization Dot plot of Slope", then the null hypothesis is* $\beta_1 = 0$. *Remember when StatKey simulates correlation we will be comparing the original r-value to all the simulated r-values in the simulation. When StatKey simulates the slope, we will be comparing the original sample slope to the simulated slopes. You will get about the same P-value from either of these. Assume the assumptions are met. Use the simulation in StatKey to answer the following questions.*

a) Write the null and alternative hypothesis for the correlation test. Address the quantitative relationship and label which is the claim. Is this a right-tailed, left-tailed, or two-tailed test?

b) Write a sentence to explain the strength and direction based on the "original sample" correlation coefficient (r).

c) Does the original sample correlation coefficient fall in a tail of the correlation simulation?

d) Write a sentence to explain the original sample slope (b_1).

e) Does the original sample slope fall in the tail of the slope simulation?

f) Is the sample slope significantly different from zero? Explain your answer.

g) Does the sample data significantly disagree with the null hypothesis? Explain your answer.

h) Put the original sample slope into the slope simulation to calculate the P-value. What is your estimated P-value? (Answers will vary.)

i) Write a sentence explaining the P-value.

j) Compare the P-value to the significance level. Could the sample data or more extreme occur by sampling variability if the null hypothesis was true or is it unlikely? Explain your answer.

k) Should we reject the null hypothesis or fail to reject the null hypothesis? Explain your answer.

l) Write a conclusion for the test addressing evidence and the claim.

m) Use the original sample slope, the estimated standard error in the simulation, and the following formula to calculate the T-test statistic. (Answers will vary.) Write a sentence to explain the T-test statistic.

T-test statistic = $\frac{(Slope-0)}{Standard\ Error}$

30. Open the "Car Data" in Excel from www.matt-teachout.org. Copy and paste the miles per gallon (mpg) and horsepower into two columns in new excel spreadsheet. The mpg should be on the left and the horsepower should be on the right. The mpg will be the explanatory variable (X) and the horsepower will be the response variable (Y). Now go to www.lock5stat.coma and click on StatKey. Under "Randomization Hypothesis Tests" click on "Test for Slope, Correlation". Under "Edit Data" paste the two columns into StatKey. Now click "Generate 1000 Samples" a few times. Use the randomized simulation in StatKey and a 1% significance level to test the claim that there is a negative (inverse) relationship between mpg and horsepower.

31. Open the "Car Data" in Excel from www.matt-teachout.org. Copy and paste the horsepower and weight into two columns in new excel spreadsheet. The horsepower should be on the left and the weight should be on the right. The horsepower will be the explanatory variable (X) and the weight in tons will be the response variable (Y). Now go to www.lock5stat.coma and click on StatKey. Under "Randomization Hypothesis Tests" click on "Test for Slope, Correlation". Under "Edit Data" paste the two columns into StatKey. Now click "Generate 1000 Samples" a few times. Use the randomized simulation in StatKey and a 10% significance level to test the claim that there is a positive (direct) relationship between the horsepower and weight of a car.

32. Open the "Health Data" in Excel from www.matt-teachout.org. Copy and paste the age of women and the height of women into two columns in new excel spreadsheet. The age of women should be on the left and the height of women should be on the right. The age of women in years will be the explanatory variable (X) and the height of women in inches will be the response variable (Y). Now go to www.lock5stat.coma and click on StatKey. Under "Randomization Hypothesis Tests" click on "Test for Slope, Correlation". Under "Edit Data" paste the two columns into StatKey. Now click "Generate 1000 Samples" a few times. Use the randomized simulation in StatKey and a 5% significance level to test the claim that there is NO relationship between the age and height of women.

Chapter 4 Review

1. Write down the definitions for the following key terms.

Correlation Coefficient (r) , R-squared , Standard Deviation of the Residual Errors , Slope , Residual , Y-Intercept , Explanatory Variable , Response Variable , Correlation , Regression , Scatterplot , Residual Plot , Regression Line , Histogram of the Residuals, Hypothesis Test , Sampling Variability (Random Chance) , P-value , Significance Level , Critical Value , Randomized Simulation , F-test statistic , Chi-Squared test statistic (χ^2) *for the Goodness of Fit or Categorical Relationship Test, T-test statistic for correlation, Z-test statistic for two-population proportion test, T-test statistic for a two-population mean test*

2. Write down the type of data and the null and alternative hypothesis for the following relationship hypothesis tests: Two-population proportion test, Goodness of Fit, Categorical Relationship Test, Two-population mean test, ANOVA, and Correlation.

3. Write down the assumptions for the following hypothesis tests: Two-population proportion test, Goodness of Fit, Categorical Association Test, Two-population mean test, ANOVA, and the Correlation Test.

4. Give the test statistic for each of the following hypothesis tests: Two-population proportion test, Goodness of Fit, Categorical Relationship Test, Two-population mean test, ANOVA, and Correlation.

5. Fill out the following table to interpret the given test statistics.

Test Statistic	Critical Value	Does the sample data significantly disagree with H_0?	Explain why.
F = 2.174	3.823		
T = −2.556	± 1.96		
χ^2 = 16.87	9.977		
F = 5.339	2.742		
T = 1.349	± 2.576		
χ^2 = 1.883	7.187		

6. Fill out the following table to interpret the given P-value.

P-value	P-value %	Significance Level	Does the sample data significantly disagree with H_0?	Could be random chance or Unlikely?	Reject H_0 or fail to reject?
0.238		5%			
0.0003		1%			
5.7 x 10^{-6}		10%			
0.441		5%			
0.138		1%			
0		10%			

7. Complete the table by writing the conclusions for the following.

P-value	Sig Level	Claim	Conclusion
0.238	5%	H_0	
0.0003	1%	H_A	
5.7 x 10^{-6}	10%	H_0	
0.441	5%	H_A	
0.138	1%	H_0	
0	10%	H_A	

8. If we want to see if two quantitative variables are related, what hypothesis test should we use? What would the test statistic be? What assumptions should we check?

9. If we want to see if two categorical variables with multiple options are related, what hypothesis test should we use? What would the test statistic be? What assumptions should we check?

10. If we want to see if categorical and quantitative variables are related, what hypothesis test should we use? What would the test statistic be? What assumptions should we check?

11. If we want to see if a categorical variables and a specific proportion are related, what hypothesis test should we use? What would the test statistic be? What assumptions should we check?

12. Suppose we want to see if the amount of money in peoples' checking accounts is related to city they live in. What hypothesis test should we use? Explain why.

13. Suppose we want to see if the percentage of people in a city that own an Android phone is related to the city they live. What hypothesis test should we use? Explain why.

14. Suppose we want to see if the amount of rainfall in areas across Europe is related to the number of fires in those areas. What hypothesis test should we use? Explain why.

15. Suppose we want to see if a person's type of health insurance is related to their education level. What hypothesis test should we use? Explain why.

16. An orthopedic surgeon that specializes in knee injuries is wondering if the proportion of knee injuries is the same for the various sports. (This would indicate that the percent of knee injuries is not related to the sport being played.) He looks through randomly selected knee injuries and finds the following data. What percentage of the knee injuries came from playing soccer? What percentage of the knee injuries came from playing tennis? What can these percentages tell us about the relationship? Since this data is looking at one proportion in six groups, what type of hypothesis test is this? Use the following Statcato printout and a 1% significance level to test the claim. Be sure to check expected values and the assumption necessary for the test. Give the chi-squared test statistic and the P-value, whether you reject the null hypothesis and a conclusion that the surgeon will understand. Write a sentence to explain the test statistic. Write a sentence to explain the P-value. Was the sample data significant? Explain why. Could the sample data have happened by random chance or is it unlikely? Explain why.

Football	Baseball	Basketball	Soccer	Hockey	Tennis
23	8	14	31	19	5

Chi-Square Goodness-of-Fit Test:
Input: C4 observed counts
Expected frequency = 16.6667

Category	Observed Frequency	Expected Frequency	Contribution to X^2
0	23.0	16.6667	2.4067
1	8.0	16.6667	4.5067
2	14.0	16.6667	0.4267
3	31.0	16.6667	12.3267
4	19.0	16.6667	0.3267
5	5.0	16.6667	8.1667

N	Number of Categories	DOF	Significance	Critical Value	Test statistics	p-Value
100.0	6	5	0.01	15.0863	28.16	$3.3869 \cdot 10^{-5}$

17. A forest ranger is looking into incidents of rabies among the animals. He thinks that the type of animal is related to whether or not they have rabies. What percentage of all the raccoons have rabies? What percentage of the squirrels have rabies? What percentage of the chipmunks have rabies? What can these probabilities show us about the relationship? This data was collected from one random sample of animals. From each animal, the type of animal was noted as well as their rabies status. Is this a Homogeneity test or an Independence test? Explain why. Use the following Statcato printout and a 5% significance level to test the claim that the type of animal is related to whether or not they have rabies. Be sure to check expected values and the assumption necessary for the test. Give the Chi-squared test statistic and the P-value, whether you reject the null hypothesis or fail to reject, and a conclusion that the ranger will understand. Write a sentence to explain the test statistic. Write a sentence to explain the P-value. Was the sample data significant? Explain why. Could the sample data have happened by random chance or is it unlikely? Explain why.

	Squirrels	Chipmunks	Raccoons
Has Rabies	17	8	7
Does not have Rabies	21	22	20

Chi-Square Test: Contingency Table:

	Squirrels	Chipmunks	Raccoons	Total
Rabies	17.0 (12.8) [1.38]	8.0 (10.11) [0.44]	7.0 (9.09) [0.48]	32.0
No Rabies	21.0 (25.2) [0.70]	22.0 (19.89) [0.22]	20.0 (17.91) [0.25]	63.0
Total	38.0	30.0	27.0	95.0

(expected frequency), [test statistic contribution]

Significance Level	DOF	χ^2	Critical value	p-Value
0.05	2	3.4670	5.9915	0.1767

18. Go to www.lock5stat.com and click on the "StatKey" tab. Then click on "χ^2 test for association". One the top left part of the page click on "one true love by gender". If it is not there, you can also click on "Edit Data" and type in the following contingency table. We want to determine if a persons belief about everyone having one true love is independent of (not related to) gender. What percent of the males believe that everyone has one true love? What percent of the females believe that everyone has one true love? What can these percentages tell us about the relationship? Use simulation and a 10% significance level to test the claim that gender and a persons' belief about one true love are independent (not related). Be sure to check expected values and the assumption necessary for the test. Give the chi-squared test statistic and the simulated P-value, whether you reject the null hypothesis and a conclusion that the music students will understand. Write a sentence to explain the test statistic. Write a sentence to explain the P-value. Was the sample data significant? Explain why. Could the sample data have happened by random chance or is it unlikely? Explain why.

[blank], Male, Female

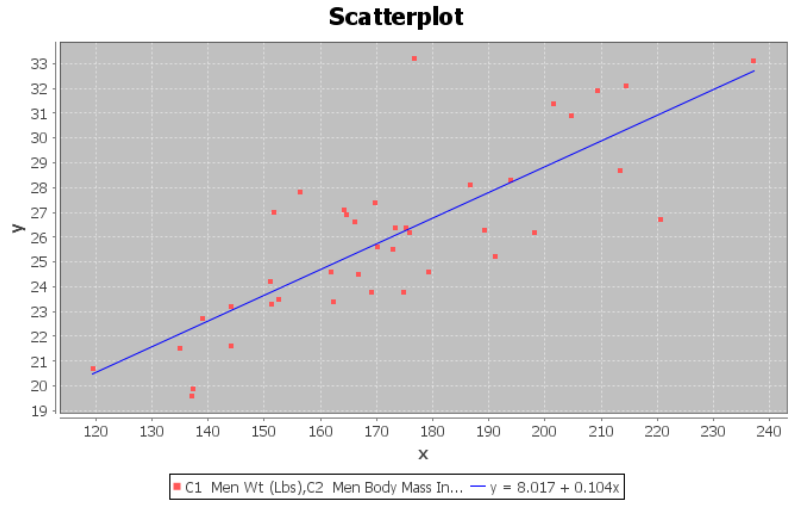
Agree, 372, 363

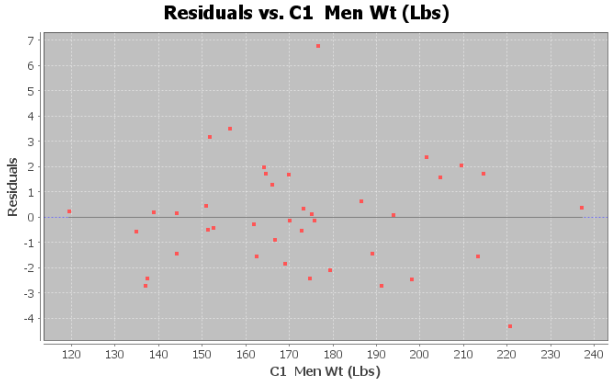
Disagree, 807, 1005

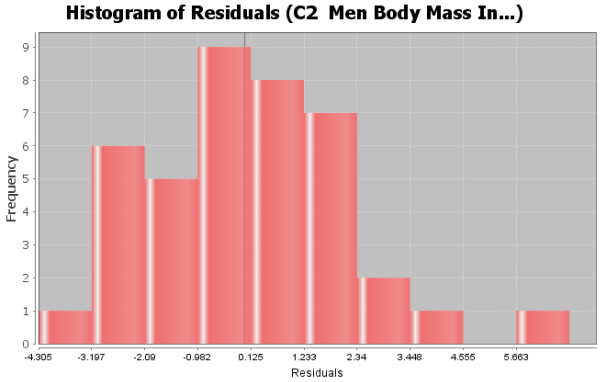
Don't Know, 34, 44

19. Go to www.matt-teachout.org and open the Math 140 Survey Fall 2019. Copy and paste the social media data and the money spent on meals data next to each other in a new Excel spreadsheet. The social media data should be on the left. Now copy the two columns together. Go to www.lock5stat.com and click on the "StatKey" button. Then click on "ANOVA for difference in means". Under "Edit Data", paste the two columns into StatKey and click "OK". Use simulation and a 5% significance level to test the claim that a math 140 students favorite social media is not related to how much they spend on meals. Be sure to check the assumptions, give the F-test statistic, null and alternative hypothesis, the simulated P-value, whether or not you reject the null hypothesis and a conclusion. Write a sentence to explain the F-test statistic. Write a sentence to explain the P-value. Was the sample data significant? Explain why. Could the sample data have happened by random chance or is it unlikely? Explain why.

20. We want to explore the relationship between the weight and body mass index for men. Let the explanatory variable (x) be the weight of the men and the response variable (y) be the body mass index (BMI) of the men. Write a sentence to explain the correlation coefficient r. Write a sentence to explain r-squared. Write two sentences to explain the two meanings of the standard deviation of the residual errors. Write a sentence to explain the meaning of the slope of the regression line. Write a sentence to explain the meaning of the y-intercept of the regression line. Use the regression line formula to predict the BMI of a man that weighs 185 pounds. How much error is there in that prediction? Use the following Statcato printout to perform a correlation hypothesis test with a 1% significance level to test the claim that there is a linear relationship between the weight of a man and his body mass index (BMI). Make sure to give make a scatterplot, residual plot, histogram of the residuals, the null and alternative hypothesis, the t-test statistic, the P-value, whether or not you reject the null hypothesis and a conclusion. Write a sentence to explain the t-test statistic. Write a sentence to explain the P-value. Was the sample data significant? Explain why. Could the sample data have happened by random chance or is it unlikely? Explain why.







Correlation and Regression: Significance level = 0.01

Series: C1 Men Wt (Lbs),C2 Men Body Mass In...
x = C1 Men Wt (Lbs)
y = C2 Men Body Mass In...
Sample size n = 40
Degrees of freedom = 38

Correlation:
H_0: ρ = 0 (no linear correlation)
H_1: ρ ≠ 0 (linear correlation)

	Test Statistic	Critical Value
r	0.7997	±0.4026
t	8.2095	±2.7116

p-Value = $6.0619 \cdot 10^{-10}$

Regression:
Regression equation $Y = b_0 + b_1x$
$b_0 = 8.0169$
$b_1 = 0.1042$

Coefficient of determination $r^2 = 0.6395$
Standard Deviation of the Residual Errors = 2.0869
