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.

<u>Null and Alternative Hypothesis for the Goodness of Fit Test</u>

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 ≠* The population % is related to a categorical variable (% is related to the groups)

<u>Expected Counts and Observed Counts</u>

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.

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

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.

<u>Chi-Squared Test Statistic (χ^2) Sentence</u>: 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 ≠* (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

<u>Calculating Chi-Squared</u>

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.

<u>Chi-Squared Test Statistic</u> (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.

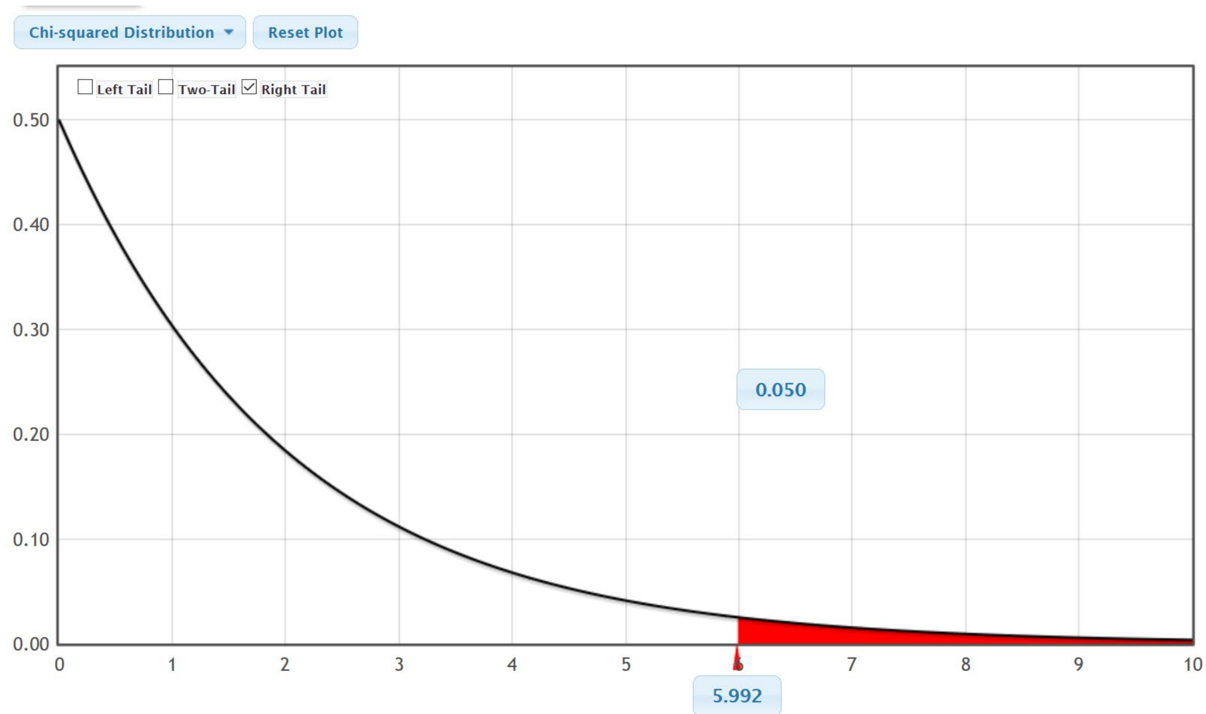
Note: Chi-squared Goodness of Fit test is always right tailed. Remember if the null hypothesis were true, then the observed and expected counts would be about the same. If that is the case then when you subtract them you should get about zero. So the center of the Chi-squared distribution should be close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

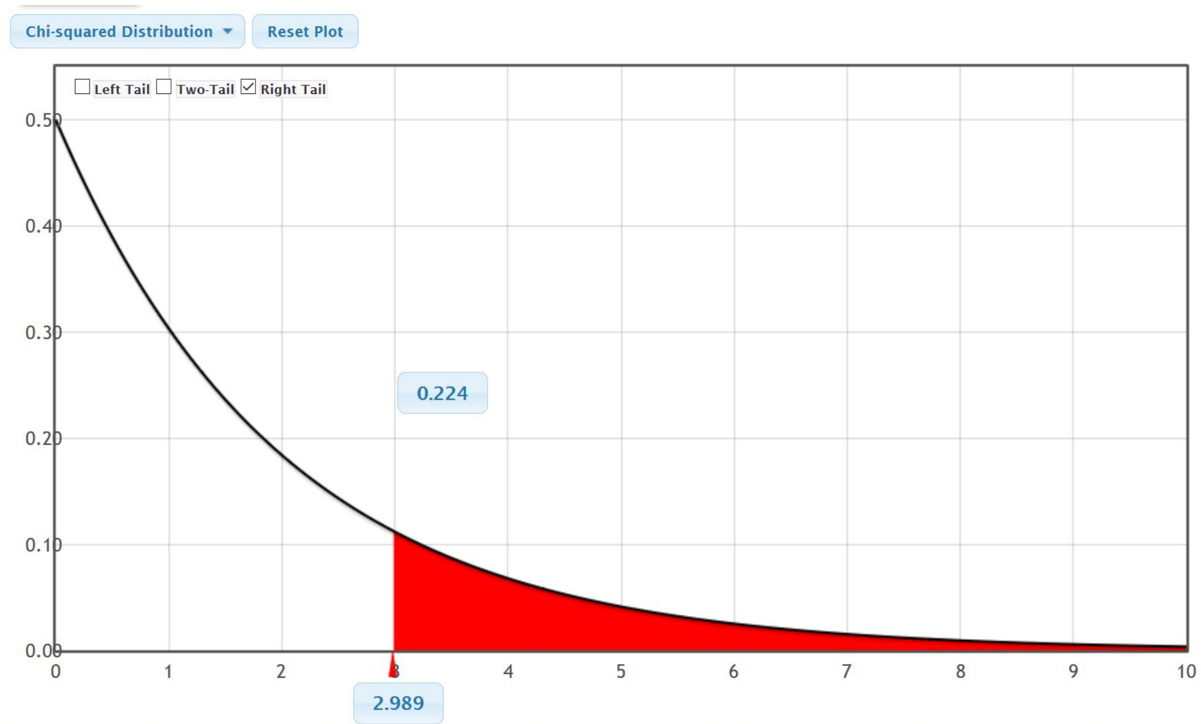
StatKey (Theoretical Chi-Squared)

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.)



↑
χ^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.

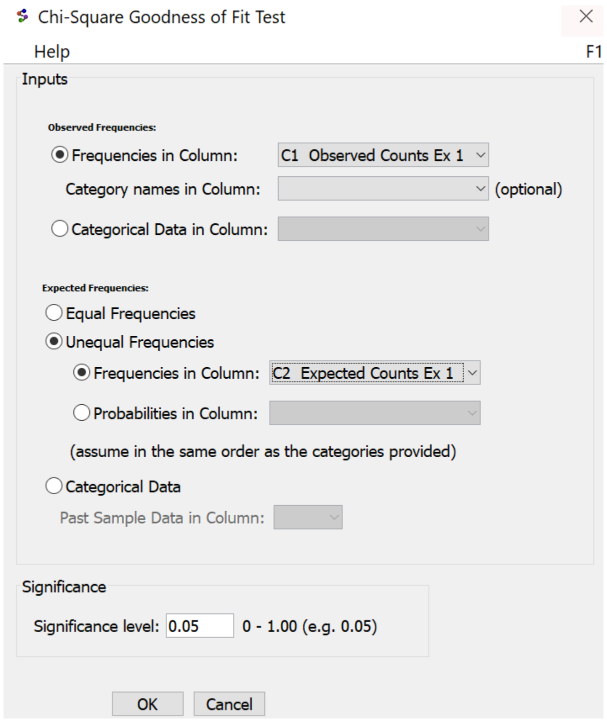
We could also have calculated the test statistic, critical value and P-value with Statcato.

<u>Statcato Directions for Goodness of Fit</u>

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

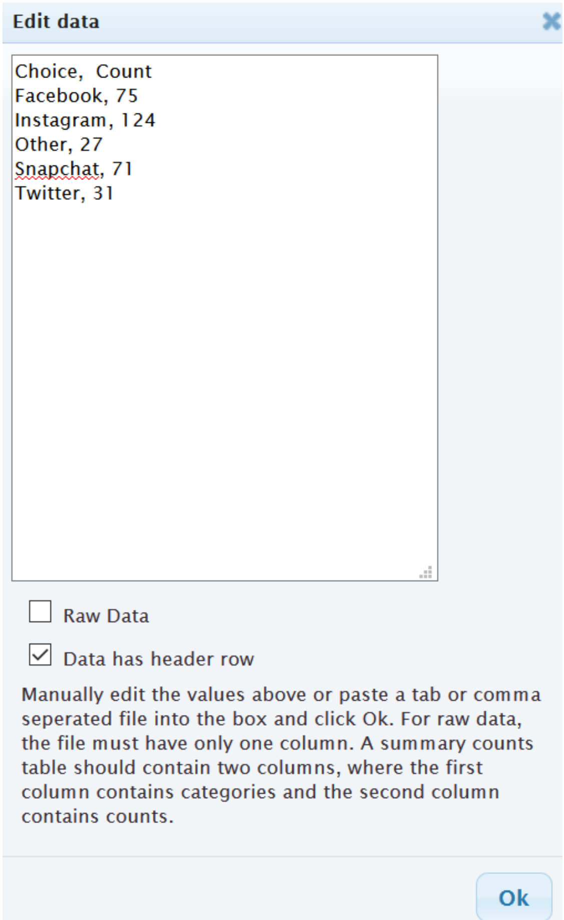
	Count
Facebook	75
Instagram	124
Other	27
Snapchat	71
Twitter	31

Using Randomized Simulation

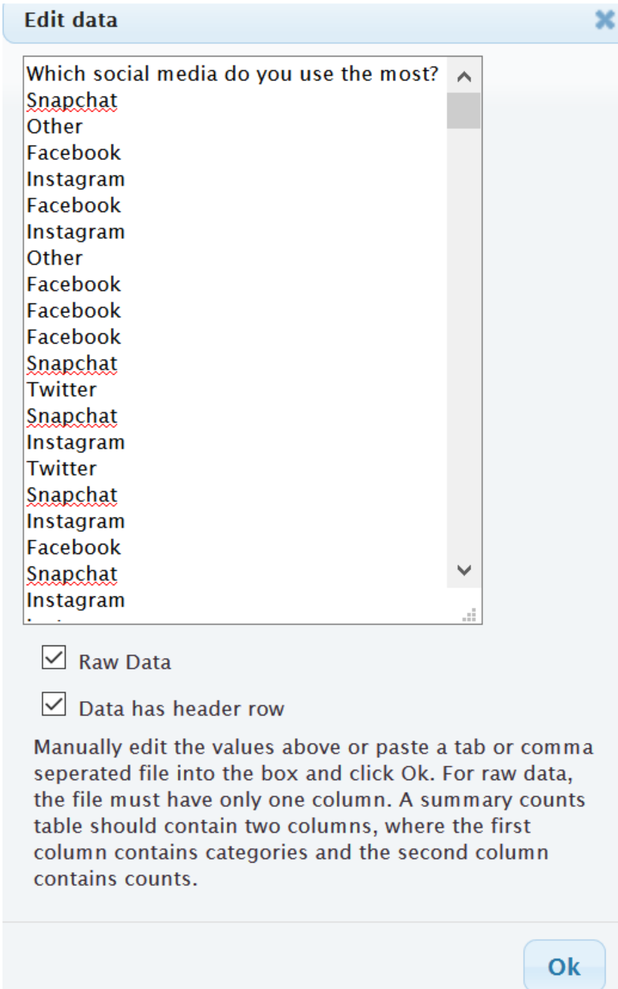
We can calculate the test statistic, critical value and P-value with a randomized simulation in StatKey. Like ANOVA, since there are three or more groups involved, we will not be able to put in the sample proportions directly into the simulation. Instead, the computer will use the Chi-squared test statistic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

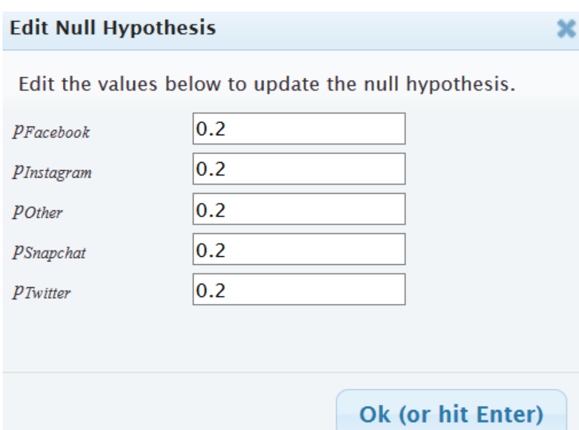
Choice, Count
Facebook, 75
Instagram, 124
Other, 27
Snapchat, 71
Twitter, 31


Edit data
Choice, Count
Facebook, 75
Instagram, 124
Other, 27
Snapchat, 71
Twitter, 31
Raw Data
Data has header row
Manually edit the values above or paste a tab or comma seperated file into the box and click Ok. For raw data, the file must have only one column. A summary counts table should contain two columns, where the first column contains categories and the second column contains counts.
Ok

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

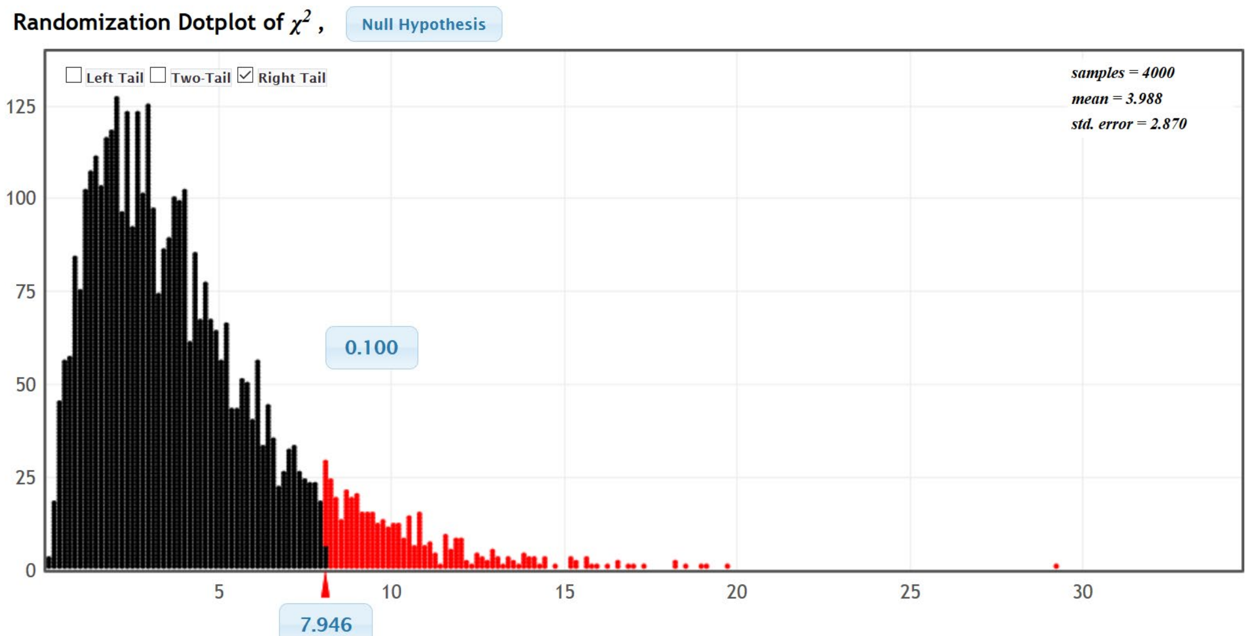
Checking the Assumptions

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.

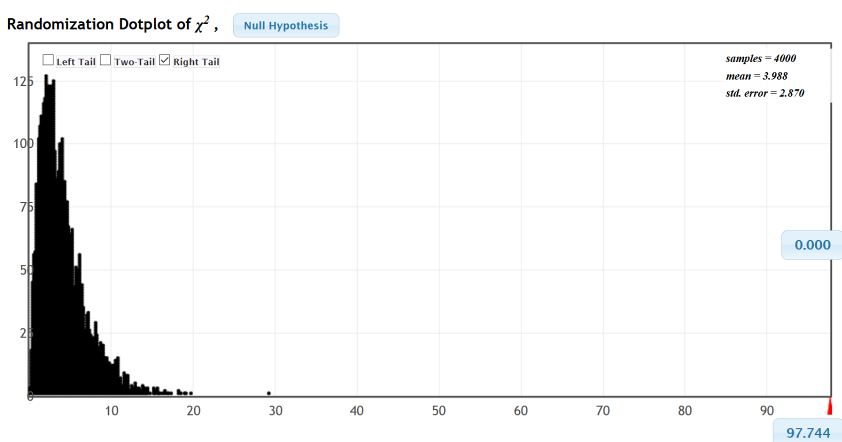
We can simulate the null hypothesis by clicking “Generate 1000 Samples” a few times. Notice the simulated distribution looks very skewed to the right. Remember the Goodness of Fit test is a right tailed test, so to calculate the Critical Value, click on “Right Tail” and put in the 10% significance level (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.



Original Sample Show Details

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

	Count
Facebook	75
Instagram	124
Other	27
Snapchat	71
Twitter	31

Randomization Sample Show Details

$n = 328, \chi^2 = 16.482$

	Count
Facebook	75
Instagram	90
Other	51
Snapchat	56
Twitter	56

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.

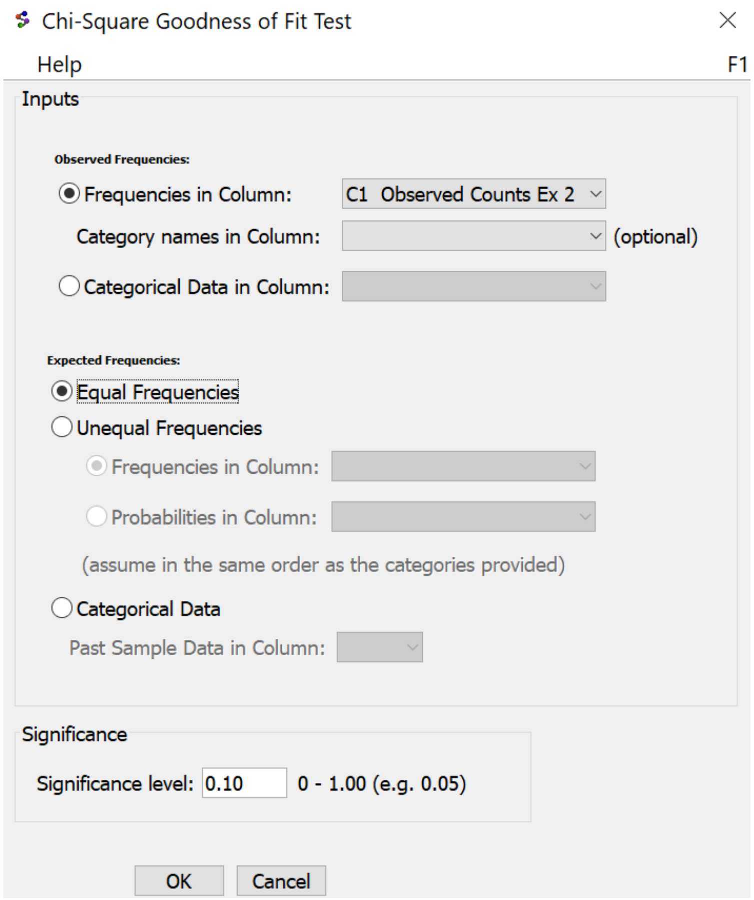
Since our P-value was low, we will reject the null hypothesis.

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

<u>Another Type of Goodness of Fit Test</u>

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.

Edit Null Hypothesis

Edit the values below to update the null hypothesis.

p_{White}	0.54
p_{Black}	0.18
$p_{Hispanic}$	0.12
p_{Asian}	0.15
p_{Other}	0.01

Ok (or hit Enter)

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

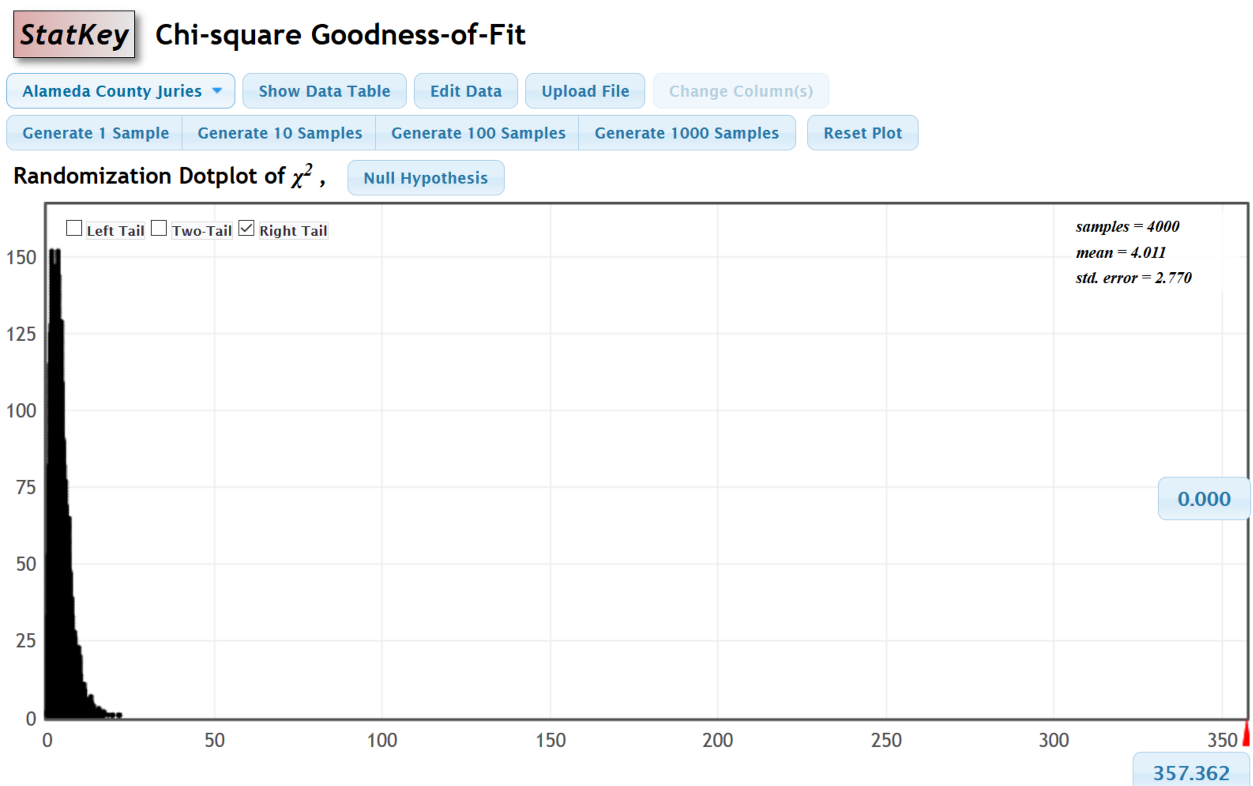
Show Details

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

	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.

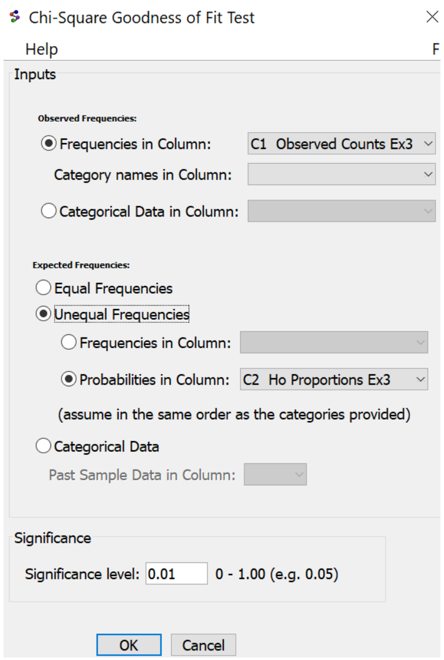


Hence we will reject the null hypothesis that the juries are in compliance and support the claim that the racial demographic of the juries in Alameda county were significantly out of compliance with the racial demographic of the county.

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.
