MATH 140 PROJECT Directions and Grading Rubric
Categorical Association Hypothesis Test / Summer 2023

Grading Rubric:

- **First and Last name (1 point)**
- **Project Title (2 points)**
- **Anti-cheating statement (2 points)**
- **Four StatKey graphs are worth 7 points each (28 total points)**
- **Sentences (#9,15,20,21) worth 4 points each (16 total points)**
- **All other questions are worth 3 points each (51 total points)**

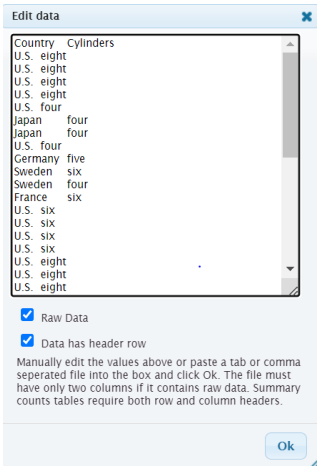
INDIVIDUAL PROJECT: Remember each student has been assigned different columns of categorical data to analyze, so do NOT work together with any other Math 140 students. This is an individual project. If your project has the same data as any other students project it will be considered cheating. The project is to perform a categorical association hypothesis test in order to see if the two categorical data sets assigned to you are related (associated) or not.

DATA and STATKEY GRAPHS: You will receive an email from Mr. Teachout assigning you two columns of categorical data from the Math 140 Survey Data Fall 2015. You will need to copy and paste the two columns assigned into another excel spreadsheet so that the two columns are next to each other. Then copy and paste the two columns together into StatKey.

How to create Graphs 1 & 2

- Go to www.lock5stat.com and open StatKey.
- Click on "χ^2 Test for Association" at the bottom middle of the StatKey page.
- Click "Edit Data" at the top of the χ^2 Test for Association page.
- Delete any data in the "Edit Data" screen.
- Copy and paste your two columns of data into the "Edit Data" screen.
- Check the boxes that say "Raw Data" and "Data has a Header Row" and then push "OK".
- Graph 1: You should not see the "Original Sample" contingency table with the χ^2 test statistic at the top right. It should look like the example below. Take a picture or screen shot of the contingency table. *(Note: Do NOT copy the table that says "Randomization Sample". That is the wrong one.)*
- Graph 2: Click the "Show Details" button next to "Original Sample". This will show the observed and expected counts that were used to calculate the χ^2 (Chi-square) test statistic. Take a picture or screen shot of the details table. It should look like the example below.

Examples below using Country and Type of Engine from the "Car Data".
(Notes: Your data will not be from the car data. Your data will be from the Math 140 Survey Fall 2015.
Also, you do NOT need to copy the "Edit Data" screen. It is shown below as an example.)



Example Graph #1 – Original Sample Contingency Table with Chi-square χ^2 test statistic

Original Sample Show Details

$n = 38, \chi^2 = 22.267$

	eight	four	five	six	Total
U.S.	8	7	0	7	22
Japan	0	6	0	1	7
Germany	0	4	1	0	5
Sweden	0	1	0	1	2
France	0	0	0	1	1
Italy	0	1	0	0	1
Total	8	19	1	10	38

Example Graph #2 – Detailed Sample Contingency Table with Observed Counts, Expected Counts and Contributions to Chi-square χ^2 test statistic

Detailed Sample Table

	eight	four	five	six	Total
U.S.	8 4.6 2.45	7 11 1.455	0 0.6 0.579	7 5.8 0.253	22
Japan	0 1.5 1.474	6 3.5 1.786	0 0.2 0.184	1 1.8 0.385	7
Germany	0 1.1 1.053	4 2.5 0.9	1 0.1 5.732	0 1.3 1.316	5
Sweden	0 0.4 0.421	1 1 0	0 0.1 0.053	1 0.5 0.426	2
France	0 0.2 0.211	0 0.5 0.5	0 0 0.026	1 0.3 2.063	1
Italy	0 0.2 0.211	1 0.5 0.5	0 0 0.026	0 0.3 0.263	1
Total	8	19	1	10	38

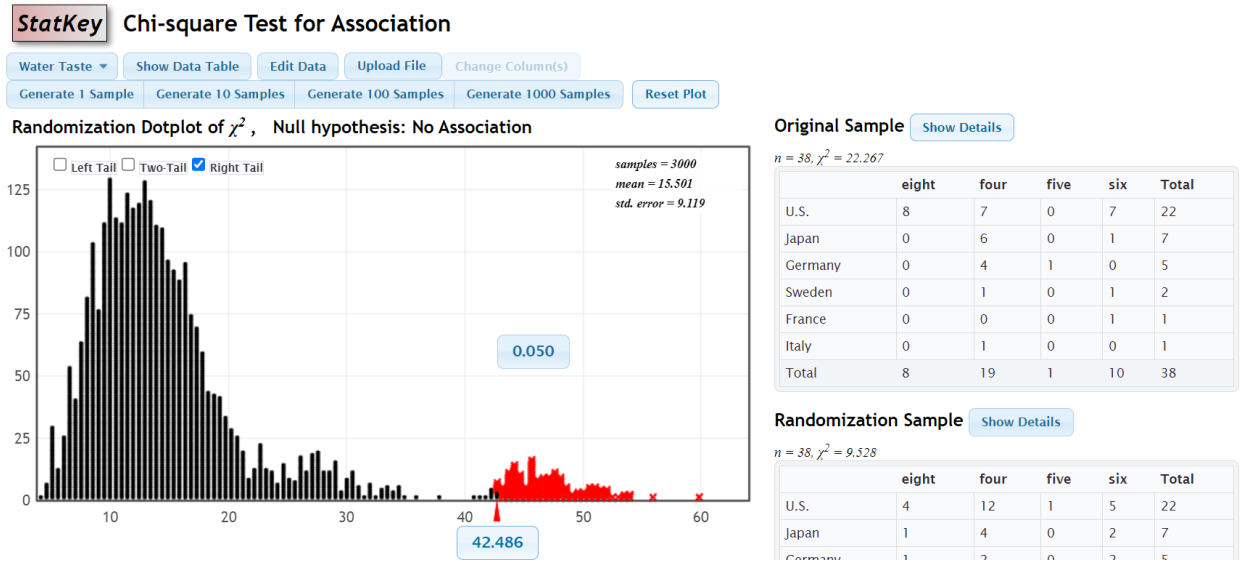
Observed, Expected, Contribution to χ^2

How to create Graphs 3 & 4

- Now that your data is in StatKey, you can create a randomization Chi-Square distribution. Click on "Generate 1000 Samples" a few times. You should now see the Chi-Square distribution.
- Graph 3: The Categorical Association Test is a right-tailed test, so click the box that says "Right Tail". Put "0.05" in the top box in the right tail (since we are using a 5% significance level). Take a picture or screen shot of this graph. The bottom box will be your critical χ^2 value! This graph shows you the tail. (See example below.)
- Graph 4: Now that we have the distribution created, we can calculate the P-value. Put your "Original Sample" χ^2 test statistic in the bottom box in the right tail. Once you do that, the top box will be the P-value! Take a picture or screen shot of this graph. (See example below.)

Examples below using Country and Type of Engine from the "Car Data".
(Notes: Your data will not be from the car data. Your data will be from the Math 140 Survey Fall 2015. Notice we put in 0.05 into the top box of the right tail in the first graph. Notice we put in the original χ^2 test statistic 22.267 into the bottom box in the second graph. You will not put in 22.267. You will put in whatever your original sample χ^2 test statistic was.

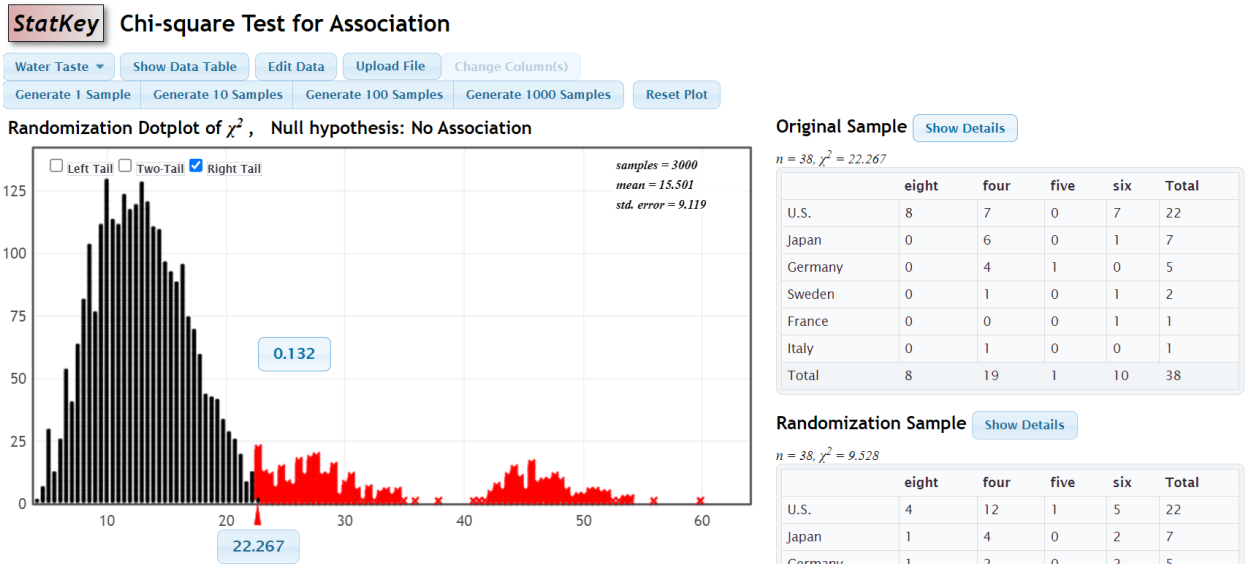
Example Graph #3 – Chi-square Test for Association distribution with 0.05 in the top box of right tail and the critical value in the bottom box of the right tail.



	eight	four	five	six	Total
U.S.	8	7	0	7	22
Japan	0	6	0	1	7
Germany	0	4	1	0	5
Sweden	0	1	0	1	2
France	0	0	0	1	1
Italy	0	1	0	0	1
Total	8	19	1	10	38

	eight	four	five	six	Total
U.S.	4	12	1	5	22
Japan	1	4	0	2	7

Example Graph #4 – Chi-square Test for Association distribution with Original Sample Chi-square χ^2 test statistic in the bottom box of right tail and the P-value in the top box of the right tail.



	eight	four	five	six	Total
U.S.	8	7	0	7	22
Japan	0	6	0	1	7
Germany	0	4	1	0	5
Sweden	0	1	0	1	2
France	0	0	0	1	1
Italy	0	1	0	0	1
Total	8	19	1	10	38

	eight	four	five	six	Total
U.S.	4	12	1	5	22
Japan	1	4	0	2	7

PROJECT REPORT

PART I: Name, Title, Anti-cheating statement

Put the following at the top of your report:

First and Last Name
Math 140 Section#???
Categorical Association Project
Summer 2023
Instructor: ???

Anti-cheating statement

I did not copy any else's project answers. I created the StatKey graphs and answered the questions myself. No one did this project for me.

PART II: StatKey Graphs

Put the pictures of the four required StatKey graphs. See examples above.

PART III: Data, Null and Alternative Hypothesis and Claim

Put the question number and the answers to these questions on the Project Report. You do NOT need to write the questions. You do NOT need to write essays or complete sentences.

1. **What are the titles of the two columns of categorical data sent to you by Mr. Teachout?**
2. **Write the null hypothesis (H_0). An example is given below.**
3. **Write the alternative hypothesis (H_A). An example is given below.**
4. **Is your claim H_0 (not related) or H_A (related)? Put the one you think is true.**

Example: H_0 *: The type of transmission in a car is NOT RELATED to the type of drive train in the car.)*

H_A *: The type of transmission in a car IS RELATED to the type of drive train in the car.)*

CLAIM: I think HA is true (The type of transmission in a car IS RELATED to the type of drive train in the car.)

PART IV: Check the following Chi-Square Categorical Association Test assumptions/conditions.

5. **Assumption #1: The data should be a random sample or represent the population. Put the following answer for question#5:** "The sample data passes the assumption of random or represents the population. The data was an unbiased census of the Math 140 students in Fall 2015 and represents the population of ALL Math 140 students in all semesters."
6. **Assumption #2: The individual people or objects in the data should not be related to each other. Put the following answer for question#6:** "We will assume the data passes the independent assumption. We will assume that the individual math 140 students are not related to each other."
7. **Assumption #3: All the expected counts must be 5 or higher. Does your data pass this assumption or fail this assumption?** List all of the expected counts in the "Detailed Sample Table". Each box gives 3 numbers. The expected count is the green number in the middle of each box in the table. (In the example table above from the "Car Data", the expected counts were 4.6, 11, 0.6, 5.8, 1.5, 3.5, 0.2, 1.8, 1.1, 2.5, 0.1, 1.3, 0.4 1, 0.1, 0.5, 0.2, 0.5, 0, 0.3, 0.2, 0.5, 0, 0.3). If all of your expected counts are 5 or higher, it passes the assumption. If any of your expected counts are below 5, it fails the assumption. **(**In the car example, some of the expected counts are below 5, so it fails the expected count assumption.)

PART V: Chi-square test statistic and significance questions.

8. **What is the χ^2 test statistic from your "Original Sample" table?** (This is listed in the Original Sample printout. For example, in the above example printout using the car data, the χ^2 -test statistic was 22.267)
9. **Write a sentence explaining the χ^2 -test statistic.** Here is an example: "The sum of the averages of the squares of the differences between the observed sample counts and the expected counts from the null hypothesis was 22.267".
10. **What is the critical value corresponding to 0.05 in the tail?** (In the example above, the Critical Value is 42.486, notice this is where the right tail begins.)
11. **Does the χ^2 test statistic listed on "Original Sample Statistics" fall in the right tail determined by the Critical Value or does the F-test statistic NOT fall in the right tail?** In the country and car engine example, the test statistic was 22.267, so this would NOT fall in the tail that starts at Critical Value of 42.486 and the test statistic 22.267 is smaller than 42.486. Would have to larger to fall in the tail.
12. **Does the sample data significantly disagree with the null hypothesis** (χ^2 -test statistic in tail) **or does the sample data not significantly disagree with the null hypothesis** (χ^2 -test statistic not in tail)**?**
13. **Are the observed sample counts significantly different than the expected counts** (χ^2 -test statistic in tail) **or are the observed sample counts NOT significantly different than the expected counts** (χ^2 -test statistic NOT in tail).

PART VI: P-Value and Sampling Variability Questions

14. **What is your P-value? Write it as a proportion and a percentage.** For example, in the simulation above the P-value is 0.132 or 13.2%.
15. **Write a sentence to explain your P-value.** For the example above: "If the null hypothesis is true and country a car is built in is not related to the type of engine, then there is a 13.2% probability of getting this sample data or more extreme by sampling variability."
16. **Is your P-value lower or higher than your 5% significance level?**
17. **Assume the null hypothesis was true. Could the sample data have occurred because of sampling variability** (High P-value) **OR is it unlikely for the sample data to have occurred because of sampling variability** (Low P-value)**?**
18. **If the sample data had met the assumptions, would your P-value be considered significant evidence** (Low P-value) **or not significant evidence** (High P-value)**?**

PART VII: Interpretations and Final Conclusion

19. **Considering the assumptions and the P-value, should we reject the null hypothesis OR fail to reject the null hypothesis?**
    - If P-value is low AND the data passed all assumptions – Reject Ho.
    - If the P-value was high AND the data passed all assumptions – Fail to reject Ho
    - If P-value is low AND the data did NOT pass all assumptions – Fail to reject Ho.
    - If the P-value was high AND the data did NOT pass all assumptions – Fail to reject Ho

20. **Write the standard conclusion sentence in context for your test addressing evidence, assumptions and the claim that your categorical variables are related or not related.** Here is an example sentence. You have to pick between "significant evidence" or "not significant evidence", "reject" or "support", "related" or "not related" depending on your P-value, passing or failing the assumptions, claim, and your null and alternative hypothesis.

    Example Conclusion: There (**is** or **is not**) significant evidence to (**reject** or **support**) the claim that the country a car is made in and the type of engine in the car are (**related** or **not related**).

    Remember use "reject" and "not related" in your sentence if Ho was your claim. Use "support" and "related" if Ha was your claim. If your P-value was higher than the 5% significance level or if the data did not pass ALL of the assumptions, then you should say there is "not significant evidence". Only if your P-value is lower than the significance level AND your data passed all the assumptions, then you can say there "is significant evidence".

21. **Does the data indicate that your categorical variables are related or NOT related?** Write the answer in context. Here are some examples though the variables will be different.

    - High P-value: Example: "Data indicates that the country a car is made in is not related to the type of engine in the car, however we do not have evidence."
    - Low P-value with Passing ALL Assumptions: Example: "Data indicates that the country a car is made in is related to the type of engine in the car. We have evidence.
    - Low P-value but One or More Assumptions did NOT pass: Example: "Data indicates that the country a car is made in may be related to the type of engine in the car, however we do not have evidence due to failed assumptions."