Chi Square in SAS® Studio

In this lesson we are going to utilize some of the SAS Studio tools to conduct Chi-squared test for categorical variables with two or more categories in each. It is used mostly to compare percentages and we can use this test under the assumptions that we have independent observations.

Dataset Description

For this example, you will be revisiting the PEW2020 data. About the PEW2020 data that you will be using:

The PEW2020 data comes from the September 2020 Pew survey in which telephone interviews were "conducted Sept. 22-28, 2020, among a national sample of 1,007 adults, 18 years of age or older, living in the United States (301 respondents were interviewed on a landline telephone, and 706 were interviewed on a mobile phone, including 487 who had no landline telephone). A combination of landline and mobile phone random-digit-dial samples were used. Interviews were conducted in English (972) and Spanish (35). The combined landline and mobile phone sample is weighted to provide nationally representative estimates of the adult population 18 years of age and older."

The questions on this survey are as follows:

1. Which country currently is the most important partner for American foreign policy?
2. In general, how would you describe relations today between the United States and Germany? Would you say they are very good, somewhat good, somewhat bad or very bad?
3. Which is more important for the United States?
    a. Having a close relationship to Germany or having a close relationship to Russia?
    b. Having a close relationship to Germany or having a close relationship to China?
4. How would you rate the likelihood of the current rivalry between China and the United States escalating into a confrontation resembling the Cold War?
5. For each of the following issues, do you see Germany as a partner or not?
    a. Protecting the environment
    b. Dealing with China
    c. Dealing with Iran
    d. Promoting free trade
    e. Protecting European security
    f. Protecting democracy and human rights around the world
6. Which of these statements comes closer to your view, even if neither is exactly right? Once the coronavirus crisis is over, do you think ...?

The demographic data collected includes:

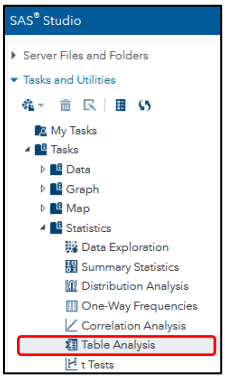
- How many of these are adults, 18 or older?
- What is your age?
- Case ID
- 'Country'
- Date
- What is the highest level of school you have completed or degree you have received?
- Are you of Hispanic or Latino origin or descent?
- Is your total annual household income from all sources, and before taxes ...?
- 'Marital status'
- Are you the parent or guardian of anyone under 18 in your household?
- As of today, do you lean more to the Republican Party or more to the Democratic Party?
- Generally speaking, would you describe your political views as ...?
- Race of Respondent
- What is your present religion, if any? Are you Protestant, Roman Catholic, Mormon, Orthodox such as Greek or Russian Orthodox, Jewish, Muslim, Buddhist, Hindu, atheist, agnostic, something else, or nothing in particular?
- 'Sex of respondent'
- State
- 'Survey'
- Including yourself, how many people are there living in your household?
- Weight

I. Select the Data Source and Performing a Chi-Square Analysis

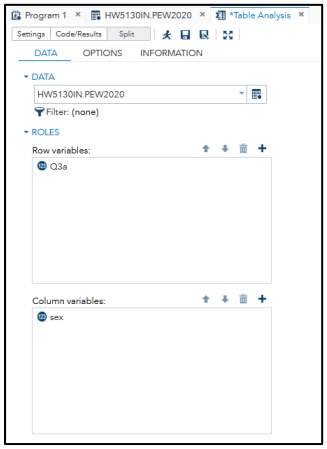
For this portion of the exercise, the question being tested is "Are men and women equally likely to prioritize US relationship to Germany over Russia?". The results from question 3 in the Pew survey and the sex variable will be used to investigate this question.

First, let's look at the counts for these two variables.

1. Download the pew2020.sas7bdat file from the Class Handouts for Lab 5 Assignment page.
2. Go to Tasks and Utilities, expand Tasks, then expand Statistics, and select Table Analysis.



3. In the Table Analysis window:
   - under Data, check that the Pew2020 data set is selected
   - select Q3a as the row variable
   - select sex as the column variable



4. To add your first and Last Name in the footer, select Code, then Edit.

5. Go to line 20 before the Run statement, hit enter and then type the following code on lines 20 and 21:

```
FOOTNOTE;
FOOTNOTE1 "First and Last Name";
```

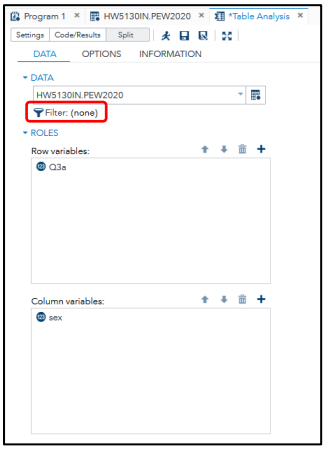
6. Click Run. You should obtain the following output:

Table of Q3a by sex

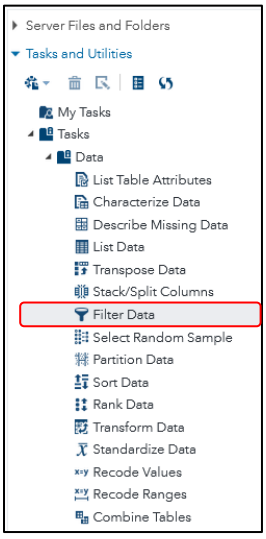
Q3a(Q3a. Which is more important for the United States? Having a close relationship to Germany or having a close relationship to Russia?)	sex('Sex of respondent')		
	Male	Female	Total
Having a close relationship to Germany	327	324	651
Having a close relationship to Russia	128	121	249
Both relationships are equally important	44	33	77
VOL: Neither	4	9	13
DK/Refused	6	11	17
Total	509	498	1007

First and Last Name

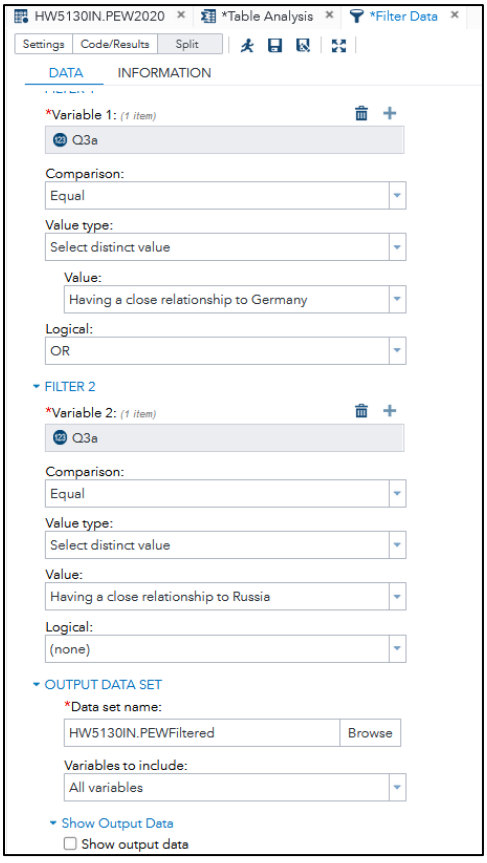
7. For this example, filter the data to only include 1. Having a close relationship to Germany and 2. Having a close relationship to Russia and exclude 3. Both relationships are equally important, 4. VOL: Neither, and 9. Don't Know / Refused. To do this, click filter.



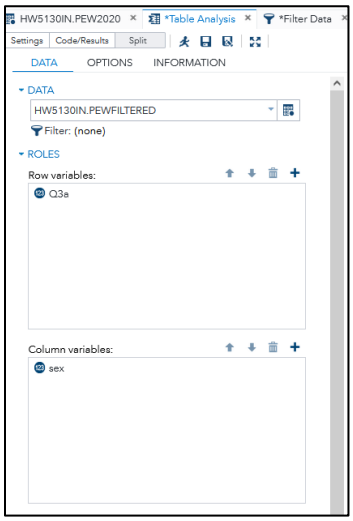
8. Under Tasks and Utilities, expand Tasks, expand Data, then select Filter Data.



9. In the filter window, select Q3a as Variable 1, set comparison to Equal, select value type to distinct value, select “Having a close relationship to Germany”, then for logical select or. Repeat these selections for variable 2 to be “Having a close Relationship to Russia”. Under Output, select the dataset name as PewFiltered. Click Run.



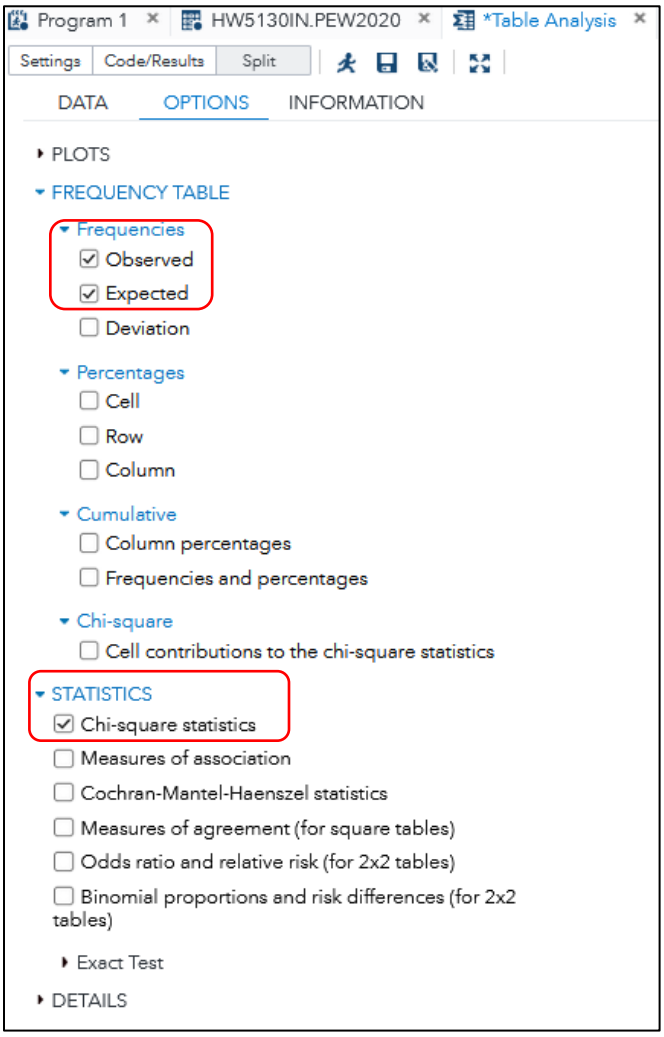
10. Now, let's perform the analysis. Go back In the Table Analysis window:
    - under Data, check that the PewFiltered data set is selected
    - select Q3a as the row variable
    - select sex as the column variable



11. Click Run. You should now have the following output.

Table of Q3a by sex			
Q3a(Q3a. Which is more important for the United States? Having a close relationship to Germany or having a close relationship to Russia?)	sex('Sex of respondent')		
	Male	Female	Total
Having a close relationship to Germany	327	324	651
Having a close relationship to Russia	128	121	249
Total	455	445	900

12. Now, go to the Options tab, and select Observed and Expected under Frequencies. Select Chi-square statistics under Statistics.



13. Click Run to obtain the following table from the output.

Table of Q3a by sex

Q3a(Q3a. Which is more important for the United States? Having a close relationship to Germany or having a close relationship to Russia?)	sex('Sex of respondent')		
	Male	Female	Total
Having a close relationship to Germany	327 329.12	324 321.88	651
Having a close relationship to Russia	128 125.88	121 123.12	249
Total	455	445	900

14. Scroll down in the output to find the following table.

Statistics for Table of Q3a by sex

Statistic	DF	Value	Prob
Chi-Square	1	0.0995	0.7524
Likelihood Ratio Chi-Square	1	0.0995	0.7524
Continuity Adj. Chi-Square	1	0.0581	0.8096
Mantel-Haenszel Chi-Square	1	0.0994	0.7525
Phi Coefficient		-0.0105	
Contingency Coefficient		0.0105	
Cramer's V		-0.0105	

The hypothesis statements for this problem is:
H_0: There is not a relationship between gender and the opinion for the preferred country that they believe the US should prioritize between Germany and Russia.
H_1: There is a relationship between gender and the opinion for the preferred country that they believe the US should prioritize between Germany and Russia.

Next, we need to look up the critical value for 1 degree of freedom at the 0.05 significance level. This critical value is 3.841. Therefore, the decision rule for this scenario is to reject H_0 if $\chi^2 > 3.841$. Since our calculated χ^2 value is less than our critical value, we fail to reject the null hypothesis and conclude that there is insufficient evidence to support that there is a relationship between gender and opinion for prioritizing Russia or Germany.

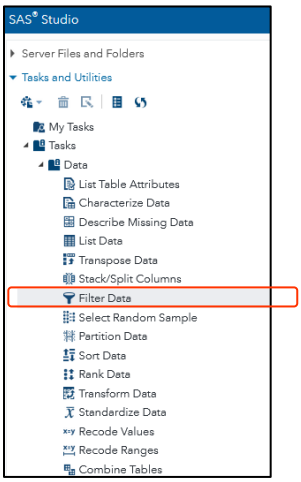
In layman's term, with an insignificant result it implies that the probabilities between gender and their opinion is therefore an independent event. This would then allow you to conclude that gender does not impact the relationship tied to the opinion. In the event that instead the Chi Square test was positive, we would infer that there is some at least one of the cells with the Chi Square frequency table that is outside the expected distribution of values. It is important to realize that as the number of categories (2×3, 2×4, 3×3, etc.) increase the Chi Square test will not identify which relationships are considered different enough from their expected probabilities on their own and so additional *post hoc* testing is required to isolate which combinations have a significant relationship.

Farther down in the output, you will see the Phi Coefficient, φ, which is a Chi-Square based measure of association. The Chi-Square coefficient depends on both the sample size and the strength of the relationship. The Phi Coefficient is independent of the sample size. When you have a 2×2 analysis, such as the one in this example, φ can be interpreted as the symmetric percent difference which measures the percent of concentration of cases on the diagonal. Additionally, in a 2×2 analysis it is identical to the correlation coefficient. The Phi Coefficient is used in the calculation of Cramers V, is also used to describe the association between two nominal variables, but the values are constrained to being between 0 (no association) to 1 (perfect association). The formula used by SAS to calculate Cramer's V also allows for the retention of the sign for φ to retain the direction of the association. In this example, the Cramer's V value is -0.0105, which indicates that while the groups may be going in different directions (i.e., making different selections), there is little to no association between the behavior of these two groups.

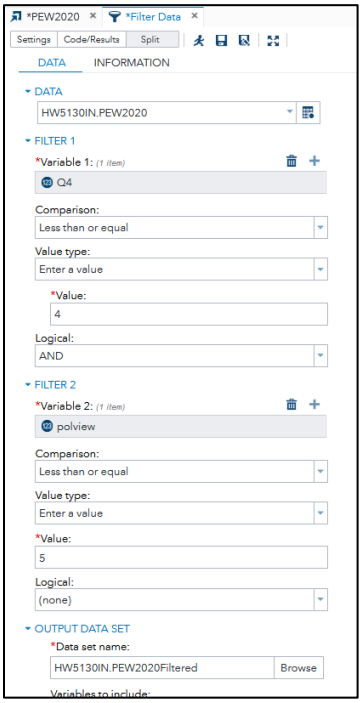
II. **Prepare the Data for the Homework**

For this example, we are interested in the research question “Are males more liberal than females?”. However, there are several values that need to be recoded first.

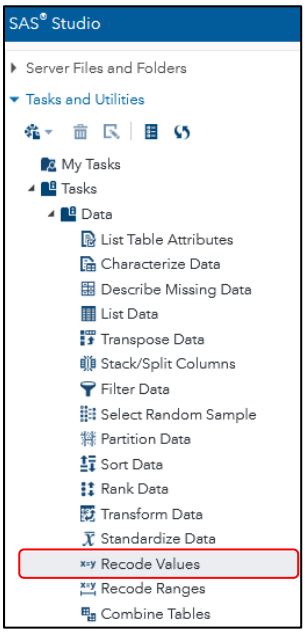
1. Select the Pew2020 Data that is in your homework library.
2. We will begin by filtering out the cases where Q4 takes the value of “DK/Refused” and where polyview variable takes the value of “Don’t Know” or “Refused”. Go to Tasks and Utilities, expand tasks, then Expand Data, and select Filter Data.



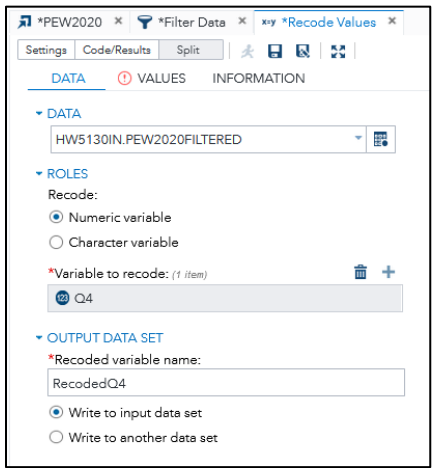
3. In the filter data window, set variable 1 to Q4 less than or equal to 4 and variable 2 as less than or equal to 5. Create a new data set called Pew2020Filtered. This will create a data set that excludes cases where Q4 takes the value of “DK/Refused” and where polyview variable takes the value of “Don’t Know” or “Refused”. Click Run.



4. Now, you need to recode some of your variables. Under Tasks and Utilities, Expand Tasks, expand data, and select Recode Values.



5. On the Data portion of the Recode Values window:
    - Make sure that you have the Pew2020Filtered file selected.
    - Under Roles, select “Numerical variable” and set Q4 as the variable to recode.
    - Under Output Data Set, set the recoded variable name to RecodedQ4 and select write to input data set.



6. On the values tab, use the following to prepare the old and new values.

Q4	Variable label	SAS entry	Suggested Recoded labels (Values)
	Very likely	1	Likely (10)
	Somewhat likely	2	
	Somewhat unlikely	3	Unlikely (11)
	Very Unlikely	4	
	DK/Refused	9	Filter out

7. Click Run. Open the Pew2020Filtered data set to see that the variable RecodedQ4 is in your list of variables.
8. Open a new Recode Values window to recode the values for polview to make a new variable names RecodedPolview in the input data set as follows:

polview	Variable label	SAS entry	Suggested Recoded labels (Values)
	Very conservative	1	Conservative (20)
	Somewhat conservative	2	
	Moderate	3	Moderate (21)
	Somewhat liberal	4	Liberal (22)
	Very liberal	5	
	Don't know	8	Filter out (9)
	Refused	9	

9. Click Run. You should now see both the RecodedQ4 variable and the recodedpolview variable in this data set.