**Analyzing Class Survey Data in Excel**

1. Choose to analyze the relationship between two of the variables included.  One or both of the variables you choose needs to be a numerical value (for instance, you might choose gender and univ GPA but not gender and high school).
2. Think about what you would expect to find and why.
3. Follow track A below if one of your variables is ***not*** numerical, but instead a category.  Follow track B below if both of your values are numerical (i.e., univ GPA and church attendance).

*Track A:* I will use gender and college GPA as an example in the description below.  Depending of what variables you choose to analyze, this might be central city/suburban/small city/rural instead of male/female and it might be political\_quant instead of GPA (or whatever variables you choose). Follow this tutorial (<http://www.kscience.co.uk/as/module5/ttest.htm>) to see details of how to do these things. It shows exactly what I’m looking for, but the data are organized a little differently so step 1 below is to reorganize the data into how tutorial shows it. In the tutorial, the fertilizer A would be like the females in the example I give and fertilizer B would be like males.

1. **Copy and paste the data of your interest into a new worksheet to match the format shown in the tutorial example.** In the new spreadsheet, you should create columns that match the categories you are interested in (male/female in this example). Then, you’ll need to copy and paste the numerical data for the variable you are interested in (GPA in this example) into the columns you’ve created. To do this, sort the categorical variable so that you have all of the females together and all males together. Then highlight the numerical data (GPA) for the females. Copy. Go to the new spreadsheet and click in the top cell for data in the column for females (or whatever category) you created, then paste the data. Repeat for each of the categories until you create a dataset that looks like the data in the tutorial.
2. **Calculate the average value of the numerical variable** (GPA in this example) **for each category** (the non-numerical variable or gender in this example).  This will tell, for instance, on average the males in this class have X GPA and the females in this class have X GPA. We can start to see if males/females in the class have higher GPA.
3. **Calculate the standard deviation of the average.** The standard deviation tells us how much variation there is in those averages. So, if the standard deviation is large it suggests that the range of GPAs within the categories (male/female) is great. If the standard deviation is relatively small, then it suggests that all males in class have similar GPAs to one another and that all females in class have similar GPAs to one another.
4. **Create a bar chart showing the relationship between these two variables.** For instance, if you look at relationship between gender and GPA, you would create a chart with two bars- one for male and one for female and the height of each bar would represent the average GPA by gender. You can skip adding the error bars (shown in tutorial) if you want to.
5. **Calculate a t-test to determine if the difference between the means is statistically significant or not.**  If the t-test value is <0.05 (or 5%) that indicates that the difference between the categories is statistically significant (or looks to be a sound relationship). If the t-test value is >0.05 (5%) then the differences between the categories is no big deal, or better put it is within a range of normal variation and you can’t draw a conclusion that the categories have a true relationship with the numerical variable.
6. **If you find there to be a relationship, think about whether this relationship is real or if maybe there is another variable out there that you didn’t consider that might be related to both of these variables you tested that could be causing you to make a spurious conclusion.**
7. **Write up results in a blog post.** Start by describing what variables you analyzed and why. What is the question you started with and what did you expect to find. Note what the average values you found were and their standard deviations. Show the chart you created. Explain what you find in this chart and why it appears to be interesting or not. Include what you found with the ttest and explain whether or not you think there is a significant relationship between the categories you examine and the numerical data. Discuss the potential for a spurious relationship (if there is another variable out there that is the real cause of what you are finding) or if you think that you’ve uncovered something that looks to be a causal relationship between the categories and the numerical data. Finally, explain why/how the results of your study make sense or not. For instance, if you find females have higher GPA than males, reflect on why this might be the case. What is it about society (and/or more specifically Michigan Tech or our class) that would lead to a situation where females have higher GPA than males?

*Track B:* I will use GPA and political\_quant as an example in the description below.  You could choose any two of the numerical data to analyze. The YouTube video shows how to follow the steps listed below.

<http://www.youtube.com/watch?v=s2TVkYmmCAs>

1. **Generate a scatterplot picturing the relationship between the two variables of interest.** If the data points look randomly scattered around, then they are likely not correlated or there is not a relationship between the two.  But, if it looks like the points generally move up to the right or down to right together in a diagonal like pattern, then that suggests there is a relationship between the two. See <http://www.dummies.com/how-to/content/how-to-interpret-a-scatterplot.html> for more information on how to interpret a scatterplot.
2. Calculate the correlation coefficient between the two variables. The result could range from zero (no relationship) to +/-1.0 (perfect correlation). Positive correlations mean that as one variable increases the other also increases. Negative correlations mean that as one variable increases the other decrease. Values with an absolute value <0.2 or so suggest that there is not a correlation in either direction. Values with an absolute value>0.4 or so suggest a strong correlation. Check out <http://www.dummies.com/how-to/content/how-to-interpret-a-correlation-coefficient-r.html> for more information on how to interpret these.
3. **If you find there to be a relationship, think about whether this relationship is real or if maybe there is another variable out there that you didn’t consider that might be related to both of these variables you tested that could be causing you to make a spurious conclusion.**
4. **Write up results in a blog post.** Start by describing what variables you analyzed and why. What is the question you started with and what did you expect to find. Show the scatterplot chart you created and explain what you think it shows. Include what you found with the correlation coefficient and explain whether or not you think there is a relationship between the variables you examined for our class. Explain how and why this makes sense or not. Discuss the potential for a spurious relationship (if there is another variable out there that is the real cause of what you are finding) or if you think that you’ve uncovered something that looks to be a causal relationship between the categories and the numerical data. Finally, explain why/how the results of your study make sense or not. For instance, if you find more conservative political orientation to be related to higher GPA, reflect on why this might be the case. What is it about society (and/or more specifically Michigan Tech or our class) that would lead to a situation where conservatism looks like increase GPA?