

# Fifth Data Assignment

This assignment will be due to be turned in **on paper** at the Final Exam. Please **print your document** and turn in at the **beginning of the Final Exam session**. Late papers will be accepted (subject to the penalties described on the syllabus) until noon on Monday, December 20.

## Part I. Methodology

From the Big Survey choose a **three pairs of questions** that you are interested in testing for association. Your choice should include “categorical vs. categorical,” “categorical vs. quantitative,” and “quantitative vs. quantitative.” Note that this means you will have selected 4 to 6 questions total.

Open the Big Survey spreadsheet and delete all columns except for those that correspond to the questions that you chose in above. **Be sure you are using a current version of the spreadsheet from CourseCompass so you have more than 150 responses.**

Make a note of data that needs to be fixed. Don't forget to be on the lookout for erroneous (i.e., impossible) quantitative values – these should be deleted the same way that nonsensical categorical answers are deleted.

Give credit to anyone you worked with, even if it was just help with the Excel details. Since we are picking many questions for this assignment, if you work with another person you should not test the exact same set of three associations.

## Part II. Perform a $\chi^2$ test.

Create a contingency table for the responses. Remember that this test is not great when any cell of the table has fewer than five responses. This may lead you to combine categories (as I did in my sample) or to just pick a different question altogether. To create the table, you should use sorting as we did in Data Assignment 2 to rearrange the rows so that is super easy to count. From the contingency table, you can copy the counts into matrix A of your calculator and run a  $\chi^2$  test at the 5% significance level. Give the test statistic and P-value, and then give the conclusion as a non-technical English sentence.

## Part III. Perform an ANOVA test

For ANOVA you will need the Analysis ToolPak – if you don't see it in Excel 2007 for Windows (i.e., the machines on campus PCs) then follow the instructions here: <http://cortlandttc.wordpress.com/2007/11/27/activating-the-anova-data-analysis-function-in-excel-2007/>

Test for association between a categorical (explanatory)

	A	B	C	D	E	F
1	Political	Paper		Conservative	Moderate	Liberal
2	1	1		1	1	
3	1	2		2	0	
4	2	1.5		1.5	2	
5	2	7		7	0	
6	2	4		4	0	
7	2	1		1	0.5	
8	2	0		0	0	
9	2	5		5	0	
10	2	1		1	1	
11	2	0		0	0	
12	2	3		3	2	
13	3	1		1	0	
14	3	2		2	0	
15	3	1		1	4	
16	3	1		1	1	
17	3	1		1	1	
18	4	1			0	
19	4	0			0	
20	4	2			2.5	
21	4	0			2	
22	4	0			0	
23	4	0.5			1	
24	4	0			0	
25	4	0			0	
26	4	1			1	
27	4	0			1	
28	4	2			1	
29	4	0			0	
30	4	0			0	
31	4	4			1.5	
32	4	1				
33	4	1				
34	4	0				
35	4	0				
36	4	2.5				
37	4	2				
38	4	0				
39	4	1				
40	4	0				
41	4	0				
42	4	1				
43	4	1				
44	4	1				
45	4	0				
46	4	0				
47	4	1.5				
48	4	0				

variable and a quantitative (response) variable by performing an ANOVA test for at the 5% significance level. The first step for this should be sorting your data so you can create a new Excel column consisting of the quantitative answers for each particular category. For example, if the two questions' data are in columns A and B, then you should select all of the data, choose "Custom Sort" under the Data tab, and then sort the data first on Column A. You will then write descriptions of the categories as new column headings and copy the numerical data for each category to lie under the appropriate heading. The screen shot above shows what this looks like on my Mac machine as I have already copied the numbers that go with "Conservative" and "Moderate" to the columns that I created and labeled "Conservative" and "Moderate."

Now the hypothesis test can be performed using ANOVA Single Factor available with the Data Analysis tool under the Data tab. The following document gives decent instructions for doing this:

[http://biology.uco.edu/PersonalPages/CButler/BIO1404/handouts/Excel\\_2007\\_ANOVA\\_directions.docx](http://biology.uco.edu/PersonalPages/CButler/BIO1404/handouts/Excel_2007_ANOVA_directions.docx)

Note that the output numbers are pretty badly formatted so you might decide to just copy them into the sample file to avoid having to reformat them in Word.

## Part IV. Regression Line

To create a scatterplot, choose the two columns of numbers, choose the Insert tab, and then Scatter. You will get a basic scatterplot. Right-click on any data point, and choose Add Trend Line with settings to give the equation and  $R^2$  value on the graph itself. You might have to move the textbox with equation/ $R^2$  so that it's readable on the chart.

Make the extra effort to label axes and make a title for the chart. See <http://www.brighthub.com/computing/windows-platform/articles/17778.aspx> for instructions for all the steps to making a nicely formatted scatterplot.

Make a conclusion about the strength of the correlation from the  $R^2$  value.

## Grading Rubric (20 points total)

### Methodology (3 points)

#### Chi-Squared Test (6 points)

- Contingency table
- Statement of hypotheses
- Test statistic and P-value
- Conclusion in plain words

#### ANOVA Test (6 points)

- Table of rearranged data
- Statement of hypotheses
- ANOVA output from Excel
- Conclusion in plain words

#### Regression (5 points)

- Well-labeled scatterplot with trendline, equation, and  $R^2$  value
- Statement about strength of correlation (association) in plain words