

The American Community Survey (<http://www.census.gov/acs/>) is a survey conducted by the US Census Bureau to gather information about communities in the US. A subset of this data have been extracted to a file on Oncourse in the “Resources” section and also on the course webpage¹.

The goal of this project is to introduce you to using a spreadsheet to process larger amounts of data.

The Project

Answer the following questions using a spreadsheet application. Tutorials on using Excel² and OpenOffice³ are available on Oncourse in the “Resources” section and also on the course webpage. A more detailed description of the data in each column is given at the end of this document.

Turn in:

1. A typed document answering each question (except question 1) thoroughly and completely. This project is the only project for which you may answer each question independently⁴.
2. Your spreadsheet file via email. **Do not print out your spreadsheet!**

Even though you are submitting your spreadsheet, all relevant content (graphs, computed values, ...) should be included in your report. Any relevant content which appears only in your spreadsheet will only count for half credit. Solutions must be described completely and in full paragraphs with complete ideas. Be sure that your solutions include enough information that the reader is able to follow your thought process and is able to determine the point of each question and response.

1. Find the mean, standard deviation, and five number summary for all the original (numerical) columns of data. It is enough to compute these and leave the values in your spreadsheet.
2. Find the number of people over 65, number of people below the poverty level, and number of men and the number of women for each county. Include an explanation of the computation required to compute the number of men and the number of women in each county.
3. Find the total number of people over 65, number of people below the poverty level, and number of men and women for all the counties. Find also the total number of people in the United States (according to this data). Use these to get the overall percentage of each.
4. Get histograms and boxplots for at least two columns (do not choose men/women never married, columns O and P). Describe the histograms and compare them to their boxplots.

If a variable has outliers a modified boxplot can be drawn which uses the LOW or HIGH value from the 1.5 IQR rule on the side where the outliers lie⁵. Plot both the boxplot and this boxplot with outliers removed on the same graph. Does the removal of outliers affect skewness of the graph?

¹<http://dean.serenevy.net/teaching/classes/Fall2007/M111-1/Project1Data.csv>

²Office Enterprise Edition 2007 is available as a free 500MB download through the IUware website (<http://iuware.iu.edu/>) or is available for \$27 from the bookstore.

³Available as a free 120MB download from the OpenOffice website (<http://www.openoffice.org/>).

⁴You are welcome to merge questions if you find that easier (especially problems 2 and 3).

⁵For example, if min = 3, max = 17, LOW = -2, and HIGH = 13 draw the boxplot using 3 for the low tail but 13 for the high tail.

Include the histograms and their corresponding boxplots in your solution. Use the values computed in problem 1 to draw the 1SD, 2SD, and 3SD lines on the histograms (by hand or using “Layout” → “Shapes” → “Line” in Excel). Describe each graph, its spread, skewness, and compare it to its boxplot. Which, if any, histograms seem to follow the 68/95/99.7 rule (and thus approximate normal curves)? Be sure to identify your graphs properly by choosing descriptive titles.

5. How do the percent of men and women who have never married compare? Get histograms and/or boxplots for each. You should talk about the shape, center, and spread of the two groups.
6. Find two different scatterplots for two pairs of variables you think might be related. Explain why you think they might be related, how strongly you expect them to be related, and whether you expect a positive or negative association. For each graph, describe its shape (is it football-shaped or not) and compare the correlation coefficient to the graph.

You may also find it interesting, if you wish, to compare each scatterplot to the scatterplot of the same variables with the axes reversed. Comparing the outliers shown in the boxplot to the (visually determined) outliers in the scatterplot may also be interesting.

Column Descriptions

- A: the county name
- B: the state
- C: the FIPS state code (Federal Information Processing Standard)
- D: the percentage of people who do not speak English at home
- E: the median age in the county
- F: the mean size of a household
- G: the median household income
- H: the mean time it takes a working adult to travel to work
- I: the percent of adults who have finished high school
- J: the percent of adults who have a bachelor’s degree
- K: the percent of households where the head(s) of the household is married
- L: the percent of people below the federal poverty level
- M: the sex ratio — the number of males per 100 females
- N: the percent of people 65 or over
- O: the percent of adult males who have never married
- P: the percent of adult females who have never married
- Q: the age dependency ratio ($100 \times (\# \text{ of people under } 18 \text{ or } 65 \text{ and over}) / (\# \text{ of people } 18\text{--}65)$)
- R: the percent of people in the county who were born in that state
- S: the percent of adult workers that use public transport to commute to work
- T: the population of the county

Extra Credit

Due 16 Oct — will not be accepted late

Find out how the American Community Survey (ACS) is conducted. Write a paper explaining in detail how the sample is chosen. Compare methodology of the ACS to that of the Gallup Poll.