

CE311S Lab 0

Data Analysis with Microsoft Excel

Purpose

The purpose of this lab is to acquaint yourself with some of the statistical tools available in Microsoft Excel.

Introduction

The data set used in this lab is taken from the 2000 United States Census data. The data consists of seven variables (population, number of males, number of females, number of households, number of workers, median household income, median household rent) for seven Texas counties (Bexar, Guadalupe, Comal, Hays, Travis, Williamson, and Bell) which make up the I-35 Corridor extending from Temple to San Antonio:



Map of counties included in data set "Lab0.xls"

The Census data is divided among "block groups," which are geographic areas which, in the hierarchy of Census divisions, are one step above Census "blocks." Generally, in a dense city area, a block group will consist of a dozen or so city blocks, though in a less dense or rural area, the block group can be much larger.

In the data set, each row represents one block group. There are two identifiers for each block group: the Geographic Identifier, which is a number uniquely associated with that block group (with respect to all block groups in the United States), and the Geography Description, which gives the location of that block group in terms of its County and Census Tract (which is another, larger, Census division).

Procedure

1. Save the file "Lab0.xls" to your desktop.
2. Open "Lab0.xls."
3. Answer question 1 below.
4. In Excel, under the "Tools" menu, look for option "Data Analysis..." If you see it, click on it and continue to step 4. If you don't see it, click on "Add Ins..." and check the box with "Analysis ToolPak" and click "OK." Now, under the "Tools" menu, select "Data Analysis..."

5. Select the option “Descriptive Statistics” and click OK.
6. Select as input “Median household income in 1999” and “Median rent” (columns H and I). If you have selected the whole column, check the box marked “Labels in First Row.” Check the “Summary Statistics” box and the “New Worksheet Ply” button. Name your worksheet ply “Stats.” Click OK. Repeat the procedure for the variables “Males” and “Females” (columns D and E) (place these results in worksheet “Stats” by using “Select Output Range” in Descriptive Statistics and specifying the appropriate placement in “Stats”). Answer question 2 below.
7. In worksheet “Sheet 1,” label the first row in column J “Average Workers per Household.” For all of the block groups with zero households, type in zero for the corresponding value in column J (rows 2 – 21). Make the first block group with some households equal to “Workers” divided by “Households.” If you don’t know how to do this, write in the text (in column J, row 22): =G22/F22 Copy this value and paste it for the rest of column J. Now run descriptive statistics as in part 6. Answer question 3.
8. The statistics for these variables do not necessarily represent the statistics for the actual population of the seven Counties. That is, the mean of the “Median household income in 1999” doesn’t necessarily equal the actual Median household income in 1999 for the seven Counties. Why is this? (Assume the Census is a complete and accurate 100% survey of the population) You may answer this in the space provided under question number 4.
9. In the worksheet "Sheet1", select the menu "Data". Select the option "Sort...". Sort the data by "Geography Identifier", making sure the bullet marked "Header Row" is selected (a problem box converting numbers formatted as text may appear, simply click "ok"). Run Descriptive Statistics on the variables "Males" and "Females" only for those block groups in Bell County. Write your thoughts on how Bell County is different from or similar to the counties viewed aggregately. You may use the space for question 5.
10. Create a new variable like you did in part 7 above, for instance, ratio of Median rent to Median household income. Why, as an engineer did you decide to study this variable? If you were given the option to collect data on some other variable, which would you choose? Answer this under question 6.
11. Print out your worksheet “Stats” and turn it in with this lab report. Please DO NOT print the data itself.

Questions

Note: We are looking for ideas for the most part. Try to be as succinct as you can.

- 1) The data is a sample of size 1980 with values of a set of seven variables for each block group in the sample. Examine the variables carefully for any redundancies. How many effective variables are there? (5 pts)

2)

a) Look at the descriptive statistics for the “median household income”. From the values for the minimum, maximum, and the median (this is the halfway point), what can you infer about the clustering of data, i.e. do the values tend to be regularly spaced, or bunched up? Give reasons for your answer. (15pts)

b) What, in your opinion, are the major differences between the statistics for males and females from the seven counties? You may be brief here. (25pts)

3) When a value is well away from the other values of the sample, we call it an “outlier”. Do you see any from the statistics for “Average workers per household”? How can you tell? Do outliers matter? Answer this last question in the viewpoint of mean and median values -in other words, would you expect a big change in the values of the mean and median if you discarded a heavy outlier? (25pts)

4) (See part 8 of procedure) (10pts)

5) (See part 9 of procedure) (10pts)

6) (See part 10 of procedure) (10pts)