Statistical Thinking for Forensic Practitioners

Excel Lab on Part 3: Data Collection

We will practice various sampling techniques on a dataset to determine how they affect a quantity of interest. Download the "Part 3 Lab - Excel" from the course website. The file contains the base pay for individuals working at construction companies in the United States. Each row represents an individual. The column descriptions are as follows:

(a) `state`: state in which the individual works

(b) `basepay`: individual's income

(c) `census region`: region of the United States in which the individual works, according to the US Census Bureau.

(d) `bracket` (%): the tax rate of the individual based on the 2020 single income tax rates (we won't use this column in this exercise).

We are interested in the average income. Suppose, for the sake of this exercise, that the individuals from which these data were collected are our population of interest. We make this assumption so that we can compare the average estimates we obtain from sampling to the true population average. Complete the following problems. Similar to the Excel lab on Part 2, there are cells in the spreadsheet that contain a number followed by three question marks (e.g., 1. ???). You can replace the contents of these cells to perform calculations when prompted.

1. Calculate the population mean. To do this quickly, type `=AVERAGE(`, hover your cursor over column `B` header (above the `basepay` column title) until it turns into a black arrow pointing downwards, and left-click. This should select the entire column. Press `)` followed by `enter` to complete the formula.

   $61733.78

2. Complete the following steps to draw a samples from the data set. If you have problems executing these steps, please refer to this article.

   (a) Add a column to the left of the `basepay` column by right clicking on the `B` column header and selecting Insert. Name this column `random`.

   (b) In cell `B2`, enter the formula `=RAND()`. This will generate a random number.

   (c) Left-click on cell `B2` and note the small square in the bottom-right corner of the cell. Your cursor should turn into a black `+` symbol when you hover over this square. Double-click the square and the remaining rows in the table should be filled in with random numbers. (Note: this is a very useful Excel trick for repeating formulas)

   (d) Now select column `B` (left-click on the column `B` header) and copy the contents (either `Ctrl+C` or right-click and select Copy). Click the arrow below `Paste` in the Home tab at the top of the spreadsheet. Select the first option under `Paste Values`. This will ensure that Excel treats the cells' contents as numbers rather than as a formula (allowing us to sort by these numbers).

   (e) Now select all 5 columns in the data set by holding left-click on the `A` column header and dragging your cursor to the `E` column header. The 5 columns should be highlighted.

(f) With the 5 columns selected, click on the `Sort` button under the Data tab at the top of the spreadsheet. In the menu that pops up, click the arrow next to `Sort by` and select `random`. Click `OK`.

(g) We will behave as if every 12 data rows constitutes a sample drawn from this population. Copy the data in rows 2 through 37 and paste it in `sampling1` tab accessible at the bottom of the spreadsheet. These 36 rows represent 3 samples.

3. Were the samples drawn in problem 2 above Simple Random Samples (SRS), Stratified Samples, or Cluster Samples? Were the samples drawn with or without replacement?

   Simple random sample drawn without replacement

4. Calculate the sample average of each sample (so use the `AVERAGE` function on each group of 12 rows). You may record the answers here, or just keep them in the spreadsheet.

   Answers will vary.

5. How do these sample averages compare to the population average from question 1? Is there any reason to suspect heterogeneity among the population based on these values? Explain.

   Answers will vary. Some individuals may find that the sample mean is far from the population mean, indicating possible heterogeneity (although the sample size drawn isn't particularly large in the first place).

6. Whatever you answer to question 5, suppose we concluded that the population may be heterogeneous. In particular, we believe that certain regions may be represented in the data more than others. To determine whether this is true, sort the rows by the `census region` column in the original, population data set. Use the `COUNTIF` function to count the number of occurrences of each region (for practice reading helper websites, read about the function here). What do you conclude?

   South seems to be over represented and Northeast under represented.

7. Based on the conclusion from question 6, suppose we decide to use stratified sampling to draw a new sample of 12 observations. This will ensure that exactly 3 observations come from each region. Clear the contents of `random` column you created above and then follow the instructions in this video to perform stratified random sampling (the steps are very similar), making sure to select 3 observations per strata to form the new sample. You can copy the contents into the `sampling2` tab at the bottom of the spreadsheet.

8. Calculate the sample average for this new sample. How does it compare to the previous sample averages and the population average?

   Answers will vary.

9. Instead of performing stratified sampling, suppose we treat regions as clusters and perform cluster sampling. The easiest way to do so is to randomly sample a number from 1 to 4 using the `RANDBETWEEN` function (read about it here). Associate Mid-West with 1, North-East with 2, South with 3, and West with 4. Calculate the sample average for whatever region is chosen. How does it compare to the sample averages previously computed and population averages?

   Answers will vary.