Stata Lab 1: Unbiasedness and Consistency of the Sample Mean

Background Information:
United States Postal Service Employee Productivity

The United States Postal Service is an independent establishment of the executive branch of the federal government, providing constitutionally-guaranteed mail service to more than 150 million addresses in the United States. It employs more than 400,000 workers across more than 30,000 retail offices, warehouses, and mail sorting facilities. The Postal Service processed and delivered more than 400 million pieces of mail *per day* in 2020. The Postal Service strives to manage its operations in a data-driven way, but given the size and scope of its operations, it is impractical to monitor productivity for its entire workforce at all times. Instead, the Postal Service employees a random-sample procedure in order to gather information about employee productivity in a cost-effective way.

In this lab, you will evaluate the properties of the sample mean of a random-sample of data as an unbiased and consistent estimator of the population mean of a random variable, employee productivity. Employee productivity is measured by the number of pieces of mail processed per hour. The dataset USPSproductivity.dta, which can be found on Canvas, contains simulated data with the following variables:

- *id*: employee id
- *prod*: employee productivity, measured as number of mail pieces processed per hour.
- *position*: employee position, either letter carrier or retail service.

In addition, this dataset contains variables that identify random samples of employees. These will be explained later in the lab.

This lab contains three parts. The first part *must be completed* ***before you come to class***. I will randomly select one to two students during the in class lab to present their findings. If you do not complete this work ahead of time, you will be behind when we begin our in-class lab and you risk being unprepared to discuss your findings in the class.

Pre-Class Assignment: Complete Before Class

1. Setup your digital workspace
   - Create a folder where you will keep files associated with this week's lab.
   - Copy the USPSproductivity.dta dataset from Canvas into the folder you've just created.
   - Set your working directory to point to this folder using the following command
     `cd "INSERT YOUR DATA PATH HERE"`
     Once you set your working directory, all of your work will be saved in this folder.
2. Start a new do-file. A do-file is a program file that executes commands within Stata. Save your do-file in your working directory with the title *Lab0_1_preClassWork.do*.
3. Start a log file by typing the following command into your do-file and executing your do-file by pressing the do-button.
   `log using "Lab0_1.smcl", replace`
4. Load in your data using the following command
   `use USPSproductivity.dta, clear`
5. Explore your dataset using the following command
   `desc id prod position`
   - How many observations are there?
   - What is the storage type of the variable *prod*?
   - What is the storage type of the variable *position*?

   You should type the answers to these questions in your log file. You can do this by including commented lines of text in your do-file. A commented line of text must begin with a `*` or a `//` symbol. Any text following this symbol, until the end of your line, will be interpreted as a comment that Stata will not try to execute.
   ```
   * For example, this is a comment
   // This is also a comment
   ```
6. Familiarize yourself with the variable *prod*.
   - What is the mean productivity? What is the standard deviation? Use the following command command to find these answers
     `summarize prod`
   - Plot a histogram of *prod* using the following command
     `histogram prod`
     What do you notice about the distribution of these data?

- Export your histogram as a .png file to refer to later on in this lab.
  `graph export productivityHistogram.png, as(png) replace`

7. Close your log file using the following command
   `log close`
8. Save your do-file

In Class Lab Part 1

1. I will group you with two other students to work through our in-class lab. This will be done randomly within Zoom breakout rooms. Your breakout room number will be your student group number.
2. Start a new do-file to capture the commands that you use to complete this part of the lab. Start a log file to save the results from your results screen.
3. Setup the dataset to contain only the variables that are relevant for your student number. Your student number identifies the set of random sample variables that you will analyze for the rest of this lab. For example, if your student number is 1, then one of your random sample variables is random1_10, which is a random sample of 10 employees to be analyzed by student number 1. Student 1 would use the following command to keep the set of variables that are relevant to you for this assignment
   `keep id prod position random1_10 random1_50 random1_100 random1_1000 random1_5000`
4. The USPS cannot easily observe the productivity of all employees within the organization. Instead, the USPS randomly captures the productivity of 10 employees. Student 1 could identify these employees based on the variable *random1_10*
   - What is the mean and standard deviation of productivity for this random sample of employees? You can calculate this answer using the following command
     `summarize prod if random1_10==1`
   - Calculate the standard error of the mean of this random sample by hand.

     Stata can be used as a calculator. For example, you could calculate the value of $\frac{100}{\sqrt{100}}$ using the following command `display 100/sqrt(100)`

     As a reminder, the standard error is calculated based on the following formula

     $$se = \frac{stddev}{\sqrt{N}}$$

   - Confirm that you have calculated the standard error correctly using the `mean` command.
     `mean prod`
   - Record the mean of your random sample.
5. USPS increases the size of its random sample to 50 employees.
   - What is the mean productivity for this random sample of employees?

- How does the standard error of your mean compare to the standard error for a random sample of size 10?
- Record the mean of your random sample.

6. USPS increases the size of its random sample to 100 employees.
   - What is the mean productivity for this random sample of employees?
   - Interpret the confidence interval of your sample mean.
   - Record the mean of your random sample.
7. USPS increases the size of its random sample to 1000 employees.
   - What is the mean productivity for this random sample of employees?
   - How does the size of the confidence interval of your mean compare to the mean from the random sample of 100 employees?
   - Record the mean of your random sample.

In-Class Lab Part 2

1. Load a dataset that corresponds to a counterfactual class with 50 groups, containing sample means for the various random sample sizes into Stata. You can do this using the following command
   `insheet using classdata.csv, comma clear`
2. Evaluate the distribution of the sample mean for repeated samples of size 10.
   - Plot a `histogram` of the sample means for these repeated random samples. Be sure to `export` your histogram as a .png file so that you can reference this later on.
   - What is the mean and the standard deviation of the sample means that we calculated across these repeated samples?
3. Evaluate the distribution of the sample mean for repeated samples of size 100.
   - Plot a `histogram` of the sample means for these repeated random samples. Be sure to `export` your histogram as a .png file so that you can reference this later on.
   - What is the mean of the sample means across these repeated samples?
   - How does the standard deviation of the sample means for random samples of size 100 compare to random samples of size 10?
4. Evaluate the distribution of the sample mean for repeated samples of size 1000.
   - Plot a `histogram` of the sample means for these repeated random samples. Be sure to `export` your histogram as a .png file so that you can reference this later on.
   - What is the mean of the sample means across these repeated samples?
   - How does the standard deviation of the sample means for random samples of size 1000 compare to random samples of size 100?
5. Explain what it means for the sample mean to be unbiased based of the population mean based on the histograms and means you've just analyzed.
6. Explain what it means for the sample mean to be a consistent estimator of the population mean based on the histograms and standard deviations you've just analyzed.