Activity_Explore sampling

June 3, 2024

1 Activity: Explore sampling

1.1 Introduction

In this activity, you will engage in effective sampling of a dataset in order to make it easier to analyze. As a data professional you will often work with extremely large datasets, and utilizing proper sampling techniques helps you improve your efficiency in this work.

For this activity, you are a member of an analytics team for the Environmental Protection Agency. You are assigned to analyze data on air quality with respect to carbon monoxide—a major air pollutant—and report your findings. The data utilized in this activity includes information from over 200 sites, identified by their state name, county name, city name, and local site name. You will use effective sampling within this dataset.

1.2 Step 1: Imports

1.2.1 Import packages

Import `pandas`, `numpy`, `matplotlib`, `statsmodels`, and `scipy`.

[ ]:
```
# Import libraries and packages

### YOUR CODE HERE ###
```

1.2.2 Load the dataset

As shown in this cell, the dataset has been automatically loaded in for you. You do not need to download the .csv file, or provide more code, in order to access the dataset and proceed with this lab. Please continue with this activity by completing the following instructions.

[ ]:
```
# RUN THIS CELL TO IMPORT YOUR DATA.

### YOUR CODE HERE ###
epa_data = pd.read_csv("c4_epa_air_quality.csv", index_col = 0)
```

Hint 1

Use the function in the `pandas` library that allows you to read in data from a csv file and load it into a DataFrame.

Hint 2

Use the `read_csv` function from the pandas `library`. Set the `index_col` parameter to `0` to read in the first column as an index (and to avoid `"Unnamed: 0"` appearing as a column in the resulting Dataframe).

1.3 Step 2: Data exploration

1.3.1 Examine the data

To understand how the dataset is structured, examine the first 10 rows of the data.

[ ]:
```
# First 10 rows of the data

### YOUR CODE HERE ###
```

Hint 1

Use the function in the `pandas` library that allows you to get a specific number of rows from the top of a DataFrame.

Hint 2

Use the `head` function from the `pandas` library. Set the `n` parameter to `10` to print out the first 10 rows.

Question: What does the `aqi` column represent?

[Write your response here. Double-click (or enter) to edit.]

1.3.2 Generate a table of descriptive statistics

Generate a table of some descriptive statistics about the data. Specify that all columns of the input be included in the output.

[ ]:
```
### YOUR CODE HERE ###
```

Hint 1

Use function in the `pandas` library that allows you to generate a table of basic descriptive statistics in a DataFrame.

Hint 2

Use the `describe` function from the `pandas` library. Set the `include` parameter passed in to this function to 'all' to specify that all columns of the input be included in the output.

Question: Based on the preceding table of descriptive statistics, what is the mean value of the `aqi` column?

[Write your response here. Double-click (or enter) to edit.]

Question: Based on the preceding table of descriptive statistics, what do you notice about the count value for the `aqi` column?

[Write your response here. Double-click (or enter) to edit.]

1.3.3 Use the `mean()` function on the `aqi` column

Now, use the `mean()` function on the `aqi` column and assign the value to a variable `population_mean`. The value should be the same as the one generated by the `describe()` method in the above table.

[ ]:
```
### YOUR CODE HERE ###
```

Hint 1

Use the function in the `pandas` library that allows you to generate a mean value for a column in a DataFrame.

Hint 2

Use the `mean()` method.

1.4 Step 3: Statistical tests

1.4.1 Sample with replacement

First, name a new variable `sampled_data`. Then, use the `sample()` dataframe method to draw 50 samples from `epa_data`. Set `replace` equal to `'True'` to specify sampling with replacement. For `random_state`, choose an arbitrary number for random seed. Make that arbitrary number `42`.

[1]:
```
### YOUR CODE HERE ###
```

1.4.2 Output the first 10 rows

Output the first 10 rows of the DataFrame.

[ ]:
```
### YOUR CODE HERE ###
```

Hint 1

Use the function in the `pandas` library that allows you to get a specific number of rows from the top of a DataFrame.

Hint 2

Use the `head` function from the `pandas` library. Set the `n` parameter to `10` to print out the first 10 rows.

Question: In the DataFrame output, why is the row index 102 repeated twice?

[Write your response here. Double-click (or enter) to edit.]

Question: What does `random_state` do?

[Write your response here. Double-click (or enter) to edit.]

1.4.3 Compute the mean value from the `aqi` column

Compute the mean value from the `aqi` column in `sampled_data` and assign the value to the variable `sample_mean`.

[ ]:
```
### YOUR CODE HERE ###
```

Question: Why is `sample_mean` different from `population_mean`?

[Write your response here. Double-click (or enter) to edit.]

1.4.4 Apply the central limit theorem

Imagine repeating the the earlier sample with replacement 10,000 times and obtaining 10,000 point estimates of the mean. In other words, imagine taking 10,000 random samples of 50 AQI values and computing the mean for each sample. According to the **central limit theorem**, the mean of a sampling distribution should be roughly equal to the population mean. Complete the following steps to compute the mean of the sampling distribution with 10,000 samples.

- Create an empty list and assign it to a variable called `estimate_list`.
- Iterate through a `for` loop 10,000 times. To do this, make sure to utilize the `range()` function to generate a sequence of numbers from 0 to 9,999.
- In each iteration of the loop, use the `sample()` function to take a random sample (with replacement) of 50 AQI values from the population. Do not set `random_state` to a value.
- Use the list `append()` function to add the value of the sample `mean` to each item in the list.

[ ]:
```
### YOUR CODE HERE ###
```

Hint 1

Review the content about sampling in Python.

1.4.5 Create a new DataFrame

Next, create a new DataFrame from the list of 10,000 estimates. Name the new variable `estimate_df`.

[ ]:
```
### YOUR CODE HERE ###
```

Hint 1

Review the content about sampling in Python.

Hint 2

Use the `mean()` function.

1.4.6 Compute the mean() of the sampling distribution

Next, compute the `mean()` of the sampling distribution of 10,000 random samples and store the result in a new variable `mean_sample_means`.

```
[ ]: ### YOUR CODE HERE ###
```

Hint 1

Use the function in the `pandas` library that allows you to generate a mean value for a column in a DataFrame.

Hint 2

Use the `mean()` function.

Question: What is the mean for the sampling distribution of 10,000 random samples?

[Write your response here. Double-click (or enter) to edit.]

Hint 3

This value is contained in `mean_sample_means`.

Hint 4

According to the central limit theorem, the mean of the preceding sampling distribution should be roughly equal to the population mean.

Question: How are the central limit theorem and random sampling (with replacement) related?

[Write your response here. Double-click (or enter) to edit.]

1.4.7 Output the distribution using a histogram

Output the distribution of these estimates using a histogram. This provides an idea of the sampling distribution.

```
[ ]: ### YOUR CODE HERE ###
```

Hint 1

Use the `hist()` function.

1.4.8 Calculate the standard error

Calculate the standard error of the mean AQI using the initial sample of 50. The **standard error** of a statistic measures the sample-to-sample variability of the sample statistic. It provides a numerical measure of sampling variability and answers the question: How far is a statistic based on one particular sample from the actual value of the statistic?

[ ]:
```
### YOUR CODE HERE ###
```

Hint 1

Use the `std()` function and the `np.sqrt()` function.

1.5 Step 4: Results and evaluation

1.5.1 Visualize the relationship between the sampling and normal distributions

Visualize the relationship between your sampling distribution of 10,000 estimates and the normal distribution.

1. Plot a histogram of the 10,000 sample means
2. Add a vertical line indicating the mean of the first single sample of 50
3. Add another vertical line indicating the mean of the means of the 10,000 samples
4. Add a third vertical line indicating the mean of the actual population

[ ]:
```
### YOUE CODE HERE ###
```

Question: What insights did you gain from the preceding sampling distribution?

[Write your response here. Double-click (or enter) to edit.]

2 Considerations

What are some key takeaways that you learned from this lab?

What findings would you share with others?

What would you convey to external stakeholders?

Congratulations! You've completed this lab. However, you may not notice a green check mark next to this item on Coursera's platform. Please continue your progress regardless of the check mark. Just click on the "save" icon at the top of this notebook to ensure your work has been logged.