Data Science Essentials

Lab 3 – Simulation and Hypothesis Testing

Overview

In this lab, you will learn how to create, run and interpret simulations using R or Python. Simulation is widely used in cases where estimates are required from complex distributions of values or a hierarchy of distributions.

After completing the simulation, you will learn how to use either R or Python to compute and understand the basics of hypothesis testing. Hypothesis testing is widely used. Any time you are trying to determine if a parameter or relationship is statistically significant you perform a hypothesis test.

What You'll Need

To complete this lab, you will need the following:

- A Web browser
- An Azure Machine Learning workspace.
- The files for this lab

Note: To set up the required environment for the lab, follow the instructions in the Setup Guide for this course.

Performing a Simulation

In this exercise, you will estimate the range of expected profitability for a lemonade stand. The profitability of the lemonade stand depends on the number of customers arriving, the profit from the drinks they order, and the tips the customer may or may not choose to leave. The distribution of possible profits is thus, the joint distribution of customer arrivals, items ordered, and tips. In practice, such a complex distribution cannot be analyzed except using simulation.

Run Code to Perform a Simulation

The code for this exercise is provided in a Jupyter notebook. Both R and Python versions of the notebook are provided.

1. Browse to https://studio.azureml.net and sign in using the Microsoft account associated with your free Azure ML account.
2. On the **Notebooks** tab, click **+NEW**. Then select the option to upload a notebook from a local file.

3. Select the **Simulation (R).ipynb** or **Simulation (Python).ipynb** file in the **Mod3** folder where you extracted the lab files for this course, accept the default name for the notebook, and select the appropriate language (R or Python 3).
4. After the notebook has been uploaded, open it and follow the instructions it contains.

Hypothesis Testing

In this exercise, you will explore and perform hypothesis tests on a famous data set collect by Frances Galton (1822-1911), who invented the regression method. Galton collected these data from Families living in late 19th century London. Galton published his famous paper in 1885, showing that the highs of children regressed to the mean of the population, regardless of the heights of the parents. From this seminal study, we get the term *regression* in statistics.

Upload the Galton Dataset

1. From the folder where you extracted the lab files for this module (for example, C:\DAT203.1x\Mod3), open the **GaltonFamilies.csv** file, using either a spreadsheet application such as Microsoft Excel, or a text editor such as Microsoft Windows Notepad.
2. View the contents of the **GaltonFamilies.csv** file, noting that it contains data on 934 cases, documenting height for family members. Then close the text file without saving any changes.
3. Browse to https://studio.azureml.net and sign in using the Microsoft account associated with your free Azure ML account.
4. On the **Datasets** tab, click **+NEW**. Then select the option to upload a dataset from a local file, and upload the **GaltonFamilies.csv** file as a new dataset named **GaltonFamilies.csv**.

Run Code to Perform Hypothesis Testing

The code for this exercise is provided in a Jupyter notebook. Both R and Python versions of the notebook are provided.

1. In Azure Machine Learning Studio, on the **Notebooks** tab, click **+NEW**. Then select the option to upload a notebook from a local file.
2. Select the **Hypothesis (R).ipynb** or **Hypothesis (Python).ipynb** file in the **Mod3** folder where you extracted the lab files for this course, accept the default name for the notebook, and select the appropriate language (R or Python 3).
3. After the notebook has been uploaded, open it and follow the instructions it contains.