**Module-B2: Basic Statistical Concepts**

**Lab by Jinlai Xu (if you have any questions, please contact** **jinlai.xu@pitt.edu****)**

# Prerequisites

Before working on the lab, we need to first setup the environment. In this lab, we need to first install the Python running environment.

There are two ways to run the jupyter notebook either on your computer or on the cloud environment (we recommend running it on your own computer that you can save the results).

## On your own computer

If running the python environment on your own computer. We choose Anaconda, which is an integrated python running environment that already includes the most used python libraries you need to use in the lab.

To install anaconda, refer to the documentation for different operating systems:

1. Windows: <https://docs.anaconda.com/anaconda/install/windows/>
2. Mac: <https://docs.anaconda.com/anaconda/install/mac-os/>
3. Linux: <https://docs.anaconda.com/anaconda/install/linux/>

After installing the environment, we can start the jupyter notebook, which has a web UI that is convenient for newbies to start with.

1. Copy the zip file provided with the lab, “Lab-ModuleB2.zip” to a folder, e.g., “C:\Users\<your user name>\Downloads\" on windows, “~/Downloads” on mac. Then, extract the zip file to the current folder.
2. Open a Terminal, on windows, it could be found on your startup menu as “Anaconda Powershell Prompt (Anaconda3)”, on Mac, you can use the default “terminal”.



1. “cd” to the above folder you extract the zip file by typing ”cd C:\Users\<your user name>\Downloads\” or “cd ~/Downloads”
2. Then, start the jupyter lab from the terminal by typing “jupyter lab”. Then, the tool will create the environment and automatically prompt to open a page in your default browser, which is a web site running on your own computer:



1. At last, simply click the “HypothesisTesting.ipynb” in the folder you extract to. Enjoy your journal!

## On cloud (colab) environment

1. Open the github repository: <https://github.com/xujinlai/sadet_lab>
2. Open the colab website: <https://colab.research.google.com/notebooks/intro.ipynb>



1. Go to File -> Open notebook -> Github
2. Enter the GitHub repository URL and click search, then open the “HypothesisTesting.ipynb”
3. Add the dataset “PSID.csv” by uploading the file to colab

#  Exploration

Make sure to put the dataset “PSID.csv” in the folder to the same place you place the notebook

Then on the jupyter lab web page, you will see the UI as below:



You can simply click the “run  button” to run the current cell in the notebook to see the output.

In this particular lab, most of the codes are already provided and what you need to do is to just fit part of the codes in corresponding places.

If you have any parts you cannot finish by yourself, please refer to the answer, which is provided in “<>-answer.ipynb”

# Hypothesis Testing

In this lab, we can learn the basic process of hypothesis testing with two examples: (i) a fair/unfair coin example, and (ii) an advanced example with a real dataset related to earning data.

The first cell needs to be run to load the libraries we may use in the remaining lab.

## 3.1 A fair/unfair coin example

In this example, we make an unfair coin that has the possibility 65% to get a head.

We then based on the hypothesis testing to test whether the coin is fair based on samples.

The code is provided and there are several parts that need to be fixed by you noted in “<Your Code>”.

Please fix the code, run the cells, and answer the questions below:

1. Based on the fair/unfair coin example, explain why we need to do the hypothesis testing.
2. Explain the basic steps we need to do for z-test.
3. Explain why we need to assume the normal distribution when using the z-test.
4. As you can see, we can also plot the distribution of the coin and intuitively get the difference between the current unfair coin with a fair coin, so why do we also need hypothesis testing, and what is the major difference between hypothesis testing and an inference (Bayesian or Frequentist).

## 3.2 An advanced example with a real dataset

In this part, we have an earning dataset on our hands. What we need to know is trying to get some conclusions from the analysis of the dataset.

Here, we are trying to test the hypothesis we have from analyzing the data that the higher education level may lead to a higher earning also.

The code is provided and several parts need to be fixed by you noted in “<Your Code>”.

Please fix the code, run the cells, and answer the questions below:

1. Explain the difference between z-test and t-test.
2. By exploring the dataset, what hypotheses you can make?
3. Explain why we need to clean the dataset, for example, we remove the entries that have an education level of more than 20.
4. From the correlation table, what you can find? And why there are both positive and negative coefficients?
5. As you can see, we use a resampling from the overall dataset to simulate the sampling from the real distribution. Explain why and what the influence of conducting random sampling.
6. In the null hypothesis, we assume high and low education will have similar earnings, but for the alternative hypothesis, the statement is not the opposite of the null hypothesis, which is the higher education will lead to higher earnings. Why we can set the alternative hypothesis like that?

## 3.3 Bonus

From the two examples above, you have learned the basic approaches of z-test and t-test.

You can make other hypotheses from analyzing the real dataset and based on the t-test template we give to test your hypotheses.

Have fun!