Assignment # 1

Dihui Lai
dlai@wustl.edu

Washington University in St Louis — February 24, 2019

1 Introduction to Statistics

1.1 Probability Distribution

You have recently joined a data science team and working on a project that needs to simulate 5 types of distributions (Bernoulli, Poisson, Gaussian, uniform distribution and Rolling-Dice distribution). Your teammates provides you with a simulated data sample "**sample_trials.csv**". In the file, each column contains 5000 data sample draw from one of the 5 distributions. However, your teammate did not label them properly. Since your teammate is on vacation, you decide to figure them out by yourself.

(1) Do the columns have discrete value or continuous value?

(2) As a first step, you investigate the distribution probability of each column by plot their histograms.

(3) Lable the column properly using the distribution name.

(4) Calculate the mean and variance of each column. Write down the distribution formula and the corresponding parameters.

1.2 Random Walk

A 1-dimension random walk is defined as a successive movements, where at each step an object can either move forward ($+1$, of probability = 0.5) or backward (-1 of probability = 0.5).

(1) Simulate a random walk of 1000 step. Set the object at position 0 initially. At each step

- Draw a random number r from the uniform distribution [0, 1)
- Update uhe position x
  $x \leftarrow x + 1$ if $r > 0.5$
  $x \leftarrow x - 1$ if $r \leq 0.5$

(2) Simulate 2000 random walk trials and make a histogram of the position at 1000^{th} step, i.e. x^i_{1000}, $i = 1, 2, 3, ...2000$.

(3) Calculate the mean and variance of x_{1000} for the 2000 trials, does it make sense?

(4) Now, simulate 2000 random walk trials and make a histogram of the position at 3000^{th} step, i.e. x^i_{3000}, $i = 1, 2, 3, ...2000$.

(5) Calculate the mean and variance of x_{3000} for the 2000 trials, does it make sense?

(6) Explain your result using central limit theorem.

2 S&P500 index

Your friend Joe, who has recently started a consluting firm. He is working on a project to help a client understand the S&P500 index. Since both Joe and his client are new to the financiel industry and had little knowledge about the market, they decide to starting with public available data and do some simple analysis. Joe learned that you are working on a master's program on dat analytics and called you for your opinion. Joe provide you with two data set in ".csv" format.

Data:
S&P500 historcal index value: "**sp500indexdaily.csv**"
Stock price of S&P500 listed companies (2013-2018): "**sp500_cmpny_all_stocks_5yr.csv**"

2.1 Data Exploration

After getting the data, you decide to explore the data by visualizing it first. You started by looking at the records in "**sp500indexdaily.csv**" and made the following plots

(1) Plot the SP500 index against trading dates (between 2009-01-01 and 2018-12-31). Instead of using actual date as x-axis, you assign an integer to each trading date and set "2009-1-1" to "0", "2009-1-2" to "1", "2009-1-3" to "2" ... etc.).

(2) In addition to the temporal property of SP500, you look into the annual statistical distribution of the index. Specifically, you create a histogram of the SP500 index for each year between 2009 and 2018.

(3) Write down the observation/conclusions you made based on the plots in (a) and (b)

2.2 Linear Regression - Single Predictor

After looking at the SP500 index change over time i.e. plots 1.1(1), you decide a linear regression can capture the trend fairly well. Therefore, you build a linear regression model with SP500 index as your target variable and time index as your predictors.

(1) How many coefficients are there in the regression model?

(2) How do you interpret the numbers?

(3) Make an in sample prediction for the SP500 index between 2009 and 2018. Compare it with the actual SP500 index. What does the model capture/not capture?

(4) Based on the model, where do you think the SP500 index will be in year 2028.

2.3 Linear Regression - Multiple Predictors

After searching the website, you learned that SP500 is an index built on the stock price 500+ companies. The largest 5 components are "Microsoft (MSFT)", "Apple Inc. (AAPL)", "Amazon.com Inc (AMZN)", "Berkshire Hathaway Inc (BRK.B)" and "Johnson & Johnson (JNJ)". You postulate that the SP500 index might be well predicted by the 5 top components already. So you did the following

(1) Build a multi-predictor linear regression model using SP500 index as the target variable and the stock price of MSFT, AAPL, AMZN, BRK.B and JNJ as the predictors (between 2013-02-08 and 2018-02-07)

(2) Check the significance of the variables by looking at the p-values of each variable. Are they significant?

(3) Do an in-sample prediction using the model and compare the predicted SP500 index with the actual SP500 index. How do you like the prediction.

(4) Drop one of the 5 variables from the model, how does the coffeicients of each variable change?