# **Project 2: Decision making Based On Historical Data**

Suraj Eswaran

University Of The Cumberlands

2024 Spring – Statistics For Data Science (MSDS -531 – A01) – First Bi-Term

Professor Dr. Danny T. Barnes

01/28/2024

**This project reflects the basics of data distribution. The project topics relate to the definitions of variance and skewness.**

**Files needed for the project are attached.**

**Cover in the project the following:**

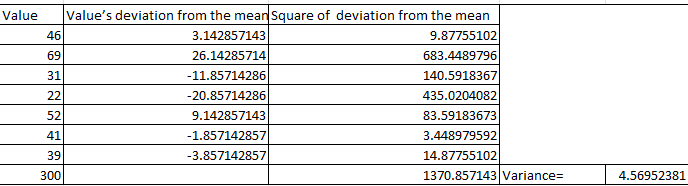
**Explain the variance and skewness.**

Variance can be measured as the variability that is reflected from the average or mean of the set of data. It is denoted by the symbol “sigma square”. It is calculated with the help of average of standard deviation of each value from the mean of the data set. Skewness can be measured as an imbalance that has been obtained from mean of set of data. Positive skewed graph will have right tail with positive skew value, where negative skewed graph will have left tail with negative skew value.

**Show a simple example of how to calculate variance and then explain the meaning of it.**

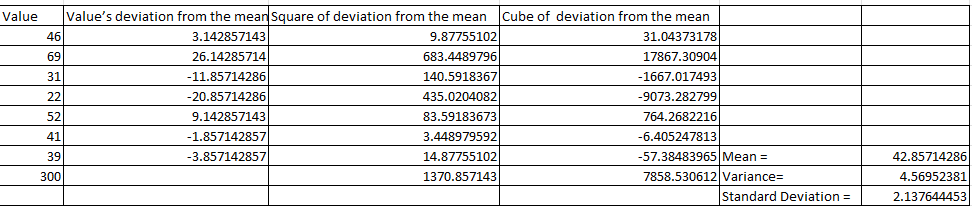
Variance is calculated by sum of square of difference between each value in a data set with mean by total number of values that are there in data set. Let’s take an example of data set, which consists of numbers {46,69,31, 22,52,41,39}, where we have number of observations as 7. First thing we need to find out mean, where we will add all the numbers and divide them by total number of observations. Thus, the mean of the following observation is 42.857.

Now, we have to find each value’s deviation from the mean, which is basically difference between each value with observed mean. After that, we have to find out square of each deviations that we have obtained from the difference and find out their total. Once the total is calculated, the variance is found out by dividing the square of difference between each deviations and total number of observations. From the above example, we have found out the variance value to be equal to 4.569.

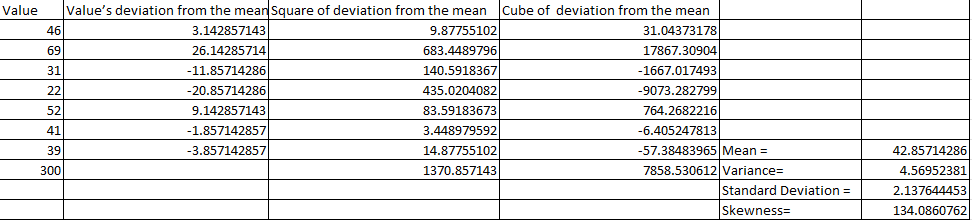


**Show a simple example of how to calculate skewness and then explain the meaning of it.**

Skewness is calculated by dividing the sum of cube of difference between each value in a data set with mean by product of total number of values and cube of standard deviation. Let’s take an example of data set, which consists of numbers {46,69,31, 22,52,41,39}, where we have number of observations as 7. First thing we need to find out mean, which is equal to 42.857. As we know standard deviation is equal to square root of variance, thus square root of 4.569 is equal to 2.137.

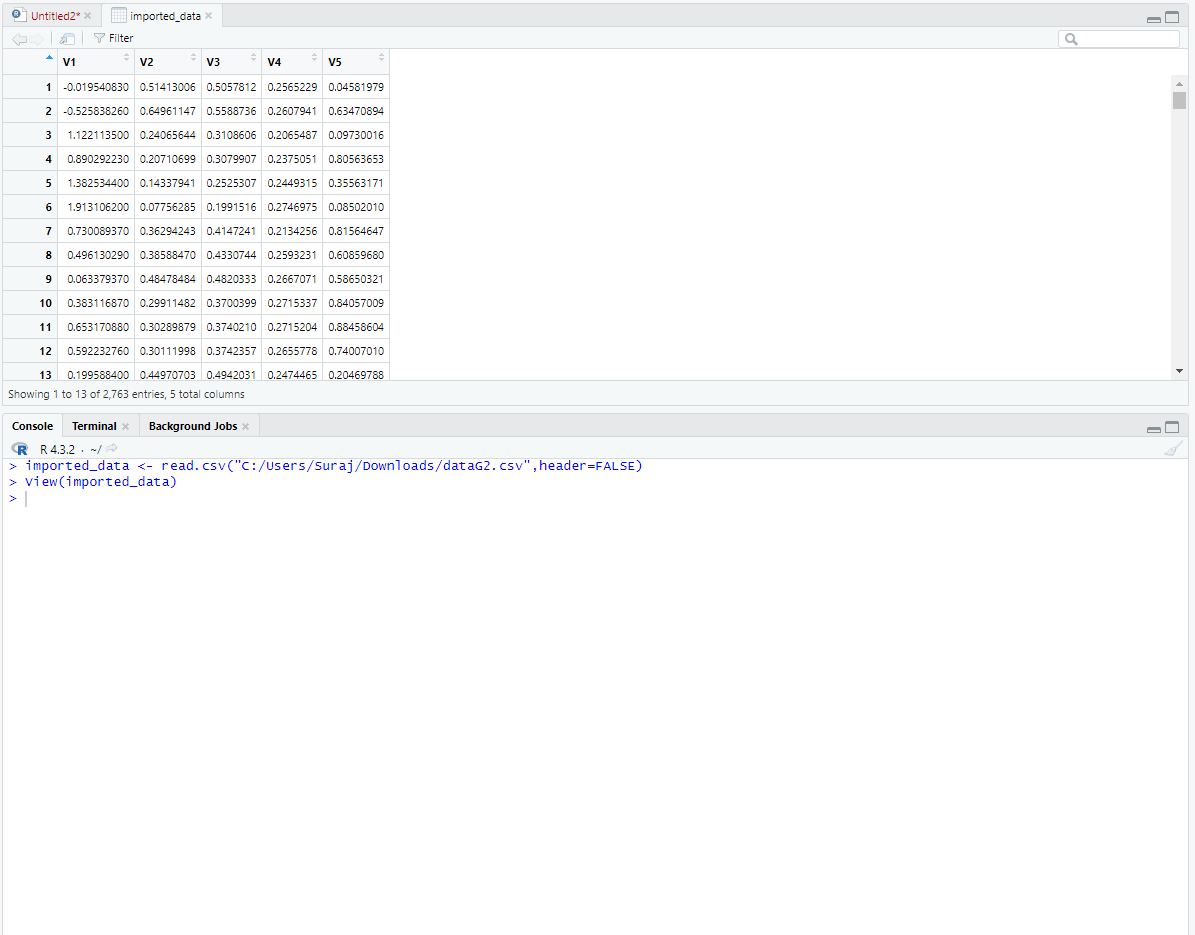


Using the formula of skewness with mean, number of observations and standard deviation, we have found out that value of skewness is equal to 134.086, which shows that it has positive skewed where most of the extreme values are present on right side of the graph and there are more values that are higher than mean, as compared to values lower than mean.



## After loading dataG2.csv into R or Octave, explain the meaning of each column or what the attributes explain. Columns are for skewness, median, mean, standard deviation, and the last price (each row describes with the numbers the distribution of the stock prices):

In dataG2.csv file, we have a list of columns of skewness, median, mean, standard deviation and last price where each row explains the number of distributions of stock prices. We are storing that csv file using “read\_csv” method where we are mentioning header as “FALSE” (Thulin, 2021, p. 68).



From the excel file, we have inferred skewness column which measures the asymmetry of a data set. If we have distribution toward right side, then that shows positive skewness, left side indicates negative skewness, and zero skewness will be towards symmetric side. The negative value in the column shows data is skewed to left side of the graph whereas positive value in the column shows data is skewed to the right side of the graph.

The second column indicates the median which is basically the middle value that are existing at the data set. Checking the column, we have observed the maximum value as 0.7557757 at index 575 and lowest value to be 0.000106703 located at index 1144.



The next column indicates mean which is actually sum of all observations by total number of observations. Larger the mean value, greater the values that are existing in the data set. With the help of R, we could find out that maximum value as 0.7019634 located at index 575(using which function) and minimum value as 0.000341332 at index 1144(Thulin, 2021, p. 109).



The fourth column represents standard deviation which is basically squared differences with respect to mean value. From the inferred table, we could obtain higher standard deviation as 0.3399379 at index 1620 and least standard deviation to be 0.01374237 at index 1144.



Inferring all the results, we have obtained highest mean and median at index 575 but the value of skewness is negative, which shows higher values are on the right side, thus they are negative skewed. But comparing to this, we have the least mean, standard deviation and median value at index 1144, having a very high positive skewness(72.729233), which results that most of the higher values are left side and thus they are positive skewed.

After standard deviation, the next column is last price which is basically the current trend price which was recorded at a particular time. We have found out that the maximum that has been recorded is 1 which is found out index 53, whereas minimum is 0 which is obtained at index 603. Comparing these values, we have noticed that skewness at the maximum stock was found out to be positive although they are not huge, which results in positive skewness, whereas at minimum index stock, it was resulting in negative skewness.



## Draw your own conclusions based on what you learned under 1. and 2.

Understanding the concept of variance and skewness, we could view that variance and skewness can be utilized to know the shape and the distribution’s spread of its data. For example, from dataG2.csv file, we could see that highest skewness value was 72.72923 but they might not be more positive skewness, as the data in a set might not have much difference to the mean, hence they are not scattered. But if you see that at index 1620, we could see standard deviation as 0.3399379(variance = 01155578), will be more positive skewed as its skewness value is also positive and the data difference between mean is also higher. With this understanding of higher variance and positive skewness can help us analyzing whether stocks might increase or decrease.

**Explain the meaning of variables 'I\_1' and 'I\_2' after you execute (after dataG2.csv is loaded in R or Octave)**

**imported\_data <- read.csv("dataG2.csv")**

**S=imported\_data[,5]-imported\_data[,3]**

**I\_1 =which.min(S) # use figure I\_1 (see attached)**

**I\_2 = which.max(S) # use figure I\_2 (see attached)**

Using those comments listed above, we have imported the CSV file using “read\_csv” file, followed by subtracting mean from last price value and finding their maximum and minimum value(I\_1 and I\_2 respectively).



Having I\_1 equal to 2050, we have obtained the following S equal to -0.5488593, which shows the last price value is lesser than mean value and median value. Since the value of S is minimum, so we could take in consideration that it is advisable to buy stocks. In the case of I\_2 equal to 1817, the value of S is equal to 0.8820354, which shows the last price value is higher than mean and median value, thus it will be advisable to sell those stock.

**Based on the results in a., which row (stock) would you buy and sell and why (if you believe history repeats)?is**

Seeing the results from I\_1 and I\_2, we could say that it is advisable to buy those stocks which has lower S value as purchasing lower price will be profitable for us, and selling those stocks which has higher S value as selling higher price will be profitable. In our case, we could buy I\_1 stocks and sell it at I\_2 rates, so that we could make profits.

**Explain how would you use the skewness (first column attribute) to decide about buying or selling a stock.**

As we know, skewness column is a measures of asymmetry that is existing in a data set. If the value of skewness is negative, then most of the data points are lower than the resultant mean value. With that information, we would understand the trend that with lower skewness value, there will be chance of decrease in the trend, which would be best time to sell the stocks. If the value of skewness is positive, then most of the data points are higher than the resultant mean value, thus we can visualize that there will be change of increase in the trend, so it would be the best time to buy the stock during that time. This indicates that depending on the value of skewness, we would decide about buying and selling a stock.

## If you want to decide, based on the historical data, which row (stock) to buy or sell, would you base your decision on skewness attribute (1st column) or the differences between the last prices with mean (differences between 5th attribute and 3rd attribute)? Explain.

Skewness (1st Column) measures the asymmetry that is existing in a data set. If the distribution having positive skewness value, then it will be having a longer tail at the right side, where if it is having negative skewness value, then it will be having it at the left side. If we have positive skewness value, there will be lots of positive outcomes, which implies that there is a chance of increase in the trend of the stock, thus giving us a clarity on selling stocks.

Difference between last price and mean (difference 5th column and 3rd column) giving us a visual that how the current trend is placed with respect to the mean value. If the difference is larger, then the current price might be higher than the mean value, thus it will be best time to sell the stocks. If the difference is smaller, then the current price will be lower than the mean value, thus it will be the best time to buy stocks.

Combining both Skewness(1st Column) and Difference between last price and mean( difference 5th column and 3rd column), will be beneficial to decide on purchasing or selling stocks as skewness helps in understanding how asymmetry from the trend and difference between last price and mean will showcase how closer the value is from the mean value. Thus, using these two keys will give us a clear decision on purchasing and selling of stocks.

## References

Thulin, M. (2021). Modern Statistics with R: From wrangling and exploring data to inference and predictive modelling. BoD-Books on Demand.