Beginner's Guide to Statistics and Probability Distribution

By Anupriya Gupta and Ishan Shah (Adapted from https://blog.quantinsti.com/statistics-probability-distribution/)

We have all realized that a working knowledge of statistics is essential for modelling different strategies when it comes to algorithmic trading. In fact, data science, one of the most sought-after skills in this decade, employs statistics to model data and arrive at meaningful conclusions. With that aim in mind we will go through some basic terminologies as well as the types of probability distributions which are employed in the domain of algorithmic trading.

We will go through the following topics:

- Historical Data Analysis
- Probability distribution
- Correlation

Historical Data Analysis

In this section, we will try to answer the fundamental question, "How do you analyse a stock's historical data and use it for strategy building?" Of course, for the analysis, we first need a data set!

Dataset

In order to keep it universal, we have taken the daily stock price data of Apple, Inc. from Dec 26, 2018, to Dec 26, 2019. You can download historical data from Yahoo Finance. If you are interested in downloading the data using python, you can visit the following link.

…

Mean? Mode? Median? What's the difference!!!

We will just take 5 numbers as an example: 12, 13, 6, 7, 19, 21, and understand the three terms.

Mean

To put it simply, mean is the one we are most used to, i.e. the average. Thus, in the above example, the 'mean' = (12 + 13 + 6 + 7 + 19 + 21)/6 = 13.

…

Mode

In a given dataset, the mode will be the number which is occurring the most. In the above example, since there is no value which is repeated, there is no mode. You can argue that every element is a 'mode.' But that doesn't help in summarizing the dataset.

…

the mode doesn't really make much sense for *some types of* data. A mode is especially useful when you want to plot histograms and visualize the frequency distribution.

This data set has a mode: 12, 13, 6, 7, 13, 21, 19, 13, 5, 3 The mode here is 13. There can be data sets with more than one mode.

…

Median

Sometimes, the data set values can have a few values which are at the extreme ends, and this might cause the 'mean' of the data set to portray an incorrect picture. Thus, we use the median, which gives **the middle value of the sorted data set**.

To find the median, you have to arrange the numbers in ascending order and then find the middle value. If the dataset contains an even number of values, you take the 'mean' of the middle two values. In our example, the median is (12 + 13)/2 = 12.5

…

Range

Range simply gives the difference between the min and max values of the data set.

In our data set: 12, 13, 6, 7, 19, 21. The max value is 21, the min value is 6, the range is 21 - 6 = 15.

…

Probability distribution

We have all gone through the example of finding the probabilities of a dice roll. Now, we know that there are only six outcomes on a dice roll, i.e. {1, 2, 3, 4, 5, 6}. The probability of rolling a 1 is 1/6. This kind of probability is called discrete, where there are a fixed number of outcomes.

For discrete probabilities, there are certain cases which are so extensively studied that their probability distribution has become standardized. Let's take, for example, Bernoulli's distribution, which takes into account the probability of getting heads or tails when we toss a coin.

…

Now, there are cases where the outcomes are not clearly defined. For example, the heights of all high school students in one grade. While the actual reason is different, we can say that it will be too cumbersome to list down all the height data and the probability. It is in this situation that the functions are essential.
Earlier, we said that for discrete values, the probability function is the probability mass function. In comparison, for continuous values, the probability function is known as a probability density function.

…

Standard Deviation

In simple words, the standard deviation tells us how far the value deviates from the 'mean.' *We need to* use *a* full dataset and try to understand how the standard deviation helps us in the 'arena of trading.' *Not on GED, the math for this is above GED levels.*

…

Wait! Normal distribution?

Normal distribution is a very simple and yet, quite profound piece in the world of statistics, actually in general life too. The basic premise is that given a range of observations, it is found that most of the values center around the mean and within one standard deviation away from the mean. Actually, it is said that 68% of the values are within this range. If we move ahead, then we see 95% of the values within two standard deviations from the mean.

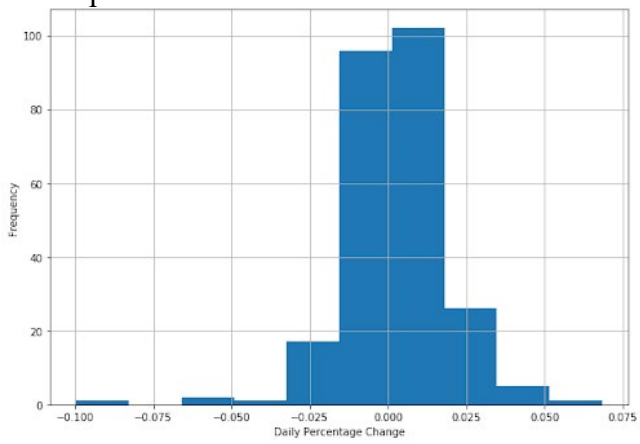
…

Histogram

Let's take an example of the heights of students in a batch. Now there might be students who have heights of 60.1 inch, 60.2 inches and so on till 60.9. Sometimes we are not looking for that level of detail and would like just to find out how many students have a height of 60 - 61 inches. Wouldn't that make our job easier and simpler? That is exactly what a histogram does. It gives us the frequency distribution of the observed values.

…

The output is as follows:



Recall how we said that the majority of the values are situated close to the mean. You can see it clearly in the histogram plotted above.
In fact, if we draw a line curve around the values, it would look like a bell.
We call this a bell curve, which is another name for the normal probability distribution, or normal distribution for short.

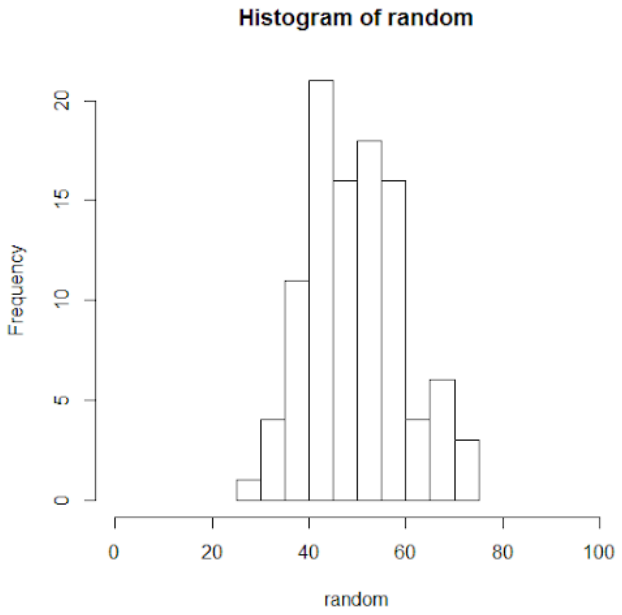
…

Normal distribution

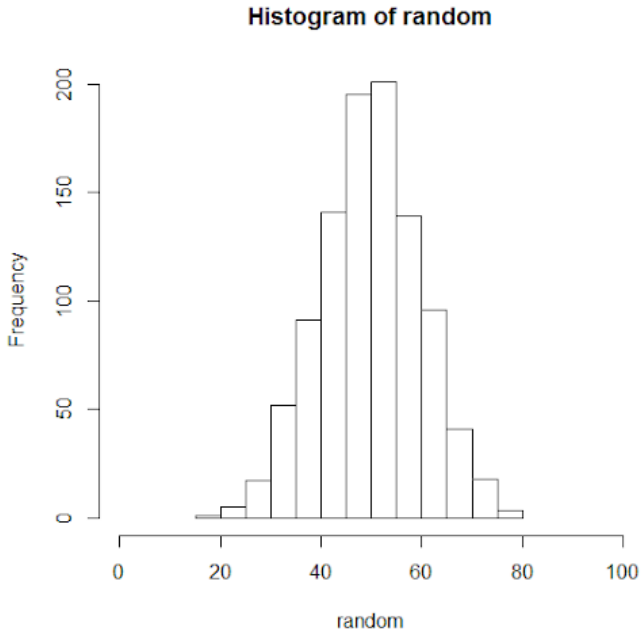
When the distribution of your data meets certain requirements, such as symmetry around the mean and bell-shaped curve, we say your data is normally distributed.

…

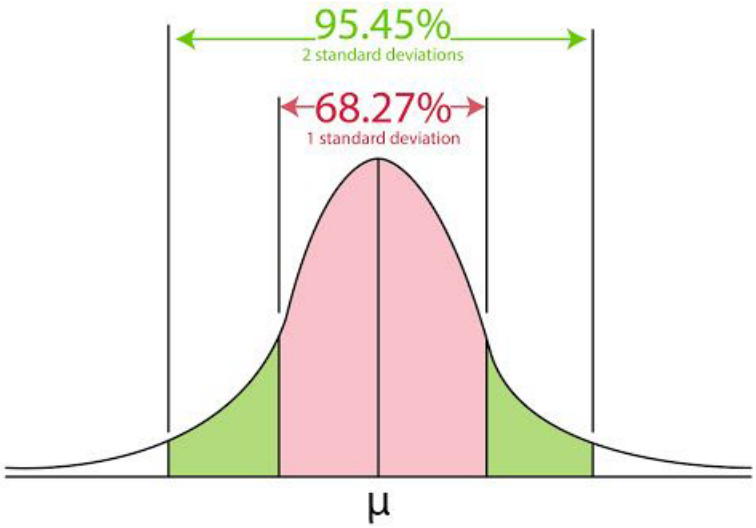
Why is it useful to know the distribution function of your dataset?
If you know that your data sample is, say, normally distributed, you can make 'predictions' about your population with certain 'confidence'.
For example, say, your data sample X represents marks obtained out of 100 in an entrance test for a sample of students. The data is normally distributed, such as X~N(50, 102). When plotted, this data would look as follows:



If you increase the number of observations in your sample data set from 100 to 1000, this is what happens:



It looks more bell-shaped!
Now that we know, X has normally distributed data with mean at 50 and standard deviation of 10, we can predict the marks of the entire student population or future students (from the same population) with a certain confidence. With almost 99.7% confidence, we can say that students would not get less than 20 or greater than 80 marks. With 95% confidence, we can say that students would get marks between 30 and 70 points.



Statistically speaking, distribution functions give us the probability of expecting the value of a given observation between two points. Hence, using distribution functions, also called probability density functions, we can ‘predict’ with certain ‘confidence’.

…

https://blog.quantinsti.com/statistics-probability-distribution/

…