Bagliotto Bastien

Machine Learning

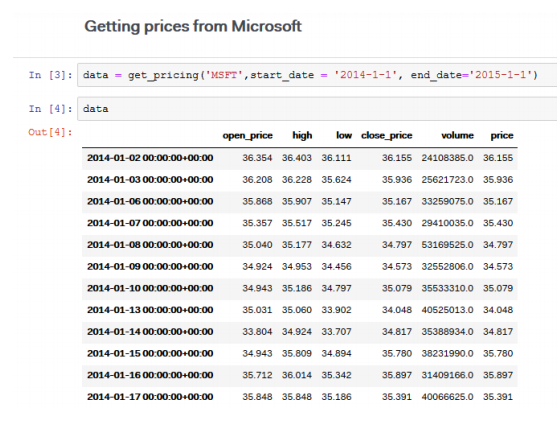
LAB 0

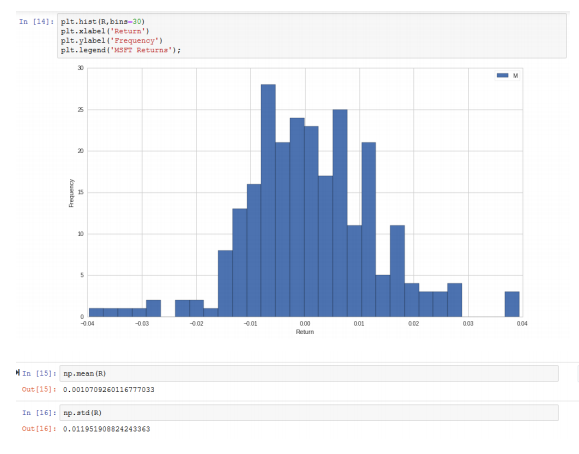
08/10/2018

Lecture 1: Introduction to Research

This lecture presents the basics of the IPython notebook. It presents various concepts like executing a command, importing libraries or plotting a graphic, which are basics ones but very useful to our course.

The most important commands are the last ones because they are used to interact directly with Quantopian and couldn’t be learned anywhere else. Theses functions are: getting the prices or the returns for example.

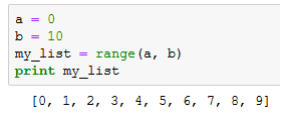


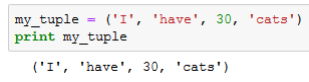


Lecture 2: Introduction to Python

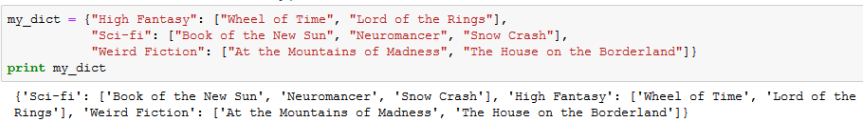
This lecture is showing us the basics of Python:

* The different basic types,
* The lists,
* The arrays,
* The tuples,
* The dictionaries,
* The functions.







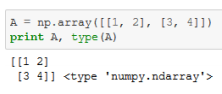


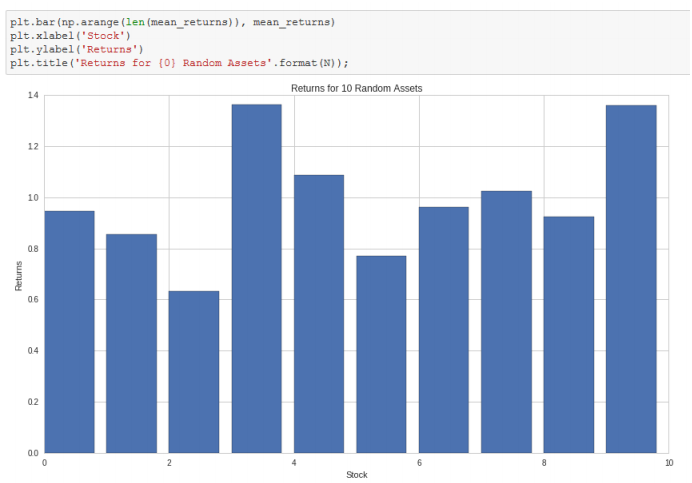
Lecture 3: Introduction to Numpy

Numpy is a Python’s library used to make mathematical manipulation over the data of Quantopian. The principal use to Numpy is to make multi-dimensional array and to manipulate very easily theses arrays.

Multidimensional arrays (matrix) can be created too, and almost every function applicable over the simple arrays can be used over matrix.

At the end of the lecture, we can find a brief foray for linear algebra, explicating matrix manipulation (multiplication) and how this could be used to calculate the portfolio variance.



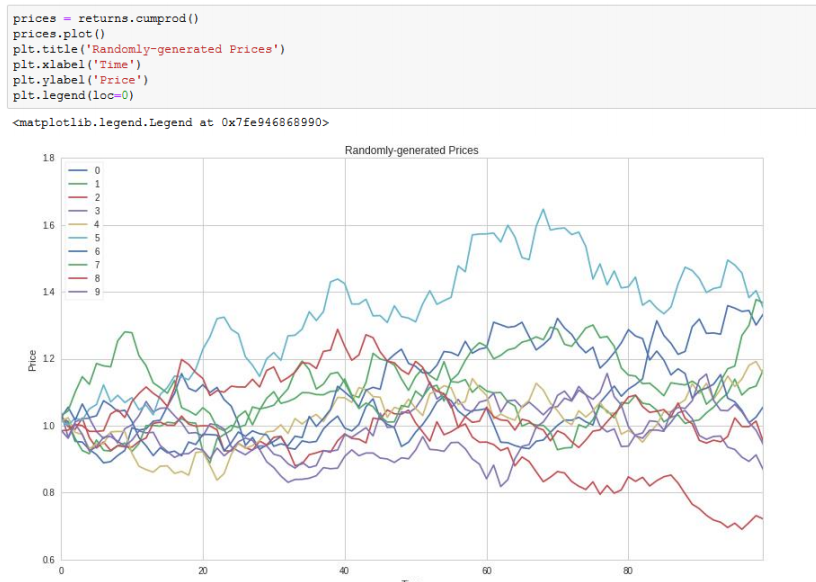


Lecture 4: Introduction to Pandas

Pandas is a Python package specifically designed to make management and analysis of data all part of the same intuitive workflow. The biggest interest of Pandas is to provide very efficient structures to manipulate data, for example the structure DataFrame or Series.

Series structure are a bit like Numpy arrays, but a Series has a name and other operators could be used on this structure, like Boolean indexing. Series can be filled by different type of data: date, float, int ... But a Series has one given type, declared during the initialization. In Quantopian they are used to do time-based analysis.

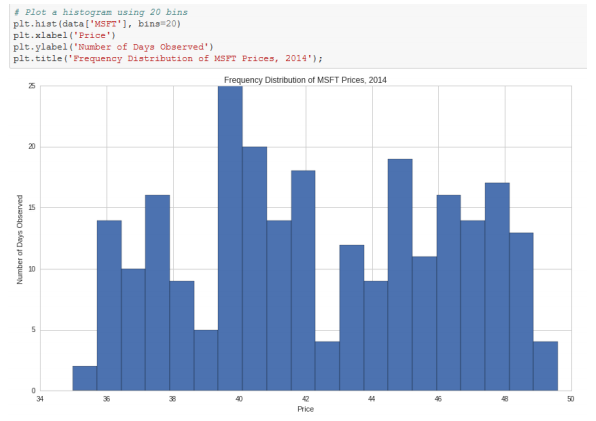
DataFrame structures are a 2-dimensional object that is a bit like Python dictionaries: for a given index, a Serie is associated with this index. As for the Series, this structure is used to do time-based analysis.

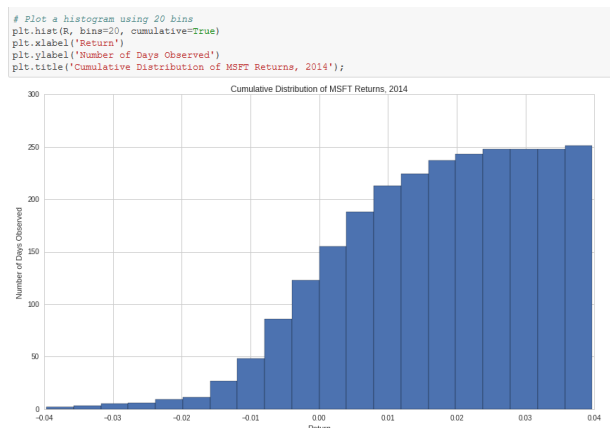


Lecture 5: Plotting data

This lecture helps use to be familiar with data plotting under his every form:

* Histograms,
* Returns histograms, used on Quantopian,
* Cumulative histograms,
* Scatter plot,
* Line graph.



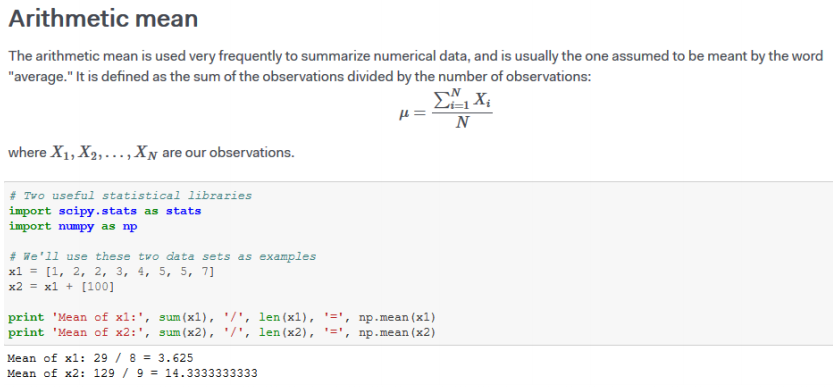


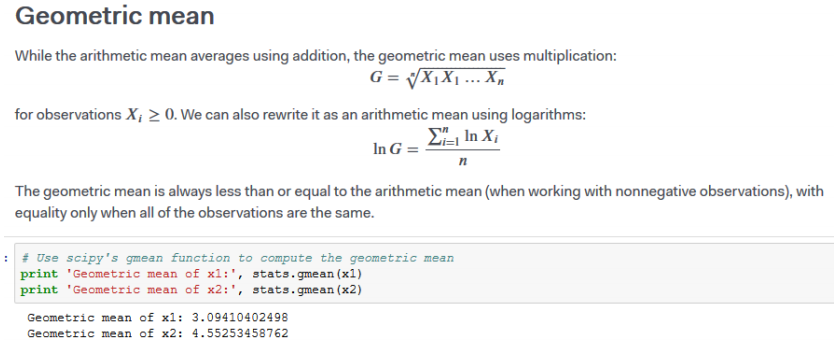
Lecture 6: Means

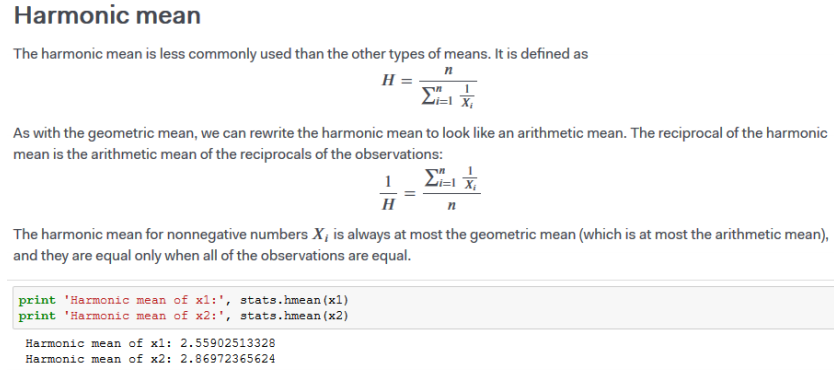
Here are explained the different means we can find:

* Arithmetic mean,
* Geometric mean,
* Harmonic mean.

Are also explained mode and median.

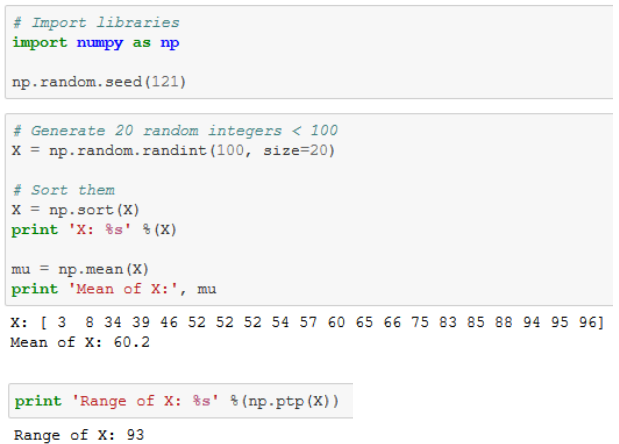


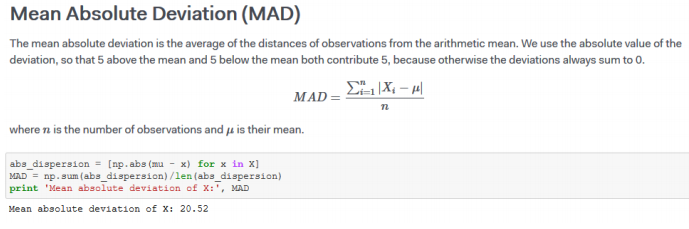


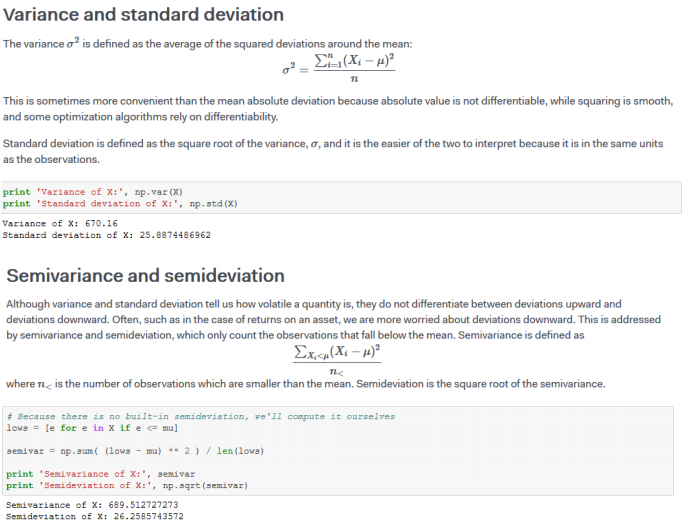


Lecture 7: Variance

The variance is a measure of dispersion, especially useful to analyze financial data and how the returns spread out historically. This lecture explores important concepts like range, mean absolute deviation, variance, standard deviation, semivariance and semideviation.







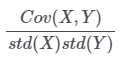
Lecture 8: Statistical moments

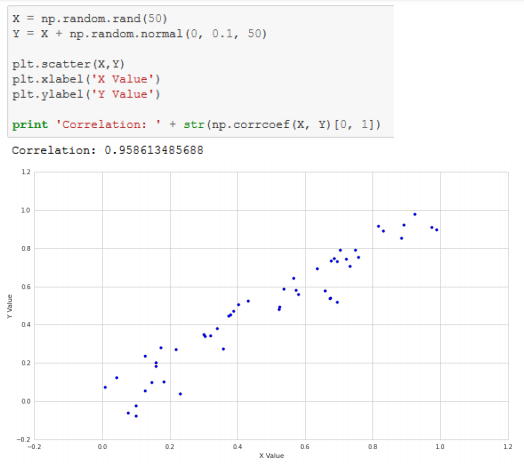
This lecture discusses how more moments than just mean and variance (the first two standard moments) can be used to describe data. To analyse that, we have to find the normal distribution of data and with the help of statistical moment, say if our data tend to stick the normal law, helping us to prevent how the futures data will come, comparing to the law.

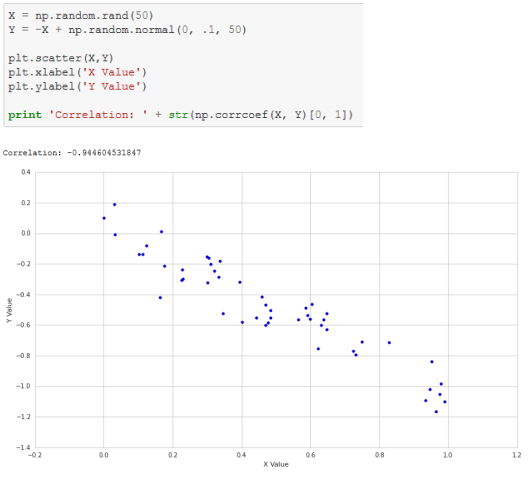
Lecture 9: Linear Correlation Analysis

The correlation coefficient measures the extent to which the relationship between two variables is linear. Its value is always between -1 and 1. A positive coefficient indicates that the variables are directly related, i.e. when one increases the other one also increases. A negative coefficient indicates that the variables are inversely related, so that when one increases the other decreases. The closer to 0 the correlation coefficient is, the weaker the relationship between the variables.

This correlation coefficient of two series X and Y is defined by:





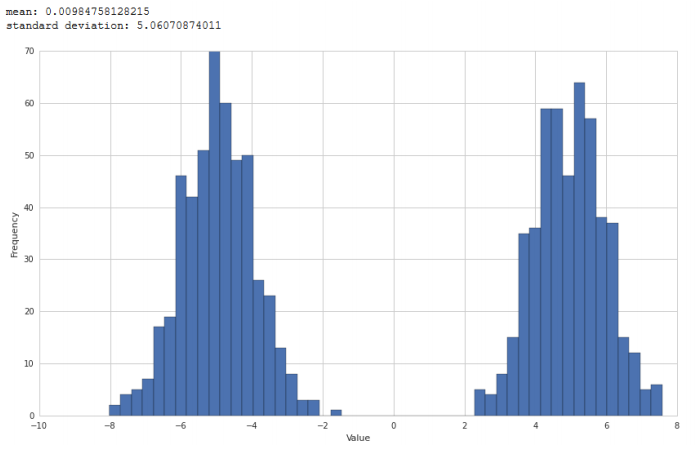


Correlation is a powerful technique, but as always in statistics, one should be careful not to interpret results where there are none.

Lecture 10: Instability of estimates

This lecture talks about the cautions when estimating parameters from distributions.

One basic example is the estimation of mean and standard deviation of a normal distribution when there are a few data.



We observe that for two graphics generated with random samples, the result is really different. That phenomena are called instability.

It is recommended to calculate also the volatility of a dataset.

Lecture 11: Random Variables

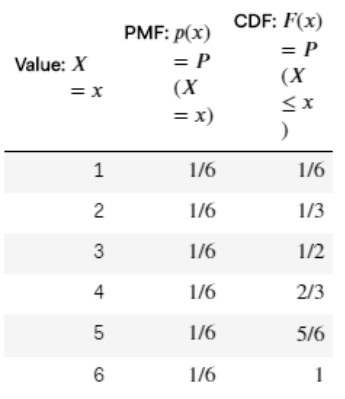
This lecture introduces the topic Random Variables. One of the most important concepts to understand random phenomena. Random variables are described in terms of probability distributions. That means the probability that each value of a random variable can come from.

There are two types: Discrete random variables and Continuous random variables. Financial assets are often expressed as moving according to deterministic and random patterns, with the random patterns being expressed with random variables. Each random variable follows a probability distribution, which assigns probabilities to all possible values of a random variable.

For a given random variable X, the probability that X is equal to a value x is P(X=x). For discrete random variables, p(x) = P(X=x), which corresponds to the probability mass function (PMF).

For continuous random variables, the probability density function is used. Another important concept is the cumulative distribution function (CDF) = P(X<=x), the probability that the random variable is less than or equal to a particular value. F(x) = P(X<=x).

To find F(x) in the discrete case, we can sum up the values of the PMF for all outcomes less than or equal to x. In the continuous case, F(x) is found by integrating the CMF.



Rolling a dice: example