**Note 1**

* In the present lesson, we will examine the statistical properties for a few examples of real world data, including test scores, earthquake magnitudes, and stock prices. Before we get into them, let me introduce my colleague, Dr. John Molina, who will also participate in the present lessons.
* [JJM]: Hello everyone, my name is John Molina. I've been following the course in the background but for this last section, Ryoichi, has asked me to present some examples outside the realm of Physics where you can use all you have learned about Stochastic Processes.

**Note 2**

* Let us begin by examining the distribution of test scores.
* Here we use the official data of the TOEIC Listening & Reading Test, which is the most popular evaluation for English communication skills in Japan. The details of the latest test were officially announced and are shown in the table here.
* [JJM]: What do you expect from this analysis?
* The distribution of test scores are often assumed to be Gaussian, which allows for the evaluation of individual scores using a single number such as the "deviation value", which measures how many standard deviations away from the mean any given student is.
* This information can be obtained by using the mean and standard deviation of the score distribution only.
* I want to examine the validity of this idea using the TOEIC test.

**Note 3**

* As usual, we begin by importing the necessary numerical and graphics libraries.
* To reduce typing, we have redefined the global plot parameters to use bigger font sizes

**Note 4**

* In the code example shown here, we define the data array with the score interval and count (as shown in the table).
* Using the officially announced mean and standard deviation of the score, we calculate the corresponding Gaussian distribution for this data.
* The exam data is then plotted as a bar chart in blue, and the Gaussian distribution is plotted as a red line.
* As you can see here, most of the score data is well represented by a Gaussian distribution.
* This shows that it is perfectly valid to represent the TOEIC score data using only the mean score and standard deviation, although noticeable differences do appear near the tails of the distributions.

**Note 5**

* Next, I want to show you the size distribution of earthquake magnitudes.
* The relationship between earthquake magnitude and frequency was first proposed by Charles Francis Richter and Beno Gutenberg in a paper published in 1956 as given in Eq.(J1).
* Here, N is the number of events having a magnitude larger than or equal to M.
* "a" is a normalization constant.
* "b" is a parameter referred to as the "b-value" which is usually close to 1, with some regional variations depending on the local subsurface structures.
* This relationship between event magnitude and frequency of occurrence is remarkably common, although the values of 'a' and 'b' may vary from region to region or over time.

**Note 6**

* First I will show you the size distribution of earthquakes in Japan, which is known to be located in one of the most seismically active areas on the planet.
* The left figure is a map of Japan showing the epicenters of earthquakes going back to 1965.
* The right figure represents the frequency-magnitude distribution of the earthquakes.
* The data periods vary depending on the magnitude ranges, but they have been correctly normalized based on the length of the data periods.
* Solid circles indicate the cumulative number of events N.
* Solid lines represent the approximate cumulative distribution as given by the G-R law with b-value = 0.87.

**Note 7**

* The second example for the size distribution of earthquakes is from the California region, which is also known to be very seismically active.
* The left figure shows a map of California with the seismicity data for Magnitude ≥3.25 from 1932 to 2008.
* The size of the symbols is proportional to the magnitude.
* The inner polygon with a dashed border indicates the area used in this article for the b-value statistics presented in the right figure.
* Earthquakes with M ≥6 are shown as stars and labeled by years.
* The right figure represents the earthquake statistics for the time period of 1932–1980.
* Here, the cumulative number of events N is plotted versus the magnitude M to determine the b-value, using a conservative cutoff value of Mc=4.2, in order to exclude missing aftershocks during the big sequences.
* In this case, we find a value of b≃1, in contrast to the value for Japan, which had b≃0.8.
* [JJM] What is the meaning of the difference in the b-values?
* There is still debate concerning the interpretation of the b-value, but it is believed to depend on the local subsurface structures.
* From a statistical point of view, we can say that a smaller b-value indicates a higher proportion of large earthquakes to small ones.

**Note 8**

* [RY] We will now examine the statistical properties of stock price variations in time. This part of the analyses is given by John. John, may I ask you to present the data analyses you have recently performed for the stock price variations?
* Yes, of course.
* If you are interested in investing in some stock or index, the instantaneous value P(t) is not what is important, what is important is how this value is changing in time.
* If you have ever seen a graph of a stock price over time you know that these changes are highly stochastic.
* It will not surprise you to learn that these changes are not Gaussian in nature.
* It is usual to work with the logarithmic change Gτ(t) defined in Eq. (J2), which gives the difference in the logarithmic price after some time τ.
* While the appearance of the logarithm might look strange, if the change in price is small enough, then Gτ(t) is approximately equal to the relative change in the price.
* You can easily prove this by using the Taylor expansion of the logarithm, as shown here.
* Thus, Gτ(t) gives you the relative return on a stock if you had bought at time t and then sold at time t+τ.
* Now we will analyze in detail how this return behaves using real-world stock data.

**Note 9**

**The lines below have to be added to the first line in In[13] in the command box.**

## You may install pandas\_datareader by typing the following command in command prompt

* To work with the financial data, we will use three new libraries: pandas, pandas\_datareader, and datetime
* If they are not yet installed on your system, you should install them using the conda installer
* Pandas can be considered as an extension of numpy that tries to make the data analysis more natural.
* Here we will mainly use it to import financial stock data from the web.

**Note 10**

* We start be defining some helper functions to help us in our analysis.
* 'logreturn' implements Eq.(J2) and computes the logarithmic return of a given data series.
* Note how we can write it as one line by simply 'slicing' the arrays in the proper manner.
* We calculate the return over time tau, by taking the elements from tau to the end, and subtracting the elements from the beginning until the end minus tau.
* The 'normalized' function simply normalizes any data series by its standard deviation. This will make it easier to compare different stocks.
* Finally, the 'pdf' function computes the normalized probability distribution function of a data series from its histogram.

**Note 11**

* Here we import the historical financial data of 3 different Japanese and American stocks, as well as the Nikkei 225 and S&P 500 market indexes.
* For this, we have used the 'pandas' DataReader function, which takes the stock 'id' or 'ticker', and the start and end time.
* Here we have tried to obtain data for the previous 30 years.
* To see what the 'pandas' DataFrame returned by the DataReader looks like, we display the 'tail' of the Toyota stock. This command prints out the last five elements in a format that is very easy to read.
* In particular, this is much easier to read than the unformatted output given by numpy arrays.
* The table shows that this 'DataFrame' is a 2D array. In particular, the rows and columns are all indexed with a unique label. In this case, the rows are labeled with the date, and the columns with the stock information.
* The benefit of 'pandas' is it allows us to reference the data by the labels, not just the row or column index number.
* For our purposes, we will be mainly interested in the 'Adj Close' price of the stocks. This is the stock price adjusted to take into account all the operations that go on after trading has stopped, and before it starts the next day.

**Note 12**

* Here we have plotted the price data for the American and Japanese stocks.
* If you look at the code, you see how easy it is to work with pandas arrays.
* The data is selected by using the name of the column, in this case 'Adj Close', and calling the plot command directly.
* To easily plot all the stocks, we make use of python's 'zip' iterator. It allows us to loop over an arbitrary number of lists simultaneously.
* In this case we define a list with the stock DataFrames, and a list with the labels. The loop variables 'stock' and 'lbl' will then simultaneously iterate over the elements of these lists.
* In the plots, you see the historical data for the different stocks. Two things are apparent.
* First, the range of the data is completely different of course. The Japanese stock are given in Yen, the American stock in dollars.
* Second, the price shows small scale (seemingly random) fluctuations, together with slower large-scale changes.
* How can we analyze such data? To begin, as we mentioned above, the price is not the most interesting quantity, but rather the return, or the changes in price over some time interval.

**Note 13**

* Let us then define the return of the stock from one trading day to the next, as defined in Eq.J2 (with τ=1).
* Given the time series data, this return can be easily computed using the 'logreturn' function we defined previously.
* In addition, we will normalize this data to have unit variance. This will allow us to easily compare different stocks, as well as to see how close to a Gaussian the distribution is.
* We have defined a 'computeReturn' function which adds this one-day logarithmic return as an additional column to the pandas DataFrames we just imported.
* You see that the frame in effect has a new column, labelled 'Return d1'. You also see that the return for the last day could not be computed, it is 'NaN' (not a number).
* One of the other benefits of using 'pandas' lies in its ability to properly handle NaN entries in your data.

**Note 14**

* Here we analyze the return of the Japanese stocks.
* On the right we have the time series data for G1, on the left the histogram of the data together with a Gaussian distribution.
* As you can easily see, the data shows very strong time variations and is clearly not Gaussian.
* In particular small returns and very large returns occur much more often than predicted by a Gaussian distribution.
* You can also notice very strong fluctuations in G1 around 2007-2010, corresponding to the global financial crisis.
* The lesson you should take for this is simple: if your model assumes that the returns are Gaussian, you cannot reasonably expect it to predict the behavior of real-world stock markets.
* Finally, even though the stock price of the different companies was considerably different, their returns show remarkably similar behavior...
* [RY] Is this just true for Japanese stock, will we see something similar for the American stocks as well?
* Let’s check now.

**Note 15**

* Here we perform the same analysis for the three well-known American companies: apple, microsoft, and hewlett-packard.
* Superficially, the same trends seem to hold as in the Japanese case. The returns are similar for the three stock and not Gaussian.
* When we look at the fluctuations in the return, we see that the data is mostly contained between -10 and 10.
* However, we see very large changes. One in particular stands out, corresponding to a return of -25 times the standard deviation!
* On this single day Apple's stock lost 50% of its value! A change this large is simply impossible under a Gaussian process.

**Note 16**

* To make a more qualitative comparison of the returns for the different stock and markets, we plot the probability distribution of the absolute value of G1(t).
* In addition to the six previous stock we have analyzed, here we also include data for two American and Japanese market indices, the S&P500 and the Nikkei 225.
* These indexes represent an average of the largest stocks traded in that particular market.
* Here, we are interested in the behavior at the tails, representing the large returns which are not captured by a Gaussian process.
* Remarkably, we see that all the data essentially collapse onto the same curve.
* By plotting on a log-log scale, we can easily identify a power law behavior, since the curves are all straight lines.
* The slope of this curve gives us the value of the exponent, which in this case seems to be around −3.
* Understanding how this universal behavior appears is currently an active area of research!
* If you develop a stochastic model for the stock market, the least it should do is reproduce this power-law behavior.