I want to analyse the stock market data and understand the statistical methods

Preamble

This case study is meant for business management students, who are willing to learn how the stock market data can be analysed. There is always a question on how one can analyse the stock market data using the most frequently used statistical methods (both descriptive and inferential) in a sequence, with a link between the outcomes of each of the methods. We make an attempt to provide an answer by analysing the BSE SENSEX (daily closing indices). We use graphical methods, summary statistics, categorical data analysis, probability (including specific models), and inferential statistics. We prefer to present the case in the form of a dialogue between the teacher and a group of students. The time horizon considered-from 2001-2016.

Discussion

Teacher: Good morning friends. I think you are eagerly waiting for today's discussion. Today's agenda is to discuss how stock market data can be analysed using the most frequently used statistical methods.

Student: Yes sir. We are ready for the class.

Teacher: Before I proceed to the stock market data, I wish to give you some idea on two aspects. The first with respect to type of study and the second with respect to scales of measurement.

The case writer(s) Srilakshminarayana G., Assistant Professor - Quantitave Methods and Operations, may be reached at lakshminarayana@sdmimd.ac.in

Basically two types of studies exists. The first one is the cross-sectional study and the second times series study. Cross-sectional data is independent of time and the data is drawn at one point of time. For example, if one is interested in studying the behavior of customers with respect to newly launched mobile phone, one needs to collect the data from the customers who have used the product. For this, the researcher collects the data from a selected group at one time point, say within one week. Similarly, if a HR manager is interested in knowing whether the employees are happy with the existing HR policies, then the manager selects a group of employees from the organization (assuming that the company has several branches) and gets information regarding their satisfaction levels.

Suppose that the marketing manager in the first example wishes to know how the customers perceive the product with the change in time, then the manager has to take the same set of customers at different time points. This study is a time series study. Similarly, the HR manager wishes to check the employee satisfaction at different time points, then she can consider a time series study.

There are four types of scaling, introduced by Stevens (1946). The first type is nominal, under which the measurements do not have any significance with respect to their order but differentiates the objects based on their names or coding. For example, the marketing manager wishes to know the employee names and employee-id, then the scale automatically is a nominal scale. Similarly, their gender, family background etc. then the scale is nominal. Whereas, if he wishes to know their educational background, societal status (high, middle, low) etc. the scale is an ordinal scale. The basic difference is, for an ordinal scale there exists natural ordering and also ordering the objects based on a criterion. That is, the objects can be ranked. Usually, when the variable under study is categorical, one uses either a nominal scale or ordinal scale. Note that, ordinal scale possesses the characteristics of a nominal scale and hence even though the scale is ordinal, it can be used as a nominal variable.

If the variable is a quantitative variable, the appropriate scales are either an interval scale or a ratio scale. I want you to look into the paper of Stevens (1946) to get more idea on these scales. Take this as an assignment. Now, let us move on to the actual discussion.

Student: Sir, Can you please give us the methods that we are going to use. Because, we read that stock market data is usually analysed using advanced statistical methods. But, we are being trained in using only the basic methods. How can we use these methods to analyse the stock market data?

Teacher: The discussion that we are going to have today will address this question. Before I proceed, I would like to clarify on certain aspects. You are right. The stock market data is usually analysed using advanced statistical methods. But, we can also analyze the same using the basic methods that I am teaching. The basic idea of using the statistical methods is to extract the hidden patterns in the data and also draw certain inferences related to the key characteristics of the situation, from where the data is drawn. I will show you in a sequence how the data can be analysed even using the basic methods. I consider the BSE SENSEX closing indices from the year 2001-2016. Then divide the same into three groups. The first from 2001-2007, second from 2008-2011, and the third from 2012-2016.

Student: Any specific reason for the division.

Teacher: You know that we had the crisis in the year 2008. So, I had divided the time horizon into three groups. This is one way of dividing the data. One can also have other divisions. But, should specify the basis for the division as, statistically it is very important (one can discuss why).

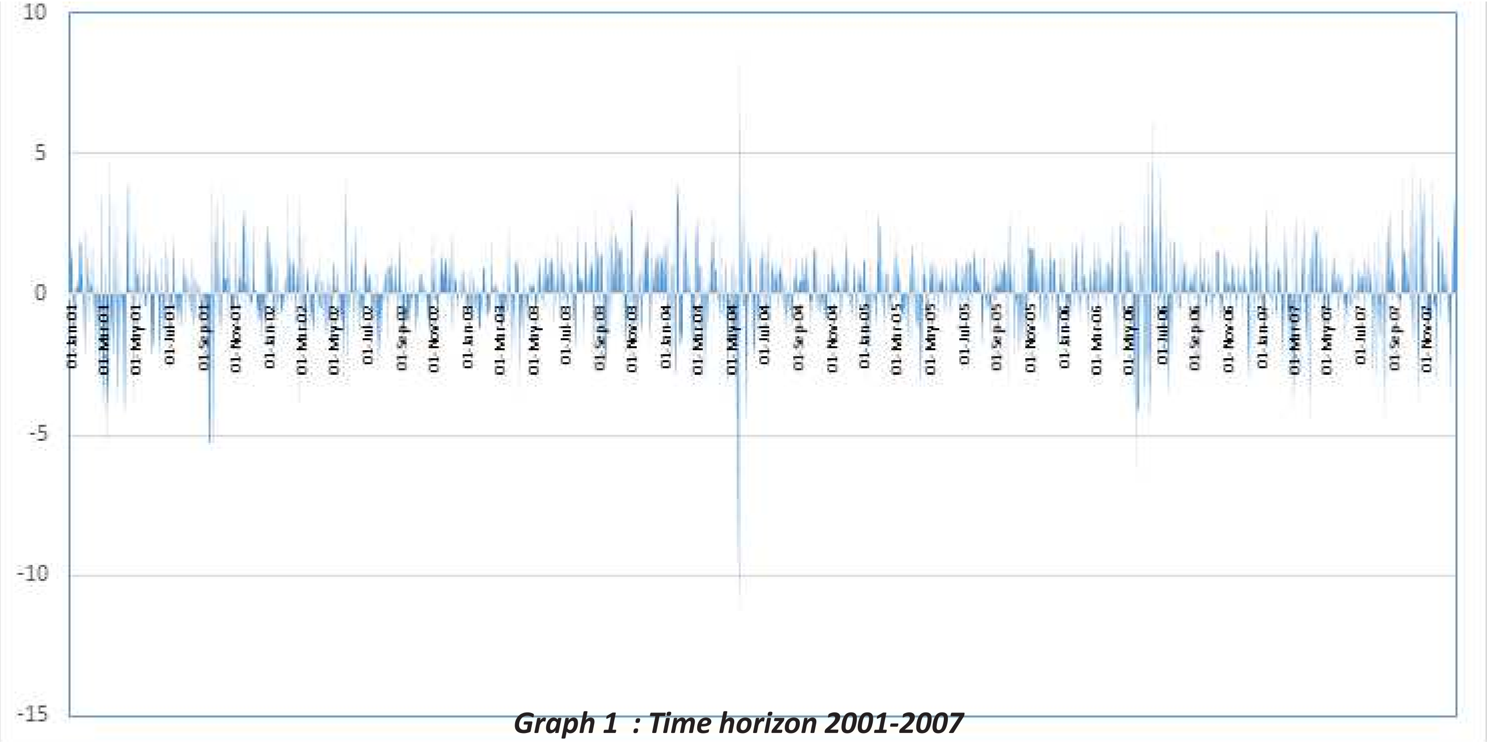
Graphical Presentation

Teacher: Can anyone let me know how to present the data graphically?

Student: Yes sir. If the data is a cross-sectional data, one can use a bar chart or a pie chart (nominal or ordinal scale), a histogram (ratio scale). If the data is a time series data, one can use a time plot.

Teacher: Yes, you are right. For the current data, I use a time plot and a comparison can be made for the three time horizons. The variable considered here is the "change", computed using the following formula.

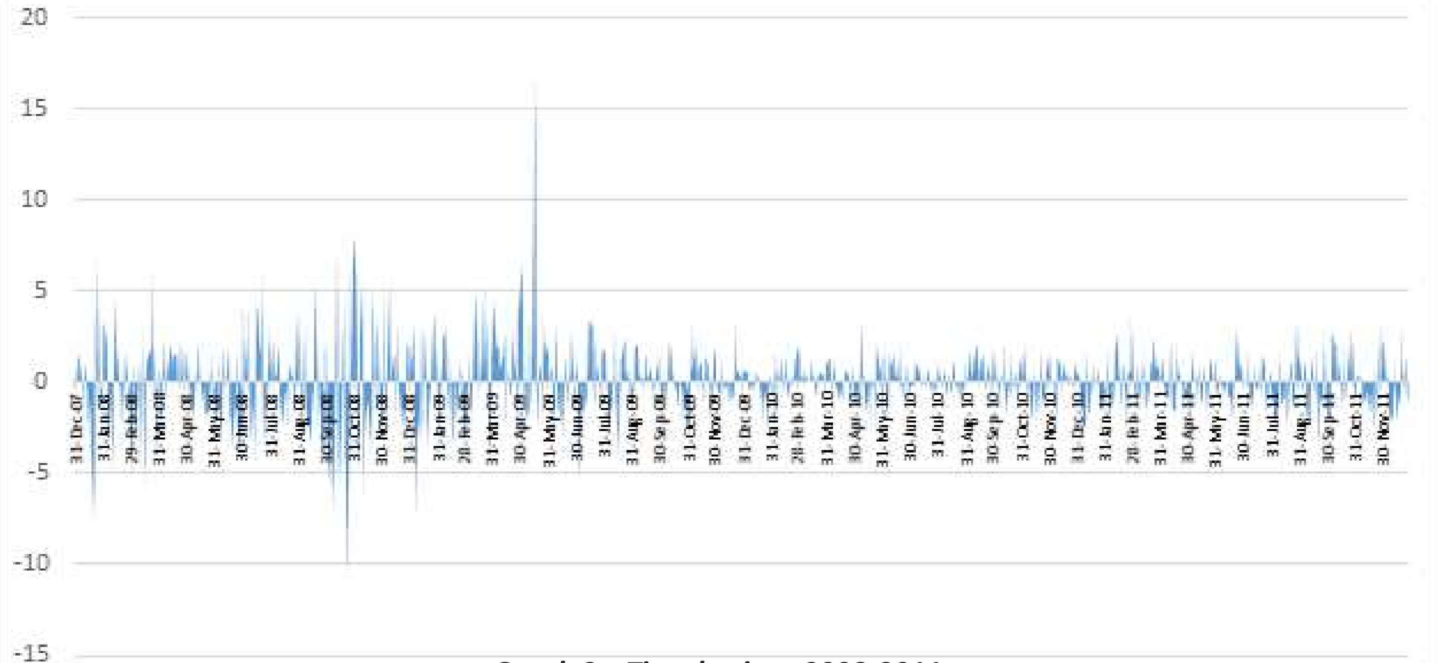
$$Change = \frac{Current - Past}{Past} * 100$$



Graph 1 : Time horizon 2001-2007

Source: From Researcher's' analysis

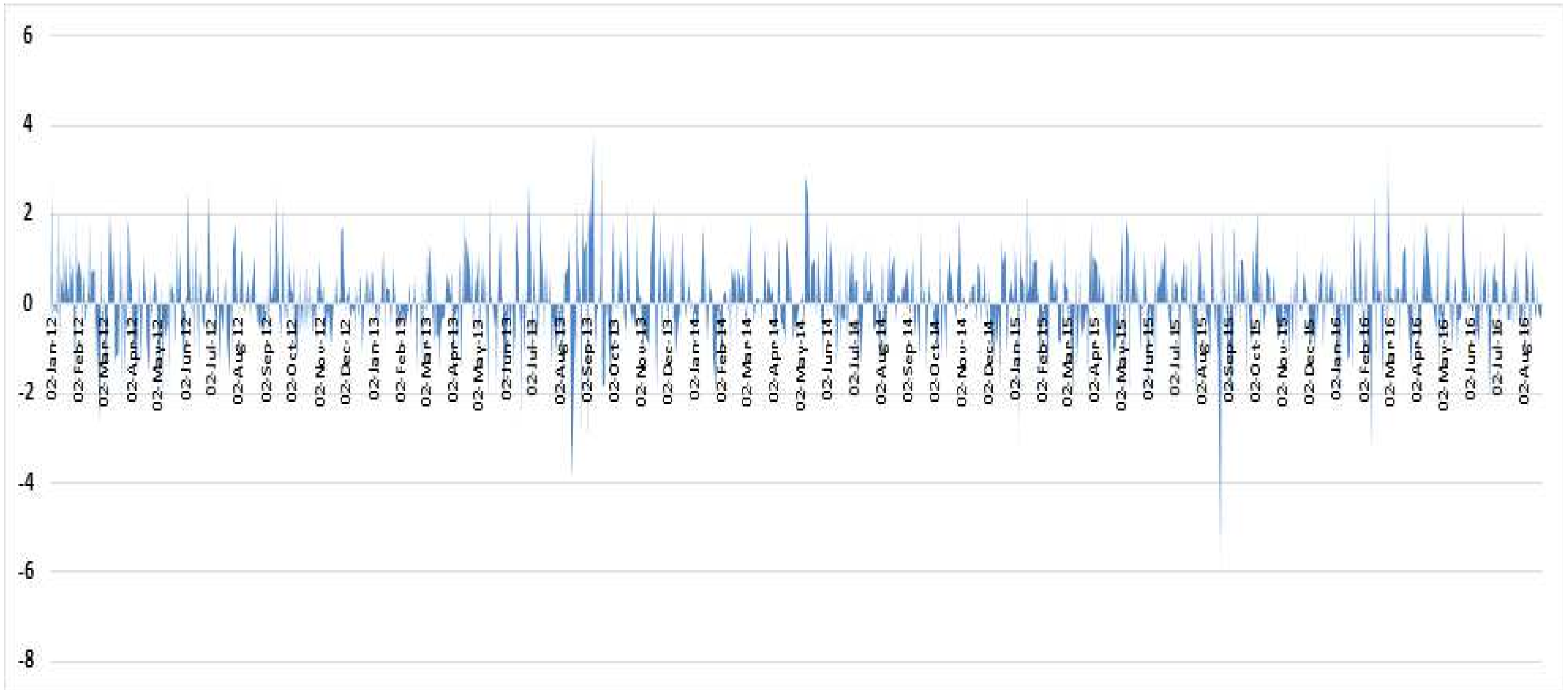
From the above graph, one can observe the change is oscillating and between May-04 and June-04, the change (either decrement or increment) is high. Similarly, in the months of May-06 and June-06 etc. To understand this better, one needs to calculate the descriptive statistics monthly. One has to take this as a motivation to compute the descriptive statistics monthly, to infer better. There is something beyond the statistical analysis. One can question on the reasons for the change in that particular month. That is, associating the events in that month to the change and this gives more insights with respect to the changes.



Graph 2 : Time horizon 2008-2011

Source: From Researcher’s’ analysis

From the above graph, one can note that during the months Jan-08, Oct-08, April-09 the change is high. A similar analysis by associating the change with the events (monthly) will give better understanding.



Graph 3 : Time horizon 2012-2016

Source: From Researcher's' analysis

In this case during the months Sep-13, Nov-13, May-14, and Nov-15 the change is high. Further analysis will give better understanding.

Student: I now understand why we have to compute the descriptive statistics. One cannot get a complete picture only based on graphical presentation. But, graphical presentation can be taken as a motivation for further analysis.

Teacher: Yes. In this case, in few months the change is high and in another the change is low, the change is fluctuating. Only from the graph one cannot get a summary of the change. That is, one cannot give the average change and the corresponding fluctuation. Also, the descriptive statistics will help one to find the month in which the change is consistent. Let us now proceed to the computation of the descriptive statistics.

One important aspect you have to note is that **doing analysis is one aspect and presenting the same is another aspect**. If you try to copy and paste the analysis, “**as it is”,** you get in SPSS (or any other tool), it only consumes pages and it may not help you much. The following tables will help you in understanding how one can present the analysis without losing the essence as well as without consuming space.

Descriptive statistics

Teacher: Under this I will be presenting the descriptive statistics month-wise for each of the time horizons*.

2001-2007

Table 1

Year -2001-2007

Month	Mean							CV						
	01	02	03	04	05	06	07	01	02	03	04	05	06	07
Jan														
Feb														
Mar														
Apr														
May														
June														
July														
Aug														
Sep														
Oct														
Nov														
Dec														

Source: From Researcher's' analysis

*The tables have been left blank as the case focuses more on how to construct the analysis for the stock market data rather than analysing the stock market data.

From the above table, one can note that in the year…….., the change is consistent as compared to other months in that year. When one links this with the graphs presented earlier, one can understand the changes better. The mean value will give the magnitude of the change and the CV can be used to calculate the deviation using the formula **SD=CV/ Mean.** Similar interpretation can be given even for other time horizons.

*Note that, we present the above analysis only for one time horizon (2001-07) and we suggest the readers to construct for other time horizons on similar lines.

Student: Sir, I appreciate what you have presented. This will help us in presenting the analysis in a precise manner. We also have understood how one can link the graphical analysis with the summary statistics and also the motivation for computing the descriptive statistics.

Teacher: The motivation for one to compute the descriptive statistics, is to know the average change with the fluctuation, consistency. Also, these descriptive statistics can be further used in testing the hypotheses that we can construct on parameters associated with the change variable. Another important part of computation of the descriptive statistics is to understand the shape of the change variable along with its behaviour. In a simple terms, is it normal or no-normal?

Student: Why should one look at the shape or normality?

Teacher: Let me explain in brief the importance of normality. Normal distribution is one continuous distribution that has got prominent place in data analysis. If the behaviour is normal, one can use the advantages of the normal distribution in computation of the probabilities as well as in testing the hypothesis **(here the teacher can explain the importance in detail).** In early stages when researchers have started analysing the stock market data, they have used normal probability model to model the change variable and later have understood that normal model is no longer an appropriate model. This made researchers to search for alternate models that can be used to study the behaviour of the change variable. Hence, if one needs to proceed further in the analysis, then normality has to be tested. This has to be done immediately after the descriptive statistics have been computed. This can be done in two ways. One, by observing the Skewness and Kurtosis of the change variable and second, by testing the hypothesis using the standard testing procedures. The following tables give the results of the analysis.

Student: I could understand the importance of Skewness and kurtosis better now. I request you to explain in brief the procedure that can be used to test the significance of normality.

Teacher: There are several tests that can be used to test the normality. I discuss here only one test. For other tests, you can refer to Thorde (2002), Sheskin (2004).

The test that I am going to discuss is the Kolmogorov-Smirnov test (K-S test). Under this the null hypothesis is that "the population is normal" against alternative "the population is non-normal". We fix the level of significance at 5%. Under this test, we first compute the probabilities of normal distribution and compare them with the probabilities computed from the empirical distribution (constructed from the data points drawn at random). The test statistic is the maximum of the distances between the two cumulative probability distributions (difference between empirical and normal). This value is compared with the critical values and then a decision is taken regarding the normality. Other details can be found in Sheskin (2004).

Table 2

Year -2001-2007

Month	**Skewness**							**Kurtosis**						
	01	**02**	**03**	**04**	**05**	**06**	**07**	**01**	**02**	**03**	**04**	**05**	**06**	**07**
Jan														
Feb														
Mar														
Apr														
May														
June														
July														
Aug														
Sep														
Oct														
Nov														
Dec														

Source: From Researcher's' analysis

Table 3
Year -2001-2007

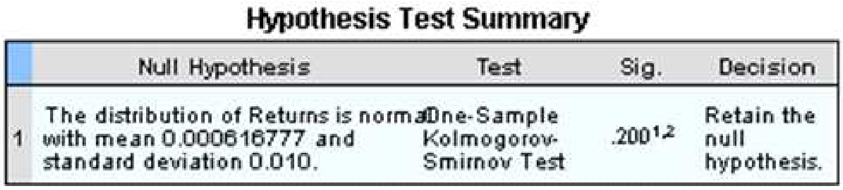
Month	Testing of normality using K-S test						
	01	02	03	04	05	06	07
Jan							
Feb							
Mar							
Apr							
May							
June							
July							
Aug							
Sep							
Oct							
Nov							
Dec							

Source: From Researcher's' analysis

From the above tables, one can note that __________,* the data follows a normal pattern. This is also evident from the values of Skewness and Kurtosis. Note that, Skewness gives the stretch of the tails and Kurtosis gives the measure of the height of the curve.

I will present the test for the year 2010.

Table 4
K-S test for normality



Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Returns is normal with mean 0.000616777 and standard deviation 0.010.	One-Sample Kolmogorov-Smirnov Test	.200^{1,2}	Retain the null hypothesis.

* The analyst has to provide the value. The blank is given purposefully.

One-Sample Kolmogorov-Smirnov Test

		Returns
N		251
Normal Parameters	Mean	0.00061677681
	Std. Deviation	0.010077621403
Most Extreme Differences	Absolute	0.051
	Positive	0.041
	Negative	-0.051
Test Statistic		0.051
Asymp. Sig. (2-tailed)		0.200

Source: From Researcher’s’ analysis

Since the p-value is more than the level of significance, one can conclude that the returns variable for the year 2010 follows a normal pattern. A Similar analysis can be presented in the form of tables for other time horizons and interpretations can be made.

Student: Sir, you have mentioned in the previous class that when normality is satisfied, we use parametric procedures and if not we use non-parametric procedures. Can you please explain the difference between both?

Teacher: Definitely. The major difference between the two is with respect to the distributional assumption. Non-parametric statistical methods do not assume any probability distribution for the variables being assessed. That is, we do not assume any model before we test a hypothesis. Whereas, in case of a parametric procedure, we assume a model before we test a hypothesis regarding the parameter of a population.

This is one reason why I always suggest that normality has to be tested before using a parametric procedure.

Student: I note this point seriously as this appears in the research papers, reports that I have read earlier. Now, I understand the motivation for testing normality.

Teacher: Yes. Unless one tests this, standard parametric procedures cannot be used for analysis. Now, the story starts from here. Most of the researchers study the consistency or normality and stop at this point. But, I would suggest to go ahead with questioning on what made the changes consistent in a particular month or year. Similarly, what made the change variable follow a normal pattern. This leads to associating the events that happen during that time period.

Student: You mean to say that we need to test for the association between the change variable and the events that takes place in that month. But, how do we do it. The data is measured on a ratio scale and quantitative analysis is used to understand the behavior. As I recall, the association are possible only if we have the categorical data. Can you please let us know how this is possible in this case?

Teacher: Definitely. This is an interesting analysis. We first learn how one can construct contingency tables for the change variable. We adopt the following steps for the same:

- Compute the descriptive statistics and test for normality for each of the years separately.
- Construct the one-sigma limits for each of the year and find those days on which the daily change is within the sigma limits, and outside the limits. Label the days on which the changes are within the limits as **W**, below the lower limit as **L** and above the upper limit as **H**. Categorize the changes as increment **(Neg)** and decrement **(POS)** based on the value. Construct a bivariate frequency table for the change variable, labelled based on one-sigma limits and labelled based on increment (decrement).
- Repeat the same for each year and note the frequencies. Now identify the news items for each of the days and construct the multivariate contingency tables for each of the years. Compute the probabilities for each year for the corresponding events and identify the events that have high probabilities.

Student: Can this procedure be adopted for other situations? I also noted that this procedure can be used to compute the probabilities in later stages.

Teacher: Yes. It can be used even for other situation as well and for computation of probabilities as well. But, needs to be constructed carefully. Now, following this procedure one can get the contingency tables as well as know the events that are very critical on a given month (year).

Student: How do you find an event as a critical event?

Teacher: Once the contingency tables are constructed using the positive or negative change and the sigma-limits, associate the events in each of the combinations. Then look at the number of times the event has recurred and then classify the event as critical.

Student: There will be many events that happen on a given day. How do we consider the events?

Teacher: You are right. There will be several events and it will be difficult one to identify one event as a critical event. One can do the following and this is one way of doing the same. Assume that the events are related to India. Classify the events as Global (G), Political (P), Business (B), and Others (O). Here, Global events may include that events that are related to the global changes, political and others. Political events mean those related to Indian context, Business means those events related to Indian context, and Others include again related to Indian context only. All others not related to Indian context can be taken as Global. Now, observe the events and categorize them into one of these events and then count the recurrence. Those events that have the most recurrence can be called as critical events. The following tables give the above mentioned process. The table is presented for only one year and for other years the same can be adopted. The following table gives you the cross-tabulation of the change and the sigma limits.

Table 5

Cross-tabulation of change and the sigma limits-year 2010

			sigma_limts_categories			**Total**
			H	**L**	**W**	
Ret_ categories	Neg	Count	0	37	78	115
		% of Total	0.0%	14.7%	31.1%	45.8%
	POS	Count	37	0	99	136
		% of Total	14.7%	0.0%	39.4%	54.2%
Total		Count	37	37	177	251
		% of Total	14.7%	14.7%	70.5%	100.0 %

Source: From Researcher’s’ analysis

The above table gives number of times there is a positive change and number of times there is a negative change and also, number of times they fall under the sigma limits and outside the sigma limits. The following table gives the cross-tabulation with respect to events.

Table 6

Summary table for events ranked-1 for the year 2010

sigma_limts_categories				**Event**				**Total**
				B	**G**	**O**	**P**	
H	Ret_categories	POS	Count	10	8	10	9	37
			% of Total	27.0%	21.6%	27.0%	24.3%	100.0%
L	Ret_categories	Neg	Count	11	10	4	12	37
			% of Total	29.7%	27.0%	10.8%	32.4%	100.0%
W	Ret_categories	Neg	Count	17	22	16	23	78
			% of Total	9.6%	12.4%	9.0%	13.0%	44.1%
		POS	Count	22	31	19	27	99
			% of Total	12.4%	17.5%	10.7%	15.3%	55.9%
	Total		Count	39	53	35	50	177
			% of Total	22.0%	29.9%	19.8%	28.2%	100.0%

Source: From researcher’s’ analysis

From the above table, one can note the following:

1. Given that the returns fall within the limits and positive, the chances that it falls under global events is 0.175, it falls under political events is 0.153. That is, the chances are high for **global events relatively**.

2. Given that the returns fall within the limits and negative, the chances that it falls under global events is 0.124, it falls under political events is 0.130. Here, the chances are high for **political events.**

3. Given that the returns fall below the lower sigma limit and negative, the chances that it falls under political events is 0.324, business events is 0.297, global events is 0.270. In this case, the chances are high for **political events.**

4. Given that the returns fall above the upper sigma limit and positive, the chances that it falls under political events is 0.243, business events is 0.27, other events is 0.270, and global events is 0.216. Here, **business events and other events have equal chances**.

From the above one can conclude that, in the year 2010 when the returns are within the limits, **global** (positive) and **political** events (negative) relatively have higher chances of association with changes. Similarly, when the returns are below the lower limit, **political events** have higher chances of association with changes and when the returns above the upper limit, **business and other events** have equal chances of association with changes. Using this, one can question on the significance of the association as well as the effect of these events on the changes. Note that, the above table cannot be used for cause and effect. It can only be used for associations. Continuing the same process for other years, one can note the repetition of these events in association with the changes variable and this ultimately hep one to summarize on the association of the events with the returns.

Student: Excellent. I understood now how one can categorize a continuous random variable.

Teacher: You need to be cautious while doing this. There may be several issues of events overlapping. Now, one can proceed further in the analysis by testing the associations between the events. The following table gives the same. In order to test the association, we use Chi-square test for associations.

Table 7

Year-2010

Chi-Square Tests				
sigma_limts_categories		**Value**	**df**	**Asymptotic Significance (2-sided)**
	Pearson Chi-Square	.873[b]	3	.832
	Likelihood Ratio	.876	3	.831
	N of Valid Cases	251		
b. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 22.45.				
d. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 15.42.				

Source: From Researcher's' analysis

From the above, one can note that the associations are not significant. That is, there is no significant association between the changes and the events.

Student: Sir, if the association is significant, can we say that these events have caused the change?

Teacher: No. Chi-square analysis cannot be used for studying the cause and effect. It can only be used for studying the associations. That is, one can only say that both the variables are moving in same direction. In other words, one can say that the increment (decrement) in the changes variables goes along with the happening of the events. But, one cannot say that because of these events the changes have taken place.

Now, let us summarize what we have done till now. We first have looked at graphical presentation of the change variable. After identifying the months that have fluctuation with respect to change variables, we have proceeded to computation of descriptive statistics. Looking at the consistency levels, we have proceeded to test the normality. Then, we have questioned on reasons for normality and consistency. To answer this we have moved onto categorical data analysis to associate the events to the movements of the change variable. Chi-Square test for association has been used to test the association.

*Note that, this is one sequence that one can look at, in analysing the stock market data. Other ways of analysis can be thought of by the students and the teachers.

Student: Sir, You have mentioned that you are going to explain the probability models in analysing the stock market data. Can you please explain the same?

Teacher: Definitely, I will explain how probability models can be used in analysing the stock market data. Let me do the same using the SENSEX data.

The probability models can be categorized into two categories (this is sufficient for a management student): Discrete and continuous. Under discrete, I will be covering the most frequently used (Bernoulli, Binomial and the Poisson) and under continuous distribution, only the normal distribution. One can think of other distributions and their applications, on lines similar to those explained in this case.

The first one which is a simple random variable is Bernoulli random variable and the associated distribution is called as the Bernoulli distribution. Under this the random variable takes only two values. In the case of SENSEX data considered, the random variable is either positive change or a negative change. That is, consider any day's change and compare it with the previous day's (next day's) change. Then one has two possibilities: positive (X=1) or negative (X=0). In such cases, the chances of change can be studied using Bernoulli random variable. For example, we consider the change variable for the year 2016. The following table gives few data points

Table 8
Bernoulli random variable

Date	Closing	Change	Bernoulli random variable (X)
01-Aug-16	28003.12	-0.17375	X=0
02-Aug-16	27981.71	-0.07646	X=0
03-Aug-16	27697.51	-1.01566	X=0
04-Aug-16	27714.37	0.060872	X=1
05-Aug-16	28078.35	1.313326	X=1
08-Aug-16	28182.57	0.371176	X=1
09-Aug-16	28085.16	-0.34564	X=0
10-Aug-16	27774.88	-1.10478	X=0
11-Aug-16	27859.6	0.305024	X=1
12-Aug-16	28152.4	1.050984	X=1
16-Aug-16	28064.61	-0.31184	X=0
17-Aug-16	28005.37	-0.21108	X=0
18-Aug-16	28123.44	0.421598	X=1
19-Aug-16	28077	-0.16513	X=0
22-Aug-16	27985.54	-0.32575	X=0
23-Aug-16	27990.21	0.016687	X=1

Source: BSE data

The table above gives one an idea about a Bernoulli random variable. For the same case, if one looks at number of positive changes or negative changes, then one has to use a binomial distribution to model the corresponding random variable. The following table gives the construction. Suppose that we consider the year 2016 and the corresponding change data indices. Let Y be the total number of positive changes. The following construction gives the details.

Table 9

Construction of Probability

	Count of P/N	Probability
N	73	0.45625
Y	87	0.54375
Grand Total	160	1

Source: From researcher's calculation

From the above table, one can note that the probability of a positive change is p=0.54375. The binomial probabilities are calculated as follows. Suppose that a sample of n=20 days changes have been drawn and the probabilities of positive changes are calculated.

Table 10

Binomial Probabilities

X=x	P(X=x)	Cumulative
0	1.52777E-07	1.52777E-07
1	3.64153E-06	3.79431E-06
2	4.12291E-05	4.50234E-05
3	0.000294816	0.00033984
4	0.001493265	0.001833105
5	0.005694863	0.007527968
6	0.016967571	0.024495539
7	0.040443251	0.06493879
8	0.078324173	0.143262963
9	0.12446033	0.267723293
10	0.163162378	0.43088567
11	0.176776175	0.607661846
12	0.158008841	0.765670687
13	0.11588425	0.881554937
14	0.069054314	0.950609251
15	0.032919043	0.983528294
16	0.012260089	0.995788382
17	0.003437962	0.999226344
18	0.000682883	0.999909227
19	8.56681E-05	0.999994895
20	5.10488E-06	1

Source: From researcher's calculation

If one wishes to find the probabilities, the above table can be used. For example, the probability of at least two positive changes is 0.999 and similarly other probabilities can be calculated. This construction can be used for other years as well to calculate the respective probabilities.

Student: This is fine. How can one construct for a Poisson random variable? I remember you saying that it is used for rare events.

Teacher: Yes. You are right. I will give you a hint and I want you to try. Suppose that one wishes to study the behaviour of a small jump or a big jump from the previous change. Then, one can use a Poisson distribution as this is case of rare event. Similarly, one can think of applying normal distribution to compute the probabilities for different events. We present one way of constructing the probabilities.

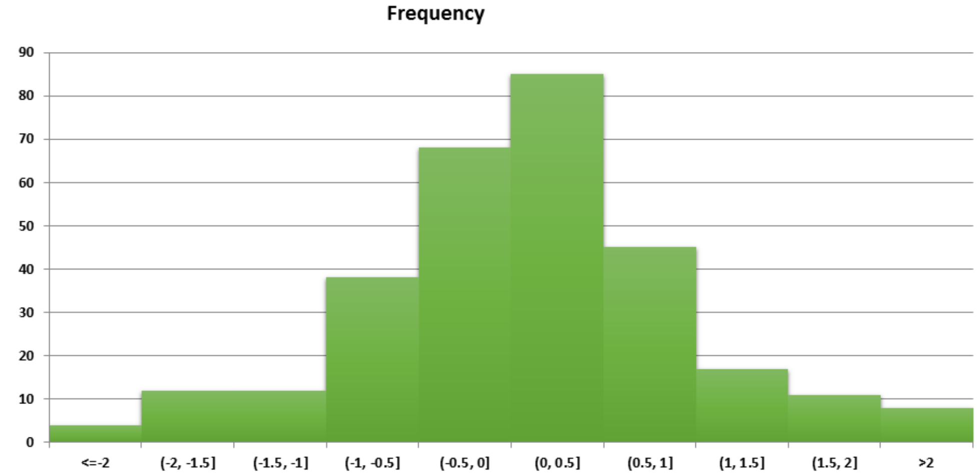
Suppose a random sample of 300 is collected from the time horizon 2012-2016. I will explain how normal distribution can be used for this sample. We first consider the sample points and construct the class intervals. Now, for the constructed class intervals, one can compute the probabilities using normal distribution. The following table gives the frequency distribution of the sample considered.

Table 11

Frequency distribution

Interval	Freq.
<=-2	4
(-2, -1.5]	12
(-1.5, -1]	12
(-1, -0.5]	38
(-0.5, 0]	68
(0, 0.5]	85
(0.5, 1]	45
(1, 1.5]	17
(1.5, 2]	11
>2	8

Source: From researcher's calculation



Graph 4 : Histogram

Source: From researcher's calculation

For the class intervals constructed, the following table gives the probabilities.

Table 12

Normal Probabilities

Interval	LL	UL	Probability
<=-2		-2	0.009219
(-2, -1.5]	-2	-1.5	0.027393
(-1.5, -1]	-1.5	-1	0.073459
(-1, -0.5]	-1	-0.5	0.144279
(-0.5, 0]	-0.5	0	0.207578
(0, 0.5]	0	0.5	0.21879
(0.5, 1]	0.5	1	0.168946
(1, 1.5]	1	1.5	0.095568
(1.5, 2]	1.5	2	0.039596
>2	2		0.015172

Source: From researcher's calculation

The following table gives the formula used to compute the above probabilities

Table 13
Formula

Probability
=NORM.DIST(R15,L5,L9,TRUE)
=NORM.DIST(R16,L$5,$L$9,TRUE)-NORM.DIST(Q16,$L$5,$L$9,TRUE)
=NORM.DIST(R17,L$5,$L$9,TRUE)-NORM.DIST(Q17,$L$5,$L$9,TRUE)
=NORM.DIST(R18,L$5,$L$9,TRUE)-NORM.DIST(Q18,$L$5,$L$9,TRUE)
=NORM.DIST(R19,L$5,$L$9,TRUE)-NORM.DIST(Q19,$L$5,$L$9,TRUE)
=NORM.DIST(R20,L$5,$L$9,TRUE)-NORM.DIST(Q20,$L$5,$L$9,TRUE)
=NORM.DIST(R21,L$5,$L$9,TRUE)-NORM.DIST(Q21,$L$5,$L$9,TRUE)
=NORM.DIST(R22,L$5,$L$9,TRUE)-NORM.DIST(Q22,$L$5,$L$9,TRUE)
=NORM.DIST(R23,L$5,$L$9,TRUE)-NORM.DIST(Q23,$L$5,$L$9,TRUE)
=1-NORM.DIST(Q24,L$5,$L$9,TRUE)

Source: From researcher's calculation

Student: So, We need to construct the class intervals first and then compute the probabilities. But, I have one doubt. How can one use the frequencies that we get under the frequency distribution?

Teacher: These frequencies along with the probabilities calculated using the normal distribution can be used to test whether the data follows a normal pattern or not. For this, one can use Chi-Square test for goodness of fit. I will show you for the above data. We first take the probabilities computed and multiply them with the total frequency. This will give the expected frequency for each of the class intervals. The following table gives the same.

Table 14

Chi-Square test for goodness of fit

Interval	LL	UL	Probability	Expected Frequencies (E)	Observed frequencies (O)	(O-E)^2/E
<=-2		-2	0.009218533	2.765559897	4	0.551007
(-2, -1.5]	-2	-1.5	0.02739318	8.217953982	12	1.740564
(-1.5, -1]	-1.5	-1	0.073459482	22.03784465	12	4.572059
(-1, -0.5]	-1	-0.5	0.144278701	43.2836103	38	0.644968
(-0.5, 0]	-0.5	0	0.207578126	62.27343782	68	0.526605
(0, 0.5]	0	0.5	0.218790284	65.63708512	85	5.712052
(0.5, 1]	0.5	1	0.16894592	50.6837759	45	0.63739
(1, 1.5]	1	1.5	0.095567571	28.67027125	17	4.750399
(1.5, 2]	1.5	2	0.039596032	11.87880972	11	0.065015
>2	2		0.015172171	4.551651371	8	2.612482
				300		Chi-square value=21.81254

Source: From researcher's calculation

The Chi-square value calculated in the last column of the above table can be used to test the hypothesis. I leave that for you to check how the hypothesis can be tested. The hint is to compare the Chi-Square value with the Chi-Square critical value and take the decision.

Now, let us proceed further for the discussion on hypothesis testing.

Student: Can you let us know the methods that you are going to teach us?

Teacher: t-test (one population, two populations), and ANOVA.

T-test one population

In order to explain this, one needs to consider one population and construct a hypothesis with respect to the parameter (average) and then test the same using the testing procedure. For the SENSEX data considered, we construct the hypothesis and test the same. The following table gives the same. Note that we are presenting the summary of the procedure and you need to look the complete procedure discussed in the class.

Table 15
Testing of hypothesis

2001-2007	2008-2011	2012-2016
H0:Mean Returns=1 p-value=0.0001 Reject H0	H0:Mean Returns=-3 p-value=0.0001 Reject H0	H0:Mean Returns=-2 p-value=0.0001 Reject H0

Source: From researcher's calculation

T-test for two independent populations

Under this, one can test the null hypothesis that there is no significant difference between any two time horizons. The following table gives the results.

Table 16
Testing of hypothesis

2001-2007 and 2008-2011	2008-2011 and 2012-2016	2012-2016 and 2001-2007
H0:Mean returns1 = Mean returns2	H0:Mean Returns2 = Mean Returns3	H0: Mean Returns3 = Mean Returns1
p-value=0.0001 Reject H0	p-value=0.0001 Reject H0	p-value=0.0001 Reject H0

Source: From researcher's calculation

ANOVA

Null Hypothesis H0: There is no significant difference between the three time horizons with respect to average returns.

Alternative Hypothesis H1: There is no significant difference between the three time horizons with respect to average returns.

Table 17
Testing of hypothesis

2001-2007, 2008-2011, and 2012-2016
H0:Mean returns1=Mean returns2= Mean Returns3
p-value=0.0001; Reject H0

Source: From researcher's calculation

I leave the discussion here and I want you to think on the above analysis and further actions based on the analysis. Good Luck.

Student: Thank you sir. You have given us an idea on how basic statistical methods can be used on stock market data.

Acknowledgement: The author thank the reviewers for their comments, which helped him to improve the presentation of the case.

References

Sheskin, D., J. (2004): Handbook of parametric and nonparametric statistical procedures. 3rd edition. Chapman and Hall /CRC.

Stevens, S., S. (1946): On the theory of scales of measurement. Vol.103, No.2684. Science.

Thorde H., C., Jr (2002): Testing for Normality. Marcel Dekker Inc.