Foundation Statistics for Data Science Beginners

What is Statistics

Statistics is the science concerned with developing and studying methods for collecting, analysing, interpreting and presenting data.

- Collection of data is done most of the time on sampling basis as it may be tedious or expensive to collect from all ie whole population. Collection of data may be from Survey, gathering of data generated by devices (measurement, surveillance video streams, stock market feeds, weather monitoring, sensor feeds etc), from corporate business application systems, from social media sites, etc.

- Analysing data involves, organizing the data (grouping, ordering, tabulating) and applying statistics to describe the characteristics of the data.

- Interpreting the data involves applying the statistics to make inferences about the information contained in the data collected.

- Presenting the data involves drawing conclusions, recommendation for business decision and presenting the information derived in a meaningful way to the people of interest. The presentation may involve in representing the data in various graphs to show the characteristics of the data, like range, spread, correlation, trend etc to support the conclusions derived.

Statistics & Data Scientists

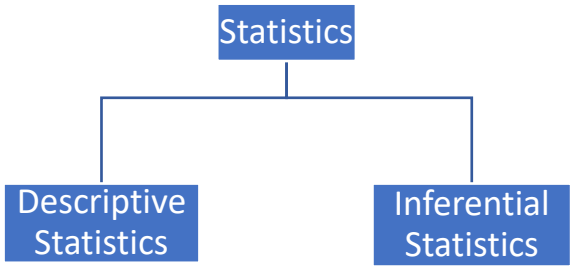
The data scientist skill was identified as a growing need in 2009 by Hal Varian, UC Berkeley Professor of Information science and the chief economist, in the Mckency&Company article.

"The ability to take data — to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it — that's going to be a hugely important skill in the next decades."- Hal Varian, chief economist at Google and UC Berkeley professor of information sciences, business, and economics

Understanding, process, derive value from data is achieved by applying the statistics. That is where Statistics becomes the base for Data Scientists.

Categories of Statistics

There are two types of Statistics, namely Descriptive Statistics & Inferential Statistics



Descriptive Statistics

Descriptive Statistics is about applying statistical and graphical techniques to presenting, organising and summarising the data being studied.

Inferential Statistics

Inferential Statistics is about drawing conclusions about the population based on the observations made by applying statistical techniques on a sample data set

Some Basic Statistical Terms

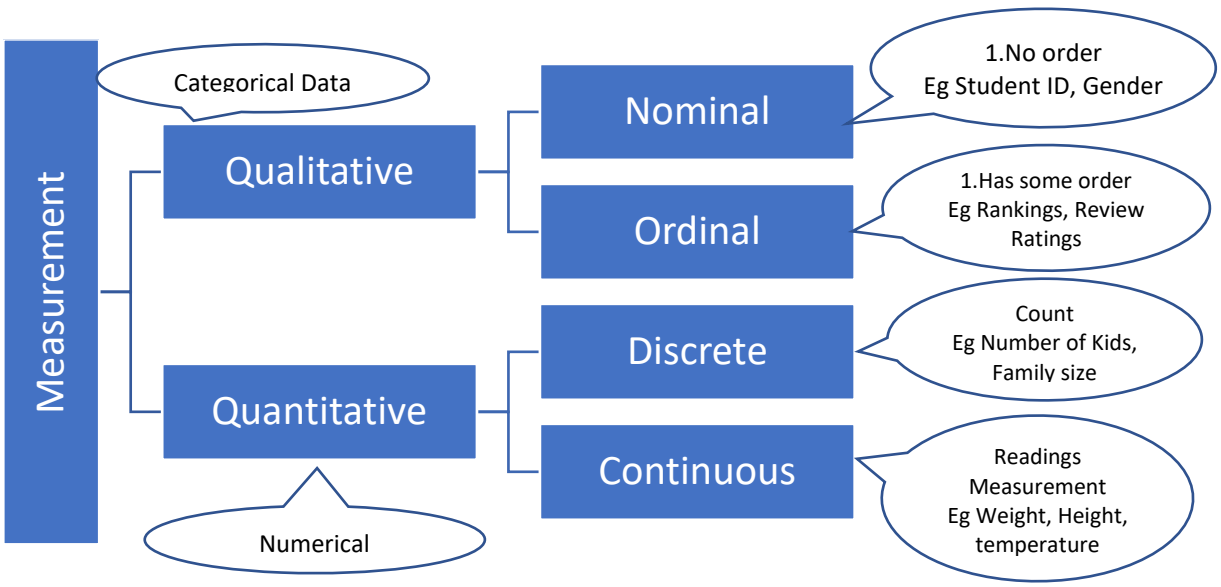
Terms	Meaning
Population	100% of all concerned entities about whom the data is being gathered
Sample	Subset of the Population from whom the data is being collected
Inference	Educated guess which is supported by the statistics derived from the samples
Descriptive Statistics	Use of statistical and graphical techniques to present information about the data set being studied
Inferential Statistics	The practice of drawing conclusions about a population by using the statistics calculated on a sample
Mean	The arithmetic average of the data set
Median	The middle value of the data set ordered in ascending order
Mode	The most repeated value among the data set
Range	The difference between the largest and the smallest number in the data set
Deviation	The difference between the value and the mean of the data set
Variance	The average of the squared deviation from the mean of the data set
Standard deviation	The square root of the variance of the data set

Measurement

Measurement is the process of assigning numbers to objects in the population and their properties. For eg, studying of student performance in exam, where students will be objects of study and their marks will be the properties. To distinguish students, we may have student id and marks gained by each, in terms of numbers / grades.

The levels of Measurement

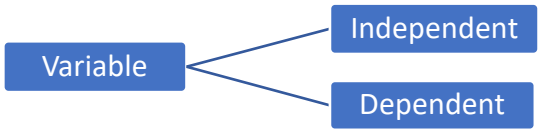
The Measurement can be majorly classified into four types under two groups.

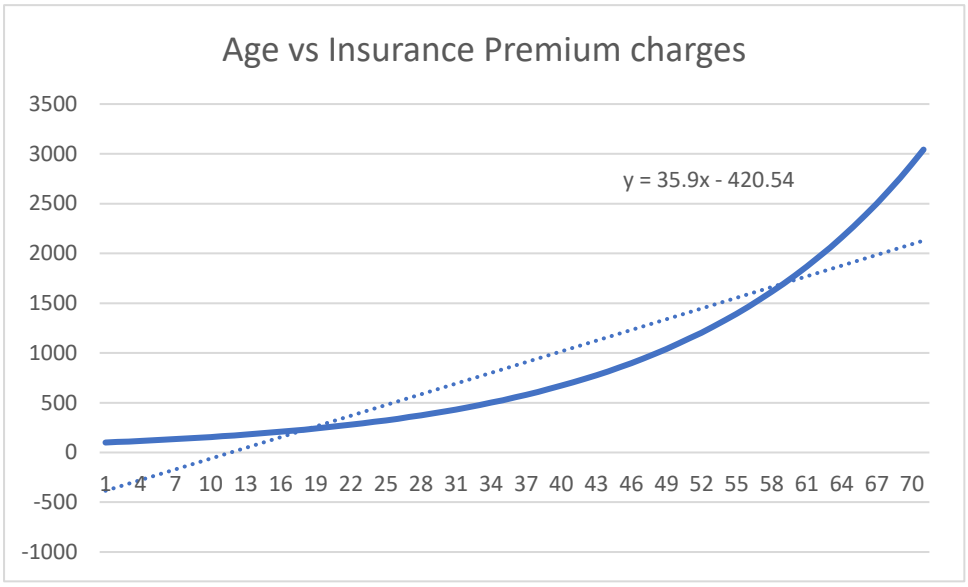


Variables

The data consists of variables, nothing but the attributes of the object we are interested in. The variables are of two types based on the dependency or relationships. ie Independent and dependent variables.

ie Dependent variables are influencing other variables or being influenced by other variables where as Independent variables are not influenced by others or influence others.



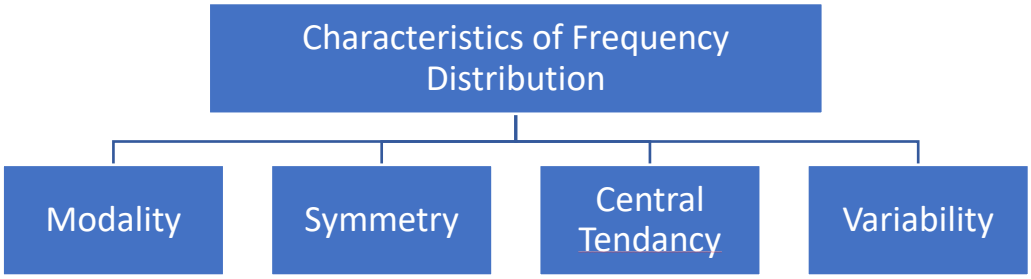


Dependency between the Age and the Insurance Premium Charges

Data Summarization

For summarization, the Statistics provides methods to describe the data in a summarised form and to represent the information contained in the data through these derived values or how it is distributed / spread.

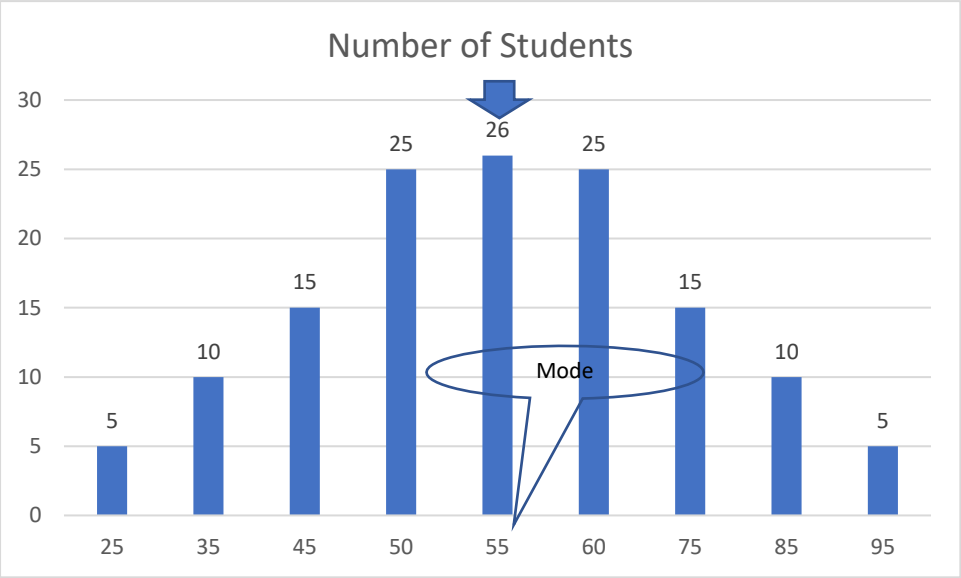
The frequency distribution table basically provides us a way to summarise the multiple occurrences of data. Based on this we also will be able to find the other characteristics of the frequency distribution, like Modality, Symmetry, Central Tendency and Variability of data.



Modality

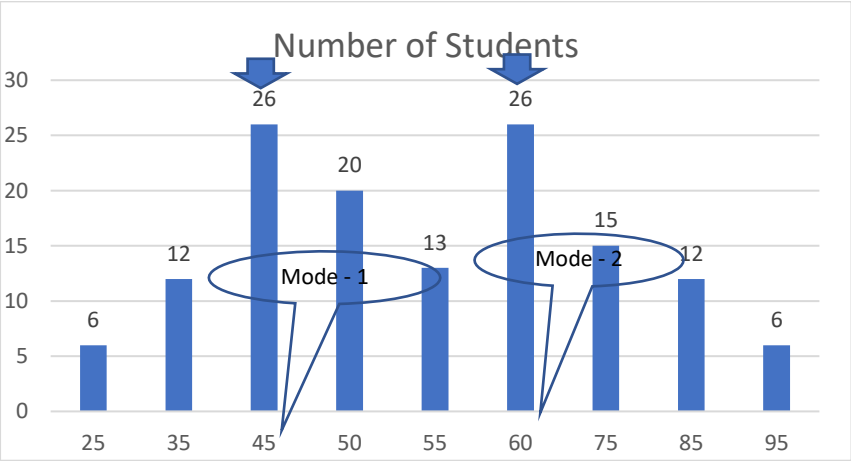
Modality indicates which value is getting repeated more in the data set being studied. If the most repeated value is only one then it is considered to be unimodal. If there are two values repeated equally i.e. two modes, then it is considered to be bi-modal.

Unimodal



Since 26 students scored 55 and highest frequency among the data set, 55 is the mode for this data set. And this is considered to be Unimodal.

Bimodal



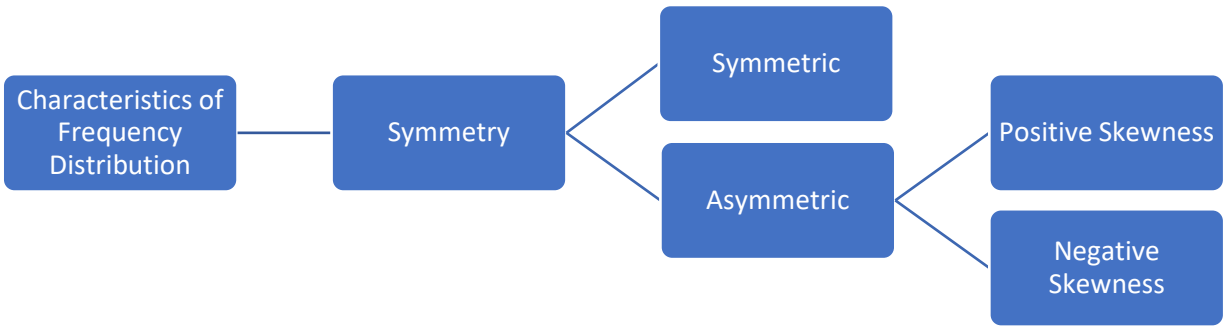
In the above example, 26 students scored marks 45 and 60. Since 26 is the highest frequency for the data set, we arrive at two modes, ie 45 and 60. Since there are two modes are there for this data set, this is considered to be bimodal.

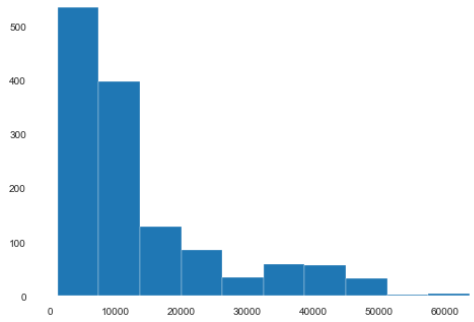
Symmetry

The Symmetry of the frequency distribution gives us the insight of how the data is distributed. When the data is distributed equally between the top and the lower ends by peeking in the mean (ie mean=median=mode), we get a normal curve or bell curved shape distribution. This is symmetric from the middle point of the curve.

In case if the distribution is not equally distributed between the top and the lower ends, the peek may be on the left or right side of the curve. This is considered to be asymmetric. If the peek is on the right ie lower values are more, then it is considered to be positive skewed. And if the peek is to the left, ie higher values are more, then it is considered to be positively skewed.

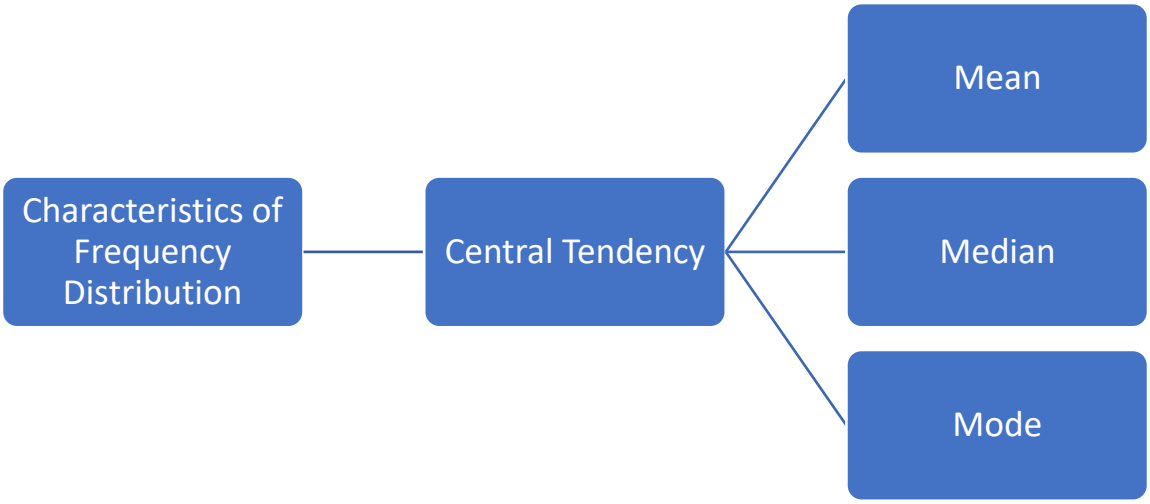
To identify the skewness, we need to draw the histogram and then super impose the normal curve on the same.





Right Skewed distribution of the insurance premiums

Central Tendency



Central Tendency of the data set can be identified by Mean, Median and Mode. These statistical values of the data set will give us the understanding of what is the common characteristic of the data set.

Mean

The mean is arrived at by summing up the data and dividing it by the total number of occurrences.

The value of the Mean will get affected by the outliers, ie the abnormal values present in the dataset. For eg, if we are trying to get the mean income of a locality and if there are 5 rich persons whose annual income in billions and the rest of the population with annual income ranging from 20 to 50k, the mean annual income will be higher than

50k because of the 5 rich persons. In this case this will not reflect the true picture of the majority of the people income.

Mean will give us a better picture when there are no outliers in the data set.

Median

The Median is arrived at by ordering the data set in the ascending order and finding the value which is there in the middle of the data set. In case if the number of elements in the data set is even, then we need to take the average of the two values which are lying at the middle point.

Median will not be affected by the outliers.

Mode

The Mode is arrived at by finding the value which is occurring frequently in the data set.

Dispersion of Measures

The Dispersion of Measures, gives the idea on how the data is spread from the centre/mean. The measures which helps us to find the dispersion are as given below

- Range
- Variance / Standard Deviation
- Mean absolute Deviation
- Inter quartile range

Range

One of the quick ways to understand the spread of data is the range of the data set, which is identified by the difference between the maximum value and the minimum value in the data set.

Range = Max – Min

Variance and Standard Deviation

The Variance and Standard Deviation provides us the information on how the data is dispersed from the mean and thus indicates the consistency of the data. This also indicates the reliability/predictability of the data

The Variance is nothing but the sum of square of the difference between the values in the data set and the mean, divided by the number of data points.

Variance = (Sum(X – Mean(X)))^2 / n

Standard Deviation = Sqrt (Variance)

Mean Absolute Deviation (MAD)

The Mean Absolute Deviation (MAD) is the average of the absolute difference between actual values and their mean. This also provides us the information on spread of the data, with respect to the mean. This measurement is used to understand the demand variability.

Mean Absolute Deviation = $(1/N)\ \Sigma\ |\ xi - \mu\ |$

Inter quartile range (IQR)

The Inter Quartile Range (IQR) indicates the spread of the data. This is identified by dividing the data into four quarters based on the percentile after arranging in order. The IQR is the difference between the Q3 and Q1

IQR = Q3 – Q1

Box–and–Whisker Plots

Box-and-Whisker Plots helps us to identify the spread and the outliers. This is also called *box plot.* Another name for this five-number summary, basically the box-and-whisker plot is derived from five numbers, ie the Minimum, first quartile, median, third quartile and the maximum.

In a box-and-whisker plot, the box is drawn from the first quartile to the third quartile and the vertical line goes through the box at the median point. The max and min points are joined with the box, which forms as whisker.

The min and maximum need to be within 1.5 times the IQR. The data points beyond the 1.5 times IQR is considered to be outliers.