**WEEK 5**

**Task 0 : Probability and Statistics**

[Data Analytics](https://dashboard.stige.in/index.php/courses/lms-data-analytics/) [Task 0 : Probability and Statistics](https://dashboard.stige.in/index.php/lessons/task-0-probability-and-statistics/)

In Progress

Probability and Statistics form the basis of Data Science. The probability theory is very much helpful for making the prediction. Estimates and predictions form an important part of Data science. With the help of statistical methods, we make estimates for the further analysis. Thus, statistical methods are largely dependent on the theory of probability. And all of probability and statistics is dependent on Data.

**Data**

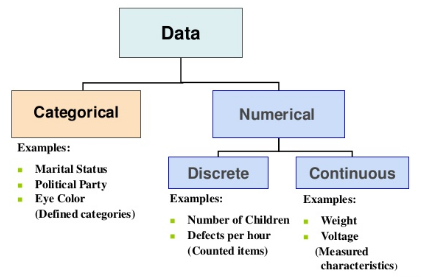
Data is the collected information(observations) we have about something or facts and statistics collected together for reference or analysis.

*Data — a collection of facts (numbers, words, measurements, observations, etc) that has been translated into a form that computers can process*

**Why does data matter?**

* Helps in understanding more about the data by identifying relationships that may exist between 2 variables.
* Helps in predicting the future or forecast based on the previous trend of data.
* Helps in determining patterns that may exist between data.
* Helps in detecting fraud by uncovering anomalies in the data.

Data matters a lot nowadays as we can infer important information from it. Now let’s delve into how data is categorized. Data can be of 2 types categorical and numerical data. For Example in a bank, we have regions, occupation class, gender which follow categorical data as the data is within a fixed certain value and balance, credit score, age, tenure months follow numerical continuous distribution as data can follow an unlimited range of values.

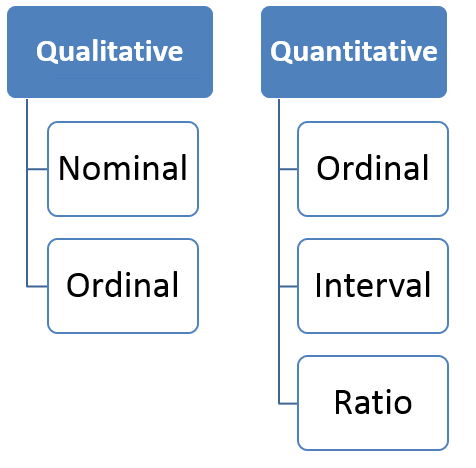


Note: Categorical Data can be visualized by Bar Plot, Pie Chart, [Pareto Chart](https://en.wikipedia.org/wiki/Pareto_chart). Numerical Data can be visualized by Histogram, Line Plot, Scatter Plot.

**Descriptive Statistics**

A descriptive statistic is a summary statistic that quantitatively describes or summarizes features of a collection of information. It helps us in knowing our data better. It is used to describe the characteristics of data.

**Measurement level of Data**



The qualitative and quantitative data is very much similar to the above categorical and numerical data.

**Nominal**: Data at this level is categorized using names, labels or qualities. eg: Brand Name, ZipCode, Gender.

**Ordinal**: Data at this level can be arranged in order or ranked and can be compared. eg: Grades, Star Reviews, Position in Race, Date

**Interval**: Data at this level can be ordered as it is in a range of values and meaningful differences between the data points can be calculated. eg: Temperature in Celsius, Year of Birth

**Ratio**: Data at this level is similar to interval level with added property of an inherent zero. Mathematical calculations can be performed on these data points. eg: Height, Age, Weight

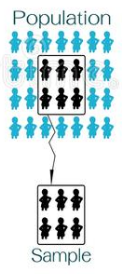
Check this out: [DATA | Mahrita Harahap](https://mahritaharahap.wordpress.com/teaching-areas/inferential-statistics/data/)

**Population or Sample Data**

Before performing any analysis of data, we should determine if the data we’re dealing with is population or sample.

**Population:**Collection of all items (N) and it includes each and every unit of our study. It is hard to define and the measure of characteristic such as mean, mode is called parameter.

**Sample:**Subset of the population (n) and it includes only a handful units of the population. It is selected at random and the measure of the characteristic is called as statistics.



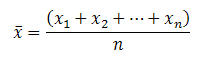
For Example, say you want to know the mean income of the subscribers to a movie subscription service(parameter). We draw a random sample of 1000 subscribers and determine that their mean income(x̄) is $34,500 (statistic). We conclude that the population mean income (μ) is likely to be close to $34,500 as well.

Now before looking at distributions of data. Let’s take a look at measures of data.

**Measure of Central Tendency**

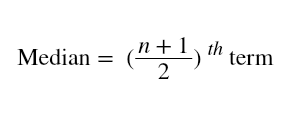
The measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data. As such, measures of central tendency are sometimes called measures of central location. They are also classed as summary statistics.

**Mean**: The mean is equal to the sum of all the values in the data set divided by the number of values in the data set i.e the calculated average. **It susceptible to outliers** when unusual values are added it gets skewed i.e deviates from the typical central value.

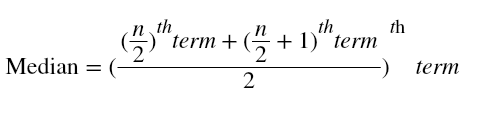


**Median**: The median is the middle value for a dataset that has been arranged in order of magnitude. Median is a better alternative to mean as it is less affected by outliers and skewness of the data. The median value is much closer than the typical central value.

If the total number of values is odd then



If the total number of values is even then

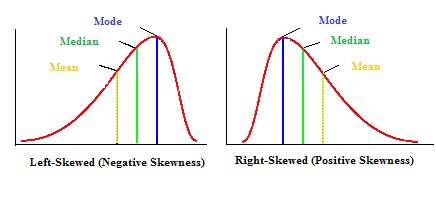


**Mode:**The mode is the most commonly occurring value in the dataset. The mode can, therefore sometimes consider the mode as being the most popular option.

For Example, In a dataset containing {13,35,54,54,55,56**,**57,67,85,89,96} values. Mean is 60.09. Median is 56. Mode is 54.

**Measures of Asymmetry**

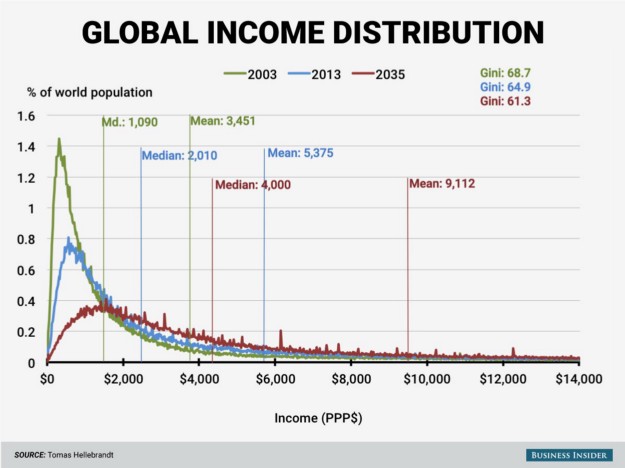
**Skewness:**Skewness is the asymmetry in a statistical distribution, in which the curve appears distorted or skewed towards to the left or to the right. Skewness indicates whether the data is concentrated on one side.



**Positive Skewness:**Positive Skewness is when the mean>median>mode. The outliers are skewed to the right i.e the tail is skewed to the right.

**Negative Skewness:**Negative Skewness is when the mean<median<mode. The outliers are skewed to the left i.e the tail is skewed to the left.

Skewness is important as it tells us about where the data is distributed.

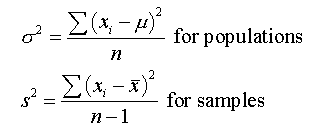


For eg: Global Income Distribution in 2003 is highly right-skewed. We can see the mean $3,451 in 2003(green) is greater than the median $1,090. It suggests that the global income is not evenly distributed. Most individuals incomes are less than $2,000 and less number of people with income above $14,000, so the skewness. But it seems in 2035 according to the forecast income inequality will decrease over time.

**Positive Skewness:**Positive Skewness is when the mean>median>mode. The outliers are skewed to the right i.e the tail is skewed to the right.

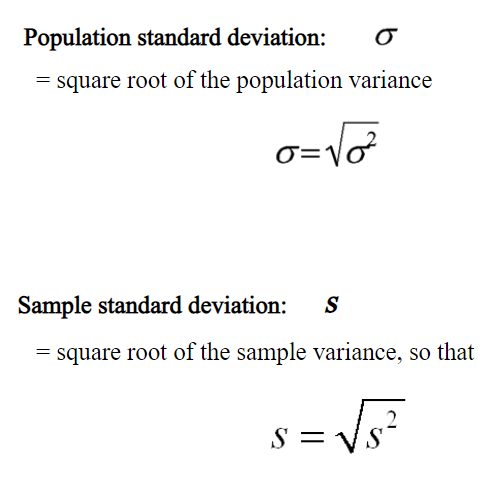
**Negative Skewness:**Negative Skewness is when the mean<median<mode. The outliers are skewed to the left i.e the tail is skewed to the left.

Skewness is important as it tells us about where the data is distributed.

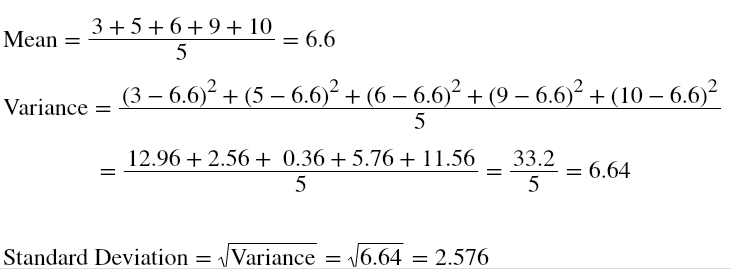


Note: *The units of values and variance is not equal so we use another variability measure.*

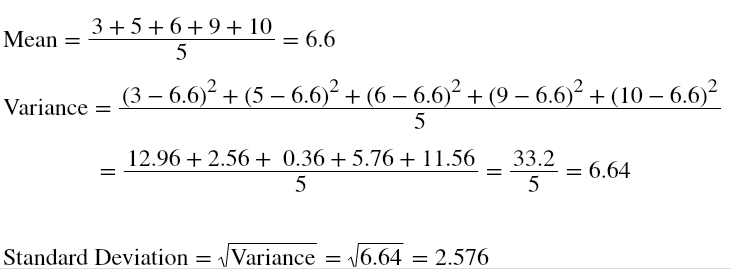
**Standard Deviation:**As Variance suffers from unit difference so standard deviation is used. The square root of the variance is the standard deviation. It tells about the concentration of the data around the mean of the data set.



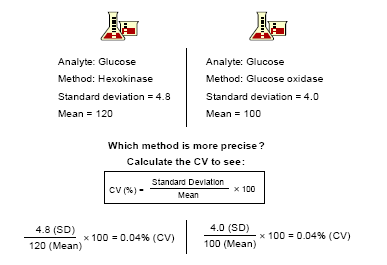
For eg: {3,5,6,9,10} are the values in a dataset.



**Coefficient of Variation(CV):** It is also called as the relative standard deviation. It is the ratio of standard deviation to the mean of the dataset.



Standard deviation is the variability of a single dataset. Whereas the coefficient of variance can be used for comparing 2 datasets.



From the above example, we can see that the CV is the same. Both methods are precise. So it is perfect for comparisons.

**Measures of Quartiles**

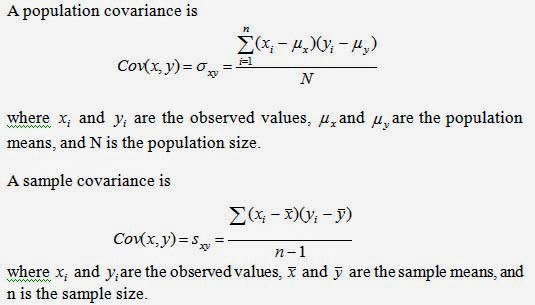
Quartiles are better at understanding as every data point considered.

Check my previous [post](https://towardsdatascience.com/data-visualization-using-matplotlib-16f1aae5ce70) — In the Boxplot [Section](https://medium.com/p/16f1aae5ce70#ec41), I have elaborated on Quartiles.

**Measures of Relationship**

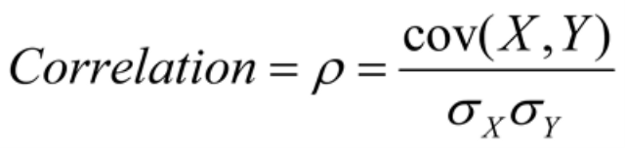
Measures of relationship are used to find the comparison between 2 variables.

**Covariance:**Covariance is a measure of the relationship between the variability of 2 variables i.e It measures the degree of change in the variables, when one variable changes, will there be the same/a similar change in the other variable.



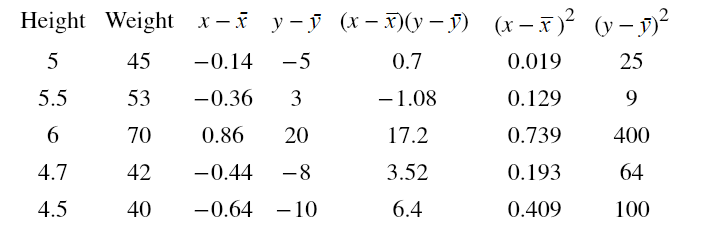
[Covariance does not give effective information about the relation between 2 variables as it is not normalized.](https://medium.com/m/signin)

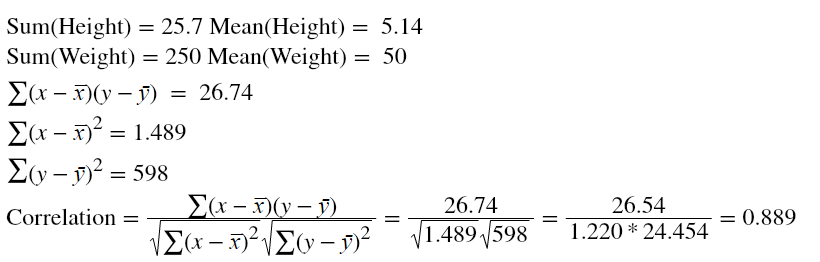
**Correlation:**Correlation gives a better understanding of covariance. It is normalized covariance. Correlation tells us how correlated the variables are to each other. It is also called as Pearson Correlation Coefficient.



The value of correlation ranges from -1 to 1. -1 indicates negative correlation i.e with an increase in 1 variable independent there is a decrease in the other dependent variable.1 indicates positive correlation i.e with an increase in 1 variable independent there is an increase in the other dependent variable.0 indicates that the variables are independent of each other.

For Example,





Correlation 0.889 tells us Height and Weight has a positive correlation. It is obvious that as the height of a person increases weight too increases.

**Note:**[**Correlation does not imply causation**](https://en.wikipedia.org/wiki/Correlation_does_not_imply_causation)**,**[**Spurious Correlation for some strange correlations**](http://www.tylervigen.com/spurious-correlations)**.**

**Conclusion**

In this article, we learnt about descriptive statistics which helps us to know about our data better by understanding crucial characteristics in a dataset.

# Task 1 : Teaser

[Data Analytics](https://dashboard.stige.in/index.php/courses/lms-data-analytics/) [Task 1 : Teaser](https://dashboard.stige.in/index.php/lessons/task-1-teaser/)

In Progress

Click on the below button to open Udacity ‘ s Course topic .

[Teaser](https://classroom.udacity.com/courses/st101/lessons/e35b6dd5-5d9e-4e81-b248-2ed3fbd64fe5/concepts/a47a2add-2e17-4e5a-be92-8fbbd19f91e1)

# Task 2 : Looking at data

[Data Analytics](https://dashboard.stige.in/index.php/courses/lms-data-analytics/) [Task 2 : Looking at data](https://dashboard.stige.in/index.php/lessons/task-2-looking-at-data/)

In Progress

Click on the below button to open Udacity ‘ s Course topic .

[Looking at Data](https://classroom.udacity.com/courses/st101/lessons/c8e43ad5-0d86-4048-969e-22a3a4b7f637/concepts/67ff012a-fbcc-4ecc-a40b-42b94e2b3ad0)

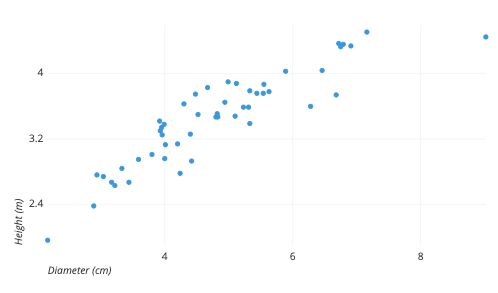
# Task 3: Scatter Plot

[Data Analytics](https://dashboard.stige.in/index.php/courses/lms-data-analytics/) [Task 3: Scatter Plot](https://dashboard.stige.in/index.php/lessons/task-3-scatter-plot/)

In Progress

**Scatter plots** (also called scatter graphs) use **dots** to represent individual pieces of data. In statistics, these plots are useful to see if two variables are related to each other .A scatter plot can suggest **various kinds of correlations** between variables with a certain [confidence interval](https://en.wikipedia.org/wiki/Confidence_interval). For example, weight and height would be on the y-axis, and height would be on the x-axis. Correlations may be positive (rising), negative (falling), or null (uncorrelated). If the dots’ pattern from lower left to upper right indicates a positive [correlation](https://en.wikipedia.org/wiki/Correlation) between the variables being studied. If the pattern of dots slopes from upper left to lower right, it indicates a negative correlation. A line of [best fit](https://en.wikipedia.org/wiki/Curve_fitting) (alternatively called ‘trendline’) can be drawn to study the **relationship between the variables.**

A **scatter plot** (aka scatter chart, scatter graph) uses dots to represent values for two different numeric variables. The position of each dot on the horizontal and vertical axis indicates values for an individual data point. Scatter plots are used to observe relationships between variables.

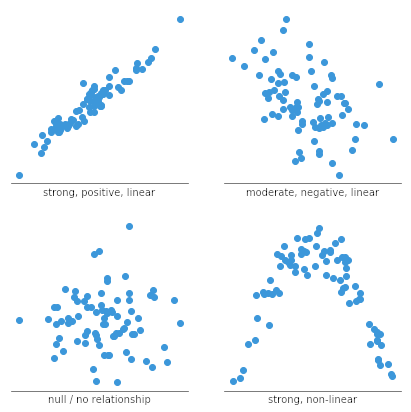


The example scatter plot above shows the diameters and heights for a sample of fictional trees. Each dot represents a single tree; each point’s horizontal position indicates that tree’s diameter (in centimeters) and the vertical position indicates that tree’s height (in meters). From the plot, we can see a generally tight positive correlation between a tree’s diameter and its height. We can also observe an outlier point, a tree that has a much larger diameter than the others. This tree appears fairly short for its girth, which might warrant further investigation.

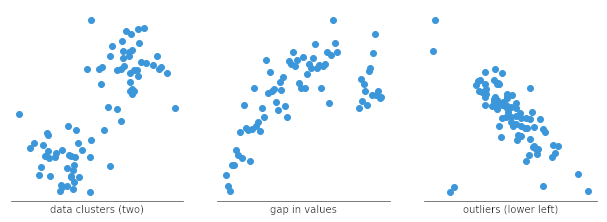
## ****When you should use a scatter plot****

**Scatter plots’ primary uses are to observe and show relationships between two numeric variables.**The dots in a scatter plot not only report the values of individual data points but also patterns when the data are taken as a whole.

Identification of **correlational relationships** is common with scatter plots. In these cases, we want to know, if we were given a particular horizontal value, what a good prediction would be for the vertical value. You will often see the variable on the horizontal axis denoted an **independent variable**, and the variable on the vertical axis the **dependent variable**. Relationships between variables can be described in many ways: **positive or negative, strong or weak, linear or nonlinear**.



A scatter plot can also be useful for identifying other patterns in data. We can divide data points into groups based on how closely sets of points cluster together. **Scatter plots can also show if there are any unexpected gaps in the data and if there are any outlier points.** This can be useful if we want to segment the data into different parts, like in the development of user personas.

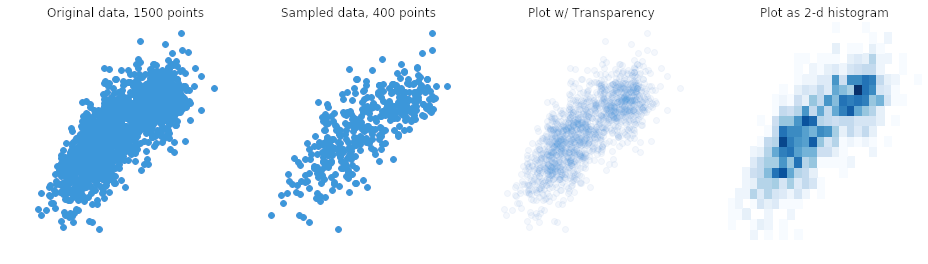


## ****Common issues when using scatter plots****

#### Overplotting

When we have lots of data points to plot, this can run into the issue of overplotting. **Overplotting is the case where data points overlap to a degree where we have difficulty seeing relationships between points and variables.** It can be difficult to tell how densely packed data points are when many of them are in a small area.

There are a few common ways to alleviate this issue. One alternative is to **sample only a subset of data points**: a random selection of points should still give the general idea of the patterns in the full data. We can also **change the form of the dots, adding transparency to allow for overlaps to be visible, or reducing point size** so that fewer overlaps occur. As a third option, we might even **choose a different chart type** like the [heatmap](https://chartio.com/learn/charts/heatmap-complete-guide/), where color indicates the number of points in each bin. Heatmaps in this use case are also known as 2-d histograms.



#### *****Interpreting correlation as causation*****

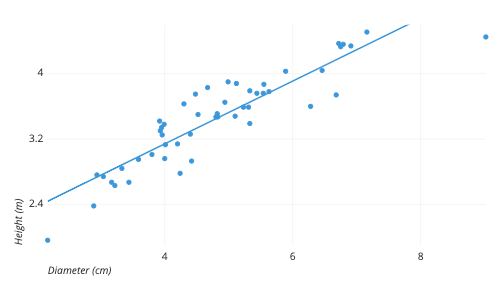
This is not so much an issue with creating a scatter plot as it is an issue with its interpretation. **Simply because we observe a relationship between two variables in a scatter plot, it does not mean that changes in one variable are responsible for changes in the other.**This gives rise to the common phrase in statistics that [correlation does not imply causation](https://en.wikipedia.org/wiki/Correlation_does_not_imply_causation). It is possible that the observed relationship is driven by some third variable that affects both of the plotted variables, that the causal link is reversed, or that the pattern is simply coincidental.

**For example**, it would be wrong to look at city statistics for the amount of green space they have and the number of crimes committed and conclude that one causes the other, this can ignore the fact that larger cities with more people will tend to have more of both, and that they are simply correlated through that and other factors. If a causal link needs to be established, then further analysis to control or account for other potential variables effects needs to be performed, in order to rule out other possible explanations.

## ****Common scatter plot options****

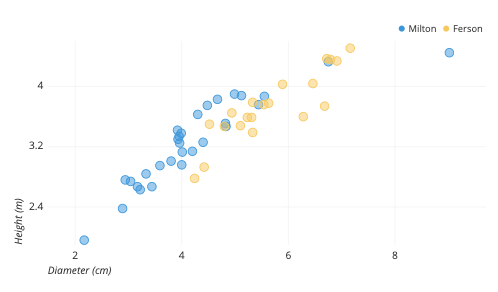
#### Add a trend line

When a scatter plot is used to look at a predictive or correlational relationship between variables, it is common to add a trend line to the plot showing the mathematically best fit to the data. This can provide an additional signal as to how strong the relationship between the two variables is, and if there are any unusual points that are affecting the computation of the trend line.

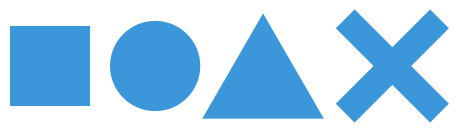


#### Categorical third variable

A common modification of the basic scatter plot is the addition of a third variable. Values of the third variable can be encoded by modifying how the points are plotted. For a third variable that indicates categorical values (like geographical region or gender), the most common encoding is through point color. Giving each point a distinct hue makes it easy to show membership of each point to a respective group.

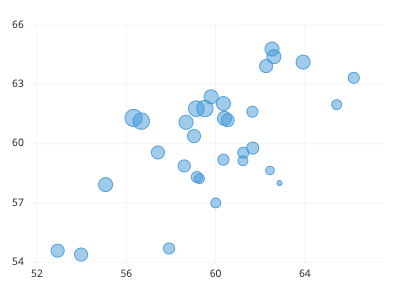
Coloring points by tree type shows that Fersons (yellow) are generally wider than Miltons (blue), but also shorter for the same diameter.

One other option that is sometimes seen for third-variable encoding is that of shape. One potential issue with shape is that different shapes can have different sizes and surface areas, which can have an effect on how groups are perceived. However, in certain cases where color cannot be used (like in print), the shape may be the best option for distinguishing between groups.

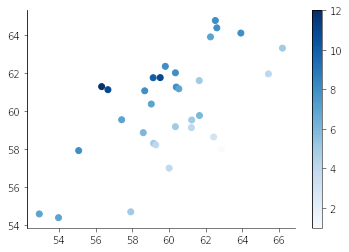
The shapes above have been scaled to use the same amount of ink.

#### Numeric third variable

For third variables that have numeric values, a common encoding comes from changing the point size. A scatter plot with a point size based on a third variable actually goes by a distinct name, the [bubble chart](https://chartio.com/learn/charts/bubble-chart-complete-guide/). Larger points indicate higher values. A more detailed discussion of how bubble charts should be built can be read in its own article.

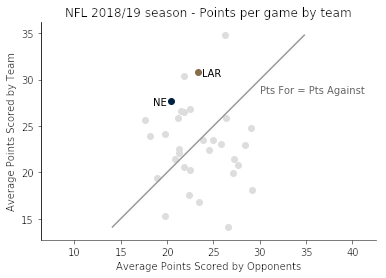


**Hue** can also be used to depict numeric values as another alternative. Rather than using distinct colors for points like in the categorical case, we want to use a continuous sequence of colors, so that, for example, darker colors indicate higher value. Note that, for both size and color, a legend is important for the interpretation of the third variable, since our eyes are much less able to discern size and color as easily as position.



#### Highlight using annotations and color

If you want to use a scatter plot to present insights, it can be good to highlight particular points of interest through the use of annotations and color. Desaturating unimportant points makes the remaining points stand out, and provides a reference to compare the remaining points against.



Click on the below button to open Udacity ‘ s Course topic .

[Scatter Plot](https://classroom.udacity.com/courses/st101/lessons/aac89651-d002-47a6-b62c-7b8ed7dfb31d/concepts/7c6dbee6-a0eb-42a3-a51b-5c630977f533)

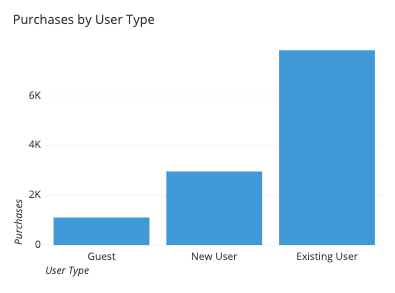
# Task 4 : Bar Chart

[Data Analytics](https://dashboard.stige.in/index.php/courses/lms-data-analytics/) [Task 4 : Bar Chart](https://dashboard.stige.in/index.php/lessons/task-4-bar-chart/)

In Progress

A **bar graph** is a chart that plots data using **rectangular bars or columns** (called bins) that represent the total amount of observations in the data for that category.Bar graphs are commonly used in financial analysis for displaying data. A stock [volume](https://www.investopedia.com/terms/v/volume.asp) chart is a commonly used type of vertical bar graph. A[**histogram**](https://www.investopedia.com/terms/h/histogram.asp)is an example of a bar graph used in statistical analysis that depicts a probability distribution in some data or sample.

A bar chart (aka bar graph, column chart)**plots numeric values for levels of a categorical feature as bars**. Levels are plotted on one chart axis, and values are plotted on the other axis. Each categorical value claims one bar, and the length of each bar corresponds to the bar’s value. Bars are plotted on a common baseline to allow for easy comparison of values.



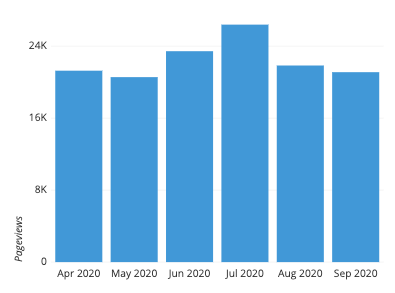
This example bar chart depicts the **number of purchases made on a site by different types of users**. The categorical feature, user type, is plotted on the horizontal axis, and each bar’s height corresponds to the number of purchases made under each user type. We can see from this chart that while there are about three times as many purchases from new users who create user accounts than those that do not create user accounts (guests), both are dwarfed by the number of purchases made by repeating users.

## ****When you should use a bar chart****

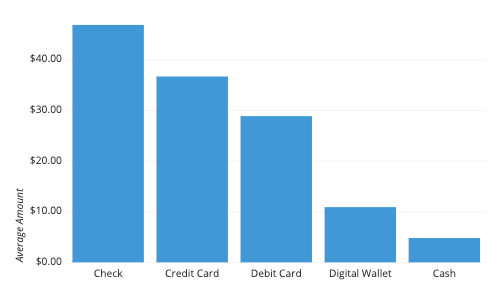
**A bar chart is used when you want to show a distribution of data points or perform a comparison of metric values across different subgroups of your data.** From a bar chart, we can see which groups are highest or most common, and how other groups compare against the others. Since this is a fairly common task, bar charts are a fairly ubiquitous chart type.

The**primary variable of a bar chart is its categorical variable**. A categorical variable takes discrete values, which can be thought of as labels. Examples include state or country, industry type, website access method (desktop, mobile), and visitor type (free, basic, premium). Some categorical variables have ordered values, like dividing objects by size (small, medium, large). In addition, some non-categorical variables can be converted into groups, like aggregating temporal data based on date (eg. dividing by quarter into 20XX-Q1, 20XX-Q2, 20XX-Q3, 20XX-Q4, etc.) **The important point for this primary variable is that the groups are distinct.**

In contrast, the **secondary variable will be numeric in nature**. The secondary variable’s values determine the length of each bar. These values can come from a great variety of sources. In its simplest form, the values may be a simple frequency count or proportion for how much of the data is divided into each category – not an actual data feature at all. For example, the following plot counts pageviews over a period of six months. You can see from this visualization that there was a small peak in June and July before returning to the previous baseline.



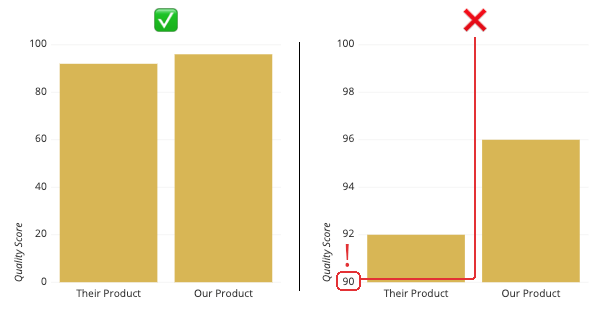
Other times, the values may be an average, total, or some other summary measure computed separately for each group. In the following example, the height of each bar depicts the average transaction size by the method of payment. **Note that while the average payments are highest with checks, it would take a different plot to show how often customers actually use them.**



## ****Best practices for using bar charts****

#### ****Use a common zero-valued baseline****

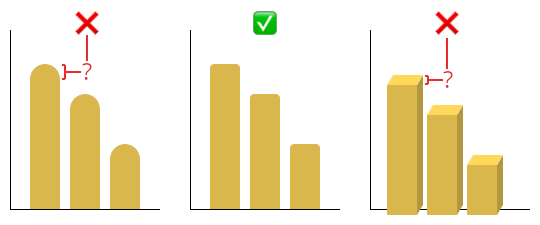
First and foremost, **make sure that all of your bars are being plotted against a zero-value baseline**. Not only does that baseline make it easier for readers to compare bar lengths, but it also maintains the truthfulness of your data visualization. A bar chart with a non-zero baseline or some other gap in the axis scale can easily misrepresent the comparison between groups since the ratio in bar lengths will not match the ratio in actual bar values.

By cutting 90 points out of the vertical axis, a small 4-point difference can be exaggerated to look like a 1:3 ratio.

#### ****Maintain rectangular forms for your bars****

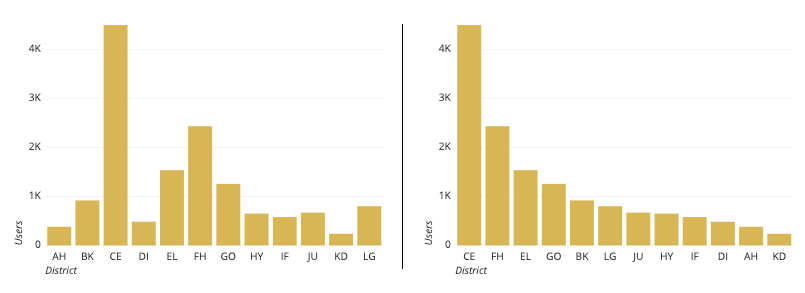
**Another major no-no** **is to mess with the shape of the bars to be plotted.** Some tools will allow for the rounding of the bar caps, rather than just have straight edges. This rounding means that it’s difficult for the reader to tell where to read the actual value: from the top of the semicircle, or somewhere in the middle? A little bit of rounding of the corners can be okay, but make sure each bar is flat enough to discern its true value and provide an easy comparison between bars.

Similarly, you should **avoid including 3-d effects on your bars**. As with heavy rounding, this can make it harder to know how to measure bar lengths, and as a bonus, might cause baselines to not be aligned (see the above point).



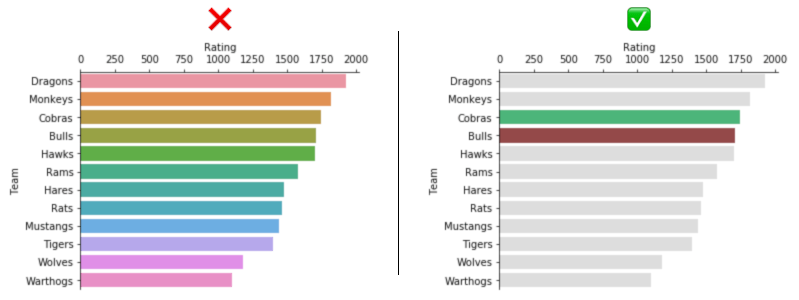
#### ****Consider the ordering of category levels****

One consideration you should have when putting together a bar chart is what order in which you will plot the bars. **A** **standard convention to take is to sort the bars from longest to shortest**: while it is always possible to compare the bar lengths no matter the order, this can reduce the burden on the reader to make those comparisons themselves. The major exception to this is if the category labels are inherently ordered in some way. In cases like that, the inherent ordering usually takes precedence.

The district codes aren’t inherently ordered, so a better representation is to sort by value.

#### ****Use color wisely****

Another consideration is **how you should use color in your bar charts**. Certain tools will color each bar differently by default, but this can distract the reader by implying additional meaning where none exists. Instead, color should be used with purpose. **For example**, you might use color to highlight specific columns for storytelling. Colors can also be used if they are meaningful for the categories posted (e.g. to match company or team colors).

The rainbow colors on the left don’t add anything meaningful to the interpretation of the plot. On the right side, most bars are neutral gray to highlight the comparison of the two colored bars.

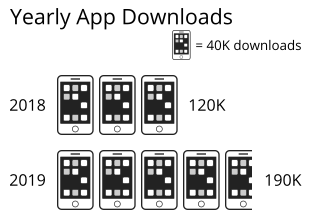
### **Common misuses**

#### ****Replacing bars with images****

It may be tempting to replace bars with pictures that depict what is being measured (e.g. bags of money for money amounts), **be careful that you do not misrepresent your data in this way**. **If your choice of symbol scales both width and height with value, differences will look much larger than they actually are, since people will end up comparing the areas of the bars rather than just their widths or heights**. In the example below, there is a 58% growth in downloads from 2018 to 2019. However, this growth is exaggerated with the icon-based representation, since the surface area of the 2019 icon is more than 2.5 times the size of the 2018 icon.



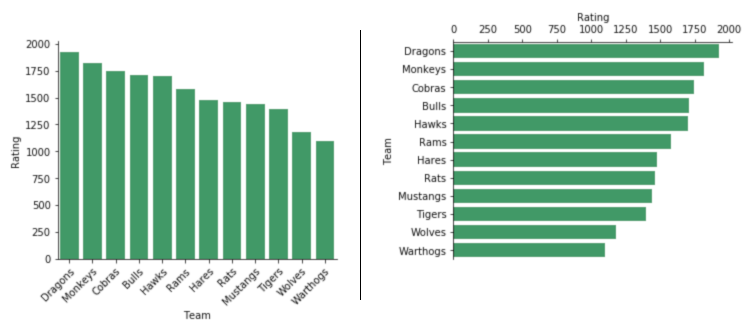
**If you feel the need to use icons to depict value, then a better – though still not great – option is to use the pictogram chart type instead.** In a pictogram chart, each category’s value is indicated by a series of icons, with each icon representing a certain quantity. In a certain sense, this is like changing the texture of its corresponding bar to a repeating image. One major caution with this chart type is that it can make values harder to read since the reader needs to perform some mental mathematics to gauge the relative values of each category.



## ****Common bar chart options****

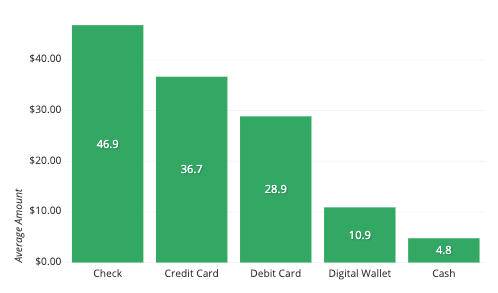
#### ****Horizontal bars vs. vertical bars****

A common bar chart variation is whether or not the bar chart should be oriented vertically (with categories on the horizontal axis) or horizontally (with categories on the vertical axis). While the vertical bar chart is usually the default, it’s a good idea to use a **horizontal bar chart**when you are faced with long category labels. In a vertical chart, these labels might overlap, and would need to be rotated or shifted to remain legible; the horizontal orientation avoids this issue.

If the bars from a previous example were vertically oriented, the Team tick labels would need to be rotated in order to be readable.

#### ****Include value annotations****

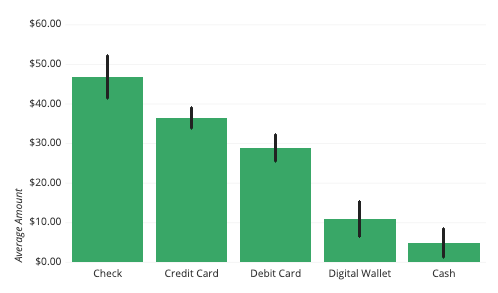
A common addition to bar charts is value annotations. While it is fairly easy for readers to compare bar lengths and gauge approximate values from a bar chart, exact values aren’t necessarily easy to state. Annotations can report these values where they are important and are usually placed in the middle of the bar or at their ends.



#### ****Include variability whiskers****

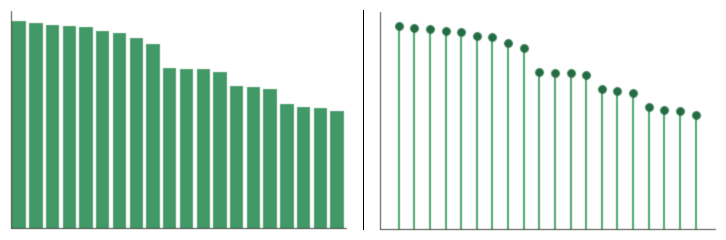
When the numeric values are a summary measure, a frequent consideration is whether or not to include error bars in the plot. **Error bars are additional whiskers added to the end of each bar to indicate variability in the individual data points that contributed to the summary measure.** Since there are many choices for uncertainty measures (e.g. standard deviation, confidence interval, interquartile range) it is important that when you display error bars, that you note in an annotation or comment what the error bars represent.

Alternatively, you may wish to depict variance within each category with a different chart type such as the [box plot](https://chartio.com/learn/charts/box-plot-complete-guide/) or [violin plot](https://chartio.com/learn/charts/violin-plot-complete-guide/). While these plots will have more elements for a reader to parse, they provide a deeper understanding of the distribution of values within each group.

Error bars indicate the standard deviation for transaction amounts for each payment type. The variability is lower for credit and debit cards compared to the others.

#### ****Lollipop chart****

One variation of the bar chart is the lollipop chart. **It presents exactly the same information as a bar chart, but with different aesthetics.** Instead of bars, we have lines topped by dots at their endpoints. **A lollipop chart is most useful when there are a lot of categories and their values are fairly close together.** By changing the aesthetic form of the plotted values, it can make the chart much easier to read.



Click on the below button to open Udacity ‘ s Course topic .

[Bar Chart](https://classroom.udacity.com/courses/st101/lessons/8da0746a-45d0-41b2-8f9c-e83c35b64e29/concepts/c041d5f4-0613-4a3a-9316-31225d265c18)

# Task 5 : Pie Charts

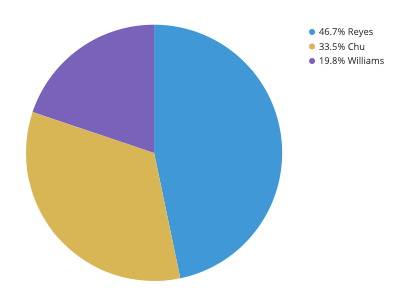
[Data Analytics](https://dashboard.stige.in/index.php/courses/lms-data-analytics/) [Task 5 : Pie Charts](https://dashboard.stige.in/index.php/lessons/task-5-pie-charts/)

In Progress

## ****What is a pie chart?****

A **pie chart** is another graphical device for **summarizing qualitative data**. The size of each slice of the pie is proportional to the number of data values in the corresponding class. In a pie chart, the [arc length](https://en.wikipedia.org/wiki/Arc_length) of each slice (and consequently its [central angle](https://en.wikipedia.org/wiki/Central_angle) and [area](https://en.wikipedia.org/wiki/Area)), is [proportional](https://en.wikipedia.org/wiki/Proportionality_(mathematics)) to the quantity it represents. While it is named for its resemblance to a [pie](https://en.wikipedia.org/wiki/Pie) which has been sliced, there are variations on the way it can be presented.

A pie chart shows how a total amount is divided between levels of a categorical variable as a circle divided into radial slices. Each categorical value corresponds with a single slice of the circle, and the size of each slice (both in area and arc length) indicates what proportion of the whole each category level takes.



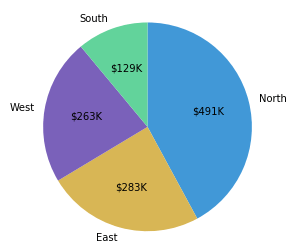
The pie chart above depicts the distribution of votes for a fictional election for a small city. We can see that Reyes, represented by the first blue slice, has just less than half of the votes. Chu (yellow) is in second, with about a third of the votes, while Williams (purple) is last, with about a fifth of the votes. The annotations in the upper right give us a more precise judgment of the proportions, but the pie chart tells the overarching story of where the votes fell.

## ****When you should use a pie chart****

Pie charts have a fairly narrow use-case that is encapsulated particularly well by its definition. In order to use a pie chart, **you must have some kind of whole amount that is divided into a number of distinct parts**. Your primary objective in a pie chart should be to compare each group’s contribution to the whole, as opposed to comparing groups to each other. If the above points are not satisfied, the pie chart is not appropriate, and a different plot type should be used instead.

The values that comprise a whole and the categories that divide the whole generally come in two major varieties. First of all, is **when the ‘whole’ represents a total count**. Examples of this include votes in an election divided by candidate, or number of transactions divided by user type (e.g. guest, new user, existing user).

A second type of **‘whole’ is when the total is a sum over an actual data variable**. For example, we might be interested not in the number of transactions, but the monetary total from all transactions. Dividing this total by an attribute like user type, age bracket, or location might provide insights as to where the business is most successful.



## ****Best practices for using a pie chart****

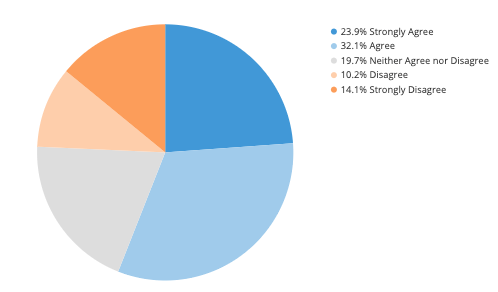
#### ****Include annotations****

It is actually very difficult to discern exact proportions from pie charts, outside of small fractions like 1/2 (50%), 1/3 (33%), and 1/4 (25%). Furthermore, if the slice values are meant to depict amounts rather than proportions, pie charts typically lack the tick marks to allow for the estimation of values directly from slice sizes. It is for these reasons that annotations are a standard inclusion for pie charts.

#### ****Consider the order of slices****

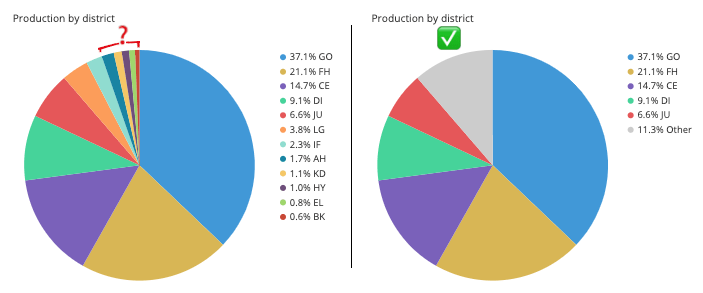
A **good order for slices can make it much easier for a reader to understand what the plot is saying**. A typical ordering goes from the largest slice to the smallest slice, very useful when there are categories with very similar values. However, if the category levels have an inherent ordering, then plotting slices in that order is usually better.

As for choosing a start point, it’s a good idea to plot slices from a cardinally-oriented direction. Visualization tools will usually start from the right or from the top. While starting from the right has a mathematical basis regarding conventions on measuring angles, starting from the top feels more intuitive, since it matches how we read from top to bottom, and how we think about progression of time on a clock or watch face.

We do not sort by size here since the labels are meaningful.

#### ****Limit the number of pie slices****

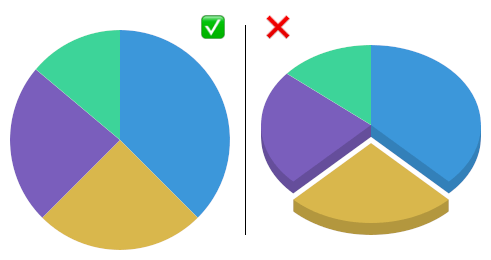
**Pie charts with a large number of slices can be difficult to read**. It can be difficult to see the smallest slices, and it can be difficult to choose enough colors to make all of the slices distinct. Recommendations vary, but if you have more than about five categories, you might want to think about using a different chart type. As another option, you might consider lumping small slices into a single ‘other’ slice, colored in a neutral gray.



#### ****Avoid distorting effects****

**Reading a pie chart accurately requires that the slices’ areas, arc lengths, and angles all point to an accurate representation of the data.** While avoiding 3-d effects is a good idea for any plot, it is especially important for pie charts. Squashing or stretching the circle or adding unnecessary depth can easily distort how large each slice compares to the whole.

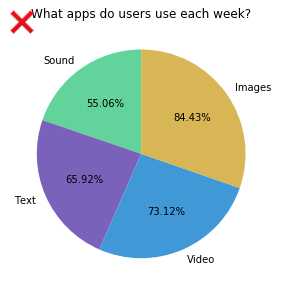
**Another distortion can come from the ‘exploded’ pie chart**, where slices are pulled out from the center for emphasis. This emphasis comes with a cost, where the gaps can make it more difficult to actually gauge the part-to-whole comparison.



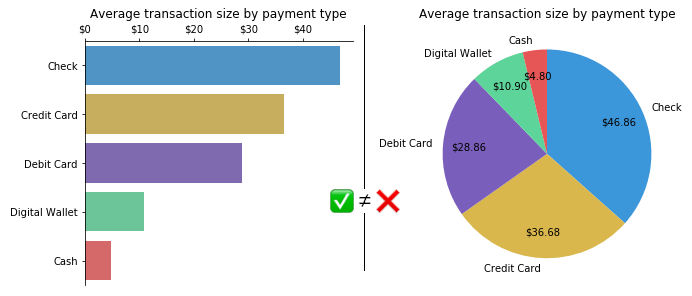
## ****Common misuses****

#### ****Fitting a pie to incompatible data****

**One of the most common mistakes with using a pie chart is to fit it to data that does not represent a parts-to-whole comparison**. This confusion occurs most often when the values to be plotted are percentages or proportions, but don’t comprise a complete whole. The example below shows how frequently the people surveyed used each of four applications, but since many people used multiple apps, the proportions sum to much more than 100%.

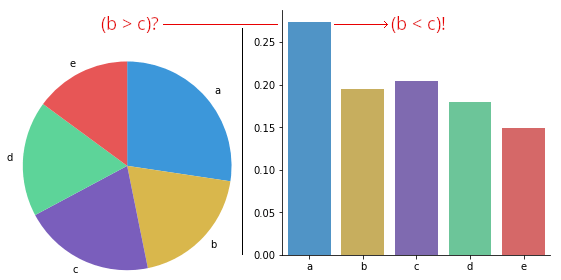


Another tricky case is if the values used for each group are a summary statistic that is not a total. The chart below was built on the average transaction amount for multiple transaction types. However, since it ignores how frequently each transaction type was used, it distorts how much revenue is coming in from each type. While checks have the highest average, they might also be fairly rare in use. **In both cases, a bar chart is an appropriate chart type to use.**



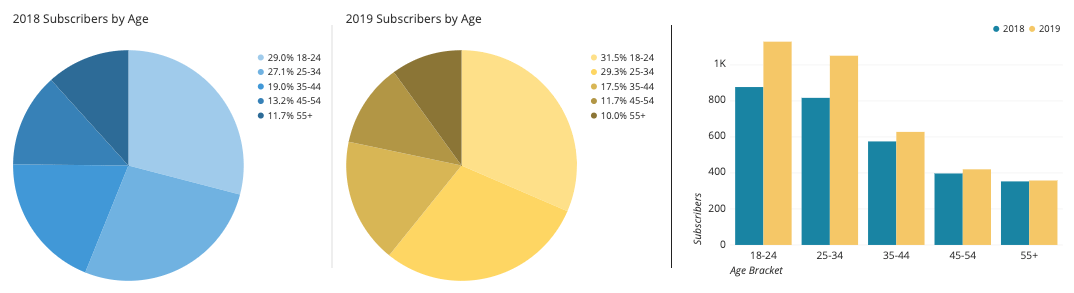
#### ****Using pie charts to compare groups to one another****

**If you want to make comparisons between groups rather than from each group to the whole, then you are better off with a different chart type.** Even when sorting slices by size, it can be difficult to tell how different two slices are, especially when they move away from the start/end point. In the below example, you might assume that the second slice is larger than the third due to order, but the corresponding bar chart actually shows the opposite. The main thing that you can really say from the pie chart is that both slices are approximately the same proportions as the whole.



#### ****Comparing values across multiple pie charts****

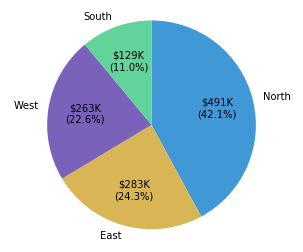
There may be cases where you will want to compare multiple pies to one another: for example, comparing user demographic distribution across multiple years. However, this runs into a similar issue as the previous section, where you want to compare groups to one another. Even worse, it’s a comparison between pies, so you can’t rely on the order of slices as easily for the comparison. Expressing the data using a different plot, like [stacked bar chart](https://chartio.com/learn/charts/stacked-bar-chart-complete-guide/), grouped bar chart, or [line chart](https://chartio.com/learn/charts/line-chart-complete-guide/), is often a better choice when this kind of group-to-group comparison is desired. Like with actual pies, pie charts are best taken one at a time.

Comparison of the pies may imply a shrinking of the older age groups by proportion, but the grouped bar chart demonstrates a growth in the younger groups.

## ****Common pie chart options****

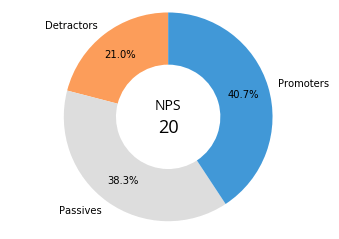
#### ****Absolute frequency vs. relative frequency****

Pie charts can be labeled in terms of absolute values or by proportions. **Labeling slices with absolute amounts and implying the proportions with the slice sizes is conventional, but consider the goals of your visualization carefully in order to decide on the best annotation style to use for your plot.** In some cases, including both numbers in the annotations can be worth the additional text.



#### ****Doughnut plot****

A doughnut plot (aka donut plot) is simply a pie chart with a central circle removed. **For the most part, there aren’t any significant differences between a pie chart and a donut chart, so the choice of a doughnut over a standard circle is mostly that of aesthetic.** One small boon for the ring shape is that the central area can be used for additional information or to report statistics.



For more knowledge, you can visit this Udacity lesson!

[Pie Charts](https://classroom.udacity.com/courses/st101/lessons/320f3ac8-d25d-4e34-ab81-515252b1856e/concepts/8dbc1c77-b172-4595-8453-a0f93a783575)

**Task 6 : Average**

[Data Analytics](https://dashboard.stige.in/index.php/courses/lms-data-analytics/) [Task 6 : Average](https://dashboard.stige.in/index.php/lessons/task-6-average/)

In Progress

In [statistics](https://www.wyzant.com/resources/lessons/math/statistics_and_probability), an**Average** is defined as the number that **measures the central tendency**of a given set of numbers. If all numbers in a list are the same number, then their average is also equal to this number. This property is shared by each of the many types of average.There are a number of different averages including but not limited to: mean, median, mode and range.

**The mean is the most commonly used average.** To get the mean value, you add up all the values and divide this total by the number of values. For example, if you wanted to find the mean of 11, 14 and 17, you would add them to give a total of 42, then divide that by the number of values you have, which is 3. So the mean of 11, 14 and 17 is 42/3 = 14.

**The other types of averages are:**

* The **median**, which places all your values in order from smallest to highest and finds the one in the middle. For example, the median of the values 3, 3, 4, 5, 9, 11 and 16 is 5.
* The **mode** is the most commonly occurring value. For example, the modal value of 1, 3, 6, 6, 6, 6, 7, 7, 12, 14 and 24 is 6 because it appears the most times.

**Using averages**

When the average of a dataset is presented to you, you need to consider which type of average has been used. Consider the average number of feet a person has. Most people in the world have two feet, so the modal value will be two. Similarly, if you were to use the median average, you would also find the answer to be two, as a very small minority of people have fewer than two feet, so two would remain the middle number.

However, if you were to calculate the mean, you would find that the answer is no longer two. There are a minority of people with fewer than two feet, for a variety of reasons, but this is enough to reduce the mean ever so slightly. As a result, almost everyone has more than the mean number of feet.

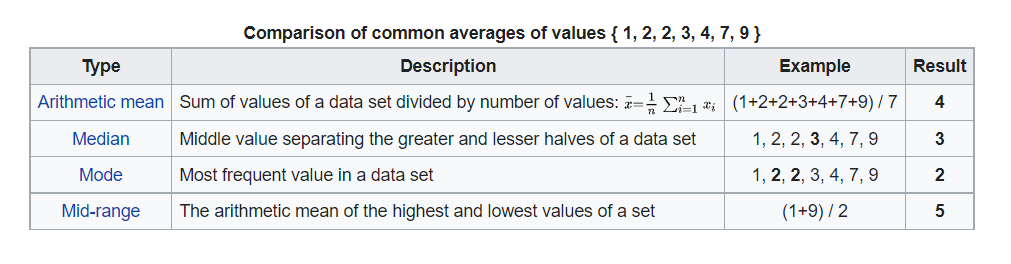
That’s enough about feet. Let’s instead consider the average of a scientific dataset. Average monthly rainfall is worked out from the recorded rainfall for that month over a specified number of years. The numbers below are the recorded January rainfall (in millimetres) for London, UK over 10 years. For simplicity these are arranged in order of smallest to highest.

17, 19, 51, 56, 69, 72, 72, 74, 75, 77

**The data show that January rainfall has ranged from 17 mm to 77 mm in the ten years that this dataset covers. The mean average rainfall is 58 mm to 2 sig figs. Do you agree that the average rainfall should be reported to 2 sig figs?**

The median rainfall is the middle value in this list. Because there are an even number of years in the sample, there is not a single middle value. Instead, both 69 and 72 are the middle numbers. The median is calculated as the number midway between these two numbers and is therefore 71 mm to 2 sig figs. The mode is the most common value, which is 72 mm.

This demonstrates perfectly the sensitivity of the mean average to extreme values. The mean average is lower than the median or mode averages due to the two very dry Januarys, which experienced only 17 and 19 mm of rain.



Click on the below button to open Udacity ‘ s Course topic .

[Average](https://classroom.udacity.com/courses/st101/lessons/e1cf0a0f-3640-4ba7-9d47-93d8cae2d932/concepts/a976cd59-47d0-487c-bb29-3643fb2f5126)

# Task 7 : Variance

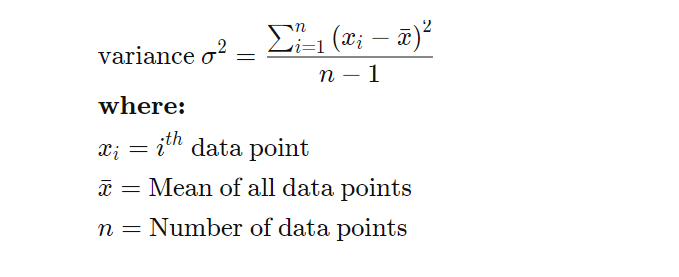
[Data Analytics](https://dashboard.stige.in/index.php/courses/lms-data-analytics/) [Task 7 : Variance](https://dashboard.stige.in/index.php/lessons/task-7-variance/)

In Progress

The term **variance** refers to a statistical measurement of the **spread between numbers** in a data set. More specifically, variance measures how far each number in the set is from the mean and thus from every other number in the set. Variance is often depicted by this symbol: σ2. It is used by both analysts and traders to determine[**volatility**](https://www.investopedia.com/terms/v/volatility.asp) and **market security.**

In statistics, **Variance** measures variability from the average or mean. **It is calculated by taking the differences between each number in the data set and the mean, then squaring the differences to make them positive, and finally dividing the sum of the squares by the number of values in the data set.**

Variance is calculated by using the following formula:



**A large variance indicates that numbers in the set are far from the mean and far from each other. A small variance, on the other hand, indicates the opposite. A variance value of zero, though, indicates that all values within a set of numbers are identical.**Every variance that isn’t zero is a positive number. A variance cannot be negative. That’s because it’s mathematically impossible since you can’t have a negative value resulting from a square.

Variance is an important metric in the investment world. **Variability is** **volatility, and volatility is a measure of risk.**It helps assess the risk that investors assume when they buy a specific asset and helps them determine whether the investment will be profitable. But how is this done? Investors can analyze the variance of the returns among assets in a portfolio to achieve the best asset allocation. In financial terms, the variance equation is a formula for comparing the performance of the elements of a portfolio against each other and against the mean.

#### ****Advantages and Disadvantages of variance****

**Statisticians** use variance to see how individual numbers relate to each other within a data set, rather than using broader mathematical techniques such as arranging numbers into quartiles. The **advantage of variance** is that it treats all deviations from the mean as the same regardless of their direction. The squared deviations cannot sum to zero and give the appearance of no variability at all in the data. One **drawback to variance**, though, is that it gives added weight to outliers. These are the numbers far from the mean. Squaring these numbers can skew the data. Another pitfall of using variance is that it is not easily interpreted. Users often employ it primarily to take the square root of its value, which indicates the standard deviation of the data set. As noted above, investors can use standard deviation to assess how consistent returns are over time.

#### ****Examples of variance****

Here’s a hypothetical example to demonstrate how variance works. Let’s say returns for stock in Company ABC are 10% in Year 1, 20% in Year 2, and −15% in Year 3. The average of these three returns is 5%. The differences between each return and the average are 5%, 15%, and −20% for each consecutive year. Squaring these deviations yields 25%, 225%, and 400%, respectively. If we add these squared deviations, we get a total of 650%. When you divide the sum of 650% by the number of returns in the data set—three in this case—it yields a variance of 216.67%. Taking the square root of the variance yields the standard deviation of 14.72% for the returns.

#### ****Related terms****

**Volatility:** Volatility measures how much the price of a security, derivative, or index fluctuates.  
**Standard Deviation:** The standard deviation is a statistic that measures the dispersion of a dataset relative to its mean. It is calculated as the square root of variance by determining the variation between each data point relative to the mean.

Click on the below button to open Udacity ‘ s Course topic .

[Variance](https://classroom.udacity.com/courses/st101/lessons/7c5c9efa-1747-4ab1-8665-da5776c52b0d/concepts/d08c6b76-1864-4a1f-9009-0ebaf3716a0d)

# Task 8 : Probability

[Data Analytics](https://dashboard.stige.in/index.php/courses/lms-data-analytics/) [Task 8 : Probability](https://dashboard.stige.in/index.php/lessons/task-8-probability/)

In Progress

**Probability** is a mathematical tool used to study **randomness.** It deals with the chance (the likelihood) of an event occurring.   
  
For example, if you toss a fair coin four times, the outcomes may not be two heads and two tails. The theory of probability began with the study of games of chance such as poker.

**Predictions take the form of probabilities**. To predict the likelihood of an earthquake, of rain, or whether you will get an A in this course, we use probabilities. Doctors use probability to determine the chance of a vaccination causing the disease the vaccination is supposed to prevent. A stockbroker uses probability to determine the rate of return on a client’s investments. You might use probability to decide to buy a lottery ticket or not. In your study of statistics, you will use the power of mathematics through probability calculations to **analyze** and **interpret** your data.

Click on the below button to open Udacity ‘ s Course topic .

[Probability](https://classroom.udacity.com/courses/st101/lessons/dae7b0a4-2939-4d44-87ab-64df5594f777/concepts/41a69bf1-5383-4dc4-9877-92857c8c782b)

# Task 9 : Conditional Probability

[Data Analytics](https://dashboard.stige.in/index.php/courses/lms-data-analytics/) [Task 9 : Conditional Probability](https://dashboard.stige.in/index.php/lessons/task-9-conditional-probability/)

In Progress

**Conditional probability** is defined as the likelihood of an event or outcome occurring, **based on the occurrence of a previous event** or outcome. Conditional probability is calculated by multiplying the [probability](https://www.investopedia.com/terms/c/compound-probability.asp) of the preceding event by the updated probability of the succeeding, or conditional, event.

Click on the below button to open Udacity ‘ s Course topic .

[Conditional Probability](https://classroom.udacity.com/courses/st101/lessons/90e97220-79e3-4a50-842a-415aba0b364f/concepts/0f9a998e-a9fb-4375-8af8-0a6337eede1c)

# Task 10 : Bayes Rule

[Data Analytics](https://dashboard.stige.in/index.php/courses/lms-data-analytics/) [Task 10 : Bayes Rule](https://dashboard.stige.in/index.php/lessons/task-10-bayes-rule/)

In Progress

**Bayes’ theorem**, named after 18th-century British mathematician Thomas Bayes, is a mathematical formula for determining [conditional probability](https://www.investopedia.com/terms/c/conditional_probability.asp). Conditional probability is the likelihood of an outcome occurring, based on a previous outcome occurring. Bayes’ theorem **provides a way to revise existing predictions or theories** (update probabilities) given new or additional evidence.   
  
In finance, Bayes’ theorem can be used to rate the [risk](https://www.investopedia.com/terms/r/risk.asp) of lending money to potential borrowers.

Click on the below button to open Udacity ‘ s Course topic .

[Bayes Rule](https://classroom.udacity.com/courses/st101/lessons/35a20143-6b94-4728-b64b-73439ec955a0/concepts/34fb5fc9-1172-4369-9b38-24467e487d93)

# Task 11 : Random Variable

## ****What Is a Random Variable?****

A **random variable** is a variable whose value is unknown or a function that assigns values to each of **an experiment’s outcomes. Random variables are often designated by letters and can be classified** as **discrete,** which are variables that have specific values, or continuous, which are variables that can have any values within a continuous range.

Random variables are often used in [econometric](https://www.investopedia.com/terms/e/econometrics.asp) or [regression](https://www.investopedia.com/terms/r/regression.asp) analysis to determine statistical relationships among one another.

### **KEY TAKEAWAYS**

* A random variable is a variable whose value is unknown or a function that assigns values to each of an experiment’s outcomes.
* A random variable can be either discrete (having specific values) or continuous (any value in a continuous range).
* The use of random variables is most common in probability and statistics, where they are used to quantify outcomes of random occurrences.
* Risk analysts use random variables to estimate the probability of an adverse event occurring.

Material - <https://www.investopedia.com/terms/r/random-variable.asp>

# Understanding Random Variable

[Data Analytics](https://dashboard.stige.in/index.php/courses/lms-data-analytics/) [Task 11 : Random Variable](https://dashboard.stige.in/index.php/lessons/task-11-random-variable/) [Understanding Random Variable](https://dashboard.stige.in/index.php/topic/understanding-random-variable/)

In Progress

In probability and [statistics](https://www.investopedia.com/terms/s/statistics.asp), **random variables** are used to **quantify outcomes** of a random occurrence, and therefore, can take on many values. Random variables are required to be **measurable** and are typically **real numbers**. For example, the letter X may be designated to represent the sum of the resulting numbers after three dice are rolled. In this case, X could be 3 (1 + 1+ 1), 18 (6 + 6 + 6), or somewhere between 3 and 18, since the highest number of a die is 6 and the lowest number is 1.

A random variable is **different from an**[**algebraic**](https://www.investopedia.com/terms/a/algebraic-method.asp)**variable**. The variable in an algebraic equation is an **unknown value** that can be calculated. The equation 10 + x = 13 shows that we can calculate the specific value for x which is 3. On the other hand, a random variable has a set of values, and any of those values could be the resulting outcome as seen in the example of the dice above.https://5672ccf87791671747694b1a6b5dd546.safeframe.googlesyndication.com/safeframe/1-0-38/html/container.html

In the corporate world, random variables can be assigned to properties such as the average price of an asset over a given time period, the [return on investment](https://www.investopedia.com/terms/r/returnoninvestment.asp) after a specified number of years, the estimated turnover rate at a company within the following six months, etc. Risk analysts assign random variables to risk models when they want to estimate the probability of an adverse event occurring. These variables are presented using tools such as scenario and [sensitivity analysis](https://www.investopedia.com/terms/s/sensitivityanalysis.asp) tables which risk managers use to make decisions concerning risk mitigation.

Video - <https://youtu.be/3v9w79NhsfI>

# Types of Random Variables

[Data Analytics](https://dashboard.stige.in/index.php/courses/lms-data-analytics/) [Task 11 : Random Variable](https://dashboard.stige.in/index.php/lessons/task-11-random-variable/) [Types of Random Variables](https://dashboard.stige.in/index.php/topic/types-of-random-variables/)

In Progress

A random variable can be either discrete or continuous. **Discrete random variables** take on a **countable number of distinct values**. Consider an experiment where a coin is tossed three times. If X represents the number of times that the coin comes up heads, then X is a discrete random variable that can only have the values 0, 1, 2, 3 (from no heads in three successive coin tosses to all heads). No other value is possible for X.

**Continuous random variables** can represent any **value within a specified range or interval** and can take on an **infinite** number of possible values. An example of a continuous random variable would be an experiment that involves measuring the amount of rainfall in a city over a year or the average height of a random group of 25 people.

Drawing on the latter, if Y represents the random variable for the average height of a random group of 25 people, you will find that the resulting outcome is a continuous figure since height may be 5 ft or 5.01 ft or 5.0001 ft. Clearly, there is an infinite number of possible values for height.

A random variable has a [**probability distribution**](https://www.investopedia.com/terms/p/probabilitydistribution.asp) that represents the **likelihood** that any of the possible values would occur. Let’s say that the random variable, Z, is the number on the top face of a die when it is rolled once. The possible values for Z will thus be 1, 2, 3, 4, 5, and 6. The probability of each of these values is 1/6 as they are all equally likely to be the value of Z.

For instance, the probability of getting a 3, or P (Z=3), when a die is thrown is 1/6, and so is the probability of having a 4 or a 2 or any other number on all six faces of a die. Note that the sum of all probabilities is 1.

Video - <https://youtu.be/dOr0NKyD31Q>

# Task 12 : Probability Distributions

A **probability distribution** is a statistical function that describes all the possible values and likelihoods that a [**random variable**](https://www.investopedia.com/terms/r/random-variable.asp)can take within a given range. This range will be bounded between the minimum and maximum possible values, but precisely where the possible value is likely to be plotted on the probability distribution depends on a number of factors.   
  
Academics, financial analysts and [fund managers](https://www.investopedia.com/terms/f/fundmanager.asp) alike may determine a particular stock’s probability distribution to evaluate the possible expected returns that the stock may yield in the future.

Click on the below button to open Udacity ‘ s Course topic .

[Probability Distributions](https://classroom.udacity.com/courses/st101/lessons/d009ab35-e56a-4610-9b35-f8ce8edea774/concepts/731c8d87-69fe-4423-8523-ea41bdaa6da2)

# Task 13 : More on Probability Distribution

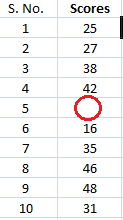
[Data Analytics](https://dashboard.stige.in/index.php/courses/lms-data-analytics/) [Task 13 : More on Probability Distribution](https://dashboard.stige.in/index.php/lessons/task-13-more-on-probability-distribution/)

In Progress

## ****Introduction****

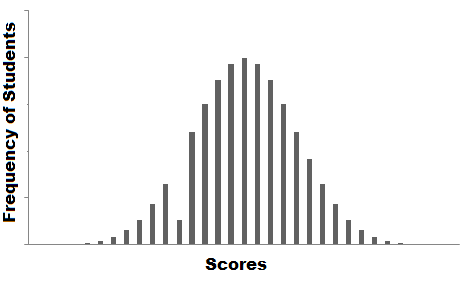
Welcome to the world of [Probability in Data Science](http://courses.analyticsvidhya.com/courses/introduction-to-data-science-2?utm_source=blog&utm_medium=6ProbabilityDistributionsarticle)! Let me start things off with an intuitive example.

Suppose you are a teacher at a university. After checking assignments for a week, you graded all the students. You gave these graded papers to a data entry guy in the university and tell him to create a spreadsheet containing the grades of all the students. But the guy only stores the grades and not the corresponding students.



He made another blunder, he missed a couple of entries in a hurry and we have no idea whose grades are missing. Let’s find a way to solve this.

One way is that you visualize the grades and see if you can find a trend in the data.



The graph that you have plot is called the [frequency distribution](http://courses.analyticsvidhya.com/courses/introduction-to-data-science-2?utm_source=blog&utm_medium=6ProbabilityDistributionsarticle) of the data. You see that there is a smooth curve like structure that defines our data, but do you notice an anomaly? We have an abnormally low frequency at a particular score range. So the best guess would be to have missing values that remove the dent in the distribution.

This is how you would try to solve a real-life problem using data analysis. For any Data Scientist, a student or a practitioner, distribution is a must know concept. It provides the basis for analytics and inferential statistics.

While the concept of probability gives us the mathematical calculations, distributions help us actually visualize what’s happening underneath.

In this article, I have covered some important [probability distributions](http://courses.analyticsvidhya.com/courses/introduction-to-data-science-2?utm_source=blog&utm_medium=6ProbabilityDistributionsarticle) which are explained in a lucid as well as comprehensive manner.

Note: This article assumes you have a basic knowledge of probability. If not, you can refer this [probability distributions](https://www.analyticsvidhya.com/blog/2017/02/basic-probability-data-science-with-examples/?utm_source=blog&utm_medium=6ProbabilityDistributionsarticle).

## ****Table of Contents****

1. Common Data Types
2. Types of Distributions
   1. Bernoulli Distribution
   2. Uniform Distribution
   3. Binomial Distribution
   4. Normal Distribution
   5. Poisson Distribution
   6. Exponential Distribution
3. Relations between the Distributions
4. Test your Knowledge!

## ****Common Data Types****

Before we jump on to the explanation of distributions, let’s see what kind of data can we encounter. The data can be discrete or continuous.

**Discrete Data**, as the name suggests, can take only specified values. For example, when you roll a die, the possible outcomes are 1, 2, 3, 4, 5 or 6 and not 1.5 or 2.45.

**Continuous Data** can take any value within a given range. The range may be finite or infinite. For example, A girl’s weight or height, the length of the road. The weight of a girl can be any value from 54 kgs, or 54.5 kgs, or 54.5436kgs.

Now let us start with the types of distributions.

## ****Types of Distributions****

### Bernoulli Distribution

Let’s start with the easiest distribution that is Bernoulli Distribution. It is actually easier to understand than it sounds!

All you cricket junkies out there! At the beginning of any cricket match, how do you decide who is going to bat or ball? A toss! It all depends on whether you win or lose the toss, right? Let’s say if the toss results in a head, you win. Else, you lose. There’s no midway.

A **Bernoulli distribution** has only two possible outcomes, namely 1 (success) and 0 (failure), and a single trial. So the random variable X which has a Bernoulli distribution can take value 1 with the probability of success, say p, and the value 0 with the probability of failure, say q or 1-p.

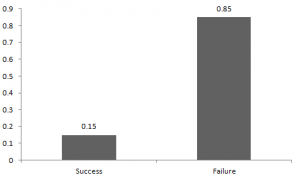
Here, the occurrence of a head denotes success, and the occurrence of a tail denotes failure.  
Probability of getting a head = 0.5 = Probability of getting a tail since there are only two possible outcomes.

The probability mass function is given by: px(1-p)1-x  where x € (0, 1).  
It can also be written as



The probabilities of success and failure need not be equally likely, like the result of a fight between me and Undertaker. He is pretty much certain to win. So in this case probability of my success is 0.15 while my failure is 0.85

Here, the probability of success(p) is not same as the probability of failure. So, the chart below shows the Bernoulli Distribution of our fight.



Here, the probability of success = 0.15 and probability of failure = 0.85. The expected value is exactly what it sounds. If I punch you, I may expect you to punch me back. Basically expected value of any distribution is the mean of the distribution. The expected value of a random variable X from a Bernoulli distribution is found as follows:

E(X) = 1\*p + 0\*(1-p) = p

The variance of a random variable from a bernoulli distribution is:

V(X) = E(X²) – [E(X)]² = p – p² = p(1-p)

There are many examples of Bernoulli distribution such as whether it’s going to rain tomorrow or not where rain denotes success and no rain denotes failure and Winning (success) or losing (failure) the game.

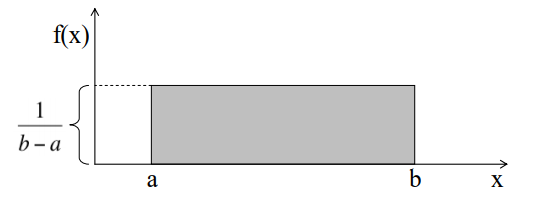
### Uniform Distribution

When you roll a fair die, the outcomes are 1 to 6. The probabilities of getting these outcomes are equally likely and that is the basis of a uniform distribution. Unlike Bernoulli Distribution, all the n number of possible outcomes of a uniform distribution are equally likely.

A variable X is said to be uniformly distributed if the density function is:



The graph of a uniform distribution curve looks like



You can see that the shape of the Uniform distribution curve is rectangular, the reason why Uniform distribution is called rectangular distribution.

For a Uniform Distribution, a and b are the parameters.

The number of bouquets sold daily at a flower shop is uniformly distributed with a maximum of 40 and a minimum of 10.

Let’s try calculating the probability that the daily sales will fall between 15 and 30.

The probability that daily sales will fall between 15 and 30 is (30-15)\*(1/(40-10)) = 0.5

Similarly, the probability that daily sales are greater than 20 is  = 0.667

The mean and variance of X following a uniform distribution is:

Mean -> E(X) = (a+b)/2

Variance -> V(X) =  (b-a)²/12

The standard uniform density has parameters a = 0 and b = 1, so the PDF for standard uniform density is given by:



### Binomial Distribution

Let’s get back to cricket.  Suppose that you won the toss today and this indicates a successful event. You toss again but you lost this time. If you win a toss today, this does not necessitate that you will win the toss tomorrow. Let’s assign a random variable, say X, to the number of times you won the toss. What can be the possible value of X? It can be any number depending on the number of times you tossed a coin.

There are only two possible outcomes. Head denoting success and tail denoting failure. Therefore, probability of getting a head = 0.5 and the probability of failure can be easily computed as: q = 1- p = 0.5.

A distribution where only two outcomes are possible, such as success or failure, gain or loss, win or lose and where the probability of success and failure is same for all the trials is called a Binomial Distribution.

The outcomes need not be equally likely. Remember the example of a fight between me and Undertaker? So, if the probability of success in an experiment is 0.2 then the probability of failure can be easily computed as q = 1 – 0.2 = 0.8.

Each trial is independent since the outcome of the previous toss doesn’t determine or affect the outcome of the current toss. An experiment with only two possible outcomes repeated n number of times is called binomial. The parameters of a binomial distribution are n and p where n is the total number of trials and p is the probability of success in each trial.

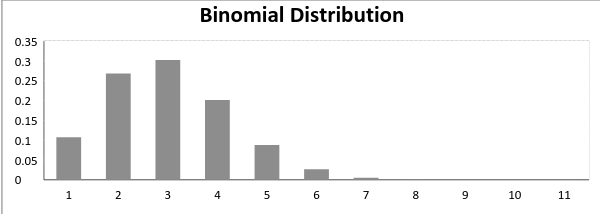
On the basis of the above explanation, the properties of a Binomial Distribution are

1. Each trial is independent.
2. There are only two possible outcomes in a trial- either a success or a failure.
3. A total number of n identical trials are conducted.
4. The probability of success and failure is same for all trials. (Trials are identical.)

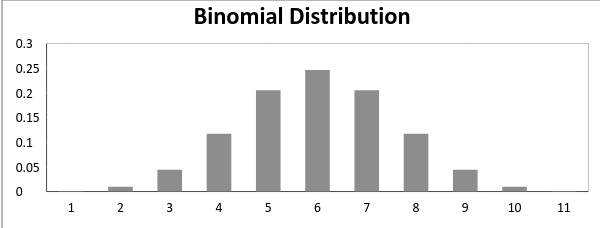
The mathematical representation of binomial distribution is given by:



A binomial distribution graph where the probability of success does not equal the probability of failure looks like



Now, when probability of success = probability of failure, in such a situation the graph of binomial distribution looks like



The mean and variance of a binomial distribution are given by:

Mean -> µ = n\*p

Variance -> Var(X) = n\*p\*q

### Normal Distribution

**Normal distribution** represents the behavior of most of the situations in the universe (That is why it’s called a “normal” distribution. I guess!). The large sum of (small) random variables often turns out to be normally distributed, contributing to its widespread application. Any distribution is known as Normal distribution if it has the following characteristics:

1. The mean, median and mode of the distribution coincide.
2. The curve of the distribution is bell-shaped and symmetrical about the line x=μ.
3. The total area under the curve is 1.
4. Exactly half of the values are to the left of the center and the other half to the right.

A normal distribution is highly different from Binomial Distribution. However, if the number of trials approaches infinity then the shapes will be quite similar.

The PDF of a random variable X following a normal distribution is given by:

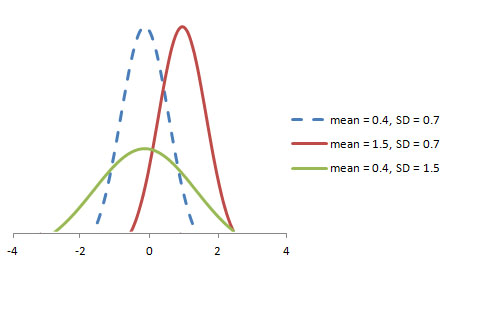


The mean and variance of a random variable X which is said to be normally distributed is given by:

Mean -> E(X) = µ

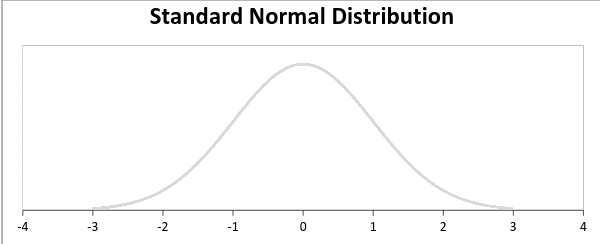
Variance -> Var(X) = σ^2

Here, µ (mean) and σ (standard deviation) are the parameters.  
The graph of a random variable X ~ N (µ, σ) is shown below.



A standard normal distribution is defined as the distribution with mean 0 and standard deviation 1.  For such a case, the PDF becomes:





### Poisson Distribution

Suppose you work at a call center, approximately how many calls do you get in a day? It can be any number. Now, the entire number of calls at a call center in a day is modeled by Poisson distribution. Some more examples are

1. The number of emergency calls recorded at a hospital in a day.
2. The number of thefts reported in an area on a day.
3. The number of customers arriving at a salon in an hour.
4. The number of suicides reported in a particular city.
5. The number of printing errors at each page of the book.

You can now think of many examples following the same course. Poisson Distribution is applicable in situations where events occur at random points of time and space wherein our interest lies only in the number of occurrences of the event.

A distribution is called **Poisson distribution** when the following assumptions are valid:

1. Any successful event should not influence the outcome of another successful event.  
2. The probability of success over a short interval must equal the probability of success over a longer interval.  
3. The probability of success in an interval approaches zero as the interval becomes smaller.

Now, if any distribution validates the above assumptions then it is a Poisson distribution. Some notations used in Poisson distribution are:

* λ is the rate at which an event occurs,
* t is the length of a time interval,
* And X is the number of events in that time interval.

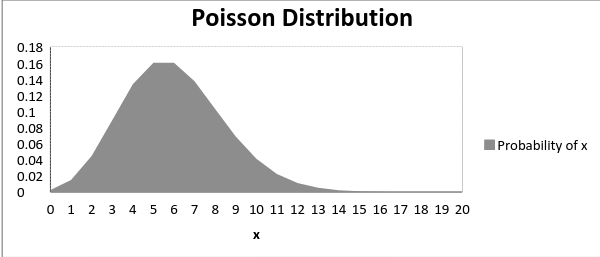
Here, X is called a Poisson Random Variable and the probability distribution of X is called Poisson distribution.

Let µ denote the mean number of events in an interval of length t. Then, µ = λ\*t.

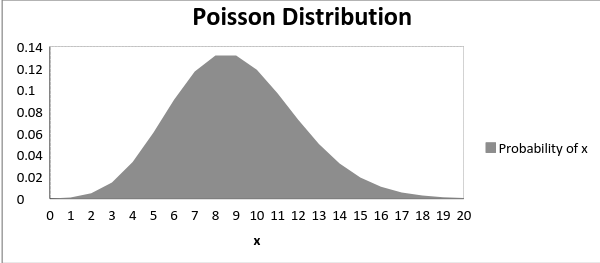
The PMF of X following a Poisson distribution is given by:



The mean µ is the parameter of this distribution. µ is also defined as the λ times length of that interval. The graph of a Poisson distribution is shown below:



The graph shown below illustrates the shift in the curve due to increase in mean.



It is perceptible that as the mean increases, the curve shifts to the right.

The mean and variance of X following a Poisson distribution:

Mean -> E(X) = µ  
Variance -> Var(X) = µ

### Exponential Distribution

Let’s consider the call center example one more time. What about the interval of time between the calls ? Here, exponential distribution comes to our rescue. Exponential distribution models the interval of time between the calls.

Other examples are:

1. Length of time beteeen metro arrivals,  
2. Length of time between arrivals at a gas station  
3. The life of an Air Conditioner

Exponential distribution is widely used for survival analysis. From the expected life of a machine to the expected life of a human, exponential distribution successfully delivers the result.

A random variable X is said to have an **exponential distribution** with PDF:

f(x) = { λe-λx,  x ≥ 0

and parameter **λ>0** which is also called the rate.

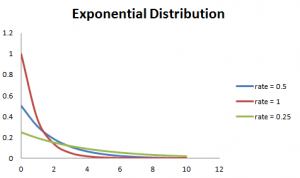
For survival analysis, λ is called the failure rate of a device at any time t, given that it has survived up to t.

Mean and Variance of a random variable X following an exponential distribution:

Mean -> E(X) = 1/λ

Variance -> Var(X) = (1/λ)²

Also, the greater the rate, the faster the curve drops and the lower the rate, flatter the curve. This is explained better with the graph shown below.



To ease the computation, there are some formulas given below.  
P{X≤x} = 1 – e-λx, corresponds to the area under the density curve to the left of x.

P{X>x} = e-λx, corresponds to the area under the density curve to the right of x.

P{x1<X≤ x2} = e-λx1 – e-λx2, corresponds to the area under the density curve between x1 and x2.

## ****Relations between the Distributions****

### Relation between Bernoulli and Binomial Distribution

1. Bernoulli Distribution is a special case of Binomial Distribution with a single trial.

2. There are only two possible outcomes of a Bernoulli and Binomial distribution, namely success and failure.

3. Both Bernoulli and Binomial Distributions have independent trails.

### Relation between Poisson and Binomial Distribution

Poisson Distribution is a limiting case of binomial distribution under the following conditions:

1. The number of trials is indefinitely large or n → ∞.
2. The probability of success for each trial is same and indefinitely small or p →0.
3. np = λ, is finite.

### Relation between Normal and Binomial Distribution & Normal and Poisson Distribution:

Normal distribution is another limiting form of binomial distribution under the following conditions:

1. The number of trials is indefinitely large, n → ∞.
2. Both p and q are not indefinitely small.

The normal distribution is also a limiting case of Poisson distribution with the parameter λ →∞.

### Relation between Exponential and Poisson Distribution:

If the times between random events follow exponential distribution with rate λ, then the total number of events in a time period of length t follows the Poisson distribution with parameter λt.

## ****Test your knowledge****

You have come this far. Now, are you able to answer the following questions? Let me know in the comments below!

1. The formula to calculate standard normal random variable is:

a. (x+µ) / σ  
b. (x-µ) / σ  
c. (x-σ) / µ

2. In Bernoulli Distribution, the formula for calculating standard deviation is given by:

a. p (1 – p)  
b. SQRT(p(p – 1))  
c. SQRT(p(1 – p))

3. For a normal distribution, an increase in the mean will:

a. shift the curve to the left  
b. shift the curve to the right  
c. flatten the curve

4. The lifetime of a battery is exponentially distributed with λ = 0.05 per hour. The probability for a battery to last between 10 and 15 hours is:

a.0.1341  
b.0.1540  
c.0.0079

## End Notes

Probability Distributions are prevalent in many sectors, namely, insurance, physics, engineering, computer science and even social science wherein the students of psychology and medical are widely using probability distributions. It has an easy application and widespread use. This article highlighted six important distributions which are observed in day-to-day life and explained their application. Now you will be able to identify, relate and differentiate among these distributions.

If you have any doubts and want to see more articles on distributions, please do write in the comment section below. For a more in-depth write up of these distributions, you can [refer this resource](http://www.math.wm.edu/~leemis/chart/UDR/UDR.html).

# Task 14 : Additional Resources on Probability

[Khan Academy](https://www.khanacademy.org/math/statistics-probability/probability-library)

**Task 16 : Mini Project**

[Data Analytics](https://dashboard.stige.in/index.php/courses/lms-data-analytics/) [Task 16 : Mini Project](https://dashboard.stige.in/index.php/lessons/task-16-mini-project/)

In Progress

Kudos !!!! You have completed half of the course and now is the time for an interesting case study .

You are given a dataset of a company where the details about people who registered for a particular post in a department of this company. You are required to use your knowledge in statistics and use different formulas in excel and draw necessary conclusions about the company .

Use the below **Steps for EDA** –

1. Understanding data columns and data
2. Check for missing data
3. Clubbing columns with multiple categories
4. Check for outliers
5. Removing outliers

Drawing Data Summary

**Answer the following Questions Based on your Analysis** –

**Q1** How many males and females are Hired ?  
**Q2** What is the average salary offered in this company ?  
**Q3** Draw the class intervals for salary in the company ?  
**Q4** Draw Pie Chart / Bar Graph ( or any other graph ) to show proportion of people working different department ?  
**Q5** Represent different post tiers using chart/graph ?

[Dataset](https://drive.google.com/file/d/10snp-2pEnDc4Wd2r1esw_rsMjRDcwYoX/view?usp=sharing)

You are required to **use Excel** to answer the above questions at the earliest. Assignment should be submitted in pdf format in the assignment tab.