**Week 1: Readings of Chapter 1**

Lesson 1-1 — Terms to Know

We start our study of statistics with some basic terms that you should know.

* **Parameter:** A characteristic of a population (often, a numerical characteristic such as a population mean, a population variance, a population standard deviation, etc.)
* **Statistic:** A characteristic of a sample (such as a sample mean or a standard deviation, etc.)
* **Data:** A collection of information (literally, data is the plural of datum; meaning: what is given)

**Data Types:**

There are two types of data:

1. **Categorical (Qualitative):**

* **Nominal:** According to Name

Examples: Data containing names, genders, races, etc.

* **Ordinal:** According to Order

Examples: Data containing ranks, data that has been organized alphabetically, etc.

1. **Numerical (Quantitative):**

* According to the ratio scale (a possible value of zero in the data is an inherent zero)

Examples: Data containing heights, weights, time durations, grades, etc.

* According to the interval scale (a zero is not inherently zero)

Example: Data containing temperatures

Statistically, a numerical set of data may be discrete or continuous.

* A **discrete** data set is one in which the measurements take a countable set of isolated values. For example, the number of chairs, the number of patients, the number of accidents, etc., are all examples of discrete data.
* A **continuous** data set is one in which the measurements can take any real value within a certain range. For example, the amount of rainfall in Charlotte in January during the last 30 years or the amount of customer waiting times at a local bank are examples of continuous data sets.

Lesson 1-2 — Types of Statistics

There are two main types of statistics, descriptive and inferential.

**Descriptive statistics** is used to describe a set of data graphically or numerically

* **Graphical Descriptive Statistics:** Describes a set of data graphically by creating bar graphs, pie charts, histograms, line plots, scatter plots, etc.
* **Numerical Descriptive Statistics:** There are a number of particular characteristics of data that are often the focus of interest to the data analyst.

These are:

* Measures describing the center of data: Examples of such measures are: mean (arithmetic average), median. mode, and the weighted mean
* Measures describing the variability (spread or dispersion) of data: Examples of these types of measures are: the **range**, the **variance**, and the **standard deviation** of data
* Measures of location: Examples of such measures are the percentile ranking and the z-score. These measures describe where a particular measurement stands compared to the rest of the data.

Measures describing the shape of the distribution of data

Skewness and kurtosis are two measures that describe the shape of the distribution of a data set

**Inferential statistics** is the process of utilizing one or more random samples in order to gain insight about the population from which those samples were selected.

* A **sample** is a part or a portion of a population.
* A **population** is the set of all individuals, objects, or measurements that are of interest.
* A **variable** is a characteristic of a population, which is of interest. A variable may be quantitative or qualitative.

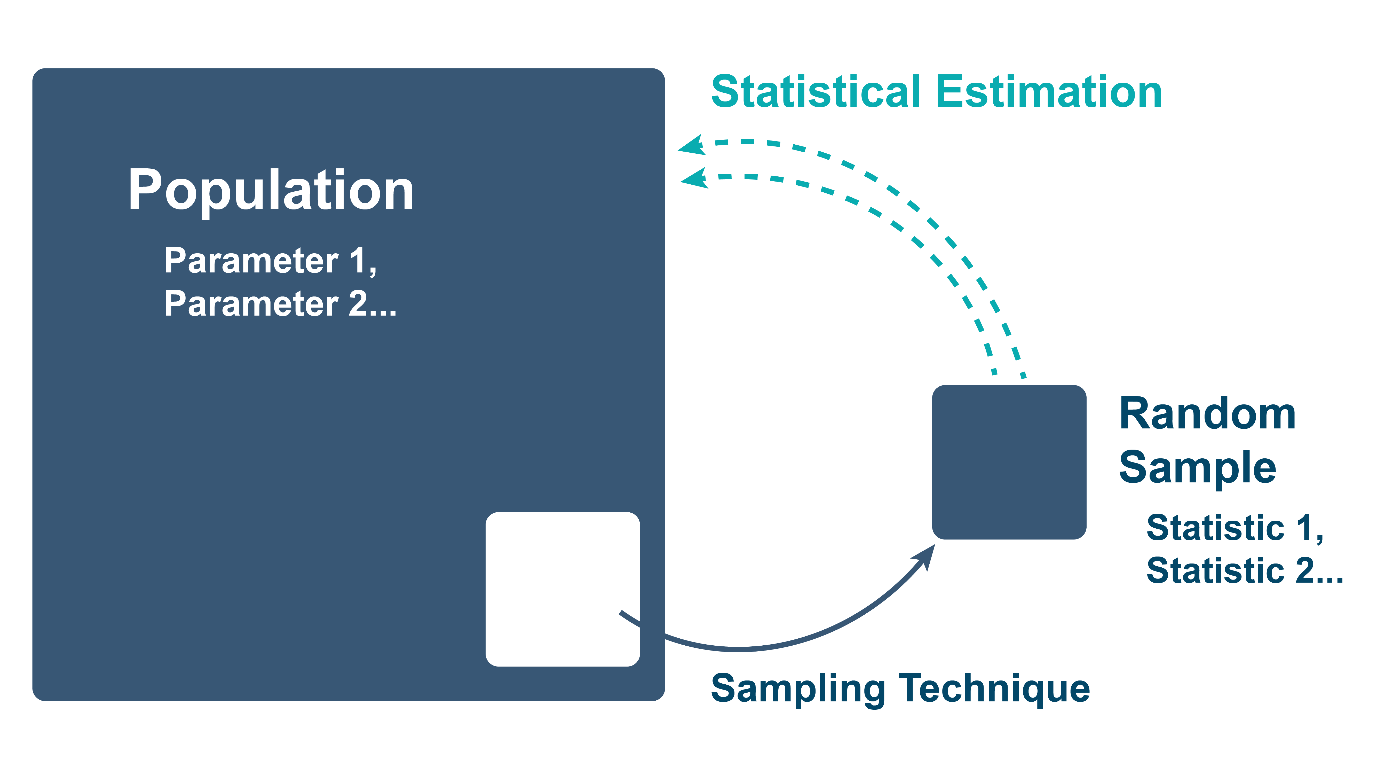
Often a data analyst may be interested in obtaining information about one or more particular parameters of a given population. However, since the entire population is not accessible in the majority of situations, the data analyst must select one or more samples from the population of interest and perform statistical analysis on these samples. Once sample characteristics have been verified or revealed, the analyst will then use the methods of inferential statistics to transform the sample information into population information.

There are three main methods of inferential statistics:

* Constructing Confidence Intervals: This is to estimate a population parameter to within two limits: a lower limit and an upper limit
* Performing Hypothesis Testing: This is to verify or to reject hypotheses or claims
* Modeling or Testing Relationships between Data sets

Lesson 1-3 — Samples, Sample Statistics and Estimation

**Samples**

This image shows that a random sample is a subset of the populations which can then be used to make estimations about the population.

A sample is a subset of elements from the set of individuals with one or more common features, known as the population, which has been selected for the study. The number of elements in a sample is denoted by n.

* n: number of elements in a sample
* N: number of elements in the population
* n < N

Samples are necessary to learn about populations, because in most real-world examples, it is impossible to measure a characteristic from every member of a population.

Here are some examples of large populations from which it would be too difficult to measure characteristics:

* The population of humans on Earth is more than 7 billion
* The population of the US is more than 300 million
* There are 1.8 billion bottles of Coca-Cola sold each day
* There may be more than 1 million pigeons in New York City alone

**Random Sampling**

A random sample is defined as a subset of a population, where the subset is chosen in such a way that every member of the population has an equal chance of being selected for the subset or sample. The number of members of a random sample is also denoted by n.

Ideally, we want samples to be representative of the populations from which they were selected. If appropriate sampling techniques are used to generate the sample, then the center, shape and spread of the population and sample distributions should be similar for any measured characteristic. This can be accomplished through random sampling techniques.

Consider the following example to illustrate the concepts of random sampling and representative samples:

population:

* {1, 2, 3, 4, 5}
* N=5
* n=2

10 possible **random samples** without replacement

* {1, 2,} {1, 3,} {1, 4}, {1, 5}, {2, 3}, {2, 4}, {2, 5}, {3, 4}, {3, 5}, {4, 5}
* The population is integers 1 through 5.

We collect a random sample without replacement from the population of size n equals two. There are ten distinct random samples that can be drawn from this population, and they are each equally likely.

Even characteristics of a small population, such as all students at a particular university would be difficult to measure because these students are rarely all in the same place at the same time. Measuring a characteristic of every student would be time intensive and costly. It would require quite a bit of organization and persistence. The number of elements in a **sample**, n, is often much less than that of the **population** from which it was selected, N, therefore measurements from a sample are easier to maintain and often more computationally manageable than measurements from an entire population.

**Examples of samples include:**

* 1,000 girls who run high school track
* 15 pigeons from New York City
* 200 dentists
* 10 state governors
* 50 members of the US Congress

**Statistics**

A **statistic** is a numerical value summarizing a measurable characteristic of a sample or subset of a set of population elements. Sample mean, sample median and sample standard deviation are examples of statistics.

If we want to know the average height for male freshmen at Northeastern University, an easily measurable characteristic, we could select, at random, 100 male freshmen to be measured at the mandatory orientation presentation. These 100 male freshmen are the random sample. The sample statistic is the average height of the 100 male freshmen who were measured. The entire set, containing all of the male freshmen at Northeastern University is the population. The **parameter** is the average height for all the male freshmen at Northeastern University.

**Statistical Estimation**

In inferential statistical analysis, we use samples to make generalizations about the populations from which they were selected. Statistical estimation is a specific type of inferential statistical analysis where we use statistics calculated from data measured from random samples to estimate population parameters. Statistical estimation is also used to quantify the uncertainty in these estimates.

Consider we are tasked with predicting the percentage of voters who will vote for a particular candidate. As a data analyst, we conduct a poll where we select a random sample from the set of all voters. Each voter has an equal chance of being selected for the sample. We ask each member of the sample who they plan to vote for in the election, and also how likely they are to vote in this election. We use this to calculate the percentage of voters who will vote for a particular candidate. This percentage is the **sample statistic**.

We then use this percentage, the sample statistic, to estimate the percentage of the entire population who will vote for the candidate in question. Because we are not getting information from all voters, our result can only be called a statistical estimate of a population parameter. It is not the true percentage of the population that will vote for the candidate.

**Statistical Notation**

We will discuss the specific values in a future module, but for now, it will be helpful to know that sample statistics and population statistics have different mathematical representations. Here are some of the notations you will see:

* **X¯**: sample mean
* **μ**: population mean
* **s2**: variance of a sample
* **σ2** (sigma squared): variance of a population
* **s**: standard deviation of a sample
* **σ**: standard deviation of the population

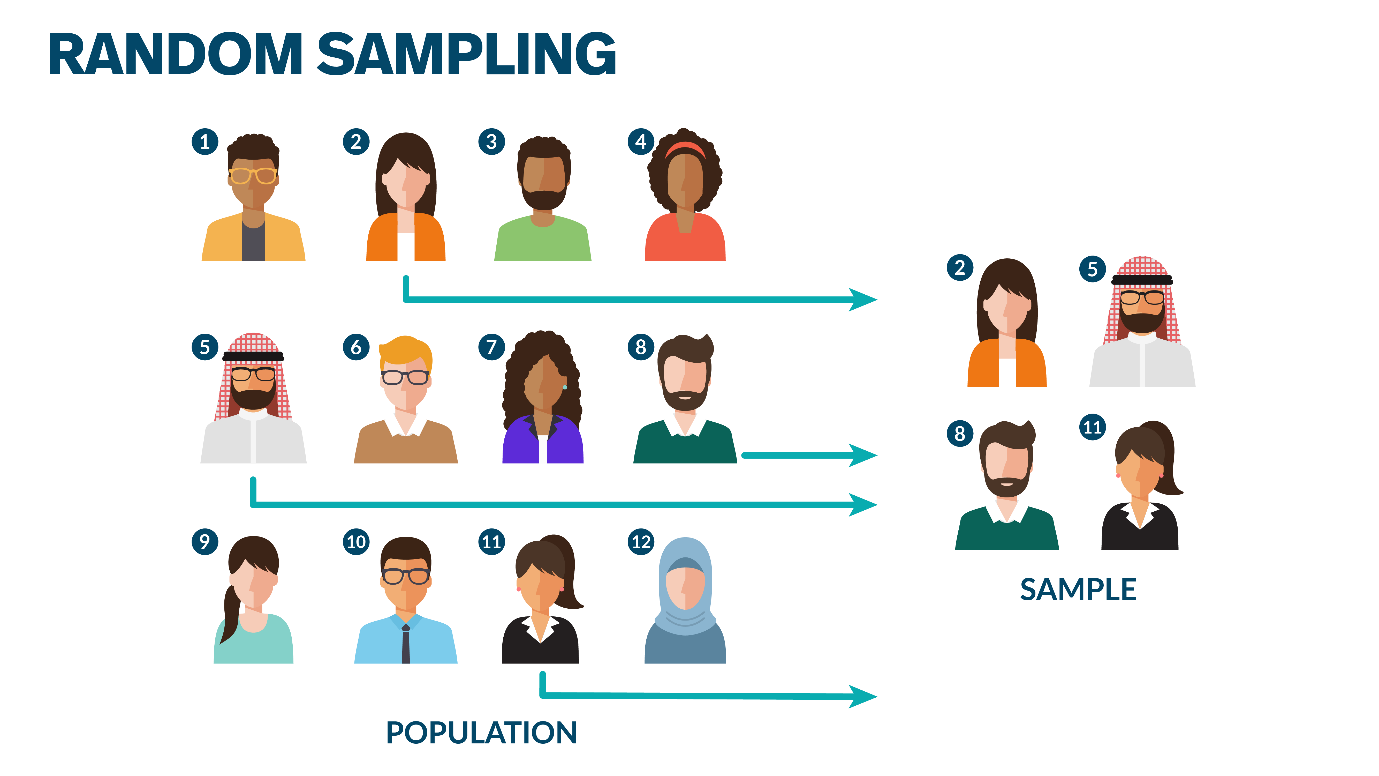
Lesson 1-4 — Sampling Methods

Selecting an appropriate sampling method is important to ensure that our sample is representative of the population. Three methods for choosing a random sample include a simple random sample, a stratified random sample and a cluster sample.

**Simple Random Sampling**

This image demonstrates how a random sample is selected from a population.

Consider the yearly census data that is collected. Every ten years, the US government does a full census of the American population. In other years, they do sampling of American households. If they do a simple random sample, they will randomly choose households from across the nation.

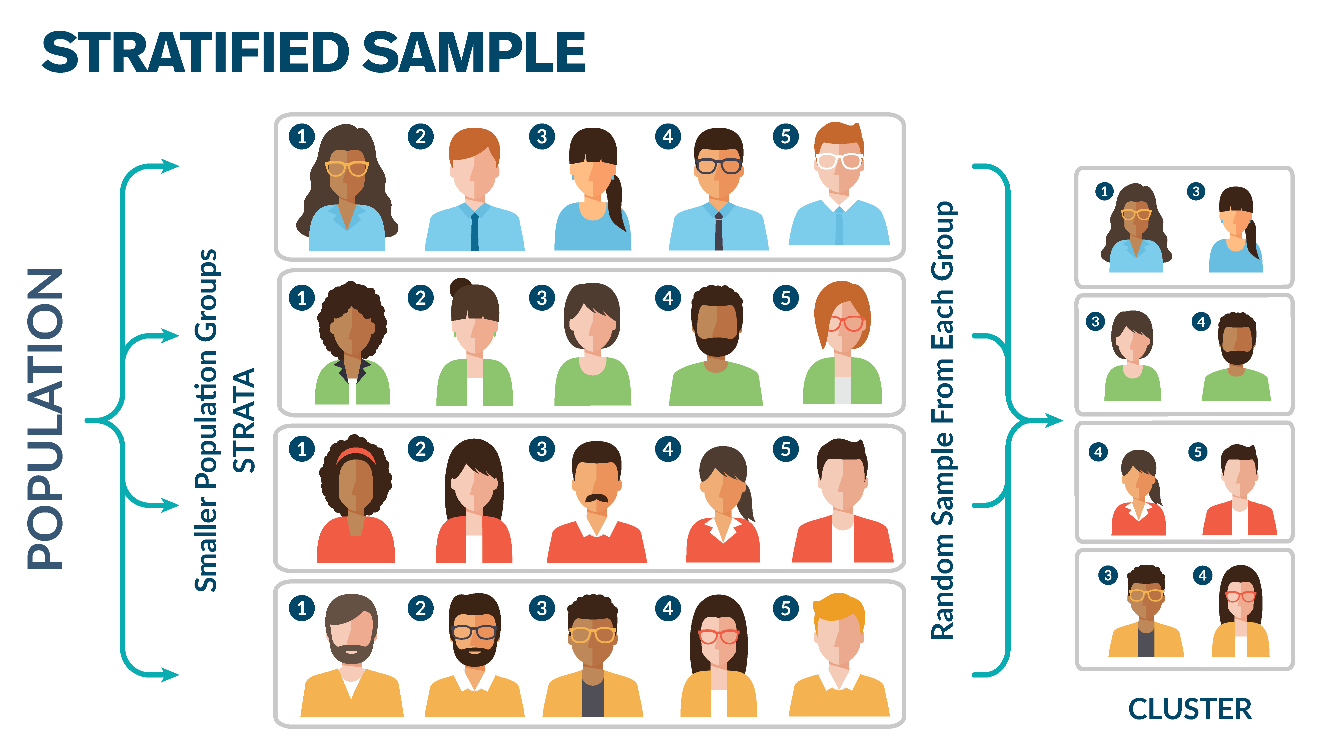
A simple random sampling method refers to any sampling method that consists of a population with N members or elements, a sample set with n members or elements, where each possible sample of n members is equally likely to occur.

In Other Words, ... In Simple Random Sampling

* Population of size N
* Sample size of n
* Each possible sample of n is equally likely to occur

**Stratified Sampling**

If the Census Bureau wants to gather data from different groups, or strata, they will use a stratified sample. For example, they may choose households based on income, or on whether they rent or own their homes, etc. In stratified sampling, you divide the population into separate groups, called strata, and then do a simple random sample from each stratum.



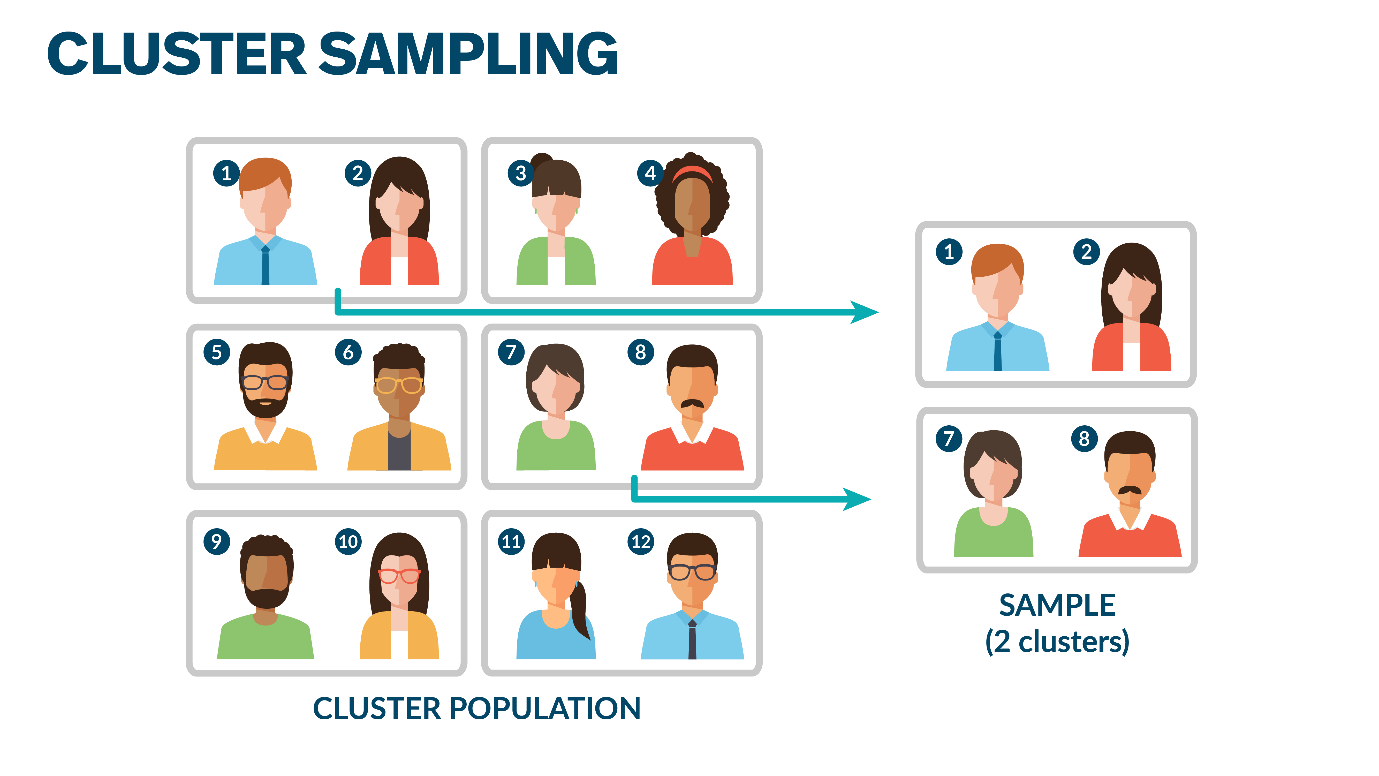
Stratified sampling has several advantages over simple random sampling. It may be possible to reduce the sample size required to achieve a given precision for a given level of confidence. It may be possible to increase the precision for a given confidence level with the same sample size.

In Other Words, ... In Stratified Sampling

* Population N is divided into H strata by stratification
* Each population member is assigned to exactly one stratum

**Cluster Sampling**

This image demonstrates how a population is clustered and a random sample is selected from the clustered population becomes a sample.



Again, a census taker may have to select households within a state. However, they cannot go to every household so they may select homes in a particular town or neighborhood. In cluster sampling, the population is divided into separate groups or clusters, and then a set of these clusters, known as a superset, is randomly selected as the final sample.

Cluster sampling performs better when the samples are heterogeneous for a particular characteristic of interest, and where each cluster is representative of the population across the characteristic of interest.

In Other Words, ... In Cluster Sampling

* Each population member is assigned to exactly one cluster
* Clusters should be heterogeneous, representative of population for characteristic of interest
* In general, greater variability and less precision that other random sampling methods
* Can be significantly cheaper than other random sampling methods

Lesson 1-5 — Additional Video Resources

Here are some additional resources that will help you gain a deeper understanding of the content and the use of R as a statistical tool.

* **Graphing- Plot Basic-mtcars**
* **Graphing- Plot Basic-mtcars2**
* **Data Management- R Scripts**
* **Basic Exploration of R**
* **Datasets-Dataframe Datasets-Factors**

**Week 2: Readings of Chapter 2**

Lesson 2-1 — Characteristics of Numerical Data

When describing numerical data, there are three characteristics we want to capture:

* the center,
* the variability or dispersion, and
* the shape.

The **measure of central tendency** of the distribution is a measure of where "the middle" is.

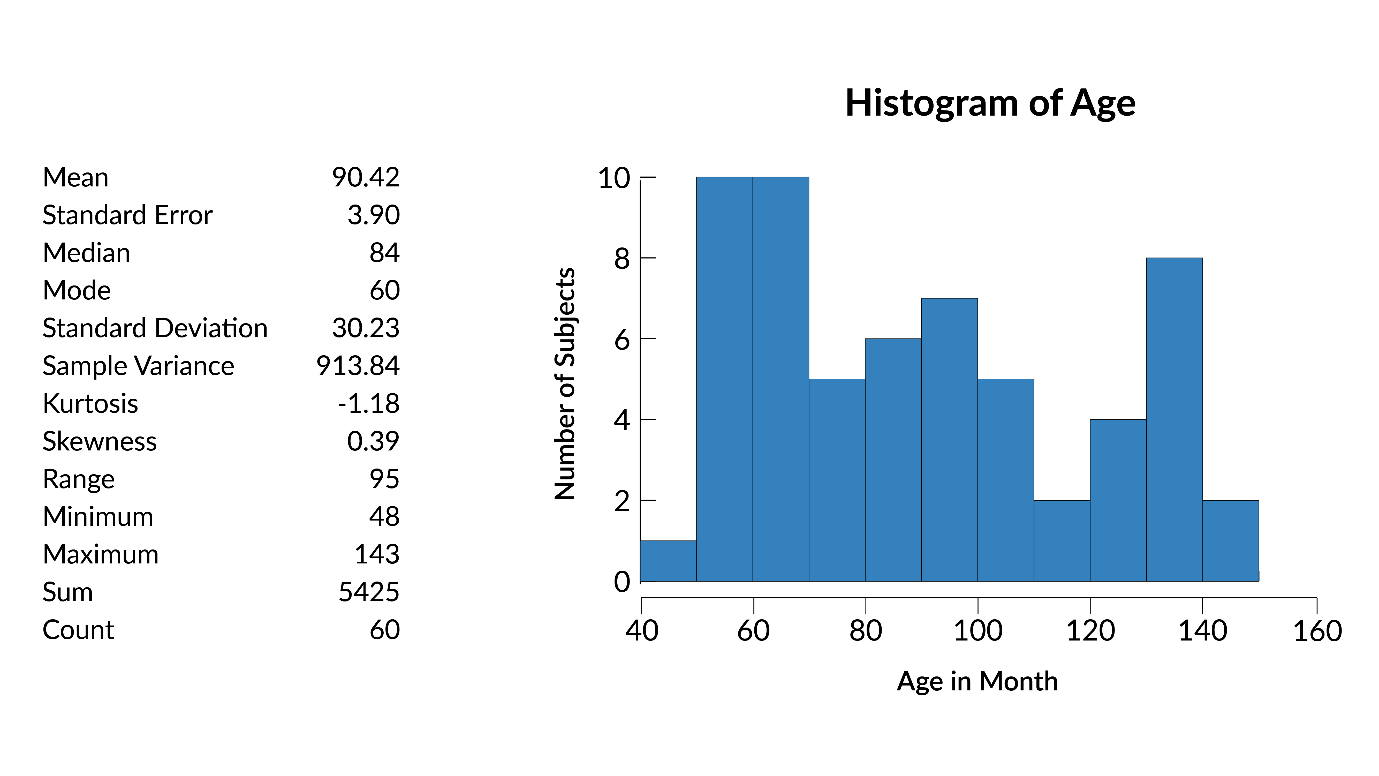
Another important metric is a **measure of the dispersion**, or of how spread out the data is. This is a measure of how far away the data is from the center.

The **shape of the data** is visually shown in the distribution curve. We consider how pointy the curve is near its peak, and the skewness of the curve or how asymmetric or lopsided the curve is.

Example: Descriptive Statistics

This table and the accompanying frequency chart display the descriptive statistics for the variable, Age, in a dataset.

Looking at the histogram, we see there are multiple peaks, one at each of 60, 90, and 130 months. We see that the mean is 90, which is consistent with what appears to be the center of the data in the frequency chart. The standard deviation is 30. We expect to see most of the data within one standard deviation of the mean, between 60 and 120.

The kurtosis and skewness indicate the pointy-ness and which way the data is leaning. We will learn more about these terms in this module. 

Lesson 2-2 — Measures of Central Tendency

The average, or the mean, is one of the most common indicators of central tendency, or of the central location of the data. It’s the sum of all the values divided by the number of values in the data set.

* n: number of data points or elements in dataset
* xi: sample data point i
* X ¯: mean of sample data (we discussed samples previously)
* X = 1/n ∑ i = 1 i = n   xi

Other types of means are geometric and harmonic, these are outside the scope of this class.

**Weighted Mean**

In some situations, the measurements in a set of data are different weights or are of different duplicities. In these types of situations, the mean is said to be a weighted mean.

Suppose that the course grade in a course is determined by the average of mid-term exam, final exam, and homework grades with the weights of 30%, 40% and 30% of the course grade respectively. Suppose a student’s grades are: 86 in the mid-term exam, 80 in the final, and 90 in the homework assignments. In this situation, the measurements have different weights. To find the average, we first multiply each measurement by the corresponding weight, and then divide the sum of all such products by the total number of weights:

Average Course Grade = 86(30) + 80(40) + 90(30) / 30 + 40 + 30

The weighted mean may also be used in situations where there are repeated measurements in a set of data. Suppose we wish to find the mean of the following 8 measurements: 2, 2, 2, 5, 5, 3, 3, 4. The measurement 2 has a weight of three, 5 and 3 each have a weight of two, and 4 has a weight of one. Therefore,

Average = 2(3) + 5(2) + 3(2) + 4(1) / 8

In general, the weighted mean of n measurements x1, x2, ..., xn with respective weights w1, w2, ..., wn is denoted by and is given by:

X-w and is given by: X-w = ∑wx / ∑w

**Median**

We have talked about the mean or average being one measure of the central tendency of a data set. The median is another common measure of the central tendency of a data set. It is simply the middle value of the sorted data set.

**To calculate the median:**

* M: median
* n: number of elements in data set
* n1: (n + 1/ 2) th element in sorted dataset
* n2: (n/2) th element in sorted dataset
* n3: (n/2 + 1) th element in sorted dataset
* if n is odd: M = n1
* if n is even: M = n2 + n3/ 2

If there are an odd number of values, the median is just the middle value of the sorted data set. If there are an even number of values, the median is the mean of the two middle values of the sorted data set. This means you add the two middle values together and divide by two.

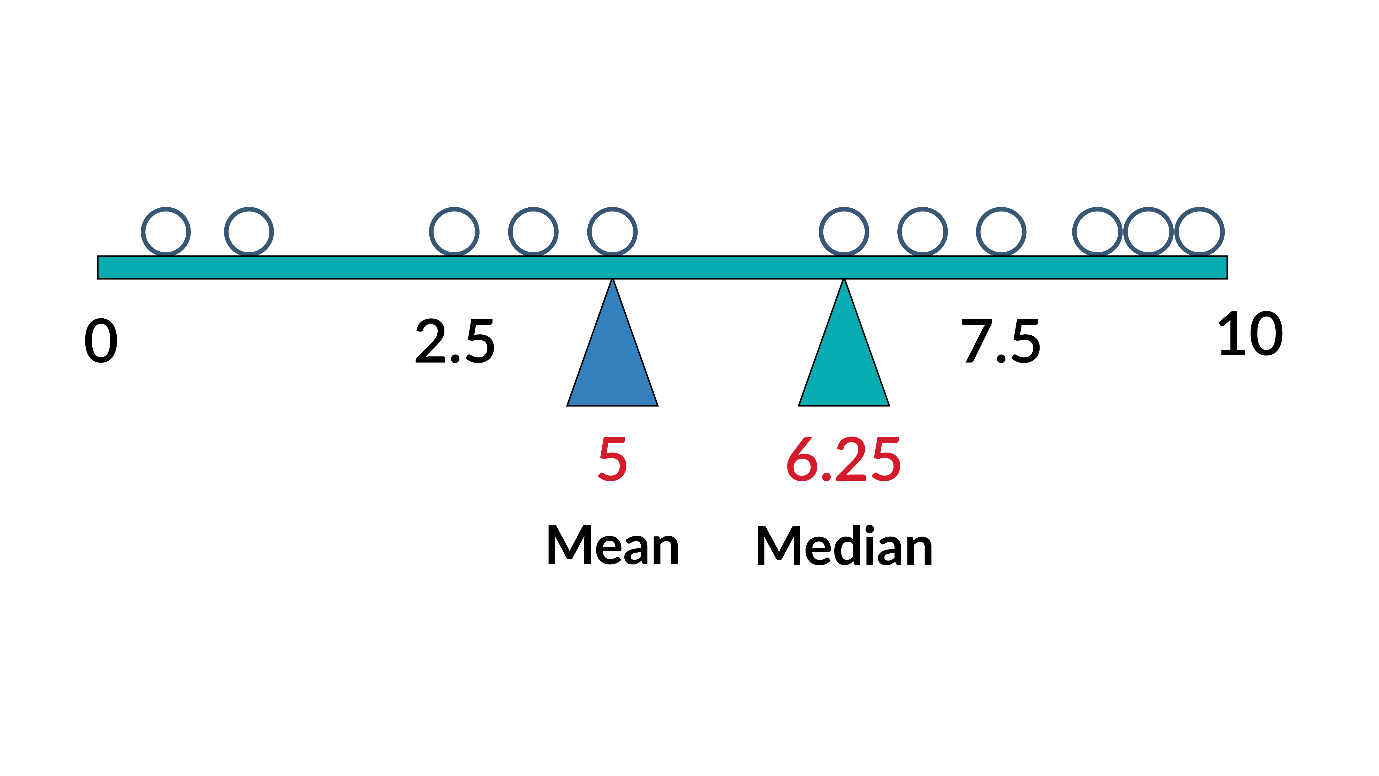
**For example:**

(45, 49, 50, 53, 60, 62, 63, 65, 66, **67, 69**, 71, 73, 74, 74, 78, 81, 85, 87, 100)

Median: 68

**Mean and Median:**

Both the mean and the median are useful in describing the central measure of a data set, but they are not always the same. The two values are usually different, and it’s up to us to determine the better measure to use. Sometimes it’s clear and sometimes it is less clear.



The graphic above shows an example where the mean and the median are different. Each ball represents a value between 0 and 10. The median is 6.25, which means half of the balls are greater than 6.25 and half of the balls are less than 6.25. The mean is 5, and the fact that the median is greater than the mean indicates that there are more data points on one side of the mean than the other.

The mean and median are often not equal and the difference between these values can tell us information about the shape of the distribution.

**Mode**

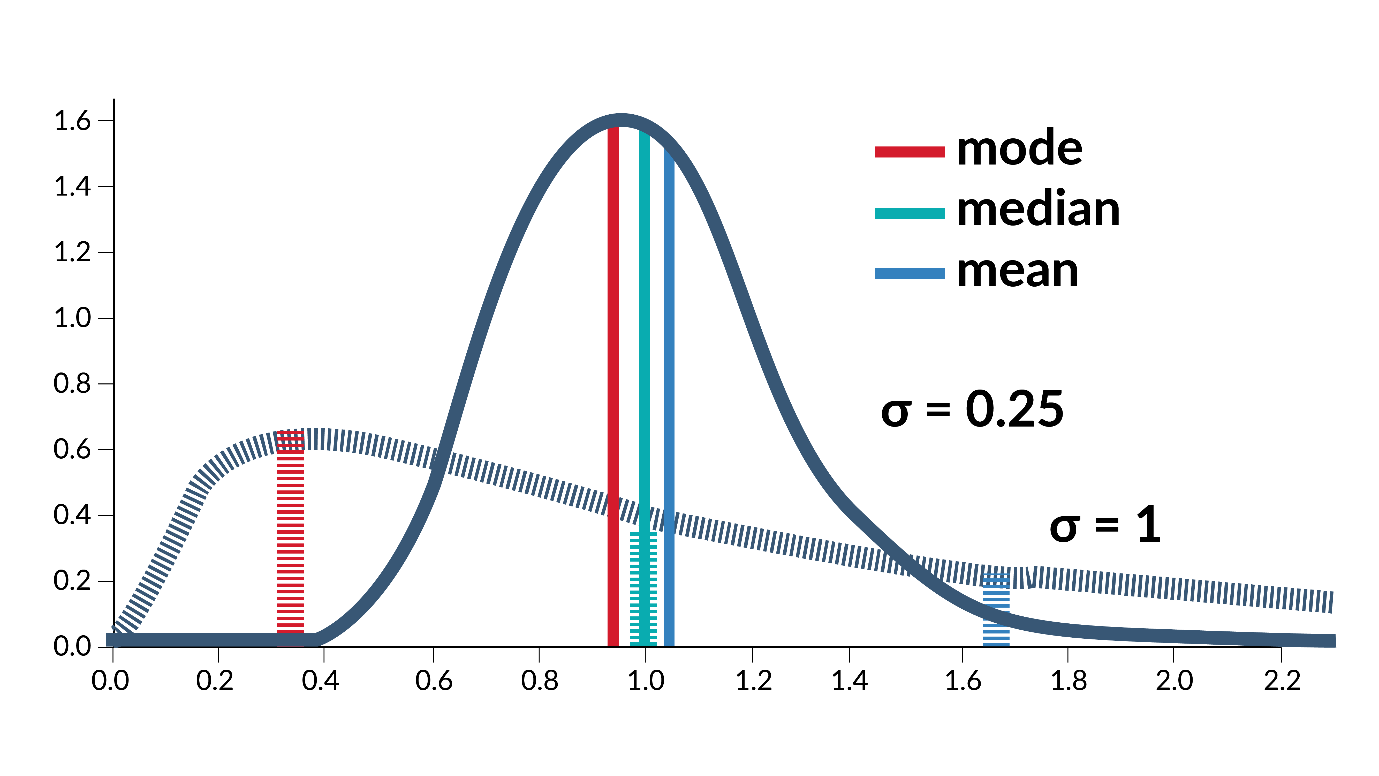
Another measure of central tendency in a data set is the mode. This is the most frequently occurring value.

In this data set, the mode is 74.

(45, 49, 50, 53, 60, 62, 63, 65, 66, 67, 69, 71, 73, 74, 74, 78, 81, 85, 87, 100)

The mode is most useful when we have a discrete variable, and it may also be applied to nominal, categorical data. For example, if we’re looking at hair color: It is reasonable to have a mode of brown hair.

**Mean, Median, and Mode**



We’ve talked about three different measures that describe the central tendency of a data set.

Let’s look at an example that shows us the differences between these measures.

We have two distributions plotted on the right. In the case of the distribution shown with a solid line, we have something that is close to a normal distribution. The mode is at the peak (the peak is always the highest frequency). The median is slightly greater than the mode because there are more data points on the right side of the distribution peak, which drives the median higher. This indicates that the distribution is slightly skewed or asymmetric. Because there are more data points on the right side of the distribution peak, this also tends to drive the mean higher.

Now, let’s look at the other distribution shown as a dashed line. It’s flatter and is clearly left leaning. Again, the mode is at the peak, approximately 0.35. The median is 1, which is much greater than the mode relative to the differences in the other curve. The mean is about 1.65, which is much higher than the median, relative to the differences in the other curve, and that is largely due to the long tail to the right. The greater disparity in these measures of central tendency shows us that the distribution is more skewed.

Lesson 2-3 — Skewness and Kurtosis

The shape and the location of the tail ends of a probability distribution are important for their influence on the sample mean and their importance for calculating the rarity of events. Skewness and kurtosis are the two measures associated with the tail ends of distributions. Skewness is a measurement of asymmetry of the probability distribution around the mean. It helps identify the direction of the influential tail region. When a distribution is perfectly symmetric, the mean, mode, and median are exactly the same and can be used as measures of central tendency. This is no longer the case with a skewed distribution. A probability distribution with a longer tail on the right side or the positive side of the distribution is described as right skewed or positively skewed. In a right skewed distribution, the mean is pulled away from the center and to the right side, while mode and median are less influenced by it. The mean becomes less centrally located than the mode or median in presence of skewness. Therefore, statisticians will often use median instead of mean to describe the center of a highly skewed distribution. An example of right skewed distribution is household income whose distribution is often skewed due to the small number of people with extremely high income. For this reason, median household income rather than mean household income is often used to describe household income. Similarly, a probability distribution with a longer tail to the left side or the negative side of the distribution is called left skewed or negatively skewed. An example of left skewed distribution is the age at death. The mean age at death is often left skewed. Due to a disproportionately high number of deaths at very young age. For this reason, average life expectancy tends to understate how long a person can expect to live. Remaining life expectancy, which excludes deaths at younger age should be used instead for estate or retirement purposes. Compared to the skewness, which helps identify the direction of the influential tail region, kurtosis is a measure of importance of tail regions without regard to the direction. A standardized normal probability distribution with mean zero and standard deviation one, has the kurtosis of three. A kurtosis higher than three is said to have positive kurtosis. Positive kurtosis has fatter or more prominent tails. For example, a Student t-distribution and logistics distribution have positive kurtosis. A kurtosis less than three is said to have a negative kurtosis. Negative kurtosis is associated with tighter distributions with less prominent tails. An example is the Bernoulli distribution with p equals 0.5. It is important to note that kurtosis indicates the relative presence or the relative lack of tails. Kurtosis differs from variance or standard deviation, which describes the overall spread or the shape of the curve. It is possible to have a wide peak with little or no tail which would be an example of high standard deviation and positive kurtosis.

Lesson 2-4 — Measures of Dispersion Part 1

**Range**

The range is straightforward: it is the difference between the maximum and minimum values in the data set. A larger range usually (but not always) indicates a large spread or deviation in the values of the data set.

{73, 66, 69, 67, 49, 60, 81, 71, 78, 62, 53, 87, 74, 65, 74, 50, 85, 45, 63, 100}

For this data set, the range is 55, usually expressed using the minimum and the maximum as: (45, 100). When saying this out loud, one usually says something like "the range is 45 to 100," or simply reports the minimum and maximum, that is, "The minimum is 45, and the maximum is 100."

**The Variance**

The variance of a sample is denoted by s2, and that of a population is denoted by σ2 (sigma squared). It is the average of the squared deviations from the mean. The formulas are given below:

* s2=∑ (X − X ¯ )2 / n – 1
* σ2 = ∑ (X − μ)2 / N

Here, n and N are respectively the sample size and the population size. For the variance of a sample, we may use an alternative formula that bypasses the mean. It is called the direct formula for the variance:

* s2= [∑x2 – (∑x)2/n] / n-1

Let's talk about expected value and variance and how they differ from sample mean and sample variance. Sometimes students believe the expected value is the same as the sample mean and vice versa. Are they the same? Sometimes we ignore the difference between variance and sample variance, they sound almost the same after all. What are expected value and variance? Expected value tells us the central area of the distribution. The formula for expected value is the sum of X times P(X). P(X) is the probability distribution of X. As you may recall, the sum of probability distribution of X is 1. Therefore, expected value is the weighted average of all the possible values that the random variable X can take. Variance tells us about the spread of the possible values of the random variable x. The formula for variance is the sum of P(x) times the square of the difference between X and the expected value of x. Therefore, variance is the weighted average of the square of the difference. Both expected value and variance can be calculated when P(x), the probability distribution of X is known. However, in real life, the probability distribution is rarely known. We do not know for example, the true distribution of stock prices or sales revenues. Even the distribution of body weight which is well measured is not readily available for calculation purposes. Because the true distribution often remains unknown or elusive, we instead use sample mean and sample variance as the estimator of the hypothetical population with unknown distribution. If you look at the formula for the sample mean, and for sample variance, you will see that P of x in the expected value and variance formulas have been replaced by 1 over n and 1 over n-1. In other words, sample mean and the sample variance assume that each outcome was equally likely to occur, which is almost never true. Therefore, we can conclude that the sample mean and sample variance are approximate estimates about the true population parameters. Fortunately, both sample mean and sample variance will approach the true population mean and population variance as the sample size grows larger. Which is implied by the law of large numbers. But that's a topic for another discussion.

**The Standard Deviation**

This quantity is the square root of the variance, and is denoted by σ for a population, and by s for a sample. The standard deviation is the most widely used measure of dispersion. An intuitive way to think of these values is that they measure the deviation from the mean of the data set.

* s = √s2
* σ = √σ2

**Example**

For the following sample data, find the mean, the mean deviation, the variance, and the standard deviation. Use both the standard and the direct formulas to compute the standard deviation: -2, 0, 1, 2, 5, 6

**Solution:** X¯ = ∑X/n = − 2 + 1 + 2 + 5 + 6/6 = 2

For the variance, using formula (2), we first construct a table as follows:

|  |  |  |
| --- | --- | --- |
| X | X − X¯ | (X − X¯)2 |
| -2 | -2 − 2 = 4 | (-4)2 = 16 |
| 0 | 0 − 2 = -2 | (-2)2 = 4 |
| 1 | 1 − 2 = -1 | (-1)2 = 1 |
| 2 | 2 − 2 = 0 | (0)2 = 0 |
| 5 | 5 − 2 = 3 | 32 = 9 |
| 6 | 6 − 2 = 4 | 42 = 16 |

Next, ∑ (X − X¯)2 can be obtained by summing the entries of the last columns:

∑ (X − X¯)2 = 16 + 4 + 1 + 0 + 9 + 16 = 46

Therefore,

s2 = ∑ (X − X¯)2 / n − 1 = 46/6 − 1 = 9.2

The standard deviation can now be calculated by taking the square root of the variance:

s = √s2 = √9.3 = 3.033

Next, we will recalculate the variance by using the direct formula (formula 3). A table such as below can be constructed to facilitate the calculations.

|  |  |
| --- | --- |
| X | X2 |
| -2 | 4 |
| 0 | 0 |
| 1 | 1 |
| 2 | 4 |
| 5 | 25 |
| 6 | 36 |
| ∑ X = 12 | ∑ X2 = 70 |

The magnitude of these two statistics is related to the values in the data. Samples and populations use different notation and equations to define mean, standard deviation and variance. A sample is a subset of a population, and this distinction will be formally defined in future modules.

Let's talk about sample mean. It is perhaps the most important tool for statistical analysis, depending on the purpose at hand, statisticians use it differently. Students sometimes become confused about the different ways that the sample mean can be used differently. Being aware of how it can be used may help prevent confusion later. In modern statistics, sample mean can be used as one: key measure of descriptive statistics. Two: estimator for inferential statistics and three: predictor for predictive analytics. When conducting descriptive statistics, sample mean is a key measure that describes an existing sample. Sample mean is a measure of central tendency that tells us where the central region of the dispersion is located. When performing inferential statistics, sample mean is an estimator for the population mean. Estimator is a statistical tool used to take an educated guess about the true value of the population from which the sample was drawn. Finally, when conducting predictive analytics, sample mean is the best first predictor for the unknown value or outcome. In the absence of additional information, using the sample mean as the predictor helps minimize the error, which is the difference between the unknown outcome and the predictor. As more information becomes available, you adjust or condition your prediction upward or downward. Linear Regression which is a method for adjusting or conditioning the sample mean of outcome is also called conditional mean for this reason. In conclusion, sample mean is a statistical tool that can be used as a measure, estimator, and predictor. In three areas of modern statistics.

* N: number of elements in a population
* μ: population mean
* σ: population standard deviation
* σ2: population variance
* n: number of elements in a sample
* X¯: sample mean
* s: sample standard deviation
* s2: sample variance

|  |  |  |  |
| --- | --- | --- | --- |
|  | Mean | Standard Deviation | Variance |
| Population | μ = 1/N ∑i = Ni = 1 Xi | σ = √ (1/N ∑i = Ni = 1 (Xi − μ)2) | σ2 = (1/N ∑i = Ni = 1 (Xi − μ)2) |
| Sample | X¯ = 1/n ∑i = n i = 1 Xi | s = √ (1/n-1 ∑i = n i = 1 (Xi − X¯)2) | s = 1/n-1 ∑i = n i = 1 (Xi − X¯)2 |

In general, the variance is the mean of the difference between each data point and the mean of the data set, squared, and then averaged. So, to calculate standard deviation, we first calculate for each data point how far it is from the mean. We square those values, sum them, and then divide that sum by the number of data points (or the number of data points minus one in the case of a sample). This is the variance. To calculate the standard deviation, we take the square root.

**Consider the example:**

For the following sample data set:

(45, 49, 50, 53, 60, 62, 63, 65, 66, 67, 69, 71, 73, 74, 74, 78, 81, 85, 87, 100)

the standard deviation is calculated as:

X¯ = 68.6

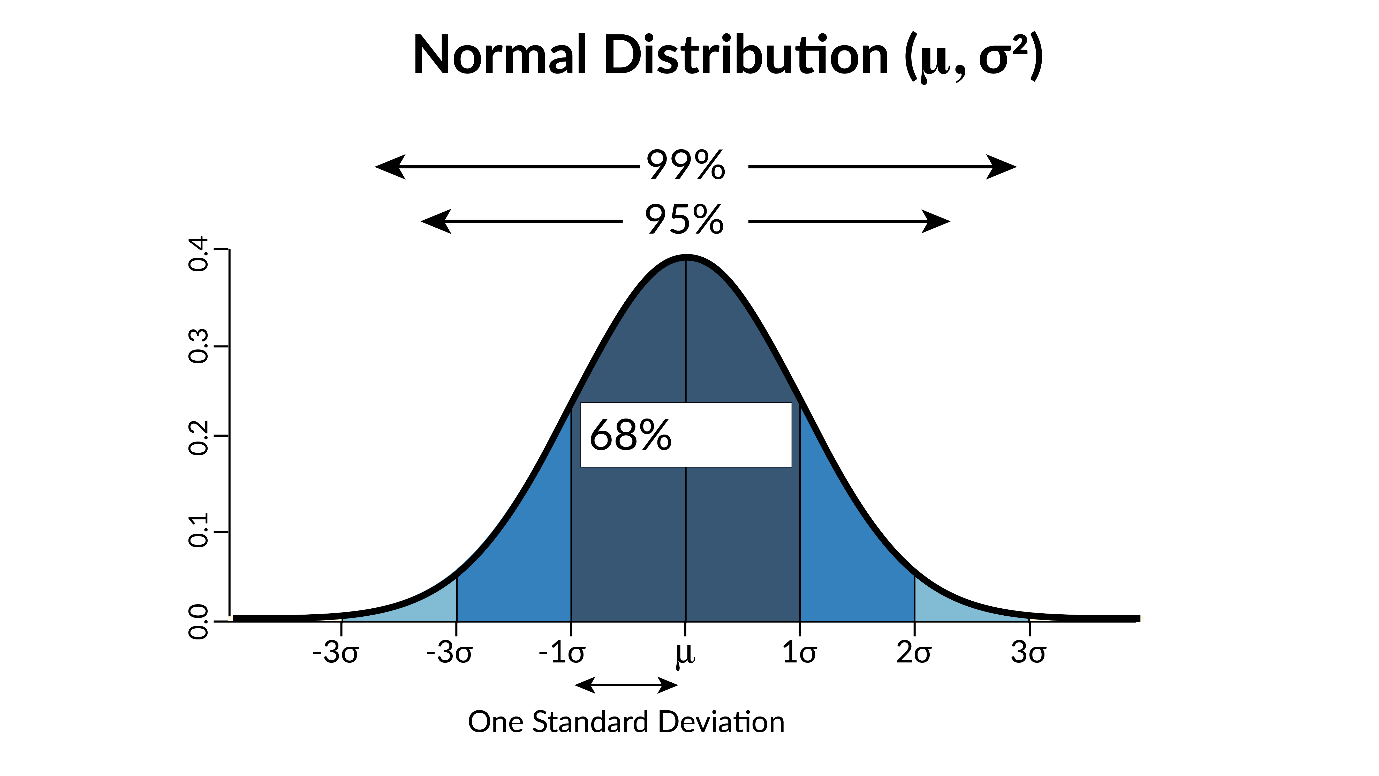
s = √ (1/20 − 1 ((73 − 68.6 )2 + (66 − 68.6)2 ± (69 − 68.6)2 ± …))

s = 13.8

Lesson 2-5 — Measures of Dispersion Part 2

**Standard Deviation and Mean - Normal Distribution**

The Normal Distribution which we will use to discuss standard deviation and the coefficient of variance is shown below. This is a graph of the normal distribution showing 68% of data falling within one standard deviation, 95% of the data falling within two standard deviations and 99% of the data falling within three standard deviations.



As we just noted, the standard deviation is helpful because it indicates the dispersion of the data. For example, in a normal distribution, 68% of the data falls within one standard deviation of the mean. 95% of the data falls within two standard deviations of the mean. 99% of the data falls within 3 standard deviations of the mean.

**Coefficient of Variation**

The coefficient of variation is simply the standard deviation divided by the mean. This is useful because the standard deviation is subject to the general magnitude of the data points and thus cannot be used directly to compare the levels of dispersion across different data sets. With this normalization of the standard deviation using the mean of the data, the coefficient of variation statistic can be used directly as a measure of dispersion across data sets. The coefficient of variation is calculated as:

* CV: coefficient of variation
* CV = α/μ
* %CV: percent coefficient of variation
* %CV = α/μ ∗ 100%

As an example, a data set of [100, 100, 100] has constant values. In this example, the standard deviation is 0, and the average is 100. That makes the coefficient of variation 0, as: 100% × 0 / 100 = 0%

A data set of [90, 100, 110] has more variability. Its standard deviation is 8.16 and its average is 100. This gives a coefficient of variation of 8.16%, as: 100% × 8.16 / 100 = 8.16%

Finally, a data set of [1, 5, 6, 8, 10, 40, 65, 88] has even more variability. Its standard deviation is 30.78 and its average is 27.875, resulting in a coefficient of variation of 110.4%, as: 100% × 30.78 / 27.875 = 110.4%

The coefficient of variation is greater than 100%, which means the standard deviation is greater than the mean. It is important to know if your data has such normalized spread.

Example sourced from "Coefficient of Variation" page on Wikipedia.com.

Lesson 2-6 — Percentiles, Quartiles and Deciles

**Percentiles**

Some other measures that are helpful when describing the distribution of data are percentiles, quartiles, and deciles. These statistics can be used as indicators of the center, spread, and even the shape of the distribution.

To calculate these, first, we order the data then divide it into a number of slots. In the case of percentiles, 100 slots.

Once we’ve done this, we can say that a score is in the "25th percentile," or that the 25th percentile is equal to a certain value. Percentiles give us a way to relate a given score to the rest of the data.

The calculation of a percentile is shown below. In a given dataset, to calculate the value of the desired percentile (or percentile of interest), XPC, we must follow the steps below:

1. Sort the data.
2. Using PC, the desired percentile, calculate the location of the value of interest, nPC
3. If nPC, is an integer, the value at that location in the sorted data set is the value of the desired percentile.
4. If nPC is not an integer, use the two nearest values to nPC in the sorted dataset to calculate nPC.

PC: desired percentile, specified as a fraction

nPC desired observation of sorted dataset

nPC = ( n + 1/100 ) \* PC

XPC: value of desired percentile in the dataset

IF nPC is not an integer AND 1 ≤ nPC ≤ n,

THEN nPC refers to specific element of dataset:

XPC = (nPC)th element of sorted dataset

IF nPC is not an integer AND 1 < nPC < n,

THEN nPC:

nPI: nPC rounded down to the nearest integer

nPR = nPC − nPI: fractional part of nPC

OI: (nPI)th element of sorted dataset

OII: (nPI+1)th element of sorted dataset

XPC = OI + (OII − OI)nPR

**Example:**

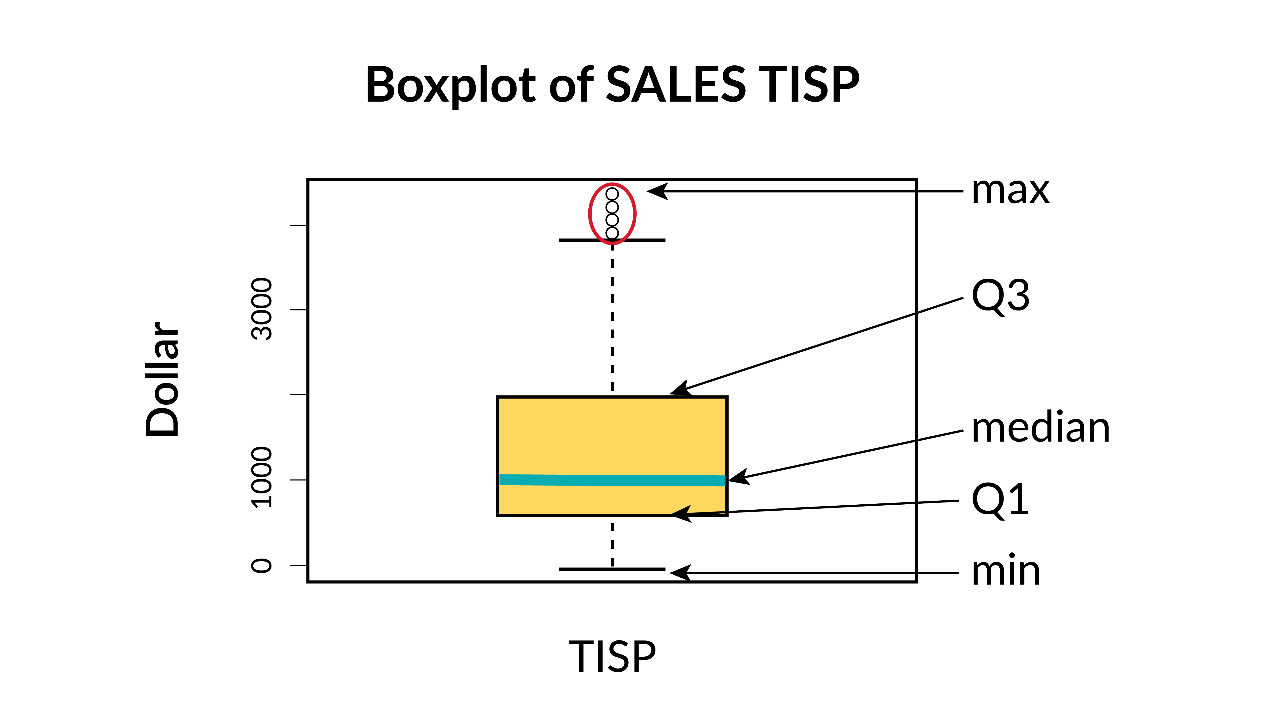
Consider the following ordered set of data: 3, 5, 7, 8, 9, 11, 13, 15. If we are interested in the 25th percentile:

nPC = (n + 1/100)\*PC = (8 + 1/100) ∗ 0.25 = 2.25

* nPC is not an integer
* nPI = 2
* nPR = 0.25
* OI = 5
* OII = 7
* XPC = OI + (OII − OI)nPR = 5 + (7−5)\*0.25 = 5.5

**Quartiles**

Quartiles refer to three values in an ordered data set: the 25th, 50th, and 75th percentiles. These values divide the data set into four equal parts, or quarters, and are used when presenting data in a box plot. Note the median is equal to the 50th percentile and is known as the second quartile, denoted Q2. The 25th percentile is known as the first quartile or the lower quartile and is denoted as Q1. The 75th percentile is known as the third quartile or the upper quartile and is denoted as Q3. These three quartiles, along with the minimum and maximum, are known as the five-number summary of descriptive statistics. Consider the example box plot, shown here:



**Deciles**

Decile analysis helps us to understand the relative importance of sections of the variable distribution. We order the data then divide it into 10 slots. These 10 groups of data are known as deciles. Since the data is broken into equal cut points, we can summarize how much contribution each decile makes to the overall values. In the case of revenue, if each data point how much a particular customer spends, we can calculate how much the customers who are in the top decile spend and compare that to how much customers in the bottom decile spend. That gives us the relative importance of the customers to the total revenue.

Lesson 2-7 — Data Presentation - Categorical Data

One of the best ways to begin to understand a new set of data is to visualize it. There are two ways to present data, graphical and tabular:

* Graphical data presentation involves some type of chart
* Tabular data presentation involves organizing data into a table

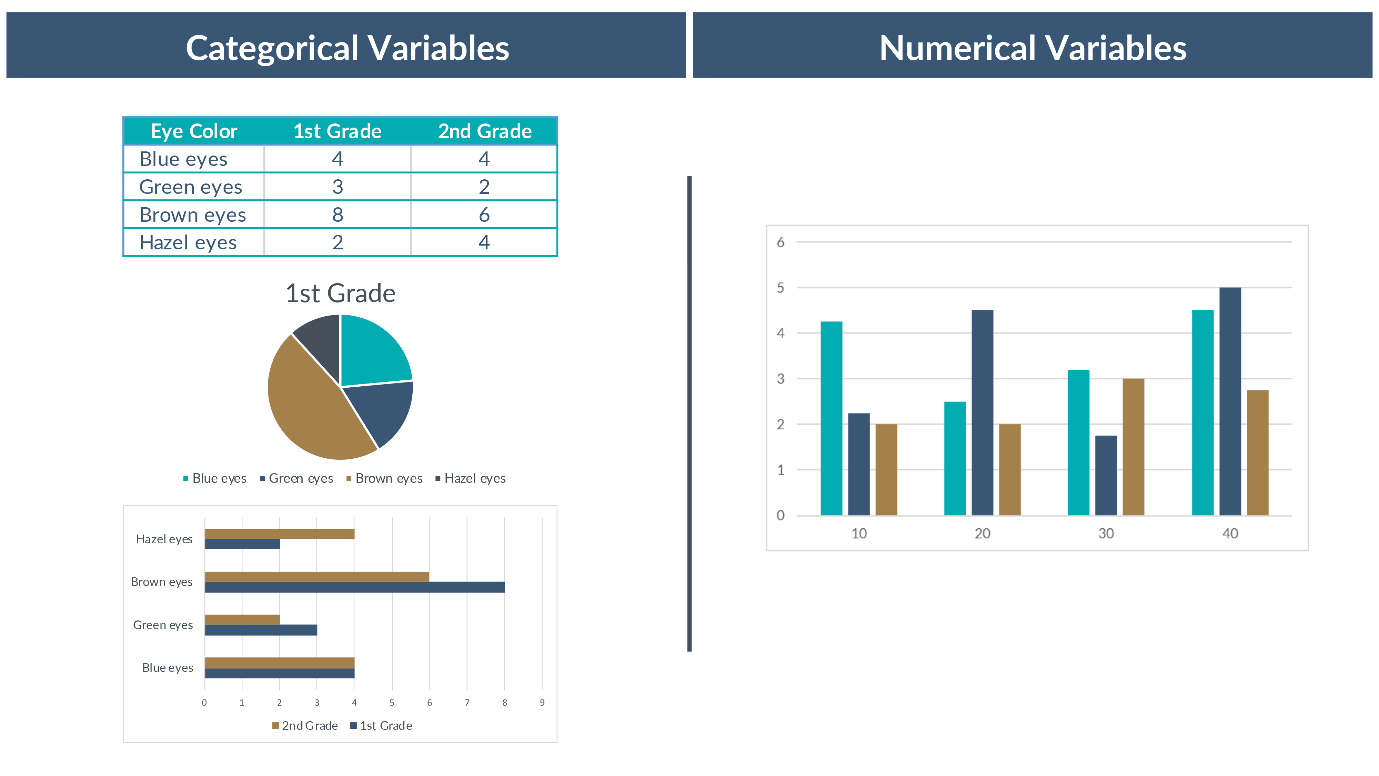
Data presentation is a way to represent the distribution of the variable so that we understand its center, shape, and spread. The type of chart or table should be selected based on the **data type**.

Recall, data can be categorized as numeric or categorical. Below is a summary of common ways to present categorical and numerical data:

* Categorical data is usually presented in a bar chart, pie chart, or table
* Numerical data is usually presented in a histogram or box plot, which are types of charts

**Example:** Data Presentation for Categorical and Numerical Variables

This image shows how different types of data are displayed. Categorical data can be displayed with tables, pie charts and bar charts. Numerical data can be shown in histograms.

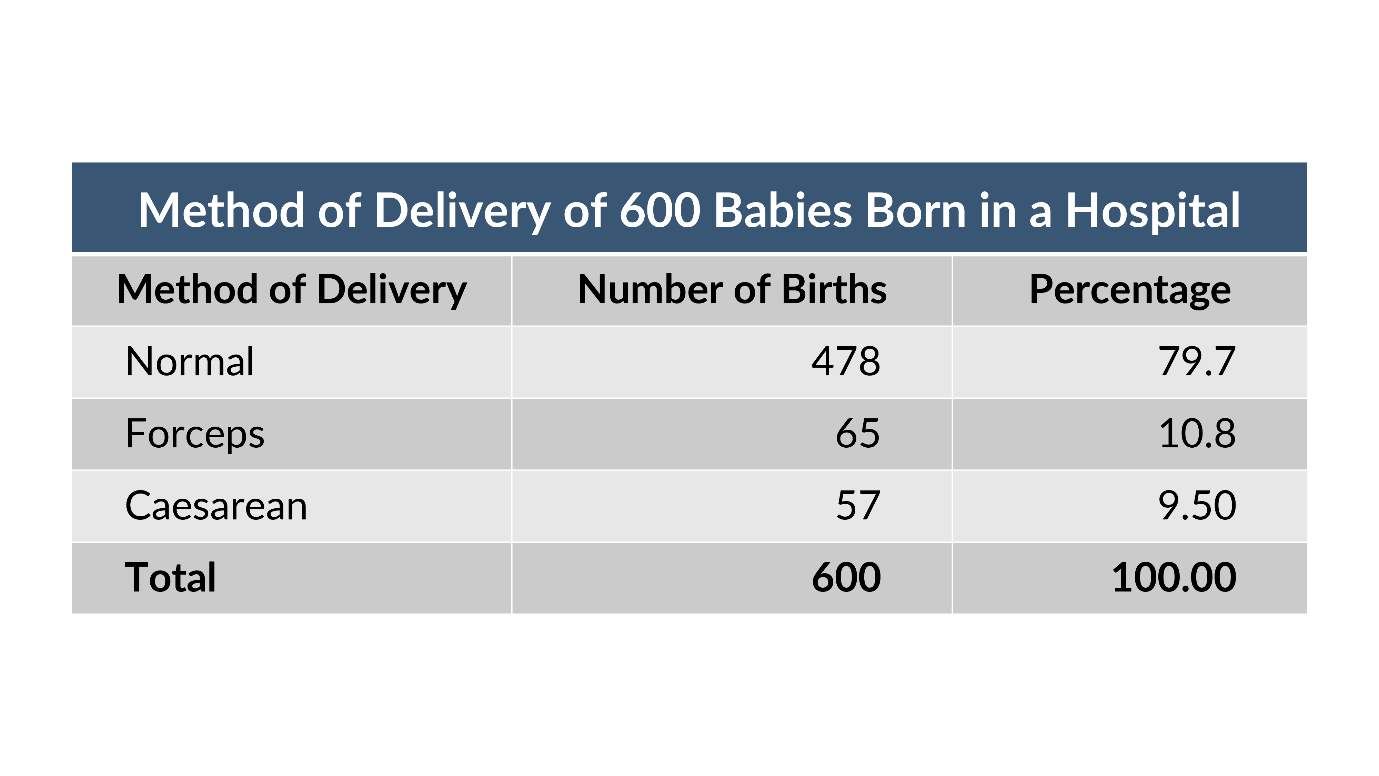


We will present tables and each of these types of charts in more detail, starting with ways to display categorical variable data.

**Categorical Variables - Table**

With categorical variables, we report the frequencies or percentages of each level of variable or category. One way to do this is in a table.

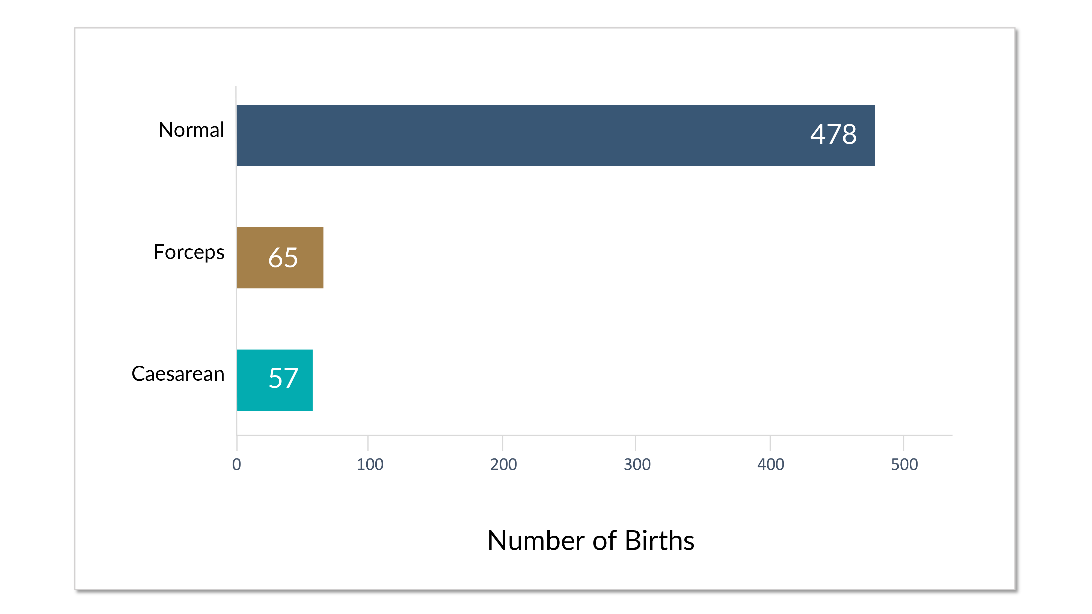
This is a table that shows the method of delivery of 600 babies born in a hospital. The columns are Method of delivery, number of births and percentage of births. The rows include the data for normal, forceps and Caesarean deliveries.



In this table, we see a few different ways to display frequencies and percentages for the same data. We have different methods of delivery of babies. The first column provides the method of delivery and the second column provides a count of the number of each method. The third column provides a decimal value that represents the percentage of the portion of deliveries in each category relative to all of the deliveries.

**Categorical Variables - Bar Chart**

This is a bar chart for the number of births for the method of delivery from the table above. 478 deliveries are normal, 65 are forceps deliveries and 57 are Caesarean deliveries.

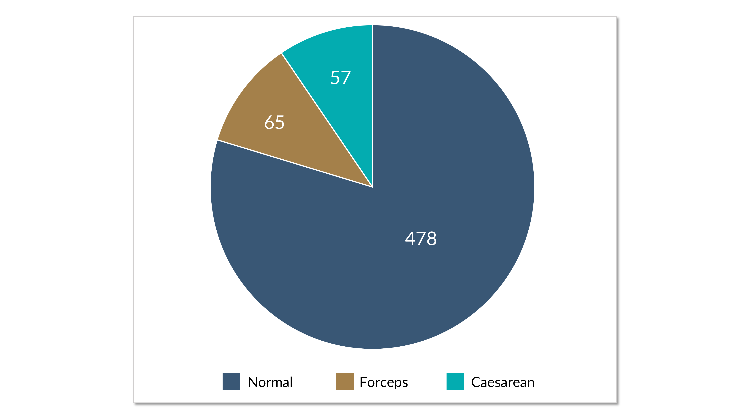


Alternatively, we can use a bar chart to represent the frequency of categorical variables graphically. This is sometimes referred to as a frequency distribution or a frequency chart and it’s very useful if we have more than 100 observations.

In this chart, we see the number of births is broken out graphically, rather than using the numbers from the table. This makes it very easy to understand the relationship between the different categories. In this presentation, we see that we have far more normal deliveries than forceps or cesarean section.

**Categorical Variables - Pie Chart**

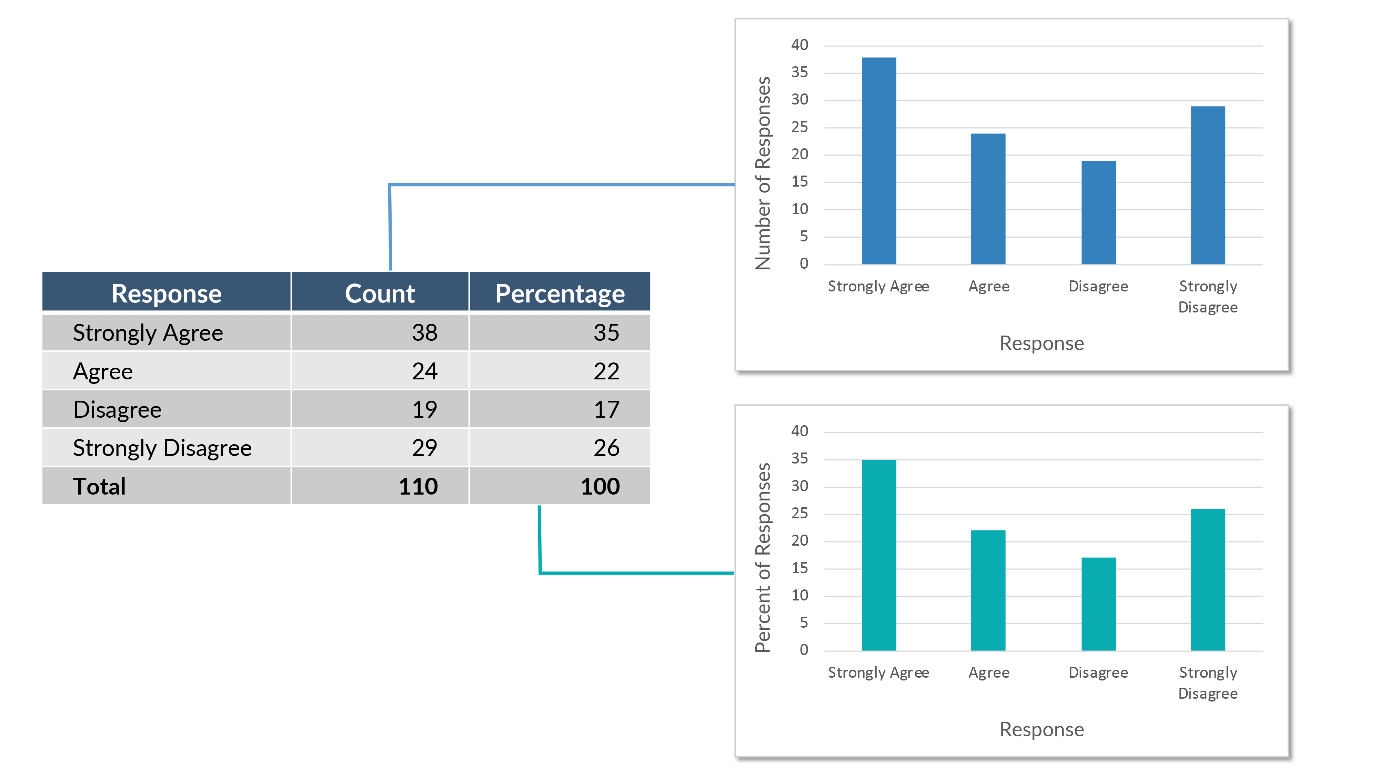
This is a pie chart of the data from the chart above. Finally, we can represent the categorical variables by using a pie chart. If we want to show the relative proportions of different pieces of the data, a pie chart is a good option.



This pie chart depicts the percentages we saw in the original table. Pie charts are good when we want to display one aspect of a variable, specifically if we want to show the different proportions of a variable. The proportion of normal deliveries relative to Forceps and Caesarean section birth is very clear in this chart. However, it’s difficult to encode any other information in this chart.

**Categorical Variables - Ordinal Data**

The previous examples showed the options for presenting nominal data. When we have ordinal data, we can use either frequency and percentage charts or a data table. A pie chart is less helpful for ordinal data because it doesn't have an inherent order, such as left to right or top to bottom.



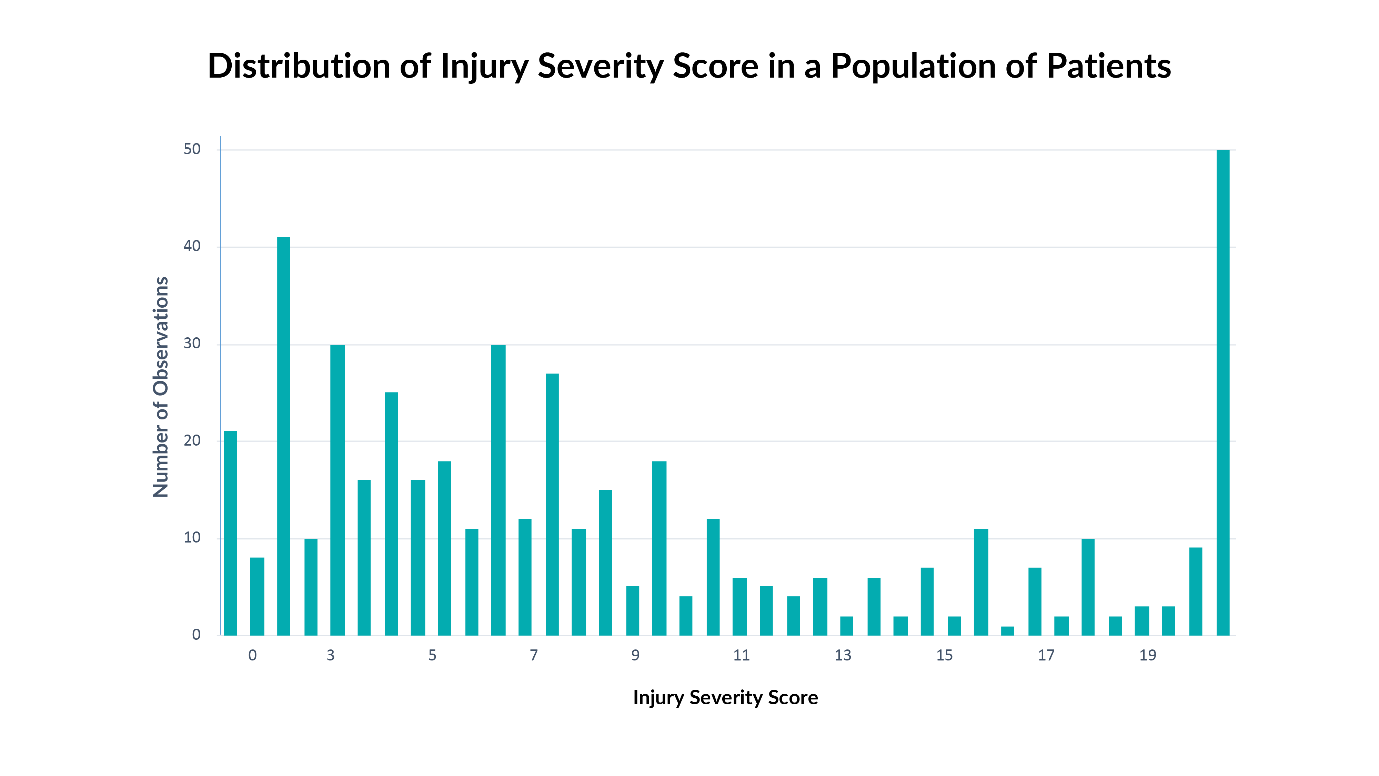
Data Chart with Number of Response bar chart and Percent of Responses bar chart. This table shows the survey result for a respondent answering the question of "Would you recommend this product?" The allowed answers are: strongly agree, agree, disagree, or strongly disagree. In addition to the raw counts, the percentage of responses for each level are also shown.

This data is also shown in a bar chart, both as frequencies (top-right) and as percentages (bottom-right). This bottom chart is also known as a percentage chart.

Lesson 2-8 — Data Presentation-Numerical Data

**Numerical Variables - Frequency and Percentage Charts**

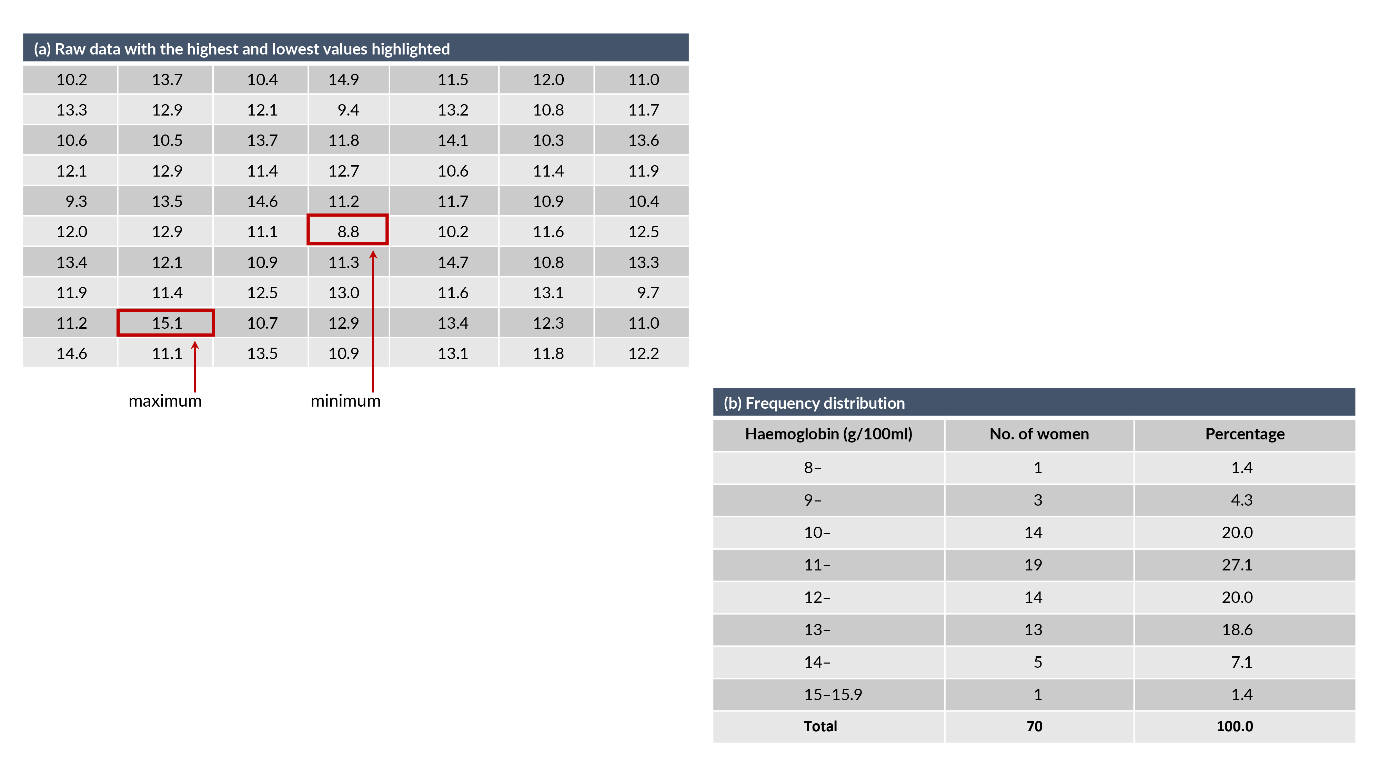
This is an image of a frequency chart showing the distribution of injury severity score in a population of patients in a hospital. The severity score ranges from 0 to 20 along the horizontal axis and the number of observations along the vertical axis.



We can also use frequency or percentage charts to display numerical data. This chart shows the frequency of Injury Severity Scores for a given population of patients.

**Numerical Variables - Tables**

This image shows two tables of numeric data. The table on the left is the raw data with the min and max values highlighted. The table on the right is a frequency table showing the raw count and the percentage.



It is possible to take continuous data and turn it into categorical data by grouping values together. Then we can calculate frequencies and percentages for each group.

It is convenient to use frequency or percentage charts to illustrate how different values of a variable are distributed. When we deal with numerical data, we can convert it to categorical data first and use a similar approach for ease of understanding.

Do this by combining the values together or binning the values. Then we can calculate frequencies and percentages for each group – this is the same as what we saw previously for categorical variables.

Lesson 2-9 — Data Presentation-Distributions

**Numerical Variables - Central Tendency and Spread**

In addition to the tabular presentation of binned numerical data (interval or ratio), frequency distributions display measures of central tendency and dispersion. A frequency distribution might show outliers or data that are far away from the middle of the data.

Distributions graphically present the dispersion or how spread out the data is. The larger these measures, the more spread out the data. Means are often presented along with standard deviations or standard errors.

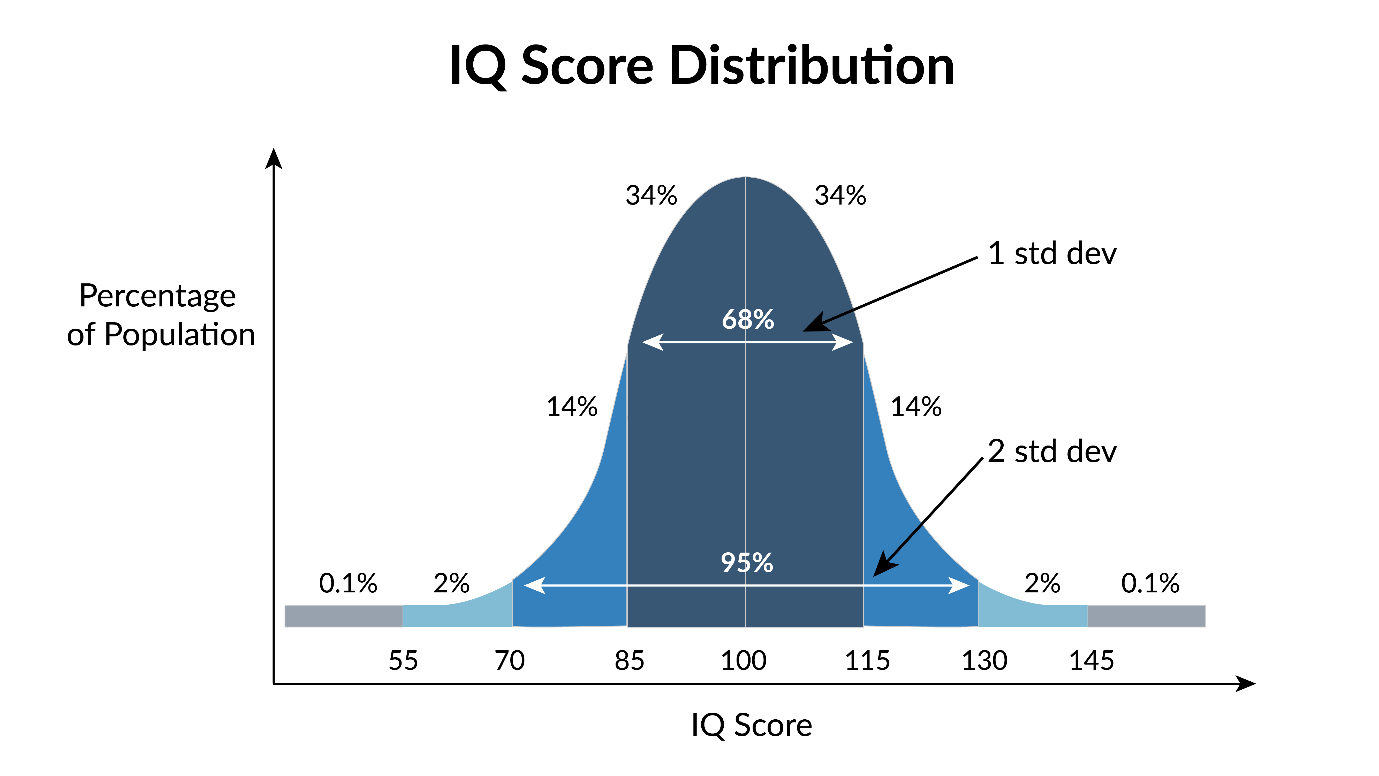
**Numerical Variables - Distributions**

The graphic below displays the distribution of numerical data. The x-axis displays the IQ score, while the y-axis shows the percentage of the population that has that score. The distribution shows a bell curve – highest in the middle, symmetric, and tapering on both ends.

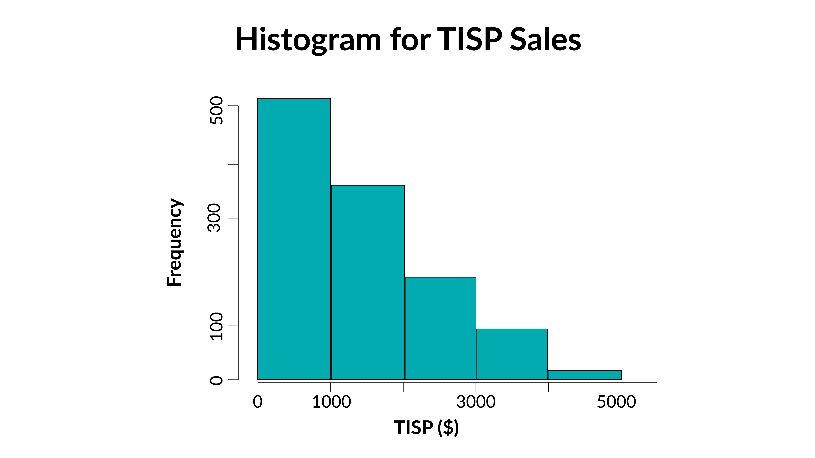
Formally, this shape represents a type of distribution known as the normal distribution, which is frequently seen in statistics. Many phenomena in life are normally distributed, such as age, height, weight, and IQ.

In this example, one measure of central tendency is the mean, which is shown as the peak in the curve. This chart shows one standard deviation and two standard deviations as measures of the dispersion of the data. Standard deviations are important for understanding how much of the data is in a given range. 68% of the data is included within one standard deviation of the mean. 95% of the data is within two standard deviations of the mean.

Normal Distribution of IQ scores showing 1, 2 and 3 standard deviations.



**Numerical Variables - Histograms**

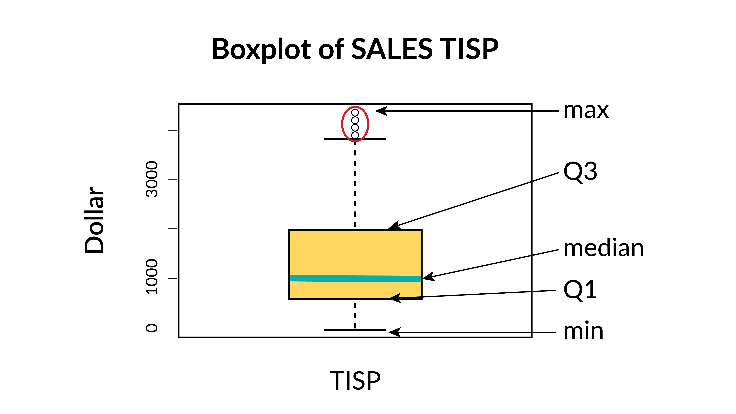


This is a histogram of TISP Sales. One of the most useful charts for numerical data is a histogram. A histogram is a bar chart presentation of the frequency table.

The data is binned, as we described for the frequency table with numerical data, and the counts for each bin are displayed graphically. In a histogram, the y-axis is always frequency or the number of data points that fall in the given bin.

**Numerical Variables - Box Plots**

This is a boxplot graph highlighting the min and max data values, and the first quartile, median and third quartile values. Another useful plot for numerical data is the box plot. It displays the five-number summary of the variable: the minimum, maximum, and median are depicted via horizontal lines on the plot. The box indicates the first, second, and third quartiles. The data points shown as circles is outliers. These terms will be described in greater detail shortly.



The upper and lower whiskers, shown as horizontal lines near the upper and lower extremes of the box plot are defined as:

upper whisker:

* min((max(X),Q3+1.5∙(Q3−Q1))
* lower whisker
* max((min(X),Q1−1.5∙(Q3−Q1))

In some cases, the outliers are ignored, and then the upper whisker and lower whisker values are considered to be the minimum and the maximum.

A box plot visually depicts the skewness of the data. If the box is in the middle of the two whiskers and the median line is in the middle of the box, this means the data is fairly symmetric. If either the box or the median line within the box leans to one side or the other, the data is skewed.

When looking at a box plot for a single variable of a data set, some important things to notice are the difference between the min and the max, the size of the box relative to the min and max, how many data points are outliers, the magnitude of the variable, and the relative location of the box and the median line within the whiskers. When we have multiple box plots, one for each variable in a dataset, we will note the differences among the various variables – how different is each box plot from the others.

**Week 3: Readings of Chapter 3**

Introduction

We use probability in many situations. Probability informs the following everyday decisions: what kind of weather to prepare for, whether or not to buy a stock and how much to bet when gambling. A better understanding of probability results in a more accurate assessment of risk. Data scientists need to be familiar with probability to answer business questions, which may influence strategic decisions managers make.

This module provides an explanation of probability for processes with a finite number of possible outcomes. It explains the meaning of probability, as well as how to calculate probabilities. It also examines the relationship between disjoint and independent events.

First, we will deal with the probability of a single event. We will look at the equation for probability, which is used to calculate the probabilities of various events.

We will also discuss the concepts of Bayes’ Rule and Simpson’s Paradox and how these concepts fit into our understanding of probability.

Finally, we will introduce some combinatorial methods, or to put it simply, ways of counting things.

Lesson 3-1 — Basic Concepts

We begin the study of probability with a straightforward example. Suppose a coin is tossed and the up face is recorded. The result is called an observation, and the process of making an observation is called an experiment. The two possible outcomes of this experiment are:

Observe a tail (T), Observe a head (H).

Each one of the above possible outcomes is called an outcome, or a simple event, or a sample point. A sample point is the most basic outcome of the experiment. The sample space of an experiment is the collection of all its sample points. In our example, the sample space, denoted by S, is: S = {T, H}.

What if we tossed a coin twice? What is the sample space of this experiment?

Even for a seemingly trivial experiment, we must be careful when listing the sample points. There are four possible outcomes, and the sample space is the collection of all above sample points:

Sample Space S = {TT, TH, HT, HH }.

An Event is the collection of one or more basic outcomes of an experiment. As we saw earlier with tossing a coin twice, the sample space included the outcomes: TT, TH, HT, HH.

Now we may define an event A as: “Observing at least one tail.” In this case, A = {TH, HT, TT}.

We are now ready to discuss probabilities of sample points. Probability is generally used synonymously with “chance,” “odds,” and similar concepts. The probability of an event A is denoted by P(A).

For a single coin toss we might state that “the chance of observing a head is 50%” or “the odds of seeing a head is 50:50.” Both these statements are based on informal knowledge of probability.

More formally, the probability of a sample point is a number between 0 and 1 (inclusive) that measures the likelihood that the outcome will occur when the experiment is performed. This number is usually taken to be the relative frequency of the occurrence of a sample point when the experiment is repeated a very large number of times.

The probabilities assigned to events in sample space must always obey two rules, also known as Probability Rules:

* All sample point probabilities must lie between 0 and 1.
* The probabilities of all the sample points within a sample space must sum to 1.

The probability of a single event P(A) is:

Probability of an outcome = The number of time the outcome is observed/The total number observable outcomes

For our example, assuming a fair coin, each sample point is equally likely to happen, i.e., the probability of each sample point is . And the probability of all the sample points to sum to

1:1/4 + 1/4 + 1/4 + 1/4 = 1

In the coin flip example where we’re flipping 2 times, the event we are interested in might be that we get 2 heads in a row. In the example of tossing a coin, the event is not the toss itself, but heads or tails.

Now we may define an event A as: “Observing at least one tail.” In this case, A = {TH, HT, TT}.

Based on the definition above, the probability of the event

P(A) : The number of times the outcome is observed/The total number of observable outcomes = ¾

Lesson 3-2 — Venn Diagrams

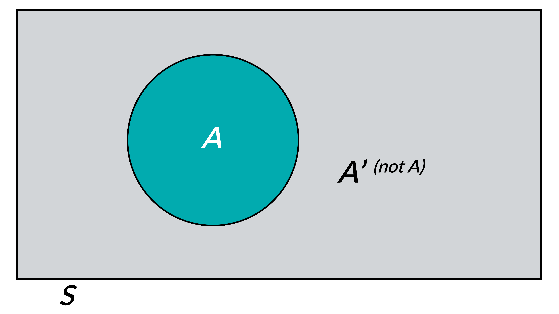
The sample space S is shown as a closed figure, labeled S, and containing all possible sample points. Such graphical representation is called a Venn diagram.

An event A belonging to a sample space S is shown as a round closed figure inside S.

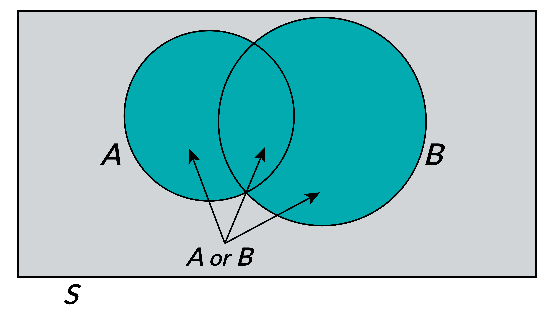
The complement of an event A, denoted by Ac (not A) or A', is the event within the sample space that occurs if A does not occur.

The figure above is the Venn diagram indicating an event A and its complement A´.

This Venn diagram shows the sample space S, with a subset A and its compliment, not A. The union of events A and B occurs when A or B or both occur.

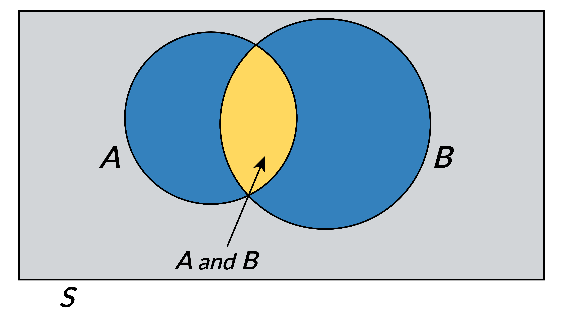


The union of the two events is usually denoted by A ∪ B or (A or B), and consists of sample points that belong to A or B or both.

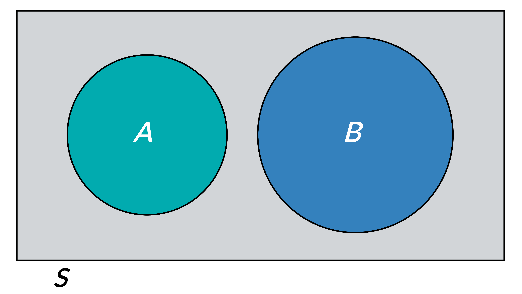


This Venn diagram shows the union of sets A and B, otherwise described as A or B. Note that A and B do intersect in this diagram.

The intersection of two events A and B, denoted by A ∩ B or (A and B), occurs if both A and B occur simultaneously. (A and B) consists of all sample points belonging to both A and B.



This Venn diagram shows the intersection of events A and B. A and B have sample points in common. Two events A and B are said to be mutually exclusive or disjoint if their intersection, contains no sample points, that is, if A and B have no sample points in common.



This Venn diagram shows two events, A and B, that have no sample points in common. They are mutually exclusive.

We can use Venn Diagrams to illustrate probability problems.

**Example:**

Consider a die-toss experiment. Define the following events:

* A: Toss an odd number
* B: Toss a number less than or equal to 3

The following Venn diagram illustrates the events A and B.

Describe A ∪ B, A ∩ B, A' and B' and find the probability of each one.

**Solution:**

S = { 1 , 2 , 3 , 4 , 5 , 6 } , A = { 1 , 3 , 5 } and B = { 1 , 2 , 3 }

The union of A and B, (is the event that occurs if we observe either an odd number or a number less than or equal to 3 or both on a single throw of the die. We find:

A ∪ B = { 1 , 2 , 3 , 5 } ⟶ P ( A ∪ B ) = P ( 1 ) + P ( 2 ) + P ( 3 ) + P ( 5 ) = 1/6 + 1/6 + 1/6 + 1/6 = 2/3

The intersection of A and B, is the event that occurs if we observe both an odd number and a number less than or equal to 3 on a single throw of the die:

A ∩ B = { 1 , 3 } ⟶ P ( A ∩ B ) = P ( 1 ) + P ( 3 ) = 1/6 + 1/6 = 1/3

The complements of A and B are given below:

A′ = { 2 , 4 , 6 } ⟶ P ( A ′ ) = P ( 2 ) + P ( 4 ) + P ( 6 ) = 1/6 + 1/6 + / 6 = 1/2

B′ = { 4 , 5 , 6 } ⟶ P ( B ′ ) = P ( 5 ) + P ( 6 ) = 1/6 + 1/6 + 1/6 = ½

Notice that events 4 and 6 are both outside of A and outside of B, thus part of both A’ and B’

Lesson 3-3 — Rules of Probability

So far, we have been discussing probability using Venn diagrams, but now is the time to formalize some of the concepts we have seen.

**The Complementary Rule of Probability**

The sum of the probabilities of complementary events equals 1: that is, P ( A ) + P ( A′ ) = 1

**Example**

Consider the experiment of tossing a fair coin twice. Use the complementary rule to calculate the probability of event A:{observing at least one tail}.

**Solution:**

The complement of A is defined as the event that occurs when A does not occur. Therefore,

A' = {observing no tails} = {HH} =1/4

P(A) = 1 − P(A′) = 1 – 1/4 = 3/4

Note: we saw this P(A) earlier and this matches the previous probability calculation.

**The Addition Rule of Probability**

The probability of the union of events A and B is the sum of the probability of events A and B minus the probability of the intersection of events A and B, that is,

P(A ∪ B) = P(A) + P(B) − P(A ∩ B)

If two events are mutually exclusive, the probability of their union equals the sum of their respective probabilities:

P(A ∪ B) = P(A) + P(B)

Note: Because A and B cannot happen simultaneously, P(A ∩ B) is 0.

**Example:**

Consider the experiment of tossing a coin twice. Suppose the coin is not balanced and the probabilities of sample points are given in the table below.

Outcome, Probability: HH, 0.16, HT, 0.24, TH, 0.24, TT, 0.36

Consider the events

* A: {Observe exactly one tail}
* B: {Observe at least one tail}

Calculate the probability of A, and the probability of B.

**Solution:** Event A contains the sample points HT and TH.

We can calculate the probability of event A by summing the probabilities of its two sample points:

* P(A) = P (HT) + P(T H) = 0.24 + 0.24 = 0.48

Similarly, since B contains the sample points HT, TH, and TT, and

* P(B) = P(HT) + P(TH) + P(TT) = 0.24 + 0.24 + 0.36 = 0.84

The above example leads us to a general procedure for finding the probability of an event:

* The probability of an event A is calculated by summing the probabilities of sample points (outcomes) in A.

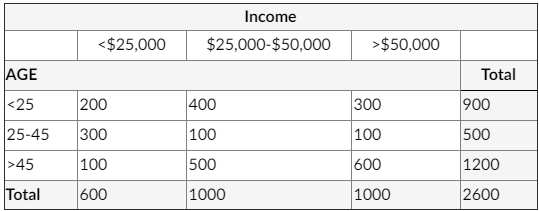
The following is a summary of steps for calculating the probability of an event:

* Define the experiment.
* List the outcomes in the sample space.
* Assign a probability to each outcome.
* Define the event(s) of interest.
* Sum the probabilities of the outcomes associated with event(s) of interest.

Let’s look at a slightly different type of problem.

**Example**:

The following table describes the income of the adult population of a small suburb of a southern city:



Consider the following events:

* A : The person is between 25 and 45
* B : The person has income less than $25,000

Calculate the probability of A. Then calculate the probability of B.

**Solution:**

The suburb has a total adult population of 2600. There are 500 adults who are between 25 and 45. Therefore P(A) = 500/2600 = 0.192.

Similarly, there are 600 adults whose income is less than $25000. Therefore, P(B) = 600/2600 = 0.231.

Consider another example

**Example**

Hospital records show that 15% of all female patients are admitted for surgical treatment, 25% are admitted for obstetrics, and 5% receive both obstetrics and surgical treatments. If a new female patient is admitted to the hospital, what is the probability that the patient will be admitted either for surgery, obstetrics, or both?

**Solution:**

* A: {A female patient admitted to the hospital receives surgical treatment}
* B: {A female patient admitted to the hospital receives obstetrics treatment}

Then, from the given information, (A) = 0.15 , P(B) = 0.25 ,and P(A ∩ B) = 0.05

Therefore, P(A ∪ B) = P(A) + P(B) − P(A ∩ B) = 0.15 + 0.25 − 0.05 = 0.35

Thus, 35% of all female patients admitted to the hospital receive either surgical treatment, obstetrics treatment, or both.

Lesson 3-4 — Conditional Probability

The event probabilities we have been discussing so far are often called unconditional probabilities since no special conditions other than those that define the experiment are assumed. Sometimes, on the other hand, we may have additional knowledge that might alter the probability of an event. A probability that reflects such additional knowledge is called the conditional probability of the event.

We represent the probability of event A, given that event B occurs by the symbol P(A | B) (it reads: the probability of A condition B) and is given by:

P(A | B) = (P(A ∩ B)) / (P(B)) (we call this formula 1)

Note that P(A ∣ B) ≠ P(B ∣ A) since

P(B ∣ A) = (P(A ∩ B)) / (P(A)) (we call this formula 2)

**The Multiplication Rule of Probability**

Formulas (1) and (2), after cross multiplication, can be written

* P(A ∩ B) = P(A ∣ B)P(B) (we call this formula 3)
* P(A ∩ B) = P(B ∣ A)P (A) (we call this formula 4)

**Example**:

The human resources director of a company constructed the following table, which describes the talent and motivation levels of company employees. The number in each cell is the number of employees that fall into that category.



(a) Suppose that an employee has a low level of motivation. What is the probability that he/she has a medium level of talent?

(b) Suppose that an employee has a low level of talent. What is the probability that he/she is highly motivated?

**Solution:**

* A = {an employee of the company is highly motivated.}
* B = {an employee of the company has a medium talent level.}
* C = {an employee of the company has a low talent level.}
* D = {an employee of the company has a low motivation level.}

1. We must find P(B ∣ D)

* P(D) = 28 / 270, and P(B ∩ D) = 5 / 270 Therefore,
* P(B ∣ D) = P(B ∩ D) / P(D) = (5 / 270) / (28 / 270) = 5 / 28 = 0.179
* This result indicates that 17.9% of those who have a low level of motivation also have medium talent levels.

1. In this part, we must find

* P(C) = 6 / 270, and (A ∩ C) = 30 / 270 Therefore,
* P(A | C) = P(A ∩ C) / P(A ) = ( 30 / 270 ) / ( 6 / 270 ) = 30 / 68 = 0.441
* This result indicates that 44.1% of the employees who have low levels of talent are also highly motivated.

Lesson 3-5 — Independent Events

Two events A and B are said to be independent if the outcome of one does not influence the outcome of the other. Mathematically, events A and B are independent if and only if

P(A | B) = P(A) (5)

This criterion means that the occurrence of event B does not affect the occurrence of event A. The "if and only if” statement implies that if events A and B are independent, then P(A | B) = P(A); and conversely, if P(A | B) = P(A), then events A and B are independent. Events that are not independent are said to be dependent.

When the Multiplication Rule of Probability (formulas (3) or (4)) and the independency criterion (5) are combined, we obtain the following alternative criterion for establishing the dependency or independency of two events:

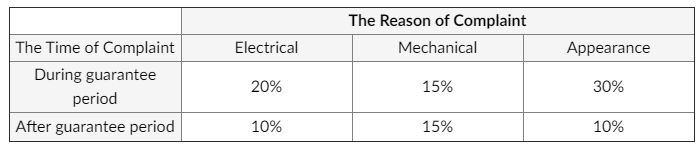
Events A and B are independent if and only if

P(A ∩ B) = P(A) P (B) (6)

In practice, formula (6) is easier to use than formula (5) because it does not include conditional probabilities. Hence the above equation may serve as a criterion to determine whether two events are dependent or independent.

**Example**

A manufacturer of an electromechanical kitchen utensil conducted an analysis of a large number of consumer complaints and found that they fell into six categories shown in the table below:



Define the following events:

* A: {Cause of complaint is product appearance},
* B: {Complaint occurred during the guarantee period}.
* Are A and B independent events?

**Solution**:

Events A and B will be independent when P(A ∩ B) = P(A) P(B) . Otherwise, they will be dependent events.

P(A ∩ B) = 0.30,

P(A) = 0.30 + 0.10 = 0.40

P(B) = 0.20 + 0.15 + 0.30 = 0.65 ⟶ P(A)P(B) = (0.40) (0.65) = 0.26

Since (A ∩ B) ≠ P(A)P(B), we conclude the A and B are dependent events.

Lesson 3-6 — Research Questions and Sample Space

Some events of interest in the real world include the event that someone is diagnosed with cancer or signs up for your mailing list. For the example of renewing a subscription to a publication, the event we are interested in might be the case that all of our customers renew their subscription.

An experiment in probability and statistics is the same as any other experiment; it is a planned and documented procedure carried out to verify, refute, or validate a hypothesis. It consists of a number of trials which are used as the basis to compute probabilities. Experiments provide insight into cause-and-effect by demonstrating what outcome might occur given certain information.

Depending on the design of the experiment, we may be able to

* discover something new,
* test a hypothesis (such as "do two different groups of people do things differently?"), or
* demonstrate something that is already known (for example, tossing a coin).

For example, to determine if someone is going to renew their subscription to a publication, the experiment might consist of collecting data about customers over a period of time (a subscription period, such as a month). This data would include whether or not they renewed their subscription that period.

In an experiment, the question is important. These questions have nothing to do with each other, they simply demonstrate the way you ask the question. How you get your sample and how a subject answer your question determines the patterns that emerge in your data?

* What percentage of all US college students spends more than one hour traveling to get to class?
* What is the probability that a randomly selected student spends more than an hour traveling to class?
* Do women typically take public transit more than men?
* What is the typical number of kindle books a student in this class owns?

Let’s start by identifying the sample space S for each research question.

**Defining Sample Space**

In order to answer the first research question, we would need to take a random sample of U.S. college students, and ask each one "Do you spend more than an hour traveling to get to class?" Each student should reply either "yes" or "no." Therefore, we would write the sample space as:

S={yes, no}

In order to answer the second research question, we would need to know how long each student spends traveling to class. One way (not the only way) of getting this information is to ask each selected student to report how long it took for them to travel to class today. In this case, if we let h denote the number of hours traveled, we would write the sample space as:

S = { h∶ h ≥ 0 hours}

If we conducted a random study to answer the third research question, how would we define our sample space? It depends on how we want to answer the question. If we asked a random sample of men and women "how many days did you take public transportation last month?" then we would write the sample space as:

S = {0,1,2,...,31}

Finally, if we were interested in estimating information about students who took the data analytics course in the past decade when trying to answer the fourth research question, we could survey all current data analytics students who own a Kindle. The question could be phrased as: "How many books do you have on your Kindle?" In that case, we would write our sample space as:

S = {0,1,2,...}

Note: the size of the sample space is often denoted as n(S), for example, a sample size with six possible outcomes would be n(S) = 6.

**Trial**

A Trial is a single execution of an experiment. In other words, an experiment may consist of several trials. Trial and experiment tend to be used interchangeably. But, for more accurate data collection, in any study, we run an experiment multiple time. In order to differentiate we can say that the experiment is the set of procedures that we use each time, and the trial is executing on those procedures. One trial is one instance of the experiment.

If our experiment is to toss a coin four times to get more thorough data, each toss would be a trial.

An outcome is the result of a single trial. In the case of tossing a coin, the potential outcome of each trial is either a head or a tail. In the case of rolling a die, the outcome is one of the six numbers represented on each side of the die.

In our example of renewing a subscription to a publication, an individual trial is a customer in a given period, and the outcome is whether the subscription was renewed or not. A single outcome is either true (the subscription is renewed) or false (it’s not), and the sample space is the set theoretic definition and not the probabilistic one.

It is true in set theory that the sample space is the collection of all elements of that set. However, in probability theory the sample space is the collection of all possible outcomes of an experiment. Therefore, from the probabilistic point of view, the sample space is the collection of all possible outcomes; that is S = {T, F}

Lesson 3-7 — Probability Distributions

Now that we know a bit about the different types of data and ways in which it can be described, we can combine that with our knowledge of probability to learn about probability distributions.

A distribution reflects the values in a data set and how often those values occur. Distributions can be used to either describe or generate data.

Given a data set that we are trying to understand, we might plot the data, and then plot various distributions (defined by continuous mathematical functions) on top of the dataset, and see which distributions most closely fits the data. This allows us to describe the data. For example, we can say: this data follows a normal distribution with specific parameters, in this case, mean and standard deviation.

Knowing which distribution, a dataset follows allows us to draw conclusions and make inferences about the data based on what we know about that particular distribution. Alternatively, we can investigate whether some data follows a certain distribution. For example, the binomial distribution models successful trials. If we run an experiment that we believe follows the binomial distribution with given parameters, we can test whether our empirical data matches the expected data based on the mathematical formula for the binomial distribution.

In many situations, the value of a variable X cannot be stated with certainty. In such cases, the variable may take one of several possible values, and there are probabilities associated with each value the random variable takes. Such a variable is called a random variable (r.v.). In other words, a random variable is a variable whose value is determined as a consequence of a random experiment.

Suppose that a banker is interested in knowing the number of customers using an ATM machine in a given day. Since this number may vary from one day to another, and it isn’t possible to say with certainty what it will be, then it’s a random variable.

There are two types of random variables: discrete and continuous. A discrete random variable is a random variable that can assume only certain and clearly separated (discrete) values. A continuous random variable, on the other hand, may assume any value in a given range or interval, and there is a continuity of the different possible values it may take.

The following are examples of discrete random variables:

* The score of a baseball team in a game.
* The number of tails observed when a coin is tossed twice.
* The average of grades in a test.
* The number of bank customers using an ATM machine in a given day.

The following are examples of continuous random variables:

* The amount of soda in a randomly chosen 12-ounce can of a particular brand.
* The amount of time it takes a randomly chosen runner to run a mile.
* The area of a randomly drawn circle.
* The amount of rainfall in a randomly chosen summer day in New Orleans.

A probability distribution for a random variable is a table, a graph, or a formula that assigns probabilities to each value the random variable takes. In the next example, we will construct a tabular and a graphical probability distribution for a certain random variable:

**Example**

Toss a coin twice and let X be the number of tails observed. Construct a probability distribution for X.

**Solution:**

Recall that the sample space of this experiment is: S = {HH, HT, TH, TT}. As far as the variable X is concerned, the possible outcomes are: either no tail will be observed (X = 0), or only one tail will be observed (X = 1), or 2 tails will be observed (X = 2). Note that X is a discrete r.v. taking separated values. Next, we will assign probabilities to each possible value of X:

The outcome of HH corresponds to the case when X = 0. Therefore, the probability that X = 0, denoted by P(X = 0) or simply P(0), is given by:

P(X=0)=P(HH)=1/4=0.25

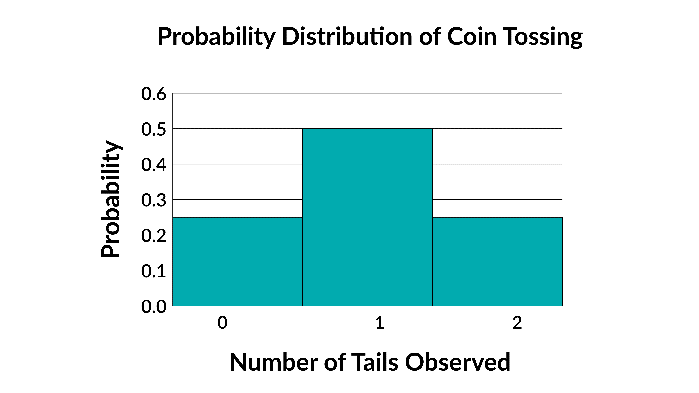
Similarly, P(X=1)=P(HT,TH)=1/4+1/4=0.5

P(X=2)=P(TT)=1/4=0.25

Number of Trials of x, Probability of Outcome P(x) – 0, 0.25, 1,0.50, 2,0.25

Lesson 3-8 — Probability Distributions: Discrete Distributions

We will now construct a probability distribution for X using the table above:



There are two important characteristics of a discrete probability distribution:

* The probability of a particular value is between 0 and 1, inclusive. That is, 0 ≤P(x)≤ 1
* The sum of the probabilities of all values is 1. That is, ∑ P(x) =1

**Example 1**

The probability distribution of a discrete r.v. X is given below.

X, P(x) ----- -3,0.15, 0,0.05, 2,0.40, 4,0.15, 5,0.10, 6,0.05, 8,0.10

Based on the above table, let’s find the following probabilities:

1. P(X = 4)
2. P(X > 5)
3. P(X < 2)

**Solution:**

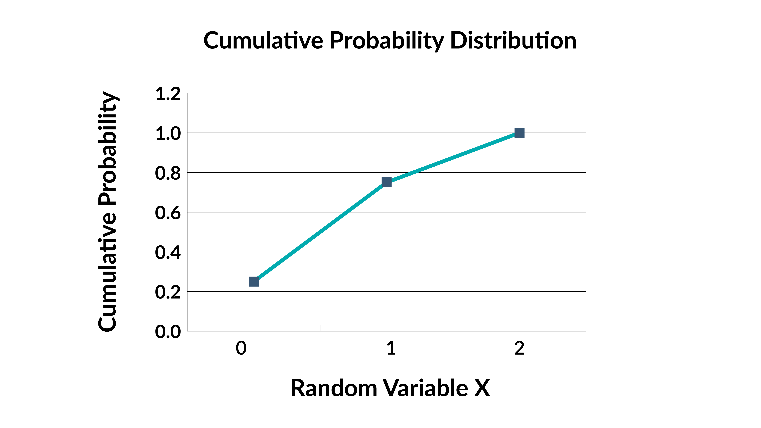
1. To find P(4), we must use the fact that all probabilities must add up to 1 (the second characteristic of a discrete probability distribution). P(4)= 1-(0.15 + 0.05 + 0.40 + 0.10 + 0.05+ 0.10)= 0.15
2. P(X>5) = P(6) + P(7) = 0.05 + 0.10 =0.15
3. P(X<2)=P(-3)+ P(0)=0.15+ 0.05 =0.20

Lesson 3-9 — Probability Distributions: Cumulative Probability Distribution

For any value x of a random variable X, the cumulative probability is given by P(X ≤ x). For example, the following is a cumulative probability distribution for the random variable X of the coin toss example.

x, P(X ≤ x) ---- 0,0.25, 1,0.75, 2,1

Graphical cumulative probability distribution for the r.v of example 1

  
The mean (or the expected value or the expectation) of a discrete probability distribution is denoted by: μ ( or by Exp(X) or by E(X)) , and is given by:

μ = E( x) = E x p( X) = ∑ x P( x)

The interpretation of μ is as follows: In the long run, if the experiment is repeated a very large number of times, and each time the observed value of the r.v. is recorded, μ will be the average of all observed values.

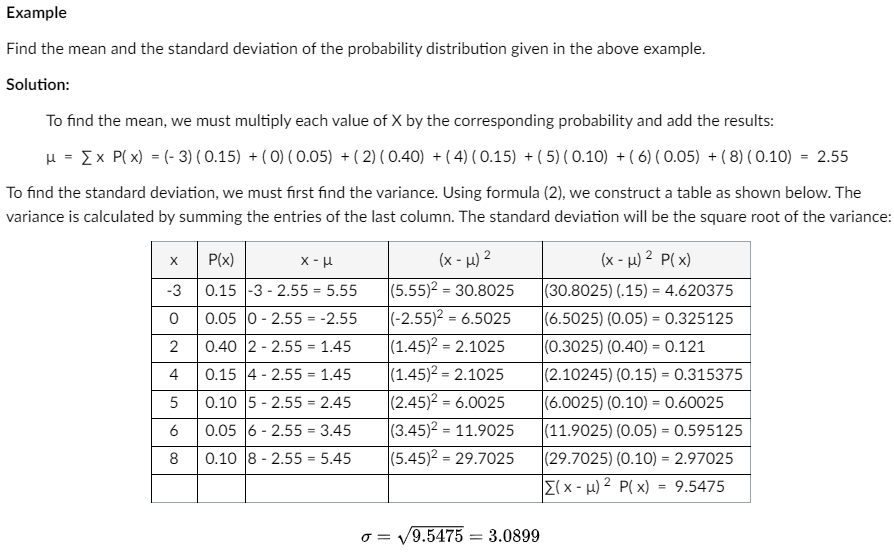
The variance and the standard deviation of a discrete r.v. are given below. The standard deviation of x is a measure of the spread of its probability distribution.

Variance: σ2 = ∑(x - μ)2 P(x)

In terms of Expectations, formula (2) can be written as:

σ2 = ∑ (x − μ)2 P(x)

**Standard Deviation:**



Instead of the formula above for σ2, the formula below (called the short-cut formula as it requires fewer number of operations than the formula above) can also be used to calculate the variance:

σ2 = (∑ x2P(x)) - μ2

This can also be written in terms of Expectations as: σ2 = E (X2) - (E(X))2 (2)

Lesson 3-10 — Discrete Probability Distributions

**Binomial Probability Distributions**

There is an important group of discrete random variables that share certain common characteristics. The random variables that possess those characteristics are called binomial random variables. The probability distribution of a binomial random variable is called the binomial probability distribution. The common characteristics of this group of random variables are:

* The experiment consists of n identical and independent trials.

For example, consider the experiment of tossing a coin 20 times. All 20 trials are identical, and the outcome of each trial does not influence the outcome of the other one.

* There are only two possible outcomes in each trial; one of which is called the Success (S), and the other is called the Failure (F).

In the coin-tossing example, there are only two possible outcomes: either observing a tail or observing a head.

* The probability of success, denoted by P(S) = p, stays the same from trial to trial. The probability of failure is denoted by q. Note that p + q = 1, or q = 1 - p.

For example, assuming a fair coin, and if success is defined to be observing a tail, then p = 0.5, and q = 1 - p = 1 - 0.5 = 0.5 , and these quantities stay the same from trial to trial.

* The binomial r.v. X is the number of successes in n trials.

**Example**

According to a recent survey, 60% of adult Charlotte residents are in favor of the new zoning regulations introduced by the city council. A random sample of 15 residents is chosen. Let X be the number of residents in the sample who are in favor of the new zoning regulations. Describe why X is a binomial random variable.

**Solution:**

The experiment consists of n =15 identical inquiries from the residents about their opinion of the new zoning regulations. Since any resident’s opinion does not influence another’s, then the trials are independent. Each inquiry may result in one of the two possible outcomes: either a randomly chosen resident is in favor of the new regulations (success) or he/she is against them (failure). Since 60% of the population is known to be in favor, then the probability of success is p = 0.6, and this probability stays the same from trial to trial. The probability of failure is q = 0.4. The number of residents among the 15 who are in favor is the binomial random variable X.

**Computing the Binomial Probability**

The probability that the binomial r.v. X takes the value x is denoted by P(X = x), or simply by P(x), and is given by:

P(x) = [n!/ X!(n − x)!]px qn – x

In this formula, n is the sample size (or the number of trials), p and q are respectively the probabilities of success and failure, and for any whole number n, the quantity n! (reads: n factorial) is computed as follows: n! = n(n - 1)(n - 2) ... (3)(2)(1), and in particular, 0! = 1 For example, 5! = (5)(4)(3)(2)(1) = 120, 15! = and (15 - 6)! = 9! = 362880.

**Poisson Distributions**

The Poisson distribution is a discrete distribution which applies when we want to calculate the probability that an event will occur a given number of times in a given interval.

Examples include how many phone calls a call center gets in a week, the number of emails someone gets in a day, or how many people visit a restaurant at lunch hour. There are some conditions for which the Poisson distribution applies:

* Each event is independent of the others
* The rate of occurrence is constant (does not vary over time)
* Two events cannot occur at the same instant
* The probability of an event occurring in an interval is proportional to the length of that interval

Let's discuss these conditions in greater detail:

* First, each event has to be independent of the others – the occurrence of one event cannot impact the occurrence of another event.
* Next, the rate of occurrence needs to be constant – it doesn't vary over time. Notice in one of the examples that we restricted the restaurant visits to the lunch hour. One can assume the rate is fairly consistent over the lunch hour, but the rate is probably very different in the afternoon hours.
* Next, two events can’t happen at the same instant; or, if it does, there’s some way to order the events. In the example of the restaurant, two people might arrive at the same time, but they self-organize into a line where the first one is served first.
* Finally, the probability of the event occurring in an interval needs to be proportional to the length of that interval. This means that if the interval were cut in half, the number of occurrences in the smaller interval should be 50% of the number of occurrences in the larger interval.

**Poisson Distribution - Formal Definition**

The Poisson distribution is defined mathematically as:

k: number of times event occurs in interval,

integer λ: average number of event occurrences in interval, λ=0

X ∼ Poisson (λ): random variable X follows Poisson distribution

P(k) = λke−λ/k!

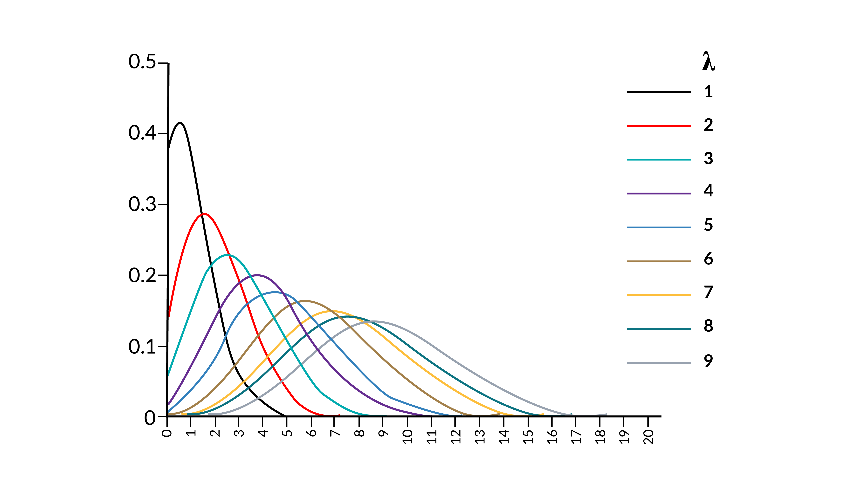
E[X] = 𝜆

**Example: Poisson Distributions**

As the rate of some event increases, the center of the curve moves to the right, as shown in the sample Poisson distribution curves below.

Because these distributions are discrete, the probability of some number of occurrences of the event can be directly read from the y-axis, without performing an integral. The x-axis represents the number of occurrences of the event.

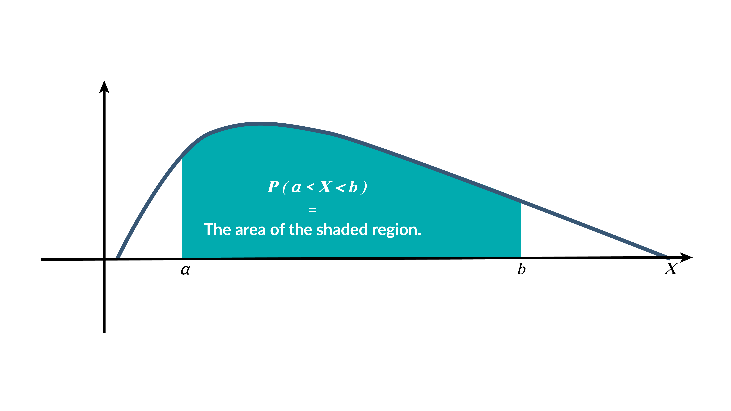
Consider the following scenario: as the average number of pieces of mail one receives in a day, λ, increases, the less likely it is that the person will receive 0 pieces of mail. This can be understood by considering the curves below for increasingly large values of λ.



This image shows the Poisson distributions associated with receiving specific number of pieces of mail.

Lesson 3-11 — Continuous Probability Distributions

A continuous random variable takes an infinite number of values in a certain range. It’s usually a result of measuring some quantity, such as the amount of time, the amount of rainfall, or the height of a person. The curve describing the probability distribution of a continuous random variable is, therefore, a smooth curve. This smooth curve is called the probability density function for the random variable X. The probability that a continuous random variable X assumes a value between two values a and b is represented by the area under the curve between the points a and b. This is a probability distribution curve where the shaded area represents the probability of the random variable X is between values a and b.



The probability distribution of a continuous random variable. The probability that the r.v. X assumes a value between any two points is represented by the area under the curve between the two points. In summary, we have: P(a < X < b) = The area under the curve between the points a and b Note that the total area under a continuous probability distribution equals 1 unit.

Furthermore, since a single point does not constitute any area, the probability that a continuous random variable assumes a particular value is zero; that is, for any real value a, P(X = a) = 0.

The above fact indicates a major difference between continuous random variables and their discrete counterparts. Consequently, in a continuous probability distribution, the inclusion or the exclusion of the end points is not significant.

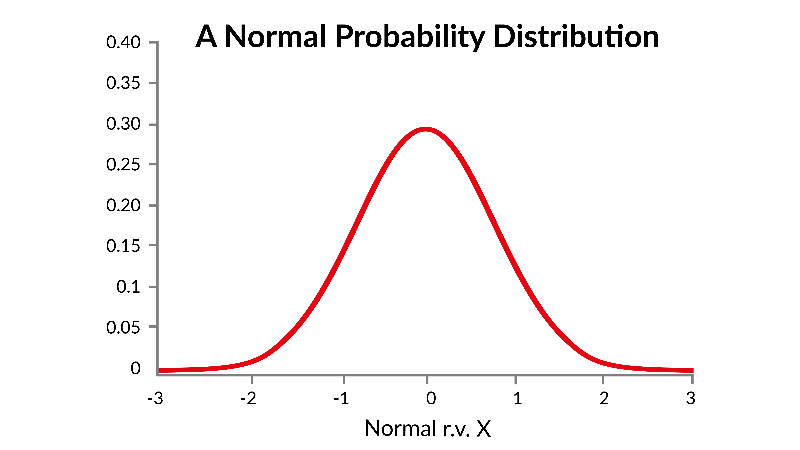
Therefore, for a continuous random variable, P(a < X < b) = P(a ≤ X < b) = P(a < X ≤ b) = P(a ≤ X ≤ b)

The most important member of the continuous probability distributions is the Normal probability Distribution.

**Normal Probability Distribution**

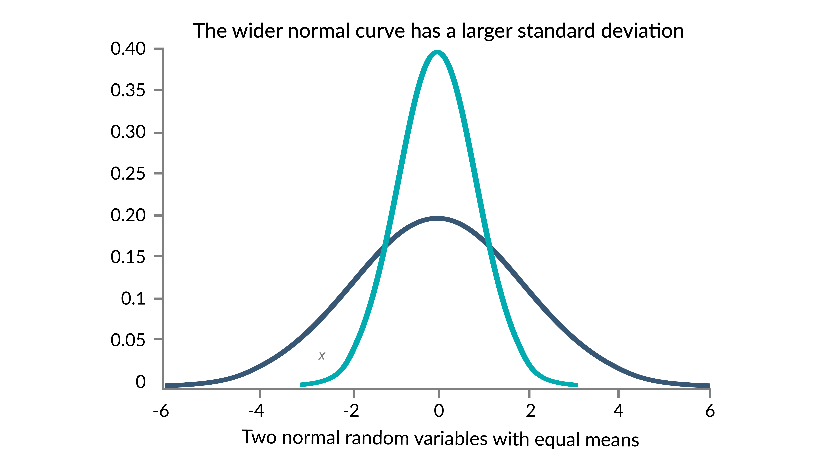
A special type of a continuous probability distribution is the normal probability distribution. A random variable possessing a normal probability distribution is called a normal random variable.

A normal probability distribution is mound-shaped and symmetric and has a single peak at the middle of the distribution, at which the mean and the median and the mode coincide. Half the area under the curve is above this center point and the other half is below it and the two halves of the curve are mirror images of each other. Also, the curve falls off smoothly in either direction and gets closer and closer to the x-axis but never touches it. The figure below shows a normal probability distribution. This is a graph of a normal probability distribution for normal random variable X.



The mean μ lies at the center of the distribution on the horizontal axis. The standard deviation σ determines how wide the distribution will be. A larger value of σ will give rise to wider normal distributions.

In the figure below, two normal distributions with equal means but different standard deviations are sketched. The wider distribution possesses the larger standard deviation. Since the total area under each normal curve must be 1 unit, the narrower distribution is taller.



This image shows the graph of two normal random variable with equal means. One has a larger standard deviation than the other, making it shorter and wider.

Since in normal distribution, finding probabilities is done by finding the areas, calculating probabilities involves finding certain portions of the area under the curve.

Having realized the significance of μ and σ in the shape of normal distribution, we must also note the common characteristics that all normal distributions share. The importance of realizing those common characteristics becomes evident when we realize that there are an infinite number of different normal distributions that possess different means and standard deviations, and it will not be possible to construct a table for each and every one, as we were able to do for discrete random variables.

Fortunately, the common characteristics that all normal distributions share allow us to choose one member of the family and use it to find probabilities for all normal distributions. This member is called the standard normal distribution. It's a very simple member of the family in a sense that it has a mean of zero and a standard deviation of one. A standard normal variable is denoted by Z.

Any normal random variable X with a mean μ and a standard deviation σ can be converted into a standard normal random variable Z by the following formula:

z = x – μ/σ

Conversely, when the above formula is solved for x, we can convert the standard normal random variable to a normal random variable with mean μ and standard deviation σ : x = σz + μ

The following facts are important to remember about normal distributions:

1. The curve is symmetric with respect to the vertical line passing through z = 0
2. The total area under the curve equals one

Lesson 4-1 — The Work of Data Analytics

**Versions of “Translation” of “Analytics”**

Business intelligence combines a broad set of data analysis applications, including ad hoc analysis and querying, enterprise reporting, online analytical processing (OLAP), mobile BI, real-time BI, operational BI, cloud and software as a service BI (SaaS BI), open source BI, collaborative BI, and location intelligence. BI technology also includes data visualization software for designing charts and other infographics, as well as tools for building BI dashboards and performance scorecards that display visualized data on business metrics and key performance indicators in an easy-to-grasp way. BI applications can be bought separately from different vendors or as part of a unified BI platform from a single vendor.

| **Date Modeling & Reporting** | **Business Intelligence & Visualization** | **General Data Analysis** | **Advanced Analytics** |
| --- | --- | --- | --- |
| **Function** |  |  |  |
| ETL (Extraction, Transformation, and Loading) process | Data reporting (KPIs, metrics) |  | Complex, large-scale analysis involving a good amount of data and Predictive Analytics / Data Mining / Machine Learning |
| Data storage (data warehouse, data lake) | Dashboard |  |  |
| Data modeling (schema design) | OLAP (cube, slice, and dice) |  | Credit risk scoring |
| Data indexing, referencing, querying and reporting | Ad hoc data analysis |  | Marketing mix modeling |
|  | Automatic monitoring |  |  |
|  | Operational, real-time BI |  |  |
| **Tools** |  |  |  |
| Oracle SQL | MS SQL | IBM Cognos| Oracle OBIEE | Excel | R / Python /  SAS / SPSS |
| MySQL, Postgre SQL | More tool options | Various tools | Various tools |
| **Roles** |  |  |  |
| IT department | IT department | Data Analyst/Financial Analyst | Statistician/Modeler/Data Scientist |
| DB analyst/Data Engineer | Data Analyst/Business Analyst |  |  |

Typical Quantitative Techniques Used in Advanced Analytics Several quantitative techniques apply to analytics projects, including:

| **Type** | **Description** |
| --- | --- |
| **Simulation** | Randomized repetitions of a set of discrete events in order to model real-world systems and phenomena (e.g., queues) |
| **Optimization** | Algorithm selects the best possible outcome, subject to satisfying constraints |
| **Matrix Algebra** | Calculations involving matrices solve multidimensional problems |
| **Fitting Functions to Data** | Also called “curve fitting,” using numerical methods to interpolate data |
| **Survival Analysis** | Originally used by life scientists, but adopted by marketers and actuaries |
| **Time Series** | When data are “auto-correlated,” such as time-dependent data (also called “Box-Jenkins”) |
| **Predictive Analytics and Machine Learning** | |
| **Classical Statistics** | * Descriptive: calculates metrics to characterize the distribution of values of data (mean, standard deviation, range, etc.) * Predictive: estimates parameters using historical data and making predictions of future outcomes (multivariate regression, generalized linear regression, etc.) |
| **Learning** | * Unsupervised learning: characterizes the data to establish classes without using explicit metrics, e.g., k-means clustering * Supervised learning: Classify and describe the data with predefined ‘labels,’ e.g., decision trees |
| **Bayesian** | Used to augment classical analysis when there is prior knowledge about how the data was generated |

Lesson 4-2 — Careers in Data Analytics

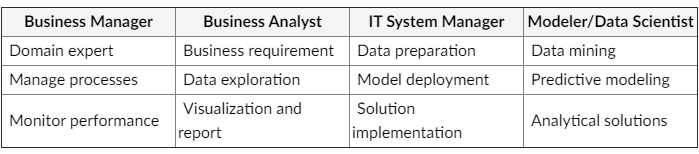
**Careers in Analytics**

Many skill sets come to play during the course of a data analysis project workflow. These include hacking skills, math and statistics knowledge, and substantive expertise. Whereas traditional research relies primarily on math/statistics and domain expertise, modern data science typically draws from all three sets. The hacking skills reflect themselves in the understanding of available tools and technologies.

**Roles**

Analytics development roles can generally fall into one of the four groups: business analyst, modeler/data scientist, IT system manager, and business manager. The business analysis defines the business requirements, performs some data exploration then creates visualizations and reports to support and evaluate the requirements. The data modeler/scientist seeks to create predictive models and mine the data for unforeseen relationships. The IT system manager is responsible for data preparation, model deployment, and solution implementation. In other words, the IT system manager manages the entry of data, the internal application of models to the data, and the customer-facing solutions.

**Roles in Data Analytics**



**Full Suite of Data Scientist Skill Set**

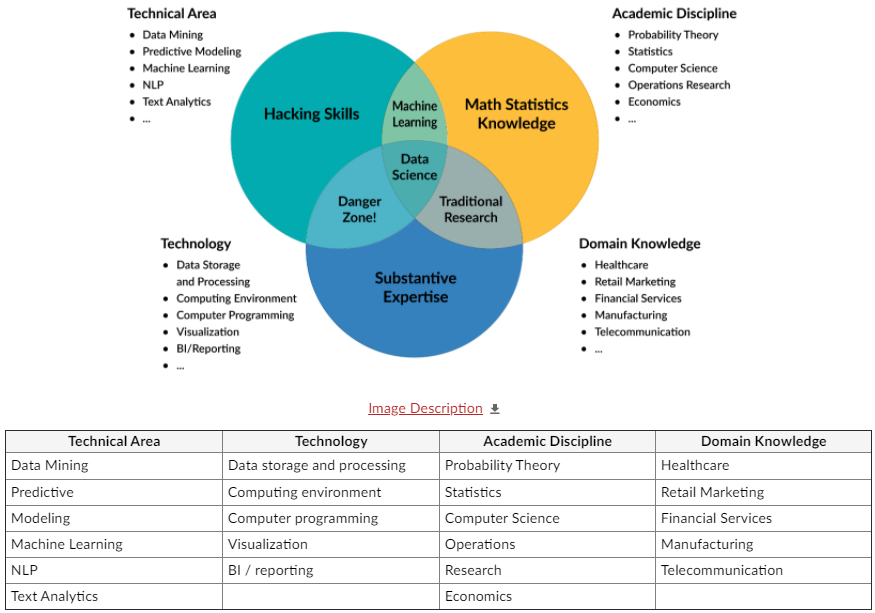


Diagram showing the types of skills needed to be a Data Scientist

**Required Resources**

* [Data Science Career Path &amp; Progression (Links to an external site.)](https://medium.com/analytics-and-data/data-science-career-path-progression-ab5140cfbc84) by Julien Kervizic from Hacking Analytics  
  This article discusses Data Science as a career and the four main axes of the work.
* [Data Analytics Market in 2020: Trends, Forecasts &amp; Challenges (Links to an external site.)](https://www.cognetik.com/blog/data-analytics-market-in-2020-trends-forecasts-challenges/#:~:text=According%20to%20the%20same%20study,to%20C-level%20business%20planning.) from Cognetik This article discusses the increase need for data scientists
* [Gartner Top 10 Data and Analytics Trends (Links to an external site.)](https://www.gartner.com/smarterwithgartner/gartner-top-10-data-analytics-trends/) from Gartner  
  This article discusses the top 10 trends in 2020.

Lesson 5-1 — Analytics Evolution

The rapid evolution of data analytics has been accelerated by advances in:

* large scale Internet connectivity
* data warehousing
* data analysis and mining algorithms development

Data analytics spans an increasing number of industries and within these industries multiple functional areas. Data-intensive industries such as retail, financial services, healthcare, and telecommunication benefit directly from data analytics. While the increasingly large online retail industry leverages data analytics for demand forecasting and merchandising, the financial industry benefits from data analytics in credit and risk estimates, the healthcare industry in drug trials, and the telecommunication industry in product subscriptions. Within each of these industries, data analytics contributes to marketing and sales promotions, customer relationships, and distributions.

|  |  |
| --- | --- |
| Data-Intensive Industries | For Various Functional Areas |
| Retail and Consumer Goods | Marketing and Sales |
| * Demand forecasting * Merchandising | * Direct and digital marketing * Product recommendation * Sales force optimization |
| Financial Services | Pricing and promotions |
| * Underwriting * Credit and risk * Fraud * Insurance premium and loss estimate | * Product pricing and promotion planning |
| Healthcare | Customer relationship management |
| * Drug trial * New drug R&D | * Loyalty program and lifecycle management * Customer acquisition and retention, cross-sell, and up-sell |
| Telecommunication | Distribution network and supply chain |
| * TV network * Product subscriptions |  |

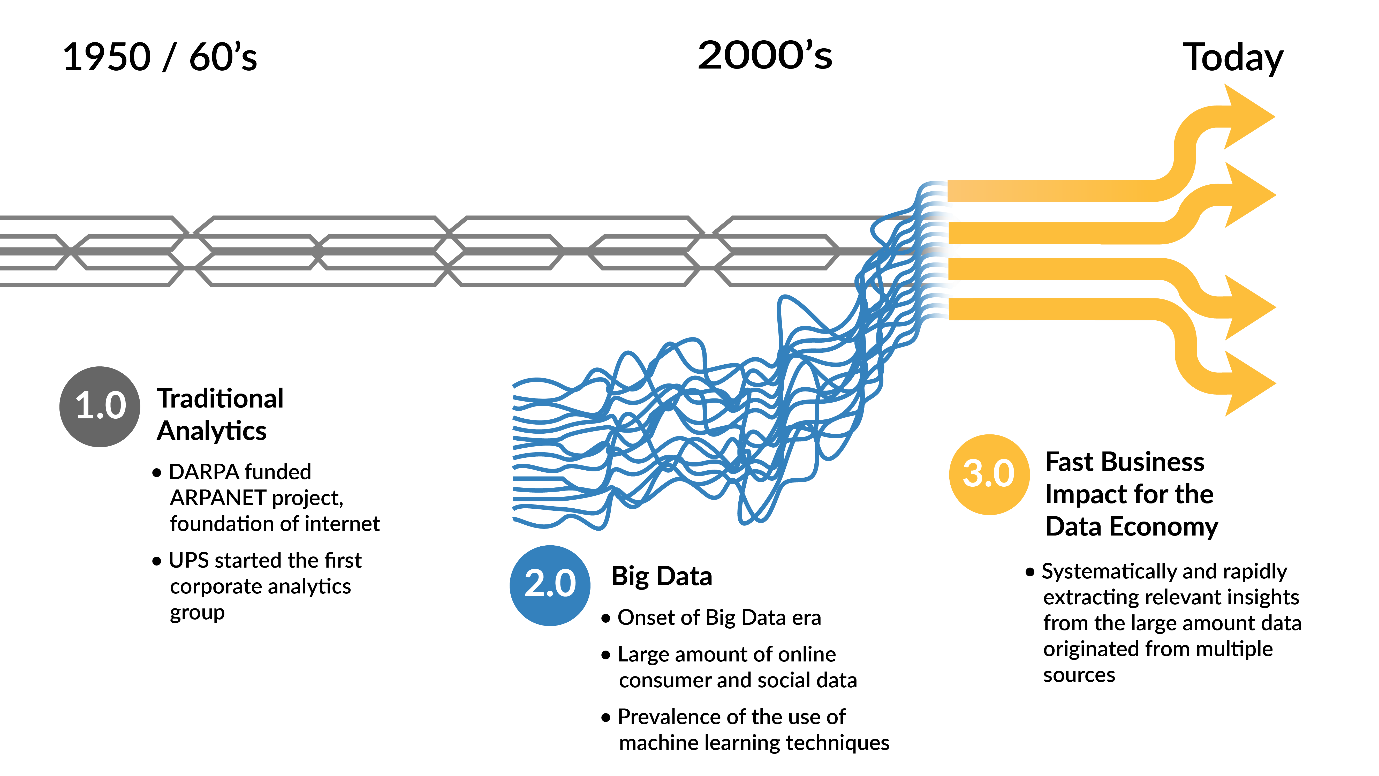
Data analytics is driven by vast collections of data gathered through the Internet. It is, therefore, appropriate to align the dawn of modern data analytics, or Analytics 1.0, to that of the Internet or ARPANET in the 1950’s, soon followed by UPS’ Analytics Group. Advances in Internet accessibility and speed through router fiber optic cables, large scale storage technologies in the 2000’s created data pools that only machine learning algorithms could effectively process. This can be considered the second period, or Analytics 2.0, in the evolution of analytics corresponding to the big data era. The use of data analytics as a continuous guide to business decisions marks the era of Analytics 3.0.

DARPA was created in 1958 as the Advanced Research Projects Agency (ARPA) by President Dwight D. Eisenhower. Its purpose was to formulate and execute research and development projects to expand the frontiers of technology and science, with the aim to reach beyond immediate military requirements.

DARPA supported the evolution of the ARPANET (the first wide-area packet switching network), Packet Radio Network, Packet Satellite Network and, ultimately, the Internet and research in the artificial intelligence fields of speech recognition and signal processing, including parts of Shakey the robot. DARPA also funded the development of Douglas Engelbart's NLS computer system and The Mother of All Demos; and the Aspen Movie Map, which was probably the first hypermedia system and an important precursor of virtual reality.

Lesson 5-2 — Evolution Timeline

This is an image of the evolution of Big Data with the three main stages, Traditional Analytics, Big Data and Fast Business Impact for the Data Economy.



The timeline above highlights the following aspects of each of the three main stages of Big Data evolution.

Three Main Stages of Big Data

* 1950s/1960s: Traditional Analytics

Darpa funded ARPANET project, foundation of Internet

UPS started the first corporate analytics group

* 2000s: Big Data

Onset of Big Data era

Large amount of online consumer and social data

Prevalence of the use of machine learning techniques

* Today: Fast Business Impact for the Data Economy

Systemically and rapidly extracting relevant insights from the large amount of data originated from multiple sources

**Analytics 1.0**

Analytics 1.0, traditional analytics, or the first phase of data analytics drew from relatively small and structured data sets. The majority of the analytical activities pertained to reporting basic characteristics of the data. The mainly descriptive analytics was formed offline as “batch” processes performed on data collected the previous day. Analytics were marginal to strategy but were available to the decision-making process, still dominated by intuition.

From a technology perspective, this was the era of the enterprise data warehouse and the data mart. Data was small enough in volume to be segregated in separate locations for analysis.

This approach was successful, and many enterprise data warehouses became uncomfortably large because of the number of data sets contained in them. However, preparing an individual data set for inclusion in a warehouse was difficult, requiring a complex ETL (extract, transform, and load) process.

For data analysis, most organizations used proprietary BI (Business Intelligence) and analytics “packages” that had a number of functions from which to select.

**Traditional Analytics**

* Data sources relatively small and structured, from internal systems
* Majority of analytical activity was descriptive analytics or reporting
* Creating analytical models was a time-consuming “batch” process
* Quantitative analysts were in “back rooms” segregated from business people and decisions
* Few organizations “competed on analytics”—analytics were marginal to strategy
* Decisions were made based on experience and intuition More than 90% of the analysis activity involved descriptive analytics or some form of reporting.

**Analytics 2.0**

* Complex, large, unstructured data sources
* New analytical and computational capabilities
* “Data Scientists” emerge
* Online firms create data-based products and services

The demand for mining the growing volume of complex, large, and unstructured data sources gave rise to Analytics 2.0. Unstructured databases (NoSQL) came to light for better management of unstructured data, and parallel servers such as Hadoop began to expedite retrieval and operations for fast-flowing data. The need for advanced analytical and computational skills prompted the emergence of “Data Scientists.” Data analytics became essential for maintaining a business competitive edge in an environment where “agile is too slow” meaning that continuous development is preferred over the typical two-week agile cycle, and where the role of a consultant generating reports is superseded by unbiased, real-time computer analysis.

**Key Developments**

* Fast flow of data necessitated rapid storage and processing
* Parallel servers running Hadoop for fast batch data processing
* Unstructured data required “NoSQL” databases
* Data stored and analyzed in public or private cloud computing environments
* “In-memory” analytics and “in-database” analytics employed
* Machine learning methods meant the overall speed of analysis was much faster (from days to minutes)
* Visual analytics often crowded out predictive and prescriptive techniques
* “Agile is too slow”
* “Being a consultant is the dead zone”
* Information (and hardware and software) wants to be free
* Share your big data tools with the community

**Analytics 3.0**

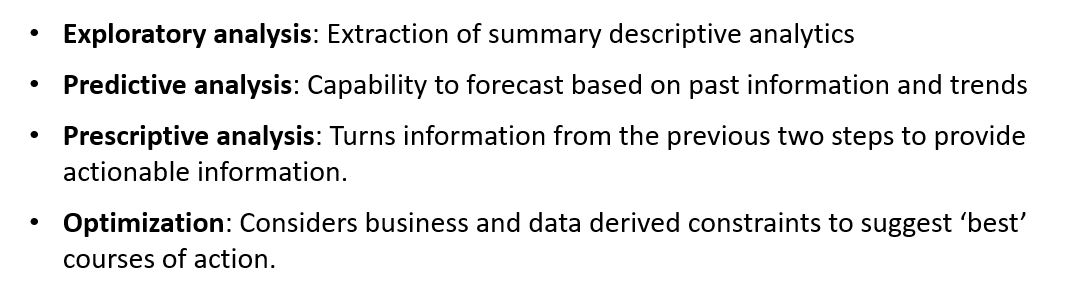
There is considerable evidence that large organizations are entering the Analytics 3.0 era. Such is the era where analytics is integral to running a business and an essential aspect of strategic planning. Trends and analyses instantly obtained for time intervals of interests provide rapid and agile insights. It’s an environment that combines the best of 1.0 and 2.0—a blend of big data and traditional analytics that yields insights and offerings with speed and impact.

**Characteristics**

* Analytics integral to running the business; strategic asset
* Rapid and agile insight delivery
* Analytical tools available at point of decision
* Cultural evolution embeds analytics into decision and operational processes
* All businesses can create data-based products and services
* "...virtually any type of firm in any industry can participate in the data economy."

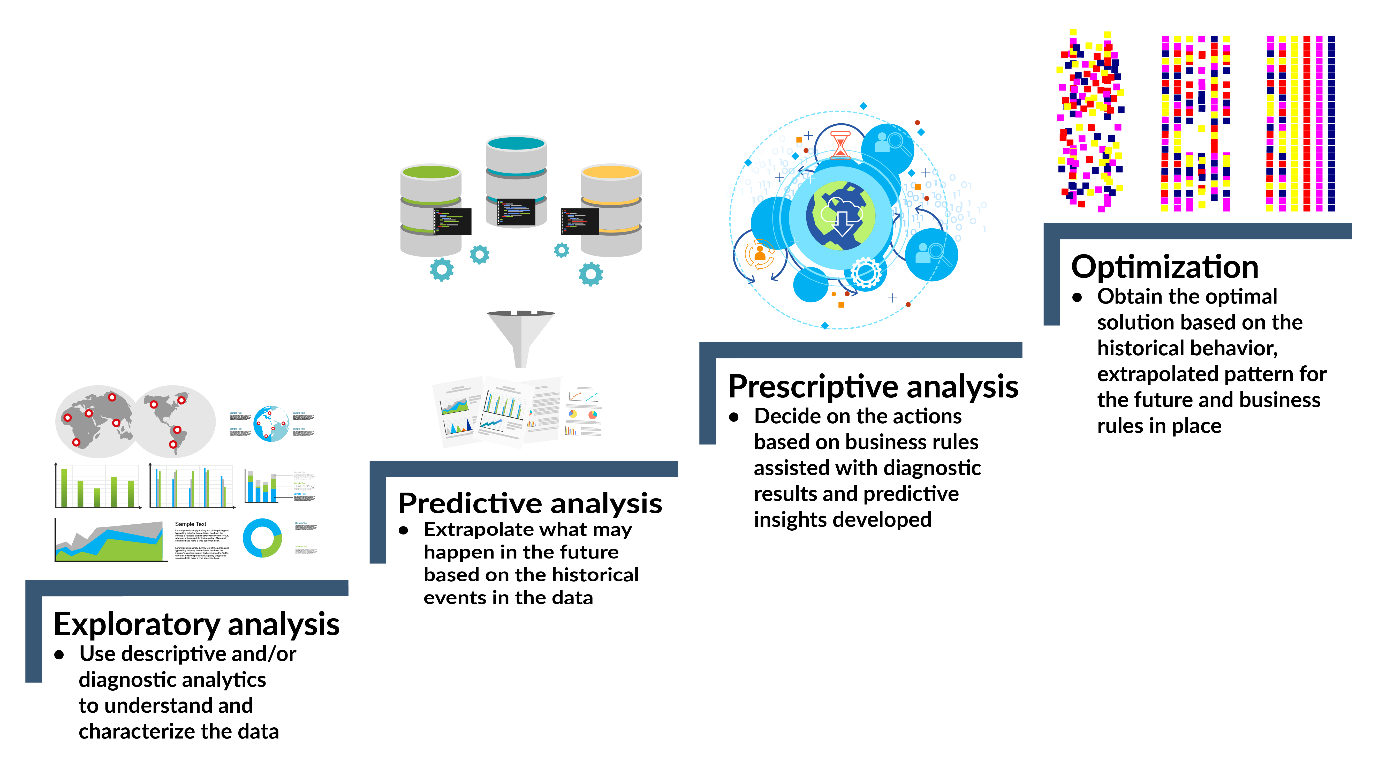
Lesson 5-3 — Analytics Development

The ladder of sophistication in data analytics can be viewed as the sequence of exploratory analysis, predictive analysis, prescriptive analysis, and optimization. Exploratory analysis pertains to the extraction of summary descriptive analytics; predictive analytics provides the capability to forecast based on past information and trends; prescriptive analysis turns information from the previous two steps to provide actionable information, and optimization considers business and data derived constraints to suggest ‘best’ courses of action.



**Analytics Applied at Various Sophistication Levels**

This image shows the 4 levels of applied analytics; Exploratory, Predictive, Prescriptive and Optimization.



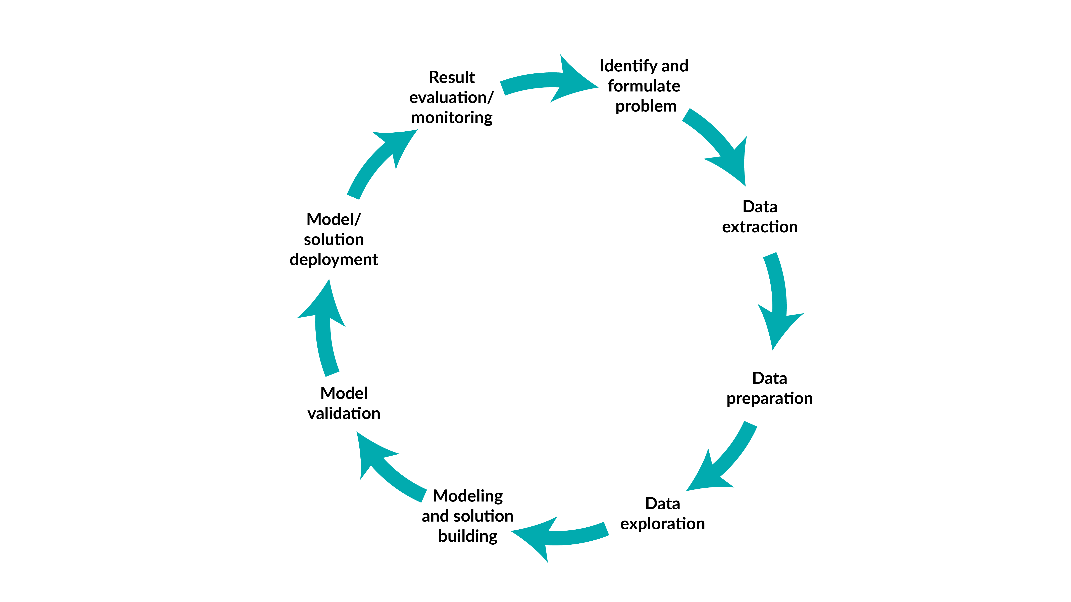
The above image outlines these levels of sophistication and provides a description of each level as follows:

* Exploratory analysis - Use descriptive and/or diagnostic analytics to understand and characterize the data.
* Predictive analysis - Extrapolate what may happen in the future based on the historical events in the data.
* Prescriptive analysis - Decide on the actions based on business rules assisted with diagnostic results and predictive insights developed.
* Optimization - Obtain the optimal solution based on the historical behavior, extrapolated pattern for the future and business rules in place.

Fundamental Elements in Analytics Development (click to open)

* Data: A data warehouse only stores data that has been modeled/structured, while a data lake is no respecter of data. It stores it all—structured, semi-structured, and unstructured.
* Processing: Before we can load data into a data warehouse, we first need to give it some shape and structure—i.e., we need to model it. That’s called schema-on-write. With a data lake, you just load in the raw data, as-is, and then when you’re ready to use the data, that’s when you give it shape and structure. That’s called schema-on-read. Two very different approaches.
* Storage: One of the primary features of big data technologies like Hadoop is that the cost of storing data is relatively low as compared to the data warehouse. There are two key reasons for this. First, Hadoop is open source software, so the licensing and community support is free. And second, Hadoop is designed to be installed on low-cost commodity hardware.
* Agility: A data warehouse is a highly-structured repository, by definition. It’s not technically hard to change the structure, but it can be very time-consuming given all the business processes that are tied to it. A data lake, on the other hand, lacks the structure of a data warehouse—which gives developers and data scientists the ability to easily configure and reconfigure their models, queries, and apps on-the-fly.
* Security: Data warehouse technologies have been around for decades, while big data technologies (the underpinnings of a data lake) are relatively new. Thus, the ability to secure data in a data warehouse is much more mature than securing data in a data lake. It should be noted, however, that there’s a significant effort being placed on security right now in the big data industry. It’s not a question of if, but when.
* Users: For a long time, the rallying cry has been BI and analytics for everyone! We’ve built the data warehouse and invited “everyone” to come, but have they come? On average, 20-25% of them have. Is it the same cry for the data lake? Will we build the data lake and invite everyone to come? Not if you’re smart. Trust me, a data lake, at this point in its maturity, is best suited for the data scientists.

Lesson 5-4 — Processes in Data Analytics



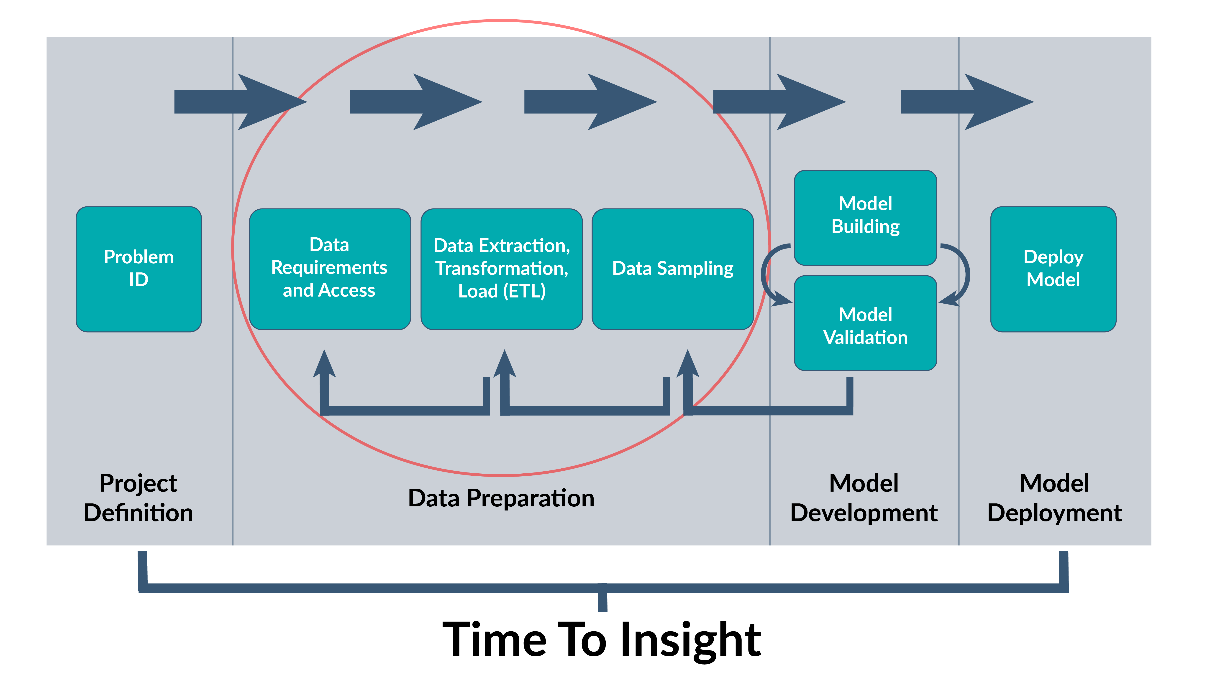
The processes in analytics development can be described as a cyclic pattern which typically begins with the identification of a problem or potential benefit from a data set. Once the data set is identified, it is extracted and prepared, or "cleaned up," for exploration. During the exploration process, the analysts develop seed ideas that will subsequently guide the modeling and solution-building tasks. Once the models are validated, the model is prepared for deployment on new (or live) data. This process will often repeat itself.

**Steps in an Analytics Development Process**

* Problem definition Define the analysis goal, scope, end deliverables, and potential approach
* Data requirement Identify the required data sources, data fields, and data coverage (e.g., temporal and geographical coverage)
* Data collection Extract, gather data from one or multiple sources and transform them into a consumable format
* Data cleansing Identify and fix the data errors, outliers, any unwanted data elements
* Data processing Transform the data format or data variables to meet the needs of downstream analysis, modeling, or development activities
* Model building Develop algorithms and build computer models based on the goal of analysis
* Model deployment Generate the results and insights from the models and implement the models in a systematic manner
* Communication Communicate the results and insights to the business users and make recommendations on subsequent actions

**Typical Analytics Project Workflow**

A typical analytics project workflow begins with the identification of a problem. The data preparation steps to follow include the data requirements and access, the data extraction, transformation, and loading (ETL), as well as data sampling (selecting a representative subset). During the model development phase, the models are built and validated. The model deployment concludes the workflow. It should be noted that significant effort on the workflow is absorbed in the data preparation phase.



**Typical Challenges and Pitfalls in an Analytics Project**

* Poorly defined problem Unclear goal of problem-solving Scope is unclear, e.g., how many SKUs to analyze Mixed objectives, e.g., economic analysis of a product category promotion for retailer versus CPG mixed
* Limited IT resources Cloud data can’t be acquired off-line within a reasonable time Can’t run the complete model due to computation limitation Too slow to generate results in real time Can’t share the data and results with network limitation
* Less-best approach Selected less effective modeling method Incremental accuracy doesn’t offset the extra complexity Inadequate or incorrect performance monitoring criteria
* Incomplete or incorrect data Primary dataset unavailable Complementary data unavailable, e.g., missing competitor pricing data Coarse data or aggregated data Very sparse data with missing value
* Insufficient communication Insufficient data dashboard to communicate the analysis result Lack of soft skills to sell the results and insights Long feedback cycle to make the results less relevant Very sparse data with missing values Isolated org structure to stifle collaboration

**Required Resources**

* [The Data Science Workflow (Links to an external site.)](https://towardsdatascience.com/the-data-science-workflow-43859db0415) by Konstantin at Towards Data Science This article discusses the process of data science and an overview of each step.
* [Analytics 4.0: it's all about taking better decisions (Links to an external site.)](https://towardsdatascience.com/analytics-4-0-its-all-about-taking-better-decisions-5b6b9d06fcea) by Lee Schlenker at Towards Data Science. This article discusses the next phase of analytics development - decision-making

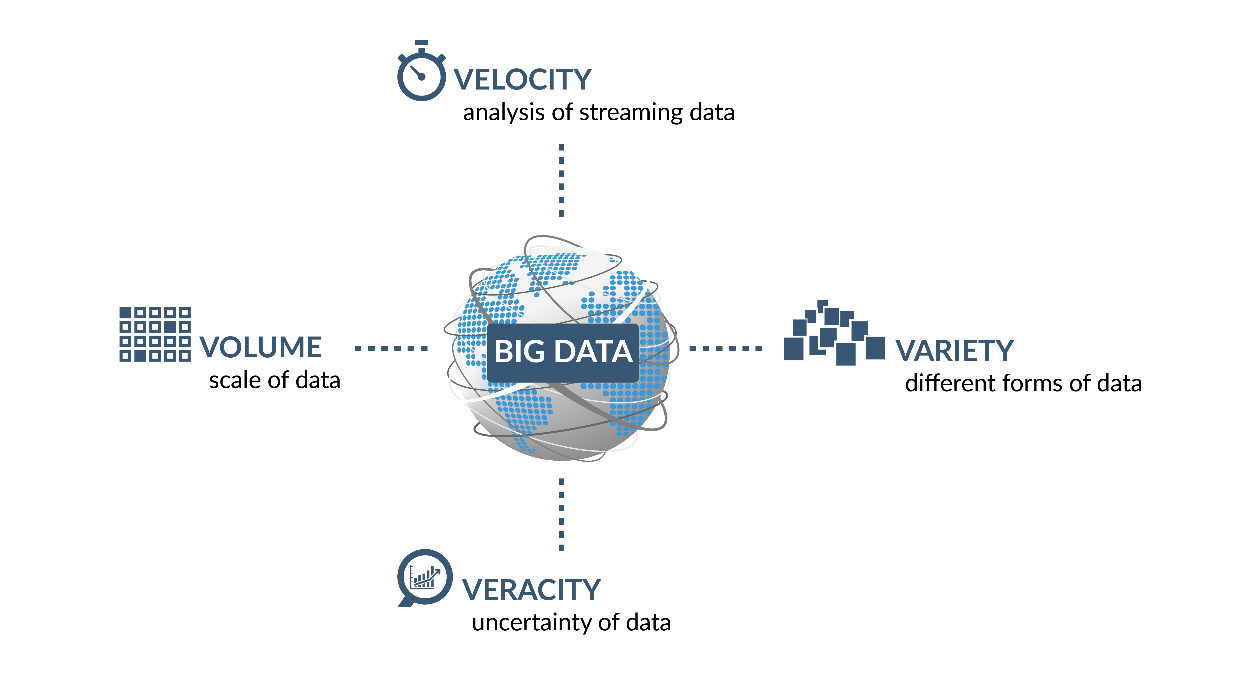
Lesson 6-1 — Introduction to Big Data

This unit examines the exponential growth of data and the challenges it poses to data analytics. A definition of analytics is proposed along with a survey of its evolution over the years. The tasks involved in modern data analytics are discussed including best industry practices and potential career paths.

**What is Big Data?**

The term ‘big data’ continuously evolves. What constituted 'big data’ a few years ago is now supplanted by a new minimum threshold. ‘Big data’ relates to the complexity of data from either the perspective of scale (size of the data), velocity (the amount of data transferred as a function of time), the variety of data (different data types), and the veracity of data (the uncertainty embedded in the data).

Greater than 40 trillion terabytes (40 zettabytes) of data is projected to be created by the end of 2020. This would constitute a 300-fold increase in the 15 years since 2005. The velocity aspect of ‘big data’ describes its increasing high transfer rates (velocity). The sources and types of data can be highly diversified. An aircraft processing center receives a variety of data generated from aircraft and monitoring equipment. Finally, a major challenge which permeates most ‘big data’ is veracity or the ‘cleanliness’ of the data. It requires the use of data correction techniques during processing.



4 Vs of Data Analysis presented in the image above are:

* Volume: scale of data
* Variety: different forms of data
* Velocity: analysis of streaming data
* Veracity: uncertainty of data

Originally defined by Gartner (2012) as "Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization,” veracity was addressed as another distinct characteristics of Big Data.

**Why Does it Matter?**

* Study shows for companies with a large amount of data, typically less than 1% are being effectively utilized.
* 89% of business leaders believe big data will revolutionize business operations in the same way the Internet did.
* 83% have pursued big data projects in order to seize a competitive edge.

'Big data' refers to the act of harnessing the information contained in the vast amounts of data being generated on a daily basis. Whether it is sensor data, intelligence data, weblogs, or any other form of data, being able to quickly process and analyze this information and display it in a coherent fashion has created great challenges across organizations.

**What Does the Field Say About Big Data?**

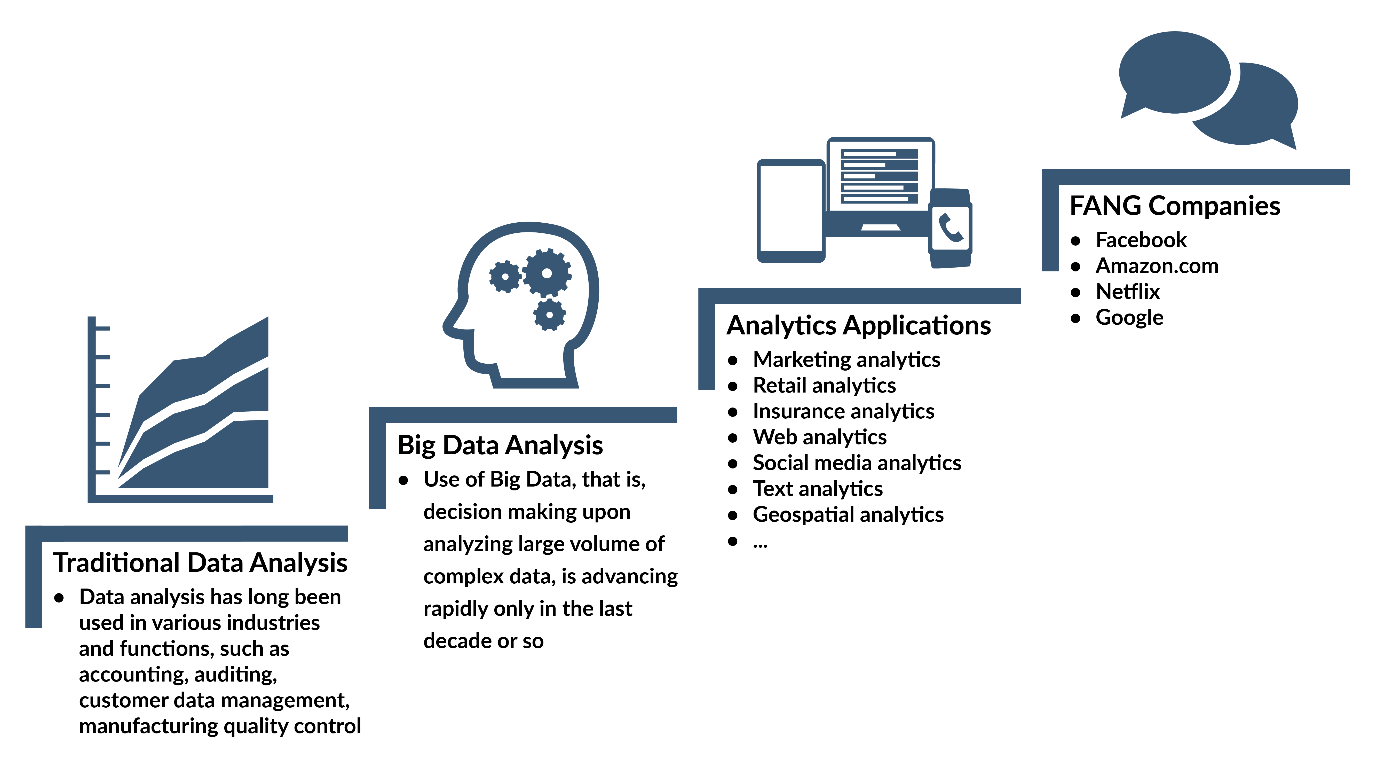
* Data-driven projects become more prevalent
* Data can’t be directly used without pre-processing
* Need data-wrangling tools
* Complex data requires a programmatic approach
* A common ”pain point” in a data organization
* Good amount of data becomes more of an asset than liability
* 90% of workers work on data analysis projects to some extent
  + Of those who do work on data projects, 93% said that the data requires some amount of pre-processing and it either takes up a large portion of time, or they do not have the necessary tools to clean the data
  + 45% said that the data is too messy and requires programming knowledge to clean
  + 52% said that the data needs to be joined from multiple disparate sources
  + Less than 20% said the data is too big

Lesson 6-2 — From Traditional Data Analytics to Big Data Analytics

Data analysis is not a new discipline. It has a long-established history in domains such as accounting, customer management, and quality control. Data analysis was often performed at regular time intervals to review and consolidate past activities. As storage increased (the cost of 10MB hard drives 20 years ago is now equivalent to 2TB), along with the availability of new tools to process the data, former time-consuming tasks have become feasible at much higher scales. The era of big data has thus ushered an exponentially increasing flow of information. More data also often signifies more complex issues. New fields have emerged to leverage the large-scale availability of data in order to increase competitiveness: marketing and retail analytics, web analytics, and social media analytics have all become mainstream. A clear trend points to increased and more sophisticated use of big data analytics.

**An Evolving Path from Traditional Data Analysis to Big Data Analytics**

This image shows the evolution from traditional analysis to big data analytics. It shows these steps: traditional data analysis, big data analysis, analytics applications and FANG companies.



**Traditional Data Analytics**:

Data analysis has long been used in various industries and functions, such as accounting, auditing, customer data management, manufacturing quality control.

**Big Data Analysis**

Use of Big Data, that is, decision making upon analyzing large volume of complex data, is advancing rapidly only in the last decade or so.

**Analytics Applications**

* Marketing analytics
* Retail analytics
* Insurance analytics
* Web analytics
* Social media analytics
* Text analytics
* Geospatial analytics

**FANG Companies (Term coined as an acronym for most popular and best performing companies**

* Facebook
* Amazon
* Netflix
* Google

**How Big Data Analytics is Different**

**It applies to a large volume of data**

In the early 1990s, a 20MB Seagate drive sold for over $500. A safer, faster 2TB drive can be purchased today for less than $100. With such availability of storage, an overwhelming amount of data is stored, much of it without adequate tools for proper use and management. It is estimated that the total amount of data generated doubles every two years.

1 MB= 1000 KB 1TB =1M \* 1M = 1012

The world’s population is currently at ~7 billion (7000 million). Considering personal use of data (home computing and mobile devices) and the massive data generation within industry, it is conceivable that 2.5 quintillion bytes (25 million terabytes) could be generated each day.

**It spans multiple disciplines**

Big data analytics is multidisciplinary in nature. It spans a wide scope of industries and disciplines, including:

* Statistics (descriptive statistics, inferential statistics)
* Computer science (information theory, computer algorithm)
* Computer engineering (computer programming, computer architecture)
* Mathematics (linear algebra)
* Operations research (simulation, mathematical optimization)
* Economics (econometric modeling)
* Business and others (domain knowledge)

Modern day data analytics spans multiple disciplines. Statistics is at the core of data analytics providing a numeric summary of large sets of data: descriptive statistics summarizing available data and inferential statistics providing a perspective of largely non-available data from samples. Computer science provides the knowledge for effectively creating algorithms while computer engineering helps understand and create complex infrastructures such as data warehousing and distributed computing. Operations research (simulation and optimization) help tackle challenges such as public health preparation and selecting the best transportation options. In addition, many domains utilize and contribute to data analytics.

**It requires specialized tool**

Big data analytics requires powerful tools and technologies, including:

* Data storage and processing (data warehouse, DB system, Cloud)
* Computing environment (multi-threading, parallel computing, Hadoop)
* Analysis and development tools (R, Python, SAS, Java, C++)
* Visualization, dashboard, and Business Intelligence (BI) supporting applications (Tableau, Oracle OBIEE, IBM Cognos)

Data analytics has dramatically shifted from simple use Lotus 1-2-3 spreadsheets to very complex data storage and processing architectures. The cloud (remote computing/storage platforms) serves a vast number of offices, home, and mobile personal devices. The open source Hadoop framework enables large scale distributed storage and computing. As such, data can be processed more efficiently and stored more safely (redundancy). Novel analysis tools have evolved both in the open source and industry communities. An example is the R language which began a little over 20 years ago and is now widely utilized in data analytics. Python is another example of a widely used, very powerful tool for data analysis. It supports parallel processing leveraging the multi-core personal computing platforms. An array of visualization tools is now available. Great visualization tools such as Tableau support interactive exploration of data as never before possible. On the open source side, great interactive visualization servers can be set up in a short amount of time through R and R Shiny.

**It presents overarching projects and org-structure**

Big data analytics requires a complex development process and organizational support for:

* Data collection, storage, and management
* Business requirement generation and translation into technical specification
* Technical development process (building, testing, and quality control of models)
* Analytic model deployment and learning
* New roles in the organization (data analyst / data scientist / chief data officer)

Data analytics often requires a complex development process which in turn requires an organizational support structure. The following tasks can require input from different entities: data collection, storage, and management require a full-time team in many organizations with specialized skills. Once a model is ready for deployment (for instance, an online marketing platform), a development team will be responsible for converting prototypes into reliable customer-facing (internal or external) modules. Thus, the roles of data analysts (which span from data cleansing to primary data exploration), data scientists (primarily responsible for mining data) and chief data officers (managing data analysts/scientists) become necessary.

The big data we produce personally (social media, personal records, mobile device data, etc.) is multiplying daily and impacts how various industries assess individualized marketing. Throughout the remainder of this course, we will explore a lot of information about how different industries are impacted by big data, how they are making sense of it all, and the ways they are putting this data to beneficial use. We will also explore the production of personal data from individuals and examine the methods of data collection and data analysis that industries perform to benefit their organizations from the data individuals provide.

**Required Resources**

* [What is big data analytics? Fast answers from diverse data sets (Links to an external site.)](https://www.infoworld.com/article/3220044/what-is-big-data-analytics-fast-answers-from-diverse-data-sets.html) by Bob Violino from InfoWorld. This article discusses what sets big data analytics apart from traditional data analytics
* [Big Data Analytics (Links to an external site.)](https://www.datamation.com/big-data/big-data-analytics.html) by Cynthia Harvey at Datamation. This article discusses how big data analytics can benefit organizations.