Lec 2: Common Plots & Sets Overview

Common Plots

Part 1

Contents

- Bar Plot
- Pie Chart
- Histograms
- Scatter Plots
- Line Graphs
- Box & Whisker Plot

Plots, graphs, and charts are vital tools in both probability and statistics. In this unit the words plot, graph, visualization, and chart will be used interchangeably. Plots are graphical techniques for representing a data set. These visualizations are integral in expressing information about a dataset and making it easier for a human brain to understand. Plots can also make it easier to detect patterns, trends, or outliers in groups of data. Throughout the next block, many different types of plots will be introduced and explained. By the end of the block there should be an understanding of how different plots can be used to express different aspects of a dataset.

In particular, the plots which will be introduced in this block are bar charts, box plots, histograms, line graphs, pie charts, and scatter plots. While there are many other types of plots which can be used to convey information, these most basic plots are often the most effective at conveying information in a manner which is easily understood.

General Guidelines for Plotting

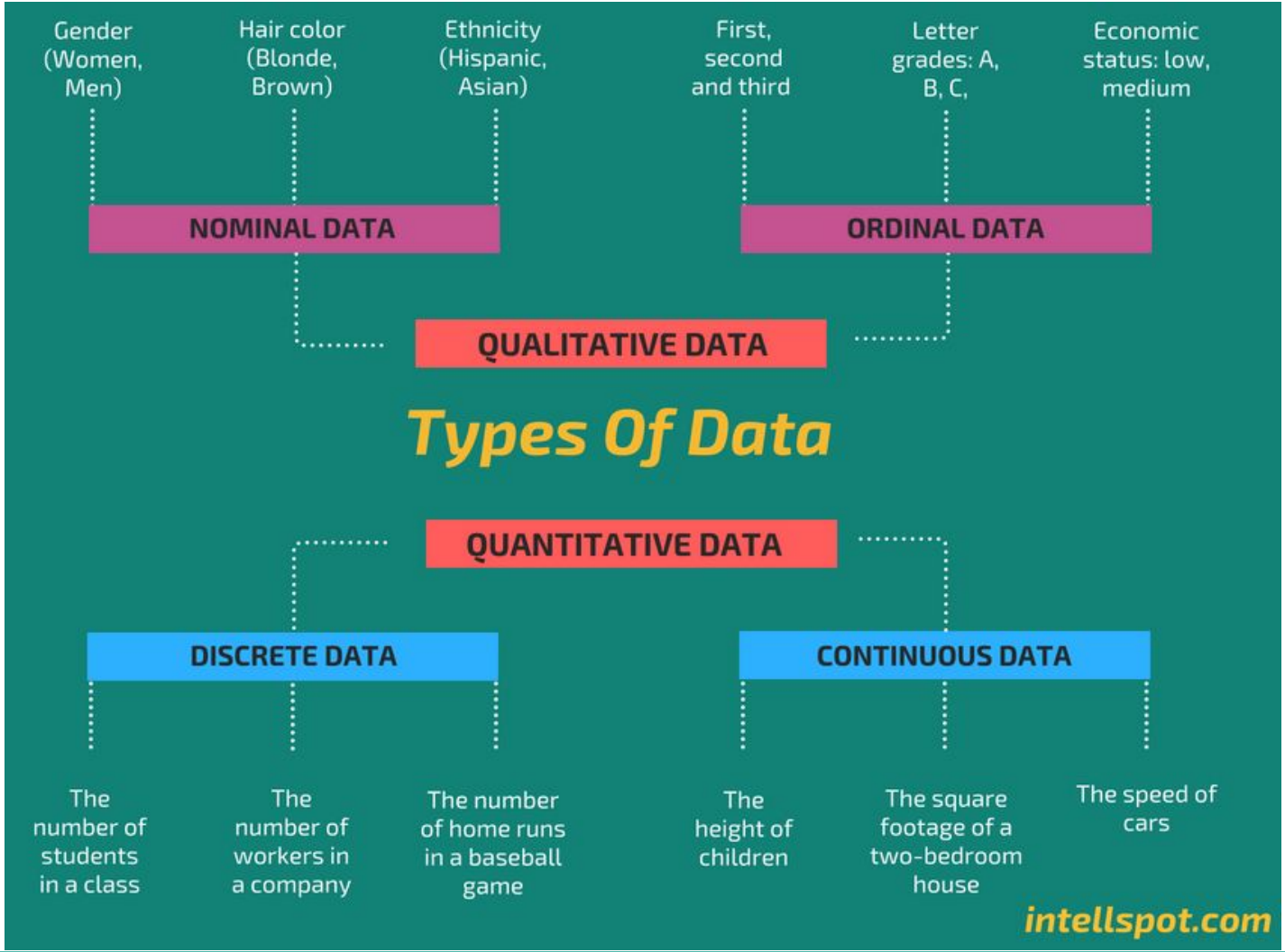
Good plots aim to present data in a meaningful way, in which a user can more easily interpret the data than if it were expressed as numbers in a table, for example. As a data scientist, you will often deal with very large datasets that cannot be conveyed explicitly; a graph or plot can be effectively used to summarize the data in a reasonable amount of space, such as an article or paper.

Good graphs convey data to the user quickly in an easily interpretable way. Plots can also show relationships in the data which would not be clear from looking at a list or a table. Plots and graphs can also be used to compare and contrast different datasets visually to identify differences.

Here are some best practices when visualizing data:

- Choose the most appropriate plot:
  - In the following lessons many of the most common plots will be introduced, it is imperative that one chooses a plot that expresses the data appropriately.
- Plots and graphs should be visually appealing, but function is always more important than form.
- Plots should not be misleading:
  - There are situations in which there is an opportunity to mislead a reader by constructing a plot in a specific manner, such as altering the scale of one or more axes.
- Clearly label your plots:
  - Plots which lack appropriate titles, axes labels, tick marks on axes, etc . . . can be very confusing or misleading to a user; without appropriate labels, plots are not very effective.

- Create plots with a less-is-more approach:
  - Avoid "chart-junk" such as pictures, thick gridlines, shadows, unnecessary 3d elements
  - Avoid the excessive use of unnecessary additional colors.
  - Avoid the use of excessively bright colors which are more difficult to see.
  - Any other elements which are not necessary to comprehend the information represented on the graph.
- Create accessible charts:
  - Consider colors and color combinations which are friendly to color-blind users.
- Only use graphs when necessary:
  - When dealing with small and simple datasets, the most effective way to describe the data to a user may be a simple list or a table.


Gender (Women, Men)
Hair color (Blonde, Brown)
Ethnicity (Hispanic, Asian)
First, second and third
Letter grades: A, B, C,
Economic status: low, medium
NOMINAL DATA
ORDINAL DATA
QUALITATIVE DATA
Types Of Data
QUANTITATIVE DATA
DISCRETE DATA
CONTINUOUS DATA
The number of students in a class
The number of workers in a company
The number of home runs in a baseball game
The height of children
The square footage of a two-bedroom house
The speed of cars
intellspot.com

Bar Plot

Figure 1: Bar Chart - GDP by Country

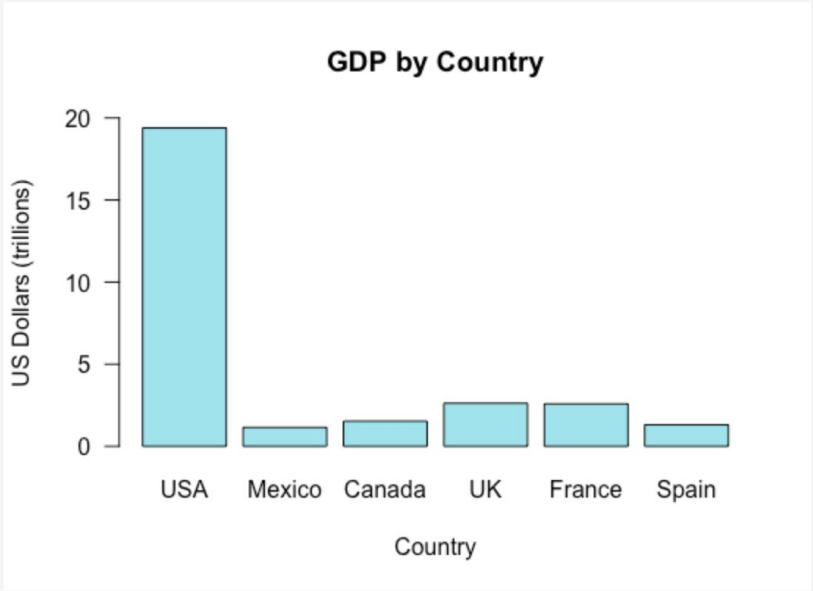
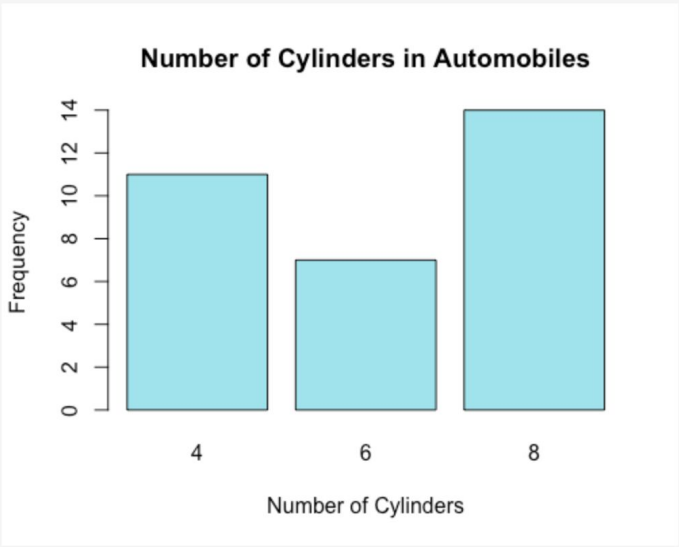


Figure 2: Bar Chart - Number of Cylinders by Frequency



Bar Plot Variations

Figure 3: Stacked Bar Chart - Total Revenue by Month and Region

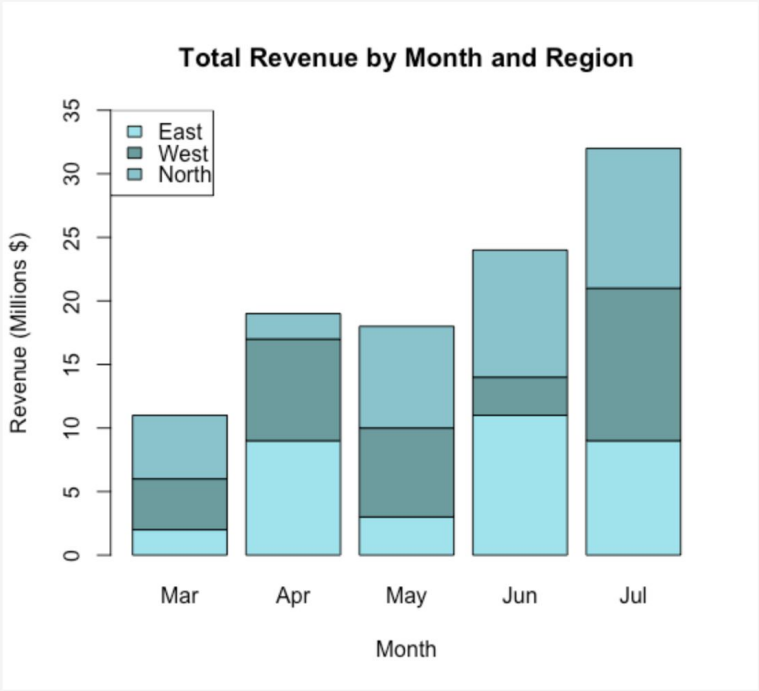
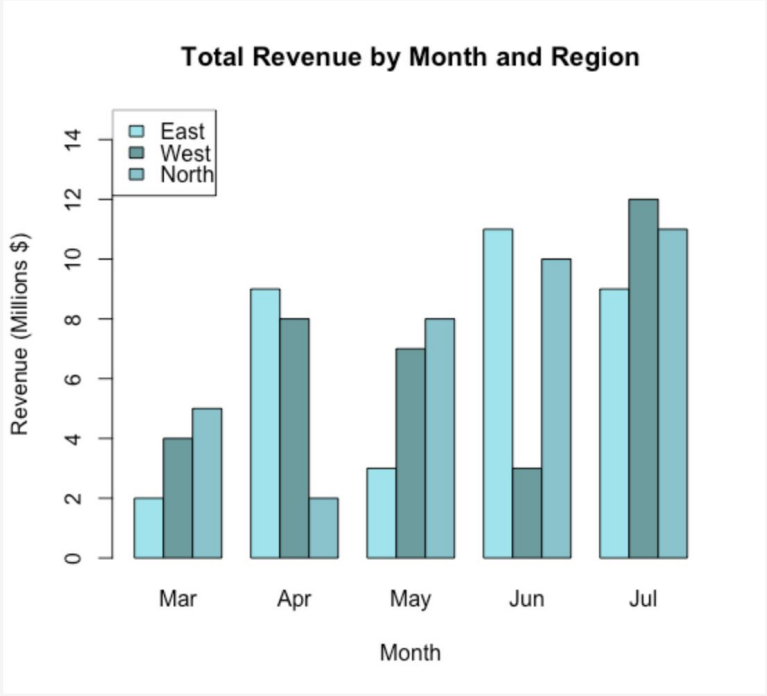
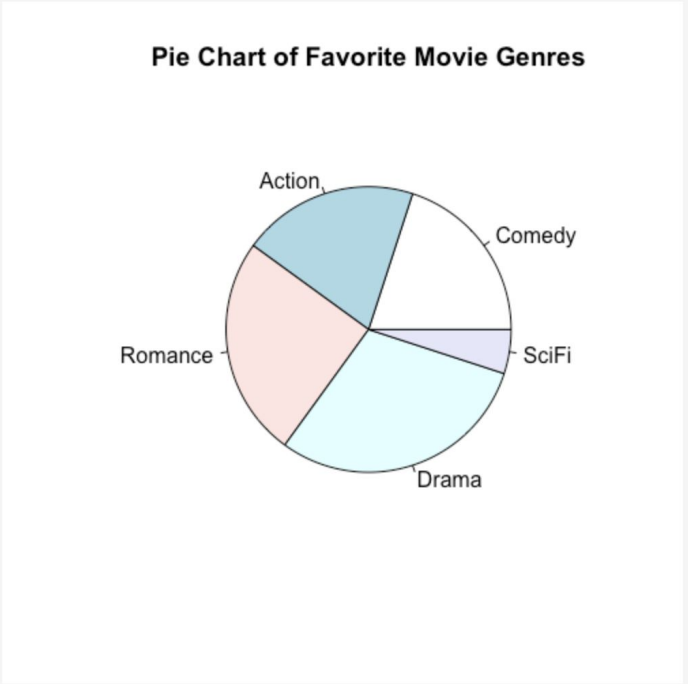


Figure 4: Side-by-Side Bar Chart - Total Revenue by Month and Region



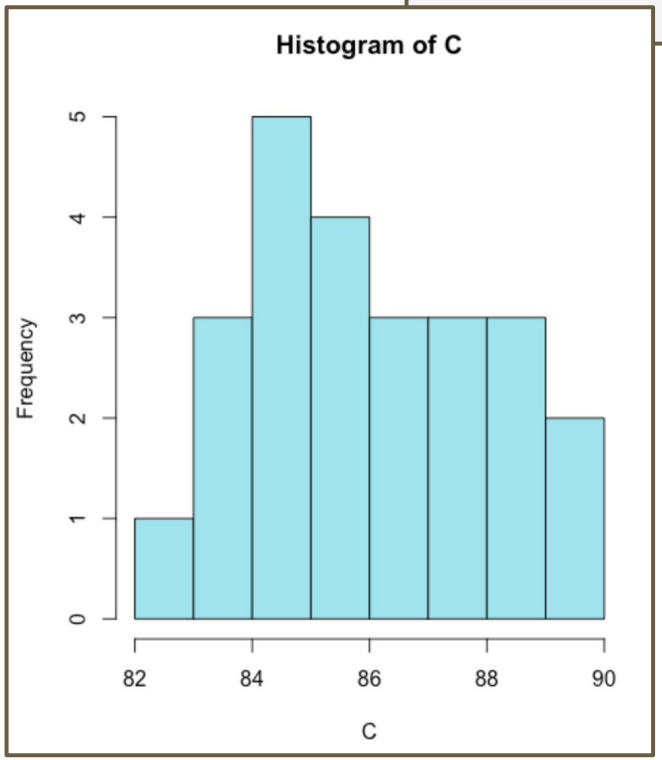
Pie Chart

Figure 1: A Simple Pie Chart



Histograms

$$C = \begin{bmatrix} 89.3 & 84.5 & 85.5 & 83.2 & 86.6 & 88.8 & 84.4 & 90.0 \\ 88.5 & 87.0 & 88.3 & 84.2 & 85.6 & 87.9 & 88.0 & 84.7 \\ 83.2 & 82.2 & 85.9 & 86.3 & 86.5 & 85.5 & 83.9 & 87.8 \end{bmatrix}$$



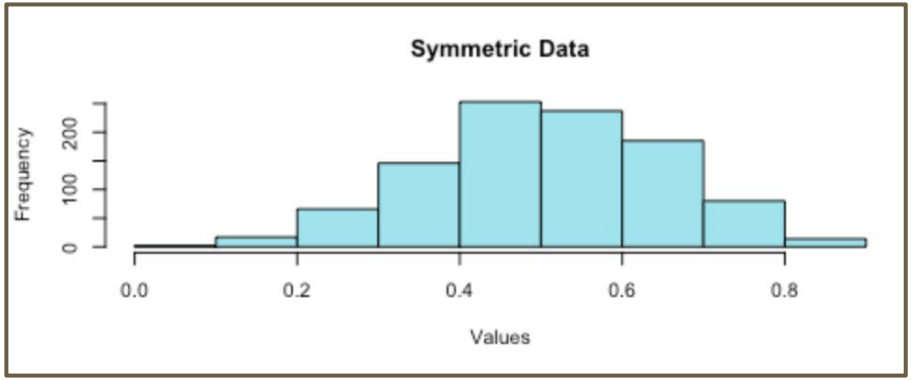
- A histogram is a chart which shows frequencies for intervals of continuous numeric values.
- These intervals are often referred to as "buckets" or a "bin". The y-axis of a histogram always represents the frequency, while the x-axis represents the values to be represented.
- The histogram is an effective way of visualizing how a numerical attribute of a dataset is distributed.
- **In other words it can be interpreted as an estimate of the probability distribution of a continuous variable.**

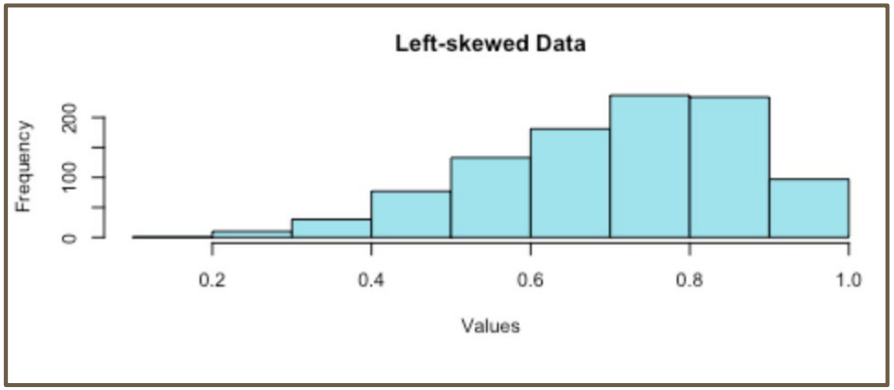
In the above plot, it can be seen that values with the highest frequency lie between 84 and 85. It is also evident that values between 82 and 83 are the least frequent. It is important to note, that convention dictates that each bin is left inclusive, and right exclusive. For example, in this plot our bins represent:

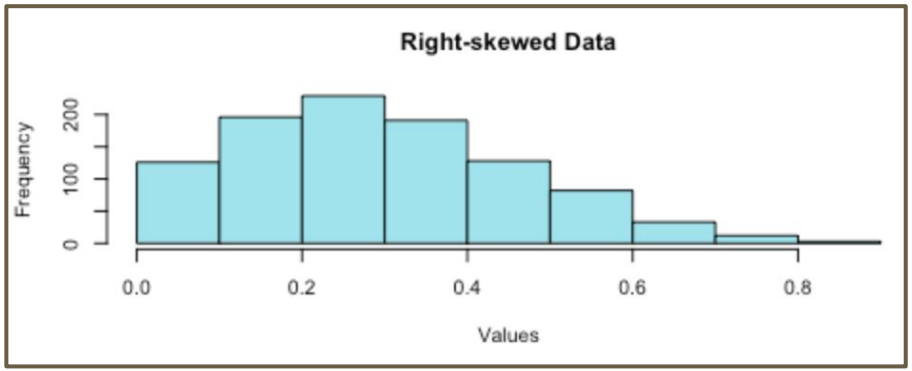
$$\{82.0 \leq C < 83.0;\ 83.0 \leq C < 84.0;\ 84.0 \leq C < 85.0;\ \ldots\}$$

Also note, that there is no space between the bars on this chart, which indicates that the data being represented is continuous. A bar chart can be used in a manner similar to a histogram to show the frequency of discrete values.

Histograms to Identify Skewness

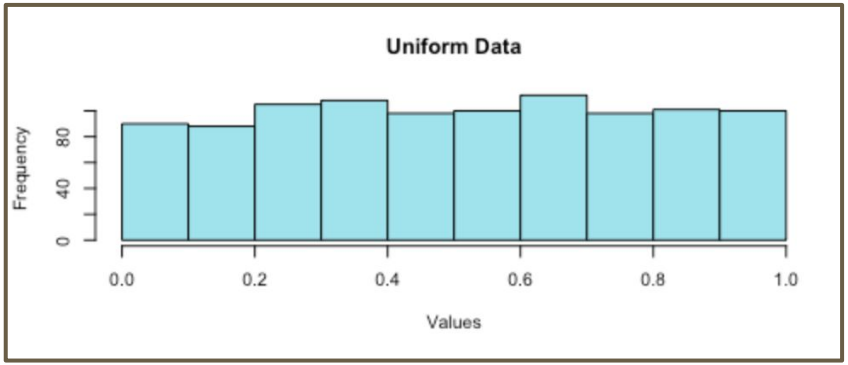


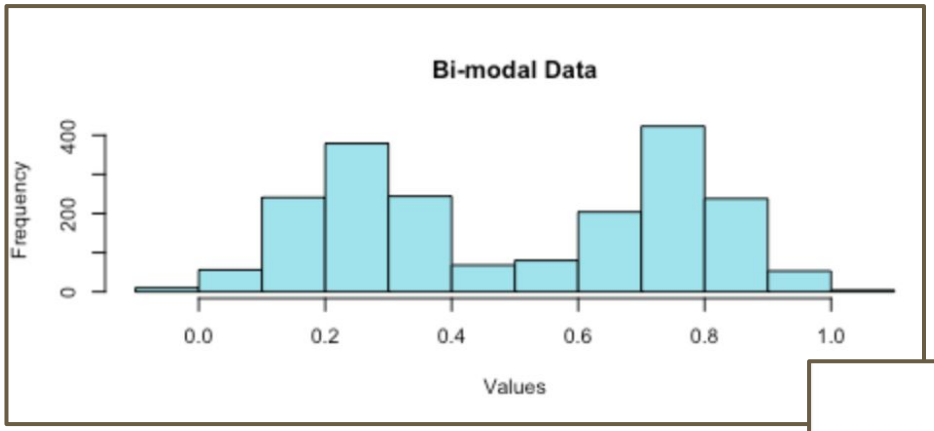


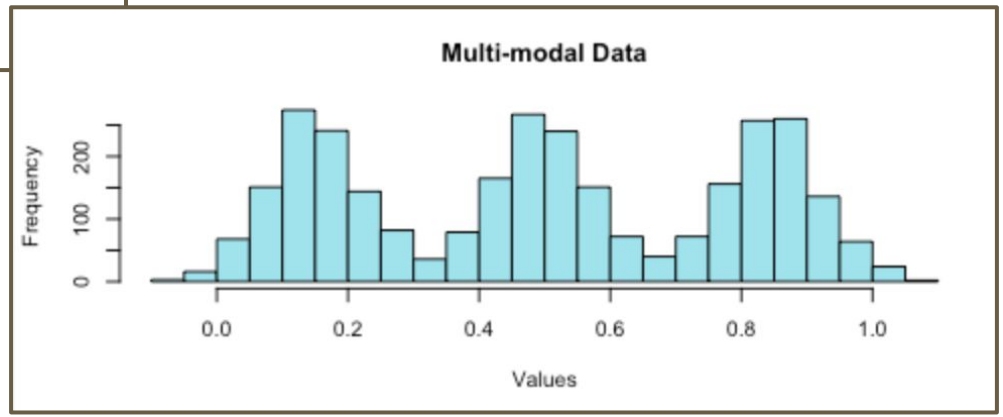


The top histogram shows data which is symetrically distributed. The middle plot shows a left-skewed histogram, this is also commonly referred to as negatively skewed. The final plot shows a right-skewed dataset, and this is also referred to as being positively skewed. All of these plots are uni-modal; in other words, there is a single concentration of values.

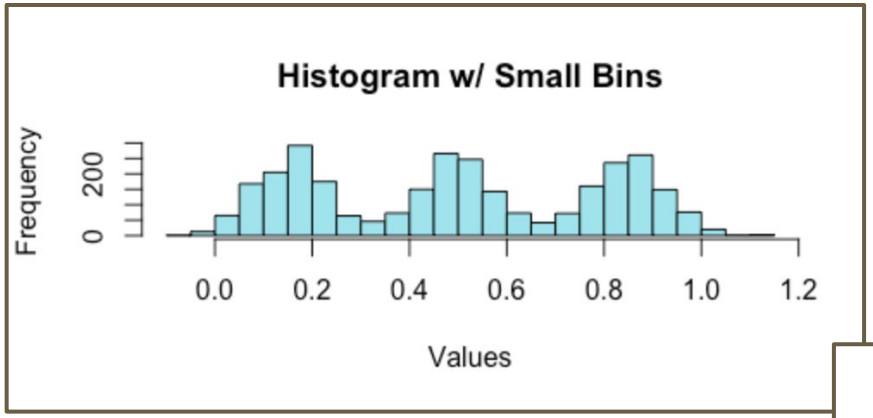
Histograms to Visualize Modality

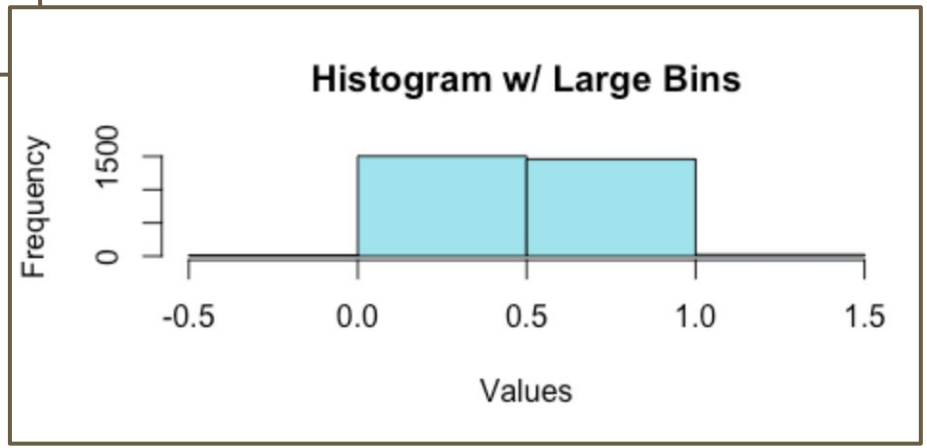






Histograms to Visualize Bin Sizes





In general, using wide bins will reduce noise created by randomness in sampling. When you have fewer data points, wider bins can be preferable. Conversely, if a data set contains a relatively large number of data points, smaller bins are more likely to give the greatest precision to a density estimation.

A common function to calculate the number of bins is as follows:

$$k = \sqrt{n}$$

Where k represents the number of bins, and n represents the number of samples in a data set.

Scatter Plots

- Scatter plots are a type of graph which utilize cartesian coordinates which typically display values for two variables for a dataset.

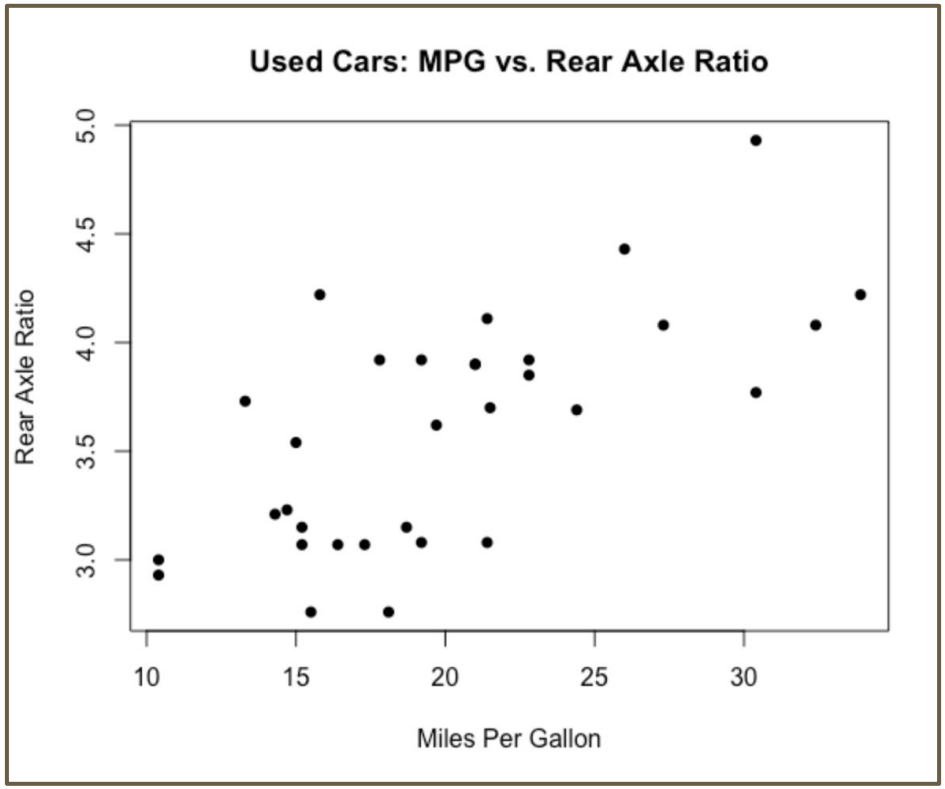
- Through the use of color, size, and different shapes, one, two, or three additional variables may be represented; the first portion of this lesson will focus on simple scatter plots which only display the two variables represented by the axes.

- In a scatter plot, each data point is represented by the intersection of the two variables with a dot (or a shape in more complex scatter plots).

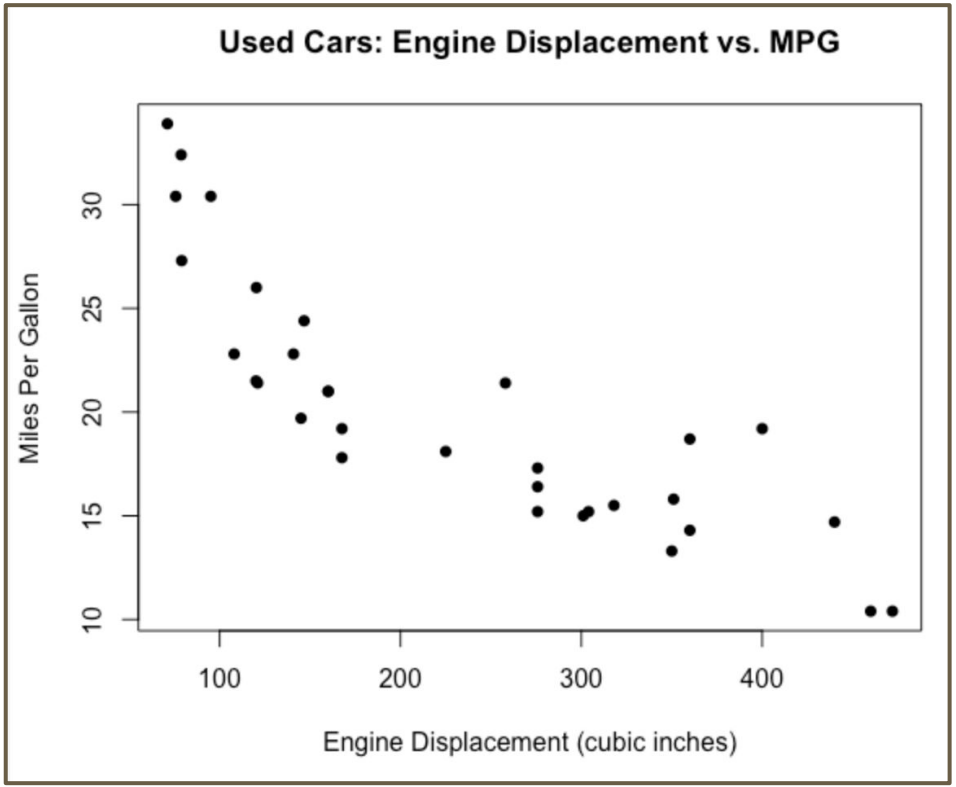
- The first variable is represented on the x-axis, while the second variable is represented on the y-axis. Scatter plots can be used to display data which is either continuous or discrete.

- It is a common that a scatter plot is used to find or display relationships between variables. If it is found that one of the variable displayed is dependent on the other, convention dictates that one assigns the independent variable to the x-axis, and the dependent variable to the y-axis.

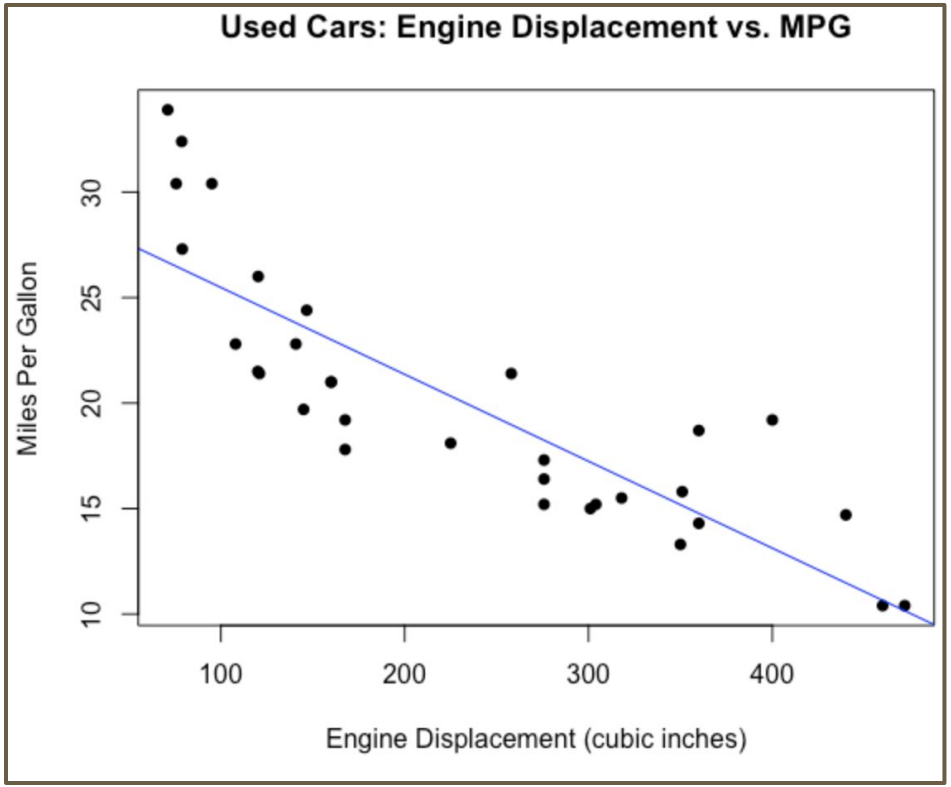
Scatter Plot of Small Correlation Between Two Variables



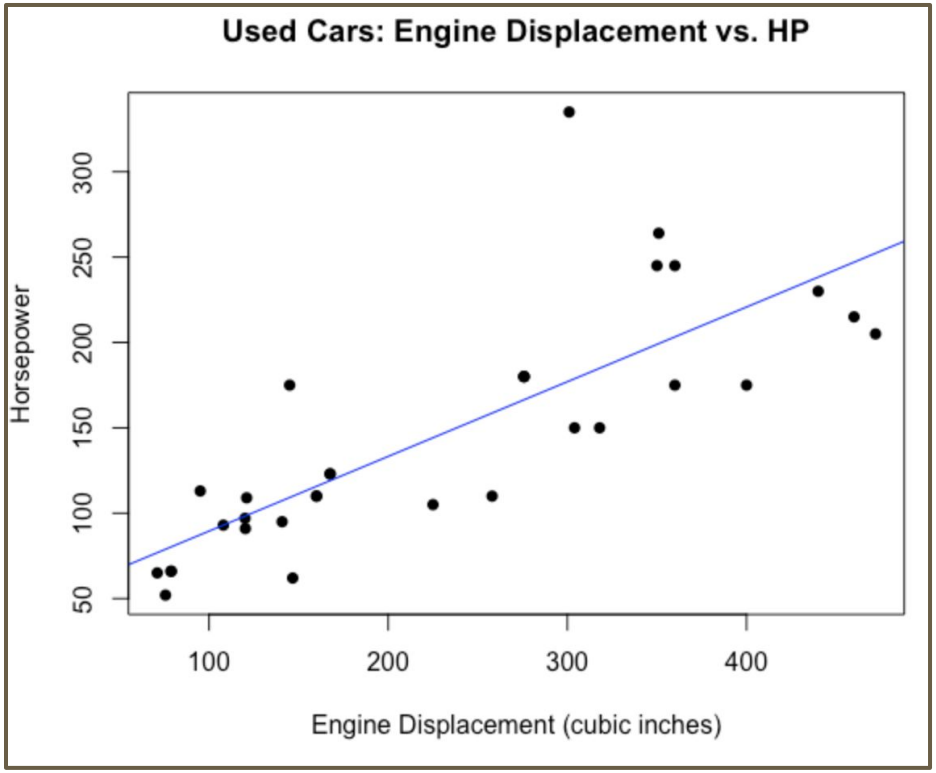
Scatter Plot of Highly Correlated Data



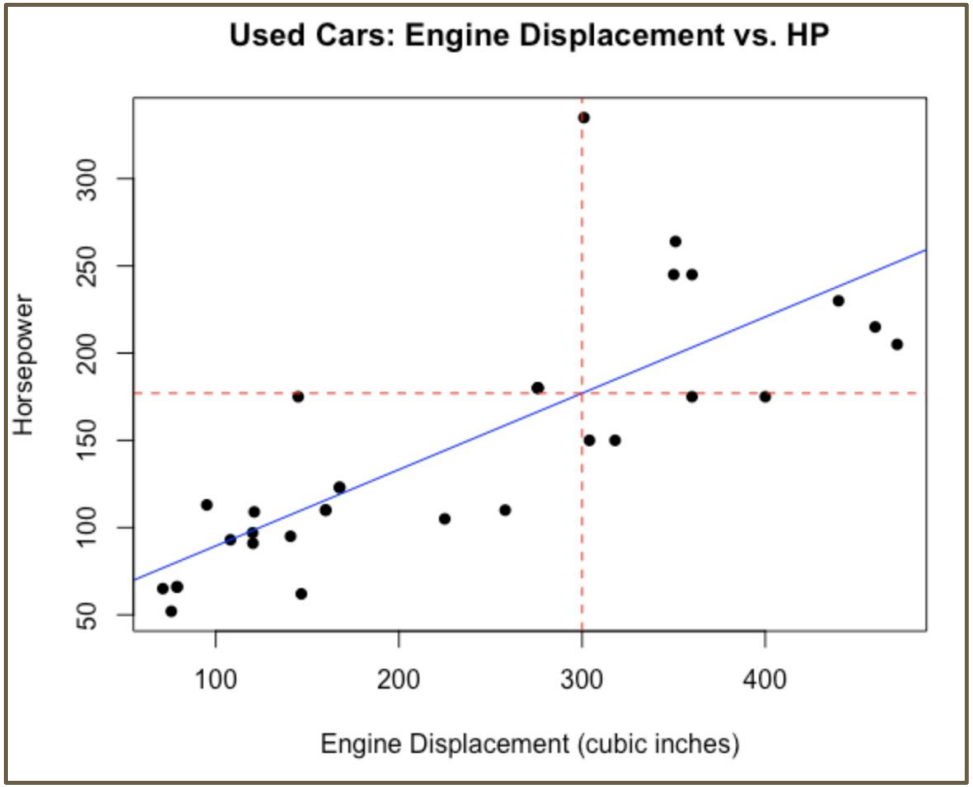
Negatively Correlated Data with a Fitted Regression Line



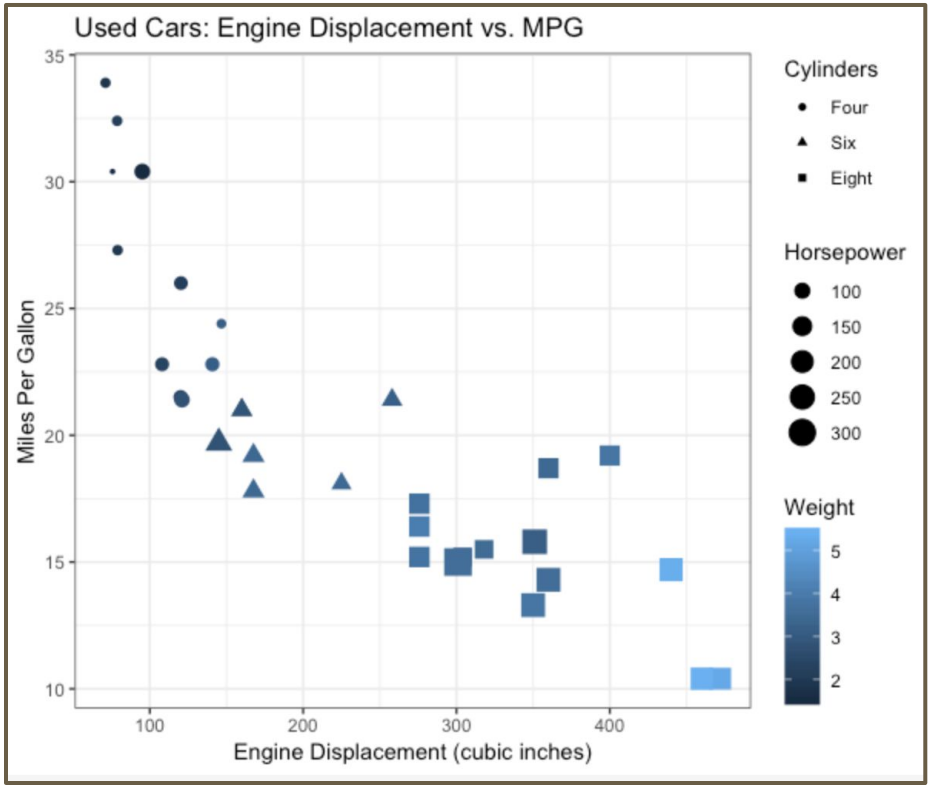
Positively Correlated Data with a Fitted Regression Line



Using a Regression Line to Make Predictions

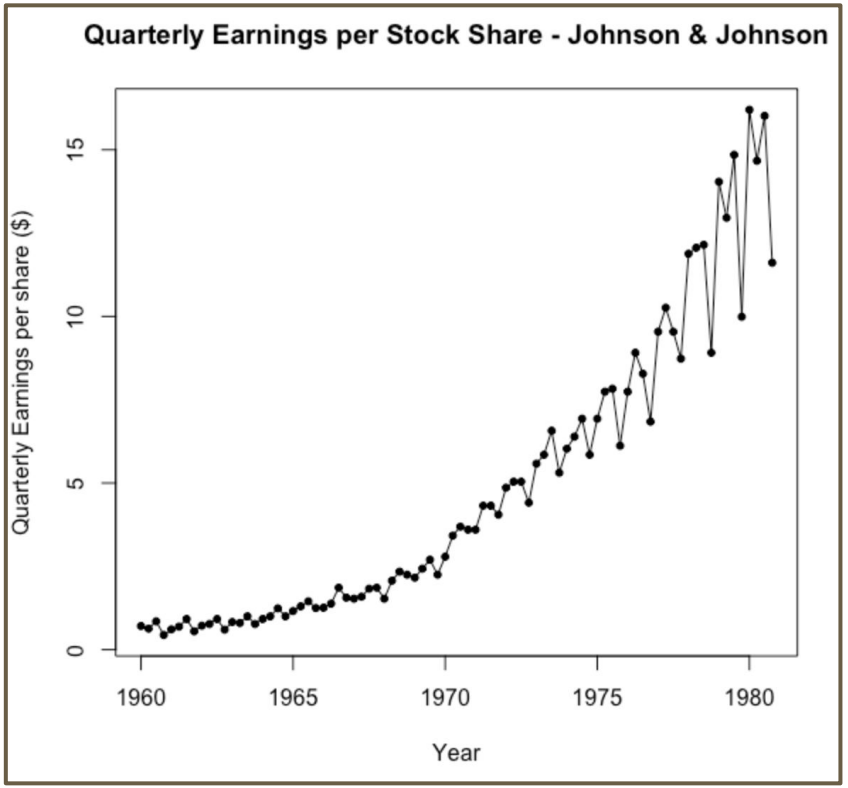


Displaying Multiple Variables on the same Scatter Plot

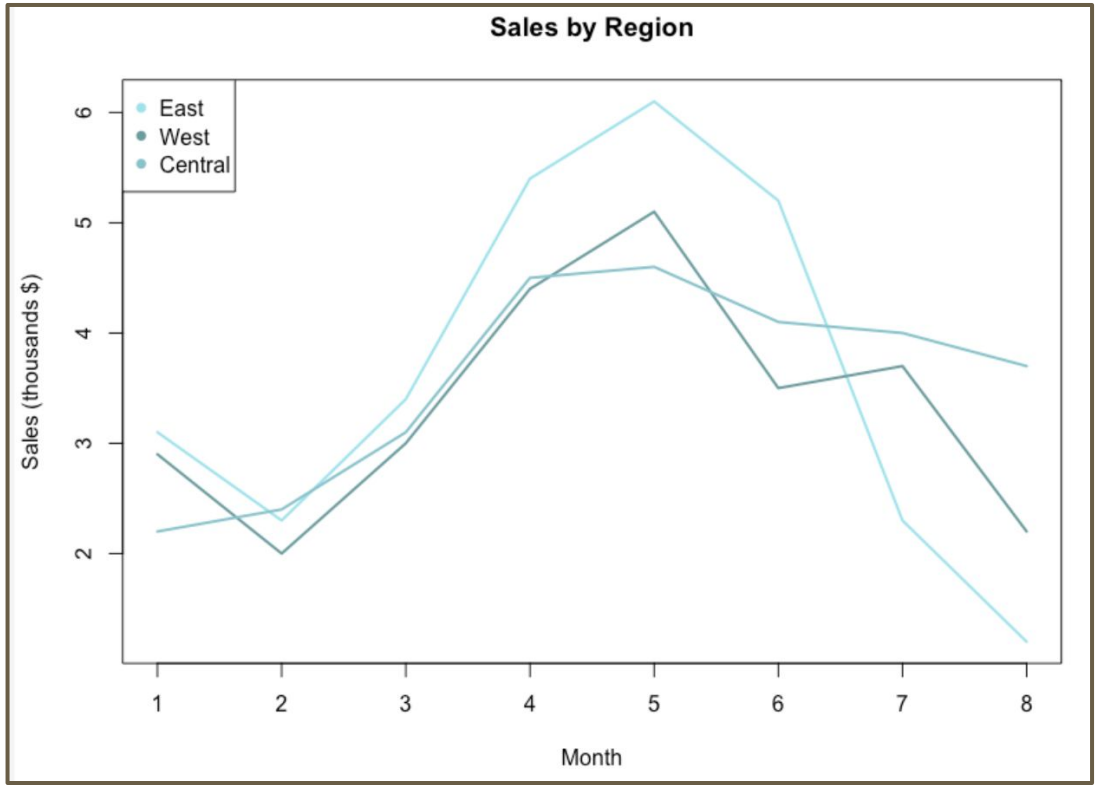


Line Graph of Time-series Data

Line Graph with Individual Data Points



Line Graph Representing Multiple Variables

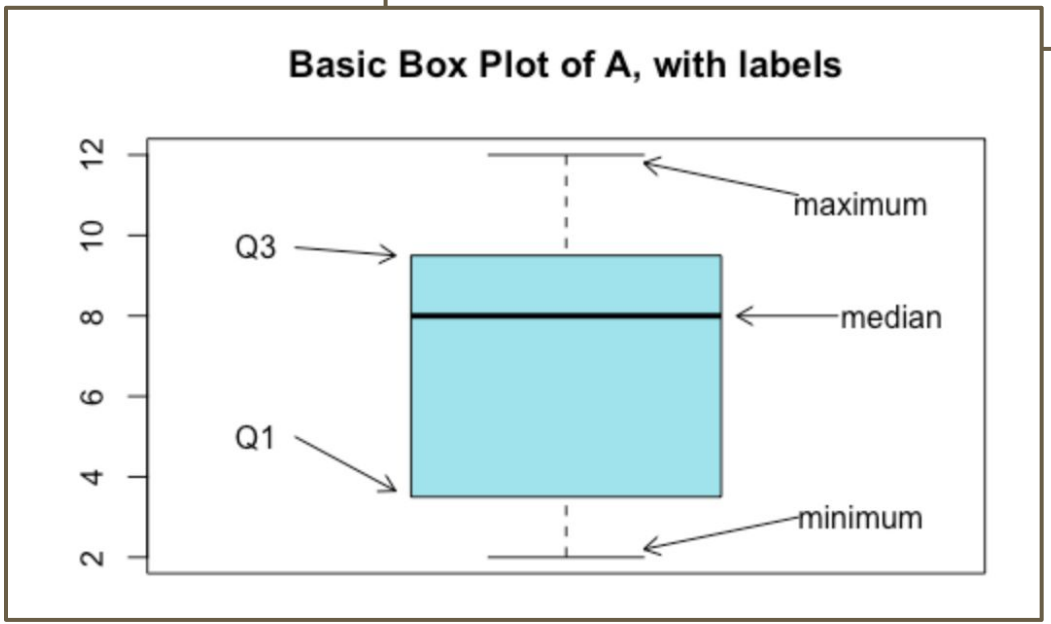


Box Plot

Example #1

Given the following dataset A, look at the box plot and the labels:

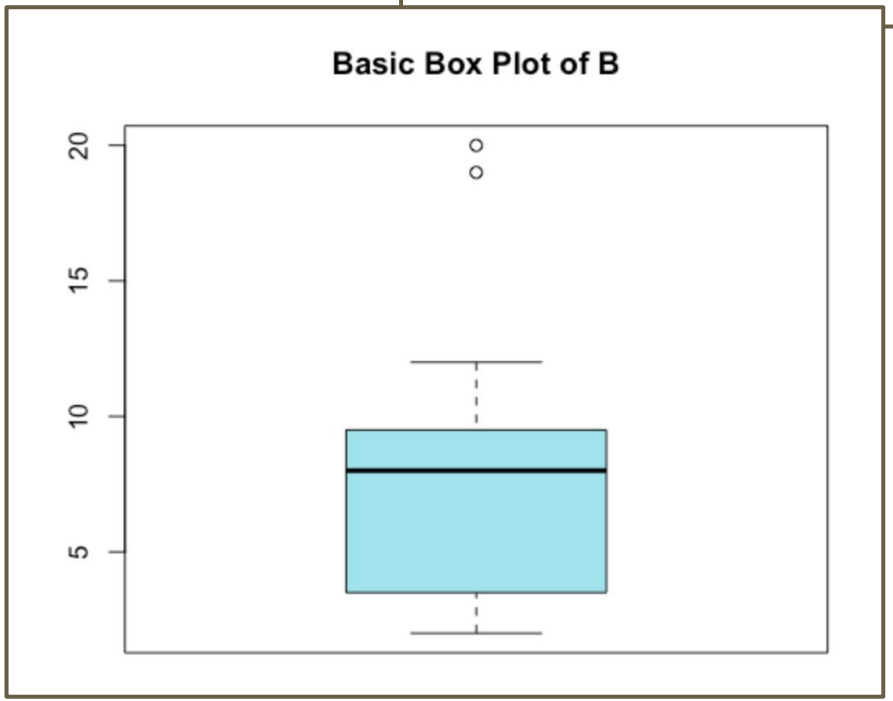
$$A = [2, 9, 3, 11, 9, 4, 8, 7, 3, 12, 5, 2, 8, 10, 11]$$



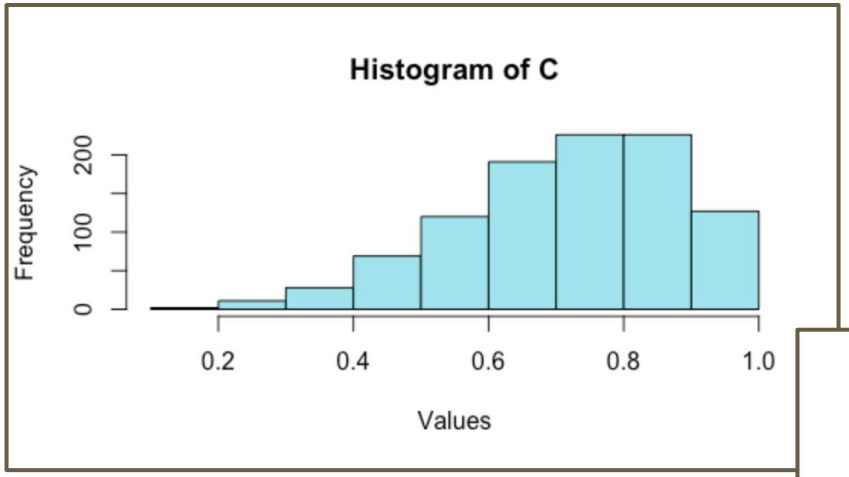
Example #2

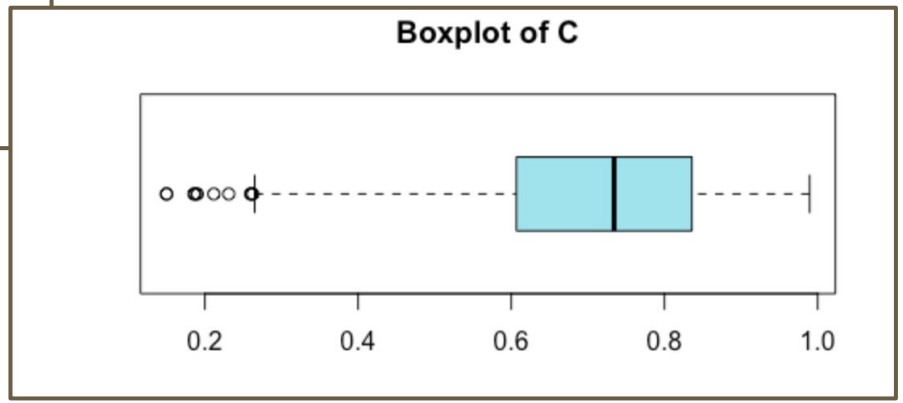
Given the following dataset B, examine the box-plot and the labels:

$$B = [2, 9, 3, 19, 9, 4, 8, 7, 3, 12, 5, 2, 8, 10, 20]$$

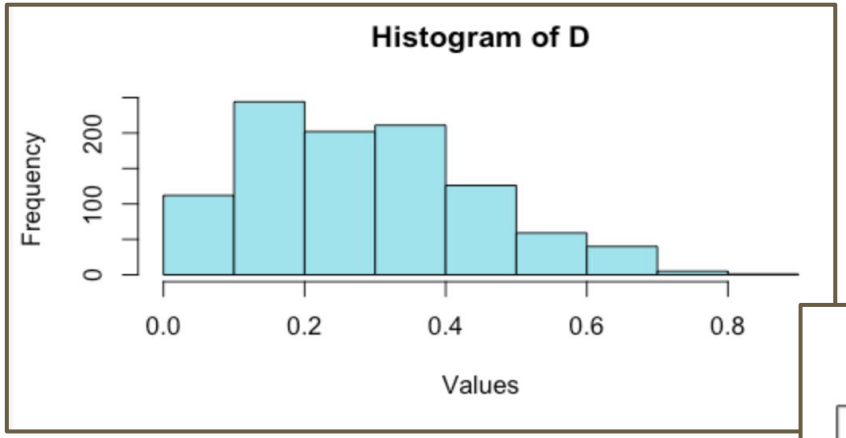


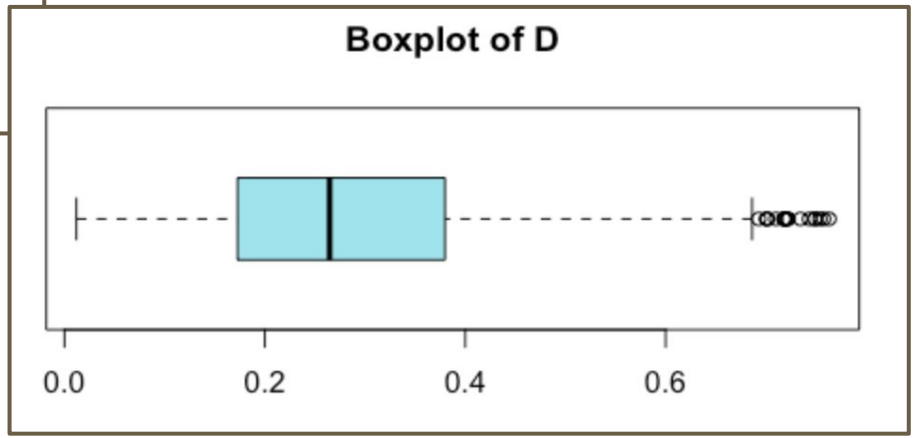
Box Plots to Identify -ve Skewness

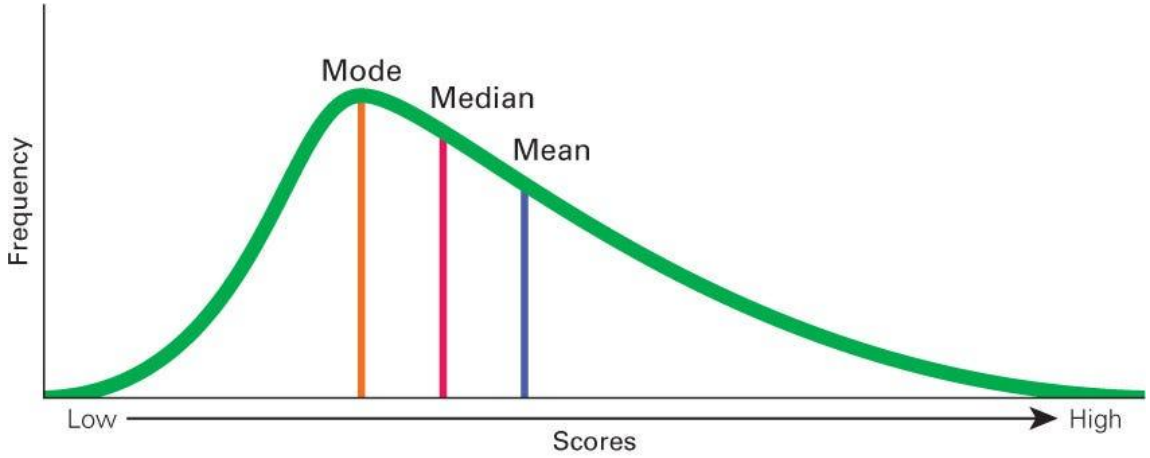




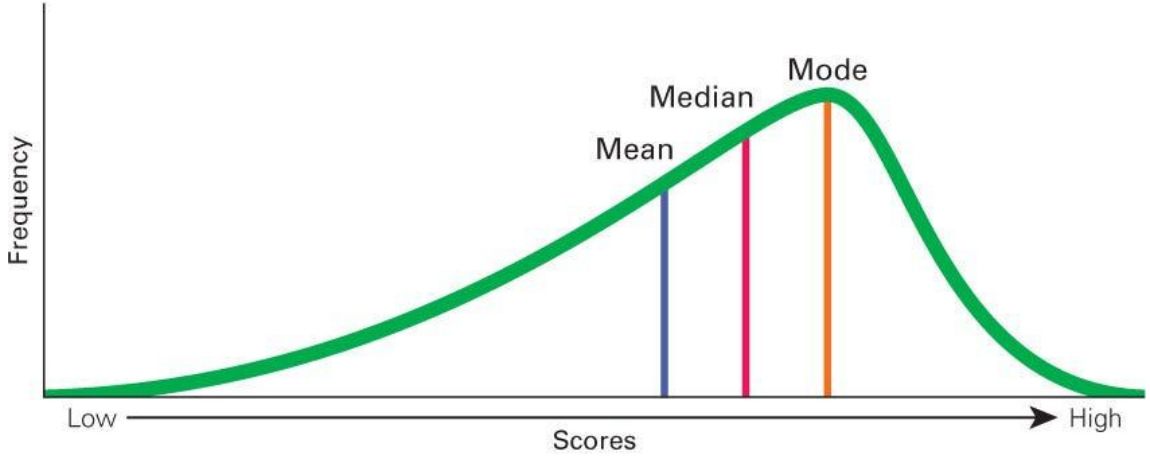
Box Plots to Identify +ve Skewness





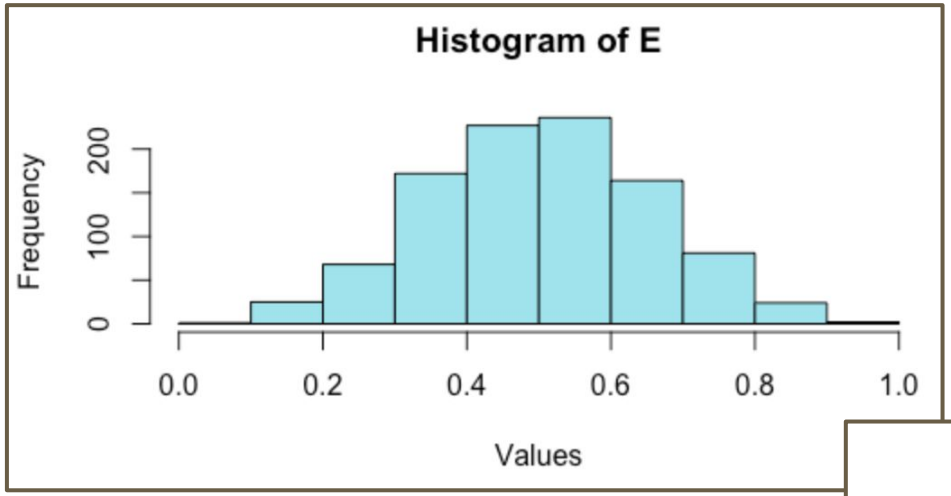


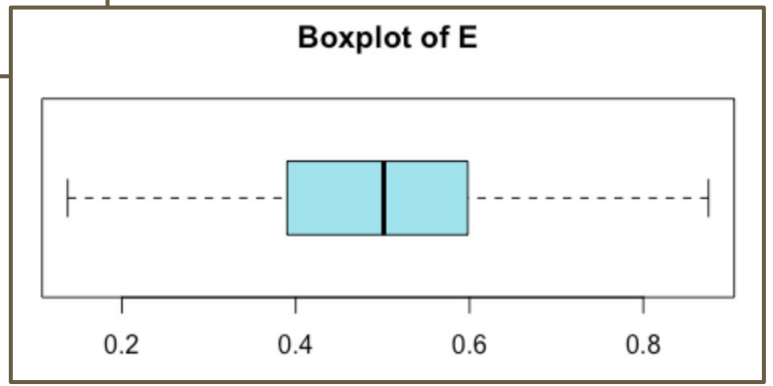
(a) Right-skewed distribution

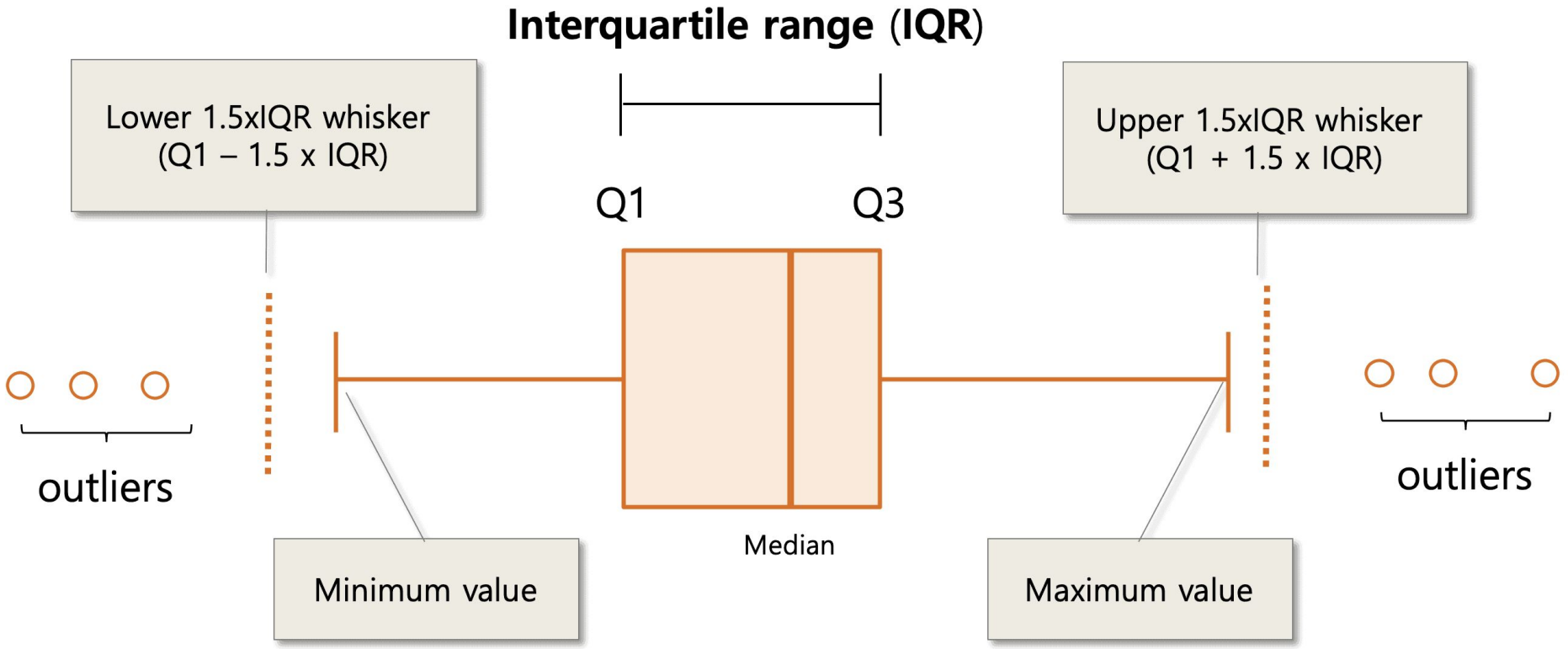


(b) Left-skewed distribution

Boxplots to Visualize Symmetry






Interquartile range (IQR)
Lower 1.5xIQR whisker
(Q1 – 1.5 x IQR)
Upper 1.5xIQR whisker
(Q1 + 1.5 x IQR)
Q1
Q3
outliers
outliers
Median
Minimum value
Maximum value

Sets Overview

Part 2

Contents

- Basic properties of a set
- The union of two events
- The intersection of two events
- The complement of an event
- The fundamental properties of set algebra
- DeMorgan's laws
- Subsets, Supersets & Inclusion rules of sets

Properties of a set

Probability Terminology

There are a few terms which need to be defined before diving directly into set theory.

In the study of the world, we often face "events" which may or may not occur; some examples could be:

- An inspector does or doesn't find a flaw in a manufactured good
- A flipped coin might or might not land with the head facing up
- A dice roll may or may not land showing three pips (pip is the name of the dot on a die)

In the world of probability we formally define these situations as **random experiments**.

Definitions

Event:

In probability theory, an event is an outcome or defined collection of outcomes of a random experiment. Typically, events are represented with a capital letter from the beginning of the alphabet, though there are some exceptions to this, such as the *sample space* as defined below.

Random Experiment:

A random experiment is an experiment or a process for which the outcome cannot be predicted with certainty.

Certain or Impossible:

If a specific outcome is guaranteed (for example, if both sides of a coin are heads, the outcome is guaranteed to be heads), then we refer to the outcome as **certain**. Lastly, if a specific outcome can never occur (as getting a tails on the aforementioned coin flip), the outcome is formally referred to as **impossible**.

Set:

A set in mathematics is simply a collection of well defined and distinct objects, and is also an object in its own right.

Sample Space:

If the outcome of a random experiment is unknown, but all of the possible outcomes are predictable in nature, then the set of all possible outcomes is known as the **sample space** and is denoted S. S represents every possible outcome of a random experiment; see an example below:

Example #1

For a random experiment which consists of the single flip of a fair coin, the sample space is:

$$S = \{H, T\}$$

That is, the set of all possible outcomes includes two elements:

1. The result of the coin flip is heads
2. The result of the coin flip is tails

Each one of these elements is referred to as a sample point in the sample space. In the terms of probability theory, certain subsets of S are referred to as **events**. That is, each event represents a subset of points of the sample space. See below for some example events of this experiment.

- Event A, where the coin lands on tails:
  - $A = \{T\}$
- Event B, where the coin lands on heads:
  - $B = \{H\}$
- Event C, where the coin lands on either heads or tails:
  - $C = \{H, T\}$

Example #2

For a random experiment which consists of:

1. The flip of a fair coin
2. The roll of a fair six-sided die

The sample space can be defined as:

$$S = \{H1,\ H2,\ H3,\ H4,\ H5,\ H6,\ T1,\ T2,\ T3,\ T4,\ T5,\ T6\}$$

Where H represents the coin landing on heads, T represents the coin landing on tails, and the integers $1,\ 2,\ 3,\ 4,\ 5,\ 6$ represent the number of pips (a pip is the name for each of the dots on a traditional die) shown on the die roll. Find the sets for the following events:

- An event A is defined as the coinflip resulting in heads
- An event B is defined as the die roll producing three pips
- An event C is defined as the coinflip resulting in heads, AND the die roll producing a value of three

Solution

$$A = \{H1,\ H2,\ H3,\ H4,\ H5,\ H6\}$$

$$B = \{H3,\ T3\}$$

$$C = \{H3\}$$

Note

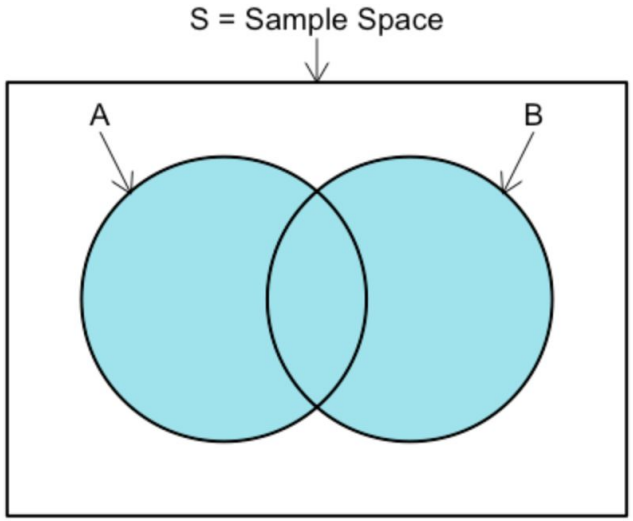
Notice that in each and every set mentioned above, there are not any duplicates. This is one of the most important properties of a set, that each object is ***distinct***, that is, no object can be represented more than one time in a set.

Set Union

Union of Two Events

In set theory, the union of two events (A, B) is defined as a set which contains all of the sample points which are in A, and also all of the sample points which are in B. The union can be interpreted as the statement "The union of events A and B contains all sample points which are in A or B." See below for a venn diagram which shows the union of two events, A and B.

The common notation for a union is $\cup$, so the union of events A and B can be expressed as $A \cup B$.



A union B

Example #1

In a random experiment where a fair six-sided die is rolled, event A is described as an outcome in which the number of pips showing is odd. Event B is defined as any outcome in which there are four or fewer pips showing. Consider the two events and find the union $(A \cup B)$.

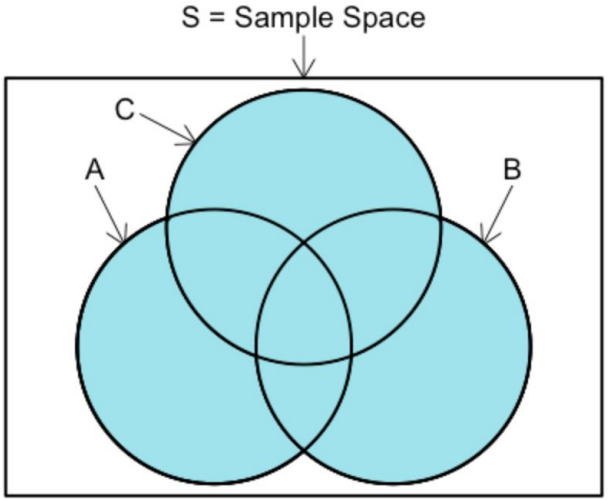
- Step 1: Clearly define the two events
  - $A = \{1, 3, 5\}$
  - $B = \{1, 2, 3, 4\}$
- Step 2: Find the union of the two events
  - $A \cup B = \{1, 2, 3, 4, 5\}$

Solution

$$A \cup B = \{1, 2, 3, 4, 5\}$$

Union of More than Two Events

The union of events is not limited to only two events, in fact, given an infinite sample space, there can be an infinite number of unions. Below, see a venn diagram which shows the union of three events, A, B, and C.



A union B union C

Example #2

In a random experiment where three fair, six-sided dice are rolled, event A is described as an outcome in which all three dice show three pips. Event B is described as an outcome in which all three dice show four pips, and event C is described as an outcome in which all three dice show five pips.

- Step 1: Define the subset of outcomes for each event
  - $A = \{333\}$
  - $B = \{444\}$
  - $C = \{555\}$
- Step 2: Define the union of the three events
  - $A \cup B \cup C = \{333, 444, 555\}$

Solution

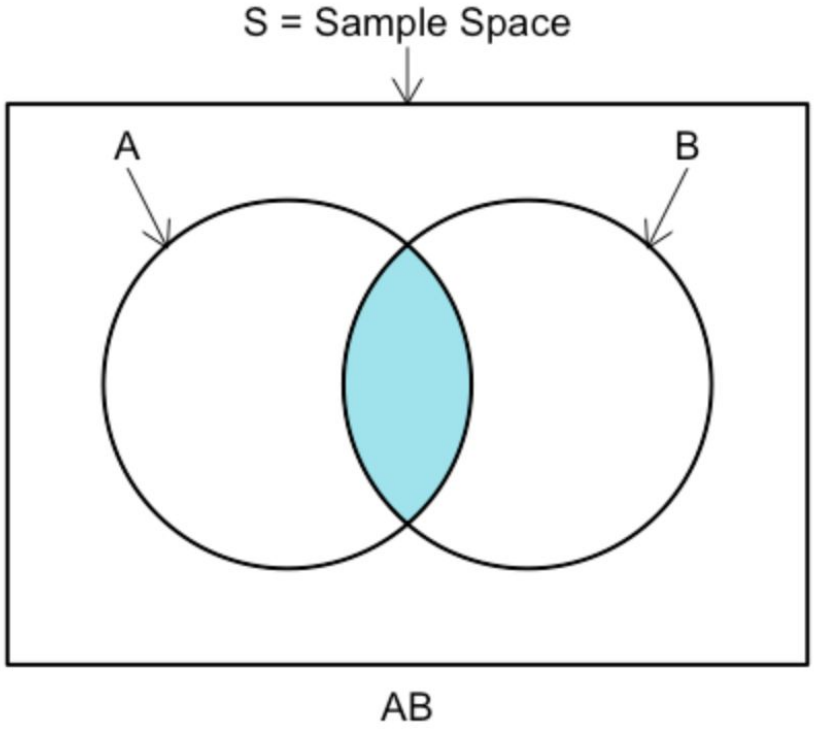
$$A \cup B \cup C = \{333, 444, 555\}$$

Intersection

In set theory, the intersection of two events, A and B, is defined as a set which contains all of the elements of A, which are also in B. The intersection of events A and B can also be defined as the set which contains all of the elements of B, which are also in A. There are two common notations for the intersection of two events shown below.

$$AB \text{ or } A \cap B$$

Venn Diagrams are commonly used to visualize set theory concepts. See below for a visual representation of an intersection between events A and B.


S = Sample Space
A
B
AB

Example #1

Find the intersection of the two events, A and B:

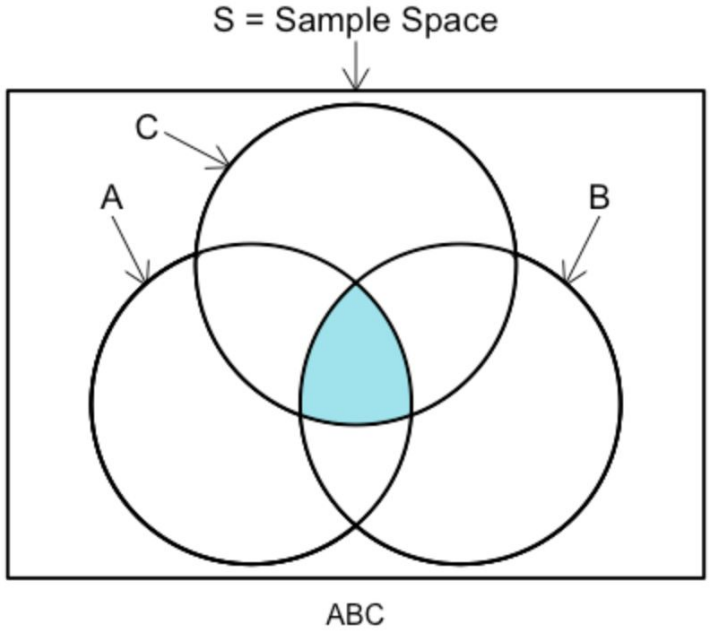
$$A = \{1, 2, 3, 4, 5, 6\} \text{ and } B = \{2, 3, 4\}$$

- Step 1: Define the two events
  - $A = \{1, 2, 3, 4, 5, 6\}$
  - $B = \{2, 3, 4\}$
- Step 2: Find the intersection (all values in ***both*** sets)
  - $A \cap B = AB = \{2, 3, 4\}$

Solution

$$A \cap B = AB = \{2, 3, 4\}$$

In the example above, the members which land in both A and B are considered the intersection of events A and B. Intersections are not limited to only two events, in fact, given an infinite sample space with an infinite number of events (each event representing a set of sample points), there can be infinite intersections. Below, we can see a venn diagram and an example of a three-way intersection.



Example #2

Given a random experiment which is comprised of a single six-sided die roll, the sample space S is represented by the integers $\{1,\ 2,\ 3,\ 4,\ 5,\ 6\}$. Event A represents an outcome where the die roll yields three or fewer pips showing; event B represents an outcome where in there are three or more pips showing; finally, event C represents the outcome in which there are an odd number of pips showing. Find the intersection of $A,\ B,\ \text{and}\ C$.

- Step 1: Define each event as a set
  - $A = \{1, 2, 3\}$
  - $B = \{3, 4, 5, 6\}$
  - $C = \{1, 3, 5\}$

- Step 2: Find the intersection
  - $A \cap B \cap C = ABC = \{3\}$

Because the sample point $\{3\}$ is the only number which occurs in all three events, it represents the entirety of the intersection between the three events.

Solution

$$A \cap B \cap C = ABC = \{3\}$$

Absorption

The theorem of absorption states that:

$$A \cup (A \cap B) = A$$

That is, a subset which contains an intersection, and is in union with one of its supersets the intersection is "absorbed".

Try drawing a venn diagram to convince yourself that this theorem is true.

Complement of a Set

Definition: Complement of a Set

In set theory, the complement of a set, A, refers to all elements not in A. When all sets under consideration are considered to be subsets of a given sample set S, the absolute complement of A is the set of elements in S but not in A.

Example #1

Consider the random experiment of rolling a single fair six-sided die. Let A be an event which is defined as all even values in the samples space, find the complement of A.

- Step 1: Define the sample space and the event
  - $S = \{1,\ 2,\ 3,\ 4,\ 5,\ 6\}$
  - $A = \{2, 4, 6\}$

Solution

$$A^c = \{1, 3, 5\}$$

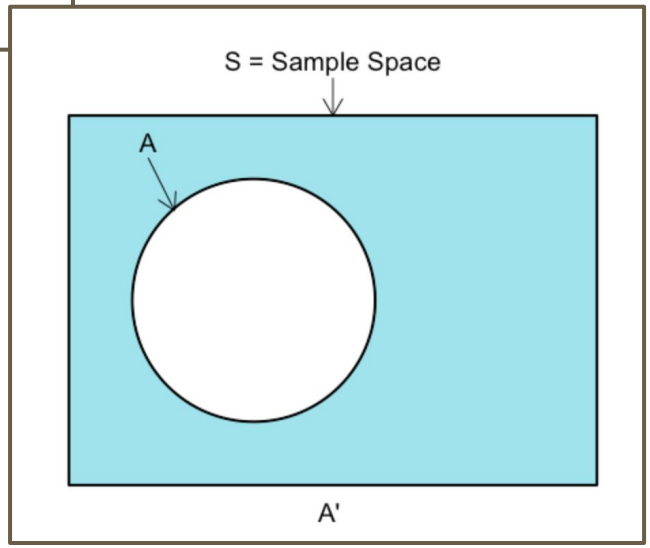
- Step 2: Find the complement of A
  - $A^c = \{1, 3, 5\}$

Notation

There are several common notations used for the complement, they are listed below:

$$complement(A) = A^c = A^0 = \overline{A} = A'$$

Throughout the materials in this statistics block, A^c will be the common notation. Much like many other concepts in set theory, venn diagrams are commonly used to visualize the complement of a set. See the venn diagram below for a simple example of a complement.



Example #2

Consider a random experiment which is described as a two-coin flip, where two unique and separate coins are flipped. Let the event B represent the event that the first coin lands on heads, and find the complement of B.

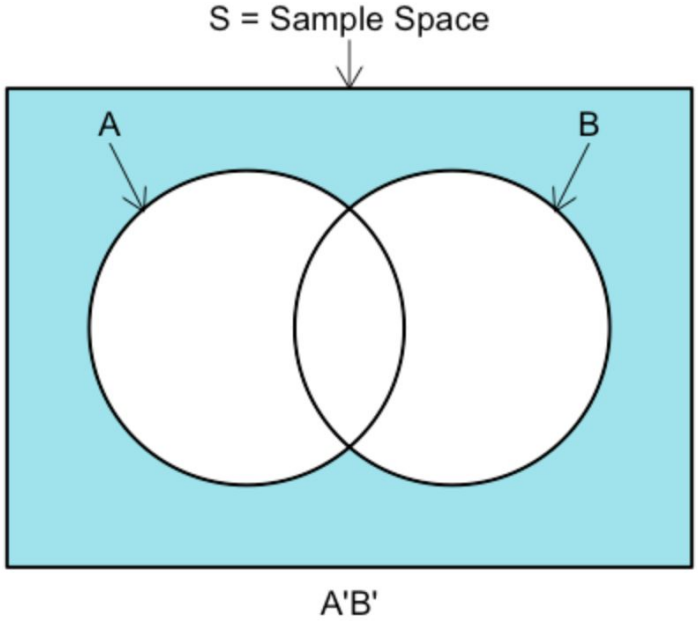
- Step 1: Define the sample space and the event B
  - $S = \{HH, HT, TH, TT\}$
  - $B = \text{ Event when the first coin is heads } = \{HH, HT\}$
- Step 2: Define B^c
  - $B^c = complement(A) = \{TH, TT\}$

Solution

$$B^c = complement(B) = \{TH, TT\}$$

In the above example, there is the sample space S and the event B. It is shown that the B^c is comprised of every element in the sample space which is not in B. This could also be interpreted as the set of all sample points in which the first coin flip does not result in heads.

Much like the other operators which have been introduced in set theory previously, the complement can be applied to multiple events. See below for a venn diagram which describes the intersection of an event A's complement, intersected with the event B's complement, that is A^cB^c:



Example #3

Consider the random experiment defined by the rolling of a fair six-sided die. An event, A, is defined as the result showing three pips; another event, B, is defined as the result of the die roll showing four pips. Using this information, find A^cB^c.

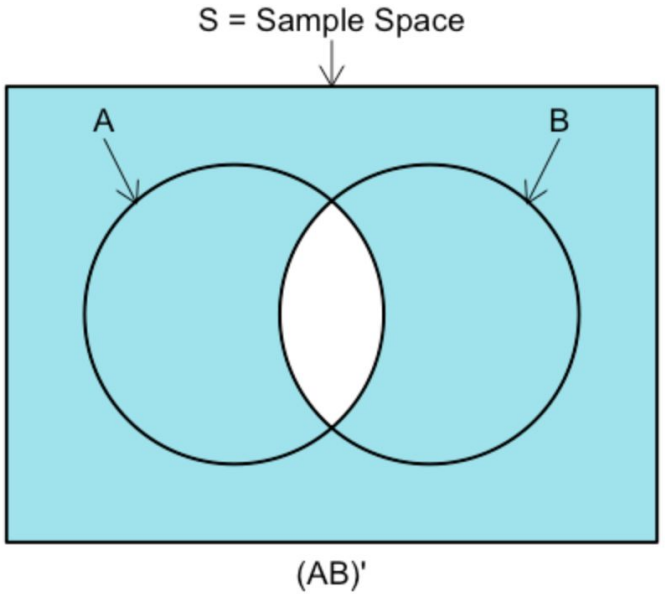
- Step 1: Define the sample space and the events
  - $S = \{1,\ 2,\ 3,\ 4,\ 5,\ 6\}$
  - $A = \{3\}$
  - $B = \{4\}$
- Step #2: Find A^c and B^c
  - $A^c = \{1,\ 2,\ 4,\ 5,\ 6\}$
  - $B^c = \{1,\ 2,\ 3,\ 5,\ 6\}$
- Step #3: Find A^cB^c
  - $A^cB^c = \{1, 2, 5, 6\}$

Solution

$$A^cB^c = \{1, 2, 5, 6\}$$

As the venn diagram in ***Figure 2*** shows A^cB^c is comprised of all sample points which are not in A and also not in B

When mixing set operators, one has to be careful and precise. Below there is an example which is very similar to the previous problem, but there is a drastically different outcome, $(AB)^c$.



Example #4

Consider again the random experiment of a two-coin flip, the event A is defined as the first coin resulting in heads; the event B is defined as the second coin resulting in heads. Find the complement of A intersect B.

- Step 1: Define the sample space and the events
  - $S = \{HH,\ HT,\ TH,\ TT\}$
  - $A = \{HH,\ HT\}$
  - $B = \{HH,\ TH\}$
- Step 2: Find AB
  - $AB = \{HH\}$
- Step 3: Find $(AB)^c$
  - $(AB)^c = \{HT,\ TH,\ TT\}$

Solution

$$(AB)^c = \{HT,\ TH,\ TT\}$$

The complement of the intersection AB, is the set of all sample points in S which are not in the intersection of events A and B.

The Difference of Sets

Another common way to express a complement is by referencing the difference between sets. See an example below:

$$A - B = AB^c$$

That is, $A - B$ can be interpreted as the set of sample points which are in A, but are not in B. This technique can also be extrapolated into multiple events as shown below:

That is, $A - BC$ can be interpreted as the set of sample points which are in A, but are not in the intersection of BC.

Throughout this introductory statistics block, we will typically avoid using the difference notation, however, it is important to know of its existence.

Laws of Set Algebra

Commutative

Just like addition and multiplication, both union (∪) and intersection (∩) are commutative. That is to say that:

- $A \cup B = B \cup A$
- $AB = BA$

Associative

Again, like addition and multiplication, both union (∪) and intersection (∩) are associative. That is to say that:

- $(A \cup B) \cup C = A \cup (B \cup C) = A \cup B \cup C$
- $(AB)C = A(BC) = ABC$

Distributive

Lastly, like addition and multiplication, both union (∪) and intersection (∩) are distributive. That is to say that:

- $A \cup (BC) = (A \cup B)(A \cup C)$
- $A(B \cup C) = (AB) \cup (AC)$

In addition to these fundamental laws, there are other common laws and properties of sets, unions, and intersections

Other Common Laws, Principles and Properties of Sets

Idempotent Laws

There are two idempotent laws of sets which might seem self-explanatory, but they should be explicitly stated.

- $A \cup A = A$
- $A \cap A = A$

Absorption Laws

There are two absorption laws, as listed below:

- $A \cup (AB) = A$
- $A(A \cup B) = A$

Domination Laws

The domination laws refer to both the universal set, as well as the null set. The universal set, is a set which contains all objects or elements and of which all other sets are subsets. A universal set that has already been referenced in this course is the sample space of a random experiment. The null set is what you may imagine it to be, and is also often called the empty set; this is a set which does not contain anything.

Notation:

- $U =$ Universal Set
- $\emptyset =$ Null Set

The two domination laws are shown below:

- $A \cup U = U$
- $A \cap \emptyset = \emptyset$

Identity Property

A set has the indentity property under a particular operation if there is an element of that set that leaves every other element of the set unchanged under a given operation. See a generalized example below:

- $A \cup \emptyset = \emptyset \cup A = A$

Complement Laws for the Universal and Null Sets

The complement of either the null set or the universal set can also be found, but it can seem slightly less intuitive than a typical complement. See below for the explicit definitions of each:

- $\emptyset^c = U$
- $U^c = \emptyset$

Involution (or double-complement) Law

The involution law covers situations where there is a double-complement. See below for an example:

- $(A^c)^c = A$

DeMorgan's Laws

- DeMorgan's First Law: $(A \cup B)^c = A^c B^c$
- DeMorgan's Second Law: $(AB)^c = A^c \cup B^c$

Subsets and Inclusion

Subset

In set theory, a set A is a ***subset*** of a set B, denoted $A \subseteq B$, if and only if every member of A is also contained in B. That is, all elements of A are also elements of B.

Superset

Equivalently to the set A being a subset of the set B above, B is a ***superset*** of A. That is B contains all members of A.

Proper Subset

If the set A is a subset of B, and B contains and at least one sample point which is not also contained in A, then A is a ***proper subset*** of B, denoted $A \subset B$.

Additional Properties of Inclusion

- A set A is a subset of B if and only if their intersection is equal to A
  - That is, $A \subseteq B \Leftrightarrow A \cap B = A$
- A set A is a subset of B if and only if their union is equal to B
  - That is, $A \subseteq B \Leftrightarrow A \cup B = B$

Equality

The equality of two events can be determined through the examination of inclusion. If an event A is a subset of event B and B is also a subset of event A; it can be determined that the events A and B are equal. See below for an example.

In the random experiment above, we have the sample space, $S = \{HH, HT, T1, T2, T3, T4, T5, T6\}$, the event E is defined as the event where there is not a die roll, $E = \{HH, HT\}$,the event F is defined as the event where there are two coin flips, $F = \{HH, HT\}$.

By the definition of a subset, E is a subset of F, and F is also a subset of E.

Solution

$E \subseteq F$ and $F \subseteq E$, therefore it can be said that $E = F$

Thank You