# Statistics Reference Series Part 1: Descriptive Statistics

Authored by: Gina Sigler

20 July 2018

Revised 5 October 2018



The goal of the STAT COE is to assist in developing rigorous, defensible test strategies to more effectively quantify and characterize system performance and provide information that reduces risk. This and other COE products are available at <u>www.afit.edu/STAT</u>.

# **Table of Contents**

Executive Summary	
Introduction2	
Types of Data2	
Graphs2	
Histogram3	5
Boxplot4	ł
Scatterplot5	)
Pie chart6	;
Bar chart6	)
Time Series Displays7	,
Summary Statistics	;
Mean	;
Median9	)
Standard Deviation/Variance9	)
Outliers/Influence Points	1
Conclusion9	1
References10	)
Appendix A: Data & Formulas11	•
Data11	
Formulas12	2

*Revision 1, 5 Oct 2018, Formatting and minor typographical/grammatical edits.* 

# **Executive Summary**

This document is designed as a refresher/summary for those seeking information or a simple example involving descriptive statistics. Descriptive statistics aid in summarizing information through either via graphical methods or numerical values. These summaries are used as stepping stones for further data analysis or inferential statistics. The graphical and numerical summaries explained are not exhaustive list, but do provide some commonly used (and occasionally misinterpreted) summary results.

Keywords: Descriptive statistics, graphical summary, numerical summary, analysis.

# Introduction

Statistics revolves around the world of data. It is used to covert data into some (hopefully) useful information. Statistics can help design studies, collect data, as well as analyze and model the data collected. Collecting all the possible desired data, referred to as the population, is impossible in almost all instances. Instead, it is possible to collect a subset of that data, which is referred to as a sample. Summarizing or displaying this sample data is called descriptive statistics.

All data used in the examples throughout this document can be found in Appendix A.

# **Types of Data**

What we can do with the sample data is largely dependent on the type of data collected. There are two overarching classifications of data known as numerical (also sometimes called quantitative) and categorical (also sometimes called qualitative).

Numerical data is simply data that consists of numbers. It can be further divided into two subclassifications: continuous and discrete. Continuous data can be any value in a range. Think of measurements where the accuracy keeps increasing such as time (hours, minutes, seconds, milliseconds, etc.) Discrete data can only have specific values. These are numbers that cannot be more accurately reported. Examples of discrete data would be the amount of physical money in someone's pocket, which can only be broken down as far as the nearest penny, or the number of puppies in a pen.

Categorical data is data that can be placed into distinct categories. It too can be divided into two subclassifications: nominal and ordinal. Nominal data has no natural order, while ordinal data has a natural order. Some examples of nominal data are colors, vehicle make, and gender. Ordinal data is usually in the form of a Likert scale such as rating satisfaction in terms of Very Satisfied, Satisfied, Unsatisfied, etc.

## Graphs

One of the best ways to explore or present data is to create a visual summary or graphic. There are far more types of graphs than those presented in this document, but this section covers most of the graphs commonly encountered in practice. Please note that all data used has been realistically fabricated.

#### Histogram

One of the most common graphs seen today is the histogram. Histograms are created using numerical data which is grouped into intervals on the x-axis. One set of data can have many different possible histograms depending how wide the intervals are; the shape of the histogram will be either shrunk or stretched depending on the chosen interval widths. There are three different kinds of histograms that simply have a different scale for the y-axis: frequency, relative frequency, and density histograms. Figure 1 is a frequency histogram, where counts are shown on the y-axis. Please note that the x and y axes can be switched, but it is conventional to have the data on the x-axis. A relative frequency histogram changes the y-axis to proportions (counts divided by the total number of data points). A density histogram changes the y-axis so that the total area under all of the bars becomes one. The density is calculated by taking the counts divided by the number of data points and again divided by the width the bar intervals. For a given set of data and bin sizes, the shapes of all three histograms will be the same – it's only the values on the y-axis that differ.





Figure 1: Frequency histogram featuring the number of miles per gallon for a set of 30 cars.

### **Boxplot**

A boxplot is another graphical tool for numerical data. A boxplot focuses on the symmetry of the data by plotting a five number summary of the data: the median, minimum, maximum, and inner quartiles. As the median divides the data into halves, the quartiles divide the data into four equal parts or quarters. A boxplot puts a box around the middle 50% of the data and lines out to the ends. This box portion of the box plot is referred to as the interquartile range (IQR). Sometimes, a data point will appear as a dot at the end of the lines, indicating that it is unusually far away from the middle, this is called an outlier. Sometimes the mean may also be displayed with a distinguishable symbol. Figure 2 is an example of multiple boxplots shown on the same graph for the car data. This time, the boxplots break out the individual makes of the cars. For example, it can be readily seen that over 75% of Toyota cars have better gas mileage than the median Ford.



Miles per Gallon by Make

Figure 2: Boxplots of the miles per gallon for the data set of 30 cars broken down by the make of the car.

# **Scatterplot**

A scatterplot is a method of displaying the relationship between two corresponding sets of numerical data. Each pair of data points (a pair consists of one data point from each of the two data sets) is plotted as a single point (x,y). A line may also be placed on the graph showing the overall trend of the data. The scatterplot in Figure 3 shows the relationship between the cars' weights and their gas mileage. The blue line shows that weight and mileage appear to follow a trend which could be represented by a line of negative slope (i.e., lighter cars tend to have better gas mileage).



Miles per Gallon Versus Weight in Pounds

Figure 3: Scatterplot of miles per gallon versus weight in pounds with different colors assigned for each make of car and a line of best fit added.

## **Pie chart**

There are two main types of graph for looking at categorical data. The first is named a pie chart; this is due to the fact that a circle is divided into slices that represent each of the categories. Figure 4 shows a pie chart of M&M colors taken in a random sample. Unfortunately, it is difficult for a reader to precisely eyeball the difference in angles between different pie pieces. Therefore, while pie charts are popular, they should be avoided whenever possible and replaced with a bar chart. The next section will illustrate the difference.



#### Figure 4: Pie chart detailing the number of each color of M&M in a random sample from a bag.

#### **Bar chart**

A bar chart is another graphical option for categorical data. It has certain similarities to the histogram in the overall look with data type on one axis and counts on the other. It is a bit different in that the bars represent completely separate categories and will not touch. There are also many possible shapes to a bar chart since the order of the categories does not matter. A bar chart is usually easier to read than a pie chart because a direct comparison of the bars can be made. Figure 5 shows the same M&M data displayed in a bar chart.



Figure 5: Bar chart showing the number of each color of M&M in a random sample from a bag.

# **Time Series Displays**

The final graph is a bit more complicated in nature than the others. A time series display is specifically used for data that has a time association. Figure 6 shows an example of the amount of money in an investment account over some months. Time is shown along the x-axis and the associated information is shown on the y-axis similar to a scatter plot. The points are often connected with line segments to show that data points are not independent from one another. This dependence is what makes the analysis of time-varying data more complicated: the amount of money in the account today is much more dependent on the amount there yesterday than how much was there 2 years ago. Many statistical techniques assume independence of all runs, an assumption which is clearly violated here.



Figure 6: Time series detailing the amount of money in an account in dollars from January to April 2018.

# **Summary Statistics**

Another great way to talk about data in aggregate is with some numerical measure. This, of course, can only be done with numerical data. There are many different kinds of numerical summaries which mostly deal with centers and spread of the sample data. There are additional summary statistics not presented in this document, but this section reviews some of the most popular summary statistics.

#### Mean

The mean, sometimes referred to as the average, is the sum of a data set divided by the number of data points. It can be thought of as the tipping point of a histogram where exactly half of the weight would be

on either side. One of the issues with a mean is that the value may not actually represent a data point within the data set. If there was a histogram with two very separate peaks, the mean would be a value belonging to neither of the distinct sets. Another issue is that the mean can be heavily influenced by outliers (see explanation below). The mean is the most commonly used measure of center, but it should never be given as the only piece of summary information for a data set. For the sample of 30 cars, the mean was 23.54 miles per gallon. A formula can be found in Appendix A.

## Median

The median is the middle value of the data when it is ordered from highest to lowest. While it is possible the median value may also not map directly to a data point from the data set, this number should at least be very close to a data point. Keep in mind, if your data has two or more distinct peaks, the median might still not be a good measures of the center. It is known as a robust statistic since it is not influenced by outliers. For the sample of 30 cars, the median was 22.55 miles per gallon.

# **Standard Deviation/Variance**

The usual reported measure of spread is the standard deviation or variance of a data set. The variance measures the average squared distance the set of points is from the mean. To calculate variance, the distances from each data point to the mean are calculated, squared, summed, and divided by the number of data points minus one. Formulas can be seen and further explained in any intro statistics textbook. Standard deviation is simply the square root of the variance, which gives a value that is back in the original units. Both of these will always be positive numbers; larger values indicate a larger spread while smaller values indicate a small spread. For the sample of 30 cars, the standard deviation was 7.55 miles per gallon. Formulas can be found in Appendix A.

# **Outliers/Influence Points**

An outlier or influential point is an extreme value that is not consistent with the rest of the data set. There can be multiple outliers in a data set. Outliers should be examined to see if they are a valuable part of the data set or a type of mistake. For example, if we had a car that weighed only 2000 lbs, but was getting only 4 mpg, that might seem a little off. Looking at the data again, it is more likely that this is a typo that was supposed to be 40 mpg or something similar, and the data point would be removed or corrected. Instead, if we measured a dump truck that weighed 25000 at 5 mpg, it would be a reasonable point that we would potentially want to keep in the data set. Removing this data point would not change our median much (from 22.30 to 22.55 mpg), but it would have a larger impact on the mean (from 22.94 to 23.54 mpg). In terms of a scatter plot, an outlier would appear as an unusual point that does not follow the trend of the data, and an influential point follow the data trend but in an extreme location.

# Conclusion

After gathering a sample of data, descriptive statistics should be the first step of analysis. Different graphs and summary information can be obtained depending on what type of data is collected. For

more information regarding any of these graphical or numerical summaries, please consult any introductory level statistics textbook. Also, please don't forget that this list is only a small snapshot of the many types of graphical and numerical summaries that can be created or calculated.

#### **References**

Ott, Lyman and Michael Longnecker. *An Introduction to Statistical Methods & Data Analysis.* 7<sup>th</sup> *ed.,* Cengage Learning, 2016.

Mendenhall, William, et al. Introduction to Probability and Statistics. 14<sup>th</sup> ed., Cengage Learning, 2013.

# **Appendix A: Data & Formulas**

## Data

#### Data for the random sample of M&M candies

Yellow	Green	Blue	Brown	Orange	Red
10	17	22	13	24	18

#### Data for the Sample of 30 cars (with added dump truck) – Note: all numbers are fictitious

Miles per Gallon	Weight (lbs)	Make
15.7	4382	Ford
16.8	3392	Chevy
21.1	2493	Chevy
21.4	2295	Ford
22.8	2395	Ford
17.1	3296	Chevy
19.4	2276	Toyota
21.3	2810	Ford
26.7	2596	Chevy
15.2	4401	Chevy
24.3	2705	Toyota
14.2	4296	Ford
30.1	2036	Chevy
27.1	2161	Toyota
13.0	4069	Chevy
12.9	3801	Ford
15.7	4169	Chevy
26.3	2261	Toyota
37.1	1864	Ford
18.1	3271	Ford
22.3	2591	Toyota
19.4	3616	Chevy
39.1	1963	Toyota
27.8	2891	Ford
23.3	2971	Ford
40.1	1774	Toyota
30.2	2388	Ford
25.4	2963	Toyota
34.6	2363	Chevy
27.6	2625	Ford
5.0	25000	Dump Truck

#### Data for the time series plot was generated through simulation.

#### **Formulas**

Mean formula:  $ar{x} = rac{\sum_{i=1}^n x_i}{n}$ 

Standard deviation formula: 
$$S = \sqrt{rac{\Sigma_{i=1}^n (x_i - ar{x})^2}{n-1}}$$

Variance formula: 
$$S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

**Note:** In each of the formulas above,  $x_i$  represents a single data point, n represents the number of pieces of data,  $\bar{x}$  represents the sample mean, s is the sample standard deviation, and  $s^2$  represents the sample variance.