1.2 | Exploratory Data Analysis in 1-D

The first step in any data analysis, regardless of dimensionality, is often exploratory in nature and includes summarizing and visualizing each of the variables or dimensions in the data set separately, without worrying about interactions between them or making any sort of inference to a hypothetical larger population. We refer to these methods as *univariate* or one-dimensional because they focus on a single variable (or dimension) at a time.

Although categorical data are common throughout most industries, the vast majority of data generated throughout the world are quantitative in nature. When working with a single quantitative dimension, we generally focus on summarizing and visualizing the *central tendency*, *dispersion*, and *shape* of the data.

1.2.1 | Measuring Central Tendency in 1-D

The most common numerical summary that is used to describe the central tendency of a distribution is the *mean*. The mean for a single quantitative variable is the numerical average of all n data values:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n} \tag{1.1}$$

We can use the functions `mean()` to calculate this numerical summary for any data set in R, or we can achieve the same result manually with vector indexing or just by using R as a calculator:

```
x <- c(3, 1, 4, 1, 5, 9, 2, 6, 5, 3, 5, 8, 9, 7, 9)
mean(x)
[1] 5.133333

(x[1]+ x[2] + x[3] + x[4] + x[5] + x[6] + x[7] + x[8] + x[9] + x[10]
+ x[11]+ x[12] + x[13] + x[14] + x[15]) / length(x)
[1] 5.133333

(3 + 1 + 4 + 1 + 5 + 9 + 2 + 6 + 5 + 3 + 5 + 8 + 9 + 7 + 9) / 15
[1] 5.133333
```

One thing to remember is that, unlike other measures of central tendency like *median* and *mode*, the mean is not *resistant* to the presence and effects of *outliers*. Conceptually, an outlier is a data point that is very different from other values and is unusually far from the central tendency. We will explore a more quantifiable definition of outliers, as well as how to measure *distance* from the central tendency, in Section 1.5.

It is also important to understand that any of the operations in R that can be applied to a vector can also be applied to a specific dimension or column in a data matrix or even a data frame.

Consider an example data set from a meta-analysis on the negative effects of timber harvest on the abundance of the southern red-backed vole (*Myodes gapperi*) in British Columbia, Canada. These data include the amount of `Area` removed during harvest in hectares, the number of `Patches` in which the removal occurred, and the corresponding `Decrease` in the abundance of voles over the next few years (when compared to nearby areas that were not harvested).

```
MG <- matrix(data = c( 9.90, 60,  3.9,
                       9.90,  9, 20.1,
                       9.90,  1, 88.6,
                       4.00, 15, 14.5,
                      10.00, 20, 17.7,
                       0.25,  1, 80.1), nrow = 6, byrow = T,
             dimnames = list(c("1","2","3","4","5","6"),
                             c("Area", "Patches", "Decrease")))
MG
   Area Patches Decrease
1  9.90      60    0.039
2  9.90       9    0.201
3  9.90       1    0.886
4  4.00      15    0.145
5 10.00      20    0.177
6  0.25       1    0.801
```

As one would expect, removing timber from a region always led to a decrease in vole abundance due to their reliance on forests for their habitat and diet, which consists primarily of seeds, nuts, roots, and berries. The goal of this meta-analysis was to synthesize results from numerous individual studies conducted across the region to provide a more robust estimate of how the amount of timber removed impacts vole populations and if the spatial configuration of this removal (e.g., one large clearcut or numerous spread-out patches) also has an effect.

Once a data matrix is available in R, we can use `mean()` to calculate the central tendency of any dimension by simply referencing the correct column index or name using matrix notation.

```
mean(MG[,3])
[1] 37.48333

mean(MG[,"Decrease"])
[1] 37.48333
```

We can see from our example data that across the six studies that were included, timber harvest resulted in an average decrease of 37.5% in the vole population when compared to nearby regions that were not harvested. Given that most data sets in R are stored using a matrix format, referencing specific dimensions within functions is a very common technique in data analysis.

1.2.2 | Measuring Dispersion in 1-D

While central tendency tells us where most of the data points lie, *dispersion* summarizes how far apart the points are from each other or from some measure of central tendency. Because data sets can have similar central tendencies but different levels of dispersion, and vice versa, it is important to consider both to get a full picture of the data.

Dispersion is commonly measured by first considering the *deviation* of a single data value x_i from the mean $\bar{x}$:

$$(x_i - \bar{x}) \tag{1.2}$$

Rather than working with this value as is, we usually use the *squared deviation* instead:

$$(x_i - \bar{x})^2 \tag{1.3}$$

Squaring the deviations helps to emphasize larger differences and ensures that all data points provide a positive value, in part so that the overall sum will not be zero. The sum of these squared deviations for all n data values in a dimension is referred to as the *sum of squares*:

$$SS = \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{1.4}$$

We can use R to quickly calculate this value:

```
ss <- sum((MG[,3] - mean(MG[,3]))^2)
ss
[1] 6778.728
```

By dividing the sum of squares by $n - 1$, which in effect averages the squared deviations that were summed together, we arrive at a common measure of dispersion known as the *variance*:

$$S^2 = \frac{SS}{n-1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \tag{1.5}$$

In R, we can use the `var()` function to find the variance or calculate it manually:

```
s2 <- ss / (nrow(MG) - 1)
s2
[1] 1355.746

var(MG[,3])
[1] 1355.746
```

Another common measure of dispersion is the *standard deviation*, S, which provides a rough estimate of the average distance of a data point from the mean:

$$S = \sqrt{S^2} = \sqrt{\frac{SS}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \tag{1.6}$$

We can again use R to calculate this value, either manually or by using the `sd()` function.

```
s <- sqrt(s2)
s
[1] 36.82045

sd(MG[,3])
[1] 36.82045
```

Much like the mean, both the variance and standard deviation are affected by outliers because they are calculated using every data point in the dimension, outliers and non-outliers alike.

A less common but nonetheless useful measure of dispersion is *quantiles*, which divide the data into equally sized consecutive sets. For example, the 50% quantile is the point at which 50% of the data have values less than this number, and can be found using the `quantile()` function and setting the `prob` argument to 0.50 (note the use of a proportion rather than a percentage).

```
quantile(MG[,3], prob = 0.50)
 50%
18.9
```

This indicates that half of the values in our example data that correspond to decreases in vole populations are less than 18.9, and the other half are greater than 18.9. Much like the mean, this value, which is known as the *median*, also provides a measure of central tendency.

Other commonly used quantiles include the 25% and 75% quantiles, which together provide a measure of the middle 50% of the data (also known as the *interquartile range*), and the 2.5% and 97.% quantiles, which together provide a measure of the middle 95% of the data.

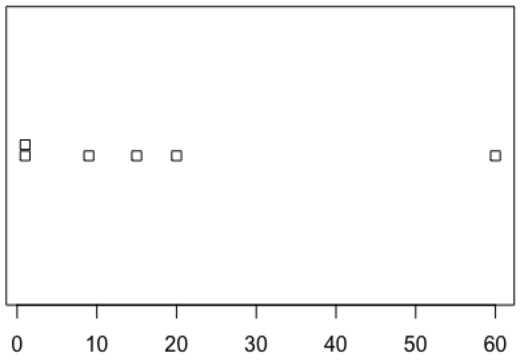
```
quantile(x, prob = c(0.25, 0.75))
 25%  75%
15.3 65.1
```

These and other quantile-based measures of dispersion are frequently used instead of the variance or standard deviation because of their resistance to the effects of outliers (something we will return to in Section 1.5).

1.2.3 | Measuring Shape in 1-D

Whereas central tendency and dispersion are commonly summarized numerically, the *shape* of a distribution is best explored visually, at least at the beginning. A common way to visualize and understand the shape of a distribution is with a *strip chart* (also known as a *dot plot*). We can create a strip chart in R by using the stripchart() function to stack a dot, square, or other symbol at the appropriate location along a horizontal axis for each case in the data set.

```
stripchart(MG[,2], method = "stack")
```



Other means of visualizing quantitative variables or dimensions include *histograms*, *boxplots*, and *stem-and-leaf displays*. Regardless of the exact approach, after a dimension has been visualized, we generally focus on describing its *modality*, *symmetry*, and *kurtosis*.

Modality describes the number of peaks in a data set. A *unimodal* distribution, such as the one pictured above, has one distinct peak that indicates the location of the most frequent values in the data set. Similarly, a *bimodal* distribution has two distinct peaks whereas a *multimodal* distribution has three or more distinct peaks.

A distribution is considered *symmetric* is we can fold the plot over a vertical center line and the two sides approximately match. Data that are not symmetric are referred to as *skewed*. If most of the data are clustered near the lower end of the horizontal axis, the data are said to be *skewed to the right*. If the pattern is reversed, as it is in the strip chart above, albeit very slightly, the data is *skewed to the left*.

In addition to visually inspecting the data for symmetry or lack thereof, we can quantify the asymmetry of a variable using *skewness*:

$$g_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{3/2}} \tag{1.7}$$

For a unimodal distribution, negative skewness indicates that the tail is on the left side of the distribution, whereas positive skewness indicates that the tail is on the right. Unimodal distributions that are perfectly symmetric will have a skewness of exactly zero.

```
g1 <- (1 / nrow(MG) * sum((MG[,2] - mean(MG[,2])) ^ 3)) /
      (1 / nrow(MG) * sum((MG[,2] - mean(MG[,2])) ^ 2)) ^ (3/2)
g1
[1] 1.344157
```

The value above indicates that the number of patches in our example data are skewed to the right, which agrees with our earlier visual inspection of the distribution.

Note that both the skewness and the variance in equation (1.5) are based on summing some power of the deviations of every data value from the mean, as was the sum of squares, further emphasizing the fundamental importance of the deviations defined in equation (1.2).

Another aspect of the shape of a distribution that is commonly summarized numerically is *kurtosis*, which measures the tailed-ness of a distribution:

$$g_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^2} \tag{1.8}$$

The kurtosis of any univariate normal distribution (discussed in more detail in Section 1.2.4 below) is 3; as such, it is common to compare the kurtosis of a distribution to this value.

Distributions with kurtosis less than 3 are said to be *platykurtic*, which means that the distribution produces fewer and less extreme outliers than does the normal distribution. Distributions with kurtosis greater than 3 are said to be *leptokurtic*, which has tails that asymptotically approach zero more quickly than a normal distribution.

```
g2 <- (1 / nrow(MG) * sum((MG[,2] - mean(MG[,2])) ^ 4)) /
      (1 / nrow(MG) * sum((MG[,2] - mean(MG[,2])) ^ 2)) ^ 2
g2
[1] 3.411018
```

This value indicates that the number of patches in our example data are leptokurtic and, as such, have a higher number of and more extreme outliers than a standard normal distribution.

Although these calculations are relatively simple, they are somewhat tedious to implement. Fortunately, many users contribute code to do all sorts of things in R, including making lengthy calculations and analyses implementable with a single command. They do this by writing bundles of code called *packages* and making them available for public download.
Accessing this code requires two steps:

1. The package must be installed onto your computer. This step can be accomplished using the function install.packages() or via **Tools -> Install Packages**. Packages only need to be installed once on each computer.
2. Each time we open R, we must tell R if we want to use any of the add-on packages that we have installed. We do this by using the library() function.

The moments package, for example, provides two functions that allow us to quickly calculate the skewness and kurtosis of a variable or dimension:

```
library(moments)
skewness(MG[,2])
[1] 1.344157

kurtosis(MG[,2])
[1] 3.411018
```

Unlike the *built-in functions* sqrt() and sum() used earlier, which are already created and defined in the programming framework, skewness() and kurtosis() will not work unless the moments package is installed and made available. This needs to be done each time we open R and is one of the more common reasons why previously-written R scripts stop working – the required packages are no longer available or haven't been installed on a new computer.

Thus far, we've focused on calculating numerical summaries for a single dimension at a time. In practice, however, we often want to calculate these values for each and every single dimension within our data. A common means of achieving this is to apply the relevant function, such as skewness() to every column in the data matrix using the apply() function:

```
apply(X = MG, MARGIN = 2, FUN = skewness)
      Area    Patches   Decrease
-0.9485507  1.3441566  0.6493345
```

The values above indicate that although both Patches and Decrease are right-skewed, the amount of Area removed is actually left-skewed.

The X and FUN arguments of the apply() function are used to denote the data matrix and the function that is to be applied to it, whereas the MARGIN argument is used to specify whether the function should be applied to the rows (value of 1) or columns (value of 2) of the matrix.

Although a number of functions have been built specifically to perform certain operations on the rows or columns of a matrix, including rowMeans() and colMeans(), the apply() function is particularly useful because it can be used on any function, even ones for which a specific row- or column-based variant has not been created. As such, we will make extensive use of this function throughout future analyses.

1.2.4 | The Normal Distribution

One distribution that is commonly encountered in data science and statistics is a *normal* or *Gaussian* distribution. Height, weight, reading ability, job satisfaction, stock prices and returns, blood pressure, and the size of parts produced by a machine in an assembly line are just a few examples of dimensions that approximately normally distributed.

Normal distributions have several characteristics that can easily be spotted using a strip chart or other visualization. The distribution is symmetric about the mean, with half the values falling below the mean. The data also follow a *bell-shaped curve*, with most values clustering around a central region and tapering off as they move further away from this central tendency.

Numerically, normal distributions have a skewness of zero, a kurtosis of three, and can be fully described using only the mean and standard deviation. The mean serves as the *location parameter* and determines where the peak of the curve is centered, whereas the standard deviation (the *scale parameter*) determines how stretched out or compressed the curve is.

Because normally distributed variables are so common, many analyses have been developed to operate under the assumption that data are normally distributed. Unfortunately, not all distributions approximate this bell-shaped pattern.

When a distribution is substantially different from normal, we can use a *transformation* to convert the data from one shape to another by applying a mathematical function to each data point in the dimension. Common data transformations include log, square root, square or other power, and reciprocal. For example, we can apply a log transformation to the number of patches in our example data using the `log()` function:

```
log(MG[,2])
       1        2        3        4        5        6
4.094345 2.197225 0.000000 2.708050 2.995732 0.000000
```

The first value of 4.09 in this new vector is simply the log of the first value of 60 in the original dimension. Because transformations are commonly used to make dimensions more normal, we would expect this newly transformed vector to have substantially less skew (i.e., closer to zero):

```
skewness(log(MG[,2]))
[1] -0.2562246
```

Once a transformation is selected, it is common practice to either add it as a new column to an existing data set or to simply replace the original untransformed data values with the newly transformed ones, thereby preserving the dimensionality of the data. Given the prevalence of non-normal data in many disciplines, transformations are likely to become an import part of your analytical toolbox!

Exercises for Section 1.2

1.5 Euler's Number. The number e, sometimes called the natural number of Euler's number, is an important mathematical constant that serves as the based for the natural logarithm. It is the limit of $(\mathbf{1} + \mathbf{1}/\boldsymbol{n})^{\boldsymbol{n}}$, an expression that arises in the study of compound interest in economics, and is equal to approximately:

$$\boldsymbol{e} = 2.71828\ 18284\ 59045\ 23536\ 02874\ 71352\ 66249\ 77572\ 47093\ 69995$$

It is named after its discoverer, the Swiss mathematician Leonhard Euler.

a. Construct a vector of the first four digits (i.e., 2, 7, 1, and 8) in R
b. Find the mean and standard deviation by using R as a calculator (without any built-in functions, which you may still use to check your answers)
c. Find the skewness by using R as a calculator
d. Find the kurtosis by using R as a calculator

1.6 Homelessness in Minnesota. While homelessness is an issue that deserves attention year-round, it often comes into sharper focus when temperatures plummet. Researchers at the Amherst H. Wilder Foundation have conducted a study of homeless individuals throughout the state of Minnesota once every three years to gather as much data as possible about the prevalence, causes, and effects of homelessness. An excerpt of the resulting data is below:

	Homelessness
1991	3100
1994	4600
1997	5600
2000	7700
2003	7800
2006	7800
2009	9700
2012	10200
2015	9300
2020	10200

These data show the total number of people (rounded to the nearest hundred) experiencing homelessness in the state over the past thirty years.

a. Construct this data matrix in R with the proper row and column names
b. Create a strip chart of the number of people experiencing Homelessness
c. Indicate whether the distribution appears to be skewed to the left, skewed to the right, or symmetrical
d. Without doing any calculations, indicate whether the skewness is likely to be positive, negative, or close to zero
e. Without doing any calculations, indicate whether the kurtosis is likely to be greater than 3, less than 3, or close to 3

f. Use `skewness()` and `kurtosis()` to find the actual values of these summary statistics and describe how they compare to your answers above

1.7 Natural Gas Prices. Natural gas is a naturally occurring hydrocarbon gas mixture that is used as a source of energy for heating, cooking, and electricity generation throughout the world. The U.S. Energy Information Administration (EIA) has been monitoring trends in the monthly price of this non-renewable resource since 1997. An excerpt of these data is below (additional data and information are available online at https://datahub.io/core/natural-gas):

	Price
2017-05	3.15
2017-06	2.98
2017-07	2.98
2017-08	2.9
2017-09	2.98
2017-10	2.88
2017-11	3.01
2017-12	2.82
2018-01	3.87
2018-02	2.67
2018-03	2.69
2018-04	2.8
2018-05	2.8
2018-06	2.97
2018-07	2.83
2018-08	2.96
2018-09	3
2018-10	3.28
2018-11	4.09
2018-12	4.04
2019-01	3.11
2019-02	2.69
2019-03	2.95
2019-04	2.65
2019-05	2.64
2019-06	2.4
2019-07	2.37
2019-08	2.22
2019-09	2.56
2019-10	2.33
2019-11	2.65
2019-12	2.22
2020-01	2.02
2020-02	1.91
2020-03	1.79
2020-04	1.74
2020-05	1.75
2020-06	1.63
2020-07	1.77
2020-08	2.3

The values above denote the average monthly price of natural gas over a period of 40 months.

a. Construct this data matrix in R with the proper row and column names
b. Create a strip chart and find the skewness of the monthly `Price`
c. Find the skewness of these data after they've undergone a logarithmic transformation
d. Find the skewness of these data after they've undergone a square transformation
e. Find the skewness of these data after they've undergone a square root transformation
f. Describe which transformation, if any, was most effective at reducing the skewness of the original data

1.8 Multidimensional Poverty Measure (continued). Exercise 1.2 introduced a sample of data collected by the World Bank Group on different South American countries' access to education, basic infrastructure, and other dimensions of poverty, which is recreated below:

	Water	Electricity	Sanitation	Education
Argentina	0.3	0.0	0.4	1.5
Bolivia	7.4	4.9	16.3	13.2
Brazil	1.7	0.2	34.2	16.0
Chile	0.1	0.3	0.6	4.0
Colombia	2.4	1.3	8.2	5.1
Ecuador	4.3	1.4	3.6	3.9
Paraguay	2.1	0.3	9.0	6.3
Peru	6.2	4.1	12.1	5.4
Uruguay	0.5	0.1	1.0	2.0

The values above denote what percent of each country's population did not have access to the designated dimension of poverty in 2019.

a. Construct this data matrix in R with the proper row and column names
b. Use `apply()` to find the skewness of each of the four dimensions
c. Use `apply()` to find the kurtosis of each of the four dimensions
d. Describe which dimension appears to be closest to being normally distributed