1

Exploring and Visualizing Data

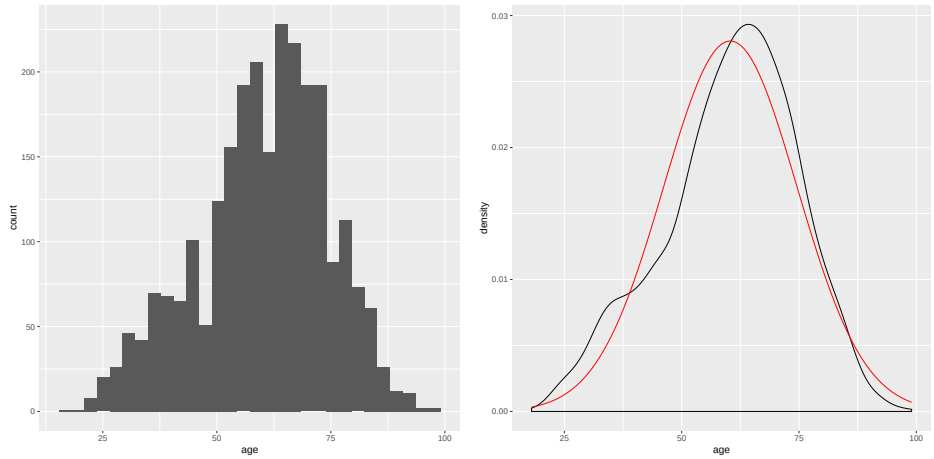
You have your plan, you carry out your plan by getting out and collecting your data, and then you put your data into a file. You are excited to test your hypothesis so you immediately run your multiple regression analysis and look at your output. You can do that (and probably will even if we advise against it), but what you need to do before you can start to make sense of that output is to look carefully at your data. You will want to know things like "how much spread do I have in my data" and "do I have any outliers". (If you have limited spread, you may discover that it is hard to explain variation in something that is nearly a constant and if you have an outlier, your statistics may be focused on trying to explain that one case.)

In this chapter, we will identify the ways to characterize your data before you do serious analysis both to understand what you are doing statistically and to error-check.

1.1 Characterizing Data

What does it mean to characterize your data? First, it means knowing how many observations are contained in your data, and the distribution of those observations over the range of your variable(s). What kinds of measures (interval, ordinal, nominal) do you have, and what are the ranges of valid measures for each variable? How many cases of missing (no data) or mis-coded (measures that fall outside the valid range) do you have? What do the coded values represent? While seemingly trivial, checking and evaluating your data for these attributes can save you major headaches later. For example, missing values for an observation often get a special code – say, "-99" – to distinguish them from valid observations. If you neglect to treat these values properly, R (or any other statistics program) will treat that value as if it were valid and thereby turn your results into a royal hairball. We know of cases in which even seasoned quantitative scholars have made the embarrassing mistake of failing to properly handle missing values in their analyses. In at least one case, a published paper had to be retracted for this reason. So don't skimp on the most basic forms of data characterization!

The dataset used for purposes of illustration in this version of this text is taken from



(a) Histogram (b) Density Curve

Figure 1.1: Distribution of Age

a survey of Oklahomans, conducted in 2016, by OU's Center for Risk and Crisis Management. The survey question wording and background will be provided in class. However, for purposes of this chapter, note that the measure of **ideology** consists of a self-report of political ideology on a scale that ranges from 1 (strong liberal) to 7 (strong conservative); the measure of the **perceived risk of climate change** ranges from zero (no risk) to 10 (extreme risk). **Age** was measured in years.

It is often useful to graph the variables in your data set to get a better idea of their distribution. In addition, we may want to compare the distribution of a variable to a theoretical distribution (typically a normal distribution). This can be accomplished in several ways, but we will show two here—a histogram and a density curve—and more will be discussed in later chapters. Here we examine the distribution of the variable measuring the perceived risk of climate change.

A histogram creates intervals of equal length, called bins, and displays the frequency of observations in each of the bins. To produce a histogram in R simply use the geom$_h istogramcommandinthe$

```
library(ggplot2)
```

```
ggplot(ds, aes(age)) +
  geom_density() +
   stat_function(fun = dnorm, args = list(mean = mean(ds$age, na.rm = T),
                                          sd = sd(ds$age, na.rm = T)), color = "red")
dev.off()
```

You can also get an overview of your data using a table known as a frequency distribution. The frequency distribution summarizes how often each value of your variable occurs in the dataset. If your variable has a limited number of values that it can take on, you can report all values, but if it has a large number of possible values (e.g., age of respondent), then you will want to create categories, or bins, to report those frequencies. In such cases, it is generally easier to make sense of the percentage distribution. Table 1.1 is a frequency

distribution for the ideology variable. From that table we see, for example, that about one-third of all respondents are moderates. We see the numbers decrease as we move away from that category, but not uniformly. There are a few more people on the conservative extreme than on the liberal side and that the number of people placing themselves in the penultimate categories on either end is greater than those towards the middle. The histogram and density curve would, of course, show the same pattern.

The other thing to watch for here (or in the charts) is whether there is an unusual observation. If one person scored 17 in this table, you could be pretty sure a coding error was made somewhere. You cannot find all your errors this way, but you can find some, including the ones that have the potential to most seriously adversely affect your analysis.

Table 1.1: Frequency Distribution for Ideology

Ideology	Frequency	Percentage	Cumulative Percentage
1 Strongly Liberal	122	4.8	4.8
2	279	11.1	15.9
3	185	7.3	23.2
4	571	22.6	45.8
5	328	13.0	58.8
6	688	27.3	86.1
7 Strongly Conservative	351	13.9	100.0
Total	2524	100	

In R, we can obtain the data for the above table with the following functions:

```
# frequency counts for each level
table(ds$ideol)

##
##   1   2   3   4   5   6   7
## 122 279 185 571 328 688 351

# To view percentages
library(dplyr)
table(ds$ideol) %>% prop.table()

##
##          1          2          3          4          5          6
## 0.04833597 0.11053883 0.07329635 0.22622821 0.12995246 0.27258320
##          7
## 0.13906498

# multiply the numbers by 100
table(ds$ideol) %>% prop.table() * 100

##
##         1         2         3         4         5         6         7
##  4.833597 11.053883  7.329635 22.622821 12.995246 27.258320 13.906498
```

Having obtained a sample, it is important to be able to characterize that sample. In particular, it is important to understand the probability distributions associated with each variable in the sample.

1.1.1 Central Tendency

Measures of central tendency are useful because a single statistic can be used to describe the distribution. We focus on three measures of central tendency; the mean, the median, and the mode.

Measures of Central Tendency

- The Mean: The arithmetic average of the values
- The Median: The value at the center of the distribution
- The Mode: The most frequently occurring value

We will primarily rely on the mean, because of its efficient property of representing the data. But medians – particularly when used in conjunction with the mean - can tell us a great deal about the shape of the distribution of our data. We will return to this point shortly.

1.1.2 Level of Measurement and Central Tendency

The three measures of central tendency – the mean, median, and mode – each tell us something different about our data, but each has some limitations (especially when used alone). Knowing the mode tells us what is most common, but we do not know how common and, using it alone, would not even leave us confident that it is an indicator of anything very *central*. When rolling in your data, it is generally a good idea to roll in all the descriptive statistics that you can to get a good feel for them.

One issue, though, is that your ability to use a statistic is dependent on the level of measurement for the variable. The mean requires you to add all your observations together. But you cannot perform mathematical functions on ordinal or nominal level measures. Your data must be measured at the interval level to calculate a meaningful mean. (If you ask R to calculate the mean student id number, it will, but what you get will be nonsense.) Finding the middle item in an order listing of your observations (the median) requires the ability to order your data, so your level of measurement must be at least ordinal. Therefore, if you have nominal level data, you can only report the mode (but no median or mean) so it is critical that you also look beyond central tendency to the overall distribution of the data.

1.1.3 Moments

In addition to measures of central tendency,"moments" are important ways to characterize the shape of the distribution of a sample variable. Moments are applicable when the data

measured is interval type (the level of measurement). The first four moments are those that are most often used.

The First Four Moments

1. *Expected Value*: The expected value of a variable, $E(X)$ is its mean.

$$E(X) = \bar{X} = \frac{\sum X_i}{n}$$

2. *Variance*: The variance of a variable concerns the way that the observed values are spread around either side of the mean.

$$s_x^2 = \frac{\sum (X - \bar{X})^2}{(n-1)}$$

3. *Skewness*: The skewness of a variable is a measure of its asymmetry.

$$S = \frac{\sum (X - \bar{X})^3}{(n-1)}$$

4. *Kurtosis*: The kurtosis of a variable is a measure of its peakedness.

$$K = \frac{\sum (X - \bar{X})^4}{(n-1)}$$

1.1.4 First Moment – Expected Value

The *expected value* of a variable is the value you would obtain if you could multiply all possible values within a population by their probability of occurrence. Alternatively, it can be understood as the mean value for a population variable. An expected value is a theoretical number , because we usually cannot observe all possible occurrences of a variable. The mean value for a sample is the average value for the variable X, and is calculated by adding the values of X and dividing by the sample size n:

$$\bar{X} = \frac{(x_1 + x_2 + x_3 + x_n)}{n} \tag{1.1}$$

This can be more compactly expressed as:

$$\bar{X} = \frac{\sum X_i}{n} \tag{1.2}$$

The mean of a variable can be calculated in `R` using the `mean` function. Here we illustrate the calculation of means for our measures of `ideology`, `age`, and `perceived risk of climate change`[1].

[1]The "na.rm=TRUE" portion of the following code simply tells R to exclude the missing (NA) values from calculation

```r
mean(ds$ideol, na.rm=TRUE)

## [1] 4.652932

mean(ds$age, na.rm=TRUE)

## [1] 60.36749

mean(ds$glbcc_risk, na.rm=TRUE)

## [1] 5.945978
```

1.1.5 The Second Moment – Variance and Standard Deviation

The *variance* of variable is a measure that illustrates how a variable is spread, or distributed, around its mean. For samples, it is expressed as:

$$s_x^2 = \frac{\sum(X - \bar{X})^2}{(n-1)} \tag{1.3}$$

The population variance is expressed as: σ_X^2.

Variance is measured in `squared` deviations from the mean, and the sum of these squared variations is termed the **total sum of squares**. Why squared deviations? Why not just sum the differences? While the latter strategy would seemingly be simpler, but it would always sum to zero. By squaring the deviations we make them all positive, so the sum of squares will always be a positive number.

Total Sum of Squares Is the squared summed total of the variation of a variable around its mean

This can be expressed as:

$$TSS_x = \sum(X_i - \bar{X})^2 \tag{1.4}$$

therefore;

$$s_x^2 = \frac{TSS_x}{(n-1)} \tag{1.5}$$

The square root of variance, σ_x^2, is the *standard deviation* (s.d.) of a variable, σ_x. The sample s.d. is expressed as:

$$s_x = \sqrt{\frac{\sum(X - \bar{X})^2}{(n-1)}} \tag{1.6}$$

This can also be expressed as $\sqrt{s_x^2}$. The standard deviation of a variable can be obtained in `R` with the `sd` function.[2]

[2]What's with those (n-1) terms in the denominators? These represent the "degrees of freedom" we have to calculate the average squared deviations and variance. We "use up" one of our observations to be able to calculate the first deviation – because without that first observation, what would there be to deviate from?

```
sd(ds$ideol, na.rm=TRUE)

## [1] 1.731246

sd(ds$age, na.rm=TRUE)

## [1] 14.20894

sd(ds$glbcc_risk, na.rm=TRUE)

## [1] 3.071251
```

1.1.6 The Third Moment – Skewness

Skewness is a measure of the asymmetry of a distribution. It is based on the third moment and is expressed as:

$$\frac{\sum(X - \bar{X})^3}{(n-1)} \tag{1.7}$$

Skewness is calculated by dividing the third moment by the the cube of the s.d.

$$S = \frac{\frac{\sum(X-\bar{X})^3}{(n-1)}}{(\sqrt{\frac{\sum(X-\bar{X})^2}{(n-1)}})^3} \tag{1.8}$$

Specifically, skewness refers to the position of the expected value (i.e., mean) of a variable distribution relative to its median. When the mean and median of a variable are roughly equal, $\bar{Y} \approx Md_Y$, then the distribution is considered approximately symmetrical, $S = 0$. This means that an equal proportion of the distribution of the variable lies on either side of the mean. However, when the mean is larger than the median, $\bar{Y} > Md_Y$, then the distribution has a *positive* skew, $S > 0$. When the median is larger than the mean, $\bar{Y} < Md_Y$, this is a *negative* skew, $S < 0$. This is illustrated in Figure 1.2. Note that for a normal distribution, $S = 0$.

1.1.7 The Fourth Moment – Kurtosis

The *kurtosis* of a distribution refers to the the peak of a variable (i.e., the mode) and the relative frequency of observations in the tails. It is based on the fourth moment which is expressed as:

$$\frac{\sum(X - \bar{X})^4}{(n-1)} \tag{1.9}$$

Kurtosis is calculated by dividing the fourth moment by the square of the second moment (i.e., variance).

$$K = \frac{\frac{\sum(X-\bar{X})^4}{(n-1)}}{(\frac{\sum(X-\bar{X})^2}{(n-1)})^2} \tag{1.10}$$

In general, higher kurtosis is indicative of a distribution where the variance is a result of low frequency yet more extreme observed values. In addition, when $K < 3$, the distribution is

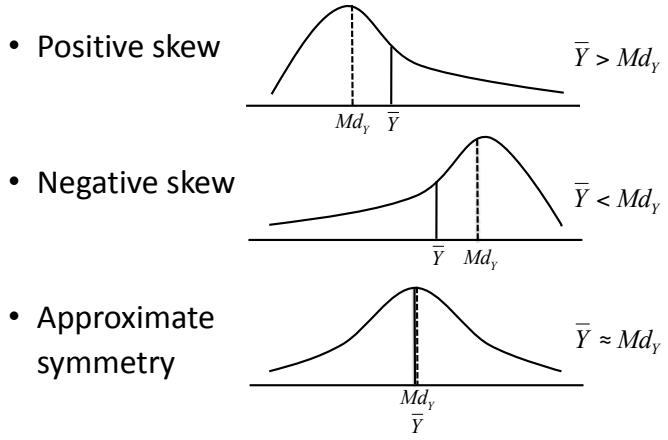


Figure 1.2: Distributional Shapes

platykurtic, which is flatter and/or more "short-tailed" than a normal distribution. When $K > 3$ the distribution is *leptokurtic*, which is a slim, high peak and long tails. In a normal distribution $K = 3$.

1.1.8 Order Statistics

Apart from central tendency and moments, probability distributions can also be characterized by **order statistics**. Order statistics are based on the position of a value in an ordered list. Typically, the list is ordered from low values to high values.

Order Statistics

Summaries of values based on position in an ordered list of all values. Types of order statistics include the minimum value, the maximum value, the median, quartiles, and percentiles.

- *Minimum Value*: The lowest value of a distribution
- *Maximum Value*: The highest value of a distribution
- *Median*: The value at the center of a distribution
- *Quartiles*: Divides the values into quarters
- *Percentiles*: Divides the values into hundredths

Median

The *median* is the value at the center of the distribution, therefore 50% of the observations in the distribution will have values above the median and 50% will have values below. For samples with a n-size that is an odd number, the median is simply the value in the middle. For example, with a sample consisting of the observed values of $1, 2, 3, 4, 5$, the median is 3. Distributions with an even numbered n-size, the median is the average of the two middle values. The median of a sample consisting of the observed values of $1, 2, 3, 4, 5, 6$ would be $\frac{3+4}{2}$ or 3.5.

The the median is the order statistic for central tendency. In addition, it is more "robust" in terms of extreme values than the mean. Extremely high values in a distribution can pull the mean higher, and extremely low values pull the mean lower. The median is less sensitive to these extreme values. The median is therefore the basis for "robust estimators," to be discussed later in this book.

Quartiles

Quartiles split the observations in a distribution into quarters. The first quartile, $Q1$, consists of observations whose values are within the first 25% of the distribution. The values of the second quartile, $Q2$, are contained within the first half (50%) of the distribution, and is marked by the distribution's median. The third quartile, $Q3$, includes the first 75% of the observations in the distribution.

The interquartile range (IQR) measures the spread of the ordered values. It is calculated by subtracting $Q1$ from $Q3$.

$$IQR = Q_3 - Q_1 \tag{1.11}$$

The IQR contains the middle 50% of the distribution.

We can visually examine the order statistics of a variable with a boxplot. A boxplot displays the range of the data, the first and third quartile, the median, and any outliers. To obtain a boxplot use the `boxplot` command. It is shown in Figure 1.3.

```
ggplot(ds, aes("", glbcc_risk)) +
  geom_boxplot()
```

Percentiles

Percentiles list the data in hundredths. For example, scoring in the 99th percentile on the GRE means that 99% of the other test takers had a lower score. Percentiles can be incorporated with quartiles (and/or other order statistics) such that:

- First Quartile: 25th percentile
- Second Quartile: 50th percentile (the median)
- Third Quartile: 75th percentile

Another way to compare a variable distribution to a theoretical distribution is with a quantile-comparison plot (qq plot). A qq plot displays the observed percentiles against those that would be expected in a normal distribution. This plot is often useful for examining the tails of the distribution, and deviations of a distribution from normality. This is shown in Figure 1.4.

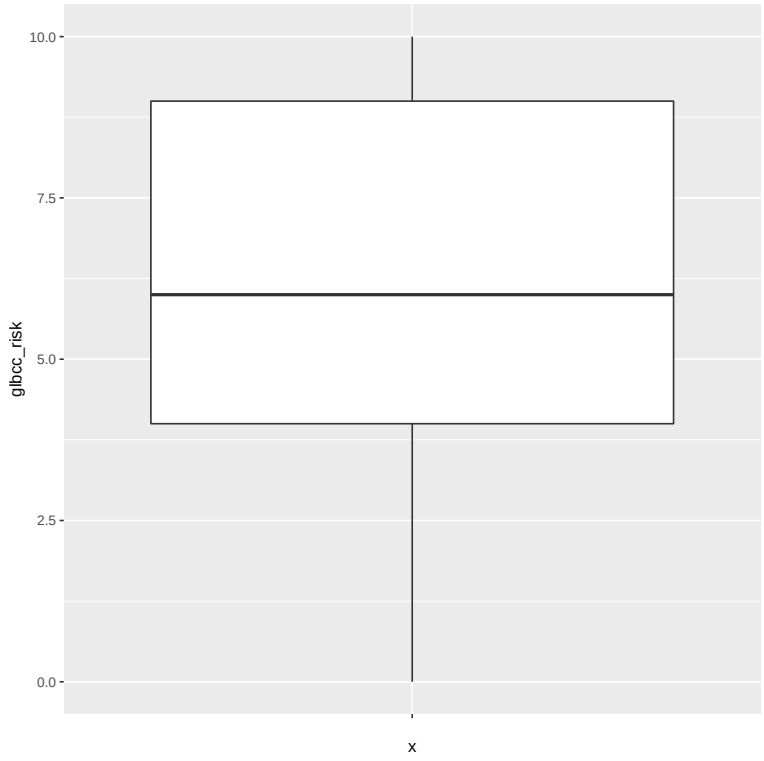


Figure 1.3: Box-plot of Climate Change Risk

```
ggplot(ds, aes(sample = glbcc_risk)) +
  stat_qq()
dev.off()
```

The qq plot provides an easy way to observe departures of a distribution from normality. For example, the plot shown in Figure 1.4 indicates the perceived risk measure has more observations in the tails of the distribution than would be expected if the variable was normally distributed.

R provides several ways to examine the central tendency, moments, and order statistics for variables and for entire data sets. The `summary` function produces the minimum value, the first quartile, median, mean, third quartile, max value, and the number of missing values (Na's).

```
summary(ds$ideol, na.rm=TRUE)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.000   4.000   5.000   4.653   6.000   7.000      23

summary(ds$age, na.rm=TRUE)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   52.00   62.00   60.37   70.00   99.00

summary(ds$gcclb_risk, na.rm=TRUE)
```

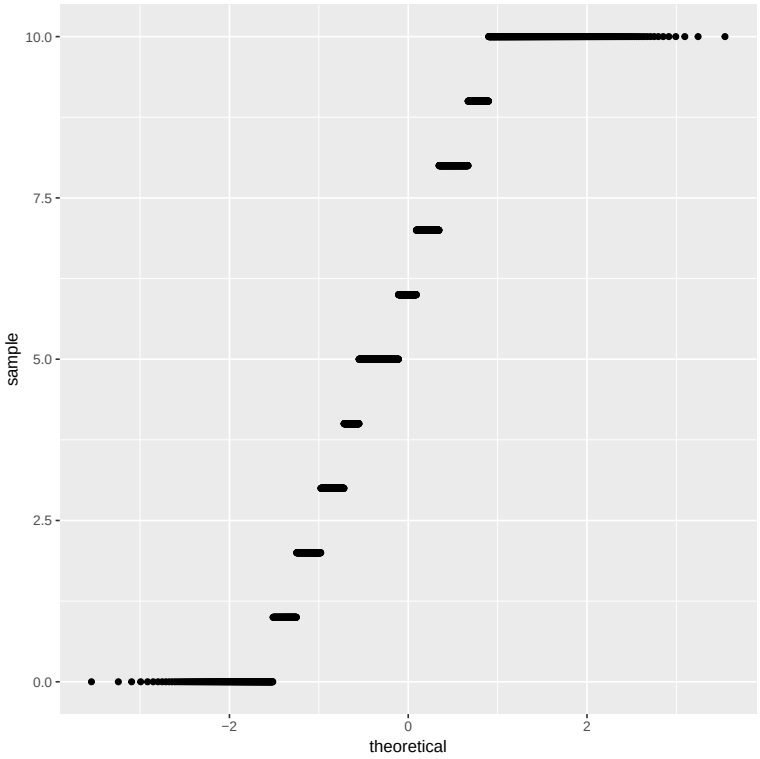


Figure 1.4: QQ Plot of Climate Change Risk

```
## Length  Class   Mode
##      0   NULL   NULL
```

We can also use the `describe` function in the `psych` package to obtain more descriptive statistics, including skewness and kurtosis.

```
library(psych)
describe(ds$ideol)

##    vars    n mean   sd median trimmed  mad min max range  skew kurtosis
## X1    1 2524 4.65 1.73      5    4.75 1.48   1   7     6 -0.45     -0.8
##      se
## X1 0.03
```

1.1.9 Summary

It is a serious mistake to get into your data analysis without understanding the basics of your data. Knowing their range, the general distribution of your data, the shape of that distribution, their central tendency, and so forth will give you important clues as you move through your analysis and interpretation and prevent serious errors from occurring. Readers also often need to know this information to provide a critical review of your work.

Overall, this chapter has focused on understanding and characterizing data. We refer to the early process of evaluating a data set as rolling in the data – getting to know the

characteristic shapes of the distributions of each of the variables, the meanings of the scales, and the quality of the observations. The discussion of central tendency, moments, and order statistics are all tools that you can use for that purpose. As a practicing scholar, policy analyst or public administration practitioner, this early stage in quantitative analysis is not optional; a failure to carefully and thoroughly understand your data can result in analytical disaster, excruciating embarrassment, and maybe even horrible encounters with the Killer Rabbit of Caerbannog.

Think of rolling in the data, then, as your version of the Holy Hand Grenade of Antioch.