3.1: 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 miscoded (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 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 dataset 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. For now, we examine the distribution of the variable measuring age. The red line on the density visualization presents the normal distribution given the mean and standard deviation of our variable.

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_histogram` command in the `ggplot2` package. Next, we plot the density of the observed data along with a normal curve. This can be done with the `geom_density` command in the `ggplot2` package.

```
library(ggplot2)
ggplot(ds, aes(age)) +
  geom_histogram()
```

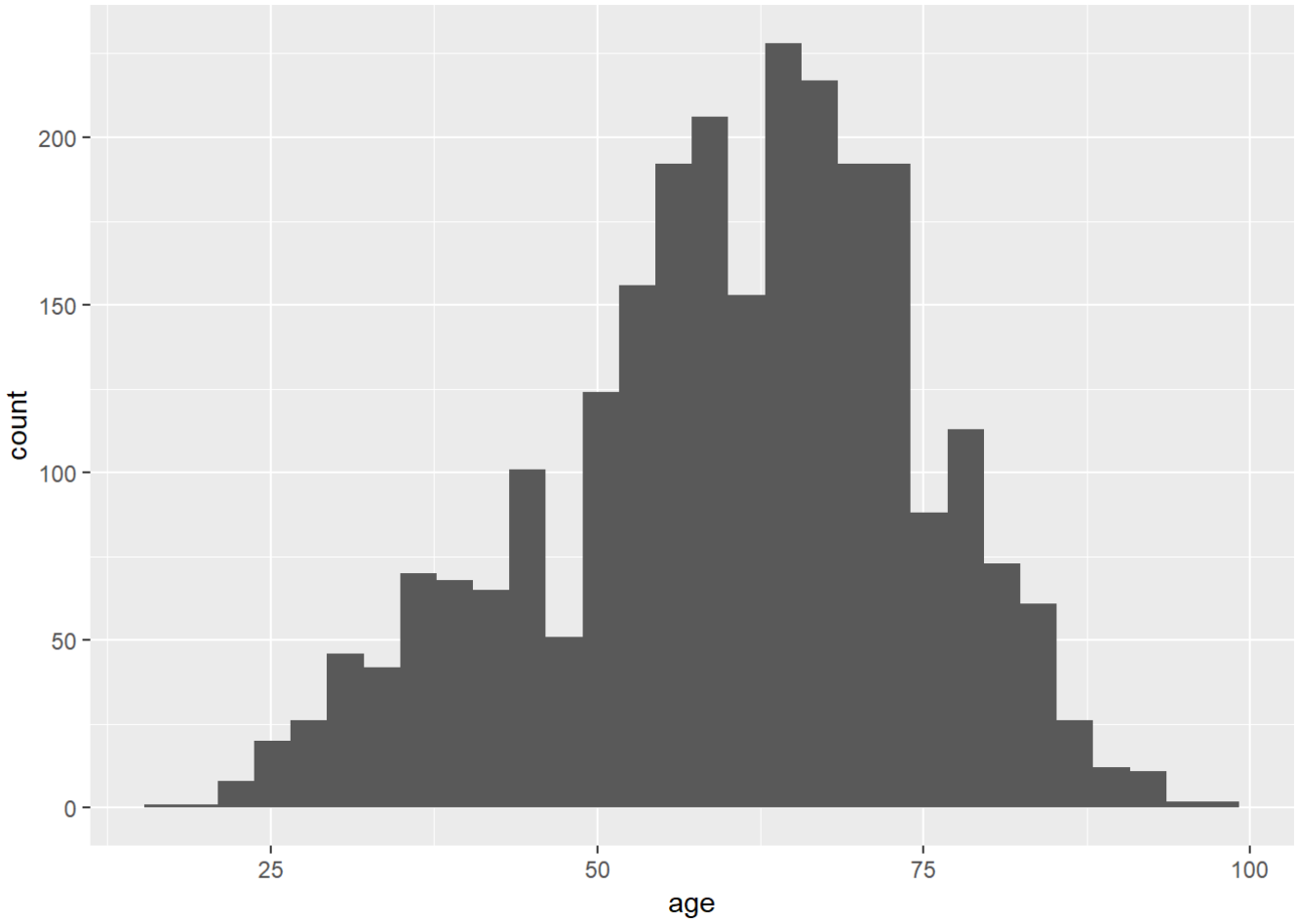

Figure 3.1.1: Histogram

```
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
```

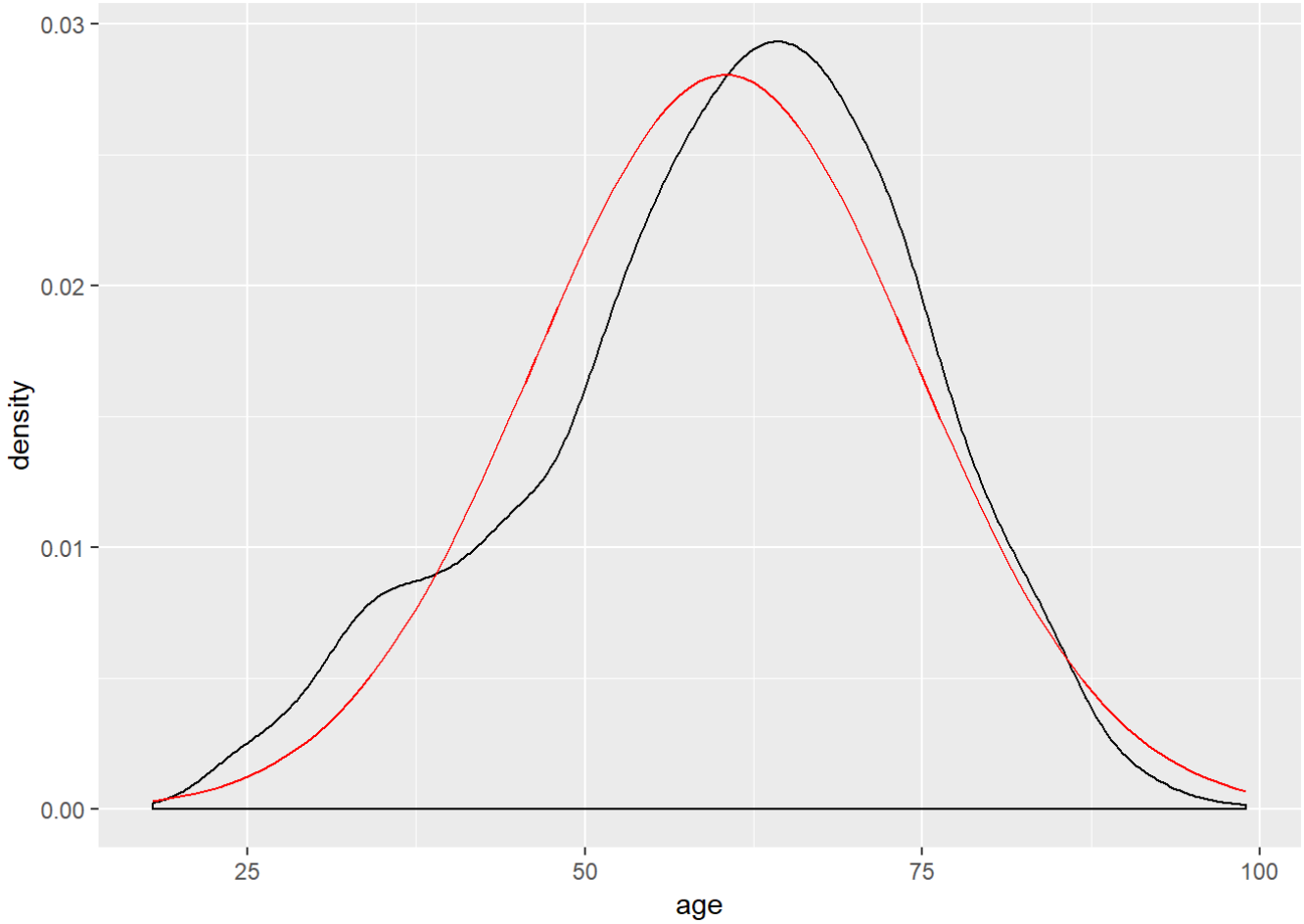

Figure 3.1.2: Density Curve

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 3.3 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.

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	

Figure 3.1.3: Frequency Distribution for Ideology

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.

3.1.1 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.

3.1.2 Level of Measurement and Central Tendency

The three measures of central tendency – the mean, median, and mode – each tells us something different about our data, but each has some limitations as well (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 any 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 ordered 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 the central tendency to the overall distribution of the data.

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

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

$E(X)=\bar{X}=\sum Xin$