Chapter 4
Descriptive Statistics

Before developing any models with or attempting to draw any inferences from a data set, the user should first get a sense of the features of the data. This can be accomplished through the data visualization methods described in Chap. 3, as well as through descriptive statistics of a variable's central tendency and dispersion, described in this chapter. Ideally, the user will perform both tasks, regardless of whether the results become part of the final published product. A traditional recommendation to analysts who estimate functions such as regression models is that the first table of the article ought to describe the descriptive statistics of all input variables and the outcome variable. While some journals have now turned away from using scarce print space on tables of descriptive statistics, a good data analyst will always create this table for him or herself. Frequently this information can at least be reported in online appendices, if not in the printed version of the article.

As we work through descriptive statistics, the working example in this chapter will be policy-focused data from LaLonde's (1986) analysis of the National Supported Work Demonstration, a 1970s program that helped long-term unemployed individuals find private sector jobs and covered the labor costs of their employment for a year. The variables in this data frame are:

treated: Indicator variable for whether the participant received the treatment.
age: Measured in years.
education: Years of education.
black: Indicator variable for whether the participant is African-American.
married: Indicator variable for whether the participant is married.
nodegree: Indicator variable for not possessing a high school diploma.

Electronic supplementary material: The online version of this chapter (doi: 10.1007/978-3-319-23446-5_4) contains supplementary material, which is available to authorized users.

J.E. Monogan III, *Political Analysis Using R*, Use R!,
DOI 10.1007/978-3-319-23446-5_4

re74: Real earnings in 1974.
re75: Real earnings in 1975.
re78: Real earnings in 1978.
hispanic: Indicator variable for whether the participant is Hispanic.
u74: Indicator variable for unemployed in 1974.
u75: Indicator variable for unemployed in 1975.

4.1 Measures of Central Tendency

Our first task will be to calculate centrality measures, which give us a sense of a typical value of a distribution. The most common measures of central tendency are the mean, median, and mode. The inter-quartile range, offering the middle 50 % of the data, is also informative. To begin with some example calculations, we first must load LaLonde's data (named LL). These are available as part of the **C**oarsened **E**xact **M**atching package (cem), which we will discuss at greater length in Chap. 8. As with any other user-defined package, our first task is to install the package:

```
install.packages("cem")
library(cem)
data(LL)
```

After installing the package, we load the library, as we will have to do in every session in which we use the package. Once the library is loaded, we can load the data simply by calling the data command, which loads this saved data frame from the cem package into working memory. We conveniently can refer to the data frame by the name LL[1].

For all of the measures of central tendency that we compute, suppose we have a single variable x, with n different values: $x_1, x_2, x_3, \ldots, x_n$. We also could sort the values from smallest to largest, which is designated differently with *order statistics* as: $x_{(1)}, x_{(2)}, x_{(3)}, \ldots x_{(n)}$. In other words, if someone asked you for the second order statistic, you would tell them the value of $x_{(2)}$, the second smallest value of the variable.

With a variable like this, the most commonly used measure of centrality is the sample *mean*. Mathematically, we compute this as the average of the observed values:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{4.1}$$

Within R, we can apply Eq. (4.1)'s formula using the mean function. So if x in this case was the income participants in the National Supported Work Demonstration earned in 1974, we would apply the function to the variable re74:

[1] These data also are available in comma-separated format in the file named LL.csv. This data file can be downloaded from the Dataverse on page vii or the chapter content link on page 53.

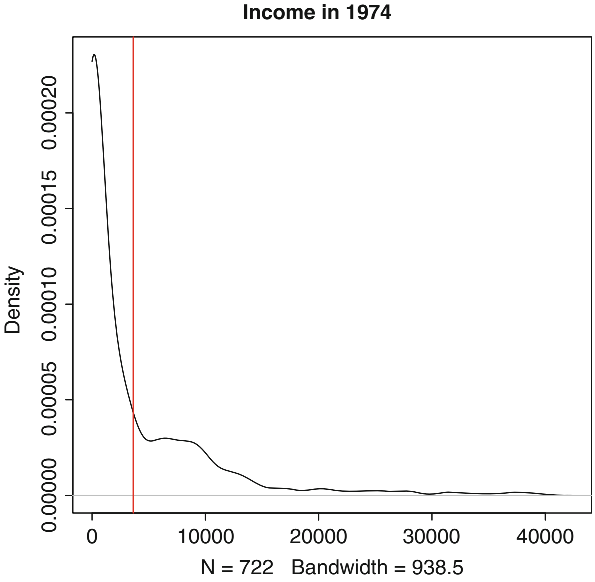

Fig. 4.1 Density plot of real earnings in 1974 from the National Supported Work Demonstration data

```
mean(LL$re74)
```

R responds by printing `[1] 3630.738`, so we can report the sample mean as $\bar{x} = 3630.738$.

Of course, it is advisable to continue to visualize the data using the tools from Chap. 3. Besides computing the mean of real earnings in 1974, we also can learn a lot just from drawing a density plot. We could do this using the `lattice` code described in the last chapter or with a little more user control as follows:

```
dens.74<-density(LL$re74,from=0)
plot(dens.74,main="Income in 1974")
abline(v=mean(LL$re74),col="red")
```

On the first line, the `density` command allows us to compute the density of observations at each value of income. With the `from` option, we can specify that the minimum possible value of income is 0 (and the `to` option would have let us set a maximum). On the second line, we simply plot this density object. Lastly, we use `abline` to add a vertical line where our computed mean of $3,630.74 is located.

The resulting graph is shown in Fig. 4.1. This figure is revealing: The bulk of the data fall below the mean. The mean is as high as it is because a handful of very large incomes (shown in the long right tail of the graph) are drawing it upward. With the picture, we quickly get a sense of the overall distribution of the data.

Turning back to statistical representations, another common measure of central tendency is the sample *median*. One advantage of computing a median is that it is more robust to extreme values than the mean. Imagine if our sample had somehow included Warren Buffett—our estimate of mean income would have increased substantially just with one observation. The median, by contrast, would move very little in response to such an extreme observation. Our formula for computing a median with observed data turns to the order statistics we defined above:

$$\tilde{x} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{where } n \text{ is odd} \\ \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(1+\frac{n}{2}\right)}\right) & \text{where } n \text{ is even} \end{cases} \tag{4.2}$$

Note that notation for the median is somewhat scattered, and $\tilde{x}$ is one of the several commonly used symbols. Formally, whenever we have an odd number of values, we simply take the middle order statistic (or middle value when the data are sorted from smallest to largest). Whenever we have an even number of values, we take the two middle order statistics and average between them. (E.g., for ten observations, split the difference between $x_{(5)}$ and $x_{(6)}$ to get the median.) R will order our data, find the middle values, and take any averages to report the median if we simply type:

```
median(LL$re74)
```

In this case, R prints `[1] 823.8215`, so we can report $\tilde{x} = 823.8215$ as the median income for program participants in 1974. Observe that the median value is *much* lower than the mean value, $2,806.92 lower, to be exact. This is consistent with what we saw in Fig. 4.1: We have a positive skew to our data, with some extreme values pulling the mean up somewhat. Later, we will further verify this by looking at quantiles of our distribution.

A third useful measure of central tendency reports a range of central values. The *inter-quartile range* is the middle 50 % of the data. Using order statistics, we compute the lower and upper bounds of this quantity as:

$$\text{IQR}_x = \left[x_{\left(\frac{n}{4}\right)}, x_{\left(\frac{3n}{4}\right)}\right] \tag{4.3}$$

The two quantities reported are called the first and third *quartiles*. The first quartile is a value for which 25 % of the data are less than or equal to the value. Similarly, 75 % of the data are less than or equal to the third quartile. In this way, the middle 50 % of the data falls between these two values. In R, there are two commands we can type to get information on the inter-quartile range:

```
summary(LL$re74)
IQR(LL$re74)
```

The `summary` command is useful because it presents the median and mean of the variable in one place, along with the minimum, maximum, first quartile, and third quartile. Our output from R looks like this:

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0     0.0   823.8  3631.0  5212.0 39570.0
```

This is handy for getting three measures of central tendency at once, though note that the values of the mean and median by default are rounded to fewer digits that the separate commands reported. Meanwhile, the interquartile range can be read from the printed output as $\text{IQR}_x = [0, 5212]$. Normally, we would say that at least 50 % of participants had an income between \$0 and \$5212. In this case, though, we know no one earned a negative income, so 75 % of respondents fell into this range. Finally, the `IQR` command reports the difference between the third and first quartiles, in this case printing: `[1] 5211.795`. This command, then, simply reports the spread between the bottom and top of the interquartile range, again with less rounding that we would have gotten by using the numbers reported by `summary`.

In most circumstances, rounding and the slight differences in outputs that these commands produce pose little issue. However, if more digits are desired, the user can control a variety of *global options* that shape the way R presents results with the `options` command that was mentioned in Chap. 1. The `digits` argument specifically shapes the number of digits presented. So, for example, we could type:

```
options(digits=9)
summary(LL$re74)
```

Our output would then look like this:

```
  Min.    1st Qu.    Median       Mean    3rd Qu.       Max.
0.000      0.000   823.822   3630.740   5211.790 39570.700
```

Hence, we can see that discrepancies are a function of rounding. Bear in mind, though, that changes with the `options` command apply to *all* outputs in the session. For instance, it turns out that if we re-ran the `mean` command the output would now show even more digits than we had before.

We also could get a general summary of all variables in a data set at once just by typing the name of the data frame alone into the `summary` command:

```
summary(LL)
```

This reports the same descriptive statistics as before, but for all variables at once. If any observations are missing a value for a variable, this command will print the number of `NA` values for the variable. Beware, though, not all of the quantities reported in this table are meaningful. For indicator variables such as **treated**, **black**, **married**, **nodegree**, **hispanic**, **u74**, and **u75**, remember that the variables are not continuous. The mean essentially reports the proportion of respondents receiving a 1 rather than a 0, and the count of any missing values is useful. However, the other information is not particularly informative.

4.1.1 Frequency Tables

For variables that are measured nominally or ordinally, the best summary of information is often a simple table showing the frequency of each value. In R, the `table` command reports this for us. For instance, our data include a simple indicator coded 1 if the respondent is African-American and 0 otherwise. To get the relative frequencies of this variable, we type:

```
table(LL$black)
```

This prints the output:

```
  0   1
144 578
```

Hence, 144 respondents are not African-American, and 578 respondents are African-American. With a nominal indicator such as this, the only valid measure of central tendency is the *mode*, which is the most common value a variable takes on. In this case, the most frequent value is a 1, so we could say that the mode is African-American.

As another example, these data measure education in years. Measured in this way, we could think of this as a continuous variable. Yet, the number of values this variable takes on is somewhat limited, as no one in the sample has fewer than three or more than 16 years of education. We therefore may wish to look at a frequency table for this variable as well. Also, even if the mean and the median are informative values, we may wish to compute the mode to know what the most common number of years of education is in these data. To compute the frequency table and to have R automatically return the mode, we type:

```
table(LL$education)
which.max(table(LL$education))
```

The table we see is:

```
  3   4   5   6   7   8   9  10  11  12  13  14  15  16
  1   6   5   7  15  62 110 162 195 122  23  11   2   1
```

At a glance, few respondents never went to high school at all, and only a handful have more than a high school education. We also could scan the table to observe that the mode is 11 years of education, which describes 195 respondents. However, if we have many more categories than we do for **education**, doing this will become difficult. Hence, feeding our table into the `which.max` command returns **which** label in the table that has the **max**imum frequency. Our resulting printout is:

```
11
 9
```

The first line, `11`, prints the value label of the cell with the highest frequency—this is our mode. The second line adds the additional detail that the value of `11` is the ninth cell of the table (a detail we usually can ignore).

Another way we could present our frequencies is in a bar plot based on the above table. We could do this with the following code:

```
barplot(table(LL$education),xlab="Years of Education",
     ylab="Frequency",cex.axis=.9,cex.names=.9,ylim=c(0,200))
abline(h=0,col='gray60')
box()
```

On the first line, we specify that we are drawing a bar plot of `table(LL$education)`. Notice that we use `cex.axis` and `cex.names` to reduce the size

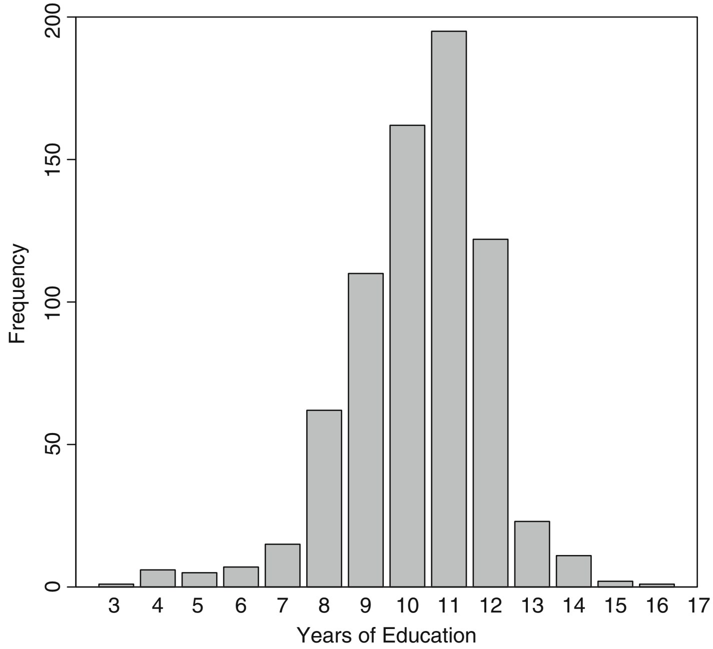

Fig. 4.2 Distribution of number of years of education from the National Supported Work Demonstration data

of the text on the vertical and horizontal axis, respectively. Afterward, we add a baseline at 0 and draw a box around the full figure. The result is shown in Fig. 4.2. With this plot, we can easily spot that the highest bar, our mode, is at 11 years of education. The graph also gives us a quick sense of the spread of the other values.

As a side point, suppose an analyst wanted not just a table of frequencies, but the percentage of values in each category. This could be accomplished simply by dividing the table by the number of cases and multiplying by 100. So for the percent of respondents falling into each category on the education variable, we type:

```
100*table(LL$education)/sum(table(LL$education))
```

R then will print:

```
         3          4          5          6 
 0.1385042  0.8310249  0.6925208  0.9695291 
         7          8          9         10 
 2.0775623  8.5872576 15.2354571 22.4376731 
        11         12         13         14 
27.0083102 16.8975069  3.1855956  1.5235457 
        15         16 
 0.2770083  0.1385042 
```

This output now shows the percentage of observations falling into each category.

4.2 Measures of Dispersion

Besides getting a sense of the center of our variable, we also would like to know how spread out our observations are. The most common measures of this are the variance and standard deviation, though we also will discuss the median average deviation as an alternative measure. Starting with the sample variance, our formula for this quantity is:

$$\mathrm{Var}(x) = s_x^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{4.4}$$

In R, we obtain this quantity with the function `var`. For income in 1974, we type:

```
var(LL$re74)
```

This prints the value: `[1] 38696328`. Hence, we can write $\mathrm{Var}(x) = 38696328$. Of course, the variance is in a squared metric. Since we may not want to think of the spread in terms of "38.7 million squared dollars," we will turn to alternative measures of dispersion as well. That said, the variance is an essential quantity that feeds into a variety of other calculations of interest.

The standard deviation is simply the square root of the variance:

$$\mathrm{SD}(x) = s_x = \sqrt{\mathrm{Var}(x)} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{4.5}$$

This simple transformation of the variance has the nice property of putting our measure of dispersion back onto the original scale. We could either take the square root of a computed variance, or allow R to do all steps of the calculation for us:

```
sd(LL$re74)
```

In this case, R prints: `[1] 6220.637`. Hence, $s_x = 6220.637$. When a variable is shaped like a normal distribution (which our income variable is not), a useful approximation is the 68-95-99.7 rule. This means that approximately 68 % of our data fall within one standard deviation of the mean, 95 % within two standard deviations, and 99.7 % within three standard deviations. For income in 1974, a heavy concentration of incomes at $0 throws this rule off, but with many other variables it will observationally hold.

A very different measure of dispersion is the *median absolute deviation*. We define this as:

$$\mathrm{MAD}(x) = \mathrm{median}(|x_i - \mathrm{median}(x)|) \tag{4.6}$$

In this case, we use the median as our measure of centrality, rather than the mean. Then we compute the absolute difference between each observation and the median. Lastly, we compute the median *of the deviations*. This offers us a sense of a typical deviation from the median. In R the command is typed:

```
mad(LL$re74)
```

Here, R returns a value of 1221.398. Like the standard deviation, this is on the scale of the original variable, in dollars. Unlike the standard deviation, this statistic turns out to be much smaller in this case. Again, extreme values can really run-up variances and standard deviations, just as they can distort a mean. The median absolute deviation, by contrast, is less sensitive to extreme values.

4.2.1 *Quantiles and Percentiles*

As a final topic, *quantiles* and *percentiles* allow us to gain a sense of a variable's overall distribution. Quantiles are the relative placement of data values in a sorted list, scaled $[0, 1]$. For a value q the quantile for that value would be the order statistic $x_{(q \cdot n)}$. Percentiles are the same thing, scaled $[0, 100]$, so for a value p the pth percentile would be $x_{\left(\frac{p \cdot n}{100}\right)}$. Hence, the median is the 0.5 quantile and the 50th percentile. Special cases of quantiles include the previously introduced *quartiles* (dividing the data into four groups), *quintiles* (dividing into five groups), and *deciles* (dividing into ten groups).

In R, the command `quantile` can give us any quantile we wish. By default, R prints the quantiles for $q \in \{0.00, 0.25, 0.50, 0.75, 1.00\}$. We have the option of specifying whichever quantiles we want, though. The syntax is:

```
quantile(LL$re74)
quantile(LL$re74, c(0,.1,.2,.3,.4,.5,.6,.7,.8,.9,1))
```

The first command prints our default quantiles, though it reports them with the rescaled percentile labels:

```
       0%        25%        50%        75%       100%
   0.0000     0.0000   823.8215  5211.7946 39570.6797
```

Essentially, this information repeats the quartile information that `summary` provided us earlier. On our second line of code, we add a vector of 11 quantiles of interest to request deciles, which give us the cut points for each additional 10 % of the data. This result is:

```
       0%        10%        20%        30%
   0.0000     0.0000     0.0000     0.0000
      40%        50%        60%        70%
   0.0000   823.8215  1837.2208  3343.5705
      80%        90%       100%
6651.6747 10393.2177 39570.6797
```

This is revealing as it shows that at least 40 % of our respondents had an income of $0 in 1974. Further, going from the 90th percentile to the 100th percentile (or maximum), we see a jump from $10,393 to $39,570, suggesting that some particularly extreme values are in the top 10 % of our data. Hence, these data do have a substantial positive skew to them, explaining why our computed median is so different from the mean.

In this chapter, we have covered the various means by which we can compute measures of centrality and dispersion in R. We have also discussed frequency tables and quantiles. Together with the graphing techniques of Chap. 3, we now have a big basket of tools for assessing and reporting the attributes of a data set. In the coming chapter, we will turn to drawing inferences from our data.

4.3 Practice Problems

Consider again Peake and Eshbaugh-Soha's (2008) analysis of drug policy coverage, which was introduced in the practice problems for Chap. 3. Recall that the comma-separated data file is named `drugCoverage.csv`. If you do not have it downloaded already, please visit the Dataverse (see page vii) or this chapter's online content (see page 53). Again, the variables are: a character-based time index showing month and year (**Year**), news coverage of drugs (**drugsmedia**), an indicator for a speech on drugs that Ronald Reagan gave in September 1986 (**rwr86**), an indicator for a speech George H.W. Bush gave in September 1989 (**ghwb89**), the president's approval rating (**approval**), and the unemployment rate (**unemploy**).

1. What can you learn simply by applying the `summary` command to the full data set? What jumps out most clearly from this output? Are there any missing values in these data?
2. Using the `mean` function, compute the following:
   - What is the mean of the indicator for George H.W. Bush's 1989 speech? Given that this variable only takes on values of 0 and 1, how would you interpret this quantity?
   - What is the mean level of presidential approval? How would you interpret this quantity?
3. What is the median level of media coverage of drug-related issues?
4. What is the interquartile range of media coverage of drug-related issues?
5. Report two frequency tables:
   - In the first, report the frequency of values for the indicator for Ronald Reagan's 1986 speech.
   - In the second, report the frequency of values for the unemployment rate in a given month.
   - What is the modal value of unemployment? In what percentage of months does the mode occur?
6. What are the variance, standard deviation, and median absolute deviation for news coverage of drugs?
7. What are the 10th and 90th percentiles of presidential approval in this 1977–1992 time frame?