

CHAPTER OVERVIEW

6: Describing Data With Numbers Using R

Any time that you get a new data set to look at, one of the first tasks that you have to do is find ways of summarising the data in a compact, easily understood fashion. This is what *descriptive statistics* (as opposed to inferential statistics) is all about. In fact, to many people the term “statistics” is synonymous with descriptive statistics. It is this topic that we’ll consider in this chapter, but before going into any details, let’s take a moment to get a sense of why we need descriptive statistics. To do this, let’s load the `aflsmall.Rdata` file, and use the `who()` function in the `lsr` package to see what variables are stored in the file:

```
load( "../data/aflsmall.Rdata" )
library(lsr)
```

```
## Warning: package 'lsr' was built under R version 3.5.2
```

```
who()
```

```
##      -- Name --      -- Class --      -- Size --
##      afl.finalists   factor         400
##      afl.margins     numeric        176
```

There are two variables here, `afl.finalists` and `afl.margins`. We’ll focus a bit on these two variables in this chapter, so I’d better tell you what they are. Unlike most of data sets in this book, these are actually real data, relating to the Australian Football League (AFL)⁶⁴. The `afl.margins` variable contains the winning margin (number of points) for all 176 home and away games played during the 2010 season. The `afl.finalists` variable contains the names of all 400 teams that played in all 200 finals matches played during the period 1987 to 2010. Let’s have a look at the `afl.margins` variable:

```
print(afl.margins)
```

```
##      [1] 56 31 56 8 32 14 36 56 19 1 3 104 43 44 72 9 28
##      [18] 25 27 55 20 16 16 7 23 40 48 64 22 55 95 15 49 52
##      [35] 50 10 65 12 39 36 3 26 23 20 43 108 53 38 4 8 3
##      [52] 13 66 67 50 61 36 38 29 9 81 3 26 12 36 37 70 1
##      [69] 35 12 50 35 9 54 47 8 47 2 29 61 38 41 23 24 1
##      [86] 9 11 10 29 47 71 38 49 65 18 0 16 9 19 36 60 24
##     [103] 25 44 55 3 57 83 84 35 4 35 26 22 2 14 19 30 19
##     [120] 68 11 75 48 32 36 39 50 11 0 63 82 26 3 82 73 19
##     [137] 33 48 8 10 53 20 71 75 76 54 44 5 22 94 29 8 98
##     [154] 9 89 1 101 7 21 52 42 21 116 3 44 29 27 16 6 44
##     [171] 3 28 38 29 10 10
```

This output doesn’t make it easy to get a sense of what the data are actually saying. Just “looking at the data” isn’t a terribly effective way of understanding data. In order to get some idea about what’s going on, we need to calculate some descriptive statistics (this chapter) and draw some nice pictures (Chapter 6). Since the descriptive statistics are the easier of the two topics, I’ll start with those, but nevertheless I’ll show you a histogram of the `afl.margins` data, since it should help you get a sense of what the data we’re trying to describe actually look like. But for what it’s worth, this histogram – which is shown in Figure 5.1 – was generated using the `hist()` function. We’ll talk a lot more about how to draw histograms in Section 6.3. For now, it’s enough to look at the histogram and note that it provides a fairly interpretable representation of the `afl.margins` data.

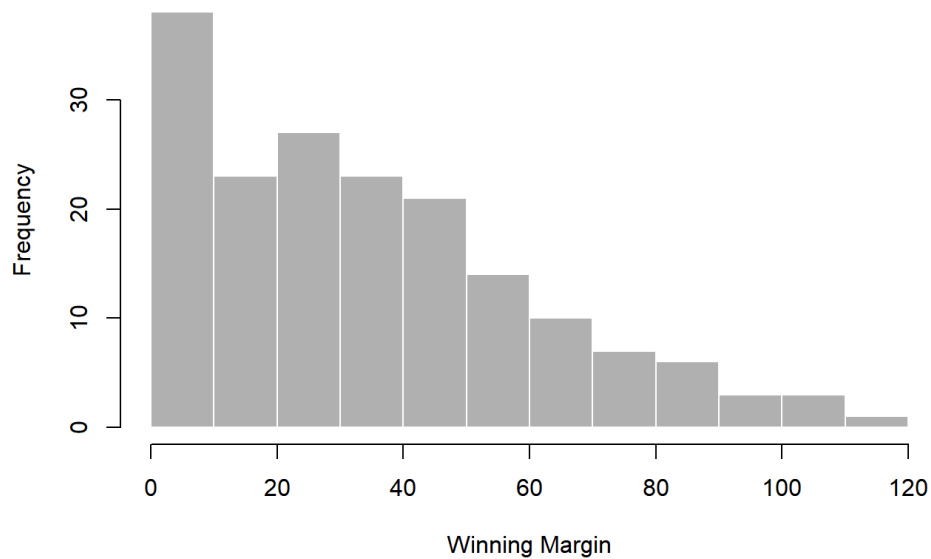


Figure 5.1: A histogram of the AFL 2010 winning margin data (the `afl.margins` variable). As you might expect, the larger the margin the less frequently you tend to see it.

[6.1: Measures of Central Tendency](#)

[6.2: Measures of Variability](#)

[6.3: Skew and Kurtosis](#)

[6.4: Getting an Overall Summary of a Variable](#)

[6.5: Descriptive Statistics Separately for each Group](#)

[6.6: Standard Scores](#)

[6.7: Epilogue- Good Descriptive Statistics Are Descriptive!](#)

This page titled [6: Describing Data With Numbers Using R](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by [Danielle Navarro](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.