CHAPTER 4

Summary Statistics

Simply throwing a bunch of numbers at your audience will only confuse them. Part of a statistician's job is to *explain* their data. In this chapter, we'll show you some of the tools R offers to let you do so, with minimum fuss.

Mean 4.1

Determining the health of the crew is an important part of any inventory of the ship. Here's a vector containing the number of limbs each member has left, along with their names.

```
limbs <- c(4, 3, 4, 3, 2, 4, 4, 4)
names(limbs) <- c('One-Eye', 'Peg-Leg', 'Smitty', 'Hook', 'Scooter', 'Dan', 'Mikey', 'Blackbeard')
```

A quick way to assess our battle-readiness would be to get the average of the crew's appendage counts. Statisticians call this the "mean". Call the mean function with the `limbs` vector.

```
> mean(limbs)
[1] 3.5
```

An average closer to 4 would be nice, but this will have to do.

Here's a barplot of that vector:

```
> barplot(limbs)
```

If we draw a line on the plot representing the mean, we can easily compare the various values to the average. The `abline` function can take an `h` parameter with a value at which to draw a horizontal line, or a `v` parameter for a vertical line. When it's called, it updates the previous plot.

Draw a horizontal line across the plot at the mean:

```
> abline(h = mean(limbs))
```

Median 4.2

Let's say we gain a crew member that completely skews the mean.

```
> limbs <- c(4, 3, 4, 3, 2, 4, 4, 14)
> names(limbs) <- c('One-Eye', 'Peg-Leg', 'Smitty', 'Hook',
                    'Scooter', 'Dan', 'Mikey', 'Davy Jones')
> mean(limbs)
[1] 4.75
```

Let's see how this new mean shows up on our same graph.

```
> barplot(limbs)
> abline(h = mean(limbs))
```

It may be factually accurate to say that our crew has an average of 4.75 limbs, but it's probably also misleading.

For situations like this, it's probably more useful to talk about the "median" value. The median is calculated by sorting the values and choosing the middle one (for sets with an even number of values, the middle two values are averaged).

Call the median function on the vector:

```
> median(limbs)
[1] 4
```

That's more like it. Let's show the median on the plot. Draw a horizontal line across the plot at the median.

```
> abline(h = median(limbs))
```

Standard Deviation 4.3

Some of the plunder from our recent raids has been worth less than what we're used to. Here's a vector with the values of our latest hauls:

```
> pounds <- c(45000, 50000, 35000, 40000, 35000, 45000, 10000, 15000)
> barplot(pounds)
> meanValue <- mean(pounds)
```

Let's see a plot showing the mean value:

```
> abline(h = meanValue)
```

These results seem way below normal. The crew wants to make Smitty, who picked the last couple ships to waylay, walk the plank. But as he dangles over the water, wily Smitty raises a question: what, exactly, is a "normal" haul?

Statisticians use the concept of "standard deviation" from the mean to describe the range of typical values for a data set. For a group of numbers, it shows how much they typically vary from the average value. To calculate the standard deviation, you calculate the mean of the values, then subtract the mean from each number and square the result, then average those squares, and take the square root of that average.

If that sounds like a lot of work, don't worry. You're using R, and all you have to do is pass a vector to the sd function. Try calling sd on the pounds vector now, and assign the result to the deviation variable:

```
> deviation <- sd(pounds)
> deviation
[1] 14500.62
```

We'll add a line on the plot to show one standard deviation above the mean (the top of the normal range)...

```
> abline(h = meanValue + deviation)
```

Hail to the sailor that brought us that 50,000-pound payday!

Now try adding a line on the plot to show one standard devation below the mean (the bottom of the normal range):

```
> abline(h = meanValue - deviation)
```

We're risking being hanged by the Spanish for this? Sorry, Smitty, you're shark bait.

Chapter 4 Completed

Land ho! You've navigated Chapter 4. And what awaits us on the shore? It's another badge!

Summary statistics let you show how your data points are distributed, without the need to look closely at each one. We've shown you the functions for mean, median, and standard deviation, as well as ways to display them on your graphs.

Continue