Summary statistics

Put your name here

Put the date here

Introduction

In this module, we'll learn about summary statistics, which are numerical summaries calculated from data.

Instructions

Presumably, you have already created a new project and downloaded this file into it. Please knit the document and work back and forth between this R Markdown file and the PDF output as you work through this module.

When you are finished with the assignment, knit to PDF one last time, proofread the PDF file **carefully**, export the PDF file to your computer, and then submit your assignment.

Sometimes you will be asked to add your own R code. That will appear in this document as a code chunk with a request for you to add your own code, like so:

```
## Add code here to [do some task]...
```

Be sure to remove the line `## Add code here to [do some task]...` when you have added your own code.

Sometimes you will be asked to type up your thoughts. That will appear in the document as follows:

Please write up your answer here.

Again, please be sure to remove the line "Please write up your answer here" when you have written up your answer. In these areas of the assignment, please use contextually meaningful full sentences/paragraphs (unless otherwise indicated) and proper spelling, grammar, punctuation, etc. This is not R code, but rather a free response section where you talk about your analysis and conclusions. If you need to use some R code as well, you can use inline R code inside the block between `\begin{answer}` and `\end{answer}`, or if you need an R code chunk, please go outside the `answer` block and start a new code chunk.

Load Packages

We load the `mosaic` package as well as the `MASS` package so that we can work with data on risk factors associated with low birth weight.

```
library(mosaic)
library(MASS)
```

Mean and standard deviance

Summary statistics come in two general flavors: measures of center and measures of spread.

The first pair we'll consider is the mean and the standard deviation. The *mean*—denoted $\bar{y}$—of a variable y is calculated by summing all the values of the variable, and dividing by the total number of observations. In formula form, this is

$$\bar{y} = \frac{\sum y}{n}.$$

This is a measure of center since it estimates the "middle" of a set of numbers.

It is calculated in R using the `mean` command. For example, if we want to calculate the mean weight in pounds of the mother at the last menstrual period, we type the following:

```
mean(birthwt$lwt)
```

```
## [1] 129.8148
```

The corresponding measure of spread is the *standard deviation*. Usually this is called s and is calculated using a much more complicated formula:

$$s = \sqrt{\frac{\sum (y - \bar{y})^2}{n - 1}}.$$

This is a measure of spread because the $(y - \bar{y})$ term measures the how far away each data point is from the mean.

In R, this is calculated with the `sd` command.

```
sd(birthwt$lwt)
```

```
## [1] 30.57938
```

The mean and the standard deviation should always be reported together.

Another related measurement is the *variance*, but this is nothing more than the standard deviation squared:

$$s^2 = \frac{\sum (y - \bar{y})^2}{n - 1}.$$

(Compare this formula to the one for the standard deviation. Nothing has changed except for the removal of the square root.) We rarely use the variance in an introductory stats class because it's not as interpretable as the standard deviation.[1] If you need to do this in R, the command is `var`.

```
var(birthwt$lwt)
```

```
## [1] 935.0985
```

(You can check and see that the number above really is just 30.5793804 squared.)

Median and IQR

Another choice for measuring the center and spread of a data set is the median and the IQR. The median is just the middle value if the list of values is ordered. In R, it is calculated using the `median` command.

[1] The main reason for this is units. If the data variable is the mother's weight in pounds, then both the mean and the standard deviation are also reported in pounds. The variance has units of "pounds squared", but what does that even mean?

```
median(birthwt$lwt)
```

```
## [1] 121
```

You can check that this is correct: if we print out the entire `lwt` variable, you can see all 189 values, and if we're clever about it, we can see them in order.

```
sort(birthwt$lwt)
```

```
##   [1]  80  85  85  89  90  90  90  91  92  94  95  95  95  95  95  95  96
##  [18]  97  98 100 100 100 100 100 101 102 102 103 103 103 105 105 105 105
##  [35] 105 105 105 107 107 108 109 109 110 110 110 110 110 110 110 110 110
##  [52] 110 110 112 112 112 112 113 113 113 115 115 115 115 115 115 115 116
##  [69] 117 117 118 118 119 119 119 120 120 120 120 120 120 120 120 120 120
##  [86] 120 120 120 120 120 120 120 121 121 121 121 122 122 123 123 123 124
## [103] 124 125 125 125 127 128 128 129 130 130 130 130 130 130 130 130 130
## [120] 130 130 130 130 131 132 132 132 133 133 134 134 134 135 135 135 135
## [137] 137 138 138 140 140 140 141 142 142 147 147 148 150 150 150 150 150
## [154] 153 154 154 155 155 155 158 158 160 160 165 167 168 169 169 170 170
## [171] 170 170 175 182 184 185 186 187 187 189 190 190 200 202 215 229 235
## [188] 241 250
```

Exercise

If there are 189 mothers in this data set, in which position in the list will the median value appear? (Be careful: you can't just divide 189 by 2!) Verify that the median value 121 does appear in the position you calculated.

Please write up your answer here.

Calculating the *interquartile range*—or *IQR*—requires first the calculation of the first and third quartiles, denoted Q1 and Q3. If the median is the 50% mark in the sorted data, the first and third quartiles are the 25% and the 75% marks, respectively.[2] Then the IQR is just

$$IQR = Q3 - Q1$$

In R, you can get the IQR by using—are you ready for this?—the `IQR` command.

```
IQR(birthwt$lwt)
```

```
## [1] 30
```

The IQR is a measure of spread because the distance between Q1 and Q3 measures the span of the "middle 50%" of the data.

[2] One way to compute these is to calculate the median of the lower and upper halves of the data separately. Then again, it's hard to know how to split the data set into halves if there are an odd number of observations. There are many different methods for computing percentiles in general, but you don't need to worry too much about the particular implementation in R.

Exercise

A general function for computing any percentile in R is the `quantile` function. For example, since Q1 is the 25th percentile, you can compute it as follows:

```
Q1 <- quantile(birthwt$lwt, 0.25)
unname(Q1)
```

```
## [1] 110
```

(The `unname` command just helps clean up the output here. Don't worry too much about it.)

Now compute Q3. Also check that the IQR calculated above matches the value you get from subtracting Q3 minus Q1.

```
## Add code here to compute and print out Q3.
## You should use the unname command as in the above chunk.
```

```
## Add code here to compute Q3 - Q1
```

The median and the IQR should always be reported together.

Robust statistics

Some statistics are more sensitive than others to features of the data. For example, outliers are data points that are far away from the bulk of the data. The mean and especially the standard deviation can change a lot when outliers are present. Also, skewness in the data frequently pulls the mean too far in the direction of the skew while simultaneously inflating the standard deviation.

On the other hand, the median and IQR are "robust", meaning that they do not change much (or at all) in the presence of outliers and they tend to be good summaries even for skewed data.

Exercise

Explain why the median and IQR are robust. In other words, why does an outlier have little or no influence on the median and IQR?

Please write up your answer here.

Five-number summary

A *five-number summary* is the minimum, Q1, median, Q3, and maximum of a set of numbers.

The `summary` command in R gives you the five-number summary, and throws in the mean for good measure.

```
summary(birthwt$lwt)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    80.0   110.0   121.0   129.8   140.0   250.0
```

You can, of course, isolate the various pieces of this. You already know most of the commands below.

```
min(birthwt$lwt)
```

```
## [1] 80
```

```
median(birthwt$lwt)
```

```
## [1] 121
```

```
max(birthwt$lwt)
```

```
## [1] 250
```

```
mean(birthwt$lwt)
```

```
## [1] 129.8148
```

```
quantile(birthwt$lwt)
```

```
##   0%  25%  50%  75% 100%
##   80  110  121  140  250
```

Exercise

What is the difference between the way `quantile` was used in a previous exercise versus the way it was used here? How did that change the output?

Please write up your answer here.

The `mosaic` package also has a summary command called `favstats` that has a little more information, including the standard deviation, the sample size, and a count of any cases that are missing data. This command is much more useful than the standard `summary` command.

```
favstats(birthwt$lwt)
```

```
##  min  Q1 median  Q3 max     mean       sd   n missing
##   80 110    121 140 250 129.8148 30.57938 189       0
```

Also, don't forget about the trick for using R commands inline. If you need to mention a statistic in the middle of a sentence, there is no need to break the sentence and display a code chunk. Be sure you're looking at the R Markdown document to note that the numbers in the next sentence are not manually entered, but are calculated on the fly:

There are 189 births represented in this data and the median weight of the women as of their last menstrual period is 121 pounds.

Your turn

Type a full sentence using inline R code (as above) summarizing the minimum and maximum baby weights (in grams) in our data set.

Please write up your answer here.

Summary statistics by group

Using base R, it's not so easy to get summary statistics for each group separately. Fortunately, the `mosaic` package comes to the rescue, allowing for more flexibility. For example:

```
favstats( ~ lwt | race, data = birthwt)
```

```
##   race min  Q1 median     Q3 max     mean       sd  n missing
## 1    1  90 112  129.5 143.25 235 132.0521 29.09381 96       0
## 2    2  98 120  129.0 179.00 241 146.8077 39.63939 26       0
## 3    3  80 105  119.0 130.00 250 120.0149 25.13026 67       0
```

The notation is a little weird. Don't worry about the tilde for now. Just learn that it needs to be there. The important part is the `lwt | race`. This says, "Look at the numerical variable `lwt` broken down by `race`." Indeed, the output has three lines, one for each race in the data. This extended notation works for lots of commands, like `mean`, `sd`, `median`, `IQR`, `quantile`, etc. as long as the `mosaic` package is loaded.

Your turn

Choose one numerical variable and one categorical variable from the `birthwt` data set. Find the five-number summary of your numerical variable grouped by your categorical variable.

```
## Add code here to compute the five-number summary of your
## numerical variable grouped by your categorical variable.
```

Conclusion

Summary statistics are simple numbers that describe and summarize data sets. Measures of center tell us where the "middle" of our data lies, and measures of spread tell us how spread out our data is. These measures should always be reported in pairs, for example the mean/standard deviation, or the median/IQR. Sometimes it can be useful to report summary statistics with the data separated by a categorical grouping variable.