Calculating Simple Descriptives

Rachael Smyth

Load Libraries

Different "libraries" or "packages" are required to perform different analyses. You must install and load the package before you can run the functions that are a part of each package. To load packages, the command is library(x) where x is the name of the package you want to load. We will eventually use some functions from the `psych` package, and so we'll load that now. Except where noted, all of the functions in this presentation are from the basic packages that are loaded by default in R.

```
library(psych)
```

The Data

The data used for this demonstration looked at the effect of two different sleep aid drugs on increasing amount of sleep compared to controls. The dataset comes as part of R and is called "sleep". You can load this data into R to use in an ongoing fashion.

```
data(sleep)
```

Viewing the Data

The `head` function presents the variable names for all of columns in your dataset, as well as the first 6 subjects worth of data.

```
head(sleep)
```

```
##   extra group ID
## 1   0.7     1  1
## 2  -1.6     1  2
## 3  -0.2     1  3
## 4  -1.2     1  4
## 5  -0.1     1  5
## 6   3.4     1  6
```

Calculating Measures of Central Tendency

Mean

Intuitively enough, the `mean` function calculates the arithmetic mean of your data.

```
mean(sleep$extra)
```

```
## [1] 1.54
```

Note that we specified the `extra` variable in the `sleep` dataset, by using a dollar sign. Alternatively, we could have referred to the `extra` variable, using indices within the data frame. We want all of the rows in the dataset, and so we would leave the first part of this index specification blank - and we want only the first column, thus making the second part of this index specification "1".

```
mean(sleep[,1])
```

```
## [1] 1.54
```

Median

Likewise, the `median` function produces the median of your data.

```
median(sleep$extra)
```

```
## [1] 0.95
```

Mode

Curiously enough, there is no `mode` function in the basic R library. There is, however, a `max` function that can be used to evaluate the value that occurs most often within a set of data. In order to leverage this function to determine the mode(s) of your data, you must first generate a frequency table from your table, using the `table` function. Then, you can use the `max` function to determine which value(s) occur most often. Because there can be multiple modes within a dataset, this approach will generate a list of all of the values that occur "the most often", as well as the number of times that they occur.

```
extrasleep<-table(sleep$extra)
subset(extrasleep, extrasleep==max(extrasleep))
```

```
##
## -0.1  0.8  3.4
##    2    2    2
```

Calculating Measures of Dispersion

Range

The `range` function produces two values for your data. The first value is the bottom end of the range (the lowest number in your data set), and the second value is the top end of the range (the highest number in your data set).

```
range(sleep$extra)
```

```
## [1] -1.6  5.5
```

Or... you can use some R arithmetic to generate the distance that your variable spans:

```
range(sleep$extra)[2] - range(sleep$extra)[1]
```

```
## [1] 7.1
```

Standard Deviation

The `sd` function calculates the standard deviation of your data. This is the "unbiased estimate" of standard deviation (sometimes called the sample standard deviation), and so it uses a denominator of **n - 1**.

```
sd(sleep$extra)
```

```
## [1] 2.01792
```

Coefficient of Variation

There is no coefficient of variation function in the basic R packages (i.e., the ones that are loaded by default). Although it is possible that there are other packages that may have such a function (in fact, there is a coefficient of variation function, `cv` in the `raster` package), this is such a simple calculation that it would be more informative to illustrate the writing of a simple function.

```
cv <- function(variable) { sd(variable) / mean(variable) }
```

Now, we can simply call this function, and it will automatically calculate the standard deviation of the variable that we specify, and divide it by the variable mean.

```
cv(sleep$extra)
```

```
## [1] 1.310337
```

Obviously, this wouldn't be a huge time-saver if we were only calculating one coefficient of variation, but it might be worthwhile if we were performing the same calculation numerous times.

Generating a "Full Set" of Descriptives

Rather than generating individual descriptive statistics (i.e., mean, standard deviation), there are a few methods that you can employ to look at the overall descriptives of your data. The `summary` function provides you with some useful descriptive information about each variable (mean, median, min, max), but does not provide you with the standard deviation, or any distributional properties (e.g., skewness, kurtosis, etc.).

```
summary(sleep)
```

```
##      extra        group        ID
##  Min.   :-1.600   1:10   1      :2
##  1st Qu.:-0.025   2:10   2      :2
##  Median : 0.950          3      :2
##  Mean   : 1.540          4      :2
##  3rd Qu.: 3.400          5      :2
##  Max.   : 5.500          6      :2
##                          (Other):8
```

A better option for a "full set" of descriptives is the `describe` function from the `psych` package. This function provides you with the sample size, mean, SD, median, min, max, range, skewness and kurtosis.

Because this function is part of the `psych` package, you need to load the `psych` package before running the function.

```
describe(sleep$extra)
```

```
##   vars  n mean   sd median trimmed  mad  min max range skew kurtosis   se
## 1    1 20 1.54 2.02   0.95    1.47 1.56 -1.6 5.5   7.1 0.39    -1.09 0.45
```

Generating a Set of Descriptives for Each Group

You can also use a very similar command (also from the `psych` package) to calculate these descriptives separately across groups. Because the `sleep` dataset contains information that allows you to differentiate which of two soporific drugs that participants took, we can use this information to calculate descriptives for each of the experimental groups in our study.

```
describeBy(sleep$extra, sleep$group)
```

```
## group: 1
##   vars  n mean   sd median trimmed  mad  min max range skew kurtosis   se
## 1    1 10 0.75 1.79   0.35    0.68 1.56 -1.6 3.7   5.3 0.42     -1.3 0.57
## ------------------------------------------------------------
## group: 2
##   vars  n mean sd median trimmed  mad  min max range skew kurtosis   se
## 1    1 10 2.33  2   1.75    2.24 2.45 -0.1 5.5   5.6 0.28    -1.66 0.63
```