Foundations of Data Analysis

The University of Texas at Austin

R Tutorials: Week 2

Univariate Descriptive Statistics

In this R tutorial, we're going to learn how to run various descriptive statistics on a single numerical variable. So just like in the histogram tutorial videos, we're going to be using the animal shelter data set. And we're going to be specifically looking at the age-at-intake variable, which is how old these animals were when they arrived at the shelter. So there are a couple different measures of the center of a numeric distribution that we're going to learn how to calculate. The first one we are already familiar with – the function – from previous videos. And it's the "**mean()**" function. So again, you can calculate the mean of a vector of numbers by just typing the word mean, open parentheses, and then giving it a data frame, dollar sign, variable name that you'd want to calculate the mean for.

```
mean(animaldata$Age.Intake)
```

```
## [1] 2.348837
```

And we can see that the mean age of animals entering the shelter is about 2.3 years. The other measure of center that we talk about typically is the median, which is the halfway point, or where 50% of the data fall below it and 50% fall above it. It works just like the mean function, and it's just called "**median()**". We can again just give it our age-at-intake variable.

```
median(animaldata$Age.Intake)
```

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

And we can see that the median age of animals at the shelter when they arrive is only one year old.

So along with the measure of center, we typically want to talk about the spread, or the amount of variability in a variable. And there are also a couple ways to describe that. Usually when we report a mean, we also report the standard deviation. And the function for that in R is just "**sd()**". And again, it takes one argument, which is the variable.

```
sd(animaldata$Age.Intake)
```

```
## [1] 3.099837
```

So we can see that the standard deviation of the age-at-intake variable is 3.1 about.

Another way to describe the amount of spread in a variable is to report the five-number summary, which consists of the minimum, the maximum, the median or halfway point, and then also the first and third quartiles, which are what you use to calculate the interquartile range, which is sometimes reported as a measure of spread. R also has a really simple function that will give you the five-number summary, and it's just called "**fivenum()**".

```
fivenum(animaldata$Age.Intake)
```

```
## [1]  0  0  1  3 17
```

If we ask for fivenum of our age-at-intake variable, we're going to see the output of the five numbers in our five-number summary, starting with the min, then the first quartile, then the median which we see is 1 again, the third quartile which is 3, and then finally the maximum age which is 17.

So with just a few simple functions we can pretty fully describe the single numerical variable age-at-intake by either reporting the mean and standard deviation together or the median and including the five-number summary.