Lab 2: Descriptive statistics

STAT218

Let's explore how some of the other descriptive statistics we've discussed behave in response to outliers. Specifically, measures of spread: standard deviation and IQR.

The variable you just looked at — dominant arm percent change in strength — has a group of observations at 100%. If these are removed, the standard deviation increases by 10%, but the IQR only increases by 5%.

```
# extract dominant arm percent change in strength
drm <- famuss$drm.ch

# drop the observations over 80%
drm.drop <- drm[drm < 100]

# compute the numeric summary with and without outliers
summary(drm)
summary(drm.drop)

# compare standard deviations
sd(drm)/sd(drm.drop)

# compare interquartile ranges
IQR(drm)/IQR(drm.drop)
```

This may not seem very notable, so let's make up an example that's a bit more extreme: let's add a very large positive observation, say, 1000. Then, the IQR does not change at all, but the standard deviation more than doubles!

```
# add a large observation
drm.add <- c(drm, 1000)

# compare IQR with and without
IQR(drm.add)/IQR(drm)
```

```r
# compare SD with and without
sd(drm.add)/sd(drm)
```

The differences in robustness between IQR and standard deviation, and between mean and median, are largely why *both* the five-number summary *and* the mean and standard deviation are reported. When these statistics differ dramatically, it is most likely due to the presence of outliers!

Your turn

Compute the numeric summary for a variable from a different dataset and *based on this alone* attempt a guess at whether there are outliers. If so, are they more likely outliers to the left or right?

```r
# load a new dataset (census)
data(census.2010)

# number of doctors per state (thousands)
doctors <- census.2010$doctors

# compute numeric summary -- guess whether there are outliers?

# make a histogram or boxplot to confirm your guess

# bonus: which state??
```