Lab 3: More descriptive statistics

STAT218

This lab covers two separate topics: measures of spread and bivariate graphical summaries. There are two goals for the activity:

- learn to calculate measures of spread (IQR and standard deviation) and explore their robustness to outliers
- learn to produce joint summaries of two variables for identifying relationships
  - contingency tables
  - proportional barplots
  - scatterplots
  - side-by-side boxplots

We will use the FAMuSS dataset again.

```
# openintro biostat package
library(oibiostat)

# famuss data
data(famuss)
```

Robustness and measures of spread

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.

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

# calculate standard deviation
sd(drm)

# interquartile range
```

```r
IQR(drm)

# average deviation
mean(abs(drm - mean(drm)))

# range
max(drm) - min(drm)

# range endpoints
range(drm)
```

The variable you just looked at — dominant arm percent change in strength — has a group of observations over 60%.

```r
# boxplot of percent change in dominant arm strength
boxplot(drm, horizontal = T, range = 2)
```

If these are removed, the standard deviation increases by 24%, but the IQR only increases by 18%.

```r
# drop the observations over 60%
drm.drop <- drm[drm < 60]

# 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!

```r
# 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!

> **i** 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: can you figure out which state??
> ```

Bivariate graphics

This part of the lab is organized according to which types of variables are being compared as potentially related. In each sub-part, you'll see a series of examples that illustrate how to produce a given graphic or other summary, and then you'll have an opportunity to try it with a different pair of variables from the FAMuSS study data.

Categorical/categorical

Consider: *is there differential expression of the ACTN gene region of interest between sexes?*

This can be answered by comparing the proportions of study participants of each genotype by sex. The steps are:

1. Start by making a contingency table

2. Convert to proportions using the appropriate row/column sums
3. Visualize and compare genotype composition by group

The examples below illustrate how to perform these steps. As you're walking through them, consider which summary answers the question — do you want to compute proportions using the genotype totals or the sex totals?

```
# retrieve the genotype and sex columns
genotype <- famuss$actn3.r577x
sex <- famuss$sex

# construct a contingency table
table(genotype, sex)

# stacked bar plots -- automatically groups by column
tbl <- table(famuss$actn3.r577x, famuss$sex)
barplot(tbl, legend = T)

# turn the table on its side with t() to group by row
barplot(t(tbl), legend = T)

# row and column margins
rowSums(tbl)
colSums(tbl)

# proportions, grouping by row
tbl_row <- tbl/rowSums(tbl)
tbl_row

# proportional stacked bar plot, grouped by row
barplot(t(tbl_row), legend = T)

# proportions, grouping by column (a little trickier)
tbl_col <- t(t(tbl)/colSums(tbl))
tbl_col

# proportional stacked bar plot, groupde by column
barplot(tbl_col, legend = T, horiz = T)
```

Your turn

Is there differential expression of the ACTN gene region by racial group?
Follow the examples above to make a contingency table and bar plot of genotype composition for each racial group. Do you see differences?

```
# retrieve the genotype and race columns

# make a contingency table of genotype and race

# are there apparent genotype differences by race? make an appropriate bar plot
```

Numeric/numeric

Taller people tend to be heavier. We should expect a relationship between weight and height. The example below shows a scatterplot of the two variables, and computes the correlation (measure of linear relationship).

```
# retrieve height and weight columns
height <- famuss$height
weight <- famuss$weight

# basic scatterplot
plot(height, weight)

# correlation
cor(weight, height)
```

A rough rule of thumb for interpreting correlations is as follows:

- $|r| < 0.3$: no relationship
- $0.3 \leq |r| < 0.6$: weak to moderate relationship
- $0.6 \leq |r| < 1$: moderate to strong relationship

In this case, the correlation of 0.53 indicates a moderate positive relationship.

Your turn

Is there a relationship between nondominant and dominant percent change in arm strength? Make a scatterplot and compute the correlation.

```
# retrieve the percent change variables

# construct a scatterplot

# compute the correlation
```

Categorical/numeric

Consider one of the main questions for the study:

> *Were differences on the ACTN gene region associated with differential change in arm strength after resistance training?*

This is a comparison between a categorical variable (genotype) and numeric variable (percent change in arm strength). Of course, we have measurements for both dominant and non-dominant arms; while there are other ways of handling this, we'll just make comparisons separately for each arm.

The examples below produce boxplots for a quick comparison of the summary statistics of percent change in arm strength between genotypes. Recall that the summary statistics are summarizing the frequency distribution.

```
# side-by-side boxplots for non-dominant arm
boxplot(ndrm.ch ~ actn3.r577x, data = famuss)

# change the orientation
boxplot(ndrm.ch ~ actn3.r577x, data = famuss, horizontal = T)

# change the whisker length (range = multiples of IQR)
boxplot(ndrm.ch ~ actn3.r577x, data = famuss, horizontal = T, range = 2)

# side-by-side boxplots for dominant arm
boxplot(drm.ch ~ actn3.r577x, data = famuss, horizontal = T)
```

There are some slight observed differences for the non-dominant arm, but it's unclear whether they are meaningful. We'll return to that later, but for now, try the graphical technique with a different set of variables.

Your turn

Investigate whether BMI seems to differ by racial group among the FAMuSS study participants.

```
# make side-by-side boxplots of BMI by race
```