Summarizing Numerical Data

Stat 145

Student Survey

`Lock5withR` includes a data set with data from a student survey. It includes the following variables.

- `Year` Year in school
- `Gender` Student's gender: 'F' or 'M'
- `Smoke` Smoker? 'No' or 'Yes'
- `Award` Prefered award: 'Academy' 'Nobel' 'Olympic'
- `HigherSAT` Which SAT is higher? 'Math' or 'Verbal'
- `Exercise` Hours of exercsie per week
- `TV` Hours of TV viewing per week
- `Height` Height (in inches)
- `Weight` Weight (in pounds)
- `Siblings` Number of siblings
- `BirthOrder` Birth order, 1=oldest
- `VerbalSAT` Verbal SAT score
- `MathSAT` Math SAT score
- `SAT` Combined Verbal + Math SAT
- `GPA` College grade point average
- `Pulse` Pulse rate (beats per minute)
- `Piercings` Number of body piercings

```
names(StudentSurvey)
```

```
##  [1] "Year"       "Gender"     "Smoke"      "Award"      "HigherSAT"
##  [6] "Exercise"   "TV"         "Height"     "Weight"     "Siblings"
## [11] "BirthOrder" "VerbalSAT"  "MathSAT"    "SAT"        "GPA"
## [16] "Pulse"      "Piercings"  "Sex"
```

```
nrow(StudentSurvey)
```

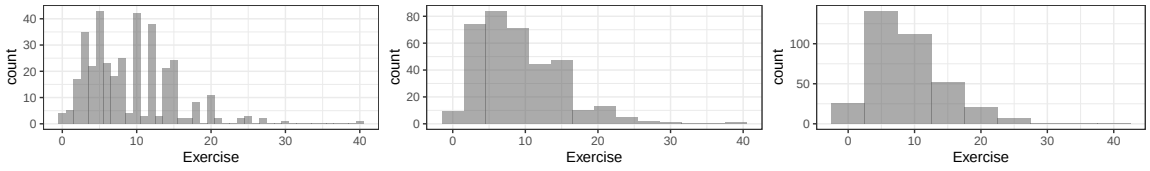
```
## [1] 362
```

Question: Which variables are categorical, which quantitative?

Exercise

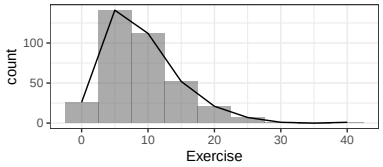
Let's take a look at one of the variables: `Exercise`. Here are some histograms of the `Exercise` variable.

```
gf_histogram( ~ Exercise, data = StudentSurvey, binwidth = 1, alpha = 0.5)
gf_histogram( ~ Exercise, data = StudentSurvey, binwidth = 3, alpha = 0.5)
gf_histogram( ~ Exercise, data = StudentSurvey, binwidth = 5, alpha = 0.5)
```



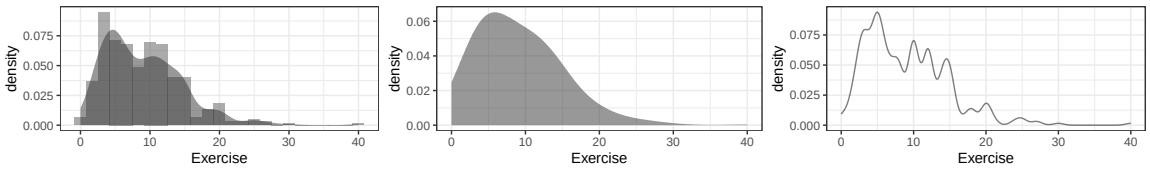
1. Which binwidth do you like best? Why?
2. What does the tallest bar in the histogram with binwidth 5 represent?
   (Be as specific as you can given the resolution of the plot.)
3. Why is there a gap with no bars near the right end of the graph? What does that represent?
4. How would you describe the shape of the histogram(s)?
5. Use the histograms to estimate the mean and the median. Which one should be larger? Why? A lot larger or just a little larger? Now compute the mean and median and see if you are right.
6. Sketch what you think a boxplot for this data set looks like. Then create one and see how your sketch compares.
7. Use your boxplot to estimate the IQR. Then compute the IQR to see how close your estimate is.
8. An alternative to a histogram is called a **frequency polygon**.
   a. Look at the example below and describe in words how a frequency polygon compares to a histogram?
   b. Why might someone prefer a frequency polygon to a histogram?

```
gf_histogram( ~ Exercise, data = StudentSurvey, alpha = 0.5, binwidth = 5) %>%
 gf_freqpoly( ~ Exercise, data = StudentSurvey, binwidth = 5)
```



8. This plots below are "smooth versions" of a histogram. What the advantages and disadvantages of these density plots vs. histograms?

```
gf_dhistogram( ~ Exercise, data = StudentSurvey, alpha = 0.5) %>%
gf_density( ~ Exercise, data = StudentSurvey)
gf_density( ~ Exercise, data = StudentSurvey, adjust = 2)  # twice as smooth
gf_dens( ~ Exercise, data = StudentSurvey, adjust = 0.5)   # half as smooth (& "open")
```



9. How would you expect the shape to change if you made a histogram for `log(Exercise)`? Try it and see if you are right. What happens if you use `log10()` instead of `log()`? [`log()` is natural log and `log10()` is log base 10. You can also use `log2()` for log base 2.]
10. Create several plots to compare `Exercise` for men and women. Describe what your plots tell you.
11. Create several plots to compare `Exercise` for students in different academic years. Describe what your plots tell you.

Finished? Try looking at some other quantitative variables, like `Piercings` or `Pulse`. Make some plots and see what they tell you about these variables.