STAT 630, HW 3
Due: Thursday, September 16

Reading: OpenIntro, Sections 2.1 and 2.2

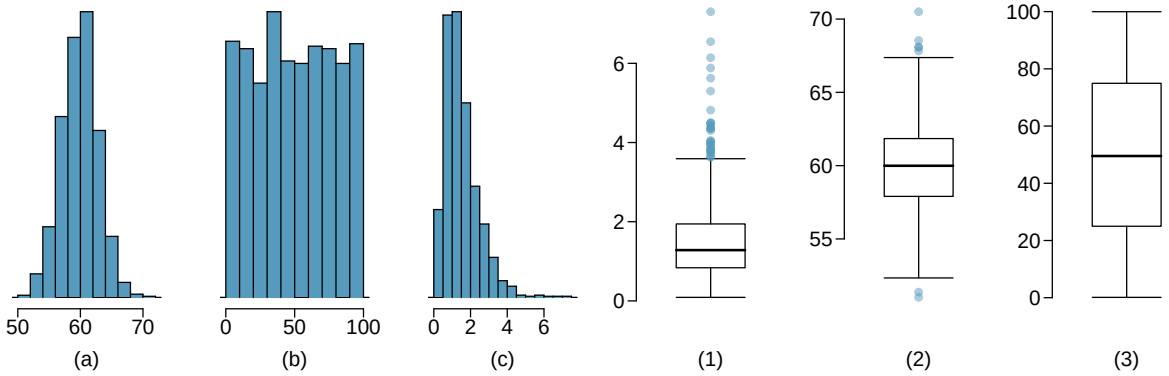
Directions: Please submit your completed assignment to Blackboard. For the concept questions, your solutions may be typed, or handwritten and then scanned. The data analysis questions should be completed using R Markdown and then rendered to HTML, PDF, or Word format.

Concept Questions

Exercise 1. An athlete preparing for a bicycle competition records the number of miles spent biking each day over a month long period. The sample mean $\bar{x} = 15.4$ miles and standard deviation $s = 6.2$ miles. What are the sample mean and standard deviation when converting distance to kilometers (km). Note that the conversion formula is 1 mile = 1.6 km.

Exercise 2. Test scores for a math class with 30 students have a sample mean $\bar{x} = 84.5$ points (out of 100) and a standard deviation $s = 7.2$ points. The teacher decides to give every student an extra 5 points for redoing questions that they missed. What are the sample mean and standard deviation after adding 5 points to each exam score?

Exercise 3. Describe the distribution in the histograms below and match them to the box plots.



Exercise 4. Without doing any calculations determine whether the standard deviation of Set 1 is larger than, smaller than, or equal to Set 2. Give a brief explanation.

(a) Set 1: $-1, 0, 0, 2, 2, 5, 7, 9, 8, 100$
Set 2: $-1, 0, 0, 2, 2, 5, 7, 9, 8, 101$

(b) Set 1: $-1, 0, 0, 2, 2, 5, 7, 9, 8, 100$
Set 2: $1, 2, 2, 4, 4, 7, 9, 11, 10, 102$

(c) Set 1: $-1, 0, 0, 2, 2, 5, 7, 9, 8, 100$
Set 2: $-2, 0, 0, 4, 4, 10, 14, 18, 16, 200$

Exercise 5. Compute the mean $\bar{x}$ and median m of the seven numbers 3, 5, 8, 15, 20, 21, 24. Apply the logarithm to the data, and then compute the mean $\bar{x}'$ and median m' of the transformed data. Is $\ln(\bar{x}) = \bar{x}'$? Is $\ln(m) = m'$? You can use R to answer this.

Data Analysis Questions

Exercises 6 and 7 use the CDC data set. Run the following command to load this data set into R:

```
cdc <- readRDS(url("https://ericwfox.github.io/data/cdc.rds"))
```

Exercise 6.

(a) Make a contingency table for the `exerany` and `genhlth` variables (the table should have 2 rows and 5 columns). Recall that the `exerany` variable is 1 if the respondent exercised in the last month, and 0 otherwise. Make sure that the levels of `genhlth` are ordered correctly by entering the command:

```
cdc$genhlth <- factor(cdc$genhlth,
  levels=c("poor", "fair", "good", "very good", "excellent"))
```

(b) Make a stacked bar plot of the contingency table in created in part (a). The plot should have the categories of `genhlth` along the horizontal axis. The segments for each bar should correspond to the `exerany` variable. Add a legend to the plot. You can use either the base R or ggplot2 approach.

(c) Use `prop.table()` to make a contingency table with the column proportions for the `exerany` and `genhlth` variables (the table should have 2 rows and 5 columns, and the columns should sum to 1).

(d) Make a stacked bar plot of the contingency table of column proportions created in part (c). Comment on any trend you notice in the plot.

Exercise 7. Make a new variable called `wtdiff` which is the difference between each person's desired weight, `wtdesire`, and actual weight, `weight` (that is, desired weight - actual weight). Plot a histogram and compute summary statistics for `wtdiff`. Describe the distribution in terms of its center, shape, and spread. Feel free to adjust the number of bins in the histogram.

Exercise 8.

For this exercise, load the `maps` package and read in the EPA stream (NRSA) data set into R:

```
library(maps)
nrsa <- readRDS(url("https://ericwfox.github.io/data/nrsa.rds"))
```

(a) Make a map of the US and superimpose those stream sites that are in `Good` condition. Color the points blue.

(b) Make another map of the US and superimpose those stream sites that are in `Poor` condition. Color the points red.