STAT 630, HW 2
Due: Thursday, September 9

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

Please refer to the lecture 2 notes.

Exercise 1. Suppose {a,b,c,d,e} is a population of size $N = 5$. Write out all possible samples of size $n = 3$ from this population. For simple random sampling, what is the probability of selecting each sample of size $n = 3$?

Exercise 2. Suppose a population consists of $N = 40$ individuals. How many possible samples of size $n = 5$ can we select from this population? Assume that sampling is done without replacement (i.e., once an individual is selected for the sample, that individual cannot be selected again).

Exercise 3. A college with 3000 students enrolled is interested in surveying its students about a change in administrative policy. In each of the following descriptions of the method of selecting students for the survey, identify the type of sampling method used (SRS, stratified, cluster, or systematic).

(a) Randomly sample 100 students, where each student at this college has the same chance of being included in the sample.

(b) There are 100 different classes that are currently in session. 10 classes are randomly selected, and every student attending each of those 10 classes is included in the sample.

(c) Every 25th student is selected from a roster that lists the names and IDs of all students attending the college.

(d) The students are divided up based on class level (freshman, sophomore, junior, senior). A simple random sample of 50 students is taken from each class level.

Exercise 4. Discuss any sources of bias in the following sampling scenarios:

(a) An administrator is interested in the number of hours students spend studying. The administrator proceeds to randomly ask 20 students at the school's library how long they study.

(b) A polling agency is interested in predicting the percentage of voters that support a certain candidate. They call or email 1500 random voters. Only 200 voters actually respond to their phone calls and emails.

Exercise 5. Consider a population of 1000 voters: 400 are Democrats, and the rest are Republicans. If you randomly interview 10 voters from this population, what is the probability that exactly 4 are Democrats?

Data Analysis and R Questions

The following exercises use the CDC data set discussed in lab 2. Run the following command to load this data set into your R workspace:

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

Exercise 6. Use `plot()` to make a scatterplot with `weight` on the x-axis, and `wtdesire` on the y-axis. Label the x-axis "Weight" and the y-axis "Desired Weight". Superimpose a 1-1 line on your scatterplot by entering the command `abline(0,1)` after creating the plot. Write a couple sentences describing the association between the two variables.

Exercise 7. Based on the scatterplot created in the previous exercise, you should notice two outliers. Use the `subset()` function to identify the outliers by extracting the two rows of the `cdc` data frame corresponding to respondents with desired weights above 500 pounds. What are the actual weights of these two respondents?
Next, create a new data frame called `cdc2` that has the outliers removed (Hint: use `subset()` again to do this). Then make another scatterplot with `weight` on the x-axis, and `wtdesire` on the y-axis, but this time with the outliers removed.

Exercise 8. Create a new data frame that contains the subset of respondents who are male *and* have exercised in the last month. Use the `summary()` function to compute summary statistics for the weight and desired weight of this subset of respondents.

Exercise 9. Use the `table()` function to make a contingency table between the general health, `genhlth`, and exercise, `exerany`, variables. Use `addmargins()` to include the row and column totals for this table. What proportion of respondents who reported to be in excellent health exercised in the past month? What proportion of respondents who reported to be in poor health exercised in the past month?