HOMEWORK 1

Due Feb 1st, start of class

Your first homework covers Chapters 1.1, 1.3, 1.10-1.13, 2.4.3, and 2.5. You should be working on your homework throughout these first two weeks. If you can't solve some of the problems, please come to office hours. Email is fine only for very short questions.

THEORETICAL PORTION

The theoretical problems should be **neatly** numbered, written out, and solved. Do not turn in messy work.

1. Say your research deals with social networks. Your first step is to study the properties of the Facebook network of college students at CU-Boulder campus. The next step is to compare your findings to the national college student Facebook network.
   (a) What are the populations you are concerned with?
   (b) What is the relationship between these populations?
   (c) What are some of the characteristics of the networks you might consider? Pick three as an example.
   (d) If you had infinite time and resources, would you be able to measure these characteristics for every member of these populations?
   (e) Say you don't have infinite time and resources – how would you go about estimating those population characteristics?

2. You're working for a US public health surveillance team, keeping an eye on infectious diseases such as flu in the US.
   (a) If your goal is to estimate the average yearly infection rate of flu among those over 65 years of age in the US, what is the population you would like to be working with?
   (b) Given that surveillance is done only via doctor's offices, what is the actual sample of people whose infection rates you'll be observing?
   (c) What kind of estimates will you get? Can they be generalized to the entire population you'd like to be working with? Under what assumptions the answer is yes?

3. (a) What is the difference between mean, median, and mode? When would you prefer to use one and not the others?
   (b) What is the difference between standard deviation and range? When would you report one and not the other to communicate how variable the data are?

4. One out of 50 people has disease D. You're a doctor and suspect that one of your patients may have that disease. You order a diagnostic test T for that patient. The test is not perfect - it accurately claims the disease is "present" in 85% of the patients who actually have it, and accurately declares the disease as "absent" in 60% of the patients who indeed don't have the disease.
   (a) What is the probability that the test result comes back negative (the test says "absence of disease")? Does this mean your patient definitely does not have the disease?
   (b) If the test is negative, how likely is he to actually have disease D?
   (c) How about if the test comes back positive (the test says "disease")?
   (d) What do you think about the accuracy of this test?

5. Below is a frequency histogram of a sample of 250 error measurements from a calibration tool. Use this histogram and the table of counts in each bin to answer the following questions:
   (a) Convert the frequency histogram to a density histogram using the same bin breaks. Draw the histogram and report the height of the bars for each bin in a table.

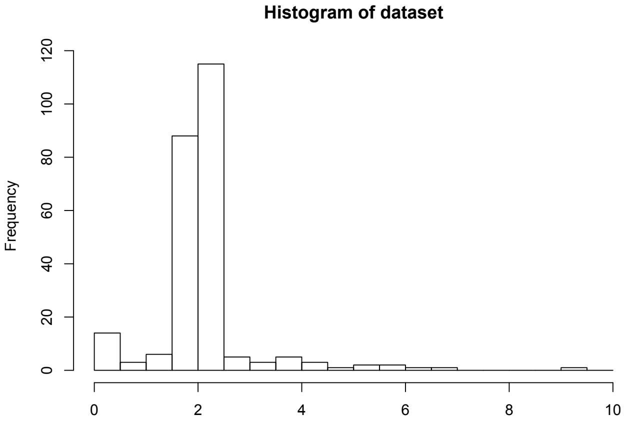

Figure 1: Frequency histogram of the error measurements from a calibration tool.

Table 1: Counts in each bin for the histogram

Bin	0.0-0.5-	0.5-1.0-	1.0-1.5-	1.5-2.0-	2.0-2.5-	2.5-3.0-	3.0-3.5-	3.5-4.0-	4.0-4.5-	4.5-5.0-
Count	14	3	6	88	115	5	3	5	3	1

Bin	5.0-5.5-	5.5-6.0-	6.0-6.5-	6.5-7.0-	7.0-7.5-	7.5-8.0-	8.0-8.5-	8.5-9.0-	9.0-9.5-	9.5-10.0-
Count	2	2	1	1	0	0	0	0	1	0

(b) Draw an approximate boxplot based on the histogram. Label the values of the median, quartiles, and IQR bounds.

COMPUTATIONAL PORTION

The computational portion of your homework should be neatly done and include all graphs, code, and comments, labeled and in order based on the problem you are addressing. Do *not* put graphs in at the end, stick code in random locations, or do anything else that will make this homework difficult to read and grade. **LABELS ARE YOUR FRIEND, USE THEM.** If you turn in something that is messy or out of order, it will be returned to you with a zero. All computations should be done using R, which can be downloaded for free at https://cran.r-project.org/. This is your first opportunity to get familiar with R, so please take your time on the problems that require it.

1. Simulations to solve a conditional probability problem: A bread factory uses 4 machines to produce its bread. Machine A produces 20% of the factory's bread, machine B produces 40% of the factory's bread, machine C produces 10% of the factory's bread, and machine D produces 30% of the factory's bread.

   Of the loaves produced by machine A, 20% are "bad", while 10% of those from machine B are bad, 5% of those from machine C are bad, and 50% of those from machine D are bad.

   Lastly, among the loaves that are "bad" and produced by machine A, 75% can be sold for 1/2 price, while 50% of the bad loaves from machine B can be sold at 1/2 price, 90% of the bad loaves from machine C can be sold at 1/2 price, and 60% of the bad loaves from machine D can be sold for half price. The other loaves cannot be sold at all.

   This problem can be solved analytically. However, I want you to use simulations to answer the following questions (be sure to justify how you decided what "enough" simulations was):

   (a) What is the probability that the factory produces a bad loaf?

(b) If a loaf is bad, what is the probability it came from machine B?

(c) If a loaf is not bad, what is the probability it came from machine C?

(d) What is the (approximate) probability that machine A produces a loaf that will be sold for 1/2 price?

(e) If a loaf is being sold for 1/2 price, what is the probability it came from machine D?

(f) **APPM 5570 students only:** The loaves produced by this factory cost the factory \$0.50 to make and are sold for \$4.00 when sold at full price. What *profit* can the factory expect after producing 1000 loaves of bread?

2. Download the data set labeled "hw1data.txt", which contains 2 datasets (one in each column), each of sample size 700.

   (a) Create a boxplot in R of the first column of data. Are there any particularly interesting features in the data that you can observe in the boxplot?

   (b) Create a histogram in R of the first column of data, making sure there are at least 40 bins in the histogram. What do you notice now?

   (c) Now repeat parts (a) and (b) for the second column of data. What do you notice now?

   (d) Do you think histograms or boxplots are better? Why?

3. There is an urn that has 100 marbles in it: 1 is red, 74 are yellow, 10 are green and 15 are blue. Simulate data to imitate someone drawing a marble from the urn (replacing after each draw), with the color recorded at each draw.

   (a) Plot the fraction of times a yellow marble is drawn in the first 500 draws. On the x-axis, include the number of draws, and on the y-axis, include the cumulative fraction of yellow draws. Why does this number fluctuate more when the number of draws is smaller?

   (b) How many draws are needed before the fraction of yellow marbles drawn in the simulated data is close to the probability of drawing a yellow marble? *You need to justify how you are defining "close" and what the optimal number of draws is.*

   (c) Repeat the two steps above for the red marble. What do you notice that is different, and why do you think it is different?