9.07
Introduction to Statistical Methods
Homework 1

Name: __________________________________

1. **Spurious or illusory correlations.**
A spurious correlation occurs when two variables are statistically related (correlated), yet there is no causal relationship between the two variables. This is usually because the relationship between the two variables is caused by a third variable.

An extreme example: A researcher collects data on all fires in Boston over the last 10 years. They find that the number of fire engines at each fire is highly correlated with the damages, in dollars, at each fire.

Should they conclude that fire engines cause the damage? What is an alternative explanation? What might be a hidden or "lurking" third variable?

2. A researcher looks at two groups of high school seniors at a local high school – one that had scores in the top 5% on their SAT's, and another that scored nearer the national average. The researcher finds that the high scorers have much larger wardrobes (i.e. they own lots more clothes) than the students with lower scores. He concludes that dressing well is important for success in life.

a. What is the "sample," in this example? What is a "population" to which the researcher might hope to generalize?

b. What is an alternative explanation for these results? If there is a lurking third variable in this explanation, what is it?

c. In looking to see whether size of wardrobe had an effect on SAT scores, what two groups did the researcher compare? Why might the students with smaller wardrobes not be the right comparison group for the students with bigger wardrobes? (Hint: might the students with smaller wardrobes be disadvantaged in some way that would also cause them to do worse on the SAT?)

d. What would be better comparison groups to test the hypothesis that bigger wardrobe size caused higher SAT scores? Describe a better experiment to test the hypothesis.

3. Compute the average, median, and mode for the following numbers:

5 4 4 5 4 9 6 1 3 8 2 5 7 8 9 1 4 5 3 8 8
2 7 6 8

4. Plot, by hand, a histogram for the numbers in question 3. Show the results below. Plot a histogram of these numbers using MATLAB. Show the results. If A is the vector of numbers, hist(A) will probably look different from the histogram you plotted by hand. However, if you specify which bins to use, with hist(A, 1:9), then the MATLAB histogram should look like yours. Show this.

5. p. 64 of your book shows three histograms – one with a long right-hand tail, one symmetric, and one with a long left-hand tail. The book gives the rule of thumb that for histograms with a long right-hand tail, the average is larger than the median. Give an intuition for why this is so.

6. Compute the standard deviation of the numbers in problem 3 (show intermediate steps, and don't just use MATLAB's standard deviation function). What fraction of the numbers are within +/- one standard deviation of the mean?

7. Generating random numbers with MATLAB's randn and rand functions. randn samples from a normal distribution with mean 0 and standard deviation 1, known as the standard normal distribution. rand samples from a uniform distribution between 0 and 1. Show histograms for the following cases. In each case, try it several times, to see how much the histograms can vary from trial to trial.

9.07
Introduction to Statistical Methods
Homework 1

Name: ___________________________________

a. 50 samples from a standard normal distribution.

b. 100 samples from a standard normal distribution.

c. 1000 samples from a standard normal distribution.

d. 1000 samples from a normal distribution with mean 5 and standard deviation 3 (Hint: if you have numbers from a normal distribution with standard deviation 1 and multiply them by s, you get numbers from a normal distribution with standard deviation s. If you add m to those numbers, you get numbers from a normal distribution with mean m.)

e. 50 samples from a uniform distribution.

f. 100 samples from a uniform distribution.

g. 1000 samples from a uniform distribution.

What do you notice about the relationship between the number of samples and how much the distribution looks like the theoretical “population” distribution? Between the number of samples and how much the distribution varies from trial to trial?

8. Generate random coin flips in MATLAB. You can do this by generating a bunch of random numbers with rand. These numbers will then vary from 0 to 1. Threshold the numbers so that any number below threshold is a head, and any number above is a tail. Where should the threshold be for a fair coin? Test this by generating a large number of coin flips, and confirming that you get roughly ½ heads. What would the threshold be for an unfair coin which has a 75% chance of coming up tails?

100 times, toss 3 coins. Show the histogram for all combinations of heads and tails (3 heads, 3 tails, 2 heads and a tail, 2 tails and a head). Show the histogram if you had tossed the trio of coins 1000 times.

9.07
Introduction to Statistical Methods
Homework 1

Name: ________________________________

9. Download the file DataHW1.mat from the assignment section. (under the materials for 9.07). Load it into MATLAB by typing "load DataHW1.mat". Type "who" and you will see a new variable called DataHW1. Plot a histogram of this data with 10 bins, and with 100 bins. This is data sampled from a bimodal distribution. Plot a frequency polygon of this data, using 20 bins. (Hint: use hist to output the bin centers and the number of points in each bin, and then use plot to plot the frequency polygon.)

10. Jurassic Park errors.

In his novel, "Jurassic Park," Michael Crichton shows the following histogram (frequency polygon) of the number of procompsognathid dinosaurs in the park of each height (height binned to the nearest cm).

Image removed due to copyright considerations. Please see histogram charting the height distribution of procompsognathids in: Crichton, Michael. *Jurassic Park.* New York, NY: Ballantine Books, 1993.

The people who work at Jurassic Park are thrilled because the histogram shows a nice normal distribution, which they say is what you'd expect from height data from a typical, healthy, biological population. (This is a reasonable expectation.) However, the mathematician, Malcolm, is disturbed that this plot looks normal instead of bimodal. He says that Jurassic Park isn't supposed to be a typical biological population. It's like a zoo

– animals were introduced in two batches of baby dinosaurs, so you'd expect two modes, one for the older, taller dinosaurs, and one for the younger, shorter ones.

Malcolm thinks that the creators of Jurassic Park have made an error. However, the author, Michael Crichton, has also made an error.

There are two things wrong with this plot. One is blatantly, absolutely, 100% wrong. The other is just extremely unlikely.

a. What two things are wrong with this histogram? You may assume that the heights are, in fact, drawn from a normal distribution.

b. Roughly how many dinosaurs of this type are in the park?

c. Assuming that we do theoretically expect the heights to have a normal distribution, help Michael Crichton out by showing an example of what this histogram should have looked like. (Make the mean and variance of the distribution similar to what Crichton had in mind, and make sure you label the axes of your plot.)