Math 115 Statistics Practice First Midterm

You can bring a *handwritten* 4x6 index card to the exam. You should bring a calculator, but you won't be allowed to use any of its statistical functions. I will give you a copy of Table A on the exam. Of course, your actual exam will be shorter than this one. At least one question on your exam will be taken directly off of your homework.

1. Explain in words what the mean, variance, and standard deviation measure about a dataset. Find the mean, variance, and standard deviation of the following dataset: 14, -3, -11, 20.
The mean is the average. It's one measure of the center of the data. The variance is a measure of the spread of the data around the mean. Roughly, it's the average of the squared differences from the mean. The standard deviation is the square root of the variance, and thus is another measure of the spread. It has the same units as the data (unlike the variance, where the units are squared).
Here, the mean is (14+(-3)+(-11)+20)/4 = 5.
The variance is $((14-5)^2+(-3-5)^2+(-11-5)^2+(20-5)^2)/(4-1)=208.67$.
The s.d. is $\sqrt{208.67}=14.45$.

2. Draw a stemplot of the following data. Give the five-number summary. Without calculating the mean, explain how you can predict whether it will be greater than or less than the median.

0, 1, 1, 2, 3, 4, 4, 5, 7, 10, 40

stem & leaf
0*011234457
1*0
2*
3*
4*0

The five-number summary (min, Q1, median, Q2, max) is (0, 1, 4, 7, 40). This distribution is skewed right, so the mean will be greater (farther right) than the median.

3. The phone company offers two long-distance calling plans. Plan A charges $3.50 a month plus $0.05 (5 cents) per minute for all long-distance calls. Plan B has no monthly fee, but charges $0.07 (7 cents) per minute for all long distance calls. The number of minutes Jane spends on long-distance calls each month follows a distribution that is approximately Normal, with a mean of 70 minutes and a standard deviation (SD) of 20 minutes. Let X represent Jane's long-distance minutes in a randomly chosen month.

a) Let Y_A and Y_B represent Jane's long distance charges in a randomly chosen month under plan A and under plan B, respectively. Compute the mean and standard deviation for Y_A and Y_B.

Notice that Y_A=3.50 + .05X, and Y_B=.07X. So (see the box on p. 54), we have that
$\overline{Y}_A = 3.50 + .05\overline{X} = 7, s_A = .05 s_X = 1, \overline{Y}_B = .07\overline{X} = 4.9, s_B = .07 s_X = 1.4$

b) Give a range of values that has approximate probability .95 of containing the amount of Jane's long-distance bill under Plan A in a randomly chosen month.

95% of values of normally distributed data fall within 2 s.d.'s of the mean, so approximately 95% of Jane's bills will be in the interval (7 – 2*1, 7 + 2*1) = (5, 9). That is, 95% of her bills will be between $5 and $9 under Plan A.

c) If she chooses Plan B, in approximately what percentage of months will Jane's long-distance bill be between $3.60 and $5.00?

We want to convert 3.60 and 5.00 into s.d.'s above and below the mean. (3.60 – 4.9)/1.4 = -.93 and (5.00-4.9)/1.4=.071, so 3.60 is .93 s.d.'s below the mean and 5.00 is .07 s.d's above the mean. Consulting

Table A, we see that probability of being between -.93 and .07 s.d.'s from the mean is .5279-.1762=.3517. Thus Jane's bill under Plan B will be between $3.60 and $5.00 in approximately 35.17% of months.

d) According to our approximation, in what percentage of months will Jane spend less than 0 minutes on long-distance calls? Does this make sense? How do we interpret this?

This is the probability that X is less than 0. (0 – 70)/20= -3.5, so 0 is 3.5 s.d.'s below the mean for X. Table A goes up only to -3.49, but we can estimate that about .0002 is the fraction of months with negative time on long distance. Obviously, this can't happen; you can't spend negative time on the phone. Remember, our data are only approximately normal. Here, the fraction predicted to be negative is so small that we can safely ignore this discrepancy between our approximation and reality.

4. True or false?

a) The following is an observational study, not an experiment: Among a group of disabled women aged 65 and older who were tracked for several years, those who had a vitamin B_{12} deficiency were twice as likely to suffer severe depression as those who did not. (*American Journal of Psychology* 15 (2000): 715)
True.

b) If the correlation between two variables is close to 0, then there is only a weak association between them.
False. That means only that there's a weak linear association.

c) The area under a density curve is 1.
True.

d) For any set of data, approximately 68% will be within 1 standard deviation of the mean.
False. This is true if the data are normal, but usually not otherwise.

e) The slope of the regression line is always between -1 and 1.
False. The correlation r is always between -1 and 1, but the slope is rs_Y/s_X, which can be anything.

f) The median is resistant to outliers.
True.

5. A small town in Vermont has only one store, the Glove Shack. Each month, they sell some number of pairs of mittens. Also each month, local residents lose some number of fingers to frostbite. The data for a few months are listed below.

mittens sold	0	2	15	11
fingers lost	0	0	4	4

a) Compute the correlation between pairs of mittens sold and fingers lost to frostbite. How do you interpret your answer?

r=.967 (see section 2.2 for how to calculate this). Thus there is a very strong positive linear association between mittens sold and fingers lost.

b) Find the equation of the regression line.
F=.312M - .18 (see 2.3).

c) Evaluate the following statement: "These mittens are terrible! People are MORE likely to get frostbite when they buy them. My hatred for these terrible mittens is as deep as the ocean." What is a possible lurking variable?

Perhaps in cold months, people are more likely both to buy mittens and get frostbite. (There are certainly other explanations.)

6. A scatterplot of house prices (in thousands of dollars) versus house size (in thousands of square feet) shows a relationship that is straight, with only moderate scatter and no outliers. The correlation between house price Y and house size X is 0.85, and the equation of the regression line is Y=9.564 + 122.74X.

a) What does the slope of 122.74 mean?

An increase of home size of 1000 square feet is associated with an increase of $122,740, on average, in price.

b) How much can a homeowner expect the value of her house to increase if she builds on an additional 2000 square feet?

2*122.74 = $245,480.

c) If a house is 1 s.d. above the mean size, how many SDs above the mean would you predict its price to be?

r s.d.'s, where in this case r is .85.

d) How much of the total variability in house prices can be explained by the variability in house sizes?

r^2, or .72.

e) If we measured the size in thousands of square meters instead of thousands of square feet, how would the correlation r change? Would the slope of the regression line change?

r wouldn't change, because it doesn't depend on the units. The slope of the regression line (rs_Y/s_X) would change, because s_X would change.

f) What does the regression line predict will be the price of a house that is only 60 square feet? (That's 60, not 60 thousand.) Do you trust this prediction? Why or why not?

Y=9.564+122.74*(60/1000)= $16, 928. I don't really trust it, because 60 square feet is far smaller than the usual house.