Math 10 2017 practice midterm

Problem 1

You are given this set of data of paired (X,Y) values.

X	1	5	4	6	8
y	2	3	5	7	10

You are also given that the correlation coefficient for this set of data is r = 0.8847.

a) Without doing any calculations, state the correlation coefficient for the new set of data below. (1 pt)

X	2	6	5	7	9
y	2	3	5	7	10

b) Without doing any calculations, state the correlation coefficient for the new set of data below. (1 pt)

X	-1	-5	-4	-6	-8
y	2	3	5	7	10

You are given a **new** set of data of paired (A,B) values.

A	1	2	3	4	5
B	6	11	16	21	26

c) Without doing any calculations, state the correlation coefficient for the set of data of paired (A,B) values above. (2 pt)

d) You are given three correlation coefficients r = 0.7, r = -0.6, r = 0. Match to these descriptions of scatter diagrams. Each value of r should match to one and only one description. (3 pts)

i) Cloud of points slopping upwards.
ii) Cloud of points slopping downwards.
iii) A diamond centered at origin, filled with points.

Problem 2 *26 points*

A biologist was interested in determining whether sunflower seedlings treated with an extract from *Vinca minor* roots resulted in a lower average height of sunflower seedlings than the standard height of 15.7 cm. The biologist treated a random sample of n = 16 seedlings with the extract. The average of the sample: 13.6 cm.

a) Should she plan on using a z or t test to analyze the difference? Why?

She follows your advice to get:

SD or SD+ (she followed your advice) of sample: 2.8 cm.

b) What is the population and what is the parameter that she is trying to estimate? (*2 points*)

c) State the null and alternative hypothesis for a **one-tailed** (z or t) test for the difference from standard height.

d) What is the **one-tailed** (z or t) score for the difference from standard height?

e) Using the appropriate table (which did you use?), find out if the difference is statistically significant at the $p < 0.05$ level. Is it?

f) Interpret this p value by writing the sentence that ought to go in the research paper describing this experiment.

g) Later you find out she has actually done similar experiments a dozen times, but this is the first time she got such a low p value. She attributes this to the improvement of her experimental design. What's your explanation?

Problem 3
Using expected value and standard error, we will formalize the prosecutor's fallacy.

Suppose a crime has been committed by 1 person from a small town of 10,001 people. A DNA sample from the perpetrator was taken from the crime scene. A random person's DNA would match this sample with 0.001 probability. The perpetrator's DNA would always match this sample with probability 1.

For a random innocent individual, we can imagine this test to be a box model with 999 tickets labeled "no match" and 1 ticket labeled "match".

We can replace the labels "no match" with the number 0, and "match" with the number 1. So that we can sum the values of draws.

a) When applied to a single random individual, what is the probability of a match? *Hint: you're drawing a ticket from the box model at random, with replacement.*

b) When the DNA matching test is applied to the 10,000 people who are innocent, how many matches would you expect?

c) Write down a formula for the standard error of a sum of 10,000 draws from a box with 1/1,000 of the tickets having value 0, and 999/1,000 of the tickets having value 1. State the formula in terms of the numbers given in this paragraph but do not simplify or evaluate your formula.

d) You are given that the standard error of the sum of draws in part b is about 3. Using normal approximation, what is the probability that the number of matches when the test is applied to the 10,000 people is between 7 and 13 inclusive?

e) If someone from the small town was accused of being the perpetrator because his DNA matched the sample using the test in this question, would you say that the probability that he is guilty is 0.001? Since the probability a random person would match is 0.001. Why or why not?

Problem 4

This really happened. There are lots of mathematical models for plant growth that are used to predict crop returns in agriculture. A group of researchers using these noticed that none of the models gave results that were that close to their data, but if they averaged the results of all the different models, the answer was not far off from reality. They got really excited about this and consulted a mathematician/statistician (friend of Dr. Wallace's) to see if any light could be shed on this result. The mathematician/statistician said he wasn't surprised. He further said this:

"Think of each model output as a measurement that is equal to the true answer plus an error . . ."

a) Finish the explanation. What is the box model? What do we know or not know about the box model? How are draws from the box model used in our explanation?

b) Suppose we choose 25 models at simple random, and they predict an average plant weight of 5.1 grams dry mass after 70 days of growth. The standard deviation of the model outputs was 1.5 grams. Construct a 95.45% confidence interval for the "true answer" that these models should give.

c) In this context, what does the confidence interval actually mean