EPI 513-607C Final Examination June 21, 1990

Unless specifically asked to in a particular question, you do not need to complete any detailed calculations. However, especially if there is a possibility of ambiguity, be clear about hypotheses, levels of confidence, degrees of freedom, whether 2 sample or 1 sample procedures, which table or formula from your textbook is relevant, etc.

Q1-Q7(best 6 of 7):7; Q8-Q11(Best 3 of 4):11;Q12:25

1 What proportion of the area of the t distribution falls (a) above (ie. to right of) t = 3.169 where df = 10, (b) below t = -1.725 where df = 20, (c) between t = $\pm$ 2.462 where df = 29.

2. If the correlation coefficient is -0.80, below-average values of the dependent variable tend to be associated with below-average values of the independent variable. True or false? Explain.

3. If the dependent variable is usually less than the independent variable, the correlation coefficient will be negative. True or false? Explain.

4. In each case, which correlation is likely to be higher? Why?
   (a) height at age 4 and height at age 18, **or**
   height at age 16 and height at age 18.
   (b) height at age 4 and height at age 18, **or**
   weight at age 4 and weight at age 18.
   (c) height and weight at age 4, **or**
   height and weight at age 18.

5. True or false? Explain or give examples.
   (a) The median and the average of any sample are always close together.
   (b) Half of the values in a sample are always below average.
   (c) With a large, representative sample, the histogram is bound to follow the normal curve quite closely.
   (d) If two samples of numbers have exactly the same average of 50 and the same SD of 10, then the percentage of entries between 40 and 60 must be exactly the same for both samples.

6. In a study of men, the correlation between height and weight was 0.43. One man in the study was both one SD above average in height and one SD above average in weight. His weight will be
   (i) larger than (ii) smaller than (iii) equal to the average weight of all men of his height in the study. Explain your choice.

7 How is the standard error of the mean affected by tripling sample size?

8. Hospital A has 218 live births during the month of January. Hospital B has 536. Which hospital is more likely to have 55% or more male births? Or is it equally likely? Explain. (There is about a 52% chance for a live-born infant to be male.)

9. A study on the relationship between height and blood pressure yields the following results:

average age (X) = 40 years, SD = 10
average blood pressure (Y) = 90 mmHg, SD = 20
The correlation coefficient was r = 0.70

(a) What is the regression equation?
(b) What would you predict the blood pressure to be for an individual aged 50 years old?

10. The effectiveness of vitamin C in orange juice and in synthetic ascorbic acid was compared in 20 guinea pigs (divided at random into two groups of 10) in terms of the length of the odontoblasts after 6 weeks, with the following results;

Orange juice:

8.2	9.4	9.6	9.7	10.0	14.5	15.2	16.2	17.6	21.5

Ascorbic acid:

4.2	5.2	5.8	6.4	7.0	7.3	10.1	11.2	11.3	11.5

We wish to test the hypothesis of no difference against the alternative that the orange juice tends to give rise to larger values.

(a) Is this a 1-sided or 2-sided test? Clearly write down the null and alternative hypothesis. Is this a matched or an unmatched study?
(b) Rank the observations from lowest rank = 1 to highest rank = 20, and find the sum of ranks in the ascorbic acid group.
(c) Using the Normal approximation to the appropriate non-parametric test, carry out the test and find the p-value.

11. Refer to the letter entitled "More on sex and racial bias in pharmaceutical advertisements."

(a) Assume that the data set containing the n1=371 and n2=426 advertisements was available to you. Describe how you would carry out (but do not carry out) a test of whether the 9.7% of advertisements picturing physicians in 1979 differs from the 4.9% figure of 1989.
(b) Focus on the 3.1% vs 4.4% in the advertisements picturing patients from minority members only. Is is possible to use a test based on the normal distribution or the chi-square to test for equality of percentages?

12. Seat belt use on interstate highways [AJPH 1990; 80:741-742]

a) The observed seat belt use in Massachusetts was 49%, the N was 404, and the Standard Error of Percent Belt Use was 2.5 [Table 1, 6th entry from bottom].
Show how the authors calculated the 2.5 (use the "rounder" numbers of 50% and N=400 to simplify the calculations).

b) "Only differences alpha = 0.05 are reported" [last sentence of methods section]

(i) Rewrite this sentence to make it clearer.
(ii) Comment on the implications/dangers of using this reporting policy.

c) "In the United States, use was higher in states with laws (60%) than states without laws (56%)" [1st half of 3rd sentence of Results section]

(i) Set up (but do not perform) the calculations for the statistical test implied in this statement.
(ii) Use the 60% vs 56% figures to illustrate the difference between statistical significance and real ("clinical") significance (you may wish to refer to the last portion of the last sentence of the abstract)

d) "but the range of use rates for states with and without laws overlapped considerably" [2nd half of 3rd sentence of Results section]
Illustrate how, without any distributional assumptions, one can test whether the states with laws tended to have higher usage measurements than those without. Give all of the measurements equal weight.

e) Belt use was higher among drivers (than among right front passengers)" [last sentence of discussion, referring to data in last section of Table 2].
This compared 9,519 drivers with 4,835 passengers. If you restricted the comparison to the 4,835 drivers and their 4,835 right front passengers, how would you set up the appropriate 2x2 table and perform the appropriate statistical test.

f) Although it is not stated explicitly in the Methods section, one presumes that when a car contained a front right passenger, both the driver and the passenger were observed, (ie. that the 14,354 occupants were in 9519 cars). Does this presumption affect the calculation of standard errors in Table 1? Why/why not? If yes, what impact does it have?