STA 6126 **Practice questions: EXAM 2**

1. A study was conducted of the effects of a special class designed to aid students with verbal skills. Each child was given a verbal skills test twice, both before and after completing a 4-week period in the class. Let Y = score on exam at time 2 - score on exam at time 1. Hence, if the population mean μ for Y is equal to 0, the class has no effect, on the average. For the four children in the study, the observed values of Y are 8-5=3, 10-3=7, 5-2=3, and 7-4=3 (e.g. for the first child, the scores were 5 on exam 1 and 8 on exam 2, so Y = 8-5=3). It is planned to test the null hypothesis of no effect against the alternative hypothesis that the effect is positive, based on the following results from a computer software package:

Variable	Number of Cases	Mean	Std. Dev.	Std. Error
Y	4	4.000	2.000	1.000

   a. Set up the null and alternative hypotheses.

   b. Calculate the test statistic, and indicate whether the P-value was below 0.05, based on using the appropriate table.

   c. Make a decision, using $\alpha = .05$. Interpret.

   d. If the decision in (c) is actually (unknown to us) incorrect, what type of error has been made? What could you do to reduce the chance of that type of error?

   e. True or false? When we make a decision using $\alpha = .05$, this means that if the special class is truly beneficial, there is only a 5% chance that we will conclude that it is not beneficial. _____

2. For a random sample of University of Florida social science graduate students, the responses on political ideology had a mean of 3.18 and standard deviation of 1.72 for 51 nonvegetarian students and a mean of 2.22 and standard deviation of .67 for the 20 vegetarian students.

   a. Defining appropriate notation, state the null and alternative hypotheses for testing whether there is a difference between population mean ideology for vegetarian and nonvegetarian students.

   b. When we use software to compare the means with a significance test, we obtain the printout

Variances	T	DF	Prob>|T|
Unequal	2.9146	41.9	0.006
Equal	1.6359	69.0	0.107

Interpret.

3. (13 pts.) A geographer conducts a study of the relationship between the level of economic development of a nation (measured in thousands of dollars for per capita GDP) and the birth rate (average number of children per adult woman). For one analysis of the data, a part of the computer printout reports

Statistic	Value	SE
Correlation	-0.460	0.170

   a. Explain how to interpret the reported value of the correlation.

   b. Can you tell whether the sign of the slope in the corresponding prediction equation would be positive or negative? Why?

   c. Suppose the prediction equation were $\hat{y} = 5.6 - 0.1x$. Interpret the slope, and show how to find the predicted birth rate for a nation that has $x = 40$.

   d. Sketch a scatterplot for which it would be inappropriate to use the correlation to describe the association.

4. A study on educational aspirations of high school students (S. Crysdale, *International Journal of Comparative Sociology*, Vol. 16, 1975, pp. 19–36) measured aspirations using the scale (some high school, high school graduate, some college, college graduate). For students whose family income was low, the counts in these categories were (9, 44, 13, 10); when family income was middle, the counts were (11, 52, 23, 22); when family income was high, the counts were (9, 41, 12, 27). Software provides the results shown below.

Table 1:

Statistic	DF	Value	Prob
Chi-Square	6	8.871	0.181

   a. Find the sample conditional distribution on aspirations for those whose family income was high.

b. Give all steps of the chi-squared test, explaining how to interpret the P-value.

c. Explain what further analyses you could do that would be more informative than a chi-squared test.

The following questions are true or false. Indicate T or F next to each.

5. _____ For a given set of data on two quantitative variables X and Y, the slope of the least squares prediction equation and the correlation must have the same sign.

6. _____ For a given set of data on two quantitative variables X and Y, the prediction equation and the correlation do not depend on the units of measurement.

7. _____ The standardized residual that follows up the chi-squared test for a contingency table has value 3.2 in the cell in row 1 and column 1. This means that if the variables were independent, it would be very unusual to observe so many observations in that cell.

8. _____ Suppose that a study reports that a 95% confidence interval for the difference $\mu_1 - \mu_2$ between the population mean annual incomes for whites (μ_1) and for Hispanics (μ_2) having jobs in home construction is (\$5000, \$5400). Then, a 95% confidence interval for the difference $\mu_2 - \mu_1$ between the population mean annual incomes for Hispanics and for whites having jobs in home construction is (-\$5400, -\$5000).

9. _____ A study of medical utilization compares mean stay in the hospital for heart transplant operations in 1999 to the mean stay in 1995, for two separate samples of such operations in the two years. In the comparison, since the same variable ("length of stay in the hospital") is measured for each sample, the data should be analyzed using methods for **dependent** samples (such as the paired-difference t test) rather than **independent** samples.

10. _____ Refer to the previous question. Suppose the data in 1995 were summarized by $\bar{y}_1 = 10.5, s_1 = 8.9$ ($n_1 = 54$), and the data in 1999 were summarized by $\bar{y}_2 = 8.0, s_1 = 7.8$ ($n_2 = 48$). These statistics suggest that the variable "length of stay in the hospital" does not have a normal distribution. Therefore, even though the samples are relatively large, we cannot use the formula $(\bar{y}_1 - \bar{y}_2) \pm t(se)$, which assumes normal population distributions.

11. _____ For large samples, the reason we refer the z test statistic for testing a proportion or difference of proportions to the normal distribution is because these methods assume that the population distribution is normal.

12. _____ Most statisticians believe that social scientists and other users of statistics put too much emphasis on confidence intervals and not enough on significance tests, since significance tests give us more information about the value of a parameter than confidence intervals.

13. _____ When we make a decision in a significance test, the reason we say "Do not reject H_0" instead of "Accept H_0" is because a confidence interval for the parameter would show that the number in H_0 is one of only many plausible values for the parameter.

14. _____ We decide to conduct a significance test to see if there is enough evidence to predict whether a majority or a minority of Floridians are in favor of affirmative action. Letting π denote the population proportion of Floridians in favor of affirmative action, we would set up the hypotheses as $H_0 : \pi \neq .50$ and $H_a : \pi = .50$.

15. _____ In the 1982 General Social Survey, 350 subjects reported the time spent every day watching television. The sample mean was 4.1 hours, with standard deviation 3.2. In a more recent General Social Survey, 1965 subjects reported a mean time spent watching television of 2.8 hours, with standard deviation 2.0. The standard error used in inference (confidence intervals and significance tests) about the true difference in population means in these two years equals $3.2 + 2.0 = 5.2$.

16. _____ Refer to the previous question. Based on the means and standard deviations reported here, it seems very plausible that the true population distribution of time spent watching television was very close to the normal distribution in both of these years.

17. _____ Refer to the previous two questions. If, in fact, the population distributions are not normal in these years, then it is invalid to construct t confidence intervals and tests about the difference of means.

Select the correct response(s) in the following problems.

18. To compare the population mean annual incomes for Hispanics (μ_1) and for whites (μ_2) having jobs in construction, we construct a 95% confidence interval for $\mu_2 - \mu_1$.

    a. If the confidence interval is (3000, 6000), then at this confidence level we conclude that the mean income for whites is lower than for Hispanics.

    b. If the confidence interval is (–1000, 3000), then the corresponding test of H_0: $\mu_1 = \mu_2$ against H_a: $\mu_1 \neq \mu_2$ has a P-value above .05.

c. If the confidence interval is (–1000, 3000), then we can conclude that $\mu_1 = \mu_2$.

d. If the confidence interval is (–1000, 3000), then we can conclude that 95% of the white subjects in the population have income between \$1000 less and \$3000 more than 95% of the Hispanic subjects in the population.

19. In using a t test for a mean, we assume that

a. The population distribution is normal, although in practice the method is robust to this assumption.

b. The sample is selected randomly.

c. $\overline{y}$ is a dependent variable, and μ_0 is an independent variable.

d. Each expected frequency is at least about 5.

e. None of the above, because the t test is robust against violations of *all* its assumptions.

STA 6126: Formulas – Exam 2

$$t = \frac{\overline{y}-\mu_0}{se} \quad df = n-1 \qquad SE = \frac{s}{\sqrt{n}} \qquad s = \sqrt{\frac{\sum(y-\overline{y})^2}{n-1}}$$

$$z = \frac{\hat{\pi}-\pi_0}{se_0}, \qquad \sigma_{\hat{\pi}} = \sqrt{\frac{\pi_0(1-\pi_0)}{n}}$$

$$(\hat{\pi}_2 - \hat{\pi}_1) \pm z(se), \quad se = \sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}$$

$$t = (\overline{y}_2 - \overline{y}_1)/se, \quad (\overline{y}_2 - \overline{y}_1) \pm t(se), \quad se = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$\chi^2 = \sum \frac{(f_0 - f_e)^2}{f_e}, \quad df = (r-1)(c-1), \quad f_e = (\text{row total})(\text{col. total})/n$$

$$\text{standardized residual } = \frac{f_o - f_e}{\sqrt{f_e(1-\text{row prop})(1-\text{col prop})}}$$

Bivariate regression models

$$E(Y) = \alpha + \beta x \qquad \hat{y} = a + bX \qquad r = b(s_X/s_Y) \qquad r^2 = (TSS - SSE)/(TSS)$$