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 report the P-value.

   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. Show how to compare the corresponding population means using a 90% confidence interval. Interpret.
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. A multiple-choice test question has four possible responses. The question is designed to be very difficult, with none of the four responses being obviously wrong, yet with only one correct answer. It first occurs on an exam taken by 400 students. The designers test whether more people answer the question correctly than would be expected just due to chance (i.e., if everyone randomly guessed the correct answer).
a. Set up the hypotheses for the test.
b. Of the 400 students, 125 correctly answer the question. Find the P-value, and interpret.
c. Make a decision about H_0, using $\alpha = .05$. Based on this decision, what can you conclude about the parameter?
d. As an alternative inference, construct a 95% confidence interval for the parameter. Contrast what you learn using these two methods.

4. (13 pts.) A geographer conducts a study of the relationship between the level of economic development of a nation and the birth rate. In one presentation of the data, a 3×3 contingency table is used for measures of each variable as high, medium or low. Part of the computer printout reports

Statistic	Value	ASE
Gamma	-0.460	0.170

a. Explain how to interpret the reported value of gamma.
b. Assuming the data refer to a random sample of nations, construct a 95% confidence interval for the population gamma. Interpret.
c. Show an example of a 3×3 contingency table, labeling the three categories, for which the sample value of gamma would equal -1.

5. 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. Explain how to analyze and interpret these data.

Table 1:

Statistic	DF	Value	Prob
Chi-Square	6	8.871	0.181

Statistic	Value	ASE
Gamma	0.163	0.080
Kendall's Tau-b	0.108	0.053

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

6. ______ 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).

7. ______ 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.

8. ______ 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 $(\bar{y}_1 - \bar{y}_2) \pm z\hat{\sigma}_{\bar{y}_1 - \bar{y}_2}$, which assumes normal population distributions.

9. ______ For large samples, the reason we refer the test statistics z for testing a mean or proportion or difference of means or difference of proportions to the normal distribution is because all these methods assume that the population distribution is normal.

10. ________ 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.

11. ________ 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.

12. ________ 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 $H_a : \pi = .50$.

13. ________ When we use the t distribution to construct a confidence interval for a mean for small samples, we must assume that the population distribution has the t distribution, whereas for large-sample confidence intervals for a mean we must assume that the population distribution has a normal distribution.

14. ________ 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 the 1994 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 $3.2 + 2.0 = 5.2$.

15. ________ 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.

16. ________ Refer to the previous two questions. If, in fact, the population distributions are not normal in these years, then it is invalid to construct large-sample confidence intervals and tests about the difference of means.

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

17. When the rows and columns of a contingency table are ordered, analyses based on an ordinal measure of association such as gamma are often preferred over conducting the Pearson chi-squared test of independence because

a. Gamma treats the variables as ordinal whereas chi-squared treats them as nominal.

b. Gamma can detect whether there is a positive or negative trend, whereas chi-squared is not designed to detect the direction of the association.

c. The large-sample z test based on gamma usually gives a smaller P-value than the large-sample chi-squared test.

d. None of the above. For contingency table data, the chi-squared test is always the preferred inferential method.

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

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

$z = \frac{\overline{Y}-\mu_0}{\hat{\sigma}_{\overline{Y}}}$ $\qquad \hat{\sigma}_{\overline{Y}} = \frac{\hat{\sigma}}{\sqrt{n}}$ $\qquad \hat{\sigma} = s = \sqrt{\frac{\Sigma(Y-\overline{Y})^2}{n-1}}$

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

$$t = \frac{\overline{Y}-\mu_0}{s/\sqrt{n}}, \quad df = n-1$$

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

$$z = (\overline{Y}_2 - \overline{Y}_1)/\hat{\sigma}_{\overline{Y}_2-\overline{Y}_1}, \quad (\overline{Y}_2 - \overline{Y}_1) \pm z\hat{\sigma}_{\overline{Y}_2-\overline{Y}_1}, \quad \hat{\sigma}_{\overline{Y}_2-\overline{Y}_1} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

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

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

$$\hat{\gamma} = \frac{C-D}{C+D} \qquad \hat{\gamma} \pm z\hat{\sigma}_{\hat{\gamma}} \qquad z = \hat{\gamma}/\hat{\sigma}_{\hat{\gamma}}$$