Homework 9

Comments and Instructions:

A. For every problem calling for you to estimate a parameter or test a claim about a parameter, an early step is to identify the parameter(s) needed, including the correct notation, and, if it is a hypothesis test, to write both hypotheses correctly. That skill is required to begin the steps where you use the data appropriately to form the confidence interval or find the *p*-value for the hypothesis test.
This course does not provide specific questions in the homework solely about choosing the correct parameters and writing hypotheses correctly. Applied statistics texts and elementary statistics texts can be used to find such exercises. Particularly useful are review chapters in elementary statistics texts. In the Lock text, Chapters 3 and 4 address problems of all types, since the method of solution is always the same – use the StatKey simulations. Thus, that is a good place to start.

B. Correct interpretations of both confidence intervals and *p*-values are essential. Those were discussed in the Week 5, and you'll see it in later assignments in the course.

C. For ALL statistical computation exercises in this week, use simulation-based techniques with the software from http://www.lock5stat.com/StatKey. Do the simulation and choose your answer according to the instructions you have been given during the previous lessons using simulations. Solutions are prepared using StatKey in the Google Chrome browser.

D. In StatKey, in the hypothesis tests for the difference of two parameters (means and proportions,) you are given a choice of the type of randomization. The solution keys for all assignments are made using the choice that appears by default. (Has "reallocate" in the name.) In real applications, another choice may be appropriate. We choose not to ask students to make those distinctions in graded work in this course.

E. The datasets used in this assignment are all from the 3rd edition of the Lock textbook, which can be downloaded from www.lock5stat.com/. We believe these particular datasets are the same in all of editions 1, 2, and 3. (If you find that is not true, please tell us so that we can clarify that to everyone.) The datasets used here are **StudentSurvey**, **SalaryGender**, and **CocaineTreatment**.
    a. Assume the **StudentSurvey** data was collected from a random sample of students at a particular college, and the inferences discussed here are about the population from which that was drawn.
    b. The description of the **SalaryGender** dataset indicates that it is US census data from 2010 and is from a random sample of College Teachers. Assume that the inferences here are about the population of US college teachers in 2010.
    c. For the **CocaineTreatment** dataset, the subjects were volunteers and they were randomly assigned to the three groups.

F. Answering questions about Confidence Intervals: Find the confidence interval and then find the length of it by subtracting : Right-hand endpoint – Left-hand endpoint.

G. Answering questions with numerical answers: Do the appropriate calculations and then choose the closest number among the choices. If two choices are equally close to the correct value, then both are counted as correct.

Actual Homework Problems:

1. StatKey dataset **SalaryGender**: Test the claim that the average age of the college teachers in the population is less than 49 years. Find the p-value.

   Choices a. 0.019 b. 0.025 c. 0.040 d. 0.050 e. 0.080 f. 0.120 0.081

2. StatKey dataset: **StudentSurvey.** Find a 90% confidence interval for the mean Math SAT score from this population. What is the length?

   Choices: a. 6.0 b. 8.0 c. 10.0 d. 12.0 e. 14.0 615.267 - 603.398 = 11.869

3. For the **CocaineTreatment** dataset, find the following pairs of sample proportions:
   i. The proportion of those in the study who did not relapse and the proportion of those in the study who did relapse.
   ii. The non-relapse proportions for (1) those who received the placebo and (2) those who received lithium treatment.
   iii. The non-relapse proportions for (1) those who received the lithium and (2) those who received desipramine.

   To do inference on any of these, we must correctly identify whether each pair involves a question about "one proportion" or about "two independent proportions."

   The distinction is: if the denominator of the two proportions is composed of exactly the same individuals, then the question is really about only one proportion *p*. That's because the other proportion is *1-p*, which depends on *p.*

   Which of those pairs of proportions you computed is/are really two independent proportions?

   Choices
   a. Only (i) b. Only (ii) c. Only (iii) d. Both (i) and (ii) e. Both (ii) and (iii)
   f. Both (i) and (iii) g. All three of (i) (ii) and (iii)

4. For the **CocaineTreatment** data, test the claim that, the population proportion not receiving either drug treatment, the proportion of "non-relapse" is less than 0.30. Find the p-value.

   Choices: (a) 0.025 (b) 0.050 (c) 0.075 (d) 0.100 (e) 0.200 0.118

5. For the **CocaineTreatment** datasesett, test the claim that the population proportion of "non-relapse" people is higher for those treated with lithium than from those not treated (i.e. treated with a placebo in our sample.) Find the p-value.
   Choices a. 0.042 b. 0.054 c. 0.165 d. 0.341 e. 0.413 f. 0.720 0.360

6. For the **CocaineTreatment** dataset find an 83% confidence interval for the difference in the proportion of people in the population who are expected to have no relapse between those treated with desipramine and those treated with lithium. What is the length?
   i. 0.29 b. 0.49 c. 0.55 d. 0.69 e. 0.82 0.333

7. StatKey dataset: **SalaryGender**. Consider the regression equation to predict Salary from Age. Find a 92% confidence interval for the slope coefficient in the population. What is the length?

   Choices: (a) 0.112 (b) 0.145 (c) 0.261 (d) 0.500 (e) 0.770 (f) 0.910 1.747 - 0.894 = 0.876

8. StatKey dataset: **SalaryGender**. Find the 97% confidence interval for the standard deviation of the salaries in the population. What is the length?
   Choices: (a) 7.0 (b) 9.1 (c) 10.1 (d) 13.1 (e) 14.6 (f) 16.1 (g) 18.0 49.099 - 34.536 = 14.563

9. StatKey dataset **SalaryGender**:
   9a. Find a 95% confidence interval for the median of the ages in the population. What is the length?
   Choices: (a) 7.0 (b) 9.1 (c) 10.1 (d) 13.1 (e) 14.5 (f) 16.1 (g) 18.0 53.5 - 44.5 = 9

   9b. Your opinion: (not graded) Does the method for providing a 95% confidence interval for the population median seem to be reasonable for this data and this confidence level? Discuss why or why not.

10. Suppose we have data on two quantitative variables (X and Y) from a random sample from a population. When we want to determine whether the variable X is useful in predicting values of the variable Y using linear regression. We can test the claim that the slope is not equal to 0. (A conclusion that the slope coefficient is not zero means that the variable X is useful in predicting values of the variable Y, using linear regression.) $H_0 : \beta = 0$ $H_A : \beta \neq 0$

    For the **StudentSurvey** dataset, determine which one of these variables shows the strongest evidence of being useful to predict weight using simple linear regression.

    Choices a. SAT b. Pulse c. Birth Order d. GPA e. Exercise

Additional Comments about using hypothesis testing on a slope coefficient in linear regression:

A. A variable can be quite useful in predication without being the "cause" of the change in Y. Review the idea of "confounding variables, "discussed in Week 6. (This particular question was picked to remind you of that!)
B. As you know, a line with slope 0 is a horizontal line. A horizontal line indicates that, for all values of X, we obtain the same value of Y. Thus the variable X has no impact on the

variable Y.
Thus, graphically, X is a NOT a good predictor of Y if the graph of the regression line "looks" really close to horizontal. But "looks" can be misleading, as can simply observing the numerical value of the slope. We have to take into account the units of measurement of the variables and the scale of the graph we're looking at. The statistical analyses take those into account through the standard errors.

C. You will obtain the same p-value if you use the correlation coefficient as the test statistic instead of the slope coefficient as the test statistic.
(This comes from the fact that the sample slope coefficient is simply the product of the sample correlation coefficient times the ratio of the standard deviations of the two variables.)

D. Often you can make better predictions of the response variable by using more than one explanatory variable. Do that with multiple regression techniques.

E. Try several things with the StudentSurvey dataset to predict Weight from one or multiple explanatory variables. (StatKey does not have an app for Multiple Regression, so if you want to try that, you'll have to do it with some different software.)
If you have not done anything with multiple regression before, and have time, it would be useful to read through an explanation of this in your elementary statistics book whether you actually try it using software or not. Many elementary statistics students have commented that doing this really helps them see how the ideas of statistics inference "come together" for them.

11. A particular restaurant is interested in whether changing to a different strategy of lighting in the dining areas will affect the tips. They have gathered information about tips from the time where they had their usual lighting for comparison purposes. That data is not needed to address the question in this problem.
For this current study, they changed the lighting for two months and recorded the data during that time. They are treating this new data as if it is a random sample of the tips of all their customers while this new strategy of lighting is used in the future (defined as the next three years.) A 95% confidence interval for the average percentage tip in this population was correctly computed to be 13.7% to 16.1%. .

Based on the information provided in this course about interpreting confidence intervals, which two of these is/are fully correct statements interpreting this result? (To earn **any** credit you must correctly identify both of the correct interpretations given.)

(Be able to explain what is incomplete or incorrect about the others.)

a. There is a 95% probability that the average percentage tip for all the customers of this restaurant during the time of the revised lighting method is used will be between 13.7% and 16.1%.
b. I am 95% sure that the average percentage tip for all the customers of this restaurant during the time the revised lighting method is used will be between 13.7% and 16.1%.
c. I have 95% confidence that the average percentage tip for all the customers of this restaurant in our sample is between 13.7% and 16.1%.

d. I have 95% confidence that the average percentage tip for all the customers of this restaurant during the time the revised lighting method is used will be between 13.7% and 16.1%.
e. I have 95% confidence that the average percentage tip for a different sample of these customers of this restaurant during the time the revised lighting method is used will be between 13.7% and 16.1%.

12. In the **CocaineTreatment** experiment, we tested the claim that the population proportion of no-relapse cases among those subjects who took desipramine is greater than 0.50 and found that the p-value = 0.28. The statement of hypotheses is $H_0 : p = 0.5$ $H_A : p > 0.5$.

Which one of these is a fully correct statement interpreting this p-value?

a. The probability of finding a sample proportion as extreme or less extreme than the sample proportion of our data is 0.28.
b. The probability of finding a sample proportion as extreme or less extreme than the sample proportion of our data is 0.28 if the null hypothesis is true.
c. The probability of finding a sample proportion as extreme or less extreme than the sample proportion of our data is 0.28 if the alternative hypothesis is true.
d. The probability of finding a sample proportion as extreme or more extreme than the sample proportion of our data is 0.28.
e. The probability of finding a sample proportion as extreme or more extreme than the sample proportion of our data is 0.28 if the null hypothesis is true.
f. The probability of finding a sample proportion as extreme or more extreme than the sample proportion of our data is 0.28 if the alternative hypothesis is true.

Recall the information in this course (and the Lock text, Sec. 1.3) about the design of a study (methodology of the study) that supports generalizing to the population and/or can be used as evidence about causality. Use that in answering any questions about these topics in this course.

13. For the two questions below, consider the StatKey 3rd ed. dataset **CocaineTreatment.** Assume that the subjects were volunteers who were randomly assigned to the treatments.

13a. Is the design of this study appropriate to generalize to the population of cocaine users?
Choices: (a) Yes (b) No

13 b. Is the design of this study appropriate to give evidence for causality if a sufficiently strong effect is found?
Choices: (a) Yes (b) No

14. (Not graded) Consider the StatKey 3rd ed. dataset **CocaineTreatment.** Why do you suppose the subjects were volunteers rather than being randomly selected from the appropriate population?
In other words, what would random selection from the population be like?
How do medical studies handle this difficulty?