Homework 11.

This week concludes our coverage of elementary/applied statistics as needed for the later assessments in this course and later courses in the program.

- Pay VERY CAREFUL attention to all information **about conditions for using theoretical dist'ns in the 6-page handout – all the way to the end of page 6!**
- The questions here for which you are expected to give numerical answers (Part 2) are specific about whether to use a simulation-based method or a theoretical dist'n method.

When you use a theoretical method, use the appropriate summary statistics as given by StatKey and any probability calculations as given by StatKey. **Your answers will be graded from an answer key using those and not values from any other tables or calculations.** (This is true for ALL remaining statistics calculations in this course for any probability distributions provided in StatKey.)

There are many other questions here (Part 1.) Even though you are not required to give numerical answers for these other questions, this would be a good place to practice doing the procedures in both ways (simulation-based and theoretical dist'n) and compare your work. Either type might appear on an exam.

Different parts of the homework

> **Part 1. Problems 1-13. Consult together on the class Discussion Board, and in your small groups.** Note that full solutions are **not** required to answer the questions for credit in this part. Nevertheless, feel free to share complete solutions to these problems using either or both methods and discussing any similarities and differences you see. (All of Part 1 together will count somewhat less than 40% of the grade on HW11)
>
> **Part 2. Problems 14-18. Do not consult** in the same way as in Part 1. Instead, **follow the usual homework rules for the Discussion Board.** (All of Part 2 together will count a bit more than 60% of the grade on HW 11.)
>
> **Datasets**: As before**,** use the dataset as downloaded from the 3rd edition datasets. Do not use any dataset in the dropdown menus, as it is possible some of them may differ from the full dataset.

Part 1.

Consult together freely, even on Piazza, even though these are for some credit. Review the beginning discussion in last week's homework about conditions. (See the discussion of the Wetsuit problem.)It Is very important here.

1. (1 point) If you were to carry out the procedure here, are the conditions for using a theoretical dist'n strictly met, according to our Statistical Formula document?
   MustangPrice: Find a 90% confidence interval for the population mean price.
   Choices: (a) Yes (b) Either they might be considered to meet the conditions after additional investigation or No

2. (1 point) If you were to carry out the procedure here, are the conditions for using a theoretical dist'n strictly met, according to our Statistical Formula document?
   CommuteAtlanta: Test a claim about whether the population mean commute time is less than 30 minutes.
   Choices: (a) Yes (b) Either they might be considered to meet the conditions after additional investigation or No

3. (1 point) If you were to carry out the procedure here, are the conditions for using a theoretical dist'n strictly met, according to our Statistical Formula document?
   In **ICUAdmissions**, Find a 90% confidence interval for the difference between the proportion of Cancer patients who entered ICU with an infection and the proportion of patients without Cancer who entered the ICU with an infection. (Look at the Data Description file to see how to interpret the coding for the variables Cancer and Infection.)
   Choices: (a) Yes (b) Either they might be considered to meet the conditions after additional investigation or No

4. (1 point) If you were to carry out the procedure here, are the conditions for using a theoretical dist'n strictly met, according to our Statistical Formula document?
   In manufacturing plant A, we want to find a 95% confidence interval for the population proportion of defective items. In a sample of size 200, we find 12 defective items.
   Choices: (a) Yes (b) Either they might be considered to meet the conditions after additional investigation or No

5. (1 point) If you were to carry out the procedure here, are the conditions for using a theoretical dist'n strictly met, according to our Statistical Formula document?
   In manufacturing plant B, we want to test the claim that the proportion of defective items coming out of the process is greater than 0.06. In a sample of size 200, we find 14 defective items.
   Choices: (a) Yes (b) Either they might be considered to meet the conditions after additional investigation or No

6. (1 point) If you were to carry out the procedure here, are the conditions for using a theoretical dist'n strictly met, according to our Statistical Formula document?
   In manufacturing plant C, we want to test the claim that the proportion of defective items coming out of the process is greater than 0.04. In a sample of size 200, we find 12 defective items.
   Choices: (a) Yes (b) Either they might be considered to meet the conditions after additional investigation or No

7. (1 point) If you were to carry out the procedure here, are the conditions for using a theoretical dist'n strictly met, according to our Statistical Formula document?
   In the **ExerciseHours** dataset, test the claim that the population mean exercise hours differs for males and females.
   Choices: (a) Yes (b) Either they might be considered to meet the conditions after additional investigation or No

8. (1 point) If you were to carry out the procedure here, are the conditions for using a theoretical dist'n strictly met, according to our Statistical Formula document?
   In the **TrafficFlow** dataset, we want to decide how much of a difference in delay time there is between two different scenarios. We have the results of one experiment that simulated busses moving along a street and recorded the delay time (in seconds) for both a fixed-time system for traffic lights and a flexible system of traffic lights. (Flexible time means that sensors monitor the flow of traffic and use that to adjust the timing of the lights.) The simulation was repeated under both conditions for a total of 24 trials. The data is paired because we have two values for each simulation run. (Reference: Lock text.)
   Pay careful attention to which of the variables (one or more) must be used to check the conditions.
   Choices: (a) Yes (b) Either they might be considered to meet the conditions after additional investigation or No

9. (1 point) If you were to carry out the procedure here, are the conditions for using a theoretical dist'n strictly met, according to our Statistical Formula document?
   In the **CocaineTreatment** dataset, test whether the two variables are associated.
   Choices: (a) Yes (b) Either they might be considered to meet the conditions after additional investigation or No

10. (1 point) If you were to carry out the procedure here, are the conditions for using a theoretical dist'n strictly met, according to our Statistical Formula document?
    In the RestaurantTips dataset, test whether the mean PercentTip for the servers are different.
    Choices: (a) Yes (b) Either they might be considered to meet the conditions after additional investigation or No

11. (1 point) Consider the **FootballBrain** dataset with its three categories: No football, football with no concussion, football with concussion.
    a. Investigate whether these is a significant difference in mean hippocampus volume among the three categories.
       i. Give appropriate comparative graphs.
       ii. Test appropriate hypotheses using a randomization test.
       iii. Do the data strictly meet the conditions for using a theoretical distribution to test these hypotheses? If so, do that test as well.
    b. Investigate whether there is a difference in mean cognition between the two categories of football players.
       (Hint: You will find this MUCH easier if you first try to use the **FootballBrain** dataset as it is provided in StatKey before trying something that involves editing the dataset. Discuss that in Piazza as needed.)
       i. Give appropriate comparative graphs.
       ii. Test appropriate hypotheses using a randomization test.
       (This is a useful problem for you to discuss with each other what it took for you to decide upon an answer to actual graded question for this problem.)
       iii. Do the data strictly meet the conditions for using a theoretical distribution to test these hypotheses?

    What to answer to be graded: Is the test in part b. ii. significant at the 5% significance level?
    Choices: Yes No

12. (1 point) Benford's law has various uses, including in auditing. Read about it in the Lock text, Sec. 7.1 in exercise 7.33 (3rd ed) and close to that is the other editions. Or read about Benford's law on Wikipedia.
    Look at the StatKey dataset **Benford.** It has the probabilities for Benford's law and two different datasets.
    Which of the two different sets of data included in that file is inconsistent with the probabilities given by Benford's law? What does that suggest about the situation from which the other set of data was collected?
    To submit this, simply answer this question "Which of the two different sets of data included in that file is inconsistent with the probabilities given by Benford's law?

    Choices: Addresses Invoices

13. (1 point) In the RestaurantTips dataset, is there an association between the Day of the week and whether the bill was paid by a credit card or not? Do we reject the null hypothesis at the 10% significance level?

    Choices: Yes No

Part 2. (Usual homework rules on the Discussion Board)

For the new material this week, "test statistic" refers to the chi-squared statistic or F statistic.

All the problems below require precise numerical answers.
The solution key is prepared using StatKey and the Visualize applets for all computations and for the probabilities from theoretical distributions.

In the several-step problems, always use your precise numerical answer, not a rounded answer, as you proceed to the next part of the problem.

Give all the required numerical answers which are not exact whole numbers in these problems **using three decimal places**.

14. (6 points) In the StudentSurvey dataset, students chose which of the 3 awards they would prefer to win: Olympic medal, Nobel Prize, or Academy Award. Test the claim that 9% of the students in the population this represents prefer the Academy Award and the remaining students are evenly split between the other two awards. Use a theoretical distribution.

    a. (2 points) What are the degrees of freedom?

    b. (4 points) Find the p-value for the test using that theoretical distribution.

15. (6 points) In the **StudentSurvey** dataset, students chose which of the 3 awards they would prefer to win: Olympic medal, Nobel Prize, or Academy Award. The dataset also includes a column **HigherSAT** indicating the part of the SAT in which the student made a higher score. **Using exactly that column of values**, test the claim that there is an association between the type of award a student prefers and their value in that **HigherSat** data column. Use a theoretical dist'n.

    a. (3 points) What is the value of the test statistic?
    b. (3 points) Find the p-value for the test.

16. (6 points) In the **TextbookCosts** dataset, the variable **Cost** is the "Total Cost (in dollars) per course for the required books." Test the claim that there is a difference in the average of the **Cost** variable in the several "fields" of study. Do this using a theoretical dist'n.

    a. (3 pts) What is the value of the test statistic?
    b. (3 pts) Find the p-value for the test.

17. (4 points) A survey from a random sample of households in a suburb included two questions "Have you visited a public library in the last year?" and "Are there children under 12 in your household?" The table below summarizes these results.

	Children under 12	No children under 12	Total
Visited a library	29	15	44
Didn't visit a library	40	105	145
Total	69	120	189

    a. (2 points) To construct a confidence interval, using theoretical methods, for the difference of the proportions of households with children under 12 who have visited a library in the last year and the proportion of households without children under 12 who have visited a library in the last year, we must compute the appropriate standard error, called SE. What is the SE?

    b. (2 points) To test the hypotheses, using theoretical methods, that there is a difference between the proportion of households, using theoretical methods, with children under 12 who have visited a library in the last year and the proportion of households without children under 12 who have visited a library in the last year, we must compute the appropriate standard error, called SE. What is the SE?

18. (4 points) A survey is planned to estimate the proportion of voters who support gun control. We want a 90% confidence interval with a margin of error of 3%.

    a. (2 points) We have no information about the proportion of the voters who support gun control. How many people need to be included in the sample?

    b. (2 points) Suppose a prior study indicated that about 70% of the voters support gun control. How many people need to be included in the sample?