Week 11 and Homework 10 – Discussion

The main challenge this week is to keep the overview firmly in mind as you go through this material. There are MANY formulas and rules. In the Week 0 Handouts, find a multi-page handout of the formulas for you to refer to, so you don't have to LEARN them – just make sure you see the overview and the relationships.

1. Print (or keep an electronic copy handy) the 6-page document of Applied Statistics formulas. (Week 11 covers the first five of the eleven rows. Week 12 covers the rows 8, 9, and 10. Rows 6, 7 and 11 are about regression. Regression topics will be covered in the next course in our program and so the formulas are not covered here.) Problems from earlier weeks will appear on homework in later weeks. It is important for you to learn to read a problem and choose the appropriate solution method from ALL those you have learned so far.
2. Go through the lectures for this week.
3. Note that you DO NOT have to compute "by hand" (with a calculator) most of the Standard Error (SE) formulas. (When my elementary stat students try to do these with a calculator, they waste too much time trying to deal with round-off errors that are keeping them from getting exactly the same answer as "the book.") All of that is basically a waste of time. My applets for SEs are at https://visualize.tlok.org/intro-stat/SE/ . In the future, if you had to do very much of this, I'd recommend making yourself a spreadsheet workbook to compute these formulas.
4. **For all material in the course from this point forward**, do NOT statistical tables from other sources. **Use the theoretical distributions in StatKey**. (They provide exactly that precision that is used for grading your answers.)
5. You DO have to have a method of using these formulas to work all the CI and HT problems on single means, single proportions, difference of means, and difference of proportions. Decide what that method will be. I recommend mostly by hand with a calculator and the applet listed in number 2 for the SE calculations. If you want to use statistical software, be sure you understand the various settings and choices you can make in it – and then make them to match the methods of the formulas on our Applied Statistics formula pages.
6. Use that method RIGHT AWAY to re-work the problems 1-6 from Homework 9. Post your questions, solutions, and your thoughts about your results in Piazza. Notice how closely (or not) they agree with your answers using simulation methods. Notice whether each actually STRICTLY meets the conditions to use the theoretical dist'n methods. Discuss this on Piazza! (Notice how such questions will be graded. See the wording of the HW question 16. Discuss this as needed.)
7. For all problems in the course, if you have some ideas about better methods to do various computations (maybe using continuity corrections or something like that) **DON'T use those on the homework (or quiz or exam) problems**. While it may not be clear to you yet, it is true that, in virtually every applied problem, all of our methods are approximations, to some extent, and so the different methods you learn are somewhat different approximations. The machine-graded answers we are using in the course don't allow for all of the possible approximations you might find to use.
8. What are our objectives for you in Weeks 5, 6, 10, 11, and 12?
    a. Carry out hypothesis tests and find confidence intervals on a variety of parameters – all those which StatKey supports – including medians and standard deviations.

b. Clearly understand, and be able to explain to others, the actual meanings of the conclusions of confidence intervals and hypothesis tests and how these arise from probability considerations.

c. Recognize and be able to explain the difficulties we discussed in previous weeks about interpreting the results of hypothesis tests

d. Be able to do HT and CI on all the different parameters in the StatKey menu by simulation and the methods based in theoretical dist'ns which are on our sheets of formulas by using those formulas. Be able to do most (all the relevant ones) BOTH ways. Observe (and think about explaining) differences you see in the results.

Before you start question 7 on HW 10: Matched Pairs Data: Discussion

1. On the StatKey website, read the description of the Wetsuit dataset. (Go to the end of the dataset page, open the pdf file of descriptions, in alphabetical order by title of the dataset, go to Wetsuit, and then read about it.) We want to determine whether this provides evidence of a difference in swimming speeds due to wearing a wetsuit. (Do you suppose this is of interest to competitive swimmers??)
2. Download the dataset. Open it in a spreadsheet. Notice that, if we want to analyze it as "difference of two means" we need it organized differently. That is, we need to lose the information about gender and type of athlete and also organize it as a table with 24 rows of data, like this:

Time	Wetsuit or not

I have attached such a file on the same page from which you downloaded this HW file, called Wetsuit-modification-1

3. Also notice that, if we want to analyze it as "matched pairs" data, then, according to our Applied Statistics formula sheet, we need the differences of the two values for each individual swimmer and with only 12 rows.

Time with wetsuit	Time without wetsuit	Difference	Gender	Type

I have attached such a file on the same page from which you downloaded this HW file, called Wetsuit-modification-2

4. Do two different hypothesis tests: one as in question 2 above and one as in question 3 above. For each of numbers 2 and 3 above, test the hypotheses. Do these give the same answer to our question in question 1 above? If not, discuss which analysis uses all the appropriate information in the data? Discuss this on the Discussion Board.
5. You can also see both analyses, and discussion, in the Lock text, Section 6.5.

Homework 10. Actual Questions

This assignment is organized very differently from previous assignments. **The first 10 questions will not be graded and no solutions will be posted.** These are practice problems for you to discuss freely on the Discussion Board. Use these as models for creating additional practice problems for yourself and obtaining feedback about whether you are working them correctly. For these, report fully on how you did it – not just the numerical answers. That's what is needed for useful discussion.

Only questions 11-18 will appear on the homework to be submitted. Only they will have solutions posted.

When I ask you to use the theoretical dist'n methods, for EACH, check whether the data STRICTLY meet the conditions for the theoretical dist'n method and comment on that when you provide a solution. (Look now at Question 16 near the end of this homework to see the choices for a graded question about conditions.)

1. Redo HW 9 Question 1 using the appropriate theoretical dist'n method of this week and determine whether your answer is consistent with your solution from HW 9.
2. Redo HW 9 Question 2 using the appropriate theoretical dist'n method of this week and determine whether your answer is consistent with your solution from HW 9.
   Three parts:
3. Find the required sample sizes – in whole numbers. **Use StatKey theoretical dist'ns** instead of other statistical tables.
   (**Decimal answers to questions of "find the sample size needed" must always be a whole number no less than the computed number.** That means if they need 22.08 items, they must obtain 23 items because 22 items will not give the needed result.)
   a. In a day-care center, the director is interested in finding a 90% confidence interval, with margin of error 3.5 hours, for the average number of hours per month that a child was actually in attendance. From a preliminary look at 18 randomly selected records, they found that the average was 124.6 hours and the standard deviation was 13.3 hours. How large a sample is needed to find the interval?
   b. Union organizers were considering working in a particular university to form a faculty union. In order to decide whether to pursue this, they decided to survey the faculty members, asking whether, if there were a vote at the current time, the faculty member would expect to vote to form a union. How many faculty members should they survey if they want a 90% confidence interval with a margin of error of 6%?
   c. Union organizers were considering working in a particular university to form a faculty union. In order to decide whether to pursue this, they decided to survey the faculty members, asking whether, if there were a vote at the current time, the faculty member would expect to vote to form a union. How many faculty members should the survey if they want a 90% confidence interval with a margin of error of 6% and they have some evidence that the percentage who would say yes is around 25%?
4. Redo HW 9 Question 4 using the appropriate theoretical dist'n method of this week and determine whether your answer is consistent with your solution from HW 9.
5. From HW9, Question 5. Test the claim that one proportion is greater than the other.

6. Redo HW 9 Question 6 using the appropriate theoretical dist'n method of this week and determine whether your answer is consistent with your solution from HW 9.
7. Topic: Matched Pairs. Look at the "Wetsuit" data from the StatKey website. Analyze the data two different ways and compare the results. Consider various questions. (See page 2 of this handout for further description of how to approach this.)
    a. Is there a difference between the results of the analyses? If so, which seems like a more correct analysis?
    b. When you did the two analyses, did you notice that the data DO NOT STRICTLY meet the conditions to use theoretical methods in some parts?
    c. As you become sufficiently experienced, you might develop modifications of these conditions you are willing to use (and explain, when you are using them, what characteristics you are considering and why.) But, In THIS COURSE, use only the given conditions, VERY STRICTLY. The characteristics of our grading process do not allow for using individual judgment and giving such justifications.

8. Topic: What proportions are relevant?
   In StatKey > Two Categorical Variables, look at the "PainKillers and Miscarriage" dataset in the dropdown menu..
   Question: Find the difference between these two proportions: proportion of women who had taken aspirin and then had a miscarriage and the proportion of women who took no painkiller and then had a miscarriage. (I'm not asking for either a CI or HT here – just asking that you identify the appropriate proportions and find the difference.)
   Comment 1: It is often tricky for people to determine which is the denominator with these questions. If the question is carefully worded, the name that follows this phrase "the proportion of subjects who …" forms the denominator. That is, the proportion is a proportion of that group. Generally speaking, the question involves two independent proportions and then those denominators are not counts of the same group.
   Comment 2: In your work, or consulting, it is possible that the questions are not carefully constructed. Another way of looking at this is to think about which is the "explanatory variable" and which is the "response variable." In that case, you will generally want to compare the conditional dist'ns of the response variable for the different values of the explanatory variable.
9. Topic: What proportions are relevant?
   In StatKey > Two Categorical Variables, look at the "NutritionStudy: Vitamin Use by Smoke" dataset in the dropdown menu. (Or download the NutritionStudy dataset and obtain the information from that.) From there, consolidate the three categories of "Vitamin Use" into just "Yes" and "No."
    a. If you want to investigate the question "Does Vitamin Use affect whether subjects smoke or not?" what two proportions would you compare? Give the description in words and the fractions.
    b. If you want to investigate the question "Does Smoking or not affect whether subjects use vitamins?" what two proportions would you compare? Give the description in words and the fractions.

10. Review the implications of the design of the study on what types of conclusions can be drawn from the results. What type of study allows generalization to the population? Why? What type of study can support conclusions about causality? Why? (Don't get elaborate on this. Think about the basic ideas.)

HW 10: Graded Portion

Each of these 9 questions is worth 1 point.

Use the same instructions for reporting multiple choice values as in the previous weeks about choosing the closest value to your actual answer. For confidence intervals, find the endpoints and then compute the length and report that.

0. This is a graded question in edX. Obviously, the correct answer is Yes.

   In the rest of this master's degree program, you are expected to be able to do (and understand) all of these statistical calculations covered this week, and similar ones. That is MANY formulas. The point is to fully understand the overview, so that you can apply the ideas to all such similar problems. Different students will need different amounts of practice to do this. The purpose of their being ungraded this week is that these are particularly suited to learning from discussion during the week the material is being presented.

   Here is the question:

   Do you understand that questions similar to any or all of these "ungraded" problems in 1-10 may appear on graded assignments in this course and other courses in our program.

   (a) Yes (b) No

11. The manufacturers are interested in estimating the percentage of defective light bulbs coming from a certain process. They want a 97% confidence interval with a margin of error of 3.6%. How many light bulbs must they test?

12. Same question as in the previous problem, but assume they had a reason to believe the proportion is fairly close to 7%. How many light bulbs must they test?

13. An airline has a regular flight between two cities. From a previous study, we estimate the standard deviation of the flight times to be 8.34 minutes. We want a 90% confidence interval for the average flight time with a margin of error of 2.5 minutes. How many flights must they include in the study to find that confidence interval?

Next three questions: **ImmuneTea** dataset on the Lock5 website. Researchers suspect that drinking tea enhances the production of interferon gamma, which is a molecule that enhances the immune system. It was known from previous studies that consumption of coffee has no effect on the amount of interferon gamma present in the human body. A recent study involved 21 healthy adults who did not normally drink either tea or coffee. The subjects were randomly divided into two groups. One group was assigned to drink 5-6 cups of tea a day and the other assigned to drink 5-6 cups of coffee per day. After two weeks, blood samples were drawn and analyzed to determine

the amount of interferon gamma present. The question tested is "Do these data support the claim that presence of interferon gamma is enhanced in tea drinkers when compared to coffee drinkers?"

14. Is the design of this study appropriate to provide evidence about causality?
    Choices: (a) Yes (b) No

15. Carry out a simulation procedure and report the p-value.
    Choices: (a) 0.04 (b) 0.06 (c) 0.08 (d) 0.10 (e) 0.12

16. Do the data strictly meet the conditions to use the theoretical dist'n method? Choices:
    (a) Yes
    (b) Either they might be considered to meet the conditions after additional investigation was done or No.

    (Not required, but recommended. **Not for submission**.
    Carry out a "theoretical dist'n" procedure and compare your result with that of your simulation procedure. Do not submit anything for this. And this should not affect your answer to Question 16 to be submitted.)

17. An appropriately designed and administered survey of adults is done. Among other things, they want to estimate the difference in the average age of married women and the average age of married men in marriages between a woman and a man. Three questions on the survey are "Are you currently married to a person of the opposite sex?" "What is your current age in years?" "What is your gender?" From the data on the survey, they estimate the average difference in ages of married women and married men, among the population from which the sample is drawn of married adults in marriages to a person of the opposite sex. Which of these is the correct method of analysis?
    Choices:
    (a) Hypothesis test on a difference of independent means
    (b) Hypothesis test on matched pairs
    (c) Confidence interval for a difference of independent means
    (d Confidence interval for the difference of means from matched pairs.

18. An appropriately designed and administered survey is done of married couples in a marriage between a woman and a man. Two questions on the survey are "What is your current age in years?" "What is your gender?" From the data from the survey, they will estimate the average difference in the ages of the woman and man in a married couple in the population from which the sample is drawn. Which of these is the correct method of analysis?
    Choices:
    (a) Hypothesis test on a difference of independent means
    (b) Hypothesis test on matched pairs
    (c) Confidence interval for a difference of independent means
    (d Confidence interval for the difference of means from matched pairs