Chapter 20. Inference on One Proportion

1. What is a proportion and how is it different from a mean?
2. What is the difference between the symbol p and the symbol $\hat{p}$ in the material in this chapter?
3. What does the sampling distribution of a proportion look like? And what are the mean and the standard deviation of the sampling distribution of a proportion?
4. What are the conditions needed to apply the techniques of this chapter to problems?
5. Do we use the t-procedures in this chapter? Why or why not?
6. What kinds of questions do we answer about proportions?
    a. If a population proportion is 0.70, what is the probability that the sample proportion for a sample of size 50 is above 0.80?
    b. Find the confidence interval for a population proportion.
    c. Test a hypothesis about a population proportion.
    d. How large a sample do we need to have a confidence interval for the population proportion with margin of error less than or equal to 0.02?
7. Do I need to draw a picture of the sampling distribution of the sample proportion for every hypothesis test on one proportion?
8. Is there anything special to know about using software for inference on one proportion?
9. How do these formulas fit the overall pattern of these four chapters?
10. Additional practice in class, as we have time: 20.26 (Use software and do it both with large-sample methods and plus-four methods.) 20.42

Discussion:

1. What is a proportion and how is it different from a mean?

    Activity 1: Look at three exercises from Chapter 20, such as 20.25, 20.26, and 20.29. For each exercise, say what the two outcomes are, what the population of interest is, and which proportion the question is about. Discuss whether you understand why someone might be interested in the question that is asked about the population proportion.

    Summary: When the original data can only take on two values – often non-numerical values - (e.g. male / female, or yes/no, or head / tail, or some other pair of values) then we name the one we are most interested in a "success" and summarize the sample data by giving the proportion of successes in the sample. In contrast, when we discussed population means, the original data took on various numerical values and so we computed sample averages (means) of those values. Read several exercises in Chapter 14, 15, or 18 for examples.

2. What is the difference between the symbol p and the symbol $\hat{p}$ in the material in this chapter?

   Answer: The population proportion is called p and the sample proportion is

   $$\hat{p} = \frac{\text{number of successes}}{\text{number of trials}}.$$

   Activity 2: For each of exercises 20.25, 20.26, and 20.29, identify what the problem calls a "success" and give the sample proportion in both fraction form and in decimal form with at least three decimal places.

3. What does the sampling distribution of a proportion look like? And what are the mean and the standard deviation of the sampling distribution of a proportion?

   Activity 3: As a class, we will start to approximate a sampling distribution of the sample proportion of "heads" when tossing a penny from your pocket, for samples of size 20. Each student should do 20 tosses and report the number of heads from those.

   Activity 4: Software Simulation. Use http://www.lock5stat.com/statkey/, choose Sampling Distribution, One Proportion, to simulate at least 2000 students doing this same activity for samples of size 20. (We will assume the coin is "fair" so the population proportion of heads is 0.5.) Write here the shape, center (mean) and spread (standard deviation) of that sampling distribution.

   Approximately how far out in the tails would a sample proportion have to be in order to be "unusual" in this distribution? (Answers should be values like 0.85 or 0.10 – that is, possible values for sample proportions. Different students will have different answers.)

   Activity 5: Software Simulation. Re-do Activity 4, but with a sample size of 100. (Do you see why I am not asking you to do one of these samples by hand first?)

   Activity 6: Look in the text in the section "The Sample Proportion $\hat{p}$" on pages 494-495. Do you see that this exact formula tells you that the mean (center) of both sampling distributions in Activities 4 and 5 would be 0.500 if we found the entire sampling distribution rather than just using software simulation to get 2000 values from it? You might want to get up to 10,000 samples or so and see if that gets even closer to 0.5.

1) Do you see that the text tells you what the standard deviation of the sampling distribution should be for each of those sample sizes? That's the SE.

2) Compute the SE for size 20 using the formula.

3) How close did your software simulation get to that?

4) Do you understand why you wouldn't necessarily expect your software simulation to get the exact value given by the formula?

5) Compute the SE for size 100 using the formula.

6) How close did your software simulation get to that?

Activity 7: Work exercise 20.2 in the text, from this same section. Do not use the software to simulate. Just draw the picture of the shape you have learned, use the formulas to compute the mean and standard deviation and put them on your picture. Then use the 68-95-99.7 rule to answer the question.

(Answer: $0.68 \pm 2 \cdot 0.0135$, which is 0.653 to 0.707)

4. What are the conditions needed to apply the techniques of this chapter to problems?

   Answer: The sample must be an SRS, and it must be "large enough." Chapters 20 and 21 are a little tricky. "Large enough" is different for the different techniques.

5. Do we use the t-procedures in this chapter? Why or why not?

   Answer: **No.** We use the normal distribution here, not t-procedures. (page 497.) The reason is that we are using the sample to estimate only one value – the proportion. We needed the t-distribution before because we were using the data to estimate two parameters – μ and σ.

6. What kinds of questions do we answer about proportions?

   a. If a population proportion is 0.70, what is the probability that the sample proportion for a sample of size 50 is above 0.80? (The equivalent questions for means are in Ch. 11.)

      Solution: Draw the normal distribution, label it with the mean (0.70) and the standard deviation (using the formula), put the score of 0.80 on the picture, shade in the area, and use a normal distribution calculation to answer the question. (Answer: z=1.54, probability = 0.0618.)

   b. Find the confidence interval for a population proportion.

      Large- sample: If both the successes and failures in the sample are at least 15, use the very straightforward formulas. See page 497 "Large-Sample Confidence Intervals …" Example 20.4.

Activity 8: More accurate confidence intervals: It is more accurate to use an approximation that, essentially, adds four trials to the data, 2 of which are successes and 2 of which are failures. See the section called "Accurate confidence intervals…" pages 499-501. These are called "plus-four" methods and can be used for any number of successes and failures, as long as the the total sample size is at least 10 and the confidence level is at least 90%. This is a LOT less strict on the sample size, so it is much more useful as well as more accurate. Do Exercise 20.7 by following Example 20.5.

Which to use: See recommendations at the beginning of the problems at the end of the chapter.

c. Test a hypothesis about a population proportion.

Activity 9: See the section called "Significance tests for a proportion. Notice that the condition about sample size here DOES NOT use the number of successes, but uses the sample size AND the value of p_O in the Ho. Do Exercise 20.13 by following Example 20.7. Pay careful attention to the State and Plan steps. Be sure to draw your picture in the Solve step.

d. How large a sample do we need to have a confidence interval for the population proportion with margin of error less than or equal to 0.02? (The equivalent questions for means are in Ch. 16.)
Answer: This is rather similar to what we did for means in Chapter 16 EXCEPT that we don't have a value for the standard deviation, because the standard deviation depends on the unknown p . So, in the formula, we have to estimate p . We do that EITHER by using 0.5 if we don't have any better information, or, if we have the result of a prior sample or something, then we use that result.

Activity 10: I have a light-bulb manufacturing plant and want to estimate the proportion of defective light-bulbs are coming out of the manufacturing process. I want 95% confidence and a margin of error of no more than 0.01.

a. What is the sample size I need if I don't know anything about the population proportion?

b. But, of course, my manufacturing process must be making mostly good bulbs. So let's assume the population proportion of defectives is approximately 0.04. Use that value as a guessed value for p and find the sample size needed.

c. What do you see about the value of doing a preliminary sample?

7. Do I need to draw a picture of the sampling distribution of the sample proportion for every hypothesis test on one proportion?

Answer: Yes. The pictures are very similar to those in Chapters 14, 15, and 18, except that now our statistic is $\hat{p}$ instead of $\bar{X}$. I think it is very important that you draw these pictures for all the one-sample problems, in order to keep it clearly in mind.

8. Is there anything special to know about using software for inference on one proportion?

   Answer: We only use software to find confidence intervals and do hypothesis tests. We do not use it to find probabilities in the sampling distribution, nor to find sample sizes needed to get a given margin of error.

   Both Minitab and CrunchIt. There **are no data files** for the proportion problems in Ch. 20 and 21. So you must open the software directly rather than clicking on the dataset icon from the text.
   The data files would just be a long list of all the individual data values, and it is much easier to simply tell you the "number of successes" and the "number of trials." When you look in the software, you will choose "summarized data" so that you can enter those.
   The software does give you a place to say what column the data was in, if you were ever given the full data set.

   CrunchIt: Statistics > Proportion > 1-sample
   Fill in the boxes as you did when you learned about doing confidence intervals and hypothesis tests in Chapter 18.

   Minitab: Stat > Basic Statistics > 1-Proportion
   Fill in the boxes as you did when you learned about doing confidence intervals and hypothesis tests in Chapter 18 here. Use the Options button. In Options, DO check the box which says to "Use test and interval based on the normal distribution."

9. How do these formulas fit the overall pattern of these four chapters?

   Answer: Just as before, these formulas follow these patterns:

Confidence interval	Hypothesis test
$\text{statistic} \pm \ z \cdot \text{SE}$	$z = \dfrac{\text{statistic} - \text{parameter from } H_O}{\text{SE}}$

10. Additional practice in class, as we have time: 20.26 (Use software and do it both with large-sample methods and plus-four methods.) 20.42

Quiz 11: Due at the beginning of class Wed. April 10.

19.40, 20.2, 20.28, 20.34, 20.38
As you do your homework, be sure to do enough "by hand" to make sure that you can find t-scores and use the t-table and the normal table appropriately on Test 4. Then you can do the rest using software.