

# Introduction to Inference

## Confidence Intervals for Proportions

1

## Estimating with Confidence

- On the one hand, we can make a general claim with 100% confidence, but it usually isn't very useful; on the other hand, we can also make claims that are very specific, but have little to no confidence in the claim. There is always tension between certainty and precision. Fortunately, in most cases, we can be both sufficiently certain and sufficiently precise to make useful statements.
- There is no simple answer to the conflict. **You must choose a confidence level yourself.** The data can't do it for you. The choice of the confidence level is somewhat arbitrary, but the most common levels are 90%, 95%, and 99%. Although *any* percentage can be used, percentages such as 92.3% or 97.6% are suspect and people will think that you're up to no good.

4

## Estimating with Confidence

When we select a sample, we know the responses of the individuals in the sample. Often we are not content with information about the sample. We want to *infer* from the sample data some conclusion about a wider population that the sample represents.

**STATISTICAL INFERENCE**

Statistical inference provides methods for drawing conclusions about a population from sample data.

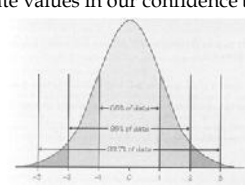
**THE CONCEPT OF CONFIDENCE**

If you randomly select a sample from the population, we know that the statistic will vary. The statistic can be close to the center or far from it. We're not completely sure, but as long as our sample is large enough, we know that it will fall somewhere on the Normal curve (by the CLT). If you use this statistic, how confident can you be that it is a good representation of the population parameter? This is the idea that we will build on to create "Confidence Intervals."

2

## Estimating with Confidence


- Let's recall an old friend that will be useful in dealing with confidence: the 68-95-99.7 rule.
  - This informal rule can help us to make a couple of quick and easy generalizations. However, using a table or technology will help us determine more appropriate values in our confidence testing.



5

## Estimating with Confidence

- Consider the following statements and determine how confident you can be in each claim:
  - I am positive that the \_\_\_ will win the World Series this year!
  - Tomorrow, it's not going to rain.
  - A Democrat is definitely going to win the next Presidential election



- Tomorrow's high temperature in San Francisco will be between 55 degrees Fahrenheit and 75 degrees Fahrenheit.

3

## Statistical Confidence (Example)

- In May of 2007, a Gallup Poll found that in a random sample of 1003 adults in the United States, 110 approved of attempts to clone humans (or about 11%). From this sample, what can we say about how adults in America feel about cloning?
- Since this data comes from a sample, we must use a particular notation to make sure that everyone knows that we have a proportion from a sample...

$$\hat{p} = \frac{110}{1003} \approx 0.11$$

6

### Statistical Confidence (Example)

- It is important to note that the data gathered was collected from only one sample. If we were to gather all possible samples from all of the adults in the US, we would have a Normal distribution (under certain conditions, of course...do you know what the conditions are? There's two of them).
- Where does this sample fall on the Normal curve?
- Does it fall on the high end, the low end, or right in the center?
- How confident are you that this sample represents all of the adults in the US? Using this sample, can we say that 11% of all US adults support cloning?

7

### Statistical Confidence (Example)

- Great, now what does that tell us?
  - ❖ Since  $\hat{p} = 0.11$  and  $SE_{\hat{p}} = 0.01$ 

$$\hat{p} \pm 2SE_{\hat{p}}$$

$$= 0.11 \pm 2(0.01)$$

$$= 0.11 \pm 0.02$$

We get an interval from  
0.09 to 0.13
- What does this mean in context of this problem?
  - ❖ In order to answer this question, we must be very careful and choose our words wisely...

10

### Statistical Confidence (Example)

- If we were to say that 11% of all US adults support cloning, our confidence would be extremely low since we're basically saying that the mean of our sample is exactly on the center of our sampling distribution which is not likely.
- So what do we do? We come up with a range that we are somewhat confident will contain the true parameter. The standard deviation of this specific sampling distribution is about 0.01 or 1%. From the sampling distributions point of view, if we go two standard deviations to the left or right of the true proportion, we will have 95% of all the possible samples. From the sample's point of view, if we go two standard deviations from the sample's proportion, we have a 95% chance of capturing the true parameter

8

### Statistical Confidence (Example)

- Correct language is an absolute must here. Here are a list of things that people like to say:
  - ❖ *"11% of all US adults support cloning."*
    - **WRONG!!!** It would be nice to be able to make this absolute claim, but we just don't have enough information to do that.
  - ❖ *"It is probably true that 11% of all US adults support cloning."*
    - **WRONG!!!** Whatever the true parameter may be, it is more than likely not going to be 11% exactly.
  - ❖ *"We don't know the exact proportion of US adults that support cloning but we know that it is in the interval 11% plus or minus 2% or between 9% and 13%."*
    - **WRONG!!!** This is closer but we don't know anything about the parameter for certain.

11

### Statistical Confidence (Example)

- Great, now what does that tell us?
  - ❖ Given any sample within our distribution, we have a 95% chance that it will be within the following range:  $\hat{p} \pm 2SE_{\hat{p}}$
  - ❖ Where  $\hat{p}$  is the sample statistic and  $SE_{\hat{p}}$  is something called the **standard error**. Why don't we just call it the standard deviation of the sampling distribution?
    - In order to find the standard deviation of the sampling distribution, we need to know the population's parameter. Since we don't know this (and can not know this without doing a census), we find the standard deviation using the sample statistic...and since we can't call it the standard deviation of the sampling distribution, we call it the standard error of the sampling distribution.
- In any case, I'm 95% sure that the population parameter will be within my grasp. Now, I've got him! Probably.

9

### Statistical Confidence (Example)

- Correct language is an absolute must here. Here are a list of things that people like to say:
  - ❖ *"We don't know the exact proportion of US adults that support cloning, but the interval from 9% and 13% probably contains the true proportion."*
    - **Correct, but not the best way to say it!!!** This statement is correct, but it is not the best statement. It is a bit too wishy-washy. We would like to quantify the word "probably."
  - ❖ *"We are 95% confident that between 9% and 13% of US adults support cloning."*
    - **YES!!!** This statement is called a confidence interval and it is the best that we can do.

12

### Confidence Intervals

- A level C confidence interval for a parameter has two parts:
  - ❖ An interval calculated from the data, usually in the form of: estimate  $\pm$  margin of error
  - ❖ A confidence level C, which gives the probability that the interval will capture the true parameter value in repeated samples.

### Critical Values

- We used the 68-95-99.7 rule to obtain a 95% confidence within two standard deviations – but this is just an informal rule. For 95% confidence, a more accurate z-score would be 1.96 standard deviations to the left and right. Our critical value would actually be  $z^* = 1.96$ .
- Take a look at Table C on the Statistics chart. By utilizing this table, we can find all of the  $z^*$  for a variety of specified Confidence level C.

### Confidence Intervals

- There are two assumptions that must be met:
  - ❖ **Independence Assumption** – Once again, since there is no way to check this for sure, we check it with two conditions:
    - ⇒ **Randomization condition** – the data come from a random sample or suitably randomized experiment.
    - ⇒ **10% condition** – the sample is no more than 10% of the population
  - ❖ **Normal Population/Sample Size Assumption** – We know that according to the CLT that the sampling distribution will be approximately normal as long as the sample is **large enough**:  $np \geq 10$  and  $nq \geq 10$ 
    - ⇒ **Success/failure condition** – we must expect that there will be at least 10 “success” and at least 10 “failures”.

### Confidence Intervals

Draw an SRS of size  $n$  from a population having **unknown** proportion  $p$  and a **unknown** standard deviation  $\sigma$ . A level C confidence interval for  $p$  is  $\hat{p} \pm ME$

ME is the Margin of Error  $ME = z^* SE(\hat{p})$

SE is the Standard Error  $SE(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{n}}$

Another way to write CI would be:  $\hat{p} \pm z^* \sqrt{\frac{\hat{p}\hat{q}}{n}}$

### Critical Values

- The critical value  $z^*$  with probability  $p$  lying to its right under the standard normal curve is called the upper  $p$  critical value of the standard normal distribution. It basically tells us how many standard deviations to the right or to the left we are from the mean for a particular confidence level.



### The 4 Step Process: C.I

- **Step 1: Determine what the question is asking** and state what you want to know. Be sure to identify the **population** of interest and the **parameter** of which you wish to draw conclusions.
- **Step 2: Choose the appropriate inference procedure.** Verify the conditions for using the selected procedure.
- **Step 3:** State the procedure that you will use. If the conditions are met, carry out the inference procedure. **Do the work!**
  - ❖ CI = estimate  $\pm$  margin of error
- **Step 4: Interpret your results in the context of the problem.**

## STARTER Ch. 19

- In January 2007, Consumer Reports conducted a study of bacteria in frozen chicken sold in the US. They purchased a random selection of 525 packages of frozen chicken of various brands from different food stores in 23 different states. They tested them for various types of bacteria that cause food-borne illnesses. They found that 83% were infected with **Campylobacter** and 15% were infected with Salmonella.
- Construct a 95% CI (or Confidence Interval) for the proportion of chickens infected with **Campylobacter**.

19

## Checking for understanding

- Construct a 95% CI (or Confidence Interval) for the proportion of chickens infected with Campylobacter.
- Follow the FOUR-STEP PROCESS...
  - ❖ **Third**, state the parameters and **show your work** – since we know that we satisfy our conditions, we will have an approximately normal distribution.
    - ⇒ The sample proportion was given:  $\mu = \hat{p} = .83$
    - ⇒ The standard deviation can be found using the formula:
 
$$\sigma_{\hat{p}} = \sqrt{\frac{\hat{p}\hat{q}}{n}} = \sqrt{\frac{(.83)(.17)}{525}} \approx .0164$$
    - ⇒ Provide a graph and solve
 
$$CI = \hat{p} \pm z^*SE = 0.83 \pm (1.96)(0.0164) = (0.7979, 0.8621)$$

22

## Checking for understanding

- Construct a 95% CI (or Confidence Interval) for the proportion of chickens infected with Campylobacter.
- Follow the FOUR-STEP PROCESS...
  - ❖ **First**, state what you want to know and determine what the question is asking
    - ⇒ We want to find an interval that is likely, with 95% confidence, to contain the true proportion,  $p$ , of frozen chickens that are infected with Campylobacter.

20

## Checking for understanding

- Construct a 95% CI (or Confidence Interval) for the proportion of chickens infected with Campylobacter.
- Follow the FOUR-STEP PROCESS...
  - ❖ **Fourth**, last but not least, state your conclusion in context of the problem:
    - ⇒ We are 95% confident that between 79.79% and 86.21% of all frozen chicken sold in the US are infected with Campylobacter.
    - ⇒ OR
    - ⇒ We are 95% confident that all frozen chicken sold in the US infected with Campylobacter lies between 79.79% and 86.21% .

23

## Checking for understanding

- Construct a 95% CI (or Confidence Interval) for the proportion of chickens infected with Campylobacter.
- Follow the FOUR-STEP PROCESS...
  - ❖ **Second**, examine the assumptions and check the conditions:
    - ⇒ **Independence Assumption**
      - + Randomization condition: We are given that the sample is a random selection
      - + 10% condition: Of all the possible packages of frozen chicken, there are probably more than 5250 packages, so it is safe to assume that the samples are independent.
    - ⇒ **Normality (Large Enough Sample) Assumption**
      - + Success/Failure condition:  $np = (525)(.83) \approx 436$  and  $nq = 525(.17) \approx 89$ . Both are greater than 10.

21

## Checking for understanding

- A spokesperson for the US Department of Agriculture dismissed the Consumer Reports finding, saying, "That's 500 samples out of 9 billion chickens slaughtered a year...With the small number they [tested], I don't know that one would want to change one's buying habits." Is this criticism valid? Why or why not?
- **The size of the population is irrelevant!!!** If Consumer Reports had a random sample, 95% of all intervals generated by studies like this are expected to capture the true contamination level.
- Now it's your turn, construct a 95% CI for the proportion of chickens infected with Salmonella. (Recall that 15% of our sample was infected with Salmonella).

24

### Checking for understanding

- Construct a 95% CI (or Confidence Interval) for the proportion of chickens infected with Salmonella.
  - ❖ **First, state what you want to know and determine what the question is asking**
    - ⇒ We want to find an interval that is likely, with 95% confidence, to contain the true proportion,  $p$ , of frozen chickens that are infected with Salmonella.

25

### Checking for understanding

- Construct a 95% CI (or Confidence Interval) for the proportion of chickens infected with Salmonella.
- Follow the **FOUR-STEP PROCESS...**
  - ❖ **Fourth**, last but not least, **state your conclusion** in context of the problem:
    - ⇒ We are 95% confident that between 11.9% and 18.1% of all frozen chicken sold in the US are infected with Salmonella.
    - ⇒ OR
    - ⇒ We are 95% confident that all frozen chicken sold in the US infected with Salmonella lies between 11.9% and 18.1% .

28

### Checking for understanding

- Construct a 95% CI (or Confidence Interval) for the proportion of chickens infected with Salmonella.
  - ❖ **Second, examine the assumptions and check the conditions:**
    - ⇒ **Independence Assumption**
      - + Randomization: We are given that the sample is a random selection
      - + 10% condition: Of all the possible packages of frozen chicken, there are probably more than 5250 packages, so it is safe to assume that the samples are independent.
    - ⇒ **Normality (or Large Enough Sample Assumption)**
      - + Success/Failure condition:  $np = (525)(.15) \approx 79$  and  $nq = 525(.85) \approx 446$ . Both are greater than 10.

26

### Choosing the sample size

You may need to choose a sample size large enough to achieve a specified margin of error. However, because the sampling distribution of  $\hat{p}$  is a function of the population proportion  $p$  this process requires that you guess a likely value for  $p$ .

$$p \sim N\left(\hat{p}, \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) \rightarrow n = \hat{p}(1-\hat{p})\left(\frac{z^*}{ME}\right)^2$$

The margin of error will be less than or equal to  $ME$  if  $\hat{p}$  is chosen to be 0.5.

Remember, though, that sample size is not always stretchable at will. There are typically costs and constraints associated with large samples.

29

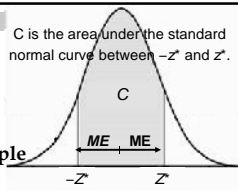
### Checking for understanding

- Construct a 95% CI (or Confidence Interval) for the proportion of chickens infected with Salmonella.
  - ❖ **Third**, state the parameters and **show your work** – since we know that we satisfy our conditions, we will have an approximately normal distribution.
    - ⇒ The sample proportion was given:  $\hat{p} = .15$
    - ⇒ The standard deviation can found using the formula:
 
$$\sigma_{\hat{p}} = \sqrt{\frac{\hat{p}q}{n}} = \sqrt{\frac{(.15)(.85)}{525}} \approx .0156$$
    - ⇒ Provide a graph and solve
 
$$CI = \hat{p} \pm z^* SE = 0.15 \pm (1.96)(0.0156) = (0.1194, 0.1806)$$

27

### CI Need To Know...

C is the area under the standard normal curve between  $-z^*$  and  $z^*$ .



- For a given sample size, higher confidence means a larger **ME**.
- Size of interval is based on sample size and level of confidence
- Larger sample size = smaller interval, smaller errors, less variable  $\phi$  smaller ME (more accurate/more confident that a given CI succeeds in catching the population proportion)
- Large confidence level = larger (wider) interval  $\phi$  need more room for error

30

### BAD EXAMPLES

Conditions

- ① Random: stated as a random sample (SRS) ✓
- ② Success/Failure: at least 10 success/failure ~~X~~ -1 -1
- ③ 10% Condition: 75 is less than all possible locations ~~X~~

All conditions have been met. \_\_\_\_\_ -1.5

CI:  $\hat{p} \pm z^* \sqrt{\frac{\hat{p}\hat{q}}{n}}$  -2  
 (0.088; 0.259) ✓  
 95% of samples are between 8.8% and 25.9%  
 $\hat{p}$  ~~X~~ -2

### BAD EXAMPLES

10% Condition ✓  
Random ✓  
Success/Failure ✓  
Large Enough ✓

Statement -2

CI:  $\hat{p} \pm z^* \sqrt{\frac{\hat{p}\hat{q}}{n}}$  -2  
 (0.101; 0.245) -1  
 The proportion of infected locations is between 10.1% and 24.5%. -1

### BAD EXAMPLES

Conditions

- ① 10% Condition: 75 locations is less than all possible locations in the county's postal routes ~~X~~ -1/2
- ② Randomness: stated as a random sample
- ③ Success/Failure:  $n\hat{p} = 13 \geq 10$  At least 10 successes/failures.  
 $n\hat{q} = 62 \geq 10$

All conditions have been met to use the Normal model for  $\pm$  prop z-interval.

$\hat{p} = \frac{13}{75}$   $n = 75$   $z^* = 1.96$

CI:  $\hat{p} \pm z^* \sqrt{\frac{\hat{p}\hat{q}}{n}}$   
 (0.08766; 0.259)

I am 95% confident that the actual proportion of all possible locations that are infected lies between 8.8% and 25.9%

### BAD EXAMPLES

(c)

$\hat{p} \pm z^* \sqrt{\frac{\hat{p}\hat{q}}{n}}$  -2  
 (0.08766; 0.259)

- ① 10% Condition: 75 locations is less than 10% ~~X~~ -1
- ② Random: Representative sample
- ③ Success/Failure:  $n\hat{p} = 13 \geq 10$  -1/2  
 $n\hat{q} = 62 \geq 10$

Statement -2

95% of samples are between 8.8% and 25.9% ~~X~~ -2