Estimating with Confidence: Developing Students' Understanding

Beth Chance and Allan Rossman
Department of Statistics, Cal Poly – San Luis Obispo
bchance@calpoly.edu, arossman@calpoly.edu

presented for California Mathematics Council – Community Colleges
Fall Conference

December 6, 2014
Monterey, CA

Example 1: Kissing Right? (adapted from *Introduction to Statistical Investigations* by Tintle et al., www.math.hope.edu/isi)

Most people are right-handed, and even the right eye is dominant for most people. Developmental biologists have suggested that late-stage human embryos tend to turn their heads to the right. In a study reported in *Nature* (2003), German bio-psychologist Onur Güntürkün conjectured that this tendency to turn to the right manifests itself in other ways as well, so he studied kissing couples to see which side they tended to lean their heads while kissing. He and his researchers observed kissing couples in public places such as airports, train stations, beaches, and parks. They were careful not to include couples who were holding objects such as luggage that might have affected which direction they turned. For each kissing couple observed, the researchers noted whether the couple leaned their heads to the right or to the left. They observed 124 couples, age 13-70 years.

We will first use Güntürkün's data to test his conjecture that kissing couples tend to lean their heads to the right. Use the symbol π to denote the proportion of all kissing couples in these countries that lean their heads to the right, or the probability that a randomly selected couple leans to the right.

a) Is π a parameter or a statistic? Explain how you are deciding.

b) Do we know the exact value of π based on the observed data? Explain.

c) State the appropriate null and alternative hypotheses, both in words and in terms of the parameter π, for testing the conjecture that kissing couples *tend to* lean their heads to the right.

Of the 124 couples observed, 80 leaned their heads to the right while kissing.

d) Calculate the sample proportion of the observed couples who leaned their heads to the right while kissing. Also indicate the symbol used to denote this value.

e) Conduct a simulation analysis (using the **One Proportion** applet) to assess the strength of evidence that the sample data provide for Güntürkün's conjecture that kissing couples tend to lean their heads to the right more often than they would by random chance. Report the approximate p-value and summarize your conclusion about this strength of evidence.

Your simulation analysis should convince you that the sample data provide very strong evidence to believe that kissing couples lean their heads to the right more than half the time in the long run. That leads to a natural follow-up question: How much more than half the time? In other words, we have strong evidence that the long-run probability of leaning to the right is greater than one-half, but can we now estimate the value for that probability? We will do this by testing many different (null) values for the probability that a couple leans to the right when kissing.

f) Now test whether the data provide evidence that the long-run probability that a couple leans their heads to the right while kissing (π) is **different** from 0.60. Use the **One Proportion** applet

to determine the p-value for testing the null value of 0.60. Report what you changed in the applet and report your p-value.

Recall that a p-value of 0.05 or less indicates that the sample data provide strong evidence against the null hypothesis and in favor of the alternative hypothesis. Thus, we can *reject* the null hypothesis when the p-value is less than or equal to 0.05. Otherwise, when the p-value is greater than 0.05, we do not have strong enough evidence against the null hypothesis and so we consider the null value to be a *plausible* (i.e., believable) value for the parameter.

g) Is the p-value for testing the null value of 0.60 less than 0.05? Can the value 0.60 be rejected, or is the value 0.60 plausible for the long-run probability that a couple leans their heads to the right while kissing?

Recall that the 0.05 criterion we are using is called the significance level. The p-value you found in the previous question should not have been smaller than 0.05. Hence, you do not reject the null hypothesis at the 0.05 level of significance and therefore you do not reject 0.60 as a plausible value for π. Thus, it is plausible that the long-run probability that a kissing couple leans their heads to the right is 0.60.

h) Does this mean that you've *proven* that exactly 60% of kissing couples lean right? Is it probable that the population parameter equals 0.6? Why or why not?

Because there are still other plausible values, now we want to "zoom in" on which values for the long-run probability are plausible and which can be rejected at the 0.05 significance level.

i) Use the applet to test the probability values given in the following table.

- Each time, change the **Probability of success** to match the value that you are testing (keeping the observed sample proportion that you count beyond the same).
- Everything else should stay the same; press **Draw Samples** and then **Count** to see the new two-sided p-value (with the **Two-sided** box checked).

Probability under H_0	**0.54**	**0.55**	**0.56**	**0.57**	**0.58**	**0.59**	**0.60**
(Two-sided) p-value							
Reject or plausible?							
Probability under H_0	**0.70**	**0.71**	**0.72**	**0.73**	**0.74**	**0.75**	**0.76**
(Two-sided) p-value							
Reject or plausible?							

j) Using a 0.05 significance level and your results from above, provide a list of plausible values for π, the long-run probability that a kissing couple leans their heads to the right. (This is called a **95% confidence interval** for π.)

Example 2: American Exceptionalism (adapted from *Introduction to Statistical Investigations* by Tintle et al., www.math.hope.edu/isi)

The Gallup organization conducted a survey with a random sample of 1019 adult Americans on December 10-12, 2010. They found that 80% of the respondents agreed with the statement that the United States has a unique character that makes it the greatest country in the world.

a) Identify the population and sample in this survey.

b) Is it reasonable to believe that the sample of 1019 adult Americans is representative of the larger population? Explain why or why not.

c) Explain why 80% is a statistic and not a parameter. What symbol would you use to represent it?

d) Identify (in words) the parameter that the Gallup organization was interested in estimating.

e) Is it reasonable to conclude that exactly 80% of all adult Americans agree with the statement about American exceptionalism? Explain why or why not.

f) Although we expect π to be close to 0.80, we realize there may be other plausible values for the population proportion as well. First consider the value of 0.775. Is this a plausible value for π? Use the **One Proportion** applet to simulate 10,000 random samples of 1019 people from such a population. (*Hint*: Keep in mind that 0.775 is what we are assuming for the population proportion and 0.80 is the observed sample proportion.) What do you estimate for the two-sided p-value? Would you reject or fail to reject the null hypothesis at the 5% level of significance?

g) Also check the **Summary Stats** box and report the mean and standard deviation of this null distribution.

h) Now consider 0.5. Is this a plausible value for π? Repeat the above and record the mean and standard deviation for this *null distribution* of sample proportions as well. Comment on how the null distribution of sample proportions has changed.

Clearly 0.5 is going to be "too far" from $\hat{p} = 0.80$ to be plausible. But how far is too far? We could use the Plausible Values Method to produce a confidence interval for the proportion of all adult Americans who agree with the statement. But that approach is somewhat cumbersome and time-intensive, so we'll now learn some shortcut approaches. Keep in mind that our driving question is whether our observed statistic falls in the tail of the null distribution corresponding to the hypothesized parameter value. So our main concern is really how spread out the null distribution is. Another way to measure how unusual our observed result is would be to standardize.

i) Reconsider our first guess of $\pi = 0.775$. How many standard deviations is 0.80 from 0.775? (*Hint*: Standardize the value by looking at the difference between 0.775 and 0.80 and divide by the standard deviation of the null distribution.)

You should notice that 0.775 and 0.80 are about 2 standard deviations apart AND that the two-sided p-value is around 0.05, so this value is close to the edge of values that can be considered plausible. Values between 0.80 and 0.775 are considered plausible and values smaller than 0.775, or more than two standard deviations below 0.80, will not be plausible values for the population proportion.

Key Idea: When a distribution is bell-shaped, as your null distribution should be for this study, approximately 95% of the statistics in the null distribution will fall within two standard deviations of the mean. This implies that 95% of sample proportions will fall within two standard deviations of the long-run probability (π), which means that π is within two standard deviations of the observed sample proportion for 95% of all samples.

We can then extend this idea to construct a 95% confidence interval.

Key idea: We can construct a 95% confidence interval of plausible values for a parameter by including all values that fall within 2 standard deviations of the sample statistic. This method is only valid when the null distribution follows a bell-shaped, symmetric distribution. We call this the ***2SD Method***. Thus we can present the 95% confidence interval for the long-run probability (or population proportion), π, in symbols as:

$$\hat{p} \pm 2 \times \text{SD}(\hat{p})$$

where $\hat{p}$ is the sample proportion and SD($\hat{p}$) is the standard deviation of the null distribution of sample proportions. The value of 2 × SD, which represents half the width of the confidence interval, is called the ***margin-of-error*** for 95% confidence.

This leads to the next question – how do we find the standard deviation of the sample proportions, SD($\hat{p}$)?

j) How did the null distribution standard deviations you found in with $\pi = 0.775$ and with $\pi = 0.5$ compare?

You should see that the standard deviation changes slightly when we change π, but not by much. The variability in the sample proportions is in fact largest when $\pi = 0.5$. So one conservative approach would be to carry out one simulation (with lots of trials) using $\pi = 0.5$, and use that value of the standard deviation of that null distribution to approximate the margin-of-error.

k) Determine a 95% confidence interval for π using the 2SD Method. First calculate 2 × (standard deviation for your null distribution of sample proportions) using 0.5 in the simulation to estimate the SD. Use this SD to approximate a 95% confidence interval for π. (*Hint*: Subtract this margin-of-error from $\hat{p}$ to determine the lower endpoint of the interval and then add this margin-of-error to $\hat{p}$ to determine the upper endpoint of the interval.)

l) In fact, we can further simplify this approximation for the margin-of-error to be $1/\sqrt{n}$. Calculate this value and compare to your calculation in k).

m) Using your answer to (l), if I wanted my margin-of-error to be .01, approximately how large does my sample size need to be?

n) Interpret this confidence interval: You are 95% confident that *what* is between what two values?

One limitation to this method is that it only applies to 95% confidence. What if we wanted to be 90% or 99% confident instead? We can extend this 2SD method to a more general theory-based approach. Instead of assuming the worst case scenario for the standard deviation of $\hat{p}$ using π = .5, we can use the sample proportion.

Definition: An estimate of the standard deviation of a statistic, based on sample data, is called the ***standard error (SE)*** of the statistic. In this case $\sqrt{\hat{p}(1-\hat{p})/n}$ is the standard error of a sample proportion $\hat{p}$.

o) Calculate the standard error for this study. How does it compare to the standard deviations you found above?

So, a more general formula for using the 2SD Method to estimate a population proportion is: $\hat{p} \pm 2\sqrt{\hat{p}(1-\hat{p})/n}$. But then how do we change the confidence level?

The 2SD Method was justified by saying 95% of samples yield a sample proportion within 2 standard deviations of the population proportion. If we want to be more confident that the parameter is within our margin-of-error, we can create a larger margin-of-error by increasing the multiplier. In fact a multiplier of 2.576 gives us a 99% confidence level, whereas a multiplier of 1.645 gives us only 90% confidence.

Key idea: The one proportion *z*-interval for π is: $\hat{p} \pm z^* \sqrt{\hat{p}(1-\hat{p})/n}$. This method is valid if the sample is randomly selected and the sample size is large.

Because we have a large sample size here, the one proportion *z*-interval approach should produce very similar results to the Plausible Values Method and the 2SD Method. In such a case, the theory-based approach is often the most convenient, especially if our confidence level is not 95%.

p) Go to the **Theory-Based Inference** applet. Enter the sample size and sample proportion for this survey. Then change the confidence level in the applet from 95% to 99% and press the **Calculate CI** button. Report the 99% confidence interval given by the applet. How does it compare to the 95% interval? (Compare both the midpoint of the interval = (lower endpoint + upper endpoint)/2 and the margin-of-error = (upper endpoint – lower endpoint)/2.)

Example 3: Reese's Pieces (adapted from *Workshop Statistics* by Rossman and Chance)
What does it mean to be "95% confident"? What does it mean to say the confidence interval method is valid? We will turn to an applet called **Simulating Confidence Intervals** to illustrate this. Consider using a random sample of Reese's Pieces candies to estimate the proportion of all such candies that are orange. The applet will simulate taking a large number of random samples and generating a confidence interval based on each sample. First make sure that the method is set to "Proportions" and "Wald." (Wald is the name for the conventional one-proportion *z*-interval.) We'll also suppose that 45% of all Reese's Pieces candies are orange, so set the population proportion to be .45, the sample size to be 75, and the confidence level to be 95%.

a) As we take new samples, what do you notice about the intervals? Are they all the same? Are any colored red? What does that denote?

b) Does the value of the population proportion change as we take new samples?

c) As we take hundreds and then thousands of samples and construct their intervals, about what percentage seem to be successful at capturing the population proportion?

d) Sort the intervals, and comment on what the intervals that fail to capture the population proportion have in common.

e) In practice, you only take one sample and construct one confidence interval. Can you be *sure* that the confidence interval successfully captures the true (but unknown) value of the parameter? In what sense can you be *confident* of this?

f) Now change the confidence level to 80%. Before pressing the Recalculate button, what changes do you expect to see? Then press the button. What two things change about the intervals?

g) Now change the sample size to 300 (with a confidence level of 95%). Does this produce a dramatically higher percentage of successful intervals? What does change about the intervals?

h) Is it desirable to have larger or smaller confidence levels? Explain.

i) Is it desirable to have wider or narrower confidence intervals? Explain.

j) What's a drawback of using a very high confidence level such as 99.9%?

k) What would it take to achieve a very high confidence and a very narrow confidence interval? Why is this so difficult to achieve?

Key idea: Interpreting confidence levels correctly requires us to think about what would happen if we took random samples from the population over and over again, constructing a CI for the unknown population parameter from each sample. The *confidence level* (e.g., 95%) indicates the long-run percentage of random samples that would produce a CI that successfully contains the actual value of the population parameter.

Example 4: Can Chimpanzees Solve Problems? (adapted from *Introduction to Statistical Investigations* by Tintle et al., www.math.hope.edu/isi)

In a 1978 study published in *Science*, Premack and Woodruff asked "To what extent does the chimpanzee comprehend the elements of a problem situation and potential solutions?" An adult chimpanzee (Sarah) was shown 30-second videotapes of a human actor struggling with one of several problems (for example, not able to reach bananas hanging from the ceiling, a record player not playing). Then Sarah was shown two photographs, one that depicted a solution to the problem (like stepping onto a box, plugging in the record player) and one that did not. She was then instructed to pick one of the photos and place it under the television monitor. (Sarah had been raised in captivity since age one and had extensive prior exposure to photographs and television.) The order in which the scenes were presented to Sarah was randomized, as was the left/right position of the photos presented to her. Sarah was shown eight different scenarios. For each scenario, researchers recorded whether or not Sarah selected the correct (solution) photo or not. It turned out that Sarah identified the correct photo for 7 of the 8 scenarios.

a) Identify the observational units for this study.

b) Identify the variable for this study. Also classify it as categorical (also binary?) or quantitative.

c) The relevant parameter can be denoted with the symbol π. Describe what π represents for this study.

d) Explain why using a one-proportion z-interval to estimate the value of π would not be appropriate with this study.

e) Use the **Simulating Confidence Intervals** applet to simulate taking a large number of random samples of size $n = 8$ with a success probability of .875 generating a large number of 95% one-proportion z-intervals based on the samples. What proportion of the intervals succeed in capturing the parameter value of .875? Is this close to 95%? If not, why not? (*Hint*: Sort the intervals.)

A relatively simple adjustment can be made to produce a confidence interval procedure that does successfully capture the parameter about 95% of the time. A "plus four" interval (proposed by Agresti and Coull, 1998) adds two additional "successes" and two additional "failures" to the sample and then uses the conventional z-interval procedure with an adjusted sample size of n+4.

f) Use the applet to produce a large number of "plus four" intervals. What proportion of these intervals succeed in capturing the parameter value of .875? Is this close to 95%? (What is changing about the intervals? Why does that lead to a higher success rate for the intervals?)

g) Use the "plus four" procedure to estimate with 95% confidence the long-run probability that Sarah identifies the correct photo.

Example 5: Cat Households (adapted from *Workshop Statistics*, by Rossman and Chance)

A sample survey of 47,000 households in 2007 found that 32.4% of American households own a pet cat.

a) Is this number a parameter or a statistic? Explain, and indicate the symbol used to represent it.

b) Should the one proportion *z*-confidence interval be valid with these data? Use technology (**Theory-Based Inference** applet) to conduct a test of whether the sample data provide evidence that the population proportion who own a pet cat differs from one-third. State the hypotheses, and report the test statistic and p-value. Draw a conclusion in the context of this study.

c) Use technology to produce a 99% confidence interval (CI) for the population proportion who own a pet cat. Interpret this interval.

d) Are the test result and CI consistent with each other? Explain how you can tell.

e) Do the sample data provide *very* strong evidence that the population proportion who own a pet cat is not one-third? Explain whether the p-value or the CI helps you to decide.

f) Do the sample data provide strong evidence that the population proportion who own a pet cat is *very* different from one-third? Explain whether the p-value or the CI helps you to decide.

Key idea: This example illustrates the distinction between *statistical* significance and *practical* significance. Especially with large sample sizes, a small difference that is of little practical importance can still be statistically significant (unlikely to have happened by chance alone). Confidence intervals should accompany significance tests in order to estimate the size of an effect/difference.

Example 6: Female Senators (adapted from *Workshop Statistics*, by Rossman and Chance)

Suppose that an alien lands on Earth, notices that there are two different sexes of the human species, and sets out to estimate the proportion of humans who are female. Fortunately, the alien had a good statistics course on its home planet, so it knows to take a sample of human beings and produce a confidence interval. Suppose that the alien happened upon the members of the 2014 U.S. Senate as its sample of human beings, so it finds 20 women and 80 men in its sample.

a) Use this sample information (with technology) to form a 95% one-proportion *z*- confidence interval for the actual proportion of all humans who are female.

b) Is this confidence interval a reasonable estimate of the actual proportion of all humans who are female?

c) Explain why the confidence interval procedure fails to produce an accurate estimate of the population parameter in this situation.

d) It clearly does not make sense to use the confidence interval in (a) to estimate the proportion of women on Earth, but does the interval make sense for estimating the proportion of women in the 2014 U.S. Senate? Explain your answer.

Key ideas: This example illustrates two important limitations of inference procedures. First, they do not compensate for the problems of a biased sampling procedure. If the sample is collected from the population in a biased manner, the ensuing confidence interval will be a biased estimate of the population parameter of interest. A second important point to remember is that confidence intervals and significance tests use sample statistics to estimate population parameters. If the data at hand constitute the entire population of interest, then constructing a confidence interval from these data is meaningless. In this case, you know precisely that the proportion of women in the population of the 2014 U.S. Senators is 0.20 (exactly!), so it is senseless to construct a confidence interval from these data.

Example 7: Body Temperatures (adapted from *Investigating Statistical Concepts, Applications, and Methods*, by Chance and Rossman)

A study examined whether 98.6^0F really was the typical healthy body temperature (Mackowiak, Wasserman, and Levine, *Journal of the American Medical Association*, 1992).

a) The study involved 130 healthy men and women, aged 18-40, who were involved in a Shigella vaccine trial in Maryland. We could analyze whether a majority of them have a body temperature lower than 98.6^oF. What important feature of this data would this be ignoring?

Instead of a confidence interval for a proportion, we often want a confidence interval for a population mean, denoted by μ. We know from the Central Limit Theorem that the distribution of sample means often follows a normal distribution with mean μ. So again we can expect our statistic (now the sample mean $\bar{x}$) to fall within 2 standard deviations of the population parameter μ and therefore we could estimate a 95% confidence interval for μ by $\bar{x} \pm 2$ SD($\bar{x}$). Or more generally, we could use $\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$.

b) What's unrealistic about applying this procedure in practice?

c) Suggest a reasonable approximation, based on the sample data, to use in place of $\sigma/\sqrt{n}$ in this confidence interval expression.

- The quantity $s/\sqrt{n}$ is called the **standard error** of the sample mean.

It seems that a reasonable confidence interval for a population mean μ is given by: $\bar{x} \pm z^* \frac{s}{\sqrt{n}}$.

d) To investigate what's wrong with this procedure, we again turn to the **Simulating Confidence Intervals** applet. First make sure that the method is set to "Means" and "z with s." Set the population mean to be 98.6, the population standard deviation to be .7, the sample size to be 130, and the confidence level to be 95%. As we take hundreds and then thousands of samples and construct their intervals, about what percentage seem to be successful at capturing the population mean?

e) But what if the sample size had only been 13? Generate thousands of intervals with this sample size. What two things change about the intervals? What if the sample size had only been 5?

- To fix this problem, we need to use a different multiplier (not z^*) of the $s/\sqrt{n}$ term.
- This multiplier comes from a probability distribution known as the *t*-distribution, producing the CI procedure for a population mean μ: $\bar{x} \pm t^* \frac{s}{\sqrt{n}}$.

Does this *t*-procedure really work? Let's see how well it does through a simulation.

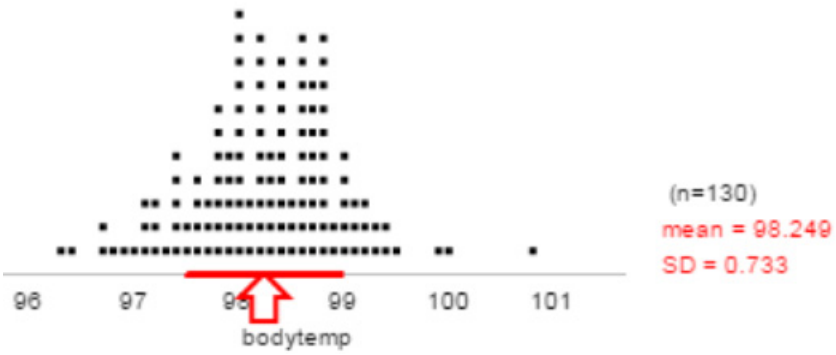
f) Change the applet's method to the "t" procedure. Generate a large number of random samples (for different sample sizes) and confidence intervals. Now is the success rate of these intervals very close to 95%? How does the success rate change as you change the sample size?

g) What happens to the *t* intervals for different sample sizes when the population distribution is not normal to begin with? (The applet lets you explore uniform and exponential populations).

Key idea: The one sample *t*-interval for a population mean μ is: $\bar{x} \pm t^* \, s/\sqrt{n}$, where $\bar{x}$ is the sample mean and *s* is the sample standard deviation. This method is valid if the sample is randomly selected and either the sample size is large or if the population distribution is normal.

Note: when the confidence level is 95%, the multiplier will be close to 2. The multiplier increases as the sample size decreases.

Below are the body temperature data from the study of 130 healthy adults.



h) Would you consider the one-sample *t*-interval to be valid for these data? Justify your answer.

i) Use the **Theory-Based Inference** applet (or a 2SD approximation) to estimate a 95% confidence interval. Write a one-sentence interpretation of the interval.

j) Does the confidence interval suggest that the sample data provide convincing evidence that the population mean body temperature is not 98.6 degrees? Explain how you can tell.

l) Do you expect that about 95% of the 130 individuals' body temperatures fall within the confidence interval? Explain why or why not.

m) How many of the temperature values in the sample fall within the confidence interval? What proportion of the sample values is this? Is this close to 95%? Should it be? Explain.

Key idea: A confidence interval estimates a population *mean*, not individual values.