Exploration 3.2 Malaria in Yemen cont.

2SD and Theory-Based Confidence Intervals for a Single Proportion

LEARNING GOALS

- Compute a confidence interval for a proportion written in terms of its endpoints from a confidence interval written in terms of center plus or minus the margin of error and vice versa.
- Approximate a 95% confidence interval for a proportion by using the 2SD method.
- Compute a confidence interval for a proportion using a theory-based approach, including checking validity conditions.
- Infer the relative width of a confidence interval when changing the confidence level.

Introduction

In Statistics, we often use the observed sample statistic to estimate an unobserved population parameter. It is important for that estimation to include an indicator of how "accurate" we think the estimate is, as well as how reliable our method is. A ***confidence interval*** provides a set of "plausible" values for the parameter, along with a ***confidence level***. In this exploration, we will explore two methods for constructing a confidence interval.

In countries like Yemen, the population has been dealing with endemics such as malaria long before the COVID-19 pandemic hit the world. To estimate the proportion of malaria cases in Yemen, a cross-sectional study was conducted on febrile patients (patients presenting with a fever) from November 2018 to April 2019. Patients included were from three districts of Hodeidah City, the second largest city in Yemen, who were referred to the laboratories of the hospitals; 355 volunteered to participate. Of the 355 participants, 115 (32.4%) were diagnosed with Malaria.

1. Identify the population and sample in this study.

   Population:

   Sample:

2. Is it reasonable to believe that the sample of 355 volunteer patients is representative of the larger population? Explain why or why not.
3. Explain why 32.4% is a statistic and not a parameter. What symbol would you use to represent it?
4. Identify (in words) the parameter that this study was interested in estimating.

5. Is it reasonable to conclude that exactly 32.4% of Yemen's population have Malaria? Explain why or why not.

   Although we expect π to be close to 0.324, we realize there may be other plausible values for the population proportion as well. First consider the value of 0.375. Is this a plausible value for π?

6. Use the **One Proportion** applet to simulate random samples of 355 people from a population with π = 0.375. (*Hint:* Keep in mind that 0.375 is what we are assuming for the population proportion and 0.324

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?

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

8. Now consider 0.50. Is this a plausible value for π? Repeat #6 and record the mean and standard deviation for this null distribution as well.

Clearly 0.50 is going to be "too far" from $\hat{p}$ = 0.324 to be plausible. But how far is too far?

9. Reconsider our first guess of π = 0.375. How many standard deviations is the observed sample proportion of 0.324 from 0.375? (*Hint*: Standardize the 0.324 value by looking at the difference between 0.324 and 0.375 and divide by the standard deviation you found in #7.)

You should notice that 0.375 and 0.324 are about 2 standard deviations apart *and* that the two-sided p-value is around 0.05, so this value (0.375) is close to the edge of values that can be considered plausible for π. Values between 0.324 and 0.375 are considered plausible and values larger than 0.375, or more than 2 standard deviations above 0.324, will not be plausible values for the population proportion.

Key Idea

When a sampling distribution of a statistic is bell-shaped, as your null distribution should be for this study, approximately 95% of the statistics in the sampling distribution will fall within 2 standard deviations of the mean. This implies that 95% of sample proportions will fall within 2 standard deviations of the parameter (π), which means that π is within 2 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 sampling 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 proportion (or population proportion), π, in symbols as

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

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

Think About It
So how do we find the standard deviation to use for the 2SD method?

10. How did the standard deviations you found in #7 (with $\pi = 0.375$) and in #8 (with $\pi = 0.50$) 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 $\pi = 0.50$. So one approach would be to carry out one simulation (with lots of trials) using $\pi = 0.50$ and use that value of the standard deviation to estimate the margin of error.

11. Determine a 95% confidence interval using the 2SD method:

- **a.** First calculate 2 × (*standard deviation for your sampling distribution of sample proportions*) using 0.5 in the simulation to estimate the standard deviation. (This is the margin of error.)
- **b.** Use this margin of error to produce a 95% confidence interval for π. (*Hint:* Subtract the margin of error from $\hat{p}$ to determine the lower endpoint of the interval and then add the margin of error to $\hat{p}$ to determine the upper endpoint of the interval.)
- **c.** Interpret the confidence interval: You are 95% confident that *what* is between what two values?

One limitation to this method is that it only applies for 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.

Theory-Based Approach

We don't always need to simulate a sampling distribution—not if we can accurately predict what would happen if we were to simulate. Instead, we can predict the standard deviation of the distribution of sample proportions by the formula $\sqrt{\pi\,(1-\,\pi)/n}$. But what do we use for the value of π in this formula? When constructing a confidence interval, we will substitute the observed 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, $\hat{p}$

12. Calculate the standard error for this study. How does it compare to the standard deviations you found in #7 and #8?

So a more general formula for using the 2SD method to estimate a population proportion would be

$$\hat{p} \pm 2\sqrt{\hat{p}\,(1-\,\hat{p})/n}$$

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.

13. We will rely on technology to find the multiplier appropriate for our confidence interval.

a. In the **Theory-Based Inference** applet specify the sample size (n) of 355 and the sample proportion of 0.324 and press **Calculate**. (The applet will fill in the count or you can specify the sample count 115 and the applet will fill in the sample proportion when you press Calculate.)
b. Check the box for **Confidence interval**, confirm the confidence level is 95%. Report the 95% theory-based confidence interval.

14. Is this theory-based confidence interval similar to the one you obtained using the 2SD method?

Validity Condition
The theory-based approach for finding a confidence interval for π (called a ***one- sample z-interval***) is considered valid if there are at least 10 observational units in each category of the categorical variable (i.e., at least 10 successes and at least 10 failures).

Because we have a large sample size here, the theory-based approach should produce very similar results to a simulation-based approach. In such a case, the theory-based approach is often the most convenient, especially if our confidence level is not 95%.

15. Change the confidence level in the applet from 95% to 99% and press the **Calculate CI** button again. 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.)

Exploring Further

16. Suppose instead of the highland Hodeidah City, a sample of 45 febrile patients from a coastal-plains city was taken where 9 were diagnosed with Malaria. It is known that the prevalence of Malaria is lower in the costal-plains cities than in the highland cities. Determine a theory-based confidence interval for the proportion of Yemen's population who have Malaria. Do you think this interval is valid?

17. How does your interval compare to the interval you found for the published study on Hodeidah City?

18. Based on characteristics of the sample of 45, does it make sense that the two intervals compare as you described in #17? Explain.

Rashad Abdul-Ghani ,Mohammed A. K. Mahdy, Sameer Alkubati, Abdullah A. Al-Mikhlafy, Abdullah Alhariri, Mrinalini Das, Kapilkumar Dave, Julita Gil-Cuesta. **Malaria and dengue in Hodeidah city, Yemen: High proportion of febrile outpatients with dengue or malaria, but low proportion co-infected**

Published: June 25, 2021
https://doi.org/10.1371/journal.pone.0253556