Lab Confidence

Abhi Thanvi, Paul Holaway

July 12th, 2022

Contents

Lab Confidence **2**

Welcome . . . 2

The Idea of this Lab . . . 2

Problem 1: Z vs. t, You Tell Me . . . 2

Problem 2: Age Is Just an Interval . . . 3

Problem 3: Shoes! . . . 4

Problem 4: Narrower or Wider . . . 5

Project Questions . . . 5

Submission . . . 5

Lab Confidence

Welcome

Just like learning a new spoken language, you will not learn the language without practice. Labs are an important part of this course. Collaboration on labs is **extremely encouraged**. If you find yourself stuck for more than a few minutes, ask a neighbor or course staff for help. When you are giving help to your neighbor, explain the **idea and approach** to the problem without sharing the answer itself so they can figure it out on their own. This will be better for them and for you. For them because it will stick more and they will have a better understanding of the concept. For you because if you can explain it to other students, that means you understand it better too.

The Idea of this Lab

I can confidently say that Confidence Interval is a huge chunk of statistic that falls under the statistical inference section. Through this lab, you will be exploring the application and calculation of Confidence Intervals using `R`. Understanding Confidence Interval (CI) is crucial to understand what we do for the time that is left for us. I am not going to huge description for today's lab as you probably understand the importance of this topic and I also want to give you time to do this lab and maybe ask questions for project. LET'S GO!

"Confidence Interval is like giving an estimated range for answer, which you aren't 100% sure about" - Woke Abhi

Problem 1: Z vs. t, You Tell Me

Question 1: You are trying to find the 98% confidence interval from a population whose size is 100. You know the Standard Deviation too! Do you use Z-Interval or t-interval?

Answer: Since we know the population standard deviation σ and the population size n is huge (way bigger than 30), we use Z-interval.

Question 2: You are trying to find the 85% confidence interval from a population whose size is 1000. You do not know the Standard Deviation. Do you use Z-Interval or t-interval? If you choose Z-interval, explain why (Hint: Something we learned in the last few classes)?

Answer: Here, since you don't know the population standard deviation value (σ), you should technically go with t-interval. However, since your population size (n) is HUGE (it's 1000, which is way bigger than 30), you can still use Z-Interval, because for large enough n values, Z-intervals and t-intervals will result in a very similar confidence interval.

Question 3: You are trying to find the 95% confidence interval from a population whose size is 10. You do not know the Standard Deviation. Do you use Z-Interval or t-interval?

Answer: Here, since you don't know the population standard deviation value (σ) AND the population size (n) is super small (10, which is smaller than 30), you need to use t-interval. Using Z-interval is incorrect here, since its relatively small resultant confidence interval wouldn't account for the discrepancies caused by sampling from a tiny population; however, the augmentation of the t distribution creates an artificially larger confidence interval to account for these discrepancies, making the t-interval a better option for this scenario.

Question 4: You are trying to find the 90% confidence interval from a population whose size is 5. You do know the Standard Deviation. Do you use Z-Interval or t-interval?

Answer: Here, even though you're sampling from a super small population (5 is less than 30), since you know the standard deviation, the small sample size is irrelevant, and you'd use the Z-interval to create your confidence interval.

Problem 2: Age Is Just an Interval

Question 1: Abhi wants you to create a 90% and 98% confidence interval for the average age for a batter in MLB. Again, I don't know Abhi's fascination with age of people, but let's just ignore that. The data for this is in the `MLB_Batters` data set. Create the two confidence intervals.

Answer: NOTE: Since the population includes everyone in the MLB, the population standard deviation is KNOWN, so you need to use Z-interval.

```
#Imports
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.7     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
MLB_Batters <- read.csv("~/Desktop/DPI 2022/MLB_Batters.csv", stringsAsFactors=TRUE)
#Common Setup Work for both CIs
xbar = mean(MLB_Batters$Age)
n = nrow(MLB_Batters)
sigma = sd(MLB_Batters$Age)*sqrt((n-1)/n)
#90% CI Work
c90 = (1-0.90)/2
Z90 = qnorm(c90, lower.tail = F)
#98% CI Work
c98 = (1-0.98)/2
Z98 = qnorm(c98, lower.tail = F)
#Print Answers
xbar + c(-1, 1)*Z90*sigma/sqrt(n) # 90% confidence interval
```

```
## [1] 28.36029 28.82008
```

```
xbar + c(-1, 1)*Z98*sigma/sqrt(n) # 98% confidence interval
```

```
## [1] 28.26504 28.91533
```

Question 2: With your group, discuss what confidence interval mean? Write the answer to what your second confidence interval means in context from question 1.

Answer: A confidence interval means that you can be confident (up to a certain extent) that your true population parameter (for example, a mean) lies within a certain interval, defined by a range centered around a sample statistic and using range boundaries based on the statistic standard deviation and the confidence level (usually 90% or 95%) modeled after a specific distribution (such as Z or t).

For instance, in the context of MLB Batters' Age (from question 1), the second confidence interval (the 98% one) means that we can be 98% sure that the true population average of the age of MLB Batters (μ) lies somewhere between 28.26504 years and 28.91533 years.

Problem 3: Shoes!

Question 1: Abhi is trying to convince Paul that people own more than 3 pairs of shoes on average. Help Abhi do so by creating a 95% confidence interval for the average number of shoes that students own. Use the `hello` data set. Remember that this is a sample of UIUC STAT107 students.

Answer: (Either Z or t is fine here.)

```
#Imports
hello <- read.csv("~/Desktop/DPI 2022/hello.csv", stringsAsFactors=TRUE)
#Common Setup Work
xbar = mean(hello$Shoes)
n = nrow(hello)
c = (1-0.95)/2
#CI
s = sd(hello$Shoes)
df = n-1
t = qt(c, df, lower.tail = F)
#Print Answer
xbar + c(-1, 1)*t*s/sqrt(n)
```

```
## [1]  7.552353 10.385147
```

Question 2: With your group, discuss what confidence interval mean? Write the answer to what your confidence interval means in context from question 1.

Answer: This confidence interval means that the true population average of the number of shoes owned by UIUC STAT107 students is somewhere between 7.552353 shoes and 10.385147 shoes.

Problem 4: Narrower or Wider

Question 1: Assuming the standard deviation remains the same. If I increase my confidence level from 90% to 95%, would my confidence interval be wider or narrower? In other words, will the range of my interval be larger or smaller?

The answer should be contextualized or in your own words. Providing numbers or math is not recommended

Answer: By increasing my confidence level, my confidence interval will become wider, because to be more sure that a true population parameter is within a confidence interval, you need to widen the interval so that it includes a wider range of possibilities for the values of the true population parameter. This is shown in the formulas, because in the margin of error formula (which is $Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ or $Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$ or $t_{\frac{\alpha}{2},df}\frac{s}{\sqrt{n}}$), the confidence level part is in the numerator, so increasing it will increase the confidence interval's margin of error, and therefore increase its range.

Question 2: Assuming the standard deviation and CI remains the same. If I increase my sampling size, would my confidence interval be wider or narrower? In other words, will the range of my interval be larger or smaller?

The answer should be contextualized or in your own words. Providing numbers or math is not recommended

Answer: If you increase your sampling size, the confidence interval would become narrower. This is because if you have a large sample size, because of the Law of Large Numbers, your sample statistic will get closer to the true population parameter than if you used a small sample size, so the range of your interval will get smaller. This is shown in the formulas, because in the margin of error formula (which is $Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ or $Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$ or $t_{\frac{\alpha}{2},df}\frac{s}{\sqrt{n}}$), the sampling size ($n$) is in the denominator, so increasing it will decrease the confidence interval's margin of error, and therefore decrease its range.

Question 3: After finishing the entire lab. With your group, discuss and report how to change the width of your confidence interval. Keep in mind, we cannot change the population and it's standard deviation.

Hint: Think about how Confidence Level and Sample Size affect the specific parts of Confidence intervals?

Answer: To change the width of your confidence interval WITHOUT changing the population and standard deviation, you need to change the confidence level and sample size. Specifically, if you want to make your confidence interval narrower, you can increase your sample size or decrease your confidence level. Likewise, if you want to make your confidence interval wider, you can decrease your sample size or increase your confidence level. In terms of the benefits of doing this, though, increasing sample size is almost always good, because increasing it will increase accuracy in your confidence levels; however, increasing confidence level too much will create too wide of an interval (resulting in vagueness and inconclusiveness in data), while decreasing confidence level will mean you can't be too sure of your conclusions - for confidence level, it's always best to have a balanced confidence level, such as 90% or 95%.

Project Questions

Feel free to work on your project if there is any time left after the labs. Paul and I are here to answer any questions during the second half of the lab times to answer mainly project related questions, but general questions are more than welcome too. Feel free to discuss among your group about any project ideas or help each other out. Remember collaboration is promoted, plagiarism is not! :)

Submission

Once you have finished your lab...

1. Go to the top left and click `File` and `Save`.

2. Click on the `Knit` button to convert this file to a PDF.
3. Submit **BOTH** the `.Rmd` file and `.pdf` file to Blackboard by 11:59 PM tonight.