Lab 3: Confidence Interval of a Proportion

Introduction and Objective:

In the lecture, we learned how to calculate the median absolute deviation (MAD) and the interquartile range of a data set. Today, we will first learn how to use R to figure out the MAD and the interquartile range.

In the lecture next week, we will discuss how to calculate the confidence interval of a proportion using the Modified Wald Method by hand. Today, we will focus on learning how to use R to calculate the CI of a proportion, including using the Modified Wald Method.

Before we can use an R package to calculate the CI of a proportion by using the Exact Method, the Standard Wald Method, and the Modified Wald Method, we are going to install some R packages first. Once we get the results from using these different methods, we will compare these results.

In addition, we are going to use R to confirm some concepts we learned in the lecture. In the end, the result of using different shortcut rules to calculate the CI of a proportion will be compared with the results we get from using the R package.

In this lab notes/manual, except in the introduction parts, all the R command lines are in **bold**; all the notes are following a # and in blue; all the R results directly follow the R codes/command lines and are in red.

I. Find MAD and the interquartile range of a data set

# Set your working directory first, and then we use the data set we used in the lecture to find the MAD of it:

```
a<-c(1, 2, 3, 4, 5, 6, 7)
```

or

```
a<-c(1:7)
```

```
mad(a, constant=1)
```

```
[1] 2
```

```
# As we discussed in the lecture when the data set is small, you have to specify constant=1.

# We use the same data set we used in the lecture to find the interquartile range:

b<-c(1, 3, 4, 6, 8, 9, 10, 12)

summary(b)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.

  1.000   3.750   7.000   6.625   9.250  12.000

interquartile<-9.25-3.75

interquartile

[1] 5.5

# Or we can use the R built-in function IQR() to calculate the interquartile range.

IQR(b)

[1] 5.5


# As we discussed in the lecture, there are different methods to calculate the interquartile range. And in R, we can use the argument “type=” to choose different methods to calculate the interquartile range.

# For instance,

IQR(b,type=1)

[1] 6

IQR(b,type=7)

[1] 5.5

# As we can see from the above results, IQR(b, type=1) gives a different result than IQR(b, type=7).


# This is what we discussed in the lecture when the data set is small, using different methods, we may get different results for the interquartile range.
```

# If you don’t specify which method you are choosing, by default, R uses type=7.

II. Different methods to calculate the CI of a proportion

# The following are the formulae and equations we used in a lecture to calculate p’:

$$p' = \frac{\#\text{ "successes"} + 2}{\#\text{of experiment s} + 4} = \frac{S+2}{N+4}$$

$$p' - 1.96 \cdot \sqrt{\frac{p'(1-p')}{N+4}} \quad \text{to} \quad p' + 1.96 \cdot \sqrt{\frac{p'(1-p)'}{N+4}}$$

# Example:

# There were 565 people enrolled in a study to test a drug’s effect on improving sleeping quality. After they were given the drug 3 times, they were asked to rate their feelings regarding sleeping quality. There were 16 people who reported their sleeping quality was improved. At a 95% confidence level, what is the CI of the proportion of people, who feel their sleeping quality improves after being given the drug 3 times?

# We may do the calculations by hand using the Modified Wald Method:

# The number of successes is 16, and the number of experiments is 565.

16/565=0.0283

p’=(16+2)/(565+4)=18/569=0.0316

# When the confidence level is 95%, the z=1.96, thus

$$W=1.96*\sqrt{\frac{p'(1-p')}{N+4}} = 1.96*\sqrt{\frac{0.0316(1-0.0316)}{565+4}} = 1.96 * \sqrt{\frac{0.0306}{569}} = 1.96 * 0.0073 = 0.0144$$

# Lower confidence limit: 0.0316-0.0144=0.0172

# Upper confidence limit: 0.0316+0.0144=0.046

# So, the CI of the proportion at 95% confidence level is [0.0172, 0.046]

# Of course, we can perform the calculations shown above in R:

```
n<-565

s<-16

s/n
```

```
[1] 0.02831858

pprime<-(s+2)/(n+4)

pprime

[1] 0.03163445

w<-1.96*sqrt((pprime*(1-pprime))/(n+4))

w

[1] 0.01438135

lowerCL<-pprime-w

upperCL<-pprime+w

lowerCL

[1] 0.01725309

upperCL

[1] 0.0460158
```

Challenge Question 1: Can you use the “function()”, an R built-in function we learned in Lab 2, to define some functions and reduce the command lines you have to type? Try it yourself first without looking at the answer key that is at the end of this chapter.

# As you can see what was shown above was basically doing the step-by-step calculations in R based on the formulae we used for calculating the CI of a proportion by hand. It is like using R as a calculator and is still time-consuming.

# In fact, many R users had performed similar calculations many times before. So to save the time of people to type in similar codes, again and again, some R users and developers put the reproducible codes in some fundamental units. These fundamental units of R are called “packages” (as we talked about in Lab 2). People upload the packages they developed and share them with all the other R users. So other R users, if they want to perform similar tasks, can download the R packages developed by other people and use them directly instead of typing in the same or similar codes again and again.

# Last week, we learned how to install R packages by using 2 different ways and installed our first 2 R packages. Today we are going to download and install another R package, which allows us to calculate the CI of a proportion in an easier way.

# Type the following command:

```
install.packages("devtools")
```

# This is just to remind you what you did in Lab 2:

# 1) install.packages() should be plural. The package is followed by an “s”. 2) install.packages() started with a little “i”, it should not be started with a capital letter “I”.

# If you are suing base R, you will see a window “CRAN mirror” pops up. These are the mirror sites with the repositories of the packages.

# Scroll down and highlight to select a CRAN mirror, for instance, “CA1”, and then click “OK”.

# After that, the package of “devtools” is installed on your computer.

# Remember what you learned in Lab 2:

# In a new R session, before you can use any specific R packages, you have to activate the packages. The “library” is the R function that activates a package. If you close and restart a new R session, you have to activate the package again if you want to use it in the new R session. But if you stay in the same R session, once you activate the R package, you do not have to activate it for the second time before you use it.

# To activate “devtools”, we type the following command:

```
library(devtools)
```

# Remember what you learned in Lab 2:

# 1) the library started with a lower case “l”. 2) there are no quotation marks before and after the name of the package.

```
install_github("shearer/PropCIs")
```

# Prior to type install_github(“shearer/PropCIs”), you must activate “devtools” first. Otherwise, there will be an error message saying “could not find function ‘install_github’”.

```
library(PropCIs)
```

# Once we activate the package “PropCIs”. We can get the CI just by typing a single command line:

```
add4ci(x = 16, n = 565, conf.level = 0.95)
```

```
data:

95 percent confidence interval:

 0.01725336 0.04601553

sample estimates:

[1] 0.03163445
```

# The function “add4ci” is for the Modified Wald Method. The “x” equals the number of “successes”. The “n” equals the number of “experiments”. The “conf.level” equals the confidence level. It expresses in proportions, for instance, if you choose a 95% confidence level, then it equals 0.95, and if you choose a 99% confidence level, then it equals 0.99.

# As you can see the result is exactly the same as what we got from calculations by hand.

# As we discussed in the lecture, if we increase the confidence level, the CI will be wider; and if we lower the confidence level, the CI will be narrower. For this example, we can lower the confidence level from 95% to 90% to see whether the result is as we talked about in the lecture:

```
add4ci(x = 16, n = 565, conf.level = 0.9)
```

```
data:

90 percent confidence interval:

 0.01956546 0.04370344

sample estimates:

[1] 0.03163445
```

```
# You can see the CI is getting narrower.

# Function “addz2ci” allows us to calculate the CI of a proportion using Standard Wald Method:

addz2ci(x = 16, n = 565, conf.level = 0.95)

data:

95 percent confidence interval:

 0.01714956 0.04585826

sample estimates:

[1] 0.03150391

# We can see the CI calculated by this Standard Wald Method is slightly different from the CI we got using the Modified Wald Method.

# As we talked about it in the lecture, the Standard Wald Method is easy to compute by hand, while Modified Wald Method is not much harder than Standard Wald Method, but it is much more accurate.

# We also talked about it in the lecture that the Exact Method (Clopper and Pearson, 1934) calculates a CI sometimes a bit wider than necessary, and it cannot be easily computed by hand. However, using R, we can easily use the Exact Method to calculate the CI. The function we are going to use is “exactci”.

exactci(x=16,n=565,conf.level=0.95)

data:

95 percent confidence interval:

 0.01627130 0.04558102
```

Challenge Question 2: Compare what you got from using “exactci”, “add4ci”, and “addz2ci”, what are the differences? Which is wider? (The answer key is at the end of this lab chapter)

III. Rule of Three, Rule of Five, and Rule of Seven

```
# If the success number is 0, and the number of the experiments is 165, which rule should apply if you want to take a shortcut to calculate the CI of the proportion?

# Since it is 0/165, the Rule of Three will be applied.

# Thus, the upper confidence limit is 3/165=0.018 and the lower confidence limit is 0.

# We can do the step-by-step calculations using modified Wald method:
```

```
n<-165
s<-0
pprime<-(s+2)/(n+4)
pprime
[1] 0.01183432
w<-1.96*sqrt((pprime*(1-pprime))/(n+4))
w
[1] 0.01630419
lowerCL<-pprime-w
upperCL<-pprime+w
lowerCL
[1] -0.004469871
upperCL
[1] 0.02813851
```

```
# When the numerator is 0, the result from Rule of Three is closer to which result, the addz2ci, add4ci, or exactci?
```

```
addz2ci(x=0,n=165,conf.level=0.95)
data:
```

```
95 percent confidence interval:

 0.00000000 0.02737217

sample estimates:

[1] 0.01137593
```

```
add4ci(x=0,n=165,conf.level=0.95)
```

```
data:

95 percent confidence interval:

 0.00000000 0.02813821

sample estimates:

[1] 0.01183432
```

```
exactci(x=0,n=165,conf.level=0.95)
```

```
data:

95 percent confidence interval:

0.00000000 0.02210878
```

```
    # So, the result using the Rule of Three is closer to the result of the Exact Method.
```

IV. There is an online tool to calculate the CI of a proportion:

http://www.graphpad.com/quickcalcs/ConfInterval1.cfm

Answer keys to the challenge questions:

1)

```
lowerCL<-function(s,n){(s+2)/(n+4)-1.96*sqrt(((s+2)/(n+4)*(1-(s+2)/(n+4)))/(n+4))}

upperCL<- function(s,n){(s+2)/(n+4)+1.96*sqrt(((s+2)/(n+4)*(1-(s+2)/(n+4)))/(n+4))}

lowerCL(16,565)

upperCL(16,565)
```

2) At 95% confidence level, by using add4ci(), the result was [0.01725336, 0.04601553]; by using addz2ci(), we got [0.01714956, 0.04585826]; and exactci() gave the result [0.01627130, 0.04558102]. The differences between the upper and the lower limits of the CIs from using add4ci(), addz2ci(), and exactci() are 0.02876217, 0.0287087, and 0.02930972, respectively. The exactci() generated a wider CI.

Group Assignment 2

A clinical trial testing a drug for its anti-cancer effect enrolled 663 patients, including 321 patients who were given placebos, and the rest were on the experimental arm. After the treatments, there were 2 patients, who were taken placebos, were reported with undetected lymph nodes metastasis. While for the patients, who were given the drug, sixty-five had undetected lymph nodes metastasis.

1) By typing R command lines, using the modified Wald method, calculate the CI of proportion at 95% confidence level for both the placebo-treated group and the drug-treated group. **(3pts)**

2) Use the R package (PropCIs) to find the CIs (exact CI, Wald method, and modified Wald method) of proportion for both groups. Copy & paste your command lines and results. **(3pts)**

3) For which group, you can take a shortcut and estimate the result by using either Rule of Three, Rule of Five, or Rule of Seven? **(1pt)** Which rule would apply? **(1pt)** What is the result of using this rule? Show your calculations. **(1pt)** Is it the same as what you got from using R? **(1pt)**