Lab 4: Confidence Interval of a Mean

Introduction and Objective:

In lab 3, we explored writing our own command lines in R and using R packages to compute the confidence interval of a proportion. In today's lab, we are going to continue our journey of computing confidence intervals and will focus on computing the confidence interval of a mean by using a few different methods in R.

The first method we are going to practice is calculating the CI of a mean following step-by-step calculations. This helps us to understand and remember how to do the calculations by hand. Then, we will define our own function, which helps us to remember what we did in the previous two labs, to calculate the CI of a mean. After that, we are going to use a simple R built-in function to compute the confidence interval of a mean.

At the very end of the lab, we will learn how to import data into R. In the next lab, we will continue this topic.

Also, there is a data file you need to download and save to the folder you will use as the working directory:

Lab1_Mouse_DNA_methylation_small_table.txt

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.

In PC, to set the working directory, you may type:

setwd("C:/R")

In Mac, you have to remember what we did in the first lab to set up the working directory.

I. Calculate the CI of a mean by doing step-by-step calculations in R:

Use the same example we used in the lecture and Lab 2:

Sample1	Sample2	Sample3	Sample4	Sample5	Sample6	Sample7	Sample8	Sample9	Sample10	Sample11	Sample12

32.88	40.63	34.43	39.97	36.9	34.71	57.74	57.42	61.3	56.47	57.87	59.39

samples<-c(32.88, 40.63, 34.43, 39.97, 36.9, 34.71, 57.74, 57.42, 61.3, 56.47, 57.87, 59.39)

Note: make sure you use commas to separate the numbers

First, calculate the SD of the samples using SD=$\sqrt{\sum(Yi - Ymean)2/(n-1)}$

1) # Step 1, Calculate the mean (average):

a<-mean(samples)

a

[1] 47.47583

2) # Calculate the difference between each value and the mean:

b<-samples-a

b

[1] -14.595833 -6.845833 -13.045833 -7.505833 -10.575833 -12.765833

[7] 10.264167 9.944167 13.824167 8.994167 10.394167 11.914167

3) # Square each of these differences:

c<-b^2

c

[1] 213.03835 46.86543 170.19377 56.33753 111.84825 162.96650 105.35312

[8] 98.88645 191.10758 80.89503 108.03870 141.94737

4) # Add up those squared differences:

d<-sum(c)

d

[1] 1487.478

5) # Divide this sum by (n-1), where n is the number of values/observations. What you get is called the variance of the "samples".

n<-12

e<-d/(n-1)

e

[1] 135.2253

6) # Take the square root of the variance you calculated. The result is the standard deviation of the values defined in “samples”.

f<-sqrt(e)

f

[1] 11.62864

Actually, there is an R built-in function, as we introduced in the first lab, allows us to get the standard deviation directly:

f1<-sd(samples)

f1

[1] 11.62864

After we get the standard deviation, we can further calculate the margin of error (W) using the equation as follows:

$W=t^* \times SD/\sqrt{n}$

t* can be looked up in the table for t-distribution. It is determined by the degrees of freedom and the confidence level.

Here, the df=n-1, and is 11. At 95% confidence level, using the t-distribution table we can find the t* is 2.201

w<-2.201*f/sqrt(n)

w

[1] 7.388536

Upper confidence limit:

upperCL<-a+w

upperCL

[1] 54.86437

lowerCL<-a-w

lowerCL

[1] 40.0873

Then, the CI of the mean at 95% confidence level is: [40.0873, 54.86437]

II. Use R to find z and t*

As you may recall, in the lecture and one of our reference books (on page 37), for calculating the CI of a proportion, only 3 z scores are provided: for 90% CI, z=1.645; for 99% CI, z=2.576; and for 95% CI, z=1.960 (which is very close to 2). However, if we want to calculate CI at an 80% confidence level or if we want to calculate CI at a 92% level, we can't find the z scores in the textbook. Of course, you may google search and try to find these z scores using online resources. But, as I told you in the lecture, we can use R to easily find any of the z scores that we may need. The following shows you how to find a particular z score in R:

Alpha= 1-confidence level

For example, if confidence level is 95%, then alpha=1-95%=0.05

Find z using R:

z=qnorm(1-alpha/2) or z=abs(qnorm(alpha/2))

Note: abs() is an R built-in function for finding the absolute values of numbers.

(We will discuss more of this when we talk about normal distribution in the lecture next week).

z<-qnorm(1-0.05/2)

z

[1] 1.959964

or

z<-abs(qnorm(0.05/2))

z

[1] 1.959964

As for the t*, which is needed for calculating the CI of a mean, our reference book only shows a table contains some of the t* values. As we talked about in class, that table and many others you may find online (such as the one at https://ucps.instructure.com/courses/16263/files/35276?module_item_id=37133) are incomplete t-distribution tables. However, similar to the case of finding the z scores, we can easily find any of the t* that we may be interested in by using R. The following is how to find the t* using R:

t*=abs(qt(alpha/2,df)) or t*=qt(1-alpha/2,df)

t<-abs(qt(0.05/2,11))

t

[1] 2.200985

or

t<-qt(1-0.05/2,11)

t

[1] 2.200985

Because w=t* x SD/$\sqrt{n}$

therefore,

w<-t*f1/sqrt(n)

w

[1] 7.388486

upperCL<-a+w

upperCL

[1] 54.86437

lowerCL<-a-w

lowerCL

[1] 40.0873

Then, the CI of the mean is: [54.86437, 40.0873]

III. Define your own function to compute the margin of error and then the CI of a mean

If I ask you to calculate the CI of the mean of a data set B: 3, 4, 6, 10, 17 at 85% confidence level, can you do it now?

Of course, we can do it. In fact, since we learned how to define our own function, we can define our own function to make this task and any similar task easier.

So, let's define our own function to calculate the margin of error and then the CI of a mean. We may apply this function to compute the CI of the mean of the "samples", and also apply it, again and again, to figure out the CI of the mean of other data set, such as the one in the challenge question.

w<-function(alpha,a,n){qt(1-alpha/2,n-1)*sd(a)/sqrt(n)}

or

w<-function(alpha,a,n){abs(qt(alpha/2,n-1))*sd(a)/sqrt(n)}

or

w<-function(alpha,samples){abs(qt(alpha/2,length(samples)-1))*sd(samples)/sqrt(length(samples))}

alpha<-0.05

n<-length(samples)

or

n<-sum(table(samples))

Note: In Lab 1 and then in the group assignment 1 of Lab 2, we talked about how to find the number of observations either by combining the built-in function table() and sum() or just by using length().

UpperLimit<-mean(samples)+w(alpha,samples,n)

UpperLimit

[1] 54.86432

LowerLimit<-mean(samples)-w(alpha,samples,n)

LowerLimit

[1] 40.08735

Challenge Question: define your own function in R to calculate the margin of error, and then at 85% confidence level, find the CI of the mean of dataset B. (The answer is at the end of this lab note.)

IV. Use an R built-in function to find the CI of a mean

Type the following command line in your R session:

t.test(samples, conf.level=0.95)

Once you hit "Enter" / "Return", the result you get should look like:

One Sample t-test

data: samples

t = 14.1428, df = 11, p-value = 2.114e-08

alternative hypothesis: true mean is not equal to 0

95 percent confidence interval:

40.08735 54.86432

sample estimates:

mean of x

47.47583

You can see the results, the confidence interval of the mean at 95% confidence level was calculated and is the same as what you calculated using step-by-step calculations or using the

function defined by you. The t.test() is an R build-in function. We are going to use it again in a future lab when we perform the Student's t-tests. Here, it provides a simple way to figure out the confidence interval of a mean.

V. Import large data into R

We will learn more about how to import data into R. Here is something simple to get us started on this topic.

Use the data file "Lab1_Mouse_DNA_methylation_small_table.txt", and select the data for gene ncf1, calculate the CI of the mean at 90% confidence level:

```
h<-read.table("Lab1_Mouse_DNA_methylation_small_table.txt")
dim(h)
head(h)
ncf1<-h[3,2:13]
ncf1
sample2<-as.matrix(ncf1)
mode(sample2)
i<-as.numeric(sample2)
i
 [1]  5.00  5.24  2.81  5.93  2.76  3.15 17.46 15.05 16.82 19.46 16.24 16.90
mode(i)
[1] "numeric"
mean(i)
[1] 10.56833
L<-sd(i)
L
```

t.test(i, conf.level=0.9)

The answer to the challenge question:

w<-function(alpha,a,n){qt(1-alpha/2,n-1)*sd(a)/sqrt(n)}

or

w<-function(alpha,a,n){abs(qt(alpha/2,n-1))*sd(a)/sqrt(n)}

b<-c(3, 4, 6, 10, 17)

LowerLimit<-mean(b)-w(0.15,b,5)

UpperLimit<-mean(b)+w(0.15,b,5)

LowerLimit

[1] 3.466482

UpperLimit

[1] 12.53352

Group Assignment 3

Use the data from the file "Lab1_Mouse_DNA_methylation_small_table.txt". At 92% confidence level, compute the confidence interval of the mean of the DNA methylation of gene Sh2b3 of all the 12 mice.

1) Do the calculations step-by-step in R. (4pts)
2) Define your own function to compute the margin of error and then the CI of the mean. (4pts)
3) Use the built-in R function to calculate the CI of the mean. (2pts)