Confidence Intervals Using R

Orville D. Hombrebueno
odhombrebueno@nvsu.edu.ph

College of Teacher Education

Outline

- Introduction to **R**
  a. The **R** Environment
  b. RStudio
  c. Data Type and Objects
  d. Basic Commands
  e. Creating Functions in R
- Confidence Intervals for the Mean (Large Samples)
  a. Finding a Confidence Interval for a Population Mean
- Confidence Intervals for the Mean (Small Samples)
  a. Constructing a Confidence Interval for the Mean: t-Distribution
- Confidence Intervals for Population Proportions
  a. Constructing a Confidence Interval for a Population Proportion
  b. Finding a Minimum Sample Size to Estimate p
- Confidence Intervals for Variance and Standard Deviation
  a. Constructing a Confidence Interval for a Variance and Standard Deviation

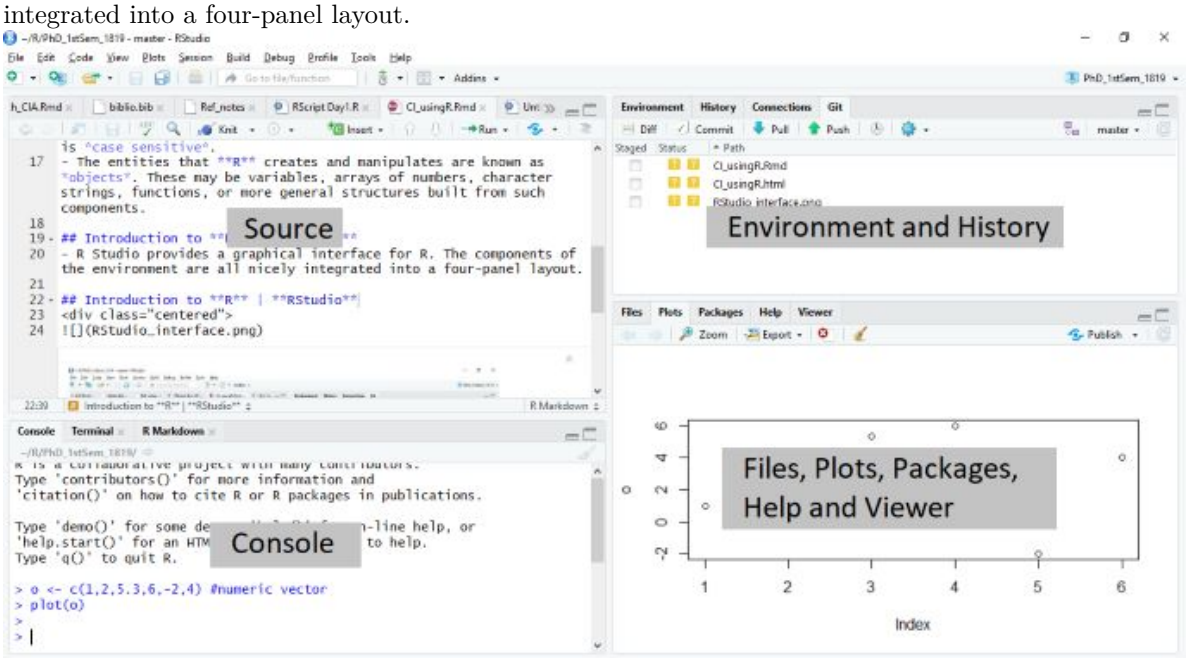
Introduction to R

The R Environment

- **R** is an integrated suite of software facilities for data manipulation, calculation and graphical display.
- It is an open-source software environment for statistical computing and graphics.
- It is an environment within which many classical and modern statistical techniques have been implemented.
- A few of these techniques are built into the base **R** environment, but many are supplied as packages.
- **R** is an *expression language* with a very simple syntax. It is *case sensitive*.
- The entities that **R** creates and manipulates are known as *objects*. These may be variables, arrays of numbers, character strings, functions, or more general structures built from such components.

RStudio

- R Studio provides a graphical interface for **R**. The components of the environment are all nicely integrated into a four-panel layout.



- The RStudio provides most of the desired features of **R** in an Integrated Development Environment (IDE).
- The **console** provides the command-line interface for interactive use of **R**. This is where users issue commands for **R** to evaluate.
- The **source** tab is a built-in text editor.
- The **Environment** tab is an interactive list of **R** objects.
- The **Files** tab displays the files and sub directories of a given directory.
- Display of graphics is rendered in the **Plots** tab.
- RStudio keeps a stack of past commands and allows one to scroll through them easily. This can be done using the up or down keys. In addition, the **History** tab allows one to scroll through past commands.
- The **Packages** tab allows users to effortlessly load, install, update, and/or delete packages in the library of packages.
- The **Help** tab is an output location for help commands and help search window.
- The **Viewer** tab is an advanced tab for local web content.

Data Type and Objects

- Scalars – atomic quantity and can hold only one value at a time.
  Examples: number, logical value, character(string)
- Vector – a sequence of data elements of the same basic type.
- Matrix – a collection of data elements in a rectangular layout.
- Data Frame – more general than a matrix, in that different columns can have different basic types.
- List – a generic vector containing other objects.

Basic Commands

- Before everything else, set your working directory. Setting the working directory is choosing a folder to save your work. You can set the working directory using the File tab or the function `setwd()`.
- Typing `?funcname` will cause **R** to open a new help file window with additional information about the function `funcname`.
- We can assign values in **R** using the "`<-`" operator.
  Example

```
> x <- 143
> x
```

```
[1] 143
```

```
> y <- 198
> y
```

```
[1] 198
```

```
> x + y
```

```
[1] 341
```

```
> z <- x + y
> z
```

```
[1] 341
```

- **R** uses *functions* to perform operations. To run a function called `funcname`, we type
  `funcname(input1, input2)`
  Example
  We use the function `c()` to create a vector of numbers.

```
> x <- c(1, 4, 3, 4, 4)
> x
```

```
[1] 1 4 3 4 4
```

- Note that the prompt, `>`, is not part of the command; rather, it is printed by **R** to indicate that it is ready for another command to be entered.
- Hitting the *up arrow key* multiple times will display the previous commands, which can then be edited.
- a plus sign, `+`, replacing `>` indicates that your code is not complete and that **R** is asking you to complete your code.
- The `ls()` function allows us to look at a list of all of the objects, such as data and functions, that we have in the environment. The `rm()` function can be used to delete any that we don't want.

Creating Functions in R

- We can write functions in R!
  Example

Suppose we want to write a function for average. The formula for average is

$$average = \frac{\sum x}{n}$$

where x is a vector and n is a scalar containing the number of elements in x.

```
> average <- function(x){
+   sum(x)/length(x)
+ }
```

```
> y <- c(4, 3, 5, 1, 7)
```

```
> average(y)
```

```
[1] 4
```

or

```
> average <- function(x){
+   a <- sum(x)
+   b <- length(x)
+   c <- a/b
+   return(c)
+ }
```

```
> average(y)
```

```
[1] 4
```

or

```
> mean(y)
```

```
[1] 4
```

Confidence Intervals for the Mean (Large Samples)

Finding a Confidence Interval for a Population Mean

1. Find the sample statistics n and $\bar{x}$.

$$\bar{x} = \frac{\sum x}{n}$$

   In **R** we have the function `mean()`.
2. Specify σ, if known. Otherwise, if $n \geq 30$, find the sample standard deviation s and use it as an estimate for σ.

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

   In **R** we have the function `sd()`.
3. Find the critical value z_c that corresponds to the given level of confidence.
   Example
   $c = 95\%$ or 0.95

```
> qnorm((1+0.95)/2)
```

```
[1] 1.959964
```

4. Find the margin of error E.

$$E = z_c \frac{\sigma}{\sqrt{n}}$$

5. Find the left and right endpoints and form the confidence interval.
   LEP

$$\bar{x} - E$$

   REP

$$\bar{x} + E$$

   Interval

$$\bar{x} - E < \mu < \bar{x} + E$$

   or

$$(\bar{x} - E,\ \bar{x} + E)$$

Example

Construct a 95% confidence interval for the population mean. Interpret your answer.

1. The stem-and-leaf plot shows the result of a random sample of airfare prices (in dollars) for a one-way ticket from Boston, MA to Chicago, IL.

```
The decimal point is 1 digit(s) to the right of the |

18 | 33
19 | 7
```

```
 20 | 99
 21 | 222333333366
 22 | 2222366888889
 23 | 88
```

Practice

2. A random sample of the closing stock prices for the Oracle Corporation for a recent year.

```
 [1] 18.41 18.32 22.86 14.47 16.91 18.65 20.86 19.06 16.83 20.71 20.74
[12] 18.42 17.72 20.66 22.05 20.85 15.54 21.04 21.42 21.43 15.56 21.74
[23] 22.34 21.97 18.01 22.13 22.83 21.81 19.11 21.96 24.34 19.79 22.16
[34] 17.97
```

Confidence Intervals for the Mean (Small Samples)

Constructing a Confidence Interval for the Mean: t-Distribution

1. Find the sample statistics n, $\bar{x}$, and s.

$$\bar{x} = \frac{\sum x}{n}, \quad s = \sqrt{\frac{\sum (\bar{x} - x)^2}{n-1}}$$

2. Identify the degrees of freedom, the level of confidence c, and the t_c.

$$\text{d.f.} = n - 1$$

Example
$c = 95\%$ or 0.95 and d.f. $= 16$

```
> qt((1+0.95)/2, df = 16)
```

```
[1] 2.119905
```

3. Find the margin of error E.

$$E = t_c \frac{s}{\sqrt{n}}$$

4. Find the left and right endpoints and form the confidence interval.
LEP

$$\bar{x} - E$$

REP

$$\bar{x} + E$$

Interval

$$\bar{x} - E < \mu < \bar{x} + E$$

or

$$(\bar{x} - E,\ \bar{x} + E)$$

Practice

Construct a 90% confidence interval for the population mean. Assume the population of each data set is normally distributed. Interpret your answer.

1. The annual earnings (in pesos) of 14 randomly selected engineers.

```
 [1] 63118 65740 72899 68500 66726 65554 69247 64963 68627 70448 71842
[12] 66873 74103 71138
```

2. The grade point averages (GPA) of 15 randomly selected college students.

```
 [1] 2.3 3.3 2.6 1.8 0.2 3.1 4.0 0.7 2.3 2.0 3.1 3.4 1.3 2.6 2.6
```

Confidence Intervals for Population Proportions

Constructing a Confidence Interval for a Population Proportion

1. Identify the sample statistics n and x.
2. Find the point estimate $\hat{p}$.
$$\hat{p} = x/n$$
3. Verify that the sampling distribution of $\hat{p}$ can be approximated by a normal distribution.
$$n\hat{p} \geq 5, \quad n\hat{q} \geq 5$$
4. Find the critical value z_c that corresponds to the given level of confidence c.
Example
$c = 95\%$ or 0.95

```
> qnorm((1+0.95)/2)
```

```
[1] 1.959964
```

5. Find the margin of error E.
$$E = z_c\sqrt{\frac{\hat{p}\hat{q}}{n}}$$
6. Find the left and right endpoints and form the confidence interval.
LEP
$$\hat{p} - E$$
REP
$$\hat{p} + E$$
Interval
$$\hat{p} - E < p < \hat{p} + E$$
or
$$(\hat{p} - E,\ \hat{p} + E)$$

Practice

1. In a survey of 7000 women, 4431 say they change their nail polish once a week. Construct a 95% confidence interval for the population proportion of women who change their nail polish once a week. Interpret your answer.
2. In a survey of 2303 U.S. adults, 734 believe in UFOs. Construct a 90% confidence interval for the population proportion of U.S. adults who believe in UFOs. Interpret your answer.

Finding a Minimum Sample Size to Estimate p

Given a c-confidence level and a margin of error E, the minimum sample size n needed to estimate p is
$$n = \hat{p}\hat{q}\left(\frac{z_c}{E}\right)^2.$$
This formula assumes that you have preliminary estimates of $\hat{p}$ and $\hat{q}$. If not, use $\hat{p} = 0.5$ and $\hat{q} = 0.5$.

Practice

You wish to estimate, with 99% confidence, the population proportion of U.S. adults who read fiction books. Your estimate must be accurate within 2% of the population proportion.

1. No preliminary estimate is available. Find the minimum sample size needed.
2. Find the minimum sample size needed, using a prior study that found that 47% of U.S. adults read fiction books.

Confidence Intervals for Variance and Standard Deviation

Constructing a Confidence Interval for a Variance and Standard Deviation

1. Verify that the population has a normal distribution.
2. Identify the sample statistic n and the degrees of freedom.

$$\text{d.f.} = n - 1$$

3. Find the point estimate s^2.

$$s^2 = \frac{\sum (\bar{x} - x)^2}{n - 1}$$

In **R** we have the `var()` function.

4. Find the critical value χ^2_R and χ^2_L that corresponds to the given level of confidence c.
In **R**,

$$\chi^2_R = \frac{1 + c}{2}, \quad \chi^2_L = \frac{1 - c}{2}.$$

Example
$c = 95\%$ or 0.95 and d.f. $= 17$

```
> qchisq((1 - 0.95)/2, df=17) #left-tail critical value
```

```
[1] 7.564186
```

```
> qchisq((1 + 0.95)/2, df=17) #righ-tail critical value
```

```
[1] 30.19101
```

5. Find the left and right endpoints and form the confidence interval for the population variance.

$$\frac{(n-1)s^2}{\chi^2_R} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_L}$$

6. Find the confidence interval for the population standard deviation by taking the square root of each endpoint.

$$\sqrt{\frac{(n-1)s^2}{\chi^2_R}} < \sigma < \sqrt{\frac{(n-1)s^2}{\chi^2_L}}$$

Practice

Assume each sample is taken from a normally distributed population and construct the indicated confidence intervals for (a) the population variance σ^2 and (b) the population standard deviation σ. Interpret the results.

1. As part of your spring break planning, you randomly selected 10 hotels in Cancun, Mexico, and record the room rate for each hotel. The results are shown in the stem-and-leaf plot. Use a 98% level of confidence.

```
The decimal point is 1 digit(s) to the right of the |

 6 | 9
 7 | 4
 8 |
 9 | 099
10 |
11 | 2
12 |
13 | 69
14 | 9
15 | 0
```

2. The weights (in pounds) of a random sample of 14 cordless drills are shown in the stem-and-leaf plot. Use a 99% level of confidence.

```
The decimal point is at the |

3 | 469
4 | 689
5 | 134579
6 | 01
```