

## Overview and Rationale

Being able to ask appropriate questions of data is an important part of the work of data analytics. It is also critical to be able to interpret the results of the analysis. This assignment is intended to familiarize you with the data sets and to get you thinking about key business questions you can ask and answer from this data.

This project will help you measure your understandings of basic concepts on analytics.

It will help you measure your skills to obtain basic descriptive statistics from a data set.

It will help you measure your skills to run hypothesis testing using different methods in R.

It will help you measure your skills on R, R Studio and R Markdown.

It will help you measure your skills to apply critical thinking to make meaningful observations of your data analysis results.

## Part 1. Title and Introduction

Create a nice title for your report.

Write a very informative introduction.

- (a) Using your own words, inform your audience what do you understand about ANOVA.
- (b) Explain the difference between one-way and two-way ANOVA.
- (c) Talk about their importance and provide practical applications on the industry of your interest.

Remember to use at least 3 different academic references.

## Part 2. Analysis section

As you have already learnt, for each task, enter all your codes in one single r chunk {r}, creating object names for all your codes and equations, and using the name of those objects to present your answers using inline R codes.

Pay special attention at the way your instructor prepared all her codes; several good examples are presented on Lecture 13. Also observe recommendations at the end of this document.

Be very organized.

### Task 1. Critical values

The main purpose of this task is not just to (1) measure your R coding skills to obtain the right answers, but also your skills to (2) prepare one single R chunk with object names for all your calculations and (3) the use of inline R codes to present your answers. Be very organized.

**\*\* Prepare your codes in one single R chunk {r task1}, and present your answers using inline R codes. \*\***

- a. For  $\alpha = 0.05$ ,  $n = 15$ , what is the critical value for a right-tailed distribution? Use  $z$  test.
- b. For  $\alpha = 0.05$ ,  $n = 15$ , what is the critical value for a right-tailed distribution? Use  $t$  test.
- c. For  $\alpha = 0.05$ ,  $n = 15$ , what is the critical value for a right-tailed chi-square test?
- d. For  $\alpha = 0.05$ ,  $df_N = 3$ ,  $df_D = 10$ , what is the critical value for a  $F$  test?
- e. For  $\alpha = 0.001$ ,  $n = 15$ , what is the critical value for a left-tailed distribution? Use  $z$  test.
- f. For  $\alpha = 0.001$ ,  $n = 15$ , what is the critical value for a left-tailed distribution? Use  $t$  test.
- g. For  $\alpha = 0.001$ ,  $n = 15$ , what is the critical value for a left-tailed chi-square test?

### Task 2. P values

The main purpose of this task is not just to (1) measure your R coding skills to obtain the right answers, but also your skills to (2) prepare one single R chunk with object names for all your calculations and (3) the use of inline R codes to present your answers. Be very organized.

**\*\* Prepare your codes in one single R chunk {r task2}, and present your answers using inline R codes. \*\***

- a. What is the p value for  $z = 2.4569$  with  $n = 256$ ?
- b. What is the p value for  $t = 2.4569$  with  $n=22$ ?
- c. What is the p value for  $z = -3.0349$  with  $n = 256$ ?
- d. What is the p value for  $t = -3.0349$  with  $n=22$ ?
- e. What is the p value for Chi square test = 5.39563 with  $n=11$ ?
- f. What is the p value for Chi square test = 5.39563 with  $n=16$ ?
- g. What is the p value for Chi square test = 5.39563 with  $n=21$ ?
- h. Compare p values among e-f-g and explain the reason of differences even the test values are the same.

### Task 3 ANOVA analysis 1

**\*\* Prepare your codes in one single R chunk {r task3}, and present your answers using inline R codes. \*\***

In the following study, weight loss was measured on 15 women after 6 weeks of receiving a particular diet. A group of volunteers used diet 1, another group diet 2 and a third group, diet 3.

The weight loss per person is presented in the following table.

Weight Loss in Females		
Diet 1	Diet 2	Diet 3
3.8	1.8	7
6	2	5.6
0.7	1.7	3.4
2.9	4.3	6.8
2.8	5.2	7.8

Using the procedure presented in class, Lecture 13: slides 14 to 21, use R to calculate the  $F$  value and test the hypothesis that there are no differences on diet results.

$$H_0: \mu_1 = \mu_2 = \mu_3$$

$H_1$ : At least one mean is different from the others.

Remember that you need to solve the following formulas:

$$\begin{aligned} \text{d.f.N.} &= k - 1 \\ \text{d.f.D.} &= N - k \end{aligned}$$

$$s_B^2 = \frac{\sum n_i (\bar{X}_i - \bar{X}_{GM})^2}{k - 1}$$

$$\bar{X}_{GM} = \frac{\sum X}{N}$$

$$s_W^2 = \frac{\sum (n_i - 1) s_i^2}{\sum (n_i - 1)}$$

$$F = \frac{s_B^2}{s_W^2}$$

Present the values for **(a)** grand mean, **(b)** between-group variance, **(c)** within-group variance, and **(d)**  $F$  test using inline R codes.

**e.** Run your hypothesis testing using  $\alpha = 0.05$  and explain your conclusions.

#### Task 4, Scheffé test

**\*\* Prepare your codes in one single R chunk {r task4}, and present your answers using inline R codes. \*\***

If your calculation on task 3 were correct, you must have rejected  $H_0$ . Now you need to find which mean is different.

For this purpose, apply the Scheffé test. The formula for Scheffé test is:

$$F_S = \frac{(\bar{X}_1 - \bar{X}_2)^2}{s_W^2 [(1/n_1) + (1/n_2)]}$$

Prepare your codes in one single R chunk and present your values for **(a)** Diet 1 versus Diet 2, **(b)** Diet 1 versus Diet 3, and **(c)** Diet 2 versus Diet 3, using inline R codes.

**d.** Find the Scheffé critical value using the r code: `qf(alpha, dfN, dfD, lower.tail = FALSE)`

**e.** Compare your three *F*s values with the Scheffé critical value, and conclude which diet caused a significant weight loss.

### Task 5. ANOVA analysis 2 using R codes.

**\*\* Prepare your codes in one single R chunk {r task5}, and present your answers using inline R codes. \*\***

For this task, import the data from file **Project7\_data.xlsx**.

**a.** Use the `mutate()` code to create a new calculated field named: `weight_loss`. You must subtract the weight at week 6 from the weight at the beginning of the study, week 0. Prepare the formula with a name but do not present the new data on your report, if it is correct, your next answers will be correct.

**b.** Create one data subset with the females. under gender it is code = 0. You can use for example code `filter(gender == "0")`. Prepare the formula with a name but do not present the new data on your report, if it is correct, your next answers will be correct.

**c.** Run ANOVA analysis for females, to test Diet versus weight loss. Present the ANOVA result on your report.

**d.** Present the summary of your ANOVA and mention the *F* test value on your report.

**\*\* Notice: To present ANOVA result and summary, do it from the R chunk, not in inline R codes.**

**e.** At  $\alpha = 0.001$ , do you have enough evidence to reject  $H_0$ ? Explain your conclusion. Lecture 13 shows how to obtain the *F* critical value.

**Important information on aov() R code:** When you run the ANOVA analysis using `aov()`, you must remember that diet is being presented as "numeric". This will affect your ANOVA analysis and prevent from running the Tukey test. To solve this issue, present Diet as a factor, I will help you with the code:

```
aov(weight_loss ~ as.factor(Diet), data = name)
```

### Task 6. Tukey test.

**a.** Using the information from task 5, run a Tukey test to find where the difference in means is located. The R code and procedure are explained in Lecture 13. Present the answer from inside the R chunk.

**b.** What are the adjusted **p values** of your Tukey (*q*) test for all Diet combinations?

**c.** Again, at  $\alpha = 0.001$ , which diet caused the most significant decrease in weight loss? Explain your conclusion.

### Part 3. Conclusions and Bibliography.

a. Write a very informative conclusions section following indications and recommendations giving by your instructor,

*Be mindful to make an overall observation of the whole project, the meaning of the results you obtained regarding the direction of the project, explain any new skills you gained.*

b. Present a bibliography section with the references you used on your report.

*Technically speaking, if you do not mention any references in the main text of your report, then it is like you did not use any, even if you add a list at the end. Present references in the main text of your reports, use either only the first author's last name and year, e.g., (Bluman, 2017) and then list them in the bibliography section in alphabetical order, or use a number in order of use or appearance, then list them in the bibliography section in that numerical order.*

### Part 4. Appendix and Acknowledgments

a. Present an appendix title to mention the Rmd document you are attaching to your report.

b. For a final project, it is a courtesy to add an acknowledgments section to thank anybody that you feel helps you in anyway during the project and whole course (classmates, TA, instructor, advisor, etc.).

### Format & Guidelines

For this week assignment you must submit 2 files:

**Submit your HTML report** containing all your findings along with important statistical issues.

**Submit the original Rmd** file you used to produce your report.

Please remember: your report is very important, make it look professional, make it as short as possible but containing all the relevant information, tell me what you learnt, and using deep critical thinking, provide examples of practical applications.

## Appendix 1. Proper use of object names in R Markdown

### Object names

Dee Chiluiza, PhD

4/22/2021

When preparing codes on R Markdown, provide a name to all your objects (codes or any equation) to prevent their automatic presentation on your report.

#### Wrong way to do it

Observe the example below, If I forget to name my objects , their outcomes are automatically presented on the HTML report. Observe that each code that does not have a name is presented individually in gray boxes and their outcome below on white boxes. This is considered un-organized for your reports.

```
pnorm(3.4531)
```

```
## [1] 0.9997229
```

```
qt(0.001, 12)
```

```
## [1] -3.929633
```

#### Correction

To prevent what you see above, I use names for all my objects, then I use inline R codes to present the answers. Also notice that I use round() only when I am ready to present my answers.

```
value1 = pnorm(-3.4531)
value2 = qt(0.001, 12)
```

The answer for value 1 is: 0.0003

The answer for value 2 is: -3.93

Here are the actual codes used for the above HTML document. Notice the use of `format(x, scientific=F)` to prevent presentation of numbers with long decimas as scientific notation:

```
1 ---
2 title: "Object names"
3 author: "Dee Chiluiza, PhD"
4 date: "4/22/2021"
5 output: html_document
6 ---
7
8 When preparing codes on R Markdown, provide a name to all your objects (codes or any equation) to prevent
9 their automatic presentation on your report.
10
11 <P><BR>
12 <FONT SIZE=4, COLOR="#A11515"><B>Wrong way to do it</B></FONT>
13 Observe the example below, If I forget to name my objects , their outcomes are automatically presented on
14 the HTML report. Observe that each code that does not have a name is presented individually in gray boxes
15 and their outcome below on white boxes.
16 This is considered un-organized for your reports.
17
18 ```{r task20}
19 pnorm(3.4531)
20 qt(0.001, 12)
21 ```
22
23 <P><BR>
24 <FONT SIZE=4, COLOR="#A11515"><B>Correction</B></FONT>
25 To prevent what you see above, I use names for all my objects, then I use inline R codes to present the
26 answers. Also notice that I use round() only when I am ready to present my answers.
27
28 ```{r task21}
29 value1 = pnorm(-3.4531)
30 value2 = qt(0.001, 12)
31 ```
32
33 <P>The answer for value 1 is: ``r format(round(value1, 4), scientific=F)``
34 <P>The answer for value 2 is: ``r round(value2, 3)``
```

## Appendix 2. Points divisions per task.

ALY2010 final project points divisions per task	Points
Title	5
Well-presented title	5
Introduction	25
a. Good information presented.	5
b. Good information presented.	5
c. Good information presented.	5
Used and presented references using: (author's last name, year).	5
Introduction is well organized.	5
Task 1. Critical values.	25
1.a. Answer is correct.	3
1.b. Answer is correct.	3
1.c. Answer is correct.	3
1.d. Answer is correct.	3
1.e. Answer is correct.	3
1.f. Answer is correct.	3
1.g. Answer is correct.	3
Prepared all codes in one single {r task1} R chunk.	1
All codes and calculations prepared using object names.	1
Used object names to present answers using inline R codes.	1
Task presentation is good and well organized.	1
Task 2. P values	28
2.a. Answer is correct.	3
2.b. Answer is correct.	3
2.c. Answer is correct.	3
2.d. Answer is correct.	3
2.e. Answer is correct.	3
2.f. Answer is correct.	3
2.g. Answer is correct.	3
2.h. Good answer and explanation.	3
Prepared all codes in one single {r task2} R chunk.	1
All codes and calculations prepared using object names.	1
Used object names to present answers using inline R codes.	1
Task presentation is good and well organized.	1
Task 3. ANOVA analysis 1	22
Solved all formulas creating object names using one single R chunk.	1
Presented all answers using inline R codes.	1
a. Grand mean is correct	4
b. Between-group variance value is correct.	4
c. Between-group variance value is correct.	4
d. F test value is correct.	4
e. Hypothesis testing conclusions are correct.	4

Task 4. Scheffé test	22
Solved formula creating object names using one single R chunk.	1
Presented answer using inline R codes.	1
a. Diet 1 versus Diet 2 value is correct.	4
b. Diet 1 versus Diet 3 value is correct.	4
c. Diet 2 versus Diet 2 value is correct.	4
d. Scheffé critical value is correct.	4
e. Observations and conclusions are correct.	4
Task 5. ANOVA analysis 2 using R codes	16
Solved all formulas creating object names using one single R chunk.	1
a. Created new calculated field.	3
b. Created data subset.	3
c. Presented ANOVA result	3
d. Presented ANOVA summary	3
e. Hypothesis testing conclusions are correct.	3
Task 5. Tukey test	14
Solved formula creating object names using one single R chunk.	1
a. Presented Tukey test results.	4
b. Presented correct adjusted p values (all three).	5
c. Hypothesis testing conclusions are correct.	4
Conclusions	25
Make overall analysis of the whole project	10
Clearly point to new skills gained during the project	10
Make substantial recommendations	5
Bibliography	10
Good presentation of references	8
Appendix	5
Mention appendix	5
Acknowledgment	5
Added an acknowledgments section.	5
Deadline submission	
If report submitted after the deadline, 10% reduction.	
	200