

FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY

SEMESTER 1 SESSION 2022/2023

BITI 2233 STATISTICS AND PROBABILITY

ASSIGNMENT 2 using Rstudio 15%

INSTRUCTIONS:

- a) This assignment must be done in a group of 3 members.
- b) Use R software (R Studio) to complete this assignment.**
- c) Your submission should include:
 - a. Your report, which includes:
 - i. Snippet of commands, answers, and respective graph in the report.
 - ii. Each figure in the report must be properly titled.
 - iii. Attach the data and the results of your calculation in the appendix.
 - iv. Include the names of each member and relevant details.
 - b. R code file in R file,
 - c. Data file in xls file.
- d) Put all files into 1 folder. Zip the folder. Submit the zip file onto the Ulearn/Assessment.
- e) The due date is on **18 January 2023 (refer to your lecturer for the exact date)**. Any late submission will be penalized.

IMPORTANT:

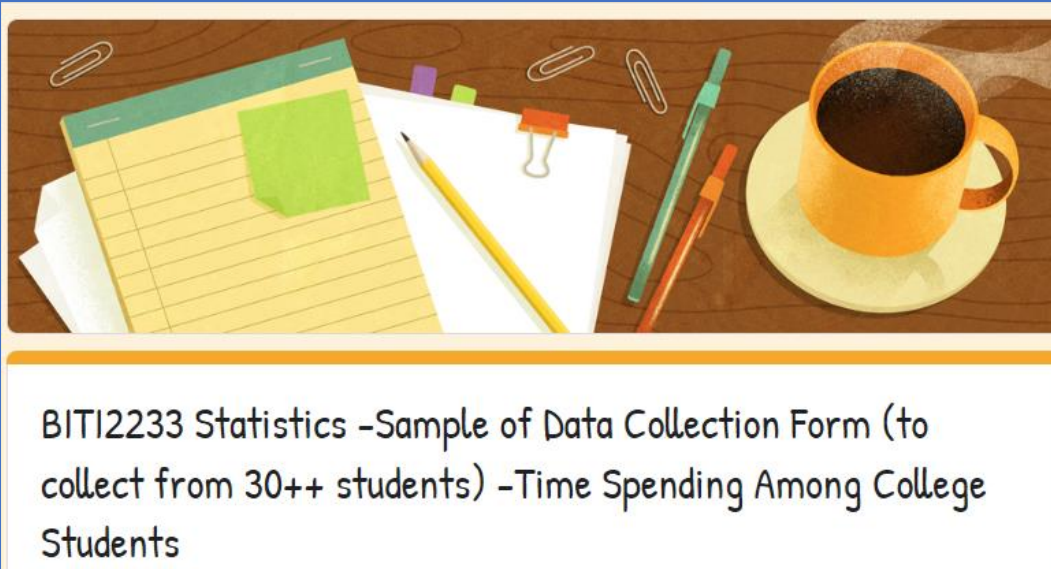
- a) You must understand your codes!
- b) Do not copy other groups' work or allow your group work to be copied.
- c) Any plagiarism detected will be penalized.
- d) Final marks are individually-based and will be given based on the lecturer's evaluation /after the presentation if required.

Question 1 (20 marks)

Estimation of population parameter using sample data

(To use data from Assignment 1)

You are required to analyze the data, then explain the finding. Use whatever data you managed to collect using the google form in Assignment 1:



1. **Assume** the data you obtained from the survey is a **sample data**; where the population would be the teenager from all universities, while you are collecting data from UTeM only. Analyze the data about **their spending of time per day** (*Choose only ONE (1) set of data, semester week OR semester break). For example:

| VARIABLE (SEMESTER WEEK) | EXAMPLE DATA | YOUR REAL DATA |
|--|--|---|
| | No of data (n): 31 | No of data (n): ____ |
| <u>Qualitative:</u> A3-Gender | <u>Proportion from 31 data:</u> Boys (p^{\wedge}): 70% Girls (q^{\wedge}): 30% | <u>Proportion from 31 data:</u> Boys (p^{\wedge}): ____ Girls (q^{\wedge}): ____ |
| <u>Quantitative:</u> C1: Class time/Going to class C2: Homework/Study related C3: Sleeping C4: Socializing C5: On screen C6: Other leisure C7: Eating C8: Grooming C9: Housework C10: Paid work C11: Volunteering C12: Errands | <u>Average from 31 data:</u> Duration-C1: 6 hours Duration-C2: 6 hours Duration-C3: 6 hours Duration-C4: 2 hours Duration-C5: 2 hours Duration-C6: 2 hours Duration-C7: 2 hours Duration-C8: 1.8 hours Duration-C9: 1.2 hours Duration-C10: 0.5 hour Duration-C11: 0.5 hour Duration-C12: 0.5 hour | <u>Average from 31 data:</u> Duration-C1: ____ hours Duration-C2: ____ hours Duration-C3: ____ hours Duration-C4: ____ hours Duration-C5: ____ hours Duration-C6: ____ hours Duration-C7: ____ hours Duration-C8: ____ hours Duration-C9: ____ hours Duration-C10: ____ hours Duration-C11: ____ hours Duration-C12: ____ hours |

| <u>Quantitative:</u> | <u>Standard deviation from 31 data:</u> | <u>Standard deviation from 31 data:</u> |
|-------------------------------|---|---|
| C1: Class time/Going to class | Duration-C1: 6 hours | Duration-C1: __ hours |
| C2: Homework/Study related | Duration-C2: 6 hours | Duration-C2: __ hours |
| C3: Sleeping | Duration-C3: 6 hours | Duration-C3: __ hours |
| C4: Socializing | Duration-C4: 2 hours | Duration-C4: __ hours |
| C5: On screen | Duration-C5: 2 hours | Duration-C5: __ hours |
| C6: Other leisure | Duration-C6: 2 hours | Duration-C6: __ hours |
| C7: Eating | Duration-C7: 2 hours | Duration-C7: __ hours |
| C8: Grooming | Duration-C8: 1.8 hours | Duration-C8: __ hours |
| C9: Housework | Duration-C9: 1.2 hours | Duration-C9: __ hours |
| C10: Paid work | Duration-C10: 0.5 hour | Duration-C10: __ hours |
| C11: Volunteering | Duration-C11: 0.5 hour | Duration-C11: __ hours |
| C12: Errands | Duration-C12: 0.5 hour | Duration-C12: __ hours |

2. Use RStudio to answer the following questions:

1. Display:

- Display all the data collected from the survey in Assignment 1.
- Display average and standard deviation for each of the variables (A3 & C1-C12).
- Display only data for semester week in a table, 'Table-semweek'.

2. Analyze:

- Compute a **90%** confidence interval for the qualitative variable (A3).
- Compute a **95%** confidence interval for the qualitative variable (A3).
- Compute a **99%** confidence interval for the qualitative variable (A3).
- Compute a **90%** confidence interval for **each** of the qualitative variables (C1-C12).
- Compute a **95%** confidence interval for **each** of the qualitative variables (C1-C12).
- Compute a **99%** confidence interval for **each** of the qualitative variables (C1-C12).

3. Explain the finding:

- Justify the difference found in the THREE (3) different confidence intervals (90%, 95%, 99%) for the qualitative variable (A3).
- Justify the difference found in the THREE (3) different confidence intervals (90%, 95%, 99%) for **each** of the qualitative variables (C1-C12).

3. Example of R codes for ONE **quantitative** variable:

The code below demonstrates how to compute a 95% confidence interval for the true population mean weight of the above data.

```
n <- 30
xbar <- 200
s <- 12
```

Let's calculate the margin of error

```
margin <- qt(0.975, df=n-1) * s / sqrt(n)
```

We can now determine the lower and upper confidence interval boundaries.

```
lowerinterval <- xbar - margin
lowerinterval
[1] 195.5191
upperinterval <- xbar + margin
upperinterval
[1] 204.4809
```

The genuine population mean weight of data has a 95% confidence interval of [195.5191, 204.4809].

*Source: <https://www.r-bloggers.com/2021/11/calculate-confidence-intervals-in-r/>

4. Report the answer. You must include the question, R codes, and output into the REPORT together with the other questions.

Question 2 (20 marks) Refer to Advertising Budget and Sales.xls

As a marketing manager, you want to analyse the cost spent on different channels of advertisement and its corresponding sales. There are three channels of advertisement which are radio ad, TV ad, and e-media ad. The analysis is to check whether more budget should be allocated to the advertisement cost to increase sale. The company claims that there is no additional budget required and believe that the sales' average is always more than 10 million. The related monthly data for 200 months is collected. The dataset captures the advertisement cost and its sales revenue in millions. Based on the dataset (Advertising Budget and Sales.xls), check the claim whether it can be accepted or not.

- a. Randomly select 50 out of 200 data. You must do the random selection in R and show how the data is selected. Please ensure that there is no duplication in the selection. You may use sample command.
- b. Based on the selected samples, conduct a basic descriptive analysis on the three channels of advertisement. Give the summary of the analyses (which should include at least a graph and the average and standard deviation of the sample).
- c. Conduct a test of hypothesis with the intent of showing that, at the 0.05 level of significance, the average of the sales is at least 10 million per month. Report the hypothesis testing process step by step and conclude the findings.
- d. Based on (b) and (c), write a short analysis on the findings. Describe the sales' performance and give several recommendations to the company on what strategy/action should be adopted in order to improve the overall sales revenue.

(3 marks)

Question 3 (20 marks)

NEW FUNCTION IN ONLINE STORE APPLICATION

A new e-commerce business company recently planned to add a new function to their online store application. The function can help the shoppers to acquire the product more quickly based on the genders to increase the purchases. Your team will work with this company to conduct a small sample survey to compare the average amount of purchases made by male and female UTeM students on various online store platforms to support this new function. On the survey day, record their purchases, the name of online store and any other related details and answer the following questions.

- A. Using RStudio, conduct a descriptive analysis with any appropriate graphs from your survey data.
- B. Using RStudio, construct a 95% confidence interval for the difference in mean purchases between all male and female UTeM students.
- C. Using RStudio, determine whether there are any differences in the purchases made by all male and female students who used online stores (you may be interested to see the difference based on the type of online stores).
- D. From the results above, what inferences may be drawn to help this company in including this new function to their online store application.

Assume that the purchases of all such male and female UTeM students are normally distributed with equal and unknown population standard deviation.

Question 4 (20 marks) Refer to production2.csv

Consider a production process in which one or more workers are engaged in a variety of tasks. For such a process, the total time spent in production varies as a function of the size of the work pool and the level of output of the various activities. At a large metropolitan department store, the number of hours worked (y) per day by the clerical staff may depend on the following variables.

x_1 – number of pieces of mail processed

x_2 – number of gift certificates sold

x_3 – Number of store charge accounts transacted

x_4 – Number of change order transactions or return processed

x_5 – Number of checks cashed

The data of working days are attached. Please refer to dataset **production2.csv**. The store's production engineer wants to model number of hours worked with simple linear regression model. Answer the following questions and provide the output as well as R scripts for the answer.

- Randomly select 35 observations without replacement from the dataset by using R.
- Provide scatter plot for each model (y vs x_1 , y vs x_2 , y vs x_3 , y vs x_4 , y vs x_5). What can you conclude from these scatter plots?
- Use the least squares formulas to fit five straight-line models-one for each independent variable-for predicting y .
- Interpret the estimated slope coefficient b in each model and test the utility of each model by testing $H_0: \beta = 0$ against $H_1: \beta \neq 0$.
- Find the coefficient correlation r and coefficient of determination r^2 , for each model. Explain and interpret the values of r and r^2 for each relationship. Indicate which independent variables predicts y best for the data.
- Conclusion – conclude and summarize your findings (you can give support/justification why the selected independent variable is the best to predict y).

Question 5 (20 marks)

Sodium Contents of Foods The amount of sodium (in milligrams) in one serving for a random sample of three different kinds of foods is listed. At the 0.05 level of significance, is there sufficient evidence to conclude that a difference in mean sodium amounts exists among condiments, cereals, and desserts?

| Condiments | Cereals | Desserts |
|------------|---------|----------|
| 270 | 260 | 100 |
| 130 | 220 | 180 |
| 230 | 290 | 250 |
| 180 | 290 | 250 |
| 80 | 200 | 300 |
| 70 | 320 | 360 |
| 200 | 140 | 300 |
| | | 160 |

The calculation of the ANOVA test must perform using R

- Use ANOVA to test for any significant differences between the means.
- What is the purpose of this study?

-QUESTIONS END-