



Australian
National
University

Research School of Finance, Actuarial Studies and Statistics

ASSIGNMENT

Semester 1, 2021

STAT7055 Introductory Statistics for Business and Finance

© 2021 ANU

INSTRUCTIONS TO STUDENTS

Due Date

- The assignment is due at 6:00pm on Thursday May 20. Note that this is different to the time and date stated in the class summary.
- Late submission of the assignment is not permitted. An assignment submitted without an extension after the due date will be awarded a mark of 0.

Obtaining your Assignment

- There are different versions of the assignment and each student will be assigned a particular version of the assignment.
- Therefore, you **must** log in to Wattle with your own ANU credentials and download your assignment directly from Wattle.

Writing your Assignment

- The assignment is an individual piece of assessment and must be completed on your own.
- You will be required to write a report in an R Markdown document that contains both R code and written text. An example of an R Markdown document, which you can use as a template, has been provided on Wattle.
- When answering the assignment questions in your report, you will need to include all your R code that you used to calculate any answers and you must also write your answers in proper sentences. For example, if you are required to calculate a sample mean, then you would include your R code for calculating the sample mean and you would also write a proper sentence in the report such as "The sample mean is equal to ...".
- Make sure to be clear and concise in your answers.

- A good way to approach writing your report is to imagine that you are a statistical consultant and that a client has asked you to do some statistical analyses. When presenting the results of your analyses to the client, you wouldn't just give them pages of R code or pages of R output. Rather, you should give them a proper report which clearly outlines and explains the results of the analyses and which also includes the R code used to produce the results.
- Once you have finished writing your report in your R Markdown document, you will need to render the document by pressing the Knit button in RStudio to create a HTML file of your report.

Submitting your Assignment

- Submission of the assignment will be through Wattle via Turnitin.
- A Turnitin link with further details regarding assignment submission will be provided on Wattle.
- For submission you will need to submit two files: the R Markdown file of your report (i.e., a “.Rmd” file) and the rendered HTML file of your report produced by pressing the Knit button in RStudio (i.e., a “.html” file). No other file types can be submitted, e.g., “.R”, “.docx”, “.RData”, etc. files will not be accepted.
- Please name your two files as “uNNNNNNN.Rmd” and “uNNNNNNN.html”, where uNNNNNNN is your student number.

Other Important Details

- You may only use built-in functions available in base R and you are not permitted to use functions in any additional R packages.
- Round all final numeric answers to 4 decimal places. However, as you will be using R, keep all decimals during all intermediate steps to ensure the accuracy of your final numeric answer.
- Please use the **help** function if you want to learn more about a particular R function, e.g., enter **help(mean)** in the R console to learn more about the **mean** function.
- For questions that require writing mathematical symbols, you are welcome to use shorthand notation, provided you make the meaning clear (e.g., using “Mu” for μ , or “!=” for \neq).
- You may use either R or the statistical tables on Wattle for finding z -values, normal probabilities, or critical values for the t and F distributions. If you use the statistical tables, follow the guidelines below:
 - When using the normal tables to look up a z -value to find the corresponding probability, round the z -value to 2 decimal places.
 - When using the normal tables to look up a probability to find the corresponding z -value, look up the closest probability in the table, or if the probability lies exactly in the middle between two probabilities in the table, choose the mid-point of the two corresponding z -values.
 - When looking up a degree of freedom in the t or F tables, if you cannot find the exact degree of freedom in the table, choose the closest degree of freedom.

Question 1 [17 marks]

A technology company has a very intensive hiring process, where each applicant is required to complete four different tests. These tests are designed to assess various qualities, skills and abilities that the company is looking for in prospective employees. The technology company has collected some data on the applicants and their test scores. Specifically, for a random selection of 500 applicants, they have recorded the following for each applicant: their age in years (**Age**), their undergraduate degree (**Degree**) and their score in each of the four tests (**Test1**, **Test2**, **Test3** and **Test4**). The data is stored in the file `AssignmentData.RData` in the dataframe `Q1.df`.

- (a) [4 marks] Use some sample statistics to describe the scores for Test 3. Specifically, based on the definitions given in the lectures, calculate the sample coefficient of variation, the sample median and the sample range of the scores.
- (b) [4 marks] Create a boxplot and a histogram of the scores for Test 3. Make sure to give each plot a proper descriptive title and label the x -axis of the histogram appropriately (do not just use the default title or labels). Based on these plots, describe the distribution of scores for Test 3. Be specific in your description, making sure to mention any interesting and/or important aspects of the distribution.
- (c) [4 marks] The company wonders whether the distribution of scores for Test 3 is different between applicants who have an undergraduate degree in Computing and applicants who have an undergraduate degree in Engineering. Create separate histograms of the scores for Test 3 for these two groups of applicants (i.e., one histogram for applicants with a Computing degree and one histogram for applicants with an Engineering degree). Make sure to give each histogram a proper descriptive title and a label for the x -axis. Based on these histograms, describe any differences or similarities in the distribution of scores for Test 3 between these two groups of applicants.
- (d) [2 marks] The company wonders whether the spread in scores might be similar for all tests. Determine whether or not the spread of the scores for Test 3 and Test 4 are similar. Do not conduct a hypothesis test, but make sure to provide a clear justification for your answer based on the data.
- (e) [3 marks] Test whether the mean score for Test 3 is greater than 18. Clearly state your hypotheses and use a significance level of $\alpha = 10\%$. **Do not** use any functions available in R or any R package that are designed to perform hypothesis tests.

Question 2 [14 marks]

For this question you will be required to generate some sample data in R.

- (a) [2 marks] First, in a single line of code, specify the seed for the random number generator in R by using the `set.seed` function with your student number (without the “u”) as the `seed` argument. For example, if your student number is `u1234567` then you would use the line of code `set.seed(1234567)`. Next, in a single line of code, create a vector consisting of 100 observations that are randomly generated from a normal distribution with mean $\mu = 65$ and variance $\sigma^2 = 182.25$ and call this vector `x` (representing the variable X). Finally, in a single line of code, create a vector consisting of 105 observations that are randomly generated from a uniform distribution between $a = 42$ and $b = 92$ and call this vector `y` (representing the variable Y). These three lines of code must be executed in succession with no other lines of code in between.

For the remaining parts of this question, assume that the values of μ , σ^2 , a and b are all unknown and all that you have available is the sample data you generated in part (a).

- (b) [4 marks] Calculate an 84% confidence interval for the population mean of the Y values. Interpret the confidence interval. **Do not** use any functions available in R or any R package that are designed to calculate confidence intervals.
- (c) [4 marks] Test whether the population proportion of X values that are greater than 75 is less than 0.353. Clearly state your hypotheses, making sure to define any parameters, and use a significance level of $\alpha = 5\%$. **Do not** use any functions available in R or any R package that are designed to perform hypothesis tests.
- (d) [4 marks] If the X values and Y values are treated as independent samples, test whether the population proportion of X values that are greater than 75 is greater than the population proportion of Y values that are greater than 88. Clearly state your hypotheses, making sure to define any parameters, and use a significance level of $\alpha = 10\%$. **Do not** use any functions available in R or any R package that are designed to perform hypothesis tests.

Question 3 [23 marks]

Some data were collected on the university entrance scores for year 12 students from four high schools over a period of four years. For each year, a sample of year 12 students were randomly selected from each high school and their university entrance scores were recorded. For a given high school, the same number of year 12 students were selected each year. However, within a given year, the number of year 12 students selected may differ between the four high schools. The data is stored in the file `AssignmentData.RData` in the dataframe `Q3.df`. For a given year, the university entrance scores for all students across all four high schools are given in the column named by the year (`Year2005`, `Year2006`, `Year2007` and `Year2008`) and the high school (1, 2, 3, or 4) to which each student belonged is given in the column named `HighSchool`.

For this question, you will be analysing the university entrance scores from 2007.

- (a) [2 marks] Calculate the sample mean university entrance score for each high school.
- (b) [2 marks] Calculate the sample variance of university entrance scores for each high school.
- (c) [3 marks] Test whether the population variance of university entrance scores is the same for high school 1 and high school 2. Clearly state your hypotheses and use a significance level of $\alpha = 5\%$. **Do not** use any functions available in R or any R package that are designed to perform hypothesis tests.
- (d) [3 marks] Test whether the population mean university entrance score for high school 2 is greater than that for high school 1 by more than 7. Clearly state your hypotheses and use a significance level of $\alpha = 5\%$. **Do not** use any functions available in R or any R package that are designed to perform hypothesis tests.

You will now conduct a one-way ANOVA on the university entrance scores from 2007 with high school as the factor.

- (e) [8 marks] Discuss whether or not the assumptions for a one-way ANOVA hold for this data. You do not need to conduct any hypothesis tests, but make sure to provide clear justifications for your answer.
- (f) [2 marks] Calculate the sum of squares for treatment for the one-way ANOVA. **Do not** use any functions available in R or any R package that are designed to perform an ANOVA.
- (g) [3 marks] Test whether the population mean university entrance score is the same for all four high schools. Clearly state your hypotheses and use a significance level of $\alpha = 5\%$. **Do not** use any functions available in R or any R package that are designed to perform hypothesis tests or an ANOVA.

Question 4 [16 marks]

A think tank has been developing aptitude tests which they are hoping could eventually be used as a replacement for IQ tests. They have conducted a long-term study where they selected a random sample of 200 people and, for each person, recorded their scores in an age-appropriate aptitude test every five years from age 5 to age 25 (`Age5`, `Age10`, `Age15`, `Age20` and `Age25`). The data is stored in the file `AssignmentData.RData` in the dataframe `Q4.df`. The think tank is interested in whether the score in the aptitude test taken at age 5 could be used to predict the score in later year aptitude tests.

For this question, you will be analysing the aptitude test scores at ages 5 (`Age5`) and 20 (`Age20`).

- (a) **[3 marks]** Create a scatter plot of the aptitude test scores at age 20 against the aptitude test scores at age 5. Make sure to give your plot an appropriate title and appropriate labels for the x and y axes. Describe the relationship between these two variables.
- (b) **[3 marks]** Test whether the correlation between the aptitude test scores at age 20 and the aptitude test scores at age 5 is greater than zero. Clearly state your hypotheses and use a significance level of $\alpha = 5\%$. **Do not** use any functions available in R or any R package that are designed to perform hypothesis tests.
- (c) **[2 marks]** Fit a simple linear regression model with aptitude test scores at age 20 as the dependent variable and aptitude test scores at age 5 as the independent variable. Write down the estimated regression model.
- (d) **[5 marks]** Discuss whether or not the assumptions for a simple linear regression model hold for this data, making sure to provide clear justifications for your answer.
- (e) **[3 marks]** Test whether the intercept is less than 10. Clearly state your hypotheses and use a significance level of $\alpha = 10\%$. **Do not** use any functions available in R or any R package that are designed to perform hypothesis tests.

END OF ASSIGNMENT