**RESEARCH SCHOOL OF FINANCE, ACTUARIAL STUDIES AND STATISTICS**

STAT1003 Statistical Techniques Assignment

Semester 1, 2024

---

INSTRUCTIONS:

- This assignment is due by 23:59 pm on Thursday, 16 May 2024.
- This assignment is worth 10% of your overall marks for this course, and is nonredeemable.
- Your assignment report should be typed. It should be submitted as a Word .docx file. You will need to include plots or outputs in your report. Make sure your plots are easy to read. Clearly label each part. Plots should include a title, a caption, axis labels and a legend (if applicable) and be self-explanatory without reading the report text.
- You are required to use R to do data analysis. Your R code and the relevant R output needs to be included for each part.
- Any references you use should be cited in-text and included in a References list at the end of your report.
- Your assignment must be your own work. Turnitin will compare similarity to online content so you can check the Turnitin similarity score. Note that Turnitin may give a high similarity score because of similarity between tables, captions or the report title, or standard wording about confidence intervals etc. So don't panic if your Turnitin score is high! Just check any links provided to make sure you haven't inadvertently over-borrowed from another source. We don't take action purely based on the Turnitin score – this would just be a trigger for me to do some further investigation of a possible issue.
- You can discuss ways to approach the problem with fellow students - just don't copy text or R code. A good general rule is that you should not be typing your assignment while looking at another student's assignment or R screen. Instead, discuss the general approach or take a look at another student's report for ideas, then perform coding and writing alone. You should not provide your report or R code electronically to another student, or give them a hard copy to take away, as this may facilitate plagiarism or cheating.
- ChatGPT and other AI or Large Language Models tools should not be used in this assignment. Use of these tools may be detected and investigated. For example, they may result in similar code or text from multiple students, often all using methods that were not taught in the course. Making cosmetic changes will not necessarily hide this. If you do decide to use a method not in the course, please justify in detail why no method from STAT1003 was applicable, and give reference/s to a textbook supporting your approach.
- You are welcome to ask me questions via the Wattle discussion forum or by email (robert.clark@anu.edu.au), but replies may take up to two business days. I won't answer

questions directly related to the assignment (e.g., how do I plot these particular variables from the assignment?) but I will probably answer generic questions (e.g., how do I plot the relationship between a continuous and a categorical variable?). Some questions I may advise that you just need to figure it out as best you can.

- Please submit your assignment on Wattle. Late submissions will attract a penalty of 5% of your mark for each day of delay. No assignments will be accepted after the post date of 23 May. Even with approved extensions, late submissions will not necessarily be marked before you sit your final exam.
- Extensions need to be applied for using the Assessment Extension Platform (a link on the Wattle site) and documentation of exceptional circumstances is normally required.

**Question 1**

[25 Marks]

The NLS (National Longitudinal Study), sponsored by the U.S. Bureau of Labor Statistics, are nationally representative surveys that follow the same sample of individuals from specific birth cohorts over time. The surveys collect data on labor market activity, schooling, fertility, program participation, health, and much, much more. You will answer some questions by exploring the heights dataset extracted from NLS. You can access the heights dataset from the R package **modelr**. (You need to install and library the R package called **modelr**. Find the help file of the dataset **heights** and understand what each variable in the dataset stands for.)

(a) [4 marks] Import the dataset to R. What is the sample size? Which are numerical variables? Which are categorical variables? Rather than analyse the complete dataset, use a subset of the data. Draw a random sample of size 100 and work with the subset in the following questions. Use your ANU student ID (without the "u") in a call to the set.seed() function at the start of your R script. Hint: you can use the commands:

*set.seed(123)*

*heights2 <- heights[sample(250,100),]*

to create a subsetted data frame called heights2. You should use your student number in place of 123, and the actual sample size in place of 250.

(b) [4 marks] Define a new variable containing *body mass index* (BMI) for each individual. This will need to be calculated from other variables in the data. Also calculate a binary variable for Obesity, and another categorical variable with three levels *obese, overweight and normal* where the last category consists of everyone not in the obese or overweight categories. You need to research the definition of BMI and work out how to code it in R. You'll also need to research how "overweight" and "obese" are defined in terms of BMI. Include reference/s for the definitions of BMI, overweight and obese. References to websites are OK for this, but please find something that looks authoratitive. Don't include references to social media pages or pages behind a paywall or requiring an account to be created. A reference to an article in an academic journal or to a textbook would be ideal.

*"Overweight" and "obese" can be sensitive terms. BMI is a very imperfect measure and is not regarded as particularly reliable as an indicator of individual health or fitness. It is still widely used to track changes in population health and may be useful for this purpose (although still flawed). Likewise, the categories obese, overweight and "normal" are based on BMI and are not particularly meaningful at an individual level.*

(c) [3 marks] Create a plot that shows the distribution of BMI for men and for women and supports comparison of these distributions. Comment on the distributions and whether/how they differ.

(d) [4 marks] Create a plot that shows the association between income and the armed forces qualification test. Also quantitatively measure this association. Provide comments based on the plot and the quantitative measure.

(e) [5 marks] Define a person to have post-school education if their years of education are 13 or more. Do people's average armed forces qualification test results vary by whether they have had post school education? Conduct a hypothesis test for this question. Make sure you include all steps of a hypothesis test. Use a significance level 0.05. Check the conditions needed to perform the test.

(f) [5 marks] Suppose you now want to investigate the relationship between education levels and armed forces qualification rest results. Do people's average armed forces qualification test results differ between the three groups: school-only (12 or less years of education), some post-school (13, 14 or 15 years of education), and substantial post-school (16 or more years of education). Conduct a single hypothesis test for this research question. Use a significance level of 0.05 and check the conditions needed to perform the test.

**Question 2** [10 Marks]

Is the probability of Type I error equal to the significance level $\alpha$? You want to confirm this by a simulation study. The simulation is designed based on the two-sided hypothesis test:

$$H_0 : \mu = 0.5$$

$$H_A : \mu \neq 0.5$$

with known variance. Assume that your sample comes from a Uniform distribution with the range 0 to 1. Assume that the sample size is n=30.

Present your code and explain how you can show that $\alpha$ is the probability of Type I error with the result. Hint: In each simulation round, you should generate a sample from the uniform distribution, then conduct a hypothesis test assuming not knowing $\mu$ and knowing $\sigma$. Find the rejection region or p-value for the test and record whether you reject or fail to reject the null.

You will need to make use of the **runif** function in R (type help(runif) into the console for more information). The first line of your R script should include a set.seed(123) command, where your individual student number should be used in place of 123.

Ideally you will make use of **for** loops in R, which have been used in some of the lectures. However, if you are not confident with coding, you can just generate samples of uniform observations one at a time in R. For each sample, then conduct the hypothesis test (by hand or using R) and record your result. Then use your recorded results to estimate the Type 1 error. No marks will be deducted for this approach as long as you simulate at least 10 samples (and so at least 10 hypothesis tests). If you can manage it, make use of R for loops to do 1000 or more simulations – this will be less manual work for you and will obviously be a more meaningful assessment of type 1 error.