Statistical Methods in R: End-Of-Class Test (Practice)

TIME ALLOWED: 3 Hours

INSTRUCTIONS TO CANDIDATES

This test contains three problems. All three problems should be completed.

The test is fully open-book, open-notes, and open-web. However, you may not consult or communicate with anyone else at any point, with the sole exception that you may seek clarification from the invigilator. You may not use any chat or messaging software during the test. You may not use any AI services. **If you use communication software (e.g. e-mail, apps, Canvas notes) or access an AI service during the test, you will be asked to leave and you will be referred to the university for academic misconduct. If it is evident from your submission that AI tools were used to answer some questions, you will receive a zero and you will be referred to the university for academic misconduct.**

Solutions should be prepared only in R / Rstudio without the use of any other software. An RMarkdown template is available and use of this to complete the questions is required. At the end of the test, you should submit your code and a compiled PDF with your responses. The content of the code and the PDF should be consistent. Be sure to give yourself plenty of time at the end to make sure that your document knits correctly.

If the question requests a specific number or group of numbers, make sure your answer is unambiguous: either by making it the sole output of a block of R code, or (if the answer appears in a large summary-type output) by explicitly adding a one-sentence written remark or formatted data table below the code chunk identifying the answer or answers. Marks will not be awarded if multiple potential "answers" could be inferred from your output and the correct one is not clearly distinguished by a remark.

A blank R code chunk is provided for most questions; however, some questions may not require any R code. Verbal responses (in place of or in addition to a code chunk) should be written as text outside the code chunk, not as comments within the code chunk. A blank "setup" chunk is provided at the start of each problem for initial operations (e.g. loading the data file).

Comment out any extraneous outputs (e.g. exploratory plots) before compiling and submitting, and please keep additional explanations to a minimum. Do *not* comment out or hide code or calculations used to answer a question. If a question is answered incorrectly, partial credit will be awarded based only on what is performed in the code. Some questions are not eligible for partial credit.

The results of hypothesis tests should specify p and/or α, regardless of whether the question specifically asks for this number. Model parameter estimates should include confidence intervals (with the associated confidence level), unless otherwise stated. Calculations of descriptive statistics do *not* require standard errors, unless otherwise stated.

Mark values for each question are indicated in brackets, e.g. **[2]**. Each complete problem is worth a maximum of 20 marks. The maximum mark for the test is 60.

The page limit for this test is **16 pages** including this cover page. Please ensure that your output file is within the page limit before submitting and remove unnecessary outputs as needed. Do *not* delete any of the instructions or questions. Marks will be deducted for submissions in excess of the page limit.

The time limit for the test is 3 hours. A late penalty of 2 marks will be applied to submissions up to 12 minutes late. Submissions more than 12 minutes late will be penalized at 2 additional marks per additional minute late. Submissions more than 30 minutes late will not be accepted.

This is a practice test. The instructions above are provided to help you know what to expect during the real test. Instructions for the real test may differ: make sure to read them carefully when you take the real test.

Problem 1. Practical Probabilities

The lengths of members of a prized fish species are thought to be normally distributed. The distribution of lengths has a mean of 20cm with a standard deviation of 3cm.

1. What is the probability that a randomly-chosen fish of this species has a length between 20 cm and 27 cm? **[2]**
2. Two fish of this species are caught independently. What is the probability that both of them are shorter than 16cm? **[2]**
3. Forty percent of all fish of this species have lengths greater than 20cm but less than X cm. What is X? **[2]**
4. In a commercial catch of 10000 fish, how many (on average) do you expect to have a length greater than 30 cm? (You do not need to supply a confidence interval.) **[2]**

From historical records stretching back 1000 years, a volcano is known to have erupted on sixteen occasions. The eruption times are thought to be completely random and unpredictable, and the average frequency of eruptions is not changing with time.

5. What is the long-term average rate of eruptions, in units of eruptions per century (per 100 years)? Provide an estimated value and an "exact" confidence interval. **[4]**
6. Assume the estimated average rate you calculated in part 5 is correct. What is the probability that the next 100 years will pass without an eruption from this volcano? **[2]**
7. Again assuming the average rate from part 4, what is the probability that the volcano will erupt *two or more* times in the next *200* years? **[2]**

A university classroom has 17 students. Six of these students are frequent readers of mystery novels. According to a national poll, 10% of the population are frequent readers of mystery novels.

8. Can you rule out the hypothesis that students in the classroom are randomly drawn from the national population, as far as interest in mystery novels is concerned? Provide a p-value. (Use an exact method to calculate the p-value for full credit, or an approximate method for partial credit.) **[4]**

Problem 2: Classroom Demographics

The file classroom.csv contains data from a sample of students in a hypothetical classroom. Heights are in centimetres, weights are in kilograms, and ages are in years.

1. Determine whether male and female students (as a population) differ in average height. Provide a p-value. **[2]**
2. Determine whether male and female students (as a population) differ in average weight. Provide a p-value. **[2]**
3. Calculate the Pearson correlation coefficient between height and weight for male students. **[2]**
4. Calculate the Spearman correlation coefficient between height and weight for students (of either gender) under the age of 25. **[2]**
5. Assess whether male and female students **of similar height** differ significantly in average weight (as a population), ignoring all other variables. Provide a p-value. (You may assume that the relation is linear, and that weight increases with height in the same way for both genders.) **[6]**
6. Produce a scatter plot that illustrates your result from part 5, using labels, symbols, lines, and/or colours as appropriate. **[6]**

Problem 3: Website Redesign

A company that sells products by mail is experimenting with a redesign for its website. Visitors are randomly shown existing version A, or one of five proposed redeveloped versions B1, B2, B3, B4, or B5. Information is collected about how long they spend on the website in seconds), whether they buy anything (Y = purchase, N = no purchase), and if so how much they spend (in £). Data is provided in file website.csv.

1. Produce a count table to summarize how many visitors from each of the six groups made, or did not make, a purchase. **[3]**
2. Can you conclude that any of the proposed redesigns influence the probability that a visitor will make a purchase? (Be sure to include a p-value.) **[4]**
3. Estimate the probability, with associated confidence intervals, of making a purchase for each of the five test groups (and the control group). **[9]**
4. Produce a plot to communicate your results from question (3), including the confidence intervals, graphically. **[4]**