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: Probability

Questions should be solved using an exact calculation, not an approximation, and expressed to at least 5 significant figures.

Part A:

Suppose you flip a coin twelve times. Assuming the coin is fair, what is the probability that you obtain:

1. Twelve heads in a row? **[2]**
2. Three tails, and nine heads (in any order)? **[2]**
3. Exactly six heads and six tails (in any order)? **[2]**
4. Suppose now that you flip this coin 12 times, and the actual result you obtain is: two heads, and ten tails. Can you rule out the hypothesis that the coin is fair? Provide an an answer (yes/no) and a p-value. **[3]**

Part B:

Suppose that you have a spinner that you can spin. When it stops spinning, it comes to rest, pointing in some direction. You measure the position of the spinner as an angle in degrees (between 0 and 360).

5. If the spinner were completely fair (equally likely to land on any angle value between 0 and 360), what type of distribution describes the PDF of spinner angles? (Give the distribution name.) **[1]**
6. Suppose you spin the spinner five times. You measure angles of: 96.6, 193.9, 253.8, 219.3, and 235.5. You notice that all five measurements are between 96.6 and 253.8 (inclusive). Assuming that the spinner is fair, calculate the probability that five completely random spins would land between these two specific values. **[2]**
7. Can you rule out the hypothesis that the spinner is fair, given the probability value you have calculated? Explain your reasoning in no more than three sentences. **[2]**

Part C:

From the moment it hatches, a particular species of fish has a constant probability of surviving or perishing. From extensive previous measurements in the literature, you know that the mean lifetime for this fish species is 100 days.

8. For any randomly chosen individual living fish of this species, what is the probability that it will *not* survive the next day (next 24 hours)? **[2]**
9. What is the probability that a newly-hatched fish will live for more than a year (365 days)? **[2]**
10. Someone asks you to quote the lifetime of a fish of this species that survives longer than exactly 99% of all other members of its species. What is the lifetime of such a fish? **[2]**

Problem 2: Runner's Regimen

A group of 50 runners are enrolled in a study on the effects of taking a particular supplement on their fitness. During March, they are asked to run a 5-km run on 10 separate occassions and record their times, which were averaged together and the value recorded in the file runners.csv. The supplement, or a placebo, is taken daily during April. In May, they are again asked to run a 5-km run on 10 occasions and record their average times.

For parts 1-3, you may assume that run times are distributed normally within each group.

1. Did the times of runners in the placebo group become significantly shorter, or significantly longer, between March and May? Provide a p-value. **[3]**
2. Did the times of runners taking the supplement become significantly shorter, or significantly longer, between March and May? Provide a p-value. **[3]**
3. Can you conclude that the supplement was effective at improving run times, in comparison to the placebo? Provide a p-value. **[3]**

For parts 4-5, make no assumptions about the distribution of run times.

4. Calculate the sample interquartile range of run times for all runners in March. **[2]**
5. Use a percentile bootstrap (or another bootstrap) to provide a 95% confidence interval on the population interquartile range for all runners in March. Use at least 10000 bootstrap iterations. **[9]**

Problem 3: Tomato Farmer

A company is researching growing methods for tomato plants. Different watering frequencies, watering mechanisms, insecticides, and fertilizers are tried in various combinations, and the total tomato yield (in g) for each plant is measured. The company uses ten large greenhouses, labeled A-J. The results are saved in the file tomato.csv.

1. Use a multi-factor ANOVA to investigate the impact of growing method (and greenhouse) on yield. Start with a general model including all explanatory variables and (two-way) interactions between variable pairs, and simplify it by removing unneeded parameters. After simplification, write explicitly which variables are significant and which variable-pair interactions are significant. You may assume that errors are distributed normally. **[10]**
2. Considering them individually, which of the five potential explanatory variables (watering frequency, watering type, insecticide, fertilizer, or greenhouse) is the most important for determining the yield in this experiment? **[3]**
3. Recommend a growing and treatment regime (list the combination of factor levels which provide the highest yield, under your preferred model). **[3]**
4. Predict the average yield per plant under your recommended regime and provide a confidence interval on this model average. **[4]**