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, Distributed

You may use geometric reasoning for the questions of this problem. If you do, make sure to explain your reasoning in place of providing R code.

A distribution is defined by the following probability density function:

$PDF(x) = \frac{1}{2}e^{-|x|}$

1. Plot this function over the range $-10 < x < +10$, using an appropriate plot type. **[2]**
2. Determine the cumulative distribution function CDF(x) for this function, and plot CDF(x) over the range $-10 < x < +10$. **[4]**
3. Determine the quantile function QF(p) for this function, and plot QF(p) over the domain $0 < p < 1$. **[4]**
4. Determine the exact mean of this PDF. **[2]**
5. Determine the exact median of this PDF. **[2]**
6. Determine the exact skewness of this PDF. **[2]**
7. Determine the 99th percentile of this PDF. **[2]**
8. Is this a platykurtic, leptokurtic, or mesokurtic distribution? **[2]**

Problem 2: Pollution and Health

As part of an environmental health study, levels of a certain trace chemical pollutant are measured in blood samples collected from two groups of people: one group near an industrial site, and a control group in an isolated suburb. Each individual was measured twice (on separate occasions) to ensure consistency.

The data are in the file pollution.csv. Each row of the table corresponds to a single measurement. The reading number for each individual is denoted by 1 (first reading) or 2 (second reading). The concentration of the pollutant (in units of nanograms per mL) is provided, along with the unique ID number of each individual (sorted in ascending order).

You may assume that the distribution of pollution concentration is normal within each group.

1. Produce two histograms showing the distributions of measured pollutant concentrations: one for the control group, and one for the industrial site group. Align the histograms vertically (one on top of the other) within the same figure. Use the same bin breaks, and the same x- and y-axis ranges, for both histograms. **[3]**
2. Calculate the mean pollutant level for the control sample and for the industrial-site sample. Provide a 95% confidence interval on the population mean for each group, correcting for any pseudoreplication. **[5]**
3. Are the population standard deviations of the pollutant concentrations consistent with being the same for the control group and the industrial-site group? **[3]**
4. Compare the mean blood pollutant concentrations of the control and industrial site groups. Can you conclude that one group has a higher population mean than the other group? (If so, specify which group, and provide a p-value.) **[3]**
5. A pollutant concentration above 15 is generally considered to be unsafe. Estimate the proportion of all people in the industrial site population that, if their pollutant levels were to be measured, would have a level of 15 or greater. (You do not have to provide a confidence interval.) **[3]**
6. A follow-up study wishes to improve the estimate of the mean pollution levels of both groups. Specifically, their goal is to obtain a standard error of 0.3 on both measurements (the mean of the control group, and the mean of the industrial-site group). Estimate how many total individuals in each group they need to measure to achieve this goal. **[3]**

Problem 3: Oscillator Pattern Modeling

The file oscillate.csv contains measurements of an oscillating source. The displacement D (response variable) is thought to be a sinusoidal function of the measurement phase, w (explanatory variable). Mathematically, the underlying behaviour is expected to be governed by one of the following four equations:

Model 0: $D(w) = k$
Model 1: $D(w) = k + A_1 \sin(w) + B_1 \cos(w)$
Model 2: $D(w) = k + A_1 \sin(w) + B_1 \cos(w) + A_2 \sin(2w) + B_2 \cos(2w)$
Model 3: $D(w) = k + A_1 \sin(w) + B_1 \cos(w) + A_2 \sin(2w) + B_2 \cos(2w) + A_3 \sin(3w) + B_3 \cos(3w)$

(You may want to refer to the PDF, or Knit the Rmarkdown file, to render the LaTeX equations above.)

1. Is this a linear model or a nonlinear model, in regards to the connection between the response variable and the model parameters? **[2]**
2. Is this series of models nested or non-nested? **[2]**
3. Which of these four models, if any, can you rule out? Provide a p-value associated with the most complex model that you can rule out. **[10]**
4. Provide estimates of the parameter values, and the standard errors on those values, for all terms in the preferred model (i.e., the simplest model that is not ruled out). **[6]**