University of Strathclyde

Department of Mathematics and Statistics

MM959: Foundations Of Probability And Statistics

This project is worth 100% of the overall mark for MM959.

The project is released on Monday 11th March at 12pm (UK time). You must submit your answers by 4pm (UK time) on Tuesday 2nd April 2024 via Myplace.

PLEASE TURN OVER

Project Overview

This project is worth 100 % of the overall mark for the module. You are required to answer the questions below and report on your results. The assessment will involve some internet research, some probability calculations and some data analysis using the Minitab software package. You should produce and upload the following **three files**:

1. **Calculations Document**: a document showing your calculations/working and final answers for each question. Include Minitab output, calculations and interpretation as instructed for each question. For calculations you can type up or include a photograph of hand written work.
2. **Report**: a scientific report produced using a word processor. This report will relate to the questions indicated by * only. It will assess your ability to present findings in a clear and professional way, suitable for communicating to other professionals and clients in a consultancy type situation. The main body of your report should be no longer than 4 typed pages (excluding references), in 11 point font. Figures can be included in an appendix if necessary to keep the main report within 4 pages.
3. **Data File**: an Excel worksheet of your dataset (see Data Analysis Question 1).

Note: The calculations and report should be your own work. Marks will be deducted from students whose reports indicate collusion. All report submissions will be submitted to the plagiarism checker Turnitin to identify any similarity between student submissions and published literature including websites. The Calculations and Report documents should be submitted in PDF format.

Marks will be awarded as follows:

- Appropriate calculations for the data analysis and probability questions as outlined below. **All working and answers should be submitted in the Calculations document** [50 marks].
- Written report, including quality of written presentation: appropriate sections, background research, properly numbered, labelled and referenced figures and tables where appropriate. **Only questions marked by * should be included**. [30 marks].
- A suitable structure for the report should follow that of a scientific report, including the sections below:

- **Introduction**: Background information on pregnancy, maternal health and breastfeeding. Finish by summarising the aims of your work (think about the questions marked by *).
- **Methods**: Short section summarising the dataset, indicating the variables appropriate to the * questions and the probability and statistical inference methods that you have used for these questions
- **Results**: A written summary of the results that you have found. With the exception of figures, please do not include copy/pasted Minitab output in the results section - put into words or create your own tables. Explain your findings in words, making reference to appropriate figures and tables. All figures and tables should have appropriate captions.
- **Discussion and Conclusion**: Summarise your main findings generally, i.e. what are the interesting findings? What implications do they have for the health related issue? How do they compare to other findings in the literature? What are the strengths and limitations of your work? Recommend any future work that could be undertaken. End with an overall conclusion.
- **References**: List of appropriate references. All references included should be cross referenced in the text. Students should reference the data set and the software used.
- **Appendix**: Include figures here to ensure main text is 4 or less pages. If you can fit all figures in main text an appendix is not necessary.

Marks will be awarded for the Introduction (5 marks), Methods (8 marks), Results (8 marks), Discussion and Conclusion (7 marks) and References (2 marks).

Background

You are a researcher working at Public Health Scotland. You have been asked to do some research into the health of new mothers (questions are outlined fully in the data analysis and probability sections which follow).

The data come from a study of new mothers in the Glasgow City Council area of the NHS Greater Glasgow and Clyde Healthboard. Data were collected from new mothers at two time points: 35 weeks into their pregnancy and at the first visit with their health visitor (the health visitor appointment occurs after the baby is born).

The data were simulated, based on the patterns observed in the Births in Scotland (`https://www.opendata.nhs.scot/dataset/births-in-scottish-hospitals`) and Infant Feeding study (`https://www.opendata.nhs.scot/dataset/infant-feeding`). The dataset `maternity_data.csv` contains information on 6150 individuals.

For the purposes of this project you should assume these data are real. The data set contains the following variables:

- **Mother Age**: The mother's age at the time of the the birth: Under 20, 20-24, 25-29, 30-35, 35-39, 40 and Over, Unknown.
- **Feeding Type**: Way in which the mother is feeding her baby: Breast, Formula, Mixed, Other.
- **Tried Breastfeeding**: Whether or not the mother tried to breastfeed her baby: Yes (coded as 1), No (coded as 0).
- **Smoked**: Whether or not the mother smoked during pregnancy: Yes (coded as 1), No (coded as 0).
- **Type of Birth**: The type of birth the mother had: Breech, Caesarean-Elective, Caesarean-Emergency, Forceps, Vaginal, Vacuum, Not known.
- **Maternal BMI**: BMI classification of the mother at 35 weeks: Underweight, Healthy, Overweight, Obese, Unknown.
- **Pulse**: Pulse of the mother at 35 weeks (measured in bpm).
- **Bumpsize**: Size of the mother's bump at 35 weeks (measured in cm).
- **Iron level**: Iron level of the mother's blood at 35 weeks (measured in g/dL).
- **Time:** Length of time a mother is in hospital after giving birth (measured in days).

Data Analysis

The analyses should be conducted in Minitab, with appropriate output, interpretation, tables and figures included in the Calculations document, numbered by question.

1. Read in the full data set and generate your own set of data for analysis by using the following sequence of steps in Minitab.
   Set the base for the random numbers.
   Calc > Set base and enter your student number – this is a 9 digit number beginning the year you registered, so will be like 202312345. You can find your student number in PEGASUS if you do not know it.
   Calc > Random Data > Bernoulli and enter 6150 for the number of rows, C11 for the storage column and 0.6 for the probability.

Data > Subset worksheet and enter column C11 for the column to select on and pick 1 as the value.

File > Save worksheet as and navigate to the folder you are using and save your file as a .xlsx with a filename such as myprojectdata.xlsx. You should include your student number in the filename.

Upload your dataset to the Myplace page alongside your Calculations and Report PDF documents.

(2 marks)

2. Present a description of the following variables in the dataset using appropriate graphical and statistical summaries, briefly justifying your choice of graph and summary statistics for each variable: Iron Level and Type of Birth.

(6 marks)

3. * Determine whether or not there is any evidence of a significant difference in Bump Size for those who have a vaginal birth compared to those who have an emergency caesarean. Note: method and results for this question should be included in the report.

(6 marks)

4. What proportion of patients smoked whilst pregnant? Give the estimate and 95% confidence interval of this proportion. What do you conclude from the width of this interval?

(4 marks)

5. * Determine whether or not there is any association between Feeding Type and the Maternal BMI? Note: method and results for this question should be included in the report.

(8 marks)

PLEASE TURN OVER

Probability

You may choose to use Minitab to do the probability calculations or you can do them by hand using statistical tables. In either case, please outline the working that you would use if using the hand-worked approach in the Calculations Document.

1. One of the reasons for collecting this dataset is that it can be used to predict outcomes for mothers in the future. It is of interest to find out if breastfeeding is related to smoking during pregnancy.

   (a) Create a two-by-two table to show the observed counts of mothers who smoked during pregnancy and whether or not they had tried breastfeeding.

   (2 marks)

   (b) Use the table created in (a) to construct a probability tree to describe the outcomes when considering breastfeeding in relation to whether or not a mother has smoked.

   (4 marks)

   (c) For the probability tree constructed in question (b), list the elementary events in the sample space S and write down their associated probabilities.

   (3 marks)

   (d) Use Bayes' Theorem to calculate the probability that a new mother smoked during her pregnancy given that she tried breastfeeding.

   (3 marks)

   (e) In a sample of 40 new mothers that are known to have tried breastfeeding, use an appropriate probability distribution to calculate the probability that at least one of these mothers smoked during pregnancy. Interpret this probability.

   (2 marks)

2. * The variable `Time` contains data on the time a mother spent in hospital after giving birth. The time is recorded in days. Note: method and results for this question (all parts) should be included in the report.

   (a) Produce a histogram and descriptive statistics for the time spent in hospital after giving birth. Describe the empirical distribution.

   (3 marks)

   (b) Suggest a possible theoretical probability distribution which might be used to describe time spent in hospital after giving birth. Write down the probability distribution function (PDF or PMF) of this theoretical distribution. Your distribution should be described in terms of the numerical parameter values(s) derived from the summary statistics in part (a).

(2 marks)

(c) Fit a selection of plausible probability distributions to the `Time` variable and determine which, if any, provide a good fit to the data. Justify your choice of good-fit distributions using appropriate plots and statistical tests; specifically look at the theoretical distribution you suggested in part (b). Comment on which distribution you would select to model this variable.

(5 marks)

80 MARKS

END OF PAPER

(Louise Kelly and Connor Watret)