**Statistics for Health Data Science - Assessment 1**

This assessment is worth 20% of your final grade for the whole unit and 20 marks are available. It aims to test your understanding of univariable distributions, sampling distribution and hypothesis testing.

**Your report should be a maximum of 10 pages (font size at least 11, standard margins), including any R code and references that you include.**

* All solutions should be clearly set out with any hypotheses carefully set up and described.
* You are encouraged to provide your R code to document your work. However, statistical outputs should be interpreted and described, with some demonstration of translating back to the problem domain. Relation to previous literature (etc) is not required.
* The open-ended nature of the assignment means that it is possible to make some choices during the analysis – mark scheme will be flexible to reflect this.
* Please present as you would a report by ensuring the submission has a structure and you are not just providing screenshots (or copy and paste) of code and results.

You are working for a local organisation that has undertaken a diabetes screening programme. Individuals were invited to attend clinics undertaking the programme over a month period and you have been given the data collected for each individual visit. You are asked to produce a report summarising the data collected and to complete 2 tasks;

1. Please provide a summary of the data, with the aid of graphs and tables (5 marks)
2. Another screening programme took place last year in a different region and they found 14% of individuals had diabetes and the average BMI and age of those attending clinics were 28kg/m2 and 55 years, respectively. They also found a significant relationship between smoking and diabetes. Investigate if your data support these results and provide possible explanations to your findings. (15 marks)

You have been provided with a dataset (SHDS\_Assessment1\_data.csv) from the screening programme that contains baseline information of patients collected during the screening visit and each patient’s diabetes diagnosis.

**Dataset information**:

The dataset contains the following variables, as described here.

|  |  |
| --- | --- |
| Diabetes\_012 | Weather the patient was diagnosed with diabetes (0 = no-diabetes, 1 = diabetes) |
| HighChol | Whether the patient has high cholesterol (0 = no, 1 = yes) |
| CholCheck | Whether the patient had cholesterol checked (0 = no, 1 = yes) |
| BMI | Body mass index (kg/m2) |
| Smoker | Whether the patient is a smoker (0 = non-smoker, 1 = smoker) |
| Stroke | Whether the patient has ever had a stroke (0 = no, 1 = yes) |
| Sex | Sex of the patient, binary (0 = female, 1 = male) |
| Age | Age of patient in years |
| MentHlth | Number of days in the past 30 days the patient says they have suffered with mental health (including stress, depression, etc) |

**Additional Hints:**

* Don’t forget to conduct an exploratory/descriptive analysis.
* Please showcase what you have learnt, but DO NOT simply perform every test you can. Carefully think about the assumptions and which tests are appropriate to perform when.
* When analysing the data, you should carefully think about any hypothesis and assumptions
* You should structure your report so that it is easy to follow what you have done, and the results of your analysis.
* Note: the dataset and scenario are entirely fictitious, and so you are not expected to provide background motivation/reading within your report, BUT you should consider the results from the fictitious scenarios presented.

**Mark scheme information**:

Approximately half of the marks are allocated for correct and appropriate analysis, with the other half being allocated for clear descriptions and justifications of the analyses, their interpretation, discussion of the implications, and presentation. It is particularly important that you clearly explain what it is you are doing, your assumptions, and conclusions. Simply presenting code and output with no discussion and explanation is inadequate and will result in a fail.

The below table gives an indication of how the 50 marks available will be awarded:

|  |  |
| --- | --- |
| **Mark (% of the marks available for question 1)** | **Descriptor** |
| 0-19% | Analysis makes little or no attempt to address the brief, is generally inaccurate and is hard to follow. Conclusions are either nonexistent or incorrect. |
| 20-29% | Analysis addresses the brief in a limited way but is generally inaccurate. Report is generally incoherent. Conclusions often incorrect, and no attempt is made to discuss assumptions of analysis. |
| 30-39% | Analysis addresses only parts of the brief and is often inaccurate. Report is generally incoherent. Conclusions often incorrect, and no attempt is made to discuss assumptions of analysis. |
| 40-49% | A partly correct analysis that has addressed at least some of the brief. Report lacks coherence on the whole. Conclusions are sometimes incorrect, and little or no attempt is made to discuss assumptions of analyses. |
| 50-59% | A mostly correct analysis that has addressed the brief. Report is coherent but may be hard to follow in places. Conclusions and interpretation generally correct but superficial. Some limited effort is made to discuss assumptions of the analysis. |
| 60-69% | Analyses are correct besides minor errors and typos. Report is coherent throughout. Conclusions and interpretation are correct but lack depth. Reasonable discussion of assumptions and limitations of analysis. |
| 70-79% | Analyses are correct, besides minor errors and typos, and address the brief in considerable detail. Conclusions and interpretation are correct and detailed. Comprehensive discussion of assumptions and limitations. |
| 80-89% | Analyses are correct, address and often go beyond the brief, demonstrating independent thought and judgement. Conclusions and interpretation are correct and detailed. Comprehensive discussion of assumptions and limitations. |
| 90-100% | Marks in this category reflect a report of publishable quality with advanced insights and going well beyond the brief. Reports may demonstrate novel and innovative approaches to address the brief, and interpretations and conclusions will demonstrate a deep understanding and judgement. |