**Fundamentals of mathematics and statistics - Assessment 1**

This assessment is worth 50% of your final grade for the whole unit. It aims to test your understanding of univariable distributions, sampling distribution and hypothesis testing.

* 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.

You have been provided with a subset of the data from the Second Manifestations of ARTerial disease (SMART) study. Here is a [link](https://link.springer.com/article/10.1023/A:1007621514757) to more information about the study.

The columns provide data on patient ID numbers (‘ID’), sex (‘SEX’), age (‘AGE’), age group (‘AGE\_GRP’), diabetes (‘DIABETES’ (0=non-diabetic, 1=diabetic)), body mass index (‘BMIO’), smoking status (‘SMOKING’), systolic blood pressure (‘SYSTBP’) and whether they experienced a cardiovascular event during follow-up (‘EVENT’ (1=yes, 0=no)).

1. **Using the SMART (‘FMS\_assessment1.csv’’) data**. **(11 marks)**
2. Produce appropriate initial summaries of sex and smoking. (2 marks)
3. Remove missing data (coded as NA) from systolic blood pressure and produce appropriate initial summaries of this variable. Discuss whether there is evidence that systolic blood pressure follows a normal distribution. (6 marks)
4. Calculate the 95% confidence interval for the true population mean blood pressure (using the sample mean, standard deviation, and the sample size *n*). The average blood pressure in the population is thought to be 145. Does the confidence interval contain 145? How would you interpret this? (3 marks)
5. **There is a standard therapy used to treat breast cancer, with the cure rate 0.3 (i.e., 30% of patients with breast cancer are cured by the therapy). A new therapy has been developed by a pharmaceutical company. To test if it has a higher cure rate than the standard one, the company will allocate the new therapy to 1000 patients who are randomly drawn from those with breast cancer. The company will conclude that the new therapy is superior if at least 330 of the 1000 patients will cure from breast cancer.** **(14 marks)**
6. Suppose *X* is the number of patients cured from the new therapy. What is the distribution of *X*? Please specify the null and alternative hypotheses. (4 marks)
7. Use the central limit theorem to calculate the probability of type I error. Please interpret the result. (4 marks)
8. If the true cure rate of the new treatment is 0.35, what is the power of the test? Please interpret the result. (2 marks)
9. If the new treatment has resulted in 400 patients cured from breast cancer. What is the *p*- value of the test? What conclusion can you draw from the *p*-value? (4 marks)
10. **Using the SMART (‘FMS\_assessment1.csv’’) data again.** **(11 marks)**
11. Test if the true population proportion of male patients is 0.5. (4 marks)
12. Calculate odds ratio of association between sex and cardiovascular event (EVENT). How would you interpret it? Is there statistical evidence of an association between sex and cardiovascular event? (7 marks)
13. **Using the SMART (‘FMS\_assessment1.csv’’) data again, perform tests to investigate the following.** **(14 marks)**
14. Is there statistical evidence of a difference in BMI between men and women? (6 marks)
15. Is there statistical evidence of a difference in BMI between age groups? (8 marks)

**Total marks available: 50**