EMAT30007 Applied Statistics 2022/23

Coursework

Nikolai Bode

(nikolai.bode@bristol.ac.uk)

Due: 26th April 2023, 01:00pm (week 23)

Submit on BLACKBOARD

General information:

Attempt to answer all questions. This coursework has three parts with one question each. The questions are released successively as the material is covered in the course.

Submit a Matlab script (.m file) file with your answers. This should run when copied into the same folder as the data files (see below) and should only use commands and Matlab packages used in the worksheets. Clearly annotate your code and include the required discussion of your findings directly in the script.

The limit for your submission is 900 lines of standard Matlab script with at most 100 characters per line, in addition to a restriction on the number of figures/plots for each question (see below). Just for indication: the model solution is 600 lines long with a lot of empty lines.

There are additional files available on Blackboard for this piece of coursework that contain data to be used in the questions. The contents of the files are described in more detail below.

The only way I will answer questions about the coursework is via the dedicated coursework discussion forum on Blackboard. This is to ensure that the entire class has access to the same information.

There are a total of 60 marks for the coursework, 20 marks for each part of the coursework.

Question 1 (20 marks):

A research team in a pharmaceutical company wants to test the efficacy of a new drug. The drug is very expensive and preliminary tests have shown that due to an unknown physiological process, the active substance of the drug is not delivered to the bloodstream of all people. Fortunately, there is a separate, much cheaper screening test to determine with high accuracy whether the active substance of the drug will appear in the blood stream of individuals. The screening test produces a score, and higher scores correlate with a higher chance of the active substance appearing in the blood stream. The research team devise the following study. First, 100 participants are recruited through adverts and screened using the cheap test. Second, the 40 participants with the highest scores from the screening test are randomly divided into two groups of 20 participants each. One randomly chosen group receives a placebo, the other the real drug. The researchers and participants do not know during the trial who gets the placebo and who gets the drug. The trial concludes with the measurement of an appropriate physiological variable in the 40 participants. The research team hopes that the drug will increase the value of this variable and conduct an appropriate statistical test to assess this. They find that in the group receiving the drug, the variable is 20% higher on average, and they claim, based on their statistical test, that this finding is statistically significant. They write a report in which they recommend the introduction of the drug to market.

Due to a very expensive administrative error, the drug is actually also administered to half of the 60 participants not included in the second part of the study, with the other 30 participants receiving a placebo. This data is found years later by a student doing an internship at the company.

The files `drugs20.txt` and `placebo20.txt` contain the final measurements for the groups of 20 participants who received the drug or the placebo, respectively. The files `drugs30.txt` and `placebo30.txt` contain the data the intern finds.

(a) Conduct the statistical test the scientists performed and conduct the same test on the data the intern has found. State what you find.

(b) Extending the material covered in weeks 13-16, devise and briefly justify a method to approximate a 95% confidence interval for the percentage difference in the variable measured between the drug and the placebo group under the assumption that instead of the initial screening, 40 participants were selected at random from the initial participant pool for the second part of the study. State what you find and discuss why this can only be an approximation.

(c) Fit an appropriate probability distribution to the data for all 100 participants using maximum likelihood estimation. Note that a standard probability distribution may not be appropriate, so you may have to define your own. Show qualitatively that the fit of the probability density function of your chosen distribution to the data is good (1 figure).

(d) Write a statistical critique of the study described above, referring to your answers in questions 1(a)-(c). Provide your critique in bullet points, not continuous text. You should discuss positives, negatives, and possible improvements for the study.

_____