Mock Exam

Applied Statistics 2022, IT University of Copenhagen

11th May 2022

This was the real exam in Applied Statistics in the Spring 2021.

Instructions

This mock exam is divided into five parts and each part have sub-questions. A number of points (pts) are assigned to each question and the total number of points in this exam is 32. If this was a real exam, you would have to hand in both a PDF document (`.pdf`) generated using R Markdown and the corresponding R Markdown code file (`.Rmd`). To simulate a real exam situation, please hand in both files in through LearnIT. When answering the questions, you must explain how you obtained the answers and use correct mathematical notation.

1. Probability Theory (6 pts)

Consider an experiment where you flip a fair coin four times.

(a) (1 pt) Define a natural sample space Ω for this experiment.

(b) (2 pts) Write down the set of outcomes corresponding to each of the following events

- A: "We get exactly one tails"
- B: "The coin always comes with the same side up."

(c) (1 pt) Summarise in words the meaning of the event $A \cup B$.

(d) (2 pts) Compute the probability for the event $C = (A \cup B)^c$.

2. Expectation, Variance, Discrete Distributions (6 pts)

Let us consider the experiment of independently throwing two fair dice characterised by the discrete random variables X and Y, respectively. The discrete random variables X and Y take the values $a = 1, 2, 3, \ldots, 6$ and $b = 1, 2, 3, \ldots, 6$, respectively.

(a) Compute the expected value for the product $Z = XY$. (2pts)

(b) Write down the probability mass function for the the discrete random variable Z defined by the product $Z = XY$. (2pts)

(c) Compute the variance of the product $Z = XY$. (2pts)

3. Maximum likelihood (4pts)

Let x_1, x_2, ..., x_n be a dataset that is a realisation of a random sample from a $U(\alpha, \beta)$ distribution, where α and β are the unknown parameters.

(a) Write down the likelihood function of the parameters. (2pts)

(b) Determine the maximum likelihood estimates for the parameters α and β. (2pts)

4. Small R Problems (8 pts)

(a) The data set `firstchi (UsingR)` contains the age of the mother at birth of the first child. Investigate the data set by computing several simple statistics on this data set. Summarise your findings. (2 pts)

(b) The data set `rat (Using R)` contains the survival times of 20 rats exposed to radiation. Visualise the data in appropriate means and discuss in the light of the data and your knowledge, what kind of parametric model would you choose for the distribution of the survival times. (2 pts)

(c) Consider the dataset `kid.weights (UsingR)`, that reports information about a sample of 250 kids. Select the kids up until 9 years old (i.e. with an age strictly less than 108 months). Plot a scatter plot of the weight versus the height. Compute a linear regression model and add the regression line to the plot. What conclusions can you derive? (2 pts)

(d) Assume that you have implemented a scientific method that you compare to the state-of-the-art published elsewhere. You use the evaluation metric G where a bigger value refers to a better outcome. Using independent experiments you get five scores for the state-of-the-art (0.908,0.915,0.908,0.905,0.904) and five for your method (0.910,0.914,0.909,0.914,0.910). State the null and alternative hypothesis and test if there is statistical evidence that your method is better than the state-of-the-art. (2 pts)

5. Bootstrapping and confidence intervals (8 pts)

Let us look at the datasets `female.inc (UsingR)` that contain income distribution for females in 2001. You may ignore the information about the race of the individuals. Your goal is to estimate the mean female income together with its 95% confidence intervals.

(a) Why is bootstrapping a good strategy for finding the confidence intervals for the mean female income? (1pt)

(b) Write a computer program that computes the bootstrap estimates for the 95% confidence intervals for the mean income of all the females in the dataset. (5pts)

(c) Report your numerical results for the mean and confidence interval together with a graphic illustrating the bootstrapped sample. (2pts)