X_400004 - Statistics
Mock Midterm Exam

14 October 2020

I recommend that you try to solve this mock Midterm exam only after you have prepared for the Midterm and that you time yourself when solving it. The instructions for the Midterm Exam follow.

Instructions:

- The exam is to be solved **individually**.
- Please **write clearly and in an organised way**: we can't grade illegible answers.
- Please change pages when starting a new question.
- This is an exam on a mathematical subject, so support your answers with **computations**, rather than words, whenever possible.
- You should report **all relevant computations** and **justify** non-trivial steps.

- This is a **closed book exam**; you are only allowed to have one A4 sheet with **handwritten** notes with you.
- You may use a calculator.

- There are 5 pages in the exam questionnaire (including this one) and you have two hours (120 minutes) to complete the exam.
- The exam consists of 10 questions spread throughout 3 problems.
- The number of points per question is indicated next to it for a total of 100 points.

- The problems are not necessarily ordered in term of difficulty. I recommend that you quickly read through all problems first, then do the problems in whatever order suits you best.
- Remember to **identify** at least one of your answer sheets with you name and student number.

At the actual Midterm, after completing your exam, you will have to digitalise your answer sheets (with the correct order and orientation) and submit them as **a single PDF** on Canvas for grading. **You will have 10 minutes following the end of the exam to do this** after which any submission will be marked as late. These instructions do not replace the VU's *Protocol of online examination for 2020-2021* that you can find on Canvas.

Prob.I: Suppose that you come across a dataset whose distribution is well modelled by the following probability density function

$$f(x) = \begin{cases} \frac{2\left(1-\frac{x}{\sqrt{\lambda}}\right)}{\sqrt{\lambda}} & \text{if } 0 \le x \le \sqrt{\lambda} \\ 0 & \text{otherwise} \end{cases},$$

where $\lambda > 0$ is a parameter that we would like to estimate. Suppose that you know that if X is a random variable with the density above, then

$$\mathbb{E}(X) = \frac{\sqrt{\lambda}}{3}, \qquad \mathbb{E}(X^2) = \frac{\lambda}{6}, \qquad \mathbb{E}(X^3) = \frac{\lambda\sqrt{\lambda}}{10}, \qquad \mathbb{E}(X^4) = \frac{\lambda^2}{15}.$$

Suppose that you observe two independent samples X_1 and X_2 from the above distribution and consider the following two possible estimators of λ:

$$\hat{\lambda} = 3(X_1^2 + X_2^2) \qquad \text{and} \qquad \tilde{\lambda} = X_1^2 + X_2^2 + 2X_1X_2.$$

(a) Compute the bias of the two estimators. Is any of the estimators unbiased?

(b) Compute the MSE of estimator $\hat{\lambda}$.

(c) It can be shown that the MSE of $\tilde{\lambda}$ is given by $\frac{41}{90}\lambda^2$. Given this and your answers to the previous questions which of the two estimators would you prefer? Justify your answer.

Prob.II: Let X be the amount of time that your toaster takes to toast a slice of bread since the moment you press the slider on the toaster. This random variable X has two components, an exponential amount of time that you have to wait until the heating element is warm, plus a deterministic amount of time θ for the toasting process to be completed. More specifically, $X = Y + \theta$, where $Y \sim Exp(1)$, and $\theta > 0$ is some unknown parameter. The probability density function f of X is then

$$f(x) = \begin{cases} e^{-(x-\theta)} & \text{, if } x \geq \theta, \\ 0 & \text{, if } x < \theta. \end{cases}$$

(This is the density of a so called *shifted exponential distribution.*)

Consider a random sample $X_1, \ldots, X_n$ of measurements distributed like X.

(a) Write down the likelihood of the data. Can you get the Maximum Likelihood estimator (MLE) of θ by differentiating the log-likelihood of the data? Justify your answer.

(b) Instead of pursuing the MLE, say that we instead go for the Method of Moments estimator (MME). Start by computing the expectation and the variance of X.

(c) Based on the expectation, compute the MME for θ.

(d) Is the MME biased? What is the mean squared error (MSE) of the MME?

Prob.III: Computer security companies must always be on the lookout for new threats. Most often than not, security breeches are unexpected. For instance, in a *timing attack* the attacker attempts to compromise a cryptosystem by analysing the time taken to execute cryptographic algorithms. Every logical operation in a computer takes time to execute, and the time can differ based on the input; with precise measurements of the time for each operation, an attacker can work backwards to the input. This information can provide the attacker with information about the CPU running the system, the type of algorithm used, etc...

To check if a certain system is secure 40 login attempts were conducted with randomly chosen passwords of diverse lengths. **The amount of time in milliseconds (ms) that the system took to deny access was recorded and can be found at the end of the questionnaire, together with a collection of descriptive statistics and various graphical representations of the data.**

(a) Determine the sample mean, sample variance, sample standard deviation, and range of the dataset. (Don't forget to report the units.)

(b) Briefly explain how **each of the plots** below supports/contradicts the possibility that the data comes from a Normal distribution.

(c) Assume that the data comes from a Normal distribution. Construct a two-sided 90% confidence interval for the variance of the system's response time and compute its realisation from the data at hand. (**This means that you need to derive the expression for the interval from an appropriate pivot, not just write down the interval**.) In light of your answer to (b), is it sensible to compute such an interval in this case? Justify your answer. (You can find some quantiles that you may need to answer this question at the end of the questionnaire.)

```
Data (time in ms):
1.777763, 2.223587, 6.869387, 10.78203, 2.332443, 4.312676, 6.440998, 2.023269,
5.531647, 3.481276, 2.965045, 2.84759, 7.710503, 3.821837, 9.726404, 2.507011,
0.5685252, 3.123111, 0.8075089, 4.79541, 4.850344, 2.119353, 8.628205, 7.345244,
3.226172, 6.608948, 2.298478, 3.180011, 5.660042, 3.385601, 5.243938, 9.7317,
3.716626, 5.093568, 2.026606, 4.956999, 8.205676, 5.941724, 3.704429, 6.257464
```

$$n = 40, \qquad \sum_{i=1}^{40} X_i = 186.829149, \qquad \sum_{i=1}^{40} X_i^2 = 1125.008716$$

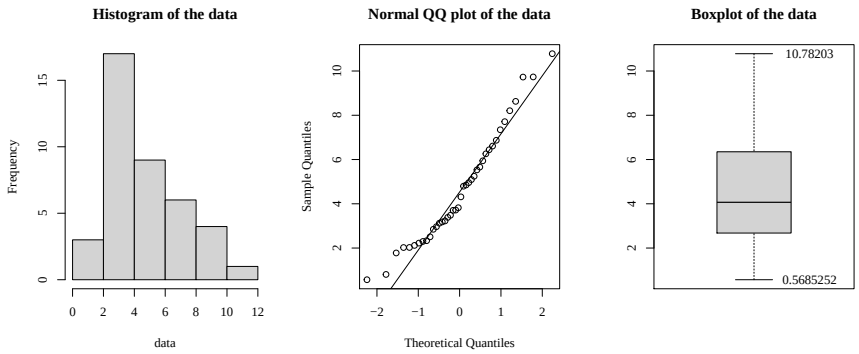

Figure 1: Some summary plots for the dataset.

Some quantiles from the Normal distribution:

$$z_{0.01} = -2.33, z_{0.025} = -1.96, z_{0.05} = -1.64, z_{0.095} = 1.64, z_{0.0975} = 1.96, z_{0.099} = 2.33.$$

Some quantiles from the t_{39} distribution:

$$t_{39;0.01} = -2.43, t_{39;0.025} = -2.02, t_{39;0.05} = -1.68, t_{39;0.095} = 1.68, t_{39;0.0975} = 2.02, t_{39;0.099} = 2.43.$$

Some quantiles from the t_{40} distribution:

$$t_{40;0.01} = -2.42, t_{40;0.025} = -2.02, t_{40;0.05} = -1.68, t_{40;0.095} = 1.68, t_{40;0.0975} = 2.02, t_{40;0.099} = 2.42.$$

Some quantiles from the χ^2_{39} distribution:

$$x^2_{39;0.01} = 21.43, x^2_{39;0.025} = 23.65, x^2_{39;0.05} = 25.70, x^2_{39;0.095} = 54.57, x^2_{39;0.0975} = 58.12, x^2_{39;0.099} = 62.43.$$

Some quantiles from the χ^2_{40} distribution:

$$x^2_{40;0.01} = 22.16, x^2_{40;0.025} = 24.43, x^2_{40;0.05} = 26.51, x^2_{40;0.095} = 55.76, x^2_{40;0.0975} = 59.34, x^2_{40;0.099} = 63.70.$$