X_400004 - Statistics
Midterm Exam

23 October 2020

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 6 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.
- Your final grade is $\max(1, \text{score}/10)$, where "score" is the number you points you get.

- 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.

After completing your exam, digitalise your answer sheets (with the correct order and orientation) and submit them as **a single PDF** on Canvas for grading. **You 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: When modelling the amount of time spent in baggage handling procedures at an airport, you come across the following statistical problem. Let $X_1, \ldots, X_n$ be a random sample distributed like X which has a probability density function given by

$$f(x) = \begin{cases} \frac{x}{\lambda} e^{-x/\sqrt{\lambda}} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases},$$

where $\lambda > 0$ is an unknown parameter. Note that (you can just take these as facts)

$$\mathbb{E}[X] = 2\sqrt{\lambda}, \qquad \mathbb{E}[X^2] = 6\lambda, \qquad \mathbb{E}[X^3] = 24\lambda^{3/2}, \qquad \mathbb{E}[X^4] = 120\lambda^2.$$

Consider two different estimators for λ, namely

$$\hat{\lambda} = \frac{(\bar{X})^2}{4}, \qquad \text{and} \qquad \tilde{\lambda} = \frac{1}{6n} \sum_{i=1}^{n} X_i^2.$$

15 pts (a) Compute the bias of $\hat{\lambda}$ and of $\tilde{\lambda}$. Is any of the two estimators unbiased? What happens to the bias as n increases?

10 pts (b) Compute the mean squared error (MSE) of $\tilde{\lambda}$.

7 pts (c) It can be shown that $\mathbb{V}(\hat{\lambda}) = 2\lambda^2/n$. Use this fact to compute the MSE of $\hat{\lambda}$. Assuming $n = 100$ which estimator is better in terms of MSE?

Prob.II: Consider a random sample $X_1, \ldots, X_n$ from a gamma distribution with parameters $\alpha > 0$, and $\beta > 0$, which are unknown to you. The probability density function of each observation is

$$f(x) = \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)}, \quad x \geq 0,$$

and $f(x) = 0$ if $x < 0$. The function Γ is the gamma function. What is relevant for you here is that if X is distributed like f, then

$$\mathbb{E}X = \frac{\alpha}{\beta}, \qquad \text{and} \qquad \mathbb{V}X = \frac{\alpha}{\beta^2}.$$

10 pts (a) Write down the system of equations that you would have to solve to compute the Maximum Likelihood estimator (MLE) for α and β. (**You don't need to solve the system and you don't need to compute** $\Gamma'(\alpha)$.)

10 pts (b) The system from (a) looks complicated, so instead we compute the Method of Moments estimator (MME). Compute the MME of α and β based on the first two moment of X.

5 pts (c) Assume now that $\beta = \alpha$. Is it possible to get an MME of α based on the first moment? Justify your answer.

6 pts (d) Still for the case when $\beta = \alpha$, derive the MME of α based on the second moment.

Prob.III: Digital security and bot detection is nowadays an important concern for businesses and companies, and monitoring keystroke dynamics gives a way to prevent improper access to a computer. By checking if the typing behaviour of the current user is "compatible" with that of the legitimate user one can potentially detect an attacker and react accordingly.

We asked the legitimate user of a certain system to type a short text (with 235 characters), while the time *between* consecutive keystrokes was recorded. **The corresponding data of the 234 inter-keystroke times (in seconds) can be found at the end of the questionnaire, together with a collection of descriptive statistics, various graphical representations of the data, as well as quantiles for different distributions.**

Although not entirely plausible, assume that the data is the realisation of a random sample from some unknown distribution.

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

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

18 pts (c) Construct an approximate, two-sided, 90% confidence interval for the expectation of the inter-keystroke 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 quantiles that you may need in this question at the end of the questionnaire.)

Data (time in seconds):
0.1746, 0.4319, 0.2015, 0.4939, 0.2023, 0.2065, 0.1692, 0.1526, 0.1651, 0.1133,
0.1307, 0.1697, 0.1915, 0.1128, 0.3344, 0.2132, 0.2841, 0.2166, 0.1503, 0.1612,
0.0973, 0.1221, 0.3645, 0.0176, 0.0954, 0.1681, 0.1816, 0.1176, 0.2608, 0.177,
0.2935, 0.4291, 0.2228, 0.1249, 0.4702, 0.1368, 0.1205, 0.1228, 0.203, 0.3926,
0.2743, 0.112, 0.1623, 0.1488, 0.2185, 0.136, 0.2828, 0.1177, 0.0961, 0.4464,
0.2117, 0.2364, 0.2203, 0.5848, 0.1606, 0.1754, 0.3835, 0.1097, 0.4526, 0.0857,
0.1024, 0.2436, 0.1932, 0.4019, 0.0468, 0.12, 0.2701, 0.1181, 0.0213, 0.2189,
0.1934, 0.1529, 0.1264, 0.0947, 0.2867, 0.1699, 0.2312, 0.1651, 0.1615, 0.0924,
0.1276, 0.272, 0.1693, 0.2693, 0.2044, 0.1903, 0.3496, 0.1291, 0.1055, 0.5198,
0.2675, 0.2789, 0.124, 0.3314, 0.0809, 0.1115, 0.1012, 0.1225, 0.1403, 0.2691,
0.1466, 0.0781, 0.1562, 0.1477, 0.1162, 0.2018, 0.1644, 0.1852, 0.2912, 0.2202,
0.2528, 0.1537, 0.1127, 0.1973, 0.0968, 0.1232, 0.1603, 0.1013, 0.1945, 0.2184,
0.15, 0.1201, 0.179, 0.1497, 0.0714, 0.3035, 0.0355, 0.1144, 0.2885, 0.1711,
0.2089, 0.1969, 0.1926, 0.32, 0.2484, 0.2346, 0.0477, 0.2503, 0.1885, 0.1429,
0.2544, 0.1772, 0.1831, 0.1497, 0.1693, 0.1289, 0.1731, 0.1595, 0.2684, 0.1315,
0.1445, 0.3315, 0.1895, 0.18, 0.1254, 0.1393, 0.1574, 0.1938, 0.3278, 0.0729,
0.578, 0.1005, 0.1975, 0.206, 0.168, 0.133, 0.0808, 0.082, 0.3079, 0.1512, 0.4028,
0.1466, 0.2502, 0.1869, 0.2485, 0.1, 0.128, 0.1343, 0.1428, 0.2834, 0.1803, 0.1558,
0.1723, 0.116, 0.1859, 0.2682, 0.2796, 0.0173, 0.1367, 0.2287, 0.1406, 0.1643,
0.1997, 0.2752, 0.0013, 0.3155, 0.0389, 0.2236, 0.0837, 0.2119, 0.0915, 0.2715,
0.2208, 0.192, 0.1946, 0.2639, 0.1211, 0.1812, 0.1874, 0.2448, 0.1898, 0.5933,
0.119, 0.1937, 0.1153, 0.2348, 0.2047, 0.144, 0.332, 0.0683, 0.1022, 0.1438,
0.2002, 0.407, 0.1228, 0.5608, 0.2065, 0.2541, 0.067, 0.3595, 0.1416, 0.1851,
0.3603, 0.7438

$$n = 234, \qquad \sum_{i=1}^{234} X_i = 46.616, \qquad \sum_{i=1}^{234} X_i^2 = 12.098$$

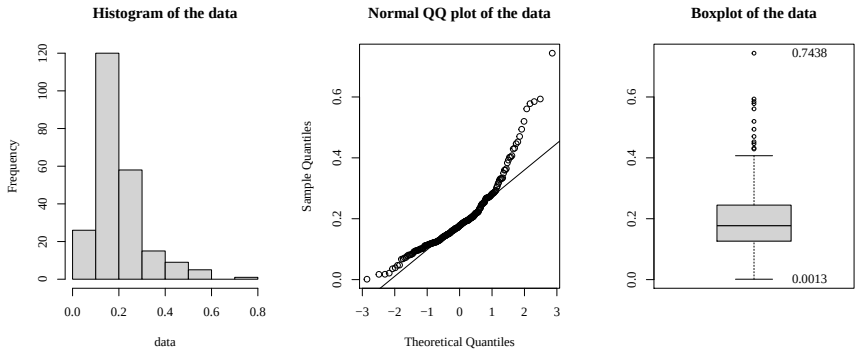

Figure 1: Some summary plots for the keystroke dataset.

Quantiles from different distributions follow below.

Some quantiles from the Normal distribution:

$$z_{0.01} = -2.33, z_{0.025} = -1.96, z_{0.05} = -1.64, z_{0.95} = 1.64, z_{0.975} = 1.96, z_{0.99} = 2.33.$$

Some quantiles from the t_{233} distribution:

$$t_{233;0.01} = -2.34, t_{233;0.025} = -1.97, t_{233;0.05} = -1.65, t_{233;0.95} = 1.65, t_{233;0.975} = 1.97, t_{233;0.99} = 2.34.$$

Some quantiles from the t_{234} distribution:

$$t_{234;0.01} = -2.34, t_{234;0.025} = -1.97, t_{234;0.05} = -1.65, t_{234;0.95} = 1.65, t_{234;0.975} = 1.97, t_{234;0.99} = 2.34.$$

Some quantiles from the χ^2_{233} distribution:

$$x^2_{233;0.01} = 185.74, x^2_{233;0.025} = 192.62, x^2_{233;0.05} = 198.67, x^2_{233;0.95} = 269.61, x^2_{233;0.975} = 277.17, x^2_{233;0.99} = 286.14.$$

Some quantiles from the χ^2_{234} distribution:

$$x^2_{234;0.01} = 186.63, x^2_{234;0.025} = 193.52, x^2_{234;0.05} = 199.59, x^2_{234;0.95} = 270.68, x^2_{234;0.975} = 278.26, x^2_{234;0.99} = 287.25.$$