X_400004 - Statistics
Solutions to the Midterm Exam

23 October 2020

Below are answers to the exam questions. Some of these are slightly abbreviated, while others include extra comments. These are for your reference only but should inform the level of detail that is expected from your answers in the exam. Also keep in mind that there might be different ways to approach each question. If you find typos and or omissions, please report them to the lecturer so they can be corrected.

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?

Solution: Let us start with $\hat{\lambda}$. Recall the properties of the sample mean, namely $\mathbb{E}[\bar{X}] = \mathbb{E}[X_i]$ and $\mathbb{V}(\bar{X}) = \mathbb{V}(X_i)/n$. Therefore

$$\mathbb{E}\hat{\lambda} = \mathbb{E}\left[\frac{(\bar{X})^2}{4}\right] = \frac{1}{4}\mathbb{E}\left[\bar{X}^2\right] = \frac{\mathbb{V}(\bar{X}) + (\mathbb{E}[\bar{X}])^2}{4} = \frac{\mathbb{V}(X_i)/n + (\mathbb{E}[X_i])^2}{4}.$$

Now note that $\mathbb{V}(X_i) = \mathbb{E}[X_i^2] - (\mathbb{E}[X_i])^2 = 2\lambda$ and so

$$\mathbb{E}\hat{\lambda} = \frac{2\lambda/n + 4\lambda}{4} = \lambda + \frac{\lambda}{2n}.$$

In conclusion, the bias of this estimator is

$$\mathbf{bias}_{\hat{\lambda}}(\lambda) = \mathbb{E}[\hat{\lambda}] - \lambda = \frac{\lambda}{2n},$$

and so this estimator is biased but the bias converges to zero as $n \to \infty$. For $\tilde{\lambda}$ the derivation is simpler. Note that

$$\mathbb{E}[\tilde{\lambda}] = \mathbb{E}\left[\frac{\sum_{i=1}^n X_i^2}{6n}\right] = \frac{1}{6n}\sum_{i=1}^n \mathbb{E}\left[X_i^2\right] = \frac{1}{6n}\sum_{i=1}^n 6\lambda = \lambda.$$

In conclusion, $\tilde{\lambda}$ is unbiased since (regardless of n) its bias is

$$\textbf{bias}_{\tilde{\lambda}}(\lambda) = \mathbb{E}[\tilde{\lambda}] - \lambda = 0.$$

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

Solution: The MSE of $\tilde{\lambda}$ is equal to its variance, since the estimator is unbiased :

$$\mathbb{V}(\tilde{\lambda}) = \mathbb{V}\left(\frac{\sum_{i=1}^n X_i^2}{6n}\right) = \frac{1}{36n^2}\sum_{i=1}^n \mathbb{V}(X_i^2) = \frac{\mathbb{E}[X_i^4] - (\mathbb{E}[X_i^2])^2}{36n} = \frac{120\lambda^2 - 36\lambda^2}{36n} = \frac{7}{3n}\lambda^2,$$

so that $\textbf{MSE}_{\tilde{\lambda}}(\lambda) = \frac{7}{3n}\lambda^2$.

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?

Solution: By the bias-variance decomposition , the MSE of $\hat{\lambda}$ is

$$\textbf{MSE}_{\hat{\lambda}}(\lambda) = \textbf{bias}^2_{\hat{\lambda}}(\lambda) + \textbf{Var}_{\hat{\lambda}}(\lambda) = \frac{\lambda^2}{4n^2} + \frac{2\lambda^2}{n}.$$

For $n = 100$ we have $\textbf{MSE}_{\hat{\lambda}}(\lambda) < \textbf{MSE}_{\tilde{\lambda}}(\lambda)$ and so the first estimator is preferable in that regard. Actually, the first estimator is better provided $n \geq 3/4$, which effectively means it is always better in terms of MSE, regardless of the sample size.

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** $\Gamma'(\alpha)$.)

Solution: For a random sample, the likelihood is the product of the density of each observation evaluated at that observation:

$$L(\theta) = \frac{\beta^\alpha X_1^{\alpha-1} e^{-\beta X_1}}{\Gamma(\alpha)} \times \cdots \times \frac{\beta^\alpha X_n^{\alpha-1} e^{-\beta X_n}}{\Gamma(\alpha)} = \frac{\beta^{n\alpha} \left(\prod_{i=1}^n X_i\right)^{\alpha-1} e^{-\beta \sum_{i=1}^n X_i}}{\Gamma(\alpha)^n}.$$

To get the MLE we take the take the natural logarithm to get the log-likelihood:

$$\ell(\theta) = n\alpha \log \beta + (\alpha - 1) \log \left(\prod_{i=1}^n X_i \right) - \beta \sum_{i=1}^n X_i - n \log \Gamma(\alpha).$$

The system that we would have to solve to get the MLE would then be

$$\begin{cases} \frac{\partial \ell(\theta)}{\partial \alpha} & = 0 \\ \frac{\partial \ell(\theta)}{\partial \beta} & = 0 \end{cases} \Leftrightarrow \begin{cases} n \log \beta + \log \left(\prod_{i=1}^n X_i\right) - n \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} & = 0 \\ \frac{n\alpha}{\beta} - \sum_{i=1}^n X_i & = 0 \end{cases}.$$

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.

Solution: Since $\mathbb{V}X = \mathbb{E}(X^2) - (\mathbb{E}X)^2$**, then** $\mathbb{E}(X^2) = \mathbb{V}X + (\mathbb{E}X)^2 = (\alpha + \alpha^2)/\beta^2$**. The MME of** α **and** β **then satisfy** $\overline{X} = \hat{\alpha}/\hat{\beta}$ **and** $\overline{X^2} = (\hat{\alpha} + \hat{\alpha}^2)/\hat{\beta}^2$**. From the first relation we have that** $\overline{X}\hat{\beta} = \hat{\alpha}$ **which when plugged into the second relation gives** $\overline{X^2} = \overline{X}/\hat{\beta} + \overline{X}^2$**, or** $\hat{\beta} = \overline{X}/(\overline{X^2} - \overline{X}^2)$**. Plugging this into the first relation gives** $\hat{\alpha} = \overline{X}^2/(\overline{X^2} - \overline{X}^2)$**.**

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.

Solution: If $\beta = \alpha$**, then** $\mathbb{E}X = 1$ **which does not depend on** α**. So it is not possible to get an MME of** α **from the first moment.**

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

Solution: If $\beta = \alpha$**, then the second moment is** $\mathbb{E}(X^2) = 1/\alpha + 1$**. So the MME satisfies** $\overline{X^2} = 1/\hat{\alpha} + 1$**. Solving for** $\hat{\alpha}$ **gives** $\hat{\alpha} = 1/(\overline{X^2} - 1)$**.**

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**.)

Solution: The sample Mean is $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 46.616/234 = 0.199213675213675$ **(seconds) , the sample variance is** $s^2 = \frac{n}{n-1}\left(\overline{x^2} - \bar{x}^2\right) = (234/233) \times \left(12.098/234 - 0.199213675213675^2\right) = 0.0120683698551044$ **(seconds^2) , the standard deviation is** $s = \sqrt{0.0120683698551044} = 0.109856132532983$ **(seconds) , and the range is (read from the box-plot)** $0.7438 - 0.0013 = 0.7425$ **(seconds).**

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

Solution: It is not reasonable to assume normality. The box-plot is not symmetric and has a large number of outliers on one side. The histogram is not bell-shaped, not symmetric, and the tails look differently thick. The points in the QQ plot are clearly not arranged in a line. All these indicate that the normality assumption is not reasonable at all.

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.)

Solution: Since we are told that we can think of our sample as being a random sample (i.i.d. sample) by the CLT we know that

$$\frac{\bar{X} - \mathbb{E}X}{\sqrt{\mathbb{V}X/n}} \approx N(0,1),$$

and by the law of large numbers we know that $S^2 \approx \mathbb{V}X$**, where** S^2 **is the sample variance, so we conclude that**

$$T = \frac{\bar{X} - \mathbb{E}X}{\sqrt{S^2/n}} = \sqrt{n}\frac{\bar{X} - \mathbb{E}X}{S} \approx N(0,1),$$

is a near-pivot for the expectation of X. (Since n is large you could also use the t_{233} quantiles because the two are close.)

By the definition of the quantiles,

$$\mathbb{P}\left(z_{0.05} \leq T \leq z_{0.95}\right) \approx 0.95 - 0.05 = 0.9,$$

so that

$$\mathbb{P}\left(z_{0.05} \leq \sqrt{n}\frac{\bar{X} - \mathbb{E}X}{S} \leq z_{0.95}\right) \approx 0.9.$$

Isolating $\mathbb{E}X$ in the middle (and eventually using that $z_{0.05} = -z_{1-0.05}$) leads to the approximate 90% CI for $\mathbb{E}X$,

$$\left[\bar{X} - z_{0.95}\frac{S}{\sqrt{n}}, \bar{X} + z_{0.95}\frac{S}{\sqrt{n}}\right].$$

Plugging in all of the information we get the following realisation of the CI:

$$\left[\bar{x} - z_{0.95}\frac{s}{\sqrt{n}}, \bar{x} + z_{0.95}\frac{s}{\sqrt{n}}\right] = \left[0.1992 - 1.64\frac{0.1099}{\sqrt{234}}, 0.1992 - 1.64\frac{0.1099}{\sqrt{234}}\right] = [0.1874, 0.2110] \text{ (sec.)}.$$

The interval is reasonable since even though the data doesn't seem to be Normal, n is large and therefore the CLT and LLN apply.

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

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

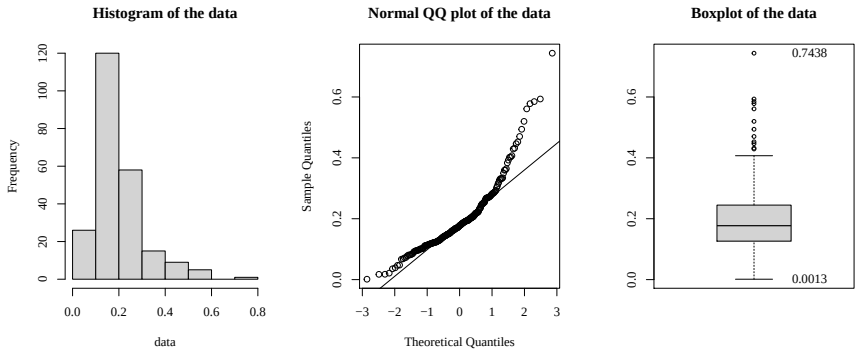

Figure 1: Some summary plots for the keystroke dataset.

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:

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