X_400004 - Statistics
Solutions to the Mock Midterm Exam

14 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: 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?

Solution: First compute the expectation of $\hat{\lambda}$:

$$\mathbb{E}\hat{\lambda} = \mathbb{E}\left(3(X_1^2 + X_2^2)\right) = 3\left(\mathbb{E}(X_1^2) + \mathbb{E}(X_2^2)\right) = 3\left(\frac{\lambda}{6} + \frac{\lambda}{6}\right) = \lambda.$$

Conclude that the estimator $\hat{\lambda}$ is unbiased, since its bias is

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

The expectation of $\tilde{\lambda}$ is

$$\begin{aligned} \mathbb{E}\left(\tilde{\lambda}\right) &= \mathbb{E}\left(X_1^2 + X_2^2 + 2X_1X_2\right) = \mathbb{E}(X_1^2) + \mathbb{E}(X_2^2) + 2\mathbb{E}(X_1X_2) \\ &= \mathbb{E}(X_1^2) + \mathbb{E}(X_2^2) + 2\mathbb{E}(X_1)\mathbb{E}(X_2), \textbf{ since } X_1 \textbf{ and } X_2 \textbf{ are independent} \\ &= \frac{\lambda}{6} + \frac{\lambda}{6} + 2\frac{\left(\sqrt{\lambda}\right)^2}{3^2} = \frac{\lambda}{3} + \frac{2}{9}\lambda = \frac{5}{9}\lambda. \end{aligned}$$

It follows that this estimator is biased since its bias is not zero:

$$\textbf{bias}_{\tilde{\lambda}}(\lambda) = \mathbb{E}\left(\tilde{\lambda}\right) - \lambda = \frac{5}{9}\lambda - \lambda = -\frac{4}{9}\lambda,$$

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

Solution:

$$\mathbb{V}(\hat{\lambda}) = \mathbb{V}\left(3(X_1^2 + X_2^2)\right) = 3^2\mathbb{V}\left(X_1^2 + X_2^2\right) \overset{\textbf{independence}}{=} 9\left(\mathbb{V}(X_1^2) + \mathbb{V}(X_2^2)\right) = 18\mathbb{V}(X_1^2).$$

Now $\mathbb{V}(X_1^2) = \mathbb{E}(X_1^4) - \left(\mathbb{E}(X_1^2)\right)^2 = \frac{\lambda^2}{15} - \left(\frac{\lambda}{6}\right)^2 = \frac{7}{180}\lambda^2$**. Therefore Var**$_{\hat{\lambda}}(\lambda) = \mathbb{V}(\hat{\lambda}) = \frac{7}{10}\lambda^2$**. Since** $\hat{\lambda}$ **is unbiased the MSE is just the variance, therefore MSE**$_{\hat{\lambda}}(\lambda) = \frac{7}{10}\lambda^2$**.**

(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.

Solution: Since

$$\textbf{MSE}_{\hat{\lambda}}(\lambda) = \frac{7}{10}\lambda^2 = \frac{63}{90}\lambda^2 > \frac{41}{90}\lambda^2 = \textbf{MSE}_{\tilde{\lambda}}(\lambda),$$

the second estimator is preferable, at least in terms of mean squared error. In this case, the small amount of bias is compensated by a significantly smaller variance.

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 $X_1, \dots, X_n$ 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.

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

$$L(\theta) = e^{-(X_1-\theta)}1\{X_1 \geq \theta\} \times \cdots \times e^{-(X_n-\theta)}1\{X_n \geq \theta\} = e^{-\sum_{i=1}^n X_i + n\theta} \prod_{i=1}^n 1\{X_i \geq \theta\}$$

$$= e^{-n(\bar{X}-\theta)}\, 1\{X_{(1)} \geq \theta\},$$

where $\bar{X}$ is the sample mean and $X_{(1)}$ is the sample minimum. We cannot simply compute the MLE by differentiating the log-likelihood because the log-likelihood is not differentiable with respect to θ. The comes from the fact that the range of values that X can take depends on θ.

(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.

Solution: You can compute $\mathbb{E}X$ and $\mathbb{V}X$ from the density but there is a more direct solution: $\mathbb{E}X = \mathbb{E}(Y+\theta) = \theta + \mathbb{E}Y = \theta + 1$, and $\mathbb{V}X = \mathbb{V}(Y+\theta) = \mathbb{V}Y = 1$. This is because Y has an exponential distribution with parameter $\lambda = 1$ and the expectation and variance of an $Exp(\lambda)$ random variable are respectively $1/\lambda$ and $1/\lambda^2$ (both 1 here.)

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

Solution: From (b) we have to solve $1 + \theta = \bar{X}$, where $\bar{X}$ is the sample mean. Solving for θ gives us $\hat{\theta} = \bar{X} - 1$.

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

Solution: We have that $\mathbb{E}\hat{\theta} = \mathbb{E}\bar{X} - 1 = \mathbb{E}X - 1 = 1 + \theta - 1 = \theta$ and so the estimator is unbiased. In this case, by the bias-variance decomposition, MSE$_{\hat{\theta}}(\theta) = 0^2 + \mathbb{V}\hat{\theta} = \mathbb{V}(\bar{X} - 1) = \mathbb{V}(\bar{X}) = (\mathbb{V}X)/n = 1/n$, since the data is i.i.d..

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.)

Solution: The sample Mean is $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 186.829149/40 = 4.6707287275$ **(ms), the sample variance is** $s^2 = \frac{n}{n-1}\left(\overline{x^2} - \bar{x}^2\right) = (40/39) \times \left(1125.008716/40 - 4.6707287275^2\right) = 6.47129338186001$ **(ms^2), the standard deviation is** $s = \sqrt{6.47129338186001} = 2.54387369612959$ **(ms), and the range is (read from the box-plot)** $10.78203 - 0.5685252 = 10.2135048$ **(ms).**

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

Solution: There is not strong evidence against the data being normal. The box-plot is relatively symmetric and has no outliers. From the histogram it is hard to say something. Most points in the normal QQ plot are close to a straight line, which is an indication that the normality assumption might be reasonable.

(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.)

Solution: The derivation of this CI can be found in the lecture slides and is not repeated here. The desired CI is given by

$$\left[\frac{(n-1)S^2}{x^2_{39;0.95}}, \frac{(n-1)S^2}{x^2_{39;0.05}}\right].$$

(This is a CI for the variance of a Normal sample with unknown expectation.) The quantiles are $x^2_{39;0.05} = 25.70$ **and** $x^2_{39;0.95} = 54.57$ **and therefore, a two-sided 90% CI for the variance is**

$$\left[\frac{39 \times 6.47129338186001}{54.57}, \frac{39 \times 6.47129338186001}{25.70}\right] = [4.624705, 9.822012].$$

The computed interval is as reasonable as the normality assumption is reasonable. Based on (b) we concluded that it is reasonable to assume that the data is Normal, and so we conclude here that the CI is reasonable.

```
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
```

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

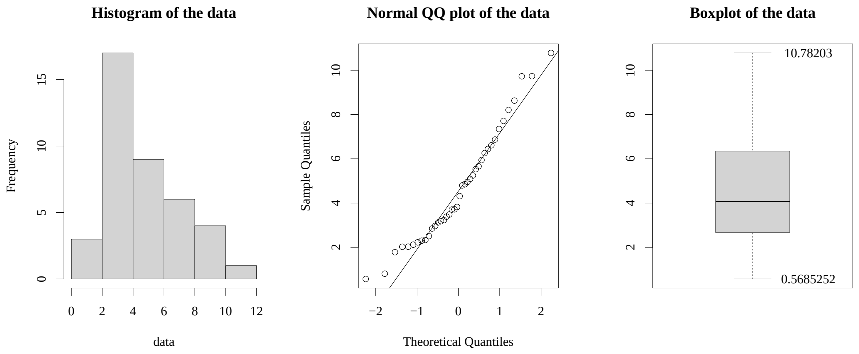

Figure 1: Some summary plots for the dataset.

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:

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