X400004 - Statistics
Solutions to the Midterm

28 October 2022

Below are answers to the exam questions. Some of these are slightly abbreviated, while others include extra comments. These are for your reference only. Your answers during the actual exam should be complete and your steps justified. Keep that in mind when preparing for upcoming exams, and when you write your own answers while preparing for them. As usual, take solutions with a grain of salt: there might be typos, and there are typically different ways to approach each question.

Prob.I: A company manager is curious about the fraction of time that workers are coming to the corporate offices. If this proportion is low, then she may consider renting out less office space.

The manager looked through the literature and decided to model the proportion of workers coming to the office using a *triangular distribution* – so named for the (triangular) shape of its density. This is what the densities in the model that she picked look like:

$$f_\theta(x) = \begin{cases} \frac{2}{\theta}x, & 0 \le x \le \theta, \\ \frac{2}{1-\theta}(1-x), & \theta \le x \le 1, \\ 0, & \text{otherwise.} \end{cases}$$

where $0 < \theta < 1$ is an unknown parameter.

A sample $X_1, \dots, X_n$, was collected where X_i represents the proportion of workers that was at the office in week i so that 0 means no one came to the office and 1 means everyone was at the office. You can assume that this is a random sample from the model above.

Answer the following questions.

4 pts (a) It can be shown that for a random variable X that is distributed like f_θ,

$$\mathbb{E}(X^p) = \frac{2}{(p+1)(p+2)} \frac{1-\theta^{p+1}}{1-\theta}, \quad p = 1, 2, \cdots .$$

What are the first two moments of a random variable X distributed like f_θ?

Solution: The p-th moment of X is $\mathbb{E}(X^p)$ so we just have to plug in p is 1 and 2 into the formula to get, respectively:

$$\mathbb{E}[X] = \frac{2}{6}\frac{1-\theta^2}{1-\theta} = \frac{1}{3}\frac{(1-\theta)(1+\theta)}{1-\theta} = \frac{1+\theta}{3}$$

for the first moment and, for the second moment,

$$\mathbb{E}[X^2] = \frac{2}{12}\frac{1-\theta^3}{1-\theta} = \frac{1}{6}\frac{1-\theta^3}{1-\theta}.$$

10 pts (b) Suppose that you are considering using estimators of the form

$$\hat{\theta} = a\bar{X} + b = \frac{a}{n}\sum_{i=1}^{n} X_i + b, \quad a, b \in \mathbb{R},$$

where a, b are some constants that you still have to pick. Answer the following: (i) what is the expectation of $\hat{\theta}$, and (ii) Find a choice of a and b that you could actually use in practice that makes $\hat{\theta}$ an unbiased estimator for θ.

Solution: (i) The expectation of the estimator $\hat{\theta}$ is

$$\mathbb{E}(\hat{\theta}) = \mathbb{E}(a\bar{X} + b) = a\mathbb{E}(\bar{X}) + b = a\mathbb{E}(X) + b = a\frac{1+\theta}{3} + b.$$

(ii) The estimator is unbiased if $\mathbb{E}(\hat{\theta}) = \theta$. We have that, for instance,

$$a\frac{1+\theta}{3} + b = \theta \qquad \Leftarrow \qquad a = 3, \; b = -1.$$

15 pts (c) Answer the following questions: (i) what is the variance of the estimator $\hat{\theta}$, and (ii) how do a, b affect the variance of $\hat{\theta}$?

Solution: (i) The variance of the estimator $\hat{\theta}$, using (a) is

$$\mathbb{V}(\hat{\theta}) = \mathbb{V}(a\bar{X} + b) = a^2\mathbb{V}(\bar{X}) = \frac{a^2}{n}\mathbb{V}(X) = \frac{a^2}{n}\left[\mathbb{E}(X^2) - (\mathbb{E}X)^2\right] = \frac{a^2}{18n}\left(1 - \theta + \theta^2\right).$$

(ii) From above, we see that the variance increases with $|a|$ and it is not affected at all with b.

15 pts (d) Answer the following: (i) suppose that you decide to set $a = 3$; what is the Mean Squared Error (MSE) of the corresponsing estimator $\hat{\theta}$, and (ii) for that choice of a, what is the choice of b that leads to $\hat{\theta}$ having the smallest MSE?

Solution: (i) Using the bias-variance decomposition, the MSE is the square of the bias plus the variance :

$$\text{MSE}(\theta) = \left(\mathbb{E}\hat{\theta} - \theta\right)^2 + \mathbb{V}\hat{\theta} = \left(a\frac{1+\theta}{3} + b - \theta\right)^2 + \frac{a^2}{18n}\left(1 - \theta + \theta^2\right).$$

Setting $a = 3$ this becomes,

$$\text{MSE}(\theta) = \left(1 + \theta + b - \theta\right)^2 + \frac{3^2}{18n}\left(1 - \theta + \theta^2\right) = \left(1 + b\right)^2 + \frac{1}{2n}\left(1 - \theta + \theta^2\right).$$

(ii) You can take derivatives with respect to b and solve but in this case it is clear that the best choice is to just take $b = -1$.

Prob.II: Someone is modelling the number of trucks that stops at a certain depot during a given hour using the probability mass function

$$f_\theta(x) = \frac{e^{-\frac{1}{\theta}}}{\theta^x\, x!}, \quad x = 0, 1, 2, \cdots,$$

where $\theta > 0$ is some unknown parameter.

In this question, you'll be asked to apply the three methods that you learned in class to find different estimators for the unknown parameter θ based on a random sample $X_1, \dots, X_n$, from the model specified above.

4 pts (a) The expectation of a random variable X distributed like f_θ is $\mathbb{E}X = 1/\theta$. Use this information to find a moment estimator for θ.

Solution: A moment estimator for θ is some quantity $\hat{\theta}$ that satisfies, for some q,

$$\overline{X^q} = \frac{1}{n}\sum_{i=1}^{n} X_i^q = g_q(\hat{\theta}),$$

where $g_q(\theta) = \mathbb{E}(X^q)$. We are given enough information to solve the above for $q = 1$:

$$\bar{X} = 1/\hat{\theta} \Leftrightarrow \hat{\theta} = 1/\bar{X}.$$

This is a moment estimator for θ.

8 pts (b) Find the Maximum Likelihood estimator for θ.

Solution: The density (PMF) of an observation is $\theta^{-x}e^{-1/\theta}/x!$, so the likelihood function is just

$$L(\theta; x_1, \dots, x_n) = \theta^{-x_1}e^{-1/\theta}/x_1! \times \cdots \times \theta^{-x_n}e^{-1/\theta}/x_n! = \theta^{-\sum_{i=1}^n x_i}e^{-n/\theta}/\prod_{i=1}^{n} x_i!.$$

Noting that $\sum_{i=1}^n X_i = n\bar{X}$, the likelihood of the data can be written as

$$L(\theta) = \frac{e^{-n/\theta}}{\theta^{n\bar{X}}\prod_{i=1}^n X_i!},$$

leading to the log-likelihood

$$\ell(\theta) = -\frac{n}{\theta} - n\bar{X}\log\theta - \sum_{i=1}^{n}\log(X_i!).$$

Taking derivative with respect to θ and solving for 0, we get

$$\frac{d\ell(\theta)}{d\theta} = \frac{n}{\theta^2} - \frac{n\bar{X}}{\theta} = 0 \Leftrightarrow \frac{n}{\theta^2} = \frac{n\bar{X}}{\theta} \Rightarrow \hat{\theta} = \frac{1}{\bar{X}}.$$

So the MLE coincides with the MME.

12 pts (c) Suppose that you put an inverse-gamma(α, β) prior on θ. Derive a Bayesian estimator for θ from the respective posterior. Be explicit about what the posterior distribution is.

Hint: If Y has a inverse-gamma(α, β) distribution, $\alpha, \beta > 0$, then the probability density function of Y satisfies $f_{\alpha,\beta}(y) \propto y^{-\alpha-1}e^{-\frac{\beta}{y}}$, $y > 0$, such that the expectation of Y is $\beta/(\alpha - 1)$.

Solution: (i) the prior density on θ is inverse-gamma(α, β) so that $\pi(\theta) \propto \theta^{-\alpha-1}e^{-\frac{\beta}{\theta}}$. In (b) we saw that the likelihood of the data was proportional to $e^{-n/\theta}\theta^{-n\bar{X}}$, so that the posterior density, being proportional to the likelihood times the prior, satisfies

$$\pi(\theta; X_1, \ldots, X_n) \propto \theta^{-\alpha-1}e^{-\frac{\beta}{\theta}} \times e^{-n/\theta}\theta^{-n\bar{X}} = e^{-(n+\beta)/\theta}\theta^{-\alpha-n\bar{X}-1},$$

which we recognise as being proportional to the density of an inverse-gamma(α', β') distribution , with parameters

$$\alpha' = \alpha + n\bar{X}, \qquad \beta' = \beta + n.$$

A possible Bayesian estimator is the posterior expectation which is the expectation of the posterior, i.e., the expectation of an inverse-gamma(α', β') distribution:

$$\frac{\beta'}{\alpha' - 1} = \frac{n + \beta}{\alpha - 1 + n\bar{X}} = \frac{1 + \beta/n}{(\alpha - 1)/n + \bar{X}}.$$

You can see from the above the the Bayes estimator is also not so different from 1 over the sample mean.

Prob.III: Suppose that Tom works at a small business that sells construction supplies. His least favourite job is splitting screws into boxes of 100 units: Tom has to count them one-by-one...

Tom would like to expedite the process and *weigh* the screws (rather than count them) but so he doesn't get into trouble with his boss, he needs to know quite precisely how much 100 screws weighs.

At the end of the questionnaire you can find data about a sample of size 30 of weights of sets of 100 screws (in grams) which Tom did count by hand. You can also find there a collection of descriptive statistics and various graphical representations of the data, as well as some quantiles from commonly used distributions. **Have a look at this information before answering the questions below.**

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

Solution: From the information we are given, the sample mean, sample variance, and sample standard deviation are respectively

$$\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{30} X_i, \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{30}(X_i - \hat{\mu})^2 = \frac{1}{n}\sum_{i=1}^{30} X_i^2 - (\bar{X})^2, \quad \hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{30} X_i^2 - (\bar{X})^2},$$

which, in this particular case, evaluate to

$$\frac{2996.467}{30} \approx 99.882\, g, \ \frac{299468.299}{30} - (99.882)^2 \approx 5.8627\, g^2, \ \sqrt{5.8627} \approx 2.421\, g,$$

the units being respectively grams, grams squared, and grams. It would also be ok to report the following estimate of the variance

$$S^2 = \frac{n}{n-1}\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{30}(X_i - \hat{\mu})^2 = \frac{1}{29}174.4389 = 6.015,$$

with corresponding estimate S of the standard deviation being 2.453 g.

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

Solution: The plots seem to be compatible with normality. The histogram is reasonably symmetric and has appropriate tails, the normal QQ plot seems to fit the diagonal line quite closely, and the box-plot is fairly symmetric and has no outliers.

14 pts (c) Irrespectively of your answer to (b), assume that the normal model is appropriate. Construct an exact two-sided, 95% confidence interval for the expectation of the weight of **one screw** 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; you can find quantiles that you may need in this question at the end of the questionnaire.)*

Solution: We are told to assume that the data is normal, we do not know $\mu = \mathbb{E}X$ or σ^2, and we want a pivot for μ so clearly we need to work with the t-pivot:

$$T = \sqrt{n}\frac{\bar{X} - \mathbb{E}X}{S} \sim t_{n-1},$$

where $S = \sqrt{S^2}$; this is an exact pivot for $\mathbb{E}X$. (Note that the above is incorrect if you replace S with $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$.) As such, we know that if $t_{29;0.975}$ is the 0.975-quantile of a t_{29}-distribution such that if $T \sim t_{29}$, then $\mathbb{P}(T \leq t_{29;0.975}) = 0.975$, we can write

$$0.95 = \mathbb{P}(t_{29;0.025} \leq T \leq t_{29;0.975}) = \mathbb{P}(-t_{29;0.975} \leq T \leq t_{29;0.975}) = \mathbb{P}(-t_{29;0.975} \leq \sqrt{n}\frac{\bar{X} - \mathbb{E}X}{S} \leq t_{29;0.975}),$$

so that by solving for $\mathbb{E}X$, the above probability is

$$\mathbb{P}\left(\bar{X} - \frac{t_{29;0.975}\, S}{\sqrt{n}} \leq \mathbb{E}X \leq \bar{X} + \frac{t_{29;0.975}\, S}{\sqrt{n}}\right)$$

Note however that we want a confidence interval for the expected weight of 1 screw, so we want a confidence interval for $\mu/100$, but immediately,

$$\mathbb{P}\left(\frac{\bar{X}}{100} - \frac{t_{29;0.975}\, S}{100\,\sqrt{n}} \leq \frac{\mathbb{E}X}{100} \leq \frac{\bar{X}}{100} + \frac{t_{29;0.975}\, S}{100\,\sqrt{n}}\right)$$

meaning that the following is a 95% confidence interval for $\mathbb{E}X/100$:

$$\left[\frac{\bar{X}}{100} - \frac{t_{29;0.975}}{100}\frac{S}{\sqrt{n}}, \frac{\bar{X}}{100} + \frac{t_{29;0.975}}{100}\frac{S}{\sqrt{n}}\right].$$

Plugging in the estimates from (a) and the fact that $t_{29;0.975} = 2.045$ we get

$$\left[\frac{99.882}{100} - \frac{2.045}{100}\frac{2.453}{\sqrt{30}}, \frac{99.882}{100} + \frac{2.045}{100}\frac{2.453}{\sqrt{30}}\right] \approx [0.9897,\ 1.0008]\ \textbf{gram}.$$

6 pts (d) How large should the sample size be so that you are 95% confident that you get the expected weight of one screw off by strictly less than 0.001g?

Solution: The actual expected weight is in the CI from (c) with 95% chance so if the size of the interval is strictly less than 0.001, then the above requirement is fulfilled. The size of the interval corresponds to the upper bound minus the lower bound and so we want

$$\left(\frac{\bar{X}}{100} + \frac{t_{29;0.975}}{100}\frac{S}{\sqrt{n}}\right) - \left(\frac{\bar{X}}{100} - \frac{t_{29;0.975}}{100}\frac{S}{\sqrt{n}}\right) = 2\frac{t_{29;0.975}}{100}\frac{S}{\sqrt{n}} < 0.001.$$

Solving the above for n, this is true if,

$$n > \left(\frac{2 \times t_{29;0.975} \times S}{0.001 \times 100}\right)^2 \approx \left(\frac{2 \times 2.045 \times 2.453}{0.001 \times 100}\right)^2 \approx 10064.43$$

So we should use a sample of at least 10065 boxes to reach such precision.

```
Sorted data (weight in grams):
95.08  95.78  96.84  97.15  97.33  97.43  98.18  98.28  98.44  98.60  98.61
98.82 98.89  99.42  99.46 100.18 100.28 100.32 100.38 100.90 101.00 101.15
101.24 101.75 102.09 103.06 103.13 103.90 104.29 104.47
```

Each observation is the weight of a box of 100 screws, not including the weight of the box.

$$n = 30, \qquad \sum_{i=1}^{30} X_i = 2996.467, \qquad \sum_{i=1}^{30} X_i^2 = 299468.299.$$

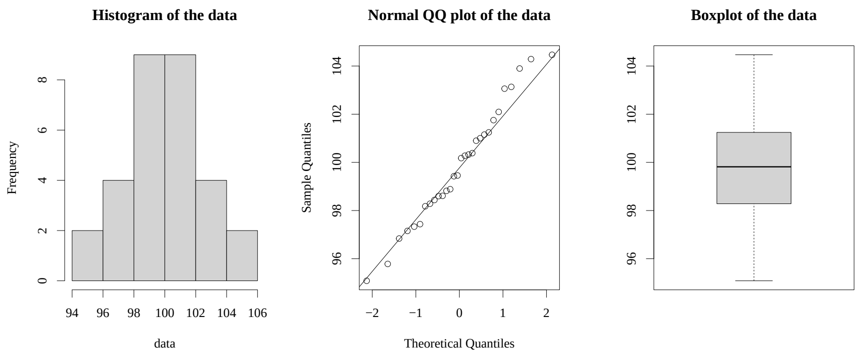

Figure 1: Some summary plots for the dataset.

Some quantiles from the Normal distribution:

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

Some quantiles from the t_{29} distribution:

$t_{29;0.01} = -2.462, t_{29;0.025} = -2.045, t_{29;0.05} = -1.699, t_{29;0.95} = 1.699, t_{29;0.975} = 2.045, t_{29;0.99} = 2.462.$

Some quantiles from the t_{30} distribution:

$t_{30;0.01} = -2.457, t_{30;0.025} = -2.042, t_{30;0.05} = -1.697, t_{30;0.95} = 1.697, t_{30;0.975} = 2.042, t_{30;0.99} = 2.457262.$

Some quantiles from the χ^2_{29} distribution:

$x^2_{29;0.01} = 14.256, x^2_{29;0.025} = 16.047, x^2_{29;0.05} = 17.708, x^2_{29;0.95} = 42.557, x^2_{29;0.975} = 45.722, x^2_{29;0.99} = 49.588.$

Some quantiles from the χ^2_{30} distribution:

$x^2_{30;0.01} = 14.953, x^2_{30;0.025} = 16.791, x^2_{30;0.05} = 18.493, x^2_{30;0.95} = 43.773, x^2_{30;0.975} = 46.979, x^2_{30;0.99} = 50.892.$