X400004 - Statistics
Midterm

28 October 2022

Instructions:

- The exam is to be solved **individually**.
- Please **write clearly and in an organised way**: illegible answers cannot be graded.
- 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 notes exam**; you are only allowed to have one A4 sheet with **handwritten** notes with you.
- You may use a calculator; no cellphone, tablet, computer, smart watch or other such device is allowed.

- There are 4 pages in the exam questionnaire (including this one) and you have 2 hours (120 minutes) to complete the exam.
- The exam consists of 11 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** the answer sheets with your name and student number.

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 $X_1, \ldots, X_n$, 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?

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.

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 $\hat{\theta}$?

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?

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 $X_1, \ldots, X_n$, 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.

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

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 $\beta/(\alpha - 1)$.

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.)

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

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.)*

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?

```
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.

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

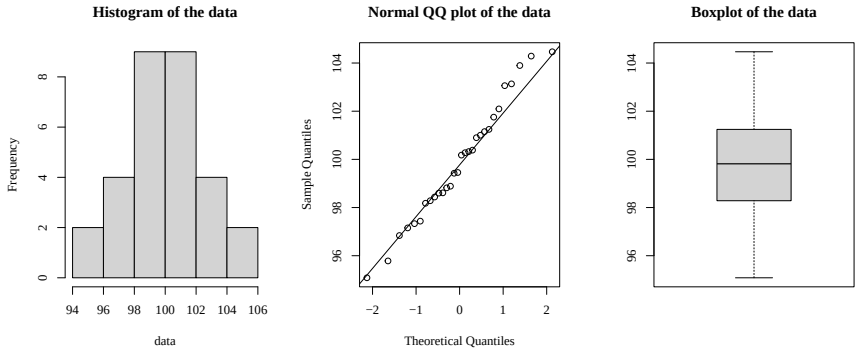

Figure 1: Some summary plots for the dataset.

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:

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