X400004 - Statistics
Midterm

27 October 2023

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 12 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: An administrator is busy trying to figure out the rate at which orders come into a website used to book car rentals. This rate has logistic implication so it is crucial to get as sharp an estimate of it as possible. This is what the PDF's and CDF's in the model that she picked for the time between two consecutive rentals look like respectively:

$$f_\theta(x) = \frac{1}{\theta}\exp\big(-x/\theta\big), \qquad \text{so that} \qquad F_\theta(x) = 1-\exp\big(-x/\theta\big), \quad x \geq 0,$$

where $\theta > 0$ is an unknown parameter. (You will recognise these as corresponding to an exponential distribution with parameter $1/\theta$.)

A sample $X_1, \ldots, X_n$, was collected where each X_i represents the time between two consecutive bookings. You can assume that this is a random sample from the model above.

You are considering using one of the two following estimators for the unknown parameter θ,

$$\hat{\theta} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \text{or} \qquad \tilde{\theta} = n \cdot X_{(1)} = n \cdot \min_{i=1,\ldots,n} X_i.$$

In this problem you will compare the two estimators.

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

$$\mathbb{E}(X^p) = p! \cdot \theta^p, \quad p = 1, 2, \cdots .$$

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

12 pts (b) Do the following: i) show that $X_{(1)}$ is distributed like an exponential random variable with parameter n/θ, and ii) compute the first two moments of $X_{(1)}$. **Hint:** compute $\mathbb{P}(X_{(1)} > x)$ and relate it to the CDF of $X_{(1)}$.

8 pts (c) Compute the mean squared error (MSE) of $\hat{\theta}$. Is the estimator biased?

8 pts (d) Compute the mean squared error (MSE) of $\tilde{\theta}$. Is the estimator biased?

4 pts (e) Justify which estimator is preferred. (In case you did not compute the MSE in (c) and (d), explain instead how you would reach a conclusion.)

Prob.II: Suppose that your company has put you in charge of a screening process for candidates for a given position. You are given a list of essential requirements for the candidates to fulfill, and you have access to a seemingly *infinite* pool of candidates. You pick random candidates from the pool until you find 5 candidates that fulfill all requirements; these 5 will later be interviewed for the position.

You think that a negative binomial distribution with parameters (r, p) with $r = 5$, is a good model for the total number of candidates X that you have to screen until you find your 5 suitable candidates. The probability mass function of the number of candidates that you'll have to screen in total to find a batch of 5 suitable candidates is therefore

$$f_p(x) = \binom{x-1}{x-5}(1-p)^{x-5}p^5, \quad x = 5, 6, \cdots,$$

where $p \in (0, 1)$ is the unknown probability that a randomly chosen person fulfils the requirements. You are interested in estimating p as this tells you something about the quality of the pool of candidates that you are using.

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 p based on a random sample $X_1, \ldots, X_n$ from the model above that you have collected so far.

6 pts (a) The expectation of a random variable X distributed like f_p is $\mathbb{E}X = 5/p$. Use this information to find a moment estimator for p.

10 pts (b) Find the Maximum Likelihood estimator for p.

12 pts (c) Suppose that you put a beta(α, β) prior on p. Identify the posterior distribution and compute the posterior expectation of p under the Bayesian model.

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

Prob.III: Suppose that someone has just become very aware of climate related issues, particularly about water consumption. They read that the expected amount of water that someone spends in a typical shower is 40 litre. They are now very curious about their own expected consumption...

At the end of the questionnaire you can find data making up a sample of size 60 of the amount of water spend (in litres) over roughly two months worth of showers. 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 sample. (Don't forget to report the units.)

9 pts (b) For each of the three plots, briefly explain how it supports/contradicts the possibility that each observation comes from a Normal distribution.

15 pts (c) Regardless of what you conclude in (b), suppose that you use the following

$$T = \sqrt{n}\frac{\bar{X} - \mu}{\sigma} \sim N(0, 1),$$

as an exact pivot for μ, where $\sigma = 5$. Derive from the pivot an exact two-sided, 95% confidence interval for the expectation of the amount of water that is spent on a typical shower from the data at hand and compute its realisation. *(This means that you need to derive the expression for the interval from the pivot, not just write down the interval and plug in all known information, including the data; you can find quantiles that you may need in this question at the end of the questionnaire.)*

4 pts (d) What do you conclude from (c) about the possibility that this person's expected water consumption is the typical 40 litre? (In case you did not solve (c) explain instead how you *would* conclude something from a 95% confidence interval for μ.)

```
Sorted data:
31.64 31.89 31.91 32.85 34.33 34.82 35.13 35.52 35.72 35.82 36.20 36.35 36.52
37.43 37.53 38.84 39.95 39.99 40.39 40.48 40.62 40.93 41.79 41.98 42.59 42.65
43.00 43.27 43.61 43.63 43.80 43.85 43.95 44.24 44.71 45.17 45.32 46.02 46.51
46.59 46.73 47.19 47.35 47.53 47.55 47.92 48.02 48.79 49.34 49.92 51.05 51.08
51.91 54.24 54.39 54.54 55.16 58.95 62.05 63.00
```

Each observation is the about of water (in litres) used in a typical shower.

$$n = 60, \qquad \sum_{i=1}^{60} X_i = 2634.266, \qquad \sum_{i=1}^{60} X_i^2 = 118783.093.$$

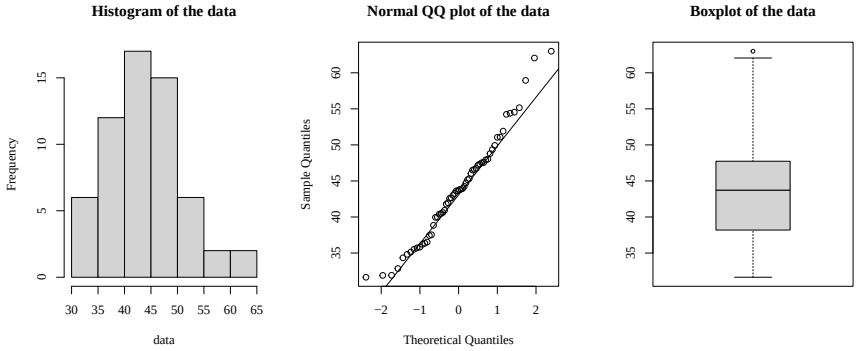

Figure 1: Some summary plots for the dataset.

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:

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