X400004 - Statistics
Midterm

29 October 2021

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, or other such device is allowed.

- There are 5 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: Suppose that a company is developing a new type of milk carton. At this stage of the development, they are choosing the material to make the carton out of. After talking to their materials engineer, they are confident that they have a good model for the amount of pressure that a material can take before bursting. This is what the densities in that model look like:

$$f_\theta(x) = \frac{3}{\theta^3}x^2, \quad 0 \le x \le \theta,$$

where $\theta > 0$ is an unknown parameter that depends on the specific material being stress-tested.

Clearly, they would like to go with a material with large θ, but the only way forward is to estimate θ from data. Suppose that the company collected a random sample $X_1, \ldots, X_n$ from a candidate material. Answer the following questions.

4 pts (a) Show that the p-th moment of a random variable X that is distributed like f_θ satisfies

$$\mathbb{E}(X^p) = \frac{3\theta^p}{p+3}, \quad p = 1, 2, \cdots .$$

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

$$\hat{\theta} = c\bar{X} = \frac{c}{n}\sum_{i=1}^{n} X_i, \quad c > 0,$$

where c is some constant that you still have to pick. Answer the following: (i) what is the expectation of $\hat{\theta}$, and (ii) what choice of c makes $\hat{\theta}$ an unbiased estimator for θ?

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

18 pts (d) Answer the following: (i) what is the Mean Squared Error (MSE) of the estimator $\hat{\theta}$, and (ii) what is the choice of c that leads to $\hat{\theta}$ having the smallest MSE?

Prob.II: Suppose that you are conducting interviews to fill a low level job opening at your company. Company policy is that you interview candidates until you find one that fulfils certain minimal requirements. As such, you think that a geometric distribution is a good model for the number of interviews you have to hold until you find one suitable candidate. The probability mass function of the number of interviews that you'll have to hold to find one candidate is therefore

$$f_p(x) = (1-p)^{x-1}p, \quad x = 1, 2, \cdots,$$

where $p \in (0, 1)$ is some unknown parameter. You expect to have to conduct many such rounds of interviews, so it is important for you to accurately estimate p.

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.

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

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

12 pts (c) Suppose that you put a beta(α, β) prior on p. Derive a Bayesian estimator for p from the respective posterior.

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: Ticket scalping is the practice of buying tickets with the express intent of reselling them later at a higher price. This task it often delegated to bots that automatically purchase tickets when they become available, resulting in many legitimate customers being left ticketless.

Suppose that you work for a company that sells tickets online and you have been tasked with understanding how long an actual person (i.e., not a bot) takes to complete the purchase of a ticket. The goal is to later use this for bot detection.

At the end of the questionnaire you can find data about the time (in seconds) that it takes a legitimate client to purchase a ticket from first hitting the *buy* button, until finishing the payment. You can also find there a collection of descriptive statistics and various graphical representations of the data. **Have a look at this information before answering the questions below.**

8 pts (a) Determine the sample mean, sample variance, sample standard deviation, and range of the dataset. (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 a(n exact) two-sided, 90% confidence interval for the expectation of the amount of time it takes a legitimate user to purchase a ticket 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.)** You can find quantiles that you may need in this question at the end of the questionnaire.

6 pts (d) Is it correct to say that (under the normal model) the realisation of the confidence interval that you got in (c) has exactly a 90% chance of containing the true expectation of the data inside? Justify your answer.

```
Sorted data (time in s):
12.501 16.700 23.024 24.928 25.983 26.610 31.067 31.697 32.624 33.593
33.662 34.908 35.315 38.547 38.730 43.058 43.660 43.939 44.301 47.397
48.012 48.914 49.468 52.520 54.567 60.361 60.807 65.381 67.726 68.804
```

$$n = 30, \qquad \sum_{i=1}^{30} X_i = 1238.803, \qquad \sum_{i=1}^{30} X_i^2 = 57434.253.$$

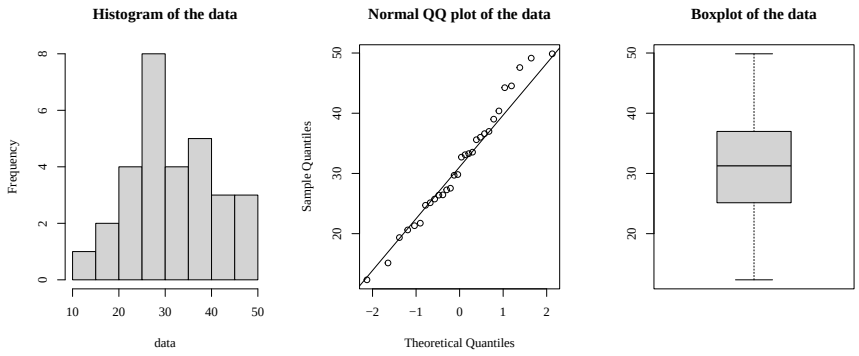

Figure 1: Some summary plots for the dataset.

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:

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