# X400004 - Statistics
# Solutions to the Midterm

## 27 October 2023

**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:** 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 $\theta$,

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

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 $f_\theta$,

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

What are the first two moments of a random variable $X$ distributed like $f_\theta$?

**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:**

$$\mathbb{E}[X] = 1! \cdot \theta^1 = \theta,$$

**for the first moment and, for the second moment,**

$$\mathbb{E}[X^2] = 2! \cdot \theta^2 = 2\theta^2.$$

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

**Solution: (i) Following the hint, we compute the CDF of $X_{(1)}$:**

$$\mathbb{P}(X_{(1)} > x) = \mathbb{P}(X_1 > x, \cdots, X_n > x) = \mathbb{P}(X_1 > x) \cdots \mathbb{P}(X_n > x) = \mathbb{P}(X_1 > x)^n,$$

**using the equivalence of the two events, independence, and the fact that all $X_i$ have the same distribution. Then,**

$$\mathbb{P}(X_1 > x)^n = \big(1 - \mathbb{P}(X_1 \leq x)\big)^n = \exp(-nx/\theta),$$

**by using the supplied CDF, leading to**

$$F_{X_{(1)}}(x) = 1 - \mathbb{P}(X_{(1)} > x) = 1 - \exp(-nx/\theta),$$

**which we identify as the CDF of an exponential distribution with parameter $n/\theta$. (ii) From (b) and since we now know that $X_{(1)}$ has an exponential distribution, we automatically get that $\mathbb{E}(X_{(1)}) = \theta/n$, and $\mathbb{E}(X_{(1)}^2) = 2\theta^2/n^2$.**

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

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

$$\mathbb{E}(\hat{\theta}) = \mathbb{E}(\bar{X}) = \mathbb{E}(X_1) = \theta,$$

**where we use that fact that we have a random sample so the estimator is unbiased. The corresponding variance is**

$$\mathbb{V}(\hat{\theta}) = \mathbb{V}(\bar{X}) = \frac{1}{n}\mathbb{V}(X_1) = \frac{1}{n}(\mathbb{E}(X_1^2) - (\mathbb{E}(X_1))^2) = \frac{1}{n}(2\theta^2 - \theta^2) = \theta^2/n,$$

**where we use that fact that we have a random sample. Putting the two together, the MSE is**

$$\mathrm{MSE}_{\hat{\theta}}(\theta) = (\mathbb{E}(\hat{\theta}) - \theta)^2 + \mathbb{V}(\hat{\theta}) = 0^2 + \theta^2/n = \theta^2/n.$$

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

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

$$\mathbb{E}(\tilde{\theta}) = \mathbb{E}(nX_{(1)}) = n\mathbb{E}(X_{(1)}) = n\frac{\theta}{n} = \theta,$$

**so the estimator is unbiased. The corresponding variance is**

$$\mathbb{V}(\tilde{\theta}) = \mathbb{V}(nX_{(1)}) = n^2\mathbb{V}(X_{(1)}) = n^2(\mathbb{E}(X_{(1)}^2) - (\mathbb{E}X_{(1)})^2) = n^2(2\theta^2/n^2 - (\theta/n)^2) = \theta^2.$$

**Putting the two together, the MSE is**

$$\mathrm{MSE}_{\tilde{\theta}}(\theta) = (\mathbb{E}(\tilde{\theta}) - \theta)^2 + \mathbb{V}(\tilde{\theta}) = 0^2 + \theta^2 = \theta^2.$$

**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.)

**Solution: Clearly, $\hat{\theta}$ is the better estimator since it always has smaller MSE when compared to $\tilde{\theta}$. In fact, the MSE of $\tilde{\theta}$ does not vanish as $n$ grows.**

**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 $p$.

**Solution: A moment estimator for $p$ can be obtained by solving**

$$\overline{X^q} = \frac{1}{n}\sum_{i=1}^{n} X_i^q = \mathbb{E}(X^q),$$

**for some $q$. We are given enough information to solve the above for $q = 1$:**

$$\bar{X} = 5/\hat{p} \Leftrightarrow \hat{p} = 5/\bar{X}.$$

**This is a moment estimator for $p$.**

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

**Solution: The density (PMF) of an observation is $\binom{x-1}{x-5}(1-p)^{x-5}p^5$, so the likelihood function satisfies**

$$L(p; x_1, \ldots, x_n) = \binom{x_1-1}{x_1-5}(1-p)^{x_1-5}p^5 \times \cdots \times \binom{x_n-1}{x-5}(1-p)^{x_n-5}p^5 \propto (1-p)^{\sum_{i=1}^n x_i - 5n}\, p^{5n},$$

**for some proportionality constant that does not depend on $p$ (and therefore does not affect the maximum of the likelihood function.) This means that the likelihood of the data satisfies**

$$L(p) \propto (1-p)^{\sum_{i=1}^n X_i - 5n}\, p^{5n},$$

**leading to the log-likelihood**

$$\ell(p) = \Big(\sum_{i=1}^{n} X_i - 5n\Big)\log(1-p) + 5n\log(p) + \textbf{const.}.$$

**Taking derivative with respect to $p$ and solving for $0$, we get**

$$\frac{d\ell(p)}{dp} = -\Big(\sum_{i=1}^{n} X_i - 5n\Big)\frac{1}{1-p} + \frac{5n}{p} = 0 \Leftrightarrow \frac{5n}{p} = \Big(\sum_{i=1}^{n} X_i - 5n\Big)\frac{1}{1-p} \Leftrightarrow \frac{1-p}{p} = \frac{\bar{X}}{5} - 1,$$

**which can now be easily solved for $p$ since the above is**

$$\frac{1}{p} - 1 = \frac{\bar{X}}{5} - 1 \Leftrightarrow \frac{1}{p} = \frac{\bar{X}}{5} \Rightarrow \hat{p} = 5/\bar{X}.$$

**So the MLE coincides with the MME.**

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

**Hint:** If $Y$ has a beta$(\alpha, \beta)$ 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 $\alpha/(\alpha+\beta)$.

**Solution: The prior density on $p$ is beta$(\alpha, \beta)$ so that $\pi(p) \propto p^{\alpha-1}(1-p)^{\beta-1}$. In (b) we saw that the likelihood of the data was proportional to $(1-p)^{\sum_{i=1}^n X_i - 5n} p^{5n}$, so that the posterior density, being proportional to the likelihood times the prior, satisfies**

$$\pi(p; X_1, \ldots, X_n) \propto (1-p)^{\sum_{i=1}^n X_i - 5n} p^{5n} \times p^{\alpha-1}(1-p)^{\beta-1} = (1-p)^{\sum_{i=1}^n X_i + \beta - 5n - 1} p^{5n+\alpha-1},$$

**which we recognise as being proportional to the density of a beta$(\alpha', \beta')$ distribution, with parameters**

$$\alpha' = \alpha + 5n, \qquad \beta' = \beta + \sum_{i=1}^{n} X_i - 5n.$$

**The posterior expectation is the expectation of the posterior, i.e., the expectation of a beta$(\alpha', \beta')$ distribution:**

$$\frac{\alpha'}{\alpha' + \beta'} = \frac{\alpha + 5n}{\alpha + \beta + \sum_{i=1}^n X_i}.$$

**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.)

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

$$\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{60} X_i, \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{60}(X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^{60} X_i^2 - (\bar{X})^2, \quad \hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{60} X_i^2 - (\bar{X})^2},$$

**which, in this particular case, evaluate to**

$$\frac{2634.266}{60} \approx 43.90\, litre, \quad \frac{118783.093}{60} - (43.90)^2 \approx 53.00\, litre^2, \quad \sqrt{53.00} \approx 7.28\, litre,$$

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

$$S^2 = \frac{n}{n-1}\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{60}(X_i - \bar{X})^2 = \frac{60}{59}53.00 = 53.90,$$

**with corresponding estimate $S$ of the standard deviation being $7,34$ litre.**

**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.

**Solution: The plots seem to be incompatible with normality. The histogram is not symmetric and has a thick left tail, the normal QQ plot seems to fit the diagonal line quite poorly at the tails, and the box-plot is asymmetric.**

**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 $\mu$, 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.)*

**Solution:** **Using the pivot that we are given, we know that if $z_{0.975}$ is the $0.975$-quantile of a standard normal distribution such that if $T \sim N(0,1)$, then $\mathbb{P}(T \leq z_{0.975}) = 0.975$, we can write**

$$0.95 = \mathbb{P}(z_{0.025} \leq T \leq z_{0.975}) = \mathbb{P}(-z_{0.975} \leq T \leq z_{0.975}) = \mathbb{P}(-z_{0.975} \leq \sqrt{n}\frac{\bar{X} - \mu}{\sigma} \leq z_{0.975}),$$

**so that by solving for $\mu$, the above probability is**

$$\mathbb{P}\left(\bar{X} - \frac{z_{0.975}\,\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + \frac{z_{0.975}\,\sigma}{\sqrt{n}}\right),$$

**meaning that the following is a $95\%$ confidence interval for $\mu$:**

$$\left[\bar{X} - z_{0.975}\frac{\sigma}{\sqrt{n}},\ \bar{X} + z_{0.975}\frac{\sigma}{\sqrt{n}}\right].$$

**Plugging in the quantile, $\sigma = 5$, and $z_{0.975} = 1.96$ we get**

$$\left[43.90 - 1.96 \cdot \frac{5}{\sqrt{60}},\ 43.90 + 1.96 \cdot \frac{5}{\sqrt{60}}\right] \approx [42.63,\ 45.17] \text{ litre}.$$

**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 $\mu$.)

**Solution:** **By definition, the confidence interval is very likely (95% sure) to contain the expected amount of water used in a typical shower by this person. Since the confidence interval is fully above 40, we have to conclude that this person is likely expected to spend more than 40 litre of water in a typical shower.**

```
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.

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

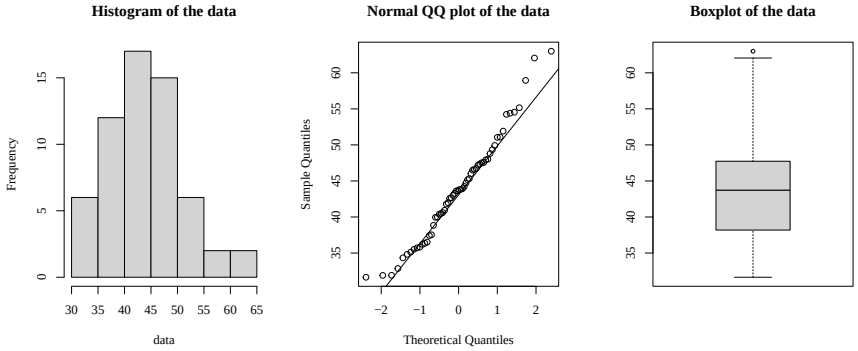


Figure 1: Some summary plots for the dataset.

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 $\chi^2_{29}$ 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 $\chi^2_{30}$ distribution:

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