X400004 - Statistics
Solutions to the Midterm

29 October 2021

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: 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

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

Solution: The p-th moment $\mathbb{E}(X^p)$ is the integral

$$\int_0^\theta x^p f_\theta(x)\,dx = \int_0^\theta x^p \frac{3}{\theta^3}x^2\,dx = \frac{3}{\theta^3}\int_0^\theta x^{p+2}\,dx = \frac{3}{\theta^3}\left[\frac{x^{p+3}}{p+3}\right]_0^\theta = \frac{3}{\theta^3}\left[\frac{\theta^{p+3}}{p+3} - 0\right] = \frac{3\theta^p}{p+3}.$$

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?

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

$$\mathbb{E}(\hat{\theta}) = \mathbb{E}(c\bar{X}) = c\mathbb{E}(\bar{X}) = c\mathbb{E}(X) = c\frac{3\theta^1}{1+3} = c\frac{3\theta}{4}.$$

(ii) The estimator is unbiased if $\mathbb{E}(\hat{\theta}) = \theta$ which only happens if we take $c = 4/3$.

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 $\hat{\theta}$?

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

$$\mathbb{V}(\hat{\theta}) = \mathbb{V}(c\bar{X}) = c^2\mathbb{V}(\bar{X}) = \frac{c^2}{n}\mathbb{V}(X) = \frac{c^2}{n}\left[\mathbb{E}(X^2) - (\mathbb{E}X)^2\right] = \frac{c^2}{n}\left[\frac{3}{5} - \left(\frac{3}{4}\right)^2\right]\theta^2 = \frac{c^2}{n}\frac{3}{80}\theta^2.$$

(ii) From above, we see that the variance increases with c.

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?

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

$$\text{MSE}(\theta) = \left(\mathbb{E}\hat{\theta} - \theta\right)^2 + \mathbb{V}\hat{\theta} = \left(c\frac{3\theta}{4} - \theta\right)^2 + \frac{c^2}{n}\frac{3}{80}\theta^2 = \left\{\left(\frac{3c-4}{4}\right)^2 + \frac{c^2}{n}\frac{3}{80}\right\}\theta^2.$$

(ii) To get the smallest MSE we take the derivative of the MSE (with respect to c) and solve for 0:

$$\frac{d}{dc}\text{MSE} = 0 \Leftrightarrow 3\frac{3c-4}{8}\theta^2 + 3\frac{c}{40n}\theta^2 = 0 \Leftrightarrow (30n+2)c = 40n \Leftrightarrow c = \frac{20n}{15n+1}.$$

Note that the second derivative of the MSE with respect to c is

$$\left(\frac{9}{8} + \frac{3}{40n}\right)\theta^2 > 0,$$

so that indeed we have found a minimiser of the MSE. So the estimator of the form $\hat{\theta} = c\bar{X}$, $c > 0$, with the smallest MSE is

$$\frac{20}{15n+1}\sum_{i=1}^{n} X_i.$$

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.

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 $q = 1$:

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

This is the moment estimator for p.

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

Solution: The density (probability mass function) of an observation is $(1-p)^{x-1}\,p$, so the likelihood function is just

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

This means that the likelihood of the data is

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

leading to the log-likelihood

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

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

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

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

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

So the MLE coincides with the MME.

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 $\alpha/(\alpha+\beta)$.

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

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

which we recognise as being proportional to the density of a beta(α', β') distribution , with parameters

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

A possible Bayesian estimator is the posterior expectation which is the expectation of the posterior, i.e., the expectation of a beta(α', β') distribution:

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

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.)

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

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

which in this particular case evaluate to

$$\frac{1238.803}{30} \approx 41.293, \ \frac{57434.253}{30} - (41.293)^2 \approx 209.328, \ \sqrt{209.328} \approx 14.468, \ 68.804 - 12.501 = 56.303,$$

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

$$S^2 = \frac{n}{n-1}\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{30}(X_i - \hat{\mu})^2 = \frac{30}{29}209.328 = 216.546,$$

with corresponding estimate S of the standard deviation being 14.716.

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 thin tails (the fatter right-most tail is likely because of the bin choice), 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 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.

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 $t_{29;0.95}$ is the 0.95-quantile of a t_{29}-distribution such that if $Z \sim t_{29}$, then $\mathbb{P}(Z \leq t_{29;0.95}) = 0.95$, we can write

$$0.9 = \mathbb{P}(t_{29;0.05} \leq T \leq t_{29;0.95}) = \mathbb{P}(-t_{29;0.95} \leq T \leq t_{29;0.95}) = \mathbb{P}(-t_{29;0.95} \leq \sqrt{n}\frac{\bar{X} - \mathbb{E}X}{S} \leq t_{29;0.95}),$$

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

$$\mathbb{P}(\bar{X} - t_{29;0.95}S/\sqrt{n} \leq \mathbb{E}X \leq \bar{X} + t_{29;0.95}S/\sqrt{n})$$

meaning that the following is a 90% confidence interval for $\mathbb{E}X$:

$$\left[\bar{X} - t_{29;0.95}\frac{S}{\sqrt{n}},\ \bar{X} + t_{29;0.95}\frac{S}{\sqrt{n}}\right].$$

Plugging in the estimates from (a) and the fact that $t_{29;0.95} = 1.699$ we get

$$\left[41.293 - 1.699\frac{14.716}{\sqrt{30}},\ 41.293 + 1.699\frac{14.716}{\sqrt{30}}\right] = [36.728,\ 45.858] \text{ seconds.}$$

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.

Solution: The statement is not correct. The realisation of the interval that we got, $[36.728,\ 45.858]$, is not random and, as such, either contains the true expectation of the data or not.

```
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
```

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

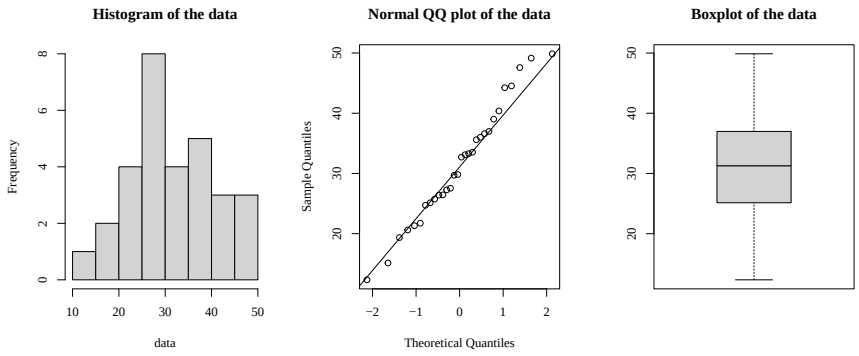

Figure 1: Some summary plots for the dataset.

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:

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