X_400004 - Statistics
Solutions to the Resit Exam

15 February 2022

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: The vending machine at the station never works properly – when you scan your bank card, there is only a chance $p \in (0, 1)$ that the payment goes through, and if it doesn't, then you have to try again. For this problem, think of the number of times that you have to scan your card until you have a successful payment as having a geometric distribution with parameter p.

Suppose that $X_1, \dots, X_n$ is a random sample distributed like $X \sim \text{geom}(p)$, for some parameter $p \in (0, 1)$. These correspond to the number of payment attempts for each of your last n independent purchases.

Remember that the geometric distribution is discrete, having probability mass function

$$f_p(x) = \mathbb{P}(X = x) = p(1 - p)^{x-1}, \quad x = 1, 2, \dots,$$

such that $\mathbb{E}X = 1/p$, and $\mathbb{V}X = (1 - p)/p^2$.

To have a sharp estimate of the expected number of tries you need a sharp estimate of p. This might help you miss the train less often (because you are definitely not leaving until you get your snack.)

4 pts (a) Find the method of moments estimator of p.

Solution: We have that $\mathbb{E}X = 1/p$ so the MME solves $\bar{X} = 1/\hat{p}$. This leads to $\hat{p} = 1/\bar{X}$.

6 pts (b) What is the maximum likelihood estimator of p?

Solution: The likelihood function is

$$f_p(x_1, \dots, x_n) = p(1 - p)^{x_1-1} \times \cdots \times p(1 - p)^{x_n-1} = p^n(1 - p)^{\sum_{i=1}^n x_i - n} = p^n(1 - p)^{n\bar{x}-n}.$$

Plugging in the random variables we get the likelihood of the data:

$$L(p) = p^n(1 - p)^{n\bar{X}-n},$$

and the respective lo-likelihood:

$$\ell(p) = n \log(p) + n(\bar{X} - 1) \log(1 - p);$$

taking derivatives and setting to zero we get

$$\frac{d\ell(p)}{dp} = \frac{n}{p} - \frac{n(\bar{X}-1)}{1-p} = 0 \Leftrightarrow \frac{1-p}{p} = \bar{X} - 1 \Leftrightarrow \frac{1}{p} - 1 = \bar{X} - 1 \Leftrightarrow p = 1/\bar{X},$$

leading to the MLE $\tilde{p} = 1/\bar{X}$ which happens to coincide with the MME $\hat{p}$.

6 pts (c) Suppose that you put a beta(α, β) prior on p. What is the posterior distribution? and what is the posterior expectation?

Solution: From the previous question, we know that the likelihood is

$$L(p) = p^n(1-p)^{n\bar{X}-n};$$

multiplying this with the prior $\pi(p) \propto p^{\alpha-1}(1-p)^{\beta-1}$ we get that the posterior is proportional to

$$\pi(p \mid X_1, \ldots, X_n) \propto p^{\alpha+n-1}(1-p)^{\beta+n\bar{X}-n-1}.$$

We recognise this as being proportional to the density of a beta$(\alpha+n, \beta+n\bar{X}-n)$ distribution. The posterior expectation is the expectation of a beta$(\alpha+n, \beta+n\bar{X}-n)$ distribution which is $(\alpha+n)/(\alpha+\beta+n\bar{X})$.

9 pts (d) Suppose that you have two estimators for p, call them, $\hat{p}$ and $\tilde{p}$. (Not necessarily the MME or the MLE from before.) If the bias of the two estimators is respectively 0 and $\sqrt{p/n}$, and the variance of the two estimators is respectively p^2/n and $p(1-p)/n$, which of the two estimators is the best? and in what sense?

Solution: Using the bias variance decomposition,

$$\begin{aligned} MSE_{\hat{p}}(p) &= (0)^2 + \frac{p^2}{n} = \frac{p^2}{n}, \\ MSE_{\tilde{p}}(p) &= \left(\sqrt{\frac{p}{n}}\right)^2 + \frac{p(1-p)}{n} = \frac{p}{n} + \frac{p(1-p)}{n}. \end{aligned}$$

We always prefer the estimator with the smallest MSE so let us for instance check when it is true that $MSE_{\hat{p}}(p) \leq MSE_{\tilde{p}}(p)$. This is equivalent to

$$\frac{p^2}{n} \leq \frac{p}{n} + \frac{p(1-p)}{n} \Leftrightarrow p \leq 1 + 1 - p \Leftrightarrow 2p \leq 2 \Leftrightarrow p \leq 1.$$

Since it is always true that $p \leq 1$ and this is equivalent to $MSE_{\hat{p}}(p) \leq MSE_{\tilde{p}}(p)$, we prefer $\hat{p}$ over $\tilde{p}$ as estimator for p since it always (meaning for any combination of n and p) has smaller MSE.

Hint: If $Y \sim$ beta(α, β), for $\alpha, \beta > 0$, then $f_{\alpha,\beta}(y) \propto y^{\alpha-1}(1-y)^{\beta-1}$, and $\mathbb{E}Y = \alpha/(\alpha+\beta)$.

Prob.II: With the proliferation of bots, online businesses are now faced with the prospect that some of the reviews of their products are artificial. Basing market analyses on such review data would lead to poor outcomes... Suppose that a company has developed a score X for how *artificial* a review seems.

The score X is computed in a very complex way leading to it having some distribution you have never heard of. All you know is that $\mathbb{E}X = \theta$ and $\mathbb{V}X = 4\theta^2$, for some unknown $\theta > 0$, and that the Central Limit theorem may be applied to these data.

Let $X_1, \dots, X_n$ be a random sample of scores for actual persons. An online business would like you to construct a confidence interval for θ since they believe that they can use such a confidence interval to build a filter for the reviews.

6 pts (a) Use the Central Limit theorem to show that the distribution of

$$T = \sqrt{n}\frac{\bar{X}_n/\theta - 1}{2},$$

is close to being $N(0,1)$ so that T is a near-pivot for θ. (Above, $\bar{X}_n$ is the sample mean where we emphasise the dependence on the sample size n.)

Solution: The CLT tells us that if $X_1, \dots, X_n$ is a random sample from some distribution with expectation $\mathbb{E}X$ and variance $\mathbb{V}X$, then the following quantity has approximately a standard normal, i.e., $N(0,1)$, distribution:

$$\sqrt{n}\frac{\bar{X} - \mathbb{E}X}{\sqrt{\mathbb{V}X}},$$

where $\bar{X} = \bar{X}_n$ is the sample mean of the n observations. In our particular case, we are told that we have a random sample with $\mathbb{E}X = \theta$ and $\mathbb{V}X = 4\theta^2$ so, by plugging in the expectation and the variance, it must be true that the following quantity has approximately a standard normal distribution

$$\sqrt{n}\frac{\bar{X} - \theta}{\sqrt{4\theta^2}} = \sqrt{n}\frac{\bar{X} - \theta}{2\theta} = \sqrt{n}\frac{\bar{X}/\theta - 1}{2} = T,$$

thus proving the claim.

9 pts (b) Use the near-pivot T to derive a two-sided confidence interval of level (approximately) 0.95 for θ. (To answer this question you may need one or more of the following quantiles: $z_{0.01} = -2.33$, $z_{0.0125} = -2.24$, $z_{0.025} = -1.96$, $z_{0.05} = -1.64$.)

Solution: Since T is a near-pivot and is approximately standard normal distributed, we know that because $z_{0.975} = -z_{1-0.975} = -z_{0.025} = 1.96$,

$$\begin{aligned}0.95 &\approx \mathbb{P}\left(-z_{0.975} \le T \le z_{0.975}\right) = \mathbb{P}\left(-z_{0.975} \le \sqrt{n}\frac{\bar{X}/\theta - 1}{2} \le z_{0.975}\right) \\ &= \mathbb{P}\left(1 - \frac{2z_{0.975}}{\sqrt{n}} \le \bar{X}/\theta \le 1 + \frac{z_{0.975}}{\sqrt{n}}\right) = \mathbb{P}\left(\frac{\bar{X}\sqrt{n}}{\sqrt{n} + 2z_{0.975}} \le \theta \le \frac{\bar{X}\sqrt{n}}{\sqrt{n} - 2z_{0.975}}\right),\end{aligned}$$

which leads to a confidence interval for p of level (approximately) 0.95:

$$\left[\frac{\bar{X}\sqrt{n}}{\sqrt{n} + 2z_{0.975}}, \frac{\bar{X}\sqrt{n}}{\sqrt{n} - 2z_{0.975}}\right].$$

We can then plug in the quantile to get the confidence interval.

4 pts (c) Suppose that you are given a one sided, upper confidence interval of level (exactly) 0.95 of the form $[0, \bar{X}_n + s/\sqrt{n}]$ for θ, for some $s > 0$. Use the fact that $\mathbb{V}X = 4\theta^2$, to find a one sided, upper confidence interval of level (exactly) 0.95 for $\mathbb{V}X$.

Solution: We are told to suppose that for some $s > 0$, we have

$$\mathbb{P}\left(0 \leq \theta \leq \bar{X}_n + s/\sqrt{n}\right) = \mathbb{P}\left(\theta \leq \bar{X}_n + s/\sqrt{n}\right) = 0.95.$$

This immediately implies that

$$\mathbb{P}\left(\theta^2 \leq (\bar{X}_n + s/\sqrt{n})^2\right) = 0.95, \qquad \text{and so,} \qquad \mathbb{P}\left(4\theta^2 \leq 4(\bar{X}_n + s/\sqrt{n})^2\right) = 0.95,$$

which, since $\mathbb{V}X = 4\theta^2$, tells us that the following is a confidence interval of level (exactly) 0.95 for $\mathbb{V}X$:

$$\left[0,\ 4(\bar{X}_n + s/\sqrt{n})^2\right].$$

8 pts (d) Consider now a confidence interval of level exactly 0.95 for θ of the form $[\bar{X}_n - r/\sqrt{n}, \bar{X}_n + r/\sqrt{n}]$, for some $r > 0$. Suppose that you collected a sample of size 100 and you got a confidence interval of length 0.7. How much more data would you need to reduce the length of the confidence interval to **strictly less** than half of that lenght?

Solution: We are told that $[\bar{x}_{100} - r/\sqrt{100}, \bar{x}_{100} + r/\sqrt{100}]$ has length 0.7, but this is the same as saying that the difference between the upper and the lower bound of the interval is

$$\bar{x}_{100} + \frac{r}{10} - \left(\bar{x}_{100} - \frac{r}{10}\right) = 2\frac{r}{10} = \frac{r}{5} = 0.7,$$

so that we must have $r = 3.5$. The confidence interval must therefore be $[\bar{X}_n - 3.5/\sqrt{n}, \bar{X}_n + 3.5/\sqrt{n}]$. We want to find n such that the length of the confidence interval reduced to strictly less than half of what it is when $n = 100$ (i.e., so that it is strictly less than $0.7/2 = 0.35$), then we want n such that

$$\bar{x}_n + \frac{3.5}{\sqrt{n}} - \left(\bar{x}_n - \frac{3.5}{\sqrt{n}}\right) < 0.35 \Leftrightarrow 2\frac{3.5}{\sqrt{n}} < 0.35 \Leftrightarrow 20 < \sqrt{n} \Leftrightarrow n > 400,$$

so we conclude that we need to take n equal to at least 401.

Prob.III: Suppose that a company is preparing to launch a new model of their most popular mobile phone. The company would like to be able to claim that charging the new model from 40% charge to 80% charge takes less than 15 minutes. Let us denote by X the amount of time (in minutes) that it takes to perform such a charge.

The company models $X \sim \mathrm{Exp}(\theta)$, $\theta > 0$, so that $\mathbb{E}X = 1/\theta$. Since they want to claim that $\mathbb{E}X = 1/\theta < 15$ (i.e., expected charge time is less than 15 minutes), they actually want to test

$$H_0 : \theta = 1/15 \quad \text{against} \quad H_1 : \theta > 1/15.$$

If they reject the null hypothesis, then they can indeed say that they have data to support the claim that the average charging time is less than 15 minutes ($\Leftrightarrow \theta > 1/15$.)

The company collects a random sample $X_1, \dots, X_n$ of charging times and uses $T = X_{(1)}$ as a test statistic. They reject the null hypothesis if $T < C$ for some appropriate critical value $C > 0$.

6 pts (a) Show that the critical value $C = 15\, e_\alpha/n$ ensures that the probability of a type I error for the test that rejects H_0 if $T < C$ is exactly α.

Solution: We just want to check if the probability that $T < 15\, e_\alpha/n$ when $\theta = 1/15$ is indeed α. That probability is

$$\mathbb{P}_{1/15}(T < 15\, e_\alpha/n) = \mathbb{P}_{1/15}\left(n\frac{1}{15}T < e_\alpha\right) = F_1(e_\alpha) = \alpha,$$

where we use the fact that when the data comes from $\mathrm{Exp}(\theta)$, then $n\theta T \sim \mathrm{Exp}(1)$, so that in particular when the data comes from $\mathrm{Exp}(1/15)$, then $nT/15 \sim \mathrm{Exp}(1)$.

10 pts (b) Suppose that the company decided to go with the test with significance level 0.01. If indeed the average charging time is below 15 minutes but only by one minute, i.e., $\theta = 1/14$, then what is the power of the test? Comment on the power that you got.

Solution: We are told to consider the test of level 0.01 which means that we must take $C = 15e_{0.01}/n = 15 \times 0.0101/n = 0.1515/n$. We then reject at significance level 0.01 if $T < 0.1515/n$. The power of a test is the probability of rejecting the null hypothesis under the assumption that θ falls under the alternative, such as is the case when $\theta = 1/14 > 1/15$. So the power of the test when $\theta = 1/14$ is given by

$$\pi(1/14) = \mathbb{P}_{1/14}(T < 0.1515/n) = \mathbb{P}_{1/14}\left(n\frac{1}{14}T < \frac{0.1515}{14}\right) = F_1(0.1515/14),$$

where we again use the fact that when the data comes from $\mathrm{Exp}(\theta)$, then $n\theta T \sim \mathrm{Exp}(1)$, so that in particular when the data comes from $\mathrm{Exp}(1/14)$, then $nT/14 \sim \mathrm{Exp}(1)$. We can easily compute the above since $F_1(x) = 1 - \exp(-x)$, so that

$$\pi(1/14) = F_1(0.1515/14) = 1 - \exp(-0.1515/14) = 0.0108.$$

This is barely above the significance level 0.01, so this means that it is difficult for this test to reject the null when the expected changing time is just one minute less than 15 minutes.

8 pts (c) An experiment was conducted where $n = 40$ *chargings* were conducted and the value $t = 0.011$ for the test statistic was recorded. Compute the p-value of this test. Would you reject the null hypothesis at significance level $\alpha = 0.05$?

Solution: The p-value of a test is the smallest significance level for which the null hypothesis is rejected. We reject the null at significance level α if $t = 0.011 < 15\, e_\alpha/40$, where we take $n = 40$. Remembering that $F_1(e_\alpha) = \alpha$, we see that we reject if

$$0.011 < 15\, e_\alpha/40 \Leftrightarrow 0.011\frac{40}{15} < e_\alpha \Leftrightarrow F_1\left(0.011\frac{40}{15}\right) < \alpha.$$

So if we reject whenever α is above $F_1\,(0.011 \times 40/15) = 1 - \exp(-0.011 \times 40/15) = 1 - \exp(-0.0293) = 0.02891$, but then the p-value is just, by definition, 0.0293. Since the p-value is less than 0.05, we would reject the null at significance level 0.05.

Hints: If $X \sim \text{Exp}(\theta)$, $\theta > 0$, then you are reminded that for $x > 0$,

$$f_\theta(x) = F'_\theta(x) = \theta e^{-\theta x} \qquad \text{and} \qquad F_\theta(x) = \mathbb{P}_\theta(X \le x) = 1 - e^{-\theta x},$$

so that $\mathbb{E}X = 1/\theta$. Also remember that $n\theta X_{(1)} \sim \text{Exp}(1)$, where $X_{(1)} = \min\{X_1, \ldots, X_n\}$.

You may also need one or more of the following quantiles, $e_{0.01} = 0.0101$, $e_{0.05} = 0.0513$, $e_{0.95} = 2.9957$, $e_{0.99} = 4.6052$, each of which has the property that $F_1(e_\alpha) = \alpha$.

Prob.IV: Suppose that you work at the marketing department of a company and that you have been tasked with analysing the effectiveness of several of the company's email based ad campaigns in the last few years. To perform this task, you are given the data in Table 1.

In the pairs (x_i, y_i), $i = 1, \ldots, 12$: the x_i represents the amount of money (in thousands of euro) invested in the campaign, and y_i represents the respective revenue (in thousands of euro) during the period that the campaign ran. (Both quantities represent daily averages.)

Variables	Values
$(x_1, \ldots, x_n)$	1.231 1.555 1.976 1.161 1.476 1.540 1.712 1.428 2.095 1.458 1.625 1.795
$(y_1, \ldots, y_n)$	11.076 11.174 11.666 10.849 11.326 11.273 11.457 11.257 11.757 11.109 11.472 11.593

Table 1: Daily investment and corresponding average daily revenue for 12 ad campaigns.

From the observations in Table 1 we see that $n\bar{x} = 19.0533$, $n\bar{y} = 136.0100$, $SS_{xx} = 0.8360$, $SS_{yy} = 0.7775$, and $SS_{xy} = 0.7670$. There are $n = 12$ measurements in total.

Figure 1, where we plot the data, seems to suggest that the relation between invested money and revenue might be linear.

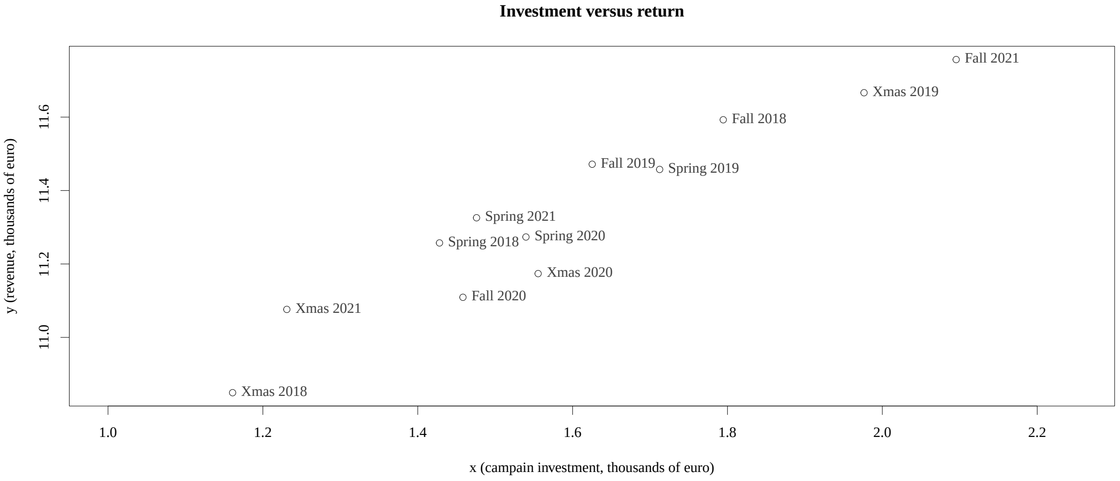

Figure 1: Daily investment and corresponding daily revenue for 12 ad campaigns.

6 pts (a) Suppose that you would like to use the Simple Linear Regression model to derive a formula that allows you to model the relation between investment (x) and the corresponding revenue (Y). In a linear regression model you assume that

$$Y_i = \alpha + \beta\, x_i + \sigma\epsilon_i, \quad i = 1, \ldots, n,$$

where $\alpha, \beta, \sigma \in \mathbb{R}$ are unknown, and the ϵ_i are random error terms. In order for Simple Linear Regression to be an adequate model here, what should you assume about: i) the relation between

(x_i, Y_i) and (x_j, Y_j), $i \neq j$; ii) the expectation of the noise terms ϵ_i; iii) the variance of the noise terms ϵ_i.

Solution: i) These should be independent; ii) the ϵ_i should have expectation 0; iii) the variance of the ϵ_i should be 1.

6 pts (b) Consider the data from Table 1 and suppose that the assumptions from (a) hold. Based on the data, what are your estimates of the intercept and the slope of the line in your model?

Solution: We have that $\hat{\beta} = S_{xy}/SS_{xx} = 0.7670/0.8360 = 0.9175$, **and** $\hat{\alpha} = \bar{y} - \bar{x} \times SS_{xY}/SS_{xx} = \bar{y} - \bar{x} \times \hat{\beta} = 136.0100/12 - 19.0533/12 \times 0.9175 = 9.8774$.

6 pts (c) (If you do not manage to compute the estimates, use the prediction formula $\hat{y} = 12 + 0.8\,x$.) What is the interpretation of the intercept? and of the slope? Assuming that these estimates are fairly accurate, do you think that the campaigns are profitable for the company?

Solution: The intercept tells us the expected revenue (under the SLR model) if no money is invested , while the slope tells us for every (thousands of) euro invested by how many (thousands of) euro the expected profit rises . (The value of $\hat{\alpha}$ (as a proxy for α) does not tell us anything about profitability of the ad campaigns, only how much we would be profiting without them.) The value of $\hat{\beta}$ (which we are told to assume is an accurate estimate of β) tells us that for each euro we invest we only get back about less than 0.92 euro back in revenue. This suggests that these ad campaigns are not working as far as increasing revenue and, therefore, do not seem to be profitable.

6 pts (d) Estimate the variance of the noise σ^2, and the coefficient of determination R^2 under the SLR modelling assumption.

Solution: The estimator for the variance of the noise is $\hat{\sigma}^2 = SS_{yy}/n - \hat{\beta}^2 SS_{xx}/n = 0.7775/12 - (0.9175)^2 \times 0.8360/12 = 0.0061$. **As for the coefficient of determination,**

$$R^2 = \frac{SS_{TOT} - SS_{RES}}{SS_{TOT}} = \frac{SS_{yy} - n\hat{\sigma}^2}{SS_{yy}} = (0.7775 - 12 \times 0.0061)/0.7775 = 0.9059.$$