X_400004 - Statistics
Solutions to the Final

18 December 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: Suppose that NS is simplifying the schedules for its Sprinters: rather than sprinters arriving at specific times, they now just arrive at random times. Denote by X the amount of time (in minutes) that it takes for a sprinter to come since the moment you arrive at the train stop. You model $X \sim \text{Exp}(\theta)$, $\theta > 0$, so that $\mathbb{E}X = 1/\theta$, where θ depends on the specific stop.

For your train stop, the NS is claiming that you should expect to wait (on average) at most 15 minutes for a sprinter but you would like to test the validity of this out. The claim is that $\mathbb{E}X = 1/\theta \leq 15$ (i.e., expected waiting time is at most 15 minutes), so you want to test

$$H_0 : \theta \geq 1/15 \quad \text{against} \quad H_1 : \theta < 1/15.$$

If you reject the null hypothesis, then you can conclude that you have data to support the claim that the average waiting time is more than 15 minutes ($\Leftrightarrow \theta < 1/15$.)

You collect a random sample $X_1, \ldots, X_n$ of waiting times and use $T = X_{(1)}$ as a test statistic. You reject the null hypothesis if $T > C$ for some appropriate critical value $C > 0$.

4 pts (a) Show that $n\theta X_{(1)} \sim \text{Exp}(1)$, where $X_{(1)} = \min\{X_1, \ldots, X_n\}$.

Solution: It suffices to show that $\mathbb{P}(n\theta X_{(1)} > x) = e^{-x}$ which is 1 minus the CDF of a $\text{Exp}(1)$ random variable. This probability is by definition

$$\mathbb{P}\Big(\min\{X_1, \ldots, X_n\} > \frac{x}{n\theta}\Big) = \mathbb{P}\Big(X_1 > \frac{x}{n\theta}, \ldots, X_n > \frac{x}{n\theta}\Big) = \mathbb{P}\Big(X_1 > \frac{x}{n\theta}\Big) \times \cdots \times \mathbb{P}\Big(X_n > \frac{x}{n\theta}\Big).$$

Since each $X_i \sim \text{Exp}(\theta)$ product is then

$$e^{-\theta\frac{x}{n\theta}} \times \cdots \times e^{-\theta\frac{x}{n\theta}} = e^{-\frac{x}{n}} \times \cdots \times e^{-\frac{x}{n}} = e^{-n\frac{x}{n}} = e^{-x}.$$

8 pts (b) Show that the critical value $C = 15\, e_{1-\alpha}/n$ ensures that the test that rejects H_0 if $T > C$ has significance level exactly α. (See definition of e_α below among the hints.)

Solution: We want to check that the probability that $T > 15\, e_{1-\alpha}/n$ (rejection) for any $\theta \geq 1/15$ (under the null) is at most α. That probability is by (a)

$$\mathbb{P}_\theta\big(T > 15\, e_{1-\alpha}/n\big) = \mathbb{P}_\theta\left(n\theta X_{(1)} > 15\theta e_{1-\alpha}\right) = 1 - F_1(15\theta e_{1-\alpha}) = e^{-15\theta e_{1-\alpha}}.$$

Since $\theta \geq 1/15$ and the above probability decreases with θ, we conclude that if $\theta \geq 1/15$, then $15\theta \geq 1$ and so

$$\mathbb{P}_\theta\big(T > 15\, e_{1-\alpha}/n\big) \leq e^{-e_{1-\alpha}} = 1 - F_1(e_{1-\alpha}) = 1 - (1-\alpha) = \alpha.$$

where we use the definition of the quantile.

8 pts (c) Suppose that you decided to go with the test with significance level 0.01. If indeed the average waiting time is above 15 minutes but only by one minute, i.e., $\theta = 1/16$, then what is the power of the test when $n = 20$? Interprete the power that you got.

Solution: We are told to consider the test of level 0.01 and a sample of size 20 which means that we must take $C = 15e_{0.99}/20 = 15 \times 4.6052/20 = 3.4539$. We then reject at significance level 0.01 if $T > 3.4539$. 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 $\theta = 1/16 < 1/15$. So the power of the test when $\theta = 1/16$ is given by

$$\pi(1/16) = \mathbb{P}_{1/16}(T > 3.4539) = \mathbb{P}_{1/16}\left(20\frac{1}{16}X_{(1)} > 20\frac{1}{16}3.4539\right) = 1 - F_1(2.1587),$$

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 $\mathrm{Exp}(1/16)$, then $nT/16 \sim \mathrm{Exp}(1)$. We can easily compute the above since $F_1(x) = 1 - \exp(-x)$, so that

$$\pi(1/16) = 1 - F_1(2.1587) = \exp(-2.1587) = 0.1155.$$

This is not so large so it tells us that with a sample of size 20 if the expected waiting time is not 15 minutes or less but instead 16 minutes, then this test will have a hart time detecting it.

6 pts (d) Suppose that, still in the case where $n = 20$, the test statistic took the value $t = 0.011$. Compute the p-value. 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 $t = 0.011 > 15\, e_{1-\alpha}/20$, where we took $n = 20$. Remembering that, by definition, $F_1(e_\alpha) = \alpha$, we see that we reject if

$$0.011 > 15\, e_{1-\alpha}/20 \Leftrightarrow 0.011\frac{20}{15} > e_{1-\alpha} \Leftrightarrow F_1\left(0.01467\right) > 1-\alpha. \Leftrightarrow \alpha > 1-(1-e^{-0.01467}) = 0.9854$$

If we reject whenever α is above 0.9854 then the p-value is just, by definition, 0.9854. Since the p-value is more than 0.05, we would not 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 $\mathbb{E}X = 1/\theta$.

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.II: Suppose that you get a random sample $X_1, \ldots, X_n \ge 0$. You are told that $\mathbb{E}X = \gamma$ and $\mathbb{E}(X^2) = 5\gamma^2$, for some unknown $\gamma > 0$, and that the Central Limit theorem may be applied to these data.

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

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

is close to being $N(0,1)$ so that T is a near-pivot for γ. (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 $X_1, \ldots, X_n$ 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 $\mathbb{E}X = \gamma$ and $\mathbb{V}X = \mathbb{E}(X^2) - (\mathbb{E}X)^2 = 5\gamma^2 - \gamma^2 = 4\gamma^2$ so, by plugging in the expectation and the variance, it must be true that the following quantity has approximately a standard normal distribution

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

thus proving the claim.

10 pts (b) Use the near-pivot T to derive a two-sided confidence interval of level (approximately) 0.9 for γ. (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, and we know that $z_{0.95} = -z_{1-0.95} = -z_{0.05} = 1.64$, then

$$0.9 \approx \mathbb{P}\left(z_{0.05} \le T \le z_{0.95}\right) = \mathbb{P}\left(-z_{0.95} \le \sqrt{n}\frac{\bar{X}/\gamma - 1}{2} \le z_{0.95}\right)$$

$$= \mathbb{P}\left(1 - \frac{2z_{0.95}}{\sqrt{n}} \le \bar{X}/\gamma \le 1 + \frac{2z_{0.95}}{\sqrt{n}}\right) = \mathbb{P}\left(\frac{\bar{X}\sqrt{n}}{\sqrt{n} + 2z_{0.95}} \le \gamma \le \frac{\bar{X}\sqrt{n}}{\sqrt{n} - 2z_{0.95}}\right),$$

which leads to a confidence interval for γ of level (approximately) 0.9:

$$\left[\frac{\bar{X}\sqrt{n}}{\sqrt{n} + 2z_{0.95}}, \frac{\bar{X}\sqrt{n}}{\sqrt{n} - 2z_{0.95}}\right].$$

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

6 pts (c) Suppose that you are given a one-sided, upper confidence interval of level (exactly) 0.9 of the form $[0, \bar{X}_n + s/\sqrt{n}]$ for γ, for some $s > 0$. Express $\mathbb{V}X$ as a function of γ and use that to find a one-sided, upper confidence interval of level (exactly) 0.9 for $\mathbb{V}X$.

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

$$\mathbb{P}\left(0 \leq \gamma \leq \bar{X}_n + s/\sqrt{n}\right) = \mathbb{P}\left(\gamma \leq \bar{X}_n + s/\sqrt{n}\right) = 0.9.$$

This immediately implies that

$$\mathbb{P}\left(\gamma^2 \leq (\bar{X}_n + s/\sqrt{n})^2\right) = 0.9, \qquad \textbf{and so,} \qquad \mathbb{P}\left(4\gamma^2 \leq 4(\bar{X}_n + s/\sqrt{n})^2\right) = 0.9,$$

which, since $\mathbb{V}X = 4\gamma^2$, tells us that the following is a confidence interval of level (exactly) 0.9 for $\mathbb{V}X$:

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

10 pts (d) Consider now a confidence interval of level exactly 0.95 for γ of the form $[\bar{X}_n - r/\sqrt{n}, \bar{X}_n + r/\sqrt{n}]$, for some constant $r > 0$. 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: A company is considering integrating AI into some of their workflows to replace external consultants but a transition would be costly so they want to make sure that there are benefits to the change. In a pilot study, 20 new projects are individually handled by external consultants and, in parallel, those same 20 projects are handled internally by a team using instead an AI tool.

In the end, a separate team assesses, for each project the outcomes of the two approaches (without knowing which is which) and gives a score. The data are summarised in Figure 1.

In the pairs (x_i, y_i), $i = 1, \ldots, 20$: the x_i represents the score of the external consultant approach and y_i represents the score of the AI approach. (Higher score means better approach.) A few numerical summaries for the data: $n\bar{x} = 160.841$, $n\bar{y} = 148.265$, $SS_{xx} = 28.16$, $SS_{yy} = 26.407$, and $SS_{xy} = 24.578$. The size of the sample is $n = 20$.

To answer the questions below you may need one or more of the following quantiles: $t_{18,0.01} = -2.55238$, $t_{18,0.0125} = -2.445006$, $t_{18,0.025} = -2.100922$, $t_{18,0.05} = -1.734064$.

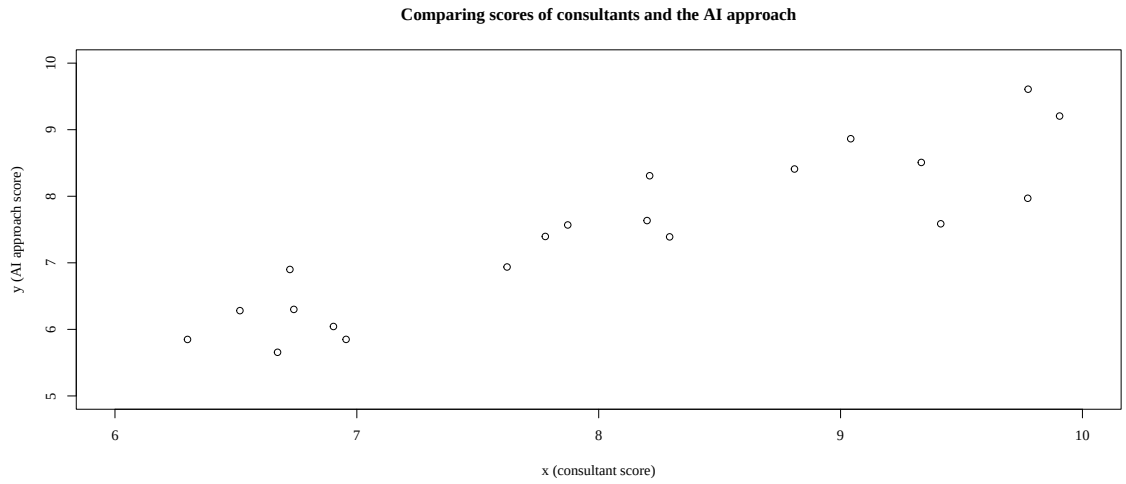

Figure 1: Score comparison for the two different approaches. (Higher score is better.)

8 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 the score for the consultant (X) and the corresponding score for the AI approach (Y). In a Simple Linear Regression (SLR) 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. State if the following **must** be true in order for SLR to be an adequate model here: i) the (X_i, Y_i) need to be i.i.d., $i = 1, \ldots, n$; ii) the expectation of the noise terms ϵ_i is zero; iii) the standard deviation of the noise terms ϵ_i is 1. (If you say a statement is false, then present the correct statement.)

Solution: i) False : these should be independent but don't have to be i.i.d.; ii) True; iii) True.

4 pts (b) Consider the data from Figure 1 and suppose that the SLR model is adequate. (i) Based on the data, what are your estimates for α and β, the parameters of the model? (ii) What insight do the estimates of α and β give you?

Solution: (i) We have that $\hat{\beta} = S_{xy}/SS_{xx} = 24.578/28.16 \approx 0.8727983$, **and** $\hat{\alpha} = \bar{y} - \bar{x} \times SS_{xY}/SS_{xx} = \bar{y} - \bar{x} \times \hat{\beta} = 148.265/20 - 160.841/20 \times 0.8727983 \approx 0.3941$. **The interpretation of the estimate of** α **is that on average, an AI scores about** 0.3941 **when the consultant scores** 0, **and the interpretation of** β **is that every extra point that a consultant scores, there is approximately and extra** 0.872 **points that the AI approach scores.**

6 pts (c) 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 $\hat{\sigma}^2 = SS_{yy}/n - \hat{\beta}^2 SS_{xx}/n = 26.407/20 - (0.8728)^2 \times 28.16/20 \approx 0.2478$. **As for the coefficient of determination,**

$$R^2 = \frac{SS_{TOT} - SS_{RES}}{SS_{TOT}} = \frac{SS_{yy} - n\hat{\sigma}^2}{SS_{yy}} = (26.407 - 20 \times 0.2478)/26.407 \approx 0.8123.$$

8 pts (d) It seems quite important to test if we can conclude if $\beta > 1$. Test $H_0 : \beta = 1$ against $H_1 : \beta > 1$ at significance level 0.05. What do you conclude from performing this test? You should use the fact that

$$\sqrt{SS_{xx}}\frac{\hat{\beta} - \beta}{\hat{\sigma}} \sim t_{n-2},$$

for any $\beta \in \mathbb{R}$, $2 < n \in \mathbb{N}$.

Solution: We can use $T = \hat{\beta}$ (or just the pivot above) as test statistic and should reject if $T > c^*$. Under the null we know that $\sqrt{SS_{xx}}(\hat{\beta} - 1)/\hat{\sigma} \sim t_{n-2}$ so for a test of level α we need

$$\alpha = \mathbb{P}_{\beta=1}(\hat{\beta} > c^*) = \mathbb{P}_{\beta=1}(\sqrt{SS_{xx}}(\hat{\beta} - 1)/\hat{\sigma} > \sqrt{SS_{xx}}(c^* - 1)/\hat{\sigma}) = 1 - F_{t_{n-2}}\left(\sqrt{SS_{xx}}(c^* - 1)/\hat{\sigma}\right).$$

This is the same as

$$F^{-1}_{t_{n-2}}(1 - \alpha) = \sqrt{SS_{xx}}(c^* - 1)/\hat{\sigma} \Leftrightarrow c^* = 1 + \hat{\sigma}F^{-1}_{t_{n-2}}(1 - \alpha)/\sqrt{SS_{xx}}.$$

From this we conclude that using the critical value $c^* = 1 + \hat{\sigma}t_{n-2;1-\alpha}/\sqrt{SS_{xx}}$ leads to a test of level α . Setting $\alpha = 0.05$ we get the critical value

$$c^* = 1 + \sqrt{0.2478} * (1.734064)/\sqrt{28.16} = 1.1627,$$

where we use the fact that by symmetry $t_{n-2;1-\alpha} = -t_{n-2;\alpha}$. Since $\hat{\beta} = 0.8728 < 1.1627$ we do not reject the null hypothesis at level 0.05. This means that based on the data, we cannot conclude that $\beta > 1$.

4 pts (e) Irrespectively of you answer to (d), suppose that you reject $H_0 : \beta = 1$ in favour of $H_1 : \beta > 1$ at significance level 0.05. Does that necessarily mean that you could conclude (at significance level 0.05) that one should prefer the AI approach?

Solution: In principle no. A possible reason why not, is that it could be that α is negative so that $\beta > 1$ just reflects that when a consultant scores higher, then the AI approach also scores higher but the consultants still score higher (on average.)