X_400004 - Statistics
Final

18 December 2023

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 4 pages in the exam questionnaire (including this one) and you have 2 hours (120 minutes) to complete the exam.
- The exam consists of 13 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.
- **Only hand in what needs grading** (so don't hand in scratch paper, your cheat-sheet.)
- **Number individual sheets that you hand in.** (For instance 1/3, 2/3, 3/3.)

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 $X_{(1)} = \min\{X_1, \ldots, X_n\}$.

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.)

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.

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 $\alpha = 0.05$?

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

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

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 $X_1, \ldots, X_n \geq 0$. 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.)

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: $z_{0.01} = -2.33$, $z_{0.0125} = -2.24$, $z_{0.025} = -1.96$, $z_{0.05} = -1.64$.)

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 $\mathbb{V}X$.

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?

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.

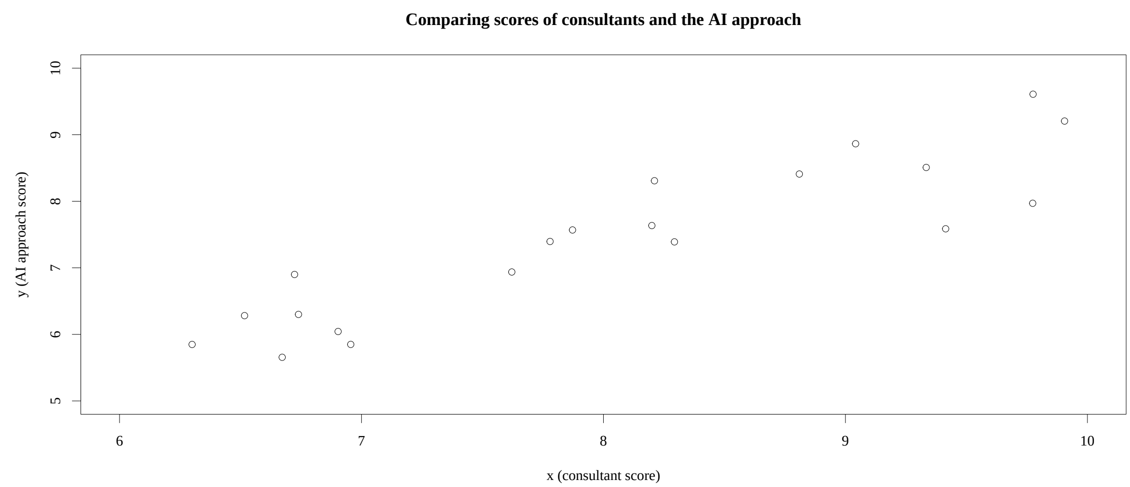

Figure 1: Score comparison for the two different approaches. (Higher score is better.)

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: $t_{18,0.01} = -2.55238$, $t_{18,0.0125} = -2.445006$, $t_{18,0.025} = -2.100922$, $t_{18,0.05} = -1.734064$.

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., $i = 1, \dots, n$; 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.)

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?

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

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

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?