X_400004 - Statistics
Solutions to the Final

20 December 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: Pivots are useful tools to build statistical tests but they are not always readily available. The Central Limit Theorem (CLT) provides another tool to build (approximate) pivots.

Consider a random sample $X_1, \ldots, X_n$ of *counts*. The Poisson distribution is usually a good model for counts, so assume that your observations are distributed like $X \sim \text{Poisson}(\theta)$, where $\theta > 0$ is some unknown model parameter. You are reminded that for $i = 0, 1, 2, \ldots,$

$$f_\theta(i) = \frac{\theta^i}{i!}e^{-\theta}, \qquad \text{and} \qquad F_\theta(i) = \mathbb{P}_\theta(X \leq i) = \sum_{j=0}^{i} \mathbb{P}(X = j),$$

are the probability mass function of X and the cumulative distribution function of X, respectively. Remember that this means that $\mathbb{E}_\theta X = \mathbb{V}_\theta X = \theta$.

8 pts (a) Let $X \sim \text{Poisson}(\alpha)$ and $Y \sim \text{Poisson}(\beta)$, two independent random variables and $\alpha, \beta > 0$. Use the fact that $\mathbb{P}(X+Y=i) = \sum_{j=0}^{i} \mathbb{P}(Y=j)\mathbb{P}(X=i-j)$ to show that $X+Y \sim \text{Poisson}(\alpha+\beta)$.
Hint: Remember the binomial theorem which says that $\sum_{j=0}^{i} \binom{i}{j} x^j y^{i-j} = (x+y)^i$.
Solution: Using the tip on how to compute $\mathbb{P}(X+Y=i)$ we see that it is

$$\sum_{j=0}^{i} \frac{\alpha^j}{j!}e^{-\alpha} \times \frac{\beta^{(i-j)}}{(i-j)!}e^{-\beta} = e^{-(\alpha+\beta)}\frac{\beta^i}{i!}\sum_{j=0}^{i}\binom{i}{j}\left(\frac{\alpha}{\beta}\right)^j \cdot 1^{i-j} = e^{-(\alpha+\beta)}\frac{\beta^i}{i!}\left(\frac{\alpha}{\beta}+1\right)^i = e^{-(\alpha+\beta)}\frac{(\alpha+\beta)^i}{i!},$$

which we recognise as the PMF of a $\text{Poisson}(\alpha+\beta)$. Above we just got everything that does not depend on j out of the sum, used the fact that $\binom{i}{j} = i!/\{j!(i-j)!\}$ and recognise that since $1 = 1^{i-j}$ we can apply the binomial theorem.

6 pts (b) From (a), $n\bar{X} = \sum_{i=1}^{n} X_i \sim \text{Poisson}(n\theta)$. Is $T = n\bar{X} - n\theta$ a pivot for θ? Justify your answer.
Solution: The quantity T is not a pivot. (It is a function of the data (and n) and θ, which it is allowed to be) but it is not a pivot because its distribution depends in θ which is an unknown parameter. You can demonstrate that the distribution of T depends on θ by noting that $\mathbb{V}(n\bar{X} - n\theta) = n\theta$, or that $\mathbb{P}(T = -n\theta) = \exp(-n\theta)$, etc..

6 pts (c) What does the CLT tell you about the distribution of $\bar{X}$? Be explicit and justify your answer.

Solution: Since $\bar{X}$ is the average of a random sample (with finite variance) we can apply the CLT to it. We have $\mathbb{E}_\theta \bar{X} = \theta$, and $\mathbb{V}_\theta \bar{X} = \theta/n$. The CLT says that the following has (approximately) a standard normal distribution:

$$\frac{\bar{X} - \mathbb{E}_\theta \bar{X}}{\sqrt{\mathbb{V}\bar{X}}} = \frac{\bar{X} - \theta}{\sqrt{\theta/n}} = \sqrt{n}\frac{\bar{X} - \theta}{\sqrt{\theta}} = \sqrt{n}\frac{\bar{X} - \mathbb{E}_\theta X}{\sqrt{\mathbb{V}X}}.$$

10 pts (d) Show that the test that rejects $H_0 : \theta = 1$ against $H_1 : \theta > 1$ when $\bar{X} > 1 + z_{1-\alpha}/\sqrt{n}$ has significance level approximately α for large n. Here, z_α represents the quantile of level $\alpha \in (0, 1)$ from standard Normal distribution.

Solution: We just have to check if $\mathbb{P}_{\theta=1}(\bar{X} > 1 + z_{1-\alpha}/\sqrt{n}) \approx \alpha$. This is true since if $Z \sim N(0, 1)$, then

$$\mathbb{P}_{\theta=1}\left(\bar{X} > 1 + \frac{z_{1-\alpha}}{\sqrt{n}}\right) = \mathbb{P}_{\theta=1}\left(\sqrt{n}\frac{\bar{X} - 1}{\sqrt{1}} > z_{1-\alpha}\right) \approx \mathbb{P}_{\theta=1}\left(Z > z_{1-\alpha}\right) = 1 - \Phi(z_{1-\alpha}) = 1 - (1 - \alpha) = \alpha.$$

The approximation gets better as n increases due to the CLT.

Prob.II: Suppose that you start a small company that wants to develop a more environmentally friendly alternative to styrofoam peanuts; cf. Figure 1. This material is used as a filler in packages to protect the contents from impact.

Figure 1: Example of styrofoam peanuts.

You are interested in testing the compressibility of the material: you want this material to be compressible to absorb impact but not too compressible otherwise you need to use a lot of it per package.

You have arrived at what you think is a good product and run some tests to measure the compressibility of the material when about 25% of the package consists of peanuts. At this stage you want to do some quality control. You collect some compressibility data $X_1, \ldots, X_n$ which you model as a random sample distributed like $X \sim N(\mu, \sigma^2)$. (Units of each observation is megapascal, MPa.)

4 pts (a) The following hypothesis are to be tested,

$$H_0 : \mu = 33\,\text{MPa}, \qquad \text{vs} \qquad H_1 : \mu \neq 33\,\text{MPa}.$$

Explain the (i) meaning of picking these specific null/alternative, and (ii) what conclusions can be drawn from a statistical test for them.

Solution: (i) Under the null, the expectation μ is exactly 33MPa and, under the alternative, μ is either more- or less than 33MPa. (ii) A rejection of the null means that the data suggests that the compressibility of the peanuts is *not* the target 33MPa, and no rejection means that the data does not provide evidence supporting that the compressibility is not 33MPa.

8 pts (b) Suppose that you use the following rule for rejection: $|\bar{X} - 33| > c$, for some appropriate c. (i) Justify why this is appropriate as a rejection rule, and (ii) what is the distribution of the test statistic $\bar{X}$ under the null if σ^2 were known?

Solution: (i) this is appropriate since under the null we expect $|\bar{X} - 33|$ to be close to 0, and under the alternative we expect $|\bar{X} - 33|$ to be large, so rejecting when you see an appropriately large value of $|\bar{X} - 33|$ is reasonable. (ii) Under the null, $\bar{X} \sim N(33, \sigma^2/n)$.

12 pts (c) It is not realistic to assume that you know σ^2 so you cannot proceed using the test statistic from (b). Instead, use the pivot $|\bar{X} - 33|/S_n$, where S_n is some appropriate estimator for σ. Compute the probability of rejection using $|\bar{X} - 33|/S_n > c$ under the null **exactly** without using knowledge of σ^2. (This probability will depend on c and eventually other known quantities.)

Solution: The relevant fact to be able to compute the probability is that under the null

$$\sqrt{n}\frac{\bar{X} - 33}{S_n} \sim t_{n-1}, \quad \textbf{where} \quad S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

We reject if $|\bar{X} - 33|/S_n > c$, so

$$\begin{aligned}\alpha(33, c) &= \mathbb{P}_{\mu=33}(|\bar{X} - 33|/S_n > c) = 2\mathbb{P}_{\mu=33}((\bar{X} - 33)/S_n > c) \\ &= 2\left\{1 - \mathbb{P}_{\mu=33}\left(\sqrt{n}\frac{\bar{X} - 33}{S_n} \leq c\sqrt{n}\right)\right\} = 2\left\{1 - F_{t_n-1}(c\sqrt{n})\right\},\end{aligned}$$

where F_{t_n-1} represents the (symmetric) CDF of a t_{n-1} distribution.

8 pts (d) Show that setting c to the specific choice $c_\alpha = F^{-1}(1 - \alpha/2)/\sqrt{n}$ gives you a test of level α where F is an appropriate CDF. (Be explicit about what F is.)

Solution: We just have to check that $\alpha(33, c_\alpha) \leq \alpha$. From before, if we take $F = F_{t_{n-1}}$, the CDF of a t_{n-1}, then

$$\begin{aligned}\alpha(33, c_\alpha) &= 2\left\{1 - F_{t_n-1}(\sqrt{n}c_\alpha)\right\} = 2\left\{1 - F_{t_n-1}(\sqrt{n}F^{-1}_{t_{n-1}}(1 - \alpha/2)/\sqrt{n})\right\} \\ &= 2\left\{1 - F_{t_n-1}(F^{-1}_{t_{n-1}}(1 - \alpha/2))\right\} = 2\left\{1 - (1 - \alpha/2))\right\} = \alpha.\end{aligned}$$

8 pts (e) Define the power $\pi(\mu)$ of the test of level α from (d). (You do not need to compute the underlying probability but make your expression explicit.) What is the value of $\pi(33)$ and why?

Solution: The power of the test is the probability of rejecting under the alternative as a function of μ, i.e.,

$$\pi(\mu) = \mathbb{P}_\mu\left(|\bar{X} - 33|/S_n > F^{-1}_{t_{n-1}}(1 - \alpha/2)/\sqrt{n}\right),$$

When $\mu = 33$ we are under the null, so $\pi(33)$ is just the probability of a type I error. Since the test is calibrated to have level α, we have $\pi(33) \leq \alpha$ but in this particular case we actually have $\pi(33) = \alpha$.

Prob.III: Suppose that you work at the HR department of a company and that you have been tasked with analysing the effectiveness of a specific training that some workers at the company follow. To perform this task, you are given the data in Figure 2.

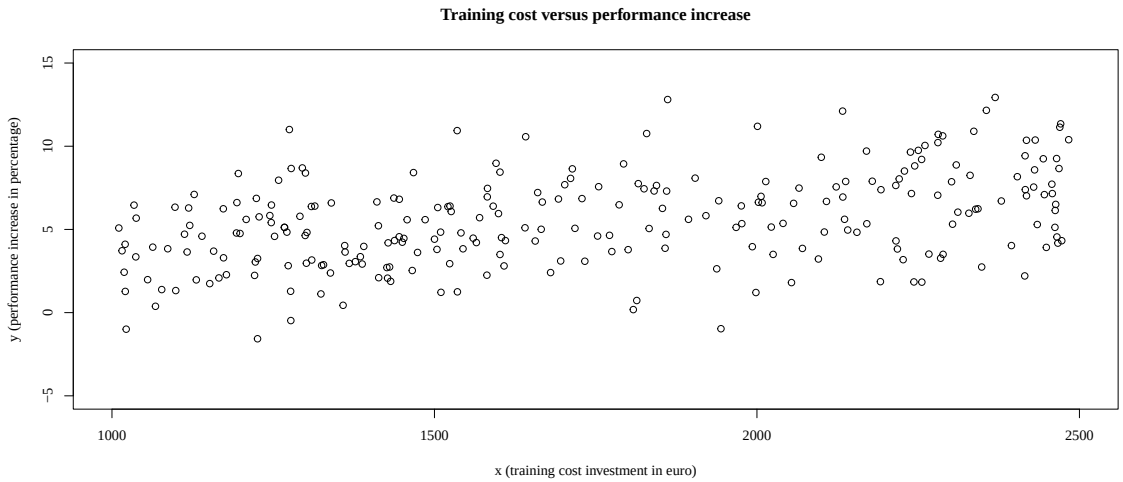

Figure 2: Training cost and respective performance increase for 256 persons.

In the pairs (x_i, y_i), $i = 1, \ldots, 256$: the x_i represents the cost of a training (in euro) for an employee, and y_i represents respective perceived gain in productivity for that same employee. A few numerical summaries for the data: $n\bar{x} = 445574$, $n\bar{y} = 1417.917$, $SS_{xx} = 52451946$, $SS_{yy} = 1944.043$, and $SS_{xy} = 136344.9$. There are $n = 256$ measurements in total.

To answer the questions below you may need one or more of the following quantiles: $t_{254,0.01} = -2.341118$, $t_{254,0.0125} = -2.254768$, $t_{254,0.025} = -1.969348$, $t_{254,0.05} = -1.650875$.

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 performance gain (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.

8 pts (b) Consider the data from Figure 2 and suppose that the assumptions from (a) hold. (i) Based on

the data, what are your estimates for α and β, the parameters of the model? (ii) What insight does the estimate of β give you?

Solution: (i) We have that $\hat{\beta} = S_{xy}/SS_{xx} = 136344.9/52451946 \approx 0.0026$, and $\hat{\alpha} = \bar{y} - \bar{x} \times SS_{xY}/SS_{xx} = \bar{y} - \bar{x} \times \hat{\beta} = 1417.917/256 - 445574/256 \times 0.0026 \approx 1.0144$. The interpretation of the estimate of β is that every euro invested in training seems to result, on average, in an increase of about 0.0026% in productivity, so quite a small increase.

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 $\hat{\sigma}^2 = SS_{yy}/n - \hat{\beta}^2 SS_{xx}/n = 1944.043/256 - (0.0026)^2 \times 52451946/256 \approx 6.2095$. As for the coefficient of determination,

$$R^2 = \frac{SS_{TOT} - SS_{RES}}{SS_{TOT}} = \frac{SS_{yy} - n\hat{\sigma}^2}{SS_{yy}} = (1944.043 - 256 \times 6.2095)/1944.043 \approx 0.1823.$$

10 pts (d) It seems quite important to test if $\beta = 0$. Test if $H_0 : \beta = 0$ against $H_1 : \beta \neq 0$ at significance level 0.05. What do you conclude in terms of the test and what is the implication in the context of the SLR model? You should use the fact that

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

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 $\sqrt{SS_{xx}}(\hat{\beta} - 0)/\hat{\sigma} \sim t_{n-2}$ so for a test of level α we need

$$\alpha = \mathbb{P}_{\beta=0}(|\hat{\beta}| > c^*) = \mathbb{P}_{\beta=0}(|\sqrt{SS_{xx}}\hat{\beta}/\hat{\sigma}| > \sqrt{SS_{xx}}c^*/\hat{\sigma}).$$

Using that $F_{t_{n-2}}(z) = 1 - F_{t_{n-2}}(-z)$, this is the same as

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

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

$$c^* = -6.2095 * (-1.969348) * \sqrt{52451946} = 0.001688.$$

Since $|\hat{\beta}| = 0.0026 > 0.001688$ we reject the null hypothesis at level 0.05. This means that based on the data, we reject the hypothesis that $\beta = 0$; however, the amount of investment seems to have a very small influence on performance.