# X_400004 - Statistics
# Final

20 December 2022

**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.

**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 $X_1, \dots, X_n$ 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 $i = 0, 1, 2, \dots,$

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

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 $\sum_{j=0}^{i} \binom{i}{j} x^j y^{i-j} = (x+y)^i$.

**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 $\theta$? Justify your answer.

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

**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 $\alpha$ for large $n$. Here, $z_\alpha$ represents the quantile of level $\alpha \in (0,1)$ from standard Normal distribution.

**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 $X_1, \dots, X_n$ 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.

**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 $\sigma^2$ were known?

**12 pts** (c) It is not realistic to assume that you know $\sigma^2$ 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 $\sigma$. Compute the probability of rejection using $|\bar{X} - 33|/S_n > c$ under the null **exactly** without using knowledge of $\sigma^2$. (This probability will depend on $c$ and eventually other known quantities.)

**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 $\alpha$ where $F$ is an appropriate CDF. (Be explicit about what $F$ is.)

**8 pts** (e) Define the power $\pi(\mu)$ of the test of level $\alpha$ 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?

**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.

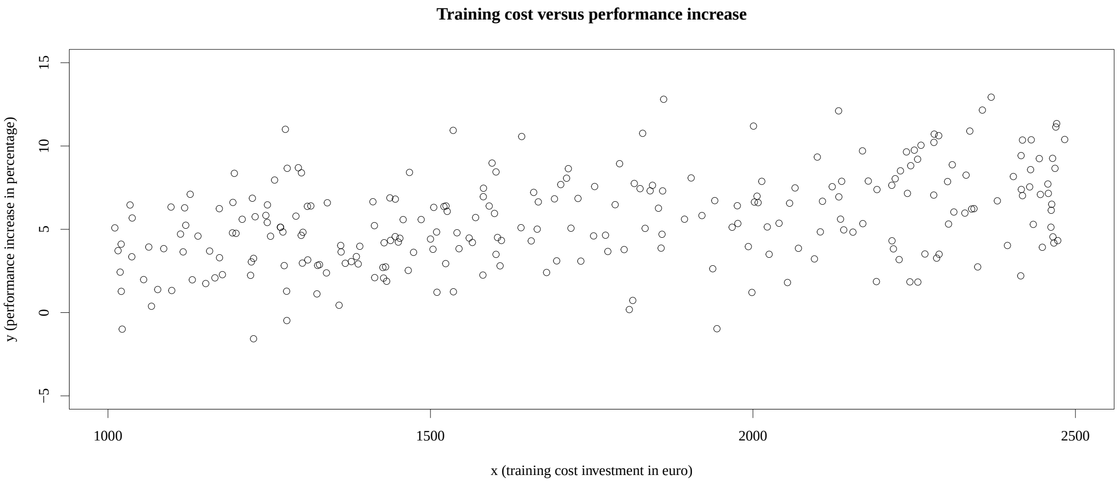

Figure 2: Training cost and respective performance increase for 256 persons.

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

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

where $\alpha, \beta, \sigma \in \mathbb{R}$ are unknown, and the $\epsilon_i$ 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 $\epsilon_i$; iii) the variance of the noise terms $\epsilon_i$.

**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 $\alpha$ and $\beta$, the parameters of the model? (ii) What insight does the estimate of $\beta$ give you?

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

**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

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