# X_400004 - Statistics
# Final

21 December 2021

**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:** Nobody likes waiting, but constantly checking to see if something is ready is also no fun... It might therefore be useful to construct upper confidence interval for expected waiting times. In this question you are asked to do just this.

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

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

are the probability density function of $X$ and the cumulative distribution function of $X$, respectively. Note that with this parametrisation, the expected waiting time is $\mathbb{E}X = 1/\theta$.

**8 pts** (a) Use the fact that for any random sample $Y_1, \ldots, Y_n$ distributed like $Y$ you have that

$$\mathbb{P}(Y_{(1)} > y) = \mathbb{P}(Y_1 > y, \ldots, Y_n > y) = \mathbb{P}(Y_1 > y) \times \cdots \times \mathbb{P}(Y_n > y) = \{1 - \mathbb{P}(Y \leq y)\}^n,$$

to show that $T = n\theta X_{(1)} \sim \text{Exp}(1)$. (Remember that $X_{(1)}$ is just shorthand notation for the sample minimum, i.e., $X_{(1)} = \min\{X_1, \ldots, X_n\}$.)

**4 pts** (b) Justify why $T$ is a pivot for $\theta$.

**8 pts** (c) Let $e_\alpha$, $\alpha \in [0, 1]$, represent the quantile of level $\alpha$ of an Exp(1) distribution so that, by definition, $F_1(e_\alpha) = \alpha$. Prove that the interval $\left[0, nX_{(1)}/e_\alpha\right]$ is a $(1 - \alpha) \times 100\%$ upper confidence interval for the **expected waiting time**.

**8 pts** (d) Suppose that you have collected the sample $\{1.12, 0.05, 1.07, 0.82\}$ of waiting times. Use the confidence interval from (c) to test at significance level 0.05 the hypotheses: $H_0 : 1/\theta = 2$ vs $H_0 : 1/\theta > 2$. (To answer this question you may need one or more of the following quantiles: $e_{0.01} = 0.0101$, $e_{0.05} = 0.0513$, $e_{0.95} = 2.9957$, $e_{0.99} = 4.6052$.)

**Prob.II:** Suppose that you run a small company and that, right now, when there is an incident at work you deal with it personally. You do enjoy directly helping your staff but this can be quite time consuming...

Suppose that you model the weekly number of incidents at work using the Poisson distribution. After $n$ weeks, you have collected a dataset $X_1, \ldots, X_n$ which you model as a random sample distributed like $X \sim \text{Poisson}(\theta)$, $\theta > 0$. Remember that this means that $\mathbb{E}_\theta X = \mathbb{V}_\theta X = \theta$, and that the probability mass function of $X$ is

$$f_\theta(x) = \mathbb{P}_\theta(X = x) = \frac{\theta^x}{x!} e^{-\theta}, \quad x = 0, 1, 2, \ldots.$$

Suppose that your patience breaks down if you have to deal with more than 5 incidents, on average, per week. To decide, in a data-driven way, if indeed there are on average more than 5 incidents per week, you will test the hypotheses:

$$H_0 : \theta = 5, \quad \text{vs} \quad H_1 : \theta > 5.$$

(Remember that $\theta$ is just the expected waiting time, $\mathbb{E}_\theta X$.) If you reject the null hypothesis (and thus accept the alternative), you will go ahead and will start delegating this work to someone else.

Consider, throughout, the following statistical test: you reject the null hypothesis $H_0$ if, for some appropriately chosen $C$,

$$T > C, \qquad \text{where} \qquad T = \sum_{i=1}^{n} X_i.$$

(To answer the following questions you may need one or more of the following quantiles: $z_{0.01} = -2.33$, $z_{0.0125} = -2.24$, $z_{0.025} = -1.96$, $z_{0.05} = -1.64$.)

**4 pts** (a) What is the distribution of the test statistic $T$ under our modelling assumption? and what is the distribution of $T$ specifically under the null? **Hint:** use the fact that if $X \sim \text{Poisson}(\theta_1)$, $Y \sim \text{Poisson}(\theta_2)$, and $X$ and $Y$ are independent, then $X + Y \sim \text{Poisson}(\theta_1 + \theta_2)$.

**12 pts** (b) Write down the probability of rejecting $H_0$ as a sum of probabilities. What choices for $C$ lead to a test with probability of type I error at most $\alpha$? Among those choices, which choice leads to a test with the lowest probability of a type II error? (In both cases just describe $C$.)

**6 pts** (c) Defining $C$ as in (b) is not very explicit. Under our modelling assumption, the Central Limit theorem tells us that if the data comes from Poisson($\theta$), then we can approximate

$$\mathbb{P}_\theta\left(\frac{T - n\theta}{\sqrt{n\theta}} > z_{1-\alpha}\right) = \mathbb{P}_\theta\left(\frac{T - \mathbb{E}T}{\sqrt{\mathbb{V}T}} > z_{1-\alpha}\right) \approx 1 - \Phi(z_{1-\alpha}) = 1 - (1-\alpha) = \alpha, \qquad \alpha \in (0,1),$$

with the approximation getting better as $n$ grows. Here, $\Phi$ represents the CDF of the standard normal distribution so that, by definition, $\Phi(z_\alpha) = \alpha$. Use this to show $C = 5n + 1.64\sqrt{5n}$ ensures that the type I error of the resulting test is (approximately) $\alpha = 0.05$.

**12 pts** (d) Use the approximation from (c) to approximate the power of the test. Make a sketch of the power $\pi(\theta)$ as a function of $\theta$. You don't need to get the shape right, just make sure that: i) you are explicit about what $\pi(\theta)$ is when $\theta$ is close to 5, and that ii) you get the monotonicity of $\pi(\theta)$ right.

**10 pts** (e) Suppose now that you collected data for two years ($n = 104$ weeks) and that the test statistic took the value $t = 581$. Use the approximation from ($c$) to approximate the p-value of the test. Based on this, what is the conclusion of the test at significance level $\alpha = 0.05$?

**Prob.III:** The literature suggests that there might be a relation between (the logarithm of the) heart-rate and longevity in mammals. In Table 1 you can see some data for 14 different mammals. In the pairs $(x_i, y_i)$, $i = 1, \ldots, 14$: the $x_i$ represents the (logarithm of the) average heart-rate in (log-)BMP, and $y_i$ represents the respective average longevity in years. We plot the data in Figure 1 where we also label each point with the name of the respective mammal.

Figure 1 seems to suggest that the relation between log-heart-rate and longevity might be linear. The actual data follows in Table 1.

From the observations in Table 1 we see that $n\bar{x} = 61.6315$, $n\bar{y} = 301.2000$, $SS_{xx} = 14.3109$, $SS_{yy} = 3477.5370$, and $SS_{xy} = -198.9952$. There are $n = 14$ measurements in total.

**6 pts** (a) Suppose that you would like to use a Simple Linear Regression model to derive a formula that allows you to model the relation between log-heart-rate ($x$) and the corresponding longevity ($Y$). In a linear regression model you assume that

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

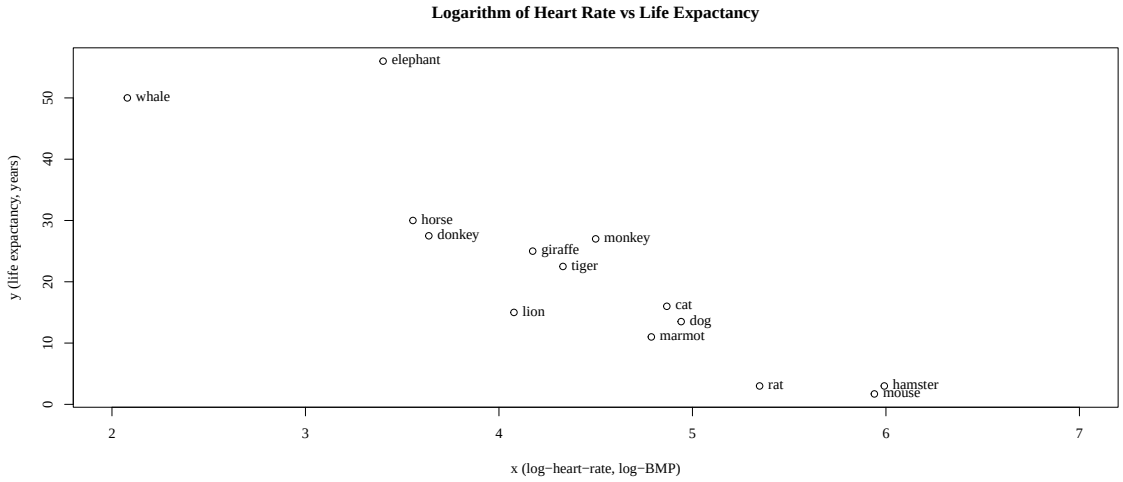


Figure 1: Average longevity- and heart-rate for 14 different mammals.

| Variables | Values |
|---|---|
| $(x_1, \ldots, x_n)$ | 5.9402 5.9915 5.3471 4.4998 4.7875 4.8675 4.9416 4.1744 4.3307 3.6376 3.5553 4.0775 3.4012 2.0794 |
| $(y_1, \ldots, y_n)$ | 1.7 3 3 27 11 16 13.5 25 22.5 27.5 30 15 56 50 |

Table 1: The longevity and log-heart-rate in mammals dataset.

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 $\epsilon_i$?

**10 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? (If you do not manage to compute the estimates, assume in the subsequent questions that your prediction formula is $\hat{y} = 80 - 10\,x$.)

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

**6 pts** (d) Say that you would like to make a prediction for the average life expectancy of a human that has an average heart-rate of 65; what would this prediction be? Does it seem like this model is picking up on an actual relation between (log-)heart-rate and life expectancy?