X_400004 - Statistics
Solutions to the Final

21 December 2021

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: 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 $X_1, \dots, X_n$ 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 $x > 0$,

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

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 $Y_1, \dots, Y_n$ distributed like Y you have that

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

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

Solution: If we follow the recommendation and apply the relation to $Y_i = n\theta X_i$ so that $Y_{(1)} = n\theta X_{(1)}$ we can conclude that

$$\mathbb{P}_\theta(n\theta X_{(1)} > y) = \{\mathbb{P}_\theta(n\theta X > y)\}^n = \{\mathbb{P}_\theta\big(X > y/(n\theta)\big)\}^n = \{1 - \mathbb{P}_\theta\big(X \le y/(n\theta)\big)\}^n.$$

Since $X \sim \text{Exp}(\theta)$, we conclude that the above is just

$$\{1 - F_\theta\big(y/(n\theta)\big)\}^n = \{\exp\big(-\theta y/(n\theta)\big)\}^n = \exp\big(-n\theta y/(n\theta)\big) = \exp(-y) = 1 - F_1(y)$$

From here we conclude that $\mathbb{P}_\theta(n\theta X_{(1)} \le y) = F_1(y)$, which is equivalent to $n\theta X_{(1)} \sim \text{Exp}(1)$ as we wanted to show.

4 pts (b) Justify why T is a pivot for θ.

Solution: The quantity T is a function of the data and θ and has a distribution that is known to us and does not depend on any unknown parameters . This is exactly the definition of T being a pivot for θ.

8 pts (c) Let e_α, $\alpha \in [0,1]$, represent the quantile of level α 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**.

Solution: Since $\left[0, nX_{(1)}/e_\alpha\right]$ is clearly an *upper* confidence interval , we have to show that it has the correct level. Since the *expected waiting time* is $1/\theta$, this means that we have to show that $\mathbb{P}_\theta\big(\left[0, nX_{(1)}/e_\alpha\right] \ni 1/\theta\big) = 1-\alpha$. This probability is the same as:

$$\mathbb{P}_\theta\big(0 \leq 1/\theta \leq nX_{(1)}/e_\alpha\big) = \mathbb{P}_\theta\big(1/\theta \leq nX_{(1)}/e_\alpha\big) = \mathbb{P}_\theta\big(e_\alpha \leq n\theta X_{(1)}\big) = 1 - F_1(e_\alpha) = 1-\alpha,$$

where we use that, by definition, $F_1(e_\alpha) = \alpha$, and that $n\theta X_{(1)} \sim \mathrm{Exp}(1)$.

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: $e_{0.01} = 0.0101$, $e_{0.05} = 0.0513$, $e_{0.95} = 2.9957$, $e_{0.99} = 4.6052$.)

Solution: By definition of the confidence interval, for any combination of n, α, θ, if the data comes from $\mathrm{Exp}(\theta)$, then the chance that $\left[0, nX_{(1)}/e_\alpha\right]$ does *not* contain $1/\theta$ is exactly α . To test the hypotheses above at significance level 0.05, we just have to check if 2 belongs to the interval $\left[0, nX_{(1)}/e_\alpha\right] = \left[0, 4 \times 0.05/e_{0.05}\right] = [0,\ 3.8991]$; since it does, we cannot reject H_0 at level 0.05.

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 $X \sim \mathrm{Poisson}(\theta)$, $\theta > 0$. 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 θ 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 $X + Y \sim \text{Poisson}(\theta_1 + \theta_2)$.

Solution: Using the tip $n-1$ times, since $X_1, \dots, X_n \sim \text{Poisson}(\theta)$, then $T = \sum_{i=1}^{n} X_i \sim \text{Poisson}(n\theta)$. Specifically under H_0, where $\theta = 5$, $T \sim \text{Poisson}(5n)$.

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 α? Among those choices, which choice leads to a test with the lowest probability of a type II error? (In both cases just describe C.)

Solution: If we reject when $T > C$, then the probability of rejecting under H_0 is the probability that $T > C$ computed under the assumption that $T \sim \text{Poisson}(5n)$. This is

$$\mathbb{P}_5(T > C) = \sum_{i>C}^{\infty} \mathbb{P}_5(T = i) = \sum_{\substack{i=0 \\ i>C}}^{\infty} \mathbb{P}_5(T = i) = \sum_{\substack{i=0 \\ i>C}}^{\infty} \frac{(5n)^i}{i!} e^{-5n}.$$

So, under H_0, $\mathbb{P}_5(T > C)$ is 1 if $C < 0$, and decreases as C increases. This means that if we pick C so that $\mathbb{P}_5(T > C) \le \alpha$ then also $\mathbb{P}_5(T > C + 1) \le \alpha$. If m is the smallest integer so that $\mathbb{P}_5(T > m) \le \alpha$, then any $C \in \{m, m + 1, \dots, n\}$ gives us a test with probability of type I error below α . The larger the rejection region, the smaller the probability of a type II error so picking $C = m$ leads to the test with the smallest probability of a type II error.

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 $\text{Poisson}(\theta)$, 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, Φ 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) $\alpha = 0.05$.

Solution: From what we are told, if we are under H_0 so that $\theta = 5$,

$$\mathbb{P}_5\left(\frac{T - 5n}{\sqrt{5n}} > z_{1-0.05}\right) \approx 0.05, \qquad \textbf{so that} \qquad \mathbb{P}_5\left(T > 5n + z_{1-0.05}\sqrt{5n}\right) \approx 0.05.$$

We can then just take $C = 5n + z_{1-0.05}\sqrt{5n} = 5n - z_{0.05}\sqrt{5n} = 5n + 1.64\sqrt{5n}$, where we use that $z_{1-\alpha} = -z_\alpha$.

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 θ. You don't need to get the shape right, just make sure that: i) you are explicit about what $\pi(\theta)$ is when θ is close to 5, and that ii) you get the monotonicity of $\pi(\theta)$ right. **Solution: The power of the test is the probability of rejecting under the alternative as a function of θ, i.e., $\pi(\theta) = \mathbb{P}_\theta(T > 5n + z_{1-0.05}\sqrt{5n})$ with this has**

to be computed under the assumption that $T \sim \text{Poisson}(n\theta)$, so by centering with $n\theta$ and scaling with $\sqrt{n\theta}$,

$$\pi(\theta) = \mathbb{P}_\theta\left(\frac{T - n\theta}{\sqrt{\theta n}} > \frac{(5-\theta)n}{\sqrt{\theta n}} + z_{1-0.05}\sqrt{\frac{5}{\theta}}\right) \approx 1 - \Phi\left(\frac{(5-\theta)n}{\sqrt{\theta n}} + z_{1-0.05}\sqrt{\frac{5}{\theta}}\right).$$

About the sketch, the argument of Φ above clearly decreases with θ, and Φ is an increasing function so $\pi(\theta)$ increases with θ. Furthermore, for $\theta \approx 5$ the above simplifies to $1 - \Phi(z_{1-0.05}) = 0.05$. So a sketch of π as a function of θ should start at the point $(5, 0.05)$, and then increase as θ increases .

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 $\alpha = 0.05$? **Solution: By definition, the p-value is the smallest significance level α at which we would reject H_0 when observing $t = 581$. We reject at significance level α if $581 > 5n + z_{1-\alpha}\sqrt{5n}$. If α leads to a rejection then we must have**

$$581 > 5 \times 104 + z_{1-\alpha}\sqrt{5 \times 104} \Leftrightarrow \frac{581 - 520}{\sqrt{520}} > z_{1-\alpha} \Leftrightarrow \Phi(2.6750) > \Phi(z_{1-\alpha}) = 1 - \alpha.$$

We conclude that all $\alpha > 1 - \Phi(2.6750)$ leads to rejection, so the p-value must be $1 - \Phi(2.6750)$. (This is ≈ 0.0037 but you could not get this with a simple calculator.) Since the p-value is smaller than 0.05 we would reject the null at significance level 0.05: indeed $1 - \Phi(2.6750) < 0.05$ because this is exactly the same as $\Phi^{-1}(1 - 0.05) = z_{0.95} = 1.64 < 2.6750$.

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.

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.

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

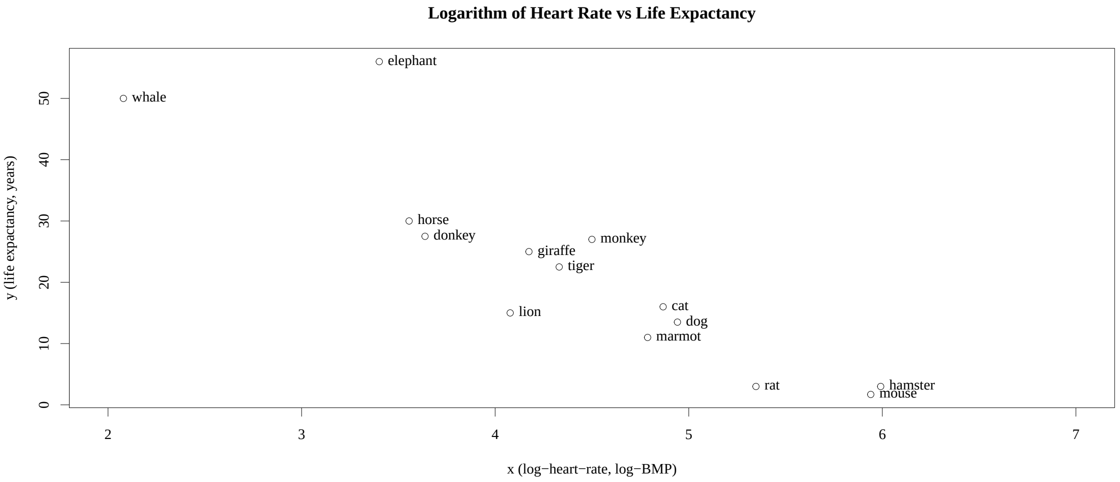

Figure 1: Average longevity- and heart-rate for 14 different mammals.

allows you to model the relation between log-heart-rate (x) and the corresponding longevity (Y). In a linear regression model you assume that

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

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 noise should have expectation 0; iii) the variance of the ϵ_i should be 1.

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 $\hat{y} = 80 - 10\,x$.)

Solution: We have that $\hat{\beta} = S_{xy}/SS_{xx} = -198.9952/14.3109 = -13.9052$ and $\hat{\alpha} = \bar{y} - \bar{x} \times SS_{xY}/SS_{xx} = \bar{y} - \bar{x} \times \hat{\beta} = 301.2000/14 - 61.6315/14 \times (-13.9052) = 82.7282$.

6 pts (c) Estimate the variance of the noise σ^2 under the SLR modelling assumption.

Solution: The estimator for the variance of the noise is $\hat{\sigma}^2 = SS_{yy}/n - \hat{\beta}^2 SS_{xx}/n = 3477.5370/14 - (-13.9052)^2 \times 14.3109/14 = 50.7471$.

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?

Solution: Using our prediction formula $\hat{y} = \hat{\alpha} + \hat{\beta} \times x = 82.7282 - 13.9052 \times x$ and plugging in $x = \log(65) = 4.1744$ we get the prediction $\hat{y} = 24.6871$. (Using the prediction formula $\hat{Y} = 80 - 10 \times x$ you would get $\hat{Y} = 38.2561$.) We know that average life expectancy for humans is much more than about 25 years (or 38 years,

for that matter) so it seems like either humans do not follow this model, or there is only correlation between x and y but no (causal) relation.