Statistiek (WISB263)

Final Exam

January 31, 2018

Schrijf uw naam op elk in te leveren vel. Schrijf ook uw studentnummer op blad 1.
(The exam is an *open–book* exam: notes and book are allowed. The scientific calculator is allowed as well).
The maximum number of points is 105 (5 extra BONUS points!!).
Grade= $min(100, \text{points})$.
Points distribution: 30–20–25–25 (+5 extra BONUS points!!)

1. An experiment is recording the number of particles emitted each minute by a radioactive source. The experiment is then repeated independently n times with different observation periods. Hence, a sample $\mathbb{X} = \{X_1, \ldots, X_n\}$ is collected and it is modelled as a sequence of *independent* Poisson random variables such that $X_i \sim \text{Poi}(i\,\mu)$, for $i \in \{1, \ldots, n\}$, i.e:

$$\mathbb{P}(X_i = k) = \frac{1}{k!} e^{-(i\,\mu)} (i\,\mu)^k$$

where μ is an unknown positive parameter.

(a) [6pt] Show that the maximum likelihood estimator for μ is given by:

$$\hat{\mu} = \frac{2}{n(n+1)} \sum_{i=1}^{n} X_i$$

(b) [8pt] Calculate $p = \mathbb{P}(\max_i X_i = 0)$ and find its maximum likelihood estimator $\hat{p}_{MLE}$.

We perform now a **different** experiment, because we are interested in whether we can detect at least one particle in a fixed time interval. Therefore, we perform n independent measurements whose outcomes are modelled by n **i.i.d.** Poisson random variables (i.e. $X_i \sim \text{Poi}(\mu)$), and we record *only* whether we detect at least one particle ($X_i > 0$).

(c) [8pt] What is the maximum–likelihood estimate of μ if for r times we detect zero particle in the n measurements?

(d) [8pt] Find an *approximated* 95% confidence interval for the maximum–likelihood estimate of μ of point (c).

2. We want to decide if a continuous random variable has probability density function $f_1(x)$ or $f_2(x)$, with

$$f_1(x) = \frac{1}{\theta_1} x^{\frac{1}{\theta_1}-1} \mathbf{1}_{(0,1)}(x), \quad f_2(x) = \frac{1}{\theta_2} x^{\frac{1}{\theta_2}-1} \mathbf{1}_{(0,1)}(x),$$

by using only one observation X. The parameters θ_1 and θ_2 are fixed and positive; we have denoted with $\mathbf{1}_{(0,1)}$ the indicator function of the open interval $(0, 1)$.

(a) [8pt] If $f(x)$ is the probability density function of the random variable X, find the most powerful test for testing:

$$\begin{cases} H_0: & f(x) = f_1(x) \\ H_1: & f(x) = f_2(x) \end{cases}$$

at the α level of significance.

(b) [6pt] Calcolate the p-value of the test if $x = 0.7$, $\theta_1 = 1$ and $\theta_2 = 5$.

(c) [6pt] Calculate the power of the test if $\alpha = 0.05$, $\theta_1 = 1$ and $\theta_2 = 5$.

3. In an experiment mice were placed in a wheel that is partially submerged in water. If they keep the wheel moving, they will avoid the water. The response is the number of wheel revolutions per minute. **Group 1** is a placebo group while **Group 2** consists of mice that are under the influence of a drug containing performance–enhancing substances. The data are:

Table 1:

Group1 X	0.3	1.3	1.2	0.8	18.2	0.5	1.4	0.6	1.6	0.0
Group2 Y	1.7	2.8	4.0	2.4	1.1	4.9	6.2	7.5	1.8	1.9

(a) [5pt] Calculate the difference in sample means. Comment this result.

(b) [12pt] Test at 0.05 level of significance the hypothesis that the *drug has no effect*. Critically justify the choice of the test. Do you think that the result of your analysis depends on this choice? Try to justify quantitatively the previous answer.

(c) [8pt] Estimate the probability that the drugged mice are performing better than the mice in the placebo group and estimate an *approximated* 95% CI for this probability.

4. Two laboratories take n measurements on the same quantity μ with two different instruments. We can model the measurements performed by the two laboratories as realisations of the following two linear models:

$$\begin{aligned} Y_i &= \mu + \epsilon_i, \\ Z_i &= \mu + \delta_i, \end{aligned}$$

where $i \in \{1, \dots, n\}$, ϵ_i are i.i.d. random variables such that $\mathbb{E}(\epsilon_i) = 0$, $\mathrm{Var}(\epsilon_i) = \sigma^2$; δ_i are i.i.d. random variables such that $\mathbb{E}(\delta_i) = 0$, $\mathrm{Var}(\delta_i) = 4\sigma^2$ and $\epsilon_i \perp \delta_j \ \forall i, j$.

In order to estimate μ, we pool the measurements together, so that we define a new random variable U_i such that:

$$U_i := \begin{cases} Y_i & \text{if } i \in \{1, \dots, n\}, \\ Z_{i-n} & \text{if } i \in \{n+1, \dots, 2n\}, \end{cases}$$

with $i \in \{1, \dots, 2n\}$.

(a) [3pt] Does the data $\mathbb{U} = \{U_1, \dots, U_{2n}\}$ satisfy the standard assumptions of linear regression? If not, which of the assumptions is violated?

(b) [7pt] Think of a method of transforming these data to a linear regression model satisfying the standard assumptions.

(c) [4pt] Using the transformed data, find the least squares estimator $\hat{\mu}_{LS}$ of μ.

(d) [4pt] Find the expected value and the variance of $\hat{\mu}_{LS}$.

(e) [7pt] In case σ^2 is unknown, propose an unbiased estimator for estimating σ^2 .

5. **BONUS POINTS** [5pt]: Let T be a test statistic with a continuous probability distribution F_0 under H_0. Then $1 - F_0(t)$ is the p–value of the test that rejects H_0 for large values of t.

(a) [2pt] Show that under H_0, the p–value $1 - F_0(T)$ is uniformly distributed in $[0, 1]$.

(b) [3pt] Suppose now that you are not a fair scientist, so that you perform each experiment twice and you report just the highest p-value of each couple of measurements. Which is in this case the distribution of the reported p-values under H_0?