Statistiek (WISB263)

Resit Exam

July 11, 2022

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 use of a laptop is allowed as well, under the restriction that the invigilator can look at the screen at all times, and that students are not allowed to type on the computer and wifi is off. The scientific calculator is also allowed).

The maximum number of points is 110 (10 extra BONUS points!!).
Grade= $min(100, \text{points})$.
Points distribution: 25–25–23-15-12 (+10 extra BONUS points!!)

1. Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a sample with $X_i \overset{i.i.d.}{\sim} \text{Unif}[\theta, 3\theta]$ for any $i \in \{1, \ldots, n\}$, with $\theta > 0$.

   (a) [8pt] Find the the Maximum Likelihood Estimator (MLE) $\hat{\theta}_{\text{MLE}}$ of θ and state the conditions on the sample for its existence. Is $\hat{\theta}_{\text{MLE}}$ biased?

   (b) [4pt] Is $\hat{\theta}_{\text{MLE}}$ consistent?

   (c) [2pt] Can we use the asymptotic theory for showing the asymptotic normality of $\hat{\theta}_{\text{MLE}}$? If yes, which is its asymptotic variance?

   (d) [7pt] Find the Method of Moment Estimator (MoM) $\hat{\theta}_{\text{MoM}}$ of θ.

   (e) [4pt] Given the sample $\mathbf{x}$ of size $n = 100$ and sample mean $\overline{\mathbf{x}}_{100} = 2.10$, estimate an *approximate* 95%–confidence interval for θ based on the MoM estimator $\hat{\theta}_{\text{MoM}}$.

2. Suppose we have three independent samples $\mathbf{X} = (X_1, \ldots, X_{n_1})$, with $X_i \overset{i.i.d.}{\sim} N(\mu_1, \sigma^2)$; $\mathbf{Y} = (Y_1, \ldots, Y_{n_2})$, with $Y_i \overset{i.i.d.}{\sim} N(\mu_2, \sigma^2)$ and $\mathbf{W} = (W_1, \ldots, W_{n_3})$, with $W_i \overset{i.i.d.}{\sim} N(\mu_3, \sigma^2)$. In this exercise we want to estimate a function θ of the parameters:

$$\theta := a_1\mu_1 + a_2\mu_2 + a_3\mu_3$$

   with a_1, a_2, a_3 known constants.

   (a) [6pt] Show that the MLE $\hat{\theta}$ of θ is:

$$\hat{\theta} = a_1\overline{\mathbf{X}}_{n_1} + a_2\overline{\mathbf{Y}}_{n_2} + a_3\overline{\mathbf{W}}_{n_3}$$

   where we denoted with $\overline{\mathbf{X}}_{n_1}, \overline{\mathbf{Y}}_{n_2}, \overline{\mathbf{W}}_{n_3}$ the sample means of the samples $\mathbf{X}, \mathbf{Y}, \mathbf{W}$ respectively.

   (b) [5pt] Which is the distribution of the estimator $\hat{\theta}$?

   (c) [4pt] If we denote with S_X^2, S_Y^2, S_W^2 the sample variances of the three samples (e.g., $S_X^2 := \frac{1}{n_1-1}\sum_{i=1}^{n_1}(X_i - \overline{\mathbf{X}}_{n_1})^2$), and with

$$S_p^2 := \frac{1}{n_1 + n_2 + n_3 - 3}((n_1-1)S_X^2 + (n_2-1)S_Y^2 + (n_3-1)S_W^2)$$

   the pooled sample variance, find the distribution of $\frac{n_1+n_2+n_3-3}{\sigma^2}S_p^2$ and of

$$U := \frac{\hat{\theta} - \theta}{S_p\sqrt{\frac{a_1^2}{n_1} + \frac{a_2^2}{n_2} + \frac{a_3^2}{n_3}}}$$

   (d) [4pt] Find a $(1-\alpha)100\%$-confidence interval for θ.

   (e) [6pt] Develop a test for $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$.

3. Suppose that $\mathbf{Y} = (Y_1, \ldots, Y_n)$, are i.i.d. discrete random variables the probability mass function (pmf) given by:

$$p_{Y_i}(y) := \mathbb{P}(Y_i = y) = \begin{cases} \theta_1, & \text{if } y = 1; \\ \theta_2, & \text{if } y = 2; \\ \theta_3, & \text{if } y = 3; \\ 0, & \text{otherwise.} \end{cases}$$

where θ_j with $j \in \{1, 2, 3\}$ are unknown parameters.

(a) [8pt] Find the MLEs of θ_1, θ_2 and θ_3.

(b) [8pt] We want to test the hypotheses:

$$\begin{cases} H_0: & \theta_1 = \theta_2 = \theta_3, \\ H_1: & H_0 \text{ is not true.} \end{cases}$$

at α level of significance. Derive a test statistic and give the general expression of the rejection region. Clearly state whether the test is an exact test or an asymptotic test.

(c) [7pt] If we collect the sample:

$$\mathbf{y} = \{1, 2, 3, 2, 2, 2, 1, 2, 3, 3, 3, 3, 1, 2, 3, 1, 1, 2, 1, 3, 1, 2, 1, 2, 3, 1, 1, 1, 1, 2, 3, 1\}$$

can you reject at 0.05 level of significance H_0 of point (b)?

4. In order to study the pollution in drinking water due to metals, several measurements have been taken in six different river locations. In each location the level of zinc concentration (mg/ℓ) for both surface water and bottom water was determined, obtaining the following data:

	Location 1	Location 2	Location 3	Location 4	Location 5	Location 6
Zinc in bottom water	0.43	0.27	0.57	0.53	0.71	0.72
Zinc in surface water	0.42	0.24	0.39	0.41	0.61	0.62

(a) [15pt] Design a statistical test for answering the research question:

`Does the concentration of zinc in bottom water exceed that of surface water?`

Perform the proposed test at 0.05 level of significance. State all the assumptions of the test and discuss their plausibility.

5. Let Y be a χ^2 distributed random variable with n degrees of freedom. We know that $\mathbb{E}(Y) = n$ and $\mathrm{Var}(Y) = 2n$.

(a) [6pt] Show that:

$$\frac{Y - n}{\sqrt{2n}} \xrightarrow{d} Z \sim N(0, 1), \text{ as } n \to \infty.$$

(b) [6pt] A machine produces steel rods of length Z, where $Z \sim N(\mu, \sigma^2)$, with $\mu = 6$ (in cm) and $\sigma^2 = 0.2$ (in cm^2). The cost L of repairing a rod that is not exactly 6 (cm) in length is given by $L = 4(Z - \mu)^2$ (in euro). If 50 rods with independent lengths are produced in a given day, approximate the probability that the total cost for repairs for that day is larger than 48.

BONUS (a) [6pt] Given a continuous random variable X, show that the random variable $Z := -\log(1 - F_X(X))$ is χ^2 distributed with two degrees of freedom. We denoted with F_X the cumulative distribution function of X (i.e., $F_X(x) = \mathbb{P}(X \le x)$).

(b) [4pt] Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a sample i.i.d. random variables sampled from a continuous distribution, whose probability density function is:

$$f_{X_i}(x) = \theta(x + 1)^{-(\theta+1)},$$

for $x \ge 0$ and $\theta > 0$ an unknown parameter. Construct a $(1 - \alpha)100\%-$ confidence interval for θ, by using the result in point (a).