Statistiek (WISB361)

Retake exam

August 21, 2014

Schrijf uw naam op elk in te leveren vel. Schrijf ook uw studentnummer op blad 1.
The maximum number of points is 100.
Points distribution: 17–23–20–18–22

1. Let $X_1, \dots, X_n$ be i.i.d. normal random variables with mean θ and variance θ^2. Let

$$T_1 = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

and let

$$T_2 = c_n\, S = c_n \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}}$$

where the constant c_n is such that the expectation $\mathbb{E}(T_2) = \theta$ (Let op: it is not needed to calculate c_n!!!). Consider the estimator $W(\alpha)$ of θ of the form:

$$W(\alpha) = \alpha T_1 + (1-\alpha)T_2$$

where $0 \le \alpha \le 1$.

(a) [7pt] Find the variance $\text{Var}(W(\alpha))$ of $W(\alpha)$.

(b) [3pt] Find the mean square error (MSE) of $W(\alpha)$ in terms of α, $\text{Var}(T_1)$ and $\text{Var}(T_2)$.

(c) [5pt] Determine in terms of α, $\text{Var}(T_1)$ and $\text{Var}(T_2)$ the value of α that gives the smallest MSE of $W(\alpha)$.

(d) [2pt] In case we have the approximation:

$$\text{Var}(T_2) \approx \frac{\theta^2}{2n},$$

find the the value of α that gives the smallest MSE of $W(\alpha)$.

2. Suppose we have a sample $\mathbb{X} = \{X_1, \dots, X_n\}$ of i.i.d. random variables $X_i \sim \text{Unif}(\theta, \theta + |\theta|)$, with $i = 1, \dots, n$ and where θ is an unknown parameter. Calculate the maximum likelihood estimator (MLE) of θ in the following cases:

(a) [7pt] $\theta > 0$

(b) [5pt] $\theta < 0$

(c) [5pt] $\theta \neq 0$

Consider now the same sample $\mathbb{X} = \{X_1, \dots, X_n\}$ of i.i.d. random variables $X_i \sim \text{Unif}(a\,\theta, b\,\theta)$, with $i = 1, \dots, n$, where a, b are constants such that $0 < a < b$.

(d) [4pt] Calculate the MLE of θ.

(e) [2pt] Find a two–dimensional sufficient statistics for θ.

3. Let $\mathbf{y} = \{y_1, \dots, y_n\}$ the realization of the random vector $\mathbf{Y} = \{Y_1, \dots, Y_n\}$ with independent components and such that $Y_i \sim N(\theta x_i, 1 + x_i^2)$, with $i = 1, \dots, n$, where $\theta \in \mathbb{R}$ is an unknown parameter and x_i are known constants such that:

$$\sum_{i=1}^{n} \frac{x_i^2}{1 + x_i^2} = 1.$$

Consider a size α test:

$$\begin{cases} H_0: & \theta = 0, \\ H_1: & \theta = 1. \end{cases}$$

For $c \in \mathbb{R}$, let R_c be the region:

$$R_c = \{\mathbf{y} \in \mathbb{R}^n : t(\mathbf{y}) > c\},$$

where

$$t(\mathbf{y}) = \sum_{i=1}^{n} \frac{x_i\, y_i}{1 + x_i^2}.$$

(a) [10pt] Show that the choice of the *rejection* region R_c *maximizes* the power of the test, for any fixed α.

(b) [3pt] Find the distribution of $t(\mathbf{Y})$ and derive an expression for the power of the test.

(c) [7pt] Show that $t(\mathbf{Y})$ is the maximum likelihood estimator (MLE) of θ. Is this estimator unbiased?

4. A study was performed in order to observe the variation on dietary habits between summer and winter among the female population. For this reason, a sample of 12 women was screened during the months of January and July 2009, measuring for each individual the percentage of the total caloric intake that comes from fat. These percentages were measured twice for each woman: one measurement X_i was taken in January and the second Y_i in July, with $i = 1, \ldots, 12$. The results for the pairs (X_i, Y_i) are shown in the following table:

	percentage of calories coming from fats
X_i	30.5, 28.4, 40.2, 37.6, 36.5, 38.8, 34.7, 29.5, 29.7, 37.2, 41.5, 37.0
Y_i	32.2, 27.4, 28.6, 32.4, 40.5, 26.2, 29.4, 25.8, 36.6, 30.3, 28.5, 32.0

We assume that X_i and Y_i are normally distributed and that *different* pairs are independently distributed.

(a) [10pt]. Test the hypothesis that the percentage of calories coming from fat is higher in January than in July, at $\alpha = 0.05$ level of significance.

(b) [8pt] Estimate the probability that for a randomly chosen woman, the percentage of calories coming from fat in July is less than the percentage of calories coming from fat in January.

5. For a certain rubber manufacturing process, the random variable Y_x (the amount in kilograms manufactured per day) has mean $\mathbb{E}Y_x = \alpha x + \beta x^2$ and *known* variance $\mathrm{Var}(Y_x) = \sigma^2$, where x is a constant, denoting the amount of raw material in kilograms used per day in the manufacturing process. The n data pairs (x, Y_x), $x = 1, \ldots, n$, are collected in order to estimate the unknown parameters of interest α and β. We assume that $Y_1, \ldots, Y_n$ are n mutually independent random variables. Furthermore, let

$$S_k = \sum_{x=1}^{n} x^k$$

with $k \in \mathbb{N}$ (S_k are non–stochastic quantities with known values).

(a) [8pt] Derive explicit expressions (in terms of S_k) for the least squares estimators $\hat{\alpha}$ and $\hat{\beta}$ of the unknown parameter α and β.

(b) [8pt] Derive explicit expressions for $\mathbb{E}(\hat{\beta})$ and $\mathrm{Var}(\hat{\beta})$, i.e. the mean and the variance of $\hat{\beta}$.

(c) [6pt] If $Y_x \sim N(\alpha x + \beta x^2, \sigma^2)$, with $x = 1, \ldots, n$, and Y_x being mutually independent random variables, compute an exact 95% –confidence interval for β if $n = 4$, $\hat{\beta} = 2$, and $\sigma^2 = 1$.