Statistiek (WISB361)

Final exam

June 29, 2015

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: 23–20–20–20–17

1. Consider the sample $\mathbb{X} = \{X_i\}_{i=1}^n$ of i.i.d. random variables with probability density function:

$$f(x|\theta) = \begin{cases} \dfrac{2\theta^2}{x^3} & \text{for } x \geq \theta, \\ 0 & \text{otherwise} \end{cases}$$

for $\theta > 0$.

(a) [7pt] Determine the maximum likelihood estimator $\hat{\theta}_{MLE}$ and $\widehat{\theta^2}_{MLE}$ of θ and θ^2 respectively.
Solution:
The likelihood for a realization x of the sample can be written as:

$$lik(\theta) = \prod_{i=1}^n \frac{2\theta^2}{x_i^3} \mathbf{1}_{(\theta,+\infty)}(x_i) = \frac{2^n\theta^{2n}}{\prod_{i=1}^n x_i^3} \prod_{i=1}^n \mathbf{1}_{(0,x_i)}(\theta) = \frac{2^n\theta^{2n}}{\prod_{i=1}^n x_i^3} \mathbf{1}_{(0,x_{(1)})}(\theta)$$

The likelihood is increasing in θ, hence:

$$\widehat{\theta}_{MLE} = X_{(1)}$$

where $X_{(1)} := \min_i X_i$. By the *invariance principle*:

$$\widehat{\theta^2}_{MLE} = (\widehat{\theta}_{MLE})^2 = X_{(1)}^2$$

(b) [4pt] Determine the distribution of $\hat{\theta}_{MLE}$.
Solution:
The probability distribution function (PDF) of $\widehat{\theta}_{MLE}$ is the PDF of $X_{(1)}$. Therefore

$$F_{X_{(1)}}(x) = 1 - \mathbb{P}(X_{(1)} > x) = 1 - \prod_{i=1}^n \mathbb{P}(X_i > x) = 1 - (1 - F_{X_1}(x))^n,$$

with

$$F_{X_1}(x) = \int_\theta^x 2\theta^2/y^3 dy = 1 - \frac{\theta^2}{x^2}$$

for $x > \theta$. Therefore

$$F_{X_{(1)}}(x) = \left(1 - \frac{\theta^{2n}}{x^{2n}}\right) \mathbf{1}_{(\theta,\infty)}(x)$$

and the probability density function (pdf) is:

$$f_{X_{(1)}}(x) = \left(2n\frac{\theta^{2n}}{x^{2n+1}}\right) \mathbf{1}_{(\theta,\infty)}(x)$$

(c) [3pt] Is $\hat{\theta}_{MLE}$ a biased estimator? Is $\hat{\theta}_{MLE}$ *asymptotically* unbiased?
Solution:
We have:

$$\mathbb{E}(\widehat{\theta}_{MLE}) = \mathbb{E}(X_{(1)}) = 2n \int_\theta^\infty x \frac{\theta^{2n}}{x^{2n+1}} dx = \frac{2n}{2n-1}\theta \tag{1}$$

so that $\widehat{\theta}_{MLE}$ is biased. However,

$$\lim_{n\to\infty} \mathbb{E}(X_{(1)}) = \theta$$

so that $\widehat{\theta}_{MLE}$ is asymptotically unbiased.

(d) [3pt] Find the method of moment estimator $\hat{\theta}_{MoM}$ of θ.
Solution:
Since $\mathbb{E}(X_{(1)}) = \mathbb{E}(X_1)$, by (1) with $n = 1$, we have that $\mathbb{E}(X_1) = 2\theta$. Thus:

$$\hat{\theta}_{MoM} = \frac{\bar{X}_n}{2}$$

where we denote with $\bar{X}_n$ the sample mean.

(e) [3pt] Find the variance of $\hat{\theta}_{MoM}$. Is it finite?
Solution:
$\text{Var}(\hat{\theta}_{MoM}) = \frac{\text{Var}(X_1)}{4n} = +\infty$, because

$$\mathbb{E}(X_1^2) = \int_\theta^\infty x^2 \frac{\theta^2}{x^3} dx = +\infty$$

(f) [3pt] Compare the *mean squared errors* (MSE) of $\hat{\theta}_{MLE}$ and $\hat{\theta}_{MoM}$. Which estimator is the most efficient?
Solution:
By point (d) it follows that $\hat{\theta}_{MoM}$ is unbiased. Hence $MSE(\hat{\theta}_{MoM}) = \text{Var}(\hat{\theta}_{MoM}) = +\infty$. As regards the $MSE(\widehat{\theta}_{MLE})$, we have:

$$\widehat{\theta}_{MLE} = \text{Var}(\widehat{\theta}_{MLE}) + (\mathbb{E}(\widehat{\theta}_{MLE}) - \theta)^2 = \text{Var}(X_{(1)}) + \left(\frac{\theta}{2n-1}\right)^2$$

Since

$$\mathbb{E}(X_{(1)}^2) = \int_\theta^\infty x^2 2n \frac{\theta^{2n} x^{2n+1}}{d} x = \frac{n}{n-1}\theta^2$$

when $n \geq 2$, we get:

$$MSE(\widehat{\theta}_{MLE}) = \frac{\theta^2}{(n-1)(2n-1)}$$

Therefore if the sample size is larger than 1, $\widehat{\theta}_{MLE}$ is more efficient than $\widehat{\theta}_{MoM}$.

2. Coffee abuse is often considered related to an increase of heart rate (i.e. number of poundings of the heart per unit of time). In order to support this claim, an experiment was planned on a sample of 10 subjects. The hearth rate was measured twice for each subject: the first time *at rest*, the second after having drank a cup of coffee. The measurements *at rest* are:

$$\mathbf{x} = \{x_1, \ldots, x_{10}\} = \{74, 68, 67, 71, 69, 65, 70, 70, 66, 67\}$$

and after a cup of coffe:

$$\mathbf{y} = \{y_1, \ldots, y_{10}\} = \{71, 72, 69, 66, 73, 77, 69, 68, 71, 78\}$$

We can assume that the random variables $Z_i = Y_i - X_i$, denoting the difference between the heart rate after the cup of coffee and the heart rate at rest, are i.i.d. normal random variables.

(a) [12pt] Test the hypothesis that the coffee increase the heart rate at $\alpha = 0.05$ level of significance.
Solution:
We have paired normal observations and we will perform a paired t–test. Since $Z_i := Y_i - X_i \sim N(\mu_\Delta, \sigma^2)$, with $\mu_\Delta := \mathbb{E}(Y_i - X_i)$, and unknown variance σ^2, we want to test the hypotheses:

$$\begin{cases} H_0: & \mu_\Delta = 0, \\ H_1: & \mu_\Delta > 0. \end{cases}$$

at $\alpha = 0.05$ level of significance. Given the test statistics T:

$$T = \frac{\bar{Z}}{S/\sqrt{n}}$$

with $S^2 = \frac{1}{n-1}\sum_{i=1}^n (Z_i - \bar{Z})^2$, under H_0 $T \sim t(n-1)$. With the data of the problem, we have the following realization t of the test statistic:

$$t = \frac{\bar{z}}{s/\sqrt{n}} = \frac{2.7}{5.697\sqrt{10}} = 1.450$$

We perform a one–sided t–test, so that the rejection region B is $B = [t_{9,0.05}, +\infty) = [1.833, +\infty)$. Since $1.450 \notin B$, we do not reject H_0 at 0.05 level of significance.

(b) [5pt] Calculate the p-value of the test.
Solution:

$$p - \text{value} = \mathbb{P}(T \geq 1.450 | H_0).$$

From the tables, we can just say that $0.05 < p - \text{value} < 0.10$.

(c) [3pt] In case the normality assumption does not hold, how you could test the hypothesis of point (a)? (It is enough to explain which test is the most appropriate, without performing the analysis).
Solution:
Signed rank test.

3. Let X_1 and X_2 be i.i.d. random variables such that $X_i \sim \text{Unif}[\theta, \theta+1]$, for $i = 1, 2$, where $\text{Unif}[\theta, \theta+1]$ denotes the uniform distribution in the interval $[\theta, \theta+1]$. In order to test:

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

at $\alpha = 0.05$ level of significance. we have two competing tests, with the following rejection regions:

$$\begin{array}{l} \textbf{TEST1} : \text{Reject } H_0 \text{ if } X_1 > 0.95, \\ \textbf{TEST2} : \text{Reject } H_0 \text{ if } X_1 + X_2 > C, \end{array}$$

with $C \in \mathbb{R}$.

(a) [2pt] Find the significance level α of **TEST1**.
Solution:

$$\alpha_1 = \mathbb{P}(X_1 > 0.95 | \theta = 0) = 0.05$$

(b) [4pt] Find the value of C so that **TEST2** has the same significance level α of **TEST1**.
Solution:
Since $\mathbb{P}(X_1 + X_2 > C | \theta = 0)$ is the area of region inside the unit square (with vertices $(0,0)$, $(0,1)$, $(1,0)$ and $(1,1)$) above the line $x_2 = C - x_1$, we have:

$$\alpha_2(C) = \mathbb{P}(X_1 + X_2 > C | \theta = 0) = \begin{cases} 1 - C^2/2 : & \text{if } 0 \leq C \leq 1, \\ (2-C)^2/2 & \text{if } 1 < C \leq 2, \\ 0 & \text{otherwise} \end{cases}$$

We see that the equation $\alpha_2(C) = \alpha_1$ has solution only for $1 < C \leq 2$. Therefore, the solution is

$$C = 2 - \sqrt{2\alpha_1} = 2 - \sqrt{0.1} \approx 1.68$$

(c) [4pt] Calculate the power function of each test.
Solution:
For the first test:

$$\pi_1(\theta) = \mathbb{P}(X_1 > 0.95 | \theta) = \begin{cases} 0 & \text{if } \theta \leq -0.5, \\ \theta + 0.5 & \text{if } -0.5 < \theta \leq 0.95, \\ 1 & \text{otherwise} \end{cases}$$

For the second test, we notice that $\pi_2(\theta) = \mathbb{P}(X_1 + X_2 > C|\theta)$ is the area of the unit square with vertices (θ, θ), $(\theta, \theta+1)$, $(\theta+1, \theta)$ and $(\theta+1, \theta+1)$, over the line $x_2 = C - x_1$. Hence,

$$\pi_2(\theta) = \begin{cases} 0 & \text{if } \theta \leq C/2 - 1, \\ (2\theta + 2 - C)^2/2 & \text{if } (C/2) - 1 < \theta \leq (C-1)/2, \\ 1 - (C - 2\theta)^2/2 & \text{if } (C-1)/2 < \theta \leq C/2 \\ 1 & \text{if } \theta > C/2. \end{cases}$$

(d) [6pt] Is **TEST2** more powerful than **TEST1**?
Solution:
From point (c) it follows that TEST1 is more powerful for θ near 0, but TEST2 is more powerful for larger θ. Hence, TEST2 is not uniformly more powerful than TEST1.

(e) [4pt] Show how to get a test that has the same significance level but more powerful than **TEST2**.
Solution:
If either $X_1 \geq 1$ or $X_2 \geq 1$ we should reject H_0, because if $\theta = 0$ then $\mathbb{P}(X_i < 1) = 1$. Thus, if we consider the rejection region:

$$B := \{(x_1, x_2) : x_1 + x_2 > C\} \cup \{(x_1, x_2) : x_1 > 1\} \cup \{(x_1, x_2) : x_2 > 1\}$$

The first set is the rejection region for TEST2. The test with this rejection region B has the same significance as TEST2 because the last two sets both have probability 0 if $\theta = 0$. But for $0 < \theta < C - 1$ the power of this test is strictly larger than $\pi_2(\theta)$.

4. Let the independent normal random variables $Y_1, Y_2, \ldots, Y_n$ be such that $Y_i \sim N(\mu, \alpha^2 x_i^2)$, for $i = 1, \ldots, n$, where the given constants x_i are not all equal and no one of which is zero.

(a) [13pt] Derive the least squares estimators of μ and α^2, after you have properly rescaled the random variables Y_i.
Solution:
The linear model for Y_i is:

$$Y_i = \mu + \epsilon_i, \quad i = 1, 2, \ldots, n,$$

where ϵ_i are independent normal RV with zero mean and **different** variances: $\mathbf{Var}(\epsilon_i) = \alpha^2 x_i^2$. Let

$$Z_i = Y_i / x_i$$

Therefore, we have the linear model for Z_i:

$$Z_i = \mu \frac{1}{x_i} + \tilde{\epsilon}_i \quad i = 1, 2, \ldots, n,$$

where $\tilde{\epsilon}_i$ are **i.i.d.** RV with zero mean and common variance: $\mathbf{Var}(\tilde{\epsilon}_i) = \alpha^2$. Minimizing the sum of the squares $S(\mu) = \sum_{i=1}^n (Z_i - \frac{\mu}{x_i})^2$, we obtain the LS estimator $\hat{\mu}_{LSE}$ of μ:

$$\hat{\mu}_{LSE} = \frac{\sum_{i=1}^n Z_i / x_i}{\sum_{i=1}^n 1/x_i^2} = \frac{\sum_{i=1}^n Y_i / x_i^2}{\sum_{i=1}^n 1/x_i^2}$$

and, from the residual sum of the squares (RSS), we get the estimator of the variance:

$$\widehat{\alpha^2} = \frac{1}{n-1} \sum_{i=1}^n (Z_i - \hat{\mu}_{LSE})^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i / x_i - \hat{\mu}_{LSE})^2$$

(b) [4pt] Which is the distribution of $(n-1)\hat{\alpha}^2/\alpha^2$?
Solution:
It follows that $(n-1)\hat{\alpha}^2/\alpha^2$ has a χ^2_{n-1} distribution.

(c) [3pt] Discuss the test of hypotheses:

$$\begin{cases} H_0: & \alpha = 1, \\ H_1: & \alpha \neq 1. \end{cases}$$

Solution:
By point (b) we can perform a two–sided χ^2_{n-1} test, using the test statistics $(n-1)\hat{\alpha}^2$.

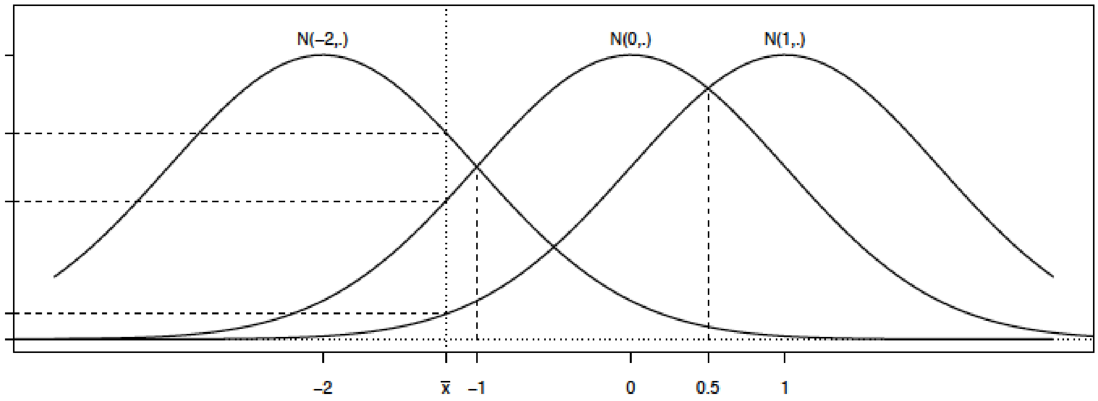
5. Consider the sample $\mathbb{X} = \{X_i\}_{i=1}^n$ of i.i.d. random variables such that $X_i \sim N(\theta, \sigma^2)$ with σ^2 **known** and $\theta \in \Omega$, where the parameter space is $\Omega = \{-2, 0, 1\}$.

(a) [2pt] Show that $\bar{X} = 1/n \sum_{i=1}^n X_i$ is a sufficient statistics for θ and that the likelihood $lik(\theta)$ can be factorized in $lik(\theta) = h(x) g_\theta(\bar{x})$, where x is a realization of $\mathbb{X}$, $\bar{x}$ is a realization of $\bar{X}$ and

$$h(x) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \bar{x})^2\right\}, \qquad g_\theta(\bar{x}) = \exp\left\{-\frac{n}{2\sigma^2}(\bar{x}-\theta)^2\right\}$$

Solution:

$$\begin{aligned} lik(\theta) &= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \theta)^2\right\} \\ &= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \bar{x} + \bar{x} - \theta)^2\right\} \\ &= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \bar{x})^2\right\} \exp\left\{-\frac{n}{2\sigma^2}(\bar{x}-\theta)^2\right\} \end{aligned}$$



In the figure above the functions $g_\theta(y)$ are plotted for the three possible values of the parameter θ.

(b) [8pt] Find a maximum likelihood estimator $\hat{\theta}_{MLE}$ of θ.
Solution:
From the Figure follows that:

$$\hat{\theta}_{MLE} = \begin{cases} -2 & \text{if} \quad \bar{x} < -1, \\ 0 & \text{if} \quad -1 \leq \bar{x} \leq 0.5 \\ 1 & \text{if} \quad \bar{x} > 0.5 \end{cases}$$

(c) [4pt] Find the probability mass function of $\hat{\theta}_{MLE}$.
Solution:

If we denote with θ_0 the *true* value of the parameter θ, we have:

$$\mathbb{P}(\hat{\theta}_{MLE} = t|\theta_0) = \begin{cases} \Phi((-1-\theta_0)\sqrt{n}/\sigma) & \text{if} \quad t=-2, \\ \Phi((0.5-\theta_0)\sqrt{n}/\sigma) - \Phi((-1-\theta_0)\sqrt{n}/\sigma) & \text{if} \quad t=0, \\ 1-\Phi((0.5-\theta_0)\sqrt{n}/\sigma) & \text{if} \quad t=1, \end{cases}$$

for $\theta_0 \in \{-2,0,1\}$ and where $\Phi(\cdot)$ denote the CDF of the standard normal distribution.

(d) [3pt] Is $\hat{\theta}_{MLE}$ a biased estimator?
Solution:
$\hat{\theta}_{MLE}$ is a biased estimator. In fact, in case $\theta_0 = -2$, we have:

$$\mathbb{E}(\hat{\theta}_{MLE}|\theta_0 = -2) = -2\Phi(\sqrt{n}/\sigma) + 1 - \Phi(2.5\sqrt{n}/\sigma) = -2\Phi(\sqrt{n}/\sigma) + \Phi(-2.5\sqrt{n}/\sigma) > -2\Phi(\sqrt{n}/\sigma) > -2$$

because $0 < \Phi(\cdot) < 1$.