Statistiek (WISB361)

Sketch of solutions for the 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 $X_1, \ldots, X_n$ 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 $0 \leq \alpha \leq 1$.

(a) [7pt] Find the variance $\mathrm{Var}(W(\alpha))$ of $W(\alpha)$.
Solution:
In normal samples, $\bar{X}$ and S are independent (see Chapter 6), hence:

$$\mathrm{Var}(W(\alpha)) = \alpha^2 \mathrm{Var}(T_1) + (1-\alpha)^2 \mathrm{Var}(T_2).$$

(b) [3pt] Find the mean squared error (MSE) of $W(\alpha)$ in terms of α, $\mathrm{Var}(T_1)$ and $\mathrm{Var}(T_2)$.
Solution:
($W(\alpha)$ is an unbiased estimator of θ. Hence, $\mathrm{MSE}(W(\alpha) = \mathrm{Var}(W(\alpha))$, which is found in point a).

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

$$\partial_\alpha \mathrm{MSE}(\alpha) = 2\alpha \mathrm{Var}(T_1) - 2(1-\alpha)\mathrm{Var}(T_2) = 0$$

Hence

$$\hat{\alpha} = \frac{\mathrm{Var}(T_2)}{\mathrm{Var}(T_1) + \mathrm{Var}(T_2)},$$

since

$$\partial_\alpha^2 \mathrm{MSE}(\alpha) = 2[\mathrm{Var}(T_1) + \mathrm{Var}(T_2)] > 0.$$

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

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

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

$$\hat{\alpha} = \frac{\mathrm{Var}(T_2)}{\mathrm{Var}(T_1) + \mathrm{Var}(T_2)} \approx \frac{\frac{\theta^2}{2n}}{\frac{2\theta^2}{2n} + \frac{\theta^2}{2n}} = 1/3$$

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

   (a) [7pt] $\theta > 0$
   Solution:
   We have:
   $$\theta > 0 \implies X_i \overset{\text{i.i.d.}}{\sim} \text{Unif}(\theta, 2\theta)$$
   and the likelihood is:
   $$lik(\theta) = \theta^{-n}\mathbf{1}(\theta \leq X_{(1)} \leq X_{(n)} \leq 2\theta) = \theta^{-n}\mathbf{1}(X_{(n)}/2 \leq \theta \leq X_{(1)}).$$
   Since the likelihood is decreasing in θ, one has:
   $$\hat{\theta}_{MLE} = \frac{1}{2}X_{(n)}$$
   provided that $X_{(n)}/2 \leq X_{(1)}$.

   (b) [5pt] $\theta < 0$
   Solution:
   We have:
   $$\theta < 0 \implies X_i \overset{\text{i.i.d.}}{\sim} \text{Unif}(\theta, 0)$$
   and the likelihood is:
   $$lik(\theta) = (-\theta)^{-n}\mathbf{1}(\theta \leq X_{(1)})$$
   Since the likelihood is increasing in θ, we have:
   $$\hat{\theta}_{MLE} = X_{(1)}$$

   (c) [5pt] $\theta \neq 0$
   Solution:
   Notice that if $\theta > 0$, $X_i \overset{\text{i.i.d.}}{\sim} \text{Unif}(\theta, 2\theta)\, \forall i \implies X_i > 0, \forall i$. Moreover, if $\theta < 0$, $X_i \overset{\text{i.i.d.}}{\sim} \text{Unif}(\theta, 0)\, \forall i \implies X_i < 0$. But when $\theta > 0$ we found that $\hat{\theta} = \frac{1}{2}X_{(n)}$ and when $\theta < 0$, we have $\hat{\theta} = X_{(1)}$ Hence,
   $$\hat{\theta}_{\text{MLE}} = \begin{cases} \frac{1}{2}X_{(n)} & if\ x_i > 0 \\ X_{(1)} & if\ x_i < 0 \end{cases}$$

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

   (d) [4pt] Calculate the MLE of θ.
   Solution:
   This time we have:
   $$\begin{aligned} lik(\theta) &= \frac{1}{\theta^n\,(b-a)^n}\mathbf{1}(a\theta \leq X_{(1)} \leq X_{(n)} \leq b\theta) \\ &= \frac{1}{\theta^n\,(b-a)^n}\mathbf{1}(X_{(n)}/b \leq \theta \leq X_{(1)}/a). \end{aligned}$$
   Since $lik(\theta)$ is decreasing in θ, then $\hat{\theta}_{\text{MLE}} = X_{(n)}/b$, provided that $X_{(n)}/b \leq X_{(1)}/a$.

   (e) [2pt] Find a two–dimensional sufficient statistics for θ.
   Solution:
   Since
   $$lik(\theta) = \frac{1}{\theta^n\,(b-a)^n}\mathbf{1}(a\theta \leq X_{(1)} \leq X_{(n)} \leq b\theta)$$
   if we pose $h(x) = \frac{1}{(b-a)^n}$, the rest of the expression is a function of only θ and of $T(Y_1, Y_2) = (X_{(1)}, X_{(n)})$. Hence, by the Fisher–Neyman factorization theorem, we have that a sufficient two–dimensional statistic is: $T = (X_{(1)}, X_{(n)})$

3. Let $\mathbf{y} = \{y_1, \ldots, y_n\}$ the realization of the random vector $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$ with independent components and such that $Y_i \sim N(\theta x_i, 1+x_i^2)$, with $i = 1, \ldots, n$, 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.
Solution:
For the independent random variables $Y_1, \ldots Y_n$, with $Y_i \sim N(\mu_i, \sigma_i^2)$, with $\mu_i = \theta\, x_i$, $\sigma_i^2 = 1 + x_i^2$, the likelihood is proportional to (non considering the factors not depending on θ):

$$\begin{aligned} lik(\theta) &\propto \exp\left\{-(1/2)\sum_{i=1}^{n}(1/\sigma_i^2)(y_i-\mu_i)^2\right\} \\ &\propto \exp\left\{-(1/2)\sum_{i=1}^{n}(\mu_i^2/\sigma_i^2)\right\}\exp\left\{(1/2)\sum_{i=1}^{n}(2y_i\,\mu_i/\sigma_i^2)\right\} \\ &\propto \exp\left\{-(\theta^2/2)\sum_{i=1}^{n}x_i^2/(1+x_i^2)\right\}\exp\left\{\theta\sum_{i=1}^{n}y_i\,x_i/(1+x_i^2)\right\} \\ &= \exp(-\theta^2/2)\exp\left\{\theta\sum_{i=1}^{n}y_i\,x_i/(1+x_i^2)\right\} \end{aligned}$$

Therefore, the likelihood ratio can be rewritten as:

$$\Lambda = \frac{lik(\theta=0)}{lik(\theta=1)} = \exp(1/2)\exp\left[-t(\mathbf{y})\right]$$

so that rejecting H_0 for small value of Λ is equivalent to reject for large values of $t(\mathbf{y})$. Hence by Neyman–Pearson Lemma, the choice of the rejection region R_c maximizes the power.

(b) [3pt] Find the distribution of $t(\mathbf{Y})$ and derive an expression for the power of the test.
Solution:
$t(\mathbf{Y})$ is normally distributed since it is a linear combination of independent normal random variables. $Y_i \sim N(\theta x_i, 1+x_i^2)$, so that $x_i\, Y_i/(1+x_i^2) \sim N(\theta x_i^2/(1+x_i^2), x_i^2/(1+x_i^2))$. Therefore, $t(\mathbf{Y}) \sim N(\theta, 1)$. Under H_1, $t(\mathbf{Y}) \sim N(1,1)$, so that

$$\pi = \mathbb{P}(t(\mathbf{Y}) > c | H_1) = 1 - \Phi(c-1)$$

(c) [7pt] Show that $t(\mathbf{Y})$ is the maximum likelihood estimator (MLE) of θ. Is this estimator unbiased?
Solution:
The log–likelihood can be written as:

$$l(\theta)) = -\theta^2/2 + \theta\sum_{i=1}^{n}(y_i x_i/(1+x_i^2))$$

so that

$$\partial_\theta l(\theta) = -\theta + \sum_{i=1}^{n}(y_i\, x_i/(1+x_i^2))$$

and $\hat{\theta}_{MLE} = t(\mathbf{Y})$. Moreover since $t(\mathbf{Y}) \sim N(\theta, 1)$, $\hat{\theta}$ is 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.
Solution:
We perform a *paired* t–test to compare the population means $\overline{X}$ and $\overline{Y}$: each observations X_i in one sample indeed is paired with observation Y_i in the other sample. We define $D_i := X_i - Y_i$ and under the normality assumption we have $D_i \sim N(\mu_D, \sigma_D^2),\ i = 1, \ldots, 12$. Hence, we want to test:

$$\begin{cases} H_0: & \mu_D = 0, \\ H_1: & \mu_D < 0. \end{cases}$$

at $\alpha = 0.05$ level of significance. The test statistics:

$$T = \sqrt{12}\frac{\bar{D}}{\sqrt{S_D^2}}$$

assumes the value $t = -2.34$. Since the critical value for the one–sided t–test with 11 degrees of freedom is -1.796, we can reject H_0 with 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.
Solution:
We have that:

$$\mathbb{P}(D < 0) = \Phi\big(\frac{0 - \mu_D}{\sqrt{\mathrm{Var}(D)}}\big) \approx \Phi(0.7) \approx 0.7.$$

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.
Solution:
By definition, the least squares estimators are the values α and β that minimize the function:

$$S(\alpha, \beta) = \sum_{x=1}^{n} [Y_x - (\alpha\, x + \beta\, x^2)]^2.$$

The equation:

$$\frac{\partial S}{\partial \alpha} = -2 \sum_{x=1}^{n} x[Y_x - (\alpha\, x + \beta\, x^2)] = 0$$

gives that:

$$\sum_{x=1}^{n} x\, Y_x - \alpha S_2 - \beta S_3 = 0$$

or

$$\hat{\alpha} = \frac{\sum_{x=1}^{n} xY_x - \hat{\beta} S_3}{S_2}$$

Similarly:

$$\frac{\partial S}{\partial \beta} = -2 \sum_{x=1}^{n} x^2[Y_x - (\alpha\, x + \beta\, x^2)] = 0$$

implies

$$\sum_{x=1}^{n} x^2\, Y_x - \alpha S_3 - \beta S_4 = 0$$

that can be rewritten as:

$$\hat{\beta} = \frac{\sum_{x=1}^{n} x^2\, Y_x - S_2^{-1} S_3 \sum_{x=1}^{n} x\, Y_x}{S_4 - S_2^{-1} S_3^2}$$

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

$$\begin{aligned} \mathbb{E}(\hat{\beta}) &= \frac{\sum_{x=1}^{n} x^2\, (\alpha\, x + \beta\, x^2) - S_2^{-1} S_3 \sum_{x=1}^{n} x\, (\alpha\, x + \beta\, x^2)}{S_4 - S_2^{-1} S_3^2} \\ &= \frac{\alpha S_2\, S_3 + \beta S_2\, S_4 - \alpha S_2\, S_3 - \beta S_3^2}{S_2\, S_4 - S_3^2} = \beta \end{aligned}$$

As regards the variance:

$$\begin{aligned} \mathrm{Var}(\hat{\beta}) &= \mathrm{Var}\left\{ \frac{\sum_{x=1}(S_2\, x^2 - S_3\, x) Y_x}{S_2 S_4 - S_3^2} \right\} \\ &= \frac{\sigma^2 \sum_{x=1}^{n} (S_2 x^2 - S_3 x)^2}{(S_2 S_4 - S_3^2)^2} \\ &= \frac{\sigma^2 \sum_{x=1}^{n} (S_2^2 x^4 - 2 S_2 S_3 x^3 + S_3^2 x^2)}{(S_2 S_4 - S_3^2)^2} \\ &= \frac{\sigma^2 [S_2^2 S_4 - 2 S_2 S_3^2 + S_2 S_3^2)]}{(S_2 S_4 - S_3^2)^2} \\ &= \frac{\sigma^2 [S_2^2 S_4 - S_2 S_3^2)]}{(S_2 S_4 - S_3^2)^2} \\ &= \frac{\sigma^2 S_2 [S_2 S_4 - S_3^2)]}{(S_2 S_4 - S_3^2)^2} \\ &= \frac{\sigma^2 S_2}{(S_2 S_4 - S_3^2)} \end{aligned}$$

(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 $\sigma^2 = 1$.

Solution:

Since $\hat{\beta}$ is a linear combination of mutually independent normal random variables, it follows that it is itself normally distributed. Thus, being $\frac{\hat{\beta}-\beta}{\sqrt{\mathrm{Var}(\hat{\beta})}} \sim N(0,1)$ since $\mathbb{E}(\hat{\beta}) = \beta$, this gives an exact 95% confidence interval for β of the form:

$$\hat{\beta} \pm 1.96\sqrt{\frac{S_2\sigma^2}{(S_2S_4 - S_3^2)}}$$

Since $n = 4$, $S_2 = 30$, $S_3 = 100$ and $S_4 = 354$. Thus, since $\sigma^2 = 1$, the CI is: $(1.57, 2.43)$.