Statistiek (WISB263)

Sketch of Solutions (Final Exam)

January 30, 2017

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 100.
Points distribution: 25-20-30-25

1. Given two parameters $a > 0$ and $k > 0$, let $\mathbf{X} = \{X_1, \dots, X_n\}$ be a random sample of n i.i.d. observations sampled from the random variable X with density function:

$$f_X(x; a, k) := \begin{cases} k\, e^{-k\,(x-a)} & x \geq a, \\ 0 & x < a \end{cases}$$

(a) (8pt) Find sufficient statistics for a, k and for the couple (a, k).
Solution: We can write the likelihood of the sample as:

$$L(\boldsymbol{X}; a, k) = k^n e^{-k \sum_{i=1}^n X_i} e^{nka} \mathbf{1}(X_{(1)} \geq a)$$

By the factorization theorem, we have the following sufficient statistics:

$$(X_{(1)}, \sum_{i=1}^n X_i) \text{ for } (a, k) \quad [\text{e.g. } h(\boldsymbol{X}) = 1,\ g(T(\boldsymbol{X}), k, a) = L(\boldsymbol{X}; a, k)]$$

$$X_{(1)} \text{ for } a, \quad [\text{e.g. } h(\boldsymbol{X}) = k^n e^{-k \sum_{i=1}^n X_i},\ g(T(\boldsymbol{X}), k, a) = e^{nka} \mathbf{1}(X_{(1)} \geq a)]$$

$$\sum_{i=1}^n X_i \text{ for } k, \quad [\text{e.g. } h(\boldsymbol{X}) = e^{nka} \mathbf{1}(X_{(1)} \geq a),\ g(T(\boldsymbol{X}), k, a) = k^n e^{-k \sum_{i=1}^n X_i}]$$

(b) (5pt) Determine, in case it exists, the maximum likelihood estimator of a in case k is known.
Solution:
Since

$$L(\boldsymbol{X}; a, k) \propto e^{nka} \mathbf{1}(X_{(1)} \geq a)$$

the likelihood is null for $X_{(1)} < a$ and increasing in a for $X_{(1)} \geq a$, it follows that $\hat{a}_{MLE} = X_{(1)}$.

(c) (5pt) Determine, in case it exists, the maximum likelihood estimator of k in case a is known.
Solution:
Provided that $X_{(1)} \geq a$, it follows that $na - \sum_{i=1}^n X_i \leq 0$. Therefore

$$L(\boldsymbol{X}; a, k) \propto e^{k(na - \sum_{i=1}^n X_i) + n \log k} =: f(k; a)$$

and for a fixed a, we have to maximize in k the positive, continuous and differentiable function $\log f(k; a)$. Since there is only one critical point and since $\frac{d^2}{dk^2} \log f(k) < 0$, it follows that:

$$\hat{k}_{MLE} = \frac{n}{\sum_{i=1}^n X_i - na}$$

(d) (7pt) Determine, in case it exists, the maximum likelihood estimator of the couple (a, k).
Solution:
Let us consider $k > 0$, $a \leq X_{(1)}$:

$$L(\boldsymbol{X}; a, k) = k^n e^{-k \sum_{i=1}^n X_i} e^{nka} \leq k^n e^{-k \sum_{i=1}^n X_i + knX_{(1)}} = e^{k(nX_{(1)} - \sum_{i=1}^n X_i) + n \log k} = f(k; X_{(1)}) \leq f(\tilde{k}; X_{(1)})$$

where $\tilde{k} = \frac{n}{\sum_{i=1}^n X_i - nX_{(1)}}$. Therefore $\left(\frac{n}{\sum_{i=1}^n X_i - nX_{(1)}}, X_{(1)}\right)$ is the MLE of (k, a).

2. We consider the following three random samples of size 100:

$$\mathbb{X}_i := \{X_{i,1}, X_{i,2}, \dots X_{i,100}\},$$

with $i \in \{1,2,3\}$. Each sample $\mathbb{X}_i$ consists of i.i.d. normal random variables, such that $X_{i,j} \sim N(50, \sigma_i^2)$ for any $j \in \{1,\dots,100\}$. Moreover the samples are independent (i.e. $X_{i,j} \perp X_{\ell m}$, for any $i \neq \ell$). We want to test:

$$\begin{cases} H_0: & \sigma_1^2 = \sigma_2^2 = \sigma_3^2, \\ H_1: & \text{the variances are not equal.} \end{cases}$$

(a) [10pt] Show that the Generalized Likelihood Ratio Test (GLRT) statistic Λ is such that:

$$-2\log\Lambda = 300\log\left(\frac{1}{3}\sum_{i=1}^{3} s_i^2\right) - 100\sum_{i=1}^{3}\log s_i^2$$

where $s_i^2 := 1/100\sum_{j=1}^{100}(X_{i,j}-50)^2$, with $i \in \{1,2,3\}$.
Solution:
The likelihood can be written as:

$$L(\sigma_1^2,\sigma_2^2,\sigma_3^2) = \prod_{i=1}^{3}\prod_{j=1}^{100}\frac{1}{(2\pi\sigma_i^2)^{1/2}}\exp\left(-\frac{1}{2\sigma_i^2}(X_{i,j}-50)^2\right) = \frac{C}{(\sigma_1^2\,\sigma_2^2\,\sigma_3^2)^{50}}e^{-50\left(\frac{S_1^2}{\sigma_1^2}+\frac{S_2^2}{\sigma_2^2}+\frac{S_3^2}{\sigma_3^2}\right)}$$

and the log–likelihood is:

$$\ell(\sigma_1^2,\sigma_2^2,\sigma_3^2) = \log(C) - 50\sum_{i=1}^{3}\log\sigma_i^2 - 50\sum_{i=1}^{3}\frac{S_i^2}{\sigma_i^2}$$

By definition of GLRT statistic, we have:

$$\log\Lambda = \max_{\sigma^2}\ell(\sigma^2,\sigma^2,\sigma^2) - \max_{\sigma_1^2,\sigma_2^2,\sigma_3^2}\ell(\sigma_1^2,\sigma_2^2,\sigma_3^2)$$

By standard calculations we find that:

$$\hat{\sigma}^2 := \text{argmax}_{\sigma^2}\ell(\sigma^2,\sigma^2,\sigma^2) = \frac{1}{3}\sum_{i=1}^{3}S_i^2$$

and that

$$(\hat{\sigma}_1^2,\hat{\sigma}_2^2,\hat{\sigma}_3^2) := \text{argmax}_{\sigma_1^2,\sigma_2^2,\sigma_3^2}\ell(\sigma_1^2,\sigma_2^2,\sigma_3^2) = \left(S_1^2,S_2^2,S_3^2\right)$$

Hence

$$\begin{aligned} -2\log\Lambda &= 2(\ell(\hat{\sigma}_1^2,\hat{\sigma}_2^2,\hat{\sigma}_3^2) - \ell(\hat{\sigma}^2,\hat{\sigma}^2,\hat{\sigma}^2)) \\ &= 300\log\left(\frac{1}{3}\sum_{i=1}^{3}S_i^2\right) - 100\sum_{i=1}^{3}\log S_i^2 \end{aligned}$$

(b) [10pt] If the collected data $\mathbf{x}_i = \{x_{i,1},\dots,x_{i,100}\}$, with $i \in \{1,2,3\}$, are such that:

$$\textstyle\sum_{j=1}^{100} x_{1,j} = 5040, \qquad \sum_{j=1}^{100} x_{2,j} = 4890, \qquad \sum_{j=1}^{100} x_{3,j} = 4920,$$

$$\textstyle\sum_{j=1}^{100} x_{1,j}^2 = 264200, \quad \sum_{j=1}^{100} x_{2,j}^2 = 250000, \quad \sum_{j=1}^{100} x_{2,j}^2 = 251700$$

perform a GLRT at $\alpha = 0.05$ level of significance (you can consider the sample size $n = 100$ large enough for applying large sample results).
Solution:
By asymptotic results we have that $-2\log\Lambda \approx \chi_2^2$. For the given data:

$$-2\log\Lambda = 0.283 < \chi_2^2(0.05)$$

so that ca cannot reject the null hypothesis at the 5% level of significance.

3. The life times (in hours) of $n = 30$ batteries have been measured from a company interested in the performances of a new product. In this way, a sample $\mathbb{X} = \{X_1, \dots X_{30}\}$ of i.i.d. random variable X_j, representing the life time of the j–th battery, has been collected. In the following table the empirical cumulative distribution function $\hat{F}_{30}(x)$ (i.e. $\hat{F}_n(x) = 1/n \sum_{j=1}^n \mathbf{1}(X_j \le x)$) is reported:

x (in hours)	1	2	4	6	8	11	13	27	29	42
$\hat{F}_{30}(x)$	7/30	12/30	16/30	20/30	23/30	26/30	27/30	28/30	29/30	1

(a) [6pt] Determine an estimator of the probability that the battery produced lasts more than 9 hours (i.e. $\mathbb{P}(X > 9)$).
Solution:
We want to estimate $p := \mathbb{P}(X > 9)$. A non–parametric unbiased estimator is given by:

$$T = 1 - \hat{F}_{30}(9) = 1 - \hat{F}_{30}(8) = 7/30$$

(b) [8pt] Derive an approximated 95% confidence interval for the probability that the battery produced lasts more than 9 hours.
Solution:
Since

$$T \approx N(p, p(1-p)/30)$$

A 95% CI for p is given by $(0.082, 0.394)$.

Due to previous statistical analyses performed on similar batteries, we can assume now that the sample is a collection of 30 i.i.d. exponential random variable with expected value θ (i.e. $X_i \sim \text{Exp}(1/\theta)$).

(c) [8pt] Under these parametric assumptions, calculate the maximum likelihood estimator of the probability that the battery produced lasts more than 9 hours.
Solution:
We want to estimate $p(\theta) := \mathbb{P}_\theta(T > 9) = e^{-9/\theta}$. Since for an exponential distribution $\hat{\theta}_{MLE} = \bar{X}$, by the invariance principle, it follows that $\hat{p}_{MLE} = e^{-9/\bar{X}}$.

(d) [8pt] If we denote with $p(\theta)$ the probability that the battery produced lasts more than 9 hours, propose a test for testing the hypotheses:

$$\begin{cases} H_0: & p = 0.32 \\ H_1: & p = 0.16. \end{cases}$$

at the α level of significance.
Solution:
Since the H_0 and H_1 are simple hypotheses, we can use the Neyman Pearson Lemma in order to construct the most powerful test with the α level of significance. Note that $p = 0.16$ iff $\theta = -9/\log 0.16 =: \theta_0$ and $p = 0.32$ iff $\theta = -9/\log 0.32 =: \theta_1$.The LRT statistics can be written as:

$$\Lambda = \frac{L(\theta_0)}{L(\theta_1)} = \exp\left(n\bar{X}(1/\theta_1 - 1/\theta_0)\right)$$

so that the test rejects for $\bar{X} < k$. By the CLT, $\bar{X} \approx N(\theta_0, \theta_0^2/30)$, so that the rejection region can be determined.

4. Let the independent random variables $Y_1, Y_2, \dots, Y_n$ be such that we have the following linear model:

$$Y_i = \alpha + \beta x_i + \epsilon_i$$

for $i = 1, \dots, n$, where ϵ_i are i.i.d. normal random variables such that $\epsilon_i \sim N(0, \sigma^2)$. Let $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ be the model in the matrix formalism. After we collected a sample of size $n = 42$, we have that:

$$(\mathbf{X}^\top\mathbf{X})^{-1} = \begin{pmatrix} 0.03 & -0.015 \\ -0.015 & 0.04 \end{pmatrix}$$

Furthermore, we know that the least squares estimate is $\hat{\boldsymbol{\beta}}^\top = (\hat{\beta}_0, \hat{\beta}_1) = (1.90, 0.65)$ and that the residual sum of squares $\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 = 160$.

(a) [8pt] Compute the 95% confidence intervals for β_0 and β_1
Solution:
We know that:

$$T_0 := \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\hat{\sigma}^2 (\mathbf{X}^\top \mathbf{X})^{-1}_{1,1}}} \sim t(42-2) = t(40)$$

and

$$T_1 := \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{\sigma}^2 (\mathbf{X}^\top \mathbf{X})^{-1}_{2,2}}} \sim t(42-2) = t(40)$$

where $\hat{\sigma}^2 = \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2/40$. With the given data:

$$\hat{\sigma}^2 = 160/40 = 4$$

So that a 95% CI for β_0:

$$1.90 \pm t_{0.975}(40)\sqrt{\hat{\sigma}^2 (\mathbf{X}^\top \mathbf{X})^{-1}_{1,1}} = 1.90 \pm 2.021\sqrt{0.12} = [1.20, 2.60]$$

β_1:

$$0.65 \pm t_{0.975}(40)\sqrt{\hat{\sigma}^2 (\mathbf{X}^\top \mathbf{X})^{-1}_{2,2}} = 0.65 \pm 2.021\sqrt{0.16} = [-0.16, 1.46],$$

and a 99% CI for β_0:

$$1.90 \pm t_{0.995}(40)\sqrt{\hat{\sigma}^2 (\mathbf{X}^\top \mathbf{X})^{-1}_{1,1}} = 1.90 \pm 2.74\sqrt{0.12} = [0.95, 2.85]$$

for β_1:

$$0.65 \pm t_{0.995}(40)\sqrt{\hat{\sigma}^2 (\mathbf{X}^\top \mathbf{X})^{-1}_{2,2}} = 0.65 \pm 2.74\sqrt{0.16} = [-0.45, 1.75],$$

(b) [10pt] Consider the test:

$$\begin{cases} H_0: & \beta_0 = 2, \\ H_1: & \beta_0 \neq 2. \end{cases}$$

Will H_0 be rejected at a significance level of 5%? And at a significance level of 1%?
Solution:
By duality of two sided test and CI, from the previous point, we do not reject the H_0 at both 5% and 1% since $2 \in$ CI in both cases.

(c) [7pt] Under the previous H_0, it holds that $\mathbb{P}(\hat{\beta}_0 > 1.90) = 0.61$ and that $\mathbb{P}(\hat{\beta}_0 < 1.90) = 0.39$. For which values of the significance level α, the null hypothesis H_0 will be rejected with the given data?
Solution:
Since from the previous points, under H_0, the distribution of $\hat{\beta}_0$ is symmetric around 2, we have that the p value of the two sided test is $p = 2\mathbb{P}(\hat{\beta}_0 < 1.90) = 0.78$. So that H_0 will be rejected for any $\alpha > 0.78$.