Statistiek (WISB263)

Sketch of Solutions for the Resit Exam

April 19, 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: 32-20-26-22

1. Let $\mathbf{X} = \{X_1, \ldots, X_n\}$ be a random sample of n i.i.d. Poisson random variables with parameter λ.

   (a) (8pt) Find the maximum likelihood for λ and its asymptotic sampling distribution.
   Solution:
   The log–likelihood can be written as:

   $$\ell(\mathbf{X};\lambda) = -n\lambda + (\sum_{i=1}^n X_i)\log\lambda - \log\left(\prod_{i=1}^n X_i!\right)$$

   so that

   $$\dot{\ell}(\mathbf{X};\lambda) = -n + \frac{\sum_{i=1}^n X_i}{\lambda}$$

   and

   $$\ddot{\ell}(\mathbf{X};\lambda) = -\frac{\sum_{i=1}^n X_i}{\lambda^2} < 0$$

   so that the MLE of λ is

   $$\hat{\lambda} = \frac{\sum_{i=1}^n X_i}{n} = \overline{X}_n$$

   By CLT,

   $$\frac{\sqrt{n}(\overline{X}_n - \lambda)}{\sqrt{\lambda}} \xrightarrow{\mathcal{D}} N(0,1)$$

   as $n \to \infty$. Therefore:

   $$\hat{\lambda} \approx N(\lambda, \lambda/n)$$

   (b) (8pt) Find the maximum likelihood estimator for the parameter $\mu = e^{-\lambda}$.
   Solution:
   By the invariance principle the MLE of μ is:

   $$\hat{\mu} = e^{-\hat{\lambda}} = e^{-\overline{X}_n}$$

   Suppose now that, rather than observing the actual values of the random variables X_i, we are just able to register whether they are null or positive. More precisely, only the events $X_i = 0$ or $X_i > 0$ for $i = 1, \ldots, n$ are observed.

   (c) (8pt) Find the maximum likelihood for λ for these new observations.
   Solution:
   Our sample now can be seen as n realizations of a Bernoulli variable Y with parameter $p = e^{\lambda}$, i.e. $\mathbb{P}(Y = 0) = p$ and $\mathbb{P}(Y = 1) = 1 - p$. Hence,

   $$\ell(\mathbf{X};\lambda) = (n - \sum_{i=1}^n Y_i)\log p + \sum_{i=1}^n Y_i \log(1-p)$$

   By standard calculations we have that the MLE of p is:

   $$\hat{p} = (n - \sum_{i=1}^n Y_i)/n$$

Therefore, by the invariance principle, the MLE of λ is:

$$\hat{\lambda} = -\log\left((n - \sum_{i=1}^{n} Y_i)/n\right)$$

that exists only for $n \neq \sum_{i=1}^{n} Y_i$, i.e. there is at least one null observation.

(d) (8pt) When does the maximum likelihood estimator not exist? Assuming that the true value of λ is λ_0, compute the probability that the maximum likelihood estimator does not exist.

Solution:

The MLE exists for $n \neq \sum_{i=1}^{n} Y_i$. Therefore we have to calculate the probability:

$$\mathbb{P}_{\lambda_0}\left(n = \sum_{i=1}^{n} Y_i\right) = \prod_{i=1}^{n} \mathbb{P}_{\lambda_0}(Y_i = 1) = (1 - e^{-\lambda_0})^n$$

2. Let $\mathbf{X} = \{X_1, \dots, X_n\}$ be a random sample of n i.i.d. random variables with densities:

$$f_X(x;\theta) = \begin{cases} \frac{\theta^3}{2} x^2 e^{-\theta x} & \text{if } x > 0, \\ 0 & \text{otherwise} \end{cases}$$

with $\theta > 0$ is an unknown parameter. Moreover, consider another random sample $\mathbf{Y} = \{Y_1, \dots, Y_n\}$ of n i.i.d. random variables with densities:

$$f_Y(y;\mu) = \begin{cases} \frac{\mu^3}{2} y^2 e^{-\mu y} & \text{if } y > 0, \\ 0 & \text{otherwise} \end{cases}$$

with $\mu > 0$ is another unknown parameter. We further assume that the two sample are independent (i.e. $X_i \perp Y_j$, for all i, j).

(a) [10pt] Find the Generalized Likelihood Ratio Test (GLRT) statistic for testing:

$$\begin{cases} H_0: & \theta = \mu, \\ H_1: & \theta \neq \mu. \end{cases}$$

Solution:

Let us denote with:

$$\mathbf{V} = \{X_1, \dots, X_n, Y_1, \dots Y_n\}$$

the sample of size $2n$ obtained pooling together the samples $\mathbf{X}$ and $\mathbf{Y}$. The log–likelihood corresponding to $\mathbf{V}$ is:

$$lik(\mathbf{V};\theta,\mu) = lik(\mathbf{X};\theta) lik(\mathbf{Y};\mu) = \frac{\theta^{3n}\mu^{3n}}{2^{2n}} e^{-\theta \sum_{i=1}^{n} X_i} e^{-\mu \sum_{i=1}^{n} Y_i} \prod_{i=1}^{n} X_i^2 Y_i^2$$

The GLRT can be written as:

$$\Lambda(\mathbf{V}) = \frac{\sup_{\theta_0} lik(\mathbf{V};\theta_0,\theta_0)}{\sup_{\theta,\mu} lik(\mathbf{X};\theta) lik(\mathbf{Y};\mu)} = \frac{lik(\mathbf{V};\hat{\theta}_0,\hat{\theta}_0)}{lik(\mathbf{X};\hat{\theta})\, lik(\mathbf{Y};\hat{\mu})}$$

where the hat denotes the MLE. Since

$$\partial_\theta \ell(\mathbf{X};\theta) = \frac{3n}{\theta} - \sum_{i=1}^{n} X_i$$

and

$$\partial^2_{\theta\theta} \ell(\mathbf{X};\theta) = -\frac{3n}{\theta^2} < 0$$

the MLE of θ is $\hat{\theta} = \frac{3n}{\sum_{i=1} X_i}$. Analogously, we have $\hat{\mu} = \frac{3n}{\sum_{i=1} Y_i}$ and $\hat{\theta}_0 = \frac{6n}{\sum_{i=1} Y_i + \sum_{i=1}^{n} X_i}$. Hence,

$$\Lambda(\mathbf{V}) = \frac{\hat{\theta}_0^{6n} \exp(-\hat{\theta}_0 \sum_{i=1}^{n}(X_i + Y_i))}{\hat{\theta}^{3n}\hat{\mu}^{3n} \exp(-\hat{\theta} \sum_{i=1}^{n} X_i - \hat{\mu} \sum_{i=1}^{n} Y_i)} = \frac{\hat{\theta}_0^{6n}}{\hat{\theta}^{3n}\hat{\mu}^{3n}}$$

Let us define now the following statistic:

$$T := \frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} X_i + \sum_{j=1}^{n} Y_j}$$

(b) [10pt] Show that the GLRT rejects H_0 if $T(1-T) < k$, for a suitable constant k.

Solution:
The GLRT statistics reject for $\Lambda(\mathbf{V}) < c$, for a suitable constant c. Then

$$\begin{aligned}
\Lambda(\mathbf{V}) &= \frac{\hat{\theta}_0^{6n}}{\hat{\theta}^{3n}\hat{\mu}^{3n}} = \frac{\left(\frac{6n}{\sum_{i=1} Y_i + \sum_{i=1}^{n} X_i}\right)^{6n}}{\left(\frac{3n}{\sum_{i=1}^{n} X_i}\right)^{3n}\left(\frac{3n}{\sum_{i=1}^{n} Y_i}\right)^{3n}} = 2^{6n}\frac{1}{\left(\frac{\sum_{i=1}^{n}(Y_i+X_i)}{\sum_{i=1}^{n} X_i}\right)^{3n}\left(\frac{\sum_{i=1}^{n}(Y_i+X_i)}{\sum_{i=1}^{n} Y_i}\right)^{3n}} \\
&= 2^{6n}\frac{1}{\left(\frac{1}{T}\right)^{3n}\left(\frac{1}{1-T}\right)^{3n}} = 2^{6n}\left(T(1-T)\right)^{3n}
\end{aligned}$$

so that we reject for $T(1-T) < k$, with $k = c^{1/3n}/4$.

3. A company wants to monitor the efficiency of two employees in completing an assigned task. For this reason, the performances of two employees (denoted by **A** and **B**) were measured by recording the times needed to complete the assigned tasks. Hence, the following two samples have been collected:

$$\mathbf{x_A} = \{5.18, 13.43, 6.31, 3.18, 4.91, 11.07\},$$

$$\mathbf{x_B} = \{5.50, 18.16, 8.14, 9.14, 14.24, 10.72\}$$

where the duration of each task is measured in hours.

(a) [10pt] Perform a test at 10% of significance for testing the hypothesis that *employee* **A** *is faster than* **B**. Discuss critically the choice of the test used.

Solution:
Since we do not have any information on the distribution of the data, we can use the non–parametric Mann–Whitney for testing:

$$\begin{cases} H_0: & F_{\boldsymbol{A}}(x) = F_{\boldsymbol{B}}(x), \quad \forall x \\ H_1: & F_{\boldsymbol{A}}(x) \geq F_{\boldsymbol{B}}(x) \end{cases}$$

We have that the sum of ranks are $T_{\boldsymbol{A}} = 30$ and $T_{\boldsymbol{A}} = 48$. The critical value for the one–tailed test is 31, so that $T_{\boldsymbol{A}} < 31$, we can reject then H_0 at 10% of significance.

Suppose now that the time T needed by an employee for completing a task can be modeled by a continuous random variable with the following probability density function:

$$f_T(t;\theta) = \begin{cases} \frac{1}{2\theta\sqrt{t}}\, e^{-\frac{\sqrt{t}}{\theta}} & \text{if } t > 0, \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

with $\theta > 0$ an unknown parameter.

(b) [8pt] Given a sample $\mathbb{T} = \{T_1, \ldots, T_n\}$ of i.i.d random variables sampled from $f_T(t;\theta)$, determine the maximum likelihood estimator of the probability $\mathbb{P}_\theta(T > 7)$.
Solution:

$$\mathbb{P}_\theta(T > 7) = \int_7^\infty \frac{1}{2\theta\sqrt{t}}\, e^{-\frac{\sqrt{t}}{\theta}}\, dt = \int_{\sqrt{7}/\theta}^\infty e^{-y} dy = e^{-\sqrt{7}/\theta} \tag{2}$$

Hence, by invariance principle, the MLE of $\mathbb{P}_\theta(T > 7)$ is $e^{-\sqrt{7}/\hat{\theta}}$, where $\hat{\theta}$ is the MLE of the parameter θ. By standard calculations or by noting that $\sqrt{T} \sim \text{Exp}(\theta)$, we can derive that the MLE of θ is:

$$\hat{\theta} = \frac{\sum_{i=1}^n \sqrt{T_i}}{n} \tag{3}$$

so that the MLE of $\mathbb{P}_\theta(T > 7)$ is $\mathbb{P}_{\hat{\theta}}(T > 7)$.

(c) [8pt] Under the parametric model (1) for the random variable T and given the samples $\mathbf{x_A}$, $\mathbf{x_B}$, estimate the probability that the time needed by an employee for completing a task is larger than 7 hours, under the further assumption that 55% of the employees are similar to employee **A** and 45% to employee **B**.

Solution:
Using the samples $\mathbf{x_A}$ and $\mathbf{x_B}$, by (3) we find that following MLE estimates for the parameter θ:

$$\hat{\theta}_{\mathbf{A}} \simeq 2.63, \qquad \hat{\theta}_{\mathbf{B}} \simeq 3.26 \tag{4}$$

Therefore, by (2),(3) and (4), we have:

$$0.55\,\mathbb{P}_{\hat{\theta}_{\mathbf{A}}}(T > 7) + 0.45\,\mathbb{P}_{\hat{\theta}_{\mathbf{B}}}(T > 7) \simeq 0.42$$

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

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 (x_i - 3.5)_+ + \epsilon_i$$

for $i = 1, \ldots, n$, where ϵ_i are i.i.d. normal random variables such that $\epsilon_i \sim N(0, \sigma^2)$ and with $(y)_+$ we denoted the positive part of the real number y (i.e. $(y)_+ := \max(0, y)$). We collect the following sample of observations

$$\mathbf{y} = \{1, 2, 4, 5, 4, 3, 1\}$$

corresponding to the predictors:

$$\mathbf{x} = \{0, 1, 2, 3, 4, 5, 6\}$$

(a) [8pt] If we rewrite the linear model using the usual matrix formalism

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

write down the design matrix $\mathbf{X}$ of the linear model.
Solution:

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 2 & 0 \\ 1 & 3 & 0 \\ 1 & 4 & 0.5 \\ 1 & 5 & 1.5 \\ 1 & 6 & 2.5 \end{pmatrix}$$

(b) [6pt] Given that

$$(\mathbf{X}^\top \mathbf{X})^{-1} = \begin{pmatrix} 0.65 & -0.24 & 0.35 \\ -0.24 & 0.14 & -0.26 \\ 0.35 & -0.26 & 0.65 \end{pmatrix}$$

estimate the model coefficients and write down the fitted model.
Solution:
Since the LSE can be written an:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$$

we have:

$$\hat{\boldsymbol{\beta}} = (1.27, 1.54, -3.27)^{\top}$$

and

$$\hat{y} = 1.27 + 1.54\,x - 3.27\,(x - 3.5)_{+}$$

(c) [8pt] Calculate the prediction of the fitted model at $x = 4.5$. Assuming that the sum of squared residuals equals 7.8, calculate a 95% confidence interval for this prediction.
Solution:
The prediction is:

$$\hat{y} = 1.27 + 1.54 \cdot 4.5 - 3.27\,(4.5 - 3.5)_{+} = 4.93$$

The estimated covariance matrix of the fitted coefficient is:

$$\Sigma_{\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\beta}}} = s^2 (\mathbf{X}^{\top}\mathbf{X})^{-1}$$

with $s^2 = RSS/(7-3) = 7.8/4 = 1.95$. Then

$$\begin{aligned} \mathrm{Var}\hat{Y} &= \mathrm{Var}\hat{\beta}_0 + x^2\mathrm{Var}\hat{\beta}_1 + (x-3.5)_{+}^2\mathrm{Var}\hat{\beta}_2 + 2x\mathrm{Cov}(\hat{\beta}_0,\hat{\beta}_1) + 2(x-3.5)_{+}\mathrm{Cov}(\hat{\beta}_0,\hat{\beta}_2) + 2x(x-3.5)_{+}\mathrm{Cov}(\hat{\beta}_1,\hat{\beta}_2) \\ &= \Sigma_{1,1} + 4.5^2\Sigma_{2,2} + \Sigma_{3,3} + 9\Sigma_{1,2} + 2\Sigma_{1,3} + 9\Sigma_{2,3} \end{aligned}$$

Therefore a 95% CI for the prediction is:

$$4.93 \pm t_{4,0.024}\sqrt{\mathrm{Var}\hat{Y}} = 4.93 \pm 4.47$$