Statistiek (WISB361)

Retake exam

July 20, 2015

Schrijf uw naam op elk in te leveren vel. Schrijf ook uw studentnummer op blad 1.
The exam is open–book and the use of the calculator is permitted.
The maximum number of points is 100.
Points distribution: 25–20–15–20–20.

1. Suppose that $X_1, X_2, \ldots, X_n$, are independent random variables with

$$X_i \sim N(i\,\theta, 1)$$

for $i = 1, \ldots, n$.

(a) [7pt] Find the maximum likelihood estimator $\hat{\theta}_{MLE}$ of θ
Solution:
The likelihood can be written as:

$$lik(\theta) = f(x_1, \ldots, x_n|\theta) = \frac{1}{(2\pi)^{n/2}} \exp\left\{ -\frac{1}{2}\sum_{i=1}^{n}(x_i - i\theta)^2 \right\}$$

so that the log–likelihood $\ell(\theta)$ is:

$$\ell(\theta) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(x_i - i\theta)^2$$

The score equation for θ is:

$$\partial_\theta \ell(\theta) = \sum_{i=1}^{n} i(x_i - i\theta) = \sum_{i=1}^{n} i x_i - \theta \sum_{i=1}^{n} i^2$$

Hence, $\partial_\theta \ell(\theta) = 0$ iff

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

and, since $\partial^2_\theta \ell(\theta) = -\sum_{i=1}^{n} i^2 < 0$, we finally have:

$$\hat{\theta}_{MLE} = \frac{\sum_{i=1}^{n} i X_i}{\sum_{i=1}^{n} i^2}$$

(b) [7pt] Find the variance of $\hat{\theta}_{MLE}$.
Solution:
Since X_i are independent RV, we have:

$$\mathrm{Var}(\hat{\theta}_{MLE}) = \frac{1}{\left(\sum_{i=1}^{n} i^2\right)^2} \sum_{i=1}^{n} i^2 \mathrm{Var}(X_i) = \frac{1}{\left(\sum_{i=1}^{n} i^2\right)}$$

(c) [6pt] Compare the variance calculated in (b) with the Cramer–Rao lower bound for an unbiased estimator of θ. Is $\hat{\theta}_{MLE}$ an efficient estimator?
Solution:
The total Fisher information $I(\theta)$ is:

$$I(\theta) = -\mathbb{E}_\theta(\partial^2_\theta \ell(\theta))$$

Since

$$\partial_\theta^2 \ell(\theta) = -\sum_{i=1}^{n} i^2$$

the Cramer–Rao lower bound for an unbiased estimator T of θ is:

$$\mathrm{Var}(T) \geq \frac{1}{\sum_{i=1}^{n} i^2}$$

We have that $\hat{\theta}_{MLE}$ is unbiased:

$$\mathbb{E}(\hat{\theta}_{MLE}) = \frac{\sum_{i=1}^{n} i\mathbb{E}(X_i)}{\sum_{i=1}^{n} i^2} = \theta \frac{\sum_{i=1}^{n} i^2}{\sum_{i=1}^{n} i^2} = \theta$$

Since, the Cramer–Rao lower bound is attained, then $\hat{\theta}_{MLE}$ is an efficient estimator for θ.

Suppose that we have now another sample $Y_1, \ldots, Y_n$ of i.i.d. random variables $Y_i \sim N(\mu, 1)$, with $i = 1, \ldots, n$ and where $\mu \in \mathbb{R}$ is an unknown parameter. Suppose we do not observe the exact values of Y_i but only their signs, i.e., we only observe $Z_i = \mathrm{sgn}(Y_i)$ for $i = 1, \ldots, n$.

(d) [5pt] Obtain the maximum likelihood estimator of μ
Solution:
In we define

$$V_i := \frac{1 + Z_i}{2}$$

V_i are i.i.d. Bernoulli RV with parameter $p = \mathbb{P}(Y_i > 0) = \mathbb{P}(Y_i - \mu > -\mu) = 1 - \Phi(-\mu) = \Phi(\mu)$, where $\Phi(\cdot)$ is the CDF of a standard normal RV. Therefore, the likelihood can be written as:

$$lik(\mu) = \Phi(\mu)^{\sum_{i=1}^{n} V_i} (1 - \Phi(\mu))^{n - \sum_{i=1}^{n} V_i}$$

By differentiating the log–likelihood, one finds the usual MLE estimator for i.i.d. binomial observations:

$$\widehat{\Phi(\mu)}_{MLE} = \frac{1}{n} \sum_{i=1}^{n} V_i = \frac{1 + \bar{Z}_n}{2}$$

Bt the invariance principle:

$$\hat{\mu}_{MLE} = \Phi^{-1}(\frac{1 + \bar{Z}_n}{2})$$

2. Let us suppose to have only **one** observation y from the discrete random variable Y, such that $Y \in \{10, 20, 30, 40, 50, 60\}$. The probability mass function (pmf) $p(y|\theta)$ of Y depends on the unknown parameter θ belonging to the discrete parameter space $\Omega := \{1, 2, 3, 4, 5, 6\}$. The pmf $p(y|\theta)$ is given by the following table:

y	10	20	30	40	50	60
$p(y\|\theta = 1)$	0.5	0.2	0.1	0.1	0.1	0
$p(y\|\theta = 2)$	0.2	0.5	0.1	0.1	0.1	0
$p(y\|\theta = 3)$	0.1	0.2	0.5	0.1	0.1	0
$p(y\|\theta = 4)$	0.1	0.1	0.2	0.5	0.1	0
$p(y\|\theta = 5)$	0.1	0.1	0.1	0.2	0.5	0
$p(y\|\theta = 6)$	0	0.1	0.1	0.1	0.2	0.5

(a) [7pt] Find the maximum likelihood estimator $\hat{\theta}_{MLE}$ of θ.
Solution:
By looking at the table we find:

$$\hat{\theta}_{MLE} = Y/10$$

(b) [4pt] Is $\hat{\theta}_{MLE}$ unbiased?
Solution:
If we calculate for $\theta = 1$ the expected value of $\hat{\theta}_{MLE}$, from the table we have:

$$\mathbb{E}_{\theta=1}(\hat{\theta}_{MLE}) = 1/10\mathbb{E}_{\theta=1}(Y) = 2.1 \neq 1$$

Thus, $\hat{\theta}_{MLE}$ is biased.

(c) [5pt] Suppose we want to test:

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

at $\alpha = 0.03$ level of significance. Propose a test statistic and find the rejection region of the test.
Solution:
We use the generalized likelihood–ratio test statistics:

$$\lambda = \frac{lik(\theta_0)}{lik(\hat{\theta}_{MLE})} = \frac{p(y|\theta = 1)}{p(y|\hat{\theta}_{MLE})}$$

The possible values of this test statistics are:

	$y = 10$	$y = 20$	$y = 30$	$y = 40$	$y = 50$	$y = 60$
λ	1	0.4	0.2	0.2	0.2	0

We reject H_0 for small values of λ. Since we have $\mathbb{P}(\lambda < 0.4|\theta = 1) = 0.3$, it follows that we reject H_0 at $\alpha = 0.03$ level of significance if $\lambda < 0.4$. Therefore, we reject H_0 for any y in the rejection region: $B = \{30, 40, 50, 60\}$.

(d) [4pt] In case the observation $y = 20$, calculate an estimate of $\mathrm{Var}(\hat{\theta}_{MLE})$.
Solution:
If $y = 20$, then $\hat{\theta}_{MLE} = 2$. Hence,

$$\mathbb{E}_{\theta=2}(\hat{\theta}_{MLE}) = 0.2 + 1 + 0.3 + 0.4 + 0.5 + 0 = 2.4$$

and

$$\mathbb{E}_{\theta=2}(\hat{\theta}^2_{MLE}) = 7.2$$

Therefore, an estimate for the variance is:

$$\widehat{\mathrm{Var}}(\hat{\theta}_{MLE}) = 7.2 - 2.4^2 = 1.44$$

3. Let $\mathbb{Y} = \{Y_1, \dots Y_n\}$ a random sample of i.i.d. random variables Y_i with probability density function:

$$f(y|\theta) = (1 + \theta\, y)/2$$

with $-1 < y < 1$ and depending on the parameter θ such that $-1 < \theta < 1$.

(a) [7pt] Derive the rejection region B of the general most powerful test with significance α for testing:

$$\begin{cases} H_0: & \theta = 0, \\ H_1: & \theta = 1/2. \end{cases}$$

Solution:
We can write the likelihood for a realization of the sample as:

$$lik(\theta) = \prod_{i=1}^{n} \frac{1}{2}(1 + \theta y_i) = 2^{-n} \prod_{i=1}^{n} (1 + \theta y_i)$$

so that the likelihood ratio test statistics rejects for:

$$\frac{lik(\theta = 0)}{lik(\theta = 1/2)} = \frac{1}{\prod_{i=1}^{n}(1 + \theta y_i)} < k$$

that is

$$\prod_{i=1}^{n}(1+\theta y_i) > k^{-1}$$

A likelihood–ratio rejection region B will be of the form:

$$B = \{(y_1, y_2, \ldots, y_n) : \prod_{i=1}^{n}(1+\theta y_i) > k_\alpha\}$$

with k_α such that $\mathbb{P}((Y_1, Y_2, \ldots, Y_n) \in B|\theta = 0) = \alpha$. By Neyman–Pearson Lemma, B is the rejection region of the most powerful test for testing the hypotheses at level of significance at most α.

(b) [4pt] When $n = 1$, find the critical value for the test statistics of the test in point (a) such that the significance $\alpha = 0.05$.
Solution:
When $n = 1$

$$B = \{y : (1+\theta y) > k_{0.05}\}$$

so that the test rejects for $y > 2(k_{0.05} - 1) = \tilde{k}$, with $\tilde{k}$ such that:

$$\mathbb{P}(Y \in B|\theta = 0) = \int_{\tilde{k}}^{1} \frac{1}{2} dy = \frac{1-\tilde{k}}{2} = 0.05$$

It follows that $\tilde{k} = 0.9$. Thus, the test rejects when $y > 0.9$.

(c) [4pt] When $n = 1$, find the power π of the test developed in point (b).
Solution:

$$\pi = \mathbb{P}(Y \in B|\theta = 1) = \mathbb{P}(Y > 0.9|\theta = 1) = \int_{0.9}^{1} \frac{1}{2}\left(1 + \frac{y}{2}\right) dy = 0.074$$

4. Two different types of injection–molding machines are used to form plastic parts. A part is considered defective if it has excessive shrinkage or is discolored. Two random samples, each of size 300, are selected, and 15 defective parts are found in the sample from **machine 1**, while 8 defective parts are found in the sample from **machine 2**.

(a) [9pt] Test the hypothesis that both machines produce the same fraction of defective parts (i.e. $p_1 = p_2$), at $\alpha = 0.05$ level of significance (You can consider the sample large enough for applying large sample results).
Solution:
The parameters of interest are the proportion of defective parts of the two machines (p_1 and p_2), so that we want to test:

$$\begin{cases} H_0: & p_1 = p_2 \\ H_1: & p_1 \neq p_2 \end{cases}$$

at level $\alpha = 0.05$ of significance. The two samples collected $\mathbb{X} = \{X_1, \ldots, X_{300}\}$ and $\mathbb{Y} = \{Y_1, \ldots, Y_{300}\}$ are independent and such that X_i are i.i.d. RV such that $X_i \sim Ber(p_1)$ and Y_i are i.i.d. RV such that $Y_i \sim Ber(p_2)$. Thus $Z_1 := \sum_{i=1}^{300} X_i \sim Bin(p_1, 300)$ and $Z_2 := \sum_{i=1}^{300} Y_i \sim Bin(p_2, 300)$. If we pose $\hat{p_1} = Z_1/300$ and $\hat{p_2} = Z_2/300$ we can define the test statistics:

$$Z = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{300} + \frac{p_2(1-p_2)}{300}}}$$

Under H_0, $p_1 = p_2 = p$, so that:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{p(1-p)}{150}}}$$

and $Z \approx N(0,1)$. Since we do not know p, we can estimate p using the pooled estimator:

$$\hat{p} = \frac{1}{n_1 + n_2}(n_1\hat{p}_1 + n_2\hat{p}_2)$$

that is the proportion of *successes* in the two samples combined. Thus, the test statistics is

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}(1-\hat{p})}{150}}} = \frac{\hat{p}_1 - \hat{p}_2}{\sigma_{\hat{p}_1 - \hat{p}_2}}$$

which is $Z \approx N(0,1)$ for large samples. With our data:

$$\hat{p}_1 = 0.05,\ \hat{p}_2 = 0.0267,\ \ \hat{p} = 0.0383,\ \sigma_{\hat{p}_1 - \hat{p}_2} = 0.015$$

and the realization of the test statistics is $z = 1.49$. Since the rejection region of the two–sided approximated test is $B = \{z : |z| \geq z_{0.025}\}$, with the standard normal quantile $z_{0.025} \cong 1.96$, then $z \notin B$. Therefore we do not reject H_0 at 0.05 level of significance.

(b) [4pt] Find the (approximated) p–value for the test of point (a).
Solution:

$$\text{p–value} \approx 2(1 - \mathbb{P}(Z < 1.49|H_0)) = 2(1 - \Phi(1.49)) = 0.14$$

Suppose now that $p_1 = 0.05$ and $p_2 = 0.01$.

(c) [7pt] What is the (approximated) power of the test?
Solution:
Under H_1, $\tilde{Z} \approx N(0,1)$, with:

$$\tilde{Z} = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{300} + \frac{p_2(1-p_2)}{300}}} = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sigma}$$

where $p_1 = 0.05$ and $p_2 = 0.01$ and $\sigma =$ Therefore:

$$\pi = \mathbb{P}(Z \in B|H_1) = 1 - \Phi\left((z_{0.025}\sigma_{\hat{p}_1 - \hat{p}_2} - (p_1 - p_2))/\sigma\right) + \Phi\left((-z_{0.025}\sigma_{\hat{p}_1 - \hat{p}_2} - (p_1 - p_2))/\sigma\right)$$

where we rewrote the rejection region as $B = \{|\hat{p}_1 - \hat{p}_2| \geq z_{0.025}\sigma_{\hat{p}_1 - \hat{p}_2}\}$. With our data:

$$\pi \approx 0.81$$

5. Consider the linear model $\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$, where X is the $n \times p$ design matrix, and $\mathbf{e}$ is the vector whose components e_i are i.i.d. random variables with $\mathbb{E}e_i = 0$ and $\text{Var}(e_i) = \sigma^2$. The least squares estimator of β is given by $\hat{\beta}_{LS} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$. Let $\mathbf{P} := \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ and $\mathbf{Q} := \mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$, where $\mathbf{I}$ is the n–dimensional identity matrix. Let $\widehat{\mathbf{Y}} := \mathbf{X}\hat{\beta}_{LS}$ be the fitted model and $\hat{\mathbf{e}} := \mathbf{Y} - \widehat{\mathbf{Y}}$ the vector of the residuals.

(a) [7pt] Knowing that $\widehat{\mathbf{Y}} = \mathbf{P}\mathbf{e} + \mathbf{X}\beta$, $\hat{\mathbf{e}} = \mathbf{Q}\mathbf{e}$ and that $\mathbf{PQ}$ is the zero matrix, prove that $\text{Cov}(\widehat{\mathbf{Y}}, \hat{\mathbf{e}}) = \mathbb{E}(\widehat{\mathbf{Y}}\hat{\mathbf{e}}^\top)$.
Solution:
Given two random vectors $\mathbf{U}$ and $\mathbf{V}$, by the definition of covariance, we have:

$$\text{Cov}(\mathbf{U}, \mathbf{V}) = \mathbb{E}(\mathbf{U} - \mathbb{E}(\mathbf{U}))(\mathbf{V} - \mathbb{E}(\mathbf{V}))^\top = \mathbb{E}(\mathbf{U}\mathbf{V}^\top) - \mathbb{E}(\mathbf{U})(\mathbb{E}(\mathbf{V})^\top$$

In our case, $\mathbf{U} = \widehat{\mathbf{Y}}$ and $\mathbf{V} = \hat{\mathbf{e}} = \mathbf{Q}\mathbf{e}$. Since $\mathbb{E}(\mathbf{e}) = \mathbf{0}$, it follows that $\mathbb{E}(\hat{\mathbf{e}}) = \mathbf{Q}\mathbb{E}(\mathbf{e}) = \mathbf{0}$. Hence, the claim follows.

(b) [8pt] Suppose now that we add an additional row vector (i.e. a vector of dimensions $1 \times p$) of variables x_{n+1} to the design matrix $\mathbf{X}$. The corresponding variable Y_{n+1} is then predicted by $\widehat{Y}_{n+1} = x_{n+1}\hat{\beta}_{LS}$. Calculate the expectation $\mathbb{E}(\widehat{Y}_{n+1})$ and the variance $\mathrm{Var}(\widehat{Y}_{n+1})$ of $\widehat{Y}_{n+1}$.
Solution:

$$\hat{Y}_{n+1} = x_{n+1}\hat{\beta}_{LS} = \sum_{i=1}^{p} x_{n+1,i}\,\hat{\beta}_{LS}^{(i)}$$

Hence,

$$\mathbb{E}(\hat{Y}_{n+1}) = \sum_{i=1}^{p} x_{n+1,i}\,\mathbb{E}(\hat{\beta}_{LS}^{(i)}) = \sum_{i=1}^{p} x_{n+1,i}\,\beta_i = x_{n+1}\beta$$

with β_i the components of the vector β.
As regards the variance:

$$\mathrm{Var}(\hat{Y}_{n+1}) = \sum_{i=1}^{p} x_{n+1,i}^2\,\mathrm{Var}(\hat{\beta}_{LS}^{(i)}) + \sum_{i<j} x_{n+1,i}x_{n+1,j}\mathrm{Cov}(\hat{\beta}_{LS}^{(i)}, \hat{\beta}_{LS}^{(j)})$$

with $\mathrm{Cov}(\hat{\beta}_{LS}^{(i)}, \hat{\beta}_{LS}^{(j)}) = \sigma^2(\mathbf{X}^\top\mathbf{X})_{ij}^{-1}$.

(c) [5pt] Suppose we have the additional information that $e_i \sim N(0, \sigma^2)$ with **known** variance σ^2. Construct a $(1-\alpha)$–confidence interval for $x_{n+1}\beta$.
Solution:
Since

$$\hat{Y}_{n+1} \sim N(\mathbb{E}(\hat{Y}_{n+1}), \mathrm{Var}(\hat{Y}_{n+1}))$$

the CI for Y_{n+1} easily follows.