Statistiek (WISB361)

Midterm exam

April 19, 2013

Schrijf uw naam op elk in te leveren vel. Ook schrijf uw studentnummer op blad 1.

The maximum number of points is 100. Points distribution: 20–37–23–20

1. A physical quantity is independently measured two times using two different instruments with two known different precisions. We can model this experiment with two independent random samples $\mathbf{X} = \{X_1, X_2, \ldots, X_m\}$ and $\mathbf{Y} = \{Y_1, Y_2, \ldots, Y_n\}$ of size m and n respectively. We assume that X_i are i.i.d. normal random variables with mean μ and standard deviation σ_X and that Y_i are i.i.d. normal random variables with mean μ and standard deviation σ_Y. We assume that σ_X and σ_Y are known and we want to estimate μ. We consider the following estimator for μ:

$$T = a\bar{X}_m + (1-a)\bar{Y}_n, \quad a \in \mathbb{R}$$

where $\bar{X}_m = 1/m \sum_{i=1}^{m} X_i$ and $\bar{Y}_n = 1/n \sum_{i=1}^{n} Y_i$.

(a) [5pt] Calculate the expected value and the variance of T.
Solution:

$$\mathbb{E}(T) = a\mathbb{E}(\bar{X}_m) + (1-a)E(\bar{Y}+n) = a\mu) + (1-a)\mu = \mu$$

$$\mathrm{Var}(T) = a^2\mathrm{Var}(\bar{X}_m) + (1-a)^2\mathrm{Var}(\bar{Y}_n) = a^2\frac{\sigma_X^2}{m} + (1-a)^2\frac{\sigma_Y^2}{n}$$

(b) [10pt] Find a such that the mean squared error MSE(a) is minimized. For this value of a calculate the variance of T.
Solution: Since T is unbiased, $MSE(a) = \mathrm{Var}(T)$. Hence,

$$MSE(a) = a^2(\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}) - 2a\frac{\sigma_Y^2}{n} + \frac{\sigma_Y^2}{n} = f(a) + \frac{\sigma_Y^2}{n}$$

In order to minimize $MSE(a)$ we have to minimize $f(a)$. Therefore, the solution of $f(a)' = 0$ is

$$\bar{a} = \frac{\frac{\sigma_Y^2}{n}}{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}$$

Since $f''(a) > 0\ \forall a$, $\bar{a}$ is the minimizer.

(c) [5pt] Give the distribution of T for the value of a found in (b).
Solution: We have:

$$1 - \bar{a} = \frac{\frac{\sigma_X^2}{m}}{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}$$

For $a = \bar{a}$ we have:

$$\begin{aligned}\mathrm{Var}(T) &= \bar{a}^2\frac{\sigma_X^2}{m} + (1-\bar{a})^2\frac{\sigma_Y^2}{n} \\ &= \frac{\frac{\sigma_X^2\,\sigma_Y^4}{m\,n^2} + \frac{\sigma_Y^2\,\sigma_X^4}{n\,m^2}}{\left(\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}\right)^2} \\ &= \frac{\frac{\sigma_Y^2}{n}\frac{\sigma_X^2}{m}}{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}\end{aligned}$$

therefore $T \sim N(\mu, \bar{\sigma}^2)$, where $\bar{\sigma}^2 = \frac{\frac{\sigma_Y^2}{n}\frac{\sigma_X^2}{m}}{\frac{\sigma_X^2}{m}+\frac{\sigma_Y^2}{n}}$

2. Consider the sample $\mathbf{X}_n$ of n i.i.d. random variables distributed with density function:

$$f(x|\theta) = \frac{1}{\theta^2}\,x\,\exp\left\{-\frac{x}{\theta}\right\}$$

with $\theta > 0$ and $x > 0$.

(a) [5pt] Write the likelihood function $lik(\theta)$ for the realization $\mathbf{x}_n = \{x_a, \ldots, x_n\}$ of the sample.
Solution:

$$lik(\theta) = \left(\prod_i^n x_i\right)\frac{1}{\theta^{2n}}\exp\left\{-\sum_{i=1}^n \frac{x_i}{\theta}\right\}$$

(b) [7pt] Find a sufficient statistic for θ.
Solution:
By the expression of $lik(\theta)$ and by the Factorization Theorem ($h(\mathbf{x}) = \prod_i^n x_i$ and $g(\mathbf{x}|\theta) = \theta^{-2n}\exp(-\sum_{i=1}^n x_i/\theta)$), it follows that $T(\mathbf{x}) = \sum_{i=1}^n x_i$ is a sufficient statistic for θ.

(c) [10pt] Given the realization $\mathbf{x}_n$, find the maximum likelihood estimator $\hat{\theta}_{MLE}$ for θ. Moreover, calculate the Fisher information $I(\theta)$
Solution:
The log–likelihood is:

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

and

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

Hence:

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

being $\ell''(\hat{\theta}_{MLE}) < 0$.

$$I(\theta) = -\mathbb{E}\left(\frac{\partial^2}{\partial\theta^2}\log f(x|\theta)\right) = -\frac{2}{\theta^2} + \frac{2\mathbb{E}(X)}{\theta^3} = -\frac{2}{\theta^2} + \frac{2\mathbb{E}(X)}{\theta^3} = \frac{2}{\theta^2}$$

by double integration by parts ($\mathbb{E}(X) = 2\theta$).

Consider the observed sample $\mathbf{x}_{30}$ of $n = 30$ observations:

$$\begin{aligned}\mathbf{x}_{30} = \; & \{0.56, 0.47, 0.30, 0.60, 0.22, 0.41, 0.76, 0.38, 0.08, 0.29, 0.57, 0.97, 0.81, 0.87, 0.36 \\ & 0.20, 1.27, 0.20, 1.38, 1.12, 0.46, 0.52, 1.17, 0.32, 0.21, 0.61, 0.61, 1.47, 0.64, 0.08\}\end{aligned}$$

we have that $\sum_{i=1}^{30} x_i = 17.91$, $(\sum_{i=1}^{30} x_i)^2 = 320.77$ and $(\sum_{i=1}^{30} x_i)^3 = 5744.96$.
With respect to this sample:

(d) [5pt] calculate the value of the maximum likelihood estimator $\hat{\theta}_{MLE}$ and determine the value of the observed Fisher information $n\, I(\hat{\theta}_{MLE})$.
Solution:

$$\hat{\theta}_{MLE} = \frac{\sum_{i=1}^{n} x_i}{2n} = \frac{17.91}{60} = 0.2985$$

$$n\, I(\hat{\theta}_{MLE}) = \frac{8\, n^3}{(\sum_{i=1} x_i)^2} = \frac{216000}{320.77} = 673.38$$

(e) [10pt] Give a 95% confidence interval for $\hat{\theta}_{MLE}$. (**Hint**: we can consider $n = 30$ large enough for using asymptotic results). Moreover, in case we want to test $H_0 : \theta = 1/4$ against $H_1 : \theta \neq 1/4$ at 5%–level of significance, can we reject the null hypothesis?
Solution:
A CI for $\hat{\theta}_{MLE}$ can be derived via normal approximation of the the distribution of the maximum likelihood estimator for θ:

$$\hat{\theta}_{MLE} \pm z(0.025) \frac{1}{\sqrt{nI(\hat{\theta}_{MLE})}} = 0.2985 \pm \frac{1.96}{\sqrt{673.38}} = 0.2985 \pm 0.0755$$

Hence, $CI = (0.22, 0.37)$.
By duality of CI and two sided Hypotheses tests, since $0.25 \in CI$ we can't reject H_0.

3. Let $X \sim N(\mu, \sigma^2)$ a random variable with $\sigma^2 = 3$. In order to test the hypothesis

$$\begin{cases} H_0 : & \mu = 2 \\ H_1 : & \mu = 1 \end{cases}$$

a sample $\mathbf{X} = \{X_1, X_2, X_3\}$ of i.i.d. observations distributed as above is collected. Given the rejection region:

$$B = \{(x_1, x_2, x_3) : 2x_1 - 2x_2 + x_3 < 1.2\}$$

(a) [5pt] Derive the distribution of the test statistics $T(\mathbf{X}) = 2X_1 - 2X_2 + X_3$.
Solution:

$$T(\mathbf{X}) \sim N(\mu, 27)$$

In fact X_1, X_2, X_3 are i.i.d $N(\mu, 3)$ random variables and

$$E(T) = \mu$$

and

$$\text{Var}(T) = 3(4 + 4 + 1) = 27$$

(b) [10pt] Calculate the significance level α of the test.
Solution:

$$\begin{aligned}
\alpha &= \mathbb{P}(T \in B|H_0) = \mathbb{P}(2X_1 - 2X_2 + X_3 < 1.2|H_0) = \mathbb{P}(T < 1.2|H_0) \\
&= \mathbb{P}(\frac{T-2}{\sqrt{27}} < \frac{1.2-2}{\sqrt{27}}|H_0) = \mathbb{P}(Z < -0.15|H_0) \\
&= 1 - \Phi(0.15) = 0.44
\end{aligned}$$

since $Z \sim N(1, 0)$.

(c) [8pt] Calculate the power of the test.
Solution:

$$\begin{aligned}
1 - \beta &= \mathbb{P}(T \in B|H_1) = \mathbb{P}(2X_1 - 2X_2 + X_3 < 1.2|H_1) = \mathbb{P}(T < 1.2|H_1) \\
&= \mathbb{P}(\frac{T-2}{\sqrt{27}} < \frac{1.2-1}{\sqrt{27}}|H_1) = \mathbb{P}(Z < 0.038|H_1) \\
&= \Phi(0.038) = 0.516
\end{aligned}$$

4. Suppose to have a **single** observation y sampled from a discrete random variable Y. The sample space of Y is $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$, and its probability mass function $p(y|\theta)$ depends on the unknown parameter $\theta \in \Omega$, where $\Omega = \{\theta_1, \theta_2, \theta_3\}$.
The values of $p(y|\theta)$ are specified in the following table:

$p(y\|\theta)$	$y = 1$	$y = 2$	$y = 3$	$y = 4$	$y = 5$	$y = 6$
$p(y\|\theta = \theta_1)$	0.02	0.03	0.04	0.02	0.03	0.86
$p(y\|\theta = \theta_2)$	0.07	0.08	0.02	0.05	0.1	0.68
$p(y\|\theta = \theta_3)$	0.2	0.05	0.03	0.15	0.54	0.03

(a) [10pt] Calculate the maximum likelihood estimator $\hat{\theta}_{MLE}$ of θ.
Solution:
Since we have only one observation y:

$$\hat{\theta}_{MLE} = \text{argmax}_{\theta \in \Omega}\, p(y|\theta)$$

Hence, from the table we have:

$$\hat{\theta}_{MLE} = \begin{cases} \theta_1 & y \in \{3, 6\}, \\ \theta_2 & y = 2, \\ \theta_3 & y \in \{1, 4, 5\} \end{cases}$$

(b) [7pt] Using the Likelihood Ratio, construct the most powerful test at 0.05–level of significance for testing:

$$\begin{cases} H_0: & \theta = \theta_1 \\ H_1: & \theta = \theta_2 \end{cases}$$

Solution:
By Neyman–Pearson Lemma, the most powerful test at α–level of significance, has rejection region of the type:

$$\frac{p(y|\theta = \theta_1)}{p(y|\theta = \theta_2)} < k_\alpha$$

where the constant k_α has to be determined. If we evaluate the ratio we have:

$p(y\|\theta)$	$y = 1$	$y = 2$	$y = 3$	$y = 4$	$y = 5$	$y = 6$
$p(y\|\theta = \theta_1)$	0.02	0.03	0.04	0.02	0.03	0.86
$p(y\|\theta = \theta_2)$	0.07	0.08	0.02	0.05	0.1	0.68
$p(y\|\theta = \theta_1)/p(y\|\theta = \theta_2)$	0.29	0.38	2	0.4	0.3	1.26

If R is the rejection region:

$$0.05 = \mathbb{P}(Y \in R|H_0) = \mathbb{P}(Y \in R|\theta = \theta_1) = \sum_{\substack{i \in \{1,\dots,6\}: \\ p(i|\theta_1)/p(i|\theta_2) < k_{0.05}}} p(i|\theta_1)$$

From the table we see that $k_{0.05} = 0.38$ and that the rejection region is $R = \{1, 5\}$.

(c) [3pt] Calculate the power of the test derived in point (b).
Solution:
By definition, the power π of the test is:

$$\pi = 1 - \beta = \mathbb{P}(Y \in R|H_1) = \mathbb{P}(Y \in \{1, 5\}|\theta = \theta_2) = 0.07 + 0.1 = 0.17$$