Statistiek (WISB361)

Final exam

July 3, 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: 20–20–20–20–10–10

1. Consider the random variable X with probability density function:

$$f(x;\theta) = \begin{cases} c(\theta)x^2 e^{-\theta\, x} & x > 0, \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta > 0$ and θ is assumed unknown.

(a) [2pt] Show that $c(\theta) = \theta^3/2$.
Solution:
First of all, I remind you that for the Gamma function: $\Gamma(n) = (n-1)!$, with n integer and $\Gamma(t) = \int_0^\infty x^{t-1}e^x dx$ [In this way you avoid a lot of integrations by parts!!!]. Hence,

$$\frac{1}{c(\theta)} = \int_0^\infty x^2 e^{-\theta x} dx = \Gamma(3)/\theta^3 = 2/\theta^3$$

(b) [5pt] Show that $\tilde{\theta} = 2/X$ is an unbiased estimator of θ and find its variance.
Solution:
We have:

$$\mathbb{E}(\tilde{\theta}) = \mathbb{E}(2/X) = \theta^3 \int_0^\infty x e^{-\theta x} dx = \theta^3\theta^{-2}\Gamma(2) = \theta$$

(c) [5pt] Find the Fisher information $I(\theta)$ for the parameter θ and compare the variance of θ to the Cramer-Rao lower bound.
Solution:
In order to calculate the Fisher information, we first calculate:

$$\frac{\partial}{\partial\theta} \ln f(x;\theta) = -x + 3/\theta$$

$$-\frac{\partial^2}{\partial\theta^2} \ln f(x;\theta) = 3/\theta^2$$

Hence, $\text{Var}_{Rao} = \theta^2/3$.
For the variance:

$$\mathbb{E}(\tilde{\theta}^2) = 2\theta^3 \int_0^\infty e^{-\theta x} dx = 2\theta^2$$

so that $\text{Var}(\tilde{\theta}) = 2\theta^2 - \theta^2 = \theta^2$. Thus,

$$\frac{\text{Var}(\tilde{\theta})}{\text{Var}_{Rao}} = 3$$

(d) [3pt] Let $\mu = 1/\theta$ and show that $\hat{\mu} = X/3$ is an unbiased estimator of μ.
Solution:

$$\mathbb{E}(\hat{\mu}) = \frac{\theta^3}{6} \int_0^\infty x^3 e^{\theta x} dx = \frac{\theta^3\Gamma(4)}{6\theta^4} = 1/\theta$$

(e) [5pt] Find the variance of $\hat{\mu}$ and show that it attains the Cramer-Rao lower bound.
Solution:

$$\mathbb{E}(\hat{\mu}^2) = \frac{\theta^3}{18} \int_0^\infty x^4 e^{\theta x} dx = \frac{\theta^3 \Gamma(5)}{6\theta^8} = 4/(3\theta^2)$$

so that $\text{Var}(\hat{\mu}) = \frac{1}{3\theta^2}$. For Cramer–Rao lower bound, we have that $g(\theta) = 1/\theta$, so that $g'(\theta) = -\theta^{-2}$, so that:

$$\text{Var}_{Rao} = I(\theta)^{-1}(g'(\theta))^2 = \frac{\theta^2}{3}\theta^{-4} = \frac{1}{3\theta^2} = \text{Var}(\hat{\mu})$$

2. Consider this time the sample $\mathbb{X} = \{X_i\}_{i=1}^n$ of i.i.d. random variables with probability density function:

$$f(x;\theta) = \begin{cases} \exp -(x-\theta), & x \geq \theta \\ 0 & \text{otherwise} \end{cases}$$

where $\theta > 0$ is the parameter of the distribution.

(a) [7pt] Find a maximum likelihood estimator (MLE) of θ.
Solution:
The likelihood function

$$lik(X_1, \dots, X_n; \theta) = \exp\left\{ -(\sum_{i=1}^n X_i - n\theta) \right\} \mathbf{1}_{\min(X_1,\dots,X_n) > \theta}$$

is an increasing function for any θ less than or equal to $\min(X_1, \dots, X_n)$, and it is zero for $\theta > \min(X_1, \dots, X_n)$. As a result, the MLE:

$$\hat{\theta}_{MLE} = \min(X_1, \dots, X_n) = X_{(1)}$$

.

(b) [3pt] Modify the MLE to get an unbiased estimator of θ.
Solution:
The order statistic density with $k = 1$ shows that $X_{(1)} - \theta \sim Exp(n)$ [look for instance Example 3.7.1], so that:

$$\mathbb{E}(\hat{\theta}_{MLE}) = \mathbb{E}(X_{(1)}) = \theta + 1/n$$

Therefore, $\hat{\theta}_{MLE} - 1/n$ is unbiased.

(c) [6pt] Find a method of moments estimator (MOM) of θ.
Solution:
Let $Y = X - \theta$. Then $Y \sim Exp(1)$ and $\mathbb{E}(Y) = \mathbb{E}(X) - \theta$. Since $\mathbb{E}(Y) = 1$, we have $\mathbb{E}(X) = 1 + \theta$ and

$$\hat{\theta}_{MOM} = \bar{X} - 1$$

(d) [4pt] Calculate the mean squared error (MSE) of $\hat{\theta}_{MOM}$.
Solution:
We have that $\mathbb{E}(\hat{\theta}_{MOM}) = 1 + \theta - 1 = \theta$. Moreover:

$$\text{Var}(\hat{\theta}_{MOM}) = \text{Var}(\bar{X} - 1) = \text{Var}(\bar{X}) = 1/n.$$

Hence:

$$MSE((\hat{\theta}_{MOM}) = 0 + 1/n = 1/n$$

3. Let $\mathbb{X}_1 = \{X_{1,1}, X_{1,2}, \ldots, X_{1,n_1}\}$ be a random sample of size n_1 of i.i.d observations $X_i \sim N(\mu_1, \sigma_1^2)$ and $\mathbb{X}_2 = \{X_{2,1}, X_{2,2}, \ldots, X_{2,n_2}\}$ a random sample of size n_2 of i.i.d. observations $X_i \sim N(\mu_2, \sigma_2^2)$. The parameters μ_1 and μ_2 are **unknown** while σ_1^2 and σ_2^2 are **known**. Due to logistical constraints, suppose that it is only possible to select a total sample size of N from these two normal populations, so that the constraint $n_1+n_2 = N$ holds.

(a) [15pt] Subject to this sample size constraint, find expressions (as a function of σ_1, σ_2 and N) for the optimal values $n_1^\star$ and $n_2^\star$ of n_1 and n_2 that *maximize* the power of a size α test:

$$\begin{cases} H_0: & \mu_1 = \mu_2, \\ H_1: & \mu_1 > \mu_2 \end{cases}$$

using as a test statistic the difference of the sample means $\bar{X}_1$ and $\bar{X}_2$, where $\bar{X}_1 := \frac{1}{n_1}\sum_{i=1}^{n_1} X_{1,i}$, $\bar{X}_2 := \frac{1}{n_2}\sum_{i=1}^{n_2} X_{2,i}$. Provide an interpretation for your findings.

Solution:

Under the H_0, the random variable:

$$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

If we denote with $z(\alpha)$ the quantile such that:

$$\mathbb{P}(Z > z(\alpha)) = \alpha$$

when $Z \sim N(0,1)$, it follows that:

$$\begin{aligned} \pi &= \mathbb{P}\left\{ \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} > z(\alpha) | H_1 \right\} \\ &= \mathbb{P}\left\{ \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} > z(\alpha) | \mu_1 > \mu_2 \right\} \\ &= \mathbb{P}\left\{ \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} > z(\alpha) - \frac{(\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} | H_1 \right\} \\ &= \mathbb{P}\left\{ Z > z(\alpha) - \frac{(\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} | \mu_1 > \mu_2 \right\} \\ &= 1 - \Phi\left(z(\alpha) - \frac{(\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \right) \end{aligned}$$

Thus, in order to maximize the power π, we need to minimize the quantity $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$ with the constraint $n_1 + n_2 = N$. One possible solution is via the method of Lagrange multipliers. We consider:

$$L = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} + \lambda(n_1 + n_2 - N)$$

Hence, $\frac{\partial Q}{\partial n_1} = -\sigma_1^2 n_1^{-2} + \lambda = 0$, $\frac{\partial Q}{\partial n_2} = -\sigma_2^2 n_2^{-2} + \lambda = 0$ and $\frac{\partial Q}{\partial \lambda} = n_1 + n_2 - N = 0$. From the first two equations we have:

$$\frac{n_1}{n_2} = \frac{\sigma_1}{\sigma_2}$$

From the third equations we have:

$$\left(1 + \frac{\sigma_1}{\sigma_2}\right) n_1 = \frac{\sigma_1}{\sigma_2} N$$

so that $n_1^\star = (\frac{\sigma_1}{\sigma_1 + \sigma_2})N$ and $n_2^\star = (\frac{\sigma_2}{\sigma_1 + \sigma_2})N$.

(b) [5pt] If $N = 100$, $\sigma_1^2 = 4$ and $\sigma_2^2 = 9$, find the numerical values of $n_1^\star$ and $n_2^\star$.
Solution:
$n_1^\star = 40$ and $n_2^\star = 60$.

4. **AZT** was the first antiretroviral drug approved by the American *Food and Drug Administration* (FDA) in order to reduce the viral load in HIV–positive patients. The standard AZT's daily dose is $300mg$. A study claims that higher daily dose of AZT are not more beneficial, since several side effects may occur. In order to verify the hypothesis that a daily dose of $600mg$ has the same efficiency of the $300mg$ dose, levels of $p24$ antigen were measured for two groups of patients: the first treated with $300mg$ of AZT and the second with $600mg$. High levels of $p24$ indicate high viral replication. The results are contained in the following table:

Dose	p24 antigen levels measured
300 mg	283, 284, 285, 286, 288, 289, 291, 295, 303
600 mg	287, 292, 293, 296, 298, 310, 314

(a) [10pt]. Test the hypothesis that the two doses have the same effect in controlling the disease against the hypothesis that the two doses do not have the same effect in controlling the disease at the significance level $\alpha = 0.01$. Does the conclusion change if $\alpha = 0.10$?
Solution:
Since we do not have any information about the distribution of the data, we perform a nonparametric two–sided Mann Whitney test. In order to use Table 8 of the textbook: we have $n_1 = 7$, $R = 80$, $R' = 39$, $R^\star = \min(80, 39) = 39$. From Table 8, the critical value R_{cr} for $R^\star$ with $\alpha = 0.01$ is 35. Hence, we do not reject the null hypothesis of no difference at 0.01 level of significance. For $\alpha = 0.10$, $R_{cr} = 43$ so that the test rejects H_0 at 0.10 level of significance.

(b) [10pt] Let us assume now that the measurements of $p24$ antigen come from two independent samples, normally distributed and with equal variances. With this additional information, perform the test of point a) at $\alpha = 0.05$ level of significance.
Solution:
Since the two samples are two independent normal samples $\{X_1 \ldots, X_9\}$ and $\{Y_1 \ldots, Y_7\}$ with equal variance, we can perform a two–sided t test with $\alpha = 0.05$. We reject H_0 whenever

$$T = |\bar{X} - \bar{Y}_2|/\sqrt{S_p^2(1/n + 1/m)} > t_{14}(0.975) = 2.145.$$

We have: $\bar{x} = 289.33$, $\bar{x} = 298.57$, $s_p^2 = 64.70$, so that $T = 2.279 > 2.145$. Therefore, we reject H_0 at $\alpha = 0.05$.

5. Consider two simple linear regression models. The first one has x_i as independent variables and Y_i as its dependent variable, i.e.:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \ldots n,$$

with ϵ_i i.i.d. random variables such that $\mathbb{E}(\epsilon_i) = 0$.
The second model uses $\tilde{x}_i = (x_i - a)/b$ as its independent variable, and $\tilde{Y}_i = (Y_i - c)/d$ as its dependent variable, where a, b, c and d are known, non–zero and fixed constants, i.e.:

$$\tilde{Y}_i = \tilde{\beta}_0 + \tilde{\beta}_1 \tilde{x}_i + \tilde{\epsilon}_i, \quad i = 1, 2, \ldots n$$

with $\tilde{\epsilon}_i$ i.i.d. random variables such that $\mathbb{E}(\tilde{\epsilon}_i) = 0$.

(a) [10pt]. What is the relationship between the least squares estimators of (β_0, β_1) in the first model and the least squares estimators of $(\tilde{\beta}_0, \tilde{\beta}_1)$ in the second model?

Solution:
We have:

$$\bar{\tilde{x}} = \frac{1}{n}\sum_{i=1}^{n}\tilde{x}_i = \frac{1}{bn}\sum_{i=1}^{n}(x_i - a) = \frac{\bar{x} - a}{b}$$

and

$$\bar{\tilde{y}} = \frac{\bar{y} - c}{d}.$$

Moreover:

$$S_{\tilde{x}\tilde{x}} = \sum_i (\tilde{x}_i - \bar{\tilde{x}}_i)^2 = \sum_i \big(\frac{x_i - a}{b} - \frac{\bar{x}_i - a}{b}\big)^2 = \frac{1}{b^2}S_{xx}$$

and similarly:

$$S_{\tilde{x}\tilde{y}} = \frac{1}{bd}S_{xy}$$

Thus,

$$\hat{\tilde{\beta}}_1 = \frac{S_{\tilde{x}\tilde{y}}}{S_{\tilde{x}\tilde{x}}} = \frac{S_{xy}/(bd)}{S_{xx}/b^2} = \frac{b}{d}\frac{S_{xy}}{S_{xx}} = \frac{b}{d}\hat{\beta}_1$$

and

$$\hat{\tilde{\beta}}_0 = \bar{\tilde{y}} - \hat{\tilde{\beta}}_1\bar{\tilde{x}} = \frac{\bar{y} - c}{d} - \frac{b}{d}\hat{\beta}_1\frac{\bar{x} - a}{b} = \frac{\bar{y} - c}{d} - \frac{\hat{\beta}_1(\bar{x} - a)}{d} = \frac{1}{d}(\hat{\beta}_0 - c + a\hat{\beta}_1)$$

6. Consider the multivariate regression model $\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$, where $\mathbf{X}$ is the *design matrix* of dimensions $n \times p$, and where the components e_i of the vector $\mathbf{e}$ are i.i.d. random variables with $\mathbb{E}(e_i) = 0$ and $\mathrm{Var}(e_i) = \sigma^2$. As you know from the lectures, the least squares estimator of β is given by $\hat{\beta} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$. Consider the projector matrix $\mathbf{P} := \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ and the matrix $\mathbf{R} := \mathbf{I}_n - \mathbf{P}$, where $\mathbf{I}_n$ is the n -dimensional identity matrix.

(a) [10pt] Let $\hat{Y} = \mathbf{X}\hat{\beta}$ and the residual vector $\hat{\mathbf{e}} = \mathbf{Y} - \hat{\mathbf{Y}}$. Show that $\hat{\mathbf{Y}} = \mathbf{Pe} + \mathbf{X}\beta$, that $\hat{\mathbf{e}} = \mathbf{Re}$ and that $\mathbf{PR}$ is the zero matrix.
Solution:

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y} = \mathbf{PY} = \mathbf{Pe} + \mathbf{X}\beta;$$

$$\hat{\mathbf{e}} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\hat{\beta} = \mathbf{Y} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y} = (\mathbf{1} - \mathbf{P})\mathbf{Y} = \mathbf{RY} = \mathbf{R}(\mathbf{Y} - \mathbf{X}\beta) = \mathbf{Re},$$

since $\mathbf{PX}\beta = \mathbf{X}\beta$;

$$\mathbf{PR} = \mathbf{P}(\mathbf{1} - \mathbf{P}) = \mathbf{P} = \mathbf{P}^2 = \mathbf{P} - \mathbf{P} = \mathbf{0}$$