University of Ljubljana

Doctoral Programme in Statistics

Methodology of Statistical Research

Written examination

January 31st, 2013

Name and surname: _____ ID number: | | | | | | | | | |

Instructions

Read carefully the wording of the problem before you start. There are four problems altogeher. You may use a A4 sheet of paper and a mathematical handbook. Please write all the answers on the sheets provided. You have two hours.

Problem	a.	b.	c.	d.	
1.			•	•	
2.				•	
3.			•	•	
4.				•	
Total					

1. (25) For purposes of sampling the population is divided into K strata of sizes $N_1, N_2, \ldots, N_K$. The sampling procedure is as follows: first a simple random sample of size $k \leq K$ of strata is selected. The selection procedure is independent of the sizes of strata. The second step is then to select a simple random sample in each of the selected strata. If stratum i is selected then we choose a simple random sample of size n_i in this stratum for $i = 1, 2, \ldots, K$. Assume the selection process on the second step is independent of the selection process on the first step.

a. (10) Find an unbiased estimator of the population mean. Explain why it is unbiased.

Solution: Define

$$I_i = \begin{cases} 1 & \textit{if stratum } i \textit{ is chosen,} \\ 0 & \textit{else.} \end{cases}$$

From the above it follows that $E(I_i) = P(I_i = 1) = k/K$ *for all* i. *Let* $\bar{Y}_i$ *be the sample average for the sample chosen in stratum* i. *We have*

$$E(I_i\bar{Y}_i) = E(I_i)E(\bar{Y}_i) = \frac{k}{K} \cdot \mu_i \,.$$

If we put

$$\bar{Y} = \sum_{i=1}^{K} w_i \cdot \frac{K}{k} \cdot I_i\bar{Y}_i$$

we have

$$E(\bar{Y}) = \sum_{i=1}^{K} w_i\mu_i = \mu \,.$$

b. (15) Find the standard error of your unbiased estimator.

Solution: We have

$$\mathrm{var}(\bar{Y}) = \sum_{i=1}^{K} w_i^2 \mathrm{var}(I_i\bar{Y}_i) + 2\sum_{i<j} w_i w_j \mathrm{cov}(I_i\bar{Y}_i, I_j\bar{Y}_j) \,.$$

By independence of I_i *and* $\bar{Y}_i$ *we have*

$$\mathrm{var}(I_i\bar{Y}_i) = E(I_i)E(\bar{Y}_i^2) - E(I_i)^2E(\bar{Y}_i)^2 \,.$$

We have

$$E(\bar{Y}_i^2) = \mathrm{var}(\bar{Y}_i) + E(\bar{Y}_i)^2 = \frac{\sigma_i^2}{n_i} \cdot \frac{N_i - n_i}{N_i - 1} + \mu_i^2 \,.$$

By independence of (I_i, I_j) *and* (Y_i, Y_j) *we have*

$$\mathrm{cov}(I_i\bar{Y}_i, I_j\bar{Y}_j) = E(I_iI_j)E(Y_i)E(Y_j) - \frac{k^2}{K^2}\mu_i\mu_j \,.$$

By definition

$$E(I_i I_j) = P(I_i = 1, I_j = 1) = \frac{k}{K} \cdot \frac{k-1}{K-1}\,.$$

It follows that

$$\mathrm{cov}(\bar{Y}_i, \bar{Y}_j) = \frac{k}{K}\mu_i\mu_j \left(\frac{k-1}{K-1} - \frac{k}{K} \right)\,.$$

Simplifying we find

$$\mathrm{cov}(\bar{Y}_i, \bar{Y}_j) = -\frac{(K-k)k}{(K-1)K^2}\mu_i\mu_j\,.$$

Putting all the pieces together gives the standard error.

2. (25) The Rayleigh distribution is given by the density

$$f(x,\theta) = \frac{x}{\theta^2}\, e^{-\frac{x^2}{2\theta^2}}$$

for $x > 0$ and $\theta > 0$. Assume your observed values $x_1, x_2, \ldots, x_n$ are like independent random variables $X_1, X_2, \ldots, X_n$ with the Rayleigh density.

a. (5) Let X have the Rayleigh density. Check that X^2 has exponential distribution with density

$$f(x) = \frac{1}{2\theta^2}\, e^{-\frac{x}{2\theta^2}}\,.$$

Solution: We compute

$$P(X^2 \le x) = P(X \le \sqrt{x}) = 1 - e^{-\frac{x}{2\theta^2}}\,.$$

This proves the assertion.

b. (10) Find the maximum likelihood estimator for the parameter θ and compute its standard error.

Solution: The log-likelihood function is

$$\ell(\theta) = \sum_{k=1}^{n} \log x_k - 2n\log\theta - \frac{1}{2\theta^2}\sum_{k=1}^{n} x_k^2\,.$$

Taking derivatives we get the equation

$$\ell'(\theta) = -\frac{2n}{\theta} + \frac{1}{\theta^3}\sum_{k=1}^{n} x_k^2 = 0\,.$$

We get

$$\hat{\theta} = \sqrt{\frac{1}{2n}\sum_{k=1}^{n} x_k^2}\,.$$

The Fisher information is

$$I(\theta) = -E\left(\frac{2}{\theta^2} - \frac{3X_1^2}{\theta^4}\right)\,.$$

We have that $E(X_1^2) = 2\theta^2$. It follows

$$I(\theta) = \frac{4}{\theta^2}\,.$$

We have

$$\mathrm{se}(\hat{\theta}) = \frac{\theta}{2\sqrt{n}}\,.$$

c. (10) Assume as known that

$$E(\hat{\theta}) = \theta \binom{2n}{n} \cdot \frac{\sqrt{\pi}}{2^{2n}\sqrt{n}}\,.$$

Compute the standard error of θ directly without referring to the Fisher information. Is the standard error the right measure of accuracy given that the estimator is biased? Explain.

Solution: By definition

$$\mathrm{var}(\hat{\theta}) = E(\hat{\theta}^2) - E(\hat{\theta})^2\,.$$

The second term is given. To compute the first we need

$$E(\hat{\theta}^2) = E\left(\frac{1}{2n}\sum_{k=1}^{n} X_k^2\right) = \theta^2\,.$$

When an estimator is biased the standard error may not be the best measure of accuracy. In such a case the mean square error is more appropriate.

3. (25) Suppose that the random variables $X_1, X_2, \ldots, X_n$ are independent equally distributed random variables taking only values 0, 1 and 2 with probabilities

$$P(X_1 = 0) = \theta^2, \quad P(X_1 = 1) = 2\theta(1-\theta) \quad \text{and} \quad P(X_1 = 2) = (1-\theta)^2$$

where $0 < \theta < 1$. We would like to test

$$H_0 \colon \theta = 1/2 \qquad \text{against} \qquad H_1 \colon \theta \neq 1/2\,.$$

a. (10) Find the likelihood ratio statistic given the observed values $x_1, x_2, \ldots, x_n$ from the above distribution.

Hint: Express the likelihood function with n_0, n_1 and n_2 where n_k is the number of occurences of k among the observed values for $k = 0, 1, 2$.

Solution: By independence the likelihood function is equal to

$$L(\theta) = 2^{n_1}\,\theta^{2n_0+n_1}(1-\theta)^{n_1+2n_2}$$

where n_k is the number of occurences of k among the observed values. The MLE estimate of θ is the solution of

$$\frac{2n_0+n_1}{\theta} - \frac{n_1+2n_2}{1-\theta} = 0\,.$$

We have

$$\hat\theta = \frac{2n_0+n_1}{2n}\,.$$

We have

$$\Lambda = \frac{\hat\theta^{2n_0+n_1}(1-\hat\theta)^{n_1+2n_2}}{\left(\frac{1}{2}\right)^{2n}}\,.$$

It follows

$$\lambda = 2\log\Lambda = 2n\log 2 + (2n_0+n_1)\log\hat\theta + (n_1+2n_2)\log(1-\hat\theta)\,.$$

b. (10) How would you determine the critical value if the confidence level is α.

Solution: By Wilks's theorem under H_0 the distribution of λ is $\chi^2(1)$. The critical level is the number c_α such that

$$P(\chi^2(1) > c_\alpha) = \alpha\,.$$

4. (25) Suppose that we have the regression model

$$\begin{aligned} Y_{i1} &= \alpha + \beta x_{i1} + \epsilon_i \\ Y_{i2} &= \alpha + \beta x_{i2} + \eta_i \end{aligned}$$

where $i = 1, 2, \ldots, n$ and we have $E(\epsilon_i) = E(\eta_i) = 0$, $\text{var}(\epsilon_i) = \text{var}(\eta_i) = \sigma^2$ and $\text{cov}(\epsilon_i, \eta_i) = \rho\sigma^2$ for some correlation coefficient $\rho \in (-1, 1)$. Further assume that the pairs $(\epsilon_1, \eta_1), (\epsilon_2, \eta_2), \ldots, (\epsilon_n, \eta_n)$ are independent.

a. (5) Denote

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} \\ 1 & x_{12} \\ 1 & x_{21} \\ \vdots & \vdots \\ 1 & x_{n1} \\ 1 & x_{n2} \end{pmatrix} \quad \text{and} \quad \mathbf{Y} = \begin{pmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ \vdots \\ Y_{n1} \\ Y_{n2} \end{pmatrix}.$$

Is

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

an unbiased estimator of the two regression parameters? Explain.

Solution: By the assumptions

$$E(\mathbf{Y}) = \mathbf{X} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}.$$

Using this and the rules for expectations it follows that the estimate is unbiased.

b. (10) Suggest an unbiased estimator of σ^2.

Solution: One possibility is to use only every second observation and use the usual unbiased estimator for σ^2.

c. (10) Suppose that ρ is known and define new pairs

$$\begin{aligned} \tilde{Y}_{i1} &= (\sqrt{1-\rho} + \sqrt{1+\rho})Y_{i1} + (\sqrt{1-\rho} - \sqrt{1+\rho})Y_{i2} \\ \tilde{Y}_{i2} &= (\sqrt{1-\rho} - \sqrt{1+\rho})Y_{i1} + (\sqrt{1-\rho} + \sqrt{1+\rho})Y_{i2} \end{aligned}$$

$$\begin{aligned} \tilde{x}_{i1} &= (\sqrt{1-\rho} + \sqrt{1+\rho})x_{i1} + (\sqrt{1-\rho} - \sqrt{1+\rho})x_{i2} \\ \tilde{x}_{i2} &= (\sqrt{1-\rho} - \sqrt{1+\rho})x_{i1} + (\sqrt{1-\rho} + \sqrt{1+\rho})x_{i2} \end{aligned}$$

and

$$\begin{aligned} \tilde{\epsilon}_i &= (\sqrt{1-\rho} + \sqrt{1+\rho})\epsilon_i + (\sqrt{1-\rho} - \sqrt{1+\rho})\eta_i \\ \tilde{\eta}_i &= (\sqrt{1-\rho} - \sqrt{1+\rho})\epsilon_i + (\sqrt{1-\rho} + \sqrt{1+\rho})\eta_i \end{aligned}.$$

Define $\tilde{\mathbf{Y}}$ and $\tilde{\mathbf{X}}$ accordingly. The new pairs satisfy the equations

$$\begin{aligned} \tilde{Y}_{i1} &= \alpha_1 + \beta\tilde{x}_{i1} + \tilde{\epsilon}_i \\ \tilde{Y}_{i2} &= \alpha_1 + \beta\tilde{x}_{i2} + \tilde{\eta}_i \end{aligned}$$

where $\alpha_1 = 2\sqrt{1-\rho}\,\alpha$. Argue that this new model satisfies the usual conditions for the regression models. What is then the best linear unbiased estimator of the regression parameters α and β. Explain.

Solution: We need to prove $E(\tilde{\epsilon}_i) = E(\tilde{\eta}_i) = 0$ *which follows easily. By a computation we prove that* $\mathrm{var}(\epsilon_i) = \mathrm{var}(\eta_i) = 4(1-\rho^2)\sigma^2$ *and* $\mathrm{cov}(\tilde{\epsilon}_i, \tilde{\eta}_i) = 0$. *The best linear unbiased estimator for* α_1 *and* β *is given by the Gauss-Markov theorem. But because* α *and* α_1 *differ by a known constant it follows that* $\alpha/(2\sqrt{1-\rho})$ *is the best unbiased estimate for* α.