University of Ljubljana

Doctoral Programme in Statistics

Methodology of Statistical Research

Written examination

February 13th, 2020

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 sampling purposes a population of size N is divided into K strata of sizes $N_1, N_2, \ldots, N_K$. Let μ and σ^2 be the population mean and the population variance. For $i = 1, 2, \ldots, K$ let μ_i and σ_i^2 be the population means and the population variances for individual strata. Assume that a stratified sample is selected such that the sample sizes for individual strata are n_i for $i = 1, 2, \ldots, K$. Denote $w_i = N_i/N$ for $i = 1, 2, \ldots, K$.

a. (10) Let $\bar{Y}_i$ be the sample mean for the i-th stratum. Let $\bar{Y}$ be the unbiased estimator of the population mean

$$\bar{Y} = \sum_{i=1}^{K} w_i \bar{Y}_i \,.$$

Show that

$$E\left[(\bar{Y}_i - \bar{Y})^2\right] = \mathrm{var}(\bar{Y}_i) + \mu_i^2 + \mathrm{var}(\bar{Y}) + \mu^2 - 2\sum_{j=1}^{K}\left(w_j \mu_i \mu_j\right) - 2w_i \mathrm{var}(\bar{Y}_i)\,.$$

Solution: Compute

$$\begin{aligned} E\left[\left(\bar{Y}_i - \bar{Y}\right)^2\right] &= E(\bar{Y}_i^2 - 2\bar{Y}_i\bar{Y} + \bar{Y}^2) \\ &= \mathrm{var}(\bar{Y}_i) + \mu_i^2 + \mathrm{var}(\bar{Y}) + \mu^2 - 2E(\bar{Y}_i\bar{Y})\,. \end{aligned}$$

By independence of $\bar{Y}_1, \bar{Y}_2, \ldots, \bar{Y}_K$ we get

$$\begin{aligned} E(\bar{Y}_i\bar{Y}) &= \sum_{j=1}^{K} w_j E(\bar{Y}_i\bar{Y}_j) \\ &= \sum_{j=1, j\neq i}^{K} w_j\mu_i\mu_j + w_i E(\bar{Y}_i^2) \\ &= \sum_{j=1, j\neq i}^{K} w_j\mu_i\mu_j + w_i(\mathrm{var}(\bar{Y}_i) + \mu_i^2) \\ &= \sum_{j=1}^{K}\left(w_j\mu_i\mu_j\right) + w_i\mathrm{var}(\bar{Y}_i)\,. \end{aligned}$$

b. (15) Let

$$\gamma = \sum_{i=1}^{K} w_i(\mu_i - \mu)^2 = \sum_{i=1}^{K} w_i\mu_i^2 - \mu^2\,.$$

Let

$$\hat{\gamma} = \sum_{i=1}^{K} w_i(\bar{Y}_i - \bar{Y})^2 \,.$$

be an estimator of γ. Modify this estimator to make it an unbiased estimator of γ.

Solution: We compute

$$\begin{aligned}
E(\hat{\gamma}) &= \sum_{i=1}^{K} w_i E\left(\bar{Y}_i - \bar{Y}\right)^2 \\
&= \sum_{i=1}^{K} w_i \left(\operatorname{var}(\bar{Y}_i) + \mu_i^2 + \operatorname{var}(\bar{Y}) + \mu^2 - 2\left(\sum_{j=1}^{K} (w_j \mu_i \mu_j) + w_i \operatorname{var}(\bar{Y}_i) \right) \right) \\
&= \sum_{i=1}^{K} w_i \operatorname{var}(\bar{Y}_i) + \sum_{i=1}^{K} w_i \mu_i^2 + \operatorname{var}(\bar{Y}) + \mu^2 - \\
&\qquad -2 \operatorname{var}(\bar{Y}) - 2 \sum_{i=1}^{K} \sum_{j=1}^{K} w_i w_j \mu_i \mu_j \\
&= \sum_{i=1}^{K} w_i \operatorname{var}(\bar{Y}_i) + \sum_{i=1}^{K} w_i \mu_i^2 + \operatorname{var}(\bar{Y}) + \mu^2 - 2\operatorname{var}(\bar{Y}) - 2\mu^2 \\
&= \gamma + \sum_{i=1}^{K} w_i \operatorname{var}(\bar{Y}_i) - \operatorname{var}(\bar{Y}) \,.
\end{aligned}$$

Both additional terms in the expectation can be estimated in an unbiased way. Subtracting these unbiased estimates from $\hat{\gamma}$ gives an unbiased estimate of γ.

2. (25) The Pareto distribution has the density

$$f(x, \alpha, \lambda) = \frac{\alpha \lambda^\alpha}{(\lambda + x)^{\alpha+1}}$$

for $x > 0$ where $\alpha, \lambda > 0$. Assume the data $x_1, x_2, \ldots, x_n$ are an i.i.d. sample from the Pareto distribution.

a. (10) Write down the equations for the maximum likelihood estimates of the parameters α and λ.

Solution: The log-likelihood function is

$$l(\mathbf{x}, \alpha, \lambda) = n \log(\alpha) + n\alpha \log(\lambda) - (\alpha + 1) \sum_{i=1}^{n} \log(\lambda + x_i)\,.$$

Equate partial derivatives to 0 *to get the equations*

$$\begin{aligned} \frac{\partial l(\mathbf{x},\alpha,\lambda)}{\partial \alpha} &= \frac{n}{\alpha} + n\log(\lambda) - \textstyle\sum_{i=1}^n \log(\lambda + x_i) &= 0 \\ \frac{\partial l(\mathbf{x},\alpha,\lambda)}{\partial \lambda} &= \frac{n\alpha}{\lambda} - (\alpha+1)\textstyle\sum_{i=1}^n \frac{1}{\lambda + x_i} &= 0\,. \end{aligned}$$

b. (15) Compute the approximate standard error of the maximum likelihood estimator $\hat{\alpha}$.

Solution: The second partial derivatives of the density are

$$\begin{aligned} \frac{\partial^2 l(x,\alpha,\lambda)}{\partial \alpha^2} &= -\frac{1}{\alpha^2} \\ \frac{\partial^2 l(x,\alpha,\lambda)}{\partial \lambda^2} &= -\frac{\alpha}{\lambda^2} + \frac{\alpha+1}{(\lambda+x)^2} \\ \frac{\partial^2 l(x,\alpha,\lambda)}{\partial \alpha \partial \lambda} &= \frac{x}{\lambda(\lambda+x)}\,. \end{aligned}$$

Integrating we get

$$I(\alpha, \lambda) = \begin{pmatrix} \frac{1}{\alpha^2} & -\frac{1}{\lambda(\alpha+1)} \\ -\frac{1}{\lambda(\alpha+1)} & \frac{\alpha}{\lambda^2(\alpha+2)} \end{pmatrix}$$

The approximate standard error is

$$se(\hat{\alpha}) = \frac{1}{\sqrt{n}} \sqrt{I_{11}^{-1}}\,,$$

where I_{11}^{-1} *is the element in the upper left corner of the inverse* $I^{-1}(\alpha, \lambda)$.

3. (25) Assume the data $x_1, x_2, \ldots, x_n$ are an i.i.d.sample from the normal distribution. Assume the parameter σ^2 is known. We test $H_0 : \mu = 0$ versus $H_1 : \mu \neq 0$.

a. (10) The null-hypothesis H_0 with a given confidence level α can be tested in two ways:

- H_0 is rejected if $|\bar{X}| > c$ for the value c such that the probability of Type I error if H_0 holds is α.
- Estimate μ and set up a $(1-\alpha)$-confidence interval as $\bar{x} \pm z_{(1-\alpha)/2} \cdot \frac{\sigma}{\sqrt{n}}$ where

$$P(-z_{(1-\alpha)/2} \leq Z \leq z_{(1-\alpha)/2}) = 1 - \alpha$$

  fro $Z \sim \mathrm{N}(0, 1)$. If the interval does not contain 0 reject H_0.

Are the two tests equal? Explain.

Solution: Yes, the two tests are the same since σ^2 is known.

b. (15) Compute the likelihood ratio tests statistics for the testing situation described above. What is the distribution of λ? Is the likelihood ratio test exact? Explain.

Solution: The computation of Λ gives

$$\Lambda = \exp\left(\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2 - x_i^2}{2\sigma^2}\right) .$$

$$\Lambda = \exp\left(-\frac{n\bar{x}^2}{2\sigma^2}\right) .$$

Since σ^2 is known H_0 is rejected if $|\bar{x}| > c$ for a suitable c. The distribution of the test statistic under H_0 is exactly $\chi^2(1)$. The test is exact.

4. (25) The model for the data is described by two sets of regression equations

$$Y_i = \alpha_1 + \beta x_i + \epsilon_i$$

for $i = 1, 2, \ldots, m$ and

$$Z_j = \alpha_2 + \beta w_j + \eta_j$$

for $j = 1, 2, \ldots, n$. For both sets of equations the standard linear regression assumptions hold. This means for all i, j we have $E(\epsilon_i) = E(\eta_j) = 0$, $\mathrm{var}(\epsilon_i) = \sigma^2$ and $\mathrm{var}(\eta_i) = \tau^2$, and all ϵ_i and η_j are uncorrelated. Further assume that

$$\sum_{i=1}^{m} x_i = 0 \qquad \text{in} \qquad \sum_{j=1}^{n} w_j = 0$$

ter

$$\sum_{i=1}^{m} x_i^2 = 1 \qquad \text{in} \qquad \sum_{j=1}^{n} w_j^2 = 1\,.$$

a. (10) Give an unbiased estimate of β based on all the data. What is the standard error of your estimate?

Solution: The two sets of equations are combined into one.

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_m \\ Z_1 \\ Z_2 \\ \vdots \\ Z_n \end{pmatrix} = \begin{pmatrix} 1 & 0 & x_1 \\ 1 & 0 & x_2 \\ \vdots & \vdots & \vdots \\ 1 & 0 & x_m \\ 0 & 1 & w_1 \\ 0 & 1 & w_2 \\ \vdots & \vdots & \vdots \\ 0 & 1 & w_n \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \beta \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_m \\ \eta_1 \\ \eta_2 \\ \vdots \\ \eta_n \end{pmatrix}.$$

Under the assumptions the OLS estimator of β is unbiased. We compute

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} m & 0 & 0 \\ 0 & n & 0 \\ 0 & 0 & 2 \end{pmatrix}.$$

The inverse is

$$\left(\mathbf{X}^T\mathbf{X}\right)^{-1} = \begin{pmatrix} 1/m & 0 & 0 \\ 0 & 1/n & 0 \\ 0 & 0 & 1/2 \end{pmatrix}.$$

We have

$$\mathbf{X}^T\mathbf{Y} = \begin{pmatrix} \sum_{i=1}^m Y_i \\ \sum_{j=1}^n Z_j \\ \sum_{i=1}^m x_i Y_i + \sum_{j=1}^n w_j Z_j \end{pmatrix}.$$

It follows that

$$\hat{\beta} = \frac{1}{2}\left(\sum_{i=1}^{m} x_i Y_i + \sum_{j=1}^{n} w_j Z_j\right).$$

The standard error is

$$\mathrm{se}(\hat{\beta}) = \frac{\sqrt{\sigma^2 + \tau^2}}{2}.$$

b. (15) Assume that $\sigma^2/\tau^2 = \lambda$ for known $\lambda > 0$. Compute the best unbiased linear estimate of β. What is its standard error?

Solution: If we multiply the second set of equations by $\sqrt{\lambda}$ and denote

$$\tilde{Z}_j = \sqrt{\lambda} Z_j, \quad \tilde{\alpha}_2 = \sqrt{\lambda}\alpha_2, \quad \tilde{w}_j = \sqrt{\lambda} w_j \quad \textit{and} \quad \tilde{\eta}_j = \sqrt{\lambda}\eta_j$$

for $j = 1, 2, ..., n$ and combine the two sets of equations into one we get the standard regression model. In this case the OLS estimator is the best unbiased linear estimator of β. However, the matrix $\mathbf{X}$ changes and we get

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} m & 0 & 0 \\ 0 & m & 0 \\ 0 & 0 & 1+\lambda \end{pmatrix}.$$

and

$$\mathbf{X}^T\mathbf{Y} = \begin{pmatrix} \sum_{i=1}^{m} Y_i \\ \sum_{j=1}^{n} \tilde{Z}_j \\ \sum_{i=1}^{m} x_i Y_i + \sum_{j=1}^{n} \tilde{w}_j \tilde{Z}_j \end{pmatrix}.$$

It follows

$$\hat{\beta} = \frac{1}{1+\lambda}\left(\sum_{i=1}^{m} x_i Y_i + \sum_{j=1}^{n} \tilde{w}_j \tilde{Z}_j\right).$$

The standard error is compute directly as

$$\mathrm{se}(\hat{\beta}) = \frac{\sigma}{\sqrt{1+\lambda}}.$$