University of Ljubljana

Doctoral Programme in Statistics

Methodology of Statistical Research

Written examination

June 29th, 2021

Name and surname: ____________________ ID number: | | | | | | | | |

Solutions

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) Suppose the population is stratified into K strata of sizes $N_1, \ldots, N_K$. Denote by μ_k the population mean in stratum k and by σ_k^2 the population variance in stratum k for $k = 1, 2, \ldots, K$. Let μ be the population mean for the whole population and σ^2 the population variance for the whole population. Suppose a stratified sample is taken with sample sizes in each stratum equal to $n_1, n_2, \ldots, n_K$. Let $\bar{X}_k$ be the sample mean in stratum k and let

$$\bar{X} = \sum_{k=1}^{K} \frac{N_k}{N} \bar{X}_k = \sum_{k=1}^{K} w_k \bar{X}_k \,.$$

a. (5) Compute $E\left[\left(\bar{X}_k - \bar{X}\right)^2\right]$.

Solution: We compute

$$\begin{aligned}
E\left[\left(\bar{X}_k - \bar{X}\right)^2\right] &= \operatorname{var}\left(\bar{X}_k - \bar{X}\right) + \left(E\left(\bar{X}_k - \bar{X}\right)\right)^2 \\
&= \operatorname{var}(\bar{X}_k) + \operatorname{var}(\bar{X}) - 2\operatorname{cov}(\bar{X}_k, \bar{X}) + (\mu_k - \mu)^2 \\
&= \frac{\sigma_k^2}{n_k} \cdot \frac{N_k - n_k}{N_k - 1} + \sum_{i=1}^{K} w_i^2 \cdot \frac{\sigma_i^2}{n_i} \cdot \frac{N_i - n_i}{N_i - 1} \\
&\qquad -2w_k \cdot \frac{\sigma_k^2}{n_k} \cdot \frac{N_k - n_k}{N_k - 1} + (\mu_k - \mu)^2 \,.
\end{aligned}$$

b. (10) Suggest an unbiased estimator for the quantity

$$\gamma^2 = \sum_{k=1}^{K} w_k (\mu_k - \mu)^2 \,.$$

Explain why the suggested estimator is unbiased.

Solution: Since we have unbiased estimators for σ_k^2 the quantity

$$\hat{\gamma}_k^2 = \left(\bar{X}_k - \bar{X}\right)^2 - \frac{\hat{\sigma}_k^2}{n_k} \cdot \frac{N_k - n_k}{N_k - 1} - \sum_{i=1}^{K} w_i^2 \cdot \frac{\hat{\sigma}_i^2}{n_i} \cdot \frac{N_i - n_i}{N_i - 1} + 2w_k \cdot \frac{\hat{\sigma}_k^2}{n_k} \cdot \frac{N_k - n_k}{N_k - 1}$$

is an unbiased estimator of $(\mu_k - \mu)^2$. Multiplying γ_k^2 by w_k and summing over k we get an unbiased estimator of γ^2.

c. (10) Suggest an unbiased estimator of the population variance σ^2. Explain why your estimator is unbiased.

Hint: check that

$$\sigma^2 = \sum_{k=1}^{K} w_k \sigma_k^2 + \sum_{k=1}^{K} w_k (\mu_k - \mu)^2 \,.$$

Solution: We write

$$\sigma^2 = \sum_{k=1}^{K} w_k \sigma_k^2 + \gamma^2 \,.$$

Since both terms on the right can be estimated in an unbiased way we have that

$$\hat{\sigma}^2 = \sum_{k=1}^{K} w_k \hat{\sigma}_k^2 + \hat{\gamma}^2$$

is an unbiased estimator of $\hat{\sigma}^2$.

2. (25) Assume the data $x_1, x_2, \ldots, x_n$ are an i.i.d. sample from the distribution with density

$$f(x) = \frac{\alpha}{2}|x|^{\alpha-1}e^{-|x|^\alpha}$$

for $\alpha > 0$.

a. (15) Write the equation for the MLE estimate of α. Compute the Fisher information $I(\alpha)$. Assume as known that

$$\int_0^\infty x^{2\alpha-1} \log^2 x \, e^{-x^\alpha} \, \mathrm{d}x = \frac{\pi^2}{6\alpha^3} - \frac{(2-\gamma)\gamma}{\alpha^3}$$

where $\gamma = 0.577216$ is the Euler constant.

Solution: The log-likelihood function is given by

$$\ell(\alpha|x_1, \ldots, x_n) = n\log(\alpha) - n\log 2 + (\alpha - 1)\sum_{k=1}^n \log|x_k| - \sum_{k=1}^n |x_k|^\alpha .$$

Setting the derivative to 0 *we get the equation*

$$\frac{n}{\alpha} + \sum_{k=1}^n \log|x_k| - \sum_{k=1}^n |x|^\alpha \log|x_k| = 0 .$$

For the Fisher information we compute

$$\ell'' = -\frac{1}{\alpha^2} - |x|^\alpha \log^2|x| .$$

We get

$$\begin{aligned} I(\alpha) &= \frac{1}{\alpha^2} + \frac{\alpha}{2}\int_{-\infty}^\infty |x|^{2\alpha-1}\log^2|x|e^{-|x|^\alpha} \\ &= \frac{1}{\alpha^2} - \frac{\pi^2}{12\alpha^2} - \frac{(2-\gamma)\gamma}{2\alpha^2} . \end{aligned}$$

b. (10) Suppose you knew the MLE estimate $\hat{\alpha}$. Write explicitely the approximate 99%-confidence interval for α.

Rešitev: The approximate standard error is given by

$$\mathrm{se}(\hat{\alpha}) = \sqrt{\frac{1}{nI(\hat{\alpha})}}$$

and $z_\alpha = 2.56$. *The approximate confidence interval is*

$$\hat{\alpha} \pm 2.56 \cdot \mathrm{se}(\hat{\alpha}) .$$

3. (25) Assume the observations $x_1, \ldots, x_n$ are an i.i.d.sample from the $\Gamma(2, \theta)$ distribution with density

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

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

a. (5) Find the maximum likelihood estimator for the parameter θ.

   Solution: The log-likelihood function is

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

   Equating the derivative to 0 *we get*

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

b. (10) For the testing problem $H_0\colon \theta = 1$ versus $H_1\colon \theta \neq 1$ find the Wilks's test statistic λ. Describe when you would reject H_0 given that the size of the test is $1 - \alpha$ with $\alpha \in (0, 1)$.

   Solution: By definition

$$\lambda = 2\ell(\hat{\theta}) - 2\ell(1) \,.$$

   Using the maximum likelihood estimator $\hat{\beta}$ we get

$$\lambda = -4n \log\left(\frac{\bar{x}}{2}\right) + 2n\left(\bar{x} - 2\right) \,.$$

   By Wilks's theorem under H_0 the distribution of the test statistic λ is approximately $\chi^2(1)$. The null-hypothesis is rejected when $\lambda > c_\alpha$ where c_α is such that $P(\chi^2(1) \geq c_\alpha) = \alpha$.

c. (10) The function

$$f(y) = -4n \log\left(\frac{y}{2}\right) + 2n(y - 2)$$

   is strictly decreasing on $(0, 2)$ and strictly increasing on $(2, \infty)$. Assume for all $c > \min_{y>0} f(y)$ you can find the two solutions of the equation $f(y) = c$. Can you use this information to give an exact test given $\alpha \in (0, 1)$? Describe the procedure. No calculations are required.

   Hint: by properties of the gamma distribution $\bar{X} \sim \Gamma(2n, \theta/n)$.

   Solution: Given the assumptions we can find such a c_α that under H_0 we have

$$P_{H_0}\left(f(\bar{X}) \geq c_\alpha\right) = \alpha \,.$$

   Let $x_1 < x_2$ be the solutions of the equation $f(x) = c_\alpha$. The test that rejects H_0 when either $\bar{X} < x_1$ or $\bar{X} > x_2$ is exact.

4. (25) Assume the regression model with

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $E(\boldsymbol{\epsilon}) = 0$ and $\operatorname{var}(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{\Sigma}$ where $\boldsymbol{\Sigma}$ is an invertible known matrix and σ^2 is an unknown parameter.

a. (5) Show that

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$$

is an unbiased estimate of the parameter $\boldsymbol{\beta}$.

Solution: We compute

$$E\left(\hat{\boldsymbol{\beta}}\right) = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T E(\mathbf{Y})\,.$$

Since $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ *we have*

$$E\left(\hat{\boldsymbol{\beta}}\right) = \boldsymbol{\beta}\,.$$

b. (5) Show that

$$\tilde{\boldsymbol{\beta}} = \left(\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{Y}$$

is an unbiased estimate of the parameter $\boldsymbol{\beta}$.

Solution: We compute

$$E\left(\tilde{\boldsymbol{\beta}}\right) = \left(\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{\Sigma}^{-1}E(\mathbf{Y})\,.$$

Since $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ *we have*

$$E\left(\tilde{\boldsymbol{\beta}}\right) = \boldsymbol{\beta}\,.$$

c. (5) Compute the covariance matrix

$$\operatorname{cov}\left(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\beta}}\right)\,.$$

Solution: Denote

$$\mathbf{A} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$$

and

$$\mathbf{B} = \left(\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\,.$$

In this notation

$$\operatorname{cov}\left(\mathbf{AY} - \mathbf{BY}, \mathbf{BY}\right) = (\mathbf{A} - \mathbf{B})\operatorname{cov}(\mathbf{Y}, \mathbf{Y})\mathbf{B}^T\,.$$

Note that $\operatorname{cov}(\mathbf{Y}, \mathbf{Y}) = \sigma^2\boldsymbol{\Sigma}$. *It is straightforward to check that*

$$(\mathbf{A} - \mathbf{B})\boldsymbol{\Sigma}\mathbf{B}^T = 0\,.$$

d. (10) Which of the two estimators for $\boldsymbol{\beta}$ is better? Explain.

Solution: Write as in the Gauss-Markov theorem

$$\begin{aligned}\operatorname{var}(\hat{\boldsymbol{\beta}}) &= \operatorname{var}(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}} + \tilde{\boldsymbol{\beta}}) \\ &= \operatorname{var}(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}) + \operatorname{var}(\tilde{\boldsymbol{\beta}}) + 2\operatorname{cov}\left(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\beta}}\right) \\ &= \operatorname{var}(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}) + \operatorname{var}(\tilde{\boldsymbol{\beta}})\,.\end{aligned}$$

This means that $\tilde{\boldsymbol{\beta}}$ is the better estimator of $\boldsymbol{\beta}$.