University of Ljubljana

Doctoral Programme in Statistics

Methodology of Statistical Research

Written examination

February 1st, 2024

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.	Total
1.					
2.				•	
3.			•	•	
4.				•	
Total					

1. (20) Suppose the population of size N is divided into M subpopulations of size K so that $N = MK$. A sample is selected in two steps: first m subpopulations are selected among the M by simple random sampling. On the second step k units are selected in each subpopulation selected by simple random sampling. The final sample is of size $n = mk$.

a. (5) Is the sample mean an unbiased estimate of the population mean? Explain.

   Solution: every unit in the population will be selected with the same probability. This means that the sample average is an unbiased estimate.

b. (5) Denote for $j = 1, 2, \ldots, M$ by μ_j the j-th subpopulation mean and by σ_j^2 the population variance in the j-th subpopulation and let

$$I_j = \begin{cases} 1 & \text{if the } j\text{-th subpopulation is selected} \\ 0 & \text{else} \end{cases}$$

   and let $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_M$ be the sample means for samples selected in subpopulations. Assume that $\bar{X}_1, \ldots, \bar{X}_M$ are independent and independent of $I_1, \ldots, I_M$. Argue that the sample mean can be written as

$$\bar{X} = \frac{1}{m}\left(\bar{X}_1 I_1 + \bar{X}_2 I_2 + \cdots + \bar{X}_M I_M\right).$$

   Show that

$$\operatorname{var}(\bar{X}_j I_j) == \frac{m}{M}\left(\operatorname{var}(\bar{X}_j) + \frac{M-m}{M}\mu_j^2\right)$$

   and

$$\operatorname{cov}(\bar{X}_j I_j, \bar{X}_l I_l) = -\frac{m}{M}\mu_j \mu_l \cdot \frac{M-m}{M(M-1)}.$$

   Solution: we know that

$$\operatorname{var}(\bar{X}_j) = \frac{\sigma_j^2}{k} \cdot \frac{K-k}{K-1}$$

   and

$$\operatorname{cov}(I_j, I_l) = \frac{m}{M} \cdot \frac{m-1}{M-1} - \left(\frac{m}{M}\right)^2 = -\frac{m(M-m)}{M^2(M-1)}.$$

   We compute

$$\begin{aligned} \operatorname{var}(\bar{X}_j I_j) &= E(\bar{X}_j^2 I_j) - E(\bar{X}_j I_j)^2 \\ &= E(\bar{X}_j^2) E(I_j) - E(\bar{X}_j)^2 E(I_j)^2 \\ &= (\operatorname{var}(\bar{X}_j) + \mu_j^2) \cdot \frac{m}{M} - \mu_j^2 \left(\frac{m}{M}\right)^2 \\ &= \frac{m}{M}\left(\operatorname{var}(\bar{X}_j) + \frac{M-m}{M}\mu_j^2\right). \end{aligned}$$

and

$$\begin{aligned}\operatorname{cov}(\bar{X}_j I_j, \bar{X}_l I_l) &= E(\bar{X}_j I_j \bar{X}_l I_l) - E(\bar{X}_j I_j) E(\bar{X}_l I_l) \\ &= E(\bar{X}_j) E(\bar{X}_l) E(I_j I_l) - E(\bar{X}_j) E(I_j) E(\bar{X}_l) E(I_l) \\ &= \mu_j \mu_l (\operatorname{cov}(I_j, I_l) + E(I_j) E(I_l)) - \mu_j \mu_l \left(\frac{m}{M}\right)^2 \\ &= \mu_j \mu_l \cdot \left(\frac{m(m-1)}{M(M-1)} - \left(\frac{m}{M}\right)^2\right) \\ &= -\frac{m}{M} \mu_j \mu_l \cdot \frac{M-m}{M(M-1)}\,.\end{aligned}$$

c. (10) Show that

$$\operatorname{var}(\bar{X}) = \frac{1}{Mm}\left(\sum_{j=1}^{M} \operatorname{var}(\bar{X}_j) + \frac{M-m}{M-1} \sum_{j=1}^{M} (\mu_j - \mu)^2\right)$$

where μ is the population mean. Assume as known that

$$\sum_{j=1}^{M} \mu_j^2 - \frac{2}{M-1} \sum_{j<l} \mu_j \mu_l = \frac{M}{M-1} \sum_{j=1}^{M} (\mu_j - \mu)^2\,.$$

Solution: we have

$$\begin{aligned}\operatorname{var}(\bar{X}) &= \operatorname{var}\left(\frac{1}{m}\left(\bar{X}_1 I_1 + \bar{X}_2 + \cdots + \bar{X}_M I_M\right)\right) \\ &= \frac{1}{m^2}\left(\sum_{j=1}^{M} \operatorname{var}(\bar{X}_j I_j) + 2 \sum_{j<l} \operatorname{cov}(\bar{X}_j I_j, \bar{X}_l I_l)\right) \\ &= \frac{1}{m^2}\left(\sum_{j=1}^{M} \frac{m}{M}\left(\operatorname{var}(\bar{X}_j) + \frac{M-m}{M}\mu_j^2\right) - 2\sum_{j<l}\left(-\frac{m}{M}\mu_j\mu_l \cdot \frac{M-m}{M(M-1)}\right)\right) \\ &= \frac{1}{Mm}\sum_{j=1}^{M} \operatorname{var}(\bar{X}_j) + \frac{M-m}{M^2 m}\left(\sum_{j=1}^{M} \mu_j^2 - \frac{2}{M-1}\sum_{j<l}\mu_j\mu_l\right) \\ &= \frac{1}{Mm}\left(\sum_{j=1}^{M} \operatorname{var}(\bar{X}_j) + \frac{M-m}{M-1}\sum_{j=1}^{n}(\mu_j - \mu)^2\right)\end{aligned}$$

d. (5) How would you estimate the standard error from the data? Just give the idea with no calculations.

Solution: For the quantities $\text{var}(\bar{X}_j)$ *we only have estimates for* m *selected subpopulations. Multiplying their sum by* m/M *would give an estimate for the average*

$$\frac{1}{Mm}\sum_{j=1}^{M}\text{var}(\bar{X}_j)\,.$$

The sum $\sum_{j=1}^{n}(\mu_j-\mu)^2$ *could be estimated by*

$$c\sum_{j=1}^{m}(\bar{X}_j-\bar{X})^2$$

for some appropriate constant.

2. (25) Assume that our observations are pairs $(x_1, y_1), \ldots, (x_n, y_n)$. We assume that the pairs are independent samples from the distribution with density

$$f(x, y) = e^{-x} \cdot \frac{1}{\sigma\sqrt{2\pi x}} \, e^{-\frac{(y-\theta x)^2}{2\sigma^2 x}}$$

for $x > 0$, $-\infty < y < \infty$ and $\sigma^2 > 0$. Assume as known that the random variable

$$Z = \frac{Y_1 - \theta X_1}{\sqrt{X_1}}$$

is distributed normally as $\mathrm{N}(0, \sigma^2)$ and is independent of X_1.

a. (5) Find maximum likelihood estimates for the parameters θ and σ.

Solution: the log-likelihood function is

$$\ell(\theta, \sigma | \mathbf{x}, \mathbf{y}) = \sum_{k=1}^{n} \left(-\frac{n}{2} \log 2\pi - n \log \sigma - \frac{(y_k - \theta x_k)^2}{2\sigma^2 x_k} \right) .$$

Taking partial derivatives we have

$$\frac{\partial \ell}{\partial \theta} = \sum_{k=1}^{n} \frac{(y_k - \theta x_k)}{\sigma^2}$$
$$\frac{\partial \ell}{\partial \sigma} = -\frac{n}{\sigma} + \sum_{k=1}^{n} \frac{(y_k - \theta x_k)^2}{\sigma^3 x_k}$$

Equating the derivatives to zero, it follow from the first equation that

$$\hat{\theta} = \frac{\sum_{k=1}^{n} y_k}{\sum_{k=1}^{n} x_k} s ,$$

and from the second

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{k=1}^{n} \frac{(y_k - \hat{\theta} x_k)^2}{x_k} .$$

b. (10) Compute the Fisher information matrix and give approximate standard errors for the above estimators.

Solution: compute for $n = 1$*:*

$$\frac{\partial^2 \ell}{\partial \theta^2} = -\frac{x}{\sigma^2} ,$$
$$\frac{\partial^2 \ell}{\partial \theta \, \partial \sigma} = -2\frac{y - \theta x}{\sigma^3} ,$$
$$\frac{\partial^2 \ell}{\partial \sigma^2} = \frac{1}{\sigma^2} - 3\frac{(y - \theta x)^2}{\sigma^4 x} .$$

Replace x by X and y by Y. From the forst part we infer that $X \sim \exp(1)$, hence

$$E\left[\frac{\partial^2 \ell}{\partial \theta^2}(\theta, \sigma | X, Y)\right] = -\frac{E(X)}{\sigma^2} = -\frac{1}{\sigma^2}\,.$$

We compute the other two expectations using the hint:

$$\begin{aligned} E\left[\frac{\partial^2 \ell}{\partial \theta\,\partial \sigma}(\theta, \sigma | X, Y)\right] &= -\frac{2E\big(Z\sqrt{X}\big)}{\sigma^3} = 0\,, \\ E\left[\frac{\partial^2 \ell}{\partial \sigma^2}(\theta, \sigma | X, Y)\right] &= \frac{1}{\sigma^2} - \frac{3E(Z^2)}{\sigma^4} = -\frac{2}{\sigma^2}\,. \end{aligned}$$

The Fisher matrix is

$$I(\theta, \sigma) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{pmatrix}\,,$$

and the approximate standard errors

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

c. (10) Compute the exact standard error of the maximum likelihood estimator $\hat{\theta}$. Assume as known that the density of the pair $(\sum_{k=1}^n X_k, \sum_{k=1}^n Y_k)$ is

$$f(x, y) = \frac{1}{(n-1)!} x^{n-1} e^{-x} \cdot \frac{1}{\sqrt{2\pi x \sigma}} e^{-\frac{(y-\theta x)^2}{2\sigma^2 x}}\,.$$

Solution: we have $\mathrm{var}\big(\hat{\theta}\big) = \dfrac{\sigma^2}{n-1}$. *The standard error follows.*

3. (25) A computer generated m values of a random variable X with values $k = 1,2,3,4$ and n values of a random variable Y with values $k = 1,2,3,4$. Assume that all the values can be considered to be independent. We would like to test whether the two random variables X and Y have the same distribution. Denote by m_1, m_2, m_3, m_4 the numbers of appearances of values $k = 1,2,3,4$ among the generated values for X and similarly denote by n_1, n_2, n_3, n_4 the numbers of appearances of values $k = 1,2,3,4$ among the generated values of Y.

a. (15) Let $p_{1,k} = P(X = k)$ and $p_{2,k} = P(Y = k)$ for $k = 1,2,3,4$. Find the maximum likelihood estimates for the probabilities.

Solution: The log-likelihood function is

$$\ell = n_1 \log p_{1,1} + n_2 \log p_{1,2} + n_3 \log p_{1,3} + n_4 \log p_{1,4} \,.$$

with the side condition $p_{1,1} + p_{1,2} + p_{1,3} + p_{1,4} = 1$. *By the Lagrange method we get that*

$$\hat{p}_{1,k} = \frac{m_k}{m} \,.$$

Similarly

$$\hat{p}_{2,k} = \frac{n_k}{n}$$

for $k = 1,2,3,4$.

b. (10) Find a test statistic to test whether X and Y have the same distribution. Describe the testing procedure to be used.

Solution: We are testing

$$H_0\colon p_{1,k} = p_{2,k} \text{ for k=1,2,3,4} \quad \text{versus} \quad H_1\colon p_{1,k} \neq p_{2,k} \text{ for some } k \,.$$

If H_0 *is true we can "pool" the data and the maximum likelihood estimates are*

$$\hat{p}_k = \frac{m_k + n_k}{m+n} \,.$$

The Wilks' λ is then

$$\lambda = 2\left(\sum_{k=1}^{4} m_k \log \frac{m_k}{m} + n_k \log \frac{n_k}{n} - (m_k + n_k) \log \frac{m_k + n_k}{m+n}\right) \,.$$

The dimensions of parameters are 6 *in the unrestricted case and* 3 *in the restricted case. The λ statistic has the* $\chi^2(3)$ *distribution. Once α is chosen we reject the null-hypothesis if λ is above the critical value.*

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 $E(\epsilon_i) = E(\eta_i) = 0$, $\mathrm{var}(\epsilon_i) = \mathrm{var}(\eta_i) = \sigma^2$ and $\mathrm{cov}(\epsilon_i, \eta_i) = \rho\sigma^2$ 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.