University of Ljubljana

Doctoral Programme in Statistics

Methodology of Statistical Research

Written examination

September 3$^{\text{rd}}$, 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) Assume that every unit in a population of size N has two values of statistical variables. Denote these pairs of values by $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$. The average of all the values

$$\lambda = \frac{1}{2N}\sum_{k=1}^{N}(x_k + y_k)\,.$$

is to be estimated. If the k-th unit is selected, she responds with the value x_k with probability $\frac{1}{2}$, and with value y_k with probability $\frac{1}{2}$ independently of other units and independently of the sampling procedure. The pollsters do not know which of the two values is given.

Assume that a simple random sample of size n is selected from the population. The quantity λ is estimated by the sample average. The estimator is expressed as

$$\hat{\lambda} = \frac{1}{n}\sum_{k=1}^{N} I_k\left(x_k J_k + y_k(1-J_k)\right)\,,$$

where I_k is the indicator that the k-th unit is selected, and J_k is the indicator that the k-th unit's response is x_k. The assumptions imply that the vectors $(I_1, \ldots, I_N)$ and $(J_1, \ldots, J_N)$ are independent, and that the indicators $J_1, \ldots, J_n$ are independent.

a. (5) Show that the estimator $\hat{\lambda}$ is unbiased.

Solution: Use independence and linearity of the expected value to get

$$E(\hat{\lambda}) = \frac{1}{n}\sum_{k=1}^{N} E(I_k)\left(x_k E(J_k) + y_k E(1-J_k)\right) = \lambda\,.$$

b. (10) Show that for $k = 1, 2, \ldots, N$

$$\operatorname{var}\left(I_k(x_k J_k + y_k(1-J_k))\right) = \frac{n}{N}\left(\frac{x_k^2 + y_k^2}{2}\right) - \frac{n^2}{N^2}\left(\frac{x_k + y_k}{2}\right)^2\,.$$

Solution: From simple random sampling we know that $E(I_k) = \frac{n}{N}$. This implies that

$$E\left[I_k\left(x_k J_k + y_k(1-J_k)\right)\right] = \frac{n}{N}\left(\frac{x_k + y_k}{2}\right)\,.$$

Using the facts that $I_k^2 = I_k$, $J_k^2 = J_k$ and $J_k(1-J_k) = 0$ we get

$$\begin{aligned} E\left[I_k^2\left(x_k J_k + y_k(1-J_k)\right)^2\right] &= E\left[I_k\left(x_k^2 J_k + y_k^2(1-J_k)\right)\right] \\ &= \frac{n}{N}\left(\frac{x_k^2 + y_k^2}{2}\right)\,. \end{aligned}$$

The formula for the variance follows.

c. (10) Show that for $k \neq l$

$$\operatorname{cov}\left(I_k(x_k J_k + y_k(1 - J_k)), I_l(x_l J_l + y_l(1 - J_l))\right) = \frac{n(n-1)}{4N(N-1)}(x_k + y_k)(x_l + y_l)\,.$$

Solution: From simple random sampling we know that

$$\operatorname{cov}(I_k, I_l) = -\frac{n(N-n)}{N^2(N-1)}\,.$$

This implies that

$$E(I_k I_l) = -\frac{n(N-n)}{N^2(N-1)} + \frac{n^2}{N^2} = \frac{n(n-1)}{N(N-1)}\,.$$

Use the linearity of expected value and independence assumptions to compute

$$\begin{aligned}
& E\left[\left(I_k(x_k J_k + y_k(1 - J_k)\right)\left(I_l(x_l J_l + y_l(1 - J_l)\right)\right] \\
= \quad & \frac{x_k x_l}{4} E(I_k I_l) + \frac{x_k y_l}{4} E(I_k I_l) + \frac{x_l y_k}{4} E(I_k I_l) + \frac{y_k y_l}{4} E(I_k I_l) \\
= \quad & \frac{n(n-1)}{4N(N-1)}(x_k + y_k)(x_l + y_l)\,.
\end{aligned}$$

2. (25) Let the observed values $x_1, x_2, \ldots, x_n$ be generated as independent, identically distributed random variables $X_1, X_2, \ldots, X_n$ with distribution

$$P(X_1 = x) = \frac{(\theta - 1)^{x-1}}{\theta^x}$$

for $x = 1, 2, 3, \ldots$ and $\theta > 1$.

a. (10) Find the MLE estimate of θ based on the observations.

Solution: We find

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

Taking the derivative we have

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

It follows that

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

b. (15) Write an approximate 99%-confidence interval for θ based on the observations. Assume as known that

$$\sum_{x=1}^{\infty} x a^{x-1} = \frac{1}{(1-a)^2}$$

for $|a| < 1$.

Solution: We have

$$\ell''(\theta, x) = -\frac{x-1}{(\theta - 1)^2} + \frac{x}{\theta^2} \, .$$

To find the Fisher information we need

$$E(X_1) = \sum_{x=1}^{\infty} x \frac{(\theta - 1)^{x-1}}{\theta^x} \, .$$

Using the hint we get

$$E(X_1) = \frac{1}{\theta} \cdot \left(1 - \frac{\theta - 1}{\theta}\right)^{-2} = \theta \, .$$

We have

$$I(\theta) = \frac{1}{\theta(\theta - 1)} \, .$$

An approximate 99%-confidence interval is

$$\hat{\theta} \pm 2.56 \cdot \sqrt{\frac{\hat{\theta}(\hat{\theta} - 1)}{n}} \, .$$

3. (25) Assume that the observed values $x_1, x_2, \ldots, x_m$ and $y_1, y_2, \ldots, y_n$ were created as independent random variables $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ with $X_k \sim \exp(\mu)$ for $k = 1, 2, \ldots, m$ and $Y_k \sim \exp(\nu)$ for $k = 1, 2, \ldots, n$. The hypothesis

$$H_0\colon \mu = \nu \qquad \text{versus} \qquad H_1\colon \mu \neq \nu$$

is to be tested. Assume that $\mu, \nu > 0$.

a. (15) Find the Wilks likelihood ratio statistics λ for this testing problem.

Solution: The log-likelihood functions is

$$\ell(\mu, \nu | \mathbf{x}, \mathbf{y}) = m \log \mu - \mu \sum_{k=1}^{m} x_k + n \log \nu - \nu \sum_{k=1}^{n} y_k \,.$$

If μ and ν can vary freely, the maximum is attained at

$$\hat{\mu} = \frac{m}{\sum_{k=1}^{m} x_k} = \frac{1}{\bar{x}} \quad \textit{and} \quad \hat{\nu} = \frac{n}{\sum_{k=1}^{n} y_k} = \frac{1}{\bar{y}} \,.$$

Evaluating the log-likelihood function at the MLE estimates gives

$$\ell(\hat{\nu}, \hat{\mu} | \mathbf{x}, \mathbf{y}) = m \log \hat{\mu} - m + n \log \hat{\nu} - n \,.$$

If $\nu = \mu$ the MLE turns out to be

$$\tilde{\mu} = \tilde{\nu} = \frac{m+n}{\sum_{k=1}^{n} x_k + \sum_{k=1}^{n} y_k}$$

and

$$\ell(\tilde{\mu}, \tilde{\nu} | \mathbf{x}, \mathbf{y}) = (m+n) \log \tilde{\mu} - m - n \,.$$

It follows that

$$\lambda = 2m \log \hat{\mu} + 2n \log \hat{\nu} - 2(m+n) \log \tilde{\mu} \,.$$

b. (5) What is the approximate distribution of the Wilk's likelihood statistics?

Solution: Bt Wilks' theorem the approximate distribution is $\chi^2(1)$.

4. (25) Assume the following regression model

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

for $i = 1, 2, \ldots, n$. Assume that the pairs $(\epsilon_1, \eta_1), \ldots, (\epsilon_n, \eta_n)$ are independent and identically distributed with $E(\epsilon_i) = E(\eta_i) = 0$, $\mathrm{var}(\epsilon_i) = \mathrm{var}(\eta_i) = \sigma^2$ and $\mathrm{corr}(\epsilon_i, \eta_i) = \rho$. Assume that ρ is known.

a. (5) Let

$$\hat{\beta} = \frac{\sum_{i=1}^n (Y_{i1}x_{i1} + Y_{i2}x_{i2})}{\sum_{i=1}^n (x_{i1}^2 + x_{i2}^2)}\,.$$

Is this estimator unbiased? Compute its standard error.

Solution: All the estimators in the sequel are of the form

$$\hat{\beta} = \sum_{i=1}^n (a_i Y_{i1} + b_i Y_{i2})$$

for suitable a_i and b_i. We have

$$E(\hat{\beta}) = \beta \sum_{i=1}^n (a_i x_{i1} + b_i x_{i2})$$

and

$$\mathrm{var}(\hat{\beta}) = \sum_{i=1}^n \mathrm{var}(a_i Y_{i1} + b_i Y_{i2}) = \sigma^2 \sum_{i=1}^n (a_i^2 + b_i^2 + 2\rho a_i b_i)\,.$$

Plugging in the respective a_i and b_i we find that all the estimators are unbiased and we derive the formulae for standard errors.

b. (5) Adding we get

$$Y_{i1} + Y_{i2} = \beta(x_{i1} + x_{i2}) + \xi_i\,,$$

where $\xi_i = \epsilon_i + \eta_i$. The terms $\xi_1, \ldots, \xi_n$ are uncorrelated with $E(\xi_i) = 0$ and $\mathrm{var}(\xi_i) = \sigma^2(2 + \rho)$. The parameter β can be estimated as

$$\hat{\beta} = \frac{\sum_{i=1}^n (Y_{i1} + Y_{i2})(x_{i1} + x_{i2})}{\sum_{i=1}^n (x_{i1} + x_{i2})^2}\,.$$

Is this estimator unbiased? Compute ist standard error.

Solution: See a.

c. (5) Replace for each $i = 1, 2, \ldots, n$ the second equation by

$$\frac{Y_{i2} - \rho Y_{i1}}{2(1-\rho)} = \beta\left(\frac{x_{i2} - \rho x_{i1}}{2(1-\rho)}\right) + \left(\frac{\eta_i - \rho\epsilon_i}{2(1-\rho)}\right)\,.$$

Denote

$$\tilde{Y}_{i2} = \frac{Y_{i2} - \rho Y_{i1}}{2(1-\rho)} \qquad \text{in} \quad \tilde{x}_{i2} = \frac{x_{i2} - \rho x_{i1}}{2(1-\rho)}\,.$$

Estimate β by

$$\hat{\beta} = \frac{\sum_{i=1}^n (Y_{i1}x_{i1} + \tilde{Y}_{i2}\tilde{x}_{i2})}{\sum_{i=1}^n (x_{i1}^2 + \tilde{x}_{i2}^2)}\,.$$

Is this estimate unbiased? Compute its standard error.

Solution: See a.

d. (10) Which of the above estimators has the smallest standard error? Explain.

Solution: Let

$$\tilde{\eta}_i = \frac{\eta_i - \rho\epsilon_i}{2(1-\rho)}\,.$$

This random variable is uncorrelated with ϵ_i *and* $E(\tilde{\eta}_i) = 0$ *and* $\mathrm{var}(\tilde{\eta}_i) = \sigma^2$. *The model in c. satisfies all the assumptions of the Gauss-Markov theorem which means that the estimator in c. is the best linear unbiased estimator of the parameters.*