University of Ljubljana

Doctoral Programme in Statistics

Methodology of Statistical Research

Written examination

February 10th, 2023

Name and surname: ____________________ ID number: ☐☐☐☐☐☐☐☐

Instructions

Read carefully the wording of the problem before you start. There are four problems altogether. 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 a stratified sample is taken from a population of size N. The strata are of size $N_1, N_2, \ldots, N_K$, and the simple random samples are of size $n_1, n_2, \ldots, n_K$. Denote by μ the population mean and by σ^2 the population variance for the entire population, and by μ_k and σ_k^2 the population means and the population variances for the strata.

a. (5) Show that

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

where $w_k = \frac{N_k}{N}$ for $k = 1, 2, \ldots, K$.

Solution: by definition we have

$$\sigma^2 = \frac{1}{N} \left(\sum_{k=1}^{K} \sum_{i=1}^{N_k} (y_{ki} - \mu)^2 \right)$$

where y_{ki} is the value for the i-th unit in the k-th stratum. Note that

$$\begin{aligned}
&\sum_{i=1}^{N_k} (y_{ki} - \mu)^2 = \\
&\quad = \sum_{i=1}^{N_k} (y_{ki} - \mu_k + \mu_k - \mu)^2 \\
&\quad = \sum_{i=1}^{N_k} (y_{ki} - \mu_k)^2 + \sum_{i=1}^{N_k} (\mu_k - \mu)^2 + 2(\mu_k - \mu) \sum_{i=1}^{N_k} (y_{ki} - \mu) \\
&\quad = \sum_{i=1}^{N_k} (y_{ki} - \mu_k)^2 + \sum_{i=1}^{N_k} (\mu_k - \mu)^2 \\
&\quad = N_k \sigma_k^2 + N_k (\mu_k - \mu)^2 .
\end{aligned}$$

Using this in the above summation gives the result.

b. (10) Let $\bar{Y}_k$ be the sample average in the k-th stratum for $k = 1, 2, \ldots, K$ and $\bar{Y} = \sum_{k=1}^{K} w_k \bar{Y}_k$ the unbiased estimator of the population mean. The estimators $\bar{Y}_1, \ldots, \bar{Y}_n$ are assumed to be independent. To estimate σ^2, we need to estimate the quantity

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

The estimator

$$\hat{\sigma}_b^2 = \sum_{k=1}^{K} w_k \bar{Y}_k^2 - \bar{Y}^2$$

is suggested. Show that

$$E(\hat{\sigma}_b^2) = \sum_{k=1}^{K} w_k(1-w_k)\mathrm{var}(\bar{Y}_k) + \sum_{k=1}^{K} w_k \mu_k^2 - \mu^2 \,.$$

Solution: we know that

$$E(\bar{Y}_k^2) = \mathrm{var}(\bar{Y}_k^2) + \mu_k^2$$

and

$$E(\bar{Y}^2) = \mathrm{var}(\bar{Y}) + \mu^2 \,.$$

We have

$$E(\hat{\sigma}_b^2) = \sum_{k=1}^{K} w_k \left(\mathrm{var}(\bar{Y}_k^2) + \mu_k^2\right) - \mathrm{var}(\bar{Y}) - \mu^2 \,.$$

Taking into account that

$$\mathrm{var}(\bar{Y}) = \sum_{k=1}^{K} w_k^2 \mathrm{var}(\bar{Y}_k) \,,$$

the result follows.

c. (10) Is there an unbiased estimator of σ^2? Explain your answer.

Solution: we know that

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

We have unbiased estimators for σ_k^2. The second term can be estimated by

$$\sum_{k=1}^{K} w_k \bar{Y}_k^2 - \bar{Y}^2 - \sum_{k=1}^{K} w_k (1-w_k) \frac{\hat{\sigma}_k^2}{n_k} \cdot \frac{N_k - n_k}{N_k - 1} \,.$$

This last term is an unbiased estimator of the second term.

2. (25) Gauss's gamma distribution is given by the density

$$f(x,y) = \sqrt{\frac{\nu}{2\pi}}\,\sqrt{y}\,e^{-y}\,e^{-\frac{\nu y(x-\mu)^2}{2}}\,.$$

for $-\infty < x < \infty$ and $y > 0$ and $(\mu,\nu) \in \mathbb{R} \times (0,\infty)$. Assume that the observations are pairs $(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$ generated as independent random pairs $(X_1,Y_1),\ldots,(X_n,Y_n)$ with density $f(x,y)$.

a. (10) Compute the maximum likelihood estimates of the parameters.

   Solution: the log-likelihood function is

$$\ell = \frac{n}{2}\log\left(\frac{2\nu}{\pi}\right) + \sum_{k=1}^{n}(\frac{1}{2}\log y_k - y_k) - \frac{\nu}{2}\sum_{k=1}^{n} y_k(x_k-\mu)^2\,.$$

   Set the partial derivatives to 0 *to get*

$$\frac{n}{2\nu} - \frac{1}{2}\sum_{k=1}^{n} y_k(x_k-\mu)^2 = 0$$

   and

$$\nu\sum_{k=1}^{n} y_k(x_k-\mu) = 0\,.$$

   The second equation gives

$$\hat{\mu} = \frac{\sum_{k=1}^{n} x_k y_k}{\sum_{k=1}^{n} y_k}\,.$$

   Insert $\hat{\mu}$ *into the second equation to get*

$$\hat{\nu} = \frac{n}{\sum_{k=1}^{n} y_k(x_k-\hat{\mu})^2}\,.$$

b. (10) Find the Fisher information matrix. Assume as known that $E(XY) = \mu$. Compute $E(Y)$ yourself by computing the marginal density of Y.

   Solution: we compute the second partial derivatives of the likelihood function for $n = 1$*:*

$$\begin{aligned}
\frac{\partial^2\ell}{\partial\mu^2} &= -\nu y_1 \\
\frac{\partial^2\ell}{\partial\nu^2} &= -\frac{1}{2\nu^2} \\
\frac{\partial^2\ell}{\partial\mu\partial\nu} &= y_1(x_1-\mu)
\end{aligned}$$

Integrating the density with respect to x gives that $Y \sim \exp(1)$, and hence $E(Y_1) = 1$. It follows that

$$I(\mu, \nu) = \begin{pmatrix} \nu & 0 \\ 0 & \frac{1}{2\nu^2} \end{pmatrix} .$$

c. (5) Give the approximate standard error of the maximum likelihood estimates.

Solution: using the Fisher's information matrix gives

$$\mathrm{se}(\hat{\mu}) \approx \frac{1}{\sqrt{n\nu}} \quad \textit{and} \quad \mathrm{se}(\hat{\nu}) \approx \frac{\sqrt{2}\nu}{\sqrt{n}} .$$

3. (20) Assume that your observations are pairs $(x_1, y_1), \ldots, (x_n, y_n)$. Assume the pairs are an i.i.d. sample from the bivariate normal density

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-\frac{(x-\mu)^2 - 2\rho(x-\mu)(y-\nu) + (y-\nu)^2}{2(1-\rho^2)}} .$$

Assume that $\rho \in (-1, 1)$ is known. We would like to test the hypothesis

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

a. (10) Find the maximum likelihood estimates for μ and ν.

Solution: derivation, after cancelling constants, gives the equations

$$\begin{aligned} \sum_{k=1}^n (x_k - \mu) - \rho \sum_{k=1}^n (y_k - \nu) &= 0 \\ -\rho \sum_{k=1}^n (x_k - \mu) + \sum_{k=1}^n (y_k - \nu) &= 0 \end{aligned}$$

Dividing by n and rearranging yields

$$\begin{aligned} \mu - \rho\nu &= \bar{x} - \rho\bar{y} \\ -\rho\mu + \nu &= -\rho\bar{x} + \bar{y} \end{aligned}$$

The solutions are $\hat{\mu} = \bar{x}$ and $\hat{\nu} = \bar{y}$. If $\mu = \nu$, the log-likelihood function becomes

$$\log\left(\frac{1}{2\pi\sqrt{1-\rho^2}}\right) - \frac{1}{2(1-\rho^2)} \sum_{k=1}^n \Big((x_k - \mu)^2 - 2\rho(x_k - \mu)(y_k - \mu) + (y_k - \mu)^2 \Big) .$$

Taking derivatives we get

$$\frac{1}{2(1-\rho^2)} \sum_{k=1}^n \Big(-2(x_k - \mu) + 2\rho(y_k - \mu) + 2\rho(x_k - \mu) - 2(y_k - \mu) \Big) .$$

Equating to zero yields

$$2n(1-\rho)\mu = (1-\rho) \sum_{k=1}^n (x_k + y_k) ,$$

and

$$\tilde{\mu} = \tilde{\nu} = \frac{1}{2n} \sum_{k=1}^n (x_k + y_k) .$$

b. (10) Find the likelihood ratio statistic for testing the above hypothesis. What is the approximate distribution of the test statistic under H_0?

Solution: we have

$$\lambda = 2\ell(\hat{\mu}, \hat{\nu}) - 2\ell(\tilde{\mu}, \tilde{\nu})\,.$$

Denote

$$\bar{z} = \frac{\bar{x} + \bar{y}}{2}\,.$$

Using the above estimates yields

$$\begin{aligned}\lambda \;&=\; \frac{1}{(1-\rho^2)}\Big((x_k - \bar{x})^2 - 2\rho(x_k - \bar{x})(y_k - \bar{y}) + (y_k - \bar{y})^2\Big) \\ &\quad - \frac{1}{(1-\rho^2)}\Big((x_k - \bar{z})^2 - 2\rho(x_k - \bar{z})(y_k - \bar{z}) + (y_k - \bar{z})^2\Big)\,.\end{aligned}$$

After some manipulation we get

$$\lambda = \frac{1}{1-\rho^2}\left(-n(\bar{x}^2 - 2\rho\bar{x}\bar{y} + \bar{y}^2) + 2n(1-\rho)\bar{z}^2\right)\,.$$

The approximate distribution of λ under H_0 is $\chi^2(1)$.

c. (5) What is the distribution of $\bar{X} - \bar{Y}$ if H_0 holds? Can you use the result to give an alternative test statistic to test the above hypothesis? What is the distribution of your test statistic under H_0?

Solution: if H_0 holds, we have $\sqrt{n}\left(\bar{X} - \bar{Y}\right) \sim \mathrm{N}(0, 2(1-\rho))$. An alternative test statistic would be

$$Z = \frac{\sqrt{n}\left(\bar{X} - \bar{Y}\right)}{\sqrt{2(1-\rho)}}$$

which is standard normal. We reject H_0 if $|Z| \geq z_\alpha$ where z_α is such that $P(|Z| \geq z_\alpha) = \alpha$.

4. (25) Assume the regression equations are

$$Y_k = \alpha + \beta x_k + \epsilon_k$$

for $k = 1, 2, \ldots, n$. The error terms satisfy the assumptions that

$$E(\epsilon_k) = 0 \quad \text{and} \quad \text{var}(\epsilon_k) = \sigma^2(1 + \tau^2)$$

for $k = 1, 2, \ldots, n$, and

$$\text{cov}(\epsilon_k, \epsilon_l) = \sigma^2\tau^2$$

for $k \neq l$, where τ^2 is assumed to be a known constant. Assume that $\sum_{k=1}^{n} x_k = 0$.

a. (10) Denote $\bar{Y} = \frac{1}{n}\sum_{k=1}^{n} Y_k$. Compute

$$\text{cov}\left(Y_k - c\bar{Y}, Y_l - c\bar{Y}\right)$$

for $k \neq l$. Here c is an arbitrary constant.

Solution: from the assumptions we have

$$\text{cov}\left(Y_k, \bar{Y}\right) = \frac{\sigma^2}{n}\left(1 + n\tau^2\right)$$

and

$$\text{cov}\left(\bar{Y}, \bar{Y}\right) = \frac{\sigma^2}{n}\left(1 + n\tau^2\right) .$$

We have

$$\begin{aligned} & \text{cov}\left(Y_k - c\bar{Y}, Y_l - c\bar{Y}\right) \\ = \ & \text{cov}(Y_k, Y_l) - 2c \cdot \text{cov}\left(Y_k, \bar{Y}\right) + c^2 \cdot \text{cov}\left(\bar{Y}, \bar{Y}\right) \\ = \ & \sigma^2\left(\tau^2 - \frac{2c}{n}\left(1 + n\tau^2\right) + \frac{c^2}{n}\left(1 + n\tau^2\right)\right) . \end{aligned}$$

b. (10) Find an explicit formula for the best linear unbiased estimator of β.

Hint: choose

$$c = 1 - \sqrt{\frac{1}{1 + n\tau^2}} .$$

Solution: with the above choice of c we have that $c \in (0, 1)$ and

$$\text{cov}\left(Y_k - c\bar{Y}, Y_l - c\bar{Y}\right) = 0$$

for $k \neq l$. Define

$$\tilde{Y}_k = Y_k - c\bar{Y} ,$$

$$\tilde{\epsilon}_k = \epsilon_k - c\bar{\epsilon}$$

and

$$\tilde{\mathbf{X}} = \begin{pmatrix} 1-c & x_1 \\ 1-c & x_2 \\ \vdots & \vdots \\ 1-c & x_n \end{pmatrix} .$$

We have

$$\tilde{Y}_k = \alpha(1-c) + \beta x_k + \tilde{\epsilon}_k$$

for $k = 1, 2, \ldots, n$. The new regression equations satisfy the usual assumptions of the Gauss-Markov theorem. The best linear estimators of the regression parameters are

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = \begin{pmatrix} n(1-c)^2 & 0 \\ 0 & \sum_{k=1}^n x_k^2 \end{pmatrix}^{-1} \begin{pmatrix} (1-c)\sum_{k=1}^n \tilde{Y}_k \\ \sum_{k=1}^n x_k \tilde{Y}_k \end{pmatrix} .$$

We get

$$\hat{\beta} = \frac{\sum_{k=1}^n x_k \tilde{Y}_k}{\sum_{k=1}^n x_k^2} . = \frac{\sum_{k=1}^n x_k Y_k}{\sum_{k=1}^n x_k^2} .$$

The last equality follows from the assumption $\sum_{k=1}^n x_k = 0$.

c. (5) Compute the variance of the best linear unbiased estimator $\hat{\beta}$.

Solution: we compute directly

$$\begin{aligned} \mathrm{var}(\hat{\beta}) &= \mathrm{var}\left(\frac{\sum_{k=1}^n x_k Y_k}{\sum_{k=1}^n x_k^2}\right) \\ &= \frac{\sigma^2}{\left(\sum_{k=1}^n x_k^2\right)^2}\left(\sum_{k=1}^n x_k^2(1+\tau^2) + \sum_{\substack{k,l \\ k\neq l}} x_k x_l \tau^2\right) \\ &= \frac{\sigma^2}{\left(\sum_{k=1}^n x_k^2\right)^2}\sum_{k=1}^n x_k^2(1+\tau^2) \\ &= \frac{\sigma^2(1+\tau^2)}{\sum_{k=1}^n x_k^2} \end{aligned}$$