University of Ljubljana, School of Economics

Quantitative finance and actuarial science

Probability and statistics

Written examination

February 10^{th}, 2023

Name and surname: ______________________ ID: | | | | | | | | |

Instructions

Read the problems carefull before staring your work. There are four problems. You can have a sheet with formulae and a mathematical handbook. Write your solutions on the paper 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

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

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) Assume the data $x_1, x_2, \ldots, x_n$ are an i.i.d. sample from the distribution given by

$$P(X_1 = x) = \binom{2x}{x} \frac{\beta^x}{4^x (1+\beta)^{x+\frac{1}{2}}}$$

for $x = 0, 1, \ldots$ and $\beta > 0$.

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

Solution: the log-likelihood function is given by

$$\ell(\beta|\mathbf{x}) = \sum_{k=1}^{n} \log \binom{2x_k}{x_k} + \log\beta \sum_{k=1}^{n} x_k - \log 4 \sum_{k=1}^{n} - \log(1+\beta) \sum_{k=1}^{n} \left(x_k + \frac{1}{2}\right) .$$

Taking derivatives and equating with 0 *we get the equation*

$$\frac{1}{\beta} \sum_{k=1}^{n} x_k - \frac{1}{1+\beta} \sum_{k=1}^{n} \left(x_k + \frac{1}{2}\right) = 0 .$$

Hence

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

b. (5) Convince yourself that

$$\begin{aligned} E(X_1) &= \sum_{k=0}^{\infty} kP(X_1 = k) \\ &= \frac{2\beta}{4(1+\beta)} \sum_{k=1}^{\infty} [2(k-1)+1] P(X_1 = k-1) \\ &= \frac{2\beta}{4(1+\beta)} 2E(X_1) + \frac{2\beta}{4(1+\beta)} . \end{aligned}$$

Use this to show that the maximum likelihood estimator is unbiased.

Solution: the equality can be checked by a straightforward computation. The equality transforms into

$$E(X_1) = \frac{\beta}{1+\beta} E(X_1) + \frac{\beta}{2(1+\beta)}$$

or

$$E(X_1) = \frac{\beta}{2} .$$

We have

$$E\left(\hat{\beta}\right) = E\left(\frac{2\sum_{k=1}^{n} X_k}{n}\right) = \beta ,$$

hence the estimator is unbiased.

c. (5) Use the Fisher information to give an approximate standard error for the maximum likelihood estomator.

Solution: compute for $n = 1$:

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

hence

$$E\left(-\ell''\right) = \frac{1}{2\beta} + \frac{\frac{\beta}{2} + \frac{1}{2}}{(1+\beta)^2} = \frac{1}{2} \cdot \frac{1}{\beta(\beta+1)}\,.$$

It follows that

$$\hat{\mathrm{se}}(\hat{\beta}) = \frac{\sqrt{2\beta(1+\beta)}}{\sqrt{n}}\,.$$

d. (10) Convince yourself that

$$\begin{aligned} E(X_1^2) &= \sum_{k=0}^{\infty} k^2 P(X_1 = k) \\ &= \frac{\beta}{4(1+\beta)} \sum_{k=1}^{\infty} \left[4(k-1)^2 + 6(k-1) + 2\right] P(X_1 = k-1) \\ &= \frac{\beta}{4(1+\beta)} \left(4E(X_1^2) + 6E(X_1) + 2\right)\,. \end{aligned}$$

Compute the exact standard error of the maximum likelihood estimator.

Solution: the equality is checked by a straightforward caclulation. We get the equation

$$E(X_1^2)(1+\beta) = \beta E(X_1^2) + \frac{3\beta}{2} E(X_1) + \frac{\beta}{2}$$

or

$$E(X_1^2) = \frac{\beta(2+3\beta)}{4}$$

and as a consequence

$$\mathrm{var}\,(X_1) = \frac{\beta(1+\beta)}{2}\,.$$

the exact variance of the estimator $\hat{\beta}$ is

$$\mathrm{var}(\hat{\beta}) = \frac{2\beta(1+\beta)}{n^2}$$

3. (25) 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. (5) 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. (10) What is the distribution of $X - Y$ 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 $X - Y \sim \mathrm{N}(0, 2(1-\rho))$. An alternative test statistic would be

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

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

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

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

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

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

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

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

Solution: from the assumptions we have

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

and

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

We have

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

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

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

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

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