School of Economics

Doctoral Programme in Economics

Probability and statistics

Written examination

July 7th, 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. (20) The population of interest has N units. For every unit there are two statistical variables: denote their values by $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$, where $y_k \in \{0, 1\}$ for all $k = 1, 2, \ldots, N$. Assume that $x_1, x_2, \ldots, x_N$ are known in advance from a full census. The quantity of interest is

$$\gamma = \frac{\sum_{k=1}^{N} x_k y_k}{\sum_{k=1}^{N} x_k}.$$

To estimate γ, we take a simple random sample of size $n \leq N$. Denote

$$I_k = \begin{cases} 1 & \text{if unit } k \text{ is chosen;} \\ 0 & \text{else;} \end{cases}$$

a. (5) Let

$$\hat{\gamma} = \frac{N}{n} \frac{\sum_{k=1}^{N} x_k y_k I_k}{\sum_{k=1}^{N} x_k}.$$

Show that $\hat{\gamma}$ is an unbiased estimator of γ.

Solution: we know that $E(I_k) = n/N$. Using this and the linearity of expectation gives that $\hat{\gamma}$ is unbiased.

b. (5) Compute the standard error of $\hat{\gamma}$.

Solution: if we denote

$$z_k = \frac{x_k y_k}{\sum_{i=1}^{N} x_k}$$

then the sampling procedure is just like simple random sampling from the population with the statistical variable with values $z_1, z_2, \ldots, z_N$. We know that

$$\operatorname{var}\left(\frac{1}{n} \sum_{k=1}^{N} z_k I_k\right) = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$$

where

$$\sigma^2 = \frac{1}{N} \sum_{k=1}^{N} (z_k - \bar{z})^2.$$

It follows that

$$\operatorname{var}(\hat{\gamma}) = \frac{N^2 \sigma^2}{n^3} \cdot \frac{N-n}{N-1}.$$

c. (10) Let

$$p = \frac{1}{N} \sum_{k=1}^{N} y_k$$

and

$$\hat{p} = \frac{1}{n} \sum_{k=1}^{N} y_k I_k \,.$$

Assume that $J_1, J_2, \ldots, J_N$ are indicators which, given $I_1, \ldots, I_N$, are conditionally independent with

$$P(J_k = 1 | I_1, \ldots, I_N) = \frac{1}{n} \sum_{l=1}^{N} y_l I_l \,.$$

Assume as known that

$$E\left(I_k J_k\right) = \frac{p(Np-1)}{n(N-1)} + \frac{y_k p}{n}$$

and

$$E(J_k) = p \,.$$

Consider the alternative "bootstrap" estimator

$$\tilde{\gamma} = \frac{\sum_{k=1}^{N} x_k y_k I_k + x_k (1 - I_k) J_k}{\sum_{k=1}^{n} x_k} \,.$$

Is $\tilde{\gamma}$ is an unbiased estimator of γ?

Solution: we compute

$$\begin{aligned} E\left((1 - I_k) J_k\right) &= E(J_k) - E(I_k J_k) \\ &= p - \frac{p(Np-1)}{n(N-1)} + \frac{y_k p}{n} \,. \end{aligned}$$

Compute

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

In general, $\tilde{\gamma}$ is not an unbiased estimator.

d. (5) Is it possible to adjust $\tilde{\gamma}$ to make it an unbiased estimator? Just give the idea. No calculations necessary.

Solution: since the sum $\sum_{k=1}^{N} x_k$ is assumed known, the question is whether we can produce an unbiased estimate of p and $\hat{p}$. The answer is positive in both cases, as we know from simple random sampling theory. With this, the above estimator can be adjusted.

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

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

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 $n = 1$*:*

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

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

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

We get

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

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

Solution: we compute directly

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