School of Economics

Doctoral Programme in Economics

Probability and statistics

Written examination

January 26th, 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 $n \le N$. 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

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

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 bothe cases, as we know from simple random sampling theory. With this, the above estimator can be adjusted.

2. (25) Assume the observed values $x_1, x_2, \ldots, x_n$ were generated as random variables $X_1, X_2, \ldots, X_n$ with density

$$f(x) = \frac{1}{\sqrt{2\pi x^3}}\, e^{-\frac{(1-\mu x)^2}{2x}}$$

for $x, \mu > 0$.

a. (5) Find the maximum likelihood estimate of μ.

Solution: the log-likelihood function is

$$\ell = \frac{n}{2}\log 2\pi - \frac{3}{2}\sum_{k=1}^{n}\log x_k - \sum_{k=1}^{n}\frac{(1-\mu x_k)^2}{2x_k}\,.$$

Taking derivatives gives the equation

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

which yields

$$\hat{\mu} = \frac{n}{x_1 + x_2 + \cdots + x_n} = \frac{1}{\bar{x}}\,.$$

b. (5) Can you fix the maximum likelihood estimator to be unbiased? Assume as known:

- The density of $X_1 + \cdots + X_n$ is

$$f_n(x) = \frac{n}{\sqrt{2\pi x^3}}\, e^{-\frac{(n-\mu x)^2}{2x}}$$

  for $x > 0$.

- Assume as known that for $a, b > 0$ we have

$$\int_0^\infty x^{-5/2} e^{-ax-\frac{b}{x}}\,\mathrm{d}x = \frac{\sqrt{\pi}\left(1+2\sqrt{ab}\right)}{2b^{3/2}}\, e^{-2\sqrt{ab}}\,.$$

Solution: let X have density $f_n(x)$. We compute

$$\begin{aligned} E\left(\frac{n}{X}\right) &= n\int_0^\infty \frac{1}{x}\, f_n(x)\,\mathrm{d}x \\ &= n^2\frac{e^{n\mu}}{\sqrt{2\pi}}\int_0^\infty x^{-5/2} e^{-\frac{\mu^2}{2}x-\frac{n^2}{2x}}\,\mathrm{d}x \\ &= n^2\frac{e^{n\mu}}{\sqrt{2\pi}}\sqrt{2\pi}\,\frac{1+n\mu}{n^3}\, e^{-n\mu} \\ &= \mu + \frac{1}{n}\,. \end{aligned}$$

The unbiased estimate is

$$\tilde{\mu} = \frac{1}{\bar{X}} - \frac{1}{n}\,.$$

c. (10) Compute the variance of the maximum likelihood estimator of μ. Assume as known that for $a, b > 0$ we have

$$\int_0^\infty x^{-7/2} e^{-ax-\frac{b}{x}}\,\mathrm{d}x = \frac{\sqrt{\pi}\left(3 + 6\sqrt{ab} + 4ab\right)}{4b^{5/2}}\, e^{-2\sqrt{ab}}\,.$$

Solution: for X with density $f_n(x)$ we compute

$$\begin{aligned} E\left(\frac{n^2}{X^2}\right) &= \int_0^\infty \frac{n^2}{x^2}\, f_n(x)\,\mathrm{d}x \\ &= n^3 \frac{e^{n\mu}}{\sqrt{2\pi}} \int_0^\infty x^{-7/2} e^{-\frac{\mu^2}{2}x - \frac{n^2}{2x}}\,\mathrm{d}x \\ &= n^3 \frac{e^{n\mu}}{\sqrt{2\pi}} \frac{\sqrt{2\pi}(3 + 3n\mu + n^2\mu^2)}{n^5}\, e^{-n\mu} \\ &= \frac{3}{n^2} + \frac{3\mu}{n} + \mu^2\,. \end{aligned}$$

The variance is

$$\mathrm{var}(\hat{\mu}) = E(\hat{\mu}^2) - \left(E(\hat{\mu})\right)^2 = \frac{\mu}{n} + \frac{2}{n^2}\,.$$

d. (5) What approximation the the standard error of the maximum likelihood estimator do we get if we use the Fisher information? Assume as known that

$$\int_0^\infty x^{-1/2} e^{-ax-\frac{b}{x}}\,\mathrm{d}x = \frac{\sqrt{\pi}}{\sqrt{a}}\, e^{-2\sqrt{ab}}\,.$$

Solution: taking second derivatives for $n = 1$ we get

$$\ell'' = -x\,.$$

Sledi, da je

$$\begin{aligned} I(\mu) &= E(X) \\ &= \frac{e^{\mu}}{\sqrt{2\pi}} \int_0^\infty \frac{1}{\sqrt{x}}\, e^{-\frac{\mu^2 x}{2} - \frac{1}{2x}}\,\mathrm{d}x \\ &= \frac{e^{\mu}}{\sqrt{2\pi}} \cdot \sqrt{2\pi\mu e^{\mu}} \\ &= \frac{1}{\mu}\,. \end{aligned}$$

Approximate standard errors using the Fisher information matrix are

$$\frac{\mu}{n},$$

which is the leading term in the expression for the exact variance.

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

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

for $-\infty < x < \infty$ and $y > 0$ and $(\mu, \lambda) \in \mathbb{R} \times (0, \infty)$. 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 $f(x, y)$. We would like to test

$$H_0\colon \mu = 0 \qquad \text{versus} \qquad H_1\colon \mu \neq 0 .$$

a. (15) Compute the maximum likelihood estimates of the parameters. Compute the maximum likelihood estimate of λ when $\mu = 0$.

Solution: the log-likelihood function is

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

Equating the partial derivatives with 0 *we get*

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

and

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

The second equation yields

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

Substituting into the second equation gives

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

When $\mu = 0$*, the MLE is determined by the first equation. Substituting* $\mu = 0$ *in the equation, we get*

$$\tilde{\lambda} = \frac{n}{\sum_{k=1}^{n} x_k^2 y_k} .$$

b. (10) Find the likelihood ratio statistics for the above testing problem. What is its approximate distribution under H_0?

Solution: the test statistic is

$$\begin{aligned} \lambda &= 2\Big[\ell(\hat{\lambda}, \hat{\mu} | \mathbf{x}, \mathbf{y}) - \ell(\tilde{\lambda}, 0 | \mathbf{x}, \mathbf{y})\Big] \\ &= n(\log\hat{\lambda} - \log\tilde{\lambda}) - \hat{\lambda}\sum_{k=1}^{n} y_k (x_k - \hat{\mu})^2 + \tilde{\lambda}\sum_{k=1}^{n} x_k^2 y_k . \end{aligned}$$

From equations for the test statistics we get that

$$\hat{\lambda}\sum_{k=1}^{n} y_k(x_k-\hat{\mu})^2=\tilde{\lambda}\sum_{k=1}^{n} x_k^2 y_k = n\,,$$

so

$$\lambda = n\log\frac{\hat{\lambda}}{\tilde{\lambda}}\,.$$

By Wilks's theorem the approximate distribution of the test statistic under H_0 is $\chi^2(1)$.

4. (25) Assume the regression equations are

$$\begin{aligned} Y_{k1} &= \alpha + \beta x_{k1} + \epsilon_{k1} \\ Y_{k2} &= \alpha + \beta x_{k2} + \epsilon_{k2} \end{aligned}$$

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

$$E(\epsilon_{k1}) = E(\epsilon_{k2}) = 0$$

$$\text{var}(\epsilon_{k1}) = \text{var}(\epsilon_{k2}) = 2\sigma^2$$

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

$$\text{cov}(\epsilon_{k1}, \epsilon_{k2}) = \sigma^2$$

for $k \neq l$. Assume that $\sum_{k=1}^{n}(x_{k1} + x_{k2}) = 0$. The vectors $(\epsilon_{k1}, \epsilon_{k2}), \ldots, (\epsilon_{n1}, \epsilon_{n2})$ are independent.

a. (5) Show that

$$\text{cov}((3+\sqrt{3})Y_{k1} + (-3+\sqrt{3})Y_{k2}, (-3+\sqrt{3})Y_{k1} + (3+\sqrt{3})Y_{k2}) = 0$$

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

Solution: compute

$$\begin{aligned} &\text{cov}((3+\sqrt{3})Y_{k1} + (-3+\sqrt{3})Y_{k2}, (-3+\sqrt{3})Y_{k1} + (3+\sqrt{3})Y_{k2}) \\ &= \sigma^2\left(-12 - 12 + (3+\sqrt{3})^2 + (-3+\sqrt{3})^2\right) \\ &= 0\,. \end{aligned}$$

b. (5) Compute

$$\text{var}\left((3+\sqrt{3})Y_{k1} + (-3+\sqrt{3})Y_{k2}\right)$$

and

$$\text{var}\left((-3+\sqrt{3})Y_{k1} + (3+\sqrt{3})Y_{k2}\right)\,.$$

Solution: both variances are the same by symmetry. For the first we compute

$$\begin{aligned} &\text{var}\left((-3+\sqrt{3})Y_{k1} + (3+\sqrt{3})Y_{k2}\right) \\ &= (-3+\sqrt{3})^2\text{var}(Y_{k1}) + (3+\sqrt{3})^2\text{var}(Y_{k1}) \\ &\qquad +2(-3+\sqrt{3})(3+\sqrt{3})\text{cov}(Y_{k1}, Y_{k2}) \\ &= \sigma^2(48 - 12) \\ &= 36\sigma^2\,. \end{aligned}$$

c. (10) Compute the best unbiased linear estimator $\hat{\alpha}$ of α as explicitly as possible.

Solution: we replace the pair (y_{k1}, y_{k2}) *by the pair*

$$(\tilde{y}_{k1}, \tilde{y}_{k2}) = \big((3+\sqrt{3})y_{k1} + (-3+\sqrt{3})y_{k2}, (-3+\sqrt{3})y_{k1} + (3+\sqrt{3})y_{k2}\big)$$

and the pair (x_{k1}, x_{k2}) *by*

$$(\tilde{x}_{k1}, \tilde{x}_{k2}) = \big((3+\sqrt{3})x_{k1} + (-3+\sqrt{3})x_{k2}, (-3+\sqrt{3})x_{k1} + (3+\sqrt{3})x_{k2}\big)\,.$$

The regression model is transformed into

$$\tilde{\mathbf{Y}} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \tilde{\boldsymbol{\epsilon}}$$

where

$$\tilde{\mathbf{X}} = \begin{pmatrix} 2\sqrt{3} & \tilde{x}_{11} \\ 2\sqrt{3} & \tilde{x}_{12} \\ \vdots & \vdots \\ 2\sqrt{3} & \tilde{x}_{n1} \\ 2\sqrt{3} & \tilde{x}_{n2} \end{pmatrix}$$

The transformed model satisfies the assumptions of the Gauss-Markov theorem so the best unbiased estimator is

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = \left(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}^T\tilde{\mathbf{Y}}\,.$$

The assumptions imply that

$$\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} = \begin{pmatrix} 4\sqrt{3}n & 0 \\ 0 & \sum_{k=1}^n(\tilde{x}_{k1}^2 + \tilde{x}_{k2}^2) \end{pmatrix}\,.$$

Further we get

$$\tilde{\mathbf{X}}^T\tilde{\mathbf{Y}} = \begin{pmatrix} 2\sqrt{3}\sum_{k=1}^n(\tilde{y}_{k1} + \tilde{y}_{k2}) \\ \sum_{k=1}^n(\tilde{x}_{k1}\tilde{y}_{k1}^2 + \tilde{x}_{k2}\tilde{y}_{k2}^2) \end{pmatrix}\,.$$

It follows that

$$\hat{\alpha} = \frac{1}{2n}\sum_{k=1}^n(\tilde{y}_{k1} + \tilde{y}_{k2}) = \bar{y}\,.$$

d. (5) Compute the standard error of $\hat{\alpha}$.

Solution: we have

$$\begin{aligned} \mathrm{var}(\hat{\alpha}) &= \frac{n}{4n^2}(2\sigma^2 + 2\sigma^2 + 2\sigma^2) \\ &= \frac{3\sigma^2}{2n}\,. \end{aligned}$$