University of Ljubljana

Doctoral Programme in Statistics

Methodology of Statistical Research

Written examination

January 30$^{\text{th}}$, 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. (20) From a population of size N we select a simple random sample of size n. We would like to estimate the proportion θ od individuals with a certain property. It is not possible to determine whether an individual has the property directly. Each individual selected will answer two questions with possible responses (YES,YES),(YES,NO),(NO,YES) and (NO,NO). If an individual has the property she will give the response (YES,YES) with probability p_1 and a mixed response with probability $1 - p_1$. If an individual does not have the property she will give the response (NO,NO) with probability p_3 and a mixed response with probability $1 - p_3$. We assume that the probabilities p_1 and p_3 are known.

Let N_1 be the number of individuals who will respond (YES,YES) and N_3 the number of individuals who will respnd (NO,NO).

For mathematical purposes we can assume that units of the population are labelled in such a way that the first $M = N\theta$ have the property and the subsequent ones do not. Let I_k be the indicator of the event that the k-th unit is selected and $I_{k,1}$ the indicator that the k-th unit will respond (YES,YES). Let $I_{k,3}$ be the indicator of the event that the k-th unit selected will repond (NO,NO). Assume that all the indicators $I_{k,1}$ and $i_{k,3}$ are independent and independent of $(I_1, I_2, \ldots, I_N)$. We can write

$$N_1 = \sum_{k=1}^{M} I_k I_{k,1} \qquad \text{in} \qquad N_3 = \sum_{k=M+1}^{N} I_k I_{k,3}\,.$$

a. (5) Compute $E(N_1)$ and $E(N_3)$.

Solution: We know that $E(I_k) = \frac{n}{N}$, and by assumption $E(I_{k,1}) = p_1$ and $E(I_{k,3}) = p_3$. By independence and linearity we have

$$E(N_1) = \frac{Mnp_1}{N} = n\theta p_1 \qquad \textit{and} \qquad E(N_3) = \frac{(N-M)np_3}{N} = n(1-\theta)p_3\,.$$

b. (10) Compute $\mathrm{var}(N_1)$, $\mathrm{var}(N_3)$ and $\mathrm{cov}(N_1, N_3)$.

Solution: If $I \sim \mathrm{Bernoulli}(p)$ then $\mathrm{var}(I) = p(1-p)$. By independence assumptions we get for $k, l \le m$

$$\begin{aligned}
\mathrm{cov}\left(I_k I_{k,1}, I_l I_{l,1}\right) &= E\left(I_k I_{k,1} I_l I_{l,1}\right) - E\left(I_k I_{k,1}\right) E\left(I_l I_{l,1}\right) \\
&= E(I_k I_l) E(I_{k,1}) E(I_{l,1}) - E(I_k) E(I_{k,1}) E(I_l) E(I_{l,1}) \\
&= p_1^2 \mathrm{cov}(I_k, I_l) \\
&= -\frac{np_1^2(N-n)}{N^2(N-1)}
\end{aligned}$$

It follows

$$\mathrm{var}(N_1) = M\frac{np_1}{N}\left(1 - \frac{np_1}{N}\right) - M(M-1)\cdot\frac{np_1^2(N-n)}{N^2(N-1)}$$

and similarly

$$\mathrm{var}(N_3) = (N-M)\frac{np_3}{N}\left(1-\frac{np_3}{N}\right) - (N-M)(N-M-1)\cdot\frac{np_3^2(N-n)}{N^2(N-1)}\,.$$

The same way we compute

$$\mathrm{cov}\,(N_1, N_3) = -M(N-M)\frac{np_1p_3(N-n)}{N^2(N-1)}$$

c. (5) Suggest an unbiased estimate of θ.

Solution: There are several possibilities. Two of them are

$$\hat{\theta}_1 = \frac{N_1}{np_1}$$

or

$$\hat{\theta}_3 = 1 - \frac{N_3}{np_3}\,.$$

By the first part both estimators are unbiased and so are their linear combinations

$$t\,\hat{\theta}_1 + (1-t)\hat{\theta}_3\,.$$

d. (5) Compute the standard error of your estimate.

Solution: The standard errors can be computed form variances of N_1, N_2 and their covariances.

2. (25) Suppose that the observed values $x_1, x_2, \ldots, x_n$ are an i.i.d. sample from the distribution with density

$$f(x,\theta) = \begin{cases} \frac{1}{\theta} r x^{r-1} e^{-\frac{x^r}{\theta}} & \text{for } x > 0 \\ 0 & \text{else.} \end{cases}$$

We assume that $\theta > 0$ and that r is a known positive constant.

a. (5) Find the maximum likelihood estimator for θ.

Solution: The log-likelihood function has the form

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

Taking derivatives we get the equation

$$-\frac{n}{\theta} + \frac{1}{\theta^2}\sum_{k=1}^{n} x_k^r = 0\,.$$

Solving for θ gives the MLE as

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

b. (5) Determine the distribution of X_1^r. Is the MLE estimator unbiased?

Rešitve: Let X_1 have density $f(x,\theta)$. A simple change of variables gives that

$$P(X_1 \leq x) = 1 - e^{-\frac{x^r}{\theta}}\,.$$

It follows that

$$P\left(X_1^r \leq y\right) = P\left(X_1 \leq y^{\frac{1}{r}}\right) = 1 - e^{-\frac{y}{\theta}}\,.$$

It follows that $X_1^r \sim \exp(1/\theta)$. This implies that $E(X_1^r) = \theta$, and by linearity

$$E\left(\hat{\theta}\right) = \theta\,.$$

c. (10) Find the exact standard error of the estimator.

Solution: The $X_1^r, \ldots, X_n^r$ are independent exponential random variables. This implies that the sum $\sum_{k=1}^{n} X_k^r \sim \Gamma(n, 1/\theta)$. For a $\Gamma(a,\lambda)$ random variables the variance equals $a\lambda^{-2}$. In our case this means that

$$\mathrm{var}(\hat{\theta}) = \frac{\theta^2}{n}$$

and consequently

$$\mathrm{se}(\hat{\theta}) = \frac{\theta}{\sqrt{n}}\,.$$

d. (5) Find the approximate standard error using Fisher information.

Solution: Taking the second derivative of the log-likelihood function for $n = 1$ *gives*

$$\ell'' = \frac{1}{\theta^2} - \frac{2X_1^r}{\theta^3}\,.$$

Taking expectations we get

$$I(\theta) = \theta^2\,.$$

It follows that

$$\mathrm{se}(\hat{\theta}) = \frac{\theta}{\sqrt{n}}\,.$$

3. (25) Bartlett's test is a commonly used test for equal variances. The testing problem assumes that all observations $\{x_{ij}\}$ for $i = 1, 2, \ldots, k$ and $j = 1, 2, \ldots, n_i$ for each i are like independent random variables where $X_{ij} \sim N(\mu_i, \sigma_i^2)$. One tests

$$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$$

against

$$H_1 : \text{the } \sigma_i^2 \text{ are not all equal.}$$

Assume we have samples of size n_i from the i-th population, $i = 1, 2, \ldots, k$, and the usual variance estimates from each sample

$$s_1^2, s_2^2, \ldots, s_k^2$$

where

$$s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$$

with $\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$ for $i = 1, 2, \ldots, k$. Introduce the following notation $\nu_i = n_i - 1$ and

$$\nu = \sum_{i=1}^{k} \nu_i$$

and

$$s^2 = \frac{1}{\nu} \sum_{i=1}^{k} \nu_i s_i^2$$

The Bartlett's test statistic M is defined by

$$M = \nu \log s^2 - \sum_{i=1}^{k} \nu_i \log s_i^2 \,.$$

a. (15) Assume that the maximum likelihood estimates for parameters μ_i and σ_i^2 are

$$\hat{\mu}_i = \bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij} \quad \text{and} \quad \hat{\sigma}_i^2 = \frac{1}{n_i} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$$

for $i = 1, 2, \ldots, k$. Write down the likelihood ratio statistic for the testing problem in question. What is its approximate distribution?

Hint: If you assume $\sigma_1^2 = \sigma^2 = \cdots = \sigma_k^2$, the MLE estimates for μ_i are still the means $\bar{x}_i$ for $i = 1, 2, \ldots, k$.

Solution: The log-likelihood function is

$$\ell = \sum_{i=1}^{k} \left(\frac{n_i}{2} \log 2\pi - n_i \log \sigma_i - \frac{1}{2\sigma_i^2} \sum_{j=1}^{n_i} (x_{ij} - \mu_i)^2 \right) \,.$$

If there are no restrictions the maximum is attained for $\hat{\mu}_i = \bar{x}_i$ *and* $\hat{\sigma}_i^2 = \frac{1}{n_i}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2$. *The maximum value of the log-likelihood function is*

$$\ell_1 = \sum_{i=1}^{k}\left(\frac{n_i}{2}\log 2\pi - n_i \log \hat{\sigma}_i - \sum_{i=1}^{k}\frac{n_i}{2}\right).$$

If all σ_i^2 *are assumed to be equal to* σ^2 *the log-likelihood function simplifies to*

$$\ell = \frac{n}{2}\log 2\pi - n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \mu_i)^2 .$$

where $n = n_1 + \cdots + n_k$. *The maximum will be attained when* $\hat{\mu}_i = \bar{x}_i$ *as in the unrestricted case. Taking the derivative over* σ *gives the equation*

$$-\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2 .$$

Solving we get

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2$$

Substituting into the log-likelihood function we get that the restricted maximum is

$$\ell_2 = \frac{n}{2}\log 2\pi - n\log\hat{\sigma} - \frac{n}{2} .$$

The likelihood ratio statistic is

$$\lambda = 2(\ell_1 - \ell_2)$$

or explicitly

$$\lambda = n\log\hat{\sigma}^2 - \sum_{i=1}^{k} n_i \log\hat{\sigma}_i^2 .$$

The approximate distribution of the λ *statistics under the null-hypothesis is* $\chi^2(r)$ *where* $r = 2k - (k+1) = k - 1$.

b. (10) The approximate distribution of Bartlett's M under the null-hypothesis is $\chi^2(r)$. What is in your opinion r? Explain why.

Solution: The Bartlett's test is almost equal to the likelihood-ratio test. Therefore the same approximate distribution will hold for the Bartlett's test under the null-hypothesis.

4. (25) Assume the regression model

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

for $k = 1, 2, \ldots, n$ where $\epsilon_1, \ldots, \epsilon_n$ are uncorrelated, $E(\epsilon_k) = 0$ and $\mathrm{var}(\epsilon_k) = \sigma^2$ for $k = 1, 2, \ldots, n$. Assume that $x_k > 0$ for all $k = 1, 2, \ldots, n$. Consider the following linear estimators of β:

$$\begin{aligned}
\hat\beta_1 &= \frac{\sum_{k=1}^n x_k Y_k}{\sum_{k=1}^n x_k^2} \\
\hat\beta_2 &= \frac{1}{n}\sum_{k=1}^n \frac{Y_k}{x_k} \\
\hat\beta_3 &= \frac{\sum_{k=1}^n Y_k}{\sum_{k=1}^n x_k}
\end{aligned}$$

a. (5) Are all estimators unbiased?

Solution: The assumptions imply that $E(Y_k) = \beta x_k$ for all $k = 1, 2, \ldots, n$. Using this we see that all estimates are unbiased.

b. (10) Which of the estimators has the smallest standard error? Justify your answer.

Solution: By Gauss-Markov the best unbiased linear estimator of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$. In the model above $\mathbf{X}$ is just a column vector. The best unbiased estimator is $\hat\beta_1$.

c. (5) Write down the standard errors for all three estimators.

Solution: The computation of variances is, given that $Y_1, \ldots, Y_n$ are by assumption uncorrelated,

$$\begin{aligned}
\mathrm{var}(\hat\beta_1) &= \frac{\sigma^2}{\sum_{k=1}^n x_k^2} \\
\mathrm{var}(\hat\beta_2) &= \frac{\sigma^2 \sum_{k=1}^n x_k^{-2}}{n^2} \\
\mathrm{var}(\hat\beta_3) &= \frac{n\sigma^2}{\left(\sum_{k=1}^n x_k\right)^2}.
\end{aligned}$$

d. (5) How would you estimate the variances of the three estimators? Are your estimators unbiased?

Solutione: We need un unbiased estimator of σ^2. We know that

$$\hat\sigma^2 = \frac{1}{n-1}\sum_{k=1}^n (Y_k - \hat\beta_1 x_k)^2$$

is such an unbiased estimator. Using this the above formulae for variance gives unbiased estimators of the variances of the three estimators.