PROBABILITY & STATISTICS

Homework assignment 3 - due February 15th, 2024

Instructions

Please turn in the homework with this cover page. You do not need to edit the solutions. Just make sure the handwriting is legible. You may discuss the problems with your peers but the final solutions should be your work.

Statement: With my signature I confirm that the solutions are the product of my own work. Name: _____ Signature: _____.

1. Do the following problems from Rice's book:

   Chapter 7: 12, 24, 28, 54, 67.

   Chapter 8: 3, 16, 27, 45.

   CHOOSE TWO SAMPLING AND TWO ESTIMATION PROBLEMS BELOW

2. Suppose a population of size N is divided into $K = N/M$ groups of size M. We select a sample of size km the following way:

   - First we select k groups out of K groups by simple random sampling *with* replacement.
   - We then select m units in each group selected on the first step by simple random sample *with* replacement.
   - The estimate of the population mean is the average $\bar{Y}$ of the sample.

   Let μ_i be the population average in the i-the group for $i = 1, 2, \ldots, K$. Let

   $$\sigma_u^2 = \frac{1}{K} \sum_{i=1}^{K} (\mu_i - \mu)^2 ,$$

   where $\mu = \sum_{i=1}^{K} \mu_i/K$. Let

   $$\sigma_w^2 = \frac{1}{N} \sum_{i=1}^{K} \sum_{j=1}^{M} (y_{ij} - \mu_i)^2 ,$$

   where y_{ij} denotes the value of the variable for the j-the unit in the i-th group.

   a. Let $k = 1$. Show that we can write the estimator as

   $$\bar{Y} = \sum_{i=1}^{K} I_i Y_i ,$$

   where

   $$I_i = \begin{cases} 1 & \text{if the } i\text{-th group is selected.} \\ 0 & \text{otherwise} \end{cases}$$

   and $\text{var}(Y_i) = \sigma_i^2/m$. Argue that it is reasonable to assume that Y_i and I_i are all independent. Let σ_i^2 be the population variance for the i-th subgroup. Compute $\text{var}(\bar{Y})$.

b. If we repeat the procedure we get independent estimators $\bar{Y}_1, \bar{Y}_2, \ldots, \bar{Y}_k$, and estimate the population average by

$$\bar{Y} = \frac{1}{k} \sum_{i=1}^{k} \bar{Y}_k \,.$$

Show that

$$\mathrm{var}(\bar{Y}) = \frac{\sigma_u^2}{k} + \frac{\sigma_w^2}{km} \,.$$

Argue that this expression is the variance of the estimator described in the introduction.

c. The assumption that we sample with replacement is unrealistic. Let $k = 1$ and assume that the sample of size m is selected by simple random sample *without* replacement. Argue that

$$\bar{Y} = \sum_{i=1}^{K} I_i Y_i \,,$$

where

$$I_i = \begin{cases} 1 & \text{if we select the } i\text{th subgroup.} \\ 0 & \text{otherwise} \end{cases}$$

Compute the variance of the estimator in this case.

d. Assume that the k groups are selected by simple random sample *without* replacement. In this case the estimator is

$$\bar{Y} = \frac{1}{k} \sum_{i=1}^{K} I_i Y_i \,,$$

where

$$I_i = \begin{cases} 1 & \text{if we select the } i\text{th subgroup.} \\ 0 & \text{otherwise} \end{cases}$$

Argue that it is reasonable to assume that $I_1, \ldots, I_K$ and $Y_1, \ldots, Y_K$ are independent. Compute the standard error of the estimator.

e. Explain why the sampling distribution in d. is approximately normal.

3. In a population od size N there are three types of units: A, B and C. We would like to estimate the proportions a, b and c of these units. When a unit is chosen it does not necessarily respond truthfully but chooses one of the three types at random. If a unit is of type $X \in \{A, B, C\}$ it will respond that it is of type $Y \in \{A, B, C\}$ with probability p_{XY}. Assume that the probabilities p_{XY} are known.

We choose a simple random sample of size n. Assume that the units choose their responses independently of each other and independently of the sampling procedure.

a. Let N_X be the number of units in the sample of type X and M_X the number of units in the sample who respond X for $X \in \{A, B, C\}$. Compute

$$E(M_X | N_A, N_B, N_C)\,.$$

b. Suggest unbiased estimates for a, b and c. When is it possible to estimate the proportions?

c. Compute $\operatorname{cov}(M_X, M_Y | N_A, N_B, N_C)$ for $X, Y \in \{A, B, C\}$.

d. Give standard errors for the unbiased estimates of a, b and c.

4. Assume that in a population of size N there are two statistical variables belonging to every unit. Denote these pairs by $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$. The problem is to estimate the average

$$\lambda = \frac{1}{2N} \sum_{k=1}^{N} (x_k + y_k)\,.$$

Assume that a simple random sample od size n is selected. A selected unit will repond with x_k with probability $\frac{1}{2}$ and y_k with probability $\frac{1}{2}$ independenly of the sampling procedure and independently of other units.

a. Find un unbiased estimator $\hat{\lambda}$ of λ.

b. Let I_k be the indicator of the event that the k-th unit is selected into the sample. Compute

$$E(\hat{\lambda} | I_1, I_2, \ldots, I_N) \qquad \text{and} \qquad \operatorname{var}(\hat{\lambda} | I_1, \ldots, I_N)\,.$$

c. Let $z_k = (x_k + y_k)/2$ and $w_k = (x_k - y_k)^2/4$. Express the standard error of $\hat{\lambda}$ with the population variance σ_z^2 of $z_1, z_2, \ldots, z_N$ and the average θ of the values $w_1, w_2, \ldots, w_N$.

d. Can the standard error of $\hat{\lambda}$ be estimated in practice? Explain.

5. Often it is difficult to obtain honest answers from sample subjects to questions like "Have you ever used heroin" or "Have you ever cheated on an exam". To reduce bias the method or *randomized response* is used. The sample subject is given one of the two statements below at random:

   (1) "I have property A.'
   (2) "I do not have property A.'

   The subject responds YES or NO to the given question. The pollster does not know to which of the two statements the subject is responding. We assume:

   - The subjects are a simple random sample of size n from a larger population of size N.
   - The statements are assigned to the chosen subjects independently.
   - The assignment of statements is independent of the sampling procedure.
   - The subjects respond honestly to the statements they are given.

   Let

   - p be the probability the a subject will be assigned the statement (1). This probability is known and is part of the design.
   - q be the proportion of subjects in the population with property A.
   - r be the probability that a randomly selected subject responds YES to the statement assigned.
   - R be the proportion of subjects in the sample who respond YES.

   a. Justify that the probability that a randomly selected subject in the population responds YES to the statement assigned is equal for all subjects. Express this probability with p and q. Show that R is an unbiased estimate of r. Take into account that the assignment of statements is independent of the selection procedure.

   b. Suggest an unbiased estimator of q. When is this possible? Express the variance of the estimator with $\mathrm{var}(R)$.

c. Let N_A be the random number of sample subjects with property A, and let N_Y be the random number of sample subjects who respond YES. Compute $E(N_Y|N_A)$ and $\text{var}(N_Y|N_A)$.

d. Compute $\text{var}(R)$. Give the standard error for the unbiased estimate of q.

6. Suppose $\{p(\mathbf{x},\theta), \theta \in \mathbf{\Theta} \subset \mathbb{R}^k\}$ is a (regular) family of distributions. Define the vector valued *score function* $\mathbf{s}$ as the column vector with components

$$\mathbf{s}(\mathbf{x},\theta) = \frac{\partial}{\partial\theta}\,\log(p(\mathbf{x},\theta)) = \text{grad}(\log(p(\mathbf{x},\theta))\,.$$

and the Fisher information matrix as

$$\mathbf{I}(\theta) = \text{var}(\mathbf{s})\,.$$

Remark: If $p(\mathbf{x},\theta) = 0$ define $\log\left(p(\mathbf{X},\theta)\right) = 0$.

a. Let $\mathbf{t}(\mathbf{X})$ be an unbiased estimator of θ based on the likelihood function, i.e.

$$E_\theta(\mathbf{t}(\mathbf{X})) = \theta\,.$$

Prove that

$$E(\mathbf{s}) = \mathbf{0} \qquad \text{and} \qquad E(\mathbf{s}\mathbf{t}^T) = \mathbf{I}\,.$$

Deduce that $\text{cov}(\mathbf{s},\mathbf{t}) = \mathbf{I}$.

Remark: Make liberal assumptions about interchanging integration and differentiation.

b. Let $\mathbf{a},\mathbf{c}$ be two arbitrary k–dimensional vectors. Prove that

$$\text{corr}^2\left(\mathbf{a}^T\mathbf{t}, \mathbf{c}^T\mathbf{s}\right) = \frac{(\mathbf{a}^T\mathbf{c})^2}{\mathbf{a}^T\text{var}(\mathbf{t})\mathbf{a}\cdot \text{c}^T\mathbf{I}(\theta)\mathbf{c}}\,.$$

The correlation coefficient squared is always less or equal 1. Maximize the expression for the correlation coefficient over $\mathbf{c}$ and deduce the Rao-Cramér inequality.

7. Assume the data pairs $(y_1,z_1),\ldots,(y_n,z_n)$ are an i.i.d. sample from the distribution with density

$$f(y,z,\theta,\sigma) = e^{-y}\cdot\frac{1}{\sqrt{2\pi y}\sigma}e^{-\frac{(z-\theta y)^2}{2y\sigma^2}}$$

for $y > 0$ and $\sigma > 0$.

a. Find the maximum likelihood estimators of θ and σ^2. Are the estimators unbiased?

b. Find the exact standard errors of $\hat{\theta}$ and $\hat{\sigma}^2$.

c. Compute the Fisher information matrix.

d. Find the standard errors of the maximum likelihood estimators using the Fisher information matrix. Comment on your findings.

8. Assume the data pairs $(x_1, y_1), \ldots, (x_n, y_n)$ are an i.i.d. sample from the bivariate normal distribution with parameters

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}.$$

a. Find the maximum likelihood estimators of the parameters. Fix the estimators if necessary so that they will be unbiased and compute their variances.

b. Suppose the parameters μ_1, σ_{11} and σ_{12} are known. Can you use this information to improve the estimator of μ_2. Compute the variance of the improved estimator.

c. Repeat the argument if only μ_1 and σ_{11} are known. What would you do? Can you compute the variance of the new estimator?

d. Suppose the parameters μ_1, μ_2, σ_{11} and σ_{12} are known. Can you give an improved estimate of σ_{22}? Prove that it is better than the maximum likelihood estimator.