Statistical Learning and Stochastic Processes
Assignment 2 (Part A: Statistical Inference)

Instructions

This assignment is to be completed in teams of two students. Every team should register itself in Brightspace. To offer you an optimal learning experience it is not allowed to use generative AI to answer these questions directly. The assignment was designed to be within the scope of the abilities of students that have the required background knowledge for this course and studied the course material. Use of gen AI (or Mathematica for that matter) to help with simplifying algebraic expressions or computing derivatives is allowed though. In any case, your report should include a separate section stating for what purposes (if any) and to what extent you used gen AI to write the report. You may be invited to answer questions in person about your report afterwards.

When asked for a derivation or proof, clearly show and justify the steps in your reasoning.

This is part A (statistical inference) of the second assignment. Part B (data analysis) will be released on January 5, 2026. Both parts should be handed in together in a single report in pdf format, ultimately on Friday, January 23, 2026.

Question 1: Randomized Response

In some surveys, questions involve topics like illegal behavior (drug use, tax evasion), stigmatized actions or beliefs, or socially undesirable traits.

If asked directly, respondents may lie to protect themselves, refuse to answer, or drop out of the survey altogether. This creates systematic bias, not just random noise.

Randomized response introduces a known random mechanism (e.g., a coin flip, dice roll, or random number generator) that the respondent observes privately and the researcher never observes. Depending on the random outcome, the respondent either answers the sensitive question truthfully, or gives a prescribed answer (or answers a different question). Because the researcher knows the probabilities of the randomization, they can recover unbiased population estimates without knowing any individual's true status.

We formalize the problem as follows. Suppose that every person in a population belongs to either group A (the sensitive group) or group B and we want to estimate the proportion π of people belonging to group A. To estimate π an i.i.d. sample of size n is drawn from the population. We do *not* observe whether a person belongs to group A or group B. Instead, each person in the sample is asked to spin a spinner with a face marked so that the spinner points to the letter A with known probability a, and to the letter B with probability $(1 - a)$. The person is to say "yes" or "no" according to whether the spinner points to the correct group; he or she does not report the group to which the spinner points.

We introduce the following notation:

$$\pi : \text{the proportion of people belonging to group } A \text{ in the population}$$

$$a : \text{the known probability that the spinner points to } A$$

$$X_i = \begin{cases} 1 & \text{if respondent } i \text{ answers "yes".} \\ 0 & \text{if respondent } i \text{ answers "no".} \end{cases}$$

$$m : \sum_{i=1}^{n} X_i \text{ (the number of times we observe the answer "yes" in the sample)}$$

(a) Express $P(X_i = 1)$ as a function of a and π.

(b) Give expressions for the joint probability of the observed data $X_1, X_2, \ldots, X_n$, and the corresponding likelihood function and loglikelihood function.

(c) Derive an expression for the maximum likelihood estimator π_{ML} of π.

(d) Can you think of an easier way to arrive at the same estimator as the maximum likelihood estimator?

(e) Show that π_{ML} is an unbiased estimator of π.

(f) Derive an expression for the variance of π_{ML}. Analyse its dependence on a.

To compare randomized response with direct questioning, suppose that persons belonging to the sensitive group will lie with unknown probability ℓ when asked the question "Do you belong to group A?". Suppose furthermore that respondents belonging to the non-sensitive group always answer truthfully. We naively estimate π as the proportion of people who answer "yes" to the question: "Do you belong to group A?".

(g) Give an expression for the bias of this naive estimator. Is it asymptotically unbiased?

(h) Derive expressions for the mean squared error (MSE) of the randomized response estimator and the naive estimator. Analyse the expressions to draw conclusions about when, depending on the relevant parameters (π,a,ℓ,n), one estimator should be favored over the other.

Question 2: Estimating the Size of Wildlife Populations

Suppose you want to estimate how many individuals of a particular animal species live in a given area. This may be relevant to determine whether a species is endangered and therefore should be protected. We assume there is a fixed but unknown number of animals in the area. This means we assume no animals are born, die, immigrate, or emigrate during the study. Some species are hard to observe, for example because they are rare, hard to detect (camouflage), or nocturnal. Therefore, we use traps to capture the animals in order to estimate the population size. We set the same traps multiple times (e.g., each night for several nights). Each trapping session is conducted in the same way. We assume different animals have different probabilities of being caught, for example because some animals are bold and active, whereas others are shy and cautious. We do not make any assumptions about the distribution of these trapping probabilities. Each captured animal can be uniquely identified (e.g. tags, markings, DNA). After capture, animals are released back into the population. After all trapping sessions are finished, the data is summarized as follows.

Let f_k denote the number of animals that were caught exactly k times in a trap. Let $n = \sum_{t=1}^{T} f_t$, where T is the number of trapping occasions, that is, n is the number of distinct animals that were caught in the study. Obviously, you do not know how many animals were never caught at all (f_0). If we knew f_0, then we could determine the population size to be $N = n + f_0$. The idea is now to estimate f_0 from the observed frequencies. Let N denote the unknown size of the population. The following estimator for N is proposed:

$$\hat{N} = n + \hat{f}_0, \quad \hat{f}_0 = \begin{cases} \frac{f_1^2}{2f_2} & \text{if} \quad f_2 > 0 \\ \\ \frac{f_1(f_1-1)}{2(f_2+1)} & \text{if} \quad f_2 = 0 \end{cases}$$

Intuitively, the proposed estimator can be understood as follows. If many animals are caught only once but very few are caught twice, that suggests there are many hard-to-catch animals, many of which were missed entirely. Conversely, if most animals are caught multiple times and few are caught only once, then you probably haven't missed many. Note that the basic estimator is not defined if $f_2 = 0$, in which case we will use the adjusted estimator.

Furthermore, the following estimator for the variance of $\hat{N}$ is proposed:

$$\hat{\sigma}^2_{\hat{N}} = \begin{cases} f_2 \left(\frac{1}{4} \left(\frac{f_1}{f_2} \right)^4 + \left(\frac{f_1}{f_2} \right)^3 + \frac{1}{2} \left(\frac{f_1}{f_2} \right)^2 \right) & \text{if} \quad f_2 > 0 \\ \\ \frac{f_1(f_1-1)}{2} + \frac{f_1(2f_1-1)^2}{4} - \frac{f_1^4}{4\hat{N}} & \text{if} \quad f_2 = 0 \end{cases}$$

Then, a $100(1-\alpha)\%$ confidence interval for N is constructed as follows,

$$\left[n + \frac{\hat{f}_0}{Q}, n + \hat{f}_0 \, Q \right] ,$$

where

$$Q = \exp\left(z_{\alpha/2} \sqrt{\ln\left(1 + \frac{\hat{\sigma}^2_{\hat{N}}}{\hat{f}_0^2}\right)} \right),$$

and $z_{\alpha/2}$ is the critical value of the standard normal distribution.

To assess the performance of the proposed point estimator and interval estimator for N, we perform a Monte Carlo simulation experiment. We consider population sizes $N = 500$, $N = 1,000$, and $N = 5000$. The number of trapping occasions is set to 40 ($T = 40$). The individual trapping probabilities are drawn from the following distributions

1. $p_i \sim U(0, 0.01) \qquad i = 1, \ldots, N.$
2. $p_i \sim 0.02 \times \text{beta}\ (1, 3) \qquad i = 1, \ldots, N.$
3. $p_i \sim 0.01 \times \text{beta}\ (2, 2) \qquad i = 1, \ldots, N.$
4. $p_i \sim U(0, 0.02) \qquad i = 1, \ldots, N.$

To estimate the bias, variance and mean squared error of the proposed point estimator, we perform 1,000 simulation runs for each scenario (combination of population size and distribution of capture probabilities). Furthermore, to verify if the proposed confidence intervals have the correct coverage, we compute the coverage for $\alpha = 0.05$ (95% confidence interval), and $\alpha = 0.01$ (99% confidence interval). Summarize the results of the simulation experiments in a Table, and discuss the results. What conclusions can you draw?

Also hand in the program code (`Python` or `R`) with comments, that you wrote to perform the simulation experiments.