HOMEWORK ASSIGNMENT #2
Electrical and Computer Engineering Faculty, University of Tehran, Iran

Version 1
1st Semester 2023-2024

Introduction to Statistical Inference

Melika Sadeghi, Amirreza Salamat,
melikasadeghi16@gmail.com, ar.salamat@yahoo.com
Instructor: Mohammad-Reza A. Dehaqani

Deadline:
27 Azar 1402

I. Introduction

This homework assignment tests understanding of key concepts in statistical inference like confidence intervals, hypothesis testing, sampling distributions, and parameter estimation. It contains 8 problems needing application of statistical theory to real-world data analysis and simulation experiments.

Problem 0 reviews the central limit theorem for determining probabilities of sums of random variables. Problem 1 derives estimators for sensitive statistics using randomized surveys. Problem 2 involves sample size calculation to estimate population proportions within desired precision.

Problem 3 checks conceptual knowledge on interpretation of intervals and distributions. Problem 4 asks to compute required sample size to achieve a target confidence interval width. Problem 5 proves the link between hypothesis testing and confidence intervals.

Problem 6 asks for unbiased estimation of the population mean using weighted averages. Problem 7 works with a dataset of 301 counties - simulating sampling distributions, making confidence intervals, comparing variance estimates under different sampling schemes.

Problem 8 uses simulation to estimate a quantity, specifically the area of a shape circumscribed by a quarter circle. The bonus estimates moments of a capped exponential distribution used in reliability analysis.

Overall, this homework develops theoretical knowledge, data analysis skills, and computational thinking. The problems attempt to connect statistical concepts to real-world techniques and datasets. The introduction summarizes the key themes and structure of the assignment.

II. Mixed Problems

Problem 0

We throw a dice 100 times. Using the Central Limit Theorem (CLT) find:

a) the probability that we get a 6, 15 to 20 times.
b) the probability that the sum of the numbers seen in the 100 throws is less than 300.

Problem 1

In surveys, it is difficult to obtain accurate answers to sensitive questions such as ”Have you ever cheated on an exam?” Warner (1965) introduced the method of randomized response to deal with such situations. A respondent spins an arrow on a wheel or draws a ball from an urn containing balls of two colors to determine which of two statements to respond to: (1) ”I have characteristic A,” or (2) ”I do not have characteristic A.” The interviewer does not know which statement is being responded to but merely records a yes or a no. The hope is that an interviewee is more likely to answer truthfully if he or she realizes that the interviewer does not know which statement is being responded to. Let R be the proportion of a sample answering Yes. Let p be the probability that statement 1 is responded to (p is known from the structure of the randomizing device), and let q be the proportion of the population that has characteristic A. Let r be the probability that a respondent answers Yes.

a. Show that $r = (2p-1)q + (1-p)$. [Hint: $P(\text{yes}) = P(\text{yes} \mid \text{question 1}) \cdot P(\text{question 1}) + P(\text{yes} \mid \text{question 2}) \cdot P(\text{question 2})$.]
b. If r were known, how could q be determined?
c. Show that $E(R) = r$, and propose an estimate, $\hat{Q}$, for q. Show that the estimate is unbiased.
d. Ignoring the finite population correction, show that

$$\text{Var}(R) = \frac{r(1-r)}{n}$$

where n is the sample size.
e. Find an expression for $\text{Var}(\hat{Q})$.

Problem 2

In a survey of a very large population, the incidences of two health problems are to be estimated from the same sample. It is expected that the first problem will affect about 3% of the population and the second about 40%. Ignore the finite population correction in answering the following questions.

a. How large should the sample be in order for the standard errors of both estimates to be less than .01? What are the actual standard errors for this sample size?
b. Suppose that instead of imposing the same limit on both standard errors, the investigator wants the standard error to be less than 10% of the true value in each case. What should the sample size be?

Problem 3

For each following statement, explain if it is true and determine if there is any problem and fix it.

a. In the CLT we are looking for a confidence interval for the sample mean.
b. With the same null hypothesis, it's always possible that we reject 2 side tests while 1 side test doesn't reject.
c. The central limit theorem states that the sampling distribution of sample means will closely resemble the normal distribution regardless of the sample size.
d. For a positively skewed distribution, the mean usually has a larger value than either the median or the mode.
e. For a given standard error, lower confidence levels produce wider confidence intervals.
f. The statement, "The 95% confidence interval for the population mean is (350,400)", is equivalent to the statement, "there is a 95% probability that the population mean is between 350 and 400".
g. A 95% confidence interval obtained from a random sample of 1000 people has a better chance of containing the population percentage than a 95% confidence interval obtained from a random sample of 500 people.
h. Suppose we have constructed the following confidence interval about the mean age of freshmen college students in a State: $18.4 \leq \mu \leq 21.5$. The proper interpretation is that we are 95% confident that the sample mean is in the range 18.4 and 21.5 years.
i. A confidence interval is an estimate for which there is a specified degree of certainty that the population parameter will fall within the range of the interval.

Problem 4

Consider a clinical study designed to compare the effectiveness of a treatment with a control. Assume that n measurements are taken for both the treatment group and the control group, with the standard deviation observed in both groups being equal to 10. What should the sample size n be in each group such that the 95% confidence interval for the difference in the mean outcomes between the two groups is less than or equal to 2 units?

Problem 5

Assuming we have a two-sample t-test with the following hypotheses:

- $H_0 : \mu_1 = \mu_2$
- $H_1 : \mu_1 \neq \mu_2$

Prove that the null hypothesis H_0 is rejected only if the confidence interval for the difference of the means does not include zero.

Problem 6

Consider a random sample of size n from a population of size N. The estimator μ is defined as:

$$\bar{X}_c = \sum_{i=1}^{n} c_i X_i,$$

where c_i are constants and $X_1, \ldots, X_n$ are the sample values.

Under what condition for the c_i's will $\bar{X}_c$ be an unbiased estimator of the population mean?

Problem 7

The attached Excel file contains values for breast cancer mortality from 1950 to 1960(y) and the adult white female population in 1960(x) for 301 counties in North Carolina, South Carolina, and Georgia.

a. Make a histogram of the population values for cancer mortality.
b. What are the population mean and total cancer mortality? What are the population variance and standard deviation?
c. Simulate the sampling distribution of the mean of a sample of 25 observations of cancer mortality.
d. Draw a simple random sample of size 25 and use it to estimate the mean and total cancer mortality.
e. Estimate the population variance and standard deviation from the sample of part (d).
f. Form 95% confidence intervals for the population mean and total from the sample of part (d). Do the intervals cover the population values?
g. Repeat parts (d) through (f) for a sample of size 100.
h. Suppose that the size of the total population of each county is known and that this information is used to improve the cancer mortality estimates by forming a ratio estimator. Do you think this will be effective? Why or why not?
i. Simulate the sampling distribution of ratio estimators of mean cancer mortality based on a simple random sample of size 25. Compare this result to that of part (c).
j. Draw a simple random sample of size 25 and estimate the population mean and total cancer mortality by calculating ratio estimates. How do these estimates compare to those formed in the usual way in part (d) from the same data?
k. Form confidence intervals about the estimates obtained in part (j).
l. Stratify the counties into four strata by population size. Randomly sample six observations from each stratum and form estimates of the population mean and total mortality.
m. Stratify the counties into four strata by population size. What are the sampling fractions for proportional allocation and optimal allocation? Compare the variances of the estimates of the population mean obtained using simple random sampling, proportional allocation, and optimal allocation.
n. How much better than those in part (m) will the estimates of the population mean be if 8, 16, 32, or 64 strata are used instead?

Problem 8

Implement the following code for $n = 10, 10000, 10000000$ in R. What does this function estimate? Using which method?

```
estimate <- function(n) {
return (4 * sum((runif(n)^2 + runif(n)^2) < 1) / n)
}
```

Bonus: Using a similar approach, design a simulation to estimate the area of a geometric shape (e.g., an ellipse, a triangle, or a complex polygon).

Bonus Problem

Let X be a random variable following an exponential distribution with density function $f(x) = \frac{1}{\alpha}e^{-x/\alpha}$, for $x \geq 0$ and scale parameter $\alpha > 0$. Define a new random variable Z by:

$$Z = \begin{cases} X & \text{if } X \leq T, \\ T & \text{otherwise,} \end{cases}$$

where T is a given positive constant.

Determine the mean and variance of Z.

Note: Random variable Z is used in contexts like reliability and survival analysis, representing times to events subject to right-censoring at time T.

III. Submission

For the programming section, each student is required to submit a well-structured, typed PDF report that presents a concise summary of their analysis. The report should include the figures mentioned in the problem description and offer a detailed discussion of each. Please avoid uploading theoritical problem in .jpg format and upload them in a single .pdf file.

For each section of the report, a separate script is expected, which can be written in MATLAB (.m), Python 3 (.py or .py3), or R (.r). Avoid submitting scripts in formats like MATLAB live scripts, Python notebooks, or R Markdown. It is crucial that the submitted code is compatible with the grader's system. Be sure to include all relevant functions and any non-standard libraries used in your code.

The report should be treated as an academic piece of writing, and it should not contain any code snippets or explanations of coding logic. Instead, it should provide the author's insights about the results and demonstrate a strong grasp of the reference article. Academic reports typically maintain a concise and highly formal tone.

Each section of the report should briefly outline the hypothesis being tested. The responsibility for designing and implementing the tests lies with the students, as does explaining the results. Interpretations should be comprehensive without unnecessary verbosity.

The report can be written in either Persian or English, with no preference for either. In Persian reports, use B Nazanin with a font size of 14 for the text body and B Titr with a font size of 18 for titles. English reports should use Times New Roman 12 for the body text and Times New Roman 16 for titles. Sentences should be written in the passive tense. In Persian reports, the correct usage of the zero-width non-joiner is mandatory. In all reports, equations, figures, and tables must be labeled with unique numbers and referenced accordingly. Referring to figures as ”the following figure,” ”the figure above,” and similar expressions is considered incorrect.

Every figure in the report should be accompanied by a descriptive caption below it, while tables should have captions above them. Feel free to use footnotes and citations as necessary for clarity and proper attribution.