Problem Set 2: Written Part

Your Name Goes Here

Details

Due Date

This assignment is due at 4:00 PM on Friday, Feb 8.

How to Write Up

The written part of this assignment can be either typeset using latex or hand written.

Grading

5% of your grade on this assignment is for turning in something legible. This means it should be organized, and any Rmd files should knit to pdf without issue.

An additional 20% of your grade is for completion. A quick pass will be made to ensure that you've made a reasonable attempt at all problems.

Across both the written part and the R part, in the range of 1 to 3 problems will be graded more carefully for correctness. In grading these problems, an emphasis will be placed on full explanations of your thought process. You don't need to write more than a few sentences for any given problem, but you should write complete sentences! Understanding and explaining the reasons behind what you are doing is at least as important as solving the problems correctly.

Solutions to all problems will be provided.

Collaboration

You are allowed to work with others on this assignment, but you must complete and submit your own write up. You should not copy large blocks of code or written text from another student.

Sources

You may refer to our text, Wikipedia, and other online sources. All sources you refer to must be cited in the space I have provided at the end of this problem set.

Problem I

True or false (and state why): If a sample from a population is large, a histogram of the values in the sample will have a shape that is approximately normal, even if the values in the population are not normally distributed.

Problem II

Suppose $X_1, \ldots, X_n$ are independent and identically distributed random variables with common mean μ and variance σ^2.

We will prove that $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$ is an unbiased estimator of σ^2. This is the reason for using the denominator of $n - 1$ when calculating the sample variance.

(1) Find $E(X_i^2)$

(2) Find $E(\bar{X}^2)$

(3) Find $E(X_i\bar{X}) = E\left(X_i \frac{1}{n}\sum_{j=1}^{n} X_j\right)$

In doing this, recall that if X_i and X_j are independent, $E(X_iX_j) = E(X_i)E(X_j)$.

(4) Using your results from (1) through (3), find $E\{(X_i - \bar{X})^2\}$

(5) Using your result from (4), show that s^2 **is an unbiased estimator of** σ^2**.**

(6) Using your result from (5) and Jensen's inequality, show that $s = \sqrt{s^2}$ **is a biased estimator of** σ**.**

Jensen's inequality is an inequality about the expected value of convex functions of a random variable. Here's one statement of the inequality.

Suppose that:

- X is a random variable which does not have $P(X = c) = 1$ for some constant c, and
- ϕ is a strictly convex function.

Then $\phi\{E(X)\} < E\{\phi(X)\}$.

In the statement of Jensen's theorem, use s in place of X (regarding the sample standard deviation as a random variable). In any practical setting the first condition of Jensen's theorem will be satisfied, so let's not worry about that. Also, use $\phi(x) = x^2$, which is strictly convex.

Problem III: Cluster Sampling

Often, it's too difficult to take a simple random sample. For reasons of cost or time, it may be easier to do cluster sampling: we divide the population into separate groups ("clusters") and then pick some of those clusters to be in our sample; everyone in a selected cluster will be included. For example, suppose we work for the New York City department of education, and we want to do a survey of students in schools in the city to get their opinions. In theory, the school system has a list of all students in the school system, so they could take a simple random sample. But that might not be practical. It might be easier to administer the test by randomly selecting some classrooms and administering the survey to all students in the selected classroom. In this example, each classroom is a cluster.

Let's look at a very simplified example of cluster sampling, where our population has 8 people in it, with responses $\{c_1, c_2, c_3, c_4, c_5, c_6, c_7, c_8\}$ to the question we will ask if they are selected. Suppose that we divide the population into four equally sized clusters, G_1 through G_4 (g for group, since c is already taken):

$$G_1 = \{c_1, c_2\} \qquad G_2 = \{c_3, c_4\} \qquad G_3 = \{c_5, c_6\} \qquad G_4 = \{c_7, c_8\}$$

For our sample, we will randomly select one of these clusters, with equal probability of selecting each of the four clusters. Both people in the selected cluster will be included in our sample, and we will compute the average response for those two people as our estimate of the population mean μ.

Introduce four random variables to represent the cluster that is selected:

$$U_1 = \begin{cases} 1 \text{ if cluster 1 is selected} \\ 0 \text{ otherwise} \end{cases} \quad U_2 = \begin{cases} 1 \text{ if cluster 2 is selected} \\ 0 \text{ otherwise} \end{cases} \quad U_3 = \begin{cases} 1 \text{ if cluster 3 is selected} \\ 0 \text{ otherwise} \end{cases} \quad U_4 = \begin{cases} 1 \text{ if cluster 4 is selected} \\ 0 \text{ otherwise} \end{cases}$$

Note that for a given sample, only one of U_1, U_2, U_3, and U_4 will be equal to 1 and the other three will be 0. From the set up, we have that $P(U_i = 1) = 0.25$ for each $i = 1, \ldots, 4$.

(1) Show that $E(U_i) = 0.25$

(2) Find an expression for the random variable $\bar{X}$ **(the sample mean) in terms of the constants** $c_1, \ldots, c_8$ **and the random variables** $U_1, \ldots, U_4$**. Argue that your expression is correct by showing that for a particular sample (i.e., a particular realization of the random variables** $U_1, \ldots, U_4$**) your expression gives the correct realized value of** $\bar{x}$ **for that sample. Just do this for one of the possible samples, no need to do it for all four possibilities.**

(3) Show that with this sampling scheme, $\bar{X}$ **is an unbiased estimator of** μ**.**