Homework 2

Aleck Zhao

September 20, 2016

Chapter 7: Survey Sampling

1. Consider a population consisting of five values - 1, 2, 2, 4, and 8. Find the population mean and variance. Calculate the sampling distribution of the mean of a sample of size 2 by generating all possible such samples. From them, find the mean and variance of the sampling distribution and compare the results to Theorems A and B in Section 7.3.1

   Solution. We have the population mean

   $$\mu = \frac{1}{5}(1+2+2+4+8) = \boxed{\frac{17}{5}}.$$

   Then the population variance is computed as

   $$\begin{aligned}\sigma^2 &= \frac{1}{5}\left[\left(1-\frac{17}{5}\right)^2 + \left(2-\frac{17}{5}\right) + \left(2-\frac{17}{5}\right)^2 + \left(4-\frac{17}{5}\right)^2 + \left(8-\frac{17}{5}\right)^2\right] \\ &= \frac{1}{5}\left[\left(-\frac{12}{5}\right)^2 + \left(-\frac{7}{5}\right)^2 + \left(-\frac{7}{5}\right)^2 + \left(\frac{3}{5}\right)^2 + \left(\frac{23}{5}\right)^2\right] \\ &= \frac{1}{125}(144+49+49+9+529) \\ &= \boxed{\frac{156}{25}}.\end{aligned}$$

   Now, we generate all samples of size 2:

   $$\begin{aligned}\{1,2\},\{1,2\},\{1,4\},\{1,8\} \\ \{2,2\},\{2,4\},\{2,8\} \\ \{2,4\},\{2,8\} \\ \{4,8\}\end{aligned}$$

   so the corresponding values for $\bar{X}$ are

   $$\begin{aligned}\frac{3}{2},\frac{3}{2},\frac{5}{2},\frac{9}{2} \\ 2,3,5 \\ 3,5 \\ 6\end{aligned}$$

   so the sampling distribution is summarized below:

Table 1: Distribution for $\bar{x}$

$\bar{x}$	3/2	2	5/2	3	9/2	5	6
$p_{\bar{X}}(\bar{x})$	1/5	1/10	1/10	1/5	1/10	1/5	1/10

Then the mean of the sampling distribution is

$$\begin{aligned} E[\bar{X}] &= \sum_{\bar{x}} \bar{x} p_{\bar{X}}(\bar{x}) \\ &= \frac{3}{2}\cdot\frac{1}{5} + 2\cdot\frac{1}{10} + \frac{5}{2}\cdot\frac{1}{10} + 3\cdot\frac{1}{5} + \frac{9}{2}\cdot\frac{1}{10} + 5\cdot\frac{1}{5} + 6\cdot\frac{1}{10} \\ &= \boxed{\frac{17}{5}} \end{aligned}$$

which agrees with Theorem 8.

The variance of the sampling distribution is then $\mathrm{Var}(\bar{X}) = E[\bar{X}^2] - (E[\bar{X}])^2$, where

$$\begin{aligned} E[\bar{X}^2] &= \sum_{\bar{x}} \bar{x}^2 p_{\bar{X}}(\bar{x}) \\ &= \left(\frac{3}{2}\right)^2\cdot\frac{1}{5} + 2^2\cdot\frac{1}{10} + \left(\frac{5}{2}\right)^2\cdot\frac{1}{10} + 3^2\cdot\frac{1}{5} + \left(\frac{9}{2}\right)^2\cdot\frac{1}{10} + 5^2\cdot\frac{1}{5} + 6^2\cdot\frac{1}{10} \\ &= \frac{139}{10} \end{aligned}$$

and

$$(E[\bar{X}])^2 = \left(\frac{17}{5}\right)^2 = \frac{289}{25}$$

so then

$$\mathrm{Var}(\bar{X}) = \frac{139}{10} - \frac{289}{10} = \boxed{\frac{117}{50}.}$$

According to Theorem 9, the sample variance is given by

$$\begin{aligned} \mathrm{Var}(\bar{X}) &= \frac{\sigma^2}{n}\cdot\frac{N-n}{N-1} \\ &= \frac{1}{2}\cdot\frac{156}{25}\cdot\frac{5-2}{5-1} \\ &= \frac{117}{50} \end{aligned}$$

which matches our result.

□

2. Suppose that a sample of size $n = 2$ is drawn from the population of the preceding problem and that the proportion of the sample values that are greater than 3 is recorded. Find the sampling distribution of this statistic by listing all possible such samples. Find the mean and variance of the sampling distribution.

Solution. As in the previous question, we generate all samples of size 2:

$$\begin{aligned} \{1,2\},\{1,2\},\{1,4\},\{1,8\} \\ \{2,2\},\{2,4\},\{2,8\} \\ \{2,4\},\{2,8\} \\ \{4,8\} \end{aligned}$$

which gives rise to the values of $\hat{p}$

$$\begin{aligned} 0,0,1/2,1/2 \\ 0,1/2,1/2 \\ 1/2,1/2 \\ 1 \end{aligned}$$

so the distribution is summarized below:

Table 2: Distribution for $\bar{x}$

$\hat{p}$	0	1/2	1
$p_{\hat{P}}(\hat{p})$	3/10	3/5	1/10

Now, the mean of the sampling distribution is

$$\begin{aligned} E[\hat{p}] &= \sum_{\hat{p}} \hat{p} \cdot p_{\hat{P}}(\hat{p}) \\ &= 0 \cdot \frac{3}{10} + \frac{1}{2} \cdot \frac{3}{5} + 1 \cdot \frac{1}{10} \\ &= \boxed{\frac{2}{5}.} \end{aligned}$$

Next, the variance of the sampling distribution is $\mathrm{Var}(\hat{p}) = E[\hat{p}^2] - (E[\hat{p}])^2$, where

$$\begin{aligned} E[\hat{p}^2] &= \sum_{\hat{p}} \hat{p}^2 p_{\hat{P}}(\hat{p}) \\ &= 0^2 \cdot \frac{3}{10} + \left(\frac{1}{2}\right)^2 \cdot \frac{3}{5} + 1^2 \cdot \frac{1}{10} \\ &= \frac{1}{4} \end{aligned}$$

and

$$(E[\hat{p}[])^2 = \left(\frac{2}{5}\right)^2 = \frac{4}{25}$$

so then

$$\mathrm{Var}(\hat{p}) = \frac{2}{5} - \frac{4}{25} = \boxed{\frac{6}{25}.}$$

□

11. Consider a population of size four, the members of which have values x_1, x_2, x_3, x_4.

    a. If simple random sampling were used, how many samples of size two are there?

    Solution. There are $\binom{4}{2} = \boxed{6}$ samples of size 2.

    □

    b. Suppose that rather than simple random sampling, the following sampling scheme is used. The possible samples of size two are

    $$\{x_1, x_2\}, \{x_2, x_3\}, \{x_3, x_4\}, \{x_1, x_4\}$$

    and the sampling is done in such a way that each of these four possible samples is equally likely. Is the sample mean unbiased?

    Solution. Here, let $\bar{X}$ denote the sample mean. Then

    $$\begin{aligned} E[\bar{X}] &= \frac{1}{4}\left(\frac{x_1 + x_2}{2} + \frac{x_2 + x_3}{2} + \frac{x_3 + x_4}{2} + \frac{x_1 + x_4}{2}\right) \\ &= \frac{1}{4}(x_1 + x_2 + x_3 + x_4) \\ &= \mu \end{aligned}$$

    so the sample mean is $\boxed{\text{unbiased.}}$

    □

16. True or false?

    a. The center of a 95% confidence interval for the population mean is a random variable.

    Answer. $\boxed{\text{True.}}$ The center of a confidence interval is the sample mean $\bar{X}$ which we know is a random variable.

    b. A 95% confidence interval for μ contains the sample mean with probability 0.95.

    Answer. $\boxed{\text{False.}}$ The interval is constructed so that it is centered at the sample mean, thus contains it with probability 1.

    c. A 95% confidence interval contains 95% of the population.

    Answer. $\boxed{\text{False.}}$ This is not how a confidence interval is defined.

    d. Out of one hundred 95% confidence intervals for μ, 95 will contain μ.

    Answer. $\boxed{\text{False.}}$ Since confidence intervals are random intervals, the expected number of intervals containing μ is 95, but we are not guaranteed exactly 95.

17. A 90% confidence interval for the average number of children per household based on a simple random sample is found to be (0.7, 2.1). Can we conclude that 90% of households have between 0.7 and 2.1 children?

    Answer. $\boxed{\text{No.}}$ This interval has a 90% chance of including the population average number of children per household. It says nothing about the endpoints of the interval.

25. Consider a random permutation $Y_1, Y_2, \cdots, Y_N$ of $x_1, x_2, \cdots, x_N$. Argue that the joint distribution of any subcollection, $Y_{i_1}, \cdots, Y_{i_n}$, of the Y_i is the same as that of a simple random sample, $X_1, \cdots, X_n$. In particular,

$$\mathrm{Var}(Y_i) = \mathrm{Var}(X_k) = \sigma^2$$

and

$$\mathrm{Cov}(Y_i, Y_j) = \mathrm{Cov}(X_k, X_l) = \gamma$$

if $i \neq j$ and $k \neq l$. Since $Y_1 + Y_2 + \cdots + Y_N = \tau$,

$$\mathrm{Var}\left(\sum_{i=1}^{N} Y_i\right) = 0$$

Express $\mathrm{Var}\left(\sum_{i=1}^{N} Y_i\right)$ in terms of σ^2 and the unknown covariance, γ. Solve for γ, and conclude that

$$\gamma = -\frac{\sigma^2}{N-1}$$

for $i \neq j$.

Solution. We may write the variance of the sum of random variables entirely as a double sum of the pairwise covariances. Thus,

$$\begin{aligned}
\mathrm{Var}\left(\sum_{i=1}^{N} Y_i\right) &= \sum_{i=1}^{N}\sum_{j=1}^{N} \mathrm{Cov}(Y_i, Y_j) \\
&= \sum_{k=1}^{N} \mathrm{Cov}(Y_k, Y_k) + \sum_{i \neq j} \mathrm{Cov}(Y_i, Y_j) \\
&= N\sigma^2 + N(N-1)\gamma = 0
\end{aligned}$$

and solving for γ we get

$$\gamma = -\frac{\sigma^2}{N-1}$$

as desired.

□

26. Let U_i be a random variable with $U_i = 1$ if the ith population member is in the sample and equal to 0 otherwise.

a. Show that the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{N} U_i x_i$.

Proof. The sample mean is given by

$$\bar{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$$

where $\{X_1, X_2, \cdots, X_n\}$ is a sample of the population $\{x_1, x_2, \cdots, x_N\}$. Thus we need to show that

$$\sum_{i=1}^{N} U_i x_i = \sum_{j=1}^{n} X_i.$$

This is trivial since if x_i is included in the sample, it is added onto the sum, and not otherwise. Since we eventually cover all of x_i, we will eventually add all elements of the sample, and none from elements that are not in the sample, which is precisely the right hand sum, as desired.

□

b. Show that $P(U_i = 1) = n/N$. Find $E[U_i]$, using the fact that U_i is a Bernoulli random variable.

Proof. The sample has n elements, and the population has N. To count how many samples contain x_i, we fix x_i and choose $n-1$ elements from the remaining $N-1$, so there are $\binom{N-1}{n-1}$ of interest. There are a total of $\binom{N}{n}$ possible samples of size n, so the probability our sample contains x_i is

$$\frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{\frac{(N-1)!}{(n-1)!(N-n)!}}{\frac{N!}{n!(N-n)!}} = \frac{n}{N}$$

as desired. Then $E[U_i] = \boxed{\frac{n}{N}}$ since U_i is Bernoulli.

□

c. What is the variance of the Bernoulli random variable U_i?

Solution. The variance of a Bernoulli random variable is given by $\sigma^2 = p(1-p)$, so

$$\mathrm{Var}(U_i) = \frac{n}{N}\left(1 - \frac{n}{N}\right) = \boxed{\frac{n(N-n)}{N^2}.}$$

□

d. Noting that $U_i U_j$ is a Bernoulli random variable, find $E[U_i U_j], i \neq j$.

Solution. We find the event $U_i U_j = 1$ means that both $U_i = 1$ and $U_j = 1$. We want the probability $P(U_i = 1, U_j = 1)$ which is the probability the random sample contains both x_i and x_j. Since the sample has size n, after fixing x_i and x_j, we must choose $n-2$ elements from the remaining $N-2$, which can be done in $\binom{N-2}{n-2}$ ways. There are a total of $\binom{N}{n}$ samples of size n, so the probability a randomly chosen sample contains both x_i and x_j is

$$\frac{\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{\frac{(N-2)!}{(n-2)!(N-n)!}}{\frac{N!}{n!(N-n)!}} = \frac{n(n-1)}{N(N-1)}$$

Thus

$$E[U_i Uj] = P(U_i U_j = 1) = P(U_i = 1, U_j = 1) = \boxed{\frac{n(n-1)}{N(N-1)}}$$

□

e. Find $\mathrm{Cov}(U_i, U_j), i \neq j$.

Solution. By definition,

$$\begin{aligned}\mathrm{Cov}(U_i, U_j) &= E[U_i U_j] - E[U_i]E[U_j] \\ &= \frac{n(n-1)}{N(N-1)} - \frac{n}{N} \cdot \frac{n}{N} \\ &= \frac{n}{N}\left(\frac{n-1}{N-1} - \frac{n}{N}\right) \\ &= \boxed{-\frac{n(N-n)}{N^2(N-1)}}\end{aligned}$$

□

f. Using the representation of $\bar{X}$ above, find $\mathrm{Var}(\bar{X})$.

Solution. We have

$$\begin{aligned}\mathrm{Var}(\bar{X}) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{N} U_i x_i\right) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{N} U_i x_i\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathrm{Cov}(U_i x_i, U_j x_j) \\ &= \frac{1}{n^2}\sum_{i=1}^{N}\sum_{j=1}^{N} x_i x_j \mathrm{Cov}(U_i, U_j) \\ &= \frac{1}{n^2}\left(\sum_{i\neq j} x_i x_j \cdot \left(-\frac{n(N-n)}{N^2(N-1)}\right) + \sum_{i=1}^{N} x_i x_i \mathrm{Cov}(U_i, U_i)\right) \\ &= \frac{1}{n^2}\left[\left(-\frac{n(N-n)}{N^2(N-1)}\right)\sum_{i\neq j} x_i x_j + \left(\frac{n(N-n)}{N^2}\right)\sum_{k=1}^{N} x_k^2\right] \\ &= \frac{N-n}{nN^2(N-1)}\left[(N-1)\sum_{k=1}^{N} x_k^2 - \sum_{i\neq j} x_i x_j\right] \\ &= \frac{N-n}{nN^2(N-1)}\left(N\sum_{k=1}^{N} x_k^2 - \sum_{i=1}^{N}\sum_{j=1}^{N} x_i x_j\right) \\ &= \frac{N-n}{n(N-1)}\left[\frac{1}{N}\sum_{k=1}^{N} x_k^2 - \left(\frac{1}{N}\sum_{i=1}^{N} x_i\right)\left(\frac{1}{N}\sum_{j=1}^{N} x_j\right)\right] \\ &= \boxed{\frac{N-n}{n(N-1)}\sigma^2.}\end{aligned}$$

□

28. 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, 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 $r = (2p-1)q + (1-p)$.

    Proof. Let Y be the event that a respondent answers Yes, so that $P(Y) = r$. Let A be the event the respondent has characteristic A, so that $P(A) = q$. Let W be the event that statement 1 is responded to, so that $P(W) = p$. Assuming the respondents are entirely truthful and that A and W are independent, we can write

$$\begin{aligned} P(Y) &= P(Y \cap A) + P(Y \cap A^c) \\ &= P(Y \cap A \cap W) + P(Y \cap A \cap W^c) + P(Y \cap A^c \cap W) + P(Y \cap A^c \cap W^c) \\ &= P(Y|A \cap W)P(A \cap W) + P(Y|A \cap W^c)P(A \cap W^c) \\ &\quad + P(Y|A^c \cap W)P(A^c \cap W) + P(Y|A^c \cap W^c)P(A^c \cap W^c) \\ &= 1 \cdot (pq) + 0 \cdot (1-p)q + 0 \cdot p(1-q) + 1 \cdot (1-p)(1-q) \\ &= pq + 1 - q - p + pq \\ &= (2p-1)q + (1-p) = r \end{aligned}$$

    as desired.

    □

    b. If r were known, how could q be determined?

    Solution. We can solve directly for q :

$$q = \boxed{\frac{r+p-1}{2p-1}.}$$

    □

    c. Show that $E[R] = r$, and propose an estimate, Q, for q. Show that the estimate is unbiased.

    Proof. For a sample $\{X_1, X_2, \cdots, X_n\}$, let $X_i = 1$ if the ith person answers Yes, and 0 otherwise. Then we may write R as

$$R = \frac{1}{n}\sum_{i=1}^{n} X_i$$

    since this is dichotomous. Now, $E[X_i] = P(X_i = 1) = r$ since X_i is Bernoulli, thus

$$\begin{aligned} E[R] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] &= \frac{1}{n}\sum_{i=1}^{n} E[X_i] \\ &= \frac{1}{n}\sum_{i=1}^{n} r = r \end{aligned}$$

    as desired.

Consider a sample $\{X_1, X_2, \cdots, X_n\}$, such that $X_i = 1$ if the ith respondent has characteristic A. Hence, X_i is Bernoulli with $E[X_i] = P(X_i = 1) = q$. Then let

$$Q = \frac{1}{n}\sum_{i=1}^{n} X_i$$

be an estimate for q. We claim it is unbiased. Indeed, it holds that

$$\begin{aligned} E[Q] &= E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] \\ &= \frac{1}{n}\sum_{i=1}^{n} q = q \end{aligned}$$

so Q is an unbiased estimator for q, as desired.

□

d. Ignoring the finite population correction, show that

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

where n is the sample size.

Proof. As in part c, define

$$R = \frac{1}{n}\sum_{i=1}^{n} X_i$$

where X_i are from a sample $\{X_1, X_2, \cdots, X_n\}$ and $X_i = 1$ if the ith respondent says Yes. Here, $E[X_i] = r$ and $\mathrm{Var}(X_i) = r(1-r)$ since X_i is Bernoulli. Then

$$\mathrm{Var}(R) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right)$$

Ignoring the finite population correction essentially assumes that each of X_i are independent, so that the variance of their sum is the sum of their variances:

$$\frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} r(1-r) = \frac{1}{n^2} nr(1-r) = \frac{r(1-r)}{n}$$

as desired.

□

e. Find an expression for $\mathrm{Var}(Q)$.

Solution. Since Q is to q as R is to r, that is, they are both estimates of a proportion of the population, $\mathrm{Var}(Q)$ is analogous to $\mathrm{Var}(R)$ so that

$$\mathrm{Var}(Q) \approx \boxed{\frac{q(1-q)}{n}}$$

if we ignore the finite population correction.

□

33. Two populations are independently surveyed using simple random samples of size n, and two proportions, p_1 and p_2 are estimated. It is expected that both population proportions are close to 0.5. What should the sample size be so that the standard error of the difference, $\hat{p_1} - \hat{p_2}$, will be less than 0.02?

Solution. Since $\hat{p_1}$ and $\hat{p_2}$ are independent, we have

$$\text{Var}(\hat{p_1} - \hat{p_2}) = \text{Var}(\hat{p_1}) + \text{Var}(\hat{p_2}).$$

Let p_1 and p_2 be the population proportions from the two populations. We don't assume the finite population correction because we don't know the size of the population. The variances of $\hat{p_1}$ and $\hat{p_2}$ are given by

$$\sigma^2_{\hat{p_1}} = \frac{p_1(1-p_1)}{n}$$
$$\sigma^2_{\hat{p_2}} = \frac{p_2(1-p_2)}{n}$$

so then

$$\begin{aligned}\sigma^2_{\hat{p_1}-\hat{p_2}} &= \frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{n} \\ &= \frac{1}{n}\left[p_1(1-p_1) + p_2(1-p_2)\right] \\ \implies \sigma_{\hat{p_1}-\hat{p_2}} &= \sqrt{\frac{1}{n}}\sqrt{p_1(1-p_1) + p_2(1-p_2)}\end{aligned}$$

We are given that $p_1, p_1 \approx 0.5$, and we want this to be less than 0.02. Solving for n,

$$\begin{aligned}\sqrt{\frac{1}{n}}\sqrt{0.5^2 + 0.5^2} &< 0.02 \\ \frac{1}{n}(0.25 + 0.25) &< 0.0004 \\ n &> 1250\end{aligned}$$

so the sample size should be $\boxed{\text{larger than 1250.}}$

□

34. 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 0.01? What are the actual standard errors for this sample size?

    Solution. Let p_1 and p_2 represent the proportions of the population that have problem 1 and problem 2, respectively. Let the sample size be n. Then, discounting the finite population correction, we have the following:

$$\sigma_{\hat{p_1}} = \sqrt{\frac{p_1(1-p_1)}{n}} = \sqrt{\frac{0.03 \cdot 0.97}{n}} = \sqrt{\frac{0.0291}{n}}$$
$$\sigma_{\hat{p_2}} = \sqrt{\frac{p_2(1-p_2)}{n}} = \sqrt{\frac{0.40 \cdot 0.60}{n}} = \sqrt{\frac{0.24}{n}}$$

Since we want both standard errors to be less than 0.01, this amounts to solving for the minimum value of n. Since $\sigma_{\hat{p_2}}$ is the larger of the two, we only need to consider this case.

$$\begin{aligned}\sigma_{\hat{p_2}} = \sqrt{\frac{0.24}{n}} &< 0.01 \\ \frac{0.24}{n} &< 0.0001 \\ n &> 2400\end{aligned}$$

so the sample size should be $\boxed{\text{at least 2401}}$ which results in standard errors of

$$\sigma_{\hat{p_1}} = \sqrt{\frac{0.0291}{2401}} \approx \boxed{0.0035}$$

$$\sigma_{\hat{p_2}} = \sqrt{\frac{0.24}{2401}} \approx \boxed{0.01}$$

□

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?

Solution. Since the standard error is defined as $\sigma_{\hat{p}} = \frac{\sigma_p}{\sqrt{n}}$ where n is the sample size and σ_p is the true value, we just want

$$\sigma_{\hat{p}} = \frac{\sigma_p}{\sqrt{n}} < 0.10\sigma_p$$

which means

$$\begin{aligned}\sqrt{n} &> 10 \\ \implies n &> 100\end{aligned}$$

so the sample size should be $\boxed{\text{at least 101.}}$

□

36. With simple random sampling, is $\bar{X}^2$ an unbiased estimate of μ^2? If not, what is the bias?

Solution. $\bar{X}^2$ is an unbiased estimate of μ^2 if and only if $E[\bar{X}^2] = \mu^2$.

We have the relation $\text{Var}(\bar{X}) = E[\bar{X}^2] - (E[\bar{X}])^2$. We know that $E[\bar{X}] = \mu$ and $\text{Var}(\bar{X}) = \frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right)$. Thus,

$$\begin{aligned}\text{Var}(\bar{X}^2) &= E[\bar{X}^2] - (E[\bar{x}])^2 \\ \frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right) &= E[\bar{X}^2] - \mu^2 \\ E[\bar{X}^2] &= \mu^2 + \frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right)\end{aligned}$$

so as long as $\sigma^2 \neq 0$ and $n \neq N$, it follows that $\bar{X}^2$ is $\boxed{\text{not an unbiased estimator}}$ for μ^2. The bias is

$$\begin{aligned}E[\bar{X}^2 - \mu^2] &= E[\bar{X}^2] - E[\mu^2] \\ &= \mu^2 + \frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right) - \mu^2 \\ &= \boxed{\frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right)}\end{aligned}$$

□

38. Let $X_1, \cdots, X_n$ be a simple random sample. Show that $\frac{1}{n}\sum_{i=1}^{n} X_i^3$ is an unbiased estimate of $\frac{1}{N}\sum_{i=1}^{N} x_i^3$.

Proof. The former sum S is an unbiased estimator of the latter sum T if and only if $E[S] = T$. We have

$$E\left[\frac{1}{n}\sum_{i=1}^{n} X_i^3\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i^3].$$

To calculate $E[X_i^3]$, consider the distribution of X_i for some fixed i. We have $P(X_i = x_j) = \frac{1}{N}$ for all x_j in the population, since any X_i in the sample has a $\frac{1}{N}$ chance of being the x_j. Then, applying the Law of the Unconscious Statistician, we have

$$E[X_i^3] = \sum_{j=1}^{N} x_j^3 P(X_i = x_j) = \sum_{j=1}^{N} x_j^3 \cdot \frac{1}{N},$$

and substituting back above, we have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} E[X_i^3] &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{N}\sum_{j=1}^{N} x_j^3\right) \\ &= \frac{1}{n}\cdot n\left(\frac{1}{N}\sum_{j=1}^{N} x_j^3\right) \\ &= \frac{1}{N}\sum_{j=1}^{N} x_j^3\end{aligned}$$

as desired.

□