Exercise 1 (Simple random sampling):
Let there be two correlated random variables $X$ and $Y$. A sample of size $n$ is drawn from a population by simple random sampling without replacement. The observed paired sample is $(X_i, Y_i), i = 1, 2, ..., n$. If the sample totals are $\hat{X}_{tot} = \frac{N}{n} \sum_{i=1}^{n} X_i$, $\hat{Y}_{tot} = \frac{N}{n} \sum_{i=1}^{n} Y_i$, then find the covariance between $\hat{X}_{tot}$ and $\hat{Y}_{tot}$.

Solution:
The covariance between $X_{tot}$ and $Y_{tot}$ is

$$
\text{Cov}(\hat{X}_{tot}, \hat{Y}_{tot}) = E \left[ \left( \hat{X}_{tot} - E(\hat{X}_{tot}) \right) \left( \hat{Y}_{tot} - E(\hat{Y}_{tot}) \right) \right] 
= E(\hat{X}_{tot} \hat{Y}_{tot}) - E(\hat{X}_{tot})E(\hat{Y}_{tot}).
$$

Note that

$$E(\hat{X}_{tot}) = X_{tot} = \sum_{i=1}^{N} X_i = N\bar{X}$$

$$E(\hat{Y}_{tot}) = Y_{tot} = \sum_{i=1}^{N} Y_i = N\bar{Y}$$

where $\bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_i$, $\bar{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i$,

$$E(X_i Y_i) = \frac{1}{N} \sum_{i=1}^{N} X_i Y_i$$

$$E(X_i Y_j) = \frac{1}{N(N-1)} \sum_{i,j=1}^{N} X_i Y_j.$$  

Also,

$$\sum_{i=1}^{N} X_i \sum_{j=1}^{N} Y_j = \sum_{i=1}^{N} X_i Y_i + \sum_{i,j=1}^{N} X_i Y_j$$

$$\Rightarrow \sum_{i,j=1}^{N} X_i Y_j = \frac{N^2 \bar{X} \bar{Y}}{N} - \sum_{i=1}^{N} X_i Y_i$$

where $\bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_i$, $\bar{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i$.

Then
\[
\text{Cov}(\hat{X}_{tot}, \hat{Y}_{tot}) = E\left[\left(\frac{N}{n}\right)^2 \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i\right] - N^2 \bar{XY}
\]
\[
= \left(\frac{N}{n}\right)^2 \left[\sum_{i=j=1}^{N} \sum_{i=j=1}^{N} E(X_i Y_j) + \sum_{i=j=1}^{N} \sum_{i=j=1}^{N} X_i \sum_{i=j=1}^{N} Y_i\right] - N^2 \bar{XY}
\]
\[
= \left(\frac{N}{n}\right)^2 \left[\frac{n}{N} \sum_{i=1}^{N} X_i Y_i + \frac{n(n-1)}{N(N-1)} \sum_{i=j=1}^{N} \sum_{i=j=1}^{N} X_i Y_i\right] - N^2 \bar{XY}
\]
\[
= \left(\frac{N}{n}\right)^2 \left[\frac{n}{N} \sum_{i=1}^{N} X_i Y_i + \frac{n(n-1)}{N(N-1)} \left(N^2 \bar{XY} - \sum_{i=1}^{N} X_i Y_i\right)\right] - N^2 \bar{XY}
\]
\[
= \left(\frac{N}{n}\right)^2 \left[\left(\frac{n}{N}\right) \left(1 - \frac{n(n-1)}{N(N-1)}\right) \sum_{i=1}^{N} X_i Y_i + \frac{n(n-1)}{N(N-1)} \sum_{i=1}^{N} X_i Y_i\right] - N^2 \bar{XY}
\]
\[
= \frac{N^2(N-n)}{Nn} \frac{1}{N-1} \sum_{i=1}^{N} (X_i - \bar{X})(Y_i - \bar{Y})
\]
\[
= \frac{N^2(N-n)}{Nn} S_{XY}
\]
where \(S_{XY} = \frac{1}{N-1} \sum_{i=1}^{N} (X_i - \bar{X})(Y_i - \bar{Y})\).

**Exercise 2 (Simple random sampling):**

Under the simple random sampling without replacement, find \(E(s_{xy})\) where

\[
s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{x})(Y_i - \bar{y})
\]
\[
\bar{x} = \frac{1}{n} \sum_{i=1}^{n} X_i, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} Y_i.
\]

**Solution:**

Consider

\[
s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{x})(Y_i - \bar{y})
\]
\[
= \frac{1}{n-1} \left[ \sum_{i=1}^{n} X_i Y_i - n\bar{x}\bar{y}\right]
\]
\[
= \frac{1}{n-1} \left[ \sum_{i=1}^{n} X_i Y_i - \frac{n}{n^2} \left( \sum_{i=1}^{n} X_i \right) \left( \sum_{i=1}^{n} Y_i\right)\right]
\]
\[
= \frac{1}{n-1} \left[ \sum_{i=1}^{n} X_i Y_i - \frac{1}{n} \left( \sum_{i=1}^{n} X_i Y_i + \sum_{i=j=1}^{n} \sum_{j=1}^{n} X_i Y_j\right)\right]
\]
\[
= \frac{1}{n-1} \left[ \frac{n-1}{n} \sum_{i=1}^{n} X_i Y_i - \frac{1}{n} \sum_{i=j=1}^{n} \sum_{j=1}^{n} X_i Y_j\right],
\]
Since \( E \left( \sum_{i=1}^{n} X_i Y_i \right) = \frac{n}{N} \sum_{i=1}^{N} X_i Y_i \)
\( E \left( \sum_{i=1}^{n} \sum_{j=1}^{n} X_i Y_j \right) = \frac{n(n-1)}{N(N-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} X_i Y_j. \)

Thus
\[
E(s_{xy}) = \frac{1}{n-1} \left[ \frac{n-1}{n} \frac{N}{n} \sum_{i=1}^{n} X_i Y_i - \frac{n(n-1)}{nN(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} X_i Y_j \right]
\]
\[= \frac{1}{N} \sum_{i=1}^{N} X_i Y_i - \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} X_i Y_j
\]
\[= \frac{1}{N} \sum_{i=1}^{N} X_i Y_i - \frac{1}{N(N-1)} \left[ N^2 \bar{X}\bar{Y} - \sum_{i=1}^{N} X_i Y_i \right]
\]
\[= \left( \frac{1}{N} + \frac{1}{N(N-1)} \right) \sum_{i=1}^{N} X_i Y_i - \left( \frac{N}{N-1} \right) \bar{X}\bar{Y}
\]
\[= \frac{1}{N-1} \left[ \sum_{i=1}^{n} X_i Y_i - N \bar{X}\bar{Y} \right]
\]
\[= \frac{1}{N-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})
\]
\[= S_{xy}.
\]

**Exercise 3 (Simple random sampling):**

Suppose in a population of size \( N \), \( Y_i \) and \( Y_N \) are two extreme values in the sense that \( Y_i \) denotes the extremely low and \( Y_N \) denotes the extremely high values among the sample units \( Y_1, Y_2, \ldots, Y_N \). Instead of using the sample mean \( \bar{y} = \frac{1}{n} \sum_{i=1}^{n} Y_i \) as an estimator of the population mean \( \bar{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i \) in such a case, the sample mean estimator is modified as follows:
\[
\begin{align*}
\bar{y} + k & \quad \text{if sample contains } Y_i \text{ but not } Y_N \\
\bar{y} - k & \quad \text{if sample contains } Y_N \text{ but not } Y_i \\
\bar{Y} & \quad \text{for all other samples}
\end{align*}
\]

where \( k > 0 \) is known constant. Determine if \( \bar{y} \) is an unbiased estimator of population mean \( \bar{Y} \) and find its variance.

**Solution:**
Suppose the population units $Y_1, Y_2, \ldots, Y_N$ are labeled as $1, 2, \ldots, N$. So $Y_i$ and $Y_N$ now corresponds to labels 1 and $N$ respectively.

Since the sample space corresponding to $\tilde{y}$ has three possible situations, so it can be divided into three disjoint subsets as follows:

$\Omega = \{ \text{All the units in the sample} \}$

$\Omega_1 = \{ \text{All the units in the sample are such that unit label 1 is present and unit label } N \text{ is absent} \}$

$\Omega_2 = \{ \text{All the units in the sample are such that unit label } N \text{ is present and unit label } 1 \text{ is absent} \}$

$\Omega_3 = \Omega - \Omega_1 - \Omega_2$.

When a sample of size $n$ is drawn from a population of size $N$, there are $\binom{N}{n}$ possible subsets. Similarly, the number of possible subsets of sample drawn from the population

* $\Omega_1$ are $\binom{N-2}{n-1}$.

* $\Omega_2$ are $\binom{N-2}{n-1}$.

* $\Omega_3$ are $\binom{N}{n} - \binom{N-2}{n-1} - \binom{N-2}{n-1}$.

Under SRSWOR,

$$E(\tilde{y}) = \frac{1}{\binom{N}{n}} \sum_{\Omega} \tilde{y}$$

$$= \frac{1}{\binom{N}{n}} \left[ \sum_{\Omega_1} (\tilde{y} + k) + \sum_{\Omega_2} (\tilde{y} - k) + \sum_{\Omega_3} \tilde{y} \right]$$

$$= \frac{1}{\binom{N}{n}} \left[ \left( \sum_{\Omega_1} \tilde{y} + \sum_{\Omega_2} \tilde{y} + \sum_{\Omega_3} \tilde{y} \right) + k \binom{N-2}{n-1} - k \binom{N-2}{n-1} \right]$$

$$= \frac{1}{\binom{N}{n}} \sum_{\Omega} \tilde{y} = \bar{Y}.$$  

So $\tilde{y}$ is an unbiased estimator of the population mean.
The variance is
\[
\text{Var}(\bar{y}) = \frac{1}{N} \sum_{\Omega} (\bar{y} - \bar{Y})^2
\]
\[
= \frac{1}{N} \left[ \sum_{\Omega_1} (\bar{y} + k - \bar{Y})^2 + \sum_{\Omega_2} (\bar{y} - k - \bar{Y}) + \sum_{\Omega_3} (\bar{y} - \bar{Y})^2 \right]
\]
\[
= \frac{1}{N} \left[ \left( \sum_{\Omega_1} (\bar{y} - \bar{Y})^2 + \sum_{\Omega_2} (\bar{y} - \bar{Y})^2 + \sum_{\Omega_3} (\bar{y} - \bar{Y})^2 \right) \right]
\]
\[
+ 2C^2 \left( \frac{N - 2}{n - 1} \right)^2 - 2C \left[ \sum_{\Omega_1} (\bar{y} - \bar{Y}) - \sum_{\Omega_2} (\bar{y} - \bar{Y}) \right]
\]
\[
= \frac{1}{N} \left[ \sum_{\Omega} (\bar{y} - \bar{Y})^2 + 2C^2 \left( \frac{N - 2}{n - 1} \right)^2 - 2C \left[ \sum_{\Omega_1} (\bar{y} - \bar{Y}) - \sum_{\Omega_2} (\bar{y} - \bar{Y}) \right] \right].
\]

Observe that
\[
* \quad \frac{N - 2}{n - 1} = \frac{n(N - n)}{N(N - 1)}
\]

* \(\frac{N - 2}{n - 1}\) units in \(\Omega_1\) containing the unit label 1, \(\frac{N - 3}{n - 2}\) of them have unit labels \((j = 2, 3, ..., N - 1)\) and none of them contains the unit label \(N\).

Thus
\[
\sum_{\Omega_1} (\bar{y} - \bar{Y}) = \sum_{\Omega_1} \bar{y} - \sum_{\Omega_1} \bar{Y}
\]
\[
= \frac{1}{N} \left[ \left( \frac{N - 2}{n - 1} \right) Y_i + \left( \frac{N - 3}{n - 2} \right) \sum_{j=2}^{N-1} Y_j - \left( \frac{N - 2}{n - 1} \right) \bar{Y} \right]
\]
\[
= \frac{1}{n} \left( \frac{N - 2}{n - 1} \right) \left[ Y_i + \frac{n - 1}{N - 2} \sum_{j=2}^{N-1} Y_j \right] - \left( \frac{N - 2}{n - 1} \right) \bar{Y}.
\]

Similarly
\[
\sum_{\Omega_2} (\bar{y} - \bar{Y}) = \frac{1}{n} \left( \frac{N - 2}{n - 1} \right) \left[ Y_N + \frac{n - 1}{N - 2} \sum_{j=2}^{N-1} Y_j \right] - \left( \frac{N - 2}{n - 1} \right) \bar{Y}.
\]
Finally,

\[ V(\bar{y}) = \frac{1}{n} Var(\bar{y}) + 2k^2 \frac{n(N-n)}{N(N-1)} - 2k \left\{ \frac{1}{n} \left( \frac{N-2}{N-1} \right) \left( Y_n + \frac{n-1}{N-2} \sum_{j=2}^{N-1} Y_j \right) - \left( \frac{N-2}{n-1} \right) \bar{Y} \right\} \]

\[ + 2k \left\{ \frac{1}{n} \left( \frac{N-2}{N-1} \right) \left( Y_N + \frac{n-1}{N-2} \sum_{j=2}^{N-1} Y_j \right) - \left( \frac{N-2}{n-1} \right) \bar{Y} \right\} \]

\[ = \frac{N-2}{N} \left[ \frac{S^2}{n} - \frac{2k}{N-1} (Y_N - Y_i - nk) \right]. \]

**Example 4 (Simple random sampling):**

Let a sample of size 2 is drawn from a population of size 3 having units \( Y_1, Y_2 \) and \( Y_3 \). The following estimator is used to estimate the population mean depending upon which sampling units are chosen

\[ \tilde{y} = \begin{cases} 
\frac{Y_1 + Y_2}{2}, & \text{if sample is } (Y_1, Y_2) \\
\frac{Y_1 + 2Y_2}{3}, & \text{if sample is } (Y_2, Y_1) \\
\frac{Y_2 + Y_3}{3}, & \text{if sample is } (Y_2, Y_3).
\end{cases} \]

Verify if \( \tilde{y} \) is an unbiased estimator of the population mean or not and find its variance.

**Solution:** The sample space \( \Omega = \{Y_1, Y_2, Y_3\} \).

\[ E(\tilde{y}) = \frac{1}{3} \left[ \left( \frac{Y_1}{2} + \frac{Y_2}{2} \right) + \left( \frac{Y_1}{2} + \frac{2Y_2}{3} \right) + \left( \frac{Y_2}{2} + \frac{Y_3}{3} \right) \right] \]

\[ = \frac{1}{3} (Y_1 + Y_2 + Y_3) \]

\[ = \bar{Y} \]

which implies that \( \tilde{y} \) is an unbiased estimator of \( \bar{Y} \). Its variance is

\[ Var(\tilde{y}) = \frac{1}{3} \left[ \left( \frac{Y_1}{2} + \frac{Y_2}{2} \right)^2 + \left( \frac{Y_1}{2} + \frac{2Y_2}{3} \right)^2 + \left( \frac{Y_2}{2} + \frac{Y_3}{3} \right)^2 \right] - \bar{Y}^2 \]

\[ = \frac{1}{3} \left[ Y_1^2 \left( \frac{1}{4} + \frac{1}{4} \right) + Y_2^2 \left( \frac{1}{4} + \frac{1}{4} \right) + Y_3^2 \left( \frac{4}{9} + \frac{1}{9} \right) + 2 \left( \frac{Y_1 Y_2}{4} + \frac{Y_1 Y_3}{6} + \frac{Y_2 Y_3}{6} \right) \right] - \left( \frac{Y_2 + Y_3}{3} \right)^2 \]

\[ = \frac{1}{9} \left( Y_1^2 + Y_2^2 + Y_3^2 \right) - \frac{2Y_1 Y_2}{3} - \frac{Y_2 Y_3}{2} - Y_3^2. \]
**Exercise 5 (Simple random sampling):**

The population coefficient of variation of a variable $Y$ is defined as $C_Y = \frac{S_Y}{\bar{Y}}$ where $S_Y^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left( Y_i - \bar{Y} \right)^2$, $\bar{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i$ is based on population size $N$. Let the sample mean $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$ is based on sample size $n$ by SRSWR which is usually employed to estimate $\bar{Y}$. Assuming $C_Y$ to be known, improve the sample mean in the sense that the estimator has a minimum mean squared error. Find out the relative efficiency of this estimator relative to sample mean.

**Solution:** Consider a scalar multiple of $\bar{y}$ to estimate $\bar{Y}$ as

$$\bar{y}_k = k\bar{y}$$

where $k$ is any scalar. The mean squared error of $\bar{y}_k$ is

$$M(\bar{y}_k) = E\left[ k\bar{y} - \bar{Y} \right]^2$$

$$= E\left[ k(\bar{y} - \bar{Y}) + (k-1)\bar{Y} \right]^2$$

$$= k^2 Var(\bar{y}_{SRSWR}) + (k-1)^2 \bar{Y}^2 + 0$$

$$= k^2 \frac{N-1}{N} S_Y^2 + (k-1)^2 \bar{Y}^2$$

$$= \bar{Y}^2 \left[ k^2 \left( \frac{N-1}{Nn} \right) C_Y^2 + (k-1)^2 \right].$$

Now we use the principle of maxima/minima s follows to find the optimum value of $k$ for which $M(\bar{y}_k)$ is minimum

$$\frac{dM(\bar{y}_k)}{dk} = \bar{Y}^2 \left[ 2k \left( \frac{N-1}{Nn} \right) C_Y^2 + 2(k-1) \right].$$

Substituting $\frac{dM(\bar{y}_k)}{dk} = 0$

$$\Rightarrow k \left[ 1 + \left( \frac{N-1}{Nn} \right) C_Y^2 \right] = 1$$

$$\Rightarrow k = \frac{1}{1 + \frac{N-1}{Nn} C_Y^2} = k^*, \text{ say.}$$

The second order condition for maxima/minima is satisfied.
Thus the optimum estimator is

\[ \bar{Y}_k = k^* Y = \frac{\bar{Y}}{1 + \frac{N-1}{N} C^2_y} \]

and its minimum mean squared error is

\[
M(\bar{Y}_k) = \bar{Y}^2 \left[ k^* \left( \frac{N-1}{N} \right) C^2_y + \left( k^* - 1 \right)^2 \right] \\
= \bar{Y}^2 \left[ \frac{N-1}{N} C^2_y \left( 1 + \frac{N-1}{N} C^2_y \right) + \left( \frac{N-1}{N} \right) C^4_y \left( 1 + \frac{N-1}{N} C^2_y \right) \right] \\
= \bar{Y}^2 \frac{N-1}{Nn} C^2_y \left( 1 + \frac{N-1}{Nn} C^2_y \right) \left[ 1 + \frac{N-1}{Nn} C^2_y \right] \\
= \bar{Y}^2 \frac{N-1}{Nn} C^2_y \left( \frac{1}{1 + \frac{N-1}{Nn} C^2_y} \right) \\
= Var(\bar{Y})_{SRSWR} \\
Var(\bar{Y}) = \frac{N-1}{Nn} C^2_y \\
Relative \ efficiency = \frac{M(\bar{Y}_k)}{Var(\bar{Y})} = \left[ 1 + \frac{N-1}{Nn} C^2_y \right]^{-1}.

Exercise 6 (Sampling for proportions):

Suppose we want to estimate the proportion of men employees (#P) in an organization having 1500 total employees. In addition, suppose 3 out of 10 employees are men. Suppose a sample is drawn by simple random sampling. Find the sample size to be selected so that the total length of confidence interval with confidence level 0.05 is less than 0.02 for SRSWR and SRSWOR.

Solution:

First, consider that a sample of size #n is drawn from a population of size #N by SRSWR. The 100(1−#\alpha)% confidence interval for sample mean (\bar{Y}) or equivalently the sample proportion (#p_{sr}) is given by
The total length of confidence interval is

\[ 2z_{\alpha/2} \sqrt{\frac{p_{wr}(1-p_{wr})}{n-1}}, \]

where \( p_{wr} = \frac{3}{10} \), \( \alpha = 0.05 \), and \( z_{\alpha/2} = 1.96 \).

The sample size is greater than the given population size in this case but since SRSWR is used, so it is not impossible.

Next, we consider the case when the sample is drawn by SRSWOR. The related 100(1-\( \alpha \))% confidence interval for \( \bar{y} \) or equivalently the proportion \( (p_{wor}) \) is

\[ \left[ \bar{y} - z_{\alpha/2} \sqrt{\frac{N-n}{Nn} s^2_{\bar{y}}}, \bar{y} + z_{\alpha/2} \sqrt{\frac{N-n}{Nn} s^2_{\bar{y}}} \right], \]

where \( N \) is the population size, \( n \) is the sample size, \( \bar{y} \) is the sample mean, and \( s^2_{\bar{y}} \) is the variance of the sample mean.

The total length of the confidence interval is

\[ 2z_{\alpha/2} \sqrt{\frac{N-n}{Nn} \frac{p_{wor}(1-p_{wor})}{n-1}}, \]

where \( p_{wor} = \frac{3}{10} \), \( \alpha = 0.05 \), and \( z_{\alpha/2} = 1.96 \).

The sample size is greater than the given population size in this case but since SRSWR is used, so it is not impossible.
It is given that
\[
2 \frac{z_{2a}}{2} \sqrt{\frac{N-n}{Nn} \frac{p_{wot} (1-p_{wot})}{n-1}} \leq 0.02
\]
or
\[
z_{2a}^2 \left( \frac{N-n}{Nn} \right) \frac{p_{wot} (1-p_{wot})}{n-1} \leq 0.0001
\]
or
\[
n \geq \frac{0.0001 + z_{2a}^2 \phi_{wot} (1-p_{wot})}{0.0001 + z_{2a}^2 \phi_{wot} (1-p_{wot})}. \]

Since \( p_{wot} = \frac{3}{10} = 0.3 \), \( z_{2a} = 1.96 \), \( N = 1500 \), we have
\[
n \geq \frac{0.0001 + (1.96)^2 0.3(1-0.3)}{0.0001 + (1.96)^2 0.3(1-0.3)} \cdot \frac{1500}{1500}
\]
\[
\approx 1265.
\]

**Exercise 7: (Varying Probability scheme):**

Show that the Yates-Grundy estimator is non-negative under Midzuno sampling design.

**Solution:** A sufficient condition for the Yates-Grundy estimator to be non-negative is
\[
\pi_i \pi_j - \pi_{ij} \geq 0, \ i \neq j, i, j = 1, 2, ..., N.
\]
The expression of \( \pi_i, \pi_j \), and \( \pi_{ij} \) are
\[
\pi_i = \left[ \frac{N-n X_i + n-1}{N-1 X_{tot} + n-1} \right]
\]
\[
\pi_j = \left[ \frac{N-n X_j + n-1}{N-1 X_{tot} + n-1} \right]
\]
\[
\pi_{ij} = \frac{(N-n)(n-1)}{(N-1)(N-2)} \left( \frac{X_i + X_j}{X_{tot}} \right) + \frac{(n-1)(n-2)}{(N-1)(N-2)}
\]
where \( X_{tot} = \sum_{i=1}^{N} X_i \) is the population total.

Now
\[
\pi_{i} - \pi_{j} = \left[ \frac{N-n}{N} X_i + n-1 \right] \left[ \frac{N-n}{N} X_j + n-1 \right] - \frac{(N-n)(n-1)}{(N-1)(N-2)} \frac{X_i + X_j}{X_{tot}} - \frac{(n-1)(n-2)}{(N-1)(N-2)}
\]

\[
= \left[ \frac{N-n}{N-1} \right]^2 X_i X_j + \frac{(N-n)(n-1)}{N-1} \left( \frac{X_i + X_j}{X_{tot}} \right) \left( \frac{1}{N-1} - \frac{1}{N-2} \right) + \frac{n-1}{N-1} \left[ \frac{n-1}{N-1} - \frac{n-2}{N-2} \right]
\]

\[
= \left( \frac{N-n}{N-1} \right)^2 \left( \frac{X_i X_j}{X_{tot}} \right) + \frac{(N-n)(n-1)}{(N-1)^2(N-2)} \left( 1 - \frac{X_i + X_j}{X_{tot}} \right) + \frac{(n-1)(n-1)}{(N-1)^2(N-2)}.
\]

Each of the terms on the right-hand side is nonnegative and so that Yates-Grundy estimator is always nonnegative.

**Exercise 8 (Stratified sampling):**

Suppose that the population of size \( N \) can be expressed as \( N = nk \) where \( n \) and \( k \) are integers. The population is divided into \( n \) strata where the \( h^{th} \) stratum contains units that are labeled as

\[
G_h = \left[ (h-1)k + j, j = 1,2,...,k \right], h = 1,2,...,n.
\]

Suppose only one unit is randomly selected from every stratum and a sample of size \( n \) is obtained. The values of units in the population are modelled as \( Y_i = \alpha + \beta i, i = 1,2,...,N \) where \( \alpha \) and \( \beta \) are some constants. Find the variance of population total.

**Solution:**

The estimate of population total in case of stratified sampling is

\[
\hat{Y}_{st} = \sum_{h=1}^{L} \frac{N_h}{n_h} y_h.
\]

where \( y_h = \sum_{j=1}^{k} y_{hj} \).

Compared to the notations in stratified sampling, we have

- Number of strata \( (L) = n \)
- \( h^{th} \) stratum size \( (N_h) = k \).
- sample size from \( h^{th} \) stratum \( (n_h) = 1 \).

Then
\[ \hat{Y}_s = \hat{Y}_{tot} = \frac{k}{n} \sum_{h=1}^{n} Y_{h1} \quad (j = 1) \]
\[ = k \sum_{h=1}^{n} Y_{ht}. \]

Since the variance of \( \hat{Y}_s \) is
\[ \text{Var}(\hat{Y}_s) = \sum_{h=1}^{L} \frac{N_h^2 (N_h - n_h)}{N_h n_h} S_h^2, \]

so we have
\[ \text{Var}(\hat{Y}_o) = \sum_{h=1}^{k} \frac{k^2 (k-1)}{k} \frac{1}{k-1} \sum_{j=1}^{k} (Y_{hj} - \bar{Y}_h)^2 \]
\[ = k \sum_{h=1}^{n} \sum_{j=1}^{k} (Y_{hj} - \bar{Y}_h)^2. \]

When population values are modeled by the relation \( Y_i = \alpha + \beta i, \ i = 1, 2, ..., N, \) then
\[ Y_{hj} = \alpha + \beta [(h-1)k + j] \]
\[ \bar{Y}_h = \frac{1}{k} \sum_{j=1}^{k} Y_{hj} \]
\[ = \frac{1}{k} \sum_{j=1}^{k} [\alpha + \beta ((h-1)k + j)] \]
\[ = \alpha + \beta [(h-1)k + \frac{k(k-1)}{2}] \]
\[ Y_{hj} - \bar{Y}_h = \beta \left( j - \frac{k+1}{2} \right) \]
\[ \sum_{j=1}^{k} (Y_{hj} - \bar{Y}_h)^2 = \beta^2 \sum_{j=1}^{k} \left[ j^2 + \frac{(k+1)^2}{4} - (k+1)j \right] \]
\[ = \beta^2 \left[ \frac{k(k+1)(2k+1)}{6} + \frac{k(k+1)^2}{4} - \frac{k^2(k+1)^2}{2} \right] \]
\[ = \beta^2 k(k^2 - 1) \frac{12}{12}. \]

Thus
\[ \text{Var}(\hat{Y}_{tot}) = \frac{\beta^2 nk^2 (k^2 - 1)}{12}. \]
Exercise 9 (Stratified sampling):
Suppose there are two strata of sizes $N_1$ and $N_2$ such that $N_1 + N_2 = N$, (population size). Let $\omega_1 = \frac{N_1}{N}$ and $\omega_2 = \frac{N_2}{N}$. Assume the population variances of first and second strata are $S_1^2$ and $S_2^2$ and they are equal, i.e., $S_1 = S_2$. Consider a cost function $C = C_1 n_1 + C_2 n_2$ where $C_1$ and $C_2$ are the cost of collecting an observation from first and second strata respectively. Assuming $N_1$ and $N_2$ to be large, show that

$$\frac{\text{Var} \left( \bar{y}_{\text{st}} \right)}{\text{Var} \left( \bar{y}_{\text{opt}} \right)} = \frac{\omega_1 C_1 + \omega_2 C_2}{[\omega_1 \sqrt{C_1} + \omega_2 \sqrt{C_1}]^2}$$

where $\text{Var} \left( \bar{y}_{\text{st}} \right)$ and $\text{Var} \left( \bar{y}_{\text{opt}} \right)$ are the variance of stratum mean under proportional and optimum allocations in stratified sampling.

Solution
We have

$$\text{Var} \left( \bar{y}_{\text{st}} \right) = \sum_{i=1}^{K} \omega_i^2 \frac{N_i - n_i S_i^2}{N_i n_i}$$

$$\approx \sum_{i=1}^{K} \omega_i^2 \frac{\omega_i^2 S_i^2}{n_i}$$

when $N_i$ is large where $K$ is the number of strata. If cost function, in general, is

$$C = \sum_{i=1}^{K} C_i n_i$$

then under proportional allocation

$$n_i \propto N_i$$

or

$$n_i = d N_i$$

or

$$\sum_{i=1}^{K} C_i n_i = d \sum_{i=1}^{K} C_i N_i$$

or

$$C = d \sum_{i=1}^{K} C_i N_i$$

or

$$d = \frac{C}{\sum_{i=1}^{K} C_i N_i}$$

Thus

$$n_i = \frac{C N_i}{\sum_{i=1}^{K} C_i N_i}$$
is the required sample size.

In the set up of given framework, we have $K = 2$, $S_1 = S_2 = S$, so

$$n_i = \frac{CN_i}{C_i N_1 + C_2 N_2}, \quad i = 1, 2, \ldots$$

Substituting $n_i$ in the expression for the variance

$$Var_{prop}(\bar{y}_{st}) = \frac{1}{N^2} \left[ \frac{N_1^2 S_1^2}{n_1} + \frac{N_2^2 S_2^2}{n_2} \right]$$

$$= \frac{1}{N^2} \left[ \frac{N_1^2}{\left( \frac{CN_1}{C_i N_1 + C_2 N_2} \right)} + \frac{N_2^2}{\left( \frac{CN_2}{C_i N_1 + C_2 N_2} \right)} \right] S^2$$

$$= \left( \frac{C_i N_1 + C_2 N_2}{C} \right) \frac{N_1 + N_2}{N^2} S^2$$

$$= \left( \frac{C_i N_1 + C_2 N_2}{C} \right) \frac{S^2}{N} \quad \text{(Using } N_1 + N_2 = N)$$

$$= \frac{N_i S_i}{C_i} \frac{S^2}{C}$$

The sample size under optimum allocation for the fixed cost is

$$n_k = \frac{N_i S_i}{\sqrt{C_i}}, \quad i = 1, 2, \ldots$$

The variance of $\bar{y}_{st}$ under optimum allocation when $N_i$ is large is
\[ \text{Var}_{\text{opt}}(\bar{y}_{st}) = \frac{1}{N^2} \sum_{i=1}^{2} \frac{N_i^2 S_i^2}{n_i} \]
\[ = \frac{S^2}{N^2} \left[ \frac{N_i^2 \left( \sum_{i=1}^{2} N_i S_i \sqrt{C_i} \right)}{N_i C} \right] + \frac{N^2 \left( \sum_{i=1}^{2} N_i S_i \sqrt{C_i} \right)}{N_2 C} \]
\[ = \frac{S^2}{N^2} \left[ \frac{N_1 \sqrt{C_1} + N_2 \sqrt{C_2}}{C} \right] \left[ \frac{N_1 \sqrt{C_1} + N_2 \sqrt{C_2}}{C} \right] \]
\[ = \frac{S^2}{C} \left( \omega_1 \sqrt{C_1} + \omega_2 \sqrt{C_2} \right)^2. \]

From the expression of \( \text{Var}_{\text{prop}}(\bar{y}_{st}) \) and \( \text{Var}_{\text{opt}}(\bar{y}_{st}) \), we find
\[ \frac{\text{Var}_{\text{prop}}(\bar{y}_{st})}{\text{Var}_{\text{opt}}(\bar{y}_{st})} = \frac{\omega_1 C_1 + \omega_2 C_2}{\left( \omega_1 \sqrt{C_1} + \omega_2 \sqrt{C_1} \right)^2}. \]

**Exercises 10 (Stratified sampling):**

Suppose there are two strata and a sample of sizes \( n_1 \) and \( n_2 \) are drawn from these strata. Let \( \phi_a \) be the actual ratio of \( n_1 \) and \( n_2 \), i.e., \( \phi_a = \frac{n_1}{n_2} \). Let \( \phi_{\text{opt}} \) be the ratio of \( n_1 \) and \( n_2 \) under optimum allocation, i.e.,
\[ \phi_{\text{opt}} = \frac{n_{1(\text{opt})}}{n_{2(\text{opt})}}. \]

Let \( V_a \) and \( V_{\text{opt}} \) be the variances of stratum mean under actual and optimum allocations, respectively. Show that irrespective of the values \( N_1, N_2, S_1, S_1 \) and \( S_2 \), the ratio \( \frac{V_{\text{opt}}}{V_a} \) is never less than
\[ \frac{4\phi}{(1 + \phi)^2} \] when \( N_1 \) and \( N_2 \) are large and \( \phi = \frac{\phi_a}{\phi_{\text{opt}}} \).

**Solution:**

When \( N_1 \) and \( N_2 \) are large, then
\[ \text{Var} \left( \bar{y}_n \right) = V_a = \frac{N_1^2 S_1^2}{n_1} + \frac{N_2^2 S_2^2}{n_2} \]

and variance under optimum allocation is
\[ \text{Var}_{\text{opt}} \left( \bar{y}_n \right) = V_{\text{opt}} = \frac{1}{n} \left( \frac{N_1 S_1 + N_2 S_2}{n} \right)^2. \]

Thus
\[
\frac{V_{\text{opt}}}{V_a} = \frac{1}{n} \left( \frac{N_1 S_1 + N_2 S_2}{n} \right)^2 = \frac{\frac{1}{n} \left( 1 + \frac{N_2 S_2}{N_1 S_1} \right)^2}{\frac{1}{n_1} + \frac{N_2^2 S_2^2}{n_2 N_1^2 S_1^2}}.
\]

Since under optimum allocation
\[
n_1 = \frac{N_1 S_1}{N_1 S_1 + N_2 S_2} n
\]
\[
n_2 = \frac{N_2 S_2}{N_1 S_1 + N_2 S_2} n
\]
\[
\Rightarrow \frac{n_1}{n_2} = \frac{N_1 S_1}{N_2 S_2} = \phi_{\text{opt}}
\]
\[
\phi_a = \frac{n_1}{n_2}
\]
\[
\phi = \frac{\phi_a}{\phi_{\text{opt}}} = \frac{n_1 N_2 S_2}{n_2 N_1 S_1}
\]

Thus
\[
\frac{V_{\text{opt}}}{V_a} = \frac{\frac{1}{n} \left( 1 + \frac{\phi n_2}{n_1} \right)^2}{\frac{1}{n_1} + \phi^2 \frac{n_2^2}{n_1^2 n_2}} = \frac{\frac{1}{n} \left( n_1 + n_2 \phi \right)^2}{n_1 + n_2 \phi^2}.
\]
In this expression, replace

- \( n \) by \( n_1 + n_2 \) and
- \( (n_i + n_2 \phi)^2 \) by \( (n_i - n_2 \phi)^2 + 4n_i n_2 \phi \),

then we have

\[
V_{\text{opt}} = \frac{(n_i - n_2 \phi)^2 + 4n_i n_2 \phi}{(n_i - n_2 \phi)^2 + n_1 n_2 (1 + \phi)^2}.
\]

Using the result that

\[
\frac{z + \alpha}{z + \beta} \geq \frac{\alpha}{\beta} \quad \text{whenever } z \geq 0,
\]

over the expression \( \frac{V_{\text{opt}}}{V_a} \), we notice that since \( 4n_i n_2 \phi \geq n_1 n_2 (1 + \phi)^2 \) is always true, so we have that

\[
\frac{V_{\text{opt}}}{V_a} \geq \frac{4\phi}{(1 + \phi)^2}.
\]

**Exercise 11 (Stratified sampling):**

Suppose there are two strata and equal size of samples are drawn from both the strata i.e., \( n_i = n_2 \). Another option to draw the samples is optimum allocation. Let \( V_e \) and \( V_{\text{opt}} \) denotes the variances under equal allocation \( (n_i = n_2) \) and optimum allocation respectively then assuming \( N_1 \) and \( N_2 \) are large, show that

\[
\frac{V_e - V_{\text{opt}}}{V_{\text{opt}}} = \left(\frac{\delta - 1}{\delta + 1}\right)^2 \quad \text{where } \delta = \frac{n_{\text{opt}}}{n_{2\text{opt}}},
\]

\( n_{\text{opt}} \) and \( n_{2\text{opt}} \) are the sample sizes drawn from first and second strata.

**Solution:**

The variance of \( \bar{Y}_s \) is

\[
\text{Var}(\bar{Y}_s) = \sum_{i=1}^{k} \left(\frac{N_i}{N}\right)^2 \frac{N_i - n_i}{N_i n_i} S_i^2
\]

\[
= \frac{1}{N^2} \sum_{i=1}^{k} N_i^2 \frac{n_i}{n_i} S_i^2
\]

assuming \( N_i \) to be large.
When \( n_1 = n_2 = \frac{n}{2} \), say, then the variance of \( \bar{y}_s^r \) under equal allocation is

\[
V_e = \text{Var}(\bar{y}_s^r) = \frac{2}{nN^2} \left( N_1^2 S_1^2 + N_2^2 S_2^2 \right).
\]

Under optimum allocation,

\[
n_1 = \frac{N_1 S_1}{N_1 S_1 + N_2 S_2} n
\]

\[
n_2 = \frac{N_2 S_2}{N_1 S_1 + N_2 S_2} n
\]

and

\[
V_{opt} = \text{Var}(\bar{y}_s^r) = \frac{1}{N^2} \left[ \frac{N_1^2 S_1^2}{n_{1opt}} + \frac{N_2^2 S_2^2}{n_{2opt}} \right]
\]

\[
= \frac{1}{N^2} \left[ \frac{N_1^2 S_1^2}{nN_1 S_1 + N_2 S_2} + \frac{N_2^2 S_2^2}{nN_2 S_2 + N_1 S_1} \right]
\]

\[
= \frac{1}{N^2} \left( \frac{N_1 S_1 + N_2 S_2}{n} \right)^2.
\]

Since \( \delta = \frac{N_1 S_1}{N_2 S_2} \), so

\[
V_e = \frac{2}{nN^2} N_2^2 S_2^2 \left( 1 + \delta^2 \right)
\]

\[
V_{opt} = \frac{2}{nN^2} N_2^2 S_2^2 \left( 1 + \delta \right)^2
\]

\[
\frac{V_e - V_{opt}}{V_{opt}} = \frac{2}{n} \frac{N_1 S_1^2 \left( \delta^2 - 1 \right) - \frac{1}{n} \frac{N_2^2 S_2^2}{n} \left( 1 + \delta \right)^2}{\frac{1}{n} \frac{N_1^2 S_1^2}{n} \left( 1 + \delta \right)^2}
\]

\[
= \frac{(\delta - 1)^2}{(\delta + 1)^2}.
\]
Exercise 12 (Ratio method of estimation)

Consider a generalized form of ratio estimator \( \hat{Y}_{ra} = \frac{Y}{X^\alpha} \) for estimating the population mean \( \mu \) where \( \alpha \) is some scalar. Derive the approximate bias and mean squared error of \( \hat{Y}_{ra} \). For what value of \( \alpha \), the mean squared error is minimum. Derive the minimum mean squared error. Note that for \( \alpha = 1 \), \( \hat{Y}_{ra} \) reduces to usual ratio estimator.

**Solution:**

Define

\[
\begin{align*}
\varepsilon_0 &= \frac{X - \bar{X}}{X} \Rightarrow \bar{X} = (1 + \varepsilon_0) \bar{X} \\
\varepsilon_i &= \frac{Y - \bar{Y}}{\bar{Y}} \Rightarrow \bar{Y} = (1 + \varepsilon_i) \bar{Y} \\
E(\varepsilon_0) &= 0, \quad E(\varepsilon_i) = 0 \\
E(\varepsilon_i^2) &= \frac{f}{n} C_x^2, \quad E(\varepsilon_i^2) = \frac{f}{n} C_y^2, \quad E(\varepsilon_i \varepsilon_i) = \frac{f}{n} \rho C_x C_y
\end{align*}
\]

\( f = \frac{N-n}{N}, \quad C_x^2 = \frac{S_x^2}{X^2}, \quad C_y^2 = \frac{S_y^2}{Y^2}, \)

\( \rho \) is the population correlation coefficient between \( X \) and \( Y \).

Write

\[
\hat{Y}_{ra} = \frac{\bar{Y}}{X^\alpha} \Rightarrow \bar{Y} = (1 + \varepsilon_i) (1 + \varepsilon_0)^{-\alpha} \\
= \bar{Y}(1 + \varepsilon_i) \left( 1 - \alpha \varepsilon_0 + \frac{\alpha(\alpha+1)}{2} \varepsilon_0^2 + \ldots \right) \quad \text{(assuming \( |\varepsilon_0| < 1 \))} \\
= \bar{Y} \left( 1 + \alpha \varepsilon_i - \alpha \varepsilon_0 - \alpha \varepsilon_i \varepsilon_0 + \frac{\alpha(\alpha+1)}{2} \varepsilon_0^2 + \ldots \right) \\
\hat{Y}_{ra} - \bar{Y} \approx \bar{Y} \left( \varepsilon_1 - \alpha \varepsilon_0 - \alpha \varepsilon_i \varepsilon_0 + \frac{\alpha(\alpha+1)}{2} \varepsilon_0^2 \right) \quad \text{(upto second order of approximation).}
\]

The bias \( \hat{Y}_{ra} \) is given as
\[
\text{Bias}\left(\hat{Y}_{ra}\right) = E\left(\hat{Y}_{ra} - \bar{Y}\right)
\approx \bar{Y}E\left(e_{1} - \alpha e_{0} - \alpha e_{0}e_{1} + \frac{\alpha(\alpha+1)}{2} e_{0} \right)
\]
\[
= \bar{Y}\left[0 - \alpha \frac{f}{n} \rho C_{X} C_{Y} + \frac{\alpha(\alpha+1)}{2} \frac{f}{n} C_{X}^{2}\right]
\]
\[
= \frac{\alpha \bar{Y} C_{X}}{n} \left[\left(\frac{\alpha+1}{2}\right) C_{X} - \rho C_{Y}\right].
\]

The mean squared error of \(\hat{Y}_{ra}\) is given as

\[
\text{MSE}\left(\hat{Y}_{ra}\right) = E\left(\hat{Y}_{ra} - \bar{Y}\right)^{2}
\approx E\left(e_{1}^{2} + \alpha^{2} e_{0}^{2} - 2\alpha e_{0} e_{1}\right) \text{ (upto second order of approximation)}
\]
\[
= \frac{f}{n} \bar{Y}^{2} \left[ C_{Y}^{2} + \alpha^{2} C_{X}^{2} - 2\alpha \rho C_{X} C_{Y}\right].
\]

To obtain the value of \(\alpha\) for which the \(\text{MSE}\left(\hat{Y}_{ra}\right)\) is minimum, we use the principle of maxima/minima as follows:

\[
\frac{d\text{MSE}\left(\hat{Y}_{ra}\right)}{d\alpha} = 0 \Rightarrow 2\alpha C_{X}^{2} - 2\rho C_{X} C_{Y} = 0
\]
or
\[
\alpha = \rho \frac{C_{Y}}{C_{X}} = \alpha_{\text{min}}, \text{ say.}
\]

The second-order condition for maxima/minima is satisfied.

The minimum value of mean squared error is obtained by substituting \(\alpha = \alpha_{\text{min}}\) in the expression for \(\text{MSE}\left(\hat{Y}_{ra}\right)\) as

\[
\text{Min MSE}\left(\hat{Y}_{ra}\right) = \frac{f}{n} \bar{Y}^{2} \left[ C_{Y}^{2} + \rho^{2} \frac{C_{Y}^{2}}{C_{X}^{2}} C_{X}^{2} - 2\rho \frac{C_{Y}}{C_{X}} C_{X} C_{Y}\right]
\]
\[
= \frac{f}{n} C_{Y}^{2} \left(1 - \rho^{2}\right).
\]
Exercise 13 (Varying probability scheme):

Suppose there are 5 units in a population with values

\[ y_1 = 1, y_2 = 1, y_3 = \frac{8}{3}, y_4 = \frac{8}{3}, y_5 = \frac{2}{3}. \]

The probabilities of selecting a sample of size 2 with different units as

\[ p(y_1, y_2) = \frac{1}{2}, \quad p(y_3, y_4) = \frac{1}{6}, \quad p(y_5, y_5) = \frac{1}{6}. \]

Calculate the first and second-order inclusion probabilities and find the probability distribution of the \( \pi \)-estimator of the total.

Suppose the sample total \( \sqrt{Y_{HF}} \) is used which is based on Horvitz-Thompson estimator is used to estimate the square root of population total \( \sqrt{Y_{tot}} \). Find the probability distribution of \( \sqrt{Y_{HF}} \) and find if it is unbiased for \( \sqrt{Y_{tot}} \) or not. Calculate the variance of \( \sqrt{Y_{HF}} \).

Solution:

The inclusion probabilities are

\[ \pi_1 = \frac{1}{2}, \pi_2 = \frac{1}{2}, \pi_3 = \frac{1}{3}, \pi_4 = \frac{1}{3}, \pi_5 = \frac{1}{3}, \pi_{12} = \frac{1}{2}, \pi_{34} = \frac{1}{6}, \pi_{35} = \frac{1}{6}, \pi_{45} = \frac{1}{6}, \pi_{y} = 0 \] for all other pairs of \((i, j)\).

The Horvitz Thompson estimator of population total is

\[
\hat{Y}_\pi = \begin{cases} 
\frac{1}{(1/2)} + \frac{1}{(1/2)} = 4 & \text{with probability } \frac{1}{2} \\
\frac{8}{3} + \frac{8}{3} = 16 & \text{with probability } \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}.
\end{cases}
\]

Since the sample size is 2, the estimator of variance is

\[
\hat{\text{Var}}(\hat{Y}_\pi) = \left( \frac{y_i - y_j}{\pi_i - \pi_j} \right)^2 \frac{\pi_i \pi_j - \pi_y}{\pi_y}.
\]

Thus for the

- sample \((y_1, y_2)\), \( p_s = \frac{1}{2}, \hat{\text{Var}}(\hat{Y}_\pi) = 0. \)

- sample \((y_3, y_4)\), \( p_s = \frac{1}{6}, \hat{\text{Var}}(\hat{Y}_\pi) = 0. \)
- sample \((y_3, y_3)\), \(p_s = \frac{1}{6}\), \(\text{Var}(\hat{Y}_x) = 0\).

- sample \((y_4, y_3)\), \(p_s = \frac{1}{6}\), \(\text{Var}(\hat{Y}_x) = 0\).

Next

\[
\sqrt{Y_x} = \begin{cases} 
2 & \text{with probability } \frac{1}{2} \\
4 & \text{with probability } \frac{1}{2} 
\end{cases}
\]

Thus

\[
E(\sqrt{Y_x}) = 2 \times \frac{1}{2} + 4 \times \frac{1}{2} = 3 < \sqrt{10} = \sqrt{Y_{tot}}.
\]

So \(\sqrt{Y_x}\) is a biased estimator of \(Y_{tot}\) and it underestimates it.

The variance of \(\sqrt{Y_x}\) is

\[
\text{Var}(\sqrt{Y_x}) = E(\hat{Y}_x) - E\Big[\hat{Y}_x^2\Big]
\]

\[
= 10 - 9 = 1.
\]