SRS with Repeated Measure of a Subject

**Background and Setting: The Population**

Label the subjects in the population by $s = 1, ..., N$, and represent the response for subject $s$ as $y_s$ and let $\mathbf{y} = (y_1, y_2, \ldots, y_N)'$. Define the population mean as

$$\mu = \frac{1}{N} \sum_{s=1}^{N} y_s$$

and define the population variance as

$$\sigma^2 = \frac{1}{N-1} \mathbf{y}' \left( \mathbf{J}_N - \frac{\mathbf{J}_N}{N} \right) \mathbf{y} = \frac{1}{N-1} \sum_{s=1}^{N} (y_s - \mu)^2.$$

A model for subject $s$ is given by

$$y_s = \mu + \epsilon_s.$$  

This model is not stochastic. Recall that if we represent a random permutation of the subjects by the vector $\mathbf{Y} = (Y_j)'$, then $E(\mathbf{Y}) = \mathbf{1}_N \mu$ and $\text{var}(\mathbf{Y}) = \sigma^2 \left( \mathbf{1}_N - \frac{\mathbf{J}_N}{N} \right)$. The stochastic model for the $i^{th}$ selected subject is given by

$$Y_i = \mu + E_i.$$  

Randomly select a simple random sample of $n$ subjects from the population without replacement. Suppose that for each selection, the subject is measured $p$ times, where $j = 1, ..., p$. Let us represent response for measure $j$ on the $i^{th}$ selected subject by $Y_{ij}$. Let $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{ip})'$, and $\mathbf{Y}_j = (\mathbf{Y}_1', \mathbf{Y}_2', \ldots, \mathbf{Y}_n')'$.  

**Example**

Suppose on two occasions, 3 subjects selected via simple random sampling from a population of 5000 subjects are asked to report the number of siblings they have in their family ($N = 5000$, $n = 3$ and $p = 2$). Then $\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_{5000} \end{pmatrix}$, and $\mathbf{Y}_j = \begin{pmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \\ Y_{31} \\ Y_{32} \end{pmatrix}$.  

**Expected Value and Variance When There is No Measurement Error**

When a sample subject is measured more than once, a simple model for the \( j^{th} \) measure of the \( i^{th} \) selected subject is

\[ Y_{ij} = \mu + \epsilon_{ij}. \]

Note that for different values of \( j \), since the same question is asked twice of the same subject, when there is no measurement error, the response recorded for the \( i^{th} \) selected subject will be identical each time the subject is asked the question. In this special case, \( E_{ij} = E_{i*} \) when \( j \neq j^* \). This will occur when there is no measurement error.

When there is no measurement error, we represent \( E_{ij} = E_{i*} = E_i \). Suppose that there are \( p \) measures on each sample subject. Notice that elements in the sample vector, \( Y_i \), have expected value \( E(Y_{ij}) = \mu \) for all \( i = 1, \ldots, n \) and \( j = 1, \ldots, p \). Therefore, \( E(Y_i) = (1_n \otimes 1_p) \mu = 1_{np} \mu \).

To evaluate the variance, first consider

\[
\text{var}(Y_i) = \begin{pmatrix}
\text{var}(Y_{i1}) & \text{cov}(Y_{i1}, Y_{i2}) & \cdots & \text{cov}(Y_{i1}, Y_{ip}) \\
\text{cov}(Y_{i2}, Y_{i1}) & \text{var}(Y_{i2}) & \cdots & \text{cov}(Y_{i2}, Y_{ip}) \\
\vdots & \vdots & \ddots & \vdots \\
\text{cov}(Y_{ip}, Y_{i1}) & \text{cov}(Y_{ip}, Y_{i2}) & \cdots & \text{var}(Y_{ip})
\end{pmatrix}.
\]

Recall that there is no error in measuring the value on a subject. Thus, regardless of which occasion the subject is measured, we will get the same value. This implies that \( Y_{i1} = Y_{i2} = Y_{i3} = \ldots = Y_{ip} = Y_i \). As a result,

\[
\text{var}(Y_i) = \begin{pmatrix}
\text{var}(Y_i) & \text{cov}(Y_i, Y_i) & \cdots & \text{cov}(Y_i, Y_i) \\
\text{cov}(Y_i, Y_i) & \text{var}(Y_i) & \cdots & \text{cov}(Y_i, Y_i) \\
\vdots & \vdots & \ddots & \vdots \\
\text{cov}(Y_i, Y_i) & \text{cov}(Y_i, Y_i) & \cdots & \text{var}(Y_i)
\end{pmatrix} = \text{var}(Y_i) J_p.
\]

In a similar manner, note that \( \text{cov}(Y_{ij}, Y_{i*}) = \text{cov}(Y_{i*}, Y_{i*}) \) for all \( i \neq i^* \), and hence

\[
\text{cov}(Y_i, Y_{i*}) = \begin{pmatrix}
\text{cov}(Y_i, Y_i) & \text{cov}(Y_i, Y_i) & \cdots & \text{cov}(Y_i, Y_i) \\
\text{cov}(Y_i, Y_i) & \text{cov}(Y_i, Y_i) & \cdots & \text{cov}(Y_i, Y_i) \\
\vdots & \vdots & \ddots & \vdots \\
\text{cov}(Y_i, Y_i) & \text{cov}(Y_i, Y_i) & \cdots & \text{cov}(Y_i, Y_i)
\end{pmatrix} = \text{cov}(Y_i, Y_{i*}) J_p.
\]

Combining these two expressions,
\[
\text{var}(Y_f) = \begin{pmatrix}
\text{var}(Y_1) & \text{cov}(Y_1, Y_2) & \cdots & \text{cov}(Y_1, Y_n) \\
\text{cov}(Y_2, Y_1) & \text{var}(Y_2) & \cdots & \text{cov}(Y_2, Y_n) \\
\vdots & \vdots & \ddots & \vdots \\
\text{cov}(Y_n, Y_1) & \text{cov}(Y_1, Y_2) & \cdots & \text{var}(Y_n) 
\end{pmatrix}
\]

which simplifies to

\[
\text{var}(Y_f) = \begin{pmatrix}
\text{var}(Y_1) & \text{cov}(Y_1, Y_2) & \cdots & \text{cov}(Y_1, Y_n) \\
\text{cov}(Y_2, Y_1) & \text{var}(Y_2) & \cdots & \text{cov}(Y_2, Y_n) \\
\vdots & \vdots & \ddots & \vdots \\
\text{cov}(Y_n, Y_1) & \text{cov}(Y_1, Y_2) & \cdots & \text{var}(Y_n) 
\end{pmatrix} \otimes J_p
\]

or

\[
\text{var}(Y_f) = \text{var} \begin{pmatrix}
Y_1 \\
Y_2 \\
\vdots \\
Y_n
\end{pmatrix} \otimes J_p.
\]

Now, \[
\text{var} \begin{pmatrix}
Y_1 \\
Y_2 \\
\vdots \\
Y_n
\end{pmatrix} = \sigma^2 \left( I_n - \frac{J_n}{N} \right). \]

As a result,

\[
\text{var}(Y_f) = \sigma^2 \left( I_n - \frac{J_n}{N} \right) \otimes J_p.
\]

**Example**

Suppose on two occasions, 3 subjects selected via simple random sampling from a population of 5000 subjects are asked to report the number of siblings they have in their family (\(N = 5000\), \(n = 3\) and \(p = 2\)). Then \(y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{5000} \end{pmatrix}, \quad Y_f = \begin{pmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \\ Y_{31} \\ Y_{32} \end{pmatrix}, \)

\[
E(Y_f) = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \mu = 1_n \mu, \quad \text{and}
\]

\[
E(Y_f) = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \mu = 1_n \mu, \quad \text{and}
\]
Assumption of Large Population Size. We will often assume that the population size is very large relative to the sample size, and approximate the variance by ignoring the last term in this expression. Doing so, we represent

\[ \text{var}(Y_t) = \sigma^2 \begin{pmatrix} 111 & 0 & 0 & 0 & 0 \\ 111 & 0 & 0 & 0 & 0 \\ 000 & 111 & 0 & 0 \\ 000 & 111 & 0 & 0 \\ 000 & 000 & 111 \end{pmatrix} \otimes J_2 \]

\[ \sigma^2 \begin{pmatrix} 111 & 1 & 1 & 1 \\ 111 & 1 & 1 & 1 \\ 111 & 1 & 1 & 1 \\ 111 & 1 & 1 & 1 \\ 111 & 1 & 1 & 1 \end{pmatrix} - \frac{1}{5000} \sigma^2 \begin{pmatrix} 111 & 1 & 1 & 1 \\ 111 & 1 & 1 & 1 \\ 111 & 1 & 1 & 1 \\ 111 & 1 & 1 & 1 \\ 111 & 1 & 1 & 1 \end{pmatrix} \]

An Introduction to the SRS with Response Error Model

Background and Setting: The Population

Label the subjects in the population by \( s = 1, \ldots, N \), and represent the \( k^{th} \) response for subject \( s \) as \( Y_{sk} \). Let \( E_R(Y_{sk}) = \mu_s \) denote the expected value with respect to response error. A model for \( k^{th} \) response for subject \( s \) is given by

\[ Y_{sk} = \mu_s + E_{sk} \quad (1) \]

Let us represent the \( \text{var}_k(Y_{sk}) = \sigma^2_{sk} \). If response is measured more than once on subject \( s \), the response errors may (or may not) be independent. For example, if response corresponds to the height of a subject (recorded to the nearest 1/8 inch), the response error from different measures of height may be assumed to be independent. Other assumptions may be made in different settings. For example, if a treatment is assigned to a subject prior to obtaining a measure, then the magnitude of the variance may be affected, but response errors may still be assumed to be independent.

We may not always consider response errors to be independent. If there is an unmeasured variable that affects response, it may introduce correlation in the response errors. If this correlation decreases with time of measurement, the correlation may be
represented as an auto-correlation. We assume response errors are independent subsequently.

It remains to account for simple random sampling when response error is present. We develop this explicitly in a separate document (Argentina2006-lec2-short.doc), where all the details are presented.