Study of measurement error due to interviewer

A Dissertation Prospectus Presented

August, 2007

by

Ruitao Zhang

Approved as to style and content by:

____________________________
Edward J. Stanek III, Chair

____________________________
Graduate Program Director
William Bartock
Table of Contents

1. Introduction ............................................................................................................3
  1.1 ......................................................................................................................... Error! Bookmark not defined.
  1.2 ......................................................................................................................... Error! Bookmark not defined.
  1.3 ......................................................................................................................... Error! Bookmark not defined.
  1.4 ......................................................................................................................... Error! Bookmark not defined.
  1.5 ......................................................................................................................... Error! Bookmark not defined.

2. Literature Review .................................................................................................. Error! Bookmark not defined.

3. Examples and Preliminary Results ........................................................................... 11
  3.1 ......................................................................................................................... Error! Bookmark not defined.
  3.2 ......................................................................................................................... Error! Bookmark not defined.
  3.3 Preliminary Results.......................................................................................... Error! Bookmark not defined.
    3.3.1 Population ................................................................................................ Error! Bookmark not defined.
    3.3.2 Parameterizations ....................................................................................... Error! Bookmark not defined.
    3.3.3 ................................................................................................................... Error! Bookmark not defined.
    3.3.4 Expectation and Variance: ................................................................. Error! Bookmark not defined.
    3.3.5 ................................................................................................................... Error! Bookmark not defined.
    3.3.6 Re-arranging Terms into the Sample and Remainder Error! Bookmark not defined.
    3.3.7 ................................................................................................................... Error! Bookmark not defined.

4. Proposal .................................................................................................................. Error! Bookmark not defined.
  4.1 ......................................................................................................................... Error! Bookmark not defined.
  4.2 ......................................................................................................................... Error! Bookmark not defined.
  4.3 ......................................................................................................................... Error! Bookmark not defined.
  4.4 ......................................................................................................................... Error! Bookmark not defined.
  4.5 ......................................................................................................................... Error! Bookmark not defined.
  4.6 ......................................................................................................................... Error! Bookmark not defined.

Reference .................................................................................................................. Error! Bookmark not defined.
1. Introduction

In a survey, every operation is a potential source of measurement error. It is unavoidable that interviewers can contribute, more or less, to the total errors of the survey. For example, an interviewer’s personality, gender, race, training method can greatly influence the respondents during the interview. In order to correct or reduce them, it is necessary to understand such errors. The purpose of this research is to study how such errors are generated and affect the survey results, and how to properly account for such errors in statistical inference (such as estimating population means, domain means).

To explore this, we start from the process of how the interviewer brings the errors to the responses of subjects. Are they through labeled subjects or through sample indices? Are they through labeled interviewers or through sampled interviewer’s indices? The word ‘labeled’ means that we know who the subject is, while the ‘sample indices’ means that we don’t who the subject is, but we know the subject’s position in the sample. Knowledge about this may help us to give more accurate statistical inference for the survey data with measurement errors due to interviewers. The answer of this lies in the sources of the variability in the survey.

Usually, variability can be introduced in three different processes: sampling, interviewer allocation, and interviewing. Interviewer allocation involves how to draw the interviewers from a population of interviewers and assign them to the subjects in the sample. According to Sarndal et al. (1992), there are three ways of assigning interviewers: 1. Deterministic assignment interviewers (each interviewer is linked to a
unique area and he will conduct the interviewing in this area). 2. Random assignment interviewers (interviewers are randomly assigned to a unique area and he/she will do conduct the interviewing in the area assigned to him/her.) 3. Interpenetrating subsample (randomly partition the sample into several equal size groups and assign each interviewer to a unique group).

In this research, we first propose to clearly describe and tell the difference between several ways of interviewer assignment, using the concept of a potential observable population. This is an extension work of Xu (c07bx21). Second, we plan to use a set of random indicator variables to clearly distinguish the above three sources of variability. As a result, we will be able to distinguish whether the interviewer effect is associated with labeled subjects or sample indices, or with labeled interviewers or sampled interviewer’s indices. Third, random permutation model (a design based model) and expanded permutation model will be used to estimate the interviewer effect. Fourth, we plan to find a way to estimate domain means (such as small area means) accounting for the interviewer effect. Fifth, if time permits, we plan to represent the sampling problem via permutations, instead of using indicator random variables as above. This maybe helpful to deal with the situation that nonresponse error (missing data) and measurement error both exist in a survey.

2. Literature review

2.1. Measurement Error and Sources

In general, there are two types of errors in survey: sampling error and nonsampling error. Sampling errors arise solely as a result of drawing a probability
sample rather than conducting a complete enumeration. Nonsampling errors arise mainly due to misleading definitions and concepts, inadequate frames, unsatisfactory questionnaires, defective methods of data collection, tabulation, coding, incomplete coverage of sample units etc. Brieumer and Lyberg (2003) identify five components of nonsampling error, namely specification, frame, nonresponse, measurement and processing error. In this paper, we mainly focus on measurement error and nonresponse error.

Measurement error can be defined as the difference between the true response of a subject and the observed response through some sources of measuring (say interviewing). To be able to discuss the statistical aspects of the errors due to measurement, we need a statistical definition for the true response. By Lessler (1985), there are two approaches used to define the true response in a survey. One approach was given by Hansen et al (1951) and (1953), who assume true responses exist independently of survey conditions:

1. The true value must be uniquely defined.

2. The true value must be defined in such manner that the purposes of the survey are met. For example, in a study of school children's intelligence, we would ordinarily not define the true value as the score assigned by the child's teacher on a given date although this might be perfectly satisfactory for some studies (if, for example, our purpose was to study intelligence as measured by teacher's ratings).

3. Where it is possible to do so consistently with the first two criteria, the true value should be defined in terms of operations which can actually be carried through (even though it might be difficult or expensive to perform the operations).

The other approach was used by Zarkovich (1966). He doesn't believe that true responses exist independently of survey conditions. He defines true responses only in relation to survey conditions. We are interested in the first approach in this paper.
According to Biemer and others (1991), there are four primary sources of measurement error: questionnaire, data-collection method, respondent and interviewer. We are interested in studying the source of interviewer. The interviewer may introduce error in survey responses by not reading the items as intended, by probing inappropriately when handing an inadequate response, or by adding other information that may confuse or mislead the respondent.

2.2. Randomness due to sampling

The context we consider is for a finite population with subjects labeled by $s = 1, \ldots, N$ and a finite population interviewers labeled by $a = 1, \ldots, A$. We assume that the population is located in a finite number of geographic areas labeled by $h = 1, \ldots, H$, with $N_h$ subjects in area $h$ such that $N = \sum_{h=1}^{H} N_h$. We also assume that the subjects don’t move from one area to the other area and that $N_h \geq 1$. Suppose a survey is conducted and the data will be collected by interviewers.

Usually there are two sampling designs (sandral 1992) to select the sample, (1) \textit{SI} (simple random sampling without replacement) of $n = fN$ subjects. (2) \textit{STSI} (stratified sampling with \textit{SI} sampling in each area) of $n = fN$ subjects, but $n_h = fN_h$ subjects will be selected in area $h$. Different sampling design will introduce different randomness in the model.

2.3. Randomness due to Interviewing and the model for interviewer effect

To define the randomness due to interviewing, we can think of a ‘population’ of responses for each subject by interviewing each subject a large number of times under exactly the same conditions and then an individual response can be thought of
as a random variable drawn from the ‘population’ of this subject (Hansen et al 1951
and 1953).

In the literature of measurement error model taking interviewer into account, two
main types of statistical models for survey are found. One is the analysis-of-variance
(ANOVA) type of model used by Kish (1962) and further developed by Hartley and
Rao (1978) and others. The other type of model is the Census Bureau Model,
introduced by Hansen, Hurwitz and Bershad (1961) and extended by Fellegi (1964,
1974) and others. We will use an ANOVA type of model, and our terminology will be
based on Wolter (1985), and Särndal, Swensson and Wretman (1992).

Following Biemer and Trewin (1997) we assume that the measurement error is the
sum of two components, an “interviewer error” due to the interviewer, and a
“response error” which depends on the respondent (and possibly other remaining
sources of error). Thus, the measurement error model says that when subject \( s \) is
interviewed by interviewer \( a \), the observed value \( y_s \) can be written as

\[
y_s = \mu_s + b_a + W_s
\]

where, \( \mu_s \) is the true value, assumed to be an unknown constant
associated with respondent \( s \) . And, \( b_a \) is the interviewer error, or interviewer effect,
ascribed to interviewer \( a \) . It is assumed to be a random variable with expected value
\( B_b \) and variance \( \sigma_b^2 \) the same for all interviewers. By definition, the interviewer effect
\( b_a \) is the same for all interviews made by the same interviewer \( a \) , in the same survey.
The variance \( \sigma_a^2 \) will be called the interviewer variance. \( W_s \) is the response error,
ascribed to respondent \( s \).

There are many publications about the measurement error in covariates (errors in
variables model), but only few for the measurement error in response. In this paper, we only consider the model with measurement error in response. But the results can be extended to the model including covariates with free measurement errors or even with measurement errors.

2.4. Randomness due to Interviewer Allocation

In many surveys, interviewers are available to interview only certain classes of the population and only in certain geographic areas. In the context of section 2.2, we also assume that the total number of interviewers $A$ and areas $H$ are same and that there is 1 interviewer per area. In this condition, Carl-Erik Sarndal (1992) gave three common ways to assign interviewers to subjects: 1.) Deterministic assignment interviewers. 2.) Random assignment interviewers. 3. Interpenetrating subsample.

In some surveys (for example in face to face interviewing survey), the traveling fee may be expensive. To avoid this, each interviewer only carries out the interviews with sampled subjects from his own area $h$ where he/she lived. In this case, each interviewer $a$ is linked to a unique area $h$. This method is called deterministic assignment interviewers. In this method, there is no randomness due to interviewer allocation.

However, In the way of random assignment interviewers, randomness will be introduced in the process of assigning the interviewer to the area. In random assignment interviewers, each interviewer $a$ may be randomly assigned to a unique area. The assigned interviewer carries out all interviews with sampled elements from his own area. This method may be used in telephone survey which has no traveling
cost. The randomness in this case will be clearly represented in section 3.2.

The third way of assigning interviewer is Interpenetrating subsample. The following notation will be used. Suppose we get a random sample with size \( n \) in a specific sampling design. Let the sample be partitioned at random into \( A \) non-overlapping groups of equal size \( m = n / A \). (We assume that \( A \) is an integer.) These groups are denoted \( g = 1, ..., A \). Now, the rule is that interviewer \( a \) is to make all the interviews in group \( a \). As for the concept of interpenetration, reference is given to Mahalanobis (1946), Bailar (1983) and Wolter (1985). Interpenetrating subsample is often be used to design the survey to estimate interviewer variance. A series of references such as Felliegi(1974), Biemer(1985) and Kleffe, Prasad, and Rao(1991) have been found in the literature.

2.5. Potential Observable Outcomes and missing data

The concept of potential observable outcomes is from Little and Rubin (Little and Rubin 2000). The main idea is that each subject has one non-stochastic potential value under each treatment (interviewer in this paper). There are two assumptions for this idea. First, it assumes that there is no interference between subjects, implying that none of the potential observable outcomes for a subject is affected by the treatment assignment that any other subject receives. Second, the SUTVA assumption implies that there are no hidden versions of treatments (interviewers); no matter how a subject received a treatment (interviewer), the outcome that would be observed is the same. However, different ways of assigning interviewer may cause different potential outcomes. We will illustrate this, in the section 3. If we
have $N$ subjects and $A$ interviewers, the maximum potential observable outcomes will form an $N \times A$ array. The potential outcomes for a particular way of assigning interviewer may only form a subset of $N \times A$ array.

We use this concept of potential observable outcomes in formulating interviewer assignment. If the subject is possible to be interviewed by an interviewer, we will assume that this subject could be potentially observed under this interviewer, and represent the response for subject $s$ given interviewer $a$ by the non-stochastic value $y_{sa}$ where $s = 1, \ldots, N$ denotes the label of a subject, and $a = 1, \ldots, A$ denotes the label of a interviewer.

There is a broad literature on inference in the presence of missing data involving conditional and unconditional concepts. Little and Rubin (2002, 2nd edition) refer to an earlier paper by Rubin (1976) in reference to inference with missing data in a finite population sampling context. This second edition omits two sections from the first edition Little and Rubin (1986) that discuss randomization inference with and without missing data, but include some discussion (see Section 3.3 entitled “Weighted complete-case analysis in Little and Rubin (2002, 2nd edition) that discuss survey non response). Ed Stanek(c07ed58) start to explore the non-response error (missing data) in survey by using the concept of potential observable outcomes and the permutations of the subjects in the population. Our work will be based on this idea.

2.6. Random Permutation Model

Random permutation models are used to provide a probabilistic “link” to relate a sample to its parent population (Stanek, Singer, Lencina 2004).
models were discussed more comprehensively in the earlier work of Cassel, Sarndal and Wretman (1977) and Rao and Bellhouse (1978). This framework does not require assumptions about parametric distributions. The permuted population consists of random variables corresponding to permutations of units in the finite population, where permutations occur with equal probabilities. The population is represented as a non-stochastic vector. The randomness is thus attributable to random sampling. Under this framework, the structural relationship between the population and the permuted one can be explicitly represented using matrices. (Li 2003). The random permutation models that Cassel et al (1977) mentioned is a random permutation superpopulation model which is also discussed by Mukhopadhyay (1984), Rao (1984), and Padmawar and Mukhopadhyay (1985).

Stanek, Singer, Lencina (2004) introduced the design-based approach to inference under simple random sampling of a finite population which encompasses a simple random permutation superpopulation model. A unified approach to estimation and prediction under simple random sampling in such design-based random permutation model was discussed by (Stanek, Singer, Lencina 2004). Additionally, this method has been applied to predicting random effects from finite population clustered samples with response error (Stanek and Singer 2004) and applied to rate estimation and standardization (Li 2003).

3. Examples and Preliminary Results
3.1. Deterministic assignment interviewers (DAI)

The context we considered here is the same as in section 2.4. We use $y_{iat}$ to
represent the response for subject $t$ interviewed by interviewer $a$ where, $a = 1, 2, A$ and $t = 1, 2, \cdots, N_h$. For each subject in this method, it is known beforehand who will interview him/her. Thus, the labels of interviewers are attached to the label of subjects. By assumption in section 2.4 we know that the total number of areas and interviewers are same and that 1 interviewer per area. Without lose generality, we assume that interviewer $h$ will interview the subjects located in area $h$. We can not separate the label of interviewer and area. To be simple, we will use $y_{ht}$ to represent the response for subject $t$ interviewed by interviewer $h$ where, $h = 1, 2, A$ and $t = 1, 2, \cdots, N_h$. With these assumptions, the potential observable responses (could be potentially observed) for subject $s$ will be only in area $h$ and can only be observed by interviewer $h$. Similarly, you will not observe interviewer $h$ in other areas except area $h$ (see Table1.). We start from the case that there is no response error in $y_{ht}$. Since 

$$s = \sum_{k=1}^{h-1} N_k + t,$$

$y_{ht}$ can also be represented by $y_{st}$. 

<table>
<thead>
<tr>
<th>Interviewer (a)</th>
<th>Subjects (ht)</th>
<th>Subjects (s)</th>
<th>$y_{at}(y_s)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Areas (h)</td>
<td>Subjects (ht)</td>
<td>Subjects (s)</td>
<td>$y_{at}(y_s)$</td>
</tr>
<tr>
<td>1</td>
<td>11</td>
<td>1</td>
<td>$y_{11}(y_1)$</td>
</tr>
<tr>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td>1</td>
<td>$N_1$</td>
<td>$N_1$</td>
<td>$y_{1N_1}(y_N)$</td>
</tr>
<tr>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
</tbody>
</table>

**Table1. Potential observable responses for the survey with deterministic assignment interviewers**
3.11. Parameterization

As mentioned in section 3.1, area and interviewer are completely confounded. We can not separate area and interviewer. We use \( h \) to indicate both area \( h \) and interviewer \( h \). The average for interviewer \( h \) is defined by \( \mu_h = \frac{1}{N_h} \sum_{t=1}^{N_h} y_{ht} \). The overall mean is defined by \( \mu = \frac{1}{N} \sum_{h=1}^{d} \sum_{t=1}^{N_h} y_{ht} \), which can also be defined as \( \mu = \frac{1}{N} \sum_{s=1}^{N} y_s \).

Since each subject is only interviewed by one interviewer, the effect of interviewer is combined with the effect of subject. We can not separate the effect of interviewer and the effect of subjects. In order to estimate the interviewer effect, the only way we can do is to assume there is no area and subject effects. We define \( \alpha_h = \mu_h - \mu \) as the effect of interviewer \( h \). Let us represent the response \( y_{ht} \) as

\[
y_{ht} = \mu + \alpha_h + \varepsilon_{ht}, \quad h=1,2,\cdots, A, t=1,2,\cdots, N_h
\]

where \( \varepsilon_{ht} = y_{ht} - (\mu + \alpha_h) \).

We represent the \( N_h \times 1 \) potentially observable responses for interviewer \( h \) by \( y_h \) where \( y_h = (y_{h1}, y_{h2}, \cdots, y_{hN_h})' \), and the \( N \times 1 \) potentially observable responses as \( y = (y_1', y_1', \cdots, y_A')' \). The vector of interviewer parameters is given by \( \mu = \left( \frac{1}{N} \mathbf{1}_N y \right)' \) where \( \mathbf{1}_N \) is an \( N \times 1 \) vector with all elements equal to one. We can express \( y \) in terms of \( \mu \), \( a = (\alpha_1, \cdots, \alpha_A)' \) and \( \varepsilon = (\varepsilon_{11}, \varepsilon_{12}, \cdots, \varepsilon_{AN_A})' \).
\[ y = I_N \mu + \left( \bigoplus_{k=1}^{A} I'_{N_k} \right) \alpha + \varepsilon. \]

In this model, we have \( A \) parameters and \( N \) observations. Since \( A \leq N \), it is possible to estimate the interviewer effect \( \alpha_h \).

### 3.12. Sampling Strategy and Random Permutation

Suppose two sampling designs are used to select a random sample of size \( n \), (1) \( SI \) (simple random sampling without replacement) of \( n = fN \) subjects. (2) \( STSI \) (stratified sampling with \( SI \) sampling in each area) of \( n = fN \) subjects, but \( n_h = fN_h \) subjects will be selected in area. One way to represent the simple random sampling without replacement in deterministic assignment interviewers is to think of all possible permutations of the subjects in the population. Let us index the positions in a permutation of subjects by \( i = 1, \ldots, N \). Thus, we can represent the sampling design by sets of random variables which have stochastic properties corresponding to the permutation of subjects.

In the setting of \( N \times A \) subjects and \( N \times A \) interviewers, the maximum potential observable responses is a \( N \times A \) array, \( y_p = \left( \left( y_{i,m} \right) \right) = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1A} \\ y_{21} & y_{22} & \cdots & y_{2A} \\ \vdots & \vdots & \ddots & \vdots \\ y_{N1} & y_{N2} & \cdots & y_{NA} \end{pmatrix} \). Different interviewer assignment may cause different potential observable responses. The potential observable responses in this deterministic assignment interviewers is only part of it, which is an \( N \) vector \( y = \left( y_{11}, y_{21}, \ldots, y_{N1}, \ldots, y_{NN_A} \right)' \). Then the potential observable population can also be represented as \( \left[ \bigoplus_{k=1}^{A} \left( I_{N_k} \mid 0_{N_0(N-N_k)} \right) \right] vec\left( Y_p \right) \).

The permutation of the population can be represented by a set of random
sampling indicator variables of \( U_{is} \), \( i = 1, \ldots, N \) and \( s = 1, \ldots, N \). The random variable \( U_{is} \) takes on a value of one if unit \( s \) is assigned to position \( i \) in a permutation, or zero otherwise. The matrix of indicator random variables for subjects is given by \( U_{N \times N} = ((U_{is})) \).

As a result, the population after the random permutation is represented by \( Y = U_{N \times N}Y \). \( Y \) can also be written as \( U_{N \times N} \left[ \bigoplus_{h=1}^{d} \left( I_{N_h} \mid 0_{N_h(N-N_h)} \right) \right] \text{vec}(Y_p) \).

Without lose generality, we can choose the first \( n \) elements of \( Y \) as our sample with size \( n \), which is \( \left( I_n \mid 0_{n(N-n)} \right) U_{N \times N} \left[ \bigoplus_{h=1}^{d} \left( I_{N_h} \mid 0_{N_h(N-N_h)} \right) \right] \text{vec}(Y_p) \).

The way to represent the sampling design \( RTSI \) in deterministic assignment interviewers is different from \( SI \). In this case, we can think of permuting the units in each of the areas.

The potential observable population in this setting is same as in \( SI \), which is \( \left[ \bigoplus_{h=1}^{d} \left( I_{N_h} \mid 0_{N_h(N-N_h)} \right) \right] \text{vec}(Y_p) \). The permutation of the subjects in area \( h \) can be represented by a set of random sampling indicator variables of \( U_{ish}^{h} \), \( i = 1, \ldots, N_h \) and \( s = 1, \ldots, N_h \). The random variable \( U_{ish}^{h} \) takes on a value of one if unit \( s \) in area \( h \) is assigned to position \( i \) in a permutation, or zero otherwise. The matrix of indicator random variables for units in area \( h \) is given by \( U_{N_h \times N_h}^{h} = ((U_{ish}^{h})) \). Then the matrix of indicator random variables for units in all the areas will be \( \bigoplus_{h=1}^{d} U_{N_h \times N_h}^{h} \). As a result, the population after the random permutation is represented by...
\[ Y = \left( \bigoplus_{h=1}^{d} U_{N_h \times N_h}^h \right) \left[ \bigoplus_{h=1}^{d} \left( I_{N_h} \mid 0_{N_h \times (N - N_h)} \right) \right] \text{vec}(y_p) \]

\[ = \left[ \bigoplus_{h=1}^{d} \left( U_{N_h \times N_h}^h \mid 0_{N_h \times (N - N_h)} \right) \right] \text{vec}(y_p). \]

Without lose generality, we can choose the first \( n_h \) elements of in each area as our sample, which is \[ \bigoplus_{h=1}^{d} \left( I_{N_h} \mid 0_{N_h \times (N - N_h)} \right) \left[ \bigoplus_{h=1}^{d} \left( U_{N_h \times N_h}^h \mid 0_{N_h \times (N - N_h)} \right) \right] \text{vec}(y_p) \]

where
\[
\sum_{h=1}^{d} N_h = N \quad \text{and} \quad \sum_{h=1}^{d} n_h = n.
\]

3.2. Random assignment interviewers (RAI)

We will use the notation in section 3.1 to represent \( y_{at} \). In this method, the subject won’t know who will interview him/her until the interviewer is assigned to the subject’s area. Thus, the indices of interviewers are attached to the label of subjects. With these assumptions, the potential observable responses (the responses that can be potentially observed) for subject \( s \) can be found in all the areas (see Table2.)

**Table2. Potential observable responses for the survey with random assignment interviewers**

<table>
<thead>
<tr>
<th>Areas (h)</th>
<th>Subjects (ht)</th>
<th>Subjects (s)</th>
<th>Interviewer (a)</th>
<th>( y_{at} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11</td>
<td>1</td>
<td>( y_{11} )</td>
<td>( y_{a1} )</td>
</tr>
<tr>
<td>1</td>
<td>( N_1 )</td>
<td>( N_1 )</td>
<td>( y_{1N_1} )</td>
<td>( y_{aN_1} )</td>
</tr>
<tr>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
</tr>
<tr>
<td>1</td>
<td>( \sum_{h=1}^{d-1} N_h + 1 )</td>
<td>( \sum_{h=1}^{d-1} N_h + 1 )</td>
<td>( \sum_{h=1}^{d-1} N_h + 1 )</td>
<td>( \sum_{h=1}^{d-1} N_h + 1 )</td>
</tr>
<tr>
<td>( A )</td>
<td>( \sum_{h=1}^{d-1} N_h + 1 )</td>
<td>( \sum_{h=1}^{d-1} N_h + 1 )</td>
<td>( \sum_{h=1}^{d-1} N_h + 1 )</td>
<td>( \sum_{h=1}^{d-1} N_h + 1 )</td>
</tr>
</tbody>
</table>
### 3.21 Parameterization

We still use the concept of potential observable outcomes in formulating the interviewer assignment design. Since each interviewer can only be assigned to a unique area, we can not separate the effect of the area and the interviewer. If we assume there is no area effect, the average for interviewer $a$ is defined by $\mu_a = \frac{1}{N} \sum_{s=1}^{N} y_{as}$. The average for subject $s$ is defined by $\mu_s = \frac{1}{A} \sum_{a=1}^{A} y_{as}$. The overall mean is defined by $\mu = \frac{1}{NA} \sum_{a=1}^{A} \sum_{s=1}^{N} y_{as}$. We define $\alpha_a = \mu_a - \mu$ as the effect of interviewer $a$ and $\beta_s = \mu_s - \mu$ as the effect of subject $s$.

If we assume there is no interaction between interviewer and subject, we can represent the response $y_{as}$ as

$$y_{as} = \mu + \beta_{s(a)} + \alpha_a, s = 1, 2, ..., N, a = 1, 2, ..., A.$$

We represent the $N \times 1$ potentially observable responses for interviewer $a$ by $y_a$ where $y_a = (y_{a1} \ y_{a2} \ ... \ y_{aN})'$, and the $N \times A$ potentially observable responses as $y = (y_1 \ y_2 \ ... \ y_A)$. The vector of interviewer parameters is given by $\mu = \left( \frac{1}{N} \mathbf{1}_N' y \right)'$ where $\mathbf{1}_N$ is an $N \times 1$ vector with all elements equal to one. We can express $y$ in terms of $\mu$, $a = (\alpha_1 \ ... \ \alpha_A)'$ and $\beta = (\beta_1 \ \beta_2 \ ... \ \beta_N)'$:

$$y = \mathbf{1}_N \mathbf{1}_N' \mu + \beta \mathbf{1}_N' + \mathbf{1}_N \ a'.$$

In this model, we have $N + A - 2$ parameters and $NA$ potential observations. It is possible to estimate the interviewer effect $\alpha_a$. 

| $AN_A$ | $N$ | $y_{1N}$ | $y_{AN}$ |
3.22. Sampling Strategy and Random Permutation

The potential observable population in this setting is an $N \times A$ matrix

$$y = \left( \begin{array}{cccc}
  y_{11} & y_{21} & \cdots & y_{A1} \\
  y_{12} & y_{22} & \cdots & y_{A2} \\
  \vdots & \vdots & \ddots & \vdots \\
  y_{1N} & y_{2N} & \cdots & y_{AN}
\end{array} \right).$$

We represent the $N \times 1$ potentially observable responses for interviewer $a$ by $y_a$ where $y_a = (y_{a1} \ y_{a2} \ \cdots \ y_{aN})'$. In order to keep track of the information for area $h$, we create a series of indicator vectors

$$p_h = (p_{1h} \ p_{2h} \ \cdots \ p_{Nh})', \ h = 1, 2, ..., A.$$ 

The element $p_{hs}$ takes on a value of one if unit $s$ is in area $h$. Let $p = (p_1 \ p_2 \ \cdots \ p_A)$. Then this matrix contains all the information for the label of area. We define $z = (y \ | \ p)$ to include the information for the label of area.

One way to represent the sampling design $SI$ in random assignment interviewers is to think of all possible permutations of the subjects and all possible permutations of interviewers.

Random sampling without replacement of subjects is introduced by defining a set of random variables of $U_i$, $i = 1, ..., N$ and $s = 1, ..., N$. The permutation of units can be defined in terms of $U_i$. Explicitly, the random variable $U_i$ takes on a value of one if unit $s$ is assigned to position $i$ in a permutation, or zero otherwise. The matrix of indicator random variables for units is given by $U_{N \times N} = \left( (U_i) \right)$. Similarly, let $j = 1, ..., A$ index the position of an interviewer in a permutation of interviewers. The permutation of the interviewers can be defined in terms of the indicator random variables $V_{ja}$, where $V_{ja}$ takes on a value of one if the interviewer $a$ is assigned to position $j$ in a permutation, or zero otherwise. The matrix of indicator random
variables for interviewers is $V_{A_1 A_2} = \left( (V_{ja}) \right)$.

We use $U$ and $V$ to represent a joint permutation of subjects and interviewer. Explicitly, let $U = (U_1, U_2, \ldots, U_N)' = \left( (U_i) \right)'$ where $U'_i = (U_{i1}, U_{i2}, \ldots, U_{iN})$ is the $i^{th}$ row vector of $U$ and $V = (V_1, V_2, \ldots, V_N)' = \left( (V_j) \right)'$ where $V'_j = (V_{j1}, V_{j2}, \ldots, V_{ja})$ is the $j^{th}$ row vector of $V$. Then

$$Y_j = U'_j y V_j = \sum_{s=1}^{N} \sum_{a=1}^{A} U_{is} V_{ja} y_{as}.$$ 

Similarly, we can define

$$P_j = U'_j p V_j = \sum_{s=1}^{N} \sum_{a=1}^{A} U_{is} V_{ja} P_{sa}.$$ 

As a result, the population after the joint random permutation is represented by $Y = \left( (Y_j) \right) = U y V'$. To keep track the information of area, the population after the joint random permutation is represented by $Z = \left( (Z_j) \right) = (U y V' | U z V')$. Using the property of the permutation matrix (Stanek, Argentina2006-lec1a.doc), we get probability statements for permutation matrix $U$ and $V$. The joint distribution of $Z$ could be derived.

Without lose generality, we can choose the first $n$ elements of $Z$ as our sample, which is

$$\left[ \oplus_{k=1}^{2A} \left( I_n \mid 0_{n(N-a)} \right) \right] \text{vec}(Z).$$

The way to represent the sampling design $RTSI$ in random assignment interviewers is different from $SI$. In this case, we can think of permuting the subject in each of the areas.

The permutation of the subjects in area $a$ can be represented by a set of random sampling indicator variables of $U_{ia}$, $i = 1, \ldots, N_a$ and $s = 1, \ldots, N_a$. The random
variable $U_i^a$ takes on a value of one if unit $s$ in area $a$ is assigned to position $i$ in a permutation, or zero otherwise. The matrix of indicator random variables for units in area $a$ is given by $U^a_{N_x} = \begin{pmatrix} U_i^a \end{pmatrix}$. Then the matrix of indicator random variables for units in all the areas will be $\bigoplus_{a=1}^A U^a_{N_x}$. The permutaion of the interviewers can be defined in terms of the indicator random variables $V_j^a$, which is same as in SI. The matrix of indicator random variables for interviewers is $V_{A \times A} = \begin{pmatrix} V_j^a \end{pmatrix}$.

As a result, the population after the random permutation is represented by

$$Y = \begin{pmatrix} Y \end{pmatrix} = \left( \bigoplus_{a=1}^A U^a_{N_x} \right) yV'.$$

Without lose generality, we can choose the first $n_a$ elements of in each area as our sample, which is $\bigoplus_{a=1}^A (I_n_a | 0_{a \times (N_n-a)})$ where $\sum_{a=1}^A n_a = n$.

### 3.3. Interpenetrating subsample (IS)

As described in section 2.4, Interpenetrating subsample is different from the first two methods. In this method, the sample is randomly divided into equal size random groups. Each interviewer is then linked to a unique group not area. In this case, the labels of the interviewers are associated with the sample indices of the subjects. This procedure is practical when interviewing does not entail great travel or other costs.

The potential observable population in this setting is an $N \times A$ matrix $y = \begin{pmatrix} \begin{pmatrix} Y_{11} & Y_{21} & \cdots & Y_{A1} \\ Y_{12} & Y_{22} & \cdots & Y_{A2} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{1N} & Y_{2N} & \cdots & Y_{AN} \end{pmatrix} \end{pmatrix}$. We assume that the sample is selected by simple random sampling. Then the permutation of the population can be represented by
\[ \mathbf{Y} = \left( (Y_{ia}) \right) = \mathbf{U} \mathbf{y} \] where, \( Y_{ia} \) indicates the response of position \( i \) interviewed by the interviewer \( a \). Suppose \( M = N / A \), \( A_l = n \) and \( l \leq M \), then we can represent the sample as

\[
\bigoplus_{a=1}^{d} \left( \mathbf{I}_1 \mid 0_{(l-M\cdot A)} \right) \bigoplus_{a=1}^{d} \left( \mathbf{I}_M \mid 0_{(M(4-M))} \right) \text{vec}(\mathbf{Y}).
\]

### 3.4. Three examples

#### 3.4.1 Example for Deterministic assignment interviewers

Suppose a finite population defined by a listing of 6 subjects, indexed by \( s = 1, ..., 6 \) and \( a = 1, ..., 3 \) interviewers. The population is located in \( h = 1, ..., 3 \) areas. Each area has two subjects (see Table 3.). Interviewer 1, 2 and 3 are assigned to area 1, 2 and 3 respectively. Then the potential observable responses can be

<table>
<thead>
<tr>
<th>Areas (h)</th>
<th>Subjects (t)</th>
<th>Subjects (s)</th>
<th>Interviewer (a)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>( y_{11}(y_1) )</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2</td>
<td>( y_{12}(y_2) )</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>3</td>
<td>( y_{21}(y_3) )</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>4</td>
<td>( y_{22}(y_4) )</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>5</td>
<td>( y_{31}(y_5) )</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>6</td>
<td>( y_{32}(y_6) )</td>
</tr>
</tbody>
</table>
In this example, area and interviewer are completely confounded. We can not separate area and interviewer. The average for interviewer $h$ is defined by $\mu_h = \frac{1}{2} \sum_{t=1}^{2} y_{ht}$. The overall mean is defined by $\mu = \frac{1}{6} \sum_{h=1}^{6} \sum_{t=1}^{2} y_{ht}$, which can also be defined as $\mu = \frac{1}{6} \sum_{s=1}^{6} y_s$. Since each subject is only interviewed by one interviewer, the effect of interviewer is combined with the effect of subject. We can not separate the effect of interviewer and the effect of subjects. In order to estimate the interviewer effect, the only way we can do is to assume there is no area and subject effects. We define $\alpha_h = \mu_h - \mu$ as the effect of interviewer $h$. Let us represent the response $y_{ht}$ as

$$y_{ht} = \mu + \alpha_h + \epsilon_{ht}, \quad h = 1, 2, \ldots, 6, \quad t = 1, 2$$

where, $\epsilon_{ht} = y_{ht} - (\mu + \alpha_h)$.

We represent the $2 \times 1$ potentially observable responses for interviewer $h$ by $y_h$, where $y_h = (y_{h1} \quad y_{h2})'$, and the $6 \times 1$ potentially observable responses as $y = (y'_1 \quad y'_1 \quad \ldots \quad y'_3)'$. The vector of interviewer parameters is given by $\mu = \left(\frac{1}{6} \mathbf{1}_6' \quad y\right)'$ where $\mathbf{1}_6$ is an $6 \times 1$ vector with all elements equal to one. We can express $y$ in terms of $\mu$, $\mathbf{a} = (\alpha_1 \quad \ldots \quad \alpha_3)'$ and $\mathbf{e} = (\epsilon_{11} \quad \epsilon_{12} \quad \ldots \quad \epsilon_{32})'$:

$$y = \mathbf{1}_6' \mu + \left(\bigoplus_{h=1}^{3} \mathbf{1}'_2\right) \mathbf{a} + \mathbf{e}.$$

In this model, we have 3 parameters and 6 observations. It is possible to estimate the interviewer effect $\alpha_h$.

### 3.42 Example for Random assignment interviewers

We will use the context in 3.42. But the interviewer in this case is randomly
assigned to area. The potential observable responses (the responses that can be potentially observed) for subject $s$ can be found in all the areas (see Table 2.)

**Table 4. Potential observable responses for the survey with random assignment interviewers**

<table>
<thead>
<tr>
<th>Areas (h)</th>
<th>Subjects (t)</th>
<th>Subjects (s)</th>
<th>Interviewer (a)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td>$y_{a1}$</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>$y_{a2}$</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1</td>
<td>$y_{a3}$</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2</td>
<td>$y_{a4}$</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>3</td>
<td>$y_{a5}$</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>4</td>
<td>$y_{a6}$</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>5</td>
<td>$y_{a7}$</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>6</td>
<td>$y_{a8}$</td>
</tr>
</tbody>
</table>

If we assume there is no area effect, the average for interviewer $a$ is defined by $\mu_a = \frac{1}{6} \sum_{s=1}^{6} y_{as}$. The average for subject $s$ is defined by $\mu_s = \frac{1}{3} \sum_{a=1}^{3} y_{as}$. The overall mean is defined by $\mu = \frac{1}{32} \sum_{h=1}^{3} \sum_{s=1}^{6} y_{hs}$. We define $\alpha_a = \mu_a - \mu$ as the effect of interviewer $a$ and $\beta_s = \mu_s - \mu$ as the effect of subject $s$ as the effect of subject $s$.

If we assume there is no interaction between interviewer and subject, we can represent the response $y_{as}$ as

$$y_{as} = \mu + \beta_{s(a)} + \alpha_a, s = 1, 2, \ldots, 6, a = 1, 2, \ldots, 3.$$  

We represent the $6 \times 1$ potentially observable responses for interviewer $a$ by
\[ y_a \text{ where } y_a = \left( y_{a1} \ y_{a2} \ \cdots \ y_{a6} \right)' \], and the 6\times3 potentially observable responses as 
\[ y = \left( y_1 \ y_2 \ \cdots \ y_3 \right)'. \]
The vector of interviewer parameters is given by 
\[ \mu = \left( \frac{1}{6} 1_6'y \right)' \]
where \( 1_6 \) is an 6\times1 vector with all elements equal to one. We can express \( y \) in terms of \( \mu \), \( a = (\alpha_1 \ \cdots \ \alpha_6)' \) and \( b = (\beta_1 \ \beta_2 \ \cdots \ \beta_b)' \):
\[ y = 1_6'y_0'\mu + 1_6'b_0' + 1_6'a'. \]

In this model, we have 7 parameters and 18 potential observations. It is possible to estimate the interviewer effect \( \alpha_a \).