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 such errors, it is necessary to understand the 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 when estimating a population means, domain means).

To explore this, we start from the variability in the process of interviewing. In a sample survey, variability can be introduced in three different processes: sampling, interviewer assignment, and interviewing. Knowledge about the variability will help us to appropriately represent the problem and provide more accurate statistical inference for the survey data with measurement errors due to interviewers.

The first variability is sampling. Suppose one interviewer from a group of interviewers is interviewing some of the persons in a population. Are the persons randomly selected persons from the population or determined persons in the population? Is this interviewer a randomly selected interviewer from the interviewer group or a determined interviewer in the interviewer group?

The second variability is Interviewer assignment. The context we consider is for a finite population with subjects labeled by $s = 1, ..., N$ and finite interviewers labeled by $a = 1, ..., A$. The population is located in a finite number of geographic areas labeled by
Interviewer assignment involves how to draw the interviewers from the $A$ 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/she 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).

The third variability is interviewing. Measurement error arises in this process. There are two components in the error. The first component is the variability among interviewers. Different backgrounds and performances of the interviewers may have different effects on the respondent. Suppose a subject is interviewed several times under the same conditions. The responses of this subject may be different. This kind of variability is the second component of the variability during the interviewing.

The former results don’t clearly describe the above three variability, which we will explore in this research. We first propose to clearly describe and distinguish 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 define and use a set of random indicator variables to clearly distinguish the different sources of variability. As a result, we will be able to distinguish whether the interviewer effect is associated with selected subjects or determined subjects, or with selected interviewers or determined interviewer. Third, the 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 may explore settings where nonresponse error (missing data) and measurement error both exist in a survey.

2. Context for this Research

The context we consider is for a finite population with subjects labeled by $s = 1, \ldots, N$ and finite interviewers labeled by $a = 1, \ldots, A$. The population is located in a finite number of geographic areas labeled by $h = 1, \ldots, H$. Within area $h$, there are $N_h$ subjects labeled as $t = 1, \ldots, N_h$. We assume that the subjects don’t move from one area to the other area and that $N = \sum_{h=1}^{H} N_h$ and $N_h \geq 1$. There are two ways to represent the response of a subject in the population (see Table 1.). We use $y_{at}$ to represent the response for subject $t$ interviewed by interviewer $a$ where, $a = 1, 2, A$ and $t = 1, 2, \ldots N_h$. In the case of simple random sampling, it is convenient to use $y_s$ to represent the response for subject $s$ where, $s = 1, 2, N$.

Table 1. Different notation for the response of a subject in the context of this research.

<table>
<thead>
<tr>
<th>Areas (h)</th>
<th>Interviewers (a)</th>
<th>Subjects (s)</th>
<th>Subjects within areas (t)</th>
<th>Using $y_{at}$</th>
<th>Using $y_s$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1, \ldots, N</td>
<td>1, \ldots, N_h</td>
<td>$y_{at}$</td>
<td>$y_s$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$i$, \ldots, $N_i$</td>
<td>$i$, \ldots, $N_i$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
3. Literature review

3.1. 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 factors related to measuring (say interviewing). To be able to discuss the statistical aspects of the errors due to

\[
\begin{align*}
\sum_{i=0}^{1} N_i + 1, N_h & \quad y_{a1}, y_{aN_a} \\
\vdots & \vdots \\
N_a & y_{aN_a}
\end{align*}
\]

\[
\begin{align*}
\sum_{i=0}^{H-1} N_i + 1, N_H & \quad y_{H1}, y_{HN_H} \\
\vdots & \vdots \\
N_H & y_{HN_H}
\end{align*}
\]
measurement, we need a statistical definition for the true response. By Lessler (1985), describes two approaches to define the true response in a survey. One approach was given by Hansen et al (1951, p151), 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 to define true values 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 plan to use the first definition of true values 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 measurement error due to 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.

3.2. Randomness due to sampling

The context we consider is same as in section 2. Suppose a survey is conducted and the data will be collected by interviewers.
We will discuss two sampling designs (sandral 1992) to select the sample, (1) $SI$ (simple random sampling without replacement) of $n = fN$ subjects. (2) $STSI$ (stratified sampling with $SI$ sampling in each area) of $n = \sum_{h=1}^{H} n_h$ subjects, but $n_h$ subjects will be selected in area $h$. Different sampling design will introduce different randomness in the model.

3.3. 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 formed by interviewing each subject a large number of times under exactly the same conditions. An individual response can be thought of as a random variable drawn from the ‘population’ of possible responses of this subject (Hansen et al 1951 and 1953).

In the literature of measurement error model taking interviewer into account, there are two main types of statistical models used. 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

$$Y_{as} = \mu_{as} + b_a + W_{as}$$

where, $\mu_{as}$ 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 $\mu_b$ 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. $W_{as}$ 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 review, we only consider the model with measurement error in response. But the results can be extended to the model including covariates free of measurement errors or with measurement errors.

3.4. 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.

3.5. 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, 2\(^{nd}\) 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.

3.6. 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). Permutation 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).

4. Preliminary Results and Examples

4.1. Deterministic assignment interviewers (DAI)

We use \( y_{at} \) to represent the response for subject \( t \) interviewed by interviewer \( a \) where, \( a = 1, 2, A \) and \( t = 1, 2, \ldots, N_a \). We assume that 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 one 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 \( t = 1, 2, \ldots, N_h \). 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 \( \sum_{t \in h} N_t + t \).
Table 4.1. Potential observable responses for the survey with deterministic assignment interviewers

<table>
<thead>
<tr>
<th>Areas (h)</th>
<th>Subjects (ht)</th>
<th>Subjects (s)</th>
<th>Interviewer (a)</th>
</tr>
</thead>
<tbody>
<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 N (_1)</td>
<td>N (_1)</td>
<td></td>
<td>( y_{1N_1}(y_{N_1}) )</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

4.1.1. 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 \( \mu_h = \frac{1}{N_h} \sum_{i=1}^{N_h} y_{hi} \). The overall mean is defined by \( \mu = \frac{1}{N} \sum_{h=1}^{A} \sum_{i=1}^{N_h} y_{hi} \), 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

\[
y_{ht} = \mu + \alpha_h + \epsilon_{ht}, \quad h = 1, 2, \ldots, H, \ t = 1, 2, \ldots, N_h \quad \text{where,} \quad \epsilon_{ht} = y_{ht} - (\mu + \alpha_h) .
\]

We represent the \( N_h \times 1 \) potentially observable responses for interviewer \( h \) by \( y_h \) where \( y_h = (y_{h1} \ y_{h2} \ \ldots \ y_{hN_h})' \), and the \( N \times 1 \) potentially observable responses as \( y = (y_1' \ y_2' \ \ldots \ y_A')' \). The vector of interviewer parameters is given by \( \mu = \left( \frac{1}{N} I_N y \right)' \) where \( I_N \) is an \( N \times 1 \) vector with all elements equal to one. We can express \( y \) in terms of \( \mu \), \( a = (\alpha_1 \ \ldots \ \alpha_A)' \) and \( \epsilon = (\epsilon_{11} \ \epsilon_{12} \ \ldots \ \epsilon_{AN_A})' \):

\[
y = I_N \mu + \left( \bigoplus_{k=1}^A I_{N_k} \right) a + \epsilon .
\]

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

### 4.1.2. Sampling Strategy and Random Permutation

#### 4.1. Deterministic assignment interviewers (DAI)

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 \( n = \sum H_{nh} \) subjects, but \( n_h \) subjects will be selected in area \( h \). 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.
Different interviewer assignment may cause different potential observable responses. In the setting of $N$ subjects and $A$ interviewers, the potential observable responses in the deterministic assignment interviewers is an $N$ vector $y = \left( y_{11}, y_{12}, \ldots, y_{1N_i}, \ldots, y_{AN_i} \right)'$.

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 $U_{N \times N} = \left( (U_{is}) \right)$. As a result, the population after the random permutation is represented by $Y = U_{N \times N} y$.

Without lose generality, we can choose the first $n$ elements of $Y$ as our sample with size $n$, which is $\left( I_n \mid 0_{(n-N)} \right) U_{N \times N} y$.

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 permutation of the subjects in area $h$ can be represented by a set of random sampling indicator variables of $U_{ih}$, $i = 1, \ldots, N_h$ and $t = 1, \ldots, N_h$. The random variable $U_{ih}$ takes on a value of one if unit $t$ 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 $U_{N_h \times N_h} = \left( (U_{ih}) \right)$. Then the matrix of indicator random variables for units in all the areas will be $\left( \bigoplus_{h=1}^{A} U_{N_h \times N_h} \right)$. As a result, the population after the random permutation is represented by
\[ \mathbf{Y} = \left( \bigoplus_{h=1}^{A} \mathbf{U}^h_{N_h \times N_h} \right) \mathbf{y} \]

Without lose generality, we can choose the first \( n_h \) elements of each area as our sample, which is
\[ \sum_{h=1}^{A} \mathbf{I}_{n_h} \mid 0 \quad \mathbf{0}_{n_h \times n_h} \mathbf{U}^h_{N_h \times N_h} \mathbf{y} \quad \text{where} \quad \sum_{h=1}^{A} N_h = N \quad \text{and} \quad \sum_{h=1}^{A} n_h = n. \]

4.2. Random assignment interviewers (RAI)

4.2.2. Sampling Strategy and Random Permutation

The potential observable population in this setting is an \( A \times N \) matrix
\[ \mathbf{y} = \left( \begin{array}{c} y_{11} \\ y_{12} \\ \vdots \\ y_{1N} \\ y_{21} \\ y_{22} \\ \vdots \\ y_{2N} \\ \vdots \\ \vdots \\ y_{N1} \\ y_{N2} \\ \vdots \\ y_{AN} \end{array} \right). \]

We represent the \( N \times 1 \) potentially observable responses for interviewer \( a \) by \( y_a \) where \( y_a = (y_{a1}, y_{a2}, \ldots, y_{aN})' \). In order to keep track of the information for area \( h \), we create a series of indicator vectors \( \mathbf{P}_h = (p_{1h}, p_{2h}, \ldots, p_{Nh})' \), \( h = 1, 2, \ldots, A \). The element \( p_{hs} \) takes on a value of one if unit \( s \) is in area \( h \). Let \( \mathbf{p} = (p_1, p_2, \ldots, p_A) \). Then this matrix contains all the information for the label of area. We define \( \mathbf{z} = (\mathbf{y} \mid \mathbf{p}) \) 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 \( U_{is}, \quad i = 1, \ldots, N \) and \( s = 1, \ldots, N \). The permutation of units
can be defined in terms of \( U_s \). Explicitly, the random variable \( U_s \) 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 \( U_{N \times N} = ((U_i)) \). Similarly, let \( j = 1, \ldots, A \) 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 \( V_{A \times A} = ((V_{ja})) \).

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

\[
Y_{yj} = U_i' y V_j = \sum_{x=1}^{N} \sum_{a=1}^{A} U_{ax} V_{ja} y_{as}.
\]

Similarly, we can define

\[
P_{yj} = U_i' p V_j = \sum_{x=1}^{N} \sum_{a=1}^{A} U_{ax} V_{ja} p_{as}.
\]

As a result, the population after the joint random permutation is represented by \( Y = \left( (Y_{yj}) \right) = UyV' \). To keep track the information of area, the population after the joint random permutation is represented by \( Z = \left( (Z_{yj}) \right) = (UyV' | UzV') \). 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

\[
\left[ \bigoplus_{k=1}^{2d} \left( I_n \mid 0_{n \times (N-n)} \right) \right] \text{vec}(Z).
\]

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 \( U_a^i \), \( i = 1, \ldots, N_a \) and \( t = 1, \ldots, N_a \). The random variable \( U_a^i \) takes on a value of one if unit \( t \) 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 \( U_a^{iN_a \times N_a} = \left( \begin{array}{c} U_a^i \end{array} \right) \). Then the matrix of indicator random variables for units in all the areas will be \( \left( \bigoplus_{a=1}^{A} U_a^{aN_a \times N_a} \right) \). The permutation of the interviewers can be defined in terms of the indicator random variables \( V_{ja} \), which is same as in \( SI \). The matrix of indicator random variables for interviewers is \( V_{A \times A} = \left( \begin{array}{c} V_{ja} \end{array} \right) \).

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

\[
Y = \left( \begin{array}{c} Y_j \end{array} \right) = \left( \bigoplus_{a=1}^{A} U_a^{aN_a \times N_a} \right) yV'.
\]

Without lose generality, we can choose the first \( n_a \) elements of \( n \) in each area as our sample, which is

\[
\left( \bigoplus_{a=1}^{A} \left( I_{n_a} \mid 0_{n_a \times (N_a-n_a)} \right) \right) \text{vec}(Y) \text{ where } \sum_{a=1}^{A} n_a = n.
\]

4.3. 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

\[
y = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1N} \\ y_{21} & y_{22} & \cdots & y_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ y_{AI} & y_{A2} & \cdots & y_{AN} \end{pmatrix}.
\]

We assume that the sample is selected by simple random sampling. Then the permutation of the population can be represented by

\[
Y = \begin{pmatrix} (y_{ia}) \end{pmatrix} = U y
\]

where, \( y_{ia} \) indicates the response of position \( i \) interviewed by the interviewer \( a \). Suppose \( M = N / A \), \( AI = n \) and \( l \leq M \), then we can represent the sample as

\[
\bigoplus_{a=1}^{A} \left( I_{l} \mid 0_{l(M-n)} \right) \bigoplus_{a=1}^{A} \left( I_{M} \mid 0_{M(M-M)} \right) \text{vec}(Y).
\]

4.4. Three examples

4.4.1 Example for Deterministic assignment interviewers

Suppose a finite population defined by a listing of 6 subjects, indexed by

\( s = 1, \ldots, 6 \) and \( a = 1, \ldots, 3 \) interviewers. The population is located in \( h = 1, \ldots, 3 \) 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 4.3. Potential observable responses for the survey with deterministic assignment interviewers**

<table>
<thead>
<tr>
<th>Interviewer (a)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td></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 $\mu = \frac{1}{6} \sum_{h=1}^{3} \sum_{t=1}^{2} y_{ht}$, 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

$$y_{ht} = \mu + \alpha_h + \epsilon_{ht}, \quad h = 1, 2, \cdots, 3, t = 1, 2 \quad \text{where,} \quad \epsilon_{ht} = y_{ht} - (\mu + \alpha_h).$$

We represent the $2 \times 1$ potentially observable responses for interviewer $h$ by $y_h$ where $y_h = (y_{h1} \ y_{h2})'$, and the $6 \times 1$ potentially observable responses as $y = (y_1' \ y_1' \ \cdots \ y_3')'$. The vector of interviewer parameters is given
by \( \mu = \left( \frac{1}{6} {\mathbf{1}}_{6}^\prime, {\mathbf{y}} \right) \) where \( {\mathbf{1}}_{6} \) is an \( 6 \times 1 \) vector with all elements equal to one. We can express \( {\mathbf{y}} \) in terms of \( \mu \), \( \alpha = (\alpha_1, \ldots, \alpha_3)^\prime \) and \( \varepsilon = (\varepsilon_{11}, \varepsilon_{12}, \ldots, \varepsilon_{32})^\prime \):

\[
{\mathbf{y}} = {\mathbf{1}}_{6} \mu + \left( \sum_{k=1}^{3} k {\mathbf{1}}_{2}^\prime \right) \alpha + \varepsilon.
\]

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

**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 Table2.)

**Table 4.4. 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>1</td>
<td>1</td>
<td>1</td>
<td>( y_{a1} )</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2</td>
<td>( y_{a2} )</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>3</td>
<td>( y_{a3} )</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>4</td>
<td>( y_{a4} )</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>5</td>
<td>( y_{a5} )</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>6</td>
<td>( y_{a6} )</td>
</tr>
</tbody>
</table>

If we assume there is no area effect, the average for interviewer \( a \) is defined
by \( \mu_s = \frac{1}{6} \sum_{s=1}^{6} y_{as} \). 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 \( \mu = \frac{1}{32} \sum_{k=1}^{3} \sum_{s=1}^{6} y_{ks} \). 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

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

We represent the \( 6 \times 1 \) potentially observable responses for interviewer \( a \) by \( y_a \) where \( y_a = (y_{a1} \ y_{a2} \ \cdots \ y_{a6})' \), and the \( 6 \times 3 \) potentially observable responses as \( y = (y_1 \ y_2 \ \cdots \ y_3) \). The vector of interviewer parameters is given by \( \mu = \left( \frac{1}{6} \mathbf{1}_6' y \right)' \) where \( \mathbf{1}_6 \) is an \( 6 \times 1 \) vector with all elements equal to one. We can express \( y \) in terms of \( \mu \), \( a = (\alpha_1 \ \cdots \ \alpha_6)' \) and \( \beta = (\beta_1 \ \beta_2 \ \cdots \ \beta_6)' \):

\[
y = \mathbf{1}_6' \mu + \beta \mathbf{1}_6' + \mathbf{1}_6 \alpha'.
\]

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

5. Proposal

5.1 Describe interviewer allocation

This research will investigate the statistical inference accounting for the interviewer effect in a survey. In the context of section 2, we also assume that the total number of interviewers \( A \) and areas \( H \) are same and that there is 1 interviewer per area. Carl-Erik Sarndal (1992) gave three common ways to the subjects: 1.) Deterministic assignment interviewers. 2.) Random assignment interviewers. 3.
Interpenetrating subsample. We will clearly describe these three ways of interviewer assignment by using the concept of using the concept of a potential observable population.

5.2. Estimate interviewer effect

We will use the random permutation model (a design based model) and expanded permutation model to estimate the interviewer effect.

5.3. Estimate population and domain means

In above context, we propose to develop a best linear unbiased estimator for the population and domain means accounting for the interviewer effect. The derivation of the BLUE of the mean of the population and BLUP of the mean of a domain is similar as what Ed did in 2004.

5.4. Apply Rao-Bellhouse theorem

We plan to apply a theorem due to Rao and Bellhouse (1978) to investigate the sufficiency of the space spanned by random variables for a predictor. In this proposal we are going to use three models: the expanded random permutation model, the typical random permutation model with, and the typical random permutation model collapsed to the sample and remainder means. We propose to use the Rao Bellhouse Theorem to compare these three frameworks in order to see whether a lower dimensional framework is sufficient to develop optimal predictors of linear combinations of treatment means.

5.5. General ways of interviewer assignment

We propose to explore the general ways of assigning interviewers. In the
context of section 2, we will allow more than one interviewer per area. Then we propose to explore the problems in section 5.2 and 5.3.

5.5 Missing data and interviewer effect

If time permits, we may explore the situation that nonresponse error (missing data) and measurement error due to interviewer both exist in a survey.