Conference on Applied Statistics in Agriculture 2006 - 18th Annual Conference Proceedings

AN ESTIMATOR OF TREATMENT EFFECTS UNDER COMBINED SAMPLING AND EXPERIMENTAL DESIGNS

Christina D. Smith

An Estimator of Treatment Effects Under Combined Sampling and Experimental Designs

Christina D. Smith
Kansas State University

Abstract

Sampling design and experimental design have developed relatively independently in recent statistical history. However, many studies do involve both a sampling design and an experimental design. For example, a polluted site may be exhaustively partitioned into area plots, a random sample of plots selected, and the selected plots randomly assigned to three clean-up regimens. To date there is no commonly used procedure for incorporating both the sampling design and the experimental design into the estimation of treatment effects. For this reason we will consider an estimator of treatment effect that does incorporate both sampling and experimental designs and discuss some of it's properties.

1 Introduction

In observational studies, much attention is given to the sampling design (e.g., simple random sampling, stratified sampling, cluster sampling, etc.) so that the realized sample is as representative of the population as statistically possible. Then the selected sample is observed for certain characteristics, that is, response variables. In designed experiments, usually little or no statistical attention is given to the how units become candidates for the experiment, but much statistical attention is given to the assignment of treatments to these units (e.g., completely randomized, randomized complete block, split plot, etc.). Then certain response values are observed that are thought to be influenced by the treatments. These two types of studies have very different objectives, yet many studies involve both a sampling design and

an experimental design. For example, a polluted site may be exhaustively partitioned into area plots, a random sample of plots selected, and the selected plots randomly assigned to three clean-up regimens.

More specifically, the finite sampling approach to a statistical study is based on the idea that the population of interest is finite and fixed, with N units uniquely labeled $i = 1, 2, ..., N$. The observed response values on the units in the population, $\mathbf{y} = y_1, ..., y_N$, are fixed (non-stochastic) but unknown. Often sampling is only thought to be associated with surveys, where interest is only in describing the population. However, when the researcher wishes to compare treatments but has no control over how treatments are assigned to the members of the population, proper sampling allows the researcher to make causal conclusions from the response values. This is referred to as an analytic survey (Smith and Sugden, 1988, Thompson, 2002).

In designed experiments, the primary objective is to make comparisons between treatments for the purpose of determining causal effects. For this reason experimental researches give considerable attention to procedures for assigning treatments to experimental units so that bias is reduced and generalization to the population is reasonable. Typically, experimentalists use a stochastic approach to these statistical studies that treats the population of units as infinite and the response variables, $Y_1, Y_2, ...$, as stochastic with respect to some probability density function, such as, the normal distribution in classical ANOVA.

In both sampling and experiments, a response can only be observed for one treatment level for each unit at one time. For example, one cannot observe a response for both treatment and control on a single unit at the same time (e.g. a unit cannot be exposed to a pollutant and not exposed simultaneously). So, treatment differences cannot be identified based on observing one unit. Thus, for the set of responses actually observed, there is a corresponding set of unobservable responses from treatments that might have been applied (Smith and Sugden, 1988, Thompson, 2002). Therefore, one may prefer to study the treatment means for the units in the experiment. This will be discussed further in the next section.

Recently more interest has been generated in studying the unification of sampling and experimental design approaches to statistical studies. Depending on the type of study, the researcher may have limited control over the design aspects of the the study or may be able to completely specify the sampling and experimental designs. For example, when units are

selected by simple random sampling and treatments are assigned to the resulting sample of units using a completely randomized design, the researcher has complete control of the sampling and experimental designs, and every treatment combination has equal probability of being observed (Thompson, 2002). When only a survey is conducted, such as in an observational study, the researcher has control over the sampling design, but may not have control over the treatment assignments. Often when designed experiments are conducted, the researcher controls the treatment assignments to units, but may not be able account for how the units were selected for inclusion into the experiment, for example when a sample of convenience is used. Table 1 is a reproduction from Smith and Sugden (1988) illustrating this relationship between sampling and experiments.

The current discussion will address the issue of estimating treatment effect when units have been sampled from the population via a sampling design and treatments have been applied to those units via some experimental design. First, the basics of sampling designs and experimental design will be reviewed. Then an estimator for treatment effect under combined sampling and experimental design, proposed by Thompson (2002), will be developed. Finally, we will present some properties of Thompson's estimator.

2 The Basics

In this section we will review some basics that will be a foundations for the development of Thompson's estimator in the next section. First, a concise review of finite population sampling is given. Then the Horvitz-Thompson estimator, which gives a basis for the form of Thompson's estimator, will be reviewed and discussed. Finally, we will give a brief review of experimental design.

2.1 Review of Finite Population Sampling

A finite population, $U = \{u_1, u_2, ..., u_N\}$, is a set of units, u_i, of fixed size N, where N may or may not be known, and where each unit in U is assigned a unique label, $i = 1, 2, ..., N$. Sometimes U is simply written as $U = \{1, 2, ..., N\}$. A sample, s, of size $n(s)$ is selected from the population, where $n(s)$ may be random or fixed, $n(s) = n$. Denote the observable response measured on unit i as y_i and form the response vector $\mathbf{Y} = \{y_1, y_2, ..., y_N\}$. The

y_i's are fixed for each unit and are only observed for the units in the sample.

For N known, each sample, s, of size $n(s)$, has a known, specified probability, $p(s)$, of being selected, based on the set of samples under consideration. The function $p(\cdot)$, which gives the probability of selecting s under a given selection procedure, is called the sampling design (Särndal, Swensson and Wretmen, 1992). The sampling design and the parameter of interest will, usually, indicate an appropriate estimator to be used. The combination of the sampling design and an estimator is called a sampling strategy (Särndal, Swensson and Wretmen, 1992).

The simplest way to visualize the selection of a sample is based on a draw-sequential sampling scheme. In a draw-sequential sampling scheme each unit is drawn based on a randomized experiment (Särndal, Swensson and Wretmen, 1992). The randomized experiment is applied for $n(s)$ draws. For example, consider simple random sampling without replacement for fixed sample size, n, using a draw-sequential scheme. Each unit is selected with equal probability from the units remaining in the population after the previous selection. That is,

$$\begin{aligned} P(u_{i_1} \in s) &= \frac{1}{N} \\ P(u_{i_2} \in s) &= \frac{1}{N-1} \\ &\vdots \\ P(u_{i_n} \in s) &= \frac{1}{N-(n-1)} \end{aligned}$$

So, the sampling design under simple random sampling without replacement with fixed sample size is given by

$$p(s) = n! \prod_{j=1}^{n-1} \frac{1}{N-j} = \frac{1}{\binom{N}{n}} \, . \tag{1}$$

Typically, when each possible sample has a known probability of being selected, each unit in the population has a known probability of appearing in the selected samples (Lohr, 1999). The probability that unit u_i is in s is given by its inclusion probability,

$$P(u_i \in s) = \sum_{s \ni i} p(s) = \pi_i \, .$$

The notation here implies that the sum should be taken over all samples containing unit i. When $\pi_i > 0$ for all $i = 1, 2, .., N$, the sample is referred to as a probability sample. Probability sampling guarantees that each unit in the population has a positive chance of appearing in at least one sample and reduces the potential selection bias (Lohr, 1999).

For example, consider again the case of simple random sampling without replacement. Consider the selection of unit i into the sample. Then there are $N - 1$ remaining units from which to select in order to fill the $n - 1$ available spaces left in the sample. Thus, the probability for unit i being in sample s is given by

$$\pi_i = P(u_i \in s) = \sum_{s \ni i} p(s) = \sum_{s \ni i} \frac{1}{\binom{N}{n}} = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N} , \tag{2}$$

which is the inclusion probability for unit i. Similarly, if unit i and unit j, $i \neq j$, are selected to be in the sample there are $N - 2$ remaining units from which to select in order to fill the $n-2$ available spaces left in the sample. Thus, the joint inclusion probability for units i and j, $i \neq j$, is given by

$$\pi_{ij} = \sum_{s \ni i,j} p(s) = \frac{\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{n(n-1)}{N(N-1)} . \tag{3}$$

One way to obtain more accurate (i.e. unbiased) estimates is to utilize the inclusion probability for each unit based on the given sampling design (Särndal, Swensson and Wretmen, 1992). One family of estimators that incorporates the inclusion probabilities of units was developed by Horvitz and Thompson (1952). These estimators are based on the inverse of the inclusion probabilities so that a unit with a large inclusion probability is down weighted, whereas, a unit with a small inclusion probability is up weighted. These estimators are unbiased for all sampling designs (Hedayat and Sinha, 1991).

2.2 Review of Horvitz-Thompson Estimator

The original Horvitz-Thompson estimator (HTE) was developed for the finite population total $T(\mathbf{Y}) = \sum_{i=1}^{N} y_i$ (Horvitz and Thompson, 1952). The general form of the HTE for

$T(\mathbf{Y})$, using the notation of Hedayat and Sinha (1991), is given by

$$\text{HTE}(s, \mathbf{Y}) = \sum_{i \in s} y_i/\pi_i = \sum_{i=1}^{N} Z_i y_i/\pi_i \tag{4}$$

where Z_i is an indicator variable that is one if unit i is in the sample and zero otherwise, that is, $Z_i \sim Bernoulli(\pi_i)$. Note that π_i is known from the sampling design and that Z_i is a random variable whose value depends on the given sample, s, for each i.

It can be shown that the HTE is a homogeneous, linear unbiased estimator (Hedayat and Sinha, 1991). It is homogeneous in the sense that the inclusion probabilities used to weight the observations are dependent on the sample selected (Hedayat and Sinha, 1991). Also, the HTE is unbiased for the population total since

$$E\left[\sum_{i=1}^{N} Z_i \frac{y_i}{\pi_i}\right] = \sum_{i=1}^{N} \frac{y_i}{\pi_i} E(Z_i) = \sum_{i=1}^{N} y_i \ . \tag{5}$$

Note that in general, the HTE is not a uniform minimum variance unbiased estimator (umvue) since no such estimator exists for estimators of the form $t(s, \mathbf{Y}) = \sum_{i \in s} a_{si} Y_i$, where a_{si} depends on the sample and the unit drawn (Hedayat and Sinha, 1991). However, the HTE is the unique best estimator for unicluster sampling designs, where unicluster means that any two samples taken by the given design are either disjoint or equivalent (Hedayat and Sinha, 1991).

The variance of the HTE given by Horvitz and Thompson (1952) is, using the notation of Lohr (1999),

$$\text{V}_{\text{HT}}(\text{HTE}) = \sum_{i=1}^{N} (y_i)^2 \left(\frac{1}{\pi_i} - 1\right) + \sum_{i \neq j} y_i y_j \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1\right) \ . \tag{6}$$

$\text{V}_{\text{HT}}(\text{HTE})$ is appropriate for samples of fixed or variable size, n or $n(s)$, respectively. An alternative expression for the variance of HTE is given by

$$\text{V}_{\text{SYG}}(\text{HTE}) = \sum_{i=1}^{N} \sum_{j>i} (\pi_i \pi_j - \pi_{ij}) \left[\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right]^2 \ , \tag{7}$$

which was first supplied by Sen (1953) and by Yates and Grundy (1953) (Hedayat and Sinha, 1991). $\text{V}_{\text{SYG}}(\text{HTE})$ is only appropriate for fixed-size sampling designs.

Estimates of both $\mathrm{V_{HT}}$(HTE) and $\mathrm{V_{SYG}}$(HTE) can be derived by introducing the indicator variable and the appropriate weight to make the estimate unbiased (provided $\pi_{ij} > 0$ for all i, j), as follows:

$$\widehat{\mathrm{V}}_{\mathrm{HT}}(HTE) = \sum_{i=1}^{N} \frac{(y_i)^2}{\pi_i} \left(\frac{1}{\pi_i} - 1 \right) Z_i + \sum_{i \neq j} \frac{y_i y_j}{\pi_{ij}} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \right) Z_i Z_j \ , \tag{8}$$

and

$$\widehat{\mathrm{V}}_{\mathrm{SYG}}(HTE) = \sum_{i=1}^{N} \sum_{j>i} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 \left(\frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \right) Z_i Z_j \ . \tag{9}$$

These variance estimates can be negative for both forms of the variance. However, $\widehat{\mathrm{V}}_{\mathrm{SYG}}$(HTE) does tend to be less negative and give less variable estimates. That is, $\widehat{\mathrm{V}}_{\mathrm{SYG}}$(HTE) is more stable than $\widehat{\mathrm{V}}_{\mathrm{HT}}$(HTE). This is, partly, because $\pi_i \pi_j - \pi_{ij} < 0$ is less frequent than $\pi_{ij} - \pi_i \pi_j < 0$. In some cases a careful choice of the underlying sampling design may ensure nonnegativity. For example, a fixed-size design, with $0 < \pi_{ij} \leq \pi_i \pi_j$ and $1 \leq i \neq j \leq N$, will always generate nonnegative estimates for $\mathrm{V_{HT}}$(HTE) and $\mathrm{V_{SYG}}$(HTE) (Hedayat and Sinha, 1991).

2.3 Review of Experimental Design

The objective of a designed experiment is to study the causal effect of some set of treatments. Consider the simple case of one treatment and a control, as presented by Holland (1986). The response from unit i after a treatment or control has been applied is $Y_i(t)$ or $Y_i(c)$, respectively. $Y_i(t)$ and $Y_i(c)$ cannot both be observed on the same unit at the same time. Thus, the treatment effect of t relative to c, $Y_i(t) - Y_i(c)$, cannot be estimated by observing a single unit. The statistical solution to this problem is to consider average effect

$$E(Y(t) - Y(c)) = E(Y(t)) - E(Y(c)) \ ,$$

which can be estimated. That is, information about the treatment effect, can be gained by observing different units. Then the exact mechanism, that is, the experimental design, that selects units for exposure to t or c is very important (Holland, 1986).

In order to evaluate $E(Y_i(t)) - E(Y_i(c))$ a set of experimental units must be available that is large enough to apply each treatment more than once (except for situations when non-replication is unavoidable). An experimental unit is a unit to which one treatment is applied

independent of treatment application to other units. Ideally, experimental units are obtained from a population of units by some random mechanism, such as a random sampling design. However, such a mechanism may be too complex and require too much capital (time and money) to be practical. Thus, many times experimental units are selected by convenience. Here, the case where a true random sample is available will be considered.

Designed experiments require two layers of control, use of homogeneous groupings (if possible) and randomization of treatment assignments. Homogeneous groupings of experimental units using blocking allows experimentalists to control experimental error, the variability between units that are treated alike. Statistical analysis from designed experiments is typically accompanied by an estimate of this variability (Kuehl, 2000).

Also, the randomization of treatment applications to the experimental units is used to ensure that the probability of any particular allocation of treatments to experimental units is equal for all possible allocations within a given homogeneous group (Mead, 1988). This randomization gives the experiment a sense of validity by reducing bias that could arise from a systematic assignment of treatments to units (Kuehl, 2000), and facilitates generalization to some larger population. That is, if the experiment were repeated at some future time, it is expected to give similar results.

Randomization can be achieved either by allocating a treatment to a particular unit, or by allocating a unit to a particular treatment (Mead, 1988). The evaluation of $E(Y_i(t)) - E(Y_i(c))$ will depend on the allocation mechanism (i.e. the experimental design). The careful control of the experimental design and the lack of randomly selected experimental units make it difficult to truly identify the population of inference (Mead, 1988).

3 Thompson's Estimator

Consider the case suggested by Smith and Sugden (1998) where the researcher has control over both the sampling design and the experimental design. Thompson (2002) proposed an unbiased estimator for the difference between two treatment means based on the HTE for this case. First, it is appropriate to develop the ideas behind his estimator. Then we will derive Thompson's estimator and discuss some of it's properties.

3.1 Preliminary Ideas

Let π_i be the inclusion probability that unit i is in sample s, and $\alpha_{i_s}^k$ be the probability that treatment k is assigned to unit i, given that $u_i \in s$. For example, if the sample size, $n(s)$, under the sampling design is random, $\alpha_{i_s}^k$, may be different for unit i depending if $n(s)$ is 'large' or 'small.' Given a fixed number of units for assignment to treatment k, unit i would have a greater chance of being assigned to treatment k if $n(s)$ was relatively 'small.' If the assignment of treatment k does not depend on the selected sample, then $\alpha_{i_s}^k$ can be written as α_i^k (this is the case considered here). Note that π_i does not have to be the same for all units in the finite population, and α_i^k does not have to be the same for all treatments in the experiment.

Let y_i^k be the fixed response of unit i to treatment k. Let Z_i be an indicator variable that is one if $u_i \in s$ and zero otherwise. Let W_i^k be an indicator variable that is one if treatment k is assigned to unit i and zero otherwise. That is,

$$Z_i = \begin{cases} 1 & u_i \in s \\ 0 & \text{otherwise} \end{cases} ,$$

and

$$W_i^k = \begin{cases} 1 & \text{treatment k assigned to } u_i \\ 0 & \text{otherwise} \end{cases} .$$

Then $\pi_i = P(u_i \in s) = P(Z_i = 1)$, and $\alpha_i^k = P(\text{trt } k \text{ is assigned to unit } i) = P(W_i^k = 1)$, where $i = 1, ..., n$ and $k = 1, ..., T$.

Then, under a given sampling design,

$$Z_i \sim Bernoulli(\pi_i) ,$$

$$E(Z_i) = P(Z_i = 1) = \pi_i,$$

$$E(Z_i^2) = (Z_i = 0)^2 P(Z_i = 0) + (Z_i = 1)^2 P(Z_i = 1) = \pi_i ,$$

$$E(Z_i Z_j) = P(Z_i = 1, Z_j = 1) = \pi_{ij} ,$$

$$var(Z_i) = E(Z_i^2) - [E(Z_i)]^2 = \pi_i - \pi_i^2 = \pi_i(1 - \pi_i) ,$$

and

$$cov(Z_i, Z_j) = E(Z_i Z_j) - E(Z_i)E(Z_j) = \pi_{ij} - \pi_i \pi_j .$$

Likewise, under a given experimental design,

$$W_i^k \sim Bernoulli(\alpha_i^k),$$

$$E(W_i^k) = P(W_i^k = 1) = \alpha_i^k \ ,$$

$$E[(W_i^k)^2] = (W_i^k = 0)^2 P(W_i^k = 0) + (W_i^k = 1)^2 P(W_i^k = 1) = \alpha_i^k \ ,$$

$$E(W_i^k W_j^k) = P(W_i^k = 1, W_j^k = 1) = \alpha_{ij}^k \ ,$$

$$var(W_i^k) = E[(W_i^k)]^2 - [E(W_i^k)]^2 = \alpha_i^k - (\alpha_i^k)^2 = \alpha_i^k(1 - \alpha_i^k) \ ,$$

and

$$cov(W_i^k, W_j^k) = E(W_i^k W_j^k) - E(W_i^k)E(W_j^k) = \alpha_{ij}^k - \alpha_i^k \alpha_j^k \ ,$$

where α_{ij}^k is the joint inclusion probability of units i and j, $i \neq j$, in treatment k. For the current discussion assume that Z_i and W_i^k are independent.

To cement these ideas consider the sampling design simple random sampling without replacement (srswor) to which the experimental design, completely randomized design (CRD), is imposed on the srswor sample units. Then

$$\pi_i = P(Z_i = 1) = \frac{n}{N}$$

and

$$\pi_{ij} = P(Z_i = 1, Z_j = 1) = \frac{n(n-1)}{N(N-1)}$$

as demonstrated in section 2. Similarly,

$$\alpha_i^k = P(W_i^k = 1) = \frac{n_k}{n}$$

and

$$\alpha_{ij}^k = P(W_i^k = 1, W_j^k = 1) = \frac{n_k(n_k - 1)}{n(n-1)} \ .$$

3.2 Development of Thompson's Estimator

Thompson (2002) proposed an estimator of treatment means, μ_k, where $\mu_k = \frac{1}{N}\sum_{i=1}^N y_i^k$ is the population mean under treatment k. The conventional estimator for the average response from treatment k is

$$\bar{y}_k = \sum_{i \in s, t_i = k} \frac{y_i^k}{n_k} = \frac{1}{n_k} \sum_{i=1}^N y_i^k Z_i W_i^k \ , \tag{10}$$

where $n_k = \sum_{i=1}^{N} Z_i W_i^k$ is the number of sample units assigned to treatment k (Thompson, 2002). However, under most sampling designs, this estimator is biased, since under a given sampling design, a given experimental design, and fixed treatment group size, n_k,

$$E(\bar{y}_k) = \sum_{i=1}^{N} \frac{1}{n_k} y_i^k E(Z_i) E(W_i^k) = \sum_{i=1}^{N} \frac{1}{n_k} y_i^k \pi_i \alpha_i^k \neq \mu_k \ , \tag{11}$$

since Z_i and W_i^k are assumed to be independent. However,

$$E\left[\bar{y}_k \frac{n_k}{N\pi_i\alpha_i^k}\right] = \sum_{i=1}^{N} \frac{n_k}{N\pi_i\alpha_i^k} E(\bar{y}_k) = \mu_k \ . \tag{12}$$

Thus, an unbiased estimator of the mean population treatment effect, $\mu_k - \mu_{k'}$, is $\hat{\mu}_k - \hat{\mu}_{k'}$, where

$$\hat{\mu}_k = \frac{1}{N} \sum_{i \in s, t_i = k} \frac{y_i^k}{\pi_i \alpha_i^k} = \frac{1}{N} \sum_{i=1}^{N} \frac{y_i^k}{\pi_i \alpha_i^k} Z_i W_i^k \tag{13}$$

(Thompson, 2002). Note that Thompson (2002) did not provide variances or variance estimators for the mean estimator or the difference in means estimator.

Going back to the example of simple random sampling without replacement (srswor) with a completely randomized design (CRD),

$$\begin{aligned} \hat{\mu}_k &= \frac{1}{N} \sum_{\substack{i \in s \\ t_i = k}} \frac{y_i^k}{\pi_i \alpha_i^k} \\ &= \frac{1}{N} \sum_{\substack{i \in s \\ t_i = k}} y_i^k \left(\frac{N}{n}\right)\left(\frac{n}{n_k}\right) \\ &= \frac{1}{n_k} \sum_{\substack{i \in s \\ t_i = k}} y_i^k \ . \end{aligned} \tag{14}$$

3.3 Properties of Thompson's Estimator

As shown in the previous section $\hat{\mu}_k$ is unbiased. Similar to the HTE, $\hat{\mu}_k$ is homogeneous in the sense that the inclusion probabilities used to weight the observations are dependent on the sample selected. That is, $\hat{\mu}_k$ is a linear estimator of the form $t(s, \mathbf{Y}) = \sum_{i \in s} a_{si} y_i$ where

a_{si} does not depend on y_i. Thus, $\hat{\mu}_k$ is a homogeneous, linear unbiased estimator (HULE) of μ_k and $\hat{\mu}_k - \hat{\mu}_{k'}$ is a HULE of the difference in means (Hedayat and Sinha, 1991). Note that that no UMVUE exists for estimators of this form (Hedayat and Sinha, 1991).

The variance of $\hat{\mu}_k$ is given by

$$\begin{aligned} var(\hat{\mu}_k) &= \frac{1}{N^2}\sum_{i=1}^{N}\left(\frac{y_i^k}{\pi_i\alpha_i^k}\right)^2 var(Z_iW_i^k) + \frac{1}{N^2}\sum_{i\neq j}\left(\frac{y_i^k}{\pi_i\alpha_i^k}\right)\left(\frac{y_j^k}{\pi_j\alpha_j^k}\right) cov(Z_iW_i^k, Z_jW_j^k) \\ &= \frac{1}{N^2}\sum_{i=1}^{N}\left(\frac{y_i^k}{\pi_i\alpha_i^k}\right)^2 \pi_i\alpha_i^k(1-\pi_i\alpha_i^k) + \frac{1}{N^2}\sum_{i\neq j}\left(\frac{y_i^k}{\pi_i\alpha_i^k}\right)\left(\frac{y_j^k}{\pi_j\alpha_j^k}\right)(\pi_{ij}\alpha_{ij}^k - \pi_i\pi_j\alpha_i^k\alpha_j^k) \\ &= \frac{1}{N^2}\sum_{i=1}^{N}(y_i^k)^2\frac{\pi_i\alpha_i^k(1-\pi_i\alpha_i^k)}{(\pi_i\alpha_i^k)^2} + \frac{1}{N^2}\sum_{i\neq j}(y_i^k y_j^k)\frac{\pi_{ij}\alpha_{ij}^k - \pi_i\pi_j\alpha_i^k\alpha_j^k}{\pi_i\pi_j\alpha_i^k\alpha_j^k} \\ &= \frac{1}{N^2}\left[\sum_{i=1}^{N}(y_i^k)^2\frac{1-\pi_i\alpha_i^k}{\pi_i\alpha_i^k} + \sum_{i\neq j}(y_i^k y_j^k)\frac{\pi_{ij}\alpha_{ij}^k - \pi_i\pi_j\alpha_i^k\alpha_j^k}{\pi_i\pi_j\alpha_j^k\alpha_i^k}\right], \end{aligned} \tag{15}$$

since

$$\begin{aligned} var(Z_iW_i^k) &= E[(Z_iW_i^k)^2] - [E(Z_iW_i^k)]^2 \\ &= E[(Z_i)^2]E[(W_i^k)^2] - [E(Z_i)]^2[E(W_i^k)]^2 \\ &= \pi_i\alpha_i^k - (\pi_i)^2(\alpha_i^k)^2 \\ &= \pi_i\alpha_i^k(1-\pi_i\alpha_i^k), \end{aligned} \tag{16}$$

and

$$\begin{aligned} cov(Z_iW_i^k, Z_jW_j^k) &= E(Z_iZ_jW_i^kW_j^k) - E(Z_iW_i^k)E(Z_iW_j^k) \\ &= E(Z_iZ_j)E(W_i^kW_j^k) - E(Z_i)E(Z_j)E(W_i^k)E(W_j^k) \\ &= \pi_{ij}\alpha_{ij}^k - \pi_i\pi_j\alpha_i\alpha_j. \end{aligned} \tag{17}$$

As with the HTE, $var(\hat{\mu}_k)$ can be rewritten in the Sen, Yates and Grundy form, for fixed sample and treatment sizes, as follows

$$var_{alt}(\hat{\mu}_k) = \sum_{i=1}^{N}\sum_{j>i}(\pi_i\pi_j\alpha_i^k\alpha_j^k - \pi_{ij}\alpha_{ij}^k)\left[\frac{y_i}{\pi_i\alpha_i} - \frac{y_j}{\pi_j\alpha_j}\right]^2. \tag{18}$$

An estimator of $var(\hat{\mu}_k)$ is given by

$$\begin{aligned}
\widehat{var}(\hat{\mu}_k) &= \frac{1}{N^2}\sum_{i=1}^{N}\frac{(y_i^k)^2}{\pi_i\alpha_i^k}\left(\frac{1-\pi_i\alpha_i^k}{\pi_i\alpha_i^k}\right)Z_iW_i^k \\
&\quad +\frac{1}{N^2}\sum_{i\neq j}\frac{(y_i^k y_j^k)}{\pi_{ij}\alpha_{ij}^k}\left(\frac{\pi_{ij}\alpha_{ij}^k-\pi_i\pi_j\alpha_i^k\alpha_j^k}{\pi_i\pi_j\alpha_j^k\alpha_i^k}\right)Z_iZ_jW_i^kW_j^k \\
&= \frac{1}{N^2}\sum_{i\in s}\frac{(y_i^k)^2}{\pi_i\alpha_i^k}\left(\frac{1-\pi_i\alpha_i^k}{\pi_i\alpha_i^k}\right) \\
&\quad +\frac{1}{N^2}\sum_{\substack{i,\,j\,\in\, s\\ i\,\neq\, j}}\frac{(y_i^k y_j^k)}{\pi_{ij}\alpha_{ij}^k}\left(\frac{\pi_{ij}\alpha_{ij}^k-\pi_i\pi_j\alpha_i^k\alpha_j^k}{\pi_i\pi_j\alpha_j^k\alpha_i^k}\right)\,, \qquad (19)
\end{aligned}$$

which is unbiased, since

$$\begin{aligned}
E\left[\widehat{var}(\hat{\mu}_k)\right] &= \frac{1}{N^2}\sum_{i=1}^{N}\frac{(y_i^k)^2}{\pi_i\alpha_i^k}\left(\frac{1-\pi_i\alpha_i^k}{\pi_i\alpha_i^k}\right)E(Z_i)E(W_i^k) \\
&\quad +\frac{1}{N^2}\sum_{i\neq j}\frac{(y_i^k y_j^k)}{\pi_{ij}\alpha_{ij}^k}\left(\frac{\pi_{ij}\alpha_{ij}^k-\pi_i\pi_j\alpha_i^k\alpha_j^k}{\pi_i\pi_j\alpha_j^k\alpha_i^k}\right)E(Z_iZ_j)E(W_i^kW_j^k) \\
&= \frac{1}{N^2}\sum_{i=1}^{N}\frac{(y_i^k)^2}{\pi_i\alpha_i^k}\left(\frac{1-\pi_i\alpha_i^k}{\pi_i\alpha_i^k}\right)\pi_i\alpha_i^k \\
&\quad +\frac{1}{N^2}\sum_{i\neq j}\frac{(y_i^k y_j^k)}{\pi_{ij}\alpha_{ij}^k}\left(\frac{\pi_{ij}\alpha_{ij}^k-\pi_i\pi_j\alpha_i^k\alpha_j^k}{\pi_i\pi_j\alpha_j^k\alpha_i^k}\right)\pi_{ij}\alpha_{ij}^k \\
&= \frac{1}{N^2}\left[\sum_{i=1}^{N}(y_i^k)^2\frac{1-\pi_i\alpha_i^k}{\pi_i\alpha_i^k}+\sum_{i\neq j}(y_i^k y_j^k)\frac{\pi_{ij}\alpha_{ij}^k-\pi_i\pi_j\alpha_i^k\alpha_j^k}{\pi_i\pi_j\alpha_j^k\alpha_i^k}\right]\,. \qquad (20)
\end{aligned}$$

The estimator $\widehat{var}(\hat{\mu}_k)$ is a homogeneous quadratic unbiased estimator (HUQE). That is, the estimator is of the form $t(s,\mathbf{Y}) = \sum_{i\in s} a_{si}y_i^2 + \sum_{i\neq j} a_{sij}y_i$ where the coefficients, a_{si} and a_{sij}, depend on the sample drawn but do not depend on y_i (Hedayat and Sinha, 1991). Another look at the example of simple random sampling without replacement (srswor) with a completely randomized design (CRD), gives

$$var(\hat{\mu}_k) = \frac{1}{N^2}\sum_{i=1}^{N}(y_i^k)^2\frac{N-n_k}{n_k}+\frac{1}{N^2}\sum_{i\neq j}(y_i^k y_j^k)\frac{n_k-N}{n_k(N-1)}\,, \qquad (21)$$

and

$$\widehat{var}(\hat{\mu}_k) = \frac{1}{N}\sum_{i=1}^{N}(y_i^k)^2\frac{N-n_k}{(n_k)^2}Z_iW_i^k+\frac{1}{N}\sum_{i\neq j}(y_i^k y_j^k)\frac{n_k-N}{(n_k)^2(n_k-1)}Z_iZ_jW_i^kW_j^k\,. \qquad (22)$$

As mentioned in section 2.2 estimates of the variance of the HTE can be negative. The same problem exists for $\widehat{var}(\hat{\mu}_k)$. That is, $\widehat{var}(\hat{\mu}_k)$ can be negative, in particular, when $\pi_{ij}\alpha_{ij}^k - \pi_i\pi_j\alpha_i^k\alpha_j^k < 0$. As in the case of the HTE, the alternative form of the variance estimate is less likely to be negative if $\pi_i\pi_j\alpha_i^k\alpha_j^k - \pi_{ij}\alpha_{ij}^k > 0$. For the current example of simple random sampling without replacement (srswor) with a completely randomized design (CRD)

$$\pi_i\pi_j\alpha_i^k\alpha_j^k - \pi_{ij}\alpha_{ij}^k = \frac{n_k}{N}\left(\frac{N-n_k}{N(N-1)}\right) > 0 \ ,$$

since $N > n_k$.

4 Summary

This manuscript is an introduction to estimation of treatment effect under combined sampling and experimental design. Thompson's estimator is one estimator that accomplishes this objective. There is considerably more work that needs to be done to develop these ideas and in particular Thompson's estimator. Here we have derived variance estimation for Thompson's estimator but these variance estimates can be negative under some designs. Work still needs to be done for estimation of treatment difference, interval estimation and hypothesis testing methodology. Also, inclusion probabilities for more complicated sampling designs (cluster, multistage, adaptive) and assignment probabilities for more complicated experimental designs (Latin square, split-plot) still need to be pursued. If this methodology is ever going to be implemented software for calculating Thompson's estimator needs to be developed. Thus, this discussion has just begun but hopefully much more discussion will ensue.

References

Hedayat, A.S. and Sinha, B.K. (1991). *Design and Inference in Finite Population Sampling.* Wiley-Interscience: New York, NY.

Holland, P.W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, **81**, 945-960.

Horvitz, D.G. and Thompson, D.J. (1952). A Generalization of Sampling Without Replacement From a Finite Universe. *Journal of the American Statistical Association*, **47**, 663-685.

Kuehl, R.O. (2000). *Design of Experiments: Statistical Principles of Research Design and Analysis*. Duxbury Press: Pacific Grove, CA (page 21).

Lohr, S.L. (1999). *Sampling: Design and Analysis*. Duxbury Press: Pacific Grove, CA (pages 25,26)

Mead, R. (1988). *The Design of Experiments: Statistical Principles for Practical Application*. Cambridge University Press: Cambridge, MA (pages 214-217).

Särndal, C., Swensson, B., and Wertman, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag: New York, NY (pages 25-30)

Sen, A.R. (1953). On the Estimate of the Variance in Sampling with Varying Probabilities. *Journal of the Indian Society of Agricultural Statistics*, **5**, 119-127.

Smith, T.M.F. and Sugden, R.A (1988). Sampling and Assignment Mechanisms in Experiments, Surveys and Observational Studies. *International Statistical Review*, **56**, 165-180.

Thompson, S.K. (2002). On Sampling and Experiments. *Envirometrics*, **13**, 429-436.

Yates, F. and Grundy, P. M. (1953). Selection without Replacement from within Strata with Probability Proportional to Size. *Journal of the Royal Statistical Society Series B*, **15**, 253-261.

Table 1: Classification of Studies with Treatments

		Treatment Assignment	
		Control	No Control
Sample Selection	Control	Experiment within a survey, or survey within an experiment	Analytic survey
	No Control	Experiment	Uncontrolled observational study