Estimating the Population Mean from a Simple Random Sample

Ed Stanek

1. Introduction

A basic problem in statistics is how to estimate the mean of a population of subjects based on the values observed for subjects selected in a simple random sample. It has been long recognized that the sample mean is the best estimate. Although the solution to the problem may not be in doubt, the fundamentals of statistics are revealed in developing the solution. Understanding these fundamentals is important when developing solutions to more complicated problems, such as estimation of random effects in mixed models. By focusing on the simpler problem, the fundamentals of statistical inference can be made clear.

Many accounts have been given of the development of inference in this simple problem, ranging in content from theoretical papers (Godambe 1955; Rao and Bellhouse 1978) to applied texts. The development presented here differs in two respects from other presentations. First, we attempt to make the level of presentation broadly accessible, while maintaining the rigor of the development. We assume the reader has basic familiarity with statistics, although not that the reader is a professional statistician. With this purpose, we sacrifice conciseness for clarity and readability. Second, we expand the usual development by identifying and describing basic random variables in the construction. Our purpose is not to integrate this presentation (via comparisons and contrast) into the literature, but rather to make the development self contained and clear. Even so, we recognize that such an integration is valuable.

In developing a solution, we highlight basic definitions, notation, assumptions, and decisions that form the basis of the solution. We limit discussion to an 'idealized' setting where subjects in the population can be enumerated (and labeled), and response for a given subject corresponds to a constant non-stochastic number. We also limit discussion to a setting where the sample of subjects can be assumed to have resulted from simple random sampling without replacement.

We begin the development with basic definitions of notation that will be used for subject's response, and parameters for the population. We note that parameters may be defined in different ways (corresponding to different parameterizations). Next, we define notation that will enable us to represent sampling. This process consists of defining random variables, a subset of which will be realized upon selection and observation of subjects in a sample. The full set of random variables can be represented in a vector. We define properties of the vector (such as the expected value and variance). We also represent a stochastic model for the vector that includes the population mean.

The estimator of the population mean is developed that satisfies familiar statistical properties. We require the estimator to be unbiased. We also require the estimator to be linear in the sample data (so that it can be evaluated) and have minimum variance among such estimators. Although such qualifications appear to make the problem trivial, we show how it leads to intriguing results.

2. Representations of the Finite Population and Sampling
We assume that the finite population consists of $s = 1, \ldots, N$ distinct subjects. These subjects are listed in the population, and a distinct label $s$ is associated with each subject. Associated with each subject is a non-stochastic response which we represent by for subject $s$ by the value $y_s$. The subject's response is a parameter for the subject. Usually in statistics, parameters are denoted using Greek characters. We follow this convention when defining combinations of the basic subject parameters (such as the mean, or variance). Our representation of the parameter for a subject as $y_s$ is consistent with notation that represents a realization of a random variable as a lower-case Latin constant.

2.1. Population Parameters.

We define the population mean as $\mu = \frac{1}{N} \sum_{s=1}^{N} y_s$ and the population variance as

$$\sigma^2 = \frac{N-1}{N} \left( \frac{1}{N} \sum_{s=1}^{N} (y_s - \mu)^2 \right) \frac{N}{N-1}.$$  The definition of the variance is the usual finite population definition. The definition of $\sigma^2$ corresponds to the survey sampling definition of the variance parameter (see (Cochran 1977)), and is commonly used to represent the variance in experimental design. We use this representation of the variance since its use simplifies other expressions.

We can represent subject response via the population mean, plus a deviation from that mean. Such a representation is in the form of a deterministic model, i.e.,

$$y_s = \mu + e_s.$$  This representation is a re-parameterization of the subject parameters. The $N$ population parameters in the sequence, $\left\{ y_s : s = 1, \ldots, N \right\}$ can be equivalently represented as $\left\{ \mu, y_s : s \neq s^*, s = 1, \ldots, N \right\}$ for any $s^* \in \{1, \ldots, N\}$.

2.2. Vector and Matrix Representation.

Vector and matrix notation can be useful to summarize the model and population parameters. Such notation is essential when problems get more complex. We use lower case bold Latin letters to represent column vectors, and bold capital Latin letters to represent matrices. The identity matrix of dimension $N \times N$ is represented by $I_N$, an $N \times N$ matrix of ones is represented by $J_N$, an $N \times 1$ column vector is represented by $1_N$, and a diagonal matrix with diagonal elements corresponding to the elements of $\mathbf{v}$ is represented by $D_v$.

Using this notation, we represent the subject specific population parameters as $y' = (y_1, y_2, \ldots, y_N)$. The population mean is given $\mu = \frac{1}{N} 1_N y'$, while the population
variance is given by \( \sigma^2 = \frac{1}{N-1} y^T \left( I_N - \frac{1}{N} J_N \right) y \). The deterministic model is represented by
\[
y = I_N \mu + e \quad \text{where} \quad e = \left( I_N - \frac{1}{N} J_N \right) y.
\]

2.3. Sampling and Creation of Random Variables

Suppose that a simple random sample of \( n \) subjects is selected from the population. We assume that the sample is selected without replacement, so that there are \( n \) different subjects in the sample. The sample may be selected using a computer algorithm using random numbers, drawing a set of labels from a hat, or in any other manner such that each possible sample has an equal chance of occurring.

One way to consider possible samples as arising is to think of all possible re-arrangements (or permutations) of the subjects in the population. Let \( i = 1, ..., N \) index the position of a subject in a permutation. In a given permutation, there will be a different subject, \( s \), at each position, \( i \). The subjects at the first \( i = 1, ..., n \) positions in a permutation are a simple random sample of subjects. This way of representing a sample is useful for defining sampling random variables.

We begin by defining random variables that enable us to represent which subject is assigned to a particular position, say position \( i \), in a permutation. Although any subject can occur at position \( i \), only one subject will occupy the position. We represent the subject that occurs by defining a set of indicator random variables, \( U_{is} \) for \( s = 1, ..., N \), where each indicator random variable can have a value of one if subject \( s \) is assigned to position \( i \) in a permutation, or zero otherwise. For any permutation, only one of the indicator random variables at position \( i \) will have a value of one. The label for the subject assigned to position \( i \) in a permutation is represented by \( \sum_{s=1}^{N} sU_{is} \). More importantly, the value of the response for the subject assigned to position \( i \) in a permutation is represented by \( Y_i = \sum_{s=1}^{N} U_{is} y_s \). The basic random variables in \( Y_i \) are the indicator random variables used to represent sampling, not the response for subject \( s \). The different definitions of \( y_s \) and \( Y_i \) are important for understanding statistical inference.

It is valuable to use a model to represent the relationship between the random variables, \( Y_i \) for \( i = 1, ..., n \), and the parameter of interest \( \mu \). The model is given by
\[
Y_i = \mu + E_i
\]
where \( E_i \) is a random variable that represents the deviation of response from the population mean for the subject assigned to position \( i \) in a permutation.

The properties of the random variables can be determined by evaluating the properties of the elementary indicator random variables. First, note that each indicator random variable \( U_{is} \) is defined to have value one with probability \( \frac{1}{N} \), and zero otherwise. Thus, \( E(U_{is}) = \frac{1}{N} \).
for all $i = 1, \ldots, N; s = 1, \ldots, N$ and hence $E(Y_i) = \sum_{s=1}^{N} E(U_{is})y_s = \mu$ for all $i = 1, \ldots, N$. This result can be summarized equivalently as $E(E_i) = 0$ for all $i = 1, \ldots, N$. We can evaluate the variance of $Y_i$ in a similar manner by working directly with the indicator random variables. However, since the random variables are not independent, it is simpler to use a matrix representation of the problem to develop expressions for the variance.

2.4. Vector and Matrix Representations of Random Variables

We represent the random variables in a random permutation as the vector $Y' = (Y_1, Y_2, \ldots, Y_N)$. This vector is similar to the representation of the vector of finite population values, $y$. In fact, the two vectors are related by the simple expression,

$$
\begin{pmatrix}
Y_1 \\
Y_2 \\
\vdots \\
Y_N
\end{pmatrix}
= 
\begin{pmatrix}
U_{11} & U_{12} & \cdots & U_{1N} \\
U_{21} & U_{22} & \cdots & U_{2N} \\
\vdots & \vdots & \ddots & \vdots \\
U_{N1} & U_{N2} & \cdots & U_{NN}
\end{pmatrix}
\begin{pmatrix}
y_1 \\
y_2 \\
\vdots \\
y_N
\end{pmatrix},
$$

or $Y = Uy$ where $U = \begin{pmatrix}
U_{11} & U_{12} & \cdots & U_{1N} \\
U_{21} & U_{22} & \cdots & U_{2N} \\
\vdots & \vdots & \ddots & \vdots \\
U_{N1} & U_{N2} & \cdots & U_{NN}
\end{pmatrix}$. Without different letters to represent the subscripts, it is easy to confuse the interpretation of $Y_i$ and $y_s$.

Using the $E(U_{is}) = \frac{1}{N}$ for all $i = 1, \ldots, N; s = 1, \ldots, N$, we can express the expected value of $Y$ compactly as $E(Y) = E(U)y = \frac{1}{N}Jy = \mu 1_N$. When evaluating the variance, we need to account for the variance and covariance of the elements in $U$. We can do this simply by evaluating the variance of a concatenation of the columns of $U$ given by $vec(U) = (U_{11} \ U_{21} \ \cdots \ \ U_{N1} \ U_{12} \ \cdots \ \ U_{N2} \ | \ \cdots \ | \ U_{1N} \ U_{2N} \ \cdots \ \ U_{NN})'$. Since

$$E(U_{is}U_{is}^*) = \begin{cases}
\frac{1}{N} & \text{when } i = i^* \text{ and } s = s^* \\
\frac{1}{N(N-1)} & \text{when } i \neq i^* \text{ and } s \neq s^* \\
0 & \text{otherwise}
\end{cases}$$

arranging terms in order, $\text{var}[vec(U)] = \frac{1}{N-1} \left(I_N - \frac{J}{N}\right) \otimes \left(I_N - \frac{J}{N}\right)$. In this representation, the symbol $\otimes$ represents the Kronecker product operator. This operator is
defined such that $A \otimes B$ consists of a matrix where each element in $A = (a_{ij})$ is multiplied by the matrix $B$ (Graybill 1983).

This representation of the variance is simple and compact. In order to use it to express the variance of $Y$, we represent $Y = (y' \otimes I_N) \text{vec}(U)$. Since $(y' \otimes I_N)$ is a constant matrix, $\text{var}(Y) = (y' \otimes I_N) \text{var}[\text{vec}(U)](y \otimes I_N)$, or

$$\text{var}(Y) = \frac{1}{N-1} y' \left( I_N - \frac{J}{N} \right) y \otimes \left( I_N - \frac{J}{N} \right).$$

Using the definition of $\sigma^2$,

$$\text{var}(Y) = \sigma^2 \left( I_N - \frac{J}{N} \right).$$

3. Estimating the Population Mean

The basic problem is how to estimate the population mean based on observations of the values of the random variables in the sample. There are many different estimators that may be considered. We limit the estimator in three ways. First, we require the estimator to be a linear function of the sample values. These sample values are realizations of the random variables in the sample. Second, we require the estimator to be unbiased. Third, we require the estimator to have minimum variance in the class of unbiased estimators.

3.1. Representing the Problem as Prediction

Commonly, when the population mean is the parameter of interest, we call the sample mean an estimate of the population mean. Looking at the problem in a slightly different way, we can consider the problem to be one of prediction. In order to consider the problem as one of prediction, we need to consider the entire population of random variables, $Y$. Any realization of these random variables will result in a vector that includes all values in the population. This fact implies that we can represent the population mean as a simple linear combination of the random variables in the population. Hence, $\mu = \frac{1}{N} I_N' Y$. Upon selecting a sample of size $n$, the values for a portion of $Y' = \left( Y_1', Y_2' \right)$ (i.e. namely those in $Y_1$) will be known. Representing the realized values in $Y_1$ as $y_1$, we find that

$$\mu = \frac{1}{N} \left( I_n' y_1 + I_{N-n}' Y_2 \right).$$

Estimation of $\mu$ requires to prediction of $Y_2$. We can describe the result as an estimator, since it represents our guess of a defined population quantity. We can also describe the result as a predictor, since it involves prediction of unknown quantities.

3.2. Deriving the Best Linear Unbiased Estimator
We use the three criteria for developing the estimator making use of the prediction framework. First, we represent the estimator as a linear function of the sample data, \( p = L'Y_i \), where \( L \) is a vector of unknown constants. Next, we require the estimator to be unbiased. This implies that \( E(L'Y_i - \mu) = 0 \). We simplify this expression by substituting the expression for \( \mu = \frac{1}{N}(1_n'Y_i + 1_{N-n}'Y) \) and using the fact that \( E(Y_i) = \mu \) resulting in
\[
L'1_n\mu - \frac{1}{N}(1_n'1_n\mu + 1_{N-n}'1_{N-n}\mu) = 0.
\]
Combining terms, the unbiasedness requirement implies that \((L'1_n - 1)\mu = 0\). In order for this constraint to hold for all possible values of \( \mu \), we require that \( 1_n'L = 1 \).

The linear constraint on the coefficients results in a single equation with \( n \) unknowns. Although there is no unique solution to the equation, we can specify the set of all possible solutions to the coefficients \( L \) in terms of a particular generalized inverse and an arbitrary vector. Taking the generalized inverse of \((1_n') = \frac{1}{n} 1_n \), this solution is given by
\[
L = \frac{1}{n} 1_n + \left(1_n - \frac{J_n}{n}\right)z
\]
where \( z \) is an arbitrary \( n \times 1 \) vector. For ease of exposition, we define
\[
a = \frac{1}{n} 1_n \quad \text{and} \quad B = \left(1_n - \frac{J_n}{n}\right) \quad \text{so that} \quad L = a + Bz.
\]
Although the coefficients are not yet uniquely specified, we have yet to impose the minimum variance criteria. This is the final step in the derivation.

The variance of the estimator is given by \( \text{var}(p - \mu) = \text{var}(p) \) since the estimator is unbiased. Since \( p = L'Y_i \) and \( \text{var}(Y_i) = \sigma^2 \left(1_n - \frac{J_n}{N}\right) \) as a result of partitioning the variance matrix for \( Y \), we can express \( \text{var}(p) = \sigma^2 L' \left(1_n - \frac{J_n}{N}\right) L \). Substituting the expression for \( L \) in terms of \( z \) result in an expression for the variance in terms of \( z \) that simplifies to
\[
\text{var}(p) = \sigma^2 \left[z'B' \left(1_n - \frac{J_n}{N}\right) Bz + 2z'B' \left(1_n - \frac{J_n}{N}\right) a + a'(1_n - \frac{J_n}{N}) a \right].
\]
We determine the values of \( z \) that minimize the variance by differentiating with respect to \( z \), and then setting the resulting derivative to zero. Taking this derivative, the estimating equations are given by
\[
B' \left(1_n - \frac{J_n}{N}\right) Bz = B' \left(1_n - \frac{J_n}{N}\right) a.
\]
The estimating equations simplify by noting that \( B' \left(1_n - \frac{J_n}{N}\right) B = \left(1_n - \frac{J_n}{n}\right) \) and
\[
B' \left(1_n - \frac{J_n}{N}\right) a = 0 \quad \text{as a result, the estimating equations simplify to} \quad \left(1_n - \frac{J_n}{n}\right) \hat{z} = 0.
\]
The form of the estimating equations is of particular interest relative to the expression for 
\[ L = \frac{1}{n} I_n + \left( I_n - \frac{J_n}{n} \right) z \]. Replacing \( z \) by any estimate \( \hat{z} \) of \( z \) that minimize the variance, 
\[ \hat{L} = \frac{1}{n} I_n + \left( I_n - \frac{J_n}{n} \right) \hat{z} \]. Then, since all minimum variance estimates of \( z \) must satisfy 
\[ \left( I_n - \frac{J_n}{n} \right) \hat{z} = 0 \], the estimate of \( L \) simplifies to \[ \hat{L} = \frac{1}{n} I_n \]. As a result, the unique minimum variance unbiased estimate of the population mean, \( \mu \), is given by the sample mean, 
\[ \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} Y_i = \bar{y} \]. Using this expression for \( \hat{L} \), 
\[ \text{var} \left( \hat{\mu} \right) = \frac{\sigma^2}{n^2} I_n \left( I_n - \frac{J_n}{N} \right) I_n = \left( 1 - \frac{n}{N} \right) \frac{\sigma^2}{n} \].
References