Examples and other Ideas Related to Samples and Positions With Discussion of Basu(1971)

Ed Stanek

Introduction

We present several examples from the literature related to sampling and inference. The examples were given as part of an active struggle with basic ideas of inference in the 1960’s and 1970’s. These examples are worth re-visiting now in light of the changes in focus in statistics, particularly with respect to mixed models, superpopulations, and Bayesian inference. Our principle focus in these examples is the representation of positions and units, particularly as it may differ between the response variables and a covariable.

Example 1. Simple Random Sampling with \( n = 1 \). (Basu 1971, p210).

“Consider the case of a simple random sample of size one from the population \( \varphi = (1, 2, ..., N) \). The sample is \((u, y)\) where \( u \) has a uniform probability distribution over the \( N \) integers \( 1, 2, ..., N \) and \( y = Y_u \). Let the population mean

\[
\bar{Y} = \frac{1}{N}(Y_1 + Y_2 + ... + Y_N)
\]

be the parameter to be estimated. Clearly, \( y \) is an unbiased estimator of \( \bar{Y} \). Let

\[
\theta_0 = (a_1, a_2, ..., a_N)
\]

and let \( \bar{a} = \frac{1}{N} \sum_{j=1}^{N} a_j \). The statistic

\[
T_0 = y - a_u + \bar{a}
\]

is an unbiased estimator of \( \bar{Y} \) with zero risk (variance) when \( \theta = \theta_0 \). The variance of \( y \) is

\[
\frac{1}{N} \sum_{j=1}^{N} (Y_j - \bar{Y})^2
\]

and that of \( T_0 \) is

\[
\frac{1}{N} \sum_{j=1}^{N} (Z_j - \bar{Z})^2
\]

where \( Z_j = Y_j - a_j \ (j = 1, ..., N) \).”

Discussion of Example 1.

Basu states that he is “taking the liberty of using the symbols \( u, y, x \) and \( \theta \) both as variables and as particular values of the variables” (p208).

We re-express Basu’s discussion using our notation.

Consider the case of a simple random sample of size one from a population \( s = 1, ..., N \), where \( y_s \) is a non-stochastic response associated with label \( s \). We index the ordered selections in the sample as \( i = 1, ..., n(=1) \). We represent the sample as \( Y_i = \sum_{s=1}^{N} U_{is} y_s \) and the subject in the sample as \( S_i = \sum_{s=1}^{N} U_{is} s \). Also, let us define a permutation of the population as
\( Y = (Y_1, Y_2, ..., Y_N)' \). The realized sample is \((s_i, y_i)\) where \( s_i = \sum_{x=1}^{N} u_{ix}s, \ y_i = \sum_{x=1}^{N} u_{ix}y_s \). \( S_i \) has a uniform probability distribution over the \( N \) integers 1,2,...,\( N \). Since \( n = 1, i = 1 \) and we could drop the subscript \( i \) in representing the sample. Notice that when \( u_{ix} = 1, u_{is} = s, \) so that we could represent \( s_i = \sum_{x=1}^{N} u_{ix}s \) by \( s = \sum_{x=1}^{N} u_{ix}s \). Similarly, when \( u_{ix} = 1, u_{is} = y_s, \) so that we could represent \( y_i = \sum_{x=1}^{N} u_{ix}y_s \) by \( y_s = \sum_{x=1}^{N} u_{ix}y_s \). Basu uses the second representation of these realizations, but then drops the subscript.

Let us represent the population mean either by \( \mu = \frac{1}{N} \sum_{s=1}^{N} y_s \) or \( \bar{Y} = \frac{1}{N} (Y_1 + Y_2 + ... + Y_N) \).

We wish to estimate this parameter (or random variables). Notice that Basu does not define what is meant by the other values of \( Y_i \) when \( i^* \neq i \). For example, if sampling is with replacement, then \( Y \) is a random variable, the values of \( Y_i \) and \( Y_{i^*} \) for \( i^* \neq i \) are uncorrelated, and \( \bar{Y} = \mu \) although \( E(\bar{Y}) = \mu \) while \( \text{var}(\bar{Y}) = \frac{\sigma^2}{N} \). If sampling is without replacement, then the values of \( Y_i \) and \( Y_{i^*} \) for \( i^* \neq i \) are correlated, and \( \bar{Y} = \mu \) is a constant. Clearly, \( Y_i \) is an unbiased predictor of \( \bar{Y} \) since \( E(Y_i) \) is equal to \( E(\bar{Y}) = \mu \), and an unbiased estimator of \( \mu \). Let \( a = (a_1, a_2, ..., a_N)' \) and let \( \bar{a} = \frac{1}{N} \sum_{s=1}^{N} a_s \). We represent the sample corresponding to \( a \) as \( A_i = \sum_{x=1}^{N} U_{is}a_s \), a realized value as \( a_i = \sum_{x=1}^{N} u_{ix}a_s \) or \( a_s = \sum_{x=1}^{N} u_{is}a_s \), and the statistic

\[
T_0 = Y_i - A_i + \bar{a}
\]

[whose realized value is \( t_0 = y_i - a_i + \bar{a} \) (or \( t_0 = y_s - a_s + \bar{a} \) ) is an unbiased estimator of \( \bar{Y} \) with zero risk (variance) when \( Y = a \). The variance of \( Y_i \) is

\[
\text{var}(Y_i) = E[(Y_i - E(Y_i))^2] = \frac{1}{N} \sum_{s=1}^{N} (y_s - \mu)^2.
\]

It is not clear to me what Basu means by his expression for the variance of \( Y_i \). For example,

\[
\frac{1}{N} \sum_{i=1}^{N} (Y_i - \bar{Y})^2 = \frac{1}{N} \sum_{i=1}^{N} ((Y_i - \mu) - (\bar{Y} - \mu))^2 = \frac{1}{N} \sum_{i=1}^{N} (Y_i - \mu)^2 - 2 \frac{1}{N} \sum_{i=1}^{N} ((Y_i - \mu)(\bar{Y} - \mu)) + \frac{1}{N} \sum_{i=1}^{N} (\bar{Y} - \mu)^2 = \frac{1}{N} \sum_{i=1}^{N} (Y_i - \mu)^2 - (\bar{Y} - \mu)^2.
\]
Terms in this expression are random variables. Suppose the random variables represent simple random sampling with replacement. Then 
\[ E\left(\frac{1}{N} \sum_{i=1}^{N} (Y_i - \mu)^2 \right) = \sigma^2, \text{ while } E\left(\bar{Y} - \mu\right)^2 = \frac{\sigma^2}{N}. \]
If the random variables represent a random permutation (sampling without replacement), then
\[ E\left(\frac{1}{N} \sum_{i=1}^{N} (Y_i - \mu)^2 \right) = \sigma^2 \] while 
\[ E\left(\bar{Y} - \mu\right)^2 = 0. \]

Basu defines \( Z_i = Y_i - A_i \) and the corresponding realized value as \( z_i = y_i - a_i \) or \( z_i = y_i - a_i \). This notation is confusing since he uses \( Y \), making it confusing as to whether this is a constant or a random variable.

Reference