Overview of Population Generation for a Simulation Study for Performance of Balanced Two Stage Predictors of a Realized Random Cluster Means

Bo Xu

Introduction

We describe a simulation study similar to the study reported by Pfeffermann and Nathan (1981, JASA 681-689). The simulation enables evaluation of predictors in two stage cluster sampling contexts such as the Seasons study. In that study, there were 3 measures on each cluster (subject) in each 3-month period. We assume that we are interested in estimating the realized cluster mean. We may be interested in a cluster mean over different time periods, such as a 7 day average, a 30 day average, and a 365 day average. An earlier description of the simulation is contained in c03ed10.doc.The population is generated in the R program cbx05p01.r.

The Population

We define the population via a set of parameters generated using random numbers prior to the simulation. The population is generated in the R program cbx05p01.r. Although the number of clusters in the population, and the number of units in each cluster is finite, we generate the values using percentiles of a distribution. Distributions used are the normal distribution, a uniform distribution, a beta distribution, and a gamma distribution. Percentiles of the distribution are selected so that values of the clusters (or units) have a relative frequency similar to the distribution. We describe the process for developing the population generated by normal random numbers next.

For all simulations, the population mean μ will be set to a constant value. The population will be composed of N clusters, with M units per cluster. We will generate a cluster mean by defining an initial variance between cluster parameters σ_P^2. Using this variance, we will generate N random variables using a $N\left(\mu,\sigma_P^2\right)$ distribution. These values will form the basis for the cluster parameters. Since the number of clusters in the population is finite, the average of the cluster parameters will not necessarily equal the population mean μ. We will force the mean of the cluster parameters to equal μ by centering the cluster parameters at μ. Each individual cluster parameters will be represented as μ_s. We round the values of these parameters to the nearest hundredth to simplify presentation of simulation results. Finally, the variance of the cluster parameters in the generated population will not equal σ_P^2 since, once again, the number of clusters is finite. We will define the variance of the cluster parameters as $\sigma^2 = \sum_{s=1}^{N} \frac{\left(\mu_s - \mu\right)^2}{N-1}$ using the simulated cluster means.

Next, we will generate unit effects for the M units for each cluster. We will force these effects to average to zero. We generate the unit effects using a normal random variable generator $N\left(0,\sigma_{Ps}^2\right)$. Since the sum of the unit effects for the finite population will not be exactly zero, we will center these effects so that their sum is exactly zero. The parameter for a cluster-unit will be formed by adding the unit effect to the cluster mean, and represented by μ_{st}. The variance of the unit parameters for a cluster will not equal σ_{Ps}^2 since once again, the number of units per cluster is finite. We will define the variance of the unit parameters as $\sigma_s^2 = \sum_{t=1}^{M} \frac{\left(\mu_{st} - \mu_s\right)^2}{M-1}$ using the generated unit parameters. Notice that we will only generate one set of unit effects for the population. These unit effects will be used to create parameters for cluster-units for each cluster. We generate the unit effects in this manner to force the variance of units within clusters to be constant for all clusters. The common within cluster variance which we represent by σ_e^2 is equal to the average within cluster variance, $\sigma_e^2 = \sum_{s=1}^{N} \frac{\sigma_s^2}{N}$.

Step 1. Generating the Clusters Means

In order to generate the population, we first set the template distribution for clusters means, and the values of μ and σ_P^2. We use percentiles of the distribution to define the initial values of the cluster means. However, since the population is finite, the variance based on these initial values will not match σ_P^2 exactly. We re-center and re-scale the initial value so that parameters calculated from the cluster means exactly match the μ and σ_P^2.

Step 2. Generating the Units Effects for Each Cluster

We generate the unit effects using a normal random variable generator $N\left(0, \sigma_{Ps}^2\right)$. We use percentiles of the distribution to define the initial values of the units for each cluster. The parameter for a cluster-unit will be formed by adding the unit effect to the cluster mean, and represented by μ_{st}. The variance of the unit parameters for a cluster will not equal σ_{Ps}^2 since once again, the number of units per cluster is finite. We will define the variance of the unit parameters as $\sigma_s^2 = \sum_{t=1}^{M} \frac{\left(\mu_{st} - \mu_s\right)^2}{M-1}$ using the generated unit parameters. These unit effects will be used to create parameters for cluster-units for each cluster. The common within cluster variance which we represent by σ_e^2 is equal to the average within cluster variance, $\sigma_e^2 = \sum_{s=1}^{N} \frac{\sigma_s^2}{N}$.

Table 1a. Variable Definitions in the Simulation for the Population of Clusters

Variable	**R Name**	**Description**
	cd	Ordinal variable indicating distribution of cluster means. (1=Normal, 2=Uniform, 3=Beta, 4=Gamma)
	ud	Ordinal variable indicating distribution of unit values. (1=Normal, 2=Uniform, 3=Beta, 4=Gamma)
N	bn	# of clusters in the population

M	bm	# of units for a cluster
μ	mu_p	Population average
σ_P^2	popv	Initial variance of cluster means.
σ_{Ps}^2	popcv	Initial variance of units in a cluster.
μ_s	mu_s	Cluster parameter
y_{st}	t	Cluster-unit parameters
σ^2	v_star	Variance of cluster parameters (over N-1)
σ_s^2	var_sr	Variance of units for a cluster (over M-1)
$\sigma_e^2 = \sum_{s=1}^{N} \frac{\sigma_s^2}{N}$	var_sra	Average variance of units for a cluster (over M-1)
s	s	Cluster label
	v_s	Variance of the units for clusters after rescaled based on target variance
	vu	Indicator Variable to control if equal (=0) or different (=1) within cluster variances
	cd_par1	Parameter 1 for Cluster Distribution (Shape 1 (for Beta) or Shape (for Gamma))
	cd_par2	Parameter 2 for Cluster Distribution (Shape 2 (for Beta))
	ud_par1	Parameter 1 for Unit Distribution
	ud_par2	Parameter 2 for Unit Distribution (Shape 2 (for Beta))

Table 1b. Variables Used to Simulate the Population, and then Dropped

Variable	**R Name**	**Description**
$\sum_{s=1}^{N} \mu_s^*$	mu_star	Sum of initial generated cluster means (that does not equal the population average
μ_s^*	ys	Initial generated cluster means
	mu_star	Initial cluster mean.
	mu_star_t	Initial unit average
$\sum_{s=1}^{N} \mu_s^{*2}$	ss_s	Sum of squared total of initial generated cluster means
$\sum_{t=1}^{M_s} \mu_{st}^{*2}$	ss_t	Sum of squared total of initial generated unit values
$\sum_{s=1}^{N} \mu_s^*$	tot_s	Sum of initial generated cluster totals
$\sum_{t=1}^{M_s} \mu_{st}^*$	tot_t	Sum of initial generated unit totals
	u	Index for a unit in a cluster
	v_star	Variance of initial cluster means
	v_star_t	Variance of initial unit values
	ys	Initial values for cluster means
	t_mn	$N \times M$ matrix containing centered and scaled units values for all clusters
	p1	$N \times (M+2)$ matrix containing cluster parameters, centered and scaled units values based on the target variance for each cluster, variance of units for each cluster