

Superpopulations and Superpopulation Models

Ed Stanek

Contents

- Overview
- Background and History
 - Generalizing from Populations: The Superpopulation
 - Superpopulations: a Framework for Comparing Statistics
 - Targets of inference (parameters of interest)
 - Problems with the Superpopulation Concept
- Superpopulations Linked to a Potentially Observable Population
 - Superpopulation Models
 - General Models
 - Exchangeable Models
 - Ericson's Superpopulation
- Stanek's Expanded Population
- Extension to Factorial Experimental Designs
- The Role of Random Sampling

Overview

A common objective of statistics is to estimate (guess) the value of a parameter (such as the average age of residents of a community) based on partial information (ie. the age of some of the residents). What constitutes the 'best' guess may be the subject of debate. Statistics offers a strategy for resolving this debate- if agreement can be made on some basic assumptions. At a minimum, the assumptions include what types of guesses are allowed (ie. linear, unbiased), and what constitutes 'best' (ie. minimum mean squared error). Once these assumptions (and possibly others) are made, the best guesses usually take the form of equations, or formulas which are expressed as functions of random variables that are to be realized (ie. observed). When the random variables are observed, the guess can be made (ie. calculated). We call the guess, along with an estimate of the uncertainty of the guess, statistical inference.

We assert that a valuable first step in statistical inference is agreement on the definition of a parameter that is the target of inference. Such agreement may not be automatic (see (Dawid, 2000 #1052)), since it may involve metaphysical concepts (such as a population, or reference set). For many (see (Dawid, 2000 #1052) for an exception), the target parameter is defined in terms of the response for units that are observed, and the response for units that are not observed (which we term potentially observed) in the population. For example, the average age of community residents is defined as the average of the ages of residents of the community.

The response for a unit not observed (and since it is not observed) may be defined in different ways. The response for such a unit may be considered to be a parameter, but one that is unknown. If the unit was observed, the parameter would be known. Alternatively, the response for such a unit may be considered to be the realization of a random variable, where a model has been specified for the random variable. If the unit was observed, although the response would be known, the parameter representing the expected response for the unit may remain unknown. Some definition of response for a unit that is not observed is necessary to define target parameters.

Once target parameters are clearly defined, the values of the units (whether observed or not observed) are represented as a set of random variables. This collection of random variables is called a superpopulation. The random variables in the superpopulation are defined in terms of an assumed probability structure. The probability structure is often defined and stated in terms of a superpopulation model. The probabilities connect the potentially observed random variables to those not observed. With such a probability structure and assumptions, a consensus can be reached as to what constitutes the 'best' guess.

We discuss in some detail the idea of a superpopulation, and various expositions of a superpopulation given in the literature. The discussion highlights the fact that the term ‘superpopulation’ may have different meaning to different authors. This discussion organizes and motivates the different definitions. The discussion is also valuable since the joint distribution of the random variables in the superpopulation may only be partially specified. Finally, we define an expanded population and related it to a common superpopulation model, the random permutation superpopulation model. We follow this with a brief mention of how similar ideas can be used in a simple factorial experimental study. Finally, we comment on the implications of this framework for the need for randomization.

Background and History

Generalizing from Populations : The Superpopulation

A concept of a more general population was first introduced when discussing estimators and analyses using repeated survey and census data. An early use of the idea was given by (Cochran, 1939 #1032) when discussing how ANOVA could be used to quantify different sources of sampling error. In the paper, Cochran summarized the ANOVA tables for similar surveys on wheat yield/acre that were conducted in five consecutive years (see below):

Table III. Analysis of variance per field of yields of wheat grain
(cwt. Per acre)

	1934		1935		1936		1937		1938	
	d.f.	m.s.	d.f.	m.s.	d.f.	m.s.	d.f.	m.s.	d.f.	m.s.
Between Districts	4	66.5	6	318.4	4	79.4	5	82.3	4	206.8
Within districts between farms	11	38.9	12	27.1	7	62.2	19	52.7	14	65.3
Within Farms between fields	-	-	15	22.8	8	31.2	11	24.2	8	12.1
Within fields	16	5.3	40	6.20	22	11.39	39	6.60	28	9.80

Sampling error										
Within fields between sets	32	2.11	80	2.18	45	2.52	78	5.09	55	4.78
Mean yield	29.1		23.3		24.3		26.2		30.7	

Source: (Cochran, 1939 #1032), p501.

Discussing these results, Cochran stated (p501):

‘The results from different years are not entirely independent, since in a number of cases the same farms were included in the samples, but the samples from different years are much more valuable than a single sample five times as large taken in a single year.’

Presumably, the samples from different years are more valuable since they account for other realistic sources of variability such as changes in annual temperature and rainfall. Over years, the population of districts, farms and fields may change, along with the yield per acre for a given field. While a population can be defined each year, a more general concept occurs when considering populations over a set of years. Cochran considers characterizing this larger collection of information to be more valuable.

The focus on the larger collection of information was evident when Cochran formulated sampling error when an appreciable fraction of the population in a year was sampled. Thus, Cochran said (p506):

‘The finite population should itself be regarded as a random sample from some infinite population; thus the sample which is taken for enumeration is regarded as a subsample from a larger sample of the same infinite population. Further, in so far as the sampling is carried out for the purpose of estimating the mean of some character in the finite population (i.e. in the larger sample), sampling errors must be measured about the mean of the larger sample. With these two points in mind, the ordinary rules of the analysis of variance may be applied. Consider the variance of the mean \bar{x}_n of a random sample drawn from a larger sample with N units and mean \bar{x}_N . Let \mathbf{s}^2 be the variance in the infinite population. Then

$$\bar{x}_n - \bar{x}_N = \left(\frac{1}{n} - \frac{1}{N} \right) (x_1 + x_2 + \dots + x_n) - \frac{1}{N} (x_{n+1} + \dots + x_N).$$

Hence

$$\begin{aligned}V(\bar{x}_n - \bar{x}_N) &= \left\{ n \left(\frac{1}{n} - \frac{1}{N} \right)^2 + \frac{N-n}{N^2} \right\} \mathbf{s}^2 \\ &= \frac{N-n}{N} \mathbf{s}^2.\end{aligned}$$

Note that this definition of the variance, \mathbf{s}^2 , is not the same as the definition of

$$S^2 = \frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N-1}$$

used in Cochran's sampling text (Cochran, 1977 #1067) (p26). The

definition of \mathbf{s}^2 refers to an infinite population.

Apart from the infinite population variance, Cochran (Cochran, 1939 #1032) did not discuss parameters of the infinite population directly. Cochran did note their obvious interest (Cochran, 1939 #1032)(p507):

'Where the population consists of a single group, the results obtained by 'finite sampling theory' agree with those obtained by the analysis of variance. The former is, however, not easily extended to the case in which the population is sub-divided into groups, at least so far as the situations arising in practice are concerned. Further, it is far removed from reality to regard the population as a fixed batch of known numbers. In economic and sociological studies the population is changing from day to day. The population at any one time is often conventional, as for example with a population of farms or carpenters, owing to the difficulty in defining a member of the population.'

These remarks indicate that parameters in the fixed finite population may not always be the target of inference. There was a need for a broader concept that corresponds to the reality of a changing population.

At the US Census Bureau, Deming and Stephan (Deming, 1941 #1033) argued that there was value in considering a broader concept than a population when conducting a census. These authors declared (p45):

'As a basis for scientific generalizations and decisions for action, a census is only a sample.'

Their reasons for this assertion are similar to those of (Cochran, 1939 #1032). The census is a sample in the following sense (p45):

‘A census describes a population that is subject to the variations of chance, because it is only one of the many possible populations that might have resulted from the same underlying system of social and economic causes.’

They considered a census an inventory at a point in time. While the census is valuable for certain functions (such as apportionment of funds), since a census at a future time will be different, the authors viewed the census as a sample valuable for predicting the future. The authors quote a letter written by Walter A. Shewhart to W.E. Deming dated 5/9/1940 (p46):

‘A so-called 100 per cent sample from the viewpoint of scientific method is, as soon as taken, a sample of the past. The usefulness of such a sample is only as a basis for drawing an inference about the future and in this case the sample (even a 100 per cent sample) is but a finite sample of a potentially infinite one that might result from the cause system existing at the time the sample was taken.’

Deming and Stephan (Deming, 1941 #1033) introduce the term ‘super-population’ in this context (p48):

‘From the point of view being expressed here, however, even a complete census, for scientific generalizations, describes a population that is but one of the infinity of populations that will result by chance from the same underlying social and economic cause systems. This infinity of populations may itself be thought of as a population, and might possibly be called a **super-population**. A sample enquiry is then a sample of a sample, and a so-called 100 per cent sample is simply a larger sample, but is still only a sample. In order to study the underlying cause systems, it is necessary to study several members of this infinity of populations; i.e., it is necessary to make sample or census enquiries on a number of different dates, preferably far enough apart to be independent, or nearly so.’ (bold added)

When summarizing concepts of superpopulations (Cassel, 1977 #1025) (p81), the first concept of a superpopulation cited by Cassel et al. agrees with (Cochran, 1939 #1032) and (Deming, 1941 #1033):

‘1. The finite population is actually drawn from a larger universe. This is the superpopulation idea in its most pure form.’

Different sets of N subjects can arise from the infinite superpopulation. These differences loosely parallel the differences between the actual population at the current time, and the conceptual population that is like those at the present time, but is of more fundamental interest (Barnard, 1973 #1073). With this concept, the superpopulation is postulated to provide an abstract representation of a broader entity from which the population values are generated (Sarndal, 1992 #1024) (p22).

Superpopulations: a Framework for Comparing Statistics from a Population

A second early use of an infinite population (or superpopulation) was as a theoretical concept to facilitate evaluating of competing survey designs or statistics. Cochran (Cochran, 1946 #576) used the idea of an infinite population to compare stratified and systematic sample designs. In order to make the comparison, Cochran postulated an (p166) ‘idealized population’ that he assumed satisfied certain properties:

‘Thus, comparisons between the systematic and stratified random samples will be made not from a single finite population, but for the average of finite populations drawn from an infinite population with monotone decreasing r .’

Similar to Cochran’s notion of a superpopulation, Hartley and Sielken (Hartley and Sielken 1975) ES 2272 formalized this idea by defining a superpopulation(p411):

‘... the super-population outlook regards the finite population of interest as a sample of size N from an infinite population and regards the stochastic procedure generating the surveyor’s sample of n units as the following two-step procedure:

Step 1. Draw a “large sample” of size N from an infinite super-population.

Step 2. Draw a sample of size $n < N$ from the large sample of size N obtained in Step 1.

Actually, Step 1 is an imaginary step, and it is usually assumed that the resulting sample elements are independent and identically distributed. Step 2 is the “real sample survey” conducted by the surveyor in accordance with his specified design p_s . The super-population theory is, therefore, concerned with repeated implementations of the two-step stochastic process consisting of Step 1 followed by Step 2. Thus, in terms of the super-population model the finite population sampling may be regarded as being based on the conditional distribution given a particular outcome of its Step 1 process.’

Both of these authors consider the superpopulation to be artificial, but use the concept as a mathematical device to derive properties of estimators of population parameters.

A different concept of a superpopulation was given by (Brewer, 1963 #1089).

Brewer considered a basic problem with studying the properties of estimators (such as ratio and regression estimators) in sampling theory was that

‘the population is treated as an entity in itself, completely independent of any stochastic process which may have generated it.’ (p93)

Brewer postulated an underlying stochastic process that gave rise to the population, motivating his idea with the following example (p95):

‘To take an imaginary example, if the problem were to estimate the production of butter in Australia each month, the value of butter production Y_t by a given factory in a given month in fact depends to a large extent on the known value of butter production Z_t by the same factory over the year covered by the last Factory Census. Alternatively, it may be regarded as dependent on the known wage bill for that factory in that month, in which case “wages” would be a useful benchmark item. In other words, given the value of the benchmark item Z_t we may make a stochastic estimate of Y_t .’

Using this as a motivating example, Brewer proposed a larger stochastic population (p95):

‘From this point of view, the actual finite population with which the sampling statistician is confronted may be regarded as one particular state of affairs (namely the state of affairs which in fact exists) from all the possible states of affairs which might have existed, given the benchmark item information. It may, in fact, be regarded as a sample of one from an infinite number of finite populations, all with the same values of Z_i but with values of Y_i varying from population to population and stochastically dependent on the Z_i .’

Brewer then used this stochastic framework to derive various properties of estimators under different sampling designs.

Isaki (Isaki, 1982 #991) cites a variety of workers who have used the superpopulation concept to compare survey designs or estimators. In addition to those previously discussed, they cite (Madow, 1944 #53)(Yates, 1949 #1099)(Godambe, 1955 #1057)(Hajek, 1959 #1094)(Rao, 1962 #1074)(Godambe, 1965 #1059)(Hanurav, 1966 #1095)(Isaki, 1970 #1096)(Rao, 1971 #1097)(Fuller, 1975 #1003)(Cassel, 1976 #997)(Brewer, 1979 #1075)and (Sarndal, 1980 #1098).

Targets of inference (parameters of interest)

The different interpretations given to a superpopulation in the 1940s and 1950s can be distinguished by a focus on different parameters of interest. Hartley et al (Hartley and Sielken 1975) ES 2272 discuss two sets of possible target parameters (finite population parameters, and infinite superpopulation parameters). The authors characterize the target parameters and sampling in Table 1 (below).

Table 1. Sampling theories classified by sampling procedures and target parameters.

Target Parameters	Sampling Procedure	
	Repeated sampling from a fixed finite population	Repeated two-step sampling from an infinite population
Parameters of Finite Population	Classical finite population sampling theory	Super-population theory for finite population sampling

	= Case 1	= Case 2
Parameters of infinite Super-population	Infeasible	Inference on infinite population parameters from two-step sampling procedure = Case 3

(Hartley, 1975 #1004) (p412)

Sarndal et. al. (Sarndal, et al. 1992) ES 2289, discuss two targets of inference similar to

Hartley and Sielken (Hartley, 1975 #1004). From Sarndal, p514:

‘To make inference, in a statistician’s language, usually means to draw conclusions, with the aid of probability statements, from a sample to a larger universe. Confidence intervals and hypothesis tests are traditional tools of inference. Suppose we have data obtained by measuring the elements k in a probability sample s drawn from the finite population U . Two important types of inference are as follows:

- a. Inference about the finite population U itself.
- b. Inference about a model or a superpopulation thought to have generated U .’

Target parameters for Sarndal’s ‘type a’ inference corresponds to Hartley’s finite population parameters. Target parameters for Sarndal’s ‘type b’ inference corresponds to Hartley’s superpopulation parameters.

When the superpopulation is artificial, or ‘purely idealized’ (as in (Cochran, 1946 #576)) there is little interest in the superpopulation parameters. Some authors (Rao, 1975 #4) have referred to superpopulations only in this idealized sense, and by default, focused inference on finite population parameters.

Other authors have focused on finite population target parameters (Sarndal, 1992 #1024), although not because the superpopulation was artificial. Sarndal continues, (Sarndal, 1992 #1024) (p514):

‘Case (a) has occupied a major part in this book. The objective is to estimate and in other ways make inferences about descriptive parameters that

characterize U . This is inference about the “now,” about the current state of the finite population. By contrast, case (b) poses questions about the process that underlies the finite population U .

Continuing, Sarndal’s (p514) further comments:

‘The model builder is interested not in the finite population U at the present moment in time, but rather in the process or the causal system relating y to z .’

When the superpopulation is considered to be a generalization of finite populations (Cochran, 1939 #1032)(Deming, 1941 #1033)(such as a census over years), the parameters of the superpopulation have meaning. Characterizing the superpopulation will correspond to characterizing the causal system. Knowing the causal system will help predict the special case (population) in the future. This idea underlies the attention given to estimating superpopulation parameters given by (Konijn, 1962 #1006)(Fuller, 1975 #1003)(Frankel, 1971 #1078) referred to target parameters corresponding to Hartley’s Case 1 and Case 3 (as attributed by Hartley (Hartley, 1975 #1004)).

For many authors, the target of inference is the finite population parameters. This was the object of Hartley, (Hartley, 1975 #1004as well as others) (Isaki, 1982 #991)(Bolfarine, 1992 #1027)) . Such parameters have clear definitions.

Problems with the Superpopulation Concept

Superpopulations have been discussed historically as a general collection of populations of interest, or an imaginary (possibly infinite) population with particular properties. When the superpopulation is a general collection of populations, the superpopulation parameter is often thought to be of most interest. The motivation for

introducing a superpopulation can be to draw inference to the superpopulation parameter, and thus, generalize the results of a given study.

Such a use of superpopulations is problematic. The idea is that ‘By imagining a bigger concept, we can draw more general conclusions.’ The statement is problematic for two reasons. If the bigger concept is meaningful (such as the effect of a drug on all adults, where the superpopulation is ‘adults’), then the link between the superpopulation and population is often arbitrary, resulting in questionable inference. In contrast, if the link between the superpopulation and the population is clear (such as the population being a simple random sample from the superpopulation), then the superpopulation parameter will have an artificial definition, and hence not provide the generalization sought for.

The uneasiness over using infinite superpopulations to generalize conclusions is avoided (at least in part) by targeting inference to finite population parameters, as opposed to superpopulation parameters. The basic approach is to define the object of the inference as a function of the realized and unrealized (or partially realized) values of the realized finite population units. Often, the population is thought to be a simple random sample from the artificial superpopulation. A model for the superpopulation is assumed, and used in estimating the finite population parameter. While the superpopulation model, and the linking of the population to the superpopulation can be questioned, this use of superpopulations does not imply generalizing beyond the finite population. We illustrate this with an example.

Suppose that a characteristic such as farm acreage (or number of family members in a health plan) is known for each unit (i.e. farm, or family) in a finite population. Suppose the target of inference is the average corn yield per acre (or average number of physician visits per member) in a season (year). The prior information (farm acreage) is treated as a sample

(of size N) from an infinite superpopulation. A model is postulated that relates the yield to acres for units in the infinite superpopulation. If this model holds in the superpopulation, then we can make use of it to obtain a more reliable finite population estimate (Isaki, 1982 #991). The specification of the units in a future finite population, or future realized yields on the finite population units are not involved in the inference. What is involved is the general relationship between the two variables (yield and acreage), which may be supported by the realized sample data. [Aside: It is not clear to me why we need an infinite superpopulation here, as opposed to simply a model in the finite population- apart from wanting to make distributional assumptions on the errors.]

Superpopulations Linked to a Potentially Observable Population

Early descriptions of superpopulation in survey sampling did not formally link labeled units in the population to labeled units in the superpopulation. Over time, this linking of the units in the population and the superpopulation became more standard. Cassel et al. defines a common framework and notation that links units in a finite population to random variables in a superpopulation such that (Cassel, et al. 1977) ES 2290 p80:

“... the vector of population values $\mathbf{y} = (y_1 \cdots y_N)$, is assumed to be the realized outcome of a vector random variable $\mathbf{Y} = (Y_1 \cdots Y_N)$. The joint distribution of $\mathbf{Y} = (Y_1 \cdots Y_N)$ will be denoted by \mathbf{x} .”

Using this notation, (p81-82), the finite population is represented by a vector of fixed constants (or parameters) denoted by \mathbf{y} , while the superpopulation is represented by a corresponding vector of random variables denoted by \mathbf{Y} . This notation is commonly used in

subsequent literature. Note that the subscript of y refers to the corresponding unit as the subscript of Y . For example y_s and Y_s both refer to unit s .

The Superpopulation Model

A superpopulation that is linked to a finite population, as defined by (Cassel, et al. 1977) ES 2290, can be thought of as defining a basic superpopulation model. Bolfarine and Zacks use this idea, where (Bolfarine, 1992 #1027) (p8):

“The superpopulation model assumes that the value of the variable of interest, associated with the i th unit of the population, $y_i, i = 1, \dots, N$, is comprised of a deterministic element h_i and a random element e_i ; that is,

$$y_i = h_i + e_i,$$

$i = 1, \dots, N$. The random vector $\mathbf{e} = (e_1 \ \dots \ e_N)$ is assumed to have zero mean and a positive definite covariance matrix, \mathbf{V} .”

Using the notation of Cassel’s (Cassel, et al. 1977) ES 2290, the superpopulation model of Bolfarine and Zack is given by:

$$Y_i = y_i + E_i,$$

for $i = 1, \dots, N$. In this model, y_i represents a parameter for the i^h unit. Additional structure can be added to the basic superpopulation model. Cassel et al., (Cassel, et al. 1977) ES 2290 p83, discusses two classes of superpopulation models (general models and exchangeable models) that we describe.

General Models

We describe two general models, the transformation model, and the regression model.

The transformation model is defined by (Cassel, 1977 #1025)(p83):

“This model defines the class of probability measures \mathbf{x} on R_N such that, for given numbers $a_k, b_k (k = 1, \dots, N)$, specified by the model maker, the transformed random variables $Z_k = (Y_k - b_k) / a_k (k = 1, \dots, N)$ have common

means, \mathbf{m} , variances, \mathbf{s}^2 , and common covariances, \mathbf{rs}^2 , for any pair $k \neq \ell$.'

The multiple regression model is defined by (Cassel, 1977 #1025) (p84) in the '... presence of known auxiliary variable measurements $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$.'. Then the multiple regression model is defined as:

'The class of probability measures \mathbf{x} on R_N such that Y_1, Y_2, \dots, Y_N are independently distributed, and

$$\mathbf{m}_k = E_{\mathbf{x}}(Y_k) = \mathbf{b}_1 + \sum_{i=1}^q \mathbf{b}_i x_{ki}, \quad \mathbf{s}_k^2 = E_{\mathbf{x}}(Y_k - \mathbf{m}_k)^2 = \mathbf{s}^2 v_k$$

where $\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_q, \mathbf{s}^2$ are unknown, and $x_{k2}, x_{k3}, \dots, x_{kq}, v_k$ is a set of known numbers for every $k (k = 1, \dots, N)$.'

An equivalent multiple regression model is defined by (Bolfarine, 1992 #1027) (p9).

Exchangeable Models

As described by (Cassel, 1977 #1025) (p83), the general superpopulation models are not necessarily exchangeable models. Cassel et al define exchangeable superpopulation models such that (p85):

"In order to be exchangeable, the distribution \mathbf{x} must be symmetric in accordance with the following definition.

Definition. Random variables Y_1, \dots, Y_N are called *exchangeable* if

Y_{r_1}, \dots, Y_{r_N} has, for every permutation r_1, \dots, r_N of $1, \dots, N$, the same joint distribution, which is called an *exchangeable distribution*."

Such models were discussed in a Bayesian context by Ericson (Ericson, 1969 #1035), but we defer such discussion (since it involves a different notion of a superpopulation).

Variations of exchangeable models are given by Cassel (Cassel, 1977 #1025)p86-87. One

variation assumes that a transformation of the random variables in the superpopulation (similar to the transformation model above) results in exchangeable random variables. Another variation is the random permutation superpopulation model, in which realizations of the superpopulation with elements corresponding to permutations of the labels are equally likely. Such models were described by (Kempthorne, 1969 #1071) and extended to two stage sampling by (Rao, 1975 #4). These models were developed without explicit definition of a superpopulation.

Ericson's Superpopulation

Ericson (Ericson, 1969 #1035) considered Bayesian estimation for a finite set of exchangeable random variables. In the previous section, the finite set of exchangeable random variables defines the superpopulation. To implement the Bayesian inference paradigm, Ericson needed to specify a prior distribution for \mathbf{Y} . Ericson defined a new superpopulation to motivate a prior distribution for \mathbf{Y} (a more detailed discussion of this is given in c00ed65.doc). As summarized by Cassel (Cassel, 1977 #1025) (p89):

‘the new superpopulation distribution (referred to by Bayesians as a “prior distribution for Y_1, Y_2, \dots, Y_N ”) can be generated as a mixing distribution, namely, if we assess a prior distribution for $\boldsymbol{\theta}$, $F(\boldsymbol{\theta}|\boldsymbol{\phi})$, where $\boldsymbol{\phi} = (\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_t)$ is a known parameter vector.’

Ericson's new superpopulation can be motivated in the following manner. Suppose a finite list of t values is constructed, where the list includes all possible values (ie. y_s) in the finite population. Some of the values in this list may occur on more than one unit in the finite population, while other values may not occur on any population units. Now, define a new superpopulation (possibly infinite) such that all the values of units in the superpopulation are contained in the same list. In this new superpopulation, let the proportion (probability) of a

value be given by q_i , and let θ represent the collection of these parameters. Ericson assumes that the distribution of θ , $F(\theta|\phi)$ is known, and then works backward to generate a prior distribution of Y , using a Dirichet distribution for θ .

In Ericson's new superpopulation, although labels are included for units in the population, and in the superpopulation, they are not linked in a one to one manner in the new superpopulation.

Stanek's Expanded Population

We describe an expanded population which is closely related to the superpopulation underlying a random permutation superpopulation model, as discussed by (Kempthorne, 1969 #1071)(Rao, 1971 #111)(Hartley, 1971 #1034)(Godambe, 1973 #1088)(Rao, 1975 #4) and (Cassel, 1977 #1025). The idea behind the superpopulation is a random labeling (or permuting) of units. We begin by repeating Hartley and Rao's description of the population, and sampling in this setting (Hartley, 1971 #1034)(p25-26):

'Step 1. Random Labeling of Units: The N units of a population are conceptually identified by a non-observable index j ($j = 1, \dots, N$). Before the sampling Step 2 commences, and unknown to the sampler, labels $i = 1, \dots, N$ are attached to the units by choosing one of the $N!$ permutations $i(j)$ with equal probabilities $1/N!$.

Step 2. The Sample Selection: Given the set of N labeled units ($i = 1, \dots, N$) a sample of fixed size n is drawn by what is called a 'size determined' design.'

We can represent this setting in terms of an adaptation of the superpopulation framework for single stage sampling considered by Stanek (c00ed27.doc), which we briefly outline here. Suppose the finite population is a labeled set of units j ($j = 1, \dots, N$). [These

labels correspond to h ($h = 1, \dots, H$) in c00ed27.doc.] Associated with each unit is an unknown value, y_j , possibly vector valued, that is associated with unit j and is made exactly known by observing unit j .

We now define a set of indicator random variables U_{ij} for ($i = 1, \dots, N$) that have a value of one if the i^{th} unit in a permutation is unit j and zero otherwise. We also define vectors $\mathbf{Y}'_j = (Y_{1j} \ \dots \ Y_{ij} \ \dots \ Y_{Nj})$ where $Y_{ih} = U_{ih}y_h$, and represent the superpopulation as $\mathbf{Y}' = (\mathbf{Y}'_1 \ \dots \ \mathbf{Y}'_h \ \dots \ \mathbf{Y}'_H)$. The superpopulation is an $N^2 \times 1$ vector of random variables.

Now consider projecting the superpopulation onto a subspace of random variables of dimension $N \times 1$ (see c00ed28.doc) by pre-multiplication by the

$$\text{matrix } \mathbf{P}' = \mathbf{1}'_N \otimes \mathbf{I}_N. \text{ Then } \mathbf{X} = ((X_i)) = \mathbf{P}'\mathbf{Y} = \begin{pmatrix} \sum_{j=1}^N U_{1j}y_j \\ \sum_{j=1}^N U_{2j}y_j \\ \vdots \\ \sum_{j=1}^N U_{Nj}y_j \end{pmatrix}. \text{ We represent elements in this}$$

$N \times 1$ vector as $X_i = \sum_{j=1}^N U_{ij}y_j$. The vector of N labeled units ($i = 1, \dots, N$) associated with the random variables X_i is the superpopulation from which Hartley and Rao (Hartley, 1971 #1034) make their sample selection.

The projected superpopulation vector, $\mathbf{P}'\mathbf{Y}$, is a vector of random variables of the same dimension as the original population. Note that it is not possible to project this vector back to the population, even though realizations of the superpopulation vector span the same space as the population.

Extension to Factorial Experimental Designs

It is possible to extend these ideas to factorial experimental designs. The key to the extension is conceiving of the potentially observable population (will all subjects receiving all treatments) representing one large structured population. For example, if there are N subjects and T treatments, the potentially observable population will have NT possible elements. In a factorial study, only a portion of the population will be observed (corresponding to the sample). The remainder is to be predicted in order to predict parameters in a model. For many (see (Dawid, 2000 #1052) for an exception), the treatment parameters are defined in terms of the difference between the treated and not treated response for all subjects (including units observed and units not observed).

An expanded population can be readily constructed. This expanded population will have N^2T^2 units. This expanded population is analogous to the expanded population resulting from simple random sampling.

The Role of Random Sampling

Superpopulations consist of random variables that are connected to units in a finite population. The connection is via a probability structure. This structure is assumed. For example, when an investigator selects a simple random sample, we assume a probability structure for random variables in a superpopulation corresponding to the simple random sample. This assumption is commonly believed and not disputed. In other settings, when an investigator claims that the data she has in hand is like a simple random sample from a larger

population, not all may believe the probability structure that connects the set of data to the larger population. Similar arguments can be raised in experimental factorial studies.

These remarks serve to underscore a general rationale for random sampling, and randomization in experimental design. The process of randomization makes the assumed probability structure that connects the population to the superpopulation believable. With the acceptance of such a connection, inference can be drawn from the realized sample to the population.

References