Survey Methodology, June 1997
Vol. 23, No. 1, pp. 3-10
Statistics Canada

A Computer Algebra for Sample Survey Theory

J.E. STAFFORD and D.R. BELLHOUSE[1]

ABSTRACT

A system of procedures that can be used to automate complicated algebraic calculations frequently encountered in sample survey theory is introduced. It is shown that three basic techniques in sampling theory depend on the repeated application of rules that give rise to partitions: the computation of expected values under any unistage sampling design, the determination of unbiased or consistent estimators under these designs and the calculation of Taylor series expansions. The methodology is illustrated here through applications to moment calculations of the sample mean, the ratio estimator and the regression estimator under the special case of simple random sampling without replacement. The innovation presented here is that calculations can now be performed instantaneously on a computer without error and without reliance on existing formulae which may be long and involved. One other immediate benefit of this is that calculations can be performed where no formulae presently exist. The computer code developed to implement this methodology is available via anonymous ftp at *fisher.stats.uwo.ca*.

KEY WORDS: *k*-statistics; Partitions; Product moments; Ratio and regression estimators; Symbolic computation; Variance estimation.

1. INTRODUCTION

In classical sampling theory two general problems concern us. These are the determination of an unbiased estimator of a parameter θ and the calculation of moments of $\hat{\theta}$, the estimator of θ.

The basic method to handle expectations and unbiased estimation is to operate on sample and population nested sums respectively through the inclusion probabilities, either single or joint probabilities as appropriate. A nested sum is a sum over the range of one or more indices such that each term in the sum depends on indices of different value. An unbiased estimator of any population nested sum is the associated sample nested sum with the quantity under the summation divided by the appropriate inclusion probability. Similarly the expectation of any sample nested sum is the associated population nested sum with the quantity under the summation multiplied by the appropriate inclusion probability.

In sampling theory, as well as several other areas of statistics, many algebraic calculations depend on a partition of some kind. With particular reference to sampling, Wishart (1952) showed that basic moment calculations under simple random sampling without replacement relied heavily on partitions. Here we will use partitions to express the sum of products of means or totals as linear combinations of nested sums and vice versa.

In the results presented here we consider the situation in which θ and $\hat{\theta}$ can be expressed as smooth functions of means or totals, population or sample as appropriate. There are two possibilities: the smooth function under consideration can be expressed as the sum of products of means or totals, or the smooth function cannot be so expressed. When the second possibility is operative the function $\hat{\theta}$ is first linearized through a Taylor expansion and θ is expressed as the root of an estimating equation. We use integer partitions to obtain terms in the Taylor linearization of a function or for the root of a function. The end result is that θ and $\hat{\theta}$ can be expressed, either exactly or approximately, as the sum of products of means or totals. These in turn can be expressed in terms of linear combinations of nested sums and vice versa. Estimation of θ or calculation of the moments of $\hat{\theta}$ is then a three step procedure: (a) Express an estimating equation for θ or the estimator $\hat{\theta}$ as the sum of products of means or totals, using Taylor linearization when necessary. (b) Transform the expression obtained in the first step to a linear combination of nested sums. Then operate on these nested sums to obtain unbiased estimates or expectations as appropriate. (c) Transform the resulting nested sums in the second step back into a sum a products of means or totals.

The key to automation of sampling theory results is the use of partitions. In general, whether these partitions are simple partitions, like that of an integer, or more complicated, like a full partition, each results from the repeated application of a fundamental rule. When the rule is identified, the possibility of automating a calculation arises. Seemingly unrelated formulae can result from the same fundamental rule and one computer algebra tool can be constructive in implementing many different calculations.

The notation used in the paper is outlined in §2. A discussion of expectation operators is given in §3. The concept of partitioning is reviewed in §4 and a rule is provided which leads to a simple recursive method for the enumeration of partitions. Integer partitions and Taylor linearization is discussed in §5. It is shown in §6 how the enumeration of partitions leads to the automatic calculation of expected values of products of sample means and *k*-statistics

[1] J.E. Stafford and D.R. Bellhouse, Department of Statistical and Actuarial Sciences, University of Western Ontario, London, Ontario, N6A 5B7.

and to the derivation of unbiased estimators of products of finite populations means and k-statistics. Also in this section we apply the methodology to ratio and regression estimation.

Automation of these calculations and derivations will provide procedures which can be performed instantaneously and without error on a computer. Also, the reliance on formulae which may be long and involved is eliminated. A great deal of hand written algebra can be avoided. All computer code for the implementation of the methodology described here was written in the symbolic package *Mathematica* 2.0 which was installed on an IBM Risc 6000 with 64 megabytes of RAM. It is available via anonymous ftp at *fisher.stats.uwo.ca*. Although we use *Mathematica*, implementation in other environments such as *Maple*, *Macsyma* or *Reduce* is no doubt possible. For example, Kendall (1993) describes a system, implemented in *Reduce*, for the identification of invariant expressions. For a complete review of computer algebra in probability and statistics prior to 1991, see Kendall (1993).

2. SOME NOTATION

Consider a finite population of size N. A measurement of interest y_j is made on each unit $j, j \in U = \{1, ..., N\}$. In addition a single auxiliary variable x_j or possibly a $P \times 1$ vector of auxiliary variables $\boldsymbol{x}_j$ may be taken on the units. The p-th entry of this vector $\boldsymbol{x}_j$ is x_{pj}, where $p = 1, ..., P$. Several kinds of finite population parameters may be defined on the measurements y_j, x_j, or $\boldsymbol{x}_j$ for $j = 1, ..., N$. We denote a finite population parameter of interest by θ. Often θ can be expressed as a smooth function of finite population means, central moments and k-statistics. For convenience here we will deal only with means and k-statistics. Note that finite population variances and covariances are also second order k-statistics.

Not all N population elements are observed. Suppose that a sample s of size n is chosen from the population U by some sampling scheme. An estimator of θ, given by $\hat{\theta}$, is a smooth function of sample means and sample k-statistics.

In order to avoid much cumbersome summation notation we adapt the index notation of McCullagh (1987) to our purposes. For any j the vector $\boldsymbol{x}_j$ contain P entries so that each of these x-variables may be associated with one of the P indices. Suppose $\{i_1, ..., i_m\}$ is a subset of m of these P indices. In our adaptation of McCullagh's notation, $x_{i_1 j}$ is now what we called the vector $\boldsymbol{x}_j$. Products of these indexed quantities become multidimensional arrays. For example the product $x_{i_1 j} x_{i_2 j} x_{i_3 j}$ is a three-dimensional array of dimension $P \times P \times P$.

Let M denote a finite population mean. The argument of M shows the structure of the summand in the mean. For example, $M(y) = \sum_{j \in U} y_j / N$ and $M(yy)$ or equivalently $M(y^2) = \sum_{j \in U} y_j^2 / N$. In index notation, for example,

$$M(x_{i_1} x_{i_2} x_{i_3}) = \sum_{j \in U} x_{i_1 j} x_{i_2 j} x_{i_3 j} / N \tag{1}$$

is a three-dimensional array. An element of this array is the mean of products in one of the permutations of the P elements taken three at a time in $\boldsymbol{x}_j$ where up to three of the elements may be alike. The (p,q,r)-th element of this array is $\sum_{j \in U} x_{pj} x_{qj} x_{rj}$ where $p, q, r = 1, ..., P$. The sample mean is denoted by m so that, for example,

$$m(x_{i_1} x_{i_2} x_{i_3}) = \sum_{j \in s} x_{i_1 j} x_{i_2 j} x_{i_3 j} / n. \tag{2}$$

For the purpose of making asymptotic expansions, since the variance of a given estimator $\hat{\theta}$ will be $O(n^{-1})$, we define a standardized variable for $\hat{\theta}$: it is the original variable $\hat{\theta}$ centered about its expectation and scaled by $1/\sqrt{n}$. That is,

$$z(\hat{\theta}) = [\hat{\theta} - E(\hat{\theta})]\sqrt{n}. \tag{3}$$

When necessary we use the summation convention of McCullagh (1987), where subscripts repeated as superscripts indicate implicit sums over that index. As a particular example, on assuming that the $\boldsymbol{x}_j$ are independent and identically distributed vectors from some infinite superpopulation, multivariate superpopulation moments can be obtained through the moment generating function which is expressed in this convention as

$$\mathrm{MGF}(\boldsymbol{t}) = 1 + \sum_{h=1}^{\infty} \mu_{i_1 \cdots i_h} \prod_{j=1}^{h} t^{i_j} / h!, \tag{4}$$

where

$$\mu_{i_1 \cdots i_h} = \frac{\partial^h}{\partial t_{i_1} \cdots \partial t_{i_h}} \mathrm{MGF}(\boldsymbol{t})|_{t=0}. \tag{5}$$

By definition, the relationship between the moment generating function and the cumulant generating function is determined by the rule $\mathrm{MGF}(\boldsymbol{t}) = \exp\{\mathrm{K}(\boldsymbol{t})\}$, where

$$\mathrm{K}(\boldsymbol{t}) = \sum_{h=1}^{\infty} \kappa_{i_1 \cdots i_h} \prod_{j=1}^{h} t^{i_j} / h! \tag{6}$$

is the cumulant generating function, where

$$\kappa_{i_1 \cdots i_h} = \frac{\partial^h}{\partial t_{i_1} \cdots \partial t_{i_h}} \mathrm{K}(\boldsymbol{t})|_{t=0}.$$

The finite population k-statistics, denoted by $K(\cdot)$, are defined as the unbiased (under the i.i.d. superpopulation model) estimators of the associated model cumulants. The number of arguments in K separated by commas denotes the order of the k-statistic. For example, the third order k-statistic $K(x_{i_1}, x_{i_2}, x_{i_3})$ is the model-unbiased estimate of (6), where

$$K(x_{i_1}, x_{i_2}, x_{i_3}) = \frac{N}{(N-1)(N-2)} \times \sum_{j \in U} [x_{i_1 j} - M(x_{i_1})][x_{i_2 j} - M(x_{i_2})][x_{i_3 j} - M(x_{i_3})]. \tag{7}$$

In the univariate case finite population k-statistics are described in Wishart (1952). In particular $K(y,y)$ and $K(y,y,y)$ in the current notation are K_2 and K_3 in Wishart's (1952) notation. The sample k-statistics, denoted by $k(\cdot)$ with the appropriate arguments, are defined as the unbiased

estimators under simple random sampling without replacement of the associated finite population k-statistics. As in Wishart (1952) the sample k-statistic can be obtained from the population k-statistic upon replacing N by n and upon taking the sum over $j \in s$ rather than all units in the finite population. For example,

$$k(x_{i_1}, x_{i_2}, x_{i_3}) = \frac{n}{(n-1)(n-2)} \times \sum_{j \in s} [x_{i_1 j} - m(x_{i_1})][x_{i_2 j} - m(x_{i_2})][x_{i_3 j} - m(x_{i_3})].$$

Note that if a comma is not present in the population or sample k-statistic, then the product of elements which appear together is required. For example, $K(xy)$ is the first order finite population k-statistic of a new variable which is the product of the measurements x_j and y_j for $j = 1, \ldots, N$; $K(x,y)$ is a second order k-statistic, in particular the finite population covariance between x and y.

3. OPERATORS

The expectation operator E can be applied directly to any sample nested sum to obtain a finite population nested sum. Likewise an unbiased estimator of any finite population nested sum is a sample nested sum. In terms of triple nested sums, for example,

$$E\left[\sum_{J_3 \in s} x_{i_1 j} x_{i_2 k} x_{i_3 l}\right] = \sum_{J_3 = 1}^{N} \pi_{jkl} x_{i_1 j} x_{i_2 k} x_{i_3 l} \tag{8}$$

and

$$\sum_{J_3 = 1}^{N} x_{i_1 j} x_{i_2 k} x_{i_3 l} \sim \sum_{J_3 \in s} x_{i_1 j} x_{i_2 k} x_{i_3 l} / \pi_{jkl}, \tag{9}$$

where J_3 is the index set $\{j, k, l\}$ such that $j \neq k \neq l$ and where π_{jkl} is a joint inclusion probability. Parallel expressions may be established for with replacement sampling schemes.

Note that m will be unbiased for the associated M under simple random sampling without replacement. In general for any sampling design of fixed size n,

$$E[m(x_{i_1} x_{i_2} x_{i_3})] = \frac{N}{n} M(x_{i_1} x_{i_2} x_{i_3} \pi)$$

and

$$M(x_{i_1} x_{i_2} x_{i_3}) \sim \frac{n}{N} m(x_{i_1} x_{i_2} x_{i_3} / \pi)$$

where $M(x_{i_1} x_{i_2} x_{i_3})$ and $m(x_{i_1} x_{i_2} x_{i_3})$ are defined in (1) and (2) respectively.

The whole operation of finding expectation of an estimator $\hat{\theta}$ or of finding an unbiased estimator for the parameter of θ may be represented schematically as

$$\sum \prod \rightarrow \sum \sum \rightarrow \sum \prod, \tag{10}$$

where $\sum\prod$ denotes the sum of products and $\sum\sum$ denotes a sum of nested sums. If θ or $\hat{\theta}$ can be expressed as a $\sum\prod$ quantity, *i.e.*, a sum of products of means, then finding an unbiased estimator of θ or moments of $\hat{\theta}$ reduces to following the schema in (10) and applying the appropriate operator, such as those given in (8) or (9), to $\sum\sum$, the middle step in the schema. If θ or $\hat{\theta}$ are smooth functions of means but cannot be expressed directly as $\sum\prod$ quantities, then an initial step is required before applying the schema in (10). For $\hat{\theta}$ the initial step is to obtain a Taylor expansion of $\hat{\theta}$. For θ the initial step is to obtain an estimating equation and then to solve it for the parameter.

We illustrate the schema in (10) by considering the simple case of finding $E[\{m(x_{i_1})\}^2]$ under simple random sampling without replacement. The first operation is to express $\{m(x_{i_1})\}^2$ in terms of nested sums. In particular,

$$\{m(x_{i_1})\}^2 = \frac{1}{n^2} \sum_{j \in s} x_{i_1 j}^2 + \frac{1}{n^2} \sum_{j \neq k \in s} x_{i_1 j} x_{i_1 k}. \tag{11}$$

This is the $\sum\prod \rightarrow \sum\sum$ step. Now the expectation operator can be applied to $\sum\sum$. On applying inclusion probabilities $\pi_j = n/N$ and $\pi_{jk} = n(n-1)/[N(N-1)]$, the expectation operation on (11) yields

$$\frac{1}{n^2} \frac{n}{N} \sum_{j=1}^{N} x_{i_1 j}^2 + \frac{1}{n^2} \frac{n(n-1)}{N(N-1)} \sum_{j \neq k = 1}^{N} x_{i_1 j} x_{i_1 k}. \tag{12}$$

Now the $\sum\sum \rightarrow \sum\prod$ step is applied. On expressing the nested sum in (12) as the sum of products, in particular $\sum_{j \neq k=1}^{N} x_{i_1 j} x_{i_1 k} = \sum_{j=1}^{N} x_{i_1 j} \sum_{j=1}^{N} x_{i_1 j} - \sum_{j=1}^{N} x_{i_1 j} x_{i_1 j}$, the third operation yields

$$E[\{(m(x_{i_1})\}^2] = \frac{N(n-1)}{(N-1)n} \{M(x_{i_1})\}^2 + \frac{N-n}{n(N-1)} M(x_{i_1}^2). \tag{13}$$

In (13), $M(x_{i_1}) = K(x_{i_1})$ and $M(x_{i_1}^2) = [N/(N-1)] K(x_{i_1}, x_{i_1}) + K(x_{i_1}) K(x_{i_1})$ so that (13) can be reexpressed as

$$E(m(x_{i_1})^2) = \{K(x_{i_1})\}^2 + (N-n) K(x_{i_1}, x_{i_1}) / (Nn). \tag{14}$$

Likewise, following the schema in (10), the operations for finding an unbiased estimator of, for example, $\{M(x_{i_1})\}^2$ is similar to (11), (12) and (13). The estimand $\{M(x_{i_1})\}^2$ is expressed in nested sums similar to (11). These sums will be nested finite population sums. Similar to (12) the inclusion probabilities are applied. In this case the finite population sums are replaced by sample sums and summand is divided by the appropriate inclusion probability. Finally, similar to (13) the resulting nested sample sums are expressed as products of sums.

Each of the elementary operations to obtain an expected value through equations (11), (13) and (14), or to obtain an unbiased estimator, can be carried out using partitions. These operations are: expressing sums of products as nested sums and vice versa, and expressing means in terms of k-statistics and vice versa.

4. PARTITIONS AND FUNDAMENTAL PROCEDURES

Central to the automation of all algebraic calculations considered here is the notion of a partition. Partitioning as a focal point gives the appearance that the automated methods presented here are nothing more than an integer partition or a partition of an index set. While we assume that a partition of an integer is understood, a full partition requires a more formal definition.

Consider a set of m indices $I_m = \{i_1, ..., i_m\}$. A single partition P_m of I_m divides the m indices into $k \le m$ mutually exclusive and exhaustive subsets or blocks of I_m. We write $P_m = (b_1 \mid b_2 \mid ... \mid b_k)$, where the $b_1, ..., b_k$ are the blocks of I_m. P_m is unique up to permutations of indices within the blocks b_i. The block b_i is comprised of a subset of the indices of I_m. Elements within a block may be constrained to an alphabetical ordering and the blocks themselves may be ordered such that leading elements of each block are ordered alphabetically. This ensures the uniqueness of the partition P_m. In this case P_m would be called a standard ordered partition. Ordering the partitions in this manner does not offer any computational advantage and hence is not a requirement in what follows. The full partition of I_m is the set $\wp_m$ of all single partitions P_m of I_m.

Now we may identify the full partition of I_m in an algorithmic way via an inclusion-exclusion rule.

i. Let $\wp_1 = \{i_1\}$.

ii. An inclusion-exclusion rule determines the contribution to $\wp_t$ by a partition $P_{t-1} \in \wp_{t-1}$. In the inclusion part of the rule, the new index i_t is added as an element in turn to each of the blocks $b_1, ..., b_k$ which comprise P_{t-1}. If P_{t-1} has k blocks, k partitions for $\wp_t$ are created. In the exclusion part of the rule a new block containing the single index i_t is added to P_{t-1}.

For example, the full partition of $I_3 = \{i_1, i_2, i_3\}$ is given by the steps

i. $\wp_1 = \{(i_1)\}$

ii. $\wp_2 = \{(i_1 i_2), (i_1 | i_2)\}$ (15)

iii. $\wp_3 = \{(i_1 i_2 i_3), (i_1 i_2 \mid i_3), (i_1 i_3 \mid i_2), (i_1 \mid i_2 i_3), (i_1 \mid i_2 \mid i_3)\}$.

From step (i) to step (ii) the inclusion rule results in the partition $(i_1 i_2)$ and the exclusion rule results in $(i_1 \mid i_2)$. From step (ii) to step (iii) the inclusion rule results in the creation of the partitions $(i_1 i_2 i_3)$, $(i_1 i_3 \mid i_2)$, and $(i_1 \mid i_2 i_3)$. The exclusion rule yields the partitions $(i_1 i_2 \mid i_3)$ and $(i_1 \mid i_2 \mid i_3)$. This type of construction is easy to automate since it depends on a simple rule. Details of automating the partition of indices into full partitions and complementary set partitions are given in Stafford (1996).

Consider, for example, the classical problem of writing the model moments of the random vector x_{i_1} in terms of its cumulants. As in (5) we can identify the h-th moment array by differentiating MGF(t) in (4) h times and setting t equal to the zero vector. The result is the h-th coefficient in the expansion of MGF(t). Equivalently we can apply the same operation to $\exp\{K(t)\}$. In this case the result is a sum that depends on the coefficients of K(t) in (6). For example, we may write the first three moments in terms of cumulants as follows:

$$\begin{aligned}
\mu_{i_1} &= \kappa_{i_1} \\
\mu_{i_1 i_2} &= \kappa_{i_1 i_2} + \kappa_{i_1}\kappa_{i_2} \\
\mu_{i_1 i_2 i_3} &= \kappa_{i_1 i_2 i_3} + \kappa_{i_1 i_2}\kappa_{i_3} + \kappa_{i_1 i_3}\kappa_{i_2} + \kappa_{i_1}\kappa_{i_2 i_3} + \kappa_{i_1}\kappa_{i_2}\kappa_{i_3}.
\end{aligned}$$

Now in each case the result is a sum over the full partitions given in (15). These partitions arise since the multiplication rule for differentiation mimics the inclusion-exclusion rule for the enumeration of the full partition.

The above result is applied to sampling theory where we consider the problem of finding the expected value of a product of sample sums. The calculation requires expanding the product of the sums to identify terms where the finite population expectation operator will behave differently due to differences in the values of inclusion probabilities and joint inclusion probabilities.

For example, the product of sums $\sum_{j\in s} x_{i_1 j} \sum_{j\in s} x_{i_2 j} \sum_{j\in s} x_{i_3 j}$ can be expressed as

$$\sum_{j\in s} x_{i_1 j} x_{i_2 j} x_{i_3 j} + \sum_{j\neq k\in s} x_{i_1 j} x_{i_2 j} x_{i_3 k} + \sum_{j\neq k\in s} x_{i_1 j} x_{i_2 k} x_{i_3 j} + \sum_{j\neq k\in s} x_{i_1 k} x_{i_2 j} x_{i_3 j} + \sum_{j\neq k\neq l\in s} x_{i_1 j} x_{i_2 k} x_{i_3 l} \cdot \quad (16)$$

The result corresponds to the full partition of the indices $I_3 = \{i_1, i_2, i_3\}$ given by $\wp_3$ in (15). The order of the partitions in $\wp_3$ is the same as the order given for the terms in (16). For each partition in $\wp_3$, the variables in the same block have the same second index in the appropriate term in (16). For example, the partition $(i_1 i_3 \mid i_2)$ corresponds to the term $\sum_{j\neq k\in S} x_{i_1 j} x_{i_2 k} x_{i_3 j}$ in (16). Each term in the result can be identified by a partition of I_3 and each partition determines the manner in which the expected value operator will behave.

In general, we want to expand products of the form $\prod_{r=1}^{m} \sum_{j\in s} x_{i_r j}$, where the product is taken over the elements i_r of the index set $I_m = \{i_1, ..., i_m\}$. As in (16), the product can be expressed in terms of the full partition of I_m. This is because the iterative rule for expanding a product of sums mimics the inclusion-exclusion rule.

The expansion of the products of sums through partitions is demonstrated inductively as follows. Assume the product of the first $t - 1$ sums can be expressed as a sum over the full partition of the index set $I_{t-1} = \{i_1, ..., i_{t-1}\}$, in particular

$$\prod_{r=1}^{t-1} \left(\sum_{j\in s} x_{i_r j}\right) = \sum_{P_{t-1}\in\wp_{t-1}} X_{P_{t-1}}. \quad (17)$$

In (17) the term $X_{P_{t-1}}$ is the sum identified by the partition $P_{t-1} = (b_1 | ... | b_k)$, $k = 1, ..., t-1$. The blocks b_j indicate groups of variables with the same second index and so P_{t-1} induces an index set $J_k = \{j_1, ..., j_k\}$ of second indices. We can express $X_{P_{t-1}}$ as

$$X_{P_{t-1}} = \sum_{j_1 \neq \ldots \neq j_k \in s} \left(\prod_{j \in J_k} X_{b_j} \right), \tag{18}$$

where X_{b_j} is a product of x's defined by the block b_j that all have the same second index. To illustrate (18), consider, for example, the third term of (16). Here $P_{t-1} = (i_1 i_3 | i_2)$ and $J_2 = \{j, k\}$ so that in (18) the sum is taken over $j \neq k \in s$ and the multiplicands of the product are $X_{b_j} = x_{i_1 j} x_{i_3 j}$ and $X_{b_k} = x_{i_2 k}$. Returning to the general discussion, when either side of (17) is multiplied by $\sum_{j \in s} x_{i_t j}$ the product of the first t sums is obtained. Now the product $X_{P_{t-1}} \sum_{j \in s} x_{i_t j}$ can be expressed as

$$\sum_{j_1 \neq \ldots \neq j_k \in s} \left(\sum_{l=1}^{k} x_{i_l j_l} \prod_{j \in J_k} X_{b_j} \right) + \sum_{j_1 \neq \ldots \neq j_k \neq j_{k+1} \in s} \left(\prod_{j \in J_k} X_{b_j} x_{i_t j_{k+1}} \right). \tag{19}$$

The first term in (19) corresponds to the inclusion part of the rule and the second term in (19) corresponds to the exclusion part of the rule. When (19) is summed over all $P_{t-1} \in \wp_{t-1}$, the result will be a sum over the full partition of the first t indices given by I_t, *i.e.*, the sum over all $P_t \in \wp_t$.

Once the product of sums, $\prod_{r=1}^{m} \sum_{j \in s} x_{i_r j}$, is expanded into a sum of nested sums, the finite population expected value operator can be applied to each term so that the expected value of this product can be obtained. The expected value under simple random sampling without replacement of the product of sums results in a weighted sum of nested sums, with each sum taken over the finite population. We then wish to evaluate these nested sums.

In general we wish to evaluate the nested sum $\sum_{J_t} Y_{J_t}$ where J_t is the index set $\{j_1, \ldots, j_t\}$. The sum is taken over all $j_1 \neq \ldots \neq j_t$ with each $j_r = 1, \ldots, N$. The summand Y_{J_t} is the product $x_{i_1 j_1} x_{i_2 j_2} \ldots x_{i_t j_t}$. In the special case when $t = 3$ or $J_3 = \{j, k, l\}$ the nested sum can be written in terms of full sums as

$$\sum_{J_3} Y_{jkl} = \sum_{j \neq k \neq l = 1}^{N} Y_{jkl} = \sum_{j \neq k \neq l = 1}^{N} x_{i_1 j} x_{i_2 k} x_{i_3 l} =$$

$$2 \sum_{j=1}^{N} x_{i_1 j} x_{i_2 j} x_{i_3 j} - \sum_{j=1}^{N} x_{i_1 j} x_{i_2 j} \sum_{j=1}^{N} x_{i_3 j} - \sum_{j=1}^{N} x_{i_1 j} x_{i_3 j} \sum_{j=1}^{N} x_{i_2 j} -$$

$$\sum_{j=1}^{N} x_{i_1 j} \sum_{j=1}^{N} x_{i_2 j} x_{i_3 j} + \sum_{j=1}^{N} x_{i_1 j} \sum_{j=1}^{N} x_{i_2 j} \sum_{j=1}^{N} x_{i_3 j}. \tag{20}$$

Note that the full sums in the rightmost expression in (20) result from the full partition $\wp_3$ in (15). The order of the partitions in $\wp_3$ is the same as the order of the terms on the right of (20). The subscripts on the right of (20) denote the block membership in $\wp_3$. For example, the partition $(i_1 i_3 | i_2)$ corresponds to the term $\sum_{j=1}^{N} x_{i_1 j} x_{i_3 j} \sum_{j=1}^{N} x_{i_2 j}$ in (20). Note also from (20) that the determination of a nested sum is complicated by the additional determination of the appropriate coefficients of the full sums.

In general the evaluation of finite population nested sums results from the repeated application of the rule

$$\sum_{j_1 \neq \ldots \neq j_t = 1}^{N} \left(\prod_{r=1}^{t} x_{i_r j_r} \right) = \sum_{j_1 \neq \ldots \neq j_{t-1} = 1}^{N} \left[\prod_{r=1}^{t-1} x_{i_r j_r} \left(\sum_{j_t = 1}^{N} x_{i_t j_t} \right) \right]$$

$$- \sum_{j_1 \neq \ldots \neq j_{t-1} = 1}^{N} \left[\sum_{l=1}^{t=1} x_{i_t j_l} \left(\prod_{r=1}^{t-1} x_{i_r j_r} \right) \right]. \tag{21}$$

This expression mimics the inclusion-exclusion rule where the first set of sums on the right follows the exclusion part of the rule and the second set follows the inclusion part of the rule. Repeated application of (21) yields

$$\sum_{j_1 \neq \ldots \neq j_t = 1}^{N} \left(\prod_{r=1}^{t} x_{i_r j_r} \right) = \sum_{P_t \in \wp_t} (-1)^{|J_t| - |P_t|}$$

$$\times \left\{ \prod_{b_k \in P_t} \left[(|b_k| - 1)! \sum_{j=1}^{N} \left(\prod_{i_r \in b_k} x_{i_r j} \right) \right] \right\}$$

where $|J_t|, |P_t|$ and $|b_k|$ are the number of indices in J_t, the number of blocks in the single partition P_t and the number of elements in the block b_k respectively.

5. INTEGER PARTITIONS AND TAYLOR LINEARIZATION

Suppose that under some sampling design an estimator $\hat{\theta}$ of a parameter θ is of interest. The methodology described in §§2 to 4 may be used in moment calculations for $\hat{\theta}$ or to find unbiased estimators of these moments. Only in the simplest cases can this methodology be applied directly. Typically $\hat{\theta}$ must be linearized so that it becomes a polynomial function of sample means or sums which are $O_p(1)$ random variables with respect to the sampling design. Once $\hat{\theta}$ is linearized in this way the methodology of §§2 to 4 is applicable.

The objective of the linearization is to write $\hat{\theta}$ as an asymptotic expansion where terms descend in order by $1/\sqrt{n}$, specifically

$$\hat{\theta} = \hat{\theta}_0 + \hat{\theta}_1 / \sqrt{n} + \hat{\theta}_2 / n + \ldots, \tag{22}$$

where $\hat{\theta}_i$ is the coefficient of the $n^{-i/2}$ term. Typically $\hat{\theta}$ is a product of quantities that can also be expanded in this way. For example, if the measurement of interest is y and one auxiliary variable x is present then θ might be $M(y)$ and the auxiliary information available is $M(x)$ as well as x_j for $j \in s$. Then $\hat{\theta} = M(x)m(y)/m(x)$, the simple ratio estimator, is a product of three quantities $M(x), m(y)$ and $1/m(x)$ all having asymptotic expansions of their own. The expansion of $M(x)$ is itself. From (3) the expansion for $m(y)$ yields $M(y) + z(m(y))/\sqrt{n}$. The expansion for $1/m(x)$ results from (3) and then applying a Taylor expansion to $[M(x) + z(m(x))/\sqrt{n}]^{-1}$.

In general any expansion of a function with sufficient regularity can be found if operators are defined to expand a function, say $g(\dot{e})$ where $\dot{e}$ is itself an expansion. We are interested in expanding functions of the form

$$g(\dot{e}) = \prod_{j=1}^{p} \dot{e}_j. \tag{23}$$

where $\dot{e}_j$ itself has the expansion $\sum_{i=0}^{\infty} e_{ij} n^{-i/2}$. In linearizing $\hat{\theta}$ the basic requirement is to define an operator that returns $\hat{\theta}_i$ in (22). The efficiency of this operator derives solely from a rule for expanding functions of the form given in (23). The calculations required are functions of integer partitions. For example the $1/n$ term in the expansion of $\prod_{j=1}^{3} \dot{e}_j$ is

$$e_{21}e_{02}e_{03} + e_{01}e_{22}e_{03} + e_{01}e_{02}e_{23} + e_{11}e_{12}e_{13} + e_{11}e_{02}e_{13} + e_{01}e_{12}e_{13}. \tag{24}$$

Collecting first indices for each term in the sum results in a list in which each element sums to 2: {(2,0,0), (0,2,0), (0,0,2), (1,1,0), (1,0,1), (0,1,1)}. On noting that the order $n^{-i/2}$ term in any expansion $\dot{e}_j$ is actually the (i + 1)-th term in the sum $\sum_{i=0}^{\infty} e_{ij} n^{-i/2}$, we may modify the list derived from (24) so that entries identify the position of terms in a sum. The modification is to add 1 to each index value in the list. In the list derived from (25) this results in all partitions of the integer 5 into 3 blocks: {(3,1,1), (1,3,1), (1,1,3), (2,2,1), (2,1,2), (1,2,2)}. In general, the i-th term in the expansion of $\prod_{j=1}^{p} \dot{e}_j$ or $\dot{e}_j^{\,p}$, where p is a positive integer, is a sum over all partitions of the integer $i + p$ into p blocks. Consequently, using this methodology any term in the expansion of, for example, the ratio estimator can be found.

We illustrate this technique with ratio and regression estimation. The ratio estimator is given by

$$M(x)\,m(y)/m(x) \tag{25}$$

and the regression estimator by

$$k(y) + b\,[K(x) - k(x)] = k(y) + \frac{k(x,y)}{k(x,x)}[K(x) - k(x)] \tag{26}$$

in the notation of k-statistics.

On using (3) the ratio estimator (25) may be expressed as

$$M(x)\left[M(y) + \frac{z(y)}{\sqrt{n}}\right]\left[M(x) + \frac{z(x)}{\sqrt{n}}\right]^{-1}. \tag{27}$$

The expression in (27) may be expressed in terms of (24) with $p = 3$. The first term in (27) is the expansion $\sum_{i=0}^{\infty} e_{i1} n^{-i/2}$ with $e_{01} = M(x)$ and $e_{11} = e_{21} = \cdots = 0$. The first term in square brackets in (28) is the expansion $\sum_{i=0}^{\infty} e_{i2} n^{-i/2}$ where $e_{02} = M(y)$, $e_{12} = z(m(y))$ and $e_{22} = e_{32} = \cdots = 0$. The second term in square brackets is the expansion $\sum_{i=0}^{\infty} e_{i3} n^{-i/2}$ where $e_{i3} = (-1)^i \{z(m(y))\}^i / \{M(x)\}^{i+1}$. To get the $1/\sqrt{n}$ term in the expansion of (27), in which case $i = 1$ and $p = 3$, we need to find the integer partitions of 4 in blocks of 3. This yields the partitions (2,1,1), (1,2,1) and (1,1,2). On subtracting 1 from each index value in the list we obtain the list (1,0,0), (0,1,0), (0,0,1). Therefore the required term in the expansion is $(e_{11}e_{02}e_{03} + e_{01}e_{12}e_{03} + e_{01}e_{02}e_{13})/\sqrt{n}$ or equivalently $[z(m(y)) - M(y)z(m(x))/M(x)]/\sqrt{n}$. The $1/n$ term is obtained from (24) which reduces to

$$[M(y)\{z(x)\}^2/\{M(x)\}^2 - z(x)z(y)/M(x)]/n.$$

The regression estimator in (26) may be expressed as

$$K(y) + \frac{z(k(y))}{\sqrt{n}} + \left[K(x,y) + \frac{z(k(x,y))}{\sqrt{n}}\right] \times \left[K(x,x) + \frac{z(k(x,x))}{\sqrt{n}}\right]^{-1}\left[\frac{z(k(x))}{\sqrt{n}}\right] \tag{28}$$

using (3). The terms in the square brackets in (28) can be expanded in a similar fashion to the ratio estimator. In this case the terms in the expansions become: $e_{01} = K(x,y)$, $e_{11} = z(k(x,y))$ and $e_{21} = e_{31} = \cdots = 0$; $e_{i2} = (-1)^i \{z(k(x,x))\}^i / \{K(x,x)\}^{i+1}$ for $i = 0, 1, 2, \ldots$; and $e_{03} = 0$, $e_{13} = z(k(x))$ and $e_{23} = e_{33} = \cdots = 0$. Consequently, the $1/\sqrt{n}$ term in the expansion of the terms in the square brackets in (28) is

$$-\frac{K(x,y)z(k(x))}{K(x,x)\sqrt{n}}$$

and the $1/n$ term is

$$-\frac{1}{n}\left[\frac{z(k(x,y))}{K(x,x)} - \frac{K(x,y)z(k(x,x))}{K(x,x)^2}\right]z(k(x)).$$

These were obtained by the same argument that was used in the ratio estimator.

6. MACHINE APPLICATIONS TO THE CALCULATION OF EXPECTED VALUES OF SAMPLE STATISTICS AND THE DERIVATION OF UNBIASED ESTIMATORS

Since the machine application to the methodology described in §§3 to 5 was done in the programming language *Mathematica* we give a brief description of the operation of *Mathematica*. Then we describe the operators that were developed in *Mathematica* to provide a computer algebra for survey sampling theory.

Programming in *Mathematica* is carried out using expressions of the form $h[e_1, e_2, \ldots]$ where the object h is called the head of the expression and the e's are the elements of the expression. We have developed a number of machine expressions in *Mathematica* in the form of $h[e_1, e_2, \ldots]$ for operators which we apply to developing a computer algebra for sampling. All of these operators have been devised to

handle vectors as their arguments as well as scalars. There are four basic operators: $EV[\cdot]$ for expected value, $Cum[\cdot]$ for calculation of cumulants, $UE[\cdot]$ for unbiased estimator, and $Aexp[\cdot]$ for asymptotic expansion. There is also an operator to switch from notation using k-statistics to notation using means and vice versa.

The expected value operator $EV[\cdot]$ on sample statistics combines and carries out in *Mathematica* the three basic operations shown in the schema in (10). $EV[\cdot]$ contains two arguments, the first is the expression for which the expected value is to be obtained and the second is the sampling design which defines the inclusion probabilities. The application in *Mathematica* of $EV[\cdot]$ to $m(x_{i_1})m(x_{i_2})m(x_{i_3})$ under simple random sampling without replacement yields

$$K(x_{i_1})K(x_{i_2})K(x_{i_3}) + \frac{(N-n)(K(x_{i_1},x_{i_2})K(x_{i_3})}{Nn} + \frac{K(x_{i_1},x_{i_3})K(x_{i_2}) + K(x_{i_1})K(X_{i_2},x_{i_3}))}{Nn} + \frac{(N^2-3Nn+2n^2)K(x_{i_1},x_{i_2},x_{i_3})}{N^2n^2}$$

in the simplest expression of the output. Note that the result is a function of the full partition of $\{i_1,i_2,i_3\}$. If the operand is changed to $\{m(x_{i_1})-M(x_{i_1})\}\times\{m(x_{i_2})-M(x_{i_2})\}\times\{m(x_{i_3})-M(x_{i_3})\}$, application of $EV[\cdot]$ yields

$$\frac{(N^2-3Nn+2n^2)K(x_{i_1},x_{i_2},x_{i_3})}{N^2n^2},$$

which was obtained by Nath (1968) for particular values of the indices i_1, i_2 and i_3. In fact, the results in Nath (1968, 1969) for the products of three and four means and the exact results in Raghunandanan and Srinivasan (1973) for up to a product of eight means can all be reproduced automatically with the software that has been developed.

To this point the sampling design used in each of the examples has been simple random sampling without replacement. Results under general sampling designs can be obtained. We illustrate these results for the operator $Cum[\cdot]$ which is used to obtain the cumulants of an estimator. Note that the second cumulant for an estimator is also the variance. The operator $Cum[\cdot]$ has three arguments. The first is an expression for the estimator, the second is the order of the cumulant and the third is the sampling design. Under general sampling designs, estimators can be expressed in terms of $\sum\prod$ in the schema given by (10) and the $\sum\prod$ can be expanded to obtain $\sum\sum$, the middle term in (10). There is, however, no general simplification to obtain the final term in (10). This is illustrated with the Horvitz-Thompson estimator of $M(y)$ given by $(n/N)m(y/\pi)$ in the notation developed here. Application of the operator $Cum[\cdot]$ under a general sampling design to obtain the third cumulant of the Horvitz-Thompson estimator yields

$$2\frac{\left\{\sum_{i=1}^{N} y_i\right\}^3}{N^3} - 3\frac{\left\{\sum_{i=1}^{N} y_i\right\}\left\{\sum_{i=1}^{N}\frac{y_i^2}{\pi_i}\right\}}{N^3} - 3\frac{\sum_{i=1}^{N}\frac{y_i^3}{\pi_i^2}}{N^3} - 3\frac{\left\{\sum_{i=1}^{N} y_i\right\}\left\{\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\pi_{ij}y_iy_j}{(\pi_i\pi_j)}\right\}}{N^3} + 3\frac{\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\pi_{ij}y_iy_j^2}{(\pi_i\pi_j^2)}}{N^3} + \sum_{i=1}^{N}\sum_{h=1}^{N}\sum_{k=1}^{N}\frac{\frac{\pi_{ijk}y_iy_jy_k}{(\pi_i\pi_j\pi_k)}}{N^3}$$

where, for example, the term π_{ii} is the single inclusion probability π_i.

The operator $Aexp[\cdot]$ has two arguments, the function for which the expansion is required and the order of the expansion. This operator is used in combination with the $EV[\cdot]$ or $Cum[\cdot]$ operators to obtain approximate expectations or cumulants. This is illustrated in the case of the multiple linear regression estimator under simple random sampling without replacement. When there are q covariates the resulting regression estimator is given by

$$k(y) + b_{i_1}[K(x^{i_1}) - k(x^{i_1})] \tag{29}$$

using index and k-statistics notation. In (29) the coefficient b_{i_1} is the vector resulting from the product $k(x_{i_1},y)\,ik(x^{i_1},x_{i_2})$ in index notation, where the $q\times q$ array $ik(x_{i_1},x_{i_2})$ is the inverse of the $q\times q$ array given by $k(x_{i_1},x_{i_2})$. Similarly we will use $IK(x_{i_1},x_{i_2})$ to denote the inverse of the finite population array $K(x_{i_1},x_{i_2})$. Derivation of the mean square error of (29) requires Taylor expansions of the elements of b_{i_1} followed by the appropriate moment calculations and collection of terms. The *Mathematica* command to obtain the approximate variance of (29) is obtained by first applying $Aexp[\cdot]$ to (29) with 2 as the order in the expansion. Then the operator $Cum[\cdot]$ is applied to the result with the following arguments: the result from the asymptotic expansion as the estimator, simple random sampling as the design and 2 for the order of the cumulant. This yields

$$\frac{(N-n)K(y,y)}{Nn} + \frac{(-N+n)K(x_{i_1},y)K(x_{i_2},y)IK(x^{i_1},x^{i_2})}{Nn}$$

in index notation as output.

Estimation is achieved through the operator $UE[\cdot]$ which has two arguments, the estimand and the sampling design. For example, application of $UE[\cdot]$ to $\{M(x)\}^2$ under simple random sampling yields

$$\frac{(Nn)\{k(x)\}^2 + (N-n)k(x,x)}{Nn}.$$

If the estimand cannot be expressed as a sum of nested sums, but instead can be expressed as the root of an estimating function, then $UE[\cdot]$ obtains a consistent estimator.

7. DISCUSSION OF FUTURE WORK

The basic building blocks to develop a comprehensive computer algebra for survey sampling theory have been given. The foundation of this algebra is based on the enumeration of partitions. Fundamental operations under partition enumeration include the evaluation of nested sums and Taylor series expansions. Once these operations have been completed then expectations of sample statistics can be calculated or unbiased estimators of population quantities can be determined.

The next phase in this work is to extend the unistage results to multistage and multiphase sampling. In both multistage and multiphase sampling the problem reduces to the computer evaluation of multiple sums under an expectation operator or the determination of an unbiased estimator of multiple finite population sums. The problem of multistage sampling is currently under investigation. Another current area of inquiry is to extend the algebra to superpopulation models.

Once the basic algebra is in place then research problems involving algebraically complex sampling formulae can be easily investigated.

ACKNOWLEDGEMENTS

The authors are grateful to David Andrews for some useful discussions on this topic. This work was supported by grants from the Natural Sciences and Engineering Research Councils of Canada and by a research contract from Statistics Canada.

REFERENCES

ANDREWS, D.F., and STAFFORD, J.E. (1993). Tools for the symbolic computation of asymptotic expansions. *Journal of the Royal Statistical Society (B)*, 55, 613-628.

KENDALL, W.S. (1993). Computer algebra in probability and statistics. *Statistica Neerlandica*, 47, 9-25.

McCULLAGH, P. (1987). *Tensor Methods in Statistics*. New York: Chapman and Hall.

NATH, S.N. (1968). On product moments from a finite universe. *Journal of the American Statistical Association*, 63, 535-541.

NATH, S.N. (1969). More results on product moments from a finite universe. *Journal of the American Statistical Association*, 64, 864-869.

RAGHUNANDANAN, K., and SRINIVASAN, R. (1973). Some product moments useful in sampling theory. *Journal of the American Statistical Association*, 68, 409-413.

STAFFORD, J.E. (1996). A note on symbolic Newton-Raphson, submitted for publication.

STAFFORD, J.E., and ANDREWS, D.F. (1993). A symbolic algorithm for studying adjustments to the profile likelihood. *Biometrika*, 80, 715-730.

WISHART, J. (1952). Moment coefficients of the k-statistics in samples from a finite population. *Biometrika*, 39, 1-13.