

A COMPUTER ALGEBRA FOR SAMPLE SURVEY THEORY

J.E. Stafford and D.R. Bellhouse¹

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 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.

RÉSUMÉ

Un système de procédures qui peut être utilisé pour automatiser les calculs algébriques compliqués qui sont fréquemment rencontrés dans la théorie de l'échantillonnage est présenté. On démontre que trois techniques de base dans la théorie de l'échantillonnage dépendent de l'application répétée de règles qui donnent lieu à des partitions: soit le calcul de valeurs espérées dans un plan de sondage à un degré, la détermination des estimateurs sans biais dans ce contexte et le calcul des moments pour le calcul des séries de Taylor. La méthodologie est illustrée ici par des applications sur le calcul des moments pour la moyenne échantillonnale, l'estimateur par le quotient et l'estimateur par régression dans le cas de l'échantillonnage aléatoire simple sans remise. L'innovation présentée ici est que les calculs peuvent être faits instantanément et sans erreurs par un ordinateur, et sans dépendre des formules existantes qui peuvent être longues et ardues. Un boni supplémentaire réside dans le fait que les calculs peuvent se faire même là où il n'existe pas de formule. Le codage informatique qui a été développé pour cette méthodologie est disponible via "ftp anonymous" à l'adresse *fisher.stats.uwo.ca*.

MOTS CLÉS: Statistiques-K; partitions; moments mixtes; estimateurs par quotient et par régression; calcul symbolique; estimation de la variance.

1. INTRODUCTION

Consider a finite population of size N . A measurement of interest y_j is made on each unit $j, j=1, \dots, N$. In addition, a single auxiliary variable x_j or possibly a $q \times 1$ vector of auxiliary variables x_j may be taken on the units. The u -th entry of this vector x_j is x_{uj} , where $u=1, \dots, q$. Several kinds of finite population parameters may be defined on the measurements y_j , x_j , or x_j for $j=1, \dots, N$. We denote a finite population parameter of interest by T . Often T can be expressed as a smooth function of 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 by some sampling scheme. An estimator of T , given by t , is a smooth function of sample means and k -statistics.

In sampling theory, two general problems concern us. These are the calculation of moments of t and the determination of an unbiased estimator of T . The basic method to handle expectations and unbiased estimation

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

is to operate on sample and population nested sums respectively through inclusion probabilities denoted by π . A nested sum is a sum over a set of indices that are all unequal. The subscripts of the inclusion probability π show the units considered for inclusion. For example, π_{jkl} is the probability of including units j , k and l in the sample. In terms of triple nested sums, for example,

$$E\left[\sum_{J_3 \in S} x_{uj} x_{vk} x_{wl}\right] = \sum_{J_3=1}^N \pi_{jkl} x_{uj} x_{vk} x_{wl} \quad (1)$$

and

$$\sum_{J_3=1}^N x_{uj} x_{vk} x_{wl} \sim \sum_{J_3 \in S} x_{uj} x_{vk} x_{wl} / \pi_{jkl} \quad (2)$$

for designs of fixed sample size, where “ \sim ” denotes “estimated unbiasedly by” and where J_3 is the index set $\{j, k, l\}$ such that $j \neq k \neq l$. Parallel expressions may be established for with replacement sampling schemes.

Now the sum of any product of means can be expressed in terms of linear combinations of nested sums and vice versa. Schematically, this may be represented as

$$\Sigma\Pi \Rightarrow \Sigma\Sigma = \Sigma\Pi, \quad (3)$$

where $\Sigma\Pi$ denotes the sum of products and $\Sigma\Sigma$ denotes a sum of nested sums. If T or t can be expressed as a $\Sigma\Pi$ quantity, *i.e.*, a sum of products of means, then finding an unbiased estimator of T or moments of t reduces to following the schema in (3) and applying the appropriate operator, such as those given in (1) or (2), to $\Sigma\Sigma$, the middle step in the schema. If T or t are smooth functions of means but cannot be expressed directly as $\Sigma\Pi$ quantities, then an initial step is required before applying the schema in (3). For t , the initial step is to obtain a Taylor expansion of t . For T , the initial step is to obtain an estimating equation and then to linearize it. An alternative approach to the estimation of T is to obtain a consistent estimator by replacing each mean in T by its unbiased estimator.

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 $\Sigma\Pi$ in terms of $\Sigma\Sigma$, or $\Sigma\Sigma$ in terms of $\Sigma\Pi$ in the schema in (3). We also use integer partitions to obtain terms in the Taylor linearization of a function.

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 concept of partitioning is reviewed in section 3 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 section 4. It is shown in section 5 how the enumeration of partitions leads to the automatic calculation of expected values of products of sample means and k -statistics and to the derivation of unbiased estimators of products of finite population 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 handwritten 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

In order to avoid much cumbersome summation notation, we adapt the index notation of McCullagh (1987) to our purposes. For any j , the vector x_j contains q entries so that each of these x -variables may be associated with one of the q indices. Suppose $\{i_1, \dots, i_m\}$ is a subset of m of these q indices. When we deal with a small number of these indices we will use the letters u , v and w so that $i_1 = u, i_2 = v$, and $i_3 = w$. We reserve the subscripts j , k and l to represent finite population units. In our adaptation of McCullagh's notation, x_{uj} is now what we called the vector x_j .

Products of these indexed quantities become multidimensional arrays. For example, the expressions in (1) and (2) are all three-dimensional arrays of dimension $q \times q \times q$.

Let M denote a finite population mean. The argument of M shows the structure of the summand in the mean. For example, $M(y) = \sum_{j=1}^N y_j / N$ and $M(y^2)$ or equivalently $M(y^2) = \sum_{j=1}^N y_j^2 / N$. In index notation, for example,

$$M(x_u x_v x_w) = \frac{1}{N} \sum_{j=1}^N x_{uj} x_{vj} x_{wj} \quad (4)$$

is a three-dimensional array. An element of this array is the mean of products in one of the permutations of the q elements taken three at a time in x_{uj} . The sample mean is denoted by m so that, for example,

$$m(x_u x_v x_w) = \frac{1}{n} \sum_{j \in s} x_{uj} x_{vj} x_{wj}$$

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 ,

$$E[m(x_u x_v x_w)] = \frac{N}{n} M(x_u x_v x_w / \pi).$$

For the purpose of making asymptotic expansions, we define a standardized variable as the original variable centered about its expectation and scaled by $1/\sqrt{n}$. In particular, for any random variable X we let

$$z(X) = [X - E(X)]/\sqrt{n}. \quad (5)$$

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 x_{uj} 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

$$MGF(t) = 1 + \sum_{h=1}^{\infty} \mu_{i_1 \dots i_h} t^{i_1} \dots t^{i_h} / h! \quad (6)$$

The third order noncentral moment array, for example, is given in index notation by

$$\mu_{uvw} = \frac{\partial^3}{\partial t_u \partial t_v \partial t_w} MGF(t) |_{t=0}, \quad (7)$$

where $i_1 = u$, $i_2 = v$ and $i_3 = w$. By definition, the

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

$$CGF(t) = \sum_{h=1}^{\infty} \kappa_{i_1 \dots i_h} t^{i_1} \dots t^{i_h} / h! \quad (8)$$

is the cumulant generating function. In (8) the summation convention of McCullagh (1987) is again used. The three dimensional array of third order cumulants, for example, is given by

$$\kappa_{uvw} = \frac{\partial^3}{\partial t_u \partial t_v \partial t_w} CGF(t) |_{t=0}, \quad (9)$$

where, $i_1 = u$, $i_2 = v$ and $i_3 = w$.

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 $K(x_u, x_v, x_w)$ is the model-unbiased estimate of (9), where

$$K(x_u, x_v, x_w) = \frac{N}{(N-1)(N-2)} \times \sum_{j=1}^N [x_{uj} - M(x_u)][x_{vj} - M(x_v)][x_{wj} - M(x_w)]. \quad (10)$$

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,

$$k(x_u, x_v, x_w) = \frac{n}{(n-1)(n-2)} \times \sum_{j \in s} [x_{uj} - m(x_u)][x_{vj} - m(x_v)][x_{wj} - m(x_w)].$$

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 $j=1, \dots, N$; $K(x, y)$ is a second order k -statistic, in particular the finite population covariance between x and y .

3. 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 $I_m = \{i_1, \dots, i_m\}$. A single partition P_m of I_m divides the m indices into $k \leq m$ mutually exclusive and exhaustive subsets or blocks of I_m . We write $P_m = (b_1 | b_2 | \dots | b_k)$, where the b_1, \dots, 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 \mathcal{P}_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 $\mathcal{P}_1 = \{i_1\}$.
- ii. An inclusion-exclusion rule determines the contribution to \mathcal{P}_t by a partition $P_{t-1} \in \mathcal{P}_{t-1}$. 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, \dots, b_k which comprise P_{t-1} . If P_{t-1} has k blocks, k partitions for \mathcal{P}_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 $I_3 = \{u, v, w\}$ is given by the steps

- i. $\mathcal{P}_1 = \{(u)\}$
- ii. $\mathcal{P}_2 = \{(uv), (u|v)\}$ (11)
- iii. $\mathcal{P}_3 = \{(uvw), (uv|w), (uw|v), (u|vw), (u|v|w)\}$.

From step (i) to step (ii), the inclusion rule results in the partition (uv) and the exclusion rule results in $(u|v)$. From step (ii) to step (iii), the inclusion rule results in the creation of the partitions (uvw) , $(uw|v)$ and $(u|vw)$. The exclusion rule yields the partitions

$(uv|w)$ and $(u|v|w)$. 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_{ij} in terms of its cumulants. As in (7), we can identify the h -th moment array by differentiating $MGF(t)$ in (6) 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 $\exp\{CGF(t)\}$. In this case, the result is a sum that depends on the coefficients of $CGF(t)$ in (7). For example, we may write the first three moments in terms of cumulants as follows:

$$\begin{aligned}\mu_u &= \kappa_u \\ \mu_{uv} &= \kappa_{uv} + \kappa_u \kappa_v \\ \mu_{uvw} &= \kappa_{uvw} + \kappa_{uv} \kappa_w + \kappa_{uw} \kappa_v + \kappa_u \kappa_{vw} + \kappa_u \kappa_v \kappa_w.\end{aligned}$$

Now in each case, the result is a sum over the full partitions given in (11). These partitions arise since the multiplication rule for differentiation mimics the inclusion-exclusion rule for the enumeration of the full partition.

In sampling theory, 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 $\sum_{j \in S} x_{uj} \sum_{j \in S} x_{vj} \sum_{j \in S} x_{wj}$ can be expressed as

$$\begin{aligned}\sum_{j \in S} x_{uj} x_{vj} x_{wj} + \sum_{j+k \in S} x_{uj} x_{vj} x_{wk} + \sum_{j+k \in S} x_{uj} x_{vk} x_{wj} \\ + \sum_{j+k \in S} x_{uk} x_{vj} x_{wj} + \sum_{j+k+l \in S} x_{uj} x_{vk} x_{wl}.\end{aligned}\quad (12)$$

The result corresponds to the full partition of the indices $I_3 = \{u, v, w\}$ given by \mathcal{P}_3 in (11). The order of the partitions in \mathcal{P}_3 is the same as the order given for the terms in (12). For each partition in \mathcal{P}_3 , the variables in the same block have the same second index in the appropriate term in (12). For example, the partition $(uw|v)$ corresponds to the term $\sum_{j+k \in S} x_{uj} x_{vk} x_{wj}$ in (12). 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 $I_m = \{i_1, \dots, i_m\}$. As in (12), 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 $I_{t-1} = \{i_1, \dots, i_{t-1}\}$, in particular

$$\prod_{r=1}^{t-1} (\sum_{j \in S} x_{i_r j}) = \sum_{P_{t-1} \in \mathcal{O}_{t-1}} X_{P_{t-1}}. \quad (13)$$

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

$$X_{P_{t-1}} = \sum_{j_1^* \dots j_k \in S} (\prod_{j \in J_k} X_{b_j}) \quad (14)$$

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 (14), consider, for example, the third term of (12). Here $P_{t-1} = (uw|v)$ and $J_2 = \{j, k\}$ so that in (14) the sum is taken over $j \neq k \in S$ and the multiplicands of the product are $X_{b_j} = x_{uj} x_{wj}$ and $X_{b_k} = x_{vk}$. Returning to the general discussion, when either side of (13) 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

$$\begin{aligned} & \sum_{j_1^* \dots j_k \in S} (\sum_{l=1}^k x_{i_t j_l} \prod_{j \in J_k} X_{b_j}) \\ + & \sum_{j_1^* \dots j_k^* j_{k+1} \in S} (\prod_{j \in J_k} X_{b_j} x_{i_t j_{k+1}}). \end{aligned} \quad (15)$$

The first term in (15) corresponds to the inclusion part of the rule and the second term in (15) corresponds to the exclusion part of the rule. When (15) is summed over all $P_{t-1} \in \mathcal{O}_{t-1}$, 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 $P_t \in \mathcal{O}_t$.

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 $\{j_1, \dots, j_t\}$. The sum is taken over all $j_1 \neq \dots \neq j_t$ with each $j_r = 1, \dots, N$. The summand Y_{J_t} is the product $x_{i_1 j_1} x_{i_2 j_2} \dots x_{i_t j_t}$.

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

$$\begin{aligned} \sum_{J_3} Y_{jkl} &= \sum_{j^* k^* l=1}^N Y_{jkl} = \sum_{j^* k^* l=1}^N x_{uj} x_{vk} x_{wl} \\ &= \sum_{j=1}^N x_{uj} \sum_{j=1}^N x_{vj} \sum_{j=1}^N x_{wj} - \sum_{j=1}^N x_{uj} x_{vj} \sum_{j=1}^N x_{wj} \\ &\quad - \sum_{j=1}^N x_{uj} x_{wj} \sum_{j=1}^N x_{vj} - \sum_{j=1}^N x_{uj} \sum_{j=1}^N x_{vj} x_{wj} + 2 \sum_{j=1}^N x_{uj} x_{vj} x_{wj}. \end{aligned} \quad (16)$$

Note that the full sums in the rightmost expression in (16) result from the full partition \mathcal{O}_3 in (10). The order of the partitions in \mathcal{O}_3 is the same as the order of the terms on the right of (16). The subscripts on the right of (16) denote the block membership in \mathcal{O}_3 . For example, the partition $(uw|v)$ corresponds to the term $\sum_{j=1}^N x_{uj} x_{wj} \sum_{j=1}^N x_{vj}$ in (16). Note also from (16) 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

$$\begin{aligned} \sum_{j_1^* \dots j_t=1}^N (\prod_{r=1}^t x_{i_r j_r}) &= \sum_{j_1^* \dots j_{t-1}=1}^N [\prod_{r=1}^{t-1} x_{i_r j_r} (\sum_{j_t=1}^N x_{i_t j_t})] \\ &\quad - \sum_{j_1^* \dots j_{t-1}=1}^N [\sum_{l=1}^{t-1} x_{i_t j_l} (\prod_{r=1}^{t-1} x_{i_r j_r})]. \end{aligned} \quad (17)$$

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 (17) yields

$$\begin{aligned} \sum_{j_1^* \dots j_t=1}^N (\prod_{r=1}^t x_{i_r j_r}) &= \sum_{P_t \in \mathcal{O}_t} (-1)^{|J_t| - |P_t|} \\ &\quad \times \left\{ \prod_{b_k \in P_t} [(|b_k| - 1)! \sum_{j=1}^N (\prod_{i \in b_k} x_{i j})] \right\} \end{aligned} \quad (18)$$

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.

The operations carried out in finding an expected value according to the schema in (3) are illustrated in

the simple case of finding $E[m(x_u)^2]$ under simple random sampling without replacement. The first operation is to express

$$m(x_u)^2 = \frac{1}{n^2} \sum_{j \in s} x_{uj}^2 + \frac{1}{n^2} \sum_{j \neq k \in s} x_{uj} x_{uk} \quad (19)$$

using (13), (14) and (15). On applying inclusion probabilities $\pi_j = n/N$ and $\pi_{jk} = n(n-1)/[N(N-1)]$, the second operation yields

$$\frac{1}{n^2} \frac{n}{N} \sum_{j=1}^N x_{uj}^2 + \frac{1}{n^2} \frac{n(n-1)}{N(N-1)} \sum_{j \neq k=1}^N x_{uj} x_{uk}. \quad (20)$$

On using (18) so that $\sum_{j \neq k=1}^N x_{uj} x_{uk} = \sum_{j=1}^N x_{uj} \sum_{j=1}^N x_{uj} - \sum_{j=1}^N x_{uj}^2$, the third operation yields

$$E(m(x_u)^2) = \frac{N(n-1)}{(N-1)n} M(x_u)^2 + \frac{N-n}{n(N-1)} M(x_u^2). \quad (21)$$

In (21), $M(x_u) = K(x_u)$ and $M(x_u^2) = [N/(N-1)]K(x_u, x_u) + K(x_u)K(x_u)$ so that (21) can be reexpressed as

$$E(m(x_u)^2) = K(x_u)^2 + (N-n)K(x_u, x_u)/(Nn). \quad (22)$$

In general, partitions may be used to express means in terms of k -statistics and vice versa.

Following the schema in (3), the operations for finding an unbiased estimator of, for example, $M(x_u)^2$ is similar to (19), (20) and (21). The estimand $M(x_u)^2$ is expressed in nested sums similar to (19). These sums will be nested finite population sums. Similar to (20), the inclusion probabilities are applied. In this case, the finite population sums are replaced by sample sums and the summand is divided by the appropriate inclusion probability. Finally, similar to (21) the resulting nested sample sums are expressed as products of sums.

4. INTEGER PARTITIONS AND TAYLOR LINEARIZATION

Suppose that under some sampling design, an estimator t of a parameter T is of interest. The methodology described in sections 2 and 3 may be used in moment calculations for t or to find unbiased estimators of these moments. Only in the simplest cases can this methodology be applied directly. Typically t 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 t is linearized in this way, the

methodology of sections 2 and 3 is applicable.

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

$$t = t_0 + t_1/\sqrt{n} + t_2/n + \dots, \quad (23)$$

where t_i is the coefficient of the $n^{-i/2}$ term. Typically, t 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 T might be $M(y)$ and the auxiliary information available is $M(x)$ as well as x_j for $j \in s$. Then $t = M(x)m(y)/m(x)$, 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 (5), the expansion for $m(y)$ yields $M(y) + z(m(y))/\sqrt{n}$. The expansion for $1/m(x)$ results from (5) 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

$$g(\dot{e}) = \prod_{j=1}^p \dot{e}_j \quad (24)$$

where \dot{e}_j itself has the expansion $\sum_{i=0}^{\infty} e_{ij} n^{-i/2}$. In linearizing t , the basic requirement is to define an operator that returns t_i in (23). The efficiency of this operator derives solely from a rule for expanding functions of the form given in (24). 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

$$e_{21}e_{02}e_{03} + e_{01}e_{22}e_{03} + e_{01}e_{02}e_{23} + e_{11}e_{12}e_{13} + e_{11}e_{02}e_{13} + e_{01}e_{12}e_{13}. \quad (25)$$

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 $n^{-1/2}$ 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 (25) 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

$$M(x)m(y)/m(x) \quad (26)$$

and the regression estimator by

$$k(y)+b[K(x)-k(x)]=k(y)+\frac{k(x,y)}{k(x,x)}[K(x)-k(x)] \quad (27)$$

in the notation of k -statistics.

On using (5), the ratio estimator (26) may be expressed as

$$M(x)\left[M(y)+\frac{z(y)}{\sqrt{n}}\right]\left[M(x)+\frac{z(x)}{\sqrt{n}}\right]^{-1} \quad (28)$$

The expression in (28) may be expressed in terms of (24) with $p=3$. The first term in (28) is the expansion $\sum_{i=0}^{\infty}e_{i1}n^{-i/2}$ with $e_{01}=M(x)$ and $e_{11}=e_{21}=\dots=0$. 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 $e_{22}=e_{32}=\dots=0$. 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 (28), 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 (25) which reduces to $[M(y)z(x)^2/M(x)^2-z(x)z(y)/M(x)]/n$.

The regression estimator in (27) may be expressed as

$$\left[K(y)+\frac{z(k(y))}{\sqrt{n}}+\left[K(x,y)+\frac{z(k(x,y))}{\sqrt{n}}\right]\right] \times \left[K(x,x)+\frac{z(k(x,x))}{\sqrt{n}}\right]^{-1} \left[\frac{z(k(x))}{\sqrt{n}}\right] \quad (29)$$

using (5). The terms in the square brackets in (29) 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 $e_{21}=e_{31}=\dots=0$; $e_{i2}=(-1)^i z(k(x,x))^i / K(x,x)^{i+1}$ for $i=0, 1, 2, \dots$;

and $e_{03}=0$, $e_{13}=z(k(x))$ and $e_{23}=e_{33}=\dots=0$. Consequently, the $1/\sqrt{n}$ term in the expansion of the terms in the square brackets in (29) is

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

and the $1/n$ term is

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

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

5. 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 sections 3 and 4 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 $h[e_1, e_2, \dots]$ where the object h is called the head of the expression and the e 's are the elements of the expression. Note that all *Mathematica* commands and output are written in a different font here to distinguish them from the regular text. Expressions may be built into the *Mathematica* programming language or may be defined by the user. Either type of expression begins with an upper case letter. For example, two built-in expressions which we use frequently here are `Expand []` and `Simplify []`. The expression `Expand []` expands out products and powers of the element in the square brackets and the expression `Simplify []` performs some algebraic manipulations on the element in the square brackets and returns the simplest form it finds.

We have developed a number of machine expressions in *Mathematica* which we apply to developing a computer algebra for sampling. In these expressions, objects associated with the finite population are denoted by the inclusion of the letter "p" and those associated with the sample by "s". For example, the population and sample sizes are given the heads N_p and N_s respectively. Here is a list of heads

for the calculation of certain sample and population statistics:

$Mp[x]$ is the population mean of the element x in the expression.

$Ms[x]$ is the sample mean of the element x in the expression.

$Kp[x, y, \dots]$ is the population k -statistic. The order of the k -statistic is given by the number of elements which are separated by commas.

$Ks[x, y, \dots]$ is the sample k -statistic.

$Z[x]$ takes a variable x , centers it about its expectation and scales it by $1/\sqrt{Ns}$ so that it is bounded in probability.

All of these machine expressions have been devised to handle vectors as their arguments as well as scalars. The vector x_{ij} is represented by $x[u]$ in machine notation. Then, for example, $Ms[x[u] x[v]]$ is a two-dimensional array containing the sample means of the products of all pairs of measurements from the vector x_{ij} for $j \in s$.

We have also developed operators to translate notation from means to k -statistics and vice versa. The translation is based on the partitioning algorithms of section 3. The translation operators are as follows.

$MpToKp$ added to an expression translates the result from population means to k -statistics.

$MsToKs$ does the same thing for sample means to k -statistics.

A notational filter called "Notation[x]" has been developed to translate the machine expression of the argument x into one suitable for display on the screen or on paper. For ease of reading, for example, it would be preferable to see N and n in the output instead of Np and Ns . Table 1 provides some examples of inputs and outputs for notational filters and translation functions.

We have defined an expected value operator "EV" of sample statistics in *Mathematica* which combines and carries out three basic operations shown in the schema in (3). In *Mathematica*, the operation to obtain, for example, the expectation operation in (21) is carried out with the command:

```
Simplify[Notation[Expand[EV[Ms[x[u]]
^2, srs]]]]]
```

Table 1
Example Inputs and Outputs for Notation and Translations Functions

Input	Output
Notation[Np]	N
Notation[Ns]	n
Notation[Mp[x y]]	$M(xy)$
Notation[Ms[x[u] x[v]]]	$m(x_u, x_v)$
Notation[Kp[x[u] y]]	$K(x_u, y)$
Notation[Ks[x, y] /. Kp->KpToMp]	$\frac{n}{n-1} (m(xy) - m(x)m(y))$
Notation[Mp[x] /. Mp->MpToKp]	$K(x)$
Notation[Ms[x ^2]]	$M(xx)$
Notation[Mp[x ^2] /. Mp->MpToKp]	$\frac{N-1}{N} K(x, x) + K(x)^2$
Notation[Z[x y]]	$z(xy)$

Note that $EV []$ 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 operation in *Mathematica* results in the output:

$$\frac{(Nn-N) M(x_u)^2 + (N-n) M(x_u, x_u)}{Nn-n}$$

In order to reexpress (21) as (22), or means in terms of k -statistics, the *Mathematica* command

```
Simplify[Notation[Expand[EV[Ms[x[u]]
^2, srs] /. Mp->MpToKp]]]
```

is used. This results in the *Mathematica* output

$$\frac{NnK(x_u)^2 + (N-n) K(x_u, x_u)}{Nn}$$

In a slightly more complicated example, application of the expected value operator and the notation operator together to evaluate $E(m(x_u)m(x_v)m(x_w)))$ in the command

```
Simplify[Notation[Expand[EV[Ms[x[u]]
Ms[x[v]]Ms[x[w]], srs] /. Mp->MpToKp]]]
```

yields

$$K(x_u) K(x_v) K(x_w) + \frac{(N-n) (K(x_u, x_v) K(x_w) + K(x_u, x_w) K(x_v) + K(x_u) K(x_v, x_w))}{Nn} + \frac{(N^2 - 3Nn + 2n^2) K(x_u, x_v, x_w)}{N^2 n^2}$$

as output. Note that the result is a function of the full partition of $I_3 = \{u, v, w\}$. If the expected value is changed to

$$E[\{m(x_u) - M(x_u)\}\{m(x_v) - M(x_v)\}\{m(x_w) - M(x_w)\}],$$

the appropriate *Mathematica* command yields

$$\frac{(N^2 - 3Nn + 2n^2) K(x_u, x_v, x_w)}{N^2 n^2},$$

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

The use of the EV operator may not always be convenient for obtaining some moment calculations. Consequently, we have defined an operator "Cum" which determines the finite population cumulant of the argument under the operator to a given order. A finite population cumulant may be defined through the finite population noncentral moments using the standard relationship between moments and cumulants illustrated in equations (6) through (9) and accompanying discussion. The Cum operator has three arguments: the estimator, the order of the cumulant and the sampling design. The two *Mathematica* commands

```
Notation[Expand[EV[(Ms[y]-Mp[y])
^2,srs] /. Mp->MpToKp]]
```

and

```
Notation[Cum[Ms[y],2,srs]]
```

result in the same output

$$\frac{(N-n) K(y, y)}{Nn}$$

for the variance of the sample mean under simple random sampling without replacement.

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. For these general designs, the estimators can be expressed in terms of $\Sigma\Pi$ in the schema given by (3) and the $\Sigma\Pi$ can be expanded to obtain $\Sigma\Sigma$, the middle term in (3). There is no general simplification to obtain the final term in (3). In order to

express this final term in (3), we introduce the *Mathematica* notation $S []$ for single or multiple sums. The arguments of S will contain subscripts given, by i, j or k and so on where values of the subscripts range over the N finite population unit numbers. The function S then denotes a multiple sum taken over the subscripts in the argument. This is illustrated with the Horvitz-Thompson estimator of $M(y)$ given by $(n/N)m(y/\pi)$ in the notation developed here. The *Mathematica* command to obtain the variance of the Horvitz-Thompson estimator is given by

```
Notation[Cum[Ns/NpMs[y/pi],2,gen]] (30)
```

where pi denote the inclusion probability π . This command yields

$$-\left(\frac{S[y_i]^2}{N^2}\right) + \frac{S\left[\frac{pi_{ij} y_i y_j}{pi_i pi_j}\right]}{N^2}.$$

Note that in the double sum in the second term, pi_{ii} is the single inclusion probability π_i . The third order cumulant or equivalently the third central moment of the Horvitz-Thompson estimator is obtained upon substituting the order 3 for 2 in the command in (30). This command in *Mathematica* yields

$$\begin{aligned} & \frac{2S[y_i]^3}{N^3} - \frac{3S[y_i]S\left[\frac{y_i^2}{pi_i}\right]}{N^3} - \frac{3S\left[\frac{y_i^3}{pi_i^2}\right]}{N^3} \\ & - \frac{3S[y_i]S\left[\frac{pi_{ij} y_i y_j}{pi_i pi_j}\right]}{N^3} \\ & + \frac{3S\left[\frac{pi_{ij} y_i y_j^2}{pi_i pi_j^2}\right]}{N^3} + \frac{S\left[\frac{pi_{ijk} y_i y_j y_k}{pi_i pi_j pi_k}\right]}{N^3}. \end{aligned}$$

For treating nonlinear estimators we define an operator AExp [] which finds the asymptotic expansion to a required order along the lines described in section 4 on integer partitions. The operator AExp has two arguments, the function for which the expansion is required and the order of the expansion. Several steps in obtaining moments of a nonlinear estimator can be put into one command so that the algebraic burden is minimized. For example, consider the ratio estimator given in (26) under simple random sampling. The expected value of this estimator to order

1/n may be obtained from the *Mathematica* command

```
Notation[EV[AExp[Mp[x]Ms[y]/Ms[x],
                2],srs]/. Mp->MpToKp].
```

The resulting output is

$$K(y) + \frac{(-n+N) K(y) K(x, x)}{nNK(x)^2} + \frac{(n-N) K(x, y)}{nNK(x)}$$

The case of the multiple linear regression under simple random sampling without replacement may now be considered. When there are q covariates, the resulting regression estimator is a generalization of the estimator in (27). In particular, the multiple linear estimator is given by

$$k(y) + b_u [K(x^u) - k(x^u)] \quad (31)$$

using index and k -statistics notation. In (31), the coefficient b_u is the vector resulting from the product $k(x_u, y) ik(x^u, x_v)$ in index notation, where the $q \times q$ array $ik(x_u, x_v)$ is the inverse of the $q \times q$ array given by $k(x_u, x_v)$. Similarly we will use $IK(x_u, x_v)$ to denote the inverse of the finite population array $K(x_u, x_v)$. Derivation of the mean square error of (31) requires Taylor expansions of the elements of b_u followed by the appropriate moment calculations and collection of terms. In *Mathematica*, we can define the multiple regression estimator through an estimating equation. This is done through two functions $f[]$ and $g[]$ in the following code:

```
g[b[u]]:=Ms[y]+b[u](Mp[x[u]]-Ms[x[u]])
f[b[u]]:=Ks[x[u],x[v]]b[u]-Ks[x[u],y]
RegEst=g[Root[f]]
```

where the function `Root[]` finds the root of a function using integer partitions as described in Stafford (1996). The expectation of this estimator to order $1/n$ is obtained from the command

```
Notation[Cum[AExp[RegEst,2],1,srs]].(32)
```

The result of this command is

$$\frac{(Nq+Nn-qn) K(y)}{Nn} + \frac{q(-N+n) IK(x_u, x_v) K(x^u) K(x^v, y)}{Nn} + \frac{(N-n) IK(x_u, x_v) IK(x_u, x_2) K(x^u, y) K(x^u, x^v, x^2)}{Nn} + \frac{(-N+n) IK(x_u, x_v) K(x^u, x^v, y)}{Nn}$$

The variance of the regression estimator to order $1/n$ on replacing the number 1 in (32) by the number 2. This results in the following *Mathematica* output

$$\frac{(N-n) K(y, y)}{Nn} + \frac{(-N-n) K(x_u, y) K(x_v, y) IK(x^u, x^v)}{Nn}$$

Note that the mean square error and the variance of the regression estimator are the same to order $1/n$.

As noted already, the steps to obtain an unbiased estimate of population product moments or k -statistics are similar to those for finding expected values. To carry out these steps in one operation, we have defined an unbiased estimator operator "UE" of population statistics in *Mathematica*. As in the situation of finding expected values, the function to find unbiased estimators, `UE[]`, contains two arguments. The first is a function of product moments or k -statistics for which the unbiased estimator is required. The second is the sampling design under which the estimator is to be unbiased.

This simplest example is to obtain the unbiased estimate of a finite population mean, say $\bar{X} = M(x)$ in the univariate case. This is obtained from the *Mathematica* command

```
Notation[UE[Mp[x],srs]]
```

which results in the simple output statement

$$m(x)$$

The unbiased estimator for $M(x)^2$ in the univariate case is obtained from the command

```
Simplify[Notation[Expand[UE[Mp[x]
                        ^2,srs]/.Ms->MsToKs]]]
```

for which *Mathematica* returns

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

6. DISCUSSION AND 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 sample sums under an expectation operator or the determination of an unbiased estimator of multiple population sums. The problem of multistage sampling is currently under investigation. Another possible 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 may easily be 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.
- Raghunandan, 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.