DOCUMENT RESUME

ED 206 671	TM 810 566
AUTHOR	Folsom, Ralph E., Jr.
TITLE	National Assessment Approach to Sampling Error Estimation. Sampling Error Monograph.
INSTITUTION	Research Triangle Inst., Research Triangle Park, N.C.
SPONS AGENCY	Education Commission of the States, Denver, Colo. National Assessment of Educational Progress.; National Center for Education Statistics (DHEW), Washington, D.C.
REPORT NO	RTI-250-796-5
PUB DATE	Apr 77
CONTRACT	OEC-0-74-0506
NOTE	100p.; Best copy available.
EDRS PRICE	MF01/PC04 Plus Postage.
DESCRIPTORS	*Educational Assessment; Elementary Secondary Education; Error of Measurement; National Competency Tests; *Research Methodology; *Research Problems; *Sampling; Statistical Bias; Testing Programs
IDENTIFIERS	*National Assessment of Educational Progress; *Sampling Error

ABSTRACT

Beginning with the planning stages of the National Assessment of Educational Progress (NAEP), careful attention has been given to the design of efficient probability sampling methods for the selection of class-age respondents and the assignment of test packages. With these methods, it is possible for NAEP researchers to make relatively precise statements about population characteristics on the basis of fairly small samples. The purpose of this monograph is to describe what is meant by relatively precise statements about population characteristics and to show how NAEP sample data are being used to gauge the accuracy of reported results. The levels of precision for Year 01 and 02 were compared, and the overall precision was improved in Year 02. The sampling error methodology developed for the Year 02 sample was applied to calculate Year 03 and 04 sampling errors. A major redesign of NAEP's primary sample was initiated for the Year 05 assessment. The Year 06 (1974-75) NAEP in-school primary sample was an independent replicate of the Year 05 sample selected from the deeply stratified primary unit frame developed for the 1973-74 survey. Four non-overlapping samples were to be used successively for Years 07 through 10. Primary type of information provided by report: Procedures (Sampling) (Evaluation). (BW)

RESEARCH TRIANGLE INSTITUTE

Sampling Error Monograph
25U-796-5

NATIONAL ASSESSMENT APPROACH TO SAMPLING ERROR ESTIMATION

by

Ralph E. Folsom, Jr.

Prepared for
National Assessment of Educational Progress

Revised April 197

CONTENTS

LIST OF TABLES

Chapter 1: INTRODUCTION

National Assessment Overview

The National Assessment of Educational Progress (NAEP) can be viewed as an annual series of large-scale sample surveys designed to measure the educational achievements of four age groups in 10 subject areas. The four specific age groups include all 9-year-olds and 13-year-olds enrolled in school at the time of assessment and all 17-year-olds and young adults, ages 26-35. The first assessment of science, writing, and citizenship spanned the 1969-70 school year. Subsequent assessments have been conducted during the 1970-71, 1971-72, and 1972-73 school years. At this writing, the Year 05 assessment of Career and Occupational Development (COD) and Writing is underway, and planning for the Year 06 assessment has begun.

National Assessment respondents answer questions and perform tasks much the same as they would on a typical achievement test. One aspect of National Assessment that distinguishes it from the typical educational testing program is the way data are reported. Instead of calculating test scores for each respondent and forming normative distributions, results on each released exercise are reported separately. Unreleased exercises are held back for reassessment in subsequent years so that trend measurements will not be biased by school systems teaching to specific NAEP exercises. The reporting of separate exercises takes the form of estimated proportions responding correctly within various subgroups of the target population. Group effects that contrast the proportion of correct answers for a specific subgroup against the corresponding national proportion are used to detect variations in knowledge, understanding, skills, and attitudes among various segments of the population. With this method of reporting, it is not necessary for each respondent to complete the entire set of exercises. Subsets of exercises, called packages, are formed, which take approximately 50 minutes each to complete. If 10 such packages are formed for a particular age-class assessment, then 10 nonoverlapping samples, each representative of the target population, are specified and assigned a particular package.

Beginning with the early planning stages of National Assessment, careful attention has been given to the design and implementation of efficient *probability* sampling methods for the selection of age-class respondents and the assignment of packages. With these methods, it is possible for NAEP researchers to make relatively precise statements about relevant population characteristics on the basis of fairly small samples. The purpose of this monograph is to describe what is meant by relatively precise statements about population characteristics and to show how National Assessment sample data are being used to gauge the accuracy of reported results.

Population Characteristics and Sample Statistics

While statisticians and other researchers familiar with survey methods are well aware of the inferential "leap" that is made when sample-based results are taken to represent population facts, many users of sample data do not readily distinguish between population parameters and sample statistics. It is the researcher's obligation, therefore, to point out that his survey results are an imperfect approximation of the truth, an approximation whose accuracy is limited by his financial resources and his sample survey skills. The sources of error that plague survey results are numerous. Many of these error sources--such as unuseable responses to vague or sensitive questions; no response from particular sample members; and errors in coding, scoring, and processing the data--are beyond the control of the sampling statistician. The nonsampling errors are also common to complete enumerations of a target population, such as the U.S. Decimal Census. One advantage of a small sample survey over a complete enumeration, in addition to the obvious cost savings, is that a smaller, more highly trained, and supervised field force followed up by careful scoring and processing of the small sample data may produce fewer nonsampling errors per respondent than the large unwieldy census operation.

In addition to poor response, nonresponse, scoring, and processing errors, sample survey results are innaccurate precisely because they are based on a sample and not on the entire population. Consider, for example, the population percentage of 9-year-olds who can answer a particular science exercise correctly. For a specified sample design and selection procedure, a very large number of possible samples could be realized. Supose that s = 1,2,... S indexes the totality of possible samples that could be drawn in accordance with a specified procedure. A *probability* sampling method is distinguished by the fact that each sample-s has a known nonzero probability of being selected. If we denote this probability of selection by $\pi(s)$ and let $\hat{P}(s)$ denote the sample-s estimate for the percentage of 9-year-olds who can answer correctly, then

$$E\{\hat{P}(s)\} = \sum_{s=1}^{S} \pi(s)\ \hat{P}(s) \qquad (1.1)$$

is the expectation, or expected value, of the sample statistic $\hat{P}(s)$. This expectation represents the average value of the estimates $\hat{P}(s)$ over a conceptually infinite sequence of repeated sample draws with $\pi(s)$ denoting the frequency of occurrence for sample-s. If this expected value does not equal the population parameter of interest, say P, then $\hat{P}(s)$ is said to be a biased estimate of P. The magnitude of this bias is specified by

$$\text{Bias}\ \{\hat{P}(s)\} = [E\{\hat{P}(s)\} - P]. \qquad (1.2)$$

Bias in a sample statistic may be attributed to nonsampling as well as to sampling sources; that is, statistics that would otherwise average out to the true population value can miss the mark if nonresponse, measurement, or processing errors are made. In the absence of nonsampling errors, probability samples provide for unbiased estimation of population totals like the numerators and denominators of NAEP P-values. On the other hand, strictly unbiased estimates for ratios of population totals are often unavailable. The sampling biases associated with ratio estimates are generally negligible when large-scale probability samples are involved. Some empirical evidence for this contention is presented in chapter 4, where the sampling biases of NAEP P-values are studied.

Besides the systematic errors that cause the sample estimate to miss the mark on the average, one must also recognize that it is possible to hit the target on the average while missing the bull's-eye substantially in some samples. To quantify these random sampling fluctuations, statisticians have defined the sampling variance of $\hat{P}(s)$ as

$$\text{Var}\ \{\hat{P}(s)\} = \sum_{s=1}^{S} \pi(s)\ [\hat{P}(s) - E\{\hat{P}(s)\}]^2. \qquad (1.3)$$

This quantity represents the squared distante between the sample values and their expectation or centroid averaged over an infinite sequence of sample draws. A more appropriate measure of sample dispersion for a biased estimator is the mean squared error, a weighted average of squared differences between sample values and the true population value P:

$$\text{MSE}\ \{\hat{P}(s)\} = \sum_{s=1}^{S} \pi(s)\ [\hat{P}(s) - P]^2\ . \tag{1.4}$$

The mean squared error of a sample statistic has an obvious relationship to its bias and variance; namely,

$$\text{MSE}\ \{\hat{P}(s)\} = \text{Bias}^2\ \{\hat{P}(s)\} + \text{Var}\ \{\hat{P}(s)\}\ . \tag{1.5}$$

The quantity most commonly used to characterize the sampling variation of a statistic is called the standard error or $\text{SE}\{\hat{P}(s)\}$, where

$$\text{SE}\ \{\hat{P}(s)\} = [\text{Var}\ \{\hat{P}(s)\}]^{1/2}\ . \tag{1.6}$$

An analogous quantity for biased statistics is

$$\text{TE}\ \{\hat{P}(s)\} = [\text{MSE}\ \{\hat{P}(s)\}]^{1/2} \tag{1.7}$$

often called the "total error" or root mean squared error.

It is apparent from the definitions in equations 1.3 through 1.7 that the true value of these sampling error measures cannot be determined from a single sample. It is possible, however, to produce valid estimates of these quantities using the data obtained from a well-designed probability sample. Probability samples which provide for estimating the sampling variability, ordinarily the standard errors, of sample statistics have been called measurable [ref. 1]. Examples of nonmeasurable probability samples include systematic random selections from lists exhibiting periodicity and stratified random samples with a single unit selected per stratum.

National Assessment is committed to the design of measurable samples, samples which provide for reasonably valid estimates of standard errors. These standard errors, used in connection with respected statistical conventions, make it possible to bridge the gap between sample estimates and population facts. A statistical framework for inferring population P-values and for inferring group effects from sample effects is outlined in the following section.

Statistical Inference

Confidence Intervals

When one makes inference from a sample about the magnitude of a population parameter, like P, by quoting a sample estimator $\hat{P}(s)$, it is common statistical practice to include a range or interval of values about $\hat{P}(s)$ which is likely to contain the true population value P. Such intervals are commonly called "confidence intervals" in the statistical literature; they frequently take the form

$$I_p(s) = \hat{P}(s) \pm k \text{ se } \{\hat{P}(s)\} \qquad (1.8)$$

where k is a constant and se $\{\hat{P}(s)\}$ is the estimated standard error for the sample statistic $\hat{P}(s)$. The "confidence coefficient" associated with such an interval is the probability that a randomly selected sample will yield an interval $I_p(s)$ that includes the true population value P. Recalling that we have S possible samples which are realized with probabilities $\pi(s)$, this confidence coefficient can be specified by defining A(P) as the set of samples where the interval $I_p(s)$ contains P and letting

$$\gamma(P) = \sum_{s \varepsilon A(P)} \pi(s) \qquad (1.9)$$

denote the probability that the interval associated with a randomly selected sample will contain P. Notice that the summation in equation 1.9 extends over all samples s which belong to the set A(P) [$s \varepsilon A(P)$ denotes s belonging to A(P)]. In empirical terms, this probability statement means that, in a conceptually infinite sequence of repeated sample draws, a fraction $\gamma(P)$ of the corresponding intervals will contain P.

In order to specify a value of k in equation 1.8 that will yield an interval with given confidence coefficient $\gamma(P)$, one must know the sampling distribution of the standardized variable

$$t(s) = [\hat{P}(s) - P]/\text{se } \{\hat{P}(s)\} . \qquad (1.10)$$

Notice that the set A(P) of samples with $I_p(s) \varepsilon P$ [$I_p(s)$ containing P] is equivalent to the set of samples with $|t(s)| \leq k$. It is clear that the sampling distribution of t(s) cannot be specified exactly without a complete enumeration of the target population. To pursue this line of inference, sampling statisticians commonly assume that the sampling distribution of t(s) can be approximated by Student's T distribution with (df) "degrees of freedom" or by the standard normal distribution when df exceeds 60. The rationale for this assumption rests on the tendency of statistics like $\hat{P}(s)$ from large probability samples to have normal-like sampling distributions. With $\hat{P}(s)$ approximately normal, the sampling distribution of t(s) will resemble Student's T with the appropriate degrees of freedom.

For a stratified multistage sample with a total of n primary sampling units (PSUs) selected from H primary strata, the degrees of freedom associated with t(s) can be approximated by $\hat{df} = (n-H)$. Some authors have recommended a more sophisticated approximation for df attributed to Satterthwaite [ref. 2]. Satterthwaite's approximation attempts to account for unequal within-stratum variance components and varying stratum sample sizes. The results of some recent empirical studies summarized in chapter 4 of this monograph seem to indicate that the naive approximation for $\hat{df}$, namely $\hat{df} = (n-H)$, is to be preferred.

A further characterization of a γ(P) confidence interval can be made in terms of its so-called Operating Characteristic (OC) curve. This OC curve summarizes the probabilities that points P* other than the true value P will be included in the interval corresponding to a randomly selected sample. If we let $\gamma(P^*) = \Pr\{I_p(s) \epsilon P^*\}$ where $I_p(s)$ has the form in equation 1.8 then

$$\gamma(P^*) = \Pr\{|t(s,\Delta^*)| \le k\} \tag{1.11}$$

where

$$\begin{aligned} t(s,\Delta^*) &= [\hat{P}(s) - P^*]/se\{\hat{P}(s)\} \\ &= t(s) + (P-P^*)/se\{\hat{P}(s)\} \\ &= t(s) + \Delta^*/se\{\hat{P}(s)\} \end{aligned}$$

has the form of Student's noncentral T statistic with df degrees of freedom and noncentrality parameter $\delta^* = \Delta^*/SE\{\hat{P}(s)\}$. For values of P* deviating considerably from the true value P, one would hope that γ(P*) would be small.

It is important to note at this point that, for a given sample design and an estimation scheme characterized by $SE\{\hat{P}(s)\}$ and the degrees of freedom-df associated with $se\{\hat{P}(s)\}$, the entire OC curve is specified once k is set. With this in mind, it is clear from equation 1.11 that, while an increase in k will raise the confidence level, $\gamma(P)$, of the associated interval, it will also inflate the probability of including unwanted values. Another way of viewing this relationship between increasing confidence and the inclusion of more unwanted values ($P^* \neq P$) is gained by observing that the expected length of a random interval such as $I_P(s)$ in equation 1.8 is directly proportional to k. Hence, the greater the confidence coefficient the wider the interval. The value of k is most commonly set to yield confidence coefficients in the neighborhood of .95 or .99.

Significance Tests

When a sizeable group effect is observed in the sample, one can ask if it is likely that such an effect could be due solely to sampling variations. To answer such questions, statisticians have devised an inferential structure known as the test of significance. We will describe this structure in the context of National Assessment "group effects":

$$\Delta\hat{P}_G(s) = [\hat{P}_G(s) - \hat{P}(s)] \qquad (1.12)$$

where $\hat{P}_G(s)$ denotes the sample-s estimate of the proportion of group G members who can answer a particular exercise correctly and $\hat{P}(s)$ depicts the corresponding proportion for the entire population. Group G could, for example, denote the 9-year-olds residing in NAEP's Northeast region, in which case $\Delta\hat{P}_G(s)$ would compare the performance of the Northeast 9-year-olds against the overall national performance of 9-year-olds.

An observed group effect $\Delta\hat{P}_G(s)$ is judged to be significantly different from zero if its absolute value exceeds a critical value C. The critical value is determined so that the probability of observing an absolute effect $\Delta\hat{P}_G(s)$ in excess of C when the true population effect ΔP_G is zero is less than some arbitrarily small probability α. This probability α of declaring an observed sample effect significant when in fact the true population effect is zero is called the significance level of the test. Commonly used significance levels are $\alpha = .01$ and $\alpha = .05$. The critical value C frequently takes the form

$$C_{\Delta P}(s) = k \text{ se } (\Delta\hat{P}(s)) \qquad (1.13)$$

where k is a constant and se$\{\Delta\hat{P}(s)\}$ is the estimated standard error for the group effect $\Delta\hat{P}(s)$. The subscript G designating a particular subgroup has been dropped from the group effect symbol in equation 1.13 to simplify our notation. If we let $A(\Delta P)$ denote the set of samples for which $\Delta\hat{P}(s)$ exceeds k se$\{\Delta\hat{P}(s)\}$ in absolute value and use

$$\alpha(\Delta P) = \sum_{s\epsilon A(\Delta P)} \pi(s) \tag{1.14}$$

to denote the probability that an observed group effect $\Delta\hat{P}(s)$ will be judged significant, then $\alpha(\Delta P)$ can be expressed as follows:

$$\alpha(\Delta P) = \Pr\{|t(s,\Delta P)|>k\} \tag{1.15}$$

where

$$\begin{aligned} t(s,\Delta P) &= \Delta\hat{P}(s)/se\{\Delta\hat{P}(s)\} \\ &= [\Delta\hat{P}(s) - \Delta P]/se\{\Delta\hat{P}(s)\} + \Delta P/se\{\Delta\hat{P}(s)\} \\ &= t(s) + \Delta P/se\{\Delta\hat{P}(s)\}. \end{aligned}$$

Notice that, as with the OC curve presented in equation 1.11 for our confidence interval, $\alpha(\Delta P)$ can be specified in terms of the sampling distribution of a statistic $t(s,\Delta P)$ which has the form of Student's noncentral T statistic. If ΔP, the true population group effect, were zero, then $\alpha(0) = \Pr\{|t(s)|> k\}$ represents the significance level of the test with $t(s)$ taking the form of Student's central T statistic. For populations with $\Delta P \neq 0$, $\alpha(\Delta P)$ gives the probability of declaring significance when the true group effect is ΔP. Taken as a function of ΔP, the curve $\alpha(\Delta P)$ described in equation 1.15 is called the power function of the significance test. As ΔP deviates increasingly from zero, one would hope that $\alpha(\Delta P)$, the probability of declaring significance, would rise sharply.

While the OC curve for our confidence interval could be completely determined if the population was fully specified, only one point of the power curve can be determined: namely, that point corresponding to the true group effect ΔP^{o}. The other points are conceptual in the sense that they specify what the probability of declaring significance would be for a similar population where the true group effect was $\Delta P^{*} \neq \Delta P^{o}$.

Prescribing a critical value for a test of significance that will yield a predetermined significance level α presumes knowledge of the sampling distribution of $t(s,\hat{\Delta P}) = \hat{\Delta P}(s)/se\{\hat{\Delta P}(s)\}$ for a conceptual population, which is like the population of interest and is characterized by a negligible group effect $\Delta P = 0$. At this point, as was the case with confidence intervals, sampling statisticians commonly assume that Student's central T distribution would be a reasonable approximation for the sampling distribution of $t(s,\hat{\Delta P})$ from a population with $\Delta P = 0$. If the degrees of freedom associated with $se\{\hat{\Delta P}(s)\}$ exceeds 60, one can effectively use the standard normal distribution to determine k such that $Pr\{|t(s,0)| > k\} = \alpha$. Typical values of k from the standard normal distribution are $k = 1.96$ for a significance level $\alpha = .05$ and $k = 2.58$ for a significance level of $\alpha = .01$. Examining the form of the "power function" in equation 1.15 makes it clear that, while one may reduce the risk of falsely declaring significance (that is reduce α) by increasing k, there will be a corresponding reduction in the power to declare significance when the true group effect ΔP deviates from zero. This same relationship was noted between increasing confidence coefficients and lengthening intervals.

In addition to the direct comparisons between subgroup and national proportions of correct answers which we have called group effects, National Assessment reports adjusted or balanced effects which attempt to correct for the masquerading of one characteristic as the effect of another. While the unadjusted group effects properly reflect the differences in achievement between specific groups of children, much of the observed difference may well be attributable to other factors on which the compared groups differ. For example, part of the deficit in achievement observed in the direct comparison of Black students with non-Blacks may be attributed to the fact that Black students tend, more than non-Black students, to have less educated parents. In the following section, the adjustment methodology used by National Assessment to compensate for some of this masquerading is presented.

Balanced Effects

The major population subgroupings used in National Assessment reports are: Age, Region, Size and Type of Community (STOC), Sex, Color, and

Parents' Education. Within the four age classes, group effects contrasting the levels of the other five factors are presented. As we have indicated, these direct comparisons across the levels of a single factor are subject to masquerading influences of the other four factors. This confusion is partially due to the unbalanced mix of these other characteristics across the levels of any single factor being examined. To balance out this disproportionality, National Assessment forms adjusted group effects (expressed in percentages) that, when combined by addition with each other and with the overall "national" percentage of success, give fitted percentages of success (P-values) that correspond with the actual sample data in the following way:

> If we choose any level of a single characteristic, say Blacks, and use the fitted P-value and estimated population size to calculate the number of successes for each Region x STOC x SEX x Parents Education subclass of Blacks, and then add these predicted numbers of successes, the predicted number of successes over all these subclasses will be the same as the total number of Black successes estimated from the sample data.

If we let $i = 1(1)4$ index NAEP's four regions; $j = 1(1)7$ the seven STOC categories; $k = 1, 2$ the two sexes; $\ell = 1(1)3$ the three color classes; and $m = 1(1)5$ NAEP's five levels of Parents' Education, then the fitted P-value for subclass $(ijk\ell m)$ has the form

$$\hat{P}_{Bal}(ijk\ell m) = \hat{P} + \Delta\hat{R}(i) + \Delta\hat{T}(j) + \Delta\hat{S}(k) + \Delta\hat{C}(\ell) + \Delta\hat{E}(m) \qquad (1.16)$$

where $\hat{P}$ is the overall (national) percent-correct and the Δ terms represent the 'Balanced' group effects for Region-i, STOC-j, Sex-k, Color-ℓ, and Parents' Education class-m. With $\hat{M}(ijk\ell m)$ denoting the estimated population size for subclass $(ijk\ell m)$ and $\hat{Y}(ijk\ell m)$ representing the estimated number of correct responses from this subclass, the balancing condition verbalized above translates into the following fitting equations:

$$\sum_{j=1}^{7}\sum_{k=1}^{2}\sum_{\ell=1}^{3}\sum_{m=1}^{5} \hat{M}(ijk\ell m)\,\hat{P}_{Bal}(ijk\ell m) = \hat{Y}(i{+}{+}{+}{+}) \text{ for } i = 1(1)4 \qquad (1.17a)$$

$$\sum_{i=1}^{4}\sum_{k=1}^{2}\sum_{\ell=1}^{3}\sum_{m=1}^{5} \hat{M}(ijk\ell m)\,\hat{P}_{Bal}(ijk\ell m) = \hat{Y}({+}j{+}{+}{+}) \text{ for } j = 1(1)7 \qquad (1.17b)$$

Similar sets of equations are produced for the other three classifications by summing over all the subgroups within a particular factor level and equating to the estimated total correct for that factor level. Notice that we have used a plus sign to denote a summed over subscript. Substituting the linear main effects model in (1.16) for $\hat{P}_{Bal}(ijk\ell m)$ the fitting equations become:

$$\hat{M}(i)\hat{P} + \hat{M}(i)\Delta\hat{R}(i) + \sum_{j=1}^{7} M(ij)\,\Delta\hat{T}(j) + \sum_{k=1}^{2} \hat{M}(ik)\,\Delta\hat{S}(k) \qquad (1.18a)$$

$$+ \sum_{\ell=1}^{3} \hat{M}(i\ell)\Delta\hat{C}(\ell) + \sum_{m=1}^{5} \hat{M}(im)\Delta\hat{E}(m) = \hat{Y}(i) \text{ for } i = 1(1)4$$

and

$$\hat{M}(j)\hat{P} + \sum_{i=1}^{4} \hat{M}(ij)\Delta\hat{R}(i) + \hat{M}(j)\Delta\hat{T}(j) + \sum_{k=1}^{2} \hat{M}(jk)\Delta\hat{S}(k) \qquad (1.18b)$$

$$+ \sum_{\ell=1}^{3} \hat{M}(j\ell)\Delta\hat{C}(\ell) + \sum_{m=1}^{5} \tilde{M}(jm)\Delta\hat{E}(m) = \hat{Y}(j) \text{ for } j = 1(1)7$$

The other three sets of fitting equations are arrived at similarly. Notice that we have suppressed the summed-over subscripts to make the expressions more compact.

Since each of the sets of fitting equations corresponding to a particular classification factor sums to the same quantity, namely

$$\hat{M}\hat{P} + \sum_{i=1}^{4} \hat{M}(i)\Delta\hat{R}(i) + \sum_{j=1}^{7} \hat{M}(j)\Delta\hat{T}(j) + \sum_{k=1}^{2} \hat{M}(k)\Delta\hat{S}(k) \qquad (1.19)$$

$$+ \sum_{\ell=1}^{3} M(\ell)\Delta\hat{C}(\ell) + \sum_{m=1}^{5} \hat{M}(m)\Delta\hat{E}(m) = \hat{Y}$$

one of the equations in each set is redundant. That is, of the 4 + 7 + 2 + 3 + 5 = 21 balancing equations produced in this fashion, only 16 are independent. To solve for our 21 balanced effects we need five additional equations. Requiring that the overall $\hat{P}$ in our model (equation 1.16) be equivalent to the unadjusted national P-value ($\hat{P} = \hat{Y} / \hat{M}$) implies in equation 1.19 that

$$\sum_{i=1}^{4} \hat{M}(i)\Delta\hat{R}(i) = \sum_{j=1}^{7} \hat{M}(j)\Delta\hat{T}(j) = \sum_{k=1}^{2} \hat{M}(k)\Delta\hat{S}(k) \tag{1.20}$$

$$= \sum_{\ell=1}^{2} \hat{M}(\ell)\Delta\hat{C}(\ell) = \sum_{m=1}^{4} \hat{M}(m)\Delta\hat{E}(m) = 0 .$$

Setting each of these sums equal to zero yields five independent equations, which can be substituted respectively for the last equation in each of the original five sets. This yields 21 independent equations, which can be solved to yield the full set of balanced effects.

While this balancing solution was not derived with the least squares principle in mind, one can view the results as a sample estimate of the least squares solution that would be obtained if the entire population of correct-incorrect (1-0) responses were predicted by a linear model with an intercept and 21 dummy variables indicating membership in the 21 factor level subgroups. The weighted restrictions in equation (1.20), with the "hats" removed from the population sizes (Ms), are commonly applied to unbalanced data sets. This dummy-variable regression view of NAEP's balanced fitting places the results in a familiar statistical setting where the adjustment of regression coefficients for unbalanced repfesentation across categories is a well-known property.

While balancing helps to correct for disproportionate numbers, this adjustment is obviously limited to the variables that are used in the analysis. Other unmeasured variables such as family income may also be causing masquerading problems. Some variables used in the adjustment, such as color, may classify respondants too coarsely; while other factors, such as parent's education, give only an indirect indication of the parents' attitude toward education or their inclination to assist the student with homework. Another potential problem with direct comparisons between subgroups is the fact that the performance of a given subgroup may differ from one subgrouping to another in the other variables. That is, the effects associated with Black students may be different in the West than in the Southeast. Such interaction effects are not accounted for in NAEP's balancing model. In spite of these deficiencies, balancing represents a big step from the outward appearances of unadjusted group effects toward the inward realities of cause and effect.

REFERENCES

1. Kish, L. (1965). Survey Sampling. New York: John Wiley and Sons.
2. Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. Biometrics 2, 110.

Chapter 2: YEAR 01 SAMPLING ERRORS

Design Description

The NAEP Year 01 sample for the three in-school age classes (9, 13, and 17) began with a highly stratified, simple random selection of 208 primary units. These primary units consisted of clusters of schools formed within selected listing units. The listing units were counties or parts of counties. Variables used to stratify these listing units included (1) Region (4 Geographic Regions), (2) SOC (4 'Size of Community' Classes), and (3) SES (2 Socio-Economic-Status Categories). Within each selected listing unit a separate set of schools was selected for each of three age groups: 9-year-olds, 13-year-olds, and 17-year-olds. For each of these age groups, schools were grouped such that every set would contain a mix of high and low SES students. Portions of some large schools were allowed to belong to more than one group. The number of schools in each of these clusters was based on the numbers of packages or questionnaires required from each PSU. The 17-year-old assessment, for example, employed 11 separate group-administered packages and 2 individually administered packages. Group administrations consisted of 12 students, while each individual package was given separately to 9 students in each PSU.

The sample was designed to yield two primary units from each of 104 strata. For the 17-year old assessment, (11 x 12) + (2 x 9) = 150 students were required from each PSU. The groups of 17-year-old schools were constructed to contain approximately 300 17-year-olds each. Once a cluster of schools was selected via simple random sampling (SRS) from those constructed, the group packages were allocated to schools. Each school in the cluster was assigned a number of group administrations roughly proportional to its enrollment of 17-year-olds. Sixteen students were selected for each group session assigned to a particular school: 12 to participate and 4 to be alternates.

The two individual packages were allocated to schools such that for each group package from 1 through 9 assigned to a school, an administration of individual package 13 was also planned. Individual package 14 administrations were similarly linked to administrations of group packages 3 through 11. For each individual package administration planned for a school, two students were selected, one to participate and one alternate. This design yields a planned sample size of 2,448 students for each group-administered package and 1,836 for each individual package.

The Year 01 out-of-school sample of young adults 26-35 and out-of-school 17-year-olds used the same basic primary sample design as the in-school sample. The same random draw was used to select PSUs in both samples; however, the out-of-school PSUs were defined in terms of a set of area segments or clusters containing an average of 35 to 40 housing units. Each of these PSUs was constructed so as to contain about 16,000 persons. The second-stage sample was a stratified random cluster sample with two clusters selected without replacement from each of five strata. The stratification was based on an ordering of segments in terms of the precent of families earning less than $3,000. The high poverty (low SES) quarter of the list was assigned two strata for a two-to-one oversampling of the low SES quarter. Each household cluster was expected to yield 12.5 eligible adult respondents. Ten packages of exercises were administered to young adults with each respondent randomly assigned a single package. Out-of-school 17-year-olds encountered in the household sample were asked to respond to a set of four or five of the 17-year-old in-school packages. Recall that there were 13 such packages. An incentive payment of 10 dollars was given for completing the set of packages.

Parameters of Interest

Proportions Correct (P-Values)

The purpose of National Assessment (Year 01) was to produce baseline estimates of the proportions of potential respondents who would answer a certain exercise in a particular way. Restricting our attention to a particular in-school age group (say 17-year-olds) and a particular exercise within one of the packages, let

$$Y_{hijk} = \begin{cases} 1 \text{ if the k-th student in school (j) of PSU (i) in} \\ \text{stratum-h answers correctly; 0 otherwise} \end{cases}$$

The population means of these 0, 1 variables are the population proportions of interest, that is

$$Y_{....} = P = \sum_{h=1}^{H} \sum_{i=1}^{N_h} \sum_{j=1}^{S_{hi}} \sum_{k=1}^{M_{hij}} Y_{hijk}/M_{+++} \qquad (2.1)$$

where

- H = the number of strata (104 planned)
- N_h = the number of PSUs in stratum (h)
- S_{hi} = the number of schools in PSU (hi)
- M_{hij} = the number of students in school (hij)

and

$$M_{+++} = \sum_{h=1}^{H} \sum_{i=1}^{N_h} \sum_{j=1}^{S_{hi}} M_{hij}\ .$$

Our sample estimates for these proportions are of the form

$$\hat{P} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{s_{hi}} \sum_{k=1}^{m_{hij}} W_{hijk}\, Y_{hijk}/W_{++++} \qquad (2.2)$$

where

- n_h = the number of PSUs selected for the sample from stratum (h) (generally $n_h = 2$)
- s_{hi} = the number of schools in PSU (hi) in which the particular package of interest was administered
- m_{hij} = the number of students from school (hij) who respond to the package of interest;

and, aside from nonresponse adjustments,

$$W_{hijk} = 1/\Pr\{\text{PSU (hi)}\} \times \Pr\{\text{Sch (j)} \mid \text{(hi)}\} \times \Pr\{\text{Kid (k)} \mid \text{(hij)}\} \qquad (2.3)$$

with

$$\Pr\{\text{PSU (hi)}\} = n_h/N_h$$
$$\Pr\{\text{SCH(j)} \mid \text{(hi)}\} = \pi_{hij}$$
$$\Pr\{\text{Kid (k)} \mid \text{(hij)}\} = m_{hij}/M_{hij}.$$

For group-administered packages the number of sample schools per PSU (s_{hi}) was always 1 in Year 01. Individual packages were administered in more than one school; that is, $s_{hi} \geq 1$ for Year 01 individual package exercises.

The estimation of out-of-school, young adult, P-values parallels the procedure presented for the Year 01 in-school sample. If we let j subscript area segments instead of schools and k young adults instead of students, the expressions in equations 2.1 and 2.2 are interchangeable. To complete the switch, we let S_{hi} denote the number of segments in the PSU-hi frame and s_{hi} the number of sample segments in PSU-hi (usually $s_{hi} = 10$). Also, let M_{hij} denote the number of eligible young adults in segment-hij and m_{hij} the number of young adults in segment-hij responding to a particular package. The out-of-school sample weights reflect the selection probabilities for young adults plus adjustments for nonresponse.

Out-of-school 17-year-olds located and tested in the household survey were combined with in-school respondents to estimate a single P-value for all 17-year-olds. The total number of out-of-school 17-year-olds and the number that could respond correctly were estimated for each 17-year-old package using weight sums for all package respondents and for all respondents answering correctly. These estimated totals were then added to the denominator and numerator of the in-school package P-Value.

Subpopulation P-Values and ΔP Values

In addition to the national P-Values discussed in the previous section, certain subpopulation breakdowns were of interest. For example, P-Values have been presented by Region, STOC, Sex, Color, and Parents' Education. These subpopulation P-Values were produced by including only those observations belonging to the subpopulation of interest in the numerator and denominator of equation 2.2. Differences between subpopulation and national P-Values were studied to assess the main effects of Region, STOC, Color, Sex, and Parents' Education. These direct comparisons were introduced as group effects or ΔP-values in chapter 1.

Balanced Effects

In chapter 1 we introduced NAEP's algorithm for adjusting group effects. This adjustment was designed to correct for the masquerading effect of ancillary variables when their distributions vary across the levels of the factor being examined. The adjustment or balancing algorithm used amounts to a set of linear equations which can be viewed as a sample approximation to the normal equations that would result from a least-squares fit to the population of 1-0 (correct-incorrect) responses based on an intercept and dummy variables indicating the levels of NAEP's five reporting categories. A set of restrictions are imposed on the balanced effects, which force the linear model intercept to equal the observed national P-Value. The left-hand sides of the balancing equations involve weighted sample estimates of population counts in the one-way and two-way margins of NAEP's Region by STOC, by Sex, by Color, and by Parents' Education classification. The right-hand sides of the balancing equations involve estimated counts of correct responses from the five one-way margins. Suppose we let X_{hijk} denote a 1 x 22 row vector for student-k (adult or out-of-school 17) in school-j (segment) of PSU-i in stratum-h with the first element equal 1 for all respondents-hijk and the remaining 21 elements taking values 1 or 0 depending on the respondent's membership in the 21 subgroups formed by NAEP's reporting categories (4 Regions + 7 STOCs + 2 Sexes + 3 Colors + 5 Parents' Education classes). Recalling that Y_{hijk} is 1 if respondent (hijk) answers correctly and 0 otherwise, we can specify the balancing equations prior to substitution with the restrictions as

$$(X^TX)\hat{\beta} = (X^TY) \qquad (2.4)$$

where

$$(X^TX)_{22x22} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{s_{hi}} \sum_{k=1}^{m_{hij}} W_{hijk}(\underset{\sim}{X}^T_{hij} \underset{\sim}{X}_{hijk})$$

$$(X^TY)_{22x1} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{s_{hi}} \sum_{k=1}^{m_{hij}} W_{hijk}(\underset{\sim}{X}^T_{hijk} Y_{hijk})$$

and

$$\hat{\beta}^T = <\hat{P}, \Delta\hat{R}(1) \ldots \Delta\hat{R}(4), \Delta\hat{T}(1) \ldots \Delta\hat{T}(7), \Delta\hat{S}(1)\ \Delta\hat{S}(2),$$
$$\Delta\hat{C}(1) . \Delta\hat{C}(3), \Delta\hat{E}(1) \ldots \Delta\hat{E}(5) > .$$

As we have noted in chapter 1, the balancing equations in equation 2.3 are not linearly independent since the sum of the 2nd through the 5th equations equals the 1st as do the 6th through the 12th, the 13th and 14th, 15th through 17th, and the 18th through 22nd. To provide for a unique solution and at the same time force the intercept $\hat{P}$ to equal the observed national P-value the final equation in each of the five blocks, which correspond to the five reporting variables, are replaced by a linear restriction on that variable's balanced effects. For example, the fifth equation in equation 2.4 is replaced by

$$\hat{M}(1{+}{+}{+}{+})\Delta\hat{R}(1) + \hat{M}(2{+}{+}{+}{+})\Delta\hat{R}(2) + \hat{M}(3{+}{+}{+}{+})\Delta\hat{R}(3) + \hat{M}(4{+}{+}{+}{+})\Delta\hat{R}(4) = 0 . \qquad (2.5)$$

This substitution can be accomplished by replacing the fifth row in $(X^T_{hijk}X_{hijk})$ with a (1 x 22) row vector with all elements except the second through the fifth set to zero. The four elements in columns two through five of the new fifth row take the values one or zero to indicate membership in regions 1 through 4 successively. The fifth row of $(X^T_{hijk}\ Y_{hijk})$ is set to zero for every respondent-(hijk). When properly weighted and summed, it is clear that the new fifth row of our individual balancing equations will yield the restriction equation 2.5. Similar substitutions of rows 12, 14, 17, and 22 with the linear equations in equation 1.20 produces NAEP's restricted set of balancing equations. In our further treatment of balancing, $(X^TX)_{hijk}$ and $(X^TY)_{hijk}$ will represent the restricted respondent-(hijk) contributions to the left- and right-hand sides of the balancing equations. Substituting these independent linear restrictions for the redundent rows of (X^TX) and setting the corresponding rows of (X^TY) to zero allows one to specify the balanced fit uniquely as

$$\hat{\beta} = (X^TX)^{-1}\ (X^TY) \qquad (2.6)$$

where

$$(X^TX)_{22x22} = \sum_{h=1}^{H}\ \sum_{i=1}^{n_h}\ \sum_{j=1}^{S_{hi}}\ \sum_{k=1}^{m_{hij}} W_{hijk}\ (X^TX)_{hijk}$$

and

$$(X^TY)_{22x1} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{S_{hi}} \sum_{k=1}^{m_{hij}} W_{hijk}\ (X^TY)_{hijk}$$

Variance Estimators

Variance Estimators for P-Values and ΔP-Values

To support the presentation of P-Values and ΔP-Values, measures of the sampling variability of these statistics were needed. A jackknife replication procedure for estimating the sampling variance of nonlinear statistics from complex multistage samples was tailored to our design. This technique is easily applied to highly stratified designs with only two primary units (PSUs) selected with replacement or without replacement from strata where the fpc (n_h/N_h) can be ignored [refs. 1,2]. The Year 01 primary sample fits this description except for a few strata containing single primary units. These singleton PSUs are accounted for in the following section.

To demonstrate the computational aspects of this technique, we can consider estimating the variance of a national P-Value. First we define expanded-up PSU totals

$$\hat{Y}_{hi} = \sum_{j=1}^{s_{hi}} \sum_{k=1}^{m_{hij}} W_{hijk}\ Y_{hijk} \qquad (2.7)$$

and

$$\hat{M}_{hi} = \sum_{j=1}^{s_{hi}} \sum_{k=1}^{m_{hij}} W_{hijk} \qquad (2.8)$$

Recalling equation 2.2, we see that the total $\hat{M}_{hi}$ represents the PSU-hi contribution to our sample estimate of the number of 17-year-olds in stratum-h while $\hat{Y}_{hi}$ is the PSU-hi contribution to the estimated number of 17-year-olds in stratum-h who could answer the question correctly. In terms of these expanded PSU totals, the P-Value becomes

$$\hat{P} = \{ \sum_{h=1}^{H} \sum_{i=1}^{n_h} \hat{Y}_{hi} \Big/ \sum_{h=1}^{H} \sum_{i=1}^{n_h} \hat{M}_{hi} \} = (\hat{Y}/\hat{M}) . \qquad (2.9)$$

The jackknife estimate of $\hat{P}$, say $\hat{P}_{JK}$, and its variance estimator are special applications of the following general result for a sample of H strata with n_h primary selections per strata [ref. 3] (with replacement or without from strata such that n_h/N_h is negligible).

Let $\hat{\Theta}^o$ depict a statistic based on data from all n_h PSUs in each stratum. Define the replication estimate Θ_{-hi} constructed from all the PSUs excluding PSU-i in stratum-h. These replication estimates should be produced as if this censored PSU had not responded; that is, reasonable nonresponse adjustments should be used in estimating Θ without PSU (hi). The jackknife pseudo-values $\hat{\Theta}_{hi}$ are then formed where

$$\hat{\Theta}_{hi} = n_h \hat{\Theta}^o - (n_h-1)\hat{\Theta}_{-hi} . \tag{2.10}$$

The jackknifed alternative for $\hat{\Theta}^o$ is

$$\hat{\Theta}_{JK} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \hat{\Theta}_{hi}/Hn_h . \tag{2.11}$$

A consistent estimate of the variance of $\hat{\Theta}_{JK}$ is

$$\mathrm{var}_{JK}(\hat{\Theta}_{JK}) = \sum_{h=1}^{H} s^2(\hat{\Theta}_{h.})/n_h \tag{2.12}$$

where

$$\Theta_{h.} = \sum_{h=1}^{n_h} \hat{\Theta}_{hi}/n_h$$

and

$$s^2(\Theta_{h.}) = \sum_{h=1} \{\hat{\Theta}_{hi} - \Theta_{h.}\}^2/(n_h-1) .$$

Commenting on an earlier draft of this report, Dr. David R. Brillinger [ref. 4] has pointed out that a pseudo-value of the form

$$\hat{\Theta}^*_{hi} = Hn_h \hat{\Theta}^o - (H-1)\hat{\Theta} - H(n_h-1)\hat{\Theta}_{-hi}$$

would be more appropriate for a stratified sample [ref. 4]. This result was obtained by approximating the expectations of $\hat{\Theta}^o$, $\hat{\Theta}_{-hi}$, and $\Theta_{-h.}$ with Taylor series of the form

$$E(\hat{\theta}^{o}) \sim \theta + \sum_{h=1}^{H} a_h/n_h \text{ + second-order terms}$$

$$E(\hat{\theta}_{-hi}) \sim \theta + a_h/(n_h-1) + \sum_{\lambda \neq h} a_\lambda/n_\lambda \text{ + second order}$$

and

$$E(\hat{\theta}_{-h\cdot}) \sim \theta + a_h/(n_h-1) + \sum_{\lambda \neq h} a_\lambda/n_\lambda \text{ + second order.}$$

Using the series approximations above, one notes that

$$(n_h-1)\, E(\hat{\theta}_{-h\cdot} - \hat{\theta}^{o}) \sim a_h/n_h \text{ + second order.}$$

Therefore, for Brillinger's alternative jackknife estimator

$$\hat{\theta}^{*}_{JK} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \hat{\theta}^{*}_{hi}/n_h H$$

$$= \hat{\theta}^{o} - \sum_{h=1}^{H} (n_h-1)(\hat{\theta}_{-h\cdot} - \hat{\theta}^{o})$$

it is clear that

$$E(\hat{\theta}^{*}_{JK}) = \theta \text{ + second-order terms.}$$

Applying a similar argument, one can demonstrate that the jackknife estimator proposed in equation (2.11) contains first-order bias terms; namely,

$$E(\hat{\theta}_{JK}) \sim \theta + \left(\frac{H-1}{H}\right) \sum_{h=1}^{H} a_h/n_h \text{ + second order.}$$

The variance approximation proposed for $\hat{\theta}^{*}_{JK}$ by Brillinger has the form

$$var_{JK}(\hat{\theta}^{*}_{JK}) = \sum_{h=1}^{H} S^2(\hat{\theta}^{*}_{h\cdot})/n_h H^2 .$$

The variance expression above is equivalent to the estimator in equation (2.12).

Applying Brillinger's result to produce a jackknifed estimate of $\hat{P} = \hat{\theta}^{o}$, we first consider the case where all $n_h = 2$. If this were the case then

$$\hat{P} = \sum_{h=1}^{H} (\hat{Y}_{h1} + \hat{Y}_{h2}) \Big/ \sum_{h=1}^{H} (\hat{M}_{h1} + \hat{M}_{h2}) . \qquad (2.13)$$

The replicate P-Values become

$$\hat{P}_{-h1} = \left\{\frac{\hat{Y} - (\hat{Y}_{h1} - \hat{Y}_{h2})}{\hat{M} - (\hat{M}_{h1} - \hat{M}_{h2})}\right\}$$

$$\hat{P}_{-h2} = \left\{\frac{\hat{Y} + (\hat{Y}_{h1} - Y_{h2})}{\hat{M} + (\hat{M}_{h1} - \hat{M}_{h2})}\right\} \qquad (2.14)$$

The first P-Value is formed by discarding PSU 1 in stratum-h and replacing its contribution to the numerator and denominator of $\hat{P}$ with the data from its companion PSU (h2). The second P-Value is formed by discarding PSU 2 in stratum-h and replacing its contribution with that from PSU (h1). The jackknife pseudo-values become

$$\hat{P}_{hi} = (H+1)\hat{P} - H\hat{P}_{-hi} \qquad (2.15)$$

and the jackknife P-Value is

$$\hat{P}_{jk} = \sum_{h=1}^{H}\sum_{i=1}^{2} \hat{P}_{hi}/2H = (H+1)\hat{P} - H\bar{\hat{P}} \qquad (2.16)$$

Equation 2.16 shows that the jackknife P-Value is (H+1) times the standard combined ratio estimate minus H times the simple average of the replicate P-Values. The variance estimate for P_{JK} is

$$var_{JK}(P_{JK}) = \sum_{h=1}^{H}\sum_{i=1}^{2} (\hat{P}_{hi} - \hat{P}_{h.})^2/2\,H^2 \qquad (2.17a)$$

Considering $\sum_{i=1}^{2} (\hat{P}_{hi} - \hat{P}_{h.})^2/2$ and recalling that

$\hat{P}_{h.} = (H+1)P - H\hat{P}_{-h.}$ we need not bother with the pseudo-values; that is,

$$\sum_{i=1}^{2} (\hat{P}_{hi} - \hat{P}_{h.})^2/2H^2 = \sum_{i=1}^{2} (\hat{P}_{-hi} - \hat{P}_{-h.})^2/2 \qquad (2.18a)$$

For $n_h = 2$ a convenient simplification for the expression in equation 2.18a is

$$\sum_{i=1}^{2} (\hat{P}_{-hi} - \hat{P}_{-h.})^2/2 = (\hat{P}_{-h1} - \hat{P}_{-h2})^2/4 \qquad (2.18b)$$

The simplified form for the jackknife variance estimator in equation 2.17a becomes

$$\text{var}_{JK}(\hat{P}_{JK}) = \sum_{h=1}^{H} (\hat{P}_{-h1} - \hat{P}_{-h2})^2/4 \,. \qquad (2.17b)$$

An analogous application of this technique produces ΔP-Values from replicates formed by successively deleting PSUs and replacing their contributions with data from their companion PSU. If these replicate ΔP-Values are denoted by $\Delta\hat{P}_{-hi}$ then the jackknife ΔP-Value is ΔP_{JK} where

$$\Delta P_{JK} = (H+1)\Delta\hat{P} - H \sum_{h=1}^{H} \sum_{i=1}^{2} \Delta\hat{P}_{-hi}/2H$$

with variance estimator

$$\text{var}_{JK}(\Delta\hat{P}_{JK}) = \sum_{h=1}^{H} (\Delta\hat{P}_{-h1} - \Delta\hat{P}_{-h2})^2/4 \,. \qquad (2.19)$$

For those unfamiliar with the jackknife linearization technique described above, it may be of interest to note the relationship between $\text{var}_{JK}(\hat{P}_{JK})$ in equation 2.17 and the standard Taylor series variance approximation for a combined ratio [ref. 5]. If we let $\delta\hat{Y}_h = (\hat{Y}_{h1} - \hat{Y}_{h2})$ and $\delta\hat{M}_h = (\hat{M}_{h1} - \hat{M}_{h2})$, the Taylor series variance approximation for var $\{\hat{P}\}$ is

$$\text{var}_{TS}(\hat{P}) = \sum_{n=1}^{H} [\delta\hat{Y}_h - \hat{P}\delta\hat{M}_h]^2/\hat{M}^2 \qquad (2.20a)$$

28

or

$$\mathrm{var}_{TS}(\hat{P}) = \sum_{h=1}^{H} \delta^2 z_h \qquad (2.20b)$$

with

$$z_{hi} = (\hat{Y}_{hi} - \hat{P}\hat{M}_{hi})/\hat{M}$$

and

$$\delta z_h = (z_{h1} - z_{h2}).$$

Examining the form of the jackknife replicate P-Values in equation 2.14, it is not difficult to see that

$$\begin{aligned} \delta\hat{P}_h &= (\hat{P}_{-h1} - \hat{P}_{-h2}) \\ &= 2\delta z_h/(1 - \delta^2\hat{M}_h/\hat{M}^2) \qquad (2.21) \end{aligned}$$

which leads to

$$\begin{aligned} \mathrm{var}_{JK}\{\hat{P}_{JK}\} &= \sum_{h=1}^{H} \delta^2\hat{P}_h/4 \\ &= \sum_{h=1}^{H} \delta^2 z_h/(1 - \delta^2\hat{M}_h/\hat{M}^2)^2 \ . \qquad (2.22) \end{aligned}$$

Comparison of the Taylor series and jackknife variance estimators in equations 2.20 and 2.22 points out the close analytic relationship between the two. The quantity $\delta^2\hat{M}_h/\hat{M}^2$ in the denominator of the jackknife variance expression in equation 2.22 is the stratum-h contribution to the estimated relative variance of $\hat{M}$, where

$$\text{rel-var}\{\hat{M}\} = \sum_{h=1}^{H} \delta^2\hat{M}_h/\hat{M}^2 \ . \qquad (2.23)$$

Since rel-var $\{\hat{M}\}$ is positive and generally much smaller than 1, we can expect the jackknife variance estimator to be slightly larger than the corresponding Taylor series variance approximation. One would also expect the difference between the two estimators to diminish as the number of strata increase since each stratum's contribution would represent a smaller fraction of rel-var $(\hat{M})$. Some numerical comparisons of the Taylor series and jackknife variance approximations will be presented in chapter 4.

Variance Estimators for Balanced Effects

Recalling the definitions of $(X^TX)_{hijk}$ and $(X^TY)_{hijk}$, the restricted respondent-(hijk) contributions to the left- and right-hand sides of our balancing equations, we begin by forming the expanded PSU-hi totals

$$(X^TX)_{hi} = \sum_{j=1}^{s_{hi}} \sum_{k=1}^{m_{hij}} W_{hijk} \quad (X^TX)_{hijk} \qquad (2.24a)$$

and

$$(X^TY)_{hi} = \sum_{j=1}^{s_{hi}} \sum_{k=1}^{m_{hij}} W_{hijk} \, (X^TY)_{hijk} \, . \qquad (2.24b)$$

These definitions allow us to specify the quantities (X^TX) and (X^TY) as stratum sums of the form

$$(X^TX) = \sum_{h=1}^{H} [(X^TX)_{h1} + (X^TX)_{h2}] \qquad (2.25a)$$

and

$$(X^TY) = \sum_{h=1}^{H} [(X^TY)_{h1} + (X^TY)_{h2}] \, . \qquad (2.25b)$$

The jackknife replicate estimator for the vector β of balanced effects, which is obtained by deleting the contribution from PSU-h1 and replacing it with the contribution from PSU-h2 can be specified as the unique solution to the following set of normal equations

$$[(X^TX) - \delta(X^TX)_h]\ \hat{\beta}_{-h1} = [(X^TY) - \delta(X^TY)_h] \qquad (2.26)$$

where

$$\delta(X^TX)_h = [(X^TX)_{h1} - (X^TX)_{h2}]$$

and

$$\delta(X^TY) = [(X^TY)_{h1} - (X^TY)_{h2}] .$$

Deleting the contribution from PSU-h2 and replacing it with the PSU-h1 contribution results in the set of normal equations

$$[(X^TX) + \delta(X^TX)_h]\ \hat{\beta}_{-h2} = [(X^TY) + \delta(X^TY)_h] \qquad (2.27)$$

which can be solved for the replicate estimate $\hat{\beta}_{-h2}$. Jackknife pseudo values

$$\hat{\beta}_{hi} = (H+1)\hat{\beta} - H\hat{\beta}_{-hi} \qquad (2.28)$$

are then formed from the replicate estimators where $\hat{\beta}$ represents the estimated vector of balanced effects based on data from all PSUs. The jackknifed estimator for β is then

$$\hat{\beta}_{JK} = \sum_{h=1}^{h} (\beta_{h1} + \beta_{h2})/2H \qquad (2.29)$$

$$= (H+1)\hat{\beta} - H\hat{\beta}_{-..}$$

To estimate the variance-covariance matrix of the jackknifed vector of balanced effects, we use

$$\text{var}_{JK}\{\hat{\beta}_{JK}\} = \sum_{h=1}^{H} \delta\hat{\beta}_h \, \delta^T\hat{\beta}_h / 4H^2 \qquad (2.30)$$

where $\delta\hat{\beta}_h = [\hat{\beta}_{h1} - \hat{\beta}_{h2}] = -H[\hat{\beta}_{-h1} - \hat{\beta}_{-h2}]$.

Notice that $\delta\hat{\beta}_h$ is a (22 x 1) column vector and $\delta^T\hat{\beta}_h$, the transpose of $\delta\hat{\beta}_h$, is a (1 x 22) row vector. The resulting quantity in equation 2.30 is therefore a (22 x 22) matrix of estimated variances and covariances among the 21 balanced effects and the corresponding national P-Value, $\hat{P}$.

In appendix A the corresponding Taylor series approximation for the variance-covariance matrix of $\hat{\beta}$ is derived. This method yields the estimator

$$\text{var}_{TS}\{\hat{\beta}\} = \sum_{h=1}^{H} \delta z_h \, \delta^T z_h \qquad (2.31)$$

where

$$\delta z_h = (X^TX)^{-1} [\delta(X^TY)_h - \delta(X^TX)_h \hat{\beta}] .$$

Although it is not immediately apparent, there is again a close relationship between the form of the jackknife and Taylor series variance-covariance estimators. Subtracting matrix equation 2.27 from 2.26 and rearranging terms, one finds that

$$(X^TX) \, \delta\hat{\beta}_h = 2H[\delta(X^TY)_h - \delta(X^TX)_h \hat{\beta}_{-h.}] \qquad (2.32)$$

where

$$\delta\hat{\beta}_h = [\hat{\beta}_{h1} - \hat{\beta}_{h2}] = -H[\hat{\beta}_{-h1} - \hat{\beta}_{-h2}]$$

and

$$\hat{\beta}_{-h.} = [\hat{\beta}_{-h1} + \hat{\beta}_{-h2}]/2 .$$

In solving the set of equations in equation 2.32 for $\delta\hat{\beta}_h$, the difference between our two jackknife pseudo values from stratum-h, yields

$$\delta\hat{\beta}_h = 2H(X^TX)^{-1}[\delta(X^TY)_h - \delta(X^TX)_h\,\hat{\beta}_{-h.}] \,. \tag{2.33}$$

Using the expression for $\delta\hat{\beta}_h$ presented in equation 2.33 shows that the jackknife variance-covariance estimator in equation 2.30 differs from the corresponding Taylor series estimator only to the extent that $\hat{\beta}_{-h.}$, the average of our two replicate estimators from stratum-h, differs from the estimate $\hat{\beta}$ based on all the data. As the number of strata increases, one would expect the difference between $\hat{\beta}$ and $\hat{\beta}_{-h.}$ to get small. For National Assessment's Year 01 sample design with 104 primary strata, there should be little difference between the two methods.

Computational Considerations

The major complication that arose in applying the procedures introduced in the previous sections to National Assessment data was strata with only one PSU. To allow these strata to contribute to variance, psuedo strata containing two or three of these singleton PSUs were formed. This collapsing of strata was done within regions and as much as possible within SOC (Size of Community) superstrata. State and county names for these PSUs were also used in the matching. When two PSUs from different strata are collapsed, some adjustment should be made for the fact that the stratum sizes (N_h's) may be quite different. One such adjustment is to replace the stratum expansion factors $(N_h/1)$ and $(N_\ell/1)$ for the two singles with a common expansion appropriate for a design with two PSUs selected from the union of strata-h and -ℓ; that is, use the common expansion factor $(N_h + N_\ell)/2$. When applied to our jackknife methodology, this adjustment amounts to replacing stratum-ℓ's contribution to $\hat{P}$ with

$$N_h(\hat{Y}_\ell/N_\ell) = N_h\bar{y}_\ell \tag{2.34a}$$

and

$$N_h(\hat{M}_\ell/N_\ell) = N_h\bar{m}_\ell \,. \tag{2.34b}$$

We have "borrowed" the data from stratum-ℓ in terms of estimated numbers of 17-year-olds per PSU ($\bar{m}_\ell$) and estimated numbers of correct respondents ($\bar{y}_\ell$), but have retained the number of PSUs appropriate for stratum-h. The contribution from stratum-ℓ is similarly replaced by stratum-h data, but its number of PSUs (N_ℓ) is retained. The adjusted replicate P-Values for collapsing singleton strata-h and -ℓ are therefore:

$$\hat{P}_{-\ell} = \left\{\frac{\bar{Y} - \hat{Y}_\ell + N_\ell \bar{y}_h}{\hat{M} - \hat{M}_\ell + N_h \bar{m}_\ell}\right\} = \left\{\frac{\hat{Y} - N_h (\bar{y}_h - \bar{y}_\ell)}{\hat{M} - N_h (\bar{m}_h - \bar{m}_\ell)}\right\} \qquad (2.35a)$$

and

$$\hat{P}_{-h} = \left\{\frac{\hat{Y} - \hat{Y}_h + N_k \bar{y}_\ell}{\hat{M} - \hat{M}_h + N_h \bar{m}_\ell}\right\} = \left\{\frac{\hat{Y} - N_h (\bar{y}_h - \bar{y}_\ell)}{\hat{M} - N_h (\bar{m}_h - \bar{m}_\ell)}\right\} . \qquad (2.35b)$$

The resulting squared difference between $\hat{P}_{-\ell}$ and $\hat{P}_{-h}$ divided by four is a conservative estimate (overestimate) of the variance contribution from strata h and -ℓ. When an odd number of singleton PSUs was available within a major region, by SOC stratum, the last three singletons were used to form three pseudo strata, each comprising one of the possible pairings among the three units. The variance contribution from three singletons was estimated by adding the three squared differences divided by eight. The division by eight results from the fact that each of the three PSUs is accounted for in two of the squared differences.

An alternative stratum size adjustment, which requires no knowledge of the separate stratum sizes, N_h, uses the estimated student population from the singleton strata in the adjustment. Assuming that the PSUs in the two collapsed strata all contain approximately the same number of students, say $\bar{M}$, then the sum of weights for a singleton strata-h estimates

$$\sum_{i=1}^{N_h} M_{hi} = N_h \bar{M}_n = N_h \bar{M} \qquad (2.36)$$

When the h-th stratum is discarded in the replicate P-Value estimation, its contribution to the numerator is replaced by its estimated population sizes ($\hat{M}_h$) times the estimated proportion correct from stratum-ℓ, that is

$$\hat{Y}_h = \hat{M}_h \cdot (\hat{Y}_\ell/\hat{M}_\ell). \qquad (2.37)$$

No change is made to the denominator with this adjustment. Such an adjustment forces an equality of PSU sizes, which is not achieved for two legitimate selections from the same stratum, and therefore would seem to underestimate the variance in this respect. As long as collapsing is not extensive, the differential effect of the two alternatives is probably negligible.

Aside from the problems surrounding stratum collapsing, the application of the jackknife technique to National Assessment data was straightforward. In one pass through the data tape, sums of weights for correct responses and total sums of weights were computed within each PSU for a specified cross-classification of subpopulations. Consider for example, the cross-classification of Region, STOC, Sex, Color, and Parents' Education, yielding a five dimensional table (4 x 7 x 2 x 3 x 4), each cell of which gets a sum of weights correct $\hat{Y}_{hi}$ (r, s, t, u, v) and a total sum of weights $\hat{M}_{hi}$ (rstuv) for each PSU (hi). All of the P-Values, ΔP-Values, and their variance estimators can be easily computed from sums and differences of these stored quantities. The balanced effects are functions of one- and two-way marginal sums from the $\hat{M}$ and $\hat{Y}$ tables. The variances and covariances of the $\hat{\beta}$s are formed from within-stratum PSU differences among these one- and two-way marginal totals.

While the jackknife replication procedure was first introduced by Quenouille [ref. 1], as a technique for reducing sampling bias in nonlinear statistics, this feature is probably not of primary importance in large samples such as National Assessment, since combined-ratio type estimators like NAEP's P-Values, ΔP-Values, and Balanced Effects should be relatively free of sampling bias. Some empirical justification for this contention can be gained by contrasting the jackknifed versions of these statistics with the original estimates. For most of NAEP's reported statistics, these differences are negligible. For these reasons, National Assessment reports unjackknifed statistics along with the associated jackknife variance estimator.

In the following section a summary analysis of Year 01 sampling errors is presented. These results are extracted from a paper by Chromy, Moore, and Clemmer [ref. 6]. The results are presented in terms of design effects or the ratio of NAEP variance estimates to simple random sampling variances.

Summary Analysis of Year 01 Sampling Errors

National design effects were estimated by Chromy et al. for 149 science and writing P-values. The median design effect estimate for the 149 exercises examined was 2.38, with the majority falling between 1.50 and 3.00. Table 2-1 shows that 82 percent were 3.50 or less, and 94 percent were 4.00 or less. Table 2-2 presents median national design effects and ranges in national design effects for various subgroups of exercises classified by age group, administration mode, and subject matter area.

Design effects for group-administered exercises were higher than those for individually administered exercises due to more clustering of the sample respondents. Each group package was administered once in each PSU to a group of 12 students selected from a single school. For individual packages, the 9 respondents selected from each PSU were spread across several schools.

The estimated design effects for 13-year-olds were smaller than those for 9-year-olds, while the 17-year-old exercise effects were smaller than those for either 9- or 13-year-olds. A plausible explanation for such a trend is that high schools are more heterogeneous in terms of students than are junior high schools, and junior high schools are more heterogeneous than the elementary schools.

Median design effects for size of community (SOC) subpopulations defined by poststratification are shown in table 2-3. As with national design effects, the median effects for SOC subpopulations are higher for group-administered exercises than for individually administered exercises. There is possibly a tendency for metropolitan and urban area median design effects to be smaller than those for more sparsely populated medium city and rural (small place) subpopulations.

The design effects discussed in the Chromy paper reflect the combined effects of clustering of the sample, unequal weighting of sample respondents, stratification, and other sample design and estimation factors. The effect of unequal weighting of sample respondents was estimated to be from 1.3 to 1.6, depending upon the exercise.

Table 2-1. Distribution of national design effects

Design Effect	Number	Percent
< 1.00	1	1%
1.00 - 1.50	16	11%
1.51 - 2.00	29	19%
2.01 - 2.50	43	30%
2.51 - 3.00	32	21%
3.01 - 3.50	8	5%
3.51 - 4.00	10	7%
4.01 - 4.50	5	3%
4.51 - 5.00	3	2%
> 5.00	2	1%
Total	149	100%

Table 2-2. Median design effects for national P-value estimates

Age	Administration Mode	Subject Area	Number of Exercises	Median Design Effects	Range of Design Effects	Mean Number of Respondents
9	Group	Science	30	2.68	1.92-4.94	2,442
13	Group	Science	27	2.26	1.31-6.01	2,415
17	Group	Science	10	1.81	.90-2.51	2,122
17	Individual	Science	1	1.13		579
26 to 35	Individual	Science	16	2.57	1.38-4.08	878
9	Group	Writing	24	2.81	1.51-3.80	2,426
13	Group	Writing	5	4.36	1.93-10.88	2,416
9	Individual	Writing	13	2.21	1.45-2.68	1,817
13	Individual	Writing	23	1.89	1.24-2.88	1,863

Table 2-3. Median design effects for size of community (SOC) subpopulation P-value estimates

				Median Design Effects for SOC Categories			
Age	Administration Mode	Subject Area	Number of Exercises	Big City	Urban Fringe	Medium City	Small Place
9	Group	Science	30	2.26	2.01	2.56	3.38
13	Group	Science	27	2.43	2.20	2.14	1.90
26 to 35	Individual	Science	16	1.91	2.25	1.47	1.86
9	Group	Writing	24	2.04	2.18	2.41	2.86
13	Group	Writing	5	3.79	2.95	3.82	3.69
9	Individual	Writing	13	1.75	1.97	2.66	1.91
13	Individual	Writing	23	1.22	1.37	1.75	2.38

Design effects for adults 26 to 35 years of age were about equal to those for 9-year-olds, possibly reflecting a similar intracluster correlation for the household sample due to small, compact clusters and variable housing patterns within PSUs.

No apparent difference was observed between design effects for science and writing exercises. This comparison is tenuous because of the small number of group-administered writing exercises and the fact that no individually administered science exercises were examined in the Chromy study.

Tables 2-3, 2-4, and 2-5 present median design effects for subpopulations defined by regional strata, and for sex and size of community subpopulations defined by poststratification. Median design effects for subpopulation estimates are of about the same magnitude or slightly smaller than the median effects for national estimates.

The largest median design effects for 9- and 13-year-old writing exercises seem to occur in the Southeast region (table 2-3).

No consistent trend was noted among the median design effects for males and females (see table 2-4).

Table 2-4. Median design effects for regional subpopulation P-value estimates (9- and 13-year-olds only)

Age	Administration Mode	Subject Area	Number of Exercises	Median Design Effects by Region Northeast	Southeast	Central	West
9	Group	Writing	24	1.89	2.93	2.32	2.65
13	Group	Writing	5	3.05	3.65	3.50	2.65
9	Individual	Writing	13	2.34	1.30	1.85	2.17
13	Individual	Writing	23	1.64	2.11	1.61	1.35

Table 2-5. Median design effects for sex subpopulation P-value estimates

Age	Administration Mode	Subject Area	Number of Exercises	Median Design Effects by Sex Males	Females
13	Group	Science	20	2.57	2.25
26 to 35	Individual	Science	15	2.08	2.20
9	Group	Writing	24	2.74	2.54
13	Group	Writing	5	2.95	4.38
9	Individual	Writing	13	2.27	2.03
13	Individual	Writing	23	1.80	1.84

Some major revisions in the National Assessment sample design occurred in Year 02. The first principal change involved doubling the within-PSU sample size and halving the number of primary units. The planned number of administrations per individual package was increased to 20 per PSU.

The second major change involved the use of controlled selection of the primary sample to permit stratification by State as well as by the previously discussed set of stratification variables. The implications of this second change for variance estimation in Year 02 are explored in the next two chapters.

References

1. Quenouille, M. H. (1956). Notes on bias in estimation. <u>Biometrika</u> 43, 353-360.

2. Tukey, J. W. (1958). Bias and confidence is not-quite large samples: Abstract. <u>Ann</u>. <u>Math</u>. <u>Statist</u>. 29, 614.

3. Arvesen, J. N. (1969). Jackknifing U-statistics. <u>Ann</u>. <u>Math</u>. <u>Statist</u>. 40, 2076-2100.

4. Brillinger, R. B. (1976). Personal communication. Math. Dept. The University of Auckland. Auckland, New Zealand.

5. Cochran, W. G. (1963). <u>Sampling Techniques</u>, 2nd. ed., New York: John Wiley and Sons.

6. Chromy, J. R., Moore, R. P., and Clemmer, A. (1972). Design Effects in the National Assessment of Educational Progress Surveys. Proceedings of the <u>Social Statistics Section of the American Statistical Association</u>. 48-52.

Chapter 3
CONTROLLED SELECTION: IMPLICATIONS FOR YEAR 02 VARIANCES

Introduction to Controlled Selection

Controlled selection can be viewed as a probability proportional to size (PPS), without-replacement sampling scheme for selecting primary sampling units (PSUs) subject to controls beyond what is possible with stratified random sampling. Stratified random samples, where the sample sizes in the various strata are required to be proportional to corresponding strata sizes, are generally more efficient than purely random samples. The effectiveness of such stratification is increased as the number of strata increases. To take full advantage of the potential gain from stratification and to guarantee representation for various subpopulations (domains) of interest, Goodman and Kish [ref. 1] developed controlled selection as a means of allocating primary units to strata proportional to size when the number of units was smaller than the number of substrata generated. Jessen [ref. 2] in his recent paper on "Probability Sampling with Marginal Constraints" presents an algorithm for selecting primary units with stratification in several directions.

Hess and Srikantan [ref. 3] considered controlled selection designs with equal probability selection of PSUs within control cells (cells of the two-way stratification array). In a Monte Carlo sampling experiment they compared variance approximations for an estimated ratio using the methods of successive differences, paired differences, and balanced half samples. It was found that these approximations substantially overestimated the variance for three of the four statistics studied.

The results presented in the following sections relate to variance estimation for a design utilizing a controlled selection algorithm to construct allocation patterns. After one of these patterns is chosen, the required number of first-stage units is selected from each control cell with PPS and without replacement. The general population structure and sample design are presented in the following section. Section 3 develops the familiar Horvitz-Thompson [ref. 4] type estimator for a population total and derives an analytic expression for the variance of such a linear statistic. Section 4 gives conditions on the set of allocation patterns and the subsequent PSU selection scheme, which provide for unbiased variance estimation. The appropriate Yates-Grundy (Y-G) [ref. 5] type variance estimator is shown to be unbiased when the aforementioned conditions are met. Chapter 4 describes a computer simulation model used to generate data for a Monte Carlo sampling experiment patterned after the Research Triangle Institute's survey design for Year 02 of National Assessment. Three variance approximations are proposed in chapter 4 as alternatives to the (Y-G) estimator. Empirical results of the Monte Carlo study are presented. The bias, mean square error, and distributional properties of four alternative variance estimators for a ratio statistic are studied.

General Population Structure and the Sample Design

Consider a population of first-stage listing units, which have been stratified in two directions. If $r = 1(1)R$ and $c = 1(1)C$ denote levels of the row and column stratification variables, then N_{rc} will represent the number of listing units in cell (rc) of this two-way stratification array. Let Y_{rck} be a characteristic of interest possessed by the k-th lising unit in cell (rc). Suppose that X_{rck} is a size measure for listing unit (rck); that is, X_{rck} represents a variable that is known for all $k = 1(1)N_{rc}$ listing units in cell (rc) and is assumed to have positive correlation with the unknown variable of interest Y_{rck}. The relative size of cell (rc) is, therefore,

$$A_{rc} = (X_{rc+}/X_{+++}), \tag{3.1}$$

where a "plus" replacing a subscript indicates summation over the levels of that subscript. An allocation strictly proportional to X of n primary sampling units (PSUs) to the RC cells of our two-way array would yield a fractional sample size $n_{rc} = nA_{rc}$ for "control cell" (rc).

Various algorithms, which will be collectively referred to as controlled selection schemes, yield samples with an expected allocation of PSUs to cells strictly proportional to their measures of size. These algorithms produce a set of S allocation patterns with the s-th pattern consisting of a set of integer allocations $\{n_{rc}(s)$; for $r = 1(1)R$ and $c = 1(1)C\}$. Each of the S patterns has a selection probability (α_s)

assigned to it, such that the expected sample size for any cell over all patterns is

$$\sum_{s=1}^{S} \alpha_s n(s)_{rc} = n_{rc} \equiv nA_{rc}, \tag{3.2}$$

the strictly proportional allocation. Additional cell and marginal constraints are usually imposed upon the allocation patterns; for example, the cell allocations $n(s)_{rc}$ are required to satisfy the following sets of inequalities:

$$| n(s)_{rc} - n_{rc} | < 1, \tag{3.3a}$$

$$| n(s)_{+c} - n_{+c} | < 1, \tag{3.3b}$$

$$| n(s)_{r+} - n_{r+} | < 1. \tag{3.3c}$$

These inequalities require that the integer allocations to cells, column margins, and row margins deviate from the strictly proportional allocations by less than one PSU.

We will consider samples with $n(s)_{rc}$ primary sampling units selected without replacement and with probabilities strictly proportional to size. That is, if the s-th pattern is chosen, then $n(s)_{rc}$ PSUs are selected from control cell (rc) with probabilities,

$$\pi(s)_{rck} = n(s)_{rc} \; (X_{rck}/X_{rc+}) \equiv n(s)_{rc} \; A_{rck}, \tag{3.4}$$

and without replacement. The unconditional probability over all patterns for selecting first-stage listing unit (rck) is, therefore,

$$\pi_{rck} = \sum_{s=1}^{S} \alpha_s \, \pi(s)_{rck} = \sum_{s=1}^{S} \alpha_s \, n(s)_{rc} \cdot A_{rck} = n_{rc} \; A_{rck}. \tag{3.5}$$

With Y_{rck} denoting the variate value of interest associated with listing unit (rck), we will be concerned with estimation for the population total

$$Y = Y_{+++} \equiv \sum_{r=1}^{R} \sum_{c=1}^{C} \sum_{k=1}^{N_{rc}} Y_{rck}. \tag{3.6a}$$

Estimation Theory

The following Horvitz-Thompson estimator for the population total (Y) will be considered in subsequent sections,

$$\hat{Y} = \sum_{r=1}^{R} \sum_{c=1}^{C} \sum_{k=1}^{n(s)_{rc}} \hat{Y}_{rck}/\pi_{rck} . \qquad (3.7a)$$

Notice that the summation in k is over those $n(s)_{rc}$ (possibly zero) listing units selected from cell (rc). The ^ over variate value Y_{rck} indicates an estimate based on subsequent stages of sampling. Recall that π_{rck} is the unconditional probability of selecting the listing unit (rck) as defined in equation (3.5).

In part of the discussion that follows, it will be convenient to use a single subscript, say $\ell = 1(1)L$, to index the two-way array of control cells. This allows one to write

$$Y = \sum_{\ell=1}^{L} \sum_{k=1}^{N_\ell} Y_{\ell k}, \qquad (3.6b)$$

in place of equation (3.6a) and

$$\hat{Y} = \sum_{\ell=1}^{L} \sum_{k=1}^{n(s)_\ell} \hat{Y}_{\ell k}/\pi_{\ell k}, \qquad (3.7b)$$

in place of equation (3.7a).

Assuming that the within-PSU stages of sampling lead to unbiased estimates of the PSU totals, it is easy to show that $\hat{Y}$ is unbiased. Notice first that $\hat{Y}$ can be rewritten as,

$$\hat{Y} = \sum_{\ell=1}^{L} n(s)_\ell \, \hat{Y}_\ell/n_\ell, \qquad (3.7c)$$

where

$$\hat{Y}_\ell = \sum_{k=1}^{n(s)_\ell} \hat{Y}_{\ell k}/\pi(s)_{\ell k}$$

is the unbiased Horvitz-Thompson estimator for the cell (ℓ) total $Y_\ell = \sum_{k=1}^{N_\ell} Y_{\ell k}$. Taking the conditional expectation of $\underline{Y}$ given the PSU allocations $n_\ell(s)$ from pattern (s) we find,

$$E\{\hat{Y}|n_\ell(s)\} = \sum_{\ell=1}^{L} n_\ell(s)\, Y_\ell/n_\ell . \qquad (3.8)$$

Recalling the definition of $n_\ell \equiv \sum_{s=1}^{S} \alpha_s\, n_\ell(s) \equiv E_s\{n_\ell(s)\}$, one sees that $\hat{Y}$ is indeed unbiased.

To derive the variance of $\hat{Y}$, the following partitioning will be useful:

$$\mathrm{Var}(\hat{Y}) = \mathrm{Var}[E\{\hat{Y}|n_\ell(s)\}] + E[\mathrm{Var}\{\hat{Y}|n_\ell(s)\}]. \qquad (3.9)$$

The first term in equation (3.9) is $\mathrm{Var}\{\sum_{\ell=1}^{L} n_\ell(s) Y_\ell/n_\ell\}$ from equation (3.7).

Therefore,

$$\mathrm{Var}[E\{\hat{Y}|n_\ell(s)\}] = \sum_{\ell=1}^{L} \mathrm{Var}\{n_\ell(s)\}\, Y_\ell^2/n_\ell^2 +$$

$$\sum_{\ell=1}^{L} \sum_{\ell'\neq\ell} \mathrm{Cov}\{n_\ell(s), n_\ell(s)\}\, Y_\ell Y_{\ell'}/\bar{n}_\ell n_{\ell'} . \qquad (3.10)$$

Letting

$$E\{n_\ell(s)\, n_{\ell'}(s)\} \equiv \sum_{s=1}^{S} \alpha_s\, n_\ell(s)\, n_{\ell'}(s) \equiv n_{\ell\ell'} ,$$

and recalling that $\sum_{\ell=1}^{L} n_\ell(s) = \sum_{\ell=1}^{L} n_\ell = n$, we find that

$$\sum_{\ell'=\ell} \mathrm{Cov}\{n_\ell(s), n_{\ell'}(s)\} = \sum_{\ell'\neq\ell} (n_{\ell\ell'} - n_\ell n_{\ell'}) = -\mathrm{Var}\{n_\ell(s)\} . \qquad (3.11)$$

The result in equation (3.11) allows one to write the between-control-cell contribution to equation (3.9) in a form reminiscent of the familiar Yates-Grundy variance expression; namely,

$$\mathrm{Var}[E\{\hat{Y}|n_\ell(s)\}] = \sum_{\ell=1}^{L-1} \sum_{\ell'=\ell+1}^{L} (n_\ell n_{\ell'} - n_{\ell\ell'}) \left(\frac{Y_\ell}{n_\ell} - \frac{Y_{\ell'}}{n_{\ell'}} \right)^2 . \quad (3.12)$$

The variance form in equation (3.12) shows clearly that the between-cell source of variation in $\hat{Y}$ is minimized when n_ℓ (the expected sample size for cell ℓ) is strictly proportional to Y_ℓ (the cell ℓ total).

Returning to the second term in our partitioning of $\mathrm{Var}(\hat{Y})$, equation (3.9) we see that,

$$\mathrm{Var}\{\hat{Y}|n_\ell(s)\} = \sum_{\ell=1}^{L} \{n_\ell(s)/n_\ell\}^2 \, \mathrm{Var}\{\hat{Y}_\ell|n_\ell(s)\} . \quad (3.13)$$

This result is an immediate consequence of equation (3.8), and the fact that PSUs from a particular control cell, (ℓ), are selected independently of those from any other cell, (ℓ'). If $\pi_{\ell kk'}(s)$ denotes the joint inclusion probability for listing units k and k' from cell (ℓ) when $n_\ell(s)$ PSUs are selected, then

$$\mathrm{Var}\{\hat{Y}|n_\ell(s)\} = \sum_{\ell=1}^{L} \{n_\ell(s)/n_\ell\}^2 \sum_{k=1}^{N_\ell-1} \sum_{k'=k+1}^{N_\ell} \{\pi_{\ell k}(s)\pi_{\ell k'}(s) - \pi_{\ell kk'}(s)\} \left(\frac{Y_{\ell k}}{\pi_{\ell k}(s)} - \frac{Y_{\ell k'}}{\pi_{\ell k'}(s)} \right)^2$$

$$+ \sum_{i=1}^{L} \{n_\ell(s)/n_\ell\}^2 \sum_{k=1}^{N_\ell} \sigma^2_{\ell k}/\pi_{\ell k}(s). \quad (3.14)$$

where $\sigma^2_{\ell k}$ denotes the conditional variance of the estimated PSU total $\hat{Y}_{\ell k}$ given that listing unit (ℓk) belongs to the first-stage sample. Since the conditional inclusion probability $\pi_{\ell k}(s) = \{n_\ell(s)/n_\ell\}\pi_{\ell k}$ where $\pi_{\ell k}$ is the corresponding unconditional inclusion probability, one can recast equation (3.14) as

$$\mathrm{Var}\{\hat{Y}|n_{\ell}(s)\} = \sum_{\ell=1}^{L} \sum_{k=1}^{N_{\ell}-1} \sum_{k'=k+1}^{N_{\ell}} \{\pi_{\ell k}(s)\pi_{\ell k'}(s)-\pi_{\ell kk'}(s)\} \left(\frac{Y_{\ell k}}{\pi_{\ell k}} - \frac{Y_{\ell k'}}{\pi_{\ell k'}}\right)^2$$

$$+ \sum_{\ell=1}^{L} \{n(\ell)/n_{\ell}\} \sum_{k=1}^{N_{\ell}} \sigma^2_{\ell k}/\pi_{\ell k} . \qquad (3.15)$$

Letting

$$E_s \{\pi_{\ell kk'}(s)\} \equiv \sum_{s=1}^{L} \alpha_s \pi_{\ell kk'}(s) \equiv \pi_{\ell kk'}$$

denote the unconditional joint inclusion probability for listing units k and k' from control cell (ℓ) and defining

$$RV_{\ell} \equiv \text{Rel-Var } \{n_{\ell}(s)\} = \mathrm{Var}\{n_{\ell}(s)\}/n_{\ell}^2 ,$$

the expectation of equation (3.15) becomes

$$E[\mathrm{Var}\{\hat{Y}|n_{\ell}(s)\}] = \sum_{\ell=1}^{L} \sum_{k=1}^{N_{\ell}-1} \sum_{k'=k+1}^{N_{\ell}} \{(RV_{\ell}+1)\pi_{\ell k}\pi_{\ell k'}-\pi_{\ell kk'}\} \left(\frac{Y_{\ell k}}{\pi_{\ell k}} - \frac{Y_{\ell k'}}{\pi_{\ell k'}}\right)^2$$

$$+ \sum_{\ell=1}^{L} \sum_{k=1}^{N_{\ell}} \sigma^2_{\ell k}/\pi_{\ell k} . \qquad (3.16)$$

Finally, combining the results in equations (3.12) and (3.16), we can specify the variance of $\hat{Y}$ as follows:

$$\mathrm{Var}(\hat{Y}) = \sum_{\ell=1}^{L-1} \sum_{\ell'=\ell+1}^{L} (n_{\ell}n_{\ell'}-n_{\ell\ell'}) \left(\frac{Y_{\ell}}{n_{\ell}} - \frac{Y_{\ell'}}{n_{\ell'}}\right)^2$$

$$+ \sum_{\ell=1}^{L} \sum_{k=1}^{N_\ell -1} \sum_{k'=k+1}^{N_\ell} \{(RV_\ell+1)\pi_{\ell k}\pi_{\ell k'} - \pi_{\ell kk'}\} \left(\frac{Y_{\ell k}}{\pi_{\ell k}} - \frac{Y_{\ell k'}}{\pi_{\ell k'}}\right)^2$$

$$+ \sum_{\ell=1}^{L} \sum_{k=1}^{N_\ell} \sigma^2_{\ell k}/\pi_{\ell k} . \qquad (3.17)$$

Although the expression in equation (3.17) neatly partitions the variance of Y into components due to between-cell, between-PSU-within-cell, and within-PSU variability, a more compact form, which combines the first two terms above, provides more insight into how this variance might be estimated. If $\pi_{\ell k;\ell' k'}$ denotes the unconditional probability that PSU (k) of cell (ℓ) and PSU (k') of cell (ℓ') both belong to the sample where $\ell \neq \ell'$, then

$$\pi_{\ell k;\ell' k'} = \sum_{s=1}^{S} \alpha_s \pi(s)_{\ell k} \pi(s)_{\ell' k'}$$

$$= \sum_{s=1}^{S} \alpha_s n(s)_{\ell} A_{\ell k} n(s)_{\ell'} A_{\ell' k'}$$

$$= n_{\ell\ell'} A_{\ell k} A_{\ell' k'} . \qquad (3.18)$$

Having defined this between-cell joint inclusion probability, one can view the first two stages of sampling (patterns and PSUs given the pattern) as a without-replacement selection of $n = \sum_{\ell=1}^{L} n_\ell$ PSUs with varying inclusion probabilities $\pi_{\ell k}$ and with joint inclusion probabilities $\pi_{\ell k;\ell' k'}$ where,

$$\pi_{\ell k;\ell' k'} = \begin{cases} \pi_{\ell kk'} \text{ if } \ell=\ell' \\ n_{\ell\ell'} A_{\ell k} A_{\ell' k'} \text{ if } \ell \neq \ell' . \end{cases} \qquad (3.19)$$

This leads one to the variance expression

$$\mathrm{Var}(\hat{Y}) = 1/2 \sum_{\ell=1}^{L} \sum_{k=1}^{N_\ell} \sum_{(\ell'k') \neq (\ell k)} (\pi_{\ell k}\pi_{\ell'k'} - \pi_{\ell k;\ell'k'}) \left(\frac{Y_{\ell k}}{\pi_{\ell k}} - \frac{Y_{\ell'k'}}{\pi_{\ell'k'}}\right)^2$$

$$+ \sum_{\ell=1}^{L} \sum_{k=1}^{N_\ell} \sigma^2_{\ell k}/\pi_{\ell k} \qquad (3.20)$$

Using the definition in equation (3.19), it is not difficult to show that the first term in equation (3.20) expands into the between-cell and between-PSU-within-cell contributions of equation (3.17). In the following section, the familiar Yates-Grundy variance form in equation (3.20) will be exploited to produce an estimator for Var($\hat{Y}$).

Variance Estimation

If an unbiased estimate, say $\hat{\sigma}^2_{\ell k}$, of the within-PSU variance is available from each sampled PSU, then,

$$\mathrm{Var}\ (\hat{Y}) = 1/2 \sum_{\ell=1}^{L} \sum_{k=1}^{n(s)_\ell} \sum_{(\ell'k') \neq (\ell k)} \left(\frac{\pi_{\ell k}\pi_{\ell'k'} - \pi_{\ell k;\ell'k'}}{\pi_{\ell k,\ell'k'}}\right) \left(\frac{\hat{Y}_{\ell k}}{\pi_{\ell k}} - \frac{\hat{Y}_{\ell'k'}}{\pi_{\ell'k'}}\right)^2$$

$$+ \sum_{\ell=1}^{L} \sum_{k=1}^{n(s)_\ell} \hat{\sigma}^2_{\ell k}/\pi_{\ell k} \qquad (3.21)$$

is an unbiased estimator for Var($\hat{Y}$) when $\pi_{\ell k;\ell'k'} > 0$ for all pairs $(\ell k;\ \ell'k')$ of listing units in the frame. This last condition requires that each pair of listing units in the frame has a chance of being in the sample and is the downfall of most controlled-selection designs when it comes to variance estimation. To satisfy this condition, the set of allocation patterns must be such that:

1. $n_{\ell\ell'} > 0$ for all pairs of nonempty cells $(\ell\ell')$ in the two-way stratification array. This implies that each pair of control cells is represented in at least one of the allocation patterns.

2. If N_ℓ (the number of listing units in cell ℓ) exceeds one, then we require that $\Pr\{n_\ell(s) \geq 2\} > 0$. That is, if a cell contains two or more listing units, then at least one pattern must assign two PSUs to the cell.

Meeting the first of these two conditions presents no major difficulty. Algorithms like Jessen's [ref. 3] "Method 3" generate an enlarged set of allocation patterns, which satisfy the inequality constraints in equation (3.1) and at the same time give all pairs of nonempty cells a positive probability of being in the sample. The second requirement for unbiased variance estimation is more severe, since it runs contrary to the basic advantage of controlled selection designs; that is, having more control cells than PSUs. If a cell's expected allocation n_ℓ is less than one, although it contains two or more listing units ($N_\ell \geq 2$), then the cell inequality $| n_\ell(s) - n_\ell | < 1$ in equation (3.1) does not allow $n_\ell(s)$ to be two or more. One way of solving this problem would be to restrict our attention to designs with $n_\ell \geq 1$ for all cells with $N_\ell \geq 2$, but this would eliminate most of the situations where control beyond stratification is desired. A more acceptable solution allows some patterns with $n_\ell(s) = 2$ when $n_\ell < 1$. Although the cell inequality is violated for cells with expected allocations less than 1, our experience indicates that it should still be possible to satisfy the marginal constraints, as long as the expected marginal allocations exceed two PSUs.

When some of the pairs $(\ell k;\ell' k')$ of listing units have no chance of appearing in the sample, the estimator in equation (3.21) will underestimate the true variance by an amount,

$$\sum^{\phi}_{(\ell k;\ell' k')} \pi_{\ell k}\pi_{\ell' k'} \left\{ \left(\frac{Y_{\ell k}}{\pi_{\ell k}} - \frac{Y_{\ell' k'}}{\pi_{\ell' k'}}\right)^2 + \frac{\sigma^2_{\ell k}}{\pi^2_{\ell k}} + \frac{\sigma^2_{\ell' k'}}{\pi^2_{\ell' k'}} \right\} , \qquad (3.22)$$

where $\sum^{\phi}_{(\ell k;\ell' k')}$ denotes summation over those pairs of listing units $(\ell k;\ell' k')$ such that $\pi_{\ell k;\ell' k'} = 0$. The portion of this downward bias due to singleton cells (cells with $N_\ell \geq 2$ but $\Pr\{n_\ell(s) \geq 2\} = 0$) can be expressed as

$$\sum_{\ell}^{(s)} \sum_{k=1}^{N_\ell} A_{\ell k} \left(\frac{Y_{\ell k}}{A_{\ell k}} - Y_\ell \right)^2 + \sum_{\ell}^{(s)} \sum_{k=1}^{N_\ell} (n_\ell - \pi_{\ell k}) \sigma^2_{\ell k} / \pi_{\ell k} \; , \tag{3.23}$$

where $\sum_{\ell}^{(s)}$ represents summation over all singleton cells. That part of equation (3.22) due to pairs of cells having no chance of being in the sample is,

$$\sum_{(\ell,\ell')}^{\phi} n_\ell n_{\ell'} \sum_{k=1}^{N_\ell} \sum_{k'=1}^{N_{\ell'}} A_{\ell k} A_{\ell k'} \left(\frac{Y_{\ell k}}{\pi_{\ell k}} - \frac{Y_{\ell' k'}}{\pi_{\ell' k'}} \right)^2 + \sum_{(\ell,\ell')}^{\phi} \left(n_{\ell'} \sum_{k=1}^{N_\ell} \sigma^2_{\ell k} / \pi_{\ell k} \right.$$

$$\left. + \, n_\ell \sum_{k'=1}^{N_{\ell'}} \sigma^2_{\ell' k'} / \pi_{\ell' k'} \right) , \tag{3.24}$$

where $\sum_{(\ell,\ell')}^{\phi}$ denotes summation over those pairs of control cells with no chance of being in the sample. As suggested earlier, this source of bias can be eliminated by adopting an algorithm like Jesson's "Method 3," which guarantees that all nonempty pairs of cells have a chance of appearing in the sample.

The underestimation that occurs as a result of singleton cells, equation (3.23), can be compensated for by a procedure analogous to collapsing strata when a single PSU is selected per stratum. One such collapsing scheme, involving successive differences, takes the following form:

$$\sum_{c=1}^{C} \sum_{t=1}^{m_c(s)} \{ \sqrt{n_{tc}}\, y_{tc1} - \sqrt{n_{t+1,c}}\, y_{t+1;c1} \}^2 / 2 \; , \tag{3.25}$$

$$\text{with } m_c(s) + 1 \equiv 1 \; ,$$

where $t = 1(1)m_c(s)$ indexes the singleton cells from column (c) represented in pattern (s), and $y_{tc1} = (\hat{Y}_{tc1}/\pi_{tc1})$ denotes the expanded PSU total from singleton cell (tc). This scheme presumes that the column classification represents the more important stratification dimension, and that the row categories can be arranged in a circular array with adjacent categories assumed to be more alike than nonadjacent categories. Adding this term to equation (3.21) and assuming for the moment that $n_{\ell\ell'} > 0$ for all pairs of cells, the bias becomes,

$$\sum_{s=1}^{S} \alpha_s \sum_{c=1}^{C} \sum_{t=1}^{\overset{m(s)}{c}} \left(\frac{Y_{tcl}}{\sqrt{n_{tcl.}}} - \frac{Y_{t+1,cl}}{\sqrt{n_{t+1,cl}}} \right)^2 /2 + \sum_{\ell}^{(s)} \sum_{k=1}^{N_\ell} \sigma^2_{\ell k} \qquad (3.26)$$

The within-PSU bias remaining in equation 3.26 can be eliminated by subtracting,

$$\sum_{\ell}^{(s)} \sum_{k=1}^{\overset{n(s)}{\ell}} \hat{\sigma}^2_{\ell k}/\pi_{\ell k} , \qquad (3.27)$$

from equation (3.21). If $\Sigma^{(\bar{s})}_{\ell}$ denotes summation over nonsingleton cells, then equation (3.21) becomes with the bias adjustments in equations (3.25) and (3.27),

$$\begin{aligned} \mathrm{var}(\hat{Y})_{adj} = & \sum_{(\ell k;\ell' k')}^{\neq} w_{\ell k;\ell' k'} (y_{\ell k} - y_{\ell' k'})^2/2 \\ & + \sum_{c=1}^{C} \sum_{t=1}^{\overset{m(s)}{c}} \{\sqrt{n_{tc}}\, y_{tcl} - \sqrt{n_{t+1,c}}\, y_{t+1,cl}\}^2/2 \qquad (3.28) \\ & + \sum_{\ell}^{(\bar{s})} \sum_{k=1}^{\overset{n(s)}{\ell}} \hat{\sigma}^2_{\ell k}/\pi_{\ell k} , \end{aligned}$$

where $w_{\ell k;\ell' k'}$ represents the Yates-Grundy "variance weight" in equation (3.21) and $y_{\ell k}$ is the expanded PSU (ℓk) total. The bias-adjusted estimator in equation (3.28) overestimates the true variance of $\hat{Y}$ by the first term of equation (3.26). If some pairs of cells have no chance of being in the sample, then the quantity in equation (3.24) makes a negative contribution to the bias of equation (3.28).

Finally, if the-within PSU variation is not estimable, as when only one second stage unit is selected per PSU, an additional negative bias is incurred; namely

$$\sum_{\ell}^{(\bar{s})} \sum_{k=1}^{N_{\ell}} \sigma^2_{\ell k} . \qquad (3.29)$$

References

1. Goodman, R. and Kish, L. (1950). Controlled Selection -- a technique in probability sampling. J. Amer. Stat. Assoc. 45, 350-372.

2. Jessen, R. J. (1970). Probability Sampling with Marginal Constraints. J. Amer. Stat. Assoc. 65, 776-795.

3. Hess, I. and Srikantan, K. S. (1966). Some Aspects of the Probability Sampling Technique of Controlled Selection. Health Services Research, Vol. I, No. 1. 8-52..

4. Horvitz, D. G., and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. J. Amer. Stat. Assoc. 47, 663-685.

5. Yates, F. and Grundy, P. M. (1953). Selection without replacement from within strata with probability proportional to size. J. Royal Stat. Soc. (B). 15, 235-261.

Chapter 4: SIMULATION STUDY OF ALTERNATIVE YEAR 02 VARIANCE ESTIMATORS

Simulation Model

In the following section, the design of a Monte Carlo sampling experiment is presented. Modeled after the Research Triangle Institute's (RTI's) sample design for Year 02 of National Assessment, 1,000 samples of 31 PSUs (counties or groups of counties) were selected from the two-way state by "major strata" grid shown in table 4-1. Table 4-1 shows the expected PSU allocations (n_{rc}) and numbers of first-stage units in the frame (N_{rc}) for the 15 States comprising NAEP's western region. The seven major strata represent a combination of size of community and socioeconomic status categories.

In order to distribute the sample proportionally across the major strata and at the same time guarantee that each State would be represented by at least one PSU (a NAEP requirement), controlled selection was used to generate 33 PSU allocation patterns that met these requirements. The 33 pattern probabilities for our design were converted into integer allocations by multiplying each by 1,000. In this fashion, the number of times a pattern was represented in the 1,000 samples was made strictly proportional to its selection probability. For each of the 221 first-stage listing units in the West, an estimate of its population of 17-year-olds was produced using 1960 census projections. These estimates were used as size measures in connection with a PPS without replacement scheme to select PSUs from the cells of table 4-1.

A data vector consisting of the actual number of 17-year-olds and the number responding correctly to several NAEP test exercises was generated for each of our 221 first-stage listing units. This data set was based on 1970 census figures for numbers of 17-year-olds and on estimated P-values (proportions correct) by State and major stratum margins for the selected NAEP exercises. To produce P-values (P_{rc}) consistent with actual census totals and the observed State and SOC marginal P-values for selected NAEP Year 02 exercises, an iterative proportional fitting technique was employed [ref. 1]. These fitted cell P-values were used along with the listing unit

Table 4-1. Primary sample allocation summary for the West

State Stratum	Expected Sample Sizes by Major Stratum Number 1	2	3	4	5	6	7	Total PSU's Allocated (No.)	Total PSU's In Frame
Alas.					.372 (1)		.628 (2)	1	(3)
Ariz.		.716 (2)				.150 (1)	.134 (4)	1	(7)
Calif.	4.000 (4)	2.189 (5)	.567 (2)	.090 (1)	2.601 (16)		.553 (13)	10	(41)
Colo.		.203 (1)	.471 (4)	.096 (1)	.099 (2)		.131 (5)	1	(13)
Hawaii		.868 (1)			.132 (1)			1	(2)
Idaho					.404 (3)		.596 (6)	1	(9)
Mont.					.482 (4)		.518 (5)	1	(9)
Nev.					.806 (2)		.194 (1)	1	(3)
N.M.		.268 (1)		.228 (2)	.144 (2)	.148 (1)	.212 (4)	1	(10)
Okla.		.735 (2)		.480 (4)	.070 (1)	.461 (4)	.254 (8)	2	(19)
Oreg.		.343 (2)	.122 (1)		.185 (2)		.350 (9)	1	(14)
Texas		2.400 (6)	.325 (3)	1.106 (13)	.530 (8)	1.241 (14)	.398 (20)	6	(64)
Utah		.447 (1)	.149 (1)		.215 (2)		.189 (3)	1	(7)
Wash.		.831 (2)	.366 (2)		.392 (4)		.411 (8)	2	(16)
Wyo.					.568 (2)		.432 (2)	1	(4)
Total	4.000 (4)	9.000 (23)	2.000 (13)	2.000 (21)	7.000 (50)	2.000 (20)	5.000 (90)	31	(221)

totals to produce a variate value Y_{rck} corresponding to the number of correct responses from listing unit (rck).

The following model was used to generate a new value for Y_{rck} each time listing unit (rck) appeared in one of our replicated samples:

$$Y_{rck} = P_{rck} M_{rck} + e_{rck} \tag{4.1}$$

where the errors e_{rck} for different listing units (k) from control cell (rc) are uncorrelated with zero expectation and variance

$$(e^2_{rck}|M_{rck}) = \sigma^2_{rc} M^g_{rck}. \tag{4.2}$$

Models similar to equation (4.1) have been used by J. Durbin [ref. 2], J. N. K. Rao [ref. 3], and numerous others to study the properties of ratio estimators for P_{rc}. We will restrict our attention to models with g=1 anc proceed to motivate a particular choice for the value of σ^2_{rc}. Our method of choosing a value for σ^2_{rc} will be to propose a plausable model for the sampling variance of the estimated cell P-value $\hat{P}_{rc}$ and then find a value of σ^2_{rc} consistent with such a model.

Suppose for the moment that listing units were selected with replacement and with probabilities strictly proportional to known sizes M_{rck}. Then the unbiased estimator

$$\hat{P}_{rc} = \sum_{k=1}^{n_{rc}(s)} P_{rck} \Big/ n_{rc}(s) \tag{4.3}$$

with

$$P_{rck} \equiv Y_{rck} \Big/ M_{rck}$$

has variance

$$\mathrm{Var}_s\,(\hat{P}_{rc}) = \Sigma^2_{rc} \Big/ n_{rc}(s) \tag{4.4}$$

where

$$\Sigma^2_{rc} = \sum_{k=1}^{N_{rc}} M_{rck} \; (P_{rck} - P_{rc})^2 \Big/ M_{rc+}$$

If, in addition, we define

$$\Sigma^2_{(rc)k} = \sum_{k=1}^{N_{rc}} M_{rck} \; P_{rck} \; (1-P_{rck}) \Big/ M_{rc+} \qquad (4.5)$$

as the within-listing unit variance component, then it is easy to see that

$$\Sigma^2_{rc} + \Sigma^2_{(rc)k} = P_{rc} \; (1-P_{rc}) \qquad (4.6)$$

Now, we define

$$\delta_{rc} = \Sigma^2_{rc} \Big/ (\Sigma^2_{rc} + \Sigma^2_{(rc)k}) \qquad (4.7a)$$

or

$$\delta_{rc} = \Sigma^2_{rc} \Big/ P_{rc} \; (1-P_{rc}) \qquad (4.7b)$$

as the within-listing unit correlation coefficient. These definitions allow us to write

$$\mathrm{Var}_s \; (\hat{P}_{rc}) = P_{rc} \; (1-P_{rc}) \; \delta_{rc} \Big/ n_{rc}(s) \qquad (4.8)$$

With the variance formula in 4.8 representing a plausable model for the between-listing unit variation in $\hat{P}_{rc}$, the estimated proportion correct from cell (rc), our goal is to specify a value for σ^2_{rc} in the data generation model equation (4.1) such that the average sampling variance of the ratio estimator

$$\hat{P}_{rc} = (\hat{Y}_{rc+} / \hat{M}_{rc+}) \tag{4.9}$$

with

$$\hat{Y}_{rc+} = \sum_{k\varepsilon s} Y_{rck} / \pi_{rck}(s)$$

and

$$\hat{M}_{rc+} = \sum_{k\varepsilon s} M_{rck} / \pi_{rck}(s)$$

is approximately equal to equation (4.8). Recalling the form of our model for Y_{rck} we note that

$$\hat{Y}_{rc+} = P_{rc} \hat{M}_{rc+} + \hat{e}_{rc+} . \tag{4.10}$$

Since $\mathcal{E}(e_{rck} | M_{rck}) = 0$, it is clear that

$$\varepsilon \, \text{Var}_s \, (\hat{Y}_{rc+}) = P_{rc}^2 \, \text{Var}_s \, (\hat{M}_{rc+}) + \varepsilon \, \text{Var}_s (\hat{e}_{rc+}). \tag{4.11a}$$

and

$$\varepsilon \, \text{Cov}_s \, (\hat{Y}_{rc+}, \hat{M}_{rc+}) = P_{rc} \, \text{Var}_s \, (\hat{M}_{rc+}) . \tag{4.11b}$$

Using these two results and the Taylor series approximation for the variance of a ratio, namely,

$$\varepsilon \, \text{Var}_s \, (\hat{P}_{rc}) \doteq M_{rc+}^{-2} \Big\{ \varepsilon \, \text{Var}_s \, (\hat{Y}_{rc+}) + P_{rc}^2 \, \text{Var}_s \, (\hat{M}_{rc+}) - 2P_{rc} \, \varepsilon \text{Cov}_s \, (\hat{Y}_{rc+}, \hat{M}_{rc+}) \Big\}, \tag{4.12}$$

we find that

$$\varepsilon \, \text{Var}_s \, (\hat{P}_{rc}) \doteq M_{rc+}^{-2} \, \varepsilon \, \text{Var}_s \, (\hat{e}_{rc+}) . \tag{4.13}$$

Now we can evaluate the expression in equation (4.13) in terms of the data generation model equation (4.1) and compare the results to our variance model in equation (4.8). First we recall the Yates-Grundy version of $\mathrm{Var}_s\ (\hat{\varepsilon}_{rc+})$; that is,

$$\mathrm{Var}_s\ (\hat{e}_{rc+}) = \sum_{k=1}^{N_{rc}} \sum_{k'\neq k} \{ \pi_{rck}(s)\ \pi_{rck'}(s) - \pi_{rckk'}(s) \}\ x$$

$$\left(\frac{e_{rck}}{\pi_{rck}(s)} - \frac{e_{rck'}}{\pi_{rck'}(s)} \right)^2 \bigg/ \ 2. \qquad (4.14)$$

Using the independence of ε_{rck} and $\varepsilon_{rck'}$ along with the result that

$$\sum_{k'\neq k}^{N_{rc}} \{ \pi_{rck}(s)\ \pi_{rck'}(s) - \pi_{rckk'}(s) \} = \pi_{rck}(s)\,[\ 1-\pi_{rck}(s)], \qquad (4.15)$$

we have

$$\varepsilon\ \mathrm{Var}_s\ (\hat{e}_{rc+}) = \sum_{k=1}^{N_{rc}} [1-\pi_{rck}(s)]\,\varepsilon(e_{rck}^2|M_{rck}) \,/\, \pi_{rck}(s)\ . \qquad (4.16)$$

Recalling our specification of the error model in equation (4.1), we have assumed that with g=1

$$\varepsilon\ (e_{rck}^2|M_{rck}) = \sigma_{rc}^2\ M_{rck}\ . \qquad (4.17)$$

With the additional assumption that the size measures A_{rck} entering into the inclusion probabilities, $\pi_{rck}(s)$ are roughly proportional to the actual listing unit totals M_{rck} (or $M_{rak} \doteq \beta_{rc}\ A_{rck}$ where β_{rc} represents the ratio $M_{rc+}\ /\ A_{rc+}$), the average variance expression in equation (4.16) reduces to

$$\varepsilon\ \mathrm{Var}_s\ (\hat{e}_{rc+}) \doteq [1-n_{rc}(s) \big/ N_{rc}]\ N_{rc}\cdot M_{rc+}\ \sigma_{rc}^2 \big/ n_{rc}(s)\ . \qquad (4.18)$$

From equation 4.13 we have

$$\varepsilon \operatorname{Var}_s (P_{rc}) \doteq [1-n_{rc}(s) / N_{rc}] N_{rc} \sigma^2_{rc} / n_{rc}(s) M_{rc+} \quad (4.19)$$

Recognizing the term in parentheses above as the finite population-correction term from a simple random sample and recalling our variance models in equations (4.4) and (4.8), we are lead to

$$\sigma^2_{rc} = M_{rc.} \Sigma^2_{rc} = M_{rc.} P_{rc} (1-P_{rc}) \delta_{rc} \quad (4.20)$$

where $M_{rc.} = M_{rc+} / N_{rc}$. This value of σ^2_{rc} yields the following approximate expression for the average variance of $\hat{P}_{rc}$:

$$\varepsilon \operatorname{Var}_s (\hat{P}_{rc}) \doteq [1-f_{rc}(s)] \Sigma^2_{rc} / n_{rc}(s)$$

$$= [1-f_{rc}(s)] P_{rc} (1-P_{rc}) \delta_{rc} / n_{rc}(s) \quad (4.21)$$

with $f_{rc}(s)$ denoting the sampling fraction for cell (rc) in pattern (s).

A variance components analysis of NAEP Year 01 in-school data suggests an average value for the within-listing unit correlation coefficient δ_{rc} of around .015. Substituting this value for δ_{rc} into our expression for σ^2_{rc}, we have arrived at the following computer simulation model for the number of correct responses Y_{rck} from listing unit (rck):

$$Y_{rck} = P_{rc} M_{rck} + d_{rck} \xi \quad (4.22)$$

where ξ is a standard normal error and

$$d_{rck} = \sigma_{rc} M^{\frac{1}{2}}_{rck} = \sqrt{M_{rc.} P_{rc} (1-P_{rc}) \times .015 M_{rck}} \quad (4.23)$$

To further assure that Y_{rck} is an acceptable estimate of the number of correct responses from listing unit rck, we have required that $Y_{rck} \varepsilon [0, M_{rck}]$. This was accomplished by censoring values of ξ, which cause Y_{rck} to exceed these limits. In order to preserve the symmetry of our error distribution, which assures that $\varepsilon \, (Y_{rck} | M_{rck}) = P_{rc} \, M_{rck}$, we have censored values of ξ in a symmetric fashion. The rules for rejecting ξ are, therefore:

1. For $P_{rc} \geq .5$:

$$\text{Reject } \xi \text{ if } |\xi| > (1-P_{rc}) \, M_{rck} \Big/ d_{rck} \, . \tag{4.23a}$$

2. For $P_{rc} < .5$:

$$\text{Reject } \xi \text{ if } |\xi| > P_{rc} \, M_{rck} \Big/ d_{rck} \, . \tag{4.23b}$$

Only for values of P_{rc} close to 0 or 1 will any censoring be required. If, for example, we make the simplifying assumption that $M_{rck} \doteq M_{rc.}$ for all k, then with $P_{rc} = .95$ or .05 the limits are approximately ± 1.87. Values of ξ would be expected to exceed these limits about 6 percent of the time. For P-values $P_{rc} = .90$ or .10 the limits become ± 2.72; these limits would be exceeded about .7 percent of the time.

The model for Y_{rck} described above deviates in two essential respects from the model used in the first version of these results presented at the ASA meetings in Montreal in August 1972. The first and probably most critical difference was in the specification of the error variance $\varepsilon(e^2_{rck} | M_{rck})$. In the Montreal model, we let $e_{rck} = d_{rck} \, \xi$ with ξ a standard normal error and

$$d_{rck} = \sqrt{M_{rc.} \; P_{rc} \; (1-P_{rc}) \; \pi_{rck}(s) \quad [1-\pi_{rck}(s)].} \tag{4.24}$$

The average variance of $\hat{e}_{rc+}$ under this model is

$$\varepsilon \, \mathrm{Var}_s \, (\hat{e}_{rc+}) = M_{rc+} \, P_{rc} \, (1-P_{rc}) \, , \tag{4.25}$$

which leads to the admittedly unappealing result

$$\varepsilon \, \text{Var}_s \, (\hat{P}_{rc}) = P_{rc} \, (1-P_{rc}) \Big/ M_{rc+} \, . \tag{4.26}$$

This unfortunate choice of error variance should have seriously underestimated the between-listing unit within control cell variation. The second change in the new model was to regenerate the value for Y_{rck} each time that listing unit (rck) appeared in one of our replicated samples. This change should cause the simulation to approximate more closely the expected level of between-listing unit variation.

To build within-listing unit variation into our simulation, we first take note of the two stages of sampling within NAEP PSUs. The typical NAEP in-school PSU has two schools representing a particular group package with 12 students selected from each school. Suppose we let $\hat{M}_{rck}$ and $\hat{Y}_{rck}$ denote estimates of the total number of 17-year-olds and the corresponding number of correct responses from listing unit (rck) based on our sample of schools and students. If we assume that schools are selected with replacement and with probabilities proportional to known numbers of 17-year-olds, say $M_{rck\ell}$ for the $\ell = 1(1) \, S_{rck}$ schools in the listing unit (rck) frame, then with simple random sampling of students within schools (ignoring the last stage finite population correction), we have

$$\text{Var}_s \, (\hat{P}_{rck} = \hat{Y}_{rck} \Big/ \hat{M}_{rck}) = \Sigma^2_{(rck)\ell} \Big/ 2 + \Sigma^2_{(rck\ell)m} \Big/ 24 \tag{4.27}$$

where

$$\Sigma^2_{(rck)\ell} = \sum_{\ell=1}^{S_{rck}} M_{rck\ell} \, (P_{rck\ell} - P_{rck})^2 \Big/ M_{rck+}$$

represents the between-school, within-listing unit variance component. The between-students, within-school component takes the form

$$\Sigma^2_{(rck\ell)m} = \sum_{\ell=1}^{S_{rck}} M_{rck\ell} \, P_{rck\ell} \, (1-P_{rck\ell}) \Big/ M_{rck+} \, . \tag{4.28}$$

With these definitions of the within-PSU variance components, it is not difficult to see that

$$\Sigma^2_{(rck)\ell} + \Sigma^2_{(rck\ell),} = P_{rck}\ (1-P_{rck}) \ . \qquad (4.29)$$

This result allows us to define the within-school correlation coefficient

$$\rho_{rck} = \Sigma^2_{(rck)\ell} / [\Sigma^2_{(rck)\ell} + \Sigma^2_{(rck\ell)m}] \qquad (4.30)$$

and to write

$$\mathrm{Var}_s\ (\hat{P}_{rck}) = P_{rck}(1-P_{rck})\ (1 + 11\rho_{rck})/24 \ . \qquad (4.31)$$

The following computer simulation model was built in accordance with the variance model in equation (4.31):

$$\hat{M}_{rck} = M_{rck} + \xi_{rck} \qquad (4.32a)$$

with

$$\varepsilon\ (\xi_{rck} \mid M_{rck}) = 0 \qquad (4.32b)$$

and

$$\varepsilon\ (\xi^2_{rck} \mid M_{rck}) = M^t_{rck} \text{ for } t = 1 \text{ or } 2 \qquad (4.32c)$$

also

$$\hat{Y}_{rck} = P_{rck}\ \hat{M}_{rck} + \eta_{rck} \qquad (4.33a)$$

with

$$\varepsilon\ (\eta_{rck} \mid \hat{M}_{rck}) = 0 \qquad (4.33b)$$

and

$$\varepsilon\ (\eta^2_{rck} \mid \hat{M}_{rck}) = M^2_{rck}\ P_{rck}\ (1-P_{rck}) \times .081 \qquad (4.33c)$$

where the constant .081 represents an average value for the quantity $(1 + 11\rho_{rck}) / 24$ with ρ_{rck} set to .086. The errors in the models above were again obtained from a censored, standard normal error generator.

In the Montreal simulation, we used errors such that

$$\varepsilon \, Var_s \, (\hat{M}_{rck}) = M^2_{rck} \tag{4.34a}$$

and

$$\varepsilon \, Var_s \, (\hat{Y}_{rck}) = P^2_{rck} \, \varepsilon \, Var_s \, (\hat{M}_{rck}) + M_{rck} \, P_{rck} \, (1-P_{rck}) \tag{4.34b}$$

which leads to

$$\varepsilon \, Var_s \, (\hat{P}_{rck}) \doteq P_{rck} \, (1-P_{rck}) \, / \, M_{rck} \, . \tag{4.35}$$

Compared to the variance model in equation (4.31), this formulation would appear to underrepresent the within-PSU variation.

Variance Approximations

In addition to estimated totals of the numbers of 17-year-olds $\hat{M}_{+++}$ and the numbers of correct responses $\hat{Y}_{+++}$, four variance estimators for $Var(\hat{P} = \hat{Y}_{+++} / \hat{M}_{+++})$ were computed from each sample. The first of these variance estimators (V1) uses the first two terms of the bias-adjusted Yates-Grundy type estimator in equation (3.29) with a jackknife pseudo value

$$p_{\ell k} = 31 \, \hat{P} - 30 \left\{ \frac{\hat{Y} - y_{\ell k}}{\hat{M} - m_{\ell k}} \right\} \tag{4.36}$$

taking the place of $31 \, y_{\ell k}$, the corresponding pseudo value for $\hat{Y}_{+++}$. The form of the jackknife linearization, which uses

$$p_{\ell k} = 2 \, \hat{P} - \left\{ \frac{\hat{Y} - (y_{\ell k} - y_{\ell' k'})}{\hat{M} - (m_{\ell k} - m_{\ell' k'})} \right\} \tag{4.37a}$$

and

$$P_{\ell' k'} = 2\hat{P} - \left\{ \frac{\hat{Y} + (y_{\ell k} - y_{\ell' k'})}{\hat{M} + (m_{\ell k} - m_{\ell' k'})} \right\} \qquad (4.37b)$$

might seem more appropriate for variance estimators involving squared differences. Recall that this was the form used with the Year 01 Design, where there were two PSUs selected per primary stratum. To simplify our calculations, we decided to use the linearization in equation (4.36) for all four of our variance estimators. While we intend to present a comparison of alternative jackknife and Taylor-Series linearizations in a subsequent section, at this point we felt that the form of our estimation equations was more crucial to the comparison than the type of linearization.

The second variance estimator (V2) studied in the simulation would be appropriate if the PSU allocation to cells was fixed and PSUs were selected with replacement. A circular successive differencing scheme was used within major strata to collapse cells where only one PSU was selected. For this purpose, the States in table 4-1 were arranged in the following circular array

$$\begin{tikzcd}[column sep=small] \text{Alas.} \arrow[r] & \text{Wash.} \arrow[r] & \text{Oreg.} \arrow[r] & \text{Calif.} \arrow[r] & \text{Nev.} \arrow[r] & \text{Ariz.} \arrow[r] & \text{N.M.} \arrow[r] & \text{Texas} \arrow[d] \\ \text{Hawaii} \arrow[u] & \text{Mont.} \arrow[l] & \text{Idaho} \arrow[l] & \text{Wyo.} \arrow[l] & \text{Utah} \arrow[l] & \text{Colo.} \arrow[l] & \text{Okla.} \arrow[l] & \end{tikzcd}$$

This particular ordering is based on geographical proximity and represents a crude attempt to make adjacent States in the array more alike than nonadjacent States. If $u=1(1)\ r_c(s)$ represents the ordered array of single PSU cells in a pattern(s) sample from column c and $\Sigma_{\ell}^{(2+)}$ denotes summation over control cells with $n_{\ell}(s) \geq 2$, we have

$$V2(s) = \sum_{\ell}^{(2+)} n_{\ell}(s) \sum_{k=1}^{n_{\ell}(s)} (p_{\ell k} - p_{\ell .})^2 \,/\, [n_{\ell}(s) - 1] \quad (31)^2$$

$$+ \sum_{c=1}^{7} \sum_{u=1}^{r_c(s)} (p_{uc} - p_{u+1,c})^2 \,/\, 2(31)^2 \qquad (4.38)$$

where

$$r_c(s) + 1 \equiv 1 .$$

Notice that the second term in equation (4.38) includes all cells with $n_\ell(s) = 1$ whereas the singleton cells in the corresponding term of V1(s) (the Yates-Grundy estimator)' must have $N_\ell \geq 2$.

Our third estimator (V3) ignores any control in the State dimension and computes the variance as if PSUs had been selected with replacement from the seven major strata. Suppose $w = 1(1)n_c(s)$ indexes all the PSUs in major stratum c. Then

$$V3(s) = \sum_{c=1}^{7} n_c(s) \sum_{w=1}^{n_c(s)} (p_{wc} - p_{.c})^2 / \{n_c(s) - 1\} (31)^2. \qquad (4.39)$$

The final estimator (V4) ignores all controls, using the formula that would be appropriate if the 31 PSUs were an unstratified selection with replacement from the entire list of 221 first-stage listing units. This estimator is computed as follows:

$$V4 = \sum_{c=1}^{7} \sum_{w=1}^{n_c(s)} (p_{wc} - p_{..})^2 / 31 \times 30 . \qquad (4.40)$$

Empirical Results

Table 4-2 presents the sampling expectations, EP, and variances, VP, for nine estimated P-values when the Montreal model is applied. The quantities (EY/EM) representing the ratio of numerator and denominator sampling expectations show that there is very little bias due to ratio estimation of the P-values. Unfortunately, this is not the case with the four variance estimators. The bias-adjusted Yates-Grundy estimator tends to underestimate VP, whereas the other three approximations tend seriously to overestimate VP. The magnitude of the various estimators tends to increase from V1 through V4 as one might expect; V1 makes an attempt to account for the between-PSU variability properly while underestimating the within-PSU variability. The other three estimators, while accounting

Table 4-2. Bias comparisons for four variance estimators (Montreal model)

P=(EY|EM)	EP	VP	EV1	EV2	EV3	EV4
10.16	10.17	3.63	2.97	5.43	5.50	5.92
38.12	38.09	4.17	2.70	6.14	6.62	7.63
40.39	40.45	4.24	2.19	6.41	6.85	7.73
62.42	62.43	5.34	4.29	8.09	8.75	9.32
69.11	69.14	4.26	3.95	7.26	7.32	7.95
72.92	72.85	3.77	4.21	6.66	6.49	7.21
77.52	77.46	3.54	3.55	5.20	5.42	6.08
91.73	91.73	3.05	2.63	2.14	2.32	2.64
93.55	93.53	3.19	3.10	2.36	2.56	3.08

for the within-PSU variation properly, tend to overestimate the between-PSU variation by ignoring any finite population corrections and overlooking various levels of control beyond stratification.

Table 4-3 shows that in terms of least total error, root mean squared error, V1 is superior to V2, which has a slight edge over V3. The unstratified estimator V4 performs poorly for all except the two largest P-values. The average performance of the four estimators over all nine P-values is summarized in table 4-4. In terms of absolute relative bias, the Yates-Grundy type estimator looks better than the other three. When one looks at relative total error, or root mean squared error divided by VP, the advantage for V1 is not as great. The stability figures in table 4-4 represent averages of estimated degrees of freedom where, for the i-th variance estimator,

$$\hat{d}f(i) \equiv 2[EV(i)]^2/Var[\hat{V}(i)] \quad \text{for } i = 1, \ldots, 4. \tag{4.41}$$

These stability measures relate directly to the shape of the T-like sampling distributions summarized in table 4-5.

The averaged frequency distributions presented in table 4-5 show the proportion of times that

Table 4-3. Root mean squared error of variance estimators (Montreal model)

EP	VP	$\sqrt{MSV1}$	$\sqrt{MSV2}$	$\sqrt{MSV3}$	$\sqrt{MSV4}$
10.17	3.63	2.95	4.70	4.65	4.86
38.09	4.17	3.02	4.58	4.69	5.36
40.45	4.24	2.89	4.55	4.80	5.65
62.43	5.34	3.65	6.48	6.65	7.12
69.14	4.26	2.56	5.41	5.37	5.66
72.85	3.77	9.35	5.98	5.61	5.95
77.46	3.54	2.66	3.57	3.55	3.95
91.73	3.05	1.64	2.18	2.14	2.03
93.53	3.19	1.58	1.96	1.79	1.68

$$\hat{T}(i) = (\hat{P} - EP) / \sqrt{F(i)\ \hat{V}(i)} \quad \text{for } i = 1,2,\ldots 4, \tag{4.42}$$

and Z-like statistics

$$\hat{Z} = (\hat{P} - EP) / \sqrt{VP}, \tag{4.43}$$

fall within the stated limits. Notice that the T-like statistics have been corrected for bias in $\hat{V}(i)$ by applying the factor $F(i) = VP/EV(i)$. Comparing $\hat{Z}$ to the normal frequencies in the last row of table 4-5, one notes that the symmetric intervals are reasonably close, especially for the larger, more critical intervals. Some positive skewness is observed in the asymmetric intervals. Of the T-like distributions, T2 appears closest to Student's T with 30 d.f. Recalling the stability measures in table 4-4, one can see that the distributions $\hat{T}(1)$ through $\hat{T}(4)$ become more like $\hat{Z}$ as the stability of $\hat{V}(i)$ increases. As estimates of approximate "degrees of freedom" these stability measures are gross underestimates. The naive approximation based on the number of PSUs selected is on the other hand reasonably accurate. Before we conclude that $\hat{V}(2)$ is superior to the other estimators when it comes to making inference about $\hat{P}$, it is important to recall that the results in table 4-5 are corrected for bias. If these corrections had not been made, none of the estimators would yield anything resembling a T distribution.

Table 4-4. Average performance of variance estimators (Montreal model)

	V1	V2	V3	V4
Rel-Bias (%)	13	50	53	63
Stability $(\hat{d}f)$	3.52	4.47	4.64	5.51
Rel-Error (%)	72	101	101	108

Table 4-5. Sampling distributions for normal and T-like statistics

Proportion w/in	±2.576	±1.960	±1.645	±1.282	±1.036	(-1.960,0)	(0,1.960)	(-1.036,0)	(0,1.036)
Student's T(30df)	.9848	.9407	.8896	.7903	.6915	.4703	.4703	.3458	.3458
$\hat{T}1$	.9367	.8941	.8522	.7662	.6829	.4379	.4562	.3323	.3506
$\hat{T}2$	.9816	.9413	.8967	.8074	.7158	.4587	.4827	.3473	.3684
$\hat{T}3$	.9861	.9480	.9064	.8142	.7211	.4628	.4852	.3500	.3711
$\hat{T}4$	.9859	.9516	.9088	.8204	.7277	.4632	.4883	.3507	.3770
$\hat{Z}$	.9906	.9605	.9246	.8423	.7509	.4756	.4849	.3702	.3807
Norm. Deviate	.9900	.9500	.9000	.8000	.7500	.4750	.4750	.3750	.3750

In figure 4-1 we have presented a plot of true and estimated variances against their corresponding P-values. The curious relationship (or lack of any) between the true variances and P-values is symptomatic of the problems inherent in the Montreal model. We should expect a strong quadratic relationship between P and VP modeled after the simple random sampling case where VP = P (100 - P)/n. Tables 4-6 and 4-7 display results based on the new model where the within-PSU variation in the estimated number of 17-year-olds, say $\hat{M}_{rck}$ is set to M^2_{rck} (or M^t_{rck} with t=2). One thing we notice immediately about these tables is that V1, the bias-adjusted Yates-Grundy estimator, has been eliminated from consideration. The reasons we have excluded V1 from further consideration at this time are two. The principal reason is the excessive cost of computing V1 relative to the

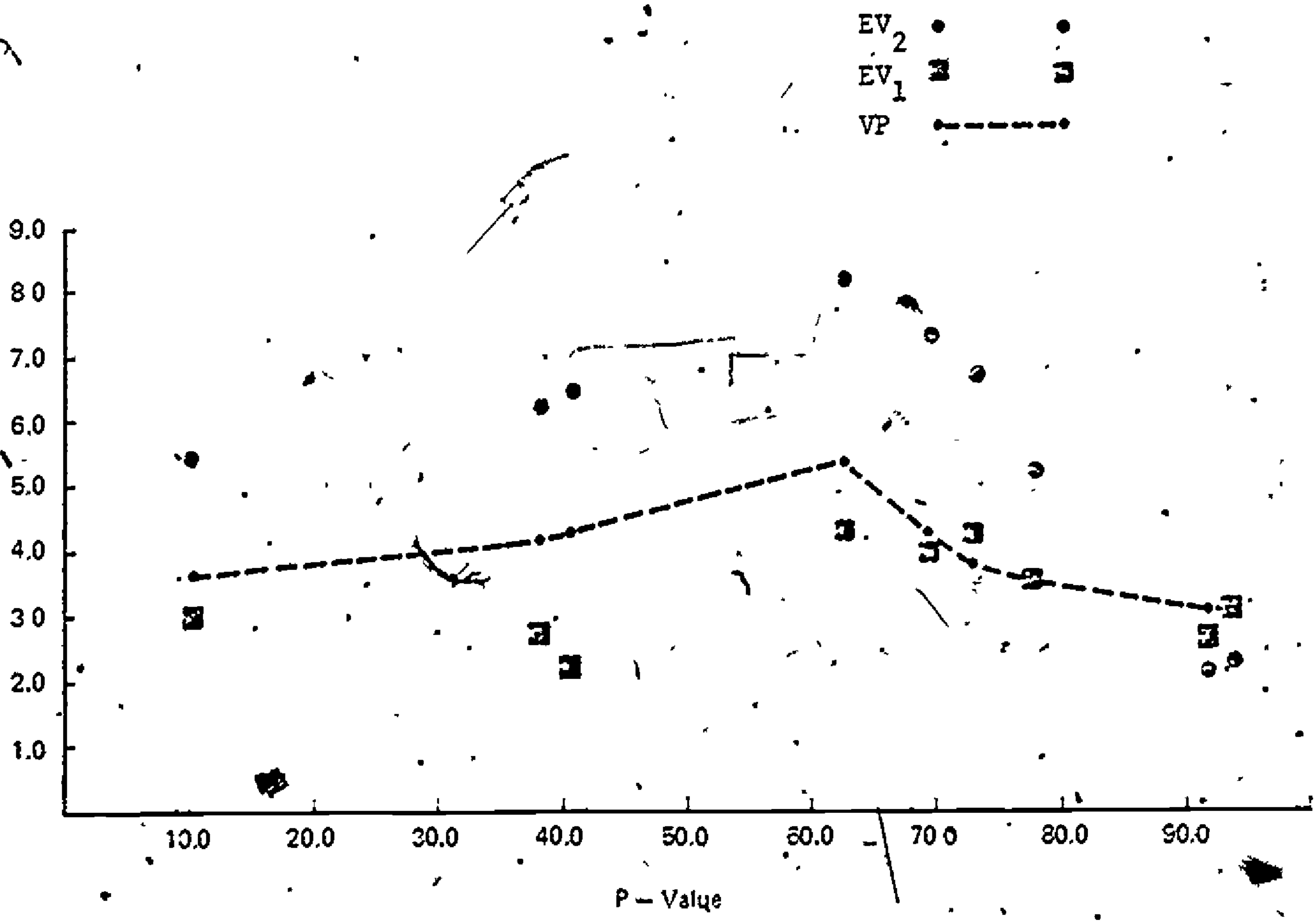

Figure 4-1. Montreal model (true variance vs. estimated variances).

Table 4-6. Bias comparisons for three variance estimators under the new model with t=2

EP	VP	EV2	EV3	EV4
10.15	3.21	5.01	4.99	5.46
38.34	8.29	10.73	10.97	12.12
40.41	8.72	10.84	11.18	12.01
62.51	9.30	11.56	12.25	12.81
69.06	8.08	10.23	10.27	10.95
72.94	7.04	9.38	9.15	9.82
77.53	6.22	7.36	7.53	8.20
91.71	3.73	2.81	2.95	3.27
93.47	3.17	2.29	2.51	3.03

Table 4-7. Design effects with N=744 for new model with t=2

EP	DEFT	D2	D3	D4
10.15	2.62	4.09	4.07	4.45
38.34	2.61	3.38	3.45	3.81
40.41	2.69	3.35	3.45	3.71
62.51	2.95	3.67	3.89	4.07
69.06	2.81	3.56	3.58	3.81
72.94	2.65	3.54	3.45	3.70
77.53	2.66	3.14	3.22	3.50
91.71	3.65	2.75	2.89	3.20
93.47	3.86	2.79	3.06	3.69

other estimators. This cost is on the average 16 times greater than the cost of V2, the cheapest estimator. Finally, although V1 has the smallest bias, the fact that it tends to underestimate the true variance is disconcerting. If an unbiased variance estimator is not available, one generally prefers an overestimate which results in conservative inferential statements. Although it may be hard to judge the dollar value of a 'good' variance estimator, excessive computing cost would seem to constitute a reasonable excuse for eliminating one of several mediocre estimates.

Figure 4-2 plots the true variance and the expected value of V2 (the least biased of the conservative estimates) against the corresponding P-value. The new model with t = 2 exhibits the desired quadratic relationship between VP and P. It is interesting to note that VP does not decline for large values of P as fast as one might expect. This is demonstrated

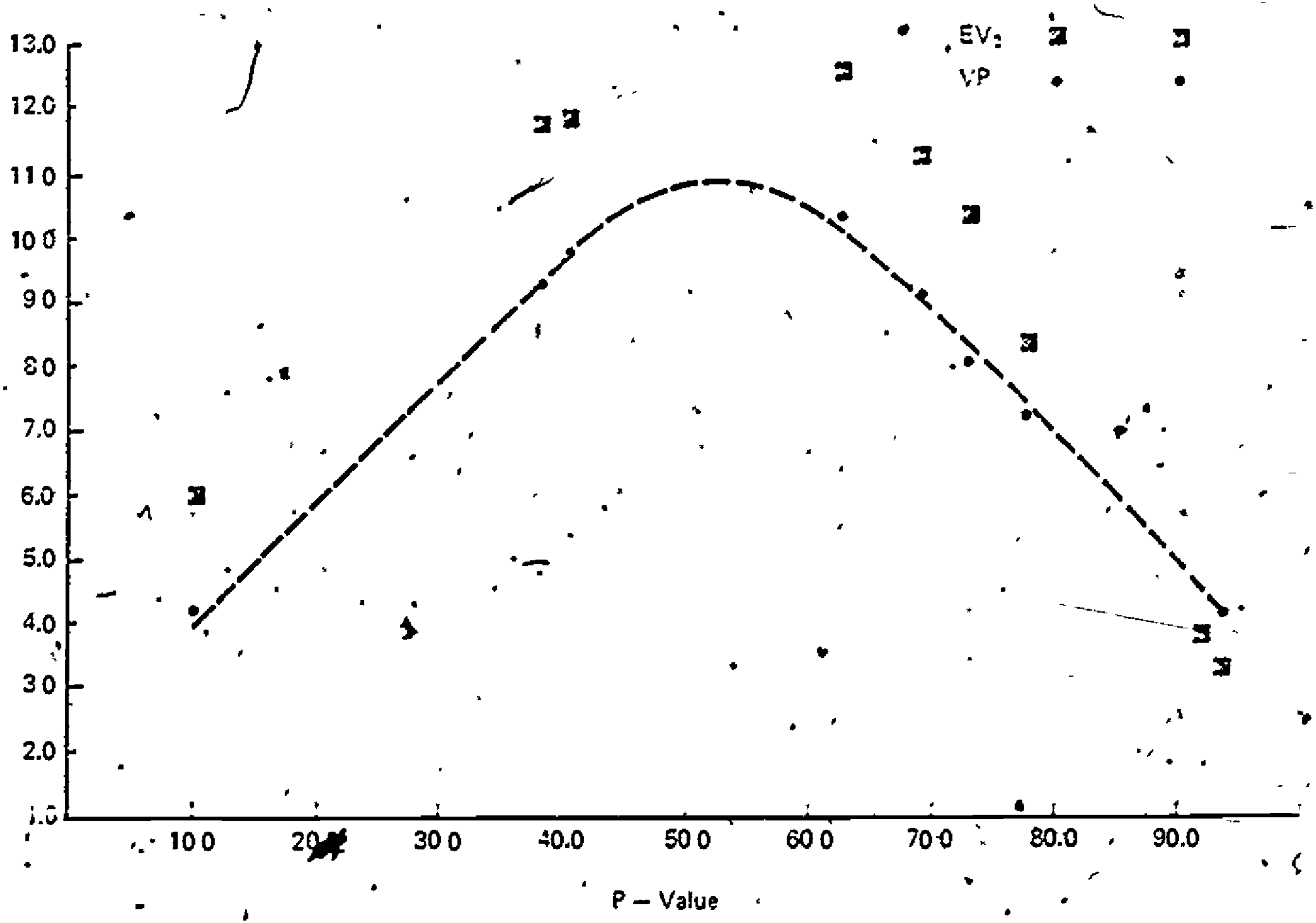

Figure 4-2. New model with t=2 (true variance vs. estimated variance).

clearly by the jump observed in design effects or DEFF values for the two P-values (in table 4-7) exceeding 90 percent. The three variance estimators behave more symmetrically dropping below VP for P > 90 percent. The median differences between those design effects estimated by V2 and the true design effects based on VP is (+) .66.

Figure 4-3 shows a plot of true variance VP and expected values of V2 (EV2) for 20 P-values, using the new model where the within-PSU variation in the estimated number of 17-year-olds $\hat{M}_{rck}$ is equal to M_{rck} (or M^{t}_{rck} with t=1). This model yields a surprisingly smooth parabolic relationship between P and VP. The slow asymmetric decline in VP for large values of P is again apparent with V2 dropping below VP near the point p = 87.5

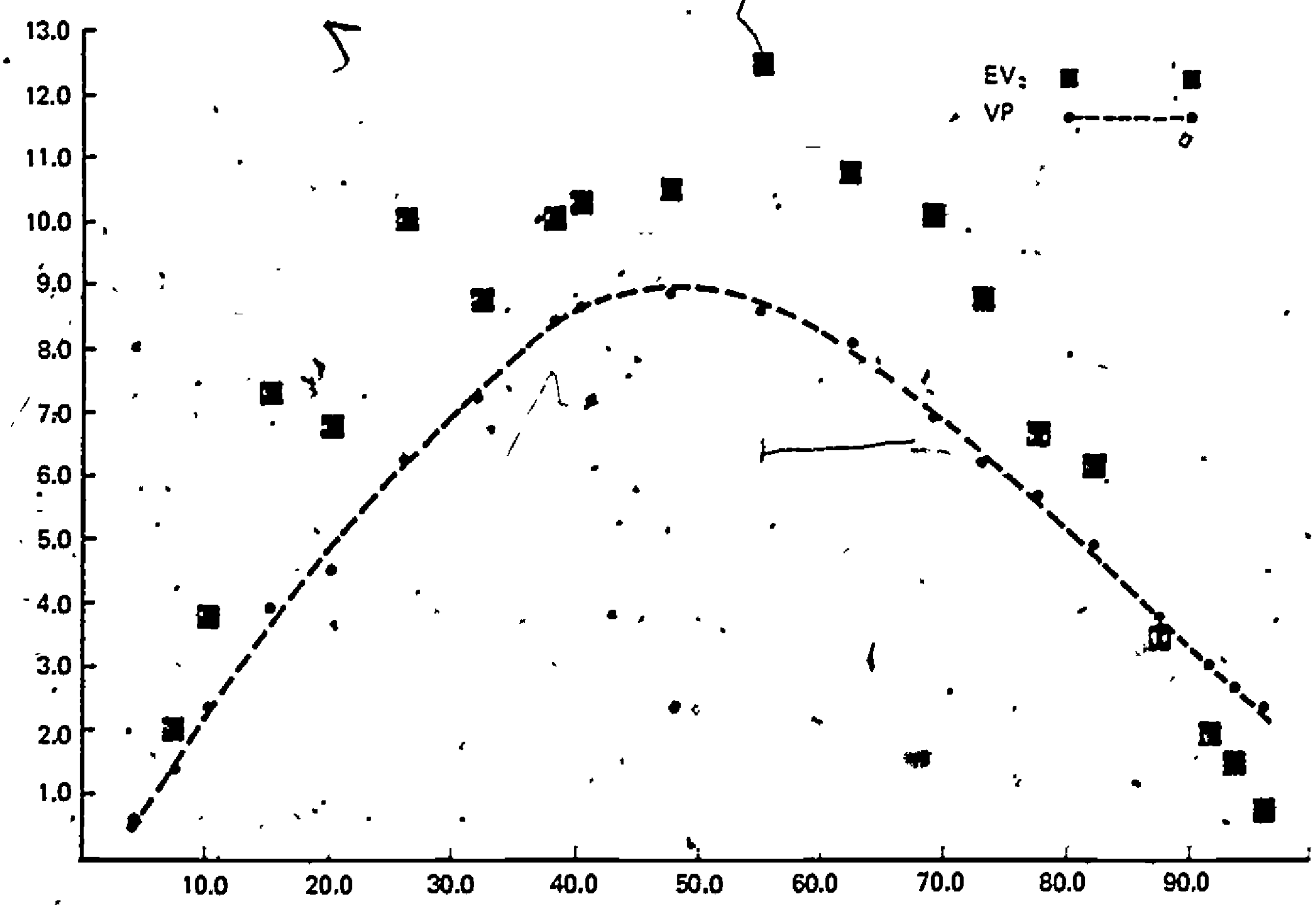

Figure 4-3. New model with t=1 (true variance vs. estimated variance).

percent. Table 4-8 shows the relationship between VP and the expected values of V-2, V-3, and V-4 for the 20 new model P-Values with t=1. Design effects D-2, D-3, and D-4, based on the three conservative variance estimators, are compared to the true design effects (DEFT) in table 4-9. The median differences between the estimated DEFF based on V2 and the true DEFF is, for this case, (+).64. Tables 4-10 and 4-11 show the average performance of our three conservative estimators for the new model with t=2, 1 respectively. Although V4 is consistently the most stable of the three estimators, V2 tends to have the smallest bias and smallest root mean square error as reflected in the average Rel-Bias and Rel-Error terms.

Table 4-8. Bias comparisons for the new model with t=1

EP	VP	EV2	EV3	EV4
4.25	.48	.56	.60	.63
7.50	1.41	2.01	3.22	3.62
10.18	2.36	3.79	3.82	4.31
15.00	3.92	7.29	7.37	8.46
20.00	4.51	6.76	7.03	7.36
26.00	6.23	10.02	10.30	10.82
32.00	7.20	8.76	8.95	9.81
38.18	8.40	10.00	10.37	11.65
40.32	8.61	10.28	10.64	11.47
47.50	8.80	10.47	10.83	14.06
55.00	8.54	12.40	12.94	13.54
62.43	8.04	10.70	11.43	12.05
69.04	6.89	10.05	10.02	10.72
72.95	6.20	8.75	8.51	9.24
77.51	5.67	6.66	6.88	7.58
82.00	4.87	6.10	6.78	7.51
87.47	3.77	3.60	4.09	4.67
91.69	3.06	1.99	2.11	2.43
93.53	2.70	1.52	1.75	2.25
96.00	2.33	.73	.79	1.18

Table 4-9. Design effects for the new model with t=1 (N=744)

EP	DEFT	D2	D3	D4
4.25	.88	1.02	1.10	1.15
7.50	1.51	2.16	3.45	3.88
10.18	1.92	3.08	3.11	3.51
15.00	2.29	4.25	4.30	4.94
20.00	2.10	3.14	3.27	3.42
26.00	2.41	3.87	3.98	4.18
32.00	2.46	3.00	3.06	3.35
38.18	2.65	3.15	3.27	3.67
40.32	2.66	3.18	3.29	3.55
47.50	2.63	3.12	3.23	4.19
55.00	2.57	3.73	3.89	4.07
62.43	2.55	3.39	3.63	3.82
69.04	2.40	3.50	3.49	3.73
72.95	2.34	3.30	3.21	3.48
77.51	2.42	2.84	2.94	3.24
82.00	2.45	3.07	3.42	3.79
87.47	2.56	2.44	2.78	3.17
91.69	2.99	1.94	2.06	2.37
93.53	3.32	1.87	2.15	2.77
96.00	4.51	1.41	1.53	2.29

Table 4-10. Average performance of variance estimates under the new model with t=2

	V2	V3	V4
Rel-Bias (%)	29.41	29.72	35.02
Rel-Error (%)	67.64	66.04	70.59
Stability	8.40	9.68	11.21

Table 4-11. Average performance of variance estimates under the new model with t=1

	V2	V3	V4
Rel-Bias (%)	37.69	44.70	55.08
Rel-Error (%)	68.05	71.28	79.33
Stability	10.46	12.75	15.32

The histograms in tables 4-12 and 4-13 are averaged over nine data sets for the first two models (Montreal and new t=2). The third histogram in both tables is based on 20 data sets. The normal or Gaussian-type distributions agree rather well, at least over the larger symmetric intervals with the standard normal. Curiously enough, the T-like distributions based on V2 have slightly fatter tails for the new model runs, while at the same time the stability measures for V2 increase from 4.47 for the Montreal model to 8.40 and 10.46 for the new model. These stability measures are still considerable underestimates when viewed as approximations for the degrees of freedom associated with the corresponding T-like distributions in table 4-13.

Comparison of Taylor Series and Jackknife Linearizations

In chapter 2 the Taylor series (TS) variance approximation formula for a NAEP P-value equation (2.20) was presented. It was shown that for a deeply stratified sample with two primary selections per stratum, the TS variance estimator was smaller than the corresponding jackknife (JK) variance approximation. To compare the TS linearization to our jackknifed variance estimators in the context of the simulation study presented in the previous sections, a Taylorized deviation

$$z_{\ell k} = 31(y_{\ell k} - \hat{P}\, n_{\ell k})/\hat{N} \qquad (4.44a)$$

Table 4-12. Sampling distributions for Gausian type statistics

Proportion w/in	±2.576	±1.960	±1.645	±1.282	±1.036	(-1.960,0)	(0,1.960)	(-1.036,0)	(0,1.036)
Gausian	.9900	.9500	.9000	.8000	.7500	.4750	.4750	.3750	.3750
Montreal Model	.9906	.9605	.9246	.8423	.7509	.4756	.4849	.3702	.3807
New Model: t=2	.9916	.9572	.9173	.8307	.7406	.4672	.4900	.3564	.3841
New Model: t=1	.9915	.9612	.9188	.8326	.7386	.4774	.4838	.3662	.3725

Table 4-13. Sampling distributions for students T-like statistics using V_2

Proportion w/in	±2.576	±1.960	±1.645	±1.282	±1.036	(-1.960,0)	(0,1.960)	(-1.036,0)	(0,1.036)
Student's T(30df)	.9848	.9407	.8896	.7903	.6915	.4703	.4703	.3458	.3458
Montreal Model	.9816	.9413	.8967	.8074	.7158	.4587	.4827	.3473	.3684
New Model: t=2	.9792	.9358	.8878	.8024	.7100	.4583	.4774	.3510	.3590
New Model: t=1	.9787	.9398	.8958	.8054	.7146	.4644	.4754	.3563	.3584

was used in place of the jackknife pseudo value

$$p_{\ell k} = 31\,\hat{P} - 30\left\{\frac{\hat{Y} - y_{\ell k}}{\hat{M} - m_{\ell k}}\right\} \qquad (4.44b)$$

when forming the estimators V2, V3 and V4 as specified in equations 4.38, 4.39, and 4.40 respectively.

Table 4-14 shows the results from an independent set of 1,000 samples generated according to the previously described simulation model with t=1. Contrasting the first two columns of table 4-14 with the corresponding columns of table 4-8 shows good agreement between the two independent (1,000 samples) replicates. The maximum standard deviation for the sampling expectation of $\hat{P}$ is .05 percent with an average across the 20 exercises of .01 percent. The simulation estimates of VP, the sampling variance of $\hat{P}$, exhibit a maximum standard deviation of .12 percent and an average deviation across exercises of .04 pertent. While it would have been desirable to make the Taylor series versus jackknife comparison on the same set of 1,000 samples, our software design was such that it was more economical to make independent runs than to incorporate both calculations in the same run.

Table 4-15 presents the average performance of the three Taylor series variance estimators. Compared with the jackknife results in table 4-11, we see a reduction of 3.3 to 4.5 percent in relative bias. The corresponding reductions in relative total error range from 4.2 percent to 8.3 percent. The stability measures for the three Taylor series estimators show a general increase of one unit over the corresponding jackknife estimators. This indication of a slight increase in stability for the TS estimators does not show up in the T-like distributions presented in table 4-16. While the tails of the TS distributions are more symmetrical than their JK counterparts, they are also fatter. The percentage of statistics outside the ± 1.960 interval is about 1 percent greater for the TS statistics.

While some consideration was given to developing a bias-correction factor for the V2 estimator based on these simulation results, it was felt that this would be necessary only if actual Year 02 sampling errors based on V2 were considerably larger than Year 01 sampling errors. While

Table 4-14. Bias comparisons for the new model with t=1 and a Taylor-series linearization

EP	VP	ET2	ET3	ET4
4.25	0.48	0.56	0.60	0.64
7.50	1.43	1.97	3.11	3.49
10.14	2.34	3.69	3.75	4.22
15.00	3.90	7.11	7.16	8.25
20.00	4.57	6.57	6.78	7.13
26.00	6.24	9.51	9.87	10.46
32.00	7.36	8.30	8.57	9.39
38.17	8.44	9.68	10.05	11.26
40.30	8.73	10.07	10.38	11.26
47.50	8.92	10.05	10.52	13.66
55.00	8.56	12.24	12.72	13.26
62.41	8.04	10.57	11.32	11.85
69.01	6.85	9.75	9.68	10.41
72.88	6.36	8.51	8.28	9.01
77.50	5.50	6.60	6.72	7.43
82.00	4.79	5.89	6.60	7.33
87.50	3.81	3.51	3.98	4.51
91.68	3.10	1.95	2.08	2.42
93.50	2.74	1.53	1.76	2.26
96.00	2.31	0.69	0.74	1.16

Table 4-15. Average performance of Taylor-series estimates under the new model with t=1

	T2	T3	T4
Rel-Bias (%)	34.42	40.66	50.58
Rel-Error (%)	63.44	66.48	71.01
Stability	11.23	13.60	16.31

Table 4-16. Sampling distributions for T-like statistics

Proportion w/in	± 2.576	±1.960	±1.645	±1.282	±1.036	(-1.960,0)	(0,1.960)	(-1.036,0)	(0,1.036)
Student's T(30df)	.9848	.9407	.8896	.7903	.6915	.4703	.4703	.3458	.3452
Taylor-Series V2	.9766	.9282	.8732	.7738	.6750	.4620	.4662	.3351	.3399
Jackknife V2	.9787	.9398	.8958	.8054	.7146	.4644	.4754	.3563	.3584
Taylor-Series V3	.9788	.9307	.8770	.7772	.6788	.4626	.4681	.3366	.3422
Jackknife V3	.9802	.9418	.8979	.8090	.7192	.4650	.4769	.3590	.3602
Taylor-Series V4	.9802	.9337	.8818	.7824	.6857	.4652	.4684	.3403	.3454
Jackknife V4	.9786	.9418	.9004	.8150	.7287	.4632	.4786	.3613	.3624

the simulation indicates strongly that such variance approximations considerably overestimate the variability of 'controlled selections,' the estimated level of precision could still be adequate for National Assessment reporting purposes. In order to resolve this issue, sampling errors were calculated for 131 Year 02 national P-values using a version of V2 with the squared difference jackknife linearization employed in Year 01 and reintroduced in equation (4.37). Letting h = 1(1)4 index NAEP's four regional strata, this variance estimator took the form

$$\mathrm{var}\,\{P_{02}\} = \sum_{h=1}^{4} \sum_{\ell}^{(2+)} \sum_{k=1}^{n_{h\ell}-1} \sum_{k'=k+1}^{n_{h\ell}} \Delta^2 p_{h\ell}(k,k')/4(n_{h\ell}-1)$$

$$+ \sum_{h=1}^{4} \sum_{c=1}^{7} \sum_{u=1}^{r_{hc}} \Delta^2 p_{hc}(u,u+1)/8 \, , \qquad (4.45a)$$

with

$$r_{hc}+1 \equiv 1$$

and

$$\Delta^2 p_{h\ell}(k,k') \equiv \left\{ \left[\frac{\hat{Y} - \Delta y_{h\ell}(k,k')}{\hat{M} - \Delta m_{h\ell}(k,k')} \right] - \left[\frac{\hat{Y} + \Delta y_{h\ell}(k,k')}{\hat{M} + \Delta m_{h\ell}(k,k')} \right] \right\}^2 \qquad (4.45b)$$

where

$$\Delta y_{h\ell}(k,k') \equiv (y_{h\ell k} - y_{h\ell k'})$$

$$\Delta m_{h\ell}(k,k') \equiv (m_{h\ell k} - m_{h\ell k'}) \, .$$

As in equation 4.37a and b, $\Sigma_{\ell}^{(2+)}$ denotes summation over State by major stratum cells-ℓ with data from two or more PSUs ($n_{h\ell} \geq 2$). The second term in equation 4.45 involves successive squared differences among the r_{hc} single PSU cells in major stratum-c of region-h. Recall that the indexing $u = 1(1)r_{hc}$ of single PSU cells follows a particular ordering of the States in region-h based on geographic proximity.

The 131 resulting variance estimates were summarized in terms of design effects. This analysis paralleled wherever possible the breakdowns presented in the Chromy et al. summary of Year 01 effects. The following chapter details this comparison of Year 01 and Year 02 sampling errors.

Chapter 5: COMPARISON OF YEAR 01 AND YEAR 02 SAMPLING ERRORS

Introduction

The computer simulation study of National Assessment's Year 02 sample described in the preceding chapter suggests that the standard variance approximations recommended for a controlled selection of primary units seriously underestimate the precision of NAEP P-values. To assess the overall impact of Year 02 sample design changes coupled with significant positive bias in the associated sampling error estimates, 60 9-year-old and 71 13-year-old reading exercises from the Year 02 Assessment were examined. Estimated sample design effects (DEFFs) for the 131 national P-values were calculated using the variance estimator described in equation 4.45. Similar sets of 131 DEFFs were computed for NAEPs four regions, two sex categories, and four SOC subpopulations.

Stem and leaf displays of these design effect distributions were formed to facilitate the calculation of median effects and other sample percentiles [ref. 1]. Table 5-1 illustrates the display of national design effects for the two age classes represented. The left most column indicates the first two significant digits of our national design effects. Associated third digits are aggregated in adjoining rows. For example, the aggregate to the right of 1.2 represents three, nine-year-old DEFFs taking values 1.26, 1.26, and 1.22. The third column is a running count from the low and high ends of the distribution toward the center. This tally facilitates location of the median DEFF and the two quartiles. These are grouped below each display along with the starred extreme values. In addition to these summary percentiles, the display provides an accurate view of the shape of our design effect distributions.

Comparative Analysis

The 131 Year 02 design effects summarized in table 5-1 range from .94 to 3.93 with a median value of 2.00. This compares to a range of .90 to 10.88 and a median of 2.38 for the 149 Year 01 exercises examined by Chromy et al. Table 5-2 compares the distribution of Year 01 and Year 02 national DEFFs.

Table 5-1. Stem and leaf display of Year 02 national design effects

/ Nines			Thirteens		Total	
.5						
.6						
.7						
.8						
.9			54	2	45	2
1.0			635	5	356	5
1.1	9	1	3	6	39	7
1.2	662	4	4592	10	2245669	14
1.3						
1.4	66793	9	50	12	0356679	21
1.5	27298	14	6	13	226789	27
1.6	0138	18	99	15	013899	33
1.7	1651	22	5629	19	11255669	41
1.8	8	23	0518214	26	01124588	49
1.9	067778	29	2738118	33	0112367777888	62
2.0	04	-2-	907930052	-9-	00002345799	-11-
2.1	15951	29	37	299	1135579	58
2.2	021	24	53	27	01235	51
2.3	73	21	470804	25	00344778	46
2.4	7111	19	564	19	114567	38
2.5	12	16	503	16	01235	32
2.6	3472	14	45	13	234457	27
2.7	019	10			019	21
2.8	5	7	4390	11	03459	18
2.9	4	6	68453	7	344568	13
3.0	4	5			4	7
3.1						
3.2	729	4	8	2	2789	6
3.3						
3.4	5	1			5	2
3.5						
3.6						
3.7						
3.8						
3.9			3	1	3	1
4.0						
	* 1.19		* 0.94		* 0.94	
	Q11.60		Q11.76		Q11.69	
	M 2.02		M 2.00		M 2.00	
	Q32.52		Q32.45		Q32.47	
	* 3.45		* 3.93		* 3.93	

Table 5-2. Distributions of Year 01 and 02 national DEFFs

	Year 01		Year 02	
Design Effect	Number	Percent	Number	Percent
< 1.00	1	1	2	2
1.00 - 1.50	16	11	19	14
1.51 - 2.00	29	19	45	34
2.01 - 2.50	43	30	34	26
2.51 - 3.00	32	21	24	18
3.01 - 3.50	8	5	6	5
3.51 - 4.00	10	7	1	1
4.01 - 4.50	5	3	-	-
4.51 - 5.00	3	2	-	-
> 5.00	2	1	-	-
Total	149	100%	131	100%

The Year 01 distribution includes 37 individually administered in-school exercises and 16 out-of-school young adult exercises, while the Year 02 distribution includes only in-school group administered exercises. While this lack of diversity in the mode of administration may explain some of the increased stability shown by the Year 02 DEFFs, we suspect that reduced variability in the Year 02 weight distributions had a more pronounced effect. It is also expected that the inclusion of individually administered exercises in the Year 02 distribution would enrich the lower end of the distribution and shift the median further below the Year 01 value.

While there appeared to be a tendency for Year 01 in-school DEFFs to decline as the age of respondents increased, no consistent trend was observed in the Year 02 data; at least, not between the ages of nine and thirteen. The regional trend observed in the Year 01 data is also obscured in the Year 02 tabulations. Table 5-3 shows that while the supremacy of

Table 5-3. Comparison of regional DEFFs

Year	Age	Mode of Administration	Subject Area	Number of Exercises	Median DEFTs by Region			
					NE	SE	C	W
01	9	Group	Writing	24	1.89	2.93	2.32	2.65
	13	Group	Writing	5	3.05	3.65	3.50	2.65
02	9	Group	Reading	60	1.53	2.52	1.73	1.71
	13	Group	Reading	71	2.04	1.66	1.77	1.75

southeastern DEFFs is maintained at age nine, at age thirteen the trend is reversed with southeast low and northeast high. As in Year 01, there was little difference between the sexes with males registering a median DEFF of 1.72 and females 1.66.

Chromy et al. also reported a possible tendency for big city and urban fringe areas to yield smaller DEFFs than the more sparsely populated medium city and small place subpopulations. This tendency is not apparent in the Year 02 DEFFs displayed in table 5-4.

In summary, one can state that Year 02 design effects are somewhat smaller and less variable than the Year 01 effects. While it can be said that the Year 02 effects varied by region and SOC, there were no consistent trends. These factors interacted in curious ways with the age class effects.

Conclusions

Comparing the indicated level of precision for Year 01 and Year 02 NAEP exercises, it was apparent that in spite of the suspected positive bias in Year 02 sampling errors, the overall level of precision was improved somewhat in Year 02. In light of this result, no bias correction was attempted for the Year 02 variances. The sampling error approximation which uses squared differences of jackknife pseudo-values within control cells and squared successive differences between single PSU cells within major size of community by SES substrata (V2) was judged to be the least biased of the computationally feasible estimators. The jackknife linearization was

Table 5-4. Year 02 DEFFs by size of community

				Median DEFTs by SOC			
Age	Mode of Administration	Subject Area	Number of Exercises	B.C.	U.F.	M.C.	S.P.
9	Group	Reading	60	2.38	1.64	2.29	2.04
13	Group	Reading	71	2.05	1.66	1.51	2.05
Total	Group	Reading	131	2.16	1.65	1.73	2.05

retained since the improvement demonstrated for the Taylor series estimators did not justify the added cost of redesigning NAEPs sampling error software.

National Assessments in-school and out-of-school designs for assessment Years 02 through 04 remained basically the same with controlled selection used at the primary stage to allocate PSUs to State by major stratum control cells within regions. The sampling error methodology developed for the Year 02 sample has been applied directly to calculate Year 03 and Year 04 sampling errors.

In view of the difficulties associated with producing reasonably unbiased sampling error estimates for controlled selections, a major redesign of NAEPs primary sample was initiated for the Year 05 assessment. The Year 05 (1973-74) primary sample included the 15 largest SMSAs in 1970 as self-representing PSUs. Sampled PSUs were stratified by region, State and size of community. NAEP's requirement that all States be represented in the sample was met by carving out a stratum within each State which was not already covered by a self-representing SMSA. These State strata were assigned two primary selections wherever size permitted. Some single PSU strata were carved out in small States. Listing units from States already covered by self-representing SMSAs and those not contained within the "carved out" State substrata were placed in a regional pool. The regional pool was stratified along size of community lines with two or three selections per stratum.

Sampled PSUs were selected with probabilities proportional to their 14-year-old population in 1970 with some adjustment to effect an oversampling of PSUs containing low income inner city areas or highly rural counties. Sampford's rejective PPS without replacement selection method was used. Aside from the 15 self-representing SMSAs, NAEP's Year 05 primary sample can be described as a deeply stratified PPS selection of two or three primary units per strata. The few single PSU strata carved out of small States were paired within regions for variance estimation. In order to account for the within-PSU variability of the self-representers, replicated school samples were drawn. With the planned collapsing of single PSU strata and the replicated school samples within self-representers, a variance approximation based on squared differences between expanded-up PSU contributions (or replicate contributions) within strata should be reasonably unbiased. Some overestimation could be expected due to ignoring the effect of without-replacement selection of PSUs and replicates. A detailed description of the Year 05 NAEP samples can be found in RTI's final report for assessment Year 05 [ref. 2].

The Year 06 (1974-75) NAEP in-school primary sample was essentially an independent replicate of the Year 05 sample selected from the deeply stratified primary unit frame developed for the 1973-74 survey. Variance estimates for Year 06 statistics were again obtained from squared differences of PSU (or replicate) level jackknife pseudo-values summed over primary strata. For the Year 07 NAEP sample, a decision was made to draw four non-overlapping samples to be used successively for Years 07 through 10. This was accomplished by adapting the deeply stratified Year 05 design strategy to select enough PSUs and replicated school samples within the self-representing primary units to serve for four years. The combined sample was then partitioned at random into four equal sized yearly samples. To preserve valid PSU replication, primary strata in the master sample were combined and the associated primarys were randomly partitioned into four sets each containing two or occasionally three units. By relaxing the all-state requirement to assure complete state coverage over the four year sample, it was possible to assure that no school would be visited more than once during the Year 07 through Year 10 assessments.

The variance estimation methodology adapted for the Year 05 and 06 samples was modified to sum squared differences between PSU (or replicate) psuedo-values and the corresponding primary stratum mean. With more primary strata containing three units, this modification was made to bring NAEP's variance estimation in line with the general jackknife approximation recommended in chapter 2, equation 2.12. A detailed description of the nonoverlapping Year 07 through Year 10 NAEP samples can be found in RTI's final report for assessment Year 07 [ref 3].

REFERENCES

1. Tukey, J. W. (1970). Exploratory Data Analysis: Volume I. Limited Preliminary Edition. New York: Addison-Wesley.

2. Chromy, J. R. et. al. (1975). Final Report on National Assessment of Educational Progress Sampling and Weighting Activities for Assessment Year 05. Prepared for NAEP by RTI's Sampling Research and Design Center.

3. Benrud, C. H. et. al. (1977). Final Report on National Assessment of Educational Progress Sampling and Weighting Activities for Assessment Year 07. Prepared for NAEP by RTI's Sampling Research and Design Center.

Appendix A: TAYLOR SERIES LINEARIZATION FOR REGRESSION COEFFICIENTS

Population Definitions

Consider a finite universe U with N units U(i). Associate with the i-th unit a variate value Y(i) and a row vector of (p-1) regressors

$$X(i) = <X_o(i)\ X_1(i)\ \ldots\ X_p(i)>. \qquad (A.1)$$

The linear prediction equation for Y(i) of the form

$$y(i) = X(i)\beta \qquad (A.2)$$

which minimizes the sum of squared deviations

$$Q^2 = \sum_{i\in U} \{Y(i) - X(i)\beta]^2 \qquad (A.3)$$

is the familiar least-squares regression equation where β is a solution to the so-called 'normal equations'

$$(X^TX)\beta = X^TY \qquad (A.4)$$

where

$$X^TX \equiv \sum_{i\in U} X^T(i)X(i)$$

and

$$X^TY \equiv \sum_{i\in U} X^T(i)\,Y(i).$$

If X^TX has rank p-1, there is a unique solution for β in (A.4); namely

$$\beta = (X^TX)^{-1}X^TY \qquad (A.5)$$

If the p-1 equations represented by the matrix equation (A.4); are not linearly independent, redundant equations can be replaced by independent linear restrictions on the βs. While the following development is limited to the full rank case, it is not difficult to extend the results directly to a particular restricted solution.

Estimation

Suppose that a sample S of n units is selected from the universe U. Let π(i) denote the probability that unit U(i) will be included in such a sample. Unbiased Horvitz-Thompson estimators for X^TX and X^TY are

$$\widehat{(x^Tx)} = \sum_{k\in s} X^T(k)\,X(k)\,/\pi(k) \qquad (A.6a)$$

and

$$(x^Ty) = \sum_{k\varepsilon s} X^T(k)\ Y(k)\ /\pi(k). \qquad \text{(A.6b)}$$

Using these estimators, we solve for b in the estimated set of normal equations

$$(x^Tx)b = (x^Ty). \qquad \text{(A.7)}$$

As an estimator of β the population vector of regression coefficients $b = (x^Tx)^{-1}\ (x^Ty)$ can be viewed as a matrix version of the combined ratio estimator $\hat{R} = (\hat{X})^{-1}(\hat{Y})$. This analogy will be strengthened by the form of the Taylor series approximation for $(b-\beta)$.

Taylor Series Variance Approximation

To generate the first order Taylor series approximation for b, we begin by evaluating the partial derivatives

$$\partial b/\partial\ (x^Ty)_j \qquad \text{for } j = 0,\ 1,\ \ldots.\ p \qquad \text{(A.8a)}$$

and

$$\partial b/\partial\ (x^Tx)_{jj'} \qquad \begin{array}{l}\text{for } j = 0,\ 1,\ \ldots.\ .p \\ j' = j,\ j+1,\ \ldots,p\end{array} \qquad \text{(A.8b)}$$

where $(x^Ty)_j$ represents the j-th element in the (P+1) x 1 column vector (x^Ty) and $(x^Tx)_{jj'}$ is the (jj')-th element of the (p+1) x (p+1) symmetric matrix (x^Tx).

First, we notice that

$$\begin{aligned}\{\partial(x^Tx)b/\partial(x^Ty)_j\} &= (x^Tx)\ \{\partial b/\partial\ (x^Ty)_j\} \qquad \text{(A.9)}\\ &= \partial(x^Ty)/\partial(x^Ty)_j\end{aligned}$$

Therefore

$$(x^Tx)\ \{\partial b/\partial(x^Ty)_j\} = \delta_j \qquad \text{(A.10)}$$

where,

$$\delta_j(r) = \begin{cases} 1 & \text{if } j = r \\ 0 & \text{otherwise}\end{cases}$$

This allows us to write

$$\{\partial b/\partial(x^Ty)_j\} = (x^Tx)^{-1}\delta_j \qquad \text{(A.11)}$$

Notice that δ_j is a (p+1) x 1 column vector with a 1 in row j and zeros elsewhere. To evaluate the partial derivative of b with respect to the elements of (X^TX) we note that.

$$\{\partial\ (x^Tx)b/\partial\ (x^Tx)_{jj'}\} = \partial\ (x^Ty)/\partial\ (x^Tx)_{jj'} = \underset{\sim}{0} \qquad (A.12)$$

Hence

$$(x^Tx)\ \{\partial b/\partial(x^Tx)_{jj'}\} + \{\partial(x^Tx)/\partial(x^Tx)_{jj'}\}\ b = \underset{\sim}{0} \qquad (A.13).$$

Recalling that (X^TX) is symmetric, we see that

$$\{\partial(x^Tx)/\partial(x^Tx)_{jj'}\} = D_{jj'} \qquad (A.14).$$

where the (r,c)-th element of $D_{jj'}$ is

$$d_{jj'}(rc) = [1 - \delta(jj')]\ \delta_j(r)\delta_{j'}(c) + \delta_{j'}(r)\delta_j(c)$$

with $\delta(jj') = 1$ if $j = j'$ and zero otherwise. This leads to

$$\{\partial b/\partial\ (x^Tx)_{jj'}\} = -\ (x^Tx)^{-1}\ D_{jj'}b \qquad (A.15)$$

where $D_{jj'}$ is a (p+1) x (p+1) symmetric matrix having 1s in positions (jj') and (j'j) and zeros elsewhere.

Evaluating the partial derivatives in equations (A.11) and (A.15) at the point $<(x^Tx) = X^TX;\ (x^Ty) = X^TY>$ we can approximate b with the first order Taylor series linearization

$$b \doteq \beta + \sum_{j=0}^{p} [(x^Ty)_j - (X^TY_j)]\{\partial b/(x^Ty)_j\} \qquad (A.16)$$

$$+ \sum_{j=0}^{p}\ \sum_{j'=j}^{p} [(x^Tx)_{jj'} - (X^TX)_{jj'}]\ \{\partial b/\partial(x^Tx)_{jj'}\}$$

which becomes

$$b \doteq \beta + (X^TX)^{-1} \sum_{j=0}^{p} [(x^Ty)_j - (X^TY)_j]\ \delta_j$$

$$- (X^TX)^{-1} \sum_{j=0}^{p}\ \sum_{j'=j}^{p} [(x^Tx)_{jj'} - (X^TX)_{jj'}]D_{jj'}\beta\ . \qquad (A.17)$$

Recalling the definitions of δ_j and $D_{jj'}$, it is easy to see that

$$\sum_{j=0}^{p} [(x^Ty)_j - (X^TY)_j] \delta_j = [(x^Ty) - (X^TY)] \tag{A.18a}$$

and

$$\sum_{j=0}^{p} \sum_{j'=j}^{p} [(x^Tx)_{jj'} - (X^TX)_{jj'}] D_{jj'} = [(x^Tx) - (X^TX)]. \tag{A.18b}$$

This allows us to rewrite (A.17) as

$$b \doteq \beta + (X^TX)^{-1} \{[(x^Ty) - (X^TY)] - [(x^Tx) - (X^TX)]\beta\}. \tag{A.19}$$

Finally, observing that $(X^TX)\beta = (X^TY)$ we have

$$b \doteq \beta + (X^TX)^{-1} [(x^Ty) - (x^Tx)\beta]. \tag{A.20}$$

To exploit the result in (A.20) in order to approximate the sampling variance of b, we can define

$$z(k) = (X^TX)^{-1} [(x^Ty)_k - (x^Tx)_k\beta] \tag{A.21}$$

where $(x^Ty)_k = X^T(k)\ Y(k)$
and

$$(x^Tx)_k = X^T(k)\ X(k).$$

The corresponding Horvitz-Thompson estimator is

$$\begin{aligned}\hat{z} &= \sum_{k\epsilon s} z(k)/\pi(k)\\ &= (X^TX)^{-1} [(x^Ty) - (x^Tx)\beta].\end{aligned} \tag{A.22}$$

Combining (A.20) and (A.22) we see that

$$\hat{z} \doteq (b - \beta). \tag{A.23}$$

The result in (A.22) leads one to the following approximation for the generalized mean-squared-error of b, our vector of p + 1 estimated regression coefficients:

$$\begin{aligned}\text{GMSE } \{b\} &\doteq E_s \{(b-\beta)(b-\beta)^T\}\\ &\doteq E_s \{\hat{z}\, \hat{z}^T\}\\ &= \text{VAR } \{\hat{z}\}.\end{aligned}$$

since $E_s\{\hat{z}\} = 0$, the $(p+1) \times 1$ null vector.

If $\mathrm{VAR}_D(\hat{T})$ is the sampling variance formula for an estimated total $\hat{T}$ from a particular sampling design (D), then the generalized mean-squared-error for a vector of (p+1) regression coefficients b estimated from D is approximately

$$\mathrm{GMSE}\{b\} \doteq \mathrm{VAR}_D\{\hat{z}\} \qquad (A.25)$$

where $\hat{z}$ is the linearized statistic

$$\hat{z} = (X^TX)^{-1} \sum_{k\epsilon s} [(x^Ty)_k - (x^Tx)_k \beta] / \pi(k) .$$

In the following section, we will exploit equation (A.25) to produce the 'Taylor Series' variance estimator or more precisely, the generalized mean-squared-error estimator for b.

Taylor Series Variance Estimator

Recalling the Taylor series approximation for GMSE{b} developed in the previous section, it is clear that if $\mathrm{var}_D(\hat{T})$ represents an appropriate variance estimator for the sample total $\hat{T}$ from design D, then $\mathrm{var}_D(\hat{z})$ is the associated estimator for GMSE {b}. Since the linearized variate value $Z(k)$ in (A.21) is a function of the unknown population quantities $(X^TX)^{-1}$ and β, one is obliged to impose another level of approximation at this point. Instead of $Z(k)$, we use

$$z(k) = (x^Tx)^{-1} [(x^Ty)_k - (x^Tx)_k b] \qquad (A.26)$$

substituting our sample estimates for the unknown population parameters. It is interesting to note at this point that (A.26) is a matrix analogue of the 'Taylorized' variate used to approximate the variance of the ratio $\hat{R} = (\hat{Y}/\hat{X})$. Expressing this ratio as $\hat{R} = (\hat{X})^{-1}(\hat{Y})$ and making the associations $(\hat{X})^{-1} <=> (x^Tx)^{-1}$ and $(\hat{Y}) <=> (x^Ty)$, the relationship between (A.26) and the familiar

$$z_{\hat{R}}(k) = [Y(k) - \hat{R}\, X(k)] / \hat{X}$$

$$= (\hat{X})^{-1} [Y(k) - X(k)\, \hat{R}] \qquad (A.27)$$

is obvious.

To illustrate the method, we can consider a stratified simple random cluster sample with h=1(1)H strata, and n(h) clusters selected from N(h) without replacement. If y(hik) is the variate value associated with the k-th unit [k=1(1) M(hi)] in cluster-i of stratum-h and X(hik) is the corresponding 1 x (p+1) row vector of regressors, we use the expanded cluster totals

$$(x^T y)_{hi} = N(h) \sum_{h=1}^{M(hi)} X^T(hik)\ Y(hik)/n(h) \qquad (A.28a)$$

and

$$(x^T x)_{hi} = N(h) \sum_{k=1}^{M(hi)} X^T(hik)\ X(hik)/n(h) \qquad (A.28b)$$

to form

$$\tilde{z}(hi) = [\ (x^T y)_{hi} - (x^T x)_{hi}\ b\ ] \qquad (A.29)$$

where

$$b = (x^T x)^{-1}\ (x^T y)$$

and

$$(x^T x) = \sum_{h=1}^{H} \sum_{i=1}^{n(h)} (x^T x)_{hi}$$

$$(x^T y) = \sum_{h=1}^{H} \sum_{i=1}^{n(h)} (x^T y)_{hi} .$$

Then, we calculate

$$\text{gmse}\ \{b\} = (x^T x)^{-1} \{ \sum_{h=1}^{H} \{1 - f(h)\}\ n(h)\ s^2_{\tilde{z}}(h)\}\ (x^T x)^{-1} \qquad (A.30)$$

where f(h) = n(h)/N(h) and

$$s^2_{\tilde{z}(h)} = \sum_{i=1}^{n(h)} [\tilde{z}(hi) - \tilde{z}(h.)][\tilde{z}(hi) - \tilde{z}(h.)]^T/[n(h)-1]$$

with $\tilde{z}(h.) = \sum_{i=1}^{n(h)} \tilde{z}(hi)/n(h)$.

Statistical Inference

Researchers are often interested in testing hypotheses about relationships among population regression coefficients. While there is no rigorous solution to the general linear hypothesis problem in a finite population context, we suggest a heuristic approach which relies on the central limiting tendency of estimated coefficient vectors $\hat{b}$ to have approximately the multivariate normal sampling distribution with mean β and variance-covariance matrix $\text{Var}_D\{\hat{z}\}$ as specified in (A.25). To the extent that this approximation for the sampling distribution of b holds, one can justify the following approach for testing

$$H_o: C^T\beta = \delta \quad \text{versus} \quad H_A: C^T\beta \neq \delta \tag{A.31}$$

Form the test statistic

$$T_c^2 = (C^Tb - \delta)^T[C^T \text{var}_D(\hat{z})\ C]^{-1}(C^Tb - \delta) \tag{A.32}$$

and reject H_o if T_c^2 exceeds the upper α percentage point of the Chi-Square distribution with c = rank (C) degrees of freedom. This test is the multivariate analogue of the common large sample normal theory test.

When the degrees of freedom (df) associated with the estimated variance-covariance matrix $\text{var}_D(\hat{z})$ drops below 60, (A.32) may be viewed as the multivariate analogue of Student's T; namely, Hotelling's T^2 statistic. The F-transformed version of Hotelling's T^2 is expressed as

$$F = \left\{\frac{df + 1 - c}{(df)c}\right\} T_c^2 \tag{A.33}$$

where df is the degrees of freedom associated with $\text{var}_D(\hat{z})$. For our example in the previous section we would recommend the approximation df $\doteq$ n(+)-H where n(+) is the total number of clusters in the sample and H is the number of strata. The transformed statistic F is compared to the upper α percentage point of Fisher's F distribution with c and (df + 1 - c) degrees of freedom.