U-STATISTICS ESTIMATION OF VARIANCE COMPONENTS FOR UNEQUAL PROBABILITY SAMPLES WITH NONADDITIVE INTERVIEWER AND RESPONDENT ERRORS

Ralph Folsom, Jr., Research Triangle Institute

1. Introduction

Since many of the large scale national household surveys conducted or sponsored by government agencies utilize multistage sample designs with probability proportional to size (PPS) selections at the initial stages, there is a need for variance component estimation methodology to separate PPS sampling error components from interviewer and respondent error components. While there have been extensions of the total survey error model developed initially by Hansen, Hurwitz, and Bershad (1961) for equal probability samples, the developments by Koch (1973) and Koop (1975) have both used the Horvitz-Thompson (HT) (1952) expression for the unequal probability without replacement sampling variance. The difficulty with the (HT) sampling variance expression as well as the alternative Yates-Grundy (1953) form is their inability to distinguish sample design parameters (sample sizes at the various selection stages) from associated. population variance component parameters. Therefore, while such approaches display the contributions of various survey error sources to the total variance, no explicit functional relationship between the sample size and the sampling variance contribution has been provided. Such a fuctional relationship in terms of sample size parameters is required if survey error models are to serve the intended purpose of total survey design optimization. In 1975, G.B. Gray presented a new variance and covariance component representation for the sampling variance of the HT total estimator that overcomes this difficulty.

The following results begin with an extension of Gray's (1975) variance-covariance representation to a wider class of proportional to size (PPS) selection methods including with replacement, minimum replacement, and without replacement schemes. Probability minimum replacement (PMR) selection as defined by James Chromy (1979) refers to methods like PPS systematic where frame units u(k) with size measures s(k) exceeding (1/n)-th of the aggregate universe size measure

$$s(+) = \sum_{k=1}^{N} s(k)$$

have a chance for multiple selections. With n(k) denoting the random selection frequency defined for each of the N universe units u(k) when n selections are made, and with E{n(k)} or En(k) denoting the expected number of hits on unit k in n selections, the condition for strict PPS selections is

$$En(k) = n\, s(k)/s(+) = n\, P(k)$$

for all universe units u(k). The PPS 'Minimum Replacement' feature of Chromy's PMR scheme is characterized as follows:

$$\mathrm{Prob}\{n(k) = \mathrm{Int}[nP(k)]\} = 1 - \mathrm{Frac}[nP(k)]$$

and

$$\mathrm{Prob}\{n(k) = \mathrm{Int}[nP(k)] + 1\} = \mathrm{Frac}[nP(k)]$$

where Int(x) denotes the integer part of x and Frac(x) is the fractional part of x. Stated simply, PMR samples are PPS samples such that

$$|n(k) - nP(k)| < 1$$

for all frame units u(k). If nP(k) < 1 for all u(k), then a PMR selection routine is a PPS without replacement scheme. While PPS with replacement selections satisfy En(k) = nP(k), the range of n(k) can extend from zero to n for any universe unit.

2. Sampling Variance Components for PPS Selections

To extend Gray's (1975) variance-covariance component expression for the (HT) universe total, assume for the moment that one could observe free of error a variate value Y(k) associated with each frame unit u(k). Now, let

$$y(k) = Y(k)/P(k) = s(+)\, Y(k)/s(k)$$

depict unit k single draw ratio estimates for the universe total

$$Y(+) = \sum_{k=1}^{N} Y(k) \quad .$$

For a PPS sample with n selections, one can define single draw random indicator variables by randomly assigning sample unit labels i = 1, 2,...., n to the n selections; that is, define

$$\lambda_k(i) = \begin{cases} 1 & \text{if frame unit } u(k) \text{ belongs to the sample and is randomly assigned sample label } i, \\ 0 & \text{otherwise.} \end{cases}$$

It is easy to see that

$$E\{\lambda_k(i)\} = En(k)/n = P(k)$$

for all i = 1,2,...., n where the expectation is over all possible samples and all possible random label assignments given the sample. Similarly, we define double draw probabilities

$$E\{\lambda_k(i)\lambda_{k'}(i')\} = P(kk') = \begin{cases} E\{n(k)[n(k)-1]/n(n-1)\} & \text{if } k = k' \ \&\ i \neq i' \\ E\{n(k)n(k')/n(n-1)\} & \text{if } k \neq k' \ \&\ i \neq i'. \end{cases}$$

Using the single draw indicators $\lambda_k(i)$, unbiased single draw estimators for Y(+) are defined by

$$y(i) = \sum_{k=1}^{N} \lambda_k(i)\, y(k) \quad .$$

One can view such statistics as random group estimators for groups of size one. With the single draw indicators defined by equally likely permutations of n sample labels, over repeated samples and repeated labelling

$$E\{y(i)\} = Y(+)$$

for all i = 1(1)n. Furthermore, all the n single draw estimators have a common variance

$$\mathrm{Var}\{y(i)\} = \sum_{k=1}^{N} P(k)[y(k)-Y(+)]^2 = \sigma^2$$

and a common covariance $\sigma^2\rho = \mathrm{Cov}\{y(i)y(i')\}$ defined by

$$\sigma^2\rho = \sum_{k=1}^{N} \sum_{k'=1}^{N} P(kk')[y(k)-Y(+)][y(k')-Y(+)] \quad .$$

The common variance σ^2 is the familiar PPS with replacement variance component. The common covariance component leads to the common correlation

$$\mathrm{Cor}\{y(i)y(i')\} = \mathrm{Cov}\{y(i)y(i')\}/\sigma^2 = \rho \quad .$$

While the covariance component $\sigma^2\rho$ is not strictly independent of the sample size parameter n since the P(kk') joint draw probabilities depend to some extent on n, it is the authors contention that for moderately large universes P(kk') is reasonable independent of n. For the without replacement case where P(kk) = 0 and P(kk') = π(kk')/ n(n-1) is the joint inclusion probability over n(n-1), Hartley and Rao (1962) have derived a Taylor series approximation for π(kk') in PPS systematic sampling from a randomly permuted frame listing. Assuming that n is much smaller than N and that P(k) = s(k)/s(+) is of order $O(N^{-1})$, Hartley and Rao show that after including all terms in the expansion of order $O(N^{-4})$ and larger, π(kk')/n(n-1) is strictly independent of n. This result suggest that it is not unreasonable to treat the covariance component $\sigma^2\rho$ as a population parameter for moderately large universes.

Since the single draw variates y(i) are identically distributed with common covariance $\sigma^2\rho$, the best linear combination of these y(i) for estimating Y(+) is their simple average; that is

$$\bar{y} = \sum_{i=1}^{n} y(i)/n = \sum_{k=1}^{N} n(k)Y(k)/En(k) \quad .$$

The simple average $\bar{y}$ above is equivalent to Chromy's PMR estimator for a PPS minimum replacement selection and reduces to the without replacement Horvitz-Thompson estimator when En(k) < 1 for all k. For with replacement PPS selections, $\bar{y}$ is the familiar unbiased estimator for Y(+). The representation of $\bar{y}$ as an average of single draw variates y(i) leads to the variance partitioning

$$\mathrm{Var}\{\bar{y}\} = \sigma^2/n + (n-1)\sigma^2\rho/n = \sigma^2[1+(n-1)\rho]/n \quad .$$

To show that this variance-covariance component representation is equivalent to the Yates-Grundy (1953) type variance expression developed by Chromy, one can use the identities

$$\sum_{k=1}^{N} P(k) = 1 \quad \text{and} \quad \sum_{k=1}^{N} P(kk') = P(k')$$

to develop alternative expressions for σ^2 and $\sigma^2\rho$; namely,

$$\sigma^2 = \sum_{k=1}^{N} \sum_{k'=1}^{N} P(k)P(k')[y(k)-y(k')]^2/2$$

and

$$\sigma^2\rho = \sum_{k=1}^{N} \sum_{k'=1}^{N} [P(k)P(k')-P(kk')][y(k)-y(k')]^2/2 \quad .$$

In this form, it is easy to show that

$$\mathrm{Var}\{\bar{y}\} = \sum_{k=1}^{N} \sum_{k'\neq k} [En(k)En(k')-E\{n(k)n(k')\}]$$
$$[Y(k)/En(k)-Y(k')/En(k')]^2/2 \quad .$$

In this form, it is clear that for with replacement selections where P(kk')=P(k)P(k') when k $\neq$ k', the covariance component $\sigma^2\rho = 0$. For minimum replacement and without replacement samples, the $[1+(n-1)\rho]$ quantity in brackets is the effect of PMR selection. This quantity is the precise analogue of the simple random sampling fpc when it is properly stated in terms of the with replacement SRS variance component σ^2 and the without replacement induced common covariance $\sigma^2\rho = -\sigma^2/(N-1)$.

The new sequential PMR selection scheme developed by Chromy provides for unbiased variance estimability while retaining the implicit stratification advantages of the PPS systematic scheme where a controlled ordering of the frame units quarantees one selection per equal sized zone or sampling interval marked off seqüentially down the ordered listing. In this case, the quantity $[1+(n-1)\rho]$ will also reflect the effect of implicit stratification due to controlled ordering.

Use of the single draw variates to develop the variance-covariance component partitioning for the variance of PPS sample statistics suggests a link to classical U-statistics estimation methods as elaborated by Hoeffding (1948). Considering the alternative expressions for σ^2 and $\sigma^2\rho$, one can define the following symetric frame kernels of degree two

$$\sigma^2(kk') = P(k)P(k')[y(k)-y(k')]^2/2$$

$$\sigma^2\rho(kk') = [P(k)P(k')-P(kk')][y(k)-y(k')]^2/2$$

where the degree of the kernel reflects the number of frame units required for its definition and the symetry is with respect to permutations of the frame unit labels k and k'. In terms of these frame kernels, the corresponding variance and covariance components are generalized totals of the form

$$\sigma^2 = \sum_{k=1}^{N} \sum_{k'=1}^{N} \sigma^2(kk') .$$

In terms of the single draw indicators, the appropriate degree 2 sample kernel for estimating σ^2 is

$$\hat{\sigma}^2(ii') = \sum_{k=1}^{N} \sum_{k'=1}^{N} \lambda_k(i)\ \lambda_{k'}(i')\ \sigma^2(kk')/P(kk')$$

The sample kernel $\hat{\sigma}^2\hat{\rho}(ii')$ is defined analogously. As long as the P(kk') double draw probabilities are positive for all frame unit pairs with $k \neq k'$ or equivalently if $E\{n(k)n(k')\} > 0$ when $k \neq k'$, the sample kernels are unbiased for every pair of sampling units with $i \neq i'$. The U-statistics estimators are generalized sample means averaging the sample kernels over all $\binom{n}{2}$ distinct pairs of sampling units; namely,

$$\hat{\sigma}^2 = \sum_{i=1}^{n} \sum_{i'>i} \hat{\sigma}^2(ii')/\binom{n}{2}$$

$$= \sum_{i=1}^{n} \sum_{i'>i} [P(i)P(i')/P(ii')][y(i)-y(i')]^2/2\binom{n}{2}$$

$$= \sum_{i=1}^{n} \sum_{i'>i} w(ii')[y(i)-y(i')]^2/2\binom{n}{2}$$

and

$$\hat{\sigma}^2\hat{\rho} = \sum_{i=1}^{n} \sum_{i'>i} [w(ii')-1][y(i)-y(i')]^2/2\binom{n}{2}$$

$$= \hat{\sigma}^2 - \sum_{i=1}^{n} [y(i)-\bar{y}]^2/(n-1)$$

$$= \hat{\sigma}^2 - s^2$$

where s^2 is the familiar PPS with replacement variance component estimator. In the next section, PPS sample U-statistics of degree m are defined, and a Yates-Grundy type unbiased variance estimator for such statistics is obtained. These results provide compact expressions for the variances of degree 2 statistics like $\hat{\sigma}^2$ and $\hat{\sigma}^2\hat{\rho}$.

3. General PPS Sample U-Statistics

The natural extension of a universe total like Y(+) to a parameter of degree m is

$$F(\underset{\sim}{+}) = \sum_{k_1=1}^{N} \ldots \sum_{k_m=1}^{N} F(k_1 \ldots k_m) = \sum_{\underset{\sim}{k}}^{N^m} F(\underset{\sim}{k})$$

where the summation above extends over all possible subsets of m universe units including those subsets with multiple representation of the same unit. While attention has been restricted to symetric functions F that are uniformly zero for subsets $\underset{\sim}{k}$ including the same unit more than once, carrying along the superfluous zero terms in the $F(\underset{\sim}{+})$ defining summation simplifies the estimation theory for with replacement and PMR selections where multiple selections of the same unit are permissable. Multiplying the single draw indicators together, an indicator for the m element subset $\underset{\sim}{k}$ of frame units belonging to the sample and being assigned the m distinct sample labels $\underset{\sim}{i} = (i_1, i_2, \ldots, i_m)$ is

$$\lambda_{\underset{\sim}{k}}(\underset{\sim}{i}) = \prod_{\ell=1}^{m} \lambda_{k_\ell}(i_\ell) \quad .$$

The expected value of $\lambda_{\underset{\sim}{k}}(\underset{\sim}{i})$ will be denoted by $P(\underset{\sim}{k})$ where

$$P(\underset{\sim}{k}) = E\ \{ \prod_{\ell=1}^{c} \prod_{j=1}^{m(k_\ell)} [n(k_\ell)-j+1]\}/m!\binom{n}{m}$$

when the subset $\underset{\sim}{k}$ contains c distinct frame units with unit $u(k_\ell)$ included $m(k_\ell)$ times in the subset $\underset{\sim}{k}$. Notice that the m dimensional subset frequencies $m(k_\ell)$ sum to m. For PPS with replacement selections where the single draw variates are independent so long as $i_\ell \neq i_{\ell'}$,

$$P(\underset{\sim}{k}) = E \prod_{\ell=1}^{m} \lambda_{k_\ell}(i_\ell) \ = \prod_{\ell=1}^{m} E\{\lambda_{k_\ell}(i_\ell)\} = \prod_{\ell=1}^{m} P(k_\ell).$$

For PPS without replacement selections where only distinct frame subsets can be included in the sample,

$$P(\underset{\sim}{k}) = \pi(\underset{\sim}{k})/m!\binom{n}{m}$$

with $\pi(\underset{\sim}{k})$ denoting the multiple inclusion probability for the distinct frame subset $(\underset{\sim}{k})$. For PMR selections, no subsets $\underset{\sim}{k}$ with frequencies $m(k_\ell)$ exceeding $\{\mathrm{Int}[nP(k_\ell)]+1\}$ can occur.

With these definitions of multiple (m) draw subsample indicators $\lambda_{\underset{\sim}{k}}(\underset{\sim}{i})$ and their corresponding expectations $P(\underset{\sim}{k})$,

$$f(\underset{\sim}{i}) = \sum_{\underset{\sim}{k}}^{N^m} \lambda_{\underset{\sim}{k}}(\underset{\sim}{i})F(\underset{\sim}{k})/P(\underset{\sim}{k}) = \sum_{\underset{\sim}{k}}^{N^m} \lambda_{\underset{\sim}{k}}(\underset{\sim}{i})f(\underset{\sim}{k})$$

is by construction an unbiased symetric kernel of degree m for estimating $F(\underset{\sim}{+})$ so long as the selection procedure quarantees that $P(\underset{\sim}{k}) > 0$ for all distinct subsets of frame units $\underset{\sim}{k}$ where $F(\underset{\sim}{k}) > 0$. Therefore, the corresponding degree m U-statistic

$$\hat{F}(\underset{\sim}{+}) = \sum_{\underset{\sim}{i}}^{\binom{n}{m}} f(\underset{\sim}{i})/\binom{n}{m}$$

is unbiased if

$$E\{ \prod_{\ell=1}^{m} n(k_\ell)\} > 0$$

for all $\binom{N}{m}$ subsets of m distinct frame units. Changing the order of summations in $\hat{F}(+)$ and $f(\underset{\sim}{i})$, the following alternative expression for a degree m U-statistic is obtained

$$\hat{F}(+) = \sum_{\underset{\sim}{k}}^{N^m} \{ \sum_{\underset{\sim}{i}}^{\binom{n}{m}} \lambda_{\underset{\sim}{k}}(\underset{\sim}{i})/\binom{n}{m}\} F(\underset{\sim}{k})/P(\underset{\sim}{k})$$

$$= \sum_{\underset{\sim}{k}}^{N^m} \hat{P}(\underset{\sim}{k})\ F(\underset{\sim}{k})/P(\underset{\sim}{k})$$

where the indicator based U-statistics in curley brackets are the unbiased estimators for $P(\underset{\sim}{k})$ obtained by removing the expectation operator from the original expression for $P(\underset{\sim}{k})$ in terms of the $n(k_\ell)$ full sample selection frequencies and $m(k_\ell)$ subset frequencies. Recalling that $F(\underset{\sim}{k}) = 0$ by definition for frame unit subsets $\underset{\sim}{k}$ with repeated elements, $\hat{F}(\underset{\sim}{+})$ can be recast in terms of a sum over distinct subsets $\underset{\sim}{k}$; namely

$$\hat{F}(\underset{\sim}{+}) = \sum_{\underset{\sim}{k}}^{\binom{N}{m}} \{ \prod_{\ell=1}^{m} n(k_\ell)\} F(\underset{\sim}{k})/E\{ \prod_{\ell=1}^{m} n(k_\ell)\} \quad .$$

Noticing that

$$\sum_{\underset{\sim}{k}}^{N^m} \lambda_{\underset{\sim}{k}}(\underset{\sim}{i}) = 1$$

for all sampling unit subsets $(\underset{\sim}{i})$, it is clear that the $P(\underset{\sim}{k})$ quantities also sum to one over all of the N^m subsets $\underset{\sim}{k}$. These summation identities lead directly to the variance expression

$$Var\{\hat{F}(\underset{\sim}{+})\}= \sum_{\underset{\sim}{k}}^{N^m} \sum_{\underset{\sim}{k}'}^{N^m} E\{\hat{P}(\underset{\sim}{k})\hat{P}(\underset{\sim}{k}')\}[f(\underset{\sim}{k})-F(\underset{\sim}{+})]$$

$$[f(\underset{\sim}{k}')-F(\underset{\sim}{+})]$$

and the Yates-Grundy type alternative

$$Var\{\hat{F}(\underset{\sim}{+})\} = \sum_{\underset{\sim}{k}}^{N^m} \sum_{\underset{\sim}{k}\neq\underset{\sim}{k}'} [P(\underset{\sim}{k})P(\underset{\sim}{k}')-E\{\hat{P}(\underset{\sim}{k})\hat{P}(\underset{\sim}{k}')\}]$$

$$[f(\underset{\sim}{k})-f(\underset{\sim}{k}')]^2/2 \quad .$$

For without replacement (wor) selections where no subsets $\underset{\sim}{k}$ with duplicate units can occur, the latter expression reduces to

$$Var\{\hat{F}(+)\}_{wor} = \sum_{\underset{\sim}{k}}^{\binom{N}{m}} \sum_{\underset{\sim}{k}'\neq\underset{\sim}{k}} [\pi(\underset{\sim}{k})\pi(\underset{\sim}{k}')-\pi(\underset{\sim}{k}\cup\underset{\sim}{k}')]$$

$$[F(\underset{\sim}{k})/\pi(\underset{\sim}{k})-F(\underset{\sim}{k}')/\pi(\underset{\sim}{k}')]^2/2$$

where $\pi(\underset{\sim}{k})$ represents the joint sample inclusion probability for subsets of m distinct frame units labeled $\underset{\sim}{k}$ and $\pi(\underset{\sim}{k}\cup\underset{\sim}{k}')$ is the corresponding joint inclusion probability for the union of the frame unit subsets labeled $\underset{\sim}{k}$ and $\underset{\sim}{k}'$.

The general variance expression for our degree m PPS sample U-statistic suggests an estimator for $Var\{\hat{F}(\underset{\sim}{+})\}$ that is itself a U-statistic of degree 2m; namely

$$var\{\hat{F}(\underset{\sim}{+})\} = \sum_{\underset{\sim}{k}\varepsilon s} \sum_{\underset{\sim}{k}'\neq\underset{\sim}{k}} \hat{P}(\underset{\sim}{k})\hat{P}(\underset{\sim}{k}')[\omega(\underset{\sim}{k}\underset{\sim}{k}')-1]$$

$$[f(\underset{\sim}{k})-f(\underset{\sim}{k}')]^2/2$$

where

$$\omega(\underset{\sim}{k}\underset{\sim}{k}') = P(\underset{\sim}{k})P(\underset{\sim}{k}')/E\{\hat{P}(\underset{\sim}{k})\hat{P}(\underset{\sim}{k}')\}$$

and the $\underset{\sim}{k}$ subsets belonging to the sample (s) can contain as many as n(k) contributions from the frame unit u(k), so long as n(k) < m. For the without replacement case the general expression reduces to

$$var\{\hat{F}(\underset{\sim}{+})\} = \sum_{\underset{\sim}{k}\varepsilon s}^{\binom{n}{m}} \sum_{\underset{\sim}{k}'\neq\underset{\sim}{k}} [\omega(\underset{\sim}{k}\underset{\sim}{k}')-1][F(\underset{\sim}{k})/\pi(\underset{\sim}{k})$$

$$-F(\underset{\sim}{k}')/\pi(\underset{\sim}{k}')]^2/2$$

with

$$\omega(\underset{\sim}{k}\underset{\sim}{k}') = \pi(\underset{\sim}{k})\pi(\underset{\sim}{k}')/\pi(\underset{\sim}{k}\cup\underset{\sim}{k}').$$

For the with replacement case, Folsom and Lessler (1980) have developed an unbiased variance-covariance component partitioning of $var\{\hat{F}(\underset{\sim}{+})\}$ that avoids the calculation of joint draw probabilities for subsets $\underset{\sim}{k}$ including more than m distinct units.

In the following section, single draw indicators are utilized in a manner analogous to that exploited by Wilk and Kempthorne (1955) to derive a random effects model for a single draw variate $y_t(hi)$ incorporating nonadditive interviewer h and sampling unit i effects.

4. Total Variance Models with Interacting Interviewer and Respondent Effects

Extending the previous results for defining single draw sampling unit variables y(i), to corresponding variates observed by a randomly assigned interviewer h on a given repeat interview trial t, we first consider a universe of A interviewers from which (a) are selected for assignment to n = ar PPS selected sampling units. Interviewer selection indicators $T_j(h)$ are defined to assume the value 1 when candidate j is selected from among the A eligible interviewers and is assigned sample interviewer label (h); otherwise $T_j(h) = 0$. Single draw sampling indicators are defined to indicate PPS sample selection and subsequent random assignment of replicate labels denoting the $m = n/a^2 = r/a$ sampling units allocated to the (hh') interviewer pair of a completely balanced cross-over design. In the proposed cross-over reinterview scheme, interviewer h visits the members of assignment (hh') on the first interview trial (t=1) and interviewer h' visits them on the second trial (t=2). All a^2 combinations including the (hh) repeat measurements are included in this perfectly balanced design. The sampling unit selection and assignment indicators $\lambda_k(hh'i)$ take the value 1 when sampling frame unit u(k) belongs to the sample and is assigned replicate label i of interviewer assignment pair (hh'); otherwise $\lambda_k(hh'i) = 0$. These random selection and assignment indicators have the following properties

$$E\{T_j(h)\} = (1/A);\ E\{T_j(h)T_{j'}(h')\} = 1/A(A-1)$$

and

$$E\{\lambda_k(hh'i)\} = P(k);\ E\{\lambda_k(hh'i)\lambda_{k'}(h''h'''i')\} = P(kk')$$

since the interviewer selection and assignment process is independent of the sampling unit selection process.

With these definitions, single draw variates which assume no residual or carry-over interviewer effects from trial to trial are defined as follows:

$$y_1(hh'i) = \sum_{j=1}^{A} \sum_{j'=1}^{A} \sum_{k=1}^{N} T_j(h)T_{j'}(h')\ \lambda_k(hh'i)\ [y(jk)+\varepsilon_1(jk)]$$

and

$$y_2(hh'i) = \sum_{j=1}^{A} \sum_{j'=1}^{A} \sum_{k=1}^{N} T_j(h)T_{j'}(h')\ \lambda_k(hh'i)\ [y(j'k)+\varepsilon_2(j'k)]$$

where the y(jk) variates represent the expected value over a conceptual series of independent repeat interviews of frame unit u(k) by candidate interviewer j; that is

$$E\{Y_t(jk)/P(k)\} = Y(jk)/P(k) = y(jk)\ .$$

Implicit in this result is the assumption that

$$E_t\{[Y_t(jk)-Y(jk)]/P(k)\} = E_t\{e_t(jk)/P(k)\} = E\{\varepsilon_t(jk)\} = 0$$

independent of the sample and interviewer selection/assignment process. These $\varepsilon_t(jk)$ errors are called 'intrinsic response errors' by Koch, Freeman, and Freeman (1975) since they are not under the direct control of the survey operation. The conceptual series of repeat interviews of unit u(k) by candidate interviewer j is assumed to be an uncorrelated process over trials so that

$$Var\{\varepsilon_t(jk)\} = E\{\varepsilon_t^2(jk)\} = \sigma_\varepsilon^2(jk).$$

The $T_j(h)$ and $\lambda_k(hh'i)$ variables, on the other hand, give rise to 'external response errors' relating to interviewer main effects and interviewer by respondent interactions. These external effects are randomized sample realizations of frame unit effects $\alpha(j)$ and $\alpha\zeta(jk)$ defined in terms of the frame unit identity

$$\begin{aligned} y(jk) &= Y(\cdot+) + [Y(j+)-Y(\cdot+)] + [y(\cdot k)-Y(\cdot+)] \\ &\quad + [y(jk)-Y(j+)-y(\cdot k)+Y(\cdot+)] \\ &= \mu + \alpha(j) + \zeta(k) + \alpha\zeta(jk) \end{aligned}$$

where

$$y(\cdot k) = \sum_{j=1}^{A} y(jk)/A = Y(\cdot k)/P(k)$$

$$Y(\cdot+) = \sum_{k=1}^{N} P(k)y(\cdot k) = \sum_{k=1}^{N} Y(\cdot k)$$

and

$$Y(j+) = \sum_{k=1}^{N} P(k)y(jk) = \sum_{k=1}^{N} Y(jk)\ .$$

Exploiting this identity, the single draw variates defined for our cross-over repeat measurement design have the following random effects representation

$$y_1(hh'i) = \mu + \alpha(h) + \zeta(i) + \alpha\zeta(hi) + \varepsilon_1(hi)$$

and

$$y_2(hh'i) = \mu + \alpha(h') + \zeta(i) + \alpha\zeta(h'i) + \varepsilon_2(h'i)$$

with the random indicators T and λ providing the link between candidate interviewer/frame unit combinations (jk) and survey interviewer/sampling unit pairs (hi).

In this development, the statistical properties of our cross-over design are fully determined by the physical probability selection/randomization process and the assumption of independence between the intrinsic response errors and the external sampling and interviewer errors. The repeat interviews and cross-over interviewer assignments are required to separate the intrinsic response error variances and covariances from the external sampling error and interactive interviewer error components. We have assumed that the intrinsic response errors are independent when $h \neq h'$ or $t \neq t'$. The explicit definitions of these error variances and covariances, derived as a consequence of the probability selection/randomization process, are as follows:

$$\sigma^2_\varepsilon = \sum_{j=1}^{A} \sum_{k=1}^{N} P(k)\sigma^2_\varepsilon(jk)/A \ ;$$

$$\sigma_{\varepsilon\varepsilon'} = \sum_{j=1}^{A} \sum_{k=1}^{N} \sum_{k'=1}^{N} P(kk')\sigma_{\varepsilon\varepsilon'}(jkk')/A \ ;$$

$$\sigma^2_{(\alpha\zeta)} = \sum_{j=1}^{A} \sum_{k=1}^{N} P(k)\alpha\zeta^2(jk)/A \ ;$$

$$\sigma_{(\alpha\zeta\zeta')} = \sum_{j=1}^{A} \sum_{k=1}^{N} \sum_{k'=1}^{N} P(kk')\alpha\zeta(jk)\alpha\zeta(jk')/A \ ;$$

$$\sigma_{(\alpha\zeta)(\alpha'\zeta')} = -\sigma_{(\alpha\zeta\zeta')}/(A-1) \ ;$$

$$\sigma^2_\zeta = \sum_{k=1}^{N} P(k)\zeta^2(k) \ ;$$

$$\sigma_{\zeta\zeta'} = \sum_{k=1}^{N} \sum_{k'=1}^{N} P(kk')\zeta(k)\zeta\ (k') \ ;$$

$$\sigma^2_\alpha = \sum_{j=1}^{A} \alpha^2(j)/A \ ;$$

$$\sigma_{\alpha\alpha'} = -\sigma^2_\alpha/(A-1) \ .$$

With these component definitions, Folsom and Lessler (1980) derived the total variance partitioning of the single draw variate mean

$$\bar{y}_{\cdot}(\cdots) = \sum_{h=1}^{a} \sum_{h'=1}^{a} \sum_{i=1}^{m} [y_1(hh'i)+y_2(hh'i)]/2m.$$

The variance of $\bar{y}_{\cdot}(\cdots)$ for the proposed completely balanced cross-over repeat interview design is

$$\begin{aligned}
\mathrm{Var}\{\bar{y}_{\cdot}(\cdots)\} &= \sigma^2_\alpha[1-(a-1)/(A-1)]/a \\
&+ \sigma^2_\zeta[1+(n-1)\rho_{\zeta\zeta'}]/n \\
&+ \sigma^2_{(\alpha\zeta)}[1+(r-1)\rho_{(\alpha\zeta\zeta')}]/n \\
&+ \sigma^2_\varepsilon[1+(r-1)\rho_{\varepsilon\varepsilon'}]/2n \\
&- (a-1)(A/A-1)\sigma^2_{(\alpha\zeta)}[1-\rho_{(\alpha\zeta\zeta')}]/2an \\
&- (a-1)\sigma^2_{(\alpha\zeta)}\rho_{(\alpha\zeta\zeta')}/a(A-1) \ .
\end{aligned}$$

Unbiased U-statistics estimators for all of the variance and covariance components in $\mathrm{Var}[\bar{y}_{\cdot}(\cdots)]$ are presented in Folsom and Lessler (1980). The component estimators are formed from linear combinations of sample mean squares derived as U-statistics from the following sample kernels:

$$\begin{aligned}
&E\{\omega(hh'i;hh'i')\ [y_1(hh'i)-y_1(hh'i')] \\
&\qquad [y_2(hh'i)-y_2(hh'i')]/2\} \\
&\quad = \sigma^2_\zeta - \sigma^2_{(\alpha\zeta)}/(A-1)
\end{aligned}$$

$$\begin{aligned}
&E\{[y_1(hh'i)-y_1(hh'i')][y_2(hh'i)-y_2(hh'i')]/2\} \\
&\quad = \sigma^2_\zeta - \sigma_{\zeta\zeta'} - \sigma^2_{(\alpha\zeta)}/(A-1) + \sigma_{(\alpha\zeta\zeta')}/(A-1)
\end{aligned}$$

$$\begin{aligned}
&E\{\omega(hh'i;hh''i')[y_1(hh'i)-y_1(hh'i')]^2/2\} \\
&\quad = \sigma^2_\zeta + \sigma^2_{(\alpha\zeta)} + \sigma^2_\varepsilon - \overset{*}{\sigma}_{\varepsilon\varepsilon}
\end{aligned}$$

$$\begin{aligned}
&E\{[y_1(hh'i)-y_1(hh''i')]^2/2\} \\
&\quad = \sigma^2_\zeta - \sigma_{\zeta\zeta'} + \sigma^2_{(\alpha\zeta)} - \sigma_{(\alpha\zeta\zeta')} + \sigma^2_\varepsilon - \sigma_{\varepsilon\varepsilon'}
\end{aligned}$$

$$E\{[y_1(hhi)-y_2(hhi)]^2/2 = \delta^2(hhi)/2\} = \sigma^2_\varepsilon$$

$$E\{\delta(hhi)\delta(hhi')/2\} = \sigma_{\varepsilon\varepsilon'}$$

$$E\{\omega(hhijhhi')\delta(hhi)\delta(hhi')/2\} = \overset{*}{\sigma}_{\varepsilon\varepsilon'}$$

$$\begin{aligned}
E\{[\bar{y}_1(h\cdot\cdot)-\bar{y}_1(h'\cdot\cdot)]^2/2\} &= \sigma^2_\alpha+\sigma^2_\zeta[1+(r-1)\rho_{\zeta\zeta'}]/r \\
&+ \sigma^2_{(\alpha\zeta)}[1+(r-1)\rho_{(\alpha\zeta\zeta')}]/r \\
&+ \sigma^2_\varepsilon[1+(r-1)\rho_{\varepsilon\varepsilon'}]/r - \sigma_{\zeta\zeta'} \\
&+ \sigma_{(\alpha\zeta\zeta')}/(A-1) \ .
\end{aligned}$$

The $\omega(hh'i;hh''i')$ variance weights are formed as $\omega(kk') = P(k)P(k')/P(kk')$ from the single and double draw probabilities linked to sample units (hh'i) and hh''i'). In practice, one could intertain either a fixed interviewer effects model, in which case a = A, or a random interviewer effects model with A = ∞.

These results can be extended directly to a variance-covariance matrix partitioning and associated component estimators for the total survey error of a vector valued single draw variate $\bar{y} = (\bar{y}_1,\ldots\bar{y}_p)$. Corresponding models for specific nonlinear statistics formed from such vector means can be obtained immediately using Taylor series linearized versions of the single draw variates. Extensions to multistage PPS samples and nested split-plot type survey agent (supervisors, coders, interviewers) randomization schemes are currently being pursued. Future research plans include specification of Jackknife pseudo-replication versions of our U-statistic variance estimators and development of a finite population central limit theorem for PPS sample U-statistics in the probability minimum replacement setting.

5. References

DUE TO LIMITATION OF SPACE REFERENCES MAY BE OBTAINED SEPARATELY FROM THE AUTHOR.