# Some Generalizations of the Horvitz-Thompson Estimator

James R. Chromy
RTI International, 3040 Cornwallis Road, P.O. Box 12194, Research Triangle Park, NC 27709

## Abstract

Horvitz and Thompson (1952) develop a theory for PPS sampling and estimation when sampling is PPS without replacement. The estimation procedure weights each selected observation by the inverse of the unit's overall selection probability and the unit's weight does not depend on the particular sample in which it was selected. They also show that nonzero pairwise probabilities are required for unbiased variance estimation. This paper shows how the Horvitz-Thompson PPS without replacement estimator and its variance are related to the PPS with replacement estimator and its variance as presented by Hansen, Hurwitz, and Madow. It also shows how this generalization can be related to minimum replacement designs which allow sampling units with large relative size measures (greater than 1/n) to be selected more than once.

**Key Words:** PPS sampling, PPS sequential sampling, PPS systematic sampling, variance approximation

I'm honored to have the opportunity to speak at this session organized in memory of Dr. Horvitz. Thanks to Ralph Folsom for organizing the session. I first met Dan while taking Dr. Charles Proctor's sampling course at N. C. State. The class was invited to hear from a practicing statistician about how sampling is really done. Dan gave a wonderful description of the multi-stage sampling procedure then used at RTI to select area household samples. Later that spring when I needed a job, I came to see Dan again and I was very pleased to receive an offer of employment. I accepted and began working at RTI in June 1966 in Dan's Survey Statistics Department of the Statistics Research Division.

When you worked for Dan, you had to earn his trust. He was a great mentor from the start and has encouraged me throughout my career. My last collaboration with him was after he retired and was living in Florida. He, Al Finkner, and I co-authored a chapter for a special National Assessment of Educational Progress history edited by Lyle Jones and Ingram Olkin (Chromy, Finkner, and Horvitz 2004).

Dan's paper was entitled "A generalization of sampling without replacement from a finite universe". In this presentation, I would like to review some of his results. Then I would like to add some additional generalizations: (1) allowing units to be drawn more than once and (2) using expectations instead of probabilities. Next, I will discuss how the Horvitz-Thompson theory can be applied to the with replacement designs of Hansen and Hurwitz. Then I would like to characterize stratification, both explicit and implicit in terms of the variance of the generalized Horvitz-Thompson (HT) estimator. Finally, I would like to discuss useful simplifications to the variance estimation problem and how they can be justified under PPS systematic and PPS sequential sampling designs.

## 1. Horvitz and Thompson (1952)

Dan developed the HT estimation methodology while completing his doctorate at Iowa State University and published it in JASA with D. J. Thompson (Horvitz and Thompson 1952). The paper was entitled "A Generalization of Sampling without Replacement from a Finite Universe." I'm trying to generalize it a little further: a generalization of a generalization. The 1952 paper had two goals:

1. "it provides a general method for dealing with sampling without replacement from a finite universe when variable probabilities of selection are used…"
2. "it examines and discusses some problems arising in the practical application …"

Under the first goal, the paper addressed unbiased estimation of totals and unbiased estimation of the variance of these estimates. It discussed single-stage sampling, but extended the results to two-stage sampling. Under the second goal, the paper presented two sampling schemes for selecting PPS samples of size 2 without replacement and provided some examples of their application.

Prior to this work, the main contenders for sampling PPS without replacement were the Hansen and Hurwitz (1943) PPS with replacement method restricted to selecting one unit per stratum and a systematic sampling scheme described by Madow (1946). Neither of these methods provided for unbiased estimation of the variance of the estimate.

The motivation for PPS sampling was the potential reduction of variance to zero or near zero if the size measures are exactly or approximately proportional to the characteristic being estimated. Intuitively, practicing statisticians would have liked to use known available auxiliary variables to take a simple average of observed ratios to adjust a known total. While this approach did work with the PPS with replacement scheme, it would produce biased estimates of the total when using equal probability sampling.

It was also known that for equal probability sampling, without replacement sampling yielded lower variance than with replacement sampling. Intuitively, this should hold for PPS sampling also.

Two relationships for taking expectations of finite sample statistics are key elements for the logic developed later in the paper. Both of them rely on theoretically knowing the probability of selection over samples admitted under the design. The first involves the expectation of a sample function of the observed values:

$$E\left[\sum_{i=1}^{n} f(x_i)\right] = \sum_{s=1}^{S} \Pr(s)\left[\sum_{i\in s} f(x_i)\right] = \sum_{i=1}^{N} f(X_i)\sum_{s=1}^{S} \Pr(s \mid u_i \in s). \qquad (1)$$

The important issue to notice is that the summation switches from being over all samples to being over the population. The final quantity in this equation is seen to be the probability of selecting unit *i*.

$$P(u_i) = \sum_{s=1}^{S} \Pr(s \mid u_i \in s).$$

Using a similar logic, he obtained this result for pairs of observations.

$$E\left[\sum_{i=1}^{n}\sum_{j\neq i}^{n} f(x_i x_j)\right] = \sum_{s=1}^{S}\Pr(s)\left[\sum_{i\in s}\sum_{\substack{j\neq i \\ j\in s}} f(x_i x_j)\right]$$

$$= \sum_{i=1}^{N}\sum_{j\neq i}^{N} f(X_i X_j)\sum_{s=1}^{S}\Pr(s\,|\,(u_i u_j \in s)\,. \qquad (2)$$

A similar switch occurs in the order of summation and the final term in the equation is the probability of selecting unit *i* and unit *j* in the same sample.

$$P(u_i u_j) = \sum_{s=1}^{S}\Pr(s\,|\,(u_i u_j)\in s)\,.$$

With these tools in hand, the paper develops unbiased estimators for the total of some variable *X*:

$$T = \sum_{i=1}^{N} X_i$$

Horvitz and Thompson considered three possible subclasses of linear estimators of a population total and settled on class 2 which applied the same coefficient to an element regardless of the order of the draw.

$$\hat{T} = \sum_{i-1}^{n} \beta_i x_i$$

where $\beta_i$ is applied to unit *i* whenever unit *i* is selected. Then noting from equation (1), that

$$E\left(\sum_{i=1}^{n}\beta_i x_i\right) = \sum_{i=1}^{N} P(u_i)\beta_i X_i$$

where $P(u_i)$ is the sum of the sample probabilities over all samples of size *n* that contain unit *i,* he concluded that the expected value can equal the population total for any set of X only if

$$\beta_i = \frac{1}{P(u_i)}\,.$$

This conclusion yielded the classic Horvitz-Thompson estimator

$$\hat{T} = \sum_{i-1}^{n}\frac{x_i}{P(u_i)}\,.$$

In order to have the opportunity to reduce variance, the selection probabilities need to be at least approximately proportional to the characteristic of interest. When considering a sample of size 1 or with replacement sampling, the arbitrary probabilities at each sample draw are denoted by $p_i$ and they sum to 1.

$$\sum_{i=1}^{N} p_i = 1.$$

When selecting samples of size n without replacement, the unit probabilities are scaled so that $P(u_i) = np_i$ and $\sum_{i=1}^{N} P(u_i) = n.$

As noted by Horvitz and Thompson, the unit probabilities must be positive and less than 1. In order to obtain unbiased estimates of the variance, all pairwise probabilities must also be positive.

Since the development of Horvitz-Thompson PPS without replacement sampling theory, a large number of sampling schemes have been developed which utilize this theory. Brewer and Hanif (1983) enumerate 50 PPS sampling schemes many of which utilize the Horvitz-Thompson theory. Ones I have found particularly useful have been Brewer's (1963) method for samples of size 2; this method provides for straightforward calculation of the pairwise probability which can then be used to select the samples. Sampford's rejective method (Sampford 1969) extends the approach to larger samples. Both methods provide for unbiased variance estimation with non-negative coefficients in the variance estimator (a topic which is discussed later).

## 2. Generalization: From Probabilities to Expected Sample Sizes

The generalization proposed here involves substituting expected sample size for unit probabilities. Horvitz and Thompson define

$$P(u_i) = \sum_{s=1}^{S} \Pr(s \mid u_i \in s) \text{, and } P(u_i u_j) = \sum_{s=1}^{S} \Pr(s \mid (u_i u_j) \in s).$$

If instead, we define a variable for the number of times unit *i* is selected in a sample of size *n* and designate it by $n_i$, then we can compute its expectation as

$$E(n_i) = \sum_{k=1}^{n} k \sum_{s=1}^{S} \Pr(u_i \in s, k \, times).$$

Similarly, we can consider the expectation of the product of sample sizes for two units as

$$E(n_i n_j) = \sum_{u=1}^{n} \sum_{v=1}^{n} uv \sum_{s=1}^{S} \Pr(u_i \in s, u \, times \, and \, u_j \in s, v \, times).$$

The generalization of the Horvitz and Thompson theory allows the same unit to be selected more than once. Different sampling schemes can be developed to control this process in various ways. In sampling without replacement when all units have relative size measures less than *1/n,* the values of $n_i$ are limited to zero and one. In this situation,

$$E(n_i) = P(u_i) \text{ and } E(n_i n_j) = P(u_i u_j).$$

The Horvitz-Thompson paper provided formulas for the variance of the estimated total and for an estimator of that variance. Most practitioners today use an alternative expression due to Yates and Grundy (1953). Both forms of the variance formulae require knowledge of the pairwise probabilities of selection and for unbiased estimation the pairwise probabilities must be positive. In terms of the generalized form of the HT estimator, the variance can be written in a form analogous to Yates and Grundy's expression as

$$V(\hat{T}) = \frac{1}{2} \sum_{i=1}^{N} \sum_{j \neq i}^{N} \{E(n_i)E(n_j) - E(n_i n_j)\} \left[ \frac{X_i}{E(n_i)} - \frac{X_j}{E(n_j)} \right]^2. \qquad (3)$$

The determination of the pairwise probabilities or pairwise expectatons can be problematical even for samples of size 2. One solution used in an example by Horvitz and Thompson noted that it failed when any unit probabilities exceeded one-half.

The unbiased variance estimator can be expressed as

$$\hat{V}(\hat{T}) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\left(\frac{E(n_i)E(n_j) - E(n_i n_j)}{E(n_i n_j)}\right)\left[\frac{x_i}{E(n_i)} - \frac{x_j}{E(n_j)}\right]^2 \qquad (4)$$

Note that the term in curly brackets of the variance, $\{E(n_i)E(n_j) - E(n_i n_j)\}$, is the negative covariance of two unit sample sizes. In without replacement sampling we expect the covariance to be negative (making this term positive), but this is not guaranteed by all selection schemes. Shortly after coming to RTI, I worked with Dan and with Dr. John Koop to select a sample of North Carolina counties and analyze the data. State personnel conducted an intensive investigation of the level of child abuse and neglect reported in those counties. The PPS sample design permitted us to draw inference about the problem for the whole state. The frame was stratified into 5 strata with 2 counties drawn from each one using a PPS without replacement scheme. The particular scheme and particular sample outcome produced a negative variance estimate for one of the strata. When combined with the estimates from the other strata, the overall variance for the state estimate was however positive. I decided to include the negative value in the calculation of the overall variance since I could do this and still get a positive estimate of variance. I recall discussing this problem with Dan and his concurring with the approach taken.

### 3. Application to Hansen and Hurwitz' With Replacement Design

Consider now the Hansen-Hurwitz with replacement sampling scheme. The individual unit sample sizes follow the multinomial distribution in PPS sampling with replacement. From the multinomial distribution, we have that

$$E(n_i) = np_i, \quad Cov(n_i n_j) = -np_i p_j, \text{ and } E(n_i n_j) = n(n-1)p_i p_j.$$

The Hansen-Hurwitz with replacement estimator is expressed as

$$\hat{T}_{wr} = \frac{1}{n}\sum_{i=1}^{n}\frac{x_i}{p_i}$$

which is algebraically equivalent to the generalized form of the HT estimator

$$\hat{T} = \sum_{i-1}^{n}\frac{x_i}{E(n_i)}.$$

The variance of the Hansen-Hurwitz with replacement sampling estimator can then be expressed by the generalized Yates-Grundy variance formula as

$$V(\hat{T}) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j\neq i}^{N}\{np_i p_j\}\left[\frac{X_i}{np_i} - \frac{X_j}{np_j}\right]^2,$$

and the variance estimate simplifies even further to

$$\hat{V}(\hat{T}) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\left(\frac{1}{n-1}\right)\left[\frac{x_i}{np_i} - \frac{x_j}{np_j)}\right]^2,$$

or equivalently to

$$\hat{V}(\hat{T}) = \sum_{i=1}^{n} \frac{n}{(n-1)} \left[ \frac{x_i}{np_i} - \frac{\hat{T}}{n} \right]^2 .$$

This demonstrates that the generalized theory for the Horvitz-Thompson estimator and its variance can be extended to with replacement sampling of Hansen and Hurwitz.

## 4. Characterization of Stratification

### 4.1 Explicit Stratification

Explicit stratification partitions the sampling frame. Samples are drawn independently in each stratum. Independent selection implies a zero covariance among pairs of unit sample sizes when each member of the pair is in a different explicit stratum. In terms of the generalized variance expression shown in equation (3), this zero covariance across explicit strata means that a positive variance contribution can arise only from units selected from within the same stratum. Even without adding a subscript for stratum, the generalized variance formula recognizes this feature. This is also true when expressed in terms of the pairwise probabilities rather than in terms of sample size expectations.

**Table 1**: Defining Explicit Strata (Solution 1)

| Unit | Size | Stratum | Stratum Total |
|---|---|---|---|
| 1 | 5 | 1 | |
| 2 | 10 | 1 | |
| 3 | 20 | 1 | |
| 4 | 30 | 1 | 65 |
| 5 | 20 | 2 | |
| 6 | 10 | 2 | |
| 7 | 5 | 2 | 35 |
| Total | 100 | | 100 |

**Table 2**: Defining Explicit Strata (Solution 2)

| Unit | Size | Stratum | Stratum Total |
|---|---|---|---|
| 1 | 5 | 1 | |
| 2 | 10 | 1 | |
| 3 | 20 | 1 | 35 |
| 4 | 30 | 2 | |
| 5 | 20 | 2 | |
| 6 | 10 | 2 | |
| 7 | 5 | 2 | 65 |
| Total | 100 | | 100 |

Deep stratification implies many strata with small sample sizes in each stratum, the extreme being one sampling unit selected per stratum. A frequent compromise which allows for unbiased estimation is to select 2 sampling units per stratum and permit unbiased variance estimation. Both one- and two-unit per stratum schemes suffer from a problem that might be called “hunking”. As an example, consider the population shown in Tables 1 and 2 with 7 sampling units of unequal size which have been ordered on some meaningful variable such as average income. The population can be divided into two strata by setting the stratum boundary between units 4 and 5 (solution 1 in Table 1) or

between units 3 and 4 (solution 2 in Table 2). The best choices result is either a 65:35 split or a 35:65 split. While PPS sampling can be defined within each stratum, the probabilities of selection can not be made close to PPS over all.

### 4.2 Implicit Stratification.

Implicit stratification resolves this problem by splitting unit 4 across two strata as shown in Table 3. Kish (1965, p. 113) refers to this type of solution as zone sampling.

**Table** : Defining Implicit Strata

| Unit | Size | Stratum | Stratum Total |
|---|---|---|---|
| 1 | 5 | 1 | |
| 2 | 10 | 1 | |
| 3 | 20 | 1 | |
| 4 (part A) | 15 | 1 | 50 |
| 4 (part B) | 15 | 2 | |
| 5 | 20 | 2 | |
| 6 | 10 | 2 | |
| 7 | 5 | 2 | 50 |
| Total | 100 | | 100 |

Many PPS designs can then be applied independently within each zone. However, unit 4 can be selected in either one of two zones or under some designs in both zones. PPS systematic sample designs can be viewed as examples of zone sampling. A general strategy followed in many large surveys is to apply explicit strata to define a small number of major strata and then to use a combination of ordering and implicit stratification within the main explicit strata.

## 5. Minimum Replacement Sampling

The basic Horvitz-Thompson estimation theory is defined in terms of specified limits on unit and pairwise probabilities with unit samples sizes limited to 0 and 1.

$$0 < P(u_i) < 1 \text{ and } 0 < P(u_i u_j) < 1 .$$

In terms of the generalized notation, Horvtiz-Thompson theory requires not only that

$$0 < E(n_i) < 1 \text{ and } 0 < E(n_i n_j) < 1,$$

but also that each unit sample size $n_i$ and each product of samples sizes, $n_i n_j$, are both limit to be either 0 or 1.

$$n_i \in \{0,1\}$$

Minimum replacement sampling (Chromy 1981) allows for larger unit sample size expectations, but limits replacement to its minimum, the achieved sample sizes may be greater than 1 but are allowed to vary by no more than 1.

$$n_i \in \{\lfloor E(n_i) \rfloor, \lfloor E(n_i) \rfloor + 1\} .$$

As a special case when $E(n_i) < 1$, PPS without replacement designs result.

Another result of the PMR limitation, is that

$$\Pr[n_i = \lfloor E(n_i) \rfloor + 1] = frac[E(n_i)].$$

Both PPS systematic sampling (Madow 1949) and a sequential PPS sample selection scheme (Chromy 1979, 1981) have the PMR properties. SAS Procedure SURVEYSELECT implements both of these methods in a PMR mode, but does not compute pairwise probabilities or expectations (Sas Institute,Inc. 2004). PMR designs also negate the need to first define self representing sampling units and to treat them as separate strata (*e.g.*, in multi-stage design applications).

### 5.1 PPS Systematic Sampling

Kish (1965, pp.234-7) gives a nice description of PPS systematic sampling as well as some other designs applied in terms of sampling from zones. In equal probability systematic sampling, the possible samples can be viewed as clearly defined clusters where only one cluster of size $n$ is selected[1]. Under the equal probability scenario of systematic sampling, $E(n_i) = 1/k$ for all i and $E(n_i n_j) = 1/k$ for $u_i$ and $u_j$ in the same systematic cluster defined by the sampling interval $k$. $E(n_i n_j) = 0$ otherwise. The pairwise product expectations follow a cyclic pattern.

For PPS systematic sampling, the exact solutions depend on the size measures and their coverage of zones. If the sampling rate is not very high, neighbouring and most nearby pairs will have zero product expectation. Variances are usually estimated under the assumption of effective implicit stratification. Either a successive difference formula or neighbor pairs with implicit strata of size 2 sample units may be used. Usually, the with replacement variance formula are used. That is the formula based on the Hansen Hurwitz with replacement sampling. An approximate finite population factor may be incorporated if the sampling rate is not quite small.

### 5.2 PPS Sequential Sampling

PPS sequential sampling utilizes a serpentine scheme for implicit stratification based on one or more categorical variables and one continuous variable (Williams and Chromy 1980). Viewing the frame as being circular and picking a random starting point is a key to obtaining positive pairwise probabilities. The sequential sampling algorithm achieves a type of zone sampling with PMR. The method is based on conditional probabilities which are functions of the cumulative size measure and the achieved sample size as the ordered frame elements are processed sequentially.

## 6. An Example

This example explores values of the negative covariance term which acts as a coefficient or weight for the squared deviations in the variance expression (equation 3).

$$W_{ij} = \{E(n_i)E(n_j) - E(n_i n_j)\} \ .$$

An artificial population with 50 units and variable size measures was generated by draws the log normal distribution with normal distribution parameters 5 and 1. All expected unit sample sizes were less than 1 when specifying a sample of size 10.

[1] The simplest case is used for discussion purposes where the number of elements in the sampling frame is an exact multiple, $k$, of the sample size, $n$. When this condition does not hold, use of a noninteger interval (same as the general approach for PPS systematic sampling) can be utilized or an integer solution may be used with allowance for the selected sample size to vary by at most 1 element.

Pairwise expectations and covariances were derived over all possible starts and averaged assuming all start points were equally probable. Note that if any of the size measures had been greater than 1, reducing them to their fractional portions would result in the same covariances since variances and covariances are location invariant.

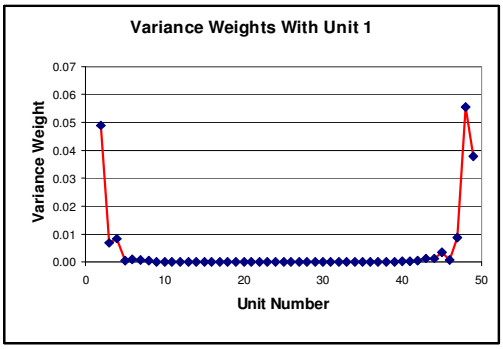


**Figure 1.** Negative Covariances with Unit 1: PPS Sequential Sampling

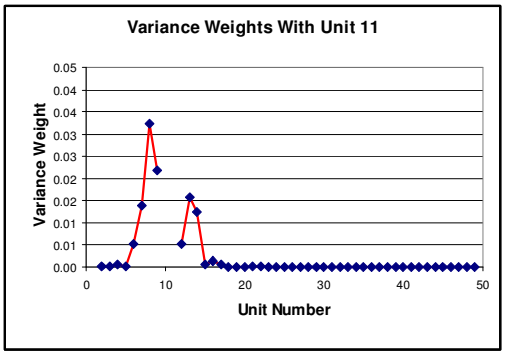


**Figure 2**. Negative Covariances with Unit 11: PPS Sequential Sampling

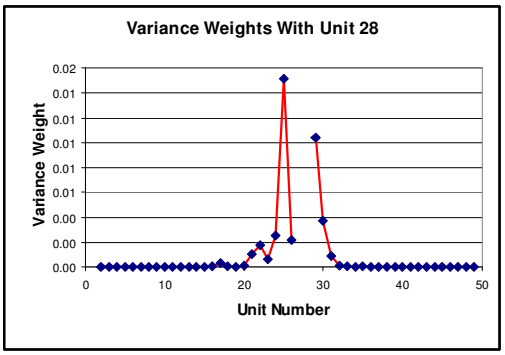


**Figure 3.** Negative Covariances with Unit 28: PPS Sequential Sampling

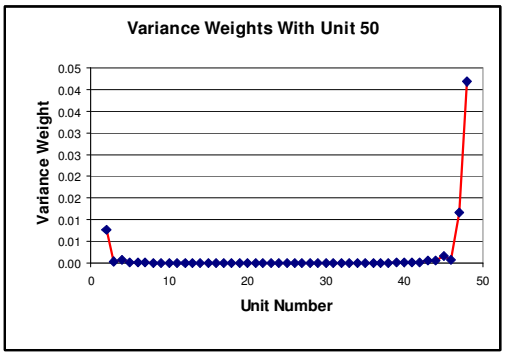


**Figure 4.** Negative Covariances with Unit 50: PPS Sequential Sampling

Figures 1 through 4 show the computed values of variance weights (negative covariances between units) for pairs matched with units 1, 11, 28, and 50 respectively. These figures illustrate that when the units of a pair are close together on the ordered list, their variance weight will be nonzero and, sometimes, quite variable. When the units of a pair are farther apart, their variance weights approach zero. This phenomenon supports the use of a successive difference variance estimator (or the use of pseudo strata of two or three sampling units) as a reasonable approximation to the unbiased variance estimator. If the approximate estimator can be based on more stable weights, it can potentially provide better estimates of the variance with reasonable bias. Typically used approximate estimators utilize the with replacement formula with an additional term for finite population correction. The bias and variance issues associated with approximate variance estimators remain to be explored possibly with additional simulation.

The appropriate variance weights for a PPS systematic sample can also be developed for this example and are shown in Figure 5. The covariances cycle between positive and negative values. Most positive values are associated with units where the pairwise probabilities are zero. For this particular example, 22 of

the pairwise probabilities with unit 1 were zero; the Horvtiz-Thompson variance still applies, but no unbiased estimator is available based on sample data from PPS systematic samples.

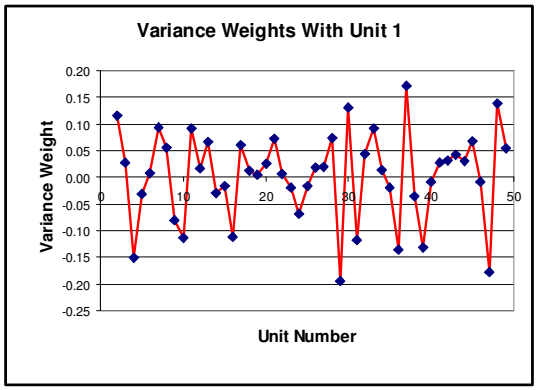


**Figure 5. Negative Covariances with Unit 1: PPS Systematic Sampling**

In summary this example shows that the generalized Horvitz-Thompson theory can provide variance formulations for both PPS systematic and PPS sequential designs. Unbiased variance estimation is possible with the PPS sequential design, but not with the PPS systematic design. Even with the PPS sequential design, the variance estimate can be expected to be quite unstable because of the variation in the variance weights. Plots of the variance weights support a near neighbour variance approximation for PPS sequential based on the actual weights damping off as units get farther apart on the ordered frame. The same phenomenon does not apply to PPS systematic samples.

## 7. Conclusions

The Horvitz-Thompson theory of estimation for PPS sampling can generalized to PPS with replacement and PPS minimum replacement designs. Probability minimum replacement designs allow use of large explicit strata with implicit stratification within these large strata based on a closed ordering. Both PPS sequential and PPS systematic sampling fit the probability minimum replacement definition. Many of the simplifying assumptions used by practitioners to obtain more stable, but biased, estimates of variance appear to be justifiable for PPS sequential sampling. This paper has been limited to the behaviour of the variance weights. More research is needed to examine the behaviour of variance estimators.

## References

Brewer, R. K. W. (1963). "A Model of Systematic Sampling with Unequal Probabilities." Australian Journal of Statistics **5**: 5-13.

Brewer, K. R. W. and M. Hanif (1983). Sampling with Unequal Probabilities. Lecture Notes in Statistics. New York, Springer-Verlag.

Chromy, J. R. (1979). Sequential Sample Selection Methods. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 401-406.

Chromy, J. R. (1981). Variance Estimators for a Sequential Selection Procedure. Current Topics in Survey Sampling. D. Krewski, R. Platek and J. N. K. Rao. New York, Academic Press**:** 329-347.

Chromy, J. R., A. L. Finkner, and D. G. Horvitz. (2004). Survey Design Issues. In The Nation's Report Card: Evolution and Perspectives. L. V. Jones and I. Olkin. Bloomington, IN, Phi Kappa Delta Educational Foundation**:** 384-425.

Horvitz, D. G. and D. J. Thompson (1952). "A generalization of sampling without replacement from a finite universe." The Journal of the American Statistical Association **47**: 663-685.

Kish, L. (1965). Survey Sampling. New York, John Wiley & Sons, Inc.

Sampford, M. R. (1969). A Comparison of Some Possible Methods of Sampling from Smallish Populations with Units of Unequal Size. In New Developments in Survey Sampling. N. L. Johnson and H. Smith Jr. New York, Wiley-Interscience**:** 170-187.

SAS Institute Inc. (2004). SAS/STAT 9.1 User's Guide. Cary, NC, SAS Institute, Inc.

Williams, R. L. and J. R. Chromy (1980). SAS Sample Selection MACROS. Proceedings of the Fifth Annual SAS Users Group International Conference, SAS Institute, pp.392-6.

Yates, F. and P. M. Grundy (1953). “Selection without Replacement from within Strata and with Probability Proportional to Size.” Journal of the Royal Statistical Society **B15**: 253-261.