REPLICATED SAMPLING AND AN APPLICATION TO LIST FRAME SURVEYS

By

FRED A. VOGEL

Sampling Studies Section
Sample Survey Research Branch
Research Division
U.S. Department of Agriculture
Washington, D. C.

January 1975

SF75-09

REPLICATED SAMPLING AND AN APPLICATION TO LIST FRAME SURVEYS

I Introduction

In recent years, SRS has been using the method of replicated sampling to select segments from the area frame. This procedure replaces the method of systematic sampling used in the past. The method of replicated sampling has been around for quite some time, but has only recently been used by SRS. Although the method of replicated sampling has been mainly used by SRS in sampling from area frames 1/, it has also been used to select samples from list frames. It is probable that this method of sampling will become more popular as time goes on.

Therefore, the purpose of this paper is to provide a better understanding of replicated sampling procedures. The principles involved will be defined. These principles will be compared to those involved in simple random sampling and systematic sampling.

An example will illustrate the concepts. Then survey data will be used to compare sampling errors obtained when using systematic selection versus replicated sampling. A replicated sample was selected from the Wyoming list frame for the 1974 Cattle Multiple Frame Survey. The data from that sample will be used to compare the alternate sampling procedures.

1/ Pratt, William L., *The Use of Interpenetrating Sampling in Area Frames*, May 1974.

II Definitions

Before defining replicated sampling, some definitions involving sampling procedures will be given.

A. Simple Random Sampling

Before selection can begin, every member of the population must be assigned a number. The population is usually numbered serially from 1 to N. Then (n) random numbers are selected. The population member whose assigned number matches the random number is then in the sample. Using this procedure, every unit in the population has an equal chance of being selected. A population in this case would be all the members of one stratum if stratified sampling were used. Simple random sampling provides unbiased estimates of means and totals and their sampling errors. It does not require that the list be in any special order before selection begins. However, it becomes cumbersome if large samples are selected.

B. Systematic Sampling

This procedure is most often used in selecting samples because of its simplicity. For example, if one needs to select 20 units out of 100, a sampling interval (100 ÷ 20= 5) is computed. Then a random number between one and the interval is selected. If the random number was 2, the second unit in the list is in the sample. The additional sample units are determined by accumulating the interval

from the random start. Again, using the random number of 2, and an interval of 5, the next units selected are the 7th, 12th, 17th, etc. Note that the units selected depend upon the ordering of the population. The main advantages of systematic sampling lie in the simplicity of the selection and with proper arrangement of the sample units in the frame, the assurance of obtaining a good distribution of the sample over the population. For example, segments can be ordered in the frame in a way to insure a good geographic distribution of the sample. The main disadvantage of systematic sampling is that the sample does not provide an unbiased estimate of the sampling error. This is important because decisions on the prevision of the data and future sample sizes are based on the sampling errors. Estimates of means and totals are unbiased however.

C. Replicated Sampling

Replicated sampling depends on the basic concepts used in both simple random and systematic sampling. The only difference is that more than one sample is selected instead of selecting one sample. For example, instead of selecting one sample of 100 farms using either simple random or systematic sampling, one could select four samples, each containing 25 units. Each sample is a replicate. The term interpenetrating sampling is synonymous with replicated sampling.

A common procedure followed when selecting a replicated sample is to divide the population into groups. This imposes an additional layer of stratification onto the universe. Suppose a sample of size 8 is to be selected from a population 32. If two replicates are to be drawn, each replicate will contain four sample units. Before selecting the four sample units for each replicate, the population is divided into four groups each containing eight elements. Within each group two units will be drawn, one unit for the first replicate, the next unit for the second replicate, etc. This is continued in each of the remaining three groups. Each one of these groups of elements is defined to be a paper stratum. The number of paper strata is the same as the size of each replicate. Again, the selection may be based on either simple random or systematic sampling procedures.

III Purpose of Replicated Sampling

1. Replicated sampling was originally developed to provide an easy way to compute sampling errors. This is important if a computer is not available. Thus, a primary use of replicated sampling is in developing countries that have no means for processing large amounts of data but do need unbiased estimates of the sampling error from their samples.

2. Systematic selection procedures can be used and still obtain unbiased estimates of the sampling errors from the sample.

3. Replicated sampling provides a way to impose an experiment in the sample design. An example was the use of two different questionnaires in the July 1974 Wyoming Cattle Multiple Frame Survey. A replicated sample was selected and replicates were assigned at random to each questionnaire version.

4. Replicated sampling provides a way to select independent samples for sample rotation.

5. It allows one to impose geographic stratification in a size group stratum or geographic substratification within land use strata.

IV Illustration of Sampling Procedures

The different methods of selecting a sample will be illustrated in this section. Table 1 shows a hypothetical population. This population can be considered to be from a single stratum if one is working with stratified samples. This population contains 32 names. We wish to select eight names from this population. Notice that each name has an identification number and a county code. The identification numbers correspond to an alphabetical listing of the names. The ID numbers are from one to 32.

First, let's select a simple random sample. There are 10,518,300 different ways or different possible samples of size 8 that can be selected from 32 using this process. It is not very difficult to select eight random numbers between one and 32.

Table 1--Alphabetical listing of sample units in the sample frame to illustrate simple random and systematic selection procedures

County		ID	Name		County		ID	Name	
10		1	William Adams		10		17	Vera Lurz	
10		2	Renos Allen		50	x	18	Edward Nissen	
20		3	Earl Anderson	*	50	x	19	Ray Person	*
30		4	Eugene Auten		10		20	Gary Ott	
50	x	5	Larry Brim		30		21	Glenn Overy	
20		6	Ben Engel		50	x	22	Alec Reed	
20		7	Willis Forks	*	20		23	Jerry Sayers	*
40		8	Leo Geise		30		24	Frank Schmidt	
40		9	Wayne Hagen		10		25	Norris Scott	
20		10	John Hanner		20		26	Emil Scow	
50	x	11	Keith Henry	*	50	x	27	Fred Shonka	*
40		12	Donald Jones		50	x	28	Claude Smith	
40		13	Floyd Keller		20		29	Max Sommer	
50	x	14	Dale Kime		30		30	Ralph Thoms	
20		15	Melvin Lavik	*	50		31	Don Wheeler	*
20		16	Roy Lowe		10		32	George Willard	

* Name selected using systematic sampling.

x A possible sample using simple random sampling.

However, consider the difficulty of selecting several hundred names from a file containing thousands of names. An important point to consider in selecting a simple random sample is that one does not have to worry about how the list is ordered before selecting the sample. Simple random sampling does not necessarily insure a good geographic distribution of the population. It is just as likely to select the entire sample from one county as it is to represent all counties.

The method of systematic sampling will be illustrated next. Again, we will select a sample of eight from the population of 32. Table 1, with the names in alphabetical order, will be used. The first step is to determine the sampling interval which is 32 divided by 8. Note that the sampling interval is the same as the expansion factor.

$$I = N \div n$$

Then select a random number between one and four. Suppose it is equal to three. Then the first name selected is the third in order. The next one selected is no. 7 (random number + interval). Our sample then becomes the third, seventh, eleventh, fifteenth, etc. There are two important points to consider:

(1) The method of systematic sampling does not provide the randomness involved with simple random sampling. There are only four different combinations of samples of size 8 that can be selected from this population. While this does not affect the probability of selection, it does affect the computation of the sampling errors. Four

possible samples compared to over 10 million possible samples do not provide a sampling distribution for systematic sampling.

(2) Systematic sampling does require that one pay attention to the ordering of the population. Consider the systematic sample we have selected from an alphabetical listing of names. Note that the entire sample was selected from only two counties.

Suppose a geographic distribution is desirable. Table 2 shows the names arranged alphabetically within counties. All names within county 10 are in alphabetical order followed by all names in county 20 and so on.

Now select another systematic sample using the same random starting point and interval used before. Again, select the third name in order followed by the seventh name in order, and so on down the line. Note how this sample is distributed. At least one name is selected from every county in the population. One should be especially concerned about the ordering of the population if the size of operation differs between counties. Using the first method of ordering the list, it is possible that the county selected represented all small farms, while the counties missed represented the large farms. The ordering of sampling units before selection has probably not been given the consideration it needs in list preparation.

Table 2--Sample units listed alphabetically within counties to illustrate systematic sampling procedures

County	ID	Name		County	ID	Name	
10	1	William Adams		30	21	Glenn Overy	
10	2	Renos Allen		30	24	Frank Schmidt	
10	17	Vera Lurz	*	30	30	Ralph Thoms	*
10	20	Gary Ott		40	8	Leo Geise	
10	25	Norris Scott		40	9	Wayne Hagen	
10	32	George Willard		40	12	Donald Jones	
20	3	Earl Anderson	*	40	13	Floyd Keller	*
20	6	Ben Engel		50	5	Larry Brim	
20	7	Willis Forks		50	11	Keith Henry	
20	10	John Hanner		50	14	Dale Kime	
20	15	Melvin Lavik	*	50	18	Edward Nissen	*
20	16	Roy Lowe		50	19	Ray Person	
20	23	Jerry Sayers		50	22	Alec Reed	
20	26	Emil Scow		50	27	Fred Shonka	
20	29	Max Sommer	*	50	28	Claude Smith	*
30	4	Eugene Auten		50	31	Don Wheeler	

Now we will use the same population and illustrate how a sample of eight can be selected using the method of replicated sampling. First look at Tables 1 and 2 again. Our sampling interval was equal to four when we selected the systematic sample. This in effect divided the population of 32 names into groups of four. The method of systematic sampling in effect selected one name from within each group. Each group is defined to be a paper stratum. If we wanted to select a sample of 16 units from this population, we could draw another sample of eight by selecting a random number between one and four and repeating the systematic process. This will provide two independent samples of size 8. The samples added together yield a sample size of 16.

However, to select a sample of 16, we were not restricted to using systematic sampling. We could have selected two sample units at random by drawing two random numbers within each paper stratum.

Next to be illustrated is how to select a replicated sample of size 8 from the population of 32 names. Table 3 shows our population of 32 names sorted alphabetically within counties. Instead of selecting one sample of size 8, we will select two independent samples, each containing four units. This can be done as follows:

(a) Compute the interval $(32 \div 4) = 8$

(b) Select two random numbers between one and eight. Say they were three and eight.

(c) Using the first random number, sample units for

replicate no. 1 are the 3rd, 11th, 19th, and 27th. Using the second random number, sample units for replicate no. 2 are the 8th, 16th, 24th, and 32nd.

Note how the sample is distributed across the list. Every county is represented at least once. Since each sample contained four units, we in effect divided our 32 names into four equal parts (or four paper strata) and selected two sample units from each. If the names were alphabetized within counties all names from one county or adjacent counties fell into one paper stratum. Note that the names were alphabetized by county, followed by names from the adjacent geographical county. Lines are drawn under every eighth time to divided the population into four groups or paper strata. Another way to select the sample would be to select two random numbers within each paper stratum.

The main advantage of replicated sampling is that it allows one to use the easier method of systematic sampling to maintain some control over the distribution of the population but yet obtain an unbiased estimate of the sampling errors.

Table 3--Sample units listed alphabetically within counties to illustrate replicated sampling procedures

County	ID	Name		County	ID	Name	
10	1	William Adams		30	21	Glenn Overy	
10	2	Renos Allen		30	24	Frank Schmidt	
10	17	Vera Lurz	1	30	30	Ralph Thoms	1
10	20	Gary Ott		40	8	Leo Geise	
10	25	Norris Scott		40	9	Wayne Hagen	
10	32	George Willard		40	12	Donald Jones	
20	3	Earl Anderson		40	13	Floyd Keller	
20	6	Ben Engel	2	50	5	Larry Brim	2
20	7	Willis Forks		50	11	Keith Henry	
20	10	John Hanner		50	14	Dale Kime	
20	15	Melvin Lavik	1	50	18	Edward Nissen	1
20	16	Roy Lowe		50	19	Ray Person	
20	23	Jerry Sayers		50	22	Alec Reed	
20	26	Emil Scow		50	27	Fred Shonka	
20	29	Max Sommer		50	28	Claude Smith	
30	4	Eugene Auten	2	50	31	Don Wheeler	2

V Comparison of Sample Procedures Using Wyoming Cattle Multiple Frame Data

A. Sample Design

The sample from the list frame for the 1974 Cattle Multiple Frame Survey in Wyoming was selected using the replicated sampling procedure. The primary purpose for replicating the sample was to provide a way to test results from two different questionnaire versions used in the survey.

Table 4 shows the sample allocation by stratum. It also shows the number of replicates selected from each stratum along with the number of paper strata. Remember the number of paper strata is the same as the size of each replicate.

Table 4--Allocation of the sample to list stratum by replicated method of sampling, Wyoming Cattle and Calf Multiple Frame Survey, June 1974

Stratum	Number replications	Number paper strata	Total number names in each paper stratum	Number selected n	Number in size group
	r_i	k_i	R_i	n_i	N_i
No cattle	4	3	145	12	436
1-99	16	28	146	448	4,103
100-199	10	25	42	250	1,058
200-299	10	20	24	200	487
300-499	10	20	21	200	431
500+	10	15	20	150	297

Notation is explained in the Appendix.

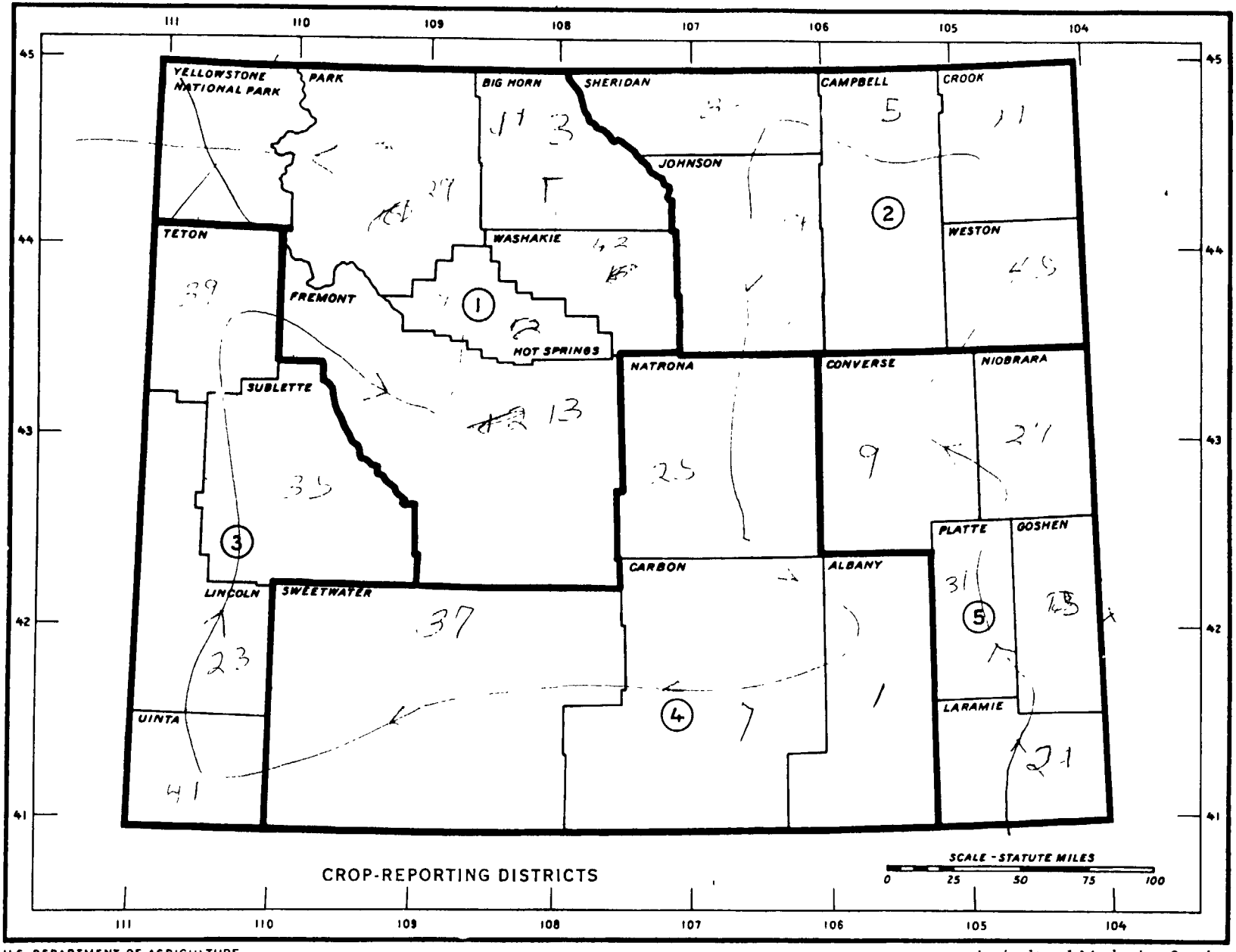
The determination of the number of replicates and their size was done somewhat arbitrarily. The decision to have more paper strata than replicates was based on the assumption that the maximum amount of geographic distribution was desired. The allocation could have required considerably more replicates than paper strata - this would decrease the geographic distribution.

Before selecting the sample, the list frame was arranged as follows:

a) The entire frame was alphabetized.

b) The alphabetical listing was sorted into county order beginning with Laramie County and ending with Park County. Figure A shows the order in which the counties were arranged. Thus, all names in Laramie County were followed by names in Goshen County.

WYOMING



CROP-REPORTING DISTRICTS

U S DEPARTMENT OF AGRICULTURE

Agricultural Marketing Service

Each name contained a measure of size indicating its stratum designation. Within each size group stratum, replicates were selected using systematic sampling. Thus, in the stratum for size group 1-99, 16 systematic samples were selected. This in effect meant there were 28 paper strata each containing 146 names. The first paper stratum contained names in the 1-99 size group and in Laramie County. If Laramie County did not contain 145 names in size group 1-99, then the paper strata also contained some names from Goshen County. The arrangement of the names in county order meant the sample would contain names distributed across the State. Remember that if the fifth unit in paper stratum no. 1 was selected, the fifth unit in every paper stratum was selected.

Note the following relationships from the table:

a. Number replicates times number paper strata = Sample size
b. Total number in stratum divided by number paper strata = Size of paper stratum.

Note also that the size of each paper stratum varies by size group.

B. Results

The survey data was used to calculate relative sampling errors based on the replicated method of sampling. To compare sampling errors with those arising from simple random sampling, sample errors were recomputed assuming that method was used. That does not provide an unbiased estimate of the sample errors. However, the sample sizes were large enough that the amount of bias should

be small. A third way was also used to calculate the relative sampling errors. This was done assuming a simple random sample was selected from within each paper stratum. Again, this is not an unbiased estimate of the sampling error but the bias should be small because of the large sample size. The formulation used in each of these calculations is shown in the Appendix.

Table 5 shows a comparison of the relative sampling errors based on each method of computation. Sampling errors computed based on the replicates generally resulted in the larger sampling error. This indicates two factors:

(1) The method of variance computation used for the systematic selection with the geographic stratification in the list frame is based on formulation used for simple random sampling. This method should have resulted in an overstatement of the relative sampling errors. Since it did not, it appears that geographic stratification had little effect on the sample errors. Most benefits were gained through the size group stratification rather than the geographic ordering of the list.

(2) More replicates should have been selected and fewer paper strata identified. This is supported by analysis done by Pratt on the Nebraska area frame which indicated that the number of replicates should be about the same as the number of paper strata.

Table 5--Sample errors for selected items by method of estimation, Wyoming Cattle and Calf Multiple Frame Survey, June 1974 1/

Item	Direct expansion	Relative method simple random sampling	Sampling errors of estimation	
			By replication	By paper strata
	(000)	(%)	(%)	(%)
Total calves born since Jan. 1	552.4	2.6	3.1	2.6
Cows and heifers expected to calve	33.5	8.4	8.4	8.4
Cows weighing less than 500 lbs.	544.4	2.6	3.2	2.6
Total cattle and calves	1,494.5	2.4	2.8	2.3

Estimates do not include extreme operators nor the nonoverlap domain.

VI Summary

There is always considerable concern about whether there is an unbiased estimate of the total inventory. Perhaps as much concern should be directed to the availability of an unbiased estimate of the sampling errors. They are used to gauge the reliability of the sample results and to determine an optimum sample size. Therefore, it is suggested additional work be done to determine how to optimize a sample allocation for replicated sampling.

SUMMARY PROCEDURES - WYOMING CATTLE MULTIPLE FRAME

n_i = Number of sample units in i^{th} stratum

r_i = Number of replications in i^{th} stratum

k_i = Number of paper strata in i^{th} stratum

= is also the number of sample units in each replication

N_i = Total number of sampling units in the i^{th} stratum

There is also the following relationship between the variables:

$n_i = r_i \times k_i$

$R_i = \frac{N_i}{K_i}$ Is the total number of replications that could be selected from the k paper strata.

This is also the expansion factor if only one replicate is to be used.

$\frac{N_i}{n_i} = \frac{R_i K_i}{r_i k_i} = \frac{R_i}{r_i}$ is the expansion factor if r_i replications are used.

Each survey item for a sample unit should be designated by:

X_ijm

$i = 1, 2 \ldots S$ Strata

$j = 1, 2 \ldots k_i$ Paper stratum in i^{th} stratum

$m = 1, 2 \ldots r_i$ Replicate in i^{th} stratum

Then $\hat{X}_ijm = \frac{R_i}{r_i} X_ijm = \frac{N_i}{n_i} X_ijm$ which is the expanded survey item.

The direct expansion for the survey item in the i^{th} stratum is

$$\hat{X}_{i..} = \sum_{j}^{k_i} \sum_{m}^{r_i} \hat{X}_{ijm} \quad \text{and}$$

$$\hat{X}_{i.m} = \sum_{j}^{k_i} X_{ijm} \quad \text{is the expanded total of the } m^{th} \text{ replication and}$$

$$\hat{\bar{X}}_{i..} = \frac{\sum_{j}^{k_i} \sum_{m}^{r_i} \hat{X}_{ijm}}{r_i} \quad \text{is the average expansion per replicate.}$$

$$\hat{\bar{X}}_{ij\bullet} = \frac{\sum_{m=1}^{r_i} \hat{X}_{ijm}}{r_i} = \text{is the mean per expanded segment total in the } j^{th} \text{ paper stratum in the } i^{th} \text{ stratum}$$

The computation of the variances can be explained by considering each replicate total as one observation. The sampling variance is measured by the variability between the replicates.

The variance of $\hat{X}_i$ can be computed by

$$\text{Var } \hat{X}_i = \frac{r_i}{r_i - 1} \sum_{m}^{r_i} (\hat{X}_{i.m} - \hat{\bar{X}}_{i..})^2$$

The direct expansion over the State is obtained by adding over strata

$$X = \sum_{i}^{S} \hat{X}_i$$

$$\text{Var } \hat{X} = \sum_{i}^{S} \text{Var } \hat{X}_i$$

Another way to obtain the sampling variance is to measure the variability between observations in each paper stratum.

$$\text{Var } \hat{X}_i \text{ (paper strata)} = \frac{r_i}{r_i - 1} \sum_{j}^{K_i} \sum_{m}^{r_i} (\hat{X}_{ijm} - \bar{\hat{X}}_{ij.})^2$$