COMPARISON OF SAMPLE DESIGNS FOR A POPULATION OF FARMS

Earl E. Houseman

U.S. Department of Agriculture
Economics, Statistics, and Cooperatives Service

ESCS-35

February, 1979

BIBLIOGRAPHIC DATA SHEET	1. Report No. ESCS-35	2.	3. Recipient's Accession No.

4. Title and Subtitle
COMPARISON OF SAMPLE DESIGNS FOR A POPULATION OF FARMS

5. Report Date
February 1979

6.

7. Author(s)
Earl E. Houseman

8. Performing Organization Rept. No. ESCS-35

9. Performing Organization Name and Address
Statistical Research Division
Economics, Statistics, and Cooperatives Service
U.S. Department of Agriculture
Washington, D.C. 20250

10. Project/Task/Work Unit No.

11. Contract/Grant No.

12. Sponsoring Organization Name and Address

13. Type of Report & Period Covered
Final

14.

15. Supplementary Notes

16. Abstracts

This report serves as supplementary training in sampling for statisticians. It compares sampling errors for alternative sampling plans. Some comparisons of interest are the relationship between sampling errors for various commodities and proportions of farms producing commodities, the efficiency of the minor civil division in the sampling unit compared to the individual farm, the efficiency of geographic stratification related to the number of strata, and the efficiency of alternate allocations of a sample to strata.

17. Key Words and Document Analysis. **17a. Descriptors**

Comparison
Farms
Population (statistics)
Sampling
Sampling theory
Statistical analysis

17b. Identifiers/Open-Ended Terms

Commodities
Geographic stratification
Sampling design
Sampling errors
Sampling unit
Strata

17c. COSATI Field Group 02-B, 12-A, 12-B

18. Availability Statement Available from:
NATIONAL TECHNICAL INFORMATION SERVICE, 5285 Port Royal Road, Springfield, Virginia 22161.

19. Security Class (This Report)
UNCLASSIFIED

20. Security Class (This Page
UNCLASSIFIED

21. No. of Pages

22. Price

CONTENTS

Washington, D.C. 20250 March 1979

COMPARISON OF SAMPLE DESIGNS FOR A POPULATION OF FARMS

Earl E. Houseman *

INTRODUCTION

Data from an annual farm census conducted by the State of Wisconsin for 1970 and 1971 were used to obtain variances for a study of many alternative sampling plans. Sixteen characteristics were selected for study with regard to patterns of distribution over the State and the proportion of farms reporting. "Number of farms reporting" is commonly used when referring to the farms that are producing a particular commodity or reporting a positive nonzero answer to a question. Thus, if $Y_1, \ldots, Y_N$ are the values of some characteristic Y for all farms in the population, the "number of farms reporting" is the number of farms for which $Y_i > 0$.

Wisconsin is a good State for purposes of this study, especially because of its wide variation in agriculture from north to south. The State is divided into nine crop reporting districts (CRDs) (fig. 1). CRDs are State subdivisions used for statistical purposes, and are generally made up of homogeneous groups of counties.

Summary data for 1970 and 1971, which include crops grown and total number of farms in the State, were derived directly from the original data (tables 1 and 2). No adjustments have been made for undereunumeration or overenumeration, definitions, or other factors. The totals and averages as shown in columns 2 and 3, for example, are not official estimates. Numbers in parentheses in the table columns correspond to algebraic descriptions in the appendix. A column that appears in more than one table always has the same number. Likewise, corresponding data for two different years also have the same column number.

Note that the proportions of farms reporting (column 7), range from less than 1 percent for potatoes and snap beans to 100 percent for farmland. In fact, the characteristics were ordered according to the proportion reporting. Population (number of persons living on farms) was included because the variation among farms is low, is reported by nearly all farms, and is distributed geographically more or less in proportion to numbers of farms. Some characteristics are more uniformly distributed than others (table 3).

*The author is statistical consultant, retired from U.S. Department of Agriculture.

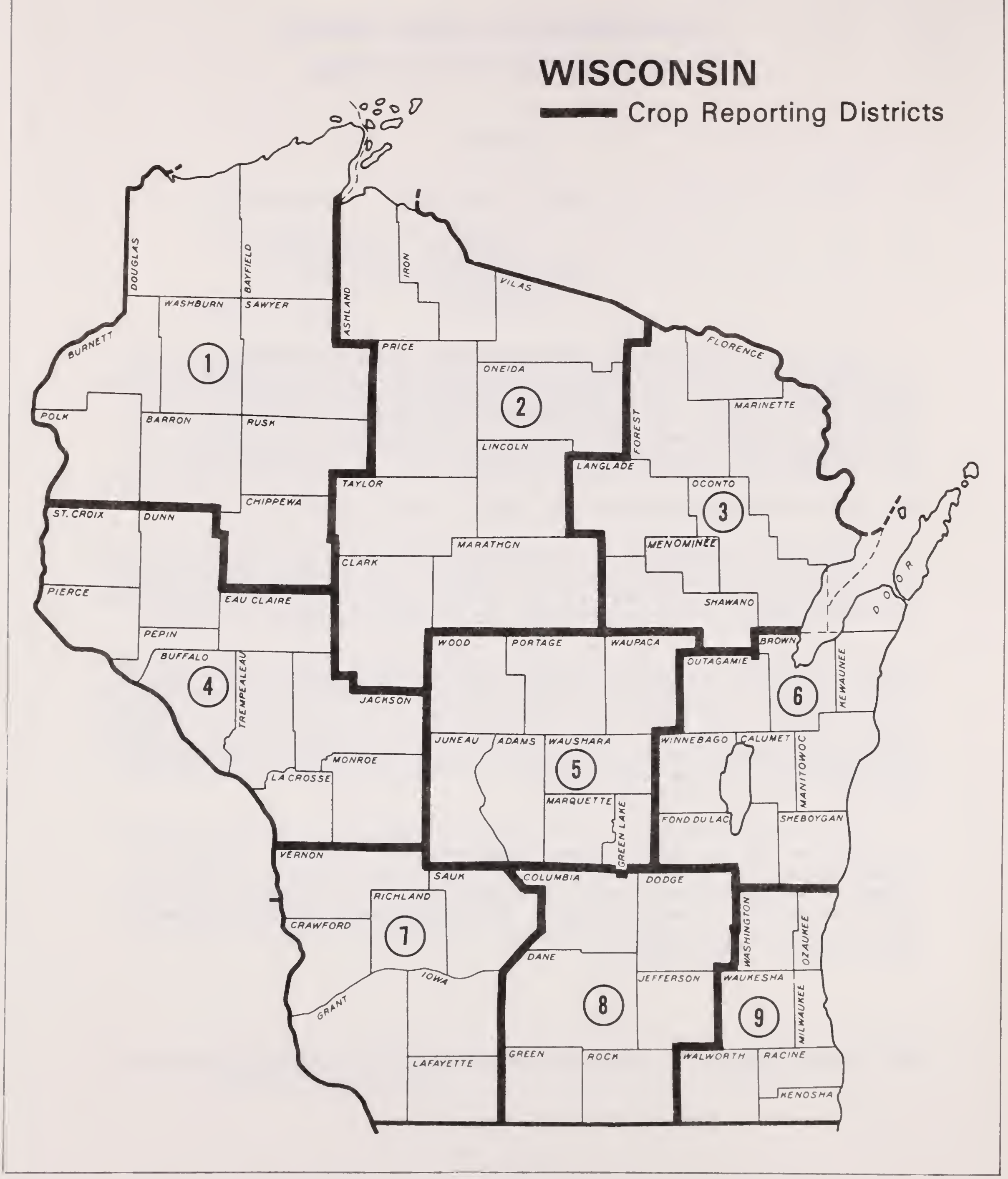
WISCONSIN
Crop Reporting Districts
DOUGLAS
BAYFIELD
ASHLAND
IRON
VILAS
WASHBURN
SAWYER
BURNETT
PRICE
FLORENCE
ONEIDA
1
2
POLK
BARRON
RUSK
MARINETTE
FOREST
LINCOLN
LANGLADE
TAYLOR
OCONTO
CHIPPEWA
ST. CROIX
DUNN
3
MARATHON
MENOMINEE
CLARK
DOOR
PIERCE
EAU CLAIRE
SHAWANO
PEPIN
BROWN
WOOD
PORTAGE
WAUPACA
BUFFALO
OUTAGAMIE
KEWAUNEE
4
TREMPEALEAU
6
JACKSON
JUNEAU
ADAMS
WAUSHARA
WINNEBAGO
CALUMET
MANITOWOC
MONROE
5
LA CROSSE
MARQUETTE
GREEN LAKE
FOND DU LAC
SHEBOYGAN
VERNON
SAUK
COLUMBIA
DODGE
RICHLAND
WASHINGTON
OZAUKEE
CRAWFORD
7
DANE
IOWA
JEFFERSON
WAUKESHA
GRANT
8
9
MILWAUKEE
GREEN
ROCK
WALWORTH
RACINE
LAFAYETTE
KENOSHA

Figure 1

Table 1--State summary, 1970

Characteristic	Unit	Total	All farms			Farms reporting				
			Average per farm	Standard deviation per farm	Relative variance per farm	Number	Proportion	Average per farm	Standard deviation per farm	Relative variance per farm
(1) 1/		(thousands) (2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Farmland	Acres	17,825	175.3	149.3	0.725	101,685	1.000	175.3	149.3	0.725
Population	No.	424	4.17	2.79	0.450	96,428	.948	4.39	2.69	0.372
Alfalfa	Acres	2,907	28.6	32.3	1.277	71,434	.702	40.7	31.5	0.593
All corn	do.	2,612	25.7	46.8	3.33	67,573	.665	38.6	52.9	1.88
All pasture	Acres	3,316	32.6	56.9	3.04	62,401	.614	53.1	64.7	1.49
Milk cows	No.	1,657	16.3	18.9	1.34	59,728	.587	27.7	17.0	0.37
Beef cattle	do.	596	5.86	23.5	16.07	26,895	.264	22.2	41.5	3.50
Clover & timothy	Acres	569	5.60	22.2	15.76	19,294	.190	29.5	43.5	2.19
Hay for silage	do.	552	5.42	19.7	13.20	16,169	.159	34.1	38.3	1.25
Cattle marketed	No.	188	1.85	23.7	163.90	7,600	.075	24.8	83.4	11.36
Soybeans	Acres	127	1.25	10.7	73.4	4,125	.041	30.7	43.6	1.99
Peas	do.	102	1.00	11.6	134.2	3,180	.031	32.0	57.4	3.24
Sheep	No.	77.7	0.76	9.0	138.2	2,742	.027	28.3	47.0	2.76
Spring wheat	Acres	15.3	0.15	3.3	488	1,194	.012	12.8	27.9	4.74
Potatoes	do.	40.1	0.39	11.1	789	741	.0073	54.1	188.0	4.76
Snap beans	do.	6.1	0.06	2.3	1501	234	.0023	25.9	40.8	2.45

1/ Numbers in parentheses correspond to algebraic descriptions in the appendix.

Table 2--State summary, 1971

Characteristic	Unit	Total (thousands)	All farms			Farms reporting				
			Average per farm	Standard deviation per farm	Relative variance per farm	Number	Proportion	Average per farm	Standard deviation per farm	Relative variance per farm
(1) 1/		(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Farmland	Acres	17,637	179.7	155.6	0.750	98,156	1.000	179.7	155.6	0.750
Population	No.	406	4.13	2.85	.476	92,669	.944	4.38	2.75	0.394
Alfalfa	Acres	3,016	30.7	35.1	1.306	70,196	.715	43.0	34.6	0.649
All corn	do.	3,102	31.6	55.1	3.04	75,068	.765	41.3	59.7	2.09
All pasture	do.	3,085	31.4	57.7	3.37	60,156	.613	51.3	66.4	1.68
Milk cows	No.	1,658	16.9	20.6	1.48	57,335	.584	28.9	19.4	0.45
Beef cattle	do.	631	6.43	24.9	15.06	28,157	.287	22.4	42.5	3.61
Clover & timothy	Acres	542	5.53	15.8	8.21	18,273	.186	29.7	25.1	0.71
Hay for silage	do.	673	6.86	23.1	11.35	17,124	.175	39.3	42.2	1.15
Cattle marketed	No.	185	1.89	27.5	212.6	7,109	.072	26.1	99.1	14.44
Soybeans	Acres	104	1.06	10.5	97.3	3,219	.033	32.3	48.2	2.23
Peas	do.	101	1.03	15.6	230.8	3,123	.032	32.3	81.5	6.36
Sheep	No.	71.4	0.728	8.12	124.6	2,564	.026	27.9	42.1	2.27
Spring wheat	Acres	17.4	0.177	2.45	190.6	1,299	.013	13.4	16.6	1.54
Potatoes	do.	42.3	0.431	12.3	809	669	.0068	63.2	134.5	4.53
Snap beans	do.	39.7	0.404	12.7	993	544	.0055	72.9	155.0	4.52

1/ Numbers in parentheses correspond to algebraic descriptions in the appendix.

Table 3--Percent distribution of each characteristic by crop reporting district (1970)

Characteristic	Crop reporting district									Total
	1	2	3	4	5	6	7	8	9	
					Percent					
Number of farms	10.6	11.0	5.8	15.1	9.5	14.9	13.4	14.1	5.7	100.1
Farmland	11.0	10.5	5.6	17.2	9.5	12.0	15.9	13.3	5.0	100.0
Farm population	9.9	11.0	5.6	14.8	8.6	15.7	13.6	14.9	5.9	100.0
Alfalfa	8.1	3.6	5.2	18.3	7.3	18.7	18.9	14.5	5.4	100.0
All corn	5.1	3.8	3.3	13.8	7.9	14.4	16.8	25.7	9.1	99.9
All pasture	16.6	13.4	3.3	17.3	6.7	5.8	24.3	10.3	2.2	99.9
Milk cows	9.7	11.7	5.5	15.0	6.7	17.0	14.6	14.7	5.1	100.0
Beef cattle	8.2	4.4	2.7	17.8	6.7	6.2	28.1	21.6	4.3	100.0
Clover & timothy	23.6	41.2	8.1	8.1	9.9	4.3	1.7	1.3	1.8	100.0
Hay for silage	8.9	9.5	4.8	14.8	6.6	16.6	16.4	18.2	4.2	100.0
Cattle marketed	1.9	0.9	0.7	7.8	7.3	9.0	23.5	40.7	8.2	100.0
Soybeans	4.0	0.4	0.0	25.0	3.8	7.8	3.8	26.3	29.0	100.1
Peas	2.3	1.1	2.6	4.9	7.4	34.1	3.4	37.2	7.0	100.0
Sheep	9.1	2.5	1.5	17.9	6.4	5.3	23.9	26.9	6.4	99.9
Spring wheat	3.2	1.2	0.9	4.2	10.5	17.4	2.2	10.1	50.3	100.0
Potatoes	4.3	13.8	19.5	1.7	42.0	3.0	1.6	2.6	11.4	99.9
Snap beans	0.0	1.3	3.5	0.5	1.5	30.4	0.4	9.2	53.2	100.0

Column (4) presents standard deviations among all farms and column (5) contains the relative variances, which may be interpreted as relative variances of a sample mean for samples of size 1. Note the inverse relation between the proportion reporting, column (7), and the relative variance, column (5). For a sample of a given size, it points to major differences in the coefficients of variation (c's of v) of sample estimates for characteristics depending on the proportion of farms reporting.

Exercise 1. Suppose a simple random sample of farms is to be selected and that the desired c of v of estimates is 2 percent. Using the 1970 data in table 1, find the sample sizes for farm population and snap beans so that the c of v of ȳ will be 2 percent, where ȳ is a simple average of the values of Y in the sample. Answer: 1,125 *and* 99,002. *The answer,* 1,125, *was obtained without using the finite population correction (fpc) since the sampling fraction is only about 1 percent. For snap beans, the fpc must be used.*

Exercise 2. With reference to exercise 1, suppose the c of v for snap beans is set at 10 percent instead of 2 percent. How large must the sample be? Answer: 60,619, *which means a sampling fraction of* 60 *percent. In a simple random sample of* 60,619 *farms, what is the expected number of growers of snap beans and what is the expected number of farms reporting farm population? Answer:* 139 *and* 57,485.

The answers to the above exercise clearly point to a sampling problem often referred to as the problem of sampling for "rare items." In the absence of special techniques to identify growers of a particular commodity prior to sampling, it might appear that a census is necessary because the required sampling fractions are very large. However, there have been strong tendencies for rare items to be incompletely enumerated unless special precautions are taken. For example, suppose that in the section of a questionnaire on crop acreages, separate questions are asked for all leading crops. Then, an "all other" question is asked to get the names and acreages of any remaining crops. Crops in the "other" category might be underenumerated by a substantial fraction. To reduce total error, efforts to reduce response error may be more important than spending additional resources on a complete census.

The mathematical relationship between columns (5) and (10) is very useful, namely,

$$V^2 \doteq \frac{V_r^2 + 1-P}{P} \qquad (1)$$

where $V^2 = \dfrac{\sum^{N}(Y_i - \bar{Y})^2}{\bar{Y}^2(N-1)}$ is the relative variance among all farms, column (5).

$V_r^2 = \dfrac{\sum^{N_r}(Y_{ri} - Y_r)}{\bar{Y}^2(N_r - 1)}$ is the relative variance among farms reporting, column (10),

N_r = is the number of farms reporting, column (6),

$P = \dfrac{N_r}{N}$ is the proportion of farms reporting, column (7),

$$\bar{Y} = \frac{\overset{N}{\Sigma}\, Y_i}{N}$$ is the average for all farms, column (3), and

$$\bar{Y} = \frac{\overset{N_r}{\Sigma}\, Y_{ri}}{N_r}$$ is the average per farm reporting, column (8)

The subscript "r" is used in reference to a subset of farms reporting, that is, the farms for which $Y_i > 0$.

Exercise 3. Equation 1 is exact if the population variances are defined by dividing sums of squares by N *and* N_r *instead of* N-1 *and* $N_r - 1$. *Show that this is true.*

Notice in tables 1 and 2 that the range of variation in V_r^2, column (10), is small compared to the variation in the values of V^2, column (5). Also note that V_r^2 is not related in a definitive way to P. The characteristics which have the largest values of V_r^2 probably have frequency distributions with a high degree of skewness; that is, a relatively small number of farms having the largest values of Y_i probably account for a substantial part of the total of Y.

Good sampling practice would call for trying to identify (prior to sampling) farms with extremely large values of Y_i and including all (or a large fraction of) such farms in the sample. If farms with large values of Y_i are identified and put in a separate stratum that is completely enumerated, the V_r^2 for the part of the population sampled will tend to be smaller and contained within rather narrow limits. In any event, equation 1 is often an important aid in forming prior judgments of sampling variances and in developing techniques for approximating sampling errors pertaining to the numerous estimates that might be produced from a sample survey.

In planning surveys, it is often helpful to have rough approximations of sampling variances available without delay. An experienced sampler can make good guesses at the values of V_r and P, and from equation 1 can make a good judgment of the magnitude of V and hence the magnitude of the sampling error for any size sample that might be under consideration.

For simple random sampling, ignoring the fpc, it follows from equation 1 that:

$$V^2\,(\bar{y}) \doteq \frac{V^2}{n} = \frac{V_r^2 + 1\text{-}P}{nP} \qquad (2)$$

where $V^2(\bar{y})$ is the relative variance of the sample mean, $\bar{y}$, which is an estimate of $\bar{Y}$. One might add a factor for design efficiency to equation 2. That is, if one judged the efficiency (variance) of the sampling plan under consideration, for example, to be 0.6 or 1.2 times the variance for simple random sampling, one could adjust $\frac{V^2}{n}$ accordingly. Of course, one's ability to make prior judgments of sampling error improves with experience and knowledge of information about variance. Even for characteristics that have not been included in a previous survey, conjecture about sampling error can provide a good indication of whether the sampling standard error of an estimate (or class of estimates) is, for example, the order of 7 or 8 percent or perhaps 3 or 4 percent. Past information and conjecture about sampling error should play or be developed to play an important role in determining the size of sample for a survey and the content of a questionnaire, and the extent of domain estimation (breakdown of the data) that any given sampling plan is likely to satisfactorily support. 1/

Equation 2, or a similar equation, is sometimes helpful in developing a method for approximating sampling errors of estimates from a sample survey. This is in lieu of computing variances for all estimates according to an exact formula for the particular sampling design involved. It should be pointed out that for simple random sampling, equation 2 extends easily to domain estimates. Suppose that $\bar{y}_d$ is an estimate of the domain mean $\bar{Y}_d$, where $\bar{Y}_d$ is the domain total divided by the total number of farms (elements) in the domain. Instead of equation 2, we have:

$$V^2\,(\bar{y}_d) = \frac{V_d^2}{n} = \frac{V_{dr}^2 + 1-P_d}{nP_d} \qquad (3)$$

where $V^2\,(\bar{y}_d)$ is the relative variance of $\bar{y}_d$,

V_{dr}^2 is the relative variance among nonzero values of Y within the domain,

and $P_d = \frac{N_{dr}}{N}$ is the number of farms in the domain with nonzero values of Y divided by the total number of farms in the entire population.

Thus, nP_d (the expected number of nonzero values of Y in the sample and in the domain) is a major factor determining the relative variance of $\bar{y}_d$.

Exercise 4. With reference to equation 2, nP *is the expected number of farms reporting in a sample of* n *farms. Study equation 2 with this in mind and with reference to data presented in tables 1 and 2. Does it appear that most of the differences in sampling variances for various commodities is explained by variation in* nP?

Note that $\bar{y}$ may be regarded as the product of two random variables, p and $\bar{y}_r$, where $p = \frac{n_r}{n}$ is the proportion reporting in a random sample of n and

$\bar{y}_r = \frac{\Sigma^{n_r} y_{ri}}{n_r}$ is the average of y for the n_r farms reporting. It follows from equation 2 that:

$$V^2(\bar{y}) = V^2(p\bar{y}_r) = \frac{V_r^2}{Pn} + \frac{(1-P)}{Pn} \qquad (4)$$

1/ Houseman, Earl E. "The Survey as a Measurement Instrument," Agricultural Economics Research, U.S. Dept. Agr., ERS, Vol. 24, No. 4, October 1972, p. 87.

Let $Pn = \bar{n}_r$ and $V_p^2 = \frac{1-P}{P}$. This gives:

$$V^2(\bar{y}) = \frac{V_r^2}{\bar{n}_r} + \frac{V_P^2}{n} \qquad (5)$$

which provides a basis for determining how much of the variance of $\bar{y}$ is associated with the variance of p and how much is associated with the variance of $\bar{y}_r$. Note that $\bar{n}_r$ is the expected value of n_r for samples of n farms and V_p^2 is the relative variance of p for n=1.

The comparative magnitudes of the two components of variance shown in equation 5 have implications on matters of sample design and estimation including the possibility of double sampling, that is, using a large sample to estimate the values of p, or N_r, for various characteristics, and subsamples of the large sample to estimate the values of $\bar{y}_r$.

Exercise 5. *From the results presented in table 1, determine the relative variance of $\bar{y}$ for soybeans assuming a simple random sample of* 1,000 *farms. Answer:* 0.0734. *Find the values of the two components given in equation 5. As a check, they should add to* 0.0734, *except for rounding errors.*

Exercise 6. *An estimator of* N_r, *the total number of farms reporting, is* Np *where* N *is the total number of farms in the population and* p *is the proportion of farms reporting in a sample of size* n. *Assuming* n = 10,000, *find the standard error of* $N\bar{y}$ *and of* Np *for potatoes using the 1970 data. Answer: The standard error of* $N\bar{y}$ *is* 10,718 *acres or* 26.7 *percent. The standard error of* Np *is* 82.2 *farms or* 11.1 *percent.*

Known Numbers of Farms Reporting

Since the c's of v are large for characteristics where P is small, consideration of all possible ways of reducing the sampling variance for such characteristics is important. How valuable would information on the number of farms reporting be in reducing sampling variance? This is part of the general question on value and cost of auxiliary data that might be incorporated in a sampling frame.

Suppose the numbers of farms reporting, N_r in column (6), are known. Then, one could use $N_r\bar{y}_r$ as an estimator of a population total. Assuming a simple random sample of all farms, how does the variance of $N_r\bar{y}_r$ compare with the variance of $N\bar{y}$?

Although $\bar{y}_r = \frac{\overset{n}{\Sigma} y_i}{n_r}$, is the ratio of random variables because n_r (the number of farms reporting in a sample of size n) is a random variable, the variance of $\bar{y}$ comes under a special condition that does not require the formula for the variance of a ratio of random variables. 2/ Among all possible samples of size n, there are samples which have the same number, n_r, of farms reporting. The relative variance of $N_r\bar{y}_r$ among such samples is $\frac{V_r^2}{n_r}$. An approximate average relative variance among all samples of n is obtained by substituting the expected value of n_r in place of its observed value. Thus $\frac{V_r^2}{\bar{n}_r}$ is a good approximation (unless $\bar{n}_r$ is very small) of the relative variance of $N_r\bar{y}_r$ among all samples of n. Therefore, since $\bar{n}_r = nP$, the relative variance of $N_r\bar{y}_r$ is approximately $\frac{V_r^2}{nP}$; and since the relative variance of $N\bar{y}$ is $\frac{V^2}{n}$, we have $\frac{V_r^2}{P}$ for comparison with V^2 where n = 1. Let $D = \frac{V_r^2}{PV^2}$, which shows the variance of $N_r\bar{y}_r$ as a proportion of the variance of $N\bar{y}$. Values of D, which will be referred to as design factors, are presented in column (11) of table 4. As an example, .66 for alfalfa means that in 1970 the variance of $N_r\bar{y}_r$ is 66 percent of the variance of $N\bar{y}$, or that knowledge of N_r could have reduced the sampling variance by 34 percent.

Exercise 7. Suppose a list of the 741 *growers of potatoes in 1970 is available and that a simple random sample of* n_r *growers is selected from this list. For comparison with the answer to exercise 6, assume that* $n_r = 73$, *which is the expected number of potato growers in a sample of 10,000 farms. The estimator is* $N_r\bar{y}_r$, *where* $N_r = 741$ *and* $\bar{y}_r$ *is the mean of a sample of* n_r *growers selected for the list. Find the standard error of* $N_r\bar{y}_r$. *Answer:* 9,717 *acres or* 24.2 *percent.*

Generally, an agricultural statistics program must be based on a list frame (a list of farms or farm operators), an area frame, or a combination of the two. For a statistically efficient basis for sampling for a wide variety of agricultural surveys, auxiliary data about sampling units is important. But there is a substantial cost in obtaining and including auxiliary data in a sampling frame. Column (11) provides some indication of the importance of having information about which farms are producing various agricultural products.

2/ Hansen, Hurwitz, and Madow. Sample Survey Methods and Theory, John Wiley & Sons, Inc., Vol. 1, p. 159.

Table 4--Design factors for two estimators and some other measures

Characteristic	Design factors for estimator, $N_r\bar{y}_r$		Ratio estimator		
			Design factors	Correlation	$\frac{V_X}{V_Y}$
1/	1970	1971	1970	1970	1970
(1)	(11)	(11)	(12)	(13)	(14)
Farmland	1.00	1.00	--	--	--
Population	0.87	0.86	1.82	.31	1.27
Alfalfa	0.66	0.69	0.82	.50	0.75
All corn	0.85	0.90	0.71	.55	.047
All pasture	0.80	0.81	0.83	.42	0.49
Milk cows	0.47	0.52	0.98	.38	0.74
Beef cattle	0.82	0.83	0.93	.27	0.21
Clover & timothy	0.73	0.46	1.00	.10	0.21
Hay for silage	0.60	0.59	0.91	.31	0.23
Cattle marketed	0.92	0.94	0.98	.20	0.067
Soybeans	0.66	0.69	0.98	.17	0.10
Peas	0.78	0.87	0.97	.21	0.074
Sheep	0.74	0.70	1.00	.044	0.072
Spring wheat	0.81	0.59	0.98	.22	0.039
Potatoes	0.83	0.80	0.98	.26	0.030
Snap beans	0.71	0.81	1.00	.036	0.022

-- = Not applicable.

1/ Numbers in parentheses correspond to algebraic descriptions in the appendix.

Farmland As An Auxiliary Variable

Owing to increasing specialization in agriculture, acres in farmland have become less effective as an auxiliary variable, except in special situations. Variance equations for the relative variances of the ratio and mean estimators in a good form for comparison and interpretation, and assuming n = 1, are:

$$V^2\left(X\frac{\bar{y}}{\bar{x}}\right) = V_Y^2 + V_X^2 - 2\rho\, V_Y V_X \text{ and} \qquad (6)$$

$$V^2(\bar{y}) = V_Y^2 \qquad (7)$$

where X is acres of farmland,

V_X^2 is the relative variance of X, which for 1970 is 0.725, the first entry in column (5),

Y is any characteristic other than farmland,

V_Y^2 is the relative variance of Y and its values are found in column (5) except for the first entry, and

ρ is the correlation between X and Y.

Thus, dividing the relative variance (or variance) of the ratio estimator by the relative variance (or variance) of the mean estimator gives:

$$D = 1 + \frac{V_X^2}{V_Y^2} - 2\rho \frac{V_X}{V_Y} \qquad (8)$$

If the value of D is 0.9, for example, the variance of the ratio estimator is 10 percent less than the variance of the mean estimator. Thus, the value of D is an inverse measure of efficiency.

Values of D, equation 8, for 1970 are listed in table 4, column (12). Values for 1971 are not shown because they are very similar. Notice that the ratio estimator is effective for only three characteristics: alfalfa, corn, and pasture. All three are acreages, each is reported by a high proportion of the farms, and each accounts for about 15 to 20 percent of the farmland. For the remaining characteristics, the ratio estimator is ineffective.

The value of D is less than 1 when $\frac{V_X}{V_Y} > 2\rho$. To help understand the conditions where the ratio estimator is effective, the values of ρ and $\frac{V_X}{V_Y}$ are given in columns (13) and (14).

Exercise 8. *Examine equation 8 and note that when* V_X *is small relative to* V_Y, *the value of* D *will be close to* 1, *especially for small to moderate values of* ρ. *On the other hand, when* V_X *is considerably larger than* V_Y, *the potential for loss or gain (value of* D*) is quite sensitive to the magnitude of the correlation. Study the results in columns 12, 13, and 14. Note that for the characteristics at the bottom of the list that* $\frac{V_X}{V_Y}$ *is small and the values of* D *are close to* 1. *Compare the results for population and hay for silage, two characteristics that have the same correlation. Suppose* $\rho = .8$ *for population, alfalfa, and potatoes. Find the values of D, assuming the values of* $\frac{V_X}{V_Y}$ *are as given in table 2, column (14). Answer:* 0.58, 0.36, *and* 0.95. *What do you conclude?*

CLUSTER SAMPLING

In the State-farm censuses of Wisconsin, farms were identified by townships, which are the smallest political subdivisions of the State. For most purposes, a township is too large to be suitable as a sampling unit. However, a study of the township as a sampling unit reveals several aspects of the general problem of choosing a sampling unit and of selecting auxiliary information about them. Some questions of interest are: How does the design efficiency of the township (compared to the individual farm) relate to P (the proportion reporting), to the geographic distribution of the characteristics, to the method of estimation, and to stratification? Stratification will be considered later.

A few townships had only one or two farms. Hence, for purposes of this analysis, townships with fewer than four farms were combined with adjacent townships, giving a total of 1,462 individual townships and township combinations which will be referred to simply as "townships." The average number of farms per township was $\frac{101,865}{1,462} = 69.6$ in 1970 and $\frac{98,156}{1,462} = 67.1$ in 1971.

Notation used in the specifications of alternative plans for sampling townships will be:

Y_{ij} is the value of Y for the j^{th} farm in the i^{th} township,

N_i is the total number of farms in the i^{th} township,

$Y_{ti} = \sum_j Y_{ij}$ is the total of Y for the i^{th} township,

$N = \sum^M N_i$ is the total number of farms in the population,

M is the number of townships in the population,

m is the number of townships in a sample of townships,

n_i is the total number of farms in the i^{th} township in a sample,

$n = \sum^m n_i$ is the number of farms in a sample of townships

$\bar{N} = \frac{N}{M}$ is the average number of farms per township,

$\bar{Y}_t = \frac{\sum_i Y_{ti}}{M}$ is the average value of Y per township,

$\bar{\bar{Y}} = \frac{\sum_i Y_{ti}}{N}$ is the population average per farm, and in a

sample of townships y_{ti}, $\bar{y}_t$, and $\bar{y}$ correspond to Y_{ti}, $\bar{Y}_t$, and $\bar{Y}$.

For the individual farm as the sampling unit, the notation will be:

Y_i is the value of Y for the i^{th} farm in the population,

$\bar{Y} = \frac{\sum Y_i}{N}$ is the population average per farm which is the same as $\bar{\bar{Y}}$ under the township notation, and f instead of n will be used for the number of farms in the sample.

Other notation follows from the above definitions.

Four alternatives have been selected for comparison:

1. A simple random sample of m townships is selected from the population of M townships. The sample townships are enumerated completely. The estimator of the population total of Y and its relative variance are given by:

$$\hat{y}_1 = M \frac{\sum_i^m y_{ti}}{m} = M\bar{y}_t \qquad (9)$$

$$V^2 (\hat{y}_1) = \frac{V^2 (Y_{ti})}{m} \qquad (10)$$

$$\text{where} \quad V^2(Y_{ti}) = \frac{\sum_i^M (Y_{ti} - \bar{Y}_t)^2}{\bar{Y}_t^2 \, (M-1)}$$

2. The same specifications as in the first alternative apply except that a ratio estimator is used, the auxiliary variable being number of farms. The estimator and its relative variance are:

$$\hat{y}_2 = N \frac{\sum_i^m y_{ti}}{\sum^m n_i} = N\bar{y} \qquad (11)$$

$$V^2 (\hat{y}_2) = \frac{1}{m} \{V^2(Y_{ti}) + V^2(N_i) - 2 \ \text{Cov} \ (Y_{ti}, N_i)\} \qquad (12)$$

$$\text{where} \quad V^2 (N_i) = \frac{\sum_i^M (N_i - \bar{N})^2}{\bar{N}^2 \ (M-1)}$$

$$\text{and} \quad \text{Cov} \ (Y_{ti}, N_i) = \frac{\sum_i^M (Y_{ti} - \bar{Y}_t) (N_i - \bar{N})}{(\bar{Y}_t) (\bar{N}) (M-1)}$$

Exercise 9. Note that $\hat{y}_2$ *is simply the average per farm in the sample multiplied by* N. *Is* n *in equation 11 a constant? Explain why the variance formula, equation 12 is the correct one to use.*

3. A random sample of m townships is selected with replacement and probabilities proportional to N_i. The estimator and its relative variance are:

$$\hat{y}_3 = N(\frac{1}{m}) \sum_{i}^{m} \frac{y_{ti}}{n_i} \qquad (13)$$

$$v^2(\hat{y}_3) = \frac{1}{m} \frac{\sum_{i}^{M} N_i (\frac{Y_{ti}}{N_i} - \bar{Y})^2}{N\bar{Y}^2} \qquad (14)$$

4. The above three plans are to be compared with a simple random sample of f farms. The estimator and its relative variance are:

$$\bar{y}_4 = N\bar{y} \qquad (15)$$

$$v^2(\hat{y}_4) = \frac{1}{f} \left[\frac{\sum^{N} (Y_j - \bar{Y})^2}{\bar{Y}^2 \ (N-1)} \right] \qquad (16)$$

To compare the alternatives, we will assume a constant sampling fraction, namely $\frac{1}{M}$. This means m=1 and that f=69.6 for 1970 and f=67.1 for 1971. The relative variances of the four estimators (table 5, columns (15), (16), (17), and (18) respectively) are used to compare townships as sampling units to farms.

Divide the variances among townships, columns (15), (16), and (17), by the variances among farms, column (18). For 1970, these design factors are shown in columns (19), (20), (21), and (22) of table 6. Take corn as an example. The design factor 26.0 means that the sampling variance for the first alternative is 26 times larger than the sampling variance for the fourth. Thus, columns (19), (20), and (21) display the high degree of inefficiency that generally exists for a "large" sampling unit. "Large" refers to the numbers of farms in the sampling units for which $Y_i > 0$. Note, for example, that the average number of farmers per township that grew potatoes was less than 1. If potato growers are widely scattered (not concentrated in a few townships), townships are small in size with regard to potato growers compared with characteristics at the top of the list. The wide differences among characteristics points up the importance of making a good choice of size of area sampling units depending on the objectives of the survey.

Exercise 10. Suppose a simple random sample of 100 *townships is selected and equation 9 is used. What is the relative variance of* $\hat{y}_1$*, for corn? Answer:* 0.01245. *In a sample of* 100 *townships, one would expect about* 6,950 *farms in 1970. Assuming a simple random sample of* 6,950 *farms, what is the sampling variance of* $\hat{y}_4$*? Answer:* 0.000479. *Do the two variances differ according to the design factor,* 26.0, *shown in column (19) of table 6?*

Exercise 11. Study columns (19), (20), and (21). Prepare a logical explanation for the reduction in the design factors as P *decreases -- that is, the loss in efficiency of the township is greatest when* P *is large. Under what conditions is this reasonable? What are the implications with regard to a measure of size of sampling units?*

It is also interesting to compare the four alternatives by using the mean per township estimator, column (15), as a base. Thus, the 1971 variances in table 5 were divided by the 1971 variances in column (15); the results are in the right half of table 6.

Table 5 -- Relative variance among townships and farms

Characteristic (1) 1/	Unit	1970				1971			
		Among townships			Among farms mean estimator	Among townships			Among farms mean estimator
		Mean estimator	Ratio estimator	PPS estimator		Mean estimator	Ratio estimator	PPS estimator	
		(15)	(16)	(17)	(18)	(15)	(16)	(17)	(18)
Farms	No.	0.508	--	--	--	0.503	--	--	--
Farmland	Acres	.516	--	0.082	0.0104	.509	--	0.088	0.0112
Population	No.	.590	0.036	.030	.0065	.584	0.038	.032	0.0071
Alfalfa	Acres	.987	.356	.267	.0184	.953	.350	.260	0.0195
All corn	do.	1.245	.681	.529	.0479	1.132	.608	.492	0.0453
All pasture	do.	1.757	1.291	.941	.0437	1.716	1.297	.944	0.0502
Milk cows	No.	.841	.211	.161	.0193	.842	.223	.168	0.0221
Beef cattle	do.	2.079	1.668	1.208	.231	1.976	1.580	1.171	0.226
Clover & timothy	Acres	4.184	4.031	3.024	.227	4.446	4.226	3.086	0.124
Hay for silage	do.	1.543	.931	.659	.19	1.534	.960	.686	0.167
Cattle marketed	No.	8.66	8.22	6.00	2.36	11.77	11.40	8.42	3.16
Soybeans	Acres	10.05	9.83	7.83	1.06	12.14	12.01	10.16	1.46
Peas	do.	9.56	8.94	9.49	1.93	11.04	10.44	33.29	3.42
Sheep	No.	4.16	3.86	3.84	1.99	4.07	3.69	2.98	1.86
Spring wheat	Acres	29.1	28.9	25.9	7.02	19.9	19.6	18.6	2.98
Potatoes	do.	36.1	36.4	51.5	11.35	35.7	36.0	53.8	12.43
Snap beans	do.	84.3	84.3	98.3	21.6	64.5	64.9	102.5	15.10

-- Not applicable.

1/ Numbers in parentheses refer to algebraic descriptions in the appendix.

Table 6 -- Design factors, township compared with individual farm

Characteristic (1) 1/	Unit	1970				1971			Among farms mean estimator
		Among townships			Among farms mean estimator	Among townships			
		Mean estimator	Ratio estimator	PPS estimator		Mean estimator	Ratio estimator	PPS estimator	
		(19)	(20)	(21)	(22)	(23)	(24)	(25)	(26)
Farmland	Acres	49.6	--	7.9	1.00	1.00	--	0.17	0.022
Population	No.	90.8	5.5	4.6	1.00	1.00	0.0065	.055	.012
Alfalfa	Acres	53.6	19.3	14.5	1.00	1.00	.367	.273	.020
All corn	do.	26.0	14.2	11.0	1.00	1.00	.537	.435	.040
All pasture	do	40.2	29.5	21.5	1.00	1.00	.76	.58	.029
Milk cows	No.	43.6	10.9	8.3	1.00	1.00	.26	.20	.026
Beef cattle	do.	9.0	7.2	5.2	1.00	1.00	.80	.59	.114
Clover & timothy	Acres	18.4	17.8	13.3	1.00	1.00	.95	.69	.028
Hay for silage	do.	8.1	4.9	3.5	1.00	1.00	.63	.45	.109
Cattle marketed	No.	3.7	3.5	2.5	1.00	1.00	.97	.72	.268
Soybeans	Acres	9.5	9.3	7.4	1.00	1.00	.99	.84	.120
Peas	do.	5.0	4.6	4.9	1.00	1.00	.95	3.02	.310
Sheep	No.	2.1	1.9	1.9	1.00	1.00	.91	.73	.46
Spring wheat	Acres	4.1	4.1	3.7	1.00	1.00	.98	.93	.15
Potatoes	do.	3.2	3.2	4.5	1.00	1.00	1.01	1.51	.35
Snap beans	do.	3.9	3.9	4.6	1.00	1.00	1.01	1.59	.23

-- = Not applicable.

1/ Numbers in parentheses refer to algebraic descriptions in the parentheses.

Exercise 12. Study column (24). For the ratio estimator, the design factor is 0.065 *for farm population and* 1.01 *for potatoes. Explain this large difference.*

Exercise 13. Assume that the township is the sampling unit and consider the following two alternatives: (1) Select a sample of townships using equal probabilities of selection. Use the ratio estimator, $\hat{y}_2$, for characteristics where the design factor in column (24) is less than 1 and the mean estimator, $\hat{y}_1$, for characteristics where the design factor is approximately 1 or larger. (2) Select the townships with probabilities proportional to N_i (number of farms) which would require using $\hat{y}_3$ as the estimator for all characteristics. Which of these two alternatives would you choose? Why?

Exercise 14. In 1970 there were 234 *growers of snap beans. Suppose there was one grower in each of* 234 *townships. In this case, how would the sampling variances for the four estimators ($\hat{y}_1$, $\hat{y}_2$, and $\hat{y}_4$) compare?*

Exercise 15. Suppose the 234 *growers of snap beans were all located within 5 townships. How would the sampling variances for the four estimators compare? Do you agree that the sampling variance would be very large for all of the estimators, even for sampling fractions as large as 25 or 50 percent?*

Exercise 16. What do the above analyses of the township as a sampling unit indicate regarding selection and use of auxiliary data for incorporation in a sampling frame? Does it appear that a substantial investment in a sampling frame, including obtaining relevant auxiliary information, might be worthwhile and perhaps necessary in some cases? Discuss.

STRATIFICATION--SAME SAMPLING FRACTIONS APPLIED TO ALL STRATA

Simple geographic stratification, using a constant sampling fraction, can generally be relied on to provide some reduction in sampling variances. Quite often the reductions are small, but the cost of stratification might also be very small. Unless one engages in a high degree of refinement, geographic stratification is generally inexpensive and easy to apply. How effective is it?

Design factors are for 1971 for three levels of stratification: 9 crop reporting districts (column 27), 72 counties (column 28), and 1,462 townships (column 29). These design factors are for stratified random sampling with a constant sampling fraction (table 7). They are sampling variances for the three levels of stratification expressed as a proportion of the sampling variance for a simple random sample of farms.

Exercise 17. For corn, find the relative standard error of the mean of a random sample of 1,000 farms stratified by counties. Use the relative variance from table 1 and the design factor from table 7. Answer: 5.2 *percent.*

Study the results in columns (27), (28), and (29) with regard to the distributions of the commodities by CRD (table 3). One might have anticipated that the gains from stratification would have been greater for commodities with the most geographic concentration. However, the last four commodities on the list are more concentrated than those at the top, but the impact of stratification was somewhat less. Remember, the comparisons being discussed assume a constant sampling fraction.

Be cautious about judging the impact of stratification from differences among stratum means. For stratified random sampling, the sampling variance is an average of within-stratum variances. In general, it is better to try judging the impact of stratification with regard to within-stratum variation. That is not easy to do when the within-stratum variances differ widely from stratum to stratum as in the case of potatoes.

Table 7 -- Design factors for stratified random sampling, 1971

Characteristic	Level of stratification (constant sampling fraction)				Optimum allocation by CRD
(1) 1/	None	CRD (27)	County (28)	Township (29)	(30)
Farmland	1.000	.981	.964	.908	--
Population	1.000	0.996	0.985	0.947	0.989
Alfalfa	1.000	0.929	0.898	0.813	0.908
All corn	1.000	0.934	0.893	0.849	0.797
All pasture	1.000	0.926	0.908	0.733	0.783
Milk cows	1.000	0.993	0.962	0.897	0.978
Beef cattle	1.000	0.981	0.968	0.935	0.837
Clover & timothy	1.000	0.826	0.780	0.636	0.616
Hay for silage	1.000	0.996	0.986	0.951	0.976
Cattle marketed	1.000	0.995	0.992	0.971	0.578
Soybeans	1.000	0.982	0.948	0.908	0.543
Peas	1.000	0.997	0.991	0.982	0.489
Stock sheep	1.000	0.997	0.995	0.988	0.806
Spring wheat	1.000	0.982	0.957	0.919	0.619
Potatoes	1.000	0.998	0.983	0.958	0.472
Snap beans	1.000	0.996	0.987	0.935	0.315
Average, all characteristics	1.000	0.969	0.950	0.896	0.714

-- = Not available.

1/ Numbers in parentheses refer to algebraic descriptions in the appendix.

Table 8 shows average alfalfa and potato acreages per farm by CRDs. This illustrates that one cannot accurately judge gains from stratification solely from information contained in this table. For alfalfa, the relative variance among the CRD means is less than 1 and for potatoes more than 12. But as shown by table 7 the gain from stratification is less for potatoes. Note the gains from optimum allocation, column (30).

Table 9 illustrates in another way the point about between-stratum variances as a basis for judging stratification. The simple analysis of variance table is used to display between- and within-stratum variances for alfalfa with stratification by CRD and township. Note the size of the mean square among CRDs compared to the mean square among townships. Defining strata with the idea of maximizing the variance among them can be misleading, especially if their number is not fixed.

Before considering alternative allocations of a sample to CRDs, it is interesting to examine the impact of stratification when the township is the sampling unit. For comparing the township with the individual farm as sampling units, the variances in columns (16) and (32) within-state and within-CRD variances, respectively (table 10), must be multiplied by 67.14 to convert the variances among townships to a basis of one farm. Multiplying column (16) by 67.14 and dividing by column (5) gives column (33), the design factors for the township compared to the farm when there is no stratification. To illustrate, the design factor for "population" is $\frac{(67.14)\ (0.038)}{0.476} = 5.36$. Column (34) is derived from columns (31) and (32) in the same way. The design factors in column (34) are somewhat less than those in column (33). In other words, the loss in efficiency when the township in the sampling unit is not as great when stratification is applied. One might also say that gains from stratification are somewhat greater for the township than for the individual farm. From another view, one might say that stratification is more important when the sampling units are large.

Table 8 -- Means per farm by CRD, alfalfa and potatoes, 1971

CRD	Average acreage per farm	
	Alfalfa	Potatoes
	Acre	
1	24.0	0.14
2	10.9	0.66
3	25.1	0.99
4	36.5	0.02
5	24.0	2.15
6	39.0	0.07
7	42.2	0.09
8	32.1	0.10
9	30.3	0.86
State average	30.7	0.43

Table 9 -- Analysis of variance, alfalfa, 1971

Source of variation	Degrees of freedom	Sum of squares	Mean square
Total	98,155	121,050,012	1,233
Among CRDs	8	8,615,115	1,076,889
Within CRDs	98,147	112,434,897	1,146
Among townships	1,461	24,066,000	16,472 1/
Within townships	96,694	96,984,000	1,003

1/ Derived. The within-township mean square 1,003 was available on the computer printout but the sums of squares for townships were not. Although these numbers were derived, they are accurate to at least three digits.

Table 10 -- Individual farm and township variances, overall (State) and within CRDs, 1971

Characteristic (1) 3/	Individual farms 1/		Townships 2/		Township design factor	
	Within State (5)	Within CRD (31)	Within State (16)	Within CRD (32)	Within State (33)	Within CRD (34)
Farmland	0.750	0.736	--	--	--	--
Population	0.476	0.474	0.038	0.041	5.36	5.81
Alfalfa	1.306	1.213	0.350	0.224	18.0	12.4
All corn	3.04	2.84	0.608	0.380	13.4	8.98
All pasture	3.37	3.12	1.297	0.931	25.8	20.0
Milk cows	1.48	1.47	0.223	0.212	10.1	9.68
Beef cattle	15.06	14.77	1.580	1.188	7.04	5.40
Clover and timothy	8.21	6.78	4.226	2.184	34.5	21.6
Hay for silage	11.35	11.31	0.960	0.910	5.68	5.40
Cattle marketed	212.6	211.6	11.40	10.24	3.60	3.25
Soybeans	97.3	95.5	12.01	9.94	8.29	6.99
Peas	230.8	230.0	10.44	9.19	3.04	2.68
Sheep	124.6	124.2	3.69	3.19	1.99	1.72
Spring wheat	190.6	187.2	19.6	15.6	6.90	5.59
Potatoes	809	807	36.0	33.5	2.99	2.79
Snap beans	993	989	64.9	63.3	4.39	4.30

-- = Not applicable.

1/ Mean per farm estimator, relative variances among individual farms.

2/ Ratio estimator (see equation 11), relative variances among township on the basis of one township.

3/ Numbers in parentheses refer to algebraic descriptions in the appendix.

Exercise 18. Column (27) in table 7 equals the entries in column (31) divided by the entries in column (5). It shows the effectiveness of stratification by CRD when the farm is the sampling unit. With reference to table 10, divide column (32) by column (16), which gives corresponding design factors for the township. Compare the results with column (27). What does this comparison show regarding the effectiveness of stratification?

Exercise 19. What is the difference between column (33) in table 6 and column (20) in table 6?

STRATIFICATION--ALLOCATION OF SAMPLE TO CRDs

To illustrate the impact on sampling variance of alternate allocations of a sample to strata, CRDs will be used as strata. Three characteristics have been selected for this purpose: alfalfa, beef cattle, and potatoes. Alfalfa and potatoes represent two widely different geographic patterns of production. Beef cattle falls between the two.

For stratified random sampling, general formulas for the estimator of a population total and its relative variance are:

$$\hat{y} = \Sigma N_h \bar{y}_h \qquad (17)$$

$$V^2(\hat{y}) = \frac{1}{N^2 \bar{Y}^2} \left[\Sigma \frac{N_h^2 S_h^2}{n_h} \right] \qquad (18)$$

where N_h is the population number of farms in stratum h,

$N = \Sigma N_h$ is the total number of farms in the population,

$\bar{y}_h$ is the sample mean for stratum h,

n_h is the size of the sample in stratum h,

$S_h^2 = \frac{\Sigma_i (Y_{hi} - \bar{Y}_h)^2}{N_h - 1}$ is the variance within stratum h,

Y_{hi} is the value of Y for the i^{th} farm in stratum h,

$Y_h = \frac{\Sigma Y_{hi}}{N_h}$ is the population mean for stratum h, and

$\bar{Y}$ is the overall population mean.

Equation 18 gives the relative variances of $\hat{y}$ for a sample of size n, where $n = \Sigma n_h$. For comparison with previous results, the relative variance of $\hat{y}$ should be expressed on the basis of a hypothetical sample of one farm. This is accomplished by multiplying the right side of equation 18 by n. Thus,

$$V^2(\hat{y}) = \frac{n}{N^2 \bar{Y}^2} \left[\Sigma \frac{N_h^2 S_h^2}{n_h} \right] \qquad (19)$$

is a general expression for the relative variance of $\hat{y}$ expressed on the basis of a sample of one farm. Equation 19 will be used to find the relative variance of $\hat{y}$ for alternate allocations.

Exercise 20. For beef cattle and 1970, verify the sample allocations shown in columns (35), (37), and (38) of table 13. Use the within-stratum standard deviations shown in column (41) of table 13 and N_h *and* Y_h *shown in table 11.*

Exercise 21. Using equation 19, verify the following two relative variances for alfalfa in table 12: 1.261, *the relative variance for 1971 when the sample is allocated in proportion to* Y_h *in 1971; and* 1.192, *the relative variance for 1971 when the sample is allocated according to the 1970 optimum.*

You probably know that sample allocations which differ by a small amount from optimum will result in a small or negligible increase in variance. This is fortunate because in practice, optimum allocation can at best only be approximated. Moreover, the optimum allocation varies among the characteristics included in a survey.

When the within-stratum standard deviations, S_h, are equal, the optimum allocation is the same as allocating the sample in proportion to N_h. It follows when the standard deviations, S_h, are moderately different that the variance for optimum allocation will be only slightly less than the variance for an allocation proportionate to N_h. And assuming small variation in the unknown values of S_h, estimates of S_h must be precise, or an effort to reduce variance by optimizing the allocation could result in an increase in variance. This suggests that rather large differences in the S_h might be necessary before optimizing the allocation is worthwhile.

Turn to tables 12, 13, and 14 and study the sample allocations in relation to the variances and design factors presented at the bottom of each table. Reference to table 11 may be helpful in understanding or interpreting the results. In particular, note the wide variation in S_h for potatoes. The largest S_h in 1971 was about 33 times larger than the smallest and the reduction in variance attributable to optimum allocation was substantial. Perhaps of greater importance, from a practical point of view, is the fact that when the 1970 optimum allocation was used in 1971, the design factor was 0.502 compared to 0.472 for the 1971 optimum. That is, the 1970 optimum allocation was nearly as effective in 1971 as the 1971 optimum.

Exercise 22. Due to an interest in allocating the sample to minimize the sampling variance for potatoes, suppose a proposal has been made to allocate the sample according to the 1970 optimum allocation for potatoes. What would the relative variances in 1971 be for alfalfa and beef cattle?

Usually some prior information about the values (or relative values) of Y_h is available. Suppose accurate estimates of the stratum totals, Y_h, exist for a previous year, but no estimates of the S_h are available. A sample could be allocated in proportion to the estimates of Y_h or in proportion to some function of the estimates of Y_h, such as the square root of the estimates. If Y_h is approximately in proportion to $N_h S_h$, a sample allocated in proportion to Y_h will be close to optimum.

Exercise 23. Show, algebraically, that the optimum size of sample from stratum h *is proportional to the stratum total* Y_h, *when* $\frac{S_h}{\bar{Y}_h}$ *is constant.*

For the three commodities, examine the differences between the allocations in proportion to Y_h in 1970 and the optimum allocations. Generally, strata which have the largest values of $\bar{Y}_h$ will have the smallest coefficients of variation, $\frac{S_h}{\bar{Y}_h}$, which means (compared to optimum) that allocating a sample in proportion to Y_h will allocate too much of the sample to strata with large values of Y_h and not enough to strata with relatively small values of Y_h. This phenomenon is apparent in tables 12, 13 and 14. With experience, and in the absence of estimates of S_h, one might decide to allocate the sample in proportion to estimates of Y_h and then arbitrarily increase the sample by 50 percent or more for strata having the smallest values of Y_h.

Table 11 -- Stratum (crop reporting district) totals for 1970

CRD	Farms N_h	Alfalfa		Beef cattle		Potatoes	
		Farms reporting	Total acres Y_h	Farms reporting	Cattle Y_h	Farms reporting	Total acres Y_h
				Number			
1	10,748	5,502	236,900	3,012	48,907	27	1,737
2	11,166	3,462	106,157	2,189	26,009	87	5,524
3	5,917	3,780	149,872	1,071	16,433	134	7,826
4	15,342	12,146	531,718	4,552	106,283	26	698
5	9,616	5,853	212,840	2,193	39,790	245	16,820
6	15,164	12,893	543,090	2,959	36,839	43	1,213
7	13,654	12,115	548,712	4,879	167,775	10	642
8	14,315	11,558	420,207	4,767	128,472	31	1,032
9	5,763	4,125	157,222	1,273	25,451	138	4,587
Total	101,685	71,434	2,906,718	26,859	595,961	741	40,079

Table 12 -- Alternative sample allocations and sampling variances, alfalfa

CRD	Allocations of a sample of 1,000 farms						Standard deviation, S_h	
	Number of farms, N_h		Item total, Y_h		Optimum, N_hS_h			
	1970 (35)	1971 (36)	1970 (37)	1971 (38)	1970 (39)	1971 (40)	1970 (41)	1971 (41)
1	106	105	81	82	114	110	33.2	34.9
2	110	111	36	39	70	72	19.6	21.7
3	58	61	52	49	56	55	29.5	30.5
4	151	150	183	179	168	166	34.0	37.1
5	94	90	73	70	88	84	28.7	31.0
6	149	150	187	190	158	154	32.6	34.5
7	134	138	189	190	147	146	33.8	35.3
8	141	138	145	144	139	137	30.3	33.3
9	57	57	54	57	60	76	32.5	44.1
Total	1,000	1,000	1,000	1,000	1,000	1,000	32.3	35.1

	Relative variances [1/]		Design factors	
Allocation of sample	1970	1971	1970	1971
No stratification	1.277 [2/]	1.306 [3/]	1.000	1.000
Proportionate to N_h	1.178 (35) [4/]	1.213 (36)	.922	.929
Proportionate to Y_h	1.228 (37)	1.261 (38)	.962	.966
Optimum	1.156 (39)	1.186 (40)	.905	.908
According to Y_h in 1970	--	1.273 (37)	--	.975
According to N_hS_h in 1970	--	1.192 (39)	--	.913

-- = Not available.

[1/] Assumes n = 1 rather than n = 1,000

[2/] From table 1

[3/] From table 2

[4/] Numbers in parentheses refer to the column numbers of the sample allocations corresponding to the relative variances.

Table 13 -- Alternative sample allocations and sampling variances, beef cattle

CRD	Allocations of a sample of 1,000 farms						Standard deviation, S_h	
	Number of farms, N_h		Item total, Y_h		Optimum, N_hS_h			
	1970 (35)	1971 (36)	1970 (37)	1971 (38)	1970 (39)	1971 (40)	1970 (41)	1971 (41)
1	106	105	82	74	73	66	14.3	14.2
2	110	111	44	52	50	52	9.6	10.7
3	58	61	27	32	36	36	12.8	13.4
4	151	150	178	185	155	150	21.4	22.8
5	94	90	67	70	73	92	16.1	23.2
6	149	150	62	72	82	99	11.5	15.1
7	134	138	281	286	242	232	37.8	38.4
8	141	138	216	186	235	201	35.0	33.5
9	57	57	43	43	54	72	19.9	28.5
Total	1,000	1,000	1,000	1,000	1,000	1,000	23.5	24.9

	Relative variance [1]		Design factor	
Allocation of sample	1970	1971	1970	1971
No stratification	16.07 [2]	15.06 [3]	1.000	1.000
Proportionate to N_h	15.74 (35) [4]	14.77	.979	.981
Proportionate to Y_h	13.02 (37)	13.30 (38)	.810	.883
Optimum, N_hS_h	12.71 (39)	12.61 (40)	.791	.837
According to Y_h in 1970	--	13.52 (37)	--	.898
According to N_hS_h in 1970	--	12.87 (39)	--	.855

-- = Not available.

[1] Assumes n = 1 rather than n = 1,000

[2] From table 1

[3] From table 2

[4] Numbers in parentheses refer to the column numbers of the sample allocations corresponding to the relative variances.

Table 14 -- Alternative sample allocations and sampling variances, potatoes

CRD	Allocations of a sample of 1,000 farms						Standard deviation, S_h	
	Number of farms, N_h		Item total, Y_h		Optimum, N_hS_h			
	1970 (35)	1971 (36)	1970 (37)	1971 (38)	1970 (39)	1971 (40)	1970 (41)	1971 (41)
1	106	105	43	34	77	70	5.8	5.6
2	110	111	138	168	155	183	11.2	13.9
3	58	61	195	139	140	94	19.1	13.0
4	151	150	17	10	32	17	1.7	1.0
5	94	90	421	451	329	352	27.6	32.8
6	149	150	30	23	71	39	3.8	2.2
7	134	138	16	29	60	89	3.6	5.4
8	141	138	26	32	63	65	3.6	4.0
9	57	57	114	114	73	91	10.2	13.4
Total	1,000	1,000	1,000	1,000	1,000	1,000	11.1	12.3

	Relative variances [1]		Design factors	
Allocation of sample	1970	1971	1970	1971
No stratification	789 [2]	809 [3]	1.000	1.000
Proportionate to N_h	787 (35) [4]	807 (36)	.997	.998
Proportionate to Y_h	535 (37)	480 (38)	.678	.593
Optimum, N_hS_h	405 (39)	382 (40)	.513	.472
According to Y_h in 1970	--	570 (37)	--	.704
According to N_hS_h in 1970	--	406 (39)	--	.502

-- = Not available.

[1] Assumes n = 1 rather than n = 1,000

[2] From table 1

[3] From table 2

[4] Numbers in parentheses refer to column number of the sample allocations corresponding to the relative variances.

APPENDIX

Table Column Number	Description
(1)	Characteristics
(2)	$\sum^{N} Y_i$, where Y_i is the value of characteristic Y for the i^{th} farm in the population of N farms.
(3)	$\frac{\sum^{N} Y_i}{N} = \bar{Y}$
(4)	$\sqrt{\frac{\sum^{N} (Y_i - \bar{Y})^2}{N-1}}$
(5)	$\frac{\sum^{N} (Y_i - \bar{Y})^2}{\bar{Y}^2 (N-1)} = V^2$
(6)	N_r, the number of farms in the population with $Y_i > 0$. All farms have some farmland so N_r for farmland is equal to N.
(7)	$\frac{N_r}{N} = P$
(8)	$\frac{\sum_{i}^{N_r} Y_{ri}}{N_r} = \bar{Y}_r$ r is used to designate a subset of farms with $Y_i > 0$. In an expression like $\sum Y_{ri}$, i is an index of farms in the subset. $\sum^{N_r} Y_{ri} = \sum^{N} Y_i$
(9)	$\sqrt{\frac{\sum^{N_r} (Y_{ri} - \bar{Y}_r)^2}{N_r - 1}}$
(10)	$\frac{\sum^{N_r} (Y_{ri} - \bar{Y}_r)^2}{\bar{Y}_r^2 (N_r - 1)} = V_r^2$
(11)	$\frac{V_r^2}{PV^2}$

Table Column Number — Description

(12) $1 + \frac{V_x^2}{V_y^2} - 2\rho \frac{V_x}{V_y}$, where V_x^2 is the relative variance of farmland and V_y^2 is the relative variance for any other characteristic. ρ is the correlation between x and y.

(13)

$$\rho = \frac{\sum^{N} (Y_i - \bar{Y}) \; (X_i - \bar{X})}{\sqrt{\sum^{N} (Y_i - Y)^2 \; \sum^{N} (X_i - \bar{X})^2}}$$

where X is acres of farmland and Y is any other characteristic.

(14)

$$\frac{V_X}{V_Y}$$

(15)

$$\frac{\sum^{M} (Y_{ti} - \bar{Y}_t)^2}{\bar{Y}_t^2 \; (M-1)} \quad \text{where } \bar{Y}_t = \frac{\sum^{M} Y_{ti}}{M}$$

The subscript t indicates that the unit is a township. There are M townships in the population, Y_{ti} is the total of Y for the i^{th} township and $\bar{Y}_t$ is the average value of Y per township. This column is the relative variance among the M values of Y_{ti}. See equation 10.

(16)

$$\frac{\sum^{M} (Y_{ti} - RN_i)^2}{\bar{Y}_t^2 \; (M-1)} \quad \text{where } R = \frac{\sum^{M} Y_{ti}}{\sum^{N} N_i}$$

and N_i is the number of farms in the i^{th} township. Column (16) is the relative variance of the ratio $\frac{Y_{ti}}{N_i}$. Relates to equation 12.

(17)

$$\frac{\sum_{i}^{M} N_i \left(\frac{Y_{ti}}{N_i} - \bar{Y}\right)^2}{\bar{Y}^2 N}$$

This column is the relative variance among townships when selected with probabilities proportionate to N_{ti}. See equation 14.

Table Column Number — Description

(18) $$\frac{1}{\bar{N}}\left[\frac{\sum^{N}(Y_i - \bar{Y})^2}{\bar{Y}^2\ (N-1)}\right] = \frac{V^2}{\bar{N}} \quad \text{where } \bar{N} = \frac{N}{M}$$

This is the relative variance of $\bar{y}$ for a random sample of $n=\bar{N}$ farms.

(19) $\frac{(15)}{(18)}$

(20) $\frac{(16)}{(18)}$

(21) (17)
(18)

(22) (18)
(18)

(23) (15)
(15)

(24) (16)
(15)

(25) (17)
(15)

(26) (18)
(15)

Stratum

(27) CRD

(28) County

(29) Township

$$\frac{\frac{1}{N}\left[\sum N_h \frac{\sum_{i}^{N_h}(Y_{hi} - \bar{Y}_h)^2}{N_h - 1}\right]}{\frac{\sum^{N}(Y_i - \bar{Y})^2}{N-1}}$$

Columns (27), (28), and (29) are average within-stratum variances divided by the overall variance.

h is the index for strata,
N_h is the number of farms in stratum h,
Y_{hi} is the value of Y for the ith farm in stratum h, and
$\bar{Y}_h$ is the average value of Y in stratum h.

(30) $$\frac{\left(\frac{n}{N^2}\right)\sum \frac{N_h^2\ S_h^2}{n_h}}{\frac{\sum(Y_i - \bar{Y})^2}{(N-1)}}$$

Table
Column
Number   Description

where n_h is the size of sample from stratum h,

$n = \Sigma n_h$, and

$$S_h^2 = \frac{\Sigma_i (Y_{hi} - \bar{Y}_h)^2}{N_h - 1}$$

The quantity $\Sigma \frac{N_h^2 S_h^2}{n_h}$ is the variance of an estimate of the population total for a sample of size n. The factor $\frac{n}{N^2}$ changes this variance to the variance of a mean of a stratified random sample assuming a hypothetical sample of n=1. For column (30), optimum allocation of n to CRD's is used to determine the n_h.

(31) $$\left[\frac{1}{\bar{Y}^2}\right] \left(\frac{\Sigma_h N_h S_h^2}{N}\right)$$

This is the average within-CRD variance among farms divided by $\bar{Y}^2$. It is the relative variance of a stratified random sample with allocation proportionate to N_h, assuming a hypothetical sample of n=1.

(32) $$\frac{1}{\bar{Y}_t^2} \left(\frac{\Sigma_h M_h S_{th}^2}{M}\right)$$

where $$S_{th}^2 = \frac{\Sigma_i (Y_{thi} - RN_{hi})^2}{M_h - 1}$$

Y_{thi} is the total of Y for the i^{th} township in stratum h,

N_{hi} is the number of farms in the i^{th} township in stratum h.

This column is the average within-CRD variance among townships for the combined ratio estimator, divided by $\bar{Y}_t^2$.

(33) $$\frac{\bar{N} \text{ times column (16)}}{\text{Column (5)}}$$

where $\bar{N}$ is the average number of farms per township,

(34) $$\frac{\bar{N} \text{ times column (32)}}{\text{Column (31)}}$$

(35) through (40) These columns show alternative allocations of a sample of 1,000 farms to CRDs.

(41) S_h , standard deviations of Y within CRDs.

*U.S. GOVERNMENT PRINTING OFFICE : 1979 0-280-931/ESCS-50

UNITED STATES DEPARTMENT OF AGRICULTURE
WASHINGTON, D.C. 20250