SURVEY DESIGN AND ESTIMATION

FOR

AGRICULTURAL SAMPLE SURVEYS

STATISTICAL REPORTING SERVICE
U.S. DEPARTMENT OF AGRICULTURE
WASHINGTON, D.C.

MAY 1986

PREFACE

The purpose of this document is to provide an overview of the primary issues encountered in the design of a sample survey for agricultural purposes.

The document is not intended to replace the well-known sampling texts nor the many journal articles dealing with specific problems. Rather, it is intended to be a reference document for two types of users.

- The statistician who is faced with attempting to analyze and understand the results of a survey already completed.

- The statistician who needs to learn about survey design principles, but has had little prior exposure and needs a quick overview of the many aspects.

The following materials were prepared as reference materials for commodity statistician. Comments and suggestions for improvement will be greatly appreciated.

FRED A. VOGEL
Director, Statistical Research Division
Statistical Reporting Service
U.S. Department of Agriculture

CONTENTS

		PAGE
I.	INTRODUCTION	1
II.	SAMPLE FRAMES	2
III.	OVERVIEW OF SELECTION PROCEDURES	9
IV.	REVIEW OF ESTIMATING PROCEDURES	23
V.	DESIGN FOR INTEGRATED MULTIPLE FRAME SURVEYS	36
VI.	IMPUTATION FOR NON-RESPONSE	47
VII.	OUTLIERS, ABERRATIONS, BUSTS, AND OTHER TROUBLEMAKERS	54
VIII.	APPROXIMATING SAMPLE SIZES	64

I. INTRODUCTION

Sampling is an application of statistical theory that relies on basic laws of probability to make inferences about a population based on characteristics of a subgroup of the population. Sampling involves more than a selection process. The overall sample design includes the choice of a frame, choice of a sampling unit, determination of the sample size, developing a selection procedure, preparing the estimators and their sampling errors that are consistent with the sample design, statistical controls for detecting and correcting non-sampling errors, and an analysis of the results.

When the term probability survey is used, it implies the ensuing survey will be based upon the following factors.

- a complete well-defined list or "frame" for the population to be surveyed will be used.

- the sample will be selected in a manner that it will be possible to state the probability of including each unit.

- the survey will be conducted in a manner to ensure that the probabilities of selection are maintained.

- the selection probabilities are used as weights to form the estimates.

II. SAMPLE FRAMES

Sometimes the most difficult part of a survey design is to determine what exactly is to be estimated. If the purpose of the survey is to measure the area planted to different crops -- is the survey to include all crops or major crops? Will estimates be needed for the country as a whole or by subareas such as states, counties, villages, etc.? Will tabulations be needed by size of farm? Are the estimates to be proportions, totals, means, totals or means over sub-populations, or ratios, etc.? What level of accuracy is needed for all of the estimates? Is there to be a series of repetitive surveys or a one-time survey?

The population needs to be well-defined. This is important for agricultural surveys. For example, if the total area of a crop is to be measured, is it the area under cultivation (vs. wild), area cultivated by all households, the area cultivated by all farms, or the area cultivated by farms that sell the product?

Once the population has been defined, it is necessary to choose the sampling frame. Two basic types of sampling frames are used for agricultural surveys, and both consist of a listing of elements of the population that allow one to select a sample with known probabilities.

A. List Frame

A list frame can be a list of all farms, a list of all farms producing crops, a list of households associated with farms, a list of villages where farmers reside, etc.

General attributes required for a list to be used for sampling purposes include the following:

- It should be complete for the population of interest. To estimate total crop area or livestock inventories, it should completely represent the population of interest.

- It should be free of duplication. A common problem with maintaining large lists of names is that undetected duplication can exist. Farms included more than once can result in biased estimates. Record linkage

methodology Fellegi and Sunter, (1969) can be used to locate duplication during the frame construction process. Gurney and Gonzalez (1972) describe a procedure to adjust for duplication detected after the sample has been selected.

- It should contain measures of size. A primary advantage of a list frame over an area frame occurs if names on the list contain measures of size that can be used in the sample design. This is especially true if the farms vary considerably in size or if only a few produce some items.

- It should be current. Names, addresses, and measures of size should be reasonably current. Data collection costs can increase sharply if interviewers spend considerable time attempting to contact people that no longer reside at the given address or are out of scope for the survey. A list of names with no measures of size in some instances can be less efficient than an area frame.

List frames can be constructed many ways. If periodic censuses are conducted, the final census list can represent a sampling frame. Care must be taken to objectively evaluate the completeness of such a list, and the quality of the measures of size. This is especially true if there is a considerable time lapse since the census. Serious consideration needs to be given to procedures and information to periodically update the list.

A snowball method can be used for specialty items. A small starter list is used -- each name is asked to give names of other operators. New names obtained are asked to report additional names. This process can continue through successive rounds until no new names are obtained. (Strand, 1970)

In some instances, lists can be constructed in more than one stage. The first stage often involves the identification of administrative areas such as counties or villages. A sample of villages can be selected, then only in the selected villages is a list of farms constructed. Some considerations for this type of list frame construction follow:

- The primary sampling units (villages, counties, townships, enumeration districts) must have well-defined boundaries, yet include the population of interest. Primary sampling units (PSU's) with overlapping boundaries will induce a bias into the resulting estimates.

- If the primary sampling units vary considerably in size in terms of the item being measured, some measure of size should be available for sampling purposes.
- All factors for list frames in general apply to the construction of a list in selected counties/villages.

B. Area Frame

An area frame is, as the name implies, the land mass of a country, state, etc. Detailed maps for the area of interest are obtained. These maps can be topographical maps, road maps, aerial photographs, or whatever other maps are available so that the land mass can be divided into small segments of land. The complete list of all segments of land constitutes a "frame." Some factors to consider in the construction of an area frame follow:

- Mapping materials need to be available that allow stratification of land areas into similar classes of agriculture or land use.

- It must be possible to delineate the land mass into segments with boundaries identifiable from the ground to minimize non-sampling errors.

- Some knowledge of the variability between farms and the frequency of occurrence of survey items is needed to determine the size of segment to be used.

The advantages of an area frame are that it is complete, i.e., covers the population of interest, provides defined boundaries for repetitive surveys, can be used for a considerable length of time without updating, and with proper photography or maps minimizes non-sampling errors.

A primary disadvantage is that an area frame becomes inefficient if farms vary considerably in size and/or some items are rare in that they appear on only a few farms. It can be said, however, that the same disadvantages can apply to a list frame if it does not contain adequate measures of size.

C. Multiple Frames

Multiple frame sampling involves the joint use of two or more sample frames. For agricultural purposes, this involves the area frame and a list frame. Both frames have inherent strengths and weaknesses. The choice of frames should attempt to capture the strengths of each.

For example, a list frame with measures of size can be efficient for sampling purposes. However, a list is generally incomplete in that it does not cover the population. An area frame can be complete, but inefficient where measures of size are needed for sampling purposes.

The joint use of an area and list frame relies upon the list for the large, unusual, and rare items, while the area frame can cover general items and also estimate for the incompleteness in the list. Two basic assumptions must be satisfied for multiple frame surveys:

- The combination of sample frames must represent the population of interest.

- It must be possible to determine for each population unit the frame or frames from which it could have been selected.

Although multiple frame sampling can sharply reduce the variance of many estimates, it should not be considered in all cases. If procedures are not followed with care, nonsampling errors will greatly exceed gains in precision. Also, multiframe sampling is not efficient for many foreign applications where large operators or list frames do not exist.

D. Sample Units Vs. Reporting Units

The sample unit is the member of the population subject to being sampled. Some examples follow:

Area Frame Sample Units	List Frame Sample Units
Counties	Counties
Villages	Villages
Clusters of Segments	Farm Operators
Segments	Household Addresses
	Telephone Numbers

The reporting unit is the element for which information is to be obtained. For example, if the household is a sample unit, what data should be associated with it? All farm operators living in the house? Only that for the oldest? Is each farmer to report for all land-operated, or only for land-owned? Definite rules and procedures are needed to ensure that the probabilities of selection are maintained throughout the survey process. Some additional examples follow:

Sampling Unit	Reporting Unit
Segment	Tract (closed). Crop acres inside the segment associated with one farm operation/operator.
	Farm (open). Crop acres, both inside and outside the segment, operated by an operator whose primary residence is inside the segment.
	Weighted. Crop acres on the entire farm are prorated to the segment by the ratio of tract acres to entire farm acres.
Name	All land operated by selected name.

The association between the sampling unit and the reporting unit is especially critical for multiple frame surveys involving an area frame and a list frame. The sampling unit for the area frame is a unit of land. Under a multiple frame context, a unique name is associated with each unit of land -- usually the operator of the land.

The sampling unit from the list frame is a name. The reporting unit for each name is all land-operated by each name. For example, the total land operated by each selected name is determined. Then all crops, livestock, etc. on that land, regardless of ownership are reported.

The overlap between these two frames (remember their different sampling units) is then determined by matching names. If the name of the operator of a tract of land in a selected area frame segment is also on the list frame, the assumption is made that the same land would be reported if the name were selected from the list frame -- thus the two frames overlap for this operation.

Critical assumptions of multiple frame sampling using an area frame along with a list frame are that

- a name can be associated with each unit of land in the area frame sample.
- an area of land can be associated with each name in the list frame sample.
- the overlap between the two frames can be determined by matching names.

REFERENCES

Fellegi, I. P. and **Sunter, A. B.** (1969)
"A Theory for Record Linkage," Journal of the Americal Statistical Association 64, pp. 1183-1210.

Gurney, Margaret and **Gonzalez, Maria Elena** (1972)
"Estimates for Samples From Frames Where Some Units Have Multiple Listings." Proceedings of the Montreal Meetings of the American Statistical Association.

Houseman, Earl E. (1975)
"Area Frame Sampling in Agriculture", SRS #20, Statistical Reporting Service, USDA.

Rao, J. N. K. (1968)
"Some Non-Response Sampling Theory When the Frame Contains an Unknown Amount of Duplication," Journal of the American Statistical Association, March 1968 (pp. 87-90).

Strand, N. D., F. A. Vogel, and **R. P. Moore** (1970)
"Frame Construction by the Snowballing Method and its Evaluation." Cooperative Project, Iowa State University Statistical Laboratory and the Statistical Reporting Service, USDA.

III. OVERVIEW OF SELECTION PROCEDURES

There are myriad ways the actual sample can be selected. Each has strengths and weaknesses depending upon the situation. The following paragraphs provide a brief overview of the primary methods. The reader should refer to a sampling text for a more complete treatment. The examples and descriptions refer to names, however, for area frame sampling, the segment may be substituted.

A. Simple Random Sampling

If a sample of 10 is to be selected from 50 names (or segments), it is necessary to give each name a number between 1 and 50. Then the selection process is merely that of selecting 10 random numbers between 1 and 50. This results in a large number of different combinations of samples that can be selected. This method of selection does assure complete randomness, but it does not assure a geographic distribution nor a size distribution. It is possible through simple random sampling to select the 10 largest operations from 50 or to select names that are all in one corner of the State or country. Therefore, pure simple random sampling is seldom used because more control over the sampling process is desired.

Summary of Simple Random Sampling

1. Total number of possible combinations of 10 that can be selected from 50 = 10,272,278,170 different samples.

2. Each name can appear in 2,054,455,600 different samples and has a chance to appear in combination with every other name in the population. This is an important consideration when dealing with outliers. Any given outlier can appear in many different samples. An important consideration is whether it will be considered an outlier in all samples in which it can appear.
3. Procedure assures complete randomization.
4. Does not assure a geographic distribution.
5. Does not assure a size distribution.

B. Systematic Sampling

Systematic sampling is more commonly used because the names in the frame can be sorted such as by size or in a geographic order. The basic procedure is

to determine a sampling interval by dividing the desired sample size into the population size. To select a sample of 10 from 50, obtain the sampling interval of 5 and then select a random number between 1 and 5 to determine the first selected unit. Then select every 5th unit thereafter. This type of sampling is restrictive in that it minimizes the amount of randomization since only one random number is drawn. It is only possible to select five different samples from the 50 names using systematic sampling.

Summary of Systematic Sampling

1. Total number of possible combinations of 10 that can be selected from 50 = 5 different samples.

2. Each name will appear in only 1 unique sample.

3. Every name does not have a chance to appear at least once with every other name.

4. Can obtain geographic or size distribution if information is available to presort the list.

5. Can be risky if have no knowledge how frame is sorted.

6. Unbiased estimates of sampling errors are not obtained because every name does not have a chance to appear with every other name.

C. Replicated Sampling

Replicated sampling is primarily a method of sampling that involves selecting several small samples instead of one large sample. For example, to select a sample of 10 from a population of 50, replicated sampling could involve selecting 2 samples of 5. The two samples can be selected by simple random selection or by systematic selection. The primary reason for using replicated sampling is to retain the advantage of systematic sampling but to allow enough randomization to estimate sampling errors correctly. Replicated sampling makes it easier to rotate samples and make adjustments in sample allocations.

Summary of Replicated Sampling

1. Total number of possible combinations of 10 that can be selected from 50 by selecting two replicates--each of size 5:

   a. 10,272,278,170 samples if select two samples of size 5 using simple random sampling.

   b. 45 different samples if select two systematic samples--each of size 5. Note that the sampling interval for each sample is 10. Two random numbers are selected between 1 and 10.

   c. 184,528,130 different samples if the 50 names are divided into 5 groups, each containing 10 names and select two random numbers between 1 and 10 from within each group (1st number for replicate 1, 2nd for replicate 2). Note that the groups can be defined or ordered in any way desired and are called paper strata.

2. Can use advantages of systematic sampling but forces more randomization.

3. Can take advantage of randomization resulting from simple random sampling but force some size or geographic distribution into the sample.

4. Simplifies rotation procedures and can adjust sample size by adding or dropping replications.

D. Sampling with probabilities proportionate to size (PPS)

In the previous examples, every name had the same chance of being selected, regardless of the method of selection or its actual size. If a measure of size can be attached to each name, a PPS sample can be drawn. The following example is used to illustrate:

Name	Measure of Size	Accumulated Measure
1	10	10
2	1	11
3	4	15
4	15	30
5	5	35

A PPS sample can be selected using either simple random, systematic or replicated sampling. For example, if a simple random sample of (2) is to be selected, two random numbers between 1 and 35 will be chosen. Any random number between 1 and 10 will select name (1). Only random number (11) will select name (2). To make sure two unique names are drawn, random numbers are selected until 2 unique names have been selected. Procedures as described above for systematic and replicated sampling can also be used to select samples proportionate to the measure of size. To select a sample of 2 using systematic sampling, first determine the interval 35/2 = 17.5. Then select a random number between 1.0 and 17.5. Again, any random number between 1.0 and 10.0 will select the first sample unit. Then add the interval to the first random number to determine the second sample unit.

Summary of PPS sampling

- PPS sampling is used for some surveys so that the sample is self weighting. Units with a measure of size larger than the sampling interval will be in the sample with certainty and maybe more than once.

- Measures of size may not be adequate to use PPS -- consider using to stratify instead.

- Expansion factors can be difficult to compute. Variance calculations can be complex.

E. Cluster Sampling

Suppose the population of 50 farms is clustered into 15 villages. The villages vary in the number of farms associated with them, but for discussion

purposes, assume they each contain 2 to 8 farms. One approach would be to select two villages and survey all farm operators associated with each village.

Summary of Cluster Sampling

1. 105 different combinations of 2 villages can be selected from the 15 villages.

2. Each farm has a chance to appear with each other farm in a sample.

3. Unbiased estimates and sampling errors can be obtained. Sampling errors compared to those from other methods of sampling will be larger if there is more variability between villages than between farms.

4. It is not possible to control the final number of farms in the sample. At the extreme, one sample of two villages could yield 4 farms while another sample of two villages could yield 16 farms. This has a direct impact on sampling variances and survey costs. Because of these reasons, cluster sampling is mainly used when
   - measures of size are available for stratification or PPS sampling, and
   - more than one stage of sampling is used.

F. Two-Stage Sampling

Two-stage sampling is often used in conjunction with cluster sampling. Again, by referring to the example of 50 farms clustered into 15 villages, suppose the decision is to select 5 villages at random, obtain a listing of all farms within each selected village, and select 2 farms from within each village.

1. Each farm has a chance to appear in the sample at least once with every one of the other farms.

2. The overall sample size and survey workload can be controlled.

3. Sampling variability will usually be larger than single stage sampling because two sources of variability are present -- between villages and between farms within villages.

4. Cost factors need to be considered, i.e., cost of building a complete frame vs. additional survey cost for a larger sample using two stages of sampling.

5. Measures of size for PPS or stratified sampling can still be important.

G. Stratification

Stratification can be used for several purposes, but each requires some information about the sample units. Sometimes stratification is used when estimates are to be made for subsets of the population such as:

1) Crop Reporting Districts
2) Milk cows and beef cows
3) Rare items

In these cases, it is not necessary to have a measure of size to stratify--all that is needed is an indication of physical location or presence or absence.

Stratification is also used when there is considerable variability between the size of sample units. The measure of size does not have to be accurate--all that is necessary is that like sample units be grouped together. For example, size codes are completely satisfactory if each one defines "like" units.

How many strata? Generally, only 4 or 5 are needed. We tend to end up with more for some surveys such as for cattle because we want to stratify by size as well as by type (Milk Cows, Cattle on Feed, etc.). The following table shows the relative efficiency of stratified sampling compared to simple random sampling. Note that even with good measures of size, little efficiency is gained with more than 4-6 strata.

TABLE A. EFFECT OF CORRELATION WITH MEASURE OF SIZE AND NUMBER OF STRATA ON SAMPLING ERRORS

NUMBER OF STRATA (L)	R^2 OF SURVEY ITEM WITH MEASURE OF SIZE .20	.40	.60	.80	.90
(STRATIFIED VARIANCE AS RATIO OF SIMPLE RANDOM SAMPLING VARIANCE)					
2	.85	.70	.55	.40	.32
4	.81	.63	.43	.25	.16
6	.80	.61	.42	.22	.11
8	.80	.60	.41	.21	.11

$$V^2_{\hat{Y}st} = \frac{N^2}{n} S^2 \left[\frac{R^2}{L^2} + (1 - R^2)\right]$$

COCHRAN SECTION 5A.8

Note that this analysis holds for one commodity. Suppose 20 distinct commodities are grown in 20 reasonably distinct areas within an area frame. Then, a specific stratum for each commodity can easily be shown to be most efficient.

Where to put stratum boundaries? If stratifying for geographic or type of farm reasons, the breakdown desired will determine the boundaries. If stratification is by size, some general rules of thumb are:

1. Attempt to equalize the total of the item being estimated across the strata.

2. Attempt to make the means as different as possible between strata.

3. Large, unusual farms or those producing rare items can be placed into separate strata.

4. Some strata can be called pre-select (to be included with certainty) if they contain units so large that they would overly influence the variance. If one considers the frequency distribution of the population as a whole, these operations will be in the skewed tail of the distribution. A rule of thumb is to include those more than 2 standard deviations from the "nearest neighbor."

Allocation to Strata? Some knowledge of the standard deviation associated with each stratum is needed.

Suppose N_h = Number of names in the h^{th} stratum and S_h is the standard deviation for the stratum, then the optimum allocation to each stratum is determined by $n_h = n (N_h S_h / \quad N_h S_h)$. Note that the stratum size (N_h) and the variability (S_h) are jointly used. In practice, the S_h values are based on previous samples and will vary from survey to survey. The usual procedure is to obtain or estimate average S_h values.

Allocation for Multiple Purpose Surveys

Many surveys obtain a wide variety of different crops and livestock at the same time. The optimum sample allocations for each item considered individually may differ widely. This requires that a compromise allocation be reached by determining which items are most important and using an allocation that minimizes their sampling error. Huddleston et al (1970) presents a procedure that jointly considers all variables and minimizes the sample size subject to the constraints imposed for individual items.

H. Sampling Errors

The purpose of this section is to discuss the concept of sampling error.
In the previous examples, it was shown many different combinations of samples will provide an estimate of the population value. The sampling error is a measure of how much variability there can be between the different estimates that could be generated from the various samples.

The remaining discussion is centered on an example in Table B which illustrates a population of 5 farms (or could be 5 segments). Note that each farm has some animals and the total number of animals in the population is 15. The population standard deviation is 1.58. This means that given the average number of animals per farm is 3, then 2/3 of the farms differ from that average by 1.58 animals. Sampling theory tells us that we can construct a similar interval around a sample estimate.

Suppose that we wish to select a sample of size 2 and use the information from the sample to estimate how many animals are in the population. If the population consists of 5 farms, and it is desired to select a sample of 2 farms using simple random sampling, then there are 10 possible samples that can be drawn. These 10 samples are listed along with the usual sample estimates. Note that the direct expansions from these samples range from a low of 7.5 to a high of 22.5. Remember that the population total being estimated is 15.0.

Note, that the average of the direct expansions from all possible samples is equal to 15 (which is equal to the true population of the five farms). More noteworthy, however, is that the standard errors also vary considerably depending upon the sample that is drawn. For example, the standard error ranges from a low of 1.94 to a high of 7.75. However, the average of the standard errors is the same as the population sampling error. Therefore, the estimated standard errors are also unbiased.

Two important points can be made from this example:

1. The degree of variability between the 10 direct expansions is measured by the sampling error. Note that 2/3 of the direct expansions are within 15 + 4.33.

2. The sampling error associated with each sample's direct expansion is also an estimate and can vary from sample to sample. Therefore, when analyzing survey results, it is important to be familiar with usual levels of sample errors. Abnormally large values may indicate the presence of an outlier. Unusually low values might indicate something is wrong.

TABLE B. POPULATION--5 FARMS--(SEGMENTS)

FARM	# ANIMALS
A	1
B	2
C	3
D	4
E	5

TOTAL # ANIMALS IN POPULATION = 15.
AVERAGE # PER FARM $\bar{y} = 3$

$$S^2 = \frac{\Sigma (Y_i - \bar{Y})^2}{N-1} = 2.50 \quad S = 1.58$$

POPULATION SAMPLING ERROR OF THE ESTIMATE OF THE POPULATION TOTAL FOR SAMPLE OF SIZE 2 =

$$s\hat{y} = \sqrt{N^2 \frac{(1 - F)}{n} S^2} = 4.33$$

SAMPLE--RANDOMLY SELECT 2 FARMS (SEGMENTS)

SAMPLE	FARM	# ANIMALS	AVERAGE PER FARM	DIRECT EXPANSION	SAMPLING ERROR	C.V.
1	A, B	1, 2	1.5	7.5	1.94	25.8
2	A, C	1, 3	2.0	10.0	3.87	38.7
3	A, D	1, 4	2.5	12.5	5.81	46.5
4	A, E	1, 5	3.0	15.0	7.75	51.6
5	B, C	2, 3	2.5	12.5	1.94	15.5
6	B, D	2, 4	3.0	15.0	3.87	25.8
7	B, E	2, 5	3.5	17.5	5.81	33.2
8	C, D	3, 4	3.5	17.5	1.94	11.1
9	C, E	3, 5	4.0	20.0	3.87	19.4
10	D, E	4.5	4.5	22.5	1.94	8.6

AVERAGE OF DIRECT EXPANSIONS = 15.0 (UNBIASED)

AVERAGE OF SAMPLING ERRORS = $\sqrt{18.75} = 4.33$

I. Summary of Sampling Procedures

It is important to understand the different options and choose the one appropriate to the situation. Since we deal with farms that vary so much in size, some form of stratification is almost always used. The method of replicated sampling within each stratum is generally recommended to inject some geographic distribution as well (assuming the file is sorted into a geographical order). A minimum of replicates are needed to stabilize variance estimates.

J. Determination of Sample Size

Decisions about the sample size become intertwined with decisions about the method of selection, the use of stratification, and level of precision desired in the resulting estimate. The purpose of the discussion here is to point out some factors to consider:

1. Level of Detail - Are total corn acres to be estimated, or is the purpose to estimate acres by variety or by District?

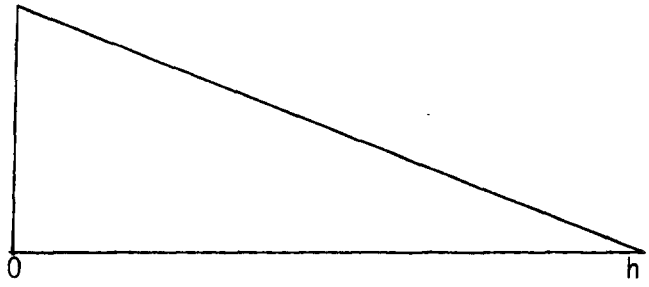
2. Quality of Sample Frame - How complete is the frame? Are control data information available to indicate presence or absence or to stratify by size or variety?

3. Frequency of Occurrence - Is the item to be estimated generally distributed across all farms or do only a few farms have it?

4. Amount of Variability - Do the farms producing the item vary considerably in size? If so, do we have some idea of the variability in size and the identification of the large units.

Cookbook Procedure to Estimate Sample Size (See Chapter VIII for greater detail)

1. Estimate amount of variability around farms expected to have the item of interest. A rule of thumb is to assume the distribution of the data looks like a right triangle (lot of small--a few large).



Then the expected mean is h/3 and the standard deviation is .24 h and V^2=.52. $V^2 = S^2/\bar{x}^2$

Deming (1960) offers an approach to estimate variance. If h is the range of the distribution with shape:

. Normal	= 6^{-1}h	$\bar{x}$ = .5h
. Equilateral triangle	= .20h	$\bar{x}$ = .5h
. Right Triangle, (skewed right)	= .24h	$\bar{x}$ = .33h
. Right Triangle, (skewed left)	= .24h	$\bar{x}$ = .67
. Uniform	= .29h	$\bar{x}$ = .5h
. Binomial (parameters)	= h pq	$\bar{x}$ = ph

2. Estimate the proportion (P) of farms in the sample that can be expected to have the item during the survey. For example, if only 10 percent of the farms produce the item, then P = .10. However, suppose 10 percent of the farms produce the item, and they are identified by a measure of size. Then previous experience can be used to determine what percent of those can be expected to have the item. This can be as high as .80 or .90.

3. Determine the precision required, i.e., the CV of the estimate.

4. The sample size ignoring the finite correction factor can be estimated by $n = (V_r^2 + 1\text{-}P)/P(CV)^2$ where $V_r^2 = \frac{S_p^2}{\bar{x}_p^2}$ which represents the rel-variance of units that have the item of interest.

The values of V_r^2, P, etc., should be for the lowest level of estimation, i.e., by variety, area etc. To adequately estimate for subgroups requires samples considerably larger than to estimate for composite groups.

Note that the sample size is highly dependent upon the value of (P), the proportion of farms having the item of interest. This emphasizes that the most important measure of size is knowing whether or not each farm has the item of interest. Then, a measure of size is only needed if there is considerable variation in the size.

The following table shows how decreasing values of P can increase the sample size to maintain a desired level of sampling variability.

CV (of Estimates)	V_r^2	P	n
.05	.53	.8	360
"	"	.5	800
"	"	.3	1600
"	"	.1	5600

REFERENCES

Cochran, W. G. (1963). Sampling Techniques, John Wiley and Sons, Inc., New York.

Claypool, P, L., R. R. Hocking and **H. F. Huddleston** (1970). "Optimum Sampling allocation to Strata." Journal of Royal Statistical Society, Series C, Vol. 19, No. 3.

Deming, W. E. (1960). Sample Design in Business Research, John Wiley and Sons, New York.

Hansen, M. H., Hurwitz, W. H. and **Madow, W. G.** (1953). Sample Survey Methods and Theory, John Wiley and Sons, New York.

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

Kisah, L and **Frankel, M.** (1968). "Balanced Repeated Replications for Analytical Statistics," Proceedings of the Social Statistics Section of the American Statistical Association.

Raj, D. (1968). Sampling Theory, McGraw-Hill Book Company.

IV. REVIEW OF ESTIMATING PROCEDURES

A. Introduction

The purpose of this review is to briefly describe estimating procedures which includes a wide variety of ratio estimates along with direct expansion estimates. Each procedure has some inherent characteristics that makes it unique. What is different about the ratio to land than the ratio to cropland? When is a direct expansion estimate better than a ratio estimate? What about large farms?

Section B provides a basic review of the direct expansion estimate. It is necessary to have a good understanding of the direct expansion before evaluating ratio estimates.

Section C below summarizes characteristics of ratio estimates.

B. Direct Expansion

The direct expansion estimate depends upon the sampling interval and the item being measured. One way to view the direct expansion is that it is the number of units in the universe (stratum) multiplied by the mean.

$$\text{Direct Expansion} = N\bar{y} = \frac{N \Sigma y_i}{n}$$

It is important to realize that the sample average (average acres per farm, etc.) is the basis of the direct expansion estimate.

A good way to evaluate the direct expansion estimate is to compare the average of positive reports with the average of all reports.

n = number of units in the sample

n_p = number of positive units in the sample

$P_r = \frac{n_p}{n}$ = proportion of positive units in the sample

The overall sample mean can be viewed as

$\bar{y} = P_r \bar{y}_p$ which is the mean of positive reports multiplied times the proportion of positive reports.

When comparing direct expansions from survey to survey, it is important to evaluate both components, that is, the proportion positive and the average of the positive reports.

P_r = Proportion positive. If this changes considerably from one survey to another, two things could have happened.

- There could have been a shift into or out of production (inners & outers).
- Something happened to the sample frame. For example, deadwood could have been cleaned out (or crept in). A change in questionnaire design can lead to more or fewer positive responses.

$\bar{y}_p \ \frac{\Sigma=y_i}{n_p}$ = average of positive reports. If this changes from survey to survey, watch for

- Operations changing size
- Outlier creeping in (or going away)

It is also important to know the difference between the average of all reports and positive reports because of procedures used to estimate for refusals. If a unit is a refusal and we know it has the item of interest, we use $\bar{y}_{p_r}$ to estimate for it. If we know nothing about the refusal, it receives the average of all reports.

The sampling error can also be expressed in terms of the average of positive reports.

$$CV^2(\bar{y}) = \frac{CV^2(\bar{y}_p)}{P_r} + \frac{(1-P_r)}{nP_r}$$

This shows that the sampling error is affected by the CV of positive reports plus the proportion of positive reports.

In fact, all positive reports could be exactly the same, but there would still be some sampling error if there were some zero reports. The following table shows what the CV of an estimate would be if the $CV(\bar{y}_p) = 0$ with different proportions of positive reports.

CV of sample estimate assuming the CV of positive reports = 0.

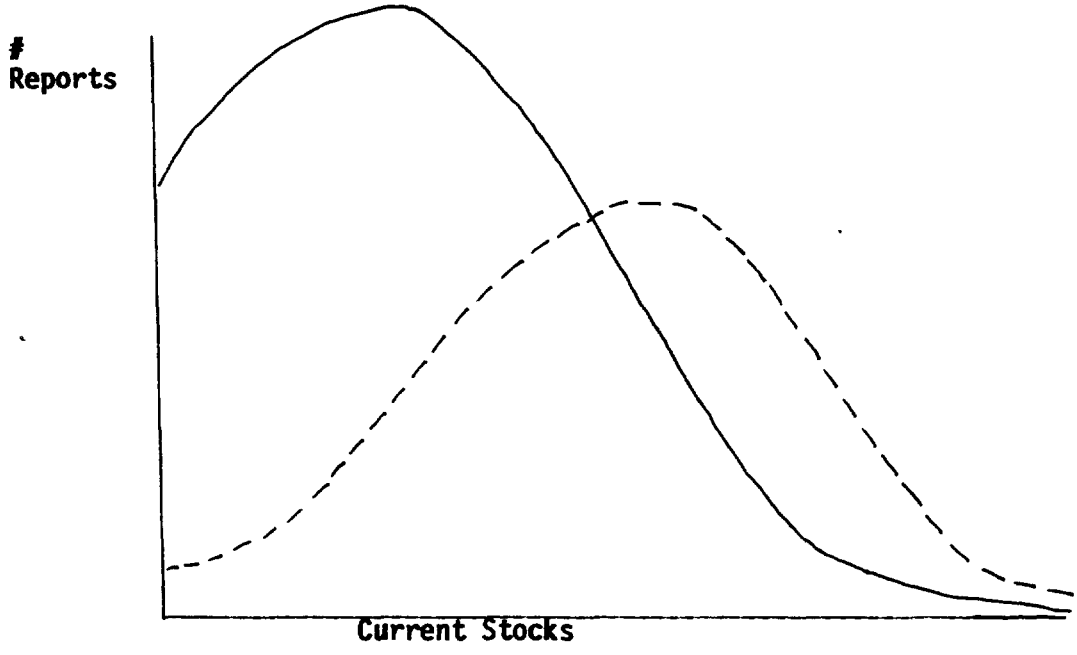
	Percent of Positive Reports			
Sample Size	**(20)**	**(40)**	**(60)**	**(80)**
100	20%	12%	8%	5%
500	9	5	4	2
1,000	6	4	3	2
1,500	3	3	2	1

Note that the effect of the number of positive reports is considerable if sample sizes are small.

Since we tend to stratify for many surveys, the number of positive reports can have considerable impact because sample sizes within a stratum are small.

The following graph shows how outliers can occur with a direct expansion. Whether a report is an outlier will depend upon whether other reports are close to it and whether there are also a lot of zero reports.

Outliers - Direct Expansion



- Large report becoming an outlier depends on how many others are close to it.

- Effect of outlier magnified as proportion of zero reports increases.

  Remember $CV^2(\bar{y}) = \frac{CV^2(\bar{y}_p)}{P_r} + \frac{(1-P_r)}{n P_r}$

- A final point about the direct expansion is that its primary use is to measure level. However, the ratio of the current and previous direct expansions does provide a measure of change. This measure of change can have a larger sampling error than that computed on a matching sample basis because when the samples are independent there is no correlation between current and previous reports. Then

  $$CV^2(R) = CV^2(\text{Current}) + CV^2(\text{Previous})$$

- Note that the negative term related to correlation between current and previous data is not present. It may not be a serious problem, though, if the correlation between matching reports would have been low anyway.

C. RATIO ESTIMATES

The ratio estimate is either the ratio of two means or the ratio of two direct expansions. Therefore all of the factors considered above for the direct expansion also apply to each variable used in the ratio estimate. In addition, the relationship of the two variables to each other needs to be considered.

Ratio estimates seem to be associated with non-probability surveys because they provide a measure of change without knowledge of the entire frame or population. However, ratio estimating procedures are based upon similar theoretical concepts that apply to probability surveys in general.

The purpose here is to discuss the general characteristics of ratio estimates and their strengths and weaknesses.

To start with, let's say we want to estimate total corn acres or total cattle inventory. When we obtain current data (y_i) for each sample unit, we also obtain some additional related information for each sample unit. This auxiliary information could be:

- Acres of land in farm
- Acres of cropland in farm
- Acres of corn last year
- Number of cattle last year/last survey
- Number of cattle during base period
- Feedlot capacity, etc.

The primary reason for using a ratio estimate is to take advantage of the correlation between the current and auxiliary information to increase the precision of the estimate.

The ratio estimate of corn acres, then is

$$\hat{Y}_R = (\bar{y}/\bar{x}) \cdot X \text{ which is}$$

the sample ratio (R) times the population total (X) for land in farm or previous corn acres. The basic assumptions are that:

- The (y_i) and (x_i) can be obtained for each member of the population (sample).

- The (X) is known without sampling error or is known independently of the survey.

- If the (X) is not a known population total, sampling errors increase by the sampling variability associated with the (X).

To examine the characteristics of ratio estimates, it is helpful to look at the coefficient of variation of the ratio (R) or ratio estimate (Y_R). We will assume that (X) is known without sampling error for the following discussion of what makes up the relative sampling error of a ratio or ratio estimate.

$$CV^2 = CV^2(\bar{y}) + CV^2(\bar{x}) - 2\rho\, CV(\bar{x})CV(\bar{y})$$

Corn Acres — Total Cropland — Correlation between X and Y

Note that the CV of the ratio estimate of corn acres is dependent not only on the amount of variability in corn, etc. but also on the variability associated with the auxiliary variable. In fact, the only way the CV of the ratio estimate can be lower than that from a direct expansion is if there is a positive correlation between corn acres and the (X) variate. The size of the correlation coefficient necessary to ensure that the CV of the ratio estimate will be lower than that from the direct expansion can be expressed as

$$\rho > \frac{CV(\bar{x})}{2\, CV(\bar{y})}$$

In other words, if the CV of the auxiliary variable is more than twice that of the ($\bar{y}$) value, then the ratio estimate is out performed by the direct expansion even with perfect correlation. If CV($\bar{x}$) is the same as CV ($\bar{y}$), then the correlation should exceed .5.

Another general requirement for any ratio estimate is that the sample size at the lowest level of summary should exceed 30 to minimize the effect of bias inherent in ratio estimates.

This bias is not that associated with the use of non-probability sampling procedures. All ratio estimates, whether from a probability or a non-probability survey contain a mathematical bias. The amount of bias is minimal if the relationship between (y_i) and (x_i) approximates a straight line through the origin. This bias also becomes minimal as the sample size becomes large. Then the overriding factor is the amount of correlation between (y_i) and (x_i).

The following paragraphs provide more detail about ratio estimates.

C/C or C/P Previous

$\bar{y}_2$ = Corn acres from sample matching with previous year's report

$\bar{y}_3$ = Corn acres from previous survey

$$E_1 = \frac{\bar{y}_2}{\bar{y}_3} \times \text{Previous acres planted}$$

$$CV^2(E_1) = CV^2(\bar{y}_2) + CV^2(\bar{y}_3) - 2\,\rho CV(\bar{y}_2)CV(\bar{y}_3)$$

Some different factors to consider are:

- It is likely that $CV(\bar{y}_2) = CV(\bar{y}_3)$ or will be similar. Than $CV^2(E_1) = 2CV^2(\bar{y}_2)(1-\rho)$ where ρ is the correlation between acres reported this year and last year.

- The correlation between this year's and last year's corn acres will have to exceed .50 for the ratio estimate to have a smaller CV than the direct expansion. This can be a problem if we are dealing with a crop that growers go into or out of producing on a yearly basis. However, the correlation can become large which will make this a good estimator.

- If the $CV(\bar{y}_2)$ is not similar to $CV(\bar{y}_3)$, two things could have happened.

  1. An error occurred and records were not properly matched.

  2. Records were matched properly. The lack of similarity between $CV(\bar{y}_2)$ and $CV(\bar{y}_3)$ indicates there has been a significant change from last year to this year (or an outlier is present). In either case, this should be a signal to dig deeper into the data. Since we are only dealing with matching reports, we must be aware that this may not represent the true picture in non-probability surveys.

- Sample size can be a problem for minor items, especially if ratios are computed at a district level. Sample size is also a problem if only current data is obtained on the current survey because reports must be matched to the previous survey. If both current and previous data are obtained on the current questionnaire, reporting burden is increased along with memory bias.

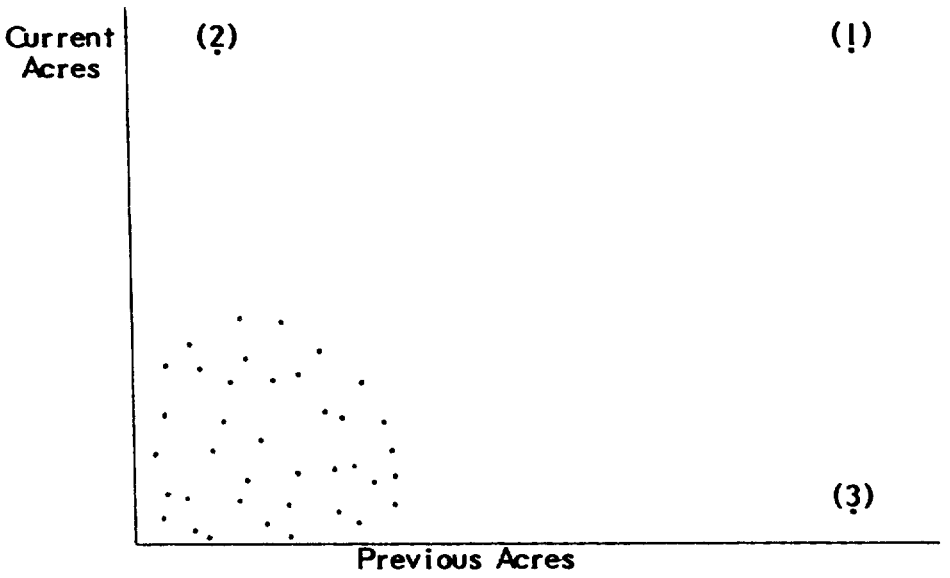
- The association of reporting units between this year and last year can be a problem. If operators change the size of their operations considerably from year to year, then the correlation will suffer. It is not necessary for the operator to report for the same reporting units. However, if the reporting units differ considerably, then the correlation will decline.

- One large report can control or move the ratio.

- The ratio does not provide a measure of level. A consistent bias in one direction can cause a departure from the proper level.

The following graph depicts how outliers can affect the current to previous estimate.

The Effect of Outliers of the C/P Ratio



1) • Will inflate correlation
   • Difficult to spot

2) • Low or negative correlation
   • Also outlier for current acres

3) • Low correlation
   • Large CV on previous acres
   • May not be outlier for current acres

Ratio to Land (R/L)

y_1 = Corn acres from sample

x = Land in farms from sample

E_2 = Ratio to Land in Farm $(\bar{y}/\bar{x})x(X)$.

$$CV^2(E_2) = CV^2(\bar{y}_1) + CV^2(\bar{x}) - 2\rho CV(\bar{y}_1)\, CV(\bar{x})$$

Let's look at some different situations:

- All farms are about the same size. Then CV(x) = 0 and the $CV(E_2)$ will be the same as that from the direct expansion. $CV^2(E_2) = CV^2(\bar{y}_1)$

- All farms have about the same corn acres (or vary within a small range) but total farm acres vary in size. Then $CV(\bar{y}_1)$ will be small. However, $CV(\bar{y})$ will be large and the correlation (ρ) will be small. Therefore, the ratio estimate will have a larger CV than the direct expansion. If this occurs, it points out a need to consider stratifying by size of farm.

- The acres in corn on each farm are proportional to total farm acres. In this situation correlation will be close to 1. If entire farm acres are not considerably more variable than corn acres, the ratio estimate will be the best. If there is good correlation between corn acres and land in farm, but the CV of the land in farm is double the CV of corn acres then it may be necessary to stratify by size of farm.

- A final point needs to be made about the ratio to land indication. Note that in the example, the R/L was multiplied by base acres. That is how the R/L should really be used. This can also be done indirectly by using a regression line with the R/L in the X axis and known total acres on the Y axis. The problem with using the regression procedure is that the R/L value can be increasing because total land is declining while corn acres are remaining constant. The increasing size of the R/L will then wrongly imply that corn acres are increasing.

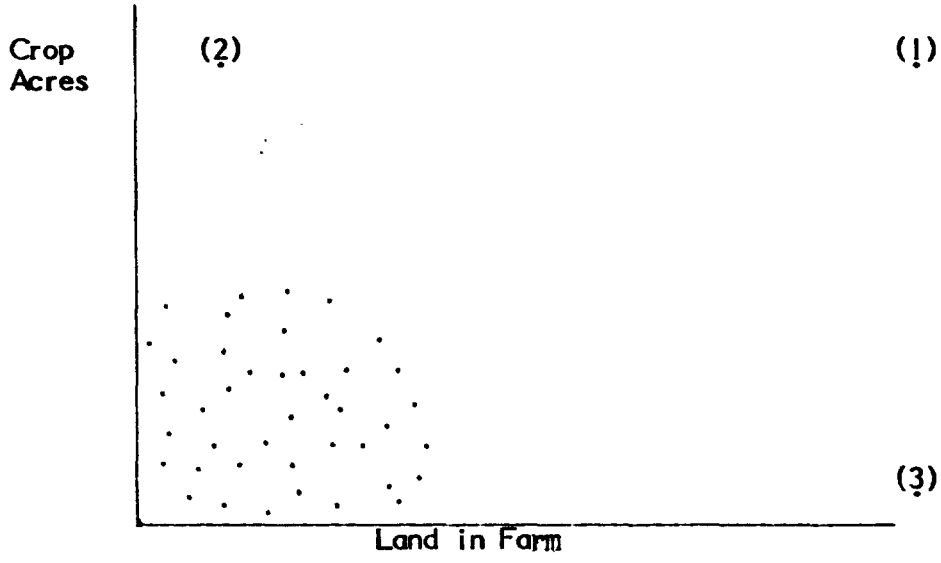
- Sample size can be less of a problem because comparable reports are not required.

- The main problem is to be able to consistently define farmland and to control the effect of outliers (large farms)

- Large farms can control the ratio and destroy whatever correlation may be present elsewhere.

The following graph depicts how outliers can affect the ratio to land estimate.

The Effect of Outliers on the Ratio to Land Estimate



1)
- Difficult to spot -- will inflate correlation
- Can have undue impact on ratio

2)
- Probably an error
- Low or negative correlation
- Will also be outlier for D.E.

3)
- Correlation will be decreased
- Large CV on land acres
- May not be outlier for crop acres

R/Cropland

Y_i = Acres in specific Crop in i^{th} farm.

X_i = Acres in Cropland in i^{th} farm.

- Estimated acres in a specific crop are

  $Y = (\bar{y}/\bar{x}) \times$ Total cropland acres in State. This requires a knowledge of the total acres of cropland in the State. Another way to use the ratio is to use a regression relationship between the ratio and the acres in the crop.

- It may be easier for an operator to define total cropland than total farmland. Total cropland can be the sum of individual crops on the questionnaire.

- There should be less variability in cropland acres than in entire farm acres. This also eliminates the problem of what to do with rangeland, etc. It is still important that

$$\rho > \frac{CV(\bar{x})}{2CV(\bar{y})}$$

- Outliers (large farms) can significantly sway the level of the ratio.

Ratio Relative

R_1 = R/Land or R/Cropland this year

R_2 = R/L or R/Cropland last year for a specific crop. The estimated acreage for a specific crop is

$\hat{Y}$ = (R_1/R_2) x Previous year's acres in specific crop.

$$CV^2(\hat{Y}) = CV^2(R_2) + CV^2(R_2) - 2COV(R_1R_2)$$

- The Ratio Relative incorporates all of the problems with the R/L and R/Cropland plus some more. Unless R_1 and R_2 are computed from matching samples, and there is some correlation between the current and previous years, the ratio relative will always be a weaker indication that the R/L and R/Cropland.

- If we do not compute the R/R using matching samples, its CV will be larger than those of the R/L or R/Cropland because $COV(R_1R_2)$ 0. The Ratio Relative can be influenced by outliers.

Harvested/Planted (H/P)

y_i = Acres for harvest on i^{th} farm

x_i = Acres planted on i^{th} farm

$\hat{Y}$ = $(\bar{y}/\bar{x})$ x Total Acres planted

- The acres to be harvested should be based on the same reporting unit as the planted acres.

- The correlation between planted and harvested acres is probably larger than the correlation to total land or cropland, therefore, the CV of harvested acres should be less than either the R/L or R/CL.

- Since the H/P relies on the same reporting unit, attempts are often made to revise planted acres to maintain a reasonable H/P relationship. This is a special concern when the C/P indicates a different level of harvested acres than the H/P. Several factors need to be considered:

- Planting intentions. What other evidence is available to indicate intentions were not realized?

- C/P ratio more affected by outliers than H/P and probably has a larger CV.

- In a nutshell, the use of the H/P to justify changes in planted acres is questionable.

D. **Summary of C/P, R/L, R/CL, R/R, H/P**

- The R/CL is favored over the R/L because of the problems defining all farmland, especially with grazing and woodland. The R/CL probably has a smaller CV.

- C/P - Relies on matching with previous reports--a disadvantage compared with the R/CL. However, it would be expected to have a smaller CV than the R/CL. However, correlation between current and previous years should be greater than .50.

- R/R - The use of the R/R should be discontinued--its use encompasses all the disadvantages of the R/L and R/CL with no advantages of its own.

- H/P - Probably "best" estimator of group for harvested acres. It is a measure of the percent of planted acres that are to be harvested--it says nothing about the level of planted acres.

What to do about "Large" Farms?

- The above indications, i.e., R/L, etc. are computed at the Crop Reporting District (CRD) level and weighted to a State indication. The size of a large farm (acres reported) may exceed the sum of all other reports combined in the district. As a result, the one farm determines the ratio in a district.

How handle them?

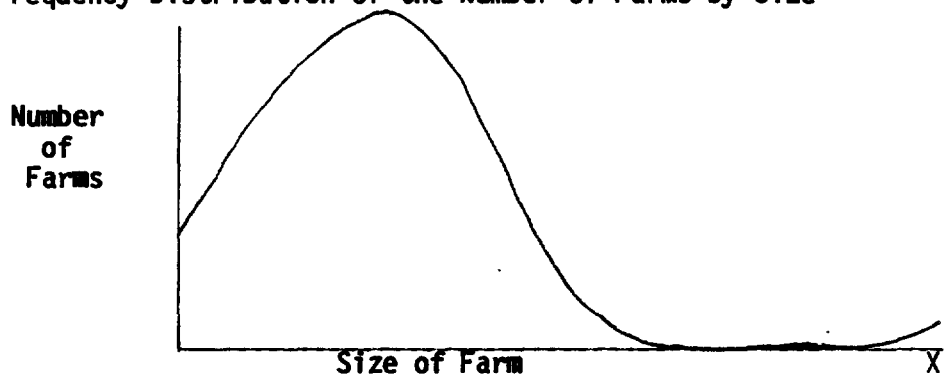
- Create a "large farm" district and place such reports in it.

- Determine the weight to assign to that district. This will require some estimate of the total land encompassed by large farms. This land should be removed from each CRD and assigned to the large farm district for weighting purposes. Census data showing acreage by size of farm will be helpful.

What is Large?

- This will vary by State and crop and requires some knowledge of the distribution of farms by size. If possible, construct frequency distributions from previous surveys or census data.

Frequency Distribution of the Number of Farms by Size



- Attempt to identify a size beyond which the curve becomes very skewed, i.e., beyond which there are few farms (5% of farms). These are the farms which should be in the "large farm" district.

E. **Area Frame Ratio Estimates (R/L, C/P)**

y_1 = Crop acres in i^{th} segment this year

x_1 = Total acres in i^{th} segment this year

$\hat{Y}_1(R/L) = (\bar{y}_1/\bar{x}_1)$ X Total land in State

y_1 = Crop acres in i^{th} segment this year

y_2 = Crop acres in i^{th} segment last year

$\hat{Y}_2$ $=(\bar{y}_1/\bar{y}_2)$ X Last year's Direct Expansion of acres harvested in matching segment

- Both are probability estimates.
- The C/P ratio ($\hat{y}_2$) is probably the "best" overall measure of change from the previous year and should be applied to previous acres. The CV of the C/P should be less than the R/L. The CV of $\hat{Y}_2$ will generally exceed the CV of the direct expansion because there is also sampling error involved with the previous year's Direct Expansion.
- The C/P ratio is based on matching segments (80% of the sample). If it indicates a different level or a different change than does the direct expansion, then the new segments rotated into the frame need to be reviewed carefully. Each rotation group of segments is an independent sample with its own level and sampling variability. A new set of segments can cause a different level from th old set of segments even though the estimates from both sets are within sampling error of each other.

F. Livestock Estimates - Ratio to Base - Ratio to Previous Survey.

The concepts and procedures discussed above apply. Ratio estimates are not used for the cattle and hog multiple frame surveys because they generally have larger sampling errors than those from the direct expansions. This is caused by the small correlation between current and previous survey indications.

The correlations between current reports and the base period or previous survey need to be considered. Also, remember that the estimate for the base period is usually based upon a larger sample than the subsequent monthly or quarterly surveys. The use of the current to previous indication can be helpful to evaluate change from the previous month, but it can cause a departure from the proper level. Therefore, the ratio to base should always receive considerable attention as a measure of level. If the two indications show different "signals" the sample should be reviewed for coverage, presence of outliers, etc.

G. Summary

The direct expansion and ratio estimating procedures both have advantages and disadvantages. The ratio estimates are good for measuring change--however, there is a risk of drifting away from the correct level.

The direct expansion estimates establish a level independent of other surveys. However, if completely new samples are used from survey to survey, the sampling variability from each sample can mask a directional change from a previous time period.

For most surveys we often first attempt to establish a level. However, the surveys are also of a repetitive nature which means we are also attempting to measure the change occurring from a previous period.

Therefore, both ratio and direct expansion estimates should be used, but their strengths and weaknesses should be understood.

V. DESIGN FOR INTEGRATED MULTIPLE FRAME SURVEYS

A. Area Frame

The area frame is stratified into different land-use categories. The sample surveys that use the area frame are multi-purpose surveys. In other words, many items such as specific crop acreages and livestock inventories are obtained at the same time.

Different reporting units are used (Tract, Farm, Weighted). Each has its strength and weaknesses. Exhibit A shows some area and multiple frame estimators. Exhibit B shows how the basic variance can be divided into components representing the proportion of positive reports and the mean of positive reports.

The variance for the different area frame estimators (3), (4), and (5) is based upon the segment as the sampling unit. The data are summed to segment totals for the variance computations.

The variance is influenced by two factors as shown in equations (23) and (28).

- The variability between segments containing the item of interest.
- The number of segments containing the item of interest.

The influence of these factors is relevant when evaluating the segment size and the choice of reporting unit.

As the proportion of the sample that contains an item of interest approaches 1.0, the contribution to the variance comes from the variation in the amount in each segment. Then the design consideration is to determine a segment size that will minimize this variability. On the other hand, if $\bar{P}$ is small, additional variation results from the small number of segments that have the item of interest. Then the design consideration is to define a segment size that will increase the proportion positive. It will still depend upon how the s^2 is affected when the segment size is increased. At this time, the number of strata and how the strata are defined also become important

considerations. In addition, it is necessary to consider a cost function as it relates to the optimum segment size. For example, we can define the cost within a stratum to be

$C_{ah} = C_1 n_{ah} + C_2 n_{ah} t_{ah}$ where

C_1 represents costs related to the number of segments and C_2 include costs caused by the number of tracts (t) in a segment.

The overall problem is to minimize the Var X_a for a fixed total cost. This requires the joint determination of an optimum segment size, the optimum reporting unit, and the allocation to strata.

Basic considerations for an area frame design for a multiple purpose survey follow:

- Definition of Strata -- number and boundaries
- Determination of sample unit size and reporting unit to be used
- Allocation to strata

B. List Frame

The same factors affecting the area frame apply except that there is only one reporting unit.

C. Multiple Frame

The estimators shown (11), (12), and (17) need to be carefully evaluated to fully appreciate factors that need to be considered.

By now the overall design problem should become more clear. The problem involves:

- Area sample unit definition and area reporting unit definition.

- The optimum allocation between frames. For example, Q_h can be small or equal to zero for some strata which reduces the size of the overall list. This can be highly dependent upon the items to be included in the survey.

- Number of strata and stratum definitions for the list and area frames and the allocation to strata. The best design for frames considered independently may not be optimal in the multiple frame sense.

- Optimum weights (P_h and Q_h).

- An overall cost function.

Two additional problems need to be considered. First, the above situation only considered the unbiased direct estimate of the total. With repetitive surveys relying upon replicated sampling, ratio and other estimators can also be used. This is important because the estimates of change can be as important as the estimates of level.

With the implementation of an Integrated Survey Program, we also need to consider the design for the periodic surveys that may follow the initial survey. Should the initial survey be subsampled or new replicates be used? In either case, alternative estimation procedures need to be evaluated. Some considerations for follow-on surveys are summarized below.

- Subsample base survey vs. using newly selected replicates for the follow-on surveys.

- How should the design of the monthly follow-on surveys affect the design of the initial multiple frame survey?

- What are alternative estimators for the initial integrated survey and the follow-on surveys?
  - Model based
  - Composite
    - combine separate indicators
    - weight replicates

EXHIBIT A

AREA FRAME DIRECT EXPANSION ESTIMATOR

$$\hat{Y}_A = \sum_{i=1}^{S^a} \sum_{j=1}^{P_i} \sum_{k=1}^{r_{ij}} e_{ijk} \cdot y'_{ijk} \qquad (1)$$

y'_{ijk} = Value of the survey item in the k^{th} segment, j^{th} paper stratum, and i^{th} land use stratum.

e_{ijk} = Inverse of the probability of selecting the k^{th} segment, j^{th} paper stratum, and i^{th} land use stratum.

r_{ij} = Number of sample replicates or segments in the j^{th} paper stratum, i^{th} land use stratum.

P_i = Number of paper strata in i^{th} land use stratum.

S^a = Number of land use strata in the area frame.

AREA FRAME ALTERNATE REPORTING UNITS

$$y'_{ijk} = \sum_{\ell=1}^{t} w_{ijk\ell} \cdot y_{ijk\ell} \qquad (2)$$

$y_{ijk\ell}$ = Value of the survey item on the ℓ^{th} farm with land in the k^{th} segment, j^{th} paper stratum, i^{th} land use segment.

$$w_{ijk\ell} = \frac{jk^{th} \text{ tract total for farm } \ell}{\text{total for farm } \ell} \qquad (3)$$

tract (closed) estimator

$$w_{ijk\ell} = \begin{cases} 1 & \text{if operator of farm } \ell \text{ lives in the } jk^{th} \text{ segment.} \\ 0 & \text{otherwise} \end{cases} \qquad (4)$$

Farm (open) estimator

$$w_{ijk\ell} = \frac{\text{acres of farm } \ell \text{ in segment } jk}{\text{acres in farm } \ell} \qquad (5)$$

Weighted estimator

$$y_{ijk}^{(nc\ell)} = \sum_{\ell=1}^{t} w_{ijk\ell} \cdot y_{ijk\ell} \cdot f^{A}_{ijk\ell} \qquad (6)$$

$$f^{A}_{ijk\ell} = \begin{cases} 1 & \text{if the operator of the } \ell^{th} \text{ farm did not have chance to be selected from the list.} \\ 0 & \text{otherwise} \end{cases}$$

= Non Overlap Domain -- Domain of interest not represented by list frame.

$$y'^{(o\ell)}_{ijk} = \sum w_{ijk\ell} \cdot y_{ijk\ell} \cdot f^{S^L}_{ijk\ell} \qquad (7)$$

$$f^{S^L}_{ijk\ell} = \begin{cases} 1 & \text{if operator of the } \ell^{th} \text{ farm could also have been selected from the list.} \\ 0 & \text{otherwise} \end{cases}$$

$$\hat{y}_A = \sum_{i=1}^{S^a} \sum_{j=1}^{P_i} \sum_{k=1}^{r_{ij}} e_{ijk} \cdot y'^{(nc\ell)}_{ijk} + \sum_{i=1}^{S^a} \sum_{j=1}^{P_i} \sum_{k=1}^{r_{ij}} e_{ijk} \cdot y'^{(o\ell)}_{ijk} \qquad (8)$$

$$\hat{y}_A = \hat{y}_A^{nc\ell} + \hat{y}_A^{(o\ell)} \qquad (9)$$

LIST FRAME ESTIMATION

$$\hat{Y}_L = \sum_{h=1}^{S^L} \sum_{i=1}^{n_h} e_{hi}\, y_{hi} \qquad (10)$$

y_{hi} = Value of the survey item reported by i^{th} farm in h^{th} list stratum.

S^L = Number of strata in the list frame.

e_{hi} = Inverse of the probability of selecting the i^{th} name in the h^{th} stratum.

MULTIPLE FRAME ESTIMATION

$$\hat{Y}^S_{MF} = \hat{Y}^{no\ell}_A + \hat{Y}_L \quad \text{(Screening Estimator)} \qquad (11)$$

$$\hat{Y}^H_{MF} = \hat{Y}^{no\ell}_A + P\,\hat{Y}^{o\ell}_A + Q\hat{Y}_L \quad \text{(Hartley Estimator)} \qquad (12)$$

$$P + Q = 1 \qquad (13)$$

A more general form of the multiple frame estimator can be evaluated. For example, use $f^{L}_{ijk\ell}$ as described above except that it associates each area overlap farm with a list frame stratum.

$$f^{S_1^L}_{ijk\ell} = \begin{cases} 1 & \text{if the } \ell^{th} \text{ farm in the } k^{th} \text{ segment is also on the list in } \underline{\text{stratum 1}} \\ 0 & \text{otherwise} \end{cases}$$

$$f^{S_2^L}_{ijk\ell} = \begin{cases} 1 & \text{if the } \ell^{th} \text{ farm in the } k^{th} \text{ segment is also on the list in } \underline{\text{stratum 2}}. \\ 0 & \text{otherwise} \end{cases}$$

$$f^{S_h^L}_{ijk\ell} = \begin{cases} 1 & \text{if the } {}^{th} \text{ farm in the } k^{th} \text{ segment is also on the list in } \underline{\text{stratum h}}. \\ 0 & \text{otherwise} \end{cases}$$

Then $$y'^{(o\ell)S_h^L}_{ijk} = \sum_{\ell}^{t} w_{ijk\ell} \cdot y_{ijk\ell} \cdot f^{S_h^L}_{ijk\ell} \qquad (14)$$

$$\hat{y}_h^{(o\ell)S_h^L} = \sum_{i=1}^{S^a} \sum_{j=1}^{P_i} \sum_{k=1} e_{ijk} \cdot y'^{(o\ell)S_h^L}_{ijk} \qquad (15)$$

$$\hat{y}^{(o\ell)} = \sum_{h=1}^{S^L} \hat{y}_h^{(o\ell)S_h^L} \qquad (16)$$

Each area overlap sub-domain can be weighted with the corresponding list stratum as suggested by Bosecker and Ford (1976).

$$\hat{y}^{B}_{MF} = \hat{y}_A^{no\ell} + \sum_{h=1}^{S^L} \left(P_h \hat{y}_h^{(o\ell)S_h^L} + Q_h \hat{y}_h^{S_h^L} \right) \qquad (17)$$

EXHIBIT B

VARIANCE COMPONENTS

The general form for the variance of stratified estimates for the direct expansion is

$$\text{Var}\,\hat{y} = \Sigma N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) s_h^2 \qquad (18)$$

For sample design considerations,it can be helpful to evaluate the variance in terms of variability between all sample units vs. the variability between sample units with the item of interest. The following example shows how the variance can be expressed in terms of positive sample units.

$$s^2 = \frac{\Sigma(y_i - \bar{y})^2}{n-1} = \frac{\Sigma y_i^2 - n\bar{y}^2}{n-1} \qquad (19)$$

n = number in sample

n_p = number of positive sample units

$\bar{y}_p$ = mean of positive reports

s_p^2 = variance of positive reports

P = $\frac{n_p}{n}$ proportion of positive reports

$$\bar{y} = \frac{n_p}{n} \cdot \bar{y}_p \qquad (20)$$

$$s^2 = P\,(s_p^2 + \bar{y}_p^2\,(1-P)) \qquad (21)$$

$$CV^2(\bar{y}) = \frac{s^2}{n\bar{y}^2} = \frac{s_p^2}{nP\bar{y}_p^2} + \frac{(1-P)}{nP} \qquad (22)$$

$$= \frac{CV^2(\bar{y}_p) + (1-P)}{np} \qquad (23)$$

Proof:

$$s^2 = \frac{n_p - 1}{n-1}\left(\frac{\Sigma y_{pi}^2 - n_p^2 \bar{y}_p^2/n}{n_{p-1}}\right) \tag{24}$$

Assume $\frac{n_p - 1}{n-1} \doteq P$ and $\frac{n_p}{n_{p-1}} = 1$ then

$$s^2 = P\ \frac{(\Sigma y_{pi}^2 - Pn_p\bar{y}_p^2 + n_p\bar{y}_p^2 - n_p\bar{y}_p^2)}{n_p - 1} \tag{25}$$

$$= P\ \frac{(\Sigma y_{pi}^2 - n_p\bar{y}_p^2 + n_p\bar{y}_p^2 - Pn_p\bar{y}_p^2)}{n_p - 1} \tag{26}$$

$$= P\, s_p^2 + P\,\bar{y}_p^2\,\frac{(1-P)\,n_p}{n_p - 1} \tag{27}$$

$$= P\, s_p^2 + y_p^2\, P\,(1-P)$$

$$CV^2 = \frac{s^2}{n\bar{y}^2} = \frac{P s_p^2}{nP^2\bar{y}_p^2} + \frac{\bar{y}_p^2\, P(1-P)}{nP^2\,\bar{y}_p^2} \tag{28}$$

$$= \frac{s^2\bar{y}_p}{\bar{y}_p^2} + \frac{(1-P)}{np} = CV^2(\bar{y}_p) + CV^2(P) \tag{29}$$

REFERENCES

Bosecker, Raymond R. and **Ford, Barry L.** (1976)
"Multiple Frame Estimation with Stratified Overlap Domain," Proceeding of the Social Statistics Section, Annual Meeting of the American Statistical Association, 1976.

Fuller, Wayne A. and **Burmeister, Leon F.**
"Estimators for Samples Selected From Two Overlapping Frames," Proceedings of the Social Science Section of the Montreal Meetings of the American Statistical Association, 1972.

Hartley, H.O
"Multiple Frames Surveys," (1962) paper given at Minneapolils Meetings of the American Statistical Association, September 1962.

Huddleston, H. F., Claypool, P.L., and **Hocking, R.R.** (1970)
"Optimal Sample Allocation to Strata Using Convex Programming," Journal of the Royal Statistical Society," Vol. 19, No. 3, pp. 273-278.

VI. IMPUTATION FOR NONRESPONSE

A. Introduction

There is a critical need to emphasize the problem of missing data.

The non-response rate is a valid indicator of survey quality--as valid as coefficients of variation and standard errors. What do 5 percent coefficients of variation mean when the nonresponse rate is 25 percent? They probably do not mean very much.

There are two types of missing data:

1. missing records -- all of the values for a sample unit are missing except for a control number

2. partially complete records -- only a few values are missing for a sample unit.

B. Missing Records -- The Problem of Information

The basic problem with missing records--refusals and inaccessibles--is an information problem. *What information does one have on missing records?* By deleting the missing records from the sample, the assumption is that there is no useful additional information. Thus, the assumption is implicitly made that the missing records are distributed the same as the reported records. With low non-response rates, the impact on survey estimates when the assumption did not hold was minimal. Assumptions that were reasonable when a few records were missing are no longer reasonable as the non-response rate increases.

If it is unreasonable to assume that the missing records are distributed the same as the reported records, what is the best assumption one can make? This question is really based on the more fundamental question of what information

---

Material extracted from a working paper written by Barry L. Ford, "A General Overview of the Missing Data Problem," August 1978.

is available for the missing records. With regard to list surveys for livestock estimates, there can be two types of information:

1. a control variable (measure of size) used to stratify the list
2. geographical information from the mailing address.

Because the control variable is the most important information available for a missing record, control data of a high quality is necessary to improve upon the assumption that reported and missing records have the same distributions. Logically, procedures which adjust for missing records are highly dependent on good control data. This dependency is so strong that before deciding which procedure is the best, one must answer "Is the quality of the control data good enough to warrant the adoption of any procedure over the operational one?"

Some examination has shown that the correlations within each stratum between the control variable and reported variables were usually below 0.30. These low correlations do not necessarily mean that the control variable is inadequate for stratification. However, they do restrict the effectiveness any missing record procedure might have in compensating for nonrespondents. Analysis of missing record procedures indicates that at least a 0.60 correlation within each stratum between the control and reported variables may be needed.

C. Procedures to Adjust for Missing Records

Almost any missing record procedure may be an imputation or a summarization procedure depending on its use. For instance, once a regression has produced an equation representing the relationship between a control variable and a survey variable, this equation may then be applied to the

estimate (a summarization process) or to each missing unit in the sample (an imputation process).

Before one should decide on a missing data procedure, one should decide if a summarization or an imputation procedure is desired. Summarization procedures are usually the more direct approach and, therefore, easier to apply -- especially when the variables are quantitative and the sample design is as uncomplicated as in a stratified simple random sample. On the other hand an imputation procedure produces a "clean" data set (i.e., data with no errors or gaps) and this facilitates further analysis. However, summarization may be ineffective in multi-stage sampling and imputation procedures usually provoke the accusation of "making up" data. Statistics Canada, for example, uses an imputation procedure because one of its primary functions is to produce "clean" data sets which other government agencies use for their own analysis.

Two types of procedures have been used:

- hot deck procedures which rely on a post-stratification of the reported data in order to substitute values from "similar" records
- regression procedures which use regression relationships among the variables to adjust the estimates.

Hot deck procedures are imputation methods while regression procedures can be summarization or imputation methods.

Imputation methods can cause underestimates of standard errors, but replication is a useful tool to correct this defect. If the sample design is complex, even a regression procedure must often be used as an imputation method, and thus, the sample design must be replicated. Although yielding

unbiased estimates of standard errors, replication does complicate the sample design. Therefore, statisticians should be aware that in many situations where a missing record procedure is desired, replication may also be required.

D. Additional Information for Missing Records

All of the previous discussion has been strictly concerned with using existing information to adjust for missing records. However, there is the alternative of collecting additional information.

A good example of this technique is currently being tested by the Statistical Research Division and has already been the subject of one working paper, "A Study of Nonrespondents in Nebraska March Hogs Survey, 1978". This paper suggested using an estimator which only requires knowledge of whether the non-respondent had any hogs or not. This estimator recognizes that a larger proportion of non-respondents without hogs receive zeros. Often non-respondents will give this information in spite of refusing to give specific hog numbers. The main problem is that there is still a subgroup of non-respondents for whom one might not find out even that much information.

Observational data is another example of additional information. On surveys where only personal interviews are used, the enumerators can observe whether livestock or livestock equipment (thus indirectly indicating livestock) are present.

E. Example of Nonresponse Estimator *

The following estimator for the nonresponse domain is based on two assumptions:

1. It will be possible to determine for nonrespondents whether or not they have the item of interest.
2. The distribution for respondents with the item of interest will also represent the nonrespondents.

The following paragraphs provide a short overview of how a direct expansion estimate can be divided into the components used to obtain the final estimate. First, some terms will be defined.

N_h = population number in the h^{th} stratum

n_h = number selected in h^{th} stratum

n_h^p = number of positive reports in the h^{th} stratum

n_h^o = number of valid zero reports (excludes refusals, inaccessibles, etc.)

$n_h^u = n_h^p + n_h^o$ = number of usable reports.

n_h^{rk} = number of refusals, and also known to have item of interest

n_h^{ru} = number of refusals whose status is unknown

$\frac{\sum^{n_h} y_{hi}}{n_h^p} = \bar{y}_h^p$ = mean of positive reports

$\frac{\sum^{n_h} y_{hi}}{n_h^u} = \bar{y}_h^u$ = mean of usable reports

* Material extracted from SRS Report, "Crop Reporting Board Standards."

The direct expansion for the h^{th} stratum can be written as follows:

$$y_h = \frac{N_h}{n_h}\left(n_h^{p} \cdot \bar{y}_h^{p} + n_h^{rk} \cdot \bar{y}^{p} + n_h^{ru} \cdot \bar{y}_h^{u}\right)$$

→Contributions to estimates from refusals, etc., whose status is unknown.

→ Contributions to estimate from refusals, etc., who are known to have the item of interest.

→Contribution to estimate from sample units reporting the item of iterest.

These components can be tabulated at the stratum, State, regional, and U.S. levels if the overall survey is at that level.

One can see after careful examination of the components that the overall estimate is sensitive to the breakdown between refusals whose status is known and those whose status is unknown in addition to the values used to estimate for them. Another procedure that should be developed would involve an estimate standardized for a number of refusals. In other words, how would the indication Y_h react if the number of refusals were constant from survey to survey?

The use of a new sample or a change in survey procedures can change the number of refusals and also the number identified to have the item of interest. Commodity statisticians should have access to these components when evaluating the level of an estimate and the change from a previous survey.

REFERENCES

Bosecker, Raymond R., Data Imputation Study on Oklahoma DES. U.S. Department of Agriculture, Statistical Reporting Service, October 1977.

Crank, Keith N. The Use of Current Partial Information to Adjust for Nonrespondents. U.S. Department of Agriculture, Statistical Reporting Service, April 1979.

Ford, Barry L. Missing Data Procedures: A Comparative Study, U.S. Department of Agriculture, Statistical Reporting Service, August 1976.

Ford, Barry L. Nonresponse to the June Enumerative Survey. U.S. Department of Agriculture, Statistical Reporting Service, Augsut 1978.

Platek, R. and **G. B. Gray.** Some Aspects of Nonresponse Adjustments, Survey Methodology, Volume 11, Number 1, June 1985.

VII. OUTLIERS, ABERRATIONS, BUSTS, AND OTHER TROUBLE MAKERS

A. Introduction

One of the more perplexing problems faced with a sample survey is to follow the sample design and survey concepts with all due care only to end up with an estimate many times larger than could be reasonably expected. In these instances, the presence of an outlier is usually obvious and the trouble maker can be easily located.

Often times, however, the presence of an outlier may not be as noticeable. However, the presence of an outlier can cause a survey estimate to show an increase while the remainder of the sample is pointing to a decline. In some of these instances, an outlier may go undetected or be difficult to locate.

Outliers are a serious problem because they have been dealt with in a subjective manner. First, a search for outliers is usually undertaken only after the survey results appear to be "suspicious". What constitutes a suspicious appearance is not defined. Procedures followed to find the "guilty" observation are mainly in the area of "looking at the data". The search involves reviewing data listings and sometimes ends when a "guilty" unit is found. The search for other "guilty" units may or may not continue. After the outlier is found, considerable debate then ensues about what to do about it. Procedures to handle the outliers are usually just as subjective as procedures used to initially identify them. Worse yet, the degree of subjectivity in the handling of the outliers varies from survey to survey and is usually tailored to meet a particular situation.

Problems related to dealing with outliers are about as old as the study of statistics. One of the earliest recorded discussions of outliers dates back to Bernoulle when he condemned the practice of discarding outliers (Beckman, 1983). However, by now there is a large body of theory related to the identification and handling of outliers. A recent article in Technometrics lists over 150 references . Therefore, the purpose of this paper is to provide a brief description of the outlier problem and to recommend procedures to handle them.

B. **Effect of Outliers in Survey Expansions**

The following example typifies the problem as it often occurs. Suppose the population consists of 5 farms and a random sample of 2 is to be selected to estimate the total number of acres.

Population of Farms	Number of Acres
Farm A	1
" B	2
" C	3
" D	4
" E	40
Population Total =	50 acres.

In practice, many different combinations of samples of 2 from the five can be selected. The goal in sampling is to ensure that whatever combination of 2 is selected, the sample will provide an efficient estimate of the population value.

A total of 10 different samples of size 2 can be selected from the 5 farms. The following table shows the direct expansion and sampling error resulting from each of the 10 samples.

Sample	Farms In Sample	Direct Expansion	Sampling Error	C.V.
1	A, B	7.5	1.9	25.3
2	A, C	10.0	3.9	39.0
3	A, D	12.2	5.8	47.5
4	A, E	102.5	75.5	73.6
5	B, C	12.5	1.9	15.2
6	B, D	15.0	7.7	51.3
7	B, E	105.0	73.6	70.0
8	C, D	17.5	1.9	10.9
9	C, E	107.5	71.6	66.6
10	D, E	110.0	69.7	63.4

Remember we are trying to estimate the population total which is 50. The average of the 10 direct expansions is 50, therefore, the sampling process is unbiased.

However, two things need to be noted

1. The outlier in this example results in a large over estimate when it falls in the sample. However, when it does not appear in a sample, the result is an under-estimate.

2. The level of the sampling error is related to the presence or absence of the outlier which means that it should be used to detect the presence of outliers in repetitive surveys.

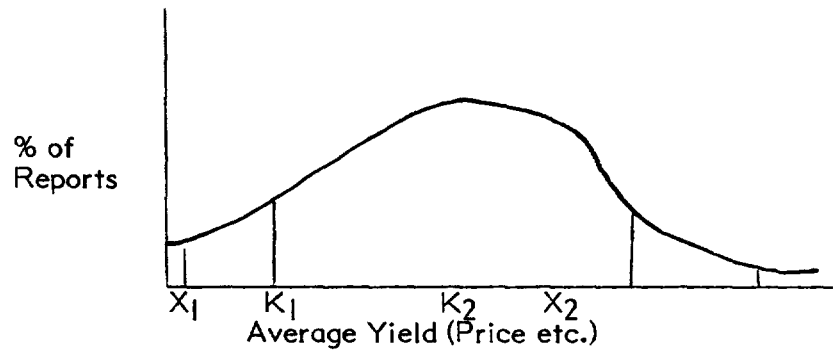
The basic dilemma is while the presence of an outlier plays havoc on a State estimate, the procedure to reconcile its effect will be different at the State level than at the national level. When national estimates are concerned, the survey is probably underestimating if some outliers are not found somewhere.

Several different strategies have been recommended to handle outliers. One approach has been to first identify the extreme values in a data set. Then the nearest neighbor rule is used, i.e., make the extreme values equal to their nearest neighbor or equal to a predetermined constant 5/. This approach assumes the sample has been selected from a symmetric distribution such as a normal distribution. In practice, this is not a reasonable solution for surveys designed to estimate totals such as total acres, livestock inventories etc., because we are generally selecting samples from populations with distributions skewed to the right. However, this approach is feasible for prices paid and received where the average prices follow more of a normal distribution. The following sections will address the problems associated with estimating averages first, then problems with direct expansions will be addressed.

C. Outlier Detection Procedures

1. Estimating Averages

The following frequency distributions provide a picture of the problem with the nearest neighbor rule.



Average Price (Average Condition or Yield)

The outliers (X_1 & X_2) are identified because they fall outside the values of K_1 and K_2. The values of K_1 & K_2 can be determined using different ways.

a) The nearest neighbor rule. This procedure is questionnable because it is dependent on the current sample and the values of K_1 & K_2 can change from survey to survey or from sample to sample.

b) The constant rule where K_1 and K_2 are determined from previous experience. Then values exceeding K_1 & K_2 are made equal to K_1 and K_2, respectively. In this case, careful evaluation is needed to make sure the K_1 & K_2 values do not mask any shifts in the distribution over time.

c) The ESD rule (Extreme Studentized Deviate) [4] which is

$R_i = (\text{Max. } X_i - \bar{X})/s.$

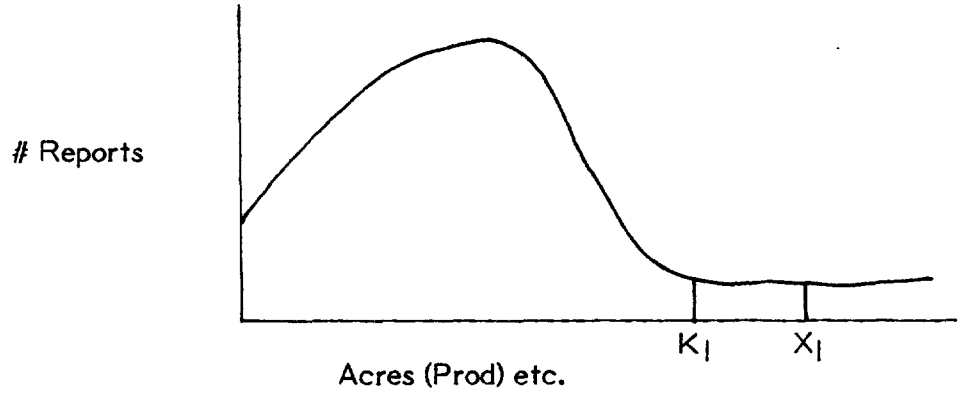
R_1 is computed from the entire sample

R_2 is computed from the sample after the max $|X_i - \bar{X}|$ variable is deleted. Each R_i is compared to a critical value and identified to be an outlier if R_i exceeds the table value. The process of computing

the R_i continues until no more outliers are detected. This is a statistical process that identifies the outliers, but is dependent upon the current sample. An alternate procedure would be to compute (s) from previous historic surveys.

2. Estimating Totals (Direct Expansions)

The next frequency distribution describes the situation when estimating total acres, production, or inventories.



A lot of reports do not have the item at all followed by a cluster of "average" reports. Then the extremes are those large operations that we attempt to sample with probability 1 but which become outliers if they are not classified correctly or whose size has changed from previous surveys. Again, rules can be devised to identify a value K_I beyond which values are determined to be outliers. However, the problem is that since the distribution is one-sided, we cannot identify reports from both ends of the distribution to keep things in balance. Therefore, procedures to adjust for the outliers can result in a negative bias in the estimate if the outlier observation is deleted.

A related problem occurs when the reported data can be small (not an outlier), but because of the probability of selection can become an outlier. This again involves a classification problem and rules can be devised to identify them.

3. Recommended Procedures to Identify Outliers

The following procedures apply when estimating both averages and totals. Since many surveys are repetitive, it is possible to determine (s) values from previous surveys. Although the level of an estimate can move over time, the (s) values should remain fairly consistent unless an outlier is present.

After historic (s) values have been determined, the ESD rule should be used to identify outliers. Another way to use the ESD rule is to identify any report to be an outlier if it differs from the average by more than a pre-determined number of standard deviations.

$$X_i > \bar{X} + sR \quad \text{or} \quad X_i < \bar{X} - sR$$

The R value to use will need to be based upon an analysis of previous survey data to identify the point at which an individual report makes a change in level or accounts for such a large part of the estimate that it has also individually affected the sampling error.

D. Estimation with Outliers Present

After an outlier has been identified, it still is necessary to compute an estimate. The detection procedure recommended in paragraph (c) does not require much information about the distribution of the data--except to determine the R values.

If the item being estimated is an average and its frequency distribution represents a normal distribution, the outlier observations can be made equal to pre-determined cut-off values. This is recommended rather than deleting them because there may not be an equal number of small and large outliers.

The situation is different, however, when estimating totals.

Some different procedures have been recommended. Suppose a sample of n has been selected from the population of N. Also, assume that t outliers were

identified using some of the rules discussed above. One estimate 3/ that can be generated is

$$\hat{Y}_1 = \sum_{i=1}^{t} y_i + \frac{N-t}{n-t} \sum_{t+1}^{n} y_i$$

This involves identifying the outliers and assigning them a weight of 1 assuming they were pre-selected. The remaining observations are expanded using expansion factors adjusted by the number of outliers found. The bias in this estimator is dependent on the number of outliers and the relationship between the means of outlier units to the mean of the non outlier unit.

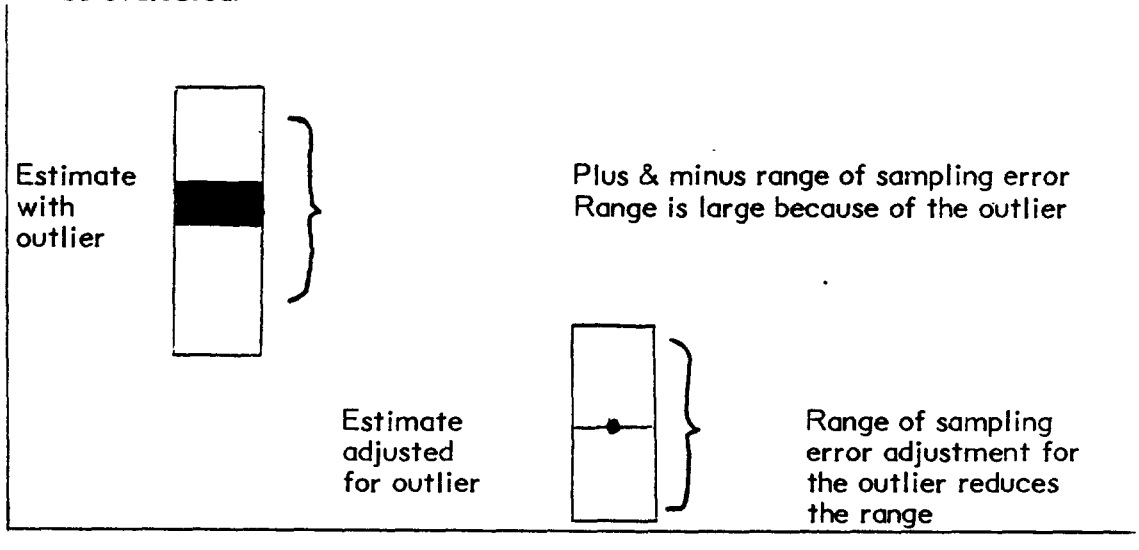
Another estimate is

$$\hat{Y}_2 = r \sum_{i=1}^{t} y_i + \frac{(N-rt)}{(n-t)} \sum_{t+1}^{n} y_i$$

This is similar to Y except that a weight (r) is applied to the outlier units.

The estimator $\hat{Y}_1$ is appropriate when the outlier is caused by an extremely large report while $\hat{Y}_2$ is appropriate when the outlier is caused by large expansion factors. Then the (r) value can be the weight the unit should have received if it had been classified correctly.

In all cases the estimate and the sampling error with outliers present should be computed. After adjustments for the outliers have been made, new estimates and sampling errors should be computed. These two estimates--unadjusted and the adjusted need to be evaluated together when determining the final estimate. The following graph illustrates how the two different estimates can be evaluated.



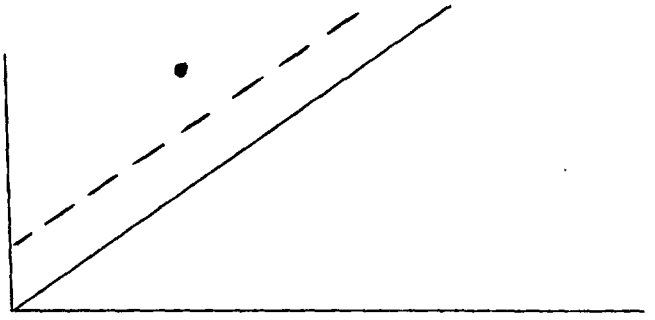
Hopefully, the ranges of the sampling errors of the adjusted and unadjusted estimates will overlap. The overlapping area should represent a compromise between the two. Remember that even though the outlier will probably cause the initial estimate to be too large, the adjustment procedure may cause an under estimate.

A final factor to remember is that what constitutes an outlier at the state level may not be an outlier at the regional or national level. As we saw from the initial example, the absence of an outlier can cause an under estimate just as its presence causes an over estimate. The point is that the outliers identified may not be outliers as far as the regional or national estimates are concerned and need to remain unadjusted at that level of summary. This then poses a dilemma, because if the state containing the outlier makes an adjustment, the remaining items have to be divided up among the other states.

D. Model Building and Regression Analysis

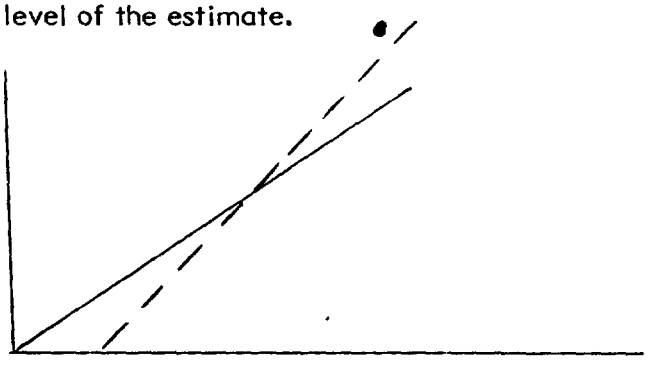
An outlier can affect regression analysis two ways

Figure a)



One observation can change can change the intercept of the line and thus affect the level of the estimate.

Figure b)



Here an outlier can change the slope of the line. This affects the level of the estimate and the level of change of y relative to X.

The outlier detected in figure (a) above was identified using the "deleted residual" (Gunst, 1980) procedure to determine how the predictor equation would change if the point in question were deleted. Basically this procedure involves computing repeated regression lines, each one derived with the i^{th} observation deleted. When each regression line is computed, a statistical test is conducted to determine if the deleted observation is part of the population represented by the line based on the other observations. If it is part of the population, it remains in the data set, if not, it is deleted.

The test to locate the outlier shown in figure (b) is based upon a very similar procedure. Again separate regressions are computed by deleting each i^{th} data point in turn. However, this time the statistical test compares the regression coefficient (B) based on the model using all data against the coefficient resulting after the ith data point is deleted. It the observation fails the test, i.e., the B value from the deleted data set is significantly different from the B value from the full data set, then the observation is deleted.

The use of such tests is fairly new because the numerous computations are only feasible with the use of high speed computers.

REFERENCES

Beckman, R. J. and **Cook, R.D.**, (1983), "Outliers" Technometrics, 25, 119-149

Gunst, Richard F. and **Mason, Robert L.**, Regression Analysis and Its Application Marcel Dekker, Inc., New York 1980

Hidiroglou, Michael A. and **Spirnath, Kodaba P.**, (1981) "Some Estimators of a Population Total from Sample Random Samples Containing Large Units, JASA, 76, 690-695

Roser, Bernard, (1983) "Percentage Points for a Generalized ESD Many Outlier Procedure", Technometrics, 25, 165-173

Searls, Donald T., (1966) "An Estimator for a Population Mean which Reduces the Effect of Large True Obervations", JASA, 61, 1200-1204

VIII. APPROXIMATING SAMPLE SIZES

Since sample sizes for many surveys can represent more than five percent of the population, sample size computations as described in many textbooks (which ignore the finite population) are not applicable. The sample size formulae which will be derived here can be used any time, but should be used when an indicated sampling rate exceeds five percent of the population. These derivations will ignore any stratification of the population (although the same technique would apply within a stratum) and will postulate a distribution of the characteristic within the sub-domain of operators processing the characteristic.

A. Estimating Sample Sizes for Means and Totals

To begin the derivation, recognize that the relative variance of an original variate X_i is

$$V^2 = S^2/\bar{X}^2 \text{ where } S^2 = \frac{\sum_{i=1}^{N}(X_i-\bar{X})^2}{N-1}$$

also the relative variance of $\bar{X}$ is

$$V_{\bar{x}}^2 = \frac{\sigma_{\bar{x}}^2}{\bar{x}^2} = \frac{(N-n)}{N}\, S^2/n\bar{x}^2 = \frac{(N-n)}{N}\, V^2/n$$

Working paper prepared by Dale Atkinson, Methods Staff, Statistical Reporting Service, U.S. Department of Agriculture

the relative variance of a total is

$$V^2_{x'} = \frac{\sigma^2_{x'}}{(X')^2} = \frac{N^2\frac{(N-n)}{N}S^2}{n(X')^2} = \frac{\frac{(N-n)}{N}S^2}{n\bar{X}^2} = \frac{(N-n)}{N}V^2/n$$

Therefore $V^2_{\bar{x}} = V^2_{x'}$ and the sample size calculations for means and totals are identical.

Specifically

$$V^2_{\bar{x}} = \frac{(N-n)}{N}V^2/n$$

$$n = \frac{(N-n)}{N}V^2/V^2_{\bar{x}}$$

$$= \frac{NV^2}{NV^2_{\bar{x}}} - \frac{nV^2}{NV^2_{\bar{x}}}$$

$$n + \frac{nV^2}{NV^2_{\bar{x}}} = \frac{NV^2}{NV^2_{\bar{x}}}$$

$$\frac{nNV^2_{\bar{x}} + nV^2}{NV^2_{\bar{x}}} = \frac{NV^2}{NV^2_{\bar{x}}}$$

$$n(NV^2_{\bar{x}} + V^2) = NV^2$$

$$n = \frac{NV^2}{NV^2_{\bar{x}} + V^2}$$

and since $V^2 \simeq \frac{V^2_r + 1-p}{p}$

(where V_r^2 is the relative variance of the original variate in the restricted domain of population units with the characteristic of interest and p is the proportion of population units possessing the characteristic)

$$n = \frac{N\left(\frac{V_r^2 + 1-p}{p}\right)}{NV_{\bar{x}}^2 + \left(\frac{V_r^2 + 1-p}{p}\right)} = \frac{N(V_r^2 + 1 - p)}{pN(CV)^2 + V_r^2 + 1-p}$$

At this point in the derivation it is necessary to postulate a distribution of the original variate within the domain of population units with the characteristic. Assuming a symmetric triangular distribution we have

$$V_r^2 = 1/6 \text{ (See Table 1)}$$

so that

$$n = \frac{N(\frac{1}{6} + 1-p)}{pN(CV)^2 + \frac{1}{6} + 1-p} = \frac{N(\frac{7}{6} - p)}{pN(CV)^2 + \frac{7}{6} - p}$$

Therefore, if we know the population size and can estimate the proportion of population units with the characteristic, we can calculate the required sample size for a desired level of precision. In order to calculate the sample size that would have been obtained by ignoring the finite population correction (i.e., assuming the population is essentially infinite) we can look at the limiting value of n as $N \to \infty$

$$\lim_{N\to\infty} n = \frac{\partial n}{\partial N} = \frac{\frac{7}{6} - p}{p(CV)^2}$$

To illustrate how much different the results of using or not using the finite population correction can be, consider an example in which the population size is 10,000, the proportion of population units with the characteristic of interest is .05. and the desired coefficient of variation is .05. Using the fpc we get 4719; however, if we ignore the fpc the indicated sample size is 8933.

* Deming, W. Edwards. Sample Design in Business Research. John Wiley & Sons, Inc. Copyright 1960.

TABLE 1. REPLICATED SAMPLING DESIGNS

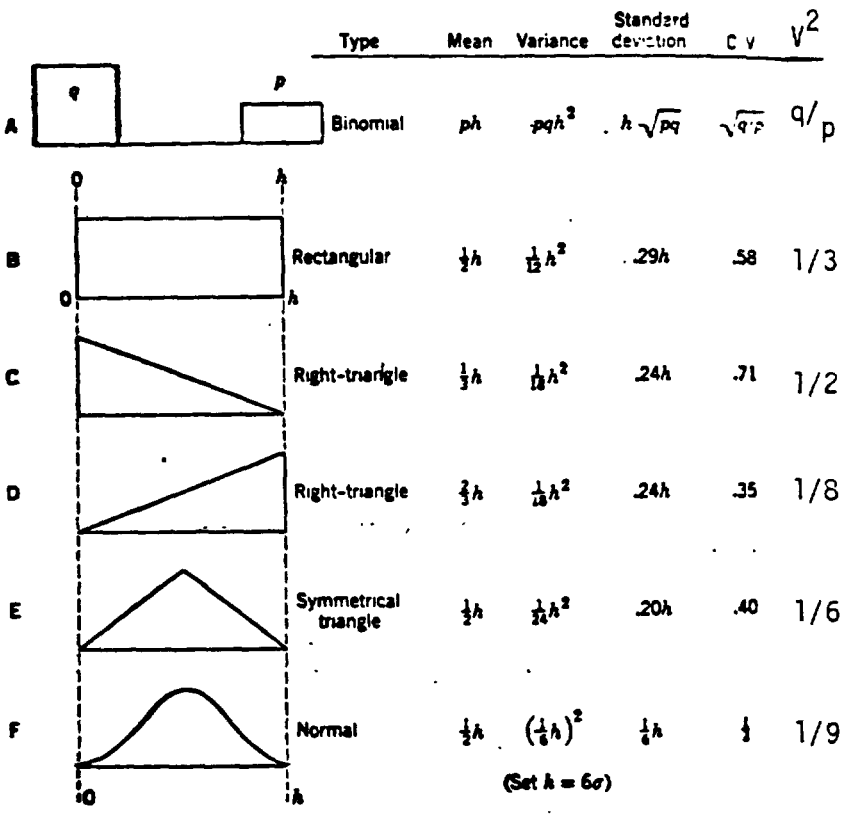


Fig. 16. New material can often be classed in advance roughly as binomial, rectangular, right-triangular, or triangular, and boundaries placed on the extremes. Quick reference to this figure will give a conservative value of σ on which to plan a sample. The range of variation is in every panel from 0 to h.

Estimating Sample Sizes for Ratios

The determination of sample sizes required to control the variation of a sample ratio for desired levels of precision will be broken into two cases, both of which occur in practical situations.

Case I - The Variables in the Numerator and Denominator Can be Assumed to Have the Same Relative Variance

Starting with an approximation for the relative variance of a ratio we have

$$V^2_{ratio} \approx \left(\frac{N-n}{N}\right)\left(\frac{V_x^2 + V_y^2 - 2\rho V_x V_y}{n}\right)$$

Assuming variables X and Y have the same relative variance and a correlation of .7.

$$= \left(\frac{N-n}{N}\right)\left(\frac{2V^2 - 1.4V^2}{n}\right)$$

$$= \left(\frac{N-n}{N}\right)\left(\frac{.6V^2}{n}\right)$$

$$= .6\left(\frac{N-n}{nN}\right)\frac{V_r^2 + 1-p}{p}$$

further assuming a symmetric triangular distribution in the restricted space

$$V_r^2 = 1/6$$

Therefore,

$$V^2_{ratio} = .6\left(\frac{N-n}{nN}\right)\frac{\frac{7}{6}-p}{p}$$

$$= .6\left(\frac{\frac{7}{6}N - Np - \frac{7}{6}n + np}{nNp}\right)$$

$$n = \frac{.7N - .6Np - .7n + 6np}{pNV^2_{ratio}}$$

$$\frac{pNV^2_{ratio}n + .7n - .6np}{pNV^2_{ratio}} = \frac{N(.7 - .6p)}{pNV^2_{ratio}}$$

$$\frac{n(pNV^2_{ratio} + .7 - .6p)}{pNV^2_{ratio}} = \frac{N(.7 - .6p)}{pNV^2_{ratio}}$$

Equating numerators and dividing both sides by the coefficient of n we have

$$n = \frac{N(.7 - .6p)}{pN(CV)^2 + .7 - .6p}$$

In order to calculate the sample size that would have been obtained by ignoring the fpc factor note that

$$\lim_{N \to \infty} \frac{\partial n}{\partial N} = \frac{.7 - .6p}{p(CV)^2}$$

Suppose as before that our population size is 10,000, the proportion of population units with the characteristic of interest is .05 and the desired coefficient of variation on the sample ratio is .05. Then using the fpc factor the indicated sample size is 3490. However, ignoring the fpc yields an indicated sample size of 5360.

CASE II - The Variable Y in the Denominator is an Auxiliary Characteristic Possessed by Virtually All Population Units

In this case we again start with

$$V^2_{ratio} \approx \left(\frac{N-n}{N}\right)\left(\frac{V_x^2 + V_y^2 - 2\varrho V_x V_y}{n}\right)$$

Assuming a correlation of variate X to auxiliary variable Y of .5 and a symmetric triangular distribution of Y, we have

$$= \left(\frac{N-n}{N}\right)\left(\frac{V_x^2 + \frac{1}{6} - 2\,(.5)\sqrt{\frac{1}{6}V_x^2}}{n}\right)$$

$$= \left(\frac{N-n}{nN}\right)\left[\left(\frac{V_r^2 + 1-p}{p}\right) + \frac{1}{6} - 2(.5)\sqrt{\left(\frac{1}{6}\right)\frac{V_r^2 + 1-p}{p}}\right]$$

$$= \left(\frac{N-n}{nN}\right)\left[\frac{V_r^2 + 1-p}{p} + \frac{1}{6} - \sqrt{\frac{V_r^2 + 1-p}{6p}}\right]$$

Further assuming a symmetric triangular distribution of X in the restricted space

$$= \left(\frac{N-n}{nN}\right)\left[\left(\frac{\frac{7}{6} - p}{p}\right) + \frac{1}{6} - \sqrt{\frac{\frac{7}{6} - p}{6p}}\right]$$

$$= \left(\frac{N-n}{nN}\right)\left[\left(\frac{7 - 5P}{6p}\right) - \sqrt{\frac{\frac{7}{6} - p}{6p}}\right]$$

Solving for n yields

$$n = \frac{N\left[\frac{7-5p}{6p} - \sqrt{\frac{\frac{7}{6} - p}{6p}}\right]}{N(CV)^2 + \frac{7 - 5p}{6p} - \sqrt{\frac{\frac{7}{6} - p}{6p}}}$$

To determine the sample size indication that would have been obtained if the fpc were ignored, again consider

$$\lim_{N \to \infty} n = \frac{\partial n}{\partial N} = \frac{\frac{7 - 5p}{6p} - \sqrt{\frac{7}{6} - \frac{p}{6p}}}{(CV)^2}$$

For our example in which N = 10,000 and cv = p = .05, considering the fpc we would get n = 4515. However, ignoring the fpc we would get n = 8228.

Estimating Sample Sizes for Proportions

Starting with an approximation for the relative variance of a proportion, we have

$$V_p^2 = \left(\frac{N-n}{N}\right) q/np \qquad \text{Where p is the proportion of successes and } q = 1 - p.$$

$$n = \left(\frac{N-n}{N}\right) \frac{q}{V_p^2 p} = \frac{q}{V_p^2 p} - \frac{nq}{NV_p^2 p}$$

$$n\left[1 + \frac{q}{NV_p^2 p}\right] = q/V_p^2 p$$

$$n = \frac{Nq}{NV_p^2 p + q}$$

$$\lim_{N \to \infty} n = \frac{\partial n}{\partial N} = \frac{q}{V_p^2 p}$$

Suppose for example we are sampling a frame of 10,000 population units, attempting to estimate with a cv of .05 a proportion whose true value is .05. Then considering the fpc we get n = 4318. However, if we were to ignore the fpc, the indicated sample size would be n = 7600.

TABLE 2. ESTIMATED SAMPLE SIZE OF ALL CROP PRODUCERS NECESSARY FOR A .05 LEVEL OF PRECISION FOR A SPECIFIC VARIETY 1/

Population Size	Proportion of All Growers Growing a Specific Variety							
	.05	.10	.20	.30	.40	.50	.75	.90
1000	900	811	660	537	434	348	182	106
1500	1285	1110	845	653	508	394	194	110
2000	1635	1362	984	733	555	422	200	112
2500	1954	1577	1091	791	587	440	205	114
3000	2246	1762	1176	835	611	453	207	115
5000	3206	2302	1394	939	665	482	213	116
7500	4077	2720	1537	1001	696	498	216	117
10000	4718	2990	1620	1036	712	506	217	117
14000	5453	3270	1699	1067	727	514	219	118

1/ To obtain the necessary sample size for a different level of precision, the following formula must be used:

$$n = \frac{N\left(\frac{1.16667-P}{P}\right)}{N(CV)^2 + \left(\frac{1.6667-P}{P}\right)}$$

WHERE N = population size
P = proportion of growers growing a specific variety
CV = desired precision or coefficient of variation

For example if the desired precision is .10, N = 3000, and P = .05,

$$n = \frac{3000\left(\frac{1.16667-.05}{.05}\right)}{3000(.10)^2 + \left(\frac{1.16667 - .05}{.05}\right)} = 1280$$

TABLE 3. SAMPLE SIZES REQUIRED OF THE OVERALL POPULATION FOR A .05 LEVEL OF PRECISION ON THE MEAN OR TOTAL OF A SUBDOMAIN (CHARACTERISTIC), ASSUMING THAT THE CHARACTERISTIC WITHIN THE SUBDOMAIN HAS A SYMMETRIC TRIANGULAR DISTIRUBITON

POPULATION SIZE	PROPORTION OF POPULATION POSSESSING CHARACTERISTIC OF INTEREST							
	.05	.10	.20	.30	.40	.50	.75	.90
100	99	98	96	93	89	85	69	55
500	474	448	398	349	303	259	154	96
1000	900	811	660	537	434	348	182	106
2500	1954	1577	1091	791	587	440	205	114
5000	3206	2303	1395	939	665	482	213	116
10000	4719	2991	1621	1036	713	507	218	118
15000	5599	3322	1713	1073	730	516	219	118
20000	6176	3517	1763	1093	739	520	220	118
30000	6884	3736	1817	1113	748	525	221	119
50000	7580	3932	1862	1130	756	528	222	119
80000	8036	4051	1888	1140	760	530	222	119
100000	8201	4093	1897	1143	761	531	222	119
120000	8315	4121	1903	1145	762	531	222	119
150000	8432	4149	1909	1147	763	532	222	119
200000	8552	4178	1915	1149	764	532	222	119

TABLE 4. SAMPLE SIZES REQUIRED OF THE OVERALL POPULATION FOR A .05 LEVEL OF PRECISION ON THE MEAN OR TOTAL OF A SUBDOMAIN (CHARACTERISTIC), ASSUMING THAT THE CHARACTERISTIC WITHIN THE SUBDOMAIN HAS A RIGHT TRIANGULAR DISTRIBUTION WITH BASE AT ZERO

POPULATION SIZE	PROPORTION OF POPULATION POSSESSING CHARACTERISTIC OF INTEREST							
	.05	.10	.20	.30	.40	.50	.75	.90
100	100	99	97	95	92	89	80	73
500	480	460	420	381	344	308	223	174
1000	921	849	723	616	524	445	286	211
2500	2057	1729	1275	976	764	607	345	241
5000	3494	2642	1711	1213	902	690	371	254
10000	5371	3590	2064	1380	991	741	385	260
15000	6542	4078	2216	1446	1025	760	390	263
20000	7342	4375	2301	1482	1043	770	393	264
30000	8366	4720	2393	1519	1062	780	395	265
50000	9416	5036	2472	1551	1077	788	397	266
80000	10132	5234	2519	1569	1086	793	399	266
100000	10395	5304	2535	1575	1089	794	399	266
120000	10578	5351	2545	1579	1091	795	399	267
150000	10768	5399	2556	1584	1092	796	399	267
200000	10965	5448	2567	1588	1094	797	400	267