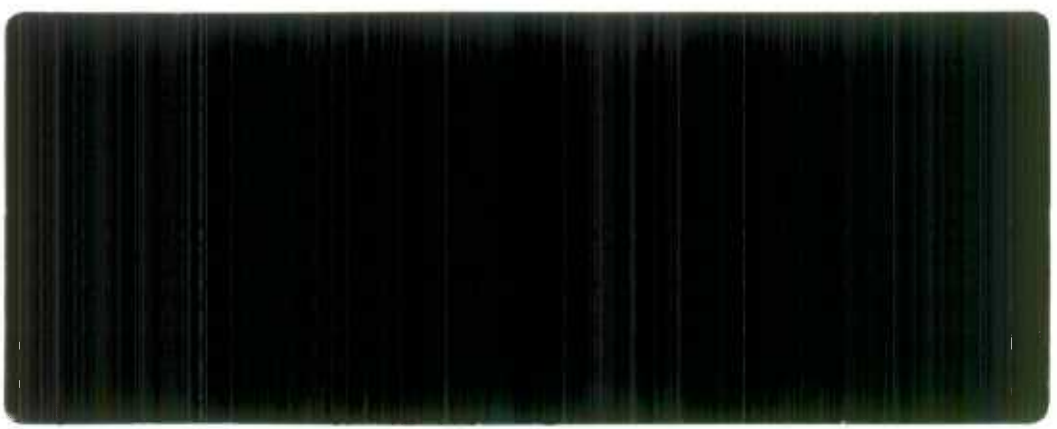
Agriculture Division

Division de l'agriculture



Statistics Canada Statistique Canada

Canada

Statistics Canada
Agriculture Division

WORKING PAPER #11

The Ratio Estimator:

an intuitive explanation
and its use in estimating agricultural variables

François Maranda, Business Survey Methods Division
Stuart Pursey, Agriculture Division
Statistics Canada
1992

Cat. No.: 21-6010MPE11200

The responsibility for the analysis and interpretation of the data is that of the authors and not of Statistics Canada.

Table of Contents

1 Introduction

During the 1989-90 fiscal year, the Business Survey Methods Division (BSMD), and staff from the Agriculture Division, began to analyze the usefulness of ratio estimation for the measurement of stocks of grain and egg production.

This objective of this note is to provide an intuitive explanation of ratio estimation to accompany the analysis completed by BSMD. Questions such as the following are addressed: What are the characteristics of a population that is well suited, or not well suited, for ratio estimation? How does stratification (as a sample design technique) compare and relate to ratio estimation (as an estimation technique)? How does ratio estimation fare when our sampling frame is no longer as good a representation of the population as it was originally?

Numerical (not statistical) examples are used to illustrate the use of ratio estimation. Finally, in section 7, basic conclusions about the usefulness of ratio estimation are listed.

2 The estimation objective

The objective is to estimate the acreage of wheat within a defined population. Throughout this report various notation and formulae are used:

True Values of the Population	
$Y =$	the (unknown) total acreage of wheat
$X =$	the (known) acreage of total land
$R =$	$Y / X =$ the (unknown) ratio of total wheat acreage to total land acreage
$N =$	the (known) number of farms

The sample	
i =	the index number of the farm
y_i =	the measured acreage of wheat on farm i
x_i =	the measured acreage of total land on farm i
n =	the number of farms sampled from the list frame

Estimators	
$\hat{Y}$ =	the simple estimate of Y
$\hat{Y}_S$ =	the stratified simple estimate of Y
$\hat{Y}_R$ =	the ratio estimate of Y (simple random sampling)
$\hat{Y}_{RS}$ =	the separate ratio estimate of Y (stratified sampling)
$\hat{Y}_{RC}$ =	the combined ratio estimate of Y (stratified sampling)

Stratified sampling	
L =	the number of defined strata in the frame
h =	the index number of the stratum
N_h =	the true (known) number of farms in stratum h
n_h =	the number of farms sampled in stratum h

The mean of the population is

$$\bar{Y} = \sum_{i=1}^{N} y_i / N \ ,$$

the total of the population is

$$Y = \sum_{i=1}^{N} y_i \ ,$$

and the variance of the population is

$$S^2 = \sum_{i=1}^{N} (y_i - \bar{Y})^2 / (N-1) \ .$$

3 Sampling and Estimation

We have decided on a probability sample, using simple random sampling, based on a (perfect) list frame.

3.1 Simple random sampling and the simple estimator

Under simple random sampling, each farm has an equal chance of being selected in the sample. The simple estimate of Y is:

$$\hat{Y} = (N/n) \sum_{i=1}^{n} y_i \ .$$

The simple estimate is unbiased. (An estimator is deemed unbiased if the average value of the estimate, taken over all possible samples of a given size, is exactly equal to the true population value.) Its sampling variance is:

$$V(\hat{Y}) \ = \ N^2 \left(1 - \frac{n}{N}\right) \frac{S^2}{n} \ .$$

This is the formula for the true value of the variance of $\hat{Y}$. Since $\hat{Y}$ is unbiased, we can state that if variance of the estimate is small then the simple estimate is "good". Thus the smaller the variance of $\hat{Y}$, the "better" the estimate. $V(\hat{Y})$ is small if its numerator is small or its denominator is large.

What type of population is such that $V(\hat{Y})$ is small? The y_i's appear in the numerator, within S^2. Therefore we want each of the quantities $(y_i-\bar{Y})^2$ in the summation to be small. This happens when each $(y_i-\bar{Y})$ is close to zero. So, set $(y_i-\bar{Y})=0$ and solve for y_i. This gives $y_i=\bar{Y}$. Therefore, S^2 is small if the y_i's are all close to the population mean, $\bar{Y}$. Our conclusion? The simple estimator works well when the farms in the population have similar acreages of wheat.

In a real survey situation we would want to develop an estimate for $V(\hat{Y})$.

3.2 Simple random sampling and the ratio estimator

Suppose we have two other pieces of information. We know, X, the total land acreage of the province. We know that the acreage of wheat on the farm, y_i, is related to the total land acreage on the farm, x_i. Also, we either know or are able to measure x_i on each farm in the sample. Given this knowledge, we can use it to improve the estimate of the acreage of wheat.

In our case, we assume that the relation between y_i and x_i is linear through the origin. By assuming a mathematical relation (i.e. a mathematical model) between y_i and x_i, we have changed our estimation objective. Instead of trying to estimate Y directly, we estimate the parameters of the chosen model. This leads indirectly to an estimate of Y -- we get it by arithmetic calculations based on the estimated parameters of our chosen model.

Clearly there are an infinite number of possible mathematical models. We choose a model because we feel it provides a correct structure or framework for reality. If the structure is "correct enough", our estimates improve, if not, they do not improve -- and they may be worse than if we had ignored "structure" altogether.

In our case, we have chosen the model:

$y_i = R x_i$, where the parameter R is to be estimated.

When we have estimated R, by $\hat{R}$, we will calculate $\hat{Y}_R$ by the formula:

$\hat{Y}_R = \hat{R} X$, where $\hat{R}$ has been estimated and X is known.

In our case, y_i is the acres of wheat and x_i is the acres of total land. Therefore R must be some value between 0 and 1 (inclusive). We can interpret R as a percentage -- the percentage that the total wheat acreage makes up of the total land acreage.

If we work through the derivation of an estimate for R, we get a very simple result. It is the simple estimate of Y divided by the simple estimate of X (as discussed in section 3.1 under simple random sampling). Thus we estimate Y and X by:

$$\hat{Y} = (N/n) \sum_{i=1}^{n} y_i \ ,$$

$$\hat{X} = (N/n) \sum_{i=1}^{n} x_i \ .$$

Therefore $\hat{R} = \hat{Y} / \hat{X} =$

$$= \frac{(N/n) \sum_{i=1}^{n} y_i}{(N/n) \sum_{i=1}^{n} x_i}$$

which reduces to

$$\frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} .$$

Thus the estimate of R is the ratio of wheat acres to total land acres (based on farms selected in the sample). Again, in our situation, we can interpret this as a percentage. Note that the raising factors cancel each other out and thus play no part in the estimation. Following on, we get the ratio estimate of Y as:

$$\hat{Y}_R = \hat{R} X, \text{ where } X \text{ is known.}$$

How well does this estimator work? Or equivalently, how well does the model work? We need to test the adequacy of the model we have assumed. We can do this using the traditional methods of regression analysis. In this note we examine "adequacy" by reference to the variance of $\hat{Y}_R$. Also, we are going to examine ratio estimation by way of the artificial example in section 4.

Statistically, the ratio estimator is biased, although the bias is negligible in large samples. (Recall that an estimator is deemed unbiased if the average value of the estimate, taken over all possible samples of a given size, is exactly equal to the true population value.)

The sampling variance of this estimator is:

$$V(\hat{Y}_R) = N^2 \frac{\left(1-\frac{n}{N}\right)}{n} \frac{\left(\sum_{i=1}^{N}(y_i - Rx_i)^2\right)}{(N-1)} .$$

What type of population is such that $V(\hat{Y}_R)$ is small? The y_i's appear in the numerator: thus we want each of the quantities $(y_i - Rx_i)^2$ in the summation to be small. This happens when each $(y_i - Rx_i)$ is close to zero. So, set $(y_i - Rx_i) = 0$ and solve for y_i/x_i. This gives $y_i/x_i = R$. R is the true ratio of total wheat acreage to total land acreage in the population. Therefore we want each ratio, $R_i = y_i/x_i$, on each farm, to be R. Thus the ratio estimator works well when the ratio is similar from farm to farm.

3.3 Stratified simple random sampling

Finally, let's suppose that the acreage of wheat is similar within strata, as defined by some stratification variable (such as a size of farm variable). As well, we are able to assign each farm in the province (i.e. the frame) to a stratum and thus correctly count the number of farms, N_h, in each stratum.

The most obvious strategy is to take an independent simple random sample from each stratum. This estimator considers each of the strata to be its own "sub-population". The simple

estimate is calculated for each stratum and the results are added up to give $\hat{Y}_S$. The simple estimate $\hat{Y}_h$ for each of the strata: $h = 1, 2,$ and is 3.

$$\hat{Y}_1 = (N_1/n_1) \sum_{i=1}^{n_1} y_{1i} .$$

$$\hat{Y}_2 = (N_2/n_2) \sum_{i=1}^{n_2} y_{2i} .$$

$$\hat{Y}_3 = (N_3/n_3) \sum_{i=1}^{n_3} y_{3i} .$$

Therefore

$$\hat{Y}_S = \sum_{h=1}^{L} \left((N_h/n_h) \sum_{i=1}^{n_h} y_{hi} \right) .$$

Note that we must know the N_h and n_h for each stratum -- that is, the population size of the stratum h and the number of farms we sample from the stratum h. Thus we must be able to assign each farm in the frame a stratum.

The stratified sample design is useful if we know that within a stratum farms are fairly consistent -- but between strata, the farms are not consistent. The sample design ensures that we do, in fact, get a sample from each strata.

This design is also useful if we must publish statistics by stratum -- where the strata are defined by province, type of farm, etc. This sample design ensures that we get a sample for each stratum of interest.

The variance of this estimator is:

$$V(\hat{Y}_S) = \sum_{h=1}^{L} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_h^2}{n_h} .$$

Note that the variance $V(\hat{Y}_S)$ depends on the S_h^2, the variances of each stratum. It does not depend on S^2, the variance of the population as a whole.

We want each of the S_h^2's to be small. Use the same reasoning as for simple estimation: S_h^2 is small if the y_i's within strata are similar from farm to farm. Our conclusion? The stratified simple estimator works well when the farms in the population have similar acreages of wheat within strata. The y_i's do not have to be similar among farms of different strata. This conclusion is more or less the same as we got for the simple estimate -- we have just qualified that conclusion "by stratum".

4 The example

The examples have been set in such a way that they illustrate the types of populations that are "best" for the particular estimators. We choose a specific sampling design and estimator because it is "best" in some sense. Our definition of "best" is influenced by a number of issues -- the budget, the type of user needs, our computer data processing capability, etc.

We must also define "best" in a number of statistical ways. For example, we may ask that the estimator be unbiased, be consistent, provide the smallest variance under certain conditions, etc.

Defining "best" is not a simple procedure. First, the mathematical statistical theory can be intricate and subtle. Second, statisticians themselves do not always agree on what is best.

For the purpose of the illustrations here, we take "best" to be the estimator that provides the lowest Mean Square Error (MSE), taken over all possible samples of a given size. The squared error is: (the estimate minus the true value)^2. The MSE is the average of all the squared errors. We leave it to the reader to calculate the MSE and thus "prove" my conclusions.

4.1 ABOTINAM -- the simple estimator

ABOTINAM 12 farms wheat 27 acres land 64 acres												
	A	B	C	D	E	F	G	H	I	J	K	L
Wheat	12	2	4	3	1	0	.5	1	.5	2	1	0
Land	16	8	8	16	2	2	2	2	2	2	2	2
% wht	75%	25%	50%	19%	50%	0%	25%	50%	25%	100%	50%	0%

In ABOTINAM, it is apparent that there is no pattern to the data. In other words, there is no "extra information" that we can use to improve the estimate. We resort to simple random sampling and the simple estimate. Suppose that we had picked up farm B, farm D, and farm H by chance. From the formula in section 3.1 our estimate is:

$$\hat{Y} = \frac{N}{n} \sum_{i=1}^{n} y_i$$

$$= (12/3) * (2 + 3 + 1) = 24 \text{ acres.}$$

This estimate happens to be in error by 3 acres.

4.2 ATREBLA -- the ratio estimator

ATREBLA 12 farms wheat 32 acres land 64 acres												
	A	B	C	D	E	F	G	H	I	J	K	L
Wheat	8	4	4	8	1	1	1	1	1	1	1	1
Land	16	8	8	16	2	2	2	2	2	2	2	2
% wht	50%	50%	50%	50%	50%	50%	50%	50%	50%	50%	50%	50%

In ATREBLA it is apparent that each farm has a similar percentage of wheat. It is always 50% of its total land. Also, we know the total land of ATREBLA: 64 acres. This type of situation works well for ratio estimation. The auxiliary variable, which we are able to measure on each sampled farm, is total land, denoted by x_i.

Let's again suppose that we had picked up farms B, D, and H by chance. From section 3.2 the ratio estimate is:

$$\hat{Y}_R = \hat{R} X, \text{ where } X \text{ is known}$$

$$= \left(\sum_{i=1}^{n} y_i \ / \ \sum_{i=1}^{n} x_i \right) X$$

$$= ((4 + 8 + 1) / (8 + 16 + 2)) * 64$$

$$= (13/26) * 64 = .5 * 64 = 32 \text{ acres.}$$

Notice that our estimate has no error. In fact, given any sample our estimate will always equal 32 acres. This is because each farm has exactly the same percentage of acres of wheat, (50%). From this, one might guess that the more consistent the percentage of wheat among the population, the better the ratio estimator works. More generally, one might guess that ratio works best when y_i and x_i are highly correlated.

One might judge how well a ratio estimator is going to work by checking to see how consistent the ratio y_i/x_i is within the population. The more "statistical methods way" is to use the tools of regression analysis to judge the adequacy of our chosen model -- in this case a "linear regression through the origin" model.

One might wonder if ratio estimator generally provides better estimates than the simple estimator. This, it turns out, is not the case. There is a point, based on the value of the correlation between y_i and x_i and the coefficients of variation of y_i and x_i, where the ratio estimator and the simple estimator are equally "best". At this point, if y_i and x_i are any less correlated it is best to use the simple estimator and if y_i and x_i are any more correlated, it is best to use the ratio estimator. See page 158 and 159 of the reference for a detailed discussion of this issue.

In our situation we could analyze the 1986 Census of Agriculture data to see if the amount of wheat is a relatively consistent percentage of total land from farm to farm within a province. More rigorously, we could examine the population using statistical methods and then choose the best approach.

4.2.1 Using a "not best" estimator in ATREBLA

Suppose we had used the simple estimate instead of the ratio estimate for ATREBLA, again with farms B, D, and H. Our estimate would be

$$\hat{Y} = \frac{N}{n} \sum_{i=1}^{n} y_i ,$$

$$= (12/3) * (4+8+1) = 52 .$$

This estimate is in error by 52 - 32 = 20 acres. By trying other samples of 3 using the simple estimate and the ratio estimate, one can see that the ratio estimator works best when the population is like that in ATREBLA.

4.2.2 Using a "not best" estimator in ABOTINAM

In ABOTINAM, suppose we had not used the simple estimate. Instead we had used the ratio estimator -- despite there being no evidence to assume our "linear regression through the origin" model. The ratio estimate (again assuming farms B, D, and H were selected) is:

$$\hat{Y}_R = \hat{R} X$$

$$= (2+3+1)/(8+16+2) * 64 = (6/26) * 64 = 14.77 \text{ acres.}$$

This estimate is in error by 27 - 14.77 = 12.23 acres -- worse than our error of 3 acres using the simple estimator.

Perhaps we may get a lucky sample and get a good estimate despite using an inappropriate estimator. ABOTINAM has 27 acres of wheat. This is 42% of its total land. A lucky sample (from the point of view of the ratio estimator) will provide us with farms such that the total of the sampled wheat acreage divided by the sampled total land acreage is close to 42% Then, by luck alone, we will get $\hat{R}$ close to 42% and thus get close to an estimate of 27 acres for Y.

4.3 NAWEHCTAKSAS -- the stratified simple estimator

NAWEHCTAKSAS 12 farms wheat 36 acres land 68 acres												
	A	B	C	D	E	F	G	H	I	J	K	L
Wheat	12	4	4	12	.5	.5	.5	.5	.5	.5	.5	.5
Land	16	8	10	20	2	2	2	1	2	2	1	2
% wht	75%	50%	40%	60%	25%	25%	25%	50%	25%	25%	50%	25%

In NAWEHCTAKSAS we have defined three strata: (small farms with 1 or 2 acres of land, medium farms with 8 or 10 acres of land, and large farms with 16 or 20 acres of land. It is apparent that within the strata farms have similar acreages of wheat. We should take advantage of this knowledge and design a sample that ensures that we get a sample from each strata.

We still plan to sample 3 farms in total, but we will ensure that we sample from each of the strata. We decide on a simple random sample within each stratum -- a sample size of 1 from each stratum. Suppose we get farms B, D, and H by chance (a sample size of 1 from each stratum).

Determine the simple estimate for each stratum:

$$\hat{Y}_{S_1} = (8/1) * .5 = 4,$$

$$\hat{Y}_{S2} = (2/1) * 12 = 24,$$

$$\hat{Y}_{S3} = (2/1) * 4 = 8,$$

and then add them up to get the stratified simple estimate:

$$\hat{Y}_{Sh} = \sum_{h=1}^{L}\left((N_h/n_h)\sum_{i=1}^{n_h} y_i\right)$$

$$= 4 + 24 + 8 = 36 \text{ acres.}$$

In this population, the acreage of wheat is constant within stratum. Because of this, no matter what farms get picked in our stratified sample and no matter how inconsistent farms are between strata, we will always get an estimate with no error. Note that the raising factors differ among the strata. As always in simple random sampling, the raising factor is the number of farms in the population (or sub-population) divided by the number sampled in the population (or sub-population).

5 Stratified simple random sampling and the ratio estimate

In NAWEHCTAKSAS, the situation was perfect for stratified simple random sampling because the acreage of wheat was constant within strata. Note that the percentage of wheat (used in ratio estimation) was not constant within the population. It varied from a low of 25% to a high of 75%. Thus a ratio estimator will not work perfectly here. In fact, given that the percentages vary quite widely, one would guess that the ratio estimate would not work well at all.

However, we notice that among the "small farm" stratum the percentage varied from 25% to 50%, among the "medium farm" stratum the percentage varied from 40% to 50%, and among the "large farm" stratum the percentage varied from 60% to 75%. The percentage of wheat is somewhat consistent within strata -- but varies widely between strata. In this type of situation, a "stratified ratio estimator" may be useful. (Of course, in this case we had "fixed it", so that the stratified simple estimate was best.)

5.1 The separate ratio estimator

Suppose that we had the following population -- we have fixed it so that it will work perfectly with a "stratified ratio estimator". we have changed the data farms "A" "C", "D", "H", and "J". The three strata are defined as before: (small farms with 1 or 2 acres of land, medium farms with 8 or 10 acres of land, and large farms with 16 or 20 acres of land. The acreages of wheat are no longer constant within strata -- but the percentage of wheat is. (The true total for the acreage of wheat is now 40 acres and total land is now 70 acres.)

12 farms wheat 40 acres land 70 acres												
	A	B	C	D	E	F	G	H	I	J	K	L
Wheat	15	4	5	12	.5	.5	.5	.5	.5	.5	.5	.5
Land	20	8	10	16	2	2	2	2	2	2	2	2
% wht	75%	50%	50%	75%	25%	25%	25%	25%	25%	25%	25%	25%

As with the stratified simple random sample, described above in section 5.3, select a sample of size one from each of the three strata. However, within strata, do not use the simple estimator -- use the ratio estimator instead. We have, as farms B, D, and H are again selected by chance in the sample:

$$\hat{Y}_{SR_1} = \hat{R}_1 X_1, \text{ where } X_1 \text{ is known}$$

$$= (4/8) * 18 = 9 ,$$

$$\hat{Y}_{SR_2} = \hat{R}_2 X_2, \text{ where } X_2 \text{ is known}$$

$$= (12/16) * 36 = 27 ,$$

$$\hat{Y}_{SR_3} = \hat{R}_3 X_3, \text{ where } X_3 \text{ is known}$$

$$= (.5/2) * 16 = 4 .$$

Therefore,

$$\hat{Y}_{RS} = \hat{Y}_{SR_1} + \hat{Y}_{SR_2} + \hat{Y}_{SR_3}$$

$$= 9 + 27 + 4 = 40 \text{ acres.}$$

This estimator is actually called the separate ratio estimator, probably because separate ratio estimates are calculated for each stratum. This estimator works well when y_i and x_i are highly correlated within stratum -- they do not have to be correlated at all between strata.

Note that the raising factors do not come into play. They cancel out (being in both the numerator and the denominator). Note also that we must know the true value of X_1, X_2, and X_3. These are the true and known values of the total land acreage within each stratum.

5.2 The combined ratio estimator

In practice, it is often the case that we do not know the true values of the acreage of land for each stratum. In this situation we use the "combined" ratio estimator. This estimator first calculates the stratified simple estimates $\hat{X}_S$ and $\hat{Y}_S$. Then the estimate of the ratio is $\hat{Y}_S / \hat{X}_S$. In our example we will get:

$$\hat{Y}_S = \hat{Y}_{S_1} + \hat{Y}_{S_2} + \hat{Y}_{S_3}$$

$$= (2/1)*4 + (2/1)*12 + (8/1)* .5$$

$$= 8 + 24 + 4 = 36 \text{ acres}$$

and

$$\hat{X}_S = \hat{X}_{S_1} + \hat{X}_{S_2} + \hat{X}_{S_3}$$

$$= (2/1)*8 + (2/1)*16 + (8/1)*2$$

$$= 16 + 32 + 16 = 64 \text{ acres.}$$

The combined stratified ratio estimator is:

$$\hat{Y}_{CR} = \hat{R}_{CS} X$$

$$= (36/64) * 70 = .56 * 70 = 39.20 \text{ acres.}$$

This estimate is in error by 0.8 acres -- not too bad an estimate considering that we don't know the true values of total land by stratum.

It is important to recognize that if we do not have the true value for X_h for each stratum, we must use the combined stratified ratio estimator. It requires knowledge of X, not X_h. Note also that the stratum raising factors must be used in the formula for the combined ratio estimator.

(However, there is one stratified sample design in which the raising factors will cancel out in the numerator and denominator of the estimate of R. This occurs if we allocate our sample proportionally among the strata according to the size of the sub population, N_h. It is called a "self-weighting" sample design. However, this type of proportional allocation is usually inefficient for skewed populations, as is often the case in agriculture.)

6 Stratified sampling versus the ratio estimator

Sometimes we may run into a population that is somewhat between ATREBLA (best for ratio estimation) and NAWEHCTAKSAS (best for stratified sampling). The population is not "perfect" for either approach. Should we use ratio estimation or stratification? Cochran, page 169, (see reference) notes that:

"Stratification by size of farm accomplishes the same general purpose as a ratio estimate in which the denominator [x_i] is farm size." Both devices diminish the effect of variations in farm size on the sampling error of the estimated mean corn per acres per farm."

Cochran is pointing out that in this case, stratified sample design and the ratio estimator do the same thing. When an auxiliary variable is useful (as a stratification variable or for use in the ratio estimator) we have a choice -- use stratified sampling or use ratio estimation.

Which one should we use? Should we approach the problem from the sample design end or the estimation methods end? There are some factors to consider.

Geographic location is more easily addressed through the sample design than the estimation method.

Ratio estimation also depends on the nature of the relation between y_i and x_i. We assume a linear through origin relation between y_i and x_i. Cochran, page 169, notes:

> "With a complex or discontinuous relation, stratification may be more effective, since, if there are enough strata, stratification will eliminate the effects of almost any kind of relation between y_i and x_i."

However, suppose a variable of interest (such as acres of wheat) is related to a certain auxiliary variable (such as total land), and another variable of interest (such as number of bulls) is related to a different auxiliary variable (such as total cattle and calves). It may better, in this case, to use two ratio estimators within one sample design -- rather than to stratify on total land or on total cattle and calves.

Finally, we can use both at the same time: the combined ratio estimator as described in section 5.1.

7 The list frame

7.1 Deficiencies

Up until now we have assumed that the list frame is perfect. Let us consider three cases. As before, our objective is to estimate Y, based on the y_i selected in the sample. Again we use simple random sampling with the simple estimate.

Case 1: There are farms in the population that do not appear on the frame.

We might consider that the population has "split" into two sub-populations: "ON" (those farms on the frame) and "OFF" (those farms off the frame). "ON" has a sample; thus an estimate is produced. However, "OFF" has no sample and consequently an estimate is not produced.

The result is that the estimate of the population total is biased downwards: we want $\hat{Y}_{OFF}+\hat{Y}_{ON}$ but get $\hat{Y}_{ON}$. Since $\hat{Y}_{ON}$ is unbiased (under simple random sampling with the simple estimate), then the bias of this estimate is $Y-Y_{ON}=Y_{OFF}$.

As well, note that the sample design was based entirely on the farms in "ON", not "ON"+"OFF", and thus is optimally designed for the estimation of $\hat{Y}_{ON}$.

Case 2: There are farms on the frame that do not exist in the population.

Here we might consider that the frame has "split" into two sub-groups: "IN" (those farms **in** the population) and "OUT" (those farms **out** of the population). The sample design is based on "IN"+"OUT" and therefore is not optimal.

The farms in "OUT", that happen to be selected in the sample, provide no data. (Hopefully, we will not impute for the "missing data" -- it's "missing" because it's not there.)

Recall from the description of the simple estimate in section 3.1 that we used N, n, and y_i . Here we would wish to do the calculations based on those y_i belonging to "IN". This requires knowing N_ϵ and n_ϵ, where the subscript ϵ refers to counts based on the y_i belonging to "IN". We are unable, however, to calculate the simple estimate because we do not know the value of N_ϵ.

The alternative method of estimation is as follows. For any farm in "OUT" selected in the sample, we set $y_i = 0$. For any farm in "IN", selected in the sample, we leave the data value y_i unchanged.

Then the estimate of Y is:

$$\hat{Y}_\epsilon = (N/n) \sum_{i=1}^{n} y_i$$

The sampling variance of Y_ϵ is:

$$V(\hat{Y}_\epsilon) = N^2 \left(1 - \frac{n}{N}\right) \frac{S^2}{n}$$

These formulae are identical to those in section 3.1, where y_i that has been modified, as described above.

The approach here is identical to the approach taken to provide estimates for "domains of study"; as described by Cochran, (see reference 1; pages 34 to 38). As Cochran explains, the estimator $\hat{Y}_\epsilon$ is unbiased.

If we were able to count the number of farms that should not be in the frame, we could replace N and n by N_ϵ and n_ϵ. This is an advantage because it makes sampling variance of Y_ϵ smaller. Cochran, on page 38, shows the gain in efficiency.

Case 3: The frame, although it contains all the farms in the population -- no more and no less -- has become out of date. Through time, some of the farms have jumped to a new stratum. With the change in reality, the sample design is no longer optimal.

Normally, in agriculture surveys, the small farms are put in a stratum with a low sampling rate (and thus a large raising factor). The large farms are put in a stratum with a high sampling rate (and thus a small raising factor).

Suppose a farm, originally quite small, had been correctly placed a lightly sampled stratum. But, through time, it has grown into a large farm. We are stuck with its original raising factor (quite a high one) and must therefore multiply its now large data values by an already high raising factor. Often these farms come out on the "top contributor to the estimate" list and we are tempted to artificially reduce the data value or its raising factor.

Is this the right thing to do? This farm, that jumped to a new stratum, represents itself and farms in the population that were not sampled. If its raising factor was 200 (a sampling rate of 1/200) then it is meant to represent 200 farms in the population. Let's suppose that there were 30,000 farms in the stratum and 8,000 of them became large and jumped into the new stratum. With a sampling rate of 1/200 we expect to pick up 40 of these "stratum jumpers" in the sample. These 40 farms represent themselves and the other 7,960 non-sampled farms in the population that grew and jumped to a new stratum. Therefore, we do not want to lower -- or change -- data values or raising factors. If we lower them, we no longer represent those other 7,960 farms and consequently we bias the estimate downwards. Although still unbiased, this sample design with its original raising factors is no

longer optimal; we would prefer to redesign the sample based on the new information and take a fresh and independent sample.

Is there ever a situation where it is better to change the data values or raising factors? Our sampling rate is meant to be high enough to provide "good enough" estimates. Suppose that in the example only 5 farms (not 8,000) jumped to a new large stratum. We are depending on a sampling rate of 1/200 to estimate for this small sub-population of just five stratum jumpers. This is really too small a sampling rate to handle the situation. By "bad luck" we may actually pick up one of these farms in the sample and go ahead and use the raising factor of 200. The sample design "thinks" that there are 199 other stratum jumpers in the population but in reality there are just 4 others. In this case it is better that we intervene and make subject matter adjustments to the estimate.

Finally it is possible that a mixture or any or all of the above cases may occur.

7.2 The impact of these deficiencies on ratio estimation

The ratio estimator, under simple random sampling, does not use the raising factors. If we can assume that errors in the list frame (missing farms and incorrect information) are not related to the "percentage of wheat acres on the farm", then one might say that the ratio estimator is robust to problems in a list frame. What helps us here is X. In a sense, X is a control total -- something that helps keep the estimate of the total where it should be, despite difficulties with the frame. Still, a key point remains -- the population must be suited for ratio estimation in the first place -- then we can take advantage of this robustness.

Most often, it is not desirable to use simple random sampling. Instead we stratify the sample for a number of reasons. Often, we must publish by geographic regions and thus must stratify by geographic region. As well, it is sometimes risky to assume a "linear through the origin" relationship. Instead we usually opt for stratification -- that is, we look for consistency within strata -- rather than assume a specific mathematical relation.

Still, it is well worth exploring our population to see if there are good gains to be made by an estimator such as the ratio estimator. Experience has shown, over many years and many populations, that the ratio estimator is often an appropriate estimator. Recently BSMD staff developed a ratio estimator for the estimation of egg production numbers. As well, BSMD is analyzing the use of the ratio estimator for the estimation stocks of grain -- preliminary results are encouraging.

8 Conclusions

1. Stratification, if done correctly, may largely accomplish the purpose of the ratio estimation. Conversely, ratio estimation may largely accomplish the purpose of stratification.

   The choice depends on a number of factors: the characteristics of the population that is sampled, the benefit-to-cost ratio of implementing and maintaining the sample designs and estimation methods are two examples.

   Often, in multipurpose surveys, stratification serves as a good general purpose tool. Ratio estimation is a more specific tool, being most useful when there is a high linear correlation between y_i and x_i and the total X is known.

2. It is sometimes worthwhile to combine the use of stratification (in the sample design) and ratio estimation (in the estimation method). Ratio estimation, in this case, might serve to further improve the results obtained through stratification.

3. In surveys at Statistics Canada, it is rarely advisable to employ a simple random sampling design. With reference to agriculture, we usually stratify to allow for estimates by geographic region and types of agricultural commodities.

4. Efficient and accurate stratification requires an accurate frame. This implies that adequate resources be allocated to the development and maintenance of the quality of the frame.

5. Ratio estimation, under simple random sampling with no stratification, does not require the use of raising factors. It may be argued that the ratio estimator, under simple random sampling, is resistant to certain problems with the frame difficulties -- those that affect the quality of the raising factors, such as a shrinking frame. However, we continue to require knowledge of X (or at least a good estimate of it) and a "strong enough" linear correlation between y_i and x_i.

REFERENCE

Cochran, William G. (1977); Sampling Techniques, 3rd Edition; John Wiley & Sons

Statistics Canada

Agriculture Working Papers

Registration Number	*Title of Agriculture Working Paper (Product No. 21-601OMPE)*	*Price*
01000	Stuart Pursey, A Description of Theil's RMSPE Method in Agricultural Statistical Forecasts	$5.00
03000	Bernard Rosien and Elizabeth Leckie, A Review of the Livestock Estimating Project with Recommendations for the Future	$5.00
04000	Glenn Lennox, An Overview of the Canadian Oilseed Industry	$5.00
05000	Lambert Gauthier, Preliminary Analysis of the Contribution of Direct Government Payments to Realized Net Farm Income	$5.00
06000	Jean B. Down, Characteristics of Farm Entrants and their Enterprises in Southern Ontario for the years 1966 to 1976 (1984)	$5.00
07000	Allister Hickson, A Summary of Commodity Programs in the United States (1984)	$5.00
08000	Les Macartney, Prairie Summerfallow Intensity: An Analysis of 1981 Census Data (1984)	$5.00
09000	Mike Shumsky, The Changing Profile of the Canadian Pig Sector (1985)	$5.00
10000	Mike Trant, Revisions to the Treatment of Imputed House Rents in the Canadian Farm Accounts 1926-1979 (1986)	$10.00
11200	François Maranda and Stuart Pursey, The Ratio Estimator: an intuitive explanation and its use in estimating agricultural variables (1992)	$10.00
12100	Rick Burroughs, The Impact of Geographic Distortion due to the Headquarters Rule (1991)	$5.00
13100	Stuart Pursey, The Quality of Agriculture Data: Strengths and Weaknesses (1991)	$5.00
14200	Professor A.M. Fuller, Derek Cook and Dr. John Fitzsimons, Alternative Frameworks for Rural Data (1992)	$10.00
15300	Brian Biggs, Ray Bollman and Michael McNamee Trends and Characteristics of Rural and Small Town Canada (1993)	$10.00
16200	Philip Ehrensaft and Ray Bollman The Microdynamics and Farm Family Economics of Structural Change in Agriculture	$10.00
17100	Livestock and Animal Products Section Grains and Oilseeds Consumption by Livestock and Poultry Canada and Provinces 1992	$50.00
18000	Ray Bollman, Leslie A. Whitener, Fu Lai Tung Trends and Patterns of Agricultural Structural Change : A Canada / U.S. Comparison	$5.00
19000	Sayed Rizvi, David Culver, Patti Negrave and Lina DiPietro Total Farm Family Income by Farm Type, Region and Size, 1991	$10.00
20000	George McLaughlin Adjustment in Canadian Agriculture	$10.00
21000	Fred Gale and Stuart Pursey Microdynamics of Farm Size Growth and Decline: A Canada-United States Comparison	$5.00
22000	Leonard Apedaile, Charles Barnard, Ray Bollman and Blaine Calkins The Structures of Agricultural Household Earnings in North America: Positioning for Trade Liberalization	$5.00
23000	Glenn Zepp, Charles Plummer and Barbara McLaughlin Potatoes: A Comparison of Canada/USA Structure	$5.00
24000	Victor J. Oliveira, Leslie A. Whitener and Ray Bollman Farm Structure Data: A U.S.-Canadian Comparative Review	$5.00
25000	Karen Gray Grain Marketing Statistics Statistical Methods Working Paper Version 2	$10.00
26000	W. Steven Danford Farm Business Performance: Estimates from the Whole Farm Database	$5.00
27000	Brian Biggs An Attempt to Measure Rural Tourism Employment	$5.00

Statistics Canada

Agriculture Working Papers

MAIL TO:

Please ship to: (Please print)

Organization: _____

Department: _____

Attention: _____

Address: _____

City: _____

Province: _____ Postal Code: _____

Telephone: _____ Facsimile: _____

METHOD OF PAYMENT

☐ My remittance made payable to the Receiver General for Canada is enclosed.

☐ Charge my MASTERCARD

☐ Charge my VISA

Account No. _____

Expiration Date _____

Name of Card Holder (print) _____

_____

Signature _____

Registration Number	*Title of Agriculture Working Paper (Product No. 21-6010MPE)*	*Price*	*Quantity*	*Total $*
17100	Livestock and Animal Products Section Grains and Oilseeds Consumption by Livestock and Poultry Canada and Provinces 1992	$50.00		
18000	Ray Bollman, Leslie A. Whitener, Fu Lai Tung Trends and Patterns of Agricultural Structural Change : A Canada / U.S. Comparison	$5.00		
19000	Sayed Rizvi, David Culver, Patti Negrave and Lina DiPietro Total Farm Family Income by Farm Type, Region and Size, 1991	$10.00		
20000	George McLaughlin Adjustment in Canadian Agriculture	$5.00		
21000	Fred Gale and Stuart Pursey Microdynamics of Farm Size Growth and Decline: A Canada-United States Comparison	$5.00		
22000	Leonard Apedaile, Charles Barnard, Ray Bollman and Blaine Calkins The Structures of Agricultural Household Earnings in North America: Positioning for Trade Liberalization	$5.00		
23000	Glenn Zepp, Charles Plummer and Barbara McLaughlin Potatoes: A Comparison of Canada/USA Structure	$5.00		
24000	Victor J. Oliveira, Leslie A. Whitener and Ray Bollman Farm Structure Data: A U.S.-Canadian Comparative Review	$5.00		
25000	Karen Gray Grain Marketing Statistics Statistical Methods Working Paper Version 2	$10.00		
26000	W. Steven Danford Farm Business Performance: Estimates from the Whole Farm Database	$10.00		
27000	Brian Biggs An Attempt to Measure Rural Tourism Employment	$5.00		
			Subtotal	
			GST (7 %)	
			GRAND TOTAL	

Statistics Canada Statistique Canada

Canada