United States
Department of
Agriculture

National
Agricultural
Statistics
Service

Research and
Applications
Division

SRB Staff Report
Number SRB-89-01

February 1989

Improving Variance Estimates for Livestock Surveys

Philip S. Kott
Jerry Thorson

IMPROVING VARIANCE ESTIMATES FOR LIVESTOCK SURVEYS, by Phillip S. Kott, Research and Applications Division, and Jerry Thorson, Estimates Division, National Agricultural Statistics Service, U.S. Department of Agriculture, Washington, DC 20250, January 1989. Staff Report No. SRB-89-01.

ABSTRACT

This paper proposes three slight adjustments in the way variances are estimated for the "adjusted" livestock indications. The new variance formula is shown to be unbiased under the same assumptions that render the adjusted indication itself unbiased. Moreover, the formula becomes identical to the standard variance formula for a direct expansion when the adjusted estimator collapses into a direct expansion.

KEY WORDS

Product estimator, adjusted estimator, unbiased, random, livestock.

ACKNOWLEDGMENTS

The authors would like to thank Bill Iwig and Ned Jones for their assistance in bringing the issues underlying the adjusted livestock estimator to our attention. We would further like to thank Jack Nealon, Ron Fecso, Jim Mergerson, George Hanuschak, and Bob Tortora for their helpful comments.

February 1989

Washington, DC

TABLE OF CONTENTS

SUMMARY

The National Agricultural Statistics Service (NASS) currently calculates what it calls adjusted indications for livestock totals based on data from its multiple frame survey. These indications generally employ product estimators in place of direct expansions when estimating list stratum livestock totals. The rationale for this approach, which cleverly incorporates partial information on the presence of livestock on sampled farms, is given in Crank (1).

Crank's paper also proposed the estimation formula for the variance of an adjusted indication that is currently in use. There are three small errors in that variance formula making it biased even when the adjusted indication itself is unbiased. This paper corrects those three errors. Moreover, the formula introduced here collapses to the standard variance for a direct expansion when there is no partial information.

An empirical example suggests that adopting the new corrected variance formula may have only a minor effect on state level variance estimates for major livestock states. The effect in minor states may be more pronounced.

IMPROVING VARIANCE ESTIMATES FOR LIVESTOCK SURVEYS

By Phillip S. Kott and Jerry Thorson

INTRODUCTION

The National Agricultural Statistics Service (NASS) uses product estimators proposed by Crank (1) to derive its so-called adjusted multiple frame livestock indications. Within each list stratum, an estimate of the fraction of farms having the livestock type in question (say hogs and pigs) is multiplied by an estimate of the mean of interest (say hogs and pigs for breeding) among those farms having the livestock type. This product is then multiplied by the number of farms in the stratum to estimate the stratum total.

The product estimator described above allows NASS to use partial information about sampled farms. In particular, NASS often knows whether a sampled farmer possesses livestock of a certain type but not the actual quantities. Thus, the effective sample size for estimating (within a list stratum) the fraction of farms possessing the livestock type can be larger than the effective sample size for estimating a livestock total by direct expansion. As a result, the product estimate often allows NASS the use of additional information not available with direct expansion.

One of the nice properties of NASS's product estimator is that in a stratum where no sampled farm provides only partial information (a presence/absence indication without actual quantities), the product estimator is identical to the direct expansion. A good estimator for the variance of the product estimator should, in that situation, also collapse into the variance estimator for the direct expansion. Unfortunately, the variance estimator currently in use does not.

The variance estimator currently in use was also developed by Crank (1). It has three small mistakes in it which make it biased when the livestock indication itself is unbiased. All three will be discussed and corrected. The new variance estimator for a particular stratum does collapse into the variance estimator for the direct expansion when there are no sampled farms in the stratum providing partial livestock information.

An empirical investigation of the total September hogs and pigs indication for Nebraska shows that adopting the variance estimation formula proposed here reduces the estimated variance of the list frame total by only 4.1%

THE VARIANCE FORMULAE

It is impossible to address the subject of the NASS livestock indication and its variance without a lot of complicated notation. To simplify matters somewhat, we will restrict our attention to a particular list stratum, a particular livestock type (say hogs and pigs), and a particular item of interest within that type (say hogs and pigs for breeding).

Let

- N be the total number of farms (in the stratum),
- N^+ be the total number of farms with some positive values for the livestock type,
- n_F be the number of sampled farms fully responding to questions about the livestock type (all farms with zeros for the entire livestock type are considered fully responding),
- n_P be the number of sampled farms only partially responding, all of which are known positives; that is, are known to possess some of the livestock type in question,
- n^+ be the number of fully and partially responding sampled farms that are known positives,
- r be the number of fully responding sampled farms that are positives, and
- n be the total number of full and partial respondents; i.e., $n = n_F + n_P$ (note that this number may be less than the original sample size).

Suppose x_i is the value of interest for farm i. The population total (for the stratum) is $X^T = \Sigma^N x_i$. The population mean is $\bar{X} = X^T/N$, and the population mean among farms possessing some of the livestock type is $\bar{X}_{(+)} = X^T/N^+$. The population fraction of farms possessing some of the livestock type is $P = N^+/N$.

Observe that $X^T = NP\bar{X}_{(+)}$. The product estimator for X^T is $x^T = Np\bar{x}_{(+)}$, where $p = n^+/n$ estimates P, and $\bar{x}_{(+)} = \Sigma^r x_i/r$ estimates $\bar{X}_{(+)}$. If the n partial and full respondents can be treated as a simple random sample, and we will assume they can be, then p is an unbiased estimator of P when conditioned on a fixed n. If the r fully responding positives can be treated as a simple random sample from the N^+ population positives, and again we will assume they can be, then $\bar{x}_{(+)}$ is an unbiased estimate of $\bar{X}_{(+)}$ when conditioned on a fixed r.

These two assumptions, while somewhat strong (and in need of either empirical confirmation or rejection), are actually <u>weaker</u> than the assumption needed to establish the unbiasedness of the direct expansion, $x_{DE} = N\Sigma^r x_i/n_F$ when $n_F > 0$. For x_{DE} to be unbiased, the n_F full respondents must be treated as a simple random sample. That means that positives are assumed as likely to <u>fully</u> respond as "zeros." Crank (1) discusses the evidence that this is not always the case. The product estimator's comparable assumption, while still somewhat heroic, is nonetheless more reasonable: positives are assumed as likely to at least partially respond as zeros. (Note: for both x^T and x_{DE} to be unbiased the r positive full respondents must be assumed equivalent to a simple random sample of the N^+ population positives.)

Accepting the two assumptions stated above, $\bar{x}_{(+)}$ conditioned on a fixed r must be independent of the random variable n^+ (because the r fully responding positives are a subset of the n^+ positives). Now $p = n^+/n$, so $\bar{x}_{(+)}$ and p must be independent random variables. As a result, $E(x^T) = E(Np\bar{x}_{(+)}) = NE(p)E(\bar{x}_{(+)}) = NP\bar{X}_{(+)} = X^T$; that is, x^T is an unbiased estimator of X^T.

Crank (1) suggested the following estimator for the variance of x^T:

$$\text{var}_C = N^2\{(\bar{x}_{(+)})^2(1 - n/N)p(1-p)/n + p^2(1 - r/N)s_{x+}{}^2/r\},$$

where $s_{x+}{}^2 = (\Sigma^r x_i{}^2 - [\Sigma^r x_i]^2/r)/(r-1)$.

We demonstrate the unbiasedness of the following alternative formulation in the appendix:

$$\mathrm{var}_A = N^2\{[(\bar{x}_{(+)})^2 - s_{x+}{}^2/r](1 - n/N)p(1-p)/(n-1) + p^2(1 - r/[pN])s_{x+}{}^2/r\}.$$

There are three minor "corrections" incorporated into var_A. The value $(\bar{x}_{(+)})^2$ in var_C has been replaced by $(\bar{x}_{(+)})^2 - s_{x+}{}^2/r$, the value $p(1-p)/n$ by $p(1-p)/(n-1)$, and the value r/N by $r/[pN]$. The first and third correction can only lower the estimated variance, while the second correction may raise it. The interested reader can verify that when $n_P = 0$, so that $n_F = n$ and $r = n^+$, var_A collapses into the standard variance estimator for a direct expansion:

$$\mathrm{var}_{DE} = N(N - n)\{\Sigma^n x_i{}^2 - (\Sigma^n x_i)^2/n\}/[n(n-1)].$$

AN EMPIRICAL EXAMPLE

In order to assess the impact of replacing var_C by var_A, we looked at the adjusted indications for total hogs and pigs in Nebraska based on the 1988 September Agricultural Survey. The results by list stratum are given in Table 1. The total effect was to reduce the estimated variance (from the list frame) by 4.1%. This translates into only a 2.1% decrease in the estimated coefficient of variation.

The impact on particular non-hog strata is often more pronounced (note strata 61, 63, 74 and 77 in particular). This suggests that the change in the variance formula may have a greater effect on states having less of a particular livestock type than Nebraska has hogs (or more non-hog strata that in fact have farms with hogs).

Three of the strata, 96, 97, and 98, have estimated variances of zero because all of the farms in those strata are sampled, and all must "respond" (missing values are filled in by the state office in practice). In 14 of the remaining 16 strata, the variance estimate went down, and in one it stayed the same. Only one hog stratum (83) had an increase in its variance estimate. Thus, it seems that the

Table 1 -- The two variance estimates for total September 1988 hogs and pigs in Nebraska by list stratum

Stratum Number	Description	Crank Variance Estimator (in millions)	New Variance Estimator (in millions)
61	Cropland 1-199	18.2	16.4
63	Capacity 1-9,999	1,769.4	1,549.7
65	Cropland 200-649	2,218.5	2,110.9
66	Capacity 10K*-59,999	733.6	711.4
67	Hogs 1-99	1,653.1	1,614.6
69	Cropland 650-9,999	138.9	129.8
70	Hogs 100-199	1,137.6	1,124.1
71	Hogs 200-299	843.0	835.5
72	Hogs 300-599	962.6	954.4
73	Capacity 60K-399,999	812.4	765.0
74	Capacity 400K-999,999	13.1	10.2
77	Rye 50+	40.6	34.0
80	Hogs 600-1,249	890.4	883.3
82	Hogs 1,250-2,999	315.1	312.0
83	Hogs 3,000-4,999	211.7	215.5
84	Hogs 5,000-9,999	172.1	172.1
96	Cropland 10,000+	0.0	0.0
97	Capacity 1,000K+	0.0	0.0
98	Hogs 10,000+	0.0	0.0
	LIST TOTAL	11,930.4	11,439.0

* K denotes 000 (e.g., 10K = 10,000)

effect of our two downward corrections usually outweigh that of our one upward correction.

CONCLUSIONS

We have shown that making three small corrections to the current variance estimation formula for a list frame adjusted livestock indication will make it unbiased when the indication itself is.

An empirical analysis of hogs and pigs in Nebraska in September 1988 suggests that the impact of making the proposed changes in large livestock states will be downward and small.

Although the product estimator is unbiased under weaker assumptions than those necessary for the direct expansion estimator to be unbiased, research is still needed into the correctness of these assumptions and on the reasons for nonresponse in general.

RECOMMENDATION

The livestock variance estimating formula proposed here should be incorporated into the 1989 SAS Summary System.

APPENDIX

The variance of x^T is

$$\begin{aligned}\mathrm{Var}(x^T) &= E[(x^T - X^T)^2] \\ &= E[(Np\bar{x}_{(+)} - NP\bar{X}_{(+)})^2] \\ &= N^2E[((P + (p-P))\{\bar{X}_{(+)} + (\bar{x}_{(+)}-\bar{X}_{(+)})\} \\ &\qquad - P\bar{X}_{(+)})^2]\end{aligned}$$

(since $p = P + (p-P)$ and $\bar{x}_{(+)} = \bar{X}_{(+)} + (\bar{x}_{(+)}-\bar{X}_{(+)})$)

$$\begin{aligned}&= N^2E[(P\bar{X}_{(+)} + P\{\bar{x}_{(+)}-\bar{X}_{(+)}\} + \{p-P\}\bar{X}_{(+)} + \\ &\qquad \{p-P\}\{\bar{x}_{(+)}-\bar{X}_{(+)}\} - P\bar{X}_{(+)})^2]\end{aligned}$$

$$\begin{aligned}&= N^2E[(P\{\bar{x}_{(+)}-\bar{X}_{(+)}\} + \bar{X}_{(+)}\{p-P\} + \\ &\qquad \{p-P\}\{\bar{x}_{(+)}-\bar{X}_{(+)}\})^2].\end{aligned} \tag{A1}$$

Since p and $\bar{x}_{(+)}$ are assumed to be independent, the expectation of all the cross terms in the last line of (A1) are zero. In addition,

$$E[(\{p-P\}\{\bar{x}_{(+)}-\bar{X}_{(+)}\})^2] = \mathrm{Var}(p)\mathrm{Var}(\bar{x}_{(+)}).$$

As a result,

$$\begin{aligned}\mathrm{Var}(x^T) = N^2[P^2\mathrm{Var}(\bar{x}_{(+)}) + \{\bar{X}_{(+)}\}^2\mathrm{Var}(p) + \\ \mathrm{Var}(p)\mathrm{Var}(\bar{x}_{(+)})].\end{aligned} \tag{A2}$$

The variance of p is well known to be $[(N-n)/(N-1)]P(1-P)/n$. An unbiased estimator for this variance is $(1 - n/N)p(1-p)/(n-1)$ (Cochran, 2, p. 52).

The variance of $\bar{x}_{(+)}$ is $(1 - r/N^+)S_{X+}{}^2/r$, where $S_{X+}{}^2 = [\Sigma^N x_i{}^2 - (\Sigma^N x_i)^2/N^+]/(N^+-1)$ is the population variance of x_i among the N^+ farms with some positive values for the livestock type in question. Although $s_{X+}{}^2$ in an unbiased estimator of $S_{X+}{}^2$, there is no simple unbiased estimator for the variance of $\bar{x}_{(+)}$ because N^+ in the finite population correction term, $1 - r/N^+$, is not known. Since $N^+ = NP$, Np is an unbiased estimator for N^+. Unfortunately, $r/(Np)$ is <u>not</u> an unbiased estimator of r/N^+. (Crank (1) incorrectly had $1 - r/N$ as the finite population correction term for $Var(\bar{x}_{(+)})$.)

Due to this roadblock in trying to estimate $Var(x^T)$ piece by piece, we will go about the business backwards. We will show that the final result, var_A, is an unbiased estimator for $Var(x^T)$. First, however, note the following two equalities (not included in Crank's (1) derivation):

$$
\begin{aligned}
E(p^2) &= P^2 + Var(p), \text{ and} \\
E((\bar{x}_{(+)})^2) &= (\bar{X}_{(+)})^2 + Var(\bar{x}_{(+)}).
\end{aligned}
\tag{A3}
$$

The expected value of our proposed variance estimator is

$$
\begin{aligned}
E(var_A) &= N^2E\{[(\bar{x}_{(+)})^2 - s_{x+}{}^2/r](1 - n/N)p(1-p)/(n-1) \\
&\qquad + p^2(1 - r/[pN])s_{x+}{}^2/r\}. \\
&= N^2([E\{(\bar{x}_{(+)})^2\} - E\{s_{x+}{}^2/r\}] \cdot \\
&\qquad E\{(1 - n/N)p(1-p)/(n-1)\} \\
&\qquad + [E(p^2) - E(p)r/N]E\{s_{x+}{}^2\}/r).
\end{aligned}
$$

(since the random variable $s_{x+}{}^2$ is conditioned on a fixed r, it is independent of $p = n^+/n$.)

$$
\begin{aligned}
&= N^2([(\bar{X}_{(+)})^2 + Var(\bar{x}_{(+)}) - S_{x+}{}^2/r]Var(p) \\
&\qquad + [P^2 + Var(p) - Pr/N]S_{x+}{}^2/r). \\
&= N^2([(\bar{X}_{(+)})^2 + Var(\bar{x}_{(+)})]Var(p) + \\
&\qquad [P^2\{1 - r/(NP)\}S_{x+}{}^2/r)
\end{aligned}
$$

$$= N^2[\{\overline{X}_{(+)}\}^2 Var(p) + P^2 Var(\overline{x}_{(+)}) + Var(p) Var(\overline{x}_{(+)})].$$

$$= Var(x^T). \qquad QED.$$

REFERENCES

(1) Crank, Keith, N. The Use of Current Partial Information to Adjust for Nonrespondents. U.S. Department of Agriculture, Economics, Statistics, and Cooperatives Service, 1979.

(2) Cochran, William G. Sampling Techniques (3rd edition). New York: Wiley, 1977.