ESTIMATION OF POPULATION TOTALS FOR HIGHLY SKEWED POPULATIONS IN REPEATED SURVEYS

Harold F. Huddleston
Statistical Reporting Service
U.S. Department of Agriculture

March 1965

ESTIMATION OF POPULATION TOTALS FOR HIGHLY SKEWED POPULATIONS IN REPEATED SURVEYS 1/

Introduction

The purpose of this report is to describe the investigation of alternative estimators for extremely skewed distributions based on prior knowledge of the distributions. The data considered in detail is the basic population of area segments within a State. However, the methods may be applied to any type of sampling unit. It is believed that the distributions associated with the segments for a particular State or group of States are similar for the same characteristic since standard rules of segment construction were used. Procedures for combining the information for the samples over years in order to improve results are developed.

The alternative estimators require considerably more knowledge of the distributions than the mean and variance which are sufficient information for determining sample size. In fact, the basic objective is to make use of the additional information which is available on the distribution from the same or similar populations. While certain additional information is required about the nature of the distribution, apriori knowledge about individual sampling units is not required. In the past, these estimators have been used primarily to develop improved State or small area estimates.

These estimators may have small biases for individual States but the biases can be made to sum to zero over a group of States. A similar approach for making estimates by districts or subdivisions within States has been attempted on a limited basis.

One of the primary sources of variation in the estimators is the contribution to the variance resulting from observations in the right-hand tail of positively skewed populations. All the estimators studied are similar in

---

1/ W. E. Kibler and Charles E. Caudill of the Research Branch assisted in compiling the results reported.

that they involve special procedures for estimating the contribution of the extreme observations to the population total for the characteristic. The three principal advantages to these estimators are: (1) In general they have smaller variances than the unbiased estimator based on the reciprocal of the probability of selection, (2) the estimated population total and variance are not subject to large changes between successive surveys (or years) due to a few extreme observations, and (3) the distribution information for the individual characteristic can be used in the estimators even though it may not be practical to design a multivariate survey which will have minimum variance for each characteristic.

Review of Literature

A great deal of statistical research and application has dealt with the problem of estimation when sampling from a "contaminated" population or when estimating from incomplete sample data; however, much of this research is concerned with special problems or is based on the assumption that the population under consideration is normally distributed. Also, many of the estimators which have been developed are highly biased, although the mean square error of the estimators are generally less than the variance of other unbiased estimators such as the mean per unit.

A great volume of statistical literature is concerned with problems of "extreme values or outliers," which are somewhat related to the problem at hand. However, no attempt has been made here to present a comprehensive review of these topics for the purpose of this paper. Instead, only a very few contributions which are closely related to the problem under consideration will be mentioned.

Hald (1949) in a paper dealing with "Incomplete Sample Data" introduces the terms "censored" and "truncated" to distinguish between the cases when the number of "missing observations" are known and unknown, respectively. The term "censored" as used in this paper will be somewhat different in that the sample observations which are censored are all known and all have very large values, i.e., they all are found in the right-hand tail of positively skewed distributions.

Krane (1957) develops a procedure for estimation from incomplete sample data when sampling from continuous distributions. The approach developed in his paper is to use maximum likelihood estimators along with an iterative process where representative weights and locations are used to transform the original set of maximum likelihood equations to an equivalent set corresponding to the likelihood equations appropriate to some complete sample from the same distribution. Iteration is necessary since the location and weight coefficients are in general functions of the parameter vector. Krane points out "that the user of the method of representative weights and locations is free to utilize any means he may desire in order to "force convergence rapidly."

In an unpublished paper, Hendricks and Huddleston (1960) have examined a procedure for discarding a small percentage of the values corresponding to the largest sample observations and replacing them by their expected values from a Pearson Type III Distribution. The earlier paper is directly related to the present paper in that most of the ideas and theory have been incorporated into the current approach. In addition, Searles and Cavallini have developed alternative approaches as a result of the earlier work.

Searles (1963, 1964) developed a series of seven alternative estimators for handling such observations all of which are biased but have smaller mean square errors than the variance of the sample mean when certain specified conditions are met. He defines "extreme observations" as those having values greater than some predesignated cutoff point, $\underline{t}$. Results were developed primarily for a particular class of continuous distributions. Distributions belonging to this class have finite first and second moments, are unimodal, and are not negatively skewed. Searles used the exponential distribution --

$$f(x) = \frac{1}{\theta} \quad e^{-x/\theta} \qquad 0 < x < +\infty$$

as an example of the class of distributions under consideration.

Searles shows that gains in efficiency can be quite marked for small sample size and that the stability of estimates are improved when the extreme observations are weighted by some factor different from 1/n in the case of equal probability of selection.

Cavallini (1963) investigates alternative estimators for extremely skewed populations for samples from the same or similar populations. Cavallini's work is also directly related to this report in that he specifically considers in detail the same basic population of area segments within a State. He develops three alternative estimators for improving the estimates of one- and two-way classifications. These estimators are of the adjustment type, i.e., certain sample marginal totals are accepted and these totals are distributed to individual cells. Marginal totals are obtained by summing State totals over years and/or substrata within States. The State totals are obtained by expanding the sample data by the reciprocal of the sampling rate. Three possible estimators of the individual cell totals (States) were considered. These were as follows:

(1) A ratio estimator defined as the ratio of the estimated total over years to the estimated total for observations less than or equal to the cutoff for the cell (State by year) under consideration.

(2) An estimator based on an additive model, defined as the sum of the row and column marginal means of the observations greater than the cutoff for the corresponding cell minus the grand mean of the observations greater than the cutoff plus the estimated total for observations less than or equal to the cutoff for that particular cell (State by year), and

(3) An estimator based on a multiplicative model defined as the product of the row and column marginal totals of the observations greater than the cutoff for the corresponding cell divided by the total sum of the expanded observations greater than the cutoff plus the estimated total based on expanded observations less than or equal to the cutoff.

These estimators may be biased for any particular cell (State-year) but the biases sum to zero over States and/or years. When these estimators were applied using June Enumerative Survey data for 1958-60, the mean square error of the estimators was estimated to be less than the variance of the unbiased estimator (Direct Expansion) in about four-fifths of the cases.

Alternative Methods of Estimating Population Totals

The various models discussed are based on dividing the characteristics for the survey sample units into two groups based on some predetermined value (or criterion). One group contains all the sampling units for which the value of the characteristic is less than or equal to some predetermined constant (X_o) referred to as the cutoff or censoring point. The other group is composed of extreme observations, that is, it contains all the values greater than the constant (X_o).

The various methods of estimating differ primarily in these respects: (1) the kind of prior information available, (2) the type of distribution, and (3) the type of estimator which is used for the total of the extreme observations.

The estimators will be discussed based on the kind of prior knowledge used in estimating the portion above the cutoff for the characteristic of interest; that is: (1) Sufficient previous results are available above and below the cutoff point for the characteristic being estimated that the contributions to the population total for the portion above the cutoff can be determined empirically, (2) previous information is available so a theoretical distribution is known which gives a good fit for the characteristic of interest, and (3) results are available for the characteristic (either from a current or a previous sample) to indicate the distribution at some higher level than the cell or substrata being estimated (i.e., substrata within States or States within a region.)

Estimators for Population Totals

When sufficient previous sample or census data are available the contribution to the population total for the portion above the cutoff can be determined empirically. A censored estimate of the total for a State for the current survey can be made by using an estimator of the following form:

$$\hat{X} = \sum_{\text{Dist.}}^{k} \left[\sum_{i=1}^{n_1} (1/P_i) X_{bi} + \hat{K} \sum_{i=1}^{n_2} (1/P_i) \right] \qquad (1)$$

where X_o = the preselected cutoff value for the characteristic,

$X_i \leq X_o$ is defined as X_{bi} and P_i is the probability of selection associated with the i sampling unit,

$\sum_{i=1}^{n_2} \frac{1}{P_i}$ = estimated population number of sampling units ($\hat{N}_a$) with values of X_i greater than X_o,

$\hat{K}$ = expected value or mean of the X_i for all $X_i > X_o$ based on prior data,

$\hat{X}$ = estimated total for the characteristic being estimated, and

$n = n_1 + n_2$.

An estimator of this form with K determined empirically assumes that there is a stable population with a finite range which can be accurately characterized by a distribution. From this population a probability sample is drawn in which the probability of $X_i > X_o$ is small (generally less than .01). The expectation is that a few of these extreme units will fall into the sample in repeated surveys, but it is not known in advance for which units X_i will be greater than X_o. The name censored estimate is used to describe this estimator since the number of units greater than X_o in a particular sample is known but the observed value of the characteristic is replaced by the expected value for units greater than X_o.

A modification of the estimator $\hat{X}$ can be made where a stable subpopulation exists and a theoretical distribution is fitted to only that portion for which $X_i > X_o$. For certain characteristics which one may wish to estimate, the population being sampled can be divided into two distinct distributions. The sample elements corresponding to the one distribution are used as observed while the expected or assumed distribution is used for the range of values where $X_o < X_i < \infty$. In such cases as this, a censored

estimator such as (1) may be used by determining the constant K from the theoretical distribution which exists for the range of values greater than the cutoff value. However, the choice of X_o is dictated largely by the point where the two populations separate and the probability is small that they overlap. As an example, PARETO'S DISTRIBUTION may be appropriate for the upper tail of the distribution and has the following form:

$$f(x) = \begin{cases} 0 & , \text{ for } x \leq X_o \\ \frac{a}{X_o} \left(\frac{X_o}{x}\right)^{a+1} & , \text{ for } x > X_o \end{cases}$$

where a is some predetermined constant > o, and $X_o < x_i \leq \infty$.

Hence, Pareto's Distribution may fit the portion of the original population above X_o, and the constant K can be determined by obtaining the expected value for this theoretical distribution. This approach has been used on the assumption that two distinct distributions are being sampled for different ranges of the characteristic which one desires to estimate.

When a series of t samples or repeated surveys are made from the same frame, the knowledge of the units for which $x_i > X_o$ can be used in the estimator by retaining these units in the current sample for the portion $x_i > X_o$. The expression

$$\frac{1}{t}\left[\sum_{j=1}^{t}\sum_{i=1}^{n_{2j}} 1/P_{ij}\right]$$

gives an average number of units above the cutoff and the total quantity above the cutoff is

$$\frac{K}{t}\left[\sum_{j=1}^{t}\sum_{i=1}^{n_{2j}} 1/P_{ij}\right] \quad .$$

Using this modification (1) can be rewritten as follows:

$$\hat{X} = \sum_{\text{Dist.}}^{k} \left[\sum_{i=1}^{n_1} \frac{1}{P_i} \cdot X_{bi} + \frac{\hat{K}}{t} \left(\sum_{j=1}^{t} \sum_{i=1}^{n_{2j}} \frac{1}{P_{ij}} \right) \right] \qquad (1a)$$

(1a) generally leads to improved estimates for the population by limiting the variability in the number of sampling units which have characteristic values greater than X_o.

The mechanics of computing a censored estimate based on either (1) or (1a) are fairly simple. Other than the sample data, only two other items, X_o and K, are necessary for computing the censored estimate. Each sample observation is compared with X_o, and if the observed value is greater than X_o, the observed value for the characteristic is discarded but the number of such elements is counted and accumulated to the population level. The constant $\hat{K}$ is multiplied by the estimated population number, $\hat{N}_a$, to obtain the estimated total for the particular item above the cutoff value X_o. This in turn is added to the estimated total for the item based on the expanded sample values below or equal to X_o.

While the estimate can be computed very easily, calculations involved in computing the variance of the estimate are somewhat more difficult. The variance of the censored estimate (1a) is computed by the following formula:

$$V(\hat{X}) = \frac{N^2 p}{n} \left[\frac{q(\bar{X}_b)^2}{t} + s_b^2 + \frac{qK^2}{t} - \frac{2q\,K\,\bar{X}_b}{t} \right] \qquad (1b)$$

N = total number of sampling units (segments) in population

n = sample size

p = proportion of sampling units with values for the item less than or equal to the cutoff (X_o)

q = 1 - p

$\bar{X}_b$ = the mean for all sampling units from the sample with values less than or equal to X_o

s_b^2 = the estimated variance of the individual observations below or equal to the cutoff for the current survey

t = number of surveys used to determine K and p

$$\frac{\bar{t}}{\bar{t}_o} = \frac{\frac{1}{V^2}}{\frac{1}{V^2} - \frac{e^{-t_o}(t_o)^{1/V^2}}{F(t_o)\ \Gamma(\frac{1}{V})}} = \frac{F(t_o)}{F(t_o) - \frac{e^{-t_o}(t_o)^{1/V^2}}{\Gamma(\frac{1}{V^2}+1)}}$$

where $V = \frac{\sigma}{m}$ = coefficient of variation

t_o = Censoring point for the variable t

$F(t_o)$ = Proportion of the population less than t_o

$\Gamma(\frac{1}{V^2}+1)$ = Value of the incomplete gamma function

$\bar{t}$ = Mean of population

$\bar{t}_o$ = Mean of population less than t_o

The ratio "H" in formula (2) can be derived from the above results as follows:

$$H = \frac{\bar{t}}{\bar{t}_o} - 1$$

The estimated variance of estimator (2) can be derived by subdividing the sums of squares as follows:

$$\sum_{i=1}^{n} (X_i - m)^2 = \sum_{i=1}^{n_a} (X_{bi} - m)^2 + \sum_{i=1}^{n_b} (X_{ai} - m)^2$$

and the variance of the total can be given in the following form:

$$\hat{V(X)} = \frac{N^2 S^2}{n} = \frac{N^2}{n}\left[\frac{n_b[S_b^2 + (m_b - m)^2]}{n-1} + \frac{n_a[S_a^2 + (m_a - m)^2]}{n-1}\right] \qquad (2a)$$

Since only a small number of sampling units will have characteristic values which are greater than the cutoff, in practice, m_a and S_a^2 are determined from the theoretical distribution rather than estimated from the sample data. However, the variation due to n_a (or $\frac{n_a}{n}$) would be based on the current survey. When using estimator (2) the proportion to be censored should be specified in advance and the cutoff determined from the theoretical distribution. If the percent cutoff is fixed in advance, the sample estimate of variance can be computed from formula applicable to post-stratification estimators. When the cutoff value X_o is fixed in advance, the sample data would be divided into two domains by the fixed cutoff. Then the variance of the domain total in stratified sampling would be:

$$V(\hat{{}_jX}) = \sum_h^L N_h^2 \left(\frac{N_h - n_h}{N_h} \right) \left[\frac{{}_jn_h - 1)\,{}_jS_h^2}{(n_h - 1)n_h} + \frac{{}_jP_h\,{}_jq_h}{(n_h - 1)}\,{}_j\bar{X}_h^2 \right] \qquad (2b)$$

and the variance of the population total would be the sum of the two domain variances, or

$$V(\hat{X}) = \sum_{j=1}^{2} V(\hat{{}_jX}) \quad .$$

Where the cutoff X_o is fixed in advance and the constant K is used for the mean of the domain above X_o, only the second term in the formula for domain 2 will be present. This is the variation due to $\frac{n_a}{n}$ or the binomial variation due of ${}_jP$.

Censored estimate III in Table 2 is an example of this approach using estimator (2) and the above variance estimator for all hogs and pigs in the June 1963 Enumerative Survey. This estimator is essentially the same as estimator (1a), if K and the expression for the number of sampling units in the population are estimated from the theoretical distribution.

For relatively small samples, both estimators (1) and (2) have been used over years based on 98-99 percent of the current sample observations and are not subject to random fluctuations due to the presence or absence of

a few large values. Where the two techniques described have seemed appropriate at a State or regional level, similar estimators are being investigated for substrata, usually consisting of geographic areas with sample of sizes 25-50. The expected gains in efficiency for samples of this size are even greater based on Searles' work, though the bias for substrata estimators may be somewhat larger.

If the State population is visualized as divided into subpopulations and each of these censored independently, estimator (1) may be used to estimate totals for these subpopulations or strata. Since the strata used are generally the Crop Reporting Districts of SRS, large samples from previous surveys or censuses may not be available to approximate the distribution which is to be censored. Hence, the required information to establish the population parameters for the censored estimator may not be known very accurately; however, these parameters can be made to coincide with State level parameters by a technique similar to those outlined in Cavallini's paper.

Examples of Estimated Totals and Variances for Selected Characteristics

In multipurpose and multivariate surveys for crop and livestock characteristics, Pearson Type III distributions with coefficients of variation of 100 to 300 percent are common. Broiler, livestock feeding, or holding operations or other highly specialized units may occur which contaminate the distribution, resulting in a distinct subset which may constitute a second distribution embedded within the population. For a group of States (region) as a whole, a sample of 2,000 units will generally be representative of the distribution of all characteristics with values from the upper tail occurring in about the expected frequency. For individual States, the means and variances for certain characteristics are frequently quite sensitive to extremely large values of the characteristics associated with only a few sampling units. These sampling units may influence the mean as much as 30 percent and the sampling error as much as 100 percent for some items though for most characteristics and States the

influence of these units may not be important for the particular sample selected. However, in repeated sampling both the means and variances will be highly sensitive to the presence or absence of these extreme values in the selected sample.

"Number of Farms" will illustrate a situation for a characteristic with a low coefficient of variation and where a few area segments (sampling units) with extremely large values for this characteristic may cause large changes in the level of the estimated mean (or total) from successive surveys. Since the number of farms per segment was controlled in the frame construction a single distribution should be indicated. However, due to changes in farm composition and residential developments, a second distribution or subset of segments may have developed for which the controlled information is now in gross error. For such a situation, Pareto's Distribution may fit the distribution of number of farms per segment.

Values for K were derived from Pareto's Distribution and from those segments with number of farms greater than eight over States and years. Results using these two techniques for determining the parameter K in estimator (1) are shown in Table 1 along with the estimate based on the reciprocal of the probability of selection.

Censored Estimate I (with K determined empirically) appears to be the most efficient estimator in terms of variance; however, when mean square errors are compared the relative efficiency of Censored Estimate II (K derived theoretically) is 110 percent while the relative efficiency of Censored Estimate I is 80 percent when compared with the Direct Expansion Estimate. The gains from the use of the Censored Estimate for "Number of Farms" as compared with a stratified sample are quite small except in Oklahoma and Texas which are States with relatively high variability.

Table 1. Estimated Number of Farms - June 1963 Enumerative Survey

State	Direct Expansion 1/		Censored Estimated 2/			
			I		II	
	Estimate	C.V.	Estimate	C.V.	Estimate	C.V.
	(000)	(%)	(000)	(%)	(000)	(%)
OHIO	155	4.9	152	4.3	155	4.6
IND	120	4.8	123	4.8	120	4.8
ILL	148	6.7	143	4.3	146	5.1
MICH	114	4.9	114	4.7	114	4.9
WIS	135	4.8	135	4.5	136	4.7
MINN	131	3.7	131	3.8	131	3.7
IOWA	166	3.8	166	3.7	166	3.8
MO	159	4.8	160	4.5	160	4.7
N DAK	60	5.0	59	4.4	60	4.5
S DAK	53	5.2	52	4.7	53	5.1
NEBR	94	6.3	88	4.4	94	6.2
KANS	96	5.6	96	4.4	98	5.9
REGION	1,432	1.5	1,420	1.3	1,433	1.4
VA	94	5.5	94	5.3	94	5.4
N C	195	4.2	198	4.0	199	3.9
S C	73	6.5	71	5.4	75	6.3
GA	89	5.9	88	5.1	89	6.2
KY	146	4.8	149	4.3	146	4.8
TENN	171	4.4	169	3.9	172	4.3
ALA	105	5.4	104	5.1	107	5.4
MISS	120	4.8	124	4.7	122	4.9
ARK	91	5.7	92	5.3	93	5.9
LA	64	5.6	63	5.3	65	5.7
OKLA	107	8.9	102	5.5	104	6.7
TEX	277	8.3	247	3.7	264	6.3
REGION	1,536	2.0	1,504	1.4	1,530	1.7
24 STATES	2,968	1.3	2,924	0.9	2,963	1.1

1/ Based on the reciprocal of the probability of selection.

2/ Censored Estimate I: Based on observed values larger than the cutoff value (X_o) (X_o determined independently for each State) averaged over years to determine K for each State.

Censored Estimate II: Pareto's Distribution with $a = 2.06$ and $X_o = 8$, used to determine K.

A similar approach is shown for examples of characteristics which in general have somewhat higher variability. Table 2 gives some results for the same estimators used in Table 1, and an additional estimate using estimator (2) under the assumption that the observations "fit" a Type III distribution and is called estimator III. In general, estimator I has smaller sampling errors than estimators II and III but somewhat larger biases are present. In terms of relative efficiency censored estimator II for the region is the most efficient of the three estimators. Pareto's Distribution for $a = 5$ and $X_o = 540$ was used to determine the parameter K in formula (1), which was used to compute censored estimator II for all hogs and pigs.

Censored Estimate III was computed by using formula (2) for the total and (2b) for the variance. A fixed cutoff ($X_o = 540$) was assumed for each State and the proportion censored was allowed to vary. Estimator (2) is more efficient if the proportion censored is fixed and the cutoff is computed or is allowed to vary. Also, because of available tables the population sampled was assumed to be a Type III distribution with a coefficient of variation of 200 percent for Ohio, Indiana, Michigan, Wisconsin, Minnesota, North Dakota, South Dakota, and Kansas. For Illinois, Iowa, Missouri, and Nebraska the coefficient of variation was assumed to be 142 percent. These assumptions are generally supported by the sample data with some minor exceptions. For example, the Iowa data gives a "better fit" if the assumed distribution has a coefficient of variation of 110 percent for estimator III as shown below.

Example:

All hogs and pigs: Assume a coefficient of variation of 1.1 and a selected cutoff point corresponding to $P = .984$.

1. Estimated total number of hogs and pigs based on complete sample = 14,805,362
2. Total number of sampling units in State = 111,410
3. Estimated mean number of hogs and pigs per segment based on complete sample = 14,805,362 ÷ 111,410 = 132.9

4. X_o = (132.9) x (4.14*) = 550.2 (Use 550)

5. Segments with total hogs and pigs greater than 550

Strata Identification	Number of hogs and pigs reported	1/P	Expanded number of hogs and pigs	No. of segments in universe greater than 550
10	649	274.044	177,855	274.0
10	630	274.044	172,648	274.0
50	770	343.909	264,810	343.9
90	643	352.970	226,960	353.0
State totals	XX	XX	842,273	1,244.9

6. Mean number of hogs and pigs per segment below the cutoff

$\bar{X}_b$ = (14,805,362 - 842,273) ÷ (111,410 - 1,245)

$\bar{X}_b$ = (13,963,089) ÷ (110,165) = 126.7

7. Fraction of distribution censored = 1245/111,410

= .011

8. Mean number of hogs and pigs per segment for complete distribution

$\bar{t}_o$ = (126.7) x (1.053*) = 133.4

(* Ratio of mean below the cutoff to the mean for the entire distribution)

9. Estimated total hogs and pigs for complete distribution

= 133.4 x 111,410 = 14,862,094

The estimated total from the preceding example shows that censored estimate III to be relatively unbiased when the sample data "fits" the assumed distribution. The estimated coefficient of variation for the complete sample was 112 percent. When a coefficient of variation of 142 percent was assumed, the censored estimate for Iowa (III in Table 2) was 105.4 percent of the direct expansion estimate, compared with 100.4 percent for the estimate computed in the example above where a coefficient of variation 110 percent was assumed.

Table 2. All Hogs and Pigs - June 1963 Enumerative Survey

State	Direct Expansion 2/		Censored Estimate 2/					
			I		II		III	
	Estimate	C.V.	Estimate	C.V.	Estimate	C.V.	Estimate	C.V.
	(000)	(%)	(000)	(%)	(000)	(%)	(000)	(%)
OHIO	3,550	10.6	3,316	8.7	3,595	10.9	3,534	10.6
IND	5,459	10.4	5,168	7.9	5,390	9.9	5,510	9.8
ILL	7,895	8.7	7,909	7.8	7,984	8.6	7,850	8.7
MICH	569	17.6	558	19.2	569	17.6	569	18.3
WIS	2,009	13.0	2,076	12.4	2,021	13.1	1,890	12.6
MINN	4,474	9.3	4,399	8.6	4,451	9.0	4,497	11.4
IOWA	14,805	6.0	15,124	5.7	14,849	6.1	15,597	6.9
MO	4,506	7.8	4,443	7.8	4,506	7.8	4,506	9.1
N DAK	535	14.2	536	16.8	535	14.2	535	12.3
S DAK	2,529	13.7	2,283	8.4	2,328	9.1	2,368	9.2
NEBR	3,687	7.2	3,739	7.0	3,687	7.2	3,687	9.5
KANS	1,604	13.1	1,523	9.5	1,610	13.2	1,614	13.4
REGION	51,622	3.0	51,074	2.7	51,525	2.9	52,157	3.2

1/ Based on the reciprocal of the probability of selection.

2/ Censored Estimate I : Based on observed values larger than the cutoff value averaged over years to determine K for each State.

Censored Estimate II : Pareto's Distribution with a = 5 and X_o = 540 used to determine K.

Censored Estimate III : Censored Estimate, assuming (1) sample observations come from a Pearson Type III Distribution, (2) a fixed cutoff value (X_o = 540), (3) a population C.V. of 1.42 for ILL, IOWA, MO, and NEBR. For other States, a C.V. of 2.00 was assumed.

In practice, censored estimates and variances generally have been completed using a single cutoff value and mean above this value for an entire State or group of States, without taking advantage of the geographic stratification employed in the sample design, (e.g. censored estimates in Tables 1 and 2). The computed sampling error for the censored estimates using this approach corresponds to a simple random sample within States. This procedure has been used under the assumption that there is one basic distribution within each State from which the sample has been selected, and because parameters can be determined more precisely based on the larger sample at the State level. There is evidence that further gains in efficiency may be realized by using different cutoff values and means for substrata within States. Some examples using this approach are shown in Table 3. The censored estimates shown in this table were all computed by using estimator (1) with cutoff values and parameters determined empirically. Estimate I and its estimate of variance were computed by determining from previous survey data a single cutoff value for each item and a mean above this value (K) and using these values in estimator (1) without considering geographic substrata in the computation of the estimate and its variance. In the variance computations (estimator (1b)), the proportion (P) below the cutoff was determined by averaging over several surveys.

Estimate IA was computed by using a common cutoff value and mean above for all substrata. However, estimates of totals and variances were computed for each substrata (Crop Reporting Districts) and were added to obtain State estimates. Actual observed values for "P" and "N_a" were used for the computations within each substrata.

Censored Estimate IV was derived by first determining independent cutoff values and other parameters for each substrata. Estimates of totals and variances were computed independently for each substrata and added to obtain State estimates. The parameters "P" and "N_a" for each substrata were updated by averaging observed values for the current survey with previous surveys. This procedure is particularly effective for acreages when the size of sampling units (segments) vary between strata. The main disadvantage is the amount of effort required to establish cutoff values and parameters for each strata. In terms of the mean square errors censored, estimate IA appears to be the most efficient for the items shown in Table 3.

Table 3. Estimated Acreages for Selected States and Crops - June 1963

State	Characteristic	Direct Expansion		Censored Estimate 1/					
				I		IA		IV	
		Estimate	C.V.	Estimate	C.V.	Estimate	C.V.	Estimate	C.V.
		(000)	(%)	(000)	(%)	(000)	(%)	(000)	(%)
ILL	Soybean-Acres planted for beans	5,549	5.4	5,591	5.2	5,503	5.0	5,555	4.1
KANS	Wheat-Acres planted	9,655	4.1	9,565	5.9	9,608	4.2	9,403	3.9
IOWA	Corn-Acres planted for grain	11,204	3.3	11,200	3.5	11,168	3.2	10,877	3.0
N C	Peanuts-Acres planted	210	13.2	179	19.0	204	7.3	2/	
TEX	Cotton-Acres planted	6,413	7.2	6,127	6.4	6,141	4.2	6,346	4.6

1/ Censored Estimate I : Computed by using a single cutoff value and mean above for the entire State disregarding geographic substrata.

Censored Estimate IA : Same as Estimate I except the estimate and variance was computed for each substrata and added to the State level.

Censored Estimate IV : A different cutoff and mean above was used in each substrata, with the estimate and variance being added to the State level.

2/ This crop is confined to only a few substrata in the State, hence separate cutoff for each stratum was not attempted.

SUMMARY

This paper has been concerned with the problem of "extreme observations" which occur when sampling from the population of area segments within a State. Procedures are discussed for handling the "extreme values" in the estimation process.

Two estimators are given which have been used in practice, both of which are generally more efficient than the Direct Expansion estimator if the assumptions under which they are used are valid and the geographic stratification in the sample design is considered. Several different approaches are presented for using apriori knowledge to determine the parameters or constants necessary for using the two censored estimators. These approaches were: (1) empirical, (2) using a theoretical distribution for the sample observations greater than the cutoff value, and (3) fitting a theoretical distribution for all sample observations and using the ratio of the theoretical mean (or total) below the cutoff to the mean of the whole distribution.

The empirical approach has been used most in practice because of the effort involved in finding and "fitting" theoretical distributions to the sample data.

Some results are shown for different characteristics in Tables 1, 2, and 3. Although sampling errors for the censored estimates are generally smaller than for the direct expansion estimator, these estimators may be biased. The degree of bias will depend upon the precision with which the constants used in the estimating formulas can be determined. Despite the fact that the estimators may be biased, censored estimates are appropriate for controlling level of estimates, particularly when sampling from the same population over time.

REFERENCES

Krane, Scott A., "Maximum Likelihood Estimation From Incomplete Data From Continuous Distributions," Masters Thesis, Iowa State University, 1957

Hendricks, W. A. and Huddleston, H. F., "Estimation of Characteristics from Restricted or "Censored" Sample Data," Unpublished, 1960

Searles, Donald T., "On the Large Observation Problem," Ph.D. Thesis, North Carolina State College, 1963

Cavalline, Carlos Manuel, "Estimation of Population Totals Given Samples From Similar Populations," Masters Thesis, Iowa State University, 1963

Searles, Donald T., "The Utilization of a Known Coefficient of Variation in the Estimation Procedure," JASA, Vol. 59, Dec. 1964