ESTIMATION OF CHARACTERISTICS FROM RESTRICTED
OR "CENSORED" SAMPLE DATA

W. A. Hendricks
H. F. Huddleston

Research and Development Staff
Agricultural Estimates Division
Agricultural Marketing Service
U.S. Department of Agriculture

1960
1. Introduction

A restricted estimator from a random sample may be viewed as a special type of partial coverage, in which all cases above a specified size criterion are omitted. The technique is essentially one of estimation from incomplete data using either a "truncated" or "censored" sample. However, the technique proposed is an alternative to using the complete data for skewed distributions. While the estimated sample characteristics for the complete data is properly assessed by the sampling errors, the apriori knowledge of the distribution other than the coefficient of variation (in setting sample size) is not used. In some cases this is due to the fact that individual units cannot be placed in size strata, or due to the lack of control which can be exercised over individual characteristics in a multi-purpose survey. The extreme upper tail of the sample distribution is censored in favor of the expected value (or a maximum likelihood value) where the inclusion of the extreme values are expected to make the variance unduly large and in all likelihood increase the mean.

The establishment of censored points at specified levels based on a negligible portion of the upper tail ($\frac{m}{n} \leq 0.02$) has proven useful in surveys. In those situations where all the sampling units with extremely large values of the characteristic cannot be identified in advance and sampled with certainty, a relatively few sampling units may exert considerable influence on the mean and the variance. The device described appears promising where the type of distribution is known in advance. The procedure has been applied to Type III distributions with coefficient of variations of 100 to 200 percent. It is also contemplated to study this procedure for distributions similar to $f(x) = ke^{-x}$.

2. Estimation of Population Total for "Censored" Sample

In the typical censored case, a number of units ($m$) in a sample of size $n$ are not measured or for some reason the value of each of the $m$ units is not available. The technique employed proceeds as if the characteristic values for $m$ sampling units in the cumulative distribution function for $F(x) \geq 0.99$ are not known. It is assumed that the sample extreme values beyond $F(x) = 0.99$ occur with the expected frequency but the observed values are seldom representative of the entire upper tail of the distribution. In addition, it is assumed that the characteristic has a Type III distribution and the coefficient of variation is known from previous information. In effect, only the absolute magnitude of the mean of the distribution is unknown.

The procedure used is to relate the mean of the total distribution to the mean of the distribution below the cutoff or censoring point. The cutoff point is determined (approximately) for the sample data by multiplying the mean of the complete sample by the ratio of the value of the cutoff point to the mean of a Type III distribution for $\bar{x} = 1.0$ for a given coefficient of variation.
a. Relationship of Mean of Population to Mean of Censored Distribution

The distribution function considered is:

\[
\text{d}F(x) = ke^{\frac{x}{\bar{x}} - \frac{1}{2\bar{x}^2}} \frac{1}{(\frac{x}{\bar{x}})^{\frac{1}{2}}} \text{d}x
\]

where each x is expressed as a fraction of the mean, k is a constant, and v is the coefficient of variation. If the ratio \(\frac{x}{\bar{x}}\) is denoted by R we can write (1) as follows:

\[
\text{d}F(R) = ke^{-\frac{R}{v^2}} \frac{1}{2} R^{-\frac{1}{2}} \text{d}R
\]

and the distribution is determined by v since

\[
E(R) = E(\frac{x}{\bar{x}}) = \frac{\bar{x}}{\bar{x}} = 1.
\]

If we make the substitution \(t = \frac{R}{v^2}\), we obtain (2) in a form easier to evaluate.

\[
\text{d}F(t) = k e^{-\frac{t}{v^2}} \frac{1}{2} t^{-\frac{1}{2}} \text{d}t
\]

\[
F(t) = \frac{1}{\Gamma\left(\frac{1}{v^2}\right)} \int_0^t e^{-\frac{t}{v^2}} \frac{1}{2} t^{-\frac{1}{2}} \text{d}t
\]

The censored mean of t is

\[
\bar{t}_0 = \frac{1}{\Gamma\left(\frac{1}{v^2}\right)} \int_0^{t_0} e^{-\frac{t}{v^2}} \frac{1}{2} t^{-\frac{1}{2}} \text{d}t
\]

\[
1 - p_0
\]

where \(t_0\) denotes the value of t corresponding to the cutoff point corresponding to \(p_0\). \(p_0\) will be taken as .01 for illustrative purposes.
We integrate the numerator by parts

let \( dw = e^{-t} \, dt \)
\[ u = t \]
\[ W = -e^{-t} \]
\[ du = \frac{1}{v^2} t \]
\[ \frac{1}{v^2} - 1 \]

\[
\frac{1}{\Gamma(\frac{1}{v^2})} \left\{ wu - \int wdu \right\} = \frac{-e^{-t\frac{1}{v^2}}}{\Gamma(\frac{1}{v^2})} \left[ \frac{t}{v^2} + \int_{0}^{t} e^{-t\frac{1}{v^2}} \, dt \right]_{0}^{t} = \frac{-e^{-t\frac{1}{v^2}}}{\Gamma(\frac{1}{v^2})} \left[ \frac{t}{v^2} \right]_{0}^{t} + \frac{1}{v^2}
\]

(6) \[ \tilde{t}_0 = \frac{\frac{1}{v^2} - 1}{1 - .01} \]

\[
\tilde{t}_0 = \frac{-e^{-t_0} \frac{1}{v^2} + \frac{1}{v^2} (\cdot99)}{.99 \Gamma(\frac{1}{v^2})} = \frac{-e^{-t_0} \frac{1}{v^2}}{.99 \Gamma(\frac{1}{v^2})} + \frac{1}{v^2}
\]

(7) \[
\tilde{t}_0 = \frac{-e^{-t_0} \frac{1}{v^2}}{.99 \Gamma(\frac{1}{v^2})} + \frac{1}{v^2}
\]

The mean of \( t \) for the entire distribution is \( \frac{1}{v^2} \).

The ratio of the mean of the entire distribution to the censored distribution is:

(8) \[
\frac{\tilde{t}}{\tilde{t}_0} = \frac{\frac{1}{v^2}}{1 - \frac{e^{-t_0} \frac{1}{v^2}}{.99 \Gamma(\frac{1}{v^2})}}
\]

or multiplying by \( .99/v^2 \)

(9) \[
\frac{\tilde{t}}{\tilde{t}_0} = \frac{.99}{e^{-t_0} \frac{1}{v^2}} = \frac{.99}{\Gamma(\frac{1}{v^2} + 1)}
\]
The value of (9) can be found by specifying \( v \), the coefficient of variation. The value of \( t_0 \) corresponding to given percentage points \( (P_0) \) are obtained from tables of the Incomplete Gamma Function. The value of \( \Gamma\left(\frac{1}{2} + 1\right) \) is found from tables of the Logarithm of the Gamma Function.

Table 1 at the end of the paper gives the computed values for coefficients of variation of 100, 145, 200 and 315 percent.

3. An Illustrative Example

A survey made in June 1959 in Mississippi will be used to illustrate the procedures. The computations are shown below for two characteristics. The direct estimate for all hogs and pigs based on the reciprocal of the probability of selection for a stratified random area sample of segments was computed in the usual way. A coefficient of variation of 200 percent is assumed and a cutoff point corresponding to \( P_0 = .010 \) was used. The average number of hogs per segment for the complete survey was 12.52 and \( t_0 \) is found from Table 1 to be 9.80. Therefore, the approximate value which is expected to cut off 1 percent of the upper tail is 12.52 x 9.80 or 123. The approximate value of \( t_0 \) has been found to be satisfactory since individual values of the characteristic are widely spaced in the upper tail. The actual number of segments cut off by 123 was 1.2 percent rather than the expected 1.0. The adjustment \( \bar{t}/\bar{t}_0 \) corresponding to .012 is used rather than .010 since it corresponds to the actual percent cutoff by \( t_0 = 123 \). The same procedures are used for the second characteristic - all farm chickens.

Example 1

All Hogs and Pigs: Assume coefficient of variation 2.00 and selected cutoff point corresponding to \( P(t_0) = .99 \).

1. Estimated total number of hogs and pigs based on complete sample 946,074
2. Total number sampling units in State = 75,583
3. Estimated mean number hogs and pigs per segment based on complete sample \( \frac{946,074}{75,583} = 12.52 = \bar{t} \)
4. \( t_0 = 12.52 \times 9.80 = 122.7 \) use 123
5. Segments with total hogs and pigs greater than 123
<table>
<thead>
<tr>
<th>Strata and segment</th>
<th>No. of pigs &amp; hogs</th>
<th>Expanded number</th>
<th>Number of segments in universe</th>
</tr>
</thead>
<tbody>
<tr>
<td>ident.</td>
<td>reported:</td>
<td>number:</td>
<td>greater 123</td>
</tr>
<tr>
<td>1 - 1</td>
<td>150</td>
<td>97.980</td>
<td>14,697</td>
</tr>
<tr>
<td>4 - 1</td>
<td>150</td>
<td>137.429</td>
<td>20,614</td>
</tr>
<tr>
<td>5 - 1</td>
<td>150</td>
<td>279.486</td>
<td>41,923</td>
</tr>
<tr>
<td>6 - 1</td>
<td>162</td>
<td>383.864</td>
<td>62,186</td>
</tr>
<tr>
<td>State totals</td>
<td>XX</td>
<td>XX</td>
<td>139,420</td>
</tr>
</tbody>
</table>

6. Mean number of hogs and pigs per segment for censored distribution

   Total is: 946,074 - 139,420 = 806,654
   Mean is: 806,654 \( \div (75,583 - 898) \) = 10.8007

7. Fraction of distribution censored \( \frac{898}{75,583} = .012 \)

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

   \[ \bar{t}_0 = 10.8007 \times 1.162 = 12.5504 \]

9. Total hogs and pigs for complete distribution is:

   \[ 12.5504 \times 75,583 = 948,597 \]

   In this case, the estimated total is essentially unchanged.

**Example 2**

All Farm Chickens: Assume coefficient of variation 2.00 and selected cutoff point corresponding to \( F(t_0) = .99 \)

1. Estimated total number farm chickens based on complete sample = 10,535,872

2. Total number of sampling units in State = 75,583.

3. Estimated mean number of farm chickens per segment based on complete sample \[ \frac{10,535,872}{75,583} = 139.4 \]
4.  \( t_o = 139.4 \times 9.80 = 1366 \)

5. Segments with total farm chickens greater 1366

| Strata : No. of and :farm with : Expanded : Number of segment : chickens : \( \frac{1}{P_i} \) : number : segments in ident. : reported : farm : universe greater chickens 1366 (= 1/P_i) |
|-----------------|-----------------|-----------------|-----------------|
| 1 - 1           | 28,000          | 97.980          | 2,743,440       | 98             |
| 3 - 1           | 3,030           | 216.704         | 656,613         | 217            |
| 8 - 1           | 4,215           | 278.720         | 1,174,805       | 279            |
| 9 - 1           | 1,430           | 440.514         | 629,935         | 441            |
| State totals:   | XX              | XX              | 5,204,793       | 1,035 = M      |

6. Mean number of farm chickens per segment is obtained from the total as follows:
   - Total is: \( 10,535,872 - 5,204,793 = 5,331,079 \)
   - Mean is: \( \frac{5,331,079}{(75,583 - 1,035)} = 71.512 \)

7. Fraction of distribution censored \( 1,035/75,583 = 0.0136 \) (approx. 0.01)

8. Mean number farm chickens for the complete distribution is:
   \( \bar{t}_o = 71.512 \times 1.182 - 84.527 \)

9. Total farm chickens for complete distribution is:
   \( 84.527 \times 75,583 = 6,388,804 \)

In this case, the estimated total and variance (not shown) are markedly reduced.
Table 1. Values of $t_o$ and $\bar{t}/\bar{t}_o$ for specified values of $P$ corresponding to the portion of the upper tail censored for a Pearson Type III Distribution with $\bar{t} = 1.0$

<table>
<thead>
<tr>
<th></th>
<th>$v = 1.00$</th>
<th></th>
<th>$v = 1.42$</th>
<th></th>
<th>$v = 2.00$</th>
<th></th>
<th>$v = 3.17$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_o$</td>
<td></td>
<td>$P_o$</td>
<td>$\bar{t}/\bar{t}_o$</td>
<td>$t_o$</td>
<td>$P_o$</td>
<td>$\bar{t}/\bar{t}_o$</td>
<td>$t_o$</td>
</tr>
<tr>
<td>6.21</td>
<td>0.002</td>
<td>1.012</td>
<td>9.55</td>
<td>0.002</td>
<td>1.021</td>
<td>15.10</td>
<td>0.002</td>
</tr>
<tr>
<td>5.52</td>
<td>0.004</td>
<td>1.022</td>
<td>8.29</td>
<td>0.004</td>
<td>1.038</td>
<td>12.76</td>
<td>0.004</td>
</tr>
<tr>
<td>5.12</td>
<td>0.006</td>
<td>1.032</td>
<td>7.55</td>
<td>0.006</td>
<td>1.053</td>
<td>11.40</td>
<td>0.006</td>
</tr>
<tr>
<td>4.83</td>
<td>0.008</td>
<td>1.041</td>
<td>7.04</td>
<td>0.008</td>
<td>1.068</td>
<td>10.44</td>
<td>0.008</td>
</tr>
<tr>
<td>4.61</td>
<td>0.010</td>
<td>1.049</td>
<td>6.63</td>
<td>0.010</td>
<td>1.081</td>
<td>9.80</td>
<td>0.010</td>
</tr>
<tr>
<td>4.42</td>
<td>0.012</td>
<td>1.057</td>
<td>6.31</td>
<td>0.012</td>
<td>1.095</td>
<td>9.16</td>
<td>0.012</td>
</tr>
<tr>
<td>4.27</td>
<td>0.014</td>
<td>1.065</td>
<td>6.05</td>
<td>0.014</td>
<td>1.107</td>
<td>8.66</td>
<td>0.014</td>
</tr>
<tr>
<td>4.14</td>
<td>0.016</td>
<td>1.072</td>
<td>5.80</td>
<td>0.016</td>
<td>1.120</td>
<td>8.24</td>
<td>0.016</td>
</tr>
<tr>
<td>4.02</td>
<td>0.018</td>
<td>1.079</td>
<td>5.60</td>
<td>0.018</td>
<td>1.133</td>
<td>7.98</td>
<td>0.018</td>
</tr>
<tr>
<td>3.91</td>
<td>0.020</td>
<td>1.086</td>
<td>5.41</td>
<td>0.020</td>
<td>1.145</td>
<td>7.54</td>
<td>0.020</td>
</tr>
<tr>
<td>3.51</td>
<td>0.030</td>
<td>1.121</td>
<td>4.71</td>
<td>0.030</td>
<td>1.204</td>
<td>6.22</td>
<td>0.030</td>
</tr>
<tr>
<td>3.22</td>
<td>0.040</td>
<td>1.155</td>
<td>4.22</td>
<td>0.040</td>
<td>1.261</td>
<td>5.48</td>
<td>0.040</td>
</tr>
</tbody>
</table>

References: Tables for Statistician and Biometricians, Part I, Table XXXI - Karl Pearson (1924) Cambridge University Press

Tables of the Incomplete Gamma Function, Table II - Karl Pearson (1922) Cambridge University Press