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

ESTIMATION OF CHARACTERISTICS FROM RESTRICTED OR "CENSORED" SAMPLE DATA

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 ($\frac{m}{n} \leq .02$) 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 $F(x) \geq .99$ are not known. It is assumed that the sample extreme values beyond $F(x) = .99$ occur with the expect 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:

$$(1) \qquad dF(x) = ke^{-\frac{1}{v^2}(\frac{x}{\bar{x}})} \quad (\frac{x}{\bar{x}})^{\frac{1}{v^2} - 1} \quad dx$$

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:

$$(2) \qquad dF(R) = ke^{-\frac{R}{v^2}} \quad R^{\frac{1}{v^2} - 1} \quad dR$$

and the distribution is determined by v since

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

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

$$(3) \qquad dF(t) = k'e^{-t} \quad t^{\frac{1}{v^2} - 1} \quad dt$$

$$(4) \qquad F(t) = \frac{1}{\Gamma(\frac{1}{v^2})} \int_0^t e^{-t} \quad t^{\frac{1}{v^2}} \quad dt$$

The censored mean of t is

$$(5) \qquad \bar{t}_o = \frac{\frac{1}{\Gamma(\frac{1}{v^2})} \int_0^{t_o} e^{-t} \quad t^{\frac{1}{v^2}} \quad dt}{1 - P_o}$$

where t_o denotes the value of t corresponding to the cutoff point corresponding to P_o. P_o will be taken as .01 for illustrative purposes.

We integrate the numerator by parts

$$\text{let} \quad dw = e^{-t}\,dt \qquad \mu = t^{\frac{1}{v^2}}$$

$$W = -e^{-t} \qquad d\mu = \frac{1}{v^2}\; t^{\frac{1}{v^2} - 1}$$

$$(6) \qquad \bar{t}_o = \frac{\frac{1}{\Gamma(\frac{1}{v^2})}\{w\mu - \int w d\mu\}}{1 - .01} = \frac{\frac{-e^{-t}t^{\frac{1}{v^2}}}{\Gamma(\frac{1}{v^2})}\Big|_o^{t_o} + \frac{1}{v^2}\int_o^{t_o} \frac{e^{-t}t^{\frac{1}{v^2}-1}\,dt}{\Gamma(\frac{1}{v^2})}}{.99}$$

$$(7) \qquad \bar{t}_o = \frac{-e^{-t_o}\; t'^{\frac{1}{v^2}}}{.99\;\Gamma(\frac{1}{v^2})} + \frac{\frac{1}{v^2}\,(.99)}{.99} = \frac{-e^{-t_o}\; t_o^{\frac{1}{v^2}}}{.99\;\Gamma(\frac{1}{v^2})} + \frac{1}{v^2}$$

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:

$$(8) \qquad \frac{\bar{t}}{\bar{t}_o} = \frac{\frac{1}{v^2}}{\frac{1}{v^2} - \frac{e^{-t_o}\; t_o^{\frac{1}{v^2}}}{.99\;\Gamma\;(\frac{1}{v^2})}}$$

or multiplying by $.99/v^2$

$$(9) \qquad \frac{\bar{t}}{\bar{t}_o} = \frac{.99}{.99 - \frac{e^{-t_o}\; t_o^{\frac{1}{v^2}}}{\frac{1}{v^2}\,\Gamma\;(\frac{1}{v^2})}} = \frac{.99}{.99 - \frac{e^{-t_o}\; t_o^{\frac{1}{v^2}}}{\Gamma\;(\frac{1}{v^2} + 1)}}$$

The value of (9) can be found by specifying v, the coefficient of variation. The value of t_o corresponding to given percentage points $(P_o$ are obtained from tables of the Incomplete Gamma Function. The value of $\Gamma(\frac{1}{v^2} + 1)$ 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_o = .010 was used. The average number of hogs per segment for the complete survey was 12.52 and t_o 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_o 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 $\bar{t}/\bar{t}_o$ corresponding to .012 is used rather than .010 since it corresponds to the actual percent cut-off by t_o = 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 $F(t_o)$ = .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_o = 12.52 x 9.80 = 122.7 use 123

5. Segments with total hogs and pigs greater than 123

Strata and segment ident.	No. of hogs & pigs reported	$\frac{1}{P_i}$	Expanded number hogs & pigs	Number of segments in universe greater 123 (= $1/P_i$)
1 - 1	150	97.980	14,697	98
4 - 1	150	137.429	20,614	137
5 - 1	150	279.486	41,923	279
6 - 1	162	383.864	62,186	384
State totals	XX	XX	139,420	898 = M

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

   Total is: 946,074 - 139,420 = 806,654

   Mean is: 806,654 ÷ (75,583 - 898) = 10.8007

7. Fraction of distribution censored 898/75,583 = .012

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

   $\bar{t}_o$ = 10.8007 x 1.162 = 12.5504

9. Total hogs and pigs for complete distribution is:

   12.5504 x 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 $F(t_o)$ = .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 x 9.80 = 1366

5. Segments with total farm chickens greater 1366

Strata and segment ident.	No. of farm with chickens reported	$\frac{1}{P_i}$	Expanded number farm chickens	Number of segments in universe greater 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: 5,331,079 ÷ (75,583 - 1,035) = 71.512

7. Fraction of distribution censored 1,035/75,583 = .0136 (approx. .014)

8. Mean number farm chickens for the complete distribution is:

   $\bar{t}_o$ = 71.512 x 1.182 - 84.527

9. Total farm chickens for complete distribution is:

   84.527 x 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

v = 1.00			v = 1.42			v = 2.00			v = 3.17		
t_o	P_o	$\bar{t}/\bar{t}_o$	t_o	P_o	$\bar{t}/\bar{t}_o$	t_o	P_o	$\bar{t}/\bar{t}_o$	t_o	P_o	$\bar{t}/\bar{t}_o$
6.21	.002	1.012	9.55	.002	1.021	15.10	.002	1.037	28.04	.002	1.074
5.52	.004	1.022	8.29	.004	1.038	12.76	.004	1.065	22.64	.004	1.131
5.12	.006	1.032	7.55	.006	1.053	11.40	.006	1.091	19.59	.006	1.184
4.83	.008	1.041	7.04	.008	1.068	10.44	.008	1.116	17.48	.008	1.234
4.61	.010	1.049	6.63	.010	1.081	9.80	.010	1.124	15.40	.010	1.301
4.42	.012	1.057	6.31	.012	1.095	9.16	.012	1.162	14.64	.012	1.331
4.27	.014	1.065	6.05	.014	1.107	8.66	.014	1.182	13.55	.014	1.381
4.14	.016	1.072	5.80	.016	1.120	8.24	.016	1.207	12.65	.016	1.429
4.02	.018	1.079	5.60	.018	1.133	7.98	.018	1.228	11.88	.018	1.476
3.91	.020	1.086	5.41	.020	1.145	7.54	.020	1.250	11.19	.020	1.526
3.51	.030	1.121	4.71	.030	1.204	6.22	.030	1.366	8.67	.030	1.775
3.22	.040	1.155	4.22	.040	1.261	5.48	.040	1.463	7.01	.040	2.040

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