BIOMETRICS INFORMATION

(You're 95% likely to need this information)

PAMPHLET NO. # 43 DATE: January 5, 1993

SUBJECT: Standard Error Formulas for Cluster Sampling (Unequal Cluster Sizes)

Simple random sampling is not always the most practical or inexpensive method of sampling. The population might be so extensive that construction of a sampling frame (list of sampling units) would be difficult, or the sampling units might be so widely dispersed that it would be prohibitively expensive to collect the data for a simple random sample. In such cases, it may be more convenient to sample groups or clusters of sampling units - plots or rows of trees, for example. This approach is known as **cluster sampling** and is common in forestry surveys.

Measurements of the sampling units within a cluster are often **correlated**. In forestry, correlations typically arise because neighbouring trees compete for the same resources or tend to be exposed to the same pests and environment. Since the data from a cluster cannot be considered independent, the usual formulas for calculating standard errors do not apply to cluster sampling. The correct formulas (for unequal cluster sizes) are given below.

Notation

Assume **n** clusters are selected at random from a population of **N** clusters and define the following symbols, for i = 1, 2,..., n:

m_i = number of sampling units in cluster i

$m = \sum_{i=1}^{n} m_i$, total number of sampling units in all clusters

$\overline{m}$ = m/n, average number of sampling units per cluster

y_{ij} = observed value or the variable of interest for sampling unit j in cluster i ($j = 1, 2,..., m_i$)

$y_{i.} = \sum_{j=1}^{m_i} y_{ij}$ sum of the observations in cluster i

$\overline{y}_i = y_{i.}/m_i$, mean of the observations in cluster i.

Ratio estimate of a mean

The population mean is estimated by the overall average,

$$\bar{y} = \sum_{i=1}^{n} y_{i.} / m \qquad (1)$$

which can also be written as a weighted average of the cluster means,

$$\bar{y} = \sum_{i=1}^{n} m_i \bar{y}_i / m$$

In many practical applications, the number of sampling units, $\mathbf{m_i}$, varies from one cluster to the next. Therefore, unlike simple random sampling, for which the sample size **m** is fixed, the denominator in the estimate (1), as well as the numerator, is subject to random variation. In this situation, $\bar{y}$ is called a **ratio estimate** because it is a ratio of two random variables. Other estimates of the mean are sometimes used when the cluster sizes are unequal, but only the ratio estimate will be considered here.

The estimated standard error of the ratio estimate $\bar{y}$ is:

$$\text{std. err. } (\bar{y}) = \sqrt{\left(1 - \frac{n}{N}\right) \frac{s^2}{n\bar{m}^2}} \qquad (2)$$

with s^2 defined as follows:

$$s^2 = \sum_{i=1}^{n} m_i^2 (\bar{y}_i - \bar{y})^2 / (n-1)$$

An equivalent expression for s^2, which is useful in computations, is:

$$s^2 = \left[\sum_{i=1}^{n} y_{i.}^2 - 2\bar{y} \sum_{i=1}^{n} m_i y_{i.} + \bar{y}^2 \sum_{i=1}^{n} m_i^2 \right] / (n-1) \qquad (3)$$

The standard error in (2) is based on a large sample approximation and should be applied only when the number of clusters in the sample, **n**, is reasonably large. If **n** is small relative to the total number of clusters, **N**, then the finite population correction, 1-n/N, is approximately equal to 1 and can be ignored.

Ratio estimate of a proportion

Formulas (1) to (3) apply to the estimation of a general mean. The corresponding formulas for a proportion are obtained as a special case by defining an indicator variable: $y_{ij} = 1$, if the respective sampling unit has the attribute of interest; otherwise, $y_{ij} = 0$. Therefore, the ratio

estimate of the proportion of the population with a particular attribute is:

$$\hat{p} = \sum_{i=1}^{n} a_i / m \qquad (4)$$

with $a_i = y_{i.}$ = the number of sampling units that are in cluster **i** and have the attribute.

The standard error of $\hat{p}$ is estimated by (2). In this case, the computational formula for s^2 is, from (3):

$$s^2 = \left[\sum_{i=1}^{n} a_i^2 - 2\hat{p} \sum_{i=1}^{n} m_i a_i + \hat{p}^2 \sum_{i=1}^{n} m_i^2 \right] / (n-1) \qquad (5)$$

For more information about the statistical properties of the ratio estimate, and other estimates that can be used when the cluster sizes are equal or unequal, refer to Cochran (1977, pp. 233-273).

Example

To estimate the average height of the trees in a spruce stand, and the proportion with leader weevil, a cluster sample of trees was selected by randomly locating 17 circular plots (3.99m radius) throughout the stand. Height and leader weevil presence or absence were recorded for all trees found within the plot boundaries. In this sampling design, the trees in a plot constitute a cluster and each tree is a sampling unit. The total number of trees in the sample is 85, with the number of trees per plot ranging from 2 to 9.

Formulas (1) to (5) were used to calculate the required estimates and their estimated standard errors. A SAS program for performing the necessary computations is listed at the end of the pamphlet. For comparison, the standard errors were also calculated with `PROC MEANS`, which assumes the data are generated by a simple random sample. The sampling fraction n/N was assumed to be negligible so that 1-n/N = 1 for both computations. This is a reasonable assumption if the surveyed area ($850m^2$) is small relative to the total area of the stand.

The results of the computations are given after the program listing. Notice that, because `PROC MEANS` assumes that there are 85 independent observations when there are in fact only 17 independent clusters, it under-estimates the standard errors by about 50%.

Reference:

Cochran, W.G. 1977. Sampling techniques (third edition). Wiley, New York.

Contact: Amanda F. Linnell Nemec
652-4517

SAS program

```
/*****************************************************************/
/* Read data from input file;                                    */
/* compute and print plot summary statistics.                    */
/*                                                               */
/* Input variables:  PLOT   = plot number                        */
/*                   TRNO   = tree number                        */
/*                   HT     = tree height                        */
/*                   WEEVIL = 1 if tree has weevil, 0 otherwise  */
/*****************************************************************/
DATA TREES;
  INFILE 'A:TREES.DAT' MISSOVER;
  INPUT PLOT TRNO HT WEEVIL;

PROC SORT DATA=TREES;                        /* Calculate plot sums. */
  BY PLOT;
DATA PLOTS;
  SET TREES;
  BY PLOT;
  IF FIRST.PLOT THEN DO;
    SUMY=0; SUMM1=0; SUMMY=0; SUMYY=0; SUMMM1=0;
    SUMA=0; SUMM2=0; SUMMA=0; SUMAA=0; SUMMM2=0;
  END;
  IF HT^=. THEN DO;
    SUMY+HT; SUMM1+1;
  END;
  IF WEEVIL^=. THEN DO;
    SUMA+WEEVIL; SUMM2+1;
  END;
  IF LAST.PLOT THEN DO;
    SUMMY=SUMY*SUMM1;
    SUMYY=SUMY*SUMY;
    SUMMM1=SUMM1*SUMM1;
    SUMMA=SUMA*SUMM2;
    SUMAA=SUMA*SUMA;
    SUMMM2=SUMM2*SUMM2;
    OUTPUT;
  END;
  DROP TRNO HT WEEVIL;

                                     /* Print plot summary statistics. */
PROC PRINT DATA=PLOTS NOOBS LABEL UNIFORM;
  TITLE 'PLOT SUMMARY STATISTICS';
  VAR PLOT SUMM1 SUMMM1 SUMY SUMYY SUMMY SUMA SUMAA SUMMA;
  SUM SUMM1 SUMMM1 SUMY SUMYY SUMMY SUMA SUMAA SUMMA;
  LABEL PLOT  = 'Plot'
     SUMM1 = 'No. Trees (m)'
     SUMMM1 = 'm*m'
     SUMY  = 'Sum Heights (y.)'
     SUMYY = 'y.*y.'
     SUMMY = 'm*y.'
     SUMA  = 'No. With Weevil (a)'
     SUMAA = 'a*a'
     SUMMA = 'm*a';
```

```
/*****************************************************************/
/* Calculate means and standard errors                           */
/* Method 1: Cluster sample formula                              */
/*                                                               */
/* Output variables:                                             */
/*                                                               */
/* MHT        = mean height (ratio estimate)                     */
/* PWEEVIL    = proportion of trees with weevil (ratio estimate) */
/* MAVG1      = average cluster size excluding trees with missing*/
/*              height data                                      */
/* MAVG2      = average cluster size excluding trees with missing*/
/*              weevil data                                      */
/* VHT        = s^2 for height                                   */
/* VWEEVIL    = s^2 for the proportion of trees with weevil      */
/* SMHT       = standard error of MHT                            */
/* SPWEEVIL   = standard error of PWEEVIL                        */
/*****************************************************************/
                                        /* Sum over plots. */
PROC SUMMARY DATA=PLOTS;
  VAR SUMY SUMM1 SUMMY SUMYY SUMMM1 SUMA SUMM2 SUMMA SUMAA SUMMM2;
  OUTPUT OUT=SUMS SUM= N=NPLOTS;
                                        /* Calculate ratio estimates */
DATA ESTIMATE;                          /* and standard errors.      */
  SET SUMS;
  MHT=SUMY/SUMM1;
  MAVG1=SUMM1/NPLOTS;
  PWEEVIL=SUMA/SUMM2;
  MAVG2=SUMM2/NPLOTS;
  VHT=(SUMYY-2*MHT*SUMMY+MHT*MHT*SUMMM1)/(NPLOTS-1);
  VWEEVIL=(SUMAA-2*PWEEVIL*SUMMA+PWEEVIL*PWEEVIL*SUMMM2)/(NPLOTS-1);
  SMHT=SQRT(VHT/NPLOTS)/MAVG1;
  SPWEEVIL=SQRT(VWEEVIL/NPLOTS)/MAVG2;
                                        /* Print estimates. */
PROC PRINT LABEL NOOBS;
  TITLE 'METHOD 1 - CLUSTER SAMPLE FORMULA';
  VAR NPLOTS SUMM1 MHT SMHT PWEEVIL SPWEEVIL;
  LABEL NPLOTS  = 'No. Plots'
     SUMM1   = 'No. Trees'
     MHT    = 'Average Height (cm)'
     SMHT   = 'Std. Err.'
     PWEEVIL = 'Prop. with Weevils'
     SPWEEVIL = 'Std. Err.';

/*****************************************************************/
/* Calculate means and standard errors                           */
/* Method 2: Random sample formula                               */
/*****************************************************************/
PROC MEANS DATA=TREES MEAN STDERR N;
  TITLE 'METHOD 2 - SIMPLE RANDOM SAMPLE FORMULA';
  VAR HT WEEVIL;
RUN;
```

Program output

PLOT SUMMARY STATISTICS

Plot	No. Trees (m)	m*m	Sum Heights (y.)	y.*y.	m*y.	No. With Weevil (a)	a*a	m*a
1	9	81	2214	4901796	19926	0	0	0
2	3	9	262	68644	786	0	0	0
3	2	4	206	42436	412	2	4	4
4	3	9	270	72900	810	1	1	3
5	4	16	386	148996	1544	3	9	12
6	4	16	455	207025	1820	4	16	16
7	2	4	264	69696	528	0	0	0
8	2	4	248	61504	496	0	0	0
9	3	9	350	122500	1050	3	9	9
10	6	36	1766	3118756	10596	2	4	12
11	9	81	1752	3069504	15768	8	64	72
12	8	64	1634	2669956	13072	2	4	16
13	9	81	1210	1464100	10890	3	9	27
14	4	16	419	175561	1676	4	16	16
15	7	49	680	462400	4760	1	1	7
16	5	25	1190	1416100	5950	0	0	0
17	5	25	568	322624	2840	1	1	5
	85	529	13874	18394498	92924	34	138	199

METHOD 1 - CLUSTER SAMPLE FORMULA

No. Plots	No. Trees	Average Height (cm)	Std. Err.	Prop. with Weevils	Std. Err.
17	85	163.224	17.795	0.4	0.096589

METHOD 2 - SIMPLE RANDOM SAMPLE FORMULA

N Obs	Variable	N	Mean	Std Error
85	HT	85	163.2235294	9.8521902
	WEEVIL	85	0.4000000	0.0534522