STRATA BOUNDARY DETERMINATION

William E. Winkler*, bwinkler@census.gov

Keywords: sampling, variance estimation

This paper provides a method of strata boundary determination that generalizes the method of Dalenius and Hodges. The method holds for general populations. It makes no restrictive assumptions about the population distribution or about the ignorability of the finite population correction.

1. INTRODUCTION

This paper describes a strategy for strata boundary determination that generalizes the classical approach of Dalenius and Hodges. Given a fixed sample size and a fixed number of strata, the Dalenius-Hodges method provides a quick means of determining strata boundaries that approximately minimizes coefficients of variation (*cv*). Applying their method requires assuming that the finite population correction (*fpc*) can be ignored, that the underlying population distribution is continuous, and that the probability density of the variable of interest is constant within strata. The minimization basically depends on one variable, strata boundary break points.

The method of this paper allows determining strata boundaries for general populations. For a given sample size, approximate minimization depends on five discrete variables: number of strata, sample allocations within strata, population variance within strata, population size within strata, and strata boundary break points. If sample design considerations require that certain stratification variables, say number of strata and sample allocations within strata, be restricted, minimization can still be performed. The methods of this paper depend on standard ideas in sampling theory (e.g. Cochran 1977).

The outline of the paper is as follows. In the second section, we present notation and definitions. We also describe the stratification methods of Dalenius and Hodges (1959) and of Ekman (1959). The third section contains a description of the empirical data base and the stratification method. In the fourth section, we prove the main theoretical result and present some empirical examples. The fifth section gives discussion of the relationship of the results with some of the generalizations of the Dalenius-Hodges method. In particular, the relationship with the method of Ekman is noted. The final section contains a summary.

2. BACKGROUND

In this section, we provide notation and definitions and we summarize the stratification methods of Dalenius and Hodges and Ekman.

2.1. Notation and Definitions

In this section, we summarize notation that is standard in sampling theory (e.g., Cochran 1977). The suffix h denotes the stratum and the suffix i the unit within the stratum. The following all refer to stratum h :

N_h total number of units,

n_h numbers of units in sample,

y_{hi} value obtained for the ith unit,

y_h break point between stratum h and stratum h+1,

$$Y_h = \sum_{j=1}^{N_h} Y_{hj}$$ true total,

$$\hat{Y}_h = (N_h \ / \ n_h) \cdot \sum_{i=1}^{n_h} y_{hi}$$ estimated total, and

$$S_h^2 = (N_h \ / \ N_h - 1)) \sum_{j=1}^{N_h} (y_{hi} - (Y_h \ / \ N_h))^2$$ variance of true total.

For the true total and its variance, the sum is over all N_h values. For the estimated total, the sum is over the n_h values in the sample.

2.2. Previous Stratification Methods

To assure proper understanding of the empirical comparisons, we summarize the methods of Dalenius and Hodges and of Ekman (see e.g., Cochran 1977, pp. 127-131).

Dalenius and Hodges (1959) stratify to make the quantities

$$N_h^{1/2} \cdot (y_h - y_{h-1}),\ h = 1,2,...,\ L, \qquad (1)$$

approximately equal.

Ekman (1959) stratifies to make the quantities

$$N_h \cdot (y_h - y_{h-1}),\ h = 1,2,...,\ L, \qquad (2)$$

approximately equal.

Each method ignores the finite population correction $(N_h - n_h) \ / \ N_h$ and each uses an easily computed surrogate $(y_h - y_{h-1})$ for the standard deviation S_h .

3. DATA BASE AND METHODOLOGY

This section provides descriptions of the data base that is used in the empirical analyses and the stratification methodology.

3.1. Empirical Data Base

The data base used in the empirical analyses consists of 1106 records, each having three quantitative data elements. Figures 1, 2 and 3 illustrate the skewness of each of the variables in the population. The variables are plotted on a log base 4 scale. Nonzero quantitative values vary from 61,000 to 1 for the first variable, from 76,000 to 1 for the second, and from 241,000 to 1 for the third.

Prior to stratification, the data base is sorted by descending order of size in the variable of interest.

3.2. Methodology

To minimize the variance when sample size is fixed we stratify to try to assure that the quantities

$$N_h^{1/2} \cdot (N_h - n_h)^{1/2} \cdot S_h \ / \ n_h \qquad (3)$$

are constant for h = 1, 2,..., L. As the underlying population distribution is discrete, we choose the stratification among a finite number of stratifications for which approximate equality in (3) holds. The theorem of section 4.1 proves the validity of our procedure.

The chief differences between our method and the methods of Dalenius and Hodges and of Ekman are:

(a) we account for the finite population correction

(b) we use the standard deviation S_h instead of the surrogate $(y_h - y_{h-1})$, and

(c) we allow choice among a finite number of stratifications for which approximate equality in (3) holds.

In actual practice, statisticians using the methods of Dalenius and Hodges and of Ekman also allow choice among stratifications approximately satisfying formulas (1) and (2), respectively.

4. RESULTS

The section is divided into two subsections. The first contains the theoretical results. The second contains empirical results.

4.1. Theoretical Results

The proof is a generalization of the classical proof of Neyman (1934). We will assume that a minimum of two units are sampled in each noncertainty stratum. The theorem is valid for sampling strategies in which the number of strata L is fixed. In a remark following the theorem, we will describe optimization for the case when L is allowed to vary.

Theorem. In stratified random sampling with a linear cost function of the form (5) (see below), the variance of the estimated total $\hat{Y}_{st}$ is a minimum for a specified cost, and the cost is a minimum for a specified variance $V(\hat{Y}_{st})$, when the quantities

$$N_h^{1/2} \cdot (N_h - n_h)^{1/2} \cdot S_h \ / \ n_h \qquad (4)$$

are constant for h = 1, 2,..., L.

Proof. We have

$$C = \sum_{h=1}^{L} n_h \qquad \text{and} \qquad (5)$$

$$V = \sum_{h=1}^{L} N_h \cdot (N_h - n_h) \cdot S_h^2 \ / \ n_h \ . \qquad (6)$$

Choosing n_h , h = 1, 2, ..., L and L to minimize C for fixed V and choosing n_h , h = 1, 2, ..., L; S_h , h = 1, 2, ..., L; N_h , h = 1, 2, ..., L and L to minimize V for fixed C is equivalent to minimizing the product

$$V \cdot C = (\sum_{h=1}^{L} n_h) \cdot (\sum_{h=1}^{L} N_h \cdot (N_h - n_h) \cdot S_h^2 \; / \; n_h).$$

Applying the Cauchy-Schwartz inequality, we have that

$$V \cdot C \geq (\sum_{h=1}^{L} n_h \cdot N_h \cdot (N_h - n_h) \cdot S_h^2 \; / \; n_h) \qquad (7)$$

.

with equality holding if and only if

$$N_h^{1/2} \cdot (N_h - n_h)^{1/2} \cdot S_h \; / \; n_h$$

is constant for h = 1,2,..., L.

We note that we can only expect to get exact equalities in (4) when underlying distributions are continuous. With real-world data, the underlying distributions are discrete. We can, thus, obtain true minima by examining a finite set of stratifications for which approximate equality holds.

Another stratification method we use assures that

$$N_h^{1/2} \cdot (N_h - n_h)^{1/2} \cdot S_h \; / \; n_h^{1/2} \qquad (8)$$

are constant for h = 1,2,..., L. Equality in (6) assures that each of the terms in the variance given by equation (6) are equal. If each n_h is equal, then the stratification yielding equalities in (8) agrees with the stratification yielding equalities in (4). We denote the first stratification method by S1 and the second by S2.

Remark. If we allow the number of strata L to vary, then the minimum *cv* for a fixed sample size n is obtained when each noncertainty stratum has a sample allocation of two units. Optimization, thus, primarily depends on determining the number of units sampled with certainty. The procedure for finding the minimum *cv* consists of a straightforward grid search for the L with corresponding break points obtained using the method of the theorem.

If we increase L (i.e., decrease the size of the certainty stratum) and the *cv* increases, then we reverse the direction of our search (i.e., decrease L).

There are necessarily n-2L units are in the certainty stratum. With skewed populations, the L yielding a minimum cv is generally obtained when the first noncertainty stratum is sampled at a rate less than 50 percent.

4.2. Empirical Results

We present a comparison of the methods of this paper with the methods of Dalenius and Hodges and of Ekman. The comparison uses the empirical data base. In each case, the number of strata and the sample size within strata are fixed. Strata boundaries are then chosen that minimize the variance.

The empirical results show that the two methods of this paper are roughly equivalent and slightly better than the method of Ekman. All three methods are better than the method of Dalenius and Hodges (Table 1).

With three exceptions, method S1 of this paper performs best. The first two exceptions are associated with stratification C for variable 1 and stratification A with variable 3. Method S2 of this paper performs slightly better (.1518 versus .1541 and .1985 versus .2065, respectively).

The third exception occurs with stratification B for variable 2. The method of Ekman performs slightly better (.0818 versus .0832).

The stratification methodologies were tried on six additional populations having skewness properties similar to the population used in this paper. Results were similar to the results presented here. No unusual situations occurred that might contradict the empirical findings.

5. DISCUSSION

For general skewed populations, we would use either the stratification methods of this paper or the method of Ekman. As each of the methods is easy to program, we prefer having printouts that give a side-by-side comparison. The method of Dalenius and Hodges performs poorly primarily because the underlying probability density function is not constant in stratification intervals and the finite population correction cannot be ignored.

In determining strata boundaries, we used the square root of the probability density function for the given intervals. We did not use more sophisticated techniques of approximating the cumulative sum of the square root of the probability density function (see e.g. Cochran 1977, pp. 127-128, where the probability density function is referred to as the frequency count function). Such approximating techniques are not easily programmed.

6. SUMMARY

We have provided a new stratification methodology that yields slight improvements over the method of Ekman in skewed populations. Both the method of this paper and Ekman's method perform better than the method of Dalenius and Hodges.

* This paper reflects views of the author and are not necessarily those of the U.S. Bureau of the Census.

REFERENCES

Cochran, W. G. (1977), *Sampling Techniques (Third Edition)*, John Wiley and Sons, New York.

Dalenius, T. and Hodges, J. L. (1959), "Minimum Variance Stratification," *Journal of the American Statistical Association*, 54, 88-101.

Ekman, G. (1959), "An Approximation Useful in Univariate Stratification," *Annals of Mathematical Statistics*, 30, 219-229.

Table 1. Comparison of CV
Different Stratifying Methods

Method	Variable and Stratification							
	1A	1B	1C	2A	2B	2C	3A	3B
S1	.0810	.0965	.1541	.0793	.0832	.1468	.2065	.1041
S2	.0810	.1007	.1518	.0796	.0833	.1478	.1985	.1052
EK	.0815	.1091	.1556	.0801	.0818	.1575	.2081	.1144
D-H	.1147	.1023	.1720	.1048	.1462	.1620	.2207	.1285

1A- Var 1 Sample: 8 certainty 2,2,2,2,2 noncertainty, total-18.

1B- Var 1 Sample: 3 certainty, 2,2,4,8 noncertainty, total-19.

1C- Var 1 Sample: 4 certainty, 2,3,5 noncertainty, total-14.

2A- Var 2 Sample: 8 certainty, 2,2,2,2,2 noncertainty, total-18.

2B- Var 2 Sample: 6 certainty, 4,4,8 noncertainty, total-16.

2C- Var 2 Sample: 3 certainty, 4,4,8 noncertainty, total-19.

3A- Var 3 Sample: 4 certainty, 4,8 noncertainty, total-16.

3B- Var 3 Sample: 4 certainty, 3,4,5 noncertainty, total-16.