CONFIDENCE LIMITS FOR SMALL PROPORTIONS IN COMPLEX SAMPLES

Shulamith T. Gross and Martin R. Frankel
Baruch College of The City University of New York
Statistics Department. Box 513. 17 Lex Ave. NY, NY 10010

KEYWORDS: complex design, stratified sampling, cluster sampling, confidence intervals, small proportions, least favorable allocation.

ABSTRACT

Exact upper confidence limits for small proportions in stratified samples are derived. An algorithm for their computation which employs a new normal approximation for the case of large strata and a finite number of defectives is proposed. Using selected examples it is shown that the usual confidence intervals derived from the standard normal approximation can be highly misleading. The loss of efficiency of non-proportionate designs, vis-a-vis simple random sampling or proportionate designs for setting confidence limits on small proportions is studied in a variety of examples.

Exact upper confidence limits for small proportions are also derived for simple random samples of equal-size clusters, and a similar algorithm for their derivation is presented. The loss in efficiency due to clustering is shown to be proportional to the cluster size when no defectives are found in the sample. In other cases the loss is slightly smaller.

I. The problem.

Applications which rely on probability sampling to estimate population proportions which are very small have gained prominence in recent years. Examples include: estimation of the prevalence of a disease in a population in which it is known to be rare; quality assessment, evaluation and control in highly reliable production processes; estimation of error rates, or adjustments, in financial auditing.

In this paper we present a preliminary report on some exact results and an iterative algorithm for computing upper confidence intervals for small proportions in stratified samples of elements. The approach we take is completely model-free: probabilities arise entirely from the sampling procedure. In fact, the approach taken here is similar to that taken by Sedransk and Meyer (1978), and Smith and Sedransk (1983) in their work on the estimation of quantiles from stratified samples from finite populations. They present both a conservative confidence interval, which requires a prodigious amount of computation to determine, and short-cut approximation methods which are demonstrated to be valid in a large number of examples.

We present exact results pertaining to special types of stratified samples. The essentials of the proofs are included in an appendix. Further details will be included in a longer version of this paper. An algorithm for computing the desired upper confidence limit in the general case is developed. The algorithm is initiated at a new normal approximation for the limit, that appears to yield quick conservative confidence limits that improve with sample size. These normal approximations are not to be confused with the classical ones (Cochran (1977), page 109). We show via examples, as Clopper and Pearson (1934) have done for simple random sampling, that the classical normal confidence limit is generally inappropriate for stratified samples when the proportion being estimated is very small. The algorithm is based on a highly time consuming tree search, which is significantly speeded up by simple pruning procedures. The latter are not guaranteed to yield the exact limit, but have done so in a variety of simple cases tested so far.

In section II we present the algorithm and its modification, and in section III we present selected examples that demonstrate both the need for an exact procedure for computing confidence bounds for small proportions, and the feasibility of the methods we propose. We point out that, strictly speaking, no exact confidence limits exist for population proportions from stratified samples because all the probabilities involved depend not only on the number of defectives (the parameter to be estimated) in the population, but also on their actual distribution among the strata (nuisance parameters) . We refer to a confidence limit that guarantees at least $(1-\alpha)\%$ coverage probability as an exact $(1-\alpha)\%$ confidence bound. We employ similarly loose terminology in the cluster sample case.

In section IV we briefly describe the setting of upper confidence limits for small proportions from cluster samples of equal size. The possible within-cluster homogeneity of element values is well recognized in survey sampling theory. Its positive presence (as measured by the intraclass correlation coefficient) in a population of clusters must be taken into account when inference is made within the context of cluster sampling. When the "worst case" approach is used, the defectives themselves are assumed to be clustered, and not distributed randomly in the population.

Exact results are presented for defective-free samples only. A new normal approximation, and an algorithm for computing an exact upper confidence limit are presented in some detail. The exact procedure, the classical approximate normal limit, and a new normal confidence limit are compared via selected examples in section V.

II. Upper confidence bounds from stratified samples.

Assume that the population and sample consist of H strata of sizes $\mathbf{N}=N_1, \ldots, N_H$ and samples of sizes $\mathbf{n}=n_1, \ldots, n_H$ respectively. A $(1-\alpha)\%$ upper confidence bound for the population proportion p may be based on the unbiased estimate of the number of elements in the population D with the characteristic

$$T = \Sigma_h x_h / f_h \qquad (1)$$

where x_h represents the number of defective elements in a sample of size n_h from stratum h, and $f_h = n_h/N_h$ is its sampling fraction. In the interest of clarity, we shall use the term "defectives" for elements with the characteristic of interest. Here T represents the projected total number of defectives in the population. We denote by $\mathbf{d}=(d_1, \ldots, d_H)$ the array of defective counts actually observed in the stratified sample. The observed value of T in the sample is

$$t=\Sigma_h d_h / f_h. \qquad (2)$$

The standard approximate normal confidence bound for D, the total number of defectives in the population, is given by $d+z_\alpha s$ where the estimated standard deviation s of T is given by the square root of

$$s^2 = \Sigma_h N_h^2 (d_h/n_h)(1 - d_h/n_h)(1 - f_h)/n_h \qquad (3)$$

The exact classical upper confidence bound for D is based on $P_{\mathbf{D}}[T \le t]$, the probability the T does not exceed the observed number of defectives t. This probability depends not only on the total number of defectives in the population D, but on its exact distribution $\mathbf{D}=(D_1, \ldots, D_H)$ among the strata. The largest value this probability can assume , for a given D, is

$$P_D[T<=t] = MAX \{ P_{\mathbf{D}}[T \le t] : \Sigma_h D_h = D \}. \qquad (4)$$

This extremal probability $P_D[T \le t]$ is increasing in D. The classic upper $(1-\alpha)\%$ bound for D is given by

$$D(t)= MIN \{ D : P_D[T \le t] \le \alpha \}. \qquad (5)$$

The use of this max-min procedure guarantees that P [D(t) >= D] is at least $(1-\alpha)$, regardless of the actual distribution $\mathbf{D}$ of defectives in the population. This procedure is a direct extension of the classical interval for proportions from simple random samples (see e.g. Lehmann (1959), 173-180). Here too the procedure represents a "worst-case" approach: Even under the worst of circumstances, where the D defectives are distributed in a manner that would make them least likely to be detected by the stratified sample, the a priori probability is at least $(1-\alpha)\%$

that the interval will actually include the true number of defectives D. Due to this interpretation, the distribution of defectives that achieves the maximum probability in (4) will be referred to as the least favorable distribution of D defectives among the H strata.

Except in two special cases, the least favorable distribution of defectives is hard to find.

Proposition 1: (Defective-free samples in non-proportionate designs)

In non-proportionate stratified designs, when no defectives are found in the sample, the least favorable distribution of defectives to the strata assigns all D defectives to the least sampled stratum, i.e., to the stratum with smallest sampling fraction. The maximal probability is given by:

$$P_D[T=0] = C_{n_1,(N_1-D)} / C_{n_1,N_1} \qquad (6)$$

provided $D <= N_1(1-f_1/f_2)$. Here $C_{l,k}$ denotes the number of combinations of l elements from an unordered set of k elements, and stratum 1 possesses the smallest sampling fraction among the H strata.

Proposition 2. (Defective-free samples in proportionate designs)

In proportionate stratified designs with a common sampling fraction f=n/N, when the sample is defective-free, the least favorable allocation of D defectives to the H strata $\mathbf{D}^*$ satisfies

$$(D-H+1)(N_h/N) <= D^*_h <= (D-H+1)(N_h/N)+1 \qquad (7)$$

For each h=1, ... ,H, i.e., in proportionate designs and defective-free samples, the (non-unique) least favorable allocation is proportional to the strata sizes.

The essentials of the proofs appear in the appendix. These propositions provide a simple way to determine the desired upper confidence interval, by searching for the smallest number of defectives D, denoted by D*, to yield a probability not exceeding α in (6) or (7), when no defectives are observed in the sample. A starting value D_0 for D, is provided in the proposition 3 below. It specifies the normal approximation to D* for the general case, when an array $\mathbf{d}$ of defectives is observed in the sample.

Proposition 3. (Normal approximation in the general case)

In stratified random samples, if stratum 1 has the smallest sampling fraction, and if

A. Sample and strata sizes are large, with $n_h/N_h \dashrightarrow f_h$ for all $1 <= h <= H$

B. The number of defectives in the strata are small, i.e., $D_h << N_h$ for all $1 <= h <= H$

then the normal approximation to the $(1-\delta)$% confidence limit for D is given by

$$D_0 = d + [((z_\alpha/2)(1-f_1)/f_1)^{1/2} + ((z_\alpha/2)^2(1-f_1)/f_1) + t)^{1/2}\}^2]. \qquad (8)$$

i.e., $\lim \{P_{\mathbf{D}}[D <= D_0]: n_h/N_h \to f_h, D_h/N_h \to 0$ and $N_h \to \infty$ for all $1 <= h <= H\} >= 1-\alpha$

The derivation of this approximation relies on the standard normal approximation to the distribution of T, for a fixed number of strata H, when n_h/N_h can be replaced by f_h, and D_h/N_h are considered negligible relative to 1 in the asymptotic variance. The latter is then maximized by replacing all sampling fractions by f_1, and Jensen's inequality is applied to the cumulative distribution function of the standard normal distribution.

Before we turn to the general case of stratified samples in which some defectives are found, we note great simplification in the determination of the desired upper confidence bound in proportionate designs. Unlike the defective-free case, this result is asymptotic.

Proposition 4. (Binomial approximation for proportionate designs)

In proportionate designs with sampling fraction f, under conditions A and B above, the probability $P_{\mathbf{D}}[T<=t]$ converges to the probability that a binomial random variable, with D trials and probability f of success does not exceed d, denoted by P[Bin(D,f)<=d]. In particular, when the sample and strata sizes are large, and the total number of defectives D in the sample is finite, the probability $P_{\mathbf{D}}[T<=t]$ does not depend on the specific distribution of the D defectives in the H strata. The upper $(1-\alpha)$% confidence limit for D is given by

$$MIN[\ D: D>=d \text{ and } P\ [Bin(D,f)<=t] <= \alpha] \text{ with } f=n/N \qquad (9)$$

In several examples of proportionate designs we tried, we found the exact product-hypergeometric probability

$$P_{\mathbf{D}}[T<=t] = \Pi_h \{C_{D_h,d_h} C_{N_h-D_h,n_h-d_h} / C_{N_h,n_h}\} \qquad (10)$$

to be almost independent of the distribution of the D defectives among the strata.

For non-proportionate designs, when some defectives are found in the sample, the binomial approximation is not as useful, and will not be presented here. The problem of determining the desired upper confidence limit D*(t) depends on locating the least favorable allocation of D defectives to the H strata. Note that the least favorable distribution for a fixed D is obtained from (10) as the array $\mathbf{D}$, that sums to D, and maximizes the probability

$$\sum_{x \in V(t)} \Pi_h \{C_{D_h,x_h} C_{N_h-D_h,n_h-x_h} / C_{N_h,x_h}\} \qquad (11)$$

among all such D arrays. All the H-tuplets $\mathbf{x} = (x_1, x_2, \ldots, x_H)$ satisfying $\Sigma_h x_h / f_h <= t$, $\Sigma_h x_h = n$, and $x_h <= n_h$ make up the set permissible $\mathbf{x}$'s denoted here by V(t). We also denote by H* the largest subscript <= H for which some $x_{H^*}>0$, and assume that the strata are ordered in order of increasing sampling fraction. The solution to the problem can now be described in three steps:

i. Generate the set of all $\mathbf{x}$'s satisfying: $\Sigma_h x_h / f_h <= t$, $\Sigma_h x_h = n$, and $x_h <= n_h$

The order in which the set V(t) is generated is immaterial, but in order to speed up the search for D* in step iii, V(t) is generated as a tree structure. Let [y] denote the integer part of y. At the root of the tree is the tuplet $(x_1, x_2, \ldots, x_H)=(x^*,0, \ldots ,0)$ where $x^*=\max\{[t\ f_1]\ ,n_1\}$. The root has x* offspring nodes: $(x^*-1, 1, 0,\ldots, 0)$, $(x^*-2, 2, 0,\ldots, 0)$ etc. down to $(0, [t\ f_2], 0,\ldots, 0)$,. If $[t\ f_2]>n_2$ then the generation of second level nodes will stop with $(x^*-n_2, n_2, 0,\ldots, 0)$. The offspring of a node $(x_1, x_2, 0,\ldots, 0)$ in the second level are similarly generated, except that now the total number of defectives to be divided among H-1 strata 2 through H is $t- x_1/f_1$ and the total remaining sample size is $n-x_1$. Thus the offspring of a node in the second level share the same x_1, their x_2 decreases monotonically from $t- x_1/f_1-1$ down to x_2-n_3 if $[t\ f_3]>n_3$ or down to zero otherwise. Units removed from stratum 2 are added to stratum 3. Thus x_3 increases from 1 to at most n_3. Continuing in the same manner, the tree thus generated has H levels, and up to x* offspring per node. When the tree is traversed inorder, tuplets with high entries x_h for strata with small sampling fractions are encountered first. This property of the tree will be useful in eliminating D values that are too small and lead to probabilities in (11) that are greater than α.

ii. Generate the set of allocations $\mathbf{D}$ that sum to D, and satisfy $D_h >= d_h$ for all h<=H.

The set of all possible allocations that distribute D defectives among the H strata can be generated as a tree structure that eases the process of terminating the search for the least favorable distribution in fruitless directions on one hand, and exposes the D which is too small (yields a probability in (11) that exceeds α) after computing the probability in (11) for as small a number of allocations $\mathbf{D}$ as possible. Such a tree structure is a rootless tree, with up to D_1 (as given in (12) below) siblings at its first level, starting with the extreme allocation $\mathbf{D} = (D_1,\ldots,D_H)$ with $D_1=d+\min(D-d,N-n)$ $D_h=d_h$ for $h>=2$, (12)

and ending with $D_1 = \max(0, D-(N-n)+(N_1-n_1))$, $D_2 = d_2 + (D-d-D_1)$ and $D_h = d_h$ for $h>=2$. The offspring of a level-1 node is generated in the same way, except that they all share the same D_1, and D_2 is decreased within the permissible limits etc. The tree has H-1 levels, and when it is traversed inorder, tuplets with large allocations for strata with small sampling fractions are encountered first.

iii. Search for the largest D for which the maximal probability (11) does not exceed α.

Determine the starting value D_0 using the normal approximation in (8). Starting with that value for D, provided it is permissible given the total size of the finite population N, devise a strategy of first decreasing and then increasing D, or vice versa, until the largest D for which the maximum probability in (11) does not exceed α is found. For each D considered, terminate the search for a least favorable distribution as soon as the probability in (11) exceeds α.

The search step in this procedure can be inordinately time consuming. Following a large number of examples, it was empirically determined, that when selecting the allocation of D' defectives between two strata with sampling fractions $f_1<f_2$, starting from $D'_1=D'$, $D'_2=D'-D'_1$ and ending with $D'_1=0$ and $D'_2=D'$ (neglecting the d's for the time being, which simply bound the D's away from zero), the probabilities either increased to a maximum and then decreased monotonically, or simply decreased monotonically. This statement has not been proved analytically, and may not always be true. It does however lead to an approximate procedure that speeds up the search in step iii considerably:

For a node at level i in the D-tree, compute the probability sum (11) for the offspring subtree inorder. Stop the search among the leaves of the subtree as soon as a local maximum probability is found. Repeat the search for a local maximum for all levels from H-1 up to i+1, replacing leaves by offspring subtrees in the pruning procedure described. A more detailed description of the algorithm will be given in the expanded version of this paper.

This curtailed search is extremely effective in eliminating all D values below the desired $D^*(t)$. The validity of the $D^*(t)$ thus found can be ascertained via a complete search for the largest D for which the probability obtained by this curtailed search does not exceed α.

III. Some examples and comments for stratified sampling.

We complete this section with a few examples that illustrate the performance of the normal approximation and the search algorithm in non-proportionate stratified designs. All examples in Table 1 below concern H=3 strata.

Table 1 displays the upper 95% confidence bounds obtained via the exact method, the new and the standard normal approximation based on the estimated standard deviation in (12). The table includes examples of small populations and small samples (group I, examples 1-4), moderate strata and moderate samples (group II, examples 5-9), moderate strata with small samples (group III, examples 10-14) and large strata with moderate samples (group IV, examples 15-17). The strength and limitations of the three methods are clearly evident in these examples.

A quick perusal through the last two columns reveals the fact, to be expected from the derivation of the new normal approximation, that it requires larger strata and samples than does the standard normal approximation (based on the standard deviation corresponding to the variance in (3)) when a substantial fraction of defectives is present (examples 2 and 14).

Table 1. Examples of exact and normal approximation confidence intervals in the non-proportionate stratified case.

	Strata sizes	Sample sizes (least favorable distribution)	defectives observed	sampling fractions	exact 95 bound (time in min.)	normal bound	stand. normal bound
1.	20,15,10	2,2,2 (11,6,2)	0,0,0	.1,.133,.2	.422 (1.37)	.622	.000
2.	20,15,10	2,2,2 (15,9,3)	1,0,0	.1,.133,.2	.600 (1.72)	.867	.467
3.	20,15,10	2,2,2 (11,11,3)	0,1,0	.1,.133,.2	.556 (1.87)	.800	.347
4.	20,15,10	2,2,2 (11,5,7)	0,0,1	.1,.133,.2	.511 (4.47)	.711	.227
5.	100,50,30	20,15,11 (33,0,0)	3,0,0	.2,.3,.367	.183 (.51)	.172	.149
6.	100,50,30	20,15,11 (24,1,1)	1,1,1	.2,.3,.367	.144 (.48)	.139	.112
7.	100,50,30	20,15,11 (23,3,0)	0,3,0	.2,.3,.367	.144 (.49)	.133	.095
8.	100,50,30	20,15,11 (17,4,2)	0,1,2	.2,.3,.367	.128 (.40)	.122	.082
9.	100,50,30	20,15,11 (19,0,3)	0,0,3	.2,.3,.367	.122 (.41)	.117	.075
10	100,50,30	5,5,5 (45,0,0)	0,0,0	.05,.1,.167	.250 (.55)	.322	.000
11.	100,50,30	5,5,5 (44,0,7)	0,0,2	.05,.1,.167	.283 (3.02)	.394	.122
12.	100,50,30	5,5,5 (61,5,0)	0,2,0	.05,.1,.167	.367 (3.43)	.467	.206
13.	100,50,30	5,5,5 (54,22,0)	1,1,0	.05,.1,.167	.422	.556 (10.48)	.344
14.	100,50,30	5,5,5 (58,28,0)	2,0,0	.05,.1,.167	.478	.639 (18.94)	.417
15.	1000,500 ,300	50,50,50 (146,0,0)	3,0,0	.05,.1,.167	.081	.079 (.77)	.063
16.	1000,500 ,300	50,50,50 (111,13,0)	2,1,0	.05,.1,.167	.069	.072 (6.06)	.054
17.	1000,500 ,300	50,50,500 (85,2,3)	2,1.05,.1,.167		.050	.052 (1.58)	.027

If the proportion being estimated is not small, the new normal approximation will be very conservative. In all cases it appears to yield a conservative limit, whereas the standard approximation tends to yield an underestimate of the exact confidence limit, which in the extreme case of defective-free samples is null.

Examples 1-4 also illustrate the fact that the least favorable distribution of defectives in the strata is not always the extreme allocation given in (12). We note that as a direct consequence of its construction, the new normal approximation will perform poorly whenever the least favorable allocation is very different from the extreme allocation. This will happen when the sample sizes are small and the sampling fractions substantially different among strata, since in that case proposition 4 is invalid. A similar phenomenon occurs in example 14.

The remaining examples in Table 1 portray the behavior of the exact algorithm. First, they show that the new algorithm is usable on desk-top computers. The need for the exact algorithm in problems that require accurate estimates of small proportions, is clearly demonstrated in these examples. The false sense of security (small upper confidence limit) conferred on the user by the usual normal approximation, when the proportions being estimated are small, is amply illustrated by examples 1, 2, 6, 11, 12. The fact that the usual normal approximation represents a considerable underestimate of the 95% upper confidence limit in examples 16 and 17 shows that the mere existence of large strata with large samples is no guarantee that the usual normal approximation will be adequate. In these last two examples the new normal approximation performs remarkably well.

Within each of the four example groups, examples that demonstrate the dependence of the exact upper confidence limit

on the specific pattern of defectives encountered in the stratified sample are presented. The corresponding least favorable distributions of defectives in the three strata are neither the extreme distribution, nor the distribution proportional to the distribution of defectives in the sample.

The loss of efficiency due to the use of non-proportionate stratified designs with widely diverging sampling fractions versus simple random sampling of equal total size, or proportionate sampling of equal total size, is apparent. It is easy to quantify in the extreme case when no defectives are found in the population. In that case the sample is also defective-free. The exact classical upper confidence interval would still replace the actual strata fractions by the minimum sampling fraction, thereby increasing the interval length unnecessarily.

A simple example will convey the point more cogently. Consider a population consisting of H=2 strata, with strata sizes 1000 and 2000 respectively. If a 20% proportional sample is taken and no defectives are found in the sample, the exact upper 95% confidence bound for the proportion of defectives is 14/3000=.0047. This is also the confidence limit from the two strata separately, and from a simple 20% random sample (SRS) in which no defectives are found. Thus no loss of efficiency is entailed by proportionate samples when the population proportion is null. If in the same population a non-proportionate sample of 100 and 500 is taken and no defectives are found, the least favorable allocation is the extreme allocation that assigns all defectives to stratum 1 (minimum sampling fraction=.1). The exact 95% confidence bound for the population proportion is now 28/3000=.0093, representing a 50% loss in accuracy over simple random sampling for defective free populations.

When the population proportion is not zero, and defectives are found in the sample, the theoretical loss of efficiency is difficult to compute. It is however possible to estimate this loss in specific examples. E.g., in example 14 of Table 1, the estimated population proportion of defectives is 40/180. We can therefore compare the exact 95% confidence limit of .444 with the one obtained from an SRS of size 15 in which approximately the same proportion of defectives is found. This confidence bound is obtained simply as D*/180, where D* is the smallest integer D for which the probability that a Hypergeometric variable counting the number of defectives in a sample of n=15 from a population of size N=180 with D defectives, does not exceed d, is not greater than .05.

The approximate number of defectives d for the comparable SRS is 3.3333 . If we take d=3 defectives in the sample, we obtain an exact upper 95% confidence bound of .433. If instead we take d=4 we obtain .506 for the bound. Although the situation is somewhat ambiguous, the loss of efficiency due to the non-proportionality of the design is not dramatic. Similar comparisons in example 5 show a bound of .044 for SRS to the exact .183 of the stratified sample. In example 16 the SRS bound is .0589 to the exact bound of .069.

These examples demonstrate that when the estimated proportion is small but some defectives are found in the sample, the upper confidence intervals obtained from non-proportionate samples are in general less efficient than confidence bounds derived from comparable simple random samples. The loss of efficiency depends however on the number of strata involved and the size of the disparity in sampling fractions in the different strata.

The immediate practical recommendation that results from these considerations is to avoid, if no other overriding considerations exist, the use of non-proportionate designs when planning surveys that attempt to estimate very small proportions.

IV. Upper confidence bounds for cluster samples of equal size.

The problem of setting confidence limits for small proportions from cluster samples of equal size can be treated in a manner similar to that used for stratified samples, when the sample of clusters is poststratified by the number of defectives in the sample.

Suppose that the population is partitioned into N clusters of equal size, say B, and a simple random sample of n complete clusters is taken. We denote by N_i (n_i) the number of clusters in the population (sample) containing exactly i (i=0,1 ... ,B) defective elements. We assume that the number of clusters $N=\Sigma_i N_i$ (but not the individual N_i's) in the population is known, and we wish to set an upper confidence bound for the proportion of defectives in the population, or equivalently, for the total number of defective elements in the population

$$D = \sum_{i=1}^{B} i N_i, \qquad (13)$$

for which $T^* = N \sum_{i=1}^{B} i x_i / n$

provides a natural estimate. Here x_i denotes the random number of clusters in the sample with exactly i defectives. Since the factor n/N = f is a constant, we refer instead to the statistic

$$T = \sum_{i=1}^{B} i x_i \qquad (14)$$

in the sequel. In order to set an upper confidence limit on the number of defectives in the population, the smallest number of defectives D is sought, for which the probability of the event T<=t with

$$t = \sum_{i=1}^{B} i n_i \qquad (15)$$

does not to exceed α. As in the case of stratified sampling, the probability of this event depends not only on D, but on the actual allocation of the D defective elements to clusters containing varying proportions of defectives. If one does not wish to postulate a priori any maximum number of defective elements per cluster, the possibility of "all defective" clusters cannot be ruled out. In order to obtain a confidence limit with at least ($1-\alpha$) coverage probability under all possible distributions of defectives among the population of clusters, we define the set of all admissible distributions of D defectives in the population as

$$S(D) = \{\mathbf{N}; \Sigma_i N_i = N, N_i >= n_i \ \Sigma_i i N_i = D\} \qquad (16)$$

For a given allocation $\mathbf{N}$ in S(D) the probability that T<=t is given by the multiple Hypergeometric distribution

$$P_{\mathbf{N}}(t) = P[T<=t] = \sum_{\mathbf{x}\in V(t)} \prod_{i=0} C_{x_i,N_i} / C_{n,N} \qquad (17)$$

where $V(t) = \{\mathbf{x}; \Sigma_i i x_i <= t, \Sigma_i x_i = n \text{ and } 0<=x_i<=N_i\}$ (18)

and under the "worst case", it achieves its maximum

$$P_D(t) = \mathrm{Max}\{P_{\mathbf{N}}(t); \mathbf{N}\in S(D)\}. \qquad (19)$$

The classical ($1-\alpha$) 100% confidence limit is then given by the smallest D for which (19) does not exceed α or

$$D(t) = \mathrm{Min}\{D; P_D(t) \le \alpha\}. \qquad (20)$$

We refer to the allocation $\mathbf{D}^*$ which achieves the maximum in (19) as the "least favorable" distribution as before.

PROPOSITION 4. In simple random sampling of complete clusters, when no defective elements are found in the sample, the least favorable allocation is given by

$N_B^* = [D/B]$. If $d = D - B^* N_B^* > 0$ then $N_d^* = 1$, $N_0^* = N - N_B^* - 1$
else $N_0^* = N - N_B^*$;
All remaining N_i^*'s vanish. (21)

The proof is immediate. In order to determine a good initial value for D in the search for the confidence limit when some defectives are found in the sample (t>0), and help locate its "least favorable" allocation, the normal approximation can be of some help.

The normal approximation.

The asymptotic Normal approximation to the distribution of the total number of defectives T in the sample when n,N-->∞ but n/N-->f and the cluster size B stays fixed, has mean fD with D given by (13), and variance given in our notation by

$$f(1-f)\,\Sigma\, N_i\,(i-\Sigma_i i N_i/N)^2 \qquad (22)$$
$$0 \le i \le B$$

(Cochran 1977, page 246). In the particular circumstances we envisage the proportion of defectives in the population $D/(NB)=\Sigma_{i=1} iN_i/(NB)$ is very small as N increases indefinitely. This requires that all N_i for $i>=1$ remain finite and only N_0 increase with N in such a way that $\Sigma_i iN_i/N-->0$ and $N_0/N-->1$. We formulate these requirements as follows:

Proposition 5. (Normal approximation for cluster samples)
For simple random samples of equal size clusters, if
A. Sample and population sizes are large, with $n/N -> f$.
B. The number of defectives in the population are small, i.e., $D=\Sigma_{i=1} iN_i$ remains finite as $N->\infty$ $\Sigma_i iN_i/N->0$ and $N_0/N->1$. then the normal approximation to the $(1-\alpha)$% confidence limit for D is given by

$$D^*(t)\{((1-f)/f)*((z_\alpha/2)*B1/2+((z_\alpha/2)2+t)1/2)2\} \qquad (23)$$

i.e., $\lim \{P_N[D<= D^*(t)]: n/N -> f,\ \Sigma_i iN_i/N->0$ and $N_0/N->1$ and $N->\infty\} >= 1-\alpha$

Under the our assumptions, the asymptotic variance in (22) reduces to $\Sigma_{i=1} i^2 * N_i * f * (1-f)$. Since $t<<D$, for the probability of the event $[T<=t]$ not to exceed α, regardless of the actual values of the N_i's, the maximum of the normal approximation to this probability, subject to $N_i>=n_i$ and $\Sigma_i iN_i=D$, is needed. It is achieved when the maximum possible number of defectives is placed in "all defective clusters", i.e., N_B is maximized. Thus the extreme allocation

$$N_B^* = [D/B];\ \text{if } d=D-B*N_B^*>0 \text{ then } N_d^*=n_d+1 \qquad (24)$$
$$\text{otherwise } N_d^*=N-N_B^*;$$

for all remaining $i>=1$, $N_i^*=n_i$ and $N_0^*=n_0+N-n-\Sigma_{i=1} N_i^*$, yields the desired maximum. Upon inserting this allocation into (22), the resulting Normal approximation for the upper confidence limit is obtained.

The search for the exact upper confidence bound proceeds in a manner similar to that of the stratified search described in section 2. The steps are:

1) Determine the number of defectives D in the population from the improved Normal approximation given in (23).

2) Generate the set V'(t) of all admissible x-tuplets that satisfy the first two conditions in (18) but not the third, which depends on **N**, and store it in memory. The third condition in (18) is checked with each use of this set in computing the cumulative Multiple Hypergeometric probabilities in (17). The cardinality of V'(t) increases rapidly with t, thus rendering the approach feasible only for a small number of defectives in the cluster sample-the case for which it was originally intended.

3) Use the extreme allocations for any trial value D, given by (24), to adjust the choice of D made in step 1.

Since this extreme distribution of D is not the least favorable distribution of D, except for t=0, the exact upper confidence bound for the total number of defectives in the population is actually larger than the one determined in step 3. Starting with the D determined in step 3, a complete search for the least favorable distribution for each D is carried out for successively larger values of D, until the first D to yield a probability of $[T<=t]$ that does not exceed α is found. A faster search results from the (experimental) realization that the trial value obtained in step 1 is always an upper bound, and the trial value obtained in step 3 is always a lower bound on the desired confidence limit. Therefore a binary search between these two values speeds up the determination of D*.

For each D the complete search is described in steps 4 and 5.

4) Generate the set S'(D) of all allocations **N** that satisfy $\Sigma_i N_i=N$ and $\Sigma_i iN_i=D$, in order of decreasing N_1.

5) Start computing the cumulative Multiple Hypergeometric probability (17) for successive allocations in S'(D). If the computed probability for an allocation exceeds α, increase D by 1 and go to step 4.

If the probabilities computed for all allocations in S'(D) do not exceed α stop the search. The desired upper $(1-\alpha)$ confidence bound for the proportion of defectives in the population is then D/(BN).

The search involved in the last two steps of the algorithm can be drastically reduced when t is small using considerations similar to those used in the stratified case. These will be explained fully in the expanded version of this paper.

Note that no simple case, analogous to that of proportional allocation for stratified sampling, exists for cluster sampling. One could say that the natural (post) stratification by number of defectives per cluster, stratifies the population of clusters into strata that are inherently unequal in the proportion of defectives they contain. For that reason there is a loss of efficiency due to clustering which is roughly proportional to cluster size in defective-free populations, and which is unavoidable.

Table 2. Examples of exact and normal approximation confidence limits in clustered samples of clusters of size B=10. N=500. n=50. $1-\alpha=.95$.

Sample array	Least favor. Distribution	Cv	B*	Exact limit	Old limit	New limit	time (min)
1 50,0,0,0,0 0,0,0,0,0,0	468,5,0,0,0 0,0,0,0,0,27	1	0	.0550	.0000	.0512	0.29
2 49,1,0,0,0 0,0,0,0,0,0	467,6,0,0,0 0,0,0,0,0,27	2	1	.0552	.0051	.0550	0.69
3 49,0,1,0,0 0,0,0,0,0,0	468,1,4,0,0 0,0,0,0,0,27	4	2	.0558	.0101	.0586	3.83
4 48,2,0,0,0, 0,0,0,0,0,0	468,3,2,0,0 0,0,0,0,0,27	4	2	.0554	.0082	.0586	1.69
5 48,0,2,0,0 0,0,0,0,0,0	468,0,2,1,2 0,0,0,0,0,27	12	4	.0570	.0165	.0654	3.41
6 48,0,0,0,2, 0,0,0,0,0,0	464,0,0,0,2 0,0,1,9,0,24	67	8	.0654	.0329	.0786	NA
7 45,5,0,0,0, 0,0,0,0,0,0	464,5,0,0,4 0,0,0,0,0,27	19	5	.0582	.0163	.0688	6.68
8 45,4,1,0,0, 0,0,0,0,0,0	463,4,1,0,2 3,0,0,0,0,27	30	6	.0601	.0200	.0722	17.04
9 10,0,0,0,0 0,0,0,0,0,0	361,0,0,0,0 0,0,0,0,1,138	1	0	.2578	.0000	.2128	0.33
10 9,1,0,0,0 0,0,0,0,0,0	369,2,0,0,0 0,0,0,0,0,129	2	1	.2584	.0050	.2282	0.54
11 9,0,1,0,0 0,0,0,0,0,0	361,1,11,0,0 0,0,0,0,0,127	4	2	.2586	.0101	.2430	0.98
12 8,2,0,0,0, 0,0,0,0,0,0	361,3,9,0,0 0,0,0,0,0,127	4	2	.2582	.0082	.2430	1.23
13 8,0,2,0,0 0,0,0,0,0,0	468,0,2,1,2 0,0,0,0,0,27	12	4	.0570	.0165	.0654	3.41
14 8,0,0,0,2, 0,0,0,0,0,0	464,0,0,0,2 0,0,1,9,0,24	67	8	.0654	.0329	.0786	NA
15 5,5,0,0,0, 0,0,0,0,0,0	464,5,0,0,4 0,0,0,0,0,27	19	5	.0582	.0163	.0688	6.68
16 5,4,1,0,0, 0,0,0,0,0,0	463,4,1,0,2 3,0,0,0,0,27	30	6	.0601	.0200	.0722	17.04

V. Some examples and comments for cluster sampling.

In this final section we present some examples from a hypothetical population of N=500 clusters of size b=10, from which a simple random sample of n=50 clusters is taken. In table 2, the 95% 'exact' confidence limit, with its attendant least favorable distribution, t value (see (15)), number of x-tuplets which measures indirectly the complexity of the problem, and computation time, are presented. The standard normal approximate limit and the new approximate normal confidence limit are also presented for comparison.

The standard approximation for the distribution of T is the Normal distribution with mean fD and standard deviation given by

the square root of (22) and estimated from the sample. In these examples f=.1, B=10, and the N_i's are simply estimated to be n_i/f. The upper 95% confidence bound for the number of defectives is computed from the formula $D=(z_\alpha s+t)/f$, with $z_\alpha=1.645$ and t given by (15).

In Table 2 above, The sample **n** is given by an array representing $(n_0,n_1, \ldots, n_{10})$, whose entries add up to n=50. The least favorable distribution is similarly presented as an array whose entries add up to N=500. The columns Cv and B* represent the number of x-tuplets in V'(t), and the highest index not exceeding B which has a non-zero entry for some element of V'(t) respectively. The succeeding three columns provide the 'exact', old and new normal approximation confidence limits. The last column provides the time in minutes (when available) to compute the 'exact' limit.

Several important points are brought out by this table:

1. The inadequacy of the standard normal approximation (old limit column) is evident in every instance.
2. The new normal approximation (new limit column) generally overestimates the upper 95% confidence limit in this setting. It does however provide a useful starting value for the search algorithm. The latter can be reasonably used on a desktop computer, judging from the execution time column.
3. The least favorable distribution is not the extreme distribution when some defectives are found in the sample. No simple rule appears to explain the distributions we observe in the examples.
4. The performance of the new normal approximation is again surprisingly good. In those cases where the algorithm is very slow one may be satisfied with the upper confidence limit given by this normal approximation. It is a quick and dirty tool that can be readily used on a common calculator.

REFERENCES

Clopper, C.J., and Pearson, E.S. (1934), "The Use of Confidence or Fiducial Limits illustrated in the Case of the Binomial", Biometrika, 26, 404-.

Cochran, W.G. (1977) Sampling Techniques. Third Edition. Wiley, New York.

Lehmann, E.L. (1959), Testing Statistical Hypotheses, Wiley, New York.

Sedransk, J. and Meyer, J. (1978), "Confidence Intervals for the Quantiles of a Finite Population: Stratified Random Sampling", Journal of the Royal Statistical Society, Ser B, 40, 239-252.

Smith, P.J., and Sedransk, J. (1983), "Lower Bounds for Confidence Coefficients for Confidence Intervals for Finite Population Quantiles", Communications in Statistics - Theory and Methods, 12(12), 1329-1344.

APPENDIX

Proof of proposition 1.

When no defectives are found in the sample the probability in (2.6) reduces to

$$P_{\underline{D}}[T=0] = \prod_{h=1}^{H} C_{(N_h-D_h),n_h}/C_{N_h,n_h} \qquad (A.1)$$

We assume for the sake of simplicity that the strata are ordered by increasing sampling fractions, i.e. $f_1 <= f_2 <= \ldots <= f_H$. It seems intuitively clear that the maximum of (A.1) will be achieved when all D defectives are placed in stratum 1. We shall in fact show that for all $0 <= D <= N_1[1-f_1/f_2] + 1$,

$$P_{\underline{D}}[T=0] <= C_{(N_1-D),n_1}/C_{N_1,n_1} = P[D]$$

It suffices to prove the assertion for H=2 strata, since the distribution of D defective among H strata can be accomplished by successive distribution of D' defectives between two strata. For H=2 we shall show that

$$C_{(N_1-D_1),n_1} C_{(N_2-D_2),n_2} > = C_{(N_1-(D_1-1)),n_1} C_{(N_2-(D_2+1)),n_2} \qquad (A.2)$$

for $D_1+D_2=D$ and $D_1>=1$. It is immediate that (A.2) holds if, and only if, $n_2/(N_2-D_2) >= n_1/(N_1-D_1)$. Since the l.h.s. of the last inequality is bounded below by f_2, and the r.h.s. is bounded above by $n_1/(N_1-D+1)$, (A.2) will certainly hold if $n_1/(N_1-D+1) <= f_2$. The assertion follows.

Proof of proposition 2.

In the two strata case, H=2 and the the r.h.s. of (A.1) is a function of D_1 alone when the total number of defectives D is fixed; we denote it by $P[D_1]$. Thus the ratio

$$P[D_1+1]/P[D_1] = [(N_1-D_1-n_1)/(N_1-D_1)][(N_2-D_2+1)/(N_2-D_2-n_2+1)]$$

for fixed D, and since $f_1=f_2$ implies $n_1N_2=n_2N_1$ we have the equivalence $P[D_1+1]/P[D_1] > 1$ if, and only if $D_1/N_1 < (D_2-1)/N_2$ implying that the maximum over $0<=D_1<=D$ is achieved at $D_1=[(N_1/N)(D-1)]+1$. The maximum may not be unique when $(N_1/N)(D-1)$ is integer. We have shown however that proposition 2 holds for H=2, and obtained a more specific solution for the two strata case.

For H>2, for fixed D, if 1 is added to D_i, then 1 must be subtracted from the number of defectives D_j placed in another stratum j not equal to i. Starting with a distribution $\underline{D}$ of the D defectives, the new distribution with D_i+1 and D_j-1 will be denoted by $\underline{D}_{i,j}$. We then have

$$P[\underline{D}_{i,j}]/P[\underline{D}] > 1 \text{ if, and only if } D_i/N_i < (D_j-1)/N_j$$

Thus a distribution of defectives D is least favorable if for each pair (i,j) $D_i/N_i >= (D_j-1)/N_j$

Using the weights $N_j/(N-N_i)$ for all j not equal to i and taking a convex combination of both sides we obtain

$$D_i/N_i >= \sum_{j=i}^{H} [N_j/(N-N_i)]\,(D_j-1)/N_j$$

or $D_i >= (N_i/N)(D-H+1)$. Taking now a convex combination over i not equal to j with weights $N_i/(N-N_j)$ we obtain $D_j <= (N_j/N)(D-1)+1$. This completes the proof of proposition 2.