SAMPLING DESIGN ISSUES WHEN DEALING WITH ZEROS

Benjamin King, Florida Atlantic University
Benjamin King, FAU, 777 Glades Road, Boca Raton, FL 33431

Key Words: Zeros, Optimal Stratification

Introduction and Background. This paper concerns the problem of sampling from a population of units, not all of which contribute to the calculation of the **estimated population total** that is desired. The problem arises in several forms that may seem different at first glance, but which all boil down to the same situation: For example, if one is interested in estimating parameters of a subpopulation (domain) and units that are in the domain cannot be identified in advance, the units that do not belong to the domain must be screened out after the sample has been selected and only the units that remain after screening are used in estimating domain characteristics.. Alternatively, the selected units may carry numerical values such as dollars, counts, weights, etc., but some of these have zero value and thus do not contribute to the total. Another form of the problem occurs when the original sample consists of geographic sites-- e.g., business establishments, military targets, whatever-- and when the selected sites are visited many are found to be empty.

The units that have zero value, or are non-existent, or which do not belong to the subpopulation of interest, will be called **"zeros"**. The motivation for this study was a series of surveys for the purpose of estimating damages in lawsuits involving health care insurance claims. For each of the large number of claims in the target population a dollar amount was paid for medical care, but a certain proportion of such payments were made in violation of the law. The plaintiffs sought to recover the improper payments as damages, and the statistical task was to estimate the total amount of damages owed by the defendants. The claim identifiers and the amounts paid were contained in a database to which random selection could be applied, but whether or not the payment was legal could only be determined by a detailed audit of the paper files associated with the claims, and thus estimation based on a relatively small sample was required. The zeros in this situation were the claims that were paid properly and which contributed nothing to the estimated total damages. The proportion of nonzeros in the population was not known and could only be guessed.

Relevant Formulas. Define the following:

N = *size of population*

M = *size of subpopulation of nonzeros*

$p = \frac{M}{N}$ = *proportion of nonzeros*

n = *sample size*

m = *number of nonzeros in sample*

Y = *population total damages*

y = *sample total damages*

$\bar{Y}^{*} = \frac{Y}{N}$ = *population mean*

$\bar{Y} = \frac{Y}{M}$ = *mean of nonzeros*

$$S^{*2} = \frac{1}{N-1}\left(\sum y_i^2 - N\bar{Y}^{*2}\right)$$

$$S^2 = \frac{1}{M-1}\left(\sum y_i^2 - M\bar{Y}^2\right)$$

In estimating the population total, if the size of the subpopulation of nonzeros were known, the conventional "blowup" formula, $\left(\frac{M}{m}\right)y$, would apply (Kish, 1965; Cochran, 1977). In the interesting case, however, where M is unknown, the estimator of Y is

$$\hat{Y} = \left(\frac{N}{n}\right)y \quad . \qquad (1)$$

As shown in the textbooks, the variance of the estimator is

$$V(\hat{Y}) = \frac{N^2 S^{*2}}{n}\left(1 - \frac{n}{N}\right), \qquad (2)$$

and S^{*2} is estimated by s^{*2}, its sample analogue, in which zeros in the sample are given a zero value and n is used as the denominator in the sample mean, not m. Cochran (1977) observes that "...some students seem to have a psychological objection to doing this, but the method is sound."

Cochran (p.38) shows further that if terms in $\frac{1}{N}$ and $\frac{1}{M}$ are ignored,

$$S^{*2} \doteq pS^2 + p(1-p)\bar{Y}^2 \quad . \qquad (3)$$

He leaves verification of the above equation for the reader.[1] Kish (1965, p. 136 and later on p.435) proves the corresponding relationship for the sample estimator of S^{*2}, but his discussion invites some confusion about whether or not the unit variance with zeros is greater than the variance for nonzeros only. It can be shown that the variance of the population including zeros may be smaller or larger than the variance of the subpopulation of nonzeros. It is smaller if $S^2 > p\bar{Y}^2$, or $C > \sqrt{p}$, where C is the coefficient of variation of the nonzeros. If the addition of zeros causes the variance to decrease, however, it also affects the mean, and as Jessen and Houseman (1944) show,

$$C^2 + 1 = p(C_0^2 + 1), \qquad (4)$$

with C_0 defined as the coefficient of variation of the population containing the zeros and nonzeros combined. Thus the c.v. of the population with zeros is always inflated.

Jessen and Houseman (1944) appear to be the earliest to discuss an obvious question. They are estimating farm characteristics for the State of Florida where the selection unit within strata is a small area called a "grid". Selected grids may contain zero farms. The question of interest is, "Under what conditions is it efficient to remove the zero grids from the frame before sampling?" As shown in Cochran's textbook, their analysis is based on the relationship

$$\frac{V(M \text{ known})}{V(M \text{ not known})} = \frac{C^2}{C^2 + (1-p)}. \qquad (5)$$

A Different Question. This paper deals with a problem that does not seem to have been discussed explicitly in the literature but which, we believe, is quite common. In the present case of sampling medical claims it is not economically feasible to remove the zeros from the frame *a priori*. It is desirable to stratify the frame in order to reduce the variance of the sample estimator of total damages, but the variances of the strata containing zeros are not known-- we only know the variances of the original claim payment dollar values. [Assume that if a claim was paid illegally the damage amount is the full payment.] It may be possible to guess the proportion, p, for a stratum and use equation (3), but what values should be used for $\bar{Y}$, and S^2, the mean and variance of the nonzeros? Under what conditions, for example, it is reasonable to assume that the parameters for the nonzero subpopulation are the same as those for the population of original claim amounts in the database frame?

The Zero-Generating Process. It is useful to think of the mechanism that results in the presence of zeros as follows: The original population has payment dollars attached to every claim. Then by some random process (from the point of view of the sampler) each claim payment, Y_i, is multiplied by a variable, X_i, that is either equal to one or to zero. Thus the damage amount is the product Y_iX_i.

Assume that X_i is a Bernoulli variable, with probability p of being equal to one, and distributed independently of Y_i. The variance of the product of two independent random variables is

$$\begin{aligned}\sigma^2(YX) &= \sigma^2(Y)\sigma^2(X) + E^2(Y)\sigma^2(X) + E^2(X)\sigma^2(Y) \\ &= \sigma^2(Y)p(1-p) + E^2(Y)p(1-p) + p^2\sigma^2(Y) \\ &= p\sigma^2(Y) + p(1-p)E^2(Y). \qquad (6)\end{aligned}$$

Observe that the above result is of the same form as Equation (3) above, except that $E(Y)$ and $\sigma^2(Y)$ are the mean and variance of the original random variable before "conversion" to zeros. This leads to the following rule:

> **If the conversion of the original payment amounts to zero is random and independent of the payment amounts, then the mean and variance of the nonzeros are the same as those of the original payments, and the variance with zeros can be estimated via Equation (3) using an assumed value of p.**

Under the assumption that the zeros are generated randomly and independently of the original amounts, we can divide the expression in (3) above by S^2 to obtain the ratio of the variance with zeros to the prior variance (i.e. before the zeros are known). The result is

$$\frac{\text{Variance with Zeros}}{\text{Prior Variance}} = p + \frac{p(1-p)}{C^2} \qquad (7)$$

which is plotted in Figure 1 as a function of p and C, the prior coefficient of variation .

[1] A straightforward proof involves the relationship $\sigma^2_y = E_x[\sigma^2(y|x)] + \sigma^2_x[E(y|x)]$, where x indicates whether y is 0 or >0. I call this "The Universal Theorem of Survey Sampling" because it appears in so many important theoretical proofs.

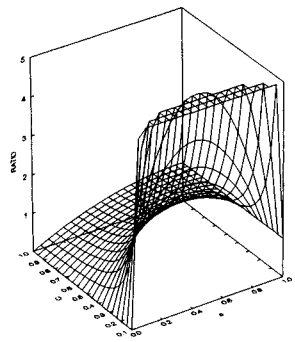

Figure 1. Ratio of Variance with Zeros to Prior Variance

As mentioned above, the ratio is less than one when C exceeds $\sqrt{p}$. As p, the proportion of nonzeros, increases from zero to one the ratio of variances is higher for all values of C.

The case of Y and X correlated. If the variables Y and X are not independent, a Taylor Series approximation for the variance of their product is

$$\sigma^2(YX) = E^2(Y)\sigma^2(X) + E^2(X)\sigma^2(Y) + 2E(Y)E(X)\sigma(X,Y)$$
$$= p\sigma^2(Y) + p(1-p)E^2(Y) - p(1-p)\sigma^2(Y)$$
$$+ 2pE(Y)\rho\sigma(Y)\sqrt{p(1-p)} \qquad (8)$$

In the example of damages for illegal medical payments, if the correlation coefficient ρ is positive, that implies that claim amounts with high dollar values would tend to be paid illegally, and the lower payments tend to be converted to zeros.

Figures 2a through 2d show the effect of correlation in the zero-generating process. In each plot the curve in the center corresponds to $\rho = 0$-- in other words the intersection of a plane fixed at the given value of p and the surface in Figure 1.[2] The curve for $\rho = 0$ is bracketed by curves for $\rho = -0.5$ (below) and $\rho = +0.5$ (above). It can be seen that although correlation may have an appreciable effect for values of p above 0.5, for $p = 0.1$ and $p = 0.2$ the three curves are practically the same. In the example of the damages for illegal medical payments, as well as in other cases with high rates of zeros (low values of p) it is therefore safe to assume that the zeros are randomly and independently distributed and that the variance for the nonzeros is the same as the prior variance.

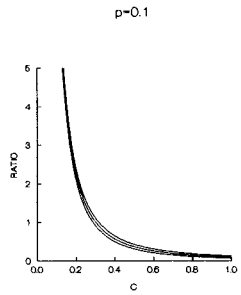

Figure2a. Effect of Correlation When p=0.1

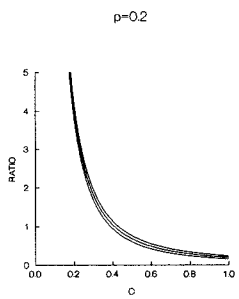

Figure 2b. Effect of Correlation When p=0.2

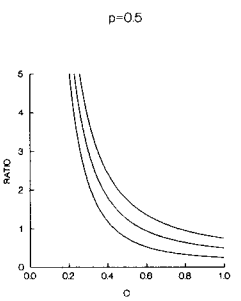

Figure 2c. Effect of Correlation When p=0.5

[2]Note that if $\rho=0$, Equation (8) does not yield the same result as Equation (6), indicating that the Taylor Series approximation is what its name implies--an approximation. The result deviates from the correct value by $p(1-p)\sigma^2(Y)$, which is negligible if p is not close to 0.5.

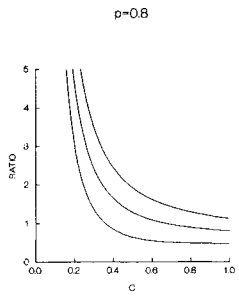

Figure 2d. Effect of Correlation When p=0.8

The Effect of the Zeros on Optimal Stratification. When the target population is stratified, both C and p will vary from stratum to stratum, with the latter difficult to estimate in advance. Table 1 (see below) provides an idea of the consequences using planning data for one of the medical payment surveys:

In this survey the strata were defined by dollar values within four major categories of claim type. This explains the four ranges from low to high of the prior means. The second to the last column shows the Neyman allocation using the variances of the original claim amounts (before conversion to zero). The last column is the Neyman allocation using the estimated p and the relationship shown in expression (3) above. The latter optimal allocation yields a standard error of the estimated total damages equal to $1,934,819, whereas the sample sizes based on the variances without zeros result in a standard error of $2,160,492. Thus in this example the efficiency of the suboptimal design (assuming the estimates of p are good) is 80 percent. Since the design was never actually executed, the prior estimate of the efficiency is all that we have.

Zeros and the *Cum* $\sqrt{f(x)}$ Rule for Determining Stratum Boundaries. A commonly used approach to constructing strata, due to Dalenius and Hodges (1959), is to form the cumulative distribution of the square root of the frequency function, $f(x)$, where x is a variable that is highly correlated with the target variable of interest, and choose stratum boundaries, x_h, that create equal intervals on the cumulative $\sqrt{f(x)}$ scale. Cochran (1997, p. 129) calls this method the " ***Cum*** $\sqrt{f(x)}$ Rule", and it is generally believed to yield strata that are a good approximation to the optimal choice for Neyman allocation. Assuming that readers are familiar with the details of the approach, we shall examine some of the implications for the present problem. In the following, we consider applying the method to the frame distribution of y, the original claim amounts before zeros are known:[3]

1. In using the ***Cum*** $\sqrt{f(x)}$ Rule, the approximation to optimal stratum construction rests on the assumption that the stratum boundaries, y_h, mark off strata that are (approximately) uniform, i.e., with constant frequency function, $f(y)$. Under this assumption, $W_h \doteq f_h(y_h - y_{h-1})$, and $S_h \doteq \frac{1}{\sqrt{12}}(y_h - y_{h-1})$. It follows that $W_h S_h \doteq \frac{1}{\sqrt{12}} f_h (y_h - y_{h-1})^2$, as shown in Cochran (1977, p.129).

2. The typical distribution of dollar amounts is either-shaped like a backward letter J or in the form of an elongated right triangle— i.e., with many low-valued amounts and few of extremely high value. Thus as one moves toward the right tail, the approximately constant stratum frequency values, f_h, **decrease**. Another feature of the ***Cum*** $\sqrt{f(x)}$ Rule is that it leads to approximately constant $W_h S_h$. Hence if f_h decreases, $(y_h - y_{h-1})$, the interval covered by Stratum h, must increase, causing S_h, the stratum standard deviation, to increase. Most readers who have applied the ***Cum*** $\sqrt{f(x)}$ Rule to similar distributions have doubtless observed this — i.e., increasing S_h as one moves further toward the extreme right tail.

3. Figure 3, below depicts two adjacent strata in the right tail of the distribution.

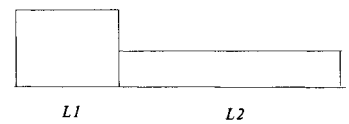

Figure 3. Two adjacent strata in right tail.

For simplicity, we call the length of the interval for the left stratum L_1 and that for the right-hand stratum L_2. Because the strata are rectangular (approximately) we can describe the change in S_h as we move toward the extreme tail as

[3]For convenience we omit subscripts when discussing individual claim amounts, assuming that there will be no confusion with the previous use of the symbol y for the sample total.

$$\Delta S = \frac{1}{\sqrt{12}}(L_2 - L_1) \qquad (9)$$

The change in $\bar{Y}$, the stratum mean, is

$$\Delta \bar{Y} = \frac{1}{2}(L_2 + L_1) \qquad (10)$$

Finally, it can easily be shown that the change in C, the coefficient of variation, is

$$\Delta C = \frac{\Delta S - C\Delta \bar{Y}}{\bar{Y} + \Delta \bar{Y}} \qquad (11)$$

Examination of Figure 3, as well as expressions (9) and (10) shows that ΔY is always greater than ΔS, with both changes positive. Consider the ratio

$$\frac{\Delta S}{\Delta \bar{Y}} = \frac{(L_2 - L_1)}{\sqrt{3}(L_2 + L_1)} \qquad (12)$$

If C, the coefficient of variation, is greater than the ratio in (12) then ΔC is negative. Another way to state this relation is that C will decrease as we move from stratum to stratum toward the right tail if $\bar{Y} < \frac{L_1(L_2 + L_1)}{2(L_2 - L_1)}$.

Since L_2 exceeds L_1 the criterion will always be positive, but we cannot prove that $\bar{Y}$ will always be less than it, and thus that C is steadily decreasing. Based on experience with the *Cum* $\sqrt{f(x)}$ Rule, however, it is our conjecture that increases are unusual and probably only occur in moving to the highest valued stratum.

4. What does this discussion have to do with the problem of zeros? The answer is that the stratum-to-stratum behavior of C bears on the issue of what happens to the relative values of the S_h for the various strata in the presence of zeros. Suppose that we have applied the *Cum* $\sqrt{f(x)}$ Rule to the distribution prior to knowing the zeros and have used Neyman allocation. To what degree is the allocation likely to become suboptimal when the zeros are taken into account? Figures 1 and 2a through 2d show the effect on the variance of changes in C for a given value of p. It is reasonable to assume in cases like the present example that p is small, that it is fairly constant from stratum to stratum, and that C stays mostly within the range in Figures 1 and 2 where the effect on S is rather flat. If in fact the effect of the zeros on the S_h is more or less constant, then the optimality of the sample allocation based on the prior variances will not be seriously affected. If, on the other hand, a decreasing C drops much below $\sqrt{(p)}$, then there may be a shift in optimal allocation toward the strata with higher dollar claim values. This would seem to be a further argument for sampling as much of the right tail with certainty as possible and thus removing that portion of the distribution from contention.

REFERENCES

Cochran, W.G. (1977). *Sampling Techniques*. 3rd Edition. John Wiley & Sons, New York.

Dalenius, T., and Hodges, J.L., Jr. (1959). Minimum variance stratification. *Journal of the American Statistical Association*, **54**, 88-101.

Jessen, R.J. and Houseman, E.E.(1944). Statistical investigations of farm sample surveys taken in Iowa, Florida and California. *Iowa Agricultural Experiment Station Research Bulletin*, **329**.

Kish, L. (1965).*Survey Sampling*. John Wiley & Sons, New York .

Table 1. A Priori Sample Allocation, with and without Zeros

N	Mean Without Zeros	Std Dev Without Zeros	C	p (Est.)	StdDev With Zeros	Neyman n Without Zeros	Neyman n With Zeros
53,391	8.86	3.82	0.43	0.15	3.49	4	7
43,706	24.17	5.39	0.22	0.15	8.88	5	14
47,002	57.48	15.71	0.27	0.15	21.41	14	36
49,397	187.56	81.19	0.43	0.15	73.99	77	131
51,919	2,899.40	2,881.98	0.99	0.15	1522.40	2,889	2,829
868,182	5.32	2.28	0.43	0.15	2.09	38	65
606,830	13.69	2.22	0.16	0.15	4.96	26	108
434,438	22.33	2.65	0.12	0.15	8.04	22	125
379,167	43.80	12.07	0.28	0.15	16.32	88	222
222,916	218.83	304.94	1.39	0.15	141.61	1,312	1,130
18,789	28.31	21.62	0.76	0.05	7.84	8	5
5,443	187.46	81.89	0.44	0.05	44.77	9	9
6,025	2,497.82	2,700.62	1.08	0.05	813.03	314	175
161,560	9.47	5.11	0.54	0.05	2.36	16	14
56,320	26.13	5.50	0.21	0.05	5.83	6	12
31,860	57.34	11.70	0.20	0.05	12.77	7	15
23,217	235.95	286.11	1.21	0.05	82.08	128	68