141

W. Edwards Deming

On Variances of Estimators of a Total Population Under Several Procedures of Sampling

Reprinted from

Contributions to Applied Statistics

Dedicated to Professor Arthur Linder

Edited by Walter Joh. Ziegler

1976 Birkhäuser Verlag, Basel und Stuttgart

W. Edwards Deming

W. Edwards Deming

On Variances of Estimators of a Total Population Under Several Procedures of Sampling

1. **Introduction**

The aim here is to compare several simple plans of sampling that often appear to be equal, but which may give widely different degrees of precision when put into use. For example, it is well known that if we draw with equal probabilities and without replacement a sample of pre-determined size n from a frame of N sampling units (Plan I ahead), and if N be known and used in the estimator $X = N\,\bar{x}$, in the notation set forth in the next section, then X is an unbiased estimator of the total population A of the frame, and the Var X is given by Equation (5) ahead.

It is also well known that if the frame contains a proportion Q of blanks (sampling units that are not members of the universe), then the variance of an estimate of the total of some extensive characteristic of the frame increases out of proportion to Q, while the variance of the ratio of two characteristics suffers only from the diminished number of sampling units that come from the universe.

Not so well known is the effect of certain tempting procedures of selection in which the size n of the sample turns out to be a random variable. The purpose here is to examine and compare some of the alternatives.

One special case of importance is where one aim of the study is to estimate the total number N of sampling units in the frame. We first of all need some notation.

Notation:

P	probability before selection that any sampling unit in the frame will fall into the sample. In Plan I, P is the so-called sampling fraction. $Q = 1 - P$.
N	number of sampling units in the frame.
n	number of sampling units in the sample in Plan I and in Plan III.
$\hat{n}$	number of sampling units in a particular sample in Plan II.

a_i — the x-population of sampling unit No. i in the frame. a_i will be 0 if sampling unit No. i is not a member of the universe. a_i may also be 0 even if sampling unit No. i is a member of the universe.

$A = \sum^{N} a_i$ — the total x-population in the frame.

$a = A/N$ — the average x-population per sampling unit in the frame, including 0-values of a_i.

$\sigma^2 = \frac{1}{N} \sum^{N} (a_i - a)^2 = a^2(C_1^2 + Q)$ the overall variance between the a_i in the frame, including the 0-values of the a_i.

C_1 — the coefficient of variation between the non-zero a_i.

$C = \sigma/a$ — the coefficient of variation between the a_i in the frame, including the 0-values of a_i.

We first compare two plans, which we shall call Plan I and Plan II, for estimation of the total x-population of a frame: later, for estimation of a ratio. In both these plans the probability that a sampling unit will be selected into the sample will be P. In both plans we presume the existence of a frame, N known in some problems, not known in others.

2. Estimates of a Total Population

Plan I. n fixed at $n = NP$. N known. To select the sample, read out n unduplicated random numbers between 1 and N. This plan is sometimes called simple random sampling. Record the sample as $x_1, x_2, \ldots, x_n$, in order of selection. Compute

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)\,, \tag{1}$$

$$X = N\bar{x}\,. \tag{2}$$

Then

$$E\,X = A\,, \tag{3}$$

$$E\,\bar{x} = a \tag{4}$$

That is, X is an unbiased estimator of A, and $\bar{x}$ is an unbiased estimator of a.

$$\operatorname{Var} X = N^2 \frac{N-n}{N-1} \frac{\sigma^2}{n} \doteq N^2 \left(\frac{1}{n} - \frac{1}{N}\right) \sigma^2\,, \tag{5}$$

$$\operatorname{Var} \bar{x} = \frac{N-n}{N-1} \frac{\sigma^2}{n} \doteq \left(\frac{1}{n} - \frac{1}{N}\right) \sigma^2 \tag{6}$$

For the rel-variances

$$C_X^2 = C_{\bar{x}}^2 \doteq \left(\frac{1}{n} - \frac{1}{N}\right) C^2 . \tag{7}$$

All this is well known. The proofs are in any book on sampling.

Plan II. P fixed, N may be known or unknown; $\hat{n}$ a random variable. To select the sample, start with sampling unit No. 1. Accept it or reject it, depending on a side-play of random numbers. For example, if $P = .01$, read out a 2-digit random number between 01 and 00, all 2-digit random numbers to have equal probabilities. Let 01 accept the unit, 02 to 00 reject it. Then go to sampling unit No. 2; read out another random number between 01 and 00 with the same side-play and same rule. Go to No. 3, then to No. 4, and onward through the whole frame to N, always with the same side-play.

$$E\,\hat{n} = N\,P\,, \tag{8}$$

$$\operatorname{Var}\hat{n} = N\,P\,Q\,. \tag{9}$$

We here define X by Equation (15) ahead and note that

$$E\,X = A\,, \tag{10}$$

so we have again an unbiased estimator of A, but here

$$\begin{aligned} \operatorname{Var} X &= \frac{N\,Q}{P}\,(\sigma^2 + a^2) \\ &= \frac{N^2\,Q}{E\,\hat{n}}\,(\sigma^2 + a^2)\,, \end{aligned} \tag{11}$$

$$C_X^2 = \frac{Q}{E\,\hat{n}}\,(C^2 + 1)\,. \tag{12}$$

The proofs will appear in a minute.

Proof of the expected value and variance of X in Plan II:

$$\begin{cases} \delta_i = 1 & \text{if sampling unit No. } i \text{ of the frame falls into the sample,} \\ \quad\;\; = 0 & \text{otherwise.} \end{cases}$$

We note that δ_i is a random variable and that

$$\delta_i^2 = \delta_i\,, \tag{13}$$

$$E\,\delta_i^2 = E\,\delta_i = P\,. \tag{14}$$

Define

$$X = \sum a_i \, \delta_i / P \, . \tag{15}$$

[Here and henceforth all sums will run from 1 to N unless marked otherwise.]

This is equivalent to

$$X = \frac{1}{P} x \, , \tag{15a}$$

where x is the total of the x-values in the sample. Then

$$E\,X = \frac{1}{P} \sum a_i \, E\,\delta_i = \sum a_i = A$$

which is Equation (10).

$$\begin{aligned}
\operatorname{Var} X &= \sum (X - E\,X)^2 \\
&= E(\sum a_i \, \delta_i/P - A)^2 \\
&= E(\sum a_i \, \delta_i/P - \sum a_i)^2 \\
&= E[\sum a_i(\delta_i/P - 1)]^2 \\
&= E \sum a_i^2(\delta_i/P - 1)^2 + E \sum_{j \neq i} a_i \, a_j(\delta_i/P - 1)\,(\delta_j/P - 1) \\
&= \sum a_i^2 \, E(\delta_i^2/P^2 - 2\,\delta_i/P + 1) + 0 \quad \text{[Because } \delta_i \text{ and } \delta_j \text{ are independent]} \\
&= \sum a_i^2(P/P^2 - 2\,P/P + 1) \\
&= \frac{Q}{P} \sum a_i^2 = \frac{Q}{P} \sum [(a_i - a) + a]^2 \\
&= \frac{N\,Q}{P} (\sigma^2 + a^2)
\end{aligned}$$

which is Equation (11)[1].

[1]) I am indebted to my friend William N. Hurwitz, deceased, for this proof of Equation (11).

Remark 1. We pause to note that the difference in variances between Plans I and II may be alarming, or it may be inconsequential. To compare their variances, we set $E\,\hat{n}$ in Equation (11) equal to n in Equation (5) and write

$$\frac{\mathrm{Var}\,(\mathrm{II})}{\mathrm{Var}\,(\mathrm{I})} = \frac{\sigma^2 + a^2}{\sigma^2} = 1 + a^2/\sigma^2 \to 1 \quad \text{as} \quad a/\sigma \to 0\,. \tag{16}$$

This equation tells us that Plan II will always yield variance higher than Plan I will yield, and that the difference will be small only if a be small compared with σ. We shall return later to this comparison when we study the effect of blanks (0-values of a_i in the frame).

Remark 2. We note that for Plan II, $E\,x_i^2 = (1/N)\,\Sigma\,a_i^2 = \sigma^2 + a^2$ for any member i of the sample. Hence any x_i^2 in the sample is an unbiased estimator of $\sigma^2 + a^2$, and a sample of size $n = 1$ provides an estimate of $\mathrm{Var}\,X$ (noted privately by my friend and colleague the late William N. Hurwitz).

Remark 3. The appendix shows for illustration all the possible samples of $n = 1$ for $P = Q = 1/2$ that can be drawn from a frame of $N = 2$ sampling units, along with calculations and comparisons with some of the formulas just learned, and with some that will appear in section 5.

Plan III. Here, we separate out in advance the blanks, or attempt to do so. This plan has advantages and disadvantages. The required separation (screening) is sometimes costly. Plan III should be chosen only after careful computation of the expected variances and costs. An example and references appear later.

An example of Plan II. The problem is to estimate the total number N of fish that traverse a channel in a season. A shunt provides an alternate path, attracting into the shunt some average fraction P of the fish. It is a fairly simple matter throughout the season to count the fish that traverse the shunt, but not so easy to count the fish that traverse the channel. However, it is possible to count on a few selected days the fish that traverse the channel. Comparison of the counts of fish that traverse shunt and channel provides an estimate of P. The variance σ^2 between sampling units would be 0, as every sampling unit in the frame has the value 1. Then under the assumption that P is constant through the season, we could set $X = \hat{N} = \hat{n}/P$ for an estimate of N, where $\hat{n}$ is the number of fish that traversed the shunt during the season. Equation (12) would then give the conditional

$$\text{Rel-Var}\,\hat{N} = \frac{1 - P}{\hat{n}} \tag{17}$$

Of course, the estimate $\hat{N} = \hat{n}/P$ would be no better than the estimate of P derived from the ancillary studies, but the estimate of the rel-variance of $\hat{N}$ derived from Equation (17) would be excellent if P be small.

Estimates could be made by direction of flow, upstream and downstream separately, and by big fish and little fish, if desired. The equation just written would give the conditional rel-variance of the estimate of any class of fish.

Another example. Any scheme for reduction of the probability of selection of sampling units that have some specified characteristic (such as certain items of low value) by use of random thinning digits or their equivalent should be examined carefully for the hazards of extra variance in the estimate of a total. One must weigh the simplicity and variance of Plan II against the lesser variance and possible extra costs of using Plan I.

A specific example of blanks may be described as follows. Suppose that the frame consists of $N = 3\,000\,000$ freight bills filed in numerical order in the office of a carrier of motor freight (possibly the inter-city hauls for one year). The management needs a sample of these shipments in order to study relations between revenues, rates, and costs as a function of weight, size, distance, and other characteristics of shipments. We suppose that the sample desired is 1 in 50 of the shipments that weigh 10000 lbs. or over, and 1 in 500 of those that weigh less than 10000 lbs. To make the selection, we list from the files on a pre-printed form 1 shipment in 50 (a systematic selection of every 50th shipment would serve the purpose); retain for the final sample every shipment listed that weighs 10000 lbs. or over, and select with probability 1 in 10 all other shipments. Suppose that the probability of 1 in 10 is achieved by preprinting the form with the symbol S on 1 line in 10, in a random pattern. Lines 1-11 on the form, when filled out, might appear as shown in Table 1.

Table 1

Line	Serial number	Weight (lbs.)	Remarks
1	CH 105474	2650	Not in sample
2	CH 105524	24450	In sample
3	CH 105574	220	Not in sample
4	CH 105624	175	Not in sample
5	CH 105674	800	Not in sample
6	CH 105724	720	Not in sample
7	CH 105774	15500	In sample
8 S	CH 105824	2750	In sample
9	CH 105874	120	Not in sample
10	CH 105924	13300	In sample
11 S	CH 105974	700	In sample
etc.			

The procedure of preprinting a form is tempting for its simplicity. But let us look at the variance of the estimate of (e.g.) the total revenue from shipments under 500 pounds. Let x be the aggregate revenue in the sample from these shipments. Then

$$X = 500\,x$$

will be an unbiased estimate of the revenue in the frame from shipments under 500 pounds. Unfortunately, Var X is afflicted with the term a^2 in Equation (11). The symbol a is the average revenue per shipment, for shipments of all weights, and $P = 1/500$. In practice, σ/a may be anywhere from .25 to .60. The term a^2 thus adds substantially to the variance of X.

A way out is to stratify into two strata the preliminary sample consisting of every 50th shipment, the two strata being (1) 10000 lbs. or over, and (2) under 10000 lbs. The 11 freight bills in Table 1 would now appear in two columns, as in Table 2. The symbol S in Table 1 is no longer needed: we take into the final sample every shipment listed under Stratum 1, and a selection of 1 from every consecutive 10 of the shipments listed under Stratum 2. We may form the estimate X as above, and the term a^2 in the variance will now disappear.

Stratification, serialization, and selection all require care, time, and supervision. Moreover, in practice, in the application to motor freight, there are 6 strata, not 2, with consequent enlargement either of errors or of care and supervision.

Table 2

Line No.	Serial number	Stratum 1 10000 lbs. or over	Stratum 2 under 10000 lbs
1	CH 105474		2650
2	CH 105524	24450	
3	CH 105574		220
4	CH 105624		175
5	CH 105674		800
6	CH 105724		720
7	CH 105774	15500	
8	CH 105824		2750
9	CH 105874		120
10	CH 105924	13300	
11	CH 105974		700

When the record of shipments is on a tape, it is possible to stratify the shipments accurately in a number of strata and to select the sample from any stratum with a fixed proportion, thus eliminating the random character of the sizes of the samples. The extra cost is negligible if the stratification and selection be carried out along with other tabulations, all in one pass of the tape.

3. Estimates of a Ratio

A sampling unit has not only an x-value but a y-value. Thus, a sampling unit might be a household, b_i the number of people therein in the labor force, a_i the number of people in the household that are in the labor force and unemployed. Then

$$B = \sum b_i \tag{18}$$

is the total number of people in the labor force, and

$$A = \sum a_i \tag{19}$$

is the total number of people in the labor force and unemployed. Put

$$a = \frac{A}{N} \tag{20}$$

[the average number of people per household in the labor force and unemployed]

as before, and

$$b = \frac{B}{N}\,. \tag{21}$$

[the average number of people per household]

Then

$$\varphi = \frac{A}{B} = \frac{a}{b} \tag{22}$$

is the overall proportion of people in the labor force unemployed. Suppose that we wish to estimate this proportion.

Plan I and Plan II both give estimates of A, B, and of $\varphi = A/B$. After seeing the possible losses in the use of Plan II for estimation of a total population, one may be astonished to learn that (so far as we carry our approximations to variances) Plan II gives for estimation of a ratio the same variance as Plan I, for a given size of sample. The proof will follow.

Plan I for a ratio. We first define the x- and y-variances between sampling units as

$$\left.\begin{aligned} \sigma_x^2 &= \frac{1}{N}\sum^{N}(a_i - a)^2 \\ \sigma_y^2 &= \frac{1}{N}\sum^{N}(b_i - b)^2 \end{aligned}\right\} \tag{23}$$

and the covariance

$$\sigma_{xy} = \frac{1}{N} \sum^{N} (a_i - a)(b_i - b) \tag{24}$$

A sample drawn and processed by Plan I gives unbiased estimates of X and of Y by use of Equation (2). It gives also the ratio

$$f = \frac{X}{Y} = \frac{\bar{x}}{\bar{y}} \tag{25}$$

as the sample analog of $\varphi = A/B$. For the variance of f, we shall be satisfied with the usual Taylor approximation wherein

$$\text{Rel-Var} f = \text{Rel-Var} \frac{X}{Y} = \left(\frac{1}{n} - \frac{1}{N}\right)(C_x^2 + C_y^2 - 2\,\varrho\, C_x\, C_y) \tag{26}$$

which is satisfactory if n is big enough. Here

$$C_x = \frac{\sigma_x}{a} \tag{27}$$

[the coefficient of variation between all the a_i in the frame],

$$C_y = \frac{\sigma_y}{b} \tag{28}$$

[the coefficient of variation of all the b_i in the frame],

and

$$\varrho = \frac{\sigma_{xy}}{\sigma_x\,\sigma_y} \tag{29}$$

is the correlation between the N pairs of values a_i and b_i.

Plan II for a ratio. Again, $E\,X = A$ by Equation (10). Also, $E\,Y = B$. Equation (11) gives

$$\text{Var}\,X = \frac{N\,Q}{P}(\sigma_x^2 + a^2)\,, \tag{30}$$

$$\text{Var}\,Y = \frac{N\,Q}{P}(\sigma_y^2 + b^2)\,, \tag{31}$$

$$\text{Var}\,\frac{X}{Y} = \text{Var}\,\frac{\Sigma\, a_i\, \delta_i}{\Sigma\, b_i\, \delta_i} \tag{32}$$

The approximation written as Equation (26) now leads to

$$\text{Rel-Var}\,\frac{X}{Y} = C_X^2 + C_Y^2 - 2\,C_{XY} \tag{33}$$

The rel-variances C_X^2 and C_Y^2 have been conquered, but we have yet to evaluate C_{XY}. First, by definition

$$\text{Cov}\,X,\,Y = E\left[\sum \frac{a_i\,\delta_i}{P} - E\sum \frac{a_i\,\delta_i}{P}\right]\left[\sum \frac{b_i\,\delta_i}{P} - E\sum \frac{b_i\,\delta_i}{P}\right]$$

$$= E\left[\sum a_i\left(\frac{\delta_i}{P} - 1\right)\right]\left[\sum b_i\left(\frac{\delta_i}{P} - 1\right)\right]$$

[all the sums run over $i = 1$ to $i = N$]

$$= E\sum a_i\,b_i\left(\frac{\delta_i}{P} - 1\right)^2 + E\sum_{j\neq i} a_i\,b_j\left(\frac{\delta_i}{P} - 1\right)\left(\frac{\delta_j}{P} - 1\right)$$

$$= \sum a_i\,b_i\,E\left(\frac{\delta_i^2}{P^2} - \frac{2\,\delta_i}{P+1}\right) + 0$$

$$= \sum a_i\,b_i\left(\frac{P}{P^2} - \frac{2\,P}{P+1}\right) = \frac{Q}{P}\sum a_i\,b_i$$

$$= \frac{N\,Q}{P}\,a\,b + \frac{N\,Q}{P}\,\varrho\,a\,b\,C_x\,C_y\,, \tag{34}$$

wherein ϱ, C_x, and C_y have already been defined. We return now to Equation (33) for Plan II, whence

$$\text{Rel-Var}\,\frac{X}{Y} \doteq C_X^2 + C_Y^2 - 2\,C_{XY} \quad \text{[Equation (33)]}$$

$$= \frac{\text{Var}\,X}{(E\,X)^2} + \frac{\text{Var}\,Y}{(E\,Y)^2} - 2\,\frac{\text{Cov}\,X,\,Y}{EX\,EY}$$

$$= \frac{N\,Q}{P(N\,a)^2}\,(\sigma_x^2 + a^2) + \frac{N\,Q}{P(N\,b)^2}\,(\sigma_y^2 + b^2)$$

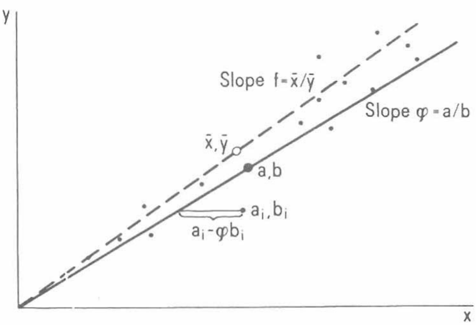
$$- 2\,\frac{N\,Q}{P\,N^2\,a\,b}\,(a\,b + \varrho\,a\,b\,C_x\,C_y)$$

$$= \frac{Q}{N\,P}\,(C_x^2 + C_y^2 - 2\,\varrho\,C_x\,C_y)\,, \tag{35}$$

which after replacement of $N P$ by $E \hat{n} = n$ appears to be precisely what we wrote in Equation 26 for Plan I. Thus, although Plan II may show a severe loss of precision for the estimates X and Y of the total x- and y-populations in the frame, it is equivalent to Plan I for the ratio X/Y.

We note in passing that, by algebraic rearrangement, Equations (26) and (35) may be written as

$$\text{Rel-Var}\,\frac{X}{Y} \doteq \left(\frac{1}{n} - \frac{1}{N}\right)\frac{1}{N}\sum\left[\frac{a_i - \varphi\, b_i}{a}\right]^2 \tag{36}$$



The complete coverage of the frame would give the centroid a, b. A line through the centroid and the origin would have slope $\varphi = a/b$. The sample of points has the centroid $\bar{x}/\bar{y}$. The line that connects it with the origin has slope $f = \bar{x}/\bar{y}$.

The factor

$$\frac{1}{N}\sum\left[\frac{a_i - \varphi\, b_i}{a}\right]^2$$

is the average square of the vertical deviations of the N points a_i, b_i from the line $x = \varphi\, y$, measured in units of a, where $\varphi = A/B = a/b$. The sample-analog

$$\text{Rel-}\widehat{\text{Var}} f = \text{Rel-}\widehat{\text{Var}}\,\frac{X}{Y} \doteq \left(\frac{1}{n} - \frac{1}{N}\right)\frac{1}{n\,\bar{x}^2}\sum_1^n (x_i - f\,y_i)^2 \tag{37}$$

may be used as an estimator of the rel-variance of X/Y, though we usually replace $n\,\bar{x}^2$ by $(n-1)\,\bar{x}^2$.

4. Effect of Blanks in the Frame

Illustration from practice

It often happens in practice that one wishes to estimate the aggregate value of some characteristic of a subclass when the total number of units in the subclass is unknown. For example, in a study of consumer research there was need for an estimate of the number of women aged 30 or over that live in a certain district, with no child under 12 years old at home: also the total disposable income of these women.

Suppose for simplicity that the frame is a list of all the occupied dwelling units in the district. Our sample will be a simple random sample of n dwelling units, drawn without replacement by reading out n random numbers between 1 and N, where N is the total number of dwelling units in the district. We depart for convenience from the notation at the front and use the subscript 1 for the specified subclass. Information is obtained on the n dwelling units in the sample, and it is noted that $\hat{n}_1$ is the count of dwelling units in this sample that contain women that belong to the specified subclass—that is, females 30 or over with no child under 12. Let $\bar{x}_1$ be the average income per dwelling unit in these $\hat{n}_1$ dwelling units. Some incomes in the specified subclass may be 0. $\hat{n}_1$ and $\bar{x}_1$ are both random variables: so is their product $\hat{n}_1 \bar{x}_1$, the total income of the women in the sample that belong to the specified subclass.

The reader will recognize the above sampling procedure as Plan I. We encounter in practice two main problems:

Problem 1. What is the variance of a ratio such as $\bar{x}_1$?

Problem 2. What is the variance of an estimator of a total, such as the total number of women in the subclass, or their total income?

We note first that the conditional expected value of $\bar{x}_1$ over all the samples that have n_1 dwelling units that contain women that belong to the specified subclass has the convenient property of being the average income of all the women in the frame that belong to the subclass. It is for this reason that the conditional rel-variance of $\bar{x}_1$ is useful for assessing the precision of a sample at hand.

What is the rel-variance of $\bar{x}_1$? Let C_1^2 be the rel-variance of incomes between the dwelling units in the frame that belong to the specified subclass. It is a fact that for the plan of sampling described here, the conditional rel-variance of $\bar{x}_1$, for samples of size $\hat{n}_1$ of the specified subclass, will be very nearly

$$\text{Rel-Var}\,\bar{x}_1 = \left(1 - \frac{n}{N}\right) \frac{C_1^2}{\hat{n}_1} \tag{38}$$

as if the dwelling units of this subclass in the frame had been set off beforehand in a separate stratum (Stratum 1), and a sample of size $\hat{n}_1$ drawn therefrom.

Incidentally, in the design-stage, for calculation of the size of sample to meet a prescribed $\text{Var}\,\bar{x}_1$, one may speculate on a value of P for the proportion of women in the frame that belong to the specified subclass, and then for the sampling procedure described above calculate the required size n of the sample by use of the formula

$$\text{Av}\,\text{Var}\,\bar{x}_1 = C_1^2\, E\,\frac{1}{\hat{n}_1} \doteq \frac{C_1^2}{n\,P}\left(1 + \frac{Q}{n\,P}\right).$$

The term Q/nP in the parenthesis arises from the fact that $\hat{n}_1$ is unpredictable, being a random variable.

We now turn our attention to Problem 2, estimation of the total number of women in the subclass. An estimator of X_1 will be

$$X_1 = \frac{N}{n}\,\hat{n}_1\,\bar{x}_1 \tag{40}$$

using N/n as an expansion factor. If we place $x_i = 1$ for dwelling unit i in the sample if it contains a woman in the specified subclass, and place $x_i = 0$ otherwise, then X_1 will be an estimate of the number of women in the frame that belong to the specified subclass. We note immediately that, as N and n are known (not random), the conditional rel-variance of X_1 for samples of size $\hat{n}_1$ is exactly equal to the rel-variance of $\bar{x}_1$.

Unfortunately, though, this estimator X_1 of the total number or total income of all the women in the frame that belong to the specified subclass will not have all the convenient properties of $\bar{x}_1$. Thus, the conditional expectation of $X_1 = \hat{n}_1\, E\,\bar{x}_1$ for samples that contain $\hat{n}_1$ members of the specified subclass is not equal to the aggregate income of all the women in the frame that belong to the subclass. One must conclude that the conditional rel-variance of X_1 for a sample at hand, although equal to the conditional rel-variance of $\bar{x}_1$, requires careful interpretation.

Instead of attempting to interpret the conditional variance of X_1, we may turn our attention to the average variance of X_1 in all possible samples of size n. We need more symbols. Subscript 1 will refer to the specified subclass; subscript 2 to the remainder. The word income will hereafter mean income from women of the specified subclass.

a_1	the average income per dwelling unit in the frame for the women that belong to the specified subclass. (Some incomes may of course be 0 in this subclass.)
$a_2 = 0$	for the remainder, because every sampling unit not in the specified subclass is a blank.
P	the proportion of the dwelling units in the frame that contain women of the specified subclass. $Q = 1 - P$.
σ_1^2	the variance in incomes between the dwelling units in the frame that belong to the specified subclass.
$a = P\,a_1$	the overall average income per dwelling unit in the frame, for the women of the specified subclass, including blanks (dwelling units with no women of the specified subclass).

We note that the overall variance between the incomes in all N dwelling units of the frame will be

$$\begin{aligned}\sigma^2 &= P\,\sigma_1^2 + Q\,\sigma_2^2 + P\,Q(a_2 - a_1)^2 \quad [\sigma_2 = 0]\\ &= P(\sigma_1^2 + Q\,a_1^2) = P\,a_1^2(C_1^2 + Q)\,.\end{aligned} \tag{41}$$

To find the average $\operatorname{Var} X_1$ over all possible samples, we may then use Equation (5), which gives

$$\operatorname{Var} X_1 = \left(1 - \frac{n}{N}\right) N^2 \sigma^2/n$$

$$= \left(1 - \frac{n}{N}\right) \frac{N^2 P a_1^2(C_1^2 + Q)}{n} \tag{42}$$

or in terms of rel-variance,

$$\text{Rel-Var } \bar{x}_1 = \left(1 - \frac{n}{N}\right) \frac{C_1^2 + Q}{n P} \tag{43}$$

in which we recognize $n P$ as $E \hat{n}_1$. The average variance of X_1 is thus afflicted by the proportion Q of blanks, whereas the average variance of $\bar{x}_1$ in Equation (38) is not.

As the proportion of blanks Q increases toward unity, Plan I becomes more and more the equivalent of Plan II with the same probability P of selection.

The problem with the variance of X_1 in Plan II arose from the assumption that N_1, the number of dwelling units in the frame with women that meet the specification of the subclass, is unknown. If N_1 were known, as sometimes it is, one could form $\bar{x}_1$ from the sample and then use the estimator

$$N_1 = N_1 \bar{x}_1 \tag{44}$$

which would have all the desirable properties of $\bar{x}_1$.

This observation suggests use of a preliminary sample by which to estimate the proportion of the total frame that belongs to the specified class. Briefly, the procedure is this: (1) to select from the frame by random numbers a preliminary sample of sufficient size N'; (2) to classify into strata by an inexpensive investigation, the units of the preliminary sample; (3) to investigate samples of sizes $\hat{n}_1$ and $\hat{n}_2$ from the two strata, to acquire the desired information. The preliminary sample furnishes estimates $\hat{P}_1$ and P_2 of the proportions of the two strata, and the final sample gives $\bar{x}_1$ and $\bar{x}_2$. The estimator is

$$\bar{x} = \hat{P}_1 \bar{x}_1 + \hat{P}_2 \bar{x}_2 . \tag{45}$$

The final sample may be selected proportionately from the strata of the preliminary sample, or (where advantageous) by Neyman allocation.

If the sorting into strata is successful, then the sample from Stratum 2 can be relatively small. It is in practice risky to reduce it to 0 for the simple reason that in most experience a few false positives in Stratum 2 are very effective in increasing the variance of $\bar{x}$.

Approximate variances for the two allocations are

$$\operatorname{Var} \bar{x} = \frac{\sigma_b^2}{N'} + \frac{\sigma_w^2}{n} \quad \text{[Proportionate allocation]} \tag{46}$$

and

$$\operatorname{Var}\bar{x} = \frac{\sigma_b^2}{N'} + \frac{(\bar{\sigma}_w^2)}{n}, \quad \text{[Neyman allocation]} \tag{47}$$

where σ_w^2 is the usual weighted average variance between sampling units within strata, and $\bar{\sigma}_w$ is the weighted average standard deviation between sampling units within strata. σ_b^2 is the variance between the means of ths strata.

There is an optimum size for the preliminary sample given by

$$\frac{n}{N'} = \frac{\sigma_w}{\sigma_b}\sqrt{\frac{c_1}{c_2}}, \quad \text{[Proportionate allocation]} \tag{48}$$

$$\frac{n}{N'} = \frac{\bar{\sigma}_w}{\sigma_b}\sqrt{\frac{c_1}{c_2}}, \quad \text{[Neyman allocation]} \tag{49}$$

where c_1 is the average cost to classify a sampling unit into a stratum, and c_2 is the average cost to investigate a unit in the final sample.

The theory is well known and need not be elaborated here. Such problems are complicated by the fact that estimation for several subclasses may be required in the same study.

Examples of blanks in the frame will be found in almost any book on sampling, one of the best being Chapter 9 in the 3rd edition of Frank Yates, *Sampling Methods for Censuses and Surveys* (Griffin, 1971). An example of calculations for a choice between Plans I and III appears in the author's book *Sample Design in Business Research* (Wiley, 1960), page 129.

We end on a further note of possible interest. If all the a_i in the specified subclass take the value 1, then $\sigma_1^2 = 0$ in Equation (41). Suppose now that the proportion Q of blanks approaches 1 and that n increases in a manner that holds $n\,P = m$. This circumstance corresponds to a count of flaws in test-panels of fixed size (fixed n, as of paint, or of a textile) in which the number of flaws in a test-panel may for practical purposes be infinite, but with an expected value of m. Equation (43) then leads to the Poisson

$$\text{Rel-Var}\,\hat{m} \to \left(1 - \frac{n}{N}\right)\frac{1}{m}, \tag{50}$$

n/N being the proportion of all panels that are observed.

5. Appendix: Illustration of Plan II with a Frame of Two Units

Table 3
The frame.

Serial numbers of sampling unit	Populations	
	x	y
	$a_1 = 1$	$b_1 = 3$
	$a_2 = 2$	$b_2 = 5$

Table 4
Statistical properties of the frame.

Total population	$A = 3$	$B = 8$
Average per sampling unit	$a = 1.5$	$b = 4$
Standard deviation	$\sigma_x = 1/2$	$\sigma_y = 1$
Coefficient of variation	$C_x = 1/3$	$C_y = 1/4$

$\varphi = A/B = a/b = 3/8 = .375,$

$\mathrm{Cov}\, x, y = \frac{1}{2}\,(.5 \cdot 1 + .5 \cdot 1) = 1/2,$

$\varrho = \mathrm{Cov}\, x, y/\sigma_x\, \sigma_y = \frac{1}{2}/\frac{1}{2} \cdot 1 = 1$ (always true with 2 points),

$C_{xy} = \mathrm{Cov}\, x, y/a\, b = \frac{1}{2}/1.5 \cdot 4 = 1/12.$

We now list the 4 possible outcomes of the sampling procedure for $P = Q = 1/2$. Their expected proportions are equal. We observe that:

1. $E\, \hat{n} = \frac{1}{4}\,(0 + 1 + 1 + 2) = 1$. $NP = 2 \cdot 1/2 = 1$, in agreement.

2. Av $X = 3 = A$, which illustrates the unbiased character of the sampling procedure. Likewise, Av $Y = 8 = B$.

3. $\mathrm{Var}\, X = \frac{1}{4}\{(0-3)^2 + (4-3)^2 + (2-3)^2 + (6-3)^2\} = 20/4 = 5.$

In comparison, the formula for $\mathrm{Var}\, X$ gives

$$\mathrm{Var}\, X = (N\, Q/P)\,(\sigma_x^2 + a^2)$$

$$= 2(\tfrac{1}{4} + 1.5^2) = 20/4 = 5\,.$$

Obviously, most of this variance comes from the term $a^2 = 1.5^2$.

Table 5
Table of all possible samples selected from Plan II from the frame shown above, with $P = Q = 1/2$.

Sampling units in sample	x-population of sample x	y-population of sample y	$X = 2x$	$Y = 2y$	X/Y	x
Both out	0	0	0	0	–	–
No. 1 out, No. 2 in	2	5	4	10	4/10	2
No. 1 in, No. 2 out	1	3	2	6	2/6	1
Both in	3	8	6	16	6/16	1.5
Average	1.5	4	3	8	133/360	1.5

4. Suppose that we know N, and that we use the estimator

$$X' = N\,\bar{x} = 2\,\bar{x}$$

for the total x-population. The three useable values of X' would then be 4, 2, 3, whose average value agrees with $A = 3$. We note that

$$\operatorname{Var} X' = \frac{1}{3}\,[(4-3)^2 + (2-3)^2 + (3-3)^2]$$

$$= \frac{2}{3}$$

which is much less than $\operatorname{Var} X = 5$, just encountered. $\operatorname{Var} X'$ has all the desirable properties of $\bar{x}$.

5. Every sample, if it contains a sampling unit, gives an estimate of $\varphi = A/B = 3/8 = .375$. The 3 possible estimates are in the table. Their average is $133/360 = .3694$. The sampling procedure for estimation of φ is therefore slightly biased, as statistical theory would lead us to expect. The bias is incidentally $.3694 - .3750 = .0056$, being only 15 parts in 1000, or only 4.7% of the standard error of X/Y.

6. Equation (34) gives the approximation

$$C^2_{X/Y} \doteq \frac{1}{E\,\hat{n}}\,[C^2_x + C^2_y - 2\,C_{xy}]$$

$$= \frac{1}{1}\left[\left(\frac{1}{3}\right)^2 + \left(\frac{1}{4}\right)^2 - 2/12\right]$$

$$= \frac{1}{9} + \frac{1}{16} - \frac{2}{12} = \frac{1}{144} = .006\,944\,.$$

7. The table of all possible samples gives

$$\sigma^2_{X/Y} = \frac{1}{3}\,[(4/10 - 133/360)^2 + (2/6 - 133/360)^2 + (6/16 - 133/360)^2]$$

$$= .002\,862/3$$

$$= .000\,954\,3$$

$$C^2_{X/Y} = .000\,954\,3/(133/360)^2$$

$$= .006\,991\,7$$

in closer agreement with .006944 than we might expect for samples of $n = 1$.

Acknowledgement

I am deeply indebted to my friend and colleague Dr. Morris H. Hansen for calling my attention years ago to Plan II, and for his continued interest and help in the theory and comparison between Plans I, II, and III in practice. I have already expressed my indebtedness to William N. Hurwitz. It is a pleasure to mention also my good fortune to work with Professor William H. Kruskal on the problem of blanks, during the preparation of my article Survey Sampling for the New Encyclopedia of the Social Sciences.

Author's address:
W. Edwards Deming, 4924 Butterworth Place, Washington 20016.