ON SAMPLE VARIANCE ESTIMATION

Francesca Gagliardi

Jean Monnet Chair

Pisa, the 9th and 16th of May 2016

The design of a random sample specifies the type of randomised procedure applied in *sample selection*. It also specifies how the population parameters are to be *estimated* from the sample results.

The selection procedure and the estimation procedure form two aspects of the *sample design*.

As to the selection procedure, many types of designs are possible and used in practice.

The procedure may for example give the same (equal) chance of appearing in the sample to all elements in the population, or some units may be given a higher chance than others.

We may select the elements individually, or first group them into larger clusters and apply the selection procedure to those clusters.

We may partition the population into strata and apply any of the above procedures separately within each stratum.

Each randomised procedure in fact determines a different set of samples which can in principle be selected using that procedure and the chance of selecting a particular sample from among those.

But to be random, any selection procedure must ensure that every unit in the population receives a specified non-zero chance of appearing in the sample to be selected.

The *estimation procedure* involves the statistical or mathematical forms in terms of sample values and possibly also of information from other sources external to the sample; it provides *estimators* which are used to produce *sample estimates* of population parameters of interest.

The procedure also includes the estimation of measures of uncertainty ('sampling error', 'confidence intervals' etc.) to which the sample results are subject.

The particular units which happen to be selected into a particular sample depends on chance, the possible outcomes being determined by the procedures specified in the sample design. This means that, even if the required information on every selected unit is obtained entirely without error, the results from the sample are subject to a degree of uncertainty due to these chance factors affecting the selection of units.

Sampling variance is a measure of this uncertainty.

The distribution of estimates from all possible samples with a given design (i.e. selection and estimation procedure) is called the *sampling distribution of the estimator*.

The average of the sampling distribution, i.e. of all possible samples estimates weighted according to their probabilities, is called the *expected value*.

Symbolically we may express this as follows. If p_s is the probability and y_s the estimate from a given sample s, the *expected value* of the estimator y is:

$$E(y) = \sum_s p_s \cdot y_s$$

where the sum is taken over all possible samples.

The *variance* of y is defined as:

$$Var(y) = \sum_s p_s \cdot [y_s - E(y)]^2$$

For various reasons, the expected or average value from all possible samples may not equal the actual population value (Y).

In the absence of measurement errors, this may arise from the particular estimation procedure, in which case it is called the technical or estimation *bias*:

$$\text{Bias} = E(y) - Y$$

The combined effect of variance and bias is the *mean squared error*, which is defined in terms of the squared differences of sample estimates y_s from the actual population value Y:

$$MSE(y) = \sum_s p_s \cdot [y_s - Y]^2 = Var(y) + (Bias)^2$$

An illustration from Simple random sampling

The simplest design is one in which every possible set of, say, *n* units from a population of *N* units receives the same chance of selection.

This is called a simple random sample (SRS).

There are $\frac{N!}{(N-n)!\cdot n!}$ such samples, and each receives a probability of selection equalling inverse of the above number.

In fact, any set of *s*, *1≤s≤n*, units receives the same chance of coming into the sample as any other set of the same size.

Different units (which corresponds to *s*=1) all receive the same chance, the chance being *n/N*. Other designs depart from simple random sampling by:

(i) suppressing some of the all possible samples noted above, and/or

(ii) by giving different units (and hence different samples) different probabilities of selection.

A Numerical Example

To illustrate some basic ideas, let us consider a very small sample to be drawn from a very small population.

Suppose the population consists of 8 units (*N*=8), from which a sample of *n*=4 units is selected with simple random sampling. Let us also assume that for the unit number *j*, from 1 to 8, the value Y_j of variable of interest equals *j* itself. That is, we have a population of 8 units with values as shown below.

Table 1. The assumed population

Unit (*j*)	1	2	3	4	5	6	7	8	mean	σ^2	S^2
Value Y_j	1	2	3	4	5	6	7	8	4.5	5.25	6.0

Two important properties of the distribution of Y_j values are the *mean*

$$\bar{Y} = \sum_j Y_j / N = 4.5 \quad (1)$$

and *population variance*, which is a measure of the variability among the Y_j values

$$\sigma^2 = Var(Y_j) = \sum_j (Y_j - \bar{Y})^2 / N = 5.25 \quad (2)$$

or a slightly different version used in sampling theory

$$S^2 = \frac{N}{N-1} \cdot \sigma^2 = \sum_j (Y_j - \bar{Y})^2 / (N-1) = 6.0 \quad (3)$$

With simple random sampling it can be seen that there are in this example

$$\frac{8!}{(8-4)! \cdot 4!} = 70$$

possible samples of distinct units, such as samples with units

(1,2,3,4), (1,2,3,5), (1,2,3,6), ... (1,2,4,5), (1,2,4,6), ... etc.

Each sample *s* appears with the same probability p_s=1/70. For each particular sample we can compute the mean of its *n* values

$$\bar{y}_s = \frac{1}{n} \cdot \sum_{i \in s} y_i \,|sample\ s. \qquad (4)$$

Table 2 represents the frequency distribution of the means of all the 70 possible samples.

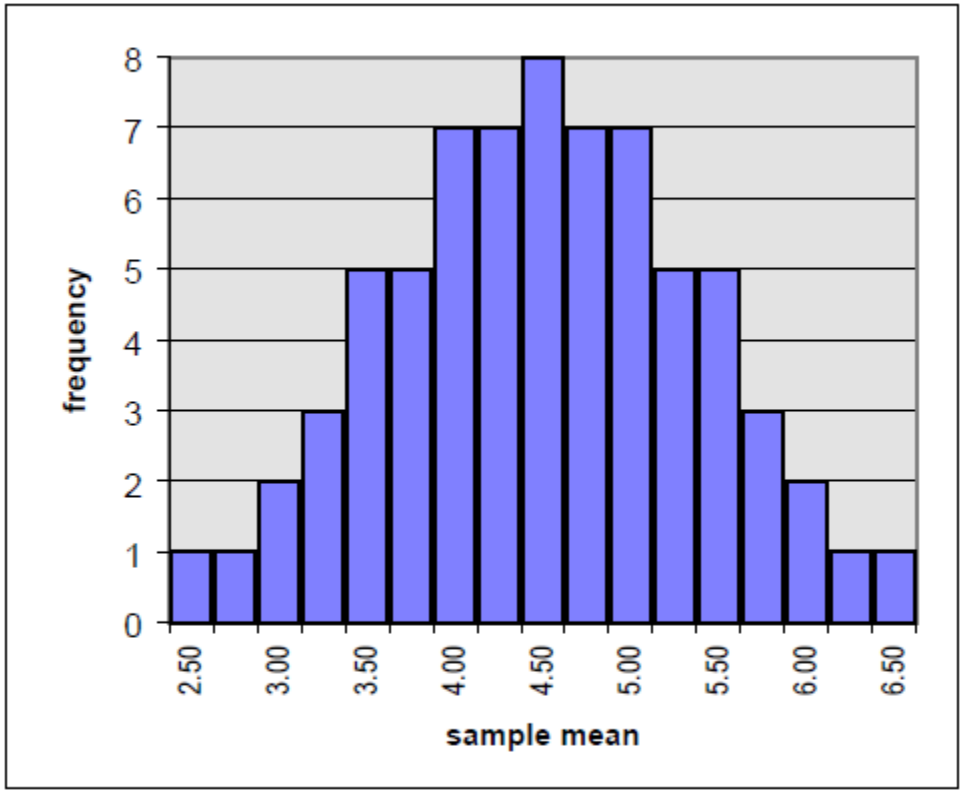
The average of all the 70 possible samples is called the *expected value* of the sampling distribution, and equals

$$E(\bar{y}) = \sum f_s \cdot \bar{y}_s / \sum f_s = 4.5,$$

which is the same as the actual population mean. We say that *the sample provides an unbiased estimator of the population mean*. Here f_s is the frequency (number of samples) with mean $\bar{y}_s$.

Table 2. Sampling distribution of the mean for simple random samples of size n=4, drawn from the population of size N=8 from Table 1.

Mean	frequency
2.50	1
2.75	1
3.00	2
3.25	3
3.50	5
3.75	5
4.00	7
4.25	7
4.50	8
4.75	7
5.00	7
5.25	5
5.50	5
5.75	3
6.00	2
6.25	1
6.50	1



$$E(\bar{y}) = 4.5; \quad Var(\bar{y}) = E\left[\bar{y} - E(\bar{y})\right]^2 = 0.75.$$

How variable are the results from one sample to another?

This is measured by *sampling variance* (i.e. variance of the sampling distribution), defined as

$$Var(\bar{y}) = E(\bar{y} - E(\bar{y}))^2, \qquad (5)$$

that is, the expected (average) value of the squared deviations of sample means from their value averaged over all samples.

In our example we can write

$$Var(\bar{y}) = \frac{1}{70} \cdot \sum_{s=1}^{70} (\bar{y}_s - 4.5)^2,$$

or in terms of the frequency distribution found above as

$$Var(\bar{y}) = \sum \; f_s \cdot (\bar{y}_s - 4.5)^2 / \sum f_s = 0.75.$$

In a simple random sample, this sampling variance is in fact related in a straightforward way to population variance and sample size:

$$Var(\bar{y}) = (\frac{N-n}{N-1}) \cdot \frac{\sigma^2}{n},$$

or more commonly written as

$$Var(\bar{y}) = (1-f) \cdot \frac{S^2}{n} \qquad (6)$$

where $f=n/N$ is the sampling rate, and S^2 is a slightly modified definition of population variance (equation 3), introduced because this form is more convenient in discussing sampling theory.

In our example, f=4/8, n=4, and S^2=6.0, giving

$$Var(\bar{y}) = (1-0.5) \cdot \frac{6.0}{4} = 0.75 \ ,$$

which is exactly as computed above from the full sampling distribution.

Population variance

More diverse the units in the population, larger in direct proportion will be the variability between estimates from samples of a given size.
It stands to reason that for a homogeneous population a small sample size may suffice, but for a heterogeneous population a large number of observations would be required to have the same confidence in the results.
Let us consider a few examples of differences in population variances.

In a population of farms, farms of different sizes may differ greatly in their production (large population variance for variables related to production), but much less so in their yields i.e. production per unit of area.
Similarly, in a population of economic establishments there may be great differences among the establishments in terms of total output, but much smaller differences in productivity i.e. in output per worker.
In either case, S^2 is likely to be smaller if the population is confined to a particular category of units, such as farms of a particular type or establishments in a particular sector of the economy.

Similar considerations apply to other populations in which units differ greatly in size or type.

In surveys of households, we often find that disparities in income are wider than those in household expenditure, and disparities in the latter in turn wider than those in expenditure on food. This is because, compared to richer people, poorer people often spend a greater part of their income (thus reducing differences in the amounts actually spent), and a grater share of that spending is on food.

For any of these variables, population variance is likely to be smaller if the statistic is considered on a per capita rather than per household basis, or if the population is restricted to households of a particular size or composition.

Estimating population variance

In practice, we need not only to estimate population parameters of interest (e.g. population mean $\bar{Y}$ from sample mean $\bar{y}$), but also to provide an estimate of the *uncertainty* to which this estimate is subject i.e. an *estimate of Var($\bar{y}$).*

How can the results of the one sample that is available be used to estimate the variability among results from all the possible samples which could have been drawn with the given design?

Procedures exist for doing this with complex samples, but let us consider for the moment an SRS.
From equation (6), its variance is given by estimating S^2. Here is an important result of sampling theory: in simple random sampling an unbiased estimator of the population parameter S^2 (Eq. 3) is provided by the sample statistic

$$s^2 = \sum_i (y_i - \bar{y})^2/(n-1). \qquad (7)$$

That is, s^2 averaged over all possible samples equals S^2: $E(s^2)=S^2$, just as for this design $E(\bar{y}) = \bar{Y}$.

Consequently, an unbiased estimate for SRS sampling variance, Var($\bar{y}$), (Eq. 6), is provided by

$$Var(\bar{y}) = (1-f) \cdot \frac{s^2}{n}. \qquad (8)$$

S^2 for proportions

For a proportion the expressions for S^2 and s^2 reduce to a very simple form and no elaborate computation is required. In general terms, a proportion implies that the population (of size *N*, say) is divided into only two categories: number N.P "yes's" and N.(1-P) "no's". We can assign the value Y_j=1 (say) to each of the former, and value Y_j=0 to each of the latter. The proportion of "yes's" is then simply the mean of these values:

$$\bar{Y} = \frac{N{\cdot}P\ ones + N{\cdot}(1-P)zeros}{N\ total\ cases} = \frac{N{\cdot}P}{N} = P,$$

and the sum of squared deviations becomes

$\sum_{j=1}^{N}(Y_j - \bar{Y})^2 = (1-P)^2$ for N.P cases with $Y_j = 1$

$+(-P)^2$ for N.(1-P) cases with $Y_j = 0$

$=N.P.(1-P)^2+N.(1-P).(-P)^2 = N.P.(1-P)$

This gives

$$S^2 = \frac{N}{N-1} \cdot P(1-P) \approx P(1-P)$$

and similarly for its estimator

$s^2 = \frac{n}{n-1} \cdot p(1-p) \approx p(1-p)$, if n is not small.

Departure from Simple Random Sampling

The sample design may depart from simple random sampling in a number of ways, the three common and important ones being the following:

- *Clustering* or multi-stage sampling, i.e. group the population elements into larger units ('clusters')
- *Stratification,* i.e. partitioning the population before sample selection.
- *Unequal selection probabilities.*

Numerical illustrations of the effect of clustering, stratification and weighting on the variance of sample estimates.

Consider again the example used for SRS on a population of 8 elements.

In fact with the SRS design, all combination of any given number s of elements, $1 \leq s \leq n$, are equally likely to appear in the sample. What clustering or stratification does is to suppress some of these possible samples, i.e. prevent particular combinations of elements from being selected.

For instance, we may divide the population into two strata, the first consisting of units (1-4) and the second of units (5-8), and select two units from each part separately. Hence samples made up of units such as (1,2,3,6) or (2,5,6,8) which contain more than two units from any part, while possible under SRS design, are not allowed under this stratification.

Similarly, the units may be clustered into groups such as (1,2),....,(7,8), and the selection done such that either no units or both units from any group appear in the sample. Hence samples like (1,3,4,6) or (1,2,6,8) which contain only one unit from any group, while possible under SRS design, are not allowed under this clustered design.

Of course, we can also have the 'restrictions' of stratification and clustering applied simultaneously.

Example of Clustering

Suppose that our 8 units are geographically located such that the two units in each pair (1+2), (3+4), (5+6) and (7+8) are close to each other.

If the survey data are to be collected by actually visiting each sample unit, it may be cheaper and more convenient to select the sample such that if one unit in a pair is selected, then the other unit is also taken into the sample automatically. In other words, a sample of size 4 elements is obtained by selecting 2 pre-defined clusters, each cluster containing 2 elements.

In this way the amount of travel required for data collection may be reduced, and the data collection process better controlled. This is the positive side of using a clustered sample design.

But what happens to the efficiency of the sample?

For this we need to consider the sampling distribution of the design.

The design is simply to select two of the four pairs listed above: pairs of original elements form our new 'units' for simple random sampling. There are only 6 possible samples. The remaining 70 - 6=64 samples, of the possible 70 samples with a simple random sampling design, have been suppressed, but without affecting the probability nature of the sample.

Table below shows the distribution of the sample means.

Their overall average is the same as the population mean i.e., with this design as well, the sample mean provides an *unbiased* estimator of the population mean.

However, compared to a SRS of elements, *sampling variance is more than doubled* (increased by a factor 1.67/0.75=2.2).

Depending upon the practical circumstances in which the survey is conducted, this loss in efficiency may be more than compensated by the increased cost-efficiency, convenience, and possible improvement in the data quality; in which case, it is better to opt for the clustered design. Or the compensation may be inadequate; in which case it would have been better to stick with simple random sampling of elements.

Sample s	mean $\bar{y}$
(1+2), (3+4)	2.5
(1+2), (5+6)	3.5
(1+2), (7+8)	4.5
(3+4), (5+6)	4.5
(3+4), (7+8)	5.5
(5+6), (7+8)	6.5
all six samples	4.5

Expected value, i.e. average of all samples

$$E(\bar{y}) = \frac{1}{6} \cdot \sum_{s=1}^{6} \bar{y}_s = 4.5 = \text{ population mean, } \bar{Y}$$

$$Var(\bar{y}) = \frac{1}{6} \cdot \sum_{s=1}^{6} \left(\bar{y}_s - 4.5 \right)^2 = 10/6 = 1.67$$

$$Deft^2 = Var(\bar{y}) / Var_0(\bar{y}) = 1.67/0.75 = 2.2$$

$$Deft = 1.5$$

Example of Stratification

Suppose that we have prior knowledge that units in the first half of the population (units j=1-4) tend to have smaller Y_j values than those in the second half (j=5-8). On the basis of this knowledge, we may divide the population into two parts, units (1,2,3,4) and units (5,6,7,8), and select a SRS of two units from each part separately. This is stratification.

First consider element sampling (i.e. without clustering) within each stratum separately.

Within each stratum we have six possible samples, giving a total of 6X6=36 full samples since each of the 6 in one part can be combined with any of the 6 in the second part. The remaining 70 - 36=34 samples, possible under unstratified simple random sampling, have been suppressed by the stratified design. We can list all the 36 samples, but it is sufficient to consider the 6 samples from each of the two strata. The overall mean is the average of the two strata means. The overall variance is half their average variance because the sample size is doubled when we put the two strata together. (Simple averaging suffices in the present example because the two strata are equal in the population and sample sizes.) The results are shown in Table below.

The population mean (4.5) is again estimated without bias. Variance is reduced with stratification by a factor 0.21/0.75=0.28 compared to SRS.

This is because the units within each stratum are homogeneous (all small or all large).

In determining the variability of the samples formed by pooling together the two strata, what matters is the variability within each stratum. With homogenous strata we are ensuring that always some small and some large units are selected, which tend to balance each other and the resulting overall sample values tend to move closer to the population values.

With such stratification we suppress statistically ‘bad’ samples, i.e. samples with too many small or too many large units.

Stratum of smaller units		Stratum of larger units	
Sample	$\bar{y}$	Sample	$\bar{y}$
1,2	1.5	5,6	5.5
1,3	2.0	5,7	6.0
1,4	2.5	5,8	6.5
2,3	2.5	6,7	6.5
2,4	3.0	6,8	7.0
3,4	3.5	7,8	7.5
average, $E(\bar{y})$	2.5		6.5
$Var(\bar{y}) = \frac{1}{6}\cdot\sum_{s=1}^{6}\left(\bar{y}_s - E(\bar{y})\right)^2$	0.42		0.42

$E(\bar{y})$, overall mean =(2.5+6.5)/2=4.5=$\bar{Y}$; $Var(\bar{y})$, overall variance =0.42/2=0.21.

$Deft^2 = Var(\bar{y})/Var_0(\bar{y}) = 0.21/0.75 = 0.28$

$Deft = 0.53.$

EFFECT OF CLUSTERING ON VARIANCE

To illustrate some basic ideas, we consider the simple case of a single stage random sample of clusters of equal size.

In place of selecting a simple random sample (srs) of n elements, the population of N elements is assumed as divided into A clusters each of size B=N/A, from which a=n/B clusters are selected with simple random sampling. All B elements in each selected cluster are taken into the sample, giving a sample of n=a.B elements.

How does the efficiency of this sample compare with that of a sample where the same number of elements have been selected with simple random sampling?

With y_{ij} as the value of a certain variable for element j in cluster i, the cluster mean and the overall mean per element are given by

$$\bar{y}_i = \sum_j y_{ij}/B \;\; ; \;\; \bar{y} \;\; = \sum_{ij} y_{ij}/a.B = \sum_j \bar{y}_i/a$$

What we have in fact is a srs of a cluster means, out of a population of A means. Hence the sample mean is an unbiased estimator of the population mean, and its variance is given by

$$S_a^2 = \frac{\sum_i (\bar{Y}_i - \bar{Y}\;\;)^2}{(A-1)}; \; Var(\bar{Y}) = (1 - \frac{a}{A}) \cdot \frac{S_a^2}{a}$$

estimated by the sample values

$$s_a^2 = \frac{\sum_i (\bar{y}_i - \bar{y}\;\;)^2}{(a-1)}; \; Var(\bar{y}) = (1 - \frac{a}{A}) \cdot \frac{s_a^2}{a}$$

EFFECT OF STRATIFICATION ON VARIANCE

If sample selection and estimation is done separately within each stratum, the same basic expressions such as equations (1)-(8) above apply to each stratum. Using subscript h to refer to a particular stratum, we have with SRS within strata:

$$Var(\bar{y}_h) = (1 - f_h) \cdot \frac{S_h^2}{n_h}, \text{ with } S_h^2 = \frac{\sum(Y_{hj} - \bar{Y}_h)^2}{N_h - 1} = \frac{N_h}{N_h - 1} \cdot \sigma_h^2$$

summed over N_h units in the stratum h, and estimated by

$s_h^2 = \frac{\sum(y_{hj} - \bar{y}_h)^2}{n_h - 1}$ summed over n_h units in the sample from h.

In putting together the results from different strata, we often do that in proportion to stratum size, e.g. $W_h = N_h/N$. For the total population $\bar{Y} = \sum_h W_h \cdot \bar{Y}_h$ and if the W_h are known, $\bar{y} = \sum_h W_h \cdot \bar{y}_h$ and $Var(\bar{y}) = \sum_h W_h^2 \cdot Var(\bar{y}_h)$

The Horvitz-Thompson Estimator

Under unequal probability sampling, the Horvitz-Thompson estimator (HT estimator) is an unbiased estimator of the population total. It is defined as

$$\hat{y}_{HT} = \sum_{i \in s} \frac{y_i}{\pi_i} = \sum_{i \in s} w_i y_i$$

where w_i is the design weight of the ith element as defined above. The HT estimator of the population mean can be expressed as

$$\hat{\bar{y}}_{HT} = \frac{1}{\hat{N}} \sum_{i \in s} \frac{y_i}{\pi_i} = \frac{1}{\hat{N}} \sum_{i \in s} w_i y_i$$

Where $\hat{N} = \sum_{i \in s} w_i$ is the estimated population size.

$$Var(\hat{y}_{HT}) = \sum_{i \in s} \frac{y^2{}_i}{\pi_i}(1 - \pi_i) + \sum_{i \in s} \sum_{j \neq i} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} (\pi_{ij} - \pi_i \pi_j)$$

and an unbiased estimator of this variance is

$$var(\hat{y}_{HT}) = \sum_{i \in s} \frac{y^2{}_i}{\pi^2{}_i}(1 - \pi_i) + \sum_{i \in s} \sum_{j \neq i} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_{ij}}$$

For a fixed-size sampling design the variance is just

$$Var(\hat{y}_{HT}) = \frac{1}{2}\sum_{i\in s}\sum_{j\neq i}(\pi_i\pi_j - \pi_{ij})\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2$$

Under SRS without replacement the Horvitz-Thompson estimator for population total Y is:

$$\hat{Y}_\pi = \frac{N}{n}\sum_{i\in s} y_i$$

The unbiased variance estimator of $Var(\hat{Y}_\pi)$ is given by

$$Var(\hat{Y}_\pi) = N^2(1-f)\frac{s^2}{N}$$

Example1.

Consider a very small population U consisting of N = 3 elements

y1 = 70 y2 = 60 y3 = 80

and we conduct a simple random sampling without replacement of size n = 2.

s	P(s)	$\bar{y}(s)$
{1,2}	$\frac{1}{3}$	65
{1,3}	$\frac{1}{3}$	75
{2,2}	$\frac{1}{3}$	70

The inclusion probability $P(1 \in s) = \frac{1}{3} + \frac{1}{3} = \frac{2}{3} = \frac{n}{N}$

$$E(\bar{y}) = \sum_{\forall s} P(s)\bar{y}(s) = \frac{1}{3} \cdot 65 + \frac{1}{3} \cdot 75 + \frac{1}{3} \cdot 70 = 70$$

$$Var(\bar{y}) = E(\bar{y}^2) - [E(\bar{y})]^2 = \frac{1}{3} \cdot 65^2 + \frac{1}{3} \cdot 75^2 + \frac{1}{3} \cdot 70^2 - 70^2 = \frac{50}{3}$$

Example2.

Lets consider now a population $U = \{1,2,3,4\}$ of dimension N=4. If we consider SRS without replacement of dimension n=3, $S = \{(1,2,3), (1,2,4), (1,3,4), (2,3,4)\}$. The sampling probability of each sample are the following:

s	p(s)
(1,2,3)	0.15
(1,2,4)	0.35
(1,3,4)	0.30
(2,3,4)	0.20
	1

First order inclusion probabilities for each unit are:

i	1	2	3	4
π_i	0.80	0.70	0.65	0.85

Suppose that the interest variable assume the following values:

i	Y_i	π_i
1	50	0.80
2	32	0.70
3	48	0.65
4	65	0.85

$$\bar{Y} = \frac{50 + 32 + 48 + 65}{4} = 48.75$$

For the first sample $= \{1,2,3\}$, the units are (50, 32, 48), so

$$\hat{\bar{Y}} = \frac{1}{4}\left(\frac{50}{0.80} + \frac{32}{0.70} + \frac{48}{0.65}\right) = 45.51$$

The sample mean $\bar{y} = \frac{50+32+48}{3} = 43.33$

For all the 4 sample we have

s	Y_i	$\hat{\bar{Y}}$	$\bar{y}$	P(s)
(1,2,3)	(50,32,48)	45.51	43.33	0.15
(1,2,4)	(50,32,65)	46.17	49.00	0.35
(2,3,4)	(32,48,65)	49.01	48.33	0.20
(1,3,4)	(50,48,65)	53.20	54.33	0.30

E($\hat{\bar{Y}}$)=45.51x0.15+46.17x0.35+49.01x0.20+53.20x0.30=48.75

V($\hat{\bar{Y}}$)=MSE($\hat{\bar{Y}}$)=(45.51-48.75)^{2}x0.15+…+(53.20-48.75)^{2}x0.30=9.85

E($\bar{y}$)=43.33x0.15+49.00x0.35+48.33x0.20+54.33x0.30=49.62≠48.75

MSE($\bar{y}$)=(43.33-48.75)^{2}x0.15+…+(54.33-48.75)^{2}x0.30=13.81

PRACTICAL PROCEDURES FOR COMPUTING SAMPLING ERRORS

Large scale household surveys are generally based on multi-stage, stratified and otherwise complex designs. A typical survey is multi-purpose in several respects: it involves many types of interrelated variables; many types of estimates such as proportions, means, ratios and differences of ratios; various types of units of analysis such as households and individuals; various levels of dis-aggregation of the sample; and diverse and numerous subclasses (subpopulations) for which estimates of levels, differences and other relationships are required. Practical procedures for estimating sampling errors therefore:

(1) must take into account the actual, complex structure of the design;
(2) should be flexible enough to be applicable to diverse designs;
(3) should be suitable and convenient for large-scale application, and for producing results for diverse statistics and subclasses;
(4) should be robust against departure of the design in practice from the ideal `model' assumed in the computation method;
(5) should have desirable statistical properties such as small mean-squared error of the variance estimator;
(6) should be economical in terms of the effort and cost involved; and
(7) suitable computer software should be available for application of the method.

The theory of ‘simple replicated variance estimators’

The theory of 'simple replicated variance estimators' provides the basis for most practical approaches to variance estimation, though in application to complex situations, additional assumptions and approximations may be involved.

The basic theory may be stated as follows. Suppose that y_j are a set of random uncorrelated variables with a common expectation Y. Then the mean $\tilde{y}$ of n values y_j $\tilde{y} = \sum_j y_j / n$ has an expected value equal to Y, and its variance is given by $\mathrm{var}(\tilde{y}) = s^2/n$, where $s^2 = \sum_j (y_j - \tilde{y})^2/(n-1)$.

The most obvious example of the above is a simple random sample (srs) of elements selected with replacement, where yj represent values of a certain variable for individual elements j.

The same idea can be applied to the more general situation when "j" refers not to individual elements but to any set of elements uncorrelated to others in the sample, and " y_j " to any complex statistic defined for each set j.

The requirement is that the y_j are uncorrelated and have a common expectation.

In practice this means that the sets should be selected and observed independently, following the same selection, measurement and estimation procedures.

Drawing on this basic idea, two broad practical approaches to the computation of sampling errors may be identified:

1) Computation from comparisons among estimates for replications of the sample, each of which reflects the structure of the full sample, including its clustering and stratification.

2) Computation from comparisons among certain aggregates for primary selections or replicates within each stratum of the sample, also known as linearization method.

The first method is simpler and computationally faster; it is normally used when applicable. There can, however be more complex situations – more complex sample designs, more complex statistics – which may require the second type of method, comparison among sample replications.

The Jack-knife Repeated Replication is a commonly used method which belongs to class (1). This is the method adopted and developed for application in EU-SILC at the EU level and also in countries that choose to use it.

The idea of replication techniques

JRR is one of the class of practical methods for variance estimation in complex samples based on *measures of observed variability among replications of the full sample*. The basic requirement is that the full sample is composed of a number of subsamples or replications, each with the same design and reflecting complexity of the full sample, enumerated using the same procedures

A replication differs from the full sample only in size. But its own size should be large enough for it to reflect the structure of the full sample, and for any estimate based on a single replication to be close to the corresponding estimate based on the full sample. At the same time, the number of replications available should be large enough so that comparison among replications gives a stable estimate of the sampling variability in practice.

With JRR, a replication is formed by dropping a small part of the total sample, such as a single PSU in one stratum; consequently each replication measures the contribution of a small part such as a single stratum.

The various re-sampling procedures available differ in the manner in which replications are generated from the parent sample and the corresponding variance estimation formulae evoked (such as the Balanced Repeated Replication (BRR) and the bootstrap, apart from JRR).

Initial formulae of JRR

Briefly, the standard JRR involves the following.

Let z be a full-sample estimate of any complexity, and $z_{(hi)}$ be the estimate produced using the same procedure after eliminating primary unit i in stratum h and increasing the weight of the remaining (a_h-1) units in the stratum by an appropriate factor g_h (see below). Let $z_{(h)}$ be the simple average of the $z_{(hi)}$ over the a_h sample units in h. The variance of z is then estimated as:

$$\mathrm{var}(z) = \Sigma_h \left[(1 - f_h) . g_h . \Sigma_i \left(z_{(hi)} - z_{(h)} \right)^2 \right].$$

Improvements of initial formulae

Originally, the factor g_h was taken as $g_h = a_h/(a_h - 1)$, recently, it has been proposed to use $g_h = w_h/(w_h - w_{hi})$, where $w_h = \Sigma_i w_{hi}, w_{hi} = \Sigma_j w_{hij}$, the sum of sample weights of ultimate units j in primary selection i. The latter form retains the total weight of the included sample cases unchanged across the replications created. With the sample weights scaled such that their sum is equal (or proportional) to some external more reliable population total, population aggregates from the sample can be estimated more efficiently, often with the same precision as proportions or means.

The design effect – deft

The design effect is defined as the ratio of the variance under the given sample design, to the variance under a simple random sample of the same size: $se = se_R \cdot deft$

Proceeding from estimates of sampling error to estimates of design effects (ratio of actual sampling error to that under equivalent simple random sampling, SRS) is essential for understanding the patterns of variation in and the determinants of magnitude of the error, for smoothing and extrapolating the results for diverse statistics and population subclasses, and for evaluating the performance of the sampling design.

Components of the design effect

Computing design effects requires the additional step of estimating sampling errors under simple random sampling.

Design effect itself can be decomposed into at least two components.

(1) the effect of sample weights on variance

(2) the effect of clustering, stratification and aspects other than weighting

The first component of the design effect (known as the *Kish effect*, or effect of the weights) could be directly calculated by the data.

For the second component, it is required to 'randomise' the data set and apply the JRR procedure, as follows:

$$se_{JRR} = se_{SRS} \cdot deft_1 \cdot deft_2$$

$$se_{JRR_{RAND}} = se_{SRS} \cdot deft_1$$

$$deft_2 = \frac{se_{JRR}}{se_{JRR_{RAND}}} = \frac{se_{SRS} \cdot deft_1 \cdot deft_2}{se_{SRS} \cdot deft_1}$$

Factor (deft1) does not depend on the structure of the sample, other than the presence of unequal sample weights for the elementary units of analysis. The main effect is the variability of these weights in the sample. The effect is also influenced by the extent to which the variable being estimated is correlated with the sample weights.

The weights are introduced in the sample to let the estimates be statistically unbiased. They are introduced in successive steps:

- Inversely to probability of selection
- For non response
- For the so-called ‘calibration’
- Other steps are present in longitudinal surveys such as SILC
- To reduce the effect of weighting (reduce deft1) it is applied the so-called ‘trimming’

Practical aspects

In order to apply the JRR technique (and any other resampling technique) it is important to clarify two practical aspects:

- Explicit and implicit stratification and computational strata
- Computational PSU (Primary Selection Units)

In many practical situations some aspects of sample structure need to be redefined to make variance computation possible, efficient and stable. Of course, any such redefinition is appropriate only if it does not introduce significant bias in variance estimation. The computational structure can differ from the actual sample structure because of various consideration such as the following.

Firstly, it is often necessary to define *computational strata and PSUs* to meet the basic requirement of practical methods of variance estimation for complex samples. Here are some common situations.

1. It may be necessary to regroup ('collapse') strata so as to ensure that each stratum has at least two sample PSUs – the minimum number required for the computation of variance.

2. Units which are included into the sample automatically ('self-representing units') are in fact strata rather than PSUs, and computational PSUs have to be defined at a lower stage within each such unit.

3. In samples selected systematically, the implied implicit stratification is often used to define explicit strata, from each of which an independent sample is supposed to have been selected. Such strata have to be formed by pairing or otherwise grouping of PSUs in the order of their selection from the systematic list, ensuring that each resulting computational stratum has at least two primary selections.

4. Sometimes non-response can result in the disappearance from the sample of whole PSUs. This can disturb the structure of the sample, such as leaving fewer than two PSUs in some strata. Variance computation requires some redefinition of the computational units to meet the basic requirement of having at least 2 PSUs per stratum.

5. The above-mentioned problem arises more frequently and seriously when computing sampling errors for subclasses (subpopulations). The risk can be reduced by aggregating PSUs and strata to create fewer, larger computational units.

Considerations such as the above apply equally irrespective of whether the JRR, Linearisation or some other form of variance computation algorithm is used.

Concluding Remark: Importance of information on sampling errors

While survey data are subject to errors from diverse sources, information on sampling errors is of crucial importance in proper interpretation of the survey results, and in rational design of sample surveys.

Of course, sampling error is only one component of the total error in survey estimates, and not always the most important component.

By the same token, it is the lower (and more easily estimated) bound of the total error: a survey will be useless if this component alone becomes too large for the survey results to add useful information with any measure of confidence to what is already known prior to the survey.

Furthermore, survey estimates are typically required not only for the whole population but also separately for many subgroups in the population.

Generally the relative magnitude of sampling error vis a vis other types of errors increases as we move from estimates for the total population to estimates for individual subgroups and comparison between subgroups.

Information on the magnitude of sampling errors is therefore essential in deciding the degree of detail with which the survey data may be meaningfully tabulated and analysed.

Similarly, sampling error information is needed for sample design and evaluation.

While the design is also determined by many other considerations (such as costs, availability of sampling frames, the need to control measurement errors), rational decisions on the choice of sample size, allocation, clustering, stratification, estimation procedures etc. can only be made on the basis of detailed knowledge of their effect on the magnitude of sampling errors of statistics obtained from the survey.

Various practical methods and computer software have been developed for computing sampling errors, and there is no justification in most situations for the continued failure to include information on sampling errors in the presentation of survey results.

References:

Verma Vijay, (1991). *Sampling Methods: Training Handbook*. Tokyo: Statistical Institute for Asia and the Pacific (SIAP).