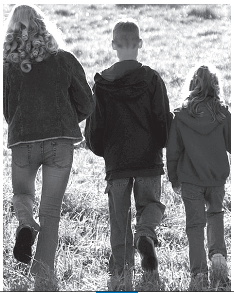


4

Replicate Weights

PISA DATA ANALYSIS MANUAL: SPSS® SECOND EDITION – ISBN 978-92-64-05626-8 – © OECD 2009

INTRODUCTION

In most cases, as mentioned in Chapter 3, national and international surveys collect data from a sample instead of conducting a full census. However, for a particular population, there are thousands, if not millions of possible samples, and each of them does not necessarily yield the same estimates of population statistics. Every generalisation made from a sample, *i.e.* every estimate of a population statistic, has an associated uncertainty or risk of error. The sampling variance corresponds to the measure of this uncertainty due to sampling.

This chapter explains the statistical procedures used for computing the sampling variance and its square root, the standard error. More specifically, this chapter discusses how to estimate sampling variances for population estimates derived from a complex sample design using replicate weights. First, the concept of sampling variance is examined through a fictitious example for simple random sampling. Second, the computation of the standard error is investigated for two-stage sampling. Third, replication methods for estimating sampling variances are introduced for simple random samples and for two-stage samples.

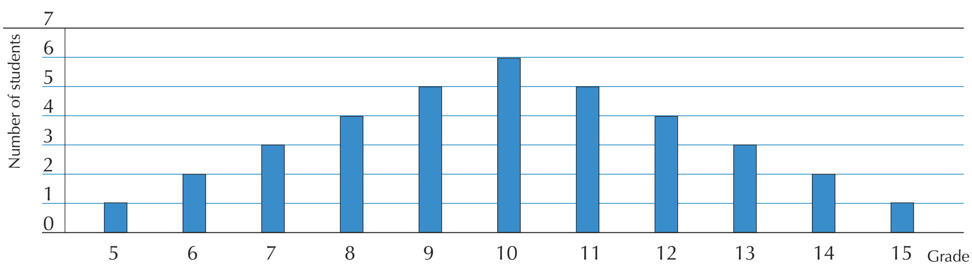
SAMPLING VARIANCE FOR SIMPLE RANDOM SAMPLING

Suppose that a teacher decides to implement a mastery learning approach in his or her classroom. This methodology requires that each lesson be followed by a student assessment. In the example given, the teacher's class has 36 students. The teacher quickly realises that it would be too time-consuming to grade all assessments and therefore decides to select a sample of tests to find out whether the material taught has been assimilated (Bloom, 1979).

However, the random sampling of a few tests can result in the selection of high achievers or low achievers only, which would introduce an important error in the class mean performance estimate. These situations are extreme examples, but drawing a random sample will always generate some uncertainty.

In the same example, before selecting some tests, the teacher grades all of them and analyses the results for the first lesson. Figure 4.1 presents the distribution of the 36 students' results. One student gets a grade 5, two students get a grade 6, and so on.

Figure 4.1
Distribution of the results of 36 students



The distribution of the student grades corresponds to a normal distribution. The population mean and the population variance are respectively equal to:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{(5+6+6+7+\ldots+14+14+15)}{36} = \frac{360}{36} = 10$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \mu)^2 = \frac{\left[(5-10)^2 + (6-10)^2 + \ldots + (14-10)^2 + (15-10)^2\right]}{36} = \frac{240}{36} = 5.8333$$

Table 4.1
Description of the 630 possible samples of 2 students selected from 36 students, according to their mean

Sample mean	Results of the two sampled students	Number of combinations of the two results	Number of samples
5.5	5 and 6	2	2
6.0	6 and 6 5 and 7	1 3	4
6.5	5 and 8 6 and 7	4 6	10
7.0	7 and 7 5 and 9 6 and 8	3 5 8	16
7.5	5 and 10 6 and 9 7 and 8	6 10 12	28
8.0	8 and 8 5 and 11 6 and 10 7 and 9	6 5 12 15	38
8.5	5 and 12 6 and 11 7 and 10 8 and 9	4 10 18 20	52
9.0	9 and 9 5 and 13 6 and 12 7 and 11 8 and 10	10 3 8 15 24	60
9.5	5 and 14 6 and 13 7 and 12 8 and 11 9 and 10	2 6 12 20 30	70
10.0	10 and 10 5 and 15 6 and 14 7 and 13 8 and 12 9 and 11	15 1 4 9 16 25	70
10.5	6 and 15 7 and 14 8 and 13 9 and 12 10 and 11	2 6 12 20 30	70
11.0	7 and 15 8 and 14 9 and 13 10 and 12 11 and 11	3 8 15 24 10	60
11.5	8 and 15 9 and 14 10 and 13 11 and 12	4 10 18 20	52
12.0	9 and 15 10 and 14 11 and 13 12 and 12	5 12 15 6	38
12.5	10 and 15 11 and 14 12 and 13	6 10 12	28
13.0	11 and 15 12 and 14 13 and 13	5 8 2	16
13.5	12 and 15 13 and 14	4 6	10
14.0	13 and 15 14 and 14	3 1	4
14.5	14 and 15	2	2
			630

The standard deviation is therefore equal to:

$$\sigma = \sqrt{\sigma^2} = \sqrt{5.833} = 2.415$$

The teacher then decides to randomly select a sample of two students after the next lesson to save grading time. The number of possible samples of 2 students out of a population of 36 students is equal to:

$$C_{36}^{2} = \frac{36!}{(36-2)!2!} = 630$$

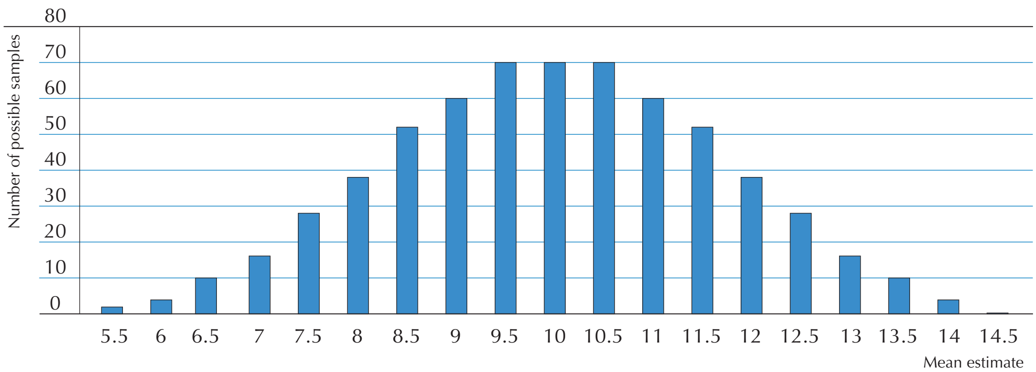
PISA DATA ANALYSIS MANUAL: SPSS® SECOND EDITION – ISBN 978-92-64-05626-8 – © OECD 2009

There are 630 possible samples of 2 students out of a population of 36 students. Table 4.1 describes these 630 possible samples. For instance, there are two possible samples which provide a mean estimate of 5.5 for student performance. These two samples are: *(i)* the student with a grade 5 and the first student with a grade 6; and *(ii)* the student with a 5 and the second student with a 6. Similarly, there are two ways of selecting a sample that would produce a mean grade of 6: the two sampled students both receive a grade 6 or one student receives a 5 and the second student receives a 7. As only two students obtained a grade 6 (4.1), there is only one possible sample with two grades 6. Since Figure 4.1 shows that there is only one student who received a grade 5 and three students who received a grade 7, there are three possible samples of two students with a grade 5 and a grade 7.

As shown in Table 4.1, there are 2 possible samples with a mean of 5.5, 4 possible samples with a mean of 6, 10 possible samples with a mean of 6.5, 16 possible samples with a mean of 7, and so on.

Figure 4.2 is a chart of the frequency of samples by their mean estimates for all possible samples of 2 students from 36.

Figure 4.2
Sampling variance distribution of the mean



As for all distributions, this distribution of the means of all possible samples can be summarised by central tendency indices and dispersion indices, such as the mean and the variance.

$$\mu_{(\hat{\mu})} = \left[(2\times 5.5)+(4\times 6)+(10\times 6.5)+(16\times 7)+(28\times 7.5)+(38\times 8)+.....+(2\times 14.5)\right]/630=10$$

The mean of all possible sample means is equal to the student population mean, *i.e.* 10. This result is not a coincidence, but a fundamental property of the mean of a simple random sample, *i.e.* the mean of the means of all possible samples is equal to the population mean. In more formal language, the sample mean is an unbiased estimate of the population mean. Stated differently, the expected value of the sample mean is equal to the population mean.

However, it should be noted that there is an important variation around this expectation. In the example considered, sample means range from 5.5 to 14.5. The variance of this distribution, usually denoted as the sampling variance of the mean, can be computed as:

$$\sigma^2_{(\hat{\mu})} = \left[(5.5-10)^2+(5.5-10)^2+(6-10)^2+\ldots+(14.5-10)^2+(14.5-10)^2\right]/630 = 2.833$$

Its square root, denoted as the standard error, is equal to:

$$\sigma_{(\hat{\mu})} = \sqrt{\sigma^2_{(\hat{\mu})}} = \sqrt{2.833} = 1.68$$

However, what information does the standard error of the mean give, or more specifically, what does the value 1.68 tell us? The distribution of the means of all possible samples follows approximately a normal distribution. Therefore, based on the mathematical properties of the normal distribution, it can be said that:

- 68.2% of all possible sample means fall between –1 standard error and +1 standard error around the mean; and
- 95.4% of all possible sample means fall between –2 standard errors and +2 standard errors.

Let us check the mathematical properties of the normal distribution on the sampling variance distribution of the mean. Remember that the mean of the sampling variance distribution is equal to 10 and its standard deviation, denoted by the term "standard error", is equal to 1.68.

How many samples have a mean between $\mu_{(\hat{\mu})} - \sigma_{(\hat{\mu})}$ and $\mu_{(\hat{\mu})} + \sigma_{(\hat{\mu})}$, *i.e.* between (10-1.68) and (10+1.68), or between 8.32 and 11.68?

Table 4.2 shows that there are 434 samples out of 630 with a mean comprised between 8.32 and 11.68; these represent 68.8% of all samples. It can also be demonstrated that the percentage of samples with means between $\mu_{(\hat{\mu})} - 2\sigma_{(\hat{\mu})}$ and $\mu_{(\hat{\mu})} + 2\sigma_{(\hat{\mu})}$, *i.e.* between 6.64 and 13.36 is equal to 94.9%.

Table 4.2
Distribution of all possible samples with a mean between 8.32 and 11.68

Sample mean	Number of samples	Percentage of samples	Cumulative % of sample
8.5	52	0.0825	0.0825
9.0	60	0.0952	0.1777
9.5	70	0.1111	0.2888
10.0	70	0.1111	0.4000
10.5	70	0.1111	0.5111
11.0	60	0.0952	0.6063
11.5	52	0.0825	0.6888
	434		

To estimate the standard error of the mean, the mean of all possible samples is computed. In reality though, only the mean of one sample is known. This, as will be shown, is enough to calculate an estimate of the sampling variance. It is therefore important to identify the factors responsible for the sampling variance from the one sample chosen.

The first determining factor is the size of the sample. If the teacher, in our example, decides to select four students instead of two, then the sampling distribution of the mean will range from 6 (the four lowest results being 5, 6, 6, and 7) to 14 (the four highest results being 13, 14, 14, and 15). Remember that the sampling distribution ranged from 5.5 to 14.5 with samples of two units. Increasing the sample size reduces the variance of the distribution.

There are 58 905 possible samples of 4 students out of a population of 36 students. Table 4.3 presents the distribution of all possible samples of 4 students for a population of 36 students. This distribution has a mean of 10 and a standard deviation, denoted standard error, of 1.155.

This proves that the size of the sample does not affect the expected value of the sample mean, but it does reduce the variance of the distribution of the sample means: the bigger the sample size, the lower the sampling variance of the mean.

Table 4.3
Distribution of the mean of all possible samples of 4 students out of a population of 36 students

Sample mean	Number of possible samples
6.00	3
6.25	10
6.50	33
6.75	74
7.00	159
7.25	292
7.50	510
7.75	804
8.00	1 213
8.25	1 700
8.50	2 288
8.75	2 896
9.00	3 531
9.25	4 082
9.50	4 553
9.75	4 830
10.00	4 949
10.25	4 830
10.50	4 553
10.75	4 082
11.00	3 531
11.25	2 896
11.50	2 288
11.75	1 700
12.00	1 213
12.25	804
12.50	510
12.75	292
13.00	159
13.25	74
13.50	33
13.75	10
14.00	3

The second factor that contributes to the sampling variance is the variance of the population itself. For example, if the results are reported out of a total score of 40 instead of 20, (*i.e.* the student results are all multiplied by two), then the mean of the student results is 20, the variance is 23.333 (*i.e.* four times 5.8333) and the standard deviation is equal to 4.83 (*i.e.* two times 2.415). The sampling variance from a sample of two students will be equal to 11.333 (*i.e.* four times 2.8333) and that the standard error of the mean will be equal to 3.3665 (*i.e.* two times 1.68).

The standard error of the mean is therefore proportional to the population variance. Based on these examples, it can be established that the sampling variance of the mean is equal to:

$$\sigma^2_{(\hat{\mu})} = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$$

and the standard error of the sample mean is equal to:

$$\sigma_{(\hat{\mu})} = \sqrt{\sigma^2_{(\hat{\mu})}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

where:

σ^2 = variance of the population,

σ = standard deviation of the population,

n = sample size,

N = population size.

Let's check this formula with the example of two students selected:

$$\sigma^2_{(\hat{\mu})} = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) = \frac{5.833}{2}\left(\frac{36-2}{36-1}\right) = 2.8333$$

As the size of the population increases, the ratio $\left(\frac{N-n}{N-1}\right)$ tends toward 1. In such cases, a close approximation of the sampling variance of the mean is given by:

$$\sigma^2_{(\hat{\mu})} = \frac{\sigma^2}{n}$$

However, in practice, the population variance is unknown and is estimated from a sample. The sampling variance estimate of the mean, just as a mean estimate, can vary depending on the sample. Therefore, being based on a sample, only an estimate of the sampling variance of the mean (or any other estimate) can be computed.

In the remainder of this manual, the concepts of sampling variance and estimations of the sampling variance will have the same symbol to simplify the text and the mathematical notations. That is, symbols depicting the estimates of sampling variance will not have a hat (^) to differentiate them from true values, but the fact that they are estimates is to be understood.

SAMPLING VARIANCE FOR TWO-STAGE SAMPLING

Education surveys and, more particularly, international surveys rarely sample students by simply selecting a random sample of students. Schools get selected first and, within each selected school, classes or students are randomly sampled.

One of the differences between simple random sampling and two-stage sampling is that for the latter, selected students attending the same school cannot be considered as independent observations. This is because students within a school will usually have more common characteristics than students from different educational institutions. For instance, they are offered the same school resources, may have the same teachers, and therefore are taught a common curriculum, and so on. Differences between students from different schools are also greater if different educational programmes are not available in all schools. For instance, one would expect to observe more differences between students from a vocational school and students from an academic school, than those that would be observed between students from two vocational schools.

Further, within a country, within subnational entities, and within cities, people tend to live in areas according to their financial resources. As children usually attend schools close to their homes, it is likely that students attending the same school come from similar socio-economic backgrounds.

A simple random sample of 4 000 students is thus likely to cover the diversity of the population better than a sample of 100 schools with 40 students observed within each school. It follows that the uncertainty associated with any population parameter estimate (*i.e.* standard error) will be greater for a two-stage sample than for a simple random sample of the same size.

The increase of the uncertainty due to the two-stage sample is directly proportional to the differences between the first-stage units, known as primary sampling units (PSUs), *i.e.* schools for education surveys. The consequences of this uncertainty are provided below for two extreme and fictitious situations:

- All students in the population are randomly assigned to schools. Therefore, there should not be any differences between schools. Randomly selecting 100 schools and then within the selected schools randomly drawing 40 students would be similar, from a statistical point of view, to directly randomly selecting 4 000 students as there are no differences between schools. The uncertainty associated with any population parameter estimate would be equal to the uncertainty obtained from a simple random sample of 4 000 students.

- All schools are different but within schools, all students are perfectly identical. Since within a particular school, all students are identical: observing only 1 student, or 40, would provide the same amount of information. Therefore, if 100 schools are selected and 40 students are observed per selected school, the effective sample size of this sample would be equal to 100. Therefore, the uncertainty associated with any population parameter estimate would be equal to the uncertainty obtained from a simple random sample of 100 students.

Of course, there is no education system in the world that can be identified with either of these extreme situations. Nevertheless, in some education systems, differences between schools appear to be very small, at least regarding the survey's measure, for example, of academic performance, while in some other educational systems, differences between schools can be quite substantial.

The academic performance of each student can be represented by a test score, or by the difference between his/her score and the country average score. In education research, it is common to split the difference between the student's score and the country average score into three parts: *(i)* the difference between the student's performance and the corresponding class mean; *(ii)* the difference between this class mean and the corresponding school mean; *(iii)* the difference between this school mean and the country mean. The first difference relates to the within-class variance (or the residual variance in terms of variance analysis). It indicates how much student scores can vary within a particular class. The second difference – the difference between the class mean and the school mean – is related to the between-classes-within-school variance. This difference reflects the range of differences between classes within schools. This between-classes-within-school variance might be substantial in educational institutions that offer both academic and vocational education. The third difference – the difference between the school average and the country average – is called the between-school variance. This difference indicates how much student performance varies among schools.

To obtain an estimate of these three components of the variance, it would be necessary to sample several schools, at least two classes per school and several students per class. PISA randomly selects 15-year-olds directly from student lists within the participating schools. Therefore, generally speaking, it is impossible to distinguish the between- and within-classes variances. PISA can only provide estimates of the between- and the within-school variances.

Table 4.4 provides the between-school and within-school variances on the mathematics scale for PISA 2003. In northern European countries, the between-school variance is very small compared to the within-school variance. In these countries, the student variance mainly lies at the within-school level. In terms of student achievement then, schools in such countries do not vary greatly. However, in Austria, Belgium, Germany and Hungary, for instance, more than 50% of differences in the student performance are accounted for at the school level. This means that the student performance differs substantially among schools. Therefore, the uncertainty associated with any population parameters will be larger for these countries when compared to the uncertainty for northern European countries, given a comparable sample size of schools and students.

As Kish (1987) noted:

> "Standard methods for statistical analysis have been developed on assumptions of simple random sampling. Assuming independence for individual elements (or observations) greatly facilitates the mathematics used for distribution theories of formulas for complex statistics. […] However, independent selection of elements is seldom realised in practice, because much research is actually and necessarily accomplished with complex sample designs. It is economical to select clusters that are natural grouping of elements, and these tend to be somewhat homogeneous for most characteristics. The assumptions may fail mildly or badly; hence standard statistical analysis tends to result in mild or bad underestimates in length of reported probability intervals. Overestimates are possible, but rare and mild."

Table 4.4
Between-school and within-school variances on the mathematics scale in PISA 2003

	Between-school variance	Within-school variance
AUS	1 919.11	7 169.09
AUT	5 296.65	4 299.71
BEL	7 328.47	5 738.33
CAN	1 261.58	6 250.12
CHE	3 092.60	6 198.65
CZE	4 972.45	4 557.50
DEU	6 206.92	4 498.70
DNK	1 109.45	7 357.14
ESP	1 476.85	6 081.74
FIN	336.24	6 664.98
FRA	3 822.62	4 536.22
GBR	1 881.09	6 338.25
GRC	3 387.52	5 991.75
HUN	5 688.56	4 034.66
IRL	1 246.70	6 110.71
ISL	337.56	7 849.99
ITA	4 922.84	4 426.67
JPN	5 387.17	4 668.82
KOR	3 531.75	5 011.56
LUX	2 596.36	5 806.97
MEX	2 476.01	3 916.46
NLD	5 528.99	3 326.09
NOR	599.49	7 986.58
NZL	1 740.61	7 969.97
POL	1 033.90	7 151.46
PRT	2 647.70	5 151.93
SVK	3 734.56	4 873.69
SWE	986.03	8 199.46
TUR	6 188.40	4 891.13
USA	2 395.38	6 731.45

Note: The results are based on the first plausible value for the mathematics scale, denoted PV1MATH in the PISA 2003 database (*www.pisa.oecd.org*).

Kish established a state-of-the-art knowledge of the sampling variance according to the type of estimator and the sampling design. The sampling variance distributions are well known for univariate and multivariate estimators for simple random samples. The use of stratification variables with a simple random sample still allows for the mathematical computation of the sampling variances, but with a substantial increase of complexity. As shown in Table 4.5, the computation of sampling variances for two-stage samples is available for some designs, but it becomes quite difficult to compute for multivariate indices.

Table 4.5
Current status of sampling errors

Selection methods	Means and total of entire samples	Subclass means and differences	Complex analytical statistics *e.g.* coefficients in regression
Simple random selection of elements	Known	Known	Known
Stratified selection of elements	Known	Available	Conjectured
Complex cluster sampling	Known for some sampling design	Available	Difficult

Note: Row 1 refers to standard statistical theory (Kish and Frankel, 1974).

Authors of sampling manuals usually distinguish two types of two-stage sampling (Cochran, 1977; Kish, 1995):
- two-stage sampling with first-stage units of equal sizes,
- two-stage sampling with first-stage units of unequal sizes.

Beyond this distinction, different characteristics of the population and of the sampling design need to be taken into account in the computation of the sampling variance, because they affect the sampling variance. Some of the factors to be considered are:
- Is the population finite or infinite?
- Was size a determining criterion in the selection of the first-stage units?
- Was a systematic procedure used for selecting first-stage or second-stage units?
- Does the sampling design include stratification variables?

The simplest two-stage sample design occurs with infinite populations of stage-one and stage-two units. As both stage units are infinite populations, PSUs are considered to be of equal sizes. If a simple random sample of PSUs is selected and if, within each selected PSU, a simple random sample of stage-two units is selected, then the sampling variance of the mean will be equal to:

$$\sigma^2_{(\hat{\mu})} = \frac{\sigma^2_{between\_PSU}}{n_{PSU}} + \frac{\sigma^2_{within\_PSU}}{n_{PSU}\, n_{within}}$$

Let's apply this formula to an education survey and consider the population of schools as infinite and the population of students within each school as infinite. The computation of the sampling variance of the mean is therefore equal to:

$$\sigma^2_{(\hat{\mu})} = \frac{\sigma^2_{between\_school}}{n_{school}} + \frac{\sigma^2_{within\_school}}{n_{students}}$$

Under these assumptions, the sampling variance of the mean and its square root, *i.e.* the standard error, in Denmark are computed as below. Table 4.6 presents the between-school and within-school variance as well as the numbers of participating schools and students in Denmark and Germany.

$$\sigma^2_{(\hat{\mu})} = \frac{1109.45}{206} + \frac{7357.14}{4218} = 5.39 + 1.74 = 7.13$$

$$\sigma_{(\hat{\mu})} = \sqrt{7.13} = 2.67$$

The sampling variance of the mean and its square root, *i.e.* the standard error, in Germany are equal to:

$$\sigma^2_{(\hat{\mu})} = \frac{6206.92}{216} + \frac{4498.70}{4660} = 28.74 + 0.97 = 29.71$$

$$\sigma_{(\hat{\mu})} = \sqrt{29.71} = 5.45$$

If both samples were considered as simple random samples, then the standard error of the mean for Denmark and Germany would be respectively equal to 1.42 and 1.51.

Table 4.6
Between-school and within-school variances, number of participating schools and students in Denmark and Germany in PISA 2003

	Denmark	Germany
Between-school variance	1 109.45	6 206.92
Within-school variance	7 357.14	4 498.70
Number of participating schools	206	216
Number of participating students	4 218	4 660

Based on these results, the following observations can be made:

- The standard error of the mean is larger for a two-stage sampling than for a simple random sampling. For example, in the case of Germany, the standard errors for simple random sampling and for two-stage sampling are 1.51 and 5.45, respectively. Considering a two-stage sample as a simple random sample will therefore substantially underestimate standard errors and consequently, confidence intervals will be too narrow. The confidence interval on the mathematic scale average, *i.e.* 503, would be equal to: [503 – (1.96*1.51);503 + (1.96*1.51)] = [500.05;505.96] in the case of a simple random sample, but equal to [503 – (1.96*5.45);503 + (1.96*5.45)] = [492.32; 513.68] in the case of a two-stage sample. This indicates that any estimated mean value between 492.32 and 500.05 and between 505.96 and 513.68 may or may not be considered as statistically different from the German average, depending on the standard error used.
- The sampling variance of the mean for two-stage samples is mainly dependent on the between-school variance and the number of participating schools. Indeed, the between-school variance accounts for 76% of the total sampling variance in Denmark, *i.e.* $\frac{5.39}{7.13} = 0.76$. In Germany, the between-school variance accounts for 97% of the total sampling variance, *i.e.* $\frac{28.74}{29.71} = 0.97$. Therefore, one should expect larger sampling variance in countries with larger between-school variance, such as Austria and Germany.

However, the PISA population cannot be considered as an infinite population of schools with an infinite population of students. Further:

- Schools have unequal sizes.
- The PISA sample is a sample without replacement, *i.e.* a school cannot be selected twice.
- Schools are selected proportionally to their sizes and according to a systematic procedure.
- Stratification variables are included in the sample design.

These characteristics of the sampling design will influence the sampling variance, so that the formula used above is also inappropriate. Indeed, *Learning for Tomorrow's World – First Results from PISA 2003* (OECD, 2004a) indicates that the standard errors for the mean performance in mathematics for Denmark and Germany are 2.7 and 3.3, respectively.

This shows that the PISA sample design is quite efficient in reducing the sampling variance. However, the design becomes so complex that there is no easy formula for computing the sampling variance or even mean.

Since the IEA 1990 Reading Literacy Study, replication or resampling methods have been used to compute estimates of the sampling variance for international education surveys. Even though these methods have been known since the late 1950s, they have not been used often as they require numerous computations. With the availability of powerful personal computers in the 1990s and the increased use of international databases by non-mathematicians, international co-ordinating centres were encouraged to use resampling methods for estimating sampling variances from complex sample designs.

According to Rust and Rao (1996):

> "The common principle that these methods have is to use computational intensity to overcome difficulties and inconveniences in utilizing an analytic solution to the problem at hand. Briefly, the replication approach consists of estimating the variance of a population parameter of interest by using a large number of somewhat different subsamples (or somewhat different sampling weights) to calculate the parameter of interest. The variability among the resulting estimates is used to estimate the true sampling error of the initial or full-sample estimate."

In the following sections, these methods will first be described for simple random samples and for two-stage samples. The PISA replication method will be presented subsequently.

REPLICATION METHODS FOR SIMPLE RANDOM SAMPLES

There are two main types of replication methods for simple random samples. These are known as the Jackknife and the Bootstrap. One of the most important differences between the Jackknife and the Bootstrap is related to the procedure used to produce the repeated subsamples or replicate samples. From a sample of *n* units, the Jackknife generates in a systematic way *n* replicate samples of *n*–1 units. The Bootstrap randomly generates a large number of repetitions of *n* units selected with replacement, with each unit having more than one chance of selection.

Since PISA does not use a Bootstrap replication method adapted to multi-stage sample designs, this section will only present the Jackknife method.

Suppose that a sample of ten students has been selected by simple random sampling. The Jackknife method will then generate ten subsamples, or replicate samples, each of nine students, as in Table 4.7.

Table 4.7
The Jackknife replicates and sample means

Student	1	2	3	4	5	6	7	8	9	10	Mean
Value	10	11	12	13	14	15	16	17	18	19	14.50
Replication 1	0	1	1	1	1	1	1	1	1	1	15.00
Replication 2	1	0	1	1	1	1	1	1	1	1	14.88
Replication 3	1	1	0	1	1	1	1	1	1	1	14.77
Replication 4	1	1	1	0	1	1	1	1	1	1	14.66
Replication 5	1	1	1	1	0	1	1	1	1	1	14.55
Replication 6	1	1	1	1	1	0	1	1	1	1	14.44
Replication 7	1	1	1	1	1	1	0	1	1	1	14.33
Replication 8	1	1	1	1	1	1	1	0	1	1	14.22
Replication 9	1	1	1	1	1	1	1	1	0	1	14.11
Replication 10	1	1	1	1	1	1	1	1	1	0	14.00

As shown in Table 4.7, the Jackknife generates ten replicate samples of nine students. The sample mean based on all ten students is equal to 14.5. For the first replicate sample, student 1 is not included in the calculation of the mean, and the mean of the nine students included in replicate sample 1 is 15.00. For the second replicate sample, the second student is not included and the mean of the other nine students is equal to 14.88, and so on.

The Jackknife estimate of sampling variance of the mean is equal to:

$$\sigma^2_{jack} = \frac{n-1}{n}\sum_{i=1}^{n}(\hat{\theta}_{(i)} - \hat{\theta})^2$$

With $\hat{\theta}_{(i)}$ representing the statistic estimate for replicate sample *i*, and $\hat{\theta}$ representing the statistic estimate based on the whole sample.

Based on the data from Table 4.7, the Jackknife sampling variance of the mean is equal to:

$$\sigma^2_{(\hat{\mu})} = \frac{9}{10}\left[(15.00-14.50)^2 + (14.88-14.50)^2 + ... + (14.11-14.50)^2 + (14.00-14.50)^2\right]$$

$$\sigma^2_{(\hat{\mu})} = \frac{9}{10}(1.018519) = 0.9167$$

The usual population variance estimator is equal to:

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 = \frac{1}{9}\left[(10-14.5)^2 + (11-14.5)^2 + ... + (18-14.5)^2 + (19-14.5)^2\right] = 9.17$$

Therefore, the sampling variance of the mean, estimated by the mathematical formula, is equal to:

$$\sigma^2_{(\hat{\mu})} = \frac{\sigma^2}{n} = \frac{9.17}{10} = 0.917$$

As shown in this example, the Jackknife method and the mathematical formula provide identical estimation of the sampling variance. Rust (1996) mathematically demonstrates this equality.

$$\hat{\mu}_{(i)} - \hat{\mu} = \frac{\left[\left(\sum_{i=1}^{n} x_i\right) - x_i\right]}{n-1} - \frac{\left[\sum_{i=1}^{n} x_i\right]}{n} = -\frac{x_i}{n-1} + \left[\sum_{i=1}^{n} x_i\right]\left[\frac{1}{n-1} - \frac{1}{n}\right]$$

$$= -\frac{1}{(n-1)}\left[x_i - \left(\sum_{i=1}^{n} x_i\right)\left(1 - \frac{(n-1)}{n}\right)\right] = -\frac{1}{(n-1)}\left[x_i - \hat{\mu}(n-(n-1)\right] = -\frac{1}{(n-1)}(x_i - \hat{\mu})$$

Therefore,

$$\left(\hat{\mu}_{(i)} - \hat{\mu}\right)^2 = \frac{1}{(n-1)^2}(x_i - \hat{\mu})^2$$

$$\Rightarrow \sum_{i=1}^{n}(\hat{\mu}_{(i)} - \hat{\mu})^2 = \frac{1}{(n-1)^2}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 = \frac{1}{(n-1)}\frac{\sum_{i=1}^{n}(x_i - \hat{\mu})^2}{(n-1)} = \frac{1}{(n-1)}\hat{\sigma}^2$$

$$\Rightarrow \sigma^2_{jack} = \frac{n-1}{n}\sum_{i=1}^{n}(\hat{\mu}_{(i)} - \hat{\mu})^2 = \frac{(n-1)}{n}\frac{1}{(n-1)}\hat{\sigma}^2 = \frac{\hat{\sigma}^2}{n}$$

The Jackknife method can also be applied to compute the sampling variance for other statistics, such as regression coefficients. As an example, in Table 4.8, the procedure consists of the computation of 11 regression coefficients: 1 based on the whole sample and 10 others based on one replicate sample. The comparison between the whole sample regression coefficient and each of the ten replicate regression coefficients will provide an estimate of the sampling variance of that statistic.

Table 4.8
Values on variables X and Y for a sample of ten students

Student	1	2	3	4	5	6	7	8	9	10
Value Y	10	11	12	13	14	15	16	17	18	19
Value X	10	13	14	19	11	12	16	17	18	15

The regression coefficient for the whole sample is equal to 0.53. The regression coefficients for ten replicate samples are shown in Table 4.9.

Table 4.9
Regression coefficients for each replicate sample

	Regression coefficient
Replicate 1	0.35
Replicate 2	0.55
Replicate 3	0.56
Replicate 4	0.64
Replicate 5	0.51
Replicate 6	0.55
Replicate 7	0.51
Replicate 8	0.48
Replicate 9	0.43
Replicate 10	0.68

The Jackknife formula, *i.e.* $\sigma^2_{jack} = \frac{n-1}{n}\sum_{i=1}^{n}(\hat{\theta}_{(i)} - \hat{\theta})^2$, can be applied to compute the sampling variance of the regression coefficient.

$$\sigma^2_{jack} = \frac{n-1}{n}\sum_{i=1}^{n}(\hat{\theta}_{(i)} - \hat{\theta})^2 = \frac{9}{10}\left[(0.35-0.53)^2 + (0.55-0.53)^2 + \ldots(0.68-0.53)^2\right] = 0.07$$

This result is identical to the result that the usual sampling variance formula for a regression coefficient would render.

REPLICATION METHODS FOR TWO-STAGE SAMPLES

There are three types of replication methods for two-stage samples:

1. the Jackknife, with two variants: one for unstratified samples and another one for stratified samples;
2. the Balanced Repeated Replication (BRR) and its variant, Fay's modification;
3. the Bootstrap.

PISA uses BRR with Fay's modification.[1]

The Jackknife for unstratified two-stage sample designs

If a simple random sample of PSUs is drawn without the use of any stratification variables, then it can be shown that the sampling variance of the mean obtained using the Jackknife method is mathematically equal to the formula provided earlier in this chapter, *i.e.:*

$$\sigma^2_{(\hat{\mu})} = \frac{\sigma^2_{between\_PSU}}{n_{PSU}} + \frac{\sigma^2_{within\_PSU}}{n_{PSU}\, n_{within}}$$

Consider a sample of ten schools and within selected schools, a simple random sample of students. The Jackknife method for an unstratified two-stage sample consists of generating ten replicates of nine schools. Each school is removed only once, in a systematic way.

For the first replicate, denoted R1, school 1 has been removed. As shown in Table 4.10, the weights of the other schools in the first replicate are adjusted by a factor of 1.11, *i.e.* $\frac{10}{9}$ or, as a general rule, by a factor of $\frac{G}{G-1}$, with *G* being the number of PSUs and the number of replicates in the sample. This adjustment factor is then applied when school replicate weights and within school replicate weights are combined to give the student replicate weights. For the second replicate, school 2 is removed and the weights in the remaining schools are adjusted by the same factor, and so on.

Table 4.10
The Jackknife replicates for unstratified two-stage sample designs

Replicate	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
School 1	0.00	1.11	1.11	1.11	1.11	1.11	1.11	1.11	1.11	1.11
School 2	1.11	0.00	1.11	1.11	1.11	1.11	1.11	1.11	1.11	1.11
School 3	1.11	1.11	0.00	1.11	1.11	1.11	1.11	1.11	1.11	1.11
School 4	1.11	1.11	1.11	0.00	1.11	1.11	1.11	1.11	1.11	1.11
School 5	1.11	1.11	1.11	1.11	0.00	1.11	1.11	1.11	1.11	1.11
School 6	1.11	1.11	1.11	1.11	1.11	0.00	1.11	1.11	1.11	1.11
School 7	1.11	1.11	1.11	1.11	1.11	1.11	0.00	1.11	1.11	1.11
School 8	1.11	1.11	1.11	1.11	1.11	1.11	1.11	0.00	1.11	1.11
School 9	1.11	1.11	1.11	1.11	1.11	1.11	1.11	1.11	0.00	1.11
School 10	1.11	1.11	1.11	1.11	1.11	1.11	1.11	1.11	1.11	0.00

The statistic of interest is computed for the whole sample, and then again for each replicate. The replicate estimates are then compared to the whole sample estimate to obtain the sampling variance, as follows:

$$\sigma^2_{(\hat{\theta})} = \frac{(G-1)}{G}\sum_{i=1}^{G}(\hat{\theta}_{(i)}-\hat{\theta})^2$$

This formula is identical to the one used for a simple random sample, except that instead of using *n* replicates, *n* being the number of units in the sample, this formula uses *G* replicates, with *G* being the number of PSUs.

The Jackknife for stratified two-stage sample designs

As mentioned at the beginning of Chapter 3, two major principles underlie all sample designs. The first is the need to avoid bias in the selection procedure, the second to achieve maximum precision in view of the available financial resources.

To reduce the uncertainty, or to minimise the sampling variance without modifying the sample size, international and national education surveys usually implement the following procedures in the sampling design:

- PSUs are selected proportionally to their size and according to a systematic procedure. This procedure leads to an efficient student sampling procedure. Equal-sized samples of students can be selected from each school. At the same time, the overall selection probabilities (combining the school and student sampling components) do not vary much.
- PISA national centres are encouraged to identify stratification variables that are statistically associated with student performance. Characteristics, such as rural versus urban, academic versus vocational, private versus public, could be associated with student performance. The sampling variance reduction will be proportional to the explanatory power of these stratification variables on student performance.

The Jackknife for stratified two-stage samples allows the reduction of the sampling variance by taking both of these aspects into consideration. Failing to do so, would lead to a systematic overestimation of sampling variances.

Table 4.11
The Jackknife replicates for stratified two-stage sample designs

Pseudo-stratum	School	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
1	1	2	1	1	1	1	1	1	1	1	1
1	2	0	1	1	1	1	1	1	1	1	1
2	3	1	0	1	1	1	1	1	1	1	1
2	4	1	2	1	1	1	1	1	1	1	1
3	5	1	1	2	1	1	1	1	1	1	1
3	6	1	1	0	1	1	1	1	1	1	1
4	7	1	1	1	0	1	1	1	1	1	1
4	8	1	1	1	2	1	1	1	1	1	1
5	9	1	1	1	1	2	1	1	1	1	1
5	10	1	1	1	1	0	1	1	1	1	1
6	11	1	1	1	1	1	2	1	1	1	1
6	12	1	1	1	1	1	0	1	1	1	1
7	13	1	1	1	1	1	1	0	1	1	1
7	14	1	1	1	1	1	1	2	1	1	1
8	15	1	1	1	1	1	1	1	0	1	1
8	16	1	1	1	1	1	1	1	2	1	1
9	17	1	1	1	1	1	1	1	1	0	1
9	18	1	1	1	1	1	1	1	1	2	1
10	19	1	1	1	1	1	1	1	1	1	2
10	20	1	1	1	1	1	1	1	1	1	0

Suppose that the list of schools in the population is divided into two parts called strata: rural schools and urban schools. Further, within these two strata, schools are sorted by size. Within each stratum, ten schools are selected systematically and proportionally to their size.

The Jackknife method for stratified two-stage sample designs consists of systematically pairing sampled schools within each stratum in the order in which they were selected. Therefore, schools will be paired with other similar schools.

Table 4.11 shows how replicates are generated for this method. Schools 1 to 10 are in the stratum of "rural", and schools 11 to 20 are in the stratum of "urban". Within each stratum, there are therefore five school pairs, or pseudo-strata (also called variance strata).

The Jackknife for stratified two-stage samples will generate as many replicates as there are pairs or pseudo strata. In this example, ten replicates will therefore be generated. For each replicate sample, one school is randomly removed within a particular pseudo-stratum and the weight of the remaining school in the pseudo-stratum is doubled. For replicate 1, denoted R1, school 2 is removed and the weight of school 1 is doubled in the pseudo-stratum 1. For replicate 2, school 3 is removed and the weight of school 4 is doubled in the pseudo-stratum 2, and so on.

As previously mentioned, the statistic of interest is computed based on the whole sample and then again based on each replicate sample. The replicate estimates are then compared to the whole sample estimate to obtain the sampling variance, as follows:

$$\sigma^2_{(\hat{\theta})} = \sum_{i=1}^{G} (\hat{\theta}_{(i)} - \hat{\theta})^2$$

This replication method is now generally used in IEA studies.

The Balanced Repeated Replication method

While the Jackknife method consists of removing only one school for each replicate sample, the Balanced Repeated Replication (BRR) method proceeds by selecting at random one school within each pseudo-stratum to have its weight set to 0, and by doubling the weights of the remaining schools as shown in Table 4.12.

Table 4.12

Replicates with the Balanced Repeated Replication method

Pseudo-stratum	School	R1	R2	R3	R4	R5	R6	R7	R8	R9	R 10	R 11	R 12
1	**1**	2	0	0	2	0	0	0	2	2	2	0	2
1	**2**	0	2	2	0	2	2	2	0	0	0	2	0
2	**3**	2	2	0	0	2	0	0	0	2	2	2	0
2	**4**	0	0	2	2	0	2	2	2	0	0	0	2
3	**5**	2	0	2	0	0	2	0	0	0	2	2	2
3	**6**	0	2	0	2	2	0	2	2	2	0	0	0
4	**7**	2	2	0	2	0	0	2	0	0	0	2	2
4	**8**	0	0	2	0	2	2	0	2	2	2	0	0
5	**9**	2	2	2	0	2	0	0	2	0	0	0	2
5	**10**	0	0	0	2	0	2	2	0	2	2	2	0
6	**11**	2	2	2	2	0	2	0	0	2	0	0	0
6	**12**	0	0	0	0	2	0	2	2	0	2	2	2
7	**13**	2	0	2	2	2	0	2	0	0	2	0	0
7	**14**	0	2	0	0	0	2	0	2	2	0	2	2
8	**15**	2	0	0	2	2	2	0	2	0	0	2	0
8	**16**	0	2	2	0	0	0	2	0	2	2	0	2
9	**17**	2	0	0	0	2	2	2	0	2	0	0	2
9	**18**	0	2	2	2	0	0	0	2	0	2	2	0
10	**19**	2	2	0	0	0	2	2	2	0	2	0	0
10	**20**	0	0	2	2	2	0	0	0	2	0	2	2

As this method results in a large set of possible replicates, a balanced set of replicate samples is generated according to Hadamard matrices in order to avoid lengthy computations. The number of replicates is the smallest multiple of four, greater than or equal to the number of pseudo-strata. In this example, as there are 10 pseudo-strata, 12 replicates will be generated.

The statistic of interest is again computed based on the whole sample, and then again for each replicate. The replicate estimates are then compared with the whole sample estimate to estimate the sampling variance, as follows:

$$\sigma^2_{(\hat{\theta})} = \frac{1}{G}\sum_{i=1}^{G}(\hat{\theta}_{(i)} - \hat{\theta})^2$$

With this replication method, each replicate sample only uses half of the available observations. This large reduction in sample might therefore become problematic for the estimation of a statistic on a rare subpopulation. Indeed, the number of remaining observations might be so small, even equal to 0, that the estimation of the population parameter for a particular replicate sample is impossible. To overcome this disadvantage, Fay developed a variant to the BRR method. Instead of multiplying the school weights by a factor of 0 or 2, Fay suggested multiplying the weights by a deflating factor k between 0 and 1, with the second inflating factor being equal to 2 minus k. For instance, if the deflating weight factor, denoted k, is equal to 0.6, then the inflating weight factor will be equal to $2-k$, *i.e.* $1-0.6=1.4$ (Judkins, 1990).

PISA uses the Fay method with a factor of 0.5. Table 4.13 describes how the replicate samples and weights are generated for this method.

Table 4.13
The Fay replicates

Pseudo-stratum	School	R1	R2	R3	R4	R5	R6	R7	R8	R9	R 10	R 11	R 12
1	1	1.5	0.5	0.5	1.5	0.5	0.5	0.5	1.5	1.5	1.5	0.5	1.5
1	2	0.5	1.5	1.5	0.5	1.5	1.5	1.5	0.5	0.5	0.5	1.5	0.5
2	3	1.5	1.5	0.5	0.5	1.5	0.5	0.5	0.5	1.5	1.5	1.5	0.5
2	4	0.5	0.5	1.5	1.5	0.5	1.5	1.5	1.5	0.5	0.5	0.5	1.5
3	5	1.5	0.5	1.5	0.5	0.5	1.5	0.5	0.5	0.5	1.5	1.5	1.5
3	6	0.5	1.5	0.5	1.5	1.5	0.5	1.5	1.5	1.5	0.5	0.5	0.5
4	7	1.5	1.5	0.5	1.5	0.5	0.5	1.5	0.5	0.5	0.5	1.5	1.5
4	8	0.5	0.5	1.5	0.5	1.5	1.5	0.5	1.5	1.5	1.5	0.5	0.5
5	9	1.5	1.5	1.5	0.5	1.5	0.5	0.5	1.5	0.5	0.5	0.5	1.5
5	10	0.5	0.5	0.5	1.5	0.5	1.5	1.5	0.5	1.5	1.5	1.5	0.5
6	11	1.5	1.5	1.5	1.5	0.5	1.5	0.5	0.5	1.5	0.5	0.5	0.5
6	12	0.5	0.5	0.5	0.5	1.5	0.5	1.5	1.5	0.5	1.5	1.5	1.5
7	13	1.5	0.5	1.5	1.5	1.5	0.5	1.5	0.5	0.5	1.5	0.5	0.5
7	14	0.5	1.5	0.5	0.5	0.5	1.5	0.5	1.5	1.5	0.5	1.5	1.5
8	15	1.5	0.5	0.5	1.5	1.5	1.5	0.5	1.5	0.5	0.5	1.5	0.5
8	16	0.5	1.5	1.5	0.5	0.5	0.5	1.5	0.5	1.5	1.5	0.5	1.5
9	17	1.5	0.5	0.5	0.5	1.5	1.5	1.5	0.5	1.5	0.5	0.5	1.5
9	18	0.5	1.5	1.5	1.5	0.5	0.5	0.5	1.5	0.5	1.5	1.5	0.5
10	19	1.5	1.5	0.5	0.5	0.5	1.5	1.5	1.5	0.5	1.5	0.5	0.5
10	20	0.5	0.5	1.5	1.5	1.5	0.5	0.5	0.5	1.5	0.5	1.5	1.5

As with all replication methods, the statistic of interest is computed based on the whole sample, and then again on each replicate. The replicate estimates are then compared to the whole sample estimate to get the sampling variance, as follows:

$$\sigma^2_{(\hat{\theta})} = \frac{1}{G(1-k)^2}\sum_{i=1}^{G}(\hat{\theta}_{(i)} - \hat{\theta})^2$$

In PISA, it was decided to generate 80 replicate samples and therefore 80 replicate weights. Therefore, the formula becomes:

$$\sigma^2_{(\hat{\theta})} = \frac{1}{G(1-k)^2}\sum_{i=1}^{G}(\hat{\theta}_{(i)}-\hat{\theta})^2 = \frac{1}{80(1-0.5)^2}\sum_{i=1}^{80}(\hat{\theta}_{(i)}-\hat{\theta})^2 = \frac{1}{20}\sum_{i=1}^{80}(\hat{\theta}_{(i)}-\hat{\theta})^2$$

OTHER PROCEDURES FOR ACCOUNTING FOR CLUSTERED SAMPLES

For the past two decades, multi-level models and software packages have been introduced in the education research field. There is no doubt that these models led to a breakthrough in the unravelling of education phenomena. Indeed, multi-level regression models offer the possibility of taking into account the fact that students are nested within classes and schools: each contributing factor can be evaluated when establishing the outcome measure.

Multi-level regression software packages, such as MLWin® or HLM®, just like any professional statistical package, provide an estimate of the standard error for each of the estimated population parameters. While SAS® and SPSS® consider the sample as a simple random sample of population elements, MLWin® and HLM® recognise the hierarchical structure of the data, but consider that the school sample is a simple random one. They therefore do not take into account the complementary sample design information used in PISA to reduce the sampling variance. Consequently, in PISA, the sampling variances estimated with multi-level models will always be greater than the sampling variances estimated with Fay replicate samples.

As these multi-level model packages do not incorporate the additional sample design information, their standard error estimates are similar to the Jackknife method for unstratified samples. For instance, the PISA 2003 data in Germany were analysed using the multi-level model proposed by SAS® and called PROC MIXED. The standard errors of the mean of the five plausible values[2] for the combined mathematical literacy scale were respectively 5.4565, 5.3900, 5.3911, 5.4692, and 5.3461. The average of these five standard errors is 5.41. Recall that the use of the formula assuming PSUs are selected as simple random sampling discussed above produces an estimate of the sampling variance equal to 5.45.

With multi-level software packages, using replicates cannot be avoided if unbiased estimates of the standard errors for the estimates need to be obtained.

CONCLUSION

Since international education surveys use a two-stage sample design most of the time, it would be inappropriate to apply the sampling distribution formulas developed for simple random sampling. Doing so would lead to an underestimation of the sampling variances.

Sampling designs in education surveys can be very intricate. As a result, sampling distributions might not be available or too complex even for simple estimators, such as means. Since the 1990 IEA Reading Literacy Study, sampling variances have been estimated through replication methods. These methods function by generating several subsamples, or replicate samples, from the whole sample. The statistic of interest is then estimated for each of these replicate samples and then compared to the whole sample estimate to provide an estimate of the sampling variance.

A replicate sample is formed simply through a transformation of the full sample weights according to an algorithm specific to the replication method. These methods therefore can be applied to any estimators[3] – means, medians, percentiles, correlations, regression coefficients, etc. – which can be easily computed thanks to advanced computing resources. Further, using these replicate weights does not require an extensive knowledge in statistics, since these procedures can be applied regardless of the statistic of interest.

Notes

1. See the reasons for this decision in the *PISA 2000 Technical Report* (OECD, 2002c).

2. See Chapter 6 for a description of plausible values.

3. Several empirical or theoretical studies have compared the different resampling methods for complex sampling design. As Rust and Krawchuk noted: "A benefit of both BRR and modified BRR over the Jackknife is that they have a sound theoretical basis for use with nonsmooth statistics, such as quantiles like the median. It has long been known that the Jackknife is inconsistent for estimating the variances of quantiles. That is, as the sample size increases for a given sample design, the estimation of the variances of quantiles does not necessarily become more precise when using the Jackknife." (Rust and Krawchuk, 2002).

References

Beaton, A.E. (1987), *The NAEP 1983-1984 Technical Report*, Educational Testing Service, Princeton.

Beaton, A.E., *et al.* (1996), *Mathematics Achievement in the Middle School Years, IEA's Third International Mathematics and Science Study*, Boston College, Chestnut Hill, MA.

Bloom, B.S. (1979), *Caractéristiques individuelles et apprentissage scolaire*, Éditions Labor, Brussels.

Bressoux, P. (2008), *Modélisation statistique appliquée aux sciences sociales*, De Boek, Brussels.

Bryk, A.S. and **S.W. Raudenbush** (1992), *Hierarchical Linear Models for Social and Behavioural Research: Applications and Data Analysis Methods*, Sage Publications, Newbury Park, CA.

Buchmann, C. (2000), *Family structure, parental perceptions and child labor in Kenya: What factors determine who is enrolled in school?* a*Soc. Forces,* No. 78, pp. 1349-79.

Cochran, W.G. (1977), *Sampling Techniques*, J. Wiley and Sons, Inc., New York.

Dunn, O.J. (1961), "Multilple Comparisons among Menas", *Journal of the American Statistical Association*, Vol. 56, American Statistical Association, Alexandria, pp. 52-64.

Kish, L. (1995), *Survey Sampling*, J. Wiley and Sons, Inc., New York.

Knighton, T. and **P. Bussière** (2006), "Educational Outcomes at Age 19 Associated with Reading Ability at Age 15", Statistics Canada, Ottawa.

Gonzalez, E. and **A. Kennedy** (2003), *PIRLS 2001 User Guide for the International Database*, Boston College, Chestnut Hill, MA.

Ganzeboom, H.B.G., P.M. De Graaf and **D.J. Treiman** (1992), "A Standard International Socio-economic Index of Occupation Status", *Social Science Research* 21(1), Elsevier Ltd, pp 1-56.

Goldstein, H. (1995), *Multilevel Statistical Models,* 2nd Edition, Edward Arnold, London.

Goldstein, H. (1997), "Methods in School Effectiveness Research", *School Effectiveness and School Improvement* 8, Swets and Zeitlinger, Lisse, Netherlands, pp. 369-395.

Hubin, J.P. (ed.) (2007), *Les indicateurs de l'enseignement*, 2nd Edition, Ministère de la Communauté française, Brussels.

Husen, T. (1967), *International Study of Achievement in Mathematics: A Comparison of Twelve Countries,* Almqvist and Wiksells, Uppsala.

International Labour Organisation (ILO) (1990), *International Standard Classification of Occupations: ISCO-88.* Geneva: International Labour Office.

Lafontaine, D. and **C. Monseur** (forthcoming), "Impact of Test Characteristics on Gender Equity Indicators in the Assessment of Reading Comprehension", *European Educational Research Journal*, Special Issue on PISA and Gender.

Lietz, P. (2006), "A Meta-Analysis of Gender Differences in Reading Achievement at the Secondary Level", *Studies in Educational Evaluation* 32, pp. 317-344.

Monseur, C. and **M. Crahay** (forthcoming), "Composition académique et sociale des établissements, efficacité et inégalités scolaires : une comparaison internationale – Analyse secondaire des données PISA 2006", *Revue française de pédagogie*.

OECD (1998), *Education at a Glance – OECD Indicators,* OECD, Paris.

OECD (1999a), *Measuring Student Knowledge and Skills – A New Framework for Assessment,* OECD, Paris.

OECD (1999b), *Classifying Educational Programmes – Manual for ISCED-97 Implementation in OECD Countries,* OECD, Paris.

OECD (2001), *Knowledge and Skills for Life – First Results from PISA 2000,* OECD, Paris.

OECD (2002a), *Programme for International Student Assessment – Manual for the PISA 2000 Database,* OECD, Paris.

OECD (2002b), *Sample Tasks from the PISA 2000 Assessment – Reading, Mathematical and Scientific Literacy,* OECD, Paris.

OECD (2002c), *Programme for International Student Assessment – PISA 2000 Technical Report,* OECD, Paris.

OECD (2002d), *Reading for Change: Performance and Engagement across Countries – Results from PISA 2000,* OECD, Paris.

OECD (2003a), *Literacy Skills for the World of Tomorrow – Further Results from PISA 2000,* OECD, Paris.

OECD (2003b), *The PISA 2003 Assessment Framework – Mathematics, Reading, Science and Problem Solving Knowledge and Skills,* OECD, Paris.

OECD (2004a), *Learning for Tomorrow's World – First Results from PISA 2003,* OECD, Paris.

OECD (2004b), *Problem Solving for Tomorrow's World – First Measures of Cross-Curricular Competencies from PISA 2003,* OECD, Paris.

OECD (2005a), *PISA 2003 Technical Report,* OECD, Paris.

OECD (2005b), *PISA 2003 Data Analysis Manual,* OECD, Paris.

OECD (2006), *Assessing Scientific, Reading and Mathematical Literacy: A Framework for PISA 2006,* OECD, Paris.

OECD (2007), *PISA 2006: Science Competencies for Tomorrow's World*, OECD, Paris.

OECD (2009), *PISA 2006 Technical Report,* OECD, Paris.

Peaker, G.F. (1975), *An Empirical Study of Education in Twenty-One Countries: A Technical report. International Studies in Evaluation VIII*, Wiley, New York and Almqvist and Wiksell, Stockholm.

Rust, K.F. and **J.N.K. Rao** (1996), "Variance Estimation for Complex Surveys Using Replication Techniques", *Statistical Methods in Medical Research,* Vol. 5, Hodder Arnold, London, pp. 283-310.

Rutter, M., *et al.* (2004), "Gender Differences in Reading Difficulties: Findings from Four Epidemiology Studies", *Journal of the American Medical Association* 291, pp. 2007-2012.

Schulz, W. (2006), *Measuring the socio-economic background of students and its effect on achievement in PISA 2000 and PISA 2003*, Paper presented at the Annual Meetings of the American Educational Research Association (AERA) in San Francisco, 7-11 April.

Wagemaker, H. (1996), *Are Girls Better Readers. Gender Differences in Reading Literacy in 32 Countries*, IEA, The Hague.

Warm, T.A. (1989), "Weighted Likelihood Estimation of Ability in Item Response Theory", *Psychometrika,* Vol. 54(3), Psychometric Society, Williamsburg, VA., pp. 427-450.

Wright, B.D. and **M.H. Stone** (1979), *Best Test Design: Rasch Measurement,* MESA Press, Chicago.

Table of contents

LIST OF BOXES

PISA DATA ANALYSIS MANUAL: SPSS® SECOND EDITION – ISBN 978-92-64-05626-8 – © OECD 2009

LIST OF FIGURES

LIST OF TABLES

User's Guide

Preparation of data files

All data files (in text format) and the SPSS® control files are available on the PISA website (*www.pisa.oecd.org*).

SPSS® users

By running the SPSS® control files, the PISA data files are created in the SPSS® format. Before starting analysis in the following chapters, save the PISA 2000 data files in the folder of "c:\pisa2000\data\", the PISA 2003 data files in "c:\pisa2003\data\", and the PISA 2006 data files in "c:\pisa2006\data\".

SPSS® syntax and macros

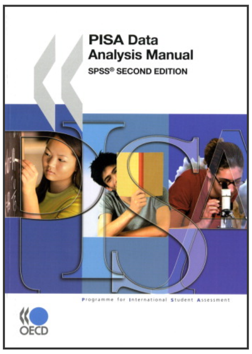
All syntaxes and macros in this manual can be copied from the PISA website (*www.pisa.oecd.org*). These macros were developed for SPSS 17.0. The 19 SPSS® macros presented in Chapter 17 need to be saved under "c:\pisa\macro\", before staring analysis. Each chapter of the manual contains a complete set of syntaxes, which must be done sequentially, for all of them to run correctly, within the chapter.

Rounding of figures

In the tables and formulas, figures were rounded to a convenient number of decimal places, although calculations were always made with the full number of decimal places.

Country abbreviations used in this manual

AUS	Australia	FRA	France	MEX	Mexico
AUT	Austria	GBR	United Kingdom	NLD	Netherlands
BEL	Belgium	GRC	Greece	NOR	Norway
CAN	Canada	HUN	Hungary	NZL	New Zealand
CHE	Switzerland	IRL	Ireland	POL	Poland
CZE	Czech Republic	ISL	Iceland	PRT	Portugal
DEU	Germany	ITA	Italy	SVK	Slovak Republic
DNK	Denmark	JPN	Japan	SWE	Sweden
ESP	Spain	KOR	Korea	TUR	Turkey
FIN	Finland	LUX	Luxembourg	USA	United States



From:

PISA Data Analysis Manual: SPSS, Second Edition

Access the complete publication at:

https://doi.org/10.1787/9789264056275-en

Please cite this chapter as:

OECD (2009), “Replicate Weights”, in *PISA Data Analysis Manual: SPSS, Second Edition*, OECD Publishing, Paris.

DOI: https://doi.org/10.1787/9789264056275-5-en