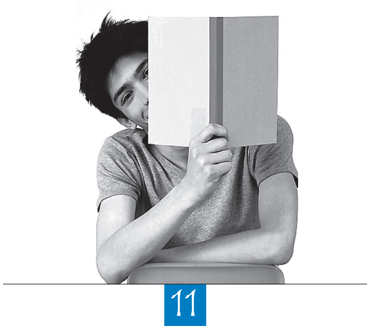
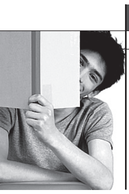


11

Standard Error on a Difference



INTRODUCTION

This chapter will discuss the computation of standard errors on differences. Following a description of the statistical issues for such estimates, the different steps for computing such standard errors will be presented. Finally, the correction of the critical value for multiple comparisons will be discussed.

STATISTICAL ISSUES AND COMPUTING STANDARD ERRORS ON DIFFERENCES

Suppose that X represents the student score for a mathematics test and Y the student score for a science test for the same sample of students. To summarise the score distribution for both tests, one can compute:

- $\mu_{(X)}$, $\mu_{(Y)}$, representing respectively the mean of X and the mean of Y,
- $\sigma^2_{(X)}$, $\sigma^2_{(Y)}$, representing respectively the variance of X and the variance of Y.

It can be shown that:

$$\mu_{(X+Y)} = \mu_{(X)} + \mu_{(Y)} \text{ and}$$

$$\sigma^2_{(X+Y)} = \sigma^2_{(X)} + \sigma^2_{(Y)} + 2\operatorname{cov}(X,Y)$$

If a total score is computed by just adding the mathematics and science scores, then according to these two formulae, the mean of this total score will be the sum of the two initial means, and the variance of the total score will be equal to the sum of the variance of the two initial variables X and Y plus two times the covariance between X and Y. This covariance represents the relationship between X and Y. Usually, high performers in mathematics are also high performers in science; thus, one should expect a positive and high covariance in this particular example.

Similarly,

$$\mu_{(X-Y)} = \mu_{(X)} - \mu_{(Y)} \text{ and}$$

$$\sigma^2_{(X-Y)} = \sigma^2_{(X)} + \sigma^2_{(Y)} - 2\operatorname{cov}(X,Y)$$

In other words, the variance of a difference is equal to the sum of the variances of the two initial variables minus two times the covariance between the two initial variables.

As described in Chapter 4, a sampling distribution has the same characteristics as any distribution, except that units consist of sample estimates and not observations. Therefore,

$$\sigma^2_{(\hat{\mu}_X - \hat{\mu}_Y)} = \sigma^2_{(\hat{\mu}_X)} + \sigma^2_{(\hat{\mu}_Y)} - 2\operatorname{cov}(\hat{\mu}_X, \hat{\mu}_Y)$$

The sampling variance of a difference is equal to the sum of the two initial sampling variances minus two times the covariance between the two sampling distributions on the estimates.

Suppose that one wants to determine whether female performance is on average higher than male performance. As for all statistical analyses, the null hypothesis has to be tested. In this particular example, it will consist of computing the difference between the male performance mean and the female performance mean, or the inverse. The null hypothesis will be:

$$H_0 : \hat{\mu}_{(males)} - \hat{\mu}_{(females)} = 0$$

To test this null hypothesis, the standard error on this difference has to be computed and then compared to the observed difference. The respective standard errors on the mean estimate for males and for females ($\sigma_{(\hat{\mu}_{males})}$, $\sigma_{(\hat{\mu}_{females})}$) can be easily computed.

What does the covariance between the two variables, *i.e.* $\hat{\mu}_{(males)}$, $\hat{\mu}_{(females)}$, tell us? A positive covariance means that if $\hat{\mu}_{(males)}$ increases, then $\hat{\mu}_{(females)}$ will also increase. A covariance equal or close to 0 means that $\hat{\mu}_{(males)}$ can increase or decrease with $\hat{\mu}_{(females)}$ remaining unchanged. Finally, a negative covariance means that if $\hat{\mu}_{(males)}$ increases, then $\hat{\mu}_{(females)}$ will decrease, and inversely.

How are $\hat{\mu}_{(males)}$ and $\hat{\mu}_{(females)}$ correlated? Suppose that in the school sample, a coeducational school attended by low performers is replaced by a coeducational school attended by high performers. The country mean will increase slightly, as well as the means for males and females. If the replacement process is continued, $\hat{\mu}_{(males)}$ and $\hat{\mu}_{(females)}$ will likely increase in a similar pattern. Indeed, a coeducational school attended by high-performing males is usually also attended by high-performing females. Therefore, the covariance between $\hat{\mu}_{(males)}$ and $\hat{\mu}_{(females)}$ will be positive.

Let us now suppose that all schools are single gender. A boys' school can replace a girls' school in the sample and therefore $\hat{\mu}_{(males)}$ and $\hat{\mu}_{(females)}$ will change. If gender is used as a stratification variable, *i.e.* all girls' schools are allocated to an explicit stratum and all boys' schools are allocated to another explicit stratum, then a girls' school can only be replaced by another girls' school. In this case, only $\hat{\mu}_{(females)}$ will change. As $\hat{\mu}_{(females)}$ might change without affecting $\hat{\mu}_{(males)}$, the expected value of the covariance between $\hat{\mu}_{(males)}$ and $\hat{\mu}_{(females)}$ is 0.

Finally, a negative covariance means that if a school is attended by high-performing males, then that school is also attended by low-performing females or the inverse. This situation is not likely.

In summary, the expected value of the covariance will be equal to 0 if the two subsamples are independent. If the two subsamples are not independent, then the expected value of the covariance might differ from 0.

In PISA, country samples are independent. Therefore, for any comparison between two countries, the expected value of the covariance will be equal to 0. The standard error on the estimate is:

$$\sigma_{(\hat{\theta}_i-\hat{\theta}_j)}=\sqrt{\sigma^2_{(\hat{\theta}_i)}+\sigma^2_{(\hat{\theta}_j)}}\text{, with } \theta \text{ being any statistic.}$$

For instance, in PISA 2003, the mean score in mathematics is equal to 503 with a standard error of 3.3 in Germany, and the mean is equal to 529 with a standard error of 2.3 in Belgium. Therefore, the difference between Germany and Belgium is 529-503=26 and the standard error on this difference is:

$$\sigma_{(\hat{\theta}_i-\hat{\theta}_j)}=\sqrt{\sigma^2_{(\hat{\theta}_i)}+\sigma^2_{(\hat{\theta}_j)}}=\sqrt{(3.3)^2+(2.3)^2}=\sqrt{10.89+5.29}=\sqrt{16.18}=4.02$$

The difference divided by its standard error, *i.e.* $\frac{26}{4.02}=6.46$, is greater than 1.96, which is significant. This means that the performance in Belgium is greater than the performance in Germany.

Similarly, the percentage of students below Level 1 in mathematics is equal to 9.2% in Germany (with a standard error of 0.8) and to 7.2% in Belgium (with a standard error of 0.6). The difference is equal to 9.2 – 7.2 = 2 and the standard error on this difference is equal to:

$$\sigma_{(\hat{\theta}_i-\hat{\theta}_j)}=\sqrt{\sigma^2_{(\hat{\theta}_i)}+\sigma^2_{(\hat{\theta}_j)}}=\sqrt{(0.8)^2+(0.6)^2}=\sqrt{0.64+0.36}=\sqrt{1}=1$$

The standardised difference is equal to 2 (*i.e.* $\frac{2}{1}$), which is significant. Thus the percentage of students below Level 1 is greater in Germany than in Belgium.

Finally, the regression coefficient of student socio-economic background index on the science performance in PISA 2006 is equal to 47.71 for Germany (with a standard error equal to 1.89), and 46.45 for Belgium

(with a standard error equal to 2.08). These two regression coefficients do not statistically differ as the standardised difference is equal to 0.44:

$$\frac{\hat{\beta}_1 - \hat{\beta}_2}{\sqrt{\sigma^2_{(\hat{\beta}_1)} + \sigma^2_{(\hat{\beta}_2)}}} = \frac{47.71 - 46.45}{\sqrt{(1.89)^2 + (2.08)^2}} = 0.44$$

While the covariance between two country estimates for any statistical parameter is expected to be 0, it differs from 0 between an OECD country and the OECD average or total, as any OECD country contributes to the computation of the OECD average or total parameter estimate. Chapter 12 will describe how the standard error on the difference between an OECD country and the OECD average can be computed.

Within a particular country, any subsamples will be considered as independent if the categorical variable used to define the subsamples was used as an explicit stratification variable. For instance, since Canada used the provinces as an explicit stratification variable, these subsamples are independent and any comparison between two provinces does not require the estimation of the covariance between the sampling distributions.

As a general rule, any comparison between countries does not require the estimation of the covariance, but it is strongly advised to estimate the covariance between the sampling distributions for within-country comparisons.

As described earlier in this section, the estimation of the covariance between, for instance, $\hat{\mu}_{(males)}$ and $\hat{\mu}_{(females)}$ would require the selection of several samples and then the analysis of the variation of $\hat{\mu}_{(males)}$ in conjunction with $\hat{\mu}_{(females)}$. Such procedure is, of course, unrealistic. Therefore, as for any computation of a standard error in PISA, replication methods using the supplied replicate weights will be used to estimate the standard error on a difference.

THE STANDARD ERROR ON A DIFFERENCE WITHOUT PLAUSIBLE VALUES

Let's suppose that a researcher wants to test whether females have higher job expectations than males in Germany.

As described in Chapter 7, the SPSS® macro MCR_SE_UNIV can be used to estimate the average job expectation for males and females respectively.

Box 11.1 **SPSS® syntax for computing the mean of job expectations by gender (*e.g.* PISA 2003)**

```
GET FILE="C:\PISA\2003\DATA\INT_STUI_2003.SAV".
SELECT IF (CNT="DEU").
SAVE OUTFILE="C:\TEMP\DEU1.SAV".

INSERT FILE="C:\PISA\MACRO\MCR_SE_UNIV.SPS".

UNIVAR STAT=MEAN /
       DEP=BSMJ /
       GRP=CNT ST03Q01/
       INFILE="C:\TEMP\DEU1.SAV"/.
```

Box 11.1 presents the SPSS® syntax for the computation of the mean for job expectations at the age of 30 (BSMJ) by gender. Table 11.1 presents the structure of the output data file as well as the results by gender.

Table 11.1
Output data file from Box 11.1

CNT	ST03Q01	Statistic	BSMJ	se_BSMJ
DEU	1	MEAN	53.05	0.57
DEU	2	MEAN	50.58	0.69

On average, job expectation is 53.05 for females and 50.58 for males. As German schools are usually coeducational and as gender is not used as an explicit stratification variable, the expected value of the covariance might differ from 0.

To compute the standard error by gender, it is necessary to compute the mean estimate for each of the 80 replicate weights. Table 11.2 presents the mean estimates by gender for 80 replicate weights.

Table 11.2
Mean estimates for the final and 80 replicate weights by gender (PISA 2003)

Weight	Mean estimate		Weight	Mean estimate	
	Females	Males		Females	Males
Final weight	**53.05**	**50.58**			
Replicate 1	53.29	50.69	Replicate 41	52.69	50.55
Replicate 2	53.16	50.53	Replicate 42	53.28	51.23
Replicate 3	53.16	50.45	Replicate 43	53.07	50.39
Replicate 4	53.30	50.70	Replicate 44	52.95	49.72
Replicate 5	52.79	50.28	Replicate 45	53.31	51.04
Replicate 6	53.14	50.76	Replicate 46	53.72	50.80
Replicate 7	53.04	50.36	Replicate 47	52.91	51.03
Replicate 8	52.97	50.11	Replicate 48	53.10	50.53
Replicate 9	53.28	51.37	Replicate 49	53.05	50.81
Replicate 10	53.01	50.55	Replicate 50	53.79	50.90
Replicate 11	53.26	50.70	Replicate 51	52.65	50.15
Replicate 12	53.16	49.86	Replicate 52	53.30	50.45
Replicate 13	52.81	50.94	Replicate 53	52.68	50.12
Replicate 14	53.21	50.71	Replicate 54	52.74	50.01
Replicate 15	53.39	50.23	Replicate 55	53.50	50.11
Replicate 16	53.06	50.46	Replicate 56	52.54	50.58
Replicate 17	53.34	50.48	Replicate 57	53.31	51.03
Replicate 18	52.71	50.42	Replicate 58	53.13	50.34
Replicate 19	53.18	50.87	Replicate 59	52.72	50.37
Replicate 20	52.82	50.44	Replicate 60	53.49	51.43
Replicate 21	53.36	50.74	Replicate 61	53.13	50.71
Replicate 22	53.15	50.72	Replicate 62	53.61	51.27
Replicate 23	53.24	50.65	Replicate 63	52.74	50.15
Replicate 24	52.68	50.51	Replicate 64	53.19	50.25
Replicate 25	52.76	50.44	Replicate 65	53.28	51.04
Replicate 26	52.79	50.43	Replicate 66	52.91	50.94
Replicate 27	53.01	50.58	Replicate 67	53.25	50.85
Replicate 28	53.24	50.12	Replicate 68	53.12	50.74
Replicate 29	52.86	50.68	Replicate 69	53.08	50.31
Replicate 30	52.85	50.02	Replicate 70	52.92	50.44
Replicate 31	52.90	50.85	Replicate 71	53.35	50.63
Replicate 32	53.25	50.60	Replicate 72	53.25	50.75
Replicate 33	53.32	50.54	Replicate 73	52.54	50.42
Replicate 34	52.42	50.55	Replicate 74	52.58	50.20
Replicate 35	52.91	50.72	Replicate 75	52.49	49.75
Replicate 36	53.06	50.36	Replicate 76	52.98	50.96
Replicate 37	52.67	50.73	Replicate 77	53.04	50.24
Replicate 38	53.36	50.16	Replicate 78	53.30	50.44
Replicate 39	52.57	50.36	Replicate 79	52.93	50.36
Replicate 40	53.07	50.58	Replicate 80	52.98	50.76

The final difference estimate will be the difference between the two final estimates, *i.e.* 53.05 – 50.58 = 2.47.

The procedure to estimate the final standard error is straightforward. It is similar to the procedure described in Chapter 7, except that θ is now a difference, and not a mean or a regression coefficient. The different steps are:

- The difference in the means between females and males is computed for each of the 80 replicates.
- Each of the 80 difference estimates is compared with the final difference estimate, then squared.
- The sum of the square is computed then divided by 20 to obtain the sampling variance on the difference.
- The standard error is the square root of the sampling variance.

These different steps can be summarised as:

$$\sigma_{(\hat{\theta})} = \sqrt{\frac{1}{20}\sum_{i=1}^{80}(\hat{\theta}_{(i)} - \hat{\theta})^2} \text{ with } \theta \text{ being a difference.}$$

Concretely:

- For the first replicate, the difference between the female mean estimate and the male mean estimate is equal to (53.29 – 50.69) = 2.60. For the second replicate, the difference estimate will be equal to (53.16 – 50.53) = 2.63 and so on for the 80 replicates. All these difference estimates are presented in Table 11.3.
- Each of the 80 replicate difference estimates is compared with the final difference estimate and this difference is squared. For the first replicate, it will be $(2.60 - 2.47)^2 = 0.0164$. For the second replicate, it will be $(2.63 - 2.47)^2 = 0.0258$. These squared differences are also presented in Table 11.3.
- These squared differences are summed. This sum is equal to (0.0164 + 0.0258 +….. + 0.0641) = 9.7360. The sampling variance on the difference is therefore equal to $\frac{9.7360}{20} = 0.4868$.
- The standard error is equal to the square root of 0.4868, *i.e.* 0.6977.

As $\frac{2.47}{0.6977}$ is greater than 1.96, job expectations for females are statistically greater than job expectations for males in Germany.

If the researcher had considered these two German subsamples as independent, then s/he would have obtained the following for the standard error on this difference

$$\sigma_{(\hat{\theta}_i - \hat{\theta}_j)} = \sqrt{\sigma^2_{(\hat{\theta}_i)} + \sigma^2_{(\hat{\theta}_j)}} = \sqrt{(0.57)^2 + (0.69)^2} = 0.895$$

In this particular case, the difference between the unbiased estimate of the standard error (*i.e.* 0.698) and the biased estimate of the standard error (*i.e.* 0.895) is quite small. The difference between the biased and unbiased estimates of the standard error, however, can be substantial, as shown later in this chapter.

A SPSS® macro of MCR_SE_DIFF has been developed for the computation of standard errors on differences. Box 11.2 presents the SPSS® syntax for running this macro. Table 11.4 presents the structure of the output data file.

Table 11.3

Difference in estimates for the final weight and 80 replicate weights between females and males (PISA 2003)

Weight	Difference between females and males (females – males)	Squared difference between the replicate and the final estimates	Weight	Difference between females and males (females – males)	Squared difference between the replicate and the final estimates
Final weight	**2.47**				
Replicate 1	2.60	0.0164	Replicate 41	2.14	0.1079
Replicate 2	2.63	0.0258	Replicate 42	2.05	0.1789
Replicate 3	2.72	0.0599	Replicate 43	2.68	0.0440
Replicate 4	2.61	0.0180	Replicate 44	3.23	0.5727
Replicate 5	2.51	0.0011	Replicate 45	2.28	0.0373
Replicate 6	2.39	0.0067	Replicate 46	2.92	0.2038
Replicate 7	2.68	0.0450	Replicate 47	1.88	0.3488
Replicate 8	2.86	0.1483	Replicate 48	2.56	0.0084
Replicate 9	1.92	0.3085	Replicate 49	2.23	0.0567
Replicate 10	2.46	0.0002	Replicate 50	2.89	0.1768
Replicate 11	2.57	0.0089	Replicate 51	2.49	0.0004
Replicate 12	3.30	0.6832	Replicate 52	2.85	0.1440
Replicate 13	1.87	0.3620	Replicate 53	2.56	0.0072
Replicate 14	2.50	0.0009	Replicate 54	2.73	0.0667
Replicate 15	3.16	0.4756	Replicate 55	3.39	0.8520
Replicate 16	2.60	0.0173	Replicate 56	1.96	0.2631
Replicate 17	2.87	0.1577	Replicate 57	2.28	0.0351
Replicate 18	2.29	0.0327	Replicate 58	2.79	0.1017
Replicate 19	2.31	0.0269	Replicate 59	2.35	0.0158
Replicate 20	2.38	0.0078	Replicate 60	2.05	0.1749
Replicate 21	2.62	0.0221	Replicate 61	2.42	0.0027
Replicate 22	2.43	0.0014	Replicate 62	2.34	0.0164
Replicate 23	2.59	0.0142	Replicate 63	2.59	0.0137
Replicate 24	2.17	0.0901	Replicate 64	2.94	0.2230
Replicate 25	2.32	0.0227	Replicate 65	2.24	0.0539
Replicate 26	2.36	0.0132	Replicate 66	1.97	0.2524
Replicate 27	2.43	0.0015	Replicate 67	2.40	0.0050
Replicate 28	3.12	0.4225	Replicate 68	2.38	0.0089
Replicate 29	2.18	0.0844	Replicate 69	2.76	0.0848
Replicate 30	2.84	0.1333	Replicate 70	2.48	0.0002
Replicate 31	2.06	0.1709	Replicate 71	2.72	0.0609
Replicate 32	2.65	0.0312	Replicate 72	2.50	0.0006
Replicate 33	2.78	0.0970	Replicate 73	2.12	0.1217
Replicate 34	1.87	0.3611	Replicate 74	2.39	0.0073
Replicate 35	2.19	0.0809	Replicate 75	2.73	0.0693
Replicate 36	2.69	0.0490	Replicate 76	2.02	0.2031
Replicate 37	1.94	0.2825	Replicate 77	2.80	0.1058
Replicate 38	3.20	0.5355	Replicate 78	2.86	0.1519
Replicate 39	2.21	0.0683	Replicate 79	2.57	0.0091
Replicate 40	2.48	0.0001	Replicate 80	2.22	0.0641
			Sum of squared differences		**9.736**

Box 11.2 **SPSS® macro for computing standard errors on differences (*e.g.* PISA 2003)**

```
INSERT FILE=“C:\PISA\MACRO\MCR_SE_DIFF.SPS”.
SET MPRINT=YES.
DIFnoPV DEP=BSMJ /
        STAT=MEAN /
        COMPARE=ST03Q01 /
        CATEG=1 2/
        INFILE=“C:\TEMP\DEU1.SAV”/
        WITHIN=CNT/                  /*default=NOWITHIN*/
        WGT=W_FSTUWT/                /*default=W_FSTUWT*/
        RWGT=W_FSTR/                 /*default=W_FSTR  */
        CONS=0.05/                   /*default=0.05    */
        NREP=80/.                    /*default=80      */
```

Beside the arguments common to most SPSS® macros, four other arguments have to be specified:

1. The DEP argument informs the macro of the numerical variable on which a mean or a standard deviation will be computed per value of a categorical variable. In this example, VAR equals BSMJ.
2. The COMPARE argument specifies the categorical variables on which the contrasts will be based.
3. The CATEG argument specifies the values of the categorical variables for which contrasts are required. As gender has only two categories, denoted 1 and 2, CATEG is set as “1 2”. If a categorical variable has four categories and if these four categories are specified in CATEG statement, then the macro will compute the standard error on the difference between:
   - category 1 and category 2;
   - category 1 and category 3;
   - category 1 and category 4;
   - category 2 and category 3;
   - category 2 and category 4;
   - category 3 and category 4.

   If only categories 1 and 2 are specified, then only the contrast between 1 and 2 will be computed, regardless of the number of categories for this categorical variable.
4. The STAT argument specifies the required statistic. See Chapter 7 for available statistics.

Table 11.4
Output data file from Box 11.2

CNT	CONTRAST	STAT	SESTAT
DEU	ST03Q01.1_2	2.47	0.6977
Total	ST03Q01.1_2	2.47	0.6977

The row denoted “Total” in Table 11.4 presents the statistic on the whole dataset regardless of the grouping variables.

For dichotomous variables, the standard error on the difference can also be computed by a regression model. Box 11.3 presents the SPSS® syntax to compute a gender difference in BSMJ and its standard error by using the MCR_SE_REG macro. Before running the syntax in Box 11.3, the gender variable of ST03Q01 needs to be recoded into a new variable denoted GENDER, with females being 1 and males being 0. Table 11.5 presents the structure of the output data file.

Box 11.3 **Alternative SPSS® macro for computing the standard error on a difference for a dichotomous variable (*e.g.* PISA 2003)**

```
GET FILE=“C:\PISA\2003\DATA\INT_STUI_2003.SAV”.
SELECT IF (CNT=“DEU”).
SELECT IF NOT (ST03Q01=8).
RECODE ST03Q01 (1=1)(2=0) INTO GENDER.
SAVE OUTFILE=“C:\TEMP\DEU2.SAV”.

INSERT FILE=“C:\PISA\MACRO\MCR_SE_REG.SPS”.

SET MPRINT=YES.
REGnoPV IND=GENDER/
        DEP=BSMJ/
        GRP=CNT/                        /* is default */
        INFILE=“C:\TEMP\DEU2.SAV”/.
```

Table 11.5

Output data file from Box 11.3

CNT	Ind	STAT	SE
DEU	INTERCEPT	50.58	0.69
DEU	GENDER	2.47	0.70
DEU	R_SQUARE	0.01	0.00

The difference estimate and its respective standard error are equal to the regression coefficient estimate and its standard error. For polytomous categorical variables, the use of the regression macro would require the recoding of the categorical variables into *h*-1 dichotomous variables, with *h* being equal to the number of categories. Further, the regression macro will compare each category with the reference category, while the macro MCR_SE_DIFF will provide all contrasts.

THE STANDARD ERROR ON A DIFFERENCE WITH PLAUSIBLE VALUES

The procedure for computing the standard error on a difference that involves plausible values consists of:

- Using each plausible value and for the final and 80 replicate weights, the requested statistic (*e.g.* a mean) has to be computed per value of the categorical variable.
- Computing, per contrast, per plausible value and per replicate weight, the difference between the two categories. There will be 405 difference estimates. Table 11.6 presents the structure of these 405 differences.
- A final difference estimate equal to the average of the five difference estimates.
- Computing, per plausible value, the sampling variance by comparing the final difference estimate with the 80 replicate estimates.
- A final sampling variance being equal to the average of the five sampling variances.
- Computing imputation variance, also denoted measurement error variance.
- Combining the sampling variance and the imputation variance to obtain the final error variance.
- A standard error being equal to the square root of the error variance.

Table 11.6

Gender difference estimates and their respective sampling variances on the mathematics scale (PISA 2003)

Weight	PV1	PV2	PV3	PV4	PV5
Final	**-8.94**	**-9.40**	**-8.96**	**-7.46**	**-10.12**
Replicate 1	-9.64	-10.05	-10.29	-8.74	-11.45
....	....	....	....	....	....
....	....	....	....	....	....
Replicate 80	-8.56	-8.52	-8.85	-7.70	-9.84
Sampling variance	**$(4.11)^2$**	**$(4.36)^2$**	**$(4.10)^2$**	**$(4.31)^2$**	**$(4.28)^2$**

Note: PV = plausible value.

A SPSS® macro has been developed to compute standard errors on differences that involve plausible values. Box 11.4 provides the SPSS® syntax. In this example, the standard error on the difference in performance in mathematics between males and females is computed. Table 11.7 presents the structure of the output data file.

Box 11.4 **SPSS® syntax for computing standard errors on differences which involve PVs (*e.g.* PISA 2003)**

```
INSERT FILE=“C:\PISA\MACRO\MCR_SE_DIFF_PV.SPS”.

SET MPRINT=YES.

DIF_PV DEP=PV1MATH PV2MATH PV3MATH PV4MATH PV5MATH/
       STAT=MEAN/
       COMPARE=ST03Q01/
       CATEG= 1 2/
       INFILE=“C:\TEMP\DEU1.SAV”/
       WITHIN = CNT/                  /*default=NOWITHIN*/
       WGT=W_FSTUWT/                  /*default=W_FSTUWT*/
       RWGT=W_FSTR/                   /*default=W_FSTR   */
       CONS=0.05/                     /*default=0.05     */
       NREP=80/.                      /*default=80       */
```

As the absolute value of the ratio between the difference estimate and its respective standard error is greater than 1.96, the null hypothesis is rejected. Thus females perform on average lower than males in Germany in mathematics. These results might also be obtained through the regression macro for plausible values.

Table 11.7
Output data file from Box 11.4

CNT	ST03Q01	STAT	SE
DEU	1-2	–8.98	4.37
Total	1-2	–8.98	4.37

Table 11.8 presents the gender difference in mean performance in mathematics for all OECD countries in PISA 2003, as well as the unbiased standard errors and the biased standard errors.

In nearly all countries, the unbiased standard error is smaller than the biased standard error, reflecting a positive covariance between the two sampling distributions. In a few countries, the difference between the two standard errors is small, but it is substantial for some other countries, such as Greece and Turkey.

These two SPSS® macros can also be used for other statistical parameters, such as variances or standard deviations. Table 11.9 presents the gender difference in the mean science performance and in the standard deviation for science performance in PISA 2006.

Surprisingly, males and females perform differently in only 8 countries out of 30. In two countries, *i.e.* Turkey and Greece, females outperform males while in Denmark, Luxembourg, Mexico, Netherlands, Switzerland and the United Kingdom, males outperforms females. On the other hand, in 23 countries, the standard deviation of the science performance for females is significantly smaller than the standard deviation of the science performance for males.

Table 11.8
Gender differences on the mathematics scale, unbiased standard errors and biased standard errors (PISA 2003)

Country	Mean difference (females – males)	Unbiased standard error	Biased standard error
AUS	-5.34	3.75	4.04
AUT	-7.57	4.40	5.59
BEL	-7.51	4.81	4.69
CAN	-11.17	2.13	2.78
CHE	-16.63	4.87	5.98
CZE	-14.97	5.08	6.11
DEU	-8.98	4.37	5.59
DNK	-16.58	3.20	4.50
ESP	-8.86	2.98	4.02
FIN	-7.41	2.67	3.24
FRA	-8.51	4.15	4.60
GBR	-6.66	4.90	4.84
GRC	-19.40	3.63	6.11
HUN	-7.79	3.54	4.69
IRL	-14.81	4.19	4.54
ISL	15.41	3.46	3.15
ITA	-17.83	5.89	5.96
JPN	-8.42	5.89	7.04
KOR	-23.41	6.77	6.90
LUX	-17.17	2.81	2.40
MEX	-10.90	3.94	5.91
NLD	-5.12	4.29	5.36
NOR	-6.22	3.21	4.04
NZL	-14.48	3.90	4.23
POL	-5.59	3.14	4.18
PRT	-12.25	3.31	5.41
SVK	-18.66	3.65	5.30
SWE	-6.53	3.27	4.30
TUR	-15.13	6.16	10.33
USA	-6.25	2.89	4.65

Table 11.9
Gender differences in mean science performance and in standard deviation for science performance (PISA 2006)

	Difference in mean (females – males)		Difference in standard deviation (females – males)	
	Difference	S.E.	Difference	S.E.
AUS	-0.05	3.76	-7.29	1.85
AUT	-7.53	4.91	0.40	3.56
BEL	-0.75	4.13	-6.81	2.54
CAN	-4.07	2.19	-5.90	1.62
CHE	-5.56	2.67	-0.67	1.67
CZE	-4.82	5.64	4.09	2.79
DEU	-7.14	3.71	-5.92	2.22
DNK	-8.93	3.24	-2.89	2.04
ESP	-4.36	2.36	-6.13	1.66
FIN	3.10	2.88	-7.87	1.93
FRA	-2.64	4.03	-8.87	2.72
GBR	-10.06	3.44	-9.45	2.19
GRC	11.41	4.68	-12.93	2.92
HUN	-6.48	4.17	-8.45	2.88
IRL	0.40	4.31	-7.07	2.12
ISL	6.17	3.44	-7.46	2.24
ITA	-3.05	3.53	-9.00	2.11
JPN	-3.26	7.40	-8.74	3.26
KOR	1.86	5.55	-7.43	2.67
LUX	-9.34	2.93	-8.71	2.15
MEX	-6.66	2.19	-3.14	1.76
NLD	-7.20	3.03	-2.80	2.35
NOR	4.37	3.39	-9.31	2.36
NZL	3.75	5.22	-8.57	2.42
POL	-3.38	2.48	-7.00	1.77
PRT	-5.04	3.33	-4.89	2.06
SVK	-6.23	4.73	-6.07	2.87
SWE	-1.28	2.97	-4.93	2.86
TUR	11.93	4.12	-4.64	2.25
USA	-0.58	3.51	-7.13	2.44

Comparisons of regression coefficients might also interest researchers or policy makers. For instance, does the influence of a student's socio-economic background on his/her performance depend on a student's gender? A regression model on the male subsample and another one on the female subsample will provide the regression coefficients but it will be impossible to compute the significance level of their difference, as the two samples are not independent. This test can, however, be easily implemented by modelling an interaction. Box 11.5 presents the SPSS® syntax for testing this interaction.

Question ST04Q01 is recoded into a new variable denoted MALE, with males being 1 and females being 0. A second variable, denoted INTER, is computed by multiplying MALE with HISEI. The INTER variable will be equal to 0 for all females and to HISEI for all males.

Box 11.5 **SPSS® syntax for computing standard errors on differences that involve PVs (*e.g.* PISA 2006)**

```
GET FILE=“C:\PISA\2006\DATA\INT_STU06_DEC07.SAV”.
SELECT IF (CNT=“BEL”).
SELECT IF NOT (ST04Q01=8).
RECODE ST04Q01 (1=0)(2=1) INTO MALES.
COMPUTE INTER=MALES*HISEI.
SAVE OUTFILE=“C:\TEMP\BEL.SAV”.

INSERT FILE=“C:\PISA\MACRO\MCR_SE_REG_PV.SPS”.

SET MPRINT=YES.

*** MODEL 1 ***.

REG_PV INFILE=“C:\TEMP\BEL.SAV”/
       IND=HISEI/
       DEP=PV1SCIE PV2SCIE PV3SCIE PV4SCIE PV5SCIE /
       GRP=CNT MALES/.

*** MODEL 2 ***.

REG_PV INFILE=“C:\TEMP\BEL.SAV”/
       IND=HISEI MALES INTER/
       DEP=PV1SCIE PV2SCIE PV3SCIE PV4SCIE PV5SCIE/
       GRP=CNT/.
```

Table 11.10 presents the regression coefficients for the male subsample regression and the female subsample regression (*e.g.* Model 1) as well as the regression coefficients for the model including males and females altogether with the interaction (*e.g.* Model 2). Standard errors are also provided.

Table 11.10
Regression coefficient of HISEI on the science performance for different models (PISA 2006)

Models	Sample	Variables	Estimates	S.E.
Model 1	Females	Intercept	405.13	7.32
		HISEI	2.21	0.13
	Males	Intercept	401.90	5.89
		HISEI	2.27	0.12
Model 2	All	Intercept	405.13	7.32
		MALE	-3.23	7.46
		HISEI	2.21	0.13
		INTER	0.06	0.15

The model with the interaction returns values for the intercept and for the HISEI regression coefficient that are identical to the corresponding estimates on the subsample of females. The regression coefficient of INTER is equal to the difference between the two HISEI regression coefficients computed on both subsamples. The standard error for the INTER regression coefficient indicates that the null hypothesis cannot be rejected. Therefore, the influence of a student's social background on his/her performance does not depend on student gender.

MULTIPLE COMPARISONS

In Chapter 4, it was noted that every statistical inference is associated with what is usually called a type I error. This error represents the risk of rejecting a null hypothesis that is true.

Let's suppose that at the population level, there is no difference in the mathematics performance between males and females. A sample is drawn and the gender difference in mathematics performance is computed. As this difference is based on a sample, a standard error on the difference has to be computed. If the standardised difference (*i.e.* the gender difference divided by its standard error) is less than –1.96 or greater than 1.96, that difference would be reported as significant. In fact, there are 5 chances out of 100 to observe a standardised difference lower than –1.96 or higher than 1.96 and still have the null hypothesis true. In other words, there are 5 chances out of 100 to reject the null hypothesis, when there is no true gender difference in the population.

If 100 countries are participating in the international survey and if the gender difference is computed for each of them, then it is statistically expected to report 5 of the 100 gender differences as significant, when there are no true differences at the population level.

For every country, the type I error is set at 0.05. For two countries, as countries are independent samples, the probability of not making a type I error, *i.e.* accepting both null hypotheses, is equal to 0.9025 (0.95 times 0.95) (Table 11.11).

Table 11.11
Cross tabulation of the different probabilities

		Country A	
		0.05	0.95
Country B	0.05	0.0025	0.0475
	0.95	0.0475	0.9025

This statistical issue is even more amplified for tables of multiple comparisons of achievement. Suppose that the means of three countries need to be compared. This will involve three tests: Country A versus Country B; Country A versus Country C; and Country B versus Country C. The probability of not making a type I error is therefore equal to:

$$(1-\alpha)(1-\alpha)(1-\alpha)=(1-\alpha)^3$$

Broadly speaking, if X comparisons are tested, then the probability of not making a type I error is equal to

$$(1-\alpha)^X$$

Dunn (1961) developed a general procedure that is appropriate for testing a set of *a priori* hypotheses, while controlling the probability of making a type I error. It consists of adjusting the value α. Precisely, the value α is divided by the number of comparisons and then its respective critical value is used.

In the case of three comparisons, the critical value for an $\alpha = 0.05$ will therefore be equal to 2.24 instead of 1.96. Indeed,

$$\frac{0.05}{3} = 0.01666$$

As the risk is shared by both tails of the sampling distribution, one has to find the z score that corresponds to the cumulative proportion of 0.008333. Consulting the cumulative function of the standardised normal distribution will return the value –2.24.

Nevertheless, the researcher still has to decide how many comparisons are involved. In PISA, it was decided that no correction of the critical value would be applied, except on multiple comparison tables.[1] Indeed, in many cases, readers are primarily interested in finding out whether a given value in a particular country is different from a second value in the same or another country, *e.g.* whether females in a country perform better than males in the same country. Therefore, as only one test is performed at a time, then no adjustment is required.

On the other hand, with multiple comparison tables, if the reader is interested in comparing the performance of one country with all other countries, the adjustment is required. For example, if one wants to compare the performance of Country 1 with all other countries, we will have the following comparisons: Country 1 versus Country 2; Country 1 versus Country 3; and Country 1 versus Country L. Therefore, the adjustment will be based on L-1 comparisons.

CONCLUSION

This chapter was devoted to the computation of standard errors on differences. After a description of the statistical issues for such estimates, the different steps for computing such standard errors were presented. The SPSS® macros to facilitate such computations were also described.

It was clearly stated that any comparison between countries does not require the estimation of the covariance. However, it is strongly advised that the covariance between the sampling distributions for any within-country comparisons should be estimated.

The two SPSS® macros can however be used for between-country comparisons. As the expected value of the covariance is equal to 0, in a particular case, one might get a small positive or negative estimated covariance. Therefore, the standard error returned by the SPSS® macro might be slightly different from the standard errors based only on the initial standard errors.

Finally, the correction of the critical value for multiple comparisons was discussed.

Note

---

[1] The Bonferroni adjustment was not presented in the PISA 2006 multiple comparison tables (OECD, 2007).

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

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

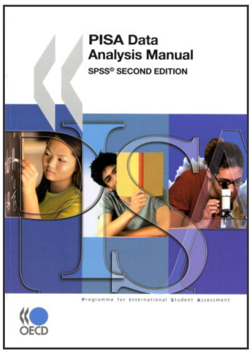
All syntaxes and macros in this manual can be copied from the PISA website (*www.pisa.oecd.org*). These macros were developed for SPSS 17.0. The 19 SPSS® macros presented in Chapter 17 need to be saved under "c:\pisa\macro\", before staring analysis. Each chapter of the manual contains a complete set 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), “Standard Error on a Difference”, in *PISA Data Analysis Manual: SPSS, Second Edition*, OECD Publishing, Paris.

DOI: https://doi.org/10.1787/9789264056275-12-en