Standard error of an estimated difference between countries when countries have different sample designs: issues and solutions

Lynn, Peter J.
Institute for Social and Economic Research
University of Essex, Wivenhoe Park
Colchester CO4 3SQ, UK
E-mail: plynn@essex.ac.uk

Kaminska, Olena
Institute for Social and Economic Research
University of Essex, Wivenhoe Park
Colchester CO4 3SQ, UK
E-mail: olena@essex.ac.uk

Abstract

Standard estimation procedures for complex sample designs assume the data relate to one population and arise from one sample design. However, in cross-country studies it is often more efficient and more practical to select respondents using different sampling strategies in different countries. As a result, there is a need for estimation procedures which correctly reflect this situation. Using data from an important cross-national study, the 2007 European Union Statistics of Income and Living Conditions (EU SILC), we identify different scenarios and suggest estimation procedures for each. Our main focus is the estimation of differences in means and proportions when only one of two countries has a clustered design. We also consider variation in the number of selected household members (all or one). Furthermore, we compare our estimation procedures with convenient alternative suboptimal approaches an analyst may take: either not taking any clustering into account or taking only households, but not higher level clustering, into account. Our results show that in a few situations the conclusion may be sensitive to the estimation procedure, mainly when the difference between countries is small but marginally significant.

Estimation of Between-Country Differences

Survey data are often used to estimate differences in parameters between countries. The parameters may be simple descriptive statistics such as means and proportions or analytic statistics such as differences in the intercept of a regression (coefficient of country dummies) or differences in the conditional association between two variables (coefficient of interaction between country dummy and *x*-variable). Such estimates of differences can be found in a wide variety of fields and contexts. However, when the data from different countries are generated using different sample designs, testing the difference is not straightforward. In this circumstance it is common to find that standard errors, and hence hypothesis tests, are estimated incorrectly.

In fact this issue is generalisable to any statistical comparison of multiple domains, where the design differs between domains. We use countries as our example of such domains as it is common for cross-national surveys to have variation in sample design between countries (Lynn et al 2007).

EU-SILC Data

For our study we use data from the European Union Statistics on Income and Living Conditions (EU-SILC) survey. EU-SILC has been carried out in all 27 EU member states since 2007 (some started earlier)

plus 4 non-member states (http://eusilc.notlong.com). Both cross-sectional and longitudinal data are collected, on income, poverty, social exclusion and other living conditions. Most items are collected through individual interviews with each adult in a household, though some items are collected through a household interview. In most countries the data are collected by means of a survey with a rotating panel design. Though the details of the design vary, a typical design involves a 4-wave rotation with annual interviews.

We use only cross-sectional data from 2007. We drop from our analysis a number of countries that either had not yet provided these data at the time the analysis was performed, or for whom the indicators of sample design parameters – which are crucial to our analysis – were either missing or did not correspond with the description of the design (and where these discrepancies could not be resolved). This left 19 countries for analysis: AT, BE, CY, CZ, EE, FI, FR, HU, IS, IT, LT, LU, LV, NL, PL, SE, SI, SK, UK.

Variables and Estimates

We estimate differences between pairs of countries in a number of descriptive parameters (means and proportions, including some subgroup means). For 5 estimates (listed in table 2) the units of analysis are households. For 15 estimates (listed in tables 3 and 4), the units of analysis are individuals. The weight applied to a sample unit is, for some countries, different in the two cases, depending on the design. Our analysis uses only design weights. No attempt was made to develop non-response adjustments to these weights. We did not utilize weights provided by Eurostat, but instead derived our own weights based on the documented description of the sample design in each country and, where relevant, the data item indicating the number of adults in the household.

Our objective is to estimate mis-specification effects, *meff* (Skinner 1989) under a range of scenarios. In all cases, we assume that weights are correctly specified in the analysis. We consider three likely forms of mis-specification: failing to take into account that samples are selected independently in each country, failing to take into account that the sample is clustered, and only partially taking into account that the sample is clustered (sub-optimal specification of clusters). In combination, this leads to five possible types of mis-specification (table 1). For each type of mis-specification, we estimate *meff* for each of 90 pairs of countries, specifically all the pairs which consist of one country with a clustered design and one with an unclustered design. (Of the 19 countries available for analysis, 10 had clustered designs and 9 had unclustered designs.) For household-level analysis, only mis-specification types 1, 2 and 3 are possible as partial consideration of clustering involves recognizing that individuals are clustered within households but not that the sample households are themselves clustered. As we have 5 household-level estimates and 15 individual-level ones, this leads to 8,100 estimates of *meff*.

Table 1: Design mis-specification scenarios

Mis-specification type	1	2	3	4	5
Ignore independence of samples	X	X		X	
Ignore clustering	X		X		
Only partially consider clustering				X	X

Estimation Procedures

We use the `svy` commands in Stata 11.0 to provide estimates that take into account aspects of the sample design. Similar approaches can be used in other software packages. We compare each mis-specified design with a correctly specified design, defined as follows:

- The independence of samples should be reflected through specification of the sample stratification. If strata are defined at sub-national level then the use of the stratum indicator will correctly signal that the sample was selected independently in each country, because countries will be aggregates of strata. But if there is no stratification within a particular country, it is necessary to specify the whole country as a single stratum. Thus, to compare a parameter between countries A and B where A has a stratified

design with 10 strata and B has an unstratified design, it is necessary first to derive a new stratum indicator which takes 11 values, 10 for country A and 1 for country B. In our data, there is no available indicator of stratum for countries with stratified designs. Thus, we simply treat countries as strata. With Stata, we specify this indicator in the `strata` option of the `svyset` command.

- Clustering should be reflected through specification of an indicator of Primary Sampling Unit (PSU). For countries with an unclustered design – such as a simple random sample or a systematic random sample of addresses or persons – the address (household) or person is the PSU. Thus, to compare a parameter between countries A and B where A has a clustered design and B has an unclustered design, it is necessary first to derive a new PSU indicator which indicates the clusters in country A and the households or persons in country B. We specify this indicator in the `psu` option of the `svyset` command.
- Variation in selection probabilities should be reflected through specification of design weights. This is done via the `pw` option of the `svyset` command.

Our Stata syntax for estimating a difference in mean value of the variable `var1` is as follows, where the variables `strata1`, `psu` and `weight1` are defined as in the three bullet points above:

```
svyset psu [pw=weight1], strata(strata1)
svy: mean var1 if cntry1==1 | cntry1==2, over(cntry1)
lincom [var1]1 - [var1]2
```

It can be seen that this correct estimation is very simple to implement once the design variables have been correctly derived.

Mis-specification type 1 (table 1) involves omitting the specification of strata and of psu.

Results

As described above, for each of 90 country pairs we carried out analysis for 5 household-level variables for each of 3 types of mis-specification and for 15 individual-level variables for each of 5 types of mis-specification. Overall we found that the *meff* is generally considerable when the clustering was not specified, whereas the effect of ignoring the independence of each national sample is negligible for most estimates. Thus, results for type 1 and type 3 mis-specification (see table 1) are very similar, as are results for types 4 and 5, while all 1,530 *meff*s for type 2 are in the range 0.98 – 1.00. Therefore, we present here only the results from mis-specification type 1.

For each of the five household variables, we present in table 2 the mean *meff* (across the 90 country pairs). These are in the range 0.70 – 0.90. However, we present also the minimum and maximum estimated *meff* for each variable and this shows that for specific pairwise comparisons *meff* can be as low as 0.07, meaning that the true variance could be 14 times the size of the estimated one if the design is mis-specified in this way and standard errors could be nearly 4 times the size of the estimated ones.

The final column of table 2 shows the number of comparisons, out of the 90, which change significance at the 0.05 level if the design is mis-specified. These are the cases where an apparently significant difference between countries ($P < 0.05$) is in fact an artifact of design mis-specification. Such cases represent 2% of all comparisons (9 out of 450).

In table 3 we present results for 12 of the 15 individual-level estimates, of which 6 are whole-sample, 3 are based on males only and 3 on females only. These 12 variables are all available for all individuals in each sample household, either because all individuals were interviewed or because only one person was interviewed but the information for other individuals was obtained from a population register. Mean *meff* (across the 90 country pairs) ranges from 0.38 for mean equivalised disposable income to 0.99 for the proportion of males who are economically active. This is a much greater range than we observed above for household-level variables, reflecting the larger intra-cluster correlation for individual variables due to the additional level of clustering (individuals within households) and the larger sample size per cluster. Failing to

correctly take clustering into account is therefore particularly problematic for individual-level estimation. Some *meff*s are very low indeed, with the smallest being 0.03 for a difference between two countries in mean equivalised disposable income, implying that standard errors could be under-estimated by a factor of 6. Overall, 36 of the 1,080 comparisons (3.3%) appear significant (P<0.05) if the design is mis-specified in this way but not significant if correctly specified.

Table 2: Results for 5 household-level variables: mis-specification type 1 over 90 country-pairs

	$\overline{y_1 - y_2}$	$\overline{meff}$	$s.d.(meff)$	$min.(meff)$	$max.(meff)$	ΔSig
Income	19160.32	0.80	0.25	0.07	1.00	3
Capacity to afford holidays	0.25	0.71	0.20	0.33	0.96	2
Capacity to afford meals	0.12	0.81	0.14	0.54	0.99	2
Ability to make ends meet	0.06	0.83	0.15	0.43	0.99	1
Number of household members	0.28	0.87	0.11	0.55	1.00	1

Table 3: Results for 12 individual-level variables: mis-specification type 1 over 90 country-pairs

	$\overline{y_1 - y_2}$	$\overline{meff}$	$s.d.(meff)$	$min.(meff)$	$max.(meff)$	ΔSig
Gender	0.024	0.44	0.06	0.33	0.58	9
Age	2.06	0.64	0.09	0.39	0.78	4
Equivalised disposable income	11,737	0.38	0.14	0.03	0.63	2
Education (ISCED)	0.099	0.55	0.19	0.06	0.81	4
Economic activity	0.070	0.76	0.16	0.27	0.97	3
Employment	0.044	0.73	0.15	0.41	1.01	3
Education (males)	0.097	0.74	0.22	0.09	0.97	8
Econ. activity (males)	0.066	0.99	0.14	0.57	1.22	0
Employment (males)	0.039	0.81	0.11	0.62	1.00	2
Education (females)	0.112	0.77	0.23	0.13	1.00	1
Econ. Act. (females)	0.075	0.94	0.18	0.37	1.14	0
Employm't (females)	0.052	0.86	0.13	0.51	1.06	0

The remaining three individual-level estimates are based on a variable, general health status, which is available only for interviewed individuals. In some countries (those with population registers from which certain EU-SILC variables can be obtained) only one person is interviewed in each household, so for estimates based on this variable some countries have a sample of individuals clustered within households, while other countries have individuals clustered within PSUs (if the selection of households was clustered) or completely unclustered. Thus, there are four possible situations when comparing a country with a clustered design with one with an unclustered design: a) both countries may have one individual observed per household, b) both may have all individuals observed per household, c) only the clustered country may have all observed, or d) only the unclustered country may have all observed. These four scenarios have potentially different implications for mis-specification so in table 4 we present results separately for each scenario.

It can be seen that *meff*s are modest when both countries interview only one person per household, but a

little more substantial when one of the countries interviews all persons. The largest *meff*s arise when both countries interview all person, as in this case an entire level of clustering is being ignored for both countries.

Table 4: Results for self-assessed general health (individual-level): mis-specification type 1

		$\overline{y_1 - y_2}$	$\overline{meff}$	$s.d.(meff)$	$min.(meff)$	$max.(meff)$	ΔSig
All individuals in both countries (48 comparisons)	All	0.071	0.72	0.06	0.61	0.85	2
	Men	0.060	0.91	0.07	0.69	1.06	0
	Women	0.080	0.89	0.06	0.77	0.98	0
One per household in both countries (6 comparisons)	All	0.057	0.95	0.01	0.93	0.97	0
	Men	0.050	0.98	0.01	0.97	0.99	0
	Women	0.063	0.97	0.03	0.94	1.00	0
All individuals in PSU country; one per household in non-PSU country (24 comparisons)	All	0.087	0.83	0.08	0.68	0.97	0
	Men	0.076	0.93	0.07	0.74	1.05	0
	Women	0.095	0.92	0.06	0.81	0.99	0
One per household in PSU country; all individuals in non-PSU country (12 comparisons)	All	0.071	0.83	0.03	0.77	0.89	0
	Men	0.063	0.96	0.02	0.93	0.98	0
	Women	0.079	0.95	0.03	0.92	1.00	0

Conclusions

Correctly specifying sample design is important. Standard errors can be seriously biased if the design is mis-specified in the ways discussed here, particularly if clustering is ignored or only partially taken into account. In testing whether differences between countries are significantly different from zero, this may lead to type I errors.

Correct specification can be easily achieved with standard software. It may be necessary to derive new indicators of strata and of PSUs, but once these are in place standard procedures can be used. A necessary prerequisite for correct specification is that indicators of sampling strata and PSUs are made available to the data analyst.

Acknowledgements

This research was carried out under the award “Analysis of Life Chances in Europe (ALICE), funded by the UK Economic and Social Research Council. The ALICE Principal Investigators were Richard Berthoud and Maria Iacovou. EU-SILC data was supplied by Eurostat. We are grateful to Vijay Verma for advice, to Alexandra Skew and Francesco Figari for helping to document and interpret the data, and to Alita Nandi and Steve Pudney for advice on Stata routines.

REFERENCES (RÉFERENCES)

Lynn, P., Gabler, S., Häder, S. & Laaksonen, S. (2007) Methods for achieving equivalence of samples in cross-national surveys. *Journal of Official Statistics* 23(1): 107-124

Skinner, C. J. (1989) Introduction to part A. In *Analysis of Complex Surveys*, ed. C. J. Skinner, D. Holt, and T. M. F. Smith, 23–58. New York: Wiley.