Survey Methodology
Part D

Fall 2015

Seppo Laaksonen

Missingness
Basic Survey Data Analysis

Handling unit missingness (nonresponse)

The methodology both for handling unit nonresponse and item nonresponse does not differ much from each other but what to do after that, it may differ more substantially. Thus the analysis itself in both cases is about as below:

1. To investigate the reasons for missing values.
2. To calculate the entire missingness rates, by reasons, and by domains (background variables).
3. To report and interpret the results, and publish the main points respectively.
4. To try to do everything better if possible in next surveys.

However, there are many differences in details as described next.

Handling unit missingness (nonresponse)

1. To investigate the reasons for missing values.

Reason	Unit nonresponse	Item nonresponse
Non-contact due to incorrect data	Possible	Not possible since this is concerned respondents
Inability to answer correctly	Due to general disability	Information difficult to get
Hard refusal	Don't participate at all. maybe in any survey	Not possible since this is concerned respondents
Soft refusal	Reply to most questions	Does not reply to all questions for various reasons, see below
Screening question	No problem	Second stage answers missing but can be used in analysis
Lost data	Possible	Should not be possible
Other or unknown reason	Possible	Possible

Handling unit missingness (nonresponse)

2. To calculate the entire missingness rates, by reasons, and by domains.

This can be made in both cases creating a response indicator so that its value = 0 for missing cases and = 1 for non-missing cases.

Then

- The respective rates can be calculated. also by domains that are available in the data set (e.g. gender, age group, region, education, industry class, socio-economic status).

and finally

- One or more response propensity models are estimated so that the response indicator is the dependent variable, and all possible domains or auxiliary variables are attempted as independent or explanatory variables.

Response propensity models

The model is most commonly a binary regression model in which the link function is either logit, probit, log-log or complementary log-log.

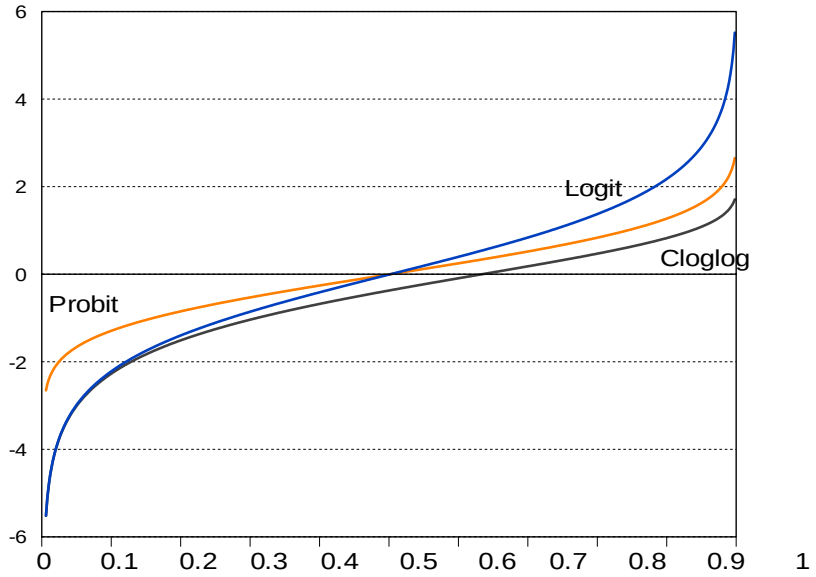
A second alternative is to build a classification tree model with similar variables, but we do not here go to details of this model. Instead. we present those about the two most common models, logit and probit.

They can be implemented as all statistical models, trying to find best explanatory variables for the response indicator, the interactios are possible to try as well.

The model estimates can be interpreted as usually but it is also good to calculate the estimated response probabilities that are called often propensity scores. We give some examples below.

Link functions of binary regression models

The graph does not include log-log that is a mirror curve for complementary log-log = Cloglog. Logit and probit are symmetric, whereas the others are asymmetric. The linear function may give unacceptable probabilities (negative or above one). You see that the curves are fairly linear within Interval (0.3. - 0.7) but far from the linearity in margins. Thus the linear function does not work at all in margin areas, if the probabilities are small or large.



Example of a fairly small unit response model with logit link

We have the following auxiliary variables, all categorical:

- Gender (2 categories)
- Agegroup (3)
- Region (7)
- Education (3).

We do not need to know the meta data of categories. The next page includes the estimates of the logistic regression model so that gender and age group are interacted.

Logistic regression for a unit response indicator

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter			DF	Estimate	Standard Error	Wald 95% Confidence Limits		Pr > Chi Sq
Intercept			1	0.4676	0.1883	0.0986	0.8367	0.0130
gender*agegroup	1	1	1	0.4061	0.1663	0.0800	0.7321	0.0146
gender*agegroup	1	2	1	0.3791	0.1461	0.0928	0.6654	0.0095
gender*agegroup	1	3	1	0.1465	0.1880	-0.2220	0.5150	0.4358
gender*agegroup	2	1	1	0.0111	0.1677	-0.3175	0.3398	0.9470
gender*agegroup	2	2	1	0.1226	0.1472	-0.1660	0.4112	0.4050
gender*agegroup	2	3	0	0.0000	0.0000	0.0000	0.0000	.
region	1		1	-0.2210	0.1448	-0.5049	0.0629	0.1271
region	2		1	0.1700	0.1855	-0.1935	0.5335	0.3594
region	3		1	0.2236	0.1491	-0.0687	0.5159	0.1337
region	4		1	0.3145	0.1575	0.0057	0.6232	0.0459
region	5		1	0.1457	0.1523	-0.1528	0.4442	0.3387
region	6		1	0.2351	0.1777	-0.1132	0.5833	0.1859
region	7		0	0.0000	0.0000	0.0000	0.0000	.
education	1		1	-0.9454	0.1048	-1.1507	-0.7400	<.0001
education	2		1	-0.6510	0.0970	-0.8411	-0.4610	<.0001
education	3		0	0.0000	0.0000	0.0000	0.0000	.

LR Statistics For Type 3 Analysis			
Source	DF	Chi-Square	Pr > ChiSq
gender* age-group	5	17.29	0.0040
region	6	25.63	0.0003
Edu-cation	2	87.61	<.0001

All are significant, Education most.
Continue the interpretation.

In order to get some understanding about the variation of propensity scores, the below table is calculated. The variations are not substantial but the scores for little educated (code = 1) are smaller as in all studies I have seen. We will later in the reweighting section try to adjust for this nuisance, or reduce the bias.

Analysis Variable : Predicted Value								
education		r	Mean	Minimum	10th Pctl	90th Pctl	Maximum	Coeff of Variation
1		974	0.46	0.33	0.37	0.54	0.56	13.1
2		1229	0.54	0.40	0.43	0.61	0.63	11.4
3		838	0.67	0.56	0.59	0.74	0.77	8.0

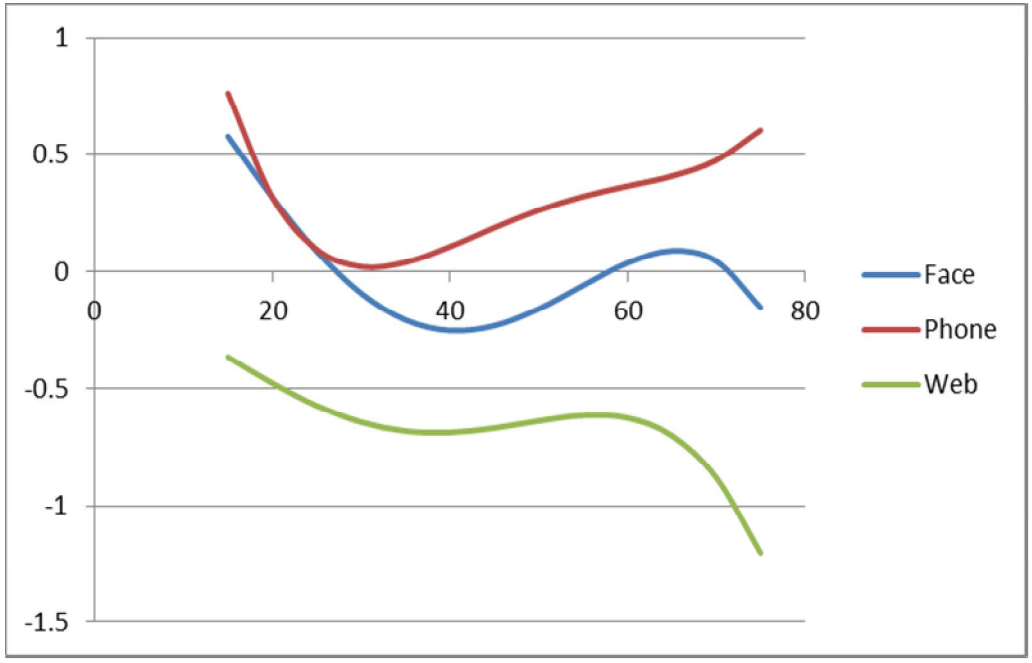
Example of a fairly broad unit response model with probit link

This example is from the Finnish security survey. We estimated the models with following explanatory (auxiliary) variables but many others were attempted:

- Mode (face-to-face, phone, web)
- Gender (male, female)
- Native language (Finnish, Russian, Other)
- Level of education (primary, secondary, tertiary, lower university, master, doctoral)
- Large region (Helsinki, Other southern towns, Other towns, Rural).
- Partnership (widowed. Single, recent marriage, medium marriage, old marriage, multi marriage)
- Children or not
- House size (one room, 2-3 rooms, more rooms).
- Age but this was made with polynomial transformation with four variables (age and its three powers). The next page shows the result.

Estimated effects of mode and age on response probabilities by age (x-axis), based on the probit model.

Note that this scale is good for understanding the differences correctly. the value 0 corresponds to the response rate = 50 per cent. A linear scale would give implausible influence.



We will come back to the missingness issues in imputation and reweighting which methods are just next methods when the missingness in its several meanings has been understood in this particular survey. Both methodologies can exploit propensity scores even though other alternatives exist.

Now we go to analysis part. starting with some points on aggregation that is always needed even in its simple forms, but the statistical model is often its final form.

Basic Aggregation

Micro data can be analyzed ordinarily using statistical tools, including frequencies, averages, medians, distributional statistics, and more demanding methods as statistical models. The outputs of such techniques can be more or less published as such if they are useful.

This kind of outputs are automatically aggregates. It is often good to save such outputs. It can be made manually, so that you type or copy most interesting results in another file such as EXCEL, but this is not always very nice to do. It is also possible to try to find more automatic tools.
All software packages give opportunity to this. Next I present examples of ESS 5 data with SAS and SPSS.

My example is concerned income that has been based as everything on interviewing so that the result is 10 categories that are deciles of income. Unfortunately, it is not possible to get income for everyone.

Aggregation example

The below is the SAS table for all ESS countries. There are three missing categories left.

Income	Frequency	Percent	Cumulative Frequency	Cumulative Percent
10th_Decile	3407	6.53	3407	6.53
1st_Decile	4439	8.51	7846	15.04
2nd_Decile	5037	9.65	12883	24.69
3rd_Decile	4675	8.96	17558	33.65
4th_Decile	4574	8.77	22132	42.42
5th_Decile	4353	8.34	26485	50.76
6th_Decile	4172	8.00	30657	58.76
7th_Decile	4031	7.73	34688	66.48
8th_Decile	3763	7.21	38451	73.69
9th_Decile	3235	6.20	41686	79.89
Dont_Know	6122	11.73	47808	91.63
Other_missing	122	0.23	47930	91.86
Refusal	4247	8.14	52177	100.00

Aggregation example. cont.

It would be nice to get some understanding what are those missing categories. One strategy is to look forward for variables without missingness or low missingness. I give an example in which two other variables are used:

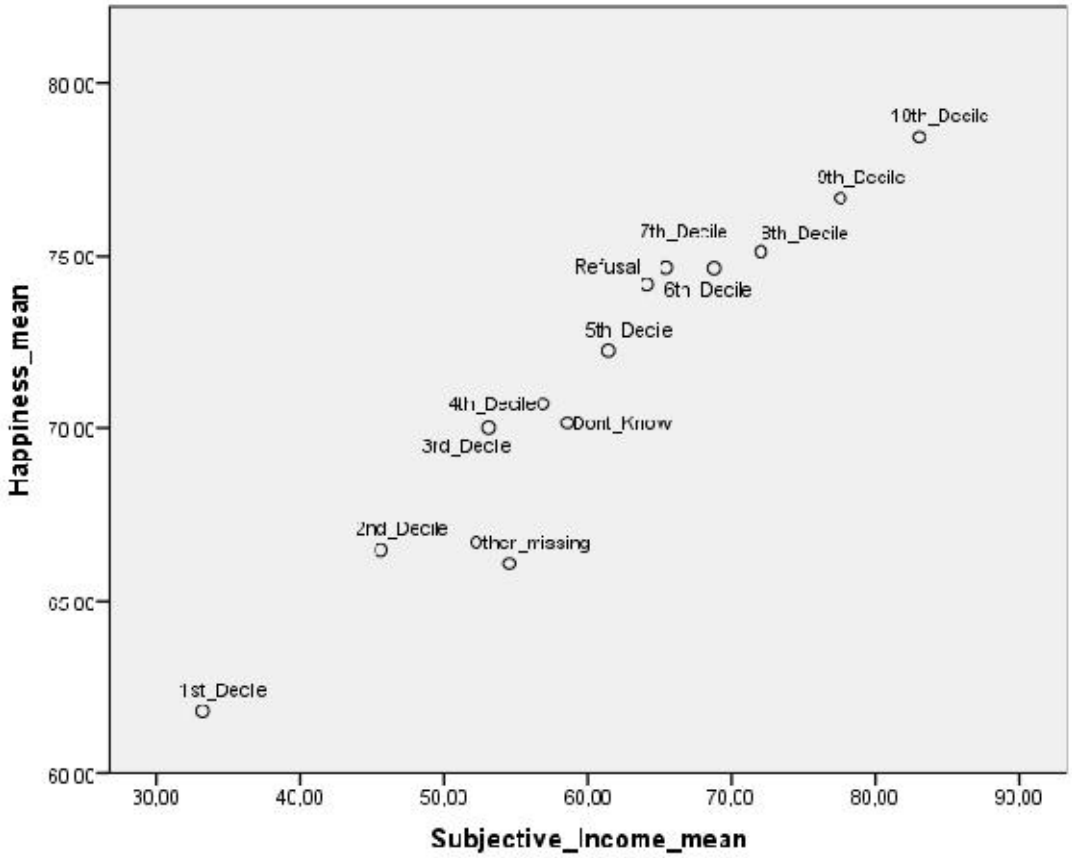
- Happiness that varies from 0 (very unhappy) to 10 (very happy)
- Subjective income that varies initially from 1 (very easy to live with such income) to 4 (much difficulties to to live).

I have rescaled both so that they vary from 0 (very bad) to 100 (very good). It is good to note that both variables have some missingness, but much less than in ordinary objective income. So, we can learn something.

Since I want to continue from basic results to graphical illustration. It is good to aggregate the output, i.e., to save it and continue then toward graphics.

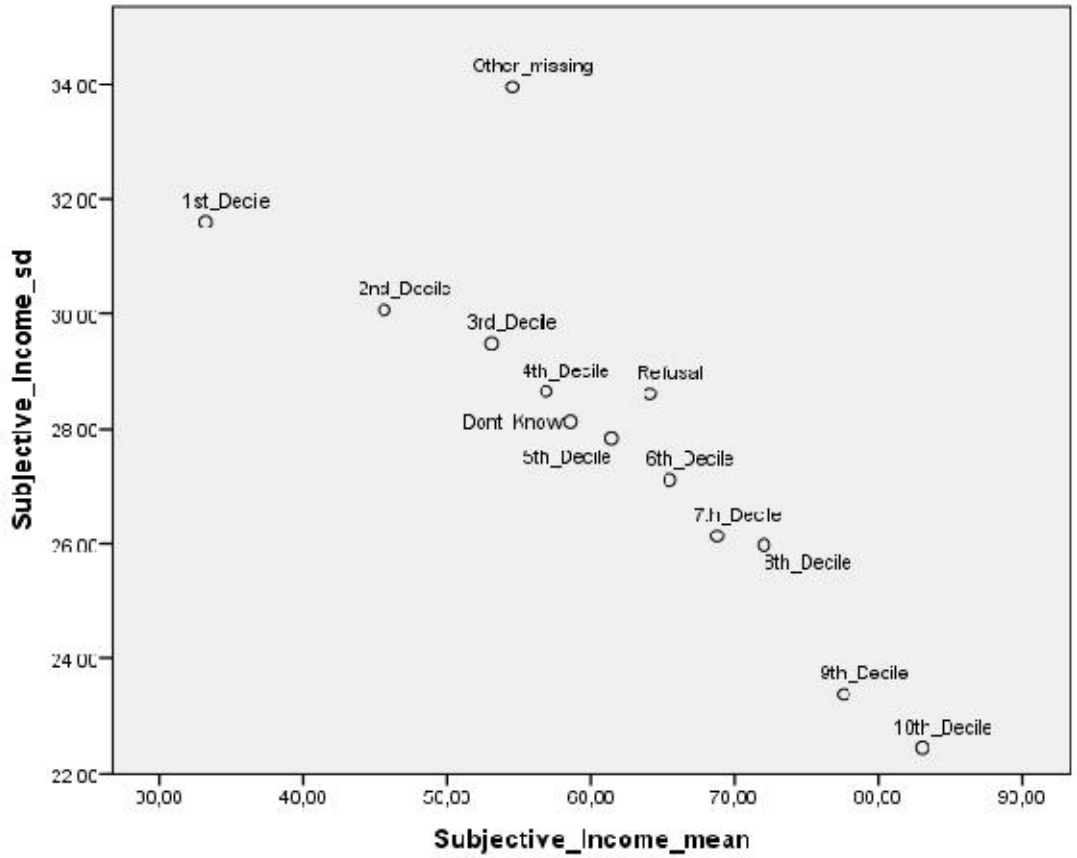
Scatter plot for all countries
X-axis = Subjective income average
Y-axis = Happiness average

You can see the missing codes. and look at which real income codes are close to those, and thus get some understanding about the average values of these missingness groups.



Scatter plot by subjective income and its standard deviation

In order to get further understanding, the second scatter is drawn so that Y-axis indicates the variability of each category by subjective income. Now we see a clear trend so that the subjective income very little varies in higher objective income groups. It thus is very clear that they can easily categorize themselves by subjective income. The small 'Other_missing' group is outlier.



Conclusions

The income example could be called as aggregate imputation since it gives some idea about objective income of missing categories. Naturally, this is not sufficient in most analysis, but if it is not possible to impute those well at micro level, it is something.

Another conclusion is that aggregation is a useful tool for reporting some results, since it also gives opportunity to continue toward statistical graphing. It is often useful to gather several results under the same file, with the same row aggregates, and this may be a good summary of the results.

Next we go to the survey data analysis of traditional methods.

Survey data analysis of traditional methods using the software package with this tool such as SPSS or SAS

In order to analyze survey data correctly, the so-called complex samples tools should be available. This tool is available for most common analyses fairly automatically. but not for very complex ones, for example, if several stages are there. We here include these basic cases. everything for cross-sectional data analysis as follows:

- Descriptive statistical analysis such as distributions and averages.
- Some statistical models such as regression and logistic regression.

Many special cases such as median. Gini-coefficient. percentiles or generalized mixed models are not considered here but such options are available.

Survey data analysis of traditional methods using the software package with this tool such as SPSS or SAS

The survey instruments thus should be included in the analysis. One instrument we have already applied, that is, the survey weights. These can be of two kinds.
- either relative weights or analysis weights whose sum = the number of the respondents
- or ordinary sampling weights whose sum is the size of the target population.

In most analyses it does not matter which of these weights to use, such as means or frequencies, but it is good to check which weight is correct for variances or similar estimates. If you are using complex survey tools, both weights should work.

The weights in general thus give the correct point estimates, thus such that can be generalized at the target population level.

Survey data analysis of traditional methods using the software package with this tool such as SPSS or SAS

It is however good to remind about the difference that the real sampling weights are working in each analysis since their sum corresponds to the target population.

The analysis weights are constructed to each specific population and their sum is the number of the respondents of this specific population that thus is often a country. Hence these weights are not working correctly in the estimations that are concerned several such specific populations, as several countries or their domains at the same time. The analysis weights thus are good for comparing the estimates of these specific populations. A statistical model such a linear or logistic regression works correctly with both weights, if this specific population is included as a control variable in the model, thus similarly to explanatory variables.

Survey data analysis of traditional methods using the software package with this tool such as SPSS or SAS

But in the case of the accuracy estimates, thus standard errors, confidence intervals or margins of error, and if the sampling design is not simple as simple random sampling or its conversion to the respondents with assumption MARS. it is needed to use the complex sample tools that have two additional instruments:

- If psu's are clusters
- If explicit stratification or post-stratification is used; the latter is considered in more details later.

The impact of stratification on standard errors is not often very big. but maybe in certain peculiar cases. Instead, if clusters are used, and their intra-class correlation is big, the impact of this instrument can be fairly big.

Survey data analysis of traditional methods using the software package with this tool such as SPSS or SAS

The survey instruments can be easily included in the data analysis so that each instrument has been included. thus

- The name of cluster variable
- The name of the stratification variable
- The name of the sampling weight variable.

In the SAS, these variables can be included into each analysis as Cluster <name>;. In the SPSS, it is first needed to create the sampling plan and when to use this in each analysis.

It should be noted that all standard methods and models are not yet available with Complex samples tools.

Example with stratified srs sampling of a business survey

Complex sample thus means that something complex has been used in sampling designing and in estimation respectively. This often means that psu is a cluster. Most surveys are not such complex samples cases, but traditional. The most common design is stratified srs that has been used very much both in human and business surveys. The difference between these is that sampling fractions may vary a lot in business surveys, thus more than in traditional human surveys. The is reason is that the data of large businesses are very important for most estimates. Hence a common strategy is that their sampling fraction is 100 per cent (e.g. above 500 employees), and the fraction is lowering dramatically to small businesses (e.g. 5-9 employees). Very small businesses, micro enterprises, are even excluded from a survey.
The response rate for largest businesses should be as high as possible, 100 per cent for largest, whereas less effort are paid to smallest businesses during the fieldwork.

Example with stratified srs sampling of a business survey

The sampling weights naturally vary substantially in such business surveys. Consequently, the use of the weights really matter. The below table is from a business survey that covers businesses of all sizes. It illustrates well how biased results may be obtained without weights.

Variable	Mean		Total	
	Unweighted	Weighted	Unweighted	Weighted
Turnover	33.9 000's	2.46 000's	85.2 billion	252.6 billion
Number of employees	3.47	283.0	13.9 million	29.0 million
Productivity	9.5	8.7	Not appropriate	Not appropriate

Example with stratified srs sampling of a survey of elderly people

It is also in some human surveys good to use unequal weights. The table below is from a survey elderly people, that is the target population is 61 to 90 years old residents of the country. It is here important to get enough respondents for older age groups as well even their number is relatively low. Hence the sampling fraction were relatively highest in these age groups. On the other hand, the fraction for males was a bit higher than for females for the same reason. The results clearly shows how biased are unweighted results but the effect is survey estimates is not usually as dramatic.

Age group	Population, count		Population, per cent	
	Unweighted	Weighted	Unweighted	Weighted
61-70	389	453907	37.5	60.0
71-80	455	249630	43.9	33.0
81-90	192	52469	18.5	6.9
All	1036	756006	100.0	100.0

Examples with complex samples tools

The first example from the Pisa

I calculated the averages of problem solving scores with four alternative options:

SRS = without any survey instruments. thus assuming that the sample is drawn by simple random sampling

Strata = stratification only has been taken into account of the three surveys instruments

Cluster = school cluster only has been included.

Weight = student weights are only included

All= all instruments (Strata. Cluster and Weight) are included in the analysis; these results thus are best ones.

The results do not vary much in means (mostly for Finland) but the standard errors are much higher in the last column. As seen, it is much due to clustering. Using design effects, DEFF's, it is easier to see the affects. See two pages forward.

The means and standard errors by four options

Analysis Variable : problem Problem Solving Scores											
		SRS		Strata		Cluster		Weight		All	
Country code 3-character	N Obs	Mean	Std Error	Mean	Std Error of Mean	Mean	Std Error of Mean	Mean	Std Error of Mean	Mean	Std Error of Mean
DEU	5001	508.4	1.4	508.4	1.4	508.4	5.1	508.7	1.4	508.7	4.8
EST	4779	516.8	1.2	516.8	1.2	516.8	3.4	515.0	1.3	515.0	3.0
FIN	8829	510.5	1.0	510.5	1.0	510.5	2.3	522.8	1.2	522.8	2.3
JPN	6351	552.2	1.0	552.2	1.0	552.2	3.7	552.2	1.0	552.2	3.7
KOR	5033	561.1	1.2	561.1	1.2	561.1	4.3	561.1	1.2	561.1	3.8
NOR	4686	502.5	1.4	502.5	1.4	502.5	3.7	503.3	1.5	503.3	3.7
RUS	5231	490.8	1.1	490.8	1.1	490.8	3.8	489.1	1.3	489.1	4.2
SWE	4736	492.1	1.3	492.1	1.3	492.1	3.3	490.7	1.3	490.7	3.3
USA	4978	509.2	1.2	509.2	1.2	509.2	4.0	507.9	1.4	507.9	4.6

Design effects based on the previous table

The design effect is the ratio of the two variances so that the denominator is the variance of simple random sampling. The variance is the square of the standard error. The DEFF's of this table thus are calculated from the standard errors of the previous table. This ensures that the DEFF's correct that is not necessarily the case in all software's. Now we see from the first column that all DEFF's of Germany, Russia and US are = 1; the reason is that the stratification variable is missing from the public Pisa data. All other stratification DEFF's are below 1, although very little. This is fairly general since it is a target in stratification, and even more important in post-stratification.

Country code 3-character	DEFF Stratifi-cation	DEFF Weights	DEFF Clusters	DEFF All
DEU	1	0,97	13,45	12,11
EST	0,937	1,19	7,98	6,35
FIN	0,972	1,38	5,61	5,25
JPN	0,978	1,04	14,28	14,31
KOR	0,919	1,02	11,94	9,4
NOR	0,999	1,06	6,85	6,78
RUS	1	1,21	11,46	13,38
SWE	0,985	1,06	6,22	6,3
USA	1	1,26	10,51	13,56

Design effects table, more comments

Now we see clearly that DEFF's due to clustering are highest, particularly in Japan and Germany. At contrast, the Finnish DEFF is fairly low. This indicates also how big differences between schools are in these countries.
DEFF due to unequal weights is highest for Finland. This thus means that the sampling fractions vary more than the other countries of the table.

Examples with complex samples tools

The second example is from the Finnish security survey that is based on stratified two-stage cluster sampling. Its clusters are areal groups created from postal codes and municipality codes. The stratum variables are region, gender and age group so that the sampling fraction for males is higher than for females. The reason of this is the target to get more accurate estimates for males.

The weights are fairly complex and their principles are told later.

The results show interesting mode effects that are much related to self-administered vs. interviewer-administered interviewing.

The cluster effects are much smaller than in the Pisa, because clusters are small areas, and there are in most cases any crime concentration within theses clusters. Crimes thus are more individual based, not areal based. Property crimes are exceptions to some extent, partner violence not at all.

Estimates and standard errors of some crime prevalence's. There are results both without weights 'Unweighted' and using 'Adjusted' weights. When comparing them you see that some times the difference is small. sometimes substantial.

	Estimates (standard errors). adjusted above and unweighted below		
	Face-to-face	Telephone	Web
Fear	Interviewer-administered items	Interviewer-administered items	Self-administered
Feeling unsafe	18.2 (1.3) 17.7 (1.2)	17.0 (0.5) 15.6 (0.5)	20.6 (0.7) 19.6 (0.6)
Fear of burglary	24.5 (1.4) 24.7 (1.3)	21.0 (0.6) 20.7 (0.6)	26.5 (0.8) 25.6 (0.7)
Fear of assault	20.0 (1.3) 18.8 (1.1)	20.3 (0.6) 18.7 (0.5)	26.1 (0.8) 25.3 (0.7)
Fear of family or friends	33.5 (1.6) 32.6 (1.4)	32.2 (0.7) 31.5 (0.6)	37.6 (0.9) 36.6 (0.8)

Estimates and standard errors of some crime prevalence's.

Property crimes	Face-to-face	Telephone	Web
Theft of car. if car owner	1.8 (0.8) 1.5 (0.7)	2.5 (0.4) 2.2 (0.4)	3.8 (0.7) 3.6 (0.6)
Damage to car	9.9 (1.7) 11.3 (1.8)	10.0 (0.8) 9.6 (0.7)	15.5 (1.5) 15.2 (1.2)
Theft of bicycle	16.9 (2.1) 16.4 (2.0)	17.1 (1.0) 15.8 (0.8)	22.6 (1.6) 21.9 (1.3)
Burglary at free-time residence	6.8 (2.4) 6.1 (2.1)	8.5 (1.2) 7.8 (1.1)	9.0 (1.6) 9.4 (1.6)
Robbery	2.0 (1.0) 1.6 (0.7)	2.5 (0.4) 2.6 (0.4)	3.5 (0.7) 3.4 (0.6)
Theft. other personal property	9.5 (1.8) 8.7 (1.5)	9.4 (0.7) 9.0 (0.7)	12.1 (1.2) 11.8 (1.0)
Burglary at home	2.3 (0.8) 2.7 (0.9)	2.7 (0.4) 2.9 (0.4)	5.0 (0.8) 5.3 (0.7)
Fraud	8.5 (1.6) 8.7 (1.5)	9.6 (0.7) 9.3 (0.7)	10.3 (1.0) 11.4 (1.0)

The differences between unweighted and well weighted results are not always big. This is also due to the sampling design that is fairly proportional, not peculiar that is usual in human surveys. In business surveys, the weights vary very substantially in most cases and the differences between such estimates as well. See an example below.

Estimates and standard errors of some crime prevalence's.

Violent crimes	Face-to-face	Telephone	Web
Physical or sexual violence. last 5 years	10.3 (1.7) 10.4 (1.6)	9.3 (0.8) 8.4 (0.6)	15.1 (1.3) 15.0 (1.1)
Physical or sexual violence. last 12 months	3.5 (1.0) 3.6 (1.0)	3.9 (0.5) 3.5 (0.4)	6.2 (0.9) 6.4 (0.8)
	Self-administered (CASI)	Interviewer-administered items	Self-administered
Sexual harassment. since the age of 15	52.8 (2.8) 46.4 (2.6)	38.0 (1.2) 33.2 (1.0)	45.3 (1.8) 42.0 (1.5)
Sexual harassment. last 12 months	21.1 (2.4) 19.9 (2.1)	10.1 (0.8) 8.4 (0.6)	22.2 (1.5) 20.6 (1.3)
Violence by stranger. since the age of 15	36.2 (2.8) 36.9 (2.5)	30.3 (1.1) 31.3 (1.0)	37.2 (1.8) 38.9 (1.6)
Violence by stranger. last 12 months	11.0 (1.8) 10.1 (1.6)	4.4 (0.5) 4.6 (0.5)	9.0 (1.0) 9.5 (0.9)
Violence by partner. since the age of 15	16.4 (3.1) 14.3 (2.5)	9.4 (1.0) 9.7 (1.0)	20.1 (2.1) 20.2 (1.8)
Violence by partner. last 12 months	4.1 (1.6) 3.7 (2.5)	1.7 (0.5) 1.5 (0.4)	4.0 (1.0) 4.4 (0.9)

These results also illustrate the effect of survey mode in confidential crimes in particular. You see that phone interviewing gives smaller crime rates than those with self-administered modes.