

**Development of a methodology for a long term strategy
on the Continuing Vocational Training Survey (CVTS)
CVTS3 M**

**Work-package 6:
Conceptual informatics framework**

Bundesinstitut für Berufsbildung (BIBB)

in cooperation with

Statistics Sweden
Statistics Finland
FÁS Training and Employment Authority
3s Research Laboratory

31 August 2005

Table of contents

- 1. Introduction..... 3
- 2. Quality components that are important to include in the quality reports..... 5
- 3. Sample Design and Allocation for CVTS3 – A Swedish example..... 10
- 4. Weighting, re-weighting and imputation..... 19

- Annex 1: Derivation of the formula, which is used for calculating sample sizes
in each stratum..... 41
- Annex 2: Excel application for calculating and allocation of sample sizes..... 43
- Annex 3: Algorithm for the calculation of the proportion of training enterprises
in a table where only the margins are known..... 55
- Annex 4: SAS-script with the algorithm for the calculation of the proportion of
training enterprises.....57

Introduction

In work-package 6 we treat the following issues:

- Quality components that are important to be included in the quality reports (chapter 2)
- Sample design and allocation for CVTS3 (chapter 3)
- Treatment of non-response (chapter 4)
- Rules and procedures for imputation (chapter 4)
- Weighting and reweighting (chapter 4).

Work-package 6 builds upon the work delivered in WP1, 2 and 3, and is related to WP5 delivered simultaneously. Chapters 2 through 4 and the related Annexes 1 through 4 have been elaborated by our Swedish partner, Statistics Sweden, and have been discussed with all partners in the consortium.

Chapter 2 deals with the components that need to be covered in the quality reports to enable users of the data to assess the quality of the data. In the ideal case the quality reports would give a full picture on relevance, accuracy, timeliness, accessibility, comparability, coherence and completeness of the data collected. This paper briefly discusses the aspect of comparability, including comparability over time. The focus is on accuracy, discussing sampling errors and non-sampling errors (coverage, measurement, processing, and non-response). The paper underlines the importance of measures reducing unit and item non-response (e.g. by re-contacting the enterprises), and argues that detailed description of the extent of non-response and of the methods used for compensating for both unit and item non-response is essential.

Chapter 3 deals with sample design and allocation and offers a tool (see annexes) for calculating the needed sample sizes contingent on different precision requirements (C-value). The sample design is simple random sampling within strata without replacement, with strata defined as cross-classification of 20 NACE categories with 3 size classes (division into substrata is allowed, and in small countries uniting similar strata might turn out to be necessary). The allocation principle is based on the expected response rates within each cell, and on the estimated proportion of training enterprises within each cell. It is suggested to use national CVTS2 results as a basis for estimating the latter proportion. The point made is that a good business register with up-to date information is most important and that the time between sample allocation and data collection must be short.

Chapter 4 (Weighting, (re-)weighting and imputation) starts with a discussion of the problems related to non-response: By reducing the sample size non-response causes less precise estimates. More essential, however, is the introduction of bias. Bias tends to increase with the rate of non-response; hence re-contacts with the enterprises are important to reduce the non-response rates as much as possible. A study of the Swedish CVTS2 data shows that high non-response rate causes uncertain survey estimates, and problems are increasing with the amount of non-response.

For treatment of item non-response, imputation of quantitative variables is recommended. In the paper rules for imputation are established, i.e. the cases when imputation is allowed, when it is not allowed (because of lack of information), and the rules for treating enterprises as unit non-response because of too many missing main variables. Based on a simulation study with Swedish data, the impact on the estimates is discussed. The paper discusses methods of imputation, and the effects on the uncertainty of estimates. As far as possible at the time of

writing the paper, detailed imputation methods for the quantitative variables are described, using ratio imputation within specific groups based on strata and, if applicable, on a variable collected in the questionnaire. The consortium does not recommend imputing qualitative variables. Contrary to quantitative variables, where imputation can be based on a model using a combination of other information on the enterprise, this is not possible with qualitative data. Therefore, there are doubts if imputation of qualitative variables will improve the quality of the estimates, compared to simply distributing the missing values in a table. If countries should decide to impute qualitative variables, too, it is recommended that the item non-response rate should not exceed 20 %.

The consortium strongly recommends delivering data files before and after imputation to Eurostat to allow further analysis. We also recommend that qualitative variables that have been imputed should not be used for further breakdown of enterprises. A further recommendation is to give countries the opportunity to place restrictions on the publication of the results of their survey should they be concerned about the appropriateness of some applications of the imputation procedures to their national data.

The paper elaborates on weighting and reweighting CVTS3. It starts off with a simple formula, and goes on taking into account non-response problems and additional information. When using additional information (e.g. updated information on the population of interest) in order to improve both bias and precision, the simple weighting procedure has to be refined. The paper provides recommendations for the construction of weights when using one or more variables as additional information. In any case, applying several weights should be avoided. In order to accommodate for this, implementing calibration weights is recommended. Both the formula for construction of the weight as well as the adapted formula for estimating the variance are provided.

Finally, the paper discusses the problems resulting from the high volatility of the population: Overcoverage, undercoverage, stratum switchers, and split or merged enterprises. Procedures to deal with these problems have to rely on assumptions, and suggestions are highly dependent on the available information. The paper provides advice on these issues.

**Quality components that are important
to include in the quality reports**

by

Statistics Sweden

Introduction

The ideal national quality reports should contain detailed descriptions of all the components in the definition of quality by Eurostat. The definition of quality is described on the web-page; <http://forum.europa.eu.int/irc/dsis/coded/info/data/coded/en/g1011043.htm>. The following quality components are included;

- Relevance
- Accuracy
- Timeliness
- Accessibility
- Comparability
- Coherence
- Completeness

This document covers accuracy and comparability and is a description of the recommended minimum content of the national quality reports. Nevertheless there are some other quality components that we still consider important but we do not describe them in this document, as we restricted our recommendations to the most essential quality components.

Why are quality reports important?

The quality report should inform the users on factors of vital importance for a correct interpretation of the statistics. Each process during the development of statistics should be described in detail, for example collection of data, editing, treatment of non-response and estimation. This information should include the concepts and methodology used in collecting and processing the data and other characteristics of the data that may affect their quality, use or interpretation. For example, the users should be able to evaluate if the objects, variables, statistical measures and reference periods correspond with his/her interests.

A general rule of thumb is that the quality report should contain information, which makes it possible for the users to evaluate if he/she can rely on the statistics from an overall point of view. A user can then analyze the survey results from his/her own particular objectives.

Accuracy

Sampling errors

- Description of the sample design
- Description of the quality in the register, which was used as sample frame
 - Describe how well the sample frame corresponds with the target population for example, existence of known overcoverage and/or undercoverage.
- Description of the estimator
 - Auxiliary information from registers, which possibly have been used in the estimation process should be described
 - Describe the calculation of the point estimator and the variance estimator.
- Calculate confidence intervals for the main indicators by NACE group and size class

Non-sampling errors

Coverage errors

During the data collection it is important that sampling units not belonging to the population are recorded since this information may be necessary in the estimation process.

Flag for:

- Overcoverage

Measurement errors

The data collection method that has been used has probably an affect on the quality in answers for different variables, unit and item non-response rates. For comparisons across countries, a description of data collection methods used in the survey is needed: Face-to-face interviews, postal questionnaires, CAPI, CATI, online questionnaire, combination of several different methods, different methods for different types of enterprises etc.

Provide detailed comments on problems with the questionnaire in general or with single questions (give comments on all variables)

Describe any problems and their effects related to:

- Errors due to the survey instrument
- Errors due to the interviewers

Processing errors

Describe the coding process in detail and give detailed comments for the questions, which have caused a lot of re-contacts with the enterprises or been flagged for often during the editing process. The most common errors in the questions should be described and also the treatment of the errors.

- Data capturing / Collection of data
- Coding
- Editing

Non-response errors

It is important to have enough resources for re-contacts with the enterprises in order to receive as high unit response rates and item response rates as possible. The countries should describe the measures that have been used for re-contacts in the quality report.

After the data collection phase, when the non-response is a fact, it is important that all available data about the sampling units (response and non-response) is saved. This information will make it possible to analyze the structure of the non-response and possibly compensate for the assumed effects of the non-response.

The quality report should contain a detailed description of methods used for compensating for unit non-response and item non-response, for example imputation and weighting.

The quality report should also contain unit non-response rates and item non-response rates for the variables A299tot, A3, A4a, B8a, B8b, C2ntot, C2ttot and C6tot¹ in CVTS3 by stratum i.e. NACE group and size class. Item non-response rates should also be calculated for the most important IVT variables, i.e. number of employed persons participating in IVT and total costs

¹ The names of the variables are according to the CVTS2 manual. At the time of writing this paper, the final list of variables was not yet established.

for IVT. This will make it possible to roughly evaluate the effects of non-response on the results in different countries.

Deliver a data set before imputation of the variables and a data set after imputation of the variables to Eurostat. This will make it possible to analyze the effects of the imputations in different countries.

We also recommend that missing data due to item non-response should be imputed for all quantitative variables. The imputations should be done according to harmonized rules described in the paper, Weighting, re-weighting and imputation.

Imputations for which other sources are available can be used as an alternative if the data sources have sufficient quality.

Comparability

If data for making imputations is coming from another survey or from administrative records then countries should ensure that the data is consistent with the concepts and definitions developed for CVTS3. The quality reports should contain such information.

Describe differences and changes in methodology and other important concepts between CVTS2 and CVTS3, and the impact on:

- Comparability between important indicators by main domains
- Comparability over time

Adequate comparisons over time are facilitated if the statistics are accompanied by information about changes in circumstances which affect the statistics, for example changes in survey methodology or changes in data collection methods.

Variables that we consider to be most important in CVTS3

Variable name CVTS2	Variable name CVTS 3 (Questionnaire V3)	Description
A1	A1	Principal NACE-code of the enterprise
A299tot	A2tot05	Total number of employees 2005-12-31
A3	A5	Total number of hours worked by persons employed in 2005
A4a	A6	Labour costs of persons employed in 2005
B8 (especially B8a/B8b)	B1 and B2 (especially B1a/B1b)	Existence of different kinds of training activities in the enterprise
C2ntot	C1tot	Employed persons participating in CVT courses
C2ttot	C4tot	Hours employed persons spent on CVT courses
C6tot (or vital parts of C6tot)	C8tot (or vital parts of C8tot)	Total costs of CVT = C8a + C8b + ... + C8e + C9a minus C9b
	E1tot05 (new)	Total number of participants in IVT
	E2tot (or vital parts of E2tot) (new)	Total costs of IVT = E2a + E2b + E2c + E3a minus E3b

Extensive resources should be set aside to allow re-contacts with the enterprises to get information for these variables if needed.

If it is still item non-response in **several** of these variables after re-contact with the enterprise that should be a clear indication that the enterprise may have to be treated as a non-response unit.

A general remark is that it is very important to have enough financial resources for re-contacts with the enterprises in order to receive high unit response rates and item response rates. It is actually better to have smaller sample sizes in order to have enough financial resources for re-contacts with the enterprises.

**Sample Design and Allocation for CVTS3 -
A Swedish example**

by

Statistics Sweden

Introduction

The aim of this paper is to give a basis for a common sampling approach in CVTS3 and also to give some guidelines on how to calculate the needed sample sizes. The CVTS3 survey is going to be carried out in a similar manner as the CVTS2 survey. Therefore, most of the guidelines and recommendations used in CVTS2 can be applied to CVTS3.

The sample design should as far as possible reflect the intended output of the survey, in order to receive sufficient quality in the estimates of the desired parameters in CVTS3. One way to achieve this is to stratify the sampling frame according to the important domains.

In CVTS2 it was agreed that a minimum of 60 cells should be used as strata. This guideline should also be used for CVTS3. The cells are defined by the cross-classification of 20 NACE categories, C, D (15-16, 17-19, 21-22, 23-26, 27-28, 29-33, 34-35, 20+36+37), E, F, G (50, 51, 52), H, I (60-63, 64), J (65-66, 67), K+O with 3 size classes (number of persons employed), 10-49, 50-249, 250-. This does not exclude the possibility of a further breakdown in the analysis of the data, though these results will probably be less robust.

It is here assumed that,

1. The sampling unit is the enterprise. In case data cannot be collected at the enterprise level but only at the local unit level, 1) the enterprise is assumed to be the primary sampling unit and the local units are the secondary sampling units, or 2) the local units are used as the primary (and only) sampling units.
2. Each enterprise can be uniquely classified into one of the 60 cells used as strata.
3. The sampling method used is simple random sampling (srs) without replacement within each stratum.

If staging or modularization of the survey is used, this means that the statistical theory for two-phase sampling should be applied when calculating the estimates for the parameters of interest and the confidence intervals.

Sampling frame

It is necessary that there exists a frame (a register or list) of all enterprises of size 10 or more persons employed (optional 5 or more) in the NACE categories C-J, K+O. It is also important that there are no other enterprises in the frame in order to avoid enterprises of no interest in the sample.

It should be possible to cross-classify the enterprises in the frame by size and NACE, which means that the cells can be used as strata. If there is no useful frame available, it will be necessary to construct one or use a multistage sampling design.

In some countries there is no useful register of enterprises but they have a fairly good register of local units. Then a sample of local units can be used to get in touch with the enterprise to which the sampled local unit is associated. The estimation procedure then has to take this into account since enterprises with many local units will have a larger probability to be included in the sample than enterprises with fewer local units. However, when this approach is used, one cannot be sure to present reliable estimates in all intended NACE categories or all size classes.

In case the quality of an existing register is too poor, e.g. not covering all sampling units of interest, important variables are missing, not up-to-date etc., it is necessary to improve on it, otherwise the survey will fail. It is not possible to give detailed recommendations on how to

improve the frame when the quality is considered too poor. The effects of the frame imperfections are likely to differ between participating countries, therefore each country should use the methods they have found as the best to improve the frame.

Especially in small countries there might be very few enterprises in some strata. Usually these strata will be totally enumerated when a minimum sample size constraint is used. If this is not the case it is wise to unite similar strata before the sample is taken. Taking a sample from united strata means that it might not be possible to present estimates for all intended cells.

Deviations from the European guidelines should be possible only if they do not reduce the quality or the comparability of the data.

Parameters of interest

There are a variety of parameters to be estimated in CVTS3. It is assumed that they are about the same as in CVTS2. They are for example the proportions of "Enterprises offering training", "Different types of training", "Enterprises with training plans and/or budgets", etc. These parameters are expressed as per cent in different NACE groups and/or size groups, with the total number of enterprises or the number of enterprises offering training, in the denominator.

Other examples of parameters of interest are "Hours spent on training courses", "Training costs". These parameters are expressed as ratios with the number of participants or the total labour costs in the denominator.

It seems that mainly all parameters of interest can be expressed as ratios within different NACE- and/or size-classes. The totals in the numerator are summed over enterprises offering training (= "training enterprises") while the totals in the denominator are either summed over all enterprises or over the "training enterprises" in the actual domain.

Of outstanding importance are of course the "training enterprises". This is a study variable and cannot be used in the sampling procedure. However, the results from CVTS2 can be used, in order to estimate the number of "training enterprises" within each of the 60 cells defined above. Using this information it will be possible to allocate the sample in such a way that at least the expected number of training enterprises within each cell will be equal to the desired number. CVTS3 has a focus on CVT, and hence training enterprises are in this case defined as those offering CVT to their staff (regardless of IVT).

Allocation principle

Most parameters of interest in CVTS2 were of the type where both numerator and denominator are summed over the "training enterprises"². This means that the working sample size is the number of "training enterprises" in the sample.

Now, what should the expected number of "training enterprises" be? Different parameters of interest will usually give different answers to this. Since there is no information which parameter is the most important, let us keep it simple and use the same idea as in CVTS2.

² Here we have taken 'all training enterprises'. It could be argued to use 'enterprises given courses'.

Assume that the parameter of interest is a proportion based on "training enterprises", for example the proportion of "training enterprises" with a training plan. The allocation is done such that the maximum length of half the confidence (95%) interval of the estimated proportion is not longer than C in each cell, where C for example is 0.05 or 0.1 or 0.2. Different Cs will give different total sample sizes and it is up to the user to make the choice according to the available resources and to the precision needed.

Usually it is good practice to use also a minimum sample size limit as an insurance against non-response. The non-response rates were quite high for most of the countries in CVTS2 and therefore each country should give particular attention to the expected non-response rates within each cell when calculating the sample sizes for CVTS3. Note; this will not solve the problems with non-response bias but it may help to ensure that there is enough enterprises in each cell for the estimation process.

The success of the allocation depends heavily on reliable estimates of the proportion of "training enterprises" within each cell. The proportion of training enterprises in each cell can, for example, be estimated from the CVTS2 data. Consequently this proportion is available for each country that participated in CVTS2. For countries that didn't participate in CVTS2 the figures from another country, which can be assumed to have the same pattern regarding the proportion of training enterprises, can be used. Obviously, the number of responding training enterprises in each cell will be overestimated if non-response increases or training incidence decreases. If there is additional information of better quality available, the proportions should of course be adjusted according to that.

If the proportion of training enterprises in each cell is not "known" these can be calculated from the marginal proportions. The preparatory work³ for CVTS2 stated that the effects of size, sector and country are largely independent of each other and this lead to the conclusion that an acceptable approximation of the proportion of "training enterprises" within each cell might be obtained by a simple multiplication of the marginal proportions. The marginal proportions for the 20 NACE categories and the 3 size classes (number of persons employed) from CVTS2 can be found in the Eurostat database NewCronos.

The allocation principle gives no possibility to allocate the sample size according to different costs for different data collection methods. Different methods cost different and each country must use its resources in the best way.

³ In the official publication, "Continuing training in enterprises: facts and figures" (European Communities, 1999), in page 21 one concludes that, "These analyses have shown that although there is some interaction between the size, sector and country of enterprises in determining whether they offer continuing training, their effects are largely independent of each other."

An illustration

In this section the figures from Sweden are used to **illustrate** the principle.

Step 1

The number of enterprises within each cell, defined by the cross-classification of NACE and size are computed from the Swedish Central Register of Enterprises, see Table 1.

Example: For NACE 21-22, (Paper and printing) the numbers of enterprises are 630, 179 and 66 within the three size classes.

Step 2

The proportions of "training enterprises" within each (20 x 3) cell are computed from the Swedish CVTS2-survey. Countries participating in CVTS2 can normally estimate these proportions from that survey, however additional information (if available) should be used if it will improve the quality in these figures. For countries that didn't participate in CVTS2 the figures from another country, which can be assumed to have similar pattern regarding the proportion of training enterprises, can be used.

An approximation of the proportion of "training enterprises" within each cell can also be obtained based on a simple multiplication of the marginal proportions for the 20 NACE categories and the 3 size classes (number of persons employed) from CVTS2 (For more information see the document CVTS3_raking-ratio.doc).

Example: In the Swedish CVTS2 the estimated proportion of "training enterprises in the combination of NACE 21-22 and the size classes are 0,91, 0,99 and 1,00.

Table 1. Number of enterprises in the Sample frame

Number of enterprises in the sample frame,
by NACE and size

NACE	Size			
	10-49	50-249	250-	All
All	26026	4408	941	31375
C	52	9	3	64
15-16	486	104	36	626
17-19	155	26	7	188
21-22	630	179	66	875
23-26	508	212	55	775
27-28	1 441	246	42	1729
29-33	1 190	353	104	1647
34-35	221	105	44	370
20, 36-37	750	200	26	976
E	147	61	17	225
F	3 181	266	43	3490
50	904	149	21	1074
51	3 029	393	64	3486
52	2 550	274	61	2885
H	1 697	151	15	1863
60-63	1 964	291	57	2312
64	60	25	25	110
65-66	161	103	34	298
67	111	24	3	138
K+O	6 789	1237	218	8244

Table 2. Estimated proportions of “training enterprises”

Estimated proportions of "training enterprises"
by NACE and size

NACE	Size			
	10-49	50-249	250-	All
All	0,88	0,99	0,99	0,903
C	0,85	1,00	1,00	0,880
15-16	0,82	1,00	0,97	0,861
17-19	0,78	0,95	1,00	0,815
21-22	0,91	0,99	1,00	0,929
23-26	0,82	1,00	1,00	0,883
27-28	0,93	0,96	1,00	0,940
29-33	0,82	0,98	1,00	0,866
34-35	0,84	1,00	0,95	0,898
20, 36-37	0,89	0,97	1,00	0,909
E	1,00	1,00	1,00	1,000
F	0,81	1,00	1,00	0,830
50	0,95	0,98	1,00	0,958
51	0,93	1,00	1,00	0,938
52	0,93	0,98	1,00	0,939
H	0,82	1,00	1,00	0,832
60-63	0,80	0,98	1,00	0,824
64	0,74	1,00	0,93	0,843
65-66	1,00	1,00	1,00	1,000
67	1,00	1,00	1,00	1,000
K+O	0,92	1,00	0,99	0,934

Step 3

The number of sampling units (enterprises) needed in the sample to get the desired maximum length of half the confidence interval are computed for each cell, a minimum number of 10 is used.

The expected response rate within each stratum is computed from CVTS2.

The non- response *bias* is another matter and has to be taken care of in other ways. It is believed that non-response biases may occur in CVTS3 and a proposal for the evaluation of these biases should be made (further information can be found in the paper *Treatment of non-response*).

Table 3. Estimated response rates.

Estimated response rates (exkl. over-coverage) by NACE and size			
--	--	--	--

NACE	Size		
	10-49	50-249	250-
C	0,57	0,42	0,67
15-16	0,45	0,42	0,71
17-19	0,47	0,53	0,89
21-22	0,47	0,41	0,64
23-26	0,49	0,61	0,65
27-28	0,49	0,55	0,67
29-33	0,47	0,52	0,67
34-35	0,50	0,59	0,72
20, 36-37	0,44	0,56	0,69
E	0,57	0,74	0,44
F	0,44	0,52	0,63
50	0,34	0,46	0,73
51	0,45	0,41	0,68
52	0,24	0,37	0,61
H	0,32	0,41	0,45
60-63	0,36	0,46	0,59
64	0,34	0,56	0,93
65-66	0,44	0,43	0,50
67	0,38	0,42	0,50
K+O	0,43	0,46	0,60

The sample sizes are computed according to the formula $n_h = 1 / (.10^2 \times te_h + 1 / N_h) / r_h$, where n_h is the number of sampling units, te_h is the proportion of "training enterprises", N_h is the total number of enterprises (training + non training) and r_h is the response rate in stratum (cell) h (see for the derivation of the formula annex 1).

Example: Assume that maximum half length should be 0.10 and the response rate is 47%, then the sample sizes in NACE 21-22 and size class 10-49 should be $1 / (.10^2 \times .905 + 1 / 630) / .47 \approx 198$.

All sample sizes for the case when the maximum half size of the confidence interval is 0.10 are shown in Table 5.

Results

Total sample sizes for different expected maximum length of confidence intervals are shown in Table 3 (estimated non-response rates according to table 3).

Table 4. Total sample sizes for different C-values.

Maximum length of half the confidence interval (C)	Sample size needed
.05	16 678
.10	7 261
.15	3 965
.20	2 497
.30	1 242

Table 5. Sample sizes for C=0.1.

NACE	Size			
	10-49	50-249	250-	All
All	4183	2284	794	7261
C	52	9	3	64
15-16	216	104	36	356
17-19	148	26	7	181
21-22	198	160	62	420
23-26	200	112	54	366
27-28	202	132	42	376
29-33	236	153	76	465
34-35	155	87	43	285
20, 36-37	223	122	26	371
E	105	51	17	173
F	266	139	43	448
50	277	131	21	429
51	232	194	57	483
52	423	200	61	684
H	361	148	15	524
60-63	326	164	57	547
64	60	25	22	107
65-66	141	103	34	278
67	111	24	3	138
K+O	251	200	115	566

For further information also see the following related documents:

- Annex 1: Derivation of the formula, which is used for calculating sample sizes in each stratum
- Annex 2: Excel application for calculating and allocation of sample sizes
- Annex 3: Algorithm for the calculation of the proportion of training enterprises in a table where only the margins are known
- Annex 4: SAS-script with the algorithm for the calculation of the proportion of training enterprises

Concluding Remarks

In countries with a low proportion of "training enterprises" the number of sampled enterprises will be higher than in this example. In strata where the proportion of "training enterprises" is low, 20% or less, the sample sizes will be very large. When the total sample size needed is too large for the available resources one might consider 1) the possibility to unite some strata that are considered as less important or 2) widening the maximum length of the confidence intervals. A third, perhaps unrealistic approach in the short perspective is to improve on the sampling frame such that the probability to find the "training enterprises" increases.

Countries should make use of their best up-to-date information on training enterprises to make an efficient sampling plan. Where such information is not available, results from CVTS2 should be used, maybe 'updated'. For countries that never carried out a survey on enterprise training, results from a country with similar conditions might be used.

We propose that each country calculate their sample sizes, on the basis of the framework above. If the sample sizes are larger than can be afforded, the country has to make a new decision about the 60 cells, based on the most important parameters to be estimated or accept a lower accuracy.

Confidentiality problems may occur at the detailed level of 60 cells. The cross tables (NACE×Size Classes) will probably only be published for the European Union and if possible for some large countries. Only marginal distributions will be published for all countries. In that case only 20+3 cells is needed in the calculations and then the needed sample sizes will be much smaller. The calculations will be made in the same way as above. The only difference is that instead of using the 60 cells in the cross-tabulation, the 20+3 marginal cells will be used.

We propose that all countries should use the proposed sampling design and allocation. However, it is recognised that different conditions in different countries may suggest different sampling approaches. If there are some special conditions in a country this should be discussed with Eurostat.

Certain plans, even though diverging from the proposal above, may be adopted on the basis of the expected quality of the estimated parameters. That is, each country can adopt the necessary modifications to the sampling plan to improve reliability, accuracy and comparability of the estimates.

A harmonised sampling design is not an objective in itself. The objective is to increase the comparability between Member States. A harmonised sampling design alone does not ensure the comparability.

We recommend that each country, on the basis of the sampling framework above, propose the sampling design, which is most suitable to the conditions in the country. Such a design is to be approved by Eurostat.

Weighting, re-weighting and imputation

by

Statistics Sweden

Non-response

Introduction

There are two types of non-response:

- **Unit non-response** arises when no survey data are collected for a unit (information is missing on all the questionnaire variables)
- **Item non-response** arises when some data are collected for a unit but values of some items are missing (information is missing on at least one, but not all, of the questionnaire variables)

Non-response causes at least two types of problems. Firstly, by reducing the sample size, non-response might cause less precise estimates for important indicators regarding different population groups. The second and more essential problem caused by non-response is the introduction of bias. Non-response can lead to over- or underrepresentation of some groups in the population. If these groups have different training pattern compared with other groups in the population, the estimates based on the respondents in the sample will be biased and therefore not representative for the entire population.

It is well known that the bias tends to increase with the rate of non-response. If the non-response rate in the survey is high, one really has reason to worry about its effects on the survey estimates. If there is no evidence, which shows the opposite, we can assume that the estimates are biased. It is very difficult to estimate the impact of non-response on the estimates. The non-response rates are useful for describing response patterns in the survey, but will not give enough information in order to analyse the effects of non-response on bias in the survey estimates.

Another matter of course is that non-response may lead to extended costs. High non-response rates will increase the administrative burden, postage fees, and so on.

The treatment of non-response: (Re-) weighting and Imputation

It is important to have enough resources for re-contacts with the enterprises in order to reduce the unit and item non-response rates as much as possible. However, in practice some amount of non-response certainly arises. It is then necessary to consider how to treat the non-response at the estimation stage of the survey. Estimation typically involves the construction of a point estimator (and weights) and an associated variance estimator. The principal methods used to correct for bias due to non-response and to make efficient use of data are imputation and re-weighting. It is recommended in CVTS3 that re-weighting is used to treat the problem of unit non-response, while imputation is used to treat problems of item non-response.

Re-weighting entails changing the weights of the respondents, compared to the weights that would have been used if no non-response had occurred. Since observations are lost by non-response, re-weighting will imply increased weights for most of the responding enterprises. Imputation entails replacing missing values by a fabricated value.

Variance estimation in the presence of imputation is a complex statistical problem, which means that more advanced formulas for estimating the variance according to statistical theory

should be used. If the variance is estimated in the same way as when imputation has not been used, it will be underestimated, which means that the length of the confidence intervals are too short. Non-response brings additional variance over and above the sampling variance.

Some experiences from CVTS2

It has not been possible to make a complete analysis of the imputation methods in CVTS2. In order to do that, it is necessary to have access to micro data before imputation of the variables and a data set after imputation of the variables for all participating countries. This will make it possible to analyse the effects of the imputations in different countries. Unfortunately, most CVTS2-countries only sent the imputed dataset to Eurostat. For CVTS3 it is recommended that both the dataset before imputation and the dataset after imputation will be sent to Eurostat.

However, a study of the Swedish CVTS2 data shows that the high non-response rates causes problems in the survey and the survey estimates are therefore very uncertain. The results from the survey are then difficult to interpret in a correct way.

It can also be assumed that the high non-response rates in CVTS2 can be an effect of difficulties for the respondents to give correct answers to some of the questions. Except for big non-response and sample errors, this means that considerable measuring errors also can be assumed to cause biased estimates in the survey.

Summary of the results from a study based on Swedish CVTS2 data

The purpose with this section is to describe problems caused by item non-response in quantitative variables where ratio imputation has been used. In order to do that a small study based on hours worked by persons employed in 1999 from Swedish CVTS2 data has been performed. This study can give some ideas about the impact on the estimates for some of the imputation methods at least for countries with similar conditions as Sweden.

It will however not be possible to make general conclusions about the effects of the imputation methods on all the variables and for all countries based on the results of this small study. The effects of the non-response are likely to differ between variables and estimates considering target parameters and also between participating countries.

The item non-response in the variable, total hours worked, was only 4 percent in the Swedish CVTS2, which was the reason for choosing this variable for performing the study.

Variable: A3 (#hours worked by persons employed in 1999)

Method of imputation: Ratio imputation according to the EU-manual.

Firstly three different levels (A, B, C) of item non-response have been simulated and secondly different estimates have been calculated. The probability that an enterprise will respond depends on which size class it belongs to. The response sets are generated according to the table below.

Response probability (%)

Size class	Level of item response rate		
	A (70 %)	B (80 %)	C (90 %)
10 - 15	0.55	0.70	0.80
16 - 30	0.60	0.75	0.88
31 - 50	0.65	0.78	0.90
51 - 100	0.70	0.82	0.92
101 - 250	0.82	0.90	0.96
251 -	0.85	0.95	0.99

Two different estimation methods have been used. In the first method imputation of item non-response and re-weighting of unit non-response was used. In the second method the item non-response was treated as unit non-response and re-weighting was used for non-response adjustment. From this study it is possible to compare the estimates calculated with those two methods with estimates based on small item non-response.

Total hours worked by persons employed in 1999 (variable A3). Differences, expressed as a percentage, between the estimates based on datasets with different level of item non-response and the estimate based on the dataset with almost no item non-response.

Field of activity	Without imputation			With imputation		
	Ratio of item nonresponse in %			Ratio of item nonresponse in %		
	30 %	20 %	10 %	30 %	20 %	10 %
Mining and quarrying	-0.3	-0.2	-0.1	0.2	0.2	0.1
Manufacture of food products, beverages and tobacco	5.2	2.6	2.5	0.6	0.5	0.1
Manufacture of textiles and leather products	1.7	2.2	2.7	-1.2	-0.2	0.0
Manufacture of pulp, paper and paper products; publishing, printing and reproduction of recorded media	1.2	1.2	1.0	-0.5	-0.1	-0.2
Manufacture of rubber, plastic, refined petroleum products, chemicals and chemical products	1.2	1.0	0.9	-0.3	-0.1	-0.1
Manufacture of basic metals and fabricated metal products	3.9	3.9	3.7	0.5	0.4	-0.1
Manufacture of machinery and equipment n.e.c; manufacture of electrical and optical equipment	2.1	1.3	0.8	0.7	0.2	0.1
Manufacture of transport equipment	2.1	1.2	0.5	-0.6	-0.4	-0.1
Manufacture of wood and wood products, manufacturing n.e.c.	3.5	3.1	2.7	0.1	0.0	0.1
Electricity, gas and water supply	2.8	2.4	2.4	0.4	0.4	0.3
Construction	2.3	2.1	2.0	-0.9	-0.1	-0.1
Sale, maintenance and repair of motor vehicles and motorcycles; retail sale of automotive fuel	8.0	3.8	3.0	-2.2	-1.8	-1.3
Wholesale trade and commission trade, except of motor vehicles and motorcycles	3.3	3.0	3.0	0.4	-0.2	-0.1
Retail trade, except of motor vehicles and motorcycles; repair of personal and household goods	8.7	8.2	6.8	-0.6	0.8	0.6
Hotels and restaurants	8.4	6.1	6.3	-6.3	-11.9	-8.8
Transport. Supporting and auxiliary transport activities	7.1	7.3	6.5	0.2	-0.1	0.0
Post and telecommunications	-1.4	-0.7	-0.8	-0.5	-0.5	-0.8
Financial intermediation. Insurance and pension funding	3.2	3.3	3.3	0.0	0.0	0.0
Activities auxiliary to financial intermediation	-148.2	-26.3	22.1	-0.3	0.0	-0.1
Other services	25.3	21.2	12.6	-1.6	-1.2	-0.9

The figures in the table above are differences, expressed as percentages, between the estimates based on the datasets with different levels of item non-response and the estimates based on the dataset with almost no item non-response. Method 1 (imputation for item non-response) produces estimates that tend to be closer to the estimates that are computed from the dataset with almost no item non-response. If imputation of item non-response isn't used, the estimates tend to be overestimated for the studied variable.

The unit non-response rate is nearly 50% and together with item non-response rates between 10-30% a lot of necessary data have been lost. This is a quality problem, which probably will result in biased estimates and comparison problems between countries. It is therefore recommended that re-contacts with the enterprises will be performed before any imputation of the quantitative variables. It is also important that the unit non-response and the item non-response are described in the national quality reports together with a description concerning treatment of non-response.

Imputation in CVTS3

Introduction

It is to be expected that the initial response from some enterprises will yield data, which are incomplete with respect to some variables (item non-response). When this occurs, it is important that countries should first try to get enterprises to provide these missing data by contacting enterprises again (especially for important quantitative variables) to see if estimates can be provided and, if necessary, to assist them in doing this. If an estimate cannot be provided, another approach is to use reliable data for the enterprise from other sources, which are compatible with the concepts and definitions adopted for the CVTS3.

If, after all other efforts have been made, some data are still missing then these data can be imputed; i.e. estimated using other information available. The purpose of imputation is to provide a sample and, hence, population estimate of a given variable which is better than that which would be obtained by simply distributing the 'not known' or 'not available' totals in a table according to the positive responses in that table. This is because imputation, normally, will take into account more information than is available in a single table. Imputation may reduce bias arising from the item non-response.

Imputation is also important in a survey such as CVTS3 because the data from national surveys will be combined to produce estimates for the EU (and larger groupings of countries). If a completed dataset is used when two tables are produced, the margins of these tables will be identical, but this might not be the case if the tables had been based upon data with different sets of missing values.

Rules for Imputation

The basic principle assumed for imputation is that, as far as possible, we should try to make use of all the information collected when interpreting the data.

The imputation methods used may have an impact on distribution of data. In general, the greater the degree and impact of imputation, the more careful you need to be in using the data. In the following situations imputation should not be made without considering the impact of imputation when analysing data:

- imputations are not allowed in the case that a record or a case (the 'questionnaire' of one enterprise) has a score on less than 50% of the variables (in particular if this concerns the core questions on training, see the list below);

- imputations are not allowed if more than 50% of the responding enterprises in a stratum has missing data on more than 25% of the quantitative variables (in particular if this concerns the core questions on training, see the list below);
- imputations for which other sources can be used as an alternative source of information are only allowed if the quality of these data can be guaranteed and the data are sufficiently recent.

It is difficult to develop detailed rules for imputation at this moment. On the one hand it requires a theoretical analysis of possible relations between variables, which can be used as a 'predictor' of the score of a particular enterprise on a particular variable. On the other hand, fully sound hypothesis regarding this can only be based on running empirical analyses on the collected data (with a potential difference in hypotheses between countries). This type of analysis can only be made if the countries submit a dataset before imputation and a dataset after imputation to Eurostat.

However, in some cases the rules listed above are too strict and should be modified with respect to the content of information provided within a record and the sample (strata). In an extreme case, for example, if an enterprise only gave quantitative information for the total number of employees and the total number of participants in courses, this is valuable information because it will contribute to the calculation of the participation rate (in courses) for the sector/size group to which it belongs and for the country as a whole. If the missing data for the enterprise are not imputed then tables including these variables will have entries for 'not available'. Usually, if there are not too many 'not availables' these tables are interpreted by ignoring the 'not availables' and assuming that the available data are representative of the population as a whole. But this is, in effect, the same as imputing to the 'not availables' the average values calculated from the available data.

It is however not enough with information on the total number of employees and the total number of participants in courses, if all other quantitative information is missing. In this proposal, the following variables is considered to be most important:

Variable name CVTS2	Variable name CVTS 3 (Questionnaire V3)	Description
A1	A1	Principal NACE-code of the enterprise
A299tot	A2tot05	Total number of employees 2005-12-31
A3	A5	Total number of hours worked by persons employed in 2005
A4a	A6	Labour costs of persons employed in 2005
B8 (especially B8a/B8b)	B1 and B2 (especially B1a/B1b)	Existence of different kinds of training activities in the enterprise
C2ntot	C1tot	Employed persons participating in CVT courses
C2ttot	C4tot	Hours employed persons spent on CVT courses
C6tot (or vital parts of C6tot)	C8tot (or vital parts of C8tot)	Total costs of CVT = C8a + C8b + ... + C8e + C9a minus C9b
	E1tot05 (new)	Total number of participants in IVT
	E2tot (or vital parts of E2tot) (new)	Total costs of IVT = E2a + E2b + E2c + E3a minus E3b

If it is still item non-response in several of these variables after re-contact with the enterprise that should be a clear indication that the record should be converted to a unit non-respondent and taken into account for the adjustment of the weight. Exceptions to this rule may be accepted, after consultation with Eurostat.

The second rule implies that imputations should not be made when the variable to be imputed is missing from too many enterprises. It is possible to imagine complex rules for determining what percentage of item response is required before imputation can be allowed. For example, if there is evidence to suggest that the likelihood of an enterprise responding to a question is independent of the answer, then a relatively low item response rate can be accepted for imputing the missing values. Conversely, if there is an association between the probability of responding and the answer to the question, then a very high response rate would be required before imputation should be used. But in such situations a low response rate will mean that the results are biased anyway and should not be used; then the item would be missing in the strata concerned. In practice these relationships are rarely known unless intensive follow-up and non-response surveys have been carried out to assess possible response biases.

The proposal is that imputation can be carried out if the item response rate for the variable to be imputed concerning responding enterprises within a stratum is 50% or more except when there is reason to believe that there is a strong association between the probability of responding and the required answer. In this latter situation an item response rate of 70% should be required. If after aggregation the item response rate within a stratum is still less than 50% (or 70% if applicable) the actual responses for this item should be converted to missing values in the whole data set supplied to EUROSTAT. However, in the light of the divergent views regarding this rule and in order to avoid as far as possible situations where results on an item (or combinations with it) would be missing for a country, especially if the country for good reasons cannot rely on the %-limit, we recommend that a distinction should be made between

- the imputations required to be done for EUROSTAT according to the rule (so that results can be produced for aggregates of the country, all countries or groups of countries), and
- the restrictions that countries would wish to place on the publication of the results of their survey because of concerns they may have about the appropriateness of some applications of the imputation procedures to their national data.

This compromise is adopted also due to the fact that only the participating countries will have all the details of their national surveys available when the imputations are made, and therefore be able to decide finally on the appropriate rules and methods.

The effect of this distinction would be twofold. First, EUROSTAT would have the data on which to make all the estimates it required. And, secondly, the national statistical services would retain some control over the quality of the results published for their country. In the tables published by EUROSTAT, therefore, some results for a country or certain strata could be suppressed because the country concerned judged that the estimate was not sufficiently reliable because of the large number of imputed values which had contributed to the estimate. This would be in addition to other reasons why data for individual cells may be suppressed; i.e. large sampling errors, number of observations too small or confidentiality reason.

All countries, therefore, should impute missing values according to the rules set out in the paper but should, afterwards, inform EUROSTAT, with reasons, if they think estimates for the country or for some strata should not be published.

Types of Imputation

Two types of data are being collected in CVTS3, quantitative and qualitative, for which different imputation procedures are required. It would be ideal if missing values for both types of data could be imputed so that enterprise records sent to EUROSTAT will be complete with respect to all variables. However, it is not possible to base the imputation of qualitative variables on a model, which uses a combination of other information in the enterprise record. This means that there exist doubts if imputation of qualitative variables will give better quality in the estimates than simply distributing the 'not known' or 'not available' totals in a table. It is therefore recommended that quantitative variables will be imputed in the first place. If qualitative variables are imputed the item non-response rate shouldn't exceed 20%. Otherwise it is better that the item non-responses for the qualitative variables are shown in the tables under the category 'not known'. It is also recommended that qualitative variables, which have been imputed, will not be used for further breakdown of enterprises within a NACE-group or size-class when the results are presented.

The procedures and proposals outlined below are offered as a model that participating countries should use since there are advantages for the comparability of the data if all countries use the same methods.

Whatever method is used, however, it is an essential requirement, following the imputation of missing data, that the enterprise record should be internally consistent. It is important, therefore, that the imputation procedures are used in such a way that this internal consistency is preserved and verified using the Data Checking Tool. It is also important that the original data records (before imputation) are kept and that a detailed description of the imputation procedure used is contained in the Quality report.

Methods of imputation

The traditional approach to imputation is to produce just one imputed value for each missing item. This is called single imputation. The length of the confidence intervals will be underestimated if the uncertainty caused by the non-response isn't taken into account in the variance expression.

A number of alternative approaches to variance estimation in the presence of imputation are possible. Some methods try to correct the "standard variance formula" by adding a suitable correction term (see Lundström & Särndal, 2001). An alternative approach is to design the imputation method in such a way that a simple variance estimator can be constructed. One such approach is multiple imputation (Rubin, 1987).

Imputation methods can be classified as either deterministic or stochastic, depending upon whether or not there is some degree of randomness in the imputed data.

Deterministic imputation methods include for example mean imputation, ratio and regression imputation and single donor nearest-neighbour imputation. These methods can be further divided into methods that rely only on deducing the imputed value from data available for the

non-respondent and other auxiliary data and those that make use of the observed data for other responding units in the survey.

Stochastic imputation methods include the hot deck, nearest neighbour imputation where a random selection is made from several “closest” nearest neighbours, regression with random residuals, and any other deterministic method with random residuals added.

In the annex detailed imputation methods for the quantitative variables are described. For the entire set of variables ratio imputation within specific groups based on strata and in some cases a variable collected in the questionnaire are used. Only one imputed value for each enterprise and variable is produced.

Quantitative variables

For the purpose of what follows, quantitative variables are all those for which a numerical response is required. Quantitative variables may be imputed using a combination of other information in the enterprise record, which contains the missing value, combined with either:

- data collected in the survey from enterprises in the same sector/size group, or
- data from another survey or other source of information, e.g. administrative records for the same sector/size group.

This paper only describes a method of imputation based entirely on information collected in the survey. If other sources are used, great care is needed to ensure that the data from the external source are as compatible as possible with the concepts and definitions adopted for CVTS3.

A principle to be adopted, and illustrated below is that whatever method is used, missing values should be imputed according to a logical sequence or order determined by the relationships between variables. This is because the ability to impute certain variables will depend on the prior imputation of other variables.

Another principle is that data to be imputed for an enterprise should be based on information from other enterprises in the same sector/size class. This will normally mean that enterprises in each of the 60 (20 sectors x 3 size classes) strata used for sample selection should be treated as a separate data set for the purpose of imputation. If sample sizes and stratification procedures permit, more strata may be used in order to improve the efficiency of the imputation. Conversely, if the number of observations in a stratum is too small or the item response rate is too low, certain strata may have to be combined before the imputations are made. These decisions will have to be based on an observation of the levels of achieved item response and item non-response in each of the strata. A first requirement, therefore, is that a count is made of both the achieved responses and the non-responses for each quantitative variable within each stratum. This will also help determine the rules and conditions for how and when imputation can be made. The same information will be useful later for the Quality report for which item response rates will be required.

The following table lists those quantitative variables for which imputation is recommended and methods for imputing missing values are specified in a table at the end of this chapter (p. 37).

Quantitative variables for which imputation is recommended and methods for imputing missing values are suggested

Note: The variable names below are according to version 3 of the questionnaire.

Variable	Description
A2m05	Total number of males employed 31-12-2005
A2f05	Total number of females employed 31-12-2005
A3m	Persons employed - Managers, professional staff and technicians
A3s	Persons employed - Skilled and semi-skilled employees
A3u	Persons employed - Unskilled employees
A4a	Persons employed - Under 25 years of age
A4b	Persons employed - 25 to 54 years of age
A4c	Persons employed – 55 years and older age
A5	Number of hours worked
A6	Total labour cost
C1m	CVT course participants - Males
C1f	CVT course participants - Females
C2m	CVT participants - Managers, professional staff and technicians
C2s	CVT participants - Skilled and semi-skilled employees
C2u	CVT participants - Unskilled employees
C3a	CVT participants - Under 25 years of age
C3b	CVT participants - 25 to 54 years of age
C3c	CVT participants – 55 years and older age
C4i	Total paid working time in hours in INTERNAL CVT courses
C4e	Total paid working time in hours in EXTERNAL CVT courses
C5tot	Total paid working time in hours in CVT courses
C5m	Total paid working time in hours in CVT courses - Male
C5f	Total paid working time in hours in CVT courses – Female
C6a	Total paid working time in hours – Languages, foreign (222) and Mother tongue (223)
C6b	Total paid working time in hours – Sales (341) and Marketing (342)
C6c	Total paid working time in hours – Accounting (344) and Finance (343), Management and Administration (345), office work (346) (including human resource management and quality management)
C6d	Total paid working time in hours – Personal skills/development (090), Working life (347) (including company knowledge and introductory courses)
C6e	Total paid working time in hours – Computer science (481) and Computer use (482)
C6f	Total paid working time in hours – Engineering (540) and Manufacturing (520)
C6g	Total paid working time in hours – Environment protection (850) and Occupational health and safety (862)
C6h	Total paid working time in hours – Personal services (810), Hotel, Restaurant and Catering services (811), Travel Tourism and Leisure services (812) and Transport services (840)
C6i	Total paid working time in hours – Other training subjects
C7a	Total paid working time in hours – Schools, colleges, universities and other higher education institutions
C7b	Total paid working time in hours – Public training institutions (financed or guided by the government; eg adult education centres)

C7c	Total paid working time in hours – Private training companies
C7d	Total paid working time in hours – Private companies whose main activity is not training (equipment suppliers), parent/associate companies)
C7e	Total paid working time in hours – Employer’s associations, chambers of commerce, sector bodies
C7f	Total paid working time in hours – Trade unions
C7g	Total paid working time in hours – Other training providers
C8a	CVT course costs – Fees and payments for courses for employees
C8b	CVT course costs – Travel and subsistence payments
C8c	CVT course costs – Labour costs of internal trainers
C8d	CVT course costs – Training centre, training premises or specific training rooms of your enterprise
C8e	CVT course costs – Teaching materials for CVT courses
C8sub	CVT course costs Sub total (a-e)
E2a	IVT costs – Labour costs of individuals registered on an IVT activity
E2b	IVT costs – Labour costs of IVT trainers or mentors
E2c	IVT costs – Other costs – training fees, travel costs, teaching materials, costs of training centre etc
E2sub	IVT costs – Sub total costs (a-c)

Note: After imputations of the different parts of a question the total must also have a value in accordance with the sum of the different parts. For instance after imputation of C5m and/or C5f the variable C5tot must be equal to the sum of C5m and C5f. Also make sure that the flagvariables have values accordingly.

The imputation procedure described in the Annex requires that each of the steps are performed sequentially following the order of the questions. This will ensure that, for the reasons mentioned above, all the prior imputations needed have been carried out before a new missing value has to be imputed. Also, it will be noted that, for the imputation of cost data, reference has to be made to whether the enterprise offers internal training only, external training only or both and whether, for enterprises offering internal training, they have a training centre. This is because answers to these questions influence the level of training costs and the distribution of the types of costs incurred. Also, for the cost data it is proposed that only the types of costs in the sub-total should be imputed.

Qualitative variables

The method proposed for the imputation of qualitative variables is sequential 'hot-decking'. The first step in this procedure is to partition the enterprise records within each sector/size cell into two groups; those in which the variable (or variables if appropriate) is missing and those which contain the variable (or variables) to be imputed.

Further sub-division will be needed, also, between enterprises offering CVT training courses (with or without other forms of training), enterprises offering other forms of CVT training and not CVT training courses, enterprises with apprentices in IVT training only and enterprises that did not offer training. This is because it may be expected that some of the qualitative answers in the questionnaire will vary considerably between these four groups. For example, the existence of a training plan in the enterprise will be highly associated with whether the enterprise offers training and what type of training.

As in the case of quantitative variables it will be necessary, first, to determine the scale of the imputation needed for each variable and whether the number of enterprises in each cell is large

enough to implement the imputation procedure. If this is not so then some aggregation of the sector/size groups will be needed. The sub-division between training and non-training enterprises and by type of training (described in the preceding paragraph) should not be violated.

There is not the same imperative, as was the case for quantitative variables, for a logical structure to be followed. This is because, generally, the logical connections between the answers are found within the qualitative questions.

Returning to the partitioning of the data in each cell as described above, the next step is for the missing responses to be replaced by values selected from an enterprise, which has answered the questions. This is preferable done in a random way but, providing that there is no systematic way in which the two data sets are ordered, it is simpler to do the replacement sequentially.

Weighting and reweighting in CVTS3

Preliminaries

In CVTS3 a sample of enterprises will be taken according to a sampling design that means Stratification with Simple random sampling within strata (STSI). The samples within strata will be taken without replacement.

Strata are constructed by dividing the sample frame, which usually is the Statistical Business Register (SBR), by NACE category and size. A minimum of 20 NACE categories and 3 size classes are demanded. However, the participating countries are allowed to divide the $20 \times 3 = 60$ “base strata” into substrata.

In the following, let index h , $h=1, 2, \dots, H$, be used for strata, where H is the total number of strata in the survey. By “strata” we mean the strata used in the sampling procedure, i.e. $H=60$ or $H>60$ when a further division of the “base strata” is done.

In stratum h , the number of enterprises in the frame is N_h .

A sample s_h of size n_h is taken from stratum h and the set r_h contains m_h responding enterprises. The total sample s contains $n = \sum_h n_h$ units and total response set r contains $m = \sum_h m_h$ units.

Let y_k be a variable of interest for enterprise k .

We want to estimate the total of y , $t_y = \sum_U y_k$ in the population U of enterprises from our sample survey.

NOTE. By defining y_k in a suitable way we can use this notation for a total in any sub-population of U . Usually, we are interested not only in totals but also in certain functions of different totals, for example proportions or means. In any case we need to state how a total should be estimated from our observed data.

The total ty is estimated by applying a weight, w_k , to each observed enterprise k in the survey.

The total ty is estimated by $\hat{t}_y = \sum_r w_k y_k$. That is, each observed y-value y_k is multiplied by a weight w_k and are summed over the observed units.

Now, the question is; How should the weights w_k be constructed?

Constructing the weights

In a world where the sample frames are perfect and there is no non-response, we could for example use the weight $w_k = N_h/n_h$ for all enterprises that are sampled in stratum h. When we have additional information about the population, for example the number of enterprises in certain groups of interest, we may use this information to improve the precision in the estimates. This can be done by post-stratification and similar techniques.

However, our world is not perfect; we will probably have a substantial amount of non-response and probably not so perfect sampling frames. We can use the sampling weight to handle the problems that are left after all efforts have been taken to improve the sampling frame, raise the response rate, edit the data etc.

Non-response problems

When the weight $w_k = N_h/n_h$ is used and we have only $m_h < n_h$ respondents in stratum h, the obtained estimate of a total will in general be too small.

The principal methods for non-response adjustments are weighting and imputation. In CVTS3 imputation will only be used for item non-response, not for unit non-response.

Much of the work on estimation in the presence of non-response has used the idea that the response is stochastic and that a response distribution exists. The response set is seen as the result of a response mechanism to which a system of response probabilities belongs.

In reality these probabilities are never known. It is then suggested that a relevant model should be formulated from which the response probabilities can be estimated. Most of the suggested models are simple. Perhaps the simplest model is that all units within a stratum respond with the same probability. This probability is then estimated by m_h/n_h and its inverse $v_k = n_h/m_h$ is used as the non-response weight. The resulting weight, $w_k = (N_h/n_h)v_k = N_h/m_h$ for enterprise k in stratum h is then used in the estimation.

This will solve the problem if the model is true or when there is no relationship between the non-response probability and the variable of interest (y). When the assumptions are not fulfilled we will get biased estimates. Another effect from non-response is that the sampling error will increase due to the fact that the number of observations is reduced.

The use of additional information

In many cases the simple model above is too simple. This means that we will need additional information about the sample and/or population of interest. There might be additional information in the sample frame that has not been used in the stratification and which can be used to construct weights that are believed to improve both bias and precision. The additional information may also consist of updated information about the population of interest, for example a newer version of the SBR that is available at the time when the estimation is to be

done. The use of additional information will have impact on the estimation of the standard errors, which has to be considered.

Example 1. Assume that the population can be divided into three regions. Each enterprise in the response set is classified according to region. The total number of enterprises in the population is known for each region. Then one possible weight is $w_k = (N_h/m_h)(N_g/\hat{N}_g)$ for a responding enterprise in stratum h that belongs to region g, where $\hat{N}_g = \sum_h (N_h/m_h) m_{hg}$ is an estimate of the known number, N_g , of enterprises in region g and m_{hg} is number of responding enterprises in stratum h that belongs to region g. The method can be used when the number m_g is large enough ($m_g > 5$, say). When there are other types of information available this technique may be used as well.

Example 2. Let t_{xg} be the known number of employees in region g, then another possible weight is $w_k = (N_h/m_h)(t_{xg}/\hat{t}_{xg})$ for a responding enterprise in stratum h that belongs to region g, where $\hat{t}_{xg} = \sum_h (N_h/m_h) t_{xrhg}$ is an estimate of the known number of employees, t_{xg} in region g and t_{xrhg} is number of employees in the responding enterprises in stratum h that belongs to region g. The number of responding enterprises in group g must be large enough also in example 2.

It may well happen that the user wants to apply the weight from example 1 to estimates where the variable of interest is qualitative while the weight in example 2 is preferred for quantitative variables. Then we have an undesired situation with two different weights that might cause problems with consistency between different estimates. There is also an increased risk for mistakes when more than one weight is available in the estimation procedure.

There is a possibility to combine the two variables used as additional information into one weight, albeit it is a bit more complicated than the suggested weights above. The weight will be calculated separately for each responding enterprise k in stratum h.

$$g_k = (N_g/\hat{N}_g) \left(1 + (x_k - \tilde{x}_g) \frac{(\bar{x}_g - \tilde{x}_g) \hat{N}_g}{\sum_h \frac{N_h}{m_h} \sum_{rhg} (x_k - \tilde{x}_g)^2} \right)$$

Define

for each responding enterprise k that belongs to group (region) g, where \hat{N}_g is computed as before, $\bar{x}_g = t_{xg}/N_g$ is the known mean value in group g, $\tilde{x}_g = \hat{t}_{xg}/\hat{N}_g$ is the estimated mean value in group g and rhg is the set of responding enterprises in stratum h, group g.

The final weight $w_k = (N_h/m_h) g_k$ has the properties that $\sum_h N_h/m_h \sum_{shg} g_k = N_g$ and $\sum_h N_h/m_h \sum_{shg} g_k x_k = t_{xg}$, that is, when the weight is used for the estimation of N_g and t_{xg} we will get the known totals exactly.

The variance for the calibration estimator is estimated according to the following formula⁴;

⁴ Estimation in the presence of Nonresponse and frame imperfections, by Sixten Lundström and Carl-Erik Särndal (Statistics Sweden 2001)

$$\hat{V}(\hat{Y}_w) = \hat{V}_{\text{SAM}} + \hat{V}_{\text{NR}}$$

The variance estimator has two components, an estimate of the sampling variance V_{SAM} , and an estimate of the non-response variance, V_{NR} . The exact expression of the variance estimator depends on the sampling design and the additional information that have been used for estimation. More information about variance estimation can be received from statistical literature, for example from “Estimation in the presence of Nonresponse and frame imperfections” by Sixten Lundström and Carl-Erik Särndal (Statistics Sweden 2001) or Model Assisted Survey Sampling by Carl-Erik Särndal, Bengt Swensson and Jan Wretman (Springer Series in Statistics 1992).

The gk-weight looks a bit cumbersome to calculate but there are different software available for this, for example CLAN97 from Statistics Sweden or CALMAR from INSEE in France. It is also possible to estimate the variance and consequently also the confidence intervals with these computer programs.

Frame problems

No doubt all countries will have problems with overcoverage, i.e. there are enterprises in the sample that no longer are in business or for other reasons do not belong to the target population. There is probably also the problem with undercoverage, i.e. enterprises that do belong to the target population but they have no chance to be included in the sample because they are not in the sampling frame.

Usually it is fairly simple to identify the overcoverage in the response set, it is more difficult to identify them among the non-respondents. One of the reasons why an enterprise is a non-respondent may be that it is no longer in business; however, we do not know that for obvious reasons. By definition it is still more difficult to know anything about the undercoverage.

These problems are due to the facts that the sampling frame is not perfect and that the population of enterprises usually is very volatile.

What we can do is to use model assumptions about the over- and undercoverage and try to adjust for them by the weighting procedure. These assumptions are made for each stratum.

- The first assumption is that the overcoverage rate among the respondents is the same as among the non-respondents.
- The second assumption is that the number of enterprises in the overcoverage in the frame is the same as the number of undercoverage enterprises, this means that the size of the population is not changing during the survey period.
- The third assumption is that the mean value of any variable of interest is the same in the part of the sample that belongs to the target population as in the undercoverage part.

These three assumptions can of course be questioned but they are simple and lead to a simple procedure. If we are willing to accept these assumptions then the number of responding enterprises in stratum h , m_h shall not contain the number of overcoverage enterprises. Note that N_h shall not be reduced for the number of discovered overcoverage. For example if there are 100 enterprises in the response set, of which 10 do not belong to the target population, then $m_h=90$ should be used in the weighting procedure in that strata.

The problem with “stratum switchers”

In practice it is not uncommon that some enterprises turn out to have changed stratum at the time for data collection. The enterprise might have grown or changed NACE-group. It is important to note that it is the stratum identity at the sampling occasion that counts, i.e. if an enterprise is sampled in NACE group E and size class 10-49 and it turns out that it has 62 employees, then N_h/m_h comes from stratum (E, 10-49). However, if the SBR has been updated and the enterprise belongs to the true size class in the updated register then it is probably a good idea to use the updated information as additional information in the construction of the final weights.

For example the updated division indexed by g may cross the original division into strata indexed by h and each responding enterprise is classified into the classes hg. The totals N_h are known from the sample procedure and the totals N_g and possibly tx_g are known from the updated register. Note that the known totals should contain data from enterprises that are members of the target population only.

The use of additional information may have the twofold effect on the estimates that the bias is decreased and the precision is increased when the information is correlated with the variable(s) of interest. Otherwise it will probably have a small or no impact on the estimates.

The problem with split or merged enterprises

It has been mentioned before in this paper that the population of enterprises usually is unstable. This does not only mean that new enterprises are born and that old ones die during a given time period. It also means that an enterprise in the sampling frame might have been split into two or more different units. Conversely, it may also happen that two or more enterprises in the sampling frame have merged into a new single unit. This is not just a weighting problem although weighting may be used as one solution to at least a part of it.

It is only possible to give general advise on how to handle some typical situations. The suggestions are highly dependent on the available information about the enterprises.

Split enterprises

The following suggestion may be used when the split occurs during the reference period (2005). When an enterprise is split into two or more separate units we can treat the “mother” enterprise from the frame as a cluster of “daughter” enterprises. The “daughters” may belong to different NACE categories and/or different size classes. We should try to collect the data for each “daughter” and include it in the final data set as separate enterprises with the same identity and weight as the “mother”.

Technically we have a situation where the “mother” is a cluster of “daughters” and we simply use the known sampling theory for that situation. This will work only if it is possible to identify and collect the data from all “daughters”. If this is not the case then we have to use imputation for the missing “daughters” or a weight that compensates for them. However, this is only possible when we know the “daughters”.

When we have no information about the missing “daughters” except their number, then a simple weight is M_k/m_k , where M_k is the total number of “daughters” and m_k is number of responding “daughters” for “mother” k. This weight is multiplied by w_k described above.

Another possible weight is the inverted proportion of the number of employees in the responding “daughters”.

When the split occurs after 2005 it should be treated as a measurement problem. The collected data about the frame enterprise is not the “true” values and should therefore be corrected.

Merged enterprises

In the case of merged enterprises the problem should be handled in different ways depending on whether the merge occurs before year 2006 or not. First, assume that it occurs before year 2006.

When two or more frame enterprises merge into one single unit it means that this unit is included in the sample if at least one of its (eligible) frame enterprises is included. The increased inclusion probability should be adjusted for. One simple way is to use the inverted number of frame enterprises that belong to the unit as the weight, which is multiplied by w_k above.

Suppose that 3 frame enterprises will be merged into one unit then the weight $a_k=1/3$ will be appropriate for the frame enterprise k . Now, suppose that two of the three frame enterprises are included in the sample, then both should be included in the final data set with identical variable values and the additional weight $1/3$, which may be included in the final weight.

Although this procedure will give unbiased estimates it may be better to use the inclusion probabilities $\pi_k=m_h/N_h$ such that $a_k = \pi_k / \sum \pi_j$, where the sum is over the merged eligible frame enterprises. If there are 3 merged enterprises, all belonging to the same stratum, then $a_k=1/3$.

It is usually an advantage to use $a_k = \pi_k / \sum \pi_j$. Suppose that a small enterprise, where m_h/N_h in general is large, is merged with a large enterprise, where m_h/N_h in general is small, and assume that the small enterprise is selected. Then the quantitative variables will usually have large values due to the influence from the large enterprise and the sampling weight w_k will also be large due to the selection of the small enterprise. By using $a_k = \pi_k / \sum \pi_j$ instead of $1/2$ the interaction effect is eliminated to a large extent.

When the merge occurred after 2005 it should be treated as a measurement problem. The collected data about the frame enterprise is not the “true” values and has to be corrected for example by splitting the values that were obtained from the merged enterprise, i.e. imputation.

Recommendations

When no additional information is used, then the weights should be constructed as $w_k = N_h/m_h$ for enterprise k in the sample stratum h , where N_h is the population size from the frame (including the overcoverage) and m_h is the number of responding enterprises with the observed overcoverage excluded.

When only one variable is used as additional information, then the weights should be constructed as

$$w_k = (N_h/m_h) \left(N_g / \hat{N}_g \right) \text{ or } w_k = (N_h/m_h) \left(t_{xg} / \hat{t}_{xg} \right),$$

for an enterprise k that belongs to stratum h , group g , where g is some arbitrary grouping of the population or the population itself and m_h is defined above.

When more than one variable is used as additional information one should avoid a situation with several weights. One way of doing this is to implement calibration weights. In the case of two variables, one that count the number of enterprises and one that contains the number of employees (according to SBR for example), the weight should be constructed as

$$w_k = (N_h/m_h)(N_g/\hat{N}_g) \left(1 + (x_k - \tilde{x}_g) \frac{(\bar{x}_g - \tilde{x}_g) \hat{N}_g}{\sum_h \frac{N_h}{m_h} \sum_{ng} (x_k - \tilde{x}_g)^2} \right),$$

for an enterprise k that belongs to stratum h , group g , where g is some arbitrary grouping of the population or the population itself and m_h is defined above.

Suggested methods for imputing missing quantitative variables

Data sets required for imputation (BOLD in 2 = data missing in 1)	Method of imputation
1. Total persons employed 2. Total persons employed and total hours worked	To impute missing hours worked calculate average hours worked per person employed from (2) and multiply by total persons employed in each enterprises of (1).
1. Total hours worked 2. Total hours worked and total labour costs	To impute missing labour costs calculate average labour cost per hours worked from (2) and multiply by total hours worked in each enterprises of (1).
1. Total persons employed 2. Total persons employed by sex	To impute persons employed by sex calculate percentage shares by sex from (2) and multiply by persons employed in each enterprise in (1)
1. Total persons employed 2. Total persons employed by occupational groups	To impute persons employed by occupational groups calculate percentage shares by occupational groups from (2) and multiply by persons employed in each enterprise in (1)
1. Total persons employed 2. Total persons employed by age categories	To impute persons employed by age categories calculate percentage shares by age categories from (2) and multiply by persons employed in each enterprise in (1)
1. Persons employed by sex and total participants in CVT courses 2. Persons employed by sex and participants in CVT courses by sex	To impute the participants by sex calculate the participation rates by sex from (2) and then, first, apply these rates to the persons employed by sex in (1) and then, secondly, scale the results to ensure that they agree with the total participants in courses for each enterprise in (1).
1. Persons employed by occupational groups and total participants in CVT courses 2. Persons employed by occupational groups and participants in CVT courses by occupational groups	To impute the participants by occupational groups calculate the participation rates by occupational groups from (2) and then, first, apply these rates to the persons employed by occupational groups in (1) and then, secondly, scale the results to ensure that they agree with the total participants in courses for each enterprise in (1).
1. Persons employed by age categories and total participants in CVT courses 2. Persons employed by age categories and participants in CVT courses by age categories	To impute the participants by age categories calculate the participation rates by age categories from (2) and then, first, apply these rates to the persons employed by age categories in (1) and then, secondly, scale the results to ensure that they agree with the total participants in courses for each enterprise in (1).
1. Total participants in CVT courses 2. Total participants in CVT courses and total hours in CVT	To impute the total hours on courses calculate the hours per participant from (2) and multiply by total participants for each enterprise in (1).
1. Total participants by sex and total hours 2. Total participants by sex and total hours by sex	To impute the hours of training by sex calculate the average hour per participants for each sex from (2) and then first, apply these rates to the participants by sex for each enterprise in (1) and then, secondly, scale the results to ensure that they agree with the total hours of training for each enterprise.
Enterprises offering both internal and external CVT courses 1. Total hours of training 2. Total hours of training and hours for internal and external courses	To impute the hours of training for internal and external courses, calculate the distribution of hours of training by internal and external courses from (2) and apply to the total hours of training for each enterprise in (1).
1. Total hours of CVT training 2. Total hours of CVT training and hours of CVT training by field of training	To impute hours of training by subject calculate the percentage distribution of hours of training by subject from (2) and apply these percentages to the total hours of training for each enterprise in (1).

Data sets required for imputation (BOLD in 2 = data missing in 1)	Method of imputation
1. Hours of CVT training in external courses 2. Hours of CVT training in external courses and hours of CVT training by provider	To impute the hours of external training by type of provider, calculate the percentage distribution of hours of training by type of provider from (2) and apply these percentages to the total hours of external training for each enterprise in (1).
Enterprises with a training centre offering only internal courses 1. Hours of CVT training 2. Hours of CVT training and total CVT training costs	To impute the total training costs, calculate the average hourly training cost from (2) and multiply by the hours of training for each enterprise in (1).
Enterprises without a training centre offering only internal courses 1. Hours of CVT training 2. Hours of CVT training and total CVT training costs	To impute the total training costs, calculate the average hourly training cost from (2) and multiply by the hours of training for each enterprise in (1).
Enterprises offering only external courses 1. Hours of CVT training 2. Hours of CVT training and total CVT training costs	To impute the total training costs, calculate the average hourly training cost from (2) and multiply by the hours of training for each enterprise in (1).
Enterprises with a training centre offering both internal and external courses 1. Hours of CVT training 2. Hours of CVT training and total CVT training costs	To impute the total training costs, calculate the average hourly training cost from (2) and multiply by the hours of training for each enterprise in (1).
Enterprises without a training centre offering both internal and external courses 1. Hours of CVT training 2. Hours of CVT training and total CVT training costs	To impute the total training costs, calculate the average hourly training cost from (2) and multiply by the hours of training for each enterprise in (1).
Enterprises with a training centre offering only internal courses 1. Total CVT training costs 2. Total CVT training costs by type of cost	To impute the types of training costs calculate the percentage distribution by type of cost from (2) and apply these percentages to the total training costs for each enterprise in (1).
Enterprises without a training centre offering only internal courses 1. Total CVT training costs 2. Total CVT training costs by type of cost	To impute the types of training costs calculate the percentage distribution by type of cost from (2) and apply these percentages to the total training costs for each enterprise in (1).
Enterprises offering only external courses 1. Total CVT training costs 2. Total CVT training costs by type of cost	To impute the types of training costs calculate the percentage distribution by type of cost from (2) and apply these percentages to the total training costs for each enterprise in (1).
Enterprises with a training centre offering both internal and external courses 1. Total CVT training costs 2. Total CVT training costs by type of cost	To impute the types of training costs calculate the percentage distribution by type of cost from (2) and apply these percentages to the total training costs for each enterprise in (1).

Data sets required for imputation (BOLD in 2 = data missing in 1)	Method of imputation
Enterprises without a training centre offering both internal and external courses 1. Total CVT training costs 2. Total CVT training costs by type of cost	To impute the types of training costs calculate the percentage distribution by type of cost from (2) and apply these percentages to the total training costs for each enterprise in (1).
1. Total number of individuals registered in IVT 2. Total number of individuals registered in IVT and total costs for IVT activities	To impute missing costs for IVT activities calculate average training costs per total number of individuals registered in an IVT activity (2) and multiply by total number of individuals registered in an IVT activity for each enterprise in (1).
1. Total costs for IVT activities 2. Total costs for IVT activities by type of cost	To impute the types of costs for IVT activities calculate the percentage distribution by type of cost for IVT activities from (2) and apply these percentages to the total costs for IVT activities for each enterprise in (1).

Annex 1:
**Derivation of the formula, which is used for calculating
sample sizes in each stratum**

by

Statistics Sweden

Derivation of the formula

The C-value is the desired maximum half-length of the confidence interval, which means that a C-value equal to 0,1 corresponds with a maximum absolute sample error equal to 0,05.

The derivation of the formula, which is used for calculating sample sizes in each stratum, follows below. In CVTS, we are mainly interested in estimating proportions, for example the proportion of training enterprises offering external courses, p_h .

$$C = 2 \sqrt{p_h (1 - p_h) \left(\frac{1}{n_h r_h} - \frac{1}{N_h} \right)}$$

$$\frac{C^2}{2^2} = p_h (1 - p_h) \left(\frac{1}{n_h r_h} - \frac{1}{N_h} \right)$$

Assume that $p_h = 0,5$

$$\Rightarrow C^2 = \frac{1}{n_h r_h} - \frac{1}{N_h}$$

$$\frac{1}{n_h r_h} = C^2 + \frac{1}{N_h}$$

$$n_h r_h = 1 / \left(C^2 + \frac{1}{N_h} \right)$$

$$n_h = 1 / \left(C^2 + \frac{1}{N_h} \right) / r_h$$

Allocate the sample sizes in such a way that the expected number of training enterprises will be roughly the same in each stratum.

Annex 2:
Excel application for calculating
and allocation of sample sizes

by

Statistics Sweden

Instructions

This EXCEL book has been worked out in order to facilitate the calculations of sample sizes for CVTS3.

Note: in this application it is possible to input both estimated response rates as well as estimated proportions of trainers in each size*nace class.

Estimated response rates should if possible not include over-coverage.

If there are any problem with the calculation, contact Eurostat.

Step 1

Change country name.

The number of enterprises within each cell, defined by the cross-classification of NACE and size is put into cells B40:D59.

These numbers should come from the sampling frame in each country.

Example: For NACE 21-22, (Paper and printing) the number of enterprises in the Swedish Central Register of Enterprises are 630, 179 and 66 within the three size classes.

Step 2

The proportion of "training enterprises" within each (20 x 3) cell is put into cells H40:J59.

Countries participating in CVTS2 can normally estimate these proportions from that survey, however, additional information (if available) should be used if it will improve the quality in these figures.

For countries that didn't participate in CVTS2 the figures from another country, which can be assumed to have similar pattern regarding the proportion of training enterprises, can be used. (For more information see the document CVTS3_raking-ratio.doc)

Example: For NACE 21-22, the proportion of training enterprises are estimated to be 0.91, 0.99 and 1.0.

Step 3

The number of sampling units (enterprises) needed in the sample to get the desired maximum half length of the (95%) confidence interval is computed for each cell (B10:E30) and the expected size of the half (95%) confidence intervals are shown in cells H10:K30.

The wanted half length of the confidence interval (the C-value) is put into cell E2.

The expected response rate (exkl. overcoverage if possible) is put into cells B74:D93.

Example: For NACE 21-22, the proportion of response units (excluding over-coverage) are estimated to be 0.47, 0.41 and 0,64.

The minimum number of sampled units in each cell is put into cell E4.

Note

The confidence intervals are only approximate and are calculated for a proportion of enterprises within the population of "training enterprises", for example the proportion of training enterprises offering external courses.

The proportion 0.5 gives this maximum confidence interval.

Illustration for Sweden

1) **C-value = 0,1**

Min. number of sampling units in stratum = 10

Sample sizes for C = 0,1

NACE	Size			
	10-49	50-249	250-	All
All	4183	2284	794	7261
C	52	9	3	64
15-16	216	104	36	356
17-19	148	26	7	181
21-22	198	160	62	420
23-26	200	112	54	366
27-28	202	132	42	376
29-33	236	153	76	465
34-35	155	87	43	285
20, 36-37	223	122	26	371
E	105	51	17	173
F	266	139	43	448
50	277	131	21	429
51	232	194	57	483
52	423	200	61	684
H	361	148	15	524
60-63	326	164	57	547
64	60	25	22	107
65-66	141	103	34	278
67	111	24	3	138
K+O	251	200	115	566

Expected (1/2) approx. 95% confidence interval

NACE	Size			
	10-49	50-249	250-	All
All	0,036	0,035	0,033	0,030
C	0,129	0,394	0,408	0,122
15-16	0,100	0,116	0,107	0,078
17-19	0,100	0,187	0,134	0,085
21-22	0,100	0,100	0,100	0,074
23-26	0,100	0,100	0,101	0,069
27-28	0,100	0,100	0,109	0,084
29-33	0,100	0,100	0,100	0,073
34-35	0,100	0,100	0,101	0,065
20, 36-37	0,100	0,100	0,132	0,078
E	0,100	0,100	0,271	0,074
F	0,100	0,100	0,116	0,090
50	0,100	0,100	0,132	0,085
51	0,100	0,100	0,100	0,087
52	0,100	0,100	0,103	0,088
H	0,100	0,100	0,285	0,090
60-63	0,100	0,100	0,110	0,083
64	0,208	0,179	0,096	0,113
65-66	0,100	0,114	0,171	0,070
67	0,120	0,242	0,577	0,106
K+O	0,100	0,100	0,100	0,083

**Number of enterprises in the sample frame,
by NACE and size**

NACE	Size			
	10-49	50-249	250-	All
All	26026	4408	941	31375
C	52	9	3	64
15-16	486	104	36	626
17-19	155	26	7	188
21-22	630	179	66	875
23-26	508	212	55	775
27-28	1.441	246	42	1729
29-33	1.190	353	104	1647
34-35	221	105	44	370
20, 36-37	750	200	26	976
E	147	61	17	225
F	3.181	266	43	3490
50	904	149	21	1074
51	3.029	393	64	3486
52	2.550	274	61	2885
H	1.697	151	15	1863
60-63	1.964	291	57	2312
64	60	25	25	110
65-66	161	103	34	298
67	111	24	3	138
K+O	6.789	1237	218	8244

Estimated proportions of "training enterprises"

by NACE and size

NACE	Size			
	10-49	50-249	250-	All
All	0,88	0,99	0,99	0,903
C	0,85	1,00	1,00	0,880
15-16	0,82	1,00	0,97	0,861
17-19	0,78	0,95	1,00	0,815
21-22	0,91	0,99	1,00	0,929
23-26	0,82	1,00	1,00	0,883
27-28	0,93	0,96	1,00	0,940
29-33	0,82	0,98	1,00	0,866
34-35	0,84	1,00	0,95	0,898
20, 36-37	0,89	0,97	1,00	0,909
E	1,00	1,00	1,00	1,000
F	0,81	1,00	1,00	0,830
50	0,95	0,98	1,00	0,958
51	0,93	1,00	1,00	0,938
52	0,93	0,98	1,00	0,939
H	0,82	1,00	1,00	0,832
60-63	0,80	0,98	1,00	0,824
64	0,74	1,00	0,93	0,843
65-66	1,00	1,00	1,00	1,000
67	1,00	1,00	1,00	1,000
K+O	0,92	1,00	0,99	0,934

**Estimated response rates
(exkl. over-coverage)
by NACE and size**

NACE	Size		
	10-49	50-249	250-
C	0,57	0,42	0,67
15-16	0,45	0,42	0,71
17-19	0,47	0,53	0,89
21-22	0,47	0,41	0,64
23-26	0,49	0,61	0,65
27-28	0,49	0,55	0,67
29-33	0,47	0,52	0,67
34-35	0,50	0,59	0,72
20, 36-37	0,44	0,56	0,69
E	0,57	0,74	0,44
F	0,44	0,52	0,63
50	0,34	0,46	0,73
51	0,45	0,41	0,68
52	0,24	0,37	0,61
H	0,32	0,41	0,45
60-63	0,36	0,46	0,59
64	0,34	0,56	0,93
65-66	0,44	0,43	0,50
67	0,38	0,42	0,50
K+O	0,43	0,46	0,60

- 2) **C-value = 0,2**
 Min. number of sampling units in stratum = 10

Sample sizes for C = 0,2

NACE	Size			
	10-49	50-249	250-	All
All	1254	822	421	2497
C	33	9	3	45
15-16	64	48	21	133
17-19	56	24	7	87
21-22	56	55	28	139
23-26	58	37	26	121
27-28	53	42	24	119
29-33	63	46	30	139
34-35	53	34	23	110
20, 36-37	62	41	18	121
E	38	24	17	79
F	69	44	25	138
50	75	47	16	138
51	59	57	26	142
52	109	63	29	201
H	95	53	15	163
60-63	85	51	29	165
64	60	23	14	97
65-66	49	47	29	125
67	53	24	3	80
K+O	64	53	38	155

Expected (1/2) approx. 95% confidence interval

NACE	Size			
	10-49	50-249	250-	All
All	0,072	0,069	0,063	0,059
C	0,198	0,394	0,408	0,170
15-16	0,199	0,200	0,200	0,153
17-19	0,200	0,203	0,134	0,162
21-22	0,200	0,200	0,202	0,148
23-26	0,201	0,199	0,201	0,138
27-28	0,201	0,201	0,197	0,169
29-33	0,201	0,200	0,200	0,146
34-35	0,199	0,202	0,199	0,130
20, 36-37	0,200	0,200	0,205	0,156
E	0,199	0,199	0,271	0,142
F	0,199	0,199	0,200	0,179
50	0,200	0,200	0,194	0,170
51	0,201	0,200	0,202	0,174
52	0,200	0,199	0,201	0,177
H	0,200	0,200	0,285	0,180
60-63	0,200	0,200	0,202	0,167
64	0,208	0,196	0,198	0,123
65-66	0,201	0,200	0,199	0,131
67	0,200	0,242	0,577	0,167
K+O	0,199	0,200	0,199	0,165

**Number of enterprises in the sample frame,
by NACE and size**

NACE	Size			
	10-49	50-249	250-	All
All	26026	4408	941	31375
C	52	9	3	64
15-16	486	104	36	626
17-19	155	26	7	188
21-22	630	179	66	875
23-26	508	212	55	775
27-28	1.441	246	42	1729
29-33	1.190	353	104	1647
34-35	221	105	44	370
20, 36-37	750	200	26	976
E	147	61	17	225
F	3.181	266	43	3490
50	904	149	21	1074
51	3.029	393	64	3486
52	2.550	274	61	2885
H	1.697	151	15	1863
60-63	1.964	291	57	2312
64	60	25	25	110
65-66	161	103	34	298
67	111	24	3	138
K+O	6.789	1237	218	8244

NACE	Size			
	10-49	50-249	250-	All
All	0,88	0,99	0,99	0,903
C	0,85	1,00	1,00	0,880
15-16	0,82	1,00	0,97	0,861
17-19	0,78	0,95	1,00	0,815
21-22	0,91	0,99	1,00	0,929
23-26	0,82	1,00	1,00	0,883
27-28	0,93	0,96	1,00	0,940
29-33	0,82	0,98	1,00	0,866
34-35	0,84	1,00	0,95	0,898
20, 36-37	0,89	0,97	1,00	0,909
E	1,00	1,00	1,00	1,000
F	0,81	1,00	1,00	0,830
50	0,95	0,98	1,00	0,958
51	0,93	1,00	1,00	0,938
52	0,93	0,98	1,00	0,939
H	0,82	1,00	1,00	0,832
60-63	0,80	0,98	1,00	0,824
64	0,74	1,00	0,93	0,843
65-66	1,00	1,00	1,00	1,000
67	1,00	1,00	1,00	1,000
K+O	0,92	1,00	0,99	0,934

**Estimated response rates)
(exkl. over-coverage
by NACE and size**

NACE	Size		
	10-49	50-249	250-
C	0,57	0,42	0,67
15-16	0,45	0,42	0,71
17-19	0,47	0,53	0,89
21-22	0,47	0,41	0,64
23-26	0,49	0,61	0,65
27-28	0,49	0,55	0,67
29-33	0,47	0,52	0,67
34-35	0,50	0,59	0,72
20, 36-37	0,44	0,56	0,69
E	0,57	0,74	0,44
F	0,44	0,52	0,63
50	0,34	0,46	0,73
51	0,45	0,41	0,68
52	0,24	0,37	0,61
H	0,32	0,41	0,45
60-63	0,36	0,46	0,59
64	0,34	0,56	0,93
65-66	0,44	0,43	0,50
67	0,38	0,42	0,50
K+O	0,43	0,46	0,60

Annex 3:
Algorithm for the calculation of the proportion
of training enterprises in a table where only
the margins are known

by

Statistics Sweden

A simple algorithm for the calculation of the proportion of training enterprises in a table where only the margins are known

Introduction

We want to calculate the proportion of "training enterprises" within each cell of a 20×3 table, defined by NACE group (20 groups) and size (3 classes). It is assumed that the number of enterprises are known for each cell.

If only the margin proportions are known, we have to estimate the body of the table from the available data. It is assumed that the proportions depends on NACE group and size class but not on the joint occurrence of NACE and size.

Notations

Let,

N_{ij} be the known number of enterprises in cell ij , $i=1, \dots, 20$ (NACE) and $j=1, 2, 3$, (size),

$N_{i.} = \sum_j N_{ij}$, i.e. number of enterprises in NACE group i ,

$N_{.j} = \sum_i N_{ij}$, i.e. number of enterprises in size class j ,

p_i be the proportion of training enterprises in NACE group i ,

p_j be the proportion in size class j ,

$M_{i.} = N_{i.} \times p_i$ be the number of training enterprises in NACE group i ,

$M_{.j} = N_{.j} \times p_j$ be the number of training enterprises in size class j .

Algorithm

1. $m_{ij}^{(0)} = N_{ij}$, $v=0$.

2. $m_{ij}^{(2v+1)} = m_{ij}^{(2v)} \frac{M_{i.}}{\sum_j m_{ij}^{(2v)}}$.

3. $m_{ij}^{(2v+2)} = m_{ij}^{(2v+1)} \frac{M_{.j}}{\sum_i m_{ij}^{(2v+1)}}$.

4. If $\max_{i,j} \left(\left| m_{ij}^{(2v)} - m_{ij}^{(2v+2)} \right| \right) < \varepsilon$ (e.g. 0.001) continue with step 5, else let $v=v+1$ and continue with step 2.

5. If $m_{ij}^{(2v+2)} > N_{ij}$ then let $m_{ij}^{(2v+2)} = 0$, $M_{i.} = M_{i.} - N_{ij}$, $M_{.j} = M_{.j} - N_{ij}$, $v=v+1$ and continue with step 2, else continue with step 6.

6. Let $p_{ij} = \frac{m_{ij}^{(2v+2)}}{N_{ij}}$, if $m_{ij}^{(2v+2)} = 0$ and $N_{ij} > 0$ then let $p_{ij} = 1$.

The algorithm should converge rather fast, step 2-4 will probably need six or seven iterations.

Annex 4:
SAS-script with the algorithm for the
calculation of the proportion of
training enterprises

by

Statistics Sweden

```

/* input the number of enterprises for each combination of NACE and size
(20*3) plus the proportion of training enterprises. Note the last row
wich contains the proportions for each size class and for the total*/

```

```

data indat;
input m1 m2 m3 pi;
cards;
52          9          3 .88
486         104        36 .861
155         26         7 .815
630         179        66 .929
508         212        55 .883
1441 246 42 .940
1190 353 104 .866
221         105        44 .898
750         200        26 .909
147         61         17 1
3181 266 43 .83
904         149        21 .958
3029 393 64 .938
2550 274 61 .939
1697 151 15 .832
1964 291 57 .824
60 25 25 .843
161         103        34 1
111         24         3 1
6789 1237 218 .934
0.88 0.99 0.99 0.903
;
run;

```

```

/* do the raking ratio to esimate the proportion of training enterprises
within each cell */

```

```

data rakrat;
set indat end=_eof;
array nij(20,3);
array ny(20,3);
array nold(20,3);
array ni(20);
array nj(3);
array m(3);
array p(3);
retain nij ny nold ni nj;
retain eps 0.001;
if _n_<=20 then do;
isum=0;
do j=1 to 3;
nij(_n_,j)=m(j); nold(_n_,j)=m(j);
nj(j)+m(j);
isum+m(j);
end;
end;

```

```

ni(_n_)=pi*isum;
end;
else do j=1 to 3;
  nj(j)=nj(j)*m(j);
end;
if _eof then do;
  put 'Iter Maxdiff';
  iter=0;
l1: iter+1;
  do i=1 to 20;
    isum=0;
    do j=1 to 3;
      isum+nold(i,j);
    end;
    do j=1 to 3;
      ny(i,j)=nold(i,j)*ni(i)/isum;
    end;
  end;
  do j=1 to 3;
    jsum=0;
    do i=1 to 20;
      jsum+ny(i,j);
    end;
    do i=1 to 20;
      ny(i,j)=ny(i,j)*nj(j)/jsum;
    end;
  end;
  check=0; maxdiff=0;
  do i=1 to 20;
  do j=1 to 3;
    maxdiff=max(maxdiff,abs(nold(i,j)-ny(i,j)));
    check+abs(nold(i,j)-ny(i,j))>eps;
    nold(i,j)=ny(i,j);
  end;
  end;
  put iter maxdiff;
  if check>0 and iter< 20 then goto l1;
  iter=0; put;
  do i=1 to 20;
  do j=1 to 3;
    if nold(i,j)>nij(i,j) then do;
      check=1;
      nold(i,j)=0;
      ni(i)=ni(i)-nij(i,j);
      nj(j)=nj(j)-nij(i,j);
    end;
  end;
  end;
  if check=1 then goto l1;
  else do;
    do i=1 to 20;

```

```
do j=1 to 3;
  if nij(i,j)>0 then do;
    p(j)=nold(i,j)/nij(i,j);
    if p(j)=0 then p(j)=1;
  end;
  else p(j)=0;
end;
output;
end;
end;
keep p1-p3;
run;
proc print round; run;
```

/* The following code can be used to transfer the proportions to EXCEL,
you may need to change the path to the EXCEL file*/

```
/*
filename dde1 dde 'EXCEL|c:\data\sas\cvts2\cvts2_allocation.xls:C=0.1!R40C8:R59C10';
filename dde2 dde 'EXCEL|c:\data\sas\cvts2\cvts2_allocation.xls:C=0.2!R40C8:R59C10';

data _temp_;
set rakrat;
file dde1;
put p1 p2 p3;
file dde2;
put p1 p2 p3;
run;
*/
```