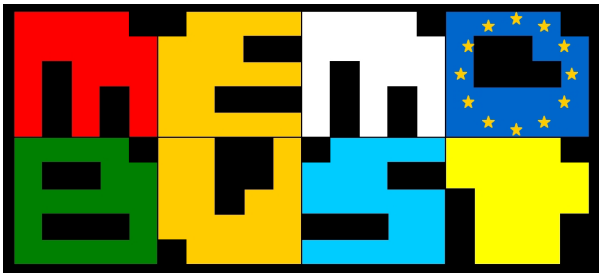

This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Different Coding Strategies

Contents

General section

1. Summary

Coding of textual responses of statistical surveys, if not made completely manually, can be done in a completely automated way ("automated coding" or batch coding – AUC) or with computer support ("computer-assisted coding" or interactive coding – CAC). The decision which is the most suitable coding approach to be adopted in a survey depends on four correlated factors: the survey technique, the amount of data to be coded, the interview length and the structure of the classification. The combination of these factors can be analysed in two alternative situations, deriving from the moment of the implementation of the coding activity: coding phase during data collection (possible only for CAC) or coding phase after data collection. Elements to define a strategy are provided.

2. General description

Generally speaking, the coding activity, if not made completely manually, can be performed according to two coding procedures (Lyberg and Dean, 1992), using computers in two possible ways:

1. "automated coding" or batch coding (AUC);
2. "computer-assisted coding" or interactive coding (CAC).

1. (AUC). The computer assigns codes to the verbal responses working in 'batch' processing. As this technique cannot be expected to assign a code to all the input statements, a manual coding or an assisted coding procedure is required after this step to assign codes to the non-coded responses.
2. (CAC). The operator assigns codes working interactively with the computer, supporting him in 'navigating' the dictionary while searching for codes to be assigned to the input descriptions. For example, when the operator fills in the verbal response on the PC, the machine will show him all dictionary descriptions that could match the input statement (only one description is shown if an exact match exists); the operator should choose one of them, assigning the most suitable code. Thus a CAC system combines the human mind with the computer potential.

The difference between the two procedures lies in their final aim and coding approach. The final aim of AUC procedure is to maximise the number of unique codes assigned automatically to the input statements, whereas the CAC aims at providing the operator with as much assistance as possible. As a consequence, the coding approach of the two systems is different:

- AUC aims at extracting a single description from the dictionary matching the input statement;
- CAC shows different descriptions (also slightly different from each other); it is important to remember that the operator works interactively with the PC and can navigate through the descriptions shown, choosing the most suitable one. Besides, CAC allows the usage of other survey information to support the assignment of codes.

These two procedures allow to manage the coding activity at two different moments of the data collection phase:

- AUC can be used after the interview, that is, when data collection is over;

- CAC can be used both after the interview (by coders) or during the interview (by the interviewer or by the respondent).

The decision which is the most suitable coding approach to be adopted in a survey depends on different correlated factors (Macchia and Murgia, 2002) that is:

1. the survey technique:
    - computer-assisted with the interviewer (CATI – *Computer Assisted Telephone Interviewing,* CAPI – *Computer Assisted Personal Interviewing*);
    - computer-assisted without the interviewer (CASI – *Computer Assisted Self-Interviewing*);
    - traditional Paper and Pencil Technique (PAPI);
2. the amount of data to be coded (Appel and Hellerman, 1983):
    - a large number (e.g., like a census);
    - a small number (like sample surveys on a few thousands of units);
3. the interview length in terms of time necessary to fill in the questionnaire:
    - short interview (less than 15 minutes);
    - long interview (more than 15 minutes);
4. the structure of the classification in conjunction with the variability of the verbal responses:
    - simple classification structure;
    - complex classification structure and high variability of verbal responses.

The structure of a classification can be represented as a tree with branches, sub-branches and leaves. Branches represent general levels of classification that are hierarchically higher than sub-branches and leaves, that represent detailed levels of classification. Therefore, a simple classification structure means a tree with branches, none or few sub-branches and no leaves, whereas a complex structure corresponds with a tree with all these components. Examples of a simple and a complex classification structure are the "*Country Classification*" and the "*Classification of economics activities*", respectively.

Table 1: Example of classifications with different levels of complexity

Simple Classification	**Complex Classification**
Country	**Economic activities**
1. France	01. Crop and animal production, hunting and related service activities
2. Germany	01.1 Growing of non-perennial crops
3. Great Britain	01.11 Growing of cereals (except rice), leguminous crops and oil seeds
4. Italy	01.12 Growing of rice
5. Spain ….	01.13 Growing of vegetables and melons, roots and tubers

Combining the above-mentioned factors, it is possible to see whether one procedure is more suitable than the other. This combination can be analysed in two alternative situations, based on the moment of the implementation of the coding activity:

1. coding phase during data collection;
2. coding phase after data collection.

2.1 Coding phase during data collection

The following table shows which is the most appropriate coding solution to adopt when computer data capturing is performed by an interviewer.

Table 2: Survey technique: computer-assisted with the interviewer (CATI, CAPI)

	Interview length	
Classification structure	**Short**	**Long**
• **Simple**	CAC	CAC
• **Complex & high response variability**	CAC	No data coding (coding after data collection)

In general, as can be seen, it is advisable to use CAC during the interview with the interviewer because:

- coded data are available for processing as soon as data collection is over;
- a higher quality of the coded data is also guaranteed by the contact with the respondent who can provide the interviewer with further explanations on the given answer, if needed;
- the previous point implies that, during this activity, the interviewer will 'train himself' in getting an answer with sufficient information to be coded.

But, if the interview is long and the coding activity during the interview would increase its duration, it is better not to use CAC and code the data at the end of data collection (even more so if the classifications are complex). In this way the following can be avoided:

- too large a number of uncompleted interviews – respondents deny their co-operation to the operator;
- errors in coding, due to the interviewer's need to speed up the interview.

As shown in table 3, the situation is different when a computer-assisted technique *without interviewer* is adopted for data capturing.

Table 3: Survey technique: computer-assisted without the interviewer (CASI)

Classification structure	
• **Simple**	CAC
• **Complex & high response variability**	No data coding (coding after data collection)

In this case, the coding activity chosen during the interview – done by the respondent himself, not being an expert of the classification – strictly depends on the classification structure. It is advisable to use CAC only if:

- the classification structure is simple;
- the codes to be assigned belong to only one branch of the classification, that is to a high hierarchical level.

2.2 Coding phase after data collection

Whatever technique is used to collect data (CATI, CAPI, CASI, or PAPI), when they are stored in a database, the amount of data to be coded plays a fundamental role in deciding which coding procedure can be adopted:

- for a large amount of data it is advisable to use AUC and subsequently CAC for the non-coded cases;
- for a small amount of data and simple classification it is better to apply AUC;
- for a small amount of data, complex classification and high response variability it is more convenient to adopt CAC.

The following table summarises what was stated before.

Table 4: Coding activity after data collection

	Quantity of data/statements to be coded	
Classification structure	**Large number**	**Small number**
• **Simple**	AUC + CAC	AUC
• **Complex & high response variability**	AUC + CAC	CAC

3. Design issues

4. Available software tools

Different tools have been developed by statistical offices to be used to code their survey data. The tow mentioned here are completely generalised, meaning that they do depend neither on the language used nor on the classification:

- for automatic coding – ACTR from Statistics Canada (Wenzowski, 1988), recently replaced by GCode;
- for computer-assisted coding – Blaise from CBS for interactive coding.

5. Decision tree of methods

6. Glossary

For definitions of terms used in this module, please refer to the separate "Glossary" provided as part of the handbook.

7. References

Appel, M. and Hellerman, E. (1983), Census Bureau Experience with Automated Industry and Occupation Coding. *Proceedings of Section on Survey Research Methods*, American Statistical Association, 32–40.

BLAISE for Windows 4.5 Developer's Guide (2002).

Lyberg, L. and Dean, P. (1992), Automated Coding of Survey Responses: an international review. Conference of European Statisticians, Work session on Statistical Data Editing, Washington DC.

Macchia, S. and Murgia, M. (2002), Coding of textual responses: various issues on automated coding and computer assisted coding. Journée d'Analyse des Données Textuelles JADT, Saint Malo.

Wenzowski, M. J. (1988), ACTR – A Generalised Automated Coding System. *Survey Methodology* **14**, 299–308.

Interconnections with other modules

8. Related themes described in other modules

1.

9. Methods explicitly referred to in this module

1.

10. Mathematical techniques explicitly referred to in this module

1.

11. GSBPM phases explicitly referred to in this module

1. GSBPM sub-process 5.2

12. Tools explicitly referred to in this module

1.

13. Process steps explicitly referred to in this module

1.

Administrative section

14. Module code

Coding-T-Different Coding Strategies

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	20-07-2012	first version	Stefania Macchia	Istat (Italy)
0.2	21-11-2012	second version (following first revision)	Stefania Macchia	Istat (Italy)
0.3	25-10-2013	third version (following EB review 04-10-2013)	Stefania Macchia	Istat (Italy)
0.3.1	28-10-2013	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 18:07