

Chapter 12

Survey Code and Tabulation



Business Research Methods Verónica Rosendo Ríos Enrique Pérez del Campo Marketing Research

Verónica Rosendo Ríos ©

"Excellence is to do a common thing in an uncommon way"

BOOKER T. WASHINGTON

CONTENTS

- 1. Questionnaire codification
- 2. Questionnaire revision and transcription
- 2. Error depuration
- 3. Statistically adjusting the data



★ CODIFICATION:

Assigning a code, usually a number, to each possible response to each question.

Code: indication of the column position (field)
Field: represents a single item of data, eg. sex
Record: related fields, such as marital status, age, household size or occupation.

Normally each record has about 80 columns, although this is not a requirement.

★ CODIFICATION:

- <u>*Respondent code*</u> and the <u>record number</u> should appear on each record in the data.
- <u>Additional codes</u> for each respondent include:
 - project code, interviewer code, data and time codes, and validation code.
- ✓ Fixed field code, which means that <u>the number of</u> <u>records for each respondent is the same and that the</u> <u>same data appear in the same column(s) for all</u> <u>respondents</u>, are very desirable.

For example, a code of 9 may be used for as a missing value code.

★ CODIFICATION:

Coding structured or closed ended questions

Coding unstructured or open ended questions



1. Coding structured or closed ended questions

✓ Coding of structured questions is relatively easy, because the respondent options are predetermined. For example,

Do you have a currently valid passport? 1. Yes 2. No

Ves No	OW	
L it K	🛞 Value Labels	×
Don	Value Labels Value: 2 Value Label: Female Add Apply Bemove	<mark>€ OK</mark> Cancel <u>H</u> elp

For this question, a "yes" response is coded 1 and a "no" response 2.

Verónica Rosendo Ríos @ Because only one response is allowed and there are only two possible responses (1 or 2), a single column is sufficient.

In general, a single column is sufficient to code a structured question with a single response if there are fewer than 9 possible responses.

2.Coding unstructured or open ended questions

- The coding of unstructured or open ended questions is <u>more complex</u>. Respondents' verbatim responses are recorded on the questionnaire.
- ✓ Codes are then developed and assigned to these responses.
- ✓ Sometimes, based on previous projects or theoretical considerations, the researcher can develop the codes <u>before beginning the fieldwork</u>. Usually, this must wait until the completed questionnaires are received.

Then the researcher lists 50 to 100 responses to an unstructured question to identify the categories suitable for coding.
 Once codes are developed, the <u>coders should be trained</u> to assign the correct codes to the verbatim responses.

2.Coding unstructured or open ended questions

- ✓ The following <u>guidelines</u> are suggested for coding unstructured questions and questionnaires in general:
- **★** <u>**Category codes:**</u> should be mutually exclusive and collectively exhaustive.</u>

Categories are mutually exclusive if *each response fits into one and* \odot only one category code.

© Categories should *not overlap*.

©Categories are collectively exhaustive if every response fits into one the assigned category codes. This can be done by adding an of additional category code of "other" or "none of the above". few (10% or less) of the responses, however, should fall into Only a this category; the vast majority of the responses should be classified into meaningful categories.

2.Coding unstructured or open ended questions

- ★ The following guidelines are suggested for coding unstructured questions and questionnaires in general:
 - Category codes should be assigned for critical issues even if no one has mentioned them.

It may be important to know that no one has mentioned a particular response. For example, the management of a major consumer goods company was concerned about the packaging for a new brand of soap. Hence, packaging was included as a separate category in coding responses to the question, "What do you like least about this bar soap?" **Transcribing data**: from the questionnaires or coding sheets onto cds, usb, tapes or directly into computers by keypunching.

- ✓ If the data has been collected via CATI or CAPI, this step is not necessary because the data are entered directly into the computer as they are collected.
- Besides keypunching, the data can be transferred by using *mark sense forms, optical scanning, or computerized sensory analysis*. Mark sense forms requires responses to be recorded with a special pencil in a predesigned area coded for that response, and the data can then be read by a machine.
- Optical scanning involves direct machine reading of the codes and
 simultaneous transcription. An example of *optical scanning* is the
 transcription of *universal product code* (UPC) data at supermarket checkout

verónica Rosendo







Marketing Research

Comienzo y Finalización = CHR(8 Separador = CHR(1

Verification

- ✓ If keypunching is used, errors can occur, and it is necessary to verify the data set and revise questionnaires for keypunching errors.
- ✓ <u>A verifier machine and a second operator</u> are used for data revision and verification. The second operator re-punches the data from the coded questionnaires. The transcribed data from the two operators are compared record by record.
- ✓ Any discrepancy between the two sets of transcribed data is investigated to identify and correct for keypunching errors. Verification of the entire data set will double the time and cost of data transcription.
- ✓ Given the time and money constraints, and that experienced keypunch
 operators are quite accurate, it is sufficient to verify only 25 or 50% of
 the data on a random process.



wurketing Research

Transciption method

The selection of data transcription method is *guided by the type of interviewing method used* and the availability of equipment.

- ✓ If CATI or CAPI are used, the data are used for ordinary telephone, inhome, mall-intercept, and mail interviews.
- ✓ The use of computerized sensory analysis systems in personal interviews is increasing with the increasing use of grid-pads, handheld computers and tablets.
 - Optical scanning can be used in structured and repetitive surveys.



Marketing Research

1. DATA CLEANING Data cleaning includes:

Consistency checksTreatment of missing responses

3. ERRORS DEPURATION: DATA CLEANING

DATA CLEANING

1. Consistency checks

✓ **Out of range data** are inadmissible and must be corrected.

For example, respondents have been asked to express their degree of agreement with a series of lifestyle statements on a 1 to 5 scale. Assuming that 9 has been designated for missing values, data values of 0, 6, 7 and 8 are out of range. The correct responses can be determined by going back to the edited and coded questionnaires.

✓ Logically inconsistent responses.

For example, a respondent may indicate that she charges long distance calls to a calling card, although she does not have one. Or a respondent reports both unfamiliarity with and frequent usage of the same product.

Extreme values should be examined.

For example, an extremely low evaluation of a brand may be the result of the respondent indiscriminately circling 1s (on a 1 to 7 rating scale) on all attributes of this brand.

2. TREATMENT OF MISSING RESPONSES

Missing responses may be due to:

- ✓ Ambiguous answers.
- ✓ No answers.
- ✓ Answers not properly recorded.



Treatment of missing responses poses **problems**, particularly if the proportion of missing responses is more than 10%.

2. TREATMENT OF MISSING RESPONSES

Missing responses: What can we do????

- 1. Substitute a Neutral Value.
- 2. Substitute an imputed response.
- 3. Casewise deletion.
- 4. Pairwise deletion.



1. Substitute a Neutral Value. A neutral value, <u>typically the mean</u> response to the variable, is substituted for the missing responses. Thus, the mean of the variable remains unchanged, and other statistics such as correlations are not affected much.

Although this approach has some followers, the logic of substituting a mean value (say 4) for respondents who, if they had answered, might have used either high ratings (6 or 7) or low ratings (1 or 2) is questionable.

Substitute an imputed response. The respondents' pattern of responses to other questions are used to calculate a suitable response to the missing questions. The researcher attempts to infer from the available data the responses the individuals would have given if they had answered the questions. This can be done statistically by determining the relationship of the variable in question to other variables based on the available data.

For example, *product usage* could be related to *household size* for respondents who have provided data on both variables. Given respondent's household size, the missing product usage response for a respondent could then be calculated. This approach, however, requires considerable effort and can introduce serious bias.

2. TREATMENT OF MISSING RESPONSES

3. Casewise deletion.

Cases or respondents with any missing responses are discarded from the analysis.

- ✓ Because many respondents may have some missing responses, this approach <u>could result in a small sample</u>.
- Throwing away large amounts of data is undesirable because it is costly and time-consuming to collect data.
- ✓ Furthermore, respondents with missing responses could differ from respondents with complete responses in systematic ways. If so, casewise deletion <u>could seriously bias the results.</u>

4. Pairwise deletion.

Instead of discarding all cases with missing values, the researcher uses only the cases of respondents with complete responses for each calculation. As a result, different calculations in an analysis may be based on different sample sizes.

This procedure may be appropriate when:

- ✓ The sample size is large
- ✓ There are few missing respondents
- ✓ The variables are not highly related
- ✓ Yet, this procedure can produce unappealing or even unfeasible results

2. TREATMENT OF MISSING RESPONSES

- The different procedures for the treatment of missing responses may yield different results, particularly when the responses are not missing at random and the variables are related.
- ✓ Hence, missing responses should be kept to a minimum.
- ✓ The researcher should carefully consider the implications of the various procedures before selecting a particular method for the treatment of nonresponse.

✓ What can we do???

- 1. Weighting
- 2. Variable Re-specification
- 3. Scale-transformation
- 4. Pairwise deletion.





1. Weighting



- Each case or respondent in the database is assigned a weight to reflect its importance relative to other cases or respondents. The value 1.0 represents the un-weighted case. *The effect of the weighting is to increase or decrease the number of cases in the sample that possess certain characteristics.*
- ✓ To make the sample data more representative of a target population on specific characteristics.
- Adjust the sample so that the greater importance is attached to respondents with certain characteristics. If a study is conducted to determine *what modifications should be made to an existing product,* the researcher might want to attach greater weight to the opinions *of heavy users* of the product. This could be accomplished by assigning weights of 3.0 to heavy users and 2.0 to medium users, and 1.0 to light users and nonusers.

Because it destroys the self-weighting nature of the sample design, weighting should **only be applied with caution**. If used, the weighting procedure should be documented and made a part of the project report.

2.Variable re-specification

- Transformation of the data to create new variables or to modify existing variables.
 - ✓ For example, imagine ten response categories of users transformed into 4 types: heavy, medium, light and nonuser.
- An important re-specification procedure involves the use of *dummy variables*.
- Dummy variables are also called *binary, dichotomous, instrumental, or qualitative* variables. They are variables that may take **on <u>only two</u>** <u>values, such as 0 or 1</u>. The general rule is that to re-specify a categorical value with K categories, K-1 dummy variables are needed.

For example, in a study of consumer preferences for frozen foods, the respondents were classified as heavy users, medium, light and non users, and they were originally assigned codes of 4, 3, 2, and 1 respectively. This coding was not meaningful for several statistical analysis. To conduct these analysis, product usage was represented by three dummy variables, X_1 , X_2 , X_3 , as shown.

2. Variable re-specification

Example of dummy variables

Product Usage Category	Original Variable Code	Dummy X ₁	Variable X ₂	Code X ₃
Nonusers	1	1	0	0
Light users	2	0	1	0
Medium users	3	0	0	1
Heavy users	4	0	0	0

3. Scale Transformation

- ✓ Manipulation of scale values to ensure comparability with other scales or otherwise to make the data suitable for analysis.
- ✓ For example, image variables may be measured on a seven point semantic differential scale, attitude variables on a continuous rating scale, and lifestyle variables on a five-point Likert scale. Therefore, it is not meaningful to make comparisons across the measurement scales for any respondent.
- ✓ A more common transformation procedure is **standardization**.
- To standardize a scale X_i, we first subtract the mean, X_bar, from each score and then divide by the standard deviation, s_x. Thus, the standardized scale will have a mean of zero and a standard deviation of
 1.
 - Standardization allows the researcher to compare variables that have been measured using different types of scales.

Mathematically, standardized scores, z_i, may be obtained as

 $Z_i = (X_i - X_{bar})$





"Building a Multipetabyte Data System"

•What is a petabyte? For those of you familiar with disk storage measured in gigabytes (GB), a petabyte is 1,000,000 GBs. Who could possibly need such a large data system? Not surprisingly, the largest retailer in the world – **Wallmart**. With over <u>800 million transactions</u> <u>tied to over 30 million customers each day</u>, the data coding and analysis needed for such a system are clear. A reliable and flexible data system design is crucial to meet Walmart's substantial data requirements. Whether it is suppliers who wish to view product movement and sales in real time, or executives who are interested in business intelligence or scenario planning, the data design aspect of Walmart's data warehouse is the key to its success. Because virtually all of the transactions are processed in real time, as the demands for richer, more robust, and timely data analysis increase, Walmart appears to have made the investments needed to grow their data warehouse into the future. In the future, there are even plans to have data marts- smaller, subject-specific data systems that can handle the needs of a particular business area.

• Sources: Zikmund et al. (2013). Hayes Weier, Mary, "Hewlett-Packard Data Warehouse Lands in Wal-Mart's shopping cart", Intelligent Enterprise (August 4, 2007).

REFERENCES

- Brown, T.J., Suter, T. (2012). *MR*. South Western, Cenage Learning. USA.
- Hair, J.; Bush, R. & Ortinau, D. (2006). *Marketing research. Within a changing environment*. Revised International Edition (3rd Ed.). McGraw Hill, New York, USA.
- Malhotra, N.K. (1996) Marketing Research. An Applied Orientation. 2nd Edn. Prentice-Hall International. USA
- Rosendo-Rios, V., de Esteban, J., Antonovica, A. (2012). MR: Development of Theoretical Concepts for Market Research I and II. South Western, Cenage Learning. USA
- Zikmund, W. G., Babin, B.J., Carr, J.C., Griffin, M.(2013) Business Research Methods. 9th Edition. South Western, Cenage Learning. USA



Chapter 12

Survey Code and Tabulation

Business Research Methods Verónica Rosendo Ríos Enrique Pérez del Campo

Marketing Research