Processing and Cleaning the Data

In: Doing Surveys Online

By: Vera Toepoel

Pub. Date: 2017
Access Date: April 9, 2019
Publishing Company: SAGE Publications Ltd
City: 55 City Road
Print ISBN: 9781446249673
Online ISBN: 9781473967243
DOI: https://dx.doi.org/10.4135/9781473967243
Print pages: 175-191

Processing and Cleaning the Data

INTRODUCTORY CASE: Data cleaning for the CentERpanel

The CentERpanel is one of the first established online probability-based panels in the world. Many longitudinal surveys have been fielded in this panel. The following steps are taken to clean the data for these longitudinal projects:

- Merge files wave 1 to wave X (in case of longitudinal surveys),
- Investigate duplicate cases and keep the best one (furthest in the questionnaire),
- Remove test cases,
- Transform randomizations/permutations when applicable,
- Remove every case that looked at the introductory screen only and did not complete any question,
- Delete redundant variables (e.g. intros and headers),
- Create/improve/check variable names,
- Create/improve/check variable labels,
- Create/improve/check value labels,
- Recode/transform variables including value labels (e.g. scales running from 1–10 in the data, but presented as 0–10 in the survey),
- Define missing values,
- Calculate the duration of the survey (and/or other paradata),
- Look at open-ended questions and clean if necessary,
- Add sociodemographic variables,
- Encipher the dataset,
- Sort data on new key,
- Save the raw AND encipher dataset (remove all identifying information from the encipher dataset).

Note that in some instances, for example, data dissemination for longitudinal surveys, it is necessary to create new variables names for every questionnaire (e.g. health 1 in wave 1 and health 2 in wave 2.

11.1 Introduction

Processing and cleaning data is probably the most thankless job there is when it comes to conducting

surveys. Nevertheless, it is of crucial importance that you take time to carefully process the data and impose cleaning strategies onto the raw data you have acquired. One of the main advantages of Web surveys is that respondents' data is in most cases already stored on a server and no data have to be entered by research agencies. Nevertheless, it takes time to create a proper codebook and clean the data. Failing to delete one or two outliers can have a significant effect on descriptive and explanatory analyses. Therefore, data should be screened and treated when necessary. This chapter discusses the steps that are necessary to process and clean the dataset once the fieldwork has ended, and before analyses can be performed.

11.2 Processing

In order to keep track of the questionnaires you need to use identification variables in your dataset. You can use simple numbers from 1 to *N*. For panel studies, it is important that you assign the same number to the same person over time. In addition, for household surveys, it is important that you keep track of both household numbers and individual numbers. You need to be able to aggregate from the individual to the household level and see how individuals in the household sample are related. Identifying respondents is also important during the fieldwork to monitor response rates, send reminders to non-respondents, and for communicating with respondents if they run into problems when completing the survey.

For mixed-mode surveys (e.g. a combination of Web and paper-and-pencil) and in the case of email surveys you need to enter the data into a database. You can add the cases (respondents' questionnaire forms) by hand to the Web survey using the survey URL. A problem that you might run into is that people in a paper-and-pencil survey could make mistakes that people in your Web survey could not. For example, if you programmed a radio button, Web survey respondents can only fill out one possible answer. Even if you requested in the instruction text that your paper-and-pencil respondents only fill out one answer, they could have ignored the instruction and filled out more than one answer. What do you do then? You want your paper-and-pencil and Web surveys to produce similar results. I often see data entry people entering only the first option, but that implies a primacy effect. Therefore, you should find some strategy to choose one of the filled out options randomly. You could, for example, throw a dice and choose the first response option when it's even and the second one when it's odd. Note that this holds only for two chosen options, not three or more!

11.3 Data cleaning

Before you start analysing your data, you should clean your data. Data cleaning is the process of identifying and often correcting 'dirty' data. This means that you should remove all unnecessary variables, entry errors, illogical answers (e.g. out of a possible range), respondent's identifying information, skip errors (good survey software already detects and deletes skip logic features), etc. In addition, you should create proper variable names, variable labels and value labels. Although many types of analyses report data errors, it is necessary to perform systematic data cleaning in order to delete all erroneous entries. This prevents you from drawing the wrong conclusions based on, for example, the relative weight of extreme high values of outliers.

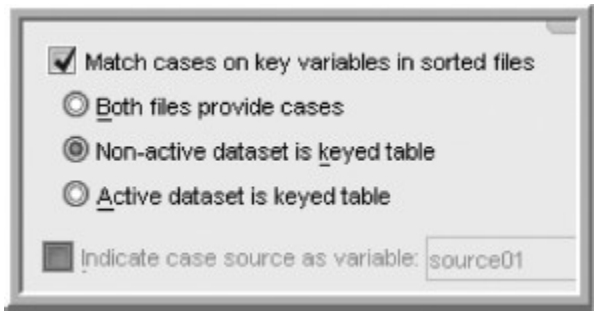
11.3.1 Screening

You will start with screening the data. This means that you are going to screen the raw data case-wise and variable-wise for redundancies, correct labelling, identifying information and inconsistencies. Often one of the first steps is merging files.

11.3.1.1 Merging files

In the case of longitudinal surveys you may want to merge files from several waves. In addition, you may want to add sociodemographic variables stored in another dataset (e.g. from a profile survey). Merging files is relatively simple. Your software will provide you with opportunities to identify a keyed dataset, which is your starting point (e.g. your particular survey). You can then add cases (e.g. in case you have split the fieldwork and want to combine two datasets), add variables (e.g. in case you want to add sociodemographic variables for the cases/respondents in your dataset only) or you can add both cases and variables at the same time.

Figure 11.1 Merging files: Add variables and/or cases (SPSS)



11.3.1.2 Duplicate, test and incomplete cases

Once you have merged all the datasets you wanted, and you have all information in a single dataset, you can start with investigating the cases in it. You need to investigate whether anyone started the questionnaire more than once and is thus a duplicate in your dataset. If you have duplicates, you need to compare the cases and keep the best one (the one that was furthest along in the questionnaire). In addition, you need to remove test cases and every case that did not complete any question.

Often people that look at the introduction screen but then drop out of the survey are stored in the database. Those people should be removed and treated as non-respondents. It is more difficult to know how to treat people who completed one or more questions, but then broke off. You can delete all **incomplete cases**, that is, all cases who did not finish the survey. This choice can be made for surveys with a relatively high number of cases and surveys where dropout rates are low. But you probably want to keep as many respondents as you can. You can also choose to keep respondents that completed most (e.g. more than half) of the survey questions or provided information that was crucial for your survey (e.g. key variables). You should report in your codebook (see Section 11.4) on how you treated incomplete cases, so that other people can see what you did.

Survey tip: It is a good idea to give test cases totally different identification numbers from normal cases, so you can find them easily in the dataset.

11.3.1.3 Randomizations or permutations

You then go further and clean the variables. When randomizations were used, you may need to put the variables or values in the original order. Some survey software packages do this for you, in other cases you need to write syntax for this yourself. Check whether your de-permutations and de-randomizations worked correctly! In addition, you need to decide whether to keep both **random** and de-randomized variables in your dataset. An argument for keeping both is completeness; you can also investigate order effects. On the other hand, it may confuse other users of the data.

11.3.1.4 Delete redundant variables

Redundant variables can be deleted. Often, introduction texts and headers are stored in the dataset, for example, with a value of 1, but they provide no information and can hence be deleted from the file.

11.3.1.5 Variable names, variable labels and value labels

Then variable names, variable labels and value labels should be checked and improved if necessary. It is important that every variable has proper variable and value labels, so that the right text is provided in the analyses. This also saves time if you want to produce tables or figures in your report: the meaning of the variables is already in the output.

11.3.1.6 Recoding

For many variables it is necessary to transform them from the original raw data file. For example, sometimes you use a scale with answer labels from 0–10, but the survey software always gives the first answer option a value of 1, with the result that your value labels run from 1–11. You then need to recode 1–11 into 0–10. In the case of binary variables, for example, gender, some people also prefer to recode values 1–2 to 0–1, making it a dummy variable.

Depending on the way you programmed a ‘don’t know’ option, it can in some cases be treated as a substantive answer (7 = don’t know) and sometimes as a missing value (don’t know = missing). You will probably want to define all ‘don’t knows’ in the same manner, therefore you might want to transfer certain values to a standard (‘don’t know’ = 99; missing). In addition, you may need to recode reverse worded answers.

11.3.1.7 Open-ended answers

You need to check open-ended answers and check that they do not contain any text that is unsuitable for public display, for example, personal identifying information, information in the survey directed at the panel

manager, offensive language, etc.

11.3.1.8 Paradata

The advantage of Web surveys is that they provide you with an enormous amount of paradata. Many researchers do not use this information, but it can provide you with valuable information. For example, the duration of the questionnaire (or question) gives insight into how seriously the respondent answered the survey. A duration of 1 minute for a 30-question survey tells you that your respondent was speeding and did not provide you with meaningful answers. You could decide to delete the case. I would suggest always including a variable representing the duration of the survey in the dataset. In addition, some researchers find it useful to look at keystroke files. Paradata such as the time of day the survey was started, the browser or device the respondent used, etc. can provide you with extra information about the response process. Some people might have had problems because they used a particular browser or completed the survey on their mobile phone. Drop outs can then be explained and anticipation of potential error is possible in future surveys.

11.3.1.9 Encipher the dataset

The cleaned dataset is then ready to be used. But before you go off and distribute the dataset, you need to encipher it so that no one can identify your respondents. You create a new respondent number and save the file under a different name. You need to store the original file with the original and enciphered respondent number, and the new file with only the enciphered number. You can distribute the enciphered dataset, but make sure you keep the key in the original file. Sort the data with the new key and off you go!

Survey tip: When cleaning data it is wise to keep at least three data files for the survey:

- The original raw dataset (e.g. survey.sav),
- The cleaned dataset (e.g. survey_1.sav),
- The enciphered (keyed) dataset (e.g. survey_1p.sav).

Don’t forget to sort the data with the new key!

11.3.2 Diagnostics: Strange patterns

Once you have screened your data, you need to diagnose strange patterns. This means that you are going to look for patterns of unintended missing data, inconsistencies, outliers (extreme values) and all other strange patterns in the response distributions. You can do this by running frequencies, graphs and descriptive statistics.

The **frequency distribution** of a variable shows the number of respondents who have selected each response option, and presents percentages and cumulative percentages for each category. Frequencies can

be presented in tables and in graphs. Graphs give you the option to visually separate the logical from the illogical answers. **Descriptive statistics** present **summary statistics** of the variables such as minimum values, maximum values, means and standard deviations. Sometimes it can also be useful to look at the **median** score (midpoint of the distribution of answers) or **mode** (most frequently endorsed answer option).

Figure 11.2 Descriptive statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Number of Brothers and Sisters	1505	0	26	3.93	3.047
Number of Children	1517	0	9	1.94	1.834
Age of Respondent	1514	18	89	45.63	17.808
Highest Year of School Completed	1517	0	99	13.28	6.553
Highest Year School Completed, Father	1069	0	20	10.88	4.129
Highest Year School Completed, Mother	1233	0	20	10.79	3.463
Highest Year School Completed, Spouse	790	0	20	12.89	3.059
R's Occupational Prestige Score (1980)	1418	17	86	42.93	13.067
Occupational Category	1418	1.00	6.00	2.9210	1.77634
Valid N (listwise)	520				

If you want to look at more than one variable at the same time, which can be useful in the case of a predicted relationship between two variables you can also use **crosstabs** to look at the frequency distributions of two variables at the same time. Note that you can find totally 'normal' values (no outliers) in separate variables, but the combination of a certain value on 'x' and a certain value on 'y' can still be a strange/erroneous answer.

Figure 11.3 Frequencies (note the strange number 9)

Respondent's Sex

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Male	636	41.9	41.9	41.9
	Female	881	58.1	58.1	100.0
	Total	1517	100.0	100.0	

Most Important Problems in Last 12 Months

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Health	91	6.0	27.1	27.1
	Finances	129	8.5	38.4	65.5
	Lack of Basic Services	4	.3	1.2	66.7
	Family	48	3.2	14.3	81.0
	Personal	19	1.3	5.7	86.6
	Miscellaneous	40	2.6	11.9	98.5
	9	5	.3	1.5	100.0
	Total	336	22.1	100.0	
Missing	System	1181	77.9		
Total		1517	100.0		

Figure 11.4 Crosstabs (note that people should not report the same answer on both variables)

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
bg09b331 My bum is too big. * bg09b332 My bum is too small.	3857	88.2%	517	11.8%	4374	100.0%

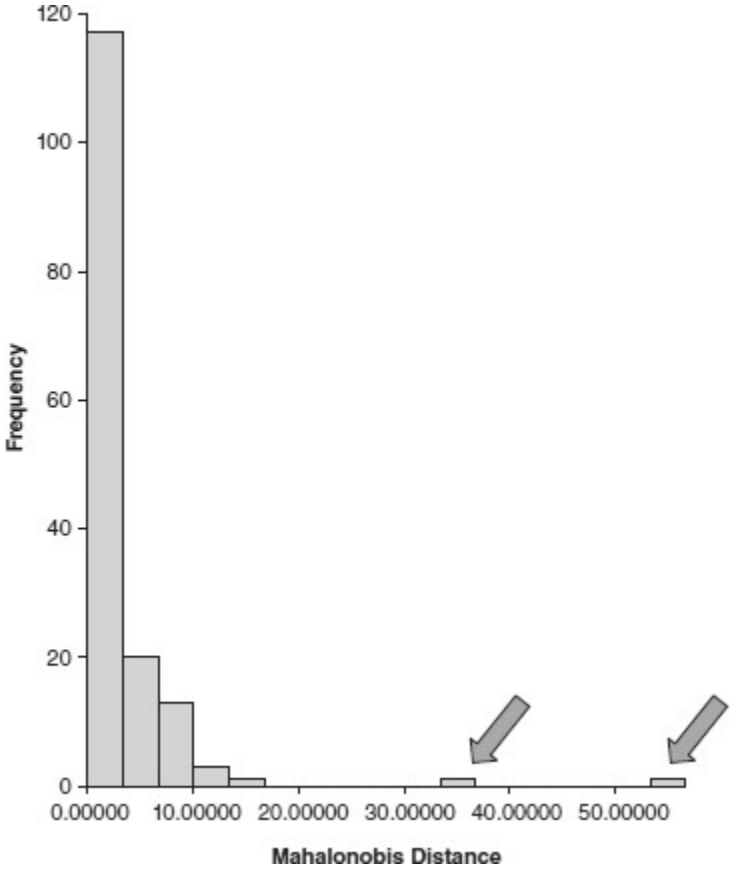
bg09b331 My bum is too big. * bg09b332 My bum is too small. Crosstabulation

Count

		bg09b332 My bum is too small.					
		1 disagree entirely	2 disagree	3 partly agree, partly disagree	4 agree	5 agree entirely	Total
bg09b331 My bum is too big.	1 disagree entirely	553	120	55	58	12	798
	2 disagree	110	1509	137	127	10	1893
	3 partly agree, partly disagree	77	246	195	3	0	521
	4 agree	124	353	11	5	1	494
	5 agree entirely	109	40	1	0	1	151
Total		973	2268	399	193	24	3857

A scatter plot can be used for continuous variables to check for bivariate outliers. If you want to check for multivariate (more than two) outliers you can use ‘Mahalonobis distance’. By doing a regression you can ask to save the Mahalonobis distance in a new variable. You can then examine high values over all variables.

Figure 11.5 Detecting outliers with Mahalonobis distance



After you have screened the data for possible errors, you want to learn the cause of the error. Was there a programming error in the questionnaire, or was the answer mistyped? In some cases it is obvious if an answer is impossible, for example, if people report doing an activity (e.g. working) for more than 24 hours a day. Or if people report their age to be an impossible value, for example, 233. The respondent could have intended to

report 23, but typed an additional ‘3’. Note that you can prevent errors like this by programming hard or soft checks.

You could search for additional information, for example, in other questionnaires or in the profile survey, and see if you can be certain that a certain error was made. You can then go to the ‘treatment’ of correcting the answer. Of course, you need to be certain what the correct answer would be. If you are not 100 per cent certain, I would advise you not to correct the data and simple define the value as missing. In some cases, unlikely answers can still be correct answers. For example, I once had a respondent who reported to earn more than a million euros a month. Although this answer is highly unlikely, it *could* be the true answer. Therefore, I did not change the answer (merely wondered which professional soccer player was in my panel).

Survey tip: Much information is provided simply by:

- Descriptives for continuous variables,
- Frequencies and cross tabs for categorical variables.

Survey tip: To check for multivariate outliers: conduct a regression and save the Mahalonobis distance. Use all variables as predictors and case number as the dependent variable. Note that the regression output is complete nonsense, but you can evaluate all variables for outliers at the same time.

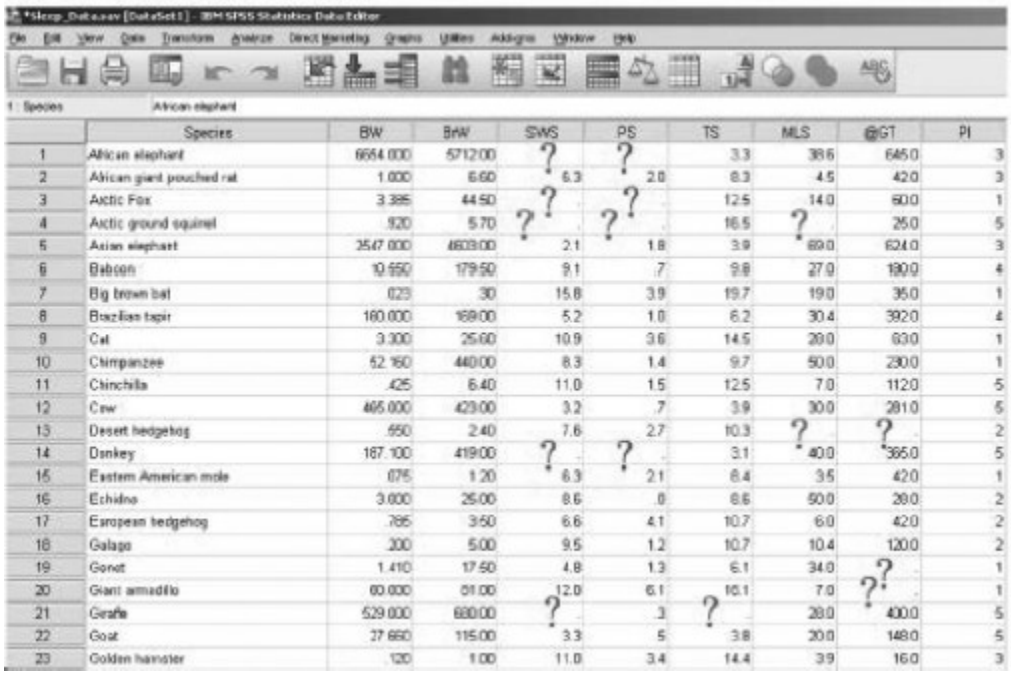
11.3.3 Treatment

Once you have diagnosed all problematic values, you will need to decide what to do with them. There are statisticians in favour of changing the raw data to potentially ‘true’ answers, but there are also people in favour of keeping the data file as original as possible. Changing data can dramatically change answer distributions and hence conclusions. You therefore need to be certain that you are changing it in the correct way. If you belong to the ‘raw data believers’, you can merely delete certain erroneous answers and report that a particular number of values were excluded from the analysis during the data cleaning phase. If you belong to the first group, you can, for example, impute missing data or change outliers to less extreme scores.

11.3.3.1 Imputations

For certain variables, such as income, you face high rates of missing values (e.g. respondents answering ‘don’t know’ or ‘won’t tell’ or simple failure to provide an answer). Especially for economists, it is important that they have values for income for every respondent. In addition, if you want to use income as a weighting variable, you will need substantive values for every respondent.

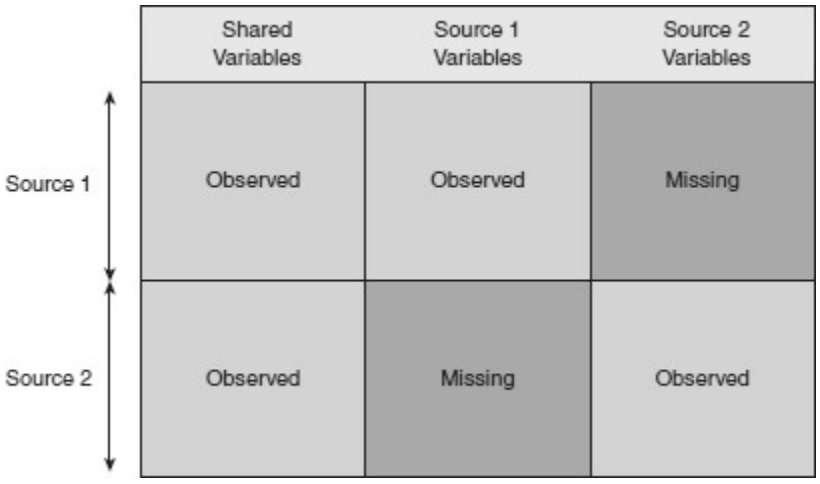
Figure 11.6 Item non-response (missing values)



	Species	BW	BrW	SWS	PS	TS	MLS	GT	PI
1	African elephant	6654.000	5712.00	?	?	3.3	38.6	645.0	3
2	African giant pouched rat	1.000	6.60	6.3	2.0	8.3	4.5	42.0	3
3	Arctic Fox	3.385	44.50	?	?	12.5	14.0	60.0	1
4	Arctic ground squirrel	.920	5.70	?	?	16.5	?	25.0	5
5	Asian elephant	3547.000	4603.00	2.1	1.8	3.9	69.0	624.0	3
6	Baboon	10.550	179.50	9.1	.7	9.8	27.0	180.0	4
7	Big brown bat	.023	.30	15.8	3.9	19.7	19.0	35.0	1
8	Brazilian tapir	160.000	169.00	5.2	1.0	6.2	30.4	392.0	4
9	Cat	3.300	25.60	10.9	3.6	14.5	28.0	63.0	1
10	Chimpanzee	52.160	440.00	8.3	1.4	9.7	50.0	230.0	1
11	Chinchilla	.425	6.40	11.0	1.5	12.5	7.0	112.0	5
12	Cow	465.000	423.00	3.2	.7	3.9	30.0	281.0	5
13	Desert hedgehog	.550	2.40	7.6	2.7	10.3	?	?	2
14	Donkey	187.100	419.00	?	?	3.1	40.0	365.0	5
15	Eastern American mole	.075	1.20	8.3	2.1	8.4	3.5	42.0	1
16	Echidna	3.000	25.00	8.6	.0	8.6	50.0	28.0	2
17	European hedgehog	.785	3.50	6.6	4.1	10.7	6.0	42.0	2
18	Galago	.200	5.00	9.5	1.2	10.7	10.4	120.0	2
19	Genet	1.410	17.50	4.8	1.3	6.1	34.0	?	1
20	Giant armadillo	60.000	81.00	12.0	6.1	18.1	7.0	?	1
21	Giraffe	529.000	680.00	?	.3	?	28.0	400.0	5
22	Goat	27.660	115.00	3.3	.5	3.8	20.0	148.0	5
23	Golden hamster	.120	1.00	11.0	3.4	14.4	3.9	16.0	3

In panels, you may want to merge data from two questionnaires. For example, if you have a survey on social integration and a survey on health, you may want to combine them and see if social integration and health are related. You will face non-response in both waves. So you have panel members for whom you have information from both surveys, panel members who completed only the survey on social integration but not health and panel members who completed the survey on health but not on social integration, as is visually demonstrated in Figure 11.7. You could then try to impute missing values as well (e.g. based on answers in a previous wave).

Figure 11.7 Missing values due to non-response in two surveys



One of the easiest ways to deal with missing items is to delete complete cases. This strategy is unbiased if items are missing completely at random (MCAR), but you miss information about means, variances (because

of the sample) and you get less power (see Chapter 4).

If you do not want to delete complete cases, imputation can be a solution. Imputation strategies can vary from very simple to very complex. One of the easiest strategies is to impute the variable mean to the cases with missing values.

Figure 11.8 Imputation with mean score

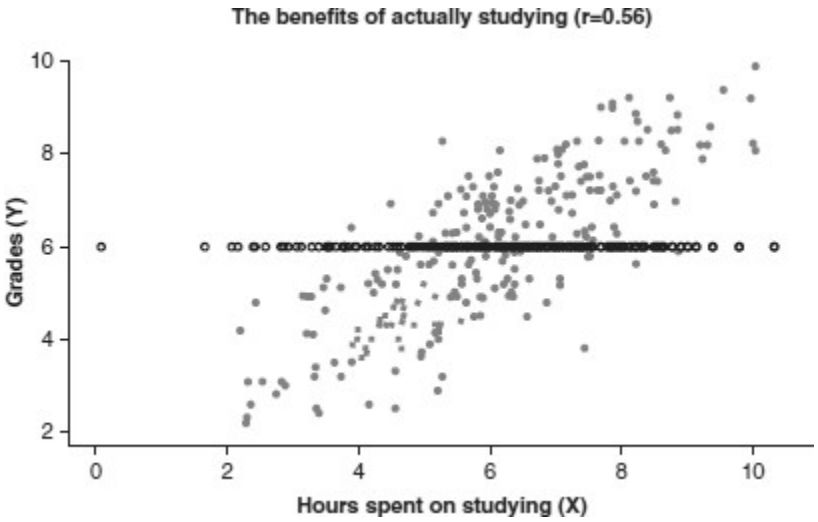
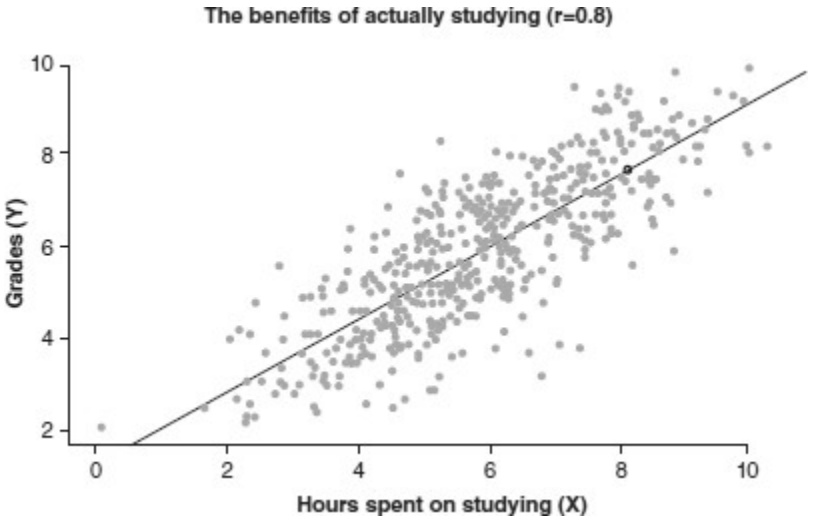


Figure 11.9 Imputation with regression estimates



Another way to impute is by using a regression. Regression imputation gives a better value than mean imputation because the imputed value is the most probable value and has minimum error. However, the true score (dependent variable, in the example in Figure 11.8 the variable 'grade') is uncertain and single predictions do not portray this uncertainty. Therefore, a commonly implemented strategy is to use multiple imputations. Instead of imputing one variable, the same missing value is imputed several times (e.g. five times). In most survey software, for example, SPSS, there is a function for multiple imputations.

Missingness mechanisms

Before you start imputing you need to realize that there are different types of missingness:

- ***Missing completely at random (MCAR):*** the probability of being missing is the same for all cases. Missingness does not depend on any data. It does not follow a pattern.
- ***Missing at random (MAR):*** the probability of being missing is not the same for all cases, but may depend on any observed information. Missingness depends on the data, but does not depend on the *missing* data. It follows a known pattern.
- ***Missing not at random (MNAR):*** the probability of being missing is not the same for all cases and may depend on the missing information. Missingness follows a pattern, but the pattern is not observed.

If the data are MCAR or MAR, the missing data mechanism can be ignored and imputation strategies can be used. If the data are MNAR, you cannot ignore the missing data mechanism, and imputation can be a problem. If you think your data are MNAR, you will need to look for additional information about imputations and contact a specialist.

11.3.3.2 Outliers

Outliers can have a significant effect on the outcome of your analyses. You can minimize their influence by:

- Changing the raw score to a less extreme value, for example,
    - Mean,
    - Mode,
    - Median,
    - Trimmed,
- Delete the extreme case from the analyses.

Note that you can question representativeness if you remove one or more respondents. You should therefore always be careful when treating outliers. Analyse your data *with and without* outliers and see if the results differ. When they do, report the analyses without outliers and carefully report why and how the cases were removed. It is important that you report precisely what you did so others can replicate or judge the decisions taken.

11.4 Codebook

A **codebook** is a document that accompanies your dataset. The codebook should contain all the relevant

information that is necessary to understand the dataset and its quality. Other researchers who might want to use the data (in a later stage) should be able to use it without consulting the original investigator. You should therefore mention how you reached your respondents, response rates, sample composition, questions, answer options, instructions, routing, variable names, value labels, etc.

In addition, I am an advocate of adding screenshots of the survey to demonstrate how questions were put on the screen visually, since it is well known that the visual layout of survey questions influences respondent's answers. For example, presenting a simple five-point Likert scale in a horizontal, vertical or matrix format can change response distributions significantly. If researchers want to compare different datasets, it is important that they know how questions were presented visually in order to be able to estimate where differences between datasets can come from.

You should start your codebook with a description of the survey objectives. In addition, it should be clear how respondents were recruited. How did you reach them and who responded? In the case of a panel survey: how were the panel members recruited, what were the response rates at all response stages (recruitment interview, registration rate, cooperation rate, see Chapter 10 for describing response rates)? In case of a desire to generalize conclusions to a greater population, you should provide statistics for the population (e.g. gender, education, age, all other relevant variables) and for your sample and indicate where they differ and how this could be a problem. You can also provide a description of how weights were built in order to make the sample (more) representative of the population.

Figure 11.10 Instruction text in a codebook (www.lissdata.nl)

3 Codebook

This codebook contains the questionnaire as administered to the LISS panel.

- Variable names: The variable names are printed in **bold** and correspond to the names in the dataset.
- Routing: The questionnaire routing is printed in *italics* for each variable concerned.
- *open:* answer box (no limit to the length of the answer).
- *string:* answer box allowing a certain number of characters (standard is 255).
- *empty:* questions could be left unanswered
- Numerical variables: If ranges were used, these are printed in *italics* in the codebook if the respondent could not see them on the screen. If the respondent was able to see the ranges, the variables are printed in normal letter. *Integer*:If a question is not subject to any limit (integer), no range is indicated.
- 'Fills' (variable text) are given between straight brackets [].
- Variables in between curly brackets {} are not part of the dataset, but the corresponding questions or text were part of the questionnaire,

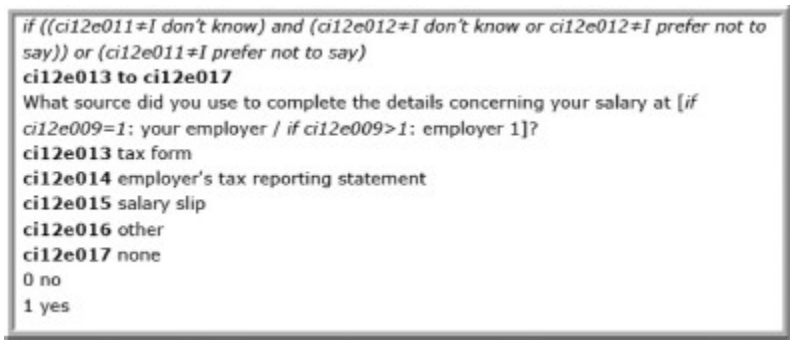
nomem_encr[1]

Number of household member encrypted

You can then describe how people should read your codebook; for example, a description of how variable

names (**bold**), instruction text (*italics*) and missings (value = 99) are coded. The first variable in the dataset is probably the identification number. Sometimes you have different identification numbers (e.g. household level and individual level). If you used randomization in the survey, it is good to report the variable (e.g. random = 1 if answer options are in incremental order, random = 2 if answer options are in decremental order) at the beginning of the dataset as well. Instruction and introduction screens are normally not variables and are therefore not present in the dataset, but they should be reported in the codebook as they were in the survey. The layout of the codebook should be clear so that you can see all variables or answer options at a glance. All answer options in the dataset should be described in the codebook (also non-substantive answer options such as 'don't know'). The routing for every question or answer option should be described in the codebook. This can be a problem if you have very difficult routing in the survey. Nevertheless, if it was possible to program the complex routing, it should also be possible to describe it in the codebook.

Figure 11.11 Example of complex routing in a codebook (www.lissdata.nl)


if ((ci12e011≠I don't know) and (ci12e012≠I don't know or ci12e012≠I prefer not to say)) or (ci12e011≠I prefer not to say)
ci12e013 to ci12e017
What source did you use to complete the details concerning your salary at [*if ci12e009=1*: your employer / *if ci12e009>1*: employer 1]?
ci12e013 tax form
ci12e014 employer's tax reporting statement
ci12e015 salary slip
ci12e016 other
ci12e017 none
0 no
1 yes

You can end the codebook with descriptive statistics of all variables in the dataset (min., max., mean, modus, number of valid observations). If you have a very large dataset you might want to produce a separate file for the descriptives, since it can make the codebook difficult to work with if you have dozens of pages with descriptive statistics. You can also provide the original questionnaire at the end of the codebook, preferably with screen shots of the online survey.

Elements in the codebook:

- Respondent/panel recruitment,
    - In case of generalizations: sample composition and population composition,
    - (Weighting description),
- ID numbers,
- (Randomization variables),
- Instruction text,
- Variables,

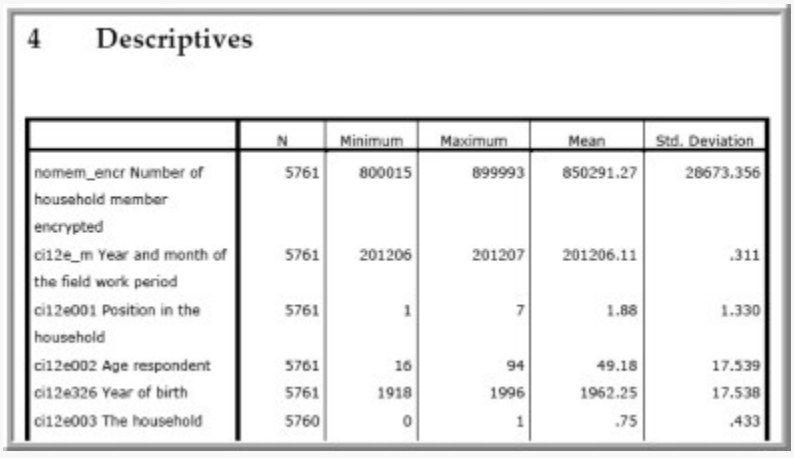
- Question text,
- Value and variable labels,
- Routing,
- (Descriptive statistics),
- (Screen shots).

Survey example

Content of the codebook:

- Research aim,
- Client,
- Response,
- Recruitment,
- Representativeness,
- Codebook questionnaire,
- Codebook profile survey/sociodemographic variables,
- Variables.

Figure 11.12 Example of descriptives at the end of the codebook



4 Descriptives

	N	Minimum	Maximum	Mean	Std. Deviation
nomem_encr Number of household member encrypted	5761	800015	899993	850291.27	28673.356
ci12e_m Year and month of the field work period	5761	201206	201207	201206.11	.311
ci12e001 Position in the household	5761	1	7	1.88	1.330
ci12e002 Age respondent	5761	16	94	49.18	17.539
ci12e326 Year of birth	5761	1918	1996	1962.25	17.538
ci12e003 The household	5760	0	1	.75	.433

Summary

Data cleaning is an important and undervalued part of the survey process. It is important that all survey information is correctly processed. In addition, raw data files should be cleaned by deleting duplicate and

test cases, variables and values should be correctly labelled, missing values need to be defined, open-ended questions need to be screened, etc. As well as this, paradata can be added to the questionnaire, for example, the duration of the survey, the duration of the question, keystroke files, browsers, devices used (e.g. mobile versus desktop), etc. This provides important information about the response process and possible errors can be detected. The dataset should be keyed so that no personal identifying information is stored in the dataset. It is important however to keep the original raw data file as well, since this is the source of the survey. You can use imputation strategies for missing data, and you can also treat outliers as well, for example, by replacing them with less extreme values. In addition to data cleaning, a proper codebook should be written that accompanies the dataset.

Key terms

Codebook
Crosstabs
Descriptive statistics
Frequency distribution
Incomplete cases
Median
Missing at random
Missing completely at random
Missing not at random
Mode
Random
Summary statistics

Exercises

1. Discuss the steps you need to take when cleaning data.
2. Why is it important to identify outliers? Discuss ways to identify them and treat them.
3. What is a bivariate outlier?
4. Discuss different ways to treat incomplete cases (drop outs).
5. Name three different forms of paradata that can be useful to add to your dataset and discuss why.

Suggested readings about imputations

De Waal, T., Pannekoek, J. and Scholtus, S. (2011) *Handbook of Statistical Data Editing and Imputation*. Hoboken NJ: John Wiley and Sons, Inc. (For more information about imputation)

Little, R.J.A. and Rubin, D.B. (2002) *Statistical Analysis with Missing Data*. Hoboken, NJ: John Wiley and Sons, Inc. (For more information about missing data)

http://dx.doi.org/10.4135/9781473967243.n11