Chapter 7. Data entry and coding

Trung Tran, Cong-Thang Ngo, Quynh-Trang B. Nguyen, Quy Khuc

©Vietkap team, Hanoi, Vietnam

Date: January 31, 2021

Preprint DOI: 10.31219/osf.io/kxp6b

Abstract

Inputting and coding data are the last step in the primary data investigation ecosystem. This task is often time-consuming given the demand for accuracy, completeness and easy-to-use of the information in the data sheet. In order to help readers, especially young scientists with little or no experience in primary data investigation, to access and practice the aforementioned step effectively, in this chapter, we will present systematically the process involved in data entry and coding. Specifically, this includes designing an input table (codebook and data sheet) and checking the data.

Keywords: primary data, input table, codebook, variables

Introduction

After an arduous and costly procedure from questionnaire design to pilot and formal survey [1–3], researchers now have a set of completed questionnaires. However, the obtained information is discrete, "dead" since a single piece can add little value to the research. In general, a typical survey receives a large number of responses, ranging from hundreds to thousands [4,5]. They are then synthesized to yield primary data that will be processed using calculations, statistical models, thereby producing scientific information and data. To achieve that dataset, collected answers must undergo a process of input and coding, adhering to certain

rules and standards so that users can understand the "literal", encoded information, and vice versa. Here are 3 steps in inputting and coding data:

- Step 1. Design an input table (codebook and data sheet)
- Step 2. Input data (data entry table)
- Step 3. Check the data (check for accuracy, completeness and comprehensibility).

Designing an input table

Designing an input table is the first step, involving (1) building the codebook and (2) building the data sheet. Although the focus of this task is on the codebook and data sheet, rechecking the questionnaire is indispensable. In other words, the questionnaire is the overarching point of all activities. It is through the questionnaire that researchers can classify information and build the codebook.

Building the codebook

Data encryption/digitization is understood as the information coding step. Each question will be encoded as a "variable" whose information corresponds with that of each question [6]. These variables will be included in the model (e.g., econometric model) for analysis. As a result, the scientist will clearly identify what type of variable each piece of information belongs to (continuous variable, ordinal variable, binary variable, or nominal variable). The codebook is responsible for explaining all the information, codes of questions, answers in the questionnaire and the input table [7,8]. The goal is to enable anyone, upon accessing the data sheet, to understand what it describes, the meaning and effects of the variables and so forth. A detailed description of the information field in a codebook is provided in Table 1.

Table 1. Information field codebook

#	Dimensions	Example 1	Example 2	Example 3
1	Question ID	A1	A2	A3
2	Question content	Which district are you living in Hanoi?	To what extent do you feel about the air quality in your area?	Does the air quality vary across districts in Hanoi?
3	Variable	Location	AirDistrict	AirVsOtherDistrict
4	Variable Characteristics	Binary	Numerical on five points 1-5 scale	Binary
5	Code	(1, 0)	(1, 2, 3, 4, 5)	(1, 0)
6	Meaning	Living district of the interviewees	(Very good, good, normal, bad, very bad)	(Yes, No)

Sources: Survey on environmental pollution in Hanoi [8]

Table 01 presents six information fields, including: question ID, content, variable, variable characteristics, code, and meaning. First, the ID. Each question has a unique ID to be distinguished from one another in the questionnaire. And each of them has content. For example, in question A1, the content is “Which district are you living in Hanoi?”. Next is the variable, representing symbols for each question. As mentioned above, each question corresponds to a variable in the econometric model. It is noted that variables should be succinct, enough for readers to understand the question. Usually it is advisable not to include more than three acronyms of three keywords for each variable. An example is the variable “Location” in question A1. The characteristics of a variable is its type that is often nominal, or ordinal, or continuous, or binary. In example 1, "Location" is a binary variable. And the code value of this variable is either 1 or 0. The last information field explains the meaning of the code value. Back to example 1, the meaning indicates where the interviewee lives, with value 1 in inner-city areas and value 0 in the suburbs. Likewise, in question A2, the content is to find out people’s opinions on the air quality in their neighborhood. The variable is “AirDistrict”, which is

categorized as ordinal, measuring from 1 to 5 corresponding to the air quality from very good to very bad.

Data sheet

A data sheet is designed to store all the useful information of collected answer sheets from the official survey. Building the input table is not as difficult and complicated as the questionnaire since it is based on the already existing content of the questionnaire. However, there is a set of requirements and standards that needs to be met: complete, clear, scientific. First, it should contain all the useful information of collected answers. Clarity means it is easy to code and does not cause confusion in the input process. The scientific characteristic is shown in the logical arrangement of rows, columns, and groups of information, ensuring ease of reading, synthesis and calculation. An input table is a matrix of $m \times n$ (m is the number of questions, n the number of sheets, or observations). The two top rows present the questions' number, the interviewees' ID and the variable name (Figure 1). Information from collected answers will be filled in accordingly and in the correct box.

	A	B	C	D	E	F	G	H
1			A1	A2	A3	A4	A5	A6
2	No.	Interview ID	Location	AirDistrict	AirVsOtherDistrict	AirCity	AirCityVsClo sedCity	AirCityVsS outhCity
3								
4								
5								
6								
7								
8								
9								
10								
11								
12								
13								

Figure 1. An unfilled-in data sheet

Data entry

After designing the input table and the codebook, the researcher instructs the input participants to understand and perform the task correctly (Figure 2). Normally, the survey participants will input the information in the questionnaire that they have investigated, or exchange sheets with each other so as to avoid subjective errors. This is quite convenient because after a period of hands-on experience, they will grasp the questions as well as the information in the questionnaire, hence assisting data entry efficiently.

	A	B	C	D	E	F	G	H
1			A1	A2	A3	A4	A5	A6
2	No.	Interview ID	Location	AirDistrict	AirVsOtherDistrict	AirCity	AirCityVsClosedCity	AirCityVsSouthCity
3	1	TH01	Tay Ho	5	1	4	3	3
4	2	TH02	Tay Ho	2	1	4	4	4
5	3	TH03	Tay Ho	3	1	4	4	4
6	4	TH04	Tay Ho	2	1	4	3	4
7	5	TH05	Tay Ho	3	1	4	3	3
8	6	TH06	Tay Ho	2	1	4	4	4
9	7	TH07	Tay Ho	4	1	3	3	1
10	8	TH08	Tay Ho	2	1	3	4	4
11	9	TH09	Tay Ho	3	1	4	4	4
12	10	TH10	Tay Ho	3	1	4	3	3

Figure 2. A filled-in data sheet [8]

Data check

Filled-in data requires checking to avoid errors or missing information. In fact, anyone can make mistakes when entering data as there exists minor errors that are hard to detect. Therefore, it takes time to meticulously check each and every information field for each and every observation. Although the scientist has “cleaned up” the data, in this step, should there be any obvious errors, they need to be checked and eliminated. Specifically, the data is checked according to each variable. Easily detectable errors include missing and wrong information, which might create a value out of the normal range or that is too large (like a value of 6

in the 1-to-5 value range, an error easy to spot); typo, not according to the codebook (like “w3” instead of “3” due to the close space between “w” and “3” in the keyboard). Any abnormal points must be checked and compared with the answer sheet, all the entered data in accordance with the value presented in the sheet as well. It is noteworthy that the data might be unusual or unreasonable, yet it will be processed in the cleaning and analysis step.

Some notes on data entry and checking

- First, the most important criterion is accurate, complete and easy to understand.
- Second, this step should be done as soon as possible, preferably within the "first week" of the official survey. The longer it is left undone, the poorer the quality of the entry.
- Third, the detection of errors (missing, wrong) should be done coincidentally with the data entry process. There is no need to wait until the end of the entry process to check the data.
- Fourth, a person simply cannot check all errors perfectly. Hence, in order to minimize errors in the data entry process, the leader should divide the team into two data entry groups. Then the two groups will work together, compare the results, find out the errors and agree on the filling methods.

In summary, data entry and coding is the last and very important step in the primary data investigation “ecosystem”. This step helps researchers to obtain a set of interconnected dataset, ready for later statistical analysis. Collected information is encoded, according to certain rules and standards, from “word” to “keywords” corresponding to each variable type. The first task is building a clear, detailed and scientific input table. The variables are encoded based on the codebook to serve statistical analysis and process. After data entry, there are two common problems: wrong and missing information. Thus, under specific circumstances, depending on the characteristics of the data, researchers need to have effective and scientific process methods.

Review questions

1. List and present steps in the data entry process.
2. Present the content of the data checking step.
3. Present some notes on data designing, data entry and data checking.

References

- [1] Kenneth D. Bailey, *Methods of Social Research*, fourth ed., The Free Press, New York, USA., 1994.
- [2] T.D. Tran, Q. Van Khuc, Primary data, OSF Prepr. (2021) 1–6. <https://doi.org/10.31219/osf.io/f25v7>.
- [3] P. Tran, T. Nguyen, T. Tran, Q. Khuc, Pilot survey, OSF Prepr. (2021) 29–33. <https://doi.org/10.31219/osf.io/dwhja>.
- [4] T. Tran, L. Pham, Q. Khuc, Sampling, OSF Prepr. (2021) 21–28. <https://doi.org/10.31219/osf.io/w9qks>.
- [5] R.B. Uma Sekaran, *Research Methods For Business: A Skill Building Approach*, 7th Edition, 2016. https://doi.org/10.1007/978-94-007-0753-5_102084.
- [6] T.D. Tran, P. Pham, Q. Van Khuc, Questionnaire design, OSF Prepr. (2021) 7–13. <https://doi.org/10.31219/osf.io/q3um6>.
- [7] P.A. Champ, K.J. Boyle, T.C. Brown, *A Primer on Nonmarket Valuation*, 2017. <http://link.springer.com/10.1007/978-94-007-7104-8>.
- [8] Q. Van Khuc, T.V. Phu, P. Luu, Dataset on the Hanoian suburbanites' perception and mitigation strategies towards air pollution, *Data Br.* 33 (2020) 106414. <https://doi.org/10.1016/j.dib.2020.106414>.