Urban Resilience and Sustainability

Urban Resilience Sustainability, 1(2): 91–106.
DOI: 10.3934/urs.2023007
Received date: 28 March 2023
Revised date: 09 June 2023
Accepted date: 19 June 2023
Published date: 29 June 2023

http://www.aimspress.com/journal/urs

Review

A researcher companion of data collection and validation methods

Abbas Ziafati Bafarasat[1,*], Eduardo Oliveira[2] and Anna Growe[3]

[1] Faculty of Technology, Design and Environment, Oxford Brookes University, Oxford, United Kingdom
[2] Entrepreneurship and Innovation and Business Sustainability Management, Thomas More University of Applied Sciences, Mechelen, Belgium
[3] Institute of Urban Development, Kassel University, Kassel, Germany

*** Correspondence:** Email: aziafati-bafarasat@brookes.ac.uk.

Abstract: This paper aims to fill a gap between textbooks and papers on data collection methods. Many undergraduate study programs include modules on data collection methods. They provide students with systematic insight into data collection methods. However, a few years later, when students want to apply this training in their research, they may have forgotten unapplied parts! Because researchers might not devote themselves to adequate re-study of data collection textbooks, they sometimes choose a data collection method without a systematic view of other methods and the wider framework of data collection. Researchers might devote their available time to selective study about applying their chosen data collection method, but a loss of overall methodological insight might have negative implications for the research. The present paper aims to help with this problem. It provides a researcher companion of data collection and validation methods in a systematic textbook style but within a concise paper with essential details and examples. The paper can be used by researchers of different disciplines as a standalone methodological guide or a legend for informed searching of more detail in relevant references.

Keywords: data collection; sampling; census; data validation; research

1. Introduction

Between the years 2017 and 2022, the author of this paper lectured at three universities in Asia and Europe. A common observation in this experience was that postgraduate students did not remember some parts of their undergraduate data collection training. Meanwhile, they could not devote time to adequate re-study of data collection textbooks for application in research. As such, these students often chose data collection methods in their research without a systematic insight into the wider framework of data collection. This had negative implications for the time, cost and quality of their research as a result of problematic or less optimum choices. The present paper aims to help with this problem. It provides a researcher companion of data collection and validation methods in a systematic textbook style but within a concise paper with essential details and examples.

The paper is structured as follows. This introduction is followed by a brief overview of fundamental terms of data collection. Then, there is the main Section 3 of the paper which presents the overall framework for collecting and validating data. There is a flowchart at the beginning of Section 3 that displays the systematic steps of data collection and validation as subsequently explained. The conclusions section guides applying this researcher companion on its own and in connection with other literature on data collection methods.

2. Fundamental terms of data collection

2.1. A variable

A variable is a name that describes data. Examples of variables are means of transportation, gender, green space size and income per capita. Variables consist of two types: (a) qualitative variables, like means of transportation or gender, and (b) quantitative variables, like blood pressure, income per capita, number of children and number of days missed from work [1].

Sometimes, qualitative variables that consist of a natural order are given quantitative data: for example, very dissatisfied = 1, relatively dissatisfied = 2, neutral = 3, relatively satisfied = 4, very satisfied = 5. These numbers are ordinal (i.e., only their order is meaningful). They cannot have a mean or standard deviation [2].

Qualitative variables could be categorized into dichotomous and polytomous variables. Dichotomous variables like coin side (heads or tails) only include two qualitative options. On the other hand, polytomous variables like means of transportation include more than two qualitative options (e.g., automobile, bus, train, bicycle) [2].

2.2. Population or target population

A population or target population is the entire set of individuals or elements about which information is sought and inferences are made [3]. Examples might include all Americans, all residents of Oxford City, or all honeybee nests in a specific region [4].

2.3. A census

A census collects data from all individuals or elements of a population [1]. A census is usually conducted by central, regional or municipal governments, for instance, to collect comprehensive data about the whole population (e.g., age, gender, number of children, employment, income, house tenure, etc.) or some data about the whole population (e.g., vacant homes).

2.4. A collection unit in a census

A collection unit in a census is the same as an element when elements can be enumerated directly: for example, heritage buildings. Otherwise, a collection unit is a container of one or more elements that cannot be enumerated directly. For example, parks can be the collection unit for skateboarders or gypsies [5].

2.5. Sampling

Sampling or *survey* collects data from some individuals or elements of a population [6]. A sample is the result of sampling. A sample should ideally be a miniature representative of the population from which it is selected [4]. If this condition is met, we can use a sample to make conclusions about its entire population. A sampling unit is the same as an element when elements can be sampled directly. Otherwise, a sampling unit is the container of one or more elements that cannot be sampled directly. For example, households can be the sampling unit of smokers or people who travel to work on foot [1].

2.6. Data validation

Data validation involves exploring and addressing errors in the collected data.

3. The framework for collecting and validating data

Data collection and validation consists of four steps when it involves taking a census and seven steps when it involves sampling (Figure 1). These steps are explained below.

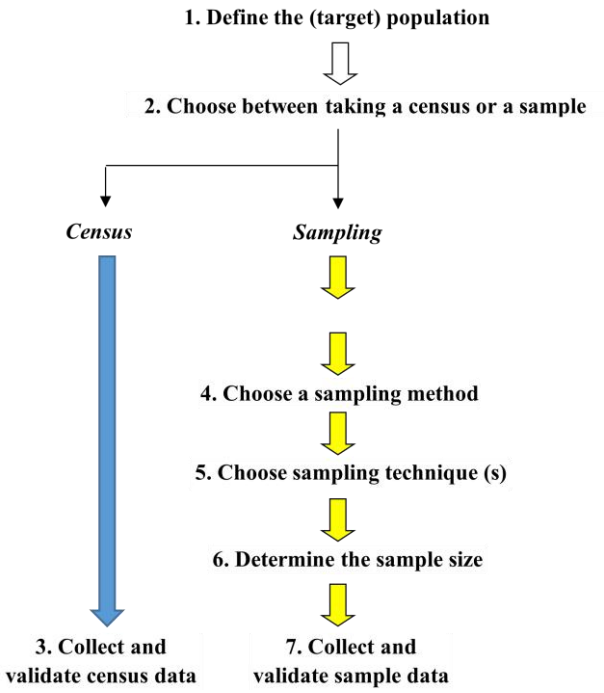

Figure 1. Data collection and validation. Author's work based on [1,7,8].

3.1. Define the (target) population

Any study should begin with a clarification of its aim and objectives. This will help define the (target) population and sampling units [4]. In a study that aims to explore the correlation between park visits and the health of older citizens, the target population might be all adults over 50 years old in city X, and the sampling units might be households in city X.

3.2. Choose between census and sample

A census or complete enumeration is costly and time-consuming. However, a census is desirable if the size of the target population is small (e.g., homeless support charities in Cairo) or the target population is very heterogeneous [9]. A census would also be desirable if the cost of sampling errors is high; for example, where the sample could miss major elements. For example, a census would be more suitable than a sample for a study of heritage building types in a city.

A sample is preferred if data collection disturbs, damages or consumes elements. For example, interview with city authorities creates a disturbance to their responsibilities. A sample is also desirable if in-depth data needs to be collected, and it is likely to observe notable similarities in data. For example, citizens are asked to describe problems in their living environment [10].

3.3. Specify the sampling frame

A sampling frame is a list of units or elements from which the sample will be selected. A sampling frame might not totally cover the population. This is called sampling frame error [6]. In the example of studying the correlation between park visits and the health of older adults in city X, the researcher

might obtain from the city's population register a full list of households in city X. This list is the sampling frame for this study. If the list misses some households, there is a sampling frame error which will lead to bias in the sample data.

Researchers should identify available sampling frames and determine which is best for their study. This involves searching for updated population registers or census lists and other databases that give good coverage of the population that the researcher wishes to survey [11]. Also, researchers should ensure that sampling frames include telephone numbers and postal addresses. Random digit dialing of phone numbers is popular in social research where a sampling frame is difficult to obtain. However, this is not suitable for areas with less access to landline phones [12]. Sometimes researchers change the sampling unit if the list of sampling units is difficult to obtain. In the previous example, the researcher might choose parks instead of households as units of sampling because the list of parks is easily accessible or because it enables better targeting of the sample. However, this might increase bias (e.g., the day and time of sampling in parks).

Treatment of sampling frame error might involve redefining the population to match the sampling frame, but this might undermine survey objectives and the generalizability of its findings. Another method to treat sampling frame error—which will be explained later—involves adjusting the collected data by weighting to counterbalance the missed part of the population [1].

3.4. Choose a sampling method: probability or non-probability

Sampling methods are usually divided into two types: probability and non-probability. Probability sampling involves random selection, which means all members of the population have an equal probability of being selected in the sample. This suggests that data collected from the sample can be generalized to the population. Often, research questions and objectives determine if probability sampling is needed. For example, whereas a study about the prevalence of domestic violence might require probability sampling, a study of community gardening may not.

Non-probability sampling involves non-random selection, which means all members of the population do not have an equal probability of being selected in the sample. This suggests that there is a limitation in generalizing data collected from a non-probability sample [4]. For instance, in a study of community gardening, the researcher might select practices that can provide richer lessons. However, the researcher will need to acknowledge in the study findings that these lessons are context specific. Table 1 indicates survey conditions in which probability sampling or non-probability sampling is more desirable.

Table 1. Non-probability sampling versus probability sampling. Author's work based on [3,6].

Survey condition	Non-probability sampling	Probability sampling
1. Application of findings in criteria setting	No	Yes
2. Heterogeneous population	No	Yes
3. Scattered population	Yes	No
4. Qualitative research design	Yes	No
5. Limited budget and time	Yes	No

3.5. Choose sampling technique(s)

Probability sampling and non-probability sampling consist of several techniques that are displayed in Figure 2 and explained below.

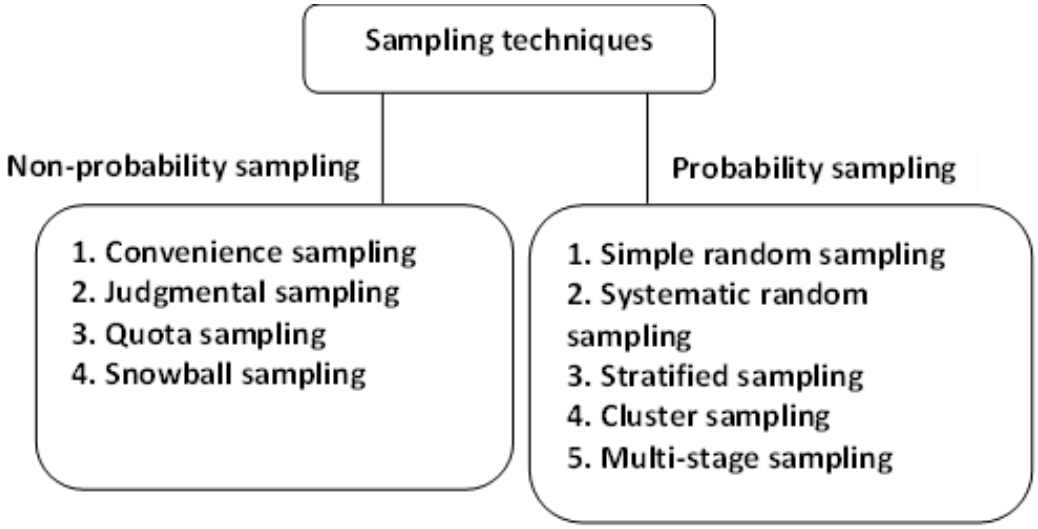

Figure 2. Sampling techniques (Author).

3.5.1. Non-probability sampling techniques

(1) Convenience sampling (sampling the most accessible elements)

In convenience sampling elements are selected because of their accessibility to the researcher [6]. Convenience sampling is often used by students in their studio projects to collect data about buildings, households, etc., in the proximity of student accommodations. A main disadvantage of convenience sampling is the lack of control over the characteristics of the sample [9,13].

(2) Judgmental or purposive sampling (sampling the most informative elements)

In this technique elements are selected from the target population based on the researcher's view about the particular usefulness of their data for the survey objectives [9]. For instance, in a survey about carbon-neutral buildings in a city, the researcher might collect data about carbon-neutral buildings that have a vernacular design. This technique provides more control over sample characteristics. However, the researcher should be more knowledgeable about the target population and study details, and the sample is subject to unknown biases [9,13].

(3) Quota sampling (sampling with proportion from the most accessible individuals)

In this technique**,** the researcher creates a sample of elements that are usually in one respect representative of a population. However, the researcher applies convenience sampling in meeting this proportion condition [9]. For instance, in a survey about community volunteering in city X, the researcher might assume that education level impacts community volunteering. The researcher knows that 63% of residents in city X do not have an academic degree, and 37% have academic qualifications. Although the researcher is conducting non-probability sampling, he/she wants to increase the representativeness of the sample. The researcher should collect a sample of residents with the same

proportion of education levels. However, in this non-probability sampling, the researcher can collect this proportionate sample from residents most accessible to him or her.

(4) Snowball sampling (sampling by referrals)

Sometimes, a sampling frame is not available for a population because elements of the population, such as traffickers or undocumented migrants, are not readily identifiable, or their identity cannot be disclosed. In such circumstances, sampling begins with a few individuals in the target population who are known to the researcher. These individuals are then asked to connect the researcher with other individuals in the target population. By obtaining referrals from referrals, this sampling process leads to a snowball effect [1,9]. Despite its benefits, in this sampling technique, the researcher will give control of sampling to individuals in the sample, which might lead to a less representative sample [9,13].

3.5.2. Probability sampling techniques

(1) Simple random sampling (sampling by random numbers)

This technique assigns a number to each element in the sampling frame and uses a random number generator to select from these elements. Random number generators are available for free on the Internet [6]. For instance, a researcher wants to study social issues of vacant lots in city X. For a simple random sampling of these lots, the researcher might obtain their list from the city's land registry, assign them numbers (e.g., 1–450) and then use a random number generator to select from these numbers. This technique is easy to use, but it requires the availability of the list of the target population, i.e., the sampling frame [9,13].

(2) Systematic random sampling (random sampling covering the whole population spectrum)

This technique assigns a number to each element in the sampling frame. The sample is created by selecting a random starting point and then picking every i^{th} element in succession from the sampling frame [14]. The sampling interval (i) is determined by dividing the population size (N) by the sample size (n) and rounding to the nearest whole number. Systematic random sampling is the most representative where elements in the sampling frame are ordered in respect of some characteristic of interest [1].

For instance, in a survey about the political views of public sector employees, the researcher wants to have a systematic random sample representing the salary spectrum in the target population. This is because the literature holds that salary plays a role in political views. There are 1456 public sector employees, and the researcher needs a sample of 40. The sampling interval is 36 (1456 ÷ 40). The researcher sorts the list of 1456 employees in ascending order of salaries. The researcher uses a random number generator to select an employee from the list. If it is employee number 296, the next individual in the sample would be employee number 332 (296 + 36), the next would be employee number 368 (332 + 36), and so forth. This sampling process continues by counting to employee number 1456 and then continuing the count from the start of the list to employee number 296.

(3) Stratified sampling (random sampling representing every group in the population)

This technique is used to sample a heterogeneous target population [15]. The population is divided into homogeneous groups or strata, and then in each stratum, simple random sampling is undertaken [6]. This means a complete list of each stratum is needed. In proportional stratified sampling, the size of the sample drawn from each stratum is proportionate to the relative size of that stratum in the population. This requires knowledge about strata proportions in the population. The elements within a stratum should be as homogeneous as possible, but the strata should be as heterogeneous as possible [1].

For instance, in December 2018, the percentage of blood donors with each blood type in the UK was as follows: O positive: 35%, O negative: 13%, A positive: 30%, A negative: 8%, B positive: 8%, B negative: 2%, AB positive: 2%, AB negative: 1%. A researcher wants to study the *prevalence* of cardiovascular disease among blood donors. There are reports that cardiovascular disease is more prevalent among people with certain blood types. The researcher, therefore, wants to undertake proportional stratified sampling of the blood donors to ensure that the study results are representative. The sample size is 164. This sample size should be distributed proportionately between donors with different blood types. This distribution is done by multiplying the sample size by the percentage of blood type in the target population. For example, 57 (164 × 35%) individuals should be randomly selected from donors with blood type O positive, and so forth.

(4) Cluster sampling (random sampling from an unknown and scattered population)

Cluster sampling can be used when a complete list of the population is unavailable. It can also be used when the population is scattered over a wide geographical expanse. Cluster sampling involves three steps as follows: (a) dividing up the map of the target population into clusters of similar size, (b) taking a simple random sample of the clusters and (c) covering all eligible elements in the sampled clusters [1,4]. Cluster sampling helps estimate the number of elements in the population by multiplying the average number of elements in the sample clusters by the number of clusters. Cluster sampling also has data collection merits. It ensures that elements of the sample are assembled in a few places, which will reduce data collection work [9,16]. However, this might decrease sample representativeness.

For instance, a researcher wants to study home-based food businesses in a metropolitan area comprising several cities and a peri-urban area. The number of these businesses is not large, but their sampling frame (complete list) is not available for the metropolitan area. The researcher divides the metropolitan area into square clusters (Figure 3). Incomplete clusters, like cluster 7, are counted together as a complete cluster. The metropolitan area is divided into 51 clusters. The researcher will need to simple random sample from the 51 clusters. If the researcher selects 4 clusters, all home-based food businesses in those 4 clusters are included in the study.

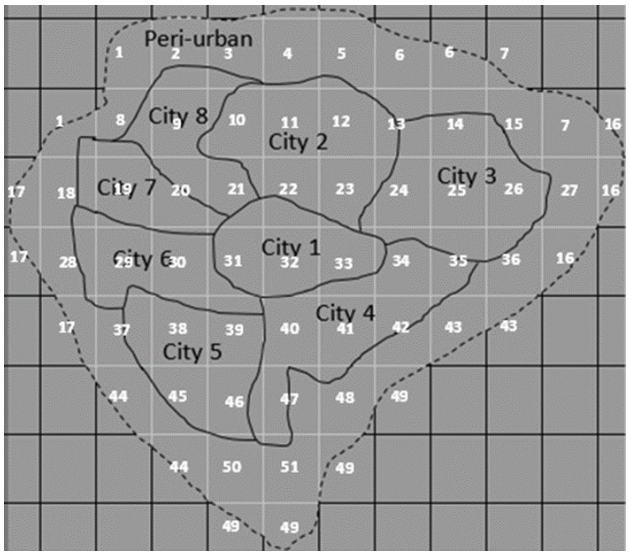

Figure 3. Dividing up a metropolitan area into clusters.

(5) Multi-stage sampling (random sampling from an unknown and large population)

Multi-stage sampling can be used when there is a large population that does not have a sampling frame. It is also used when there is a large population that has a sampling frame, but the population covers a wide geographical expanse [7,17]. A common two-stage sampling involves (a) dividing up the map of the target population into clusters of similar size, (b) taking a simple random sample of the clusters and (c) taking a simple random sample of the elements of the selected clusters [18].

Overall, multi-stage sampling might involve different stages in moving from a broad to a narrow sample, and it might combine several probability sampling techniques [9,17]. By taking samples from samples, multi-stage sampling reduces the cost and effort needed for probability sampling. For instance, 30 × 7 sampling is a two-stage sampling technique developed by the World Health Organization in 1978. In this technique, the map of the target population is divided into clusters of the same size. Then, 30 clusters are selected by simple random sampling. Then, one household is selected by simple random sampling from each of the 30 clusters. The selected households and their nearest six households will comprise the 210 elements of the sample [19].

3.6. Determine the sample size (n)

There are different methods to determine sample size, including using a formula, sample size calculator, table and sample size from another study. Sample size formulas are available in two main categories: for qualitative variables and for quantitative variables. They are explained below.

3.6.1. Sample size formulas

(1) Sample size formula (1) for qualitative variables

$$n = \frac{p(100 - p)}{SE^2} \tag{1}$$

where n = sample size, p = expected proportion of the variable in the population. It is based on previous studies or a pilot study.

$$SE\ (standard\ error) = \frac{margin\ of\ error}{1.96\ (if\ confidence\ level\ is\ 95\%) or\ 2.56\ (if\ confidence\ level\ is\ 99\%)}$$

Margin of error is a permissible degree for the inaccuracy of sample results. Most studies accept a margin of error of 5%. Confidence level is a measure of the accuracy of sample results. Most studies accept a confidence level of 95%, but some prefer 99%. Margin of error, p and $1 - p$ are used in Equation 3.1 without percent [20]. For unknown p, the researcher should include 50%. If the qualitative variable is polytomous and has several expected proportions (e.g., blood types), the p that is closest to 50% will be put in the formula [21].

For instance, a researcher wants to study immigrant households in city X. A previously published study holds that 25% of households in city X are immigrants. The researcher is willing to accept a 95% confidence level and a 5% margin of error for the study. The sample size is calculated as follows:

$$n = \frac{25\ (100 - 25)}{(\frac{5}{1.96})^2} = \frac{1875}{6.5} = 288\ (rounded)$$

In another example, a researcher wants to study building materials in city X. The variable building materials is polytomous. Prior information holds that 80% of buildings are made of brick, 15% are concrete, 3% are mud, and 2% are wooden. The researcher is willing to accept a 95% confidence level and a 5% margin of error for the study. The sample size is calculated as follows:

$$n = \frac{80\ (100 - 80)}{(\frac{5}{1.96})^2} = \frac{1600}{6.5} = 246\ (rounded)$$

If the population has a known size, and n > 5% N, the sample size is reduced by using Eq (2), in which n_0 is the initial sample size [21]. This is a *common rule* in sample size calculation by formulas.

$$n = \frac{n_0}{1 + \frac{n_0}{N}} \tag{2}$$

(2) Sample size formula (3) for quantitative variables

$$n = (\frac{SD}{SE})^2 \tag{3}$$

where *n* = sample size, and SD = the standard deviation. It may be obtained from a pilot sample of 30 elements or from a previous study.

$$SE\ (standard\ error) = \frac{margin\ of\ error}{1.96\ (if\ confidence\ level\ is\ 95\%) or\ 2.56\ (if\ confidence\ level\ is\ 99\%)}$$

Margin of error is a permissible degree for the inaccuracy of sample results. Most studies accept a margin of error of 5%. Confidence level is a measure of the accuracy of sample results. Most studies accept a confidence level of 95%, but some prefer 99%. Margin of error is used in Equation 3.3 without percent [20]. For example, in a study of physician visits per week, a pilot sample of 30 physicians provides a standard deviation of 8 visits. If the researcher is willing to accept a 95% confidence level and 2% margin of error, the sample size for the study is calculated as follows:

$$n = (\frac{8}{\frac{2}{1.96}})^2 = (\frac{8}{1.02})^2 = 7.84^2 = 61.51 = 62\ (rounded)$$

If a survey involves a mix of qualitative and quantitative variables, the sample size will be determined for the variable which plays the most important role in the study [21].

3.6.2. Sample size calculators and tables

Sample size calculators are available online and as Excel files. Tables can also be used to determine sample size. Table 2 provides a general reference for sample size in many surveys.

Table 2. Sample size table. Author's work based on [21].

Population Size	Sample size					
	Quantitative variables Margin of error = 3%			Qualitative variables Margin of error = 5% Population proportion = 50%		
	Confidence level (1−α)			Confidence level (1−α)		
	1−α = 90% Z = 1.65	1−α = 95% Z = 1.96	1−α = 99% Z = 2.58	1−α = 90% Z = 1.65	1−α = 95% Z = 1.96	1−α = 99% Z = 2.58
100	46	55	68	74	80	87
200	59	75	102	116	132	154
300	65	85	123	143	169	207
400	69	92	137	162	196	250
500	72	96	147	176	218	286
600	73	100	155	187	235	316
700	75	102	161	196	249	341
800	76	104	166	203	260	363
900	76	105	170	209	270	382
1,000	77	106	173	213	278	399
1,500	79	110	183	230	306	461
2,000	83	112	189	239	323	499
4,000	83	119	198	254	351	570
6,000	83	119	209	259	362	598
8,000	83	119	209	262	367	613
10,000	83	119	209	264	370	623

3.6.3. Sample size from a similar study

A study may use the sample size of a previous similar study. A disadvantage of this is reliance on someone else correctly determining the sample size. However, the procedures employed in the previous study to determine sample size may be reviewed to ensure they are correct [22].

3.7. Collect and validate sample data

After the sample size is determined, the researcher begins to collect data with a particular effort to avoid bias. Efforts to avoid bias in data collection might involve the following:

- Use trained data collectors;
- Identify a larger sample size than you need. Some suggest that after the sample size is determined, the researcher should increase it up to 40% [22];

- Send reminders to the recipients of mail surveys, and make repeat phone calls to potential telephone survey respondents;
- Provide gift or cash incentives to respondents; and be realistic about your target population [4].

However, we cannot ensure that there will be no errors in the data. We need to explore and address errors in the data. This is called data validation. Three types of sample data validation are explained below.

3.7.1 Inspect data by common sense and drop troublesome or suspicious elements

For instance, a researcher who surveys the incomes of a sample of 120 households might inspect the sample data and find out that incomes reported by 21 households are unbelievably low in relation to their jobs, houses, etc. If the researcher has adequately oversampled in the prediction of this bias, the researcher will drop the 21 suspicious households and study the remaining 99 households. However, if the researcher has not oversampled, dropping the problematic 21 households will lead to under-sampling. As such, the researcher will need to sample another 21 households or carry out imputation.

3.7.2 Conduct imputation

There are three types of imputation: mean imputation, dynamic imputation and regression imputation. Mean imputation involves substituting the sample mean for the missing or problematic data. In the previous example, the researcher can put the mean income of the 99 unsuspicious households for the 21 suspicious households. But given this large number of mean replacements, the sample's representativeness will decline. Therefore, other types of imputation will be more suitable in this example. Dynamic imputation involves substituting the data collected from a similar individual for the missing data. In the previous example, the researcher can put for each of the 21 suspicious households the income of an unsuspicious sample household that has a similar job or house [23].

Regression imputation involves using regression analysis to estimate the missing data [9]. In the previous example, the researcher might use regression line of the relationship between years of experience and income to estimate the incomes of the 21 suspicious households. Therefore, in a sensitive topic like income, it is good practice to collect supplemental predictor information, like years of experience, that can be used for regression imputation where needed.

3.7.3 Check and address sampling frame error

Proportions of interest in the sample should be compared with accurate population proportions. If these proportions do not match despite meeting survey requirements, there is a sampling frame error. To address this error, weighting is applied to equalize sample proportions with population proportions. The weights are obtained by dividing population proportions by the corresponding sample proportions [1].

For example, a mail survey was conducted in city X to determine the patronage of a community center. The resulting sample differed in age structure from the area population. The researcher reviewed the sampling process and concluded that there was no error in sampling. Therefore, there should be a sampling frame error. This error could not be addressed by repeating sampling. As such, the sample was weighted to equalize sample proportions with population proportions in terms of age groups. The weights applied were determined by dividing the population proportions by the

corresponding sample proportions (Table 3). For instance, the data for a respondent aged 13–18 would be overweighted by multiplying by 1.42, whereas the data for a respondent aged 75 plus would be underweighted by multiplying by 0.75 [1].

Table 3. Correcting sampling frame error [1].

Age group	Sample percentage	Population percentage	Weight
13–18	4.32	6.13	1.42
19–24	5.89	7.45	1.26
25–34	12.23	13.98	1.14
35–44	17.54	17.68	1.01
45–54	14.66	15.59	1.06
55–64	13.88	13.65	0.98
65–74	15.67	13.65	0.87
75 plus	15.81	11.87	0.75
Total	100	100	

3.8. Collect and validate census data

If a census is chosen for data collection (step 3), the process to collect census data should avoid two broad kinds of error: Population under-coverage, the exclusion of elements that should have been enumerated, and population over-coverage, the inclusion of elements in more than one enumeration. Under-coverage can occur if the list of collection units (e.g., dwellings, clinics, parks) is incomplete. Over-coverage can occur if a collection unit is listed twice or some elements are included in two collection units, such as people with part-time residences or gypsies moving between parks [24]. Census data should be checked for these errors, and where these errors are found, they should be corrected in census data. This is the validation of census data. Validation of census data has two methods: dual systems and data accounting. They are explained below.

3.8.1. Dual systems validation of census data

Dual systems validation applies sampling to explore and correct errors in census data [25]. In other words, cluster sampling of the geographical area of the census is conducted, resulting in a random selection of *n* clusters. These selected clusters are referred to as the P-sample. Eligible collection units of the P-sample are listed by fieldwork – i.e., their list is provided independently from the list used to take the census. Then, individuals of the P-sample are enumerated in the collection units listed by fieldwork. The P-sample enumeration is then compared with the census enumeration in the selected clusters known as the E-sample enumeration. This will discover census coverage errors and help identify alternative processes to prevent them in the future. It is also possible to correct census coverage errors and estimate the true population count by Eq (4) [23]:

$$DSE = (C - II)\left(\frac{E \cap P}{E}\right)\left(\frac{P}{E \cap P}\right) \tag{4}$$

DSE: the dual systems estimate of true population count
C: census enumeration
II: Enumeration of individuals in the census with suspicious eligibility

E: E-sample enumeration
P: P-sample enumeration
$E \cap P$: matching enumerations of E and P

For example, a census of the homeless population in region X guided by a directory of homeless centers enumerates 5,000 homeless individuals. The researchers want to validate the census data. It is estimated that 5% of the census enumeration relates to individuals with no precise clues for their inclusion (or exclusion) in the homeless population. Therefore, II = 5000 × 5% = 250 The researchers divide the region into 35 clusters and then randomly select 3 clusters as the P-sample. In the 3 clusters of the P-sample, the researchers collect a list of homeless centers by fieldwork. They list 8 homeless centers (including registered and informal centers). In these centers, the researchers enumerate 400 homeless individuals. Therefore, P = 400. The census enumeration in the 3 clusters is 340 homeless individuals (E = 340), and 303 homeless individuals match between the two enumerations (E ∩ P = 303).

The dual systems estimate of the true homeless population count in the region is

$$DSE = (5000 - 250)(\frac{303}{340})\left(\frac{400}{303}\right) = 4750 \times 0.89 \times 1.32 = 5580$$

3.8.2. Data accounting validation of census data

This method of validating census data includes additions to / reductions from the enumeration of a previous census to estimate the true population for a new census that needs validation [23]. For example, a recent census of student homes in city X enumerates 952 homes. A researcher wants to validate this census data. A previous census had identified 324 homes. Municipality data indicates that in the period between the two censuses, 54 student homes went out of service, and 400 homes were added to student accommodation. According to data accounting, the true count of student homes for the recent census is estimated as follows: 324 – 54 + 400 = 670. There is a notable difference between the results of data accounting (670) and the recent census (952). This suggests population over-coverage in the recent census or an error in the municipality data of changes between the two censuses. To explore this and estimate the true count of student homes in city X, the researcher can apply the dual systems method of validation.

4. Conclusion

This researcher companion can be used as a standalone methodological guide or in connection with textbooks and detailed references about data collection methods. For the second application, the paper provides three main contributions as follows: (a) it helps design the data collection process and steps with a critical view of different alternatives; (b) it helps explore detailed references for the designed data collection process; and (c) with its systematic overview, it helps better explain technical contents in textbooks and detailed references about data collection methods.

Conflict of interest

The authors declare no conflict of interest.

References

1. Malhotra NK, Birks DF (2007) *Marketing Research: An Applied Approach*. Harlow: Pearson Education.
2. Menard S (2004) *Polytomous Variable, in The Sage Encyclopedia of Social Science Research Methods*. Thousand Oaks: Sage Publications.
3. Levy PS, Lemeshow S (2008) *Sampling of Populations: Methods and Applications*. Hoboken: John Wiley and Sons.
4. Fink A (2003) *How to Sample in Surveys*. Second edition. Thousand Oaks: Sage Publications.
5. United Nations (1982) *National Household Survey Capability Programme: Non-sampling Errors, in Household Surveys: Sources, Assessment and Control*. New York: United Nations.
6. Daniel J (2012) *Sampling Essentials: Practical Guidelines for Making Sampling Choices*. Thousand Oaks: Sage Publications.
7. Taherdoost H (2016) Sampling methods in research methodology; how to choose a sampling technique for research. *Int j res manag* 5:18–27. http://dx.doi.org/10.2139/ssrn.3205035
8. Gill J, Johnson P, Clark M (2010) *Research Methods for Managers*. London: Sage Publications.
9. Australian Bureau of Statistics (2019) Samples and censuses.
10. Dattalo P (2008) *Determining Sample Size: Balancing Power, Precision, and Practicality*. New York: Oxford University Press.
11. WHO (World Health Organization) (2005) The WHO STEPS Surveillance Manual.
12. Puszczak K, Fronczyk A, Urbański M, Pashova S (2013) *Analysis of Sampling Frames*. Task force on quality of BCS data, OECD.
13. Australian Bureau of Statistics (2019) Sample design.
14. Da Costa NM, Hepp K, Martin KA (2009) A systematic random sampling scheme optimized to detect the proportion of rare synapses in the neuropil. *J Neurosci Meth* 180: 77–81. https://doi.org/10.1016/j.jneumeth.2009.03.001
15. Sampath S (2001) *Sampling Theory and Methods*. Chennai: CRC Press.
16. Thompson SK (1990) Adaptive cluster sampling. *J Am Stat Assoc* 85: 1050–1059.
17. Chauvet G (2015) Coupling methods for multistage sampling. *Ann Statist* 43: 2484–2506. https://doi.org/10.1214/15-AOS1348
18. Yates F (1960) *Sampling Methods for Censuses and Surveys*, 2 Eds. London: Charles Griffin Co. Ltd.
19. Henderson RH, Sundaresan T (1982) Cluster sampling to assess immunization coverage: a review of experience with a simplified sampling method. *Bull World Health Organ* 60: 253–260.
20. Fox N, Hunn A, Mathers N (2009) *Sampling and sample size calculation*. National Institute for Health Research.
21. Bartlett JE, Kotrlik JW, Higgins CC (2001) Organizational research: Determining appropriate sample size in survey research. *Inf technol learn perform j* 19: 43–50.
22. Israel GD (2003) *Determining sample size*. University of Florida.
23. United Nations (2010) *Handbook on Population and Housing Census Editing*, New York: United Nations. https://doi.org/10.18356/1157f3b5-en
24. Statistics Canada (2016) Population coverage error.

25. National Research Council (2007) *Research and Plans for Coverage Measurement in the 2010 Census: Interim Assessment*. Washington: The National Academies Press.