Collecting and validating data: A simple guide for researchers

Abbas Ziafati Bafarasat[1]

[1]Affiliation not available

January 28, 2021

Abstract

This paper provides an A-Z guide for sampling and census within an integrated framework of data collection and validation. It provides worked examples and inter-disciplinary exemplifications of essential methods, techniques and formulas for each step of the framework. Real research application tricks are disentangled for non-statisticians.

Hosted file

`Main Document.doc` available at https://authorea.com/users/719238/articles/704250-collecting-and-validating-data-a-simple-guide-for-researchers

Collecting and validating data: A simple guide for researchers

Abbas Ziafati Bafarasat, Faculty of Art and Architecture, Yazd University
abbas.ziafatibafarasat@yazd.ac.ir

Abstract. This paper provides an A-Z guide for sampling and census within an integrated framework of data collection and validation. It provides worked examples and inter-disciplinary exemplifications of essential methods, techniques and formulas for each step of the framework. Real research application tricks are disentangled for non-statisticians.

Keywords: sampling; census; target population; data collection and validation

Introduction

The National Academies of Science, Engineering, and Medicine emphasize the interdisciplinary nature of data science and point out that researchers from all backgrounds should have opportunities to learn data science at all levels (Gundlach and Ward, 2020). This A-Z guide aims to address four constraints in inclusive use of sampling and census science as follows:

(I) The literature on sampling and census is either too brief or abound with details;

(II) There is no concise reference about sampling and census in an integrated framework of data collection and validation;

(III) Inter-disciplinary and worked examples are limited; and

(IV) Learning hooks in terms of diagrams, tables and application tricks are not customized for non-statisticians.

Basic terms

The *population* or *target population* is the entire set of individuals about which information is sought and inferences are made (Levy and Lemeshow, 2008). Examples might include all Americans, all residents of California during the 1994 earthquake, and all honey bee nests in a region (Fink, 2003). It is often impractical to collect *data* or a set of values (e.g. granite, 2.3 meters, 4.1 hours, retired) from all *individuals or elements* of a population by taking a *census* (Malhotra and Birks, 2007). Therefore, a subset of a given population might be selected in a process called *sampling* or *survey* (Dattalo, 2008). The result of sampling is a *sample* of individuals or elements that should ideally be a miniature representative of the population from which it is selected (Fink, 2003). If this condition is met, we can use a sample to make inferences about its target population as a whole. *Variable* is the name which describes and organizes data (e.g. stone type, height, study duration, job status).

Data collection and validation framework

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

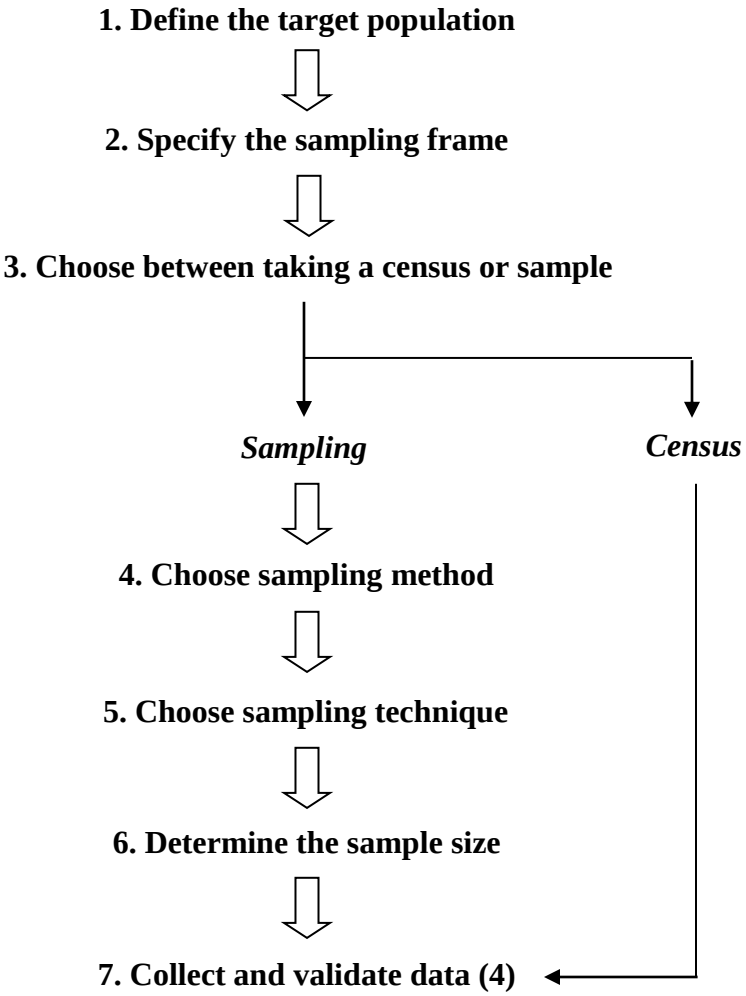

Figure 1: Data collection and validation (Adapted from Malhotra and Birks, 2007; Gill *et al*., 2010; Daniel, 2012; Taherdoost, 2016).

1 Define the target population

Survey objectives should be specified to enable applying clear and definite eligibility criteria to the target population in terms of elements, sampling units, geographical extent and time (Fink, 2003). A *sampling unit* is the same as an element when elements can be sampled directly, for example junior Chemistry students. Otherwise a sampling unit is the container of one or more elements which cannot be sampled directly. For example, households can be the sampling unit of

smoking parents and dwellings might be the sampling unit of solar panels (United Nations, 1982; Malhotra and Birks, 2007). It is less likely that an up-to-date list of all parents living in a large geographical extent will be available or easily obtainable. However, it is conceivable that a list of all households is available or can be constructed without great difficulty (Levy and Lemeshow, 2008).

2 Specify the sampling frame

A sampling frame is the *actually enumerable* or *listed* population from which the sample will be selected, which might not totally cover the population (Dattalo, 2008; Taherdoost, 2016). Suitable sampling frames of human populations consist of: lists of individuals in the population provided for administrative purposes; aggregates of census returns in a complete census; lists of households or dwellings in given areas; and lists or map of towns, villages, and administrative areas (Yates, 1960). However, such dynamic examples as census listings should not be used directly as a sampling frame except for surveys taken very soon, i.e. within six months from the census enumeration (United Nations, 1982). In other instances, the discrepancy between the population and the sampling frame might result from registration error or representation error (Zhengdong, 2011).

Worked example: Many Australian public health research studies use the telephone directory or the electoral roll as a sampling frame. In an investigation into the coverage of these sampling frames, Smith *et al* (1997) found that the telephone directory listed 82.2 per cent of people they had interviewed, and the electoral roll contained 84.3 per cent of them!

Worked example: Urban poverty sampling frames are tricky to apply not only because of their dynamism but also due to their contested exclusion and inclusion criteria. Hussain (2003) and Hao (2009) highlight this for 14 million and 22.7 million urban poor headcounts in China, respectively.

Treatment of sampling frame error might involve redefining the population in terms of the sampling frame, but this might undermine survey objectives and generalize-ability of its findings. Another method to treat sampling frame error - which will be explained later – involves adjusting the data collected by a weighted scheme to counterbalance the sampling frame error (Malhotra and Birks, 2007).

3 Choose between taking a census or sample

A census or complete enumeration is costly and time consuming, but it is preferable if size of the target population as reflected in the sampling frame is small (e.g. directory of homeless support charities in Cairo) and where variance in the characteristic to be measured is large (e.g. funding sources of charities) (Daniel, 2012). A census would also be more desirable if the cost of sampling errors is high, for example where the sample omitted a major element. On the other hand, a sample would be more desirable if the cost of non-sampling errors is high, for example where unqualified interviewers incorrectly question target respondents about satisfaction with their neighborhood (Malhotra and Birks, 2007). Also, a sample is preferred if intended measurements damage or consume elements (e.g. measuring life span of light bulbs), if many measurements need to be conducted in individual elements (e.g. in-depth interviews about educational experience), and if there are other practical and ethical considerations (Australian Bureau of Statistics, 2019). Table 1 summarizes survey conditions in which the use of a sample or a census would be more desirable.

Table 1: Sample versus census (Adapted from Malhotra and Birks, 2007; Daniel, 2012; Australian Bureau of Statistics, 2019a).

	Conditions favoring the use of	
	Sample	***Census***
1 Budget and time available	Limited	Considerable
2 Population (sampling frame) size and scatter	Large	Small
3 Variance in the characteristic	Small	Large
4 Cost of sampling errors	Low	High
5 Cost of non-sampling errors	High	Low
6 Nature of measurement	Damaging/consuming	Non-damaging/consuming
7 Multiple measurements in elements	Yes	No
8 Exploratory study and data updating	Yes	No

4 Choose sampling method

Sampling methods are usually divided into two types: *probability* sampling and *non-probability* sampling. Probability sampling implies random selection without subjectivity. In non-probability sampling, elements are chosen based on the researcher's decision (Fink, 2003). Table 2 indicates strengths and weaknesses of probability sampling and non-probability sampling.

Table 2: Non-probability sampling versus probability sampling (Adapted from Levy and Lemeshow, 2008; Daniel, 2012).

	Non-probability sampling	***Probability sampling***
1. Exploratory study	Strength	Weakness
2. Generalization of findings	Weakness	Strength
3. Heterogeneous population	Weakness	Strength
4. Scattered population	Strength	Weakness
5. Qualitative research design	Strength	Weakness
6. Limited budget and time	Strength	Weakness

5 Choose sampling technique

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

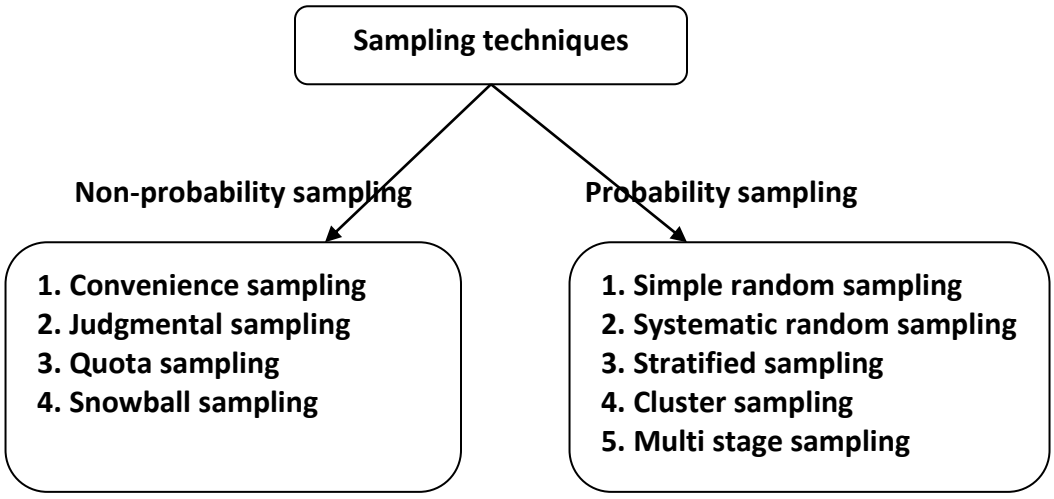

Figure 2: Sampling techniques (Adapted from Malhotra and Birks, 2007; Taherdoost, 2016).

5.1. Non-probability sampling techniques

5.1.1. Convenience (availability or haphazard) sampling: In convenience sampling - like street corner interviews and recruitment of respondents using Facebook - elements are selected because of their accessibility to the researcher (Dattalo, 2008; Zhang *et al*., 2020).

Advantages:

•Easy and less time-consuming

Disadvantages:

•No control over sample characteristics (Daniel, 2012; Australian Bureau of Statistics, 2019b).

Worked example: In convenience sampling of the urban forest of Meran (Italy), Speak *et al.*, (2018) obtained significant species diversity, incorporating rare species. This was a result of covering wide area per unit sampling time!

5.1.2. Judgmental (purposive) sampling: In this technique, elements are selected from the target population on the basis of the researcher's judgment about their typicality, variability, hypothesis confirmation or disconfirmation, and information productivity (Daniel, 2012)

Advantages:

• More control over sample characteristics

Disadvantages:

• Researcher should be more knowledgeable about the target population, the sampling sites and study details

• Sample is subject to unknown biases (Daniel, 2012; Australian Bureau of Statistics, 2019b).

Worked example: A survey was undertaken in Nigeria to examine the extent to which Treasury Single Account can block financial leakages, promotes transparency and accountability in the public financial management. To answer the relevant questionnaire consisting of a Likert scale - where 1 referred to strongly disagree, and 5 strongly agree - a sample of 72 respondents was chosen using judgmental sampling from Ministries, Departments and Agencies within Bauchi metropolis (Bashir, 2016).

5.1.3. Quota sampling: This technique seeks to ensure that the sample represents certain characteristics in proportion to their prevalence in the target population. The researcher gives data collectors a specific proportion of elements to be selected, and data collectors then use convenience sampling in selecting elements that satisfy their quota controls (Daniel, 2012).

Advantages:

• Some representativeness of the target population

Disadvantages:

• Proportions of interest in the target population should be known (Daniel, 2012; Australian Bureau of Statistics, 2019b).

Worked example: In a survey about perceptions of dissection by students in one medical school, students were approached at convenience to obtain a quota sample of 29 medical students from Years 1–5, using proportions in the student population of the medical school as the sampling frame (Lempp, 2005). The proportional quota sample was as follows:

Training stage	12 students in Years 1 and 2, 11 students in Years 3 and 4, 6 students in Year 5
Gender	17 female, 12 male
Average age	22 years
Place of birth	25 UK, 1 Europe not UK, 3 outside Europe
Entry to medical school	19 after school, 5 after a gap year, 5 mature
Religion	8 Christian, 7 Muslim, 1 Hindu, 1 Jewish, 12 none
Self-identified ethnicity	1 African-Asian, 1 Arab, 1 Bangladeshi, 2 Black, 1 Chinese, 1 Jewish, 2 Indian, 2 Pakistani, 1 Persian, 17 White

5.1.4. Snowball (respondent-assisted) sampling: Sometimes a sampling frame is not available because its individuals - such as persons with AIDS, sex workers, drug users and members of a faith group - are not readily identifiable or their identity cannot be disclosed. In such circumstances, sampling begins with studying a few individuals in the target population who are known to the researcher. These individuals are then asked to help identify other individuals in the target population and maybe encourage them to take part in the study. By obtaining referrals from referrals, this process leads to a snowballing effect (Daniel, 2012; Malhotra and Birks, 2007).

Advantages:

• Effective in sampling hard-to-reach groups

Disadvantages:

• Non-independence of observations (Daniel, 2012; Australian Bureau of Statistics, 2019b).

Worked example: Browne (2005) studied non-heterosexual women using a sample of 28 individuals snowballed from 13 of these women who were Browne's friends prior to the study. Some of the 28 participants were friends with each other and some were in relationships with each other.

5.2. Probability sampling techniques

5.2.1. Simple random sampling: This technique assigns a number to each element in the sampling frame and uses a random number generator to select elements from the sampling

frame. Random number generators are included in some commercial software and are available free on the Internet (e.g., http://www.random.org) (Dattalo, 2008).

Advantages:

•Simplicity and ease of use

Disadvantages:

•Requires a complete list of the target population, i.e. the sampling frame (Daniel, 2012; Australian Bureau of Statistics, 2019b).

5.2.2. Systematic random sampling: This technique also assigns a number to each element in the sampling frame. The sample is chosen by selecting a random starting point and then picking every *ith* element in succession from the sampling frame. 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 most representative where elements in the sampling frame are ordered in respect of some characteristic of interest (e.g. firms list ordered according to annual sales) (Malhotra and Birks, 2007).

Advantages:

•Ease of use

•Helpful in observing rare phenomena in a large expanse

Disadvantages:

•Periodicity bias

•Requires a complete list of the target population (da Costa *et al*., 2009; Daniel, 2012).

Worked example: In a survey with sampling frame of New England soil map, Schellentrager and Doolittle (1991) applied systematic sampling of map grids for a soil survey.

Da Costa *et al*., (2009) systematically sampled copper grids of coronal sections drawn from the cat's brain to detect the proportion of rare synapses in the neuropil.

5.2.3. Stratified sampling: This technique is used to increase representativeness in probability sampling from a heterogeneous target population (Sampath, 2001). The sampling frame is divided into homogeneous groups or strata (e.g., age groups, gender), and then a sample is taken from each stratum usually by simple random sampling (Dattalo, 2008). In proportional stratified sampling, the size of the sample drawn from each stratum is in proportion to the relative size of that stratum in the sampling frame. Stratification criteria consist of homogeneity, heterogeneity, relatedness and cost. The elements within a stratum should be as homogeneous as possible, but

the elements in different strata should be as heterogeneous as possible. Stratification should also be closely related to the characteristic of interest. And, variables used for stratification should be easy to measure and apply (Malhotra and Birks, 2007).

Advantages:

•Oversampling of minority groups of interest can be designed in disproportional stratified sampling

•The results are more representative

•Different selection, analysis and estimation procedures can be applied to the various strata.

Disadvantages:

• For simple random sampling in each stratum, it needs a complete list of the target population

• Increase in costs

• Danger of stratifying too finely (Daniel, 2012; Australian Bureau of Statistics, 2019b).

Worked example: In a study of domestic energy usage patterns in Hong Kong, Tso and Yau (2003) believed that these patterns vary significantly between housing types. Therefore, they stratified the target households into four different housing types: public rental (38%), government subsidized home ownership scheme (15%), private development (41%) and village housing (6%). Proportional to these percentages, a total of 1516 households were randomly sampled using strata lists to complete a questionnaire.

5.2.4. Cluster sampling: Cluster sampling is, similarly to systematic random sampling, used in surveys with large expanse. A cluster is a naturally occurring mega-unit, such as a school or university, hospital, etc. (Fink, 2003). A common form of cluster sampling involves simple random sampling from geographical areas - such as provinces, watersheds, counties, cities or residential blocks - and then covering all eligible individuals in the sampled clusters (Fink, 2003; Malhotra and Birks, 2007). Adaptive cluster sampling allows adding clusters from the neighborhood of randomly selected clusters where random clusters satisfy a condition of interest. However, the adaptive selection procedure introduces biases into sampling (Thompson, 1990).

Advantages:

• Eliminates the need for a complete list of the target population

• Ensures that the sample individuals will be closer together, thus enumeration costs and field work will be reduced.

Disadvantages:

• Less accurate when the number of sampled clusters is very small

• Less applicable without combination with other techniques (Thompson, 1990; Daniel, 2012).

5.2.5. Multi-stage sampling: This technique, which is widely used for household and health surveys when there exists no sampling frame, or when the population is scattered over a wide area, involves moving from a broad to a narrow sample, using a step by step process, which might ultimately combine several probability sampling techniques (Chauvet, 2015; Taherdoost, 2016).

Advantages:

• Eliminates the need for a complete list of the target population

• Ensures that the sample elements or individuals will be closer together, thus enumeration costs and field work will be reduced.

Disadvantages:

• Less accurate where it is just a combination of cluster sampling and simple random sampling (Daniel, 2012; Chauvet, 2015).

Worked example: The 30 by 7 sampling design was developed by WHO in 1978. Its goal was to estimate immunization coverage in children by surveying a random selection of 210 individuals using two sampling stages as follows:

1. Random selection of 30 clusters from clusters (i.e. geographical areas of interest) into which the target population (e.g. children in a country) is divided.

2. Random selection of one household within each chosen cluster and then continuing to the next nearest household until a total of 7 children of the appropriate age is obtained (Henderson and Sundaresan, 1982).

In the event that a list of households or a map that shows the location of all dwellings is not available for the chosen clusters, another stage of cluster sampling can be undertaken to enable random household selection in Stage 2. This means that the following should additionally be done, for example, if clusters are cities:

- The streets map of the city will be used to divide it up into blocks or areas bounded by streets. If blocks are of very unequal area, groupings of the smaller blocks and subdivision of the larger blocks should be performed.

- One block is randomly selected in which all the dwellings will be marked in fieldwork. This will enable random household selection by numbered dwellings in the chosen block (see Yates, 1960).

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. Selection, application and outcomes of these methods, depend on the following information and preferences.

(a) Type of survey variables including:

• ***Dichotomous*:** variables like gender and relationship status for which *two qualitative options* (e.g. single or engaged) exist.

• ***Polytomous*:** variables like skin color and job satisfaction for which *more than two qualitative options* (e.g. very dissatisfied, relatively dissatisfied, neutral, relatively satisfied, very satisfied) exist (Menard, 2004).

• ***Quantitative*:** variables like *a 7 point scale* to measure job satisfaction, *a 10 point scale* to rate residential environment, number of children, number of days missed from work, temperature, blood glucose, weight, height, distance, and market share for which *a range of number values exist that have a meaningful mathematical difference* (Malhotra and Birks, 2007); and

(b) Information about the target population:

• For dichotomous and polytomous variables which are qualitative, researchers need to know *proportions of variables* (p) *in the target population*. Information about population size (N) will also be helpful. Information about population proportions might be based on prior research, pilot study, estimates from experienced researchers or industry conventions. If there is no clue, the most conservative estimate of a population proportion is 50% (Daniel, 2012).

• For quantitative variables, researchers need to know *standard deviation (σ) of a variable in the population*. Because σ is usually unknown, researchers can replace it with standard deviation (s) of a variable from the sample which will be estimated as explained later (Israel, 2003).

(c) Desired accuracy:

Researchers need to decide about their desired survey accuracy partly through determining sample size. This decision involves selecting the *margin of error* (e) and choosing the *confidence level* ($1-\alpha$) in formulas, tables and calculators of sample size. A high confidence level and small margin of error will increase sample size (Gill *et al*., 2010). The margin of error, which is also known as the *level of precision*, is number of percent points sample values will differ from the true population values (Israel, 2003). For instance, if a researcher uses the margin of error of 5% in size determination of a sample to survey domestic violence, and in 22% of the sample there is

domestic violence, the true population value for domestic violence will fall between ± 5% from 22% (i.e. between 17% and 27%). This 17% - 27% range is called the *confidence interval*. The probability of confidence interval is the confidence level (1−α), and the probability of not falling within the confidence interval is the *significance or alpha level* (α). Confidence levels of 90%, 95% and 99% (i.e. alpha levels of 10%, 5% and 1%) and margins of error of 1%, 3% and 5% are often desirable for sample size determination.

6.1. Using a formula

Table 3 provides an overview of common formulas that are used to calculate sample size (and their needed information) based on the type of survey variable. They are explained below.

Table 3: Formulas to determine sample size (Own work)

Variable	***Formula***	***Information needed***
Dichotomous	Equation 1	Population proportion
Polytomous	Equation 1 for a proportion closest to 50%	Population proportions
Quantitative	Equation 2 for scales with up to 13 points	Estimate of population standard deviation
Quantitative	Equation 3 for many number options	Estimate of population standard deviation
Mix	Equation of the primary variable	Varies with equation

6.1.1. Formula for dichotomous variables

Equation 1

$$n_0 = \frac{Z^2 pq}{e^2}$$

Where: *e* is the desired margin of error that is often 5%, *p* is proportion of the variable in the population, *q* is 1 – p. The z-value is found in a Z table based on the desired confidence level. For confidence levels of 95% and 99%, Z values are 1.96 and 2.58, respectively (Bartlett *et al*., 2001; Israel, 2003).

Exercise example: A researcher wants to evaluate effectiveness of a provincial agriculture program which promotes a new irrigation technology. Expert opinion holds that 35% of the farmers have embraced the new technology. The researcher is willing to accept a 95% confidence level and 5% margin of error for the survey. Sample size is calculated as follows:

$n_0 = 1.96^2 \times 35\% \times 65\% \div 5\%^2 =$ **350** farmers

- If the population has a known size and n>5% N, sample size is reduced as follows:

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

Where: n_0 is the original sample size, and N is size of the target population (Bartlett *et al*., 2001).

6.1.2. Formula for polytomous variables

- If population proportions are known, sample size is calculated using Equation 1 for a population proportion that is closest to 50%.

- If population proportions are unknown, the researcher should include 50% for population proportion in Equation 1.

- Where survey response options are less or more general than categories of population proportions, the researcher can adapt survey response options to population proportions by combining or splitting them.

Exercise example: A researcher wants to conduct a race and ethnicity survey. Prior research holds that 55% of citizens are white, 25% are black, 15% are Asian, and 5% are from other backgrounds. The researcher is willing to accept a 95% confidence level and 5% margin of error for the survey. As p_1 (55%) is the population proportion closest to 50%, it is used for sample size calculation by Equation 1 as follows:

$n_0 = 1.96^2$ x **55%** x45% ÷ $5\%^2$ = **380** individuals

- If the population has a known size and n>5% N, sample size is reduced as mentioned earlier.

6.1.3. Formula for quantitative variables with up to a 13 point scale

Equation 2

$$n_0 = \frac{Z^2 (S)^2}{(n_p e)^2}$$ S=1, where n_p is even

$$s = \frac{n_p}{n_p - 1}$$, where n_p is odd

Where: *e* is the desired margin of error that is often 3%; Z value is 1.96 for confidence level of 95%; s is standard deviation in the sample as estimate of standard deviation in the population; n_p is number of points on the scale (Bartlett *et al*., 2001).

Exercise example: In a survey, housewives will be asked to rate their marriage happiness using a numerical scale from 1 to 5. The investigator is willing to accept a 95% confidence level and 3% margin of error for the survey. Sample size is calculated as follows:

s= 5 ÷ (5-1) = 1. 25

n_0= 1.96^2 x 1.25^2 ÷ $(5x3\%)^2$ = **267** housewives

- If the population has a known size and n>5% N, sample size is reduced as mentioned earlier.

- For scales of more than 13 points, Eq. 2 results in a small sample.

6.1.4. Formula for quantitative variables with many options

Equation 3

$$n_0 = \frac{Z^2 S^2}{e^2}$$

Where: *e* is the desired margin of error (often 3%) applied without percentage; Z value is 1.96 for a confidence level of 95%; s is standard deviation in the sample as estimate of standard deviation in the population. In this equation, s is obtained from a previous study or is calculated through a pilot sample. Based on guidance concerning how large a pilot sample should be by type of survey aim, the proverbial rule of thumb of n≥30 can be overall applied (Hertzog, 2008; Charan and Biswas, 2013).

Exercise example: In a survey of regional parks visitors, the researcher is willing to accept a 95% confidence level and 3% margin of error. They survey is interested in number of visits per year. The researcher undertakes a pilot study of 30 visitors and calculates a mean of 56 visits and standard deviation of 20 visits. Sample size for the survey is calculated as follows:

n_0= 1.96^2 x 20^2 ÷ 3^2 = **171** visitors

- If the population has a known size and n>5% N, sample size is reduced as mentioned earlier.

6.1.5. Formula for a mix of variables

Where a survey involves a mix of variables, sample size will be determined for the variable which plays the most important role in the study (Bartlett *et al*., 2001).

Example: In a survey of income, life satisfaction and marital status, if the researcher is more interested in income or hypothesizes that income has an important role in life satisfaction and marital status, Equation 3 should be used. If such a significant role is assumed for life satisfaction, Equation 2 should be used where life satisfaction is rated by numbers and Equation 1 should be used if life satisfaction is rated in qualitative terms.

6.2. Using a sample size calculator and table

Sample size calculators are designed for online use and as Excel files. Tables can also be used such as Table 4 which provides a general reference for sample size in many surveys.

Table 4: A general table of sample size (Adapted from Bartlett *et al*., 2001)

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

6.3. Using a sample size from a similar study

It is possible that some studies have the same objectives, research design and resources as some previous studies so that the new studies basically extend previous investigations to new contexts and sites. Such new studies can use the sample size of previous peer studies, but in so doing they will be relying on someone else correctly determining the sample size unless they review the procedures employed in previous studies (Israel, 2003).

7 Collect and validate data 4

7.1. Survey

After sample size is determined, the researcher begins to collect data with particular effort to avoid two major forms of data collection bias including *'non-response' bias* and *'response' bias*. Non-response bias occurs when there is a failure to collect data from sampled elements (Daniel, 2012). In some surveys, response rates of around 70% are considered to be adequate, but in some studies rates of between 95% and 100% are expected (Fink, 2003). Response bias occurs when one is able to collect data from sampled elements, but the data collected are inaccurate or inappropriate (Daniel, 2012). The following guidelines can be applied before and during data collection to minimize the occurrence of data collection bias.

- Use trained data collectors;

- Identify a larger *n* than you need, but pay attention to the costs;

- 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 survey eligibility criteria (Fink, 2003).

However, data collection bias will almost certainly occur to some degree. Also, sometimes there is error in sampling frame which is reflected in the data (Malhotra and Birks, 2007). As Figure 3 indicates data validation, *which involves exploring and addressing errors in the collected data*, involves two aspects in surveys as follows.

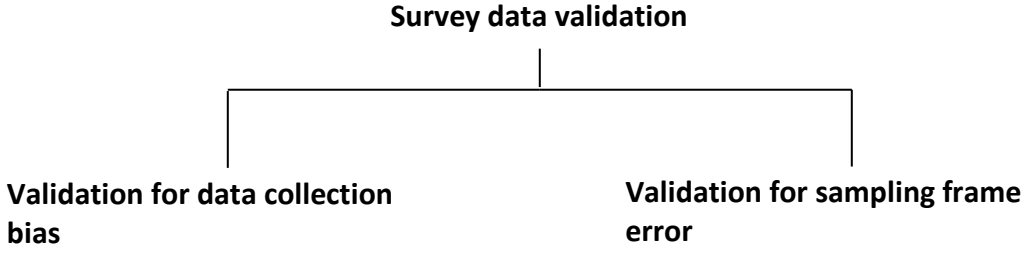

Figure 3: Survey data validation in two aspects (Own work)

7.1.1. Validation for data collection bias

(a) *Dropping the troublesome individual or element* from the sample: This will reduce sample size and is not recommended if one has a great deal of missing data; and

(b) *Conducting imputation*: It consists of three major types of mean imputation, dynamic imputation, and regression imputation. Mean imputation involves substituting the sample mean or the mean of a subgroup of the sample for the missing data. Dynamic imputation involves substituting the data collected from a similar individual or unit for the missing data. Regression imputation involves using regression analysis to estimate the missing data using variables highly correlated with the variable that has missing data as estimators (United Nations, 2010; Daniel, 2012).

7.1.2. Validation for sampling frame error

Comparisons between sample proportions and the target population proportions should be made if information about population proportions is accurate. Where it is found that these proportions do not match despite meeting survey requirements, this error is related to sampling frame that does not adequately represent the target population. Weighting is then applied to some individuals or elements to equalize sample proportions with population proportions (Malhotra and Birks, 2007). The weights are obtained through dividing population proportions by sample proportions.

7.2. Census

There are two broad kinds of census error. *Population under-coverage* refers to the exclusion of persons who should have been enumerated, and *population over-coverage* refers to the inclusion of persons who were enumerated more than once. Under-coverage can occur in the first stage of the census if the list of collection unit (e.g. dwellings, offices, parks) used is incomplete. Over-coverage can occur if a collection unit is listed twice or some individuals or elements are included in two collection units such as people with part-time residences or gypsies moving between parks. Coverage error can also occur in the data processing stage (Statistics Canada, 2016). As Figure 4 illustrates, census data validation involves two methods as explained below.

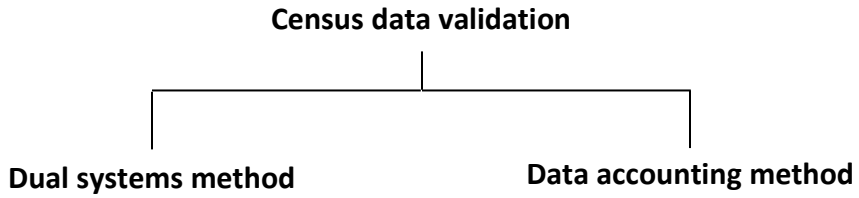

Figure 4: Census data validation methods (Own work).

7.2.1. Dual systems method

After the census has been completed, cluster sampling of geographical areas of the census is conducted resulting in a random selection of *n* clusters 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. The P-sample enumerations are then compared with the census enumerations in the P-sample referred to as the E-sample enumerations. This is intended not only to discover census coverage errors and identify alternative processes to prevent such errors in the future, but also to correct coverage errors and estimate true population count by Equation 4 as follows (National Research Council, 2007; United Nations, 2010):

Equation 4 $$DSE = (C - II)\left(\frac{CE}{E}\right)\left(\frac{P}{M}\right)$$

Where:

DSE: the dual systems estimate of true population count

C: census enumerations

II: the number of individuals lacking sufficient information for matching

CE: correct E-sample enumerations (In Figure 7, CE=4)

E: E-sample enumerations (In Figure 7, E=6)

P: P-sample enumerations (In Figure 7, P=7)

M: the number of P-sample individuals matching E-sample individuals. As Figure 5 exemplifies, M=CE.

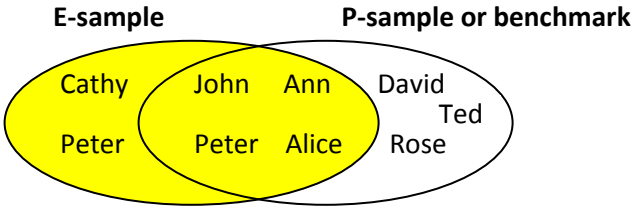

Figure 5: An example of E-sample and P-sample or benchmark (Own work)

Exercise example: A national census guided by a directory of homeless centers enumerates **80, 000** homeless people. It is estimated that 5% of the census enumeration relates to individuals who provide no precise clues for inclusion or exclusion in the homeless population. Therefore, C= 80,000, and II=80,000 x 5%= 4000.

We choose randomly three provinces as the P-sample in which we make a list of homeless centers and unregistered homeless shelters (e.g. public parking, underpass) to act as

7.2.2. Data accounting method

This method needs enumerations from a previous census as well as accurate data about additions to, and reductions from, the basic enumerations until the new census date to construct population estimates for that date. This data accounting provides consistent and logical bounds for census enumerations. For a demographic census, for example, if census enumerations give results far outside the bounds implied by Equation 5 accounting, the departures need to be field inspected and corrected (e.g. by the dual systems method) (National Research Council, 2007; United Nations, 2010).

Equation 5 $P_1 = P_0 + B - D + I - E$

Where:

P_1 = population at the census date;

P_0 = population at a previous census date;

B = births between the two census dates;

D = deaths between the two census dates;

I = immigration between the two census dates; and

E = emigration between the two census dates.

Conclusions

Sampling and census should be understood within an integrated framework of data collection and validation. This, along with worked examples, multi-disciplinary exemplifications and must-knows in each step of the framework were presented in this paper to help researchers meet three principles of scientific investigation as follows: (a) to make informed decisions about available methodological and statistical choices in relation to study objectives, subjects and resources; (b) to optimize data collection and processing with prior insight about mechanisms of error; and (c)

to correct errors in the data and be reflexive about any remaining issues in making inferences about the target population.

References

Australian Bureau of Statistics (2019a) Samples and censuses. Australian Bureau of Statistics.

Australian Bureau of Statistics (2019b) Sample design. Australian Bureau of Statistics.

Bartlett, J.E., Kotrlik, J.W. and Higgins, C.C. (2001) Organizational research: Determining appropriate sample size in survey research, *Information Technology, Learning, and Performance Journal*, 19(1), pp.43-50.

Bashir, Y.M. (2016) Effects of treasury single account on public finance management in Nigeria, *Research Journal of Finance and Accounting*, 7(6), pp. 164-170.

Browne, K. (2005) Snowball sampling: using social networks to research non-heterosexual women, *International Journal of Social Research Methodology*, 8(1), pp.47-60.

Charan, J. and Biswas, T. (2013) How to calculate sample size for different study designs in medical research?, *Indian Journal of Psychological Medicine*, 35(2), pp.121-126.

Chauvet, G. (2015) Coupling methods for multistage sampling, *The Annals of Statistics*, 43(6), pp.2484-2506.

Da Costa, N.M., Hepp, K. and Martin, K.A. (2009) A systematic random sampling scheme optimized to detect the proportion of rare synapses in the neuropil, *Journal of Neuroscience Methods*, 180(1), pp.77-81.

Daniel, J. (2012) *Sampling Essentials: Practical Guidelines for Making Sampling Choices*. Thousand Oaks: Sage Publications.

Dattalo, P. (2008) *Determining Sample Size: Balancing Power, Precision, and Practicality*. New York: Oxford University Press.

Fink, A. (2003) How to Sample in Surveys. Second edition. Thousand Oaks: Sage Publications.

Gill, J., Johnson, P. and Clark, M. (2010) *Research Methods for Managers*. Fourth edition. London: Sage Publications.

Gundlach, E. and Ward, M.D. (2020) The data mine: Enabling data science across the curriculum, *Journal of Statistics Education*. Online First.

Hao, Y. (2009) Poverty and exclusion in urban China. WZB Discussion Paper. WZB.

Henderson, R.H. and Sundaresan, T. (1982) Cluster sampling to assess immunization coverage: a review of experience with a simplified sampling method, *Bulletin of the World Health Organization*, 60(2), pp. 253-260.

Hertzog, M.A. (2008) Considerations in determining sample size for pilot studies, *Research in Nursing & Health*, 31(2), pp.180-191.

Hussain, A. (2003) *Urban Poverty in China: Measurements, Patterns and Policies*. Geneva: International Labour Office.

Israel, G.D. (2003) Determining sample size. Working Paper PEOD6. The Agricultural Education and Communication Department, Florida Cooperative Extension Service, Institute of Food and Agricultural Sciences, University of Florida.

Lempp, H.K. (2005) Perceptions of dissection by students in one medical school: beyond learning about anatomy. A qualitative study, *Medical Education*, 39(3), pp.318-325.

Levy, P.S. and Lemeshow, S. (2008) *Sampling of Populations: Methods and Applications*. Fourth edition. Hoboken: John Wiley & Sons.

Malhotra, N.K. and Birks, D.F. (2007) *Marketing Research: An Applied Approach*. Third European edition. Harlow: Pearson Eduation.

Menard, S. (2004) Polytomous variable. In: Lewis-Beck, M., Bryman, A.E. and Liao, T.F. (Eds.), *The Sage Encyclopedia of Social Science Research Methods*. Thousand Oaks: Sage Publications, p.834.

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

Sampath, S. (2001) *Sampling Theory and Methods*. Chennai: CRC Press.

Schellentrager, G.W. and Doolittle, J.A. (1991) Using systematic sampling to study regional variation of a soil map unit, *Spatial Variabilities of Soils and Landforms*, 28, pp.199-212.

Smith, W., Mitchell, P., Attebo, K. and Leeder, S. (1997) Selection bias from sampling frames: Telephone directory and electoral roll compared with door-to-door population census: Results from the blue mountains eye study, *Australian and New Zealand Journal of Public Health*, 21(2), pp.127-133.

Speak, A., Escobedo, F.J., Russo, A. and Zerbe, S. (2018) Comparing convenience and probability sampling for urban ecology applications, *Journal of Applied Ecology*, 55(5), pp.2332-2342.

Statistics Canada (2016) Population coverage error. Statistics Canada.

Taherdoost, H. (2016) Sampling methods in research methodology; how to choose a sampling technique for research, *International Journal of Academic Research in Management*.

Thompson, S.K. (1990) Adaptive cluster sampling, *Journal of the American Statistical Association*, 85(412), pp.1050-1059.

Tso, G.K. and Yau, K.K. (2003) A study of domestic energy usage patterns in Hong Kong, *Energy*, 28(15), pp.1671-1682.

United Nations (1982) *National Household Survey Capability Programme; Non-sampling Errors in Household Surveys: Sources, Assessment and Control*. New York: United Nations.

United Nations (2010) *Handbook on Population and Housing Census Editing*. New York: United Nations.

Yates, F. (1960) *Sampling Methods for Censuses and Surveys*. Third edition. London: Charles Griffin and Company Limited.

Zhang, B., Mildenberger, M., Howe, P.D., Marlon, J., Rosenthal, S.A. and Leiserowitz, A. (2020) Quota sampling using Facebook advertisements, *Political Science Research and Methods*, 8(3), pp.558-564.

Zhengdong, L.I. (2011) Error analysis of sampling frame in sample survey, *Studies in Sociology of Science*, 2(1), pp.14-21.