# Figure 8.1. SURVEY SAMPLING: An Introduction

A **sample survey** involves making inferences about a population on the basis of information obtained from a sample (*i.e.*, a subset) selected from the population; inherent in this description of *sampling* are the processes of *defining* the target population appropriately for the Question(s) to be answered, *specifying* the study population (*e.g.*, by a suitable **frame**), *selecting* units (a unit is one or more elements) and *estimating* attributes, which entails *measuring* response (and perhaps explanatory) variate(s) for each unit selected. The alternative to a sample survey is a **census**, in which *every* element of the population is investigated.

**Sample survey:**
Defining
Specifying
*Selecting*
Responding
Measuring
*Estimating*

* **Target population:** the group of elements to which the investigator(s) want Answer(s) to the Question(s) to apply.
* **Study population:** a group of elements *available* to an investigation.
* **Respondent population:** those elements of the study population that *would* provide the data requested under the incentives for response offered in the investigation;
* **Non-respondent population:** those elements of the study population that would *not* provide the data requested under the incentives for response offered in the investigation.

**Census:**
Defining
Specifying
Responding
Measuring

There is further discussion of the respondent and non-respondent populations in Section 3 overleaf on pages 8.4 and 8.5.

The choice between a sample survey and a census is usually a trade-off between their advantages and disadvantages. The *advantages* of sample surveys are that they:

+ use fewer *resources* (money and/or time);
+ are usually shorter in duration and so provide more *timely* Answer(s) to Question(s);
+ are the only feasible option when measuring is *destructive*; *e.g.*, crash tests on cars, test firing of camera flash bulbs and gun cartridges, measuring cigarette tar and nicotine levels and bursting pressures of plastic bags and condoms.

The main *dis*advantages of sample surveys are that:

– there is inherent *uncertainty* in using a sample attribute as a basis for *estimating* a population attribute;
– they usually do not provide information about *every* segment (or subgroup) of the population being investigated.

## 1. Terminology

*We* distinguish an *element* from a *unit*.

- **Elements** are the entities that make up a population; for example, a person is an element of the population of Canadians, but we recognize that many populations in data-based investigating have non-human or *in*animate elements.
- **Units** are the entities *selected* for the sample; a unit may be one element (*e.g.*, a person) or *more than* one (*e.g.*, a household).

In multi-stage selecting (discussed below), we speak of *primary* selecting units (PSUs) selected at the *first* stage, *secondary* selecting units (SSUs) selected at the *second* stage, and so on.

Major components of the Plan for a sample survey include the four matters summarized at the right; details are as follows.

**Population** – Unstratified / Stratified
**Number of stages** – One-stage / Multi-stage
**Last stage units** – Elements / Clusters
**Selecting method** – Probability / Other

- **Stratifying** is a process of subdividing a population into groups called **strata** before the units are selected; strata must be defined in such a way that every population unit belongs to *one and only one* stratum. Units for the sample are selected from *each* of the strata.
- Units can be selected from the population in one or more **stages**, the level of aggregation of the units *de*creasing through successive stages; for Canada, for example, a *five*-stage Plan could involve selecting *provinces* within the country, *counties* within the provinces selected, *urban and rural areas* within these counties, *households* within the selected areas, and a *person* within each of these households.
  – A sample survey to obtain *household* data would involve only the first *four* of these stages.
- The *last* stage of a multi-stage sample survey (or a *one*-stage survey) can select units which are either *individual* population elements or *groups* of elements (called *clusters*); obviously, stages *before* the last can *only* select *groups* of elements.
- A variety of processes is used to *select* units at each stage; the basis of the statistical theory for estimating population attributes from sample data is *probability selecting*, in which each population unit has a *known* probability of being selected for the sample.

## 2. Selecting

We distinguish the following processes for *selecting* units for sample surveys:

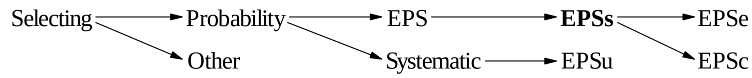
- probability selecting;
- accessibility selecting;
- convenience selecting;
- haphazard selecting;
- quota selecting;
- volunteer selecting;
- judgement selecting.

In *probability selecting* (also called 'probability sampling'), every unit of the study population has a *known* probability of being selected. Sample survey Plans are often developed so that, as far as practicable, every unit of the population has (approximately) the *same* selection probability; this is the case for *equiprobable selecting* (EPS) [but see Note 2 overleaf on page 8.4]. We need to recognize *which entities* are equiprobable in EPS – all possible *samples of a given size* (EPSs) or all population *units*



(*continued overleaf*)

(EPSu) or all population *elements* (EPSe) or all population *clusters* (EPSc); the schema below summarizes these possibilities. The meaning of EPS in *these* Course Materials [and, elsewhere, of 'simple random selecting (SRS)'] is shown in **bold** type in the schema.




**NOTES:** 1. The foregoing discussion of EPS is based on equiprobable selecting with*out* replacement (EPSWOR); survey sampling and these Course Materials have little concern with selecting *with* replacement (EPSWIR). *We* usually omit *WOR* – its inclusion in some titles (*e.g.*, in STAT 220 Figure 8.11 and STAT 332 Figure 2.10) is an occasional reminder.

2. For a respondent population frame of four units a, b, c, d, six samples of size 2 (as listed at the right) can be selected by EPS but only *two* by systematic selecting (SYS); these two samples are *equally* probable if the selecting starting point is chosen equiprobably. Thus, under *both* selecting processes, each unit has an inclusion probability of one-half. This reminds us that EPS implies equal *unit* inclusion probabilities but the converse is *not* true – here, systematic selecting with *equal* inclusion probabilities is *not* EPS because it cannot select four of the samples possible under EPS.

| **EPS** | **SYS** |
|---|---|
| ab | ac |
| ac | bd |
| ad | |
| bc | |
| bd | |
| cd | |

3. A phrase commonly used in texts dealing with survey sampling is *simple random selecting* [or, regrettably, *simple random* ***sampling***] (SRS); in the terminology of these Course Materials, it is equivalent to *EPS from an unstratified population. We* use this longer phrase because it is more evocative of the actual selecting process.

4. If there are **N** units in the respondent population and there are n units in the sample, then: $f = \frac{n}{\mathbf{N}}$ is called the **sampling fraction.** -----(8.1.1)
[ $f$ is the probability any respondent population unit is included in a sample obtained by EPS.]

Probability selecting has two advantages.

- It is the basis of statistical theory for survey sampling; *e.g.*, it is the basis of expressions for confidence intervals for respondent population attributes.
- For estimating attributes like averages, totals, proportions and frequencies, *under repetition* it eliminates sampling *inaccuracy*.
  - Zero sampling inaccuracy under repetition has limited implication for the magnitude of sample error – it only marginally reduces the risk of sample error large enough to impose unacceptable limitation on Answer(s).
  - If a sample, *after* it has been selected, is found to have sample error that imposes unacceptable limitation on an Answer, it is usually impossible to determine, from an examination of the *sample itself*, the extent to which the sample error is due to a flawed selecting process or to chance; instead, it is necessary to know the *process* by which the sample was selected.

Related matters are:

- The statistical theory of survey sampling usually deals *only* with the uncertainty that arises because of the sample data are inherently *incomplete*; it seldom takes account of *other* sources of uncertainty (sometimes called *non-sample errors* – see Figure 8.5 of these Course Materials) such as study error, non-response error, data processing mistakes, etc. (See also Note 9 at the bottom of the facing page 8.5 and recall Section 7 on pages 6.6 and 6.7 in Figure 6.1 of these Materials.)
- *Non*-probability selecting processes can yield Answers with acceptable limitations in some investigations, but the results of (probability) selecting theory (*e.g.*, to obtain confidence intervals for respondent population attributes) do *not* apply to them.
  - A common *misuse* of statistical methods is to give confidence intervals based on sample survey data obtained using *any* of the six *non*-probability selecting processes listed overleaf in Section 2 near the bottom of page 8.3.
  - The experience of sample survey statisticians over several decades provides compelling evidence that, when human judgement plays a significant role in sample selecting (as is often the case in non-probability processes), sample error is more likely to impose *un*acceptable limitation on Answers, particularly for Questions with a *descriptive* aspect.
  - The names of the six *non*-probability selecting processes listed overleaf on page 8.3 do not necessarily specify a *unique* selecting method – the first two methods overlap and the first five involve some degree of 'accessibility' and/or 'convenience'.

    Haphazard selecting is sometimes ***wrongly*** equated with 'random' selecting; *i.e.*, with our *equiprobable* selecting.

    Quota selecting is a similar idea to **covering** – see the top of page 5.24 in Figure 5.7 of the STAT 231 Course Materials.

    Volunteer selecting is *not* to be confused with **volunteer** (or **voluntary**) **response**, a phrase sometimes used to indicate that *human* elements can (usually) *choose* whether to respond, *i.e.*, whether to provide the requested data; a separate (measuring) issue is whether these responses are correct or truthful (see Note 68 on page 5.62 in STAT 231 Figure 5.7).

## 3. Responding

Many data-based investigations encounter the difficulty that not all the data called for in the Plan are acquired – this is the general problem of *missing data*. For *human* populations, this matter is usually referred to as *non-response* and it can impose *un*acceptable limitations on Answer(s). A framework for discussing non-response is shown in the diagram at the upper right of the facing page 8.5, where the *study* population (represented by the outer area) is made up of the *respondent* population and the *non-respondent* population. The set of units selected from the study population is the **selection**, and comprises the **sample** (from the respondent population) and the **non-respondents** (from the *non*-respondent population). The diagram has *two* cate-

## Figure 8.1. SURVEY SAMPLING: An Introduction (continued)

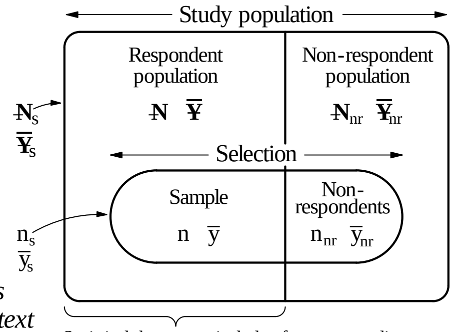
gories of symbols:

- **N**s and ns refer to *numbers of (elements or) units*;
- $\bar{\mathbf{Y}}$s and $\bar{y}$s are *averages* of a response variate **Y** of the (elements or) units.

The relationships among the numbers of (elements or) units are:

Study population = Respondent population + Non-respondent population
$\mathbf{N}_S = \mathbf{N} + \mathbf{N}_{nr}$

Selection = Sample + Non-respondents
$n_S = n + n_{nr}$




Statistical theory, particularly of survey sampling, is developed mainly in the context of the *respondent* population, often with*out* recognizing it explicitly.

**NOTES:** 5. In these Course Materials, use of '*sample*' to refer only to *respondents* is *seldom* followed elsewhere; it can be confusing when only the *context* indicates whether 'sample' means *selection* or *sample* of our terminology.

- Phrases like *intended sample* and *achieved sample* are also used in some places to convey the same distinction as our *selection* and *sample*.

6. The symbols **N** (for the number of respondent population elements or units) and n (for the number of units in the sample) could usefully be written as $\mathbf{N}_R$ and $n_r$ as a reminder that they refer to *respondents*. However, the *un*subscripted symbols **N** and n are in such widespread use that it is unrealistic to expect to change them; in addition, there would be the inconvenience of a subscript on two frequently-used symbols.

7. A major Plan consideration in a sample survey of a *human* population is how to offer (as specified in the *sampling protocol*) adequate *incentives for response* (*e.g.*, proper presentation of the survey, suitable follow-up, rewards for response); the goal of the incentives is to achieve as *high* a response *rate* as possible. An illustration of why a high response rate is important is provided in Table 8.1.1 below.

- The first column of Table 8.1.1 at the right lists seven levels of non-response for a survey question with a *Yes* or *No* answer. The entries in the second and third columns of the Table show *intervals* of possible percentages of *Yes* answers among *all* the units *selected*. The second column is based on the assumption that, in the *sample*, half the answers are *Yes* and half are *No*, regardless of the non-response rate; for the third column, it is assumed that 90% of the *sample* answer *Yes* and 10% answer *No*, regardless of the non-response rate.

**Table 8.1.1**

| NON-RESPONSE RATE | PERCENTAGE OF *Yes* ANSWERS AMONG ALL UNITS SELECTED: 50% *Yes* in Sample | 90% *Yes* in Sample |
|---|---|---|
| 0% | 50% | 90% |
| 5 | 47.5 – 52.5 | 85.5 – 90.5 |
| 10 | 45 – 55 | 81 – 91 |
| 25 | 37.5 – 62.5 | 67.5 – 92.5 |
| 50 | 25 – 75 | 45 – 95 |
| 75 | 12.5 – 87.5 | 22.5 – 97.5 |
| 90 | 5 – 95 | 9 – 99 |

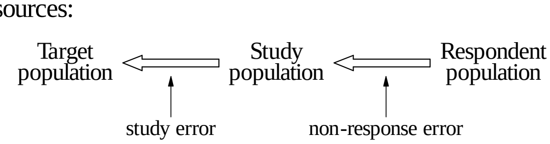
We see that every percentage point of non-response adds a percentage point to the range (*i.e.*, to the uncertainty) for the proportion of possible *Yes* answers in the *selection*. Also, for 90% *Yes* in the sample, the *centres* of the ranges differ progressively more from 90% as we go down Table 8.1.1; this reflects *non-responding inaccuracy* arising from a situation where *Yes* and *No* are not equally likely.

8. Two matters about the separation of the study population into the respondent and *non*-respondent populations are:

- Which elements of the study population fall in which of the two populations depends on the *incentives* offered for response – different incentives would presumably, in general, result in *different* sets of the study population elements in the two populations. For *given* incentives for response (as specified in the sampling protocol of a *particular* survey), *we* usually assume that a given element will *always* make the *same* decision about whether or not to respond; this *deterministic* view of the response decision is why *we* usually limit discussion to non-response *error*, rather than going on to non-responding *inaccuracy*.
- The respondent and non-respondent populations are *conceptual* in the sense that we only *encounter* subsets of them (as the sample and the non-respondents); if a unit is *not* included in the selection, we generally do *not* know (and do not *need* to know) to which of the two populations it belongs.

9. In addition to *sample* error and *measurement* error when estimating *respondent* population attributes, the inference *back* from these estimates to plausible values for *study* population attributes and, hence, for those of the *target* population is subject to the effects of error from two *other* sources:

- *non-response* error (due to differing attribute values in the respondent and study populations);
- *study* error (due to differing attribute values in the study and target populations);




the schema at the right above shows this matter pictorially. All *four* error categories impose limitations on Answers.



(*continued overleaf*)

## 4. Measuring

It is convenient to distinguish *four* components of a *measuring process*:

- the measuring instrument or gauge; • the operator; • the measuring protocol; • the element (or unit) measured.

In the context of a sample survey, the measuring *instrument* is the questionnaire. It is curious that, when sample surveys are carried out, investigators who would *never* undertake to assemble the types of instrumentation typically used in a laboratory (*e.g.*, balances, spectrophotometers) approach the task of developing a *questionnaire* (*i.e.*, a *sample survey* measuring instrument) with what seems to be little or no recognition of either the difficulty or the importance of doing so successfully.

## 5. Estimating

In *our* discussion of survey sampling, we are concerned with estimating four types of attributes:

- averages; • totals; • proportions; • frequencies.

For averages and proportions, the corresponding sample attribute (or 'statistic') is generally used to estimate the respondent population attribute, but the sample total and sample frequency do *not* estimate the respondent population total or frequency.

- ○ We speak of *estimating* from the data obtained from a *sample*, but of *imputing* to deal with *missing* data for *non-respondents*.
  - – An imputed value for a non-response is (hopefully) a 'best guess' of the missing response variate data based, for example, on known values of relevant *explanatory* variate(s) of the unit concerned; imputed values are included to facilitate data analysis. Rarely does imputing meaningfully reduce the severity of limitations on Answers by reducing incompleteness of data.
- ○ The term 'statistic', beloved of introductory texts, is *avoided* in these Material; instead, *we* use 'estimator' or 'estimate' as appropriate. For us, 'population parameter' and 'sample statistic' are 'population attribute' and 'sample attribute' – *models* have parameters.

## 6. 'Good' Sample Surveys

A summary of Plan requirements for a good sample survey (*i.e.*, one with *acceptable* limitations on Answers) is as follows.

- ⊙ A well-presented questionnaire, with unambiguous and answerable questions, which is properly administered and which gathers the information needed to address properly the Question(s) to be answered.
- ⊙ A sample of adequate size, obtained by a probability selecting method.
- ⊙ A high response (or low *non*-response) rate.
- ⊙ Correct use of statistical theory, together with accurate data processing and clear and complete presentation of the Answer(s).

Deficiencies in any *one* of these four Plan components, *even if all other components are adequate*, can impose unacceptable limitations on the Answer(s) obtained from the survey.

A *report* of a sample survey should, as a *minimum* requirement, provide the reader with enough information about the Plan to make clear the following *four* matters:

- ∗ What the target and the study populations were.
- ∗ What the non-response rate was.
- ∗ What process was used to select the units.
- ∗ What the sample size was.

The first question is concerned with *study* error, the second with *non-response* error and the third and fourth with *sample* error. The importance of *sample size* is that, *for fixed data quality* and assuming the same information content per observation, the *larger* the amount of data the *more* precise the Answer(s) the data can yield. However, it is almost always (extremely) *difficult* to maintain high quality in (very) large data sets, one reason why a sample survey is usually preferred over a census.

Assessment of *measurement* error requires, as a minimum, a copy of the questionnaire (*e.g.*, in an appendix of the report).

**NOTE:** 10. Despite a central concern with survey sampling, these Course Materials seldom uses the word **sampling** [except in the (established) phrases *survey sampling*, *sampling protocol*, *sampling inaccuracy*, *sampling imprecision* and *sampling fraction*] for three reasons:

- the general use of the word encompasses the four processes of *selecting*, *responding*, *measuring* and *estimating*, which are best *explicitly* kept as *separate* entities and considered *individually*;
- the specific use is basically a synonym for the process of *selecting*, a word whose clarity can be lost when it is instead called 'sampling';
- some uses of the word could be taken as implying that sampling (in the sense of reasoning from data derived from a *subset* of some population) is primarily the domain of sample surveys [*i.e.*, investigations with an observational Plan to answer Question(s) with a descriptive aspect] whereas, in reality, essentially *all* types of investigations [including those with observational or experimental Plans to answer Question(s) with a descriptive or causative aspect] take only a *sample* – that is, a census is *rare* in data-based investigating.

In other sources of information about survey sampling, you are likely to encounter the word *sampling*; when you do, you should identify the sense in which it is being used.