# PRINCIPLES OF DATA COLLECTION

Patrick Boily[1,2,3]

**Abstract**

Data analysis tools and techniques work in conjunction with collected data. The type of data that needs to be collected to carry out such analyses, as well as the priority placed on the collection of quality data relative to other demands, dictate the choice of data collection strategies. The manner in which the resulting outputs of these analyses are used for decision support will, in turn, influence appropriate data presentation strategies and system functionality. In this report, we present a brief overview of sampling methods, automated data collection, and web scraping.

**Keywords**

Data collection, web scraping, automatic data collection, sampling methods, questionnaire design.

**Funding Acknowledgement**

Parts of this report were funded by a University of Ottawa grant to develop teaching material in French (2019-2020). These were subsequently translated into English before being incorporated into this document.

[1]Department of Mathematics and Statistics, University of Ottawa, Ottawa
[2]Data Action Lab, Ottawa
[3]Idlewyld Analytics and Consulting Services, Wakefield, Canada
**Email**: pboily@uottawa.ca

## Contents



## 1. Introduction

> Fisher's Maxim
>
> To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of.
>
> – R.A. Fisher, Presidential Address to the *First Indian Statistical Congress*, 1938

Data analysis tools and techniques work in conjunction with collected data. The type of data that needs to be collected to carry out such analyses, as well as the priority placed on the collection of quality data relative to other demands, will dictate the choice of data collection strategies. The manner in which the resulting outputs of these analyses are used for decision support will, in turn, influence appropriate data presentation strategies and system functionality, which is an important access of the analytical process.

Although analysts should always endeavour to work with **representative** and **unbiased data**, there will be times when the available data is flawed and not easily repaired. Analysts have a professional responsibility to explore the data, looking for potential fatal flaws **prior** to the start of

the analysis and to inform their client and stakeholders of any findings that could halt, skew, or simply hinder the analytical process or its applicability to the situation at hand.

Unless a clause has specifically been put in the contract to allow a graceful exit at this point, you will have to proceed with the analysis, flaws and all. It is **EXTREMELY IMPORTANT** that you do not simply sweep these flaws under the carpet. Address them repeatedly in your meetings with the clients, and make sure that the analysis results you present or report on include an appropriate *caveat*.

### 1.1 Data Collection System

Analysts might also be called upon to provide suggestions to evaluate or fix the data collection system. The following items could help with that task.

- **Data Validity**: the system must collect the data in such a way that data validity is ensured during initial collection. In particular, data must be collected in a way that ensures sufficient accuracy and precision of the data, relative to its intended use.
- **Data Granularity, Scale of Data**: the system must collect the data at a level of granularity appropriate for future analysis.
- **Data Coverage**: the system must collect data that comprehensively, rather than only partially or unevenly, represents the objects of interest. As well, the system must collect and store the required data over a sufficient amount of time, and at the required intervals, to support data analyses that require data spanning a certain duration.
- **Data Storage**: the system must have the functionality to store the types and amount of data required for a particular analysis.
- **Data Accessibility**: the system must provide access to the data relevant for a particular analysis, in a format that is appropriate for this analysis.
- **Computational/Analytic Functionality**: the system must have the ability to carry out the computations required by relevant data analysis techniques.
- **Reporting, Dashboard, Visualization**: the system must be able to present the results of the data analysis in a meaningful, usable and responsive fashion.

A number of different overarching strategies for data collection can be employed. Each of these different strategies will be more or less appropriate under certain data collection circumstances, and will result in different system functional requirements. In this report, we will focus on survey sampling, questionnaire design, and automated data collection.

### 1.2 Formulating the Problem

The **objectives** drive all other aspects of quantitative analysis. With a **question** (or questions) in mind, an investigator can start the process that leads to **model selection**. With potential models in tow, the next step is to consider what **variates** (fields, variables) are needed, the **number** of observations required to achieve a pre-determined **precision**, and how to best go about **collecting**, **storing** and **accessing** the data.

Another important aspect of the problem is to determine whether the questions are being asked of the data in and of **itself**, or whether the data is used as a **stand-in for a larger population**. In the later case, there are other technical issues to incorporate into the analysis in order to be able to obtain generalizable results.

Questions do more than just drive the other aspects of data analysis – they also drive the development of quantitative methods. They come in all flavours and their variability and breadth make attempts to answer them challenging: no single approach can work for all of them, or even for a majority of them, which leads to the discovery of better methods, which are in turn applicable to new situations, and so on, and so on.

Not every question is answerable, of course, but a large proportion of them may be answerable partially or completely; quantitative methods can provide insights, estimates and ranges for possible answers, and they can point the way towards possible implementations of the solutions.

As an illustration, consider the following questions:

- Is cancer incidence higher for second-hand smokers than it is for smoke-free individuals?
- Using past fatal collision data and economic indicators, can we predict future fatal collision rates given a specific national unemployment rate?
- What effect would moving a central office to a new location have on average employee commuting time?
- Is a clinical agent effective in the treatment against acne?
- Can we predict when border-crossing traffic is likely to be higher than usual, in order to appropriately schedule staff rotations?
- Can personalized offers be provided to past clients to increase the likelihood of them becoming repeat customers?
- Has employee productivity increased since the company introduced mandatory language training?
- Is there a link between early marijuana use and heavy drug use later in life?
- How do selfies from over the world differ in everything from mood to mouth gape to head tilt?

How can such questions be answered? In many instances, the next step requires obtaining relevant data.

### 1.3 Data Types

Data has **attributes** and **properties**. Fields are classified as **response**, **auxiliary**, **demographic** or **classification** variables; they can be **quantitative** or **qualitative**; **categorical**, **ordinal** or **continuous**; **text-based** or **numerical**. Furthermore, data is **collected** through experiments, interviews, censuses, surveys, sensors, scraped from the Internet, etc.

Collection methods are not always sophisticated, but new technologies usually improves the process in many ways (while introducing new issues and challenges): modern data collection can occur over one pass, in batches, or continuously.

How does one decide which data collection method to use? The type of question to answer obviously has an effect, as do the required precision, cost and timeliness. Statistics Canada's *Survey Methods and Practices* [6] provides a wealth of information on probabilistic sampling and questionnaire design, which remain relevant in this day of big (and real-time) data.

The importance of this step cannot be overstated: without a well-designed plan to collect meaningful data, and without safeguards to identify flaws (and possible fixes) as the data comes in, subsequent steps are likely to prove a waste of time and resources.

As an illustration of the potential effect that data collection can have on the final analysis results, contrast the two following ways to collect similar data.

**Yes. I Mean No. ... I Think.**

The Government of Québec has made public its proposal to negotiate a new agreement with the rest of Canada, based on the equality of nations; this agreement would enable Québec to acquire the exclusive power to make its laws, levy its taxes and establish relations abroad — in other words, sovereignty — and at the same time to maintain with Canada an economic association including a common currency; any change in political status resulting from these negotiations will only be implemented with popular approval through another referendum; on these terms, do you give the Government of Québec the mandate to negotiate the proposed agreement between Québec and Canada?

– 1980 Québec sovereignty referendum question

**Do You Think They Learned Something From 1980?**

Should Scotland be an independent country?

– 2014 Scotland independence referendum question

The end result was the same in both instances, but an argument can be made that the 2014 Scottish 'No' was a much clearer 'No' than the Québec 'No' of 34 years earlier – in spite of the smaller 2014 victory margin (55.3%-44.7%, as opposed to 59.6%-40.4%).

### 1.4 Data Storage and Access

Data **storage** is also strongly linked with the data collection process, in which decisions need to be made to reflect how the data is being collected (one pass, batch, continuously), the volume of data that is being collected, and the type of access and processing that will be required (how fast, how much, by whom).

Stored data may go **stale** (e.g. people move, addresses no longer accurate), so it may be necessary to implement regular updating collection procedures.

Until very recently, the story of data analysis has been written for small datasets: useful collection techniques yielded data that could, for the most part, be stored on personal computers or on small servers. The advent of Big Data has introduced new challenges *vis-à-vis* the collection, capture, access, storage, analysis and visualisation of datasets; some effective solutions have been proposed and implemented, and intriguing new approaches are on the way (such as DNA storing [21], to name but one). We shall not discuss those challenges in detail, but be aware of their existence.

## 2. Questionnaire Design

**A Modern Paradox**

People resist a census, but give them a profile page and they'll spend all day telling you who they are.

– Max Berry, *Lexicon*, 2013

A **questionnaire** is a sequence of questions designed to obtain information on a subject from a respondent. Design principles vary according to the subject matter and the mode of data collection, but we strongly encourage pre-testing a variety of questionnaires.

### 2.1 Questionnaire Design Basics

In general, questionnaires should:

- be as brief as possible, with no wasted questions;
- be accompanied by clear and concise instructions;
- keep the respondent's interests in mind;
- emphasise confidentiality;
- be serious and courteous in tone;
- be free of mistakes and laid out attractively;
- be worded clearly;
- be designed to be accurately answered, and
- be ordered attentively.

### 2.2 Question Types

The basic questionnaire unit is, of course, the **question**, which comes in two flavours:

- **closed**, with a fixed number of pre-determined mutually exclusive and collectively exhaustive answer choices (and should always include an "Other (Please Specify)" category to counteract the loss of expressiveness of such questions), and
- **open**, which serves to identify common response choices to be used as closed question choices in subsequent questionnaires.

### 2.3 Wording Considerations

It is well known that the wording of the questions can influence a questionnaire's responses [5]; please keep the following **wording considerations** in mind when designing a questionnaire:

- avoid **abbreviations** and **jargon** ("Does your organization use any TTWQ practices?");
- do not use words and terminology that are **too complex** ("How often have you been defenestrated?" vs. "How often have you been thrown out of a window?");
- specify the **frame of reference** ("What is your income?" vs. "What was your household's total income form all sources before taxes and deductions in 2017?");
- make the question as **specific** as possible ("How much fuel did your moving company use during the last year?" vs. "How much did your moving company spend on fuel during the last year?");
- ensure that the questions can be answered by all respondents;
- avoid **double-barrelled** questions ("Do you plan to leave your car at home and take the bus to work during the coming year?" vs. "Do you plan to leave your car at home in the coming year? If so, do you plan to take the bus to work?"), and
- avoid leading questions (see the always excellent [7] for a not-so-facetious example).

### 2.4 Question Order

The order of the questions is just as important as the wording. Questionnaires should be designed to **flow smoothly** and **follow a logical sequence** (logical to the respondent, that is):

1. start with an **introduction** which provides the title, subject, and purpose of the survey;
2. request **cooperation** and explain the importance of the survey and how the results will be used;
3. indicate the degree of **confidentiality** and provide a deadline and a contact address;
4. open with a series of **easy** and **interesting questions** to establish the respondent's confidence;
5. group similar questions under a **common heading**;
6. introduce **sensitive topics** once trust and confidence are likely to have developed;
7. allow some space and/or time for **additional comments**, and
8. **thank** the respondent for their participation.

A lot more has been written about questionnaire design (see [3], for instance). It can be surprisingly easy to get lost in the jungle and spend way too much time on the "perfect" design; remember that without a sound sampling plan, whatever data is collected may not prove up to the task of drawing the actionable insights that the client is really interested in seeing answered.

## 3. Automated Data Collection

> One Man's Trash...
>
> It's been said that the streets of the Web are paved with data that cannot wait to be collected, but you'd be surprised by how much trash there is out there.
>
> – Patrick Boily, 2020

The way we **share**, **collect**, and **publish** data has changed over the past few years due to the ubiquity of the *World Wide Web*. **Private businesses**, **governments**, and **individual users** are posting and sharing all kinds of data and information. At every moment, new channels generate vast amounts of data.

There was a time in the recent past where both scarcity and inaccessibility of data was a problem for researchers and decision-makers. That is **emphatically** not the case anymore.

Data abundance carries its own set of problems, however, in the form of

- tangled masses of data, and
- traditional data collection methods and classical data analysis techniques not being up to the task anymore (which is not to say that the results they would give would be incorrect; it's rather their lack of efficiency that comes into play).

The growth and increasing popularity and power of **open source software**, such as R and Python, for which the source code can be inspected, modified, and enhanced by anyone, makes program-based automated data collection quite appealing.

One note of warning, however: time marches on and packages become **obsolete** in the blink of an eye. If the analyst is unable (or unwilling) to **maintain their extraction/analysis code** and to **monitor the sites** from which the data is extracted, the choice of software will not make much of a difference.

So why bother with automated data collection? Common considerations include:

- the sparsity of financial resources;
- the lack of time or desire to collect data manually;
- the desire to work with up-to-date, high-quality data-rich sources, and
- the need to document the analytical process from beginning (data collection) to end (publication).

Manual collection, on the other hand, tends to be cumbersome and prone to error; non-reproducible processes are also subject to heightened risks of "death by boredom", whereas program-based solutions are typically more reliable, reproducible, time-efficient, and produce datasets of higher quality (this assumes, of course, that coherently presented data exists in the first place).

### 3.1 Automated Data Collection Checklist

That being said, **web scraping** or **statistical text processing** is not always recommended. As a start, it is possible that no online and freely available source of data meets the analysis' needs, in which case an approach based on survey sampling is probably indicated.

If most of the answers to the following questions are positive, then an automated approach may be the right choice.

- Is there a need to repeat the task from time to time (e.g. to update a database)?
- Is there a need for others to be able to replicate the data collection process?
- Are online sources of data frequently used?
- Is the task non-trivial in terms of scope and complexity?
- If the task can be done manually, are the financial resources required to let others do the work lacking?
- Is the will to automate the process by means of programming there?

The objective is simple: automatic data collection should yield a collection of unstructured or unsorted datasets, at a reasonable cost.

### 3.2 Ethical Considerations

We now turn our attention to a burning question for consultants and analysts alike: is all the freely available data on the Internet ACTUALLY freely available?

A **spider** is a program that grazes or crawls the web rapidly, looking for information. It jumps from one page to another, grabbing the entire page content. **Scraping** is taking specific information from specific websites (which is the goal): how are these different?

> "Scraping inherently involves **copying**, and therefore one of the most obvious claims against scrapers is copyright infringement." [8]

```
# robots.txt
#
# This file is to prevent the crawling and indexing of certain parts
# of your site by web crawlers and spiders run by sites like Yahoo!
# and Google. By telling these "robots" where not to go on your site,
# you save bandwidth and server resources.
#
# This file will be ignored unless it is at the root of your host:
# Used:    http://example.com/robots.txt
# Ignored: http://example.com/site/robots.txt
#
# For more information about the robots.txt standard, see:
# http://www.robotstxt.org/robotstxt.html

User-agent: *
Crawl-delay: 10
# Directories
Disallow: /includes/
Disallow: /misc/
Disallow: /modules/
Disallow: /profiles/
Disallow: /scripts/
Disallow: /themes/
# Files
Disallow: /CHANGELOG.txt
Disallow: /cron.php
Disallow: /INSTALL.mysql.txt
Disallow: /INSTALL.pgsql.txt
Disallow: /INSTALL.sqlite.txt
Disallow: /install.php
Disallow: /INSTALL.txt
Disallow: /LICENSE.txt
Disallow: /MAINTAINERS.txt
Disallow: /update.php
```

**Figure 1.** `robots.txt` file for **cqads.carleton.ca**.

What can be done to minimize the risk?

- work as transparently as possible;
- document data sources at all time;
- give credit to those who originally collected and published the data;
- if you did not collect the information, you probably need permission to reproduce it, and, more importantly,
- don't do anything illegal.

A number of cases have shown that the courts have not yet found their footing in this matter (see *eBay* vs. *Bidder's Edge*, *Associated Press* vs. *Meltwater*, *Facebook* vs. *Pete Warden*, *United States* vs. *Aaron Swartz*, for instance [9]). There are legal issues that we are not qualified to discuss, but in general, it seems as though larger companies/organisations usually emerge victorious from such battles.

Part of the difficulty is that it is not clear which scraping actions are illegal and which are legal. There are rough guidelines: re-publishing content for commercial purposes is considered more problematic than downloading pages for research/analysis, say. A site's `robots.txt` (Robots Exclusion Protocol) file tells scrapers what information on the site may be harvested with the publisher's consent – heed that file (see Figure 1 for an example).

Perhaps more importantly, **be friendly!** Not everything that can be scraped needs to be scraped. Scraping programs should 1) behave "nicely"; 2) provide useful data, and 3) be efficient, in that order. When in doubt, contact the data provider to see if they will grant access to the databases or files.

Finally, note the importance of following the **Scraping Do's and Don't's**:

1. **stay identifiable**;
2. **reduce traffic** – accept compressed files, check that a file has been changed before accessing it again, retrieve only parts of a file;
3. **do not bother server with multiple requests** – many requests per second can bring smaller server downs, webmasters may block you if your scraper is too greedy (a few requests per second is fine), and
4. **write efficient and polite scrapers** – there is no reason to scrape pages daily or to repeat the same task over and over, select specific resources and leave the rest untouched.

### 3.3 Web Data Quality

Data quality issues are inescapable. It is not rare for clients to have spent thousands of dollars on data collection (automatic or manual) and to respond to the news that the data is flawed or otherwise unusable with: "well, it's the best data we have, so find a way to use it."

These issues can be side-stepped to some extent if consultants get involved in the project during or prior to the data collection stage, asking questions such as:

- what type of data is best-suited to answer the client's question(s)?
- is the available data of sufficiently high quality to answer the client's question(s)?
- is the available information systematically flawed?

Web data can be **first-hand** information (a tweet or a news article), or **second-hand** (copied from an offline source or scraped from some online location, which may make it difficult to retrace). **Cross-referencing** is a standard practice when dealing with secondary data.

Data quality also depends on its **use(s)** and **purpose(s)**. For example, a sample of tweets collected on a random day could be used to analyse the use of a hashtags or the gender-specific use of words, but that dataset might not prove as useful if it had been collected on the day of the 2018 U.S. Presidential Election to predict the election outcomes (due to **collection bias**).

An example might help to illustrate some the pitfalls and challenges. Let's say that a client is interested in finding out what people think of a new potato peeler using a standard telephone survey. Such an approach has a number of pitfalls:

- **unrepresentative sample** – the selected sample might not represent the intended population;
- **systematic non-response** – people who don't like phone surveys might be less (or more) likely to dislike the new potato peeler;
- **coverage error** – people without a landline can't be reached, say, and
- **measurement error** – are the survey questions providing suitable info for the problem at hand?

Traditional solutions to these problems require the use of survey sampling (more on this later), questionnaire design (see previous section), omnibus surveys, reward systems, audits, etc. These solutions can be **costly**, **time-consuming**, and **ineffective**.

**Proxies** – indicators that are strongly related to the product's popularity without measuring it directly, could be useful. If **popularity** is defined as large groups of people preferring a potato peeler over another one, then sales statistics on a commercial website may provide a proxy for popularity.

Rankings on `Amazon.ca` (or a similar website) could, in fact, paint a more comprehensive portrait of the potato peeler market than would a traditional survey. It would suffice, then, to build a scraper that is compatible with Amazon's **application program interface** (API) to gather the appropriate data.

Of course, there are potential issues with this approach as well:

- **representativeness** of the **listed products** – are all potato peelers listed? If not, is it because that website doesn't sell them? Is there some other reason?
- **representativeness** of the **customers** – are there specific groups buying/not-buying online products? Are there specific groups buying from specific sites? Are there specific groups leaving/not-leaving reviews?
- **truthfulness** of customers and **reliability** of reviews – how can we distinguish between paid (fake) reviews and real reviews?

Web scraping is usually well-suited for collecting data on products (such as the aforementioned potato-peeler), but there are numerous questions for which it is substantially more difficult to imagine where data could be found online: what data could you collect online to measure the popularity of a government policy, say?

### 3.4 Web Technologies 101

Online data can be found in **text**, **tables**, **lists**, **links**, and other structures, but the way data is presented in browsers is not necessarily how it is stored in HTML/XML. Furthermore, when web pages are **dynamic**, there is a "cost" associated with automated collection. Consequently, a basic knowledge of the web and web-related techs and documents is crucial. Information is readily available online (see references) and in [8, 9].

There are three areas of importance for data collection on the web:

- technologies for **content dissemination** (HTTP, HTML/XML, JSON, plain text, etc.);
- technologies for **information extraction** (R, Python, XPath, JSON parsers, Beautiful Soup, Selenium, regexps, etc.), and
- technologies for **data storage** (R, Python, SQL, binary formats, plain text formats, etc.).

Webpage content itself comes into three main categories: Hypertext Markup Language (HTML; used for web content and code), Cascading Style Sheets (CSS; used for webpage style), and JavaScript (js; used for interactivity with the webpage). HTML is, in some sense, the most fundamental; understanding the tree structure of HTML documents, for instance, will go a long way towards helping consultants get full use of the **scraping toolbox**.

### 3.5 Scraping Toolbox

From experience, we know that a number of tools can facilitate the automated data extraction process, including: *Developer Tools*, *XPath*, *Beautiful Soup*, *Selenium*, and *regular expressions*.

**Developer Tools** show the correspondence between the HTML code for a page and the rendered version seen in the browser (see Figure 2 for an example). Unlike "View Source", Developer Tools show the *dynamic* version of the HTML content (i.e. the HTML is shown with any changes made by JavaScript since the page was first received). Inspecting a page's various elements and discovering where they reside in the HTML file is **crucial** to efficient web scraping:

- **Firefox** – right click page → Inspect Element
- **Safari** – Safari → Preferences → Advanced → Show Develop Menu in Menu Bar, then Develop → Show Web Inspector
- **Chrome** – right click page → Inspect

**XPath** is a query (domain-specific) language which is used to select specific pieces of information from marked-up documents such as HTML, XML, or variants such as SVG, RSS. Before this can be done, the information stored in a marked-up document needs to be converted (or **parsed**) into a format suitable for processing and statistical analysis; this is implemented in the R package XML, for instance. The process is simple; it involves

1. specifying the data of interest;
2. locating it in a specific document, and
3. tailoring a query to the document to extract the desired info.

XPath queries require both a **path** and a **document** to search; paths consist of hierarchical addressing mechanism (succession of nodes, separated by forward slashes ("/"), while a query takes the form `xpathSApply(doc,path)`: `xpathSApply(parsed_doc,"/html/body/div/p/i")`, for instance, would find all `<i>` tags found under a `<p>` tag, itself found under a `<div>` tag in the `body` of the `html` file of `parsed_doc`. Consult [8] for a substantially heftier introduction.

**Regular Expressions** can be used to achieve the main web scraping objective, which is to extract relevant information from reams of data. Among this mostly unstructured data lurk **systematic elements**, which can be used to help the automation process, especially if quantitative methods are eventually going to be applied to the scraped data. Systematic structures include numbers, names (countries, etc.), addresses (mailing, e-mailing, URLs, etc.), specific character strings, etc. Regular expressions (regexps) are abstract sequences of strings that match concrete recurring patterns in text; they allow for the systematic extraction of the information components from plain text, HTML, and XML. Some examples that illustrate the main concepts are shown in the accompanying *Jupyter Notebooks*.

**Beautiful Soup** is a Python library that helps extract data out of HTML and XML files. It parses HTML files, even if they're broken. Beautiful Soup does not simply convert bad HTML to good X/HTML; it allows a user to fully inspect the (proper) HTML structure it produces, in a programmatical fashion. When Beautiful Soup has finished its work on an HTML file, the resulting *soup* is an API for **traversing**, **searching**, and **reading** the document's elements. In essence, it provides **idiomatic** ways of navigating, searching, and modifying the parse tree of the HTML file, which can save a fair amount of time.

For instance, `soup.find_all('a')` would find and output all `<a ...> ... </a>` tag pairs (with attributes and content) in the `soup`, whereas

```
for link in soup.find_all('a'):
print(link.get('href')
```

would output the URLs found in the same tag pairs. The Beautiful Soup documentation is quite explicit and provides numerous examples [14].

**Selenium** is a Python tool used to automate web browser interactions. It is used primarily for testing purposes, but it has data extraction uses as well. Mainly, it allows the user to open a browser and to act as a human being would:

- clicking buttons;
- entering information in forms;
- searching for specific information on a page, etc.

Selenium requires a driver to interface with the chosen browser. Firefox, for example, uses `geckodriver`. Other supported browsers have their own drivers (see [15–18]).

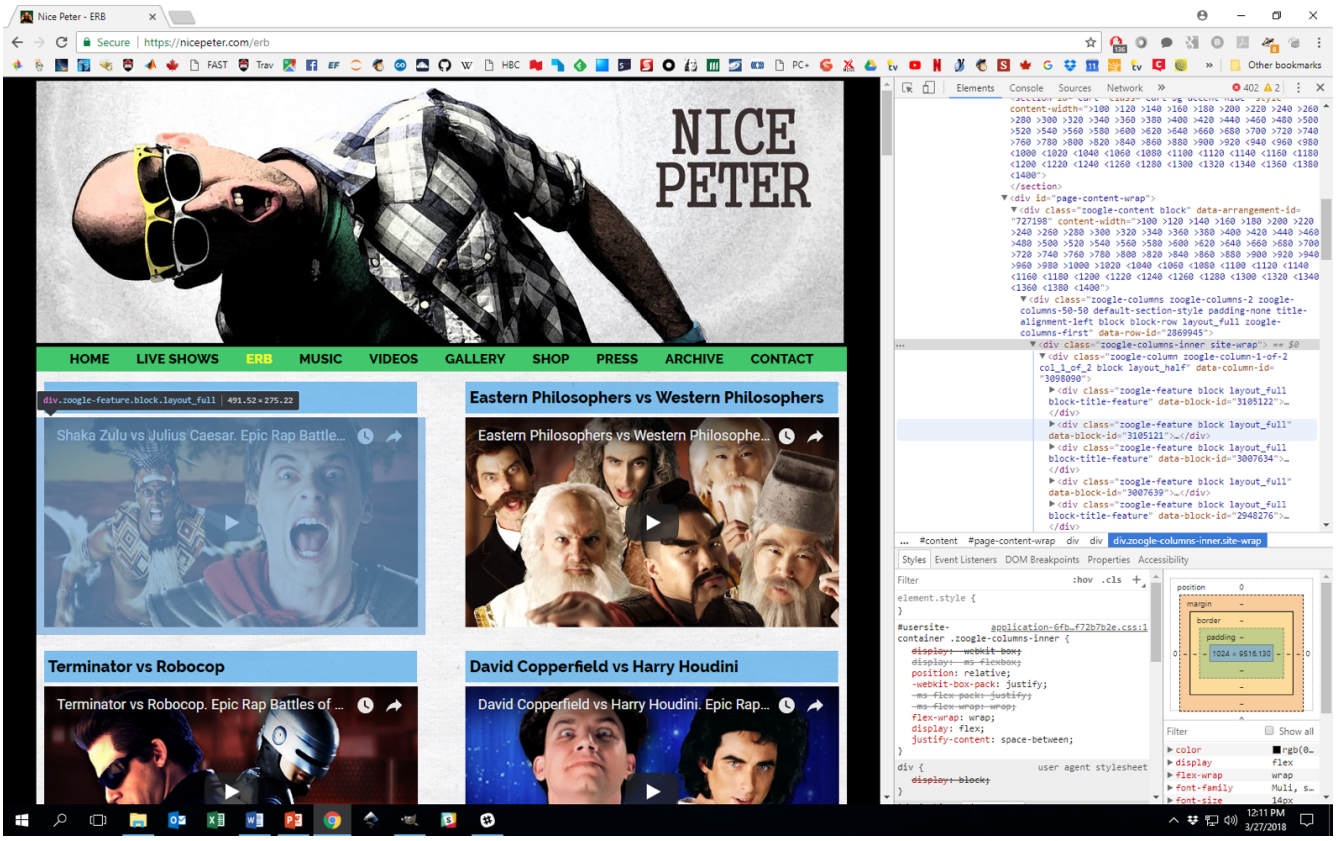


**Figure 2.** Inspecting **https://nicepeter.com/erb**'s elements using Chrome's *Developer Tools*.

Selenium automatically controls a complete browser, including rendering the web documents and running JavaScript. This is useful for pages with a lot of dynamic content that isn't in the base HTML. Selenium can program actions like "click on this button", or "type this text", to provide access to the dynamic HTML of the current state of the page, not unlike what happens in *Developer Tools* (but now the process can be fully automated). More information can be found in [12, 13].

Let us end this section by providing a short summary of the **automated data collection decision process** [8,9], as seen by quantitative consultants.

1. **Know exactly what kind of information the client needs**, either **specific** (e.g. GDP of all OECD countries for last 10 years, sales of top 10 tea brands in 2017, etc.) or **vague** (people's opinion on tea brand *X*, etc.)

2. **Find out if there are any web data sources that could provide direct or indirect information on the client's problem.** That is easier to achieve for specific facts (a tea store's webpage will provide information about teas that are currently in demand) than it is for vague facts. Tweets and social media platforms may contain opinion trends; commercial platforms can provide information on product satisfaction.

3. **Develop a theory of the data generation process when looking into potential data sources.** When was the data generated? When was it uploaded to the Web? Who uploaded the data? Are there any potential areas that are not covered, consistent, or accurate? How often is the data updated?

4. **Balance the advantages and disadvantages of potential data sources.** Validate the quality of data used – are there other independent sources that provide similar information against which to crosscheck? Can original source of secondary data be identified?

5. **Make a data collection decision**. Choose the data sources that seem most suitable, and document reasons for this decision. Collect data from several sources to validate the final choice.

## 4. Statistical Survey Sampling

> You Can't Say It's Not True
>
> The latest survey shows that 3 out of 4 people make up 75% of the world's population.
>
> – David Letterman (attributed)

While the *World Wide Web* does contain troves of data, web scraping does not address the question of data validity: will the extracted data be **useful** as an analytical component? Will it suffice to provide the quantitative answers that the client is seeking?

A **survey** (a fair amount of information for this section is taken from [1,6]) is any activity that collects information about characteristics of interest:

- in an **organized** and **methodical** manner;
- from some or all **units** of a population;
- using **well-defined** concepts, methods, and procedures, and
- compiles such information into a **meaningful** summary form.

A **census** is a survey where information is collected from all units of a population, whereas a **sample survey** uses only a fraction of the units.

### 4.1 Sampling Model

When survey sampling is done properly, we may be able to use various statistical methods to make inferences about the **target population** by sampling a (comparatively) small number of units in the **study population**. The relationship between the various populations (**target**, **study**, **respondent**) and samples (**sample**, **intended**, **achieved**) is illustrated in Figure 3.

### 4.2 Deciding Factors

In some instances, information about the **entire** population is required in order to solve the client's problem, whereas in others it is not necessary. How does one determine which type of survey must be conducted to collect data? The answer depends on multiple factors:

- the type of question that needs to be answered;
- the required precision;
- the cost of surveying a unit;
- the time required to survey a unit;
- size of the population under investigation, and
- the prevalence of the attributes of interest.

Once a choice has been made, each survey typically follows the same **general steps**:

1. statement of objective
2. selection of survey frame
3. sampling design
4. questionnaire design
5. data collection
6. data capture and coding
7. data processing and imputation
8. estimation
9. data analysis
10. dissemination and documentation

The process is not always linear, in that preliminary planning and data collection may guide the implementation (selection of a frame and of a sampling design, questionnaire design), but there is a definite movement from objective to dissemination.

### 4.3 Survey Frames

The **frame** provides the means of **identifying** and **contacting** the units of the study population. It is generally costly to create and to maintain (in fact, there are organisations and companies that specialise in building and/or selling such frames). Useful frames contain:

- identification data,
- contact data,
- classification data,
- maintenance data, and
- linkage data.

The ideal frame must minimize the risk of **undercoverage** or **overcoverage**, as well as the number of **duplications** and **misclassifications** (although some issues that arise can be fixed at the data processing stage).

Unless the selected frame is **relevant** (which is to say, it corresponds, and permits accessibility to, the target population), **accurate** (the information it contains is valid), **timely** (it is up-to-date), and **competitively priced**, the statistical sampling approach is contraindicated.

### 4.4 Survey Error

One of the strengths of statistical sampling is in its ability to provide estimates of various quantities of interest in the target population, and to provide some control over the **total error** (TE) of the estimates. The TE of an estimate is the amount by which it differs from the true value for the target population:

$$\begin{aligned}\text{Total Error} = \text{Measurement Error} + \text{Sampling Error} \\ + \text{Nonresponse Error} + \text{Coverage Error},\end{aligned}$$

where the

- **coverage error** is due to differences in the study and target populations;
- **non-response error** is due to differences in the respondent and study populations;
- **sampling error** is due to differences in the achieved sample and the respondent population;
- **measurement error** is due to true value in the achieved sample not being assessed correctly.

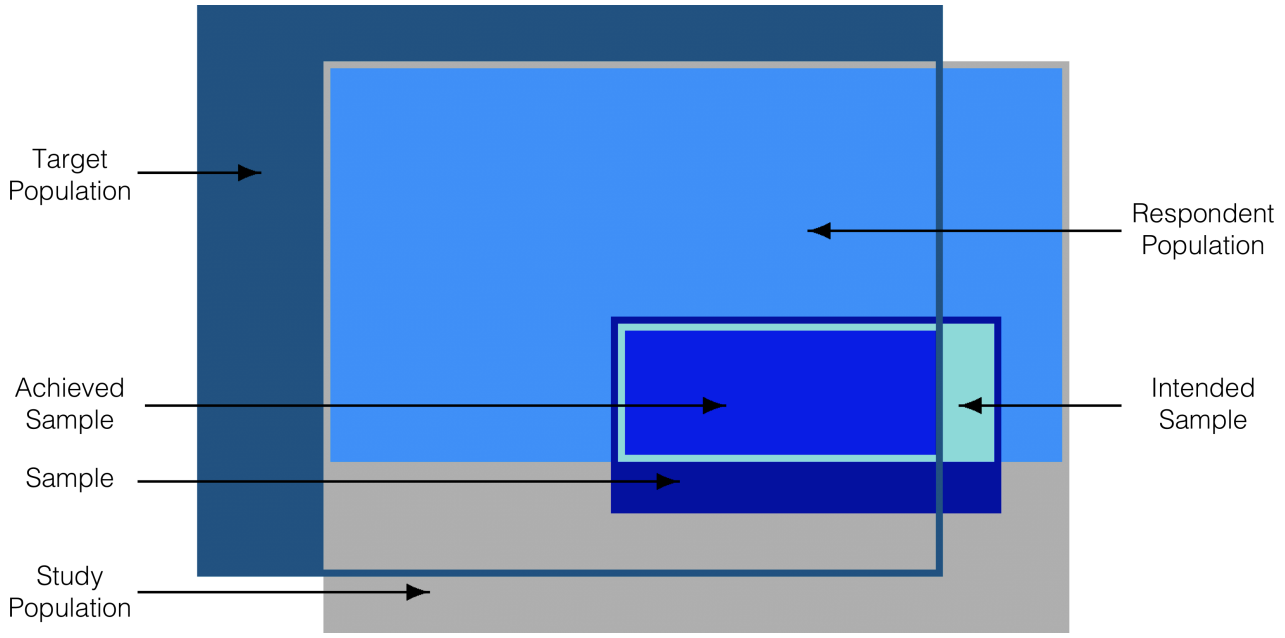


**Figure 3.** Various populations and samples in the sampling model.

If we let

- $\overline{x}$ be the computed attribute value in the achieved sample;
- $\overline{x}_{\text{true}}$ be the true attribute value in the achieved sample under perfect measurement;
- $x_{\text{resp}}$ be the attribute value in the respondent population;
- $x_{\text{study}}$ be the attribute value in the study population, and
- $x_{\text{tar}}$ be the attribute value in the target population,

then

$$\begin{aligned}\text{Total Error} &= \overline{x} - x_{\text{tar}} \\ &= (\overline{x} - \overline{x}_{\text{true}}) + (\overline{x}_{\text{true}} - x_{\text{resp}}) \\ &\quad + (x_{\text{resp}} - x_{\text{study}}) + (x_{\text{study}} - x_{\text{tar}}).\end{aligned}$$

In an ideal scenario, Total Error $= 0$. In practice, there are two main contributions to Total Error: **sampling errors** (which we will discuss shortly) and **nonsampling errors**, which include every contribution to survey error which is not due to the choice of sampling scheme. The latter can be controlled, to some extent:

- **coverage error** can be minimized by selecting a high quality, up-to-date survey frame;
- **non-response error** can be minimized by careful choice of the data collection mode and questionnaire design, and by using "call-backs" and "follow-ups";
- **measurement error** can be minimized by careful questionnaire design, pre-testing of the measurement apparatus, and cross-validation of answers.

These suggestions are perhaps less useful than one could hope in modern times: survey frames based on landline telephones are quickly becoming irrelevant in light of an increasingly large and younger population who eschew such phones, for instance, while response rates for surveys that are not mandated by law are surprisingly low. This explains, in part, the impetus towards automated data collection and the use of **non-probabilistic sampling** methods.

### 4.5 Modes of Data Collection

How is data traditionally captured, then? There are **paper-based** approaches, **computer-assisted** approaches, and a suite of other modes.

- **Self-administered questionnaires** are used when the survey requires detailed information to allow the units to consult personal records (which reduces measurement errors), they are useful to measure responses to sensitive issues as they provide an extra layer of privacy, and are typically not as costly as other collection modes, but they tend to be associated with high non-response rate since there is less pressure to respond.

- **Interviewer-assisted questionnaires** use trained interviewers to increase the response rate and overall quality of the data. Face-to-face **personal interviews** achieve the highest response rates, but they are costly (both in training and in salaries). Furthermore, the interviewer may be required to visit any selected respondents many times before contact is established. **Telephone interviews**, on the other hand produce "reasonable" response rates at a reasonable cost and they are safer for the interviewers, but they are limited in length due to respondent phone fatigue. With random dialing, 4-6 minutes of the interviewer's time

is spent in out-of-scope numbers for each completed interview.

- **Computer-assisted interviews** combine data collection and data capture, which saves valuable time, but the drawback is that not every sampling unit may have access to a computer/data recorder (although this is becomine less prevalent). All paper-based modes have a computer-assisted equivalent: **computer-assisted self-interview** (CASI), **computer-assisted interview** (CAI), **computer-assisted telephone interview** (CATI), and **computer-assisted personal interview** (CAPI).
- Unobtrusive direct observation; diaries to be filled (paper or electronic); omnibus surveys. and email, Internet, and social media.

### 4.6 Non-Probabilistic Sampling

There exists a number of methods to select sampling units from the target population that use subjective, non-random approaches (NPS). These methods tend to be **quick**, **relatively inexpensive** and **convenient** in that a survey frame is not needed. NPS methods are ideal for **exploratory analysis** and **survey development**.

Unfortunately, they are sometimes used **instead** of probabilistic sampling designs, which is problematic; the associated selection bias makes NPS methods **unsound** when it comes to **inferences**, as they cannot be used to provide **reliable estimates of the sampling error** (the only component of Total Error on which the analysts has direct control). Automated data collection often fall squarely in the NPS camp, for instance. While we can still analyse data collected with a NPS approach, we **may not generalise the results** to the target population (except in rare, census-like situations).

NPS methods include

- **Haphazard** sampling, also known as 'man on the street' sampling; it assumes that the population is homogeneous, but the selection remains subject to interviewer biases and the availability of units;
- **Volunteer** sampling in which the respondents are self-selected; there is a large selection bias since the silent majority does not usually volunteer; this method is often imposed upon analysts due to ethical considerations; it is also used for focus groups or qualitative testing;
- **Judgement** sampling is based on the analysts' ideas of the target population composition and behaviour (sometimes using a prior study); the units are selected by population experts, but inaccurate preconceptions can introduce large biases in the study;
- **Quota** sampling is very common (and is used in exit polling to this day in spite of the infamous "Dewey Defeats Truman" debacle of 1948 [19]); sampling continues until a specific number of units have been selected for various sub-populations; it is preferable to other NPS methods because of inclusion of sub-populations, but it ignores non-response bias;
- **Modified** sampling starts out using probability sampling (more on this later), but turns to quota sampling in its last stage, in part as a reaction to high non-response rates;
- **Snowball** sampling asks sampled units to recruit other units among their acquaintances; this NPS approach may help locate hidden populations, but it biased in favour of units with larger social circles and units that are charming enough to convince their acquaintances to participate.

There are contexts where NPS methods might fit a client's need (and that remains their decision to make, ultimately), but the analyst MUST still inform the client of the drawbacks, and present some probabilistic alternatives.

### 4.7 Probabilistic Sampling

The inability to make sound inferences in NPS contexts is a monumental strike against their use. While probabilistic sample designs are usually **more difficult and expensive** to set-up (due to the need for a quality survey frame), and take **longer** to complete, they provide **reliable estimates** for the attribute of interest and the sampling error, paving the way for small samples being used to draw inferences about larger target populations (in theory, at least; the non-sampling error components can still affect results and generalisation).

We shall take a deeper look at traditional probability sample designs such as **simple random**, **stratified random**, and **systematic**, – **cluster**, **probability proportional to size**, **replicated**, **multi-stage** and **multi-phase** variants also exist (see [1, 6] for details).

Let us start with some basic mathematical concepts. Consider a finite population $\mathscr{U} = \{u_1, \ldots, u_N\}$. The **mean** and **variance** of the population are given by

$$\mu = \frac{1}{N}\sum_{j=1}^{N} u_j \quad \text{and} \quad \sigma^2 = \frac{1}{N}\sum_{j=1}^{N}(u_j - \mu)^2, \quad \text{respectively.}$$

If $\mathscr{Y} = \{y_1, \ldots, y_n\}$ is a sample of $\mathscr{U}$, the **sample mean** and **sample variance** (also known as the **empirical mean** and **empirical variance**) are given by

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \quad \text{and} \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2, \quad \text{respectively.}$$

Let $X_1, \ldots, X_n$ be random variables, $b_1, \ldots, b_n \in \mathbb{R}$, and E, V, and Cov be the **expectation**, **variance** and **covariance**

operators, respectively. Recall that

$$\begin{aligned}
\mathrm{E}\left(\sum_{i=1}^{n} b_i X_i\right) &= \sum_{i=1}^{n} b_i \,\mathrm{E}(X_i) \\
\mathrm{V}\left(\sum_{i=1}^{n} b_i X_i\right) &= \sum_{i=1}^{n} b_i^2 \,\mathrm{V}(X_i) + \sum_{1\le i\ne j}^{n} b_i b_j \,\mathrm{Cov}(X_i, X_j) \\
\mathrm{Cov}(X_i, X_j) &= \mathrm{E}(X_i X_j) - \mathrm{E}(X_i)\,\mathrm{E}(X_j) \\
\mathrm{V}(X_i) &= \mathrm{Cov}(X_i, X_i) = \mathrm{E}\left(X_i^2\right) - \mathrm{E}^2(X_i).
\end{aligned}$$

The **bias** in an error component is the average of that error component if the survey is repeated many times independently under the same conditions. The **variability** in an error component is the extent to which that component would vary about its average value in the ideal scenario described above. The **mean square error** of an error component is a measure of the size of the error component:

$$\begin{aligned}
\mathrm{MSE}(\hat{\beta}) &= \mathrm{E}\left((\hat{\beta}-\beta)^2\right) = \mathrm{E}\left((\hat{\beta}-\mathrm{E}(\hat{\beta})+\mathrm{E}(\hat{\beta})-\beta)^2\right) \\
&= \mathrm{V}(\hat{\beta}) + \left(\mathrm{E}(\hat{\beta})-\beta\right)^2 = \mathrm{V}(\hat{\beta}) + \mathrm{Bias}^2(\hat{\beta})
\end{aligned}$$

where $\hat{\beta}$ is an estimate of $\beta$. Incidentally, the unusual denominator in the sample variance insures that it is an unbiased estimator of the population variance.

Finally, if the estimate is unbiased, then an approximate **95% confidence interval** (95% CI) for $\beta$ is given by

$$\hat{\beta} \pm 2\sqrt{\hat{\mathrm{V}}(\hat{\beta})},$$

where $\hat{\mathrm{V}}(\hat{\beta})$ is a sampling design-specific estimate of $\mathrm{V}(\hat{\beta})$.

In what follows, we discuss a number of sampling designs and present some of their advantages and disadvantages. We also show how to compute estimates for various population attributes (mean, total, proportion, ratio, difference, regression) and how to estimate the corresponding 95% CI. Finally, we briefly discuss how to compute sample sizes for a given **error bound** (an upper limit on the radius of the desired 95% CI), and how to determine the **sample allocation** (how many units to be sampled in various subpopulation groups), for designs where it is appropriate to do so.

In all instances, the target population consists of $N$ measurements/units, $\mathscr{U} = \{u_1, \ldots, u_N\}$, and the true population mean, population variance, population total, and population proportion for the variable of interest are $\mu$, $\sigma^2$, $\tau$, and $p$, respectively. The sample is a subset of the target population, $\mathscr{Y} = \{y_1, \ldots, y_n\} \subseteq \mathscr{U}$ from which we estimate the respective population attributes *via* $\overline{y}$, $s^2$, $\hat{\tau}$, and $\hat{p}$.

For a given characteristic, we define $\delta_i$ as 1 or 0 depending on whether the corresponding sample unit $y_i$ possesses the characteristic in question or not. Lastly, we set the error bound to $B = 2\sqrt{\hat{V}} > 0$.

In **Simple Random Sampling** (SRS), $n$ units are selected randomly from the survey frame, as in Figure 4 (top left image). It is by far the easiest sampling design to implement, and estimates for the resulting sampling errors are well known and easy to compute. Another advantage is that SRS does not require auxiliary information, which can be useful with more economical survey frames.

This can backfire however, as SRS makes no use of such information even when it is available. There is also no guarantee that the sample will be representative of the population. Note as well that SRS may be costly if the sample is widely spread out, geographically.

The SRS estimators are

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad \hat{\tau} = N\overline{y}, \quad \text{and} \quad \hat{p} = \frac{1}{n}\sum_{i=1}^{n} \delta_i$$

with respective variances

$$\mathrm{V}(\overline{y}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right), \quad \mathrm{V}(\hat{\tau}) = N^2 \cdot \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right),$$

and

$$\mathrm{V}(\hat{p}) = \frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right).$$

The 95% CI is approximated by substituting the true variance $\sigma^2$ by the unbiased estimator $\frac{n-1}{n}s^2$:

$$\hat{\mathrm{V}}(\overline{y}) = \frac{s^2}{n}\left(1-\frac{n}{N}\right), \quad \hat{\mathrm{V}}(\hat{\tau}) = N^2 \cdot \frac{s^2}{n}\left(1-\frac{n}{N}\right),$$

and

$$\hat{\mathrm{V}}(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n-1}\left(1-\frac{n}{N}\right).$$

Finally, the sample size required to achieve an upper error bound $B$ are

$$n_{\overline{y}} = \frac{4N\sigma^2}{(N-1)B^2 + 4\sigma^2}, \quad n_{\hat{\tau}} = \frac{4N^3\sigma^2}{(N-1)B^2 + 4N^2\sigma^2}$$

and

$$n_{\hat{p}} = \frac{4Np(1-p)}{(N-1)B^2 + 4p(1-p)},$$

where $\sigma^2$ and $p$ have been previously estimated (perhaps as part of a prior survey).

In **Stratified Random Sampling** (STS), $n = n_1 + \cdots + n_k$ units are selected randomly from the survey frame by first establishing $k$ natural strata (such as provinces, or age groups), and selecting $n_j$ units from the $N_j$ units in stratum $j$, with $\overline{y}_j$ and $\hat{p}_j$ the SRS estimators in stratum $j$, $j = 1, \ldots, k$. An illustration is provided in Figure 4 (top middle).

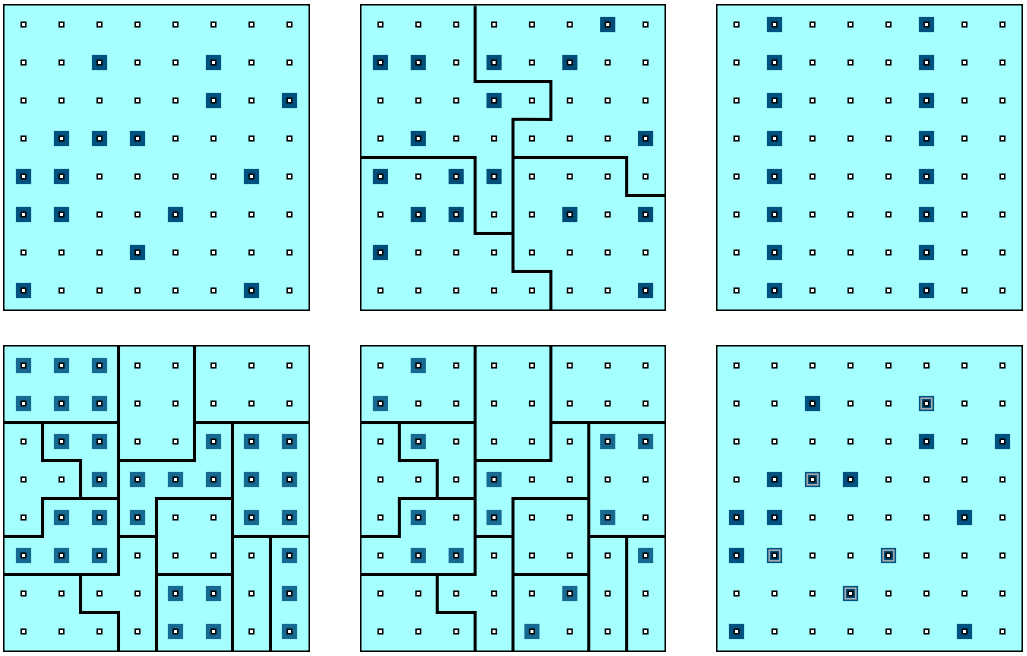

**Figure 4.** Schematics of sampling designs. Top row, from left to right: simple random sampling, stratified random sampling, systematic random sampling; bottom row, from left to right: cluster sampling, multi-stage sampling, multi-phase sampling.

STS may produce a smaller bound on the error of estimation than would be produced by a SRS of the same size, particularly if measurements within a strata are homogeneous, and it may be less expensive to implement if the elements are stratified into convenient groupings. Another added benefits is that it may provide parameter estimates for sub-populations that coincide with the strata. There are no major disadvantage to this sample design, except for the fact that there might not be natural ways to stratify the frame (in the sense that each stratum might not be homogeneous in its units), in which case STS is roughly equivalent to SRS.

The STS estimators are

$$\overline{y}_{\text{st}} = \sum_{j=1}^{k} \frac{N_j}{N} \overline{y}_j, \quad \hat{\tau}_{\text{st}} = N\overline{y}_{\text{st}}, \quad \text{and} \quad \hat{p}_{\text{st}} = \sum_{j=1}^{k} \frac{N_j}{N} \hat{p}_j,$$

with approximate variances given by

$$\hat{\text{V}}(\overline{y}_{\text{st}}) = \frac{1}{N^2} \sum_{j=1}^{k} N_i^2 \hat{\text{V}}\left(\overline{y}_j\right), \quad \hat{\text{V}}(\hat{\tau}_{\text{st}}) = N^2 \hat{\text{V}}(\overline{y}_{\text{st}}),$$

and

$$\hat{\text{V}}(\hat{p}_{\text{st}}) = \frac{1}{N^2} \sum_{j=1}^{k} N_i^2 \hat{\text{V}}\left(\overline{p}_j\right).$$

In the STS design, the sample determination question is two-fold: what size $n$ should the sample have, and how should they be allocated to each stratum $(n_j, j = 1, \ldots, k)$.

We can select $n$ based on **cost** considerations or on error bound considerations. Let $c_0$ be the fixed survey operation costs (**overhead**), $c_j$ be the **cost per response** in stratum $j$ (which may need to include the cost of trying to reach non-respondents), and $C$ be the **total cost** of conducting the survey. The sample size $n$ that minimises $\hat{\text{V}}(\overline{y}_{\text{st}})$ subject to $C = c_0 + \sum_{j=1}^{k} c_j n_j$ and $n = \sum_{j=1}^{k} n_j$ is

$$n_{\text{st},C} = (C - c_0) \frac{\sum_{j=1}^{k} \frac{N_j \sigma_j}{\sqrt{c_j}}}{\sum_{j=1}^{k} N_j \sigma_j \sqrt{c_j}}.$$

In the **general optimum allocation scheme**, the sampling weights (by strata) are

$$w_j = \frac{n_j}{n} = \frac{N_j \sigma_j c_j^{-1/2}}{\sum_{\ell=1}^{k} N_\ell \sigma_\ell c_\ell^{-1/2}}.$$

In the **Neyman allocation scheme**, we assume that the cost per response is identical in each stratum, whence

$$w_{j,N} = \frac{n_j}{n} = \frac{N_j \sigma_j}{\sum_{\ell=1}^{k} N_\ell \sigma_\ell},$$

while in the **proportional allocation scheme** we further assume that $\sigma_j = \sigma$ for all $j$, so that

$$w_{j,P} = \frac{n_j}{n} = \frac{N_j}{N}.$$

Other allocation schemes are also sometimes selected, such as the **square root proportional scheme** which fixes

$$w_{j,S} = \frac{N_j^{1/2}}{\sum_{\ell=1}^{k} N_\ell^{1/2}}$$

in order to insure that smaller strata (e.g. provinces with smaller populations, say) are allocated enough observations to produce sub-population estimates.

Note that while budgetary considerations need to be considered in practice, the preceding approach does not allow prescribed error bounds, which could prove problematic. The sample sizes required to achieve an upper error bound $B$ are

$$n_{\mathrm{st},\overline{y}} = \frac{4\sum_{j=1}^{k}\frac{N_j\sigma_j^2}{w_j}}{N^2B^2+4\sum_{j=1}^{k}N_j\sigma_j^2}, \quad n_{\mathrm{st},\hat{\tau}} = \frac{4N^2\sum_{j=1}^{k}\frac{N_j\sigma_j^2}{w_j}}{N^2B^2+4\sum_{j=1}^{k}N_j\sigma_j^2},$$

and

$$n_{\mathrm{st},\hat{p}} = \frac{4\sum_{j=1}^{k}\frac{N_jp_j(1-p_j)}{w_j}}{N^2B^2+4\sum_{j=1}^{k}N_jp_j(1-p_j)},$$

where $\sigma_j^2$ and $p_j$ have been previously estimated, and a specific allocation scheme $\{w_j\}$ has already been selected.

In **Systematic Sampling** (SYS), $n$ units are selected randomly from the survey frame by first (randomly) selecting a unit $y_1$ among the first $k = \lfloor \frac{N}{n} \rfloor$ units in the frame and s systematically adding every subsequent $k^{\text{th}}$ unit to the sample. An illustration is provided in Figure 4 (top right).

SYS is typically appropriate when the frame is already **sorted** along the characteristic of interest in which case it provides greater information by unit cost than SRS. It is simpler to implement than SRS since only one random number is required, and like SRS, it does not require auxiliary frame information. Depending on the sample size and on how the frame is sorted, SYS can produce a sample that is more widely spread (and thus perhaps more representative) than SRS, which may help eliminate other sources of bias.

On the other hand, it can introduce bias when the pattern used for the systematic sample coincides with a pattern in the population, and it makes no use of auxiliary frame information even if such information exists. Furthermore, any advantage in precision over SRS disappears if the frame is randomly ordered. Embarrassingly, SYS may lead to a variable sample size if $n$ does not evenly divide $N$; perhaps more importantly, SYS does not allow for an unbiased estimator of the sampling variance.

For all practical purposes, SYS behaves like SRS for a random population. In that case, the SRS variance formula may provide a decent approximation.

If the frame is **ordered** along the characteristic of interest, each SYS sample will contain some of the smaller values as well as some of the larger values, which would not necessarily be the case in a general SRS sample. This implies that the SRS estimators will have smaller variances than the corresponding SYS estimators, so that the use of the SRS variance formula produces an underestimate of the true sampling error in that case.

In a similar vein, a population is **periodic** if the frame is **periodic** along the characteristic of interest, a SYS sample that hits both the peaks and valleys of a cyclical trend will bring the method more in line with SRS and allow the use of the SRS variance formula as a reasonable approximation. To avoid the problem of underestimating the variation, consider changing the random starting point several times.

If $n$ divides $N$ evenly, then systematic sampling can be viewed as grouping the population into $k = N/n$ strata, and selecting one unit from each stratum. The difference between SYS and STS is that only the first unit is picked randomly in SYS – all other samples are automatically selected based on the position of the first choice.

One can also view SYS as a one-stage cluster sampling (see the next sub-section), where a primary sampling unit is defined as one of the $k = N/n$ possible systematic samples. An SRS of one unit can then be drawn from these $k$ primary sampling units. The SYS sample will consist of all of the items in the selected primary sample.

The SYS estimators are computed exactly as the corresponding SRS estimators; their variances are given by

$$\mathrm{V}(\overline{y}_{\mathrm{sys}}) = \frac{\sigma^2}{n}[1+(n-1)\rho], \quad \mathrm{V}(\hat{\tau}_{\mathrm{sys}}) = N^2\,\mathrm{V}(\overline{y}_{\mathrm{sys}}),$$

and

$$\mathrm{V}(\hat{p}_{\mathrm{sys}}) = \frac{p(1-p)}{n}[1+(n-1)\rho],$$

where $\rho$ is the **intra-cluster correlation** (which is typically impossible to compute exactly).

Other sampling schemes tend to be substantially more complicated (in the sense that the estimators and variance estimates are harder to derive), but the conceptual ideas behind those sampling schemes are still pretty straightforward; if required, in-depth details can be found in [6].

**Cluster Sampling** (CLS), for instance is typically used when the data collection cost increases with the "distance" separating the elements. The population is separated in clusters, and an SRS of clusters is selected – all units within a selected clusters are retained in the sample (see Figure 4 (bottom left) for an illustration). As an example, to sample individuals in the population without a population frame (which might be hard to come by), it might be easier to obtain a dwelling frame and to start by sampling dwellings (which are the population **clusters**), and then to select all individuals in the sampled dwellings. CLS surveys are usually less expensive and less time-consuming to conduct than SRS, and they can be used to show "regional" variations, but they will be wasteful if the cluster sizes are too large, and biased if only a few clusters are sampled.

## References


[1] Farrell, P., *STAT 4502 Survey Sampling Course Package*, Fall 2008

[2] Lessler, J. and Kalsbeek, W. [1992], *Nonsampling Errors in Surveys*, Wiley, New York

[3] Oppenheim, N. [1992], *Questionnaire Design, Interviewing, and Attitude Measurement*, St. Martin's Press

[4] Hidiroglou, M., Drew, J. and Gray, G. [1993], "A Framework for Measuring and Reducing non-response in Surveys," *Survey Methodology*, v.19, n.1, pp.81-94

[5] Gower, A. [1994], "Questionnaire Design for Business Surveys," *Survey Methodology*, v.20, n.2, pp.125-136

[6] *Survey Methods and Practices*, Statistics Canada, Catalogue no.12-587-X

[7] **Sir Humphrey's Primer on Leading Questions**, *Yes, Prime Minister*, S01, E02, BBC, 1986.

[8] Munzert, S., Rubba, C., Meissner, P., Nyhuis, D. [2015], *Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining*, Wiley

[9] Mitchell, R. [2015], *Web Scraping with Python*, O'Reilly.

[10] **XPath introduction**

[11] **Wikipedia article on XML/HTML**

[12] Taracha, R. [2017], **Introduction to Web Scraping Using Selenium**.

[13] **Selenium documentation**

[14] **Beautiful Soup documentation**

[15] ***Chrome* driver**

[16] ***Edge* driver**

[17] ***Firefox* driver**

[18] ***Safari* driver**

[19] DeTurck's, D., **Case Study 2: the 1948 Presidential Election**, retrieved on 12 July 2018.

[20] Allie, E. [2014], Canadian Vehicle Use Study: Electronic Data Collection, in *Proceedings of Statistics Canada Symposium 2014*, Beyond traditional survey taking: adapting to a changing world

[21] **Storing data in DNA is a lot easier than getting it back out**, MIT Technology Review, Jan 2018.