Survey Sampling Methods | 11

by **Patrick Boily**; inspired by **Patrick Farrell**

11.1 Background 435
Survey Sampling Generalities438
Survey Frames 440
Fundamental Concepts . . . 440
Data Collection Basics . . . 443
Types of Sampling Methods 443
11.2 Questionnaire Design . . . 446
Basic Concepts 446
Question Types 447
Design Considerations . . . 448
Question Order 449
11.3 Simple Random Sampling . 451
Basic Notions 455
Estimators and C.I. 458
Sample Size 471
11.4 Stratified Random Sampling 473
Estimators and C.I. 480
Sample Size and Allocation 490
Comparison: SRS and STS 497
11.5 Ratio, Regression, Difference499
Ratio Estimation 499
Regression Estimation . . . 510
Difference Estimation 518
Comparisons 521
11.6 Cluster Sampling 525
Estimators and C.I. 525
Sample Size 541
Comparison: SRS and CLS 543
11.7 Special Topics 544
Systematic Sampling 544
Sampling with PPS 550
Multi-Stage Sampling . . . 553
Multi-Phase Sampling . . . 557
Miscellaneous 559
11.8 Exercises 567

Simply put, data analysis requires data. In pedagogical settings, we take for granted that the data at our disposal is "perfect" (or "ideal"): it either consists of the totality of potentially available data, or it is a representative subset thereof. In practice, either of these can be difficult to achieve; it can prove costly (and sometimes impractical) to collect data from which we can infer population trends and characteristics.

While web scraping (and automated methods) are sometimes used to facilitate the data collection process (see Chapter 28, *Web Scraping and Automatic Data Collection*), the samples that they provide often fail to be representative enough to be of use in practice.

In this chapter, we discuss the principles that underlie statistical sampling methods, and show how to obtain estimates for various sampling plans.

11.1 Background

> To call in the statistician after the experiment is done may be no more than asking them to perform a post-mortem examination: at best, they may be able to 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 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.[1]

[1] Unless some clause has specifically been put in the contract/agreement to allow a graceful exit at this point, consultants will have to proceed with the analysis, flaws and all. It is **EXTREMELY IMPORTANT** that one does not simply sweep these flaws under the carpet. Address them repeatedly in meetings with the clients, and make sure that the analysis results that are presented or reported on include an appropriate *caveat*.

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?

Next steps nearly always requires obtaining relevant data.

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* [55] 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.

> 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]

> Should Scotland be an independent country? [2014 Scotland independence referendum question]

The end result was the same in both instances (no to independence), but an argument can easily 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.[2]

[2] 55.3%-44.7% in the Scotland referendum, as opposed to 59.6%-40.4% in the Québec referendum.

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 are no longer accurate, etc.), so it may be necessary to implement regular updating collection procedures.

Until very recently, the story of data analysis has only 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.[3]

We shall not discuss those challenges in detail in this module, but we urge analysts and consultants alike to be aware of their existence.

11.1.1 Survey Sampling Generalities

> The latest survey shows that 3 out of 4 people make up 75% of the world's population. [David Letterman]

While the *World Wide Web* does contain troves of data, web scraping (see Chapter 28) 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 clients and stakeholders are seeking?

A **survey** [55] 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.

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 11.1.

[3] Such as DNA storing [56], to name but one (!).

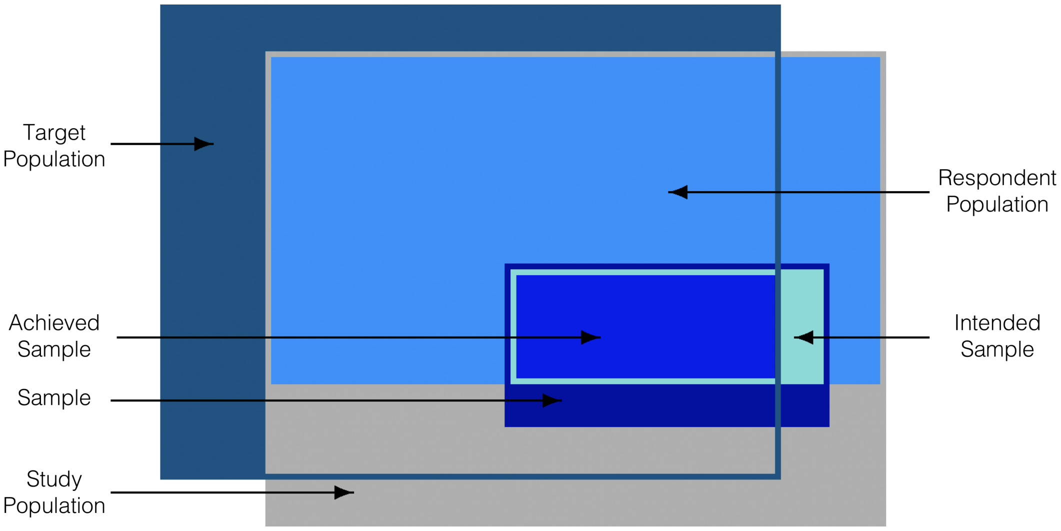


Figure 11.1: Various populations and samples in the sampling model.

- **Target population:** population for which we want to obtain information;
- **Study population** (survey population): population covered by the survey (it may be different from the target population, but ideally the two are very similar);[4] conclusions drawn from the survey results only apply to the study population;
- **Respondent population:** units of the study population that would participate in the survey if they were asked to do so; it may be different from the study population if the respondents are not representative of the study population;
- **Survey frame:** provides the means to **identify** and **communicate** with the units in the survey population; it takes the form of a list, which is linked to the population under study;
- **Intended sample:** subset of the study population targeted by the survey;
- **Achieved sample:** subset of the study population whose characteristics were in fact measured.

In general, a survey is preferred to a census if it is **expensive/laborious** to measure the characteristics of interest for each unit, or if the units are **destroyed** by measuring the characteristics.

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 do we 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.

[4] The difference may be due to the **difficulty/high cost** of data collection for some units excluded from the study population.

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.[5]

11.1.2 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 specialize 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 contra-indicated.

11.1.3 Fundamental Sampling Concepts

In general, a survey is conducted to **estimate certain attributes of a population** (statistics), such as, for example

- a **mean**;
- a **total**, or
- a **proportion**.

[5] Compare with Figure 14.4, Section 14.4.

A **population** (either target, study, or respondent) has a finite number N of members, called **units** or **items**. The **response** associated with the $j-$th unit of the population is represented by u_j.

Let $\mathcal{U} = \{u_1, \dots, u_N\}$ be a population of size $N < \infty$. If u_j represents a numerical variable,[6] the **mean**, **variance**, and **total** of the **response** in the population are respectively

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

If u_j represents a **binary variable**,[7] the **proportion** of the **response** in the population is

$$p = \frac{1}{N}\sum_{j=1}^{N} u_j.$$

We seek to estimate μ, τ, σ^2 and/or p using the values of the response variable for the units in the achieved sample $\mathcal{Y} = \{y_1, \dots, y_n\} \subseteq \mathcal{U}$. The relationship between $\mathcal{Y}$ and $\mathcal{U}$ is simple: in general, $n \ll N$ and $\forall i \in \{1, \dots, n\}, \exists! j \in \{1, \dots, N\}$ such that $y_i = u_j$.

The **empirical mean**, **empirical total**, and **empirical variance** are:

$$\overline{y}(,\hat{p}) = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2, \quad \hat{\tau} = \left(\frac{N}{n}\right)\sum_{i=1}^{n} y_i = N\overline{y}.$$

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

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

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 this scenario.

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 β. Finally, if the estimate is **unbiased**, then an approximate **95% confidence interval** (95% C.I.) for β is given by

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

[6] E.g., if u_j is the salary of the $j-$th unit in the population.

[7] E.g., 1 if the $j-$th unit earns more than \$70K per year, 0 otherwise.

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

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:

$$\text{Total Error} = \text{Measurement Error} + \text{Sampling Error} + \text{Non-response Error} + \text{Coverage Error},$$

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.[8]

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_{resp} be the attribute value in the respondent population;
- x_{study} be the attribute value in the study population, and
- x_{target} be the attribute value in the target population,

then

$$\underbrace{\overline{x} - x_{\text{target}}}_{\text{total error (TE)}} = \underbrace{(\overline{x} - \overline{x}_{\text{true}})}_{\text{meas. \& proc. error}} + \underbrace{(\overline{x}_{\text{true}} - x_{\text{resp}})}_{\text{sampling error}} + \underbrace{(x_{\text{resp}} - x_{\text{study}})}_{\text{non-response error}} + \underbrace{(x_{\text{study}} - x_{\text{target}})}_{\text{coverage error}}.$$

In an ideal scenario, TE = 0. In practice, there are two main contributions to Total Error: **sampling errors** (which are this module's main concern) 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.

[8] We sometimes also include the **processing error** in this component, due to the fact that the real value of the characteristic of interest can be affected by the data transformations performed throughout the analysis.

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.[9]

11.1.4 Data Collection Basics

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).
- Other approaches include unobtrusive direct observation; diaries to be filled (paper or electronic); omnibus surveys; email, Internet (e.g., Survey Monkey ↗), social media, etc.

11.1.5 Types of Sampling Methods

There is a large variety of methods to select sampling units from the target population.

Non-Probabilistic Sampling

Those that use subjective, non-random approaches are called **non-probabilistic sampling** (NPS) methods; these tend to be **quick, relatively inexpensive** and **convenient** in that a survey frame is not needed.

[9] This explains, in part, the impetus towards **automated data collection** and the use of **non-probabilistic sampling** methods.

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**.[10]

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 generalize the results** to the target population (except in rare, census-like situations).

NPS methods include:

- **haphazard** sampling, also known as "person 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 [57]); 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.

[10] The only component of the total error TE on which the analysts has direct control.

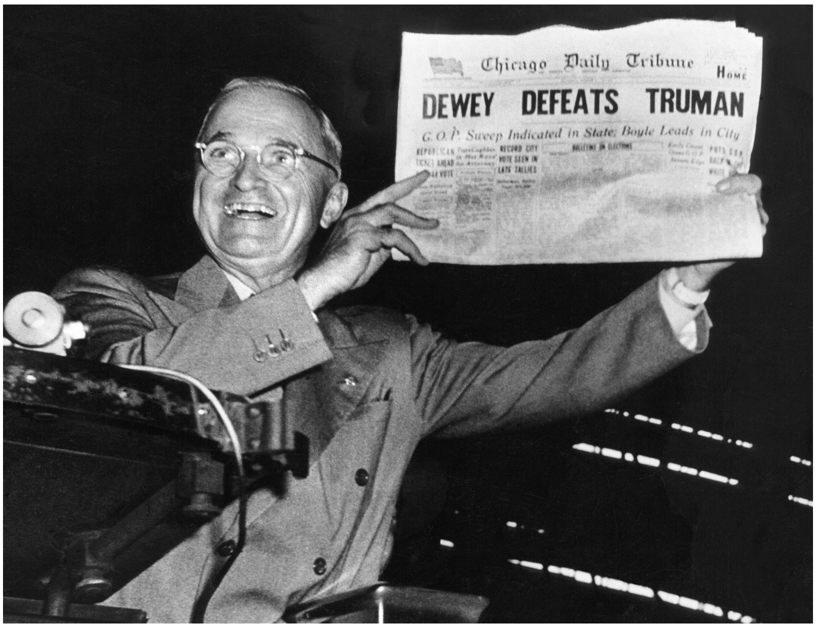


Figure 11.2: Dewey vs Truman – the aftermath: Truman victorious!

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.

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).

In this chapter, we take a deeper look at the traditional probability sample designs:

- **simple random sampling** (SRS), see Section 11.3;
- **stratified random sampling** (STS), see Section 11.4;
- **systematic random sampling** (SyS), see Section 11.7.1;
- **cluster random sampling** (CLS), see Section 11.6;
- **sampling with probability proportional to size** (PPS), see Section 11.7.2, and
- more advanced designs, see Section 11.7.

In this chapter, the analysis is made easier by assuming that the sampling error dominates the survey error, i.e., that:

- the study population is **representative** of the target population ($x_{\text{study}} \approx x_{\text{target}}$);
- the respondent population and the study population **coincide**, as are the achieved sample and the target sample ($x_{\text{resp}} \approx x_{\text{study}}$), and

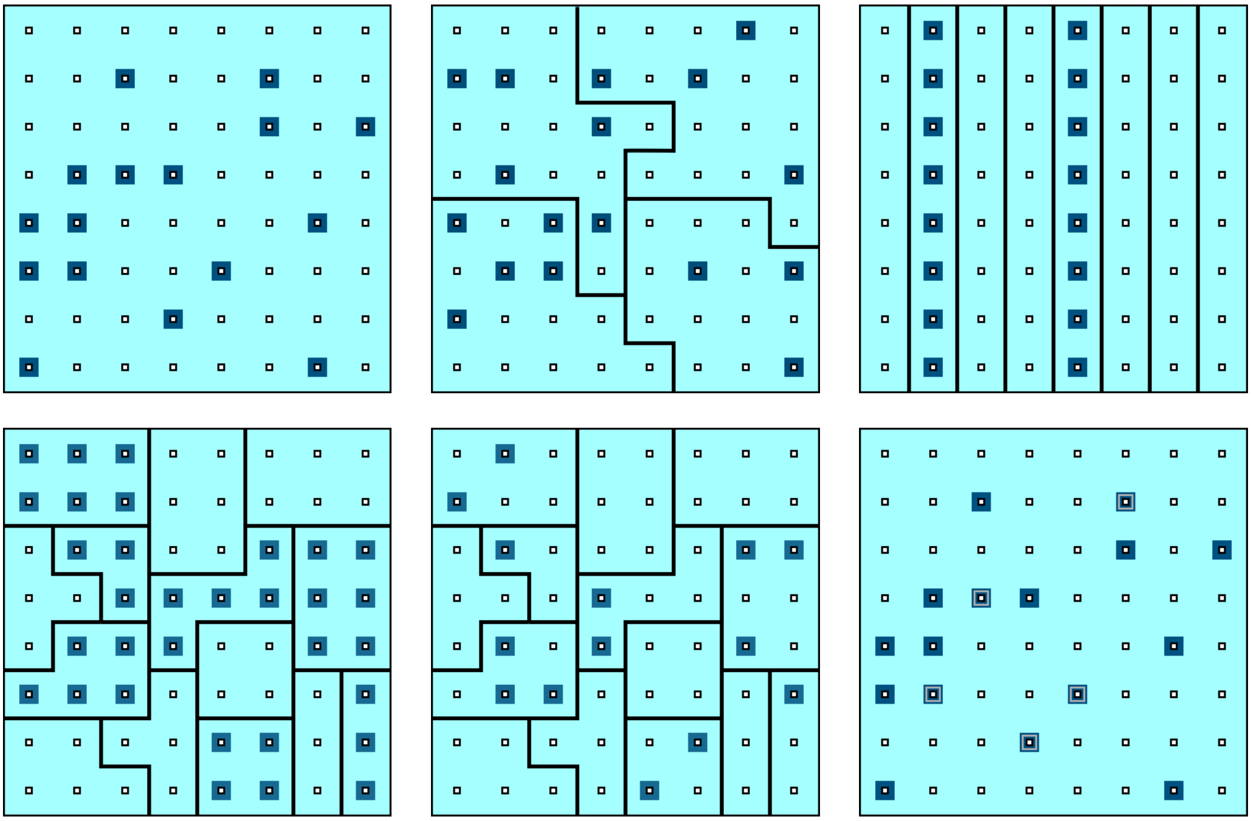


Figure 11.3: Schematics of various sampling designs (from left to right, top to bottom): simple random sampling, stratified sampling, systematic sampling, cluster sampling, multi-stage sampling, multi-phase sampling.

- the response is measured without error in the achieved sample ($\overline{x} \approx \overline{x}_{\text{true}}$).

The objective is to **control and evaluate the sampling error** ($\overline{x}_{\text{true}} - \overline{x}_{\text{resp}}$) for various random sampling designs.

11.2 Questionnaire Design

> People resist a census, but give them a profile page and they'll spend all day telling you who they are [58].

A **questionnaire** is a series of questions designed to **obtain information on a topic** from respondents. Of course, design principles vary depending on the **subject** and **method of data collection**, but it is considered good practice to test various questionnaires on **random pilot populations** before rolling it out on the study population.

11.2.1 Basic Concepts

In general, a questionnaire should:

- be as brief as possible, and free of unnecessary questions;
- be accompanied by clear and concise instructions;
- keep the respondent's interests in mind;
- emphasize confidentiality;

- keep a serious and courteous tone;
- be error-free and attractively presented;
- be clearly and precisely worded;
- be designed so that it can be answered accurately, and
- neatly arranged.

The quality of the collected data depends to a large extent on the quality of the questionnaire – **this is a practical aspect of the discipline on which much more time should be spent than on data analysis**; reputable survey firms employ **specialized teams** for questionnaire design.

There is an added challenge for Government of Canada (GoC) federal departments that are collecting and reporting information about the public and representatives of businesses or other entities, including federal public servants: see Public opinion research in the Government of Canada for details. Some of the information presented in this section will overlap with the POR guidelines, but at other times, our (generic) advice will differ.

When working with the GoC, the POR guidelines must obviously take precedence.[11]

11.2.2 Question Types

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

- **closed** questions, with a fixed number of predetermined, mutually exclusive, and collectively exhaustive answer choices (and which should always include an "Other (please specify)" category to counteract loss of expressiveness), and
- **open** questions, which are used primarily to identify common response choices for use in closed-ended questions in a subsequent questionnaire; any closed-ended question should have been an open-ended question at some point.

In everyday conversation, closed-ended questions are not appropriate:

> Asking open-ended questions is a friendly way to approach others in discussions. Knowing the difference between open and closed questions will be invaluable in your career and social life. How to ask open-ended questions, *WikiHow*

In a survey, it is rather open-ended questions that are not appropriate: closed-ended questions require less **effort** on the part of respondents, and they are generally **easier to quantify**, allowing more questions to be asked in a restricted **amount of time** and for a given **budget**.

For example, compare the two following questions.

Open question: What is the most important issue facing Ontario in 2022?

Closed-ended question: Which of these is the most important challenge for Ontario in 2022?

- economy and unemployment

[11] Fancy footwork might be required to overcome the challenges presented by the guidelines, but that is par for the course.

- impact of COVID-19
- reconciliation with indigenous communities
- taxes
- budget deficit
- the environment
- organized crime
- gang violence
- racism
- other (please specify)

However, closed-ended questions can also lead to:

- a **loss of an opportunity to test the waters** in order to obtain further clarification;
- **introducing response bias** by presenting alternatives that respondents would never have thought of, and
- a potential **loss of interest** if the choice of answers does not match a respondents' expectations.

Adding open-ended questions to the questionnaire can mitigate these risks. The use of text analysis and natural language processing methods can also help to extract the main meaning or sentiments of an answer to an open-ended question.[12]

11.2.3 Design Considerations

It is well known that the **formulation of questions** can influence the responses of a questionnaire; it is good idea to keep the following **wording considerations** in mind when developing questionnaires:

- Avoid **abbreviations** and **jargon**: "Does your organization use TTWQ practices?"
- Avoid using **complex terms** when **simpler terms** will do: "How many times have you been defenestrated?" vs. "How many times have you been thrown out a window?"
- Ensure that all respondents can answer the questions, by asking **relevant** and **appropriate-level** questions;
- Clarify the **framework**: "What is your annual income?" vs. "What was your total household income from all sources, before taxes and deductions, in 2021?
- Make the question as **accurate** as possible: "How much fuel did your moving company use last year?" (answers received: 2,500 liters, 800 gallons, $13500, more than the previous year, etc.) vs. "How much did your moving company spend on fuel last year?"
- Avoid "**double-barreled**" questions: "Do you plan to leave your car at home and take LRT to work?" vs. "Do you plan to leave your car at home? If so, do you plan to take LRT to work?", and
- Avoid **leading questions**: the always excellent *Yes, Prime Minister* gives a clear-cut example: [13] Sir Humphrey demonstrates that asking leading questions in a particular order can lead a respondent to support the reintroduction of national service:
  - Are you concerned about the number of unemployed youth?
  - Are you concerned about the increase in teenage crime?
  - Do you think there is a lack of discipline in our schools?

[12] See Chapters 29 and **??** for details and for limitations of such approaches.

[13] Which is not nearly as facetious as it appears, in the final analysis.

Yes, Prime Minister | S04xE02 | Leading Questions | *The Ministerial Broadcast*

- Do you think young people would appreciate some leadership?
- Do you think they would respond to a challenge?
- Would you support the re-introduction of national service in the UK?

The first five questions are designed and presented in such a way as to elicit support – the obvious answer to each is "yes". After this pattern of agreement, Sir Humphrey launches the crucial question, framed in such a way that it proposes national service as a supposed solution to all the above problems. In the second part of the exchange, Sir Humphrey demonstrates that another set of leading questions can lead the respondent to oppose the reintroduction of national service:

- Does the danger presented by war worry you?
- Does the arms race worry you?
- Do you think it is dangerous to arm young people and teach them to kill?
- Is it bad to force people to take up arms against their will?
- Would you oppose the reintroduction of national service?

Sir Humphrey's first four questions are deliberately designed to produce agreement. In keeping with the survey design, the fifth question does the same: a person who answers "yes" to each of these questions is necessarily opposed to the reintroduction of national service.[14]

11.2.4 Question Order

The **order** in which the questions are presented is as important as their wording. Questionnaires should be designed to be **seamless** and **follow a logical process**, from the perspective of the respondents:[15]

1. begin with an **introduction** that provides the title, topic and purpose of the survey;
2. ask for **cooperation** from respondents 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 contact address;
4. follow up with a series of **easy** and **interesting** questions to build respondent confidence;
5. group similar questions under the same heading;
6. only introduce **sensitive topics** when a relationship of trust is likely to have been established with the respondents;
7. leave some space and/or time for **additional comments**, and
8. **thank** respondents for their participation.

It is worth remembering that without a "sound sampling plan", collected data may be of such poor quality that it is impossible to use it to draw any meaningful conclusions. It is also essential to capture **demographic information** that allows classification of units into **stratas** (STS) or **clusters** (CLS); we will revisit those concepts in subsequent sections.

Example: Consider the following video.

[14] Based on an idea by Nagesh Belludi.

[15] Questionnaire design is discussed in the following references:

- Hidiroglou, M., Drew, J. and Gray, G. [1993], "A Framework for Measuring and Reducing Nonresponse in Surveys," *Survey Methodology*, v.19, n.1, pp.81-94 [59]
- Gower, A. [1994], "Questionnaire Design for Business Surveys," *Survey Methodology*, v.20, n.2, pp.125-136 [60]
- *Survey Methods and Practices* ↗ , Statistics Canada, catalogue number 12-587-X [55]

Figure 11.4: 2021 Census – How do I complete the questionnaire?

Transcription of the video

In May, your household will receive a letter to complete the 2021 Census questionnaire. On your letter, you will find a secure access code that allows you to complete the questionnaire online. Once online, you can complete the questionnaire in three easy steps. Simply log on using your secure access code, complete the questionnaire and select "Submit." If you need help or require a paper version, please call the Census Help Line. For more information or to complete the 2021 Census questionnaire, visit census.gc.ca . It's safe, quick and easy.

Message from the Chief Statistician of Canada

Thank you for taking a few minutes to participate in the 2021 Census. The information you provide is converted into statistics used by communities, businesses and governments to plan services and make informed decisions about employment, education, health care, market development and more. Your answers are collected under the authority of the Statistics Act and kept strictly confidential. By law, every household must complete a 2021 Census of Population questionnaire. Statistics Canada makes use of existing sources of information such as immigration, income tax and benefits data to ensure the least amount of burden is placed on households. The information that you provide may be used by Statistics Canada for other statistical and research purposes or may be combined with other survey or administrative data sources. Make sure you count yourself into Canada's statistical portrait, and **complete your census questionnaire today.**

Thank you,

Anil Arora
Chief Statistician of Canada

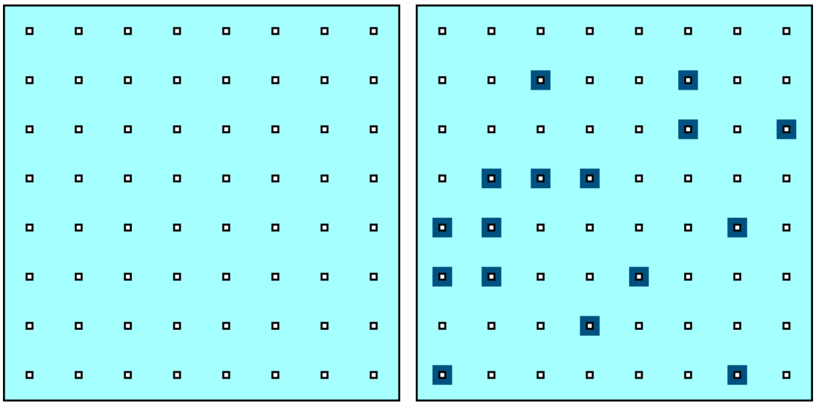


Figure 11.5: Schematics of SRS: target population (left) and sample (right).

11.3 Simple Random Sampling

Let $\mathcal{U}$ be a population composed of N units, whose responses are

$$\mathcal{U} = \{u_1, \ldots, u_N\}.$$

Suppose we are interested in the **mean** μ of this target population $\mathcal{U}$, where

$$\mu = \frac{1}{N} \sum_{j=1}^{N} u_j.$$

Since the population is of finite size, it is possible to compute μ directly... at least, in theory. In practice, we rarely have access to the response values for the entire population $\mathcal{U}$, which leads us to use **sampling methods**.

A **sample** $\mathcal{Y}$ of size n is a subset of the target population $\mathcal{U}$,

$$\mathcal{Y} \subseteq \{y_1, \ldots, y_n\} \subseteq \{u_1, \ldots, u_N\} = \mathcal{U},$$

from which we can approximate μ using the **sample mean**[16]

$$\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i.$$

A **simple random sample** (SRS) of size n is obtained by randomly selecting n units from the target population, **one at a time, without replacement**. In Figure 11.5, a SRS of size $n = 16$ is selected from a population of size $N = 64$.

At each stage of the sampling procedure, all units not yet in the sample have the same probability of being added to the sample. In an SRS, each subset of n units **has the same probability of being selected**.

How do we choose a **random** sample?

This used to be done "by hand", using tables of random numbers. Nowadays, we simply use software (SAS, R, etc.) to obtain **(pseudo-)**random samples.

Example What is the average life span, by country, in 2011?

We use the data available in the Gapminder dataset.

[16] 16: This is not the only estimator of μ.

```
library(tidyverse) # for dplyr, ggplot2
gapminder = read.csv("gapminder_SS.csv",
                     stringsAsFactors=TRUE)
gapminder <- gapminder[,c("country","year","region",
                       "continent","population",
                       "infant_mortality","fertility",
                       "gdp","life_expectancy")]
```

The structure is provided below:

```
str(gapminder)
```

```
'data.frame':   10545 obs. of  9 variables:
 $ country         : Factor w/ 185 levels "Albania","Algeria",..: 1 2 3 4
 $ year            : int  1960 1960 1960 1960 1960 1960 1960 1960 1960 196
 $ region          : Factor w/ 22 levels "Australia and New Zealand",..: 1
 $ continent       : Factor w/ 5 levels "Africa","Americas",..: 4 1 1 2 2
 $ population      : int  1636054 11124892 5270844 54681 20619075 1867396
 $ infant_mortality: num  115.4 148.2 208 NA 59.9 ...
 $ fertility       : num  6.19 7.65 7.32 4.43 3.11 4.55 4.82 3.45 2.7 5.5
 $ gdp             : num  NA 1.38e+10 NA NA 1.08e+11 ...
 $ life_expectancy : num  62.9 47.5 36 63 65.4 ...
```

A famous chart displays the relationship between 4 of the variables [61]. Our version for 2011 (built with R) can be found in Figure 11.6.

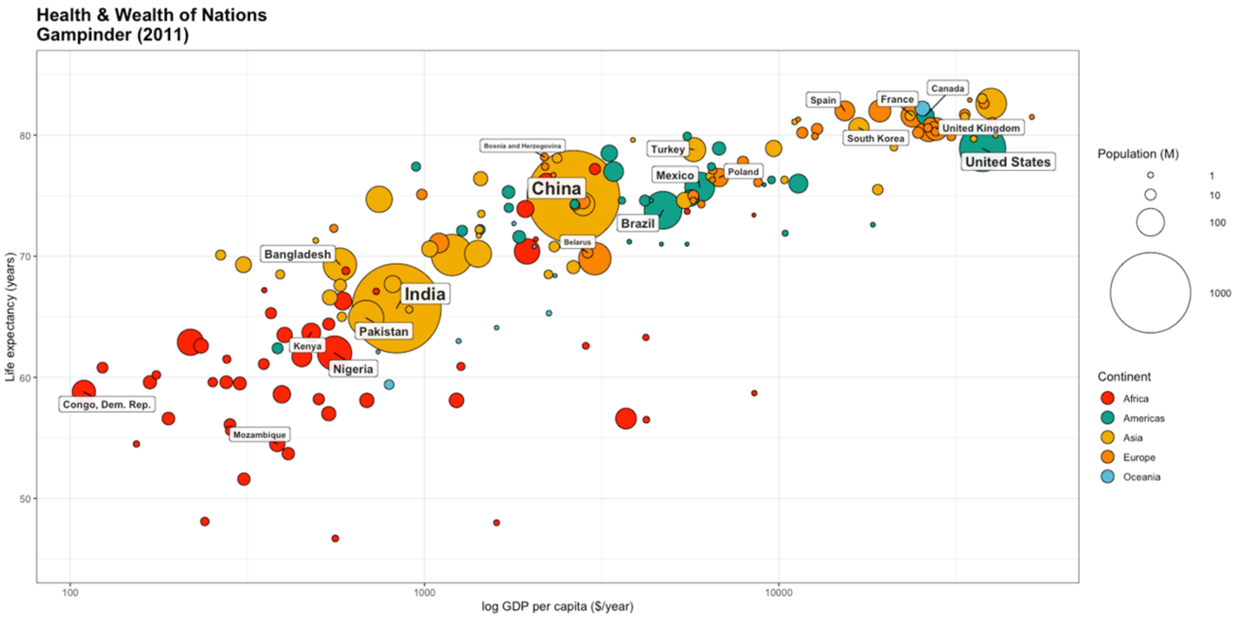


Figure 11.6: Health and wealth of nations for the 2011 Gapminder data.

We start by extracting the information of interest.

```
gapminder.SRS <- gapminder |>
  filter(year==2011) |>
  select(life_expectancy)
```

```
str(gapminder.SRS)
```

```
'data.frame':   185 obs. of  1 variable:
 $ life_expectancy: num  77.4 76.1 58.1 75.9 76 ...
```

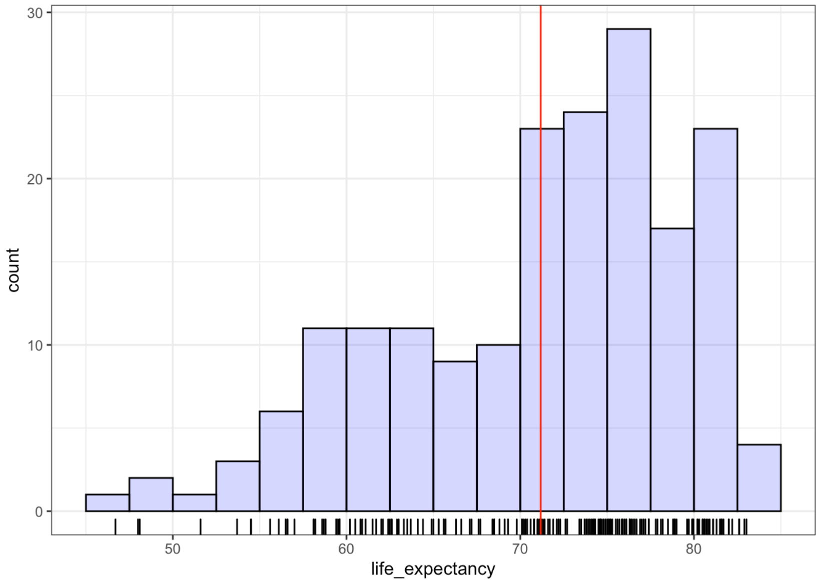
In this specific example, we know the true average life expectancy per country in 2011 (at least, for the $N = 185$ countries in the dataset).

```
mean(gapminder.SRS)
```

```
[1] 71.18
```

The distribution of the population $\mathcal{U} = \{u_1, \ldots, u_{185}\}$ is shown below (with mean in red):

```
ggplot(data=gapminder.SRS, aes(life_expectancy)) +
  geom_rug() +
  geom_histogram(col="black", fill="blue", alpha=.2,
                 breaks=seq(45, 85, by = 2.5)) +
  geom_vline(xintercept=mean(gapminder.SRS$life_expectancy),
             color="red")
```



We select a random sample of size $n = 10$ from $\mathcal{U}$. The indices are:

```
set.seed(1234) # for replicability
N = dim(gapminder.SRS)[1]
n = 10
(sample.ind = sample(1:N,n, replace=FALSE))
```

```
[1]  28  80 150 101 111 137 133 166 144 132
```

The corresponding sample $\mathcal{Y} = \{y_1, \ldots, y_{10}\}$ is obtained *via*:

```
(gapminder.SRS.n = gapminder.SRS[sample.ind,])
```

```
[1] 67.60 67.70 76.10 79.97 75.70 79.70 70.20 59.60 78.90 78.50
```

Its empirical mean $\overline{y}$ is:

```
(y.bar = mean(gapminder.SRS.n))
```

```
[1] 73.397
```

But a different sample may lead to a different estimate. Case in point, consider the following:

```
set.seed(12345) # replicability
(sample.ind = sample(1:N,n, replace=FALSE))
(gapminder.SRS.n = gapminder.SRS[sample.ind,])
(y.bar = mean(gapminder.SRS.n))
```

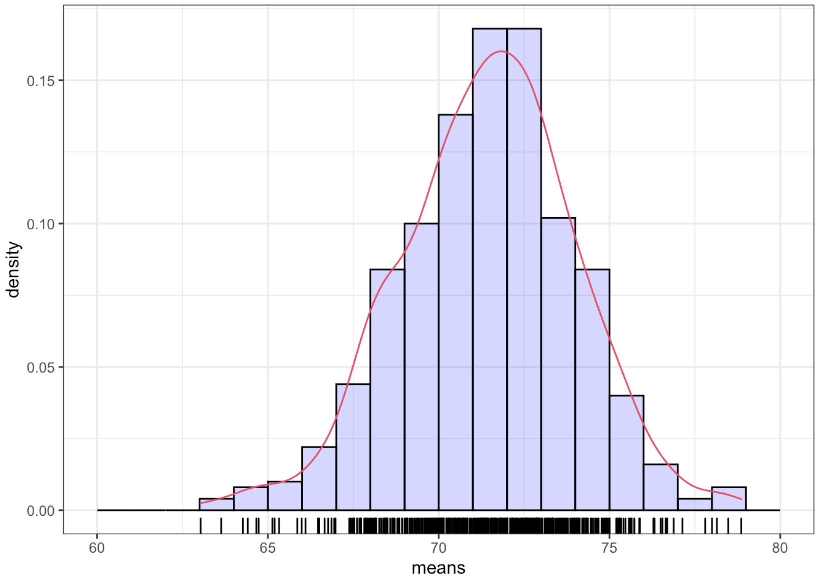
```
[1] 142  51 152  58  93  75  96   2  86 180
[1] 71.0 74.3 63.0 81.6 65.0 75.0 46.7 76.1 78.1 74.8
[1] 70.56
```

It is quite reasonable for the two estimates to be different – since each y_i in a SRS is a **random variable**, so is the mean $\overline{y}$.

The **sampling variability** explains how the estimates vary with the sample. For example, if we prepare $m = 500$ samples, each of size $n = 10$, we could obtain the empirical means below:

```
set.seed(12) # for replicability
N=dim(gapminder.SRS)[1]
n=10
m=500
means <- c()
for(k in 1:m){
 means[k] <- mean(gapminder.SRS[sample(1:N,n,
                                       replace=FALSE),])
 }

ggplot(data=data.frame(means), aes(means)) +
    geom_histogram(aes(y =..density..),
                   breaks=seq(60, 80, by = 1),
                   col="black", fill="blue", alpha=.2) +
    geom_density(col=2) + geom_rug(aes(means))
```



There is some variability, of course, but the sample means seem to congregate around the 72 mark:

```
summary(data.frame(means))
```

```
means
  Min.  :63.03
1st Qu.:69.83
Median :71.53
  Mean  :71.44
3rd Qu.:73.05
  Max.  :78.86
```

11.3.1 Basic Notions

The **population variance** σ^2 is a measure of **dispersion**, i.e., the tendency of the response values to deviate from the **population mean** μ:

$$\begin{aligned}\sigma^2 &= \frac{1}{N}\sum_{j=1}^{N}(u_j-\mu)^2 = \frac{1}{N}\sum_{j=1}^{N}(u_j^2-2u_j\mu+\mu^2)\\ &= \frac{1}{N}\left(\sum_{j=1}^{N}u_j^2-2\mu\sum_{j=1}^{N}u_j+N\mu^2\right) = \frac{1}{N}\left(\sum_{j=1}^{N}u_j^2-2N\mu^2+N\mu^2\right)\\ &= \frac{1}{N}\sum_{j=1}^{N}\left(u_j^2-N\mu^2\right) = \frac{1}{N}\sum_{j=1}^{N}u_j^2-\mu^2\end{aligned}$$

The parameters μ and σ^2 can be interpreted in terms of the **expectation** and **variance** of a random variable.

Let X be a discrete random variable whose **probability mass function** (p.m.f.) is $f(x)=P(X=x)$. Thus,

$$\text{E}[X]=\sum_x xf(x),\quad \text{V}[X]=\sum_x(x-\text{E}[X])^2f(x),\quad \text{SD}[X]=\sqrt{\text{V}[X]}.$$

For a sample of size $n = 1$ from this population, whose value is represented by the random variable Y_1, we have $f(u_j) = P(Y_1 = u_j) = \frac{1}{N}$ for $j = 1, \ldots, N$, from which we see that

$$\mathrm{E}[Y_1] = \sum_{j=1}^{N} u_j f(u_j) = \frac{1}{N}\sum_{j=1}^{N} u_j = \mu,$$

and

$$\mathrm{V}[Y_1] = \sum_{j=1}^{N}(u_j - \mu)^2 f(u_j) = \frac{1}{N}\sum_{j=1}^{N} u_j^2 - \mu^2 = \sigma^2, \quad \mathrm{SD}[Y_1] = \sqrt{\mathrm{V}[Y_1]} = \sigma.$$

In general, however, the estimator $\overline{y}$ of the population mean μ is computed using **more than one observation** – different sample sizes n could yield different values of $\overline{y}$. In order to control the sampling error associated with an SRS, one needs to know the **distribution of** $\overline{Y}$; in particular, $\mathrm{E}[\overline{Y}]$ and $\mathrm{V}[\overline{Y}]$.

If $y_1, \ldots, y_n$ are **independent and identically distributed** (i.i.d.) random variables, the **central limit theorem** (CLT) imposes

$$\overline{Y} \sim_{\text{approx.}} \mathcal{N}(\mu, \sigma^2/n).$$

Example Consider a finite population with $N = 4$ elements:

$$u_1 = 2, \quad u_2 = 0, \quad u_3 = 1, \quad u_4 = 5.$$

The population mean and variance are, respectively,

$$\mu = \frac{1}{4}(2 + 0 + 1 + 5) = 2 \quad \text{and} \quad \sigma^2 = \frac{1}{4}(2^2 + 0^2 + 1^2 + 5^2) - 2^2 = \frac{7}{2}.$$

Suppose that draw a SRS of size $n = 3$ without replacement from this population in order to approximate (estimate) the true mean μ. There are $\binom{4}{3} = 4$ such samples:

Sample	**Values**	$\overline{y}$	$P(\overline{Y} = \overline{y})$
u_1, u_2, u_3	$2, 0, 1$	1	$1/4$
u_1, u_2, u_4	$2, 0, 5$	$7/3$	$1/4$
u_1, u_3, u_4	$2, 1, 5$	$8/3$	$1/4$
u_2, u_3, u_4	$0, 1, 5$	2	$1/4$

Then

$$\mathrm{E}[\overline{Y}] = \sum_{\overline{y}} \overline{y} P(\overline{Y} = \overline{y}) = \tfrac{1}{4}\left(1 + \tfrac{7}{3} + \tfrac{8}{3} + 2\right) = 2 = \mu$$

$$\mathrm{V}[\overline{Y}] = \sum_{\overline{y}} \overline{y}^2 P(\overline{Y} = \overline{y}) - \mathrm{E}^2[\overline{Y}] = \tfrac{1}{4}\left(1^2 + \left(\tfrac{7}{3}\right)^2 + \left(\tfrac{8}{3}\right)^2 + 2^2\right) - 2^2 = \tfrac{7}{18}.$$

This is all great. . . except that $\mathrm{V}[\overline{Y}] \neq \frac{\sigma^2}{n} = \frac{7}{6}$. What is going on? ■

Here's how we can explain this discrepancy. Let $\mathcal{U} = \{u_1, \ldots, u_N\}$ be a finite population of size N. A SRS $\mathcal{Y} = \{y_1, \ldots, y_n\}$ of size n is drawn from $\mathcal{U}$ without replacement. Let Y_i be the random variable which represents the value of the $i-$th unit of the sample, respectively.

All Y_i have **identical distributions**: for any $u_j \in \mathcal{U}$, we have:[17]

$$P(Y_1 = u_j) = \frac{1}{N},$$
$$P(Y_2 = u_j) = \frac{P(Y_2 = u_j \mid Y_1 \neq u_j) \cdot P(Y_1 \neq u_j)}{P(Y_1 \neq u_j \mid Y_2 = u_j)} = \frac{\frac{1}{N-1} \cdot \frac{N-1}{N}}{1} = \frac{1}{N},$$
$$P(Y_3 = u_j) = \frac{P(Y_3 = u_j \mid Y_1, Y_2 \neq u_j) \cdot P(Y_1, Y_2 \neq u_j)}{P(Y_1, Y_2 \neq u_j \mid Y_3 = u_j)} = \frac{\frac{1}{N-2} \cdot \frac{N-2}{N-1} \cdot \frac{N-1}{N}}{1} = \frac{1}{N},$$

and so on:
$$P(Y_i = u_j) = \frac{1}{N}$$
for any $1 \leq i \leq n, 1 \leq j \leq N$, and so $\mathrm{E}[Y_i] = \mu$, $\mathrm{V}[Y_i] = \sigma^2$ for any i.

Thus, in the preceding example, we would have

$$\mathrm{E}[Y_1] = \mathrm{E}[Y_2] = \mathrm{E}[Y_3] = \mu = 2 \quad \text{and} \quad \mathrm{V}[Y_1] = \mathrm{V}[Y_2] = \mathrm{V}[Y_3] = \sigma^2 = \frac{7}{2}.$$

But the $\{Y_i\}$ are **not independent** of each other since (for example)

$$\mathrm{E}[\overline{Y}] = \mu = 2, \quad \text{but} \quad \mathrm{V}[\overline{Y}] = \mathrm{V}\left[\tfrac{Y_1+Y_2+Y_3}{3}\right] = \frac{7}{18} \neq \frac{\sigma^2}{3} = \frac{7/2}{3} = \frac{7}{6}.$$

It is in the variance that we observe a difference. The **covariance** between two (discrete) random variables X_1, X_2 is a **measure of the strength of association between** X_1 **and** X_2. If $\mathrm{E}[X_i] = \mu_i$ and $\mathrm{V}[X_i] = \sigma_i^2 < \infty$ for all i, then

$$\mathrm{Cov}[X_1, X_2] = \mathrm{E}[(X_1 - \mu_1)(X_2 - \mu_2)] = \mathrm{E}[X_1 X_2] - \mu_1 \mu_2.$$

If X_1, X_2 both take values in $\mathcal{U} = \{u_1, \ldots, u_N\}$, then their **joint expectation** is
$$\mathrm{E}[X_1 X_2] = \sum_{j=1}^{N} \sum_{k=1}^{N} u_j u_k P(X_1 = u_j, X_2 = u_k).$$

In the case where $X_1 = Y_i$ and $X_2 = Y_\ell$ (with the interpretation given before) for $1 \leq i \neq \ell \leq n$, we get

$$P(Y_i = u_j, Y_\ell = u_k) = P(Y_i = u_j) P(Y_\ell = u_k \mid Y_i = u_j) = \begin{cases} \frac{1}{N} \cdot \frac{1}{N-1} & \text{if } j \neq k \\ 0 & \text{if } j = k \end{cases}$$

But $\mathrm{E}[Y_i] = \mathrm{E}[Y_\ell] = \mu$, and so

$$\mathrm{Cov}(Y_i, Y_\ell) = \begin{cases} \frac{1}{N(N-1)} \Big[\sum_{j=1}^{N} \sum_{k=1}^{N} u_j u_k - \underbrace{\sum_{m=1}^{N} u_m^2}_{\text{doublecounting}} \Big] - \mu^2 & \text{if } i \neq \ell \\ \sigma^2 & \text{if } i = \ell \text{ (by convention)} \end{cases}$$

We use the properties $\sum u_\xi = N\mu$ and $\sum u_\xi^2 = N(\mu^2 + \sigma^2)$ to simplify the

[17] Be careful not to confuse the unit u_j with its response value u_j; we use the same notation by laziness, but they represent different concepts.

expression when $i =\neq \ell$:

$$\begin{aligned}\mathrm{Cov}(Y_i, Y_\ell) &= \frac{1}{N(N-1)}\Big[\sum_{j=1}^{N}\sum_{k=1}^{N} u_j u_k - \sum_{m=1}^{N} u_m^2 - N(N-1)\mu^2\Big] \\ &= \frac{1}{N(N-1)}\Big[\sum_{j=1}^{N} u_j\Big(\sum_{k=1}^{N} u_k\Big) - N(\sigma^2+\mu^2) - N(N-1)\mu^2\Big] \\ &= \frac{1}{N(N-1)}\Big[N\mu\sum_{j=1}^{N} u_j - N\sigma^2 - N\mu^2 - N^2\mu^2 + N\mu^2\Big] \\ &= \frac{1}{N(N-1)}\Big[N\mu\cdot N\mu - N\sigma^2 - N^2\mu^2\Big] = -\frac{\sigma^2}{N-1}.\end{aligned}$$

Using the formulas of the previous section, we thus obtain

$$\begin{aligned}\mathrm{E}[\overline{Y}] &= \mathrm{E}\Big[\frac{Y_1+\cdots+Y_n}{n}\Big] = \frac{1}{n}\mathrm{E}[Y_1+\cdots+Y_n] = \frac{1}{n}\Big(\mathrm{E}[Y_1]+\cdots\mathrm{E}[Y_n]\Big) \\ &= \frac{1}{n}\underbrace{(\mu+\cdots+\mu)}_{n\text{ times}} = \mu, \quad \text{and}\end{aligned}$$

$$\begin{aligned}\mathrm{V}[\overline{Y}] &= \mathrm{V}\Big[\frac{Y_1+\cdots+Y_n}{n}\Big] = \frac{1}{n^2}\mathrm{V}[Y_1+\cdots+Y_n] = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{\ell=1}^{n}\mathrm{Cov}(Y_i, Y_\ell) \\ &= \frac{1}{n^2}\Big[\sum_{i=1}^{n}\sigma^2 + 2\sum_{i=1}^{n}\sum_{\ell=i+1}^{n}\mathrm{Cov}(Y_i, Y_\ell)\Big] = \frac{1}{n^2}\Big[n\sigma^2 - n(n-1)\frac{\sigma^2}{N-1}\Big] \\ &= \frac{\sigma^2}{n}\left(1-\frac{n-1}{N-1}\right) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right).\end{aligned}$$

Let's go back to the above example: we have $N = 4$, $n = 3$, $\mu = 2$, and $\sigma^2 = \frac{7}{2}$. According to what we have just found, we indeed get

$$\mathrm{E}[\overline{Y}] = 2 \qquad \text{and} \qquad \mathrm{V}[\overline{Y}] = \frac{7/2}{3}\left(\frac{4-3}{4-1}\right) = \frac{7}{18}.$$

The component $\frac{N-n}{N-1}$ is the **finite population correction factor** (FPCF); it shows up because the population is not infinite. Since the SRS is constructed without replacing the units in the finite population after they have been drawn into the sample, the presence of a unit in the SRS affects the probability that another unit will also be in the SRS – **the random variables Y_i are not independent.**[18]

11.3.2 Estimators and Confidence Intervals

The estimator $\overline{y}$ is unbiased under SRS. In that case, how do we interpret the sapling variance $\mathrm{V}(\overline{y})$? Quite simply, it provides an idea of the typical distance between the **empirical mean** $\overline{y}$ and the **population mean** μ.

The **mean square error** of $\overline{y}$ under SRS is

$$\mathrm{MSE}(\overline{y}) = \mathrm{V}(\overline{y}) + (\mathrm{E}(\overline{y})-\mu)^2 = \mathrm{V}(\overline{y}) + 0 = \mathrm{V}(\overline{y}),$$

which is to say that the estimation error is entirely dominated by $\mathrm{V}(\overline{y})$.

[18] When N is "large" and the ratio $\frac{n}{N}$ is "small", the FPCF ≈ 1, in which case the situation is very similar to sampling with replacement.

When we sample with replacement,[19] the samples $y_1, \ldots, y_n$ are viewed as **independent** from one another. If they are also **indentically distributed**, we then have $\text{E}(y_i) = \mu$ and $\text{V}(y_i) = \sigma^2$, or

$$\text{E}(\overline{y}) = \mu, \quad \text{and} \quad \text{V}(\overline{y}) = \frac{\sigma^2}{n}.$$

When $n \to \infty$, the CLT states that $\overline{y} \sim_{\text{approx.}} \mathcal{N}(\mu, \sigma^2/n)$, whence

$$Z = \frac{\overline{y} - \mu}{\text{SD}(\overline{y})} = \frac{\overline{y} - \mu}{\sigma/\sqrt{n}} \sim_{\text{approx.}} \mathcal{N}(0,1).$$

Let $\alpha \in (0,1)$. Denote the $(1-\alpha)^{\text{th}}$ **quantile of a standard normal random variable** $Z \sim \mathcal{N}(0,1)$ by $z_\alpha > 0$. According to the frequentist interpretation of probability, we can expect that $\frac{\overline{y}-\mu}{\sigma/\sqrt{n}}$ will fall in the interval $(-z_{\alpha/2}, z_{\alpha/2})$ roughly $100(1-\alpha)\%$ of the time:[20]

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = P\left(-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \overline{y} - \mu \leq z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \approx 1 - \alpha.$$

The quantity

$$B_\alpha = z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = z_{\alpha/2}\text{SD}(\overline{y})$$

is the **bound on the error of estimation**, and we can build an approximate **95% confidence interval** for the mean μ:

$$\text{C.I.}(\mu; 100(1-\alpha)\%): \quad y \pm B_\alpha = y \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

However, in a SRS scenario, we are **NOT** dealing with i.i.d. random variables. How must this argument be modified when we sample without replacement from a finite population?

Sampling Context – Gapminder Data

We will illustrate the important concepts of sampling theory with the help of the 2011 Gapminder dataset, as we had done at the start of the section. In addition to average life expectancy, we are also interested in:

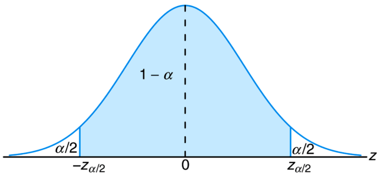
- the **total population** of the planet,
- the **average population** per country, and
- the **proportion** of countries with a population of less than 10M.

The population of 185 countries is available – it ranges from 56,641 to 1,348,174,478, with an average value $\mu = 37{,}080{,}426$.

```
gapminder.SRS <- gapminder |>
  filter(year==2011) |> select(life_expectancy,population)
str(gapminder.SRS)
summary(gapmider.SRS)
```

[19] Which is not a SRS situation.

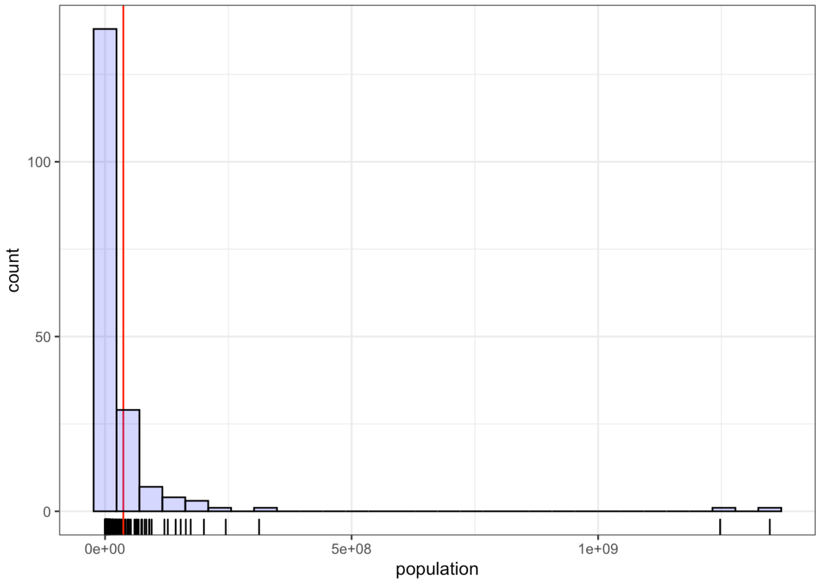
[20] The important quantiles are illustrated below:



```
'data.frame':   185 obs. of  2 variables:
 $ life_expectancy: num  77.4 76.1 58.1 75.9 76 ...
 $ population     : int  2886010 36717132 21942296 88152 41655616 2967984
```

```
life_expectancy  population
Min:    46.70       5.644e+04
1st Qu: 65.30       2.064e+06
Median :73.70       7.563e+06
Mean   :71.18       3.708e+07
3rd Qu.:77.40       2.423e+07
Max.   :83.02       1.348e+09
```

```
ggplot(data=gapminder.SRS, aes(population)) +
  geom_rug() +
  geom_vline(xintercept=mean(gapminder.SRS$population),
             color="red") +
  geom_histogram(col="black", fill="blue", alpha=.2)
```



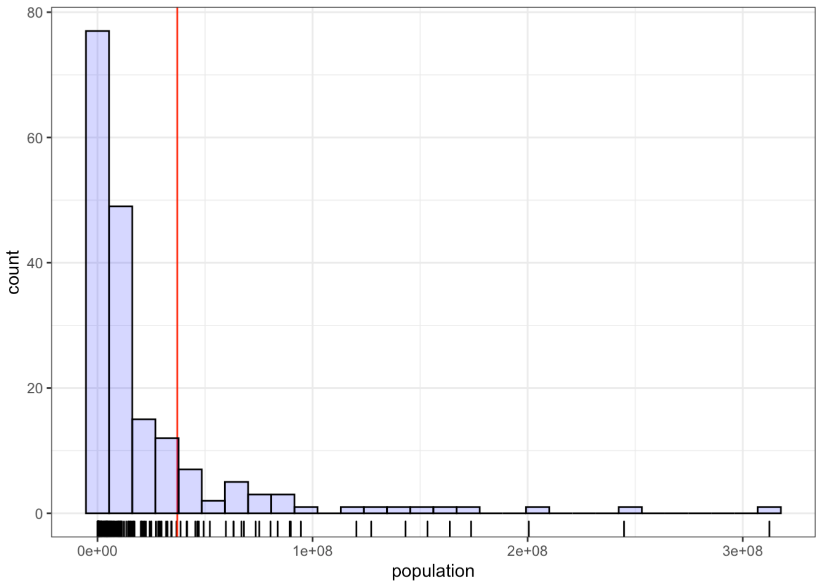
The population distribution by country is **asymmetric**, with a tail that **spreads to the right**, and two outliers (China and India). These observations will sometimes be removed from the data set.

```
gapminder.SRS.2 <- gapminder |>
  filter(year==2011) |>
  select(life_expectancy,population) |>
  filter(population<500000000)
nrow(gapminder.SRS.2)
summary(data.frame(gapminder.SRS.2$population))
```

```
[1] 183
```

```
Min.   1st Qu.  Median   Mean      3rd Qu.   Max.
56441  2061342  7355231  23301958  22242334  312390368
```

```
ggplot(data=gapminder.SRS.2, aes(population)) +
  geom_rug() +
  geom_vline(xintercept=mean(gapminder.SRS$population),
             color="red") +
  geom_histogram(col="black", fill="blue", alpha=.2)
```



The associated distribution has the same shape as the one with all countries, but the 183 populations all fall below $312,390,368$, with a mean value of $\mu = 23,301,958$.

Estimating the Mean μ

In an SRS, we have shown that the empirical mean $\overline{y}$ computed from a sample of size n is an **unbiased estimator** of the mean μ of a population of size N and variance σ^2. We have also shown that the **sampling variance** of the $\overline{y}$ estimator is

$$\mathrm{V}(\overline{y}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right).$$

What distribution can we expect $\overline{y}$ to follow? Let's go back to the example of the world population (without China and India). We produce 500 SRS samples of $n = 20$ countries from the list of $N = 183$ countries. For each sample $1 \leq i \leq 500$, we compute the **empirical mean** $\overline{y}_i$:

```
set.seed(12) # replicability
N=dim(gapminder.SRS.2)[1]
n=20
m=500

means <- c()
for(k in 1:m){
    means[k] <- mean(gapminder.SRS.2[sample(1:N,n,
                     replace=FALSE),2])
```

```
  }
```

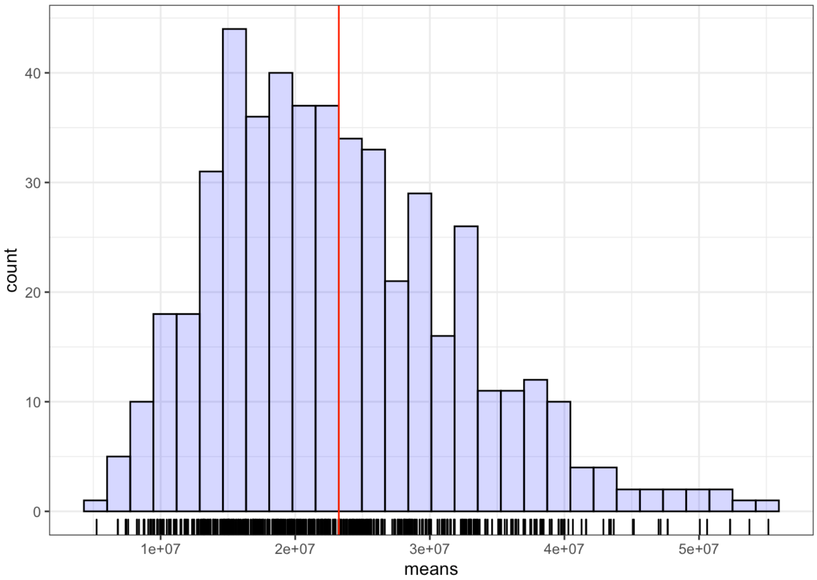
The SRS sample means are listed below:

```
summary(data.frame(means))
```

```
Min.   : 5244486
1st Qu.:16289930
Median :21986525
Mean   :23238867
3rd Qu.:28718720
Max.   :55152022
```

Their distribution (and mean) is:

```
ggplot(data=data.frame(means), aes(means)) +
  geom_rug() +
  geom_histogram(col="black", fill="blue", alpha=.2) +
  geom_vline(xintercept=mean(means), color="red")
```



Although the distribution of empirical means $\overline{y}_i$ is **asymmetric with a tail spreading to the right**, the density curve still resembles that of a **normal distribution**.

Central Limit Theorem – SRS Let $\mathcal{U} = \{u_1, \ldots, u_N\}$ be a finite population with mean μ and variance σ^2, and let $\mathcal{Y} = \{y_1, \ldots, y_n\} \subseteq \mathcal{U}$ be a simple random sample. If n and $N - n$ are both "sufficiently large", then

$$\overline{y} \sim_{\text{approx.}} \mathcal{N}\left(\text{E}(\overline{y}), \text{V}(\overline{y})\right) = \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)\right).$$

In a SRS, the **bound on the error of estimation** and the approximate **95% C.I.** are given by:

$$B_\mu = 2\sqrt{\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)} \quad \text{and} \quad P(|\overline{y}-\mu| \leq B_\mu) \approx P\left(\left|\frac{\overline{y}-\mu}{\text{SD}(\overline{y})}\right| \leq 2\right) \approx 0.9544.$$

In practice, the **population variance** σ^2 is rarely known. We usually approximate it with the **empirical variance**

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\overline{y})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - n\overline{y}^2\right], \quad \{y_i\} \text{ i.i.d.}$$

Unfortunately, s^2 is a **biased estimator** of σ^2 when the simple random sample is selected without replacement from a finite population. Indeed,

$$\begin{aligned}
\text{E}(s^2) &= \text{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\overline{y})^2\right] \\
&= \text{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\mu+\mu-\overline{y})^2\right] \\
&= \text{E}\left[\frac{1}{n-1}\left[\sum_{i=1}^{n}(y_i-\mu)^2 - n(\overline{y}-\mu)^2\right]\right] \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n}\text{E}\left[(y_i-\mu)^2\right] - n\text{E}\left[(\overline{y}-\mu)^2\right]\right] \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n}\sigma^2 - n\text{V}(\overline{y})\right] \\
&= \frac{1}{n-1}\left[n\sigma^2 - n\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)\right] = \frac{\sigma^2}{n-1}\left[n - \frac{N-n}{N-1}\right] \\
&= \frac{\sigma^2}{n-1}\left[\frac{nN-n-N+n}{N-1}\right] = \frac{\sigma^2}{n-1}\cdot\frac{N(n-1)}{N-1} = \frac{N}{N-1}\sigma^2.
\end{aligned}$$

The **unbiased estimator** of σ^2 in the SRS context is instead

$$\frac{N-1}{N}s^2$$

since

$$\text{E}\left[\frac{N-1}{N}s^2\right] = \frac{N-1}{N}\text{E}(s^2) = \frac{N-1}{N}\cdot\frac{N}{N-1}\sigma^2 = \sigma^2.$$

We can approximate the **sampling variance** by replacing σ^2 by $\frac{N-1}{N}s^2$ in the expression for $\text{V}(\overline{y})$:

$$\hat{\text{V}}(\overline{y}) = \frac{N-1}{N}\cdot\frac{s^2}{n}\left(\frac{N-n}{N-1}\right) = \frac{s^2}{n}\left(\frac{N-n}{N}\right) = \frac{s^2}{n}\left(1-\frac{n}{N}\right).$$

The **bound on the error of estimation** is thus approxiated by

$$B_\mu \approx \hat{B}_\mu = 2\sqrt{\hat{\text{V}}(\overline{y})} = 2\sqrt{\frac{s^2}{n}\left(1-\frac{n}{N}\right)},$$

from which we conclude that

$$\text{C.I.}(\mu; 0.95): \quad \overline{y} \pm 2\sqrt{\frac{s^2}{n}\left(1-\frac{n}{N}\right)}$$

is an approximate **95% confidence interval for** μ.

If the population variance σ^2 is **known**, the FPCF is $\frac{N-n}{N-1}$; if it is **unknwon**, the FPCF in $1-\frac{n}{N}$. In practice, when the **sampling rate** $\frac{n}{N}$ is below 5%, we can easily drop the FPCF ($1-\frac{n}{N} \approx 1$) without affecting the resulting quantities too greatly.

Example We draw a SRS sample $\mathcal{Y}$ of size $n = 132$ from a finite population $\mathcal{U}$ with $N = 37,444$ units. Let the sample mean and sample standard deviation be $\overline{y} = 111.3$ and $s = 16.35$, respectively. Find an approximate 95% C.I. for the population average μ.

The bound on the error of estimation is roughly

$$\hat{B}_\mu = 2\sqrt{\hat{\text{V}}(\overline{y})} = 2\sqrt{\frac{16.35^2}{132}\left(1-\frac{132}{37444}\right)} \approx 2.8,$$

which implies that

$$\text{C.I.}(\mu; 0.95) \approx 111.3 \pm 2.8;$$

the outcome is basically the same without the FPCF. ■

Example Find an approximate 95% C.I. for the average population per country in 2011 (excluding China and India) with a SRS of size $n = 20$.

We draw such a SRS sample and compute its sample mean $\overline{y}$ and sample variance s^2 (the outcomes will of course vary from one sample to another).

```
set.seed(12) # replicability
N = dim(gapminder.SRS.2)[1]
n = 20
SRS = gapminder.SRS.2[sample(1:N,n, replace=FALSE),2]
(y.bar = mean(SRS))
(s.2 = var(SRS))
```

```
[1] 35217143
[1] 5.492071e+15
```

If we do not know the population variance, the bound $\hat{B}_\mu$ and the corresponding approximate 95% C.I. for μ are given by:

```
(B.hat = 2*sqrt(s.2/n*(1-n/N)))
(IC.hat = c(y.bar-B.hat,y.bar+B.hat))
```

```
[1] 31278890
[1]  3938253 66496034
```

We can compare with the true mean μ:

```
(mu = mean(gapminder.SRS.2[,2]))
```

```
[1] 23301958
```

Sure enough, μ is in the confidence interval:

```
mu > IC.hat[1] & mu < IC.hat[2]
```

```
[1] TRUE
```

In this case, however, we also knew the population variance σ^2:

```
(sigma.2 = var(gapminder.SRS.2[,2]))
```

```
[1] 1.885224e+15
```

The bound B_μ and the corresponding approximate 95% C.I. for μ are then obtained *via*:

```
(B = 2*sqrt(sigma.2/n*(N-1)/(N-n)))
(IC = c(y.bar-B,y.bar+B))
```

```
[1] 20518160
[1] 14698984 55735303
```

Sure enough, μ is again in the confidence interval:

```
mu > IC[1] & mu < IC[2]
```

```
[1] TRUE
```

In both cases, the true mean $\mu = 23,301,958$ is contained in the confidence interval. We also notice that the C.I. when the variance σ^2 is known is contained in the 95% C.I. when the variance is not known.[21] ■

In this case, the true mean was in the confidence interval. But it could be that the 95% C.I. constructed from a sample does not contain the mean μ.

Example We repeat this procedure $m = 1000$ times (with different samples each time). If the CLT for SRS applies, how many times would we expect μ to be in the approximate 95% C.I. built from the simple random samples? Assume that σ^2 is not known.

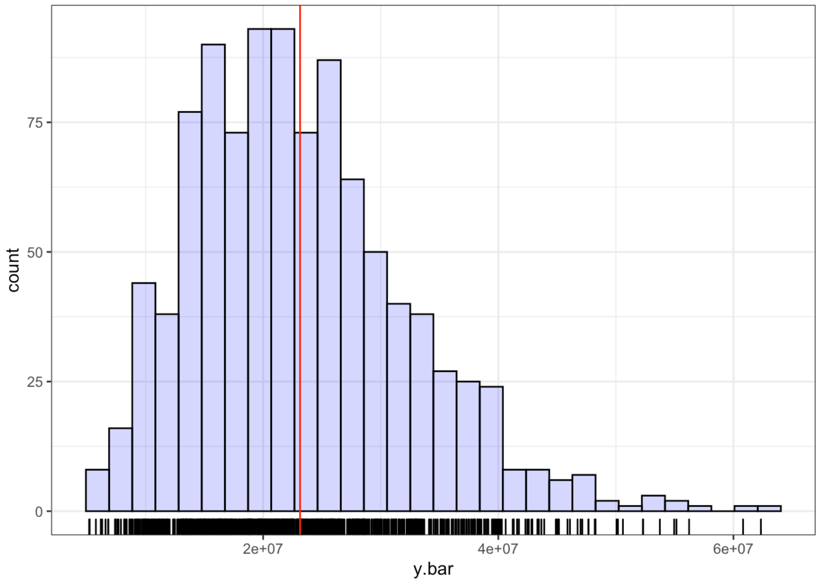
[21] Will this always be the case?

```
m = 1000
mu.in.IC = c()
y.bar = c()
for(j in 1:m){
  test = gapminder.SRS.2[sample(1:N,n, replace=FALSE),2]
  s.2 = var(test)
  B.hat = 2*sqrt(s.2/n*(1-n/N))
  y.bar[j] = mean(test)
  mu.in.IC[j] = y.bar[j]-B.hat < mu & mu < y.bar[j]+B.hat
  }
mean(mu.in.IC)
```

```
[1] 0.821
```

This is not the $\approx 95\%$ we expected; but if we increase the sample size, the proportion gets closer to 95% (see Exercises). The long tail of the population distribution for $N = 183$ units probably plays a role – the distribution of the sample measn $\overline{y}$ (with $m = 1000$ samples of size $n = 20$) does not appear to be normal.

```
ggplot(data=data.frame(y.bar), aes(y.bar)) +
  geom_rug() +
  geom_histogram(col="black", fill="blue", alpha=.2) +
  geom_vline(xintercept=mean(y.bar), color="red")
```



Estimating the Total τ

Most of the work has been done: since the **total** τ can be re-written as

$$\tau = \sum_{j=1}^{N} u_j = N\mu,$$

we can approximate τ with a SRS through the formula

$$\hat{\tau} = N\overline{y} = \frac{N}{n}\sum_{i=1}^{n} y_i.$$

This estimator is unbiased since its **expectation** is

$$\mathrm{E}(\hat{\tau}) = \mathrm{E}(N\overline{y}) = N \cdot \mathrm{E}(\overline{y}) = N\mu = \tau.$$

Its **sampling variance** is given by

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

the **bound on the estimation error** is thus

$$B_\tau = 2\sqrt{\mathrm{V}(\hat{\tau})} = 2\sqrt{N^2 \cdot \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)} = N \cdot B_\mu.$$

Since we do not usually know the true population variance σ^2 of $\mathcal{U}$, we provide an approximation by substituting σ^2 by the sample variance s^2, which needs to be multiplied by the "biased" factor $\frac{N-1}{N}$.[22] We can thus provide an **approximation of the sampling variance** using

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

this yields an **approximate bound on the estimation error** of

$$B_\tau \approx \hat{B}_\tau = 2\sqrt{\hat{\mathrm{V}}(\hat{\tau})} = 2\sqrt{N^2 \cdot \frac{s^2}{n}\left(1 - \frac{n}{N}\right)} = N \cdot \hat{B}_\mu,$$

and an **approximate 95% C.I. for** τ:

$$\text{C.I.}(\tau; 0.95): \quad \hat{\tau} \pm 2\sqrt{N^2 \cdot \frac{s^2}{n}\left(1 - \frac{n}{N}\right)}.$$

Example Consider a sample $\mathcal{Y}$ of size $n = 132$ drawn from a finite population $\mathcal{U}$ of size $N = 37,444$. Suppose the empirical mean and standard deviation of the sample are $\overline{y} = 111.3$ and $s = 16.35$, respectively. Give an approximate 95% C.I. for the total τ in $\mathcal{U}$.

The approximate bound on the error of estimation

$$\hat{B}_\tau = 2\sqrt{N^2 \cdot \hat{\mathrm{V}}(\overline{y})} = 2\sqrt{37444^2 \cdot \frac{16.35^2}{132}\left(1 - \frac{132}{37444}\right)} \approx 106,383.9643,$$

which yields

$$\text{C.I.}(\tau; 0.95) \approx 37,444 \cdot 111.3 \pm 106,383.9643 = 4,167,517.2 \pm 106,384.0,$$

or simply $(4,061,133.2; 4,273,901.2)$. ■

Example Find an approximate 95% C.I. for the population of the planet in 2011 (excluding China and India), using a SRS of size $n = 20$, assuming

[22] Recall that s^2 is a biased estimator of σ^2 in a SRS.

that

$$\overline{y} = 27{,}396{,}632 \quad \text{and} \quad \text{C.I.}(\mu; 0.95) \equiv (6{,}755{,}099; 48{,}038{,}164).$$

We have $\hat{B}_\mu \approx 48{,}038{,}164 - 27{,}396{,}632 = 20{,}641{,}532$ and

$$\hat{B}_\tau \approx N\hat{B}_\mu = 183 \cdot 20{,}641{,}532 = 3{,}777{,}400{,}356,$$

from which we conclude that

$$\text{C.I.}(\tau; 0.95): \quad N\overline{y} \pm B_\tau = 183(27{,}396{,}632) \pm 3{,}777{,}400{,}356,$$

or simply, C.I.$(\tau; 0.95) :\equiv (1{,}236{,}183{,}300; 8{,}790{,}984{,}012)$.[23] ■

Estimating a Proportion p

In a population $\mathcal{U}$ where $u_j \in \{0, 1\}$ represents a **binary response** for all $1 \leq j \leq N$,[24] the **mean** takes a particular interpretation:

$$p = \mu = \frac{1}{N}\sum_{j=1}^{N} u_j$$

is the **proportion** of the units possessing the characteristic in question.

This proportion can be estimated with a SRS *via*:

$$\hat{p} = \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \quad y_i \in \{0, 1\}.$$

It is an unbiased estimator of the proportion since its **expectation** is

$$\text{E}(\hat{p}) = \text{E}(\overline{y}) = \mu = p;$$

its **sampling variance** is

$$\text{V}(\hat{p}) = \text{V}(\overline{y}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right).$$

But $U^2 = U$ when U is a binary response, from which we see that

$$\sigma^2 = \text{E}[U^2] - \text{E}^2[U] = \text{E}[U] - \text{E}^2[U] = p - p^2 = p(1-p),$$

and so

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

The **bound on the error of estimation** is thus

$$B_p = 2\sqrt{\text{V}(\hat{p})} = 2\sqrt{\frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right)}.$$

When the population variance σ^2 is unknown (which is to say, when the true p is unknown, which is usually the case), the **sampling variation approximation** is

$$\hat{\text{V}}(\hat{p}) = \hat{\text{V}}(\overline{y}) = \frac{s^2}{n}\left(1 - \frac{n}{N}\right).$$

[23] 23: The interval is "valid", but it is perhaps too wide to be of practical use. We will discuss ways to improve the prediction in future sections.

[24] 24: For example, $u_j = 1$ when the corresponding unit possesses a certain characteristic, and $u_j = 0$ when it does not.

But recall that y_i only takes on the values 0 and 1, so that $y_i^2 = y_i$ for $1 \leq i \leq n$, from which we see that

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - n\overline{y}^2\right) = \frac{n\overline{y} - n\overline{y}^2}{n-1} = \frac{n(\hat{p} - \hat{p}^2)}{n-1} = \frac{n\hat{p}(1-\hat{p})}{n-1},$$

and

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

The **approximate estimation error bound** becomes

$$B_p \approx \hat{B}_p = 2\sqrt{\hat{\mathrm{V}}(\hat{p})} = 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}\left(1 - \frac{n}{N}\right)},$$

with the corresponding approximate 95% C.I. for p being

$$\text{C.I.}(p; 0.95): \quad \hat{p} \pm 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}\left(1 - \frac{n}{N}\right)}.$$

Example Consider a sample $\mathcal{Y}$ of size $n = 132$ drawn from a finite population $\mathcal{U}$ of size $N = 37,444$. Suppose that 25 of the observations of $\mathcal{Y}$ have a particular characteristic. Find an approximate 95% C.I. for the proportion p of the observations of $\mathcal{U}$ that possess the feature.

In this case, $\hat{p} = 25/132 \approx 0.19$. The required approximate bound is thus

$$\hat{B}_p = 2\sqrt{\hat{\mathrm{V}}(\hat{p})} = 2\sqrt{\frac{0.19(1-0.19)}{132-1}\left(1 - \frac{132}{37444}\right)} \approx 0.0684,$$

from which we get

$$\text{C.I.}(p; 0.95) \approx 0.19 \pm 0.0684 \equiv (0.121, 0.258). \quad \blacksquare$$

Example Find an approximate 95% C.I. for the proportion of countries for which the 2011 population fell below 10M, using a SRS with sample size $n = 20$.

Let's draw a SRS sample of size $n = 20$ and compute $\hat{p}$ (results will vary from one sample to when the population of a country is smaller than 10M and FALSE otherwise.

```
set.seed(1234) # replicability
N=dim(gapminder.SRS.2)[1]
n=20
thresh.10 <- gapminder.SRS.2[,2] < 10000000
SRS = thresh.10[sample(1:N,n, replace=FALSE)]
```

The proportion of countries with a population smaller than 10M in that sample is:

```
(p.hat = mean(SRS))
```

```
[1] 0.6
```

The true proportion p, amongst the $N = 185$ countries, is:

```
(p = mean(thresh.10))
```

```
[1] 0.5737705
```

If we assume that population variance is unknown, the bound $\hat{B}_p$ and the approximate 95% C.I. are given by:

```
(B.p = 2*sqrt(p.hat*(1-p.hat)/(n-1)*(1-n/N)))
(IC = c(p.hat-B.p,p.hat+B.p))
```

```
[1] 0.2121422
[1] 0.3878578 0.8121422
```

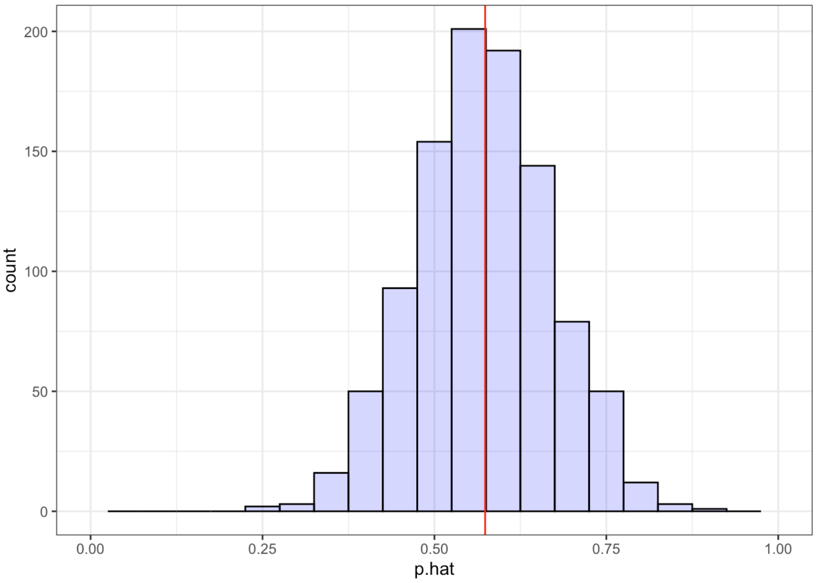
The true proportion $p \approx 0.568$ is indeed in the confidence interval. If we repeat this process $m = 1000$ times, how often is the true proportion found inside the obtained C.I.?

```
m=1000
p.in.IC = c()
p.hat = c()
for(j in 1:m){
  p.hat[j] = mean(thresh.10[sample(1:N,n, replace=FALSE)])
  B.p = 2*sqrt(p.hat[j]*(1-p.hat[j])/(n-1)*(1-n/N))
  p.in.IC[j] = p.hat[j]-B.p < p & p < p.hat[j]+B.p
  }
mean(p.in.IC)
```

```
[1] 0.963
```

Quite close to 95%, you will agree. The distribution of the $m = 1000$ estimates $\hat{p}$ is shown below, with the true proportion (red vertical line).

```
ggplot(data=data.frame(p.hat), aes(p.hat)) +
  geom_histogram(bins=21, col="black", fill="blue",
                 alpha=.2) +
  geom_vline(xintercept=mean(gapminder.SRS.2[,2]<10000000),
             color="red") + xlim(0,1)
```



11.3.3 Sample Size

Selecting an appropriate sample size is a challenge, and there is a bit of a chicken-and-egg scenario at play.

Firstly, there is a **practical** problem associated with sampling: since the cost associated with each response can be **costly** (in terms of **time**/**cost**), we often seek to **minimize the size** of the **realized** sample $\mathcal{Y}$, given a **desired error bound**.

Secondly, the SRS error bound is expressed as

$$B_\xi = 2\sqrt{\mathrm{V}(\hat{\xi})}, \quad \xi \in \{\mu, \tau, p\},$$

but the variance depends on the sample size $|\mathcal{Y}| = n$. We must then express n in terms of the (known) parameters N, σ^2, and B_ξ.

Mean μ

If we are trying to estimate the mean μ, we have:

$$B_\mu = 2\sqrt{\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)} \iff \underbrace{\frac{B_\mu^2}{4}}_{=D_\mu} = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$$

$$\frac{(N-1)D_\mu}{\sigma^2} = \frac{N-n}{n} = \frac{N}{n} - 1 \iff \frac{(N-1)D_\mu + \sigma^2}{\sigma^2} = \frac{N}{n}$$

$$\iff n_\mu = \frac{N\sigma^2}{(N-1)D_\mu + \sigma^2}.$$

However, we can only use this formula is we **know the population variance** σ^2. We could chose to use the **empirical variance** s^2 of the sample $\mathcal{Y}$ as we did when we estimated the sample variance, **but we haven't drawn $\mathcal{Y}$ from $\mathcal{U}$ yet**!

Stratagies (to obtain σ^2):

- use a **preliminary sample** (not necessarily random),
- use the empirical variance obtained in a previous study, or
- for a proportion, use a conservative estimate ($p = 0.5$).

Example Consider a finite population $\mathcal{U}$ with size $N = 37,444$. We are interested in the mean μ of the response variable in $\mathcal{U}$. In a preliminary SRS of size $n = 132$, we computed an (empirical) standard deviation of $s = 16.35$.

Using $\sigma = s$, find the minimal SRS sample size n_μ required to estimate the mean with a bound on the error of estimation at most $B_\mu = 1.7$.

We can use the formula directly to get

$$D_\mu = \frac{(1.7)^2}{4} \approx 0.73 \implies n_\mu = \frac{37444(16.35)^2}{(37444-1)(0.73)+16.35^2} = 366.39 \approx 367. \quad \blacksquare$$

Total τ

If instead, we are seeking to estimate the total τ, we have:

$$\begin{aligned} B_\tau = 2\sqrt{N^2 \cdot \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)} &\iff \underbrace{\frac{B_\tau^2}{4N^2}}_{=D_\tau} = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) \\ &\iff \frac{(N-1)D_\tau}{\sigma^2} = \frac{N-n}{n} = \frac{N}{n} - 1 \\ &\iff \frac{(N-1)D_\tau + \sigma^2}{\sigma^2} = \frac{N}{n} \\ &\iff n_\tau = \frac{N\sigma^2}{(N-1)D_\tau + \sigma^2}. \end{aligned}$$

Example Consider a finite population $\mathcal{U}$ of size $N = 37,444$. We are interested in the total τ of the response variable of $\mathcal{U}$. In a preliminary SRS of size $n = 132$, we computed an empirical standard deviation of $s = 16.35$.

Using $\sigma = s$, find the minimal SRS sample size n_τ required to estimate the total response with a bound on the error of estimation at most $B_\tau = 10000$.

We can use the formula directly to obtain

$$D_\tau = \frac{(10000)^2}{4(37444)^2} \approx 0.018 \implies n_\tau = \frac{37444(16.35)^2}{(37444-1)(0.018)+16.35^2} \approx 10706. \quad \blacksquare$$

Proportion p

If we are interested in the proportion p, we have:

$$\begin{aligned} B_p = 2\sqrt{\frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right)} &\iff \underbrace{\frac{B_p^2}{4}}_{=D_p} = \frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right) \\ &\iff \frac{(N-1)D_p}{p(1-p)} = \frac{N-n}{n} = \frac{N}{n} - 1 \\ &\iff \frac{(N-1)D_p + p(1-p)}{p(1-p)} = \frac{N}{n} \\ &\iff n_p = \frac{Np(1-p)}{(N-1)D_p + p(1-p)}. \end{aligned}$$

Example Consider a finite population $\mathcal{U}$ of size $N = 37,444$. We are interested in the proportion p of units that have a particular feature. In a preliminary SRS of size $n = 132$, we identify 25 observations possessing the feature.

Using the approximation $\sigma^2 = \frac{25}{132} \cdot \frac{107}{132}$ from the preliminary sample, find the minimal SRS sample size n_p required to estimate the true proportion with a bound on the error of estimation of at most $B_p = 0.03$.

We use the formula directly and obtain

$$D_p = \frac{(0.03)^2}{4} \approx 0.0002 \implies n_p = \frac{37444(0.189)(0.811)}{(37444-1)(0.0002) + (0.189)(0.811)} \approx 671. \quad \blacksquare$$

Example Consider a situation similar to the previous example. Using the (conservative) approximation $\sigma^2 = (0.5)^2$, find the minimal SRS sample size n_p required to estimate the true proportion with a bound on the error of estimation of at most $B_p = 0.03$.

We use the formula directly and obtain

$$D_p = \frac{(0.03)^2}{4} \approx 0.0002 \implies n_p = \frac{37444(0.5)(0.5)}{(37444-1)(0.0002) + (0.5)(0.5)} \approx 1080. \quad \blacksquare$$

11.4 Stratified Random Sampling

The theory we developed in the previous section allows us to determine the distribution of the three **unbiased** estimators $\overline{y}$, $t\hat{a}u$, and p.

For instance, we have shown that if the size N of a finite population $\mathcal{U} = \{u_1, \ldots, u_N\}$ of expectation μ and variance σ^2 and the size n of the SRS $\mathcal{Y}$ from which the estimator $\overline{y}$ is constructed are **sufficiently large**, and if moreover the responses u_j are **i.i.d.** for $1 \leq j \leq N$, then $\overline{y}$ follows **approximately** a normal distribution whose parameters are

$$\mathrm{E}(\overline{y}) = \mu \quad \text{and} \quad \mathrm{V}(\overline{y}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right).$$

The higher σ^2 is, the more the repeated SRS $\overline{y}$ values vary.

In practice, the normal approximation is:

- often **acceptable** – see average life expectancy, previous section;
- but **it is not always so**, which can lead to some challenges – cf. the C.I.$(\mu; 0.95)$ for the average population which was in fact only an 80% C.I. for a SRS of size $n = 20$ in the previous section.

In the presence of **outliers** or when n, N are **too small**, the performance of an SRS may leave something to be desired.

Example Consider a finite population with $N = 16$ elements:

$$2, 2, 2, 2, 0, 0, 0, 0, 1, 1, 1, 1, 5, 5, 5, 5.$$

The population mean and variance are, respectively:

$$\mu = \frac{1}{16}(4 \cdot 2 + 4 \cdot 0 + 4 \cdot 1 + 4 \cdot 5) = 2;$$
$$\sigma^2 = \frac{1}{16}(4 \cdot 2^2 + 4 \cdot 0^2 + 4 \cdot 1^2 + 4 \cdot 5^2) - 2^2 = \frac{7}{2}.$$

Suppose that we draw an SRS of size $n = 4$ from this population, in order to estimate the mean μ.

From what we discussed in the previous section, the expectation and sampling variance of the estimator $\overline{y}$ are, respectively:

$$\mathrm{E}(\overline{y}) = 2 \quad \text{and} \quad \mathrm{V}(\overline{y}) = \frac{\sqrt{7/2}^2}{4}\left(\frac{16-4}{16-1}\right) = \frac{7}{10}.$$

We could also restrict the sampling structure in the following manner:

1. we start by **separating the population** into 4 segments (the **strata**):

   strata 1 : 2, 2, 2, 2
   strata 2 : 0, 0, 0, 0
   strata 3 : 1, 1, 1, 1
   strata 4 : 5, 5, 5, 5

2. we then draw a SRS of size $n = 4$ **by selecting one unit per stratum**.

In such a situation (which is **NOT** a SRS$(n = 4, N = 16)$), **each achieved sample** takes the form $\{2, 0, 1, 5\}$: the empirical mean is **always** 2, and so the sampling variance **is null**.

In practice, this artificial situation rarely occurs, but if the units of the population can be grouped into **natural strata**, i.e., **sub-populations** for which:

- the response value is **homogeneous** within each stratum, but
- it is **heterogeneous** from one stratum to another, then

this approach can produce an estimator whose sampling variance **is lower** than that of the SRS estimator; as a bonus, the sample **preserves certain population structures**.

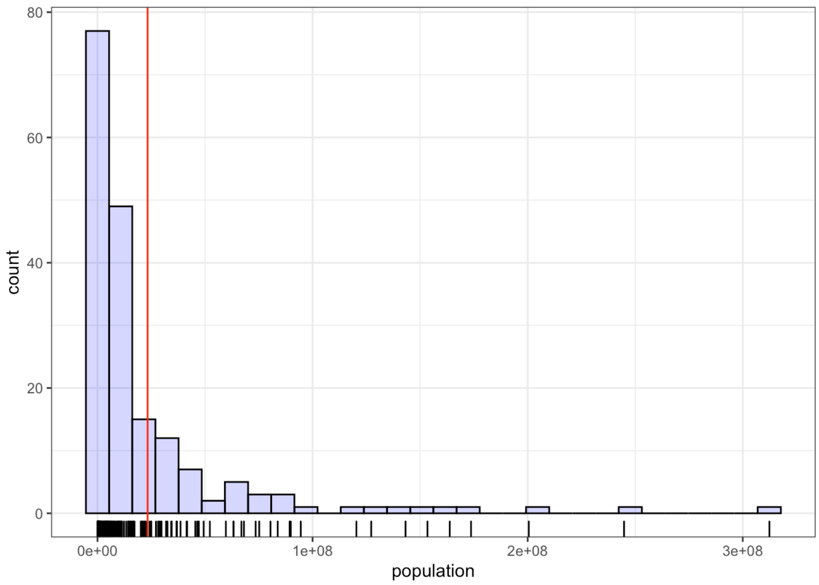
Example Find an approximate 95% C.I. for the average population per country (excluding China and India) in 2011. The population distribution in the 2011 Gapminder dataset has the following characteristics:

```
gapminder.STS <- gapminder |>
    filter(year==2011) |> select(population) |>
    filter(population < 1000000000)
summary(gapminder.STS$population)
```

```
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
    56441   2061342   7355231  23301958  22242334 312390368
```

The true average population, by country, is $\mu = 23,301,958$. Recall that the population distribution is asymmetrical:

```
N = nrow(gapminder.STS)
ggplot(data=gapminder.STS, aes(population)) +
    geom_histogram(col="black", fill="blue", alpha=.2) +
    geom_vline(xintercept=mean(gapminder.STS$population),
        color="red") + geom_rug()
```



We use the population strata $[0, 10M), [10M, 25M), [25M, 50M)$, 100M+.

```
gapminder.STS <- gapminder.STS |>
      mutate(strata = ifelse(population<10000000,"S1",
            ifelse(population<25000000,"S2",
            ifelse(population<50000000,"S3",
            ifelse(population<100000000,"S4","S5")))))

gapminder.STS <- gapminder.STS[order(gapminder.STS$population),]

gapminder.STS$strata <- as.factor(gapminder.STS$strata)
```

The number of countries in each stratum is:

```
(strata.N <- tapply(gapminder.STS$population,
                    gapminder.STS$strata, length))
```

```
 S1  S2  S3  S4  S5 
105  35  21  13   9 
```

For a sample size of $n = 20$, we use approximately n_i countries per stratum S_i:

```
strata.N/sum(strata.N)*20
```

```
        S1         S2         S3         S4         S5 
11.4754098  3.8251366  2.2950820  1.4207650  0.9836066 
```

Some practical considerations might suggest the use of a **different allocation** (more on this later). The distribution of the population by stratum has the following characteristics:

```
tapply(gapminder.STS$population, gapminder.STS$strata,
       summary)
```

```
$S1
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  56441  622957 2886010 3386819 5411377 9988846 

$S2
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
10027140 11234699 15177280 15682124 20213668 24928503 

$S3
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
25016921 29427631 34499905 36211465 41655616 49356692 

$S4
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
52237272 63268405 73517002 73841185 83787634 94501233 

$S5
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max. 
120365271 143211476 163770669 182154642 200517584 312390368 
```

In the first attempt, we draw a SRS from each stratum, using the following sizes: $(n_1, n_2, n_3, n_4, n_5) = (11, 4, 3, 1, 1)$:

```
set.seed(12345) # replicability
n=c(); n[1] = 11; n[2] = 4; n[3] = 3; n[4] = 1; n[5] = 1
ind = list()

# draw a SRS of indices in each of the 5 strata
```

```
ind[[1]] <- sample(1:strata.N[1],n[1])
ind[[2]] <- sum(strata.N[1:1]) + sample(1:strata.N[2],n[2])
ind[[3]] <- sum(strata.N[1:2]) + sample(1:strata.N[3],n[3])
ind[[4]] <- sum(strata.N[1:3]) + sample(1:strata.N[4],n[4])
ind[[5]] <- sum(strata.N[1:4]) + sample(1:strata.N[5],n[5])
```

The average population in the sample is computed as below (this value will change from one STS to another).

```
sample.STS <- gapminder.STS[unique(unlist(ind)),]
mean(sample.STS$population)
```

```
[1] 24378331
```

This naïve approach is not ideal.[25] The estimator

$$\overline{y}_{\text{STS}} = \tfrac{1}{20}(y_1 + \cdots + y_{20})$$

implies that **each observation had the same probability of being chosen**, which is not the case in reality.[26] In our second attempt, the weight of each selected observation depends on the size of the stratum.[27]

```
set.seed(123456) # replicability
cumul.n = cumsum(n); cumul.N = cumsum(strata.N)

ind = list()
ind[[1]] <- sample(1:strata.N[1],n[1])
for(j in 2:length(n)){
  ind[[j]] <- cumul.N[j-1] + sample(1:strata.N[j],n[j])
}
sample.STS <- gapminder.STS[unique(unlist(ind)),]
sample.STS = sample.STS[order(sample.STS$population),]

y.bar <- list()
y.bar[[1]] <- mean(sample.STS[1:n[1],c("population")])
for(j in 2:length(n)){
  y.bar[[j]] <- mean(sample.STS[(cumul.n[j-1]+1):cumul.n[j], c("population")])
}

y.bar.STS <- 0
for(j in 1:length(n)){
  y.bar.STS <- y.bar.STS +
            as.numeric(strata.N[j])*y.bar[[j]]
}

y.bar.STS/N
```

```
[1] 22668202
```

The estimate is very close to the actual value of μ, but a lone point estimate does not tell the full story.

[25] Despite the relative accuracy of the estimate.

[26] Remember, we are not dealing with a SRS situation.

[27] We will discuss the theoretical details in the next section.

We repeat this procedure 500 times, each time using the same size allocation $(n_1, n_2, n_3, n_4, n_5) = (9, 3, 3, 3, 2)$:

```
set.seed(12) # replicability
strata.N <- tapply(gapminder.STS$population,
                   gapminder.STS$strata, length)
cumul.N = cumsum(strata.N)

n=c(); n[1] = 9; n[2] = 3; n[3] = 3; n[4] = 3; n[5] = 2
cumul.n = cumsum(n)

m=500
means <- c()
for(k in 1:m){
    ind = list()
    ind[[1]] <- sample(1:strata.N[1],n[1])
    for(j in 2:length(n)){
      ind[[j]] <- cumul.N[j-1] +
                        sample(1:strata.N[j],n[j])
    }
    ind.STS <-unique(unlist(ind))
    sample.STS <- gapminder.STS[ind.STS,]
    sample.STS = sample.STS[order(sample.STS$population),]

    y.bar <- list()
    y.bar[[1]] <- mean(sample.STS[1:n[1],c("population")])
    for(j in 2:length(n)){
      y.bar[[j]] = mean(sample.STS[(cumul.n[j-1]+1):
                            cumul.n[j],c("population")])
    }

    y.bar.STS <- 0
    for(j in 1:length(n)){
      y.bar.STS <- y.bar.STS +
                    as.numeric(strata.N[j])*y.bar[[j]]
    }

    means[k] <- y.bar.STS/N
}
```

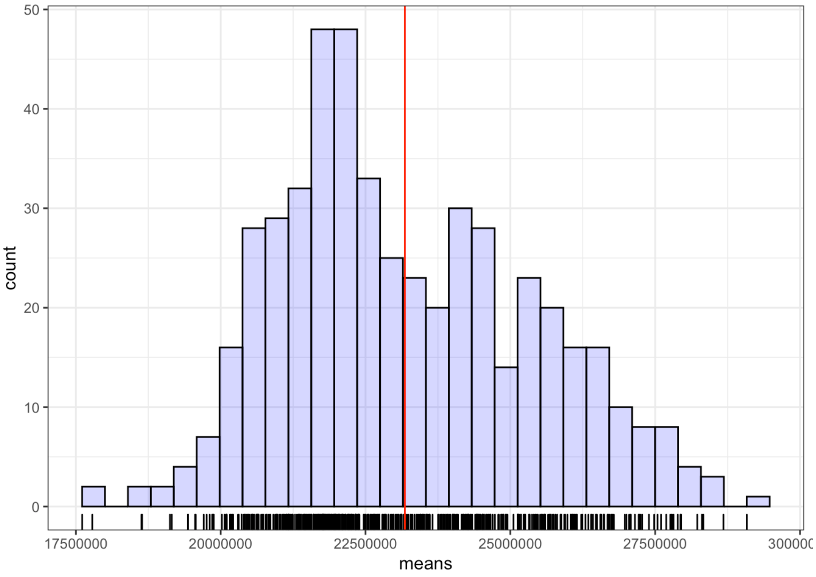
For each sample $1 \leq i \leq 500$, we then compute the **empirical mean** – their distribution has the following characteristics:

```
summary(means)
```

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
17608174 21602380 22735650 23179372 24655297 29082447
```

Finally, we plot the histogram of the STS means (with their mean in red):

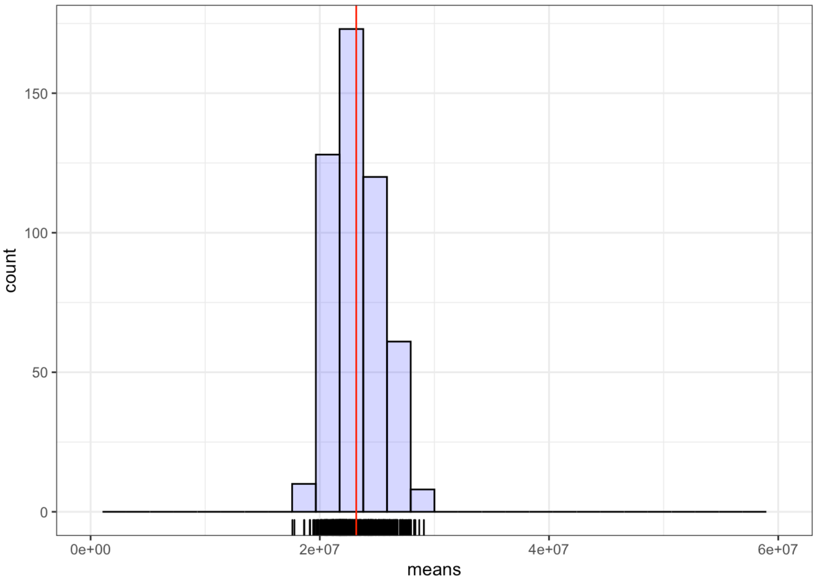
```
ggplot(data=data.frame(means), aes(means)) + geom_rug() +
    geom_histogram(col="black", fill="blue", alpha=.2) +
    geom_vline(xintercept=mean(means), color="red")
```



Not only is the shape of the distribution closer to a normal distribution, compared to the distribution of $\overline{y}$ obtained using SRS, but its variance is also much lower.

As an illustration, ccompare the following image, on the same scale as the corresponding histogram for SRS in Section 11.3.2.

```
ggplot(data=data.frame(means), aes(means)) + geom_rug() +
  geom_histogram(col="black", fill="blue", alpha=.2) +
  xlim(0,60000000) +
  geom_vline(xintercept=mean(means), color="red")
```



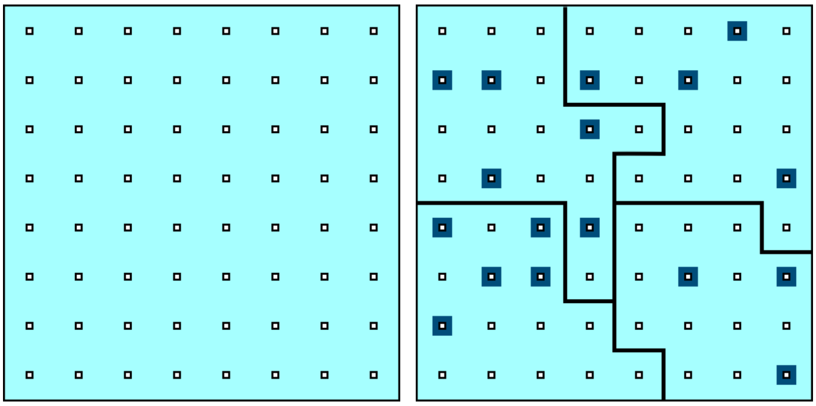


Figure 11.7: Schematics of STS: target population (left) and sample (right).

11.4.1 Estimators and Confidence Intervals

Assume that we are interested in a finite population $\mathcal{U} = \{u_1, \dots, u_N\}$, whose expectation is μ and variance is σ^2. We cover the population with M disjoint **strata**, containing, respectively, $N_1, \dots, N_M$ units:

$$\mathcal{U}_1 = \{u_{1,1}, \dots, u_{1,N_1}\}, \ \cdots \ , \ \mathcal{U}_M = \{u_{M,1}, \dots, u_{M,N_M}\},$$

with **stratum mean** and **stratum variance**

$$\mu_i = \frac{1}{N_i}\sum_{j=1}^{N_i} u_{i,j} \quad \text{and} \quad \sigma_i^2 = \frac{1}{N_i}\sum_{j=1}^{N_i} u_{i,j}^2 - \mu_i^2, \quad 1 \leq i \leq M.$$

A stratified sample $\mathcal{Y}$ of size $n \leq N$ is a subset of the target population $\mathcal{U}$, with$n_1 + \cdots + n_M = n$ and $n_i \leq N_i$ for $1 \leq i \leq M$:

$$\{\underbrace{y_{1,1}, \dots, y_{1,n_1}}_{\in \text{ strate } \mathcal{U}_1}, \dots, \underbrace{y_{M,1}, \dots, y_{M,n_m}}_{\in \text{ strate } \mathcal{U}_M}\} \subseteq \bigcup_{i=1}^{M} \mathcal{U}_i = \mathcal{U}.$$

If every sample $\mathcal{Y}_i = \{y_{i,j} \mid 1 \leq j \leq n_i\}$ is drawn from the corresponding stratum $\mathcal{U}_i$ *via* a SRS, **independently from one stratum to another**, we obtain a **stratified random sample** (STS) of size n. The **sample mean** and the **sample variance**[28] of $\mathcal{Y}_i$ are denoted by $\overline{y}_i$ and s_i^2, respectively. In a STS design, each observation in a stratum **has the same probability of being selected**, but it **may differ from one stratum to another**.

[28] Which it is important to remember is not the same thing as the "sampling variance" of an estimator.

Mean μ

In a STS, the **sample mean** of the observations of the sample $\mathcal{Y}$ falling in the stratum $\mathcal{U}_i$ is an estimator of μ_i given by

$$\overline{y}_i = \frac{1}{n_i}\sum_{\ell=1}^{n_i} y_{i,\ell}, \quad \text{where } n_i = |\mathcal{U} \cap \mathcal{Y}_i|, \ 1 \leq i \leq M.$$

The true mean μ and the **STS estimator** of μ are thus:

$$\mu = \frac{1}{N}\sum_{j=1}^{N} u_j = \frac{1}{N}\sum_{i=1}^{M}\sum_{j=1}^{N_i} u_{i,j} = \frac{1}{N}\sum_{i=1}^{M} N_i\mu_i \quad \text{and} \quad \overline{y}_{\text{STS}} = \frac{1}{N}\sum_{i=1}^{M} N_i\overline{y}_i.$$

Since $\mathcal{Y}_i$ is a SRS drawn from $\mathcal{U}_i$, we have:[29]

$$\mathrm{E}(\overline{y}_i) = \mu_i \quad \text{and} \quad \mathrm{V}(\overline{y}_i) = \frac{\sigma_i^2}{n_i}\left(\frac{N_i - n_i}{N_i - 1}\right), \quad \text{pour } 1 \leq i \leq M.$$

The **expectation** of the STS estimator is thus:

$$\mathrm{E}\left(\overline{y}_{\text{STS}}\right) = \mathrm{E}\left(\frac{1}{N}\sum_{i=1}^{M} N_i \overline{y}_i\right) = \frac{1}{N}\sum_{i=1}^{M} N_i \mathrm{E}(\overline{y}_i) = \frac{1}{N}\sum_{i=1}^{M} N_i \mu_i = \mu,$$

which is to say that $\overline{y}_{\text{STS}}$ is an **unbiased estimator** of the true mean μ for a population of size N with variance σ^2.[30]

The **sampling variance** of the estimator $\overline{y}_{\text{STS}}$ is

$$\mathrm{V}\left(\overline{y}_{\text{STS}}\right) = \mathrm{V}\left(\frac{1}{N}\sum_{i=1}^{M} N_i \overline{y}_i\right) = \frac{1}{N^2}\sum_{i=1}^{M} N_i^2 \mathrm{V}(\overline{y}_i) + \sum_{i \neq i'}^{M} N_i N_{i'} \underbrace{\mathrm{Cov}(\overline{y}_i, \overline{y}_{i'})}_{=0}$$

$$= \frac{1}{N^2}\sum_{i=1}^{M} N_i^2 \mathrm{V}(\overline{y}_i) = \frac{1}{N^2}\sum_{i=1}^{M} N_i^2 \cdot \frac{\sigma_i^2}{n_i}\left(\frac{N_i - n_i}{N_i - 1}\right).$$

Central Limit Theorem – STS If $n, N-n, n_i$, and $N_i - n_i$ are all sufficiently large, for all i, then

$$\overline{y}_{\text{STS}} \sim_{\text{approx.}} \mathcal{N}\left(\mathrm{E}(\overline{y}_{\text{STS}}), \mathrm{V}(\overline{y}_{\text{STS}})\right) = \mathcal{N}\left(\mu, \frac{1}{N^2}\sum_{i=1}^{M} N_i^2 \cdot \frac{\sigma_i^2}{n_i}\left(\frac{N_i - n_i}{N_i - 1}\right)\right).$$

In a STS, the **bound on the error of estimation** is

$$B_{\mu;\text{STS}} = 2\sqrt{\mathrm{V}(\overline{y}_{\text{STS}})} = 2\sqrt{\frac{1}{N^2}\sum_{i=1}^{M} N_i^2 \cdot \frac{\sigma_i^2}{n_i}\left(\frac{N_i - n_i}{N_i - 1}\right)}$$

and the corresponding approximate **95% C.I. for** μ is

$$\text{C.I.}_{\text{STS}}(\mu; 0.95): \quad \overline{y}_{\text{STS}} \pm B_{\mu;\text{STS}}.$$

In practice, the **population variance** σ^2 is rarely known,[31] in which case we use the **sample variance**.[32]

In each stratum, the **empirical variance** s_i^2 is

$$s_i^2 = \frac{1}{n_i - 1}\sum_{\ell=1}^{n_i}(y_{i,\ell} - \overline{y}_i)^2 = \frac{1}{n_i - 1}\left[\sum_{\ell=1}^{n_i} y_{i,\ell}^2 - n_i \overline{y}_i^2\right], \quad 1 \leq i \leq M.$$

We can then approximate the **sampling variance** in $\mathcal{U}_i$ as we did for a SRS, using

$$\hat{\mathrm{V}}(\overline{y}_i) = \frac{s_i^2}{n_i}\left(1 - \frac{n_i}{N_i}\right), \quad 1 \leq i \leq M.$$

[29] For the sake of completeness, the SRS estimator is sometimes denoted by $\overline{y}_{\text{SRS}}$.

[30] It is evidently not the one as $\overline{y}_{\text{SRS}}$ is also such an estimator.

[31] As is the **variance** σ_i^2 **in each stratum** $\mathcal{U}_i$, $1 \leq i \leq M$.

[32] And the corresponding **finite population correction factor**.

The **sampling variance** of the estimator $\overline{y}_{\text{STS}}$ is thus

$$\hat{\text{V}}(\overline{y}_{\text{STS}}) = \frac{1}{N^2}\sum_{i=1}^{M} N_i^2 \text{V}(\overline{y}_i) = \frac{1}{N^2}\sum_{i=1}^{M} N_i^2 \cdot \frac{s_i^2}{n_i}\left(1 - \frac{n_i}{N_i}\right).$$

The **bound of the estimation error** is approximated by

$$B_{\mu;\text{STS}} \approx \hat{B}_{\mu;\text{STS}} = 2\sqrt{\hat{\text{V}}(\overline{y}_{\text{STS}})} = 2\sqrt{\frac{1}{N^2}\sum_{i=1}^{M} N_i^2 \cdot \frac{s_i^2}{n_i}\left(1 - \frac{n_i}{N_i}\right)},$$

whence

$$\text{C.I.}_{\text{STS}}(\mu; 0.95): \quad \overline{y}_{\text{STS}} \pm \hat{B}_{\mu;\text{STS}} \equiv \overline{y}_{\text{STS}} \pm 2\sqrt{\frac{1}{N^2}\sum_{i=1}^{M} N_i^2 \cdot \frac{s_i^2}{n_i}\left(1 - \frac{n_i}{N_i}\right)}$$

is an **approximate 95% C.I. for μ.**

In practice, when the **stratum sampling rate** $\frac{n_i}{N_i}$ is below 5%, we can drop the FPCF in the corresponding stratum.

Example Consider a finite population $\mathcal{U}$ of size $N = 37,444$, separated in two disjoint strata $\mathcal{U}_1$ and $\mathcal{U}_2$, of respective sizes $N_1 = 21,123$ and $N_2 = 16,321$. A STS sample $\mathcal{Y}$ of size $n = 132$ is drawn from $\mathcal{U}$, with $n_1 = 82$ and $n_2 = 50$.

Suppose the empirical mean and standard deviation in $\mathcal{Y}_1$ and $\mathcal{Y}_2$ are $\overline{y}_1 = 120.7$, $\overline{y}_2 = 96.6$, $s_1 = 18.99$, and $s_2 = 14.31$, respectively. Find a 95% C.I. for the mean μ of $\mathcal{U}$.

The bound on the error of estimation is $\approx \hat{B}_{\mu;\text{STS}} = 2\sqrt{\hat{\text{V}}(\overline{y}_{\text{STS}})}$:

$$2\sqrt{\frac{21123^2}{37444^2} \cdot \frac{18.99^2}{82}\left(1 - \frac{82}{21123}\right) + \frac{16321^2}{37444^2} \cdot \frac{14.31^2}{50}\left(1 - \frac{50}{16321}\right)} \approx 2.95,$$

so $\text{C.I.}_{\text{STS}}(\mu; 0.95) \approx \left(\frac{21{,}123(120.7)}{37{,}444} + \frac{16{,}321(96.6)}{37{,}444}\right) \pm 2.95 \equiv (107.25, 113.14)$.

Example Find a 95% confidence interval for the average life expectancy by country in 2011 (including India and China), using a STS of size $n = 20$.[33]

We can basically re-use the same code:

```
LE.1 <- gapminder |> filter(year==2011) |>
    select(population,life_expectancy)
summary(LE.1)
```

```
  population          life_expectancy
 Min.   :5.644e+04   Min.   :46.70
 1st Qu.:2.064e+06   1st Qu.:65.30
 Median :7.563e+06   Median :73.70
 Mean   :3.708e+07   Mean   :71.18
 3rd Qu.:2.423e+07   3rd Qu.:77.40
 Max.   :1.348e+09   Max.   :83.02
```

[33] Stratifying using the country **populations**, as we did earlier in this section.

The average life expectancy is $\mu = 71.18$. We now prepare the strata according to the population, and we sort the observations from the smallest population to the largest:

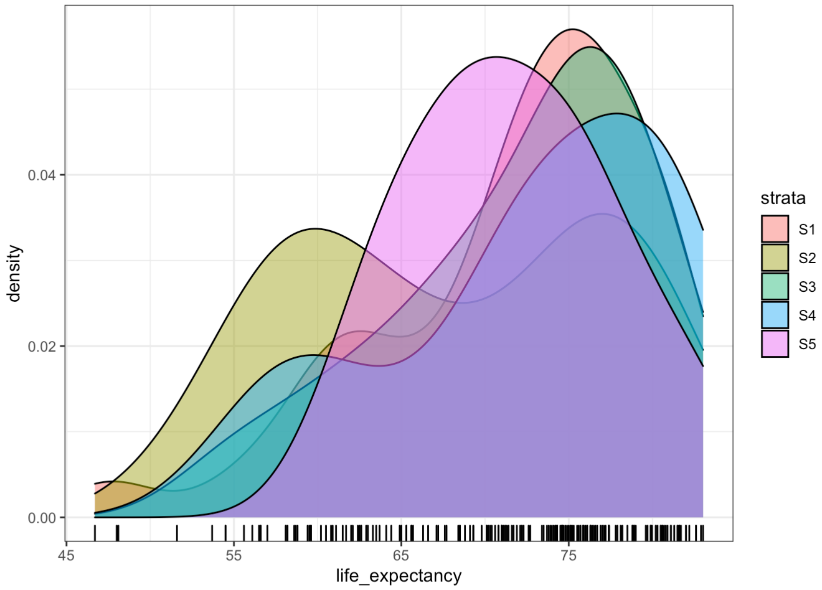
```
LE.1 <- LE.1 |> mutate(strata = ifelse(population<10000000,"S1",
    ifelse(population<25000000,"S2", ifelse(population<50000000,"S3",
    ifelse(population<100000000,"S4","S5")))))
LE.1 <- LE.1[order(LE.1$population),]
LE.1$strata <- as.factor(LE.1$strata)

# number of countries in each stratum
(strata.N <- tapply(LE.1$life_expectancy, LE.1$strata, length))
```

```
 S1  S2  S3  S4  S5
105  35  21  13  11
```

Unfortunately, the life expectancy distributions in each stratum overlap to a great extent: this is not a good sign as it suggests that a country's population is not aligned with its life expectancy.[34]

```
ggplot(LE.1,aes(x=life_expectancy,fill=strata)) +
    geom_density(alpha=0.5) + geom_rug()
```



Since there are $N = 185$ observations in the data set, a sample of size $n = 20$, allocated in such a way as to maintain the relative frequencies of the number of observations in each $\mathcal{U}_i$ (this is known as **proportional allocation**), would have the following stratum allocation:

```
N=sum(strata.N)
strata.N/sum(strata.N)*20
```

```
       S1        S2        S3        S4        S5
11.351351  3.783784  2.270270  1.405405  1.189189
```

[34] And so that the strata are heterogeneous with respect to life expectancy.

In practice, we prefer to have at least 2 observations per stratum, so we might use $(n_1, n_2, n_3, n_4, n_5) = (11, 3, 2, 2, 2)$.

```
n=c(11,3,2,2,2)
```

We select a STS sample $\mathcal{Y}$ with these characteristics *via*:

```
set.seed(123456) # replicability
cumul.n = cumsum(n)
cumul.N = cumsum(strata.N)

ind = list()
ind[[1]] <- sample(1:strata.N[1],n[1])
for(j in 2:length(n)){
    ind[[j]] <- cumul.N[j-1] + sample(1:strata.N[j],n[j])
  }

sam.LE.1 <- LE.1[unique(unlist(ind)),]
sam.LE.1 <- sam.LE.1[order(sam.LE.1$population),]
```

Next, we compute the mean $\overline{y}_i$ and the standard deviation s_i in each bucket $\mathcal{Y}_i$, $1 \leq i \leq 5$.

```
y.bar <- list()
std.dev <- list()
y.bar[[1]] <- mean(sam.LE.1[1:n[1],c("life_expectancy")])
std.dev[[1]] <- sd(sam.LE.1[1:n[1],c("life_expectancy")])

for(j in 2:length(n)){
  y.bar[[j]] <- mean(sam.LE.1[(cumul.n[j-1]+1):cumul.n[j],
                       c("life_expectancy")])
  std.dev[[j]] <- sd(sam.LE.1[(cumul.n[j-1]+1):cumul.n[j],
                       c("life_expectancy")])
}

rbind(y.bar,std.dev)
```

```
        [,1]     [,2]     [,3]     [,4]     [,5]
y.bar   70.83636 71.6     67.55    72.15    76.2
std.dev 7.551327 3.774917 18.45549 2.757716 9.050967
```

There is not much variation in the means, but the standard deviation values are all over the place: this is due to small sample sizes in some strata, and overlapping distributions of life expectancy by strata.

As we've already mentioned, **the stratification of countries by population does not align with the estimate of mean life expectancy**. We will continue the STS estimation procedure, for illustration purposes, but in practice, this is the stage at which we would require a different stratification or another sampling plan altogether.

The estimator $\overline{y}_{\text{STS}}$ is:

```
mean.LE.1 <- 0
for(j in 1:length(n)){
    mean.LE.1 <- mean.LE.1 +
        as.numeric(strata.N[j])*y.bar[[j]]
}
(mean.LE.1 <- mean.LE.1/N)
```

```
[1] 71.01902
```

This is fairly close to the true mean μ. The bound on the error of estimation $\hat{B}_{\mu;\text{STS}}$ is:

```
B=0
for(j in 1:length(n)){
    B <- B + as.numeric((strata.N[j]/N)^2*
        std.dev[[j]]^2/n[j]*(1-n[j]/strata.N[j]))
  }
(B <- 2*sqrt(B))
```

```
[1] 3.883388
```

This is quite a large bound, all things considered. The 95% C.I. is thus:

```
c(mean.LE.1 - B, mean.LE.1 + B)
```

```
[1] 67.13563 74.90241
```

Compare with the $\text{C.I.}_{\text{SRS}}(\mu; 0.95)$ obtained previously – the SRS interval was much narrower. This is no doubt due to stratification on the basis of population being a poor choice when dealing with life expectancy.

Example Find a 95% confidence interval for the average life expectancy by country in 2011 (including India and China), using a STS of size $n = 20$.[35]

We make the appropriate modifications to the code, using the following strata, say:

$$\mathcal{U}_1 = \{u_j \mid u_j < 70\}, \quad \mathcal{U}_2 = \{u_j \mid 70 \leq u_j < 80\}, \quad \mathcal{U}_3 = \{u_j \mid u_j \geq 80\}.$$

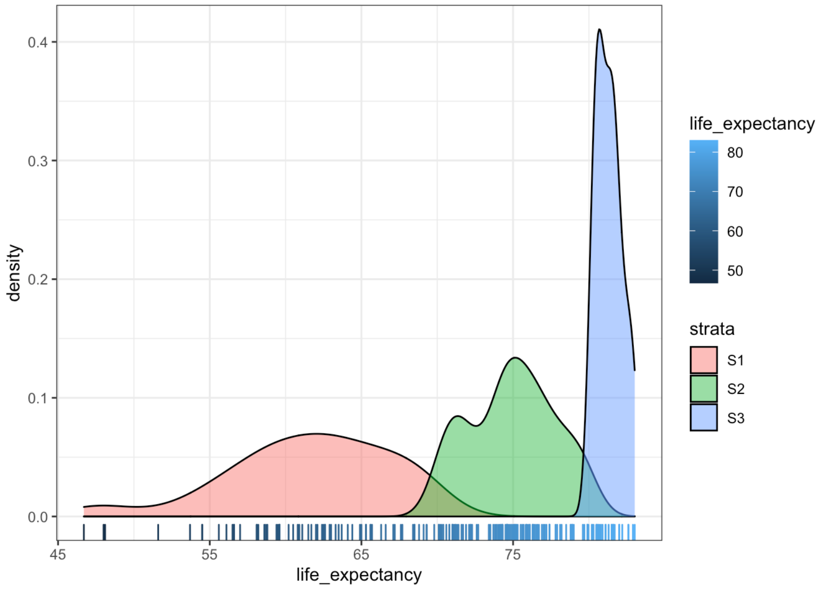
```
LE.2 <- gapminder |> filter(year==2011) |>  select(life_expectancy)
LE.2 <- LE.2 |> mutate(strata = ifelse(life_expectancy<70,"S1",
                                    ifelse(life_expectancy<80,"S2","S3")))
LE.2 <- LE.2[order(LE.2$life_expectancy),]
LE.2$strata <- as.factor(LE.2$strata)
(strata.N <- tapply(LE.2$life_expectancy, LE.2$strata, length))
```

[35] This time stratifying the data using the country **life expectations**. In general, we do not stratify with respect to the variable of interest, but with the help of auxiliary variables that are linked to the variable of interest.

```
S1 S2 S3
65 93 27
```

By construction, the life expectancy distributions do not overlap from stratum to stratum.

```
ggplot(LE.2,aes(x=life_expectancy,fill=strata)) +
    geom_density(alpha=0.5) +
    geom_rug(aes(color=life_expectancy))
```



Since there are $N = 185$ observations in the data set, with $(N_1, N_2, N_3) = (65, 93, 27)$, a sample of size $n = 20$ could be drawn according to:

```
N=sum(strata.N)
strata.N/sum(strata.N)*20
```

```
       S1        S2        S3
 7.027027 10.054054  2.918919
```

We will use $(n_1, n_2, n_3) = (7, 10, 3)$.

```
n=c(7,10,3)
```

The rest of the code runs as in the previous example.

```
cumul.n = cumsum(n)
cumul.N = cumsum(strata.N)

set.seed(123456) # replicability
ind = list()
ind[[1]] <- sample(1:strata.N[1],n[1])
```

```
for(j in 2:length(n)){
    ind[[j]] <- cumul.N[j-1] + sample(1:strata.N[j],n[j])
  }

sam.LE.2 <- LE.2[unique(unlist(ind)),]
sam.LE.2 <- sam.LE.1[order(sam.LE.2$life_expectancy),]

y.bar <- list()
std.dev <- list()
y.bar[[1]] <- mean(sam.LE.2[1:n[1],c("life_expectancy")])
std.dev[[1]] <- sd(sam.LE.2[1:n[1],c("life_expectancy")])

for(j in 2:length(n)){
  y.bar[[j]] <- mean(sam.LE.2[(cumul.n[j-1]+1):cumul.n[j],
                     c("life_expectancy")])
  std.dev[[j]] <- sd(sam.LE.2[(cumul.n[j-1]+1):cumul.n[j],
                     c("life_expectancy")])
}
```

With this sample $\mathcal{Y}$, the strata means and standard deviations are:

```
rbind(y.bar,std.dev)
```

```
        [,1]     [,2]     [,3]
y.bar   71.5     70.27    74.2
std.dev 8.469553 7.721838 7.277362
```

These quantities are more reasonable than with the previous stratification (why?), but they could change from one STS sample to the next. The values for $\overline{y}_{\text{STS}}$ and $\hat{B}_{\mu;\text{STS}}$ are:

```
mean.LE.2 <- 0
for(j in 1:length(n)){
  mean.LE.2 <- mean.LE.2 +
                 as.numeric(strata.N[j])*y.bar[[j]]
}
(mean.LE.2 <- mean.LE.2/N)

B=0
for(j in 1:length(n)){
    B <- B + as.numeric((strata.N[j]/N)^2*
                std.dev[[j]]^2/n[j]*(1-n[j]/strata.N[j]))
}

(B <- 2*sqrt(B))
```

```
[1] 71.27573
[1] 3.35133
```

The estimator is quite close to the true value $\mu = 71.18$, but it is when calculating the bound on the error of estimation that the STS approach proves its superiority. In this case, the 95% C.I. for μ is:

```
c(mean.LE.2 - B, mean.LE.2 + B)
```

```
[1] 67.92440 74.62706
```

These examples show that stratified sampling can improve SRS estimation, **but that this is not always going to be the case**.

Total τ

Most of the work has been done: since the **total** τ can be re-written as

$$\tau = \sum_{j=1}^{N} u_j = N\mu,$$

we can estimate the total with a STS using:

$$\hat{\tau}_{\text{STS}} = N\overline{y}_{\text{STS}} = \frac{N}{N}\sum_{i=1}^{M} N_i\overline{y}_i = \sum_{i=1}^{M} N_i\overline{y}_i.$$

It is an **unbiased** estimator of the total since its **expectation** is

$$\text{E}(\hat{\tau}_{\text{STS}}) = \text{E}(N\overline{y}_{\text{STS}}) = N \cdot \text{E}(\overline{y}_{\text{STS}}) = N\mu = \tau.$$

Its **sampling variance** is

$$\text{V}(\hat{\tau}_{\text{STS}}) = \text{V}(N\overline{y}_{\text{STS}}) = N^2 \cdot \text{V}(\overline{y}_{\text{STS}}) = \sum_{i=1}^{M} N_i^2 \cdot \frac{\sigma_i^2}{n_i}\left(\frac{N_i - n_i}{N_i - 1}\right),$$

assuming that we know the variance σ_i^2 in each strata $\mathcal{U}_i$, $1 \leq i \leq M$, whence the **bound on the error of estimation** is

$$B_{\tau;\text{STS}} = 2\sqrt{\text{V}(\hat{\tau}_{\text{STS}})} = 2\sqrt{\sum_{i=1}^{M} N_i^2 \cdot \frac{\sigma_i^2}{n_i}\left(\frac{N_i - n_i}{N_i - 1}\right)} = N \cdot B_{\mu;\text{STS}}.$$

Since the variances σ_i^2 are usually unknown, we often use the stratum variances s_i^2, with correction factors $\frac{N_i-1}{N_i}$, $1 \leq i \leq M$. The **approximation of the sampling variance** is thus

$$\hat{\text{V}}(\hat{\tau}_{\text{STS}}) = \hat{\text{V}}(N\overline{y}_{\text{STS}}) = \sum_{i=1}^{M} N_i^2 \cdot \frac{s_i^2}{n_i}\left(1 - \frac{n_i}{N_i}\right),$$

whence the **bound on the error of estimation** is

$$B_{\tau;\text{STS}} \approx \hat{B}_{\tau;\text{STS}} = 2\sqrt{\hat{\text{V}}(\hat{\tau}_{\text{STS}})} = 2\sqrt{\sum_{i=1}^{M} N_i^2 \cdot \frac{s_i^2}{n_i}\left(1 - \frac{n_i}{N_i}\right)} = N \cdot \hat{B}_{\mu;\text{STS}},$$

and the **approximate 95% C.I. for** τ is

$$\text{C.I.}_{\text{STS}}(\tau; 0.95): \quad \hat{\tau}_{\text{STS}} \pm \hat{B}_{\tau;\text{STS}}.$$

Example Consider a finite population $\mathcal{U}$ of size $N = 37,444$, split into two strata $\mathcal{U}_1$ and $\mathcal{U}_2$, of sizes $N_1 = 21,123$ and $N_2 = 16,321$, respectively. A STS $\mathcal{Y}$ of size $n = 132$ is drawn from $\mathcal{U}$, with $n_1 = 82$ and $n_2 = 50$.

Suppose the empirical mean and standard deviation in $\mathcal{Y}_1$ and $\mathcal{Y}_2$ are $\overline{y}_1 = 120.7$, $\overline{y}_2 = 96.6$, $s_1 = 18.99$, and $s_2 = 14.31$, respectively. Find a 95% C.I. of the total τ in $\mathcal{U}$.

The bound on the error of estimation is $\approx \hat{B}_{\tau;\text{STS}} = 2\sqrt{\hat{\text{V}}(\hat{\tau}_{\text{STS}})}$:

$$2\sqrt{21123^2 \cdot \frac{18.99^2}{82}\left(1 - \frac{82}{21123}\right) + 16321^2 \cdot \frac{14.31^2}{50}\left(1 - \frac{50}{16321}\right)} \approx 110312.3;$$

$$\text{C.I.}_{\text{STS}}(\tau; 0.95) \approx 21123(120.7)+16321(96.6)\pm 110312.3 \approx (4015842, 4236467).$$

Proportion p

If the response $u_{i,\ell} \in \{0,1\}$ represents the absence or the presence of a certain characteristic for the ℓth unit in the ith strata $\mathcal{U}_i$, the **mean**

$$p = \mu = \frac{1}{N}\sum_{i=1}^{M}\sum_{\ell=1}^{N_i} u_{i,\ell}$$

is the **proportion** of all units in $\mathcal{U}$ which posess the characteristic. This proportion can be estimated with a STS *via*

$$\hat{p}_{\text{STS}} = \frac{1}{N}\sum_{i=1}^{M} N_i\hat{p}_i, \quad \text{where } \hat{p}_i = \frac{1}{n_i}\sum_{\ell=1}^{n_i} u_{i,\ell},\ 1 \leq i \leq M.$$

This is an unbiased estimator of p since

$$\text{E}(\hat{p}_{\text{STS}}) = \text{E}(\overline{y}_{\text{STS}}) = \mu = p;$$

its **sampling variance** is:

$$\begin{aligned}\text{V}(\hat{p}_{\text{STS}}) = \text{V}(\overline{y}_{\text{STS}}) &= \frac{1}{N^2}\sum_{i=1}^{M} N_i^2 \cdot \frac{\sigma_i^2}{n_i}\left(\frac{N_i - n_i}{N_i - 1}\right) \\ &= \frac{1}{N^2}\sum_{i=1}^{M} N_i^2 \cdot \frac{p_i(1-p_i)}{n_i}\left(\frac{N_i - n_i}{N_i - 1}\right),\end{aligned}$$

where $\sigma_i^2 = p_i(1-p_i)$ is the variance of the response variable u in the stratum $\mathcal{U}_i$.

The **bound on the error of estimation** is

$$B_{p;\text{STS}} = 2\sqrt{\text{V}(\hat{p}_{\text{STS}})} = 2\sqrt{\frac{1}{N^2}\sum_{i=1}^{M} N_i^2 \cdot \frac{p_i(1-p_i)}{n_i}\left(\frac{N_i - n_i}{N_i - 1}\right)}.$$

Since the proportions p_i are not usually known, the **approximate sampling variance** is used instead:

$$\hat{\text{V}}(\hat{p}_{\text{STS}}) = \frac{1}{N^2}\sum_{i=1}^{M} N_i^2 \cdot \frac{\hat{p}_i(1-\hat{p}_i)}{n_i - 1}\left(1 - \frac{n_i}{N_i}\right).$$

The **approximate bound on the error of estimation** is thus

$$B_{p;\text{STS}} \approx \hat{B}_{p;\text{STS}} = 2\sqrt{\hat{\text{V}}(\hat{p}_{\text{STS}})} = \frac{2}{N}\sqrt{\sum_{i=1}^{M} N_i^2 \cdot \frac{\hat{p}_i(1-\hat{p}_i)}{n_i-1}\left(1-\frac{n_i}{N_i}\right)},$$

and the corresponding **approximate 95% C.I. for** p is

$$\text{C.I.}_{\text{STS}}(p;0.95): \quad \hat{p}_{\text{STS}} \pm \frac{2}{N}\sqrt{\sum_{i=1}^{M} N_i^2 \cdot \frac{\hat{p}_i(1-\hat{p}_i)}{n_i-1}\left(1-\frac{n_i}{N_i}\right)}.$$

If the sample size in a stratum is too small, we can use the conservative estimate $\hat{p}_i = 0.5$.

Example Consider a finite population $\mathcal{U}$ of size $N = 37,444$, split into two strata $\mathcal{U}_1$ and $\mathcal{U}_2$, of sizes $N_1 = 21,123$ and $N_2 = 16,321$, respectively. A STS $\mathcal{Y}$ of size $n = 132$ is drawn from $\mathcal{U}$, with $n_1 = 82$ and $n_2 = 50$.

Suppose that $n_1 = 20$ of the observations from $\mathcal{Y}_1$ and $n_2 = 5$ of the observations from $\mathcal{Y}_2$ possess a certain characteristic. Find a 95% C.I. for the proportion p of the units in $\mathcal{U}$ that possess the characteristic.

In this case, $\hat{p}_1 = 20/82 \approx 0.244$ and $\hat{p}_2 = 5/50 = 0.10$, from which we obtain

$$\hat{p}_{\text{STS}} = \frac{21123}{37444}(0.244) + \frac{16321}{37444}(0.10) = 0.181.$$

The bound on the error of estimation is thus

$$\hat{B}_p = \tfrac{2}{37444}\sqrt{21123^2\tfrac{0.244(1-0.244)}{82-1}\left(1-\tfrac{82}{21123}\right) + 16321^2\tfrac{0.1(1-0.1)}{50-1}\left(1-\tfrac{50}{16321}\right)} \approx 0.0654,$$

from which we conclude that

$$\text{C.I.}(p;0.95) \approx 0.181 \pm 0.0654 \equiv (0.116, 0.247).$$

11.4.2 Sample Size and Allocation

When determining the size of a STS sample $\mathcal{Y}$, we must also consider the problem of **allocating the number of units** n_i **in each stratum** $\mathcal{Y}_i$. If $|\mathcal{Y}_i| = n_i$, $1 \leq i \leq M$, then $n = n_1 + \cdots + n_M$. But what are the n_i?

In a STS, the sampling variance of the estimator $\overline{y}_{\text{STS}}$ is

$$\text{V}(\overline{y}_{\text{STS}}) = \frac{1}{N^2}\sum_{i=1}^{M} N_i^2 \cdot \frac{\sigma_i^2}{n_i}\left(\frac{N_i-n_i}{N_i-1}\right).$$

When $N_i \gg 1$, then $N_i \approx N_i - 1$ and so

$$\text{V}(\overline{y}_{\text{STS}}) \approx \frac{1}{N^2}\sum_{i=1}^{M} N_i^2 \cdot \frac{\sigma_i^2}{n_i}\left(\frac{N_i-n_i}{N_i}\right) = \frac{1}{N^2}\sum_{i=1}^{M} N_i^2 \cdot \frac{\sigma_i^2}{n_i} - \frac{1}{N^2}\sum_{i=1}^{M} N_i\sigma_i^2.$$

Since the sampling variance $\text{V}(\overline{y}_{\text{STS}})$ determines the bound on the error of estimation $\hat{B}_{\mu;\text{STS}}$, we can minimize the bound (and thus the error) **by minimizing the sampling variance**. The quantities N, N_i, σ_i^2, are fixed for $1 \le i \le M$; what we minimize against is the sample size n and the allocation n_i in each stratum.

The **total cost of the survey** $\tilde{C}$ can also affect the allocation. The survey budget includes the **overhead cost** (indirect costs) c_0 and the **cost per response** c_i in each stratum $\mathcal{U}_i$, $1 \le i \le M$. The total cost is thus

$$\tilde{C} = c_0 + \sum_{i=1}^{M} c_i n_i,$$

which must remain below than **available survey budget** C. The allocation problem is an optimization problem: we seek to solve

$$\arg_{(n,n_1,\ldots,n_M)} \min \mathrm{V}(\overline{y}_{\mathrm{STS}}), \quad \text{subject to } \tilde{C} \le C.$$

We use the method of **Lagrange multipliers**. The objective function becomes

$$\begin{aligned} f(n_1,\ldots,n_M,\lambda) &= \mathrm{V}(\overline{y}_{\mathrm{STS}}) + \lambda(\tilde{C} - C) \\ &= \frac{1}{N^2}\sum_{k=1}^{M} N_i^2 \cdot \frac{\sigma_k^2}{n_k} - \frac{1}{N^2}\sum_{k=1}^{M} N_k\sigma_k^2 + \lambda(c_0 + \sum_{k=1}^{M} c_k n_k - C). \end{aligned}$$

Its critical points solve

$$\begin{aligned} 0 = \frac{\partial f(n_1,\ldots,n_M,\lambda)}{\partial n_i} &= \frac{1}{N^2}\sum_{k=1}^{M} N_k^2\sigma_k^2 \frac{\partial(1/n_k)}{\partial n_i} + \lambda \sum_{k=1}^{M} c_k \frac{\partial(n_k)}{\partial n_i} \\ = -\frac{N_i^2\sigma_i^2}{N^2 n_i^2} + \lambda c_i, &\quad 1 \le i \le M, \end{aligned}$$

which is to say that

$$n_i = \frac{N_i\sigma_i}{N\sqrt{\lambda}\sqrt{c_i}}, \quad 1 \le i \le M.$$

The **strata sampling weights** w_i are

$$w_i = \frac{n_i}{n_1 + \cdots + n_M}, \quad 1 \le i \le M.$$

The **general optimal allocation** is thus

$$w_i = \frac{n_i}{n} = \frac{\frac{N_i\sigma_i}{N\sqrt{\lambda}\sqrt{c_i}}}{\sum_{k=1}^{M} \frac{N_k\sigma_k}{N\sqrt{\lambda}\sqrt{c_k}}} = \frac{\frac{N_i\sigma_i}{\sqrt{c_i}}}{\sum_{k=1}^{M} \frac{N_k\sigma_k}{\sqrt{c_k}}}, \quad 1 \le i \le M.$$

Once we have determined the size n of the sample $\mathcal{Y}$, we compute the size of the sample n_i in each $\mathcal{Y}_i$ using $w_i \cdot n$, $1 \le i \le M$. Since the product $w_i \cdot n$ is not typically an integer, we allocate $[w_i \cdot n]$ units to each $\mathcal{Y}_i$,[36] and distribute the remaining

$$n - [w_1 \cdot n] - \cdots - [w_M \cdot n]$$

units using "common sense" (while ensuring that $\tilde{C} \le C$).

[36] The **integer part** $[x]$ of x is the largest integer smaller than x.

If the cost per response in each stratum is constant, $c_1 = \cdots = c_M$, **Neyman allocation** yields the following stratum sampling weights:

$$w_i = \frac{n_i}{n} = \frac{N_i\sigma_i}{N_1\sigma_1 + \cdots + N_M\sigma_M}, \quad 1 \le i \le M.$$

If moreover the variance is the same in each stratum, $\sigma_1^2 = \cdots = \sigma_M^2$, **proportional allocation** yields the following stratum sampling weights:

$$w_i = \frac{n_i}{n} = \frac{N_i}{N_1 + \cdots + N_M} = \frac{N_i}{N}, \quad 1 \le i \le M.$$

Once the sample size and allocation have been selected, the methods in the previous section can be used to provide confidence intervals for the mean μ, for the total τ, or for a proportion p. When the variances are unknown, the usual approximations can be used.

We may at times use allocation schemes that are not necessarily **ideal** from a technical perspective, but which facilitate the preparation of reports or the dissemination of results:

$$w_i = \frac{n_i}{n} = \frac{f(N_i)}{f(N_1) + \cdots + f(N_M)}, \quad 1 \le i \le M, \ f \text{ a random function.}$$

For instance, when studying Canadian populations, we often stratify according to the provinces and use $f(x) = \sqrt{x}$; the proportional allocation and square root allocation sampling weights for the 13 Canadian jurisdictions (based on 2022 population data) are shown below.

Table 11.2: Sampling weights for Canadian provinces, under proportional allocation and square root allocation (racine, in French).

Juridiction	Prop.	Racine	Juridiction	Prop.	Racine
Ontario	38.26%	22.4%	Nouvelle-Ecosse	2.63%	5.9%
Québec	23.23%	17.4%	Nouveau-Brunswick	2.13%	5.3%
Colombie-Britannique	13.22%	13.2%	Terre-Neuve-et-Labrador	1.48%	4.4%
Alberta	11.57%	12.3%	Ile-du-Prince-Edward	0.41%	2.3%
Manitoba	3.64%	6.9%	Territoires-du-Nord-Ouest	0.12%	1.2%
Saskatchewan	3.12%	6.4%	Yukon	0.10%	1.2%
			Nunavut	0.10%	1.2%

Example Consider a finite population $\mathcal{U}$ of size $N = 37,444$, separated in two disjoint strata $\mathcal{U}_1$ and $\mathcal{U}_2$, of respective sizes $N_1 = 21,123$ and $N_2 = 16,321$. We seek to estimate the mean μ of $\mathcal{U}$ using a STS. The survey budget allows for a sample size $n = 132$.

In a preliminary study, we estimated $\sigma_1 \approx 20$ and $\sigma_2 \approx 15$. If the cost of a response in the first stratum is four times that of the cost of a response in the second stratum, find the general optimal allocation. If the response cost per stratum is constant, determine the Neyman and the proportional allocations.

In the general case, we have $c_1 = 4c_2$,

$$\frac{N_1\sigma_1}{\sqrt{c_1}} = \frac{21123(20)}{\sqrt{4c_2}} = \frac{211230}{\sqrt{c_2}}, \quad \frac{N_2\sigma_2}{\sqrt{c_2}} = \frac{16321(15)}{\sqrt{c_2}} = \frac{244815}{\sqrt{c_2}},$$

and

$$\frac{N_1\sigma_1}{\sqrt{c_1}} + \frac{N_2\sigma_2}{\sqrt{c_2}} = \frac{211230}{\sqrt{c_2}} + \frac{244815}{\sqrt{c_2}} = \frac{456045}{\sqrt{c_2}},$$

from which we conclude that

$$n_1 = 132\left(\frac{211230}{456045}\right) = 61.13 \quad \text{and} \quad n_2 = 132\left(\frac{244815}{456045}\right) = 70.87;$$

the general optimal allocation is thus $n_1 = 61$ and $n_2 = 71$.

If the cost for a response is the same in both strata, $c_1 = c_2$, then:

$$N_1\sigma_1 = 21123(20) = 422460, \quad N_2\sigma_2 = 16321(15) = 244815,$$

and

$$N_1\sigma_1 + N_2\sigma_2 = 422460 + 244815 = 667275,$$

from which we conclude that

$$n_1 = 132\left(\frac{422460}{667275}\right) = 83.57 \quad \text{and} \quad n_2 = 132\left(\frac{244815}{667275}\right) = 48.43;$$

the Neyman allocation is thus $n_1 = 84$ and $n_2 = 48$.

If we do not trust the study conducted beforehand, and we assume that the variance is constant in each stratum ($\sigma_1 = \sigma_2$), then we have

$$N_1 = 21123, \quad N_2 = 16321, \quad \text{and} \quad N_1 + N_2 = 21123 + 16321 = 37444,$$

from which we conclude that

$$n_1 = 132\left(\frac{21123}{37444}\right) = 74.46 \quad \text{and} \quad n_2 = 132\left(\frac{16321}{37444}\right) = 57.54;$$

the proportional allocation is thus $n_1 = 74$ and $n_2 = 58$. ■

Sample Size, Given a Bound on the Error of Estimation

In theory, only **analytical considerations** should influence the sample size. Recall that in a STS of size n, the sampling weight corresponding to the stratum $\mathcal{U}_i$ is $w_i = \frac{n_i}{n}$, for $1 \leq i \leq M$. When we estimate μ *via* $\overline{y}_{\text{STS}}$, the bound on the error of estimation can be written

$$B_{\mu;\text{STS}} = 2\sqrt{\frac{1}{N^2}\sum_{i=1}^{M} N_i^2 \cdot \frac{\sigma_i^2}{w_i \cdot n}\left(\frac{N_i - w_i \cdot n}{N_i - 1}\right)}.$$

We seek to express n in terms of the parameters N_i, σ_i, w_i, and $B_{\mu;\text{STS}}$. If $N_i \gg 1$,[37] then $N_i \approx N_i - 1$ and so

$$\underbrace{\frac{B^2_{\mu;\text{STS}}}{4}}_{=D_{\mu;\text{STS}}} \approx \frac{1}{N^2}\sum_{i=1}^{M} N_i^2 \cdot \frac{\sigma_i^2}{w_i \cdot n}\left(\frac{N_i - w_i \cdot n}{N_i}\right)$$

$$\Longleftrightarrow N^2 D_{\mu;\text{STS}} \approx \frac{1}{n}\left\{\sum_{i=1}^{M} \frac{N_i^2\sigma_i^2}{w_i}\right\} - \sum_{i=1}^{M} \frac{N_i^2\sigma_i^2}{w_i} \cdot \frac{w_i}{N_i}$$

$$\Longleftrightarrow \frac{N^2 D_{\mu;\text{STS}} + \sum_{i=1}^{M} N_i\sigma_i^2}{\sum_{i=1}^{M} \frac{N_i^2\sigma_i^2}{w_i}} \approx \frac{1}{n} \Longleftrightarrow n_{\mu;\text{STS}} \approx \frac{\sum_{i=1}^{M} \frac{N_i^2\sigma_i^2}{w_i}}{N^2 D_{\mu;\text{STS}} + \sum_{i=1}^{M} N_i\sigma_i^2}.$$

[37] Which is hopefully the case in practice.

Under **general optimal allocation**, the stratum sampling weights are given by

$$w_i = \frac{N_i\sigma_i}{\sqrt{c_i}}\left(\sum_{k=1}^{M}\frac{N_k\sigma_k}{\sqrt{c_k}}\right)^{-1}, \quad 1 \le i \le M,$$

and the sample size is then

$$n_{\mu;\mathrm{STS}} \approx \frac{\left(\sum_{i=1}^{M}\frac{N_i^2\sigma_i^2}{N_i\sigma_i/\sqrt{c_i}}\right) \div \left(\sum_{k=1}^{M}\frac{N_k\sigma_k}{\sqrt{c_k}}\right)^{-1}}{N^2D_{\mu;\mathrm{STS}} + \sum_{i=1}^{M}N_i\sigma_i^2} = \frac{\left(\sum_{i=1}^{M}N_i\sigma_i\sqrt{c_i}\right)\left(\sum_{i=1}^{M}\frac{N_i\sigma_i}{\sqrt{c_i}}\right)}{N^2D_{\mu;\mathrm{STS}} + \sum_{i=1}^{M}N_i\sigma_i^2}$$

Under **Neyman allocation**, the stratum sampling weights are given by

$$w_i = N_i\sigma_i\left(\sum_{k=1}^{M}N_k\sigma_k\right)^{-1}, \quad 1 \le i \le M,$$

and the sample size is then

$$n_{\mu;\mathrm{STS}} \approx \frac{\left(\sum_{i=1}^{M}\frac{N_i^2\sigma_i^2}{N_i\sigma_i}\right) \div \left(\sum_{k=1}^{M}N_k\sigma_k\right)^{-1}}{N^2D_{\mu;\mathrm{STS}} + \sum_{i=1}^{M}N_i\sigma_i^2} = \frac{\left(\sum_{i=1}^{M}N_i\sigma_i\right)^2}{N^2D_{\mu;\mathrm{STS}} + \sum_{i=1}^{M}N_i\sigma_i^2}$$

In a **proportional allocation** scenario, the stratum sampling weights are given by

$$w_i = N_i\left(\sum_{k=1}^{M}N_k\right)^{-1}, \quad 1 \le i \le M,$$

and the sample size is then

$$n_{\mu;\mathrm{STS}} \approx \frac{\left(\sum_{i=1}^{M}\frac{N_i^2\sigma_i^2}{N_i}\right) \div \left(\sum_{k=1}^{M}N_k\right)^{-1}}{N^2D_{\mu;\mathrm{STS}} + \sum_{i=1}^{M}N_i\sigma_i^2} = \frac{\sum_{i=1}^{M}N_i\sigma_i^2}{ND_{\mu;\mathrm{STS}} + \frac{1}{N}\sum_{i=1}^{M}N_i\sigma_i^2}$$

When we try to estimate the total τ using the estimator $\hat{\tau}_{\mathrm{STS}}$, we must substitute

$$D_{\mu;\mathrm{STS}} = \frac{B_{\mu;\mathrm{STS}}^2}{4} \quad \text{by} \quad D_{\tau;\mathrm{STS}} = \frac{B_{\tau;\mathrm{STS}}^2}{4N^2}.$$

When we want to estimate a proportion p using the estimator $\hat{p}_{\mathrm{STS}}$, the bound remains

$$D_{p;\mathrm{STS}} = \frac{B_{p;\mathrm{STS}}^2}{4},$$

but we have to substitute the stratum variances σ_i^2 by $p_i(1 - p_i)$. The proportions p_i can be estimated with the help of a previous study, or, **conservatively**, by using $p_i = 0.5$.

Example Consider a finite population $\mathcal{U}$ of size $N = 37,444$, separated in two disjoint strata $\mathcal{U}_1$ and $\mathcal{U}_2$, of respective sizes $N_1 = 21,123$ and $N_2 = 16,321$. We seek to estimate the mean μ of $\mathcal{U}$ using a STS, with a bound on the error of estimation of $B_{\mu;\text{STS}} = 5$. The response costs by stratum are $c_1 = 400\$$ and $c_2 = 100\$$.

In a preliminary study, we estimated $\sigma_1 \approx 20$ and $\sigma_2 \approx 15$. Determine the sample size and allocation in each of the three scenarios: general optimal allocation, Neyman allocation, and proportional allocation (in the last two cases, use $c_1 = c_2 = 100\$$).

In the general case, we have

$$\begin{aligned}
\frac{N_1\sigma_1}{\sqrt{c_1}} &= \frac{21123(20)}{\sqrt{400}} = 21123, \quad \frac{N_2\sigma_2}{\sqrt{c_2}} = \frac{16321(15)}{\sqrt{100}} = 24481.5, \\
N_1\sigma_1\sqrt{c_1} &= 21123(20)\sqrt{400} = 8449200, \quad N_2\sigma_2\sqrt{c_2} = 16321(15)\sqrt{100} = 2448150 \\
N_1\sigma_1^2 &= 21123(20)^2 = 8449200, \quad N_2\sigma_2^2 = 16321(15)^2 = 3672225, \\
\sum_{i=1}^{2}\frac{N_i\sigma_i}{\sqrt{c_i}} &= 45604.5, \quad \sum_{i=1}^{2} N_i\sigma_i\sqrt{c_i} = 10897350, \quad \sum_{i=1}^{2} N_i\sigma_i^2 = 12121425, \\
D_{\mu;\text{STS}} &= \frac{5^2}{4} = 6.25, \quad n = \frac{(10897350)(45604.5)}{(37444)^2(6.25) + 12121425} = 56.63 \approx 57 \\
n_1 &= 57\left(\frac{21123}{45604.5}\right) = 26.4 \approx 26, \quad n_2 = 57\left(\frac{24481.5}{45604.5}\right) = 30.6 \approx 31.
\end{aligned}$$

If instead the response cost per stratum is constant ($c_1 = c_2 = 100$), we have:

$$\begin{aligned}
N_1\sigma_1 &= 21123(20) = 422460, \quad N_2\sigma_2 = 16321(15) = 244815, \\
N_1\sigma_1^2 &= 21123(20)^2 = 8449200, \quad N_2\sigma_2^2 = 16321(15)^2 = 3672225, \\
\sum_{i=1}^{2} N_i\sigma_i &= 667275, \quad \sum_{i=1}^{2} N_i\sigma_i^2 = 12121425, \\
D_{\mu;\text{STS}} &= \frac{5^2}{4} = 6.25, \quad n = \frac{(667275)^2}{(37444)^2(6.25) + 12121425} = 50.74 \approx 51 \\
n_1 &= 51\left(\frac{422460}{667275}\right) = 32.30 \approx 32, \quad n_2 = 51\left(\frac{244815}{667275}\right) = 18.71 \approx 19.
\end{aligned}$$

It turns out that the exact value of $c_1 = c_2$ does not come into play.

If we look for a proportional allocation, we still have

$$\begin{aligned}
N_1\sigma_1 &= 21123(20) = 422460, \quad N_2\sigma_2 = 16321(15) = 244815, \\
N_1\sigma_1^2 &= 21123(20)^2 = 8449200, \quad N_2\sigma_2^2 = 16321(15)^2 = 3672225, \\
\sum_{i=1}^{2} N_i\sigma_i &= 667275, \quad \sum_{i=1}^{2} N_i\sigma_i^2 = 12121425, \\
D_{\mu;\text{STS}} &= \frac{5^2}{4} = 6.25, \quad n = \frac{12121425}{37444(6.25) + \frac{12121425}{37444}} = 51.72 \approx 52 \\
n_1 &= 52\left(\frac{21123}{37444}\right) = 29.33 \approx 29, \quad n_2 = 52\left(\frac{16321}{37444}\right) = 22.67 \approx 23.
\end{aligned}$$

The exact value of $c_1 = c_2$ also does not come into play. ■

Sample Size, Given a Budget

In practice, however, it is often **budgetary considerations** that play the most important role in sample size selection.

In a STS of size n, the stratum sampling weights are $w_i = \frac{n_i}{n}$, for $1 \leq i \leq M$. In this case, we seek to **maximize the size** n **allowed by the survey budget** C:

$$C = c_0 + \sum_{i=1}^{M} c_i n_i = c_0 + n \sum_{i=1}^{M} c_i w_i \implies n = \frac{C - c_0}{\sum_{i=1}^{M} c_i w_i}.$$

In a **general optimal allocation** scenario, we have

$$w_i = \frac{N_i \sigma_i}{\sqrt{c_i}} \left(\sum_{k=1}^{M} \frac{N_k \sigma_k}{\sqrt{c_k}} \right)^{-1}, \quad 1 \leq i \leq M,$$

from which we see that

$$c_i w_i = c_i \cdot \frac{N_i \sigma_i}{\sqrt{c_i}} \left(\sum_{k=1}^{M} \frac{N_k \sigma_k}{\sqrt{c_k}} \right)^{-1} = N_i \sigma_i \sqrt{c_i} \left(\sum_{k=1}^{M} \frac{N_k \sigma_k}{\sqrt{c_k}} \right)^{-1}, \quad 1 \leq i \leq M;$$

the sample size is then

$$n_{\text{STS}} = (C - c_0) \left(\sum_{i=1}^{M} \frac{N_i \sigma_i}{\sqrt{c_i}} \right) \left(\sum_{i=1}^{M} N_i \sigma_i \sqrt{c_i} \right)^{-1}.$$

In a **Neyman allocation** or **proportional allocation** scenario, the sample weights are

$$w_i = N_i \sigma_i \left(\sum_{k=1}^{M} N_k \sigma_k \right)^{-1} \quad 1 \leq i \leq M,$$

from which we see that

$$c_i w_i = c \cdot N_i \sigma_i \left(\sum_{k=1}^{M} N_k \sigma_k \right)^{-1} \quad 1 \leq i \leq M;$$

the sample size is then

$$n_{\text{STS}} = (C - c_0) \left(\sum_{i=1}^{M} N_i \sigma_i \right) \left(c \sum_{i=1}^{M} N_i \sigma_i \right)^{-1} = \frac{C - c_0}{c}.$$

Example Consider a finite population $\mathcal{U}$ of size $N = 37,444$, separated in two disjoint strata $\mathcal{U}_1$ and $\mathcal{U}_2$, of respective sizes $N_1 = 21,123$ and $N_2 = 16,321$. We seek to estimate the mean μ of $\mathcal{U}$ using a STS. The budget for the study is $C = 20,000\$$, minus $c_0 = 4,000\$$ for overhead costs. The cost of a response in each stratum are $c_1 = 400\$$ and $c_2 = 100\$$, respectively.

In a preliminary study, we estimate $\sigma_1 = 20$ and $\sigma_2 = 15$. Determine the sample size and allocation in each of the three scenarios: general optimal allocation, Neyman allocation, and proportional allocation (in the last two cases, use $c_1 = c_2 = 100\$$).

In the general case, we have

$$\begin{aligned}
\frac{N_1\sigma_1}{\sqrt{c_1}} &= \frac{21123(20)}{\sqrt{400}} = 21123, \quad \frac{N_2\sigma_2}{\sqrt{c_2}} = \frac{16321(15)}{\sqrt{100}} = 24481.5,\\
N_1\sigma_1\sqrt{c_1} &= 21123(20)\sqrt{400} = 8449200,\\
N_2\sigma_2\sqrt{c_2} &= 16321(15)\sqrt{100} = 2448150\\
\frac{N_1\sigma_1}{\sqrt{c_1}} + \frac{N_2\sigma_2}{\sqrt{c_2}} &= 21123 + 24481.5 = 45604.5,\\
N_1\sigma_1\sqrt{c_1} + N_2\sigma_2\sqrt{c_2} &= 8449200 + 2448150 = 10897350,\\
n &= (20000 - 4000)\left(\frac{45604.5}{10897350}\right) = 66.96 \approx 66,\\
n_1 = 66\left(\frac{21123}{45604.5}\right) &= 30.56 \approx 31, \quad n_2 = 66\left(\frac{24481.5}{45604.5}\right) = 35.43 \approx 35.
\end{aligned}$$

If the response cost per stratum is constant ($c_1 = c_2 = 100$):

$$\begin{aligned}
N_1\sigma_1 &= 21123(20) = 422460, \quad N_2\sigma_2 = 16321(15) = 244815,\\
N_1\sigma_1 + N_2\sigma_2 &= 422460 + 244815 = 667275,\\
n &= \frac{20000 - 4000}{100} = 160,\\
n_1 = 160\left(\frac{422460}{667275}\right) &= 101.3 \approx 101, \quad n_2 = 160\left(\frac{244815}{667275}\right) = 58.7 \approx 59.
\end{aligned}$$

If we also assume that the variances are equal in the 2 strata, the sample size remains $n = 160$, but the proportional allocation yields

$$n_1 = 160\left(\frac{21123}{37444}\right) = 90.25 \approx 90 \quad \text{and} \quad n_2 = 160\left(\frac{16321}{37444}\right) = 69.74 \approx 70. \quad \blacksquare$$

11.4.3 Comparison Between SRS and STS

Let $\mathcal{U} = \{u_1, \ldots, u_N\}$ have mean μ and variance σ^2.

Using a SRS of size n, we can construct the estimator $\overline{y}_{\text{SRS}}$, with sampling variance

$$\text{V}(\overline{y}_{\text{SRS}}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right).$$

We have studied the properties of such estimators in section 11.3.

If $\mathcal{U}$ can be split into M strata

$$\mathcal{U}_1 = \{u_{1,1}, \ldots, u_{1,N_1}\}, \ \cdots\ , \mathcal{U}_M = \{u_{M,1}, \ldots, u_{M,N_M}\},$$

with mean and variance μ_i and σ_i^2, respectively, for $1 \leq i \leq M$.

Using a STS of size $n = (n_1, \ldots, n_M)$, we can construct the estimator $\overline{y}_{\text{STS}}$, with sampling variance

$$\text{V}(\overline{y}_{\text{STS}}) = \frac{1}{N^2}\sum_{i=1}^{M} N_i^2 \cdot \frac{\sigma_i^2}{n_i}\left(\frac{N_i - n_i}{N_i - 1}\right).$$

Both samples have the same size; is there any way to determine which of the two approaches is preferable **before** computing the confidence intervals? In general, the sample design for which the **sampling variance** of the corresponding estimator is **smallest** is preferred.[38]

If $N \gg n$ and $N_i \gg n_i$ for all $1 \leq i \leq M$, then $N - n \approx N - 1$ and $N_i - n_1 \approx N_i - 1$ for all $1 \leq i \leq M$. Consequently,

$$\mathrm{V}(\overline{y}_{\mathrm{SRS}}) \approx \frac{\sigma^2}{n} = \frac{1}{nN}\sum_{i=1}^{M}\sum_{j=1}^{N_i}(u_{i,j}-\mu)^2 \quad \text{and} \quad \mathrm{V}(\overline{y}_{\mathrm{STS}}) \approx \frac{1}{N^2}\sum_{i=1}^{M} N_i^2 \cdot \frac{\sigma_i^2}{n_i}.$$

In a proportional allocation scenario, $n_i = n \cdot \frac{N_i}{N}$ for all $1 \leq i \leq M$, from which we see that

$$\mathrm{V}(\overline{y}_{\mathrm{STS}})_{\mathrm{Prop}} \approx \frac{1}{N^2}\sum_{i=1}^{M} N_i^2 \cdot \frac{\sigma_i^2 \cdot N}{nN_i} = \frac{1}{nN}\sum_{i=1}^{M} N_i\sigma_i^2.$$

In a Neyman allocation scenario, $n_i = n \cdot \frac{N_i\sigma_i}{N_1\sigma_1 + \cdots + N_M\sigma_M}$ for all $1 \leq i \leq M$, from which we see that

$$\mathrm{V}(\overline{y}_{\mathrm{STS}})_{\mathrm{Neyman}} \approx \frac{1}{N^2}\sum_{i=1}^{M} N_i^2 \cdot \frac{\sigma_i^2\left(\sum_{k=1}^{M} N_k\sigma_k\right)}{nN_i\sigma_i} = \frac{1}{nN^2}\left(\sum_{i=1}^{M} N_i\sigma_i\right)^2.$$

But

$$\begin{aligned}
\mathrm{V}(\overline{y}_{\mathrm{SRS}}) &\approx \frac{1}{nN}\sum_{i=1}^{M}\sum_{j=1}^{N_i}(u_{i,j}-\mu)^2 = \frac{1}{nN}\sum_{i=1}^{M}\sum_{j=1}^{N_i}(u_{i,j}-\mu_i+\mu_i-\mu)^2 \\
&= \frac{1}{nN}\sum_{i=1}^{M}\sum_{j=1}^{N_i}\left\{(u_{i,j}-\mu_i)^2 + 2(u_{i,j}-\mu_i)(\mu_i-\mu) + (\mu_i-\mu)^2\right\} \\
&= \frac{1}{nN}\left\{\sum_{i=1}^{M}\underbrace{\sum_{j=1}^{N_i}(u_{i,j}-\mu_i)^2}_{N_i\sigma_i^2} + 2\sum_{i=1}^{M}(\mu_i-\mu)\underbrace{\sum_{j=1}^{N_i}(u_{i,j}-\mu_i)}_{N_i\mu_i - N_i\mu_i = 0} + \sum_{i=1}^{M}(\mu_i-\mu)^2\underbrace{\sum_{j=1}^{N_i}1}_{N_i}\right\} \\
&= \frac{1}{nN}\left\{\sum_{i=1}^{M}N_i\sigma_i^2 + \sum_{i=1}^{M}N_i(\mu_i-\mu)^2\right\} = \mathrm{V}(\overline{y}_{\mathrm{STS}})_{\mathrm{Prop}} + \frac{1}{nN}\sum_{i=1}^{M}N_i(\mu_i-\mu)^2.
\end{aligned}$$

As such,

$$\mathrm{V}(\overline{y}_{\mathrm{SRS}}) \gg \mathrm{V}(\overline{y}_{\mathrm{STS}})_{\mathrm{Prop}}, \quad \text{whenever } \frac{1}{nN}\sum_{i=1}^{M}N_i(\mu_i-\mu)^2 \gg 0;$$

a STS under proportional allocation is substantially preferable to a SRS when **the variance of the stratum means is high**.

Similarly, set

$$\overline{\sigma} = \frac{1}{N}\sum_{i=1}^{M}N_i\sigma_i = \sqrt{n\mathrm{V}(\overline{y}_{\mathrm{STS}})_{\mathrm{Neyman}}}.$$

[38] This corresponds to a tighter (smaller) C.I.

As such,

$$\begin{aligned}
\mathrm{V}(\overline{y}_{\mathrm{STS}})_{\mathrm{Prop}} - \mathrm{V}(\overline{y}_{\mathrm{STS}})_{\mathrm{Neyman}} &= \frac{1}{nN}\sum_{i=1}^{M} N_i\sigma_i^2 - \frac{\overline{\sigma}^2}{n} \\
&= \frac{1}{nN}\left\{\sum_{i=1}^{M} N_i\sigma_i^2 - N\overline{\sigma}^2\right\} \\
&= \frac{1}{nN}\sum_{i=1}^{M} N_i(\sigma_i^2 - 2\sigma_i\overline{\sigma} + \overline{\sigma}^2) \\
&= \frac{1}{nN}\sum_{i=1}^{M} N_i(\sigma_i - \overline{\sigma})^2 \geq 0;
\end{aligned}$$

a STS under Neyman allocation is substantially preferable to a STS under proportional allocation **when the variance of the stratum standard deviations is high**.

Combining these, we can conclude that a STS under Neyman allocation is substantially preferable to a SRS when **stratum means and standard deviations vary greatly across strata**.

Since in practice there are other considerations at play (sampling cost, etc.), one may still decide in favor of a SRS or a STS under proportional allocation, especially if the difference in the corresponding variances is (relatively) small.

11.5 Ratio, Regression, Difference

In what follows we present ways to obtain estimates of the mean, the total, or of a proportion with the help of **auxiliary information**. So far, we have only discussed **univariate** SRS and STS estimators. Can we use more than one response per unit to obtain better approximations?

In the 2011 Gapminder dataset, there are $N = 168$ countries in 2011 for which the **life expectancy** Y and the (logarithm of the) **gross domestic product per capita** X are available. Suppose it is known that $\mathrm{E}[X] = \mu_X = 7.84$. If we draw a sample $\{(x_1, y_1), \ldots, (x_{10}, y_{10})\} \subseteq \mathcal{U}$ for which the mean of y_i/x_i is 8.67, can we expect that $\mu_Y \approx 8.67\mu_X = 68.00$?[39]

11.5.1 Ratio Estimation

Let $\mathcal{U} = \{(X_1, Y_1), \ldots, (X_N, Y_N)\}$ be a finite population of size N for which each unit u_j has 2 observed values: X_j and Y_j. The **ratio of the means** R is the ratio of the means (or totals):

$$R = \frac{\sum_{j=1}^{N} Y_j}{\sum_{j=1}^{N} X_j} = \frac{\mu_Y}{\mu_X} = \frac{\tau_Y}{\tau_X}, \quad \text{as long as } \mu_X, \tau_X \neq 0.$$

We are interested in such quotients when we try to determine the average wage Y as a function of years of schooling X in Canada, for example.

[39] See Figure 11.6.

Ratio Estimator

Let $\mathcal{Y} = \{(x_{i_1}, y_{i_1}), \ldots, (x_{i_n}, y_{i_n})\} \subseteq \mathcal{U}$ a **bivariate simple random sample** of size n. We often simplify the notation by writing

$$\mathcal{Y} = \{(x_1, y_1), \ldots, (x_n, y_n)\}.$$

The **sample ratio of means** r is an estimator of R:

$$r = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} = \frac{\overline{y}}{\overline{x}} = \frac{\hat{\tau}_Y}{\hat{\tau}_X}, \quad \text{as long as } \overline{x}, \hat{\tau}_X \neq 0.$$

Warning: this is a biased estimator!

Example Consider a finite bivariate population with $N = 4$ units:

$$u_1 = (1, 2), \quad u_2 = (1, 0), \quad u_3 = (2, 1), \quad u_4 = (4, 5).$$

The population ratio of means R is simply

$$R = \frac{2 + 0 + 1 + 5}{1 + 1 + 2 + 4} = \frac{8}{8} = 1.$$

Suppose that we want to provide an estimate of R by drawing a SRS of size $n = 3$ from $\mathcal{U}$. There are $\binom{4}{3} = 4$ such samples.

Sample	y **Values**	$\overline{y}$	x **Values**	$\overline{x}$	r	$P(r)$
u_1, u_2, u_3	$2, 0, 1$	1	$1, 1, 2$	$4/3$	$3/4$	$1/4$
u_1, u_2, u_4	$2, 0, 5$	$7/3$	$1, 1, 4$	2	$7/6$	$1/4$
u_1, u_3, u_4	$2, 1, 5$	$8/3$	$1, 2, 4$	$7/3$	$8/7$	$1/4$
u_2, u_3, u_4	$0, 1, 5$	2	$1, 2, 3$	2	1	$1/4$

We can compute the expectation of the estimator r directly:

$$\mathrm{E}[r] = \sum_r rP(r) = \frac{1}{4}(3/4 + 7/6 + 8/7 + 1) = \frac{341}{336} \approx 1.014881 \neq R = 1. \quad \blacksquare$$

What is the **sampling bias** of r as an estimator of R, then?

$$\mathrm{E}[r - R] = \mathrm{E}\left[\frac{\overline{y}}{\overline{x}} - R\right] = \mathrm{E}\left[\frac{1}{\overline{x}}(\overline{y} - R\overline{x})\right] = ??$$

Ratio Estimator Bias

In this last expression for the sampling bias, only $\overline{x}$ and $\overline{y}$ change when the sample changes: R remains constant. But there is no simple expression allowing us to compute exactly the **expectation of a quotient of random variables**; we must use approximations.

Let $f : [a, b] \longrightarrow \mathbb{R}$ be C^2 over $[a, b]$ (i.e., f, f', f'' are all continuous over $[a, b]$). According to Taylor's theorem, for all $c \in (a, b)$, there exists a ξ between c and z such that

$$f(z) = f(c) + f'(c)(z - c) + \frac{f''(\xi)}{2}(z - c)^2.$$

Since f'' is continuous over $[a, b]$, f'' is bounded on $[a, b]$: $\exists M > 0$ such that $|f''(z)| \leq M$ for all $z \in [a, b]$.

Thus, if z is **sufficiently close** to c,

$$|f(c) + f'(c)(z - c)| \gg \frac{M}{2}(z - c)^2 \geq \left|\frac{f''(\xi)}{2}(z - c)^2\right|,$$

from which we conclude that

$$f(z) \approx f(c) + f'(c)(z - c);$$

this is the linear approximation of f at $z = c$. If $f(z) = \frac{1}{z}$, we know that $f'(z) = -\frac{1}{z^2}$. Set $z = \overline{x}$ and $c = \mu_X$.

Since f is C^2 over any interval $[a, b]$ with $a > 0$, if $\overline{x}$ is sufficiently close to μ_X, then the liner approximation becomes

$$\frac{1}{\overline{x}} \approx \frac{1}{\mu_X} - \frac{1}{\mu_X^2}(\overline{x} - \mu_X)$$

(the constant approximation would be $\frac{1}{\overline{x}} \approx \frac{1}{\mu_X}$).

But $\text{E}(\overline{x}) = \mu_X$, $\text{E}(\overline{y}) = \mu_Y$ (SRS), and $\mu_Y = R\mu_X$, so that

$$\begin{aligned}
\text{E}[r - R] &= \text{E}\Big[\frac{\overline{y} - R\overline{x}}{\overline{x}}\Big] \approx \text{E}\Big[\Big(\frac{1}{\mu_X} - \frac{1}{\mu_X^2}(\overline{x} - \mu_X)\Big)(\overline{y} - R\overline{x})\Big] \\
&= \text{E}\Big[\frac{1}{\mu_X}(\overline{y} - R\overline{x})\Big] - \text{E}\Big[\frac{1}{\mu_X^2}(\overline{x} - \mu_X)(\overline{y} - R\overline{x})\Big] \\
&= \frac{1}{\mu_X}\Big(\text{E}(\overline{y}) - R \cdot \text{E}(\overline{x})\Big) - \frac{1}{\mu_X^2}\Big(\text{E}\big[\overline{xy} - \mu_X\overline{y} - R\overline{x}^2 - R\mu_X\overline{x}\big]\Big) \\
&= \frac{1}{\mu_X}\underbrace{\Big(\mu_Y - R\mu_X\Big)}_{=0} - \frac{1}{\mu_X^2}\Big(\text{E}(\overline{xy}) - \mu_X\text{E}(\overline{y}) - R\big(\text{E}(\overline{x}^2) - \mu_X\text{E}(\overline{x})\big)\Big) \\
&= -\frac{1}{\mu_X^2}\Big(\text{E}(\overline{xy}) - \mu_X\mu_Y - R\big(\text{E}(\overline{x}^2) - \mu_X^2\big)\Big)
\end{aligned}$$

We further simplify the sampling bias $\text{E}[r - R]$ with the help of $\text{E}(\overline{xy}) = \mu_X\mu_Y + \text{Cov}(\overline{x}, \overline{y})$, and $\text{E}(\overline{x}^2) = \mu_X^2 + \text{V}(\overline{x})$. Thus,

$$\text{E}[r - R] \approx -\frac{1}{\mu_X^2}\big[\text{Cov}(\overline{x}, \overline{y}) - R \cdot \text{V}(\overline{x})\big].$$

In an SRS of size n, drawn from a finite population withsize N and variance $sigma^2$, we have already seen that

$$V(\overline{x}) = \frac{\sigma_X^2}{n}\Big(\frac{N - n}{N - 1}\Big) \quad \text{and} \quad V(\overline{y}) = \frac{\sigma_Y^2}{n}\Big(\frac{N - n}{N - 1}\Big).$$

Consider the random variable $Z = X + Y$. The SRS estimator of

$$\mu_Z = \mu_X + \mu_Y$$

is

$$\overline{z} = \overline{x} + \overline{y};$$

its **sampling variance** is

$$\begin{aligned} \mathrm{V}(\overline{z}) &= \frac{\sigma_Z^2}{n}\left(\frac{N-n}{N-1}\right), \quad \text{where} \\ \sigma_Z^2 = \frac{1}{N}\sum_{j=1}^N (z_j - \mu_Z)^2 &= \frac{1}{N}\sum_{j=1}^N \left\{(x_j + y_j) - (\mu_X + \mu_Y)\right\}^2 \\ &= \frac{1}{N}\sum_{j=1}^N (x_j - \mu_X)^2 + \frac{2}{N}\sum_{j=1}^N (x_j - \mu_X)(y_j - \mu_Y) + \frac{1}{N}\sum_{j=1}^N (y_j - \mu_Y)^2 \\ &= \sigma_X^2 + 2\sigma_{XY} + \sigma_Y^2 = \sigma_X^2 + 2\rho\sigma_X\sigma_Y + \sigma_Y^2, \end{aligned}$$

where $\rho = \frac{\mathrm{Cov}(X,Y)}{\sigma_X\sigma_Y}$ is the **Pearson correlation coefficient between X and Y**.

On the one hand,

$$\mathrm{V}(\overline{z}) = \frac{\sigma_X^2 + 2\rho\sigma_X\sigma_Y + \sigma_Y^2}{n}\left(\frac{N-n}{N-1}\right);$$

on the other,

$$\begin{aligned} \mathrm{V}(\overline{z}) &= \mathrm{V}(\overline{x} + \overline{y}) = \mathrm{V}(\overline{x}) + 2\mathrm{Cov}(\overline{x}, \overline{y}) + \mathrm{V}(\overline{y}) \\ &= \frac{\sigma_X^2}{n}\left(\frac{N-n}{N-1}\right) + 2\mathrm{Cov}(\overline{x}, \overline{y}) + \frac{\sigma_Y^2}{n}\left(\frac{N-n}{N-1}\right); \end{aligned}$$

we can thus conclude that

$$\mathrm{Cov}(\overline{x}, \overline{y}) = \frac{\rho\sigma_X\sigma_Y}{n}\left(\frac{N-n}{N-1}\right).$$

Consequently,

$$\begin{aligned} \mathrm{E}[r - R] &\approx -\frac{1}{\mu_X^2}\left[\mathrm{Cov}(\overline{x}, \overline{y}) - R \cdot \mathrm{V}(\overline{x})\right] \\ &= -\frac{1}{\mu_X^2}\left[\frac{\rho\sigma_X\sigma_Y}{n}\left(\frac{N-n}{N-1}\right) - R\frac{\sigma_X^2}{n}\left(\frac{N-n}{N-1}\right)\right] \\ &= \frac{1}{\mu_X^2} \cdot \frac{R\sigma_X^2 - \rho\sigma_X\sigma_Y}{n}\left(\frac{N-n}{N-1}\right) \end{aligned}$$

But the **systematic error** is not the only way to qualify the magnitude of the error made when using r to estimate R: the **mean square error** (MSE) of r is

$$\mathrm{MSE}(r) = \mathrm{E}\left((r - R)^2\right) = \mathrm{V}(r) + \left(\mathrm{E}(r) - R\right)^2.$$

Ratio Estimator Variability

We can obtain an approximation of $\mathrm{V}(r)$ using the constant Taylor approximation (of order 0):

$$\frac{1}{\overline{\overline{x}}} \approx \frac{1}{\mu_X}.$$

Thus,

$$\mathrm{V}(r) = \mathrm{V}(r - R) = \mathrm{V}\Big[\frac{\overline{y}}{\overline{\overline{x}}} - R\Big] = \mathrm{V}\Big[\frac{\overline{y} - R\overline{x}}{\overline{x}}\Big] \approx \mathrm{V}\Big[\frac{\overline{y} - R\overline{x}}{\mu_X}\Big].$$

Consider the random variable $W = Y - RX$. Since $\mu_Y = R\mu_X$,

$$\mu_W = \mu_Y - R\mu_X = 0.$$

The SRS sample mean of W in $\mathcal{Y}$ is thus

$$\overline{w} = \overline{y} - R\overline{x} \implies \mathrm{V}(r) \approx \mathrm{V}\Big[\frac{\overline{w}}{\mu_X}\Big] = \frac{1}{\mu_X^2}\mathrm{V}(\overline{w}) = \frac{1}{\mu_X^2} \cdot \frac{\sigma_W^2}{n}\Big(\frac{N-n}{N-1}\Big),$$

where

$$\sigma_W^2 = \frac{1}{N}\sum_{j=1}^{N}(W_j - \mu_W)^2 = \frac{1}{N}\sum_{j=1}^{N}W_j^2 = \frac{1}{N}\sum_{j=1}^{N}(Y_j - RX_j)^2.$$

We thus have

$$\mathrm{V}(r) \approx \frac{1}{\mu_X^2} \cdot \frac{1}{n} \cdot \frac{1}{N}\sum_{j=1}^{N}(Y_j - RX_j)^2\Big(\frac{N-n}{N-1}\Big).$$

The ratio between the systematic error $\mathrm{E}[r - R]$ and the standard deviation of r is then

$$\frac{\mathrm{E}[r-R]}{\mathrm{SD}(r)} \approx \frac{1}{\sqrt{n}} \cdot \frac{R\sigma_X^2 - \rho\sigma_X\sigma_Y}{\sigma_W}\sqrt{\frac{N-1}{N-n}};$$

when $n, N \longrightarrow \infty$ (while $N \gg n$), we must have

$$\frac{\mathrm{E}[r-R]}{\mathrm{SD}(r)} \longrightarrow 0.$$

In other words, although it is impossible to get rid of the bias, the estimation error

$$\mathrm{MSE}(r) = \mathrm{V}(r) + (\mathrm{E}(r) - R)^2$$

is dominated by the variance $\mathrm{V}(r)$ if the sample size n is **sufficiently large**.

Example The list of countries for which both life expectancy and (logarithm of) gross domestic product per capita are available in 2011 contains $N = 168$ observations.

```
gapminder.RLD <- gapminder |>  filter(year==2011) |>
    select(life_expectancy,gdp,population)

# we keep only the observations that have both
gapminder.RLD <-  gapminder.RLD[complete.cases(gapminder.RLD),]
gapminder.RLD <- gapminder.RLD |> mutate(lgdppc=log(gdp/population))
(N=nrow(gapminder.RLD))
```

```
[1] 168
```

We draw $m = 500$ SRS samples of $n = 20$, and we compute the estimator r of the ratio R for each of these samples.

```
set.seed(12) # replicability
n=20
m=500

quotients <- c()
for(k in 1:m){
    samp <- gapminder.RLD[sample(1:N,n, replace=FALSE),c("life_expectancy","lgdppc")]
    quotients[k] <- mean(samp$life_expectancy/samp$lgdppc)
  }
```

The average of the 500 estimators is shown below:

```
quotients <- data.frame(quotients)
mean(quotients$quotients)
```

```
[1] 9.238648
```

We already know that $\mu_X = 7.84$. It would be reasonable to expect that $\mu_Y \approx \overline{r}\mu_X$:

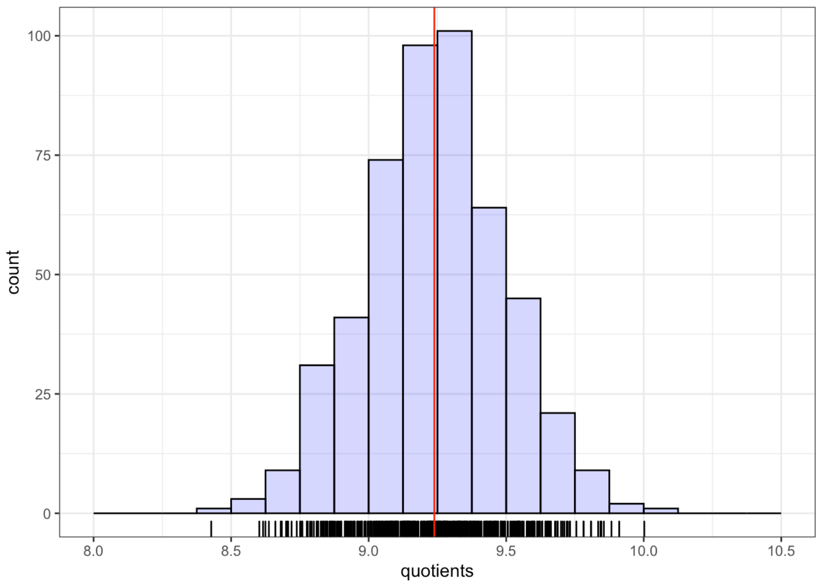
```
mean(gapminder.RLD$lgdppc)*mean(quotients$quotients)
```

```
[1] 72.45559
```

Is this a better approximation than the one we obtained at the beginning of the section: $\mu_Y \approx 68.00$? This question cannot be answered without knowing the **distribution of the estimator** r.[40]

```
ggplot(quotients, aes(quotients)) +
    geom_rug(aes(quotients)) +
    geom_histogram(breaks=seq(8, 10.5, by = .125),
                   col="black", fill="blue", alpha=.2) +
    geom_vline(xintercept=mean(quotients$quotients),
               color="red")
```

[40] Keep in mind that it is indeed a random variable since its value depends on the sample $\mathcal{Y}$ selected.



```
summary(quotients)
```

```
Min.     1st Qu.  Median   Mean     3rd Qu.  Max.
8.428    9.073    9.246    9.239    9.401    10.002
```

What do you think?

Ratio Estimator Confidence Intervals

We can show that the estimator r follows **approximately** a normal distribution $\mathcal{N}(\mathrm{E}(r), \mathrm{V}(r))$, from which we conclude that the **bound on the error of estimation is**

$$B_R \approx \hat{B}_R = 2\sqrt{\hat{\mathrm{V}}(r)} \approx 2\sqrt{\frac{1}{\mu_X^2} \cdot \frac{s_W^2}{n}\left(1 - \frac{n}{N}\right)} \approx 2\sqrt{\frac{1}{\overline{x}^2} \cdot \frac{s_W^2}{n}\left(1 - \frac{n}{N}\right)},$$

where

$$s_W^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - rx_i)^2.$$

Thus

$$\text{C.I.}(R; 0.95): \quad r \pm \hat{B}_R$$

is an **approximate 95% C.I. for** R.

Write $\rho = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}$. We notice that

$$\begin{aligned}
\sigma_W^2 &= \frac{1}{N}\sum_{j=1}^{N} W_j^2 = \frac{1}{N}\sum_{j=1}^{N}(Y_j - RX_j)^2 = \frac{1}{N}\sum_{j=1}^{N}(Y_j - \mu_Y + \mu_Y - RX_j) \\
&= \frac{1}{N}\sum_{j=1}^{N}(Y_j - \mu_Y + R\mu_X - RX_j)^2 = \frac{1}{N}\sum_{j=1}^{N}[(Y_j - \mu_Y) - R(X_j - \mu_X)]^2 \\
&= \frac{1}{N}\sum_{j=1}^{N}(Y_j - \mu_Y)^2 - 2R\frac{1}{N}\sum_{j=1}^{N}(X_j - \mu_X)(Y_j - \mu_Y) + R^2\frac{1}{N}\sum_{j=1}^{N}(X_j - \mu_X)^2 \\
&= \sigma_Y^2 - 2R\text{Cov}(X,Y) + R^2\sigma_X^2 = \sigma_Y^2 - 2R\rho\sigma_X\sigma_Y + R^2\sigma_X^2,
\end{aligned}$$

By analogy, we then have $s_W^2 = s_Y^2 - 2r\hat{\rho}s_X s_Y + r^2 s_X^2$, where

$$\begin{aligned}
s_X^2 &= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right), \quad s_Y^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - n\overline{y}^2\right), \\
s_{XY} &= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\overline{xy}\right), \quad \text{and} \quad \hat{\rho} = \frac{s_{XY}}{s_X s_Y}.
\end{aligned}$$

In practice, we can also use the following formula:

$$s_W^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - 2r\sum_{i=1}^{n} x_i y_i + r^2\sum_{i=1}^{n} x_i^2\right).$$

Example Consider a SRS $\mathcal{Y} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ of size $n = 132$, drawn from a population of size $N = 37,444$. Find a 95% C.I. for R if

$$\begin{aligned}
&\sum_{i=1}^{n} x_i = 9464.6, \quad \sum_{i=1}^{n} y_i = 14691.6, \\
&\sum_{i=1}^{n} x_i^2 = 686773.2, \quad \sum_{i=1}^{n} x_i y_i = 1062186, \quad \sum_{i=1}^{n} y_i^2 = 1670194.
\end{aligned}$$

With this sample, we have $r = \frac{14691.6}{9464.6} \approx 1.55$, so that

$$\begin{aligned}
s_W^2 &= \frac{1670194 - 2(1.55)(1062186) + (1.55)^2(686773.2)}{132 - 1} \approx 209.2, \quad \text{and} \\
\hat{\text{V}}(r) &\approx \frac{132^2}{9464.6^2}\frac{209.2}{132}\left(1 - \frac{132}{37444}\right) = 0.0003 \implies \text{C.I.}(R; 0.95) \approx 1.552 \pm 0.035.
\end{aligned}$$

Example Find a 95% C.I. for the ratio of life expectancy by the logarithm of the GDO per capita in 2011 with the help of a SRS of size $n = 20$.

The true ratio is:

```
(R = mean(gapminder.RLD$life_expectancy)/mean(gapminder.RLD$lgdppc))
```

```
[1] 9.046742
```

We draw a sample of size $n = 20$, and we calculate the intermediate sums:

```
N=nrow(gapminder.RLD); n=20
set.seed(123456) # replicability
index = sample(1:N,n, replace=FALSE)
samp = gapminder.RLD[index,c("life_expectancy","lgdppc")]

(sum.xi = sum(samp$lgdppc))
(sum.yi = sum(samp$life_expectancy))
(sum.xi.2 = sum(samp$lgdppc^2))
(sum.yi.2 = sum(samp$life_expectancy^2))
(sum.xiyi = sum(samp$lgdppc*samp$life_expectancy))
```

```
[1] 167.2794
[1] 1450.82
[1] 1430.912
[1] 106117.4
[1] 12245.93
```

Finally, we compute the estimator r and its variance, as well as the desired confidence interval.

```
r = sum.yi/sum.xi
s2.W = 1/(n-1)*(sum.yi.2-2*r*sum.xiyi+r^2*sum.xi.2)
V = n^2/sum.xi^2*(1/n)*s2.W*(1-n/N)
B = 2*sqrt(V)
c(r-B,r+B)
```

```
[1] 8.252515 9.093552
```

We would expect the quotient R to be in the interval $(8.25, 9.09)$ with 95% probability;[41] since $R = 9.046742$, it is indeed the case.[42]

Estimation of the Mean and the Total Using the Ratio Estimator

In practice, we often know τ_X and/or μ_X. It is possible to use the relation

$$\mu_Y = R\mu_X, \quad \text{where } R = \frac{\mu_Y}{\mu_X}$$

in order to approximate μ_Y (if μ_X is unknown, one uses $\mu_X \approx \overline{x}$).

Since $r = \overline{y}/\overline{x}$, the **ratio-based estimator for** $\hat{\mu}_{Y;R}$ is simply:

$$\hat{\mu}_{Y;R} = r \cdot \mu_X.$$

But we have already observed that r is a biased estimator of R, so we expect $\hat{\mu}_{Y;R}$ to be a biased estimator of μ_Y, with a normal distribution: $\hat{\mu}_{Y;R} \sim_{\text{approx}} \mathcal{N}\left(\mathrm{E}(\hat{\mu}_{Y;R}), \mathrm{V}(\hat{\mu}_{Y;R})\right)$.

It is easy to show

$$\mathrm{E}[\hat{\mu}_{Y;R} - \mu_Y] = \mu_X \mathrm{E}[r - R] \approx \frac{1}{\mu_X} \cdot \frac{R\sigma_X^2 - \rho\sigma_X\sigma_Y}{n}\left(\frac{N-n}{N-1}\right)$$

$$\mathrm{V}(\hat{\mu}_{Y;R}) = \mathrm{V}(r \cdot \mu_X) = \mu_X^2 \mathrm{V}(r) \approx \frac{\sigma_W^2}{n}\left(\frac{N-n}{N-1}\right).$$

[41] According to the frequentist interpretation of confidence intervals.

[42] As we have noticed several times, the confidence interval can of course change depending on which sample is drawn from the population.

The bound of error on the estimation of $\mu_{Y;R}$ is thus

$$B_{\mu_{Y;R}} \approx \hat{B}_{\mu_{Y;R}} = 2\sqrt{\mathrm{V}(\hat{\mu}_{Y;R})} \approx 2\sqrt{\frac{s_W^2}{n}\left(1-\frac{n}{N}\right)},\ s_W^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - rx_i)^2,$$

from which we see that $\text{C.I.}_R(\mu_Y;0.95) \equiv \hat{\mu}_{Y;R} \pm \hat{B}_{\mu_{Y;R}}$ is an **approximate 95% C.I. for** μ_Y.

It is also possible to use the relationship

$$\tau_Y = R\tau_X, \quad \text{where} \quad R = \frac{\mu_Y}{\mu_X} = \frac{\tau_Y}{\tau_X}$$

to approximate τ_Y (if τ_X is unknown, we use $\tau_X \approx N\overline{x}$).

Since $r = \overline{y}/\overline{x}$, the **ratio-based estimator for** $\hat{\tau}_{Y;R}$ is simply:

$$\hat{\tau}_{Y;R} = r \cdot \tau_X.$$

But we have already observed that r is a biased estimator of R, so we expect $\hat{\tau}_{Y;R}$ to be a biased estimator of τ_Y, which follows a normal distribution:

$$\hat{\tau}_{Y;R} \sim_{\text{approx}} \mathcal{N}\left(\mathrm{E}(\hat{\tau}_{Y;R}), \mathrm{V}(\hat{\tau}_{Y;R})\right).$$

It is easy to show

$$\mathrm{E}[\hat{\tau}_{Y;R} - \tau_Y] = \tau_X \mathrm{E}[r - R] \approx \frac{N}{\mu_X} \cdot \frac{R\sigma_X^2 - \rho\sigma_X\sigma_Y}{n}\left(\frac{N-n}{N-1}\right)$$
$$\mathrm{V}(\hat{\tau}_{Y;R}) = \mathrm{V}(r \cdot \tau_X) = \tau_X^2 \mathrm{V}(r) = N^2\mu_X^2\mathrm{V}(r) \approx N^2 \cdot \frac{\sigma_W^2}{n}\left(\frac{N-n}{N-1}\right).$$

The **bound of error on the estimation of** $\tau_{Y;R}$ is thus

$$B_{\tau_{Y;R}} \approx \hat{B}_{\tau_{Y;R}} = 2\sqrt{\mathrm{V}(\hat{\tau}_{Y;R})} \approx 2N\sqrt{\frac{s_W^2}{n}\left(1-\frac{n}{N}\right)},\ s_W^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - rx_i)^2,$$

from which we conclude that $\text{C.I.}_R(\tau_Y;0.95) \equiv \hat{\tau}_{Y;R} \pm \hat{B}_{\tau_{Y;R}}$ is an **approximate 95% C.I. for** τ_Y.

Example Consider a SRS $\mathcal{Y} = \{(x_1,y_1),\ldots,(x_n,y_n)\}$ o size $n = 132$, drawn from a population of size $N = 37,444$. Find a 95% C.I. for μ_Y using ratio-based estimation, given that

$$\sum_{i=1}^{n} x_i = 9464.6, \quad \sum_{i=1}^{n} y_i = 14691.6,$$
$$\sum_{i=1}^{n} x_i^2 = 686773.2, \quad \sum_{i=1}^{n} x_iy_i = 1062186, \quad \sum_{i=1}^{n} y_i^2 = 1670194.$$

With this sample, we have $r \approx 1.55$, $s_W^2 \approx 209.2$, $\hat{\mathrm{V}}(r) \approx 0.00031$, and $\text{C.I.}(R;0.95) \approx 1.552 \pm 0.035$. Moreover, $\overline{x} = 9464.6/132 = 71.70$. Thus

$$\text{C.I.}_R(\mu_Y;0.95) = \mu_X \cdot \text{C.I.}(R;0.95) \approx \overline{x} \cdot \text{C.I.}(R;0.95) \equiv 111.29 \pm 2.51. \quad \blacksquare$$

Example Find a 95% C.I. for the average life expectancy by country μ_Y, in 2011, using ratio estimation and the logarithm of the gross domestic product per capita in 2011 (X), with a SRS sample of size $n = 20$.

We use the same sample as in the preceding example on the topic. We have already obtained a confidence interval for the ratio:

$$\text{C.I.}(R; 0.95) = (8.25, 9.09).$$

The sample mean of X was $\overline{x} = \frac{167.2794}{20} = 8.364$. The 95% confidence interval for the average life expectancy using ratio estimation is thus

$$\text{C.I.}_R(\mu_Y; 0.95) = \mu_X \cdot \text{C.I.}(R; 0.95) \approx \overline{x} \cdot (8.25, 9.09) = (69.00, 76.03).$$

Recall that the true value is $\mu_Y = 70.95$.

Sample Size

Just as was the case with SRS and STS, we can determine the required sample size assuming that we have some information about the population distribution.

To give an estimate for R, use:

$$\begin{aligned} B_R \approx 2\sqrt{\frac{1}{\mu_X^2} \cdot \frac{\sigma_W^2}{n}\left(\frac{N-n}{N-1}\right)} &\Longleftrightarrow \underbrace{\frac{B_R^2 \mu_X^2}{4}}_{=D_R} = \frac{\sigma_W^2}{n}\left(\frac{N-n}{N-1}\right) \\ &\Longleftrightarrow \frac{(N-1)D_R}{\sigma_W^2} = \frac{N-n}{n} = \frac{N}{n} - 1 \\ &\Longleftrightarrow \frac{(N-1)D_R + \sigma_W^2}{\sigma_W^2} = \frac{N}{n} \\ &\Longleftrightarrow n_R = \frac{N\sigma_W^2}{(N-1)D_R + \sigma_W^2}. \end{aligned}$$

To give an estimate for μ_Y with ratio estimation, use:

$$B_{\mu_{Y;R}} \approx 2\sqrt{\frac{\sigma_W^2}{n}\left(\frac{N-n}{N-1}\right)} \Longleftrightarrow n_{\mu_Y} = \frac{N\sigma_W^2}{(N-1)D_{\mu_Y} + \sigma_W^2}, \quad D_{\mu_Y} = \frac{B_{\mu_{Y;R}}^2}{4};$$

for τ_Y, use:

$$B_{\tau_{Y;R}} \approx 2\sqrt{N^2 \cdot \frac{\sigma_W^2}{n}\left(\frac{N-n}{N-1}\right)} \Longleftrightarrow n_{\tau_Y} = \frac{N\sigma_W^2}{(N-1)D_{\tau_Y} + \sigma_W^2}, \quad D_{\tau_Y} = \frac{B_{\tau_{Y;R}}^2}{4N^2}.$$

Since we do not typically know σ_W^2, we often use a small preliminary sample and use the empirical variance s_W^2 as an estimator of σ_W^2.

Example Consider a SRS $\mathcal{Y} = \{(x_1, y_1), \dots, (x_n, y_n)\}$ of size n, drawn from a population of size $N = 37,444$. Assume that we have $\sigma_W^2 \approx 209.2$ and $\mu_X \approx 71.7$, perhaps from a previous study.

Determine the minimum sample size required to ensure that the bound on the error of estimation of the:

1. ratio R using r is at most 0.025;
2. mean μ_Y using $\hat{\mu}_{Y;R}$ is at most 5, and
3. total τ_Y using $\hat{\tau}_{Y;R}$ is at most 25.

We simply use the formulas.

1. since $D_R = \frac{B_R^2 \mu_X^2}{4} = \frac{0.025^2(71.7)^2}{4} \approx 0.8033$, we have
$$n_R = \frac{37444(209.2)}{(37444-1)(0.8033)+209.2} = 258.6453 \implies n_R \geq 259;$$
2. since $D_{\mu_Y} = \frac{B^2_{\mu_{Y;R}}}{4} = \frac{5^2}{4} \approx 6.25$, we have
$$n_{\mu_Y} = \frac{37444(209.2)}{(37444-1)(6.25)+209.2} = 33.443 \implies n_{\mu_Y} \geq 34;$$
3. since $D_{\tau_Y} = \frac{B^2_{\tau_{Y;R}}}{4N^2} = \frac{25^2}{4(37444)} \approx 0.001502243$, we have
$$n_{\tau_Y} = \frac{37444(209.2)}{(37444-1)(0.001502243)+209.2} = 29509.62 \implies n_{\tau_Y} \geq 29510.$$

In this last case, the desired bound $B_{\tau_{Y;R}}$ is probably too tight (the resulting sample size is way too large). ■

11.5.2 Regression Estimation

Ratio estimation is a special case of a more general method, **regression estimation**. In the `gapminder.csv` dataset for 2011, we recognize that there is a more or less linear relationship between the **life expectancy** Y and the **logarithm of the GDP per capita** X for $N = 168$ countries.

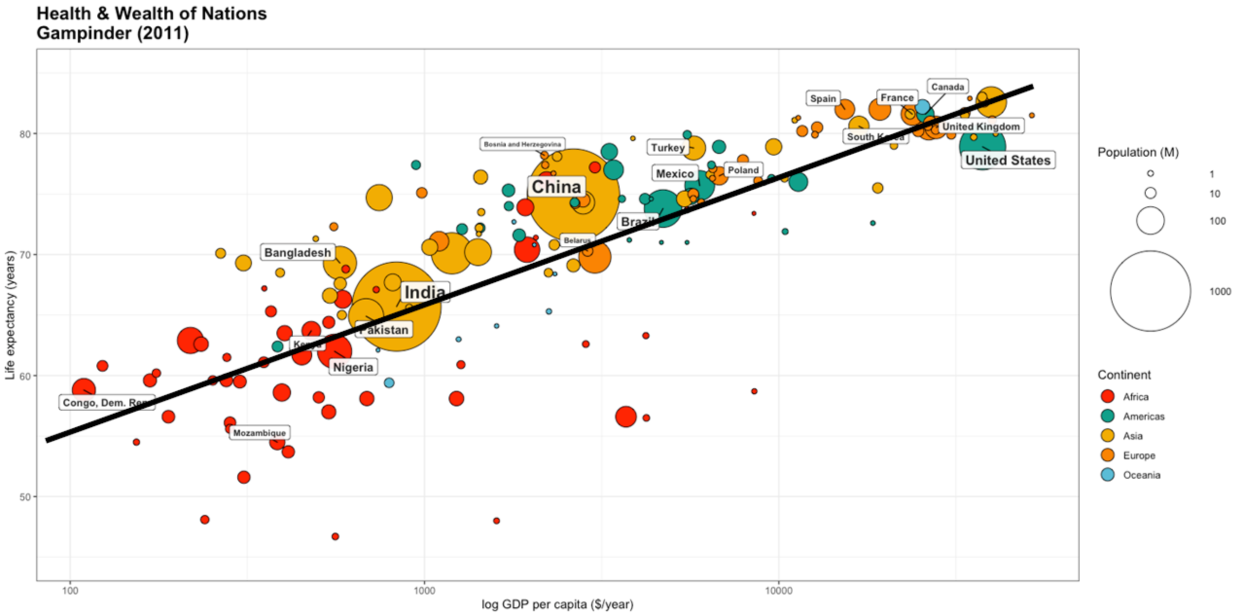


Figure 11.8: Health and wealth of nations for the 2011 Gapminder data, with superimposed line of best fit.

When we compute

$$r = \overline{y}/\overline{x}$$

using a SRS of size n, we are really assuming that the true relationship between Y and X takes the form $Y = RX \approx rX$, i.e., that it is a straight line of slope r **passing through the origin**. But this last condition does not seem to be met. What to do in this case?

Regression Estimator

As above, let $\mathcal{U}$ be a finite bivariate population of size N, and $\mathcal{Y} \subseteq \mathcal{U}$ be a finite bivariate random sample of size n. We assume that the relationship between Y and X takes the form

$$Y - \mu_Y = \beta(X - \mu_X).$$

If μ_X is known (as we had assumed was the case for ratio estimation), the **regression estimator** $\hat{\mu}_{Y;L}$ **of** μ_Y obtained with the SRS $\mathcal{Y}$ is

$$\hat{\mu}_{Y;L} = \overline{y} + \beta(\mu_X - \overline{x}).$$

For now, we treat β as an **unknown** constant (since μ_Y is also unknown). Since $\mathcal{Y}$ is drawn in a SRS context, $\text{E}(\overline{x}) = \mu_X$ and $\text{E}(\overline{y}) = \mu_Y$, so that

$$\text{E}(\hat{\mu}_{Y;L}) = \text{E}(\overline{y}) + \beta(\mu_X - \text{E}(\overline{x})) = \mu_Y + \beta(\mu_X - \mu_X) = \mu_Y.$$

Consider the random variable $W = Y + \beta(\mu_X - X)$. As β is constant, we have

$$\mu_W = \mu_Y + \beta(\mu_X - \mu_X) = \mu_Y.$$

The sample mean of W is thus

$$\overline{w} = \overline{y} + \beta(\mu_X - \overline{x}) = \hat{\mu}_{Y;L} \implies \text{V}(\hat{\mu}_{Y;L}) = \text{V}(\overline{w}) = \frac{\sigma^2_{W;L}}{n}\left(\frac{N-n}{N-1}\right).$$

But

$$\begin{aligned}\sigma^2_{W;L} &= \frac{1}{N}\sum_{j=1}^{N}(W_j - \mu_W)^2 = \frac{1}{N}\sum_{j=1}^{N}(Y_j + \beta(\mu_X - X_j) - \mu_Y)^2 \\ &= \frac{1}{N}\sum_{j=1}^{N}\left\{(Y_j - \mu_Y) - \beta(X_j - \mu_X)\right\}^2 = \sigma^2_Y - 2\beta\rho\sigma_X\sigma_Y + \beta^2\sigma^2_X,\end{aligned}$$

where $\rho = \frac{\text{Cov}(X,Y)}{\sigma_X\sigma_Y} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}$. Consequently,

$$\text{V}(\hat{\mu}_{Y;L}) = \frac{\sigma^2_Y - 2\beta\rho\sigma_X\sigma_Y + \beta^2\sigma^2_X}{n}\left(\frac{N-n}{N-1}\right).$$

In general, for a given systematic error (bias), preference is given to the estimator **with the lowest variance**. The value of β which minimizes $\text{V}(\hat{\mu}_{Y;L})$ would then satisfy

$$\frac{\partial \text{V}(\hat{\mu}_{Y;L})}{\partial\beta}(\beta^*) = \frac{1}{n}\left(\frac{N-n}{N-1}\right)(-2\rho\sigma_X\sigma_Y + 2\beta^*\sigma^2_X) = 0,$$

which is to say that

$$\beta^* = \rho\frac{\sigma_Y}{\sigma_X} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}\cdot\frac{\sigma_Y}{\sigma_X} = \frac{\sigma_{XY}}{\sigma_X^2},$$

from which we conclude that

$$\begin{aligned}
\mathrm{V}(\hat{\mu}_{Y;L}) &= \frac{\sigma_Y^2 - 2\beta^*\rho\sigma_X\sigma_Y + (\beta^*)^2\sigma_X^2}{n}\left(\frac{N-n}{N-1}\right)\\
&= \frac{\sigma_Y^2 - 2\rho\frac{\sigma_Y}{\sigma_X}\rho\sigma_X\sigma_Y + (\rho\frac{\sigma_Y}{\sigma_X})^2\sigma_X^2}{n}\left(\frac{N-n}{N-1}\right)\\
&= \frac{\sigma_Y^2 - 2\rho^2\sigma_Y^2 + \rho^2\sigma_Y^2}{n}\left(\frac{N-n}{N-1}\right)\\
&= \frac{\sigma_Y^2(1-\rho^2)}{n}\left(\frac{N-n}{n-1}\right).
\end{aligned}$$

Regression Estimator Bias

The task is to determine the coefficients α, β that "best describe" the linear relationship between X and Y,

$$y_i = \alpha + \beta x_i + \varepsilon_i, \quad i = 1, \dots, n,$$

where we assume that $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n) \sim_{\text{approx.}} \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_n)$.

There are several ways to interpret the phrase "best describe" – the **least squares estimators** $\hat{\alpha}$ and $\hat{\beta}$ are those that minimize the residual sum of squares

$$Q(\alpha,\beta) = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2.$$

We solve the system of equations

$$\frac{\partial Q}{\partial \alpha}(a,b) = \sum_{i=1}^n -2(y_i - a - bx_i) = 0, \quad \frac{\partial Q}{\partial \beta}(a,b) = \sum_{i=1}^n -2x_i(y_i - a - bx_i) = 0,$$

which yields

$$\hat{\alpha} = a = \overline{y} - b\overline{x} \quad \text{and} \quad \hat{\beta} = b = \frac{\displaystyle\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\displaystyle\sum_{i=1}^n (x_i - \overline{x})^2}.$$

In practice, it is this $b = \hat{\rho}\frac{s_Y}{s_X}$ that plays the role of the estimator β^*; note that it varies from one SRS to another. Since b is a random variable,[43] **we cannot conclude that** $\mathrm{E}(b\overline{x}) = \mathrm{E}(b)\mathrm{E}(\overline{x})$, so that

$$\mathrm{E}(\hat{\mu}_{Y;L}) = \mathrm{E}(\overline{y}) + \mu_X\mathrm{E}(b) - \mathrm{E}(b\overline{x}) \neq \mu_Y,$$

in general.

[43] in the sense that we obtain (potentially) a different slope with every SRS $\mathcal{Y}$.

However, if the sample size n is large, it is possible to show that

$$\mathrm{E}[\hat{\mu}_{Y;L} - \mu_Y]$$

is of order $\frac{1}{n}$ (as was the case for the systematic error in ratio estimation); $\hat{\mu}_{Y;L}$ is therefore a **biased estimator** of μ_Y.

Regression Estimator Variability

The sampling variance of $\hat{\mu}_{Y;L}$ is also of order $\frac{1}{n}$, and so the quotient of the bias $\mathrm{E}[\hat{\mu}_{Y;L} - \mu_Y]$ by the standard deviation of $\hat{\mu}_{Y;L}$ is of order $\frac{1}{\sqrt{n}}$.

Thus, when $n, N \longrightarrow \infty$ (assuming that $N \gg n$), we have

$$\frac{\mathrm{E}[\hat{\mu}_{Y;L} - \mu_Y]}{\mathrm{SD}(\hat{\mu}_{Y;l})} \longrightarrow 0.$$

Although it is impossible to get rid of the bias, the estimation error

$$\mathrm{MSE}(\hat{\mu}_{Y;L}) = \mathrm{V}(\hat{\mu}_{Y;L}) + (\mathrm{E}(\hat{\mu}_{Y;L}) - \mu_Y)^2$$

is dominated byt the variance $\mathrm{V}(\hat{\mu}_{Y;L})$ when n is **sufficiently large**.

Regression Estimator Confidence Intervals

The regression estimator $\hat{\mu}_{Y;L}$ follows **approximately** a normal distribution $\mathcal{N}\left(\mathrm{E}(\hat{\mu}_{Y;L}), \mathrm{V}(\hat{\mu}_{Y;L})\right)$, from which we conclude that the **bound on the error of estimation** is

$$B_L \approx \hat{B}_L = 2\sqrt{\hat{\mathrm{V}}(\hat{\mu}_{Y;L})} \approx 2\sqrt{\frac{s^2_{W;L}}{n}\left(1 - \frac{n}{N}\right)},$$

where $s^2_{W;L}$ is the **regression mean square error**,

$$s^2_{W;L} = \frac{n-1}{n-2}(s^2_Y - b^2 s^2_X) = \frac{n-1}{n-2} \cdot s^2_Y(1 - \hat{\rho}^2).$$

Consequently $\mathrm{C.I.}_L(\mu_Y; 0.95) : \hat{\mu}_{Y;L} \pm \hat{B}_L$ is an **approximate 95% C.I. for** μ_Y.[44]

Example Consider a SRS $\mathcal{Y} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ with $n = 132$, drawn from population of size $N = 37,444$. In a preceding study, we have shown that $\mu_X \approx 70.3$. Find a 95% C.I. for μ_Y using regression estimation if

$$\sum_{i=1}^n x_i = 9464.6, \quad \sum_{i=1}^n y_i = 14691.6,$$

$$\sum_{i=1}^n x_i^2 = 686773.2, \quad \sum_{i=1}^n x_i y_i = 1062186, \quad \sum_{i=1}^n y_i^2 = 1670194.$$

[44] We tackle τ_Y and p_Y in the usual manner.

We must evaluate $\overline{x}$, $\overline{y}$, s_X^2, s_{XY}, s_Y^2, and $\hat{\rho}$. But

$$\overline{x} = \frac{9464.6}{132} \approx 71.7, \quad \overline{y} = \frac{14691.6}{132} \approx 111.3,$$

$$s_X^2 = \frac{686773.2 - 132(71.7)^2}{132-1} \approx 62.2, \quad s_Y^2 = \frac{1670194 - 132(111.3)^2}{132-1} \approx 267.3$$

$$s_{XY} = \frac{1062186 - 132(71.7)(111.3)}{132-1} \approx 67.2, \quad \hat{\rho} = \frac{67.2}{\sqrt{(62.2)(267.3)}} \approx 0.521.$$

The estimator for the regression slope is therefore $b = \hat{\rho}\frac{s_Y}{s_X} = 1.08$. Moreover,

$$s_{W;L}^2 = \frac{131}{130} \cdot 267.3 \cdot (1 - 0.521^2) \approx 196.77.$$

Consequently,

$$\hat{\mu}_{Y;L} = 111.3 + 1.08(\underbrace{70.3}_{\mu_X} - 71.7) = 109.8, \quad \text{and}$$

$$\hat{B}_L \approx 2\sqrt{\frac{196.77}{132}\left(1 - \frac{132}{37444}\right)} = 2.43,$$

from which we conclude that

$$\text{C.I.}_L(\mu_Y; 0.95) \equiv 109.8 \pm 2.43.$$

Of course, if the linearity assumption is not valid, we should not expect the bound on the error of estimation using regression estimation to be substantially tighter than the one obtained in a SRS, say.

Example Find a 95% C.I. for the average life expectancy by country in 2011 using regression estimation against the logarithm of the GDP per capita, with $n = 20$, assuming that it is known that $\mu_X = 7.84$.

We draw a sample of size $n = 20$ and calculate the required quantities:

```
set.seed(123456) # replicability
N=nrow(gapminder.RLD); n=20
index = sample(1:N,n, replace=FALSE)
samp = gapminder.RLD[index,c("life_expectancy","lgdppc")]
mu.X = mean(gapminder.RLD$lgdppc)
```

The sample means are:

```
(y.bar = mean(samp$life_expectancy))
(x.bar = mean(samp$lgdppc))
```

```
[1] 72.541
[1] 8.363971
```

The intermediate sums and the correlation coefficient are:

```
sum.xi = sum(samp$lgdppc)
sum.yi = sum(samp$life_expectancy)
sum.xi.2 = sum(samp$lgdppc^2)
sum.yi.2 = sum(samp$life_expectancy^2)
sum.xiyi = sum(samp$lgdppc*samp$life_expectancy)

s2.X = (sum.xi.2-n*x.bar^2)/(n-1)
s2.Y = (sum.yi.2-n*y.bar^2)/(n-1)
s.XY = (sum.xiyi-n*x.bar*y.bar)/(n-1)

(rho = s.XY/sqrt(s2.X*s2.Y))
```

```
[1] 0.667983
```

Next, we evaluate the MSE:

```
(s2.W.L = (n-1)/(n-2)*s2.Y*(1-rho^2))
```

```
[1] 26.8736
```

The bound on the error of estimation is thus:

```
(B = 2*sqrt(s2.W.L/n*(1-n/N)))
```

```
[1] 2.175976
```

and the corresponding 95% C.I. for the mean life expectancy by country is:

```
(hat.mu.Y.L = y.bar + rho*sqrt(s2.Y/s2.X)*(mu.X-x.bar))
c(hat.mu.Y.L-B,hat.mu.Y.L+B)
```

```
[1] 70.71572
[1] 68.53974 72.89170
```

For comparison's sake, the true mean is $\mu_Y = 70.95$.

We can also compute the estimate and the confidence interval directly, with the base `lm()` function.

```
reg.lin = lm(life_expectancy~lgdppc, data=samp)
summary(reg.lin)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-16.2467   0.1592   1.6513   2.6614   5.8812

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)  43.2559     7.7768   5.562  2.8e-05 ***
lgdppc        3.5013     0.9194   3.808  0.00129 **
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 5.184 on 18 degrees of freedom
Multiple R-squared:  0.4462,    Adjusted R-squared:  0.4154
F-statistic:  14.5 on 1 and 18 DF,  p-value: 0.001287
```

The required quantities can be extracted as follows:

```
(b = as.numeric(reg.lin$coefficients[2]))
```

```
[1] 3.501336
```

```
(s2.W.L = summary(reg.lin)$sigma^2)
```

```
[1] 26.8736
```

Sample Size

If we seek an regression estimate of μ_Y, we use:

$$B_L \approx 2\sqrt{\frac{\sigma^2_{W;L}}{n}\left(\frac{N-n}{N-1}\right)} \iff \underbrace{\frac{B_L^2}{4}}_{=D_L} = \frac{\sigma^2_{W;L}}{n}\left(\frac{N-n}{N-1}\right) \iff$$

$$\frac{(N-1)D_L}{\sigma^2_{W;L}} = \frac{N-n}{n} = \frac{N}{n} - 1 \iff \frac{(N-1)D_L + \sigma^2_{W;L}}{\sigma^2_{W;L}} = \frac{N}{n},$$

$$\iff n_L = \frac{N\sigma^2_{W;L}}{(N-1)D_L + \sigma^2_{W;L}}.$$

For τ_Y, we use:

$$B_{\tau;L} \approx 2N\sqrt{\frac{\sigma^2_{W;L}}{n}\left(\frac{N-n}{N-1}\right)} \iff n_{\tau;L} = \frac{N\sigma^2_{W;L}}{(N-1)D_{\tau;L} + \sigma^2_{W;L}},$$

$$\text{where } D_{\tau;L} = \frac{B^2_{\tau;L}}{4N^2}.$$

Since we do not know usually know $\sigma^2_{W;L}$, we often draw a small preliminary sample on which we compute the sample $s^2_{W;L}$, which is used as an estimator of $\sigma^2_{W;L}$.[45]

Example Determine the sample size n required to estimate the average life expectancy μ_Y using regression estimation against the logarithm of GDP per capita in 2011, with a bound of error on the estimation of $B_L = 1$, if $\sigma_{W;L} \approx 5.194$ and $N = 168$.

Using the formula, we have:

$$n_L = \frac{168(5.194)^2}{167(1^2/4) + (5.194)^2} = 65.94498 \implies n_L \geq 66.$$

Since there are good reasons to trust that the relationship between life expectancy and log GNP per capita in 2011 is approximately linear (see Figure 11.8), the regression approach is a strong one.[46] How does it compare with the example that uses ratio estimation?

[45] **Warning:** Even if formal manipulations can still be performed, the estimate may not be valid **if the relationship between the variables X and Y is not linear**.

[46] Assuming, of course, that μ_X is known; otherwise, it is pretty much useless.

11.5.3 Difference Estimation

Difference estimation is another special case of regression estimation, where the slope β is now assumed to be 1.

If μ_X is known, the **difference estimator** $\hat{\mu}_{Y;D}$ **of** μ_Y computed from a SRS $\mathcal{Y}$ is

$$\hat{\mu}_{Y;D} = \overline{y} + (\mu_X - \overline{x}).$$

Difference estimation is a good strategy when the relationship between X and Y is approximately **linear** and of **slope** 1,[47] as long as the variance of Y along this line is **constant for all** X. Since $\mathcal{Y}$ is drawn according to a SRS, $\mathrm{E}(\overline{x}) = \mu_X$ and $\mathrm{E}(\overline{y}) = \mu_Y$, from which we conclude that

$$\mathrm{E}(\hat{\mu}_{Y;D}) = \mathrm{E}(\overline{y}) + (\mu_X - \mathrm{E}(\overline{x})) = \mu_Y + (\mu_X - \mu_X) = \mu_Y.$$

Consider the random variable $D = Y - X$, whose expectation is

$$\mu_D = \mu_Y - \mu_X.$$

The sample mean of D is thus

$$\overline{d} = \overline{y} - \overline{x} \implies \hat{\mu}_{Y;D} = \mu_X + (\overline{y} - \overline{x}) = \mu_X + \overline{d}.$$

Consequently,

$$\mathrm{V}(\hat{\mu}_{Y;D}) = \mathrm{V}(\mu_X + \overline{d}) = \mathrm{V}(\overline{d}) = \frac{\sigma_D^2}{n}\left(\frac{N-n}{N-1}\right).$$

But

$$\begin{aligned}\sigma_D^2 &= \frac{1}{N}\sum_{j=1}^{N}(D_j - \mu_D)^2 = \frac{1}{N}\sum_{j=1}^{N}\left\{(Y_j - X_j) - (\mu_Y - \mu_X)\right\}^2 \\ &= \frac{1}{N}\sum_{j=1}^{N}\left\{(Y_j - \mu_Y) - (X_j - \mu_X)\right\}^2 = \sigma_Y^2 - 2\rho\sigma_X\sigma_Y + \sigma_X^2,\end{aligned}$$

where $\rho = \frac{\mathrm{Cov}(X,Y)}{\sigma_X\sigma_Y} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}$. As such,

$$\mathrm{V}(\hat{\mu}_{Y;D}) = \frac{\sigma_Y^2 - 2\rho\sigma_X\sigma_Y + \sigma_X^2}{n}\left(\frac{N-n}{N-1}\right).$$

The difference estimator $\hat{\mu}_{Y;D}$ follows **approximately** a normal distribution $\mathcal{N}(\mathrm{E}(\hat{\mu}_{Y;D}), \mathrm{V}(\hat{\mu}_{Y;D}))$, from which we obtain the **bound on the error of estimation**

$$B_D \approx \hat{B}_D = 2\sqrt{\hat{\mathrm{V}}(\hat{\mu}_{Y;D})} \approx 2\sqrt{\frac{s_D^2}{n}\left(1 - \frac{n}{N}\right)},$$

where

$$s_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \overline{d})^2 = s_Y^2 - 2\hat{\rho}s_Xs_Y + s_X^2,$$

[47] Passing or not **through the origin**.

so that $\text{C.I.}_D(\mu_Y; 0.95) : \hat{\mu}_{Y;D} \pm \hat{B}_D$ is an **approximate 95% C.I. for** μ_Y.[48]

Example Auditors are often interested in comparing the audited value Y of items with their book value X. Suppose that $N = 180$ items in inventory have a book value of $\tau_X = 13,320$. A SRS of $n = 10$ items yields the following data:

item i	1	2	3	4	5	6	7	8	9	10
Audit y_i	9	14	7	29	45	109	40	238	60	170
Book x_i	10	12	8	26	47	112	36	240	59	167
$d_i = y_i - x_i$	−1	2	−1	3	−2	−3	4	−2	1	3

Find a 95% C.I. for the mean audit value μ_Y using difference estimation.

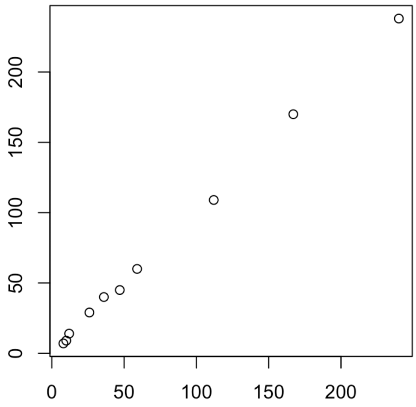


Figure 11.9: Scatterplot of X and Y.

From the scatterplot, we surmise that the slope of the linear fit of Y against X is approximately 1. We must compute $\overline{d}$ and s_D^2:

$$\sum_{i=1}^{10} d_i = 4, \quad \sum_{i=1}^{10} d_i^2 = 58, \implies \overline{d} = \frac{4}{10} \quad \text{and} \quad s_D^2 = \frac{58 - 10(0.4)^2}{10-1} = 6.27.$$

Since $\mu_X = \frac{\tau_X}{N} = \frac{13320}{180} = 74$, the difference estimator is

$$\hat{\mu}_{Y;D} = \mu_X + \overline{d} = 74 + 0.4 = 74.4$$

and the bound is

$$\hat{B}_D \approx 2\sqrt{\hat{\mathrm{V}}(\hat{\mu}_D)} = 2\sqrt{\frac{6.27}{10}\left(1 - \frac{10}{180}\right)} = 1.54,$$

from which

$$\text{C.I.}_D(\mu_Y; 0.95) : \quad 74.4 \pm 1.54 \equiv (72.86, 75.94).$$

Example Consider a bivariate SRS sample $\mathcal{Y} = \{(x_i, y_i)\}$ of size $n = 132$, drawn from a population of size $N = 37,444$. In a preceding study, we

[48] We tackle τ_Y and p_Y in the usual manner.

found that $\mu_X \approx 70.3$. Find a 95% C.I. for μ_Y using difference estimation, assuming that

$$\sum_{i=1}^{n} x_i = 9464.6, \quad \sum_{i=1}^{n} y_i = 14691.6,$$

$$\sum_{i=1}^{n} x_i^2 = 686773.2, \quad \sum_{i=1}^{n} x_i y_i = 1062186, \quad \sum_{i=1}^{n} y_i^2 = 1670194.$$

In a previous example, we have already computed

$$\overline{x} = 71.7, \ \overline{y} \approx 111.3, \ s_X^2 \approx 62.2, \ s_Y^2 \approx 267.3, \ s_{XY} \approx 67.2.$$

The difference estimator is thus

$$\hat{\mu}_{Y;D} = \overline{y} + (\mu_x - \overline{x}) = 111.3 + (70.3 - 71.7) = 109.9,$$

so that

$$\hat{B}_D \approx 2\sqrt{\frac{267.3 - 2(67.2) + 62.2}{132}\left(1 - \frac{132}{37444}\right)} = 2.427,$$

and

$$\text{C.I.}_D(\mu_Y; 0.95) \equiv 109.9 \pm 2.427.$$

Example Find a 95% C.I. for the average life expectancy by country in 2011 μ_Y using the difference method with the logarithm of GDP per capita per country (X), using a sample of size $n = 20$. Assume that $\mu_X = 7.84$ is known.

We draw a sample of size $n = 20$ and compute the various required quantities.

```
set.seed(1234567) # for replicability
N=nrow(gapminder.RLD); n=20
index = sample(1:N,n, replace=FALSE)
samp = gapminder.RLD[index,c("life_expectancy","lgdppc")]
d = samp$life_expectancy - samp$lgdppc
```

```
(mu.X = mean(gapminder.RLD[,"lgdppc"]))
(y.bar = mean(samp$life_expectancy))
(x.bar = mean(samp$lgdppc))
(d.bar = mean(d))
(s2.d = var(d))
```

```
[1] 7.842661
[1] 70.105
[1] 7.577646
[1] 62.52735
[1] 47.69057
```

Note that the regression slope does not seem to be 1 (if that was the case, we would expect $\overline{y}/\overline{x} \approx 1$). Difference estimation is not recommended in this case, but we will continue the example nonetheless.

The bound on the error of estimation and the difference estimate are computed below, and the confidence interval is:

```
B = 2*sqrt(s2.d/n*(1-n/N))
hat.mu.Y.D = y.bar + (mu.X-x.bar)
c(hat.mu.Y.D-B,hat.mu.Y.D+B)
```

```
[1] 67.47129 73.26874
```

In spite of the difference estimation assumptions not being met, the 95% C.I. for Y does contain the true value, $\mu_Y = 70.95$! A happy coincidence, no more.

Sample Size

As with the other methods, we can determine the sample size required to achieve a certain bound on the error of estimation.

In order to estimate μ_Y and τ_Y *via* difference estimation, use:

$$B_{\mu;D} \approx 2\sqrt{\frac{\sigma_D^2}{n}\left(\frac{N-n}{N-1}\right)} \iff n_{\mu;D} = \frac{N\sigma_D^2}{(N-1)D_{\mu;D}+\sigma_D^2};$$

$$B_{\tau;D} \approx 2N\sqrt{\frac{\sigma_D^2}{n}\left(\frac{N-n}{N-1}\right)} \iff n_{\tau;D} = \frac{N\sigma_D^2}{(N-1)D_{\tau;D}+\sigma_D^2},$$

where

$$D_{\mu;D} = \frac{B_{\mu;D}^2}{4} \quad \text{and} \quad D_{\tau;D} = \frac{B_{\tau;D}^2}{4N^2}.$$

As we do not know usually know σ_D^2, we often draw a small preliminary sample and use the **empirical variance** s_D^2 as an estimator of σ_D^2.

Warning! Even if formal manipulations can still be performed, **the estimate may not be valid if the relationship between the variables** X **and** Y **is not linear with slope** ≈ 1.

11.5.4 Comparisons

We have already compared the bounds on the error of estimation for SRS, STS (Prop), and STS (Neyman), and discussed contexts in which one might expect a STS to be preferable to an SRS, or a STS (Neyman) preferable to a STS (Prop).

What can be said about ratio, regression, and difference estimation, both compared to SRS and to each other?

Comparaison Between SRS and the Ratio Method

In what context can we expect ratio estimation to perform "well"? Obviously, the relationship between Y and X must at least be **linear** and **pass through the origin**, i.e.,

$$y_i = \beta x_i + \varepsilon_i, \quad i = 1, \dots, n.$$

It is generally assumed that the observations $\{x_i > 0\}$ are fixed, and that the error terms $\{\varepsilon_i\}$ are independent of each other, with

$$\mathrm{E}(\varepsilon_i) = 0 \quad \text{and} \quad \mathrm{V}(\varepsilon_i) = f(x_i)\sigma^2 > 0.$$

The question becomes: what form must $f(x_i)$ take so that the least squares solution $\hat{\beta}$ is **exactly** the estimator r of the ratio R?

If we set

$$\underbrace{\frac{y_i}{\sqrt{f(x_i)}}}_{y_i'} = \beta \underbrace{\frac{x_i}{\sqrt{f(x_i)}}}_{x_i'} + \underbrace{\frac{\varepsilon_i}{\sqrt{f(x_i)}}}_{\varepsilon_i'}, \quad i = 1, \dots, n,$$

we get

$$\mathrm{E}(\varepsilon_i') = \frac{1}{\sqrt{f(x_i)}}\mathrm{E}(\varepsilon) = 0 \quad \text{and} \quad \mathrm{V}(\varepsilon_i') = \frac{1}{f(x_i)}\mathrm{V}(\varepsilon_i') = \frac{f(x_i)\sigma^2}{f(x_i)} = \sigma^2,$$

and the assumptions of the least squares problem are satisfied. The estimator β is obtained by minimizing

$$Q(\beta) = \sum_{i=1}^n (\varepsilon_i')^2 = \sum_{i=1}^n (y_i' - \beta x_i')^2 = \sum_{i=1}^n \frac{1}{f(x_i)}(y_i - \beta x_i)^2;$$

since

$$Q'(\beta) = -2\sum_{i=1}^n \frac{x_i}{f(x_i)}(y_i - \beta x_i),$$

this is equivalent to solving

$$0 = \sum_{i=1}^n \frac{x_i}{f(x_i)}(y_i - \hat{\beta}x_i) \Longleftrightarrow 0 = \sum_{i=1}^n \left(\frac{x_i y_i}{f(x_i)} - \hat{\beta}\frac{x_i^2}{f(x_i)}\right) \Longleftrightarrow \hat{\beta} = \frac{\sum_{i=1}^n \frac{x_i y_i}{f(x_i)}}{\sum_{i=1}^n \frac{x_i^2}{f(x_i)}}.$$

If $\frac{x_i}{f(x_i)} = k > 0$ for all $i = 1, \dots, n$, the estimator $\hat{\beta}$ becomes

$$\hat{\beta} = \frac{k\sum_{i=1}^n y_i}{k\sum_{i=1}^n x_i} = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i} = r.$$

Thus, when the variance of Y along the line $Y = \beta X$ is

$$\mathrm{V}(y_i) = \mathrm{V}(\beta x_i + \varepsilon_i) = \mathrm{V}(\varepsilon_i) = x_i\sigma^2$$

(i.e., the variance of Y is **proportional to** X), the estimator r of the ratio R is exactly the least squares solution, $\hat{\beta} = r$, and we can expect ratio estimation to produce "good" results.

Of course, one can use the ratio estimation method with a SRS $\mathcal{Y}$ to obtain an estimate $\hat{\mu}_{Y;R}$ of μ_Y even if $\mathrm{V}(\varepsilon) \neq x\sigma^2$.

We have already determined the variance of this estimator:

$$\begin{aligned}\mathrm{V}(\hat{\mu}_{Y;R}) &= \mathrm{V}(r\mu_X) = \mu_X^2\mathrm{V}(r) \approx \frac{1}{n}(\sigma_Y^2 + R^2\sigma_X^2 - 2R\rho\sigma_X\sigma_Y)\left(\frac{N-n}{N-1}\right)\\ &= \underbrace{\frac{\sigma_Y^2}{n}\left(\frac{N-n}{N-1}\right)}_{\mathrm{V}(\overline{y}_{\mathrm{SRS}})} + \frac{R^2\sigma_X^2 - 2R\rho\sigma_X\sigma_Y}{n}\left(\frac{N-n}{N-1}\right).\end{aligned}$$

Consequently, $\mathrm{V}(\overline{y}_{\mathrm{SRS}}) \gg \mathrm{V}(\hat{\mu}_{Y;R})$ if and only if $R^2\sigma_X^2 - 2R\rho\sigma_X\sigma_Y \ll 0$, which is to say if

$$\rho \gg \frac{R\sigma_X}{2\sigma_Y} = \frac{\mu_Y\sigma_X}{2\mu_X\sigma_Y} = \frac{1}{2}\cdot\frac{\mathrm{CV}_X}{\mathrm{CV}_Y}.$$

Comparaison Between SRS and the Regression Method

We have already determined the variance of the estimator $\hat{\mu}_{Y;L}$ of μ_Y:

$$\begin{aligned}\mathrm{V}(\hat{\mu}_{Y;L}) &\approx (1-\rho^2)\frac{\sigma_Y^2}{n}\left(\frac{N-n}{N-1}\right) = \underbrace{\frac{\sigma_Y^2}{n}\left(\frac{N-n}{N-1}\right)}_{\mathrm{V}(\overline{y}_{\mathrm{SRS}})} - \rho^2\cdot\frac{\sigma_Y^2}{n}\left(\frac{N-n}{N-1}\right)\\ &= (1-\rho^2)\mathrm{V}(\overline{y}_{\mathrm{SRS}}).\end{aligned}$$

Consequently, $\mathrm{V}(\hat{\mu}_{Y;L}) \ll \mathrm{V}(\overline{y}_{\mathrm{SRS}})$ when $(1-\rho^2)\mathrm{V}(\overline{y}_{\mathrm{SRS}}) \ll \mathrm{V}(\overline{y}_{\mathrm{SRS}})$, which is to say that

$$1-\rho^2 \ll 1 \Longleftrightarrow 0 \ll |\rho| \leq 1.$$

Comparaison Between SRS and the Difference Method

We have already determined the variance of the estimator $\hat{\mu}_{Y;D}$ of μ_Y:

$$\begin{aligned}\mathrm{V}(\hat{\mu}_{Y;D}) &= \frac{\sigma_Y^2 - 2\rho\sigma_X\sigma_Y + \sigma_X^2}{n}\left(\frac{N-n}{N-1}\right)\\ &= \underbrace{\frac{\sigma_Y^2}{n}\left(\frac{N-n}{N-1}\right)}_{\mathrm{V}(\overline{y}_{\mathrm{SRS}})} + \frac{\sigma_X^2 - 2\rho\sigma_X\sigma_Y}{n}\left(\frac{N-n}{N-1}\right).\end{aligned}$$

Consequently, $\mathrm{V}(\hat{\mu}_{Y;D}) \ll \mathrm{V}(\overline{y}_{\mathrm{SRS}})$ when $\sigma_X^2 - 2\rho\sigma_X\sigma_Y \ll 0 \Longleftrightarrow \sigma_X^2 \ll 2\sigma_{XY}$.

Comparaison Between the Ratio, Regression, and Difference Methods

For each of the estimators $\hat{\mu}_{Y;\alpha}$, $\alpha \in \{R, L, D\}$, we have shown that the sampling variance takes the (approximate) form

$$\mathrm{V}(\hat{\mu}_{Y;\alpha}) \approx \mathrm{V}(\overline{y}_{\mathrm{SRS}}) + \frac{A_\alpha}{n}\left(\frac{N-n}{N-1}\right),$$

where

$$A_\alpha = \begin{cases} R^2\sigma_X^2 - 2R\rho\sigma_X\sigma_Y, & \alpha = R \\ -\rho^2\sigma_Y^2, & \alpha = L \\ \sigma_X^2 - 2\rho\sigma_X\sigma_Y, & \alpha = D \end{cases}$$

In general, $\mathrm{V}(\hat{\mu}_{Y;\alpha}) \ll \mathrm{V}(\hat{\mu}_{Y;\gamma})$ if and only if $A_\alpha \ll A_\gamma$; these are the terms that must be compared to one another.

For instance,

$$\begin{aligned} \mathrm{V}(\hat{\mu}_{Y;R}) \gg \mathrm{V}(\hat{\mu}_{Y;L}) &\iff R^2\sigma_X^2 - 2R\rho\sigma_X\sigma_Y \gg -\rho^2\sigma_Y^2 \\ &\iff R^2\sigma_X^2 - 2R\rho\sigma_X\sigma_Y + \rho^2\sigma_Y^2 \gg 0 \\ &\iff (R\sigma_X - \rho\sigma_Y)^2 \gg 0 \iff |R\sigma_X - \rho\sigma_Y| \gg 0 \\ &\iff R \gg \rho\frac{\sigma_Y}{\sigma_X} = \hat{\beta} \quad \text{or} \quad R \ll \hat{\beta} \end{aligned}$$

All things being equal, the regression estimator is preferable to the ratio estimator (according to their bounds on the error of estimation) when **the ratio is quite different from the slope of the regression line**.

Similarly,

$$\begin{aligned} \mathrm{V}(\hat{\mu}_{Y;D}) \gg \mathrm{V}(\hat{\mu}_{Y;L}) &\iff \sigma_X^2 - 2\rho\sigma_X\sigma_Y \gg -\rho^2\sigma_Y^2 \\ &\iff \sigma_X^2 - 2\rho\sigma_X\sigma_Y + \rho^2\sigma_Y^2 \gg 0 \\ &\iff (\sigma_X - \rho\sigma_Y)^2 \gg 0 \iff |\sigma_X - \rho\sigma_Y| \gg 0 \\ &\iff 1 \gg \rho\frac{\sigma_Y}{\sigma_X} = \hat{\beta} \quad \text{or} \quad 1 \ll \hat{\beta}. \end{aligned}$$

All things being equal, the regression estimator is preferable to the difference estimator (according to their bounds on the error of estimation) when **the slope of the regression line takes a value far from** 1.

But the regression estimator is always **at least as good as the other two** since the latter two are special cases of regression estimation.

Finally, we can also compare the estimators by the ratio and by the difference:

$$\begin{aligned} \mathrm{V}(\hat{\mu}_{Y;R}) \gg \mathrm{V}(\hat{\mu}_{Y;D}) &\iff R^2\sigma_X^2 - 2R\rho\sigma_X\sigma_Y \gg \sigma_X^2 - 2\rho\sigma_X\sigma_Y \\ &\iff |R| \neq 1 \quad \text{and} \quad \sigma_X^2 \gg \frac{2}{R+1}\sigma_{XY} \end{aligned}$$

and

$$\mathrm{V}(\hat{\mu}_{Y;D}) \gg \mathrm{V}(\hat{\mu}_{Y;R}) \iff |R| \neq 1 \quad \text{and} \quad \sigma_X^2 \ll \frac{2}{R+1}\sigma_{XY}$$

Otherwise, the variances are of relatively similar magnitude.

11.6 Cluster Sampling

In practice, collecting sample data can require a tremendous amount of **travel**. Imagine a survey where the residents of the entire country are the **target population**, and a range of demographic and health indicators are measured about the **units**:

- age, height, weight, ethnicity, neighborhood, etc;
- blood pressure, blood cholesterol and mercury levels, body-mass index, etc.

Some of the information can be **self-reported by the units** (age, ethnicity, etc.), but in many cases (body-mass index, mercury levels, etc.), data collection requires the use of **health experts** and **specialized equipment**.

If all the sample units are from the Greater Toronto Area (GTA), say, it may be efficient to move the panel of experts (with all the required equipment in a trailer) from site to site, staying 2 weeks at each site. With about 20 sites in the GTA, data collection would take about a year to complete, but the cost of the survey would be greatly reduced: each night, the interviewers would **go home**; the cost of **moving the equipment** would also be minimized because of the small distances involved.

In a national study, where units could be drawn from several jurisdictions and remote locations, this approach is no longer necessarily recommended as it is potentially very expensive. Instead, one could start by taking a **first sample of geographic areas** (cities, regional municipalities, etc.), and then select a **sub-sample of units** (residents) in each of these areas.

Such a strategy is known as **multi-stage sampling** (MnS, see Section 11.7.3). Stratified sampling, for example, is a M2S for which the first level sample is a **census** and the second level sample is a SRS.

As another example, when the first level sample comes from a SRS and the second level sample is a **census** (all units are selected), we speak of **cluster sampling** (CLS).

11.6.1 Estimators and Confidence Intervals

As it was the case in the second chapter, we are interested in a finite population $\mathcal{U} = \{u_1, \ldots, u_N\}$ of expectation μ and variance σ^2.

Suppose we can cover the population with M disjoint **clusters** containing, respectively, $N_1, \ldots, N_M$ units, so that $N_1 + \cdots + N_M = N$:

$$\mathcal{G}_1 = \{u_{1,1}, \ldots, u_{1,N_1}\}, \ \cdots \ , \ \mathcal{G}_M = \{u_{M,1}, \ldots, u_{M,N_M}\},$$

with cluster **expectation**, **total**, and **variance** given by

$$\mu_i = \frac{1}{N_i}\sum_{j=1}^{N_i} u_{i,j}, \quad \tau_i = N_i\mu_i, \quad \text{and} \quad \sigma_i^2 = \frac{1}{N_i}\sum_{j=1}^{N_i} u_{i,j}^2 - \mu_i^2, \quad 1 \leq i \leq M.$$

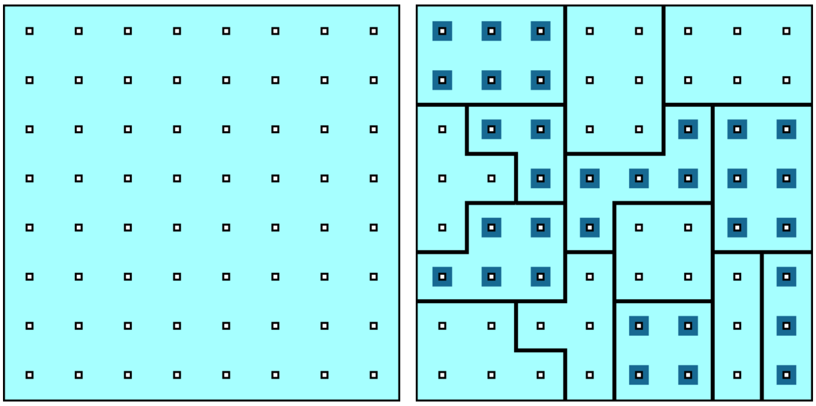


Figure 11.10: Schematics of CLS: target population (left) and sample (right).

A **cluster random sample** (CLS) $\mathcal{Y}$ is a subset of the target population $\mathcal{U}$ which is obtained by first drawing a SRS of $m > 1$ clusters, and then selecting all units in the selected clusters:

$$\mathcal{G}_{i_1} \cup \cdots \cup \mathcal{G}_{i_m} = \{\underbrace{y_{i_1,1}, \ldots, y_{i_1,N_{i_1}}}_{\text{cluster } \mathcal{G}_{i_1}}, \ldots, \underbrace{y_{i_m,1}, \ldots, y_{i_m,N_{i_m}}}_{\text{cluster } \mathcal{G}_{i_m}}\} \subseteq \bigcup_{\ell=1}^{M} \mathcal{G}_\ell = \mathcal{U}.$$

When $\mathcal{G}_{i_k}$ belongs to the CLS $\mathcal{Y}$, we denote its **mean**, **total**, and **variance** by $\overline{y}_{i_k}$, y_{i_k}, and $s^2_{i_k}$, respectively, for $1 \leq k \leq m$.

In a CLS design, each observation **has the same probability of being selected**, but the sample size **may change from one CLS to another**, unless the clusters all have the same size in the first place.

Estimating the Mean μ for Clusters of Equal Size

Let us assume that all clusters have the same size: $N_1 = \cdots = N_M = n \implies N = Mn$. The **cluster mean** of the sample observations in $\mathcal{Y}$ is an estimator of μ:

$$\overline{y}_C = \frac{1}{mn}\sum_{k=1}^{m}\sum_{j=1}^{n} y_{i_k,j} = \frac{1}{mn}\sum_{k=1}^{m} y_{i_k} = \frac{1}{m}\sum_{k=1}^{m} \overline{y}_{i_k} = \frac{1}{m}\sum_{k=1}^{m} \mu_{i_k}.$$

Therefore, the cluster average is simply the **average of the selected cluster averages**. This is not surprising since

$$\mu = \frac{1}{N}\sum_{\ell=1}^{M}\sum_{j=1}^{n} u_{\ell,j} = \frac{1}{Mn}\sum_{\ell=1}^{M}\sum_{j=1}^{n} u_{\ell,j} = \frac{1}{Mn}\sum_{\ell=1}^{M} \tau_\ell = \frac{1}{M}\sum_{\ell=1}^{M} \mu_\ell.$$

We can easily show that $\overline{y}_C$ is an **unbiased estimator** of μ:

$$\mathrm{E}(\overline{y}_C) = \frac{1}{m}\sum_{k=1}^{m} \mathrm{E}(\mu_{i_k}) = \frac{1}{m}\sum_{k=1}^{m} \mu = \mu.$$

Furthemore, its **sampling variance** is

$$\mathrm{V}(\overline{y}_C) = \frac{\sigma_C^2}{m}\left(\frac{M-m}{M-1}\right), \quad \text{where } \sigma_C^2 = \frac{1}{M}\sum_{\ell=1}^{M}(\mu_\ell - \mu)^2,$$

since clusters are drawn using an SRS. Indeed, $\overline{y}_C$ is the mean of a SRS with m:

$$\{\mu_{i_1},\ldots,\mu_{i_m}\} \subseteq \{\mu_1,\ldots,\mu_M\}.$$

Central Limit Theorem – CLS: if m and $M-m$ are sufficiently large, then

$$\overline{y}_C \sim_{\text{approx.}} \mathcal{N}\left(\text{E}(\overline{y}_C),\text{V}(\overline{y}_C)\right) = \mathcal{N}\left(\mu, \frac{\sigma_C^2}{m}\left(\frac{M-m}{M-1}\right)\right).$$

In a CLS, the **bound on the error of estimation** is thus

$$B_{\mu;C} = 2\sqrt{\text{V}(\overline{y}_C)} = 2\sqrt{\frac{\sigma_C^2}{m}\left(\frac{M-m}{M-1}\right)},$$

and the corresponding **95% C.I. for** μ is simply

$$\text{C.I.}_C(\mu;0.95): \quad \overline{y}_C \pm B_{\mu;C}.$$

In practice, the **variance of the cluster means** σ_C^2 is rarely known – the empirical variance (and the corresponding **correction factor**) is used instead:

$$\hat{\text{V}}(\overline{y}_C) = \frac{s_C^2}{m}\left(1-\frac{m}{M}\right), \text{ where } s_C^2 = \frac{1}{m-1}\sum_{k=1}^{m}(\overline{y}_{i_k}-\overline{y}_C)^2.$$

The **bound on the error of estimation** is then approximated by

$$B_{\mu;C} \approx \hat{B}_{\mu;C} = 2\sqrt{\hat{\text{V}}(\overline{y}_C)} = 2\sqrt{\frac{s_C^2}{m}\left(1-\frac{m}{M}\right)},$$

$$\Longrightarrow \text{C.I.}_C(\mu;0.95): \quad \overline{y}_C \pm \hat{B}_{\mu;C} \equiv \overline{y}_C \pm 2\sqrt{\frac{s_C^2}{m}\left(1-\frac{m}{M}\right)}.$$

Example Consider a finite population $\mathcal{U}$ of size $N = 37,444$, divided into $M = 44$ clusters $\mathcal{G}_\ell$, each of size $n = 851$. We draw a SRS of $m = 6$ clusters. The means of these clusters are:

$$\overline{y}_1 = 120.7, \overline{y}_2 = 75.2, \overline{y}_3 = 116.3, \overline{y}_4 = 111.1, \overline{y}_5 = 116.9, \overline{y}_6 = 96.6.$$

Find a 95% C.I. for the mean μ.

The bound on the error of estimation for μ is $\approx \hat{B}_{\mu;C} = 2\sqrt{\hat{\text{V}}(\overline{y}_C)}$; we see that

$$\overline{y}_C = \tfrac{1}{6}\textstyle\sum_{k=1}^{6}\overline{y}_k \approx 106.1,\ s_C^2 = \tfrac{1}{6-1}\textstyle\sum_{k=1}^{6}(\overline{y}_k-\overline{y}_C)^2 = \frac{69089.6-6(106.1)^2}{6-1} \approx 300.8,$$

from which we have

$$\text{C.I.}_C(\mu;0.95) \approx 106.1 \pm 2\sqrt{\tfrac{300.8}{6}\left(1-\tfrac{6}{44}\right)} \equiv (93.0, 119.3).$$

Estimating the Mean μ for Clusters of Different Sizes

In practice, the clusters are often all of **different** sizes, so we could write

$$\mu = \frac{\sum_{\ell=1}^{M}\sum_{j=1}^{N_\ell} u_{\ell,j}}{\sum_{\ell=1}^{M} N_\ell} = \frac{\sum_{\ell=1}^{M}\tau_\ell}{\sum_{\ell=1}^{M} N_\ell},$$

where τ_ℓ is the sum of $u_{\ell,j}$ for units in the cluster $\mathcal{G}_\ell$, $1 \leq \ell \leq M$.[49]

If we still draw m clusters from the population of M clusters using an SRS, the form of μ suggests the use of the following estimator:

$$\overline{y}_C = \frac{\sum_{k=1}^{m}\sum_{j=1}^{N_{i_k}} y_{i_k,j}}{\sum_{k=1}^{m} N_{i_k}} = \frac{\sum_{k=1}^{m} y_{i_k}}{\sum_{k=1}^{m} N_{i_k}},$$

where we are using the notation of Section 11.5.

If the **average cluster size** is $\overline{N} = \frac{N}{M}$, this is similar to the situation that leads to **ratio estimation of the mean**. By performing the mapping $(\overline{y}_C, \mu, \overline{N}, \tau_\ell, N_\ell) \leftrightsquigarrow (r, R, \mu_X, Y_j, X_j)$, we can therefore conclude that $\overline{y}_C$ is a **biased estimator** of μ, whose **sampling variance** is

$$\mathrm{V}(\overline{y}_C) \approx \frac{1}{\overline{N}^2}\cdot\frac{1}{m}\left(\frac{M-m}{M-1}\right)\cdot\frac{1}{M}\sum_{\ell=1}^{M}\underbrace{(\tau_\ell - \mu N_\ell)}_{=N_\ell(\mu_\ell-\mu)}{}^2.$$

Consequently, the **bound on the error of estimation** is given by

$$B_{\mu;C} = 2\sqrt{\mathrm{V}(\overline{y}_C)} \approx 2\sqrt{\frac{1}{\overline{N}^2}\cdot\frac{1}{m}\left(\frac{M-m}{M-1}\right)\cdot\frac{1}{M}\sum_{\ell=1}^{M}(\tau_\ell - \mu N_\ell)^2}.$$

In practice, we often only have access to the sampled clusters – we must then use the **empirical variance**:

$$\begin{aligned}
\hat{\mathrm{V}}(\overline{y}_C) &\approx \frac{1}{\overline{N}^2}\cdot\frac{1}{m}\left(1-\frac{m}{M}\right)\cdot\frac{1}{m-1}\sum_{k=1}^{m}(y_{i_k} - \overline{y}_C N_{i_k})^2 \\
&= \frac{1}{\overline{N}^2}\cdot\frac{1}{m}\left(1-\frac{m}{M}\right)(s_Y^2 + s_N^2\overline{y}_C^2 - 2\overline{y}_C\hat{\rho}s_N s_Y), \quad \text{where} \\
s_Y^2 &= \frac{1}{m-1}\sum_{k=1}^{m}(y_{i_k} - \overline{\overline{y}})^2, \quad s_N^2 = \frac{1}{m-1}\sum_{k=1}^{m}(N_{i_k} - \overline{N})^2, \\
\hat{\rho} &= \frac{\sum_{k=1}^{m}(y_{i_k} - \overline{\overline{y}})(N_{i_k} - \overline{N})}{\sqrt{\sum_{k=1}^{m}(y_{i_k} - \overline{\overline{y}})^2\sum_{k=1}^{m}(N_{i_k} - \overline{N})^2}}, \quad \overline{\overline{y}} = \frac{1}{m}\sum_{k=1}^{m} y_{i_k}.
\end{aligned}$$

Since it is not always possible to determine the average $\overline{N}$ of the clusters in the population $\mathcal{U}$, we often use $\overline{n}$, the **average cluster size in the**

[49] If $N_1 = \cdots = N_M = n$, the formulas we will develop will collapse to those seen in the preceding section.

sample $\mathcal{Y}$ instead:

$$\overline{n} = \frac{N_{i_1} + \cdots + N_m}{m}.$$

The **bound on the error of estimation** is thus

$$\hat{B}_{\mu;C} \approx 2\sqrt{\frac{1}{\overline{n}^2} \cdot \frac{1}{m}\left(1 - \frac{m}{M}\right)(s_Y^2 + s_N^2\overline{y}_C^2 - 2\overline{y}_C\hat{\rho}s_N s_Y)}$$

and the **approximate 95% C.I. for** μ is

$$\text{C.I.}_C(\mu; 0.95): \quad \overline{y}_C \pm \hat{B}_{\mu;C}.$$

Example Consider a finite population $\mathcal{U}$ of size $N = 37,444$, divided into $M = 44$ clusters $\mathcal{G}_\ell$. We draw a SRS of $m = 6$ clusters. The means of the observations in these clusters are:

k	1	2	3	4	5	6
$\overline{y}_k$	120.7	75.2	116.3	111.1	116.9	96.6
N_k	850	176	1011	1001	843	910

Find a 95% C.I. for the mean μ.

The bound on the error of estimation is $\approx \hat{B}_{\mu;C} = 2\sqrt{\hat{\text{V}}(\overline{y}_C)}$; we see that

$$\overline{y}_C = \frac{\sum_{k=1}^6 N_k\overline{y}_k}{\sum_{k=1}^6 N_k} = \frac{531073.3}{4791} \approx 110.8, \quad \overline{n} = \frac{1}{6}\sum_{k=1}^{6} N_k = \frac{4791}{6} = 798.5$$

$$\overline{\overline{y}} = \frac{\sum_{k=1}^6 N_k\overline{y}_k}{6} = \frac{531073.3}{6} = 88,512.2,$$

$$s_N^2 = \frac{1}{6-1}\sum_{k=1}^{6}(N_k - \overline{n})^2 = 98,146.7$$

$$s_Y^2 = \frac{1}{6-1}\sum_{k=1}^{6}(N_k\overline{y}_k - \overline{\overline{y}})^2 = 1,465,229,403.4,$$

$$\hat{\rho} = \frac{\sum_{k=1}^6 (N_k - \overline{n})(N_k\overline{y}_k - \overline{\overline{y}})}{\sqrt{\sum_{k=1}^6 (N_k - \overline{n})^2 \sum_{k=1}^6 (N_k\overline{y}_k - \overline{\overline{y}})^2}} \approx 0.9796$$

$$s_Y^2 + s_N^2\overline{y}_C^2 - 2\overline{y}_C\hat{\rho}s_N s_Y = 66,814,598.95$$

from which we conclude that

$$\hat{\text{V}}(\overline{y}_C) = \frac{1}{798.5^2} \cdot \frac{1}{6}\left(1 - \frac{6}{44}\right)(66,814,598.95) \approx 15.1$$

and $\text{C.I.}_C(\mu; 0.95) \approx 110.8 \pm 2\sqrt{15.1} \equiv (103.1, 118.6)$.

Example Find a 95% C.I. for the average life expectancy by country in 2011 (including India and China), using a CLS of size $m = 8$, assuming that the $N = 185$ countries have been grouped into $M = 22$ **clusters** determined by **geographic regions**.

We re-use the code from the previous sections,[50] The cluster sizes in the population are as follows.

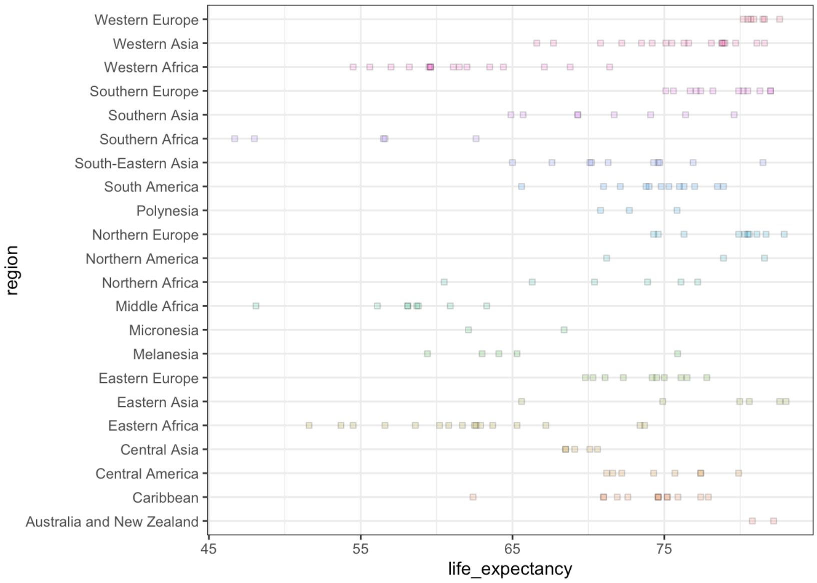
[50] With some modifications, in particular with respect to the **clusters** (`region`.

```
gapminder.CLS <- gapminder |>  filter(year==2011) |> select(life_expectancy, region)
summary(gapminder.CLS,22)
```

```
 life_expectancy                         region
 Min.   :46.70   Australia and New Zealand: 2
 1st Qu.:65.30   Caribbean                :13
 Median :73.70   Central America          : 8
 Mean   :71.18   Central Asia             : 5
 3rd Qu.:77.40   Eastern Africa           :16
 Max.   :83.02   Eastern Asia             : 6
                 Eastern Europe           :10
                 Melanesia                : 5
                 Micronesia               : 2
                 Middle Africa            : 8
                 Northern Africa          : 6
                 Northern America         : 3
                 Northern Europe          :10
                 Polynesia                : 3
                 South America            :12
                 South-Eastern Asia       :10
                 Southern Africa          : 5
                 Southern Asia            : 8
                 Southern Europe          :12
                 Western Africa           :16
                 Western Asia             :18
                 Western Europe           : 7
```

We note that the average life expectancy is $\mu = 71.18$. We can explore the distribution of life expectancy by cluster using the following code:

```
ggplot(data=gapminder.CLS, aes(x=life_expectancy, y=region, fill=region)) +
    geom_point(col="black", alpha=.2,pch=22) +
    theme(legend.title = element_blank(), legend.position="none")
```



We notice a significant variability between some clusters (Southern Africa vs. Southern Europe, for example), but there is still a lot of overlap (which is a good sign). Next, we draw a SRS of $m = 8$ clusters:

```
set.seed(12345) # for replicability
regions=unique(gapminder.CLS[,"region"])
M=length(regions); m=8
(sample.reg = sample(1:M,m, replace=FALSE))
```

```
[1] 14 19 16 11  2 21  6  7
```

We provide a summary of the observations in the sampled clusters:

```
sample.ind = gapminder.CLS$region %in% regions[sample.reg]
gapminder.CLS.n = gapminder.CLS[sample.ind,]
gapminder.CLS.n$region <- as.factor(gapminder.CLS.n$region)
(summ = gapminder.CLS.n |> group_by(region) |>
    summarise(N=n(), y.bar=mean(life_expectancy),
              total.y=sum(life_expectancy)))
```

```
# A tibble: 8 × 4
  region                        N y.barre total.y
  <fct>                     <int>   <dbl>   <dbl>
1 Australia and New Zealand     2    81.5    163
2 Central America               8    75.0    600.
3 Central Asia                  5    69.4    347.
4 Melanesia                     5    65.5    328.
5 Northern Africa               6    70.7    424.
6 Northern America              3    77.2    232.
7 South-Eastern Asia           10    72.6    726.
8 Western Asia                 18    75.8   1364.
```

We can also produce a summary of this summary:

```
(summ.final = summ |>
    summarise(sum.N = sum(N), moy.N = mean(N),
              y.bar.bar = mean(total.y),
              sum.y.bar = sum(total.y)))
```

```
# A tibble: 1 × 4
  sum.N moy.N y.barre.barre sum.y.barre
  <int> <dbl>         <dbl>       <dbl>
1    57  7.12          523.       4184.
```

We can now calculate the cluster estimator:

```
(est.y.bar.G=summ.final$sum.y.bar/summ.final$sum.N)
```

```
[1] 73.40316
```

Next, its sampling variance:

```
s2.Y = var(summ$total.y)
s2.N = var(summ$N)
rho = cor(summ$N,summ$total.y)
V.est.y.G = 1/summ.final$moy.N^2*1/m*(1-m/M)*
    (s2.Y+s2.N*est.y.bar.G^2-
    2*est.y.bar.G*rho*sqrt(s2.N*s2.Y))
```

The bound on the error of estimation and the 95% C.I. for μ are:

```
B = 2*sqrt(V.est.y.G)
c(est.y.bar.G - B,est.y.bar.G + B)
```

```
[1] 71.35310 75.45321
```

The performance of CLS is generally worse than that of SRS and/or STS – no surprise, given the discussion at the beginning of this section. The nature of the clusters may also play a role (in contrast to STS, CLS is more efficient when the cluster structure is **similar from one cluster to another**), which is not really the case here. We will discuss this further.

Estimating the Total τ

Most of the work has already been done: since the **total** τ can be rewritten as

$$\tau = \sum_{j=1}^{N} u_j = N\mu,$$

we can estimate the total with a CLS using the formula

$$\hat{\tau}_C = N\overline{y}_C.$$

There are two possibilities: either $N_1 = \cdots = N_M = n$, or the clusters are not all the same size.

If $N_1 = \cdots = N_M = n$, we have an **unbiased** estimator of τ:

$$\mathrm{E}(\hat{\tau}_C) = \mathrm{E}(N\overline{y}_C) = N \cdot \mathrm{E}(\overline{y}_C) = N\mu = \tau,$$

$$\mathrm{V}(\hat{\tau}_C) = N^2 \cdot \mathrm{V}(\overline{y}_C) = N^2 \cdot \frac{\sigma_C^2}{m}\left(\frac{M-m}{M-1}\right) \approx N^2 \cdot \hat{\mathrm{V}}(\overline{y}_C) = N^2 \cdot \frac{s_C^2}{m}\left(1-\frac{m}{M}\right).$$

If the clusters are of different sizes, we have a **biased** estimator of τ, with **sampling variance** given by

$$\begin{aligned}
\mathrm{V}(\hat{\tau}_C) &= \mathrm{V}(N\overline{y}_C) = N^2 \cdot \mathrm{V}(\overline{y}_C) \approx N^2 \cdot \hat{\mathrm{V}}(\overline{y}_C) \\
&= \frac{N^2}{\overline{N}^2} \cdot \frac{1}{m}\left(1-\frac{m}{M}\right)(s_Y^2 + s_N^2\overline{y}_C^2 - 2\overline{y}_C\hat{\rho}s_N s_Y) \\
&= M^2 \cdot \frac{1}{m}\left(1-\frac{m}{M}\right)(s_Y^2 + s_N^2\overline{y}_C^2 - 2\overline{y}_C\hat{\rho}s_N s_Y).
\end{aligned}$$

The estimator follows an approximate normal distribution

$$\hat{\tau}_C \sim_{\text{approx}} \mathcal{N}\left(\text{E}(\hat{\tau}_C), \hat{\text{V}}(\hat{\tau}_C)\right),$$

as long as the quantities m, and $M - m$ are both "large enough".

In both cases, the **bound on the error of estimation** is

$$B_{\tau;C} \approx \hat{B}_{\tau;C} = 2\sqrt{\hat{\text{V}}(\hat{\tau}_C)}$$

and the **95% C.I. for** τ takes the usual form:

$$\text{C.I.}_C(\tau; 0.95): \quad \hat{\tau}_C \pm \hat{B}_{\tau;C}.$$

Example Consider a finite population $\mathcal{U}$ of size $N = 37,444$, divided into $M = 44$ clusters $\mathcal{G}_\ell$, each of size $n = 851$. We draw a SRS of $m = 6$ clusters. The means of the observations in these clusters are:

$$\overline{y}_1 = 120.7,\ \overline{y}_2 = 75.2,\ \overline{y}_3 = 116.3,\ \overline{y}_4 = 111.1,\ \overline{y}_5 = 116.9,\ \overline{y}_6 = 96.6.$$

Find a 95% C.I. for the total τ in $\mathcal{U}$.

We have previously seen that $\text{C.I.}_C(\mu; 0.95) \equiv (93.0, 119.3)$ for this CLS, with $N_1 = \cdots = N_6 = 851$. Therefore,

$$\text{C.I.}_C(\tau; 0.95) \approx 37444(93.0, 119.3) \equiv (3481307.7, 4466805.3).$$

Example Consider a finite population $\mathcal{U}$ of size $N = 37,444$, divided into $M = 44$ clusters $\mathcal{G}_\ell$. We draw a SRS of $m = 6$ clusters. The mean of the observations in these clusters are:

k	1	2	3	4	5	6
$\overline{y}_k$	120.7	75.2	116.3	111.1	116.9	96.6
N_k	850	176	1011	1001	843	910

Find a 95% C.I. for the total τ in $\mathcal{U}$.

We have already seen in a previous example that $\text{C.I.}_C(\mu; 0.95) \equiv (103.1, 118.6)$ for this CLS with different cluster sizes. Therefore,

$$\text{C.I.}_C(\tau; 0.95) \approx 37444(103.1, 118.6) \equiv (3860476, 4440858).$$

WARNING: how do we do this **if the size** N **of the population is unknown?** Note that

$$\tau = \sum_{\ell=1}^{M} \tau_\ell = M \cdot \frac{1}{M} \sum_{\ell=1}^{M} \tau_\ell = M\overline{\tau},$$

where $\overline{\tau}$ is **mean of the cluster totals in the population**.

We could then use the estimator

$$M\overline{y}_T = M \cdot \frac{1}{m} \sum_{k=1}^{m} y_{i_k},$$

where $\overline{y}_T$ is the **mean of the m cluster totals in the CLS**.

In that case, we are dealing with a SRS of size m, drawn from M cluster totals, i.e., this is an **unbiased** estimator:

$$\begin{aligned}
&\mathrm{E}(M\overline{y}_T) = \tau \\
&\mathrm{V}(M\overline{y}_T) = M^2 \cdot \mathrm{V}(\overline{y}_T) = M^2 \cdot \frac{\sigma_T^2}{m}\left(\frac{M-m}{M-1}\right) \\
&\hat{\mathrm{V}}(M\overline{y}_T) \approx M^2 \cdot \hat{\mathrm{V}}(\overline{y}_T) = M^2 \cdot \frac{s_T^2}{m}\left(1-\frac{m}{M}\right),
\end{aligned}$$

where

$$\sigma_T^2 = \frac{1}{M}\sum_{\ell=1}^{M}(\tau_\ell - \overline{\tau})^2 \quad \text{and} \quad s_T^2 = \frac{1}{m-1}\sum_{k=1}^{m}(y_{i_k} - \overline{y}_T)^2.$$

The estimator follows an approximate normal distribution

$$M\overline{y}_T \sim_{\text{approx}} \mathcal{N}\left(\tau, \hat{\mathrm{V}}(M\overline{y}_T)\right),$$

as long as the quantities m, and $M - m$ are both "large enough".

The **bound on the error of estimation** is then

$$B_{\tau;T} \approx \hat{B}_{\tau;T} = 2\sqrt{\hat{\mathrm{V}}(M\overline{y}_T)}$$

and the 95% C.I. for τ takes the usual form:

$$\text{C.I.}_T(\tau; 0.95): \quad M\overline{y}_T \pm \hat{B}_{\tau;T}.$$

Example Consider a finite population $\mathcal{U}$ of unknown size, divided into $M = 44$ clusters $\mathcal{G}_\ell$. We draw a SRS of $m = 6$ clusters. The mean of the observations in these clusters are:

k	1	2	3	4	5	6
$\overline{y}_k$	120.7	75.2	116.3	111.1	116.9	96.6
N_k	850	176	1011	1001	843	910

Find a 95% C.I. for the total τ in $\mathcal{U}$.

Since the population size N is unknown, the bound on the error of estimation for τ is $\approx \hat{B}_{\tau;T} = 2\sqrt{\hat{\mathrm{V}}(M\overline{y}_T)}$; we see that

$$\overline{y}_T = \frac{1}{6}\sum_{k=1}^{6} N_k\overline{y}_k = \frac{531073.3}{6} \approx 88512.2, \quad M\overline{y}_T = 44(88512.2) = 3894537.5$$

and

$$s_T^2 = \frac{1}{6-1}\sum_{k=1}^{6}(N_k\overline{y}_k - \overline{y}_T)^2 = \frac{1}{5}\left(\sum_{k=1}^{6} N_k^2\overline{y}_k^2 - 6\overline{y}_T^2\right) = 1465229403,$$

from which we conclude that

$$\hat{\mathrm{V}}(M\overline{y}_T) = (44)^2 \cdot \frac{1465229403}{6}\left(1 - \frac{6}{44}\right) = 408310593755.73$$

and $\mathrm{C.I.}_T(\tau; 0.95) \approx 3894537.5 \pm 2\sqrt{408310593755.73} \equiv (2616554, 5172521)$.

The estimator is unbiased, but the confidence interval for τ is much wider than that given by $\mathrm{C.I.}_C(\tau; 0.95) \equiv (3860476, 4440858)$; this is not surprising since we have more information in the latter case (namely, the size N of the population).

Example Find a 95% C.I. for the world population in 2011 (excluding China and India), using a CLS of size $m = 8$, drawn from $M = 22$ clusters determined by geographic regions.

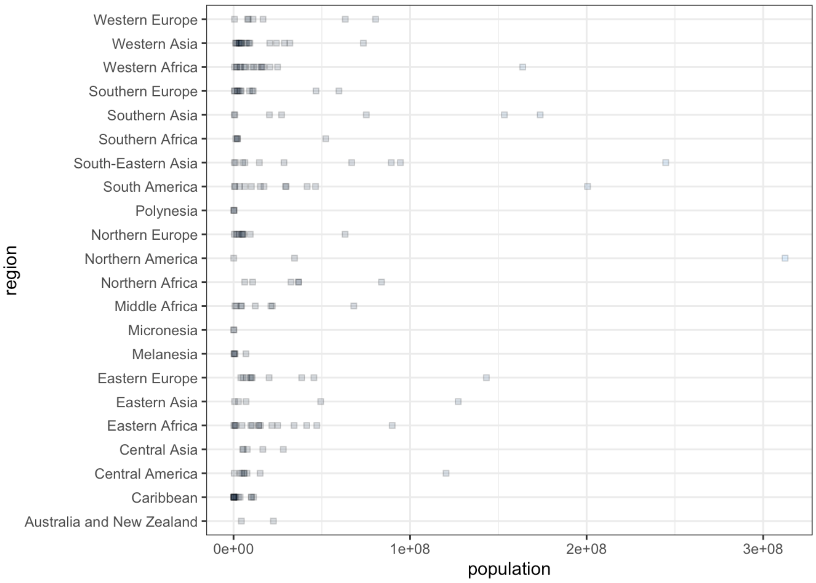
We re-use the code from the previous sections to create the clusters. The true population total is found below:

```
gapminder.CLS.pop <- gapminder |> filter(year==2011) |>
    select(population, region) |>
    filter(population < 500000000)
(sum(gapminder.CLS.pop$population))
```

```
[1] 4264258312
```

We start by studying the distribution of population by region:

```
ggplot(data=gapminder.CLS.pop, aes(x=population, y=region,
    fill=population)) +
    geom_point(col="black", alpha=.2,pch=22) +
    theme(legend.title = element_blank(),
          legend.position="none")
```



The essential statistics are calculated as follows:

```
summ.pop = gapminder.CLS.pop |> group_by(region) |>
    summarise(N=n(), y.pop=mean(population),
              tau.pop=sum(population))
```

Next we draw a SRS of clusters:

```
set.seed(22) # for replicability
regions = unique(gapminder.CLS.pop[,"region"])
M=length(regions); m=8
N=nrow(gapminder.CLS.pop)
(sample.reg = sample(1:M,m, replace=FALSE))
```

```
[1]  6  9 10 12 17  5 11  3
```

The sample is summarized as follows:

```
sample.ind = gapminder.CLS.pop$region %in%
        regions[sample.reg]
gapminder.CLS.T = gapminder.CLS.pop[sample.ind,]
gapminder.CLS.T$region <- as.factor(gapminder.CLS.T$region)
(summ.T = gapminder.CLS.T |> group_by(region) |>
    summarise(N=n(), tau=sum(population)))
```

```
# A tibble: 8 × 3
  region              N       tau
  <fct>           <int>     <int>
1 Central America     8 163510619
2 Eastern Europe     10 294249971
3 Middle Africa       8 134483803
4 Northern Europe    10  99989705
5 South America      12 401182686
6 Southern Asia       7 450825356
7 Western Africa     16 316604189
8 Western Asia       18 237909741
```

If we assume the number of units in the population to be known ($N = 183$), the estimator of the average population per country is:

```
(y.G = sum(summ.T$tau)/sum(summ.T$N))
```

```
[1] 23581529
```

The estimator for the total population (excluding China and India) is:

```
(tau.G = N*y.G)
```

```
[1] 4315419784
```

The bound on the error of estimation and the 95% C.I. for τ are:

```
s2.G =  1/(m-1)*sum((summ.T$tau-y.G*summ.T$N)^2)
V = M^2*s2.G/m*(1-m/M)
B = 2*sqrt(V)
c(tau.G-B,tau.G+B)
```

```
[1] 2441918142 6188921427
```

If we assume instead that the number of units is unknown, the estimator of the population per cluster is:

```
(y.T = sum(summ.T$tau)/m)
```

```
[1] 262344509
```

The estimator for the total population (excluding China and India) would then be:

```
(tau.T = M*y.T)
```

```
[1] 5771579192
```

The bound on the error of estimation and the 95% C.I. for τ in that case are computed below:

```
s2.T =  1/(m-1)*sum((summ.T$tau-y.T)^2)
V = M^2*s2.T/m*(1-m/M)
B = 2*sqrt(V)
c(tau.G-B,tau.G+B)
```

```
[1] 2746857662 5883981906
```

The actual value $\tau = 4,264,258,312$ is found within the 95% C.I.[51]

Estimating a Proportion p

In a population where $A_{\ell,j} \in \{0,1\}$ represents the absence or presence of a characteristic for the jth unit in the ℓth cluster, the **mean**

$$p = \frac{1}{N}\sum_{\ell=1}^{M}\sum_{j=1}^{N_\ell} A_{\ell,j} = \frac{\sum_{\ell=1}^{M} A_\ell}{\sum_{\ell=1}^{M} N_\ell}$$

is the **proportion** of the population units possessing the characteristic, where A_ℓ is the number of units with the characteristic in the ℓth cluster.

[51] But different SRS of clusters might lead to different outcomes.

If we are still drawing m clusters using a SRS from the M clusters in the population, the form taken by p suggests the use of the following estimator:

$$\hat{p}_C = \frac{\sum_{k=1}^{m}\sum_{j=1}^{N_{i_k}} a_{i_k,j}}{\sum_{k=1}^{m} N_{i_k}} = \frac{\sum_{k=1}^{m} a_{i_k}}{\sum_{k=1}^{m} N_{i_k}},$$

where a_{i_k} is the number of units with the characteristic in the kth sampled cluster.

Set $\overline{N} = \frac{N}{M}$. If N is unknown, we use $\overline{N} \approx \overline{n} = \frac{1}{m}(N_{i_1} + \cdots + N_{i_m})$. There are then two possibilities: either $N_1 = \cdots = N_M = n$, or the clusters are not all of the same size. If $N_1 = \cdots = N_M = n$, we have an **unbiased** estimator of p:

$$\mathrm{E}(\hat{p}_C) = p, \quad \mathrm{V}(\hat{p}_C) = \frac{1}{n^2}\cdot\frac{\sigma_P^2}{m}\left(\frac{M-m}{M-1}\right) \approx \frac{1}{n^2}\cdot\frac{s_P^2}{m}\left(1-\frac{m}{M}\right) = \hat{\mathrm{V}}(\hat{p}_C),$$

where

$$\sigma_P^2 = \frac{1}{M}\sum_{\ell=1}^{M}(A_\ell - pN_\ell)^2 \quad \text{and} \quad s_P^2 = \frac{1}{m-1}\sum_{k=1}^{m}(a_{i_k} - \hat{p}_C N_{i_k})^2.$$

If the clusters are of different sizes, we have a **biased** estimator of p, whose **sampling variance** is:

$$\mathrm{V}(\hat{p}_C) \approx \frac{1}{\overline{N}^2}\cdot\frac{\sigma_P^2}{m}\left(\frac{M-m}{M-1}\right), \quad \hat{\mathrm{V}}(\hat{p}_C) \approx \frac{1}{\overline{n}^2}\cdot\frac{s_P^2}{m}\left(1-\frac{m}{M}\right).$$

The estimator follows an approximate normal distribution

$$\hat{p}_C \sim_{\text{approx}} \mathcal{N}\left(\mathrm{E}(\hat{p}_C), \hat{\mathrm{V}}(\hat{p}_C)\right),$$

as long as the quantities m, and $M - m$ are both "large enough".

In both cases, the **bound on the error of estimation** is

$$B_{p;C} \approx \hat{B}_{p;C} = 2\sqrt{\hat{\mathrm{V}}(\hat{p}_C)}$$

and the **95% C.I. for** p takes the usual form:

$$\text{C.I.}_C(p; 0.95): \quad \hat{p}_C \pm \hat{B}_{p;C}.$$

Example Find a 95% C.I. for the proportion of countries whose life expectancy is above 75 years in 2011, using a CLS with $m = 8$, assuming that the countries are grouped into $M = 22$ clusters determined by geographic regions.

We re-use the code of the previous sections to create the clusters, and we create a new indicator variable for the 75 years life expectancy threshold:

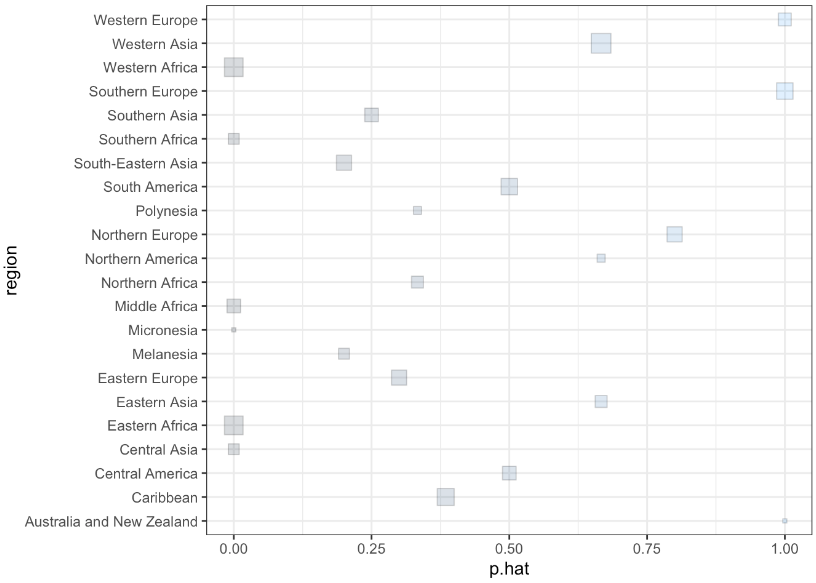
```
gapminder.CLS$life.75 <- ifelse(
        gapminder.CLS$life_expectancy>75,1,0)
gapminder.CLS.75 <- gapminder.CLS |> select(life.75,region)
(mean(gapminder.CLS.75$life.75)) # true proportion
```

```
[1] 0.3945946
```

We begin by examining the proportions in each region:

```
summ.75 = gapminder.CLS.75 |>
    group_by(region) |>
    summarise(N=n(), p.hat=mean(life.75))

ggplot(data=summ.75,aes(x=p.hat, y=region, size=N, fill=p.hat)) +
    geom_point(col="black", alpha=.2,pch=22) +
    theme(legend.title = element_blank(), legend.position="none")
```



The proportion of countries with a life expectancy of more than 75 years is found to vary greatly from region to region – this may affect the quality of the estimate.

Next, we draw a SRS of $m - 8$ clusters:

```
set.seed(0) # for replicability
regions = unique(gapminder.CLS[,"region"])
M=length(regions)
m=8
(sample.reg = sample(1:M,m, replace=FALSE))
```

```
[1] 14  4  7  1  2 11 22 18
```

Then, we provide a summary of the proportions by cluster:

```
sample.ind = gapminder.CLS$region %in% regions[sample.reg]
gapminder.CLS.G = gapminder.CLS[sample.ind,]
gapminder.CLS.G$region <- as.factor(gapminder.CLS.G$region)
(summ.75.n = gapminder.CLS.G |>
    group_by(region) |>
    summarise(N=n(), p.hat=mean(life.75)))
```

```
# A tibble: 8 × 3
  region                        N p.hat
  <fct>                     <int> <dbl>
1 Australia and New Zealand     2 1
2 Caribbean                    13 0.385
3 Central America               8 0.5
4 Micronesia                    2 0
5 Northern Africa               6 0.333
6 Northern Europe              10 0.8
7 South-Eastern Asia           10 0.2
8 Southern Europe              12 1
```

We now have enough information to compute the CLS estimator of the proportion:

```
(p.G = sum(summ.75.n$N*summ.75.n$p.hat)/sum(summ.75.n$N))
```

```
[1] 0.5555556
```

Finally, we compute the sampling variance, the margin of error, and the 95% C.I. for p (assuming that the average cluster size is not known):

```
mean.size = sum(summ.75.n$N)/m
s2.p.G =  1/(m-1)*sum((summ.75.n$N*summ.75.n$p.hat-
            p.G*summ.75.n$N)^2)
V = 1/mean.size^2*s2.p.G/m*(1-m/M)
(B = 2*sqrt(V))
c(p.G-B,p.G+B)
```

```
[1] 0.2025966
[1] 0.3529590 0.7581521
```

The actual value $p = 0.394$ is indeed within the 95% confidence interval. We assumed that the average cluster size was unknown; is this also the case if we use the known value $\overline{N} = \frac{185}{22} \approx 8.41$?

The observations of the Gapmider dataset are probably not that suitable for CLS ... at least, not if we use regions as clusters.

11.6.2 Sample Size

Depending on whether the clusters are of equal size or not, the variance formulas take different forms; however, they coincide when $N_i = n$ for all i; it is only the **nature of the estimator bias** and the **exactness of its sampling variance** that are affected.

Consequently, we will only study the situation where the clusters are assumed to be of different sizes. In what follows, we will use the notations

$$\sigma_E^2 = \frac{1}{M}\sum_{\ell=1}^{M}(\tau_\ell - \mu N_\ell)^2 \quad \text{and} \quad s_E^2 = \frac{1}{m-1}\sum_{k=1}^{m}(y_{i_k} - \overline{y}_C N_{i_k})^2.$$

Mean μ

If we want to estimate μ with $\overline{y}_C$, we use:

$$\begin{aligned}
B_{\mu;C} = 2\sqrt{\frac{1}{\overline{N}^2}\cdot\frac{\sigma_E^2}{m}\left(\frac{M-m}{M-1}\right)} &\Longleftrightarrow \underbrace{\frac{B_{\mu;C}^2\overline{N}^2}{4}}_{=D_\mu} = \frac{\sigma_E^2}{m}\left(\frac{M-m}{M-1}\right)\\
&\Longleftrightarrow \frac{(M-1)D_\mu}{\sigma_E^2} = \frac{M-m}{m} = \frac{M}{m} - 1\\
&\Longleftrightarrow \frac{(M-1)D_\mu + \sigma_E^2}{\sigma_E^2} = \frac{M}{m}\\
&\Longleftrightarrow m_{\mu;C} = \frac{M\sigma_E^2}{(M-1)D_\mu + \sigma_E^2}.
\end{aligned}$$

Obviously, we can only use this formula **if we know the variance** σ_E^2 of the cluster totals in the population $\mathcal{U}$. If that is not available, we can use the **empirical variance** s_E^2 from a **preliminary sample**, or that from **a prior survey**.[52]

Finally, note that this formula allows us to determine the **number of clusters** m to be drawn from a SRS of clusters in order to obtain some margin of error on the estimate; the sample size may change from one realization to another, depending on the size of the sampled clusters.

Example Consider a company that wants a cost inventory for the \$N=\$625 items in stock. In practice, it might be tedious to obtain a SRS of these items; however, the items are arranged on $M = 100$ shelves and it is relatively easy to select a SRS of shelves, treating each shelf as a cluster of items. How many shelves would need to be sampled in order to estimate the average value of all items in inventory with a bound on the error of estimation of at most $B_{\mu;C} = 1.25\$$, assuming $\sigma_E^2 \approx 317.53\$$?

Set $D_\mu = \frac{B_{\mu;C}^2\overline{N}^2}{4} = \frac{(1.25)^2(6.25)^2}{4} \approx 15.26$; then

$$m_{\mu;C} = \frac{M\sigma_E^2}{(M-1)D_\mu + \sigma_E^2} = \frac{100(317.53)}{(100-1)(15.26)+317.53} = 17.4 \approx 18. \quad \blacksquare$$

[52] If the average size $\overline{N}$ of the clusters of $\mathcal{U}$ is unknown, we use the **empirical average size** $\overline{n} = (N_{i_1} + \cdots + N_{i_m})/m$ from the preliminary sample.

Total τ

If we want to estimate τ with $N\overline{y}_C$, we use:

$$B_{\tau;C} = 2\sqrt{M^2 \cdot \frac{\sigma_E^2}{m}\left(\frac{M-m}{M-1}\right)} \Longleftrightarrow \underbrace{\frac{B_{\tau;C}^2}{4M^2}}_{=D_{\tau;C}} = \frac{\sigma_E^2}{m}\left(\frac{M-m}{M-1}\right)$$

$$\Longleftrightarrow \frac{(M-1)D_{\tau;C}}{\sigma_E^2} = \frac{M-m}{m} = \frac{M}{m} - 1$$

$$\Longleftrightarrow \frac{(M-1)D_{\tau;C} + \sigma_E^2}{\sigma_E^2} = \frac{M}{m}$$

$$\Longleftrightarrow m_{\tau;C} = \frac{M\sigma_E^2}{(M-1)D_{\tau;C} + \sigma_E^2}.$$

Example Consider a company that wants a cost inventory for the $N=$625 items in stock. In practice, it might be tedious to obtain a SRS of these items; however, the items are arranged on $M = 100$ shelves and it is relatively easy to select a SRS of shelves, treating each shelf as a cluster of items. How many shelves would need to be sampled in order to estimate the total value of all items in inventory with a bound on the error of estimation of at most $B_{\tau;C} = 600\$$, assuming $\sigma_E^2 \approx 317.53\$$?

Set $D_{\tau;C} = \frac{B_{\tau;C}^2}{4M^2} = \frac{(600)^2}{4(100)^2} = 9$; then

$$m_{\tau;C} = \frac{M\sigma_E^2}{(M-1)D_{\tau;C} + \sigma_E^2} = \frac{100(317.53)}{(100-1)(9) + 317.53} = 26.3 \approx 27. \quad \blacksquare$$

If we want to estimate τ with $M\overline{y}_T$, we use:

$$B_{\tau;T} = 2\sqrt{M^2 \cdot \frac{\sigma_T^2}{m}\left(\frac{M-m}{M-1}\right)} \Longleftrightarrow \underbrace{\frac{B_{\tau;T}^2}{4M^2}}_{=D_{\tau}} = \frac{\sigma_T^2}{m}\left(\frac{M-m}{M-1}\right)$$

$$\Longleftrightarrow \frac{(M-1)D_{\tau}}{\sigma_T^2} = \frac{M-m}{m} = \frac{M}{m} - 1$$

$$\Longleftrightarrow \frac{(M-1)D_{\tau} + \sigma_T^2}{\sigma_T^2} = \frac{M}{m}$$

$$\Longleftrightarrow m_{\tau;T} = \frac{M\sigma_T^2}{(M-1)D_{\tau} + \sigma_T^2}.$$

Example Consider a company that wants a cost inventory for the $N=$625 items in stock. In practice, it might be tedious to obtain a SRS of these items; however, the items are arranged on $M = 100$ shelves and it is relatively easy to select a SRS of shelves, treating each shelf as a cluster of items. How many shelves would need to be sampled in order to estimate the total value of all items in inventory with a bound on the error of estimation of at most $B_{\tau;T} = 600\$$, assuming $\sigma_T^2 \approx 682.77\$$?

Set $D_{\tau;T} = \frac{B^2_{\tau;T}}{4M^2} = \frac{(600)^2}{4(100)^2} = 9$; then

$$m_{\tau;T} = \frac{M\sigma^2_T}{(M-1)D_{\tau;T} + \sigma^2_T} = \frac{100(682.77)}{(100-1)(9) + 682.77} = 43.4 \approx 44. \quad \blacksquare$$

Proportion p

If we want to estimate p with $\hat{p}_C$, we use:

$$\begin{aligned}
B_{p;C} = 2\sqrt{\frac{1}{\overline{N}^2} \cdot \frac{\sigma^2_p}{m}\left(\frac{M-m}{M-1}\right)} &\iff \underbrace{\frac{B^2_{p;C}\overline{N}^2}{4}}_{=D_{p;C}} = \frac{\sigma^2_p}{m}\left(\frac{M-m}{M-1}\right) \\
&\iff \frac{(M-1)D_{p;C}}{\sigma^2_P} = \frac{M-m}{m} = \frac{M}{m} - 1 \\
&\iff \frac{(M-1)D_{p;C} + \sigma^2_P}{\sigma^2_P} = \frac{M}{m} \\
&\iff m_{p;C} = \frac{M\sigma^2_P}{(M-1)D_{p;C} + \sigma^2_P}.
\end{aligned}$$

11.6.3 Comparison Between SRS and CLS

Consider a ClS $\mathcal{Y}$ consisting of m clusters drawn from a population $\mathcal{U}$ of size N, distributed in M clusters. Let μ be the mean and σ^2 the variance of the population $\mathcal{U}$.

If the clusters are all of size n, we can show that

$$\mathrm{V}(\overline{y}_C) \approx \frac{\sigma^2 - \overline{\sigma^2}}{m}\left(1 - \frac{m}{M}\right), \quad \text{where} \quad \overline{\sigma^2} = \frac{1}{M}\sum_{\ell=1}^{M}\sigma^2_\ell,$$

where σ^2_ℓ is the variance in the ℓth cluster.

But we can also consider $\mathcal{Y}$ as having arisen from a SRS with size mn. In that case, we have

$$\mathrm{V}(\overline{y}_{\mathrm{SRS}}) = \frac{\sigma^2}{mn}\left(\frac{N-mn}{N-1}\right) \approx \frac{\sigma^2}{mn}\left(1 - \frac{mn}{N}\right) = \frac{\sigma^2}{mn}\left(1 - \frac{mn}{Mn}\right) = \frac{\sigma^2}{mn}\left(1 - \frac{m}{M}\right),$$

from which we conclude that

$$\begin{aligned}
\mathrm{V}(\overline{y}_C) - \mathrm{V}(\overline{y}_{\mathrm{SRS}}) &\approx \frac{1}{m}\left(1 - \frac{m}{M}\right)\left(\sigma^2 - \overline{\sigma^2} - \frac{\sigma^2}{n}\right) = \frac{1}{m}\left(1 - \frac{m}{M}\right)\left(\frac{n-1}{n}\sigma^2 - \overline{\sigma^2}\right) \\
&\approx \frac{1}{m}\left(1 - \frac{m}{M}\right)(\sigma^2 - \overline{\sigma^2}), \quad \text{si } n - 1 \approx n.
\end{aligned}$$

Consequently, $\mathrm{V}(\overline{y}_C) \gg \mathrm{V}(\overline{y}_{\mathrm{SRS}})$ if and only if $\sigma^2 \gg \overline{\sigma^2}$, which is the case when the **mean of the cluster variances is smaller than the variance in the population**.

The moral of the story is that a ClS is effective if the clusters, regardless of their size, are **as heterogeneous as the population itself**.

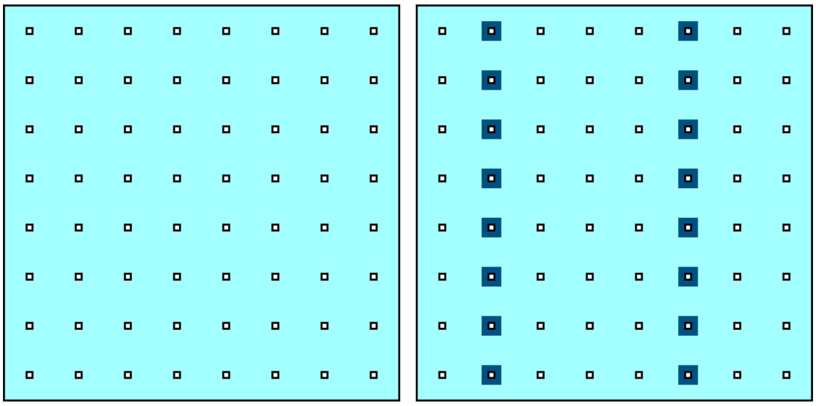


Figure 11.11: Schematics of SYS: target population (left) and sample (right).

11.7 Special Topics

We complete this introduction to survey sampling by discussing a few additional topics.[53]

11.7.1 Systematic Sampling

With the advent of easy-to-access pseudo-random number generators,[54] it is not very arduous to draw a pseudo SRS $\mathcal{Y}$ of size n from a population $\mathcal{U}$ of size N (assuming that we have an appropriate sampling frame, of course).

However, it remains possible for the obtained sample to **not be representative** of the population: a SRS of countries that do not include China or India, for example, would not be very useful if we are trying to estimate the average population of the world's countries.

In some cases, a **systematic sampling design** (SYS) can be used to maximize the probability that the random sample $\mathcal{Y}$ represents the population.

Here is how we draw a 1−in−M systematic sample of size n (or $n+1$) from an ordered list of size N:

1. determine the integer part $M = \lfloor \frac{N}{n} \rfloor$;
2. randomly select an integer γ in $\{1, 2, \ldots, M\}$;
3. the sample $\mathcal{Y}$ then contains the values corresponding to units

$$\underbrace{\gamma, \gamma + M,\ \gamma + 2M,\ \ldots\ , \gamma + (n-1)M}_{n\text{ units}},\ \underbrace{\gamma + nM}_{\text{if } \gamma + nM \leq N}.$$

If the ordering of the units in the sampling frame is fixed, there can only be M different SYS samples of size n (or $n+1$, in some cases).

Example The Gapminder dataset contains socio-economic information on $N = 185$ countries in 2011. What are the average life expectancy and population of the world's countries?

[53] A few of which could even be called advanced.

[54] Excel, R, SAS, `Python`, etc.

We modify the code allowing us to access the data set:

```
gapminder.SYS <- gapminder |> filter(year==2011) |>
    select(country, life_expectancy, population)
N=nrow(gapminder.SYS)
```

There are 185 units in the data set. If we are interested in a SYS of size $n = 20$, say, the integer M is:

```
n=20
(M=floor(N/n))
```

```
[1] 9
```

The vector of observations $0, M, 2M, \ldots, nM$ is therefore:

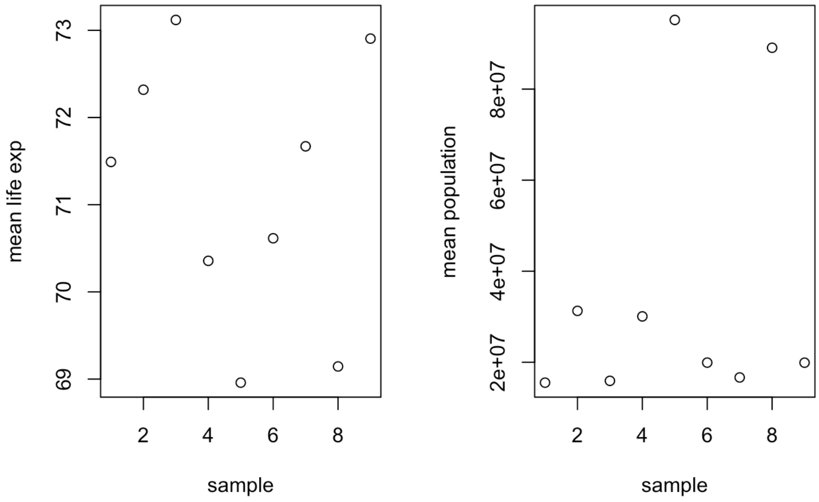
```
index = M*(0:n)
```

We construct $M = 9$ samples $\mathcal{Y}_i$, $i = 1, \ldots, 9$, assuming that the units appear in alphabetical order (by country name) in the dataset.

```
moy.SYS.life_exp = c() # initialization - life expectation
moy.SYS.pop = c()      # initialization - population

for(j in 1:M){# all SYS of size n or n+1, alpha order
    index.tmp = j + index
    index.tmp <- index.tmp[index.tmp < N+1]  # keeping indices <= N
    sample.sys = gapminder.SYS[index.tmp,2:3]
    moy.SYS.life_exp[j]=mean(sample.sys$life_expectancy)
    moy.SYS.pop[j]=mean(sample.sys$population)
  }

# charts
par(mfrow=c(1,2))
plot(moy.SYS.life_exp, xlab="sample", ylab="mean life exp")
plot(moy.SYS.pop, xlab="sample", ylab="mean population")
```

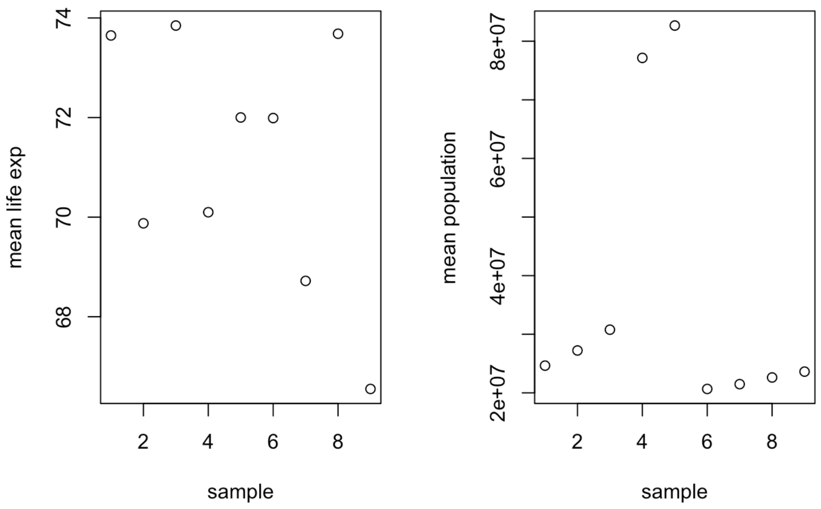


Could you identify the sample that contains China or India? What if we change the order in which the countries are listed in the dataset?

```
gapminder.SYS <- gapminder.SYS[order(gapminder.SYS$population),]

for(j in 1:M){# all SYS of size n or n+1, population order
    index.tmp = j + index
    index.tmp <- index.tmp[index.tmp < N+1]
    sample.sys = gapminder.SYS[index.tmp,2:3]
    moy.SYS.life_exp[j]=mean(sample.sys$life_expectancy)
    moy.SYS.pop[j]=mean(sample.sys$population)
  }

par(mfrow=c(1,2))
plot(moy.SYS.life_exp, xlab="sample", ylab="mean life exp")
plot(moy.SYS.pop, xlab="sample", ylab="mean population")
```



We obtain similar results when ordering the units in the dataset by life expectancy. ■

In general, if there is a correlation between the **position (rank) of the unit** in the sampling frame and the **value of the variable of interest**, the sampling variance of the SYS estimator will be **lower** than that of the SRS estimator, because the sample is **more likely** to be representative of the population.

If there is no such correlation, the SYS sample is essentially an SRS sample, and the sampling variances are comparable – a SYS is as likely to be **representative** of the population as an SRS.

Finally, if the step M is aligned with the periodicity of the values of the variable of interest, it is the opposite: the sampling variance of a SYS is larger than that of an SRS – a SYS is then **less representative** of the population than an SRS.

Some examples illustrating these situations are shown in Figure 11.12.

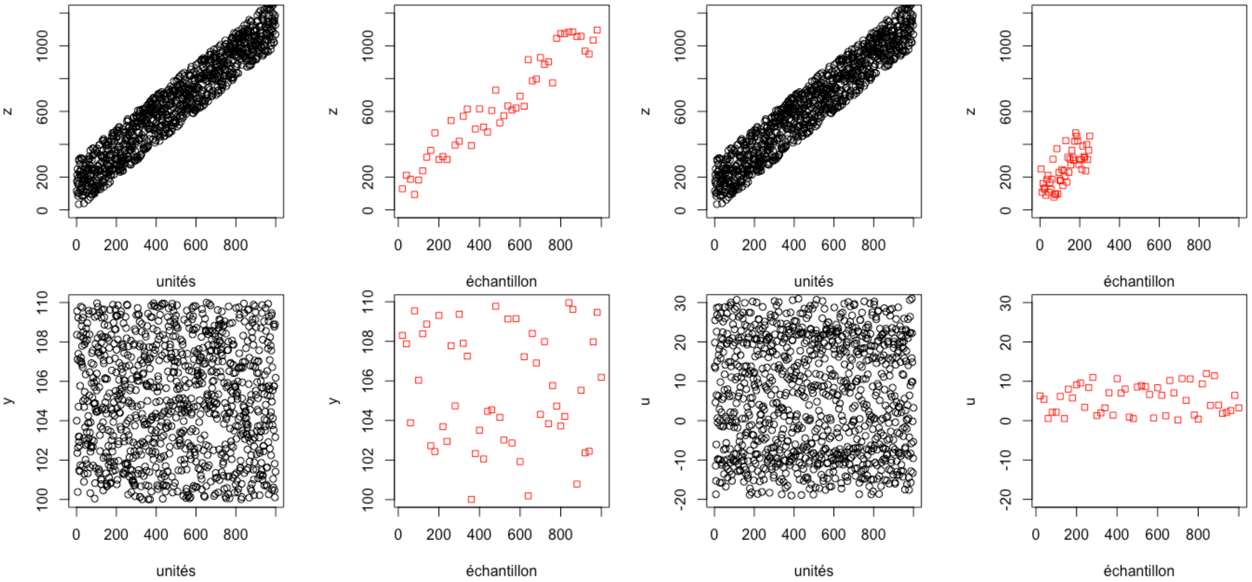


Figure 11.12: Various populations and systematic samplings: the order in which the population observations are presented may affect the representativity of the SYS sample.

SYS as SRS

If the order in which the units are listed in the sampling frame is **random**,[55] we can simply consider that the sample

$$\mathcal{Y}_{\text{SYS}} = \underbrace{\{y_1, y_2, y_3, \ldots, y_{n-1}, y_n\}}_{\{u_\gamma, u_{\gamma+M}, \ldots, u_{\gamma+(n-1)M}\}} \subseteq \mathcal{U}$$

of size $n \approx \frac{N}{M}$ is in fact a SRS of size n. In that case, the theory developed in Section 11.3 for SRS remains valid.

Estimating the Mean μ The empirical mean

$$\overline{y}_{\text{SYS}} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

is an **unbiased** estimator of the true population mean μ, with **bound on the error of estimation**

$$B_{\mu;\text{SYS}} \approx \hat{B}_{\mu;\text{SYS}} = 2\sqrt{\hat{\mathrm{V}}(\overline{y}_{\text{SYS}})} = 2\sqrt{\frac{s^2_{\text{SYS}}}{n}\left(1 - \frac{n}{N}\right)},$$

where

$$s^2_{\text{SYS}} = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y}_{\text{SYS}})^2;$$

the corresponding **95% C.I. for** μ is thus

$$\text{C.I.}_{\text{SYS}}(\mu; 0.95): \quad \overline{y}_{\text{SYS}} \pm \hat{B}_{\mu;\text{SYS}}.$$

[55] **Careful!** this is not always easy to demonstrate.

Estimating the Total τ The quantity

$$\hat{\tau}_{\text{SYS}} = N\overline{y}_{\text{SYS}} = \frac{N}{n}\sum_{i=1}^{n} y_i$$

is an **unbiased** estimator of the true population total τ, with **bound on the error of estimation**

$$B_{\tau;\text{SYS}} \approx \hat{B}_{\tau;\text{SYS}} = 2N\sqrt{\hat{\text{V}}(\overline{y}_{\text{SYS}})} = 2N\sqrt{\frac{s^2_{\text{SYS}}}{n}\left(1-\frac{n}{N}\right)};$$

the corresponding **95% C.I. for** τ is thus

$$\text{C.I.}_{\text{SYS}}(\tau;0.95): \quad \hat{\tau}_{\text{SYS}} \pm \hat{B}_{\tau;\text{SYS}}.$$

Estimating the Proportion p If $y_i \in \{0,1\}$ denotes the absence or presence of a certain characteristic, the quantity

$$\hat{p}_{\text{SYS}} = \overline{y}_{\text{SYS}}$$

is an **unbiased** estimator of the true proportion p of units with the characteristic, with **bound on the error of estimation**

$$B_{p;\text{SYS}} \approx \hat{B}_{p;\text{SYS}} = 2\sqrt{\hat{\text{V}}(\hat{p}_{\text{SYS}})} = 2\sqrt{\frac{\hat{p}_{\text{SYS}}(1-\hat{p}_{\text{SYS}})}{n-1}\left(1-\frac{n}{N}\right)};$$

the corresponding **95% C.I. for** p is thus

$$\text{C.I.}_{\text{SYS}}(p;0.95): \quad \hat{p}_{\text{SYS}} \pm \hat{B}_{p;\text{SYS}}.$$

SYS as CLS

In practice, SYS is equivalent to a CLS of size $m = 1$, where each cluster is one of the 1−in−M SYS samples.

The quantity

$$\overline{y}_C = \frac{\displaystyle\sum_{k=1}^{m}\sum_{j=1}^{N_{i_k}} y_{i_k,j}}{\displaystyle\sum_{k=1}^{m} N_{i_k}} = \frac{\displaystyle\sum_{k=1}^{m} y_{i_k}}{\displaystyle\sum_{k=1}^{m} N_{i_k}},$$

where we use the CLS notation, is thus a **biased** estimator of the **population mean,** μ.

The **average cluster size** is denoted by $\overline{N} = \frac{N}{M}$; its **sampling variance** is

$$\text{V}(\overline{y}_C) \approx \frac{1}{\overline{N}^2}\cdot\frac{1}{m}\left(\frac{M-m}{M-1}\right)\cdot\frac{1}{M}\sum_{\ell=1}^{M}\underbrace{(\tau_\ell-\mu N_\ell)^2}_{=N_\ell(\mu_\ell-\mu)} := \frac{1}{\overline{N}^2}\cdot\frac{\sigma_C^2}{m}\left(\frac{M-m}{M-1}\right),$$

and the corresponding **95% C.I. for** μ is thus

$$\text{C.I.}_{\text{G}}(\mu;0.95): \quad \overline{y}_C \pm 2\sqrt{\text{V}(\overline{y}_C)}.$$

If the average cluster size $\overline{N}$ is unknown, we simply substitute it by

$$\overline{n} = \frac{1}{n}\sum_{k=1}^{m} N_{i_k}.$$

The estimator of the **total population** τ is thus either:

- $N\overline{y}_C$, when the number of units N in the population is known, or
- $M\overline{y}_T$, where $\overline{y}_T$ is the **(empirical) mean of the sampled cluster totals**, when only M is known.

Consequently, the sampling variances are

$$\mathrm{V}(N\overline{y}_C) \approx M^2 \cdot \frac{\sigma_C^2}{m}\left(\frac{M-m}{M-1}\right) \quad \text{and} \quad \mathrm{V}(M\overline{y}_T) \approx M^2 \cdot \frac{\sigma_T^2}{m}\left(\frac{M-m}{M-1}\right),$$

where σ_C^2 and σ_T^2 are computed as for a CLS. We can then construct the **95% C.I. for** τ in the usual manner.

Pretty simple, eh?

The sample contains exactly $m = 1$ cluster, so $\overline{n} = n$. The problem doesn't end there – since we don't know σ_C^2 or σ_T^2 in general, we would use the empirical variances

$$\hat{\mathrm{V}}(\overline{y}_C) \approx \frac{1}{\overline{N}^2} \cdot \frac{1}{m}\left(1-\frac{m}{M}\right) \cdot \frac{1}{m-1}\sum_{k=1}^{m}(y_{i_k} - \overline{y}_C N_{i_k})^2$$

$$\hat{\mathrm{V}}(M\overline{y}_T) \approx M^2 \cdot \frac{1}{m}\left(1-\frac{m}{M}\right) \cdot \frac{1}{m-1}\sum_{k=1}^{m}(y_{i_k} - \overline{y}_T)^2.$$

But if $m = 1$, these variances do not exist. How do we get out of this mess? If we cannot treat the SYS as if it were a SRS (for whatever reason), the solution is to **draw additional SYS samples (replicates) and treat it as a CLS**, modifying the value of M as necessary.

11.7.2 Sampling with Probability Proportional to Size

In practice, the **size** (whether or not this is a physical characteristic) of the sample units is often quite **variable** – a SRS is not always effective since it does not take into account the **importance that larger population units** may have.

Additional information on the unit size can sometimes be used to select a sample that provides a more accurate estimator of the parameters of interest.

One possible way to do this is to assign (potentially) **equal** selection probabilities to different units, based on their size.

Example To a certain extent ($\rho = 0.46$), the larger the area of a country, the larger its population. If we are trying to estimate the population of the planet, it might be desirable to adopt a sampling scheme in which the probability of selecting a country is **proportional to its area** – in an SRS, it is very likely that neither **China** nor **India** will be selected, resulting in an underestimate of the total sought. ■

If the variable of interest is (more or less) related to the size of the unit, one can assign a **probability of selection proportional to the size** of the unit (PPS). Note that in a PPS, previously selected units are **replaced** in the population, allowing for the **multiple selection of a single unit**.

Selecting a PPS With Replacement

We consider two selection methods for a PPS sample:

- **cumulative totals**, and
- the **Lahiri method**.

In both cases, the PPS sample selection procedure consists of associating with each unit a **range of numbers**,[56] related to the **size of the unit**, and taking the units that correspond to numbers chosen **at random** from the set of numbers associated with the **entire** population of N units.

In the **method of cumulative totals**, the **size** of the $i-$th unit is denoted by x_i, $1 \leq i \leq N$. We associate a **range** to each unit as follows:

Unit		**Range**	
1	1	to	x_1
2	$x_1 + 1$	to	$x_1 + x_2$
3	$x_1 + x_2 + 1$	to	$x_1 + x_2 + x_3$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$N-1$	$x_1 + \cdots + x_{N-2} + 1$	to	$x_1 + \cdots + x_{N-2} + x_{N-1}$
N	$x_1 + \cdots + x_{N-1} + 1$	to	$x_1 + \cdots + x_{N-1} + x_N$

Finally, we draw a PPS sample by choosing n integers **at random** between 1 and $X = x_1 + \cdots + x_{N-1} + x_N$ (**with replacement**) and by selecting the units **associated with these integers**.

[56] These are often **integers**, but that is not necessary.

Example In a village, there are 8 orchards, each containing a certain number of apple trees. A sample of $n = 3$ orchards is drawn (with replacement), in proportion to the number of apple trees per orchard.

ID i	Size x_i	Cumulative Totals	Associated Range
1	50	50	1 – 50
2	30	80	51 – 80
3	25	105	81 – 105
4	40	145	106 – 145
5	26	171	146 – 171
6	44	215	172 – 215
7	20	235	216 – 235
8	35	270	236 – 270

We choose $n = 3$ integers at random between 1 and 270: 108, 140, and 201, say. The associated units are the 4th, the 4th, and the 6th. ■

In the **Lahiri method**, we still denote the size of a unit by x_i, $1 \leq i \leq N$, but without having **to calculate and report the successive cumulative totals**, which can be tedious to accomplish, even with a computer.

The method consists in selecting a pair of integers (i, j), where $1 \leq i \leq N$ and $1 \leq j \leq M = \max\{x_i \mid 1 \leq i \leq N\}$. If $j \leq x_i$, the ith unit is added to the sample. Otherwise, the pair (i, j) is rejected.

We continue in this manner until n units have been selected.[57]

Estimation

Let us revisit the orchard example, where u_i is the yield of all apple trees in the ith orchard.

ID i	# Trees x_i	π_i	Yield
1	50	50/270	$u_1 = 2250$
2	30	30/270	$u_2 = 1080$
3	25	25/270	$u_3 = 1300$
4	40	40/270	$u_4 = 1400$
5	26	26/270	$u_5 = 1196$
6	44	44/270	$u_6 = 1716$
7	20	20/270	$u_7 = 820$
8	35	35/270	$u_8 = 1680$

We are interested in the **total** apple production of the village, which we know in this case to be $\tau = 11,442$. Since **in principle** an orchard with more apple trees should produce more apples, we draw a PPS sample of $n = 3$ units (with replacement), where the number of apple trees in the orchard is used as the unit size.

In what follows, we illustrate the concepts using the sample

$$y_1 = u_4 = 1400, y_2 = u_4 = 1400, y_3 = u_6 = 1716.$$

[57] There are other ways to do this, of course; the important thing is to have a **mechanism for selecting a PPS sample**. We generally prefer sampling without replacement to sampling with replacement, but the latter is a reasonable substitute to the former if $\frac{n}{N}$ is " **sufficiently small**".

If the sample $\mathcal{Y}$, with $|\mathcal{Y}| = n$, is drawn from $\mathcal{U}$ using a PPS, the units $y_1, \ldots, y_n$ are **independent** and distributed according to

y_i	u_1	$\cdots$	u_j	$\cdots$	u_N
$p(y_i)$	π_1	$\cdots$	π_j	$\cdots$	π_N

where $0 < \pi_j < 1$ for all $1 \leq j \leq N$ and $\pi_1 + \cdots + \pi_N = 1$.

For all $1 \leq i \leq n$, there is a $1 \leq j \leq N$ such that $y_i = u_j$. Set $w_i = \frac{u_j}{\pi_j}$. The **sampling weights** w_i are also **independent** and distributed according to

$$P(y_i = u_j) = P\left(w_i = \frac{u_j}{\pi_j}\right) = \pi_j, \quad 1 \leq i \leq n,\ 1 \leq j \leq N.$$

We note that for any $1 \leq i \leq n$, the **expected weight** is

$$\mathrm{E}(w_i) = \sum_{j=1}^{N} w_j P(w_i = w_j) = \sum_{j=1}^{N} \frac{u_j}{\pi_j} \cdot \pi_j = \sum_{j=1}^{N} u_j = \tau.$$

In other words,

$$\hat{\tau}_{\text{pps}} = \overline{w} = \frac{1}{n}\sum_{i=1}^{n} w_i$$

is an **unbiased estimator of the total** τ. Its **sampling variance** is computed as follows:

$$\mathrm{V}(\hat{\tau}_{\text{pps}}) = \mathrm{V}\left(\frac{1}{n}\sum_{i=1}^{n} w_i\right) = \underbrace{\frac{1}{n^2}\sum_{i=1}^{n}\mathrm{V}(w_i)}_{\text{ind. des } w_i} = \frac{1}{n^2}\sum_{i=1}^{n}\left[\sum_{j=1}^{N}(w_j - \tau)^2 P(w_i = w_j)\right]$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{N}\left(\frac{u_j}{\pi_j} - \tau\right)^2 \pi_j = \frac{1}{n}\sum_{j=1}^{N}\left(\frac{u_j}{\pi_j} - \tau\right)^2 \pi_j = \frac{1}{n}\sum_{j=1}^{N}\left(\frac{u_j^2}{\pi^j} - \frac{2\tau u_j}{\pi_j} + \tau^2\right)\pi_j$$

$$= \frac{1}{n}\left(\sum_{j=1}^{N}\frac{u_j^2}{\pi_j} - 2\tau \underbrace{\sum_{j=1}^{N} u_j}_{=\tau} + \tau^2 \underbrace{\sum_{j=1}^{N}\pi_j}_{=1}\right) = \frac{1}{n}\left(\sum_{j=1}^{N}\frac{u_j^2}{\pi_j} - \tau^2\right).$$

In practice, we do not typically know the true value of τ, so we use the **unbiased estimator**

$$\hat{\mathrm{V}}(\hat{\tau}_{\text{pps}}) = \frac{1}{n(n-1)}\left(\sum_{i=1}^{n} w_i^2 - n\hat{\tau}_{\text{pps}}^2\right).$$

Central Limit Theorem – PPS: if n and $N - n$ are sufficiently large, then

$$\hat{\tau}_{\text{pps}} \sim_{\text{approx.}} \mathcal{N}\left(\tau, \hat{\mathrm{V}}(\hat{\tau}_{\text{pps}})\right).$$

The **bound on the error of estimation** and the **95% C.I. for** τ are therefore

$$\hat{B}_{\tau;\text{pps}} = 2\sqrt{\hat{\mathrm{V}}(\hat{\tau}_{\text{pps}})} \quad \text{and} \quad \text{C.I.}_{\text{pps}}(\tau; 0.95) = \hat{\tau}_{\text{pps}} \pm \hat{B}_{\tau;\text{pps}}.$$

Example In the orchard dataset, we have

$$\hat{\tau}_{\text{pps}} = \frac{1}{3}\Bigg[\underbrace{\frac{1400}{40/270}}_{w_1} + \underbrace{\frac{1400}{40/270}}_{w_2} + \underbrace{\frac{1716}{44/270}}_{w_3}\Bigg] = 9810;$$

$$\hat{\text{V}}(\hat{\tau}_{\text{pps}}) = \frac{1}{3(2)}\Bigg[\Bigg(\underbrace{\frac{1400}{40/270}}_{w_1}\Bigg)^2 + \Bigg(\underbrace{\frac{1400}{40/270}}_{w_2}\Bigg)^2 + \Bigg(\underbrace{\frac{1716}{44/270}}_{w_3}\Bigg)^2 - 3\cdot \underbrace{9810^2}_{\hat{\tau}^2_{\text{pps}}}\Bigg]$$

$$= 129,600.$$

Consequently, the 95% C.I. for the total apple yield in the village is

$$\text{C.I.}_{\text{pps}}(\tau; 0.95) = 9810 \pm 2\sqrt{129,600} \equiv (9090, 10530).$$

The actual total yield ($\tau = 11,442$) **does not fall** within the confidence interval – why might this be the case? Is this problematic? ■

In general, $\text{V}(\hat{\tau}_{\text{pps}}) \leq \text{V}(\hat{\tau}_{\text{SRS}})$. In the orchards example, we can show that

$$\text{V}(\hat{\tau}_{\text{SRS}}) \approx 8^2 \cdot \frac{172981.4375}{3}\left(\frac{8-3}{8-1}\right) = 2,635,907.619, \quad \text{and}$$

$$\text{V}(\hat{\tau}_{\text{pps}}) \approx \frac{1}{3}\left[\frac{2250^2}{50/270} + \cdots + \frac{1680^2}{35/270} - 11,442^2\right] = 723,912.$$

We can also give an estimate of the population average μ using

$$\hat{\mu}_{\text{pps}} = \frac{\hat{\tau}_{\text{pps}}}{N}, \quad \hat{\text{V}}(\hat{\mu}_{\text{pps}}) = \frac{\hat{\text{V}}(\hat{\tau}_{\text{pps}})}{N^2}, \quad \text{C.I.}_{\text{pps}}(\mu; 0.95) = \frac{\text{C.I.}_{\text{pps}}(\tau; 0.95)}{N}.$$

A lot more can be said on the topic; PPS usually provides a springboard to more sophisticated sampling designs and other theoretical considerations [62–64].

11.7.3 Multi-Stage Sampling

By splitting the sampling process into several stages, one can **reduce costs** and **focus the logistical aspects of sampling on a few focal points**. In **multi-stage sampling** (M*n*S), a sample of large units (**primary units**) is drawn, then sub-units (**secondary units**) are drawn from the large units, and so on.

Example Sampling units in a Canadian province could be decomposed into three steps:

1. conduct a sample of municipalities (**primary units**);
2. sample neighbourhoods in the sampled municipalities (**secondary units**), and
3. sample households in the samples neighbourhoods (**tertiary units**).

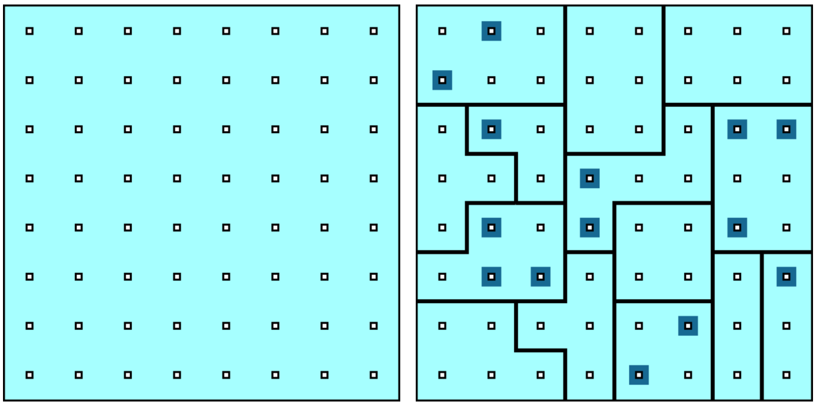


Figure 11.13: Schematics of SRS2S: target population (left) and sample (right).

In a M*n*S, the sample is concentrated around several **pivots**: in field studies, for example, this has the advantage of considerably reducing the survey area, which helps to **reduce non-sampling errors**.[58]

Furthermore, detailed information is often available for **groups** of sample units, but not for **individual** units: it is therefore not necessary to obtain a **complete** sampling frame for **all** sample units, but only for those belonging to the primary units selected in the first round, for example.

Any probability sampling method can be used at **each stage**, and they **can change from stage to stage**: e.g., a municipality SRS, a neighborhood SRS, a household SRS, etc.

Two-Stage Simple Random Sampling

In a 2-stage process, if sampling is conducted using a SRS for both stages, the method is known as **two-stage simple random sampling** (SRS2S).

Example The biomass of a plant species in a forest area can be estimated by drawing a SRS of $m = 8$ compartments (primary units) from the $M = 40$ compartments composing the population under study.

For each of these compartments $1 \leq i \leq m$, we then draw a SRS of n_i plots, and measure the biomass in the plot. Estimates of the average or total amount of biomass in the forest area can be calculated using appropriate formulas. ■

Estimation

Let be a population consisting of M primary units, having N_ℓ secondary units in the ℓth primary unit. Denote by $u_{i,j}$ the value of the response variable of the jth secondary unit in the ith primary unit.

The **population mean** is

$$\mu = \frac{\sum_{\ell=1}^{M} \sum_{j=1}^{N_\ell} u_{\ell,j}}{\sum_{\ell=1}^{M} N_\ell}.$$

[58] In addition to reducing operational costs.

Suppose we draw a SRS of m primary units, and a SRS of n_i secondary units in the ith primary unit. The total sample size is thus $n = n_1 + cdots + n_m$. We obtain an unbiased estimator of μ from:

$$\overline{y}_{\text{SRS2S}} = \frac{1}{m\overline{N}}\sum_{i=1}^{m} N_i\overline{y}_i = \frac{1}{m\overline{N}}\sum_{i=1}^{m}\frac{N_i}{n_i}\sum_{k=1}^{n_i} y_{i,k} = \frac{1}{m\overline{N}}\sum_{i=1}^{m}\sum_{k=1}^{n_i}\frac{MN_i}{mn_i}y_{i,k},$$

where

$$\overline{N} = \frac{1}{M}\sum_{\ell=1}^{M} N_\ell \approx \frac{N_1 + \cdots + N_m}{m}.$$

The sampling variance is composed of two components:

- a measure of the variation **between the primary units**, and
- a measure of the variation **within the primary units**.

When $n_i = N_i$ for all $1 \leq i \leq m$, we are dealing with a **CLS** and the variance is only given by the first component (see Section 11.6). In the case where $m = M$, we are dealing with a **STS** and the variance is only given by the second component (see Section 11.4).

When $m \neq M$ and $n_i \neq N_i$ for at least one primary unit i, the variance is a combination of these two extremes: in that case, the second component represents **the contribution of sub-sampling** (another name for MnS). We use the **law of total variance** to estimate the sampling variance:

$$\begin{aligned} \text{V}(\overline{y}_{\text{SRS2S}}) &= \text{E}[\text{V}(\overline{y}_{\text{SRS2S}} \mid m)] + \text{V}(\text{E}[\overline{y}_{\text{SRS2S}} \mid m]) \\ &= \frac{1}{\overline{N}^2}\cdot\frac{\sigma_T^2}{m}\left(\frac{M-m}{M-1}\right) + \frac{1}{mM\overline{N}^2}\sum_{i=1}^{m} N_i^2\cdot\frac{\sigma_i^2}{n_i}\left(\frac{N_i-n_i}{N_i-1}\right) \\ &\approx \underbrace{\frac{1}{\overline{N}^2}\cdot\frac{s_T^2}{m}\left(1-\frac{m}{M}\right)}_{\text{between primary units}} + \underbrace{\frac{1}{mM\overline{N}^2}\sum_{i=1}^{m} N_i^2\cdot\frac{s_i^2}{n_i}\left(1-\frac{n_i}{N_i}\right)}_{\text{within primary units}}, \end{aligned}$$

where

$$s_T^2 = \frac{1}{m-1}\sum_{i=1}^{n}\left(N_i\overline{y}_i - \overline{N}\overline{y}_{\text{SRS2S}}\right)^2, \quad s_i^2 = \frac{1}{n_i-1}\sum_{k=1}^{n_i}(y_{i,k}-\overline{y}_i)^2.$$

Example The biomass of a plant species (kg) is measured in plots of 0.025 ha (secondary units) selected from $m = 8$ compartments (primary units), randomly selected themselves among the $M = 40$ compartments of a forested area. The summary of results is shown in the following table:

Comp.	1	2	3	4	5	6	7	8
$\overline{y}_i$	118	107	109	110	120	95	93	90
s_i^2	436	516	586	456	412	497	755	496
N_i	1760	1975	1615	1785	1775	2050	1680	1865
n_i	9	10	8	9	9	10	8	9

Find a 95% C.I. for the average biomass per plot and per compartment, and for its total in the forested area.

Solution: Since we do not know $\overline{N}$, we approximate it with the mean

$$\overline{N} \approx \frac{1}{8}(1760 + \cdots + 1865) = 1813.125.$$

The totals in the selected primary units are then:

Comp.	1	2	3	4	5	6	7	8
$N_i\overline{y}_i(\times 10^5)$	2.077	2.113	1.760	1.964	2.130	1.946	1.562	1.679

The SRS2S estimators of the mean μ, of the mean of the totals in the compartments, and of the total are:

$$\overline{y}_{\text{SRS2S}} = \frac{1}{8(1813.125)}(2.077 + \cdots + 1.679) \times 10^5 = 105.01;$$

$$\overline{N}\overline{y}_{\text{SRS2S}} = 1813.125 \cdot 105.01 = 190,403.75; \quad \tau_{\text{SRS2S}} = M \cdot \overline{N}\overline{y}_{\text{SRS2S}} = 7,616,150.$$

The variance between compartments (primary units) is thus:

$$s_T^2 = \frac{1}{8-1}\sum_{i=1}^{8}(N_i\overline{y}_i - 190,403.75)^2 = 4.55 \times 10^8$$

Finally, we calculate the variance within the compartments:

Comp.	1	2	3	4	5	6	7	8
$\frac{N_i^2}{\overline{N}^2} \cdot \frac{s_i^2}{n_i}\left(1 - \frac{n_i}{N_i}\right)$	48.2	51.3	72.7	50.4	45.6	49.4	93.9	54.9

The sampling variance is thus

$$\begin{aligned}\hat{\text{V}}(\overline{y}_{\text{SRS2S}}) &= \frac{4.55 \times 10^8}{8(1813.125)^2}\left(1 - \frac{8}{40}\right) + \frac{1}{8(40)}(48.2 + \cdots + 54.9) \\ &= 14.03\end{aligned}$$

The variances of the other two estimators are easily calculated:

$$\begin{aligned}\hat{\text{V}}(\overline{N}\overline{y}_{\text{SRS2S}}) &= \overline{N}^2\hat{\text{V}}(\overline{y}_{\text{SRS2S}}) = (1813.125)^2 \cdot 14.03 = 46,141,324.55; \\ \hat{\text{V}}(\tau_{\text{SRS2S}}) &= M^2\overline{N}^2\hat{\text{V}}(\overline{y}_{\text{SRS2S}}) = (40)^2 \cdot (1813.125)^2 \cdot 14.03 = 73,826,119,284;\end{aligned}$$

the confidence intervals are thus

$$\begin{aligned}\text{C.I.}_{\text{SRS2S}}(\mu; 0.95): &\quad 105.01 \pm 2\sqrt{14.03} \equiv (97.5, 112.5) \\ \text{C.I.}_{\text{SRS2S}}(\tfrac{N_0}{M}\mu; 0.95): &\quad 190,403.75 \pm 2\sqrt{46,141,324.55} \equiv (176818, 203989.2312), \\ \text{C.I.}_{\text{SRS2S}}(\tau; 0.95): &\quad 7,616,150 \pm 2\sqrt{73,826,119,284} \equiv (7072730, 8159569)\end{aligned}$$

assuming of course that the central limit theorem remains valid in the context of a SRS2S. ■

11.7.4 Multi-Phase Sampling

Multi-stage sampling (MnP) plays a crucial role in many types of surveys, including those conducted by **remote sensing**.

In the first phase, a **selected** number of units are sampled, but only a **small** number of characteristics are captured for each unit. In each successive phase, a larger **number** of features is measured on a smaller **sub-sample** of units.

In this way, the target parameter can be estimated with **more accuracy** and at **lower cost**, by studying the relationship between the features measured in the different sampling phases.

Two-Phase Random Sampling

A MnP with only two phases is called a **two-phase sampling** (M2P). M2Ps are particularly useful in a situation where enumeration of the **main trait** is expensive (in terms of costs or labor), but in which an **auxiliary trait** correlated to the main trait can easily be observed.

Thus, it is sometimes preferable to draw a **large** SRS in the **first phase** in order to analyze the auxiliary variables, which leads to more accurate estimates of τ or μ for that auxiliary variable (at least, that is the hope). In the second phase, a **smaller** sample is drawn, usually a **sub-sample** of the characteristic, and the **auxiliary variable** are measured.

Estimates of the main characteristic are then obtained using the information obtained in the **first phase**, using the **ratio method** or the **regression method**, for instance. The precision of the final estimates can be increased by including **several correlated auxiliary variables**.

Example If we want to estimate the total volume of wood τ in a forest, we could first measure the circumference c_i and height h_i of the trees i in some sample, then the volume v_{i_k} of the trees i_k in a sub-sample. We only need to determine the statistical relationship between τ_v, τ_c, and τ_h to complete the procedure. ■

The MnP sampling method helps to reduce the **cost of enumeration** and increase the **accuracy of estimates**. It can also be used to **stratify** a population: an initial sample is taken based on the auxiliary characteristic, which is used to subdivide the population into strata in which the main characteristic is more or less **homogeneous**.

As long as the two characteristics are **correlated**, accurate estimates of the main characteristic are obtained from a second, relatively small sample.

M2P can also be paired with M2S, for example (or with any other sampling design). If both selection steps are performed with SRS, the method is called **two-phase simple random sampling** (SRS2P).

In the first phase, the population is divided into well-defined sampling units; a SRS $\mathcal{Y}_1$ of size n_1 is drawn from these units; the **auxiliary variable** x is measured on all units of $\mathcal{Y}_1$. Next, a sub-SRS $\mathcal{Y}_2$ of size n_2 is drawn from $\mathcal{Y}_1$; the **main characteristic** y is measured on all units of $\mathcal{Y}_2$.

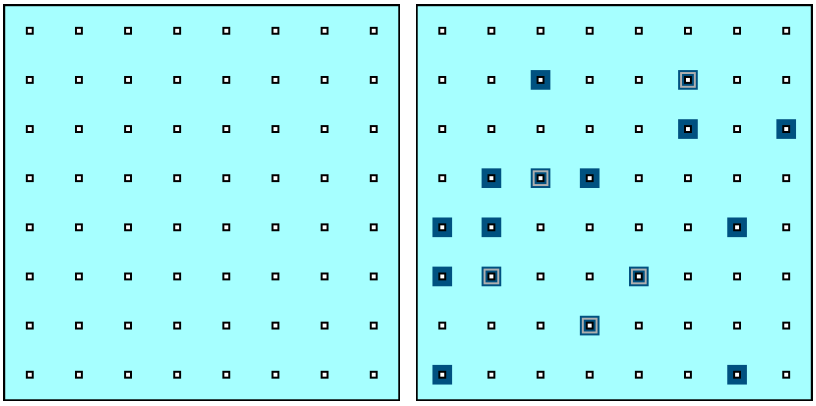


Figure 11.14: Schematics of SRS2P: target population (left) and sample (right).

We can evaluate $r_{\mathcal{Y}_2}$ or $b_{\mathcal{Y}_2}$ from the observations in $\mathcal{Y}_2$ (using either the ratio method or the regression method), which yields

$$\begin{aligned}\hat{\mu}_{Y;R;\text{SRS2P}} &= r_{\mathcal{Y}_2}\cdot\overline{x}_{\mathcal{Y}_1} \quad \text{or}\\ \hat{\mu}_{Y;L;\text{SRS2P}} &= \overline{y}_{\mathcal{Y}_2} + b_{\mathcal{Y}_2}(\overline{x}_{\mathcal{Y}_1} - \overline{x}_{\mathcal{Y}_2}).\end{aligned}$$

Estimation

Due to the **double sampling**, two terms contribute to sampling variances of the estimators (the first when going from $\mathcal{U}$ to $\mathcal{Y}_1$, and the second from $\mathcal{Y}_1$ to $\mathcal{Y}_2$):

$$\begin{aligned}\hat{\text{V}}(\hat{\mu}_{Y;R;\text{SRS2P}}) &= \frac{1}{n_2}\left(s_Y^2 - 2r_{\mathcal{Y}_2}s_{XY} + (r_{\mathcal{Y}_2})^2 s_X^2\right) + \frac{1}{n_1}\left(2r_{\mathcal{Y}_2}s_{XY} - (r_{\mathcal{Y}_2})^2 s_X^2\right)\\ \hat{\text{V}}(\hat{\mu}_{Y;L;\text{SRS2P}}) &= \frac{1}{n_2}s^2_{XY;L} + \frac{1}{n_1}\left(s^2_{XY;L} - s_Y^2\right)\end{aligned}$$

where s_Y^2, s_{XY}, and s_X^2 are the usual quantities (in $\mathcal{Y}_2$), and

$$\begin{aligned}r_{\mathcal{Y}_2} &= \frac{\overline{y}_{\mathcal{Y}_2}}{\overline{x}_{\mathcal{Y}_2}},\; b_{\mathcal{Y}_2} = \frac{s_{XY}}{s_X^2}, \quad \text{and}\\ s^2_{XY;L} &= \frac{n_2-1}{n_2-2}\cdot\left\{s_Y^2 - b^2_{\mathcal{Y}_2}s_X^2\right\}\end{aligned}.$$

Example We are interested in the biomass of any plant in a region, which is divided into plots of 0.025 ha each. First, we measure the number x of groves per unit in a SRS $\mathcal{Y}_1$ of $n_1 = 200$ plots.

Then, the biomass y of the plant in question is calculated in each unit of a sub-SRS $\mathcal{Y}_2$ of $n_2 = 40$ plots:

$$\overline{x}_{\mathcal{Y}_1} = 374.4;\quad \sum_{i=1}^{40} x_i = 15,419;\quad \sum_{i=1}^{40} y_i = 2104;$$

$$\sum_{i=1}^{40} x_i^2 = 7,744,481;\quad \sum_{i=1}^{40} x_iy_i = 960,320;\quad \sum_{i=1}^{40} y_i^2 = 125,346.$$

What would a 95% C.I. for the average biomass per plot look like?

Let us compute the required intermediate quantities:

$$\begin{aligned}
\overline{x}_{\mathcal{Y}_2} &= \frac{15419}{40} = 385.5; \quad \overline{y}_{\mathcal{Y}_2} = \frac{2104}{40} = 52.6; \quad r_{\mathcal{Y}_2} = \frac{\overline{y}_{\mathcal{Y}_2}}{\overline{x}_{\mathcal{Y}_2}} = \frac{52.6}{385.5} = 0.14; \\
s_X^2 &= \frac{1}{39}[7744481 - 40(385.5)^2] \approx 46175; \; s_Y^2 = \frac{1}{39}[125346 - 40(52.6)^2] \approx 376 \\
s_{XY} &= \frac{1}{39}[960320 - 40(385.5)(52.6)] \approx 3827.7; \quad b_{\mathcal{Y}_2} = \frac{s_{XY}}{s_X^2} = \frac{3827.7}{46175.4} \approx 0.08; \\
s_{XY;L}^2 &= \frac{39}{38}[376.3 - 0.08^2(46175.4)] \approx 82.9;
\end{aligned}$$

which gives us

$$\hat{\mu}_{Y;R;\text{SRS2P}} = 0.14(374.4) \approx 51.1; \quad \hat{\mu}_{Y;L;\text{SRS2P}} = 52.6 + 0.08(374.4 - 385.5) \approx 51.7$$

and

$$\begin{aligned}
\hat{\text{V}}(\hat{\mu}_{Y;R;\text{SRS2P}}) &= \frac{376.3 - 2(0.14)(3827.7) + (0.14)^2 46175.4}{40} \\
&\quad + \frac{2(0.14)3827.7 - (0.14)^2 46175.4}{200} \approx 5.67; \\
\hat{\text{V}}(\hat{\mu}_{Y;L;\text{SRS2P}}) &= \frac{82.9}{40} + \frac{82.9 - 376.3}{200} \approx 3.54;
\end{aligned}$$

from which we conclude that

$$\begin{aligned}
\text{C.I.}_{R;\text{SRS2P}}(\mu_Y; 0.95) &= 51.1 \pm 2\sqrt{5.67} \equiv (46.3, 55.8) \\
\text{C.I.}_{L;\text{SRS2P}}(\mu_Y; 0.95) &= 51.7 \pm 2\sqrt{3.54} \equiv (47.9, 55.5). \quad \blacksquare
\end{aligned}$$

11.7.5 Miscellaneous

We end the module by briefly discussing a few notions that did not find a natural slot in the previous sections:

- design effects;
- adjusting for non-response;
- estimating the size of a population,
- randomized responses, and
- Bernoulli sampling.

Design Effect

The **design effect** compares the estimator for a given sampling design and for a SRS. It is the ratio of the **sampling variance of the estimator under the given sampling design** to the **sampling variance of the estimator under a SRS** (assuming samples of the **same size**).

This value is often applied to compare the **efficiency** of estimators from different sampling designs. If the ratio < 1, the sampling design is more efficient than SRS; if it is > 1, it is less efficient than SRS.

We directly compared the theoretical variances of several sampling designs in sections 11.4.3, 11.5.4, and 11.6.3, but in practice we compute the design effect using the achieved samples (assuming that they had been drawn under various sampling plans).

Design effects also help to obtain approximate variance estimates for complex **sampling designs**. If a design effect estimate is available from a previous survey (that used the sampling design we will be using for this survey), it can be used to determine the **sample size** required to meet some pre-determined condition(s).

Adjusting for Nonresponse

Non-response is a problem in **all** surveys. **Total non-response** (when all or almost all data from a sampled unit are missing) occurs when:

- a sample unit **refuses to participate** in the survey;
- we cannot **establish contact with a sample unit**;
- the sampled unit cannot be **found**, or
- the information obtained form the unit is **useless/invalid**.

The simplest way to deal with such non-response is to ignore it; in some **exceptional** circumstances (when the affected observations are not in any way different from those for whom we have valid and complete measurements), proportions or means that are estimated without adjusting for non-response are **more or less the same** as those produced by applying adjustment for non-response.

If one neglects to **compensate** for nonresponding units, however, the **totals are generally underestimated** (e.g., the size of a population, total revenue, or total acres harvested, say).

The most common way to deal with total non-response is to **adjust the base sampling weights** by assuming that the responding units represent both responding and nonresponding units. If the **nonrespondents are equivalent to the respondents** for the characteristics measured in the survey, this is a reasonable approach.

The base weights for nonrespondents are then redistributed among respondents, using a **adjustment factor for nonrespondents** that is multiplied by the base weight, to obtain an adjusted weight.

Example If we draw a SRS of size $n = 25$ from a stratum of size $N = 1000$, the **probability of inclusion** of each of these units and the corresponding **basic weight** are

$$\pi = \frac{n}{N} = \frac{25}{1000} = 0.025, \quad w = \frac{1}{\pi} = \frac{1}{0.025} = 40.$$

In other words, each selected unit represents \$40 units in the stratum.

If we only get a response from $n_r = 20$ of the $n = 25$ selected units, the **non-response adjustment factor** (NRAF) and the **adjusted weight** (for non-response) become:

$$\text{NRAF} = \frac{n}{n_r} = \frac{25}{20} = 1.25$$
$$w_{\text{nr}} = w \cdot \text{NRAF} = 1.25(40) = 50;$$

each responding unit then represents 50 units in the stratum. This adjusted weighting is what we would end up working with. ■

Of course, the adjusted weight may vary from stratum to stratum, depending on the sample design and the sample size/allocation.

When we want to determine the optimal sample size/allocation across various strata, what we obtain is the **target sample size** (assuming that the target and study populations coincide). We then have to resort to **inflation** of the sample size to achieve the target.

Example The allocation of a StS of size $n = 29$ is found to be $(17, 9, 3)$. In a prior study, the non-response rates by stratum were determined to be $(16.2\%, 20.8\%, 31.2\%)$. Which allocation optimizes the likelihood of achieving the target allocation?

We only need to solve

$$n_1(1 - 0.162) = 17, \quad n_2(1 - 0.208) = 9, \quad n_3(1 - 0.312) = 3,$$

which gives a practical sample allocation of $(n_1, n_2, n_3) = (20.3, 11.3, 4.3) \approx (21, 12, 5)$, and a practical sample size of $n = 38$.

Estimating a Population Size

How do we proceed if the size N of the population $\mathcal{U}$ is unknown? When the population is large enough, we can always use the approximation $N \approx \infty$ in the sampling variance formulas.

But sometimes it is the parameter N that represents the quantity of interest; as an example, how would we find out the number N of \$5 bill in circulation?

We approach such a problem using the **catch-and-release** method (compare with the approach used in Module 26):

1. we capture n_1 bills at random (without replacement) from the population;
2. we mark them and release them back into circulation;
3. at a later time, n_2 bills are captured at random (without replacement) from the population;
4. we count the number X of marked bills, $0 < X \leq n_2$.

If we wait long enough (to let the marked bills propagate in the population, say), we obtain

$$\frac{n_1}{N} \approx \frac{X}{n_2}, \quad \text{from which we have } \hat{N} = \frac{n_1 n_2}{X},$$

where X follows a **hypergeometric** distribution with parameters $n_1, N -$

n_1, n_2, and probability mass function

$$P(X=x)=\frac{\binom{n_1}{x}\binom{N-n_1}{n_2-x}}{\binom{N}{n_2}},\quad 0\leq x\leq n_2$$

$$\mu_X=\mathrm{E}[X]=n_2\underbrace{\left(\frac{n_1}{N}\right)}_{p}=n_2p,\quad \sigma_X^2=\mathrm{V}[X]=n_2p(1-p)\left(\frac{N-n_2}{N-1}\right).$$

If $\frac{n_2}{N}<0.05$, we can ignore the FPCF term in the variance:

$$\sigma_X^2=\mathrm{V}[X]\approx n_2p(1-p).$$

We can now develop expressions for $\mathrm{E}[\hat{N}]$ and $\mathrm{V}[\hat{N}]$, using a **Taylor series of order 2 near** $X\approx\mu_X=n_2p$:

$$f(X)\approx f(\mu_X)+f'(\mu_X)(X-\mu_X)+\frac{f''(\mu_X)}{2}(X-\mu_X)^2.$$

Si $\hat{N}=f(X)=\frac{n_1n_2}{X}$, so that

$$\begin{aligned}\hat{N}&\approx\frac{n_1n_2}{\mu_X}-\frac{n_1n_2}{\mu_X^2}(X-\mu_X)+\frac{n_1n_2}{\mu_X^3}(X-\mu_X)^2\\&=\frac{n_1}{p}-\frac{n_1}{n_2p^2}(X-n_2p)+\frac{n_1}{n_2^2p^3}(X-n_2p)^3.\end{aligned}$$

Consequently,

$$\begin{aligned}\mathrm{E}[\hat{N}]&=\mathrm{E}\left[\frac{n_1n_2}{\mu_X}-\frac{n_1n_2}{\mu_X^2}(X-\mu_X)+\frac{n_1n_2}{\mu_X^3}(X-\mu_X)^2\right]\\&=\mathrm{E}\left[\frac{n_1n_2}{\mu_X}\right]-\mathrm{E}\left[\frac{n_1n_2}{\mu_X^2}(X-\mu_X)\right]+\mathrm{E}\left[\frac{n_1n_2}{\mu_X^3}(X-\mu_X)^2\right]\\&=\frac{n_1n_2}{\mu_X}-\frac{n_1n_2}{\mu_X^2}(\underbrace{\mathrm{E}[X]}_{\mu_X}-\mu_X)+\frac{n_1n_2}{\mu_X^3}\mathrm{E}\left[(X-\mu_X)^2\right]\\&=\frac{n_1n_2}{\mu_X}+\frac{n_1n_2}{\mu_X^3}\mathrm{V}[X]\approx\frac{n_1}{p}+\frac{n_1}{n_2^2p^3}\cdot n_2p(1-p)=\frac{n_1}{p}+\frac{n_1}{n_2p^2}(1-p)\\&=\frac{n_1}{p}\left(1+\frac{1-p}{n_2p}\right)=N\left(1+\frac{1-p}{n_2p}\right).\end{aligned}$$

Since $\frac{1-p}{n_2p}>0$, $\mathrm{E}[\hat{N}]\neq N$, and so $\hat{N}$ is an **asympotically unbiased estimator of** N when the sample size n_2 increases.

We can provide an approximation of the variance using a **Taylor series of order 1 near** $X\approx\mu_X=n_2p$:

$$\hat{N}\approx\frac{n_1n_2}{\mu_X}-\frac{n_1n_2}{\mu_X^2}(X-\mu_X)=\frac{n_1}{p}\left(1-\frac{X-n_2p}{n_2p}\right)=\frac{n_1}{p}\left(2-\frac{X}{n_2p}\right).$$

Putting all this together, we get

$$\begin{aligned} \mathrm{V}[\hat{N}] &\approx \mathrm{V}\left[\frac{n_1}{p}\left(2-\frac{X}{n_2p}\right)\right] = \frac{n_1^2}{p^2}\cdot\mathrm{V}\left[-\frac{X}{n_2p}\right] = \frac{n_1^2}{n_2^2p^4}\cdot\mathrm{V}[X] \\ &\approx \frac{n_1^2n_2p(1-p)}{n_2^2p^4} = \frac{n_1^2(1-p)}{n_2p^3}. \end{aligned}$$

In practice, we do not know the true p, so we use

$$\hat{\mathrm{V}}[\hat{N}] = \frac{n_1^2(1-\hat{p})}{n_2\hat{p}^3}, \quad \text{where } \hat{p} = \frac{X}{n_2}.$$

Central Limit Theorem – Population Size N**:** if n_2 and N are sufficiently large, we have

$$\hat{N} \sim_{\text{approx.}} \mathcal{N}\left(\mathrm{E}[\hat{N}], \hat{\mathrm{V}}[\hat{N}]\right) \approx \mathcal{N}\left(\frac{n_1n_2}{X}, \frac{n_1^2(1-\hat{p})}{n_2\hat{p}^3}\right),$$

and the corresponding **95% C.I. for** N is thus

$$\text{C.I.}(N; 0.95): \quad \frac{n_1n_2}{X} \pm 2\sqrt{\frac{n_1^2(1-\hat{p})}{n_2\hat{p}^3}}.$$

Example Say that $n_1 = 500$ bills were initially captured, marked, and releases; of the $n_2 = 300$ bills recaptured at a later date, $X = 127$ were marked. Give a 95% C.I. for the total number of $5.

The point estimate is $\hat{N} = \frac{500\cdot 300}{127} \approx 1181.102$. We also have $\hat{p} = \frac{X}{n_2} = \frac{127}{300} \approx 0.423$, from which we get the bound on the error of estimation

$$2\sqrt{\hat{\mathrm{V}}(\hat{N})} = 2\sqrt{\frac{500^2\cdot(1-0.42)}{300\cdot(0.42)^3}} = 159.176,$$

and

$$\text{C.I.}(N; 0.95): \quad 1181.102 \pm 159.176 \equiv (1021.9, 1340.3). \quad \blacksquare$$

Randomized Response

Let's say we ask students whether they cheated on a test or an assignment during the pandemic. If the answer is "Yes," we can likely conclude that it is the true answer. But since there is a **social cost** associated with such an answer, we can expect that some cheaters will answer "No". What can we do to reduce the measurement error for **sensitive** questions?

First approach: with such questions, the skill of the interviewer plays a crucial role – this aspect should not be overlooked.

Second approach: the **randomized response** technique requires the use of two questions:

- the **sensitive question**, and

- an **innocent** question,

as well as a **random mechanism with known parameters** (heads or tails, etc.).

Randomized responses work as follows: the respondent flips a coin (without announcing the result to the interviewer), and answers honestly one of the 2 questions:

- "**head**": "Have you ever cheated on a test?";
- "**tail**": "Were you born in January?";

Since the interviewer does not know the outcome of the draw, they do not know whether the respondent is answering the sensitive question or the innocent one. **In theory**, the anonymity provided by the randomized response is freeing (the social cost is **diminished, if not eliminated altogether**) – therefore, we could expect an honest answer, regardless of the question.

But we have to be careful: this approach can only be successful if we know the probabilities:

- θ of **observing a positive response to the innocent question**;
- ρ of **the question being answered actually being the sensitive question**, and
- ϕ of **observing a positive response, whatever the question**.

Let p be the **proportion of positive responses to the sensitive question**, which is the quantity of interest. According to the Law of Total Probability, we have

$$\begin{aligned}\phi &= P(\text{positive response}) \\ &= \underbrace{P(\text{positive} \mid \text{sensitive})}_{p} \times \underbrace{P(\text{sensitive})}_{\rho} + \underbrace{P(\text{positive} \mid \text{innocent})}_{\theta} \times \underbrace{P(\text{innocent})}_{1-\rho}, \\ &= p\rho + \theta(1-\rho)\end{aligned}$$

or

$$p = \frac{\phi - \theta(1-\rho)}{\rho}.$$

If $\hat{\phi}$ is the proportion of positive responses in the achieved sample, then the **randomized response estimator** is

$$\hat{p}_{\text{rr}} = \frac{\hat{\phi} - \theta(1-\rho)}{\rho}, \quad \theta, \rho \text{ constants},$$

whose sampling variance is

$$\text{V}(\hat{p}_{\text{rr}}) = \text{V}\left(\frac{\hat{\phi} - \theta(1-\rho)}{\rho}\right) = \text{V}\left(\frac{\hat{\phi}}{\rho}\right) = \frac{1}{\rho^2} \cdot \text{V}(\hat{\phi}).$$

Since $\hat{\phi}$ is a SRS proportion estimator obtained from a sample of size n in a population $\mathcal{U}$ of size N, its **sampling variance** is

$$\text{V}(\hat{\phi}) = \frac{\phi(1-\phi)}{n}\left(\frac{N-n}{N-1}\right),$$

from which we conclude that

$$\mathrm{V}(\hat{p}_{\mathrm{rr}}) = \frac{1}{\rho^2}\cdot\frac{\phi(1-\phi)}{n}\left(\frac{N-n}{N-1}\right).$$

As the true value of ϕ is typically not known, we instead use the unbiased estimator

$$\hat{\mathrm{V}}(\hat{p}_{\mathrm{rr}}) = \frac{1}{\rho^2}\cdot\frac{\hat{\phi}(1-\hat{\phi})}{n-1}\left(1-\frac{n}{N}\right),$$

and we build a **95% C.I. for** *p via*

$$\mathrm{C.I.}_{\mathrm{rr}}(p;0.95): \quad \hat{p}_{\mathrm{rr}} \pm 2\sqrt{\hat{\mathrm{V}}(\hat{p}_{\mathrm{rr}})}.$$

The factor $1/\rho^2$ **penalizes the uncertainty** brought by the randomized response – the higher ρ is, the lower $\hat{\mathrm{V}}(\hat{p}_{\mathrm{rr}})$ is.

There are practical considerations that limit how high ρ can get: if it is **too large**, the anonymity conferred by the approach evaporates, and we risk ruining the study by causing an increase in non-response.

Example We seek to determine the incidence of cheating in online courses among students in the Department of Mathematics and Statistics ($N = 442$), using a SRS with $n = 65$. We use the scheme described in this section with $\rho = 1/2$, and observe $\theta = \frac{52}{442}$ and $\hat{\phi} = \frac{21}{65}$. Find a 95% C.I. for the proportion of students who cheated during the pandemic.

We only need compute

$$\begin{aligned}\hat{p}_{\mathrm{rr}} &= \frac{21/65 - 52/442(1-1/2)}{1/2} = 0.53\\ \hat{\mathrm{V}}(\hat{p}_{\mathrm{rr}}) &= \frac{1}{1/2^2}\cdot\frac{21/65(1-21/65)}{65-1}\left(1-\frac{65}{442}\right) = 0.012,\end{aligned}$$

which yields $\mathrm{C.I.}_{\mathrm{rr}}(p;0.95) = 0.53 \pm 2\sqrt{0.012} \equiv (0.31, 0.74)$. ■

Bernoulli Sampling

Bernoulli sampling (BS) is a **random** sampling design – we do not know the sample size **before** it is drawn.

Each unit of the population $\mathcal{U} = \{u_1, \dots, u_N\}$ is assigned the same probability of inclusion in the sample $\mathcal{Y}$: $\pi_j = \pi \in (0,1)$, for all j. We denote the **achieved sample size** by n_a.

The BS design[59] consists of performing N independent Bernoulli trials, each with probability of success π (where a success means that the unit is **included** in the sample, and a failure means that it **rejected**).

The probability of obtaining a sample $\mathcal{Y}$ of size n_a is then:

$$P(|\mathcal{Y}| = n_a) = \pi^{n_a}(1-\pi)^{N-n_a}.$$

There are 2^N possible samples, with size varying from $n_a = 0$ to $n_a = N$.

[59] I know, I know.

The sample size follows a **binomial** distribution $n_a \sim B(N, \pi)$:

$$P(n_a = n) = \binom{N}{n}\pi^n(1-\pi)^{N-n}, \quad \mathrm{E}[n_a] = N\pi, \quad \mathrm{V}[n_a] = N\pi(1-\pi).$$

When N is sufficiently large, this distribution is **approximately normal**; the **95% C.I. for** n is thus

$$\text{C.I.}(n_a; 0.95): \quad N\pi \pm 2\sqrt{N\pi(1-\pi)}.$$

Let $\pi_{j,k}$ be the probability of inclusion of units u_j and u_k, $j \neq k$ in the smaple $\mathcal{Y}$. Since the Bernouilli trials are independent of one another,

$$\pi_{j,k} = P(\{u_j, u_k\} \in \mathcal{Y}) = P(u_j \in \mathcal{Y}) \cdot P(u_k \in \mathcal{Y}) = \pi_j \pi_k = \pi^2.$$

The estimator

$$\hat{\tau}_{\text{BS}} = \frac{1}{\pi}\sum_{i=1}^{n_a} y_i$$

is an **unbiased estimator of the total** τ in $\mathcal{U}$: indeed,

$$\mathrm{E}[\hat{\tau}_{\text{BS}}] = \frac{1}{\pi}\mathrm{E}[n_a\overline{y}] = \frac{\mathrm{E}[n_a]\mathrm{E}[\overline{y}]}{\pi} = \frac{N\pi\mu}{\pi} = N\mu = \tau,$$

as n_a and $\overline{y}$ are independent of each other.

In the same vein, the **sampling variance** of $\hat{\tau}_{\text{BS}}$ is approximately

$$\hat{\mathrm{V}}[\hat{\tau}_{\text{BS}}] = \frac{1}{\pi}\left(\frac{1}{\pi} - 1\right)\sum_{i=1}^{n_a} y_i^2.$$

If N and n_a are sufficiently large, the Central Limit Theorem comes into play again, and we build a **95% C.I. for** τ using

$$\text{C.I.}_{\text{BS}}(\tau; 0.95): \quad \hat{\tau}_{\text{BS}} \pm 2\sqrt{\hat{\mathrm{V}}[\hat{\tau}_{\text{BS}}]}.$$

The corresponding estimators for the mean $\overline{y}_{\text{BS}}$ and the proportion $\hat{p}_{\text{BS}}$ are obtained in the usual manner.

Example A teacher has to correct 600 exam papers. For each paper, she rolls a die and only corrects it (at this stage) if it shows a 6.

At the end of the process, she has graded 90 papers, of which 60 have received a passing grade. Find a 95% C.I. for the total number of passes in her class.

Let $y_i = 1$ if the ith marked examen received a passing grade, and $y_i = 0$

otherwise. We have $N = 600$, $\pi = 1/6$, $n_a = 90$,

$$\begin{aligned}
&\sum_{i=1}^{90} y_i = 60, \\
&\sum_{i=1}^{90} y_i^2 = 60, \\
&\hat{\tau}_{\text{BS}} = \frac{1}{1/6}\sum_{i=1}^{90} y_i = 6(60) = 360 \\
&\hat{\text{V}}[\hat{\tau}_{\text{BS}}] = \frac{1}{1/6}\left(\frac{1}{1/6}-1\right)\sum_{i=1}^{90} y_i^2 = 6(5)(60) = 1800.
\end{aligned}$$

The 95% C.I. is thus $\text{C.I.}_{\text{BS}}(\tau; 0.95) = 360 \pm 2\sqrt{1800} \equiv [277, 443]$. We are not going to lie... it is looking particularly bleak for the students. ■

11.8 Exercises

1. You are tasked with estimating the annual salary of data scientists in Canada. Determine the:
   - populations (target, study, respondent);
   - sampling frames;
   - samples (target, achieved);
   - information about units (units, response variable, attributes);
   - sources of error (coverage, non-response, sampling, measurement and processing) and variability (sampling, measurement).
2. We seek to estimate the average daily distance travelled by Ontario cars, as well as their daily fuel consumption. Discuss various approaches to be used. What are some of the issues and challenges that could be encountered?
3. We seek an estimate of the average daily distance travelled in Winter 2012 in Ontario, as are the average daily fuel consumption and the proportion of vehicles not in use. An SRS is selected from the Ontario fleet (size $N=$7,868,359); the responses are collected in the file `Autos.xlsx` ↗. Discuss issues that may affect the quality of the data. Provide a numerical and visual summary of the data for the sample. Give an approximate 95% C.I. for each population mean sought, with corresponding coefficient of variation.
4. We seek an estimate of the average daily distance travelled in Winter 2012 in Ontario, as are the average daily fuel consumption and the proportion of vehicles not in use. An STS is selected from the Ontario fleet (size $N=$7,868,359), with information concerning vehicle type and age (the strata); the responses are collected in the file `Autos.xlsx` ↗. Discuss issues that may affect the quality of the data. Provide a numerical and visual summary of the data for the sample. Give an approximate 95% C.I. for each population mean sought, with corresponding coefficient of variation. Conduct the same exercise for each stratum.
5. We seek an estimate of the average daily distance travelled in Winter 2012 in Ontario. An SRS is selected from the Ontario fleet (size

$N=$7,868,359). The responses, as well as the corresponding daily fuel consumption, are collected in the file Autos.xlsx . Give an approximate 95% C.I. for the characteristic of interest using quotient, regression, and difference estimation.

6. Could cluster sampling be used to provide estimates of average daily distance travelled, average daily fuel consumption, and proportion of vehicles not in use in Winter 2012 in Ontario? Treat the vehicle type and age information found in Autos.xlsx as cluster information.
7. Repeat the previous exercise using multi-phase and multi-stage sampling.
8. Draw $m = 1000$ SRS samples of size n from the $N = 183$ countries (excluding China and India) in the 2011 Gapminder dataset to estimate the average propulation by country μ. For $n = 30, 60, 90, 120$, what proportion of the m samples yield an approximate 95% C.I. containing μ? Assume that σ^2 is not known.
9. Find an approximate 95% C.I. for the average life expectancy μ of the $N = 185$ countries in the 2011 Gapminder dataset using a SRS of size $n = 20$. Is the true average life expectancy in your confidence interval? Repeat this task $m = 1000$ times, with different SRS samples. What proportion of the m samples yield approximate 95% C.I. containing μ? Assume that σ^2 is not known. Compare with the results of the previous exercise. How do you explain the discrepancy?
10. Find an approximate 95% C.I. for the proportion p of countries whose life expectancy fell below 60 years in the 2011 Gapminder dataset ($N = 185$), using a SRS of size $n = 20$. Is the true proportion in the confidence interval? Repeat this task $m = 1000$ times, with different SRS samples. What proportion of the m samples yield approximate 95% C.I. containing the true p? Assume that σ^2 is not known. Compare with the results of exercises 8 and 9.
11. Find an approximate 95% C.I. for the total population of the planet in the 2011 Gapminder dataset ($N = 185$), using a STS of size $n = 20$. What variable will you use to stratify the data? Repeat this task $m = 1000$ times, with different STS samples. What proportion of the m samples yield approximate 95% C.I. containing the true total τ? Is the distribution of the obtained totals (approximately) normal? How do you explain the shape of this distribution?
12. Find an approximate 95% C.I. for the proportion p of countries whose life expectancy fell below 60 years in the 2011 Gapminder dataset ($N = 185$), using a STS of size $n = 20$. What variable will you use to stratify the data? Is the true proportion in the confidence interval? Repeat this task $m = 1000$ times, with different STS samples. What proportion of the m samples yield approximate 95% C.I. containing the true p? Compare with the results of exercise 10.
13. Consider a sample $\mathcal{Y} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ drawn from a population of size $N = 37,444$. In a preceding study, we have found that $\sigma^2_{W;L} \approx 188.2$. Find the minimal n which ensures that the bound on the error of (regression) estimation of the mean μ_Y is at most 5. Do the same for the total τ_Y and a bound of at most 250.
14. Find a 95% C.I. for the proportion of countries in the 2011 Gapminder dataset ($N = 185$) whose life expectancy is above 75 years, using a CLS with $m = 8$, assuming that the countries are grouped

into $M = 22$ clusters determined by geographic regions. Assume further that the average cluster size is known to be $\overline{N} = 8.41$.

15. Consider a CLS $\mathcal{Y}$ consisting of m clusters drawn from a population $\mathcal{U}$ of size N, distributed in M clusters. Let μ be the mean and σ^2 the variance of the population $\mathcal{U}$. If the clusters are all of size n, show that

$$\mathrm{V}(\overline{y}_C) \approx \frac{\sigma^2 - \overline{\sigma^2}}{m}\left(1 - \frac{m}{M}\right), \quad \text{where } \overline{\sigma^2} = \frac{1}{M}\sum_{\ell=1}^{M} \sigma_\ell^2,$$

where σ_ℓ^2 is the variance in the ℓth cluster.