An introduction to sampling from a finite population

Federico Andreis
Department of Policy Analysis and Public Management
Università Commerciale Luigi Bocconi
federico.andreis@unibocconi.it

Karolinska Institutet
17 October 2017

Course structure

What is this workshop about?

Today is devoted to provide an introduction to finite population sampling designs, both from a theoretical and applied point of view.

We will:

- take a look at the theoretical framework
- introduce some common probability sampling schemes
- discuss estimation of population parameters
- learn how to implement them in `R`.
- **Goals**: after this course the participants should be able to
  - identify the components of a sampling strategy
  - construct and implement a probability sampling design
  - estimate relevant quantities and correctly interpret the results.

Agenda

The day is structured as follows:

- Morning: theory of finite population sampling
  - 9.00-10.30
  - **break!**
  - 10.45-12.00
- Lunch
- Afternoon: hands-on using R!
  - 13.00-14.30
  - **break!**
  - 14.45-16.00

I'll try to stick to swedish times!

Course material

This course's material includes:

- lecture notes, morning session
- lecture notes, afternoon session
- `R` workspaces containing the datasets for the practicals
- additional material on `R`.
- you! Feel free to raise your hand...

You can obtain everything from the stats4life website.

An introduction to survey sampling

What is survey research?

From the Encyclopedia of Survey Research Methods:

> *"Survey research is a systematic set of methods used to gather information to generate knowledge and to help make decisions."*

What is survey research?

From the Encyclopedia of Survey Research Methods:

> *"Survey research is a systematic set of methods used to gather information to generate knowledge and to help make decisions."*

Survey research encompasses **census** and **sampling** undertakings:

What is survey research?

From the Encyclopedia of Survey Research Methods:

> *"Survey research is a systematic set of methods used to gather information to generate knowledge and to help make decisions."*

Survey research encompasses **census** and **sampling** undertakings:

- a census survey seeks to collect complete information from all units in a population
- a sample survey uses a representative subgroup to determine characteristics of the entire population.

We will here focus on the latter case and refer to it as **survey sampling**.

Sample extraction and data collection

We can identify two primary defining characteristics of a survey:

1. a sample is taken from the population
2. a systematic instrument is used to gather data.

Today we are concerned with exploring 1. and the implications of different methods for extraction on the validity of studies based on survey data and generalizations thereof.

We will discuss selection methods under both a theoretical and an applied perspective, and hopefully reach an understanding of why the way data are collected is so relevant.

Sources of error

No survey effort is immune to errors. For this reason, it is extremely important to recognize their nature and handle their presence appropriately. Errors related to the data collection and processing include:

- **sampling error**: structural in the nature of sampling, probability sampling can give some control over it

Sources of error

No survey effort is immune to errors. For this reason, it is extremely important to recognize their nature and handle their presence appropriately. Errors related to the data collection and processing include:

- **sampling error**: structural in the nature of sampling, probability sampling can give some control over it
- **measurement errors**: mostly related to inadequate measurement tools (either machines, questionnaires, or the interviewers themselves)

Sources of error

No survey effort is immune to errors. For this reason, it is extremely important to recognize their nature and handle their presence appropriately. Errors related to the data collection and processing include:

- **sampling error**: structural in the nature of sampling, probability sampling can give some control over it
- **measurement errors**: mostly related to inadequate measurement tools (either machines, questionnaires, or the interviewers themselves)
- **missing responses**: one of the most important sources of bias, can arise either because of nonresponse, partial response, or bad recording

Sources of error

No survey effort is immune to errors. For this reason, it is extremely important to recognize their nature and handle their presence appropriately. Errors related to the data collection and processing include:

- **sampling error**: structural in the nature of sampling, probability sampling can give some control over it
- **measurement errors**: mostly related to inadequate measurement tools (either machines, questionnaires, or the interviewers themselves)
- **missing responses**: one of the most important sources of bias, can arise either because of nonresponse, partial response, or bad recording
- **processing errors**: mistakes in processing, editing, and analyzing sample data.

Target, frame, and survey population

When planning a survey sampling, it is important to have a clear understanding of the following distinctions

- **general population**: what contains the object of inference
- **target population**: what we want to make inference on

Target, frame, and survey population

When planning a survey sampling, it is important to have a clear understanding of the following distinctions

- **general population**: what contains the object of inference
- **target population**: what we want to make inference on
- **frame population**: what we are able to sample from

Target, frame, and survey population

When planning a survey sampling, it is important to have a clear understanding of the following distinctions

- **general population**: what contains the object of inference
- **target population**: what we want to make inference on
- **frame population**: what we are able to sample from
- **survey population**: what we can really observe.

Target, frame, and survey population

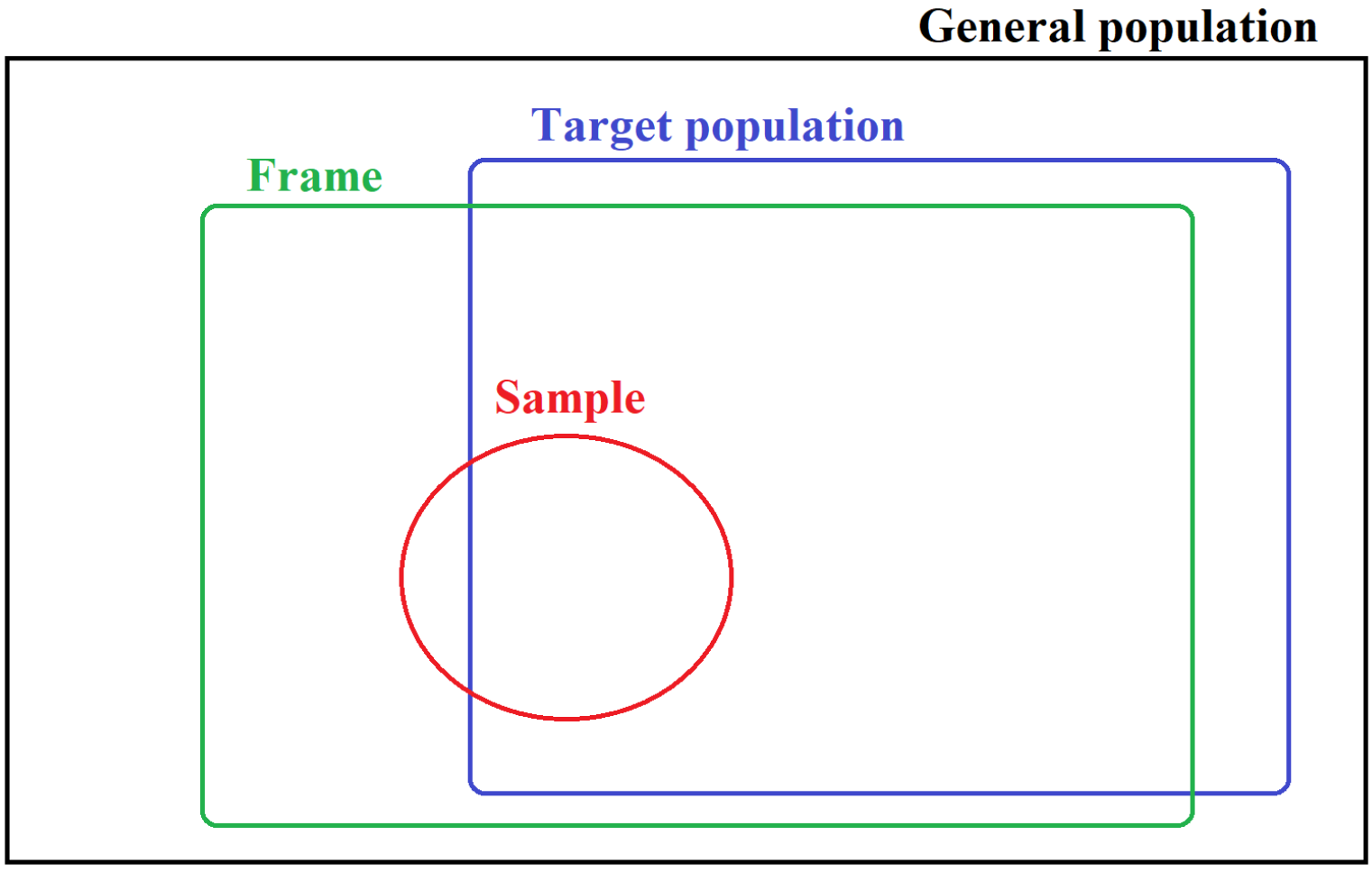
When planning a survey sampling, it is important to have a clear understanding of the following distinctions

- **general population**: what contains the object of inference
- **target population**: what we want to make inference on
- **frame population**: what we are able to sample from
- **survey population**: what we can really observe.

Ideally, there should be no distinction, meaning that all units in the target population should be eligible for extraction and all those selected to enter the sample do so.

In practice, this is seldom the case: incomplete frame coverage and unit nonresponse being the main culprits.

Sampling from a frame



Errors related to the frame

The quality of the sampling frame is extremely relevant. Common related problems include

- **undercoverage** : the frame does not contain all the units in the target population, hence some units have zero probability of being selected, potentially biasing the results

Errors related to the frame

The quality of the sampling frame is extremely relevant. Common related problems include

- **undercoverage** : the frame does not contain all the units in the target population, hence some units have zero probability of being selected, potentially biasing the results
- **duplications** : units appear more than once in the list, common when the frame is obtained by merging multiple frames

Errors related to the frame

The quality of the sampling frame is extremely relevant. Common related problems include

- **undercoverage** : the frame does not contain all the units in the target population, hence some units have zero probability of being selected, potentially biasing the results
- **duplications** : units appear more than once in the list, common when the frame is obtained by merging multiple frames
- **overcoverage** : the frame also contains units not in the target population, and while no bias should arise, it may lead to a waste of resources.

Special attention should be put on these potential issues: failing to account for frame deficiencies can lead to severe bias and invalid inference.

Probability vs nonprobability sampling

When the choice of which elements to include in a sample is not based on a randomized procedure, rather on arbitrary decisions, a **nonprobability sampling** is in place. In contrast, a **probability sampling** approach allows each unit in the population to have a (known) nonzero probability of being randomly selected.

Probability vs nonprobability sampling

When the choice of which elements to include in a sample is not based on a randomized procedure, rather on arbitrary decisions, a **nonprobability sampling** is in place. In contrast, a **probability sampling** approach allows each unit in the population to have a (known) nonzero probability of being randomly selected.

Starting with a classic 1934 paper by Neyman, the intrinsic limitations of the nonprobability approach have been eviscerated, mainly with respect to the impossibility of drawing inference from such samples 'as is'.

Probability vs nonprobability sampling

When the choice of which elements to include in a sample is not based on a randomized procedure, rather on arbitrary decisions, a **nonprobability sampling** is in place. In contrast, a **probability sampling** approach allows each unit in the population to have a (known) nonzero probability of being randomly selected.

Starting with a classic 1934 paper by Neyman, the intrinsic limitations of the nonprobability approach have been eviscerated, mainly with respect to the impossibility of drawing inference from such samples 'as is'.

Despite the weaknesses of nonprobability methods, they are still widely used, primarily for reasons of cost and convenience. Notable examples include convenience, purposive, and quota sampling. Today, the focus will be on probability methods.

Some definitions

A **population** is a set $U = \{u_1, u_2, ..., u_N\}$ composed by N elementary statistical units; we index the units with $i = 1, ..., N$. Let Y denote the survey variable and $Y_N = \{y_1, ..., y_N\} \in \mathbb{R}^N$ the (typically unknown) population vector of values of Y.

A **statistical parameter** θ is any function of Y_N that may be of interest. Important examples of $\theta(Y)$ include the population mean, total,

$$\mu_y = \frac{1}{N}\sum_{i=1}^{N} y_i, \quad T_y = \sum_{i=1}^{N} y_i$$

and variance

$$\sigma_y^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \mu_y)^2.$$

Sample, sample space, and sampling

A **sample** is a subset $\mathbf{s}$ of $n < N$ elements of U.

The **sampling fraction** $f = n/N \in (0, 1)$ indicates the proportion of units of the population that are to be sampled. This quantity is typically small.

The **sample space** Ω is the set of all the samples of n elements that can be formed from a population of N elements.

A **sampling design** is a probability distribution p of $\mathbf{s}$ over Ω, so that $p(\mathbf{s}) \in (0, 1)$ is the probability of extracting sample $\mathbf{s}$, and $\sum_{\mathbf{s} \in \Omega} p(\mathbf{s}) = 1$.

Inclusion probabilities

For a given sampling design p and sample space Ω, we define the **first-order inclusion probability** of unit i as

$$\pi_i = \sum_{\mathbf{c} \in \Omega_i} p(\mathbf{c})$$

where $\Omega_i \subseteq \Omega$ is the set of the samples that contain unit i.
$\pi_i \in [0, 1]$ is the probability that unit i enters the sample.

Inclusion probabilities

For a given sampling design p and sample space Ω, we define the **first-order inclusion probability** of unit i as

$$\pi_i = \sum_{\mathbf{c} \in \Omega_i} p(\mathbf{c})$$

where $\Omega_i \subseteq \Omega$ is the set of the samples that contain unit i.
$\pi_i \in [0, 1]$ is the probability that unit i enters the sample.

Similarly, we define the **second order inclusion probability** for units i and j as

$$\pi_{ij} = \sum_{\mathbf{c} \in \Omega_{ij}} p(\mathbf{c})$$

where $\Omega_{ij} \subseteq \Omega$ is the set of the samples that contain both i and j.
$\pi_{ij} \in [0, 1]$ is the probability that the couple (i, j) enters the sample.

Sampling weights

Ideally, the sample is a miniature of the population. In practice, it is often **unbalanced**, thus calling for an adjustment to improve representativity before inference can be drawn.

For this reason, **weights** are usually attached to sample observations before estimation is carried out. A common choice of weights is given by the inverse of the inclusion probabilities

$$w_i = \pi_i^{-1}, i \in \mathbf{s}$$

The so called **self-weighting** designs are such that all units have the same inclusion probabilities, hence no weighting is needed.

Estimators and estimates

An **estimator** $\widehat{\Theta}$ is a random variable whose values are a function of the sample realizations $(Y_1, Y_2, ..., Y_n)$. Its role is to provide information regarding a population parameter θ. **Note**: the sampling distribution of an estimator depends on the sampling design underlying the selection process.

An **estimate** $\widehat{\theta}$ is the value that an estimator $\widehat{\Theta}$ takes on in correspondence of a specific sample $(y_1, y_2, ..., y_n)$.

A **sampling strategy** for selecting a sample and estimating a population quantity is described by the triplet $\{\Omega, p, \widehat{\Theta}\}$.

Design-based and model-based estimation

Today we will restrict ourselves to the so called **design-based** (DB) approach to estimation, as opposed to **model-based** (MB).

Design-based and model-based estimation

Today we will restrict ourselves to the so called **design-based** (DB) approach to estimation, as opposed to **model-based** (MB).

In DB, the values of Y are considered to be fixed, and the variation in the estimates arises *only* by virtue of them being based on random samples from the population, rather than a census.

Design-based and model-based estimation

Today we will restrict ourselves to the so called **design-based** (DB) approach to estimation, as opposed to **model-based** (MB).

In DB, the values of Y are considered to be fixed, and the variation in the estimates arises *only* by virtue of them being based on random samples from the population, rather than a census.

In contrast, MB methods assume that the values $Y_1, ..., Y_N$ are the realizations of a random variable from a statistical model, and the observed values $y_1, ..., y_n$ can be thought of as originated from a transform of the same model, rather than due only to the random sampling component.

Q&A

1. Why isn't running a census always a good idea?

Q&A

1. Why isn't running a census always a good idea?
2. How can probability sampling give some control over sampling errors?

Q&A

1. Why isn't running a census always a good idea?
2. How can probability sampling give some control over sampling errors?
3. Why is undercoverage a potential source of large bias in estimation?

Q&A

1. Why isn't running a census always a good idea?
2. How can probability sampling give some control over sampling errors?
3. Why is undercoverage a potential source of large bias in estimation?
4. What is the difference, in terms of inclusion probabilities, between nonprobability and probability sampling designs?

Basic probability designs

Simple random sampling without replacement

This is the simplest design: units are extracted from U and are not replaced. Every sample in Ω has the same probability of being extracted, every unit in the population U has the same probability of being included in the sample, as does every pair of units.

Simple random sampling without replacement

This is the simplest design: units are extracted from U and are not replaced. Every sample in Ω has the same probability of being extracted, every unit in the population U has the same probability of being included in the sample, as does every pair of units.

Formally: given N and n, the sample space is composed by $\binom{N}{n}$ possible samples, hence for a **simple random sampling** (SRS) procedure

$$p(\mathbf{s}) = \binom{N}{n}^{-1}, \quad \forall \mathbf{s} \in \Omega$$

$$\pi_i = \frac{n}{N} = f, \quad \forall i \in U$$

$$\pi_{ij} = \frac{n(n-1)}{N(N-1)}, \quad \forall (i \neq j) \in U$$

SRS - point estimation

Given an SRS sample $\mathbf{s} = (y_1, ..., y_n)'$, an unbiased estimator of the population total is

$$\widehat{\tau}_y = \frac{N}{n} \sum_{k \in \mathbf{s}} y_k = N\bar{y}$$

with variance of $\widehat{\tau}_y$

$$V(\widehat{\tau}_y) = N^2 \frac{1-f}{n} S^2$$

where $S^2 = N/(N-1)\sigma_y^2$. The the variance of the estimator is unbiasedly estimated by

$$\widehat{V}(\widehat{\tau}_y) = N^2 \frac{1-f}{n} s^2$$

with $s^2 = (n-1)^{-1} \sum_{k \in \mathbf{s}} (y_k - \bar{y})^2$ the sample variance.

SRS - interval estimation

Under mild assumptions and with a sufficiently large n, the sample mean is approximately normally distributed

$$\bar{y} \approx N(\mu_y, \frac{1-f}{n} S^2)$$

hence a $1-\alpha$ level confidence interval for μ_y is immediately obtained as

$$\bar{y} \pm z_{1-\alpha/2} S \sqrt{\frac{1-f}{n}}.$$

If S is unknown and needs to be estimated, one then resorts to the Student t distribution

$$\bar{y} \pm t_{n-1;\alpha/2} s \sqrt{\frac{1-f}{n}}.$$

SRS - sample size determination

A relevant question in survey sampling is: how large should the sample size be, for our estimates to be reliable? Moreover: what does *reliable* even mean?

SRS - sample size determination

A relevant question in survey sampling is: how large should the sample size be, for our estimates to be reliable? Moreover: what does *reliable* even mean?

Consider an unbiased estimator $\widehat{\Theta}$ for θ. The associated $1-\alpha$ confidence interval can be expressed as

$$P\left(|\widehat{\Theta} - \theta| \leq d\right) = 1 - \alpha$$

where d is called the **margin of error** and is a function of $V(\widehat{\Theta})$.

SRS - sample size determination

A relevant question in survey sampling is: how large should the sample size be, for our estimates to be reliable? Moreover: what does *reliable* even mean?

Consider an unbiased estimator $\widehat{\Theta}$ for θ. The associated $1-\alpha$ confidence interval can be expressed as

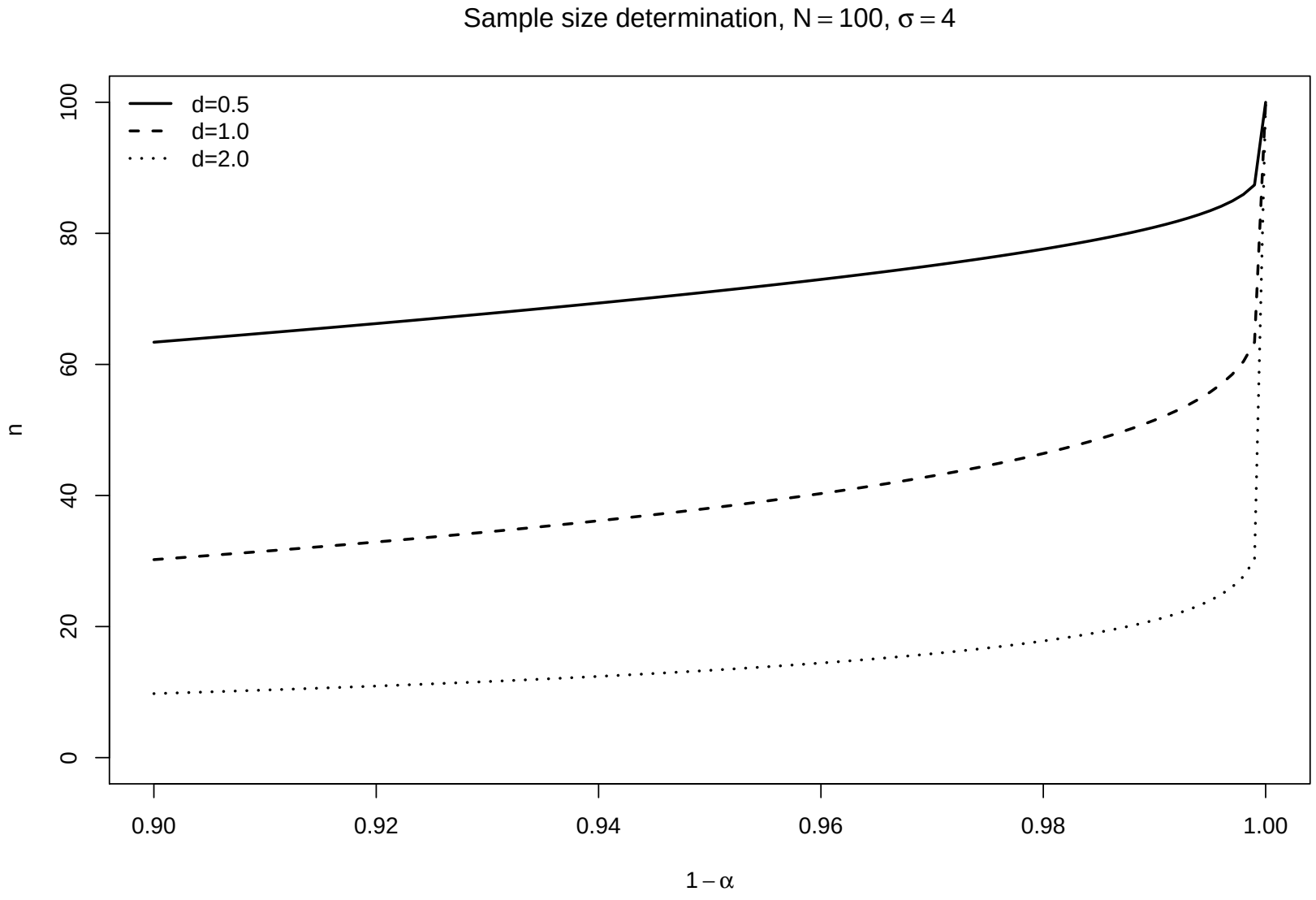
$$P\left(|\widehat{\Theta}-\theta| \leq d\right) = 1-\alpha$$

where d is called the **margin of error** and is a function of $V(\widehat{\Theta})$.

If we are estimating the mean μ_y, then $d = z_{1-\alpha/2} S \sqrt{\frac{1-f}{n}}$; it is then possible to solve for n, once α and d have been specified:

$$n = \left[\left(\frac{d}{z_{1-\alpha/2}\sigma_y}\right)^2 + \frac{1}{N}\right]^{-1} \approx_{N \gg 0} \frac{z_{1-\alpha/2}^2 \sigma_y^2}{d^2}$$

SRS - sample size



SRS - sample size

Similarly, for given d and α, it is possible to obtain expressions for the estimation of the total T_y

$$n = \left[\left(\frac{d}{N z_{1-\alpha/2} \sigma_y} \right)^2 + \frac{1}{N} \right]^{-1} \approx_{N \gg 0} \frac{N^2 z_{1-\alpha/2}^2 \sigma_y^2}{d^2}$$

and the proportion p_y

$$n \approx \frac{N p_y (1 - p_y)}{(N-1) \frac{d^2}{z_{1-\alpha/2}^2} + p_y (1 - p_y)} \approx_{N \gg 0} \frac{z_{1-\alpha/2}^2 p_y (1 - p_y)}{d^2}.$$

SRS - sample size

Similarly, for given d and α, it is possible to obtain expressions for the estimation of the total T_y

$$n = \left[\left(\frac{d}{N z_{1-\alpha/2} \sigma_y} \right)^2 + \frac{1}{N} \right]^{-1} \approx_{N \gg 0} \frac{N^2 z_{1-\alpha/2}^2 \sigma_y^2}{d^2}$$

and the proportion p_y

$$n \approx \frac{N p_y (1 - p_y)}{(N-1) \frac{d^2}{z_{1-\alpha/2}^2} + p_y (1 - p_y)} \approx_{N \gg 0} \frac{z_{1-\alpha/2}^2 p_y (1 - p_y)}{d^2}.$$

In practice, σ_y and p_y are typically unkown and must by estimated by s_y and $\widehat{p}_y$: this is an **educated guess**. For the proportion, another possibility is to be **conservative**, and replace p_y with 0.5.

The design effect

A very important concept in sampling theory is the **design effect**. Consider the estimation of the total: for a given sampling design p^* the design effect is defined as

$$deff = \frac{V(\widehat{\tau}_{p^*})}{V(\widehat{\tau})}$$

i.e., the ratio of the variance of the estimator of the total under design p^* to its variance under SRS.

The *deff* provides a measure of relative efficiency in estimating a population parameter (here the total) with a given design as compared to using an SRS. If the *deff* is < 1, then p^* is preferable from a statistical point of view.

Stratified sampling

When available auxiliary information allows to identify disjoint subpopulations (strata) that are homogeneous within and highly heterogeneous across, independent selection of SRS samples from each of them can lead to in creased accuracy in estimation. This is a **stratified sampling** approach.

Stratified sampling

When available auxiliary information allows to identify disjoint subpopulations (strata) that are homogeneous within and highly heterogeneous across, independent selection of SRS samples from each of them can lead to in creased accuracy in estimation. This is a **stratified sampling** approach.

Let $N = \sum_{h=1}^{H} N_h$, $n = \sum_{h=1}^{H} n_h$, where h indicates the strata, N_h is the strata size and n_h the strata-specific sample size. Then

$$p(\mathbf{s}) = \left[\prod_{h=1}^{H} \binom{N_h}{n_h} \right]^{-1}$$

$$\pi_{i(h)} = \frac{n_h}{N_h} = f_h, \quad \forall i \in U, h = 1, ..., H$$

$$\pi_{i(h)j(h')} = \begin{cases} \frac{n_h(n_h-1)}{N_h(N_h-1)}, & h = h' \\ \frac{n_h n_{h'}}{N_h N_{h'}}, & h \neq h' \end{cases}, \forall (i,j) \quad \in U$$

Stratified sampling - sample size allocation

The samples extracted from the strata need not have the same sample size. Indeed, that is but one of many possibilities that include, for example:

- **uniform allocation** $n_h = \left\lceil \frac{n}{H} \right\rceil, \forall h = 1, ..., H$
- **proportional allocation** $n_h = n\frac{N_h}{N}$
- **Neyman-Tschuprow allocation** $n_h = n\frac{N_h S_h}{\sum_{h=1}^{H} N_h S_h}$.

In any case, $\sum_{h=1}^{H} n_h = n$. The uniform allocation is straightforward, yet not always feasible. The Neyman-Tschuprow allocation yields the lowest variance estimators, but requires knowledge of strata-specific variances S_h^2; the proportional allocation is a common choice.

Stratified sampling - estimation

The unbiased estimator of the total is the sum of the unbiased strata-specific estimators

$$\widehat{\tau}_{str} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{k \in \mathbf{s}} y_{k(h)} = \sum_{h=1}^{H} N_h \bar{y}_h$$

with variance the sum of the variances of the strata-specific estimators

$$V(\widehat{\tau}_{str}) = \sum_{h=1}^{H} N_h^2 \frac{1 - f_h}{n_h} S_h^2$$

estimated by

$$\widehat{V}(\widehat{\tau}_{str}) = \sum_{h=1}^{H} N_h^2 \frac{1 - f_h}{n_h} s_h^2.$$

Stratified sampling - interval estimation

$1-\alpha$ confidence intervals for the mean, total, and proportion can be derived in the usual way, using the normal approximation when the strata-specific sample sizes are large enough, and resorting to Student's t otherwise.

For the mean μ_y, $\bar{y}_{str} = \widehat{\tau}_{str}/N$ and $V(\bar{y}_{str}) = V(\widehat{\tau}_{str})/N^2$, hence

$$\bar{y}_{str} \pm z_{1-\alpha/2}\sqrt{\sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2 \frac{1-f_h}{n_h} S_h^2}.$$

For the proportion p_y, $\widehat{p}_{str} = \sum_{h=1}^{H} N_h \widehat{p}_h/N$ and $V(\widehat{p}_{str}) = \sum_{h=1}^{H} N_h^2 V(\widehat{p}_h)/N^2$, hence

$$\widehat{p}_{str} \pm z_{1-\alpha/2}\sqrt{\sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2 (1-f_h)\frac{\widehat{p}_h(1-\widehat{p}_h)}{n_h-1}}.$$

Cluster sampling

A cluster sample can be defined as a SRS of **primary sampling units** (PSU) that consist of groups (clusters) of **secondary sampling units** (SSU). Once the PSUs have been extracted, all the SSUs contained therein are collected and form the final sample.

Cluster sampling

A cluster sample can be defined as a SRS of **primary sampling units** (PSU) that consist of groups (clusters) of **secondary sampling units** (SSU). Once the PSUs have been extracted, all the SSUs contained therein are collected and form the final sample.

Let M denote the number of clusters, and m the size of a sample of clusters $\mathbf{g}$. Then, in analogy with SRS,

$$p(\mathbf{g}) = \binom{M}{m}^{-1}$$

Let i index a PSU and j a SSU. Since all units in a selected cluster enter the sample, it follows that

$$\pi_{ij} = \frac{m}{M}$$

$$\pi_{(ij)(i'k)} = \begin{cases} \frac{m}{M}, & i = i' \\ \frac{m(m-1)}{M(M-1)}, & i \neq i' \end{cases} , \forall \{(ij), (i'k)\}$$

Cluster sampling - point estimation

Let $N = \sum_{i=1}^{M} N_i$ denote the number of SSUs, where N_i is the number of SSU in each PSU; y_{ij} denotes the value for SSU j in PSU i, and $y_i = \sum_{j=1}^{N_i} y_{ij}$ is the total of the survey variable for PSU i.

The population total is $\tau = \sum_{i=1}^{M} \sum_{j=1}^{N_i} y_{ij} = \sum_{i=1}^{M} y_i$, while the population mean can be intended per PSU $\mu_{PSU} = \tau/M$, and per SSU $\mu_{SSU} = \tau/N$.

Cluster sampling - point estimation

Let $N = \sum_{i=1}^{M} N_i$ denote the number of SSUs, where N_i is the number of SSU in each PSU; y_{ij} denotes the value for SSU j in PSU i, and $y_i = \sum_{j=1}^{N_i} y_{ij}$ is the total of the survey variable for PSU i.

The population total is $\tau = \sum_{i=1}^{M} \sum_{j=1}^{N_i} y_{ij} = \sum_{i=1}^{M} y_i$, while the population mean can be intended per PSU $\mu_{PSU} = \tau/M$, and per SSU $\mu_{SSU} = \tau/N$.

When the PSUs are selected with SRS, the total can be estimated as

$$\widehat{\tau}_{cl} = \frac{M}{m} \sum_{i=1}^{m} y_i = M\bar{y}_{PSU}$$

with variance estimated by

$$\widehat{V}(\widehat{\tau}_{cl}) = M^2 \frac{1 - f_1}{m} s^2_{PSU}$$

where $f_1 = m/M$ and $s^2_{PSU} = (m-1)^{-1} \sum_{i=1}^{m} (y_i - \bar{y}_{PSU})^2$.

Cluster sampling - ratio estimator

If there is reason to believe that the PSU totals y_i are strongly correlated with the cluster sizes N_i, a more efficient approach may be the **ratio estimator**. For the total,

$$\widehat{\tau}_r = \frac{\sum_{i=1}^{m} y_i}{\sum_{i=1}^{m} N_i} N = rN$$

with variance estimated by

$$\widehat{V}(\widehat{\tau}_r) = \frac{M(M-m)}{m(m-1)} \sum_{i=1}^{m} (y_i - rN_i)^2.$$

Estimators for the mean and confidence intervals can be obtained as usual.

Cluster sampling - sample size

One of the most common uses of the *deff* is as a multiplicative coefficient by which to inflate or deflate the desired sample size under SRS, to provide the same level of accuracy under different designs.

If $deff < 1$, a smaller sample will be needed, compared to SRS, larger if $deff > 1$.

Consider once again the total; if the clusters have approximately the same size, the following relationship hold

$$deff_{cl} = 1 + \frac{M(N-1)}{M-1}\delta$$

where $\delta = 1 - S_W^2/S^2 \in \left[-\frac{M-1}{M(N-1)}, 1\right]$ is the **cluster homogeneity coefficient**.

Bibliography

Bibliography

- Chambers R and Clark R (2012). *An Introduction to Model-Based Survey Sampling with Applications*. Oxford Statistical Science Series 37.
- Cochran WG (1977). *Sampling Techniques - 3rd edition*. New York: Wiley.
- Conti PL and Marella D (2012). *Campionamento da popolazioni finite: il disegno campionario*. Collana di Statistica e Probabilità Applicata, Springer Verlag Mailand. doi:10.1007/978-88-470-2577-6.
- Horvitz DG and Thompson DJ (1952). *A generalization of sampling without replacement from a finite universe*. Journal of the American Statistical Association, 47, 663-685.

Bibliography

- Lavrakas PJ - editor (2008). *Encyclopedia of Survey Research Methods - volumes 1&2*. SAGE publications.
- Neyman J (1967). *On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection*. In A selection of early statistical papers of J. Neyman. Berkeley: University of California Press. (Reprinted from Journal of the Royal Statistical Society, 1934, 97[4], 558625).