SM-4331 Advanced Statistics
Chapter 2 (Survey Sampling)

Dr Haziq Jamil

FOS M1.09

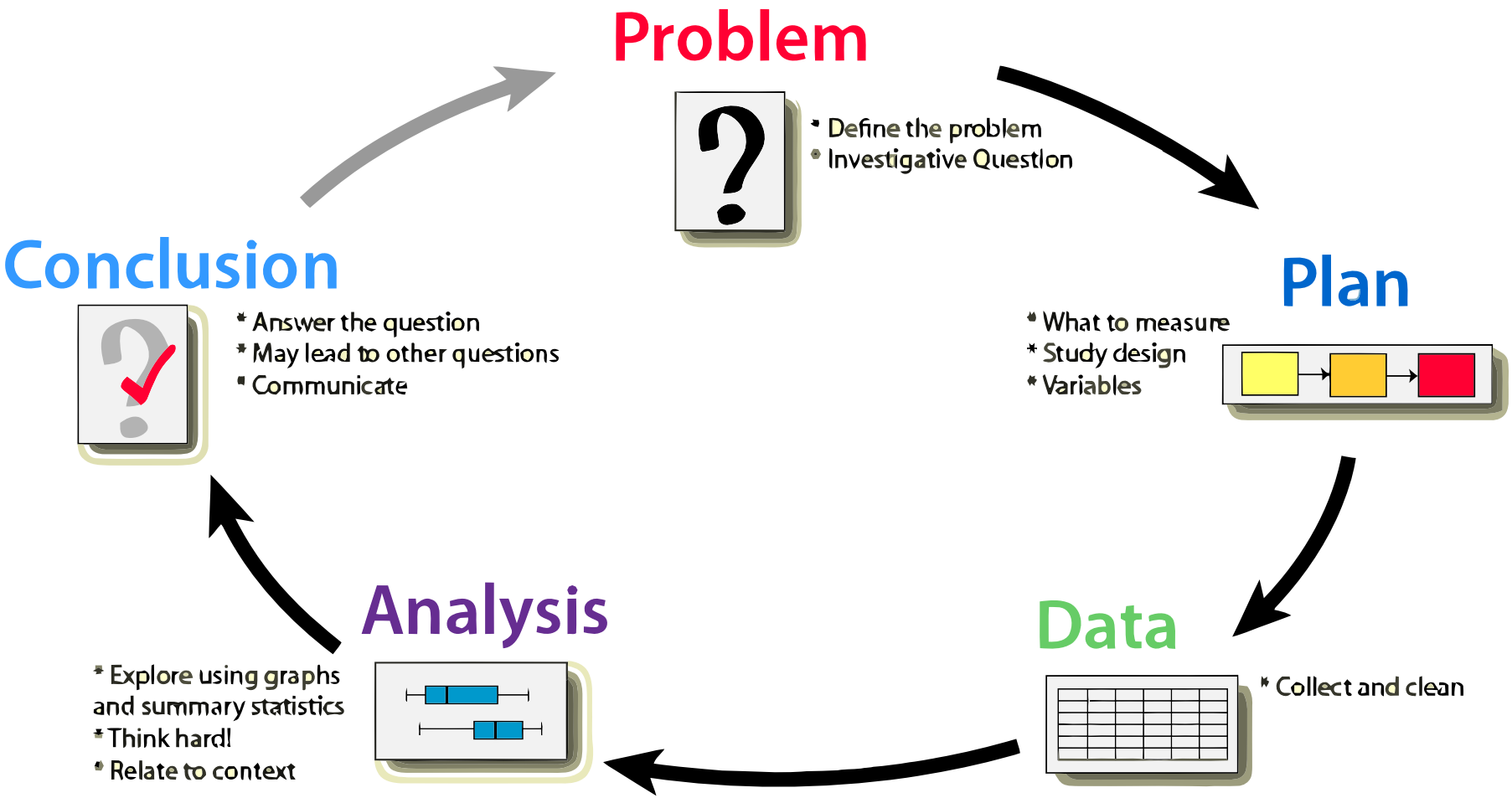
Universiti Brunei Darussalam

Semester II 2019/20

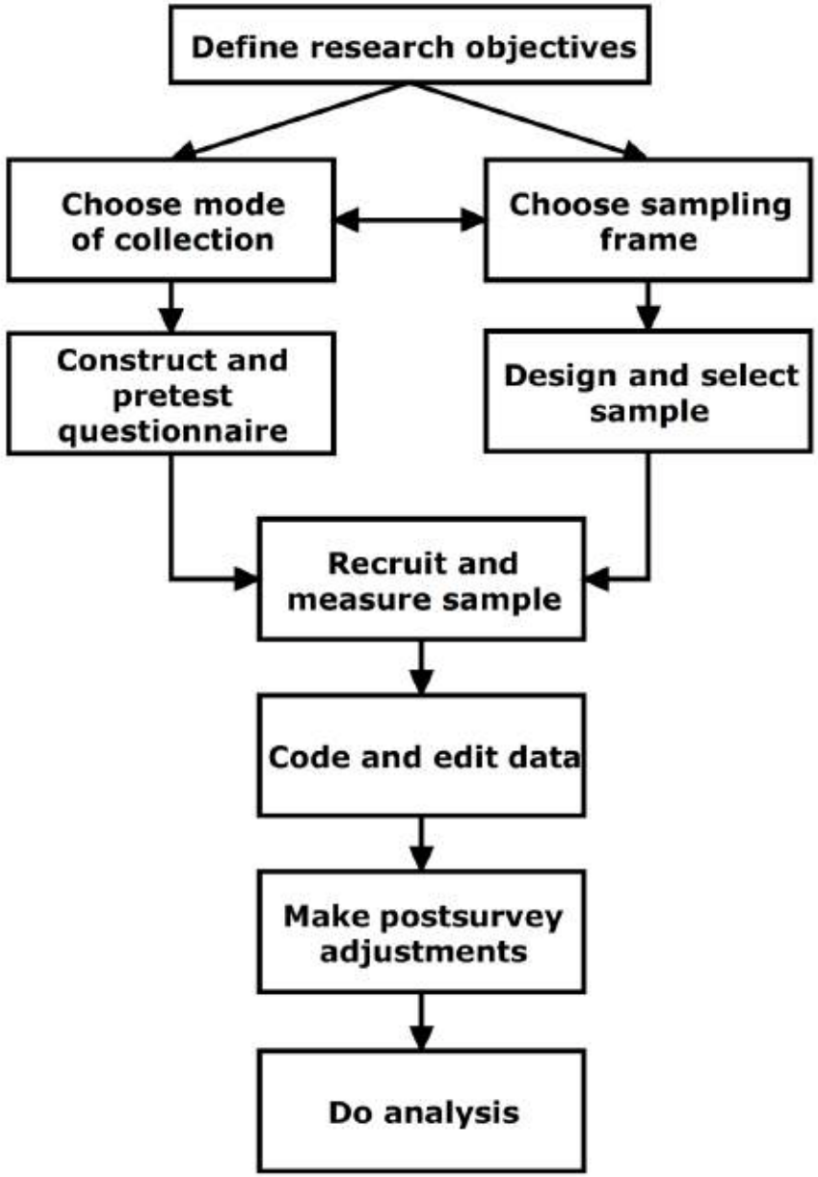
Outline

Outline

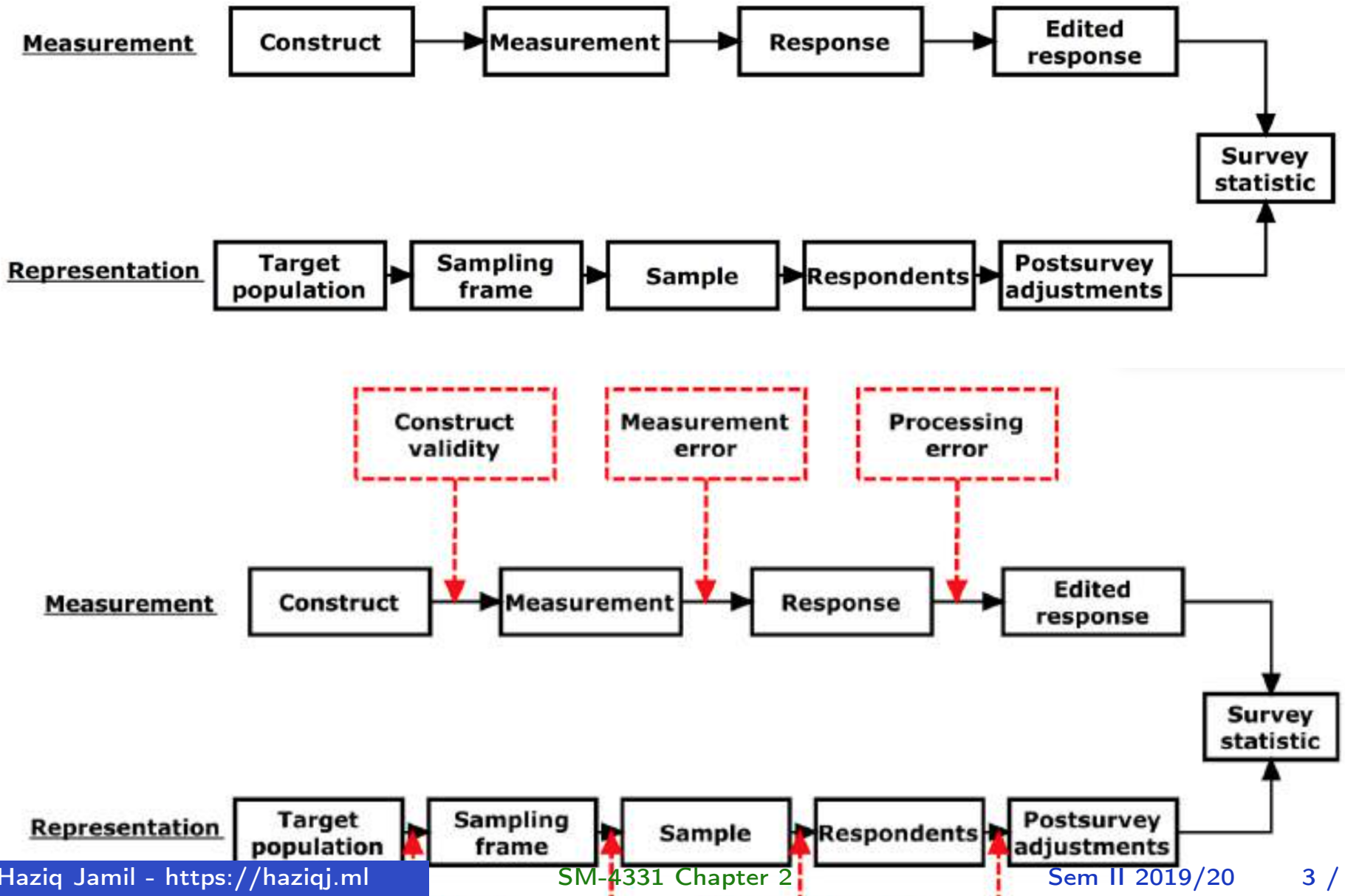
The statistical enquiry cycle



Survey research process



Survey process and potential errors



Sampling design and sampling process

Representation here concerns "how to obtain a sample which is representative of a target population".

For now, we shall take for granted other parts of the survey process, which are:

1. **Measurement**: aims, sources and quality of the variables to be studied have already been well thought of.
2. Non-response error and post-survey adjustments.

In this course, we will focus only on **coverage error** and **sampling error**, and how to minimise/avoid these.

Terminology

Definition 1 (Population)

A population U consists of N elements, i.e. $U = \{1, \ldots, N\}$. The population constitutes all of the elements of a set. It is the group to whom we intend to generalise the results of the study.

- Often, but not nearly always, the elements are people.
- For now, U is treated as a fixed set of elements, i.e. $N < \infty$ is finite.
- The population size N may be known or unknown.

We are interested in the values of variables y that would be obtained by administering survey questions to a respondent. Values in the population would be denoted $\mathbf{Y} = \{y_1, \ldots, y_N\}$.

Terminology (cont.)

Definition 2 (Sample)

A sample of size n is a subset of U of $n < N$ distinct elements from the population U.

- Values in the sample would be denoted (somewhat imprecisely) $\mathbf{y} = \{y_1, \ldots, y_n\}$–see remark below.
- Values of y_i in the sample are observed and thus known, while y_i for the rest of the population U (unsampled) are unknown.

Remark

This does not mean a one-to-one correspondence with the population, i.e. y_1 in the sample is not necessarily y_1 in the population, etc..

Sampling frames and sampling units

Definition 3 (Sampling frame)

The **sampling frame** is the materials or devices which delimit, identify and allow access to, the elements of the target population.

In other words, the sampling frame is the information we use to actually select a sample of elements and to get in touch with them. The simplest form of frame is a **list** of everyone/everything in the whole population, with contact details.

However, often sampling frames and sampling involve also lists of other kinds of **sampling units**, which differ from the elements. For instance, to sample individuals (elements in the sampling frame), sampling design calls for sampling of households (sampling unit).

Aim of sampling

The aim of our analysis is to ultimately draw conclusions about the value of some parameter $\theta = \theta(\mathbf{Y})$ of the distribution of y_i in the population U.

- Of course, if the entirety of the population set $\mathbf{Y}$ is known to use, we would use this data set.
- Almost always, population data is unavailable and we must rely on a sample instead.
- Advantages of performing sampling:
  - To enumerate the entire population is costly and time-consuming.
  - Less logistical burdens lead to better quality data.
  - With proper methodology, samples gives high accuracy.

Using the samples, we estimate θ using some estimator of θ, $\hat{\theta} = \hat{\theta}(\mathbf{y})$ say. As we know, a function of the samples e.g. $\hat{\theta}(\mathbf{y})$ is called a *statistic*.

Aim of sampling (cont.)

Therefore, given

- target population U
- target parameter θ
- available sampling frame(s); and
- other operational and cost constraints

select and use

- method of obtaining samples (the sampling design), and
- the form of estimator $\hat{\theta}$ (which may depend on the sampling design),

so that $\hat{\theta}$ used is a good estimator of θ.

Remark

In the Estimation Theory chapter, we discussed several desirable properties that an estimator should have. We will revisit some of these in the next section.

Example

Example 4

The US National Crime Victimisation Survey (NCVS): A 'rotating panel' survey collected by the US Census Bureau, on behalf of the US Bureau of Justice Statistics (`https://www.bjs.gov`). One purpose of this survey is to provide an alternative measure of the level of crime, to supplement and contrast with the numbers reported to the police (Groves et al., 2011, see).

- Population: US residents aged 12 or older ($N >$ 200 million).
- Sample: Interviewed $n \approx$78,600.
- Variable: $y_i =$ Person i has had something belonging to them stolen in the last 6 months.
- Parameter: $\theta =$ The proportion of the population who have had something belonging to them stolen in the last 6 months.

Example of sampling frame in the next subsection.

Three populations

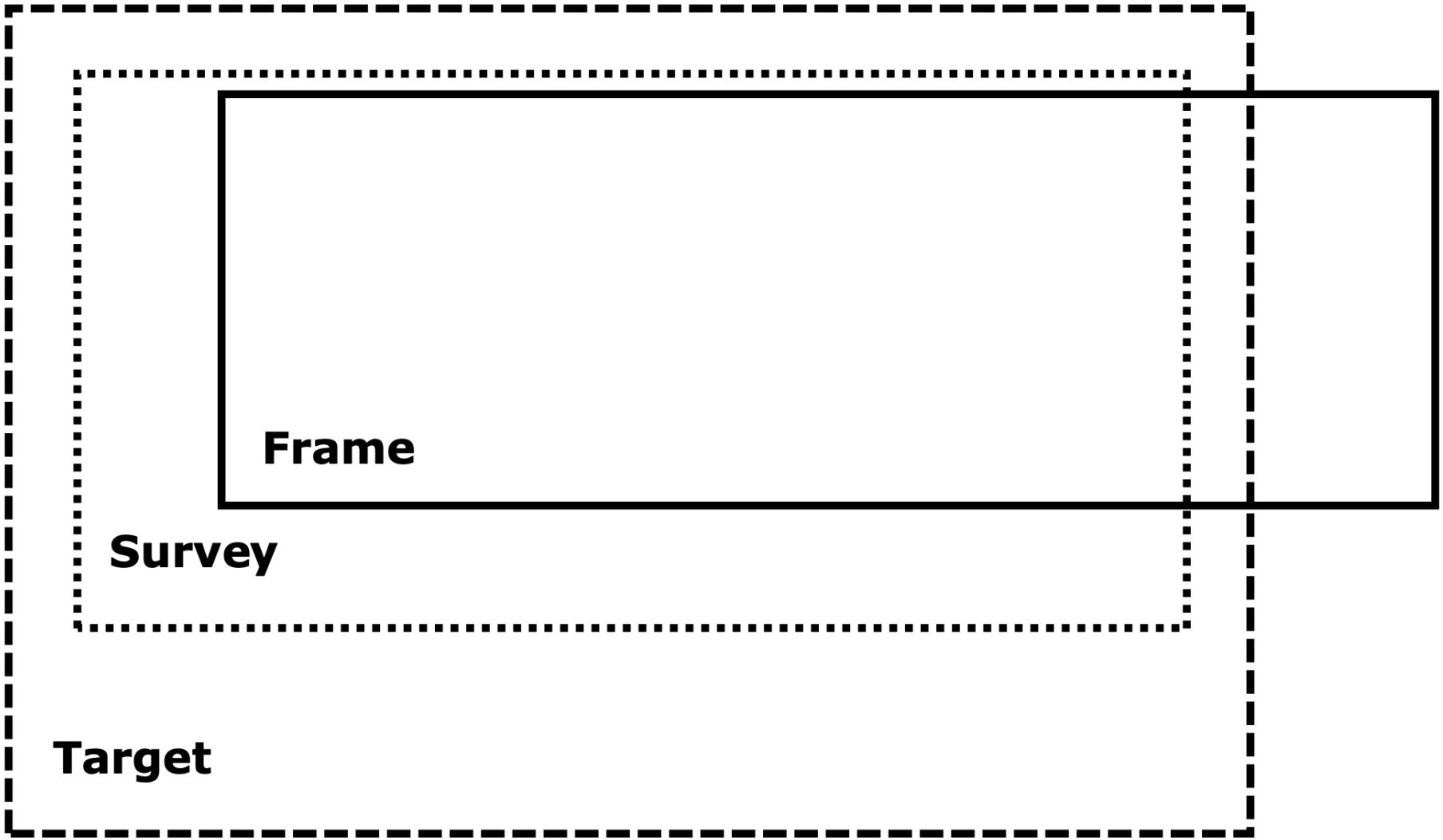
We wish to draw samples from the population U. However, the population we actually sample from is defined by the sample frame available to us.

Definition 5 (Target, survey and frame population)

- The **target population** is the actual population set that we are interested in.
- The **survey population** is the one we in practice aim to sample from.
- The **frame population** the one we actually sample from.

If the frame population is different from the target population, there is **coverage error**.

Three populations (cont.)



Target vs survey population

The **target population** is often stated very generally

- Restricted by geographical boundaries, age limits, and date only.
- Difficulties of definition mean target populations are never truly fixed.
- Most interesting analytical research questions are about such general populations, without further restrictions.

Example 6

Examples of target population:

- US residents aged 12 or over in [year].
- Individuals aged 15–64 in Brunei in [year].
- All registered postgraduate students at UBD at the start of S1 2019.

Target vs survey population (cont.)

The **survey population** essentially omits those in the target population that are difficult to be reached due to practical and cost reasons.

- The argument is that these omissions do not matter much, so that conclusions on the survey population is still relevant enough.
- The survey population effectively becomes the new target population.

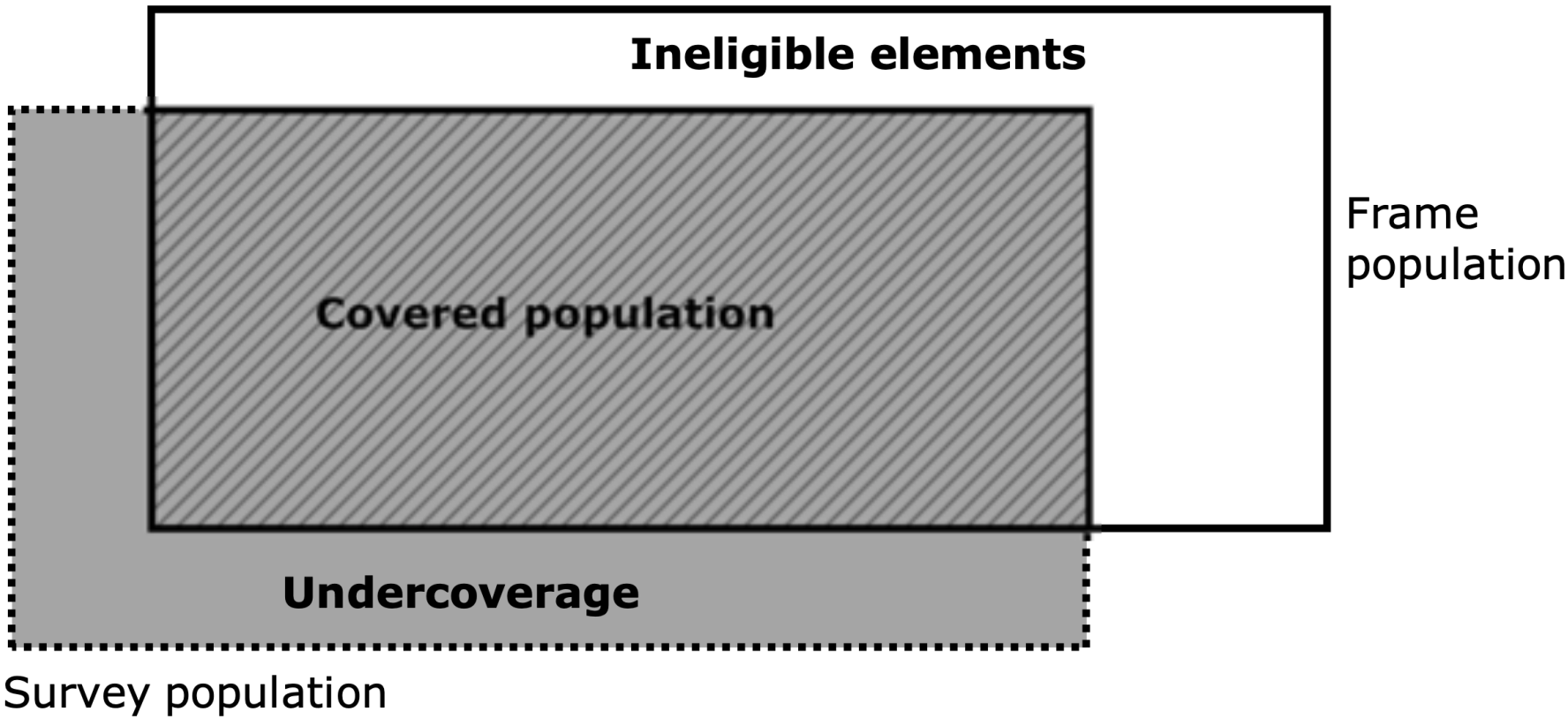
Example 7

Often, the survey population excludes those in remote or sparsely populated and/or dangerous areas. In the UK specifically, the area of Scotland "north of the Caledonian Canal" often omitted.

The frame population

The population of elements that have a chance of actually being included in the sample is defined by the **frame population**.

This "frame population" is often not identical with the survey population.



Coverage error

Coverage error causes **bias** in estimates (see Exercise sheet). The main source of coverage error is undercoverage:

Definition 8 (Undercoverage)

Undercoverage is caused by elements in the survey population which are not in the frame population, so cannot be sampled.

There are also two kinds of **overcoverage**:

- **Ineligible elements** which are included in the frame population but are not members of the survey population.
- **Duplicate elements** appearing several times in the frame population.

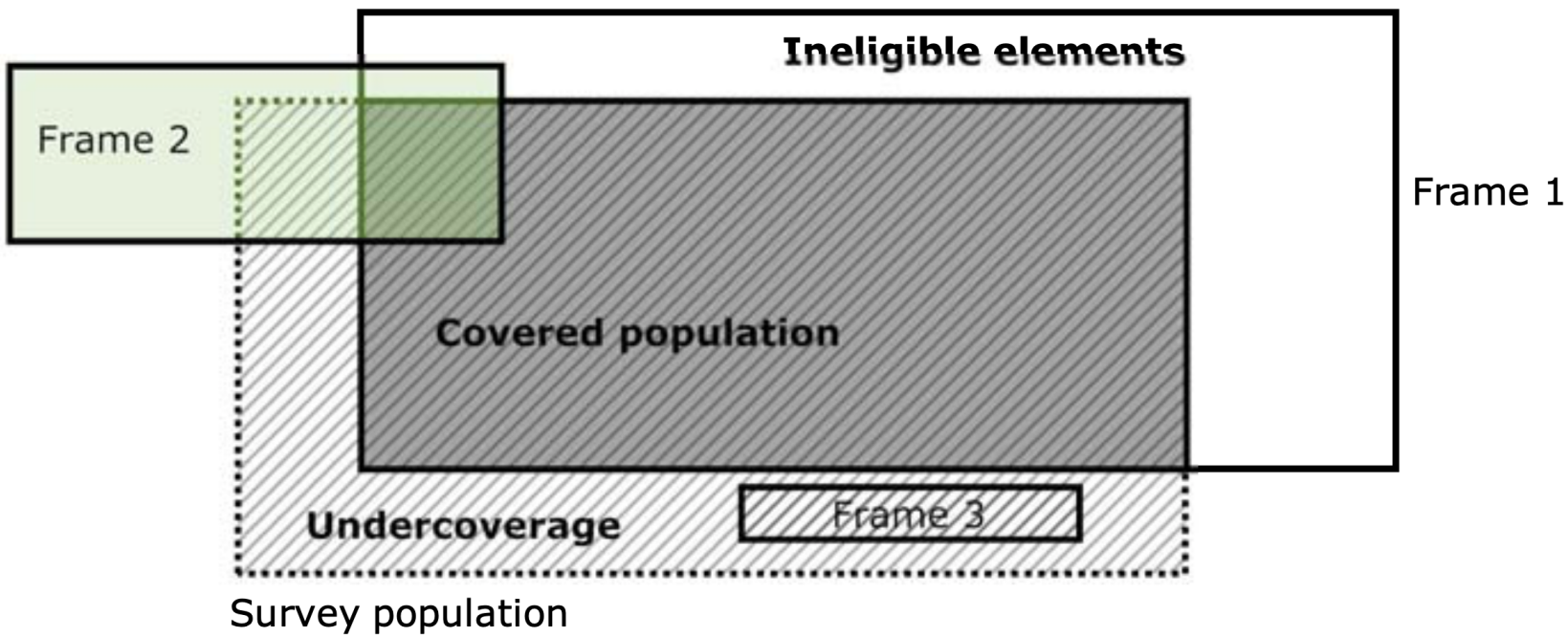
Remark

Ineligible elements can usually be identified once sampled and screened out; but if not, then it is a major problem. Duplicate elements are often harder to identify, but should be rare, unless the frame is faulty/poor.

Coverage error (cont.)

One of the ways to reduce (under)coverage error is to use **multiple frames**.

- If frames have no overlap, they are simply combined.
- If frame have overlap, the duplicate elements are removed first.



The ideal sampling frame

Definition 9 (List frame)

A **list frame** is a list of all elements of the target population.

This would be the most ideal and convenient frame to use for sampling, but it is not always available.

- In Scandinavian countries, population registers do exists.
- Recent country census data may be used, if possible.
- In the UK for example, this does not exist.
- Narrower list frames are often available, e.g. for organisations such as schools or businesses.

Multistage area sampling

The following approach is very widely used for general social surveys:

1. Sample areas;
2. Sample households (housing units, addresses) within sampled areas;
3. Sample individuals within sampled households.

If good enough frames are available, some of these steps could be skipped in frame construction

- If we had a list frame of all individuals, go straight to 3.
- If we had a list frame of all households, go straight to 2.

Remark

Regardless of list frame availability, a multistage design is often used for reasons of cost and efficiency of sampling. We will discuss this later in cluster sampling.

Multistage area sampling (cont.)

For multistage area sampling, we need at most three list frames:

1. **List frame of areas.** This is usually easily available (e.g. from Land Survey). Sometimes need to adopt multistage design within area sampling.
2. **List frame of households.** Two possibilities here:
   - Such a frame exists (e.g. In the UK, the Royal Mail's Postcode Address File (PAF) is used for major surveys).
   - Survey staff visit the sampled areas and construct an up-to-date list (map) of all housing units within each of them.
3. **List frame of individuals**. This can easily be made on the spot, if none exists.

Remark

If the target population is households, then we simply stop at stage 2. Stage 3 is replaced by selection (rather than sampling) of the reporting individual.

Example: Sampling frames in the NCVS

Example 10

A multistage area sampling for the NCVS was done. The list frames were as follows:

1. Two-stage for areas:
   (a) List of one or more adjacent **counties or metropolitan areas**.
   (b) List of (census enumeration) **districts**.
2. A multiple frame design (which are combined) for the household stage:
   (a) List of **housing units**.
   (b) List of **"group quarters"**.
   (c) List of **construction permits** for new housing.
   (d) List of **small areas** (where permit data is not available).
3. Individual stage: **All eligible individuals** within a household are interviewed.

1. Basic concepts and terminology
2. Sampling design for probability samples
3. Sampling designs in practice
4. Technical details of SRS

Sampling design

Definition 11 (Sampling design)

Let $u \subset U$ be a subset of elements in the population. Define the probabilities

$$p(u) = \mathrm{P}(\text{selected sample consists of the elements in } u) \geq 0$$

to the the **sampling design**.

We will always assume that the sample is selected non-deterministically. This allows us to leverage the laws of mathematics and probability when quantifying sampling error.

Probability sampling design

The sampling design implies also the inclusion probabilities:

Definition 12 (Inclusion probabilities)

The **inclusion probabilities** π_i for each individual element $i \in U$ in the population is defined to be

$$\pi_i = \text{P(element } i \text{ is selected into the sample)}$$

For a **probability sampling design** (i.e. random sampling),

$$\pi > 0 \text{ and } \underline{\text{known}} \text{ for every } i \in U.$$

In other words, in order to conduct random sampling, every element must have a known and non-zero probability of being included in the selected sample (though they need not be equal).

Recall: Estimation theory principles

Recall these principles:

- Elements in the sample $u \subset U$ are random. A different sample $u' \subset U$ could include different elements of the population.
- Similarly, sample values $\mathbf{y} = \{y_1, \ldots, y_n\} \subset \mathbf{Y} = \{y_1, \ldots, y_N\}$ may also vary from sample to sample.
- The value of an estimator $\hat{\theta} = \hat{\theta}(\mathbf{y})$ is a random variable and has a distribution (whatever this may be).
- The only source of randomness is the sampling design $p(u)$ (the values $y_1, \ldots, y_N$ are treated as fixed, though unknown).
- Randomness induces error in estimators; we should a "good" estimator to conduct inference.
- The mean squared error of an estimator is defined as

$$\text{MSE}(\hat{\theta}) = \text{Var}(\hat{\theta}) + \{\text{Bias}(\hat{\theta})\}^2.$$

- Small variance and small bias (i.e. small MSE) = good estimator.

Sampling error

If there no other errors (nonsampling errors) exist in the entire design process, then MSE($\hat{\theta}$) would only be due to **sampling error** only, i.e.

- Variability in $\hat{\theta}$ between different samples; and
- Systematic bias inherent in $\hat{\theta}$.

Often we can choose $\hat{\theta}$ that is unbiased (or at least approximately unbiased, i.e. $\mathrm{E}(\hat{\theta} - \theta) \approx 0$. This means

$$\mathrm{MSE}(\hat{\theta}) \approx \mathrm{Var}(\hat{\theta}),$$

so it remains that we should

- Quantify sampling variance, by estimating $\mathrm{Var}(\hat{\theta})$ accurately; and
- Minimise sampling variance, subject to cost constraints.

Remark

In essence, the aim of survey design is to obtain the maximum amount of information per amount of money spent. Nonsampling errors aside, a good sampling design helps achieve a desirable cost-MSE tradeoff.

Non-probability sampling

Some sampling designs are not probability sampling: quota, volunteer, convenience, purposive, snowball, or expert sampling–inclusion probabilites are not known!

- These are common in opinion polling and market research, but should never be used for major academic and governmental surveys.
- Impossible to evaluate bias, variance and MSE of estimators.
- Statements about properties of estimators require wishful thinking; and statistical models will have unverifiable assumptions.

Building block of sampling designs

Most probability sampling designs can be described as combinations of three basic elements:

1. Simple random sampling (SRS)
2. Cluster sampling
3. Stratified sampling

Here we consider only sampling **without replacement**, i.e. where a sampling unit may appear only once in the sample.

Simple random sampling (SRS)

Definition 13 (Simple random sampling)

SRS is a sampling design in which selecting n out of N possible elements from the population is done with equal probabilities.

SRS is an *equal probability selection method* (epsem).

For large surveys, full sampling design is very rarely SRS. Nonetheless, it is a useful starting point, for at least two reasons:

- It is a benchmark to which other designs can be compared.
- Individual stages of a more complex sampling design often use SRS, e.g. to sample areas from an area frame.

Simple random sampling (SRS) (cont.)

Estimator for the mean (SRS)

The sample mean

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad y_i \in \mathbf{y}$$

is an unbiased estimator of the population mean

$$\mu = \frac{1}{N}\sum_{i=1}^{N} y_i, \quad y_i \in \mathbf{Y}.$$

The variance of this estimator is

$$\text{Var}(\bar{y}) = \frac{N-n}{Nn} \cdot \overbrace{\frac{1}{N-1}\sum_{i=1}^{N}(y_i-\mu)^2}^{S^2} = \frac{N-n}{N} \times \frac{S^2}{n}.$$

Simple random sampling (SRS)

In practice we do not know the values of μ nor S^2, therefore we have to come up with an estimate of the variance.

Estimator for the variance of the mean (SRS)

$$\widehat{\text{Var}}(\bar{y}) = \frac{N-n}{Nn} \cdot \overbrace{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}^{s^2} = \frac{N-n}{N} \times \frac{s^2}{n}$$

is an unbiased estimated of $\text{Var}(\bar{y})$.

Simple random sampling (SRS)

- The factor $\frac{N-n}{N} =: \phi$ is called the **finite population correction (FPC)**. It is responsible for changing the variance of $\bar{y}$ when the sample is drawn from a finite population in comparison to an infinite population.
- The factor n/N is called the **sampling fraction**. Realise that $\phi = 1 - n/N$.
- Note that $\phi \to 1$ as $N \to \infty$. In practice, the FPC can be ignored if N is large enough (or as a rule of thumb, if $n/N < 0.05$).
- However, ignoring the ϕ will result in the overestimation of $\text{Var}(\bar{y})$.

Cluster sampling

Suppose elements in a population are grouped in space into natural, known clusters (e.g. people in families, people in neighbourhoods, employees in companies, etc.).

(The most basic version of) cluster sampling:

- Sample the **clusters** with SRS.
- Interview all eligible elements in a cluster.

Other generalisations of basic cluster sampling:

- **Subsampling**: Do not include every element in a sampled cluster, but only a sample of these elements.
- **Multistage sampling**: A sequence of sampling stages of nested clusters (e.g. Big area – Small area – Household – Individual).

Samples can be drawn using SRS or others (more on this later).

Cluster sampling (cont.)

Strategies for "choosing" clusters:

- Groupings are mutually homogenous (i.e. no apparent "difference" in the groups). E.g. area sampling or geographical cluster sampling–Wouldn't mind samples coming from whichever district that we end up sampling from.
- Members in the clusters should be as heterogenous as possible (i.e. should be as "different" as possible from each other internally in the group). Each cluster should be a small-scale representation of the entire population.
- Clusters should be mutually exclusive and collectively exhaustive: Each member belongs to one and only one cluster.

Cluster sampling (cont.)

Suppose that

- There are M clusters, with N_j units in each cluster $j = 1, \ldots, M$.
- We sample m of the M clusters, so that $n = \sum_{j=1}^{m} N_j$.

Estimator for the mean (Cluster sampling)

The sample mean

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i, \quad y_i \in \mathbf{y}$$

is an unbiased estimator of the population mean μ. The variance of this estimator is

$$\text{Var}_{cl}(\bar{y}) = \frac{M-m}{Mm} \cdot \overbrace{\frac{1}{M-1} \sum_{j=1}^{M} (\mu_j - \mu)^2}^{S_{cl}^2} = \frac{M-m}{M} \times \frac{S_{cl}^2}{m}.$$

Cluster sampling (cont.)

- $\mu_j = N_j^{-1} \sum_{y_i \in j} y_i$, $j = 1, \ldots, M$ are called the **cluster means**. They relate to the population mean by the formula

$$\mu = \frac{1}{N} \sum_{j=1}^{M} N_j \mu_j.$$

  Cluster means are generally not known, but can be estimated (only for the sampled clusters) by $\bar{y}_j = n_j^{-1} \sum_{y_i \in j} y_i$.

- S_{cl}^2 are known as the **variance of the within-cluster means**. It can be estimated using

$$s_{cl}^2 = \frac{1}{m-1} \sum_{j=1}^{m} (\mu_j - \bar{y})^2$$

Cluster sampling (cont.)

As before none of μ_j (at least not all of them), μ or S^2_{cl} are known.

Estimator for the variance of the mean (Cluster)

$$\widehat{\text{Var}}_{cl}(\bar{y}) = \frac{M-m}{Mm} \cdot \overbrace{\frac{1}{m-1}\sum_{j=1}^{m}(\mu_j - \bar{y})^2}^{s^2_{cl}} = \frac{M-m}{M} \times \frac{s^2_{cl}}{m}$$

is an unbiased estimator of $\text{Var}_{cl}(\bar{y})$.

Multistage sampling with unequal cluster sizes

Cluster sampling is easiest if all clusters are of the same size. However, this is rarely (if ever) the case.

Example 14

Consider a population of $N = \sum_{j=1}^{M} N_j = 1000$ elements in $M = 10$ clusters of sizes N_j:

Cluster j	1	2	3	4	5	6	7	8	9	10
Size N_j	25	50	50	75	75	100	125	150	150	200

Suppose we want to draw a cluster sample with $m = 2$ clusters, total sample size of $n = 40$, and/or equal inclusion probabilities. How is this achieved?

Multistage sampling with unequal cluster sizes (cont.)

Example 14

Option 1: Standard cluster sampling $m = 2$ using SRS.

- The probability of selection for each cluster is $2/10$.
- If all elements in each cluster are taken, the probability of inclusion for elements are also $\pi_i = 2/10$.
- If subsample elements in each cluster, decide on a constant fraction of units (e.g. $1/5$). Then $\pi_i = 2/10 \times 1/5 = 0.04$.
- However, the realised sample size n is random (it depends on the cluster size).
- Random sample size is not good, as it increases the variance of estimators, and adds logistical difficulty to data collection.

Multistage sampling with unequal cluster sizes (cont.)

Example 14

Option 2: Standard cluster sampling $m = 2$ using SRS, fixed 20 units per cluster sampled.

- The probability of selection for each cluster is $2/10$.
- Since we only subsample 20 per cluster, sample size is now fixed $n = 2 \times 20 = 40$.
- However, $\pi_i = 2/10 \times 40/N_j = 8/N_j$ (non-constant) and depends on cluster size–e.g. $\pi_i = 0.16$ for elements in Cluster 1, but $\pi_i = 0.02$ for elements in Cluster 10.
- Variable inclusion probabilities can be allowed for in estimation by weighting, but usually increases variability of estimators. Best to avoid if possible.

Multistage sampling with unequal cluster sizes (cont.)

Example 14

Option 3: Standard cluster sampling $m = 2$ with unequal probabilities, fixed 20 units per cluster sampled.

- For clusters, set inclusion **probabilities proportionate to size (PPS)**, where size is N_j.
- Sample size is now constant $n = 2 \times 20 = 40$.
- Inclusion probabilities are also constant:
  - Inclusion probability for each cluster $j \in \{1, \ldots, M\}$ is $m(N_j/N) = 2(N_j/1000)$.
  - Inclusion probability of an element in cluster j, if this cluster is sampled, is $20/N_j$.
  - Thus, overall inclusion probability for each unit is

$$\pi_i = m(N_j/N) \times 20/N_j = 2(N_j/1000) \times 20/N_j = 0.04$$

Why cluster sampling?

- **Cost reduction and increased sampling efficiency**. Think about area sampling–greatly reduce logistical costs because do not have to cover the entire country, for e.g.
- **Reduction in the total number of interviews/responses collected**, thus reducing potential errors.
- **The population is naturally divided into clusters**, so take advantage of this fact. For a fixed sample size, the expected random error is smaller when most of the variation in the population is present internally within the clusters, and not between the clusters.

Strata

Definition 15 (Strata)

Strata (plural of stratum) refer to collectively exhaustive and mutually exclusive subgroups of the population, i.e. strata should define a partition of the population. Every element in the population must be assigned to one and only one stratum.

Key differences between strata and clusters/groups/domains:

- Strata need not be substantively interesting subgroups which we would consider in eventual analysis.
- Strata **must** be identifiable from the sampling frame, e.g. stratification of UBD students by faculty requires knowing which faculty each student belongs to before sampling.
- Strata ideally should be (strongly) correlated with variable of interest y, e.g. political survey would want to reflect the diversity of the population, thus stratify by ethnicity/race/religion.

Stratified sampling

Strategies for "choosing" strata:

- When we realise that subpopulations within an overall population vary, it could be advantageous to sample each subpopulation independently.
- Members in the strata should be as homogenous as possible (unlike clusters).
- And all the things mentioned previously: strata identifiable in frame (not sample), correlation between strata and variable of interest, etc.

The basic version of stratified sampling:

- (After identifying strata), simply sample elements from **all strata**, and combine the sample.
- Note: the sampling designs can be different in each strata.

Remark

In cluster sampling, we sample clusters and take all (or subsample from) the elements in the clusters sampled. In stratified sampling, we sample from **all of the strata**.

Stratified sampling (cont.)

Suppose that

- Population of $N = \sum_{h=1}^{H} N_h$ elements is divided into H strata of sizes N_h, $h = 1, \ldots, H$.
- Sample separately and independently n_h elements **from each stratum** $h = 1 \ldots, H$.

Estimator for the mean (Stratified sampling)

An unbiased estimator for the population mean μ is given by

$$\bar{y}_{st} = \sum_{h=1}^{H} \omega_h \bar{y}_h,$$

where $\omega_h = N_h/N$ and $\bar{y}_h$ is the sample mean within stratum h.

Stratified sampling (cont.)

Variance of the estimator for the mean (Stratified sampling)

If sampling within strata is epsem, then

$$\text{Var}(\bar{y}_{st}) = \sum_{h=1}^{H} \left(\frac{N_h - n_h}{n_h N_h} \right) \cdot \overbrace{\frac{1}{N_h - 1} \sum_{i=1}^{N_h} (y_i - \mu_h)^2}^{S_h^2} = \sum_{h=1}^{H} \left(\frac{N_h - n_h}{N_h} \times \frac{S_h^2}{n_h} \right),$$

where S_h^2 and μ_h are the (population) variance and (population) mean within stratum h, respectively.

Stratified sampling (cont.)

As with cluster sampling, the within stratum mean and variance are calculated in the same way:

- $\mu_h = N_h^{-1} \sum_{y_i \in h} y_i$, $h = 1, \ldots, H$, and

$$\mu = \frac{1}{N} \sum_{h=1}^{H} N_h \mu_h.$$

  Within strata means μ_h are generally not known, but can be estimated by $\bar{y}_h = n_h^{-1} \sum_{y_i \in h} y_i$

- An estimator for the within stratum variance is given by

$$s_h^2 = \frac{1}{n_h - 1} \sum_{j=1}^{n_h} (y_i - \bar{y}_h)^2.$$

Stratified sampling (cont.)

As before, none of μ_h or S_h^2 are known.

Estimator for the variance of the mean (Stratified sampling)

$$\widehat{\text{Var}}(\bar{y}_{st}) = \sum_{h=1}^{H} \left(\frac{N_h - n_h}{n_h N_h} \right) \cdot \overbrace{\frac{1}{n_h - 1} \sum_{i=1}^{n_h} (y_i - \bar{y}_h)^2}^{s_h^2} = \sum_{h=1}^{H} \left(\frac{N_h - n_h}{N_h} \times \frac{s_h^2}{n_h} \right),$$

is an unbiased estimator of $\text{Var}(\bar{y}_{st})$.

Proportionate and disproportionate allocation to strata

How many samples should we collect from each stratum?

Proportionate allocation

Suppose we set $n_h = nN_h/N = n\omega_h$.

- Then $n_h/N_h = n/N$ and $n_h/n = N_h/N$, i.e. proportions of strata in the sample match the proportions in the population.
- If sampling is epsem within strata, then the sampling is epsem overall.

Disproportionate allocation

Suppose the sampling fractions n_h/N_h are not the same in all strata

- Then for some strata their proportions n_h/n in the sample are larger than their proportions N_h/N in the population–some strata are oversampled.
- Overall inclusion probabilities π_i will not be equal for all elements, so must be weighted in analysis.

Example

Example 16

Consider "social class" as a possible grouping, and we are interested in understanding "attitudes towards a welfare state".

Social class as a domain of interest:

- Identify each sampled respondent's social class (say, based on occupation); then
- Estimate average attitudes separately for different social classes (how do attitudes vary in different social classes?).

Example (cont.)

Example 16

Social class as a stratum:

- How to identify social class in the frame? Don't exactly have a universal social class indicator for individuals!
- Use variables that are correlated with class–e.g. area, stratified by percentage of owner-occupied households in last census. The idea is that areas of high home ownership is associated with high social class areas.
- Stratification in this manner is likely to include respondents with an appropriate distribution of respondents from different social classes.

Example (cont.)

Example 16

Consider a population of $N = \sum_{h=1}^{H} N_h = 1000$ elements in $H = 4$ strata of sizes N_h:

% home ownership in strata h	$< 10\%$	$10 - 50\%$	$50 - 90\%$	$> 90\%$
Size N_h	50	150	400	400

A sample of $n = 100$ proportionate to strata would like like this:

% home ownership in strata h	$< 10\%$	$10 - 50\%$	$50 - 90\%$	$> 90\%$
Size n_h	5	15	40	40

Example (cont.)

Example 16

Whereas a sample of size $n = 100$ disproportionate to strata size would look like

% home ownership in strata h	$< 10\%$	$10 - 50\%$	$50 - 90\%$	$> 90\%$
Size n_h	25	25	25	25

Here, a constant sample size of 25 was taken in each stratum. But this means that the low ownership people are oversampled, whereas the high ownership people are under-represented.

This sample is fine, but any statistic that is calculated must be weighted accordingly, i.e. weights to reduce the responses from low ownership strata, and inflate responses from high ownership strata.

Why stratified sampling?

- **Improved efficiency**. By fixing the proportions of strata in the sample, stratified sampling removes part of sampling variation in estimators.
- **Control of within-strata sample sizes** n_h.
  - Proportionate stratification fixes sample proportions of strata to be equal to population propotions.
  - Disproportionate stratification fixes the sample proportions to be different from population proportions, in whatever way we desire–for e.g., it can be used to try to make sample sizes in small but interesting groups (e.g. ethnic minorities) large enough.
- **Reduce costs**. Measurements become more manageable and/or cheaper when the population is grouped into strata.
- **Obtain between-strata statistics**. Often it is desirable to have estimates of population parameters for groups within the population.

How to draw a random sample

We've been talking a lot about "drawing a random sample"–but how do we actually achieve this? Several ways:

- Using random number tables
- Using computers
- Systematic sampling
- Random routes
- At the doorstep

Random number tables

In old days, random number tables such as the one below are used.

61424	20419	86546	00517
90222	27993	04952	66762
50349	71146	97668	86523
85676	10005	08216	25906
02429	19761	15370	43882
90519	61988	40164	15815
20631	88967	19660	89624
89990	78733	16447	27932

Random number tables (cont.)

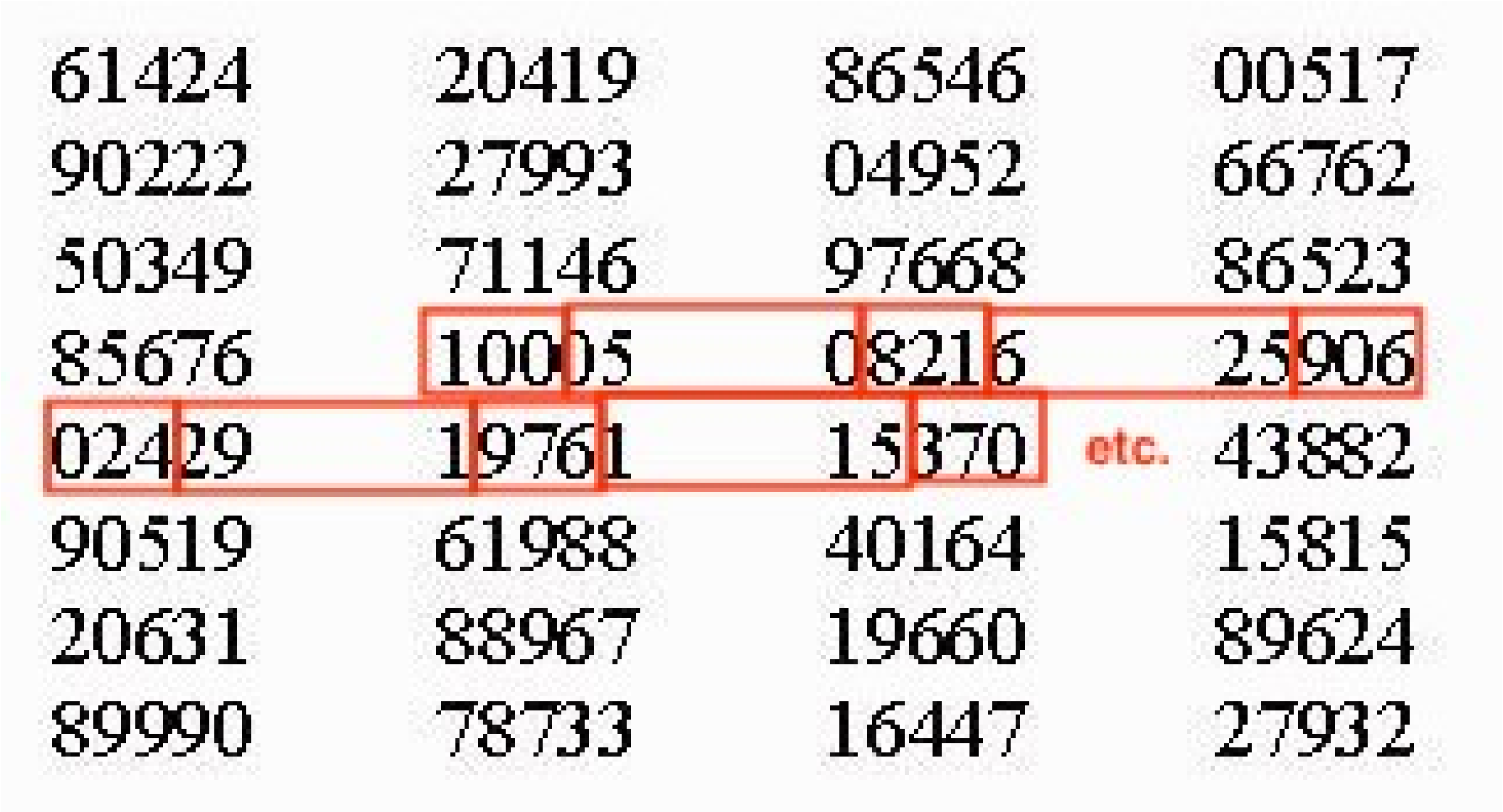
Example 17

Suppose we want to randomly sample from a population of size $N = 500$. Before we begin sampling, we take the sampling frame and label all individuals with a digit from 001 up to 500. Then,

- Select a random starting point ("seed").
- Read the digits horizontally 3-digits at a time, and continue on the next line.
- If the 3 digits form a number greater than 500, this is deleted.
- These digits will then represent the identifiers of the randomly selected individuals.

Random number tables (cont.)

E.g. 100, 50, ~~821~~, ~~625~~, ~~906~~, 24, 291, ~~976~~, 115, 370, etc.



(Pseudo-)Random number generators (RNG)

Nowadays, one can simply generate a sequence of random numbers (with equal probabilities or otherwise) using computers in software.

- In R, use the command

```
sample(500, size = 20, replace = FALSE, prob = NULL)
```

```
##  [1]  87 432  97 293   8 311 469 256 498  28
## [11] 291 317 323 260 493 227  41  31 298 413
```

- The argument `prob = NULL` implies epsem.
- R can also perform stratified sampling with the add-on package `sampling`.

Remark

In the field, computers are hardly ever brought, and therefore the if computers are used, the numbers are generated at the planning stage (before going into the field).

Systematic sampling

This method is a random sampling technique which minimises the need for random number generation.

Example 18

Suppose we want to randomly sample $n = 20$ units from a population of size $N = 500$. Arrange frame as an ordered list, with each unit numbered $1, 2, \ldots, 500$.

- Let $k = \lfloor N/n \rfloor$, e.g. $k = \lfloor 500/20 \rfloor = 25$.
- Select one number from $1, \ldots, N$ at random using random tables or a RNG–say this is 91.
- Sample this unit and every k-th unit after it, looping back to the beginning at the end of the list, until n sampled.

We get 91, 116, 141, 166, 191, 216, 241, 266, 291, 316, 341, 366, 391, 416, 441, 466, 491, 41, 66

Systematic sampling (cont.)

Systematic sampling effectively produces a simple random sample from a single random number.

Caution

Use systematic sampling only when the order of the list is uncorrelated with the variable of interest. That is to say, the list should not be ordered in any particular way.

To illustrate that systematic skip concealing a pattern, suppose we were to sample a planned neighbourhood where each street has ten houses on each block. This places houses no. $1, 10, 11, 20, 21, 30, \ldots$ on block corners. If we end up doing a systematic sample of every 10th household, then our sample will consist only of corner houses entirely, or non-corner houses entirely (depending on our starting point). This is not a representative sample at all.

Random routes

A method used for addresses-within-areas stage of multistage sampling. Often times, constructing a full frame of addresses is difficult.

Example 19

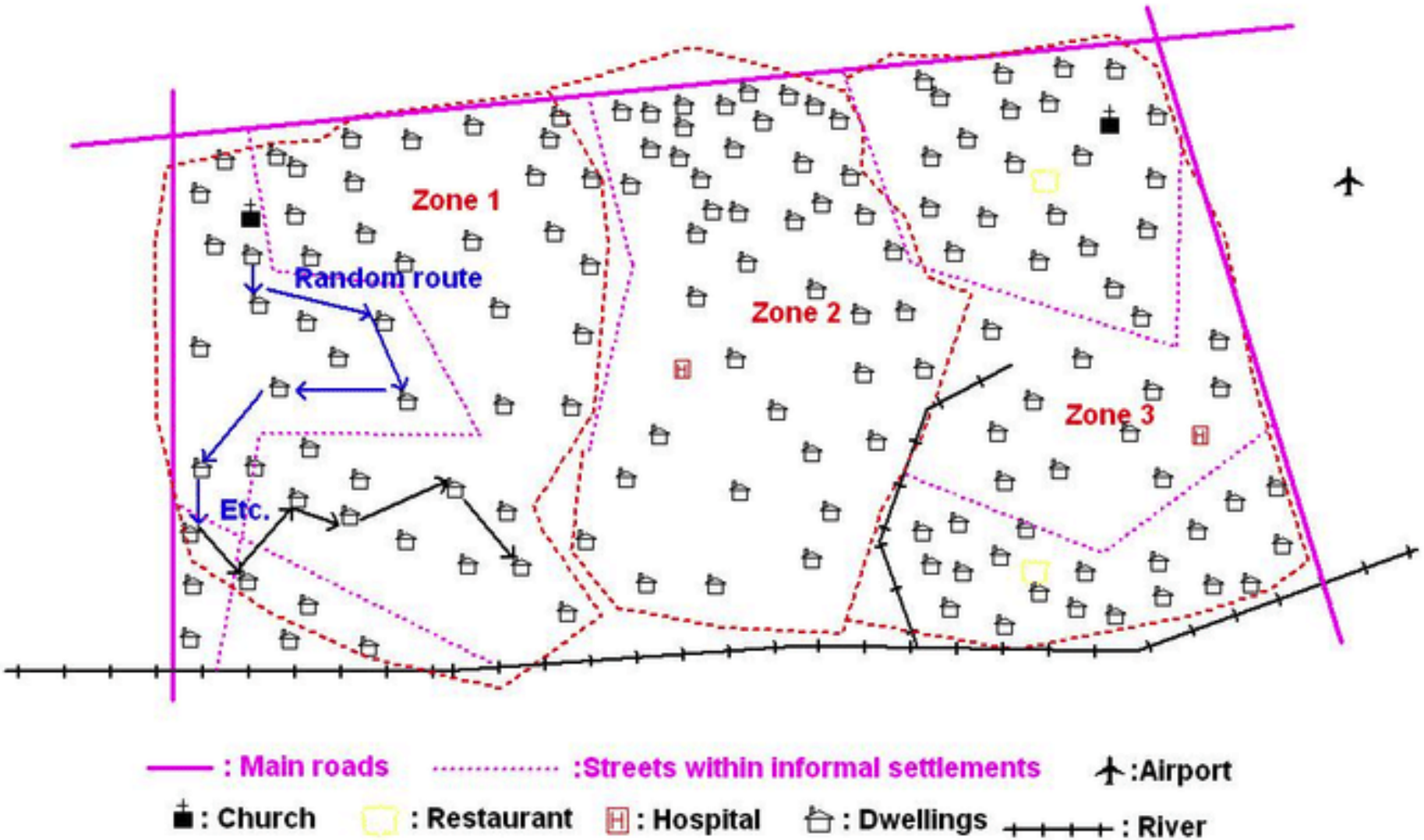
Suppose we (roughly) have delimited an area in which to sample households, but for one reason or another, we are unable to construct a full frame of addresses.

- Select a starting point–e.g. an address at random or a prominent landmark.
- Interviewer proceeds through the area according to a fixed set of rules–e.g. always turn left if possible within the area boundary.
- Sample addresses at fixed intervals–e.g. every 6th on your left.

This method is similar to systematic sampling, but not as unambiguous or easily controlled. Not ideal, but a pragmatic design for situations where alternatives are difficult.

Random routes

Rules are a bit flexible, as long as samples are selected "random"-ly enough.



At the doorstep

Recall that in multistage area sampling, last stage is often sampling of an individual from a household, using a list frame constructed there and then.

Instead of a computer, a procedure like this is often used:

- Assign a number to each household–e.g. by order of interview
- Number eligible individuals in a household in a predetermined order–e.g. by age and sex
- Use a pre-assigned table of random numbers (a "Kish grid") to find the number of the individual who should be interviewed. This depends on the household number and the number of eligible persons

Remark

In modern times, RNGs can be drawn even from a smartphone or tablet, so the Kish grid is less often used. In fact it may even be problematic: respondents may be unwilling to give information (list of names, age, sex, etc.) of everyone in their household.

Kish grid (Leslie Kish, 1949)

	Eligible People							
Household	1	2	3	4	5	6	7	8+
1st	1	1	1	1	1	1	1	1
2nd	1	2	2	2	2	2	2	2
3rd	1	1	3	3	3	3	3	3
4th	1	2	1	4	4	4	4	4
5th	1	1	2	1	5	5	5	5
6th	1	2	3	2	1	6	6	6
7th	1	1	1	3	2	1	7	7
8th	1	2	2	4	3	2	1	8
9th	1	1	3	1	4	3	2	1
10th	1	2	1	2	5	4	3	2

Complex sampling designs

Most large general surveys with face-to-face data collection use a complex sampling design which combines the elements discussed above.

Often the design is a stratified multistage design, with

- First stage a stratified sample of areas (clusters). These first-stage clusters are the primary sampling units.
- Perhaps some more stages of cluster sampling.
- Sampling of households.
- Sampling of individuals, if target elements are individuals (or selection of reporting individual, if elements are households).

JPKE Household Expenditure Survey

`http://pmo.gov.bn/SitePages/Household%20Expenditure%20Survey%20(HES).aspx`

The JPKE Household Expenditure Survey (HES) collects "the latest and comprehensive information on the expenditure patterns and income levels of households in Brunei Darussalam".

- Target population: Households in Brunei in [year].
- Survey population: Private households in Brunei in [year], so excludes army barracks, school hostels, commercial dwellings, etc.
- Frame population: Those of the survey population whose addresses are listed in the Census Enumeration Block.

JPKE Household Expenditure Survey (cont.)

The sampling conducted by JPKE is a multistage design

- **Stage 1**: Stratified sample of "sub-section" clusters, PPS with respect to number of households in each sub-section.
- **Stage 2**: SRS of households within each sub-section.
- **Stage 3**: Select reporting individual in the household.

Sample size is $n \approx 3,200$ (out of $N \approx 63,800$) households using this scheme.

Stratification of areas in Brunei

JPKE (usually) stratifies the areas in Brunei by **district** and **rurality**. Brunei is therefore divided into eight strata:

1. Brunei-Muara, urban
2. Brunei-Muara, rural
3. Belait, urban
4. Belait, rural
5. Tutong, urban
6. Tutong, rural

Recall

Seek heterogeneity between strata, and homogeneity within strata.

Clustering areas in Brunei

- Districts in Brunei are primarily subdivided into **mukims**.
- Each mukim is further subdivided into **kampongs**.
- For census work in Brunei, these kampongs are divided into segments, and each segment into sub-segments.

Therefore, these sub-segments (or so-called census enumeration blocks) are the primary sampling units in the first sampling stage.

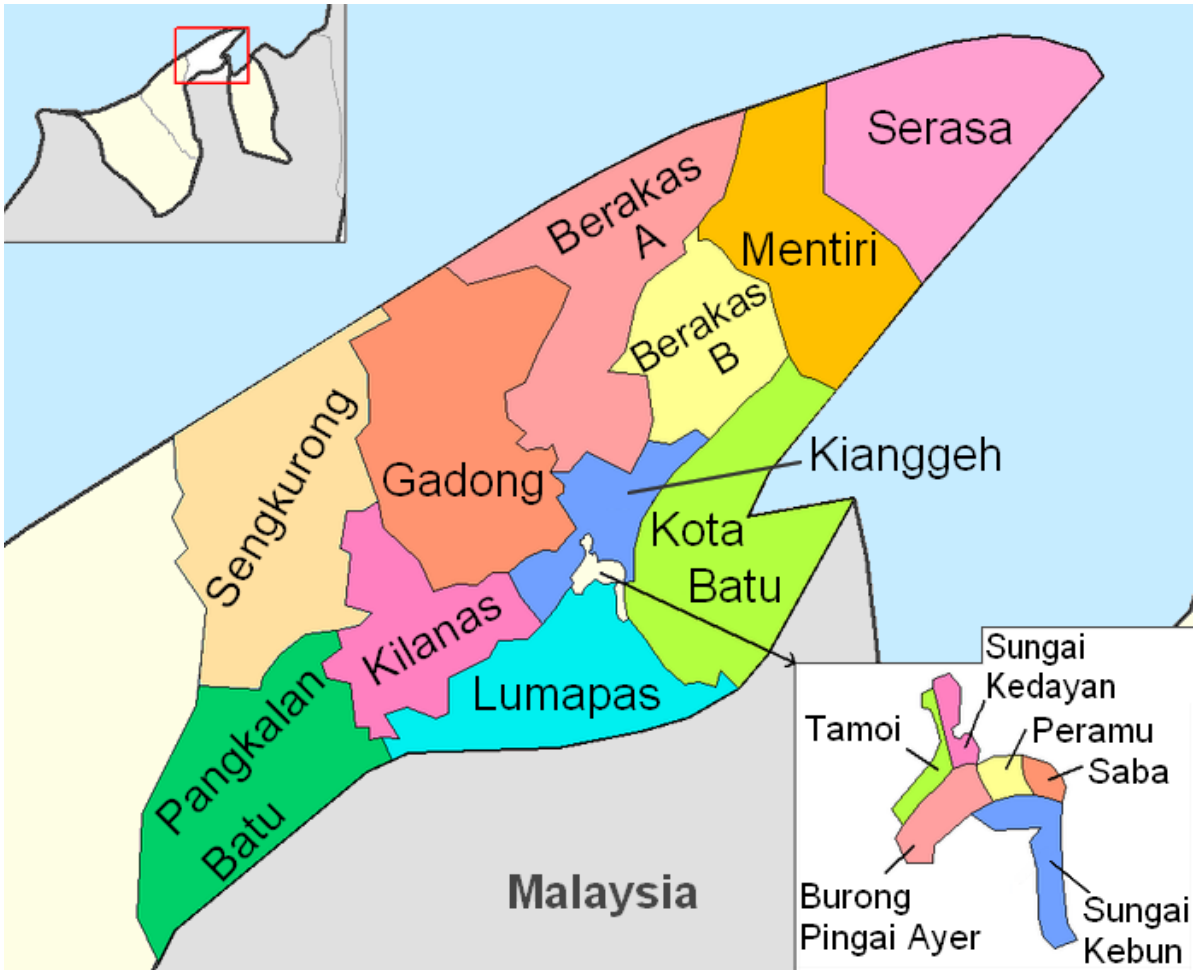
Each sub-segment contains roughly 60 households (as much as possible). This is by design, so as to reduce impact of clustering on variance of estimators and survey costs.

Recall

Seek heterogeneity within clusters.

Clustering areas in Brunei (cont.)

Brunei-Muara » Mukim Gadong A » Kg. Rimba & STKRJ Rimba » Segment » Sub-segment



KG. RIMBA & STKRJ RIMBA

Stage 1: Sampling sub-sectors

- For each of the eight strata, a list frame of sub-segments must be produced.
- This list also includes **the number of households** in each sub-segment.
- Sub-segments are selected by RNGs, with probabilities adjusted to be proportionate to size of households in each sub-segment, i.e. sub-segments with more households in them have a higher chance of being selected.

This has the effect of creating a PPS sample of sub-segments, stratified explicitly by district and rurality.

This is based on the expectation that household expenditure may well be correlated with district and type of area (rural or urban).

Roughly 244 sub-segments are sampled out of a total 1,910 sub-segments.

In between Stages 1 & 2

Once the sub-segments are identified, there is a need to produce the list frame of households for each sub-segment.

- Often, the available list is out-of-date, due to housing development, households moving, etc.
- A household listing exercise is conducted, whereby a "surveyor" would survey each sub-segment and list down all the households in it.

This process can take a long time to complete, and is costly (but necessary) to do.

Stages 2 & 3: Sampling households

Once the list frame of households is available, systematic sampling (epsem) is performed to obtain a **fixed number of households** from each sub-segment.

Households that have been identified are then visited, and a reporting individual chosen and interviewed.

Remark

Would it be possible that there are more than one household in the household selected, e.g. multi-generational families? Yes, but these would have been already considered during the household listing exercise. If not, usually select one household at random.

Probability of selection

Simple random sampling (SRS) without replacement is a method of selecting n out N units (at random). At any stage of selection, any one of the remaining units have the same chance of being selected, i.e. $1/N$.

Definition 20 (Selection probability for a unit at the kth draw)

For SRS without replacement,

$$\mathrm{P}(\text{selecting the } i\text{th unit at the } k\text{th draw}) = \frac{1}{N}$$

Probability of selection (cont.)

Proof.

Let A_j denote the event that a particular unit i from the population is selected at the jth draw, $j = 1, \ldots, n$. We are interested in the event $A_1^c \cap A_2^c \cap \cdots \cap A_{k-1}^c \cap A_k$, i.e. the unit is selected only at the kth draw.

$$\begin{aligned}
\mathrm{P}(A_1^c \cap \cdots \cap A_{k-1}^c \cap A_k) &= \mathrm{P}(A_1^c)\,\mathrm{P}(A_2^c \cap \cdots \cap A_{k-1}^c \cap A_k | A_1^c) \\
&= \mathrm{P}(A_1^c)\,\mathrm{P}(A_2^c | A_1^c)\,\mathrm{P}(A_3^c \cap \cdots \cap A_{k-1}^c \cap A_k | A_1^c, A_2^c) \\
&= \qquad \vdots \\
&= \mathrm{P}(A_1^c)\,\mathrm{P}(A_2^c | A_1^c)\,\mathrm{P}(A_3^c | A_1^c, A_2^c)\,\mathrm{P}(A_4^c | A_1^c, A_2^c, A_3^c) \cdots \\
&\qquad \cdots \mathrm{P}(A_k | A_1^c, A_2^c, \ldots, A_{k-1}^c)
\end{aligned}$$

Probability of selection (cont.)

Proof.

Now, $\mathrm{P}(A_1) = 1/N$, as any of the N units can be selected. So $\mathrm{P}(A_1^c) = 1 - 1/N = (N-1)/N$.

After one unit has been drawn, there are $N-1$ units left. So $\mathrm{P}(A_2|A_1^c) = 1/(N-1)$, and thus $\mathrm{P}(A_2^c|A_1^c) = 1 - 1/(N-1) = (N-2)/(N-1)$.

Continuing this pattern, we can show that $\mathrm{P}(A_j^c|A_{j-1}^c, \ldots, A_1^c = (N-j)/(N-j+1)$.

Now once $k-1$ units have been drawn, the probability that the unit is drawn next is simply $P(A_k|A_1^c, \ldots, A_{k-1}^c) = 1/(N-k+1)$.

Probability of selection (cont.)

Proof.

Combining all of this together,

$$\begin{aligned}
\mathrm{P}(A_1^c \cap \cdots \cap A_{k-1}^c \cap A_k) &= \frac{N-1}{N} \cdot \frac{N-2}{N-1} \cdots \frac{N-k+1}{N-k+2} \cdot \frac{1}{N-k+1} \\
&= \frac{\cancel{N-1}}{N} \cdot \frac{\cancel{N-2}}{\cancel{N-1}} \cdots \frac{\cancel{N-k+1}}{\cancel{N-k+2}} \cdot \frac{1}{\cancel{N-k+1}} \\
&= \frac{1}{N}
\end{aligned}$$

□

Probability of inclusion

Note that a unit can be selected either at the first draw, or the second draw, or the third draw, etc. But what is the probability that a unit is selected at all?

Definition 21 (Probability of inclusion)

$$\pi_i = \mathrm{P}(\text{unit } i \text{ is selected}) = \frac{n}{N}$$

Proof.

Again, let A_j be the event that a unit is selected at the jth draw. Consider the event $A = A_1^c \cap \ldots A_n^c$, i.e. the unit is not included in the sample. We are then interested in the event A^c, the unit is included in the sample.

Probability of inclusion

Proof.

$$
\begin{aligned}
\mathrm{P}(A^c) &= 1 - \mathrm{P}(A) \\
&= 1 - \mathrm{P}(A_1^c \cap \ldots A_n^c) \\
&= 1 - \mathrm{P}(A_1^c)\,\mathrm{P}(A_2^c|A_1^c) \cdots \mathrm{P}(A_n^c|A_1^c, \ldots, A_{n-1}^c) \\
&= 1 - \frac{\cancel{N-1}}{N} \cdot \frac{\cancel{N-2}}{\cancel{N-1}} \cdots \frac{N-n}{\cancel{N-n+1}} \\
&= \frac{N-(N-n)}{N} \\
&= \frac{n}{N}
\end{aligned}
$$

as required. □

Probability of a sample being selected

If n units are selected by SRS (without replacement), then the total number of possible samples are $\binom{N}{n}$. Therefore, the probability that any one of this combination is selected is $1/\binom{N}{n}$.

Definition 22 (Probability of drawing a particular sample)

$$\text{P(drawing a sample)} = \frac{1}{\binom{N}{n}} = \frac{n!(N-n)!}{N!}$$

The sample mean

Let $\mathbf{Y} = \{y_1, \ldots, y_N\}$ be the values of a variable y within a population with (unknown) mean μ. Suppose a sample $\mathbf{y} = \{y_1, \ldots, y_n\}$ is drawn using SRS. Earlier we saw that the sample mean

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i, \quad y_i \in \mathbf{y}$$

is an unbiased estimator of the population mean μ, whose variance is given by

$$\text{Var}(\bar{y}) = \frac{N-n}{Nn} \cdot \overbrace{\frac{1}{N-1} \sum_{i=1}^{N} (y_i - \mu)^2}^{S^2} = \frac{N-n}{N} \times \frac{S^2}{n}.$$

We will prove these two facts (unbiased property and form of the variance).

The sample mean is unbiased

Proof.

$$
\begin{aligned}
\mathrm{E}(\bar{y}) &= \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}(y_i) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{N} y_j \cdot \overbrace{\mathrm{P}(y_j \text{ selected at } i\text{th draw})}^{1/N}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mu = \mu
\end{aligned}
$$

Proof of variance of sample mean (SRS)

Under SRS, the variance of the sample mean is **not** σ^2/n.

Proof.

$$
\begin{aligned}
\mathrm{Var}(\bar{y}) &= \mathrm{E}\left[(\bar{y} - \mathrm{E}\,\bar{y})^2\right] \\
&= \mathrm{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}(\bar{y} - \mu)\right)^2\right] \\
&= \mathrm{E}\left(\frac{1}{n^2}\sum_{i=1}^{n}(y_i - \mu)^2 + \frac{1}{n^2}\sum_{i\neq j}(y_i - \mu)(y_j - \mu)\right) \\
&= \underbrace{\frac{1}{n^2}\sum_{i=1}^{n}\mathrm{E}\left[(y_i - \mu)^2\right]}_{(1)} + \underbrace{\frac{1}{n^2}\sum_{i\neq j}\mathrm{E}\left[(y_i - \mu)(y_j - \mu)\right]}_{(2)}
\end{aligned}
$$

Proof of variance of sample mean (SRS) (cont.)

Proof.

Consider the first term (1):

$$\begin{aligned}\frac{1}{n^2}\sum_{i=1}^{n}\mathrm{E}\left[(y_i-\mu)^2\right] &= \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 \\ &= \sigma^2/n\end{aligned}$$

But since $\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(y_i-\mu)^2 = \frac{N-1}{N}\cdot\frac{1}{N-1}\sum_{i=1}^{n}(y_i-\mu)^2 = \frac{N-1}{N}S^2$,

$$\frac{1}{n^2}\sum_{i=1}^{n}\mathrm{E}\left[(y_i-\mu)^2\right] = \frac{N-1}{Nn}S^2$$

Proof of variance of sample mean (SRS) (cont.)

Proof.

Let A_k denote the event that unit k selected at the ith draw, and $B_{k'}$ denote the event that unit k' selected at the jth draw. For $i \neq j$,

$$\begin{aligned}
\mathrm{E}(y_i y_j) &= \sum_{k \neq k'} y_k y_{k'} \cdot \mathrm{P}(A_k \cap B_{k'}) \\
&= \sum_{k \neq k'} y_k y_{k'} \cdot \mathrm{P}(A_k)\, \mathrm{P}(B_{k'} | A_k) \\
&= \sum_{k \neq k'} y_k y_{k'} \cdot \frac{1}{N} \frac{1}{N-1},
\end{aligned}$$

and thus by extension,

$$\mathrm{E}\left[(y_i - \mu)(y_j - \mu)\right] = \frac{1}{N(N-1)} \sum_{k \neq k'} (y_k - \mu)(y_{k'} - \mu) \cdot$$

Proof of variance of sample mean (SRS) (cont.)

Proof.

Furthermore, realise that $\sum_{k=1}^{N}(y_k - \mu) = 0$, and therefore

$$\left(\sum_{k=1}^{N}(y_k - \mu)\right)^2 = \sum_{k=1}^{N}(y_k - \mu)^2 + \sum_{k \neq k'}(y_k - \mu)(y_{k'} - \mu)$$

$$0 = (N-1) \cdot \overbrace{\frac{1}{N-1}\sum_{k=1}^{N}(y_k - \mu)^2}^{S^2} + \sum_{k \neq k'}(y_k - \mu)(y_{k'} - \mu)$$

which implies $\sum_{k \neq k'}(y_k - \mu)(y_{k'} - \mu) = -(N-1)S^2$, giving us

$$\mathrm{E}\left[(y_i - \mu)(y_j - \mu)\right] = \frac{-(N-1)S^2}{N(N-1)} = \frac{-S^2}{N}.$$

Proof of variance of sample mean (SRS) (cont.)

Proof.

Therefore, the result for (2) is

$$\frac{1}{n^2}\sum_{i\neq j}\frac{-S^2}{N} = \frac{1}{n^2}\times 2\cdot\frac{n!}{2!(n-2)!}\times\frac{-S^2}{N} = \frac{n-1}{n}\times\frac{-S^2}{N}$$

Combining (1) and (2) together, we get

$$\begin{aligned}\mathrm{Var}(\bar{y}) &= \frac{N-1}{Nn}S^2 - \frac{n-1}{Nn}S^2 \\ &= \frac{N-n}{Nn}S^2.\end{aligned}$$

Proof of variance of sample mean (SRS) (cont.)

Remark

The reason why the variance of the sample mean is not σ^2/n is because the samples were not drawn independently (due to non-replacement). On the other hand, if independence of samples were assumed (e.g. in SRS with replacement),

- The term (2) in the proof vanishes due to independence.
- The term (1) remains, and as we saw is equal to σ^2/n.
- Note that $\sigma^2/n = \frac{N-1}{Nn}S^2$.

Estimation of population totals

Sometimes, it is also of interest to estimate the population total, e.g. total household income, total expenditures, etc. Let τ denote the population total, i.e.

$$\tau = \sum_{i=1}^{N} y_i = N\mu.$$

A natural estimator for τ is

$$\hat{\tau} = N\bar{y}$$

which is obviously unbiased, as

$$\mathrm{E}(\hat{\tau}) = N\,\mathrm{E}(\bar{y}) = N\mu.$$

This estimator has variance

$$\mathrm{Var}(\hat{\tau}) = \mathrm{Var}(N\bar{y}) = N^2\,\mathrm{Var}(\bar{y}) = \frac{N(N-n)}{n}S^2.$$

Confidence interval for the mean (known variance)

In order to construct a confidence interval for the estimate of the population mean, we have to make an additional assumption:

- Assume that the population is normally distributed with mean μ and variance σ^2.

If each y_i is normal, then

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

is also normal. We know these two facts already: $\mathrm{E}(\bar{y}) = \mu$ and $\mathrm{Var}(\bar{y}) = \frac{N-n}{Nn} S^2$. Therefore,

$$\bar{y} \sim \mathrm{N}\left(\mu, \overbrace{\frac{N-n}{Nn} S^2}^{\mathrm{Var}(\bar{y})}\right).$$

Confidence interval for the mean (known variance) (cont.)

Using the normal distribution for the sample mean, we can construct a $100(1-\alpha)\%$ interval for the mean using the fact that

$$\frac{\bar{y}-\mu}{\sqrt{\text{Var}(\bar{y})}} \sim \text{N}(0,1)$$

and

$$\text{P}\left(-z(\alpha/2) \leq \frac{\bar{y}-\mu}{\sqrt{\text{Var}(\bar{y})}} \leq z(\alpha/2)\right) = 1-\alpha,$$

where $z(\alpha/2)$ is the $(1-\alpha/2)$th percentile of the standard normal dist.

Definition 23 (CI for mean, known variance)

Assuming known variance, the $100(1-\alpha)\%$ confidence interval for the mean is given by

$$\left(\bar{y} - z(\alpha/2)\sqrt{\text{Var}(\bar{y})}, \bar{y} + z(\alpha/2)\sqrt{\text{Var}(\bar{y})}\right).$$

Confidence interval for the mean (unknown variance)

Most likely the population variance S^2 is unknown, and has to be estimated by

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

In such cases, instead of a normal distribution for the sample mean, it follows a t-distribution with $(n-1)$ degrees of freedom:

$$\frac{\bar{y} - \mu}{\sqrt{\widehat{\mathrm{Var}}(\bar{y})}} \sim t_{n-1},$$

where $\widehat{\mathrm{Var}}(\bar{y}) = \frac{N-n}{Nn}s^2$.

Confidence interval for the mean (unknown variance) (cont.)

Thusly,

$$\mathrm{P}\left(-t_{n-1}(\alpha/2) \leq \frac{\bar{y}-\mu}{\sqrt{\widehat{\mathrm{Var}}(\bar{y})}} \leq t_{n-1}(\alpha/2)\right) = 1-\alpha,$$

from which we can construct the confidence interval.

Definition 24 (CI for mean, unknown variance)

When the population variance is unknown, the $100(1-\alpha)\%$ confidence interval for the mean is given by

$$\left(\bar{y} - t_{n-1}(\alpha/2)\sqrt{\widehat{\mathrm{Var}}(\bar{y})}, \bar{y} + t_{n-1}(\alpha/2)\sqrt{\mathrm{Var}(\bar{y})}\right).$$

Determination of sample size

How large of a sample do we need? We have discussed that

- The larger the sample size, the smaller the sampling error (MSE/variance of estimators).
- However, the larger the sample size, the costlier the survey will be.
- Other non-sampling errors also generally increase with sample size (e.g. data-entry error, human errors, etc.)

It seems that there should be a basis in which sample size is determined optimally. We shall consider four aspects:

- Bound on the variance (of the estimator)
- Bound on estimation error
- Bound on the confidence interval
- Bound on costs

Bound on the variance

Suppose we wish to collect a certain number of samples n such that the resulting variance of the estimator, $\text{Var}(\bar{y})$, does not exceed a given value B. In this case, we must find n such that

$$\begin{aligned}
\text{Var}(\bar{y}) = \frac{N-n}{Nn} S^2 &\leq B \\
\frac{1}{n} - \frac{1}{N} &\leq \frac{B}{S^2} \\
\frac{1}{n} &\leq \frac{B}{S^2} + \frac{1}{N} = \frac{NB + S^2}{NS^2} \\
\Rightarrow n &\geq \frac{NS^2}{NB + S^2} \\
&= \frac{S^2/B}{1 + N^{-1} \cdot S^2/B}
\end{aligned}$$

Bound on the estimation error

If we have some prior knowledge about what the value of μ should be (e.g. from expert opinions, past research, etc.), and we require a sample size n such that the sample mean $\bar{y}$ should not differ from μ by a specified amount ϵ. Since $\bar{y}$ is random after all, the best we can do is make the following probability statement:

$$\mathrm{P}(|\bar{y} - \mu| \leq \epsilon) = 1 - \alpha.$$

In words, "the probability of the estimate $\bar{y}$ not being ϵ units away from the mean μ is equal to $1 - \alpha$". Ideally we want ϵ to be as small as possible, and $1 - \alpha$ to be as close to 1 as possible.

Bound on the estimation error (cont.)

Since

$$\frac{\bar{y} - \mu}{\sqrt{\mathsf{Var}(\bar{y})}} \sim \mathsf{N}(0, 1),$$

this implies that

$$\begin{aligned}
\mathsf{P}(|\bar{y} - \mu| \leq \epsilon) &= \mathsf{P}\left(\frac{|\bar{y} - \mu|}{\sqrt{\mathsf{Var}(\bar{y})}} \leq \frac{\epsilon}{\sqrt{\mathsf{Var}(\bar{y})}} \right) \\
&= \mathsf{P}\left(\frac{-\epsilon}{\sqrt{\mathsf{Var}(\bar{y})}} \leq \frac{\bar{y} - \mu}{\sqrt{\mathsf{Var}(\bar{y})}} \leq \frac{\epsilon}{\sqrt{\mathsf{Var}(\bar{y})}} \right) \\
&= 1 - \alpha,
\end{aligned}$$

in which case we are looking for the sample size n which satisfies

$$z(\alpha/2) = \frac{\epsilon}{\sqrt{\mathsf{Var}(\bar{y})}}.$$

Bound on the estimation error (cont.)

Solving for n, we obtain

$$\begin{aligned}
\sqrt{\text{Var}(\bar{y})} &= \frac{\epsilon}{z} \\
\left(\frac{1}{n} - \frac{1}{N}\right) S^2 &= \frac{\epsilon^2}{z^2} \\
\frac{1}{n} &= \frac{\epsilon^2}{S^2 z^2} + \frac{1}{N} \\
&= \frac{N\epsilon^2 + S^2 z^2}{N S^2 z^2} \\
\Rightarrow n &= \frac{N S^2 z^2}{N\epsilon^2 + S^2 z^2} = \frac{S^2 z^2}{\epsilon^2 + S^2 z^2 / N}
\end{aligned}$$

Bound on the confidence interval

In a similar manner, we can figure out the smallest n required such that the width of the confidence interval from Definition 23 does not exceed a certain width B

$$2z(\alpha/2)\sqrt{\text{Var}(\bar{y})} \leq B$$

Remark

Obviously the smaller the width, the more confident we are in our estimation.

Bound on the confidence interval (cont.)

Solving for n yields

$$
\begin{aligned}
2z\sqrt{\mathrm{Var}(\bar{y})} &\leq B \\
\mathrm{Var}(\bar{y}) &\leq \frac{B^2}{4z^2} \\
\left(\frac{1}{n} - \frac{1}{N}\right) S^2 &\leq \frac{B^2}{4z^2} \\
\frac{1}{n} &\leq \frac{B^2}{4z^2S^2} + \frac{1}{N} \\
&= \frac{NB^2 + 4z^2S^2}{4Nz^2S^2} \\
\Rightarrow n &\geq \frac{4Nz^2S^2}{NB^2 + 4z^2S^2} = \frac{4z^2S^2/B^2}{1 + 4z^2S^2/(NB^2)}
\end{aligned}
$$

Bound on survey costs

Let B be the amount of money budgeted to conduct sampling for the survey. Typically, the costs of sampling can be divided into two:

- Fixed overhead costs B_0.
- Variable costs B_1, i.e. cost of obtaining one unit in the sample.

So $B = B_0 + B_1 n$.

Therefore, the required sample size is obtained by solving for n, which gives us

$$n = \frac{B - B_0}{B_1}.$$

Remarks

Remark

In all of these bounds, we required to know the true value of the population variance S^2. There are two strategies that we can adopt:

- Choose a fixed value for S^2 based on an expert decision, past research, or a pilot survey.
- Assume that S^2 is unknown, and use the estimator s^2 in its place. Slight changes are required, e.g. using t_{n-1} distribution instead of the standard normal (however, for large n, $t_{n-1} \approx \mathrm{N}(0,1)$, so it doesn't matter too much).

Remark

It is also possible to use these results if other sampling techniques are deployed (e.g. cluster sampling or stratified sampling), provided we know the formula for the variance of the estimator $\mathrm{Var}(\bar{y})$ in these instances.

References I

- Groves, R. M., F. J. Fowler Jr, M. P. Couper, J. M. Lepkowski, E. Singer, and R. Tourangeau (2011). *Survey methodology*. Vol. 561. John Wiley & Sons.
- Shalabh (Jan. 2020). *Chapter 2, Sampling Theory*. URL: http://home.iitk.ac.in/~shalab/sampling/chapter2-sampling-simple-random-sampling.pdf.
- United Nations Statistical Division (2008). *Designing household survey samples: practical guidelines*. Vol. 98. United Nations Publications. URL: https://unstats.un.org/unsd/demographic/sources/surveys/Handbook23June05.pdf.

References II

Vaessen, M., M. Thiam, and T. Le (2005). "Household Sample Surveys in Developing and Transition Countries". *United Nations, Department of Economic and Social Affairs, Statistics Division. Studies in Methods, Series F* 96. URL: https://unstats.un.org/unsd/hhsurveys/pdf/Household_surveys.pdf.