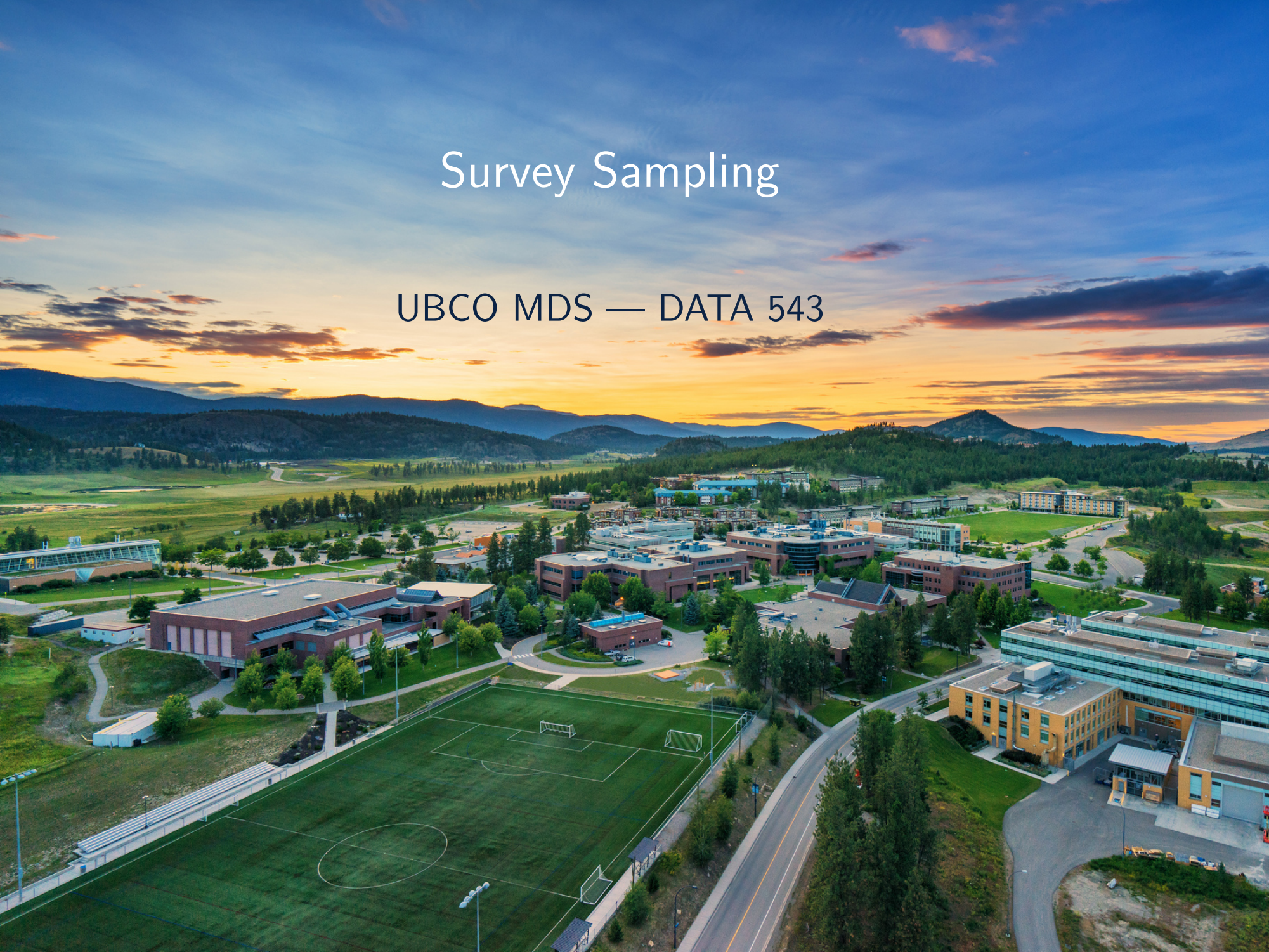
Survey Sampling

UBCO MDS — DATA 543



Survey Sampling

- Survey sampling typically refers to observational studies (i.e. passively observe and record the variable(s) of interest).

- One objective of survey sampling is to estimate some property (eg. mean, total) of a finite population based on a sample, i.e. without conducting a complete census of every item in the population.

Finite Population

- A *population* refers to a collection of 'elements' each having certain characteristics of interests.
- Let $U = \{1, \ldots, i, \ldots, N\}$ denote a *finite population*, where N is the population size, e.g. all individuals residing in Canada
- By finite population (aka *survey population*) we mean a finite set of labeled individuals.
- The elements labeled $1, 2, \ldots, i, \ldots, N$ may be called its *elements* or *units* or *individuals*.
- A *sample* is a subset of the population.

Requirements of a Good Sample

- Lohr (2009) describes a perfect sample as "a scaled-down version of the population, mirroring every characteristic of the whole population". For example,

	Population		Sample	
	M	F	M	F
Employed	10000	8000	100	80
Unemployed	500	300	5	3

- We look for a *representative* sample.
- A sample is representative when the characteristics of interest in the population can be estimated with a known degree of accuracy.

Definitions

- The *target population* is the complete collection of individuals or elements we want to study.
  - In a political poll the target population might be: all eligible voters, all registered voters, all who voted last election, ...

- The *sampled population* (aka *study population*) is the collection of all possible elements that might have been chosen in a sample. N.B. this is not necessarily the units that were actually sampled!

Target vs Sampled population

The ideal scenario in which the sampled population will be identical to the target is rarely met in practice. The sampled population is usually smaller than the target population.

- A *sampling unit* is a unit that can be selected for a sample.
  - Sampling units can be the individual elements, or clusters (eg. households)

- The *observation unit* (OU) is the unit we take measurement from (sometimes called an element)
  - In a study on humans, OU are usually individuals.

- The *sampling frame* is the list of sampling units, for example,
  - telephone survey: list of all residential phone #'s in the city
  - agricultural survey: a list of all farms in a certain area

Sampling units vs. observation units

If we want to study individuals, but do not have a list of all individuals in the target population, households may serve as the sampling units, and the observation units are the individuals living in the households.

Example 1 (UBC President approval)

The UBC President wanted to know the approval rating among current UBCO undergraduate students. To do so, we obtained a list of email addresses of students who had volunteered during orientation week. We then picked 100 students from this list, sending each an email asking whether they thought he was "a good President". All students responded.

Identify each of the following:

The target population

The sampled population

The sampling unit

The observation unit

The sampling frame

UBC President Approval Example: Answers

Identify each of the following:

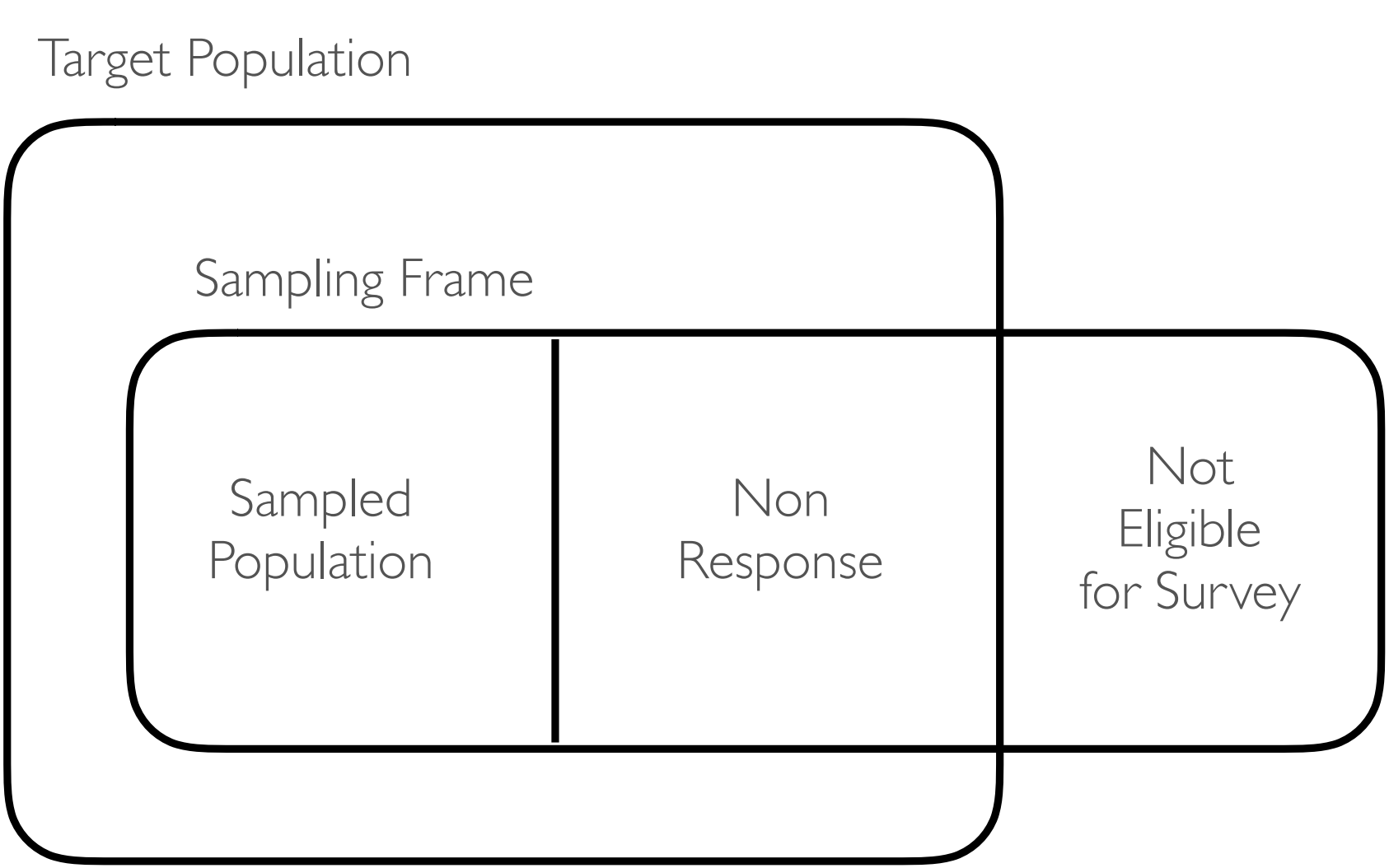
The target population: All current UBCO undergraduate students

The sampled population: All students who volunteered during orientation week

The sampling unit: The email addresses of students who volunteered during orientation week

The observation unit: The 100 UBCO undergraduate students who volunteered during orientation week

The sampling frame: The list of email addresses of students who volunteered during orientation week


Target Population
Sampling Frame
Sampled Population
Non Response
Not Eligible for Survey

Selection Bias

- *Selection bias* occurs when some population units are sampled at a different rate than intended by the investigator.
- Potential causes:

  undercoverage Failing to include all of the target population in the sampling frame.

  overcoverage Including population units in the sampling frame that are not in the target population.

  nonresponse Failing to obtain responses from all of the chosen sample.

  - Nonresponse distorts the results of many surveys, since nonrespondents differ critically from the respondents
  - The extent of that difference is unknown unless you can later obtain information about the nonrespondents.
- A good sample will be as free from selection bias as possible.

Study Errors

(Lohr, 1999, pg. 15–17), (Chaudhuri and Stenger, 2005, pp.297–325)

Recall the objective of sample surveys is primarily to estimate a population attribute (e.g. the mean age of students in the class) based on a sample. When doing this, we can generally classify our sources of error into one of three types.

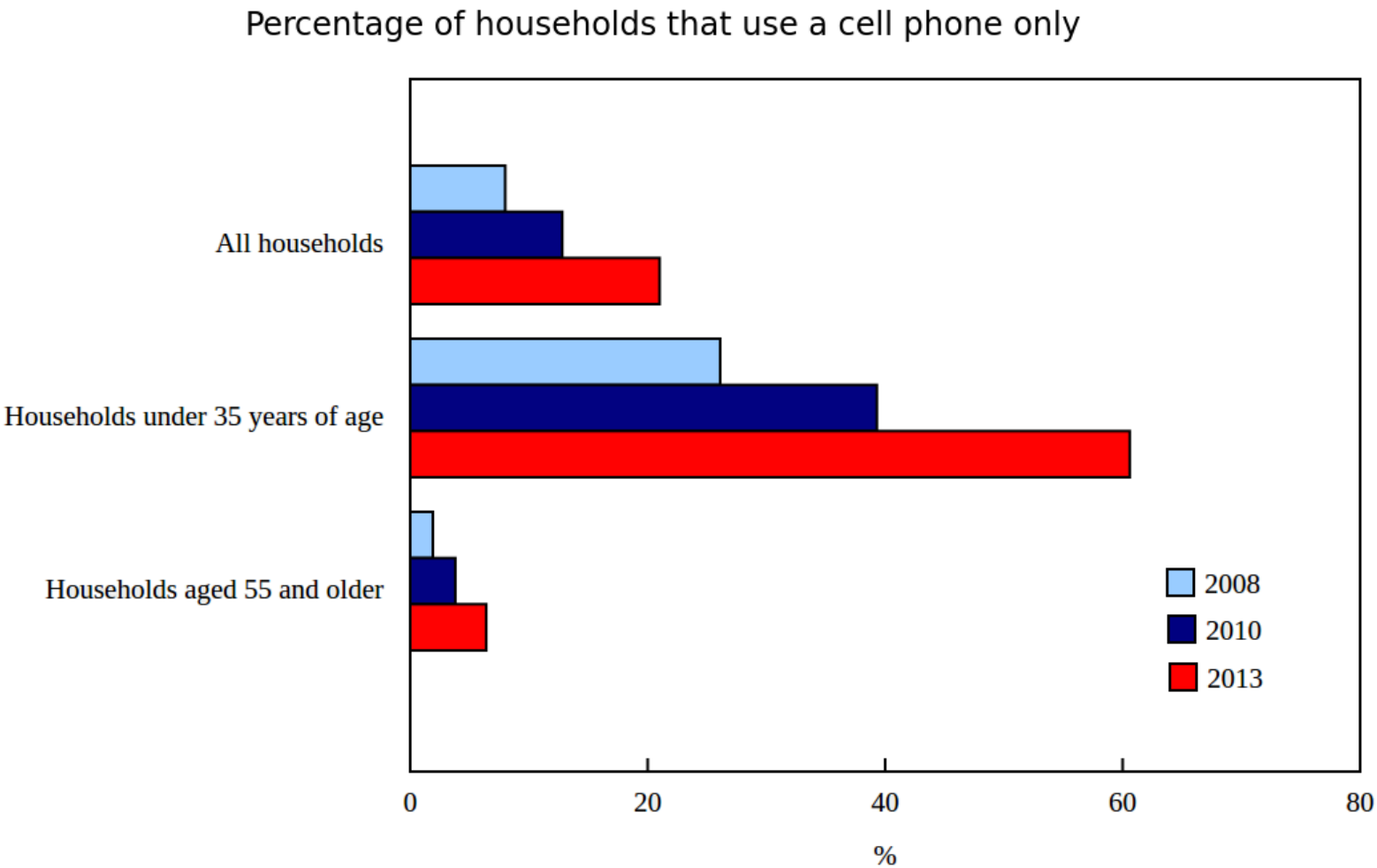
1. Frame (aka coverage) error: the difference in attributes of interest between the target population and sampling frame. A frame error occurs when the wrong sub-population is used to select a sample.
2. Sample error: the difference in attributes of interest between the sampling frame and the sample. These errors occur because of variation in the number or representativeness of the sample that responds.
3. Measurement error: when the true and measured values of the variates on the units in the sample are different.

Frame Error Example

e.g. electoral polls often involve automatically dialling limited to landlines.

- The sampling frame is (e.g.) a telephone directory, but the target population is potential voters.
- Younger people are less likely to have landlines, so our sampling frame (a list of telephone numbers) differs from the target population (all potential voters) in a manner that might be problematic (e.g. if younger/older people are more likely to vote a certain way).

Frame Error Example



Source: Statistics Canada, 2013

Sample Error Example

e.g. suppose a polling company could contact *everyone* in their target population.

- They're likely to encounter **non-response**
- If certain voters are more/less likely to respond to polling, this can lead to sample error.
  - "shy" Trump voters: people who are not willing to admit they support(ed) Trump, and therefore refuse to respond.
  - Note: while this is a 'good example' in the sense you've likely heard of this potential phenomenon, there isn't much (if any) evidence to support it[1]

[1] https://fivethirtyeight.com/features/shy-voters-probably-arent-why-the-polls-missed-trump/

Measurement Error Example

When the true and measured values of the variates on the units in the sample are different.

- Direct measurements, such as height, blood pressure, diet.
- If respondents lie (social desirability bias, discomfort with the question, bad memory, etc).
- 'Leading' questions. e.g. 'Yes Prime Minister'
- Interviewers could affect response.
- Agree/disagree questions (people tend towards agreement with any statement regardless of the content).

Measurement bias occurs when the response has a tendency to differ from the true value in one direction.

As with selection bias, measurement error and bias must be considered and minimized in the design stage of the survey.

Measurement Error

The best 'solution' to measurement error is to try and avoid it through careful design.

- e.g. phrasing questions fairly[2], or reframing the study question so it doesn't rely on difficult to measure variables.
- If data *are* measured with error, there are many statistical techniques to try and address it.
- Methods for statistically correcting measurement error are beyond the scope of this module, but you should still keep it in mind!

Selection bias and measurement error are examples of *nonsampling errors*, which are any errors that cannot be attributed to the sample-to-sample variability.

[2] See Lohr (2009) section 1.5 on how to properly design a questionnaire if interested.

Sample

Definition 2 (sample)

A sequence $s = (i_1, i_2, \ldots, i_n)$ of which each entry i_j is one of the elements of U, *not necessarily distinct* from each other but *ordered* respectively as the 1st, 2nd, ... jth, ..., nth member of the sequence s and $1 \leq n < N$ is called a *sample* from the population U.

- n is called the sample size of s or just the size of s.

Sidenote

- In survey sampling, we seek to learn about some characteristics of the finite population from the sample.
- We recognize that the population may change over time and we are only observing a snapshot.
- In this course we will assume the population as fixed, or to put another way, any changes to the population in the period of our study is not a big concern.

Sampling Protocol

- A *sampling protocol* (or *sampling design*) is the mechanism by which we choose our samples.
- A *probability sampling protocol* is where some probabilistic method is used to select the sample from the frame.
  - eg. for each student in the class list I toss a coin, if heads, that student is sampled.
- A *non-probability sampling* is where samples are selected based on the subjective judgement of the interviewer.
  - Convenience sampling,
  - Self-selection sampling,
  - Quota sampling,
  - Judgment sampling,
  - ...

Non-probability sampling

- *Convenience sampling* units are sampled based on what's easily available e.g. students who show up to class.
- *Self-selection sampling*: units choose themselves.
  - e.g. many internet polls
- *Quota sampling*: units are selected so that some attributes of the sample match known attributes in the target population.
  - e.g. if 50% of the class are computer science majors, 30% are math majors, and 20% are stat majors, I pick my sample so that it has the same distribution.
- *Judgment sampling*: units are selected so the samplers *think* the sample will be representative of the whole population.
  - e.g. I try to form a sample of students where the proportion of majors match up to those in the whole class, but I guess what major students are.

Problems with non-probability sampling

- Convenience sampling is often biased, since the units that are easiest to select or that are most likely to respond are usually not representative of the harder-to-select or nonresponding units.
  - eg. students who show up to class may be inherently different than the students who choose to skip class.

- Quota: is someone's major an important/relevant attribute to try and build into our sample?

- Judgment: is my judgment biased?
  - Eg, If we want to estimate the average amount a shopper spends and select people who look like they have spent an "average" amount we sample to confirm our prior opinion.

Problems with non-probability sampling

- Self-selection: are students who volunteer to be in a sample representative of the whole class?

  For example:

  - Suppose we are sampling adolescents to study how frequently adolescents talk to their parents and teachers about AIDS.
  - Adolescents willing to talk to the investigators about AIDS are probably also more likely to talk to other authority figures about AIDS.
  - Results from this sample will probably overestimate the amount of communication occurring between parents/teachers and adolescents in the population.

Probability Sampling

- Probability sampling: selecting units for the sample based on a probability model.
- Major advantage: if we understand the probabilistic mechanism we've used to form our sample, we can assess the sample error mathematically.
- In reality, our sample will (almost certainly) differ from the target population, so we will have uncertainty about the population parameter we're interested in.
- If we have a statistical model for how we've sampled units, we can estimate this uncertainty in the form of confidence intervals and hypothesis tests.

Probability Sampling

Here are a few probability sampling techniques:

- Simple random sampling without replacement.
- Simple random sampling with replacement.
- Systematic sampling.
- Cluster sampling.
- Stratified sampling.

Probability Sample

- In probability sampling, each element (sampling unit) in the (sample/study) population has a known, non-zero probability of being included in the sample
- Such a sampling can be specified through a probability measure defined over the set of all possible samples.
- The *sample space*, denoted $\mathcal{D} = \{s\}$, is the set of all possible samples from the population.
- A probability measure P, which is defined on the sample space, takes the value $P(s)$ s.t. $0 \leq P(s) \leq 1 \; \forall \; s$ in $\mathcal{D}$ and $\sum_{s \in \mathcal{D}} P(s) = 1$.
- P is referred to as *sampling design* or just *design*.

Sampling Design

Source Wu & Jiahua for example

Example 3

Let $N = 3$ and $U = \{1, 2, 3\}$. The collection of possible samples are $s_1 = \{1\}$, $s_2 = \{2\}$, $s_3 = \{3\}$, $s_4 = \{1, 2\}$, $s_5 = \{1, 3\}$, $s_6 = \{2, 3\}$, and $s_7 = \{1, 2, 3\}$.

An example of probability measure P is given by

s	s_1	s_2	s_3	s_4	s_5	s_6	s_7
$P(s)$	$1/9$	$1/9$	$1/9$	$2/9$	$2/9$	$2/9$	0

Selection of a sample based on above probability measure can be done using a random number generator in R.

```
> sample(1:7, 1, prob=c(1, 1, 1, 2, 2, 2, 0)/9)
```

Random sample generator

- Building from the example above, we may want to restrict our sample size to $n = 2$. In that case, our sampling design may look like this:

s	s_1	s_2	s_3	s_4	s_5	s_6	s_7
$P(s)$	0	0	0	1/3	1/3	1/3	0

In R we could generate

```
> sample(1:7, 1, prob=c(0, 0, 0, 1, 1, 1, 0)/3)
```

The output will be a number between 4 and 6 corresponding to s_4 through s_6. N.B. s_1–s_3 (samples of size one) and s_7 (the complete population) has a zero percent chance of being selected.

Inclusion Probability

The probability that element i is selected in the sample is called *inclusion probability*, denoted by

$$\pi_i = P(\text{unit } i \text{ in sample}), i = 1, 2, \ldots, N.$$

It is required that all $\pi_i > 0$. If $\pi_i = 1$, the element is guaranteed to be included in the sample.

Example pg 25 of Lohr (1999)

Example 4

Let $N = 4$ and $U = \{1, 2, 3, 4\}$. The collection of possible samples of size 2 are $s_1 = \{1, 2\}$, $s_2 = \{1, 3\}$, $s_3 = \{1, 4\}$, $s_4 = \{2, 3\}$, $s_5 = \{2, 4\}$, $s_6 = \{3, 4\}$

An example of probability measure P is given by

s	s_1	s_2	s_3	s_4	s_5	s_6
$P(s)$	1/3	1/6	0	0	0	1/2

The inclusion probability for each i is

$$\pi_1 = P(s_1) + P(s_2) + P(s_3) = 2/6 + 1/6 = 3/6 = 1/2$$
$$\pi_2 = P(s_1) + P(s_4) + P(s_5) = 1/3$$
$$\pi_3 = P(s_2) + P(s_4) + P(s_6) = 2/3$$
$$\pi_4 = P(s_3) + P(s_5) + P(s_6) = 1/2$$

Simple Random Sampling

- One of the simplest sampling designs for selecting 1 unit from a population of size N is to assume $\pi_i = \frac{1}{N}$ for $i = 1, \ldots, N$.
- Once $n > 1$ we have two choices:
  - *Simple random sampling with replacement* (SRSWR)
  - *Simple random sampling without replacement* (SRSWOR)

Usage of 'SRS'

Sometimes SRSWOR is shortcutted to SRS. Be aware, however, that SRS could also refer to 'stratified random sampling'.

SRSWR

For simple random sampling *with* replacement,

- To draw a sample of size n from a population of size N we repeat the process of randomly selecting 1 unit exactly n times
- Each time a unit is selected it is *replaced* or *returned* back to the population so that at every draw a unit's selection-probability remains at $1/N$.
- i.e. units can be selected more than once.

In R we would need to specify `replace = TRUE` in:

```
> sample(1:N, n, replace = TRUE)
```

SRSWOR

In finite population sampling, however, sampling the same person twice provides no additional information. We usually prefer to sample without replacement, so that the sample contains no duplicates

For simple random sampling with*out* replacement we will use a slightly different notation for our sample,

- Let $s^* = \{j_1, \ldots, j_K\}$ be a a set of *distinct* and *unordered* elements of U with $1 \leq K < N$.
- K is called the sample size for s^*

In R we would use `sample` with `replace = FALSE` (the default):

```
> sample(1:N, K)
```

SRSWOR

For simple random sampling with*out* replacement $n > 1$

- Draw the first unit randomly; each unit selection-probability is $1/N$.
- Draw the second unit randomly; each unit selection-probability is $1/(N-1)$.

  $\vdots$
- Draw the ith unit randomly; each unit selection-probability is $1/(N-i-1)$.

SRSWOR Example

Suppose we have a population of 100 first year students from 20 different programs (5 students from each program), we want to select a sample of size 20.

In SRSWOR every possible subset of n units in the population has the same chance of being the sampled.

- e.g. We could use the list of all the students and randomly select a sub-sample without replacement

```
> # to sample their position in the list:
> sample(100, 20)
 [1] 92 81 57 11 65 84 40 20 16 28 71
[12] 39 38 48 43 76 80 74 21 25
```

SRSWOR Example

See Lab 1 for R code

```
> head(classList)
                     names           program   englcourse
1     Glammeyer, Jonathan           History ENGL 100 001
2 French-Stegall, Antonio       Mathematics ENGL 100 001
3       Francis, Cheyenne              Arts ENGL 100 001
4           Zuni, Michael Civil Engineering ENGL 100 004
5      Vialpando, Lesslie Civil Engineering ENGL 100 005
6         Shirali, Alayna         Geography ENGL 100 004

> sample(classList$names, 20)
 [1] Lobatos, Ivan
     Leyba, Elaine
     Hamilton, Curtis
     Juarez, Malyssa
     Lampe, Kevin
     ...
```

Stratified Sampling

- Sometimes the population is divided (either naturally or imposed) into a number of distinct non-overlapping subpopulations called *strata* (from Latin words meaning "to make layers").
- A stratum (sing.) is a homogeneous subset of the population. (Plural for stratum = strata)
- If we divide the population into H strata, the strata should have no overlap, and they constitute the whole population so that each sampling unit belongs to exactly one stratum.
- We then draw an independent probability sample from each stratum, then pool the information to obtain overall population estimates.

Stratified Sampling

- U_h, $h = 1, 2, \ldots, H$ such that $U_h \cap U_{h'} = \emptyset$ for all $h \neq h'$ and $U_1 \cup U_2 \cup \cdots \cup U_H = U$
- If N_h is the size of the hth stratum, we must have $N_1 + N_2 + \cdots + N_H = N$.

Stratified Sampling Advantages

Stratification can enhance the precision of estimators.

- It can be shown that the variance of estimates under a so-called *Stratified SRSWOR* is much smaller than that using a SRSWOR.
- This happens because the variance within each stratum is often lower than the variance in the whole population.
    - e.g. A study on blood pressure may want to stratify according to age group since different ages tend to have different blood pressure.

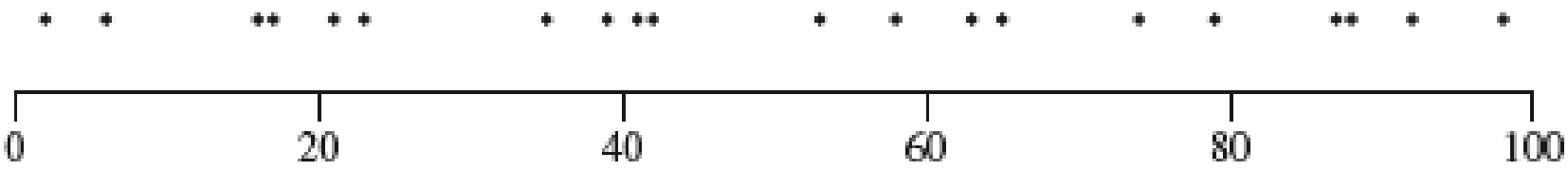
Stratified Sampling Disadvantages

Since stratified sampling almost always leads to higher precision estimates than SRSWOR why ever do SRSWOR?

- Stratification adds complexity which may not be worth a small gain in precision.
- Stratified sampling is most useful when strata means are very different.

Stratified Sampling Example

Suppose we have a population of 100 students from 20 different programs (5 students from each program), we want to select a sample of size 20.



To create a stratified random sample, we could divide the population strata and take a SRSWOR from each stratum

- e.g. We use the list of all the students divided by program and randomly select a subsample for each program

```
byProgram <- split(classList, program)
stratSamp <- lapply(byProgram, function(x) sample(x$names,1))
> stratSamp
$Accounting
[1] Todacheene, Sandora

$Arts
[1] Francis, Cheyenne

$Chemistry
[1] Xiao, Joshua

$`Civil Engineering`
[1] el-Daoud, Sanad

...
```

Cluster Sampling

- A closely related method to stratified sampling is cluster sampling.
- As before, the population is composed of $C(>1)$ *clusters* consisting of M_i $(\sum_{i=1}^{C} M_i = N)$ individuals.
- A sample of $r(<C)$ clusters is chosen by SRSWOR and *a*ll elements M_i of each of the sampled clusters are surveyed.

Cluster Sampling vs Stratification

Figure: Difference between cluster sampling and stratification: Source of image + useful discussion

Cluster Sampling Example

Suppose we have a population of 100 students from 20 different programs (5 students from each program), we want to select a sample of size 20.

Cluster sample units in the population are aggregated into larger sampling units, called clusters.

- e.g. Since all first year students take ENGL 100, randomly select one section of that course and interview all the students.

```
> byCourse <- split(classList, englcourse)
> sampCourse <- as.character(sample(unique(classList$englcourse),1))
> byCourse[sampCourse]
$`ENGL 100 001`
                      names                 program   englcourse
1      Glammeyer, Jonathan                 History ENGL 100 001
2  French-Stegall, Antonio             Mathematics ENGL 100 001
3        Francis, Cheyenne                    Arts ENGL 100 001
7        Lancaster, Joseph                 History ENGL 100 001
...
```

Cluster Sampling vs Stratification

Key difference include:

- In cluster sampling only a subset of clusters are being sampled from. In stratified sampling all strata are being sampled from.
- In cluster sampling all members from selected clusters are surveyed. In stratified sampling only a sample of the members from the selected strata are sampled.
- Clusters are assumed to be heterogeneous, while strata are homogeneous segments.

Cluster Sampling vs Stratification

Some comments:

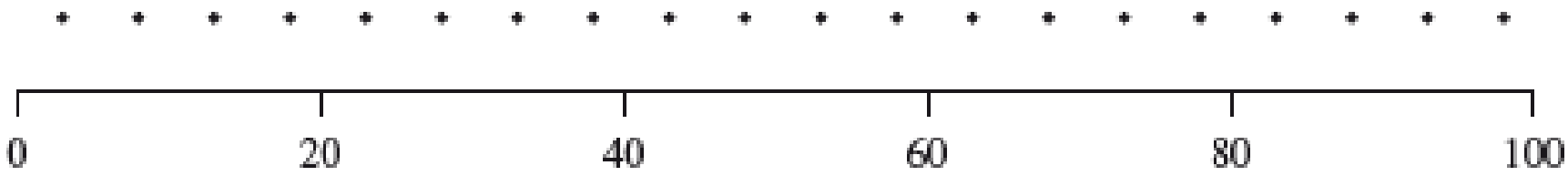
- Cluster sampling has the advantage reducing cost and improving efficiency (eg. if each cluster is a neighbourhood, less neighbourhoods visited).
- When cluster sizes are not all equal, complications will arise.

Systematic Sampling

- In *systematic sampling*, every kth unit is selected from the list of the population elements.
  - eg. a systematic sample from a class list of 250 might take every fifth student from the alphabetized list on Canvas.
- Systematic sampling is also approximately the same as SRSWOR when the population is roughly in a random order.

Systematic Sampling Example

Suppose we have a population of 100 students from 20 different programs (5 students from each program), we want to select a sample of size 20.



A starting point is chosen from a list of population members using a random number. That unit, and every k^{th} unit thereafter, is chosen to be in the sample.

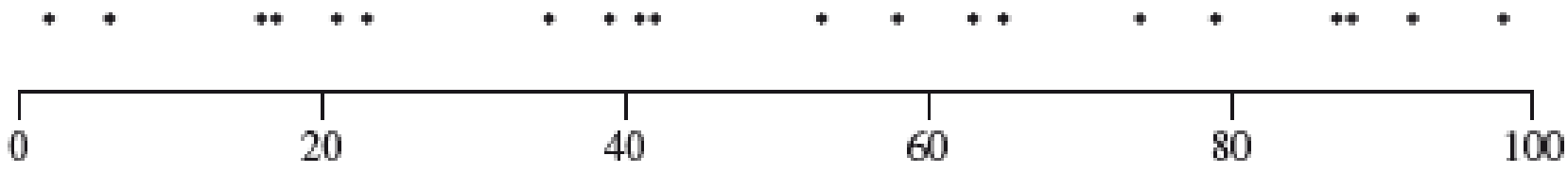
- e.g. We start from the 2^{nd} student in the list and sample every fifth student thereafter.

```
> # start at 2 and take every 5th one thereafter
> clustSampB <- seq(from=2, to =100, by = 5)
> clustSampB
 [1]  2  7 12 17 22 27 32 37 42 47 52 57 62 67 72 77 82 87 92 97
> classList$names[clustSampB]
 [1] French-Stegall, Antonio
     Lancaster, Joseph
     Stanley, Aaron
     Boisselle, Lane
     ...
```

Image source: Fig 2.1 from Lohr (2009)

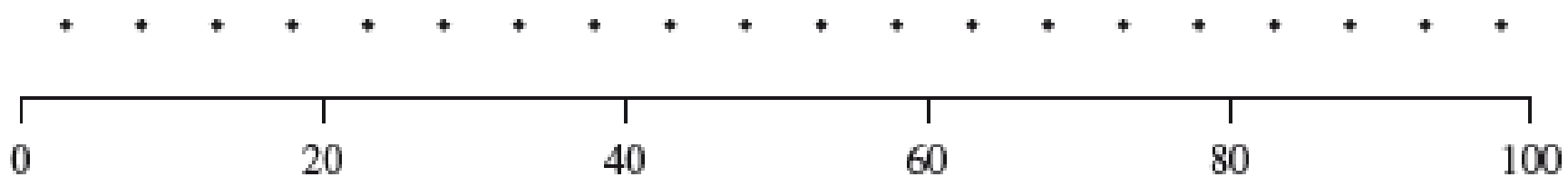
Simple random sample:

Stratified random sample:



Cluster sample:

Systematic sample:



References I

Chaudhuri, A. and Stenger, H. (2005), *Survey Sampling: Theory and Methods, Second Edition*, Statistics: A Series of Textbooks and Monographs, CRC Press.
URL: *https://books.google.ca/books?id=U_JE5mi5ohMC*

Lohr, S. (2009), *Sampling: design and analysis*, Nelson Education.

Lohr, S. L. (1999), 'Sampling: Design and analysis'.