STAT 332 - Sampling Survey Issues

- Terminology
- Sampling Protocols
- Errors
- Questionnaire Design
- Course Notes Coverage: Chapter 4

Why sample?

We conduct samples to learn about a **population**. e.g.

-
-
-

Key aspect: our population is

In theory, we could conduct a

Terminology

Terminology is **essential** in statistics.

Do not underestimate the importance of using precise language!

This isn't a survey example, but consider the statement:

> *"We're studying an eye disease and want to know if our new treatment works."*

Terminology

An important (and often overlooked) part of statistics is converting vague objectives into precise statements.

> *"Do patients in our target population (children aged 3-8 years) show a significant improvement in visual acuity after receiving 3 months of the new treatment when compared with patients who receive 3 months of standard treatment?"*

Terminology

Observational Unit:

Target Population:

-
-
-
-

Note: a ‘unit’ is not necessarily a single person!

Terminology

Sample Population:

-
-

Sampling Frame:

-
-

Sampling Unit:

Terminology

Keep in mind:

- The sample population and the target population **can** be identical.
- Observational units are sometimes referred to simply as ‘units’.
- Do not forget that ‘sample population’ does not (necessarily) mean the units that were sampled!

Example

The UW President wanted to know the approval rating among current UW 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 observational units
- The target population
- The sample population
- The sampling frame
- The sampling units

Example

Answers:

- The observational units:
- The target population:
- The sample population:
- The sampling frame:
- The sampling units:

Important: don't assume this is trivial!

e.g. 'students' would not be correct for the target population.

Population Parameters

Suppose our target is population is $\mathrm{U} = \{1, 2, 3, \ldots, N\}$

- N:
- The study variable or response of interest is :

Some population parameters of interest are:

-
-
-

Population Proportion

Suppose the study variable or response of interest is binary, e.g.

-
-

The study variable is an

the population total is

and the population average is a

Variance Properties

For variance we have two relations.

1. For any response we have

2. and for binary responses we have

Variance Property for any response

Variance Property for Binary Responses

Census

A **census** an investigation is where we

A sample survey is preferred over a census because of

- the improved quality of the estimates available from a carefully conducted survey rather than a sloppy census,
-
-

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

- A **non-probability sampling** is where samples are selected based on the subjective judgement of the interviewer.

We'll cover several types of probability sampling protocols during the course.

Non-Probability Sampling Protocols

Some non-probability sampling protocol are

- Convenience sampling,
- Self-selection sampling,
- Quota sampling, and
- Judgment sampling.

Convenience sampling

Convenience sampling: units are sampled based on what's easily available.

- e.g.

- e.g.

Self-selection sampling

Self-selection sampling: units choose themselves.

- e.g.

- e.g.

Quota sampling

Quota sampling: units are selected so that some attributes of the sample match known attributes in the target population.

- e.g.

- e.g.

Judgment sampling

Judgment sampling: units are selected so the samplers *think* the sample will be representative of the whole population.

- e.g.

Study Errors

Non-probability sampling protocols have lots of obvious problems.

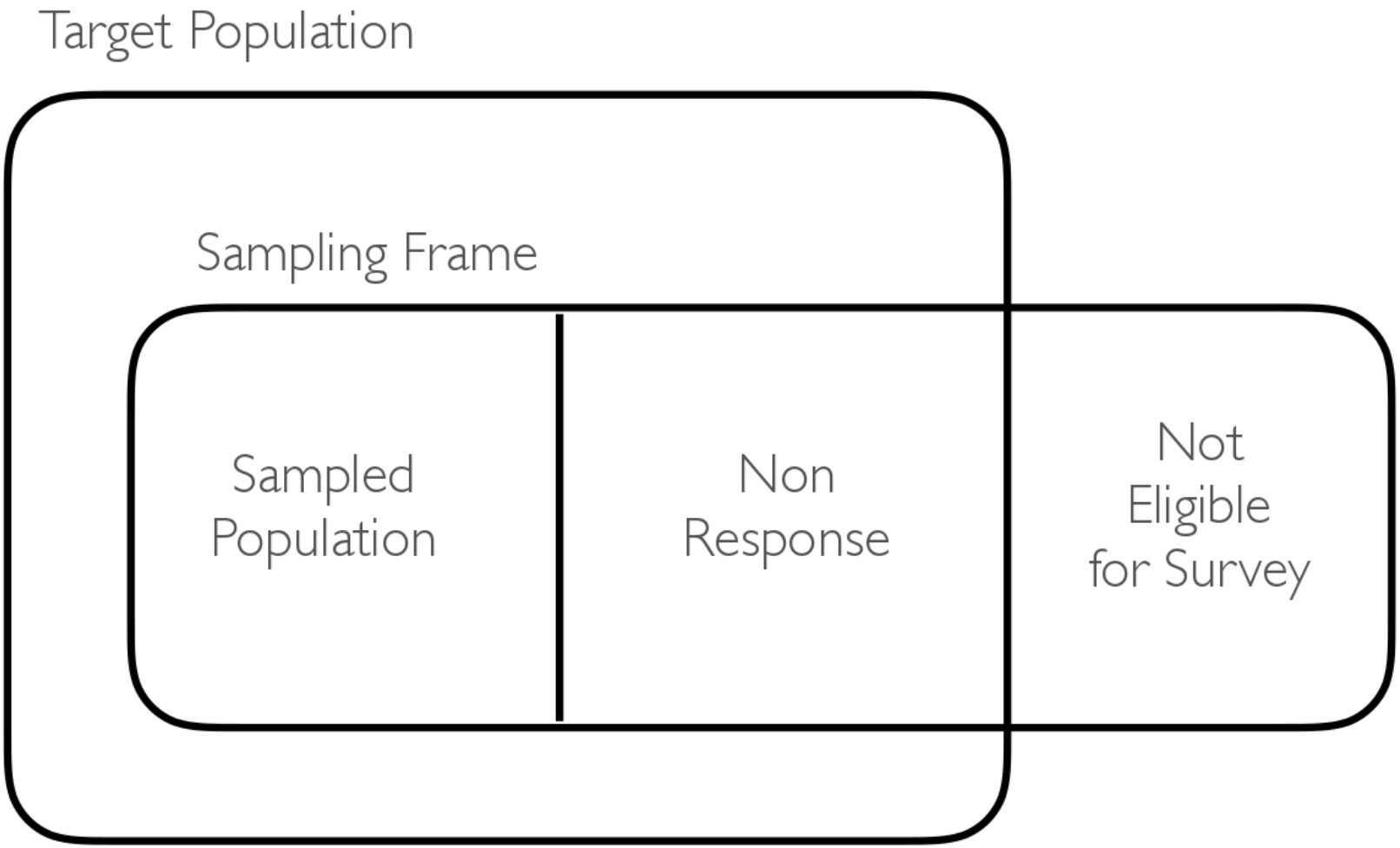
- Convenience: are students who show up to class representative of the whole class?
- Self-selection: are students who volunteer to be in a sample representative of the whole class?
- Quota: is someone's major an important/relevant attribute to try and build into our sample?
- Judgment: is my judgment biased?

Study Errors

An important part of survey design is identifying the population you're interested in. Unfortunately, this is often done long before a statistician gets involved.

> *To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of.*
>
> *Ronald Fisher*

Study Errors



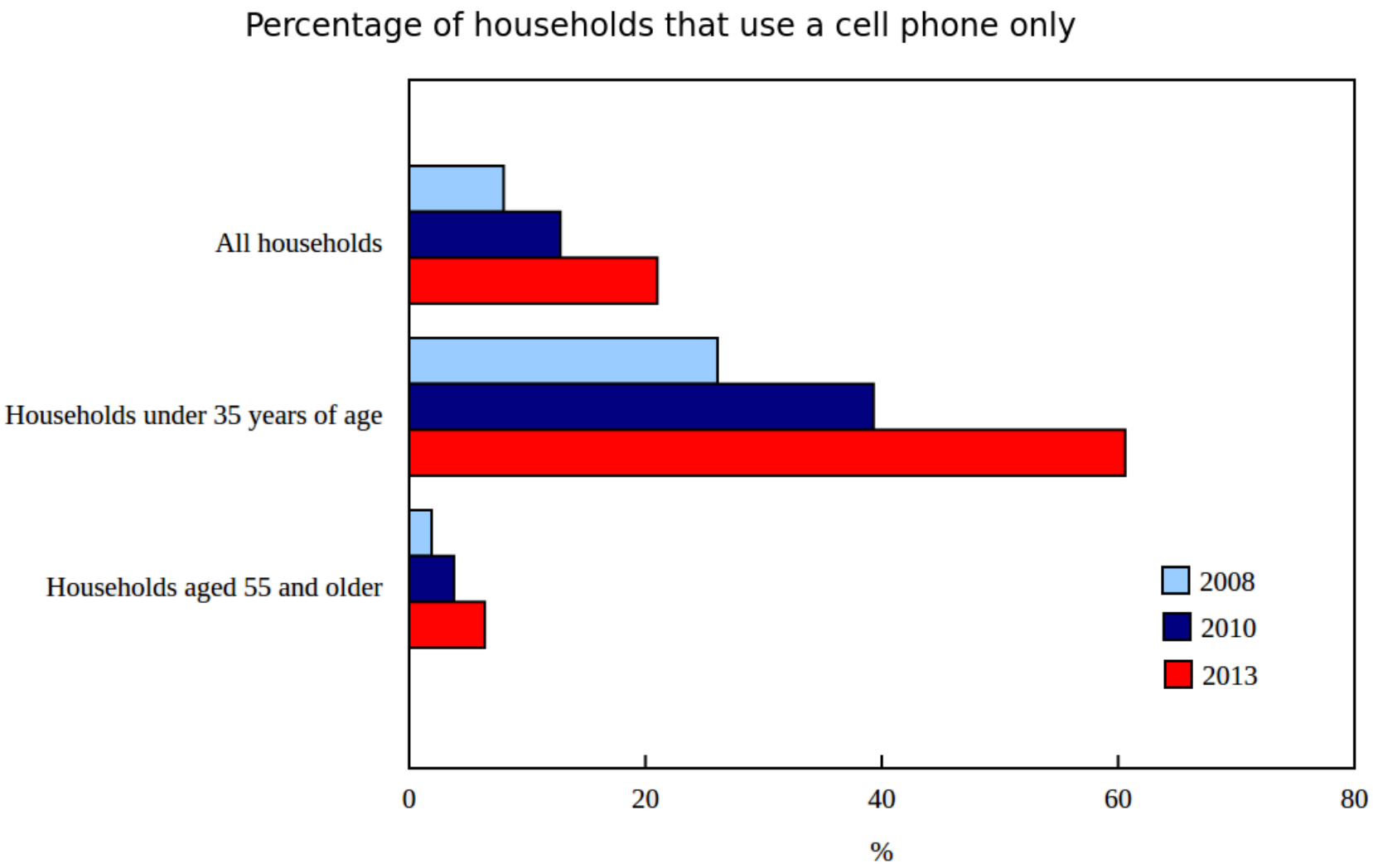
When sampling, we can usually classify errors as one of three types:

Study Errors

Remember: we're trying to estimate a population attribute (e.g. the mean age of students in the class). When sampling to do this, we can usually classify errors as one of three types:

- Frame error:
- Sample error:
- Measurement error:

Frame Error Example



Source: Statistics Canada, 2013

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).

Sample Error

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

- They're likely to encounter **non-response** (more on this later in the course).
- If certain voters are more/less likely to respond to polling, this can lead to sample error.

Measurement Error

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 (e.g. 'shy' voters).
- 'Leading' questions. e.g. 'Yes Prime Minister'
- Interviewers could affect response.
- Questions using forced choice rather than agree/disagree questions as people tend to agree with any statement regardless of the content.

Measurement Error

In 2016 the UK held a referendum on leaving the European Union. The original referendum question was:

"Should the United Kingdom remain a member of the European Union?"

but this wording could be judged to be helpful for those who wanted the UK to remain in the EU. The UK's Electoral Commission suggested changing the wording to:

Measurement Error

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

e.g. phrasing questions fairly, 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 STAT 332, but you should still keep it in mind!

STAT 332 Assumptions

Unless otherwise stated, in STAT 332 we will make the following assumptions:

-
-
-

Note: you will still be expected to be able to identify potential instances of frame, sample and measurement error.