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. students who show up to a class
- e.g. a survey of people walking on the street.

Self-selection sampling

Self-selection sampling: units choose themselves.

- e.g. if I ask the class for volunteers for my sample
- e.g. many internet polls

Quota sampling

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

- e.g. Marketing survey or panel

- e.g. if 50% were CS, 30% were math 20% we STAT majors.
  I choose or pick a sample ~~so~~ to match this distribution

Judgment sampling

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

- e.g. I choose a sample to match the previous distribution but I guess what major students are.

Study Errors

Non-probability sampling protocols have lots of obvious problems.

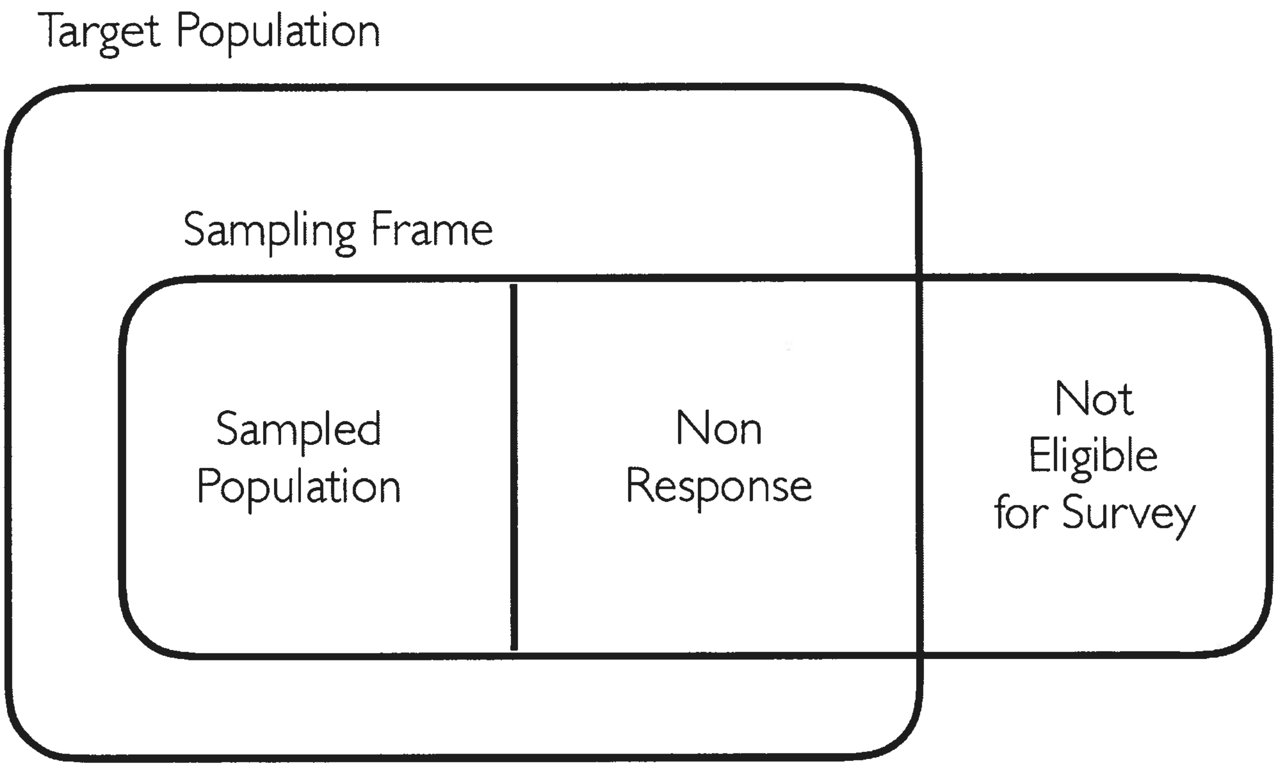
- Convenience: are students who show up to class representative of the whole class?
- Self-selection: are students who volunteer to be in a sample representative of the whole class?
- Quota: is someone's major an important/relevant attribute to try and build into our sample?
- Judgment: is my judgment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



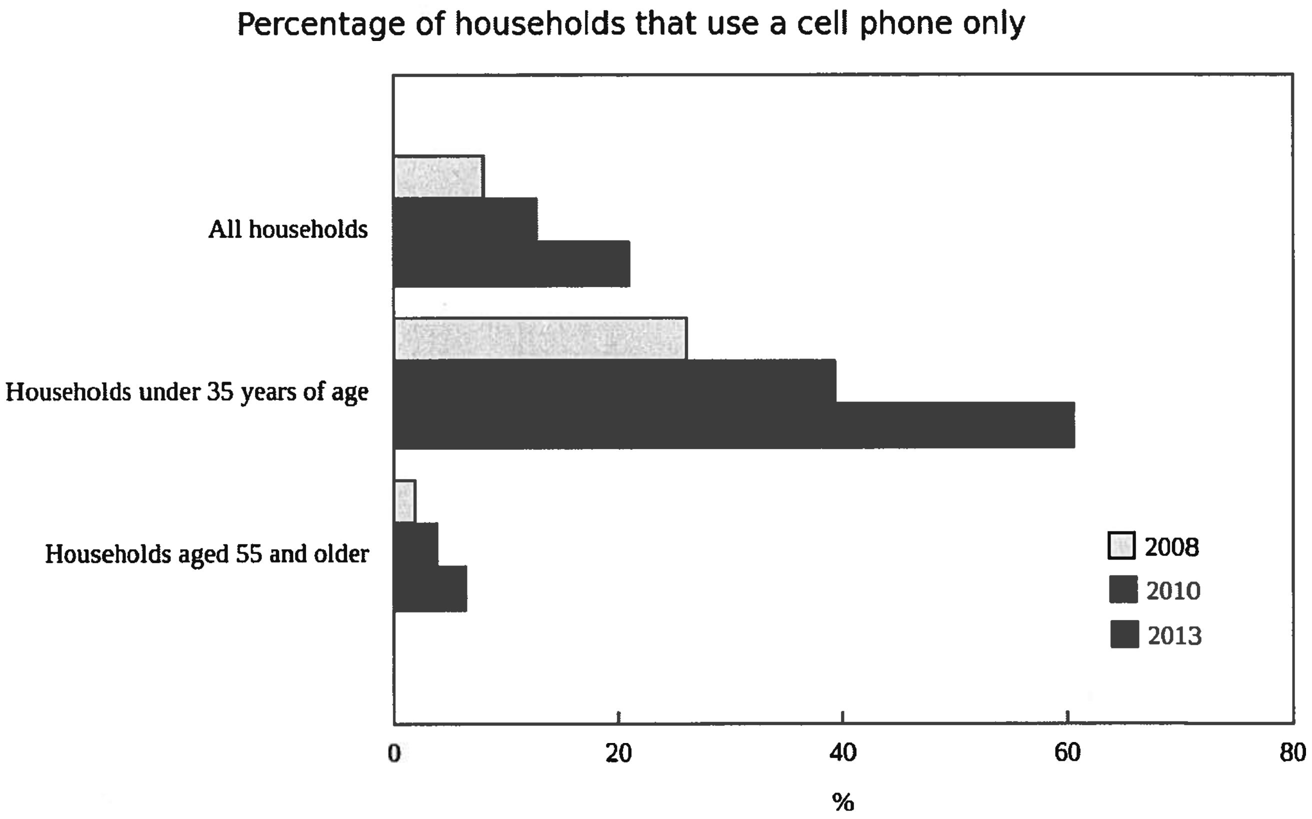
When sampling, we can usually classify errors as one of three types:

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: the difference in attributes between the target population and sampling frame
- Sample error: the difference in attributes of interest between the sampling frame and sample
- Measurement error: when true and measured values of the variates on the units in the sample are different.

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. <u>'Yes Prime Minister'</u>
- 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:

Should the United Kingdom remain a member of the European Union or leave the European Union?

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:

- the sampling frame is complete (i.e. it contains everyone in the target population
- there is no non-response.
- Measurements are accurate

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

STAT 332 - Probability Sampling

- Notation
- Probability Sampling Protocols
- Toy Example
- Sampling Estimators
- Course Notes Coverage: Beginning of Chapter 5.

Probability Sampling

Probability sampling: selecting units for the sample based on a probability model.

- Major advantage: we can understand the probabilistic mechanism we used to form the sample and we can assess the sample error mathematically.
- 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.

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.

Probability Sampling

Here are a few probability sampling techniques, some of which we'll cover in detail:

- Simple random sampling without replacement.
- Simple random sampling with replacement.
- Systematic sampling.
- Cluster sampling.
- Stratified sampling.
- Ratio and regression estimation.

Notation:

- N: # of units
- $U = \{1, \dots, N\}$: the sampling frame
- s: our sample. A ~~subes~~ subset of U
  $s \subset U$
- n: sample size
- y_i: the value of the response variate on unit i.

Sampling Protocol:

A **sampling protocol** or **sampling design** is the mechanism by which we choose our samples.

- The design is determined by assigning to each possible sample s the probability P(s)
- Let D be the set of all possible sample s with $P(s) > 0$
- $\sum_{s \in D} P(s) = 1$
- A design can equivalently be described by a step-by-step procedure.

Design Space Example

e.g. if $N = 3$, we have $U = \{1, 2, 3\}$ and consider the set of distinct samples:

- $S_1 = \{1\}$, $S_2 = \{2\}$, $S_3 = \{3\}$ $n=1$
- $S_{12} = \{1, 2\}$, $S_{13} = \{1, 3\}$, $S_{23} = \{2, 3\}$ $n=2$
- $S_{123} = \{1, 2, 3\}$ $n=3$ census.

Design Example

Some probability sampling protocols or sampling designs based on the previous set of samples are

- Pick ~~S_1, S_{123}~~ S_{12}, S_{13}, S_{23} each with prob. equal to $1/3$

  random sampling without replacement

- ~~S_1, S_2~~

  Pick S_1, S_{12}, S_{13}, S_{123} each with prob. $1/4$

  This design has varying sample size.