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.

- What is the month-by-month unemployment rate in Canada?
- What proportion of UW students eat on campus
- How many STAT students are awake at 8:30 am

Key aspect: our population is finite

In theory, we could conduct a census and survey everyone of interest but this is often ~~impractical~~ impractical

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: An object or individual that we could take a measurement on.

Target Population: A collection of units we want to study. Denote by $U = \{1, 2, \ldots, N\}$

- Students in STAT 332
- Population of Canada
- Canadian households, farms or business.
- Tax files

Note: a 'unit' is not necessarily a single person!

Terminology

Sample Population: A collection of units which we could sample

- Americans with a landline
- STAT 332 students who show up to class.

Sampling Frame: The list of units we could sample

- A telephone directory.
- STAT 332 attendance list

Sampling Unit: A unit we actually sample.

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: Individual UW undergraduate students.
- The target population: All UW undergraduate students
- The sample population: Students who volunteered during orientation week
- The sampling frame: A list of email addresses of students who volunteered.
- The sampling units: The 100 students selected from the list.

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 population size
- The study variable or response of interest is : y_i

  e.g. income, size of a farm

Some population parameters of interest are:

- the population average $\mu = \frac{1}{N}\sum_{i=1}^{N} y_i$
- the population total $\tau = \sum_{i=1}^{N} y_i = N\mu$
- the population variance

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \mu)^2$$

Population Proportion

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

- yes or no we have a good President
- small farm i.e. farm less than 231 acres.

The study variable is an indicator variable

$$z_i = \begin{cases} 1 & \text{if } \cancel{\text{true}} \text{ yes} \\ 0 & \text{otherwise} \end{cases}$$

the population total is $M = \sum_{i=1}^{N} z_i$ = # units with yes

and the population average is a proportion denoted by π

$$\mu_z = \frac{1}{N}\sum_{i=1}^{N} z_i = \frac{M}{N} = P$$

$\mu_z = \pi$

Variance Property for any response

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i-\mu)^2$$

$$= \frac{1}{N-1}\sum_{i=1}^{N}(y_i^2 - 2y_i\mu + \mu^2)$$

$$= \frac{1}{N-1}\left[\sum_{i=1}^{N}y_i^2 - 2\mu\sum_{i=1}^{N}y_i + N\mu^2\right]$$

$$= \frac{1}{N-1}\left[\sum_{i=1}^{N}y_i^2 - 2\mu N\mu + N\mu^2\right]$$

$$= \frac{1}{N-1}\left[\sum_{i=1}^{N}y_i^2 - 2N\mu^2 + N\mu^2\right]$$

$$= \frac{1}{N-1}\left[\sum_{i=1}^{N}y_i^2 - N\mu^2\right]$$

Variance Properties

For variance we have two relations.

1. For any response we have

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i-\mu)^2 = \frac{1}{N-1}\left[\sum_{i=1}^{N} y_i^2 - N\mu^2\right]$$

2. and for binary responses we have

$$\sigma_z^2 \approx p(1-p)$$

$$\sigma_z^2 \approx \pi(1-\pi)$$

for large N.

Variance Property for Binary Responses

$$\sigma_z^2 = \frac{1}{N-1} \sum_{i=1}^{N} (z_i - \mu_z)^2$$

$$= \frac{1}{N-1} \left[\sum_{i=1}^{N} z_i^2 - N\mu_z^2 \right]$$

$z_i = \{0,1\}$ $\Rightarrow z_i^2 = \{0,1\}$

$$= \frac{1}{N-1} \left[\sum_{i=1}^{N} z_i - N p^2 \right]$$

$$= \frac{1}{N-1} \left[Np - Np^2 \right]$$

$$= \frac{N}{N-1} p(1-p)$$

$$\approx p(1-p) \quad \text{for large } N.$$

$$\approx \pi(1-\pi) \quad \text{for large } N$$

where $\mu_z = \pi$

Census

A **census** an investigation is where we examine every unit

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,
- cost
- time frame