Sampling Techniques

Population: The collection of all units of a specified type in a given region at a particular point or period of time. It consists of a known, finite number N of units. There is a variable of interest y that has a value (fixed, though perhaps unknown) . For each unit is the population.

The units in the population are identified and are labeled $1, 2, \cdots, N$

Sometimes we are interested in parts or segments of a population. Such segments are considered to be - Domain of Study , or Sub - population.

Element: An object on which a measurement is taken.

Population: A collection of elements about which we wish to make an inference.

Sampling unit: Non-overlaping collections of elements form the population that cover the entire population.

Frame: A list of an sampling unit.

Sample: A collection of sampling units drawn from a frame or frames.

Note:

1. Elements and sampling units might be the same $\cdots$ single element units.
2. There may be a series of frames
   (a) Frame of Towns & cities.
   (b) Frame of housing units or city block within town or city.
   (c) Frame of households within housing unit or city block.

There are two broad categories of samples

1. Random, or probability samples.
2. Non-random samples.

Non-random samples:

Often requires use of one's judgement in determining the composition of sample $\cdots$ Hence also called judgement samples or purposive samples.

Various types of Non-random samples

- Quota samples: Units selected to get a specified composition.

  e.g. 40% males 60% females

  Quota samples don't work well because

  - they often involve subjective judgement of an interviewer.
  - there are many variables that could be used to define quotas.

    Which should be used?

    What is the effect of those not used?

- Volunteer samples - units volunteer $\cdots$ why?
- Haphazard samples - use units that happen to be available.

Errors in survey :

Sampling error - Arises because only a part of population is included in survey.

Non-sampling error:

(1) Non-observation error

(2) Observation error

(1) Non-observation errors

- coverage errors

  Frame doesn't include all population units.

- Non-response errors
    - inability of contact unit in sample.
    - sampled unit can't provide information.
    - sampled unit refuses.

There are many techniques to try to reduce effect of non-response.

- Sample non-respondents.
- looking at auxiliary information about units.
- call backs & multiple mailings.
- incentives.
- Endorsements
- Protection of privacy
- Randomized response techniques.

(2) Observation errors.

`Due to :`

interviewer

Respondent

measurement instrument

method of data collection and rewarding.

Interviewer effect:

- The interviewers manner
- Male versus female interviewer

Respondent effect

- Recall
- prestige or pride

- unable to understand questions
- lying, deliberate deception

Measurement effect of data collection

- Terms not precisely defined
- Order of questions
- Types of instruments
    - telephone
    - mail
    - direct interview
    - direct measurement
- Types of questions - open or closed
- Question wording

These types of errors can have a major impact on the results of a survey.

- Sampling error $\Rightarrow$ Need good sample design.
- Non-sampling error $\Rightarrow$ Need careful planning, execution of analysis of survey results. Quality control procedures.

A good survey requires careful planning.

1. Clearly state objectives of survey.
2. Carefully define the target population.
3. Carefully select an appropriate frame(s).
4. Choose a proper sample design.
5. Determine method used for obtaining data (or measurement).
6. Determine how and what measurement are to be made.

7. Carefully select and train field workers.

8. Do a comprehensive pre-test on a small sample.

9. Prepare a detailed field work plan.

10. Decide and describe how collected data are to be handled (including quality control procedures)

11. Determine the data analysis required.

12. Prepare final report.

13. Review what happen to learn about ways to improve your next survey.

Crucial point :

Total survey error = Non-sampling error + Sampling error

As sampling size increases $\cdots$

sampling error decreases, but non-sampling error increases.

$n \uparrow$ Sampling error $\downarrow$ Non-sampling error $\uparrow$

This is why CENSUS may not be better.

In a census, there is no sampling error. But the non-sampling error may be large.

$\cdots$ Need to strike a balance.

Some general preliminaries :

Let $U_1, U_2, \ldots, U_N$ be the N units (or elements) of the population. Let Y_i be the value of variable y determined from unit $U_i,\ i = 1, \ldots, N$. Any function of all of the values of the population units is called a population parameter.

Examples : Population total

$$Y = \sum_{i=1}^{N} Y_i = Y_1 + Y_2 + \cdots + Y_N$$

Mean $\overline{Y} = \dfrac{Y}{N} = \displaystyle\sum_{i=1}^{N} \dfrac{Y_i}{N}$

Let

$$Y_i = \begin{cases} 1 & , \text{ if } U_i \text{ has some given characteristic.} \\ 0 & , \text{ if } U_i \text{ does not have the characteristic.} \end{cases}$$

$P = \dfrac{Y}{N} = \dfrac{\sum_{i=1}^{N} Y_i}{N}$ is the population proportion.

Variance:

Murthy defines $\sigma^2 = \dfrac{\sum_{i=1}^{N}(Y_i - \overline{Y})^2}{N}$.

Cochran defines $S^2 = \dfrac{\sum_{i=1}^{N}(Y_i - \overline{Y})^2}{N-1}$.

(Murthy defines this as σ'^2) Standard deviation $\Rightarrow \sigma, \sigma'$ or S

Coefficient of Variation $C(y) = \dfrac{\sigma}{\overline{Y}}$

If there is a second measurement x made on each unit,

X_i or $U_i,\ i = 1, \ldots, N$, we define population ratio $R = \dfrac{\overline{Y}}{\overline{X}}$

Random Sample (Cochran)

A sample obtained according to sampling procedure which has the following properties:

1. We can define the set of distinct samples $S_1, S_2, \ldots, S_v$ which the procedure is capable of selection when applied to our population.
2. Each possible sample S_i is assigned a known probability of selection π_i.
3. One of the S_i is selected by a random process in which each S_i receives its proper π_i.
4. Method for computing estimate from the selected sample must be stated and must lead to a unique estimate for any specific sample.

The specification of all possible samples of a given type with their corresponding probabilities is called the *Sample design.*

Sampling units selected into the sample are called *Sample units.*

N population size

n Sample size

$\frac{n}{N}$ Sampling fraction

Sample units $\{u_1, u_2, \ldots, u_n\}$

We observe the variable y as y_i on $u_i,\ i = 1, \ldots, n$.

Sample measurements are $\{y_1, y_2, \ldots, y_n\}$

Consider the random sample selected according to a sample design : $\{y_1, y_2, \ldots, y_n\}$

Any function of these sample values that is free unknown population parameters is called a *statistic*. An *estimator* is a statistic obtained by a specified procedure for estimating a population parameter.

Let t be an estimator of a parameter θ. Let t_i be the value of t based on the sample $i,\ i = 1, \ldots, M_0$ where M_0 is the total number of possible samples from our sample design. The expected value of t is given by $\mathrm{E}(t) = \sum_{i=1}^{M_0} t_i \pi_i$.

If $\mathrm{E}(t) = \sum_{i=1}^{M_0} t_i \pi_i = \theta$, then t is an unbiased estimator of θ.

If $\mathrm{E}(t) \neq \theta$, then the bias of t is $\mathrm{B}(t) = \mathrm{E}(t) - \theta$.

Mean square error (MSE):

$\mathrm{M}(t) = \mathrm{E}\{t - \theta\}^2 = \sum_{i=1}^{M_0} (t_i - \theta)^2 \pi_i$

$\mathrm{M}(t) = \mathrm{E}\{t-\theta\}^2 = \mathrm{E}\{t-\mathrm{E}(t)+\mathrm{E}(t)-\theta\}^2 = \mathrm{E}\{t-\mathrm{E}(t)\}^2+\{\mathrm{E}(t)-\theta\}^2 = \mathrm{V}(t)+\mathrm{B}^2(t)$,

where $\mathrm{V}(t) = \mathrm{E}\{t - \mathrm{E}(t)\}^2$ is the variance of t.

MSE = Variance + (Bias)^2 if $\mathrm{B}(t) = 0 \Rightarrow$ MSE = $\mathrm{V}(t)$

The variance of t is $V(t) = \mathrm{E}\{t-\mathrm{E}(t)\}^2$ & if t is unbiased for θ, then $V(t) = \mathrm{E}\{t-\theta\}^2$

Also , $V(t) = \mathrm{E}\{t - \mathrm{E}(t)\}^2 = \mathrm{E}(t^2) - \{\mathrm{E}(t)\}^2$

Useful result : $\mathrm{E}(t^2) = V(t) + \{\mathrm{E}(t)\}^2$

We might want to find an unbiased estimator of $V(t)$.

An interesting result:

Suppose the sample is obtained as two or more subsamples each drawn according

to the sampling scheme. Suppose each subsample provides a valid estimate of the parameter $\cdots$ These subsamples are called Inter Penetrating Subsamples. Subsamples need not be independent. But sometines it is helpful.

Let $t_1, t_2, \ldots, t_k$ be k estimators of θ. Let $\mathrm{E}\{t_i\} = \theta,\ i = 1, \cdots, k$. And let the t_i be independent.

Let $\bar{t} = \frac{1}{k}\sum t_i$ ($\bar{t}$ is also an estimator of θ). Then

1. $\mathrm{E}(\bar{t}) = \theta$

   $\mathrm{E}(\bar{t}) = \mathrm{E}\{\frac{1}{k}\sum_{i=1}^{k} t_i\} = \frac{1}{k}\sum_{i=1}^{k}\mathrm{E}(t_i) = \frac{1}{k}\sum_{i=1}^{k}\theta = \theta$

2. Let $v(\bar{t}) = \frac{1}{k}\frac{\sum_{i=1}^{k}(t_i - \bar{t})^2}{k-1}$, then

$$\begin{aligned}
\mathrm{E}[v(\bar{t})] &= \frac{1}{k(k-1)}\mathrm{E}\left[\sum(t_i - \bar{t})^2\right] \\
&= \frac{1}{k(k-1)}\mathrm{E}\left\{\sum t_i^2 - k\bar{t}^2\right\} \\
&= \frac{1}{k(k-1)}\left\{\sum_{i=1}^{k}\mathrm{E}(t_i^2) - k\mathrm{E}(\bar{t}^2)\right\} \\
&= \frac{1}{k(k-1)}\left\{\sum_{i=1}^{k}[V(t_i) + \theta^2] - k[V(\bar{t}) + \theta^2]\right\},
\end{aligned}$$

   since t_i are unbiased and $V(t_i) = \mathrm{E}(t_i^2) - \theta^2$.

   So ,

$$\begin{aligned}
\mathrm{E}\{v(\bar{t})\} &= \frac{1}{k-1}\left\{k\left[\frac{\sum V(t_i)}{k^2}\right] + \theta^2 - V(\bar{t}) - \theta^2\right\} \\
&= \frac{1}{k-1}\{kV(\bar{t}) - V(\bar{t})\} = V(\bar{t})
\end{aligned}$$

   Thus, $v(\bar{t}) = \frac{1}{k(k-1)}\sum_{i=1}^{k}(t_i - \bar{t})^2$ is unbiased for $V(\bar{t})$.

We have an unbiased estimator for θ and we have an unbiased estimator for the variance of the estimator of θ.