

By pure Chance

The merits of probability sampling

Matthias Ganninger

gesis

Leibniz Institute
for the Social Sciences

RECSM, UPF — Barcelona, Spain — June 1, 2011

Definitions:

Sampling refers to the methods and procedures required to obtain a subset of elements (sample) from a larger set of elements (population)

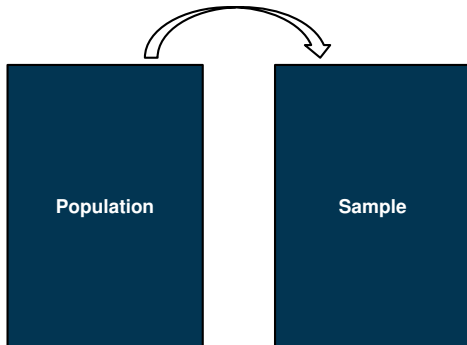
or, more broadly, following Smith, T.M.F. (2001, 167)

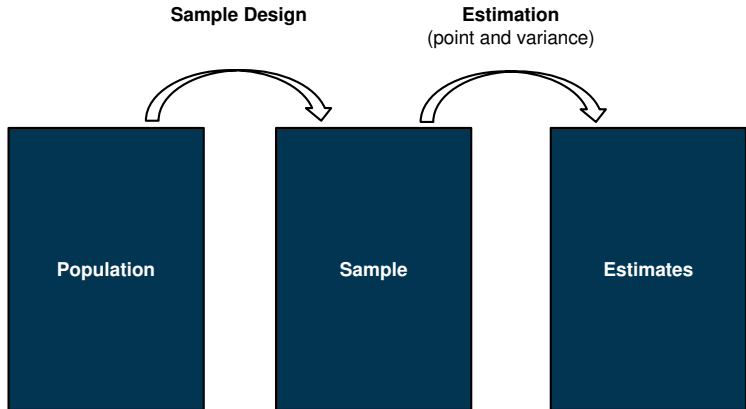
“Statistical theory deals with the relationship between samples and populations, and in this sense sampling embraces the whole of statistical inference.”

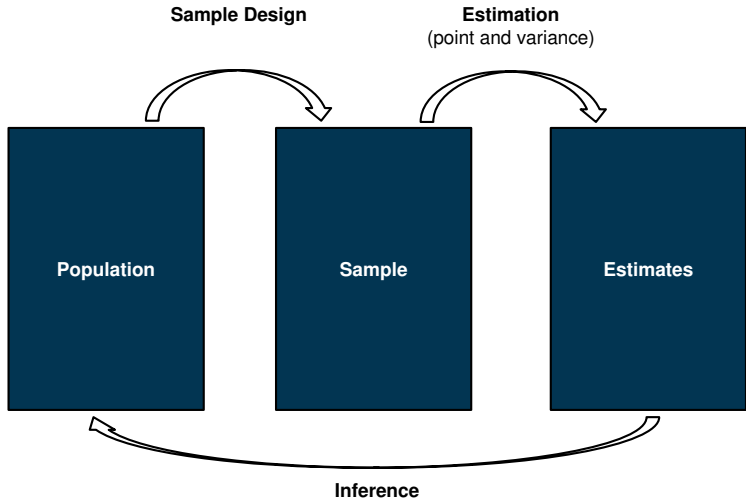


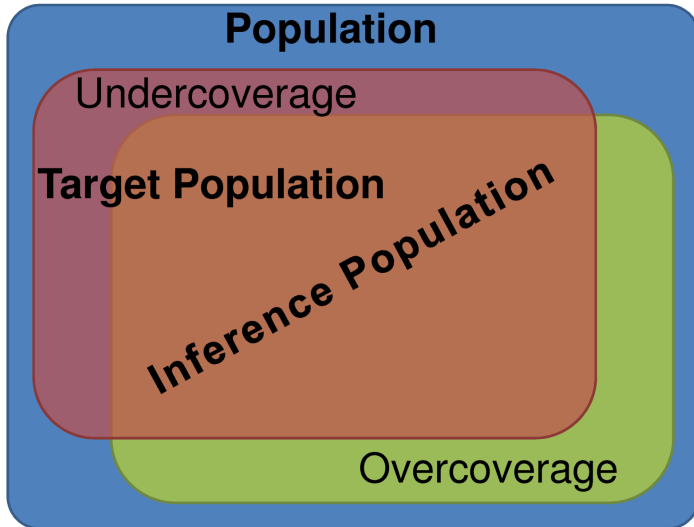
Population

Sample Design









- Each element of the population has a fixed value on the study variable
- The study variable can be surveyed only at the elements of the sample
- In many cases, surveying the **whole population** is either too **expensive**, **time consuming**, **destructive** or not possible at all
- There exist different methods to obtain a sample (**probability** and **non-probability** sample designs)

Population

i	Z_i	h
1	20	2
2	15	1
3	10	1
4	30	2
5	50	2
6	15	1
7	10	1
8	25	2
9	10	1
10	15	1

Stratification

- Define stratum boundaries as $h = \{[1; 19], [20]\}$
- Allocate the sample size $n = 5$ proportionally

$$n_1 = 5 \cdot \frac{6}{10} = 3$$

$$n_2 = 5 \cdot \frac{4}{10} = 2$$

- Calculate $\pi_i = \frac{n_h}{N_h}$
- Problem: how are π_i defined? Based on theoretical or empirical n_h ?
- Draw n_h elements by srs from the strata

- Assume we had sample elements $s = \{2, 3, 6, 4, 8\}$
- The estimated mean is

$$\hat{y} = \sum_{h=1}^H \frac{N_h}{N} \bar{y}_h = 0.6 \cdot 13.3 + 0.4 \cdot 27.5 = 19$$

and the estimated variance is

$$\begin{aligned} \hat{V}(\hat{y}) &= \sum_{h=1}^H \left(\frac{N_h}{N} \right)^2 \frac{S_h^2}{n_h} \cdot 1 - \left(\frac{n_h}{N_h} \right) \\ &= \left[0.36 \cdot \frac{8.3}{3} \cdot \left(1 - \frac{3}{6} \right) \right] + \left[0.16 \cdot \frac{12.5}{2} \cdot \left(1 - \frac{2}{4} \right) \right] \\ &= 1 \end{aligned}$$

- Comparison srs vs. strrs:

$$\begin{aligned} \hat{V}(\hat{y})^{(\text{srs})} &= \frac{S^2}{n} \left(1 - \frac{n}{N} \right) = \frac{140}{5} \cdot 0.5 = 14 \\ \hat{V}(\hat{y})^{(\text{strrs})} &= \sum_{h=1}^H \left(\frac{N_h}{N} \right)^2 \frac{S_h^2}{n_h} \cdot 1 - \left(\frac{n_h}{N_h} \right) = 7.3\bar{6} \end{aligned}$$

$$\begin{aligned}
 deff_{\text{str}}(\hat{t}) &= \frac{\text{within stratum variance}}{\text{total variance}} \\
 &= \frac{\sum_{h=1}^H \frac{N_h}{N} S_h^2}{\sum_{h=1}^H \frac{N_h}{N} [s_h^2 + (\bar{Y}_h - \bar{Y})^2]}
 \end{aligned}$$

In an example with a US Census population, ?, 63 shows that with proportional allocation

$$deff_{\text{str}}(\hat{t}) = 0.79$$

In addition to the analysis from survey data we need to investigate the quality of the results. The aim is to generalize results from the sample to the universe.

→ inferential statistics

Sampling function and estimator

Properties of estimators

- unbiasedness
- efficiency
- normality
- large sample properties (limit theorems)
- small sample properties
- Test procedures based on survey data

- We have the inclusion probabilities π_i
- Elements with a **low a priori probability** to be in the sample receive a **high design weight** and vice versa
- The HT estimator of the population mean is

$$\hat{y}_{HT} = \frac{1}{N} \sum_{i=1}^n \frac{y_i}{\pi_i}$$

and its variance

$$V\left(\hat{y}_{HT}\right) = \frac{1}{N^2} \left[\sum_{i=1}^n \frac{1 - \pi_i}{\pi_i^2} \cdot y_i^2 + \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \frac{\pi_{i,j} - \pi_i \pi_j}{\pi_{i,j} \cdot \pi_i \cdot \pi_j} \cdot y_i \cdot y_j \right]$$

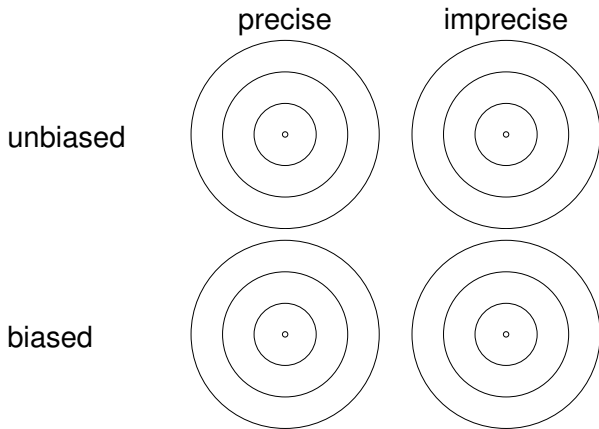
Point Estimator:

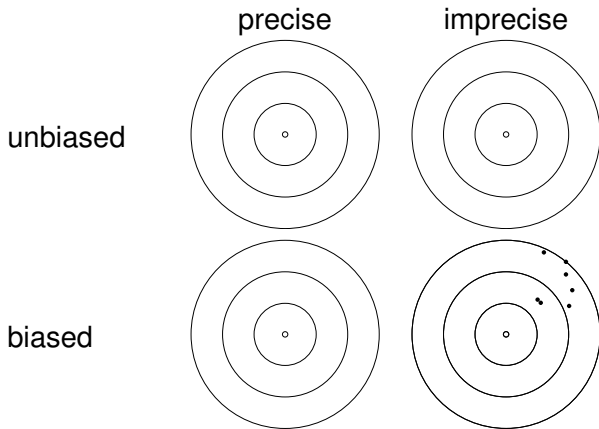
$$\hat{T}^{\text{GREG}} = \hat{T}_Y + (T_X - \hat{T}_X)' \hat{\beta}$$

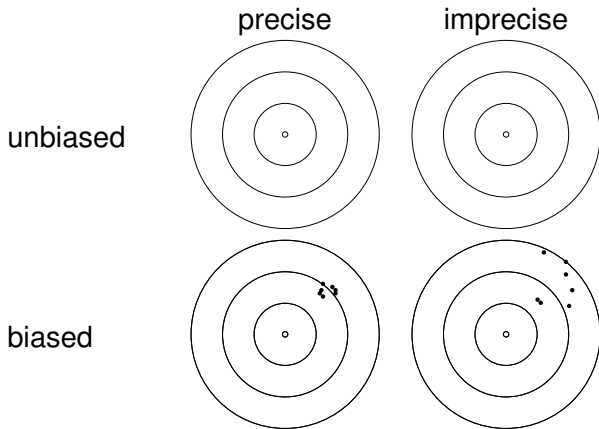
Variance:

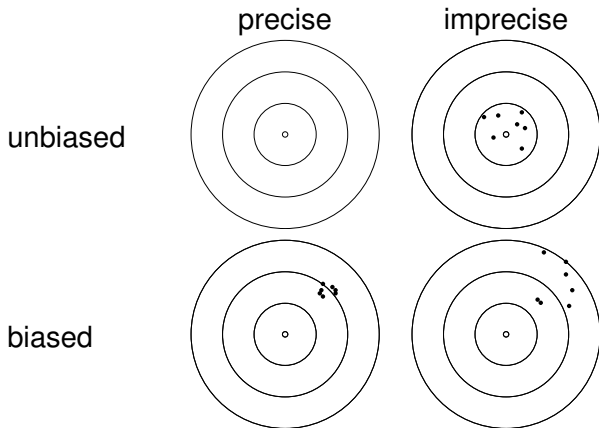
$$\text{Var}(\hat{T}^{\text{GREG}}) = \sum_{h=1}^H N_h^2 \cdot \frac{S_h^2}{n_h} \cdot \left(1 - \frac{n_h}{N_h}\right) \cdot (1 - \rho^2)$$

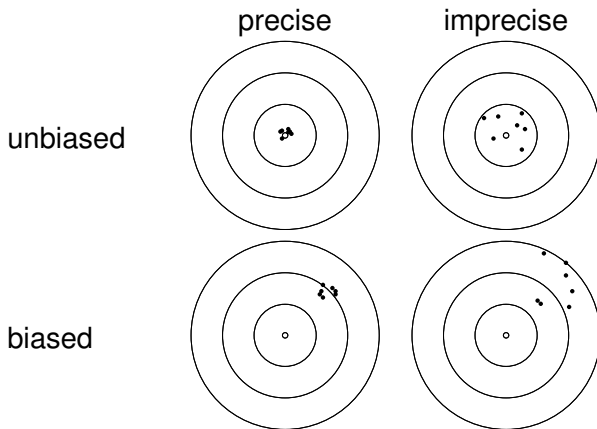
where $S_{h,Y}^2$ is the variance of the study variable in stratum h and ρ the correlation of the auxiliary variable and the study variable.











- No explicit statistical error distribution

- No explicit statistical error distribution
- Quotas defined on loose ground

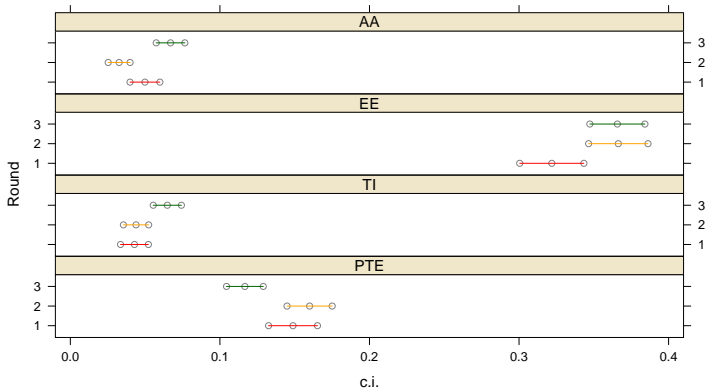
- No explicit statistical error distribution
- Quotas defined on loose ground
- Interviewer bias due to decision who belongs to which quota

- No explicit statistical error distribution
- Quotas defined on loose ground
- Interviewer bias due to decision who belongs to which quota
- Selection within quote not random

- No explicit statistical error distribution
- Quotas defined on loose ground
- Interviewer bias due to decision who belongs to which quota
- Selection within quote not random
- Interviews often conducted on streets or in office

- No explicit statistical error distribution
- Quotas defined on loose ground
- Interviewer bias due to decision who belongs to which quota
- Selection within quote not random
- Interviews often conducted on streets or in office
- Overall very little control

confidence intervals by round and variable

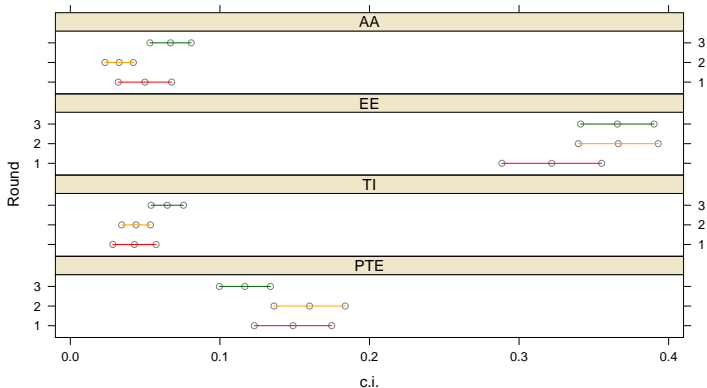


PTE: Technicians and Associate Professionals TI: Independent worker

EE: Executant Workers

AA: Wage earning farmer or agricultural worker

confidence intervals by round and variable (deff)



PTE: Technicians and Associate Professionals TI: Independent worker

EE: Executant Workers

AA: Wage earning farmer or agricultural worker



That's all Folks!