Sampling & non-response

Many figures taken from
Biemer & Lyberg (2003)

Selected slides borrowed from
Juan Muñoz, World Bank

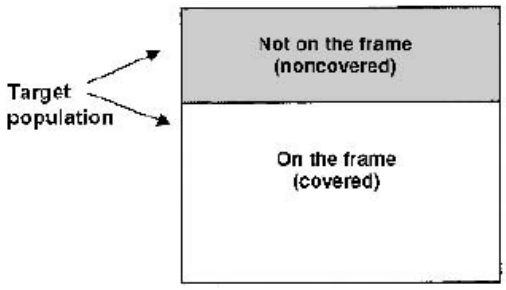
Topics

- Coverage error // unit non-response error
- Sampling frame
  - Auxiliary information
- Sampling method
  - Simple random sampling (unbiased & un-efficient)
  - Sampling error
  - Stratified, multistage & cluster
- Pre-test

Sampling

- Random/ scientific sampling
  - Each element of the population has a known, positive probability of being included in the sample
  - Allows calculation of sampling errors and confidence intervals
- Elements vs. units
- Requires a sampling frame
  - An actual list
  - A conceptual or implicit list

Coverage error



$\bar{Y}_{nc}$ = mean of the noncovered population

t_{nc} = proportion of the population not covered

$\bar{Y}_{c}$ = mean of the population covered

t_{c} = proportion of the population covered

$\bar{Y}_{p} = t_{c}\bar{Y}_{c} + t_{nc}\bar{Y}_{nc}$

Figure 3.1 Basic coverage problem. Only the unshaded region of the box is covered by the sampling frame. The shaded region is not covered and is therefore missed by the sample. To the right of the box are the components that make up the bias due to noncoverage error.

$$\mathrm{RB}_{nc} = (1 - t_c)\frac{\bar{Y}_c - \bar{Y}_{nc}}{\bar{Y}_p}$$

Coverage error

Sampling frame

- Incomplete (and missing are different)
- One to many, or many to one relationship between population and frame
- Cost increased when sampling frame is larger than target population

Reducing coverage bias

- Better frame (perhaps through combination)
- Dual frame (and identify which frame belong to)
- Post-adjustment based on external data source (with prior planning to match questions)

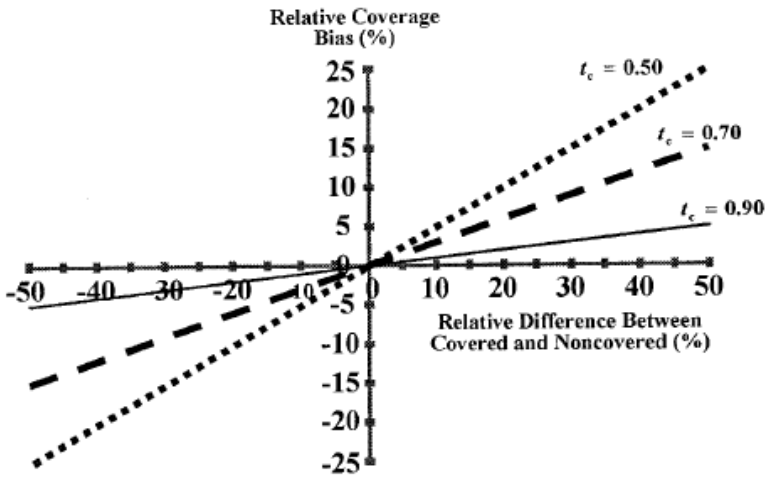


Figure 3.3 Coverage bias as a function of t_c and the relative difference between $\overline{Y}_c$ and $\overline{Y}_{nc}$.

Sampling

SRS (same probability p = n / N, independent)

- Random number generator
- Lottery
- Randomized list and systematic sampling

Advantages?

- Simple
- Easy to explain
- Equal probability → self-weighted

Disadvantages?

What's not random sampling

- Convenience
  - Aggregation Point Intercept Sampling
- Purposive
- Snowball Sampling
- Quota
  - Random Disproportionate Stratified Sampling

Random sample

- Woodland owners in associations
- NC residents
- (Hispanic) visitors to zoo

Table 9.1 Income for a Small Population of 10 Persons

Person in the Population	Actual Income
P1	$Y_1 = 60{,}000$
P2	$Y_2 = 72{,}000$
P3	$Y_3 = 94{,}000$
P4	$Y_4 = 90{,}000$
P5	$Y_5 = 102{,}000$
P6	$Y_6 = 116{,}000$
P7	$Y_7 = 130{,}000$
P8	$Y_8 = 135{,}000$
P9	$Y_9 = 141{,}000$
P10	$Y_{10} = 160{,}000$
Mean	$\bar{Y} = 110{,}000$

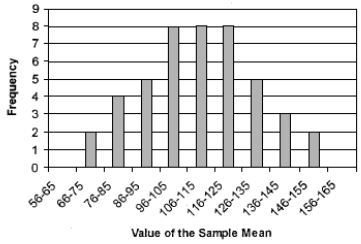


Figure 9.1 Sampling distribution of the sample mean for all possible samples of size 2 selected from the artificial population in Table 9.1.

Sampling error and sample size

Standard error e when estimating a prevalence P in a sample of size n taken from an infinite population

$$e = \sqrt{\frac{P(1-P)}{n}}$$

Confidence intervals

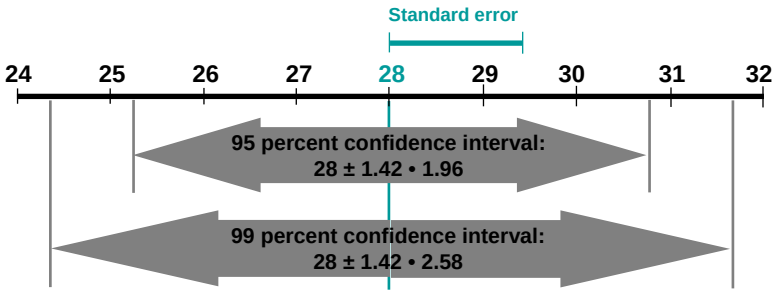
In a sample of 1,000 households, 280 households (28 percent) have preschool children.

$$e = \sqrt{\frac{0.28 \times 0.72}{1{,}000}} = 0.0142$$

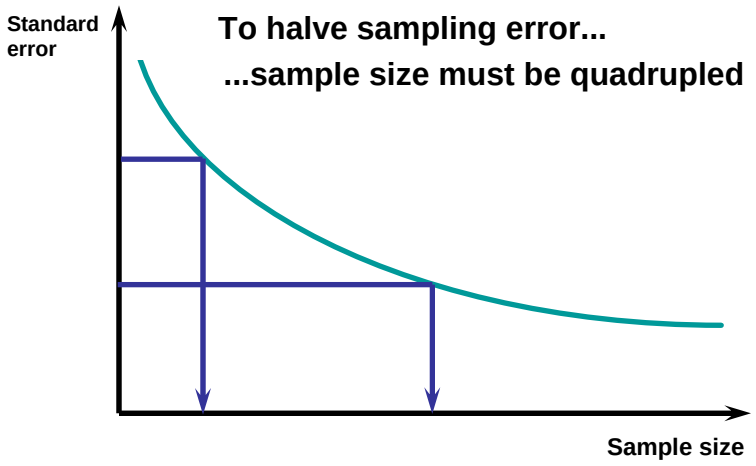
Standard error is 1.42 percent.

Confidence intervals

In a sample of 1,000 households, 280 households (28 percent) have preschool children. Standard error is 1.42 percent.



Sampling error and sample size



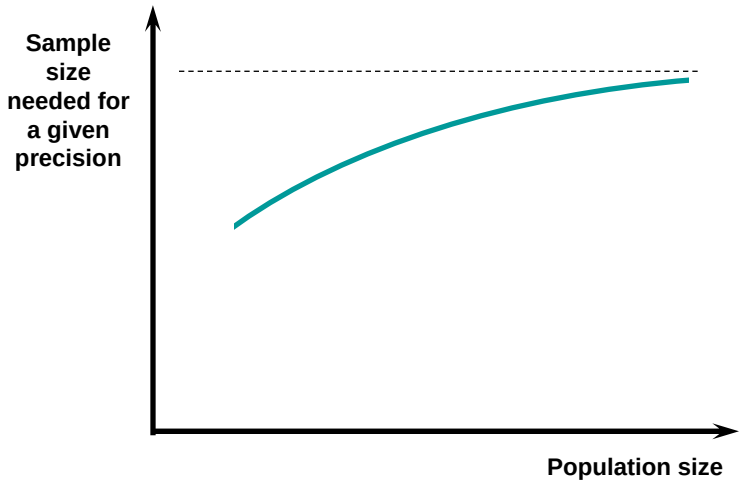
Sample size and population size

Standard error e when estimating a prevalence P in a sample of size n taken from a population of size N

$$e = \sqrt{1 - \frac{n}{N}} \sqrt{\frac{P(1-P)}{n}}$$

finite population correction

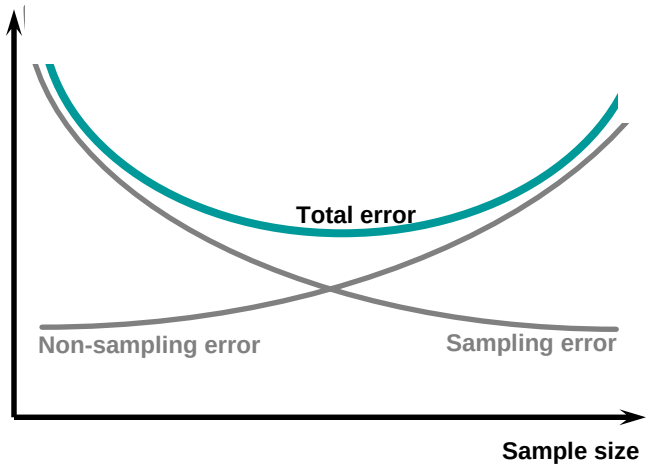
Sample size and population size



Sample size

- If you only want to know a proportion with given “margin of error” d, then
- Assume
  - maximum variance (50%)
  - infinite sample
- Required $n = 1/d^2$
- For example if d = 5%, n =400

Sampling vs. non-sampling errors



Problems with SRS

Increase precision

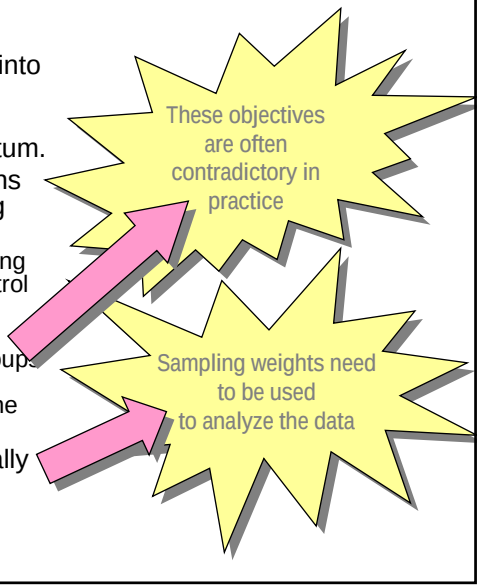
- Stratification

Reduce cost

- Multi-stage
- Cluster
- But because usually homogenous, lose precision

Stratification

- The population is divided up into subgroups or "**strata**".
- A separate sample of units is then selected from each stratum.
- There are two primary reasons for using a stratified sampling design:
  - To potentially reduce sampling error by gaining greater control over the composition of the sample.
  - To ensure that particular groups within a population are adequately represented in the sample.
- The sampling fraction generally varies across strata.



Examples of Stratification

- **Agricultural survey**
  - Agro-ecological zones
  - Land use
  - Farm size

Estimation under stratified random sampling

- Each stratum is treated as an independent population
- Estimate of stratified total is sum of stratum totals
- Estimate of stratified mean is weighted combination of stratum means
- Variance calculated independently for each stratum

Sample allocation under stratified sampling

- Three major types of sample allocation of sample units among the strata:
  - Proportional allocation
  - Equal allocation
  - "Optimum" allocation (Neyman – requires information on distribution of key variables across strata)

Proportional allocation

- The sample allocated to each stratum is proportionally to the number of units in the frame for the stratum:

$$n_h = n \times \frac{N_h}{N}$$

- Simplest form of sample allocation
- Provides self-weighting sample
- Efficient sample design when variability is similar for the different strata

Equal allocation

- Each stratum is allocated an equal number of sample units:

$$n_h = \frac{n}{L}$$

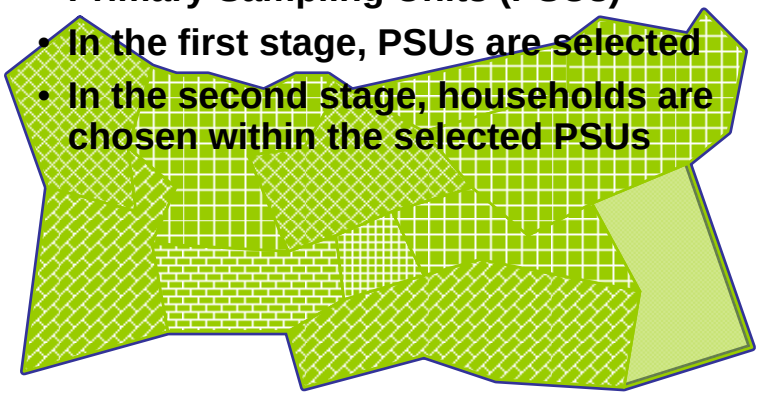
- Used when same level of precision is required for each stratum
- Example: reliable survey estimates required for each region
- Must use weights to account for different sampling fractions by stratum

Practical allocation criteria

- Allocation is often a compromise between proportional, equal and optimal; e.g. we start with a proportional allocation and then we increase the sample size in the smaller regions
- Cost consideration are also a factor; e.g., in countries with high proportion of rural population, sometimes a higher sampling rate is used for the urban stratum, to increase the urban sample size and because of the lower cost of data collection in urban areas

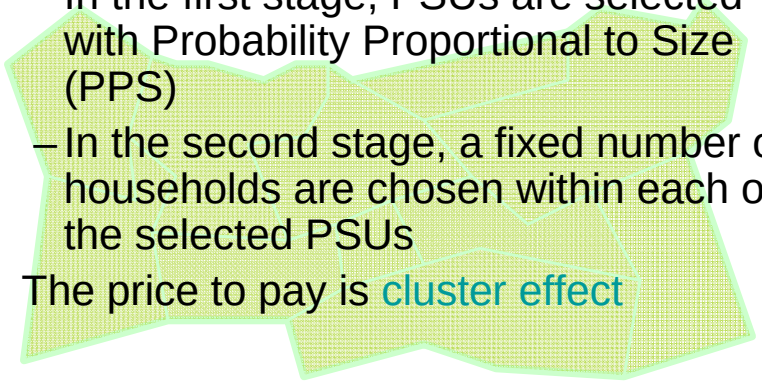
Two-stage sampling: Example from national survey

- **The country is divided into small Primary Sampling Units (PSUs)**
- **In the first stage, PSUs are selected**
- **In the second stage, households are chosen within the selected PSUs**



Two-stage sampling

- The sample can be made self-weighted if
  - In the first stage, PSUs are selected with Probability Proportional to Size (PPS)
  - In the second stage, a fixed number of households are chosen within each of the selected PSUs
- The price to pay is cluster effect



Cluster effect

Standard error grows when the sample of size n is drawn from k PSUs, with m households in each PSU ($n=k \bullet m$)

Intra-cluster correlation coefficient

$$e^2_{TSS} = e^2_{SRS}\left[1+\rho(m-1)\right]$$

Cluster effect

Two Stage Sample

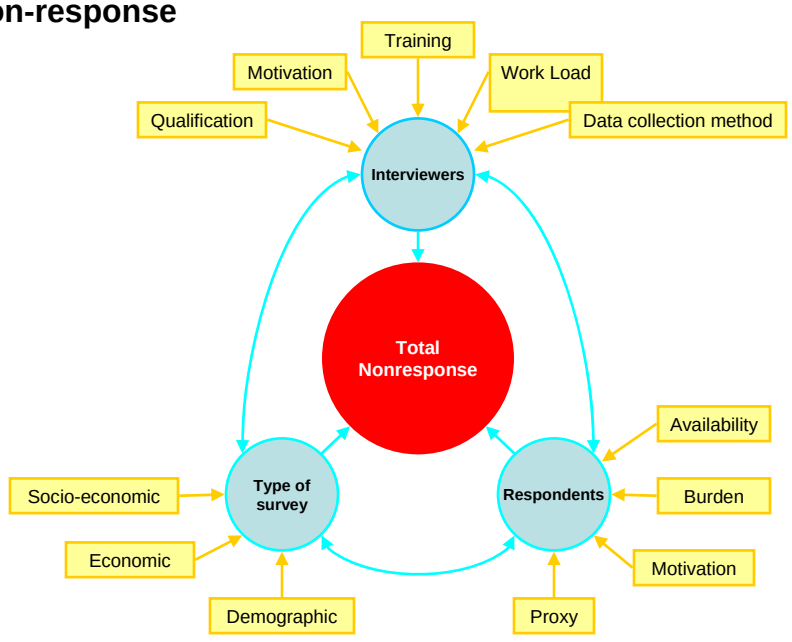
Simple Random Sample

Cluster effects

For a total sample size of 12,000 households

Number of PSUs	Number of households per PSU	Intra-cluster correlation coefficient 0.01	0.02	0.05	0.10
3000	4	1.03	1.06	1.15	1.30
2000	6	1.05	1.10	1.25	1.50
1500	8	1.07	1.14	1.35	1.70
1000	12	1.11	1.22	1.55	2.10
800	15	1.14	1.28	1.70	2.40
600	20	1.19	1.38	1.95	2.90
400	30	1.29	1.58	2.45	3.90
300	40	1.39	1.78	2.95	4.90
200	60	1.59	2.18	3.95	6.90
150	80	1.79	2.58	4.95	8.90
100	120	2.19	3.38	6.95	12.90

Non-response



Source: "Some factors affecting Non-Response." by R. Platek. 1977. *Survey Methodology*. *3*. 191-214

Unit non-response // coverage error

- Distinguish
  - No longer in frame
  - Could not contact
  - Refused

Table 3.4 Common Reasons for Nonresponse and the Type of Nonresponse It Is Likely to Generate

Reason for Nonresponse	Typical Nonresponse Category
Not motivated	Refusal
Lack of time	Refusal
Fear of being registered	Refusal
Traveling, vacation	Unable to contact
Not a good time	Refusal
Unlisted telephone number	Unable to contact
Answering machine, telephone number display	Refusal and unable to contact
Wrong address or wrong telephone number	Unable to contact
Illness or impaired	Other
Language problems	Other
Business staff changes	Refusal and unable to contact
Business owner change	Refusal and unable to contact
Business restructured	Refusal, unable to contact, and other
Survey too difficult	Refusal
Business policy not to participate in surveys	Refusal
Low priority	Refusal
Too costly	Refusal
Sensitive questions	Refusal
Boring topic	Refusal
Heavy interviewer workload	Refusal and unable to contact
Data collection period too short	Refusal and unable to contact
Screening	Refusal
Bad questions or questionnaire	Refusal
Moved	Unable to contact

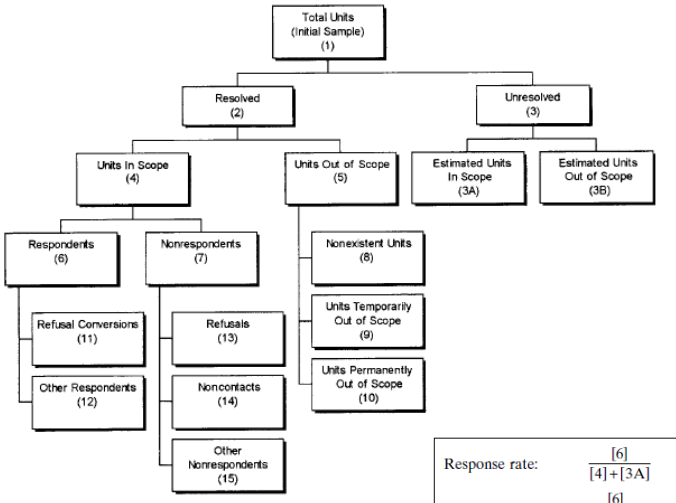


Figure 3.8 Final disposition of the sample with componer
[Adapted from Hidiroglou et al. (1993).]

Response rate: $\frac{[6]}{[4]+[3A]}$

Cooperation rate: $\frac{[6]}{[6]+[13]}$

Refusal rate: $\frac{[13]}{[4]}$

Nonresponse rate: $\frac{[7]+[3A]}{[4]+[3A]}$

Noncontact rate: $\frac{[14]+[3A]}{[4]+[3A]}$

AAPOR

http://www.aapor.org/responseratesanoverview

Standard Definitions

Final Dispositions of Case Codes and Outcome Rates for Surveys

Revised 2008

RDD Telephone Surveys
In-Person Household Surveys
Mail Surveys of Specifically Named Persons
Internet Surveys of Specifically Named Persons

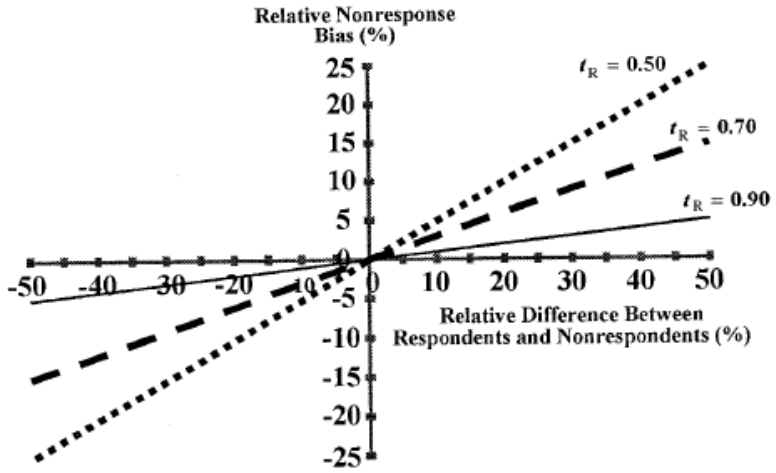


Figure 3.7 Relative nonresponse bias as a function of the response rate, t_R, and the relative difference between $\bar{Y}_R$ and $\bar{Y}_{NR}$.

Planning for non-response

- Larger sample
- Replacement

- Build in checks (for unit & item non-response)
  - Sampling frame
  - External records
  - Alternative mode screening questions
  - Alternative mode follow-up to non-respondents

Adjusting for non-response

- Weighting

- Imputation
  - Hot deck
  - Nearest neighbor
  - Strata means
  - Models

Pre-test: IRB

Research involving the use of educational tests (cognitive, diagnostic, aptitude, achievement), survey procedures, interview procedures or observation of public behavior, unless:

(i) information obtained is recorded in such a manner that human subjects can be identified, directly or through identifiers linked to the subjects; and

(ii) any disclosure of the human subjects' responses outside the research could reasonably place the subjects at risk of criminal or civil liability, or be damaging to the subjects' financial standing, employability, or reputation.

(iii) NOTE this category does not apply to research involving minors, unless the study only involves observation, and the researchers do not have any direct contact with the children.

Pre-test: IRB

- Anonymous
- Don’t interview minors
- Don’t add questions that could place respondents at risk

- Introduction (oral or written)
  - Voluntary, anonymous, # minutes, purpose is for your class assignment

Pre-test: Assignment

- Due on Friday 20 February
  - Revised questionnaire (used for pre-test)
  - Data entry template
  - At least two completed interviews
    - Questionnaires
    - Data
- Rolled over to final report
  - Descriptive statistics on total sample of $N \geq 12$

So who will be our pre-test sample?

•Students in classes (Cacao farmers)
•Students in African Students Union (Land stewards)
•Museum park visitors (Richmond recreators & NCMAT visitors)
•NCSU staff and Duke students (Raccoon relations)
•Zoo visitors (Zoo visitors)
•Land owners at meeting (Forest cooperators)
•Non-students including WRC staff, neighbors (NC non-game)

Next week

- Due on Friday 20 February
  - Revised questionnaire (used for pre-test)
  - Data entry template
  - At least two completed interviews
    - Questionnaires
    - Data
- Readings
  - None required
  - Supplemental will be posted on website