Statistical Methods for Sample Surveys (140.640)

Lecture 1

Introduction to Sampling Method

Saifuddin Ahmed

Session	Date	Description
1	01/21/09	Introduction to survey sampling methods
2	01/26/09 01/28/09	Simple random sampling, systematic sampling Lab
3	02/02/09 02/04/09	Sample size estimation Lab
4	02/09/09 02/11/09	Stratified sampling, sampling with varying probabilities (e.g., PPS) Lab
5	02/16/09 02/18/09	Cluster sampling, multistage sampling Lab
6	02/23/09 02/25/09	Weighting and imputation Lab
7	03/02/09 03/04/09	Special topics (design issues for pre-post survey sampling for program evaluation and longitudinal study design, WHO/EPI cluster sampling, Lot Quality Assurance Sampling[LQAS], sample size estimation for program evaluation) Project/Lab
8	03/09/09 03/11/09	Case Studies and Review Project review

Grading:

Lab	60%
Project	40%

Text: UN Handbook.pdf

(available at the CoursePlus website)

Text (Optional):

Sampling of Populations: Methods and Applications, 3rd ed
Paul S. Levy and Stanley Lemeshow
Wiley Interscience

Sampling: Design and Analysis
Sharon L. Lohr
Duxbury Press

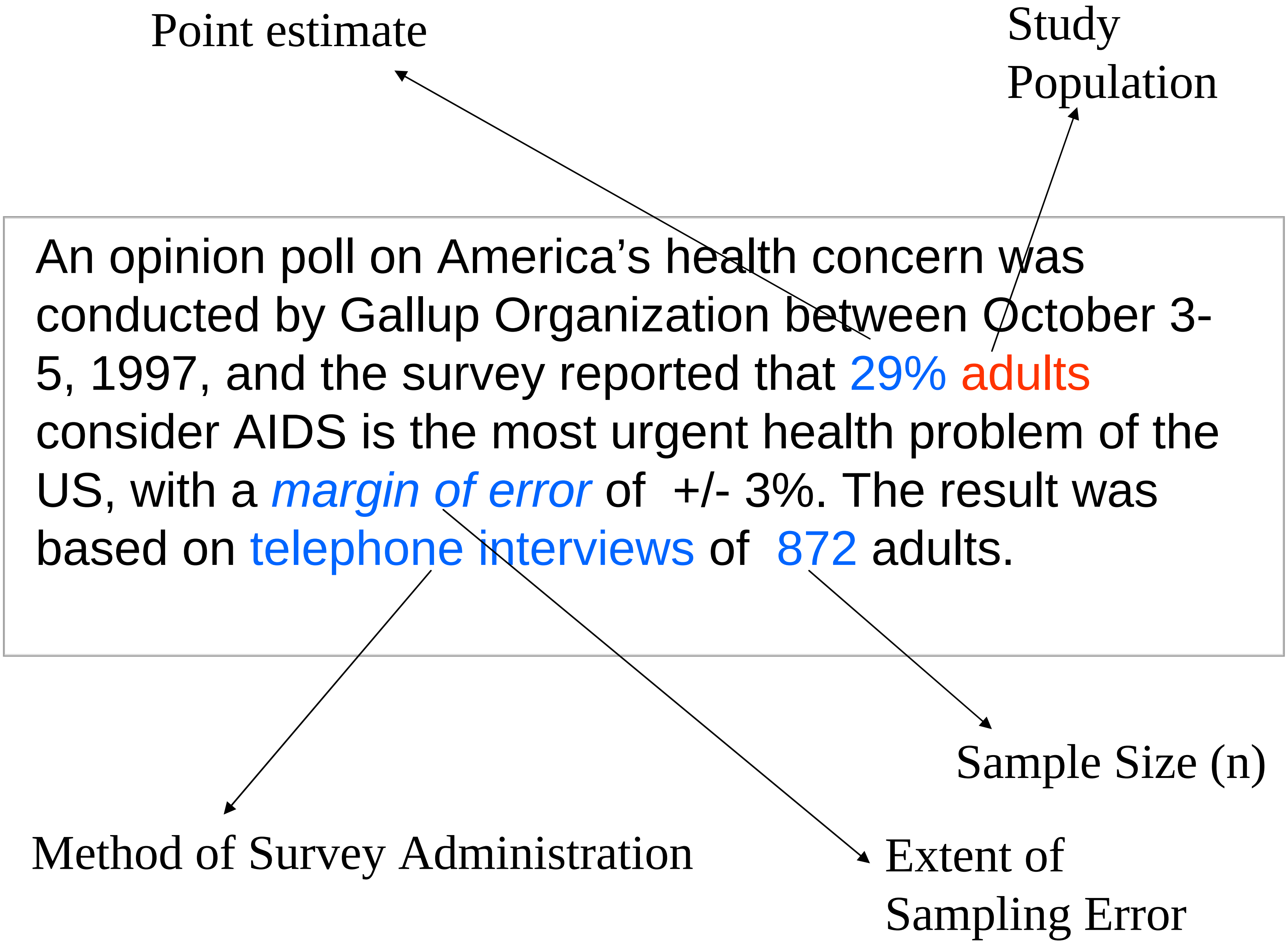
“There is hardly any part of statistics that does not interact in someway with the theory or the practice of sample surveys.

The difference between the study of sample surveys and the study of other statistical topics are primarily matters of emphasis”

W. Edwards Deming

An example for the scope of surveys:

An opinion poll on America’s health concern was conducted by Gallup Organization between October 3-5, 1997, and the survey reported that 29% adults consider AIDS is the most urgent health problem of the US, with a *margin of error* of +/- 3%. The result was based on telephone interviews of 872 adults.



Point estimate

Study Population

An opinion poll on America’s health concern was conducted by Gallup Organization between October 3-5, 1997, and the survey reported that 29% adults consider AIDS is the most urgent health problem of the US, with a *margin of error* of +/- 3%. The result was based on telephone interviews of 872 adults.

Sample Size (n)

Method of Survey Administration

Extent of Sampling Error

The outcome variable is **Bernoulli random variable.**

A binary variable have only yes/no category of responses.

“Do you consider AIDS is the most urgent health problem of the US?”

29% responded “yes”, and 71% responded “no”.

Let p = 0.29, and q = 1 – p = 0.71.

So, 872 X 0.29 = 253 responded “Yes”, and
872 X 0.71 = 619 responded “No”.

In summary, from the raw data, Gallup’s statistician *estimated* that

$$p = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum (1+0+1+1+.........+0)}{872} = \frac{253}{872} = 0.29$$

Here, 1 = “yes”, and 0 = “no” responses.

How much confidence do we have on this “point estimate” (29%) ?

From our knowledge of basic statistics, we can construct a 95% confidence interval around p as:

$$\hat{p} \pm Z_{.05} * se(\hat{p})$$

That is,

$$\hat{p} \pm 1.96 \sqrt{\frac{pq}{n}}$$

$$.29 \pm 1.96 \sqrt{\frac{(.29)(.71)}{872}}$$

$= .29 \pm 0.03$

So, 95% CI of *p* ranges between (.26 to .32).

$$\hat{p} \pm Z_{.05} * se(\hat{p})$$

The above mathematical expression could be rephrased as:

$$estimate \pm margin\_of\_error$$

(29%) (3%)

$$margin\_of\_error = 1.96\sqrt{\frac{(.29)(.71)}{872}} = 0.03$$

$$margin\_of\_error^{2} = 1.96^{2}\frac{(.29)(.71)}{872}$$

$$872 = 1.96^{2}\frac{(.29)(.71)}{margin\_of\_error^{2}}$$

$$n = 1.96^{2}\frac{pq}{d^{2}}$$

This example also shows that:

1. if we know/guess an estimated value (p), we can estimate the required sample size with a specified “margin of error”.

2. If sample size even increased significantly, the gain in efficiency would be small.

Stata output:

With n=872, margin of error=

```
di 1.96*sqrt(.29*.71/872)

.03011799
```

With n=1744 (2 times of the original sample size),

margin of error = 0.02

```
di 1.96*sqrt(.29*.71/1744)

.02129664
```

.

. **Improvement in “margin of error” = .03011799/.02129664 = 1.4142136**

With n=8720 (10 times of the original sample size),

margin of error = 0.01

```
di 1.96*sqrt(.29*.71/8720)

.00952415
```

Improvement in “margin of error” = .03011799/.00952415= 3.1622777

If sample size even increased significantly, the gain in efficiency would be small.

Check: sqrt(2) = 1.4142136 and sqrt(10) = 3.1622777

Explore the questions

- What was the target population? (US population)
- What was the sample population? (Adults whom could be reached by telephone)
- How the survey was conducted? (By telephone interview)
- How was the sample selected? (Randomly selected from telephone list)
- How reliable are the estimates? Precision of the estimates?
- How much confidence do we have on the estimates?
- Was any bias introduced in the sample selection (process)?
- Why 872 adults were selected for the survey?
- Can an estimate based on about 1000 respondents represents 187 million adults of the US?

All these issues are part of the sample survey methods.

Survey is conducted to measure the characteristics of a population.

Sampling is the process of selecting a subset of observations from an entire population of interest so that characteristics from the subset (sample) can be used to draw conclusion or making inference about the entire population.

Survey Design + Sampling Strategy = Survey sampling

Survey design:

What survey design is appropriate for my study?
How survey will be conducted/implemented?

Sampling procedure:
What sample size is needed for my study?
How the design will affect the sample size?

Appropriate survey design provides the best estimation with high reliability at the lowest cost with the available resources.

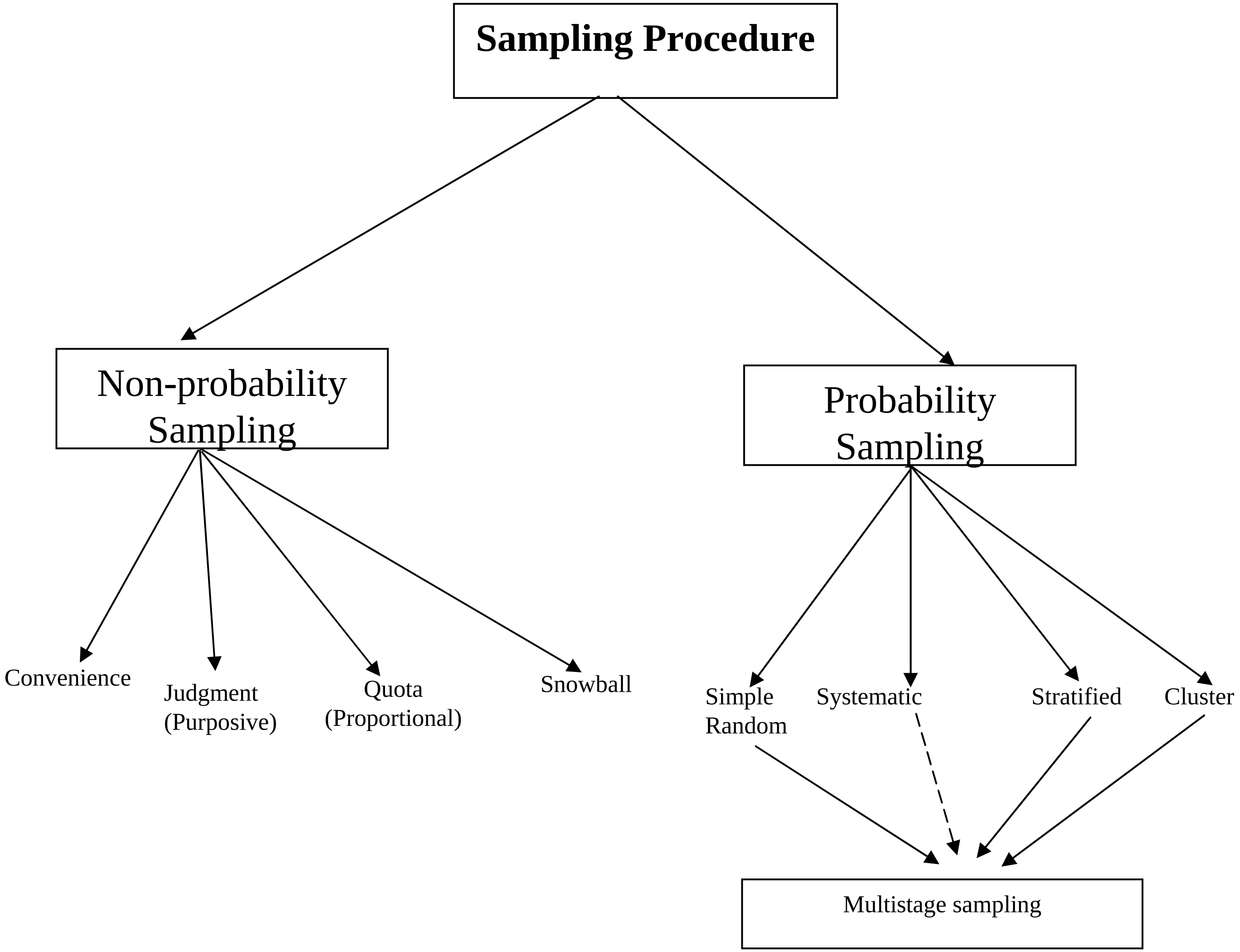
Why do we conduct surveys?

- *Uniqueness:* Data not available from other source
- *Standardized measurement:* Systematic collection of data in a structured format
- *Unbiased representativeness:* Selection of sample based on probability distribution

- contrast with census and experiment

Type of survey design

- Cross-sectional surveys: Data are collected at a point of time.
- Longitudinal surveys:
  - Trends: surveys of sample population at different points in time
  - Cohort: study of same population each time over a period
  - Panel: Study of same sample of respondents at various time points.


Sampling Procedure
Non-probability Sampling
Probability Sampling
Convenience
Judgment (Purposive)
Quota (Proportional)
Snowball
Simple Random
Systematic
Stratified
Cluster
Multistage sampling

Types of Samples

Probability Samples

- Simple random sampling
- Sampling with varying probabilities
- Systematic sampling
- Stratified sampling
- Cluster sampling

- Single-stage sampling vs. multi-stage sampling
- Simple vs. complex sampling

Non-probability samples

Judgement sample, purposive sample, convenience sample: **subjective**

Snow-ball sampling: **rare group/disease study**

Advantages of probability sample

- Provides a quantitative measure of the extent of variation due to random effects
- Provides data of known quality
- Provides data in timely fashion
- Provides acceptable data at minimum cost
- Better control over nonsampling sources of errors
- Mathematical statistics and probability can be applied to analyze and interpret the data

Disadvantages of Non-probability Sampling

- Purposively selected without any confidence
- Selection bias likely
- Bias unknown
- **No mathematical property**
- Non-probability sampling should not be undertaken with science in mind
- Provides false economy

Characteristics of Good sampling

- Meet the requirements of the study objectives
- Provides reliable results
- Clearly understandable
- Manageable/realistic: could be implemented
- Time consideration: reasonable and timely
- Cost consideration: economical
- Interpretation: accurate, representative
- Acceptability

Sampling Issues in Survey

- Produce best estimation
- Not too low, not too large
- Minimum error/unbiased estimation
- Economic consideration
- Design consideration: best collection strategy

Steps in Survey

1. Setting the study objectives

- What are the objectives of the study?
- Is survey the best procedure to collect data?
- Why other study design (experimental, quasi-experimental, community randomized trials, epidemiologic designs,,e.g., case-control study) is not appropriate for the study?
- What information/data need to be collected?

2. Defining the study *population*

Representativeness

Sampling frame

3. Decide sample design: alternative considerations

4. Questionnaire design

Appropriateness, acceptability, culturally appropriate, understandable

Pre-test: Appropriate, acceptable, culturally appropriate, will answer

5. Fieldwork

Training/Supervision

Quality monitoring

Timing: seasonality

6. Quality assurance

Every steps

Minimizing errors/bias/cheating

7. Data entry/compilation

Validation

Feedback

8. Analysis: Design consideration

9. Dissemination

10. Plans for next survey: what did you learn, what did you miss?

Advantages of Sampling

- Greater economy
- Shorter time-lag
- Greater scope
- Higher quality of work
- Evaluation of reliability
- Helps drawing statistical inferences from analytical surveys

Limitations of Sampling

- Sampling frame: may need complete enumeration
- Errors of sampling may be high in small areas
- May not be appropriate for the study objectives/questions
- Representativeness may be vague, controversial

Modes of survey administration

- Personal interview
- Telephone
- Mail
- Computer assisted self-interviewing(CASI)
  - Variants: CAPI (personal interview); CATI (telephone interview) – Replaces the papers
- Combination of methods

Interview Surveys

Interviewer selection:

- background characteristics (race, sex, education, culture)
- listening skill
- recording skill
- experience
- unbiased observation/recording

Interviewer training:

- be familiar with the study objectives and significance
- thorough familiarity with the questionnaire
- contextual and cultural issues
- privacy and confidentiality
- informed consent and ethical issues
- unbiased view
- mock interview session

Supervision of the interviewer:

- Spot check
- Questionnaire check
- Reinterview (reliability check)

Advantages

- Higher response rate, lowest refusal rates
- May record non-verbal behavior, activities, facilities, contexts
- Ensure privacy
- May record spontaneity of the response
- May provide probing to some questions
- Respondents only answer/participate
- Completeness
- Complex questionnaire may be used
- Illiterate respondents may participate

Disadvantages

- Cost (expensive)
- Time
- Less anonymity
- Less accessibility
- Inconvenience
- Interview bias
- Often no opportunity to consult records, families, relatives

Self-Administered Surveys

Advantages:

- Cost (less expensive)
- Time
- Accessibility (greater coverage, even in the remote areas)
- Convenience to the respondents (may complete any time at his/her own convenient time)
- No interviewer bias
- May provide more reliable information (e.g., may consult with others or check records to avoid recall bias)

Self-Administered Surveys

Disadvantages:

- Low response rate
- May not return questionnaire
- May not respond to all questions
- Lack of probing
- Must be literate so that the questionnaire could be read and understood
- May not return within the specified time
- Introduce self selection bias
- Not suitable for complex questionnaire

Telephone Interview

Advantages:

- **Less expensive**
- **Shorter data collection period than personal interviews**
- **Better response than mail surveys**

Disadvantages:

- **Biased against households without telephone, unlisted number**
- **Nonresponse**
- **Difficult for sensitive issues or complex topics**
- **Limited to verbal responses**

New Terminology in Computer Age

- PPI: Paper and Pencil Interview
- CAI: Computer-Assisted Interview
- CATI: Computer-Assisted Telephone Interview
- CAPI: Computer-Assisted Personal Interview
- CASI: Computer-Assisted Self-Interview
- Internet Surveys: Surveys over the WWW

Survey concerns

Selection bias:

- Selectivity
- misspecification
- undercoverage
- non-response

Survey concerns

Measurement bias:

- recall bias (just forgot)
- may not tell the truth
- may not understand the question
- Interviewer bias
- may try to satisfy the interviewer by stating what they want to hear
- misinterpret the questionnaire
- may be interpreted differently
- certain word may mean different thing
- question wording and order may have different impact
- longer questionnaire may deter appropriate response

Population and Sample

- *Population:* The entire set of individuals about which findings of a survey refer to.
- *Sample:* A subset of population selected for a study.
- *Sample Design:* The scheme by which items are chosen for the sample.
- *Sample unit:* The element of the sample selected from the population.
- *Unit of analysis:* Unit at which analysis will be done for inferring about the population. Consider that you want to examine the effect of health care facilities in a community on prenatal care. What is the unit of analysis: health facility or the individual woman?.

Finite Population:

A finite population is a population containing a finite number of items.

Compare to:

- Infinite population
- Natural population
- Homogenous population vs. heterogeneous population

Sampling Frames

For probability sampling, we should know the selection probability of an individual to be included in the sample.

Moreover, for obtaining unbiased estimators of parameters (e.g., % of mothers providing exclusive breastfeeding) we also make sure that all possible individuals are included in the selection process.

To comply with these requirements, we must sample the units from the population in such a way that we can calculate the selection probability, as well as, make sure that every one in the population is provided a chance for selection in the sample.

we must have a list of all the individuals (units) in the population. This list or sampling frame is the basis for the selection process of the sample.

“A [sampling] frame is a clear and concise description of the population under study, by virtue of which the population units can be identified unambiguously and contacted, if desired, for the purpose of the survey”

- Hedayet and Sinha, 1991

Based on the sampling frame, the sampling design could also be classified as:

Individual Surveys

- List of individuals available
- When size of population is small
- Special population

Household Surveys

- Based on the census of the households
- Convenient
- Individual level information is unlikely to be available
- In practice, limited to small geographical areas and know as “area sampling frame”
- Example: Demographic and Health Surveys (DHS)

Institutional Surveys

- Hospital/clinic lists
- 1990 National Hospital Discharge Survey
- National Ambulatory Medical Care Survey
- May be performed at two or multiple levels: Hospital list-> patient list
- Problem: different size facilities

Problems of Sampling Frame

- Missing elements
- Noncoverage
- Incomplete frame
- Old list
- Undercoverage
- May not be readily available
- Expensive to gather

Sampling frame is a more general concept. It includes physical lists and also the procedures that can account for all the sampling units without the physical effort of actually listing them.

What will you do when no list is available?

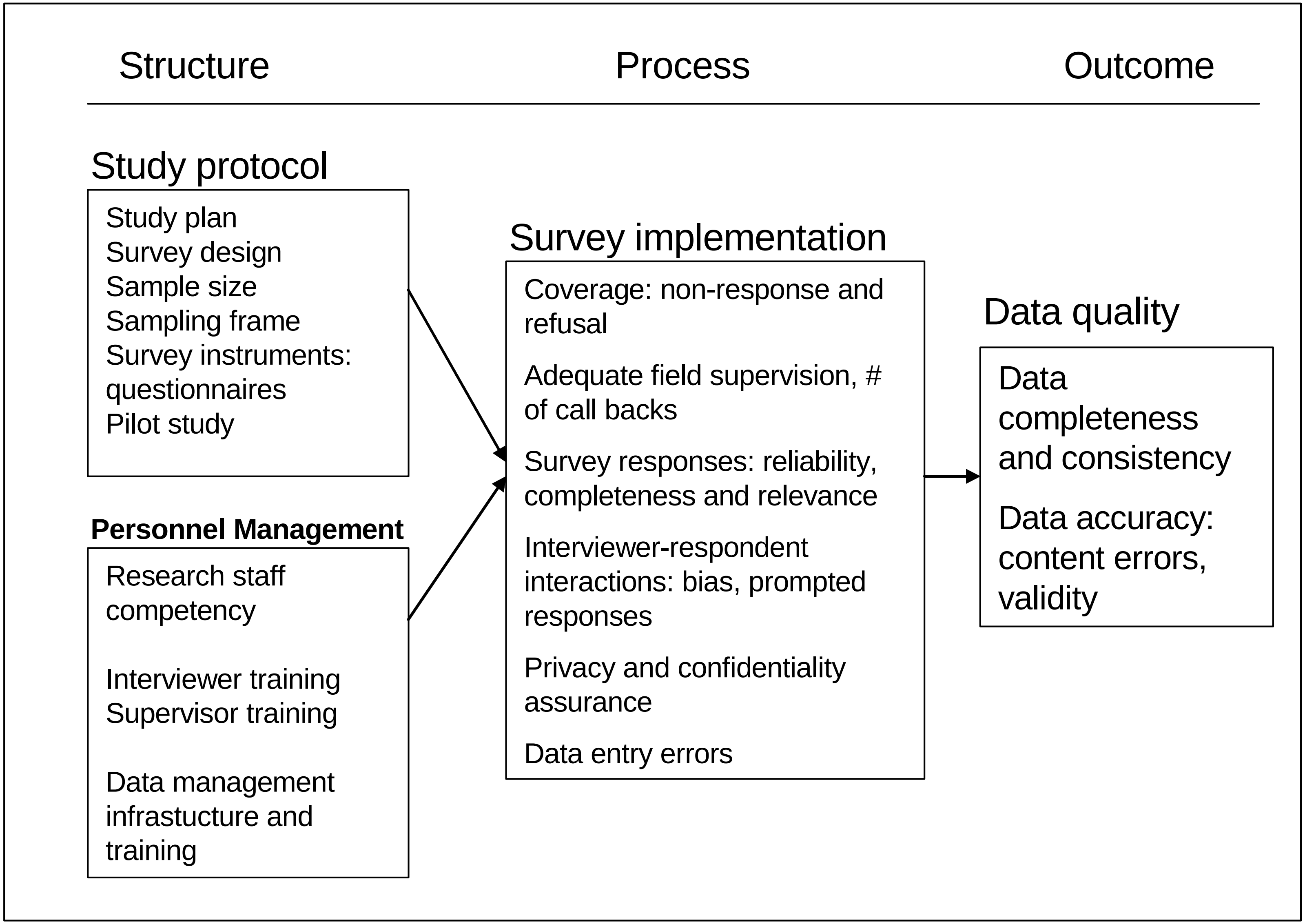
Properties of Probability Sampling

We take a small sample to infer about a larger population. We want to make sure that the representativeness of the sample and the accuracy or precision of the estimators.

What is an “estimator”?

Suppose that θ is the population characteristics (e.g., % children immunized). We want to measure it by some function, e.g., Y_i/N. Let us call it $f(\theta)$ when we consider the estimation is based on sample. θ ***is called a statistics or estimator***.

Data quality assessment framework



Structure

Process

Outcome

Study protocol

Study plan
Survey design
Sample size
Sampling frame
Survey instruments: questionnaires
Pilot study

Personnel Management

Research staff competency

Interviewer training
Supervisor training

Data management infrastucture and training

Survey implementation

Coverage: non-response and refusal

Adequate field supervision, # of call backs

Survey responses: reliability, completeness and relevance

Interviewer-respondent interactions: bias, prompted responses

Privacy and confidentiality assurance

Data entry errors

Data quality

Data completeness and consistency

Data accuracy: content errors, validity

Two important characteristics:

Unbiasedness:

One important criterion for the sample is judged by “representativeness”. In statistical term we refer this to “unbiasedness”, i.e., θ (estimated θ) should be unbiased.

Precision:

Once we have estimated θ, we want to assess the “precision” or “accuracy” of an unbiased estimator.

Smaller the variance, the more precise the estimator

The objective of sampling theory is to devise sampling scheme which is:

- Economical
- Easy to implement
- Yield unbiased estimators
- Minimize sample variations (produce less variance)

Survey Designs

Definitions of Population and Sample:

Population ***P*** consists of N elements, U_1, U_2, …., U_N.

The Size of population is N.

U's are the elements (units of universe).

Finite population parameter: a property of population ***P*** (e.g, mean). This value is independent of how P is sampled.

With each unit of U_i there is realization of some characteristics, Y_i . As an example, Y_i may be age of the i^{th} individual.

Examples of population parameters:

Population total is $$T_Y = \sum_{i=1}^{N} Y_i$$

Example: total income.

Population mean is $$\overline{Y} = \frac{T}{N} = \frac{1}{N}\sum_{i=1}^{N} Y_i$$

Example, mean age.

Population variance is $S_Y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \overline{Y})^2$

Population variance is also expressed as $\sigma_Y^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \overline{Y})^2$

σ_Y^2 is related to S_Y^2 by, $\sigma_Y^2 = \frac{N-1}{N}S^2$.

To distinguish S_Y^2 from σ_Y^2, S_Y^2 is often called “population mean square”.

Standard deviation of the Y_i values is $\sigma_Y = \sqrt{\sigma_Y^2}$

Sample statistics

Sample ***S*** consists of *n* elements selected with the associated characteristics values of y_1, y_2, …., y_n.

Note that, sample y_1 may not be the same Y_1.

Sample mean is $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

Sample variance is $s_y^2 = \frac{1}{n-1}\sum_{i=1}^{N}(y_i - \bar{y})^2$

Population parameters are usually expressed in Greek letters. Population total may be expressed as τ (tau), and mean as μ (mu).

Relationship of sample to population

	Population *Parameters* N		*Sample* *statistics* n
Variable (Y)	Y_i		y_i
Mean	$\bar{Y} = \frac{\sum Y_i}{N}$		$\bar{y} = \frac{\sum y_i}{n}$
Total	$Y = \sum Y_i = N\bar{Y}$	$\longleftarrow \boxed{N\bar{y} = \frac{N}{n} n\bar{y}} \longleftarrow$	$y = \sum y_i = n\bar{y}$

In summary, $\bar{y}$ is an unbiased estimator of μ, the population mean, and $N\bar{y}$ is an unbiased estimator of population total.

$N\bar{y}$ **is called expansion estimator**