

Geneva Workshop 2015

# SAMPLING

An Introduction

Theory vs. Practice

# **A PERSONAL STORY OF A STATISTICIAN WITH A MATHEMATICS BACKGROUND**

# A national sample survey to know opinions & attitudes of students:

A stratified two-stage (cluster) sampling design, PSU selected with pps (probability proportional to size,  $Z_{hi}$ ); strata="provinces", PSU="schools", SSU="students":

- (Textbook) estimated population total:

$$\widehat{Y}_{st} = \sum_h^L \frac{1}{n_h} \sum_i^{n_h} \frac{M_{hi} y_{hi}}{m_{hi} z_{hi}}$$

- Self-weighted estimation:

$$\widehat{Y}_{st} = \sum_h^L \frac{1}{f_{0h}} \sum_i^{n_h} y_{hi}$$



A survey and a trial

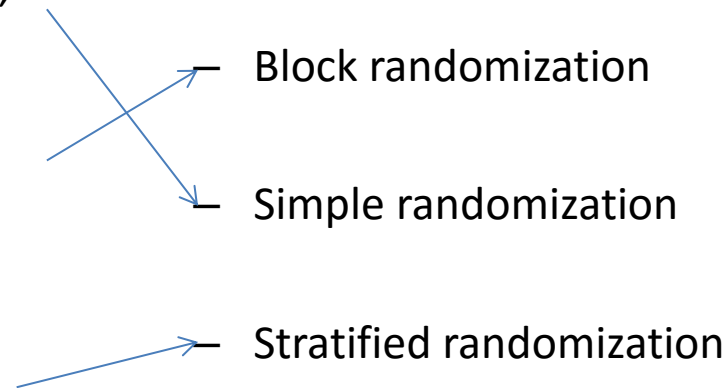
# **TWO EXAMPLES, TWO DIFFERENT WORLDS?**

**A survey**: Country “X” had 160000 deliveries last year, 100000 women  $\geq 20$  years old and 60000 women  $< 20$  years old. The MoH wants to know the incidence of “mistreatment during delivery” and thinks of taking a sample of 1000 deliveries/women; different methods/options are considered:

- List the population of mothers alphabetically and choose every 160<sup>th</sup> person
  - Separate  $\geq 20$  and  $< 20$  years old mothers and randomly select 500 from each group
  - Assign each mother a unique number from 1 to 160000, and have a computer program to randomly select 1000 numbers within the range
  - If clinics’ sizes in the country were about 500 deliveries each last year, select randomly 2 clinics
- 
- Simple random sampling
  - Cluster sampling
  - Systematic sampling
  - Stratified sampling

**A trial**: we want to assess the efficacy of pre-pregnancy Calcium on the incidence rate of pre-eclampsia. Non-pregnant women who wants to become pregnant soon (eligibility criterion), are allocated to “control” (Placebo) or “experimental” (Calcium) groups. Different allocation methods are considered:

- Numbers 0 and 1 are randomly generated for each subject recruited, 0 goes “Placebo” and 1 “Calcium”
- Every 6 women, 3 are randomly allocated to “Placebo” and the rest goes to “Calcium”
- Women are recruited at different Family Planning clinics, and within each clinic allocation of women to the treatment arms is conducted



Two different worlds?

# **SAMPLING VS. RANDOMIZATION**

# Sampling vs. Randomization?!

## Sampling:

- Is concerned with the selection of a subset of individuals from within a population to estimate characteristics of the whole population (Wikipedia)
- The process of selecting a portion, piece, or segment that is representative of a whole (The American Heritage College Dictionary, 1993)
- Random sampling: ...

## Randomization:

- Is the process of making something random (Wikipedia):
  - Generating a random permutation
  - Selecting a random sample of a population
  - Allocating experimental units to a treatment or control via random assignment
- Practice of using chance methods to assign subjects to treatments (Stat Trek)



# Sampling vs. Randomization?

- Study designs:
  - Observational (epidemiological) studies:
    - Cross-sectional (Surveys): sampling
    - Case-control: sampling \* (and matching)
    - Cohort: sampling \* (and matching)
  - Experimental studies:
    - Clinical trials: randomization

\*: restricted sampling

# Sampling and Randomization

- Randomization: the objective is to make things comparable (usually treatment groups in controlled clinical trials)
- Sampling:
  - Is a particular case of randomization
  - Sampling (in sample surveys) randomize units of the population into one of two “treatments”: “being in the sample” and “not being in the sample”
  - Through sampling we make these two groups comparable, i.e. “the sample” is similar to “the population (without the sample)”; usually said “the sample is representative of the population”

# **SAMPLING IN SURVEYS**

# Some Key Concepts:

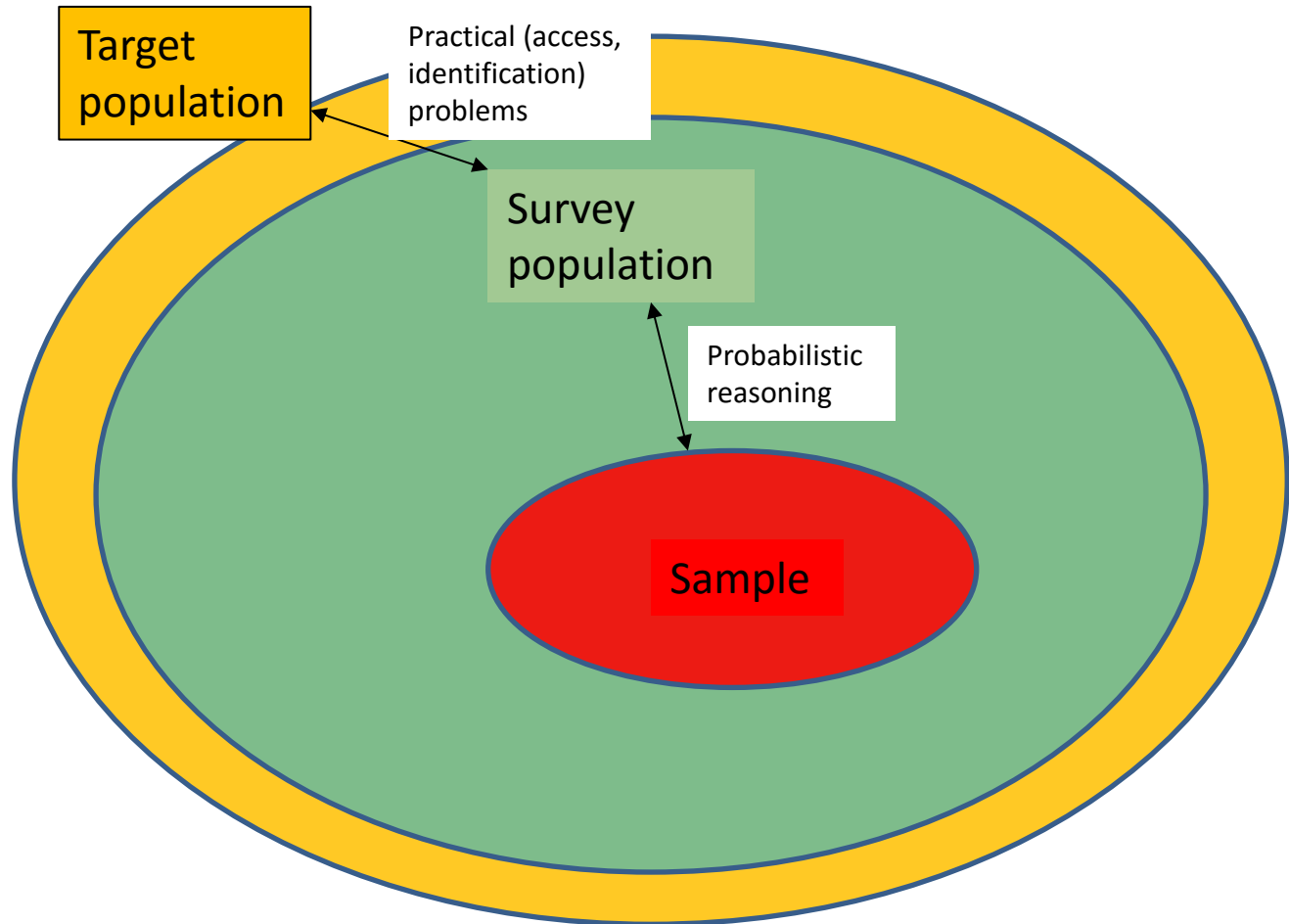
- Sampling is a fundamental tool for making inferences in observational (cross-sectional) research:
  - We use samples because usually we cannot afford censuses
  - The sample is a mean to study the population
  - Sampling makes sense for (very) large populations
  - Significance tests are for «sampling-based» studies, not for censuses
  - First think about your «population»
    - «target» population, «survey/sampled» population, «sample»
  - Larger sample sizes reduce the variability of estimates, not the variability of the corresponding variables

# «Target» and «survey» populations:

Criteria for defining the target population:

- Content
- Units
- Extent
- Time

Example: Objective of the survey is to estimate the prevalence of Hepatitis B carriers among children <18 years. Target population is all children <18 years (content), from all types of families (units) living in Thailand (extent) in 2004 (time). The specificities of the criteria for the survey population are generally different.



# Sampling frames

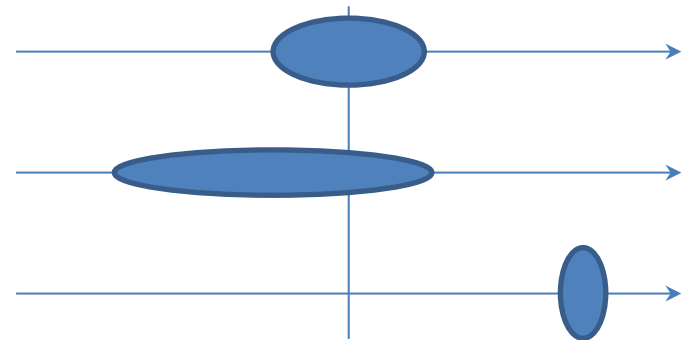
A sampling frame is a list that allows identifying every individual of the target population

- Highest possible coverage of target population
- Up to date
- Examples:
  - Households
  - Companies/enterprises
  - Schools

# The Estimation Process:

- Parameters and Estimators:
  - Parameters: characteristics of the population (e.g. «diabetes prevalence»)
  - Estimators: statistics computed on samples
  - There are «estimators» for each «parameter»

- Estimation Error components:
  - Precision: Random component
  - Bias: Systematic component (e.g. measurement instrument not calibrated)



# Sampling: the SRS paradigm

- Size:
  - Absolute, not relative
  - «Sample is 5% of population» means nothing
  - Mainly improves «precision» of estimates
- Design:
  - At least as important: gives credibility to estimates
  - May make sampling more efficient («precision») and «practical»



# **SURVEY SAMPLING**

# What's a Survey?

- A Survey is essentially a cross-sectional study.
  - Cross-sectional: data refer to events/variables all measured at the same time
  - Longitudinal: data refer to events/variables measured at different times
    - Retrospective: All events occurred (and measured) before data were collected
    - Prospective: Events and their recording occur concurrently, after study starts
- It is good for estimating:
  - incidence and/or prevalence of relevant characteristics; e.g. proportion of stillbirths among deliveries, in delivery clinics in country X
  - Correlations (exploring potential associations)
- It is not good for estimating cause-effect relationships

# Some basics:

- Identify the «target» and «sampled» populations
  - Identify the main parameter(s)
  - Precision required
  - Sampling frame
- Sample size determination
  - SRS (simple random sampling): EPI-INFO
    - Self-weighted design (equi-probable design); makes your sample representative of the population
  - More complex designs (clustered and/or stratified)
- Designs:
  - Types of:
    - Stratified
    - Cluster / multi-stage
    - With-without replacement
    - Systematic
    - Combinations: e.g. «two-stage stratified» design
  - Make sampling feasible
  - Adapted to the context
  - Aim at a «self-probability» design (close to SRS) for «representativeness»
    - Otherwise weights should be used at the analysis

# More basics:

- Increasing sample size increases precision (reduces random error) but does not effect bias
  - In a census, precision is maximum (random error is zero) but bias might remain the same
- Improving the sampling design may reduce some bias sources
  - Other bias sources are non-sampling related
- Rationale behind:
  - Stratification
    - Better representativeness (reduces bias)
    - Reduces sample size is stratification is «efficient»
  - Cluster/multi-stage
    - Make sampling more feasible
    - Variables at different levels are relevant (multilevel models)
    - Sampling less efficient
  - (Again) larger sample sizes reduces only variability of estimates:  $sem = \sigma/\sqrt{n}$

An Example

# **THE WHO GLOBAL SURVEY**

# The WHO Global Survey

- **Caesarean delivery rates and pregnancy outcomes: the 2005 WHO global survey on maternal and perinatal health in Latin America. *Villar J et al. Lancet 2006; 367: 1819–29.***
  - ***Aim: to assess the association between caesarean delivery and pregnancy outcome at the institutional level, adjusting for the pregnant population and institutional characteristics.***

# Methods («Population»):

*We designed the 2005 WHO global survey on maternal and perinatal health to explore the relation between rates of caesarean delivery and perinatal outcomes in the medical institutions of eight randomly selected countries in the region of the Americas, using a multistage stratified sampling procedure. We obtained data between Sept 1, 2004, and March 30, 2005. After country selection, we identified a representative sample of geographic areas within each country and, within these geographic areas, a representative sample of care units. We selected countries with a probability proportional to the population of the country, provinces with a probability proportional to the population of the province, and health institutions with a probability proportional to the number of deliveries per year. Here, we present results from the eight countries in Latin America; we will report results of a similar survey done in Africa separately. In 2006, we will prepare the survey for Asia. We initially stratified each country by its capital city (always included) and two other randomly-selected administrative geographic areas (provinces or states). Within these three areas, we undertook a census of hospitals that reported more than 1000 deliveries in the previous year. We then stratified data by province or state, choosing a representative sample of up to seven institutions each. If there were seven or fewer eligible institutions, we included them all. We included all women admitted to the selected institutions for delivery during a fixed data collection period of either 2 or 3 months, depending on the total number of expected deliveries per institution for the complete year (3 months if  $\leq 6000$  per year; 2 months if  $> 6000$  per year).*

# Methods («Procedures»):

- *We collected data at two levels—institutional and individual.*
  - *At the institutional level, we gathered data on one occasion only, with the aim of obtaining a detailed description of the health facility and its resources for obstetric care.*
  - *At the individual level, we obtained from all women’s medical records information to complete a two-page pre-coded form, summarising obstetric and perinatal events.*

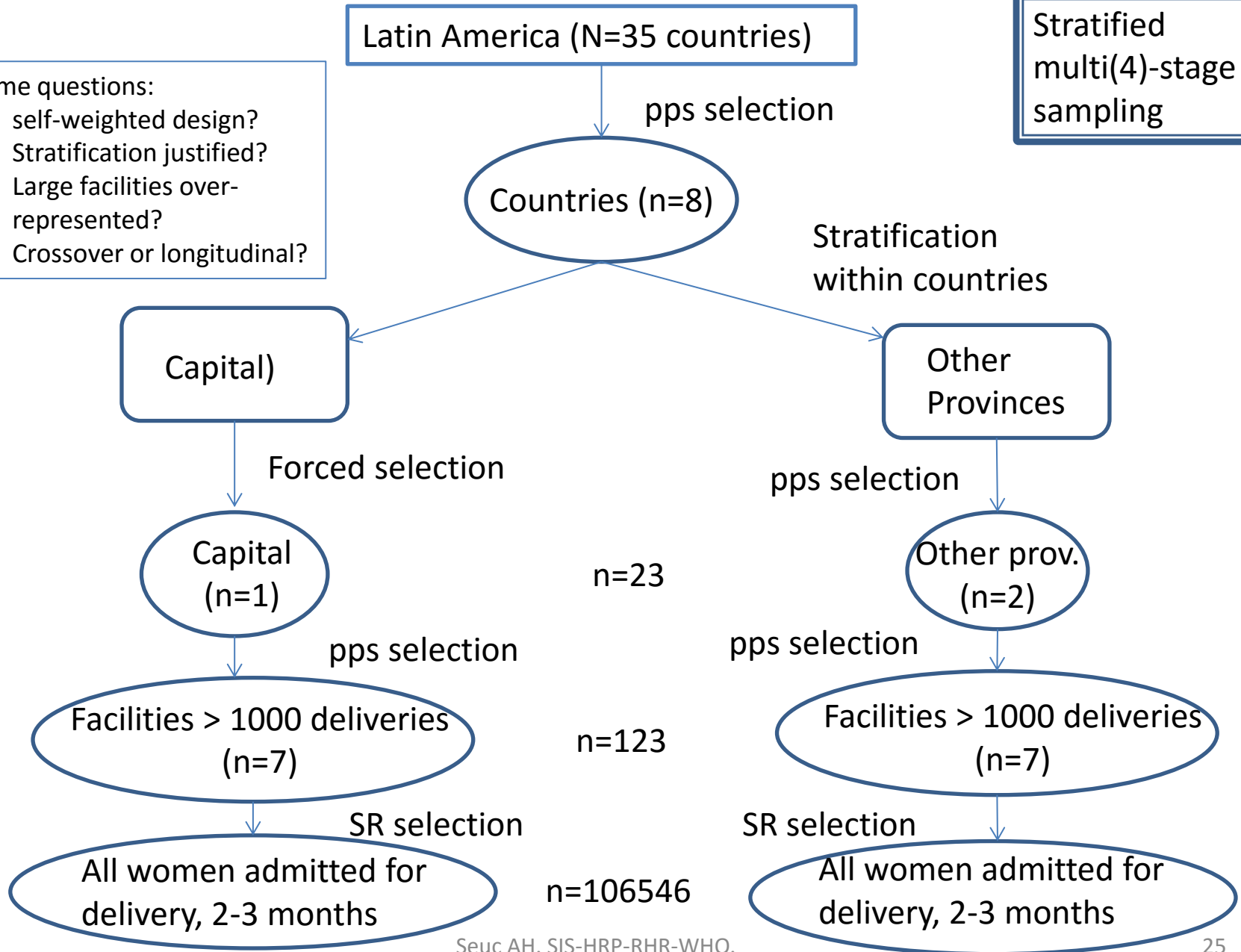
*Trained staff reviewed the medical records of all women within a day after delivery and abstracted data to their individual data collection forms, which were completed during the period that the woman and newborn baby remained in hospital.*

*We used the individual-level form to obtain information about demographic characteristics, maternal risk, pregnancy events, mode of delivery, and outcomes up to hospital discharge.*



- Some questions:
- self-weighted design?
  - Stratification justified?
  - Large facilities over-represented?
  - Crossover or longitudinal?

Stratified multi(4)-stage sampling



# Design effect (in comparison to srs)

## Definition (from Wikipedia):

$Deff = 1 + (m - 1)\rho$ , where  $m$  is the number of observations in each cluster, and  $\rho$  is the intraclass correlation coefficient (between 0 and 1).

More formally, the design effect is the ratio of two theoretical variances for an estimator: the actual variance for a given sampling design, and the variance assuming the same sample size, but using simple random sampling without replacement

Sample size calculations for surveys / cross-sectional studies

# **EPI-INFO**

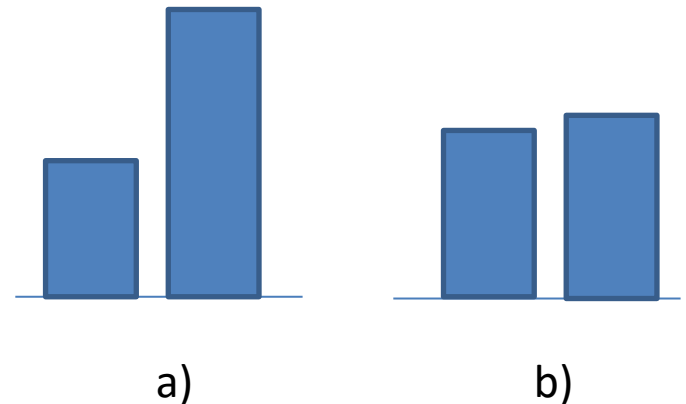
# Main Objectives in a Survey:

- Obtain precise estimations of population means (proportions or rates). Examples:
  - Prevalence of diabetes in country «X»
  - Incidence of Cesarean Section in country «Y»
  - Average BMI of teenagers in region «Z»

# **SAMPLING FOR RCT**

# Some basics:

- Goal: assess relationship between «exposure» and «outcome»
- Comparability between exposure groups is fundamental (randomization)
- Representativeness is desirable (associations tend to be «stable» across populations):
  - Inclusion/exclusion criteria
  - Stratification (multicountry trials)
- The smaller the difference we want to detect (we expect to find) between the exposure groups, the larger the sample sizes:



Sampling for Experimental Studies (RCT)

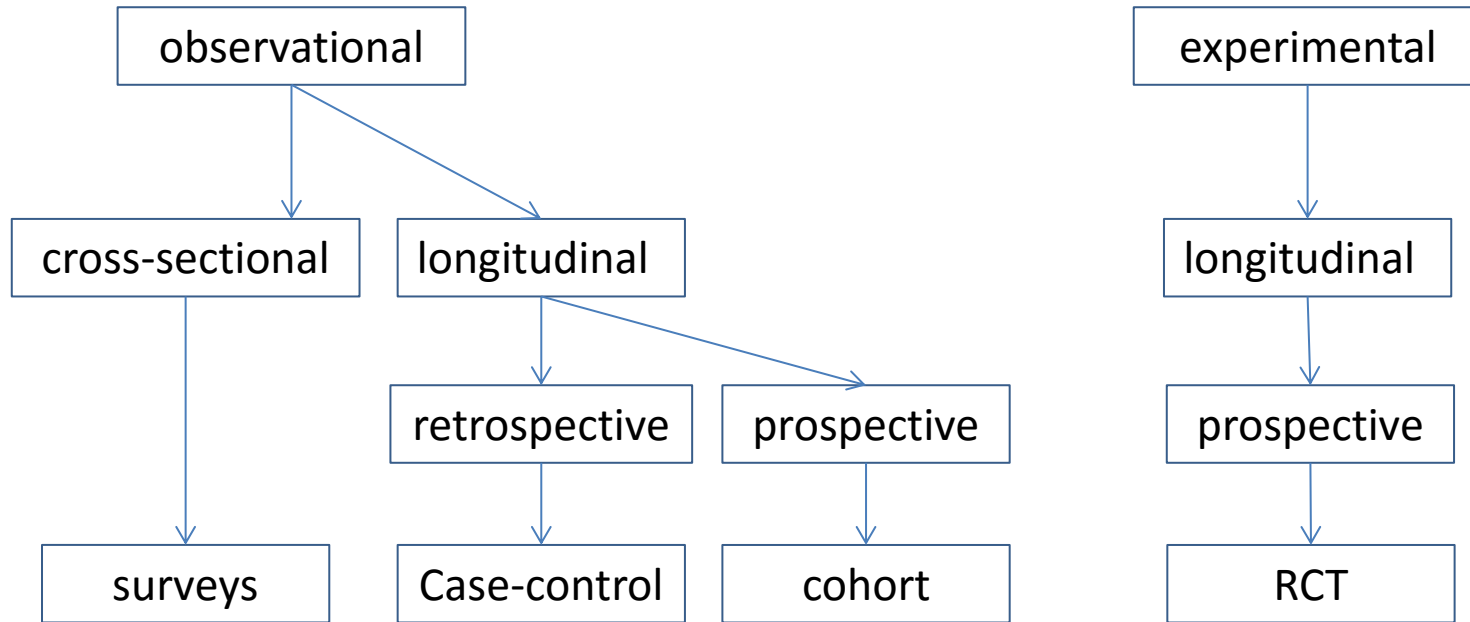
# **EPI-INFO**

# Main objectives in an Experimental Study (RCT):

- Experimental (e.g. RCT) or non-experimental studies want to assess:
  - if the outcome incidences (%) in the «exposed» and «control» groups are or not the same
  - If the outcome means in the «exposed» and «control» groups are or not the same
- Examples:
  - Calcium supplementation before pregnancy reduces the risk of pre-eclampsia
    - Calcium... pregnancy... preeclampsia
  - Baby HC intake first 4w after delivery through BF, affects development:
    - Height, weight, BMI at 2 and 4 years
    - Age at menarche in girls



# Research Designs:



# **OTHER SAMPLING SCENARIOS**

# Other sampling scenarios:

- Qualitative studies:
  - The goal is to obtain insights into a phenomenon, not to generalize to a population:
    - Maximum variation
    - Homogeneous
    - Critical cases
    - Theory-based
    - Opportunistic
    - Convenience
    - Quota
    - Snowball(Onwuegbuzie, 2007)
  - Delphi method:
    - Expert groups are selected (sampled), to discuss and reach consensus on complicated and/or sensitive issues using successive iterations (Skulmoski 2007)
- Adaptive designs in confirmatory clinical trials:
  - The study is planned involving design modifications based on the results of interim analyses, while keeping the pre-specified type I error:
    - Sample size
    - Randomization ratio
    - Number of treatment arms
    - Varying number and timing of interim analyses
    - Stopping rules(EMA 2007)

# Other sampling scenarios (cont.):

- LQAS (Lot Quality Assessment Sampling):
  - History: a small representative sample of a recently manufactured lot of goods from a production unit, such as an assembly line or machine, is inspected. If the number of defective goods in the sample exceeds a predetermined allowable number, then the lot is rejected; otherwise the lot is accepted
  - Current applications in health: units (municipalities, hospitals, etc.) are accepted or rejected if they do/don't satisfy a pre-determined coverage target for a particular intervention (e.g. immunization, contraceptive use, etc.).
    - Purpose: To identify priorities for improving health program's implementation and/or execution
- Sampling design: high probability of accepting lots with performance levels above the high level border; high probability of rejecting lots with performance levels below the low level border
- Assumptions:
  - The null hypothesis needs to be defined; main objectives of the study and ethical concerns accounted
  - Lots need to be defined
  - Low and high threshold levels of acceptance need to be defined for "lots"
  - Lots above the high threshold level are accepted; lots below the low threshold level are rejected
  - The larger the difference between the two threshold levels, the smaller the sample size required in each lot
- When it makes sense:
  - High homogeneity within lots
  - High heterogeneity between lots

(See Rhoda et al (2010))

# Other sampling scenarios (cont.):

- Sampling hard to reach populations
  - What is a hard-to-reach population?:
    - a) small
    - b) members hard to identify / no sampling frame
    - c) stigmatized
  - Objective: construct probability samples (extrapolate with calculable variances)
  - Approaches:
    - Time location sampling (TLS)
    - Respondent-driven sampling (RDS)
    - Capture-recapture

(See e.g. Marpsat 2010)
- Internet (Web/email) surveys:
  - Self selected samples make them usually not representative of the “population”(?)
  - But fast and not expensive way of getting preliminary information, useful if limitations are identified and addressed

(see e.g. Fleming 2008, Fricker 2008)
- Simulation/Monte Carlo methods
  - Estimating the  $\pi$  value, etc.
- Resampling
  - Bootstrap
  - Jackknife
  - Crossvalidation