Sampling

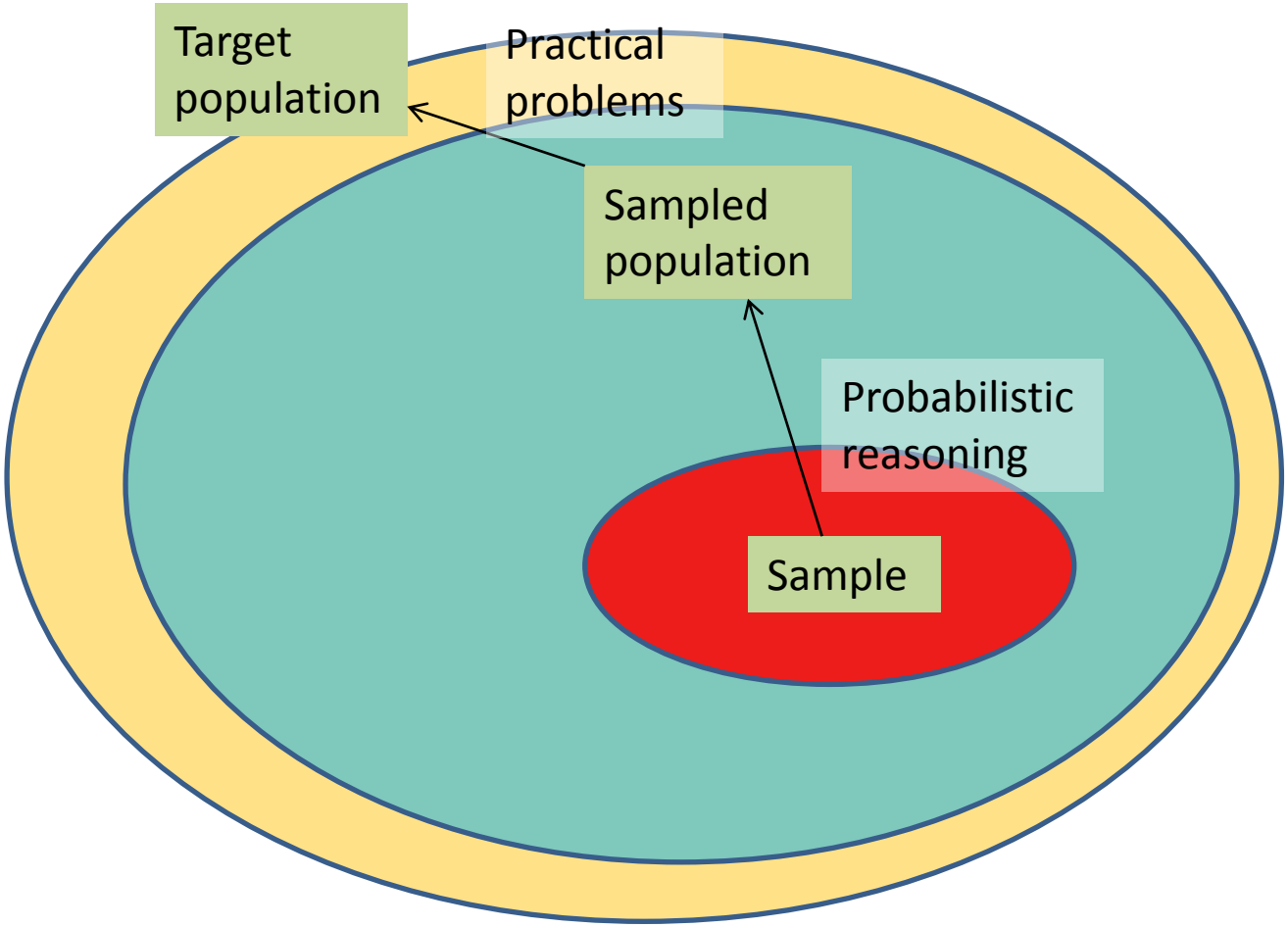
A basic introduction

Training Course in Sexual and Reproductive Health Research
Geneva Workshop 2012

Key Concepts:

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

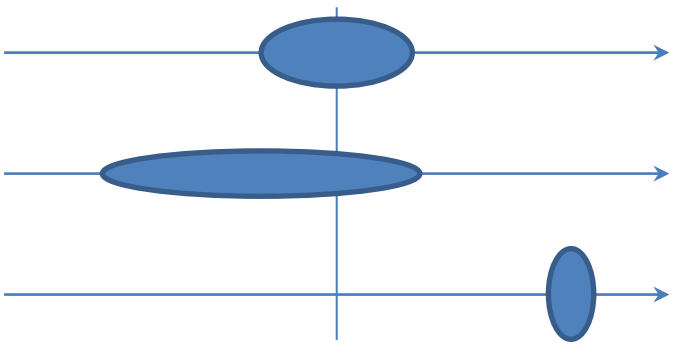
«target» and «sampled» populations:



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-weigthed 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:

- **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 ≤6000 per year; 2 months if >6000 per year).

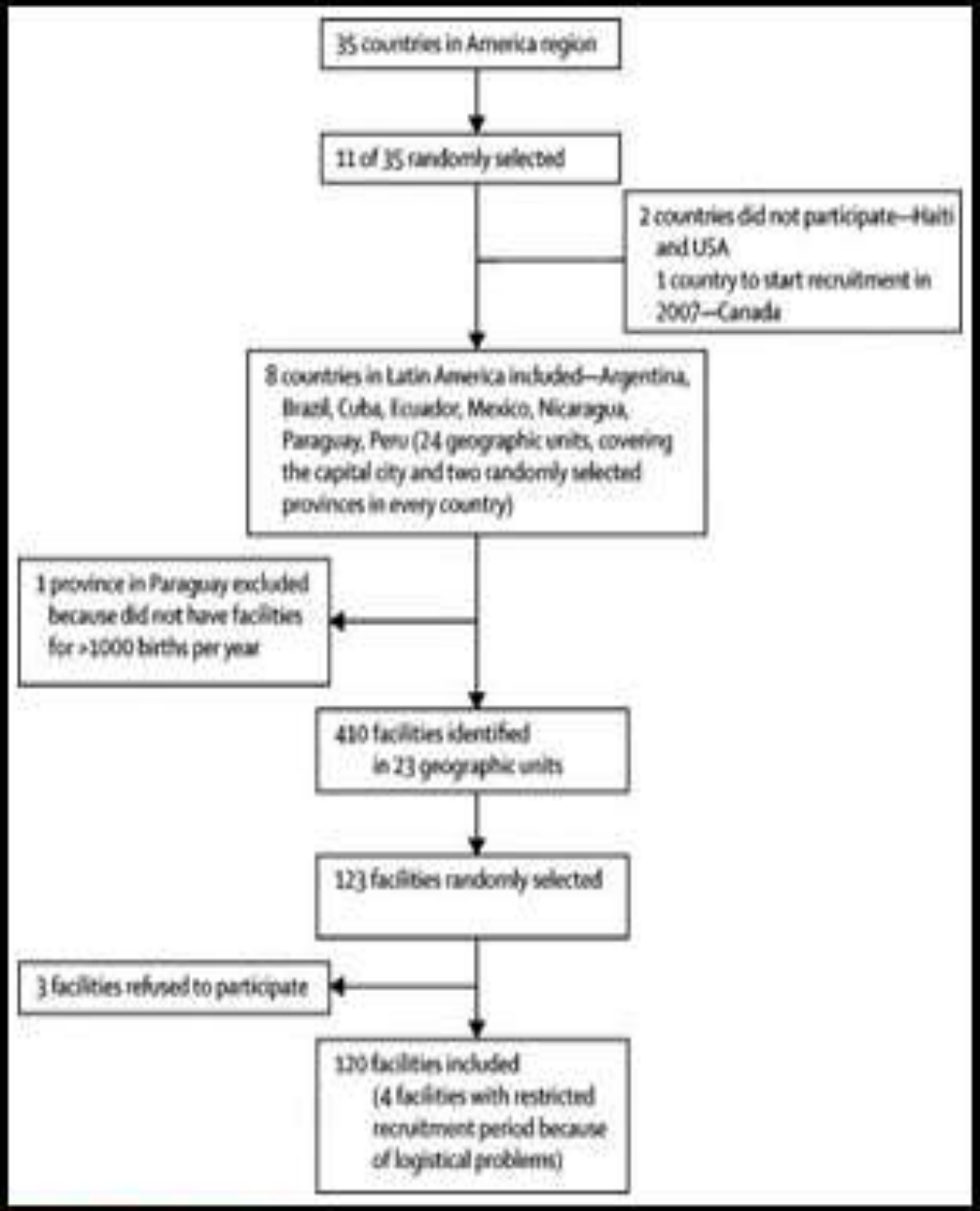
Methods («Procedures»):

- *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.*

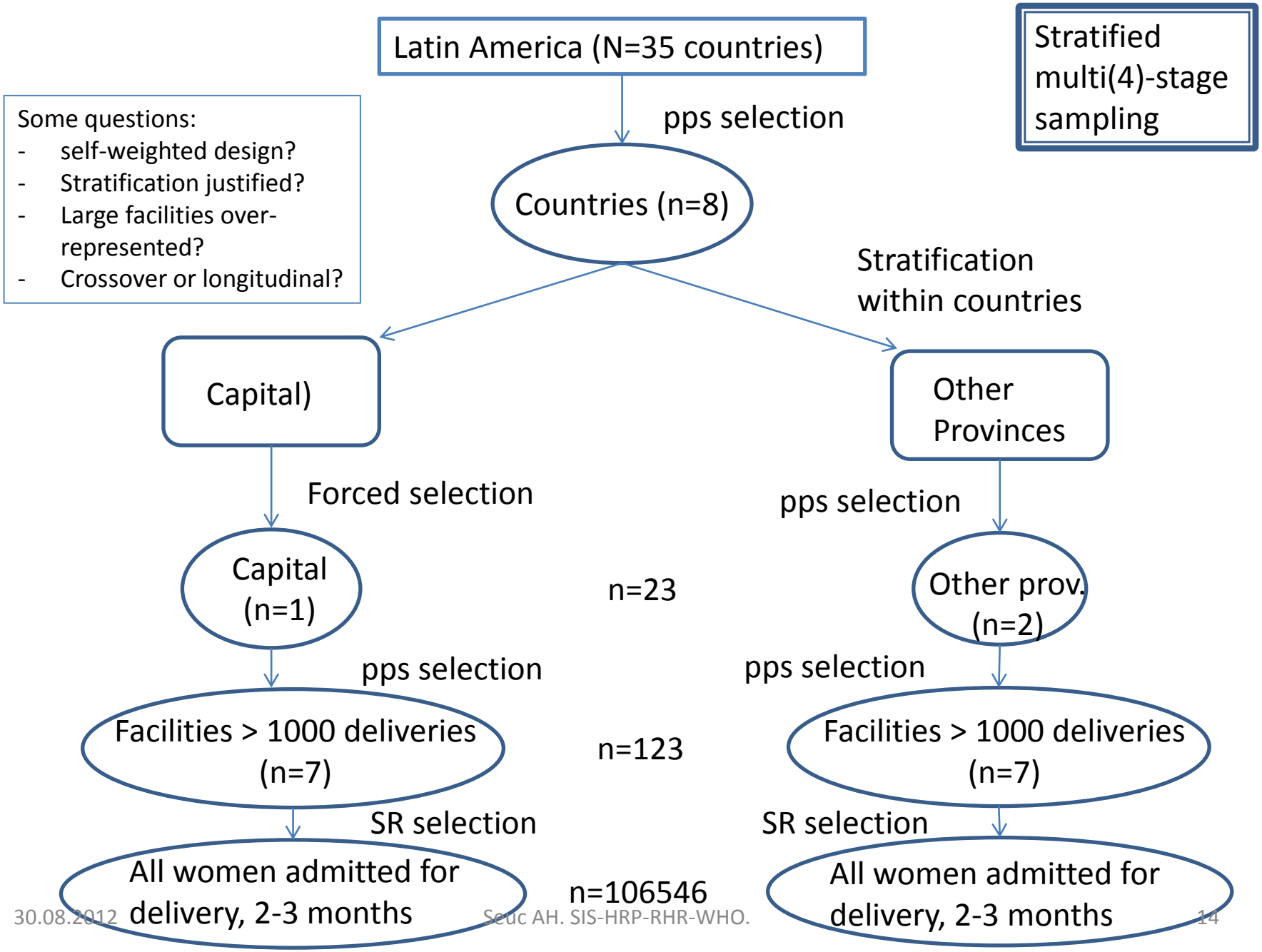
Trained staff reviewed the medical records of all women within a day after delivery and abstracted data to their individual data collection forms, which were completed during the period that the woman and newborn baby remained in hospital.

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.

Figure 1



Source


Latin America (N=35 countries)
pps selection
Countries (n=8)
Stratification within countries
Capital)
Other Provinces
Forced selection
pps selection
Capital (n=1)
n=23
Other prov. (n=2)
pps selection
pps selection
Facilities > 1000 deliveries (n=7)
n=123
Facilities > 1000 deliveries (n=7)
SR selection
SR selection
All women admitted for delivery, 2-3 months
n=106546
All women admitted for delivery, 2-3 months
Some questions:
- self-weighted design?
- Stratification justified?
- Large facilities over-represented?
- Crossover or longitudinal?
Stratified multi(4)-stage sampling

Sample size calculations for surveys / cross-sectional studies

Dean AG, Sullivan KM, Soe MM. OpenEpi: Open Source Epidemiologic Statistics for Public Health, Version 2.3.1.

www.OpenEpi.com

EPI-INFO

Main Objectives in a Survey:

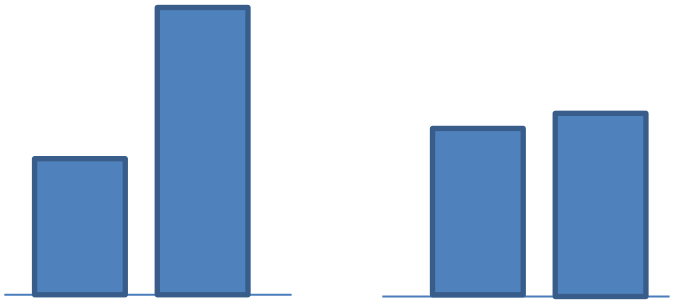
- Obtain precise estimations of means (proportions). Examples:
  - Prevalence of diabetes in country «X»
  - Incidence of Cesarean Section in country «Y»
  - Average BMI of teenagers in region «Z»

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:



a) b)

Sampling for Experimental Studies (RCT)

Dean AG, Sullivan KM, Soe MM. OpenEpi: Open Source Epidemiologic Statistics for Public Health, Version 2.3.1.

www.OpenEpi.com

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: