July 21 - 24, 2021 | Garki Hospital Abuja | Resource Person

Research Methodology Boot Camp

with Epi Info Training

Dr. Adamu Onu
MBBS, FWACP (FM)
MS Epidemiology & Biostatistics
PhD Public Health (Epidemiology)

Target Audience Clinical Researchers, Post-Part 1 Residents, and Others

Important Information

- Limited slots are available on a first come, first served basis
- Laptop running Windows 10 required
- Organized as morning lecture sessions and afternoon hands on coaching sessions

Highlights

- Research Methodology
- Research Design
- Data Management
- Sample Size Calculations
- Test Statistics
- Interpretation of Results
- Report Writing
- Hands-on training sessions
- Statistical consulting sessions

For further details contact
Email: epimetrix@gmail.com
Phone: +234 803 474 9930

epimetrix
health systems
data for health decisions

Cluster Sample Design and Analysis

Random sampling

- Each member of population has an equal chance of selection (unbiased)
- Represents target population (sampling frame)
- Requires a list of all population members (enumeration)

When random sampling is not feasible

- Population (sampling frame) is not enumerated
- Population covers a wide geographic area (e.g. country)
- Potential for contamination of interventions after randomization
- Studies of health care practices or educational or community interventions

Definition

- In the selection process, the sampling units consist of one or more mutually exclusive groups (clusters)
- Natural grouping within the population
- Each cluster has an equal chance of selection from a list of all clusters

Examples of clusters

- Spatial
  - Geographic Wards
  - Households
- Organizational
  - Schools
  - Physician practices
- Temporal
  - Day of the week

Disadvantages

- Compared with simple random sample, greater likelihood that sample is not representative of population.
- Requires larger sample size than simple random sample
- Adjustments in analysis are required

Advantages

- Only members of selected clusters need to be enumerated and studied
- Members of the sample are physically together in groups, rather than scattered all over study population
- More economical use of resources
- Easier to obtain large sample size

Multistage sampling

Stratification

- Stratification at first stage (e.g. Geopolitical zone)
- Each subgroup (stratum) of population is represented and can be analyzed separately
- Increases precision of estimates and power of study by reducing variance
- Distinguish from stratification in analysis (after data collection)

Multistage sampling

Primary sampling unit

- First stage of random selection (e.g. States within Geopolitical zones).
- Can study all members of a selected cluster or randomly select clusters within a cluster (e.g. LGAs within State → Wards within LGA).
- Only in final cluster selected are the individual units listed and sampled.

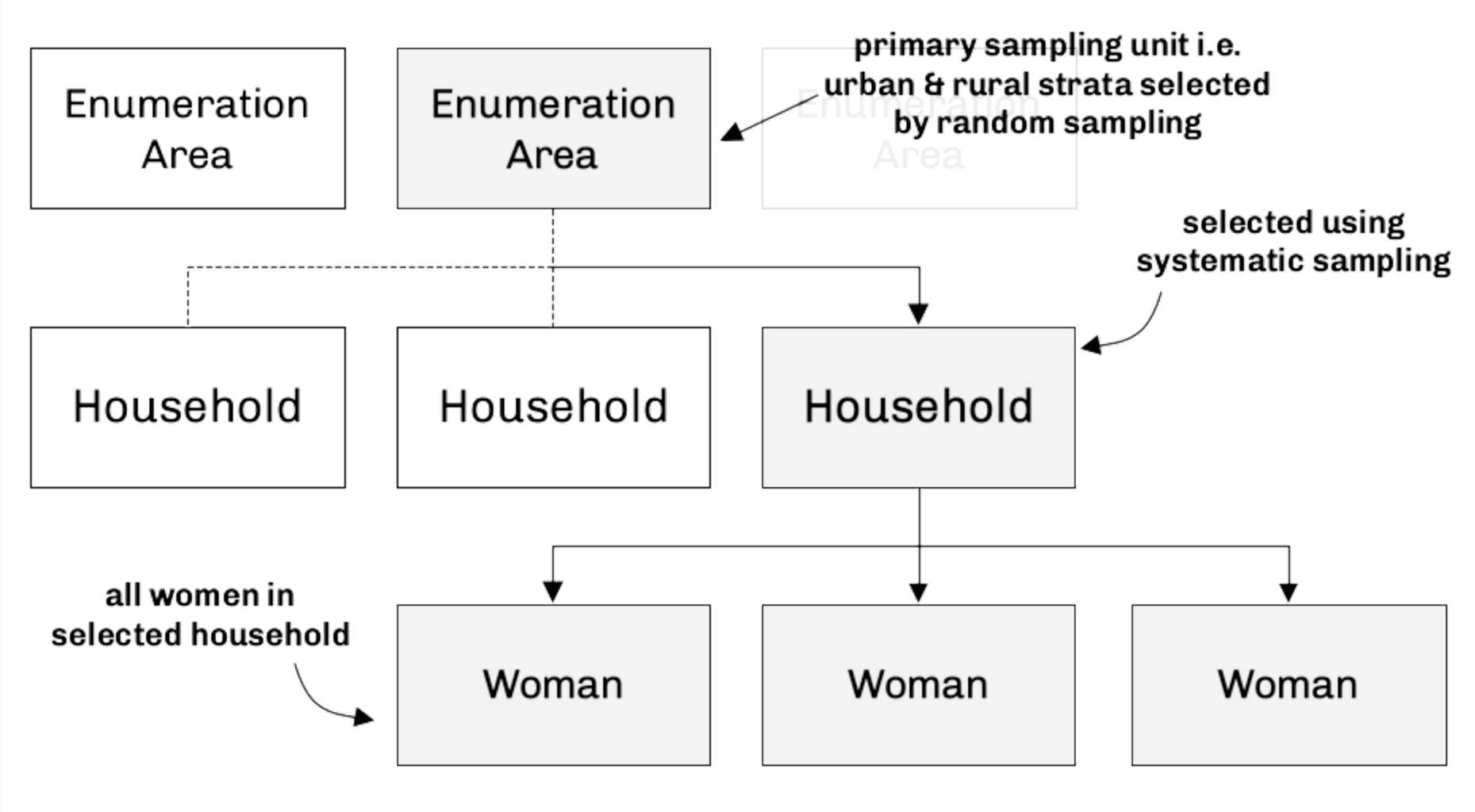
DHS Methodology

The census enumeration areas (EAs) from the 2006 Population and Housing Census were the sampling frame for the survey. These EAs were the primary sampling units (i.e., the clusters) for the surveys. The sample for the 2010 survey was selected using a two-stage stratified design.

During the first stage of the 2010 Nigeria MIS sample, the survey selected 240 clusters: 83 in the urban areas and 157 in the rural areas.

A list of all the households in the selected EAs served as the sampling frame for selecting households in the second stage. The surveys selected households in the second selection stage using equal probability systematic sampling.

All women aged 15–49 years old in the selected households were eligible for individual interviews.


Enumeration Area
Enumeration Area
primary sampling unit i.e. urban & rural strata selected by random sampling
selected using systematic sampling
Household
Household
Household
all women in selected household
Woman
Woman
Woman

Unequal selection probabilities

- Clusters typically differ in size
- Need to adjust so that sample will be representative of population
  - Proportional sampling
  - Weighted analysis

Proportional sampling

Community	Population	Cumulative Pop.
1	1000	1000
2	400	1400
3	200	1600
4	300	1900
5	1200	3100
6	1000	4100

Proportional sampling

- Divide total population by no. of clusters to be sampled
  - if 3 clusters, then 4100 ÷ 3 = 1367
- Generate random number from 1 to 1367 (e.g. 241) to select first cluster (community 1)
- Add sampling interval to select remaining clusters
  - 241 + 1367 = 1608; 1608 + 1367 = 2975 gives communities 4 & 5

Proportional sampling

Advantages

- Leads to communities being selected with probability proportional to size (PPS)
- Self-weighting, so a weighted analysis is not required
- Alternative method of selecting number of members of cluster proportional to size of cluster gives a more biased estimate of true population value

Proportional sampling

Special case

- Possible for same cluster to be chosen twice if population > sampling interval
- More likely if the number of communities selected (sampling fraction) is large
- Proper approach is to select two subsamples of chosen cluster
- Inappropriate to select another community, or to repeat entire sampling procedure until no clusters are repeated

Cluster sample design considerations

- Members of a cluster tend to be more alike than members of the population as a whole
- Members from the same cluster tend to provide less information about the population than do members from different clusters
- This reduction in information translates into estimates that are likely to be less precise than estimates obtained using simple random sampling

Cluster sample design considerations

- Cluster sampling increases the variance and widens confidence intervals of population estimates (reduces study power)
- Cluster sampling requires a larger sample size than would be calculated for simple random sampling
- Increasing the number of clusters is better than increasing the number of members sampled in each cluster

Intracluster correlation

- Intracluster correlation coefficient (r_I) defines how related members of a cluster are. These coefficients are typically small (e.g. 0.01).
- Also called *rate of homogeneity*
- Measures variability between clusters compared with variability within clusters

Intracluster correlation

- Minimal intracluster correlation of demographic variables (e.g. age, sex, marital status): ρ close to zero.
- Intracluster correlation is greater with regard to health practices and socioeconomic variables ρ = 0.1 to 0.3
- Can also be affected by variability of interviewers

Sample size calculation

- Design effect (D) is the ratio of the number of subjects required using cluster sample vs. random sample.
- For cluster sampling with *m* members per cluster and a correlation within clusters of ρ for the variable under study, the design effect is given by:

$$D = 1 + (m - 1)\rho$$

- Better to have many clusters with fewer subjects per cluster than few clusters with many subjects

Estimating ρ and m

- Number in each cluster (m) is best chosen on practical grounds (e.g. no. of units that can be surveyed in 1 day)
- Intracluster correlation (ρ) can be estimated based on type of variable studied

Number of clusters

$$c = \frac{P(1-P)D}{E^2 m}$$

- *P* = estimated proportion with condition
- *D* = design effect
- *E* = sampling error
- *m* = number of subjects in each cluster