Epidemiologic Study Design: Descriptive Cross-sectional Study (Surveys) and Sampling

Dr Ko Ko Zaw
MBBS, MPH, PhD
Professor/Head, Epidemiology Department
University of Public Health

Learning Objectives

- List and discuss the choices for survey administration
- Describe why sampling is important in applied research
- Explain what distinguishes a probability sample from a non-probability sample
- List and discuss the types of probability sampling

What is a Survey?

Survey = Observational or descriptive non-experimental study in which information is collected systematically from individuals or other units (households, businesses, etc.)

- ***Census*** = from everyone in population
- ***Sample Survey*** = from sample of population
- ***Poll*** = for political or public opinion information

Examples of Surveys

- Political opinion polls
- Population-based HIV Impact Assessment Survey (Uganda and elsewhere)
- Myanmar Demographic and Health Survey 2014-15
- Myanmar Micronutrient and Food Consumption Survey (MMFCS) (2016-2017)
- many others

12 Steps to Conducting a Survey

1. Determine study question / aims
2. Budget, schedule
3. Establish the method of data collection
4. Establish universe, working population and sampling method
5. Establish sample size and inclusion criteria and select the sample

12 Steps to Conducting a Survey

6. Design the data collection instrument
7. Pre-testing the survey
8. Training interviewers
9. Implementing the survey
10. Coding and data entry
11. Analyzing the data
12. Reporting the results

Step 3. Establish Method of Data Collection

- Face-to-face (in-person) interview
- Telephone
- Mail
- Self-administered in group setting, i.e., class
- Internet / online
- Other

Method Advantages / Disadvantages

- Face-to-face (in-person) Interview
  - High response rates; flexibility
  - More complete and accurate answers
  - Not dependent on literacy, educational level, or visual acuity
  - Time consuming; potential observer bias
- Telephone Interview
  - Inexpensive; rapid; large numbers or area
  - Non-response; no visual cues; rushed; potential observer bias

Method Advantages / Disadvantages

- Mail
  - Inexpensive; rapid; large numbers or area
  - Non-response; complexity
- Self-administered in Group Setting
  - Requires higher-level approval
  - High response rates
  - answers slanted by peer-pressure or fear of review by higher-level authority
- Online / Internet
  - Limited to skilful users of computer w/ Internet access
  - Non-response
  - Can target large numbers

Step 4. Establish Study Universe, Sampling Frame, and Sampling Method

- **Study universe / Target population**: group of people who are relevant to the study being conducted

Census

- “Exhaustive survey”
- Every member of population included
- Provides **true population value**
- With limited resources, only possible in small, geographically concentrated population
- Rare

Stankovic Camp II,
Skopje, Macedonia,

What is Sampling?

- **Sampling** = Procedure by which some members of the population are selected as representatives of the entire population
- **Objective:** to make observations or measurements on these members, and draw inferences regarding the entire population

Exhaustive Surveys vs. Sampling

Exhaustive Survey	Sample
▪ Measure all individuals	▪ Measure subset of individuals
▪ Obtain true population value	▪ Obtain estimate of value
▪ No confidence interval	▪ Calculate confidence interval

Why sample?

- Gather information from large population using smaller number of people
- Compared with census
  - Can be done at lower cost
  - Can be done in less time
  - Requires fewer resources
- Reasonable (and calculable) accuracy

Probability vs. Non-probability Sampling

Probability

- Based on statistical theory
- Uses random selection of subjects — each has known probability of being selected

Non-probability

- Not based on statistical theory
- Does not use random selection of subjects

Step 4. Establish Study Universe, Sampling Frame, and Sampling Method

- **Study universe / Target population**: group of people who are relevant to the study being conducted
- **Sampling frame**: List of people in the target population
- **Sample**: people from target population selected to be in the study

Sampling Terminology 1

- **Study Universe / Sampling Universe / Target Population / Source Population**: Population from which a sample will be selected
- **Sampling frame**: List of people in the target population

- **Sample**: people from target population selected to be in the study

Probability Sampling Methods

- **Simple random sample**: number chosen at random from random number table
- **Systematic sample**: every *n*th entry of a list (ideally, randomly sorted) chosen based on total N and number to be sampled
- **Stratified random sample**: strata are chosen and simple random samples are chosen within strata
- **Cluster sample**: sampling within randomly selected clusters

Simple Random Sampling

- **Principle**: Each unit (individual) has same, non-zero probability of being selected for sample
- **Procedure**
  - List all individuals
  - Use random numbers to select

Simple Random Sampling

Example

- Clinic satisfaction survey
- Sample size (n=50)
- Use clinic log book
- Assign random numbers
- Randomly select 50 patients
  - Table of random numbers
  - Paper slips in a bag/hat
  - Computer generated random numbers
- Conduct the survey

Simple Random Sampling – Example

Draw sample of 5 people

Number	Household
1	Kazoora
2	Amanya
3	Nsubuga
4	Bibodi
5	Musoke
6	Patel
7	Wasswa
8	Olwenyi
9	Gitta
0	Mbazzi

Random number table

7648	2352	6959	1937
2554	6804	9098	4316
4318	2346	7276	1880
7136	9603	0163	3152
7000	2865	8357	4475
9804	0042	1106	7949
2932	9958	9582	2235
1140	1164	7841	1688
4097	8995	5030	1785
5420	0125	4953	1332
5540	6278	1584	4392
3258	1374	1617	7427

Simple Random Sampling – Example

Draw sample of 5 people

Number	Household
1	Kazoora
→ 2	Amanya
3	Nsubuga
→ 4	Bibodi
5	Musoke
→ 6	Patel
7	Wasswa
→ 8	Olwenyi
9	Gitta
→ 0	Mbazzi

Random number table

7648	2352	6959	1937
2554	6804	9098	4316
4318	2346	7276	1880
7136	9603	0163	3152
7000	2865	8357	4475
9804	0042	1106	7949
2932	9958	9582	2235
1140	1164	7841	1688
4097	8995	5030	1785
5420	0125	4953	1332
5540	6278	1584	4392
3258	1374	1617	7427

Simple Random Sampling

- **Advantage**:
  - selection not biased
  - sampling error easily determined
- **Disadvantage**:
  - need complete list of individuals
  - Individuals may be scattered and poorly accessible
  - Can be expensive
- **Use**: small, geographically concentrated population

Systematic Sampling

- **Principle**: Units drawn with equal interval between units (data should not be ordered)
- **Procedure**
  - Calculate **sampling interval** (**k** = N / n)
  - Use random number <k to begin
  - Select every k^{th} unit from first unit
- **Analysis**: same as simple random sampling

Systematic Sampling – Example

Draw sample of 5 people
k = **10 / 5 = 2**

Number	Household
1	Kazoora
2	Amanya
3	Nsubuga
4	Bibodi
5	Musoke
6	Patel
7	Wasswa
8	Olwenyi
9	Gitta
0	Mbazzi

Random number table

7648	2352	6959	1937
2554	6804	9098	4316
4318	2346	7276	1880
7136	9603	0163	3152
7000	2865	8357	4475
9804	0042	1106	7949
2932	9958	9582	2235
1140	1164	7841	1688
4097	8995	5030	1785
5420	0125	4953	1332
5540	6278	1584	4392
3258	1374	1617	7427

Systematic Sampling

- **Advantage**: faster and easier than Simple RS
- **Disadvantages**:
  - need complete list of individuals
  - can be biased if list has pattern
- **Use**: small scale survey in geographically concentrated population

Stratified Random Sampling

- **Principle**: Population is divided into sub-groups (age, sex, etc.) and sample should reflect them
- **Procedure**
  - Identify homogeneous sub-groups or strata
  - Construct sampling frame in each stratum
  - Sampling in each stratum independently

More Terminology

- **Sampling unit**: Entity (individual, household, school, etc.) selected during a sampling process

- **Primary sampling unit** (PSU) = sampling unit at the **first** stage sampling in stratified and cluster surveys (e.g., district, school, household)
- **Basic or elementary or secondary sampling unit** (SSU) = sampling unit at the **second** stage sampling in stratified and cluster surveys (e.g., individual)

Stratified Random Sampling – Example

- Clinic patients: 81% women, 19% men

- Divide patients into 2 groups (“strata”)
- Create sampling frame for each group
- Select a random sample from each group
  - Can be proportional to source population
  - Can oversample small strata, but then need to use weights in analysis
- Conduct the survey on the 50 selectees

Stratified Random Sampling

- **Advantages**:
  - Each subgroup is represented in sample
  - allows for oversampling
  - can get separate estimates (such as prevalence) from the whole population <u>and</u> from individual strata
- **Disadvantage**:
  - Sampling error more difficult to measure

What if…?

- You do not have a complete list of basic sampling units

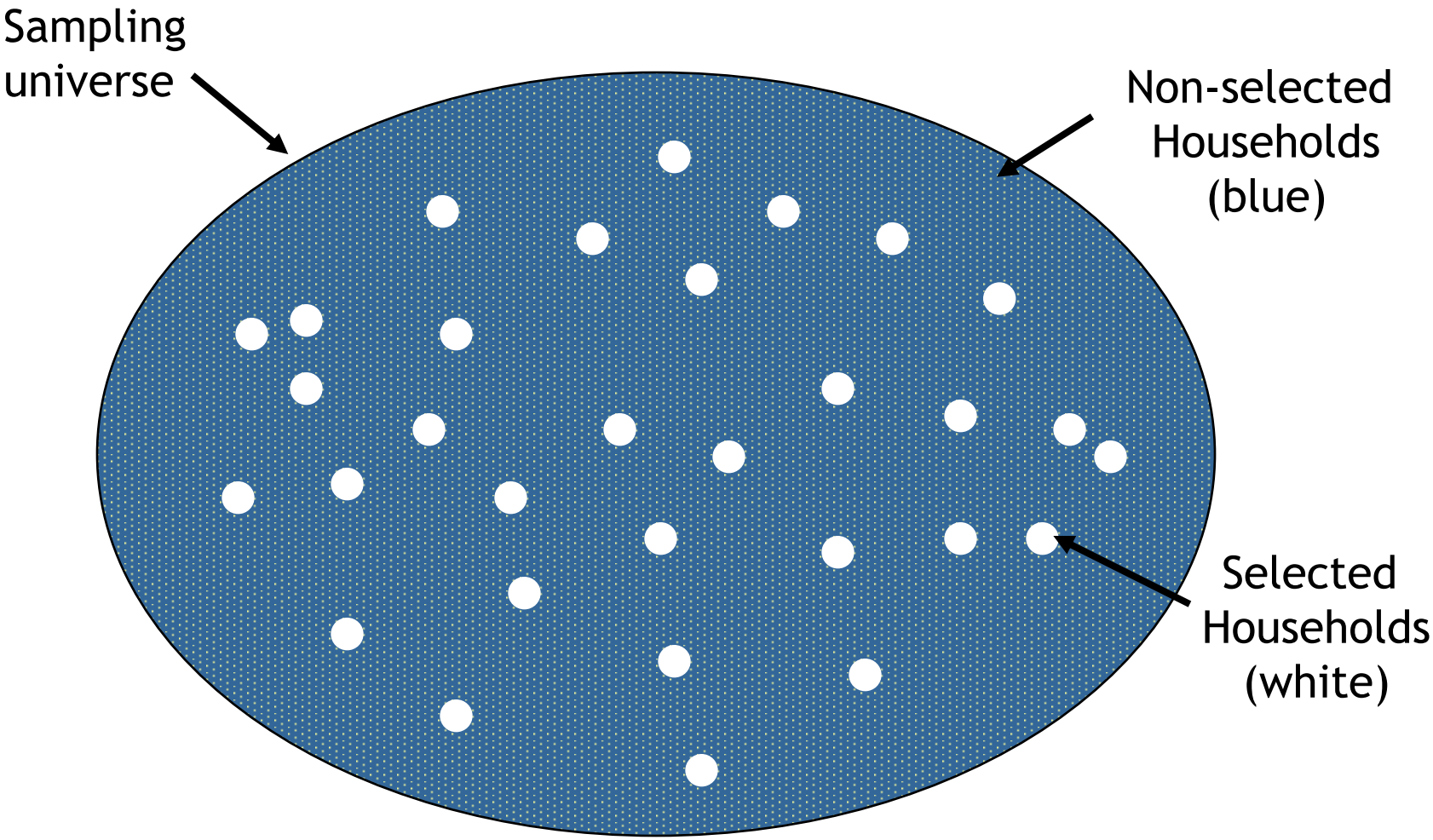
  or

- Survey population is geographically dispersed, so SRS or systematic sampling is impractical

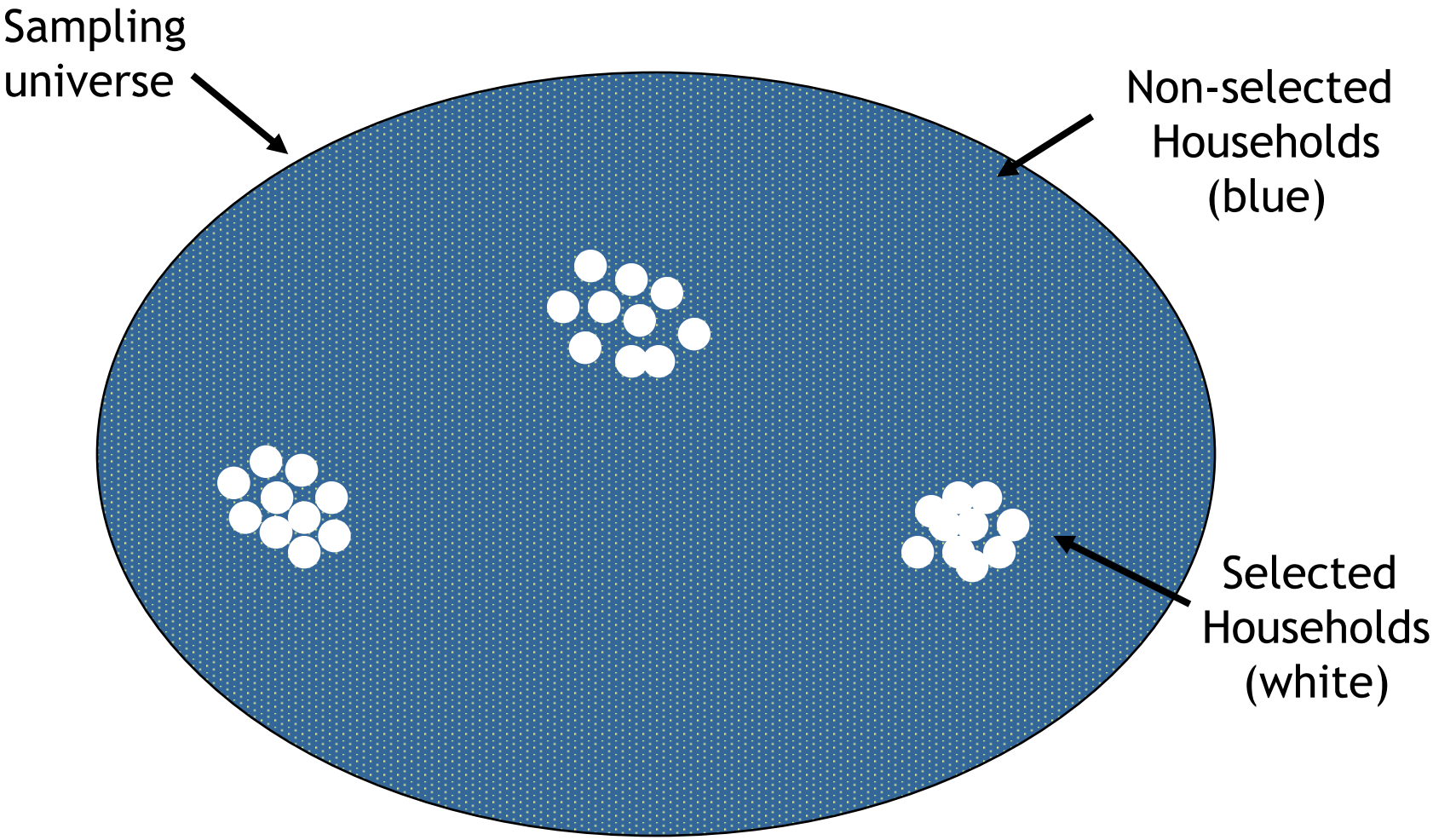
Cluster Sampling

- **Principle**: Random sample of clusters (e.g., villages, census tracks), then sample within
- **Procedure**:
  - Select PSUs from list of villages, census tracks – done during planning stage, ***in the office***
  - In selected clusters, include all or sample (SRS or systematic) of SSUs, done ***in the field***

Simple Random Sampling (30 households)



Cluster Sampling (30 households – 3 clusters with 10 HHs)



Cluster Sampling

- **Advantages**:
  - Often most feasible method in field
  - Efficient – basic sampling units closer together
  - Does not require list of every individual in pop.
- **Disadvantage**:
  - Requires larger sample
  - May require weighted analysis*

* WHO 30x7 does not!

Cluster Sampling – Stage 1

- The **probability** of each village being selected is **proportional** to the **size** of its population (PPS)
- PPS assures that each household within the survey area has an equal (known and non-zero) chance of being selected

Probability Proportional to Size (PPS)

Not PPS

Village

Tsaag-annur	Nogoon-nuur	Ulgii	Altant-sogts	Bugat	Bayan-nuur
231	912	3,099	376	484	763

PPS

Tsaag-annur	Nogoon-nuur	Ulgii	Altant-sogts	Bugat	Bayan-nuur
231	912	3,099	376	484	763

Cluster Sampling: Stage 1

1. Construct a list of primary sampling units (e.g. camp sections), and estimated population size of each
2. List the cumulative population in an adjacent column
3. Calculate **sampling interval**, by dividing total population by number of clusters
4. Pick a random start between 1 and sampling interval
5. Select first cluster
6. Add sampling interval to start number to identify 2nd cluster
7. Continue until all clusters have been selected

30x7 Cluster Survey

Stage 1

- Probability proportionate to size (self-weighting in analysis)
- Need list of villages, estimated population of each
- Determine interval by dividing total population by 30
- List villages, start at random starting point for first cluster
- Add interval, identify second cluster
- Repeat for 30 clusters

Stage 2

- Upon arrival in village choose random starting location, then select houses until 7 children are found

30x7 Cluster Survey

Village	Estimated Pop.	Cum. Pop.	Range	
A	250	250	1 – 250	
B	2,500	2,750	251 – 2750	
C	400	3,150	2751 – 3150	
D	650	3,800	3151 – 3800	
E	300	4,100	3801 – 4100	
F	1,500	5,600	4101 – 5600	
G	800	6,400	5601 – 6400	
H	750	7,150	6401 – 7150	
I	1,200	8,350	7151 – 8350	
J	900	9,250	8350 – 9250	
etc.	etc.	etc.	etc.– 30000	
Total	30,000	30,000		

30x7 Cluster Survey

Village	Estimated Pop.	Cum. Pop.	Range	
A	250	250	1 – 250	1
B	2,500	2,750	251 – 2750	2
C	400	3,150	2751 – 3150	0
D	650	3,800	3151 – 3800	1
E	300	4,100	3801 – 4100	0
F	1,500	5,600	4101 – 5600	2
G	800	6,400	5601 – 6400	1
H	750	7,150	6401 – 7150	0
I	1,200	8,350	7151 – 8350	2
J	900	9,250	8350 – 9250	1
etc.	etc.	etc.	etc.– 30000	20
Total	30,000	30,000		30

167
1,167
2,167
3,167
4,167
5,167
6,167
7,167
8,167
9,167
⋮
29,167

30x7 Cluster Survey

Village	Estimated Pop.	Cum. Pop.	Range		
A	250	250	1 – 250	1	0
B	2,500	2,750	251 – 2750	2	3
C	400	3,150	2751 – 3150	0	0
D	650	3,800	3151 – 3800	1	1
E	300	4,100	3801 – 4100	0	0
F	1,500	5,600	4101 – 5600	2	1
G	800	6,400	5601 – 6400	1	1
H	750	7,150	6401 – 7150	0	1
I	1,200	8,350	7151 – 8350	2	1
J	900	9,250	8350 – 9250	1	1
etc.	etc.	etc.	etc.– 30000	20	21
Total	30,000	30,000		30	30

613
1,613
2,613
3,613
4,613
5,613
6,613
7,613
8,613
⋮
29,613

PPS Cluster Sampling — Advantages, Disadvantages

- **Advantages**
  - Does not require rosters
  - Simple analysis (no weights required)
  - Efficient
  - Proven
- **Disadvantages**
  - Cannot analyze subgroups
  - Loss of precision due to correlation within clusters (need to account for “design effect”)

Non-probability Sampling

- Methods
  - **Subjective / Purposive / Judgment** – select key people
  - **Convenience** – invite reachable people
  - **Respondent-driven, Snowball** – ask participants to bring in friends
  - **Volunteer sampling** – invite volunteers to participate
  - **Quota sampling** – identify predetermined number of people
  - Other
- Advantages – easier, cheaper, quicker
- Disadvantages
  - Often biased, not representative of population of interest

Remaining Steps

6. Design the data collection instrument
7. Pre-test the data collection instrument
8. Train interviewers
9. Implement the survey
10. Code and enter data
11. Analyze the data
12. Report the results

Q1. What type of sampling?

a. Every 10th listing in pop. register	**Systematic**
b. Pick names out of a hat.	**Random**
c. Approach shoppers at a mall.	**Convenience**
d. Randomly select 5 students from each class in an elementary school.	**Stratified**
e. Ask each enrollee to bring in 3 acquaintances.	**Snowball / RDS**

Q2. Probability vs. Non-Probability

Cluster	**Probability**
Convenience	**Non-Probability**
Respondent-driven	**Non-Probability**
Simple random	**Probability**
Stratified random	**Probability**
Subjective	**Non-Probability**
Systematic	**Probability**
Volunteer	**Non-Probability**

Q3. Probability vs. Non-Probability

Q3a. Which is more likely to provide representative results?

A3a. **Probability**

Q3b. Which is usually easier to conduct?

A3b. **Non-Probability**

Q4. Need roster?

Q4. Which type of probability sampling requires having a roster (sampling frame) of potential participants?

Cluster	**Do not need**
Simple random	**Need**
Stratified random	**Need**
Systematic	**Need**

Q5. Which sampling method here?

Q6. Which sampling method here?

Concluding Points

- Primary reason for selecting sample is to draw inferences about a population without having to enroll every member
- Probability sampling (rather than non-probability sampling) is necessary to obtain valid results
- Several types of probability samples, each with its advantages and disadvantages
- Realities in the field usually guide choice of sampling strategy