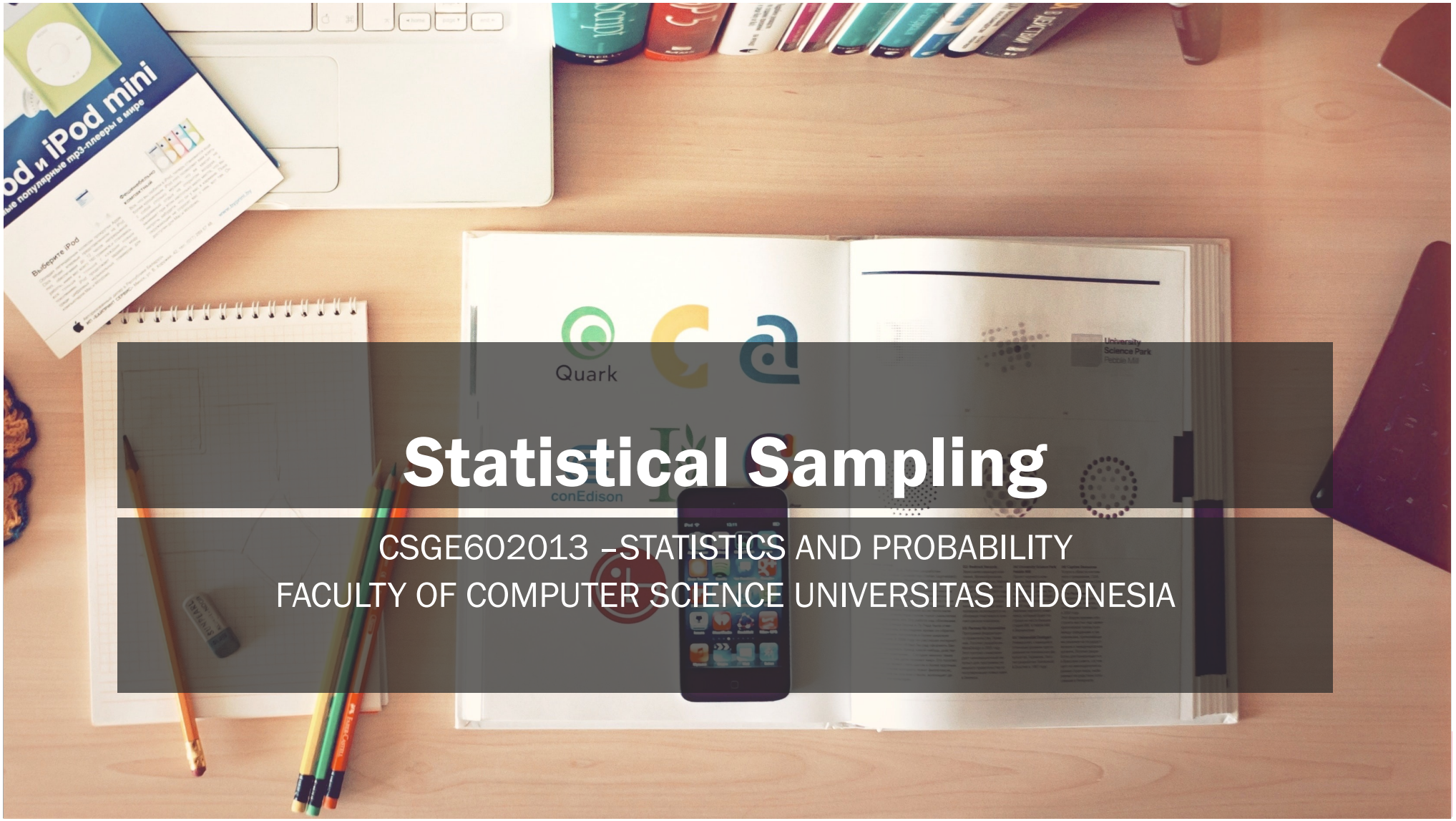


Statistical Sampling

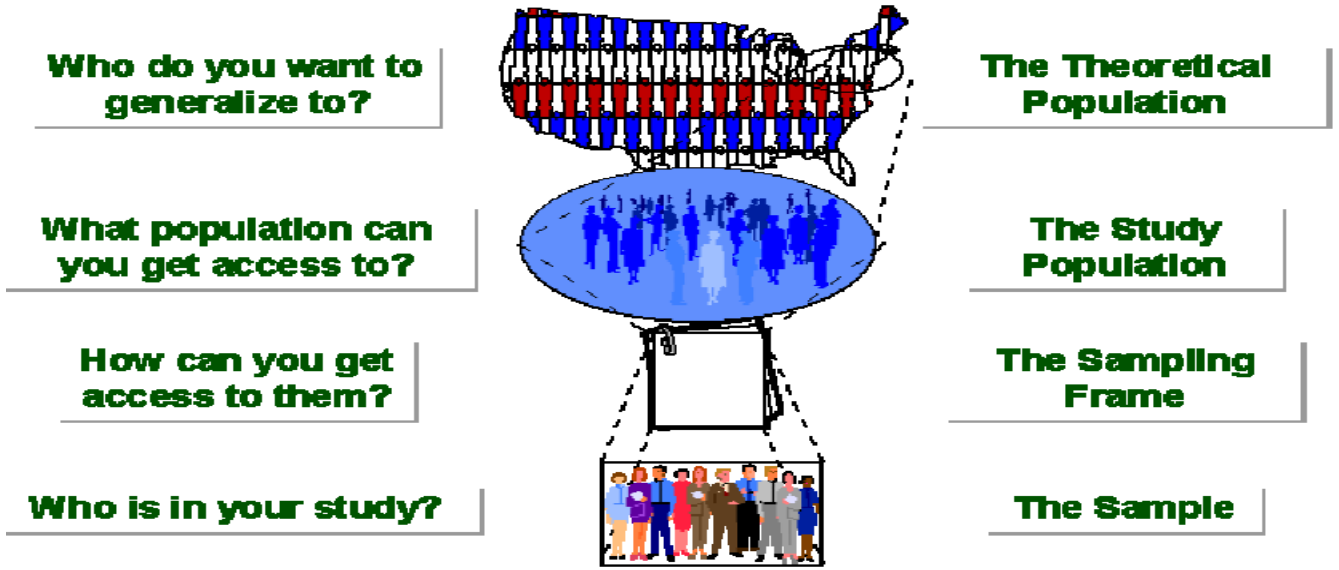
CSGE602013 –STATISTICS AND PROBABILITY

FACULTY OF COMPUTER SCIENCE UNIVERSITAS INDONESIA

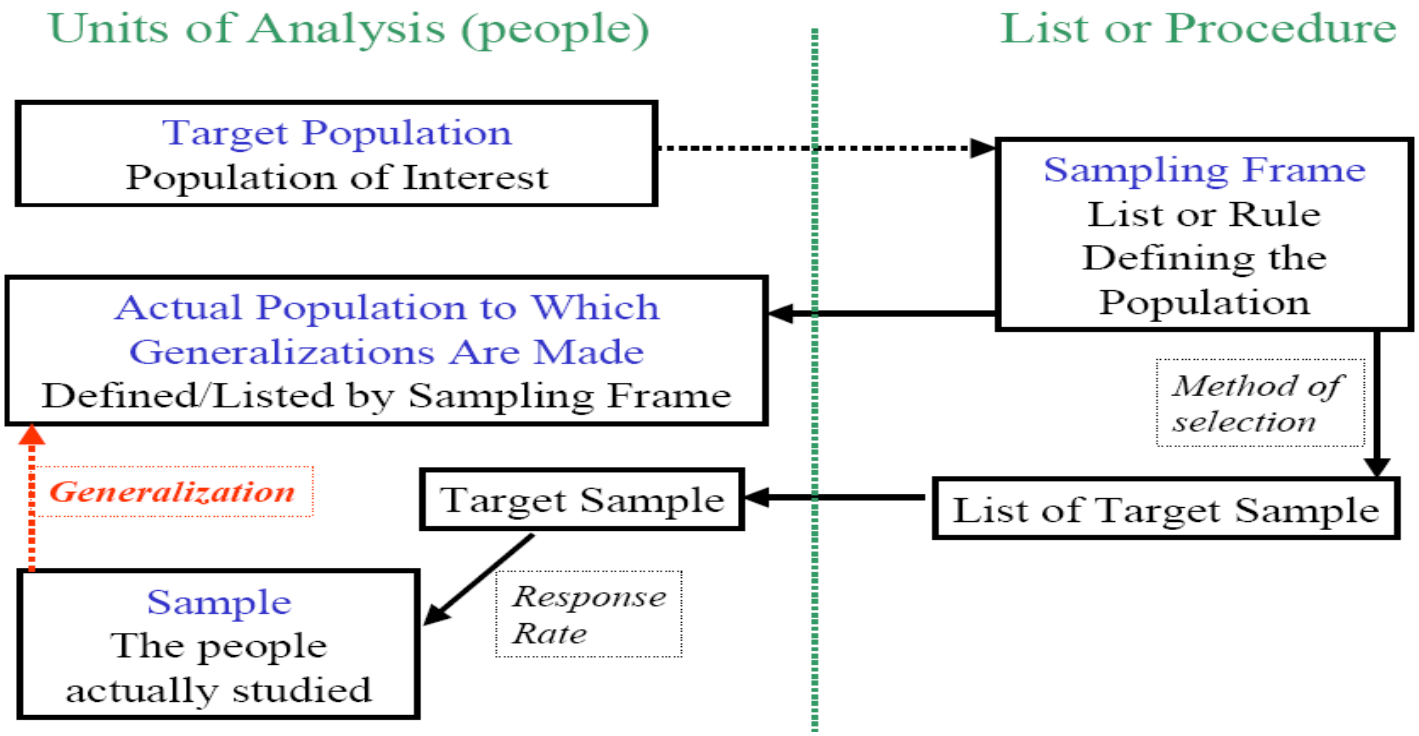
Credits

- Hawkes, James S., and William H. Marsh. *Discovering statistics*. Hawkes, 2004.
- Sarah DiCalogero - Statistical Sampling
- YouTube. (2009). *Random Sample*. Retrieved from http://www.youtube.com/watch?v=xh4zxC1OpiA&feature=related
- YouTube. (2009). *Types of Random*. Retrieved from http://www.youtube.com/watch?v=wUwH7SIfg9E&feature=related

Key Sampling Concepts



Sampling Process



- Key Ideas
  - Distinction between the population of interest and the actual population defined by sampling frame.
  - Generalizations can be made only to the actual population
  - Understand crucial role of the sampling frame
- Sampling frame
  - The list or procedure defining the population
  - Distinguish sampling frame from sample
  - Example: telephone book, voter list, random digit dialing
  - Essential for probability sampling, but can be defined for non-probability sampling

Sampling Methods

Probability Sampling

- Simple Random
- Systematic Random
- Stratified Random
- Random Cluster
- Stratified Cluster
- Complex Multi-stage Random

Non-probability Sampling

- Quota
- Purposive
- Convenience

Probability Sampling

- Each element of population has a known non-zero probability of selection

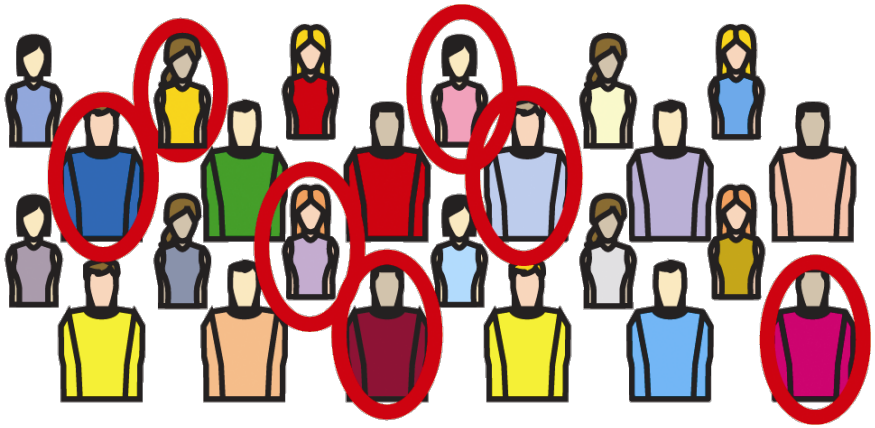
Thus,

- If some elements of population have zero probability of selection (cannot be selected) ⟶ non-probability sampling
- If probabilities of selection are not known ⟶ non-probability sampling

Five Basic Sampling Methods

Simple Random Sampling

- The “pick a name out of the hat” technique
  - Random number table
  - Random number generator



Hawkes and Marsh (2004)

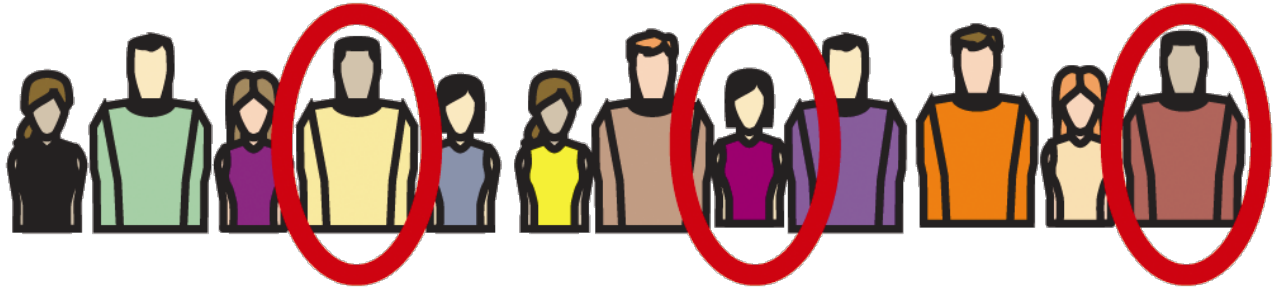
Simple Random Sampling

- Each element in the population has an equal probability of selection **and** each combination of elements have an equal probability of selection
- Each selection is independent of other selections

Hawkes and Marsh (2004)

Systematic Sampling

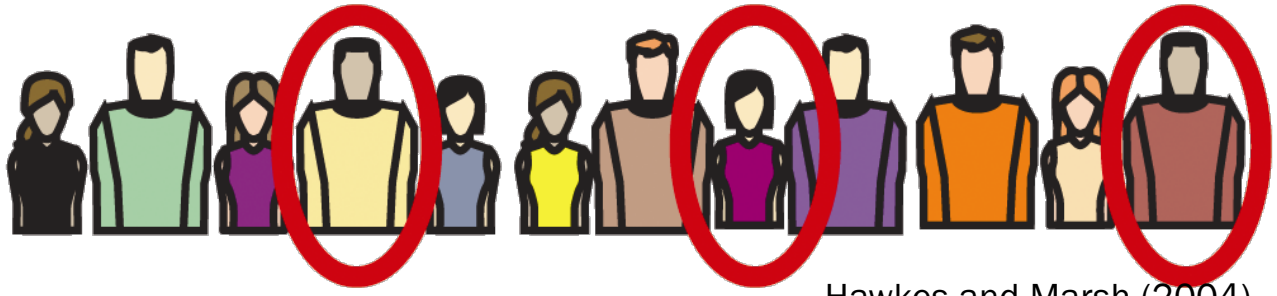
- All data is sequentially numbered
- Every *n*th piece of data is chosen
- Each element has an equal probability of selection, but combination of elements has different probabilities



Hawkes and Marsh (2004)

Systematic Sampling

- Population size = N, desired sample size = n, size of each sampling interval = N/n = k, which means number of sampling interval = n.
- Randomly select a number j between 1 and k (inclusive), then sample j-th element in each sampling interval e.g., N = 18, n = 6, k = 3, and random j = 3. Select 3rd element, 6th, 9th, and so on



Hawkes and Marsh (2004)

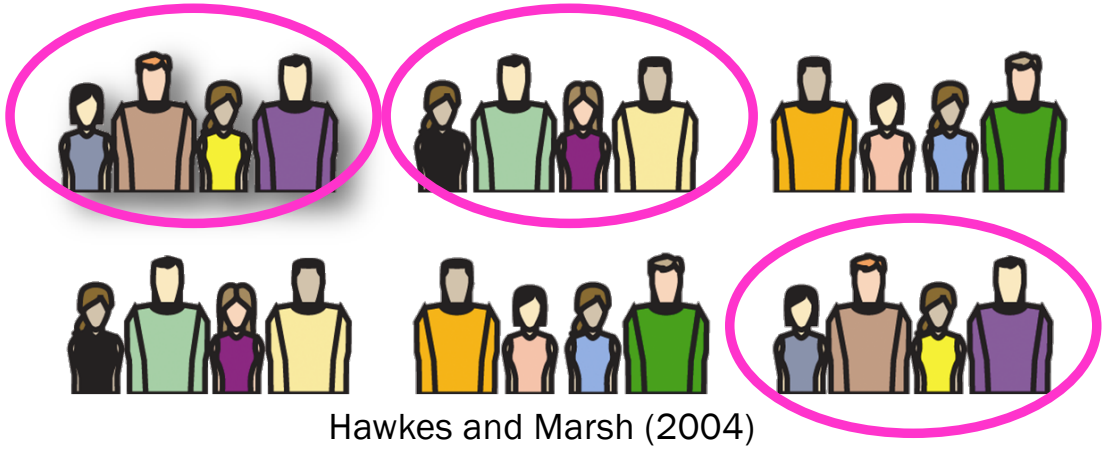
Stratified Sampling

- Data is divided into subgroups (strata)
- Strata are based specific characteristic
  - Age
  - Education level
  - Etc.
- Use random sampling within each strata
- Probabilities of selection may be different for different groups, as long as they are known

Hawkes and Marsh (2004)

Cluster Sampling

- Data is divided into clusters
  - Usually geographic
- Random sampling used to choose clusters
- All data used from selected clusters



Hawkes and Marsh (2004)

Stratified Sampling vs Cluster Sampling

- Stratified Sampling
  - Divide population into groups different from each other: sexes, races, ages
  - Sample randomly from each group
  - Less error compared to simple random
  - More expensive to obtain stratification information before sampling
- Cluster Sampling
  - Divide population into groups: schools, cities
  - Randomly sample some of the groups
  - More error compared to simple random
  - Reduces costs to sample only same areas or organizations

Convenience Sampling

- Data is chosen based on convenience
- No reason tied to purposes of research, for example: students in your class, people on State Street, friends, etc
  - BE WARY OF BIAS!

Hawkes and Marsh (2004)

Example 1: Sampling Methods

- In a class of 18 students, 6 are chosen for an assignment

Random	Pull 6 names out of a hat
Systematic	Selecting every 3rd student
Stratified	Divide the class into 2 equal age groups. Randomly choose 3 from each group
Cluster	Divide the class into 6 groups of 3 students each. Randomly choose 2 groups
Convenience	Take the 6 students closest to the teacher

Example 2: Utilizing Sampling Methods

- Determine average student age
  - Sample of 10 students
  - Ages of 50 statistics students

18	21	42	32	17	18	18	18	19	22
25	24	23	25	18	18	19	19	20	21
19	29	22	17	21	20	20	24	36	18
17	19	19	23	25	21	19	21	24	27
21	22	19	18	25	23	24	17	19	20

Example 2 – Random Sampling

- Random number generator

Data Point Location	Corresponding Data Value
35	25
48	17
37	19
14	25
47	24
4	32
33	19
35	25
34	23
3	42
Mean	25.1

Example 2 – Systematic Sampling

- Take every 5^{th} data point

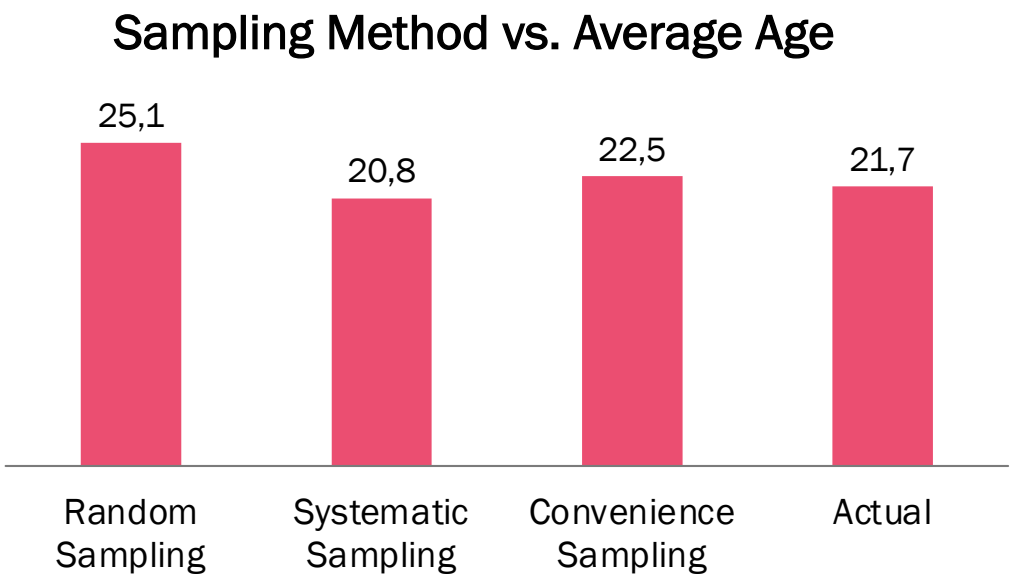
Data Point Location	Corresponding Data Value
5	17
10	22
15	18
20	21
25	21
30	18
35	21
40	27
45	23
50	20
Mean	20.8

Example 2 – Convenience Sampling

- Take the first 10 data points

Data Point Location	Corresponding Data Value
1	18
2	21
3	42
4	32
5	17
6	18
7	18
8	18
9	19
10	22
Mean	22.5

Example 2 - Comparison



Sampling in Practice

- Often a non-random selection of basic sampling frame (city, organization, etc)
- Fit between sampling frame and research goals must be evaluated
- Sampling frame as concept is relevant to all kinds of research (including non-probability)
- Probability sampling means you can generalize to the population defined by the sampling frame.
- Non-probability sampling means you cannot generalize beyond the sample.

Sample Size

- Heterogeneity: need larger sample to study more diverse population
- Desired precision: need larger sample to get smaller error
- Sampling design: smaller if stratified, larger if cluster
- Nature of analysis: complex multivariate statistics need larger samples
- Accuracy of sample depends upon sample size, not ratio of sample to population