

Course #12

Biostatistics

Nelly Ali-eldin, MD
Department of Biostatistics & Cancer
Epidemiology
NCI, Cairo Univeristy

Biostatistics

- 14 sessions each 1 hour and half
- 12 lectures plus 2 reviews
- Lecturers:
 - Prof. Inas El-attar, PHD
 - Prof. Nelly Alieldin, MD
- Notes “Biostatistics, course #12”
- Other useful references for further readings:
 - “Basic & clinical biostatistics” (Dawson-Saunders and Trapp)
 - “Medical statistics” (Bretty Kirkwood and Jonathan Sterne)
 - Presentations available on NCI, web-site:
www.nci.cu.edu.eg

Objectives

- Not to convert clinicians and health professionals into biostatisticians or experts, but:
- To understand the principles and methods used in biostatistics
- To be competent in the use of simple and basic tools of biostatistics
- Capable of exercising critical judgement when assessing results reported by others or authors in medical journals
- Have a better communication with biostatisticians

Outline of the course

- Introduction, Definitions & Types of sample
- Types of variable and scales of measurement
- Measures of central tendency and variability
- Data presentation in tables & graphs
- Probability definitions & Normal distribution
- Z score & Sampling distribution
- Hypothesis Testing
- Comparisons between two means (Z-test & T-test)
- Chi-square test
- Linear Regression & Correlation , Analysis of Variance (ANOVA)
- Survival Analysis
- Clinical trials

Lecture #1

Introductory Concepts, Definitions & Sampling

What is statistics

A field of study concerned with:

- Methods for gathering data , sampling and experimental design
- Summarizing data in estimates, tables or graphs,
- Inferential statistics: make generalization to a population when only part of it is withdrawn, a sample

Descriptive vs Inferential statistics

- ***Descriptive statistics:*** deal with the enumeration, organization, and graphical representation of data.
- ***Inferential statistics:*** The process where we can estimate the quality of a larger population by analyzing a small sample

Population and Samples

- A Population is the larger set of objects we wish to study
 - Ex: The number of democrats in the United States
- A Sample is a set of “representative” objects we choose in order to estimate the characteristics of the larger set of objects
 - Ex: Take 100 people from each state and determine whether they are democrats

Parameters and Statistics

- A Parameter is the “quality” of the population we are trying to estimate
- In order to estimate the parameter we measure the quality in a sample. This sample quality is called its statistic

Sampling

- Why we use sampling
- Definitions in sampling
- Sampling errors
- Main methods of sampling
- Sample size calculation



Why do we use sampling?

Get information from large populations with:

- Reduced costs
- Reduced field time
- Increased accuracy with enhanced methods
- Sometimes studying the whole population is impossible
- Sampling error could be estimated



Definition of sampling

Procedure by which some members of a given population are selected as representatives of the entire population

Definition of sampling terms

Sampling unit (element)

- Subject under observation on which information is collected
 - Example: children <5 years, hospital discharges, health events...

Sampling fraction

- Ratio between sample size and population size
 - Example: 100 out of 2000 (5%)

Definition of sampling terms

Sampling frame

- List of all the sampling units from which sample is drawn
 - Lists: e.g. children < 5 years of age, households, health care units...

Sampling scheme

- Method of selecting sampling units from sampling frame
 - Randomly, convenience sample...

Survey errors

- Systematic error (or bias)

Sample not typical of population

- Inaccurate response (information bias)
- Selection bias

- Sampling error (random error)

Representativeness (validity)

A sample should accurately reflect distribution of relevant variable in population

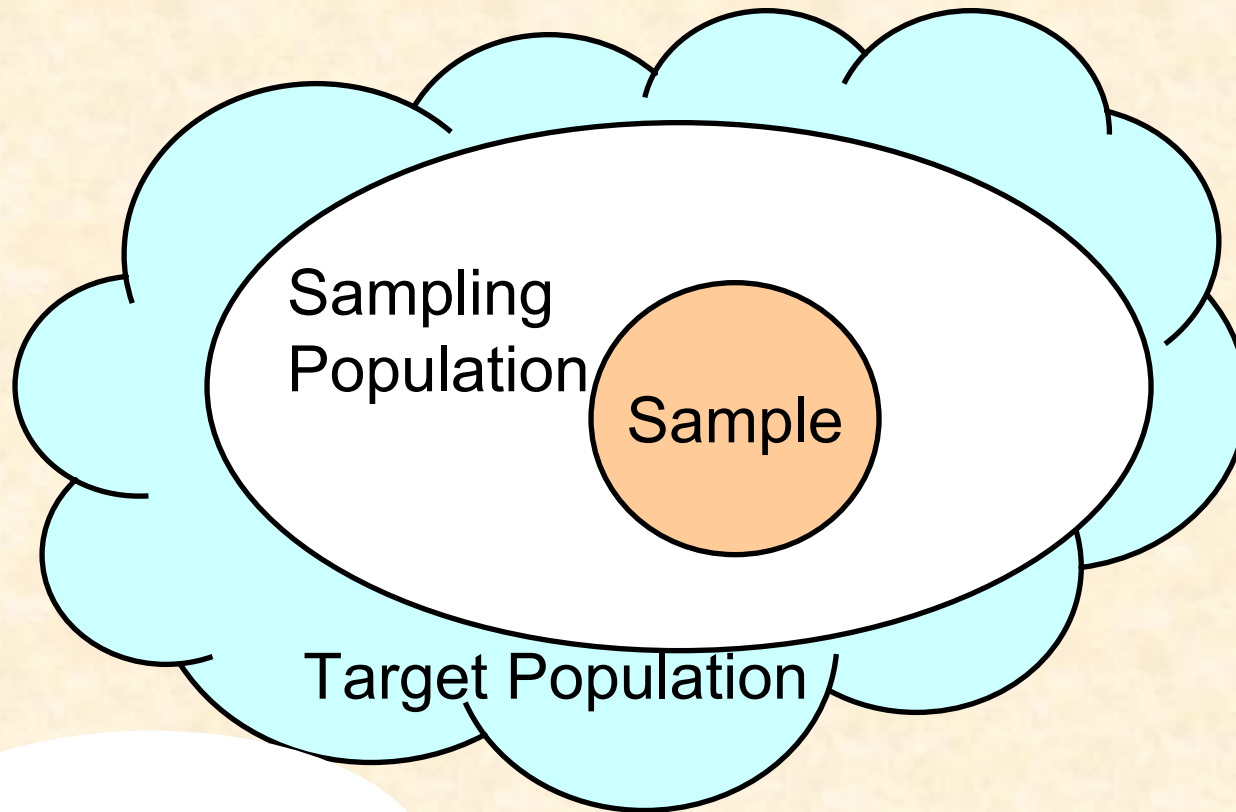
- **Person** e.g. age, sex
- **Place** e.g. urban vs. rural
- **Time** e.g. seasonality

Representativeness essential to generalise


Ensure representativeness before starting,

Confirm once completed

Sampling and representativeness



Target Population → Sampling Population → Sample



Sampling error

- Random difference between sample and population from which sample drawn
- Size of error can be measured in probability samples
- Expressed as “standard error”
 - of mean, proportion...
- Standard error (or precision) depends upon:
 - Size of the sample
 - Distribution of character of interest in population

Sampling error

When simple random sample of size 'n' is selected from population of size N, standard error (s) for population mean or proportion is:

$$\frac{\sigma}{\sqrt{n}} \quad \sqrt{\frac{p(1-p)}{n}}$$

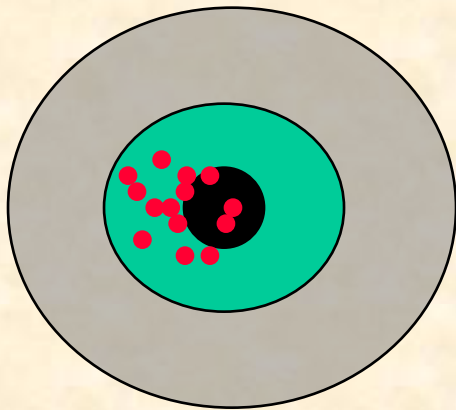
Used to calculate, 95% confidence intervals

Estimated 95%
confidence interval

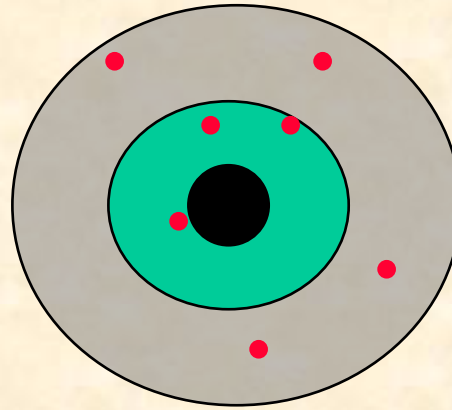
$$\bar{X} \pm 2 \times s_x$$

Quality of a sampling estimate

**Precision
& validity**

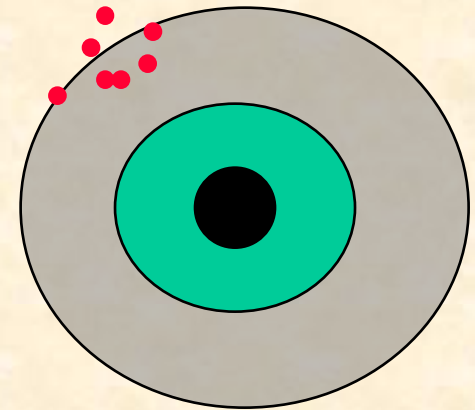


No precision



**Random
error**

**Precision but
no validity**



**Systematic
error (bias)**

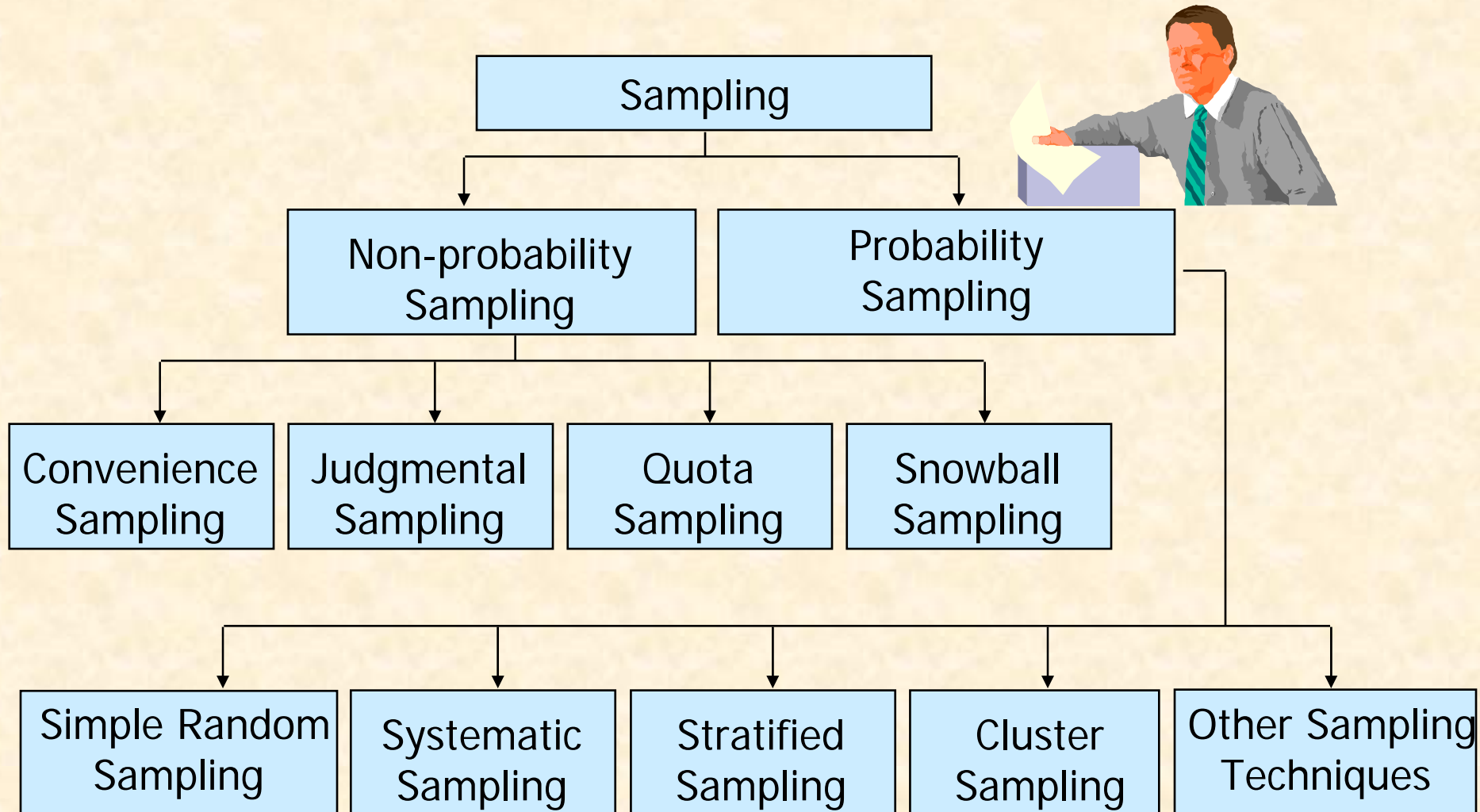
Survey errors: example

Measuring height:

- Measuring tape held differently by different investigators
 - loss of precision
 - Large standard error
- Tape shrunk/wrong
 - systematic error
 - Bias (cannot be corrected afterwards)

179
178
177
176
175
174
173

Types of sampling



Convenience Sampling

Convenience sampling: attempts to obtain a sample of convenient elements. Often, respondents are selected because they happen to be in the right place at the right time.

- use of students, and members of social organizations
- mall intercept interviews without qualifying the respondents
- “people on the street” interviews
- Patients with specific cancer diagnosis attending a clinic

Judgmental Sampling

Judgmental sampling is a form of convenience sampling in which the population elements are selected based on the judgment of the researcher.

- test markets
- expert witnesses used in court
- Patients with advanced HCC but young age with low PS score selected in uncontrolled trial for a new treatment

Quota Sampling

Quota sampling may be viewed as two-stage restricted judgmental sampling.

- The first stage consists of developing control categories, or quotas, of population elements.
- In the second stage, sample elements are selected based on convenience or judgment.

Control Characteristic	<u>Population composition</u>	<u>Sample composition</u>	
	Percentage	Percentage	Number
Sex			
Male	48	48	480
Female	52	52	520
	<hr/> 100	<hr/> 100	<hr/> 1000

Snowball Sampling

In **snowball sampling**, an initial group of respondents is selected, usually at random.

- After being interviewed, these respondents are asked to identify others who belong to the target population of interest.
- Subsequent respondents are selected based on the referrals.

Probability samples

- Random sampling
 - Each subject has a known probability of being selected
- Allows application of statistical sampling theory to results to:
 - Generalise
 - Test hypotheses

Methods used in probability samples

- Simple random sampling
- Systematic sampling
- Stratified sampling
- Multi-stage sampling
- Cluster sampling

Simple random sampling

- Principle
 - Equal chance/probability of drawing each unit
- Procedure
 - Take sampling population
 - Need listing of all sampling units (“sampling frame”)
 - Number all units
 - Randomly draw units



Simple random sampling

- Advantages
 - Simple
 - Sampling error easily measured
- Disadvantages
 - Need complete list of units
 - Does not always achieve best representativeness
 - Units may be scattered and poorly accessible

Simple random sampling

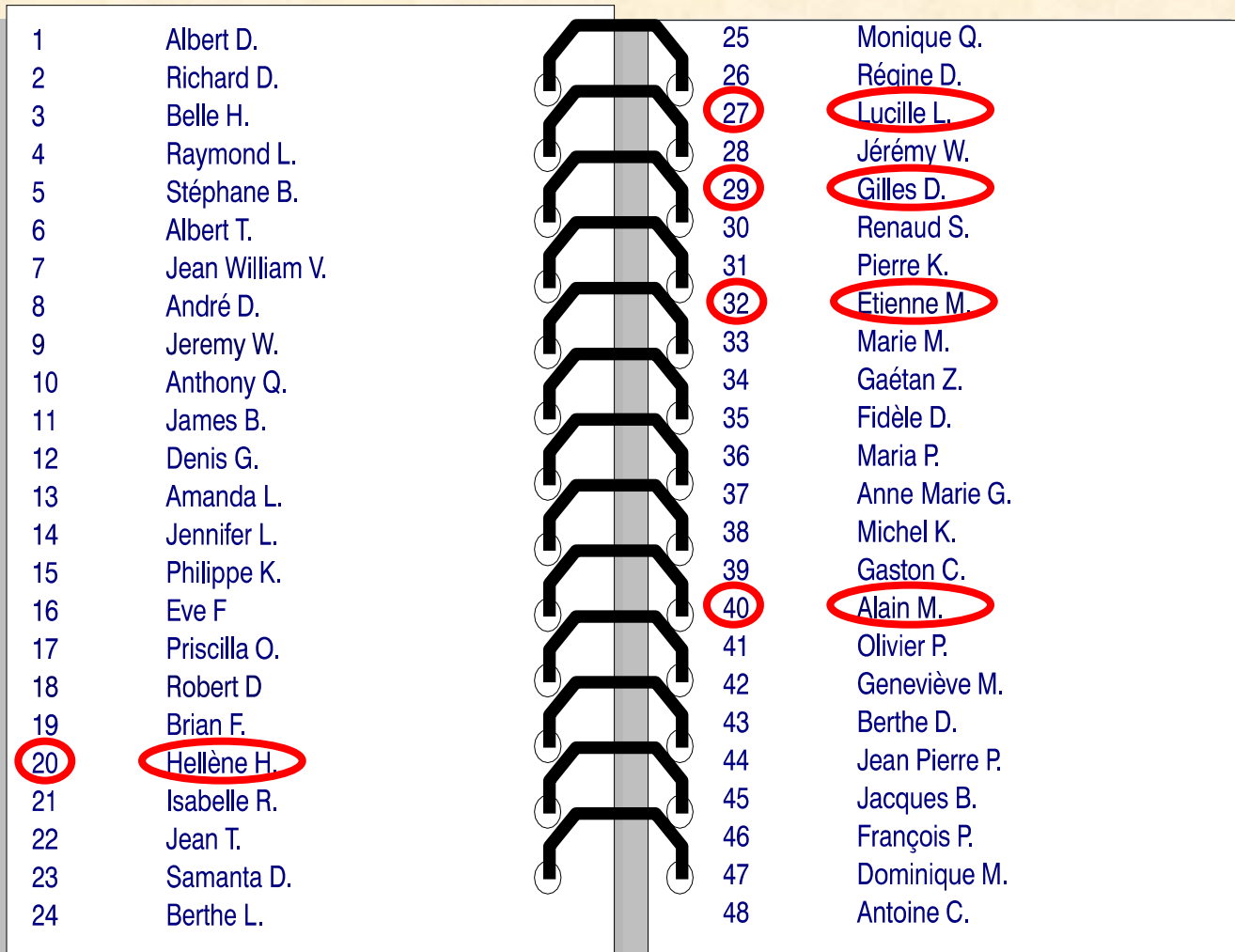
Example: evaluate the prevalence of tooth decay among 1200 children attending a school

List of children attending the school

- Children numerated from 1 to 1200
- Sample size = 100 children
- Random sampling of 100 numbers between 1 and 1200

How to randomly select?

Simple random sampling



1	Albert D.	25	Monique Q.
2	Richard D.	26	RéGINE D.
3	Belle H.	27	Lucille L.
4	Raymond L.	28	Jérémy W.
5	Stéphane B.	29	Gilles D.
6	Albert T.	30	Renaud S.
7	Jean William V.	31	Pierre K.
8	André D.	32	Etienne M.
9	Jeremy W.	33	Marie M.
10	Anthony Q.	34	Gaétan Z.
11	James B.	35	Fidèle D.
12	Denis G.	36	Maria P.
13	Amanda L.	37	Anne Marie G.
14	Jennifer L.	38	Michel K.
15	Philippe K.	39	Gaston C.
16	Eve F	40	Alain M.
17	Priscilla O.	41	Olivier P.
18	Robert D	42	Geneviève M.
19	Brian F.	43	Berthe D.
20	Hellène H.	44	Jean Pierre P.
21	Isabelle R.	45	Jacques B.
22	Jean T.	46	François P.
23	Samanta D.	47	Dominique M.
24	Berthe L.	48	Antoine C.



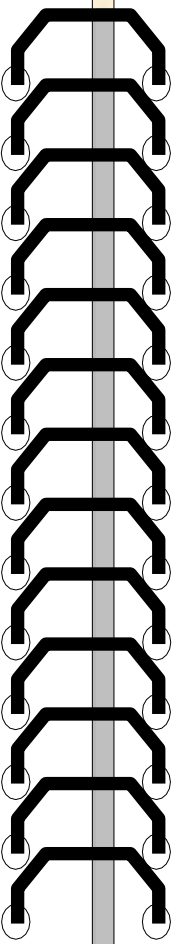
Systematic sampling

- Principle
 - Select sample at regular intervals based on sampling fraction
- Advantages
 - Simple
 - Sampling error easily measured
- Disadvantages
 - Need complete list of units
 - Periodicity

Systematic sampling

- $N = 1200$, and $n = 60$
 \Rightarrow **sampling fraction** = $1200/60 = 20$
- List persons from 1 to 1200
- Randomly select a number between 1 and 20 (ex : 8)
 \Rightarrow 1st person selected = the 8th on the list
 \Rightarrow 2nd person = $8 + 20 =$ the 28th etc

Systematic sampling



The diagram illustrates systematic sampling using a vertical line with 48 numbered points. A zigzag line connects the points, starting at 1, going up to 25, then zigzagging down to 48. The points are numbered 1 through 48, with the numbers 8, 28, and 48 circled in red. The names corresponding to these circled numbers are also circled in red.

1	Albert D.	25	Monique Q.
2	Richard D.	26	Régine D.
3	Belle H.	27	Lucille L.
4	Raymond L.	28	Jérémy W.
5	Stéphane B.	29	Gilles D.
6	Albert T.	30	Renaud S.
7	Jean William V.	31	Pierre K.
8	André D.	32	Etienne M.
9	Jeremy W.	33	Marie M.
10	Anthony Q.	34	Gaétan Z.
11	James B.	35	Fidèle D.
12	Denis G.	36	Maria P.
13	Amanda L.	37	Anne-Marie G.
14	Jennifer L.	38	Michel K.
15	Philippe K.	39	Gaston C.
16	Eve F.	40	Alain M.
17	Priscilla O.	41	Olivier P.
18	Robert D.	42	Geneviève M.
19	Brian F.	43	Berthe D.
20	Hellène H.	44	Jean Pierre P.
21	Isabelle R.	45	Jacques B.
22	Jean T.	46	François P.
23	Samanta D.	47	Dominique M.
24	Berthe L.	48	Antoine C.

Stratified sampling

- Principle :
 - Divide sampling frame into homogeneous subgroups (strata) e.g. age-group, occupation;
 - Draw random sample in each strata.

Stratified sampling

- Advantages
 - Can acquire information about whole population and individual strata
 - Precision increased if variability within strata is less (homogenous) than between strata
- Disadvantages
 - Can be difficult to identify strata
 - Loss of precision if small numbers in individual strata
 - resolve by sampling proportionate to stratum population

Multiple stage sampling

Principle:

- consecutive sampling
- example :

sampling unit = household

- 1st stage: draw neighborhoods
- 2nd stage: draw buildings
- 3rd stage: draw households

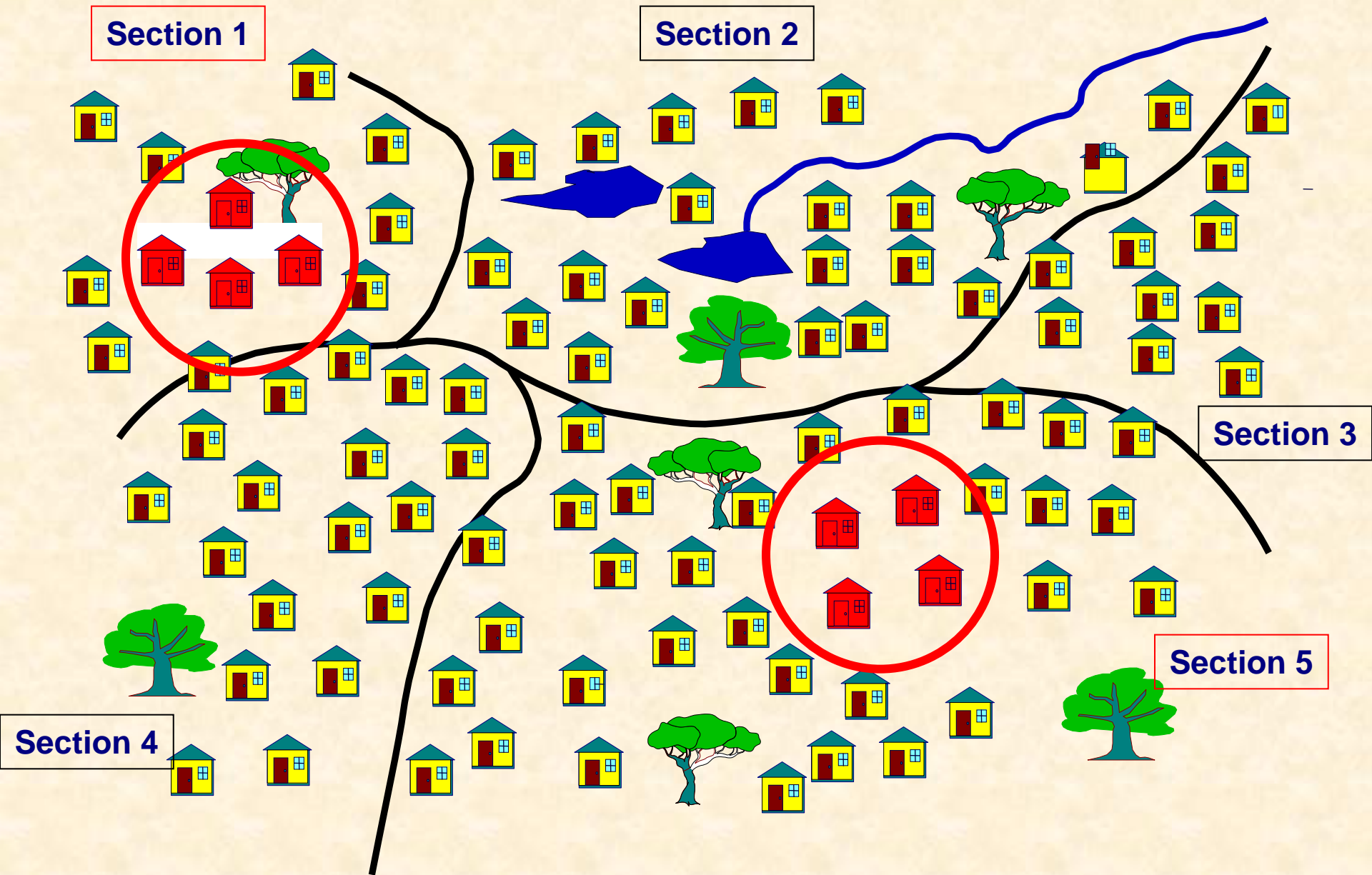
Cluster sampling

- Principle
 - Sample units not identified independently but in a group (or “cluster”)
 - Provides logistical advantage.

Cluster sampling

- Principle
 - Whole population divided into groups e.g. neighbourhoods
 - Random sample taken of these groups (“clusters”)
 - Within selected clusters, all units e.g. households included (or random sample of these units)

Example: Cluster sampling



Cluster sampling

- **Advantages**

- Simple as complete list of sampling units within population not required
- Less travel/resources required

- **Disadvantages**

- Potential problem is that cluster members are more likely to be alike, than those in another cluster (homogenous)....
- This “dependence” needs to be taken into account in the sample size....and the analysis (“design effect”)

Selecting a sampling method

- Population to be studied
 - Size/geographical distribution
 - Heterogeneity with respect to variable
- Availability of list of sampling units
- Level of precision required
- Resources available
 - Judgmental, convenience

Sample size estimation

- Estimate number needed to
 - reliably measure factor of interest
 - detect significant association
- Trade-off between study size and resources....
- Sample size determined by various factors:
 - significance level (“alpha”)
 - power (“1-beta”)
 - expected prevalence of factor of interest

Strengths and Weaknesses of Basic Sampling Techniques

Technique	Strengths	Weaknesses
<i>Nonprobability Sampling</i>		
Convenience sampling	Least expensive, least time-consuming, most convenient	Selection bias, sample not representative, not recommended for descriptive or causal research
Judgmental sampling	Low cost, convenient, not time-consuming	Does not allow generalization, subjective
Quota sampling	Sample can be controlled for certain characteristics	Selection bias, no assurance of representativeness
Snowball sampling	Can estimate rare characteristics	Time-consuming
<i>Probability sampling</i>		
Simple random sampling (SRS)	Easily understood, results projectable	Difficult to construct sampling frame, expensive, lower precision, no assurance of representativeness.
Systematic sampling	Can increase representativeness, easier to implement than SRS, sampling frame not necessary	Can decrease representativeness
Stratified sampling	Include all important subpopulations, precision	Difficult to select relevant stratification variables, not feasible to stratify on many variables, expensive
Cluster sampling	Easy to implement, cost effective	Imprecise, difficult to compute and interpret results

Type 1 error

- The probability of finding a difference with our sample compared to population, and there really isn't one....
- Known as the α (or “type 1 error”)
- Usually set at 5% (or 0.05)

Type 2 error

- The probability of not finding a difference that actually exists between our sample compared to the population...
- Known as the β (or “type 2 error”)
- Power is $(1 - \beta)$ and is usually 80%

Randomization

Randomization is the process of assigning clinical trial participants to treatment groups. Randomization gives each participant a known (usually equal) chance of being assigned to any of the groups. Successful randomization requires that group assignment cannot be predicted in advance.

Why Randomize?

- If, at the end of a clinical trial, a difference in outcomes occurs between two treatment groups (say, intervention and control) possible explanations for this difference would include:
 - the intervention exhibits a real effect;
 - the outcome difference is solely due to chance
 - there is a systematic difference (or bias) between the groups due to factors other than the intervention.

Randomization aims to obviate the third possibility.

Forms of Randomization

- Simple Randomization
- Permuted Block Randomization
- Stratified Block Randomization

Simple Randomization

- Coin Tossing for each trial participant
- Sequence of Random Numbers from statistical textbooks
- Computer generated sequence



Illustrations

The computer generated sequence:

4,8,3,2,7,2,6,6,3,4,2,1,6,2,0,.....

Two Groups (criterion:even-odd):

AABABAAABAABAAA.....

Three Groups:

(criterion:{1,2,3}~A, {4,5,6}~B, {7,8,9}~C; ignore 0's)

BCAACABBABAABA.....

Two Groups: different randomisation ratios(eg.,2:3):

(criterion:{0,1,2,3}~A, {4,5,6,7,8,9}~B)

BBAABABBABAABAA.....

Permuted Block Randomization

- Used for small studies to maintain reasonably good balance among groups
- In a two group design, Blocks having equal numbers of As and Bs (A = intervention and B = control, for example) are used, with the order of treatments within the block being randomly permuted

Illustration

With a block size of 4 for two groups(A,B), there are 6 possible permutations and they can be coded as:

1=AABB, 2=ABAB, 3=ABBA, 4=BAAB, 5=BABA, 6=BBAA

Each number in the random number sequence in turn selects the next block, determining the next four participant allocations (ignoring numbers 0,7,8 and 9).

e.g., The sequence 67126814.... will produce BBAA AABB ABAB BBAA AABB BAAB.

In practice, a block size of four is too small since researchers may crack the code and risk selection bias. Mixing block sizes of between 6 and 12 is better with the size kept unknown to the investigator. This precaution maintains concealment.

Stratified Block Randomization

Stratified block randomization can further restrict chance imbalances to ensure the treatment groups are as alike as possible for selected prognostic variables or other patient factors. A set of permuted blocks is generated for each combination of prognostic factors

Typical examples of such factors are age group, severity of condition, and treatment centre. Stratification simply means having separate block randomization schemes for each combination of characteristics ('stratum')

For example, in a study where you expect treatment effect to differ with age and sex you may have four strata: male over 65, male under 65, female over 65 and female under 65

Inappropriate randomisation methods

- Assigning patients alternately to treatment group is not random assignment
- Assigning the first half of the population to one group is not random assignment
- Assignments by methods based on patient characteristics such as date of birth, order of entry into the clinic or day of clinic attendance, are not reliably random

Issues leading to Blinding

- Most investigators have firm views about which of a range of alternative treatments is more effective and often, which is more appropriate for particular groups of patients. As a result, there is a strong temptation by investigators to channel particular groups of patients to particular treatments (channeling effect)
- There is also a risk of the investigators subconsciously losing their objectivity in their assessments of treatment effects simply because of their clear preference for particular treatments
- There is a risk of having other forms of bias, which can be satisfactorily controlled by proper blinding

Blinding

All of these **potential** problems can be avoided if **everyone** involved in the study is blinded to the actual treatment the patient is receiving.

Blinding (also called **masking or concealment of treatment**) is intended to avoid bias caused by subjective judgment in reporting, evaluation, data processing, and analysis due to knowledge of treatment.

Hierarchy of Blinding

- **open label:** no blinding
- **single blind:** patient (usually; occasionally may be assessor) blinded to treatment
- **double blind:** patient and assessors (who often are also the health care providers and data collectors) blinded to treatment
- **complete blind:** everyone involved in the study blinded to treatment

Thank You