Biostatistics

Luonan Chen

Chinese Academy of Sciences

Lecture 1

Introduction To Biostatistics

- ***Key words :***
  - *Statistics(统计) , data（数据）, Biostatistics（生物统计）,*
  - *Variable （变量）,Population（总体）,Sample（样本）*

- 总体：研究对象的全体，是具有相同性质的个体所组成的集合
- 个体：组成总体的基本单元
- 样本：从总体中抽出若干个个体的集合称为样本。
- 变量：相同性质的事物间表现差异性或差异特征的数据称为变量。
- 参数：参数也称参量，是对一个总体特征的度量。
- 准确性：是指统计数接近真知的程度。
- 精确性：指调查或试验中同一试验指标或性状的重复观测值彼此接近程度的大小。
- 误差：是试验中不可控因素所引起的观测值偏离真值的差异
- 错误：是指在试验过程中，人为因素所引起的差错。
- 统计数：从样本计算所得的数值称为统计数，它是总体参数的估计值。
- 生物统计学：是统计学在生物学中的应用，是用数理统计的原理和方法来分析解释生命现象的一门科学，是研究生命过程中以样本推断总体的一门科学。

Statistics

- The field of statistics: The study and use of theory and methods for the analysis of data arising from random processes or phenomena. The study of how we make sense of data.

_ forming hypotheses
_ designing experiments and observational studies
_ gathering data
_ summarizing data
_ drawing inferences from data, e.g. testing hypotheses

Biostatistics

- Biostatistics is the branch of applied statistics directed toward applications in the health sciences and biology
- Biostatistics is sometimes distinguished from the field of biometry based upon whether applications are in the health sciences (biostatistics) or in broader biology (biometry, e.g., agriculture, ecology, wildlife biology).

生物大数据挑战

- **体系：**为生物系统信息过程的研究，揭示生物数据之间联系，开发数据库和工具，整合、获取、使用、分析和解释数据

- **目标：** 数据 → 信息 → 知识
- 统计学： **数据 → 相关性 → 因果性**

发现新的生物学知识，阐释复杂生命过程的机制和规律，为应用提供依据和方法

生物大数据特征

小样本 （大样本）	生物医学领域，虽然群体大样本，但个体(如病人) 多为小样本
高维 （低维）	组学（或影像）数据包含大量特征信息，反映生物体中成千上万个分子等的状态
网络 （分子）	表征系统性的而非碎片化的关系
动态 （静态）	表征生命演化的本质特征之一
多样性 （共性）	生物学的进化本质特征之一。生物信息不仅含有群体的共性信息，更重要是含有其个性信息
因果 （相关）	不同于传统大数据研究的关联性，生物学需要揭示各种分子和现象的因果性
整合 （分解）	不同层次的异质信息高度融合才能系统地、定量地揭示复杂生物现象

生物大数据研究的挑战：生物数据→生物信息→生物知识

数据 → 相关性 → 因果性

1

传统大数据 →大样本低维数据

↓

传统 统计学

生物医学大数据 → 小样本高维数据

↓

非传统 统计学

大样本→相关性

↓

大数据理论和方法

统计学

相关性→因果性

↓

非传统大数据理论

Statistics → Basis

- Bioinformatics
- Big-data science
- Systems biology
- Computational biology
- Network biology
- Network medicine

Another aspect: Dynamics

Challenge

- Much of life is composed of a systematic component (i.e., signal) and a random component (i.e., error or noise)

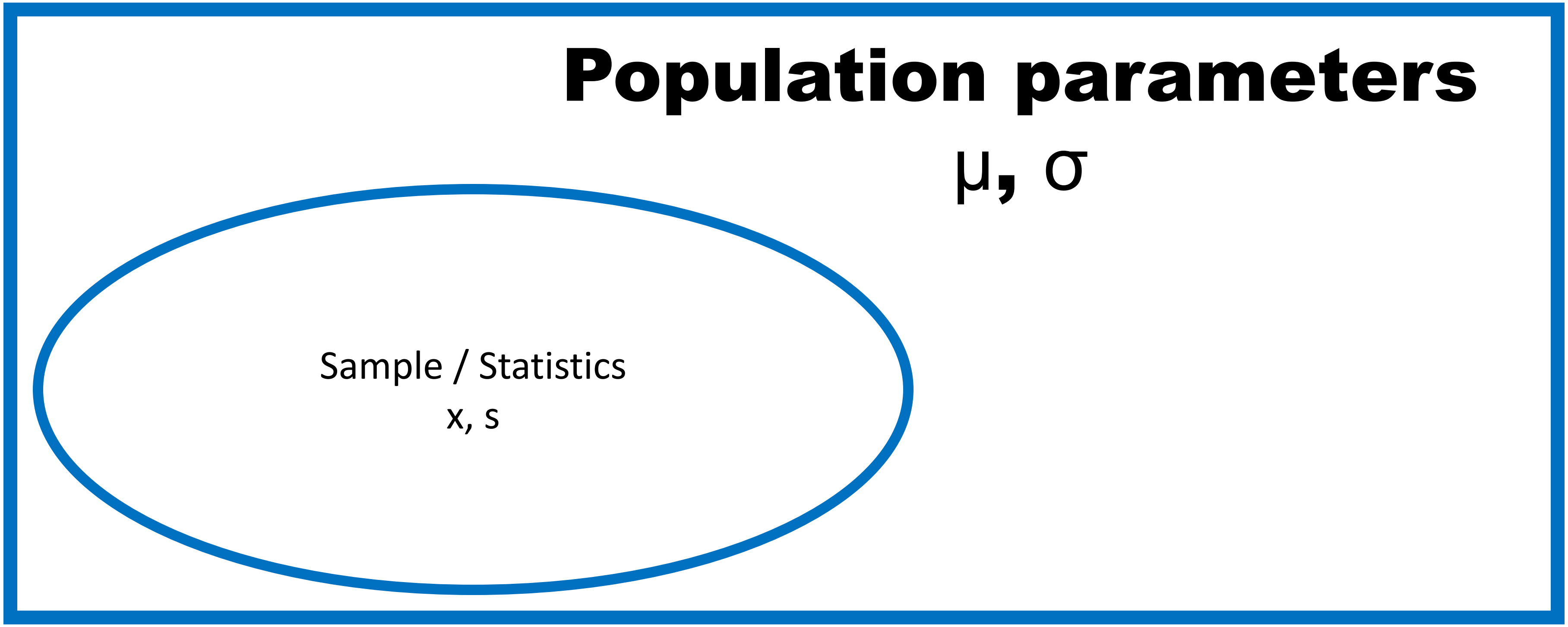
  Real data = Deterministic + Random

- Example:
  - Smoking is associated with lung cancer.
  - Yet not everyone that smokes, gets lung cancer, and not everyone that gets lung cancer, smokes
  - Yet we know that there is an association (a systematic component)
  - Longevity data in Shanghai, Tokyo and HK
  - Lung cancer ratio in Denmark > China due to air pollution ?

- Our challenge
  - Identify the systematic component (separate it from the random component), estimate it, and perhaps make inferences with it

Big Picture

Populations and Samples



Population parameters

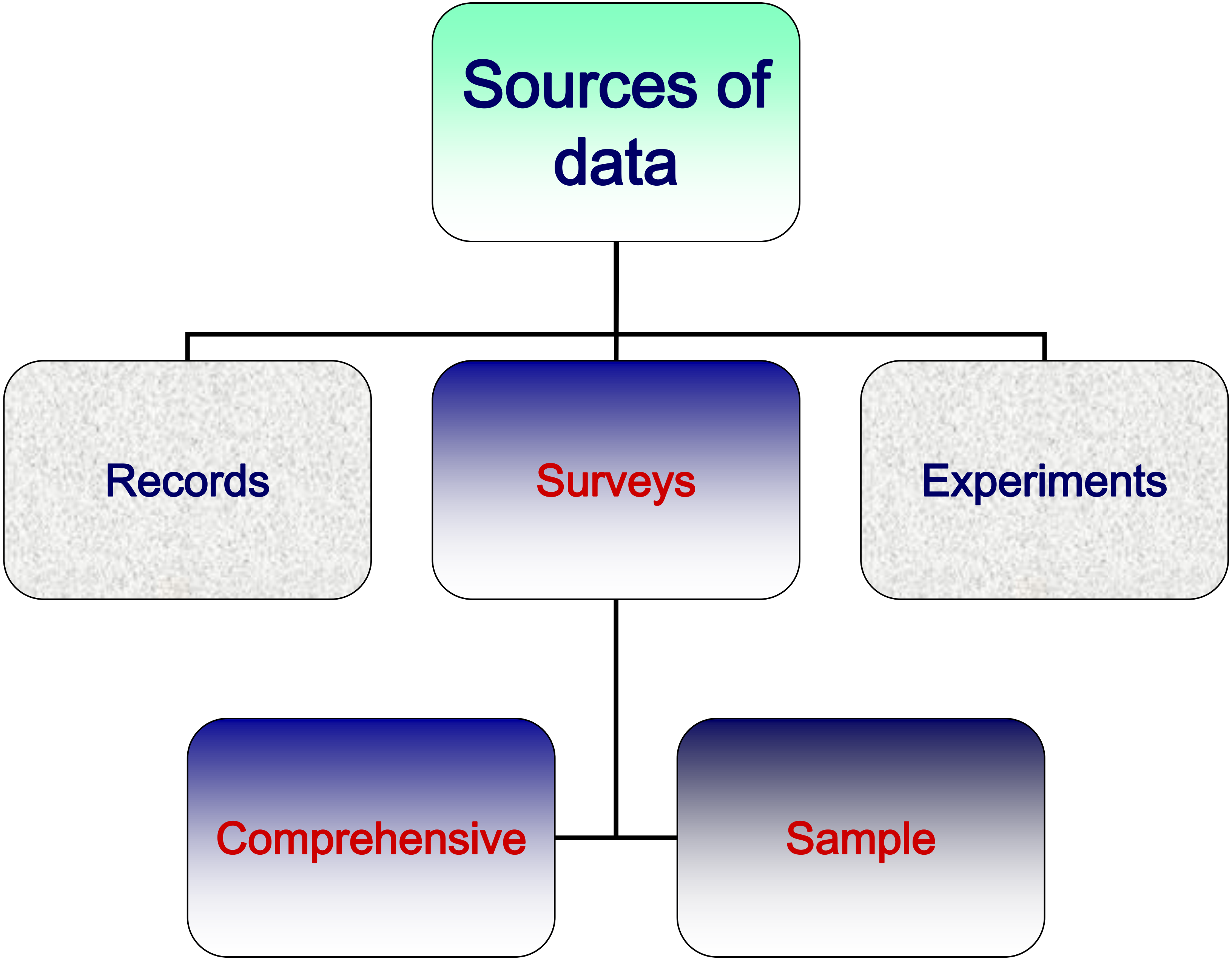
μ, σ

Sample / Statistics

x, s

Data:

- **The raw material of Statistics is data.**
- **We may define data as figures. Figures result from the process of counting or from taking a measurement.**
- ***For example:***
- **- When a hospital administrator counts the number of patients (counting).**
- **- When a nurse weighs a patient (measurement)**
- **- Gene sequences**
- **- Gene expression**
- **- Protein expression**
- **- metabolic expression**
- **- molecular interaction**



Sources of data

Records

Surveys

Experiments

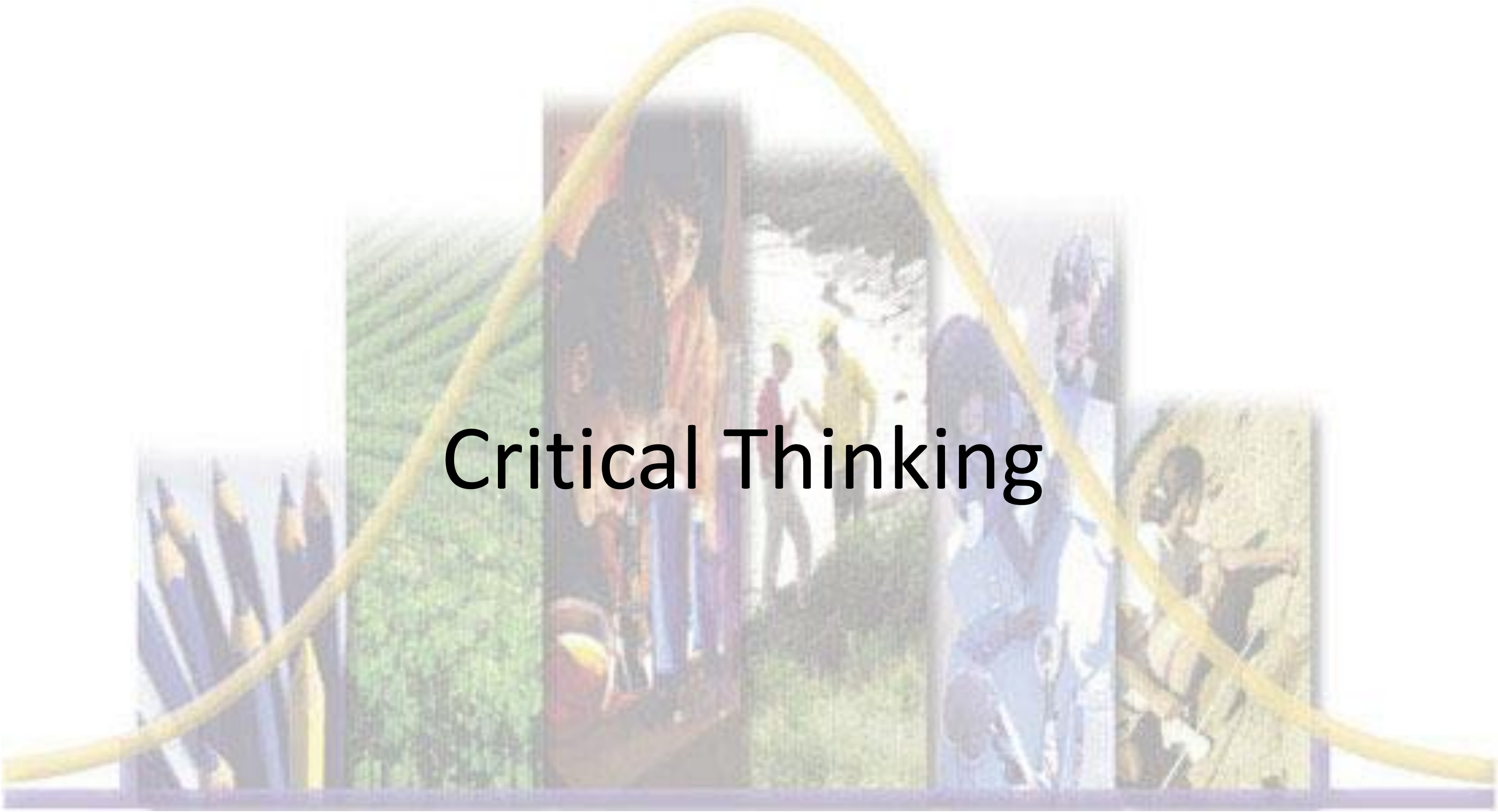
Comprehensive

Sample

Data

- Pieces of information : Information resolves uncertainty

- Scales of Measurement
  - Nominal – unordered categories
  - Ordinal – ordered categories
  - Discrete – only whole numbers are possible, order and magnitude matters
  - Continuous – any value is conceivable



Critical Thinking

Success in Statistics

- Success in the introductory statistics typically requires more common sense than mathematical expertise.

How common sense is used when we think critically about data and statistics ?

Misuses of Statistics

- Bad Samples
- Small Samples
- Misleading Graphs
- Pictographs
- Distorted Percentages
- Loaded Questions
- Order of Questions
- Refusals
- Correlation & Causality
- Self Interest Study
- Precise Numbers
- Partial Pictures
- Deliberate Distortions

To correctly interpret a graph, we should analyze the numerical information given in the graph instead of being mislead by its general shape.

* A variable:

It is a characteristic that takes on different values in different persons, places, or things.

For example:

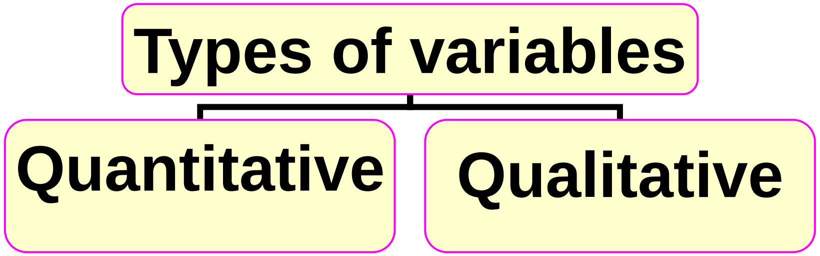
- heart rate,

- the heights of adult males,

- the weights of preschool children,

- the ages of patients seen in a dental clinic,

- the concentration of a protein.



Quantitative Variables

It can be measured in the usual sense.

For example:

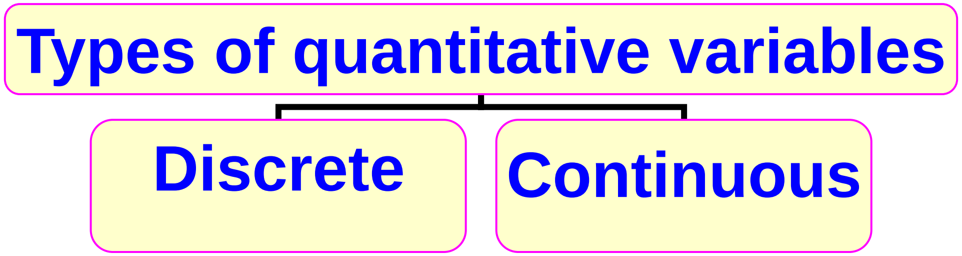
- the heights of adult males,
- the weights of preschool children,
- the ages of patients seen in a
- dental clinic.

Qualitative Variables

Many characteristics are not capable of being measured. Some of them can be ordered or ranked.

For example:

- classification of people into socio-economic groups,
- social classes based on income, education, etc.



Types of quantitative variables

Discrete

Continuous

A discrete variable

is characterized by gaps or interruptions in the values that it can assume.

For example:

- The number of daily admissions to a general hospital,
- The number of decayed, missing or filled teeth per child in an elementary school.

A continuous variable

can assume any value within a specified relevant interval of values assumed by the variable.

For example:

- Height,
- weight,
- skull circumference.

No matter how close together the observed heights of two people, we can find another person whose height falls somewhere in between.

* A population:

It is the largest collection of values of a random variable for which we have an interest at a particular time.

For example:

The weights of all the children enrolled in a certain elementary school.

Populations may be finite or infinite.

* A sample:

It is a part of a population.

For example:

The weights of only a fraction of these children.

Key Concepts

- Sample data must be collected in an appropriate way, such as through a process of random selection.

Major Points

- If sample data are not collected in an appropriate way, the data may be so completely useless that no amount of statistical tutoring can salvage them.

- Randomness typically plays a critical role in determining which data to collect.

Definitions

- Cross Sectional Study

  Data are observed, measured, and collected at one point in time.

- Retrospective (or Case Control) Study

  Data are collected from the past by going back in time.

- Prospective (or Longitudinal or Cohort) Study

  Data are collected in the future from groups (called cohorts) sharing common factors.

Definitions

- Confounding

  occurs in an experiment when the experimenter is not able to distinguish between the effects of different factors

Try to plan the experiment so confounding does not occur!

Controlling Effects of Variables

- Blinding

  subject does not know he or she is receiving a treatment or placebo （安慰药）

- Blocks

  groups of subjects with similar characteristics

- Completely Randomized Experimental Design

  subjects are put into blocks through a process of random selection

- Rigorously Controlled Design

  subjects are very carefully chosen

Replication and Sample Size

Replication

repetition of an experiment when there are enough subjects to recognize the differences in different treatments

Sample Size

use a sample size that is large enough to see the true nature of any effects and obtain that sample using an appropriate method, such as one based on randomness

Definitions

Random Sample

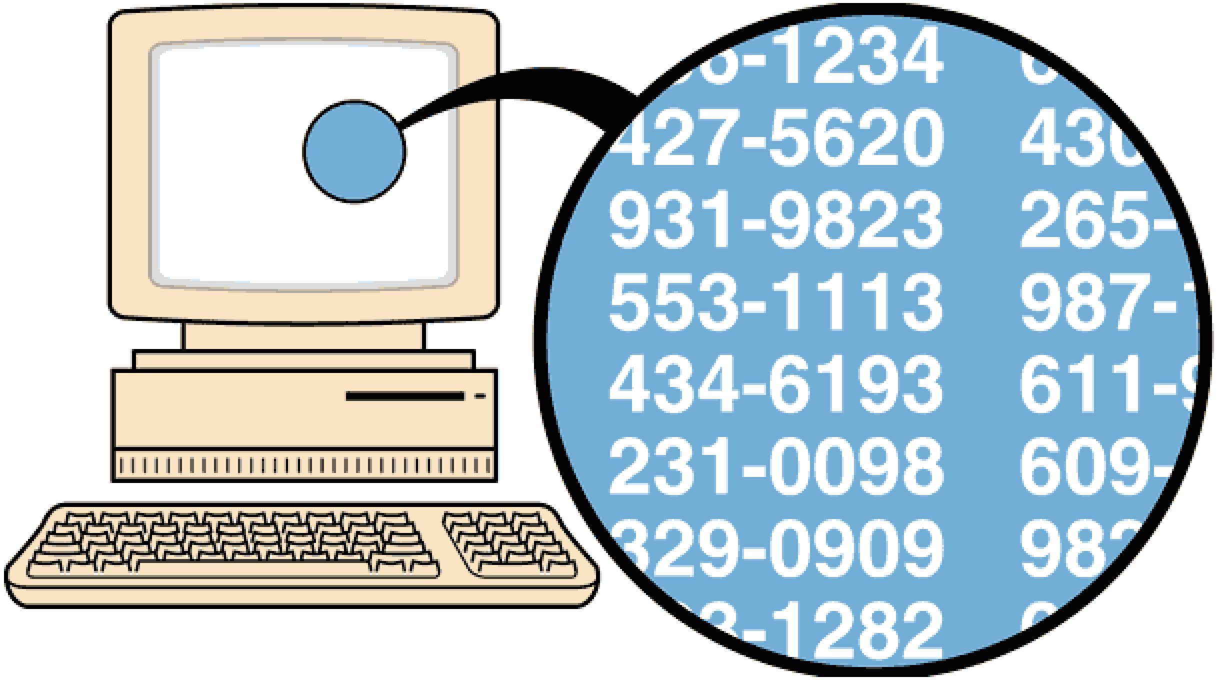
members of the population are selected in such a way that each individual member has an equal chance of being selected

Simple Random Sample (of size n)

subjects selected in such a way that every possible sample of the same size n has the same chance of being chosen

Random Sampling

selection so that each has an equal chance of being selected

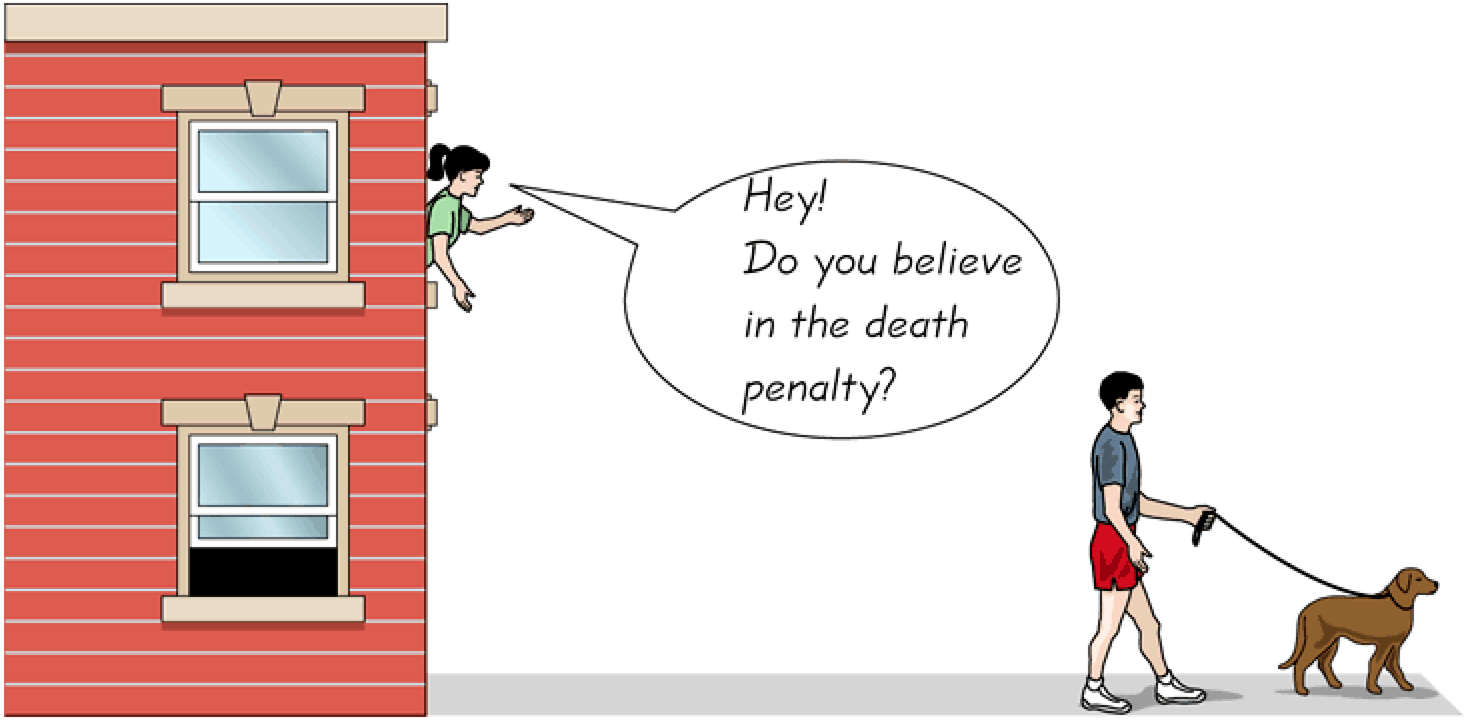


Systematic Sampling

Select some starting point and then select every K th element in the population

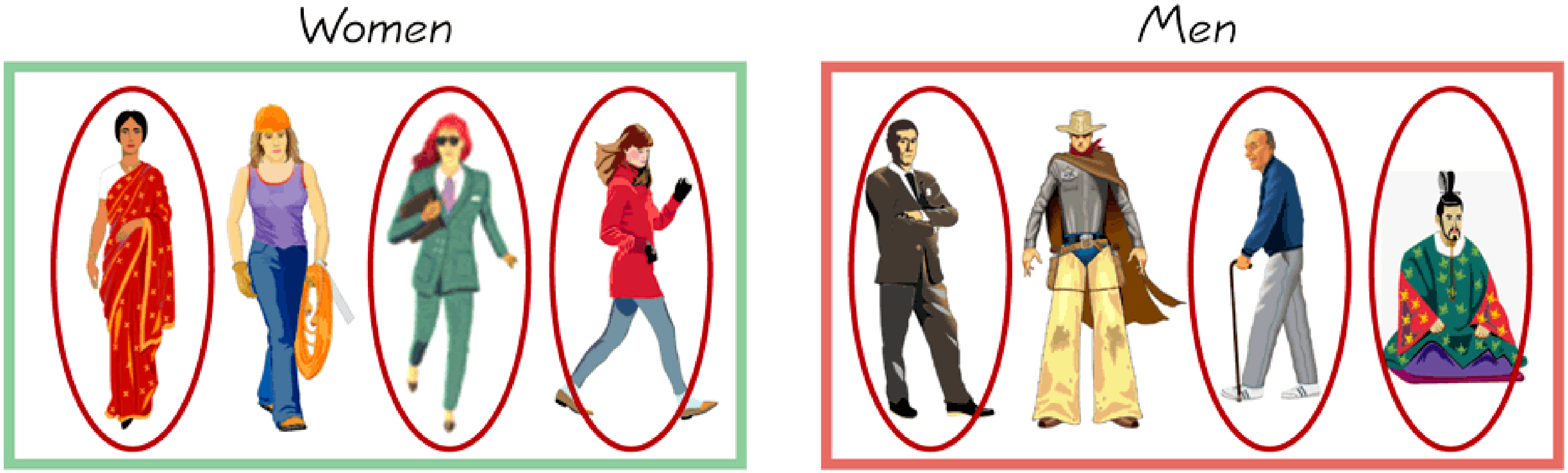
Convenience Sampling

use results that are easy to get



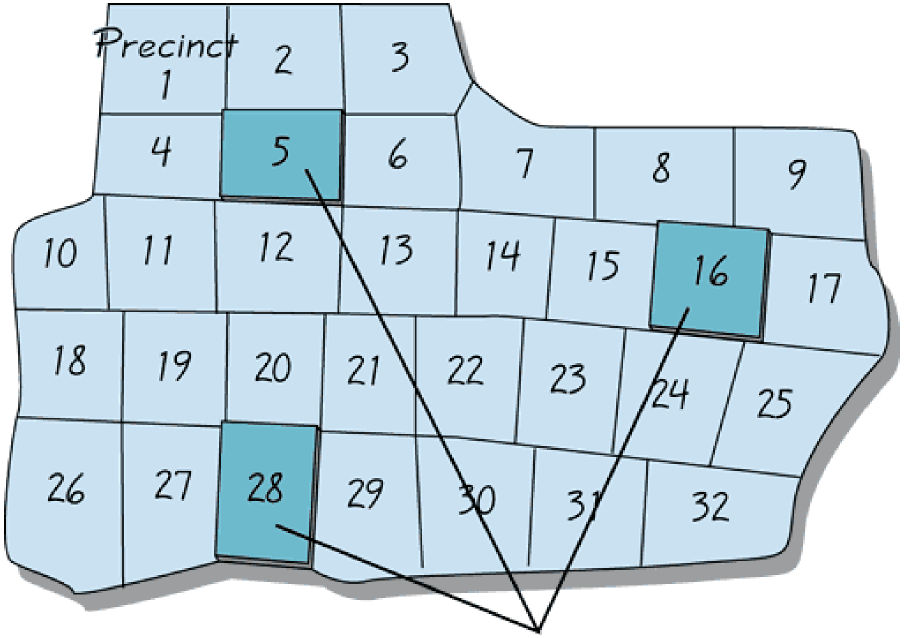
Stratified Sampling

subdivide the population into at least two different subgroups that share the same characteristics, then draw a sample from each subgroup (or stratum)



Cluster Sampling

divide the population into sections (or clusters); randomly select some of those clusters; choose all members from selected clusters



Interview all voters in shaded precincts.

Methods of Sampling

- Random
- Systematic
- Convenience
- Stratified
- Cluster

Definitions

- Sampling Error

  the difference between a sample result and the true population result; such an error results from chance sample fluctuations

- Nonsampling Error

  sample data that are incorrectly collected, recorded, or analyzed (such as by selecting a biased sample, using a defective instrument, or copying the data incorrectly)