1. INTRODUCTION TO STATISTICS

Introduction, basic terms, types of variables, etc

Course introduction

- Read proforma / course description.

Introduction

What is statistics?

The word statistics has two meanings

- Refer to numerical facts.
  - For example, number of students, age, income, etc
  - The total number of registered students for STQS1913 is 150.

- Refer to the field or discipline of study
  - Statistics is the science of collecting, analyzing, presenting, and interpreting data, as well as of making decisions based on such analyses.

What is statistics?

- We make decisions every day, and it comes with uncertainty.
- There may be no definite solution in a given problem.
- So how should we make decision?
  - Collect and use available information (data).
  - Perform (statistical) analyses.
  - Make an educated guess, decision or conclusion.

Statistics covers all of the above.

Types of statistics

1. Descriptive statistics – consists of methods for organizing, displaying, and describing data by using tables, graphs, and summary measures.
   - Statistics is used to describe data.

2. Inferential statistics – consists of methods that use sample results to help make decisions or predictions about a population.
   - Statistics is used to make decisions or conclusions.

Why study statistics?

A few reasons

- Data is everywhere.
- To understand presented statistics and graph from media.
- To make better and objective decisions.
- In science fields, to be able to effectively conduct research.

Basic terms in statistics

- Element or member of a sample or population – a specific subject or object about which the information is collected.
  - For example, a person, firm, item, state, or country

- Variable – a characteristic under study that assumes different values for different elements.
  - In contrast to a variable, the value of a constant is fixed.

Basic terms in statistics

- Observation or measurement – the value of a variable for an element.

- Data set – a collection of observations on one or more variables.

Basic terms in statistics

Table 1.1 Total Revenues for 2010 of Six Companies

Company	2010 Total Revenue (millions of dollars)
Wal-Mart Stores	421,849
Royal Dutch Shell	378,152
Exxon Mobil	354,674
BP	308,928
Sinopec Group	273,422
China National Petroleum	240,192

Source: Fortune Magazine, July 25, 2011.

2010 Total Revenue ⟵ Variable

An element or a member ⟶ Exxon Mobil

354,674 ⟵ An observation or measurement

Exercise

1.9 The following table gives the number of dog bites reported to the police last year in six cities.

City	Number of Bites
Center City	47
Elm Grove	32
Franklin	51
Bay City	44
Oakdale	12
Sand Point	3

element { 1 2 3 4 5 6 } obs

a. What is the variable for this data set? — number of Bites

b. How many observations are in this data set? → 6

c. How many elements does this data set contain? 1 6

Exercise

1.10 The following table gives the state taxes (in dollars) on a pack of cigarettes for nine states as of April 1, 2009.

State	State Tax (in dollars)	
Alaska	2.00	1
Iowa	1.36	2
Massachusetts	2.51	3
Missouri	.17	4
New Hampshire	1.33	5
New York	2.75	6
Ohio	1.25	7
South Carolina	.07	8
West Virginia	.55	9

a. What is the variable for this data set? — state tax

b. How many observations are in this data set? — 9

c. How many elements does this data set contain? — 1

9

Types of variables

Types of variables

- Quantitative variables – can be measured numerically.
  - Two types:
    - Discrete variables – countable outcome
      - Eg: number of students, number of days
    - Continuous variables – uncountable outcome
      - Eg: height of a person, weight, temperature

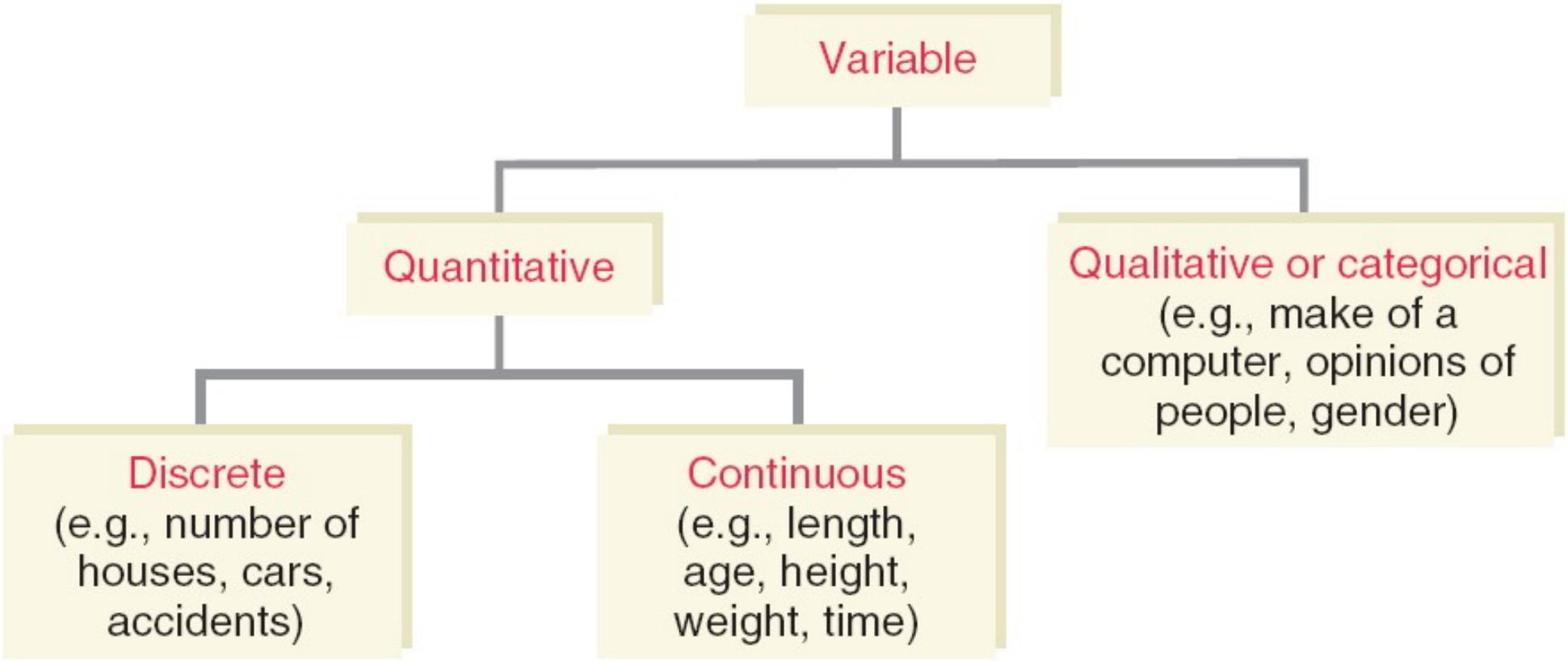
selanjar

- Qualitative or categorical variables – cannot be written in numerical values
  - Eg: blood type, exam grade, gender

→ no mean

category: 1, 2, 3 → may appear as numbers.

Types of variables



Cross-section vs time-series data

We may come across these types of data:

- Cross-section data
  - Data collected on different elements at the same point in time or for the same period of time

- Time-series data
  - Data collected on the same element for the same variable at different points in time or for different periods of time

(element vs time-period.)

Cross-section vs time-series data

Shop	Sales (RM)				
	2010	2011	2012	2013	2014
A	900	1000	1100	1050	1300
B	550	570	595	600	650
C	600	500	780	900	980
D	400	430	460	490	520
E	880	870	890	900	895

time series data

cross-sectional data

Exercise

1.14 Indicate which of the following variables are quantitative and which are qualitative.

a. Number of persons in a family Quanti - Discrete

b. Colors of cars Quali

c. Marital status of people Quali

d. Time to commute from home to work Quanti - Continous

e. Number of errors in a person's credit report Quanti - Discrete

1.16 Classify the quantitative variables in Exercise 1.14 as discrete or continuous.

Exercise

UNIVERSITI KEBANGSAAN MALAYSIA
The National University of Malaysia

1.15 Indicate which of the following variables are quantitative and which are qualitative.

a. Number of typographical errors in newspapers Quanti, discrete

b. Monthly TV cable bills Quanti, Continous

c. Spring break locations favored by college students Quali

d. Number of cars owned by families Quanti, discrete

e. Lottery revenues of states Quanti, continous

1.17 Classify the quantitative variables in Exercise 1.15 as discrete or continuous.

Exercise

1.21 Classify the following as cross-section or time-series data.

a. Average prices of houses in 100 cities cross-sectional

b. Salaries of 50 employees cross-sectional

c. Number of cars sold each year by General Motors from 1980 to 2009 time-series

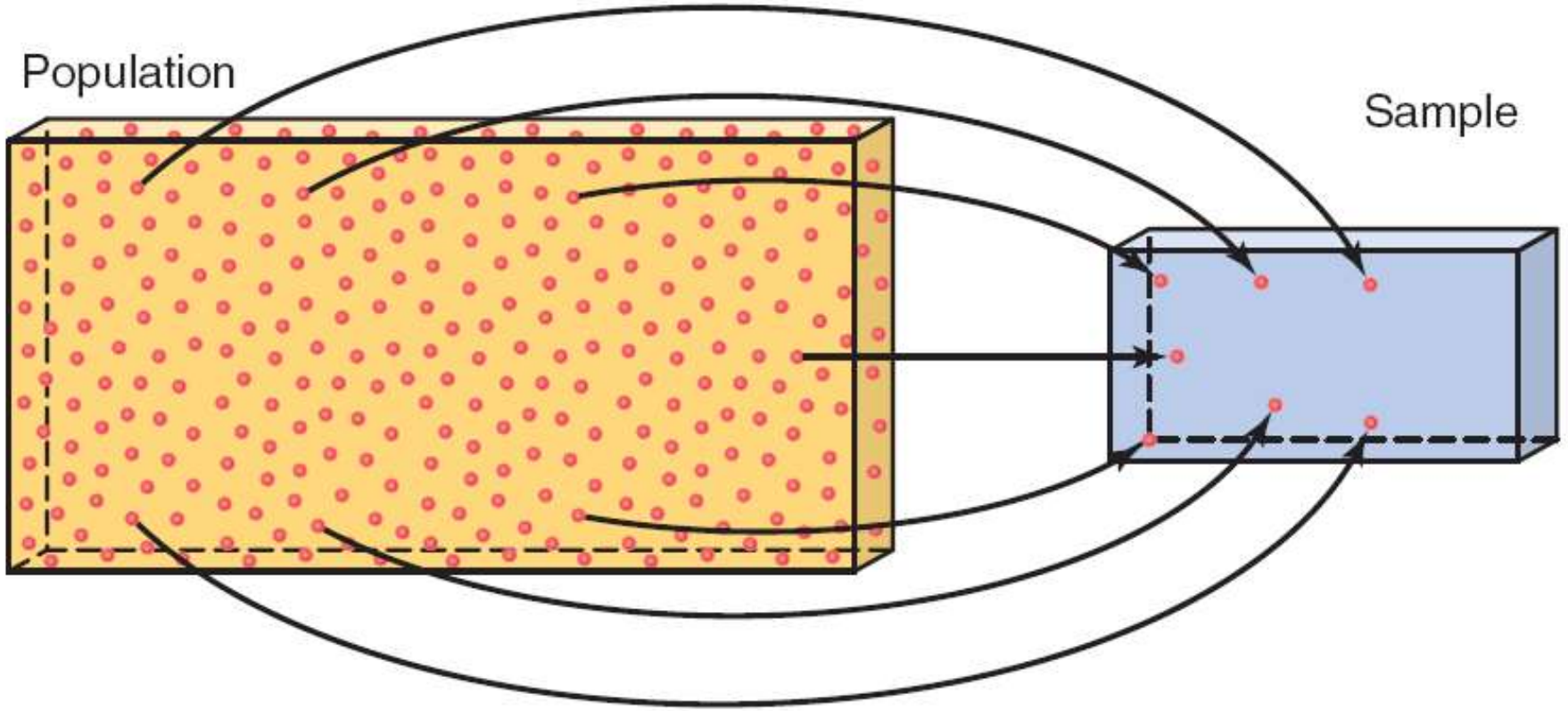
d. Number of employees employed by a company each year from 1985 to 2009 time-series

Population vs sample

Population vs sample

- Population – consists of all elements (individuals, items, or objects) whose characteristics are being studied.
- Sample – a portion of the population selected for study
- Census – a survey that includes every member of the population
- Sample survey – the technique of collecting information from a portion of the population
- Representative sample – sample that represents the characteristics of the population as close as possible

Population vs sample



Why sample?

- Time
  - Size of population can be quite large
  - Census can take time to complete
- Cost
  - Limited budget
- Impossibility of conducting a census
  - May not be possible to identify and access each member of the population

Exercise

1.7 Explain whether each of the following constitutes a population or a sample.

a. Number of personal fouls committed by all NBA players during the 2008–2009 season

b. Yield of potatoes per acre for 10 pieces of land

c. Weekly salaries of all employees of a company

d. Cattle owned by 100 farmers in Iowa

e. Number of computers sold during the past week at all computer stores in Los Angeles

population

sample

population

example

population

Random sampling techniques

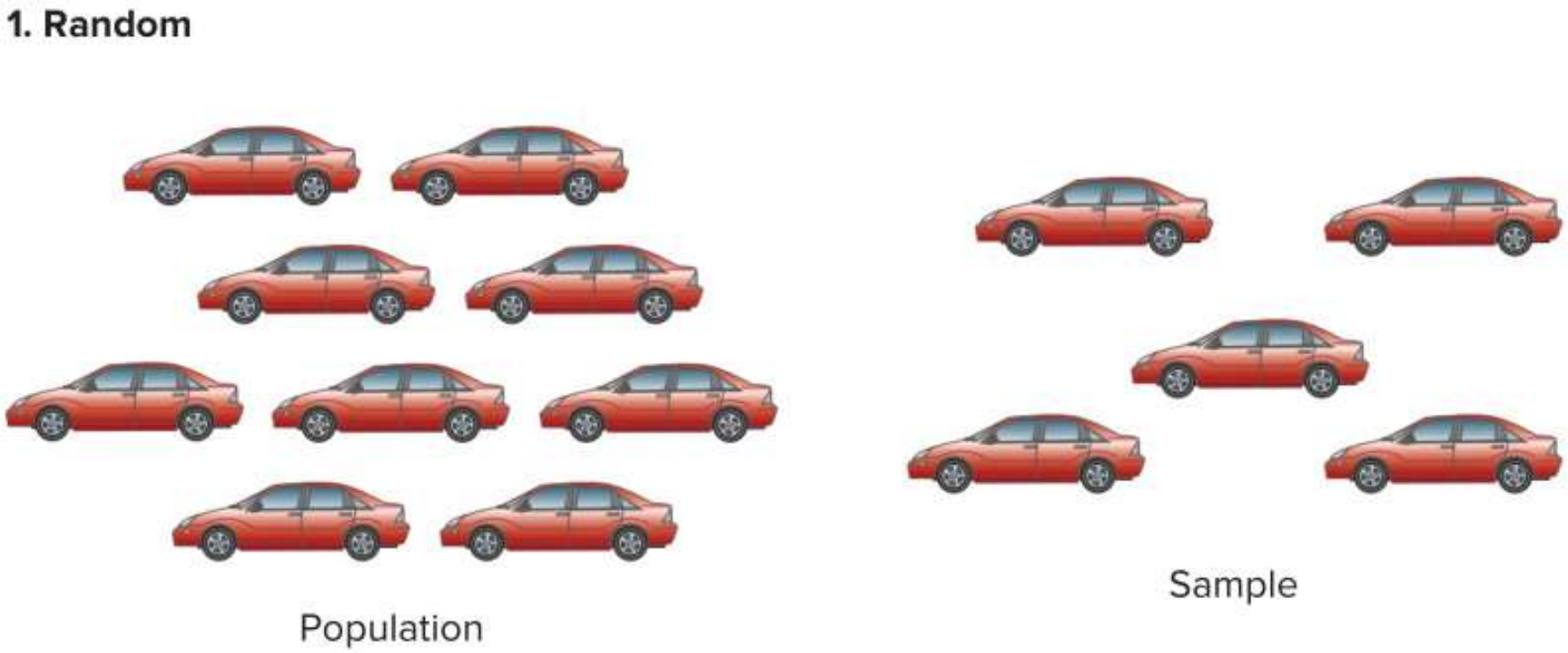
Why sample at random?

- Samples are to be collected to represent the populations.
  - Why sample at random? To avoid selection bias.

- Some of the sampling techniques:
  - Simple random sampling
  - Systematic random sampling
  - Stratified random sampling
  - Cluster sampling

Simple random sampling

- Basically, select sample at random in which each sample has the same probability of being selected.



Systematic random sampling

- Set the value of k, where k is the number obtained by dividing the population size by the intended sample size.
- Randomly select one member from the first k units.
- Then, select every kth member starting with the first selected member.

$$k = \frac{N}{n}$$

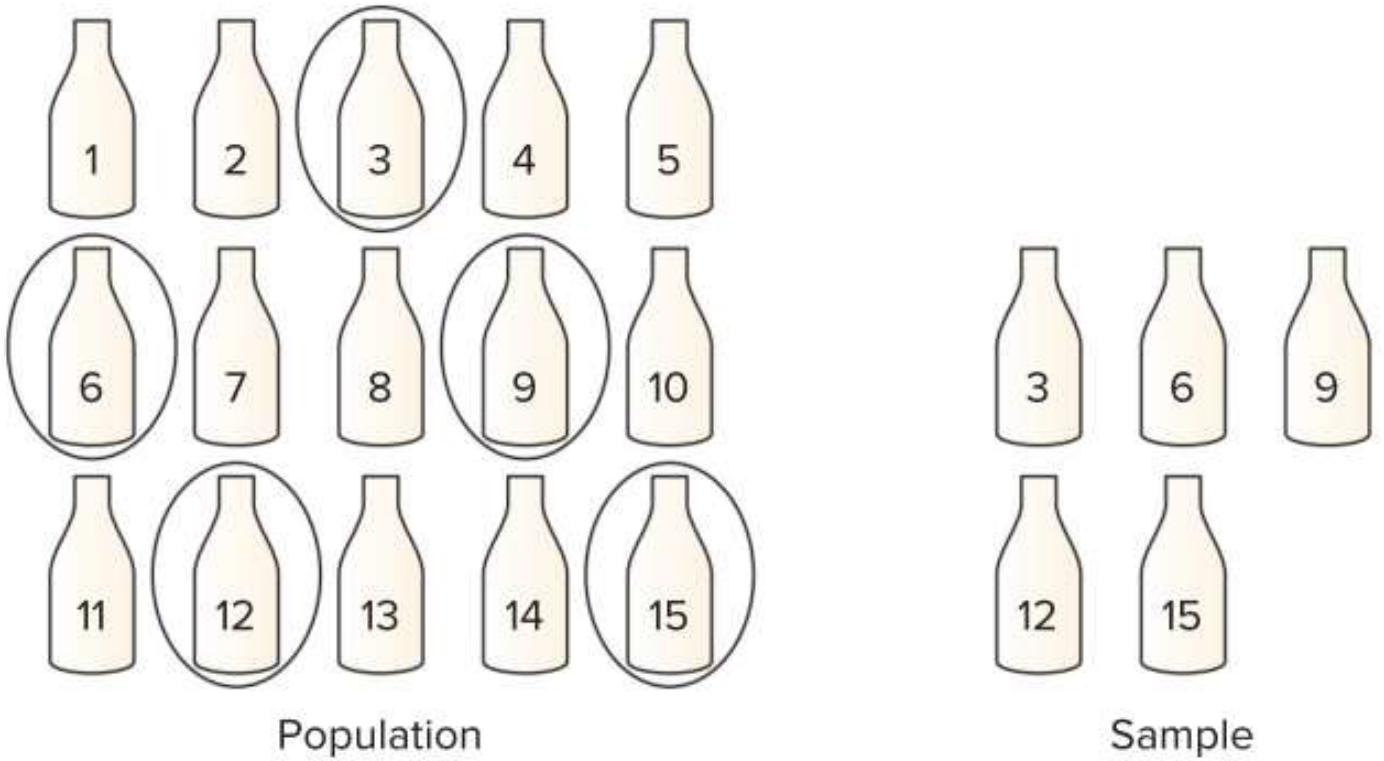
Systematic random sampling

$k = \frac{N}{n}$

$= \frac{15}{5}$

$= 3$ #

2. Systematic

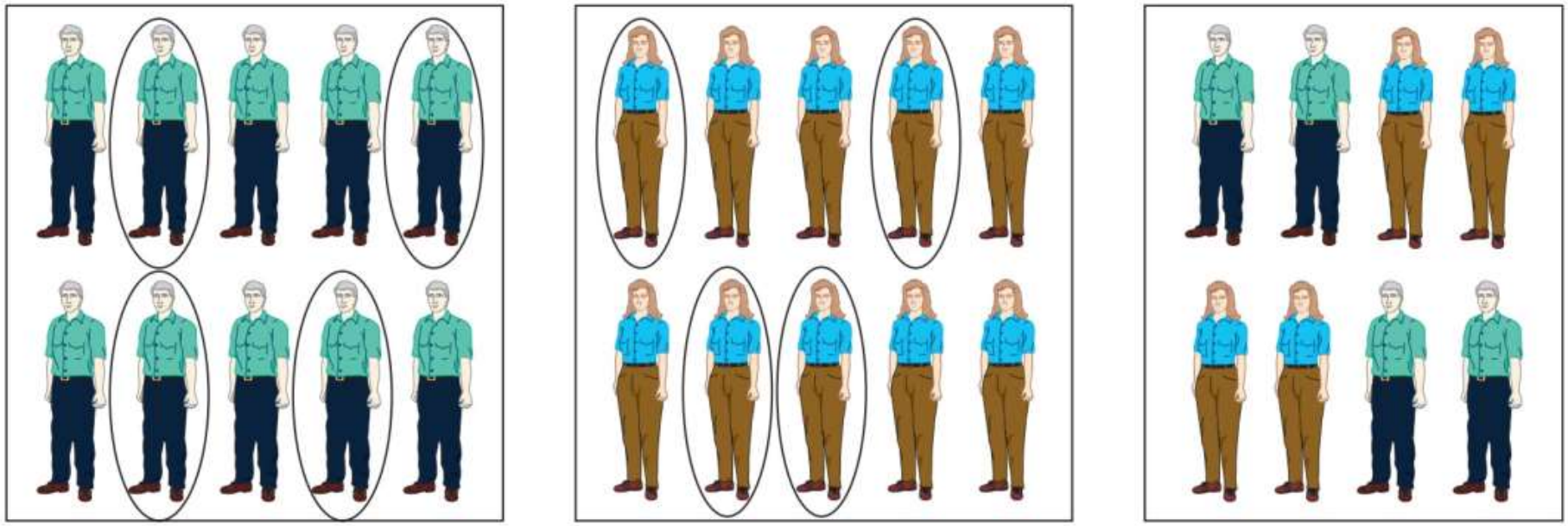


Stratified random sampling

- Divide the population into subpopulations, which are called strata.
- Then, select one sample from each of these strata.
- The collection of all samples from all strata gives the stratified random sample.

Stratified random sampling

3. Stratified

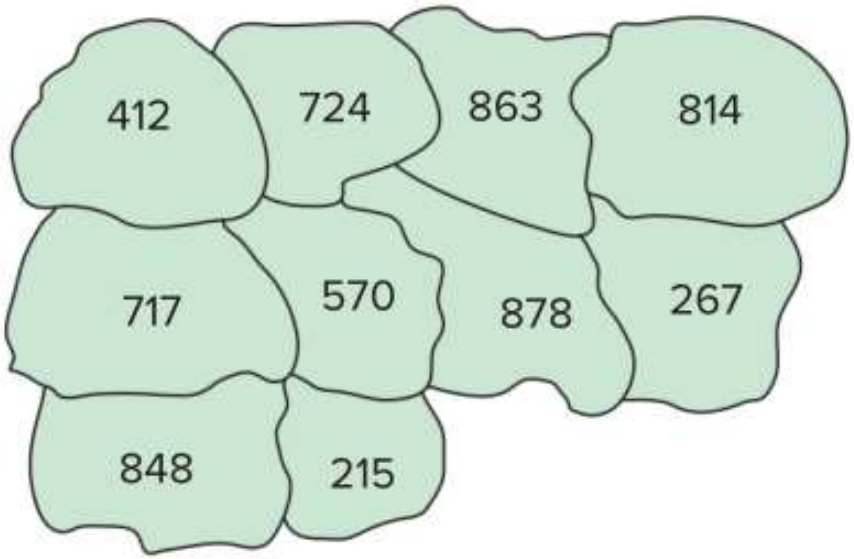


Cluster sampling

- The whole population is first divided into (geographical) groups called clusters.
- Each cluster is representative of the population.
- Then select a random sample of clusters.
- The elements of the selected clusters is our sample.

Cluster sampling

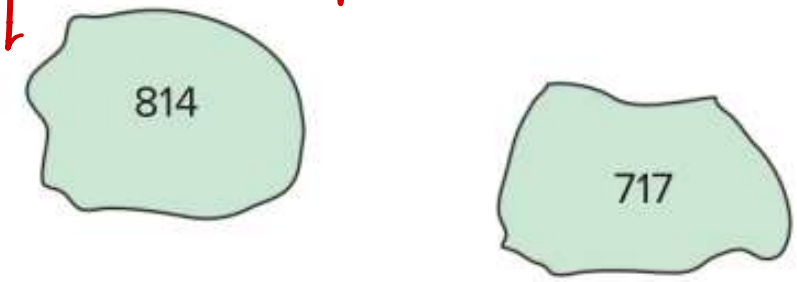
4. Cluster



Area codes

Population

take as a whole for area 717



Sample

Exercise

EXAMPLE 1–5 Sampling Methods

State which sampling method was used.

a. Out of 10 hospitals in a municipality, a researcher selects one and collects records for a 24-hour period on the types of emergencies that were treated there.

b. A researcher divides a group of students according to gender, major field, and low, average, and high grade point average. Then she randomly selects six students from each group to answer questions in a survey.

c. The subscribers to a magazine are numbered. Then a sample of these people is selected using random numbers.

d. Every 10th bottle of Energized Soda is selected, and the amount of liquid in the bottle is measured. The purpose is to see if the machines that fill the bottles are working properly.

Summary

- What is statistics?
- Basic terms in statistics – variables, element/member, observations, data set.
- Types of variables – qualitative and quantitative (discrete/continuous)
- Cross-section vs time-series data.
- Population vs sample.
- Sampling techniques – simple random, systematic random, stratified, cluster.