Statistics

Dr.Mamatha.H.R

Professor

Department of Computer Science and Engineering

PES University

Bangalore

Course material created using various Internet resources and text book

Why we Need to Know about Statistics

- To know how to properly present information
- To know how to draw conclusions about populations based on sample information
- To know how to improve processes
- To know how to obtain reliable forecasts

What Is Statistics?

1. **Collecting Data**
   e.g., Survey
2. **Presenting Data**
   e.g., Charts & Tables
3. **Characterizing Data**
   e.g., Average

Data Analysis

Why?

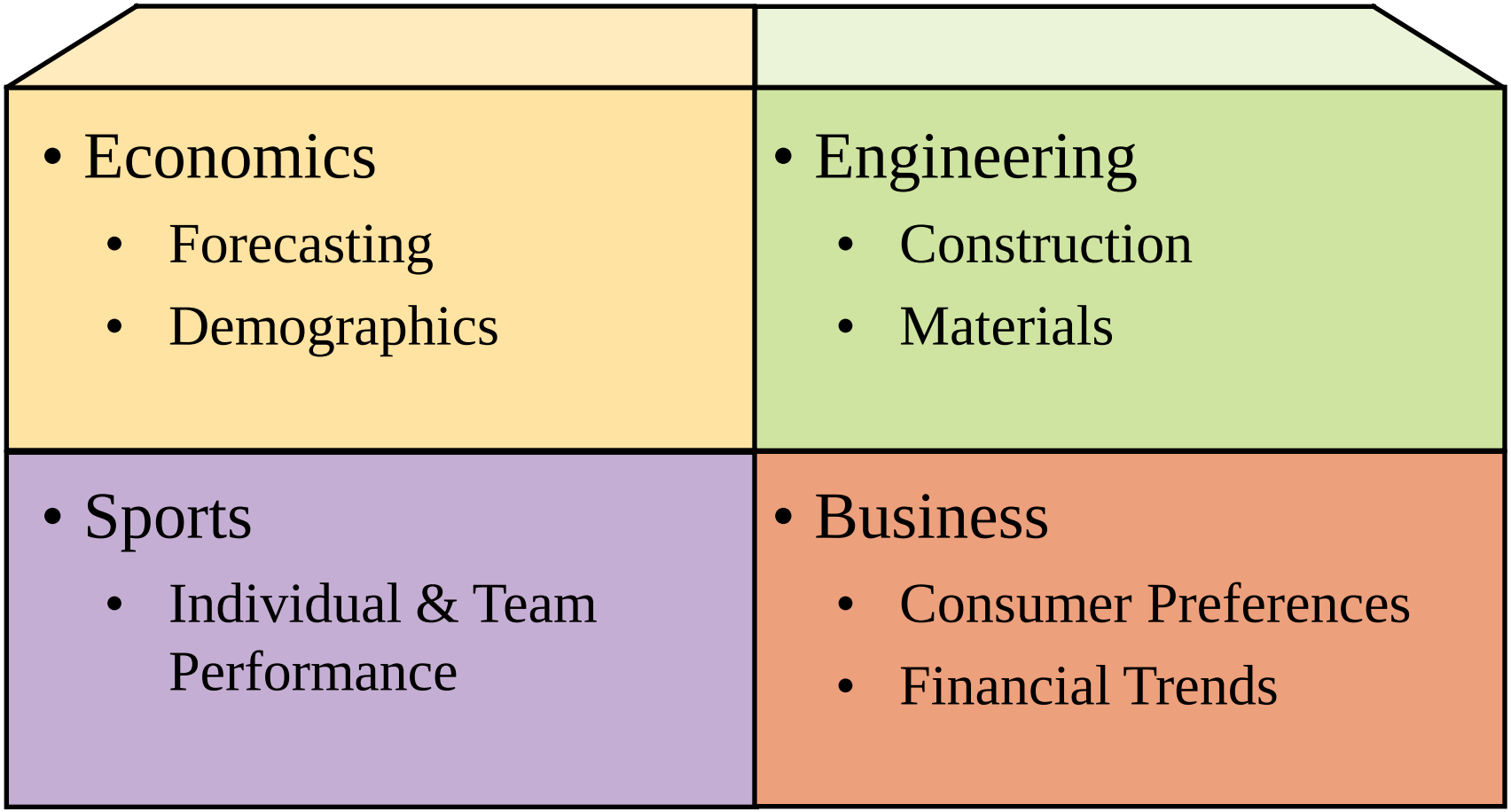
Decision-Making

What Is Statistics?

Statistics is the science of data. It involves collecting, classifying, summarizing, organizing, analyzing, and interpreting numerical information.

Types of Statistical Applications in Business

Application Areas



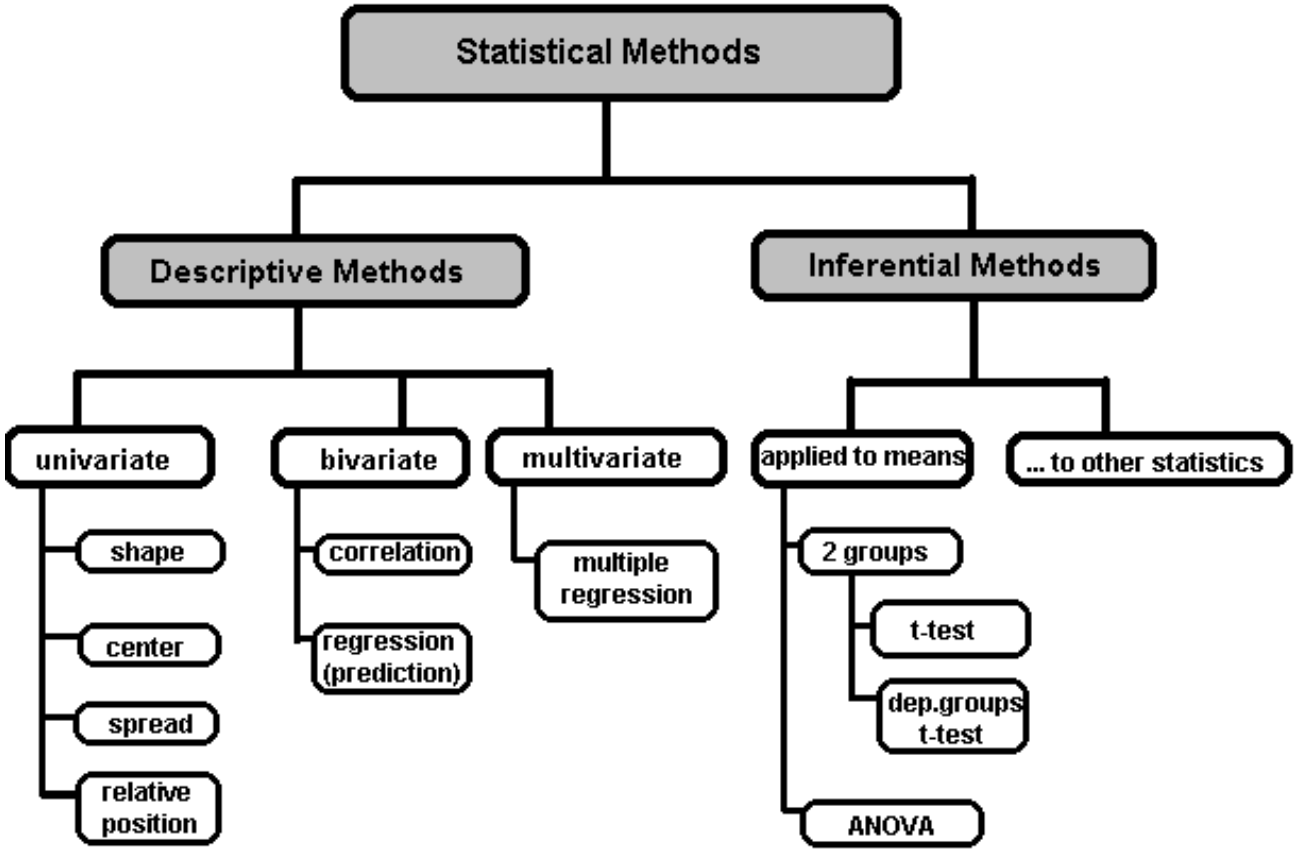
Statistics: Two Processes

Describing sets of data

and

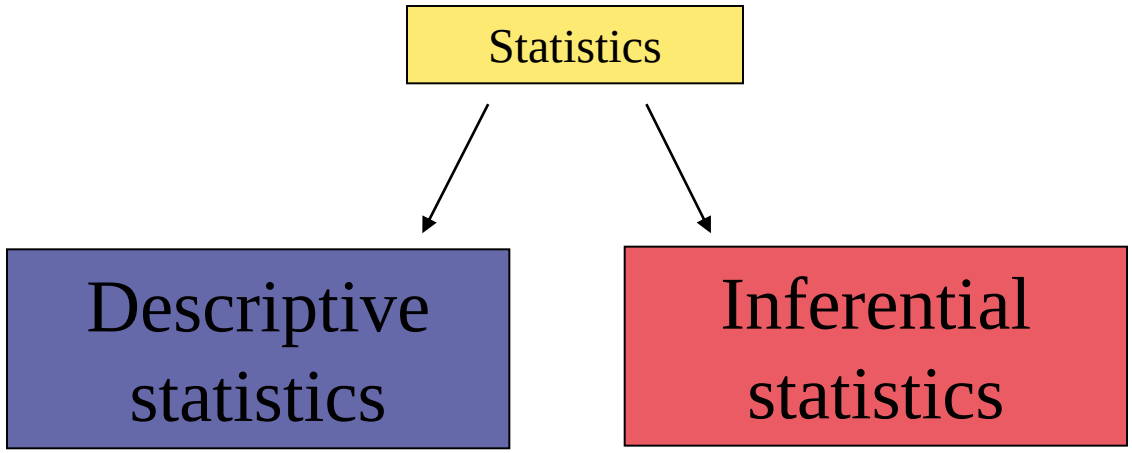
Drawing conclusions (making estimates, decisions, predictions, etc. about sets of data based on sampling)

A Taxonomy of Statistics



Branches of Statistics

The study of statistics has two major branches: **descriptive statistics** and **inferential statistics**.



Involves the organization, summarization, and display of data.

Involves using a sample to draw conclusions about a population.

Descriptive and Inferential Statistics

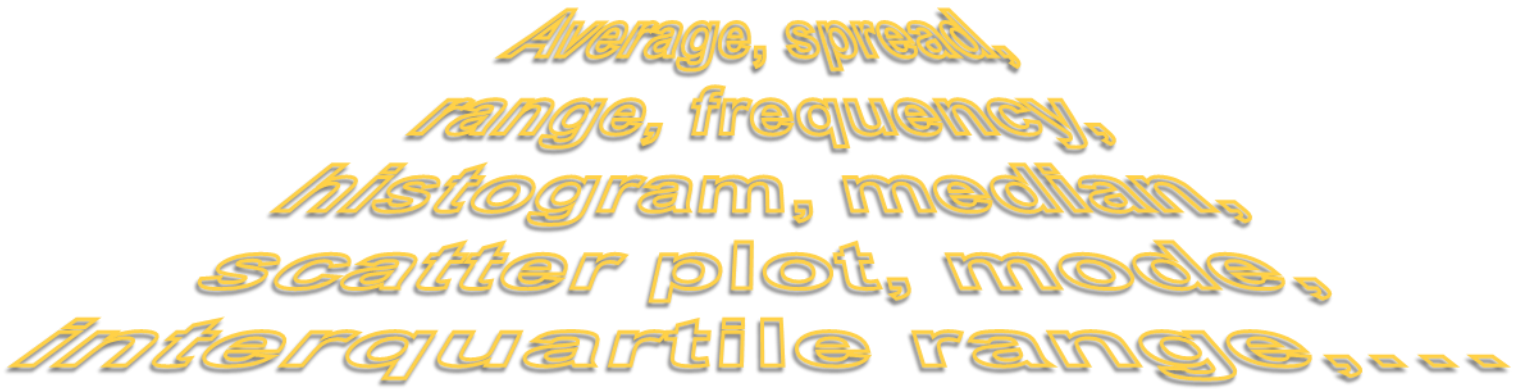
Example:

In a recent study, volunteers who had less than 6 hours of sleep were four times more likely to answer incorrectly on a science test than were participants who had at least 8 hours of sleep. Decide which part is the descriptive statistic and what conclusion might be drawn using inferential statistics.

The statement "four times more likely to answer incorrectly" is a descriptive statistic. An inference drawn from the sample is that all individuals sleeping less than 6 hours are more likely to answer science question incorrectly than individuals who sleep at least 8 hours.

Types of Statistical Applications

- **Descriptive statistics** utilizes numerical and graphical methods to look for patterns in a data set, to summarize the information revealed in a data set and to present that information in a convenient form.



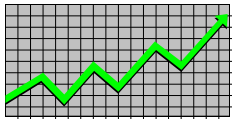
Descriptive Statistics

- **Descriptive statistics** are methods for organizing and summarizing data.
- For example, tables or graphs are used to organize data, and descriptive values such as the average score are used to summarize data.
- A descriptive value for a population is called a **parameter** and a descriptive value for a sample is called a **statistic**.

Descriptive Statistics

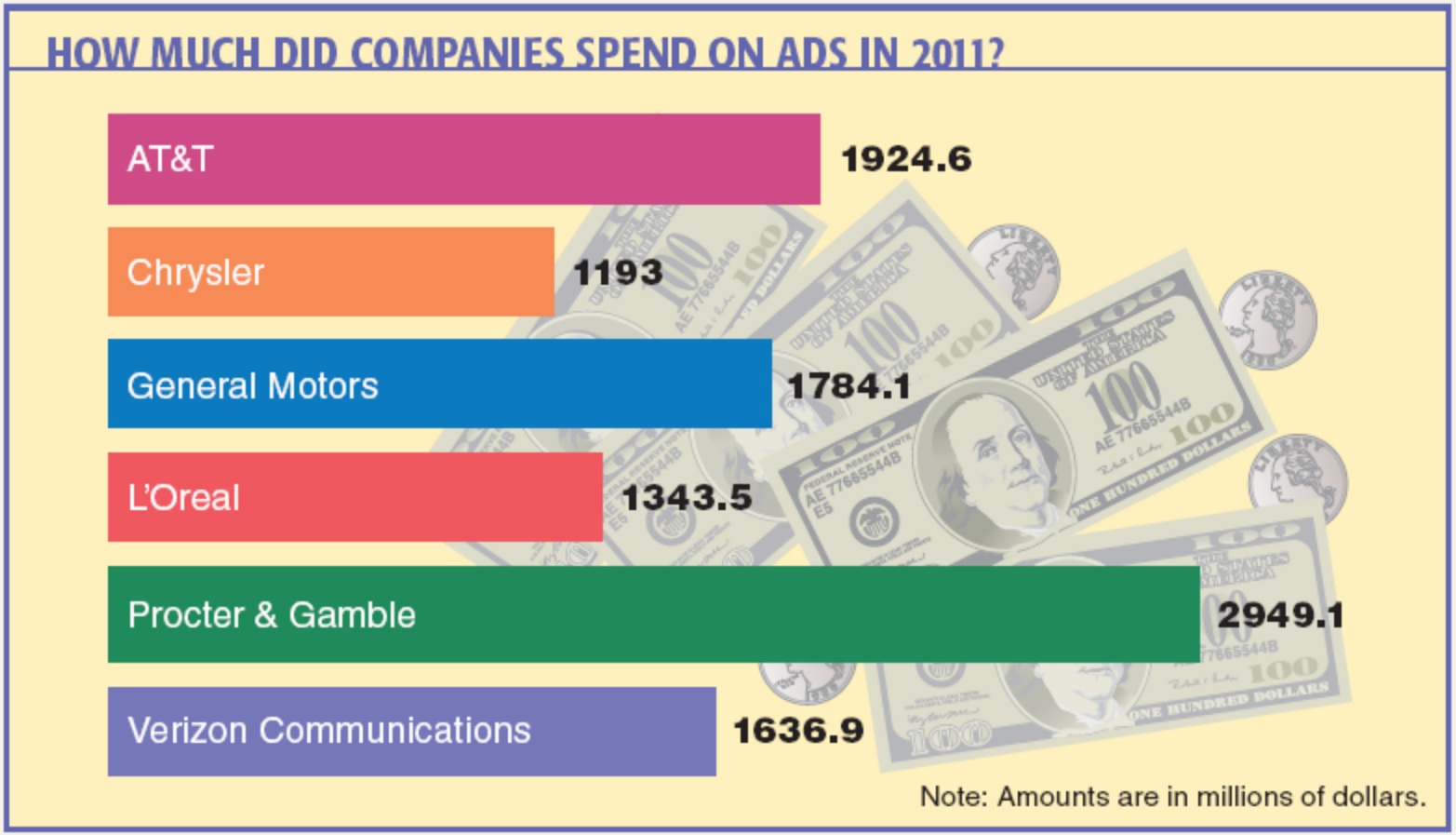
- Collect data
  - e.g. Survey

- Present data
  - e.g. Tables and graphs

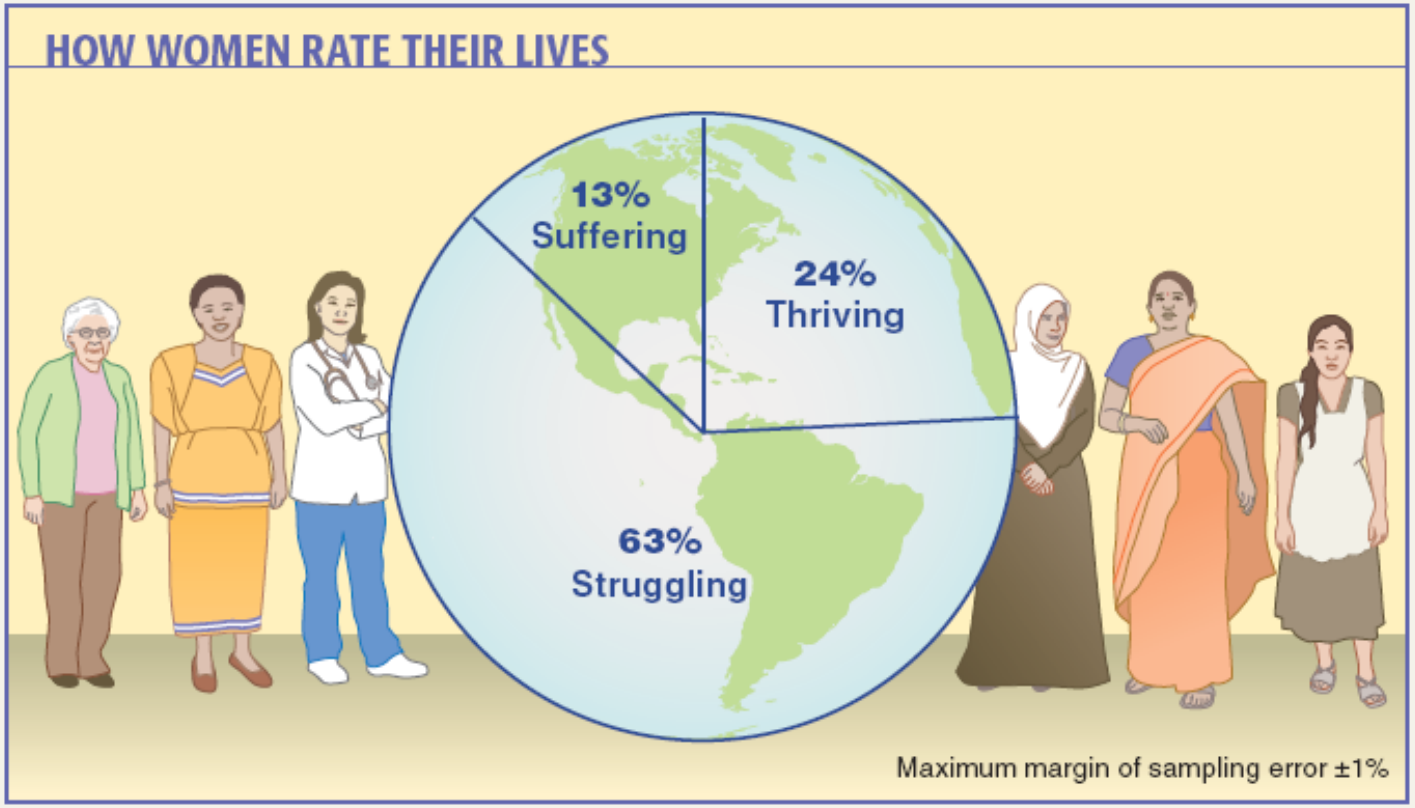


- Characterize data
  - e.g. Sample mean = $\frac{\sum X_i}{n}$

Case Study 1- How Much Did Companies Spend on Ads in 2011?



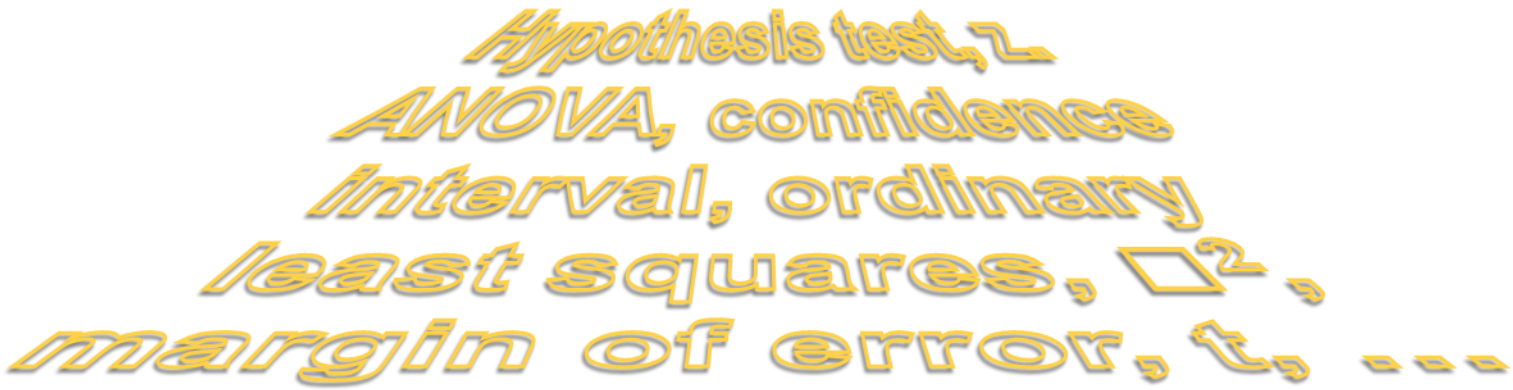
Case Study 2 How Women Rate Their Lives



Data source: Gallup poll of adult women aged 15 and older conducted during 2011 in 147 countries and areas.

Types of Statistical Applications

- **Inferential statistics** utilizes sample data to make estimates, decisions, predictions or other generalizations about a larger set of data.



Inferential Statistics

- Estimation
  - e.g.: Estimate the population mean weight using the sample mean weight
- Hypothesis testing
  - e.g.: Test the claim that the population mean weight is 120 pounds

Drawing conclusions and/or making decisions concerning a population based on sample results.

Inferential Statistics

1. Involves
   - Estimation
   - Hypothesis Testing
2. Purpose
   - Make decisions about population characteristics

Examples of Inferential Statistics

A new milk formulation designed to improve the psychomotor development of infants was tested on randomly selected infants. Based on the results, it was concluded that the new milk formulation is effective in improving the psychomotor development of infants.

Fundamental Elements of Statistics

Descriptive Statistics

- The population or sample of interest
- One or more variables to be investigated
- Tables, graphs or numerical summary tools
- Identification of patterns in the data

Inferential Statistics

- Population of interest
- One or more variables to be investigated
- The sample of population units
- The inference about the population based on the sample data
- A measure of reliability of the inference

Fundamental Elements of Statistics

- An **experimental unit** is an object about which we collect data.
  - Person
  - Place
  - Thing
  - Event

Fundamental Elements of Statistics

- A **population** is a set of units in which we are interested.
  - Typically, there are too many experimental units in a population to consider *every* one.
    - If we can examine every single one, we conduct a **census**.

  A ***population*** consists of all elements – individuals, items, or objects – whose characteristics are being studied.

  The population that is being studied is also called the ***target population***.

Fundamental Elements of Statistics

- A **variable** is a characteristic or property of an individual unit.

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

Contd.,

The value of a variable for an element is called an ***observation*** or ***measurement***.

A ***data set*** is a collection of observations on one or more variables.

Table :Total Revenues for 2010 of Six Companies

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

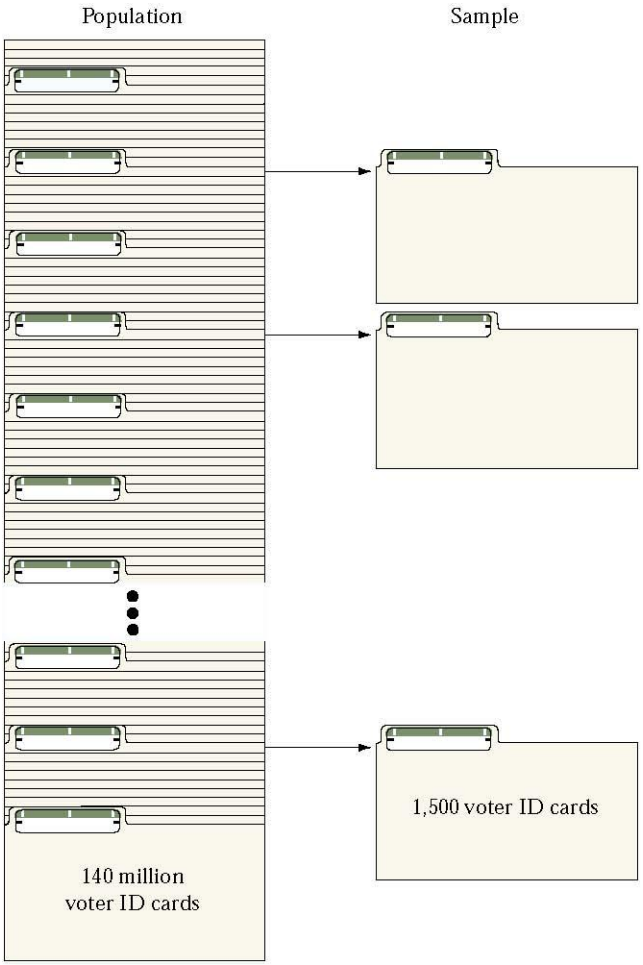
Source: Fortune Magazine, July 25, 2011.

An element or a member → Exxon Mobil

354,674 ← An observation or measurement

2010 Total Revenue (millions of dollars) ← Variable

Fundamental Elements of Statistics



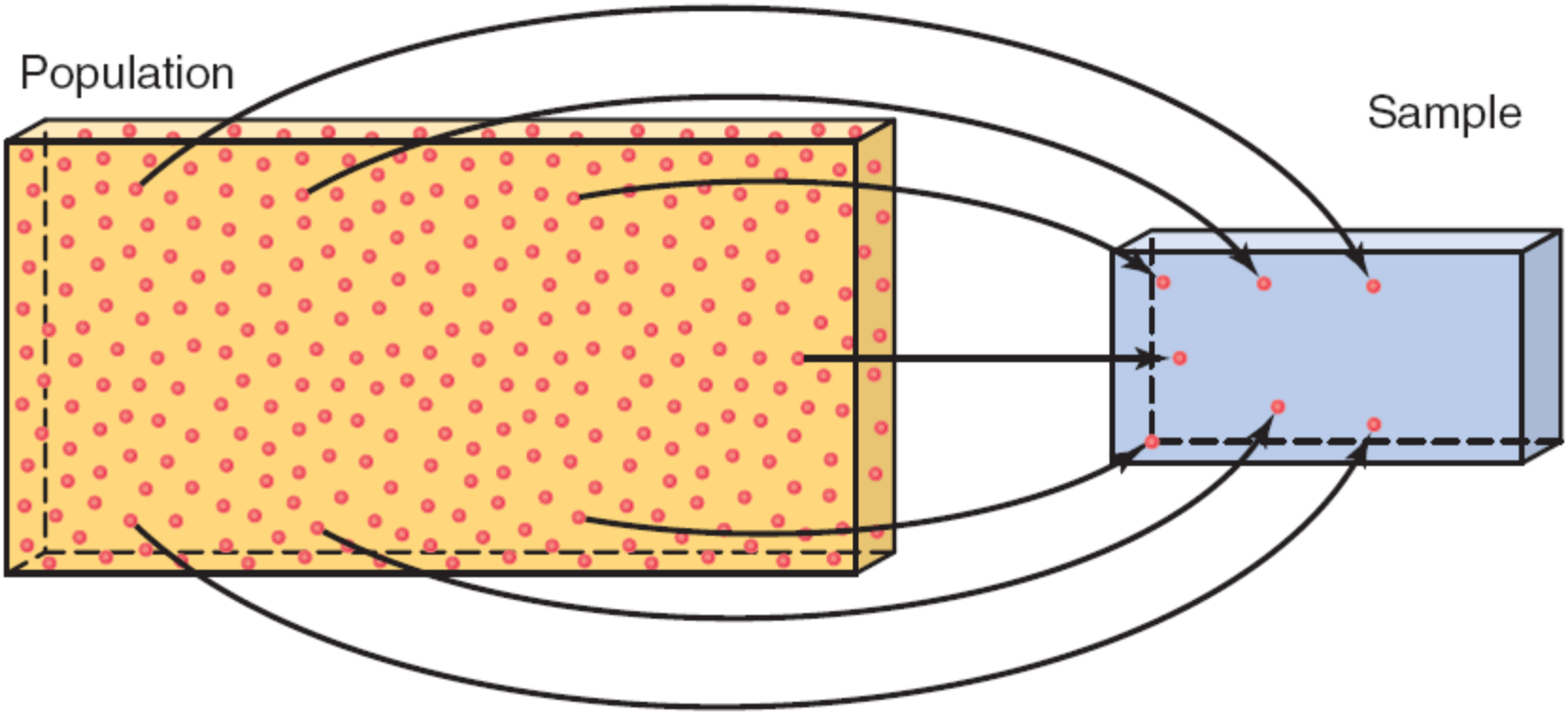
- A **sample** is a subset of the population.
- A portion of the population selected for study is referred to as a ***sample***.

Contd.,

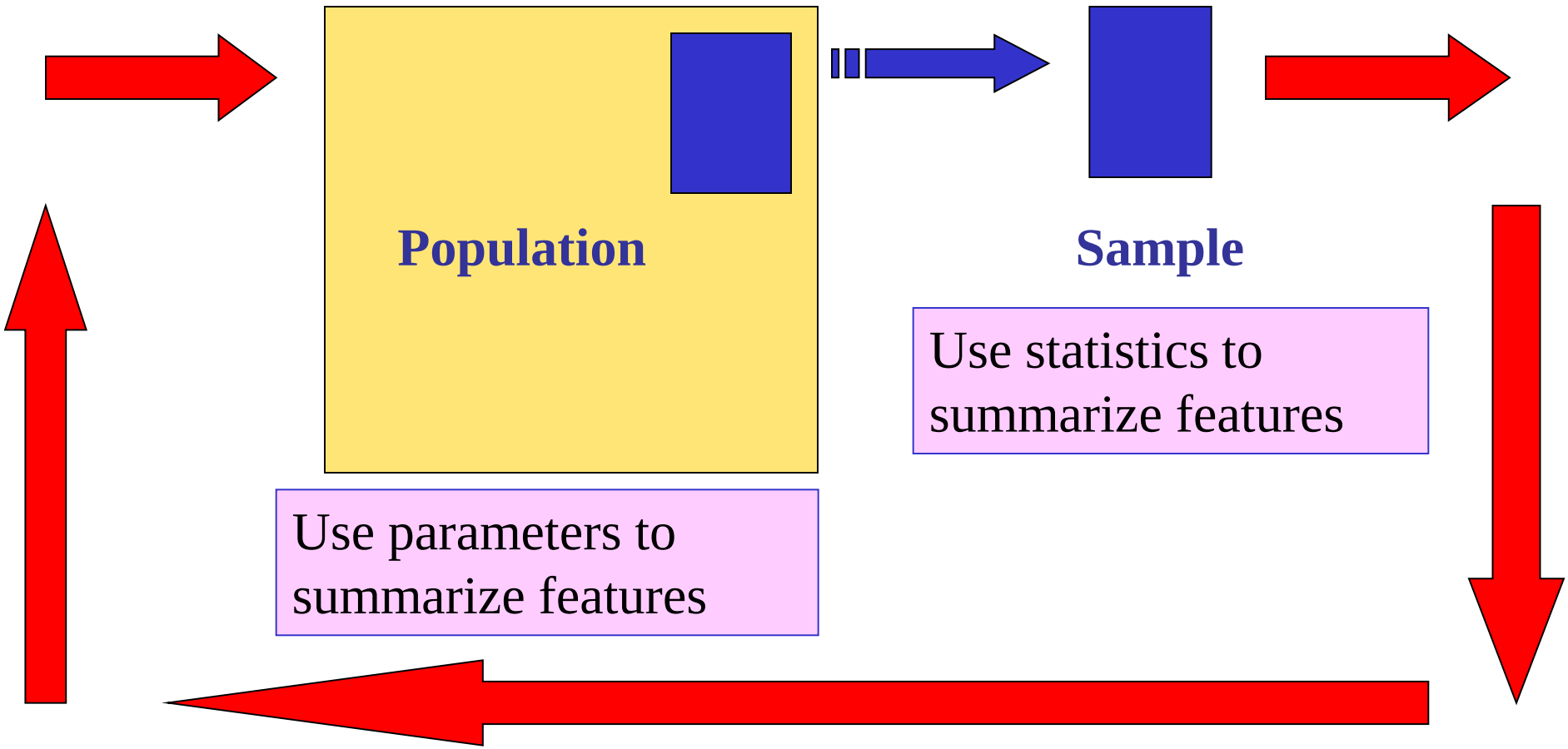
- A parameter is a summary measure computed to describe a characteristic of the population.
- - : numerical characteristic of a population
- A statistic is a summary measure computed to describe a characteristic of the sample
- - : numerical characteristic of a sample
- An element is an object on which a measurement is taken.
- - An ***element*** or ***member*** of a sample or population is a specific subject or object (for example, a person, firm, item, state, or country) about which the information is collected.

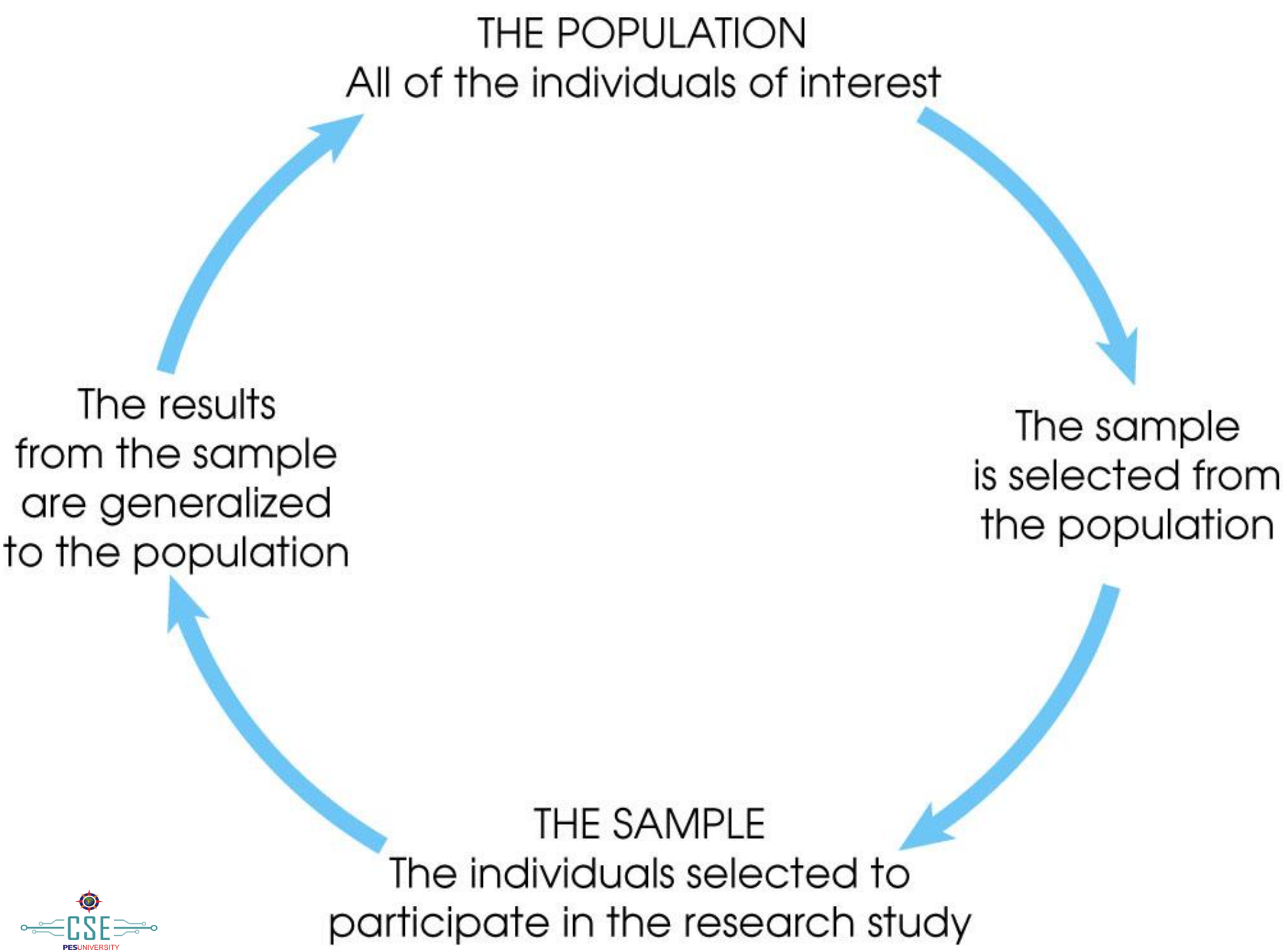
- Sampling units are nonoverlapping collections of elements from the population that cover the entire population
- A sampling frame is a list of sampling units.

Population and Sample



Population and Sample





Contd.,

A survey that includes every member of the population is called a ***census***.

The technique of collecting information from a portion of the population is called a ***sample survey***.

A sample that represents the characteristics of the population as closely as possible is called a ***representative sample***.

A representative sample exhibits characteristics typical of those possessed by the target population.

Contd.,

A sample drawn in such a way that each element of the population has a chance of being selected is called a ***random sample***.

A random sample of n units is selected in such a way that every different sample of size n has the same chance of being selected.

If all samples of the same size selected from a population have the same chance of being selected, we call it ***simple random sampling***. Such a sample is called a ***simple random sample***.

Contd.,

A sample may be selected with or without replacement.

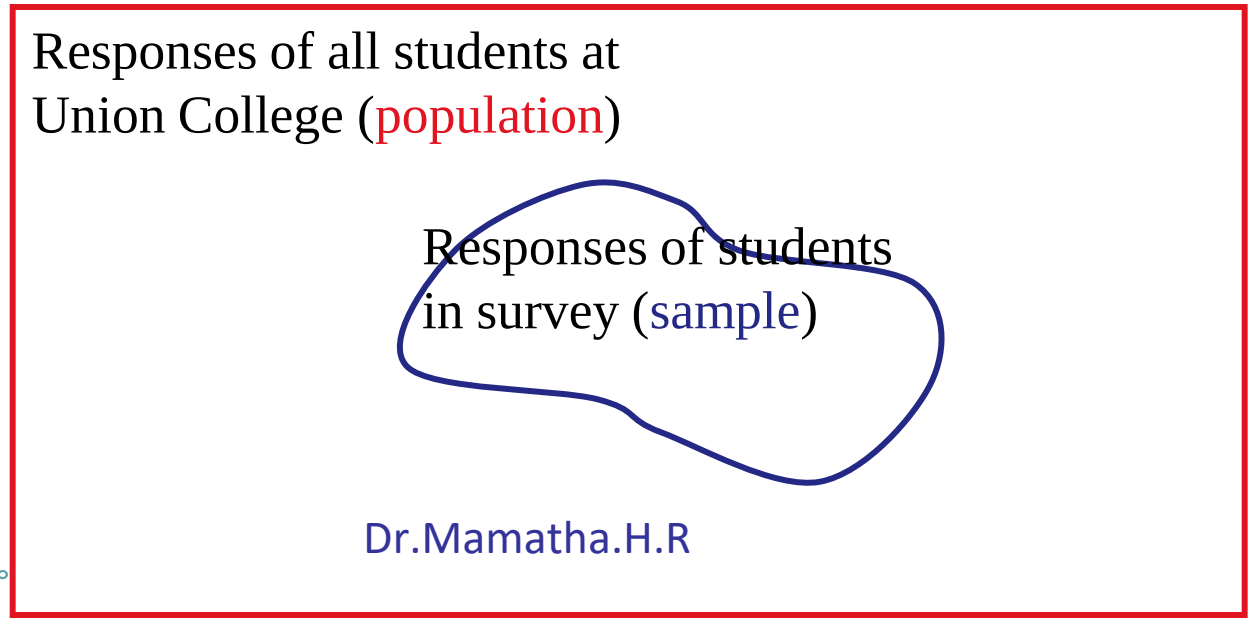
In sampling ***with replacement***, each time we select an element from the population, we put it back in the population before we select the next element.

Sampling ***without replacement*** occurs when the selected element is not replaced in the population.

Populations & Samples

Example:

In a recent survey, 250 college students at Union College Were asked if they smoked cigarettes regularly. 35 of the students said yes. Identify the population and the sample.



Parameters & Statistics

Example:

Decide whether the numerical value describes a population parameter or a sample statistic.

a.) A recent survey of a sample of 450 college students reported that the average weekly income for students is $325.

b.) The average weekly income for all students is $405.

Parameters & Statistics

Example:

Decide whether the numerical value describes a population parameter or a sample statistic.

a.) A recent survey of a sample of 450 college students reported that the average weekly income for students is $325.

Because the average of $325 is based on a sample, this is a sample statistic.

b.) The average weekly income for all students is $405.

Because the average of $405 is based on a population, this is a population parameter.

SAMPLING

LEARNING OBJECTIVES

- **Learn the reasons for sampling**
- **Develop an understanding about different sampling methods**
- **Distinguish between probability & non probability sampling**
- **Discuss the relative advantages & disadvantages of each sampling methods**

SAMPLING

- Why sample?
  - Resources (time, money) and workload
  - Gives results with known accuracy that can be calculated mathematically

SAMPLING……

- What is your population of interest?
  - To whom do you want to generalize your results?
    - All doctors
    - School children
    - Indians
    - Women aged 15-45 years
    - Other
- Can you sample the entire population?

SAMPLING…….

- 3 factors that influence sample representative-ness
  - Sampling procedure
  - Sample size
  - Participation (response)

- When might you sample the entire population?
  - When your population is very small
  - When you have extensive resources
  - When you don’t expect a very high response

Who do you want to generalize to?

The Theoretical Population

What population can you get access to?

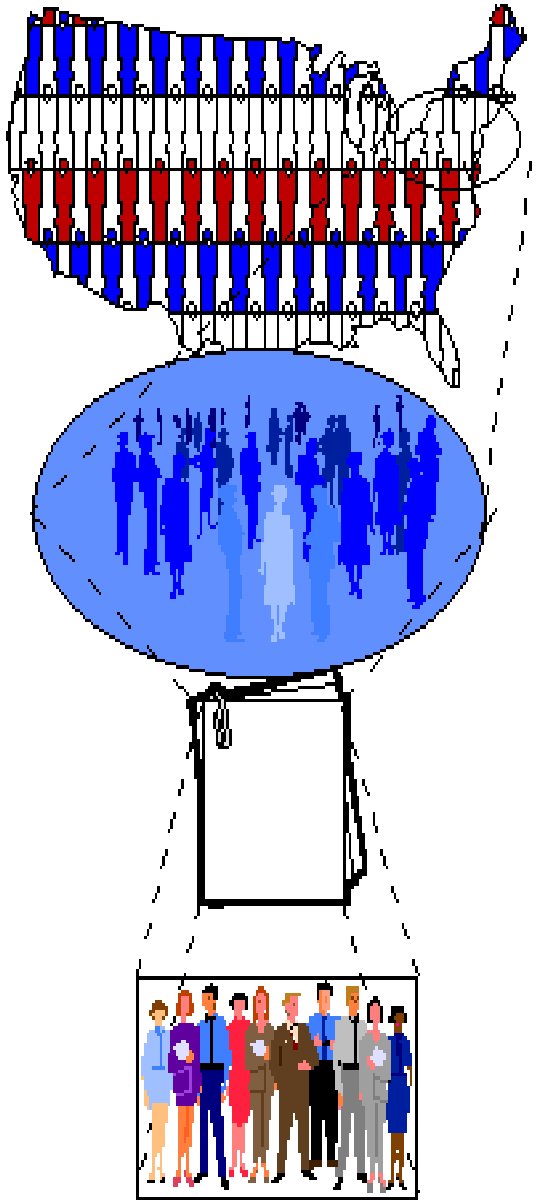
The Study Population

How can you get access to them?

The Sampling Frame

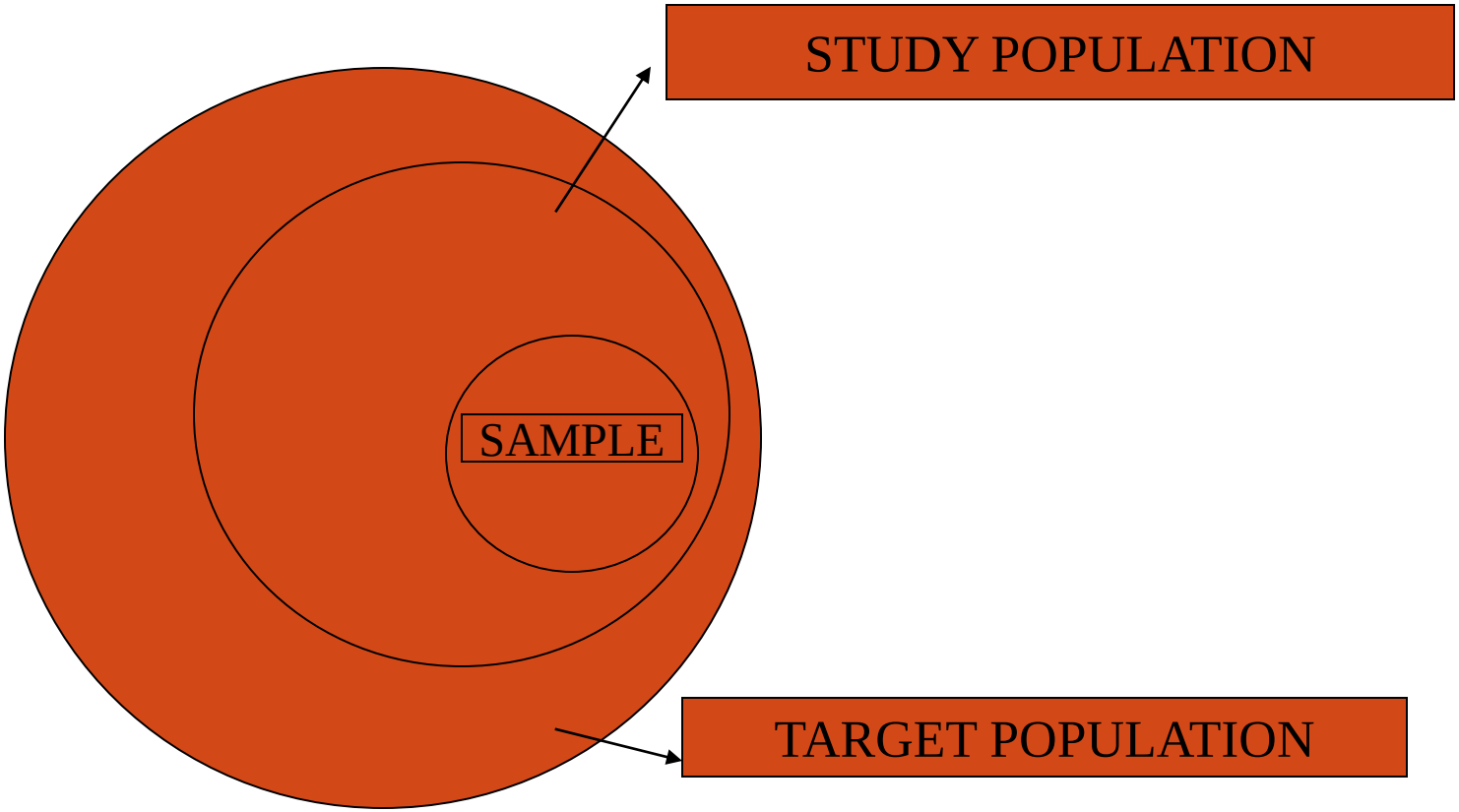
Who is in your study?

The Sample



SAMPLING BREAKDOWN

SAMPLING…….



Process

- The sampling process comprises several stages:
  - Defining the population of concern
  - Specifying a sampling frame, a set of items or events possible to measure
  - Specifying a sampling method for selecting items or events from the frame
  - Determining the sample size
  - Implementing the sampling plan
  - Sampling and data collecting
  - Reviewing the sampling process

Population versus sample

- A population can be defined as including all people or items with the characteristic one wishes to understand.
- Because there is very rarely enough time or money to gather information from everyone or everything in a population, the goal becomes finding a representative sample (or subset) of that population.

Contd.,

- Note also that the population from which the sample is drawn may not be the same as the population about which we actually want information. Often there is large but not complete overlap between these two groups due to frame issues etc .

- Sometimes they may be entirely separate - for instance, we might study rats in order to get a better understanding of human health, or we might study records from people born in 2008 in order to make predictions about people born in 2009.

SAMPLING FRAME

- In the most straightforward case, such as the sentencing of a batch of material from production (acceptance sampling by lots), it is possible to identify and measure every single item in the population and to include any one of them in our sample.
- However, in the more general case this is not possible. There is no way to identify all rats in the set of all rats. Where voting is not compulsory, there is no way to identify which people will actually vote at a forthcoming election (in advance of the election)
- As a remedy, we seek a *sampling frame* which has the property that we can identify every single element and include any in our sample .
- The sampling frame must be representative of the population

the populations consisting of actual physical objects—the students at a university, the concrete blocks in a pile, the bolts in a shipment. Such populations are called tangible populations. Tangible populations are always finite

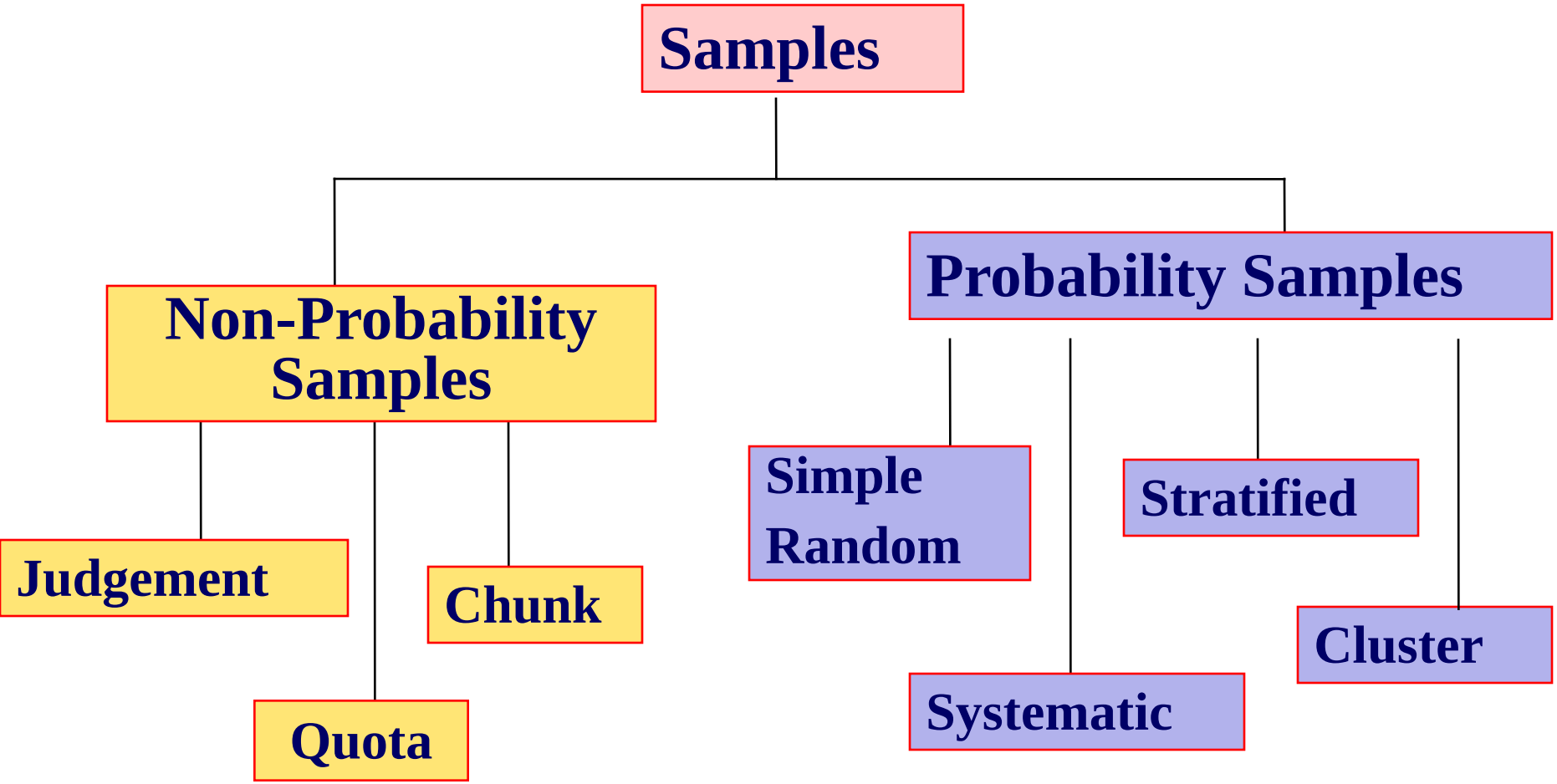
A simple random sample may consist of values obtained from a process under identical experimental conditions.

In this case, the sample comes from a population that consists of all the values that might possibly have been observed.

Such a population is called a conceptual population.

Imagine that an engineer measures the length of a rod five times, being as careful as possible to take the measurements under identical conditions. No matter how carefully the measurements are made, they will differ somewhat from one another,because of variation in the measurement process that cannot be controlled or predicted.

Types of Sampling Methods



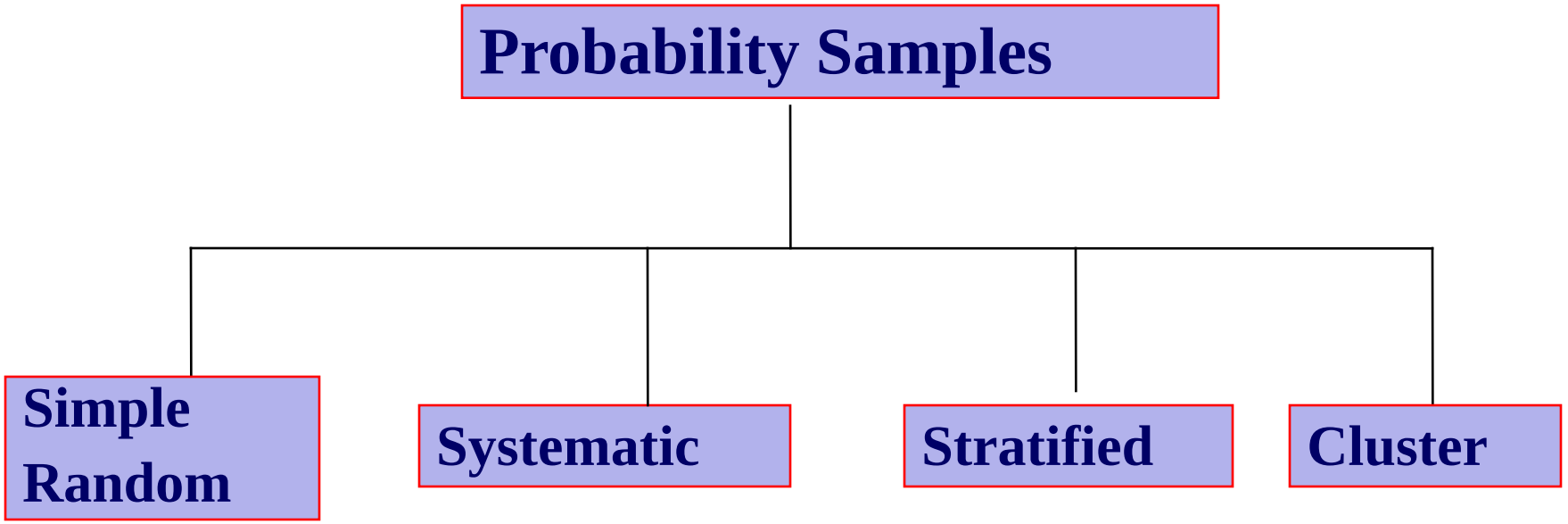
Types of Samples

- Probability (Random) Samples

  Simple random sample
  - Systematic random sample
  - Stratified random sample
  - Multistage sample
  - Multiphase sample
  - Cluster sample
- Non-Probability Samples
  - Convenience sample
  - Purposive sample
  - Quota

Probability Sampling

- Subjects of the sample are chosen based on known probabilities



PROBABILITY SAMPLING

- A **probability sampling** scheme is one in which every unit in the population has a chance (greater than zero) of being selected in the sample, and this probability can be accurately determined.

- . When every element in the population *does* have the same probability of selection, this is known as an 'equal probability of selection' (EPS) design. Such designs are also referred to as 'self-weighting' because all sampled units are given the same weight.

PROBABILITY SAMPLING…….

- Probability sampling includes:
- Simple Random Sampling,
- Systematic Sampling,
- Stratified Random Sampling,
- Cluster Sampling
- Multistage Sampling.
- Multiphase sampling

NON PROBABILITY SAMPLING

- Any sampling method where some elements of population have *no* chance of selection (these are sometimes referred to as 'out of coverage'/'undercovered'), or where the probability of selection can't be accurately determined.
- It involves the selection of elements based on assumptions regarding the population of interest, which forms the criteria for selection.
- Hence, because the selection of elements is nonrandom, nonprobability sampling not allows the estimation of sampling errors..

- *Example: We visit every household in a given street, and interview the first person to answer the door. In any household with more than one occupant,*

- *this is a nonprobability sample, because some people are more likely to answer the door (e.g. an unemployed person who spends most of their time at home is more likely to answer than an employed housemate who might be at work when the interviewer calls) and it's not practical to calculate these probabilities*.

NONPROBABILITY SAMPLING…….

- Nonprobability Sampling includes: Accidental Sampling, Quota Sampling and Purposive Sampling.
- In addition, nonresponse effects may turn *any* probability design into a nonprobability design if the characteristics of nonresponse are not well understood, since nonresponse effectively modifies each element's probability of being sampled.

SIMPLE RANDOM SAMPLING

- Applicable when population is small, homogeneous & readily available
- All subsets of the frame are given an equal probability. Each element of the frame thus has an equal probability of selection.
- It provides for greatest number of possible samples. This is done by assigning a number to each unit in the sampling frame.
- A table of random number or lottery system is used to determine which units are to be selected.

Simple Random Sample

Number each frame unit from 1 to *N*.

Use a random number table or a random number generator to select *n* distinct numbers between 1 and *N*, inclusively.

Easier to perform for small populations

Cumbersome for large populations

Simple Random Sample: Numbered Population Frame

01 Alaska Airlines	11 DuPont	21 LTV
02 Alcoa	12 Exxon	22 Litton
03 Amoco	13 Farah	23 Mead
04 Atlantic Richfield	14 GTE	24 Mobil
05 Bank of America	15 General Electric	25 Occidental Petroleum
06 Bell of Pennsylvania	16 General Mills	26 JCPenney
07 Chevron	17 General Dynamics	27 Philadelphia Electric
08 Chrysler	18 Grumman	28 Ryder
09 Citicorp	19 IBM	29 Sears
10 Disney	20 Kmart	30 Time

Simple Random Sampling: Random Number Table

9	9	4	3	7	8	7	9	6	1	4	5	7	3	7	3	7	5	5	2	9	7	9	6	9	3	9	0	9	4	3	4	4	7	5	3	1	6	1	8
5	0	6	5	6	0	0	1	2	7	6	8	3	6	7	6	6	8	8	2	0	8	1	5	6	8	0	0	1	6	7	8	2	2	4	5	8	3	2	6
8	0	8	8	0	6	3	1	7	1	4	2	8	7	7	6	6	8	3	5	6	0	5	1	5	7	0	2	9	6	5	0	0	2	6	4	5	5	8	7
8	6	4	2	0	4	0	8	5	3	5	3	7	9	8	8	9	4	5	4	6	8	1	3	0	9	1	2	5	3	8	8	1	0	4	7	4	3	1	9
6	0	0	9	7	8	6	4	3	6	0	1	8	6	9	4	7	7	5	8	8	9	5	3	5	9	9	4	0	0	4	8	2	6	8	3	0	6	0	6
5	2	5	8	7	7	1	9	6	5	8	5	4	5	3	4	6	8	3	4	0	0	9	9	1	9	9	7	2	9	7	6	9	4	8	1	5	9	4	1
8	9	1	5	5	9	0	5	5	3	9	0	6	8	9	4	8	6	3	7	0	7	9	5	5	4	7	0	6	2	7	1	1	8	2	6	4	4	9	3

$N = 30$

$n = 6$

Simple Random Sample: Sample Members

01 Alaska Airlines	11 DuPont	21 LTV
02 Alcoa	12 Exxon	22 Litton
03 Amoco	13 Farah	23 Mead
04 Atlantic Richfield	14 GTE	24 Mobil
05 Bank of America	15 General Electric	25 Occidental Petroleum
06 Bell Pennsylvania	16 General Mills	26 Penney
07 Chevron	17 General Dynamics	27 Philadelphia Electric
08 Chrysler	18 Grumman	28 Ryder
09 Citicorp	19 IBM	29 Sears
10 Disney	20 KMart	30 Time

$N = 30$

$n = 6$

SIMPLE RANDOM SAMPLING……..

- Estimates are easy to calculate.
- Simple random sampling is always an EPS design, but not all EPS designs are simple random sampling.

- **Disadvantages**
- If sampling frame large, this method impracticable.
- Minority subgroups of interest in population may not be present in sample in sufficient numbers for study.

Simple random samples always differ from their populations in some ways, and occasionally may be substantially different.

Two different samples from the same population will differ from each other as well. This phenomenon is known as **sampling variation**.

Sampling variation is one of the reasons that scientific experiments produce somewhat different results when repeated, even when the conditions appear to be identical

Independence

The items in a sample are said to be independent if knowing the values of some of them does not help to predict the values of the others.

With a finite, tangible population,the items in a simple random sample are not strictly independent, because as each item is drawn, the population changes. This change can be substantial when the population is small.

However, when the population is very large, this change is negligible and the items can be treated as if they were independent

REPLACEMENT OF SELECTED UNITS

Sampling schemes may be *without replacement* ('WOR' - no element can be selected more than once in the same sample) or *with replacement* ('WR' - an element may appear multiple times in the one sample).

- For example, if we catch fish, measure them, and immediately return them to the water before continuing with the sample, this is a WR design, because we might end up catching and measuring the same fish more than once. However, if we do not return the fish to the water (e.g. if we eat the fish), this becomes a WOR design.

SYSTEMATIC SAMPLING

- **Systematic sampling** relies on arranging the target population according to some ordering scheme and then selecting elements at regular intervals through that ordered list.
- Systematic sampling involves a random start and then proceeds with the selection of every *k*th element from then onwards. In this case, *k*=(population size/sample size).
- It is important that the starting point is not automatically the first in the list, but is instead randomly chosen from within the first to the *k*th element in the list.
- A simple example would be to select every 10th name from the telephone directory (an 'every 10th' sample, also referred to as 'sampling with a skip of 10').

Systematic Sampling

Convenient and relatively easy to administer

Population elements are an ordered sequence (at least, conceptually).

The first sample element is selected randomly from the first *k* population elements.

Thereafter, sample elements are selected at a constant interval, *k*, from the ordered sequence frame.

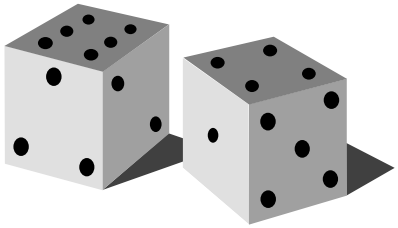
$$k = \frac{N}{n},$$

where :

n = sample size

N = population size

k = size of selection interval

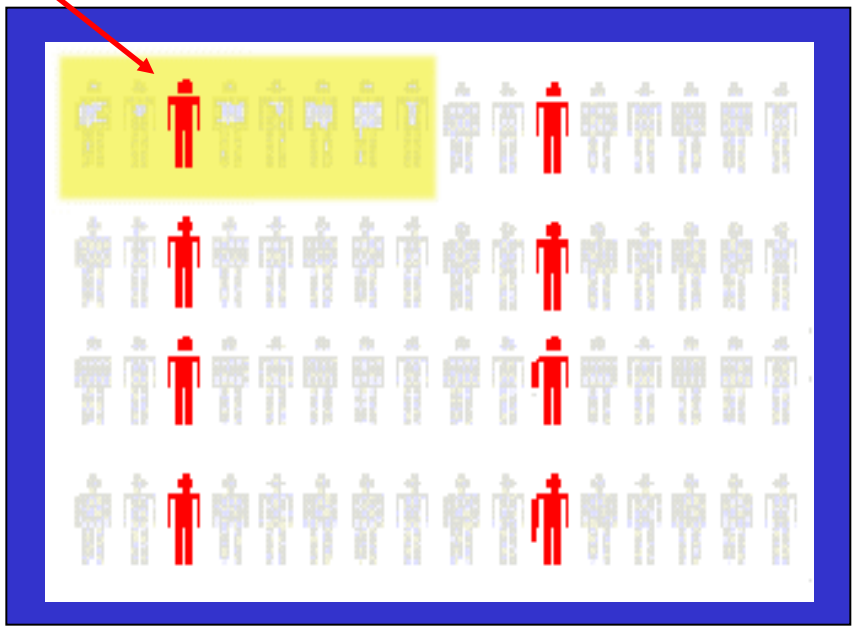


N = 64

n = 8

k = 8

First Group



Systematic Samples

- Decide on sample size: n
- Divide frame of N individuals into groups of k individuals: k=n/n
- Randomly select one individual from the 1st group
- Select every k-th individual thereafter

SYSTEMATIC SAMPLING……

As described above, systematic sampling is an EPS method, because all elements have the same probability of selection (in the example given, one in ten). It is *not* 'simple random sampling' because different subsets of the same size have different selection probabilities - e.g. the set {4,14,24,...,994} has a one-in-ten probability of selection, but the set {4,13,24,34,...} has zero probability of selection.

Systematic Sampling: Example

Purchase orders for the previous fiscal year are serialized 1 to 10,000 (N = 10,000).A sample of fifty (n = 50) purchases orders is needed for an audit.

k = 10,000/50 = 200

First sample element randomly selected from the first 200 purchase orders. Assume the 45th purchase order was selected.

Subsequent sample elements: 245, 445, 645, . . .

SYSTEMATIC SAMPLING……

- **ADVANTAGES:**
- Sample easy to select
- Suitable sampling frame can be identified easily
- Sample evenly spread over entire reference population
- **DISADVANTAGES:**
- Sample may be biased if hidden periodicity in population coincides with that of selection.
- Difficult to assess precision of estimate from one survey.

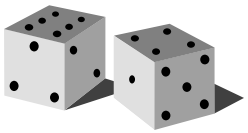
STRATIFIED SAMPLING

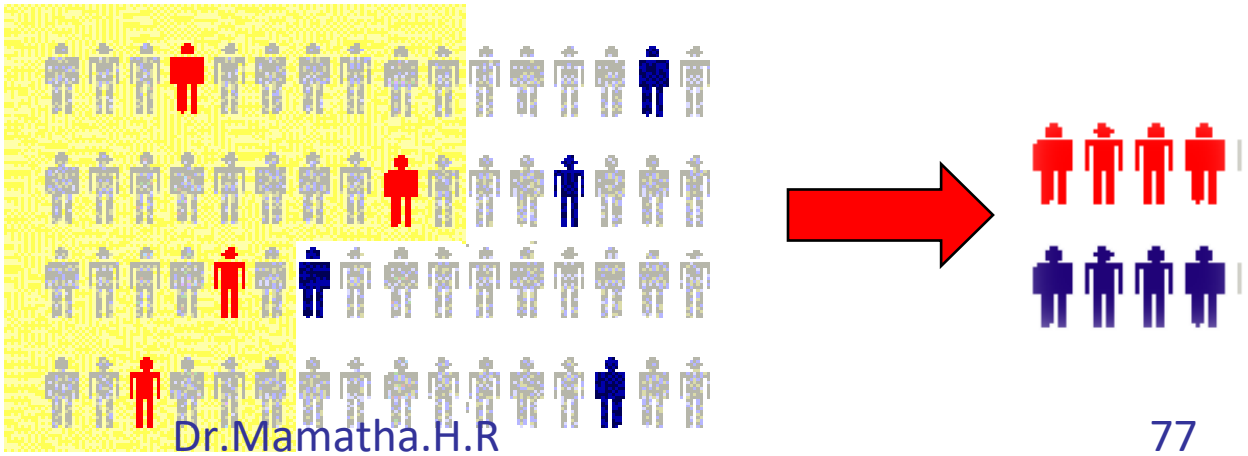
Population divided into two or more groups according to some common characteristic

Simple random sample selected from each group

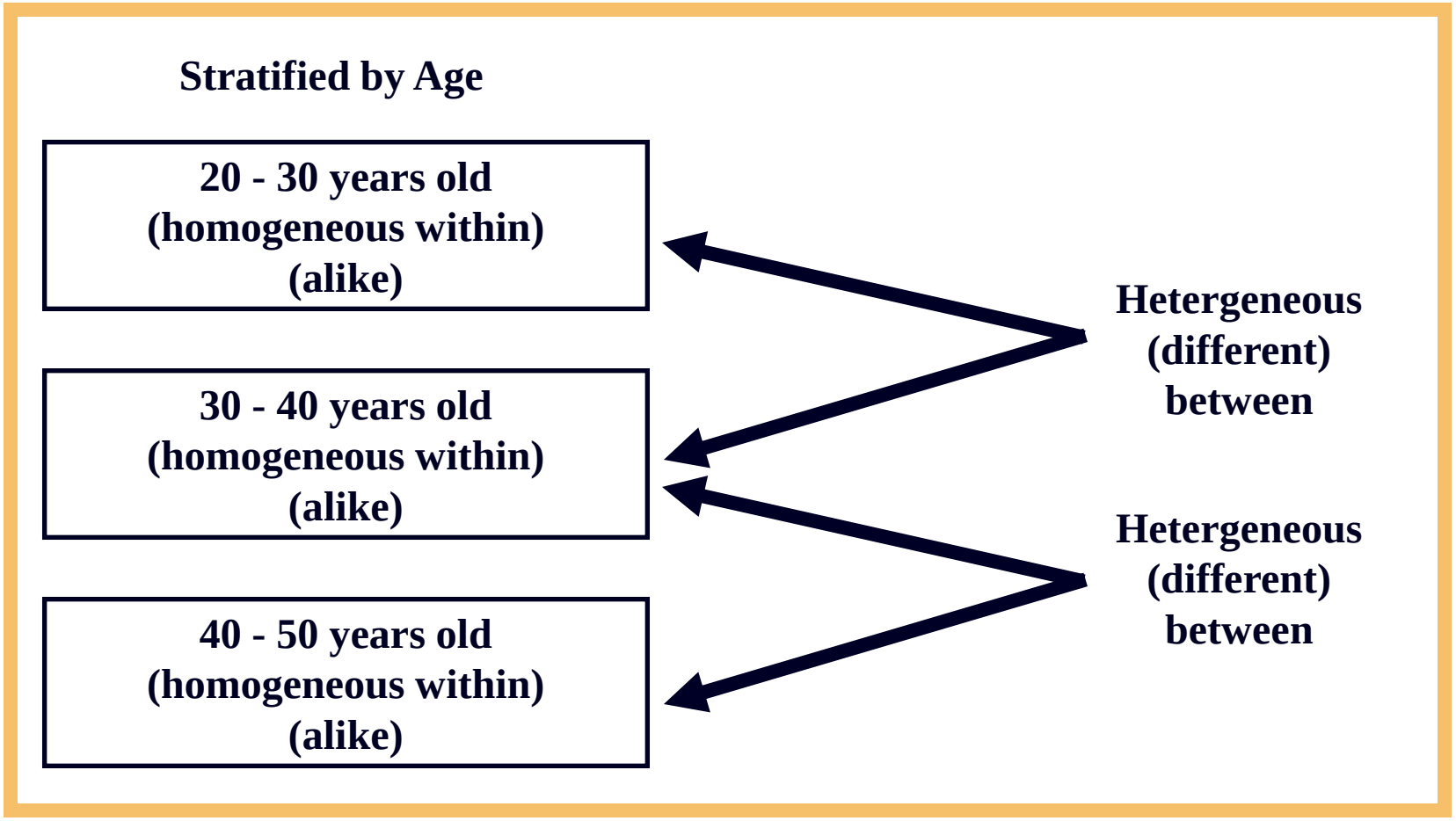
The two or more samples are combined into one

Stratified Samples





Stratified Random Sample: Population of FM Radio Listeners



STRATIFIED SAMPLING…….

Draw a sample from each stratum

STRATIFIED SAMPLING

Where population embraces a number of distinct categories, the frame can be organized into separate "strata." Each stratum is then sampled as an independent sub-population, out of which individual elements can be randomly selected.

- Every unit in a stratum has same chance of being selected.
- Using same sampling fraction for all strata ensures proportionate representation in the sample.
- Adequate representation of minority subgroups of interest can be ensured by stratification & varying sampling fraction between strata as required.

STRATIFIED SAMPLING……

- Finally, since each stratum is treated as an independent population, different sampling approaches can be applied to different strata.

Drawbacks to using stratified sampling.

- First, sampling frame of entire population has to be prepared separately for each stratum
- Second, when examining multiple criteria, stratifying variables may be related to some, but not to others, further complicating the design, and potentially reducing the utility of the strata.
- Finally, in some cases (such as designs with a large number of strata, or those with a specified minimum sample size per group), stratified sampling can potentially require a larger sample than would other methods

Cluster Sampling

Population is divided into nonoverlapping clusters or areas

Each cluster is a miniature, or microcosm, of the population.

A subset of the clusters is selected randomly for the sample.

If the number of elements in the subset of clusters is larger than the desired value of *n*, these clusters may be subdivided to form a new set of clusters and subjected to a random selection process.

Cluster Sampling: Some Test Market Cities

Cluster Sampling

- **Advantages**
  - More convenient for geographically dispersed populations
  - Reduced travel costs to contact sample elements
  - Simplified administration of the survey
  - Unavailability of sampling frame prohibits using other random sampling methods
- **Disadvantages**
  - Statistically less efficient when the cluster elements are similar
  - Costs and problems of statistical analysis are greater than for simple random sampling

CLUSTER SAMPLING…….

- **Identification of clusters**
  - List all cities, towns, villages & wards of cities with their population falling in target area under study.
  - Calculate cumulative population & divide by 30, this gives sampling interval.
  - Select a random no. less than or equal to sampling interval having same no. of digits. This forms 1st cluster.
  - Random no.+ sampling interval = population of 2nd cluster.
  - Second cluster + sampling interval = 4th cluster.
  - Last or 30th cluster = 29th cluster + sampling interval

CLUSTER SAMPLING…….

Two types of cluster sampling methods.

One-stage sampling. All of the elements within selected clusters are included in the sample.

Two-stage sampling. A subset of elements within selected clusters are randomly selected for inclusion in the sample.

CLUSTER SAMPLING…….

	Freq	c f	cluster
I	2000	2000	1
II	3000	5000	2
III	1500	6500	
IV	4000	10500	3
V	5000	15500	4, 5
VI	2500	18000	6
VII	2000	20000	7
VIII	3000	23000	8
IX	3500	26500	9
X	4500	31000	10
XI	4000	35000	11, 12
XII	4000	39000	13
XIII	3500	44000	14,15
XIV	2000	46000	
XV	3000	49000	16
XVI	3500	52500	17
XVII	4000	56500	18,19
XVIII	4500	61000	20
XIX	4000	65000	21,22
XX	4000	69000	23
XXI	2000	71000	24
XXII	2000	73000	
XXIII	3000	76000	25
XXIV	3000	79000	26
XXV	5000	84000	27,28
XXVI	2000	86000	29
XXVII	1000	87000	
XXVIII	1000	88000	
XXIX	1000	89000	30
XXX	1000	90000	

- 90000/30 = 3000 sampling interval

Difference Between Strata and Clusters

- Although strata and clusters are both non-overlapping subsets of the population, they differ in several ways.
- All strata are represented in the sample; but only a subset of clusters are in the sample.
- With stratified sampling, the best survey results occur when elements within strata are internally homogeneous. However, with cluster sampling, the best results occur when elements within clusters are internally heterogeneous.

Nonrandom Sampling

Convenience Sampling: sample elements are selected for the convenience of the researcher

Judgment Sampling: sample elements are selected by the judgment of the researcher

Quota Sampling: sample elements are selected until the quota controls are satisfied

Snowball Sampling: survey subjects are selected based on referral from other survey respondents

QUOTA SAMPLING

- The population is first segmented into mutually exclusive sub-groups, just as in stratified sampling.
- Then judgment used to select subjects or units from each segment based on a specified proportion.
- For example, an interviewer may be told to sample 200 females and 300 males between the age of 45 and 60.
- It is this second step which makes the technique one of non-probability sampling.

In quota sampling the selection of the sample is non-random.

For example interviewers might be tempted to interview those who look most helpful. The problem is that these samples may be biased because not everyone gets a chance of selection.

CONVENIENCE SAMPLING

- Sometimes known as **grab** or **opportunity sampling or accidental or haphazard sampling.**
- A type of nonprobability sampling which involves the sample being drawn from that part of the population which is close to hand. That is, readily available and convenient.
- The researcher using such a sample cannot scientifically make generalizations about the total population from this sample because it would not be representative enough.

.

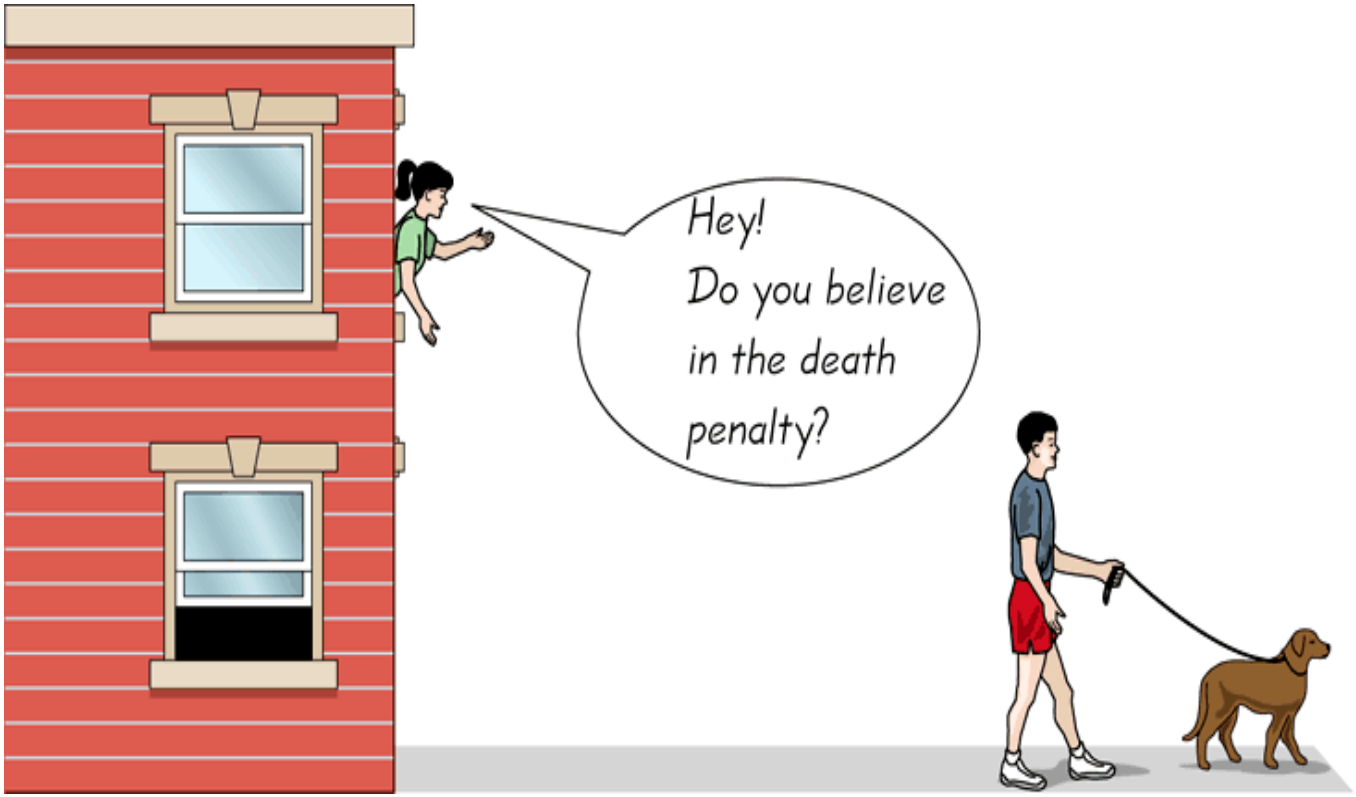
For example, if the interviewer was to conduct a survey at a shopping center early in the morning on a given day, the people that he/she could interview would be limited to those given there at that given time, which would not represent the views of other members of society in such an area, if the survey was to be conducted at different times of day and several times per week.

This type of sampling is most useful for pilot testing.

In social science research, snowball sampling is a similar technique, where existing study subjects are used to recruit more subjects into the sample

CONVENIENCE SAMPLING…….

- Use results that are easy to get



Judgmental sampling or Purposive sampling

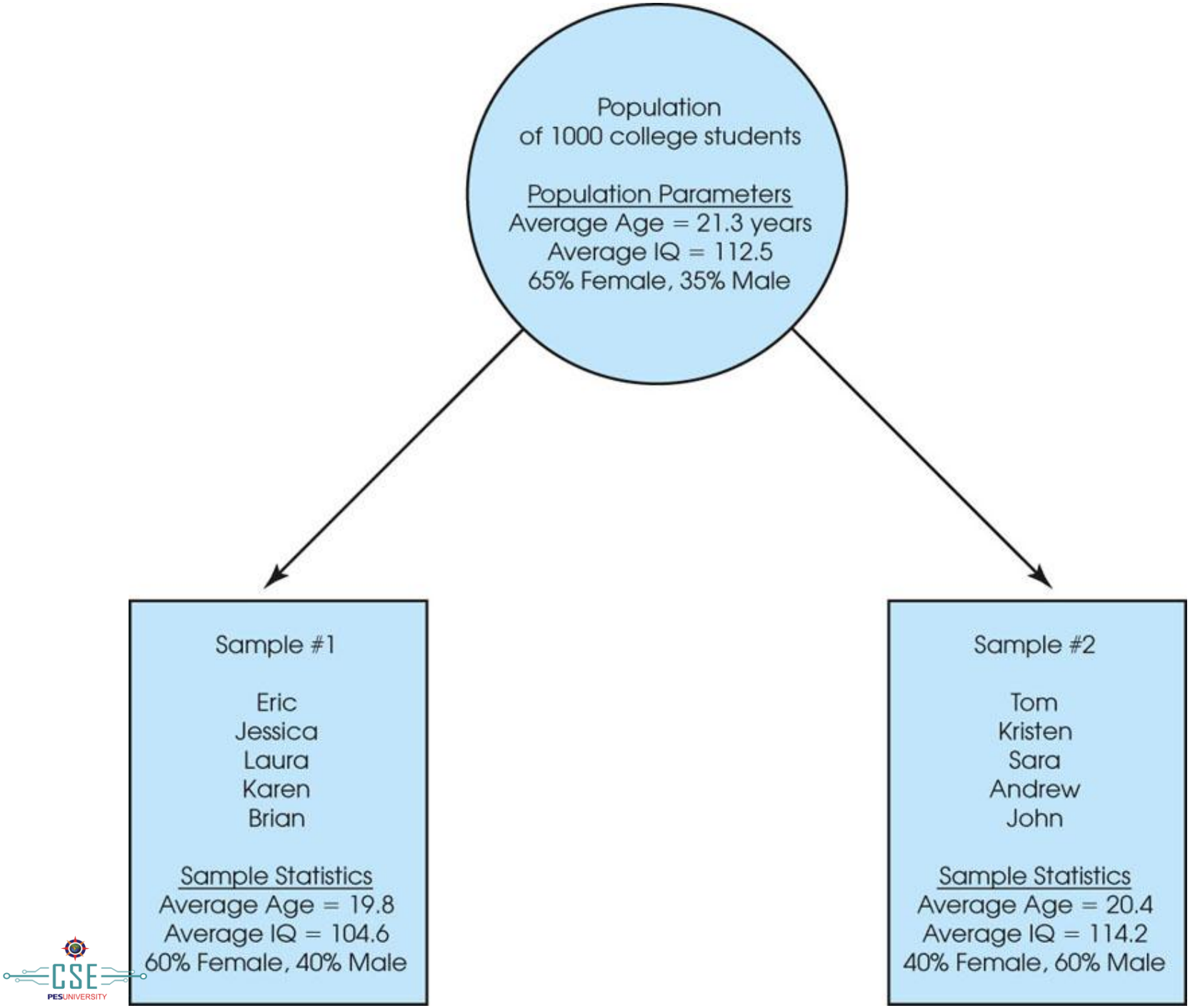
- - The researcher chooses the sample based on who they think would be appropriate for the study. This is used primarily when there is a limited number of people that have expertise in the area being researched

Sample Size?

- The more heterogeneous a population is, the larger the sample needs to be.
- For probability sampling, the larger the sample size, the better.
- With nonprobability samples, not generalizable

Sampling Error

- The discrepancy between a sample statistic and its population parameter is called **sampling error**.
- Defining and measuring sampling error is a large part of inferential statistics.


Population
of 1000 college students
Population Parameters
Average Age = 21.3 years
Average IQ = 112.5
65% Female, 35% Male
Sample #1
Eric
Jessica
Laura
Karen
Brian
Sample Statistics
Average Age = 19.8
Average IQ = 104.6
60% Female, 40% Male
Sample #2
Tom
Kristen
Sara
Andrew
John
Sample Statistics
Average Age = 20.4
Average IQ = 114.2
40% Female, 60% Male

Non-sampling errors are the results of mistakes made in implementing data collection and data processing, such as failure to locate and interview the correct household, is understanding of the questions on the part of either the interviewer or the respondent, and data entry errors.

– Major sources : Sampling Bias, Non-response Bias.

Sampling bias occurs when a chosen sample is not representative of the larger population.

– It occurs due to the sampling technique/method used to perform data collection.

– It can be either selection bias and nonresponsive bias.

Sampling Bias

A sampling method has sampling bias if all subjects in the population are not equally likely to be included in a sample.

Selection bias is a type of sampling bias that occurs when objects are selected from the population in a non-random fashion. With selection bias, the exclusion of certain objects from possible samples affects statistical results based on those samples.

Nonresponse bias is a type of sampling bias that occurs because of the absence of certain objects or subjects from a sample.

For example, some subjects don’t respond to surveys because they refuse, cannot be contacted, or have a lack of interest in the survey content.

Questions???

A physical education professor wants to study the physical fitness levels of students at her university. There are 20,000 students enrolled at the university, and she wants to draw a sample of size 100 to take a physical fitness test. She obtains a list of all

20,000 students, numbered from 1 to 20,000. She uses a computer random number generator to generate 100 random integers between 1 and 20,000 and then invites the 100 students corresponding to those numbers to participate in the study.

Solution

Yes, this is a simple random sample. Note that it is analogous to a lottery in which each student has a ticket and 100 tickets are drawn.

A quality engineer wants to inspect rolls of wallpaper in order to obtain information on the rate at which flaws in the printing are occurring. She decides to draw a sample of 50 rolls of wallpaper from a day's production. Each hour for 5 hours, she takes the 10 most recently produced rolls and counts the number of flaws on each. Is this a simple random sample?

Solution

No. Not every subset of 50 rolls of wallpaper is equally likely to comprise the sample.

To construct a simple random sample, the engineer would need to assign a number to each roll produced during the day and then generate random numbers to determine which rolls comprise the sample.

A construction engineer has just received a shipment of 1000 concrete blocks, each weighing approximately 50 pounds. The blocks have been delivered in a large pile. The engineer wishes to investigate the crushing strength of the blocks by measuring the strengths in a sample of 10 blocks. Which sampling method is suitable?

To draw a simple random sample would require removing blocks from the center and bottom of the pile, which might be quite difficult. For this reason, the engineer might construct a sample simply by taking 10 blocks off the top of the pile.

A quality inspector draws a simple random sample of 40 bolts from a large shipment and measures the length of each. He finds that 34 of them, or 85%, meet a length specification. He concludes that exactly 85% of the bolts in the shipment meet the specification. The inspector's supervisor concludes that the proportion of good bolts is likely to be close to, but not exactly equal to, 85%. Which conclusion is appropriate?

Solution

Because of sampling variation, simple random samples don’t reflect the population perfectly. They are often fairly close, however. It is therefore appropriate to infer that the proportion of good bolts in the lot is likely to be close to the sample proportion, which is 85%. It is not likely that the population proportion is equal to 85%,

Another inspector repeats the study with a different simple random sample of 40 bolts. She finds that 36 of them, or 90%, are good. The first inspector claims that she must have done something wrong, since his results showed that 85%, not 90%, of bolts are good. Is he right?

Solution

No, he is not right. This is sampling variation at work. Two different samples from

the same population will differ from each other and from the population.

A geologist weighs a rock several times on a sensitive scale. Each time, the scale gives a slightly different reading. Under what conditions can these readings be thought of as a simple random sample? What is the population?

.

Solution

If the physical characteristics of the scale remain the same for each weighing, so that the measurements are made under identical conditions, then the readings may be considered to be a simple random sample.

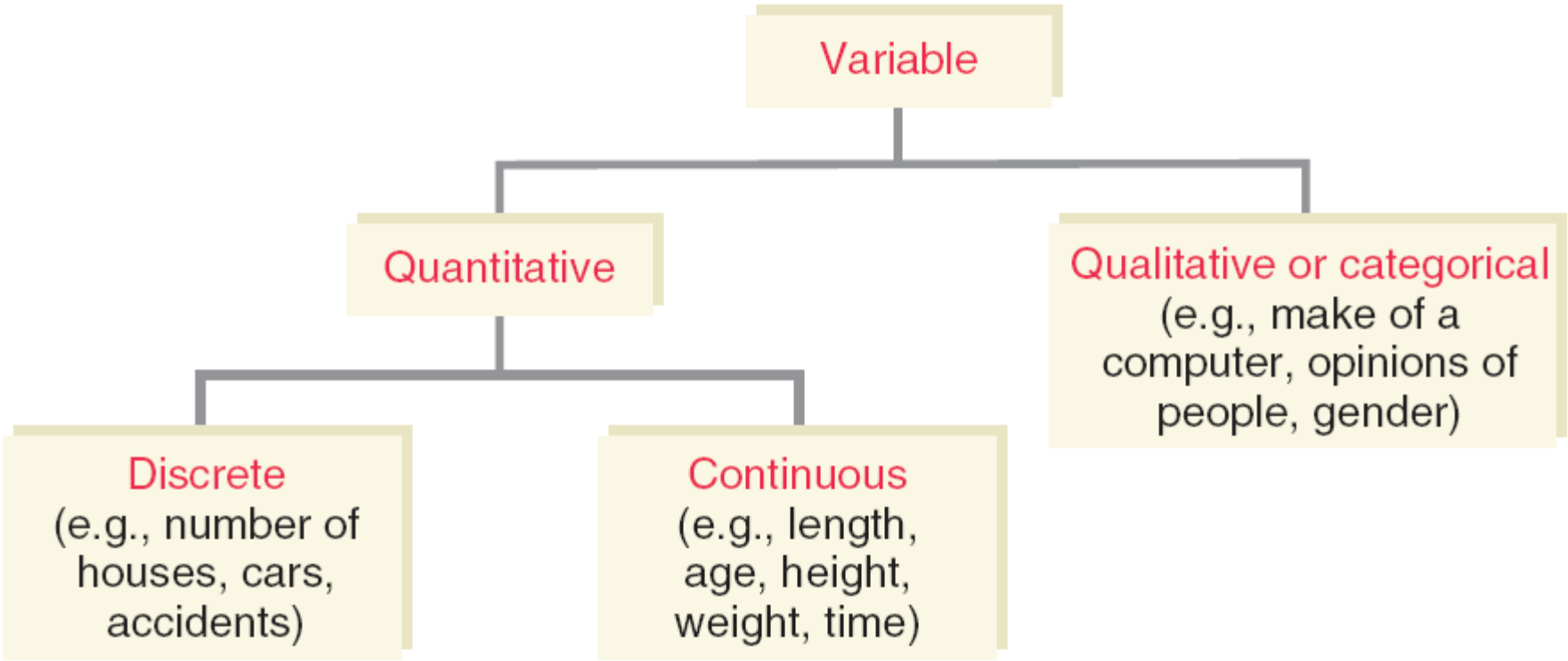
The population is conceptual.

It consists of all the readings that the scale could in principle produce

What sampling method u recommend?

- Determining proportion of undernourished five year olds in a village.
- Investigating nutritional status of preschool children.
- In estimation of immunization coverage in a province, data on seven children aged 12-23 months in 30 clusters are used to determine proportion of fully immunized children in the province.Give reasons why cluster sampling is used in this survey.

Types of Variables



Quantitative Variables

Definition

A variable that can be measured numerically is called a ***quantitative variable***. The data collected on a quantitative variable are called ***quantitative data***.

Quantitative Variables: Discrete

Definition

A variable whose values are countable is called a ***discrete variable***. In other words, a discrete variable can assume only certain values with no intermediate values.

Quantitative Variables: Continuous

Definition

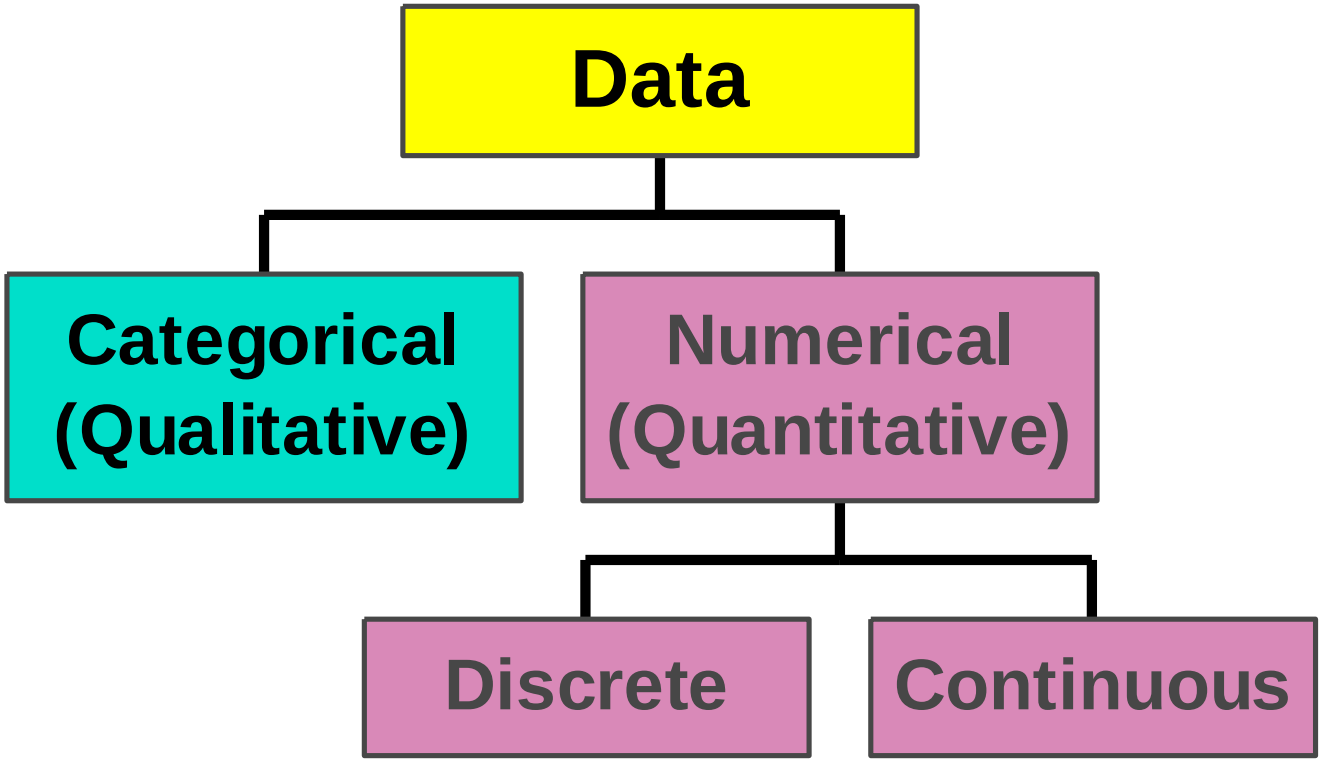
A variable that can assume any numerical value over a certain interval or intervals is called a ***continuous variable***.

Qualitative or Categorical Variables

Definition

A variable that cannot assume a numerical value but can be classified into two or more nonnumeric categories is called a ***qualitative*** or ***categorical variable***. The data collected on such a variable are called ***qualitative data***.

Types of Data



Types of Data

- **Quantitative Data** are measurements that are recorded on a naturally occurring numerical scale.
  - Age
  - GPA
  - Salary
  - Cost of books this semester

Types of Data

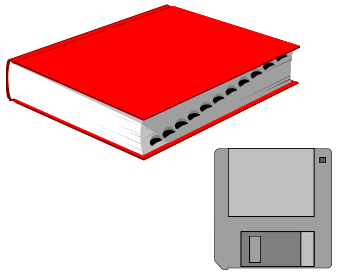
- **Qualitative Data** are measurements that cannot be recorded on a natural numerical scale, but are recorded in categories.
  - Year in school
  - Live on/off campus
  - Major
  - Gender

Data Sources

Primary
Data Collection

Secondary
Data Compilation

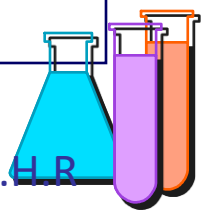
Print or Electronic



Observation

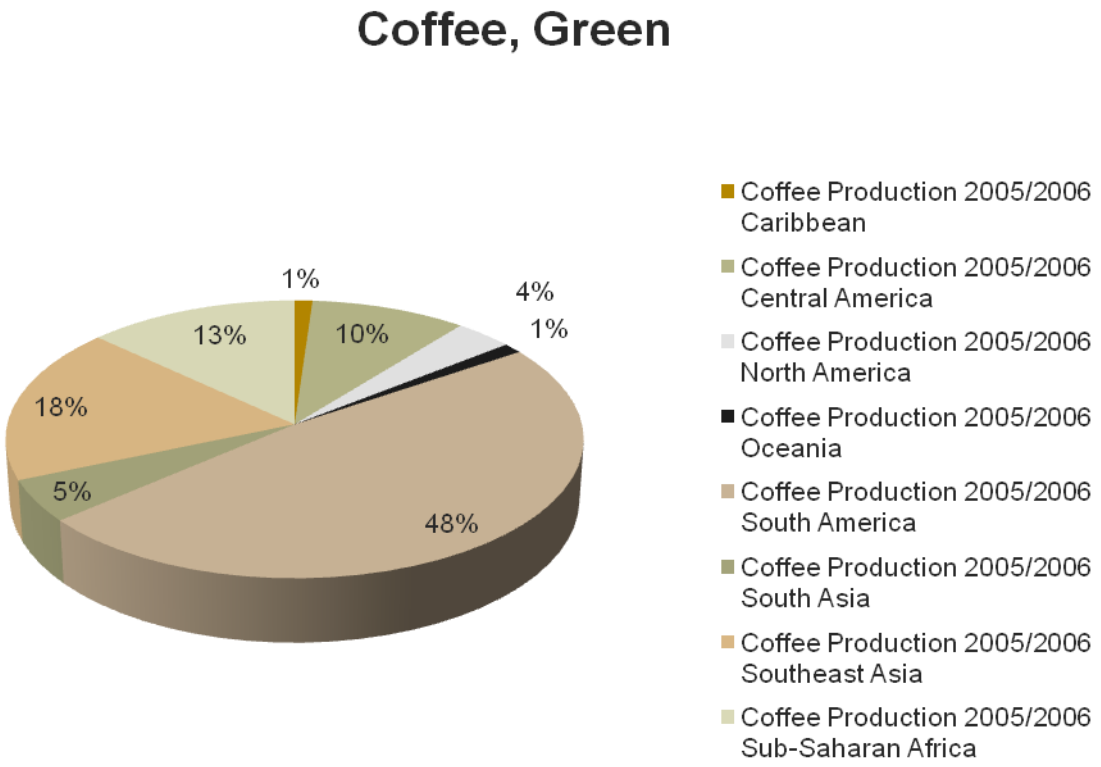
Survey

Experimentation



Collecting Data

- Published Source
- Designed Experiment
- Survey
- Observational Study



SOURCE: United States Department of Agriculture Foreign Agricultural Service

Collecting Data

- Published Source
  - Journal
  - Book
  - Newspaper
  - Magazine
  - (Reliable) Web Site

Collecting Data

- Designed Experiment
  - Strict control over the experiment and the units in the experiment

Collecting Data

- Survey
  - Gallup, Harris and other polls
  - Nielsen

Collecting Data

- Observational Study
  - Observe units in natural settings
  - No control over behavior of units

CROSS-SECTION VS. TIME-SERIES DATA

- Cross-Section Data
- Time-Series Data

Cross-Section Data

Definition

Data collected on different elements at the same point in time or for the same period of time are called ***cross-section data***.

Table Total Revenues for 2010 of Six Companies

Table 1.2 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.

Time-Series Data

Definition

Data collected on the same element for the same variable at different points in time or for different periods of time are called ***time-series data***.

Table Money Recovered from Health Care Fraud Judgments

Table 1.3 Money Recovered from Health Care Fraud Judgments

Year	Money Recovered (billions of dollars)
2006	2.2
2007	1.8
2008	1.0
2009	1.6
2010	2.5

Types of Experiments

There are many types of experiments that can be used to generate data.

In a one-sample experiment, there is only one population of interest, and a single sample is drawn from it.

In a multi-sample experiment, there are two or more populations of interest, and a sample is drawn from each population.

In many multi-sample experiments, the populations are distinguished from one another by the varying of one or more factors that may affect the outcome. Such experiments are called factorial experiments.

Outliers

- Outliers are points that are much larger or smaller than the rest of the sample points.
- Outliers may be data entry errors or they may be points that really are different from the rest.
- Outliers should not be deleted without considerable thought—

sometimes calculations and analyses will be done with and without outliers and then compared.

Types of Statistical Studies

Study design

Experimental studies
(intervention of the researcher, observation of what happens)

Observational studies
(subjects are observed, no action from the researcher)