Chapter 1: Sampling and Data

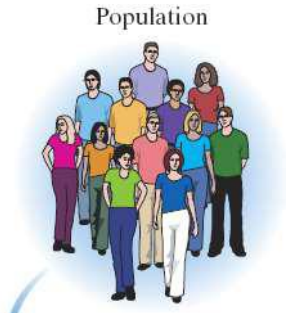
1.1: Definitions of Statistics, Probability, and Key Terms

Statistics - the science of planning studies and experiments, obtaining data, and then organizing, summarizing, presenting, analyzing, interpreting, and drawing conclusions based on the data.

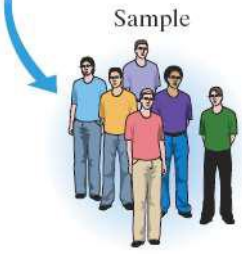
Descriptive Statistics - organizing and summarizing data; by graphing and by numerical values (such as an average).

Inferential Statistics - uses methods that take a result from a sample, extend it to the population, and measure the reliability of the result.

Population Vs Sample



The entire group of individuals (or things or objects) to be studied is called the **population**.



A **sample** is a subset of the population that is being studied.

An **individual** is a person or object that is a member of the population being studied.

Example 1:
A community college employs 86 full-time faculty members. To gain the faculty's opinions about an upcoming building project, the college president wishes to obtain a simple random sample that will consist of 9 faculty members.

- What is the **population**?

- What is the **sample**?

The researchers can use the data collected from the 9 teachers to make inferences about the population of all faculty at the college.

Sampling – selecting a portion (the sample) of the larger population to study for the purpose of gaining information about the population.

Parameter vs Statistic

A **parameter** is a numerical summary of a *population*.

48.2% of **all** students on your campus own a car.

- **48.2%** represents a parameter because it is a numerical summary of a population.

A **statistic** is a numerical summary based on a sample.

In a sample size of 100 students, we find that 46% own a car.

- **46%** represents a statistic because it is a numerical summary of a sample.

Note: A sample statistic is used to estimate a population parameter.

Representative Sample – the idea that the sample must contain the characteristic of the population. One of the main concerns in the field of statistics is how accurately a statistic estimates a parameter.

Example 2: The companies maintaining the wind turbines hire people just to drive around the wind farms and check on any turbine reporting any problems. Let us assume there are 114 such employees nationwide. A random sample of 50 employees in this position showed that their mean salary was $31, 986.

Match the following terms to their description.

Sample	a. The mean salary of the sample of 50 employees, $31, 986.
Population	b. The 114 employees.
Statistic	c. The 50 employees.
Parameter	d. The mean salary of all the 114 employees.

Example 3: Determine whether the underlined value is a **parameter** or a **statistic**.

a) Following the 2006 national midterm election, 18% of the governors of the 50 United States were female.

b) The average score for a class of 55 students taking a statistics midterm exam was 68%.

c) In a national survey of high school students (grades 9 to 12), 25% of respondents reported that someone had offered, sold, or given them an illegal drug at school.

d) In a national survey on substance abuse, 66.4% of respondents who were full-time college students aged 18 to 22 reported using alcohol within the past month.

Variable – a characteristic or measurement that can be determined for each member of the population. (Variables are usually denoted by capital letters such as *X* and *Y*)

Data – the actual values of the variables. They may be numbers, or they may be words.

Two words that often come up in statistics are **mean** and **proportion**.

Mean – or "average"

Proportion – part out of the whole/total. $\frac{\text{part out of the whole}}{\text{total}}$

Note: The words "mean" and "average" are used interchangeably. The technical terms are "arithmetic mean" and "average" and they refer to the **center of the data**.

1.2: Data, Sampling, and Variation in Data and Sampling

Data may come from a population or from a sample.

Qualitative or Categorical data are the result of categorizing or describing attributes or characteristics of a population.
Ex: gender, zip code, color of eyes, hometown.

Quantitative data are the result of counting or measuring attributes of a population.
Ex: height, weight, GPA, time.

Qualitative	Quantitative
Observations that use your senses to observe the results. (Sight, smell, touch, and hearing.)	These results are measurable. Observations with instruments such as rulers, beakers, and thermometers.
The flower is yellow. 	The leaf is 7 cm long. 

Example 4: Determine whether the following variables are **qualitative** or **quantitative**.

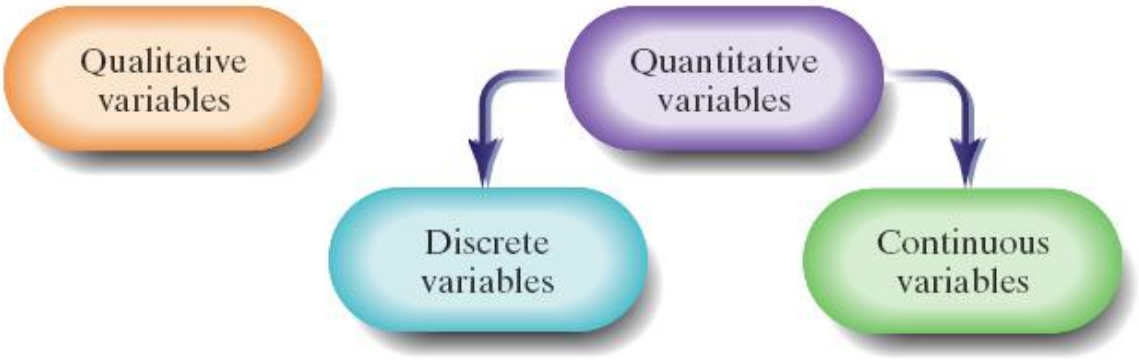
(a) Social Security Number

(b) Temperature

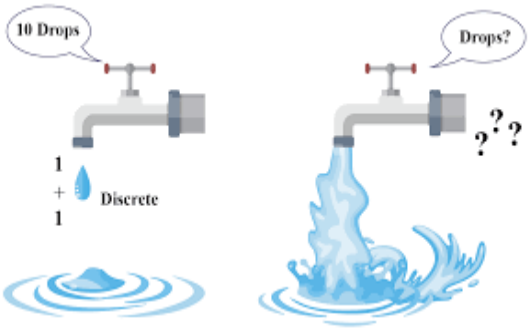
(c) Number of days during the past week that a college student posted on social media

(d) Zip code

Quantitative Data is considered **Discrete** or **Continuous**.



A **Discrete random variable** has either a finite or countable number of values. The values of a discrete random variable can be plotted on a number line with space between each point.



A **Continuous random variable** has infinitely many values. The values of a continuous random variable can be plotted on a number line in an uninterrupted fashion. It will include counting numbers but also fractions, decimals or irrational numbers.

**continuous scale that covers a range of values without gaps, interruptions, or jumps.

Example 5: Determine whether the quantitative variable is discrete or continuous.

(a) Number of text messages someone receives in a week

(b) Volume of water lost each day through a leaky faucet

(c) Length (in seconds) of a country song

(d) Number of sequoia trees in a randomly selected acre of Yosemite National Park

(e) Temperature on a randomly selected day in Denton, Texas

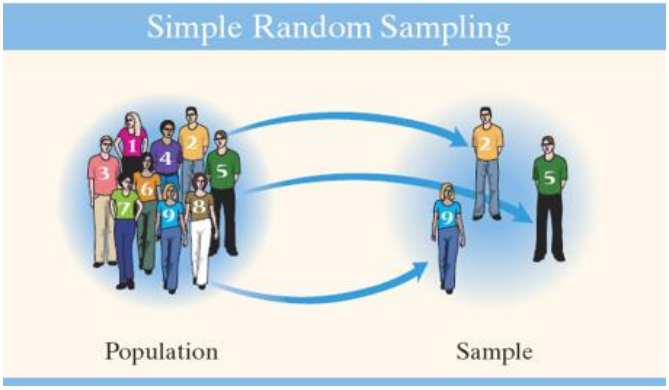
(f) Internet connection speed in kilobytes per second

(g) Points scored in an NCAA basketball game

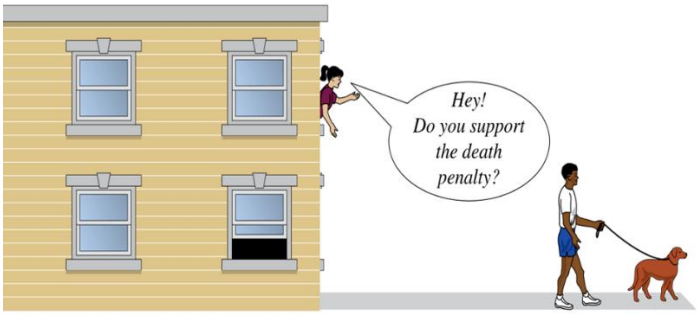
(h) Air pressure in pounds per square inch in an automobile tire

Sampling Techniques (5)

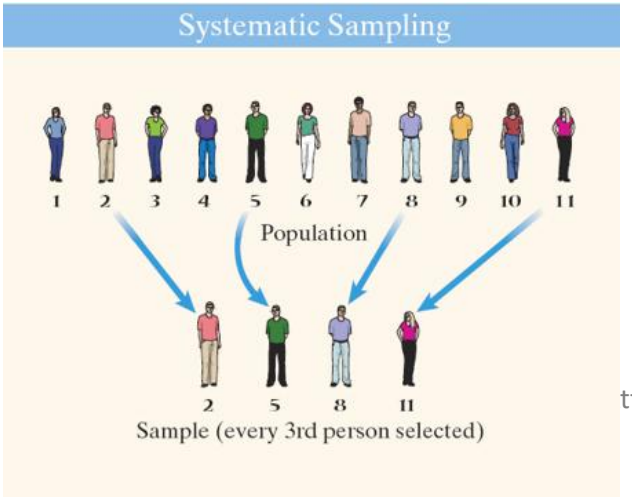
Simple Random Sampling
Randomly selects *n* subjects from a population.
Each subject has an equal chance of being chosen.

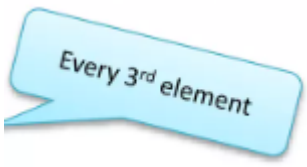


Convenience Sampling:
A sample in which the researcher uses data that was easy to get. This technique is.
NOT A VALID sampling method and will likely result in biased data.



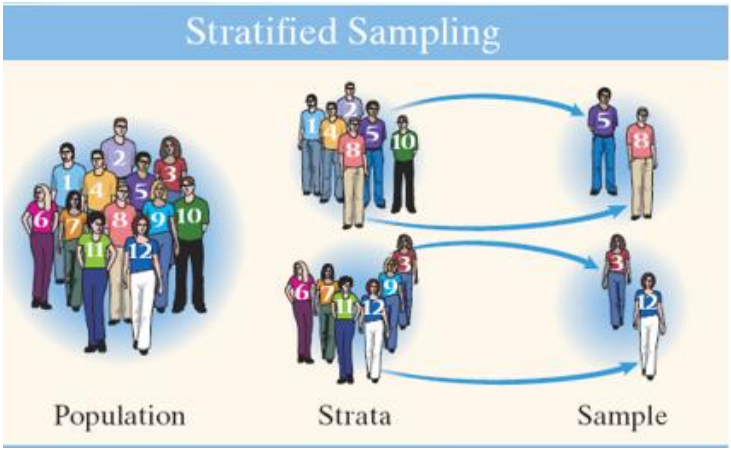
Systematic Sampling:
A sample in which the researcher selects a starting point and then selects every k^{th} element in the population.
Example: Start at the 2nd person, then select every 3rd person.





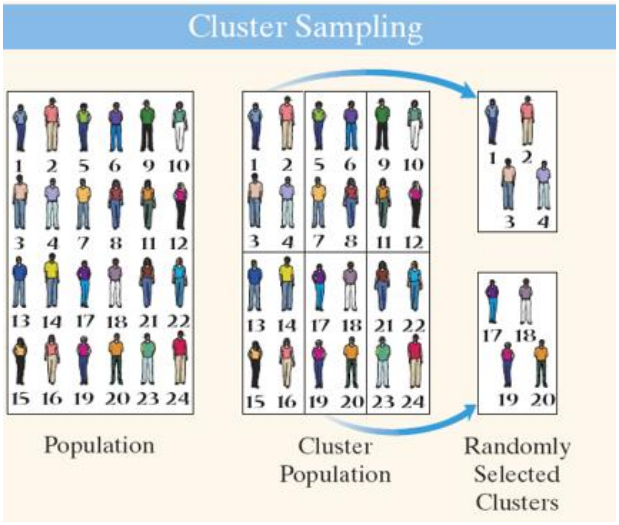
Stratified Sampling:
A sample in which the researcher subdivides the population into at least two different subgroups (or strata), and then draws a sample from each subgroup.

"Some from all"



Cluster Sampling:
A sample in which the researcher first divides the population into sections (or clusters), and then randomly selects all members from some of those clusters.

"All from some"



A statistical study suffers from **bias** if its design or conduct tends to favor certain results. A sample is biased if the sample is not representative of the population.

A **representative sample** is a sample in which the relevant characteristics of the sample members are the same as the characteristics of the population.

Example 6: Identify the sampling method.

A. Instructors teaching research methods are interested in knowing what study techniques their students are utilizing.

   a) Rather than assessing all students, the researcher draws names from a hat.

   b) Rather than assessing all students, the researchers randomly select 10 students from each of the sections to comprise their sample.

   c) Rather than assessing all students, the researcher selects all students from 2 of her sections.

B. To estimate the percentage of defects in a recent manufacturing batch, a quality-control manager at Intel selects every 8th chip that comes off the assembly line starting with the 3rd until she obtains a sample of 140 chips.

C. To determine the prevalence of human growth hormone (HGH) use among high school varsity baseball players, the State Athletic Commission randomly selects 50 high schools. All members of the selected high schools' varsity baseball teams are tested for HGH.

D. A member of Congress wishes to determine her constituency's opinion regarding estate taxes. She divides her constituency into three income classes: low-income household's middle-income households, and upper-income households. She then takes a simple random sample of households from each income class.

E. To identify if an advertising campaign has been effective, a marketing firm conducts a nationwide poll by randomly selecting individuals from a list of known users of the product.

Sampling Bias

A statistical study suffers from **bias** if its design or conduct tends to favor certain results.
A sample is biased if the sample is not representative of the population.

A **representative sample** is a sample in which the relevant characteristics of the sample members are the same as the characteristics of the population.

Three Sources of Bias

Sampling bias means that the technique used to obtain the individuals to be in the sample tends to favor one part of the population over another.

Nonresponse bias exists when individuals selected to be in the sample who do not respond to the survey have different opinions from those who do. (Low response rate)

Response bias exists when the answers on a survey do not reflect the true feelings of the respondent. (Participants may lie)

Example 7: Identify the source of bias.

(a) When a research company polls residents about their voting intentions, new Canadians are under-represented. This is an example of

(b) A recent email poll asked teens whether they used nicotine products. Only 10% of students responded. Of those, only 8% used nicotine products. This is an example of

(c) In a survey, subjects are asked to record how many alcoholic drinks they consume each day. Which type of bias could result?

1.3: Frequency, Frequency Tables, and Levels of Measurement

After collecting data, it is often useful and informative to summarize the data by constructing a frequency distribution.

- **Frequency** - count of how many observations fall into a category.
- **Frequency distribution** - a listing of all categories along with their frequencies. Frequency distributions can be used with both qualitative and quantitative data. The definitions below are used in discussing and constructing frequency distributions for quantitative data.

- **Relative frequency** - the proportion or percentage of the count in a category relative to the total number of items in all categories.
- **Relative frequency distribution** - listing of all categories along with their relative frequencies. *The sum of all the relative frequencies must be 1*.
  (If rounding was used the sum may not be exactly 1, but should be close)
- **Cumulative frequency distribution** - the sum of the frequencies for that class and all previous classes.

Example 8: The table below lists frequencies of the amount of cash each student had in his or her pocket.

Money in $	Frequency
0-4	5
5-9	3
10-14	8
15-19	6
20-24	10
25-29	7

Use the frequency distribution table to construct the relative frequency distribution and the cumulative frequency distribution.

Money in $	Frequency	Rel Freq.	Cum. Freq
0-4	5		
5-9	3		
10-14	8		
15-19	6		
20-24	10		
25-29	7		