

Probability and Statistics

Prepared by Dr. Israa Lewaa

Lecture 1

جامعة الجميع الذكية

STA 201

Introduction to Statistics



جامعة الجميع الذكية
EVERYONE'S SMART UNIVERSITY

- Chapter 1: Introduction To Statistics
- Chapter 2: Descriptive Statistics
- Chapter 3: Probability
- Chapter 4: Discrete Probability Distributions
- Chapter 5: Normal Probability Distributions
- Chapter 6: Confidence Intervals
- Chapter 7: Hypothesis Testing with One Sample
- Chapter 8: Hypothesis Testing with Two Samples
- Chapter 9: Correlation And Regression
- Chapter 10: Chi-Square Tests and The F -Distribution

1- An Overview of statistics

2- Data Classification

3- Data Collection and Experimental Design

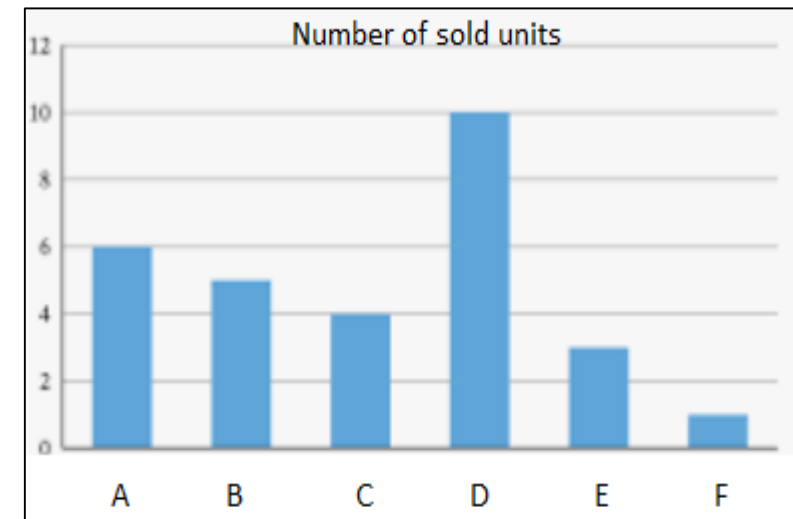
Definition of Statistics

Population and Sample

Parameter and Statistic

Branches of Statistics (Descriptive statistics and Inferential statistics)

Statistics is the science of collecting, organizing, analyzing, and interpreting data in order to make decisions.



A **population** is the collection of *all* outcomes, responses, measurements, or counts that are of interest.

A **sample** is a subset, or part, of a population.

In a recent survey, 1500 adults in the United States were asked if they thought there was solid evidence of global warming. Eight hundred fifty-five of the adults said yes. Identify the population and the sample. Describe the sample data set.

Solution

The population consists of the responses of all adults in the United States, and the sample consists of the responses of the 1500 adults in the United States in the survey. The sample is a subset of the responses of all adults in the United States. The sample data set consists of 855 yes's and 645 no's.

A **parameter** is a numerical description of a *population* characteristic.

A **statistic** is a numerical description of a *sample* characteristic.

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

1. A recent survey of 200 college career centers reported that the average starting salary for petroleum engineering majors is \$83,121. (*Source: National Association of Colleges and Employers*)
2. The 2182 students who accepted admission offers to Northwestern University in 2009 have an average SAT score of 1442. (*Source: Northwestern University*)
3. In a random check of a sample of retail stores, the Food and Drug Administration found that 34% of the stores were not storing fish at the proper temperature.

Solution

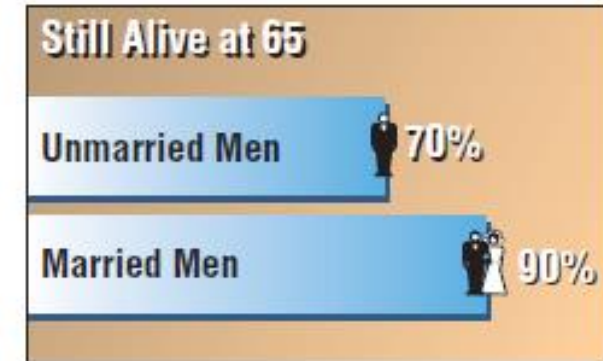
1. Because the average of \$83,121 is based on a subset of the population, it is a sample statistic.
2. Because the SAT score of 1442 is based on all the students who accepted admission offers in 2009, it is a population parameter.
3. Because the percent of 34% is based on a subset of the population, it is a sample statistic.

Descriptive statistics is the branch of statistics that involves the organization, summarization, and display of data.

Inferential statistics is the branch of statistics that involves using a sample to draw conclusions about a population. A basic tool in the study of inferential statistics is probability.

Decide which part of the study represents the descriptive branch of statistics. What conclusions might be drawn from the study using inferential statistics?.

1. A large sample of men, aged 48, was studied for 18 years. For unmarried men, approximately 70% were alive at age 65. For married men, 90% were alive at age 65. (Source: *The Journal of Family Issues*)



2. In a sample of Wall Street analysts, the percentage who incorrectly forecasted high-tech earnings in a recent year was 44%. (Source: *Bloomberg News*)

Solution

1. Descriptive statistics involves statements such as “For unmarried men, approximately 70% were alive at age 65” and “For married men, 90% were alive at 65.” A possible inference drawn from the study is that being married is associated with a longer life for men.
2. The part of this study that represents the descriptive branch of statistics involves the statement “the percentage [of Wall Street analysts] who incorrectly forecasted high-tech earnings in a recent year was 44%.” A possible inference drawn from the study is that the stock market is difficult to forecast, even for professionals.

Types of Data

Level of measurements

Qualitative data consist of attributes, labels, or nonnumerical entries.

Quantitative data consist of numerical measurements or counts.

Classify the following variables according to the type of data:

- a. Salary
- b. Gender
- c. Sales volume of digital music players
- d. Soft drink preference
- e. Temperature
- f. Number of home video screens

- a. Salary is Quantitative data.
- b. Gender is Qualitative data.
- c. Sales volume of digital music players is Quantitative data.
- d. Soft drink preference is Qualitative data.
- e. Temperature is Quantitative data.
- f. Number of home video screens is Quantitative data

1- Nominal

2- Ordinal

3- Interval

4- Ratio



Data at the **nominal level of measurement** are qualitative only. Data at this level are categorized using names, labels, or qualities. No mathematical computations can be made at this level.

Data at the **ordinal level of measurement** are qualitative or quantitative. Data at this level can be arranged in order, or ranked, but differences between data entries are not meaningful.

Data at the **interval level of measurement** can be ordered, and meaningful differences between data entries can be calculated. At the interval level, a zero entry simply represents a position on a scale; the entry is not an inherent zero.

Data at the **ratio level of measurement** are similar to data at the interval level, with the added property that a zero entry is an inherent zero. A ratio of two data values can be formed so that one data value can be meaningfully expressed as a multiple of another.

Type of data

whether they are:

- Qualitative
- Quantitative.

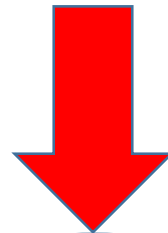
Level of measurements

whether they are:

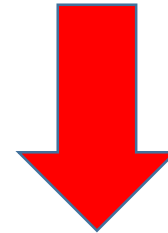
➤ Nominal

➤ Ordinal

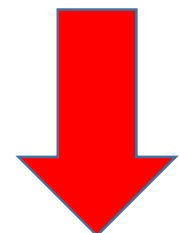
➤ Interval
➤ ratio



If Qualitative
data



If Qualitative or
Quantitative
data



If Quantitative
data

Classify the following variables according to the type of data & level of measurements:

- a. Salary
- b. Gender
- c. Sales volume of digital music players
- d. Soft drink preference
- e. Temperature
- f. Number of home video screens

- a. Salary is Quantitative data and measurement level is Ratio.
- b. Gender is Qualitative data and measurement level is Nominal.
- c. Sales volume of digital music players is Quantitative data and measurement level is Ratio.
- d. Soft drink preference is Qualitative data and measurement level is Nominal.
- e. Temperature is Quantitative data and measurement level is Interval.
- f. Number of home video screens is Quantitative data and measurement level is Ratio.

1- Design of a statistical study

2- Data Collection

3- Experimental Design

4- Sampling Techniques

The goal of every statistical study is to collect data and then use the data to make a decision. Any decision you make using the results of a statistical study is only as good as the process used to obtain the data. If the process is flawed, then the resulting decision is questionable.

Although you may never have to develop a statistical study, it is likely that you will have to interpret the results of one. And before you interpret the results of a study, you should determine whether the results are valid, as well as reliable. In other words, you should be familiar with how to design a statistical study.

We have four main methods for collecting the data:

- 1- Do an observational study
- 2- Perform an experiment
- 3- Use a simulation
- 4- Use a survey

Do an observational study: In an **observational study**, a researcher observes and measures characteristics of interest of part of a population but does not change existing conditions.

For instance, an observational study was performed in which researchers observed and recorded the mouthing behavior on nonfood objects of children up to three years old.

Experiment or a **treatment** is applied to part of a population and responses are observed. Another part of the population may be used as a **control group**, in which no treatment is applied. In many cases, subjects (sometimes called **experimental units**) in the control group are given a **placebo**, which is a harmless, unmedicated treatment, that is made to look like the real treatment.

For instance, an experiment was performed in which diabetics took cinnamon extract daily while a control group took none. After 40 days, the diabetics who took the cinnamon reduced their risk of heart disease while the control group experienced no change.

A **simulation** is the use of a mathematical or physical model to reproduce the conditions of a situation or process. Collecting data often involves the use of computers. Simulations allow you to study situations that are impractical or even dangerous to create in real life, and often they save time and money.

For instance, automobile manufacturers use simulations with dummies to study the effects of crashes on humans. Throughout this course, you will have the opportunity to use applets that simulate statistical processes on a computer.

A **survey** is an investigation of one or more characteristics of a population. Most often, surveys are carried out on *people* by asking them questions. The most common types of surveys are done by interview, mail, or telephone. In designing a survey, it is important to word the questions so that they do not lead to biased results, which are not representative of a population.

For instance, a survey is conducted on a sample of female physicians to determine whether the primary reason for their career choice is financial stability. In designing the survey, it would be acceptable to make a list of reasons and ask each individual in the sample to select her first choice.

Consider the following statistical studies. Which method of data collection would you use to collect data for each study? Explain your reasoning.

1. A study of the effect of changing flight patterns on the number of airplane accidents
2. A study of the effect of eating oatmeal on lowering blood pressure
3. A study of how fourth grade students solve a puzzle
4. A study of U.S. residents' approval rating of the U.S. president

1. Because it is impractical to create this situation, use a simulation.
2. In this study, you want to measure the effect a treatment (eating oatmeal) has on patients. So, you would want to perform an experiment.
3. Because you want to observe and measure certain characteristics of part of a population, you could do an observational study.
4. You could use a survey that asks, “Do you approve of the way the president is handling his job?”

In order to produce meaningful unbiased results, experiments should be carefully designed and executed. It is important to know what steps should be taken to make the results of an experiment valid. Three key elements of a well-designed experiment are *control*, *randomization*, and *replication*.

Because experimental results can be ruined by a variety of factors, being able to **control** these influential factors is important. One such factor is a *confounding variable*. A **confounding variable** occurs when an experimenter cannot tell the difference between the effects of different factors on a variable.

Randomization is a process of randomly assigning subjects to different treatment groups.

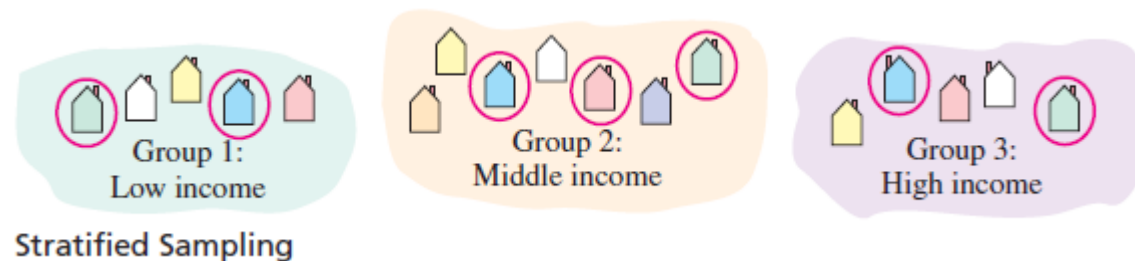
Replication is the repetition of an experiment under the same or similar conditions.

A number of Sampling Techniques which are;

- 1- Random Sample
- 2- Simple Random Sample
- 3- Stratified Sample
- 4- Cluster Sample
- 5- Systematic Sample

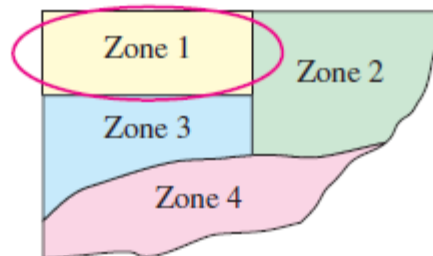
- 1- A **random sample** is one in which every member of the population has an equal chance of being selected.
- 2- A **simple random sample** is a sample in which every possible sample of the same size has the same chance of being selected.

3- Stratified Sample: When it is important for the sample to have members from each segment of the population, you should use a stratified sample.



4- Cluster Sample When the population falls into naturally occurring subgroups, each having similar characteristics, a cluster sample may be the most appropriate.

Zip Code Zones in West Ridge County



Cluster Sampling

5- Systematic Sample A systematic sample is a sample in which each member of the population is assigned a number. The members of the population are ordered in some way, a starting number is randomly selected, and then sample members are selected at regular intervals from the starting number.



Systematic Sampling

1- An Overview of statistics

Definition of Statistics

How to distinguish between a population and a sample

How to distinguish between a parameter and a statistic

How to distinguish between descriptive statistics and inferential statistics

2- Data Classification

How to distinguish between qualitative data and quantitative data

How to classify data with respect to the four levels of measurement: nominal, ordinal, interval, and ratio

3- Data Collection and Experimental Design

How data are collected: by doing an observational study, performing an experiment, using a simulation, or using a survey

How to design an experiment

How to create a sample using random sampling, simple random sampling, stratified sampling, cluster sampling, and systematic sampling