# 1 Statistics and Data

## 1.1 Objectives

- Describe the field of statistics using the terms population, sample, parameter, statistic, sampling, descriptive statistics, and inferential statistics.
- Distinguish between qualitative, quantitative, cross-sectional, and time series data and their applications in business and economics.

## 1.2 What is Statistics?

- The language of data
- The Art and Science of Getting Information from Data
- The Study of Collecting, Analyzing, and Interpreting Data

## 1.3 Data Context

Data are information, which is comprised of facts or characteristics about a subject of interest. You must know the context of the data before referencing or analyzing it.

- *Who are the data about?* Who are the subjects of the data? Subjects can be people, objects, events, etc. This includes any demographic information that describes the subject, such as country, gender, or any other group label.
- *What was measured?* What characteristics are measured about each subject, e.g., demographic information, sales information, opinions, costs, rates, time, etc.
- *Why was the data collected?* What business problem was the data collected to support?
- *When, where, and how was the data collected?* The source of the data can make the difference between insight and nonsense.
    - When were the data collected? Time – day, year, etc.
    - Where were the data collected? From websites, in person interview, etc.
    - How were the data collected? How were subjects selected? For surveys, how were subjects contacted – phone call, email?

### 1.3.1 Example Problem: Data Context

The following table shows the Fortune 500 rankings of America's largest Corporations for 2010. Next to each corporation are its market capitalization (in billions of dollars as of March 26, 2010) and its total return to investors for the year 2009. This data was obtained from Fortune.com and from each corporation's annual reports. [1]

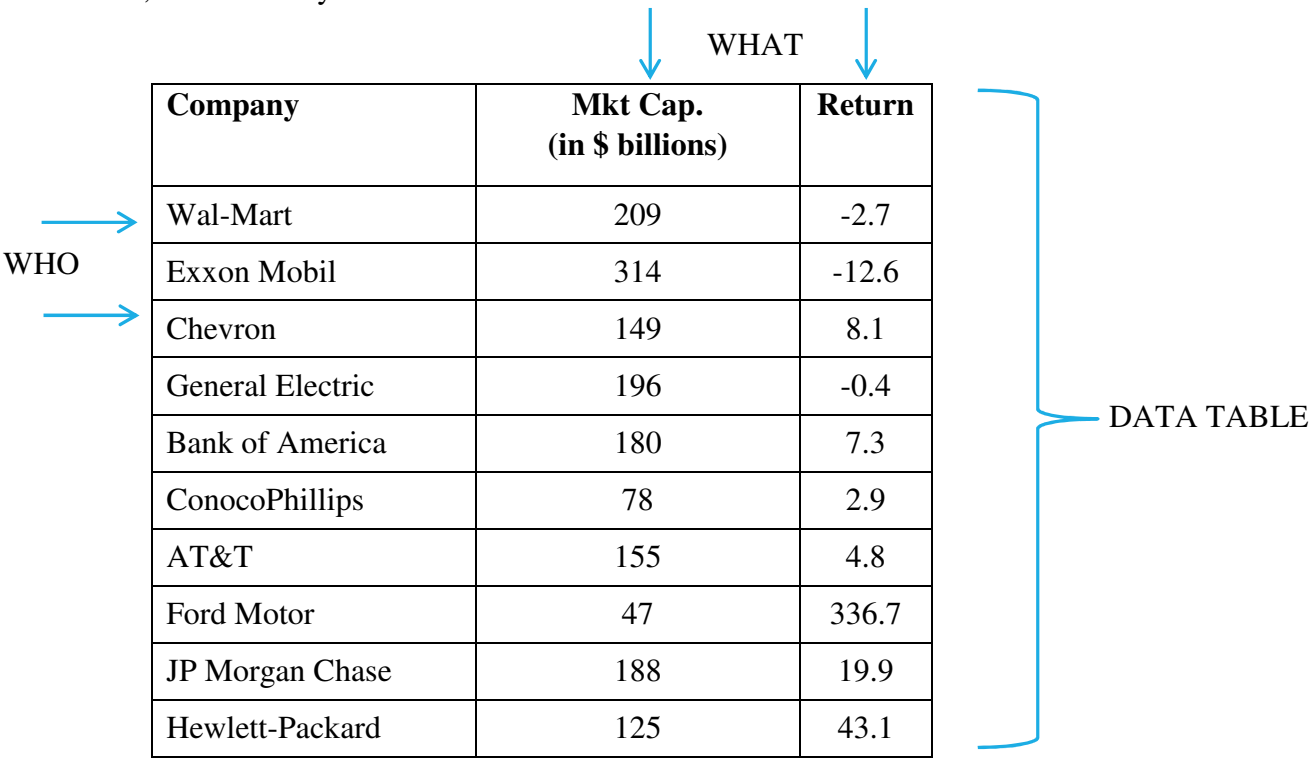
| **Company** | **Mkt Cap. (in $ billions)** | **Return** |
|---|---|---|
| Wal-Mart | 209 | -2.7 |
| Exxon Mobil | 314 | -12.6 |
| Chevron | 149 | 8.1 |
| General Electric | 196 | -0.4 |
| Bank of America | 180 | 7.3 |
| ConocoPhillips | 78 | 2.9 |
| AT&T | 155 | 4.8 |
| Ford Motor | 47 | 336.7 |
| JP Morgan Chase | 188 | 19.9 |
| Hewlett-Packard | 125 | 43.1 |

1. Who are the data about?
   **The data are about large American Fortune 500 corporations.**
2. What was measured?
   **Marketing capital in billions as of March 26, 2010**
   **Return to investors for the year 2009**
3. Why was the data collected?
   **What are some questions you could answer with this data? The problem scenario does not specify.**
4. When, where, and how was the data collected?
   **Obtained from Fortune.com and from annual reports.**
5. What are the variables?
   **Marketing capital – Mkt Cap.**
   **Return to investors – Return**

[1] Chapter 3, Section 4, Problem 26, Page 82

## 1.4 Data Organization

A characteristic observed about people, objects, or events is called a variable because the values often **differ in kind or degree among the various subjects**. The values of the characteristics are organized into a data table with each row representing a subject and each column representing a variable. Data are organized and stored in a form that supports efficient movement or processing, i.e., electronically in databases and data warehouses.




| **Company** | **Mkt Cap. (in $ billions)** | **Return** |
|---|---|---|
| Wal-Mart | 209 | -2.7 |
| Exxon Mobil | 314 | -12.6 |
| Chevron | 149 | 8.1 |
| General Electric | 196 | -0.4 |
| Bank of America | 180 | 7.3 |
| ConocoPhillips | 78 | 2.9 |
| AT&T | 155 | 4.8 |
| Ford Motor | 47 | 336.7 |
| JP Morgan Chase | 188 | 19.9 |
| Hewlett-Packard | 125 | 43.1 |

## 1.5 Types of Data

### 1.5.1 Cross-Sectional

Cross-sectional data is a set of data points collected by observing many subjects (such as individuals, firms, countries, or regions) at the same point of time, or without regard to differences in time.

*Example*

The following table shows the Fortune 500 rankings of America's largest Corporations for 2010. Next to each corporation are its market capitalization (in billions of dollars as of March 26, 2010) and its total return to investors for the year 2009. This data was obtained from Fortune.com and from each corporation's annual reports. [2]

- No time column
- Values represent one time period
- Cannot compare the characteristics over a period of time

| **Company** | **Mkt Cap. (in $ billions)** | **Return** |
|---|---|---|
| Wal-Mart | 209 | -2.7 |
| Exxon Mobil | 314 | -12.6 |
| Chevron | 149 | 8.1 |
| General Electric | 196 | -0.4 |
| Bank of America | 180 | 7.3 |
| ConocoPhillips | 78 | 2.9 |

[2] Chapter 3, Section 4, Problem 26, Page 82

### 1.5.2 Time Series

Time series data is a set of data points indexed (or listed or graphed) in time order. Most commonly, a time series is a sequence taken at successive, equally spaced points in time such as daily, weekly, monthly, quarterly, annually, etc.

*Example*

Elizabeth feels she is ready to invest some of her earnings. She investigates two mutual funds from Janus Capital Group using data her financial planner obtained directly from the company. The following table compares the annual returns (in percentages) of the two mutual funds over the past 10 years. [3]

- Has a time column
- One year represents one time period
- Can compare the characteristics over the one-year periods of time

| Year | Janus Balanced Fund | Janus Overseas Fund |
|---|---|---|
| 2000 | -2.16 | -18.57 |
| 2001 | -5.04 | -23.11 |
| 2002 | -6.56 | -23.89 |
| 2003 | 13.74 | 36.79 |
| 2004 | 8.71 | 18.58 |
| 2005 | 7.75 | 32.39 |
| 2006 | 10.56 | 47.21 |
| 2007 | 10.15 | 27.76 |
| 2008 | -15.22 | -52.75 |
| 2009 | 24.28 | 78.12 |

[3] Chapter 3, Case Studies, Case Study 3.2, Page 102

### 1.5.3 Practice Problems: Types of Data

**Note:** Cross-sectional data and time series data are equally valuable in different types of research.

Classify the following data scenarios as cross-sectional (C) or (T) time series…

_____ The test scores of students in a class

_____ The current average prices of regular gasoline in different states

_____ The sales prices of single-family homes sold last month in California

_____ GDP of the United States from 1990-2010

_____ Daily price of DuPont stock during the first quarter

_____ Quarterly housing starts collected over the last 60 years

_____ Results of market research testing consumer preferences for soda

_____ The 2011 year-end book value per share for all companies listed on the New York Stock Exchange

_____ The stock price for Google at the end of the past four quarters

_____ The price of oil over the past 10 years

_____ The sale prices of townhouses sold last year

_____ Starting salaries of recent business graduates at Penn State University

C, C, C, T, T, T, C, C, T, C, T, C

## 1.6 Variables

There are two types of variables:

**Qualitative** and **quantitative**

There are four scales of measurement, two for each type of variable:

**Nominal**, **ordinal**, **interval**, and **ratio**

Variable types and scales of measurement describe the nature of information assigned to the variables.

A qualitative (Categorical) variable assumes **labels or names** to identify the characteristic. Qualitative variables are described as either nominal or ordinal.

A quantitative variable assumes **numeric values**. Quantitative variables are described as either interval or ratio.

### 1.6.1 Types of Variables

| Qualitative | Quantitative | |
|---|---|---|
| | Continuous | Discrete |
| **Description**<br>Assumes labels or names to identify the characteristics | **Description**<br>Assumes numeric values<br>Infinitely uncountable within an interval, i.e., can take any value within an interval | **Description**<br>Assumes numeric values<br>Countable number of values |
| **Examples**<br>Last name<br>Gender<br>Marital status<br>Religious affiliation<br>Zip code<br>Social security number<br>Opinion yes/no<br>Names of companies<br>Class status<br>Performance rating<br>1 to 5<br>Poor to excellent | **Examples**<br>Time spent studying<br>Amount of money spent on groceries<br>Price of cars<br>Temperature<br>Distance<br>Speed | **Examples**<br>Number of items sold<br>Number of social media sites you use<br>Number of children<br>Number of cars you own<br>Number of bathrooms in your house |

### 1.6.2 Scales of Measurement

| Nominal | Ordinal | Interval | Ratio |
|---|---|---|---|
| **Type**<br>Qualitative | **Type**<br>Qualitative | **Type**<br>Quantitative | **Type**<br>Quantitative |
| **Description**<br>• Categorize a characteristic<br><br>Least sophisticated scale | **Description**<br>• Categorize a characteristic<br>• Rank | **Description**<br>• Categorize a characteristic<br>• Rank<br>• Meaningful difference between values | **Description**<br>• Categorize a characteristic<br>• Rank<br>• Meaningful difference between values<br>• Has a true zero point<br>Most sophisticated scale |
| **Examples**<br>Last Name<br>Gender<br>Marital status<br>Opinion yes/no | **Examples**<br>Performance rating<br>• 1 to 5<br>• Excellent to poor<br>Class status<br>• Freshman, sophomore, junior, senior | **Examples**<br>Temperature<br>• Diff between 20˚ & 30˚ is 10˚<br>• Diff between 60˚ & 70˚ is 10˚<br>Cannot construct ratios: 80˚ is not twice as hot as 40˚ | **Examples**<br>Distance<br>Price of gasoline<br>Price of gold<br>• Diff between \$20 & \$30 is \$10<br>• Diff between \$60 & \$70 is \$10<br>Construct ratios: \$40 is twice as much as \$20 |

## 1.7 Researcher

A researcher **studies a problem using statistical methods** and reports or presents the information obtained.

## 1.8 Consumer of Statistics

A consumer of statistics **reads statistical reports to obtain information** about a problem.

*Note:* Only trust research that is adequately supported by valid statistical methods and theories.

## 1.9 Population vs. Sample

| Population | Sample |
|---|---|
| • A **population** consists of all items or subjects of interest in a statistical problem.<br>• The population is typically too large to study directly.<br>• A **population parameter** describes a characteristic of the population.<br>  o Average salary of all high school teachers<br>  o Median family income for all residents in a particular state | • A **sample** is a representative subset of the population.<br>• A **sample statistic** describes a characteristic of a sample and estimates a population parameter.<br>  o Average salary for a sample of high school teachers<br>  o Median family income for a sample of families from a particular state |

## 1.10 Sampling

- The process of selecting a subset of a population to study.
- It is cost prohibitive and/or infeasible to study the entire population so statisticians study smaller subsets of the population, i.e., samples.
- Used to estimate population parameters.
- Used heavily in manufacturing and service settings to ensure high quality products and services.

## 1.11 Branches of Statistics

Descriptive statistics is the branch of statistics concerned with **numeric (averages, percentages, etc.) and graphical summaries of data**.

Inferential statistics is the branch of statistics concerned with the problem of **estimating population parameters and testing hypotheses about the parameters**.

### 1.11.1 Practice Problems: Sampling

Sampling is appropriate in settings where processes can be standardized. Select the settings below in which sampling would be appropriate…

- ❑ Computer assembly
- ❑ Custom cabinet making
- ❑ Cell phone manufacturing
- ❑ Technical support by phone

State whether the following data scenarios require sampling (Yes) or not (No)…

_____ US Unemployment rate

_____ Average salary for American high school teachers

_____ Total rainfall in Phoenix, Arizona in 2010

_____ The Cleveland Indians' hitting average in 2010

_____ The average SAT score of incoming Freshman at a university

_____ The average life of light bulbs produced by a manufacturer

_____ The percentage of US school teachers who support democrats

_____ The average height of NBA players

_____ The average content of cereal boxes produced by a manufacturer

Computer assembly, Cell phone manufacturing, Technical support by phone

Yes, Yes, No, No, No, Yes, Yes, No, Yes

### 1.11.2 Diagram: The Big Idea of Statistics