統計學

Fall 2025

政治大學統計系余清祥

2025年9月2日

第一章：資料與統計

http://csyue.nccu.edu.tw

Chapter Contents

Introduction

The term **statistics** refers to numerical facts such as averages, medians, percentages, and maximums that help us understand a variety of business and economic situations.

Statistics can also refer to the art and science of collecting, analyzing, presenting, and interpreting data.

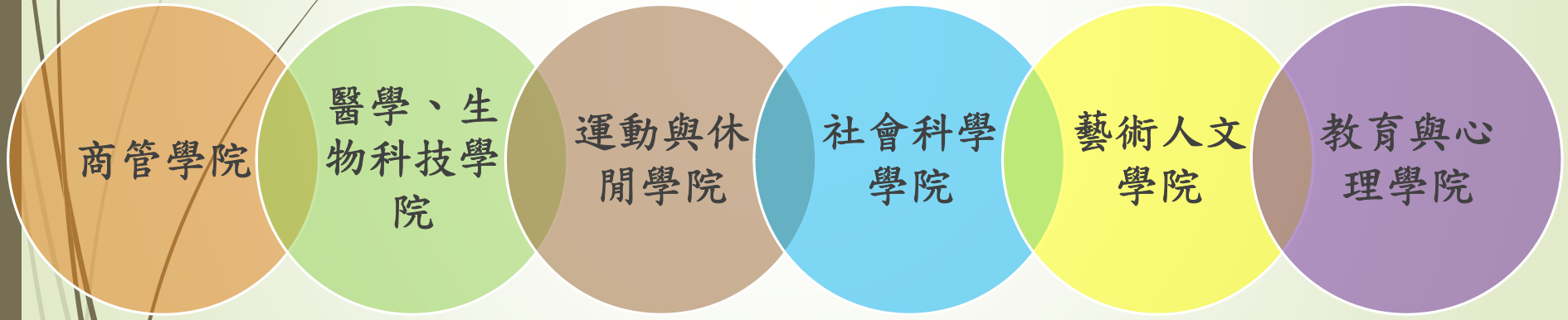
- Section 1.1 begins by introducing applications of statistics in business and economics.
- In Section 1.2, we define the term data, and introduce the concept of a data set and how it is characterized.
- Section 1.3 discusses how data can be obtained from existing sources or through surveys and experimental studies designed to obtain new data.
- Sections 1.4 and 1.5 describe the uses of data in developing descriptive statistics and making statistical inferences.
- The chapter closes with miscellaneous introductory topics in statistics.

1.1 Applications in Business & Economics

- **Accounting**: public accounting firms use statistical sampling procedures when conducting audits for their clients.
- **Economics**: economists use statistical information in making forecasts about the future of the economy or some aspect of it.
- **Finance**: financial advisors use price-earnings ratios and dividend yields to guide their investment advice.
- **Marketing**: electronic point-of-sale scanners at retail checkout counters are used to collect data for a variety of marketing research applications.
- **Production**: a variety of statistical quality control charts are used to monitor the output of a production process.
- **Information Systems**: a variety of statistical information helps administrators assess the performance of computer networks.

統計分析的應用領域

- 統計應用領域按學院分類為六大類：



商管學院

- 商業：<u>行銷領域</u>中應用普遍

→Walmart尿布與啤酒

→Target制定懷孕指數

→T-Mobile店內安裝監視器提升銷量

→Prada裝RFID紀錄衣服選購與試衣

→FB粉絲團與頁面顯示廣告等

圖片來源：http://m.101media.com.tw/content/s8LGfRYZHpQjSlFx9K01ALslCv8vRx
http://www.rfidarena.com/2013/1/3/the-%E2%80%9Csmart-fitting-room%E2%80%9D-concept.aspx

商管學院

圖片來源：http://www.sbs.ox.ac.uk/faculty-research/entrepreneurship-centre/events/fintech-founders-perspective
https://mattermark.com/sizing-the-fintech-opportunity/

■ 財金產業

1.風險控管(Risk Control)

→信用評等、信用卡盜刷、貸款審核與違約預警

2.金融科技(Fintech；Financial Technology)

→第三方支付單位(PayPal、Apple Pay、支付寶等)提供網路收款及付款服務

→網路銀行提供線上匯款、金融交易與投資理財功能(美國銀行、摩根、大通等)

1.2 Data

Data are the facts and figures collected, analyzed, and summarized for presentation and interpretation.

Elements are the entities on which data are collected.

A **variable** is a characteristic of interest for the elements.

A **data set** consists of all the data collected for a particular study.

- The total number of data values in a complete data set is the number of elements multiplied by the number of variables.

An **observation** is the set of **measurements** obtained for a particular element.

- A data set with *n* elements contains *n* observations.

1.2 Data Set for 60 Nations in the World Trade Organization

DATAfile: *Nations*

Shown here are the first five nations out of 60 (rows) of a table including several variables (columns.)

- Observation #1 (Armenia) contains the measurements: Member, 4,267, B+, and Stable.
- Observation #2 (Australia) contains the measurements: Member, 51,812, AAA, and Negative.
- And so on.

Nation	WTO Status	Per Capita GDP (S)	Fitch Rating	Fitch Outlook
Armenia	Member	4,267	B+	Stable
Australia	Member	51,812	AAA	Negative
Austria	Member	48,328	AA+	Stable
Azerbaijan	Observer	4,214	BB+	Stable
Bahrain	Member	28,608	B+	Stable

2020 U.S. Census Questionnaire

Person 1

5. Please provide information for each person living here. If there is someone living here who pays the rent or owns this residence, start by listing him or her as Person 1. If the owner or the person who pays the rent does not live here, start by listing any adult living here as Person 1.

What is Person 1's name? *Print name below.*

First Name MI

Last Name(s)

6. What is Person 1's sex? *Mark X ONE box.*

☐ Male ☐ Female

7. What is Person 1's age and what is Person 1's date of birth? *For babies less than 1 year old, do not write the age in months. Write 0 as the age.*

Print numbers in boxes.

Age on April 1, 2020 Month Day Year of birth

years

9. What is Person 1's race?

*Mark X one or more boxes **AND** print origins.*

☐ White – *Print, for example, German, Irish, English, Italian, Lebanese, Egyptian, etc.*

☐ Black or African Am. – *Print, for example, African American, Jamaican, Haitian, Nigerian, Ethiopian, Somali, etc.*

☐ American Indian or Alaska Native – *Print name of enrolled or principal tribe(s), for example, Navajo Nation, Blackfeet Tribe, Mayan, Aztec, Native Village of Barrow Inupiat Traditional Government, Nome Eskimo Community, etc.*

☐ Chinese
☐ Filipino
☐ Asian Indian
☐ Other Asian – *Print, for example, Pakistani, Cambodian, Hmong, etc.*

☐ Vietnamese
☐ Korean
☐ Japanese

☐ Native Hawaiian
☐ Samoan
☐ Chamorro
☐ Other Pacific Islander – *Print, for example, Tongan, Fijian, Marshallese, etc.*

https://www.census.gov/programs-surveys/decennial-census/technical-documentation/questionnaires.2020_Census.html

2020年人口普查問卷(部分)

一、住宅狀況【若本宅有2戶以上住戶，僅其中一戶須填寫】

[1]本宅居住及住宅使用情形

住宅指供家庭居住為目的之房屋，有單獨通往宅外的通道及住宅設備

經常居住指已實際居住或預期居住6個月以上

□1 有人經常居住的住宅，目前使用情形爲

□(1)住家專用　□(3)營業或服務業用

□(2)兼工業用　□(4)兼農業用

本住宅有 [ ][ ]房 [ ]廳 衛浴 [ ]：[ ]坐

小數點

房間數不含廚房、衛浴間、儲藏室、車庫及走廊等

只有廁所或洗澡、淋浴設施者以0.5套計

□2 無人經常居住的住宅，目前使用情形爲

□(1)偶爾自住

□(2)自住以外用途(如辦公室、倉庫等)

□(3)目前沒有使用

□3 有人經常居住的其他房屋【指住宅以外的房屋，如廠房、辦公大樓、旅館、宿舍(單身、學生)、醫院等】

□4 有人經常居住的其他處所【非屬房屋，如帳篷、路邊、地下道等】

[8]您是否爲本戶主要家計負責人　□1 是　□2 否　非親屬戶免填本問項；同一戶應只有1人勾選「1是」

[9]您是否因生病、受傷、衰老而有右列限制或困難且需他人幫忙及連續預期達6個月以上　本問項1-9可複選

□1 吃飯　□6 在室內外走動

□2 上下床　□7 備餐(煮飯)

□3 穿脫衣服　□8 洗(含晾曬)衣服

□4 上廁所　□9 處理家務(打掃、擦桌等清潔工作)

□5 洗澡　□10 不具上列各項障礙

未滿5歲(104年11月8日及以後出生)以下問項免填

[10]使用語言情形

	(1)國語	(2)閩南語	(3)客語	(4)原住民族語	(5)臺灣手語	(6)其他語言	(7)不知或無
1 兒時最早學會	□	□	□	□	□	□	□
2 現在主要使用	□	□	□	□	□	□	□
現在次要使用	□	□	□	□	□	□	□
3 父親最常使用	□	□	□	□	□	□	□
母親最常使用	□	□	□	□	□	□	□

[11]您5年前(104年11月8日)的居住地點

□1 同現住處所

□2 同現住鄉(鎮市區)不同處所

□3 其他鄉(鎮市區)→ [ ][ ]縣市 [ ][ ][ ][ ]鄉鎮市區

□4 大陸地區(含港澳)

□5 其他 _____

2000年人口普查編碼簿

項目	欄位名稱		欄位代號	資料型態	欄位長度	起	迄
1		檔案識別碼	F001	文字	1	1	1
3		FILLER	T001	文字	8	2	9
2	統一編號	卡號	C001	文字	1	10	10
4		縣市代號	T021	文字	2	11	12
5		鄉鎮市區代號	T022	文字	2	13	14
6		村里代號	T023	文字	3	15	17
7		普查區號	T024	文字	3	18	20
8		宅號	T025	文字	3	21	23
9		戶號	T026	文字	3	24	26
10		鄰號	T027	文字	3	27	29
11		人口序號	A004	數字	4	30	33
12		國籍代碼	P001	數字	3	34	36
13		性別	A010	數字	1	37	37
14		FILLER	FILLER	文字	7	38	44
15		年齡	A020	數字	3	45	47
16		FILLER	FILLER	文字	7	48	54
17		經常居住	A041	數字	1	55	55
18		FILLER	FILLER	數字	1	56	56
19		與戶長關係	A050	數字	2	57	58
20		婚姻狀況	A060	數字	1	59	59

1.2 Nominal & Ordinal Scales of Measurement

Nominal scale: the data are labels or names used to identify an attribute of the element.

Example: the WTO Status variable in the Nations table has a nominal scale of measurement because it uses the data "Member" and "Observer" are labels used to identify the status category for a nation.

Note: a numeric code may also be used. For example, we could use the label "1" to identify a "Member" status category and "2" to identify an "Observer" status category.

Ordinal scale: the data have the properties of nominal data, and the order or rank of the data is meaningful.

Example: the Fitch Rating variable in the Nations table has ordinal data, which range from AAA to F, are rank-ordered by credit rating.

Note: ordinal data can also be recorded by a numerical code, such as a student's class rank in school.

1.2 Interval and Ratio Scales of Measurement

Interval scale: the data have the properties of ordinal data, and the interval between observations is expressed in terms of a fixed unit of measure. Data with an interval scale of measurement are always numerical.

Example: College admission SAT scores have an interval scale of measurement.

Note: the difference between the values of a variable with an interval scale of measurement are always meaningful.

Ratio scale: the data have the properties of interval data, and the ratio of two values is meaningful.

Example: variables such as distance, height, weight, and time use the ratio scale of measurement.

Note: the ratio scale of measurement requires that a zero value be included to indicate that nothing exists for the variable at the zero point.

1.2 Categorical and Quantitative Data

Data can be further classified as **categorical** or **quantitative**.

The statistical analysis that is appropriate depends on whether the data for the variable are categorical or quantitative.

Categorical Data

A **categorical variable** is a variable with categorical data.

Statistical analyses are rather limited.

We can summarize categorical data by:

- counting the number of observations in each category
- computing the proportion of the observations in each category

Quantitative Data

- A **quantitative variable** is a variable with quantitative data.
- Quantitative data indicate *how many* or *how much* and are always numerical.
- Ordinary arithmetic operations are meaningful for quantitative data.
- More statistical analysis methods are available if the data are quantitative.

1.2 Cross-Sectional & Time Series Data

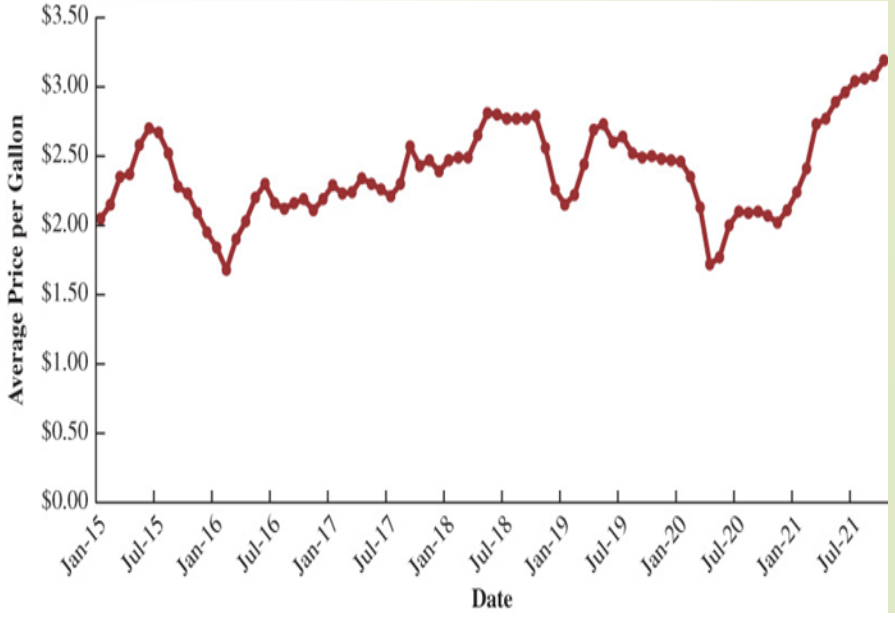
Cross-sectional data are collected at the same or approximately the same point in time.

Example: The data in the previously shown Nations data file are cross-sectional and they describe the five variables for the 60 WTO nations

Time series data are instead collected over several time periods.

Graphs of time series data, such as the U.S. average price per gallon of regular gasoline between 2015 and 2021 shown to the right, help understand:

- what happened in the past
- identify any trends over time
- project future levels for the time series



1.3 Existing Data Sources

In some cases, data needed for a particular application already exist.

The most important categories of existing data sources and related examples are:

- Internal company records – employee records, production records, inventory records, sales records, credit records, and customer profiles
- Business database services – Dun & Bradstreet, Bloomberg, and Dow Jones & Co.
- Government agencies – Census Bureau, Federal Reserve Board, Office of Management & Budget, Department of Commerce, Bureau of Labor Statistics, and DATA.gov
- Industry associations – U.S. Travel Association
- Special-interest organizations – Graduate Management Admission Council (GMAT)
- Internet – Google, Yahoo, Twitter, etc.

1.3 Statistical Studies

Observational Study

In *an observational study*, no attempt is made to control or influence the variables of interest.

Examples:

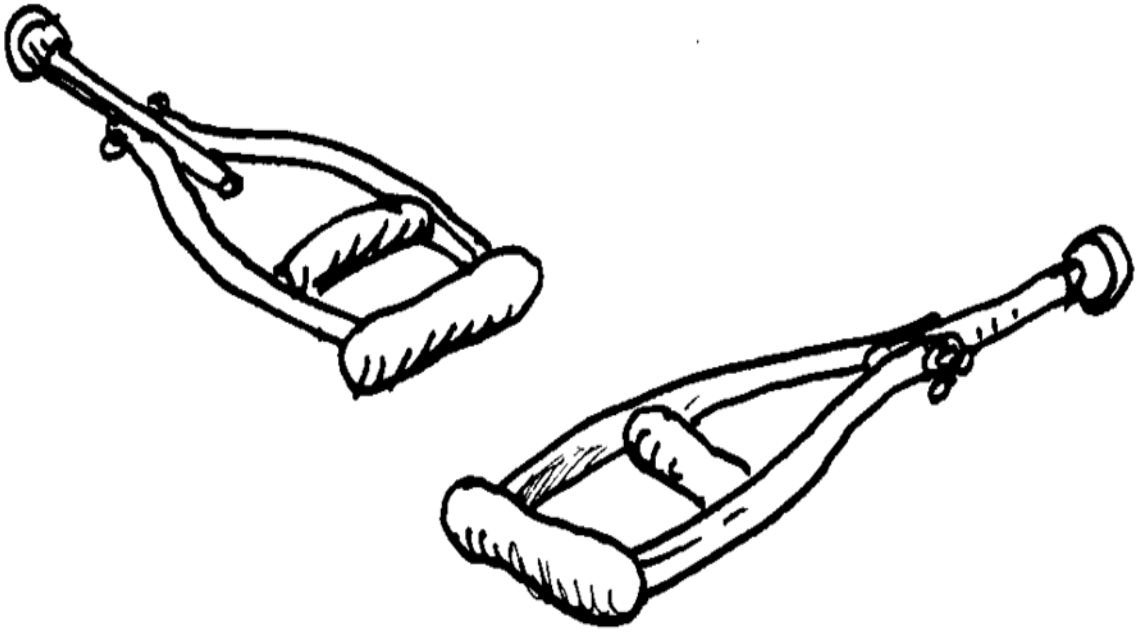
- Recording data on demographics and shopping habits of a random group of Walmart customers.
- Recording CEO gender and return on equity (ROE) for a sample of Fortune 500 companies to investigate the relationship between CEO gender and company performance.
- Surveys and public opinion polls.

Experiment

- An *experiment* is conducted under controlled conditions, and its data can provide more information compared to data obtained from existing sources or observational studies.
- In experiments, first, a variable of interest is identified, and then one or more other variables are identified and controlled. The collected data are then analyzed to investigate how they influence the variable of interest.
- Example – a pharmaceutical company administers different dosages of a new drug to groups of randomly selected individuals and monitors how they are affected by it.

世界規模最大的醫學實驗（沙克疫苗）

A MORE POSITIVE EXAMPLE IS THE *SALK POLIO VACCINE.* IN 1954, VACCINE TRIALS WERE PERFORMED ON SOME 400,000 CHILDREN, WITH STRICT CONTROLS TO ELIMINATE BIASED RESULTS. GOOD STATISTICAL ANALYSIS OF THE RESULTS FIRMLY ESTABLISHED THE VACCINE'S EFFECTIVENESS, AND TODAY POLIO IS ALMOST UNKNOWN.



第一階段
（人體藥理學）

第二階段
（治療探索）

第三階段
（治療確認）

第四階段
（治療開發）

臨床前期 動物試驗	健康受試驗者或特定疾病族群	以病人進行對某藥原有的療效和安全性進行評估	1.確認第二階段藥品目標藥原有及安全性的有效證據 2.提供核准藥品上市前依據	藥證TFDA核准上市	藥證TFDA後各類型試驗
1.評估藥物療效、觀察毒性 2.進行藥物效力學（pharmacodynamics：PD） 3.藥物動力學（pharmacokinetics：PK）	收案數：15-30人	收案數：少數~100人	收案數：100~1000人		

3-5年　10年　1-2年

Center of Clinical Trial Research . Cathay General Hospital

1.3 Considerations on Data Acquisition

Time Requirement

- Searching for information can be time-consuming.
- Information may no longer be useful by the time it is available.

Cost of Acquisition

- Organizations often charge for information even when it is not their primary business activity.

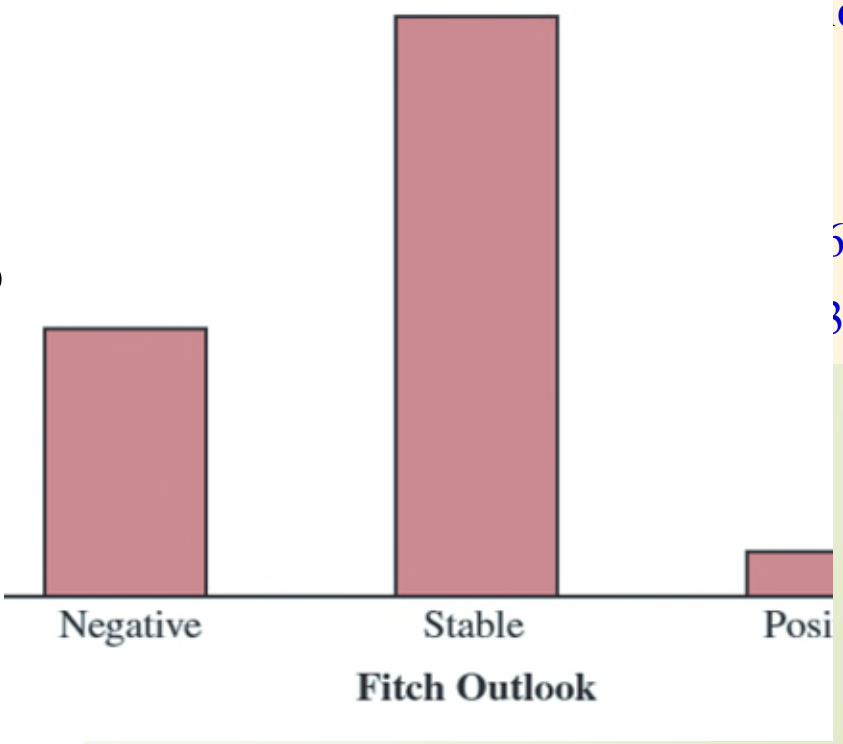
Data Errors

- Using any data that happen to be available.
- Data acquired with little care can lead to misleading information.

1.4 Descriptive Statistics

Most of the statistical information in publications consists of data that are summarized and presented in a form that is easy to understand.
Such summaries of data, which may b tabular, graphical, or numerical, are referred to as **descriptive statistics**.
Examples:

- The variable Fitch Outlook in the WTO Nations table can be summarized as a table or bar chart.
- Numerical descriptive statistics such as the mean (or average.)



1.5 Statistical Inference

When the collection of information about a large group of elements (individuals, companies, voters, households, products, & customers) is not feasible because of time, cost & other considerations, data can be collected from only a small portion of the group.

Formally, we use the following definitions of the larger and smaller groups of elements:

Population: the set of all elements of interest in a particular study

Sample: a subset of the population

We also formally define the following statistical processes:

Statistical inference: the process of using data obtained from a sample to make estimates and test hypotheses about the characteristics of a population

Census: the process of collecting data for the entire population

Sample survey: a process of collecting data for a sample

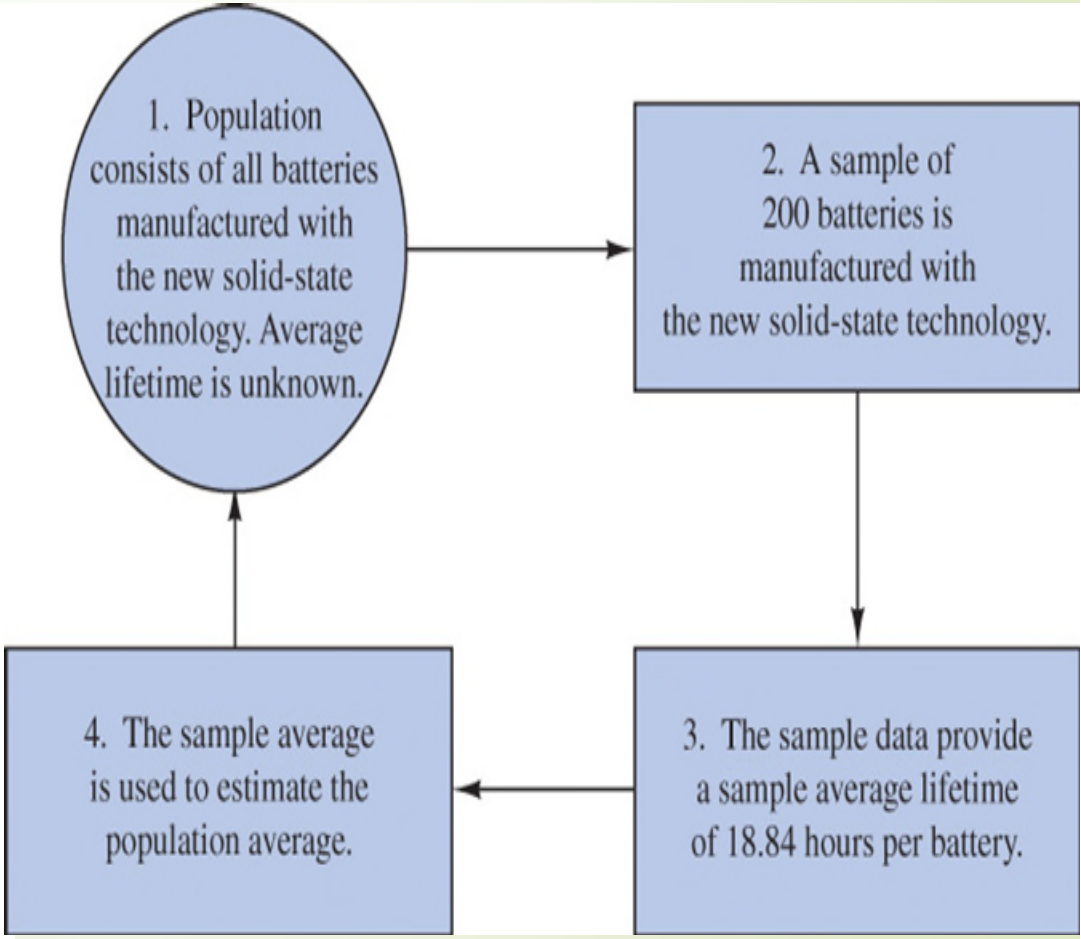
1.5 The Rogers Industries Example

DATAfile: *Rogers*

Rogers Industries has developed a new solid-state lithium battery that should last longer and be safer to use.

Researchers want to evaluate the advantages of the new battery using statistical inference.

The sample average battery life of 18.84 hours can be used as an estimate (inference) of the population average.



1.6 Analytics

Analytics is the scientific process of transforming data into insight for making better decisions.

Analytics is now generally thought to comprise three broad categories of techniques:

Descriptive analytics describes what has happened in the past.

- Examples – data queries, reports, descriptive statistics, data visualization, data dashboards, and basic what-if spreadsheet models

Predictive analytics uses models constructed from past data to predict the future or to assess the impact of one variable on another

- Examples – linear regression, time series analysis, forecasting models, and simulation.

Prescriptive analytics yield the best course of action.

- Examples – optimization models and decision analysis

1.7 Big Data and Data Mining

Big data are larger and more complex data sets that cannot be managed, processed, or analyzed with commonly available software in a reasonable amount of time.

Analysts often define *big data by* referring to the **Three V's**:

- *Volume*: the amount of available data; *velocity*: the speed at which data is collected and processed; *variety*: different data types.

Data warehousing is the process of capturing, storing, and maintaining data.

- Computing power and data collection tools have reached the point where it is now feasible to store and retrieve extremely large quantities of data in seconds.

Data mining methods help develop useful decision-making information from large databases.

- Analysts use a combination of automated procedures to "mine the data" and convert it into useful information.
- The most effective data mining procedures are multiple regression, logistic regression, and machine learning, which discover relationships in the data and predict future outcomes.

1.8 Computers and Statistical Analysis

Statisticians use computer software to perform tedious and time-consuming statistical computations and analyses.

End-of-chapter appendixes cover the step-by-step procedures for using Microsoft Excel and the statistical package JMP to implement the statistical techniques presented in each chapter.

Big data requires special data manipulation and analysis tools such as:

- Hadoop – open-source software for the distributed processing of large data sets
- R and Python – open-source programming languages
- SAS and SPSS – commercially available packages

1.9 Ethical Guidelines for Statistical Practice

In a statistical study, unethical behavior can take a variety of forms including:

- Improper sampling
- Inappropriate analysis of the data
- Development of misleading graphs
- Use of inappropriate summary statistics
- Biased interpretation of the statistical results

One should strive to be fair, thorough, objective, and neutral in collecting, analyzing, and presenting data.

Consumers of statistics should also be aware of the possibility of unethical behavior by others.

The American Statistical Association developed a report on "Ethical Guidelines for Statistical Practice" that contains 67 guidelines organized into 8 topic areas of professionalism and responsibilities that address the major stakeholders of statistical analysis and research.

Note: Ethical and Responsible Use of Data and Predictive Models_SOA

Summary

- Statistics is the art and science of collecting, analyzing, presenting, and interpreting data.
- Data consists of the facts and figures that are collected and analyzed.
- The four scales of measurement used to obtain data on a particular variable include nominal, ordinal, interval, and ratio.
- For purposes of statistical analysis, data can be classified as categorical or quantitative.
  - Categorical data use labels or names to identify an attribute of each element and use either the nominal or ordinal scale of measurement.
  - Quantitative data are numeric values that indicate how much or how many and use either the interval or ratio scale of measurement.
- Descriptive statistics are the tabular, graphical, and numerical methods used to summarize data. The process of statistical inference uses data obtained from a sample to make estimates or test hypotheses about the characteristics of a population.
- The last sections of the chapter introduced analytics, big data, data mining, the role of computers in statistical analysis, and a summary of ethical guidelines for statistical practice.