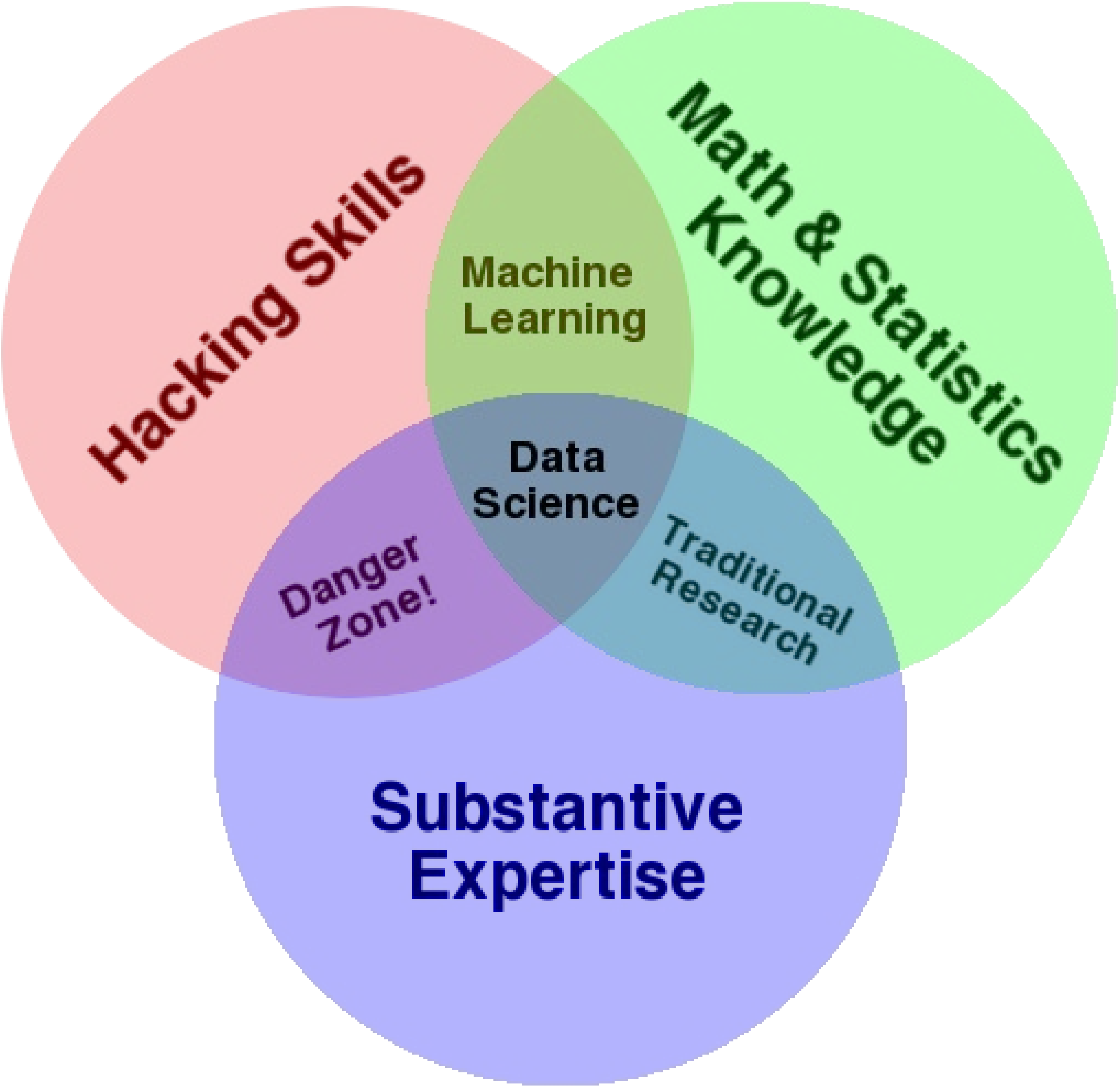
Chapter 1 - Data and Statistics

- Statistics
- Applications in Business and Economics
- Data
- Data Sources
- Descriptive Statistics
- Statistical Inference
- Big Data and Data Mining
- Computers and Statistical Analysis
- Ethical Guidelines for Statistical Practice

Learning Objectives

1. Understand the meaning of the terms elements, variables, and observations as they are used in statistics.
2. Understand the four scales of measurement
3. Obtain an understanding of the difference between categorical, quantitative, cross-sectional, and time-series data.



The Data Science VennDiagram by Drew Conway

每組蒐集一份資料

1. 分析該資料中的元素為何，有哪些變數，有多少觀察值
2. 變數屬於何種尺度
3. 資料屬於質性或量化的資料
4. 資料屬於橫斷面或縱斷面資料
5. 資料來源屬於觀察的或實驗的
6. 若你們要蒐集這份報告中的資料，你們會考慮甚麼?

評分標準

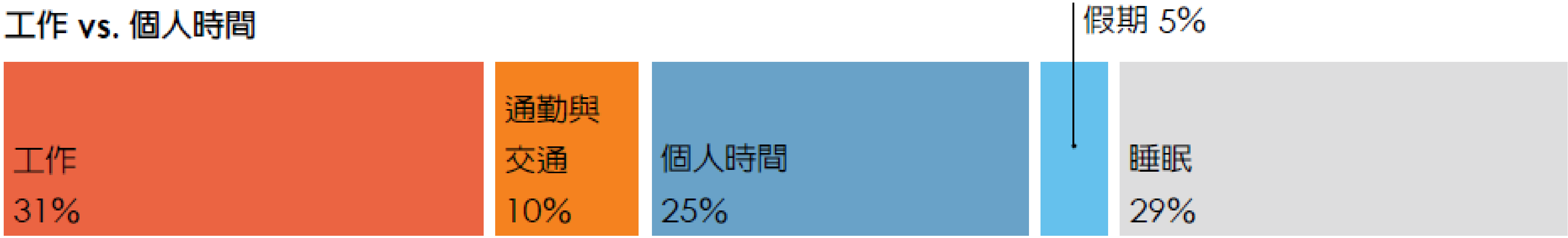
1. 分析該報告中的元素為何(1分)，有哪些變數(1分) ，有多少觀察值(1分)
2. 變數屬於何種尺度(3分)
3. 資料屬於質性或量化的資料(1分)
4. 資料屬於橫斷面或縱斷面資料(1分)
5. 資料來源屬於觀察的或實驗的(1分)
6. 若你們要蒐集這份報告中的資料，你們會考慮甚麼? (1分)

What is Statistics?

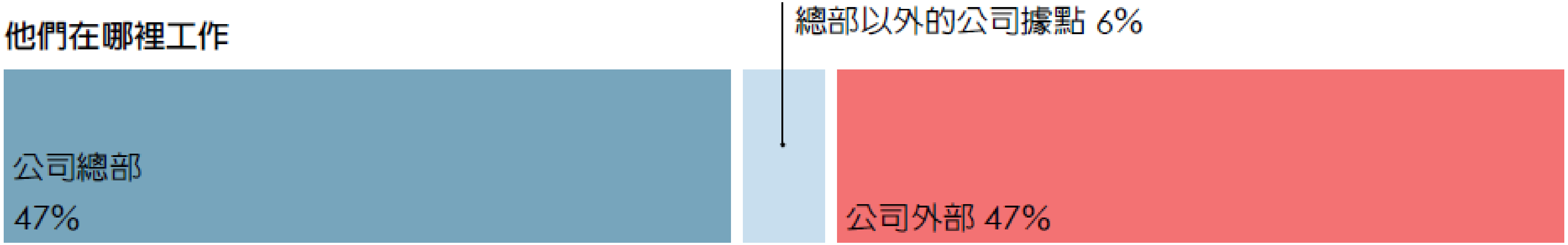
What is Statistics?

- The term statistics can refer 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.

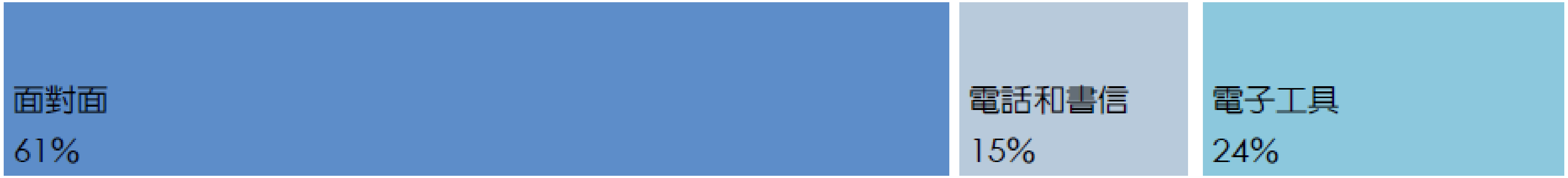
工作 vs. 個人時間



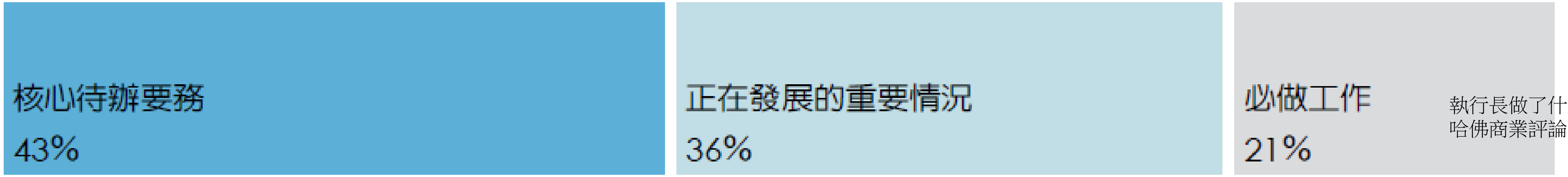
他們在哪裡工作



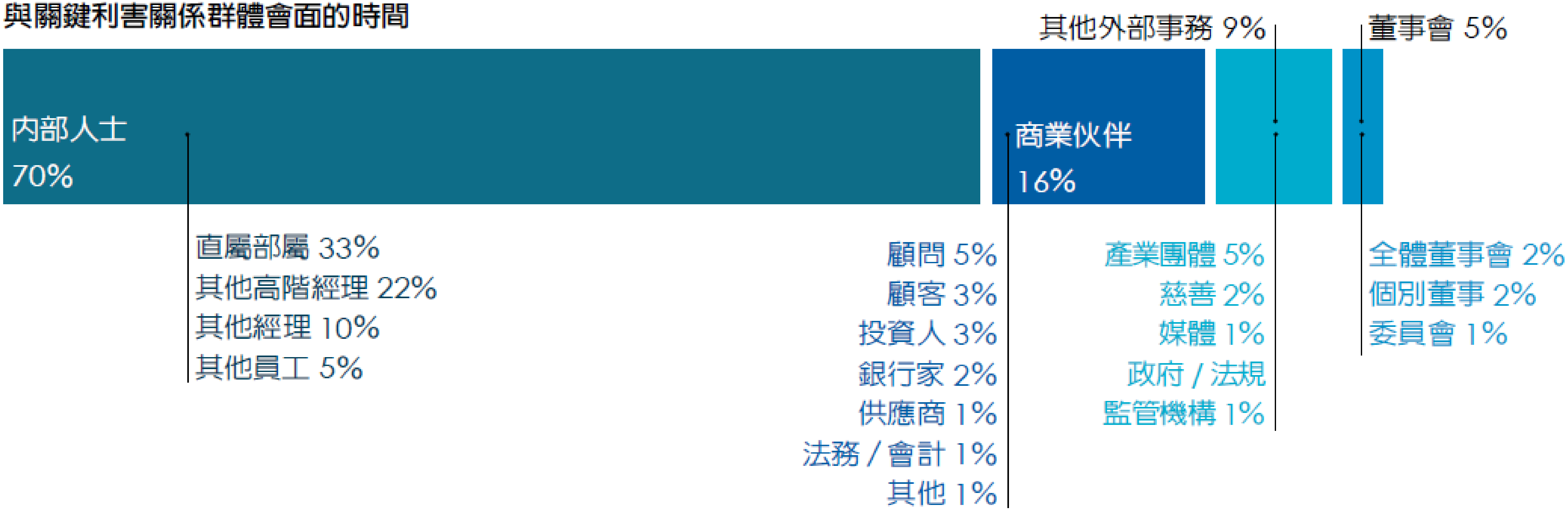
溝通模式



核心待辦要務 vs. 其他活動



與關鍵利害關係群體會面的時間



統計在商業與經濟上的應用有哪些?

- Accounting
- Economics
- Finance
- Marketing
- Production
- Information Systems

Applications in Business and 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.

Applications in Business and Economics

Marketing

- Electronic point-of-sale scanners at retail checkout counters are used to collect data for a variety of marketing research applications.

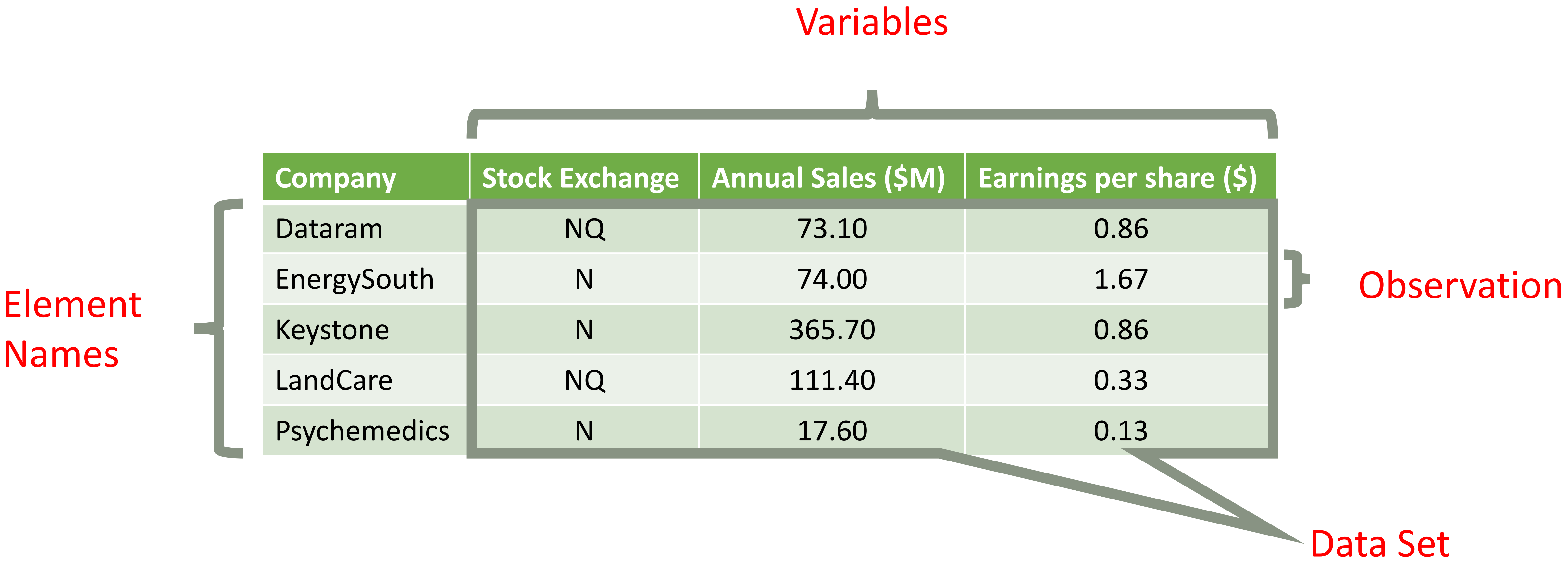
Production

- 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.

Data, Data Sets, Elements, Variables, and Observations



Variables

Element Names

Observation

Data Set

Company	Stock Exchange	Annual Sales ($M)	Earnings per share ($)
Dataram	NQ	73.10	0.86
EnergySouth	N	74.00	1.67
Keystone	N	365.70	0.86
LandCare	NQ	111.40	0.33
Psychemedics	N	17.60	0.13

Data and Data Sets

- Data are the facts and figures collected, analyzed, and summarized for presentation and interpretation.
- All the data collected in a particular study are referred to as the data set for the study.

Elements, Variables, and Observations

- Elements are the entities on which data are collected.
- A variable is a characteristic of interest for the elements.
- The set of measurements obtained for a particular element is called an observation.
- A data set with *n* elements contains *n* observations.
- The total number of data values in a complete data set is the number of elements multiplied by the number of variables.

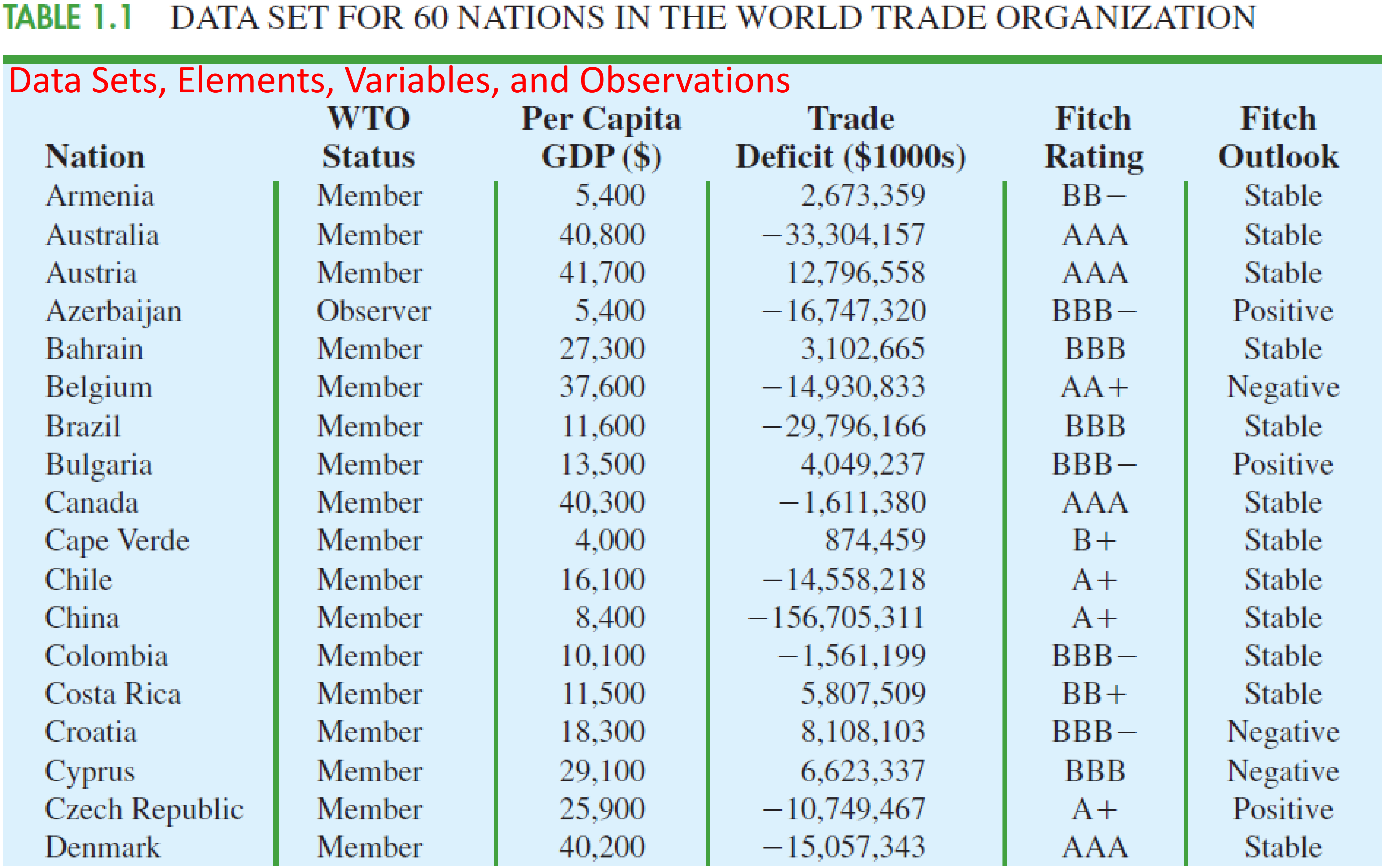


TABLE 1.1 DATA SET FOR 60 NATIONS IN THE WORLD TRADE ORGANIZATION

Data Sets, Elements, Variables, and Observations

Nation	WTO Status	Per Capita GDP ($)	Trade Deficit ($1000s)	Fitch Rating	Fitch Outlook
Armenia	Member	5,400	2,673,359	BB−	Stable
Australia	Member	40,800	−33,304,157	AAA	Stable
Austria	Member	41,700	12,796,558	AAA	Stable
Azerbaijan	Observer	5,400	−16,747,320	BBB−	Positive
Bahrain	Member	27,300	3,102,665	BBB	Stable
Belgium	Member	37,600	−14,930,833	AA+	Negative
Brazil	Member	11,600	−29,796,166	BBB	Stable
Bulgaria	Member	13,500	4,049,237	BBB−	Positive
Canada	Member	40,300	−1,611,380	AAA	Stable
Cape Verde	Member	4,000	874,459	B+	Stable
Chile	Member	16,100	−14,558,218	A+	Stable
China	Member	8,400	−156,705,311	A+	Stable
Colombia	Member	10,100	−1,561,199	BBB−	Stable
Costa Rica	Member	11,500	5,807,509	BB+	Stable
Croatia	Member	18,300	8,108,103	BBB−	Negative
Cyprus	Member	29,100	6,623,337	BBB	Negative
Czech Republic	Member	25,900	−10,749,467	A+	Positive
Denmark	Member	40,200	−15,057,343	AAA	Stable

Scales of Measurement

- Scales of measurement include
  - Nominal
  - Ordinal
  - Interval
  - Ratio
- The scale determines the amount of information contained in the data.
- The scale indicates the data summarization and statistical analyses that are most appropriate.

Scales of Measurement

Nominal scale

- Data are labels or names used to identify an attribute of the element.
- A nonnumeric label or numeric code may be used.

Example

Students of a university are classified by the school in which they are enrolled using a nonnumeric label such as Business, Humanities, Education, and so on.

Alternatively, a numeric code could be used for the school variable (e.g. 1 denotes Business, 2 denotes Humanities, 3 denotes Education, and so on).

Scales of Measurement

Ordinal scale

- The data have the properties of nominal data and the order or rank of the data is meaningful.
- A nonnumeric label or numeric code may be used.

Example

Students of a university are classified by their class standing using a nonnumeric label such as Freshman, Sophomore, Junior, or Senior.

Alternatively, a numeric code could be used for the class standing variable (e.g. 1 denotes Freshman, 2 denotes Sophomore, and so on).

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.
- Interval data are always numeric.

Example

Melissa has an SAT score of 1985, while Kevin has an SAT score of 1880. Melissa scored 105 points more than Kevin.

Scales of Measurement

Ratio scale

- Data have all the properties of interval data and the ratio of two values is meaningful.
- Ratio data are always numerical.
- Zero value is included in the scale.

Example:

Price of a book at a retail store is $ 200, while the price of the same book sold online is $100. The ratio property shows that retail stores charge twice the online price.

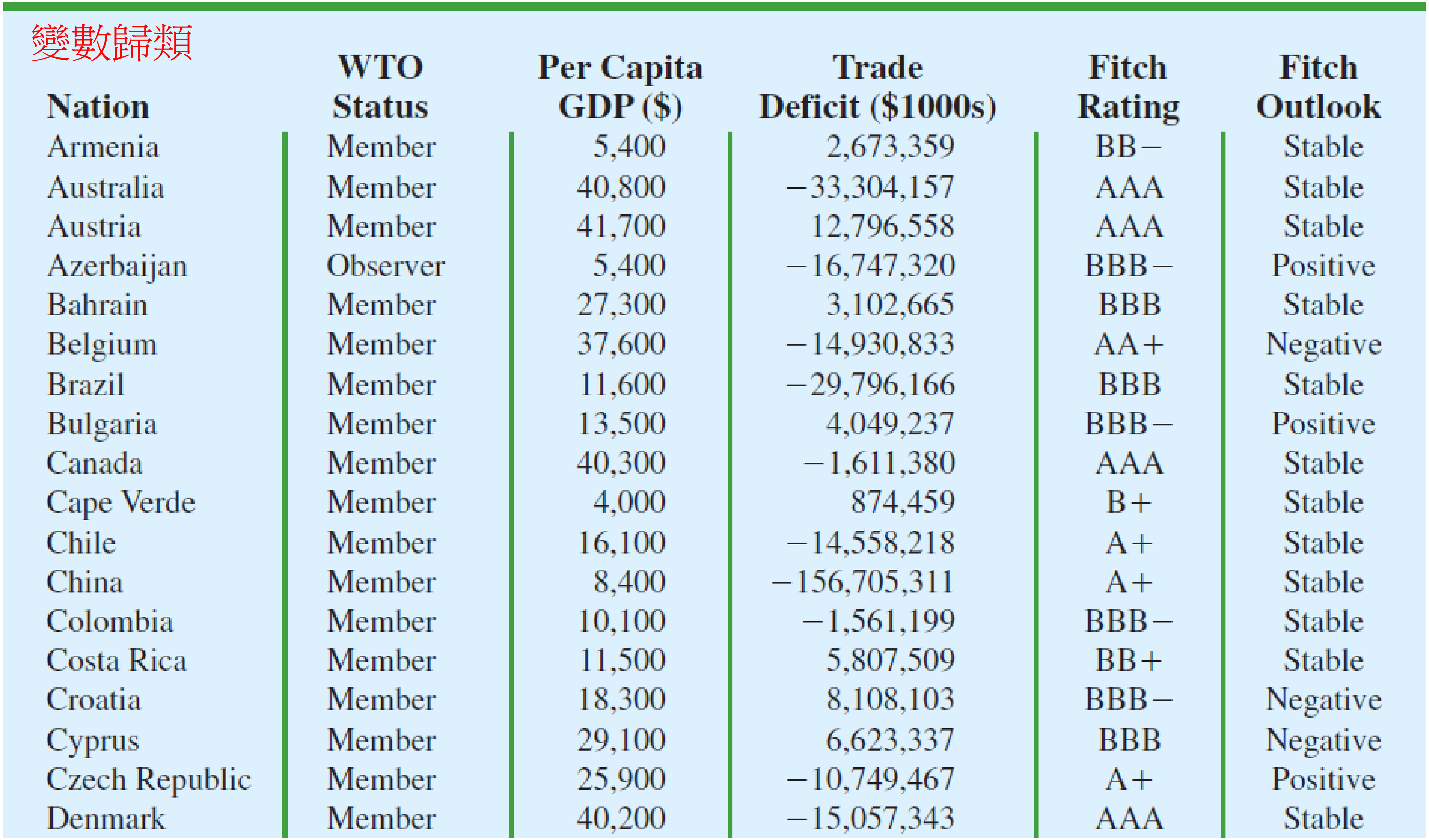
資料的類型—衡量尺度

類別/名目尺度(nominal scale)	• 宗教信仰、血型
順序尺度(ordinal scale)	• 滿意度、排名
等距/區間尺度(interval scale)	• 智商、溫度、成績
等比/比率尺度(ratio scale)	• 營業額、失業率

衡量尺度

特性	同一性	大於	固定間距	絕對零點
名目尺度	✓			
順序尺度	✓	✓		
等距尺度	✓	✓	✓	
等比尺度	✓	✓	✓	✓

TABLE 1.1 DATA SET FOR 60 NATIONS IN THE WORLD TRADE ORGANIZATION



變數歸類

Nation	WTO Status	Per Capita GDP ($)	Trade Deficit ($1000s)	Fitch Rating	Fitch Outlook
Armenia	Member	5,400	2,673,359	BB−	Stable
Australia	Member	40,800	−33,304,157	AAA	Stable
Austria	Member	41,700	12,796,558	AAA	Stable
Azerbaijan	Observer	5,400	−16,747,320	BBB−	Positive
Bahrain	Member	27,300	3,102,665	BBB	Stable
Belgium	Member	37,600	−14,930,833	AA+	Negative
Brazil	Member	11,600	−29,796,166	BBB	Stable
Bulgaria	Member	13,500	4,049,237	BBB−	Positive
Canada	Member	40,300	−1,611,380	AAA	Stable
Cape Verde	Member	4,000	874,459	B+	Stable
Chile	Member	16,100	−14,558,218	A+	Stable
China	Member	8,400	−156,705,311	A+	Stable
Colombia	Member	10,100	−1,561,199	BBB−	Stable
Costa Rica	Member	11,500	5,807,509	BB+	Stable
Croatia	Member	18,300	8,108,103	BBB−	Negative
Cyprus	Member	29,100	6,623,337	BBB	Negative
Czech Republic	Member	25,900	−10,749,467	A+	Positive
Denmark	Member	40,200	−15,057,343	AAA	Stable

Categorical(類別) and Quantitative(定量) Data

- Data can be further classified as being categorical or quantitative.
- The statistical analysis that is appropriate depends on whether the data for the variable are categorical or quantitative.
- In general, there are more alternatives for statistical analysis when the data are quantitative.

Categorical Data

- Labels or names are used to identify an attribute of each element
- Often referred to as qualitative data
- Use either the nominal or ordinal scale of measurement
- Can be either numeric or nonnumeric
- Appropriate statistical analyses are rather limited

Quantitative Data

- Quantitative data indicate how many or how much.
- Quantitative data are always numeric.
- Ordinary arithmetic operations are meaningful for quantitative data.

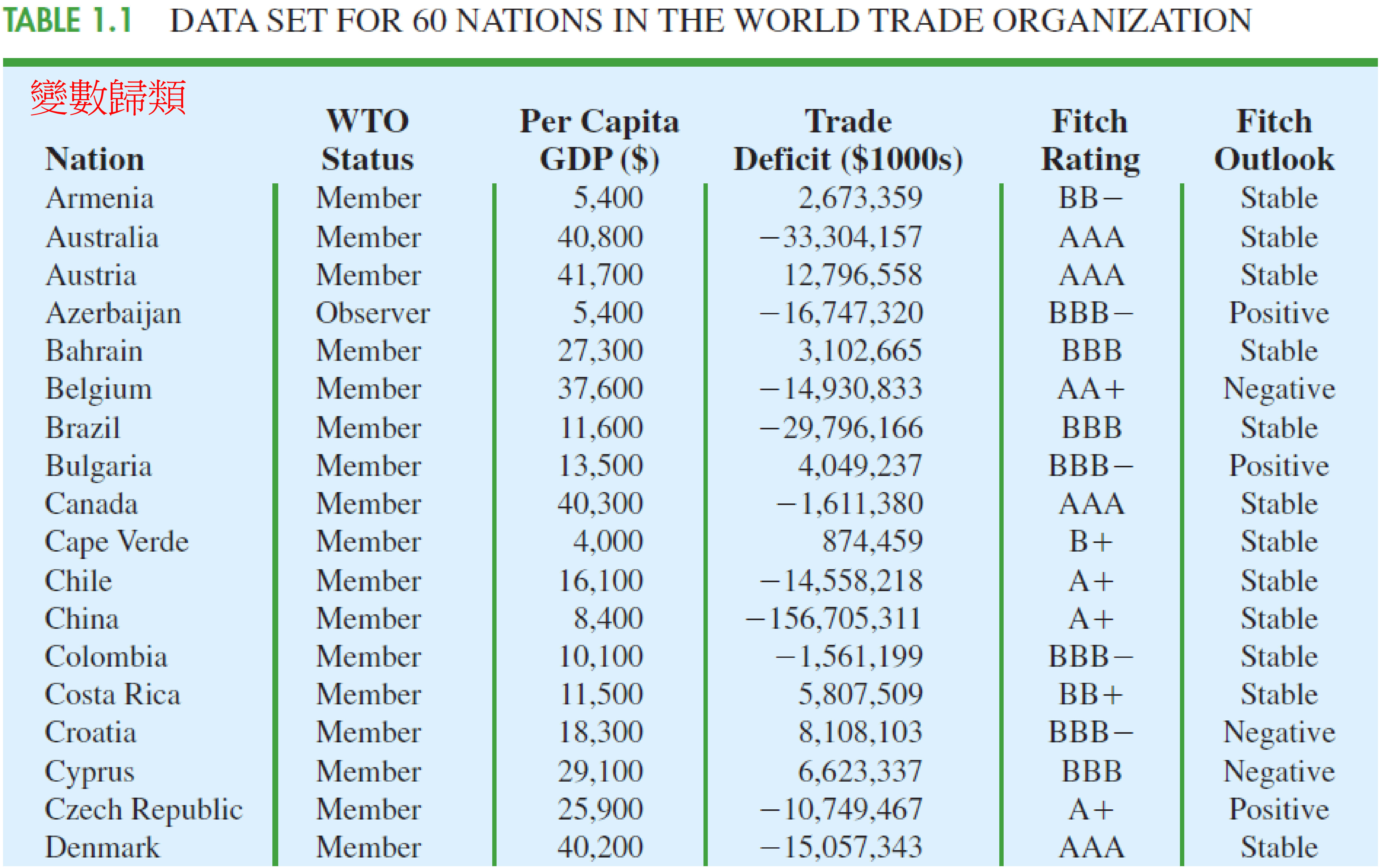


TABLE 1.1 DATA SET FOR 60 NATIONS IN THE WORLD TRADE ORGANIZATION

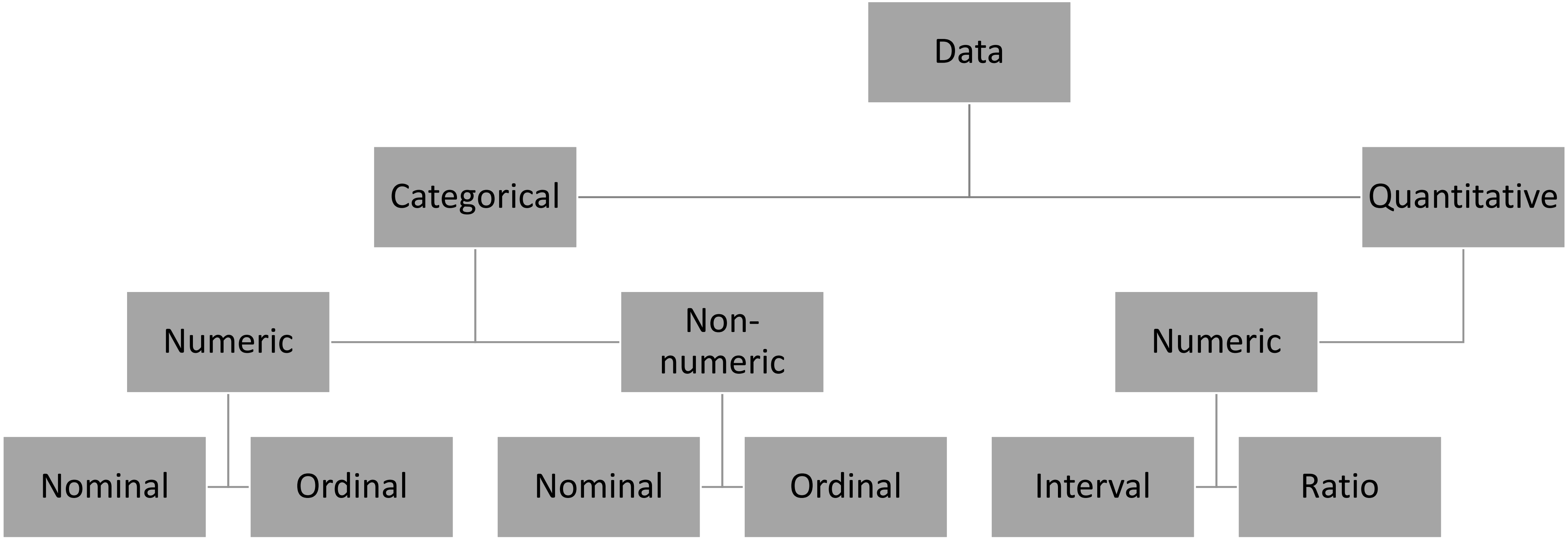
變數歸類

Nation	WTO Status	Per Capita GDP ($)	Trade Deficit ($1000s)	Fitch Rating	Fitch Outlook
Armenia	Member	5,400	2,673,359	BB−	Stable
Australia	Member	40,800	−33,304,157	AAA	Stable
Austria	Member	41,700	12,796,558	AAA	Stable
Azerbaijan	Observer	5,400	−16,747,320	BBB−	Positive
Bahrain	Member	27,300	3,102,665	BBB	Stable
Belgium	Member	37,600	−14,930,833	AA+	Negative
Brazil	Member	11,600	−29,796,166	BBB	Stable
Bulgaria	Member	13,500	4,049,237	BBB−	Positive
Canada	Member	40,300	−1,611,380	AAA	Stable
Cape Verde	Member	4,000	874,459	B+	Stable
Chile	Member	16,100	−14,558,218	A+	Stable
China	Member	8,400	−156,705,311	A+	Stable
Colombia	Member	10,100	−1,561,199	BBB−	Stable
Costa Rica	Member	11,500	5,807,509	BB+	Stable
Croatia	Member	18,300	8,108,103	BBB−	Negative
Cyprus	Member	29,100	6,623,337	BBB	Negative
Czech Republic	Member	25,900	−10,749,467	A+	Positive
Denmark	Member	40,200	−15,057,343	AAA	Stable

The *FinancialTimes*/Harris Poll is a monthly online poll of adults from six countries in Europe and the United States. A January poll included 1,015 adults in the United States. One of the questions asked was, “How would you rate the Federal Bank in handling the credit problems in the financial markets?” Possible responses were Excellent, Good, Fair, Bad, and Terrible (Harris Interactive website, January 2008).

1. What was the sample size for this survey?
2. Are the data categorical or quantitative?
3. Would it make more sense to use averages or percentages as a summary of the data for this question?
4. Of the respondents in the United States, 10% said the Federal Bank is doing a good job. How many individuals provided this response?

Scales of Measurement



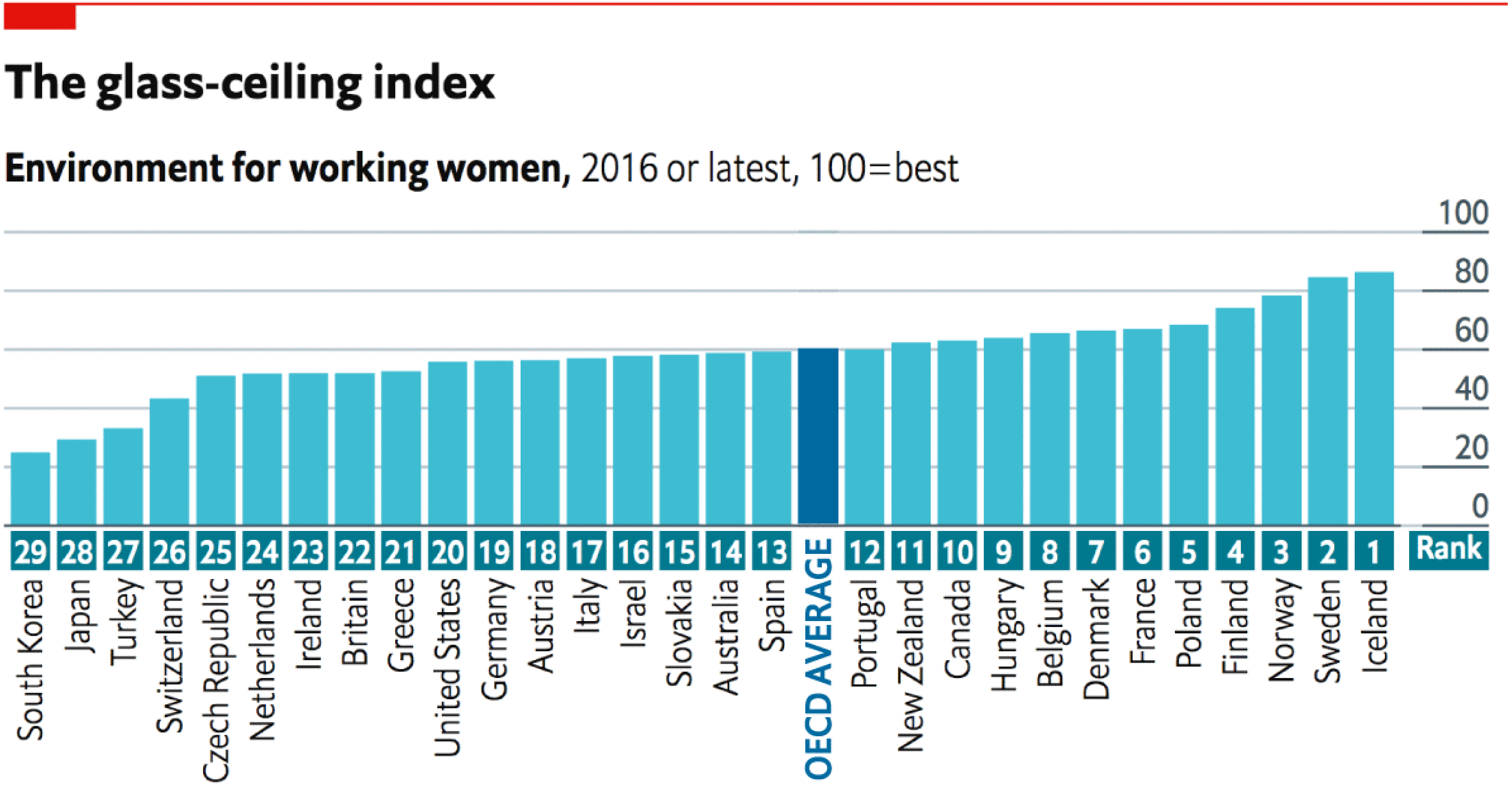
Cross-Sectional Data

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

Example

Data detailing the number of building permits issued in November 2013 in each of the counties of Ohio.

The best and worst places to be a working woman



The glass-ceiling index

Environment for working women, 2016 or latest, 100=best

Mar 8th 2017 by THE DATA TEAM (Economist) https://www.economist.com/blogs/graphicdetail/2017/03/daily-chart-0?fsrc=scn/fb/te/bl/ed/

Time Series Data

Time series data are collected over several time periods.

Example

Data detailing the number of building permits issued in Lucas County, Ohio in each of the last 36 months.

Graphs of time series data help analysts understand

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

Time Series Data

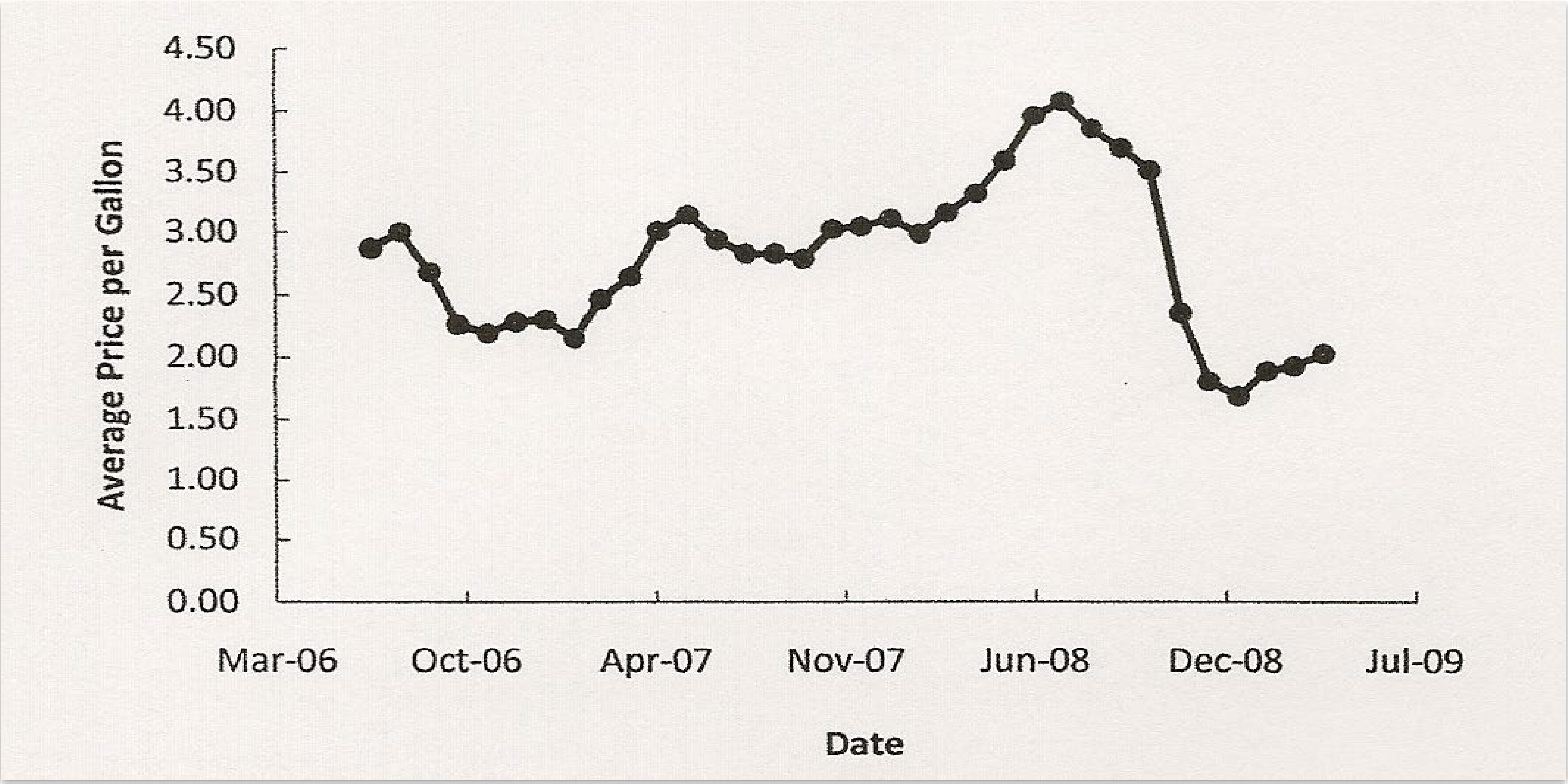
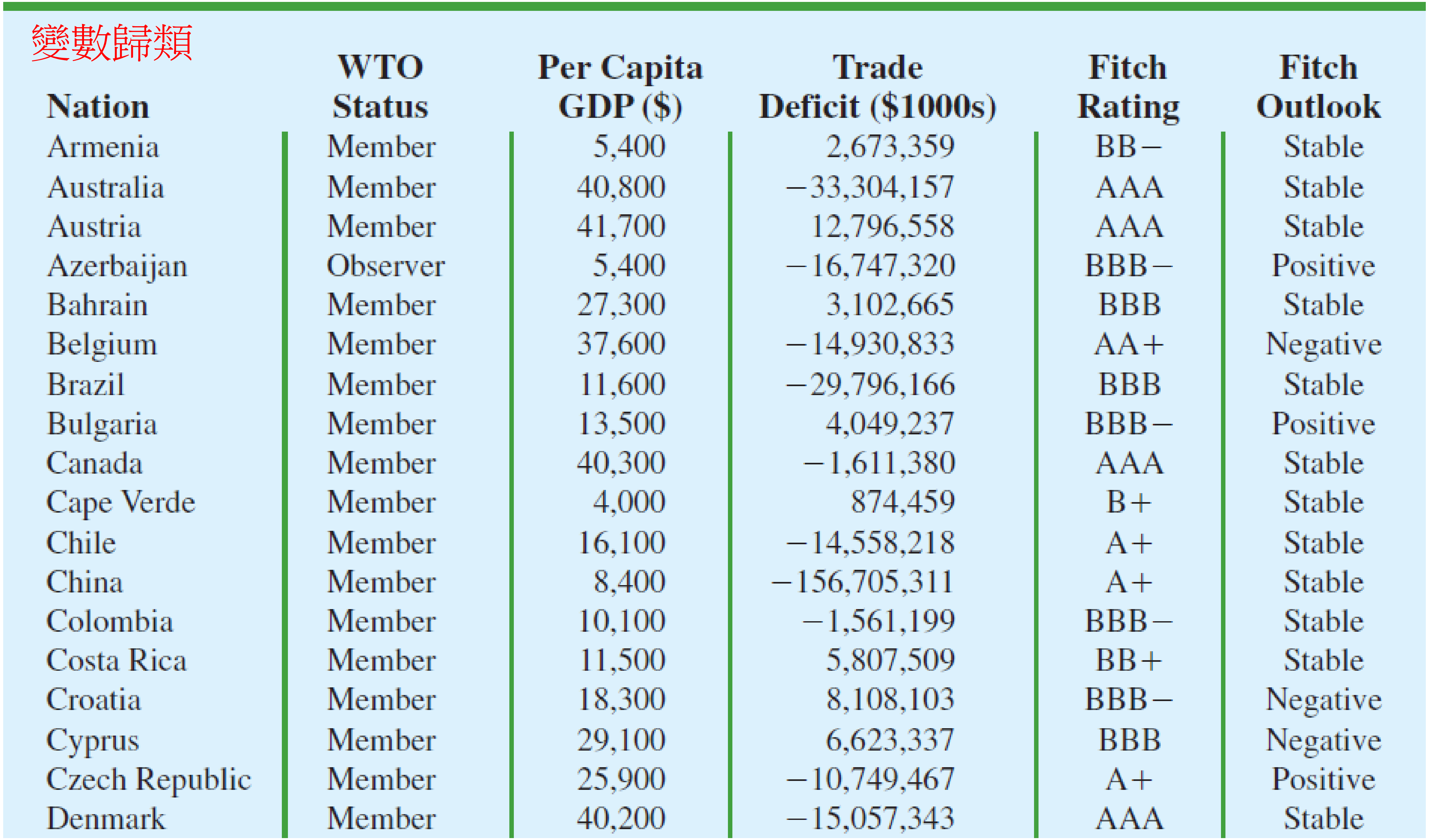


TABLE 1.1 DATA SET FOR 60 NATIONS IN THE WORLD TRADE ORGANIZATION

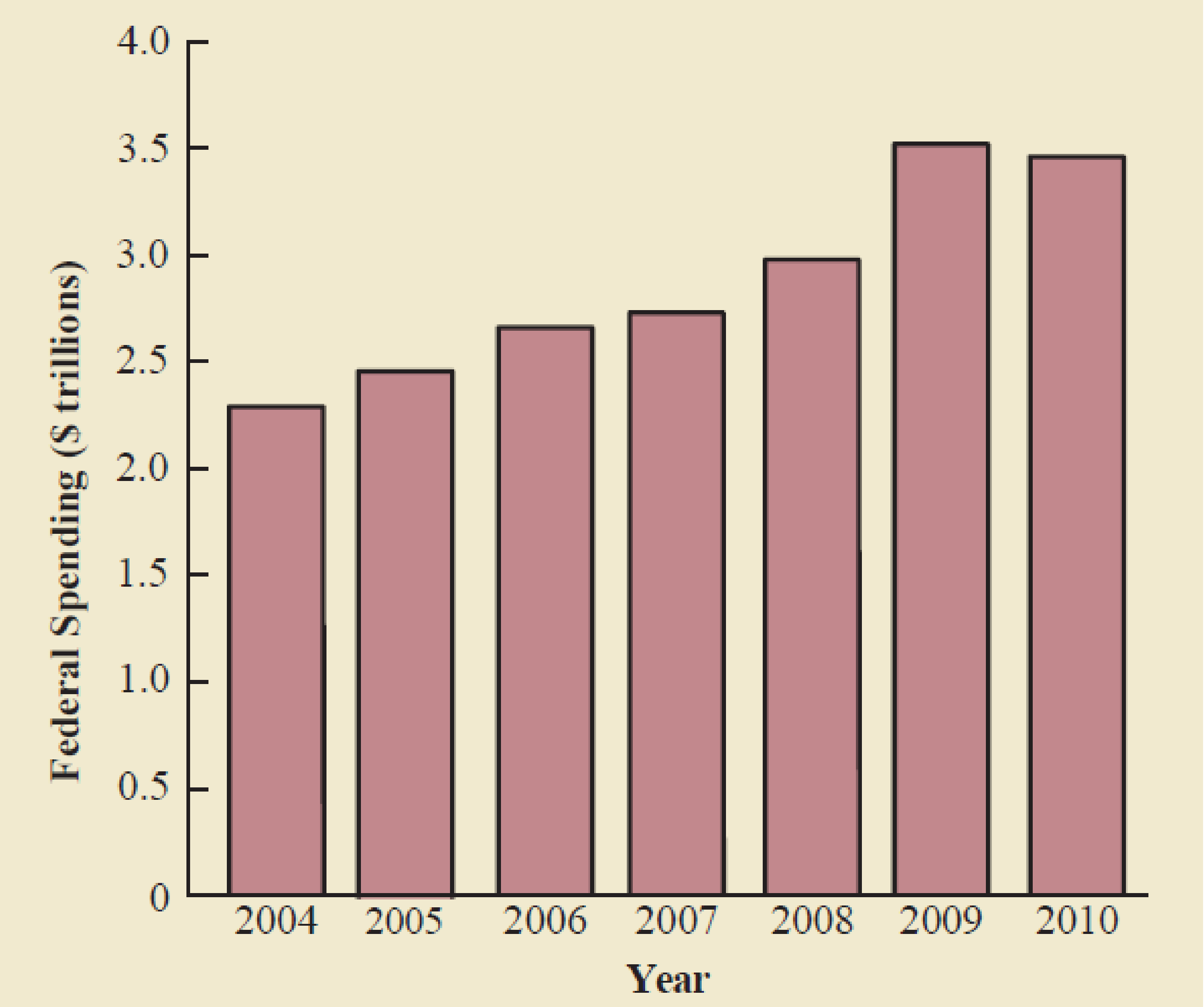


變數歸類

Nation	WTO Status	Per Capita GDP ($)	Trade Deficit ($1000s)	Fitch Rating	Fitch Outlook
Armenia	Member	5,400	2,673,359	BB−	Stable
Australia	Member	40,800	−33,304,157	AAA	Stable
Austria	Member	41,700	12,796,558	AAA	Stable
Azerbaijan	Observer	5,400	−16,747,320	BBB−	Positive
Bahrain	Member	27,300	3,102,665	BBB	Stable
Belgium	Member	37,600	−14,930,833	AA+	Negative
Brazil	Member	11,600	−29,796,166	BBB	Stable
Bulgaria	Member	13,500	4,049,237	BBB−	Positive
Canada	Member	40,300	−1,611,380	AAA	Stable
Cape Verde	Member	4,000	874,459	B+	Stable
Chile	Member	16,100	−14,558,218	A+	Stable
China	Member	8,400	−156,705,311	A+	Stable
Colombia	Member	10,100	−1,561,199	BBB−	Stable
Costa Rica	Member	11,500	5,807,509	BB+	Stable
Croatia	Member	18,300	8,108,103	BBB−	Negative
Cyprus	Member	29,100	6,623,337	BBB	Negative
Czech Republic	Member	25,900	−10,749,467	A+	Positive
Denmark	Member	40,200	−15,057,343	AAA	Stable

The figure provides a bar chart showing the amount of federal spending for the years 2004 to 2010 (Congressional Budget Office website, May 15, 2011).

1. What is the variable of interest?
2. Are the data categorical or quantitative?
3. Are the data time series or cross-sectional?
4. Comment on the trend in federal spending over time.



Data Sources

Existing Sources

- Internal company records – almost any department
- Business database services – Dow Jones & Co.
- Government agencies - U.S. Department of Labor
- Industry associations – Travel Industry Association of America
- Special-interest organizations – Graduate Management Admission Council (GMAT)
- Internet – more and more firms

Data Sources

Data Available From Internal Company Records

Record	Some of the Data Available
Employee records	Name, address, social security number
Production records	Part number, quantity produced, direct labor cost, material cost
Inventory records	Part number, quantity in stock, reorder level, economic order quantity
Sales records	Product number, sales volume, sales volume by region
Credit records	Customer name, credit limit, accounts receivable balance
Customer profile	Age, gender, income, household size

Data Sources

Data Available From Selected Government Agencies

Government Agency	Web address	Some of the Data Available
Census Bureau	www.census.gov	Population data, number of households, household income
Federal Reserve Board	www.federalreserve.gov	Data on money supply, exchange rates, discount rates
Office of Mgmt. & Budget	www.whitehouse.gov/omb	Data on revenue, expenditures, debt of federal government
Department of Commerce	www.doc.gov	Data on business activity, value of shipments, profit by industry
Bureau of Labor Statistics	www.bls.gov	Customer spending, unemployment rate, hourly earnings, safety record

Data Sources

資料共享平台，競賽平台，公開資料

- Kaggle https://www.kaggle.com/
- Data.world https://data.world/
- GitHub https://github.com/ (與Kaggle相似的簡體中文競賽平台)
- DataCastle http://www.pkbigdata.com/
- Oodata https://www.oodata.com.tw/
- 政府資料開放平臺 https://data.gov.tw/
- 臺北市資料大平臺 https://data.taipei/#/

Data Sources

Statistical Studies – Observational

- In observational (nonexperimental) studies no attempt is made to control or influence the variables of interest.
- Example - Survey
- Studies of smokers and nonsmokers are observational studies because researchers do not determine or control who will smoke and who will not smoke.

Data Sources

Statistical Studies – Experimental

- In experimental studies the variable of interest is first identified. Then one or more other variables are identified and controlled so that data can be obtained about how they influence the variable of interest.
- The largest experimental study ever conducted is believed to be the 1954 Public Health Service experiment for the Salk polio vaccine (小兒麻痺疫苗). Nearly two million U.S. children (grades 1- 3) were selected.

Data Acquisition Considerations

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 or were acquired with little care can lead to misleading information.

Descriptive Statistics (敘述統計)

- Most of the statistical information in newspapers, magazines, company reports, and other publications consists of data that are summarized and presented in a form that is easy to understand.
- Such summaries of data, which may be tabular, graphical, or numerical, are referred to as descriptive statistics.

Example

The manager of Hudson Auto would like to have a better understanding of the cost of parts used in the engine tune-ups performed in her shop. She examines 50 customer invoices for tune-ups. The costs of parts, rounded to the nearest dollar, are listed on the next slide.

Example: Hudson Auto Repair

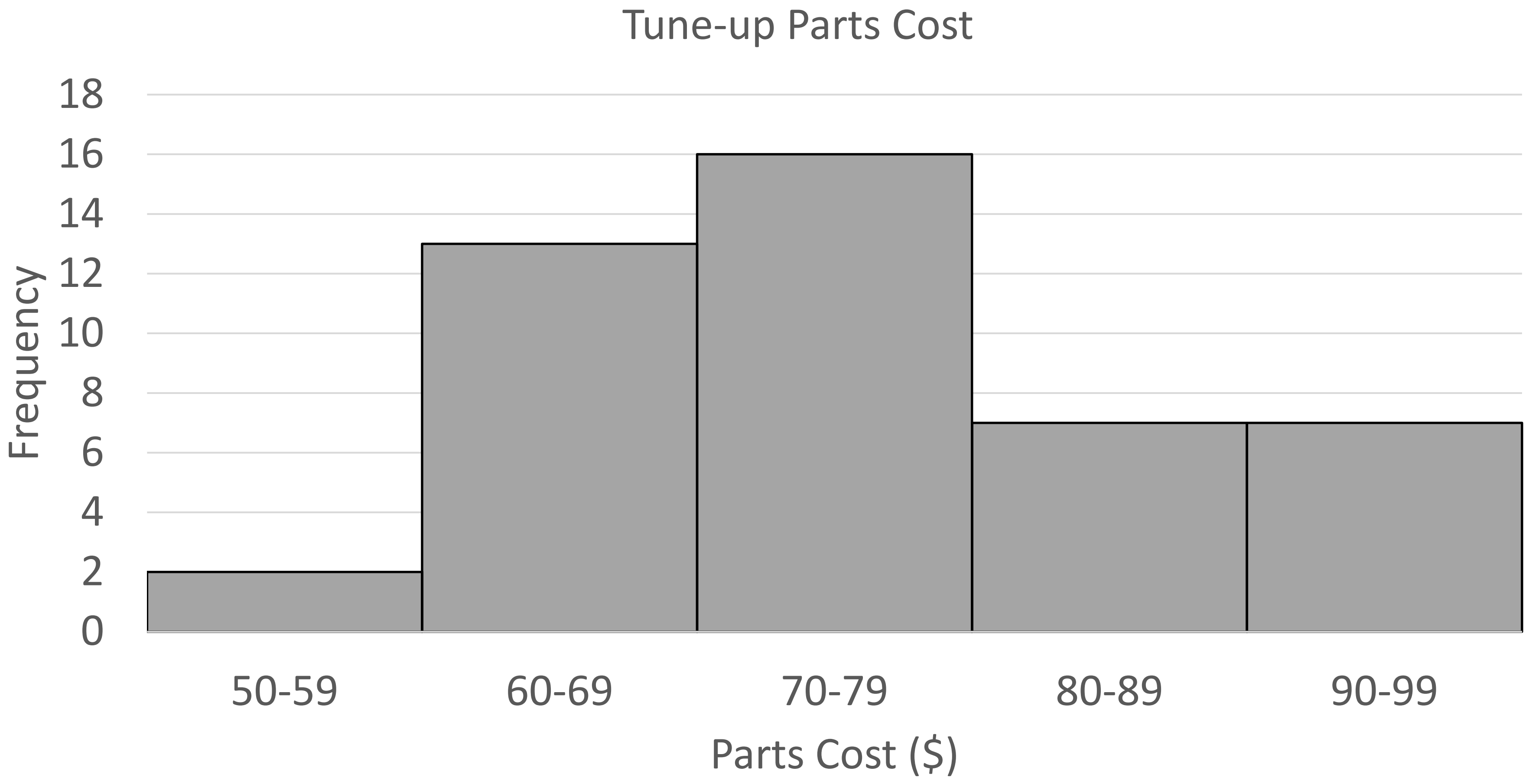
Sample of Parts Cost ($) for 50 Tune-ups

91	78	93	57	75	52	99	80	97	62
71	69	72	89	66	75	79	75	72	76
104	74	62	68	97	105	77	65	80	109
85	97	88	68	83	68	71	69	67	74
62	82	98	101	79	105	79	69	62	73

Tabular Summary: Frequency and Percent Frequency

Parts Cost ($)	Frequency	Percent Frequency
50-59	2	4%
60-69	13	26%
70-79	16	32%
80-89	7	14%
90-99	7	14%
100-109	5	10%
TOTAL	**50**	**100%**

Graphical Summary: Histogram



Numerical Descriptive Statistics

- The most common numerical descriptive statistic is the mean (or average).
- The mean demonstrates a measure of the central tendency, or central location of the data for a variable.
- Hudson's mean cost of parts, based on the 50 tune-ups studied is $79 (found by summing up the 50 cost values and then dividing by 50).

Statistical Inference (統計推論)

Population: The set of all elements of interest in a particular study.

Sample: A subset of the population.

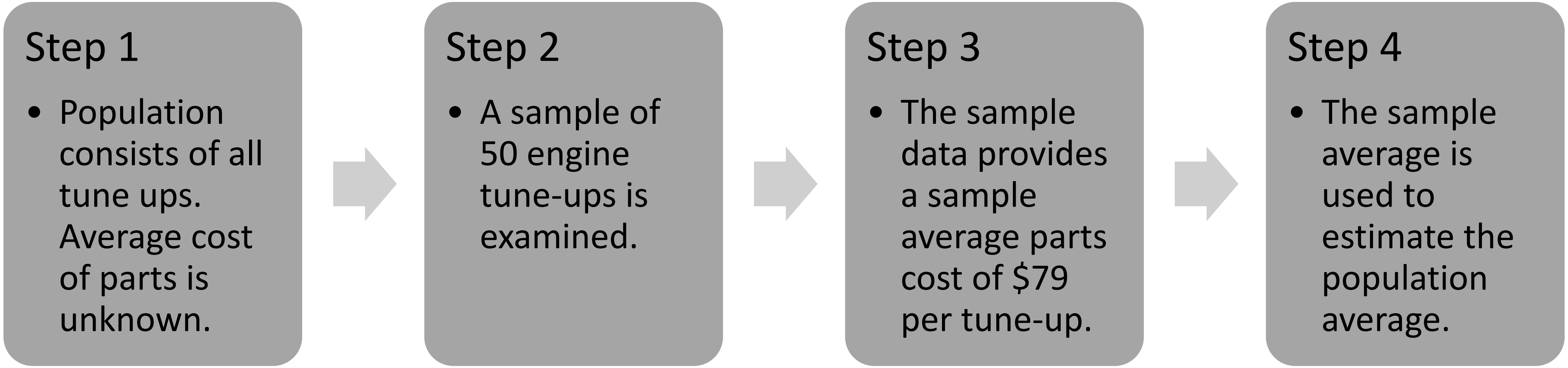
Statistical inference: The process of using data obtained from a sample to make estimates and test hypotheses about the characteristics of a population.

Census: Collecting data for the entire population.

Sample survey: Collecting data for a sample.

Process of Statistical Inference

Example: Hudson Auto



Step 1

- Population consists of all tune ups. Average cost of parts is unknown.

Step 2

- A sample of 50 engine tune-ups is examined.

Step 3

- The sample data provides a sample average parts cost of $79 per tune-up.

Step 4

- The sample average is used to estimate the population average.

我們如何得知大學生實習主要的工作內容為何，平均月薪多少?
在這個問題中甚麼是母體?甚麼是樣本?

研究的道德準則與IRB (Internal Review Board)

- 「道德準則」的緣起：第二次世界大戰結束後，德國科學家所做的一些「不人道」研究逐步的被揭露

1. 納粹相信德國人是優秀的種族，所以在「優生學」的 名義下，為了找出一個有效減少劣等民族人口的方 法，他們在不加麻醉的情況下，大量的對囚犯做「絕育手術」的實驗。
2. 他們用「減壓艙」來模擬高地的壓力，然後將囚犯關入艙內，看他們能忍受多久，才會死亡。
3. 為了模擬飛機失事的情形，德國空軍曾將猶太囚犯放入接近冰點的冷水中，看飛行員可以在冰水中存活多久 。

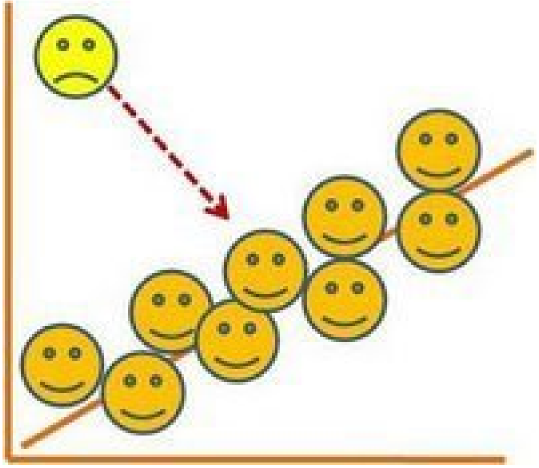
胡志偉，學術研究倫理簡介： 從國內外學術倫理案件談起

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 as you collect, analyze, and present data.
- As a consumer of statistics, one should also be aware of the possibility of unethical behavior by others.

Case

- Company A hired a consulting company to conduct a survey to understand employee satisfaction and willingness to stay, and to use the results as a publicity for recruiting next year.



個案中不合乎道德的地方有哪些?

Ethical Guidelines for Statistical Practice

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