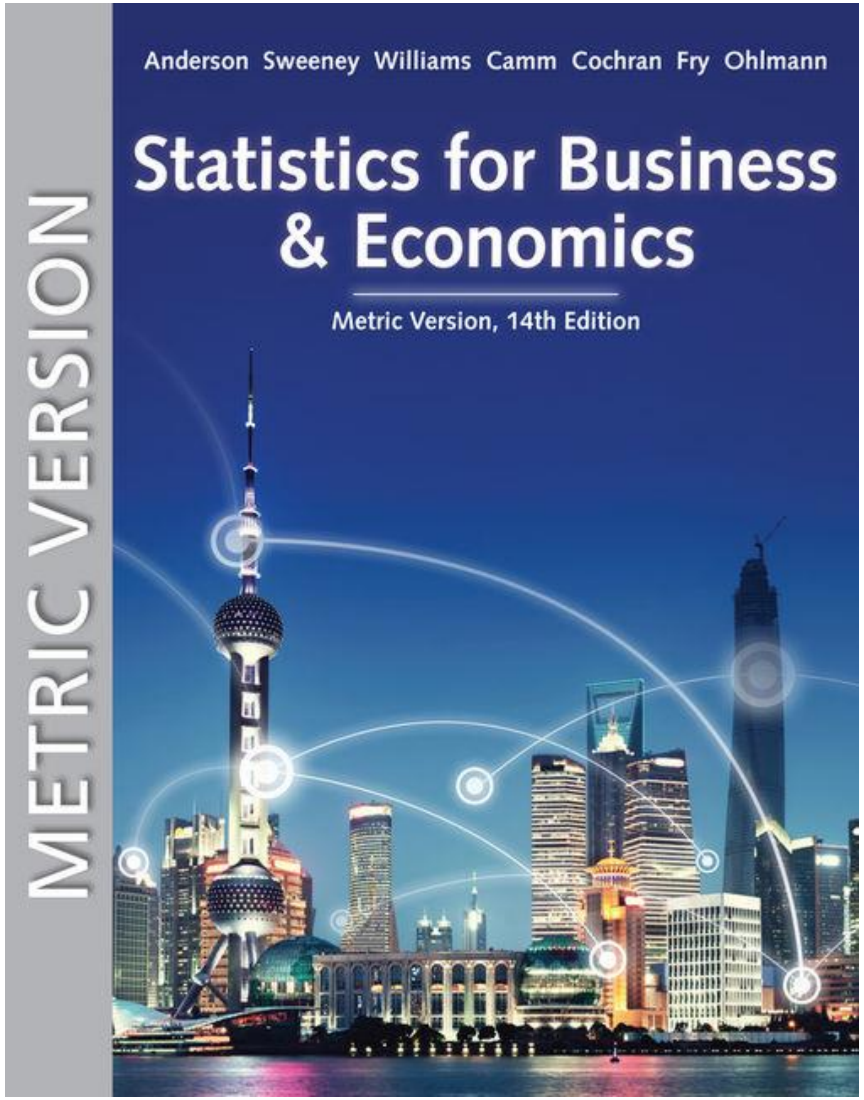


Statistics for Business and Economics (14e) Metric Version

Chapters 1~2

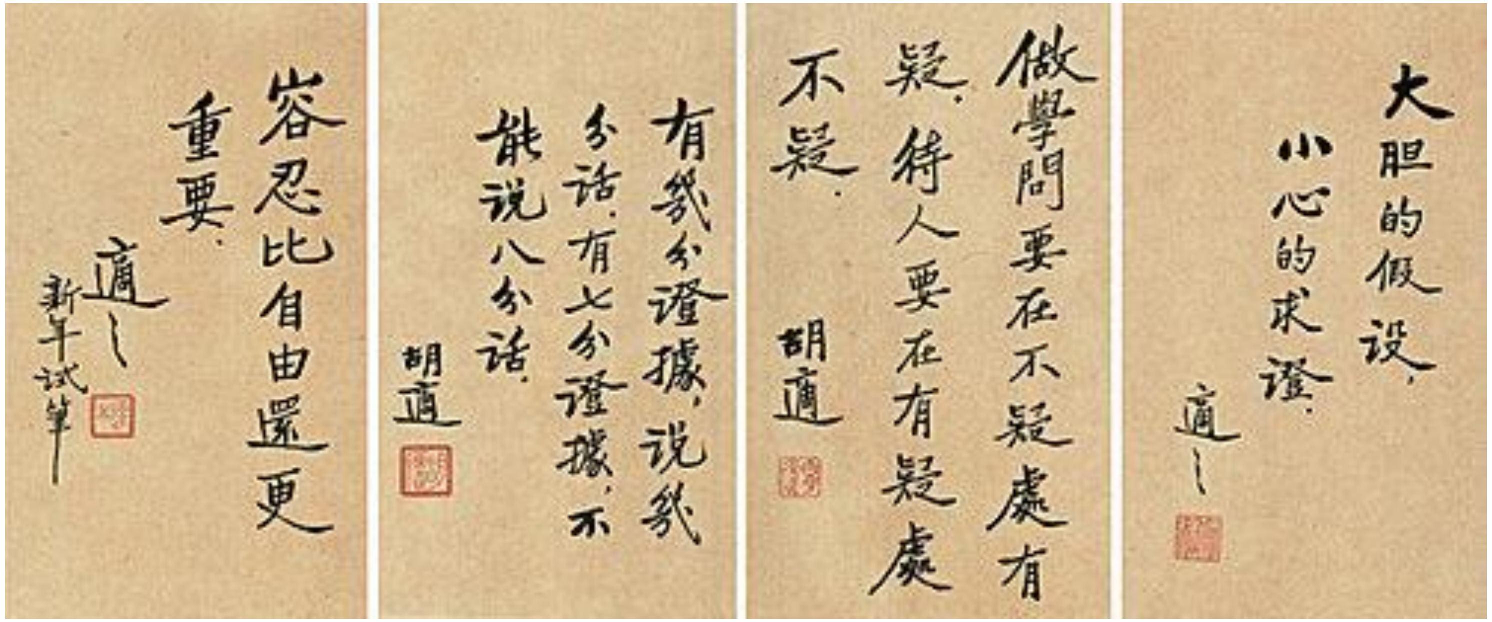
Chapter 1 - Data and Statistics

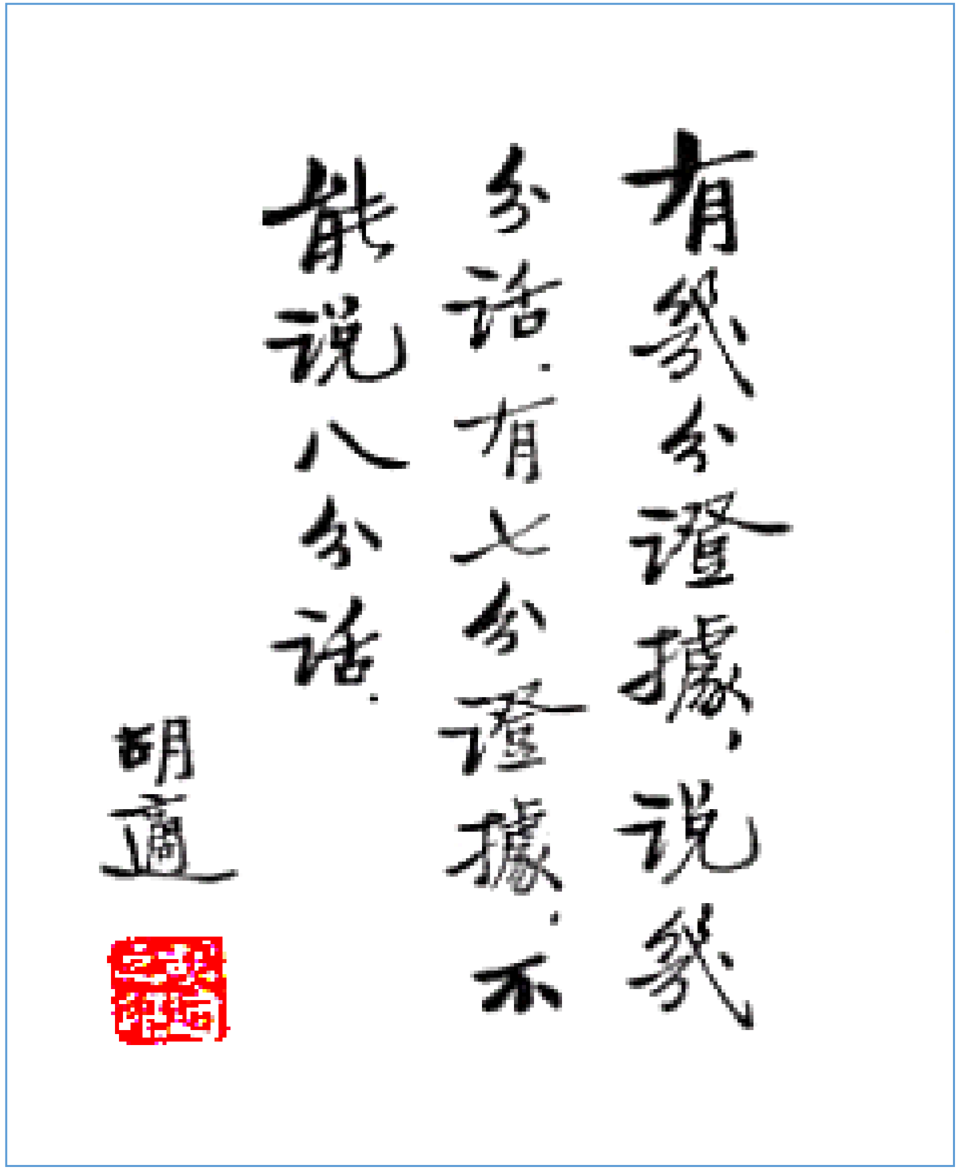
What Is Statistics?

- 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.

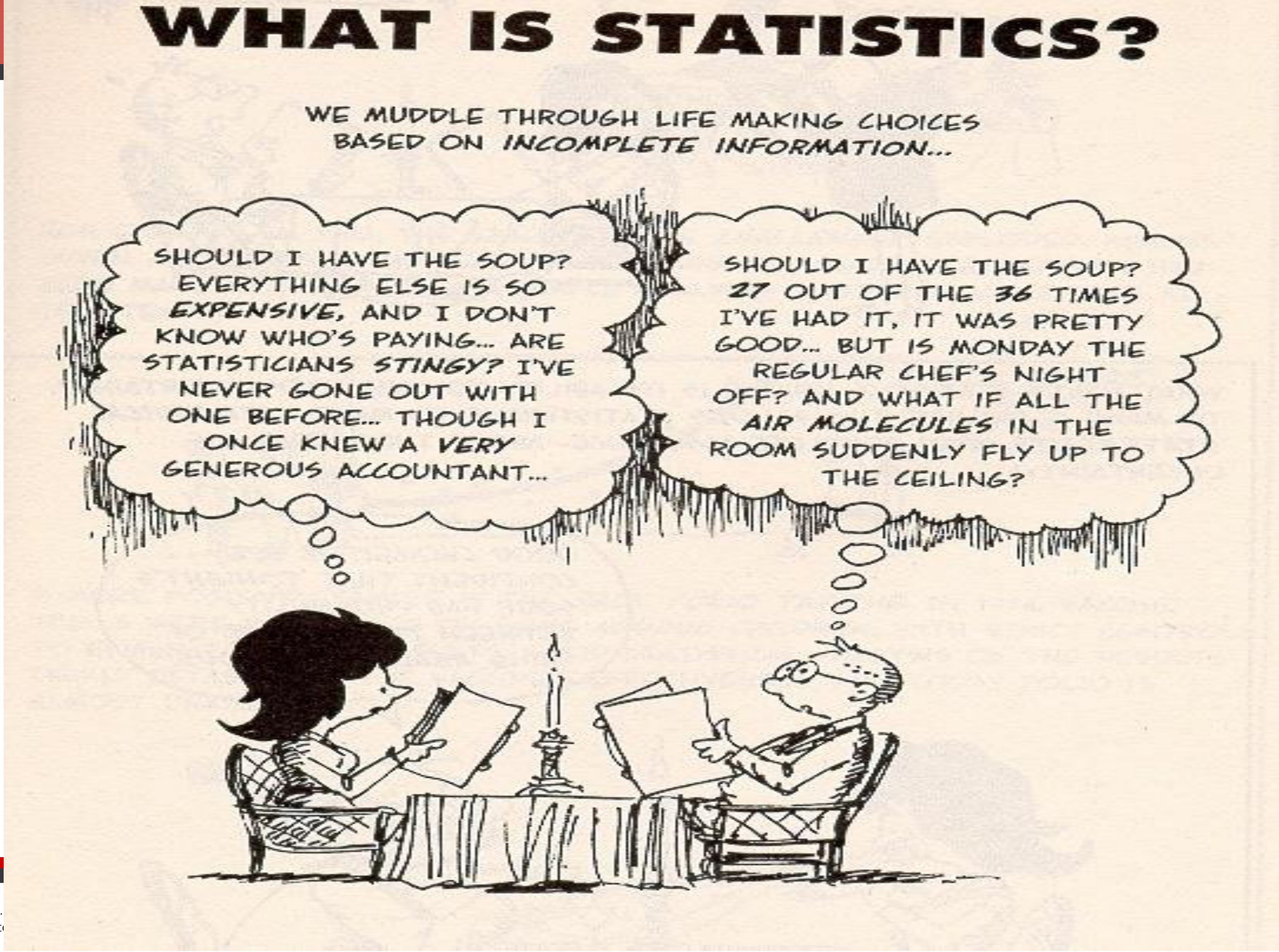
什麼是統計?

- 統計學是研究定義問題、運用資料蒐集、整理、陳示、分析與推論等科學方法，在不確定(Uncertainty)情況下， 做出合理決策的科學。





https://en.wikipedia.org/wiki/File:Hu_Shih_1960_color.jpg


WHAT IS STATISTICS?
WE MUDDLE THROUGH LIFE MAKING CHOICES BASED ON INCOMPLETE INFORMATION...
SHOULD I HAVE THE SOUP? EVERYTHING ELSE IS SO EXPENSIVE, AND I DON'T KNOW WHO'S PAYING... ARE STATISTICIANS STINGY? I'VE NEVER GONE OUT WITH ONE BEFORE... THOUGH I ONCE KNEW A VERY GENEROUS ACCOUNTANT...
SHOULD I HAVE THE SOUP? 27 OUT OF THE 36 TIMES I'VE HAD IT, IT WAS PRETTY GOOD... BUT IS MONDAY THE REGULAR CHEF'S NIGHT OFF? AND WHAT IF ALL THE AIR MOLECULES IN THE ROOM SUDDENLY FLY UP TO THE CEILING?

統計與知識

- 統計整理資訊為歸納法(Induction)，從龐雜的資料找出共同趨勢，區分資料為：

Regular（一般；規律）

↓

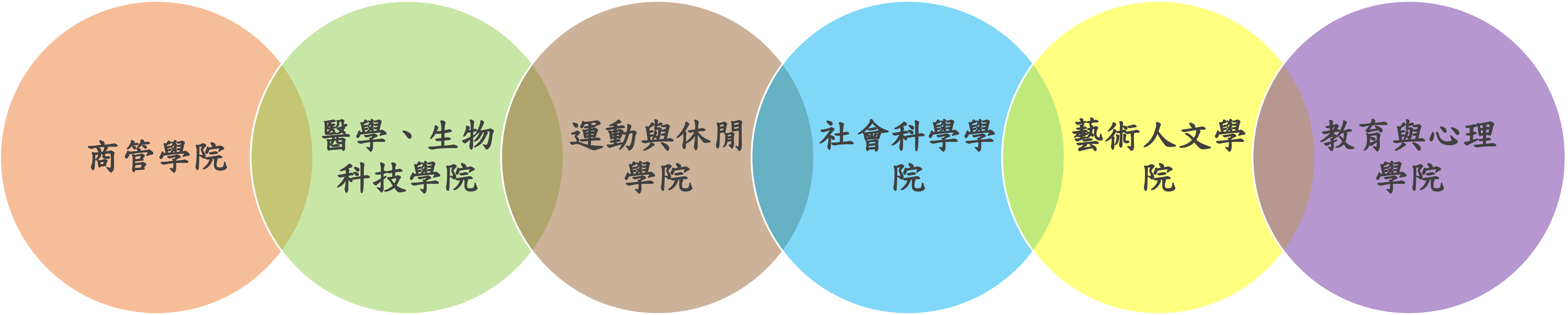
Irregular（異常）

↓

Extreme（極端）

大數據的應用領域

- 大數據應用領域按學院分類為六大類：



商管學院

- 商業：行銷領域中應用普遍

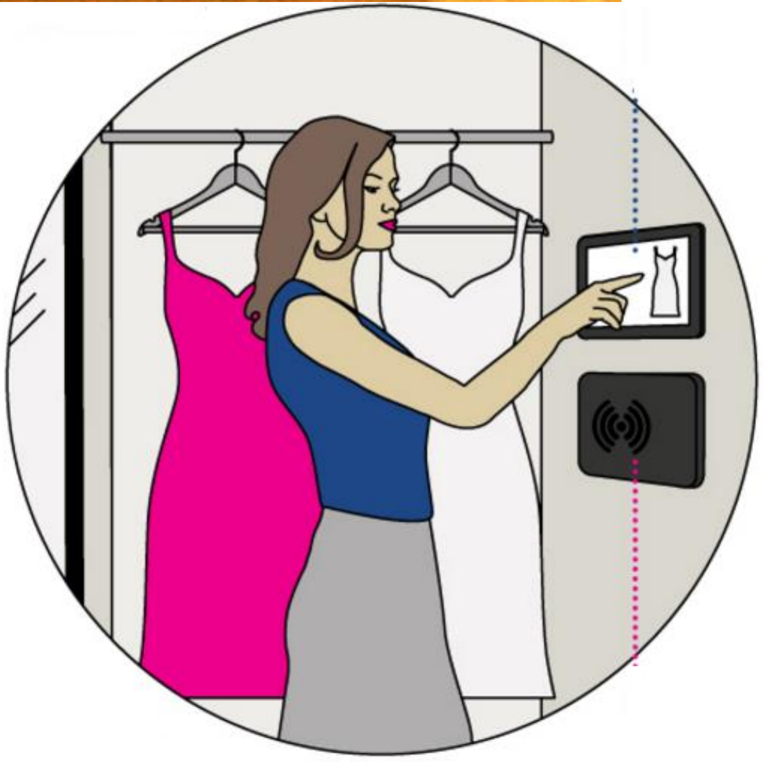
→Walmart尿布與啤酒

→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過程

→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

□ 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)

→<u>第三方支付單位</u>(PayPal、Apple Pay、支付寶等)提供網路收款及付款服務，保障買賣雙方權利。

→<u>網路銀行</u>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)

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/

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.

Data, Data Sets, Elements, Variables, and Observations

Variables

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

Element Names

Observation

Data Set

註：編碼簿（Codebook）

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* ☒ *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 ___ years | Month | Day | Year of birth

9. What is Person 1's race?

Mark ☒ *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 ☐ Vietnamese ☐ Native Hawaiian

☐ Filipino ☐ Korean ☐ Samoan

☐ Asian Indian ☐ Japanese ☐ Chamorro

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

☐ Other Pacific Islander – *Print, for example, Tongan, Fijian, Marshallese, etc.*

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

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

Scales of Measurement (1 of 6)

- 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 (2 of 6)

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 (3 of 6)

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 (4 of 6)

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 (5 of 6)

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.

Categorical and Quantitative 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.

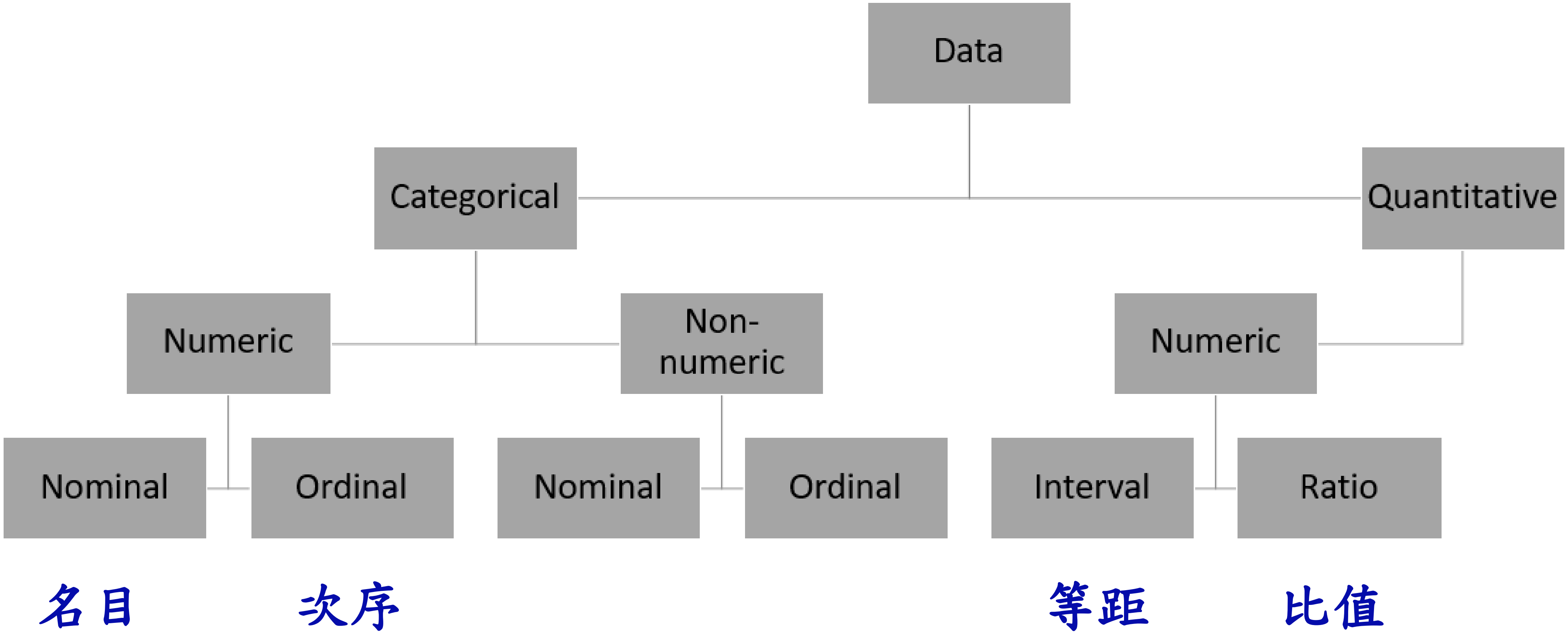
Categorical Data

- 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

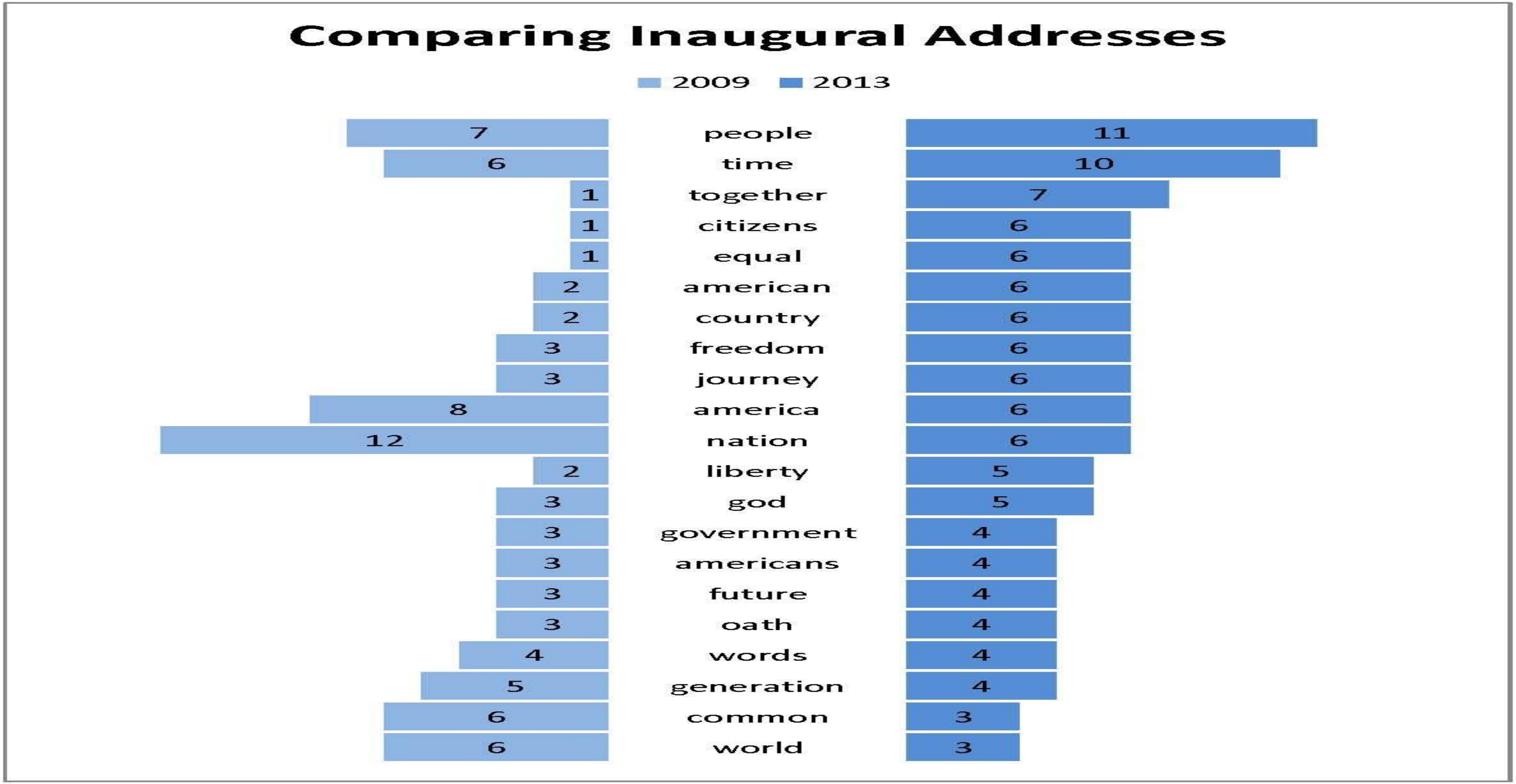
- Quantitative data indicate how many or how much.
- Quantitative data are always numeric.
- Ordinary arithmetic operations are meaningful for quantitative data.

Scales of Measurement (6 of 6)



電腦容量單位的演變（資料爆炸！）

單位	縮寫	意義
Bit	b	1 or 0
Byte	B	8 Bits
Kilobyte	KB	1,024 Bytes
Megabyte	MB	1,024 KB
Gigabyte	GB	1,024 MB
Terabyte	TB	1,024 GB
Petabyte	PB	1,024 TB
Exabyte	EB	1,024 PB
Zettabyte	ZB	1,024 EB
Yottabyte	YB	1,024 ZB



Analyzing the speech of President Obama (**Textmining**)

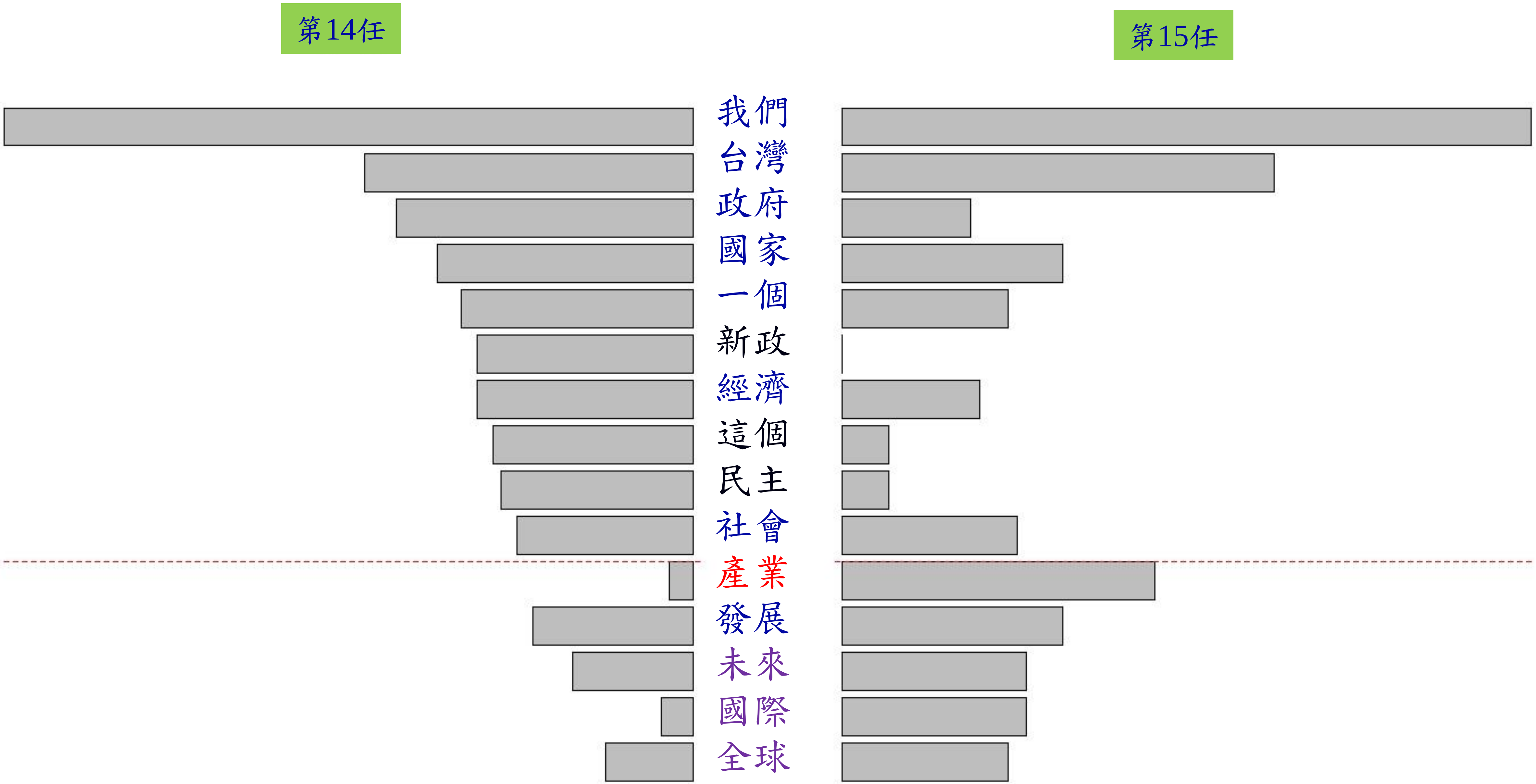
第14任蔡英文總統就職演講稿最常出現字詞

排序	單字			雙字詞		
	類別	次數	頻率	類別	次數	頻率
1	的	293	5.48%	我們	86	2.012%
2	我	114	2.13%	台灣	41	0.959%
3	們	90	1.68%	政府	37	0.866%
4	一	75	1.40%	國家	32	0.749%
5	會	74	1.38%	一個	29	0.679%
6	是	70	1.31%	新政	27	0.632%
7	個	66	1,23%	經濟	27	0.632%
8	民	63	1.18%	這個	25	0.585%
9	人	59	1.10%	民主	24	0.562%
10	國	59	1.10%	社會	22	0.515%

第15任蔡英文總統就職演講稿最常出現字詞

排序	單字			雙字詞		
	類別	次數	頻率	類別	次數	頻率
1	的	257	4.94%	我們	75	2.590%
2	我	114	2.10%	台灣	47	1.620%
3	們	92	1.77%	產業	34	1.170%
4	國	79	1.52%	國家	24	0.830%
5	人	68	1.31%	發展	24	0.830%
6	會	65	1.25%	未來	20	0.690%
7	在	63	1,21%	國際	20	0.690%
8	一	62	1.19%	社會	19	0.660%
9	是	55	1.06%	全球	18	0.620%
10	要	53	1.02%	一個	18	0.620%

蔡英文總統就職演講稿常見雙字詞



臺灣報紙頭條新聞的用字比較(2012~2019年)

《蘋果日報》

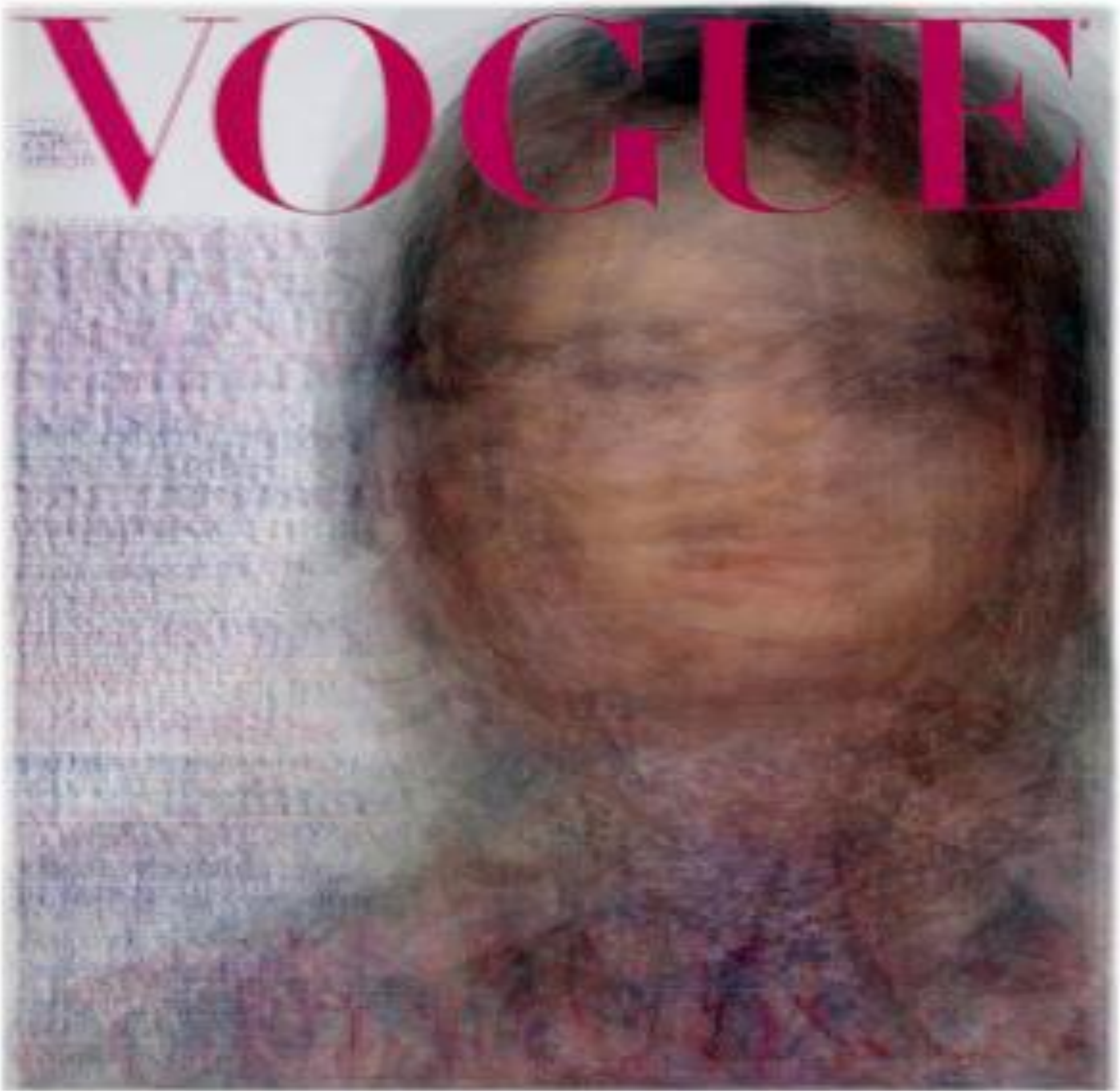
卡方指標

《中國時報》

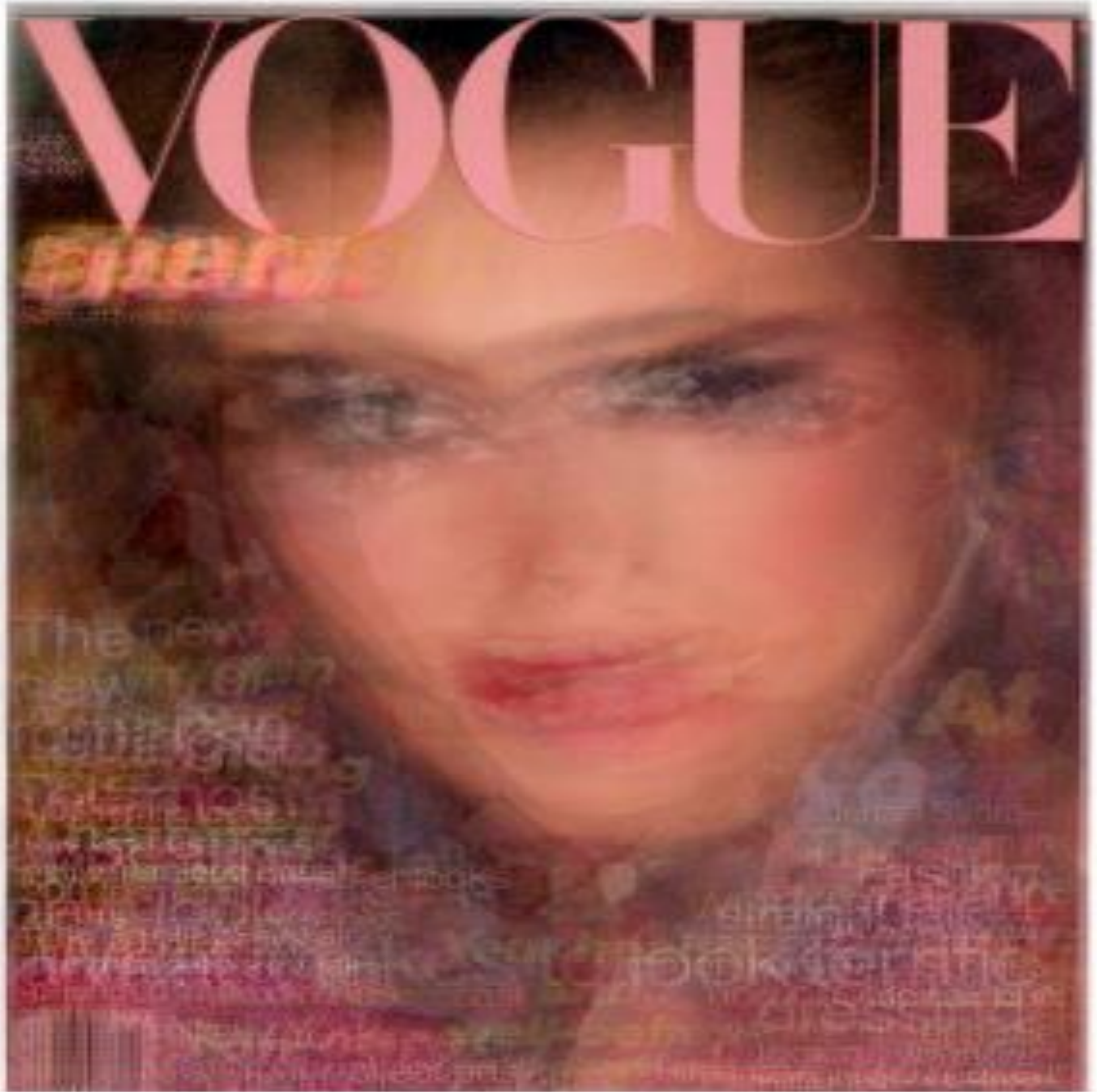
《自由時報》

耶魯大學數位人文實驗室「Robots Reading Vogue」

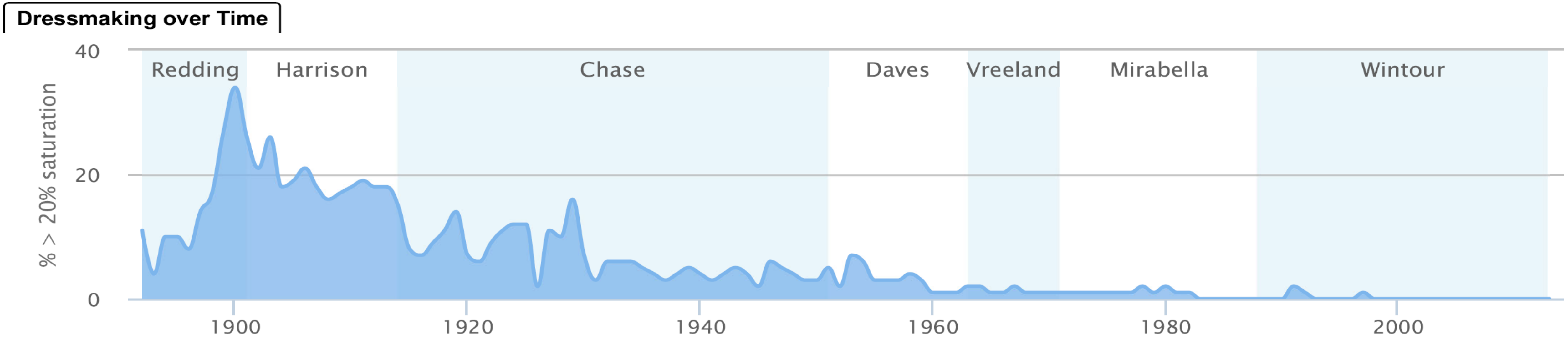
1970

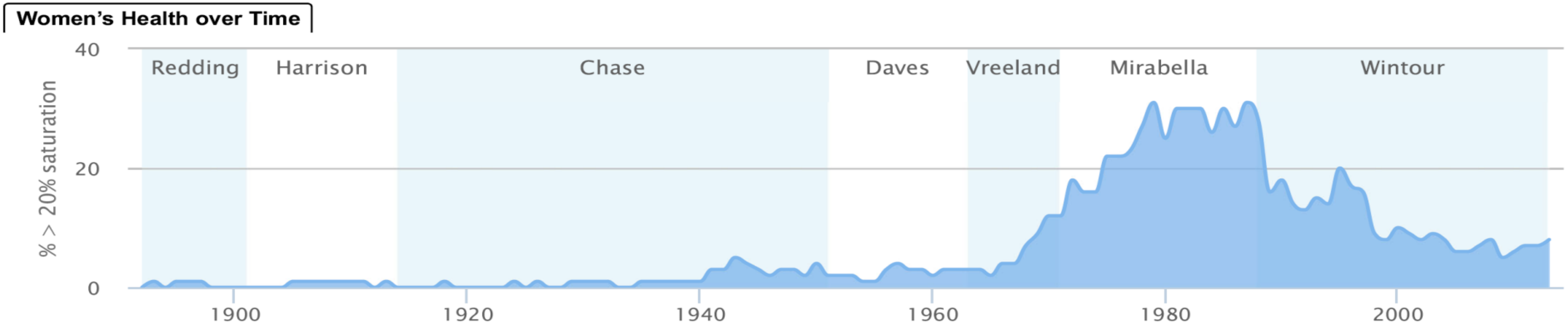


1980



https://miro.medium.com/max/3778/1*zdoQ-oKnWAPBKbUMYYL--w.jpeg





Vogue雜誌的風格趨勢變化

Cross-Sectional Data （橫斷面資料）

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

Longitudinal Data （縱斷面資料）

Example

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

https://t18.pimg.jp/039/215/248/1/39215248.jpg

https://t16.pimg.jp/072/409/306/1/72409306.jpg

Time Series Data (1 of 2)

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.

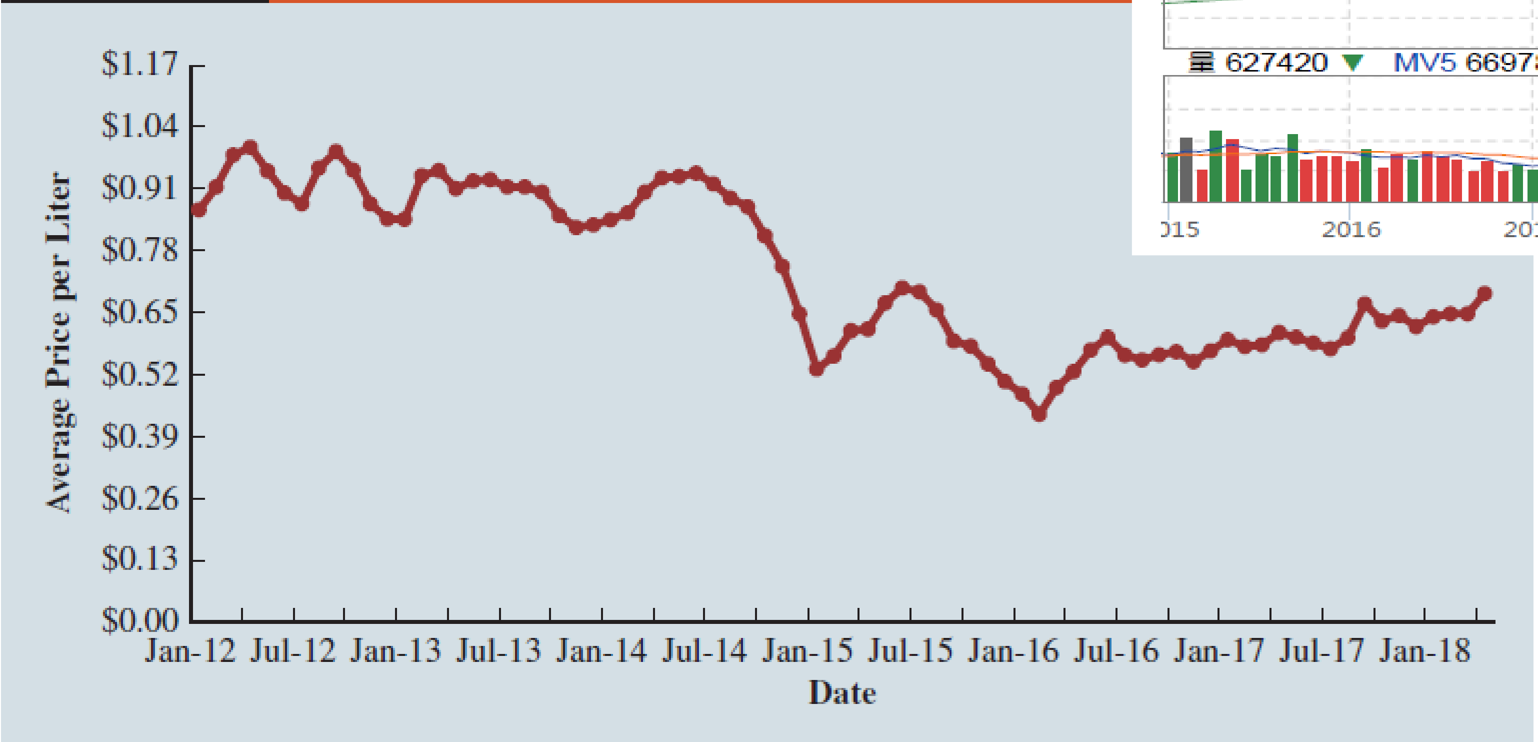
Graphs of time series data help analysts understand

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

Time Series Data (2 of 2)

Graph of Time Series Data

FIGURE 1.1 U.S. Average Price per Liter for Conventional Reg



Data Sources (1 of 5)

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
- 臺灣政府也有許多開放資料可供下載，例如：內政部統計處

  https://www.moi.gov.tw/cp.aspx?n=5590

Data Sources (2 of 5)

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 (3 of 5)

Data Available From Selected Government Agencies

U.S. 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 (4 of 5)

Statistical Studies – Observational

- In observational (nonexperimental) studies （觀察研究）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.

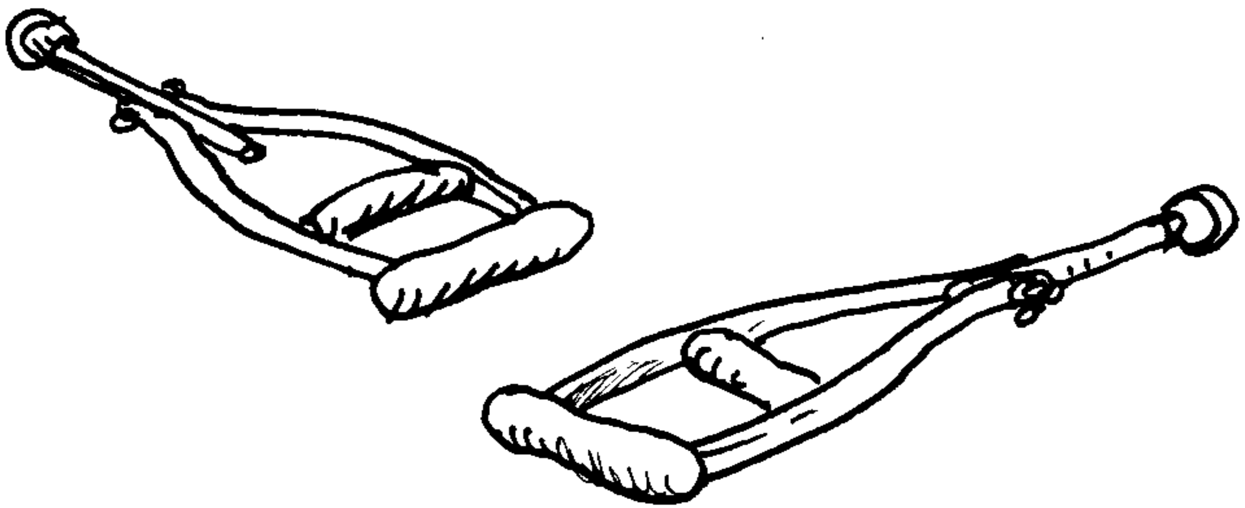
Data Sources (5 of 5)

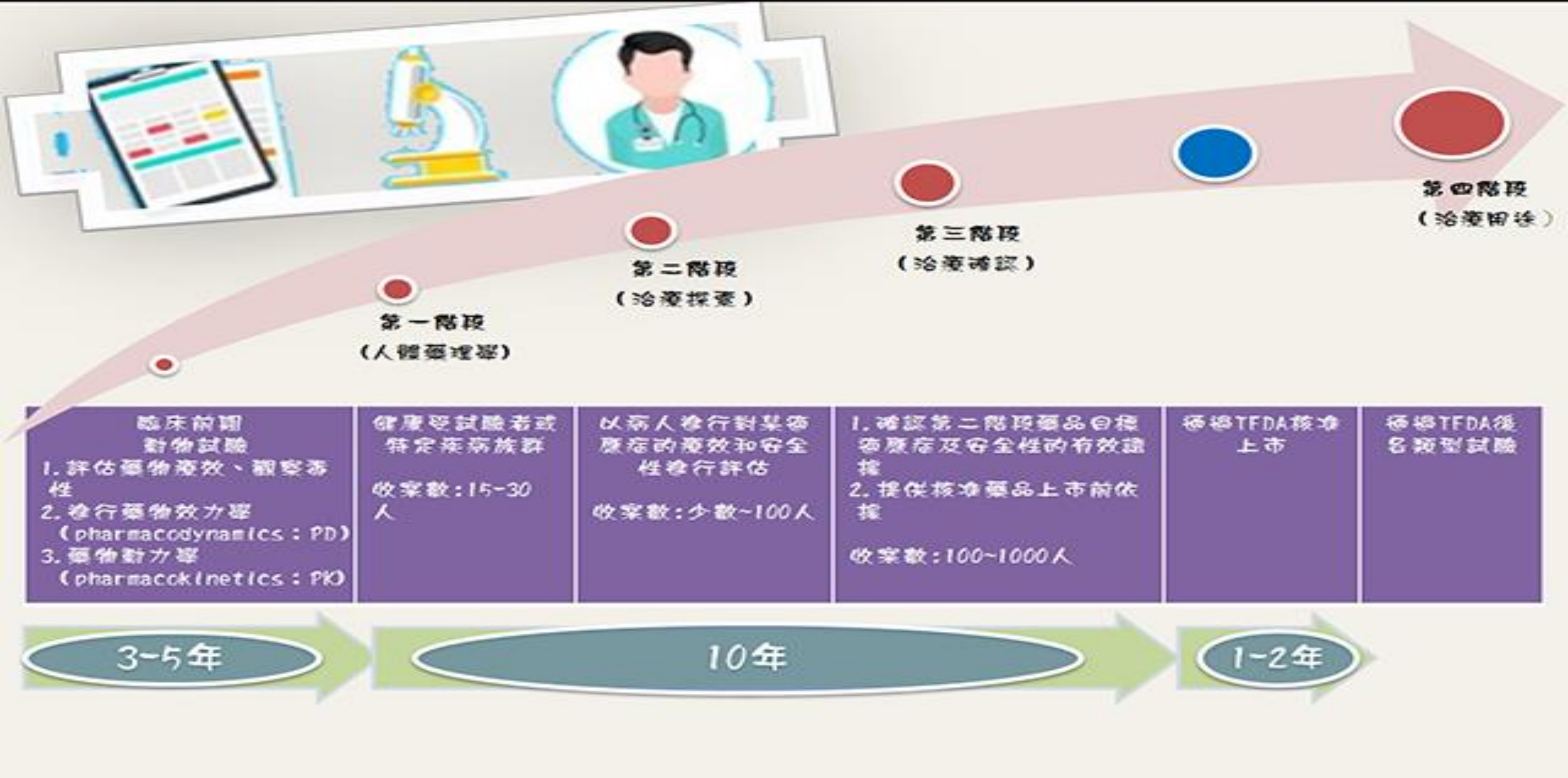
Statistical Studies – Experimental

- In experimental studies （實驗設計）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. Nearly two million U.S. children (grades 1- 3) were selected.

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.





https://www.cgh.org.tw/hipershop/rwd1320/store/F1/dmr8-1(1).png

Center of Clinical Trial Research . Cathay General Hospital

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.

Chapter 2 - Descriptive Statistics: Tabular and Graphical Displays

2.1 - Summarizing Data for a Categorical Variable

- Categorical data use labels or names to identify categories of like items.

2.2 - Summarizing Data for a Quantitative Variable

- Quantitative data are numerical values that indicate how much or how many.

2.3 - Summarizing Data for Two Variables Using Tables

2.4 - Summarizing Data for Two Variables Using Graphical Displays

2.5 - Data Visualization: Best Practices in Creating Effective Graphical Displays

Summarizing Categorical Data

- Frequency Distribution
- Relative Frequency Distribution
- Percent Frequency Distribution
- Bar Chart
- Pie Chart

Frequency Distribution

A frequency distribution is a tabular summary of data showing the number (frequency) of observations in each of several non-overlapping categories or classes.

Example: Marada Inn

Guests staying at the Marada Inn were asked to rate the quality of their accommodations as being *excellent*, *above average*, *average*, *below average*, or *poor*.

Rating	Frequency
Poor	2
Below Average	3
Average	5
Above Average	9
Excellent	1
Total	20

Relative Frequency and Percent Frequency Distributions (1 of 2)

- The relative frequency of a class is the fraction or proportion of the total number of data items belonging to the class.

$$\text{Relative frequency} = \frac{\text{Frequency}}{n}$$

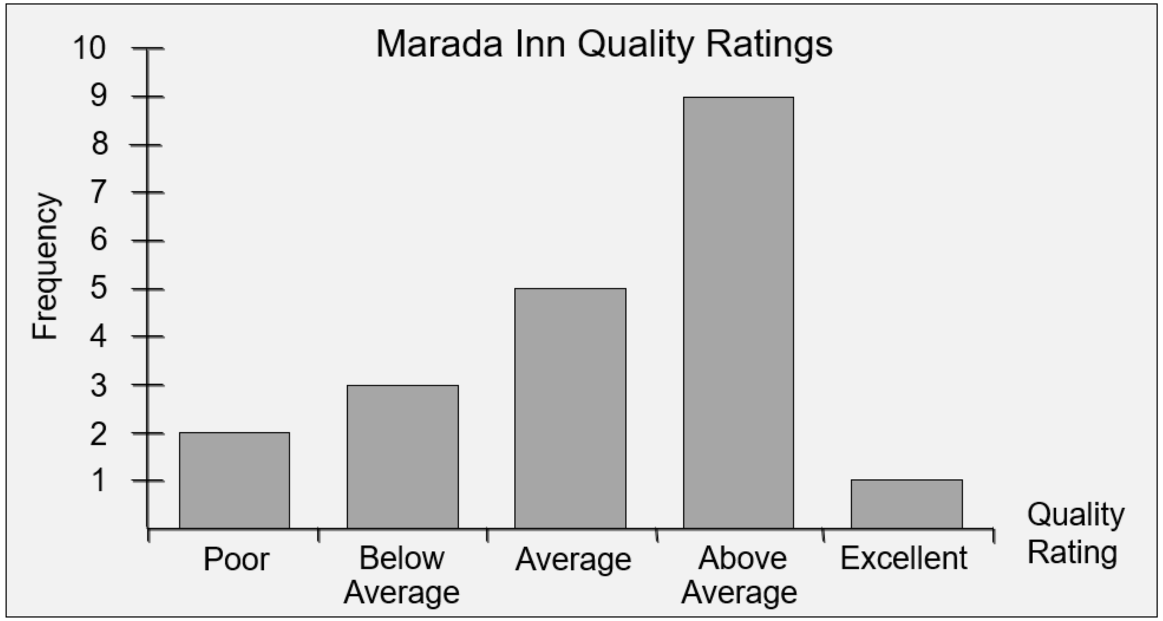
- The percent frequency of a class is the relative frequency multiplied by 100.

Example: Marada Inn

Rating	Relative Frequency	Percent Frequency
Poor	0.10	10%
Below Average	0.15	15%
Average	0.25	25%
Above Average	0.45	45%
Excellent	0.05	5%
Total	1.00	100%

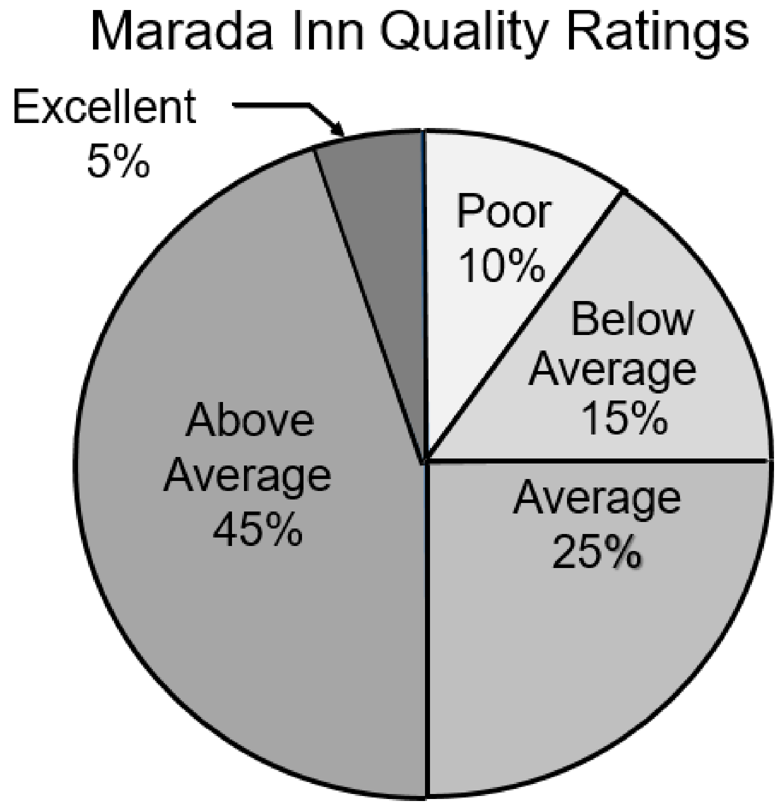
Bar Chart （長條圖）

- A bar chart is a graphical display for depicting qualitative data.
- A frequency, relative frequency, or percent frequency scale can be used for the other axis (usually the vertical axis).
- Using a bar of fixed width drawn above each class label, we extend the height appropriately.
- The bars are separated to emphasize the fact that each class is a separate category.



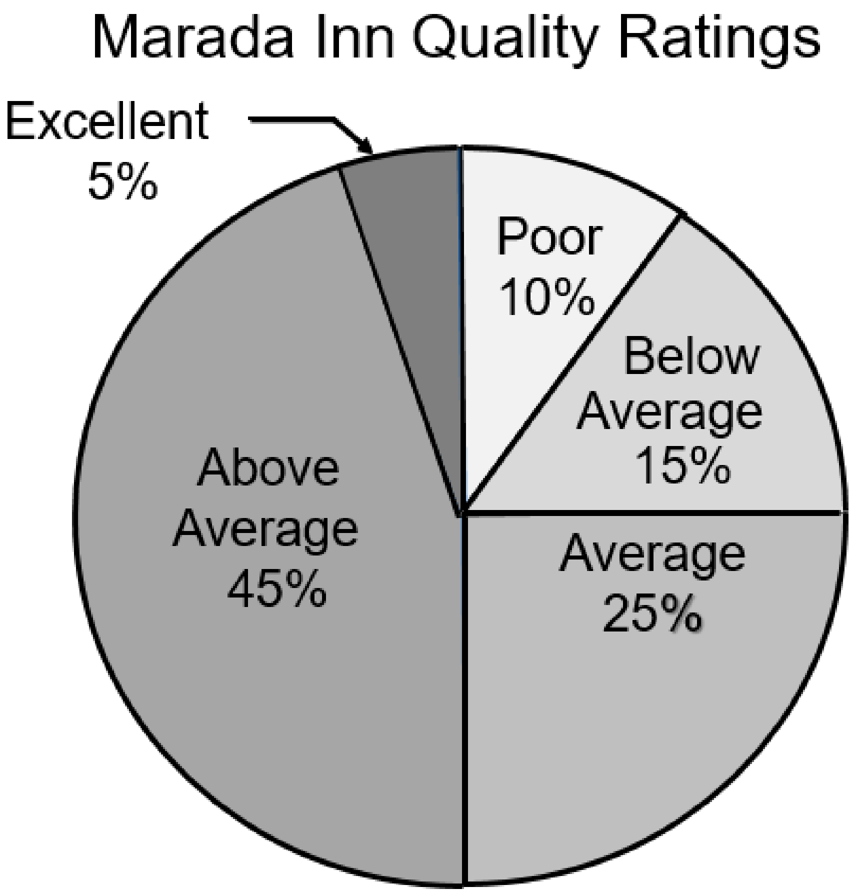
Pie Chart （圓餅圖）

- The pie chart is a commonly used graphical display for presenting relative frequency and percent frequency distributions for categorical data.
- First draw a circle; then use the relative frequencies to subdivide the circle into sectors that correspond to the relative frequency for each class.
- Because there are 360 degrees in a circle, a class with a relative frequency of 0.25 would consume 0.25(360) = 90 degrees of the circle.



Example: Marada Inn

- Half of the customers surveyed gave Marada a quality rating of “above average” or “excellent” (look at the left side of the pie). This might please the manager.
- For each customer who gave an “excellent” rating, there were two customers who gave a “poor” rating (looking at the top of the pie). This should displease the manager.



Summarizing Quantitative Data

- Frequency Distribution
- Relative Frequency and Percent Frequency Distributions
- Dot Plot
- Histogram
- Cumulative Distributions
- Stem-and-Leaf Display

Frequency Distribution – Quantitative Data (1 of 2)

The manager of Hudson Auto would like to gain a better understanding of the cost of parts used in the engine tune-ups performed in the shop. She examines 50 customer invoices for tune-ups. The costs of parts, rounded to the nearest dollar, are shown below.

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

Frequency Distribution – Quantitative Data (2 of 2)

Example: Hudson Auto Repair

If we choose six classes the approximate class width = (109 – 50)/6 = 9.83 or about 10.

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

Part Cost ($)	Frequency
50-59	2
60-69	13
70-79	16
80-89	7
90-99	7
100-109	5
Total	50

Relative Frequency and Percent Frequency Distributions (2 of 2)

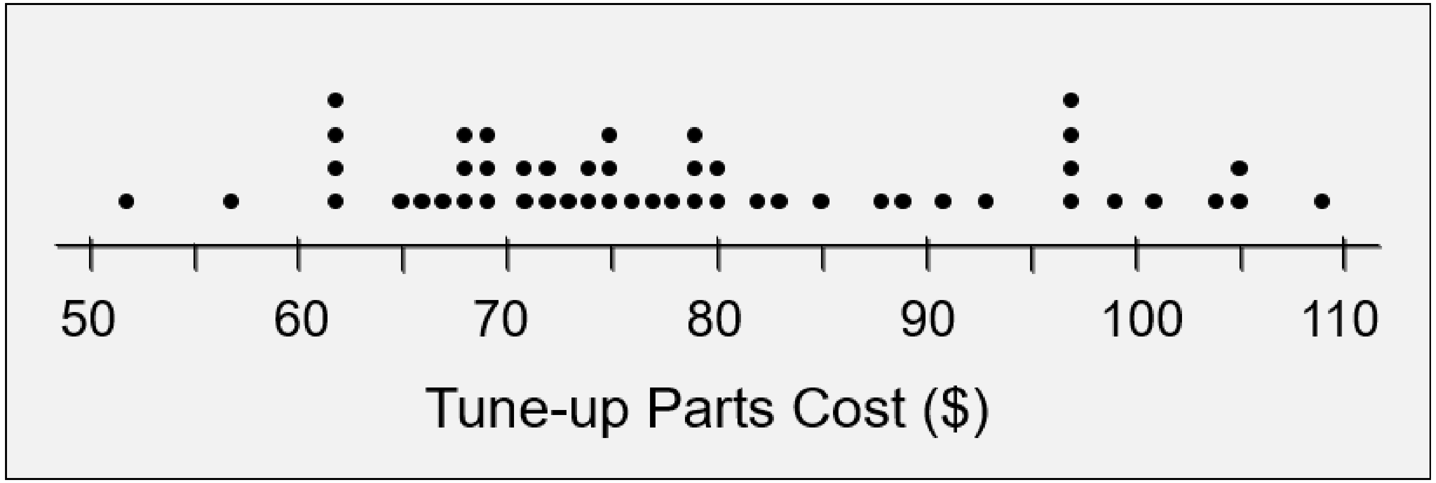
Insights

- Only 4% of the parts costs are in the $50-59 class.
- 30% of the parts costs are under $70.
- The greatest percentage (32% or almost one-third) of the parts costs are in the $70-79 class.
- 10% of the parts costs are $100 or more.

Parts Cost ($)	Relative Frequency	Percent Frequency
50-59	0.04 = 2/50	4 = .04(100)
60-69	0.26	26
70-79	0.32	32
80-89	0.14	14
90-99	0.14	14
100-109	0.10	10
Total	1.00	100

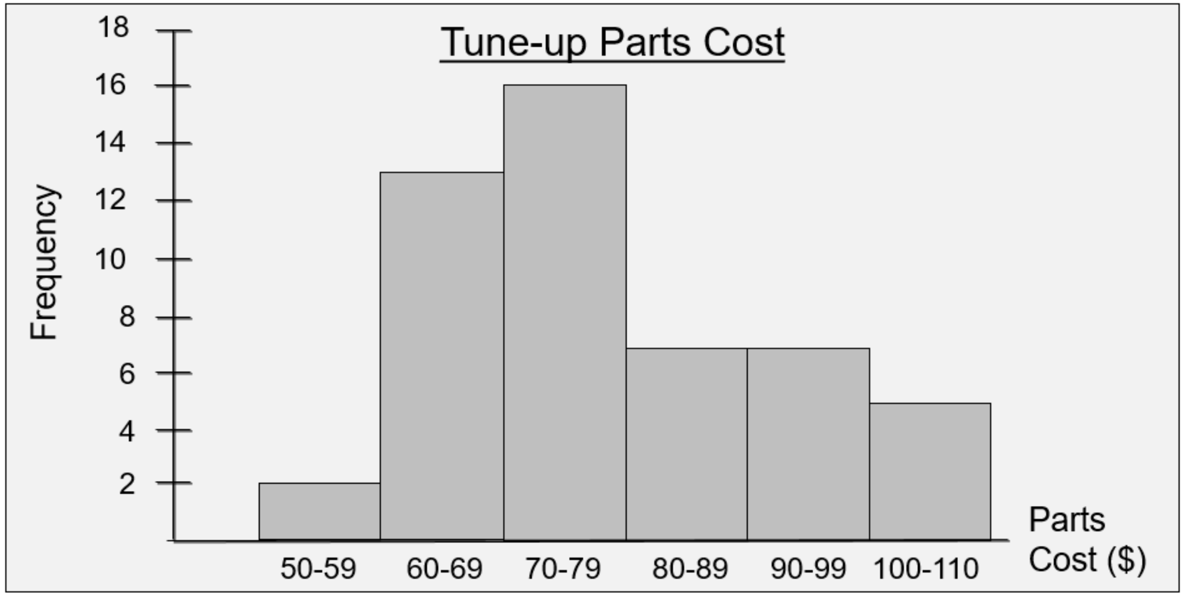
Dot Plot

- One of the simplest graphical summaries of data is a dot plot.
- A horizontal axis shows the range of data values.
- Then each data value is represented by a dot placed above the axis.



Histogram（直方圖）

- The variable of interest is placed on the horizontal axis.
- A rectangle is drawn above each class interval with its height corresponding to the interval's frequency, relative frequency, or percent frequency.
- Unlike a bar graph, a histogram has no natural separation between rectangles of adjacent classes.

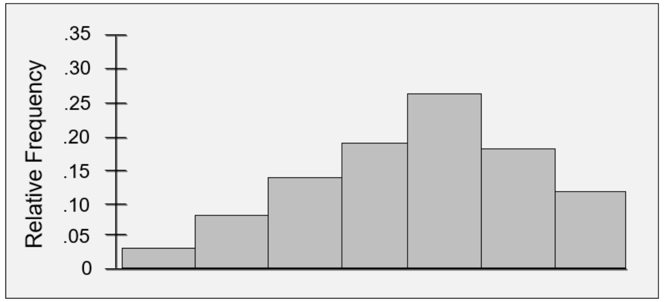


Histograms Showing Skewness （偏度）

Moderately Skewed Left

A longer tail to the left

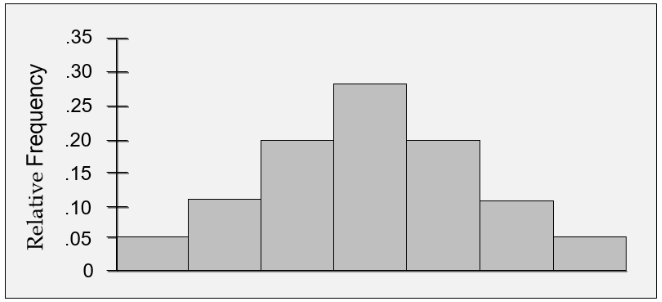
Ex: Exam Scores



Symmetric

Left tail is the mirror image of the right tail

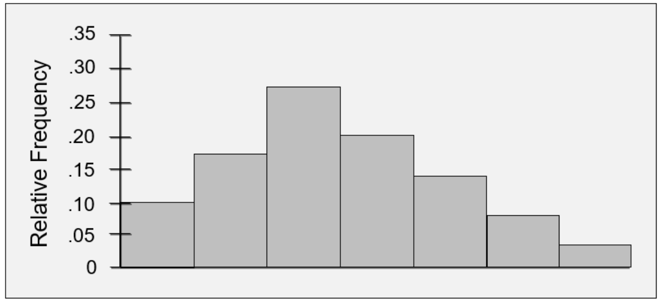
Ex: Heights of People



Moderately Right Skewed

A Longer tail to the right

Ex: Housing Values



Cumulative Distributions（累積次數分配）

Cumulative frequency distribution – shows the *number* of items with values less than or equal to the upper limit of each class.

Cumulative relative frequency distribution – shows the *proportion* of items with values less than or equal to the upper limit of each class.

Cumulative percent frequency distribution – shows the *percentage* of items with values less than or equal to the upper limit of each class.

Hudson Auto Repair

Cost ($)	Cumulative Frequency	Cumulative Relative Frequency	Cumulative Percent Frequency
≤ 59	2	.04	4
≤ 69	15 = 2+13	.30 = 15/50	30 = .30(100)
≤ 79	31	.62	62
≤ 89	38	.76	76
≤ 99	45	.90	90
≤ 109	50	1.00	100

Stem-and-Leaf Display (1 of 3)（枝葉圖）

- A stem-and-leaf display shows both the rank order and shape of a distribution of data.
- It is similar to a histogram on its side, but it has the advantage of showing the actual data values.
- The leading digits of each data item are arranged to the left of a vertical line.
- To the right of the vertical line we record the last digit for each item in rank order.
- Each line (row) in the display is referred to as a <u>stem</u>.
- Each digit on a stem is a <u>leaf</u>.

Stems	Leaves
5	2 7
6	2 2 2 2 5 6 7 8 8 8 9 9 9
7	1 1 2 2 3 4 4 5 5 5 6 7 8 9 9 9
8	0 0 2 3 5 8 9
9	1 3 7 7 7 8 9
10	1 4 5 5 9

Stem-and-Leaf Display (2 of 3)

Leaf Units

- A single digit is used to define each leaf.
- In the preceding example, the leaf unit was 1.
- Leaf units may be 100, 10, 1, 0.1, and so on.
- Where the leaf unit is not shown, it is assumed to equal 1.
- The leaf unit indicates how to multiply the stem-and-leaf numbers in order to approximate the original data.

If we have data with values such as
8.6 11.7 9.4 9.1 10.2 11.0 8.8

Leaf Unit = 0.1

8	6 8
9	1 4
10	2
11	0 7

Stretched Stem-and-Leaf Display

- If we believe the original stem-and-leaf display has condensed the data too much, we can stretch the display vertically by using two stems for each leading digit(s).
- Whenever a stem value is stated twice, the first value corresponds to leaf values of 0 - 4, and the second value corresponds to leaf values of 5 - 9.

5	2
5	7
6	2 2 2 2
6	5 6 7 8 8 8 9 9 9
7	1 1 2 2 3 4 4
7	5 5 5 6 7 8 9 9 9
8	0 0 2 3
8	5 8 9
9	1 3
9	7 7 7 8 9
10	1 4
10	5 5 9

Stem-and-Leaf Display (3 of 3)

If we have data values such as

1806, 1717, 1974, 1791, 1682, 1910, and 1838

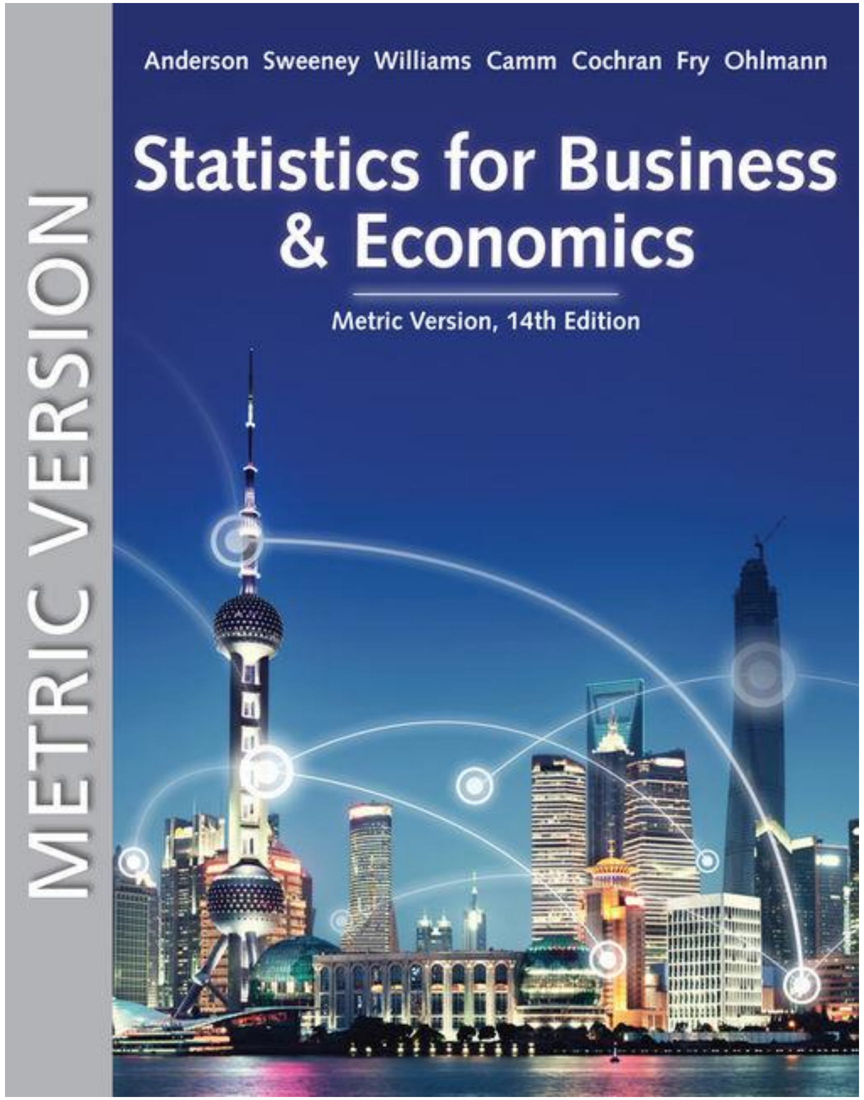
Leaf Unit = 10

16	8
17	1 9
18	0 3
19	1 7

The 82 in 1682 is rounded down to 80 and is represented as an 8.

Statistics for
Business and Economics (14e)
Metric Version

Chapters 3, 5, 6



Chapter 3 - Descriptive Statistics: Numerical Measures

3.1 - Measures of Location

3.2 - Measures of Variability

3.3 - Measures of Distribution Shape, Relative Location, and Detecting Outliers

3.4 - Five-Number Summaries and Box Plots

3.5 - Measures of Association Between Two Variables

3.6 - Data Dashboards: Adding Numerical Measures to Improve Effectiveness

統計的分析觀點

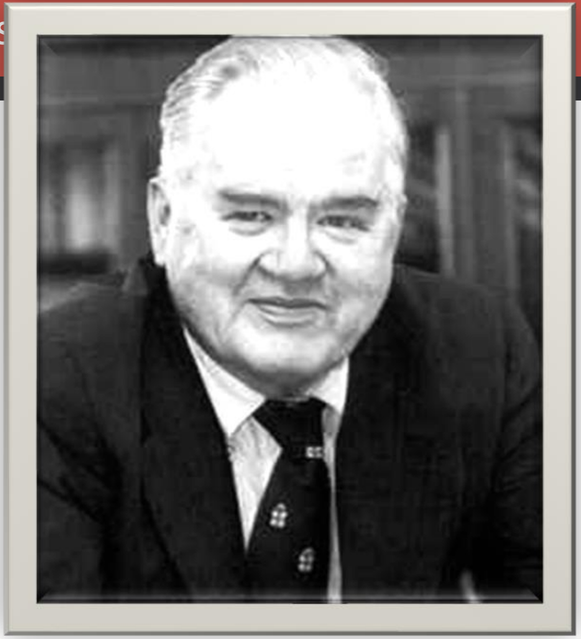
根據統計觀點，分析有以下兩類：

- 探索性資料分析(Exploratory Data Analysis)

→The role of **EDA** is to figure out the essence of data and to develop research hypothesis,

- 驗證性資料分析(Confirmatory Data Analysis)

→While the role of **CDA** is to examine evidence and test hypothesis & build models.



EDA：讓資料說話

- 資料驅動(Data Driven)

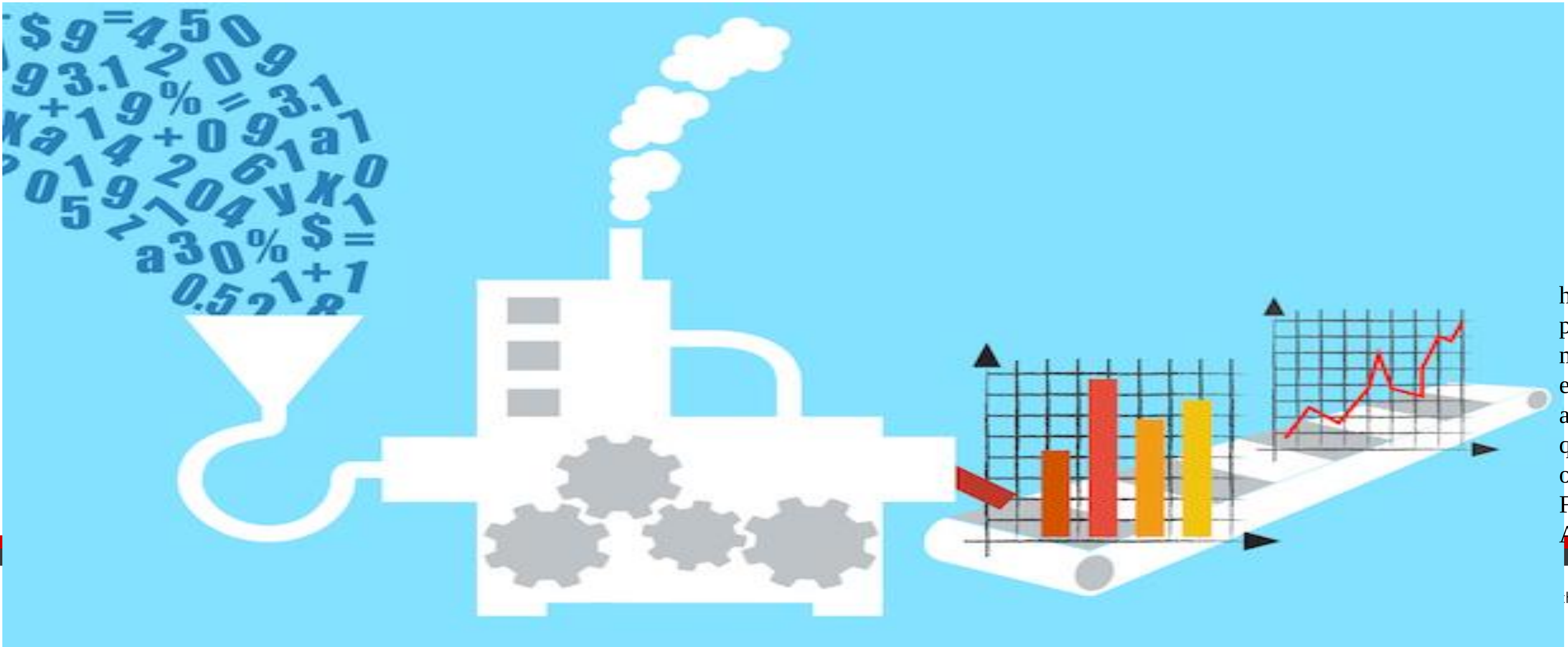
→Tukey於1970年代提出EDA，他認為

"more emphasis needed to be placed on using data to construct research hypotheses"

→EDA is not a mere collection of techniques. EDA is a philosophy as to **how we dissect a data set**; **what we look for**; **how we look**; and **how we interpret**.

探索性資料分析(資料驅動)

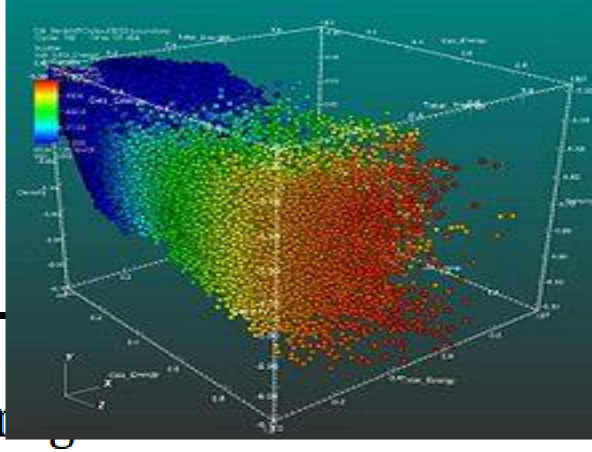
Exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics … EDA is for seeing what the data can tell us beyond the formal modeling. ---Wikipedia



https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.aiche.org%2Facademy%2Fwebinars%2Fapplied-statistics-exploratory-data-analysis&psig=AOvVaw36ZuxAJqz27dLqU5IFzBMO&ust=1570108849384000&source=images&cd=vfe&ved=0CAIQjRxqFwoTCJC1qLXV_eQCFQAAAAAdAAAAABAJ

WIKIPEDIA

https://upload.wikimedia.org/wikipedia/commons/c/c4/Scatter_plot.jpg

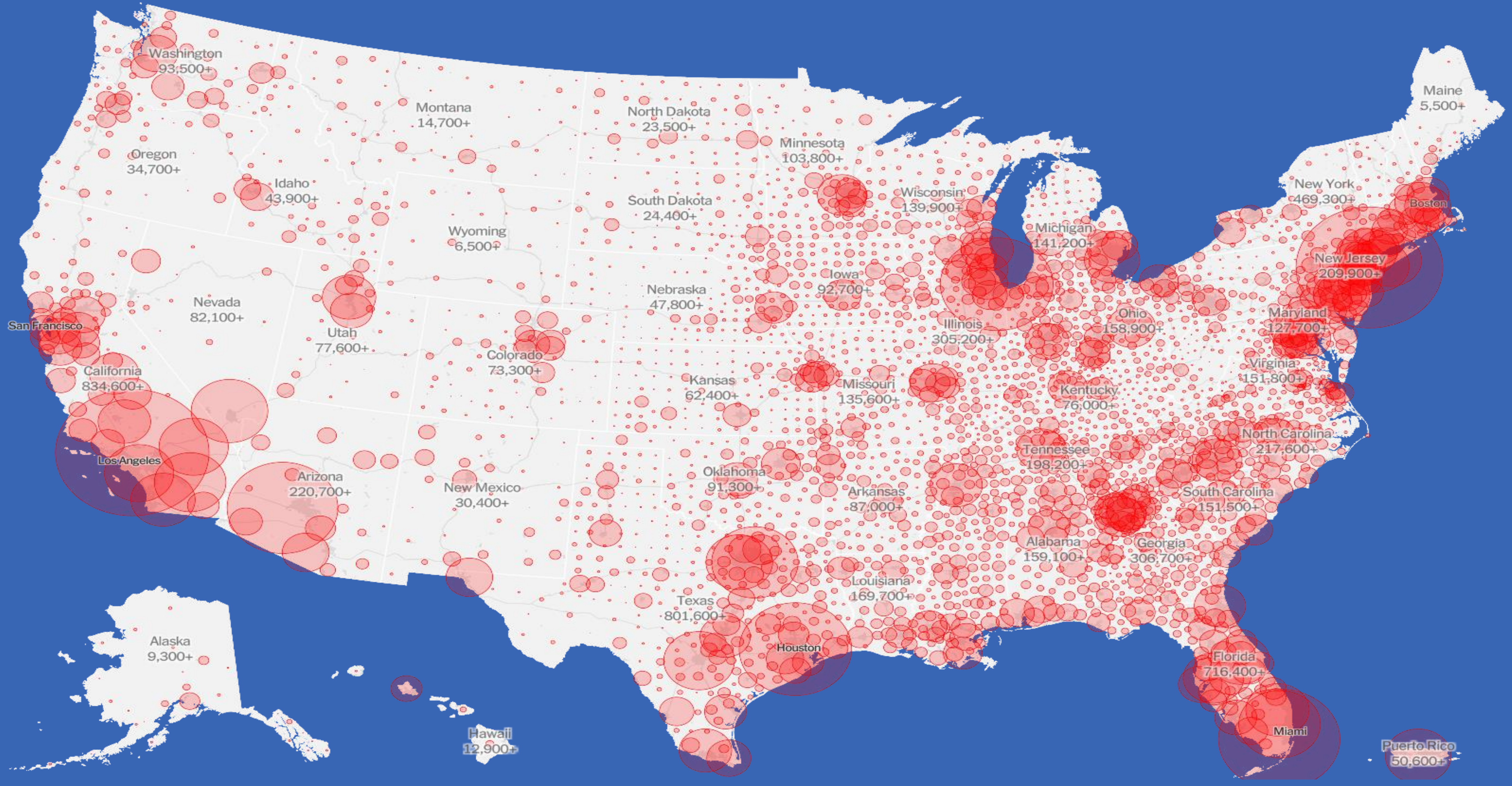


Data visualization

Data visualization is the graphic representation of data. It involves producing im relationships among the represented data to viewers of the images. This communication is achieved through the use of a systematic mapping between graphic marks and data values in the creation of the visualization. This mapping establishes how data values will be represented visually, determining how and to what extent a property of a graphic mark, such as size or color, will change to reflect changes in the value of a datum.

To communicate information clearly and efficiently, data visualization uses statistical graphics, plots, information graphics and other tools. Numerical data may be encoded using dots, lines, or bars, to visually communicate a quantitative message.[1] Effective visualization helps users analyze and reason about data and evidence. It makes complex data more accessible, understandable and usable. Users may have particular analytical tasks, such as making comparisons or understanding causality, and the design principle of the graphic (i.e., showing comparisons or showing causality) follows the task. Tables are generally used where users will look up a specific measurement, while charts of various types are used to show patterns or relationships in the data for one or more variables.

美國各地Covid-19確診數（紐約時報）



https://www.nytimes.com/interactive/2020/us/coronavirus-us-cases.html

REMAIN
48.11%

LEAVE
51.89%

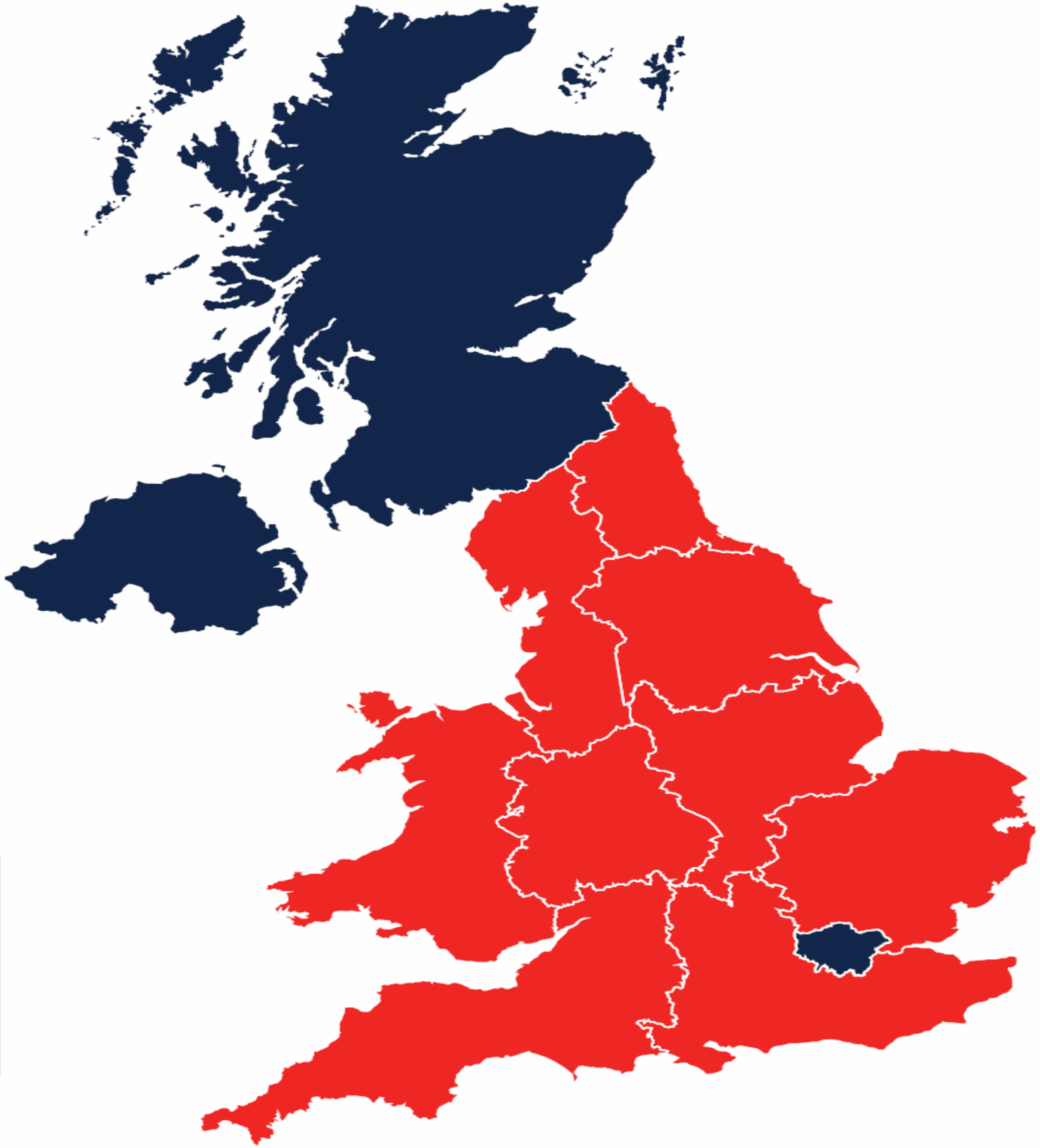
382 / 382 districts reporting

Breakdown by region

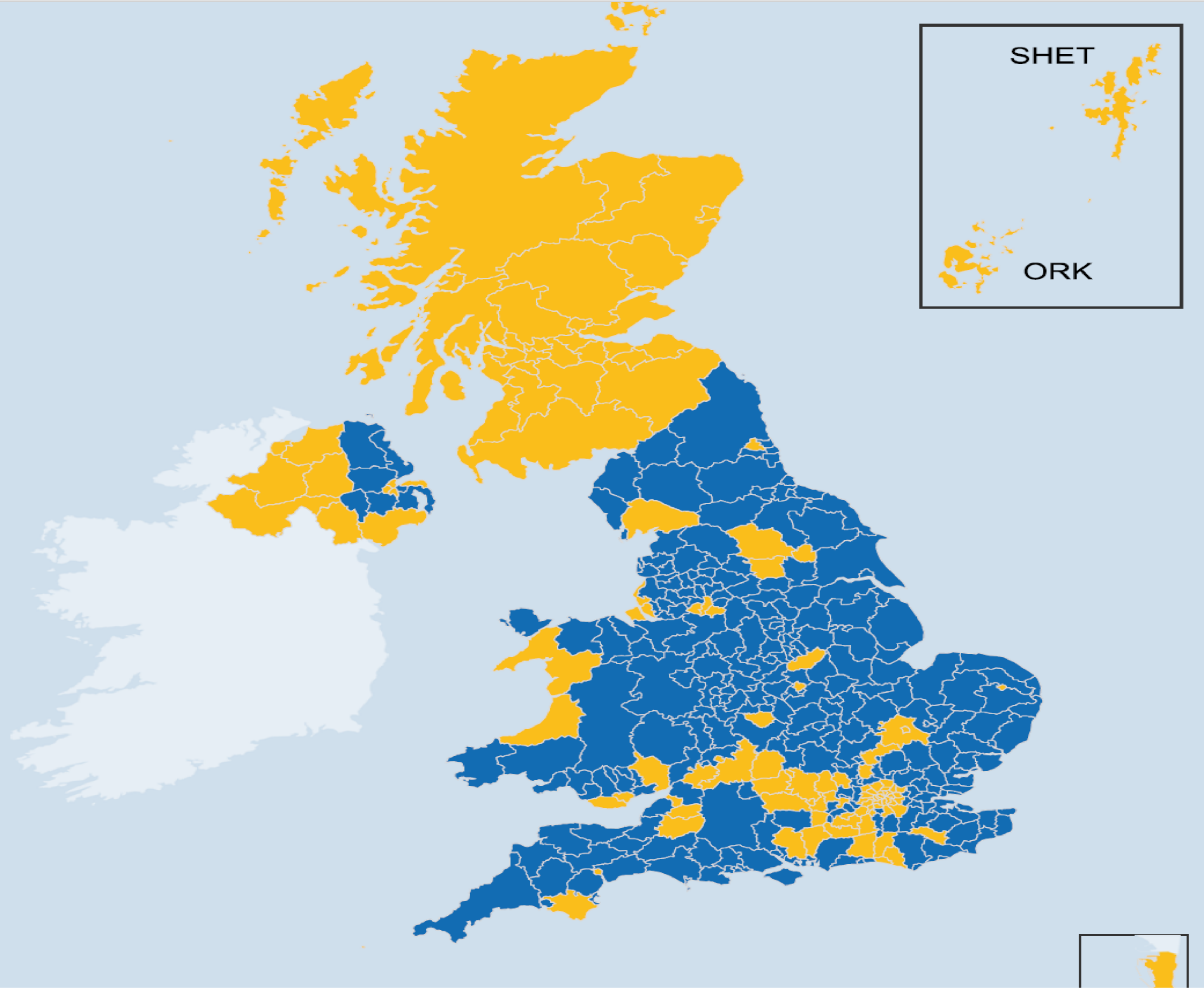
South East

Remain	2,391,718
Leave	2,567,965

67 / 67 districts reporting



英國脫歐公投結果
（CNN）



英國脫歐公投結果(BBC)

England

Leave 53.4%
15,188,406 VOTES

Northern Ireland

Leave 44.2%
349,442 VOTES

Scotland

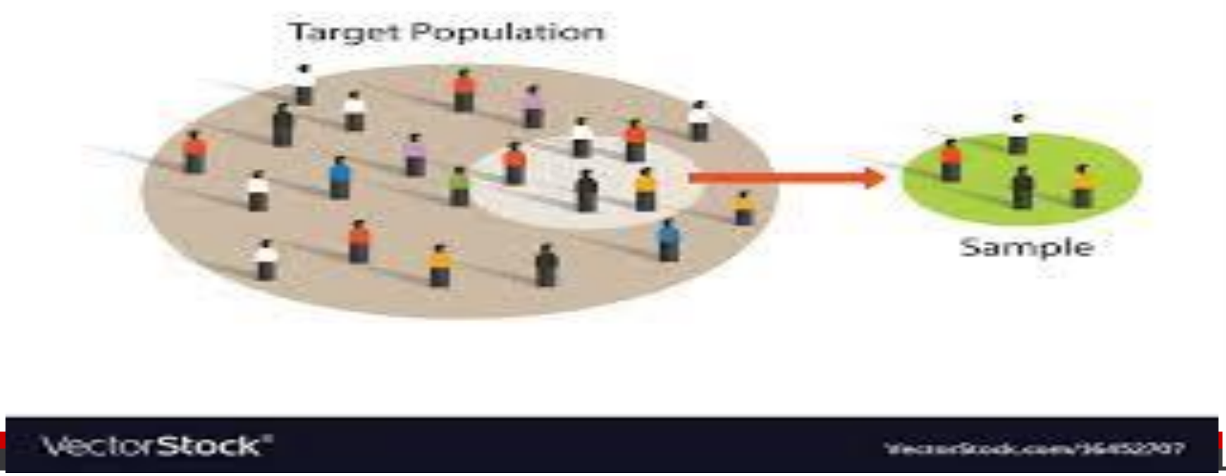
Leave 38.0%
1,018,322 VOTES

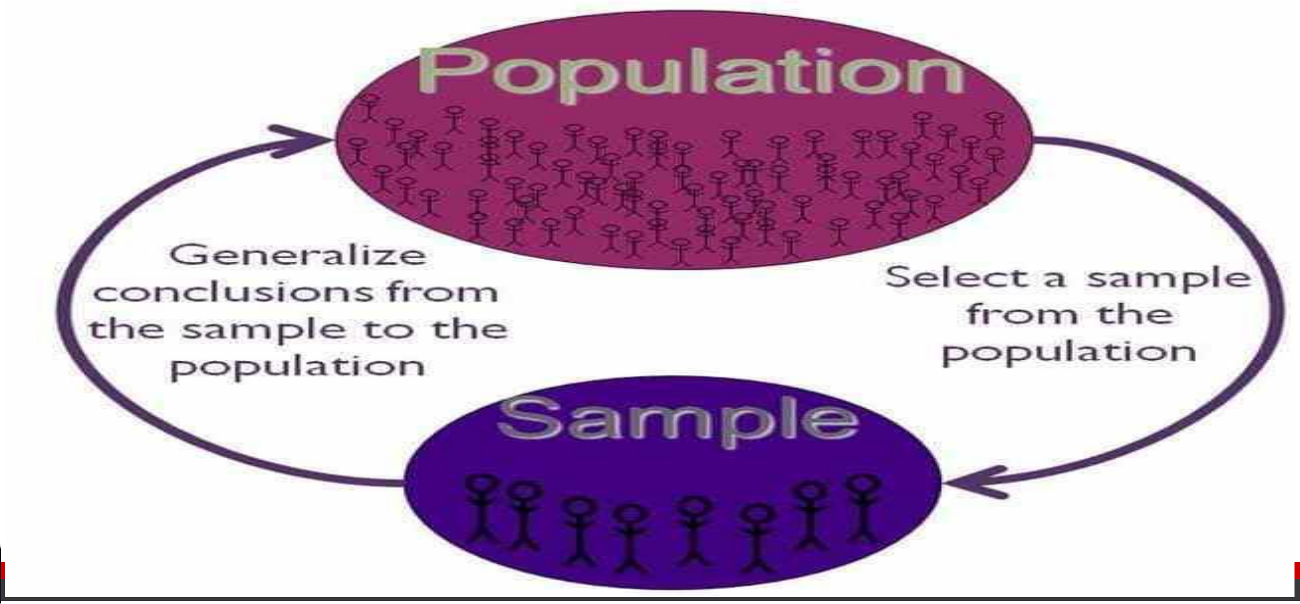
Wales

Leave 52.5%
854,572 VOTES

Numerical Measures

- If the measures are computed for data from a sample, they are called sample statistics.
- If the measures are computed for data from a population, they are called population parameters.
- A sample statistic is referred to as the point estimator of the corresponding population parameter.



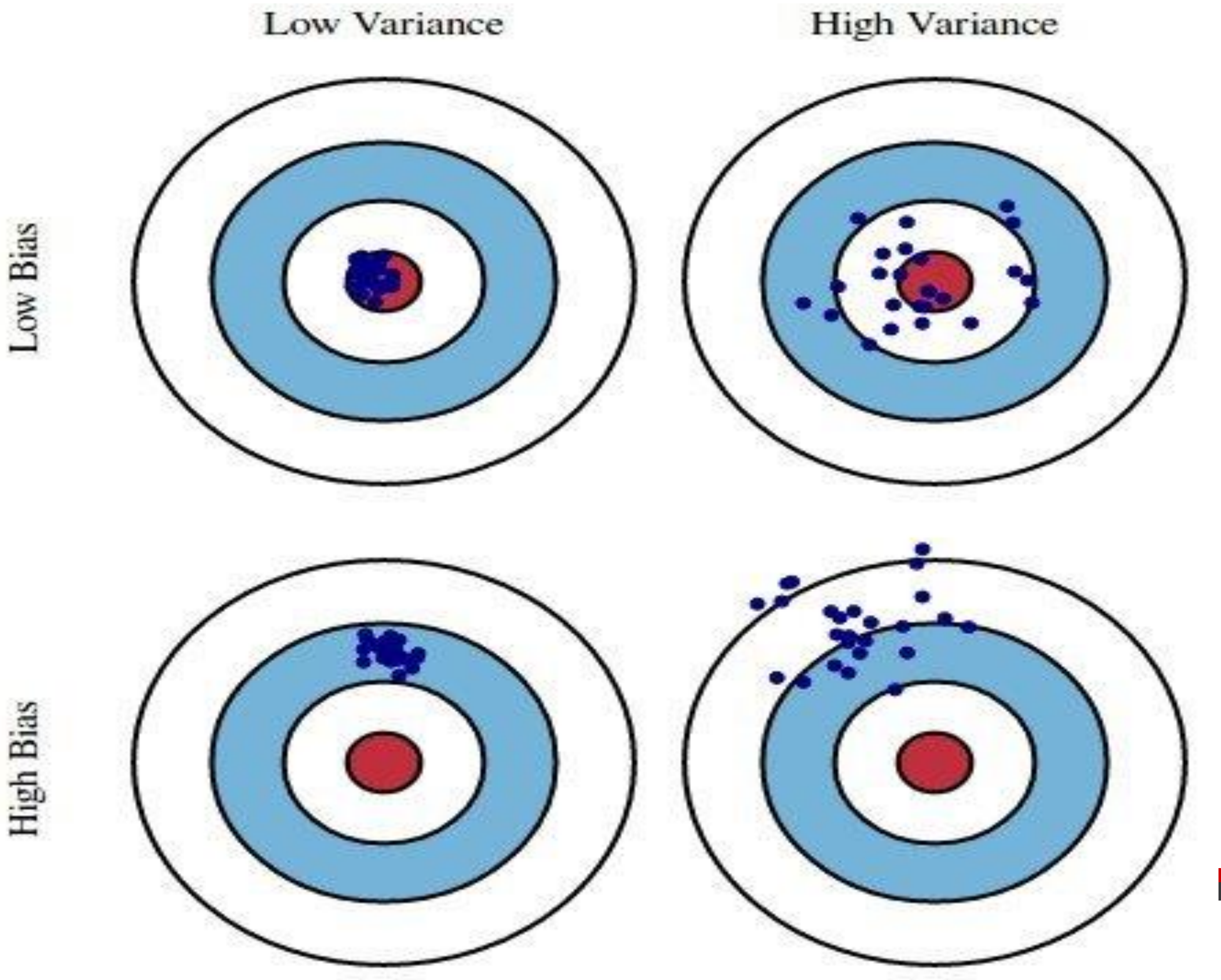


https://previews.123rf.com/images/bakhtiarzein/bakhtiarzein1708/bakhtiarzein170800001/83182787-sample-from-population-statistics-research-survey-methodology-selection-concept.jpg

https://encrypted-tbn0.gstatic.com/images?q=tbn:ANd9GcTtOH1le3m9uwwqpuIIiyeBRkSxUnu4_Tw7JA&usqp=CAU

Measures of Location（集中趨勢量數）

- Mean
- Median
- Mode
- Weighted Mean
- Geometric Mean
- Percentiles
- Quartiles



Mean

- Perhaps the most important measure of location is the mean.
- The mean provides a measure of central location.
- The mean of a data set is the average of all the data values.
- The sample mean $\bar{x}$ is the point estimator of the population mean, μ.

$$\bar{x} = \frac{\sum x_i}{n}$$

where $\sum x_i$ = the sum of the values of the n observations and

n = the number of observations in the sample.

Sample Mean $\bar{x}$

Seventy efficiency apartments were randomly sampled in a college town. The monthly rents for these apartments are listed below.

545	715	530	690	535	700	560	700	540	715
540	540	540	625	525	545	675	545	550	550
565	550	625	550	550	560	535	560	565	580
550	570	590	572	575	575	600	580	670	565
700	585	680	570	590	600	649	600	600	580
670	615	550	545	625	635	575	650	580	610
610	675	590	535	700	535	545	535	530	540

$$\bar{x} = \frac{\sum x_i}{n} = \frac{41{,}356}{70} = 590.80$$

Median (1 of 4)

- The median of a data set is the value in the middle when the data items are arranged in ascending order.
- Whenever a data set has extreme values, the median is the preferred measure of central location.
- The median is the measure of location most often reported for annual income and property value data.
- A few extremely large incomes or property values can inflate the mean.

Median (2 of 4)

Here we have an odd number of observations:

7 observations: 26, 18, 27, 12, 14, 27, and 19.

Rewritten in ascending order: 12, 14, 18, 19, 26, 27, and 27.

The median is the middle value in this list, so the median = 19.

Median (3 of 4)

Here we have an even number of observations:

8 observations: 26, 18, 27, 12, 14, 27, 19, and 30.

Rewritten in ascending order: 12, 14, 18, 19, 26, 27, 27, and 30.

The median is the average of the two middle values in this list, so the median = (19 + 26)/2 = 22.5.

Median (4 of 4)

Example: Apartment Rents

Notice that there are 70 values provided which are in ascending order.

Averaging the 35th and 36th values: Median (575 + 575)/2 = 575.

525	530	530	535	535	535	535	535	540	540
540	540	540	545	545	545	545	545	550	550
550	550	550	550	550	560	560	560	565	565
565	570	570	572	575	575	575	580	580	580
580	585	590	590	590	600	600	600	600	610
610	615	625	625	625	635	649	650	670	670
675	675	680	690	700	700	700	700	715	715

Mode

- The mode of a data set is the value that occurs with greatest frequency.
- The greatest frequency can occur at two or more different values.
- If the data have exactly two modes, the data are bimodal.
- If the data have more than two modes, the data are multimodal.

The mode is 550.

525	530	530	535	535	535	535	535	540	540
540	540	540	545	545	545	545	545	550	550
550	550	550	550	550	560	560	560	565	565
565	570	570	572	575	575	575	580	580	580
580	585	590	590	590	600	600	600	600	610
610	615	625	625	625	635	649	650	670	670
675	675	680	690	700	700	700	700	715	715

Weighted Mean (1 of 3)

- In some instances, the mean is computed by giving each observation a weight that reflects its relative importance.
- The choice of weights depends on the application.
- The weights might be the number of credit hours earned for each grade, as in GPA.
- In other weighted mean computations, quantities such as kilograms, dollars, or volume are frequently used.

$$\bar{x} = \frac{\sum w_i x_i}{\sum w_i}$$

where: x_i = value of observation i

w_i = weight for observation i

Weighted Mean (2 of 3)

Ron Butler, a home builder, is looking over the expenses he incurred for a house he just built. For the purpose of pricing future projects, he would like to know the average wage ($/hour) he paid the workers he employed. Listed below are the categories of workers he employed, along with their respective wage and total hours worked.

Worker	Wage ($/hr)	Total Hours
Carpenter	21.60	520
Electrician	28.72	230
Laborer	11.80	410
Painter	19.75	270
Plumber	24.16	160

Weighted Mean (3 of 3)

Example: Construction Wages

Worker	x_i	w_i	$w_i x_i$
Carpenter	21.60	520	11232.0
Electrician	28.72	230	6605.6
Laborer	11.80	410	4838.0
Painter	19.75	270	5332.5
Plumber	24.16	160	3865.6
		1590	31873.7

$$\bar{x} = \frac{\sum w_i x_i}{\sum w_i} = \frac{31{,}873.7}{1{,}590} = 20.0464 = \$20.05$$

FYI, the equally-weighted (simple) mean = $21.21

Geometric Mean (1 of 2)

- The geometric mean is calculated by finding the *n*th root of the product of *n* values.
- It is often used in analyzing growth rates in financial data (where using the arithmetic mean will provide misleading results).
- It should be applied anytime you want to determine the mean rate of change over several successive periods (be it years, quarters, weeks, . . .).
- Other common applications include: changes in populations of species, crop yields, pollution levels, and birth and death rates.

$$\bar{x}_g = \sqrt[n]{(x_1)(x_2)\ldots(x_n)}$$

$$= [(x_1)(x_2)\ldots(x_n)]^{1/n}$$

Geometric Mean (2 of 2)

Example: Rate of Return

Period	Return (%)
1	-6.0
2	-8.0
3	-4.0
4	2.0
5	5.4

Growth Factor
0.940
0.920
0.960
1.020
1.054

$$\bar{x}_g = \sqrt[5]{(0.94)(0.92)(1.02)(1.054)} = (0.89254)^{1/5} = 0.97752$$

The average growth rate per period is (0.97752 – 1)(100) = –2.248%.

Percentiles

- A percentile provides information about how the data are spread over the interval from the smallest value to the largest value.
- Admission test scores for colleges and universities are frequently reported in terms of percentiles.
- The p^{th} percentile of a data set is a value such that at least *p* percent of the items take on this value or less and at least $(100 - p)$ percent of the items take on this value or more.
- Arrange the data in ascending order.
- Compute L_p, the location of the p^{th} percentile.

$$L_p = \left(\frac{p}{100}\right)(n+1)$$

80th Percentile

Example: Apartment Rents $$L_p = \left(\frac{p}{100}\right)(n+1) = \left(\frac{80}{100}\right)(70+1) = 56.8$$

The 80th percentile is the 56th value plus 0.8 times the difference between the 57th and 56th values.

So the 80th percentile = 635 + 0.8(649 – 635) = 646.2.

525	530	530	535	535	535	535	535	540	540
540	540	540	545	545	545	545	545	550	550
550	550	550	550	550	560	560	560	565	565
565	570	570	572	575	575	575	580	580	580
580	585	590	590	590	600	600	600	600	610
610	615	625	625	625	635	649	650	670	670
675	675	680	690	700	700	700	700	715	715

80th Percentile, Part 2

Example: Apartment Rents

"At least 80% of the items take on a value of 646.2 or less."

"At least 20% of the items take on a value of 646.2 or more."

56/70 = .8 or 80%

14/70 = .2 or 20%

525	530	530	535	535	535	535	535	540	540
540	540	540	545	545	545	545	545	550	550
550	550	550	550	550	560	560	560	565	565
565	570	570	572	575	575	575	580	580	580
580	585	590	590	590	600	600	600	600	610
610	615	625	625	625	635	649	650	670	670
675	675	680	690	700	700	700	700	715	715

Quartiles

Quartiles are specific percentiles.

1. First Quartile = 25th Percentile
2. Second Quartile = 50th Percentile = Median
3. Third Quartile = 75th Percentile

Third Quartile (75^{th} Percentile)

Example: Apartment Rents

$$L_p = \left(\frac{p}{100}\right)(n+1) = \left(\frac{75}{100}\right)(70+1) = 53.25$$

The 75^{th} percentile is the 53^{rd} value plus 0.25 times the difference between the 54^{th} and 53^{rd} values.

The 75^{th} percentile = third quartile = 625 + 0.25(625 – 625) = 625.

525	530	530	535	535	535	535	535	540	540
540	540	540	545	545	545	545	545	550	550
550	550	550	550	550	560	560	560	565	565
565	570	570	572	575	575	575	580	580	580
580	585	590	590	590	600	600	600	600	610
610	615	625	625	625	635	649	650	670	670
675	675	680	690	700	700	700	700	715	715

Measures of Variability （散佈趨勢量數）

- It is often desirable to consider measures of variability (dispersion), as well as measures of location.
- For example, in choosing supplier A or supplier B we might consider not only the average delivery time for each, but also the variability in delivery time for each.
- Common measures of variability are:
  - Range
  - Interquartile Range
  - Variance
  - Standard Deviation
  - Coefficient of Variation

Range

- The range of a data set is the difference between the largest and smallest data value.
- It is the simplest measure of variability.
- It is very sensitive to the smallest and largest data values.

Range = largest value – smallest value = 715 – 525 = 190.

525	530	530	535	535	535	535	535	540	540
540	540	540	545	545	545	545	545	550	550
550	550	550	550	550	560	560	560	565	565
565	570	570	572	575	575	575	580	580	580
580	585	590	590	590	600	600	600	600	610
610	615	625	625	625	635	649	650	670	670
675	675	680	690	700	700	700	700	715	715

Interquartile Range (IQR)

- The interquartile range of a data set is the difference between the third quartile and the first quartile.
- It is the range for the middle 50% of the data.
- It overcomes the sensitivity to extreme data values.

3rd Quartile (Q_3) = 625

1st Quartile (Q_1) = 545

IQR = 625 – 545 = 80

525	530	530	535	535	535	535	535	540	540
540	540	540	545	545	545	545	545	550	550
550	550	550	550	550	560	560	560	565	565
565	570	570	572	575	575	575	580	580	580
580	585	590	590	590	600	600	600	600	610
610	615	625	625	625	635	649	650	670	670
675	675	680	690	700	700	700	700	715	715

Variance

- The variance is a measure of variability that utilizes all the data.
- It is based on the difference between the value of each observation (x_i) and the mean ($\bar{x}$ for a sample, μ for a population).
- The variance is useful in comparing the variability of two or more variables.
- The variance is the average of the squared deviations between each data value and the mean.
- The variance of a sample is: $s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$
- The variance for a population is: $\sigma^2 = \frac{\sum(x_i - \mu)^2}{N}$

Standard Deviation

- The standard deviation of a data set is the positive square root of the variance.
- It is measured in the same units as the data, making it more easily interpreted than the variance.
- The standard deviation of a sample is: $s = \sqrt{s^2}$
- The standard deviation of a population is: $\sigma = \sqrt{\sigma^2}$

Coefficient of Variation

- The coefficient of variation indicates how large the standard deviation is in relation to the mean.
- The coefficient of variation of a sample is: $\left[\frac{s}{\bar{x}} \text{ x } 100\right]\%$
- The coefficient of variation of a population is: $\left[\frac{\sigma}{\mu} \text{ x } 100\right]\%$

Sample Variance, Standard Deviation, and Coefficient of Variation

Example: Apartment Rents

- The variance is: $s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} = 2{,}996.16$

- The standard deviation is: $s = \sqrt{s^2} = \sqrt{2{,}996.16} = 54.74$

- The coefficient of variation is: $\left[\frac{s}{\bar{x}} \text{ x } 100\right]\% = \left[\frac{54.74}{590.80} \text{ x } 100\right]\% = 9.27\%$

Measures of Association Between Two Variables

- Thus far we have examined numerical methods used to summarize the data for one variable at a time.
- Often a manager or decision maker is interested in the relationship between two variables.
- Two descriptive measures of the relationship between two variables are covariance and correlation coefficient.

Covariance

- The covariance is a measure of the linear association between two variables.
- Positive values indicate a positive relationship.
- Negative values indicate a negative relationship.
- The covariance is computed as follows:

For samples: $$s_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

For populations: $$\sigma_{xy} = \frac{\sum(x_i - \mu_x)(y_i - \mu_y)}{N}$$

Correlation Coefficient (1 of 2)

- Correlation is a measure of linear association and not necessarily causation.
- Just because two variables are highly correlated, it does not mean that one variable is the cause of the other.
- The correlation coefficient is computed as follows:

For samples: $$r_{xy}=\frac{s_{xy}}{s_x s_y}$$

For populations: $$\rho_{xy}=\frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

Correlation Coefficient (2 of 2)

- The coefficient can take on values between –1 and +1.
- Values near –1 indicate a strong negative linear relationship.
- Values near +1 indicate a strong positive linear relationship.
- The closer the correlation is to zero, the weaker the relationship.

Covariance and Correlation Coefficient (1 of 3)

A golfer is interested in investigating the relationship, if any, between driving distance and 18-hole score.

Average Driving Distance (yards)	Average 18-Hole Score
277.6	69
259.5	71
269.1	70
267.0	70
255.6	71
272.9	69

Covariance and Correlation Coefficient (2 of 3)

Example: Golfing Study

	x	y	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
	277.6	69	10.65	-1.0	-10.65
	259.5	71	-7.45	1.0	-7.45
	269.1	70	2.15	0	0
	267.0	70	0.05	0	0
	255.6	71	-11.35	1.0	-11.35
	272.9	69	5.95	-1.0	-5.95
Average	267.0	70.0		Total	-35.40
Std. Dev.	8.2192	.8944			

Covariance and Correlation Coefficient (3 of 3)

Example: Golfing Study

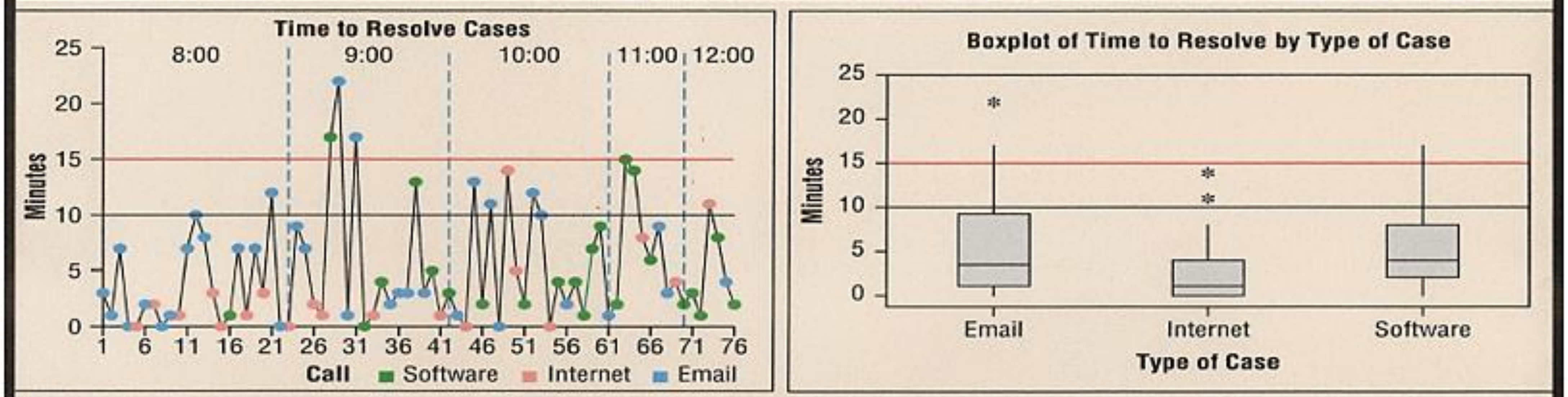
- Sample Covariance: $s_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{-35.40}{6-1} = -7.08$

- Sample Correlation Coefficient: $r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{-7.08}{(8.2192).8944)} = -.9631$

Data Dashboards: Adding Numerical Measures to Improve Effectiveness (1 of 2)

- Data dashboards are not limited to graphical displays.
- The addition of numerical measures, such as the mean and standard deviation of KPIs, to a data dashboard is often critical.
- Dashboards are often interactive.
- Drilling down refers to functionality in interactive dashboards that allows the user to access information and analyses at an increasingly detailed level.

Data Dashboards: Adding Numerical Measures to Improve Effectiveness (2 of 2)



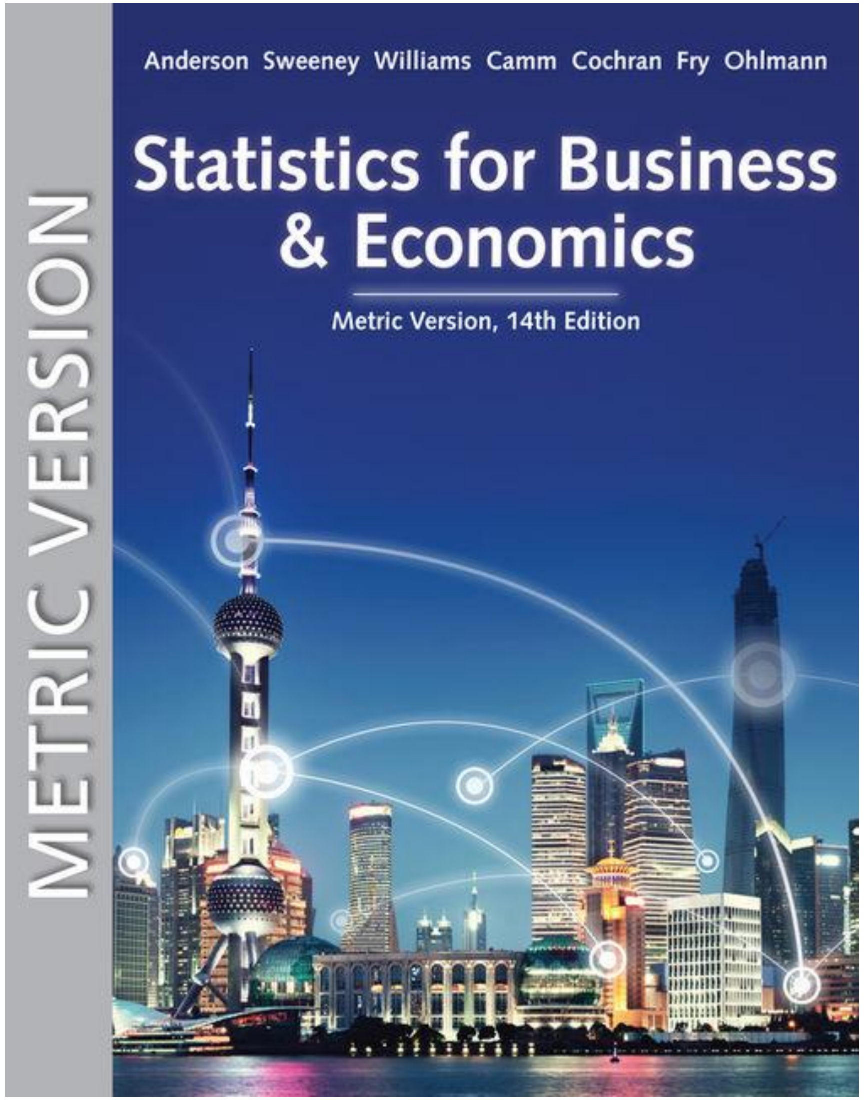
Summary Statistics - Resolved Cases

Type of Case	Cases	Mean	Median	Std. Dev
Email	34	4.6	2.0	5.6
Internet	19	5.4	3.0	4.9
Software	23	5.2	4.0	4.2

Hour	Cases	Mean	Median	Std. Dev
8:00	22	3.5	2.0	3.7
9:00	19	5.8	3.0	6.6
10:00	19	5.3	4.0	4.8
11:00	9	6.9	6.0	5.1
12:00	6	4.8	3.5	3.9

Statistics for Business and Economics (14e) Metric Version

Anderson, Sweeney, Williams, Camm, Cochran, Fry, Ohlmann



Chapter 5: Discrete Probability Distributions

5.1 - Random Variables

5.2 - Developing Discrete Probability Distributions

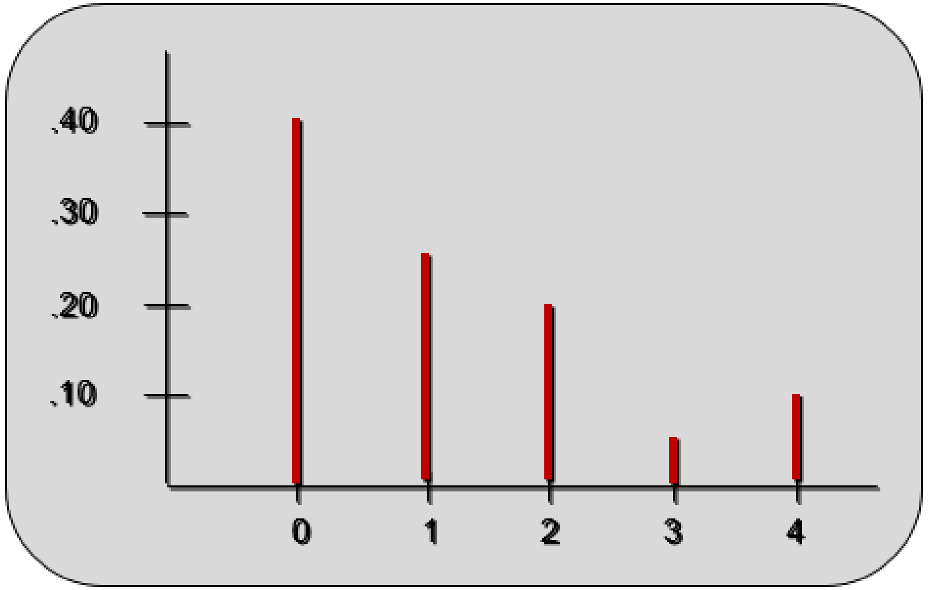
5.3 - Expected Value and Variance

5.4 - Bivariate Distributions, Covariance, and Financial Portfolios

5.5 - Binomial Probability Distribution

5.6 - Poisson Probability Distribution

5.7 - Hypergeometric Probability Distribution



Discrete Probability Distributions (1 of 7)

- The probability distribution for a random variable describes how probabilities are distributed over the values of the random variable.
- We can describe a discrete probability distribution with a table, graph, or formula.

Types of discrete probability distributions:

- First type: uses the rules of assigning probabilities to experimental outcomes to determine probabilities for each value of the random variable.
- Second type: uses a special mathematical formula to compute the probabilities for each value of the random variable.

Discrete Probability Distributions (2 of 7)

- The probability distribution is defined by a probability function, denoted by $f(x)$, that provides the probability for each value of the random variable.
- The required conditions for a discrete probability function are:

$$f(x) \geq 0 \quad \text{and} \quad \sum f(x) = 1$$

Discrete Probability Distributions (3 of 7)

- There are three methods for assigning probabilities to random variables: classical method, subjective method, and relative frequency method.
- The use of the relative frequency method to develop discrete probability distributions leads to what is called an empirical discrete distribution.

Discrete Probability Distributions (4 of 7)

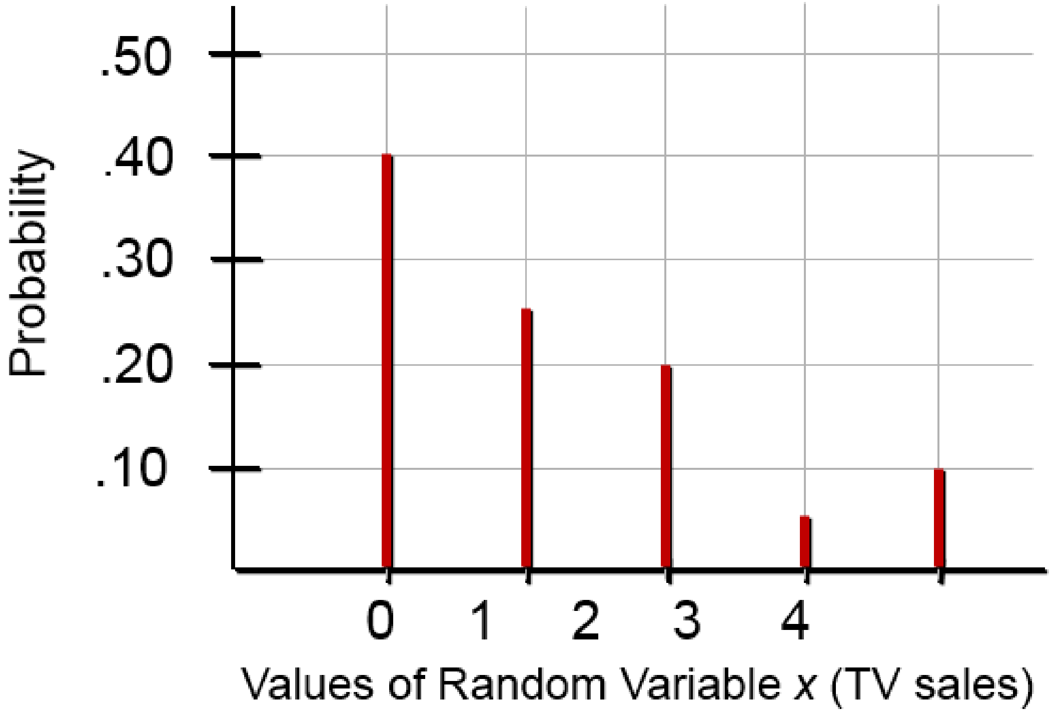
Example: JSL Appliances

Using past data on TV sales, a tabular representation of the probability distribution for sales was developed.

Units Sold	Number of Days	x	$f(x)$
0	80	0	.40 = 80/200
1	50	1	0.25
2	40	2	0.20
3	10	3	0.05
4	20	4	0.10
	200		1.00

Discrete Probability Distributions (5 of 7)

Example: JSL Appliances



Graphical representation of probability distribution

Discrete Probability Distributions (6 of 7)

- In addition to tables and graphs, a formula that gives the probability function, $f(x)$, for every value of x is often used to describe the probability distributions.
- Several discrete probability distributions specified by formulas are the discrete-uniform, binomial, Poisson, and hypergeometric distributions.

Discrete Probability Distributions (7 of 7)

- The discrete uniform probability distribution is the simplest example of a discrete probability distribution given by a formula.
- The discrete uniform probability function is

$$f(x) = 1/n$$

where: n = the number of values the random variable may assume

- The values of the random variable are equally likely.

Expected Value (1 of 2)

- The expected value, or mean, of a random variable is a measure of its central location.

$$E(x) = \mu = \sum xf(x)$$

- The expected value is a weighted average of the values the random variable may assume. The weights are the probabilities.
- The expected value does not have to be a value the random variable can assume.

Variance and Standard Deviation

- The variance summarizes the variability in the values of a random variable.

$$Var(x) = \sigma^2 = \Sigma(x - \mu)^2 f(x)$$

- The variance is a weighted average of the squared deviations of a random variable from its mean. The weights are the probabilities.
- The standard deviation, σ, is defined as the positive square root of the variance.

Expected Value (2 of 2)

Example: JSL Appliances

x	$f(x)$	$xf(x)$
0	.40	.00
1	.25	.25
2	.20	.40
3	.05	.15
4	.10	.40

$E(x) =$ 1.20 = expected number of TVs sold in a day

Variance

Example: JSL Appliances

X	$x - \mu$	$(x - \mu)^2$	f(x)	$(x - \mu)^2$ f(x)
0	−1.2	1.44	.40	.576
1	−0.2	0.04	.25	.010
2	0.8	0.64	.20	.128
3	1.8	3.24	.05	.162
4	2.8	7.84	.10	.784
				Variance of daily sales = σ^2 = 1.660

Standard deviation of daily sales = 1.2884 TVs

Bivariate Distributions (1 of 3)

A bivariate probability distribution is a probability distribution involving two random variables.

For example, here are the daily sales at the DiCarlo Motors automobile dealership in Saratoga, New York, and DiCarlo, another dealership in Geneva, New York. The table shows the number of cars sold at each of the dealerships over a 300-day period.

	Saratoga Dealership						
Geneva Dealership	**0**	**1**	**2**	**3**	**4**	**5**	**Total**
0	21	30	24	9	2	0	86
1	21	36	33	18	2	1	111
2	9	42	9	12	3	2	77
3	3	9	6	3	5	0	26
Total	54	117	72	42	12	3	300

Bivariate Distributions (2 of 3)

Let us define x = number of cars sold at the Geneva dealership and y = the number of cars sold at the Saratoga dealership. We can now divide all of the frequencies by the number of observations (300) to develop a bivariate empirical discrete probability distribution for automobile sales at the two DiCarlo dealerships.

	Saratoga Dealership						
Geneva Dealership	**0**	**1**	**2**	**3**	**4**	**5**	**Total**
0	.0700	.1000	.0800	.0300	.0067	.0000	.2867
1	.0700	.1200	.1100	.0600	.0067	.0033	.3700
2	.0300	.1400	.0300	.0400	.0100	.0067	.2567
3	.0100	.0300	.0200	.0100	.0167	.0000	.0867
Total	.18	.39	.24	.14	.04	.01	1.0000

Bivariate Distributions (3 of 3)

The table below shows the expected value for the mean total sales and the standard deviation of total sales for these two dealerships.

s	$f(s)$	$sf(s)$	$s - E(s)$	$(s - E(s))^2$	$(s - E(s))^2 f(s)$
0	.0700	.0000	–2.6433	6.9872	.4891
1	.1700	.1700	–1.6433	2.7005	.4591
2	.2300	.4600	–.6433	.4139	.0952
3	.2900	.8700	.3567	.1272	.0369
4	.1267	.5067	1.3567	1.8405	.2331
5	.0667	.3333	2.3567	5.5539	.3703
6	.0233	.1400	3.3567	11.2672	.2629
7	.0233	.1633	4.3567	18.9805	.4429
8	.0000	.0000	5.3567	28.6939	.0000
		$E(s) = 2.6433$			$Var(s) = 2.3895$

Covariance

The covariance and/or correlation coefficient are good measures of association between two random variables.

Covariance = $\sigma_{xy} = [Var(x + y) - Var(x) - Var(y)]/2.$

= (2.3895 – 0.8696 – 1.25)/2

= 0.1350

A covariance of .1350 indicates that daily sales at DiCarlo's two dealerships have a positive relationship.

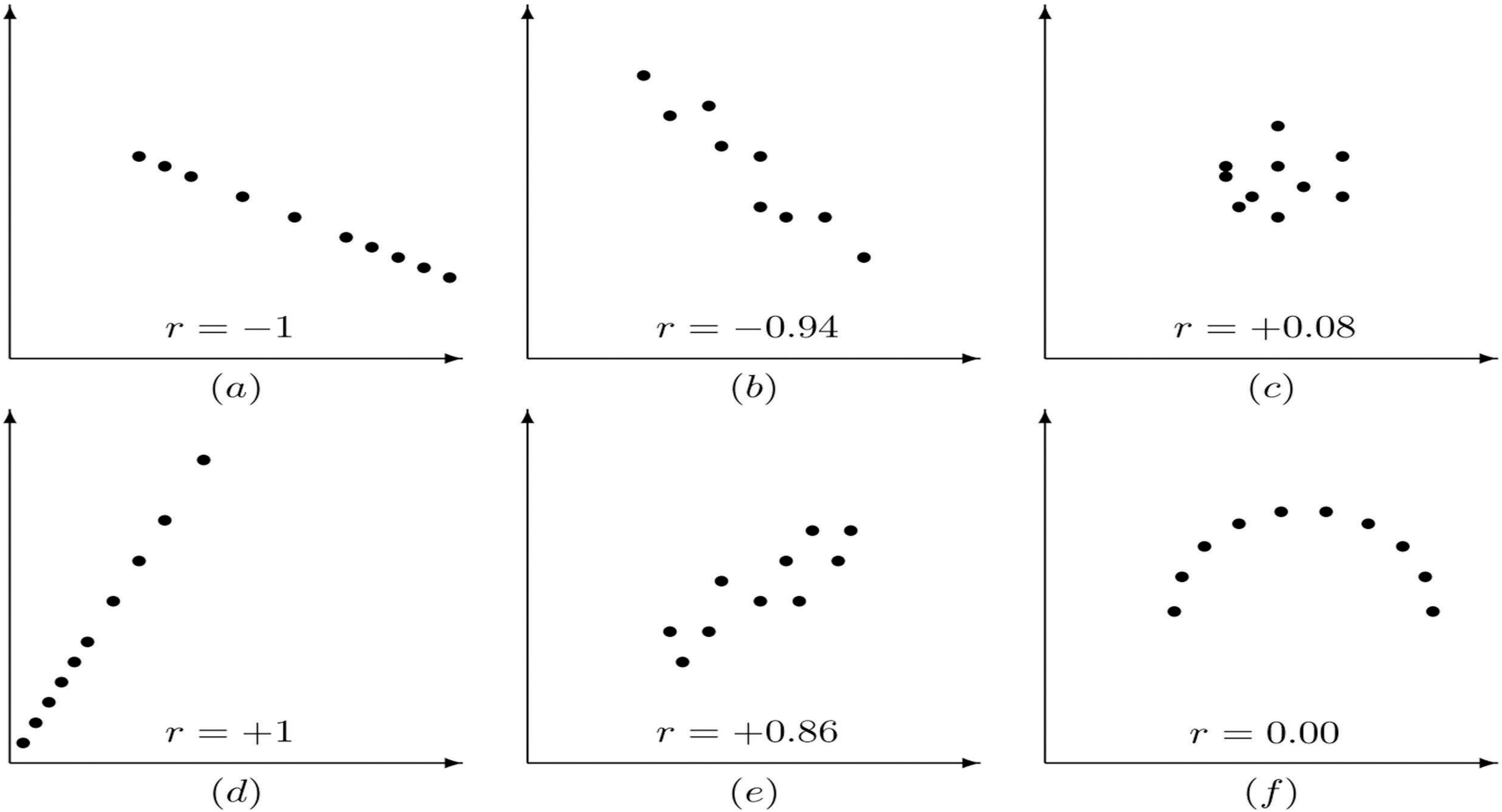
Correlation

To get a better sense of the strength of the relationship, we can compute the correlation coefficient.

$$\text{Correlation} = \rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

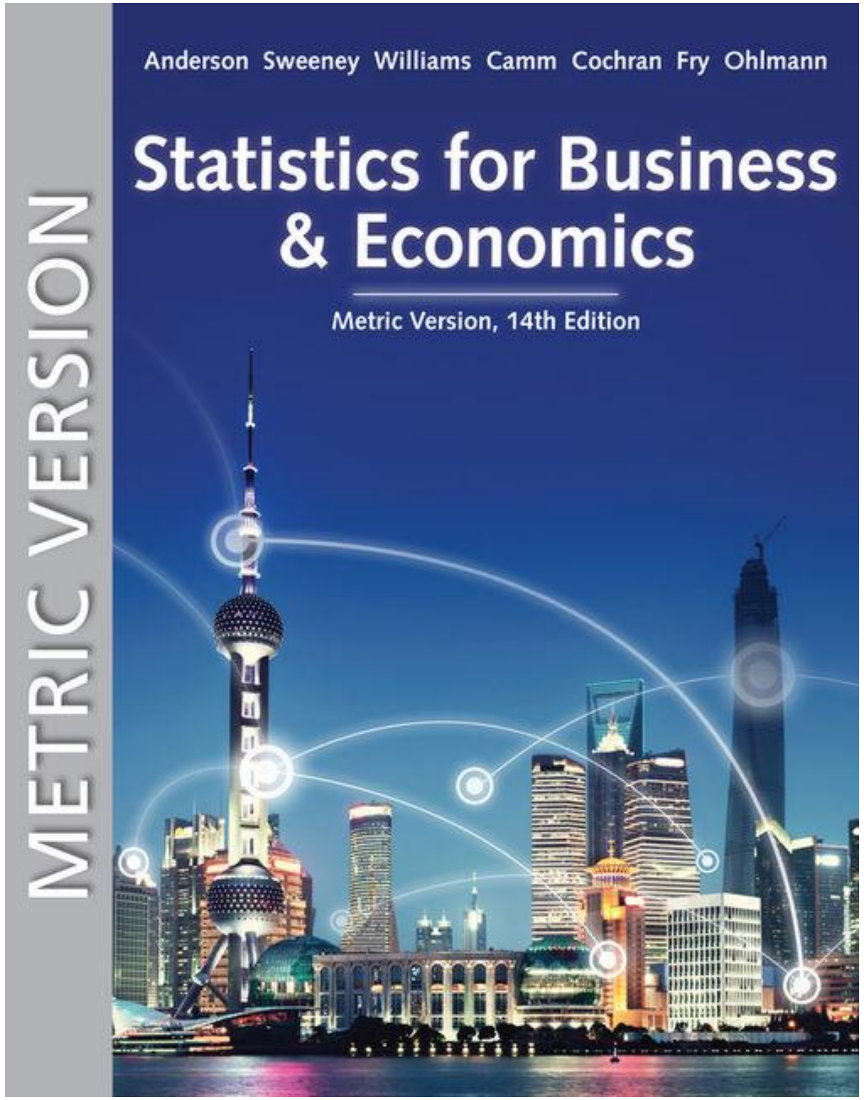
$$\rho_{xy} = \frac{0.1350}{(0.9325)(1.1180)} = 0.1295$$

The correlation coefficient of .1295 indicates there is a weak positive relationship between the random variables representing daily sales at the two DiCarlo dealerships. If the correlation coefficient had equaled zero, we would have concluded that daily sales at the two dealerships were independent.


$r = -1$
(a)
$r = -0.94$
(b)
$r = +0.08$
(c)
$r = +1$
(d)
$r = +0.86$
(e)
$r = 0.00$
(f)

Statistics for
Business and Economics (14e)
Metric Version

Anderson, Sweeney, Williams, Camm, Cochran, Fry, Ohlmann



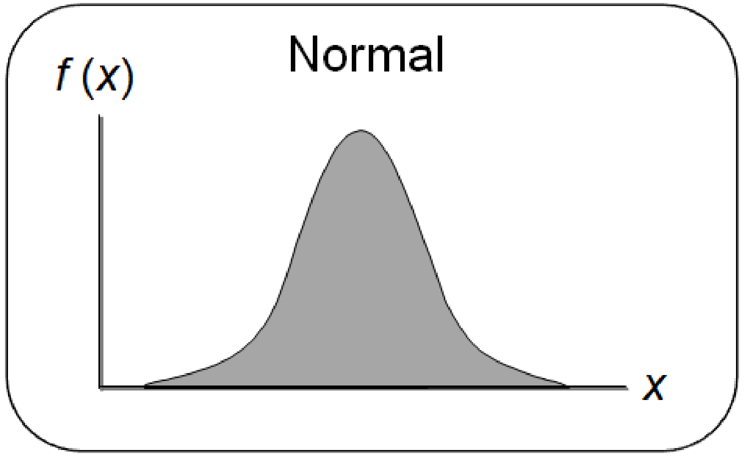
Chapter 6 - Continuous Probability Distributions

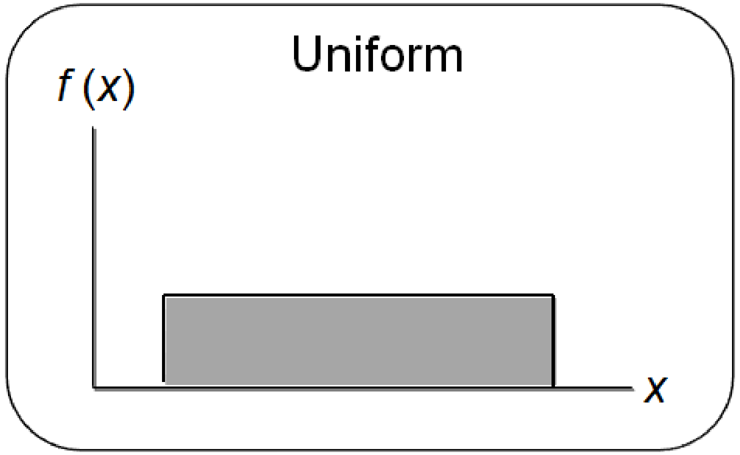
6.1 – Uniform Probability Distribution

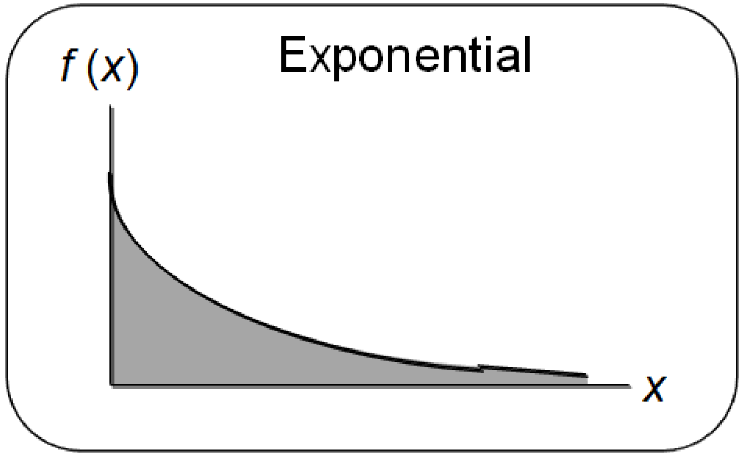
6.2 – Normal Probability Distribution

6.3 – Normal Approximation of Binomial Probabilities

6.4 – Exponential Probability Distribution







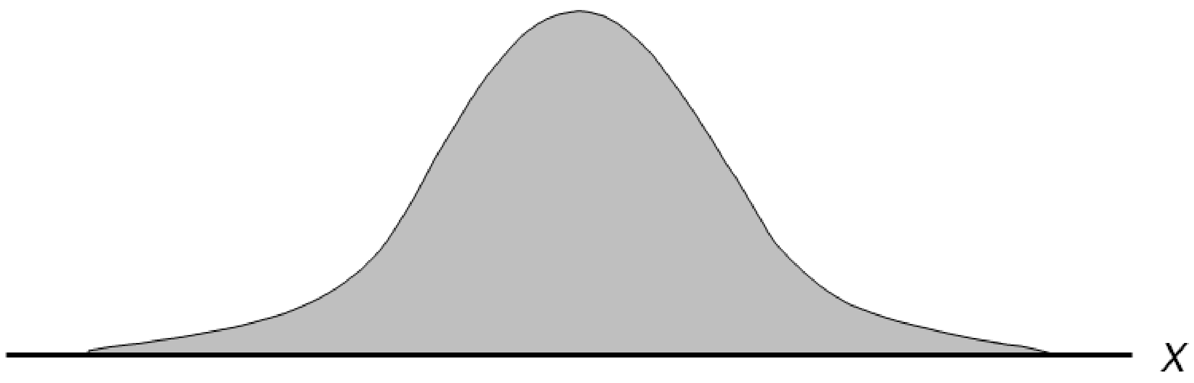
Normal Probability Distribution (1 of 7)

- The normal probability distribution is the most important distribution for describing a continuous random variable.
- It is widely used in statistical inference.
- It has been used in a wide variety of applications including:
  - Heights of people
  - Amount of rainfall
  - Test scores
  - Scientific measurements
- Abraham de Moivre, a French mathematician, published *The Doctrine of Chances* in 1733.
- He derived the normal distribution.

Normal Probability Distribution (2 of 7)

Normal Probability Density Function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



where

μ= mean

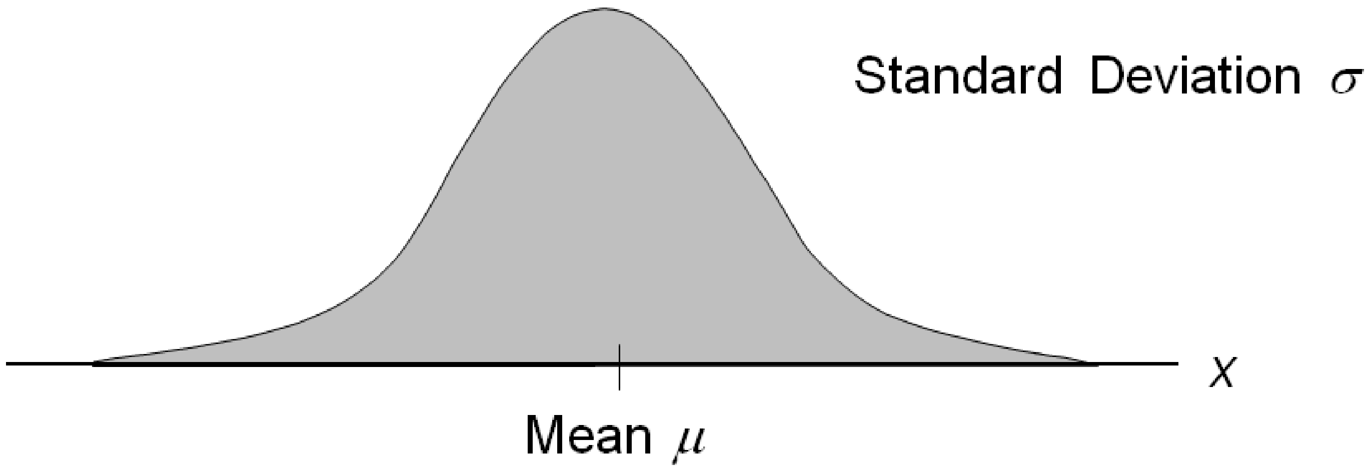
σ = Standard deviation

π = 3.14159265359…

Normal Probability Distribution (3 of 7)

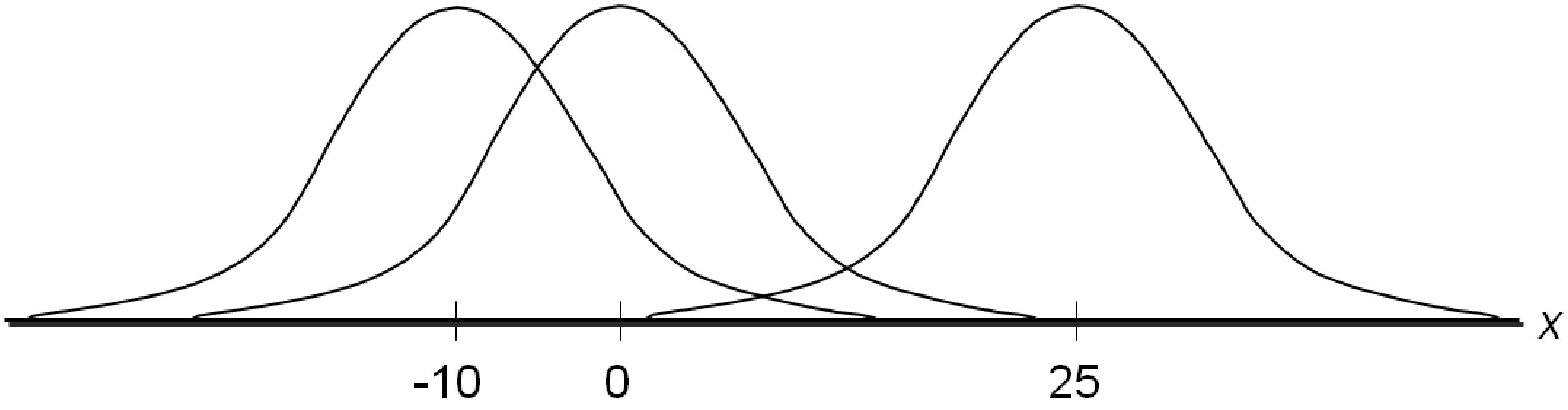
The entire family of normal probability distributions is defined by its mean μ and its standard deviation σ.

The highest point on the normal curve is at the mean, which is also the median and mode.



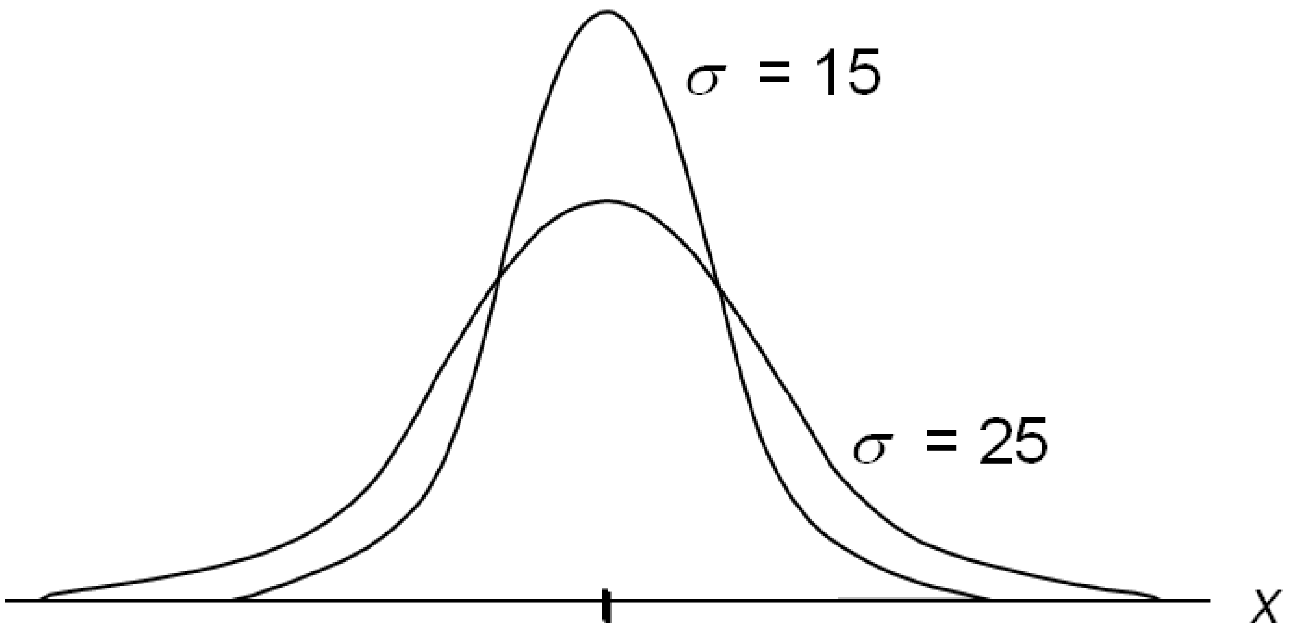
Normal Probability Distribution (4 of 7)

The mean can be any numerical value: negative, zero, or positive.



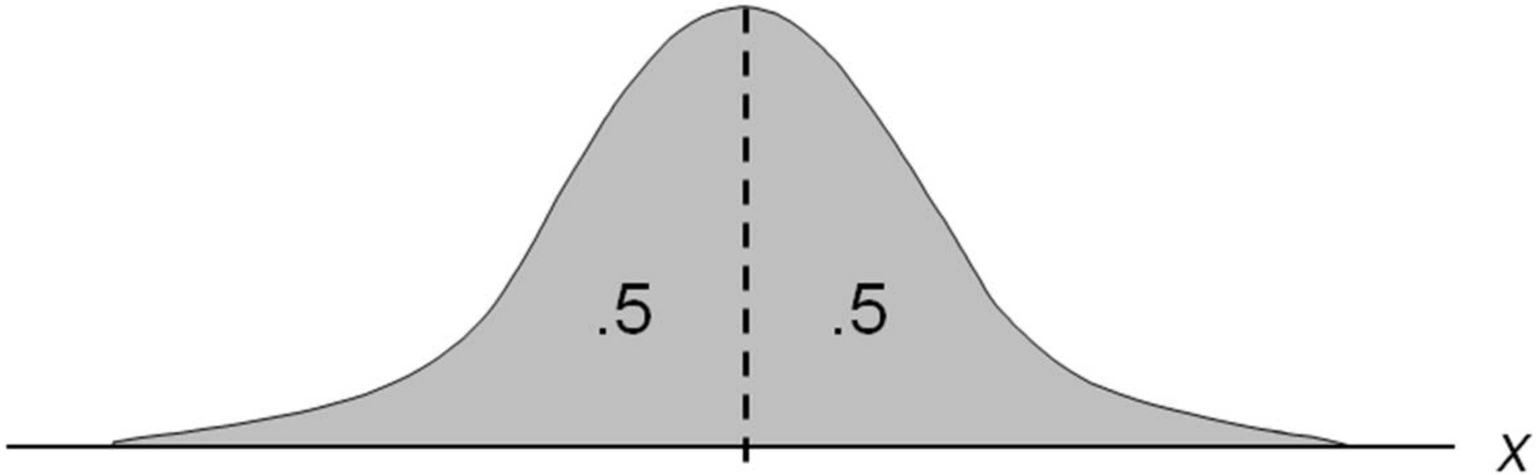
Normal Probability Distribution (5 of 7)

The standard deviation determines the width of the curve: larger values result in wider, flatter curves.



Normal Probability Distribution (6 of 7)

Probabilities for the normal random variable are given by areas under the curve. The total area under the curve is 1 (0.5 to the left of the mean and 0.5 to the right).



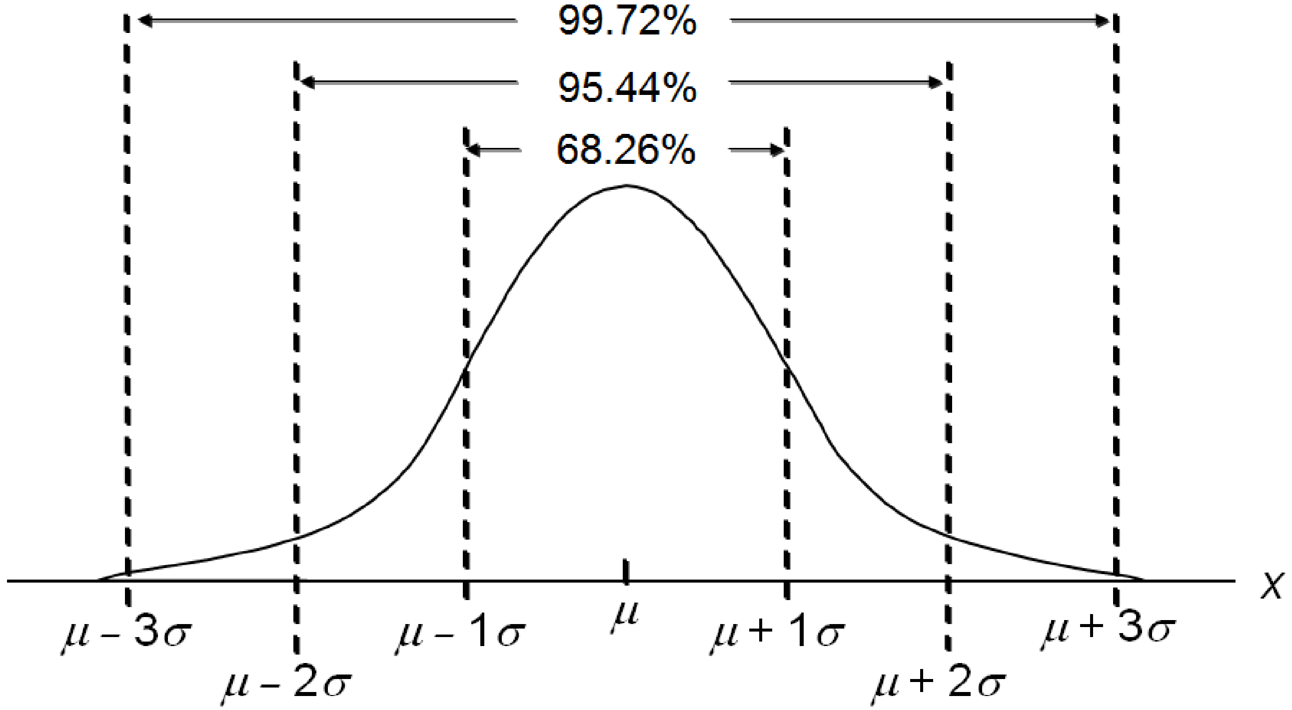
Normal Probability Distribution (7 of 7)

Empirical Rule（經驗法則）

68.26% of values of a normal random variable are within ±1 standard deviation of its mean.

95.44% of values of a normal random variable are within ±2 standard deviations of its mean.

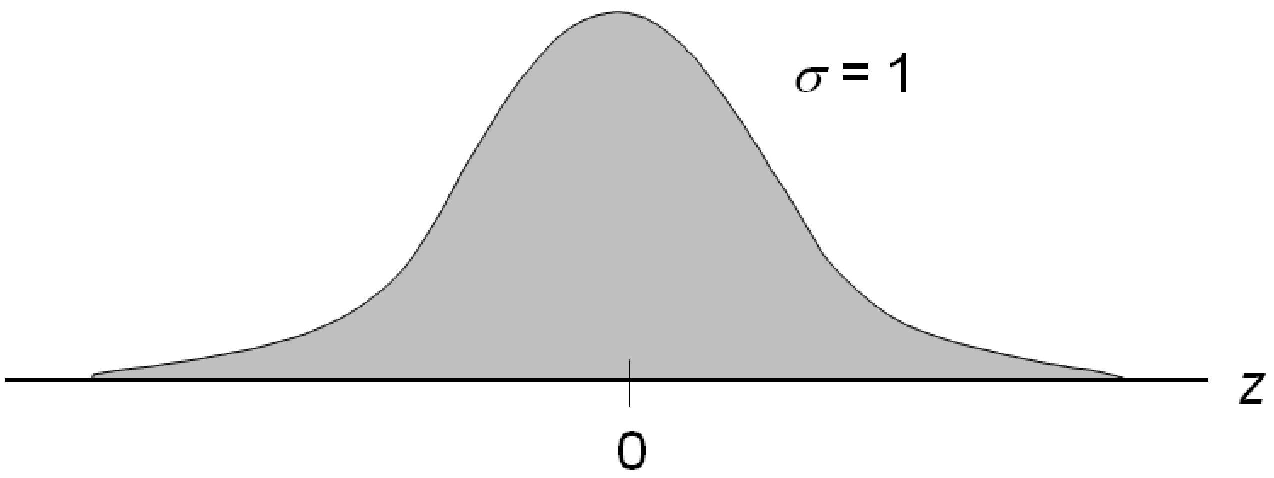
99.72% of values of a normal random variable are within ±3 standard deviations of its mean.



Standard Normal Probability Distribution (1 of 10)

A random variable having a normal distribution with a mean of 0 and a standard deviation of 1 is said to have a standard normal probability distribution.

The letter z is used to designate the standard normal random variable.



Standard Normal Probability Distribution (2 of 10)

Converting to the Standard Normal Distribution

$$z = \frac{x - \mu}{\sigma}$$

We can think of *z* as a measure of the number of standard deviations *x* is from μ.

Standard Normal Probability Distribution (3 of 10)

Example: Pep Zone

Pep Zone sells auto parts and supplies including a popular multi-grade motor oil. When the stock of this oil drops to 20 liters, a replenishment order is placed.

The store manager is concerned that sales are being lost due to stockouts while waiting for a replenishment order.

It has been determined that demand during replenishment lead-time is normally distributed with a mean of 15 liters and a standard deviation of 6 liters.

The manager would like to know the probability of a stockout during replenishment lead-time. In other words, what is the probability that demand during lead-time will exceed 20 liters?

Standard Normal Probability Distribution (4 of 10)

Solving for the Stockout Probability

Step 1: Convert *x* to the standard normal distribution.

$$z = \frac{(x-\mu)}{\sigma}$$

$$z = \frac{(20 - 15)}{6}$$

$$z = 0.83$$

Step 2: Find the area under the standard normal curve to the left of z = 0.83.

Standard Normal Probability Distribution (5 of 10)

Cumulative Probability Table for the Standard Normal Distribution

$$P(z \leq 0.83) = 0.7967$$

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.	.	.	.	.	.	.	.	.	.	.
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7795	.8023	.8051	.8078	.8106	.8133
.9	.8129	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
.	.	.	.	.	.	.	.	.	.	.

Standard Normal Probability Distribution (6 of 10)

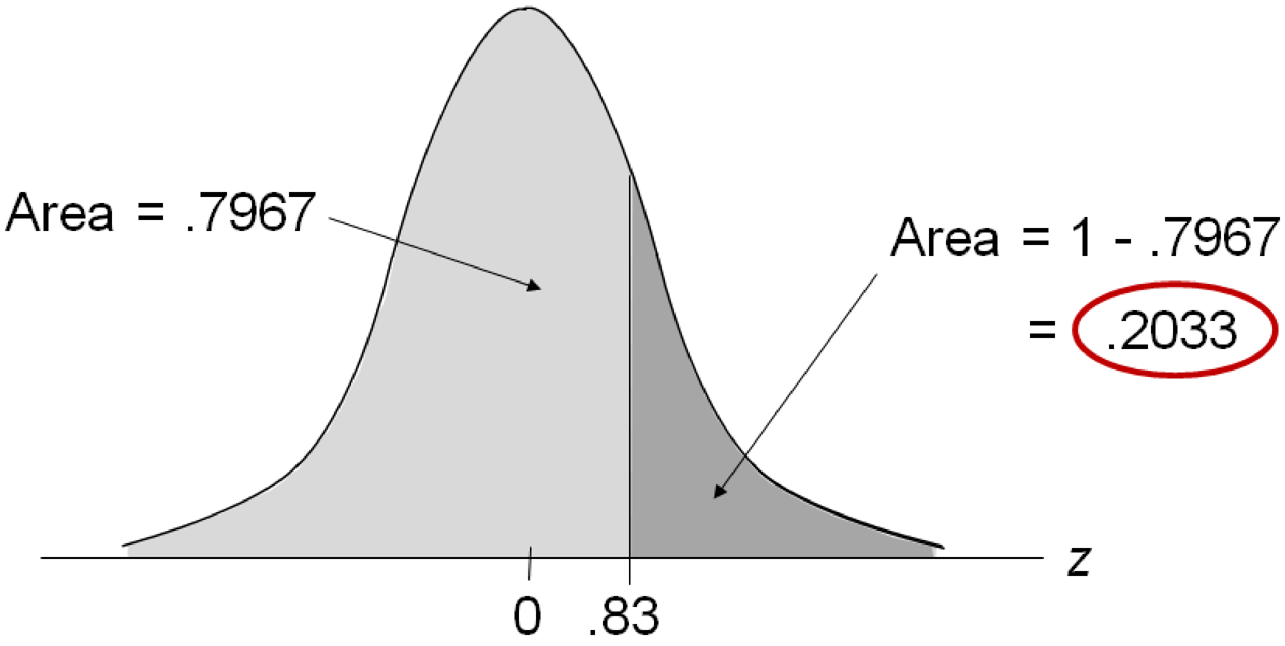
Solving for the Stockout Probability

Step 3: Compute the area under the standard normal curve to the right of z = 0.83.

$$\begin{aligned} P(z > 0.83) &= 1 - P(z \leq 0.83) \\ &= 1 - 0.7967 \\ &= 0.2033 \end{aligned}$$

Standard Normal Probability Distribution (7 of 10)

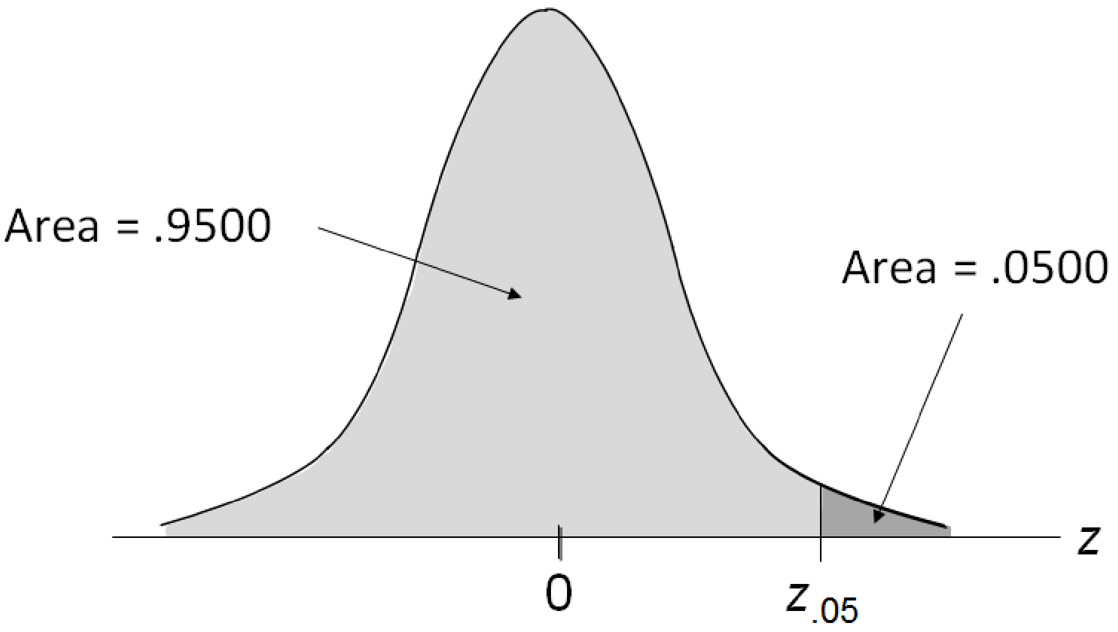
Solving for the Stockout Probability



Standard Normal Probability Distribution (8 of 10)

If the manager of Pep Zone wants the probability of a stockout during replenishment lead-time to be no more than .05, what should the reorder point be?

(Hint: Given a probability, we can use the standard normal table in an inverse fashion to find the corresponding z value.)



Standard Normal Probability Distribution (9 of 10)

Solving for the Reorder Point

Step 1: Find the z-value that cuts off an area of .05 in the right tail of the standard normal distribution by looking up the complement of the right tail area 1 – 0.05 = 0.95.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.	.	.	.	.	.	.	.	.	.	.
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
.	.	.	.	.	.	.	.	.	.	.

Standard Normal Probability Distribution (10 of 10)

Solving for the Reorder Point

Step 2: Convert $z_{.05}$ to the corresponding value of x.

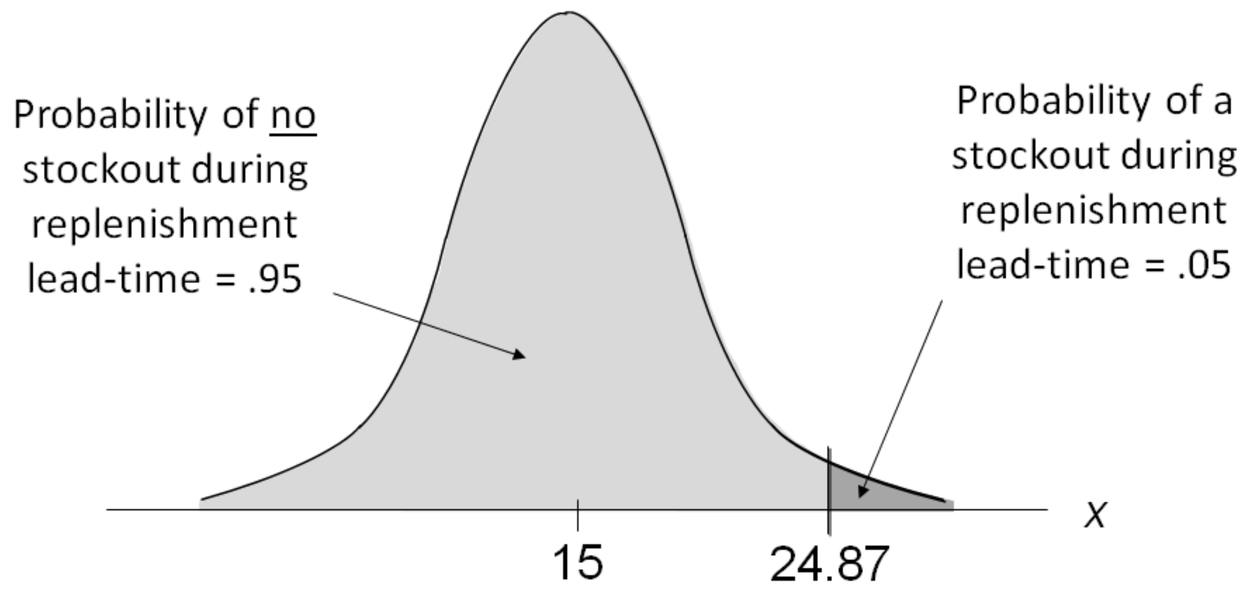
$$\begin{aligned} x &= \mu + z_{0.05}\sigma \\ &= 15 + 1.645(6) \\ &= 24.87 \end{aligned}$$

which we round to 25.

A reorder point of 25 liters will place the probability of a stockout during lead time at (slightly less than) 0.05.

Normal Probability Distribution

Solving for the Reorder Point



Standard Normal Probability Distribution

Solving for the Reorder Point

By raising the reorder point from 20 liters to 25 liters on hand, the probability of a stockout decreases from about .20 to .05.

This is a significant decrease in the chance that Pep Zone will be out of stock and unable to meet a customer's desire to make a purchase.

Using Excel to Compute Normal Probabilities

Excel has two functions for computing cumulative probabilities and *x* values for any normal distribution:

- NORM.DIST is used to compute the cumulative probability given an x value.
- NORM.INV is used to compute the x value given a cumulative probability.