巨量資料管理學院碩士在職專班

統計分析

2021/9/16

陳光宏

課程介紹

課程目標

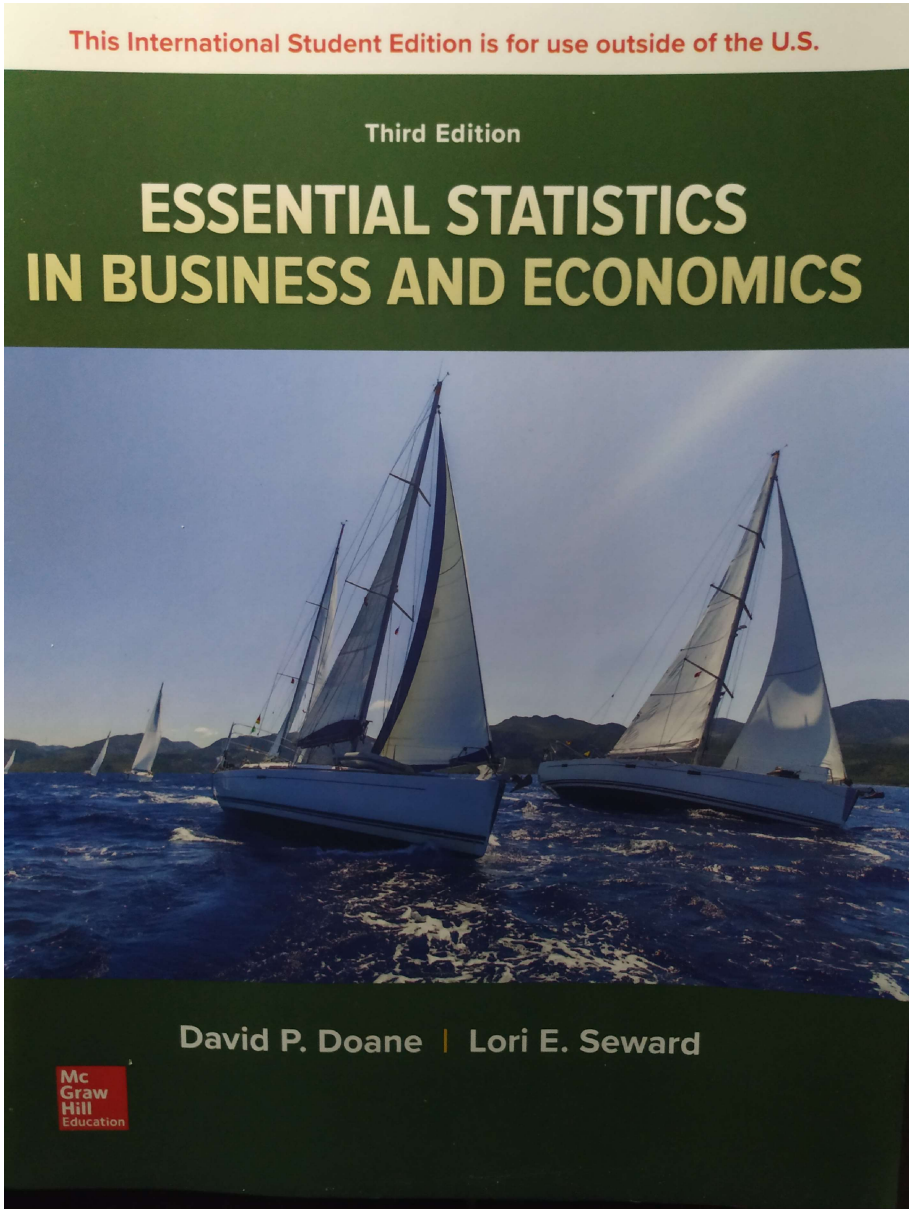
- 統計作為溝通的工具
  - 了解統計學家在想什麼
  - 讓數字說話
- 培養數據分析力及數字敏感度
  - 正確解讀統計分析結果
  - 用數據進行決策
- 職場競爭力
  - 軟體操作技巧

週次 Wk	日期 Date	課程內容 Content	授課教師
1	9月9日	中秋節	
2	9月16日	Introduction to statistics	陳光宏
3	9月23日	Descriptive statistics	陳光宏
4	9月30日	Probability	陳光宏
5	10月7日	Software Lab I	王雅蕙
6	10月14日	Probability distribution	陳光宏
7	10月21日	Estimation	陳光宏
8	10月28日	Hypothesis testing - Categorical variables	陳光宏
9	11月4日	Midterm report	陳光宏
10	11月11日	Hypothesis testing - Continuous variables	陳光宏
11	11月18日	Analysis of variance (ANOVA)	陳光宏
12	11月25日	Software Lab II	王雅蕙
13	12月2日	Linear regression I	陳光宏
14	12月9日	Linear regression II	陳光宏
15	12月16日	Categorical data analysis	陳光宏
16	12月23日	Logistic regression	陳光宏
17	12月30日	Software Lab III	王雅蕙
18	1月6日	Final term report	陳光宏

評分標準

- 出席 10%
  - 簽到
- 課堂作業與討論 30%
  - 分組討論
  - 課堂作業
  - 課後作業
- 期中報告 30%
- 期末報告 30%
  - 分組口頭與書面報告

教科書


This International Student Edition is for use outside of the U.S.
Third Edition
ESSENTIAL STATISTICS
IN BUSINESS AND ECONOMICS
David P. Doane | Lori E. Seward
Mc
Graw
Hill
Education

課堂討論與練習 – 讓數據說話

- 請參考“Week 1檔案.xlsx”業務員分析資料
- 如果你是業務部門主管，公司要求裁掉一名業務，你會選誰？

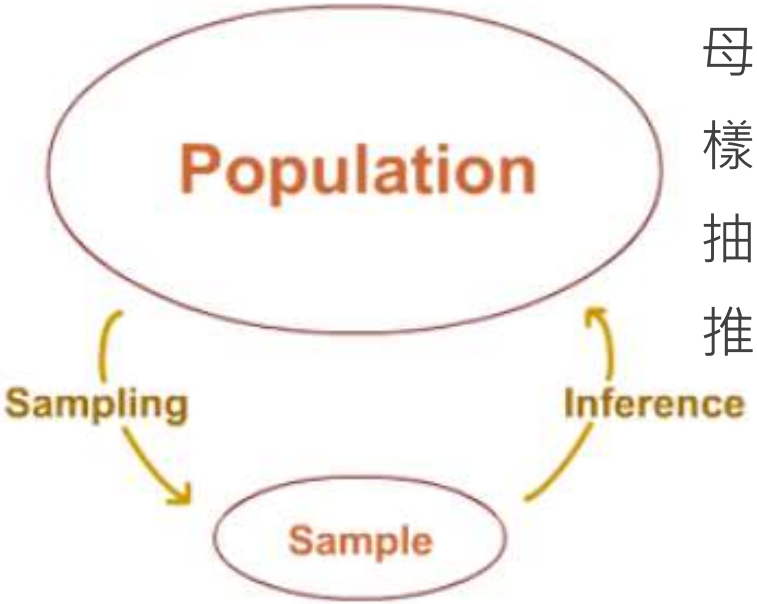
- 時間：8分鐘

日期	業務	銷售量
1月21日	SALES-A	10
2月24日	SALES-A	40
2月27日	SALES-A	120
3月29日	SALES-A	210
3月11日	SALES-A	60
3月5日	SALES-B	10
1月10日	SALES-B	140
3月1日	SALES-B	70
2月1日	SALES-B	30
2月12日	SALES-B	80
2月4日	SALES-B	20
2月8日	SALES-B	30
2月15日	SALES-B	20
3月3日	SALES-B	40
2月28日	SALES-C	50
3月8日	SALES-C	100
3月16日	SALES-C	80
2月27日	SALES-C	10
3月2日	SALES-C	20
3月3日	SALES-C	30
3月5日	SALES-C	50
3月10日	SALES-C	60

統計概論

統計學

- 是一種由資料(data) 萃取出資訊 (information) 的方法
- 母群體與樣本
- 不確定性 (Uncertainty)
  - 運用機率的概念



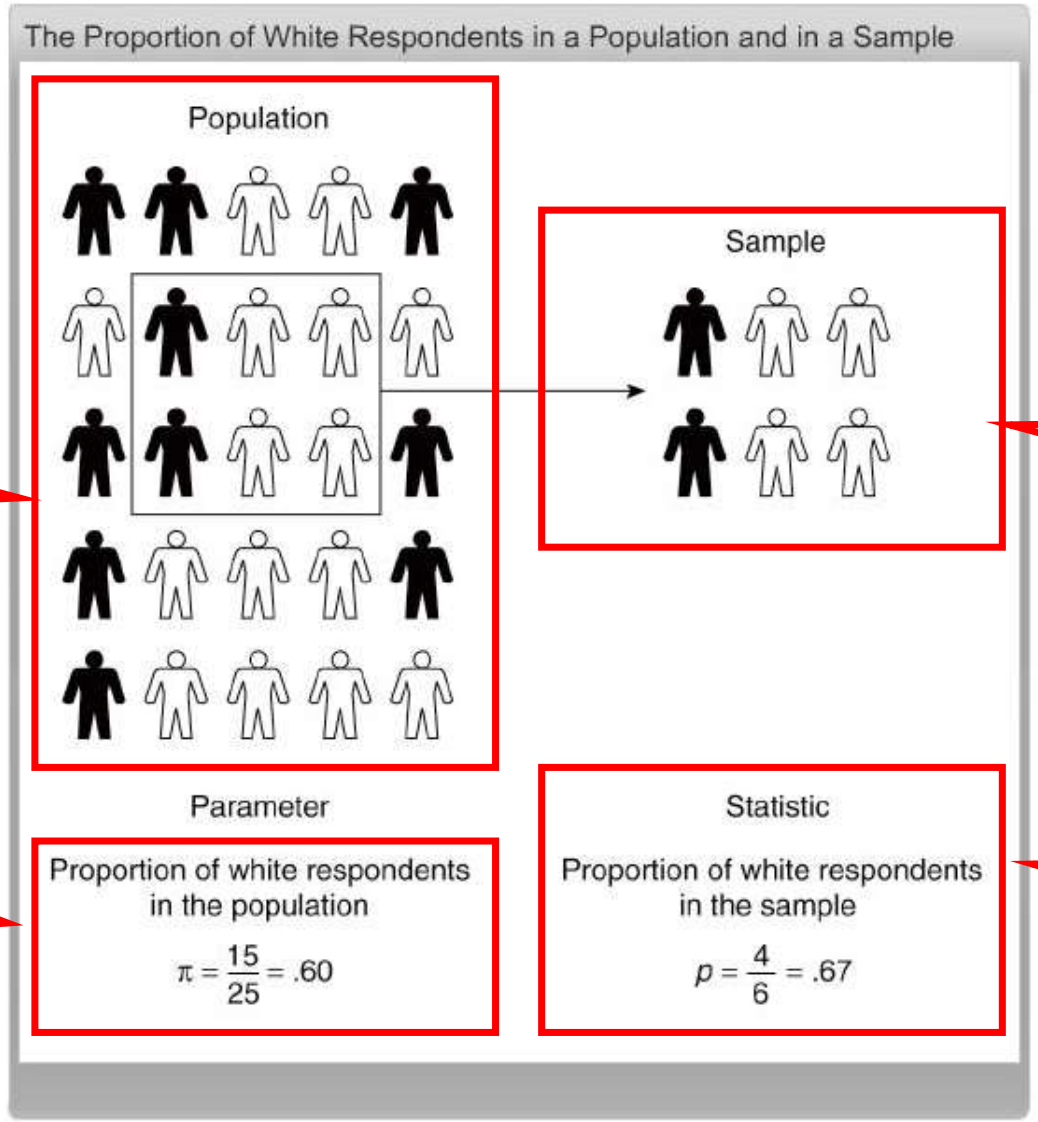
母體 Population

樣本 Sample

抽樣 Sampling

推論 Inference

舉例來說



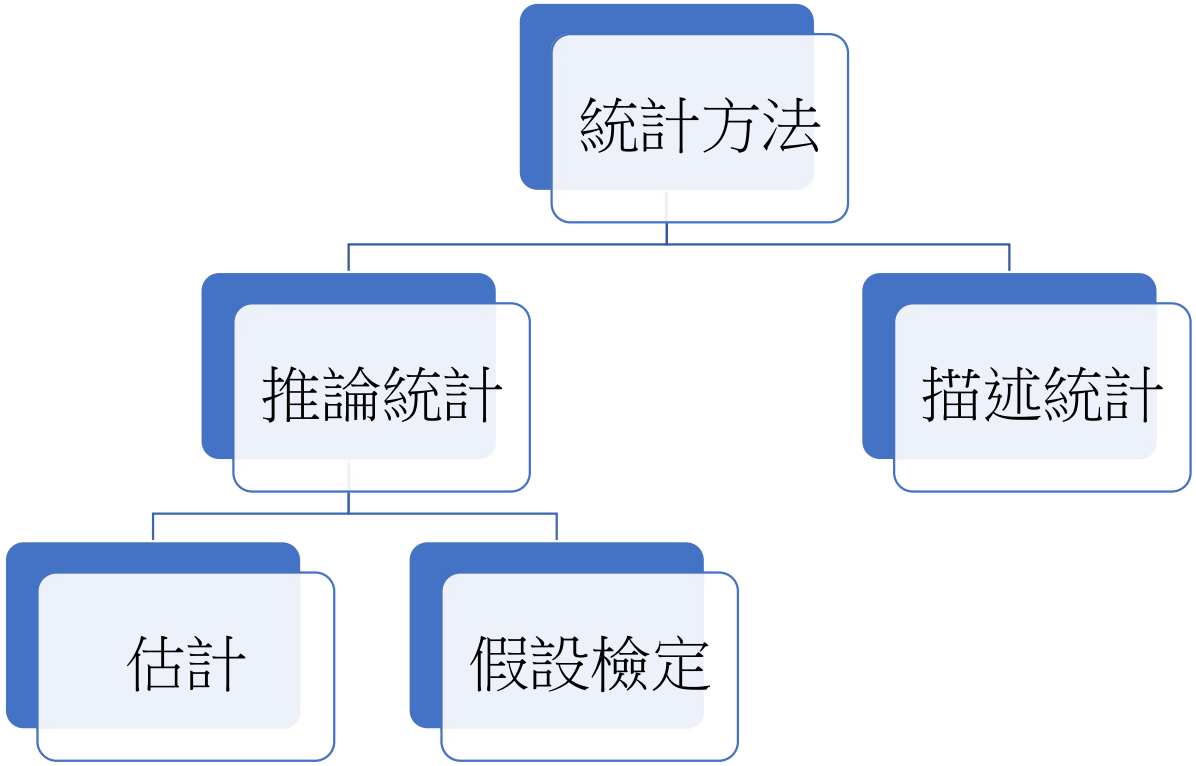
母群體

樣本

真值

估計值

統計方法的分類



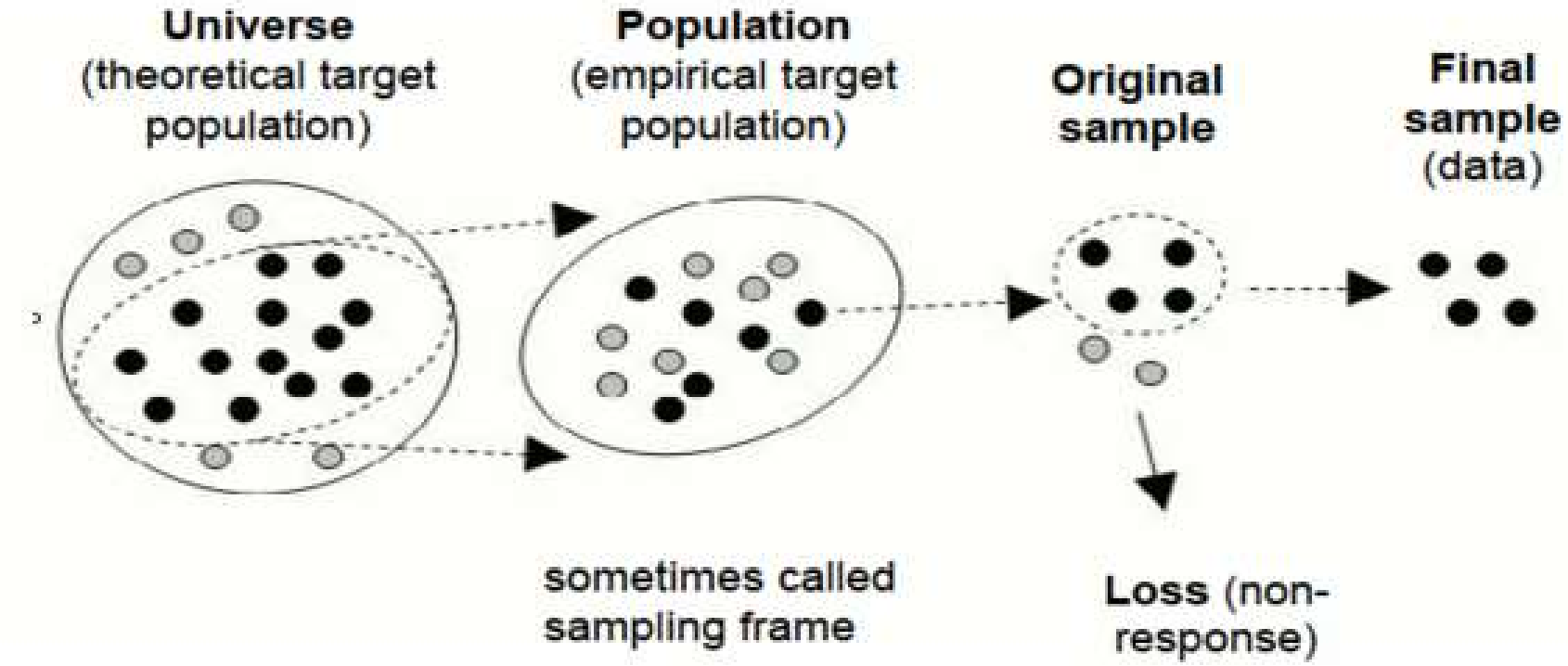
迴歸建模
資料收集
資料視覺化
資料整理

資料來源

常見的資料來源

- 初級資料 (Primary data)
  - 臨床資料
    - 臨床試驗、觀察型研究
  - 經由抽樣取得的資料
    - Nation al Health and Nutrition Examination Survey (NHANES)
    - Nutrition and Health Survey in Taiwan (NAHSIT)
- 次級資料 (Secondary data)
  - 官方行政資料
    - 健保資料庫、癌症登記、死亡登記

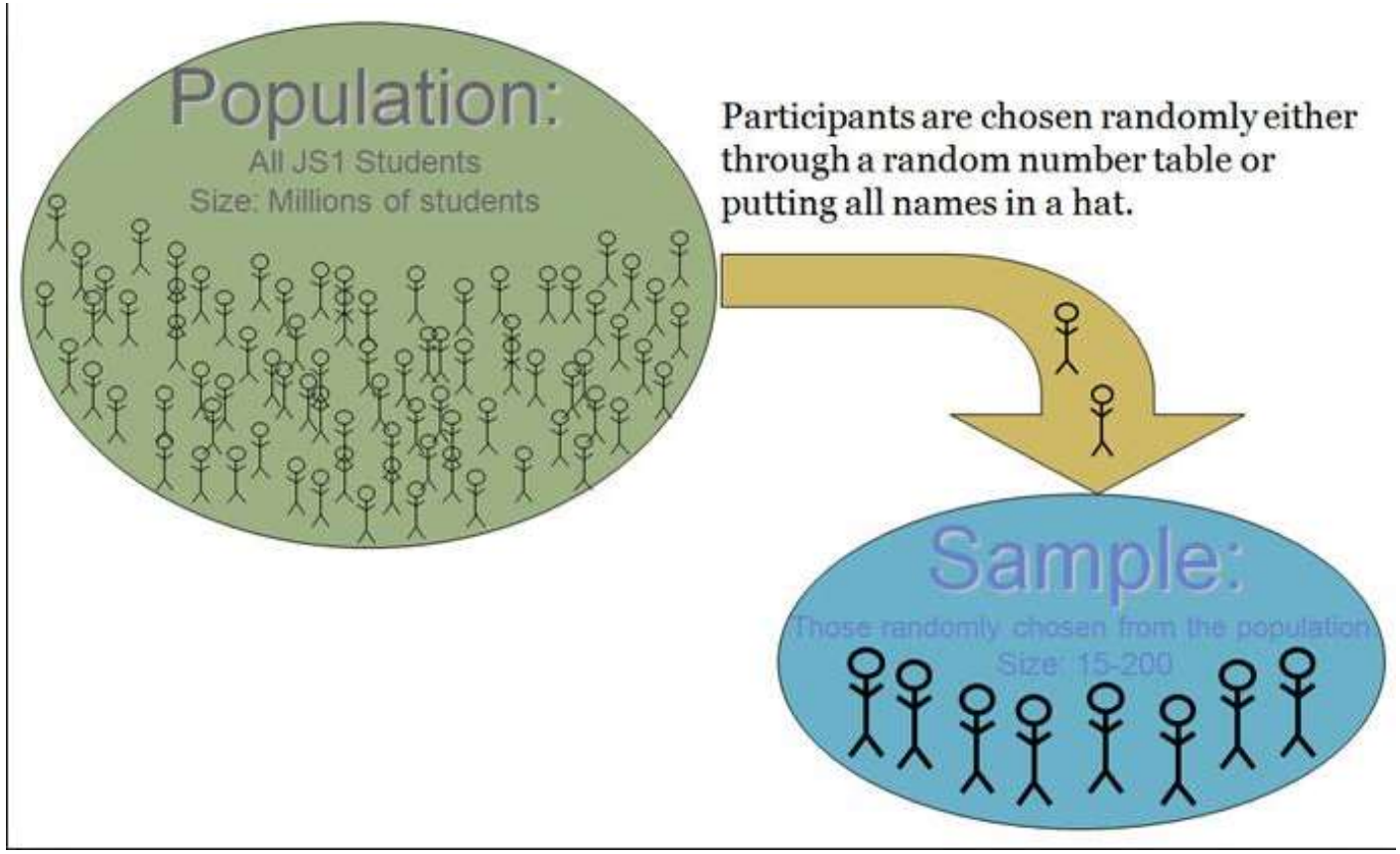
抽樣


Universe
(theoretical target population)
Population
(empirical target population)
Original sample
Final sample
(data)
sometimes called sampling frame
Loss (non-response)

機率抽樣的方法

- 簡單隨機抽樣 (Simple random sampling)
- 系統抽樣 (Systematic sampling)
- 分層抽樣 (Stratified sampling)
- 集群抽樣 (Cluster sampling)

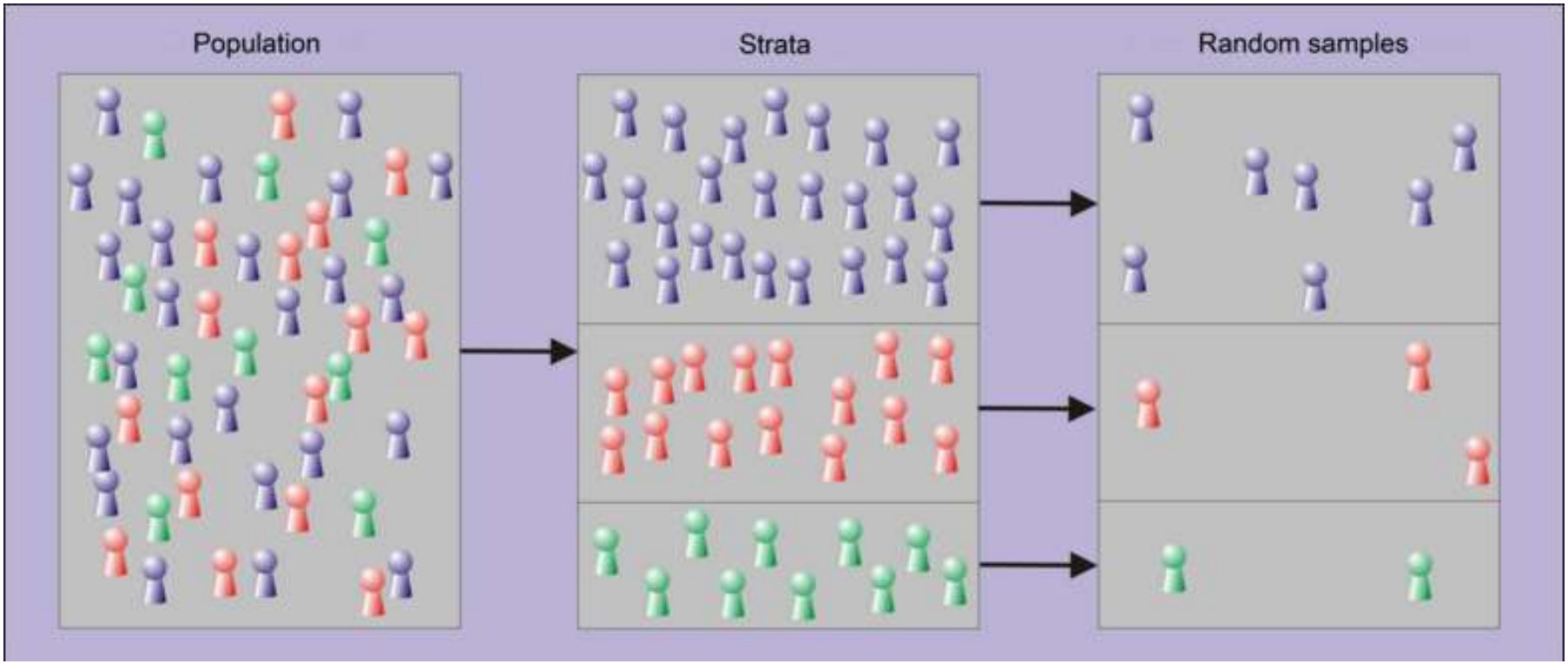
簡單隨機抽樣


Population:
All JS1 Students
Size: Millions of students
Participants are chosen randomly either through a random number table or putting all names in a hat.
Sample:
Those randomly chosen from the population
Size: 15-200

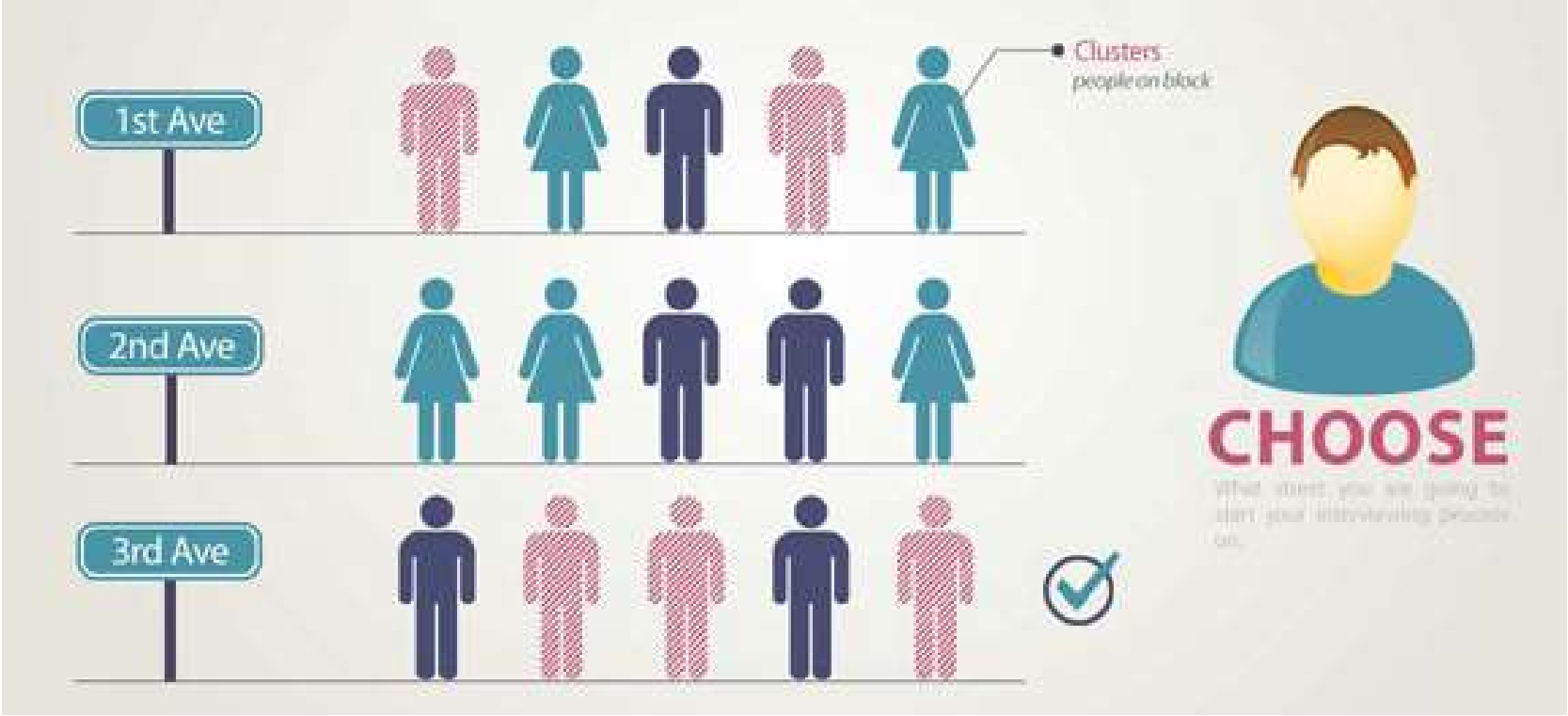
系統抽樣


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40

分層抽樣

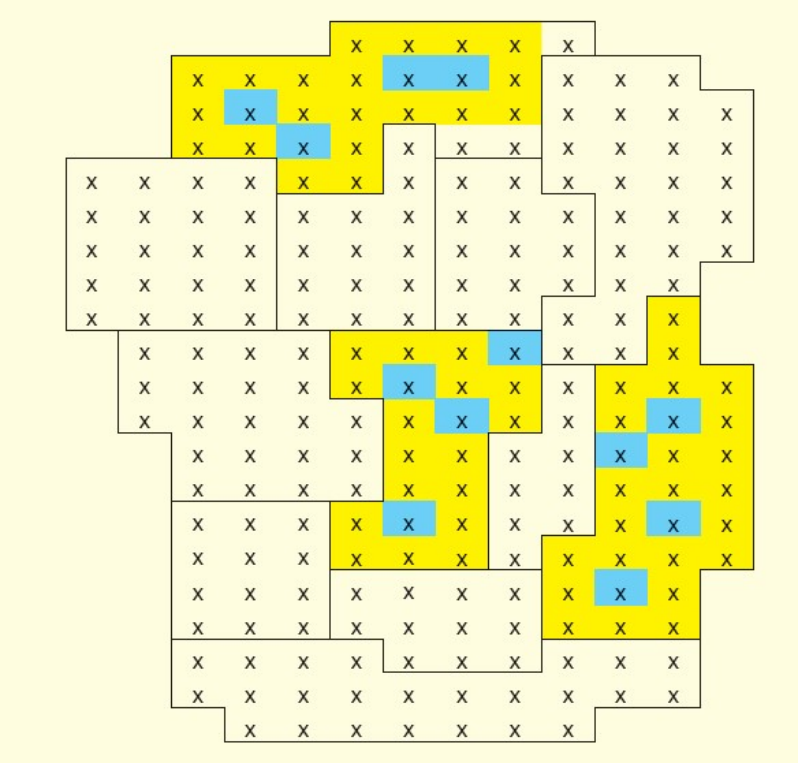


集群抽樣



混合多種抽樣方法

- 兩階段集群抽樣



- 分層多階段機率抽樣

表 4-1、各縣市之各層各階段抽出單位數

縣市別	層別	鄉鎮市區數	人口數	佔該縣市人口	預定樣本數	各階段抽出單位數				調整後樣本數
						鄉鎮市區	村里	鄰	人	
合計	49	358	23,110,923		30,800	168				30,960
台北市	一	8	1,705,671	64.34%	1,029	8		256	4	1,024
	二	4	945,297	35.66%	571	4	18	8	4	576
	小計	12	2,650,968	100.00%	1,600	12				1,600
新北市*	一	6	2,195,543	56.06%	1,009	6		252	4	1,008
	二	23	1,720,908	43.94%	791	6		22	6	792
	小計	6	3,916,451	100.00%	1,800	6				1,800
台中市	一	5	611,203	56.04%	785	5		196	4	784
	二	3	479,470	43.96%	615	3	20	8	4	640
	小計	8	1,090,673	100.00%	1,400	8				1,424
台中市*	一	11	1,135,201	72.13%	1,010	6		42	4	1,008
	二	10	438,520	27.87%	390	2		32	6	384
	小計	21	1,573,721	100.00%	1,400	8				1,392
台南市	一	3	404,621	52.25%	679	3		170	4	680
	二	3	369,709	47.75%	621	3	20	8	4	640
	小計	6	774,330	100.00%	1,300	6				1,320
	一	6	529,824	48.05%	673	4		42	4	672
	二	10	309,863	28.10%	393	4		16	6	384

非機率抽樣

Judgment Sample	Use expert knowledge to choose "typical" items (e.g., which employees to interview).
Convenience Sample	Use a sample that happens to be available (e.g., ask co-workers' opinions at lunch).
Focus Groups	In-depth dialog with a representative panel of individuals (e.g., iPod users).

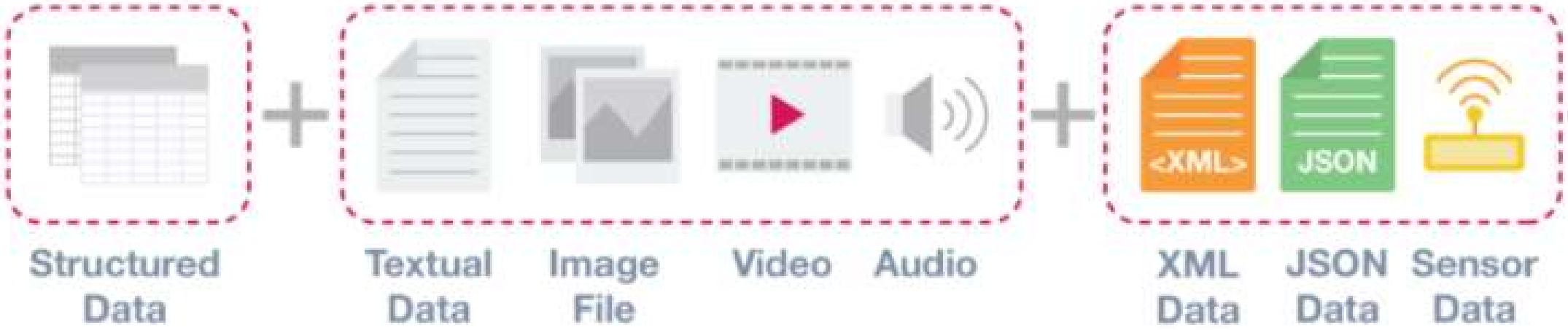
資料與變項

Structured data

- defined rows and columns
- SAS, Microsoft Access and Excel, Oracle, Teradata, Hadoop, and others
- engines enable SAS to read structured data

Unstructured data

- no defined columns
- text, delimited, JSON, weblogs, and others
- must be imported into SAS


<XML>
JSON
Structured Data
Textual Data
Image File
Video
Audio
XML Data
JSON Data
Sensor Data

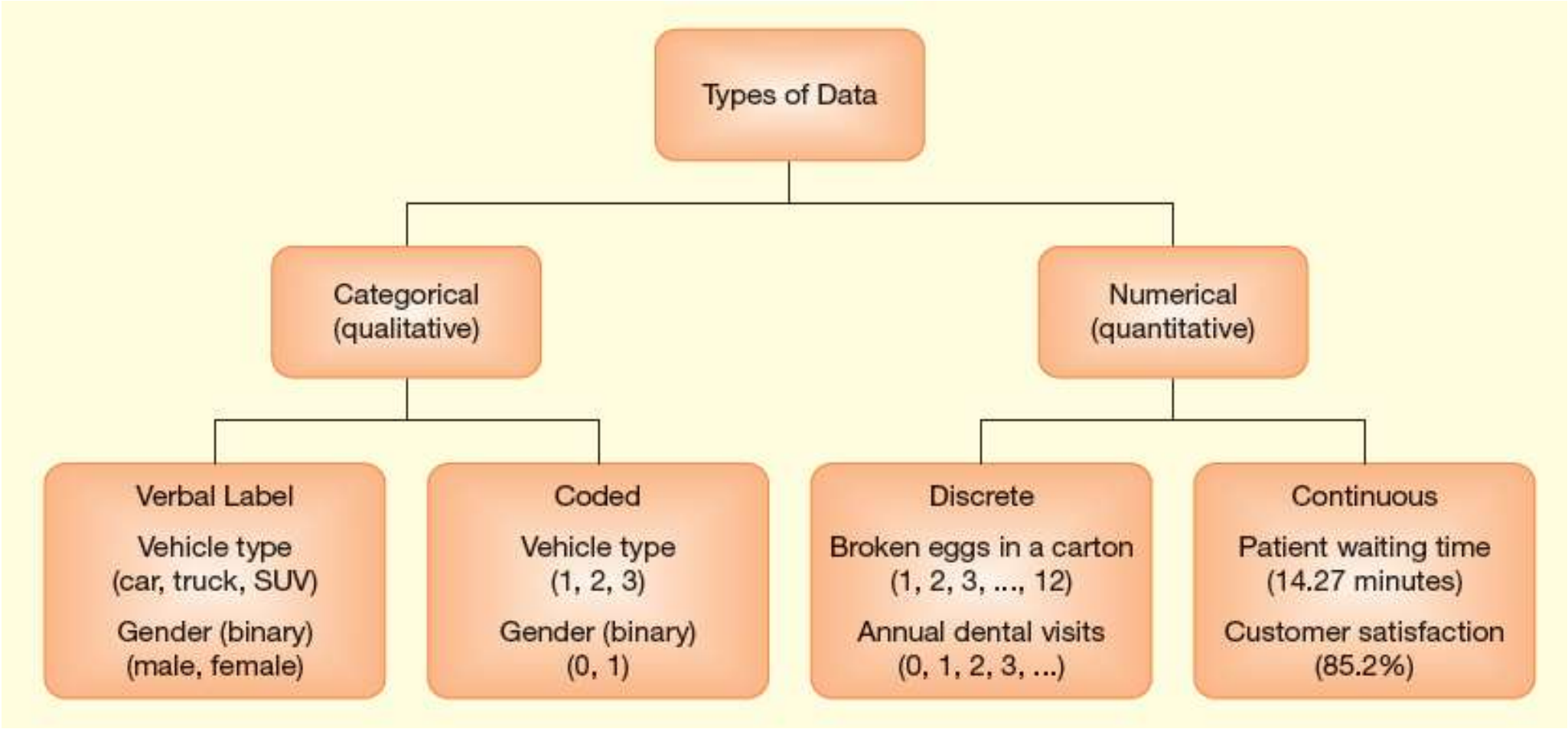
結構化資料範例

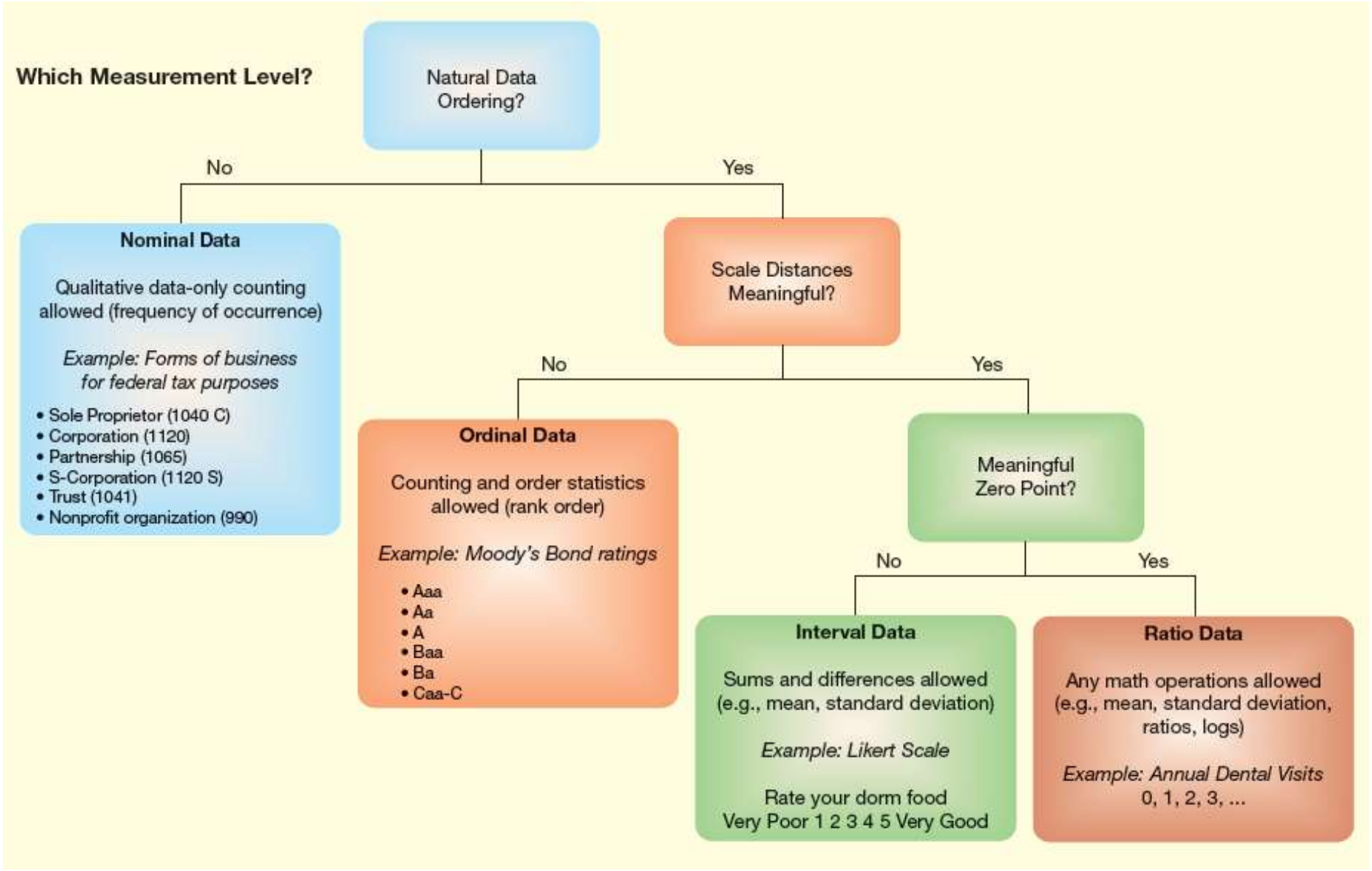
ID	SEX	YR_BRTH	YEAR_DX	RACE	Agedx	Surv_time	RADIATN
9781609(	1	1942	2004	1	61	41	1
9782257[	1	1942	2005	2	62	25	1
9784915[	2	1940	2004	1	64	0	1
9792544(	1	1941	2007	1	66	57	1
9794374[	1	1925	2007	1	82	5	1
9850897(	2	1923	2007	1	83	82	1
9851009(	2	1938	2008	2	70	5	1
9852263[	1	1944	2008	1	63	66	1
9853457[	2	1949	2008	1	58	66	1
9853899[	1	1932	2009	1	76	59	1
9853913[	2	1941	2008	1	66	39	1
9854757[	2	1930	2008	1	78	61	1

- **橫列**表示
  觀察值
  Observation
  Row
- **直行**表示
  變項
  Variable
  Column

常見的變項類型

- 連續變項 (continuous variables)
  - 例如氣溫、身高體重
- 類別變項 (categorical variables)
  - 名目型 (nominal)
    - 二分類 (Binary/dichotomous)、多組 (multinomial)
    - 例如性別、年齡分組
  - 次序型 (ordinal)
    - 數值大小有順序的分別
    - 例如滿意度、疾病嚴重度
- 日期 / 時間


Types of Data
Categorical (qualitative)
Numerical (quantitative)
Verbal Label
Vehicle type (car, truck, SUV)
Gender (binary) (male, female)
Coded
Vehicle type (1, 2, 3)
Gender (binary) (0, 1)
Discrete
Broken eggs in a carton (1, 2, 3, ..., 12)
Annual dental visits (0, 1, 2, 3, ...)
Continuous
Patient waiting time (14.27 minutes)
Customer satisfaction (85.2%)


Which Measurement Level?
Natural Data Ordering?
No
Yes
Nominal Data
Qualitative data-only counting allowed (frequency of occurrence)
Example: Forms of business for federal tax purposes
• Sole Proprietor (1040 C)
• Corporation (1120)
• Partnership (1065)
• S-Corporation (1120 S)
• Trust (1041)
• Nonprofit organization (990)
Scale Distances Meaningful?
No
Yes
Ordinal Data
Counting and order statistics allowed (rank order)
Example: Moody's Bond ratings
• Aaa
• Aa
• A
• Baa
• Ba
• Caa-C
Meaningful Zero Point?
No
Yes
Interval Data
Sums and differences allowed (e.g., mean, standard deviation)
Example: Likert Scale
Rate your dorm food
Very Poor 1 2 3 4 5 Very Good
Ratio Data
Any math operations allowed (e.g., mean, standard deviation, ratios, logs)
Example: Annual Dental Visits
0, 1, 2, 3, ...

關聯式資料庫

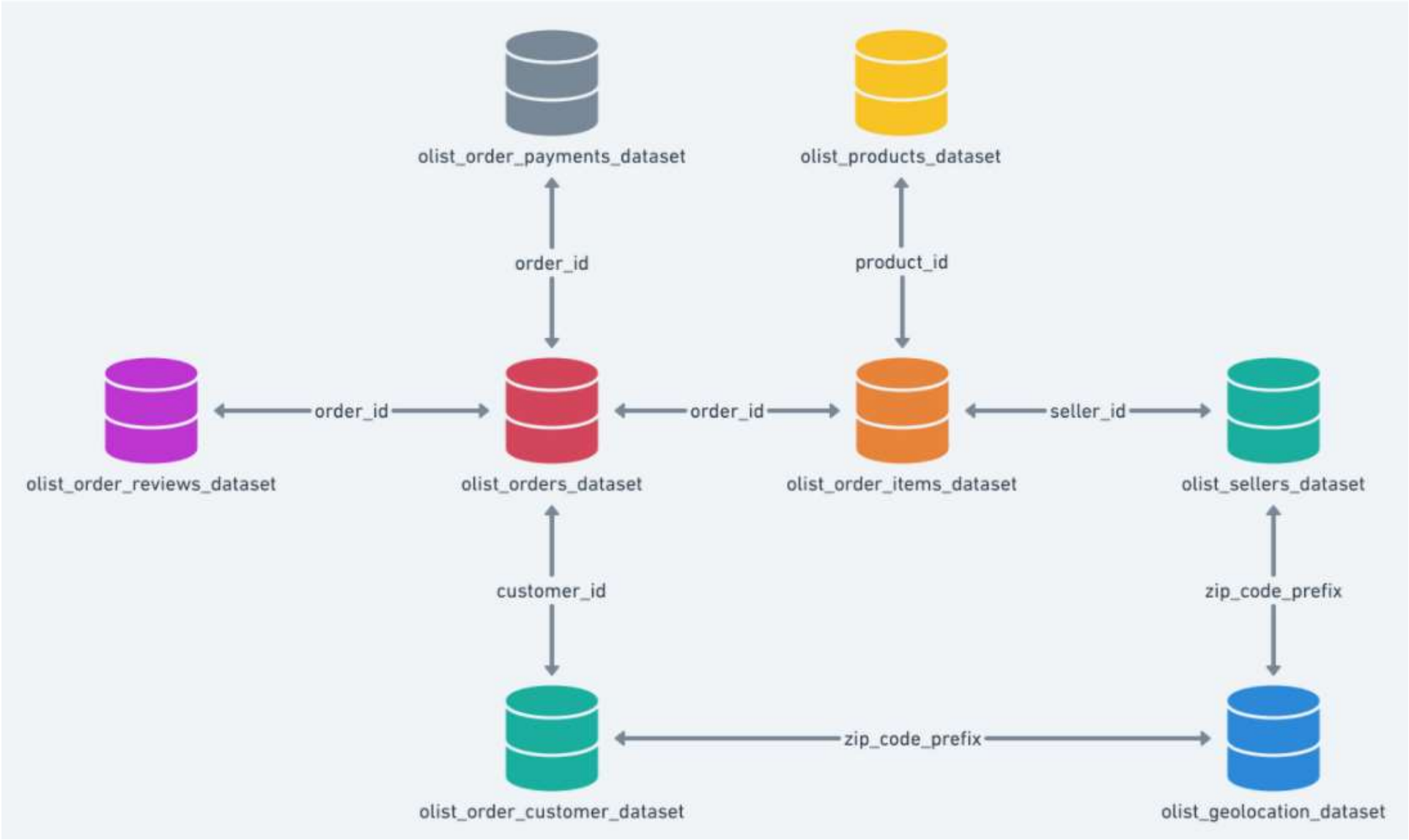
- https://www.kaggle.com/olistbr/brazilian-ecommerce
- Provided by Olist (www.olist.com)

資料描述

- Olist
  - 巴西購物平台
  - 尋找賣方，在平台上銷售商品
  - 利用Olist物流廠商
- 將近100,000筆交易紀錄
- 2016年~2018年

檔案

Customers Dataset
Geolocation Dataset
Order Items Dataset
Payments Dataset
Order Reviews Dataset
Order Dataset
Products Dataset
Sellers Dataset
Category Name Translation


olist_order_payments_dataset
olist_products_dataset
order_id
product_id
order_id
order_id
order_id
seller_id
olist_order_reviews_dataset
olist_orders_dataset
olist_order_items_dataset
olist_sellers_dataset
customer_id
zip_code_prefix
zip_code_prefix
olist_order_customer_dataset
olist_geolocation_dataset

課堂討論與練習 – 什麼是資料？

請參考Week 1檔案.xlsx” 來店人數分析

1. 請問這是報表還是資料？
2. 請嘗試修改成一個可用來分析的資料

- 時間：5分鐘

課後作業

- 請具體寫出一個今天學習到的統計概念 (字數不限)