巨量資料管理學院碩士在職專班

統計分析

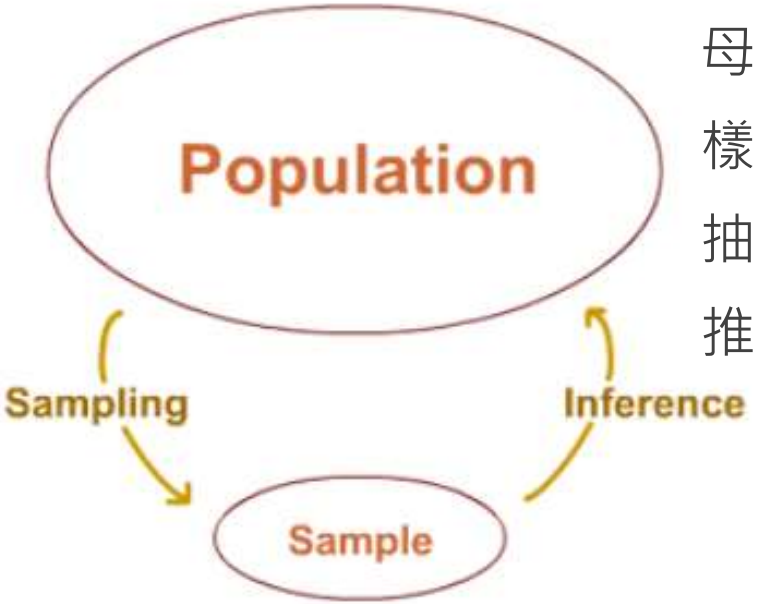
2022/9/23

陳光宏

統計基本概念

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

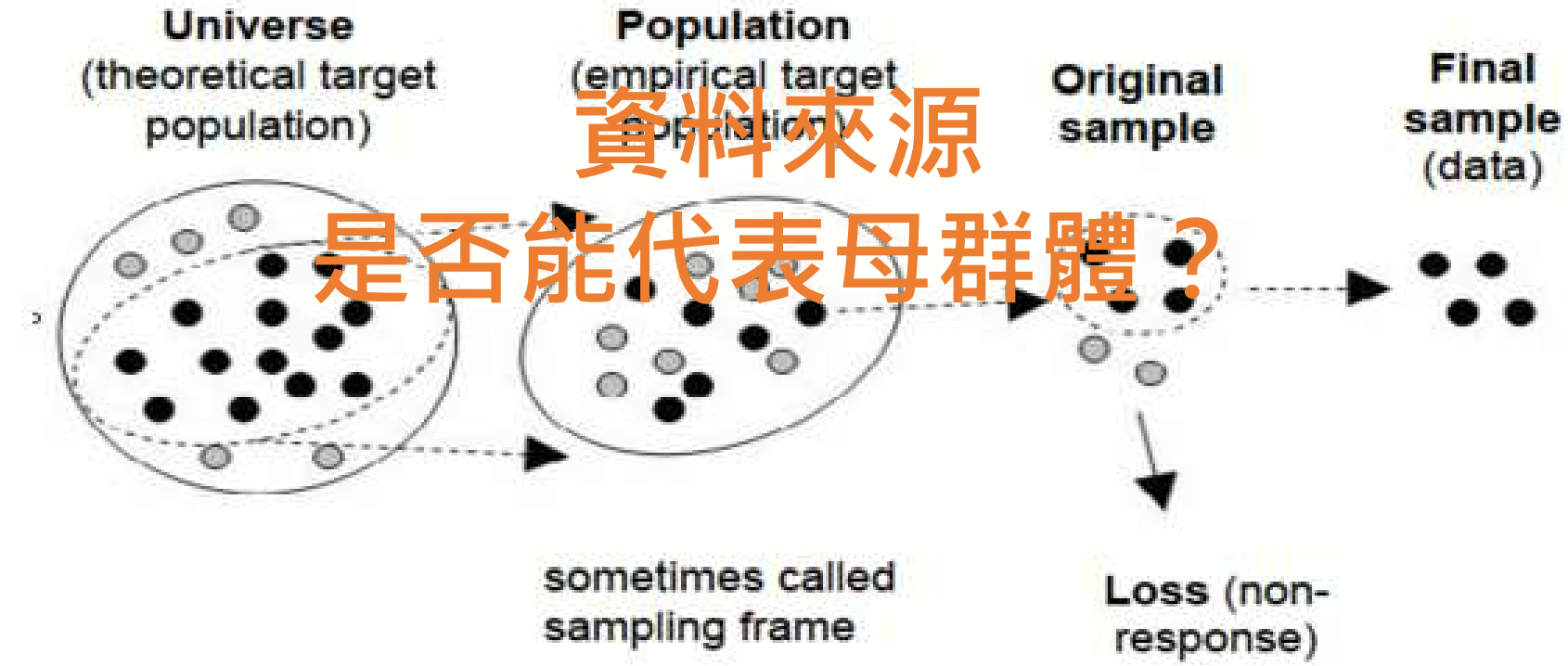
抽樣 Sampling

推論 Inference

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



資料來源
是否能代表母群體？

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

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

報表 vs. 資料

	各時段來店人數分析					
時段	12月28日	12月29日	12月30日	12月31日	1月1日	1月2日
10am	61	95	74	89	182	191
11am	178	184	113	183	289	356
12pm	213	207	114	240	366	433
1pm	273	233	141	245	422	495
2pm	297	276	211	295	499	561
3pm	345	326	278	321	588	553
4pm	255	222	243	251	621	528
5pm	140	107	120	115	488	381
6pm	90	84	51	70	379	320
7pm	99	75	67	31	325	279
8pm	83	72	59	72	255	254

date	time_slot	num
12月28日	10am	61
12月28日	11am	178
12月28日	12pm	213
12月28日	1pm	273
12月28日	2pm	297
12月28日	3pm	345
12月28日	4pm	255
12月28日	5pm	140
12月28日	6pm	90
12月28日	7pm	99
12月28日	8pm	83
12月29日	10am	95
12月29日	11am	184
12月29日	12pm	207
12月29日	1pm	233
12月29日	2pm	276
12月29日	3pm	326
12月29日	4pm	222
12月29日	5pm	107
12月29日	6pm	84
12月29日	7pm	75
12月29日	8pm	72

課堂討論與練習 – 資料

請參考 “Week 3檔案.xlsx” Apple資料

1. 請問這是資料還是報表？
2. 請問這個檔案作為資料，有什麼不足的地方？

廠牌	機型	配備	價格	銷售數量	銷售業績
Apple	iPhone 7	簡配	$11,688	2	$23,376
Apple	iPhone 8	全配	$12,999	1	$12,999
Apple	iPhone 6S	簡配	$7,388	3	$22,164
Apple	iPhone SE	簡配	$7,999	4	$31,996
Apple	iPhone X	簡配	$11,788	3	$35,364
Apple	iPhone 5	簡配	$3,888	2	$7,776
Apple	iPhone 6S	簡配	$7,388	1	$7,388
Apple	iPhone 7S	簡配	$11,366	1	$11,366
Apple	iPhone SE	簡配	$7,999	2	$15,998
Apple	iPhone 5	簡配	$3,888	2	$7,776
Apple	iPhone 7S	簡配	$11,366	2	$22,732
Apple	iPhone 8	全配	$12,999	2	$25,998
Apple	iPhone X	簡配	$11,788	1	$11,788
Apple	iPhone 7	簡配	$11,688	3	$35,064
Apple	iPhone 6S	簡配	$7,388	2	$14,776
Apple	iPhone 8	全配	$12,999	2	$25,998
Apple	iPhone 6S	簡配	$7,388	1	$7,388
Apple	iPhone 5	簡配	$3,888	3	$11,664
Apple	iPhone SE	簡配	$7,999	1	$7,999
Apple	iPhone 6S	簡配	$7,388	4	$29,552
Apple	iPhone SE	簡配	$7,999	2	$15,998
Apple	iPhone 8	全配	$12,999	2	$25,998

譯碼簿 Coding book

欄位序號	英文名稱	中文名稱	屬性	長度	備註
1	ID	交易序號	num	10	
2	date	交易日期	date	10	yymmdd10.
3	brand	廠牌	char	15	
4	type	機型	char	20	
5	equip	配備	char	8	中文
6	price	價格	num	6	
7	qty	銷售數量	num	4	
8	total	銷售業績	num	15	價格x銷售數量

描述統計

概念

- 目的：了解資料狀況
- 區分變項類型
  - 類別、連續
- 連續變項
  - 中心位置的測量
  - 散佈的測量
  - 分佈的形狀
- 類別變項
  - 列聯表

Where are the data values concentrated? What seem to be typical or middle data values? Is there central tendency?
How much dispersion is there in the data? How spread out are the data values? Are there unusual values?
Are the data values distributed symmetrically? Skewed? Sharply peaked? Flat? Bimodal?

類別變項

- 呈現各評分等級的百分比
- n, %

Table 1.1 U.S. Education Rating by 400 Educators

Rating	Frequency
A	35
B	260
C	93
D	12
Total	400

Table 1.2 Calculations for the Pie Chart in Example 1.3

Rating	Frequency	Relative Frequency	Percent	Angle
A	35	35/400 = .09	9%	.09 × 360 = 32.4°
B	260	260/400 = .65	65%	234.0°
C	93	93/400 = .23	23%	82.8°
D	12	12/400 = .03	3%	10.8°
Total	400	1.00	100%	360°

範例

Table 1. Baseline characteristics in patients with hepatitic C virus infection

Variable	n	%
Age[a]	63.1 ± 12.9	
Sex		
Female	340	62.4
Male	205	37.6
Hypertension	220	40.4
Diabetes	156	28.6
Hyperlipidemia	100	18.4
HCV genotype		
1	270	49.5
2 or 3	199	36.5
4 or 5 or 6	43	7.9
Multiple	31	5.7
High HCV viral load	374	68.6

列聯表 (Contingence tables)

- 類別變項

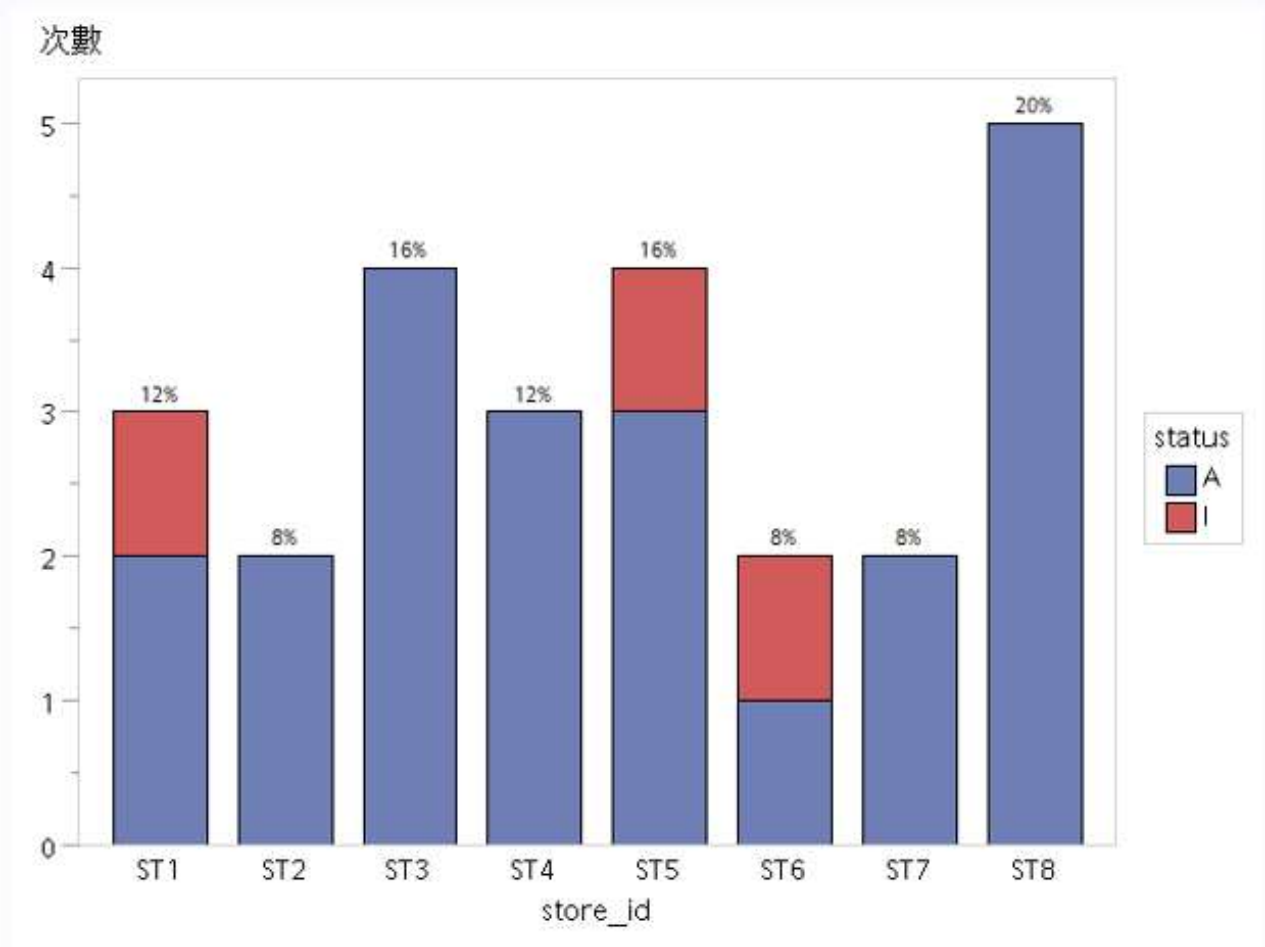
- 交叉表 (cross-table)
- 樞紐分析表
- n, %

	Dog	Cat	Total
Male	42	10	52
Female	9	39	48
Total	51	49	100

educational level	*smoking status*			
	never smoked	currently smoke	former smoker	totals
did not finish high school	25	40	30	95
high school graduate	30	30	40	100
BS degree	50	10	60	120
totals	105	80	130	315

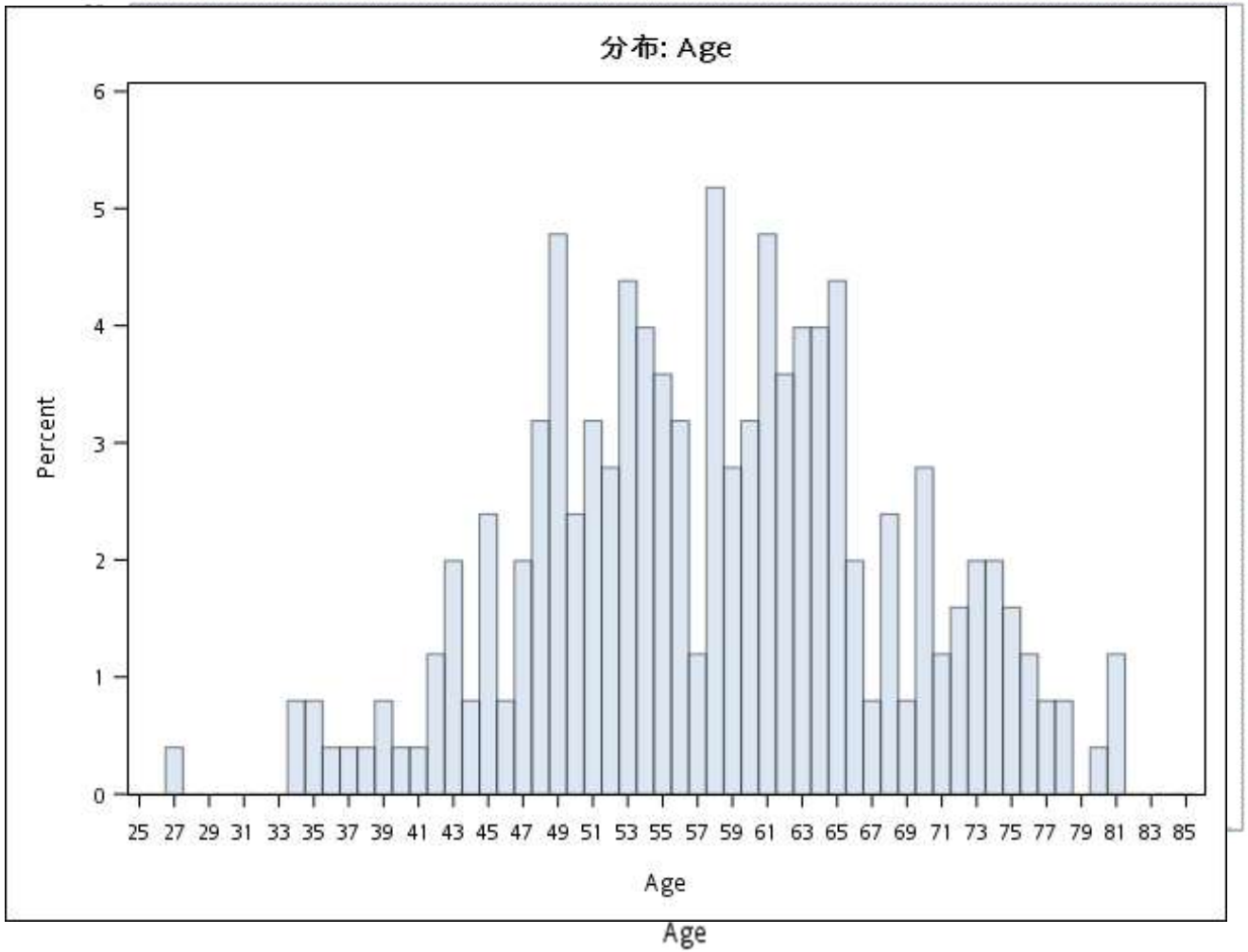
長條圖

- 適用於類別資料
  - 性別、BMI分組

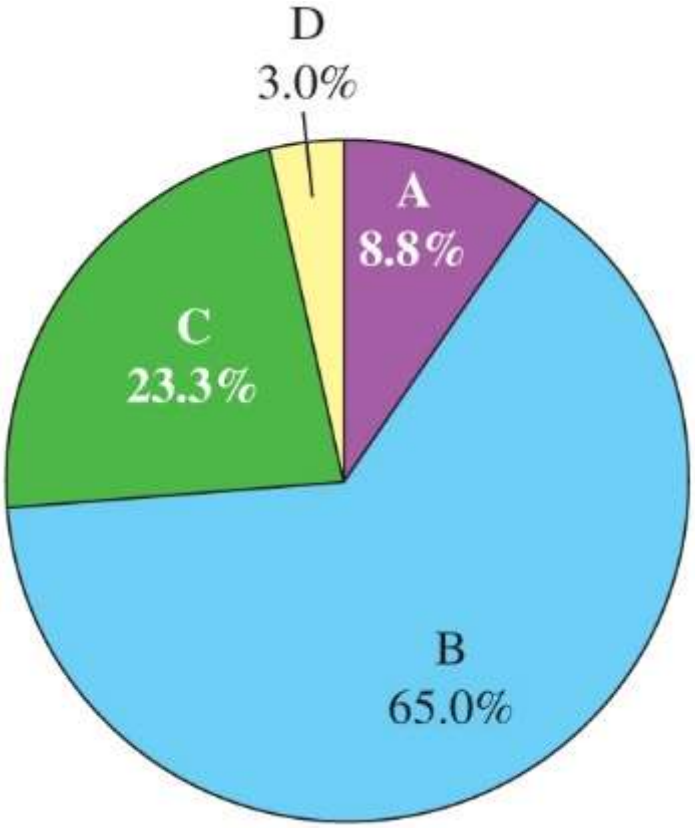


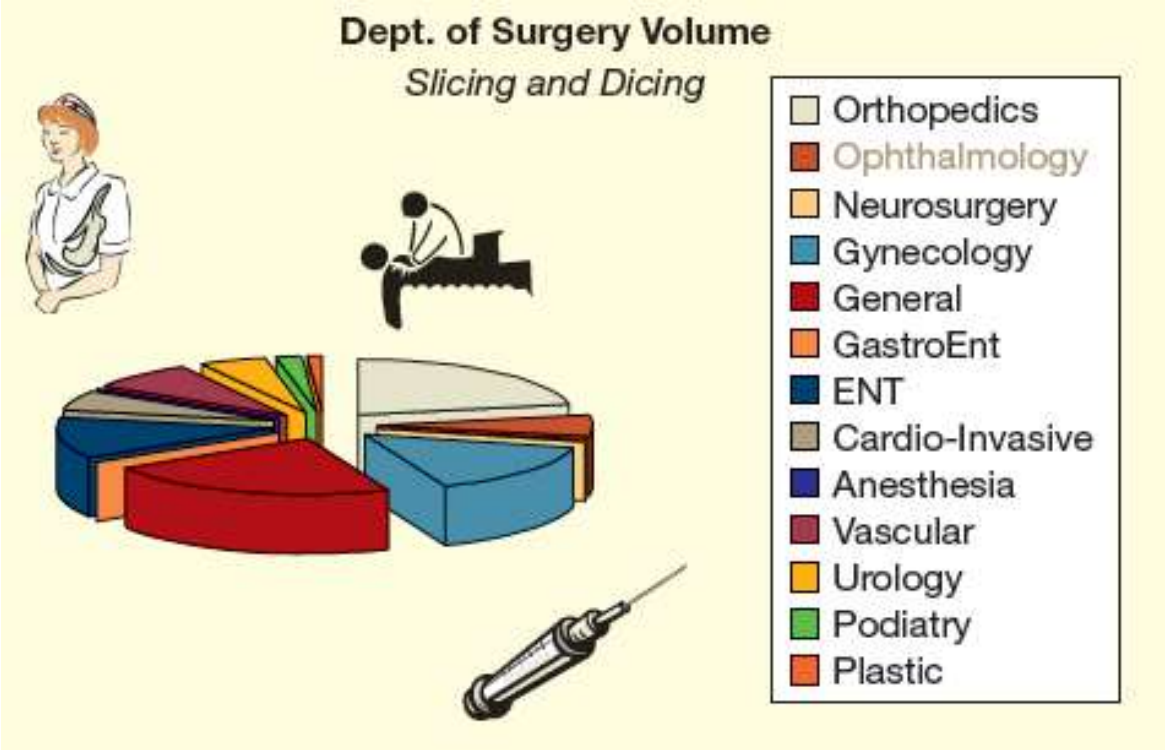
長條圖

- 也可適用於數值資料
  - 例如年齡
  - 但統計軟體通常會自動將年齡分組 (例如 11-20, 21-30, 31-40...)
  - 不建議



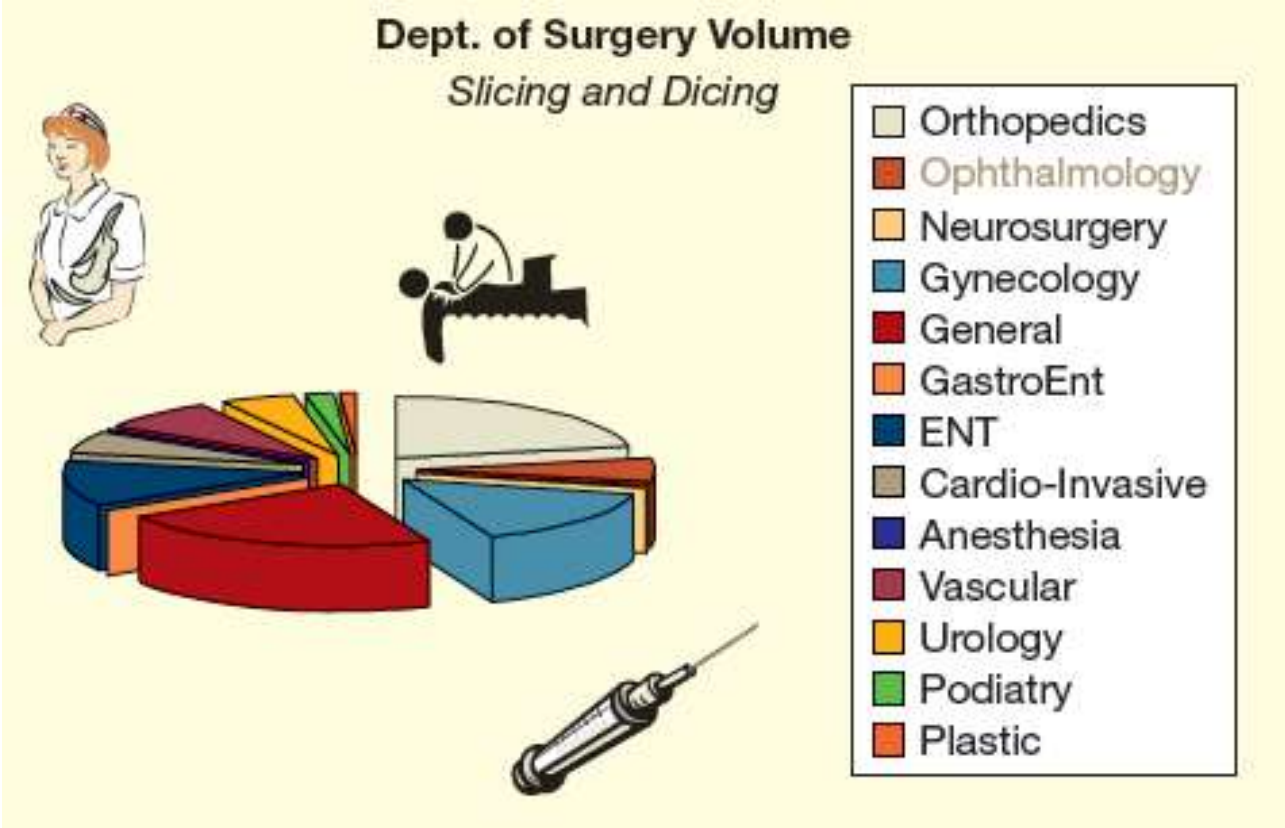
圓餅圖


D
3.0%
A
8.8%
C
23.3%
B
65.0%


Dept. of Surgery Volume
Slicing and Dicing
Orthopedics
Ophthalmology
Neurosurgery
Gynecology
General
GastroEnt
ENT
Cardio-Invasive
Anesthesia
Vascular
Urology
Podiatry
Plastic

課堂討論與練習

- 請討論右圖有什麼不夠好的地方？
- 您會建議如何修改？



連續變項

- 中心位置的測量 (Measures of location)
  - 平均值 (Mean)
  - 中位數 (Median)
- 散佈的測量 (Measures of spread)
  - 標準差 (Standard deviation)
  - 範圍 (Range)
  - 四分位 (Quartiles)
  - 四分位距 (Inter-quartile range)
- 分佈的形狀 (Shape)

中心位置的測量

- 平均值

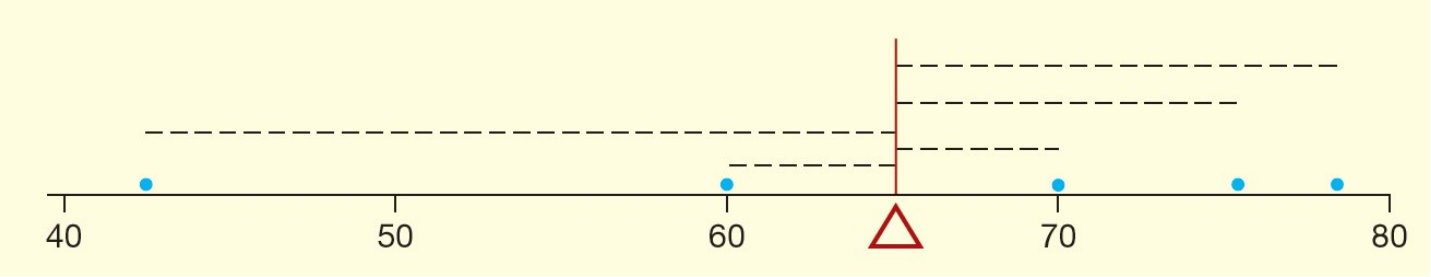
1 2 3 4 5

→ (1+2+3+4+5) / 5 = 3

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- 各點到平均值的距離總和最小

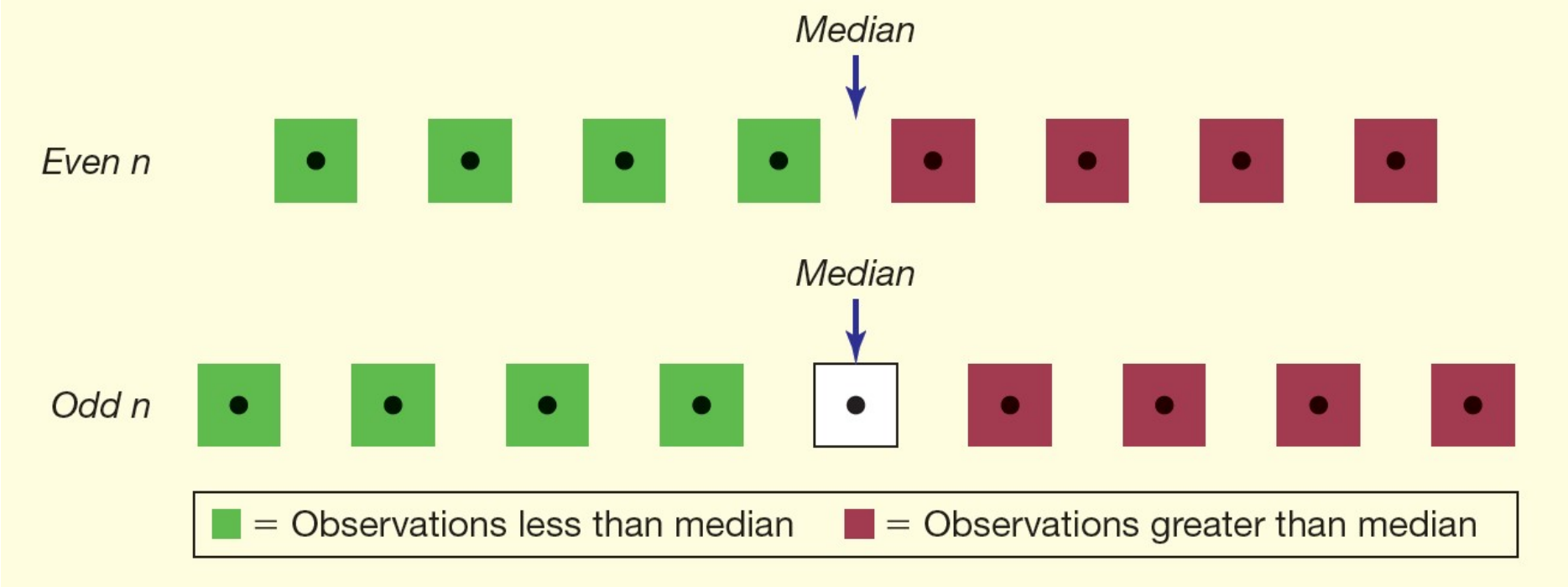
$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0$$



$$\sum_{i=1}^{n} (x_i - \bar{x}) = (42 - 65) + (60 - 65) + (70 - 65) + (75 - 65) + (78 - 65)$$
$$= (-23) + (-5) + (5) + (10) + (13) = -28 + 28 = 0$$

中心位置的測量

- 中位數


Median
Even n
Median
Odd n
= Observations less than median
= Observations greater than median

軟體操作 - EXCEL

- 平均值
  =AVERAGE(資料範圍)

- 中位數
  =MEDIAN(資料範圍)

課堂討論與練習 – 平均值與中位數

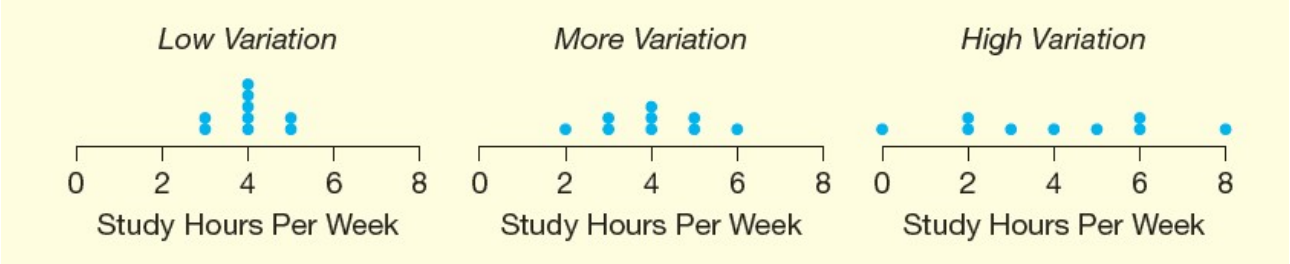
請參考“Week 3檔案.xlsx”Apple資料

1. 請用excel計算銷售業績的平均值與中位數
2. 現在新增一筆交易，售出100隻iPhone 8，請計算銷售業績的平均值與中位數
3. 承2，請問新增該筆交易後，平均值和中位數有什麼變化？

廠牌	機型	配備	價格	銷售數量	銷售業績
Apple	iPhone 7	簡配	$11,688	2	$23,376
Apple	iPhone 8	全配	$12,999	1	$12,999
Apple	iPhone 6S	簡配	$7,388	3	$22,164
Apple	iPhone SE	簡配	$7,999	4	$31,996
Apple	iPhone X	簡配	$11,788	3	$35,364
Apple	iPhone 5	簡配	$3,888	2	$7,776
Apple	iPhone 6S	簡配	$7,388	1	$7,388
Apple	iPhone 7S	簡配	$11,366	1	$11,366
Apple	iPhone SE	簡配	$7,999	2	$15,998
Apple	iPhone 5	簡配	$3,888	2	$7,776
Apple	iPhone 7S	簡配	$11,366	2	$22,732
Apple	iPhone 8	全配	$12,999	2	$25,998
Apple	iPhone X	簡配	$11,788	1	$11,788
Apple	iPhone 7	簡配	$11,688	3	$35,064
Apple	iPhone 6S	簡配	$7,388	2	$14,776
Apple	iPhone 8	全配	$12,999	2	$25,998
Apple	iPhone 6S	簡配	$7,388	1	$7,388
Apple	iPhone 5	簡配	$3,888	3	$11,664
Apple	iPhone SE	簡配	$7,999	1	$7,999
Apple	iPhone 6S	簡配	$7,388	4	$29,552
Apple	iPhone SE	簡配	$7,999	2	$15,998
Apple	iPhone 8	全配	$12,999	2	$25,998

散佈/變異的測量
Dispersion/Spread/Variability

- 標準差 (SD)
  - 資料的散佈情況



1 2 3 4 5, mean=3

$$SD = \sqrt{\frac{(1-3)^2+(2-3)^2+(3-3)^2+(4-3)^2+(5-3)^2}{5-1}}$$

- 變異數 (Variance) = $SD^2 = \frac{\text{平方和 } Sum\ of\ squares\ (SS)}{\text{自由度 } Degree\ of\ freedom\ (df)}$

標準差 Standard deviation

DEFINITION

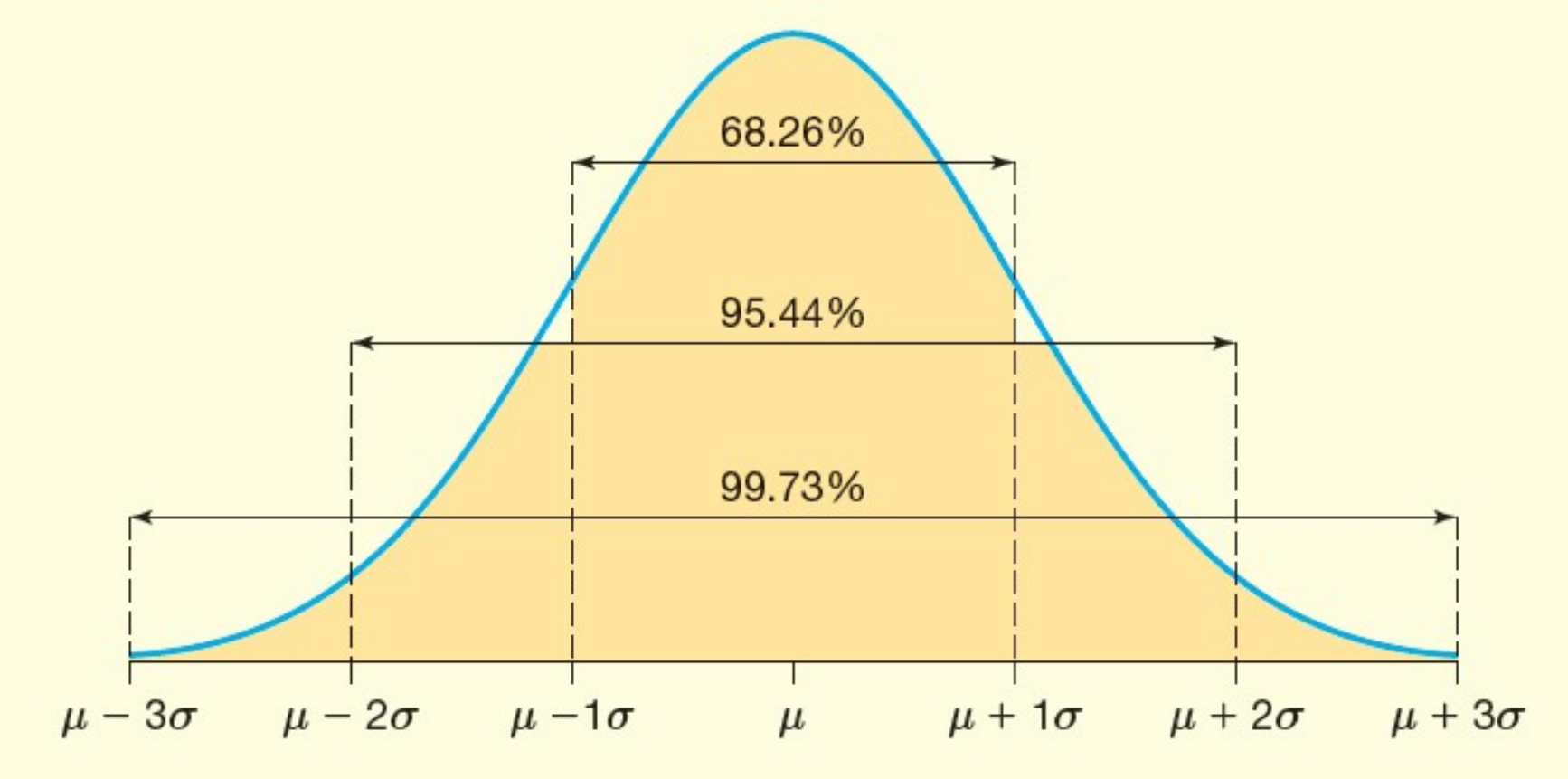
The **variance of a sample** of n measurements is the sum of the squared deviations of the measurements about their mean $\bar{x}$ divided by $(n-1)$. The sample variance is denoted by s^2 and is given by the formula

$$s^2 = \frac{\Sigma\left(x_i - \bar{x}\right)^2}{n-1}$$

1 2 3 4 5, mean=3

$$SD = \sqrt{\frac{(1-3)^2+(2-3)^2+(3-3)^2+(4-3)^2+(5-3)^2}{5-1}}$$

從標準差了解資料散佈狀況



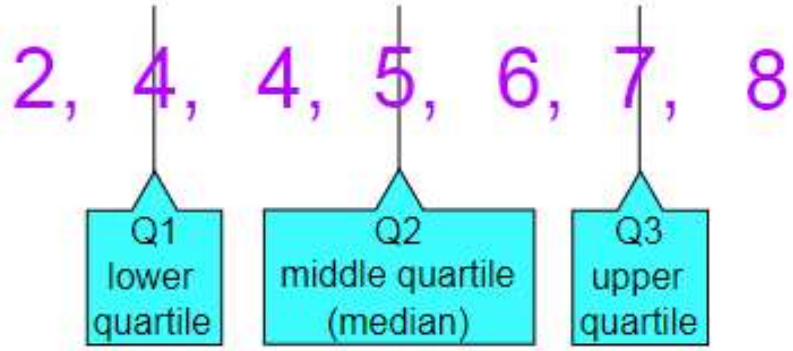
平均値±標準差

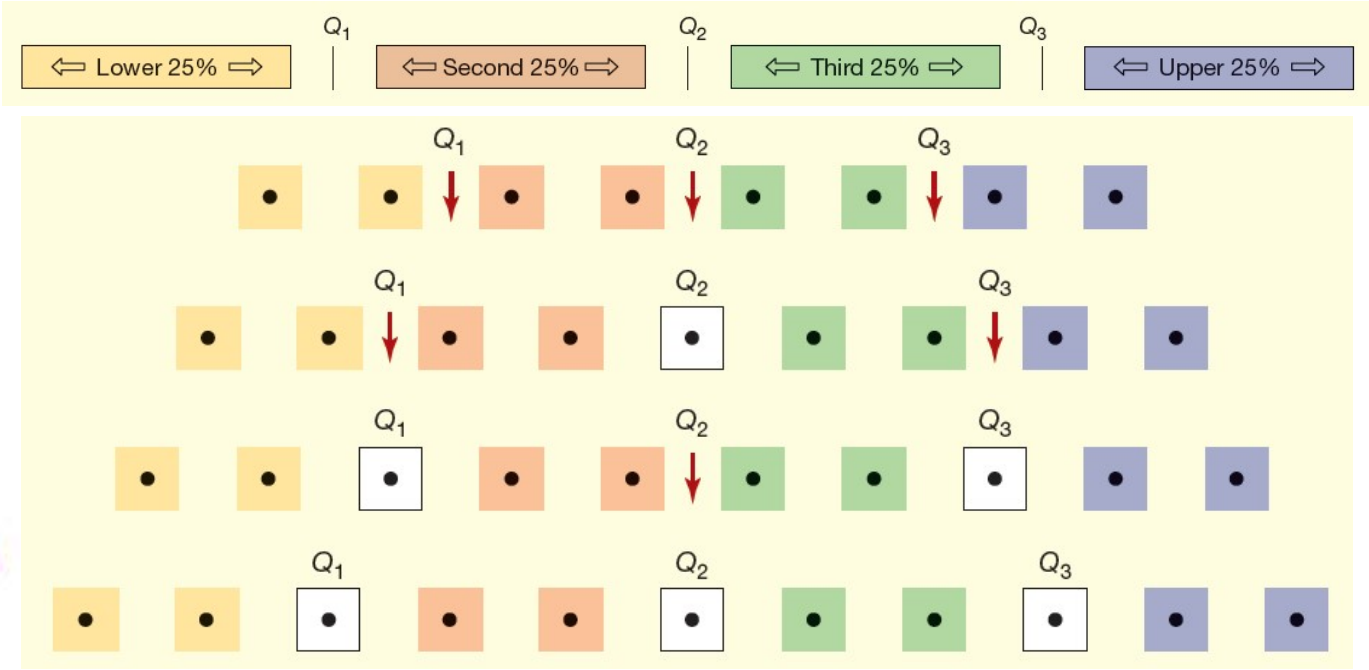
Table 1. Comparisons of baseline serum levels of biomedical measurements between patients in polycystic ovarian syndrome (PCOS) and control group, stratified by obesity status

	Control group		PCOS group		
Variables[a]	Nonobese (n=111)	Obese (n=102)	Nonobese (n=213)	Obese (n=232)	p-value[b]
BMI (kg/m^2)	23.18 ± 1.92	30.15 ± 0.84*	23.57 ± 2.28	30.67 ± 1.35§	<0.001
Testosterone (nmol/	1.93 ± 0.32	1.98 ± 0.37	2.91 ± 0.54†	2.87 ± 0.33‡	<0.001
LH/FSH ratio	0.65 ± 0.09	0.81 ± 0.11*	1.83 ± 0.32†	1.93 ± 0.39‡,§	<0.001
HOMA-IR	2.00 ± 0.33	2.08 ± 0.34	3.17 ± 1.48†	3.75 ± 2.47‡,§	<0.001
TSP-1 (ng/mL)	174.79 ± 41.44	224.55 ± 46.72*	102.06 ± 18.66†	112.96 ± 27.77‡,§	<0.001
TGF-β1 (ng/mL)	165.94 ± 32.66	191.21 ± 40.10	491.31 ± 149.74†	616.13 ± 118.19‡,§	<0.001
NF-κB (ng/mL)	276.15 ± 35.70	349.94 ± 53.17*	648.97 ± 296.02†	768.59 ± 158.77‡,§	<0.001

散佈/變異的測量
Dispersion/Spread/Variability

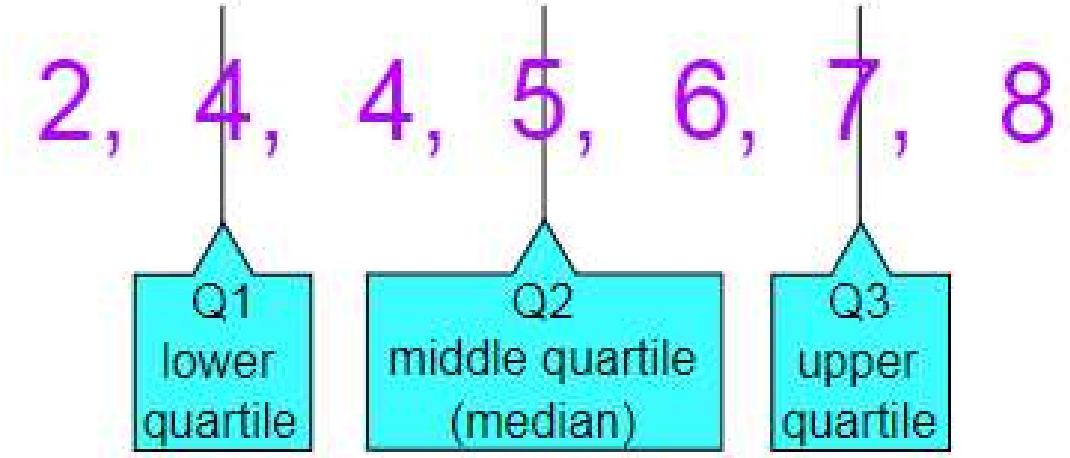
- 四分位數 (Quartile)
  - Q1: 25th 百分位
  - Q2: 50th百分位
  - Q3: 75th百分位



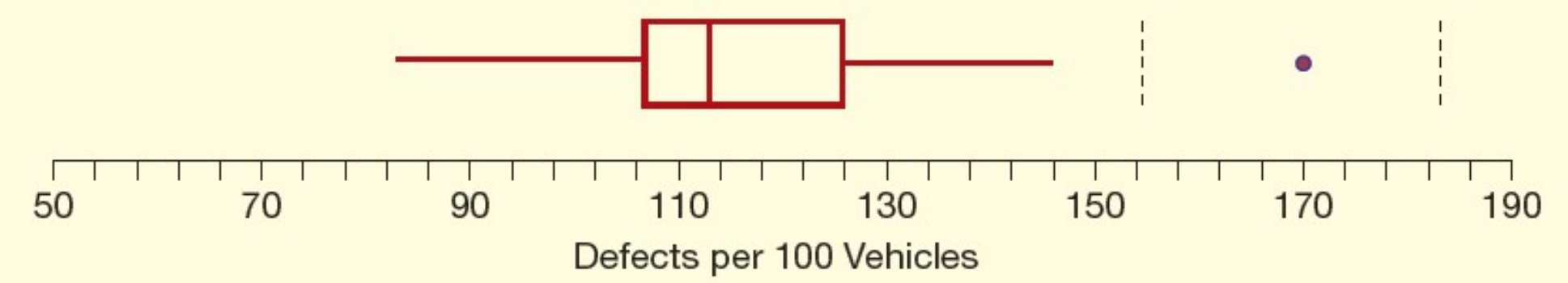


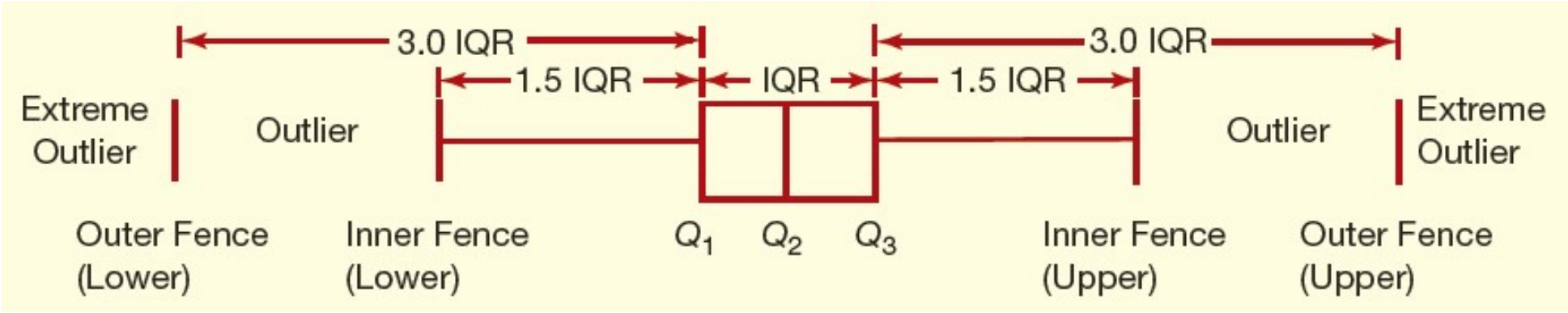
散佈/變異的測量
Dispersion/Spread/Variability

- 四分位距 (Inter-quartile range)
  - Q3-Q1: 7 – 4 = 3
  - Q1~Q3: 4, 7



盒型圖

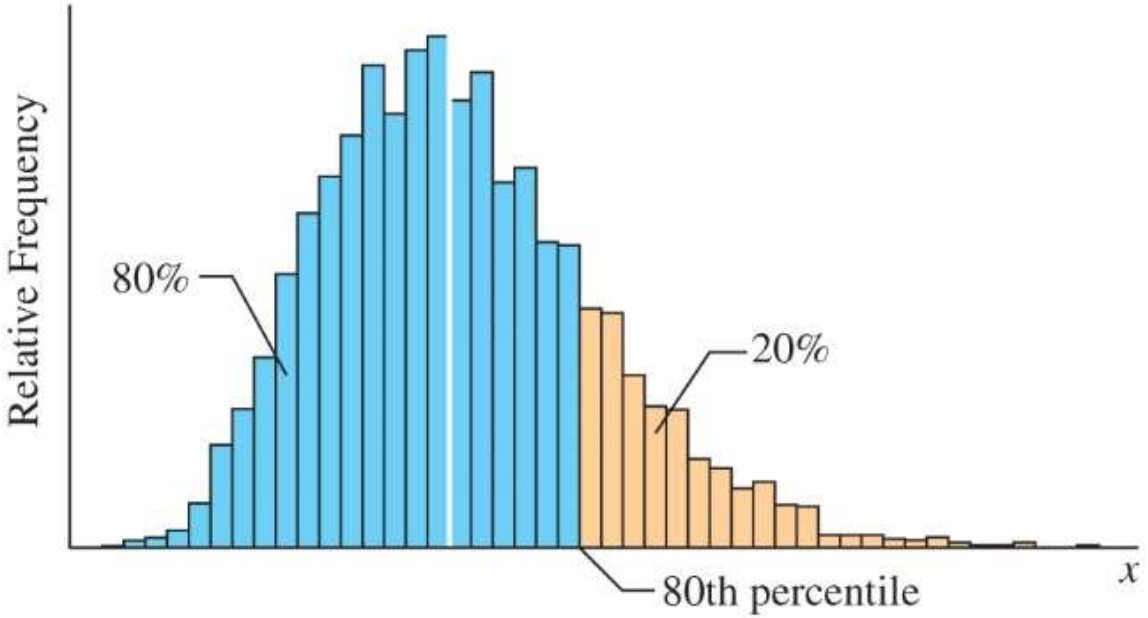

50
70
90
110
130
150
170
190
Defects per 100 Vehicles


3.0 IQR
3.0 IQR
1.5 IQR
IQR
1.5 IQR
Extreme Outlier
Outlier
Outlier
Extreme Outlier
Outer Fence (Lower)
Inner Fence (Lower)
Q_1
Q_2
Q_3
Inner Fence (Upper)
Outer Fence (Upper)

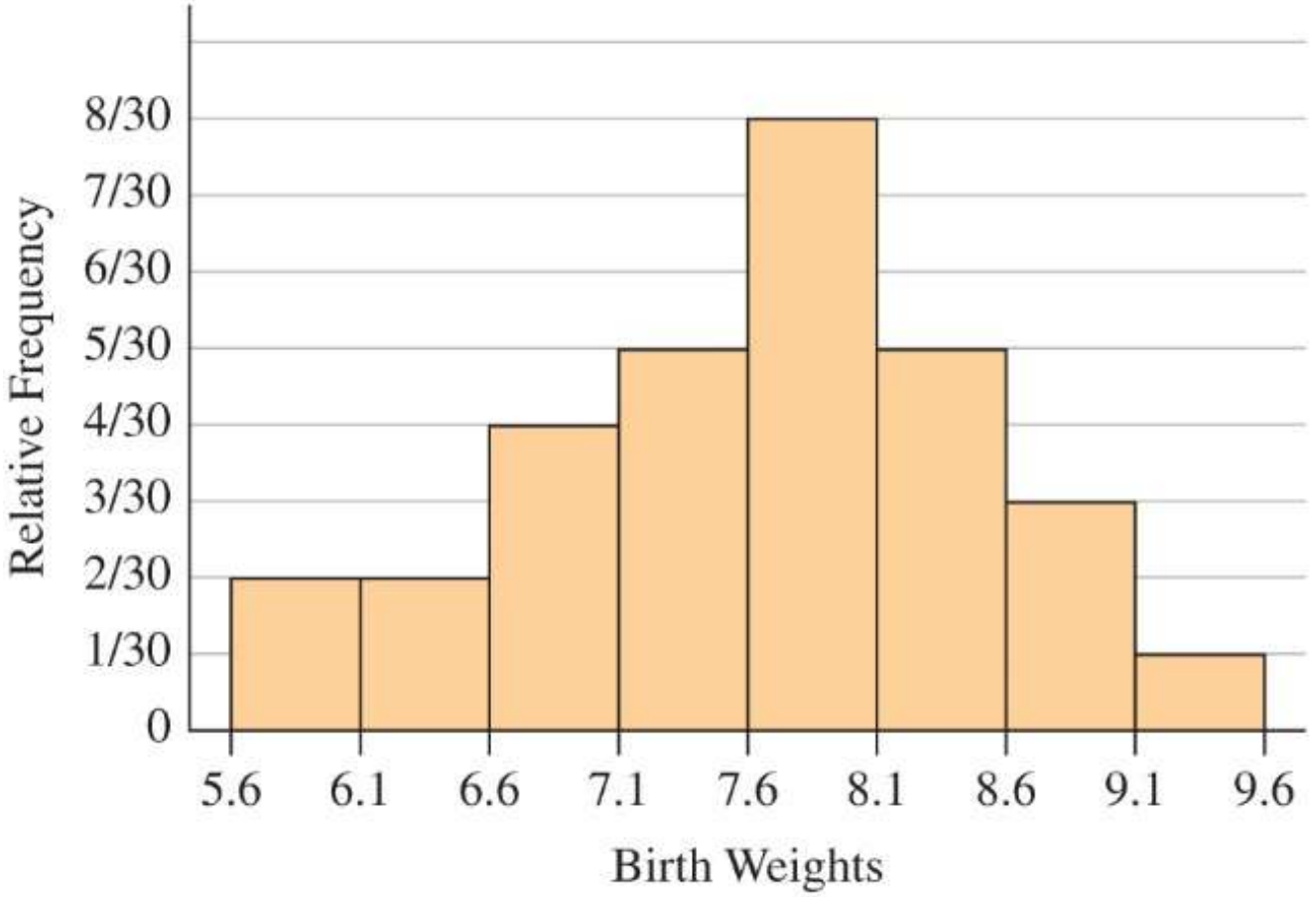
百分位 (Percentile)

DEFINITION

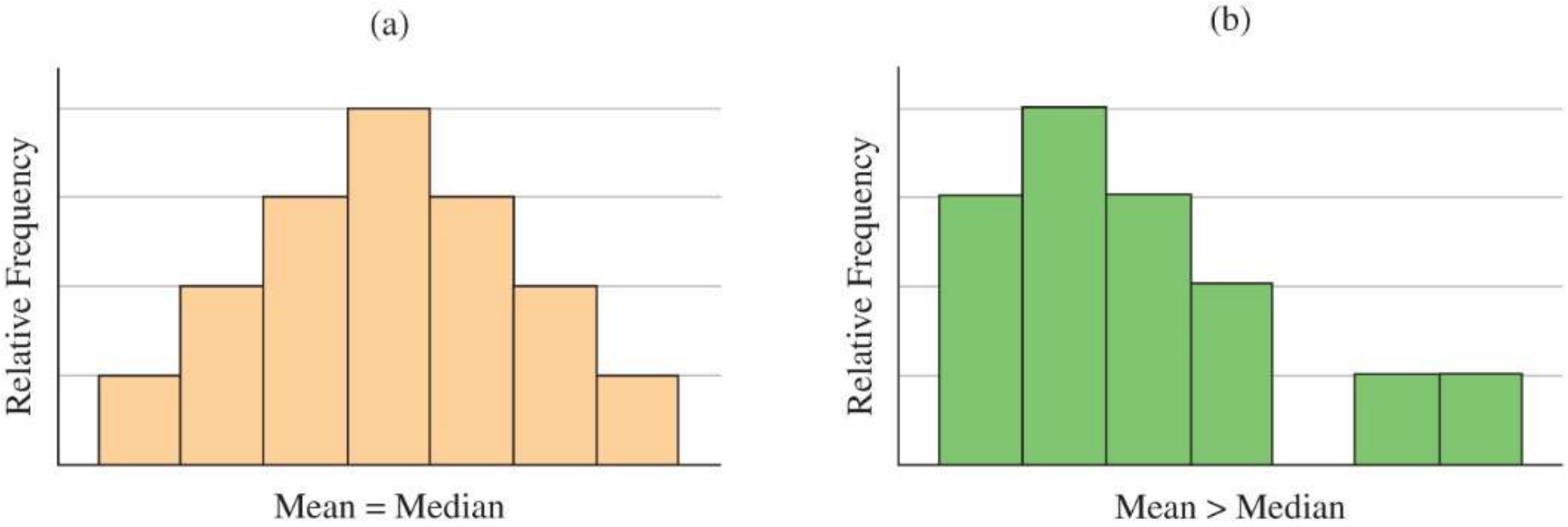
A set of n measurements on the variable x has been arranged from smallest to largest. The ***p*th percentile** is the value of x that is greater than $p\%$ of the measurements and is less than the remaining $(100 - p)\%$.



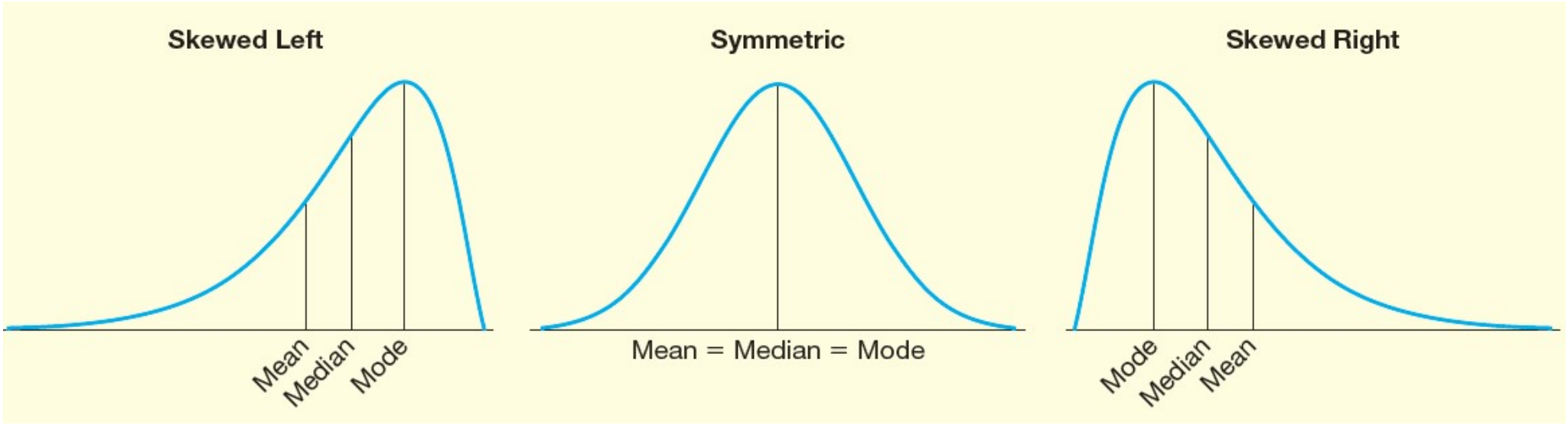
直方圖 (Histogram)



判斷資料分布狀況

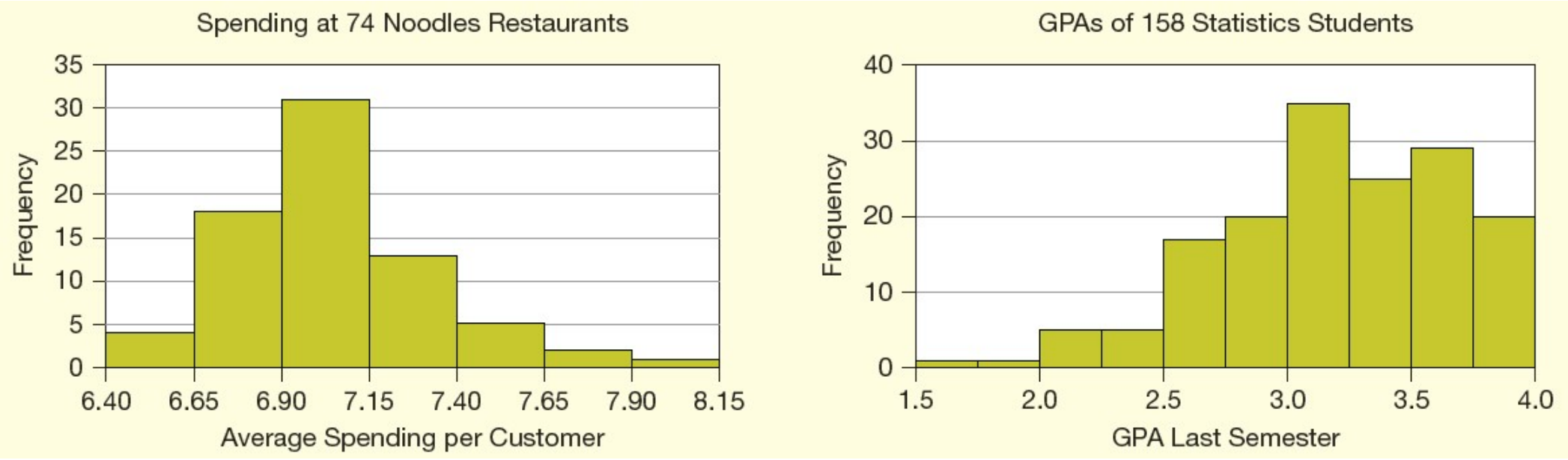

(a)
Relative Frequency
Mean = Median
(b)
Relative Frequency
Mean > Median

分佈的形狀



左偏　　對稱　　右偏

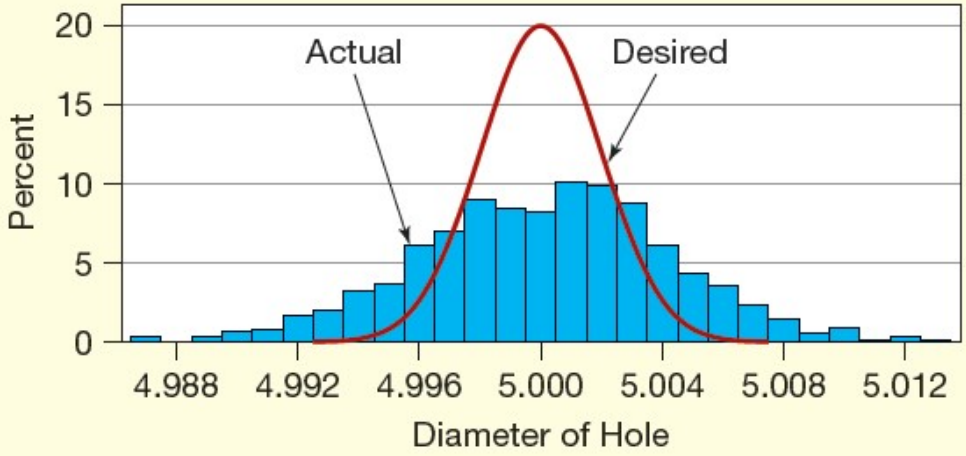
用直方圖呈現


Spending at 74 Noodles Restaurants
Frequency
35
30
25
20
15
10
5
0
6.40
6.65
6.90
7.15
7.40
7.65
7.90
8.15
Average Spending per Customer
GPAs of 158 Statistics Students
Frequency
40
30
20
10
0
1.5
2.0
2.5
3.0
3.5
4.0
GPA Last Semester

中心位置與散佈的應用

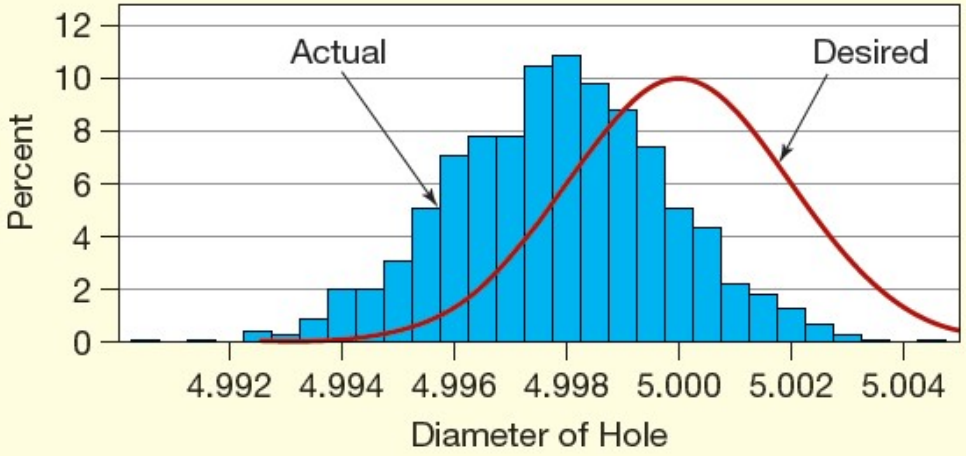
Machine A

Process is correctly centered,
but variation is excessive



Machine B

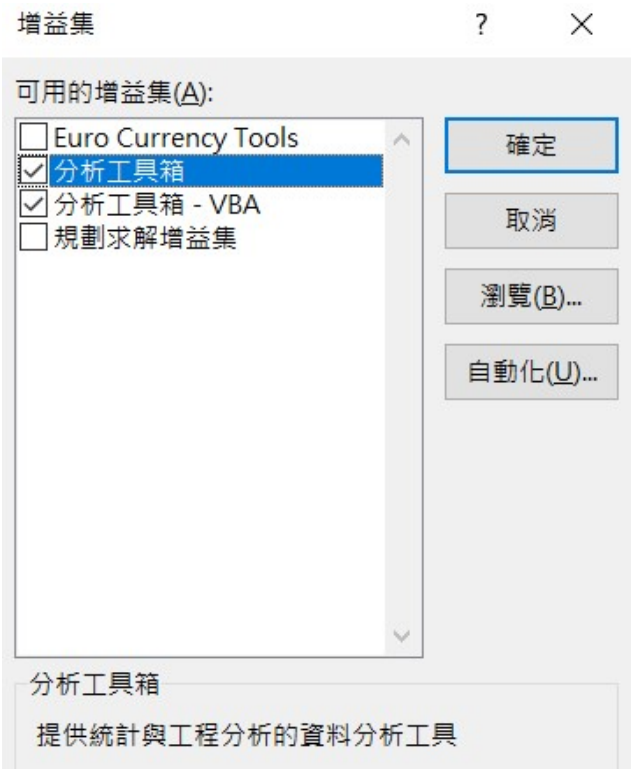
Variability is acceptable, but
process is incorrectly centered



軟體操作 - EXCEL

- 標準差
  =STDEV(資料範圍)
  =STDEV.S(資料範圍)
- 變異數
  =VAR(資料範圍)
  =VAR.S (資料範圍)
- 四分位數
  =QUARTILE.EXC(資料範圍,1)
  =PERCENTILE.EXC(資料範圍,.25)
- 最大值、最小值
  MAX (資料範圍)
  MIN (資料範圍)

- 增益集



兩變項的相關性

- 一個變項X變大 (或變小)，另一個變項Y會不會跟著變大變小

皮爾森相關係數

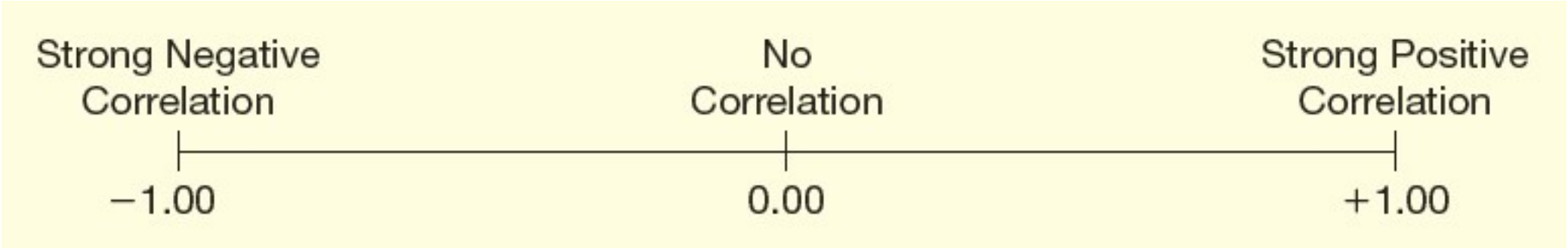
$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

X和Y共同的變異

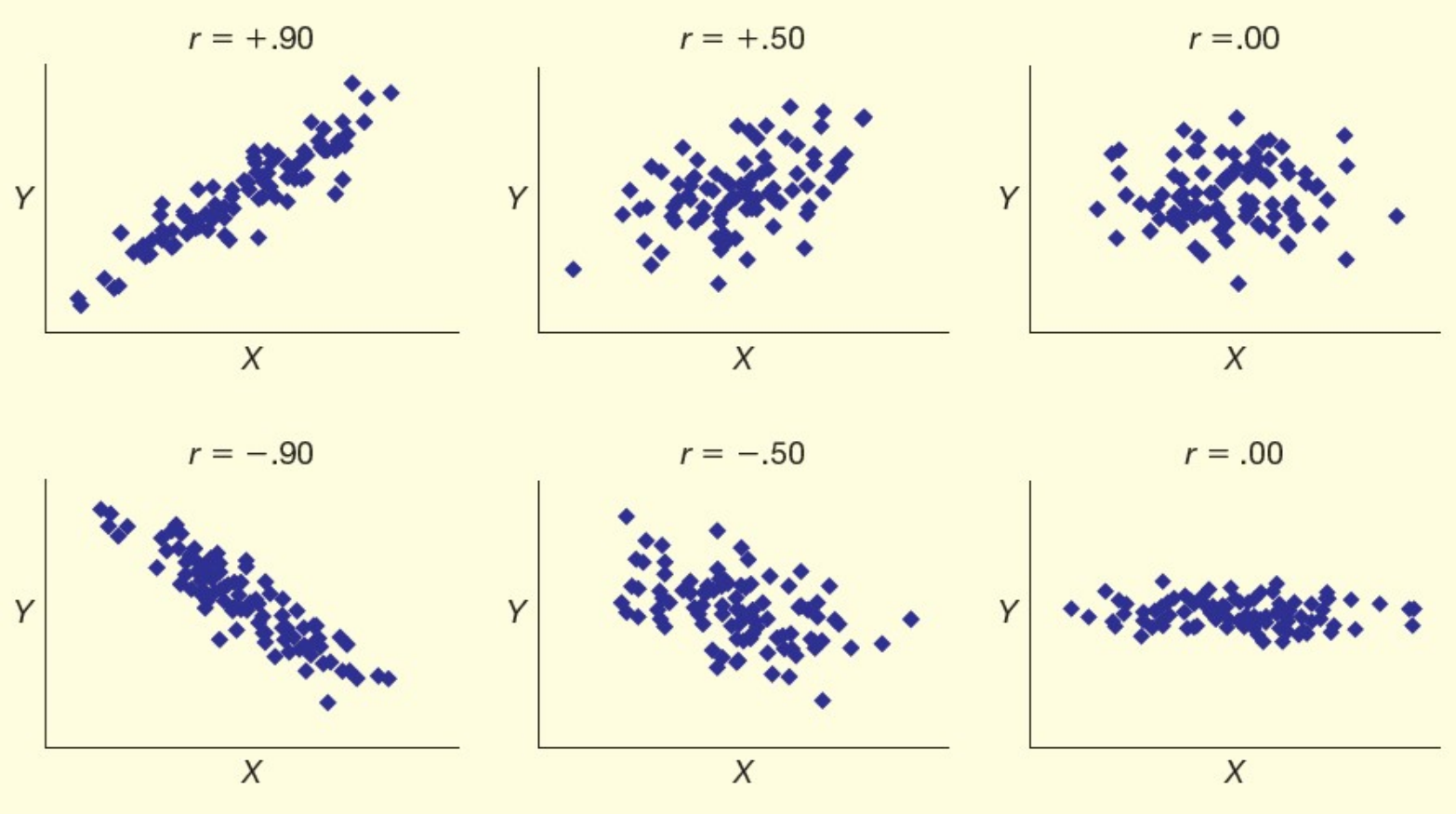
X變項的變異(平方和)

Y變項的變異(平方和)

相關係數的解讀


Strong Negative Correlation
No Correlation
Strong Positive Correlation
−1.00
0.00
+1.00

散佈圖 (Scatter plot)

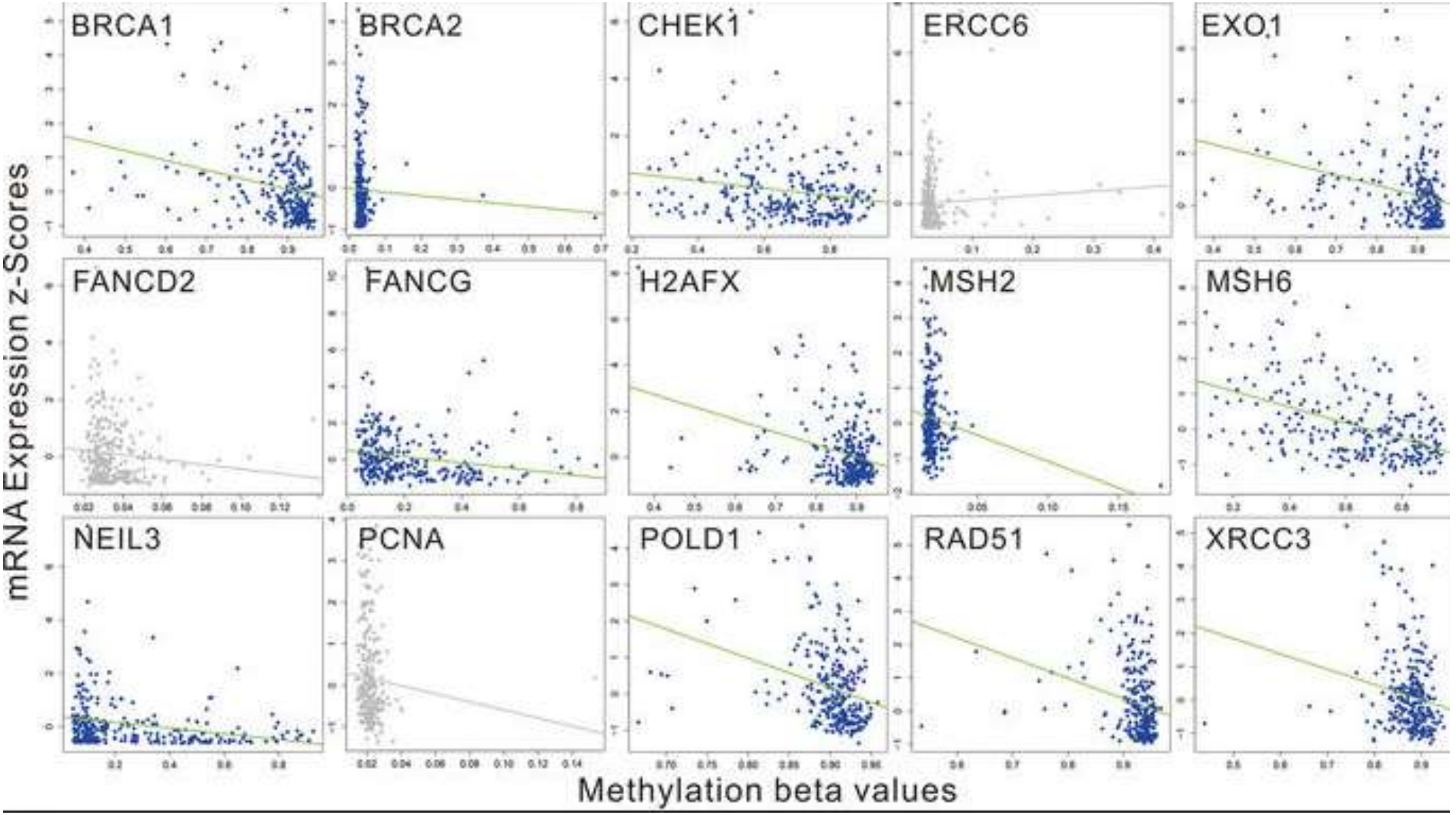


Correlation matrix (相關係數矩陣)

Table I. Pearson correlation matrix for the forest structure components, forest tree abundance (TA), forest logs abundance (LA), snag abundance (AS), leaflitter depth (LD), elevation (ALT), proximity to streams (PS), canopy opening (CO) recorded in 56 areas at Reserva Ducke.

	TA	LA	AS	LD	ALT (m)	PS (m)	CO (%)
TA	1.000						
LA	-0.137	1.000					
AS	0.173	-0.309	1.000				
LD	0.039	0.313	-0.351	1.000			
ALT (m)	0.079	-0.079	0.114	-0.331	1.000		
PS (m)	0.001	0.119	0.088	-0.141	0.493*	1.000	
CO (%)	0.220	0.522*	-0.419*	0.533*	-0.294	-0.150	1.000

* Statistical significance of $p < 0.05$ resulting from the Bonferroni probability matrix used to evaluate how strong and significant were the correlations among the independent variables (forest structure components).


BRCA1
BRCA2
CHEK1
ERCC6
EXO1
FANCD2
FANCG
H2AFX
MSH2
MSH6
NEIL3
PCNA
POLD1
RAD51
XRCC3
mRNA Expression z-Scores
Methylation beta values

討論

- 遺漏值
  - Missing value
- 極端值
  - Outlier, extreme value

課後作業1 – 描述資料

請參考“Week 3檔案.xlsx”Apple資料

1. 請利用excel的資料分析工具，描述銷售業績變項

廠牌	機型	配備	價格	銷售數量	銷售業績
Apple	iPhone 7	簡配	$11,688	2	$23,376
Apple	iPhone 8	全配	$12,999	1	$12,999
Apple	iPhone 6S	簡配	$7,388	3	$22,164
Apple	iPhone SE	簡配	$7,999	4	$31,996
Apple	iPhone X	簡配	$11,788	3	$35,364
Apple	iPhone 5	簡配	$3,888	2	$7,776
Apple	iPhone 6S	簡配	$7,388	1	$7,388
Apple	iPhone 7S	簡配	$11,366	1	$11,366
Apple	iPhone SE	簡配	$7,999	2	$15,998
Apple	iPhone 5	簡配	$3,888	2	$7,776
Apple	iPhone 7S	簡配	$11,366	2	$22,732
Apple	iPhone 8	全配	$12,999	2	$25,998
Apple	iPhone X	簡配	$11,788	1	$11,788
Apple	iPhone 7	簡配	$11,688	3	$35,064
Apple	iPhone 6S	簡配	$7,388	2	$14,776
Apple	iPhone 8	全配	$12,999	2	$25,998
Apple	iPhone 6S	簡配	$7,388	1	$7,388
Apple	iPhone 5	簡配	$3,888	3	$11,664
Apple	iPhone SE	簡配	$7,999	1	$7,999
Apple	iPhone 6S	簡配	$7,388	4	$29,552
Apple	iPhone SE	簡配	$7,999	2	$15,998
Apple	iPhone 8	全配	$12,999	2	$25,998

課後作業2

- 請具體寫出一個今天學習到的統計概念 (字數不限)