「商業資料分析與管理決策」—分析與決策

Spring 2023

授課教師：統計系余清祥

2024年3月31日

授課內容：統計分析

課程下載：csyue.nccu.edu.tw

資料分析策略

- 「觀察」、「推論」、「驗證」三步驟

→首先檢查資料品質，避免Garbage in, garbage out 的窘境 ，通常會佔用一半以上的時間。

→接著是探索性資料分析(EDA)，逐步找出資料重要特性，作為進一步推論(如迴歸分析)的依據。

→驗證性資料分析(CDA)則是最後步驟，分析結果應與EDA接近，否則需重頭檢查。

資料偵錯
資料輸入錯誤、尋找可能的離群值。

初步探索資料特性
資料的集中、散佈趨勢

驗證已知的結果
是否與已知的結果相同？

EDA and CDA

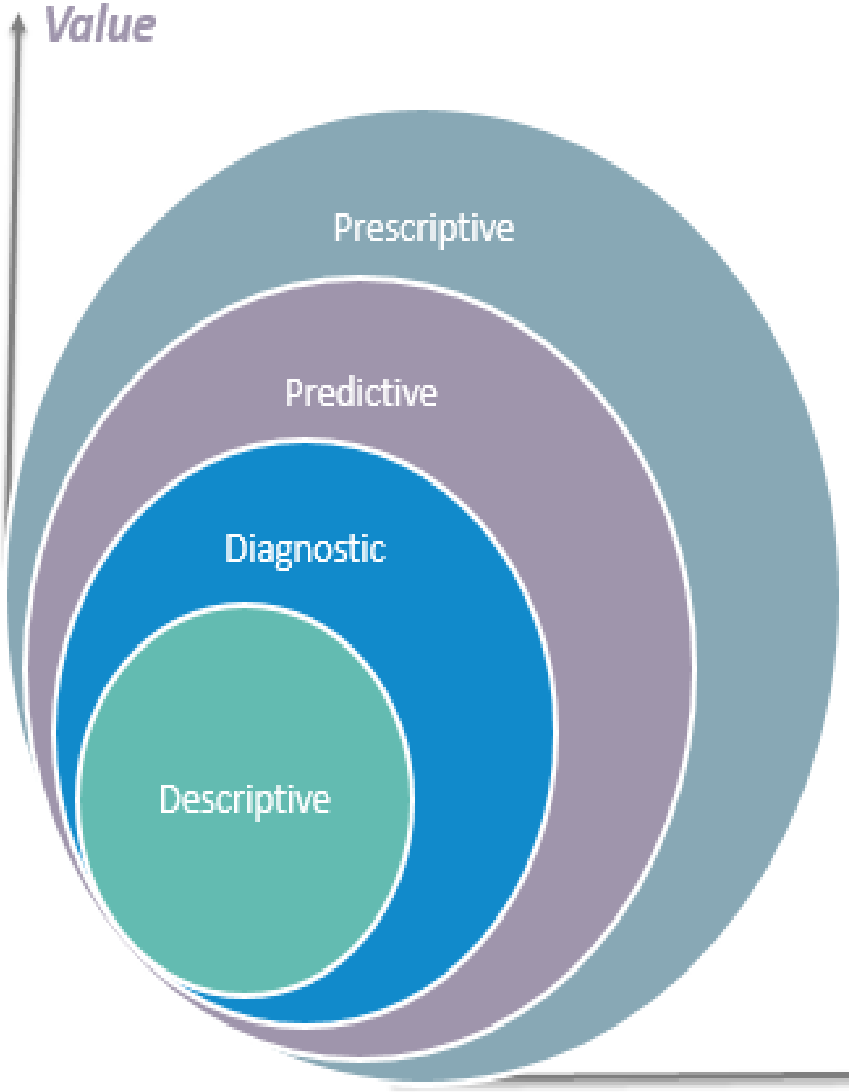
- EDA (or **Descriptive analytics**) tells us what happened up from the data.
- CDA can be separated into two parts:

→**Predictive analytics** give us clues about the future, given data and domain knowledge.

→**Prescriptive analytics** provide suggestions for optimizing the future.

1. **Descriptive**. Traditional HR metrics are largely efficiency metrics (turnover rate, time to fill, cost of hire, number hired and trained, etc.). The primary focus here is on cost reduction and process improvement. Descriptive HR analytics reveal and describe *relationships* and *current and historical data patterns*. This is the foundation of your analytics effort. It includes, for example, dashboards and scorecards; workforce segmentation; data mining for basic patterns; and periodic reports.
2. **Predictive**. Predictive analysis covers a variety of techniques (statistics, modeling, data mining) that use current and historical facts to make predictions about the future. It's about probabilities and potential impact. It involves, for example, models used for increasing the probability of selecting the right people to hire, train, and promote.
3. **Prescriptive**. Prescriptive analytics goes beyond predictions and outlines decision options and workforce optimization. It is used to analyze complex data to predict outcomes, provide decision options, and show alternative business impacts. It involves, for example, models used for understanding how alternative learning investments impact the bottom line (rare in HR).

4 types of Data Analytics



What is the data telling you?

Descriptive: *What's happening in my business?*

- Comprehensive, accurate and live data
- Effective visualisation

Diagnostic: *Why is it happening?*

- Ability to drill down to the root-cause
- Ability to isolate all confounding information

Predictive: *What's likely to happen?*

- Business strategies have remained fairly consistent over time
- Historical patterns being used to predict specific outcomes using algorithms
- Decisions are automated using algorithms and technology

Prescriptive: *What do I need to do?*

- Recommended actions and strategies based on champion / challenger testing strategy outcomes
- Applying advanced analytical techniques to make specific recommendations

Principa
www.principa.co.za

http://www.kdnuggets.com/2017/07/4-types-data-analytics.html

探索性資料分析(EDA)

- EDA首要目的在於資料偵錯、獲得資料的大略資訊、驗證已知結果。

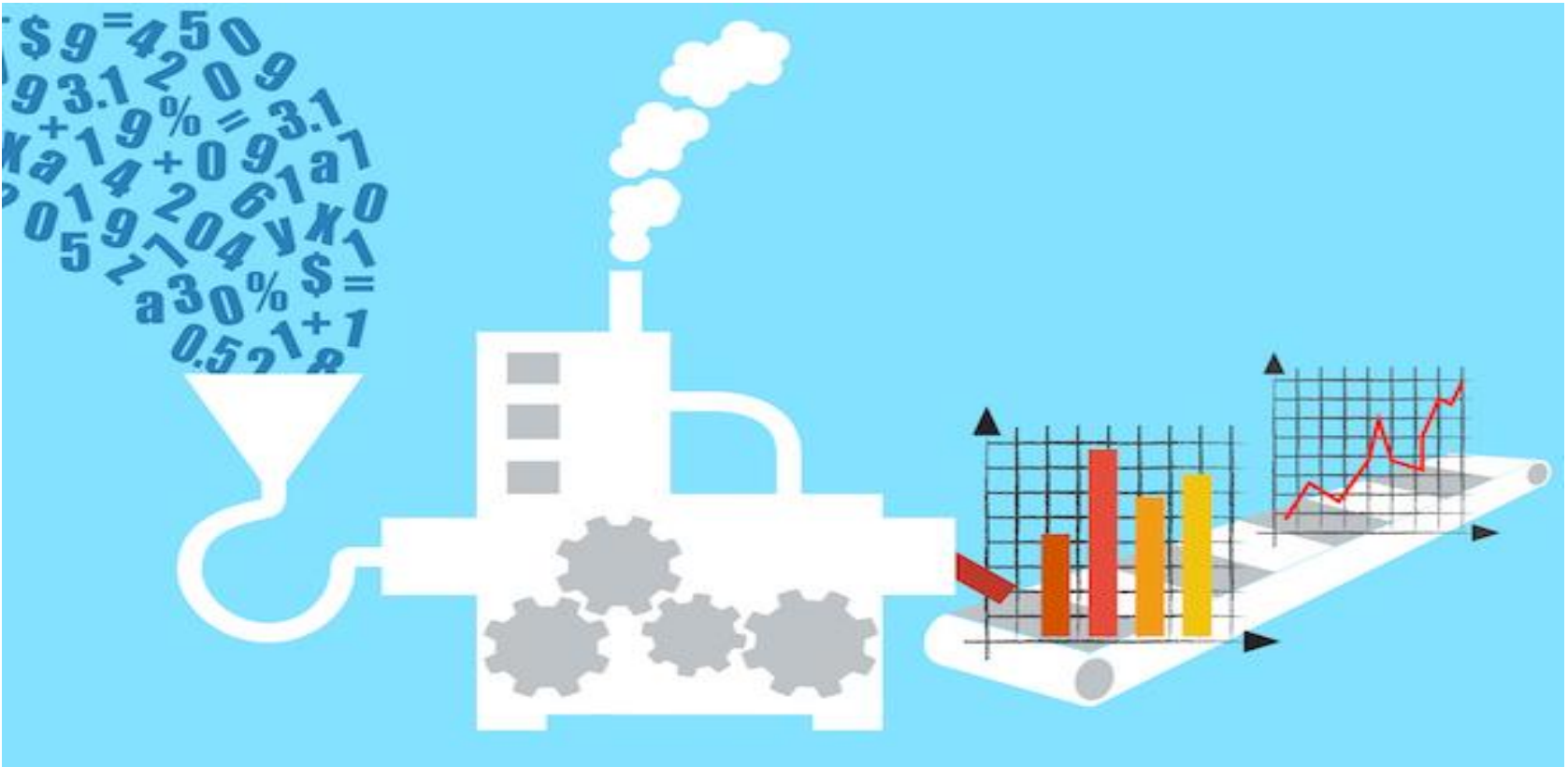
→圖形、表格在EDA中扮演重要的角色；並由分析結果中尋找合適的下一步分析方法。

→使用任何的統計方法前，先確定該方法需要的假設條件是否滿足。

https://github.com/souravbose1991/Auto-EDA

探索性資料分析(資料驅動)

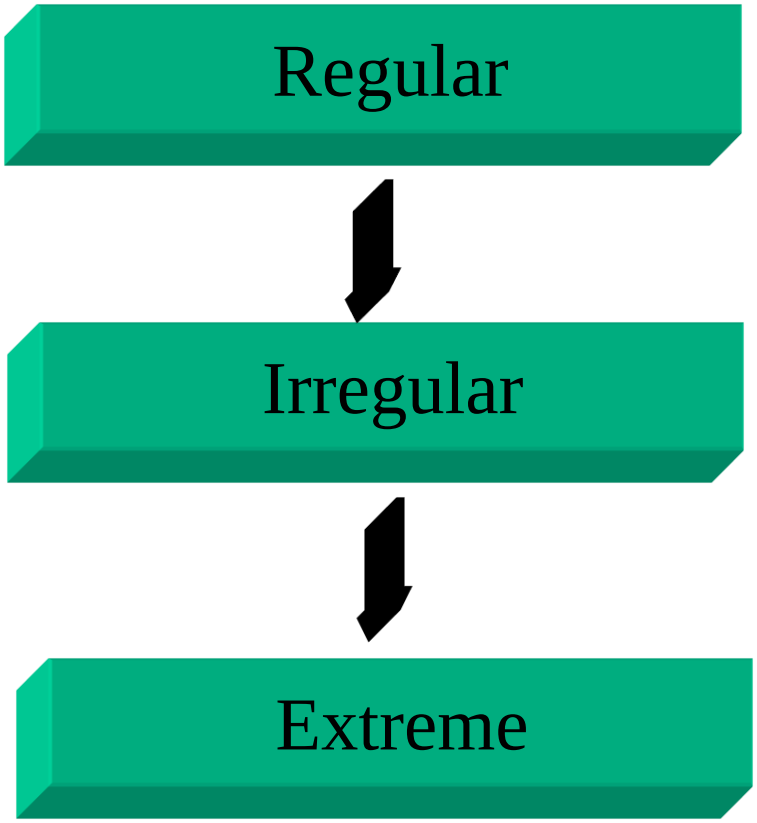
Exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics … EDA is for seeing what the data can tell us beyond the formal modeling. ---Wikipedia



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

統計與知識

- 統計整理資訊的方法屬於歸納法(Induction)，從龐雜的資料找出共同趨勢，並區分資料具有以下哪一種特性：



一般、常見

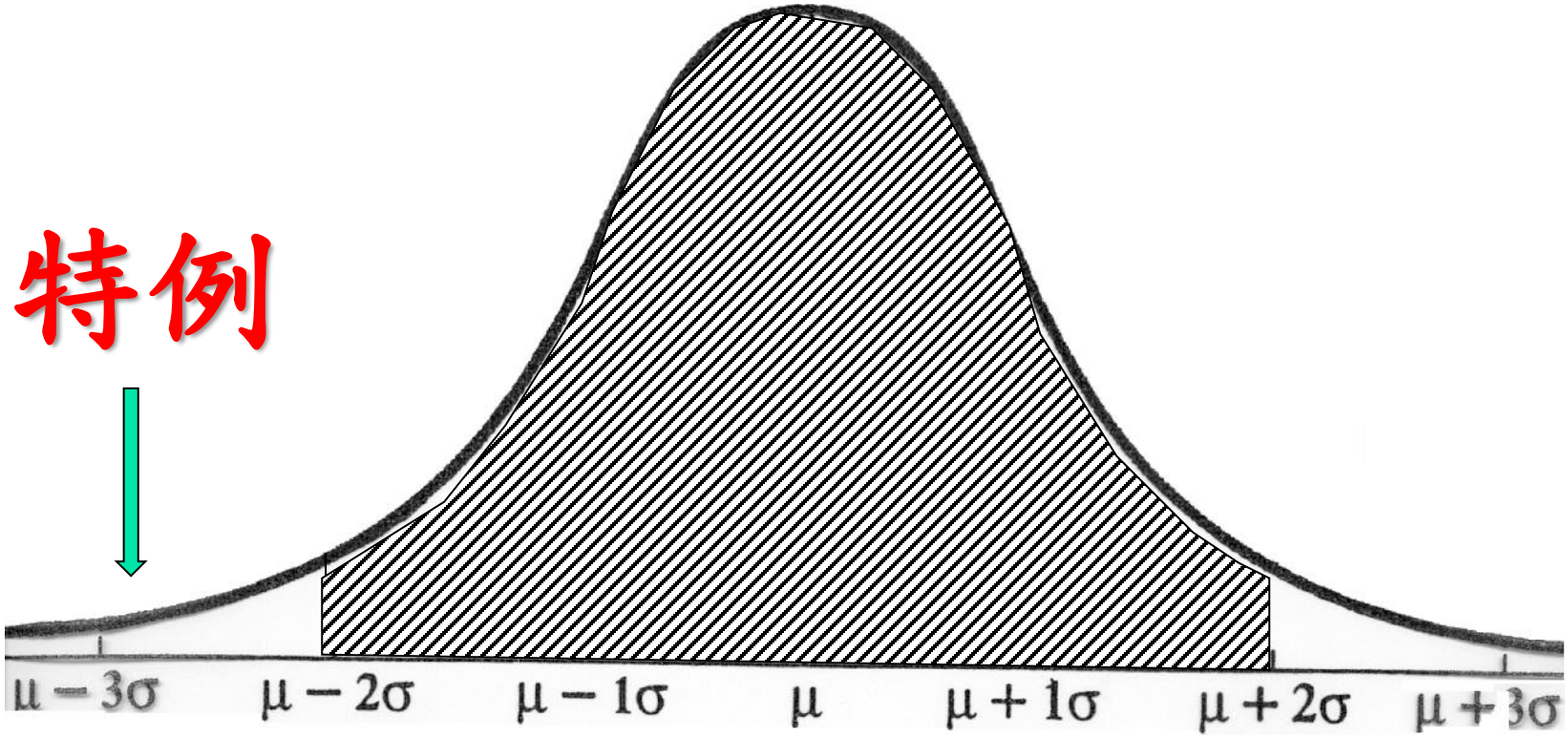
不尋常、不規則

離群值(Outlier)

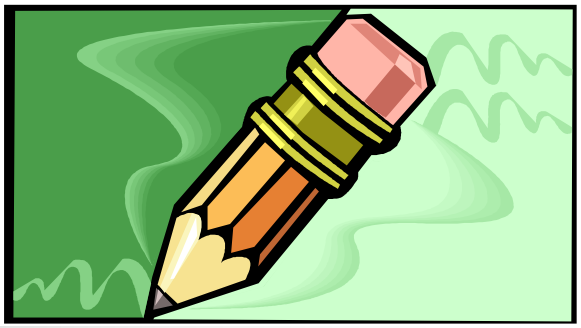
罕見、極端

原則!!

特例



敘述性統計量



- 探索性或基本資料分析(Explanatory Data Analysis或Initial Data Analysis)是資料分析中最基本、也是非常重要的一個步驟，資料分析的成敗往往在這個步驟中決定。

→敘述性統計量包括資料的基本特性，例如：平均數、標準差、所佔比例(圖表)等 。

→ 一般的分類方式為：

集中趨勢量數、差異量數

基本資料分析

- 資料偵錯

  → 資料輸入錯誤、尋找可能的離群值。

- 初步探索資料的特性

  → 資料的集中、散佈趨勢。

- 驗證已知的結果

  → 是否與已知的結果相同？

統計大師對資料分析的建議

- John Tukey: “An approximate answer to the right question is worth a great deal more than a precise answer to the wrong question”

→Q: Check EDA (Exploratory Data Analysis)

- George Box: “All models are wrong; but some are useful.”

→Q: Check Box-Cox transformation

資料類型

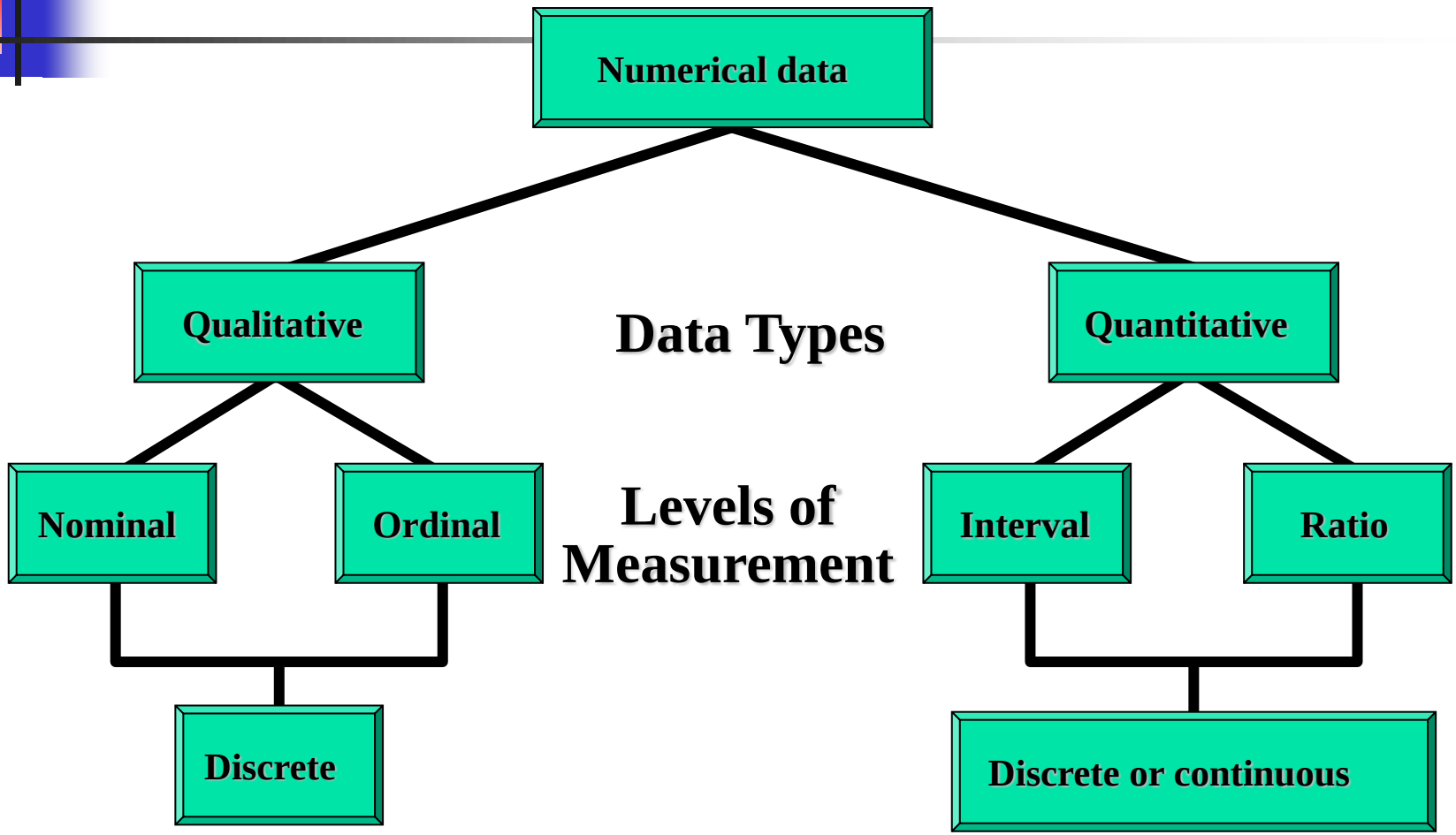
資料通常可分成四種類型：

- Nominal（名目）
- Interval（區間）
- Ordinal（順序）
- Ratio（比值）

蘊含的資訊與資料類型有關！

資料分析方式也與其類型有關！

Types of Data (資料類型)



- 資料類型將直接影響分析方法的選取，並非所有資料都適合常見的統計方法，任意使用分析方法可能會得出令人啼笑皆非的結果。

→A > B, B > C 是否代表A > C？（遞移律！）

	甲城市	乙城市	丙城市
A候選人	1	2	3
B候選人	2	3	1
C候選人	3	1	2

註：1代表最喜歡，3代表最不喜歡。

問卷資料範例：

- 請問您本次購買的機車是

  什麼廠牌_____ 汽缸大小_____c.c.

- 請問您打算幾年後換購新機車？

  □1. 1年以下 □2. 1～2年 □3. 3～4年

  □4. 5年以上 □5. 其他_____（請說明）

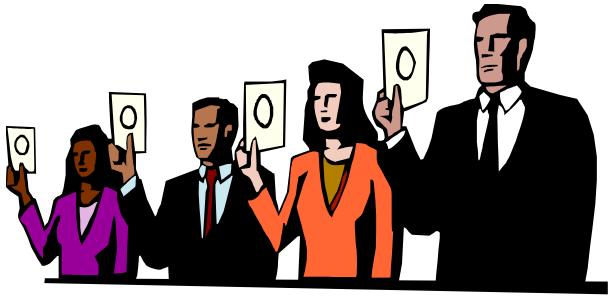
- *請問您對本郵局的滿意程度為何？*

  □1.不滿意 □2.普通 □3. 滿意

- *請問您對本郵局的滿意程度為何？*

  □1.非常不滿意 □2.不滿意 □3.普通

  □4.滿意 □5. 非常滿意



- 數字定義可能引起的問題：

→ 評審給候選人A、B、C、D的評分。

評審	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*總分*
A	4	1	2	4	1	2	4	18
B	3	4	1	3	4	1	3	19
C	2	3	4	2	3	4	2	20
D	1	2	3	1	2	3	1	13

- 候選人D分數明顯最低，刪除後評審重新對候選人A、B、C評分。

  →A的分數反而最高。

評審	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*總分*
A	3	1	2	3	1	2	3	15
B	2	3	1	2	3	1	2	14
C	1	2	3	1	2	3	1	13

柯侯5：侯柯2

11/15 美麗島73波

侯柯配+1

11/14 ETtoday

侯柯配+1

11/14 聯合報

誤差2.9%，柯領先3%

柯侯配+1

11/14 美麗島72波

重複，採納最新版11/15

ETtoday是採網路問卷，主動填寫回覆民調，容易有灌票可能，也非經驗證的科學民調

11/13 匯流

柯侯配+1

11/11 美麗島71波

重複，採納最新版11/15

11/10 美麗島70波

重複，採納最新版11/15

11/9 美麗島69波

重複，採納最新版11/15

11/9 ETtoday

重複，採納最新版11/14

11/8 美麗島68波

重複，採納最新版11/15

11/8 趨勢

柯侯配+1

11/8 世新大學

柯侯配+1

11/8 求真

柯侯配+1

11/7 美麗島67波

重複，採納最新版11/15

https://tw.news.yahoo.com/%E8%97%8D%E7%99%BD%E5%90%88%E7%B5%90%E5%B1%80%E6%90%B6%E5%85%88%E6%9B%9D-%E9%BB%83%E5%85%89%E8%8A%B9%E6%95%B4%E7%90%86%E6%97%A5%E8%87%B3%E4%BB%8A%E6%B0%91%E8%AA%BF-%E6%9F%AF%E4%BE%AF%E9%85%8D5%E5%88%86vs-%E4%BE%AF%E6%9F%AF%E9%85%8D2%E5%88%86-093008725.html

統計分析：誤差＋不確定性

- 四組資料為$N(\mu,1)$的100筆亂數，期望值分別為0、0.1、0.2、0.3。

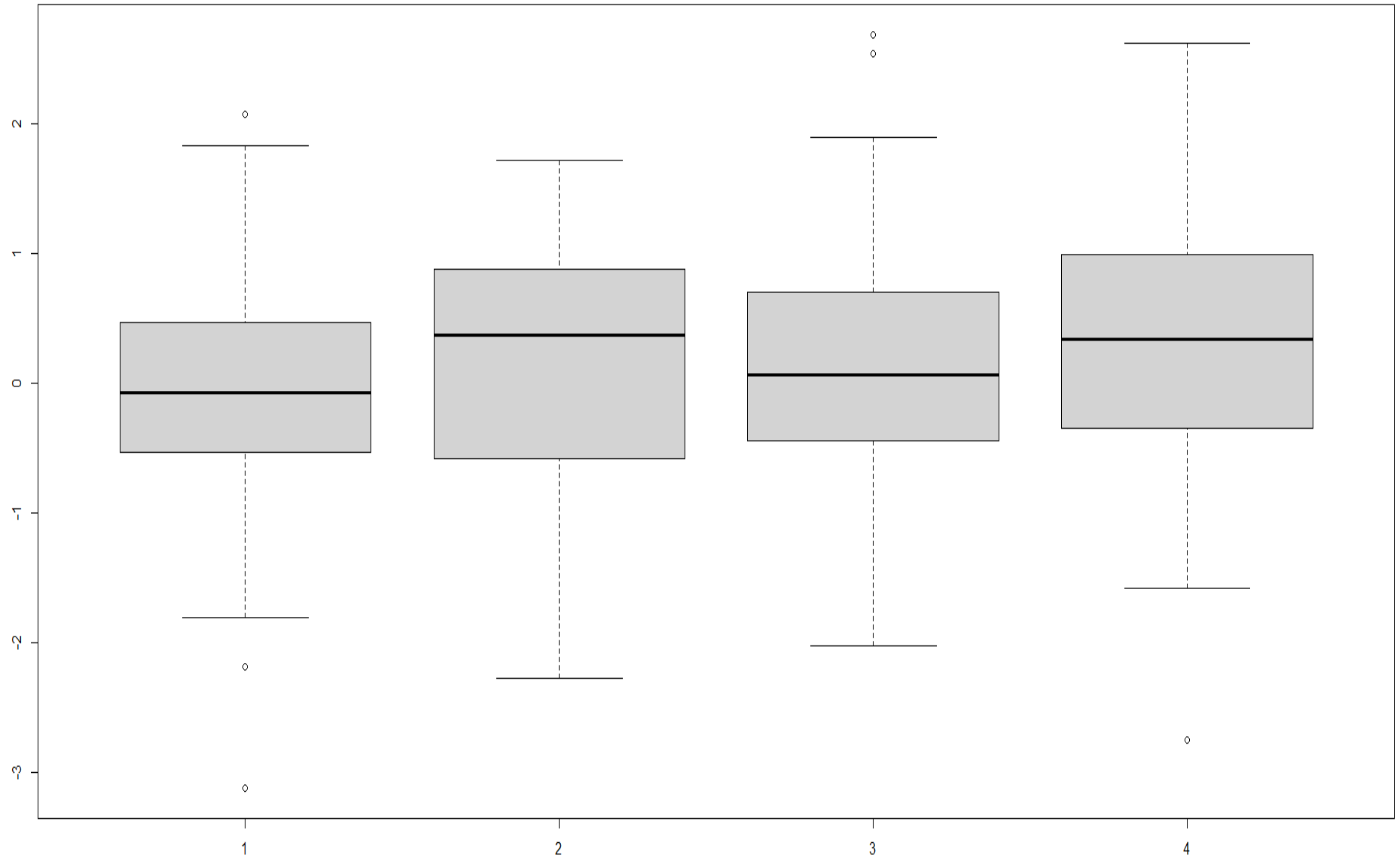
→兩兩 t 檢定的p-value，僅有第一組及第四組有差異（p-value < 0.01）。

- 遞移律未必成立 ($\mu_1 = \mu_2$ & $\mu_2 = \mu_3$ → $\mu_1 = \mu_3$)

→統計數值不是必然結果，其中隱含不確定性，無法套用一般數學計算的規則！

- 延伸問題：如何由統計表達「$\mu_1 > \mu_2$」？

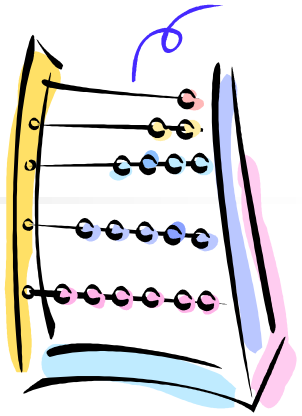
統計思維的範例



集中趨勢量數

- Mean (平均數)
- Median (中位數)
- Mode (眾數)
- Percentiles (百分位數)
- Quartiles (四分位數)

集中趨勢量數



- 平均數(Mean;期望值)

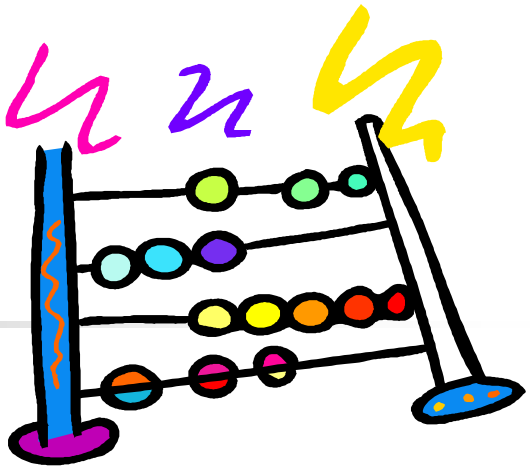
→算術平均數(Average)；

→加權平均數(Weighted Average)；

→其他(幾何平均數、調和平均數)。

- 中位數(Median): 一半的數值比中位數大，一半的數值比中位數小。
- 眾數(Mode):出現次數最多的數值

差異量數



- 全距(Range):

→最大與最小數值之差(Range = Max－Min)

- 四分位差(Quartile Deviation):

→四分位數(Quartile;Q_1): 3/4的數值比大Q_1，1/4的數值比Q_1小。

→四分位差 = $Q_3 - Q_1$

- 變異數(Variance; σ^2)與標準差(Standard Deviation;σ)

敘述統計量(範例)

例題一、試以文字詮釋以下隨機抽出某公司業務部門20位員工的年齡：

41	25	25	33	27	31	42
35	36	32	36	41	34	29
34	31	34	35	32	35	

→ 平均數＝33.4，中位數＝34.0，標準差＝4.75，全距＝17。

敘述統計量(續)

例題二、試以文字詮釋以下隨機抽出某公司20位員工去年請假的天數：

0 0 0 0 0 0 0 0 1 1

1 2 2 3 4 5 5 6 7 42

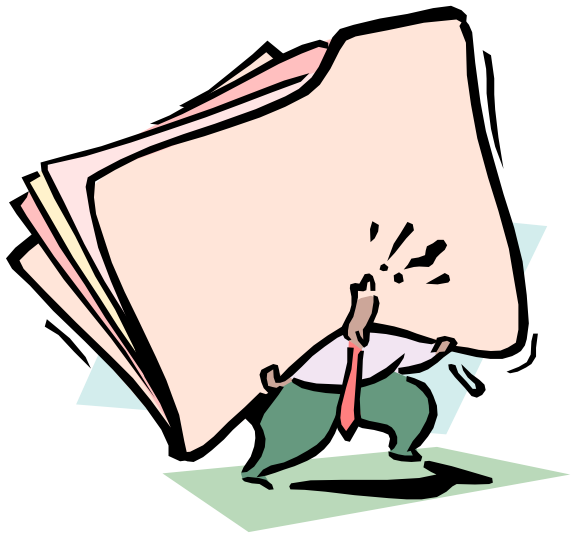
→ 你/妳 看到了甚麼現象？

敘述統計量(續)

例題三、街頭隨機訪問20位成年受訪者去年閱讀某月刊的期數：

0 1 11 0 0 0 2 12 0 0

12 1 0 0 0 0 12 0 11 0



→ 請問這是甚麼樣的月刊？

敘述統計量(續)

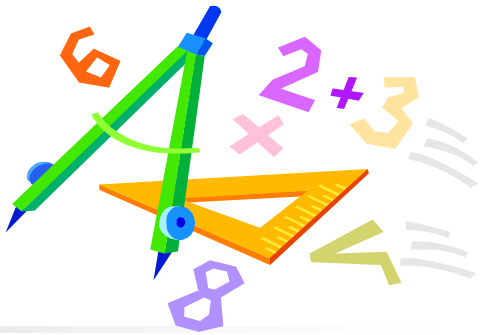
例題四、以下為隨機抽出某地區16位孕婦的身高(單位：公尺)：

1.57 1.55 1.60 1.52 1.68 1.57 1.62 1.55

1.65 1.52 2.55 1.60 1.55 1.60 1.62 1.57

→ 請問你/妳看到資料有何特性？

基本資料分析



- 基本資料分析的首要目的在於資料偵錯、獲得資料的大略資訊、驗證已知結果。(例如：正常 vs. 異常！)

→因此，圖形、表格在基本資料分析中扮演重要的角色；並由基本資料分析的結果中尋找合適的下一步分析方法。

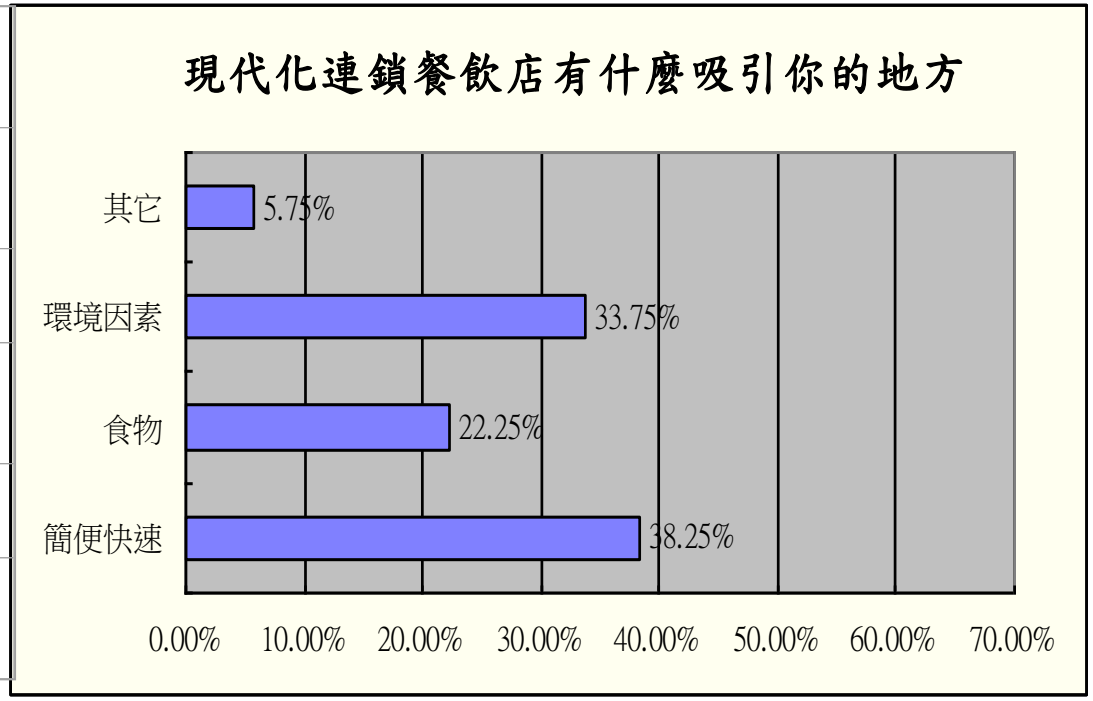
→ 使用任何的統計方法前，先確定該方法需要的假設條件是否滿足。

資料分析範例

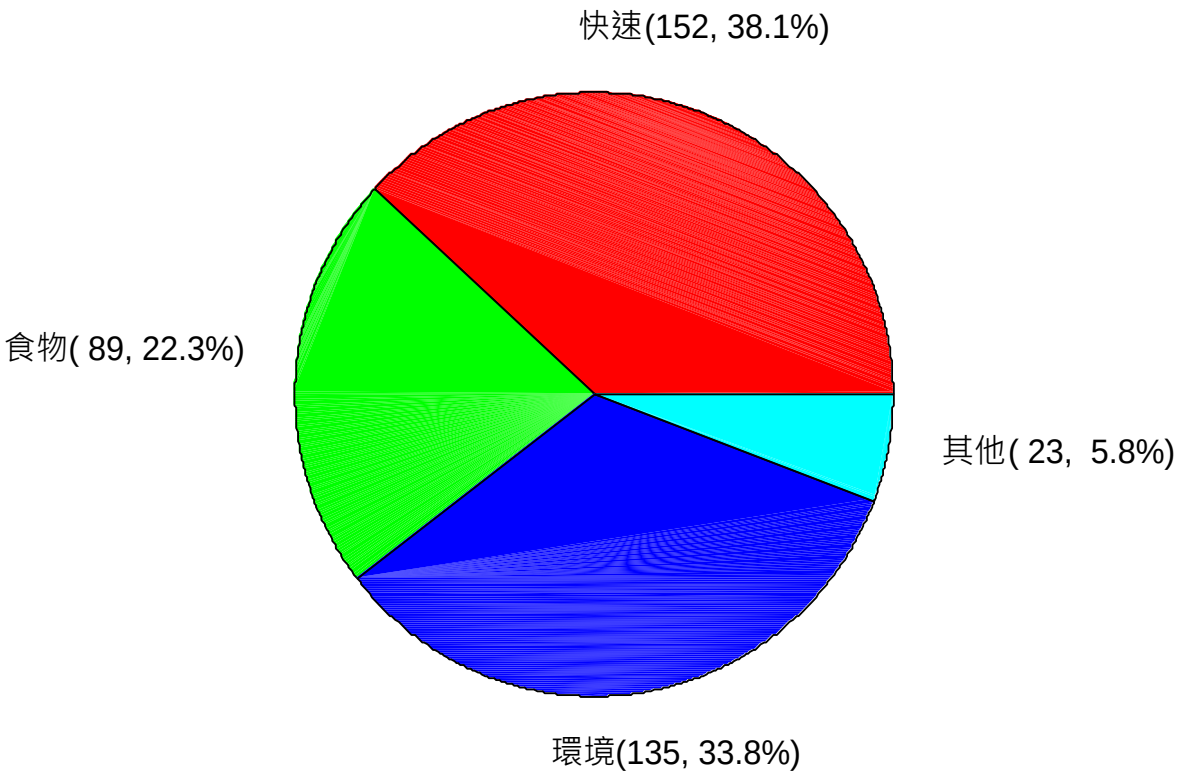
- 以表格、圖形展示資料更有效果！

長條圖(Bar Chart)

	選項	人數	百分比
1	簡便快速	153	38.25
2	食物	89	22.25
3	環境因素	135	33.75
4	其它	23	5.75
	N＝	400	

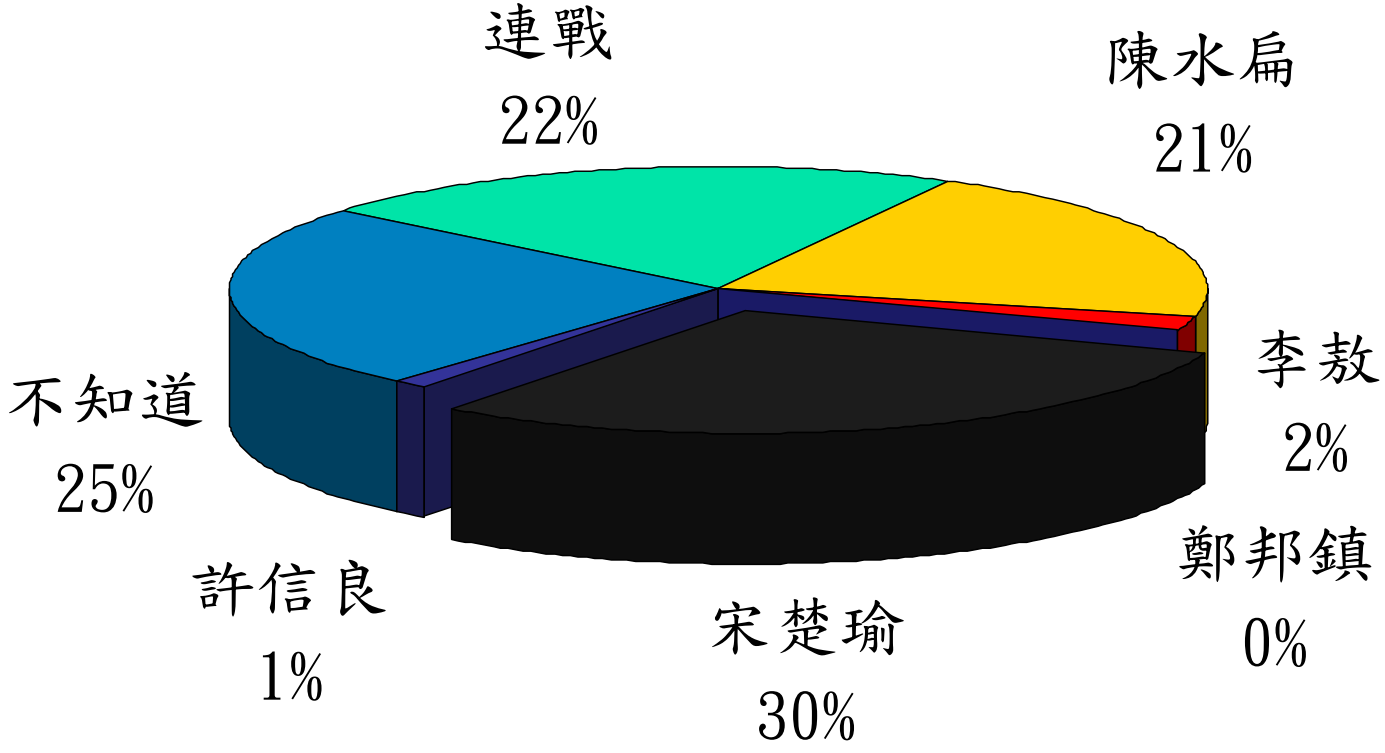


- 圓餅圖(Pie Chart)也是另一種圖形表示法。



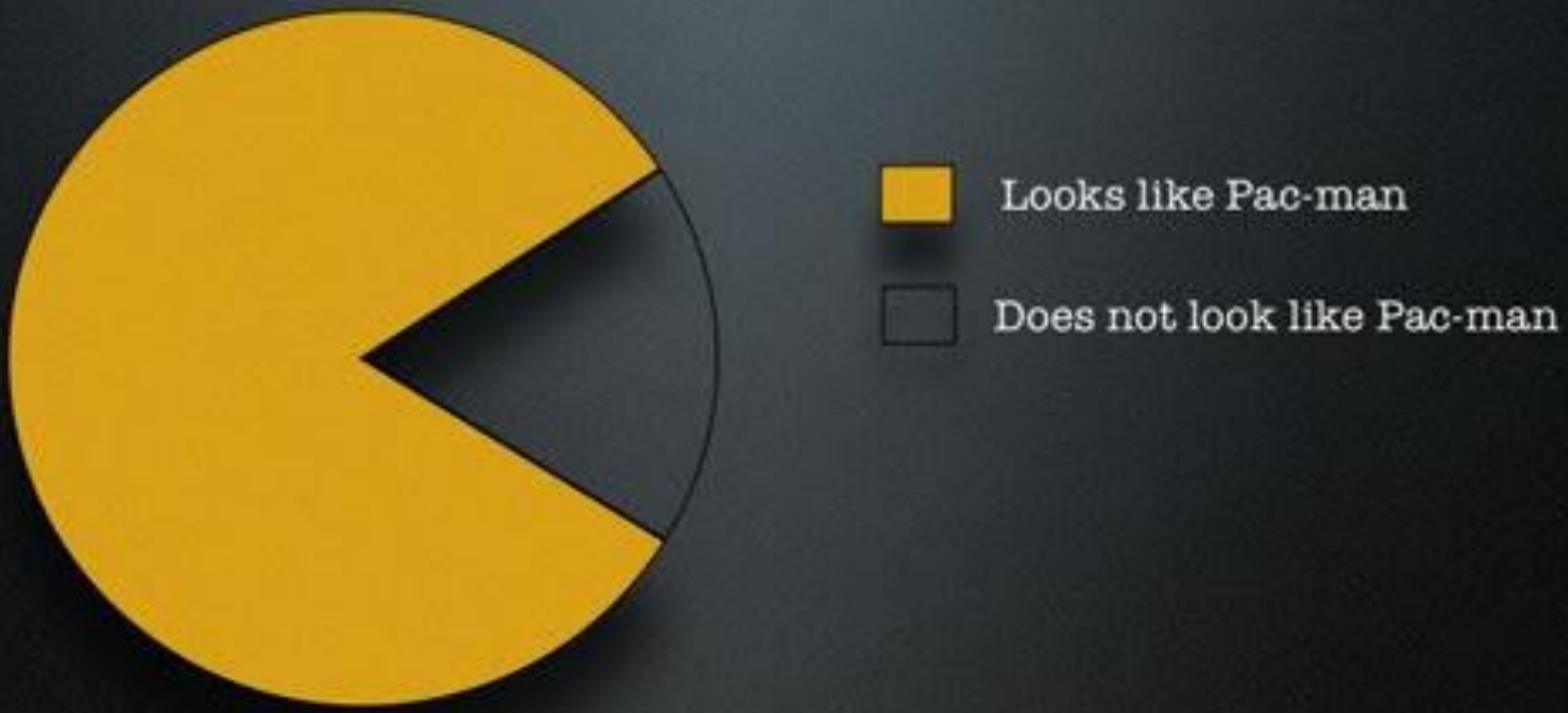
圓餅圖的範例(立體)

總統候選人的支持比例(88年11月)




Pie I have eaten
Pie I have not yet eaten

Percentage of chart which looks like Pac-man

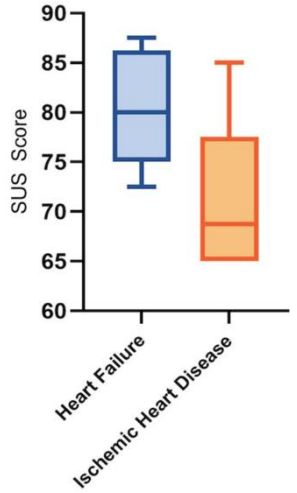


Box Plot 範例(Apartment Rents)

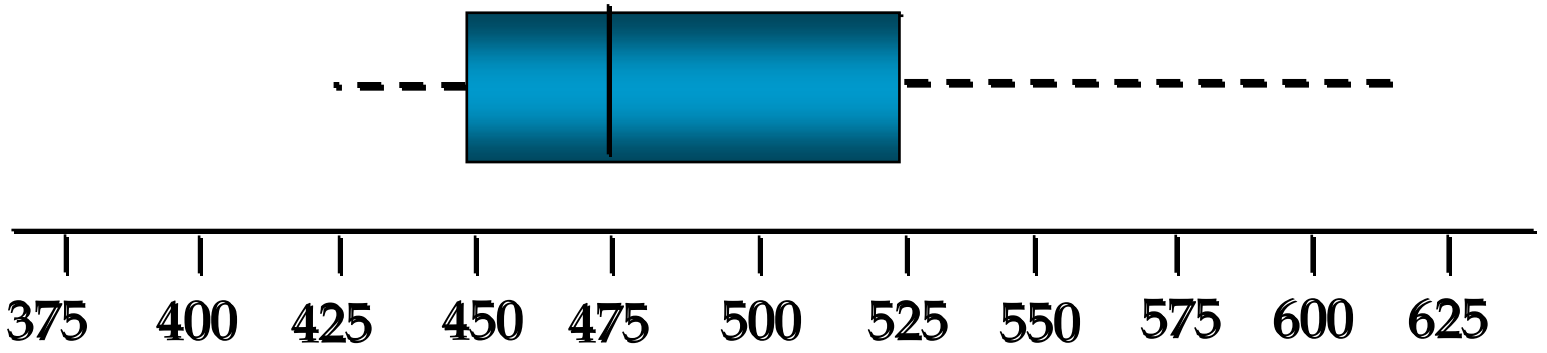
Lower Limit: Q1 - 1.5(IQR) = 450 - 1.5(75) = 337.5

Upper Limit: Q3 + 1.5(IQR) = 525 + 1.5(75) = 637.5

→There are no outliers.



https://www.mdpi.com/2075-4418/12/4/926



Example: Singer Heights Story

Each singer in the NY Choral Society in 1979 self-reported his or her height to the nearest inch. Their voice parts in order from highest pitch to lowest pitch are Soprano, Alto, Tenor, Bass. The first two are typically sung by female voices and the last two by male voices.

→What is the best way(s) to describe the heights of these singers?

Note: You may use find the data from web site

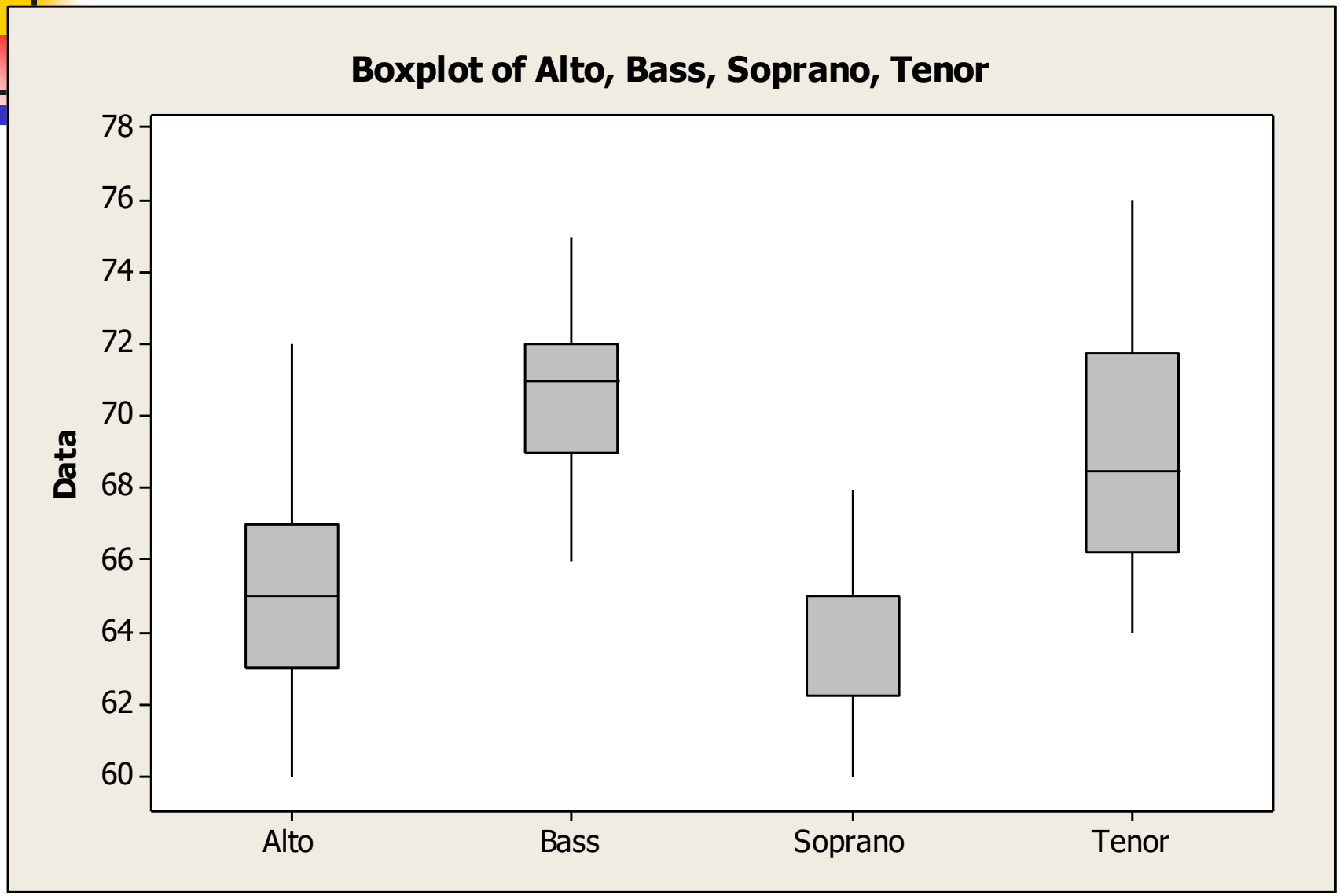
http://lib.stat.cmu.edu/DASL/

Some descriptive statistics

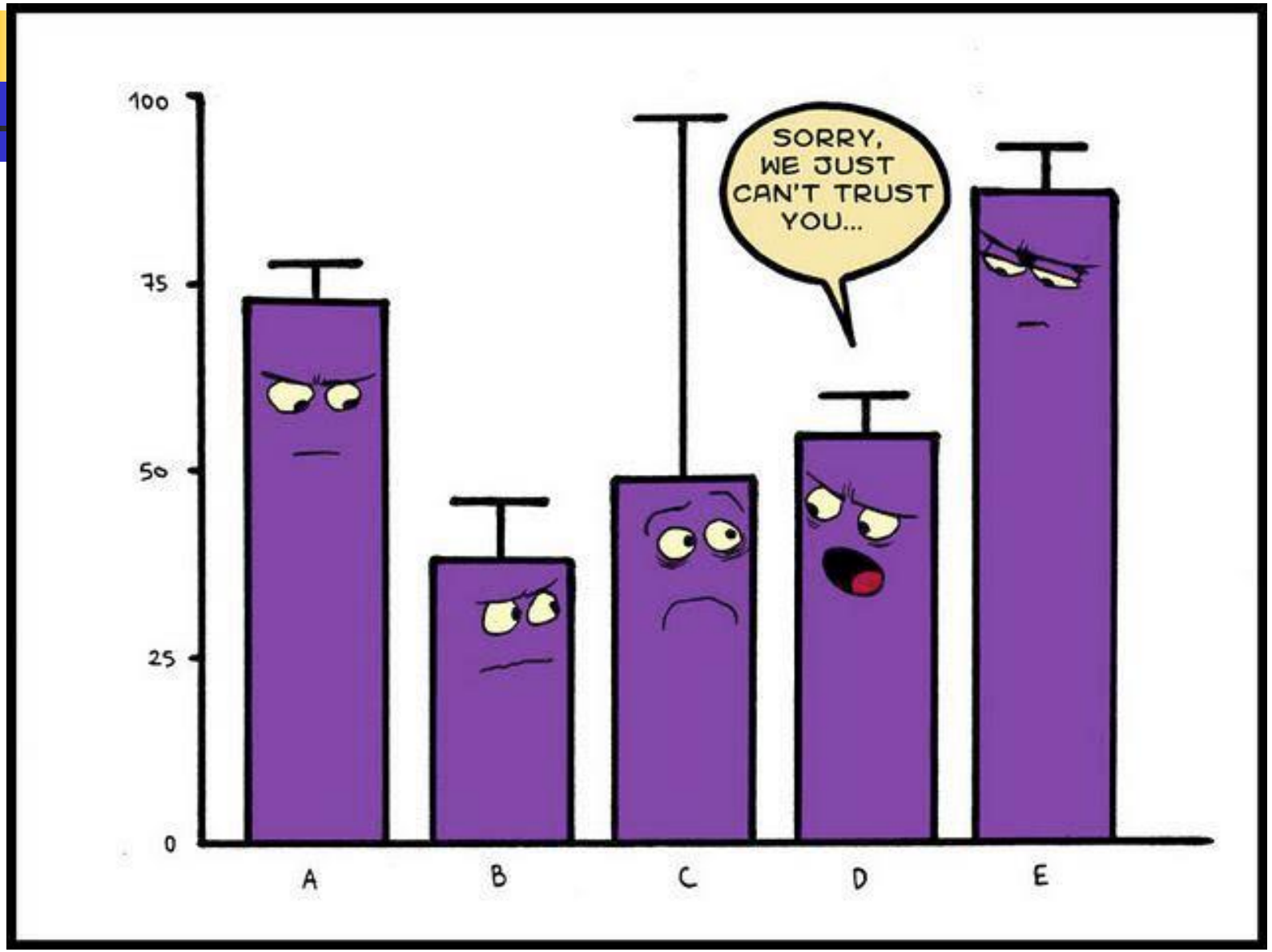
Variable	N	N*	Mean	SE Mean	StDev
Soprano	36	0	64.250	0.312	1.873
Alto	35	0	64.886	0.472	2.795
Tenor	20	0	69.150	0.719	3.216
Bass	39	0	70.718	0.378	2.361

Variable	Minimum	Q1	Median	Q3	Maximum
Soprano	60.000	62.250	65.000	65.000	68.000
Alto	60.000	63.000	65.000	67.000	72.000
Tenor	64.000	66.250	68.500	71.750	76.000
Bass	66.000	69.000	71.000	72.000	75.000

Compare the differences!

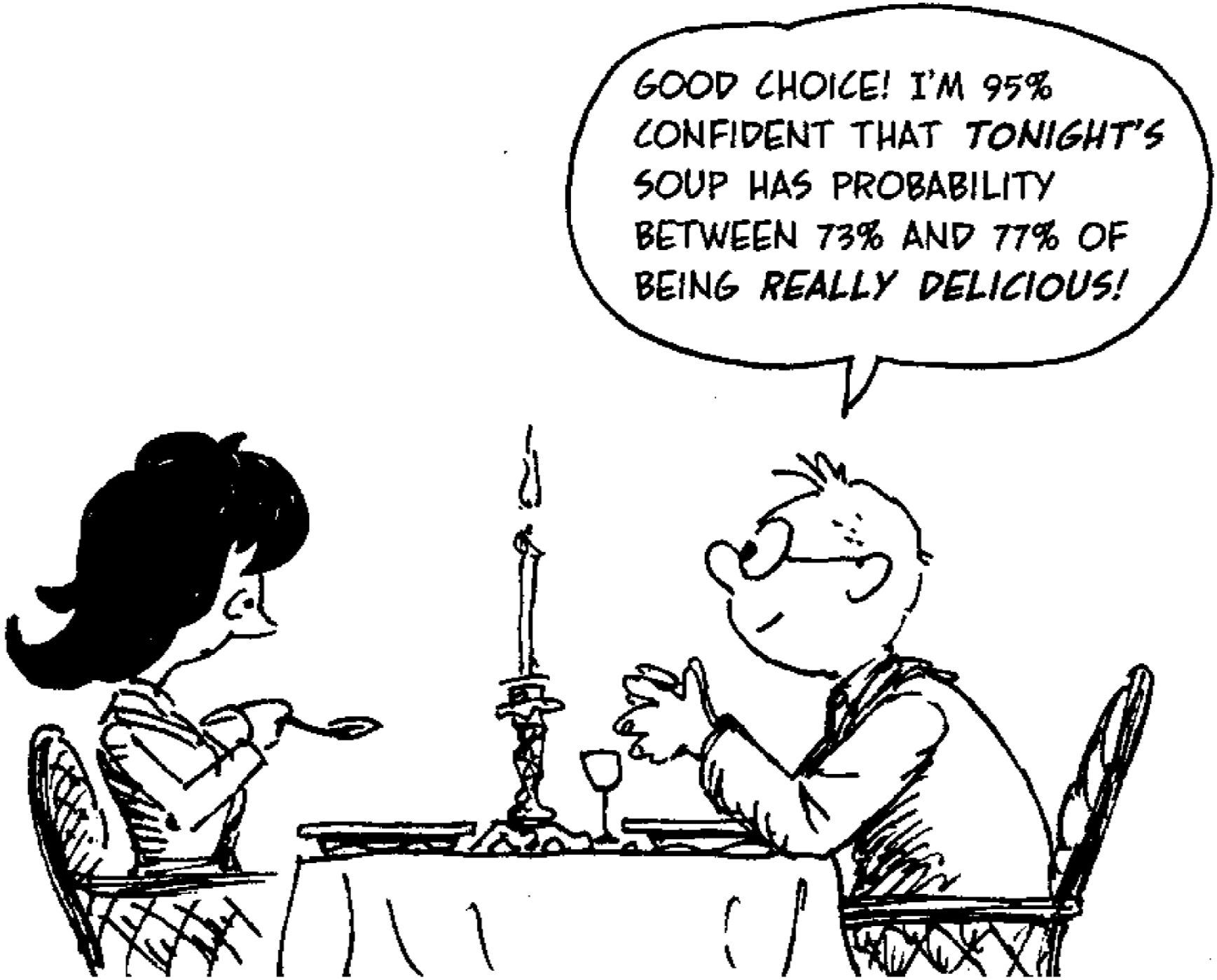


非我族類其心必異！

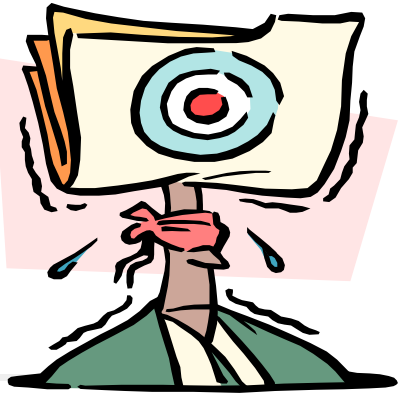


結果詮釋與推論限制

- 分析結果若能合理闡述，可達到「畫龍點睛」之效，但最忌諱忽略關鍵點，純粹就數字面來詮釋，反而變成「畫虎不成反類犬」。
- 研究結果的推論也需注意，不能僅從統計結果來看，也需瞭解推論結果代表的意義。


GOOD CHOICE! I'M 95% CONFIDENT THAT TONIGHT'S SOUP HAS PROBABILITY BETWEEN 73% AND 77% OF BEING REALLY DELICIOUS!

驟下結論(範例)



- 多數車禍發生在車速40～60公里/時，僅有少數在車速超過100公里。

→開快車比較安全？

- 美國亞歷桑那州死於肺結核的比例最高。

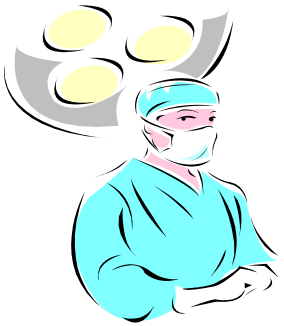
→亞歷桑那州的天氣易於感染肺結核？

- 調查小學生的拼字能力，發現腳愈大的拼字能力也較強。

→腳的大小影響拼字能力？

驟下結論(續)

- 2001年大陸調查發現長壽者中，排行老大者最多。



→排行老大較長壽？

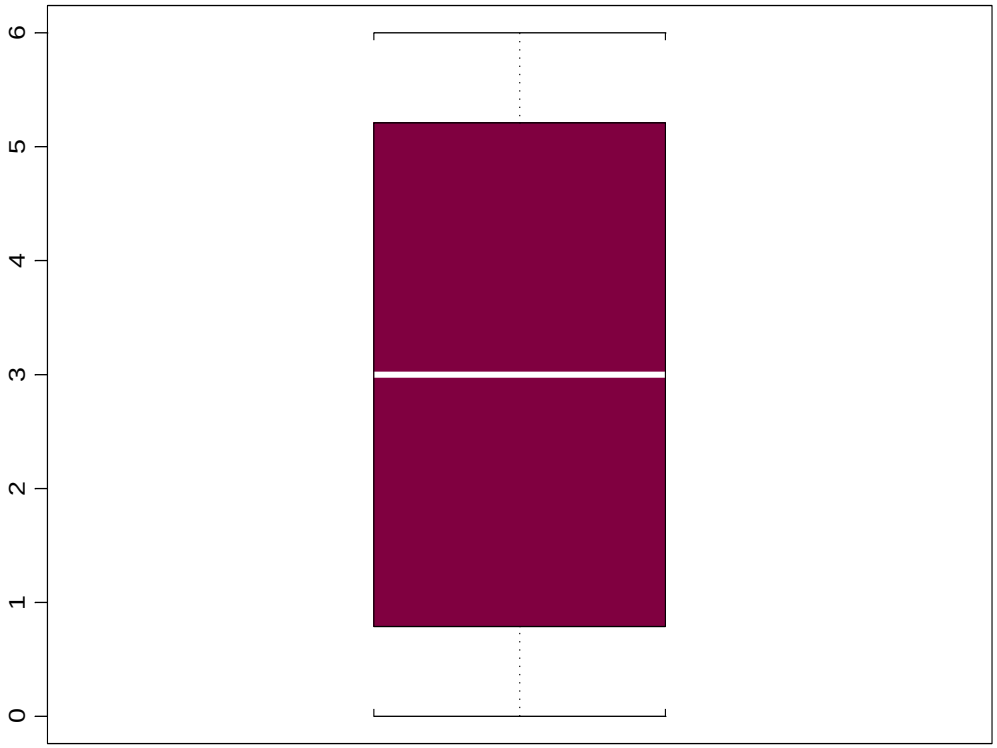
→抑或是排行老大者佔了多數？

- 英國公務統計顯示在家裡生產者，發生意外的比例較在醫院生產者高，因此孕婦都應該在醫院生產。

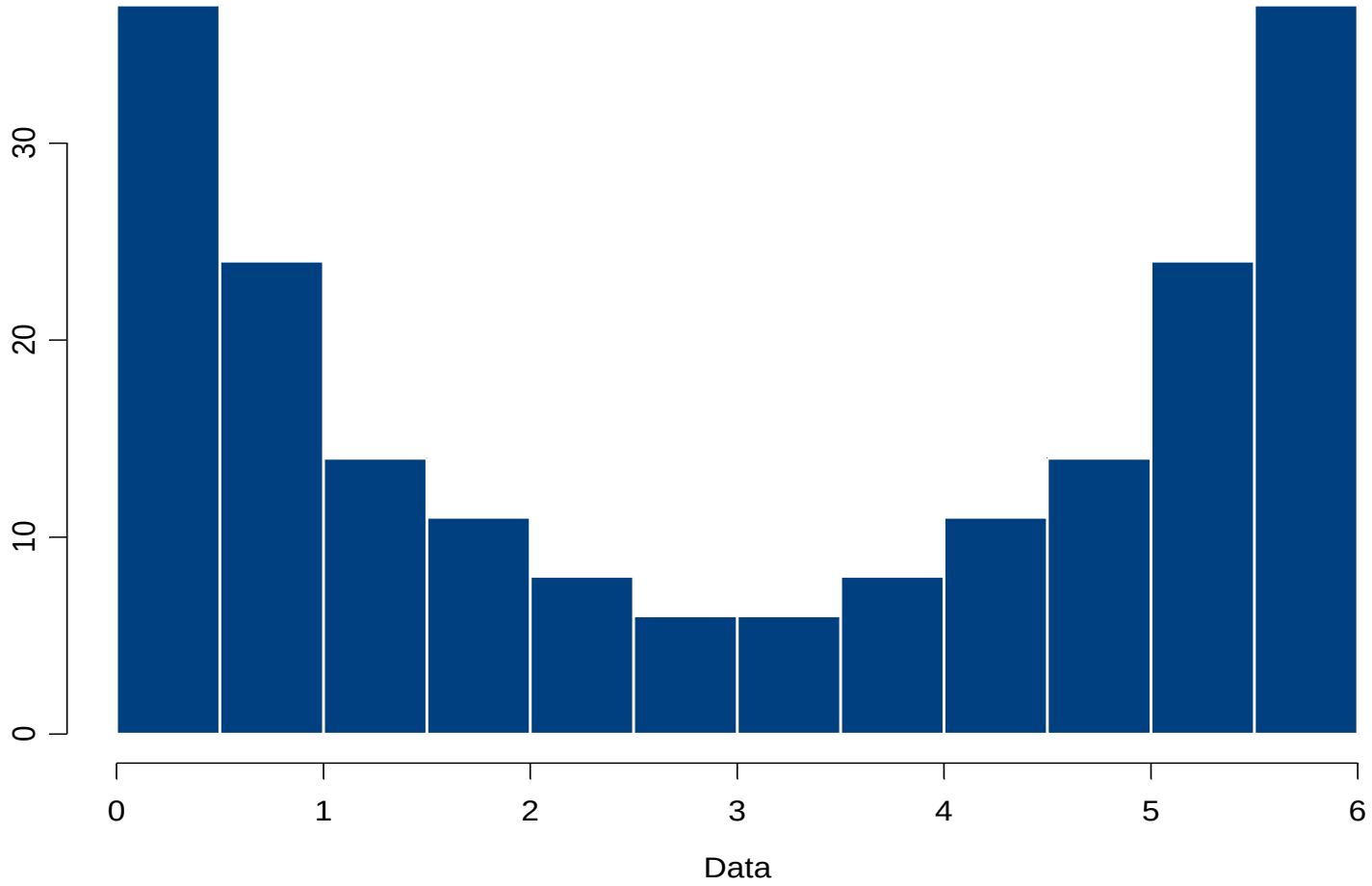
→為什麼有些孕婦會在醫院以外的地方生產？

表格與圖形(範例)

例題五、大略敘述下圖資料的特性。



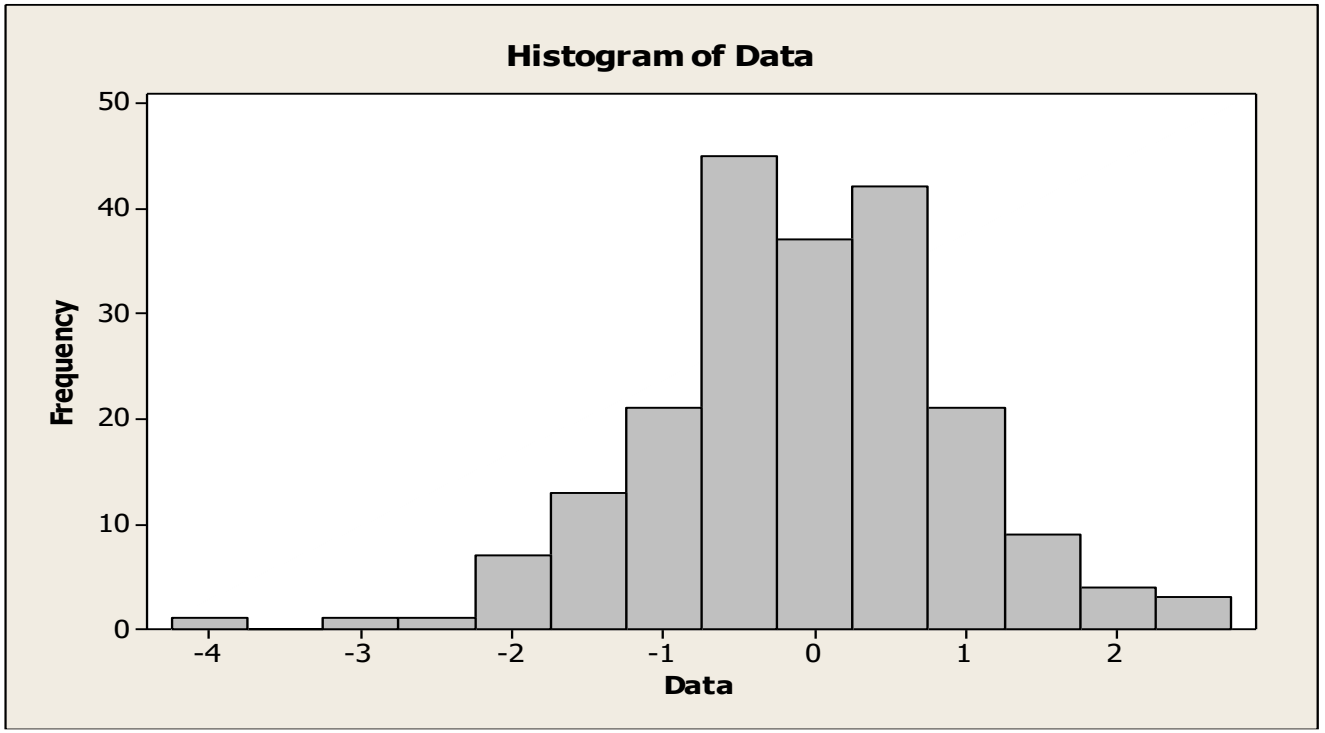
原始資料(你/妳猜對了嗎？)

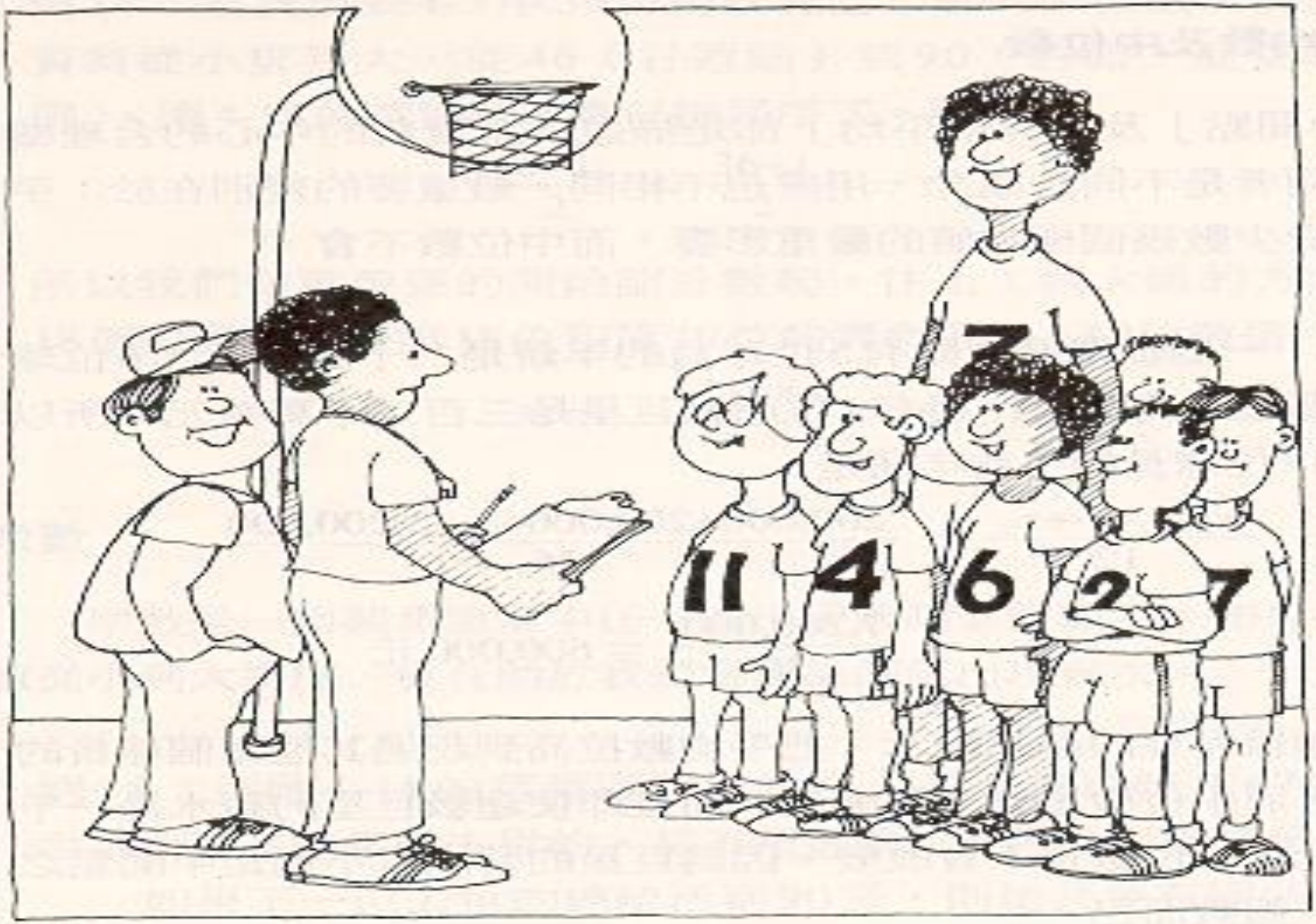


問題：常態分配的箱型圖具有哪些特性？

平均數與中位數

- 左偏(Left-skewed)：當少數觀察明顯小於一般觀察值 (如下圖)，平均數將被這些觀察值拉下，但中位數較不受影響，此時中位數大於平均數。
- 少數觀察值較大時稱為右偏(Right-skewed)。





「我們是應該宣布我們的平均高度來嚇死對手，還是宣布我們的中位數高度來消除他們的戒心呢？」

老謝曝「很多優秀球員是姜豐年去求來的」：無所求的無私奉獻最感動人！

如何藉由統計獲取資訊？

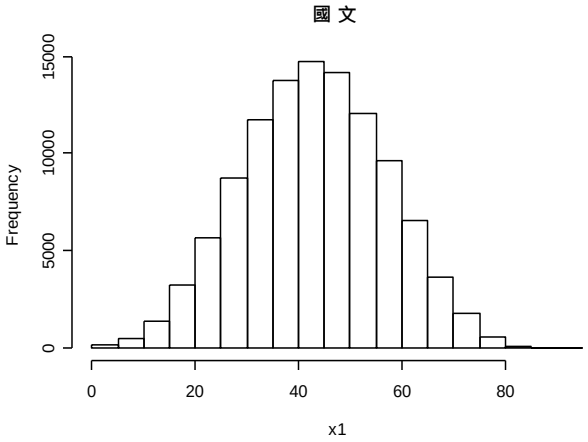
- 如果想瞭解民國94年指定考試各科的特性，可以藉助哪些工具？

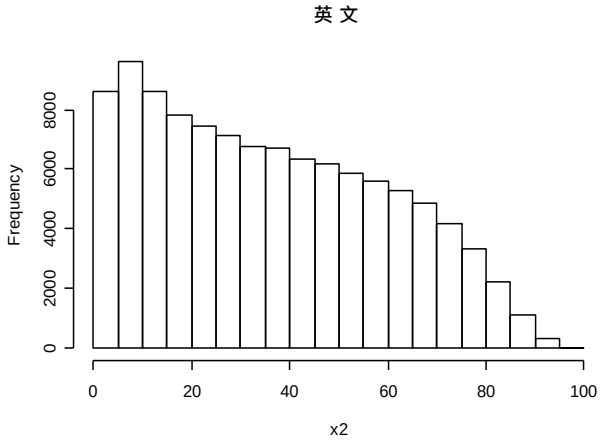
→例如：那一科的分數最不平均(哪一科大多數人都考得不好，只有少數人分數分高)。

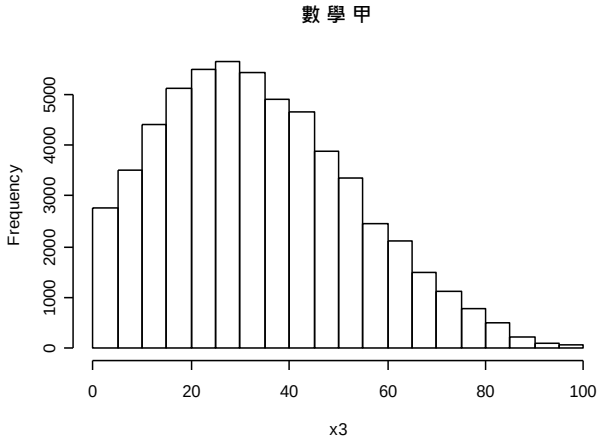
→平均數明顯大於中位數，稱為右偏；若平均數明顯小於中位數，稱為右偏。平均數等於中位數，則為兩側對稱。

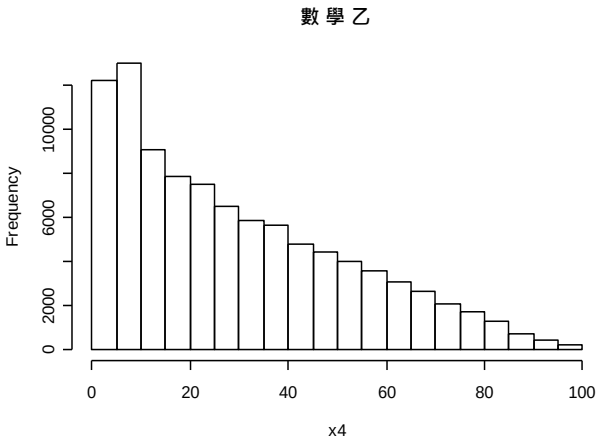
民國 94 年大學指定考試各科成績

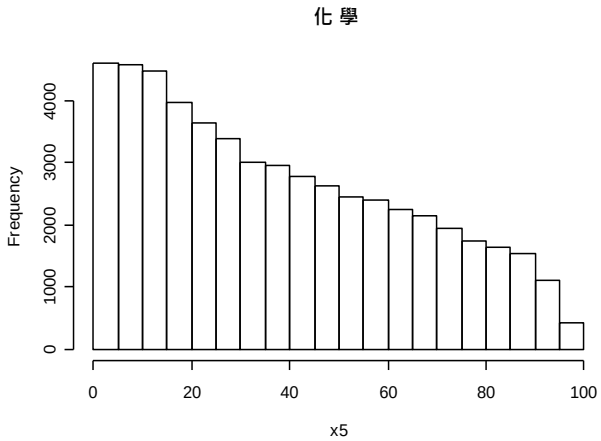
	國文	英文	數學甲	數學乙	化學	物理	生物	歷史	地理
Min.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0	0.00
12%	27.00	8.00	11.00	4.00	8.00	6.00	22.00	13.0	18.00
1st Qu.	34.00	16.00	22.00	12.00	15.00	12.00	32.00	28.0	30.00
Median	44.00	34.00	34.00	29.00	34.00	23.00	45.00	39.0	39.00
Mean	43.56	36.68	36.36	34.36	38.88	28.75	46.16	38.7	39.51
3rd Qu.	53.00	56.00	49.00	56.00	60.00	41.00	60.00	50.0	49.00
88%	60.00	69.00	59.00	61.00	76.00	57.00	71.00	56.0	55.00
Max.	93.00	98.00	100.00	100.00	100.00	100.00	99.00	89.0	90.00
st.d.	13.88	23.88	18.72	25.97	27.00	21.50	19.39	16.20	14.46

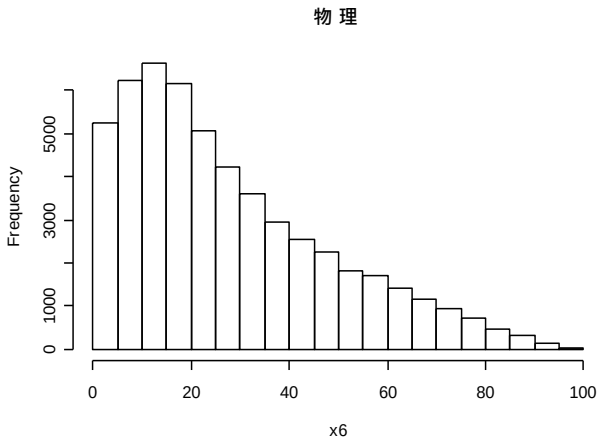

國文
Frequency
0
5000
10000
15000
0
20
40
60
80
x1

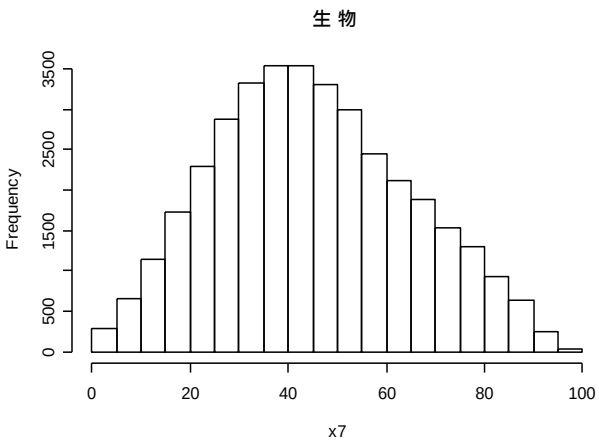

英文
Frequency
0
2000
4000
6000
8000
0
20
40
60
80
100
x2

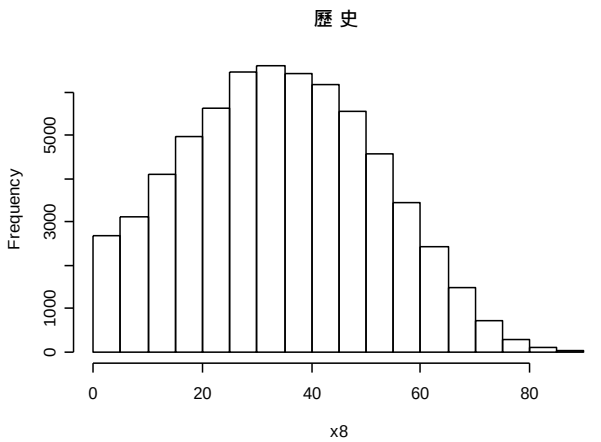

數學甲
Frequency
0
1000
2000
3000
4000
5000
0
20
40
60
80
100
x3

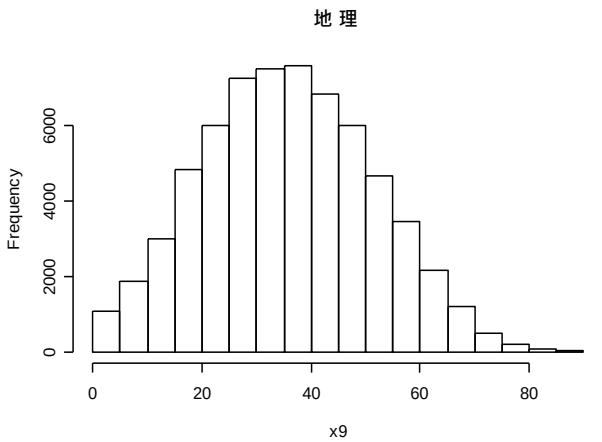

數學乙
Frequency
0
2000
6000
10000
0
20
40
60
80
100
x4

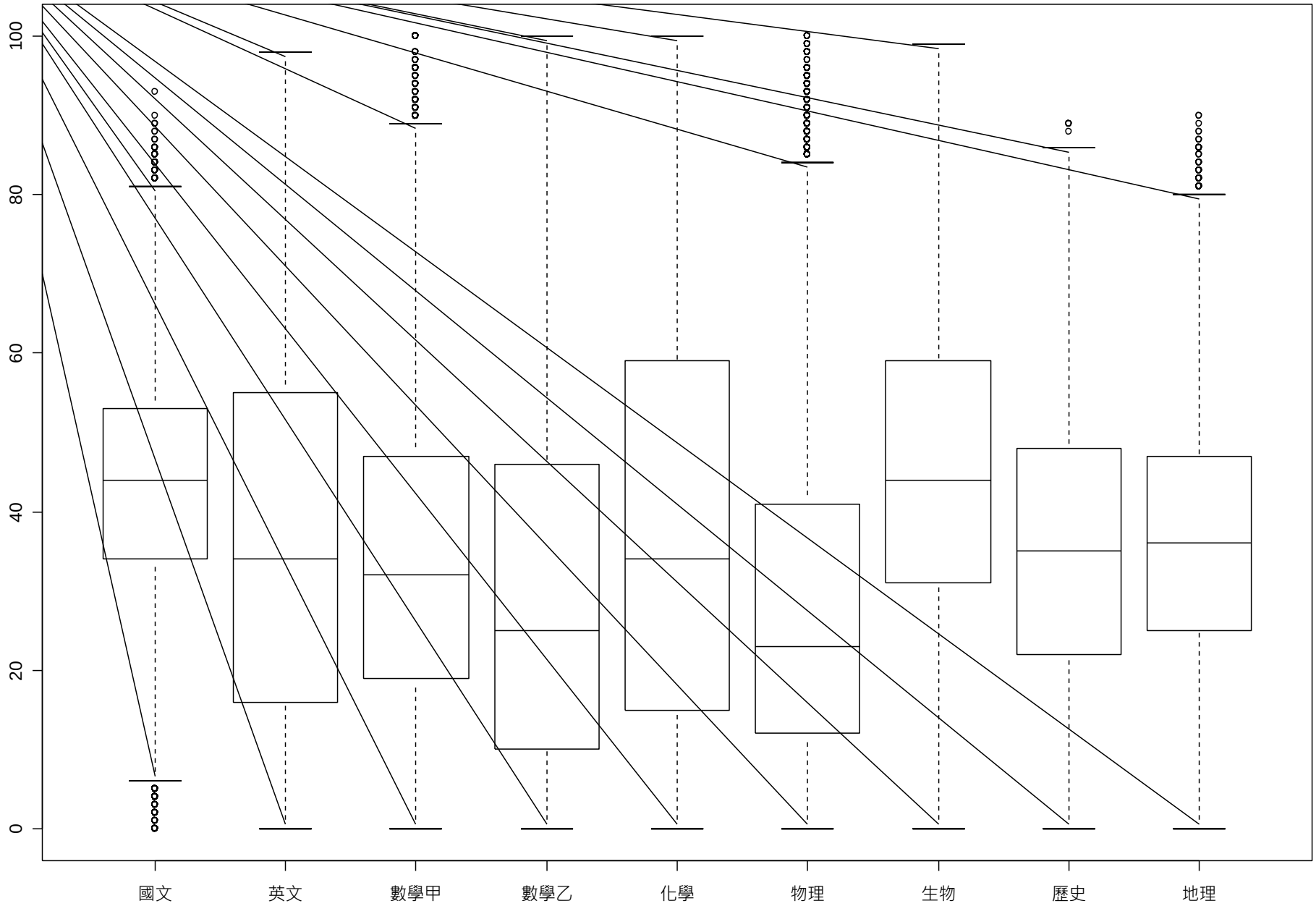

化學
Frequency
0
1000
2000
3000
4000
0
20
40
60
80
100
x5


物理
Frequency
0
1000
3000
5000
0
20
40
60
80
100
x6


生物
Frequency
0
500
1500
2500
3500
0
20
40
60
80
100
x7


歷史
Frequency
0
1000
3000
5000
0
20
40
60
80
x8


地理
Frequency
0
2000
4000
6000
0
20
40
60
80
x9


100
80
60
40
20
0
國文
英文
數學甲
數學乙
化學
物理
生物
歷史
地理

以下 100 筆資料乃自常態分配 $N(\mu,\sigma^2)$ 抽出的隨機樣本(已排序)：

12.8	42.7	51.2	62.5	73.9
14.8	42.8	51.7	62.7	74.2
21.5	43.5	51.9	62.9	74.8
22.2	44.2	52.4	63.1	75.2
23.8	44.2	53.0	63.8	75.5
30.0	45.1	53.3	63.9	75.8
30.5	45.1	53.6	63.9	76.1
32.2	45.7	56.1	65.5	76.9
32.7	47.2	57.4	65.9	77.2
33.7	47.4	57.5	67.0	78.4
34.1	48.0	57.5	67.1	79.6
34.9	48.3	58.4	67.2	80.1
35.6	49.1	59.1	67.3	80.9
36.1	49.1	59.5	68.6	82.6
36.9	49.2	59.7	69.0	83.7
37.5	49.4	60.1	70.8	83.7
40.8	49.6	60.2	71.2	84.0
41.3	49.7	60.8	72.2	84.3
42.0	50.8	61.4	73.2	84.7
42.5	50.9	61.8	73.8	85.7

→請問這些資料有什麼特性？

(a) 在不借助於樣本平均數及樣本變異數，估計期望值 μ 、變異數 σ^2 。(註：也就是純粹藉由目視來判斷！)

(b) 由你/妳從(a)估計出的期望值及變異數，驗證這 100 筆資料是否服從常態分配。(註：不能使用圖表，建議使用平均數與樣本標準差！)

(c) 以這些資料繪製 Boxplot，並以繪出的圖形驗證資料是否具有常態分配的特性。

(d) 假設這些資料在記錄時，意外地將 90 筆來自 $N(\mu_1,\sigma^2)$ 分配、10 筆來自 $N(\mu_2,\sigma^2)$ 分配混在一起，其中 $\mu_1 \neq \mu_2$ 。請說明如何判斷 μ_1,μ_2 兩者何者較大，同時大略估計 μ_1,μ_2 兩者的差異大小。

Note: 上述資料是由 90 筆來自 $N(60,15^2)$ 分配、10 筆來自 $N(30,15^2)$ 分配組成。

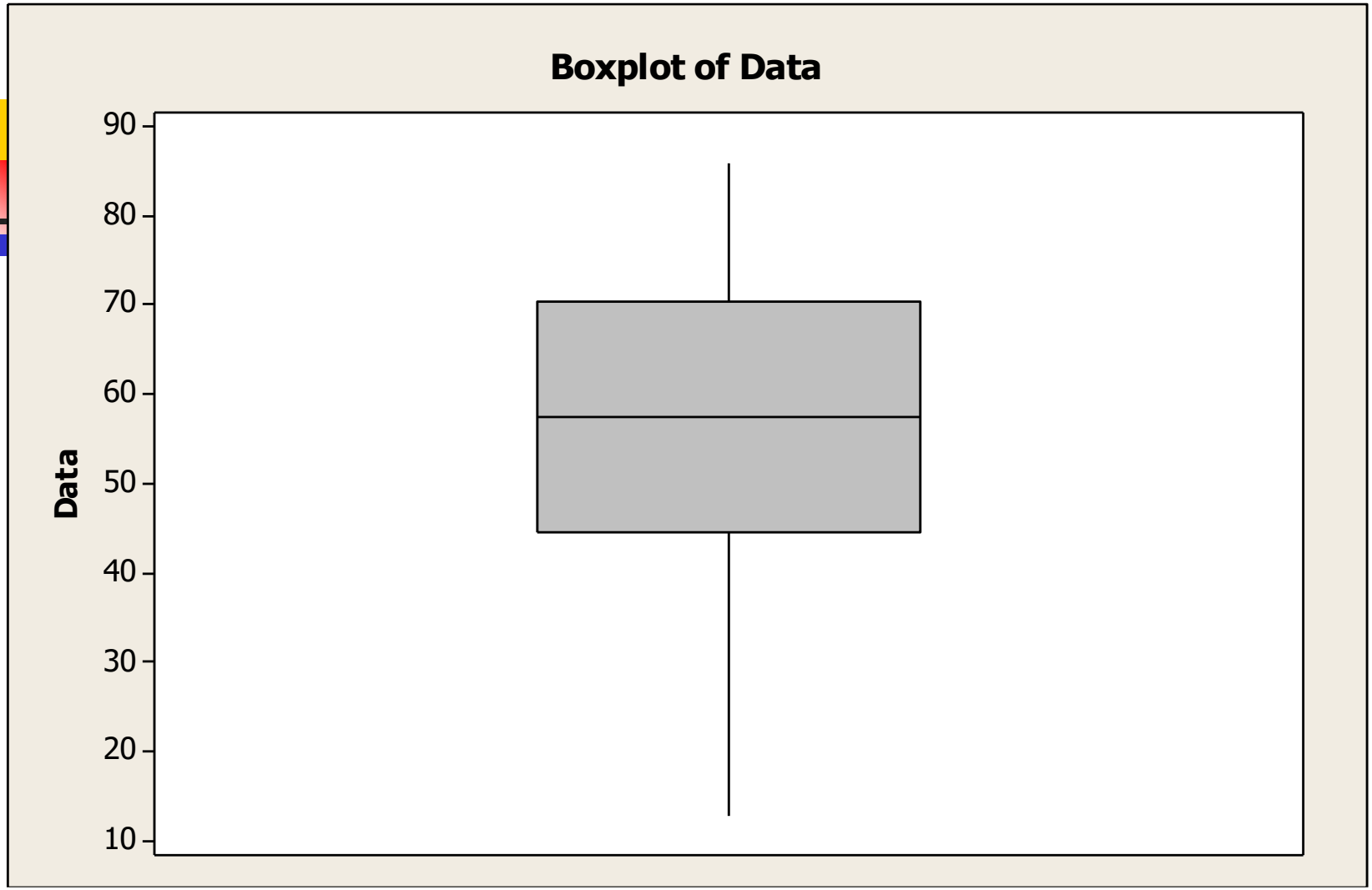
Variable	N	Mean	Median	TrMean	StDev	SE Mean
第一組	90	59.45	59.94	59.53	14.75	1.55
第二組	10	29.76	26.93	28.40	13.92	4.40
合併	100	56.49	57.56	57.01	17.13	1.71

Variable	Minimum	Maximum	Q1	Q3
第一組	30.55	85.79	48.26	72.52
第二組	12.89	57.54	19.90	38.56
合併	12.89	85.79	44.51	70.38

(a) 平均數(期望值)可由中位數代替，標準差可由 全距/4 或全距/4 近似。本題中的樣本平均數為56.49與中位數57.56很接近；樣本標準差為17.13與全距/4＝18.23，但與全距/6 ＝12.15較為接近，而與相去較遠。

(b) 樣本平均數加減一倍標準差約可涵蓋68%的觀察值，樣本平均數加減兩倍標準差約可涵蓋95%的觀察值。依此想法檢查，我們發現：各有48個、84個觀察值落入各區間，與預期有一段相當大的差距。(若代入18.23為標準差則有較佳的結果，分別有69、98個點落在區間內。)

(c) 由下圖可知資料數值較小的散佈較為分散，雖然沒有明顯的離群值、中間的Box較為集中，常態分配可能有問題。（QQ plot及常態分配檢定不拒絕！）

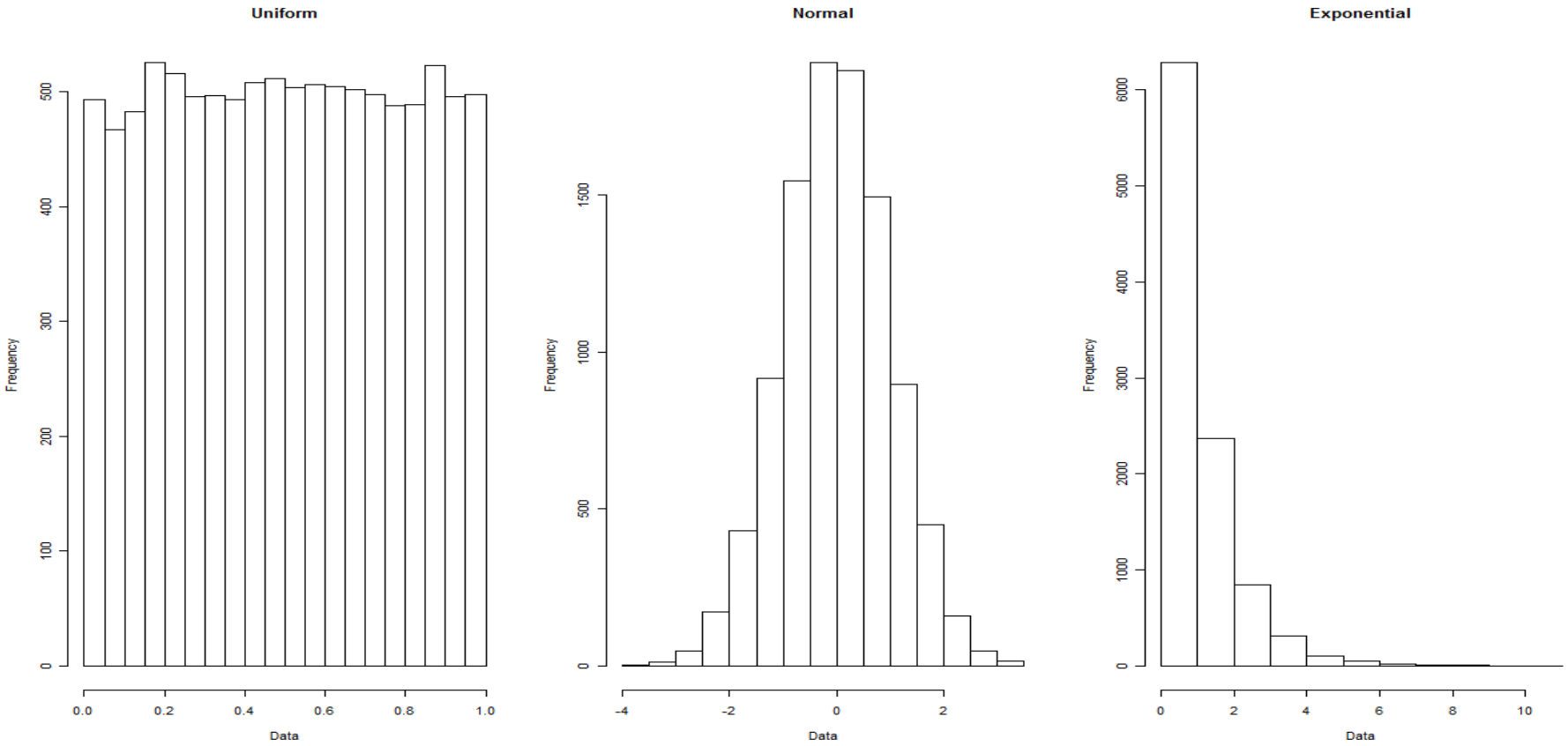


除了資料較多者多一些外，無法確定是否不為常態分配

例題：區隔不同分配的資料

- 如何區隔來自連續型均勻分配、常態分配、指數分配的資料？

→下圖為10,000個亂數繪出的Histogram。



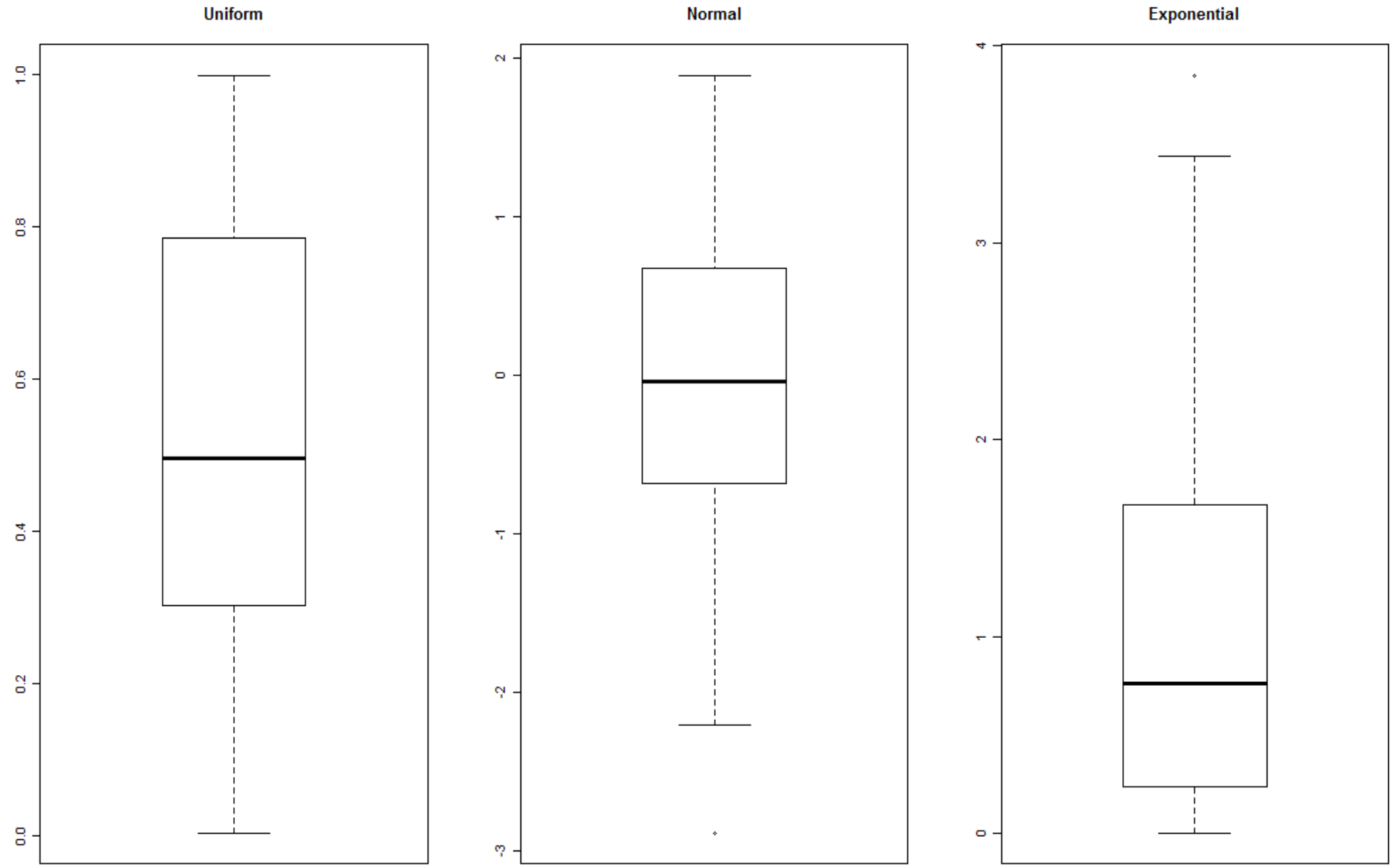
選擇有代表性的統計數據

- 如果有足夠觀察值，藉由Histogram足以區分這三個分配；但若資料量不足，可透過統計量確認觀察值的特性。

→首先可比較平均數、中位數，若兩者差異大（以標準差判斷），資料應屬指數分配。

→常態分配較均勻分配更集中，四分位數間距離較為一致(Min, Q1, Median, Q3, Max)。

註：另一種可能是藉助於Boxplot。



註：上述圖形為100個亂數的結果。

基本統計量

- 以下是三種分配100個亂數的基本統計量：

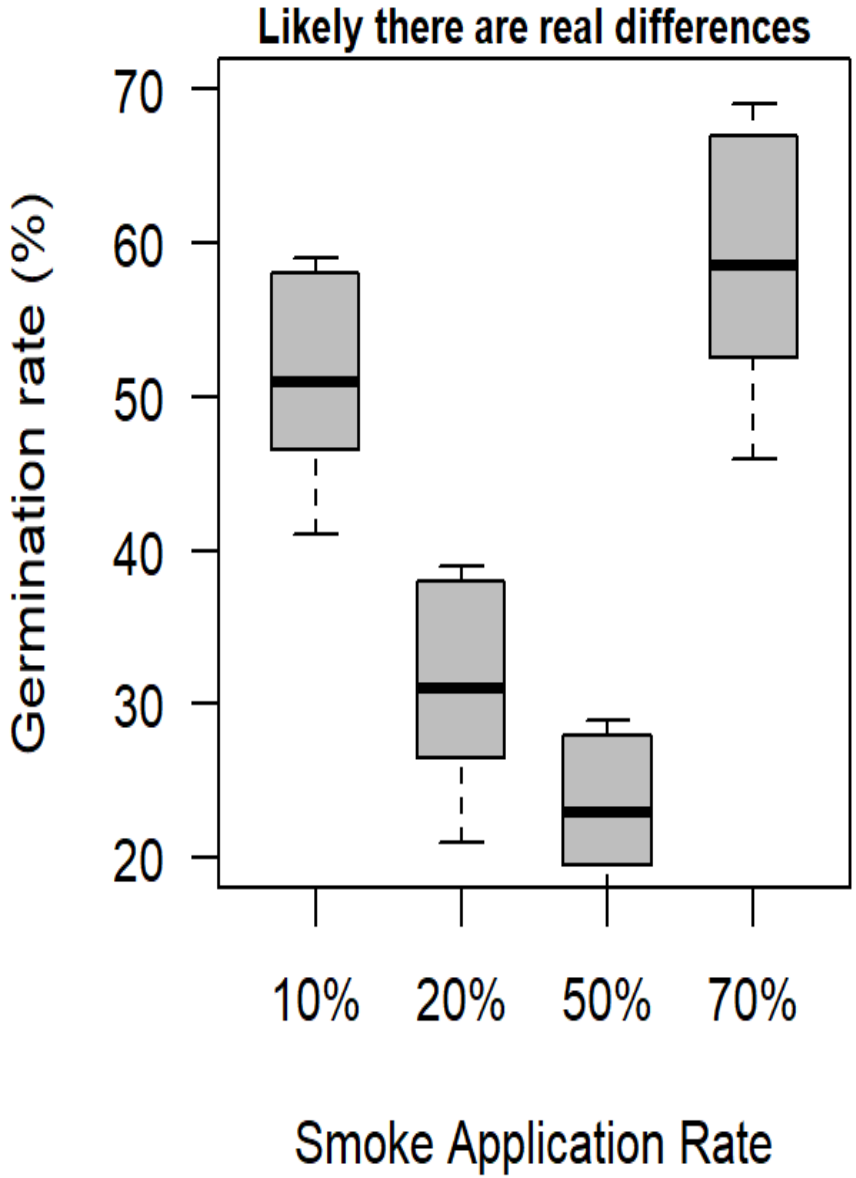
(1) Min. 1st Qu. Median Mean 3rd Qu. Max. St.d.
.0037 .1967 .4577 .4601 .6920 .9707 .2821

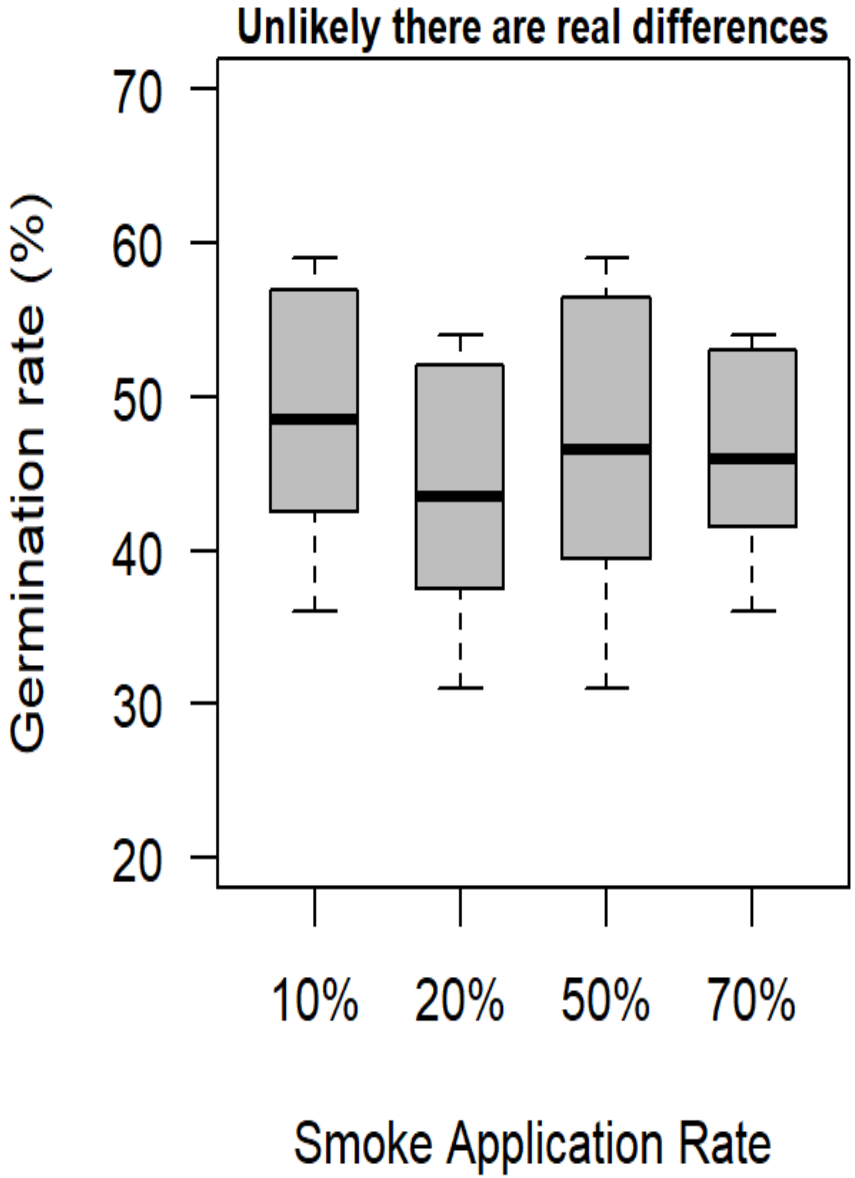
(2) Min. 1st Qu. Median Mean 3rd Qu. Max. St.d.
-2.2060 -.4512 .1167 .1298 .9329 2.2570 .9507

(3) Min. 1st Qu. Median Mean 3rd Qu. Max. St.d.
.0214 .2483 .5749 .9590 1.2900 4.2170 .9856

註：第三筆資料明顯右偏；第一筆資料比第二筆資料更為「均勻」。

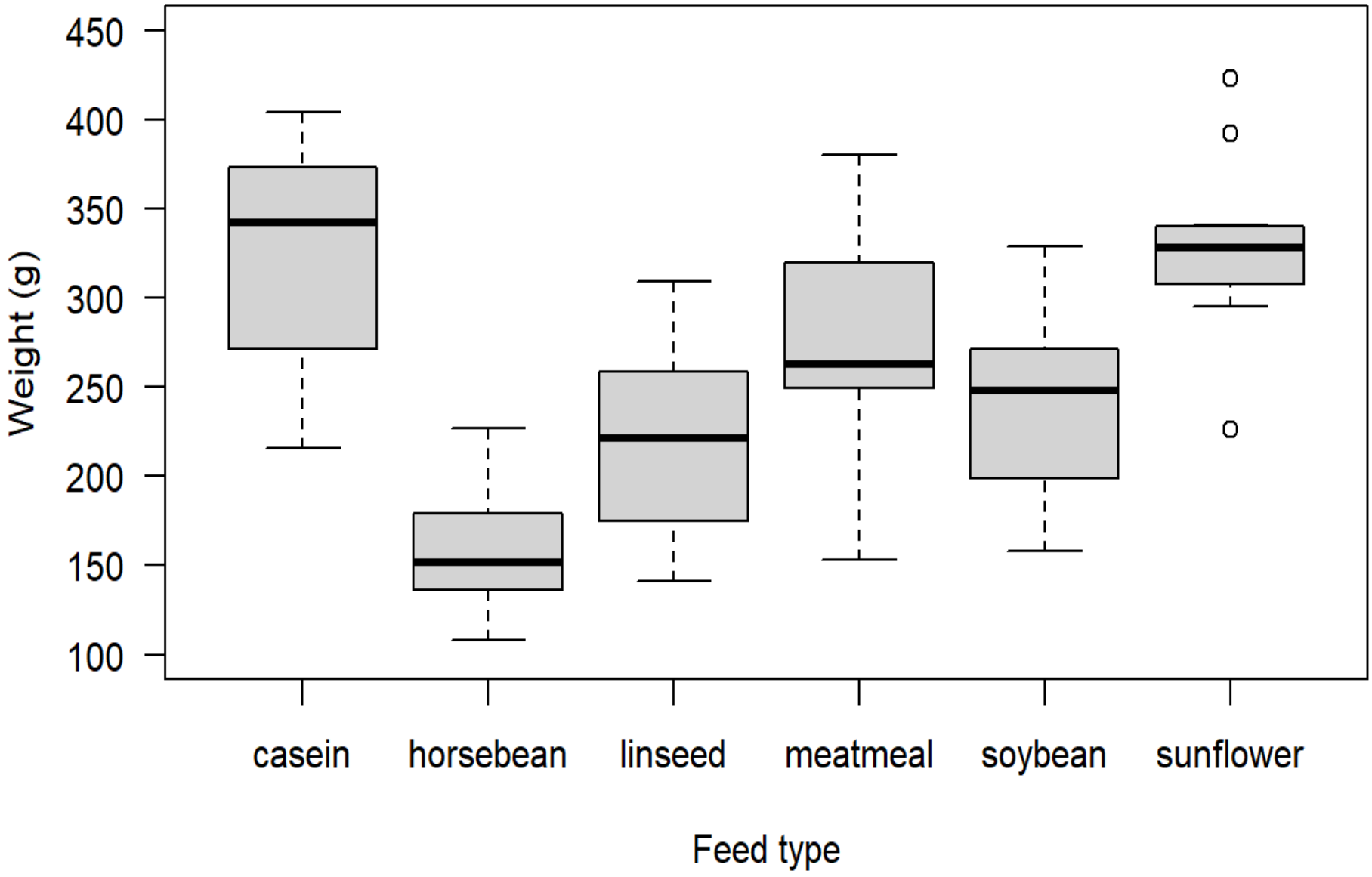
Boxplot can detect between variations!





https://saeappliedstats.tech/analysis-of-variance.html

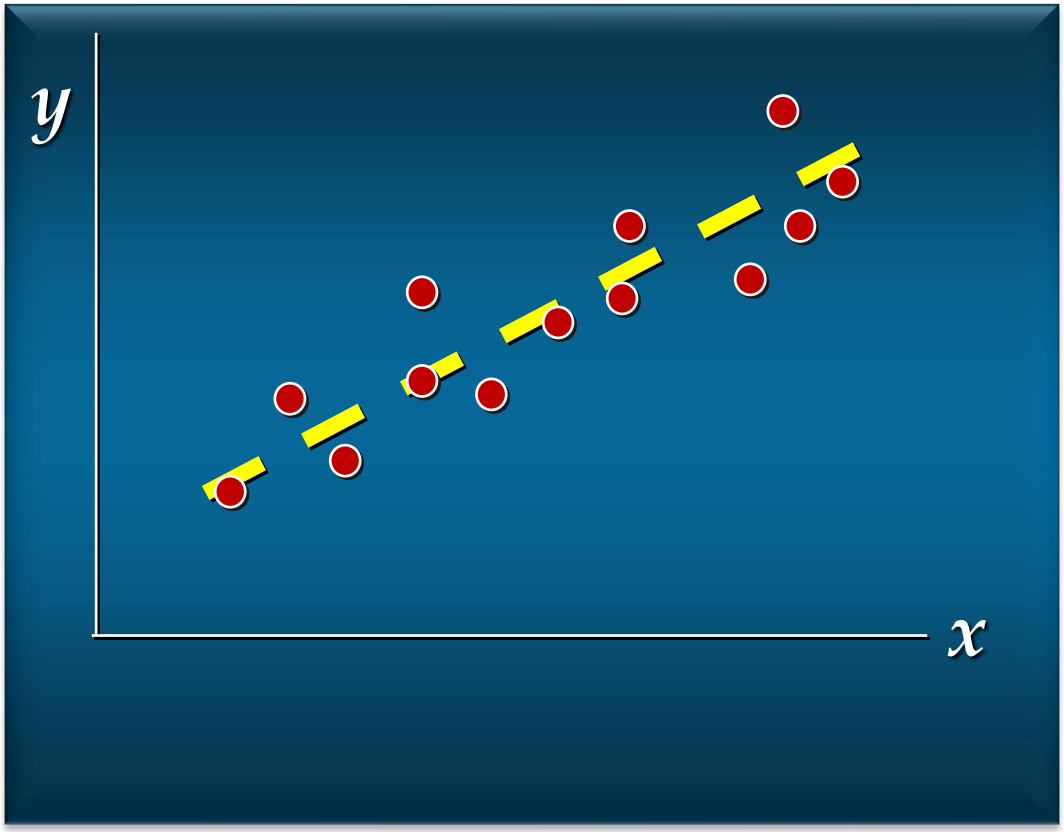
R data “chickwts”, weights of chickens fed



https://saeappliedstats.tech/GitHub-Bookdown_files/figure-html/unnamed-chunk-237-1.png

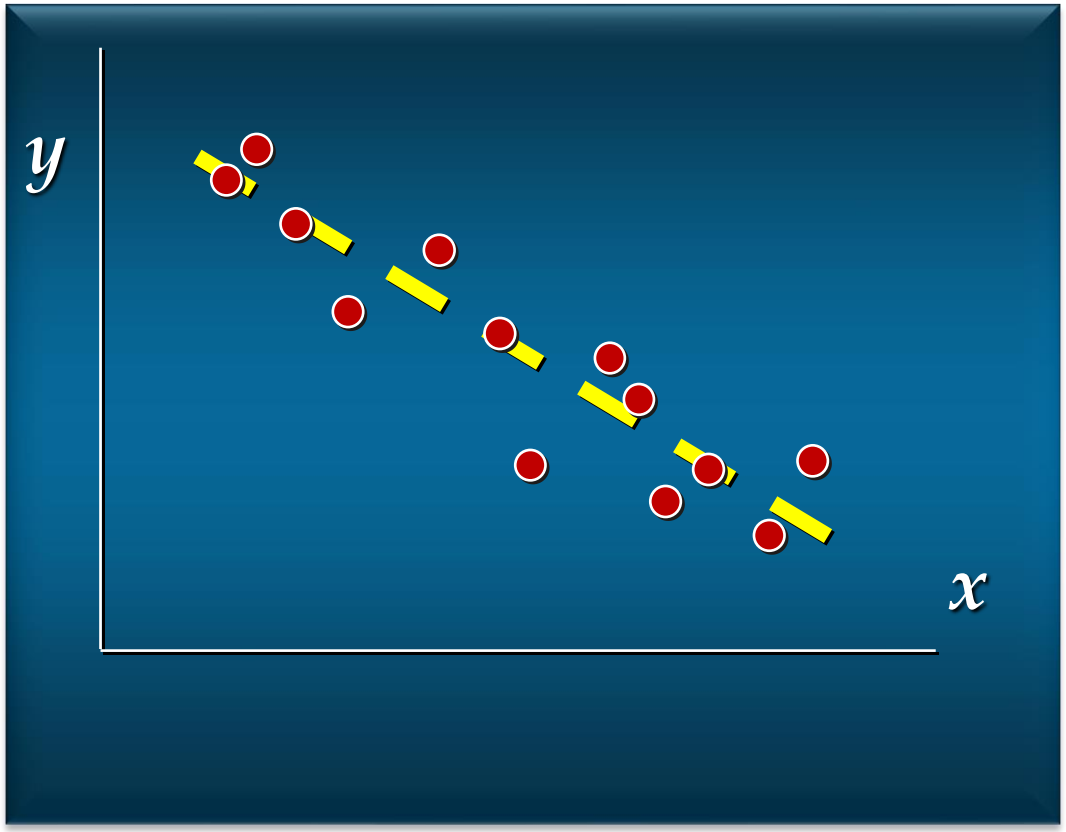
散佈圖(Scatter Diagram)

- 正向關係



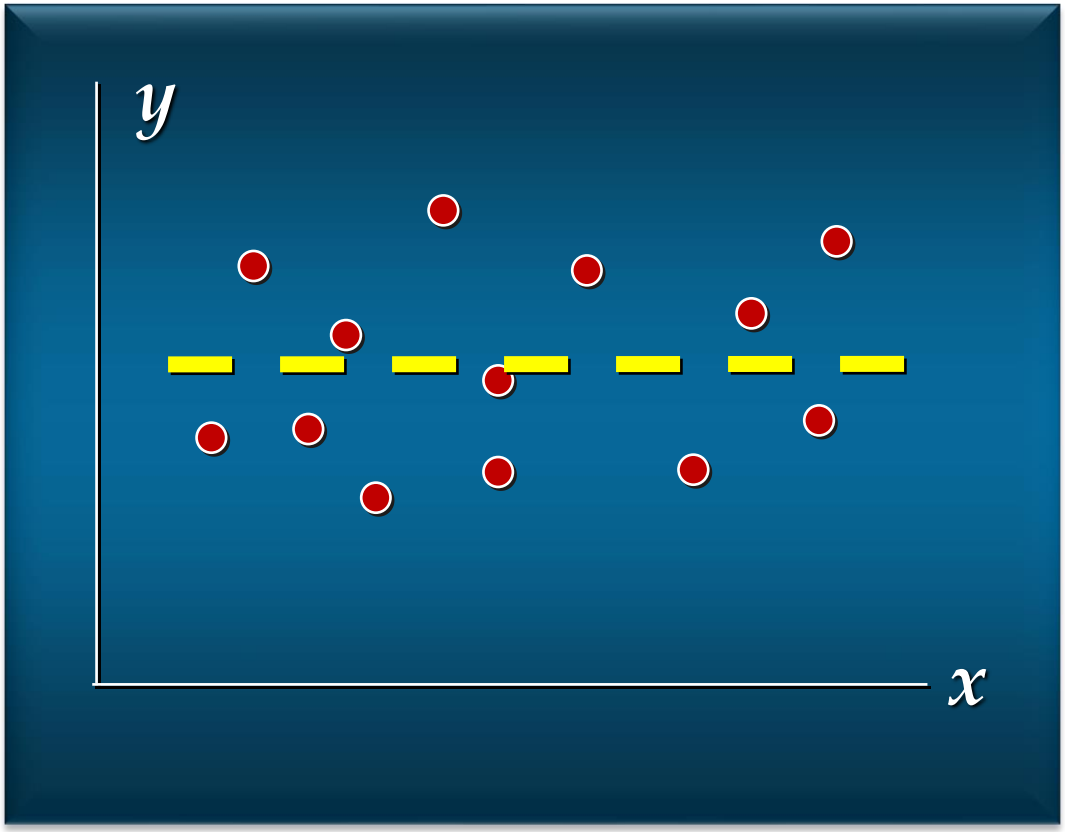
散佈圖

- 負向關係

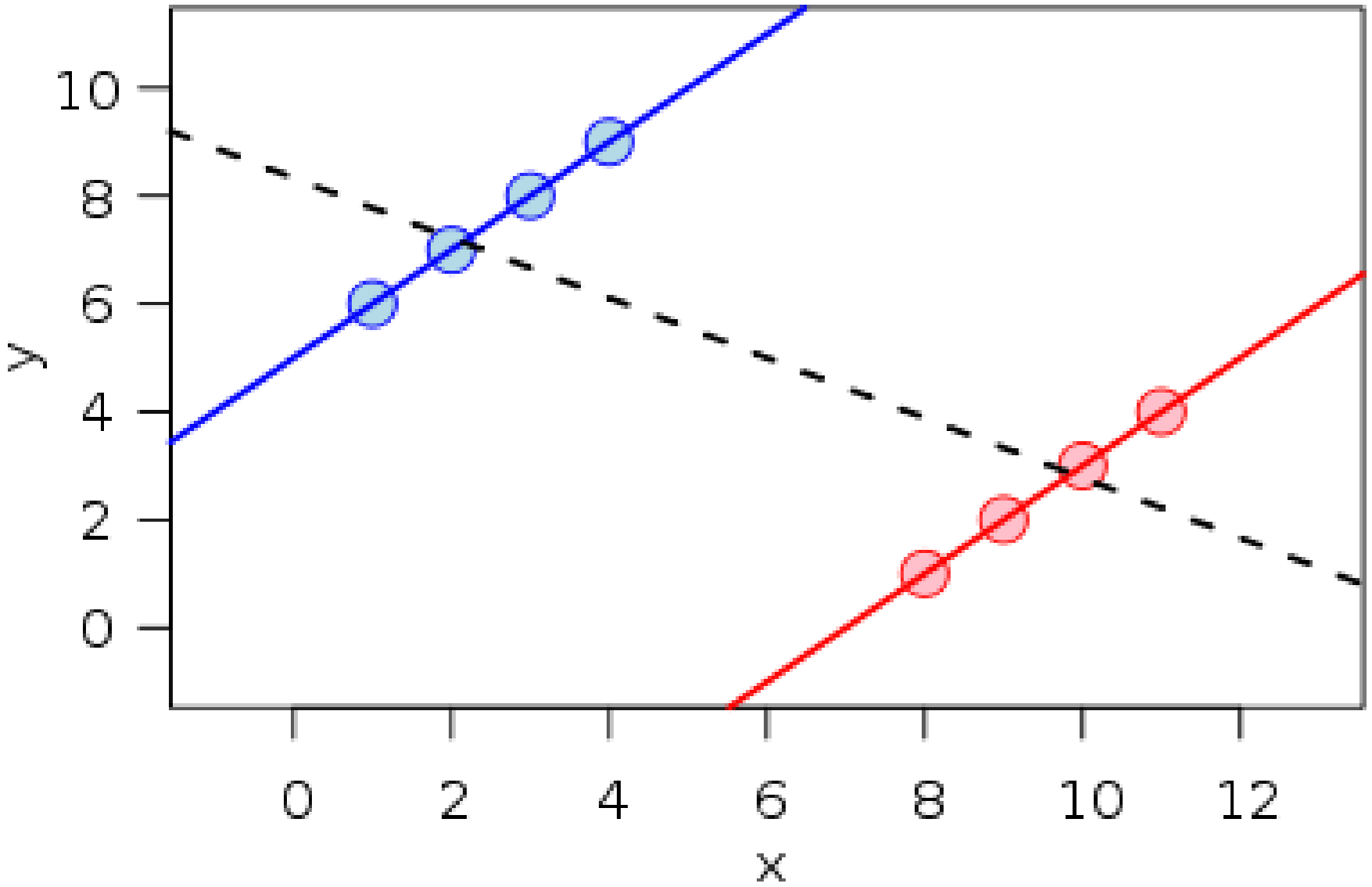


散佈圖

- 沒有明顯關係



辛普森謬論(Simpson Paradox)



抽樣分配(Sampling Distribution)

- 樣本所衍生的資訊稱為統計量(Statistic)，而統計量的機率分配稱為抽樣分配。
- 較常見的參數統計量為期望值、變異數的估計。

$$\rightarrow \hat{\mu} = \bar{x}_n = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\rightarrow \hat{\sigma}^2 = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$$

中央極限定理 (Central Limit Theorem)

- 若 $x_1, x_2, \ldots, x_n$ 為互相獨立、且具有同樣分配的觀察值(或稱為一組隨機樣本)，則當樣本數 n 趨近於無窮大：

$$\frac{\bar{x}_n - \mu}{s/\sqrt{n}} \rightarrow N(0,1)$$

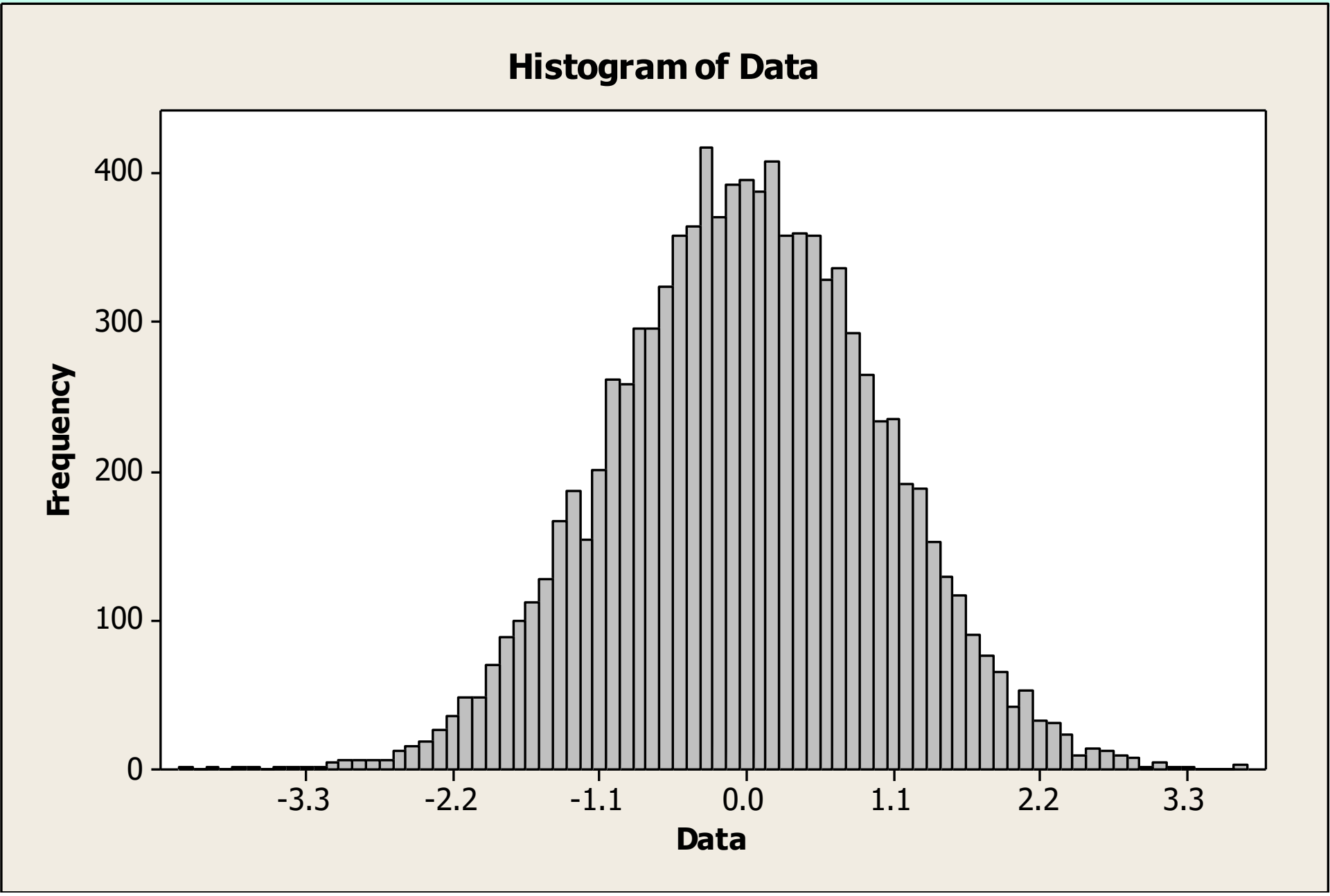
註：互相獨立且來自於同一分配可記為i.i.d. (Identically and Independently distributed)與隨機(Random)樣本。

中央極限定理

- 中央極限定理(Central Limit Theorem, CLT)是統計非常重要的定理，指的是「當觀察值個數非常大時，其平均數的抽樣分配接近常態分配」。

→常態分配(Normal or Gaussian Distribution)是一個形狀接近鐘形(Bell-shaped)的統計分配，與常見的「20-80定律」有關。

鐘型分配曲線的範例


Histogram of Data
Frequency
400
300
200
100
0
-3.3
-2.2
-1.1
0.0
1.1
2.2
3.3
Data

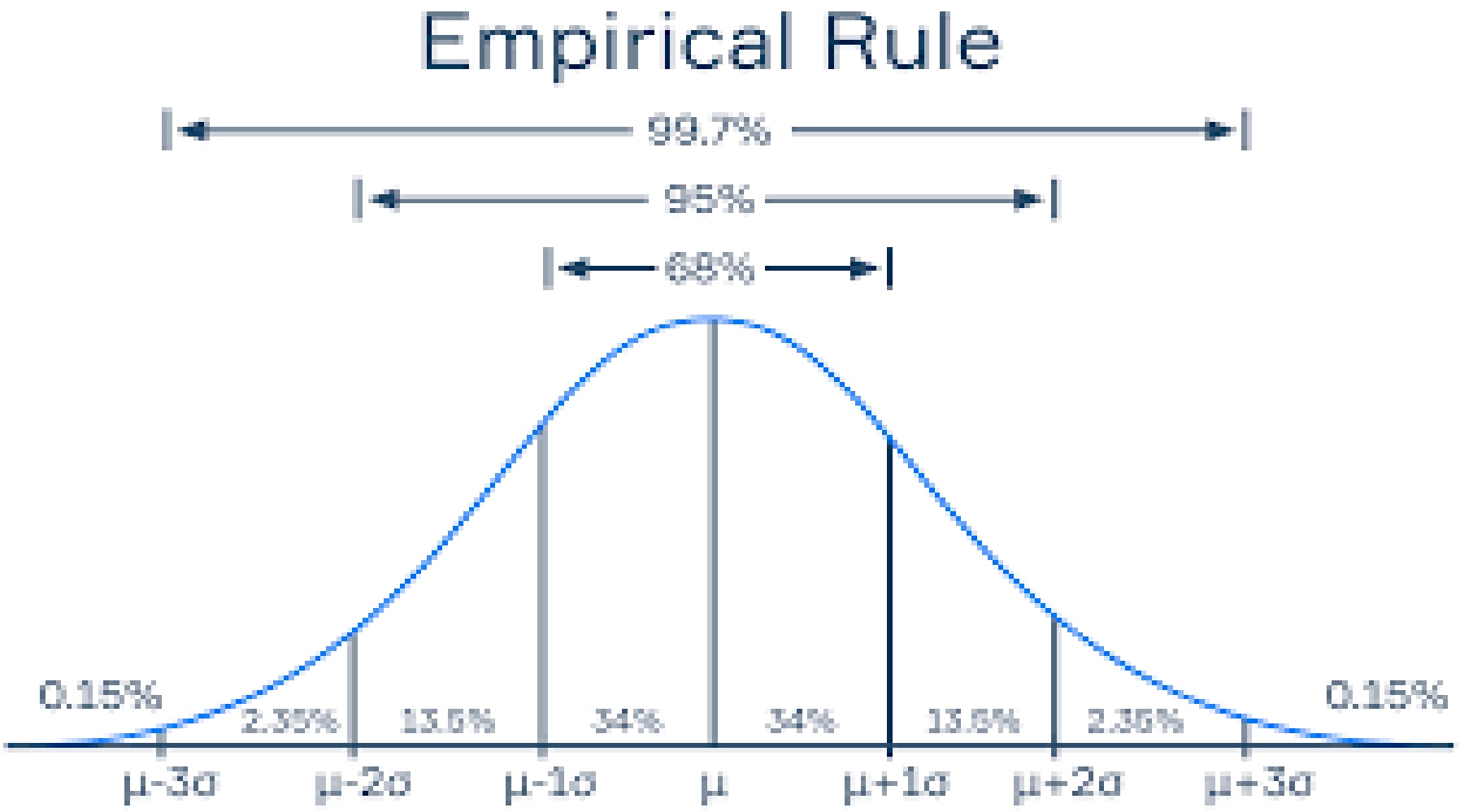
經驗法則: 若為鐘型分配資料

- 1.約有68%的資料落入 $(\mu-\sigma,\mu+\sigma)$
- 2.約有95%的資料落入 $(\mu-2\sigma,\mu+2\sigma)$
- 3.約有99.72%的資料落入 $(\mu-3\sigma,\mu+3\sigma)$

- 註: 1. 以上各母體參數可用樣本數值代替，例如 $s^2 \to \sigma^2, \bar{x} \to \mu$

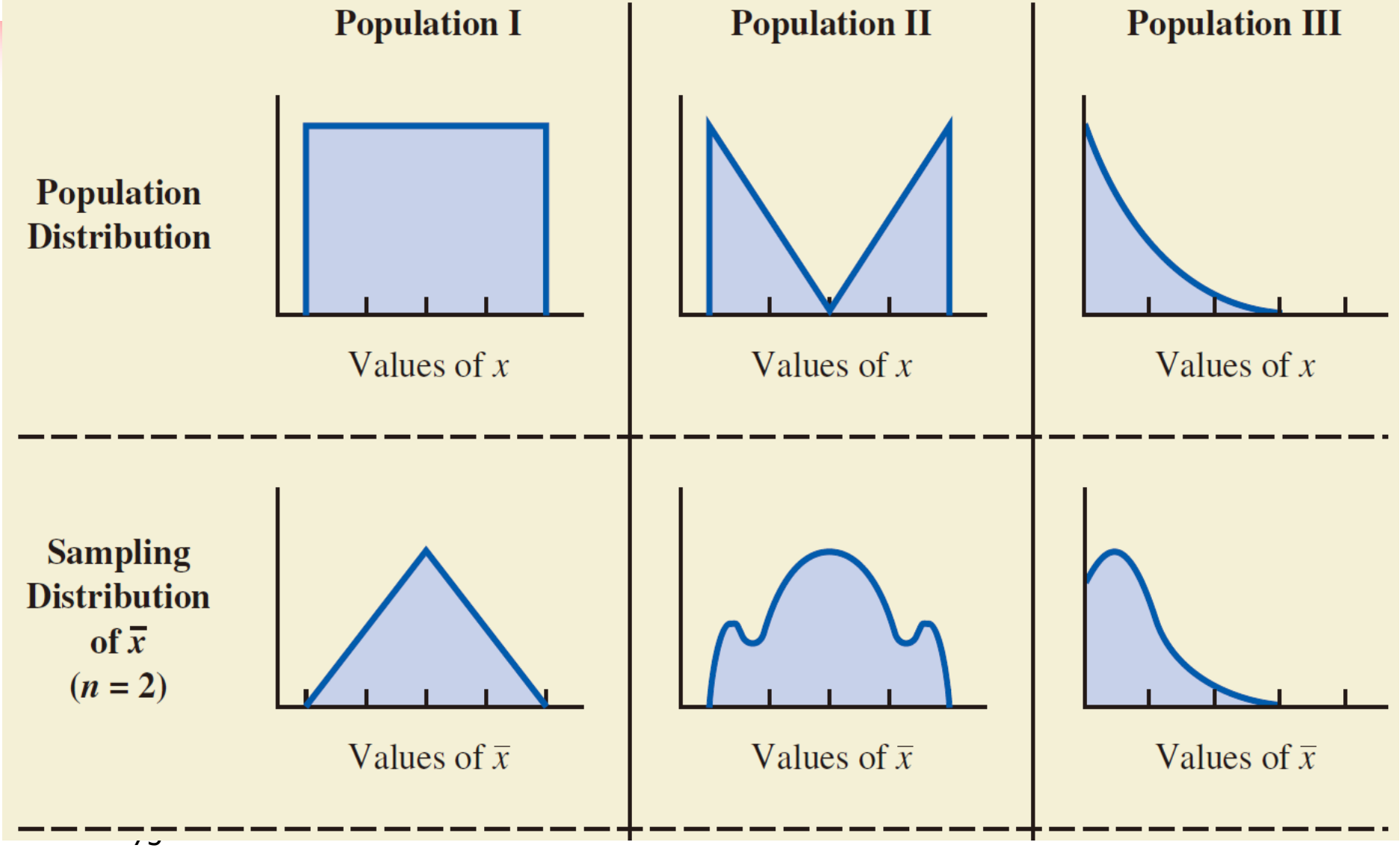
  2.也有人用全距/4或全距/6代替s。

什麼是經驗法則？

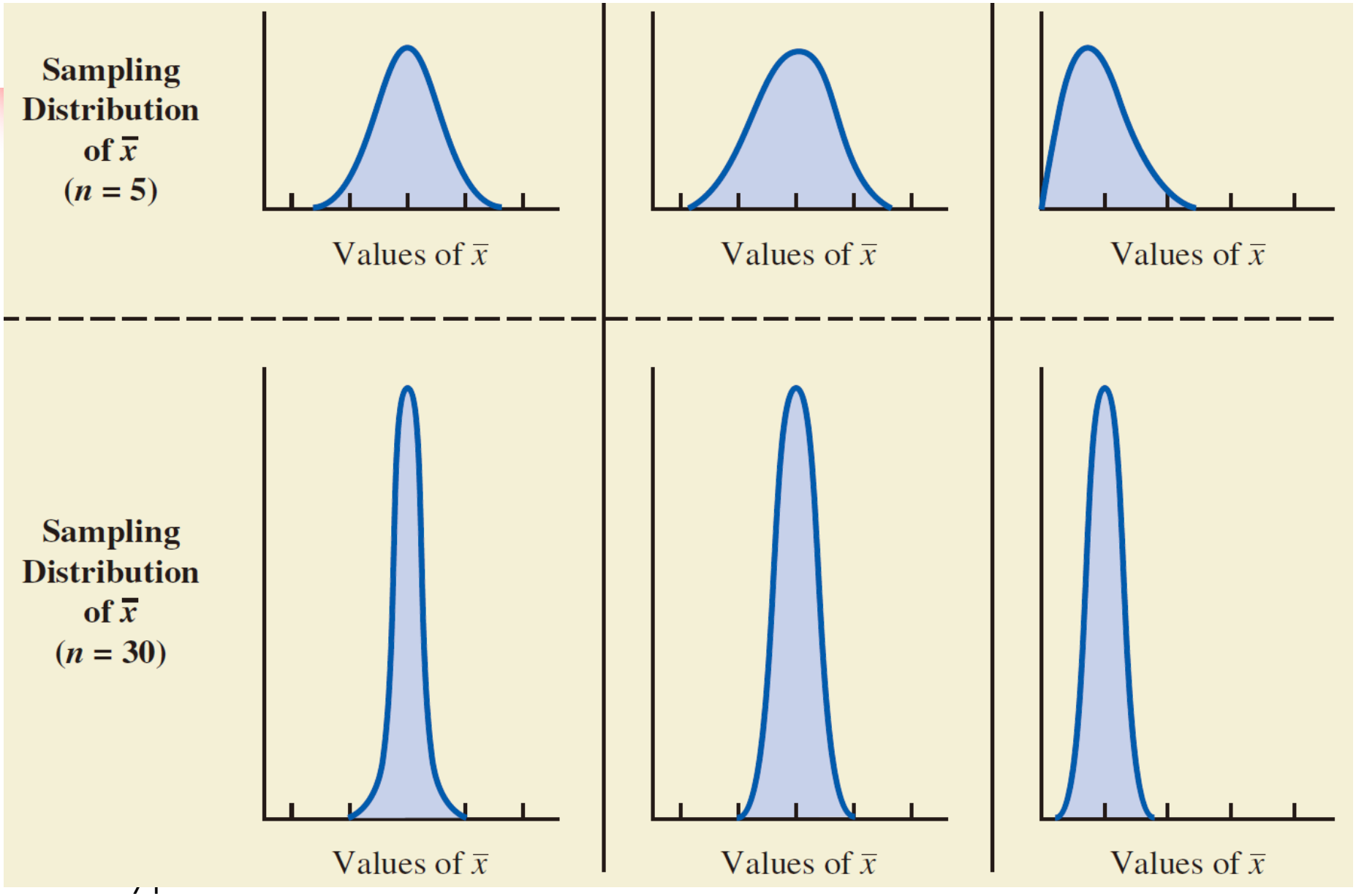


https://hyperskill.org/learn/step/24676

Central Limit Theorem



Central Limit Theorem (Conti.)



Detecting Outliers (離群值)

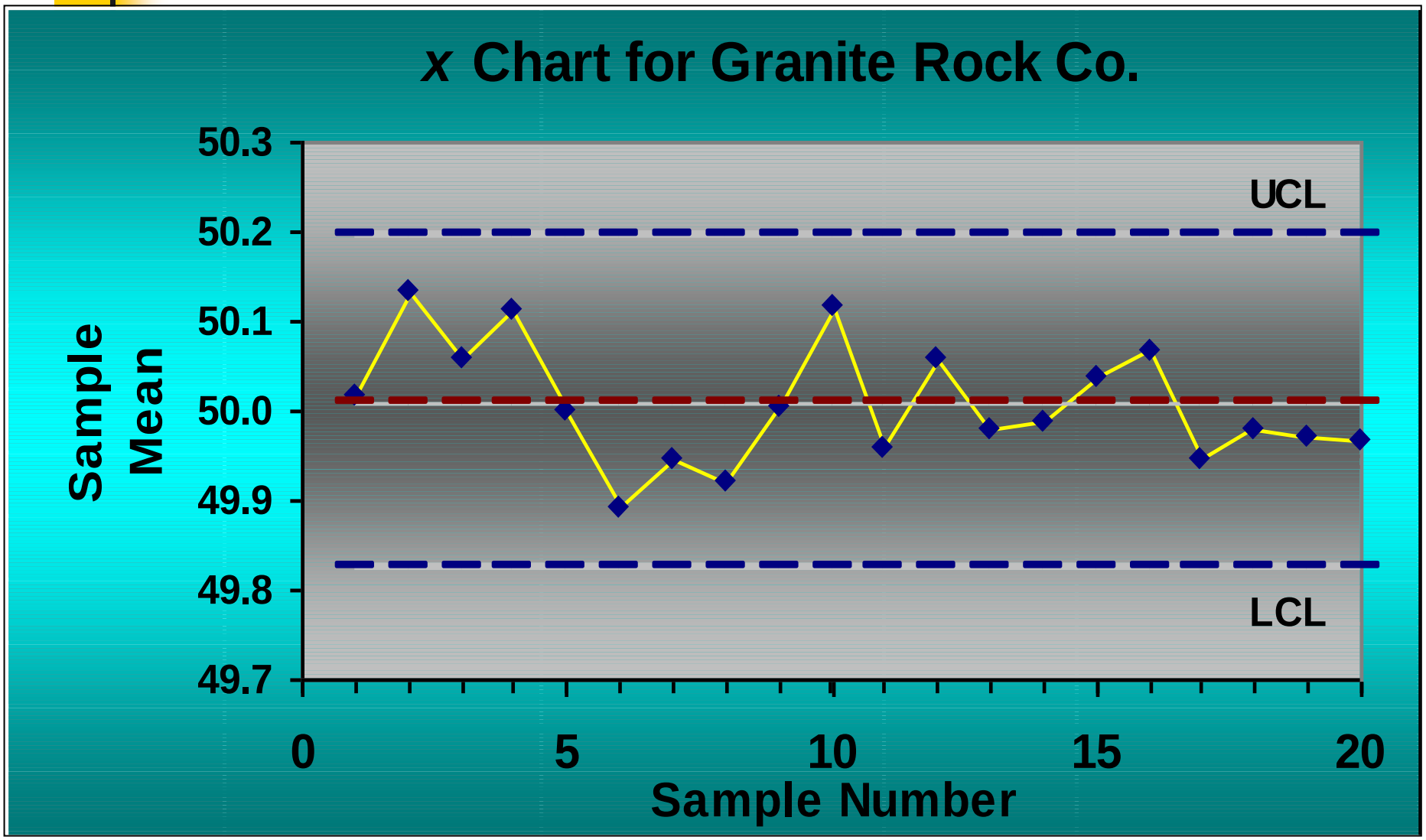
- 離群值是資料中異常大或異常小數值。

→常見定義是Z分數(z-score)小於–3或大於 +3的觀察值。（註：$Z = \frac{x - \mu}{\sigma}$）

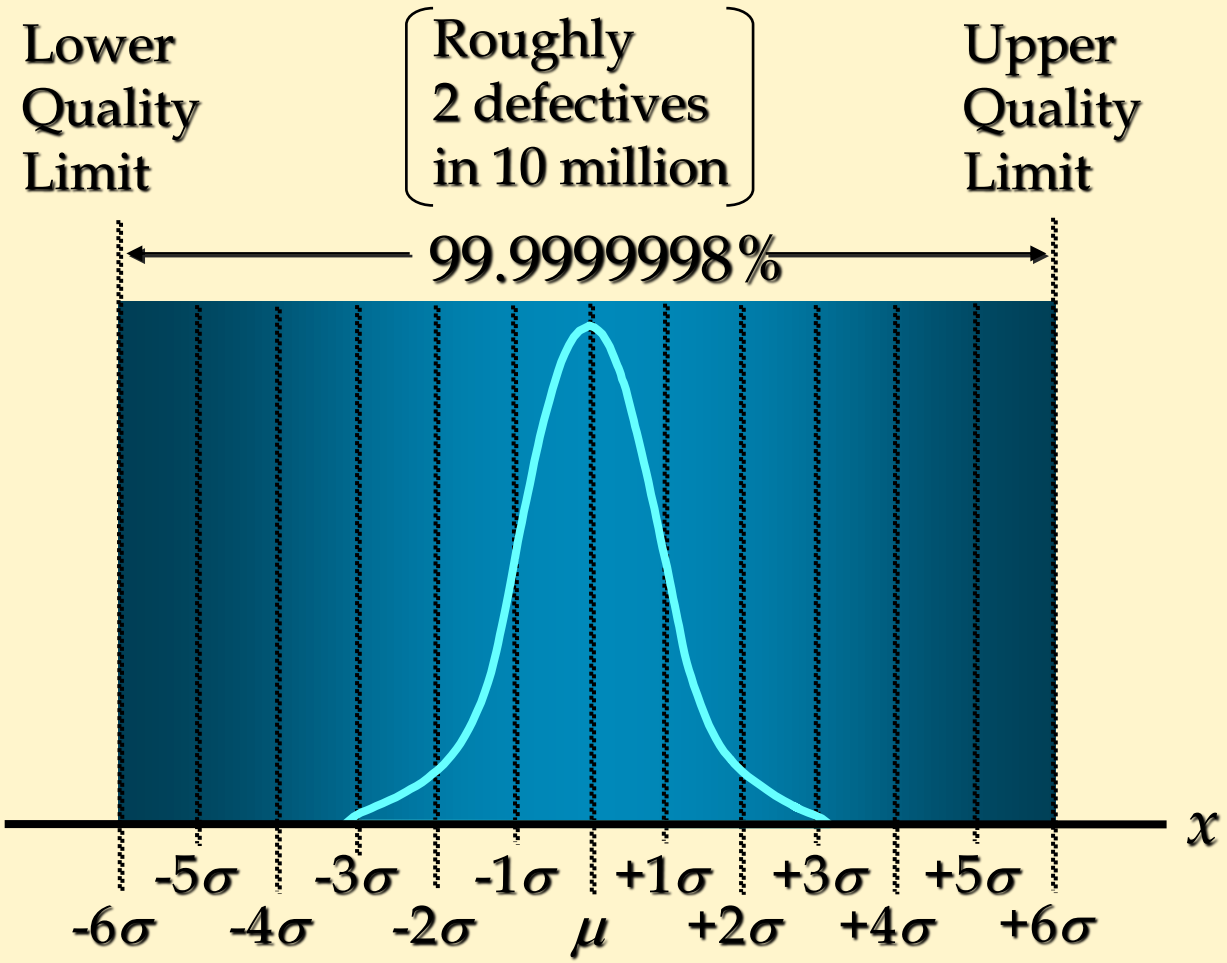
- 離群值的發生原因包括：記錄（書寫）錯誤、或是歸檔錯誤。

→離群值的發生也是正常現象，以Z分數大於3、小於–3為例，發生可能約千分之三。

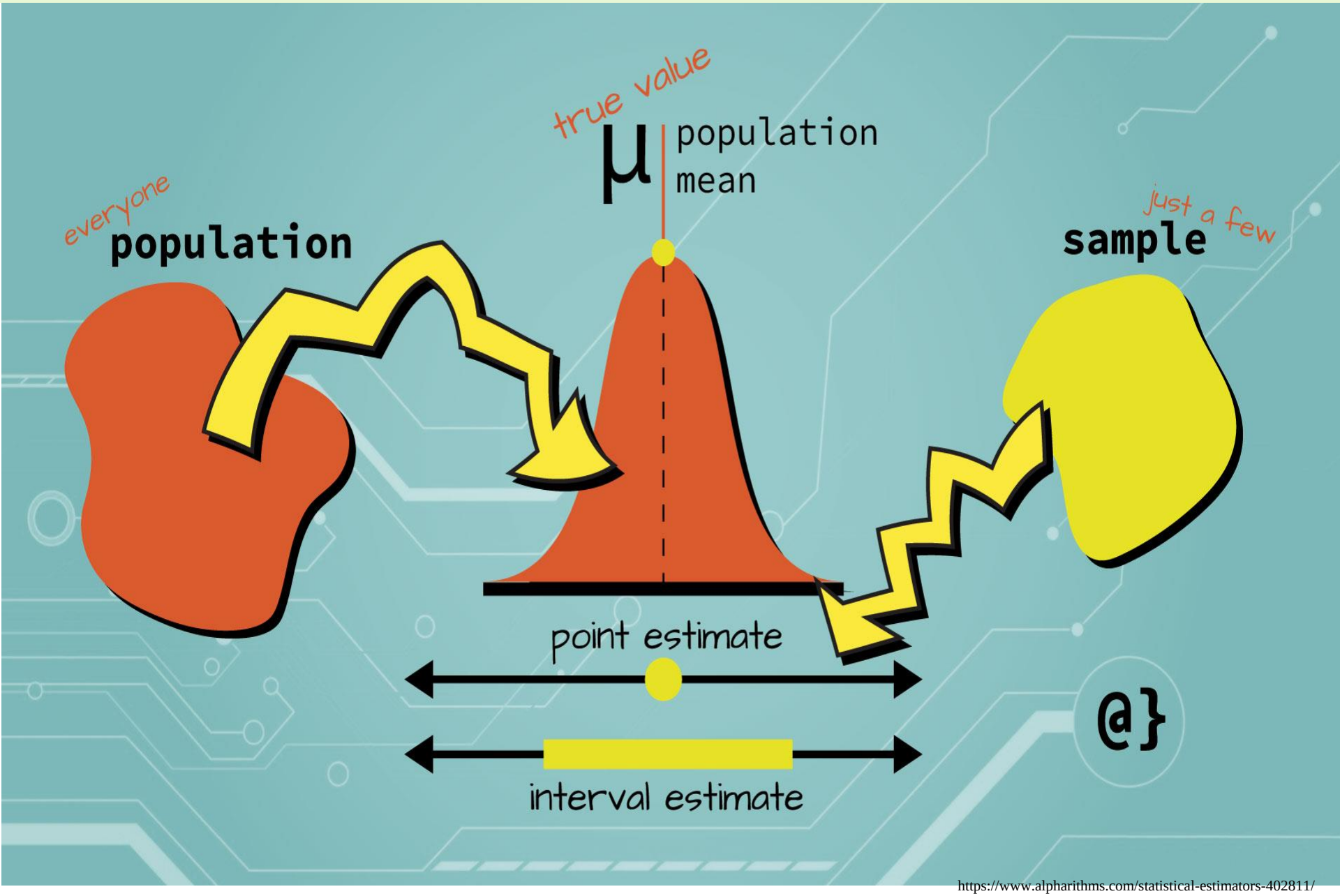
離群值的實例應用


x Chart for Granite Rock Co.
50.3
50.2
50.1
50.0
49.9
49.8
49.7
Sample Mean
UCL
LCL
0
5
10
15
20
Sample Number

現代品管—Six Sigma



關於統計估計



https://www.alpharithms.com/statistical-estimators-402811/

統計估計

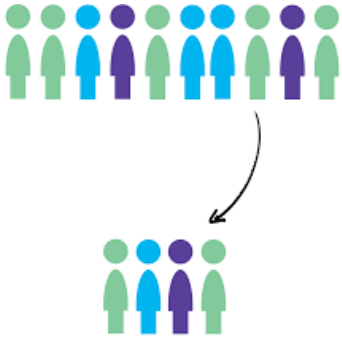
- 對於母體參數的估計，可分為點估計、區間估計兩種。

→點估計多半會是最有可能的數值，或是樣本平均數、樣本變異數。

→區間估計則是**點估計**加上**最大誤差**(Marginal Error)為範圍，例如：

$$\bar{x} \pm \text{ Margin of Error}$$

最大誤差與**容許錯誤**、**誤差分配**有關。



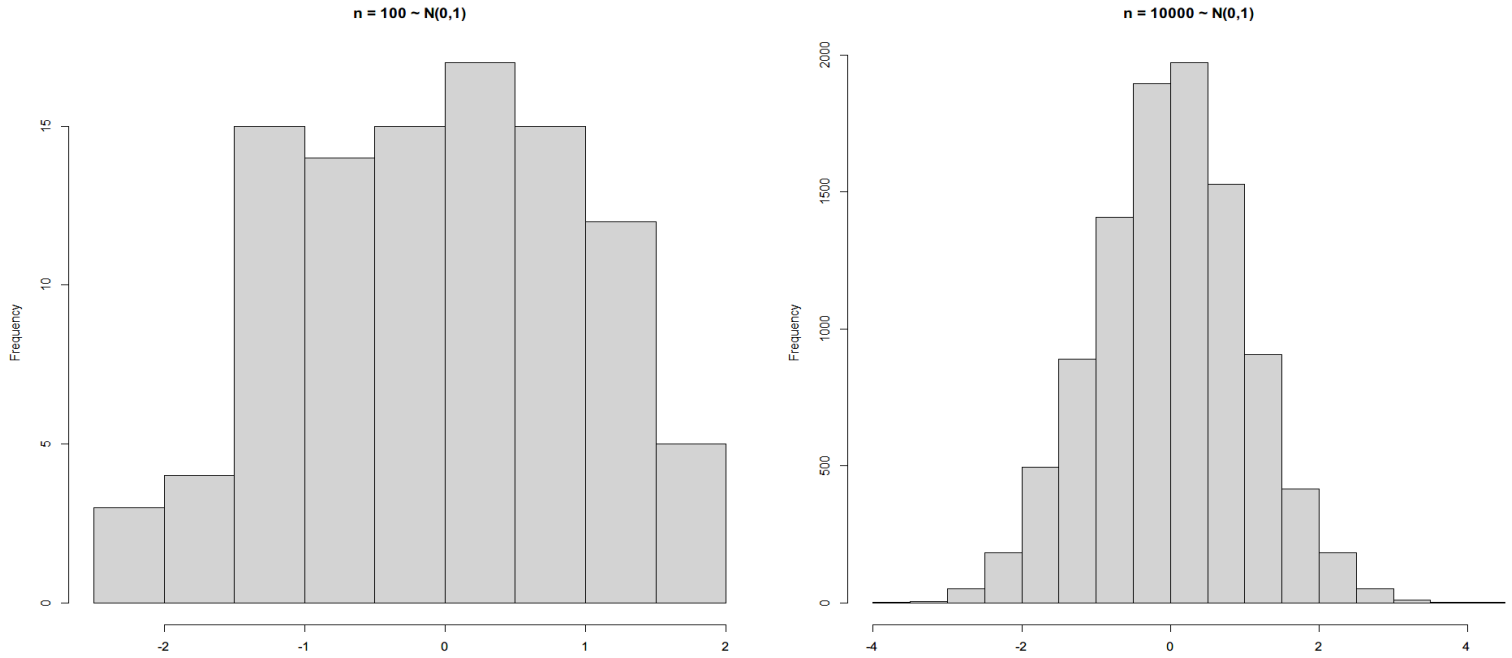
母體與樣本

https://www.qualtrics.com/experience-management/research/population-vs-sample/

- 觀察值反映母體特徵：

$$\text{Observations} = \text{Truth} + \text{Error}$$

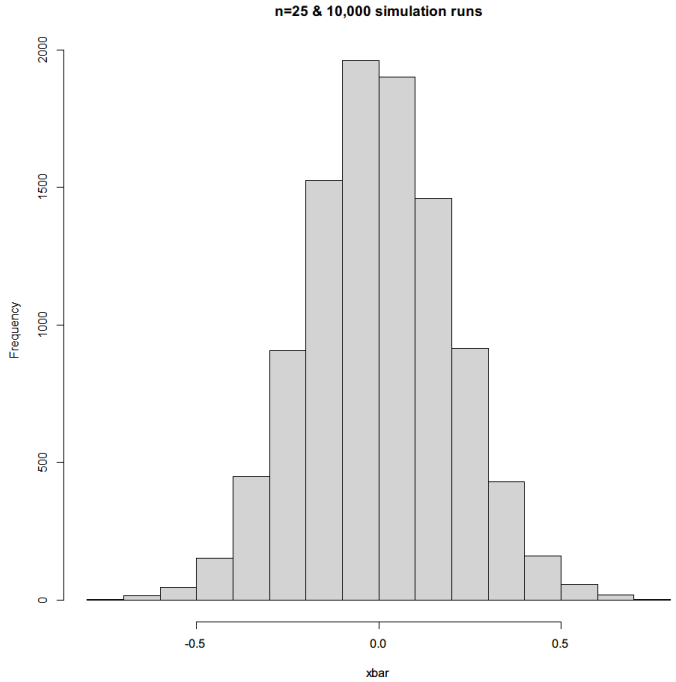
→觀察值愈多、愈能瞭解母體特性。

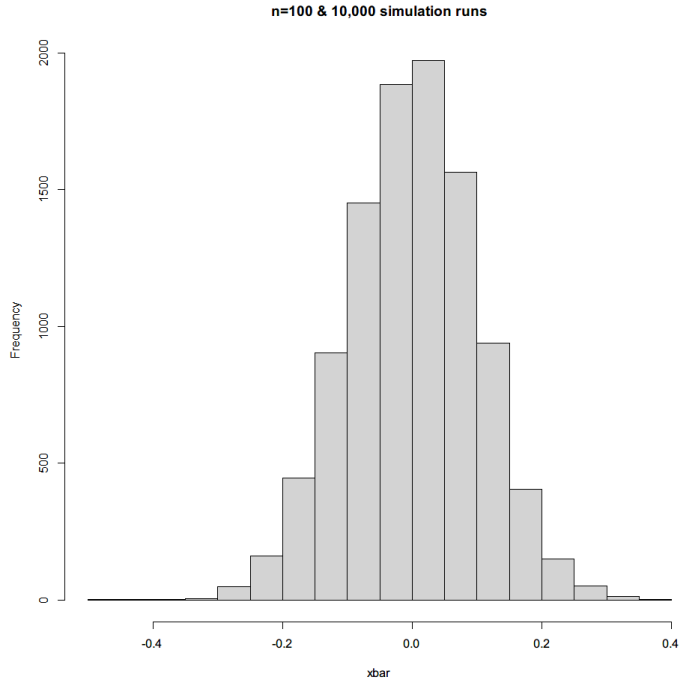


樣本數與估計誤差

- 如果估計期望值，觀察值服從常態分配，樣本數愈多、平均數分佈愈集中！

變異數為0.0401 & 0.0098 (約4倍)





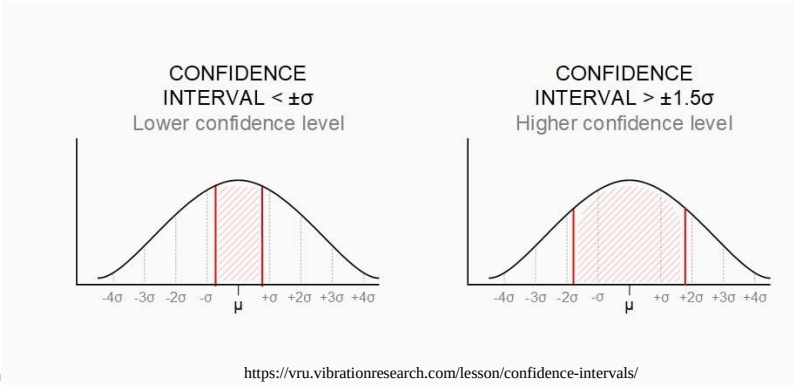
關於信賴區間的幾個疑問

- 95%信賴區間的詮釋：不代表每次建立的信賴區間，有95%的機會涵蓋真實的參數。

→問題：那又該如何解釋？

- 天氣預測時，經常會提到下雨機率，例如30%下雨機率如何詮釋？
- 颱風未來行進方向的預測，通常隨著時間而擴大，如何以統計角度解釋？另外，為什麼預測範圍會擴大？

抽樣誤差與區間估計



https://vru.vibrationresearch.com/lesson/confidence-intervals/

- 抽樣誤差伴隨的不確定性，通常以區間(範圍)估計因應，精確度要求愈高、愈無法達成。

→信心係數愈大、信賴區間也愈寬！

- 如何詮釋信賴區間？

→95%不代表建立信賴區間後，涵蓋真實參數值的機率！！（註：參數值未知且**固定**，只有落在信賴區間之內或之外兩種結果。）

註：貝氏統計對參數的認知略有不同。

生活中關於機率的案例

- 預測臺北市明天下雨機率為30%，如何詮釋？

→過去有100次和明天類似的氣候環境，平均有30次下雨。(問題：氣象署的下雨定義？)

→明天臺北市有30%的面積會下雨。

- 某位棒球選手打擊率為三成，如何詮釋？

→選手維持三成打擊率，每次上場打擊時有30%出現安打，並非預測下次擊出安打的正確性為三成。(隨機猜中選擇題答案也類似。)

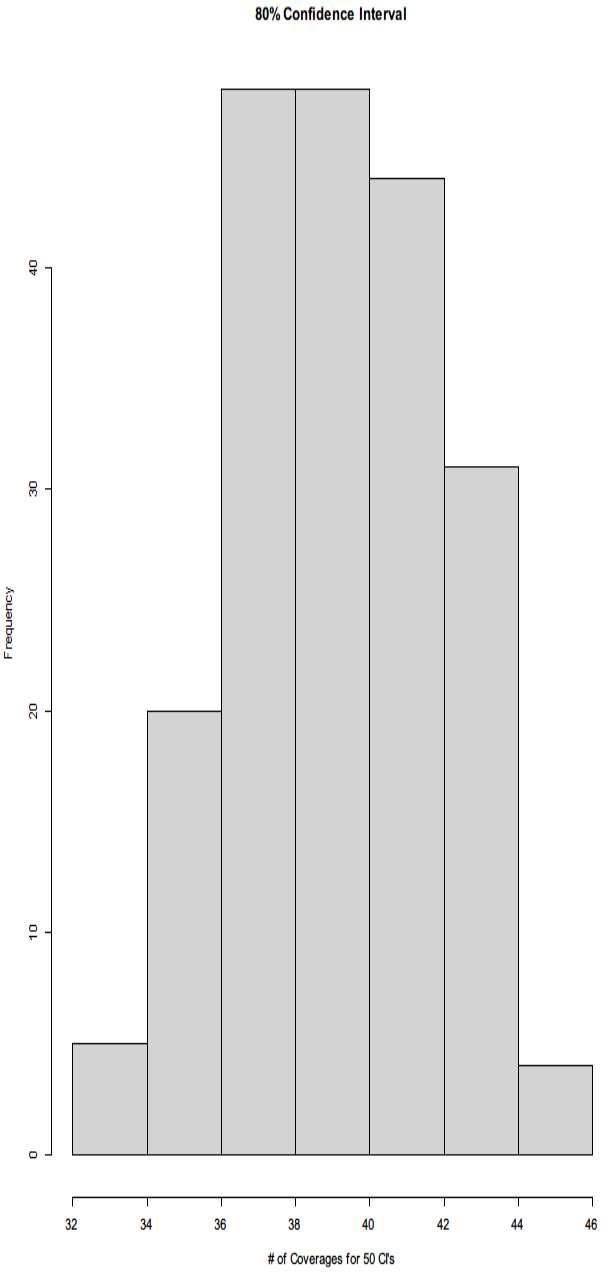
如何詮釋信心係數？

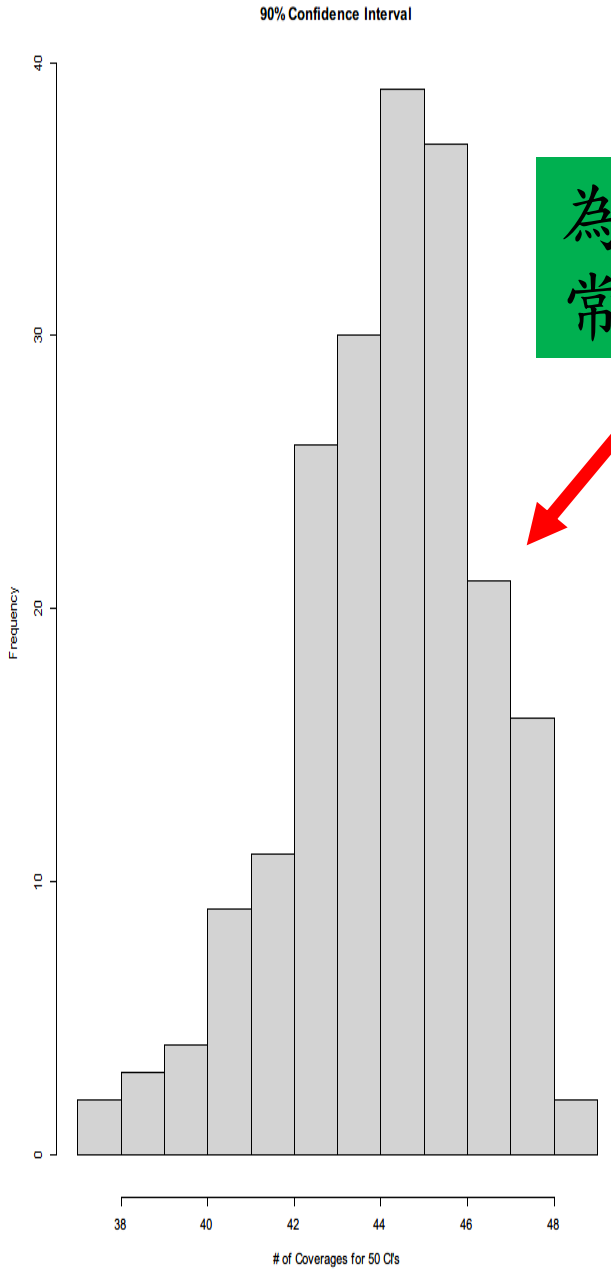
- 信賴區間為**程序機率**，不用於單次解釋。

→一旦取得樣本資料，點估計、區間估計就已確定，沒有所謂的機率、或是信心水準。

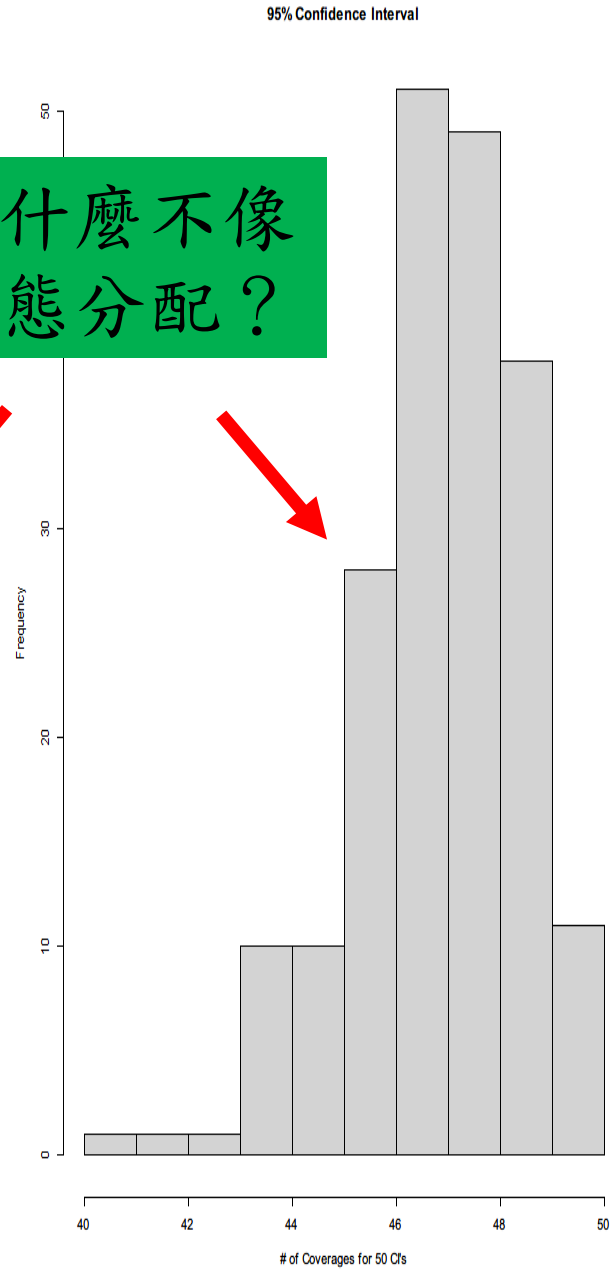
- 再次以電腦模擬驗證信心係數，隨機產生100個標準常態分配的亂數，80%、90%、95%對應臨界值為1.281552、1.644854、1.959964，重複一萬次電腦模擬，涵蓋真實期望值的比例分別為79.33%、89.30%、94.64%。

每50個信賴區間涵蓋真實值的個數(200次)

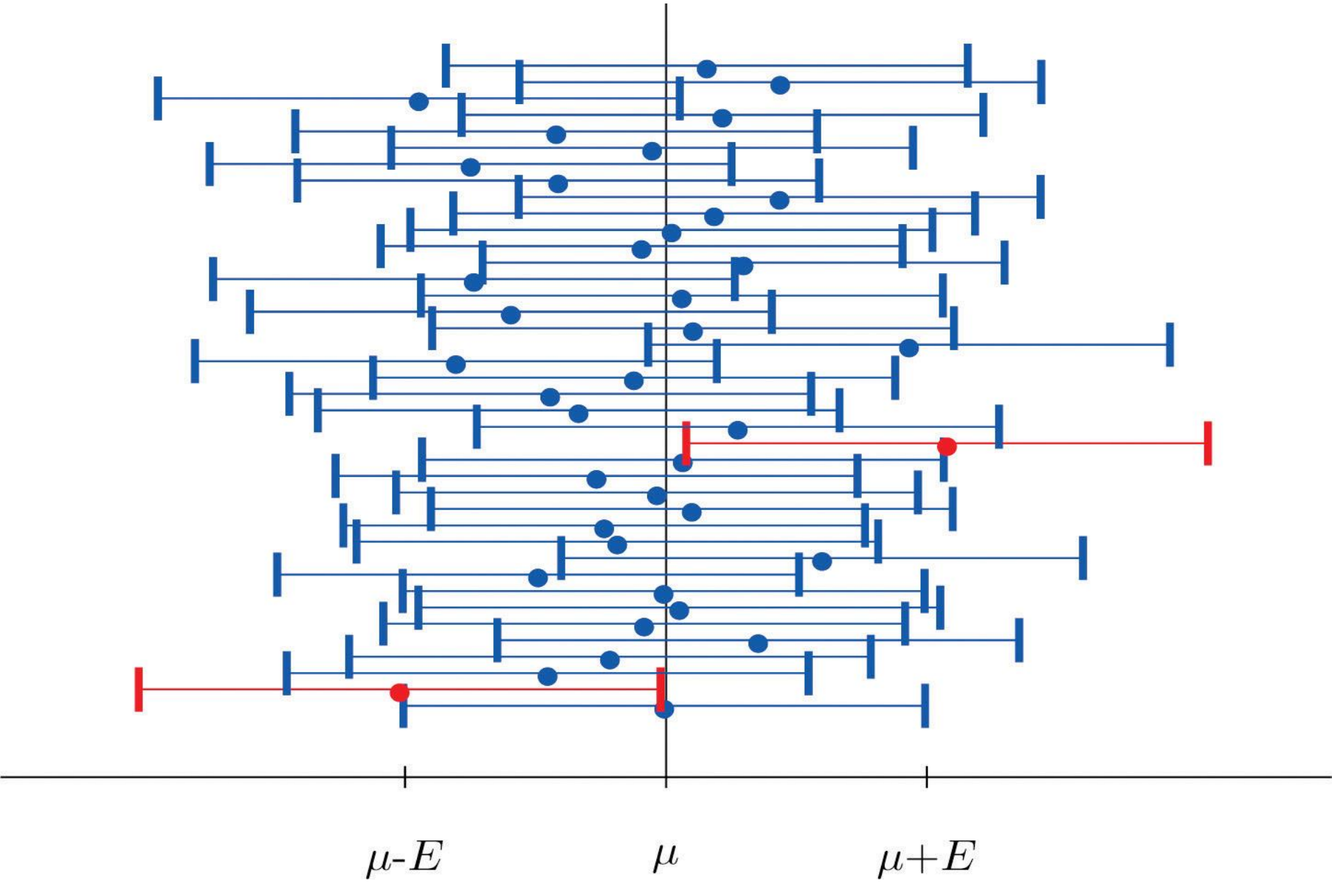




為什麼不像
常態分配？



95%信賴區間的範例




颱風路徑潛勢預報圖
2008/07/27 14:00 LST
30日14時
29日14時
28日14時
28日02時
27日14時
27
26
25
22

虛無假設與對立假設的設定

- 假設檢定（Hypothesis Testing）通常以虛無假設為中心，確定H_0是否為真。

→例如：探討汽車油耗 μ（公里/公升），通常有三種可能設定：

1. One-tailed, lower tail: $H_0: \mu \geq \mu_0$ $H_a: \mu < \mu_0$
2. One-tailed, upper tail: $H_0: \mu \leq \mu_0$ $H_a: \mu > \mu_0$
3. Two-tailed: $H_0: \mu = \mu_0$ $H_a: \mu \neq \mu_0$

Hypothesis Testing（假設檢定）

- Hypothesis testing is to determine whether a statement about the value of a population parameter should or should not be rejected.
- Null hypothesis (虛無假設), H_0 , is a tentative assumption about a population parameter.
- Alternative hypothesis (對立假設), H_a, is the opposite of what is stated in the null hypothesis.

P-value與假設檢定

- p值（p-value）：當虛無假設為真時，出現與樣本觀察值類似（或更極端）結果的機率。

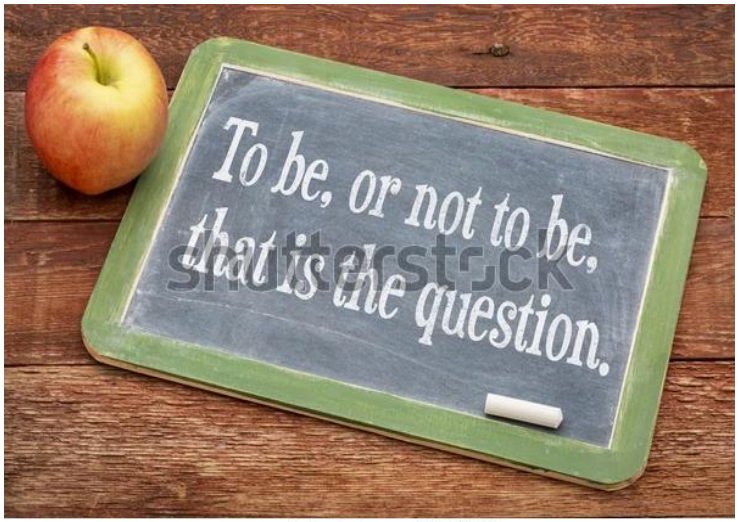
→p值可視為在虛無假設下，樣本觀察值發生的機率。**（註：機率愈小、虛無假設愈不可能）**

- 統計分析從結果反推原因**（Inverse Probability）**

→在不知道真實狀況下，只能以發生機率判斷（或排除）可能原因。

Suggested Guidelines for p-Values

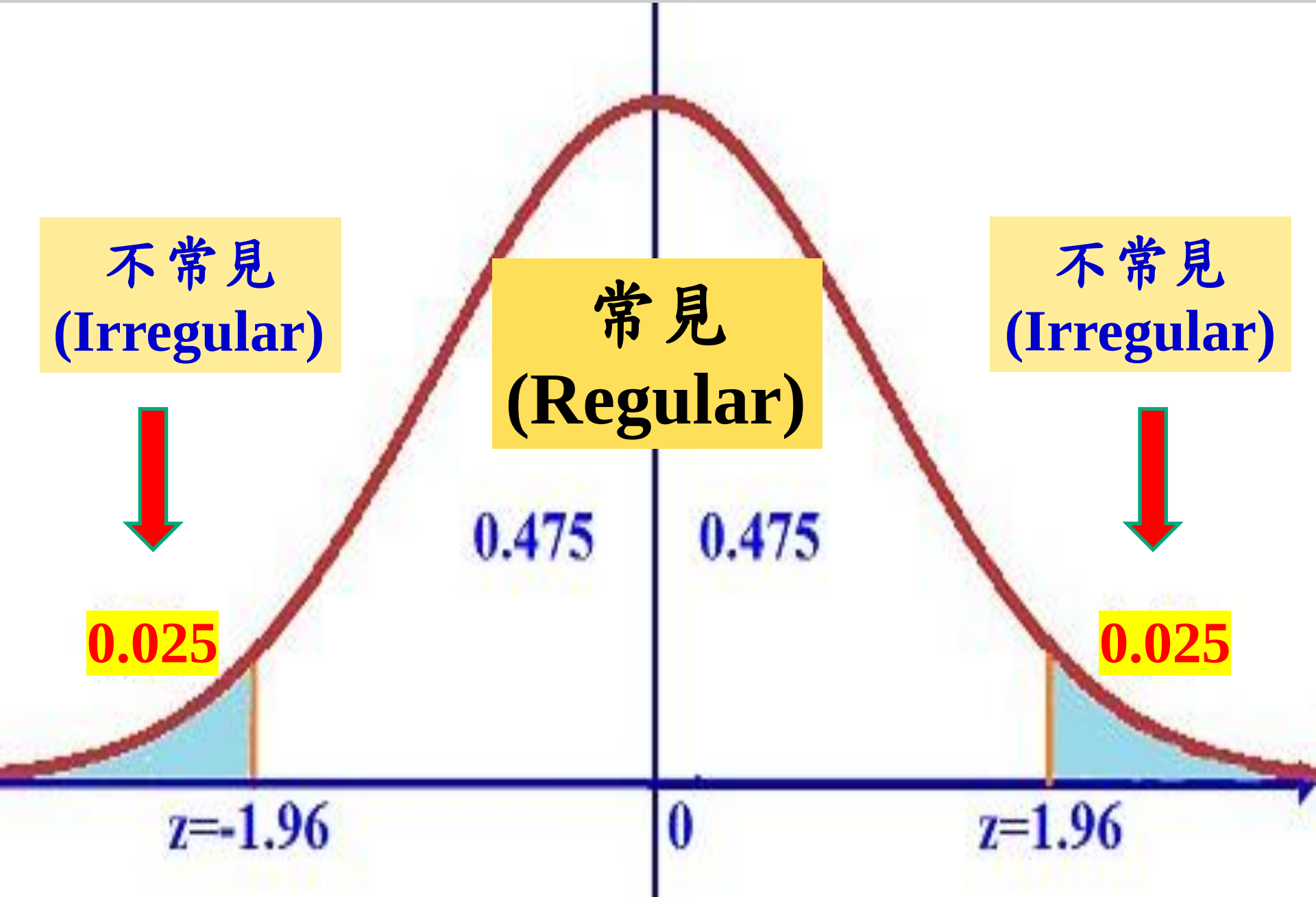
- Less than 0.01: Overwhelming evidence to reject H_0
- Between 0.01~ 0.05: Strong evidence to reject H_0
- Between 0.05 ~ 0.10: Weak evidence to reject H_0
- Greater than 0.10: Insufficient evidence to reject H_a



https://image.shutterstock.com/image-photo/be-not-that-question-text-600w-246702544.jpg

https://image.shutterstock.com/image-vector/vector-illustration-hamlet-play-isolated-600w-131784596.jpg

P-value與信賴區間的機率意涵



Type I Error

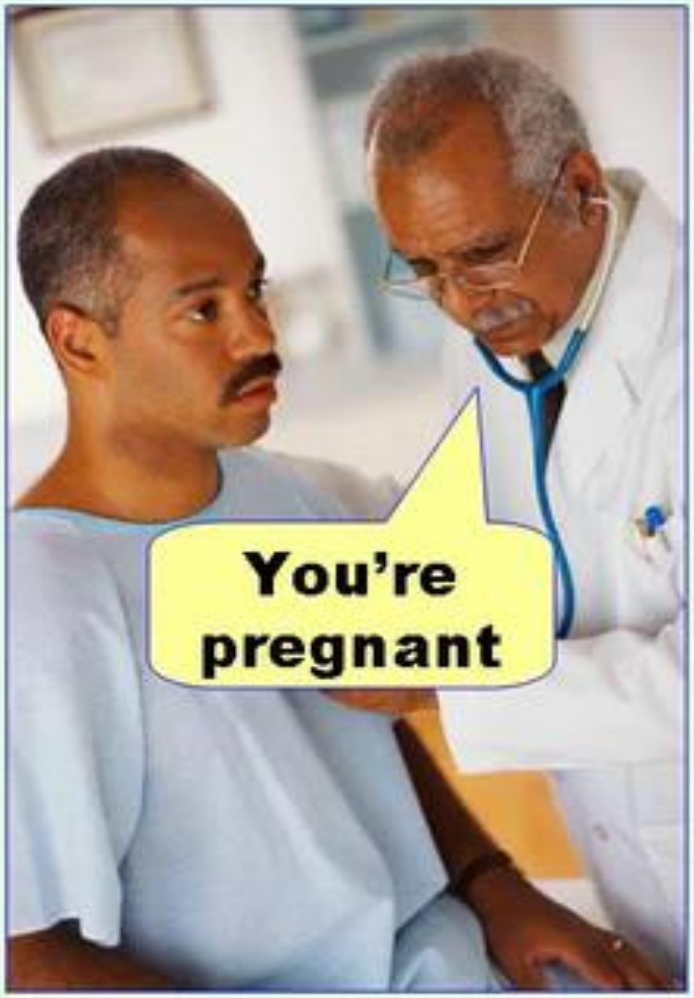
- Because hypothesis tests are based on sample data, we must allow for the possibility of errors.
- A Type I error is rejecting H_0 when it is true.
- The probability of making a Type I error when the null hypothesis is true as an equality is called the level of significance.
- Hypothesis testing that only control for Type I error is often called significance test.

Type II Error

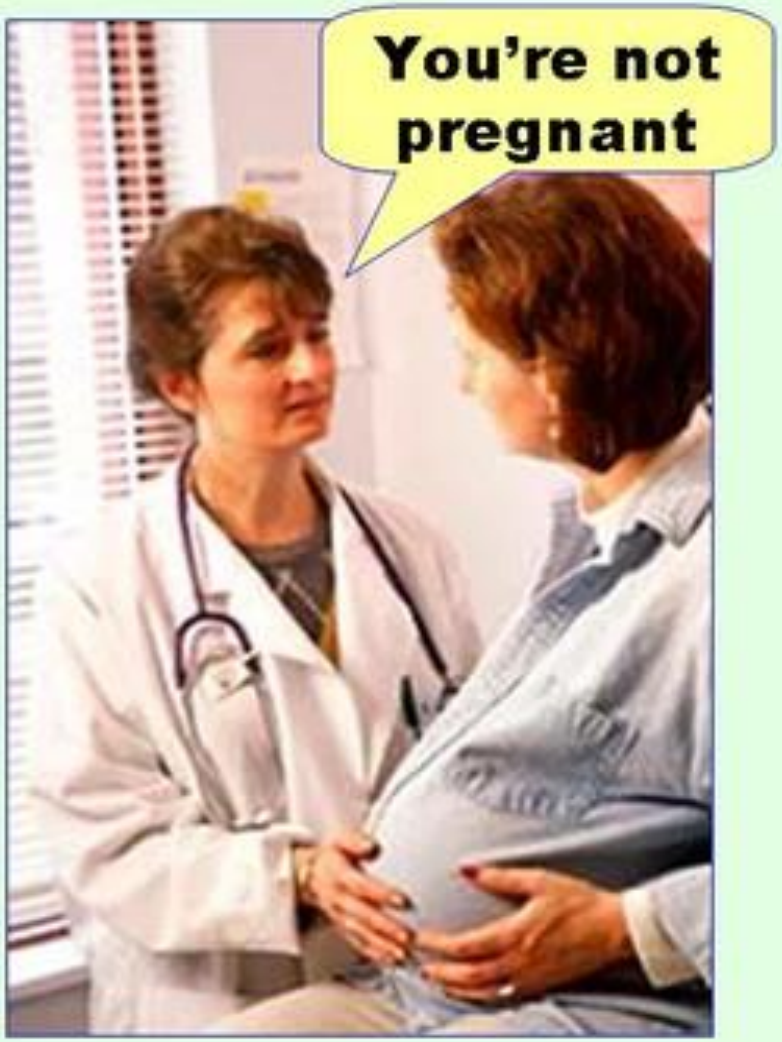
- Type II error: Do not reject H_0 when it is false.
- It is difficult to control for the Type II error.
- Statisticians avoid the risk of making a Type II error by using "**do not reject H_0**" rather than "accept H_0".

註：檢定結論只有「拒絕H_0」和「不拒絕H_0」。
→無法得知真實狀況，根據資料分析結果排除可能性較低者！（拒絕與對立假設有關。。。）

Type I error
(false positive)



Type II error
(false negative)



https://cdn.shopify.com/s/files/1/0882/2472/files/type-i-and-type-ii-errors_grande.jpg?1529108046624782202

Type I and Type II Errors

	Population Condition	
Conclusion	H_0 True ($\mu \leq 12$)	H_0 False ($\mu > 12$)
Accept H_0 (Conclude $\mu \leq 12$)	Correct Conclusion	Type II Error
Reject H_0 (Conclude $\mu > 12$)	Type I Error	Correct Conclusion

	有病者	無病者
檢驗結果 陽性 +	真陽性 a	偽陽性 c
檢驗結果 陰性 -	偽陰性 b	真陰性 d

https://epaper.ntuh.gov.tw/health/201606/images/health_5_clip_image002.jpg

敏感性 $= \frac{a}{a+b}$ · 真陽性率：有病者檢驗結果為陽性的比率

特異性 $= \frac{d}{c+d}$ · 真陰性率：無病者檢驗結果為陰性的比率

虛無假設H_0為沒病，偽陽性為型一誤差、偽陰性為型二誤差。

- **型一誤差(Type-1 Error)等於P(拒絕H_0|H_0為真)**
- **型二誤差(Type-2 Error)等於P(不拒絕H_0|H_0不為真)**