「商業資料分析與管理決策」—分析與決策

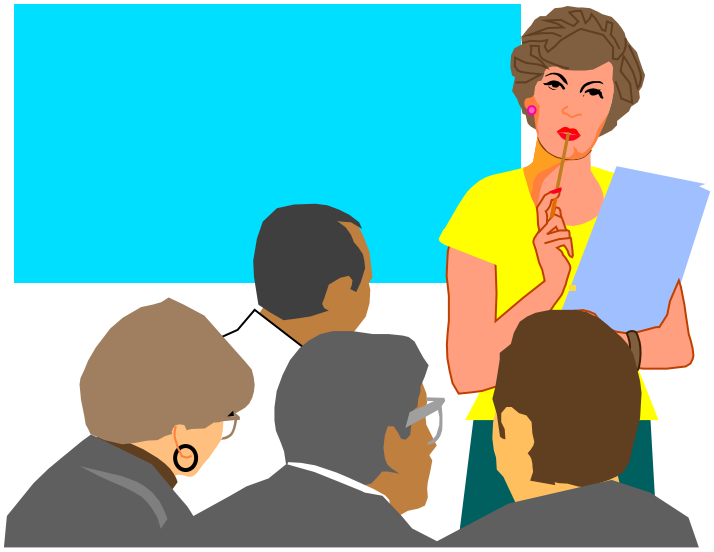
Spring 2025

授課教師：統計系余清祥

2025年4月27日

授課內容：統計分析_Part 2

課程下載：csyue.nccu.edu.tw



如何藉由統計獲取資訊？

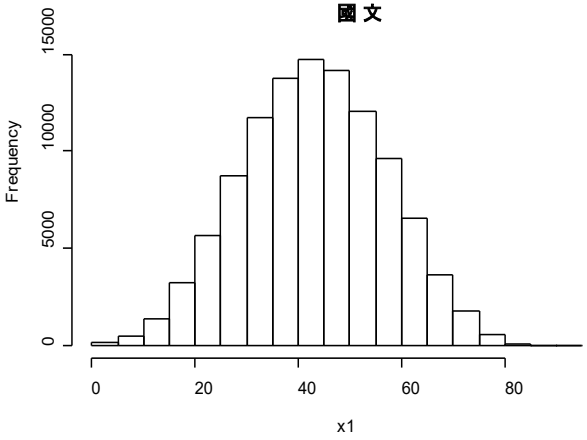
- 如果想瞭解民國94年指定考試各科的特性，可以藉助哪些工具？

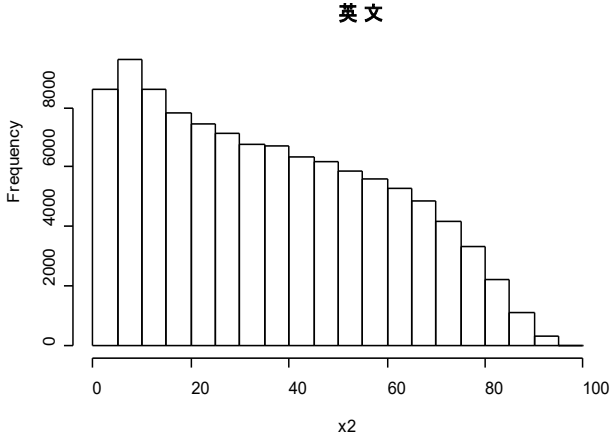
→例如：那一科的分數最不平均(哪一科大多數人都考得不好，只有少數人分數分高)。

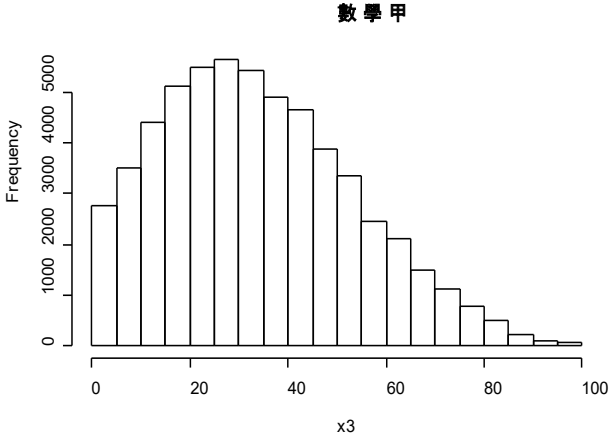
→平均數明顯大於中位數，稱為右偏；若平均數明顯小於中位數，稱為右偏。平均數等於中位數，則為兩側對稱。

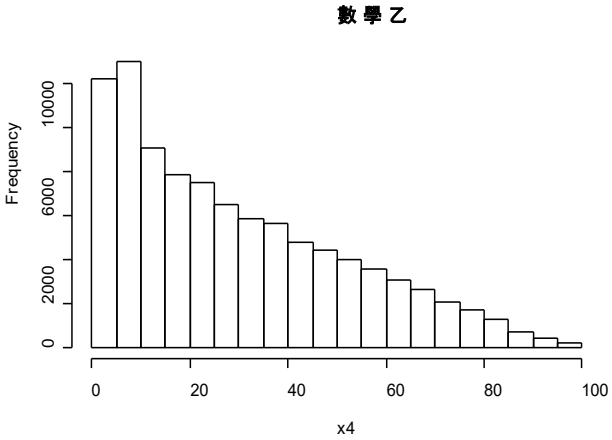
民國 94 年大學指定考試各科成績

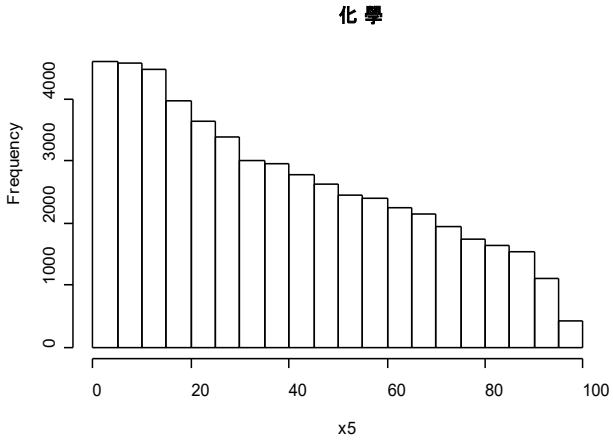
	國文	英文	數學甲	數學乙	化學	物理	生物	歷史	地理
Min.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0	0.00
12%	27.00	8.00	11.00	4.00	8.00	6.00	22.00	13.0	18.00
1st Qu.	34.00	16.00	22.00	12.00	15.00	12.00	32.00	28.0	30.00
Median	44.00	34.00	34.00	29.00	34.00	23.00	45.00	39.0	39.00
Mean	43.56	36.68	36.36	34.36	38.88	28.75	46.16	38.7	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

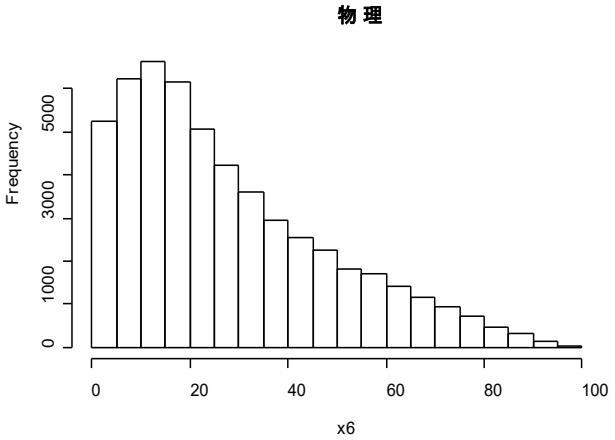

國文
Frequency
0
5000
10000
15000
0
20
40
60
80
x1

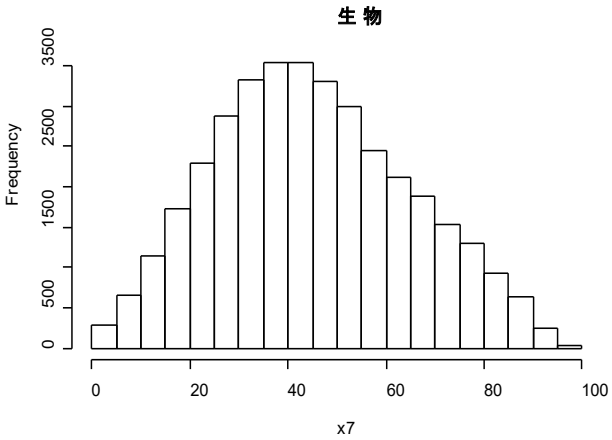

英文
Frequency
0
2000
4000
6000
8000
0
20
40
60
80
100
x2

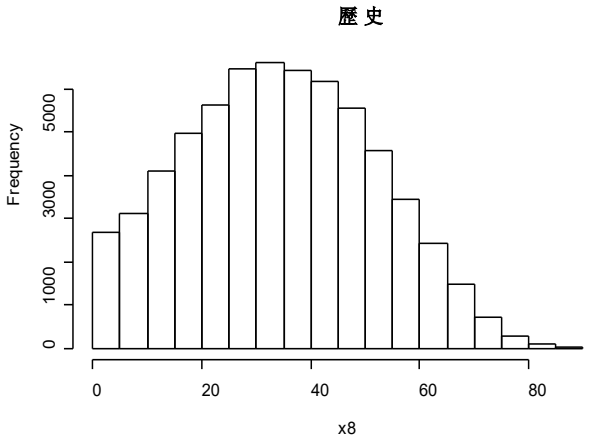

數學甲
Frequency
0
1000
2000
3000
4000
5000
0
20
40
60
80
100
x3

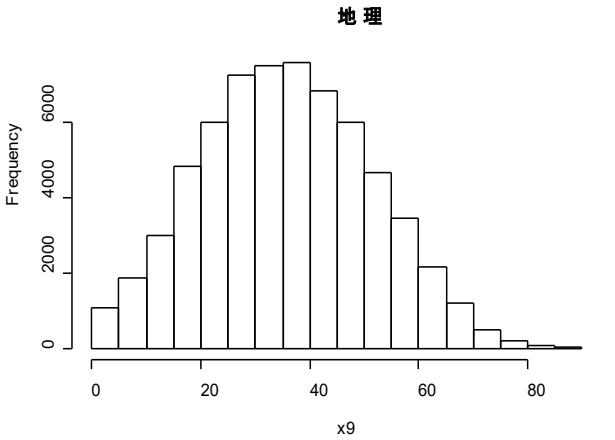

數學乙
Frequency
0
2000
6000
10000
0
20
40
60
80
100
x4

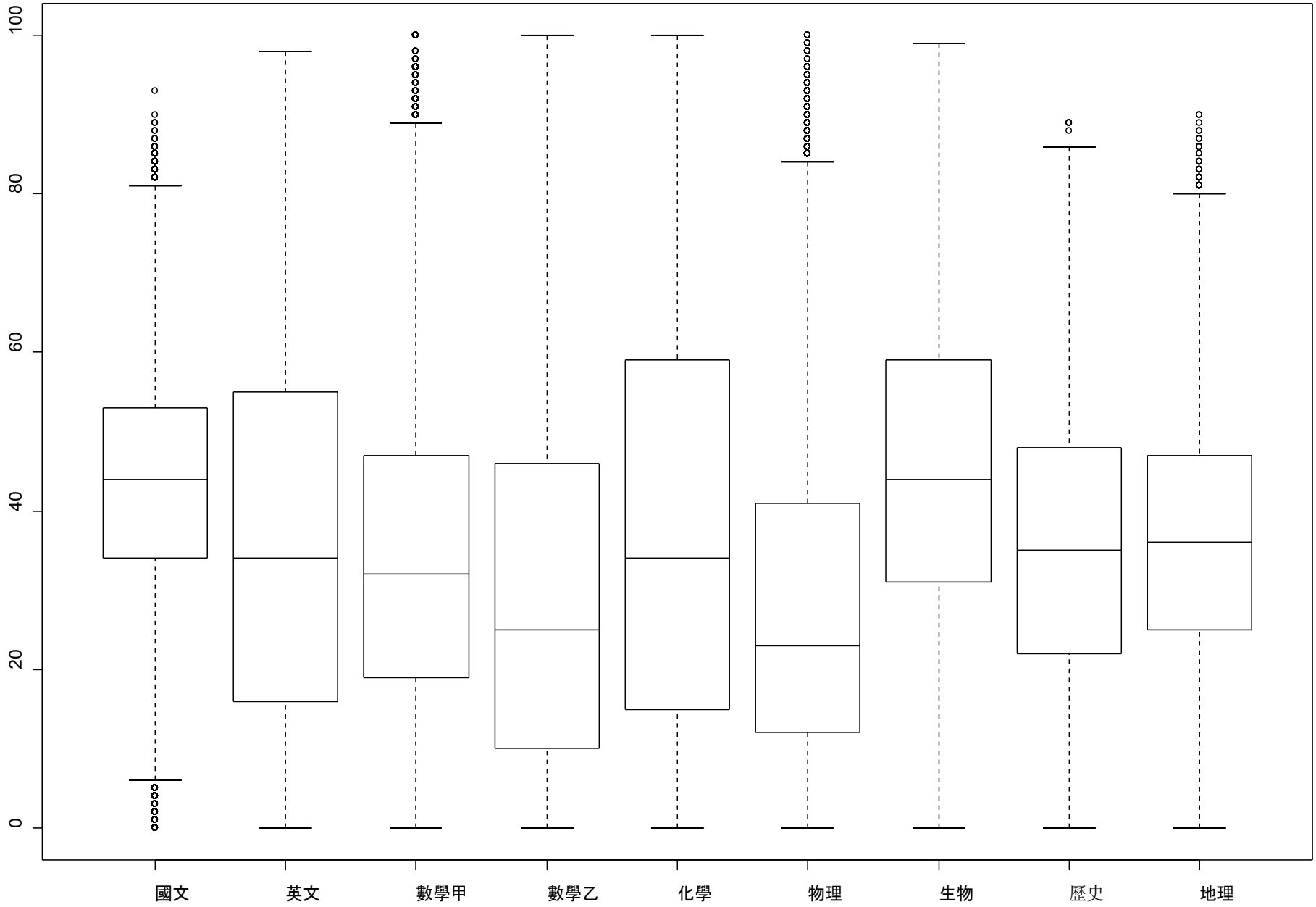

化學
Frequency
0
1000
2000
3000
4000
0
20
40
60
80
100
x5


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

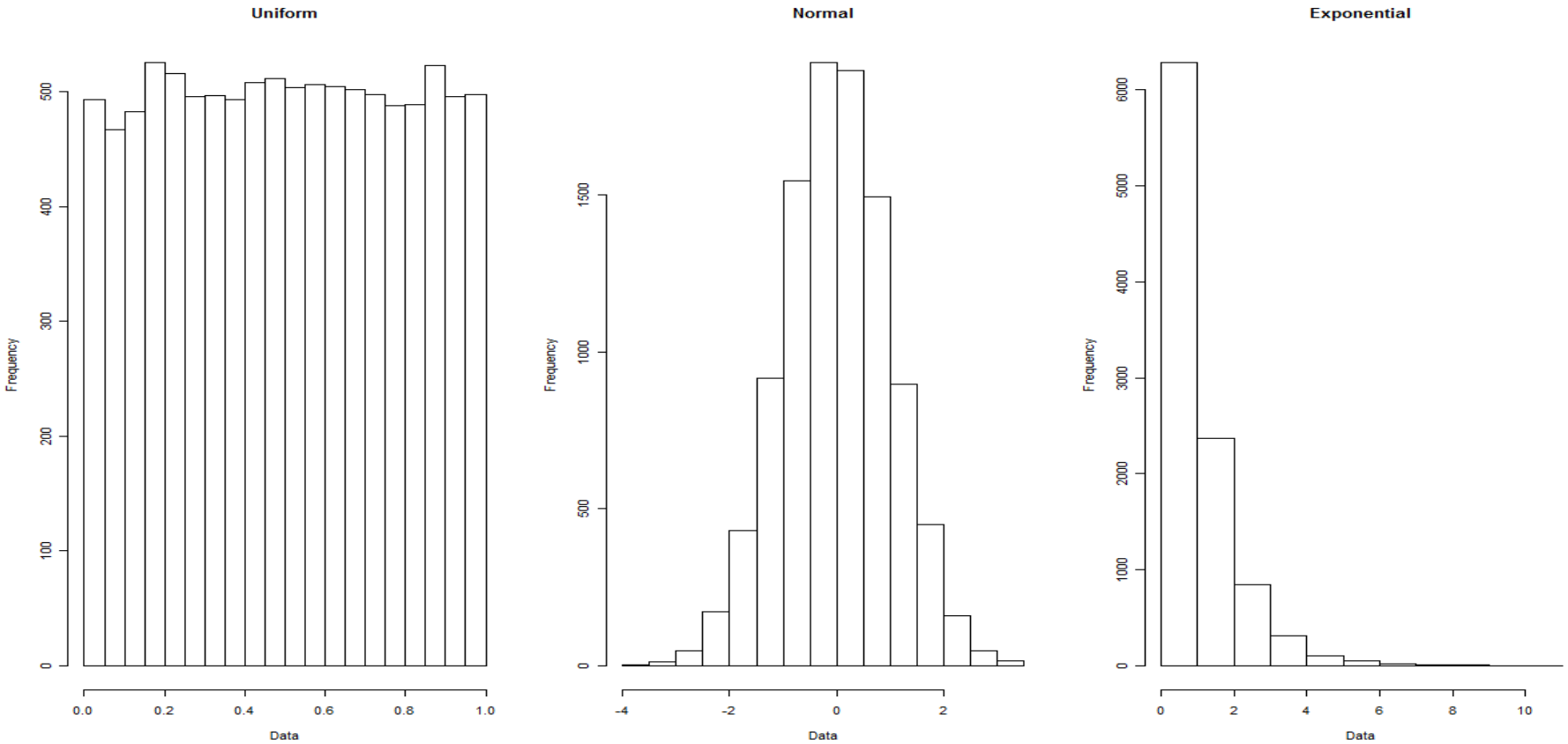

地理
Frequency
0
2000
4000
6000
0
20
40
60
80
x9


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

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。



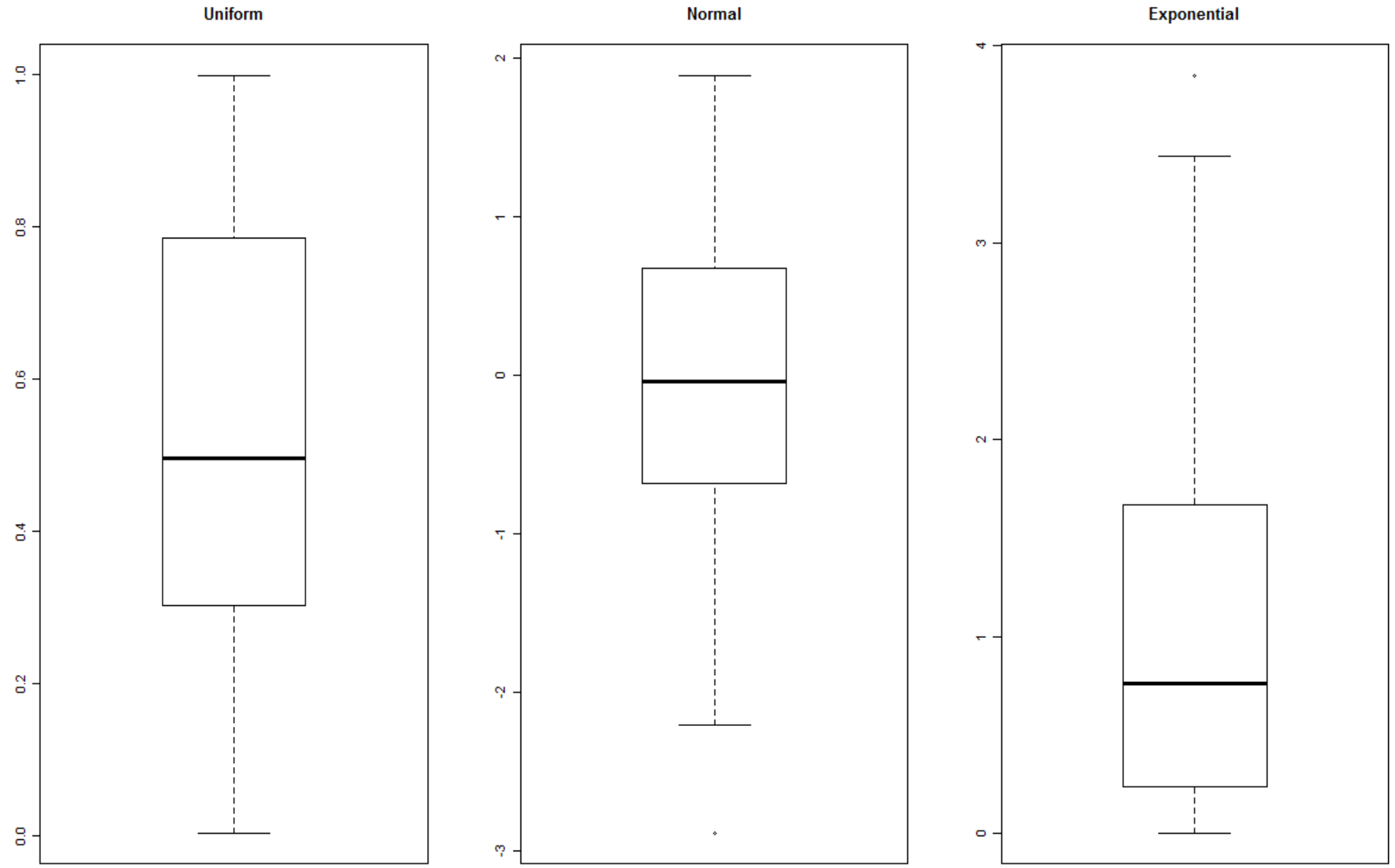
選擇有代表性的統計數據

- 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：

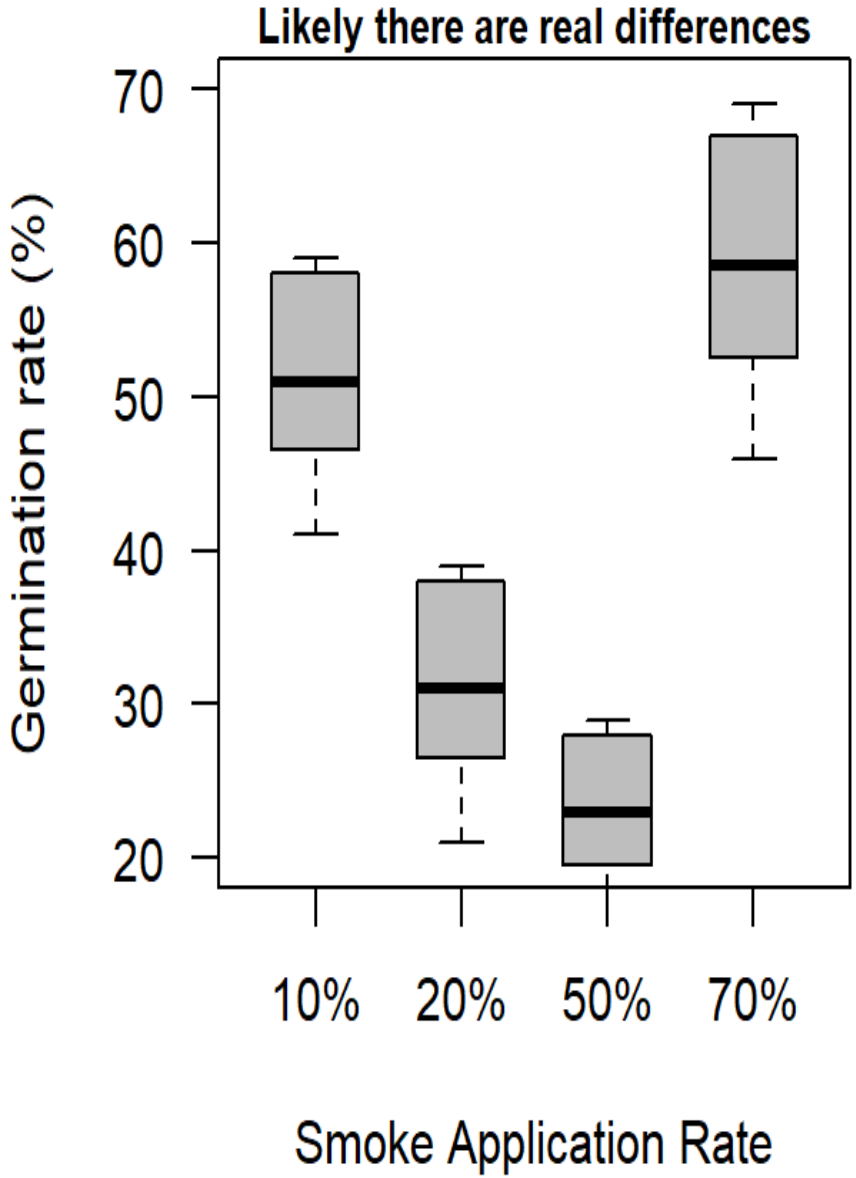
(1) Min. 1st Qu. Median Mean 3rd Qu. Max. St.d.
.0037 .1967 .4577 .4601 .6920 .9707 .2821

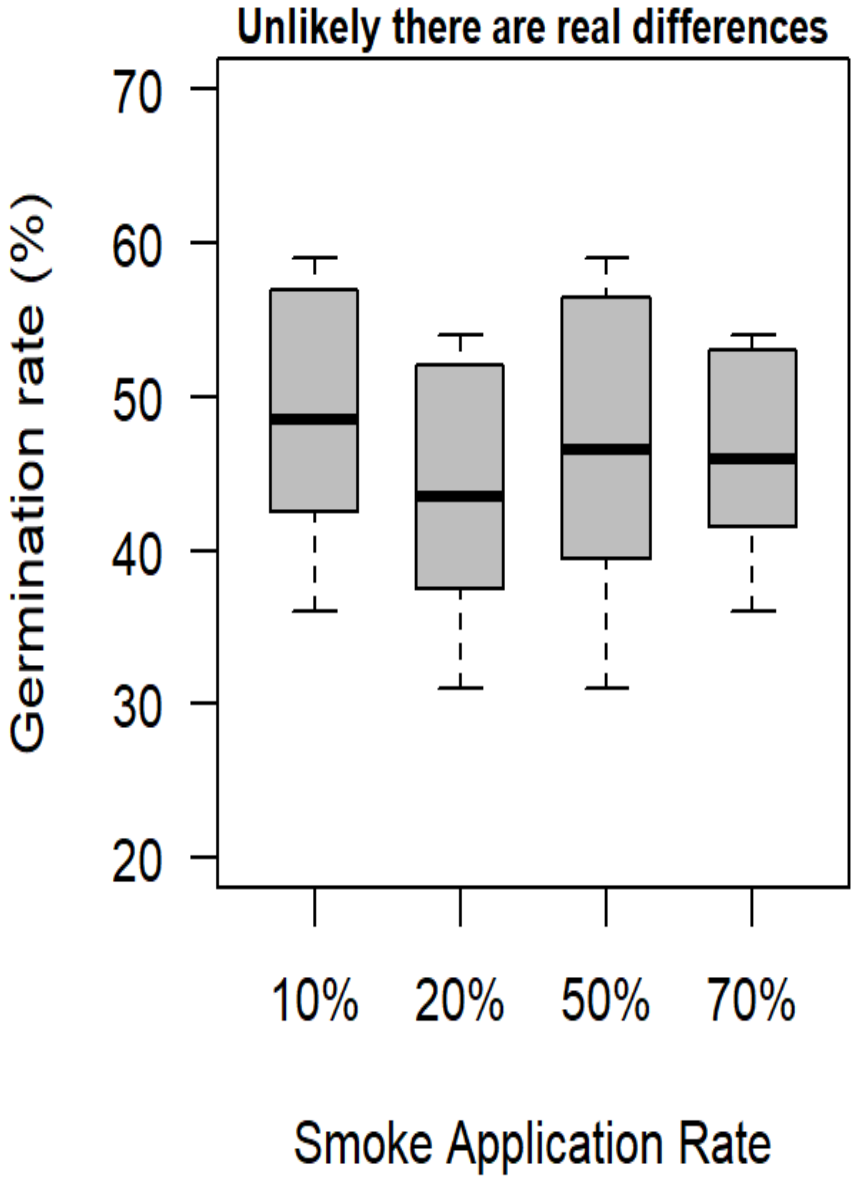
(2) Min. 1st Qu. Median Mean 3rd Qu. Max. St.d.
-2.2060 -.4512 .1167 .1298 .9329 2.2570 .9507

(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」。

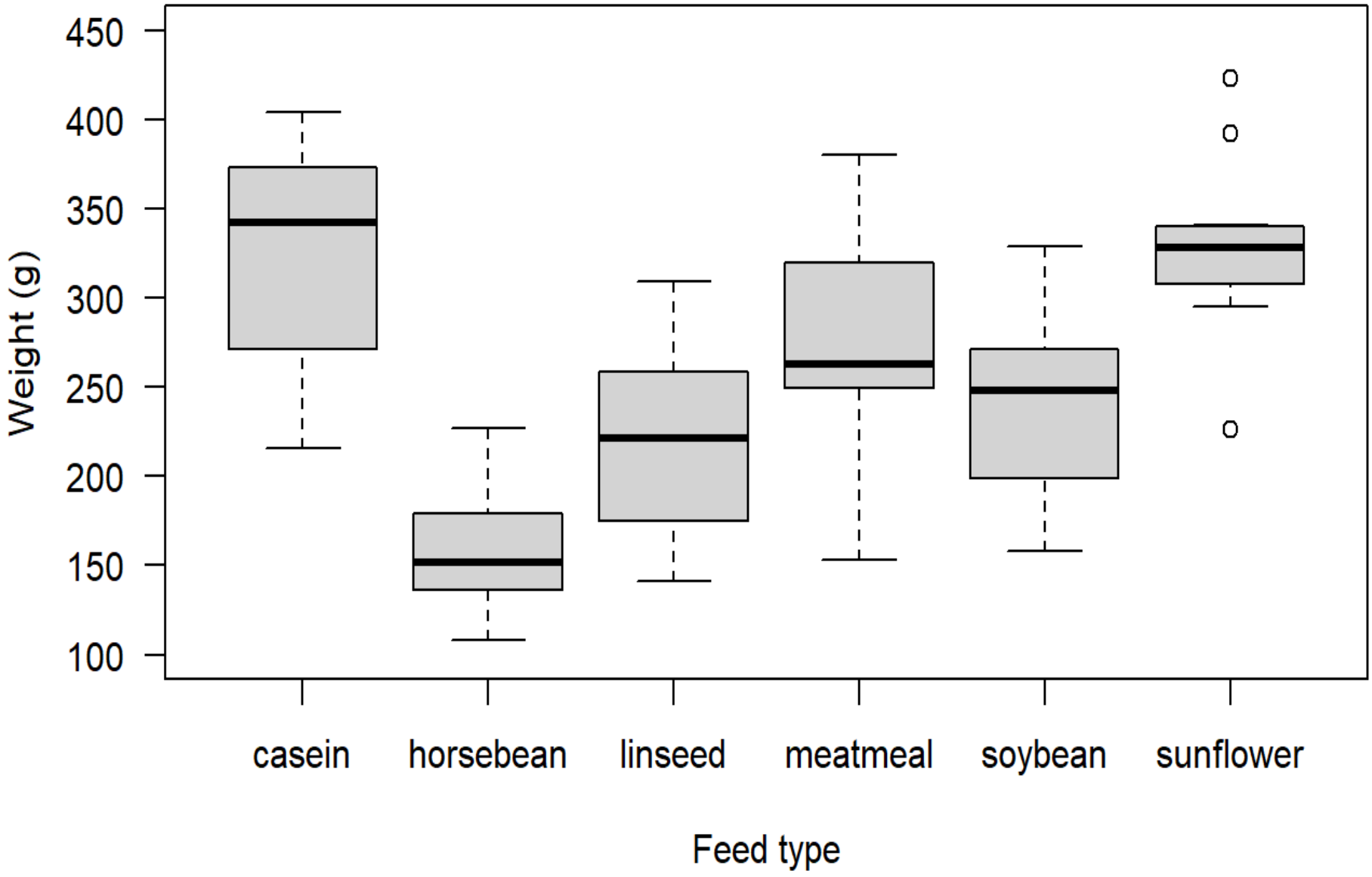
Boxplot can detect between variations!





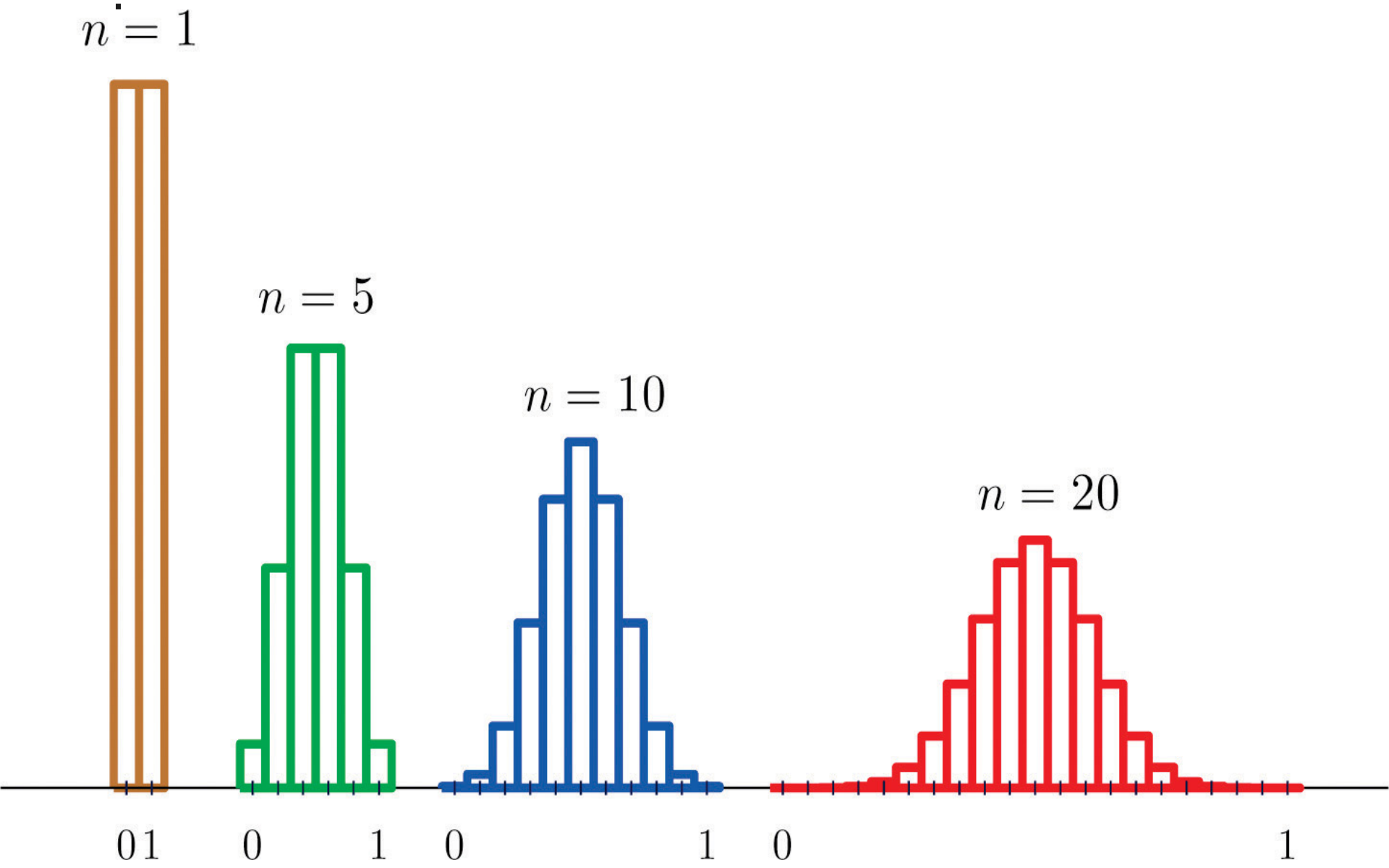
https://saeappliedstats.tech/analysis-of-variance.html

R data “chickwts”, weights of chickens fed



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

抽樣分配(Sampling Distribution)



抽樣分配

- 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。

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

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

中央極限定理

- 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。

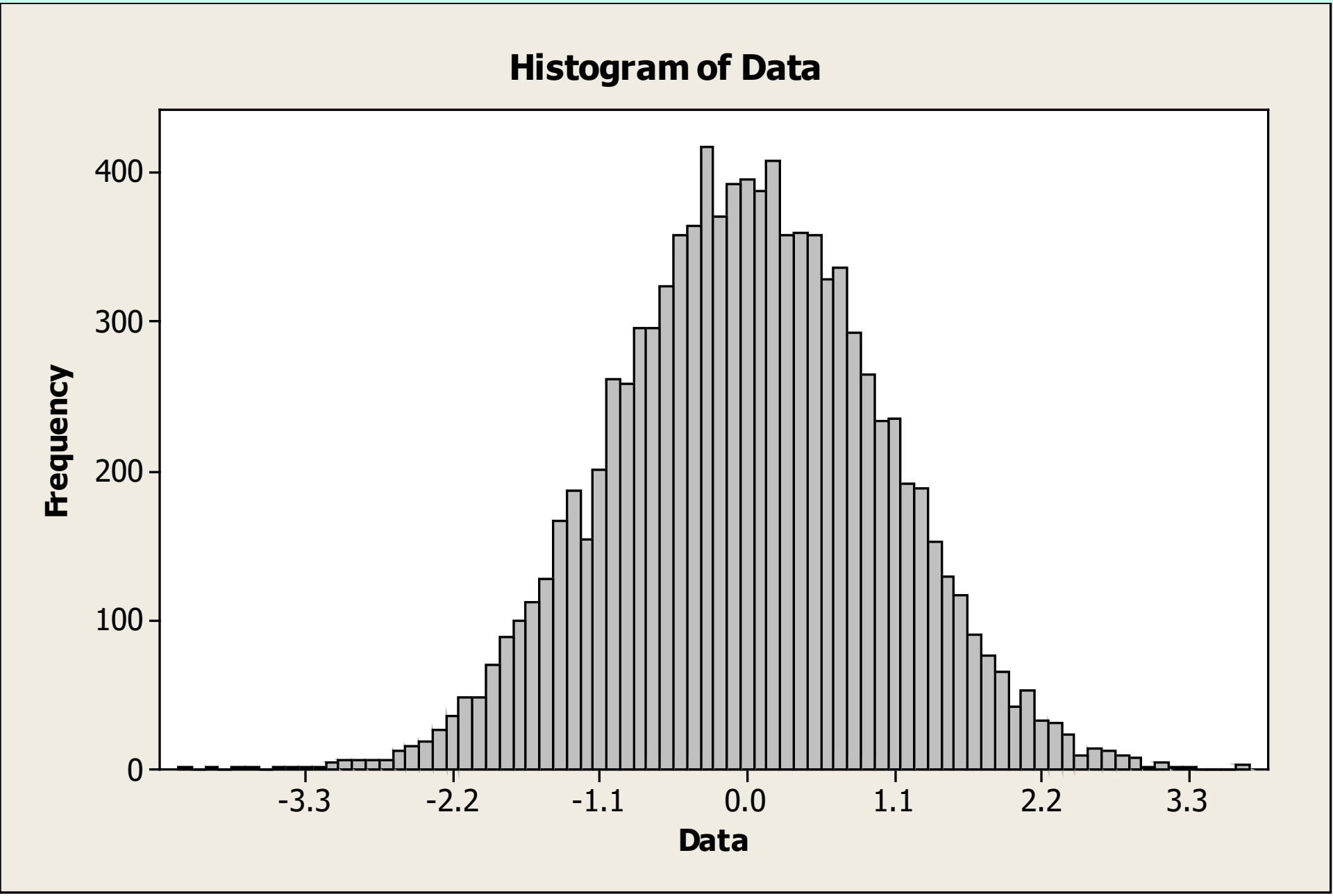
什麼是中央極限定理？

- 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：

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

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。

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

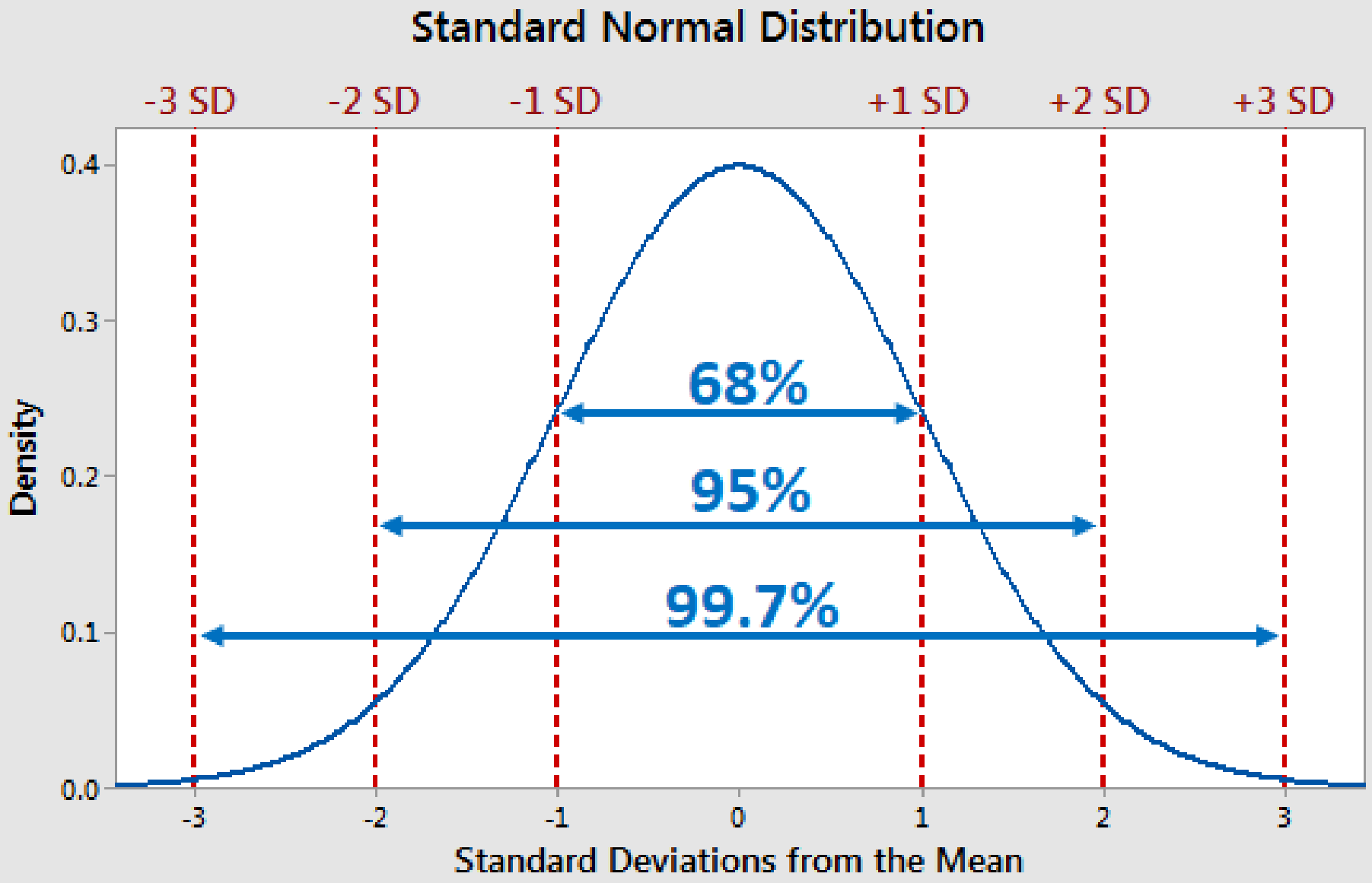
經驗法則: 若為鐘型分配資料

- 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 $(\mu-3\sigma,\mu+3\sigma)$

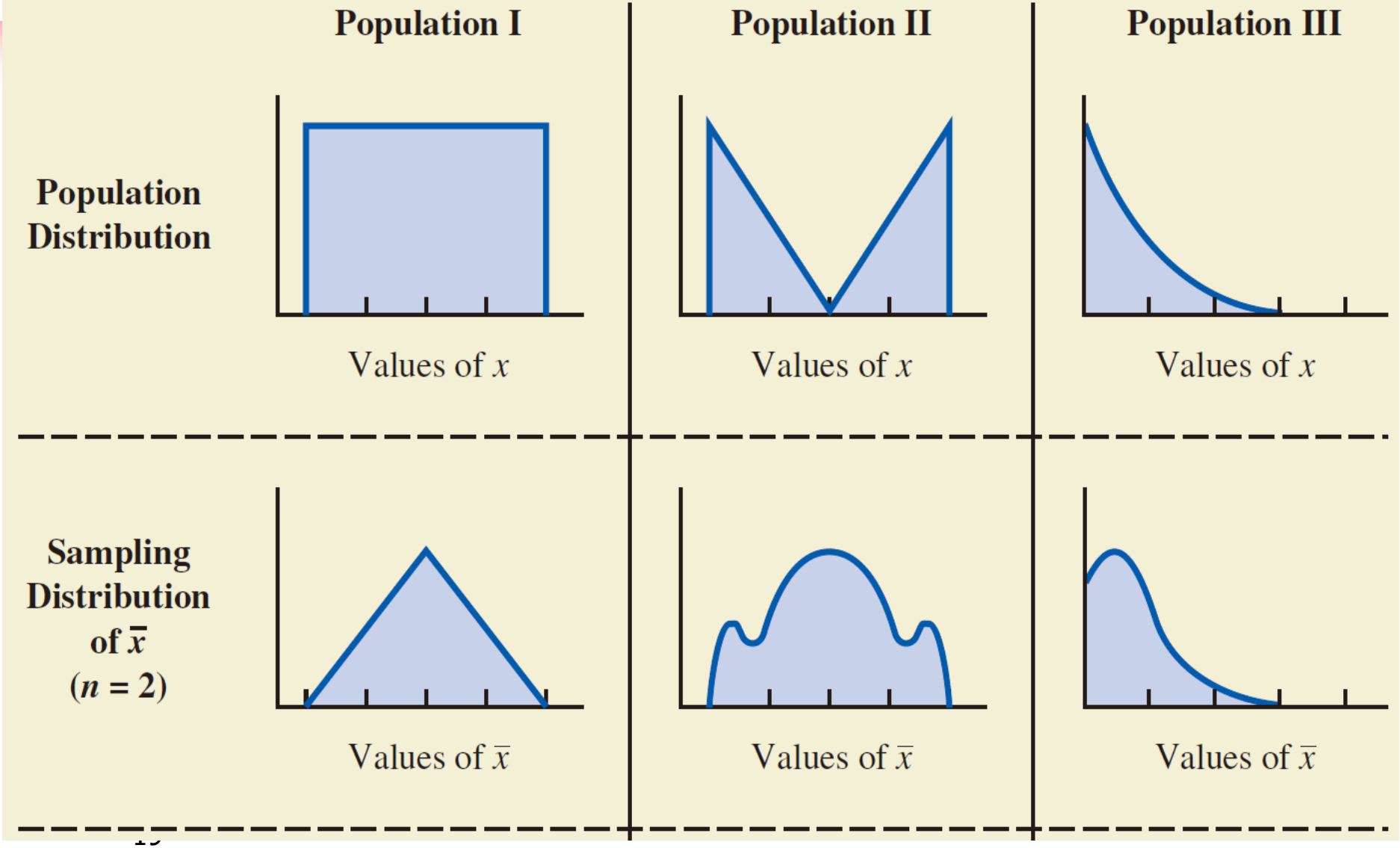
- 註: 1. 以上各母體參數可用樣本數值代替，例如 $s^2 \to \sigma^2, \bar{x} \to \mu$

2.也有人用全距/4或全距/6代替s。

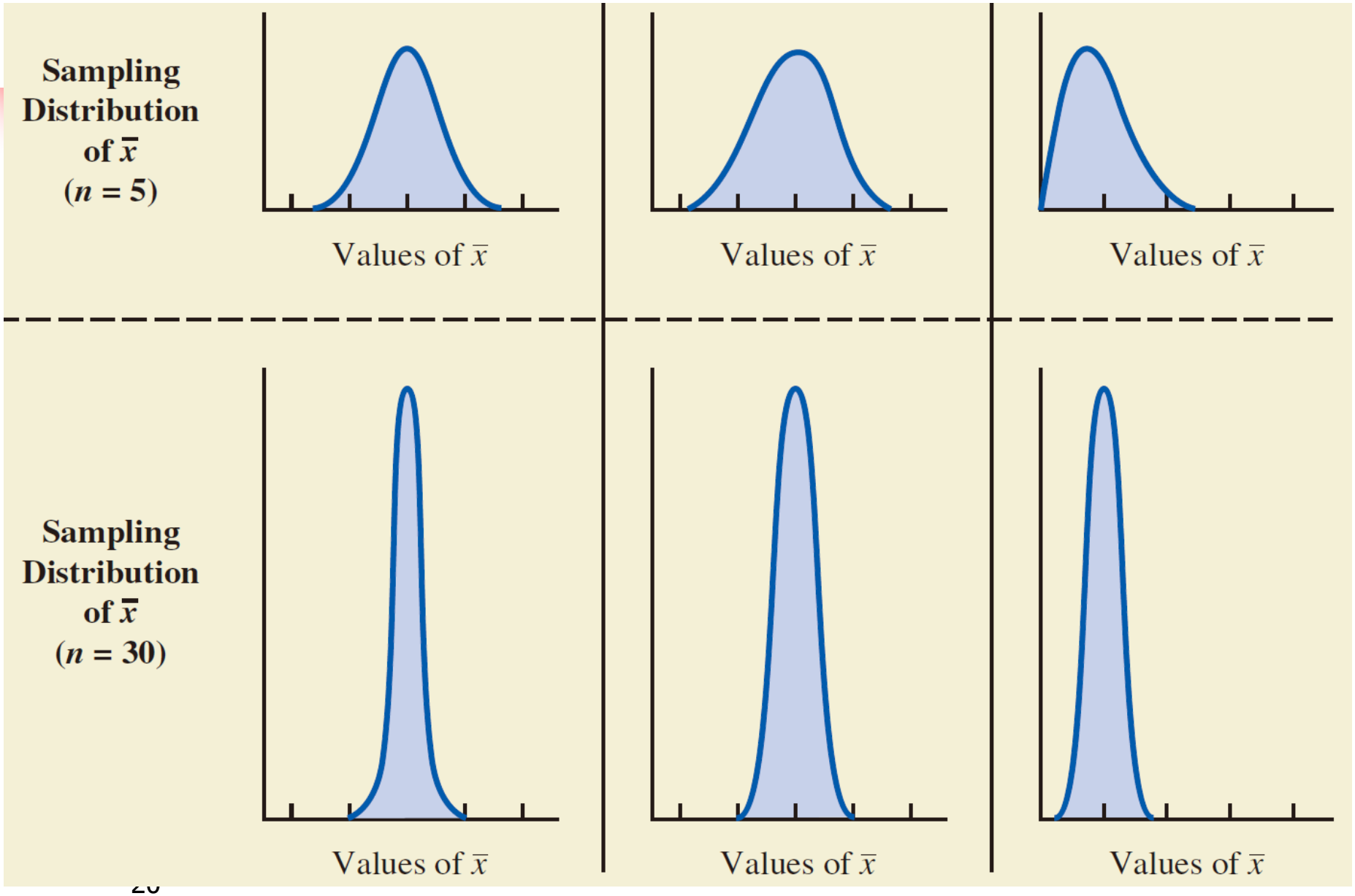
什麼是經驗法則？


Standard Normal Distribution
-3 SD
-2 SD
-1 SD
+1 SD
+2 SD
+3 SD
68%
95%
99.7%
Density
0.4
0.3
0.2
0.1
0.0
-3
-2
-1
0
1
2
3
Standard Deviations from the Mean

Central Limit Theorem



Central Limit Theorem (Conti.)



Detecting Outliers (離群值)

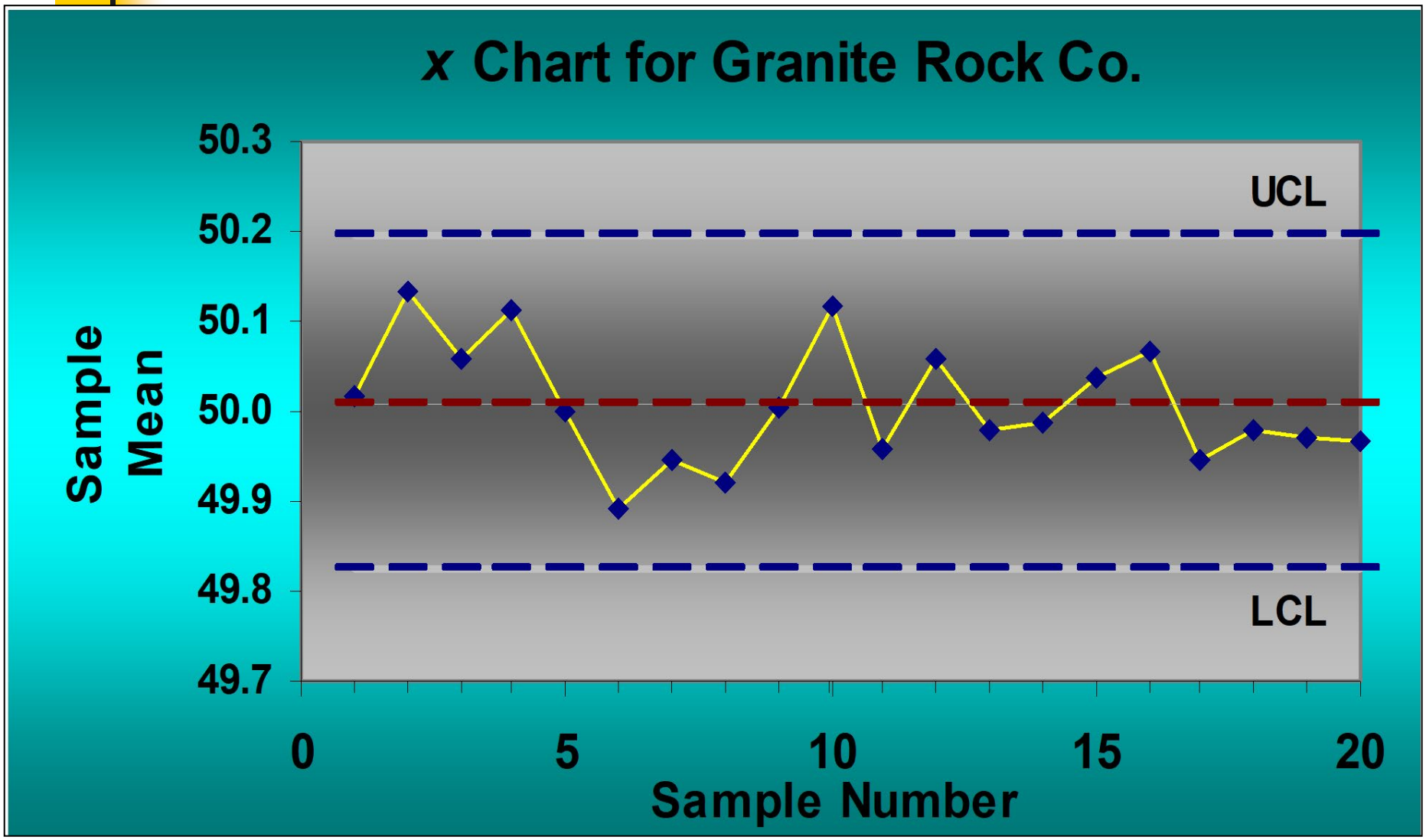
- 離群值是資料中異常大或異常小數值。

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

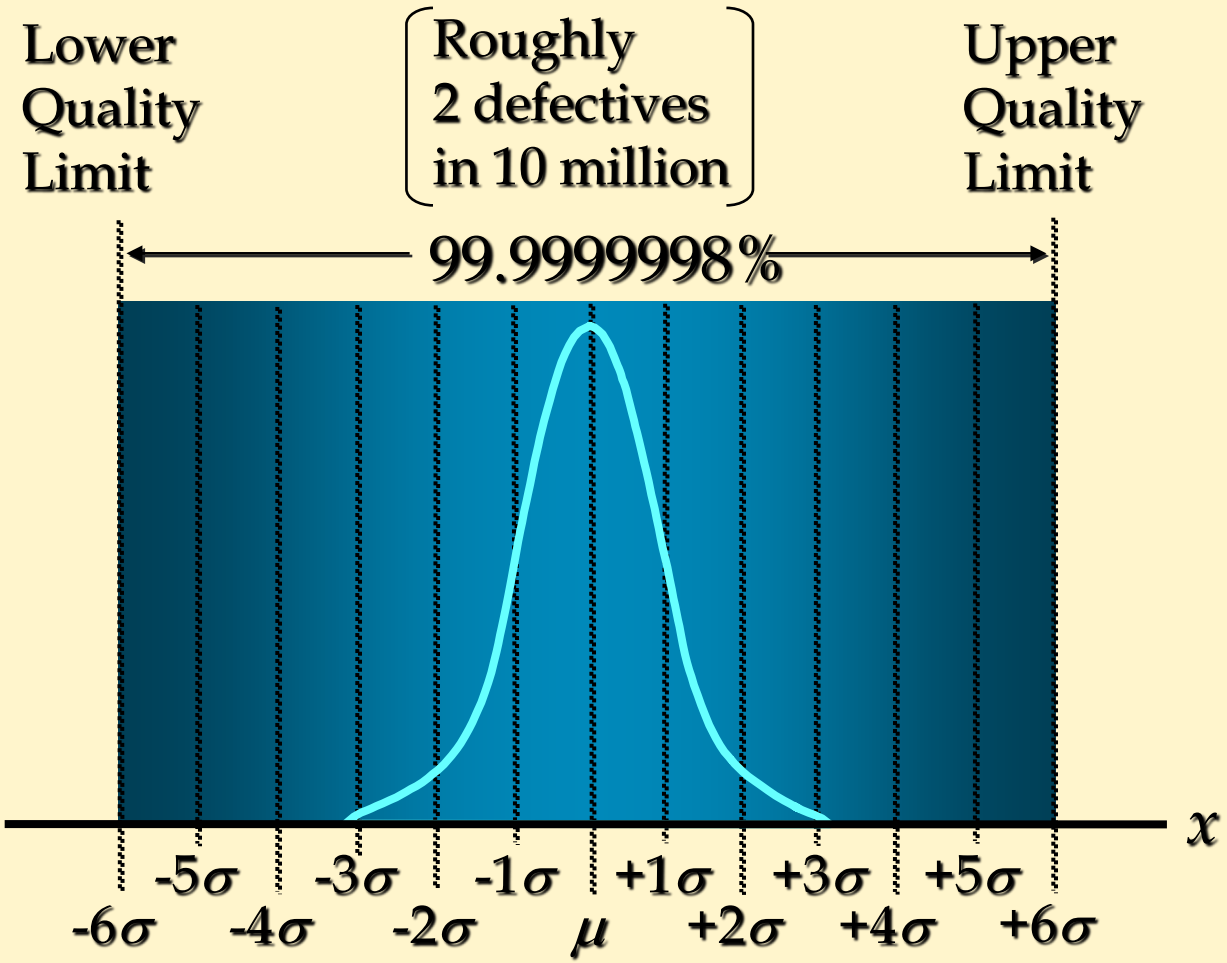
- 離群值的發生原因包括：記錄（書寫）錯誤、或是歸檔錯誤。

→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。

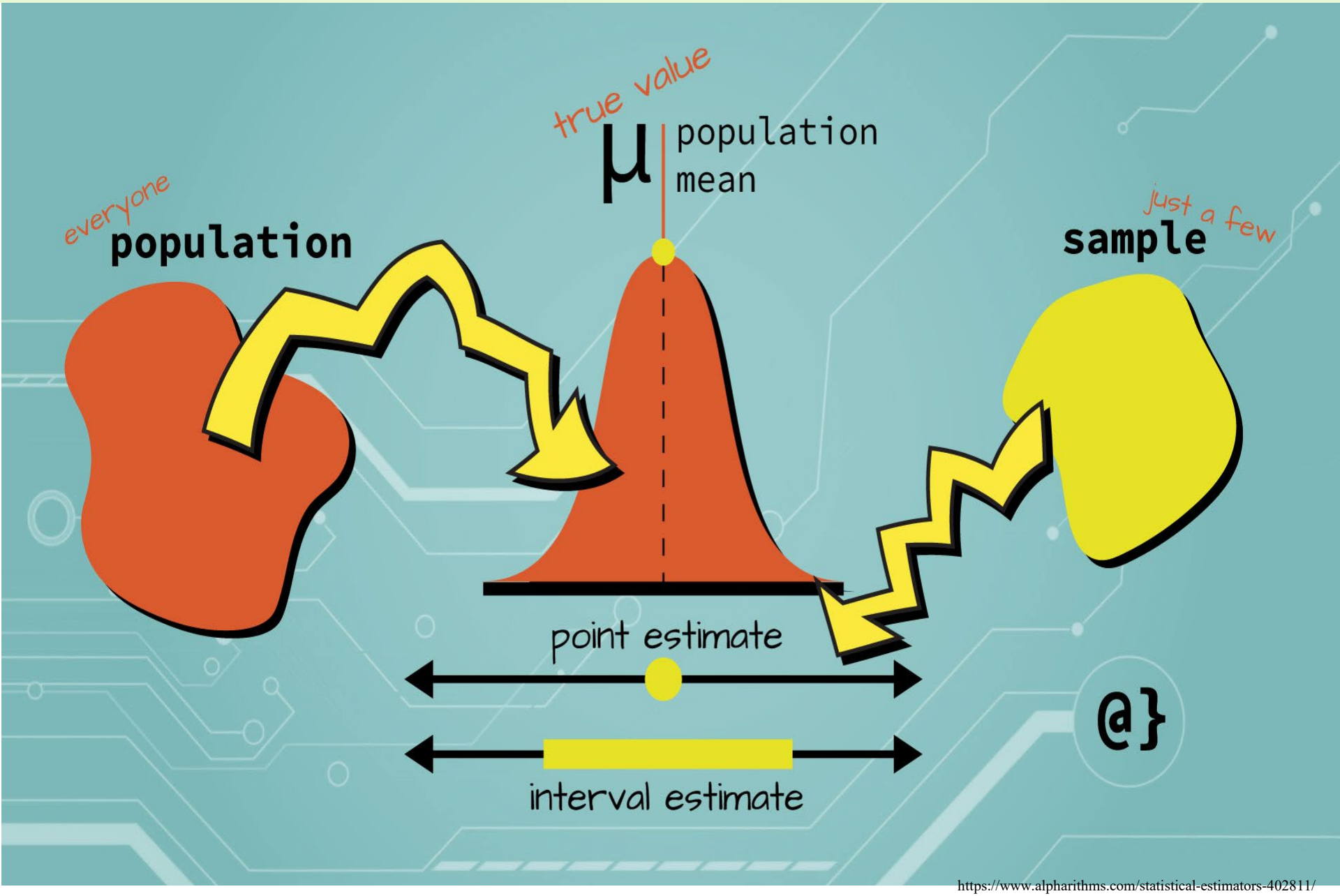
離群值的實例應用


x Chart for Granite Rock Co.
Sample Mean
50.3
50.2
50.1
50.0
49.9
49.8
49.7
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/

參數估計的種類

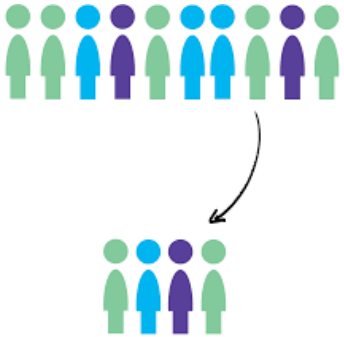
- 母體參數的估計：點估計、區間估計。

→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。



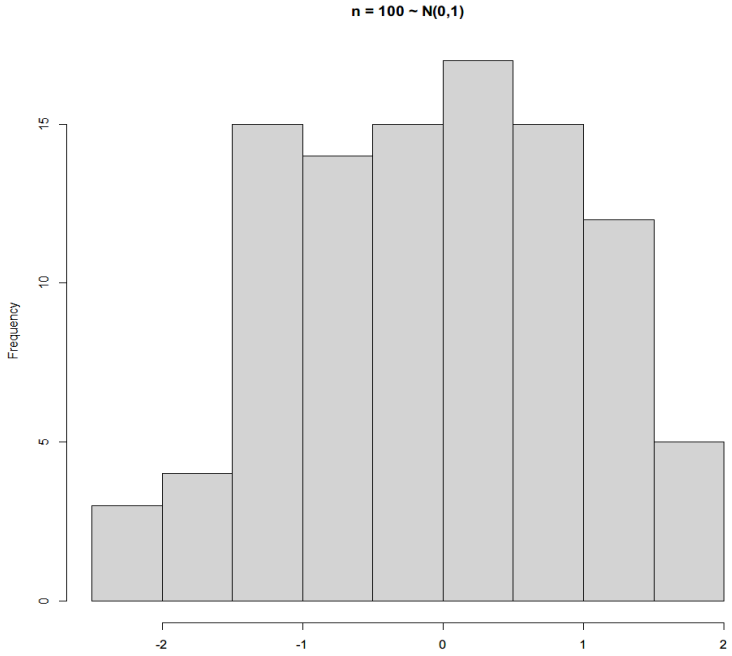
https://www.qualtrics.com/experience-management/research/population-vs-sample/

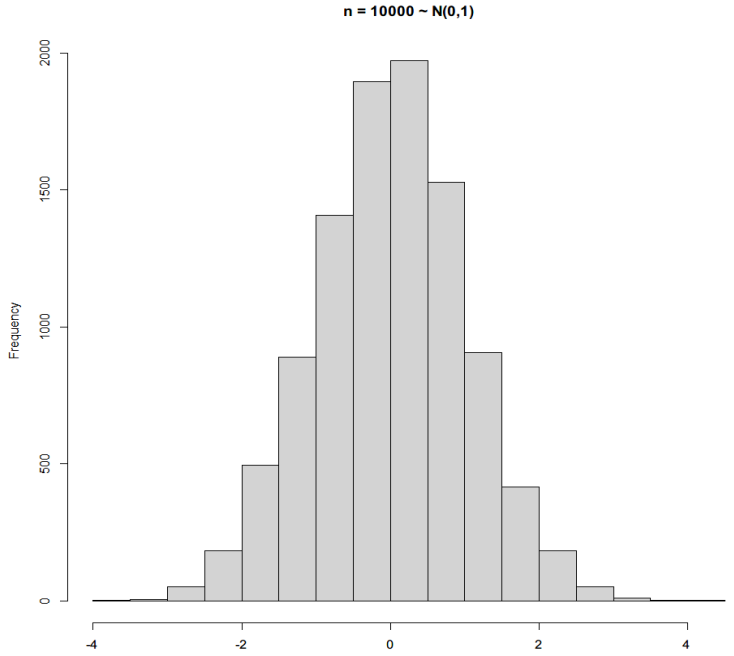
母體與樣本

- 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。

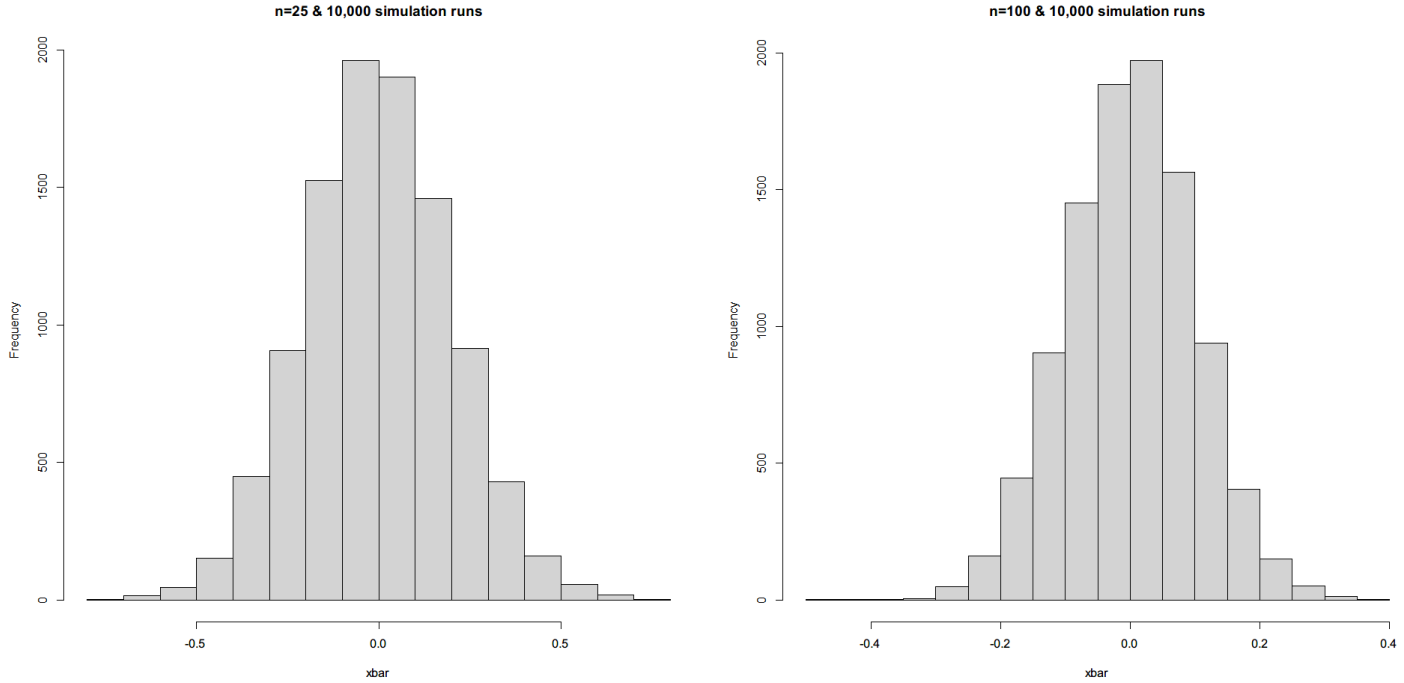




樣本數與估計誤差

- 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)



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。

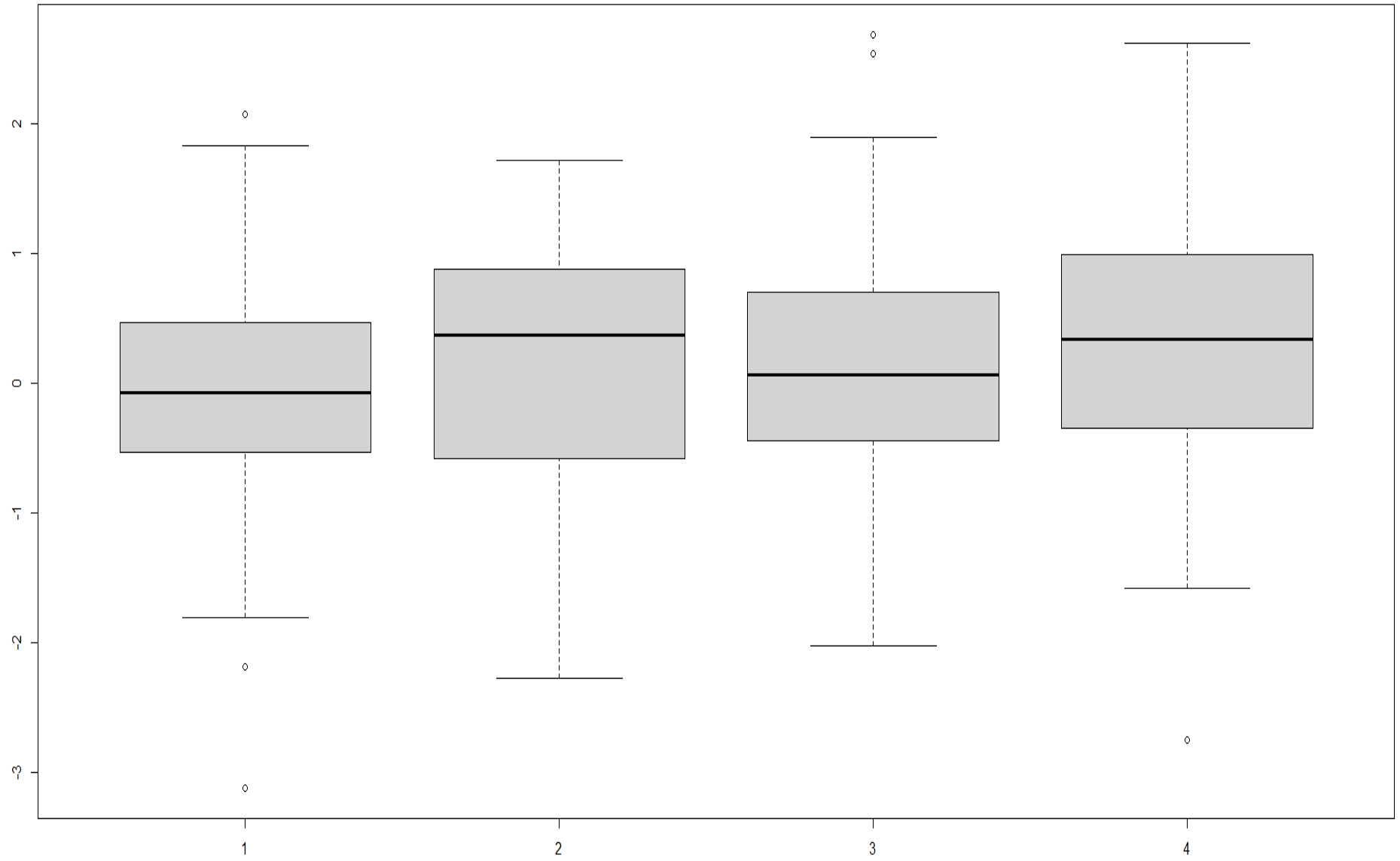
→兩兩 t 檢定的p-value，僅有第一組及第四組有差異（p-value < 0.01）。

- 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

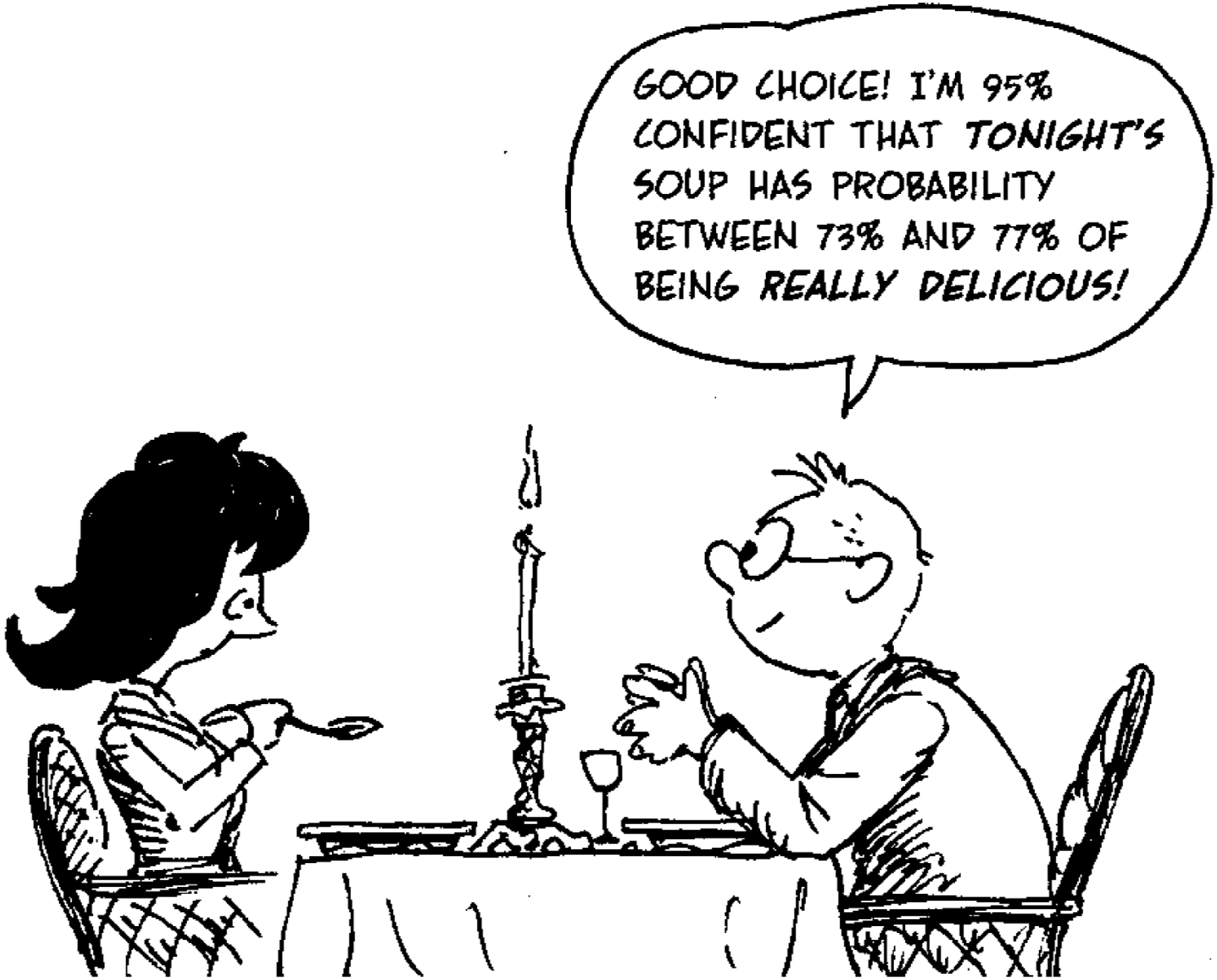


關於信賴區間的幾個疑問

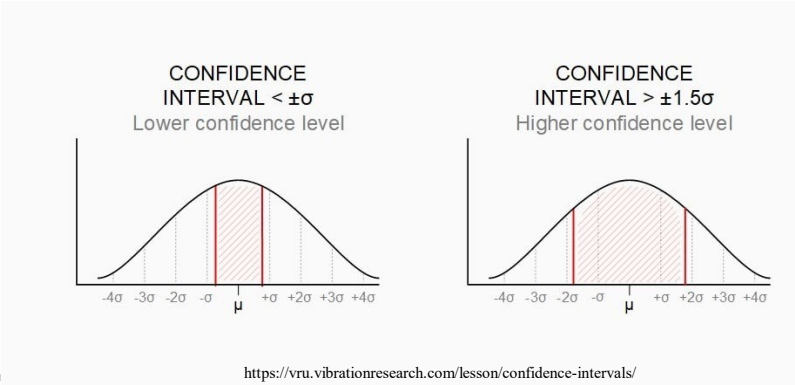
- 95%信賴區間的詮釋：不代表每次建立的信賴區間，有95%的機會涵蓋真實的參數。

→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？


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

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？

→某位選手有三成打擊率，亦即每次上場打擊有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。)

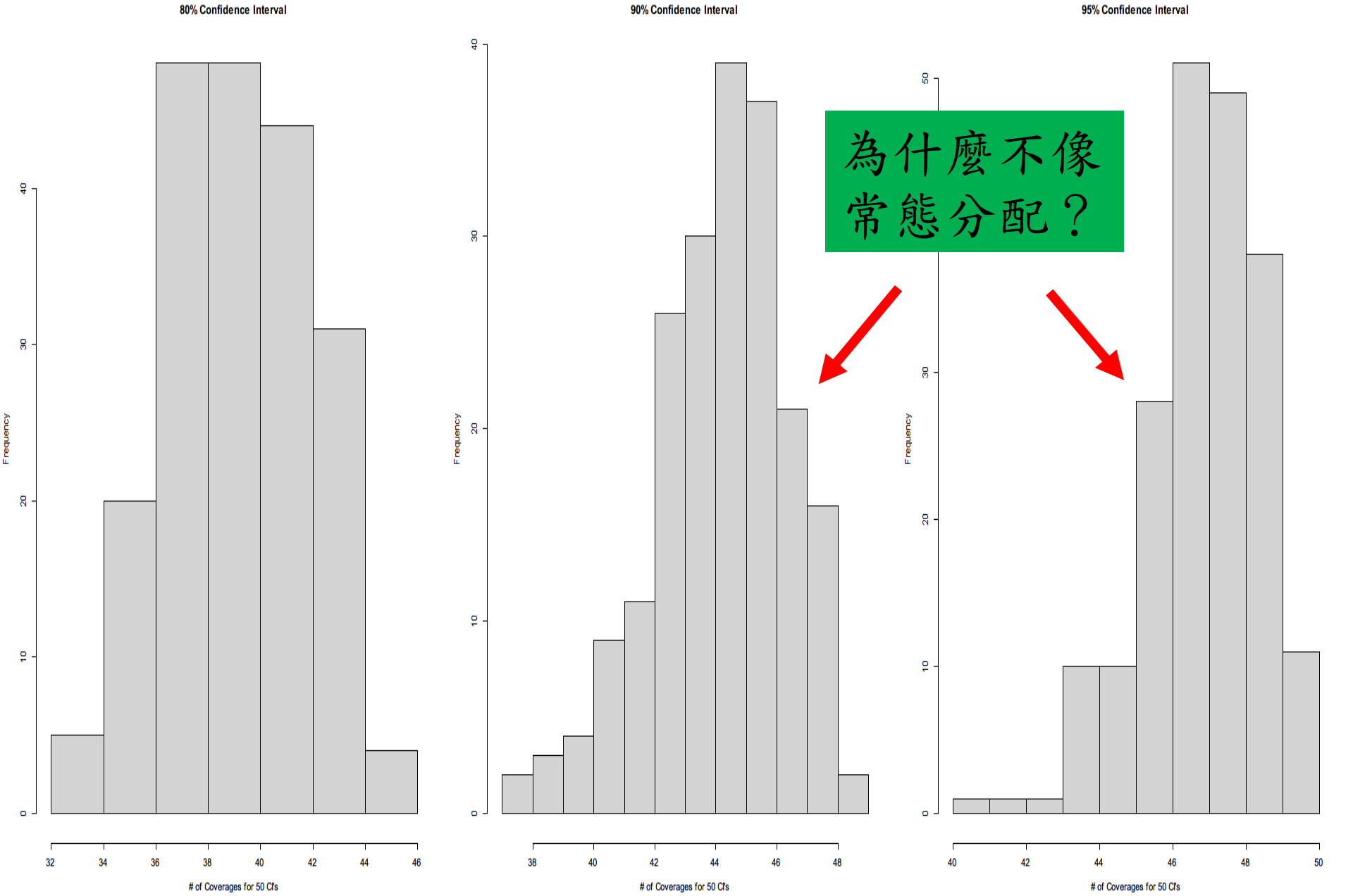
如何詮釋信心係數？

- 信賴區間為**程序機率**，不用於單次解釋。

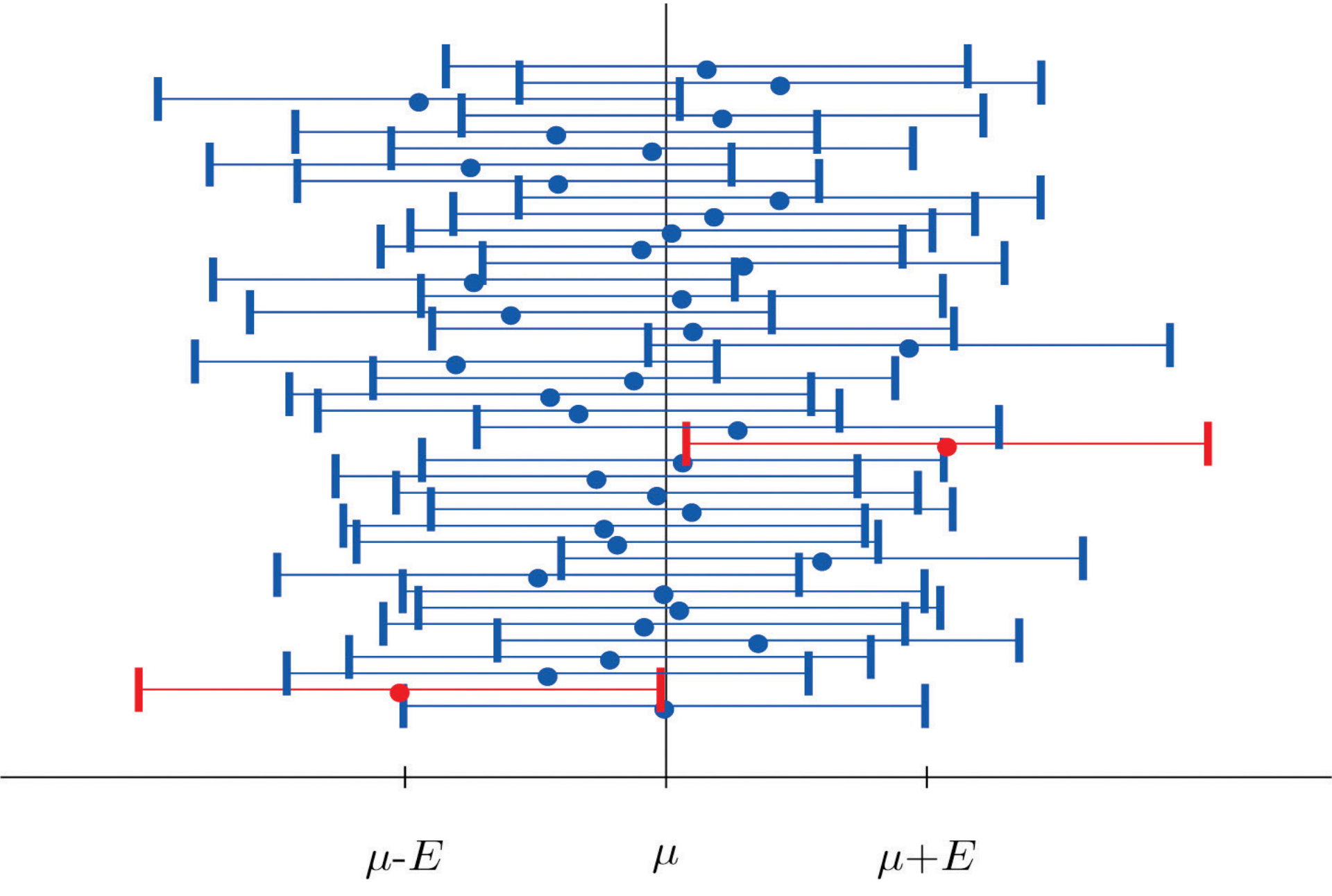
→一旦取得樣本資料，點估計、區間估計就已確定，沒有所謂的機率、或是信心水準。

- 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)



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
24
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。

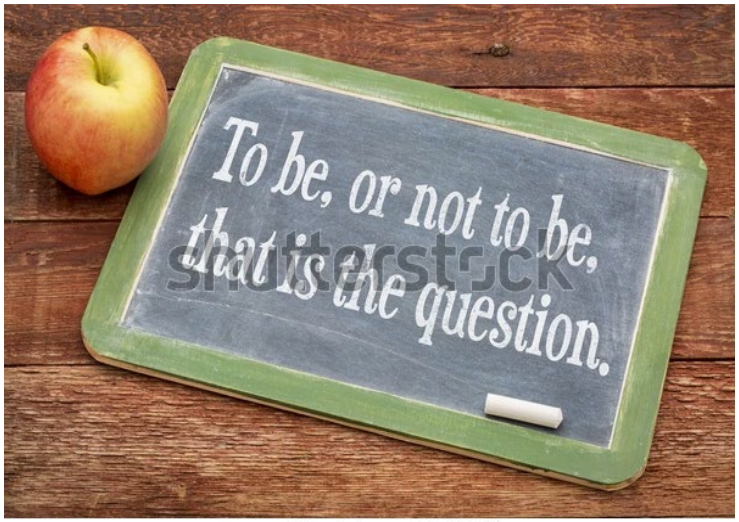
→p值可視為在虛無假設下，樣本觀察值發生的機率。**（註：機率愈小、虛無假設愈不可能）**

- 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

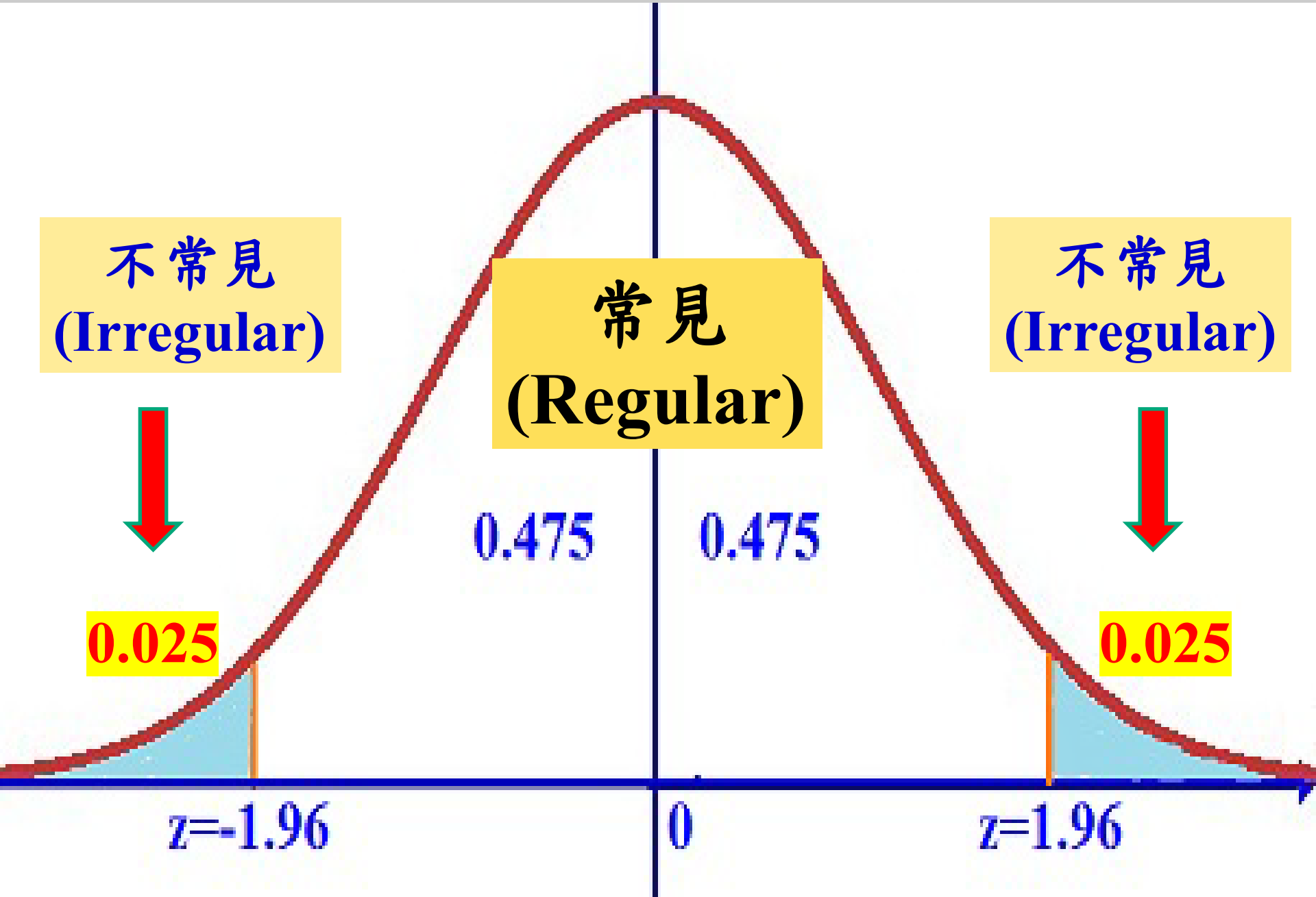
- Less than 0.01: Overwhelming evidence to reject H_0
- Between 0.01~ 0.05: Strong evidence to reject H_0
- Between 0.05 ~ 0.10: Weak evidence to reject H_0
- Greater than 0.10: Insufficient evidence to reject H_a



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

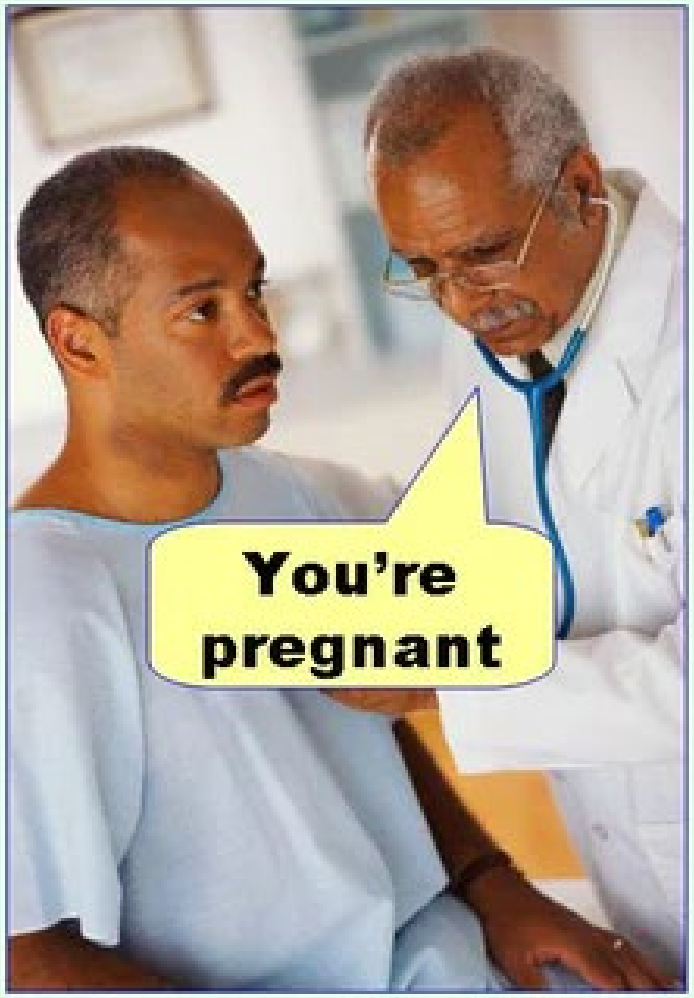
- Because hypothesis tests are based on sample data, we must allow for the possibility of errors.
- A Type I error is rejecting H_0 when it is true.
- The probability of making a Type I error when the null hypothesis is true as an equality is called the level of significance.
- Hypothesis testing that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

Type II Error

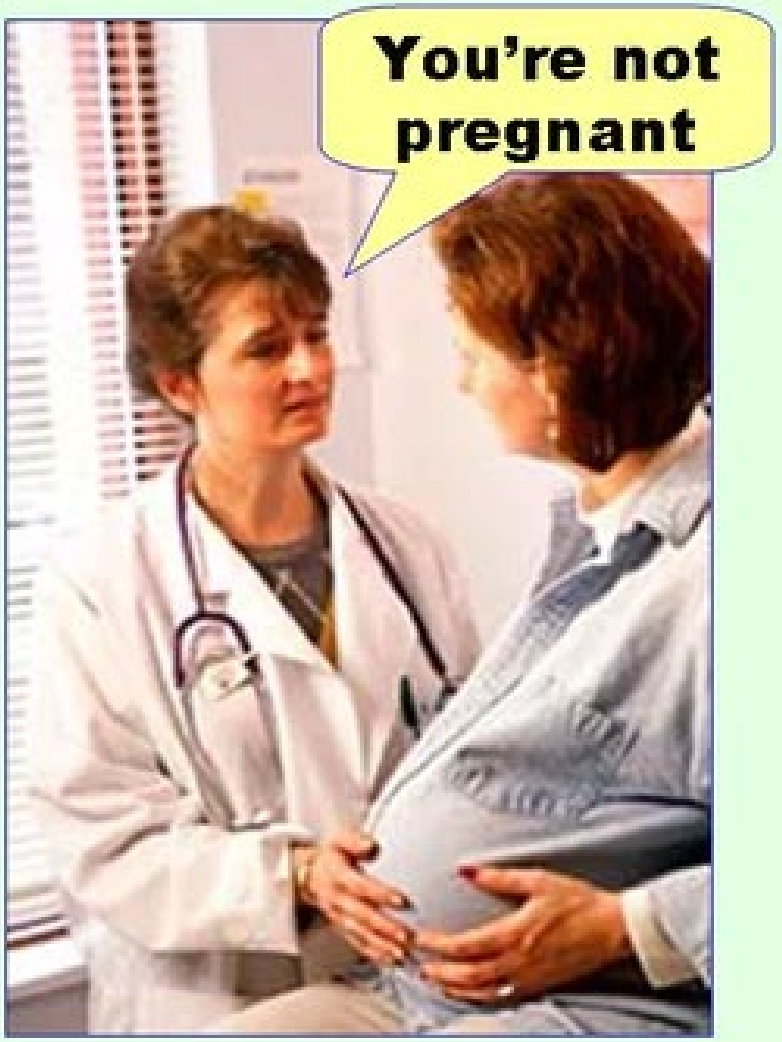
- Type II error: Do not reject H_0 when it is false.
- It is difficult to control for the Type II error.
- Statisticians avoid the risk of making a Type II error by using "**do not reject H_0**" rather than "accept H_0".

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)**