統計實務

Spring 2024

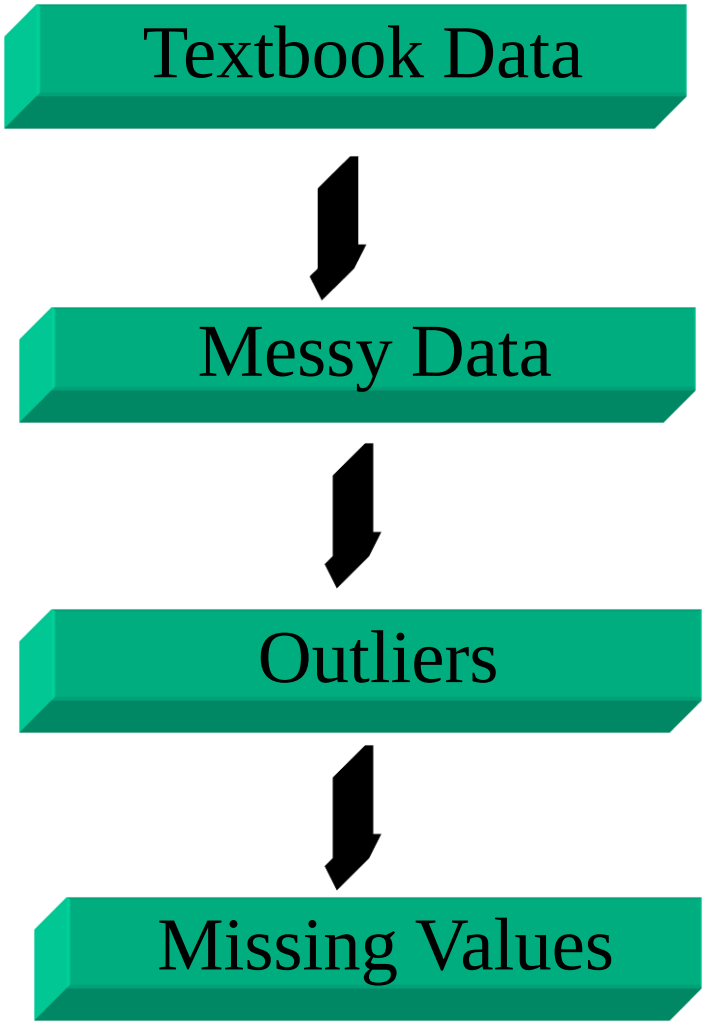
授課教師：統計系余清祥

日期：2024年5月1日

第十一週：探索性資料分析

資料分析的難易度

- 以分析難易度區分，資料大致可分為四種不同層次：（勿隨意刪除離群值！）



資料分析的方向

- 視研究目的，分析大致分為兩個角度：

→尋找資料的整體趨勢；

→偵測較為異常的現象。

- 舉例而言：

→整體趨勢包括平均數、變異數、相關係數等，能反映整筆資料特質的數值。

→異常現象包括異常觀察值（如：離群值）、整體特性的改變、資料是否同質等。

資料分析與研究目的

- 在此以兩個案例示範可能的分析方向：

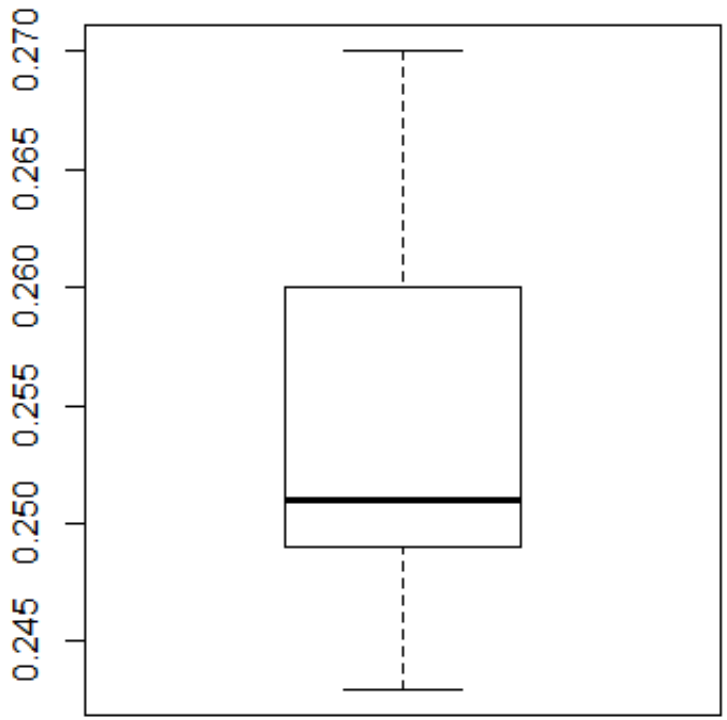
1. 建立迴歸模型（y vs. x's）

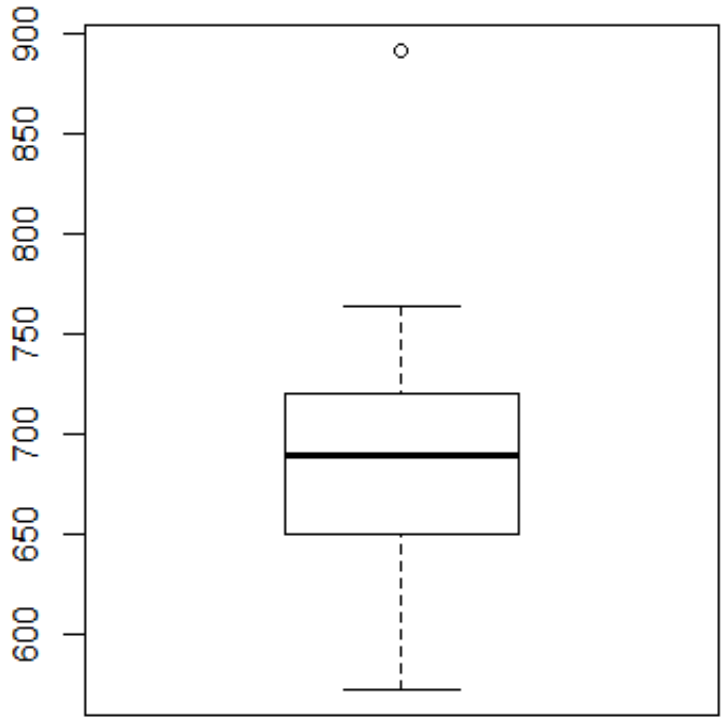
→目標在於與目標變數有關的解釋變數，可透過散佈圖、相關係數（或表格）等，確定有關連的解釋變數及其函數型態。

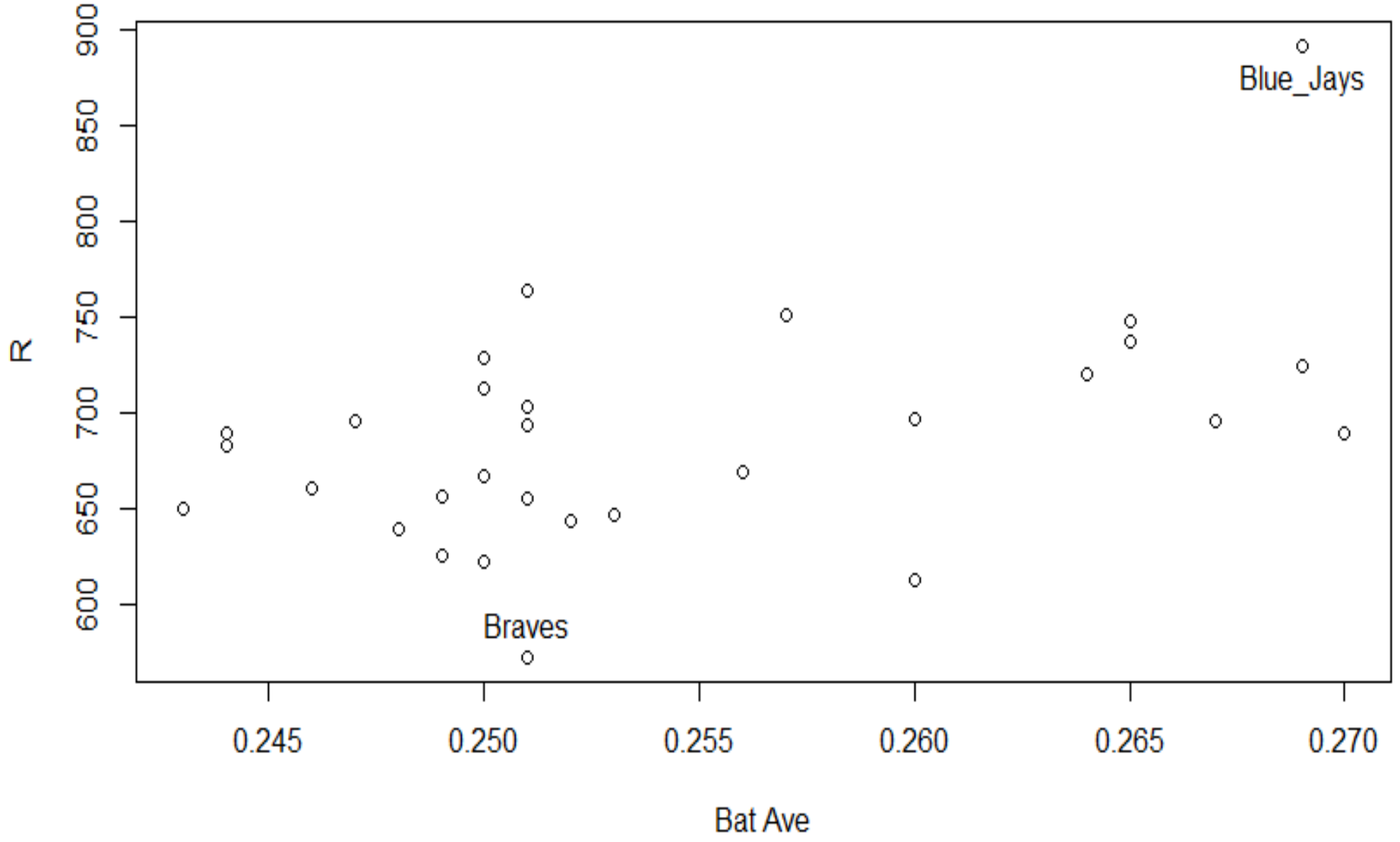
2. 尋找迴歸分析的離群值

→離群值與假設條件有關(Outliers are model based!)，需先決定離群值的定義，例如：Influential outliers及High leverage points。

分析範例：2015年MLB各隊統計







Q：兩個觀察值是Influential outliers或High leverage points？

Q：哪些因素與是否進入季後賽有關？（91%，防守）

Team	AVG	R	H	HR	2B	3B	RBI	SB	OBP	SLG	ERA	H	BB	K	SV	WHIP
Detroit Tigers	.270	689	1515	151	289	49	660	83	.328	.420	4.64	1491	489	1100	35	1.37
Toronto Blue Jays	.269	891	1480	232	308	17	852	88	.340	.457	3.80	1353	397	1117	34	1.21
Kansas City Royals	.269	724	1497	139	300	42	689	104	.322	.412	3.73	1372	489	1160	56	1.28
San Francisco Giants	.267	696	1486	136	288	39	663	93	.326	.406	3.72	1344	431	1165	41	1.23
Boston Red Sox	.265	748	1496	161	294	33	706	71	.325	.415	4.31	1486	478	1218	40	1.36
Colorado Rockies	.265	737	1479	186	274	49	702	97	.315	.432	5.04	1579	579	1112	36	1.51
Arizona Diamondbacks	.264	720	1494	154	289	48	680	132	.324	.414	4.04	1450	500	1215	44	1.33
Pittsburgh Pirates	.260	697	1462	140	292	27	661	98	.323	.396	3.21	1392	453	1338	54	1.24
Miami Marlins	.260	613	1420	120	236	40	575	112	.310	.384	4.02	1374	508	1152	35	1.32
Texas Rangers	.257	751	1419	172	279	32	707	101	.325	.413	4.24	1459	508	1095	45	1.36
Cleveland Indians	.256	669	1395	141	303	29	640	86	.325	.401	3.67	1274	425	1407	38	1.19
St. Louis Cardinals	.253	647	1386	137	288	39	619	69	.321	.394	2.94	1359	477	1329	62	1.25
Tampa Bay Rays	.252	644	1383	167	278	32	612	87	.314	.406	3.74	1314	477	1355	60	1.23
Washington Nationals	.251	703	1363	177	265	13	665	57	.321	.403	3.62	1366	364	1342	41	1.21
Atlanta Braves	.251	573	1361	100	251	18	548	69	.314	.359	4.41	1462	550	1148	44	1.41
Oakland Athletics	.251	694	1405	146	277	46	661	78	.312	.395	4.14	1402	474	1179	28	1.30
Milwaukee Brewers	.251	655	1378	145	274	34	624	84	.307	.393	4.28	1432	517	1260	40	1.36
New York Yankees	.251	764	1397	212	272	19	737	63	.323	.421	4.05	1417	474	1370	48	1.30
Baltimore Orioles	.250	713	1370	217	246	20	686	44	.307	.421	4.05	1406	483	1233	43	1.32
Los Angeles Dodgers	.250	667	1346	187	263	26	638	59	.326	.413	3.44	1317	395	1396	47	1.18
Chicago White Sox	.250	622	1381	136	260	27	595	68	.306	.380	3.98	1443	474	1359	37	1.32
Houston Astros	.250	729	1363	230	278	26	691	121	.315	.437	3.57	1308	423	1280	39	1.20
Seattle Mariners	.249	656	1379	198	262	22	624	69	.311	.411	4.16	1430	491	1283	45	1.31
Philadelphia Phillies	.249	626	1374	130	272	37	586	88	.303	.382	4.69	1592	488	1153	35	1.45
Cincinnati Reds	.248	640	1382	167	257	27	613	134	.312	.394	4.33	1436	544	1252	35	1.36
Minnesota Twins	.247	696	1349	156	277	44	661	70	.305	.399	4.07	1506	413	1046	45	1.33
Los Angeles Angels	.246	661	1331	176	243	21	621	52	.307	.396	3.94	1355	466	1221	46	1.26
New York Mets	.244	683	1351	177	295	17	654	51	.312	.400	3.43	1341	383	1337	50	1.18
Chicago Cubs	.244	689	1341	171	272	30	657	95	.321	.398	3.36	1276	407	1431	48	1.15
San Diego Padres	.243	650	1324	148	260	36	623	82	.300	.385	4.09	1371	516	1393	41	1.31

註：Rays、Yankee、Angels三個球隊的預測錯誤。

統計分析的方向

- Exploratory Data Analysis (EDA；探索性) vs. Confirmatory Data Analysis (CDA；驗證性)

→Supervised Learning屬於CDA，而Unsupervised Learning屬於EDA。

- 科學研究通常包括「觀察」、「推論」、「驗證」幾個步驟，EDA及CDA分屬於前兩個、第三個步驟。

→EDA通常比較不容易（以及SOP）！

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。

→同時進行其他探索性資料分析(EDA)，一步一步找出資料的重要特性，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？

統計的分析觀點

根據統計觀點，分析有以下兩類：

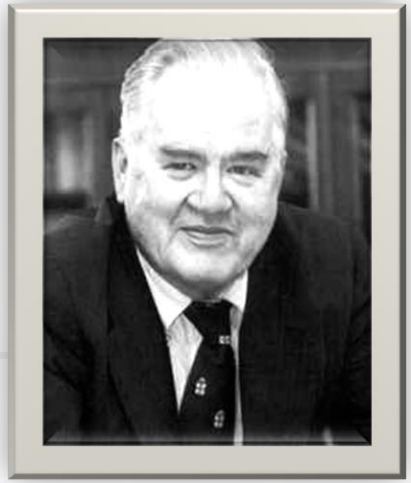
- 探索性資料分析(Exploratory Data Analysis)

→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”

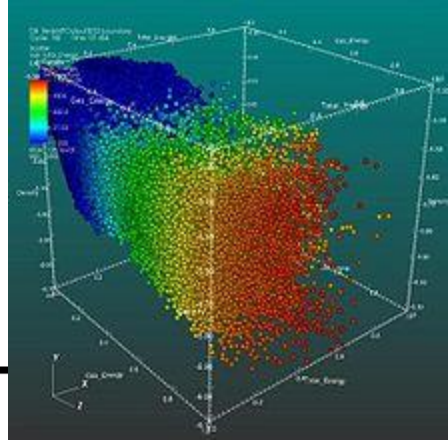
→EDA is not a mere collection of techniques. EDA is a philosophy as to **how we dissect a data set**; **what we look for**; **how we look**; and **how we interpret**.

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

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



WIKIPEDIA

Data visualization

Data visualization is the graphic representation of data. It involves producing images that communicate 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.

整體趨勢分析

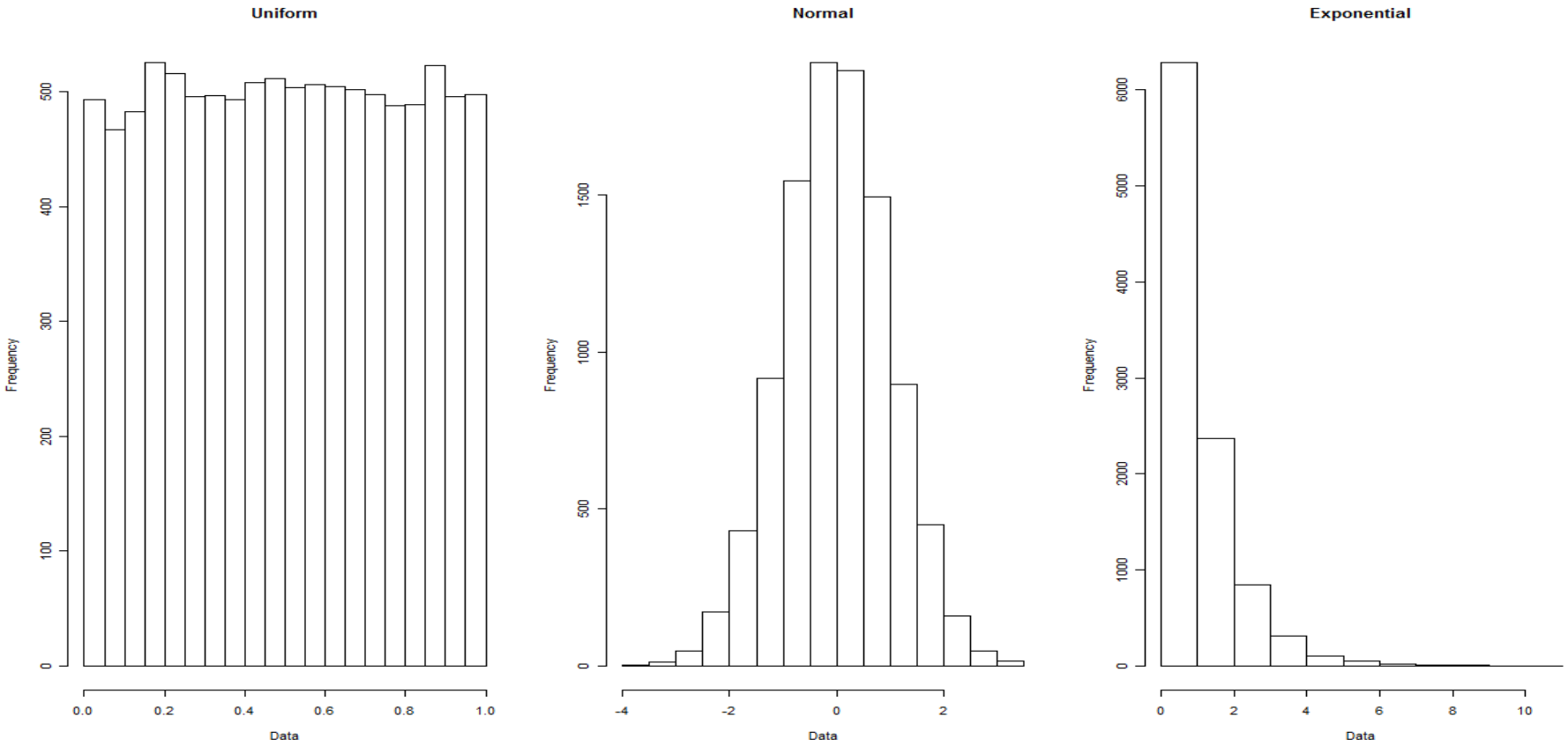
- 教科書資料的數量通常較少，很少出現異常觀察值，但實際資料經常會有「意外」，需要視情況而定，調整分析步驟及項目。
- 整體趨勢的分析可仿造「集中趨勢」及「分散趨勢」，計算具有代表性的數據，接著再輔以圖形、表格，以另一角度驗證這些結果，作為進一步分析的參考。

→問題：若統計數據與圖表不一致，如何進行下一步？

範例一：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。



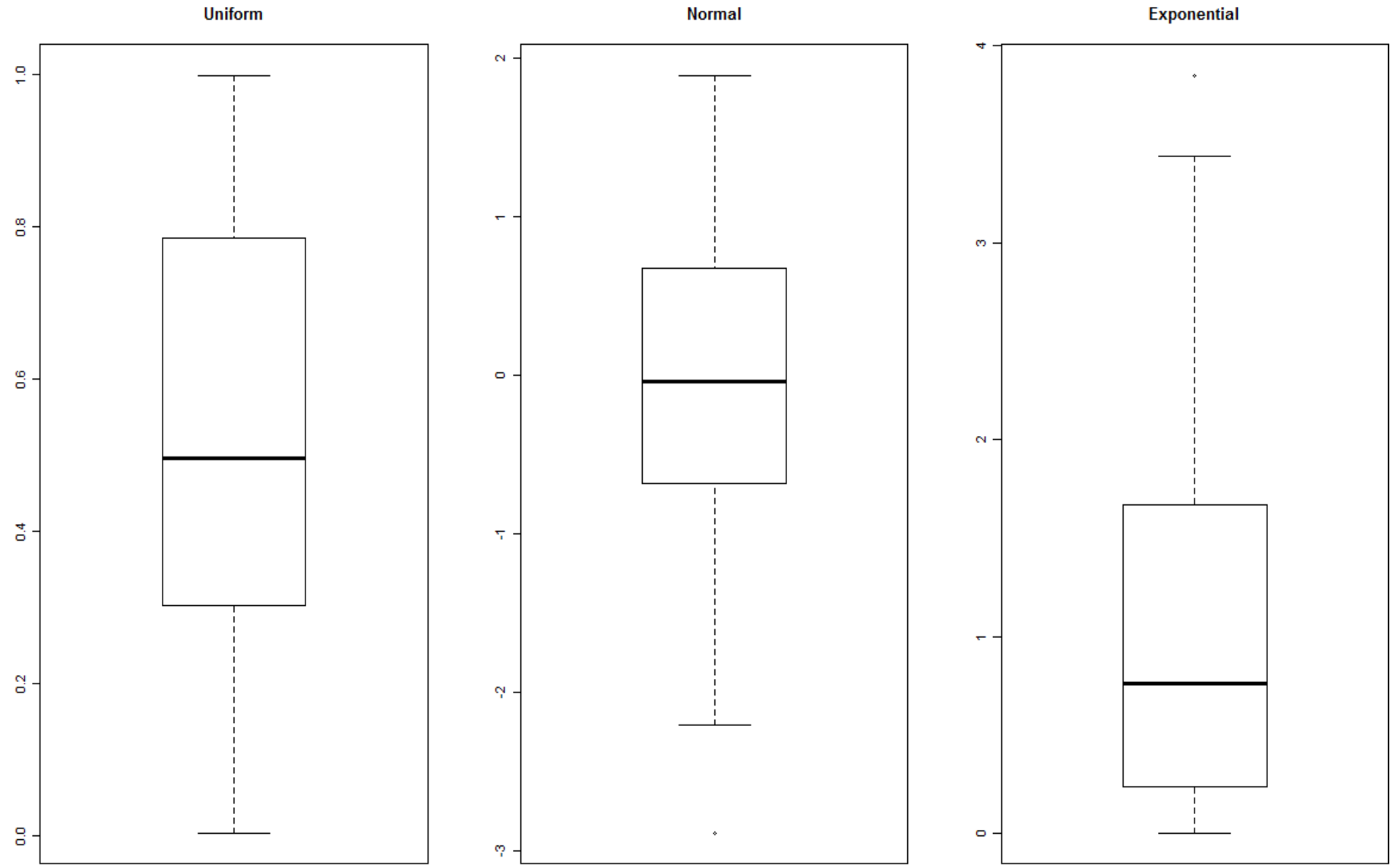
選擇有代表性的統計數據

- 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。

範例一（續）：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)

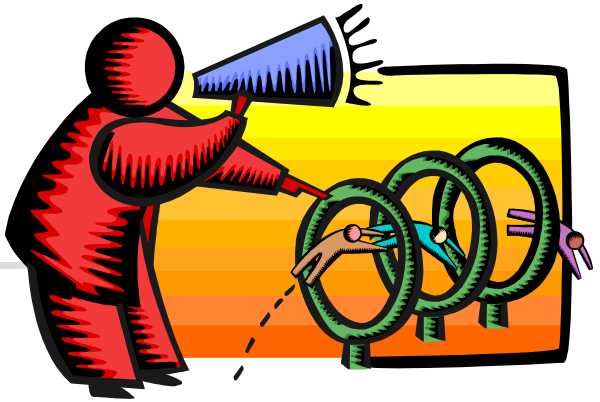
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	St.d.
-2.2060	-.4512	.1167	.1298	.9329	2.2570	.9507

(3)

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

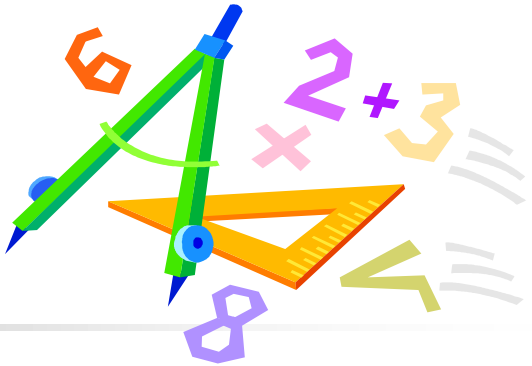
註：第三筆資料明顯右偏；第一筆資料比第二筆資料更為「均勻」。

統計分析的原則



- 確定問題的定義
- 化繁為簡（反璞歸真）
- 結合相關知識
- 發揮聯想力（大膽假設）
- 勿驟下結論（小心求證）

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。

→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。

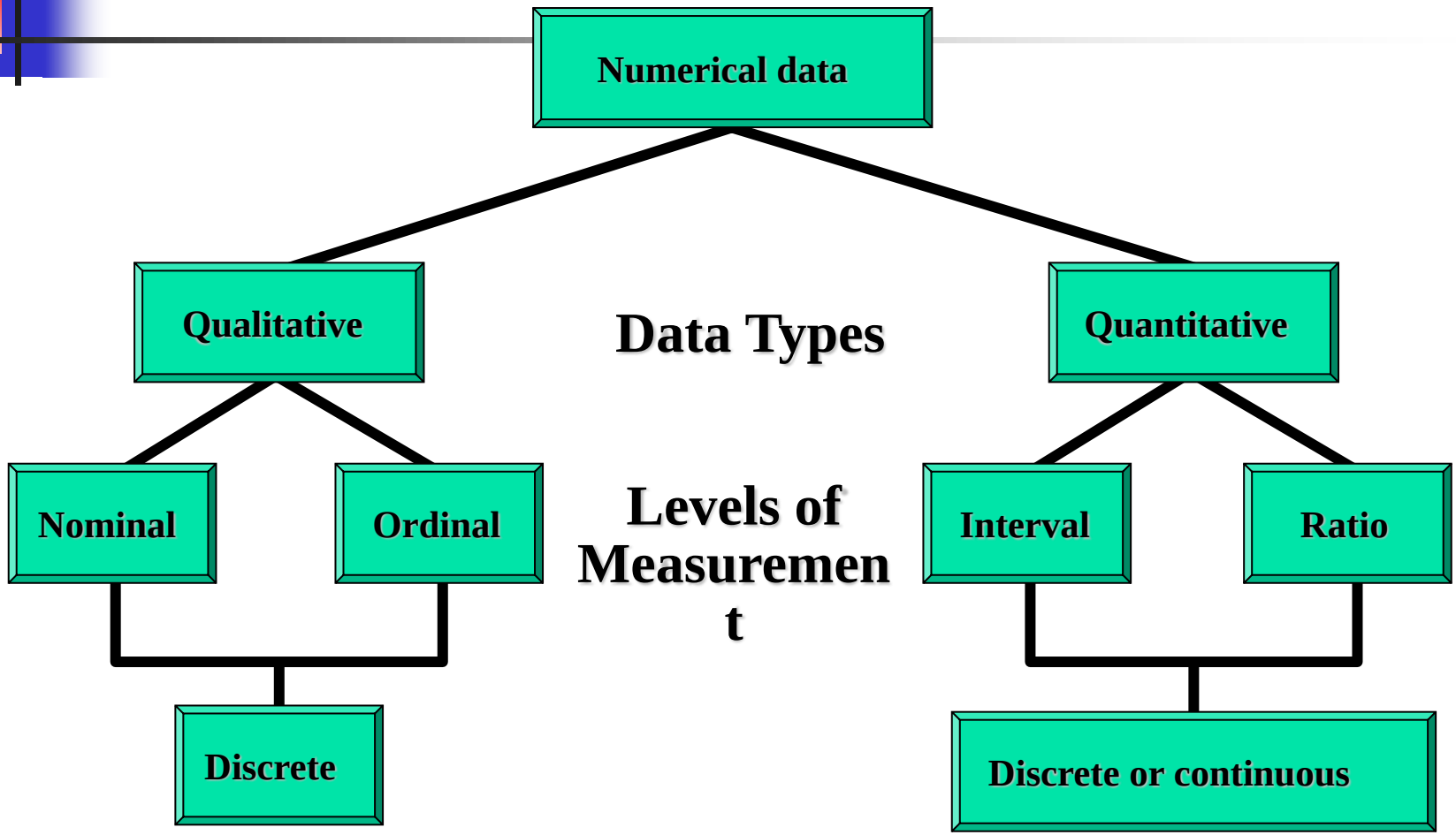
- 資料類型將直接影響分析方法的選取，並非所有資料都適合常見的統計方法，任意使用分析方法可能會得出令人啼笑皆非的結果。

  →已知A > B, B > C 是否代表A > C？

	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。

Types of Data(資料類型)



圖形與表格

- 除了基本的敘述統計量外，圖形與表格可以輔助判斷資料的特性。

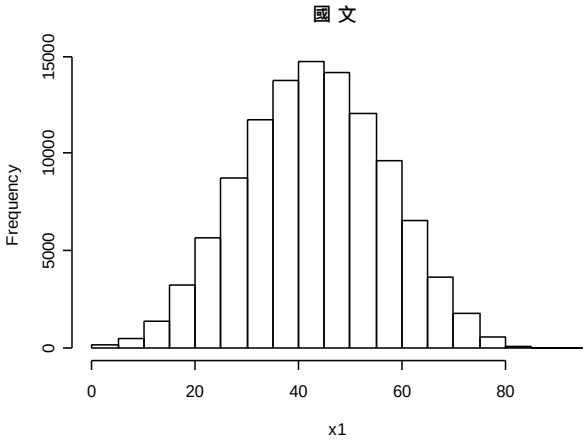
→常見的圖形：Boxplot、Histogram

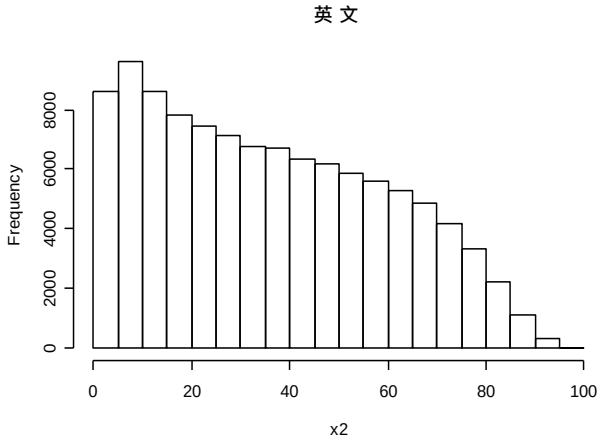
- 這些圖表看似簡單，但仔細判讀仍可發現重要訊息，甚至不需進階統計分析，即能約略猜出分析的結論。

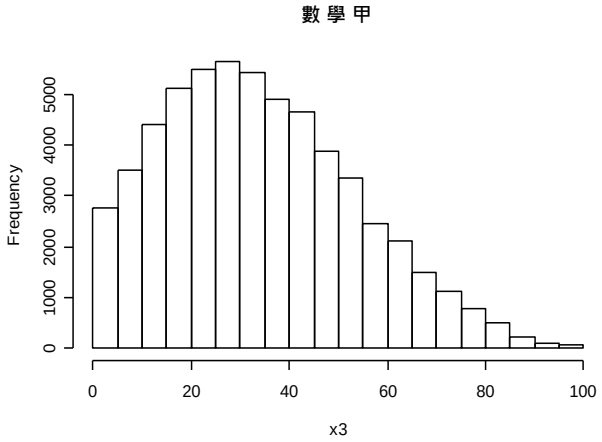
→以民國94年大學指定科目考試的成績為例，判斷各科分數的特性。

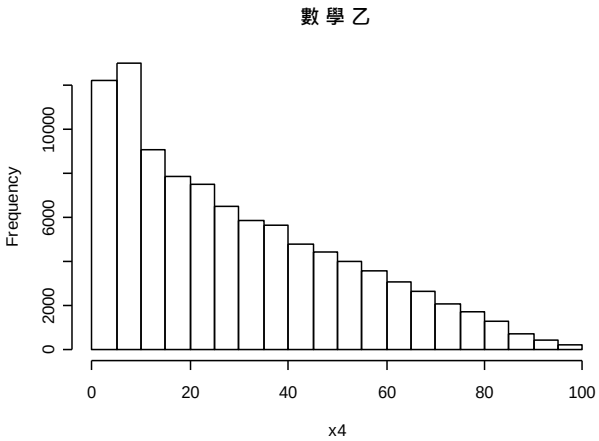
民國 94 年大學指定考試各科成績

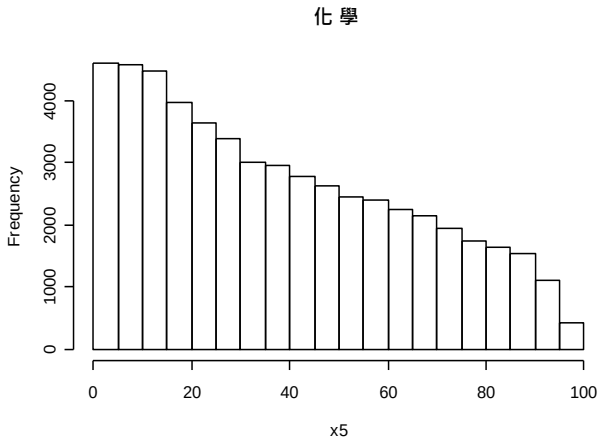
	國文	英文	數學甲	數學乙	化學	物理	生物	歷史	地理
Min.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0	0.00
12%	27.00	8.00	11.00	4.00	8.00	6.00	22.00	13.0	18.00
1st Qu.	34.00	16.00	22.00	12.00	15.00	12.00	32.00	28.0	30.00
Median	44.00	34.00	34.00	29.00	34.00	23.00	45.00	39.0	39.00
Mean	43.56	36.68	36.36	34.36	38.88	28.75	46.16	38.7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

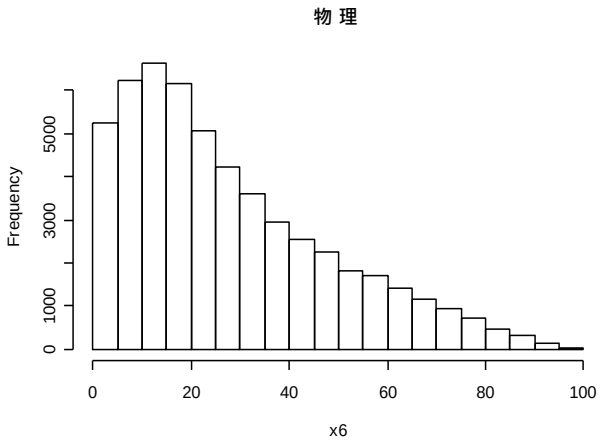

國文
Frequency
0
5000
10000
15000
0
20
40
60
80
x1

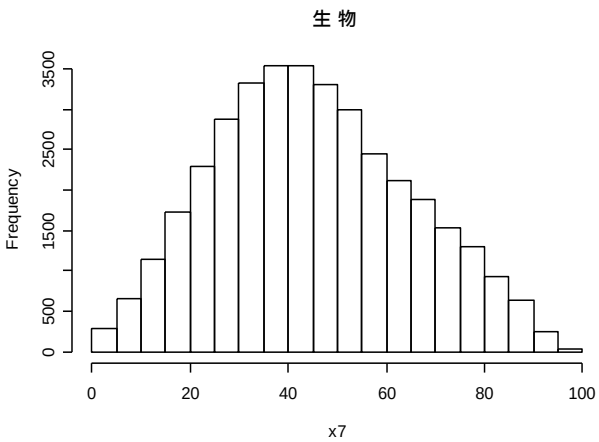

英文
Frequency
0
2000
4000
6000
8000
0
20
40
60
80
100
x2

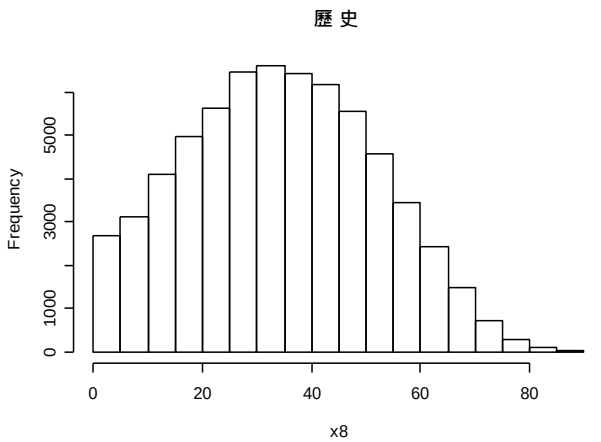

數學甲
Frequency
0
1000
2000
3000
4000
5000
0
20
40
60
80
100
x3

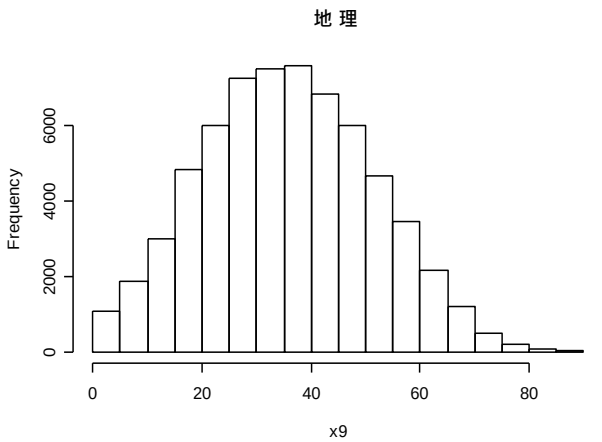

數學乙
Frequency
0
2000
6000
10000
0
20
40
60
80
100
x4

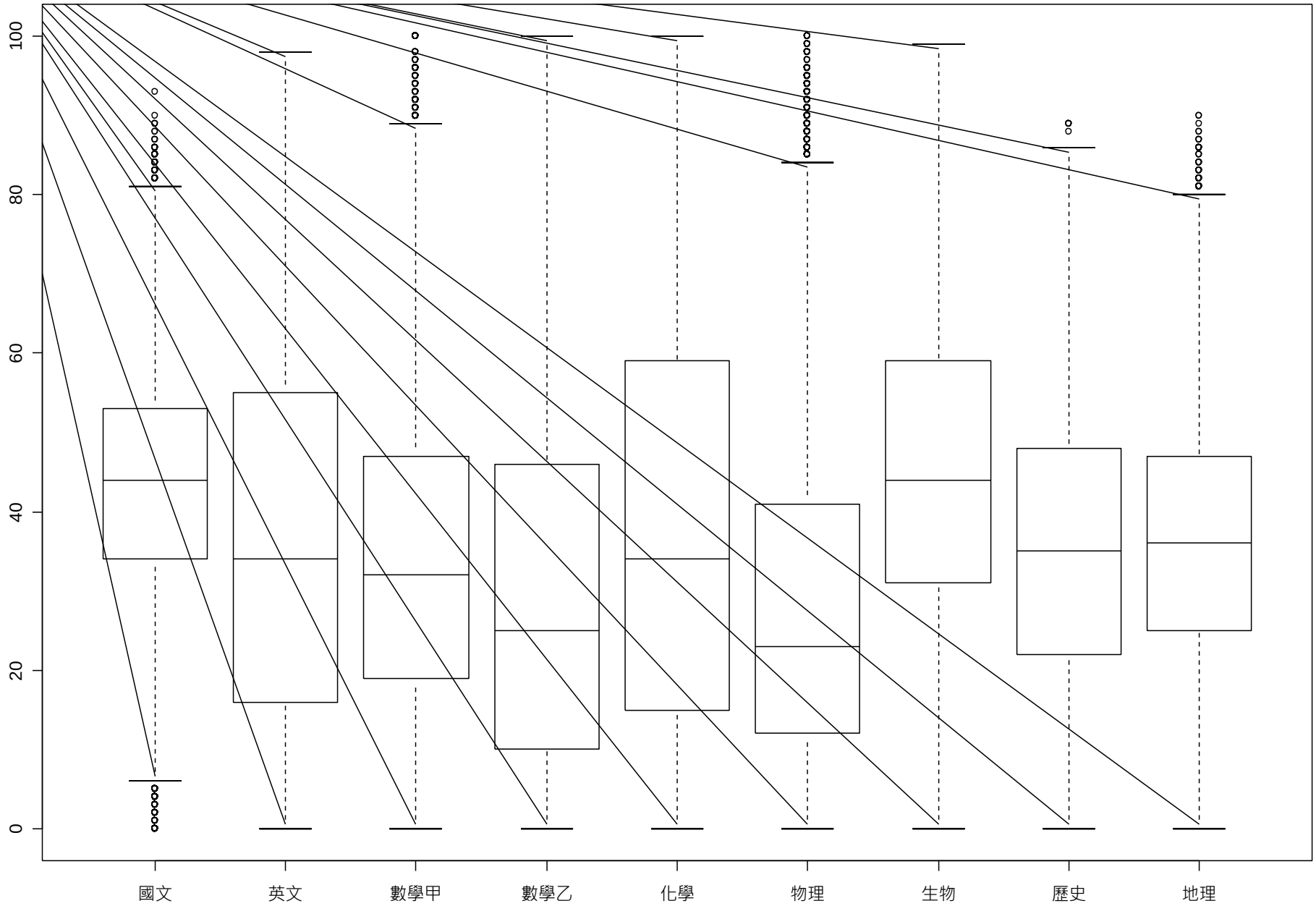

化學
Frequency
0
1000
2000
3000
4000
0
20
40
60
80
100
x5


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

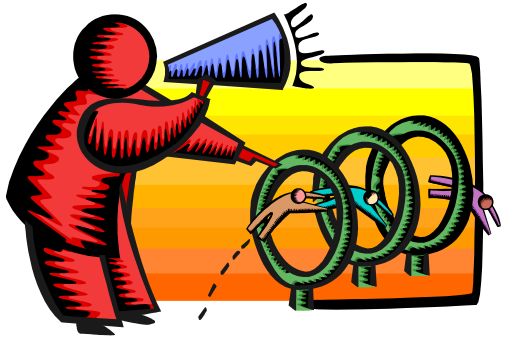

歷史
Frequency
0
1000
3000
5000
0
20
40
60
80
x8


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

巨量資料與統計研究

- 過去統計研究多仰賴理論推導、電腦模擬，巨量資料讓實證分析的角色更多元，解除資料大小及範圍的限制。
- **資料驅動**（**Data Driven**；讓資料說話）提供另一種角度的思維，有別於由專家意見導引研究方向，藉由基本資料分析篩選出資料的重要特質。



資料vs.理論（或模型）

- 資訊呈現幾何級數成長，不少人認為資料足以取代理論，但數字可以自己說話嗎？

→資料被人們賦予意義，但也會因特定目的而曲解，因而失去其客觀性。

- 輕易、過度迷信模型是另一種極端，全球金融危機即是教訓，也應搭配重要訊息。

→資料與理論（模型）的配合，經常需要經歷一連串的錯誤學習，透過試誤、整合等階段，反反覆覆的累積經驗。

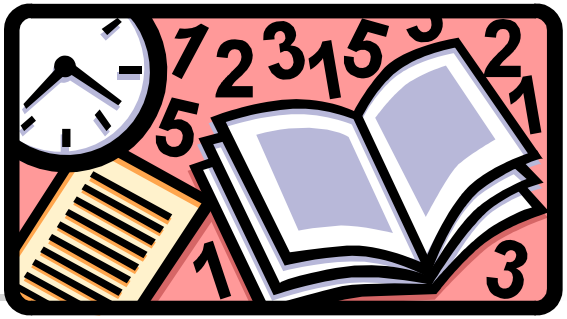
定義問題

蒐集資料

分析資料

詮釋結果

迴歸分析



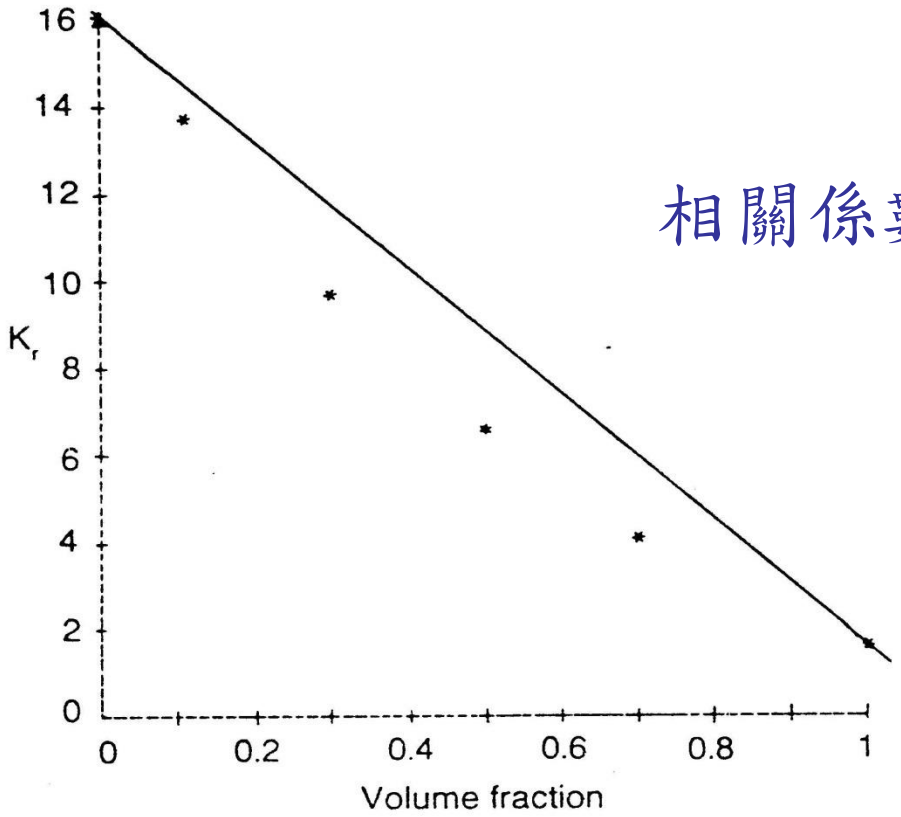
- 迴歸分析是最常見的統計方法之一，但也是最容易被誤用的方法。

→ 迴歸測量解釋變數與反應變數間的線性關係，但兩者不見得具有因果關係。

→ R^2 測量資料配適的吻合度，數值大者不代表迴歸模型正確；變數顯著者也不代表整體的迴歸模型可用。

迴歸分析(續)

- 猜猜看 R^2 的大小。



相關係數＝–0.985 !!!

Figure C.1 Observations on two variables for the chemical hexane-benzonitrile-dimesulfoxide.

迴歸分析(續)：何者 R^2 較大？

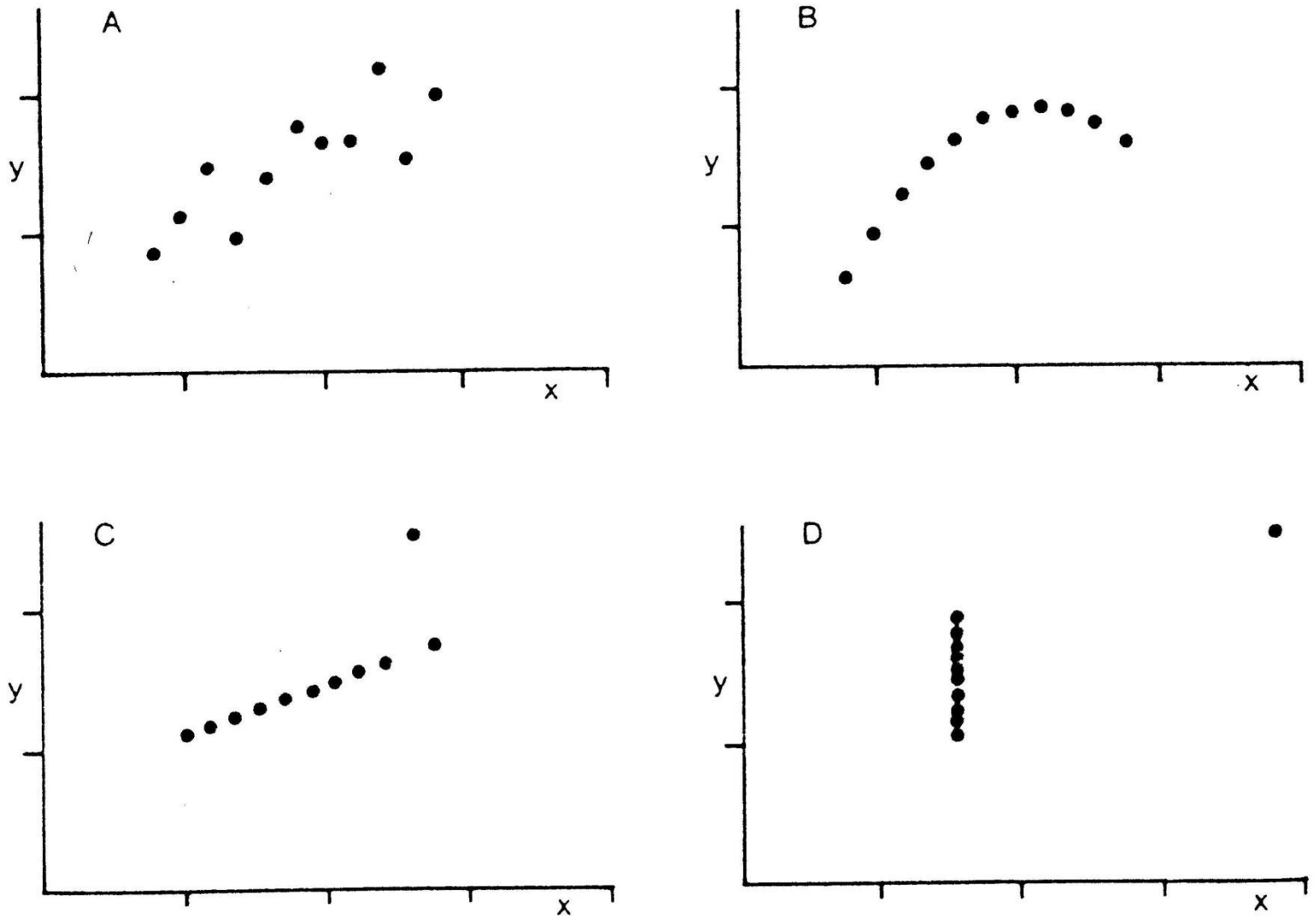


Figure C.2 Four bivariate sets of data.

迴歸分析(續)

- 你相信嗎？！

→ 由四個圖形中的資料，配適出截距及斜率完全相同的迴歸方程式。($R^2 \cong 0.67$)

(Anscombe, 1973, 文章中的資料)

- 但你/妳覺得哪一個圖形中的資料較合適使用迴歸分析？

迴歸分析(Regression Analysis)

- 迴歸分析以自變數（獨立變數 *X*；Independent Variable）的函數型態，描述與因變數（被解釋變數*Y*；Dependent Variable）的關係：

$$Y = f(X_1, X_2, X_3, \ldots, X_n) + e$$

其中：

- *f*( ) 代表系統變異(*systematic* variation)
- e 則為非系統或隨機變異(*unsystematic* variation or random error)

- 如果選取了適合的解釋變數X，被解釋變數Y的迴歸模型中

$$Y = f(X_1, X_2, X_3, \ldots, X_n) + e$$

觀察值＝模型＋誤差

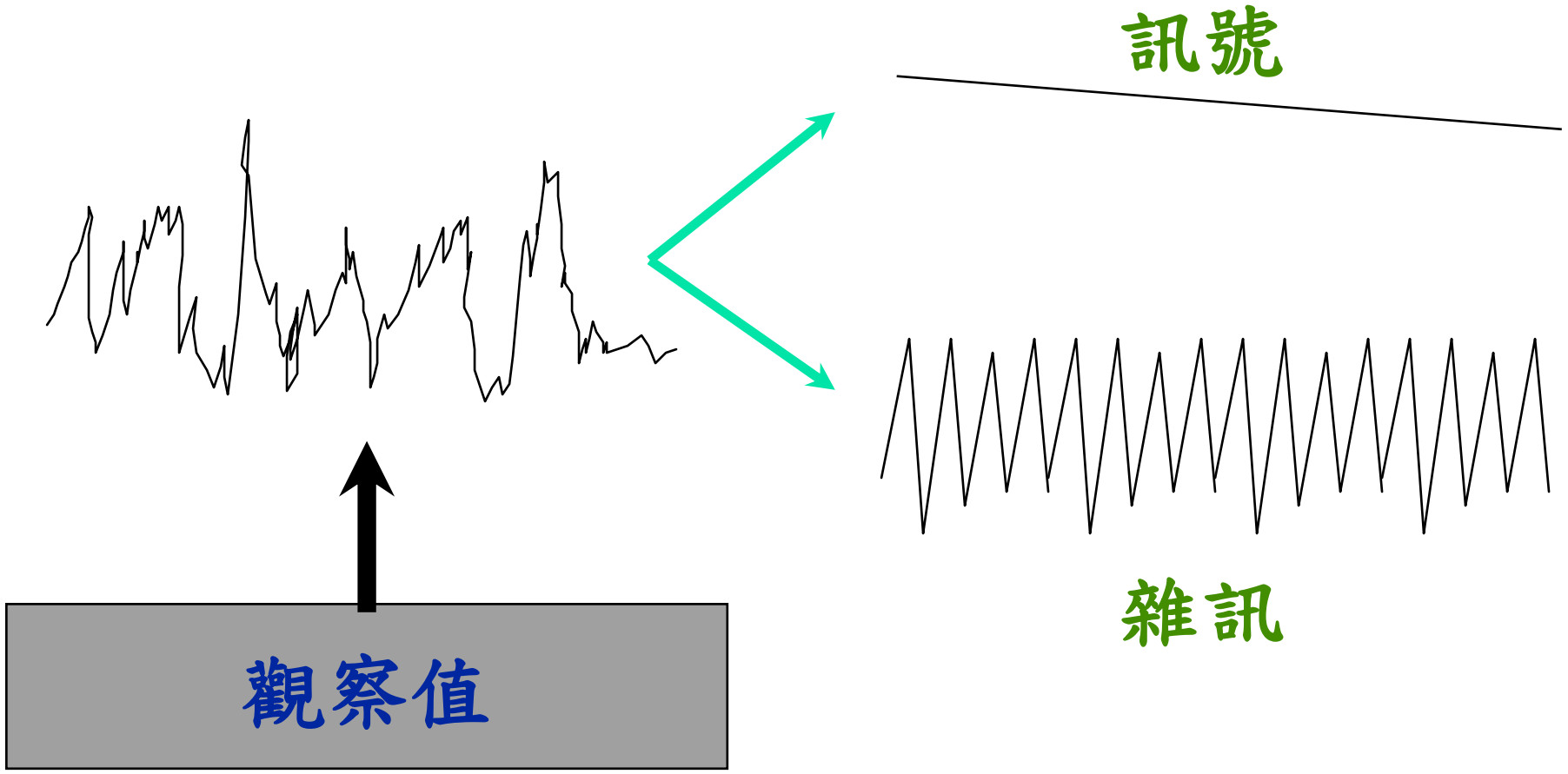
誤差項應該很小。

註： Observations = Model + Error

Observations = Signal + Noise

註：迴歸分析中，不代表X與Y間存在因果關係，因果關係需由其他資訊決定。

訊號與雜訊(Signal to Noise)



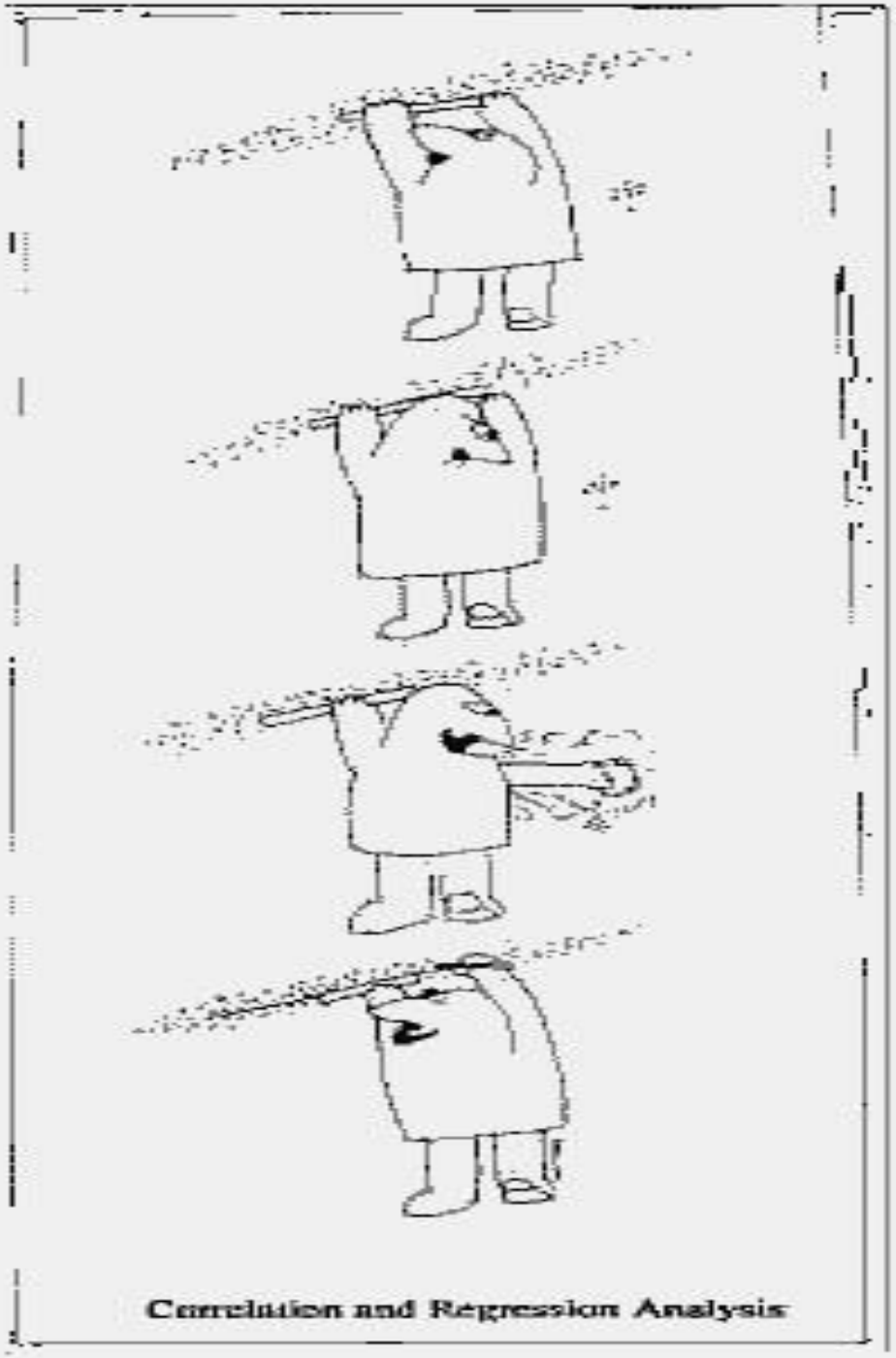
迴歸模型(續)

如果真實模型是：

$$y_i = B_0 + B_1X_i + B_2Z_i$$

$$y_i = b_0 + b_1X_i + b_2Z_i + e_i$$

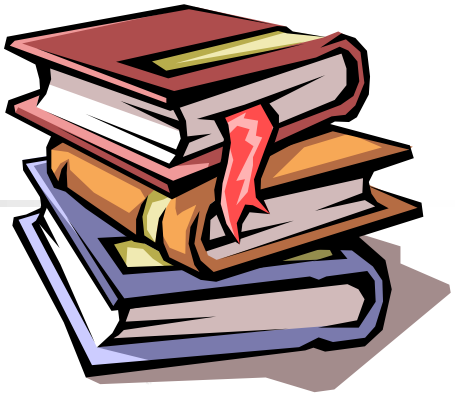
如果正確地找出X與Y的函數關係，
剩餘的誤差將不具任何有系統的資訊。



https://www.slideshare.net/alexandrdoro/outliers-43285182

Outliers are model-based!

分析範例(一)



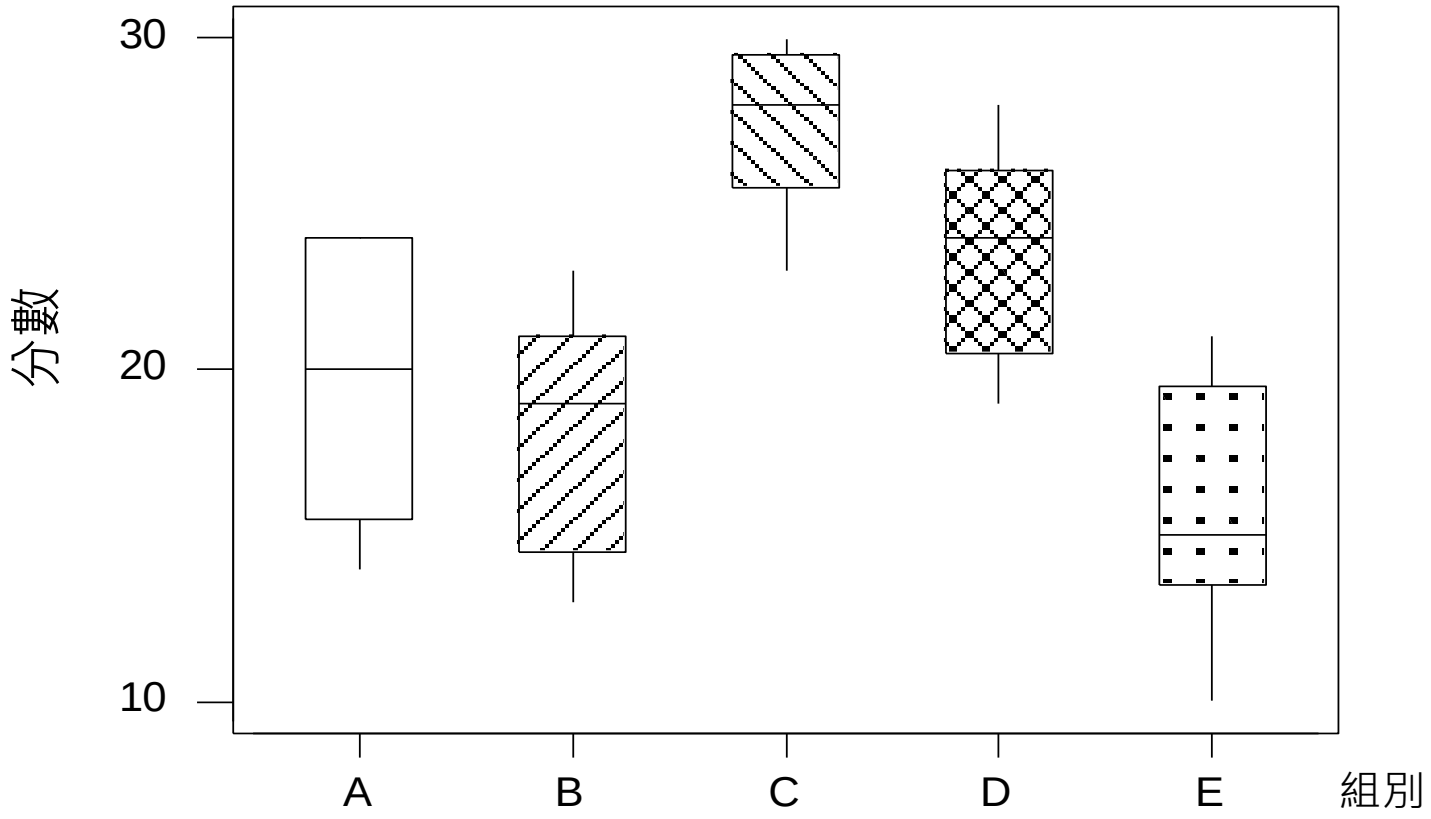
- 教學方法是否存有差異？

Table B.1 Test results for 45 students

Group A (control)	17	14	24	20	24	23	16	15	24
Group B (control)	21	23	13	19	13	19	20	21	16
Group C (praised)	28	30	29	24	27	30	28	28	23
Group D (reproved)	19	28	26	26	19	24	24	23	22
Group E (ignored)	21	14	13	19	15	15	10	18	20

→ 進行變異數分析前可用的方法？

箱型圖是不錯的選擇！

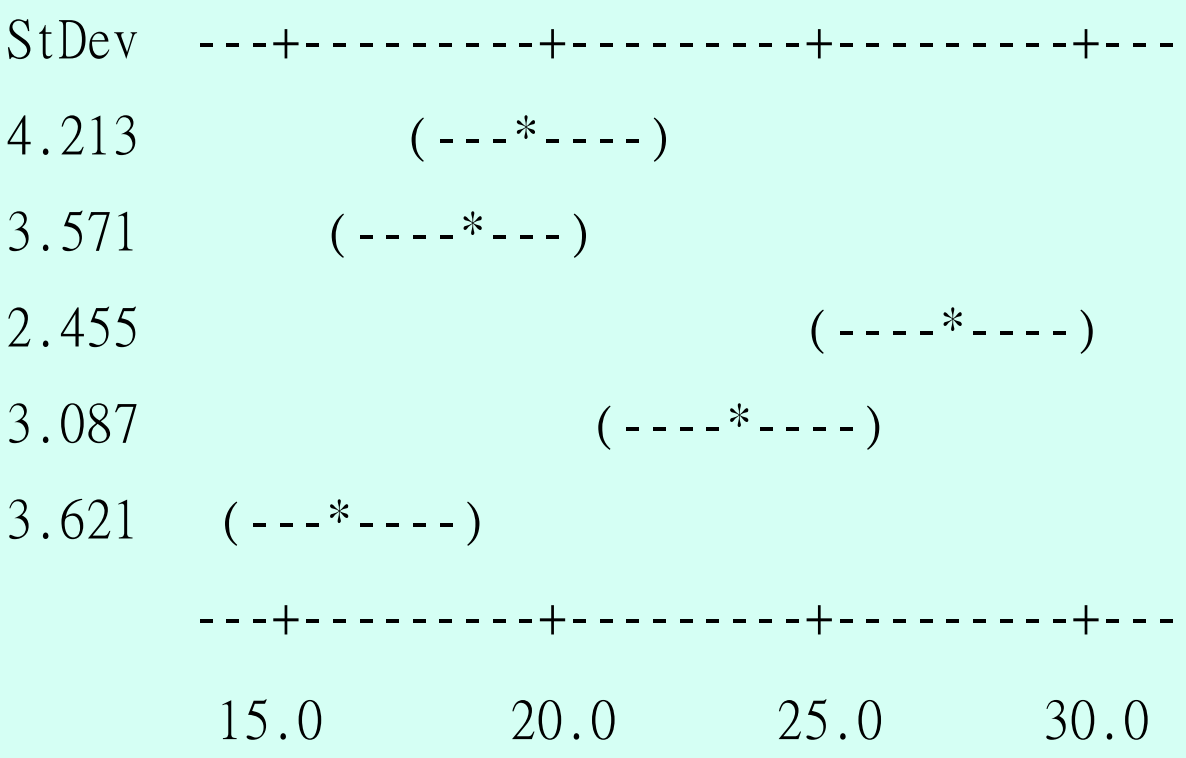


Analysis of Variance

Source	DF	SS	MS	F	P
Factor	4	722.7	180.7	15.27	0.000
Error	40	473.3	11.8		
Total	44	1196.0			

Individual 95% CIs For Mean Based on Pooled StDev

Level	N	Mean	StDev
A	9	19.667	4.213
B	9	18.333	3.571
C	9	27.444	2.455
D	9	23.444	3.087
E	9	16.111	3.621



Pooled StDev = 3.440

分析範例(二)

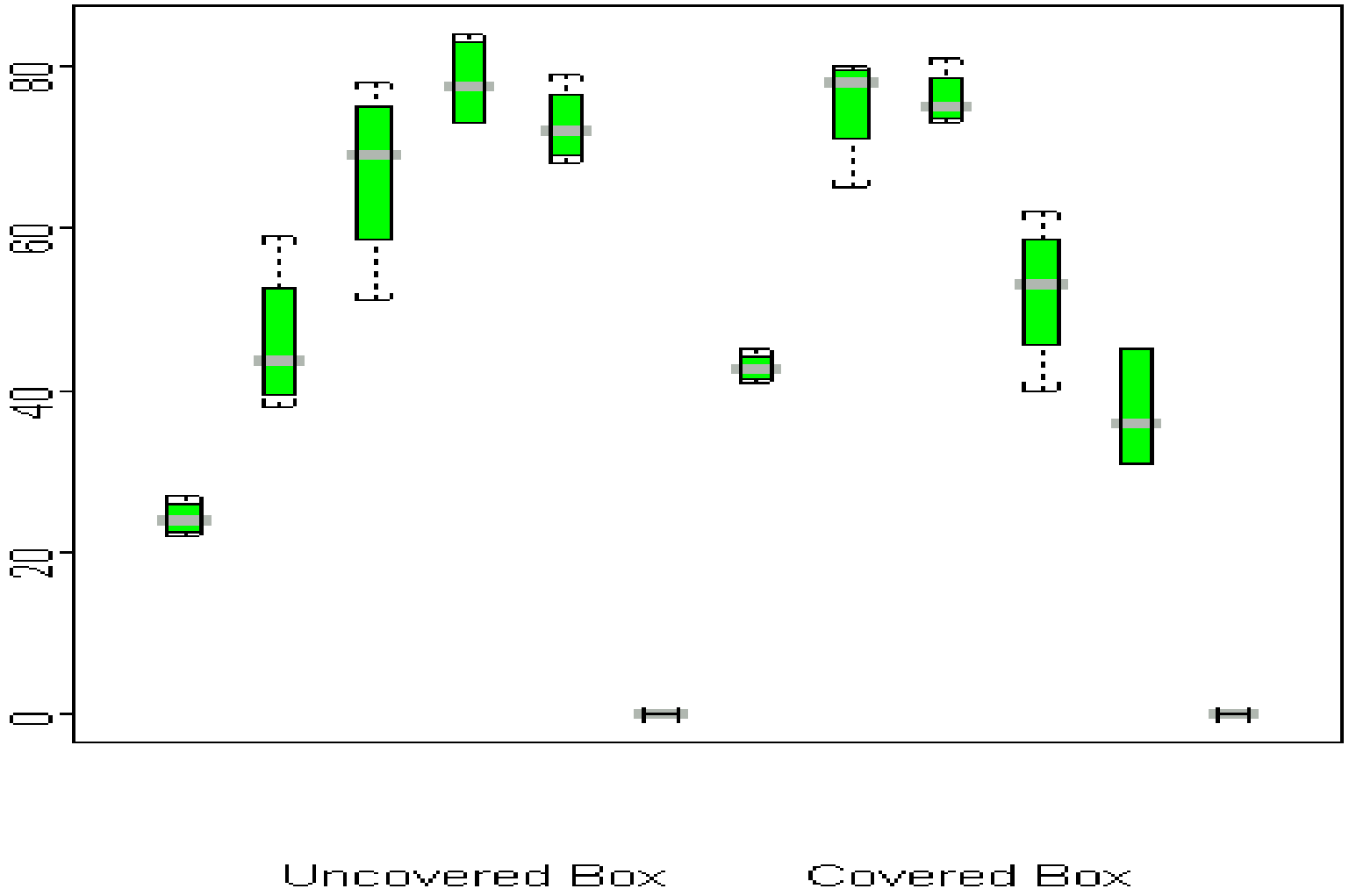
- 種子發芽與哪些因素有關？

Table B.2 Numbers of seeds germinating

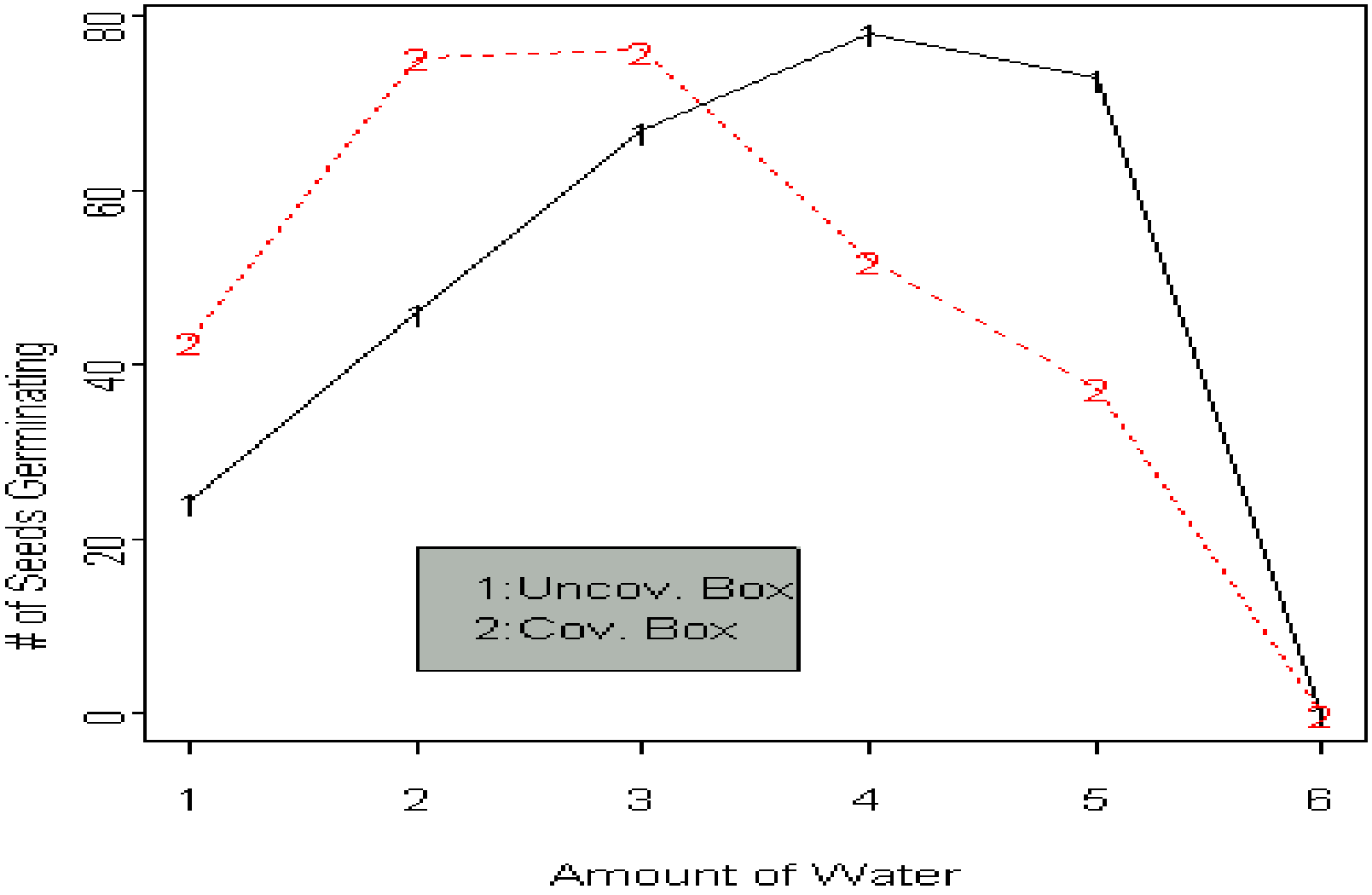
	Moisture content					
	1	3	5	7	9	11
Boxes uncovered	22	41	66	82	79	0
	25	46	72	73	68	0
	27	59	51	73	74	0
	23	38	78	84	70	0
Boxes covered	45	65	81	55	31	0
	41	80	73	51	36	0
	42	79	74	40	45	0
	43	77	76	62	*	0

* Denotes missing observation.

箱型圖中先找出趨勢！



折線圖中看出交互作用！


# of Seeds Germinating
0
20
40
60
80
1
2
3
4
5
6
Amount of Water
1:Uncov. Box
2:Cov. Box

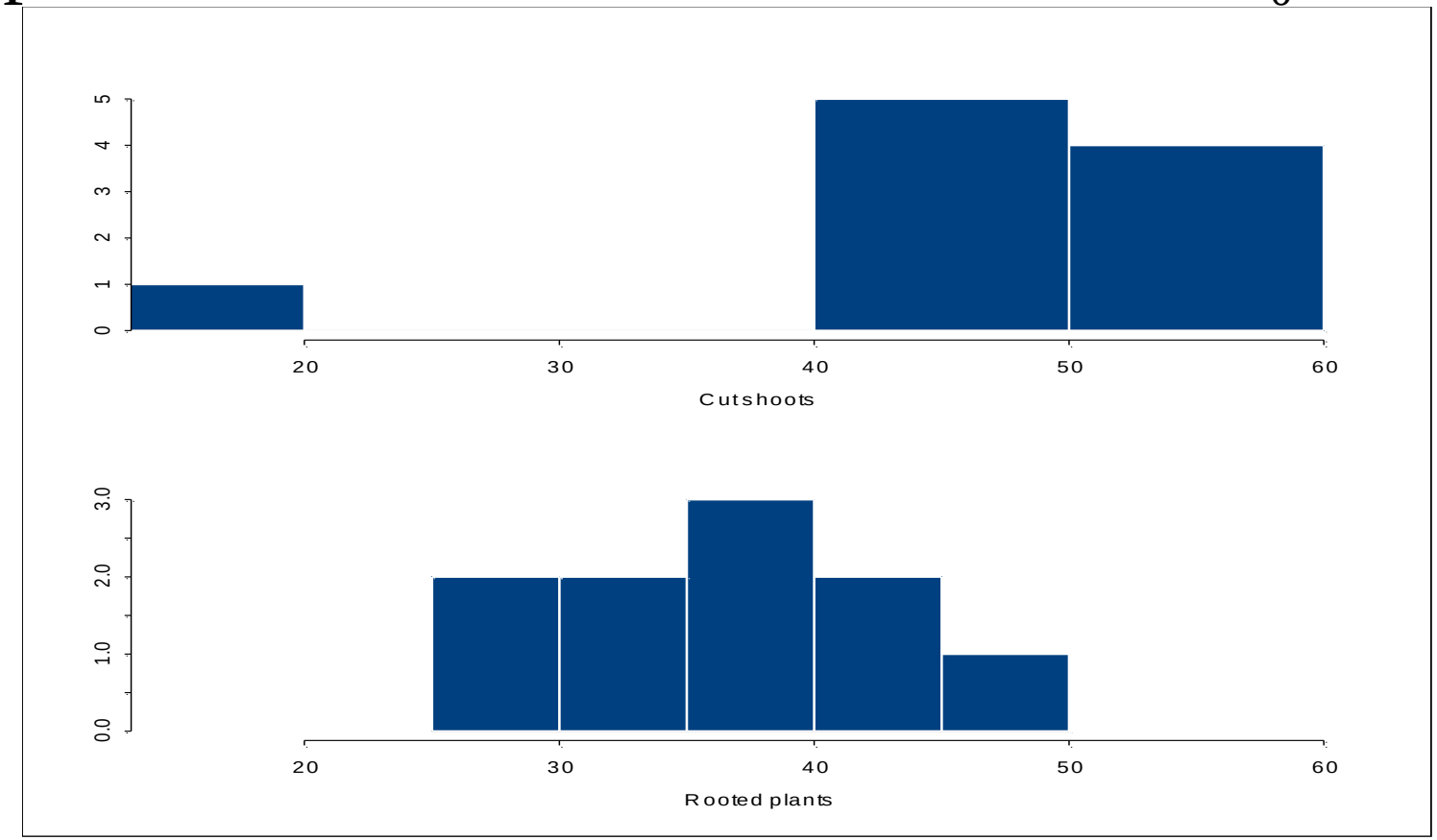
分析範例(三)

Cut shoots:	53	58	48	18	55	42	50	47	51	45
Rooted plants:	36	33	40	43	25	38	41	46	34	29

- 兩種栽種方式是否有相同的效果？

→ p-value = 0.11 > α = 0.01，不拒絕 H_0

→



分析範例(四)

→頭髮顏色是否與眼睛顏色相關？
(不藉助統計分析工具的前提下！)

Table B.3 Observed frequencies of people with a particular combination of hair and eye colour

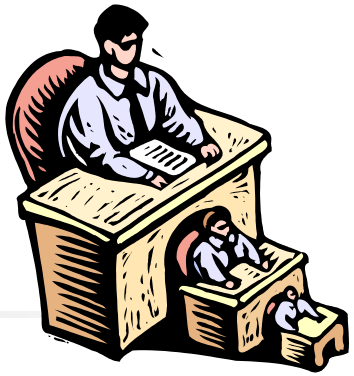
	Hair colour			
Eye colour	Black	Brunette	Red	Blond
Brown	68	119	26	7
Blue	20	84	17	94
Hazel	15	54	14	10
Green	5	29	14	16

另一個範例：

Table B.4 Observed frequencies of Swedish families with a particular yearly income and family size

Number of children	Yearly income (units of 1000 kronor)				
	0–1	1–2	2–3	3+	Total
0	2 161	3 577	2 184	1 636	9 558
1	2 755	5 081	2 222	1 052	11 110
2	936	1 753	640	306	3 635
3	225	419	96	38	778
≥4	39	98	31	14	182
Total	6 116	10 928	5 173	3 046	25 263

時間數列(Time Series)



- 蒐集的資料具時間先後順序，且每一個觀察值的數值大小與其前後的觀察值有關。

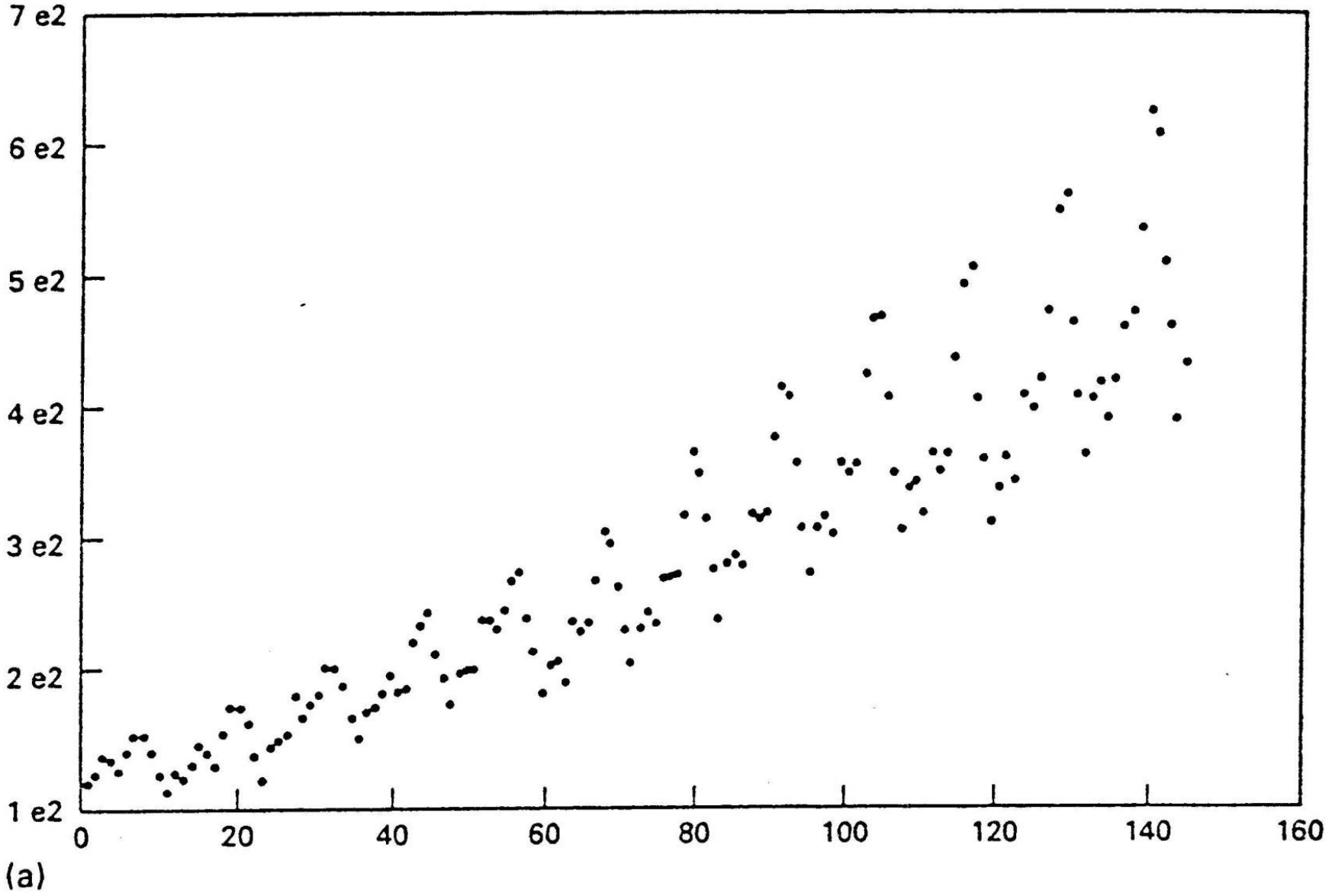
→ 股票價格

→ 木柵的每天最高、低、平均溫度

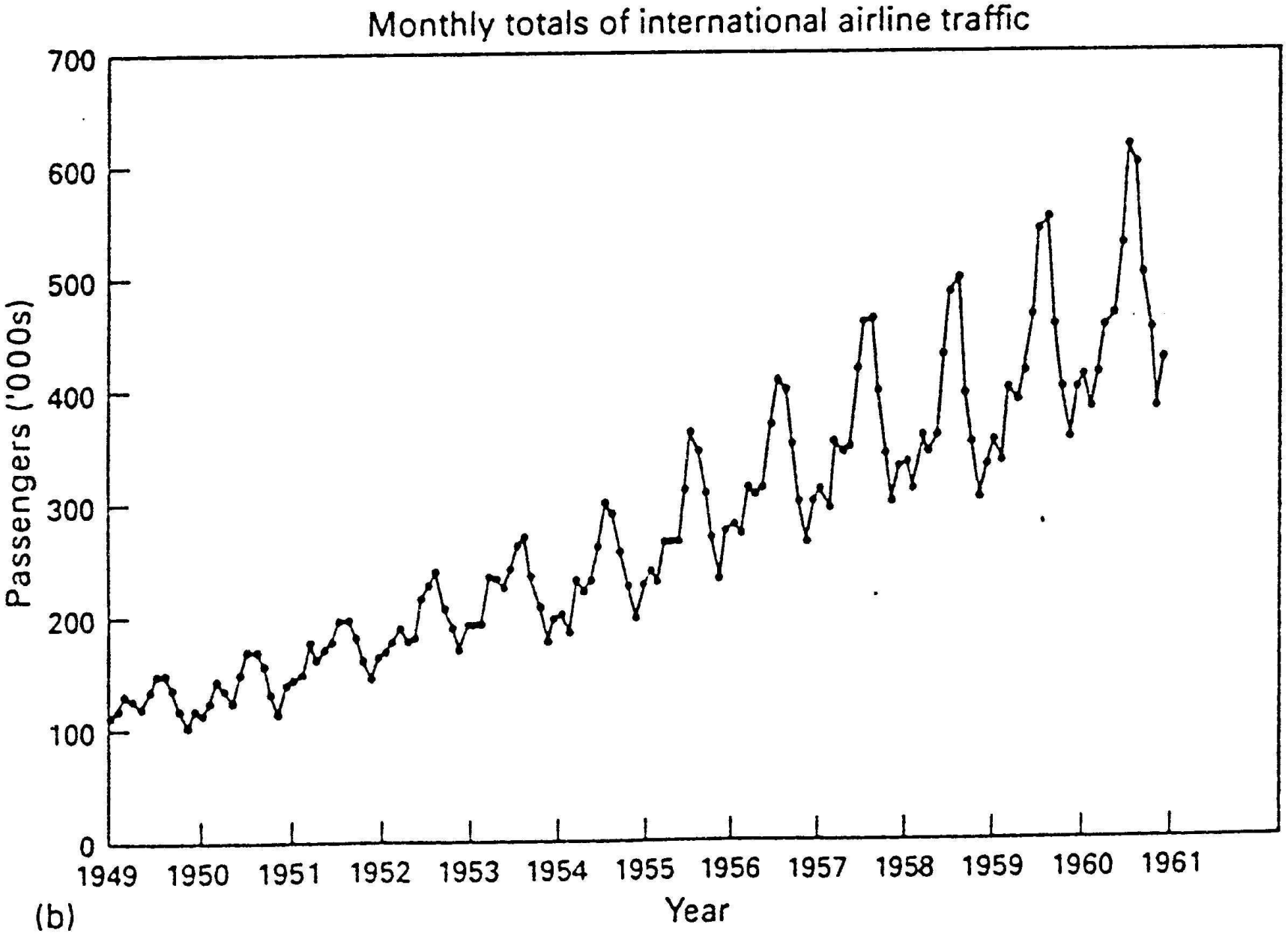
→ 飯店的住房率(淡、旺季之分)

- 如何判斷資料具前後相關性？

什麼分析方法合適？

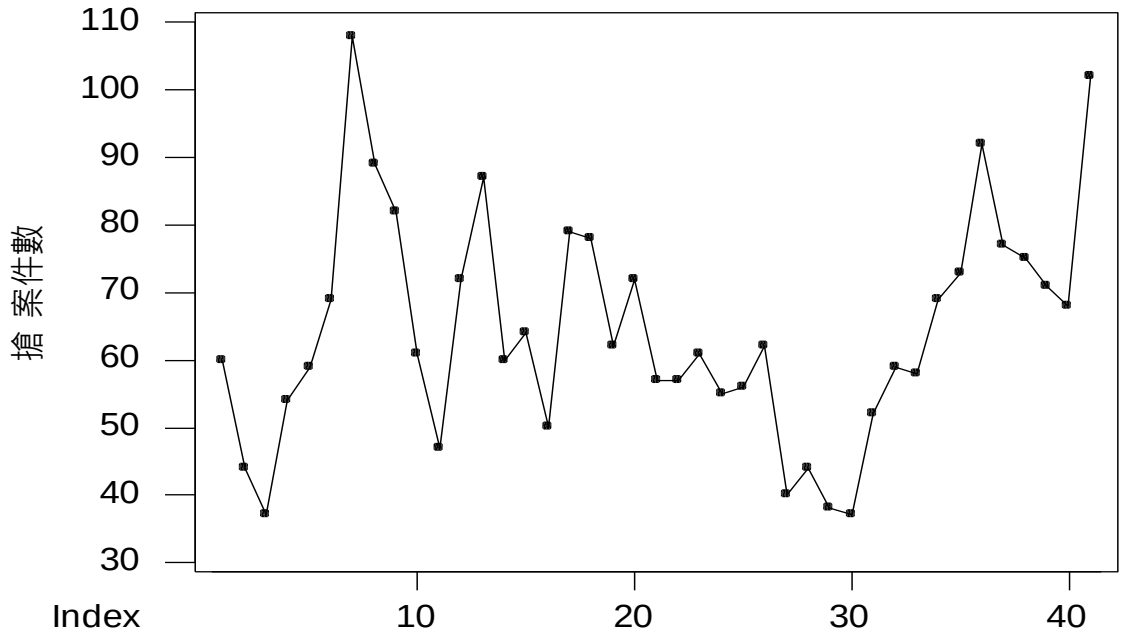


連線之後較易看出端倪！


Monthly totals of international airline traffic
Passengers ('000s)
700
600
500
400
300
200
100
0
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
Year
(b)

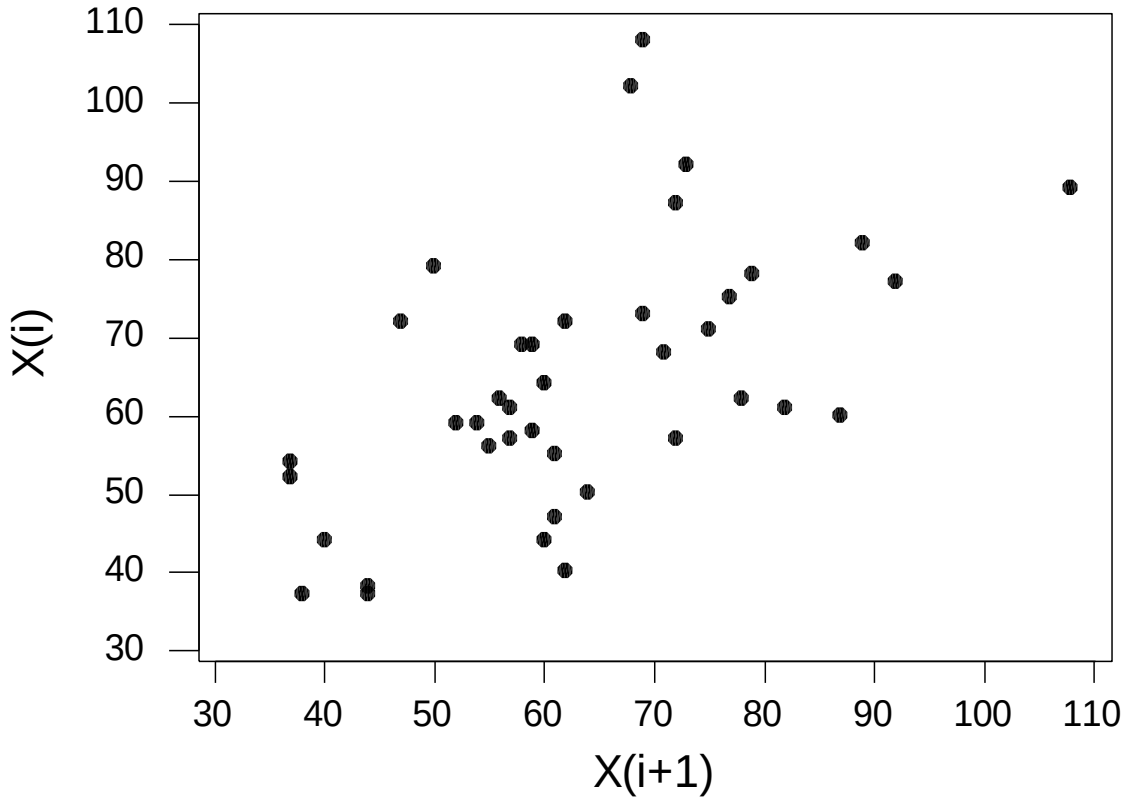
時間數列(範例)

- 如何判斷下列資料具時間先後關係？



試試看畫出X_i vs. X_{i+1}！

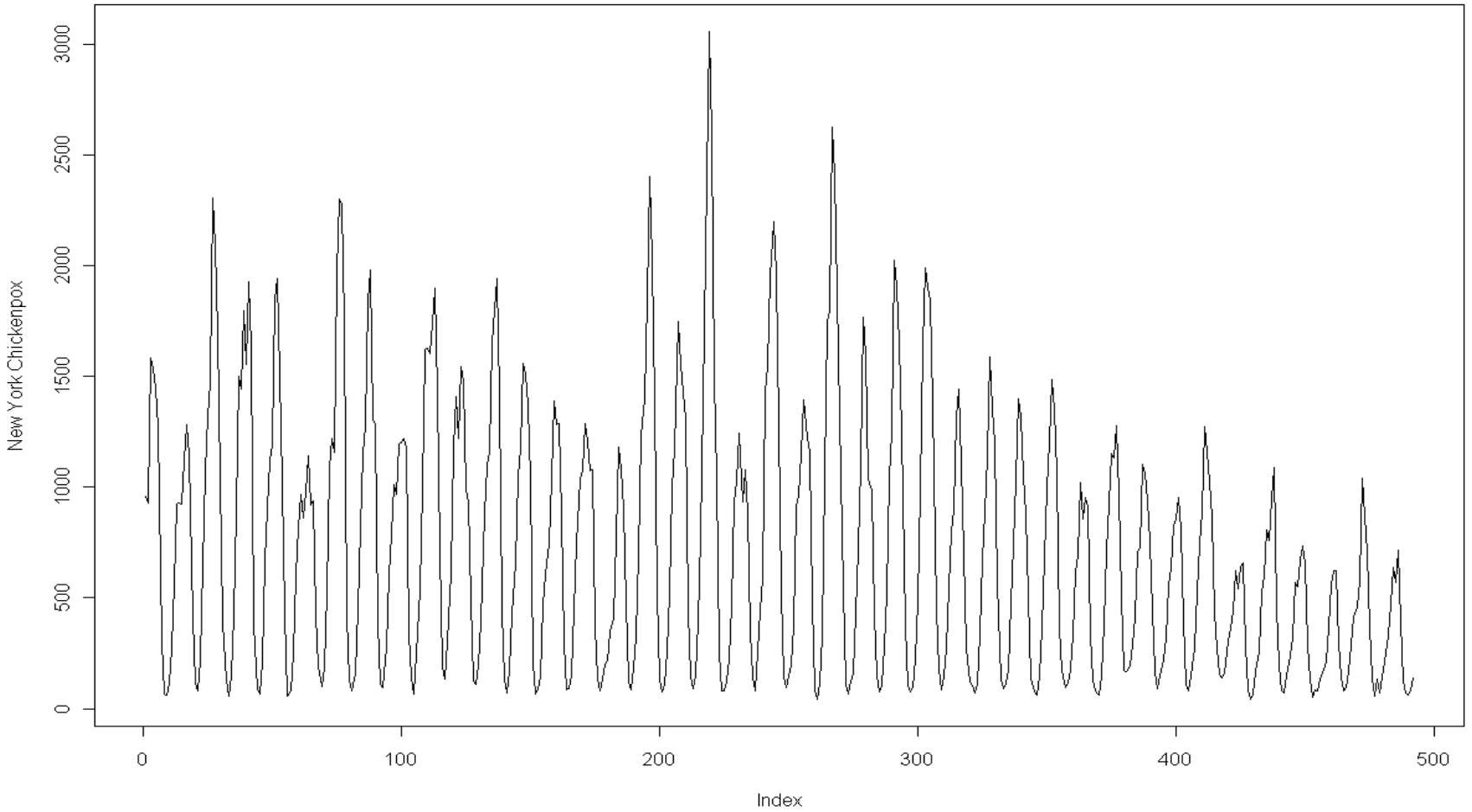
→ 是否有明顯的正相關？

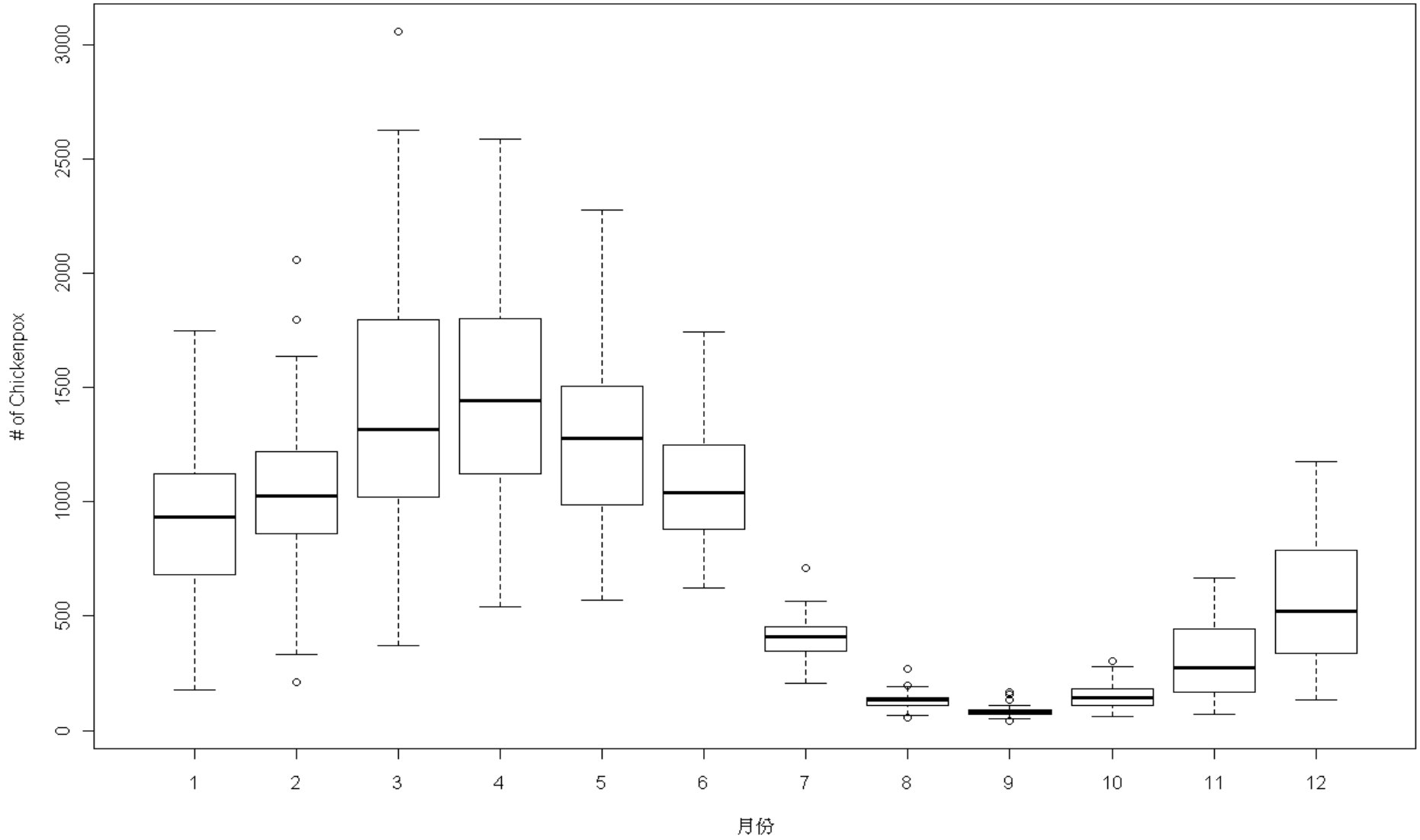


例題五：時間相關的資料分析

- 如何分析與時間有關的資料？（品質管制）

→下圖為美國紐約1931-1972每月水痘發病數。





→明顯可知春天病例數較多、秋天較少。