什麼是統計？

政治大學統計系余清祥

2008年8月31日

Email: csyue@nccu.edu.tw

Website: csyue.nccu.edu.tw

商學院有3G (三「計」)

會計— 很快忘記！

經濟—經常忘記！！

統計—通通忘記！！！

註：為什麼這三門課特別棘手，統計似乎最難上手？

統計是什麼、為什麼學統計？

- 我們對日常生活中許多事物都不確定，需要充分的資訊才能做出合理的判斷。

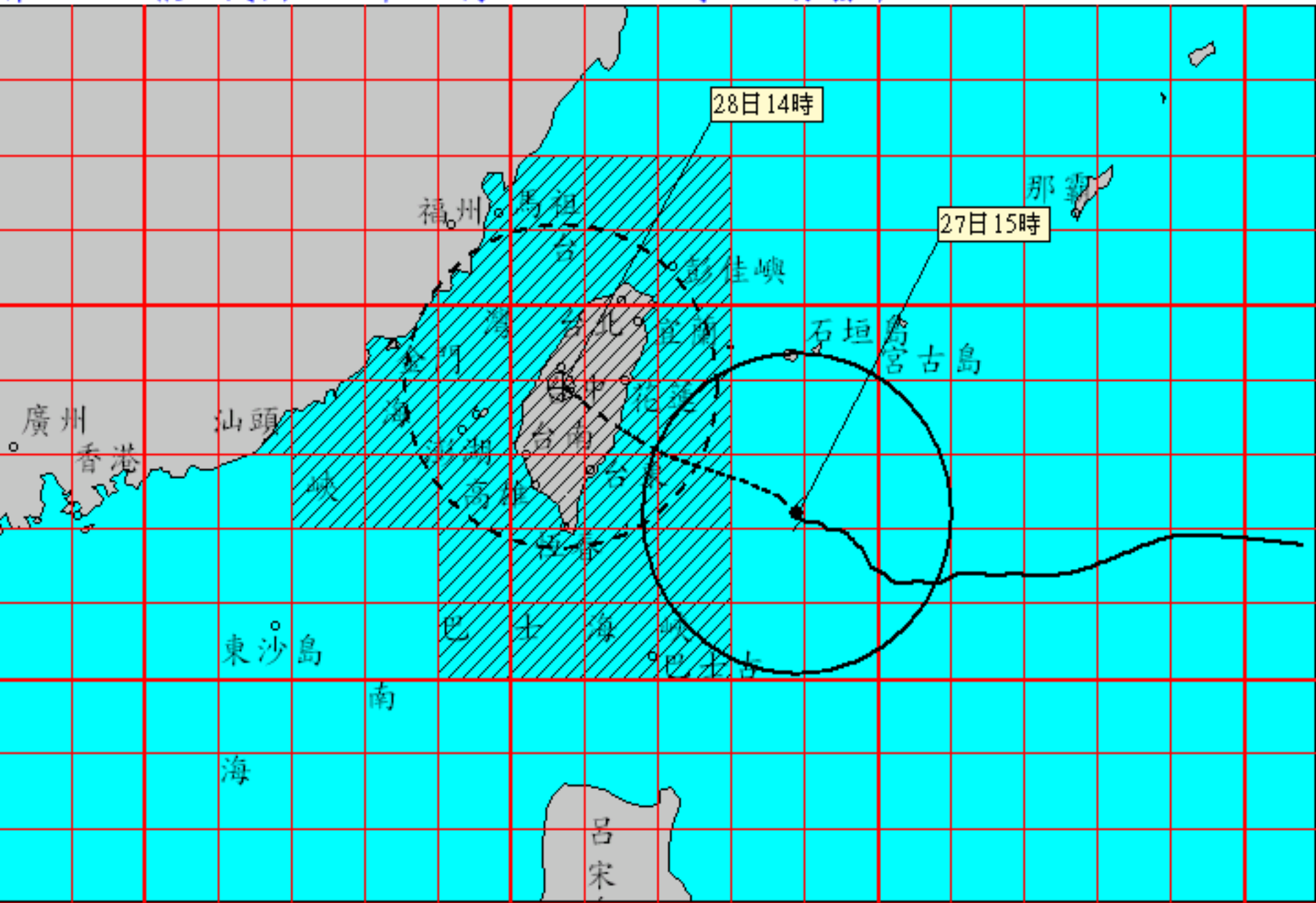
→例如：颱風會不會侵襲臺灣，可能帶來多少雨量？(歷史資料！)

→根據你/妳的大學基測分數，選擇最有可能上榜、又有興趣的科系。(分數落點！)

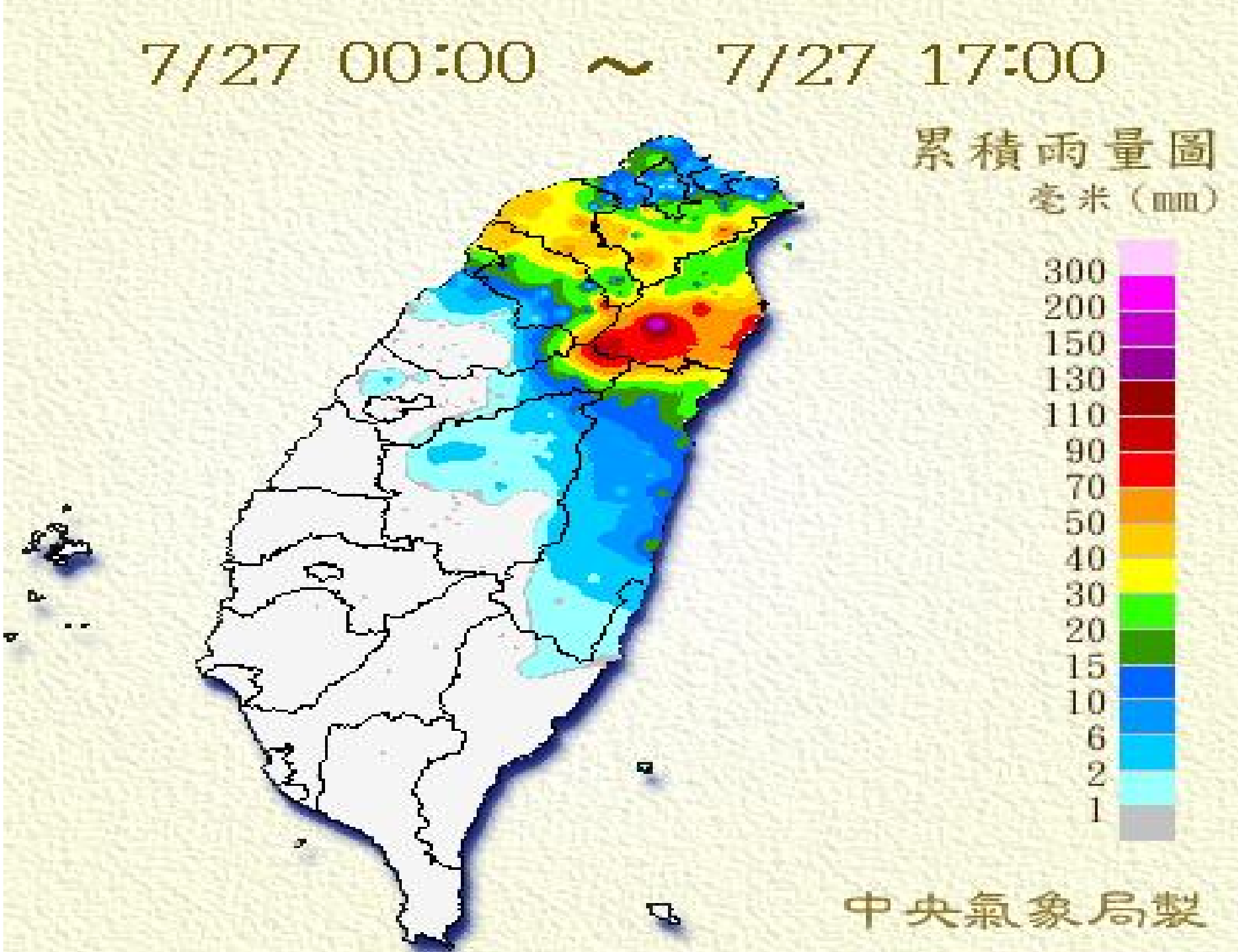
→如何選課，避免必修課程被當，或是免除被1/2的危險？(道聽途說？)

→王建民前兩年為何沒拿到賽揚獎？

中度颱風 （編號第8號 國際命名：FUNG-WONG，中文譯名：鳳凰）
第 10-1 報 民國 97 年 7 月 27 日 15 時 15 分發布




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


7/27 00:00 ～ 7/27 17:00
累積雨量圖
毫米（mm）
300
200
150
130
110
90
70
50
40
30
20
15
10
6
2
1
中央氣象局製

美國職業運動

山姆大叔(Uncle Sam)應該是全世界使用統計數據最頻繁的國家，由他們職業運動統計數字的多樣性、詳細程度可見一斑。

- 以最近風靡全台的美國職棒為例，你們覺得有那些資料會在統計之列？

→例如：有人認為王建民應該是賽揚獎的當然候選人，你們覺得哪些項目應該列入評比、比重又各是多少？

Sortable Stats: Pitching

League Leaders | Stats by Position

Batting Pitching Fielding

Filter: AL | 2006 Season

Name	Team	G	GS	W	L	SV	CG	SHO	IP	H	R	ER	HR	BB	K	ERA	WHIP	BAA
Johan Santana	MIN	34	34	19	6	0	1	0	233.2	186	79	72	24	47	245	2.77	1.00	.216
Chien-Ming Wang	NYY	34	33	19	6	1	2	1	218.0	233	92	88	12	52	76	3.63	1.31	.277
Jon Garland	CWS	33	32	18	7	0	1	1	211.1	247	112	106	26	41	112	4.51	1.36	.294
Kenny Rogers	DET	34	33	17	8	0	0	0	204.0	195	97	87	23	62	99	3.84	1.26	.253
Justin Verlander	DET	30	30	17	9	0	1	1	186.0	187	78	75	21	60	124	3.63	1.33	.266
Freddy García	PHI	33	33	17	9	0	1	0	216.1	228	116	109	32	48	135	4.53	1.28	.267
Randy Johnson	ARI	33	33	17	11	0	2	0	205.0	194	125	114	28	60	172	5.00	1.24	.250
Kevin Millwood	TEX	34	34	16	12	0	2	0	215.0	228	114	108	23	53	157	4.52	1.31	.272
Ervin Santana	LAA	33	33	16	8	0	0	0	204.0	181	106	97	21	70	141	4.28	1.23	.241
Josh Beckett	BOS	33	33	16	11	0	0	0	204.2	191	120	114	36	74	158	5.01	1.29	.245

註：WHIP代表Walk & Hit per Inning Pitched, 一般1≤WHIP≤1.75。

Sortable Stats: Pitching

League Leaders | Stats by Position

Batting | Pitching | Fielding

Filter: AL | 2007 Season

Name	Team	G	GS	W	L	SV	CG	SHO	IP	H	R	ER	HR	BB	K	ERA	WHIP	BAA
Josh Beckett	BOS	29	29	20	6	0	1	0	194.2	179	71	68	15	40	188	3.14	1.13	.240
Chien-Ming Wang	NYY	30	30	19	7	0	1	0	199.1	199	84	82	9	59	104	3.70	1.29	.265
Fausto Carmona	CLE	32	32	19	8	0	2	1	215.0	199	78	73	16	61	137	3.06	1.21	.248
C.C. Sabathia	CLE	33	33	18	7	0	4	1	234.0	230	91	83	19	36	205	3.19	1.14	.258
Justin Verlander	DET	31	31	18	6	0	1	1	195.2	177	86	80	19	65	176	3.68	1.24	.235
John Lackey	LAA	32	32	18	9	0	2	2	217.0	217	87	75	18	52	177	3.11	1.24	.258
Kelvim Escobar	LAA	29	29	17	7	0	3	1	189.2	177	78	73	11	64	156	3.46	1.27	.249
Tim Wakefield	BOS	30	30	16	12	0	0	0	182.0	185	100	97	20	64	109	4.80	1.37	.265
Roy Halladay	TOR	31	31	16	7	0	7	1	225.1	232	101	93	15	48	139	3.71	1.24	.268
Johan Santana	MIN	33	33	15	13	0	1	1	219.0	183	88	81	33	52	235	3.33	1.07	.225

Sortable Stats: Batting

League Leaders | Stats by Position

AVG Leaders: Minimum 490 plate appearances

Batting Pitching Fielding

Filter: Qualified leaders | AL | 2007 Season

Name	Team	G	AB	R	H	2B	3B	HR	RBI	BB	K	SB	CS	AVG	OBP	SLG	OPS
Magglio Ordóñez	DET	155	587	116	211	52	0	28	136	75	79	4	1	.359	.430	.591	1.022
Ichiro Suzuki	SEA	157	662	110	232	22	7	6	68	48	75	37	7	.350	.396	.432	.828
Plácido Polanco	DET	139	577	104	196	35	3	9	66	37	29	7	3	.340	.388	.458	.845
Jorge Posada	NYY	142	500	90	168	42	1	20	89	72	97	2	0	.336	.423	.544	.967
Chone Figgins	LAA	112	434	79	146	24	6	3	58	49	78	40	12	.336	.397	.440	.837
Mike Lowell	BOS	150	573	75	187	36	2	20	116	53	69	3	2	.326	.381	.501	.882
David Ortiz	BOS	146	539	112	175	50	1	33	114	109	101	3	1	.325	.440	.605	1.045
Derek Jeter	NYY	154	631	100	203	38	3	12	71	55	97	14	8	.322	.387	.448	.836
Vladimir Guerrero	LAA	149	572	88	184	45	1	26	123	71	62	2	3	.322	.401	.540	.941
Dustin Pedroia	BOS	135	508	86	161	38	1	8	50	47	42	6	1	.317	.381	.443	.824

註：SLG＝total bases/ at bats、OPS＝OBP＋SLG。

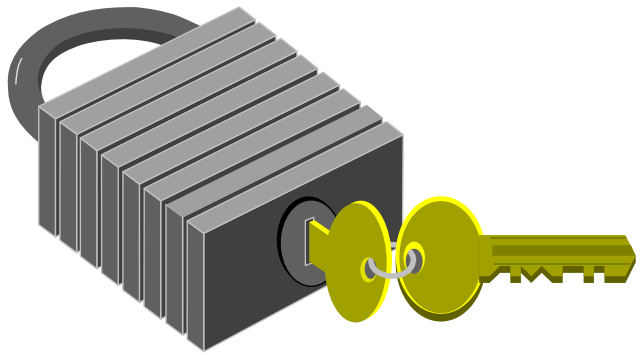
資訊與決策

- 過去擁有土地、資金等的資本家，透過勞力及資本密集而具有絕對優勢；21世紀有另一個重要轉變，將成為知識經濟的時代，早一步擁有充分資訊的人，會握有絕對優勢。(Yahoo、Google！)

→簡單地說，統計就是一門蒐集資料、整理與分析資料，協助我們做出合理判斷的一門科學。

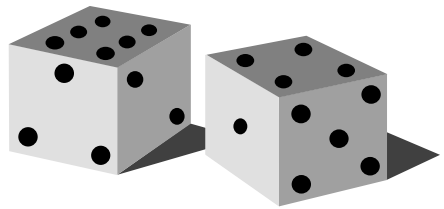
什麼是統計?

- 統計學是研究定義問題、運用資料蒐集、整理、陳示、分析與推論等科學方法，在不確定(Uncertainty)情況下，做出合理決策的科學。



學統計就像在當偵探

- 資料搜集 → 搜集線索
- 資料分析 → 思考分析
- 決策推論 → 推理判斷



統計可能的應用領域

資訊與知識的價值(資料採礦)

- 資料挖掘(Data Mining)的範例：$ $ $ $

→ 協助超級市場促銷及陳設商品

Milk, eggs, sugar, bread

Customer1

Milk, eggs, cereal, bread

Customer2

Eggs, sugar

Customer3

沃馬特量販店(Wal-Mart)

- 沃馬特最先蒐集、分析顧客資料，並以整理所得的資訊，提高銷售業績。

→分析發現美國消費者在週末購物時，許多人會同時購買尿布及啤酒。

- **問題：**為什麼這兩種商品會一起購買？又如何將這份資訊轉變為業績？

註：沃馬特從美國西南部發跡，剛開始只是一家五金行，現在已是全美最大的百貨零售業者。

數 據 (Data)

資 訊 (Information)

事 實 (Fact)

知 識 (Knowledge)

馬克吐溫對統計的想法

There are three kinds of lies:

Lies,

Damned lies,

and Statistics!!

統計可協助理性的決策(貝氏定理)

A patient takes a lab test and the result comes back positive. The test returns a correct positive result in 99% of the cases in which the disease is actually present, and a correct negative result in 98% of the cases in which the disease is not present. Furthermore, .001 of all people have this cancer.

P(cancer) =	.001	P(~ cancer) =	.999
P(+ | cancer) =	.99	P(– | cancer) =	.01
P(+ | ~cancer) =	.02	P(– | ~cancer) =	.98

$$P(\text{cancer} \mid +) = \frac{P(+ \mid cancer)\,P(cancer)}{P(+)} = \mathbf{.047}$$

計算細節：

- 假設某地區有一百萬人：

  → 999,000人健康，1,000人罹患癌症

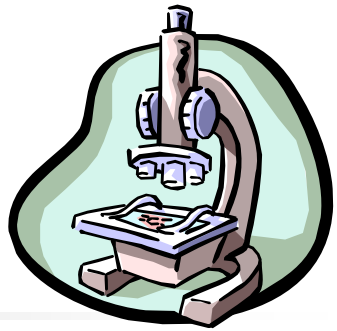
  → 檢查出陽性反應者：

  (1)健康者中有 $999{,}000 \times 2\% = 19{,}980$

  (2) 癌症患者中有 $1{,}000 \times 99\% = 990$

  因此，癌症患者佔陽性反應者的比例：

$$P(cancer \mid +) = \frac{990}{19{,}980 + 990} = \frac{990}{20{,}970} \cong 4.72\%$$



Suppose a second test for the same patient returns a positive result as well. What are the posterior probabilities for cancer?

P(cancer) =	.001	P(~cancer) =	.999
P(+ | cancer) =	.99	P(– | cancer) =	.01
P(+ | ~cancer) =	.02	P(– | ~cancer) =	.98

$$P(\text{cancer} \mid +_1 +_2) = \frac{P(+_1 +_2 \mid \text{cancer}\ )\, P(\text{cancer}\ )}{P(+_1 +_2)} = \mathbf{.710}$$

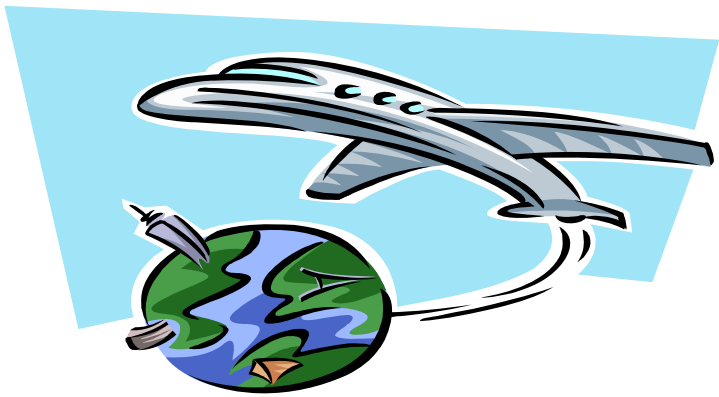
規避不必要的風險

- 1986年美國挑戰者號太空梭的爆炸

→ O形環(O-Ring)在低溫下無法正常運作，造成燃料外洩而爆炸。

→分析過去各種溫度下的失敗比例

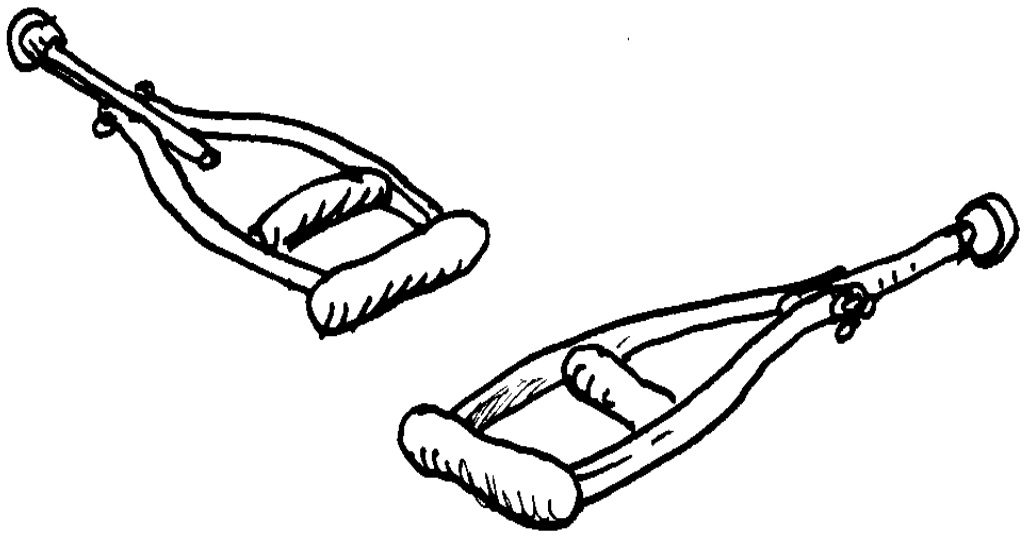
(羅吉士迴歸；logistic regression)



(參考書籍：天下文化「你管別人怎麼想」→費曼)

世界最大的醫學實驗(沙克疫苗)

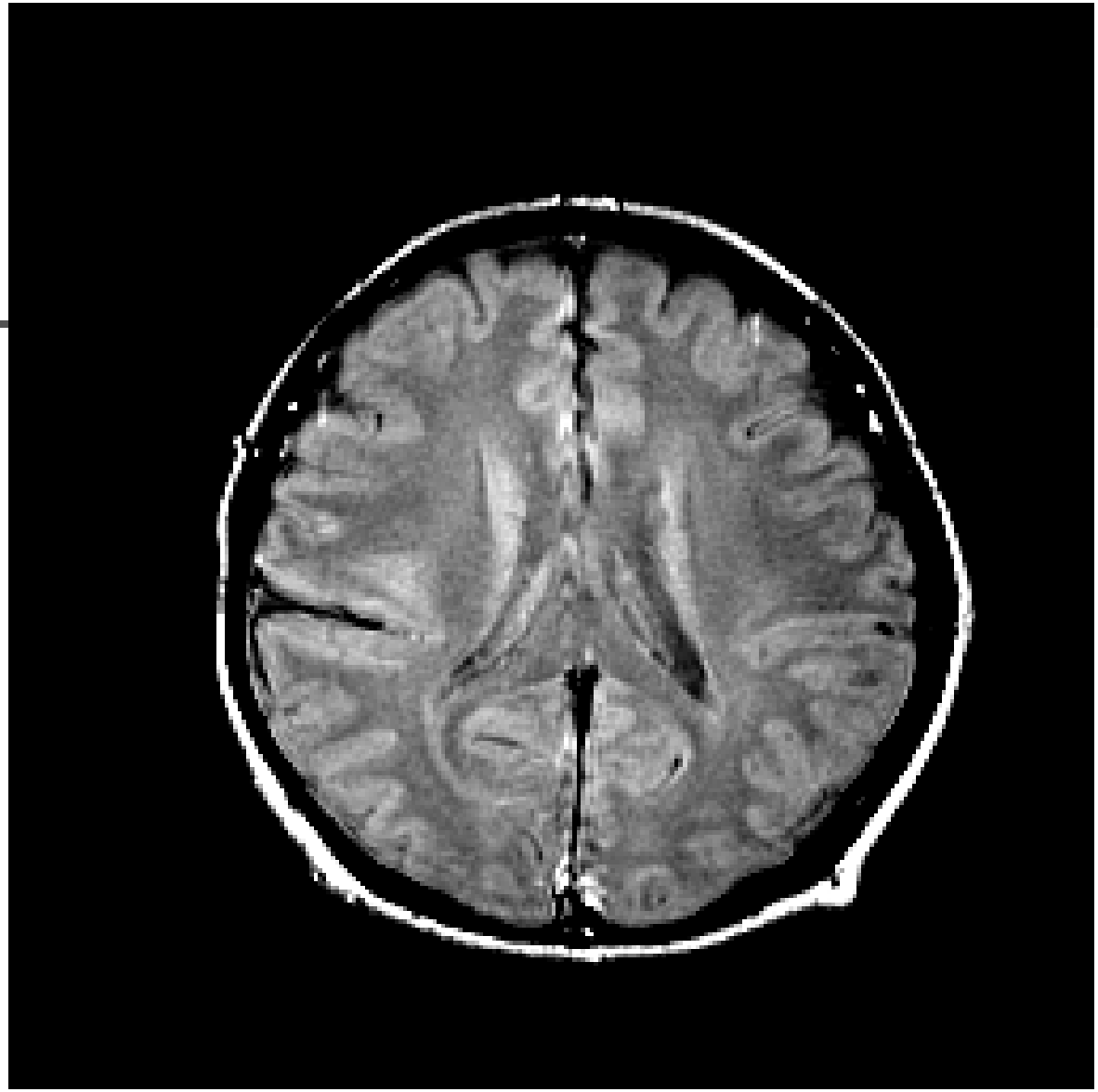
A MORE POSITIVE EXAMPLE IS THE *SALK POLIO VACCINE*. IN 1954, VACCINE TRIALS WERE PERFORMED ON SOME 400,000 CHILDREN, WITH STRICT CONTROLS TO ELIMINATE BIASED RESULTS. GOOD STATISTICAL ANALYSIS OF THE RESULTS FIRMLY ESTABLISHED THE VACCINE'S EFFECTIVENESS, AND TODAY POLIO IS ALMOST UNKNOWN.



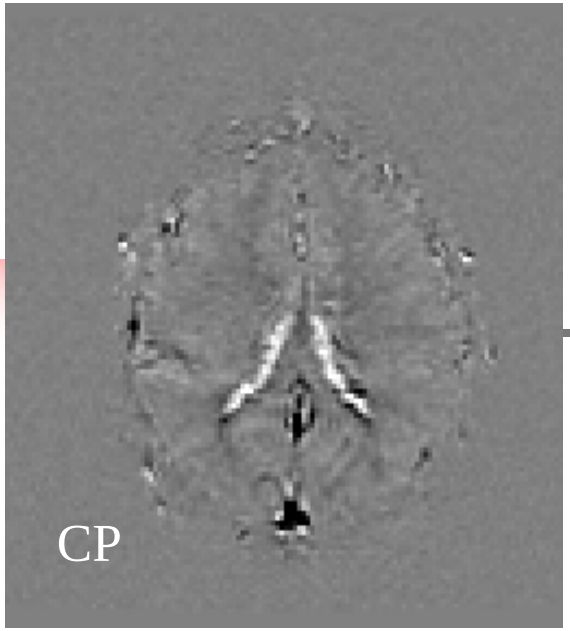
統計與醫學應用

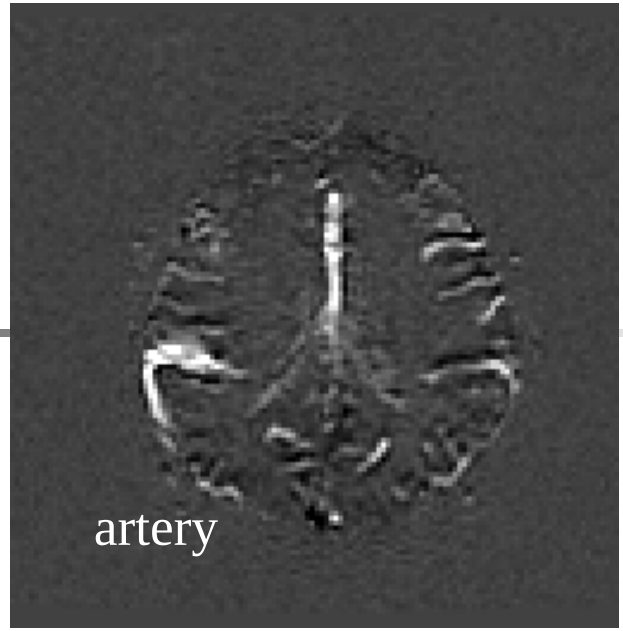
- 健康檢查時經常會抽血、驗尿、或萃取某個身體組織附近的樣本，再從檢體中判定是否罹患某種疾病。

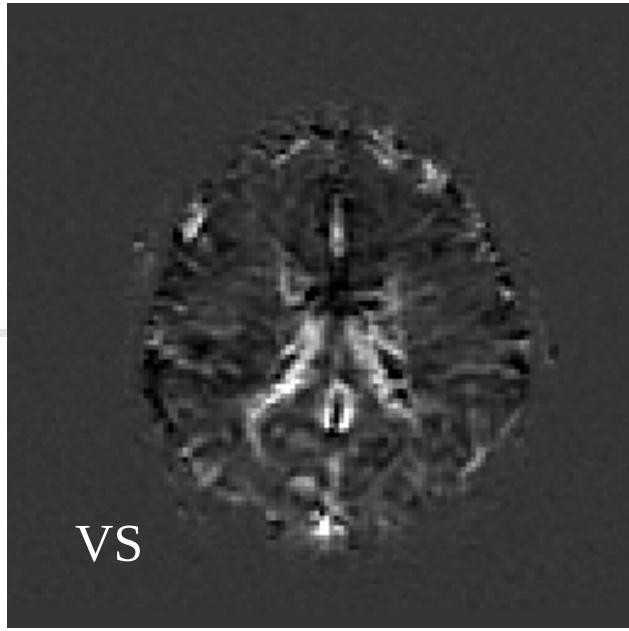
→檢查的結果以圖像表示較為清楚，例如：下圖即為中風病人的檢查結果。

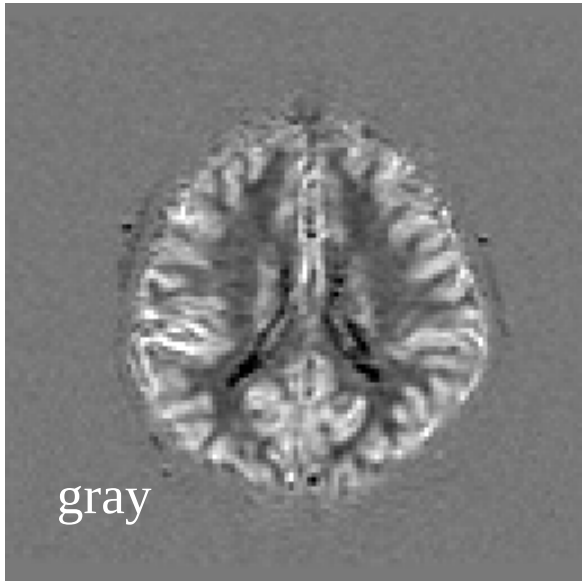


Anatomical proton-density-weighted image of human brain





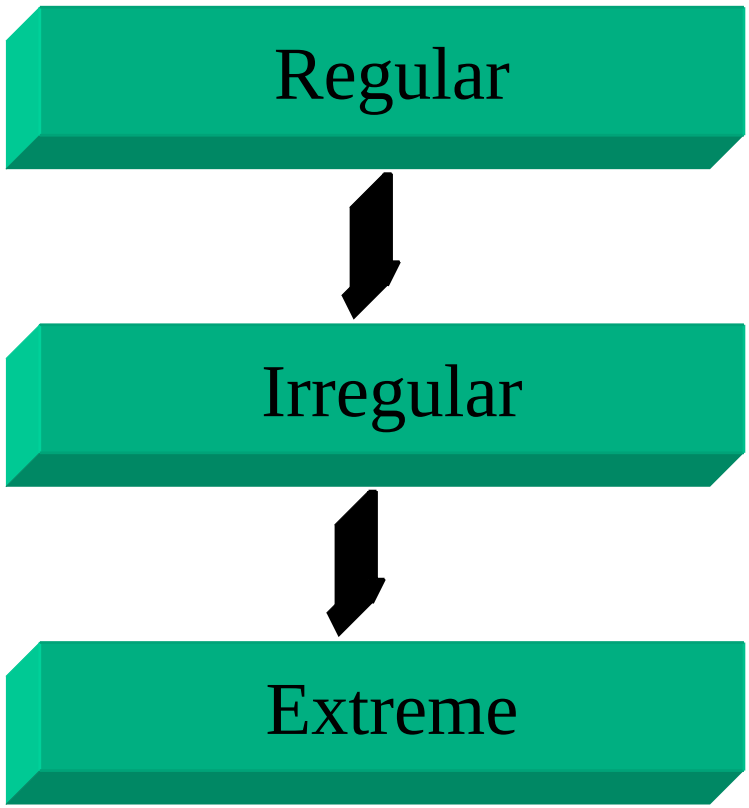




通常可分為5個圖。

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：



金融保險的範例：

- 避免信用不佳客戶的用卡核准，以減少發卡銀行損失。(預測瑕疵戶)

→舊卡戶在使用終止後是否發給新卡

→考慮發新卡的對象

- 盜刷信用卡，尋找異常(Irregularities)現象。
- 新舊銀行林立，且提供的金融商品差異不大，銀行為了尋找自己的競爭利基，需要找出可以為自己保留顧客，甚至挖掘潛在顧客的關鍵資訊。

R、F、M、T問題需求表

Factor	Method
R	最近來的比率（最近一個月來的次數）
F	一個月的平均次數 （一年來的總次數 /12）
M	平均每次交易金額（提款、轉帳）
T	存摺餘額

顧客分群表

顧客分群	顧客群組1	顧客群組2	顧客群組3	顧客群組4
問題需求	利益高 ↑ 風險低 ↓	利益低 ↓ 風險低 ↓	利益高 ↑ 風險高 ↑	利益低 ↓ 風險高 ↑
R/F/M/T	↓ ↓ ↓ ↑	↓ ↓ ↓ ↓	↑ ↑ ↑ ↑ 或 ↓ ↓ ↑ ↑	↑ ↑ ↑ ↓ 或 ↑ ↑ ↓ ↓
說明	此為最好的顧客	此為忠誠的顧客，但其價值低	此顧客對企業價值高，相對地風險也高，易變成別人的顧客	對銀行的利益很少
方法	保留住，並增加對銀行價值	保留住，提升其對銀行價值	盡量把顧客保留住	花少量的投資於此

- 問題：以你/妳的角度而言，在審查信用卡申請，或是判斷信用卡是否遭到盜刷，如何設計整個程序？

→觀念：正常 vs.異常

→資料：有哪些個人資料需要蒐集，哪些資料的優先順序較高？

→分析：如何判斷現在出現的結果確實是「異常」，而非巧合？

→決策：如果分析的結果有商議的空間，又該如何處理？

如何學好統計？

如何學好統計？

- 統計、經濟、會計等基礎科目，之所以讓大家覺得頭痛，主因之一在於內容變化較多，需要思考、記憶才能熟練。
- 問題：學習是否可以只靠理解？

→以九九乘法表(記憶)與建構式數學(思考)為例，兩者各有優缺點，兩者適度的搭配更能相得益彰。建議先瞭解內容，加上適量的背誦，可以有較好的效果。

如何自我提升？

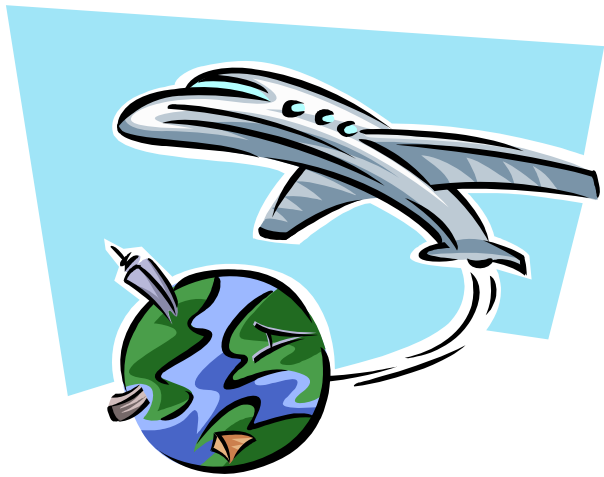
- 知識的交流具有教學相長的特質，投入愈多、回收也愈多，也不會因為分享心得而失去知識，反而因此會累積地更多、更廣！

→「打破沙鍋問到底！」

→「處處留心皆學問！」

→「他山之石可以攻錯！」

→「你管別人怎麼想！」



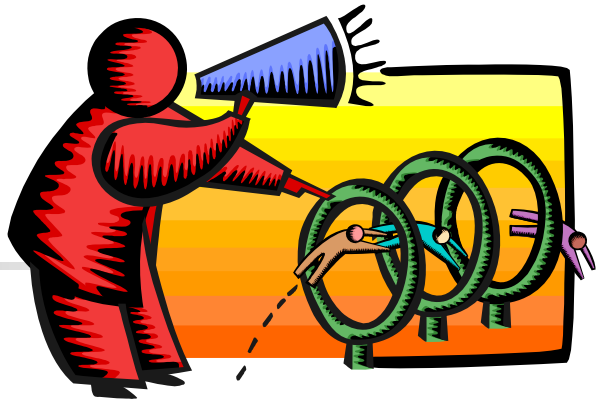
定義問題

蒐集資料

分析資料

詮釋結果

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）

範例：汽車大保養的費用

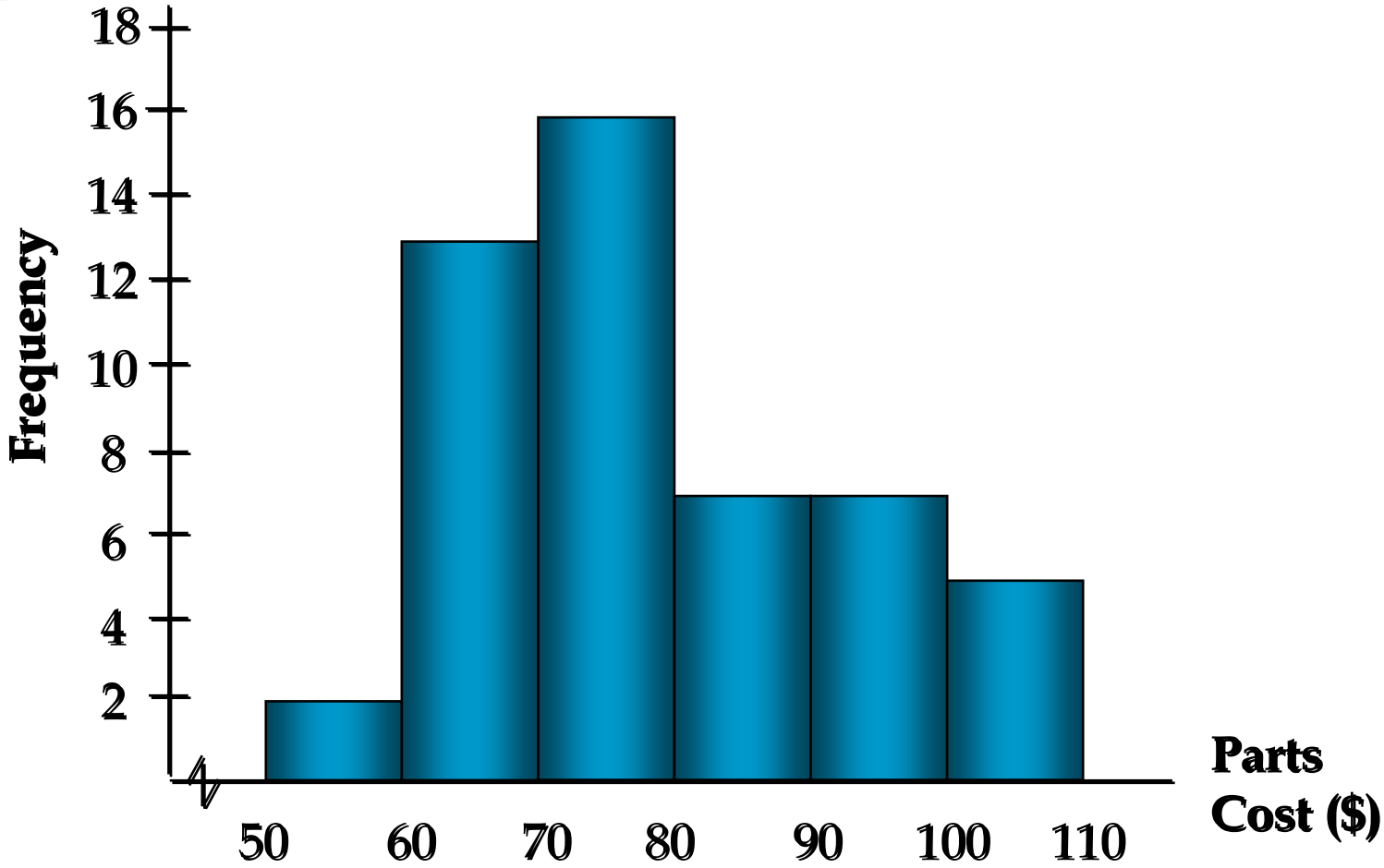
→某汽車修理廠統計上個月50位客戶，汽車大保養的所需費用，發現(單位：百元)

平均數＝79.0、中位數＝75.5、

標準差＝14.0。

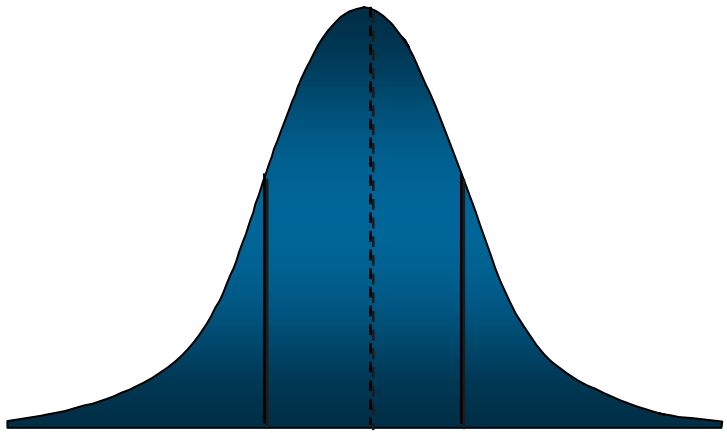
91	78	93	57	75	52	99	80	97	62
71	69	72	89	66	75	79	75	72	76
104	74	62	68	97	105	77	65	80	109
85	97	88	68	83	68	71	69	67	74
62	82	98	101	79	105	79	69	62	73

直方圖(Histogram)



經驗法則(Empirical Rule)

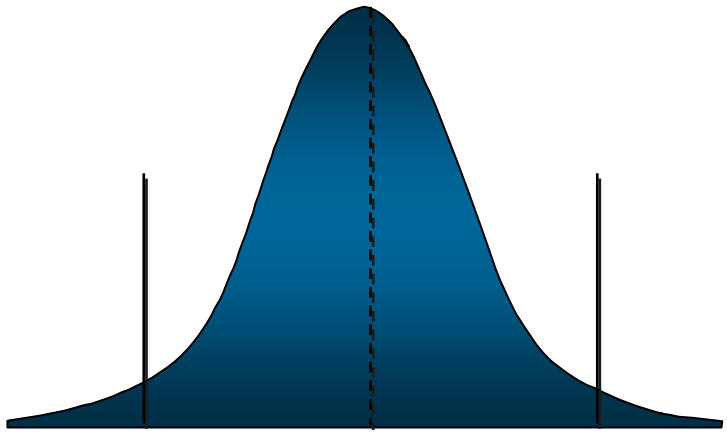
資料的分佈比例為鐘型(bell-shaped)曲線：



→大約有 **68%** 的資料與期望值(Mean)距離在一個標準差(**standard deviation**)之內。

經驗法則(Empirical Rule)

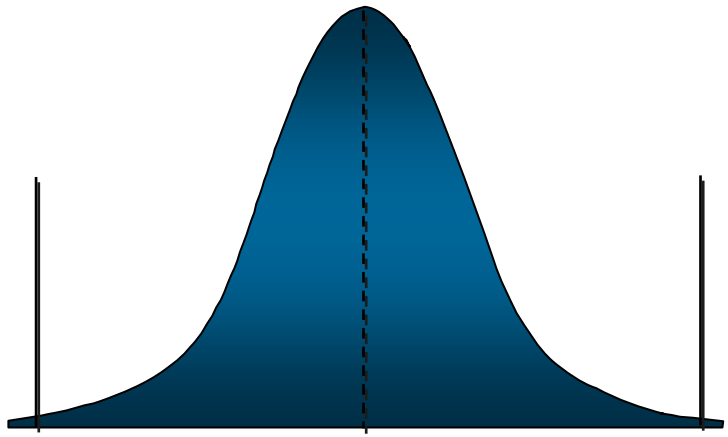
資料的分佈比例為鐘型(bell-shaped)曲線：



→大約有 **95%** 的資料與期望值(Mean)距離在兩個標準差(**standard deviation**)之內。

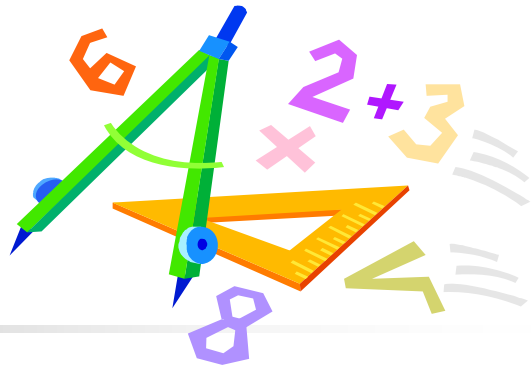
經驗法則(Empirical Rule)

資料的分佈比例為鐘型(bell-shaped)曲線：



→大約幾乎全部資料(**99.7%**) 的資料與期望值(Mean)距離在三個標準差之內。

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。

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。

如何藉由統計獲取資訊？

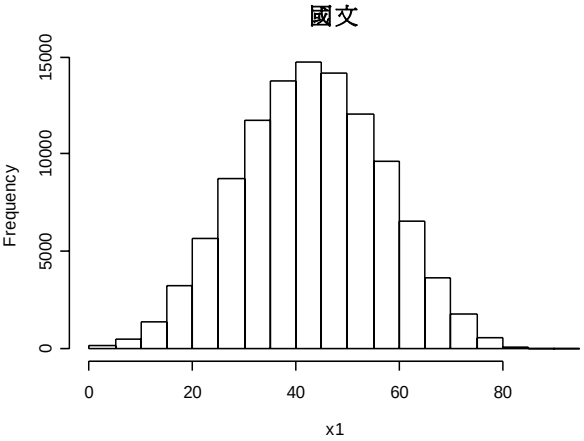
- 如果想瞭解民國94年指定考試各科的特性，可以借助哪些工具？

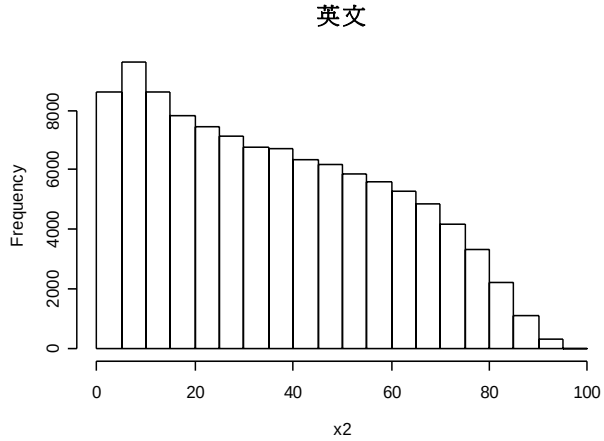
→例如：那一科的分數最不平均，像是哪一科大多數人都考得不好，只有少數人分數分高。

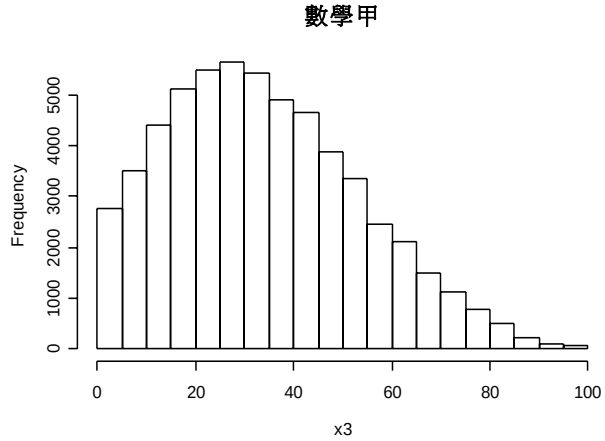
→平均數明顯大於中位數，稱為右偏(skewed to the right)；反之，若平均數明顯小於中位數，稱為左偏(skewed to the left)。平均數等於中位數，則為兩側對稱。

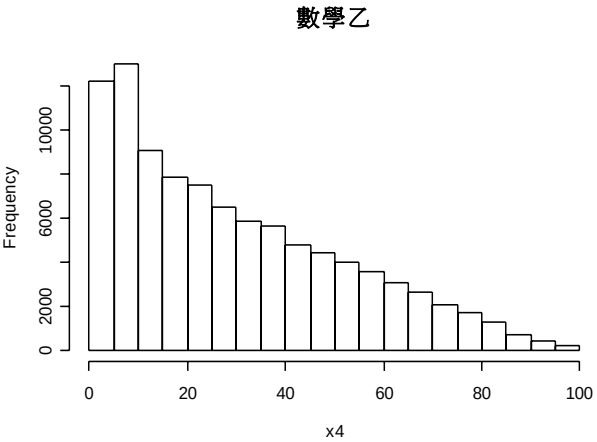
民國 94 年大學指定考試各科成績

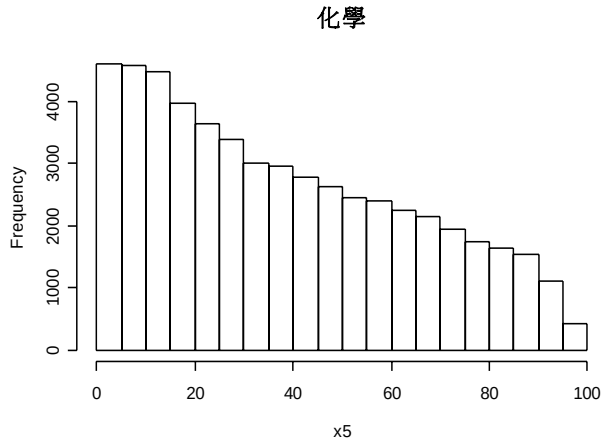
	國文	英文	數學甲	數學乙	化學	物理	生物	歷史	地理
Min.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0	0.00
12%	27.00	8.00	11.00	4.00	8.00	6.00	22.00	13.0	18.00
1st Qu.	34.00	16.00	22.00	12.00	15.00	12.00	32.00	28.0	30.00
Median	44.00	34.00	34.00	29.00	34.00	23.00	45.00	39.0	39.00
Mean	43.56	36.68	36.36	34.36	38.88	28.75	46.16	38.7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

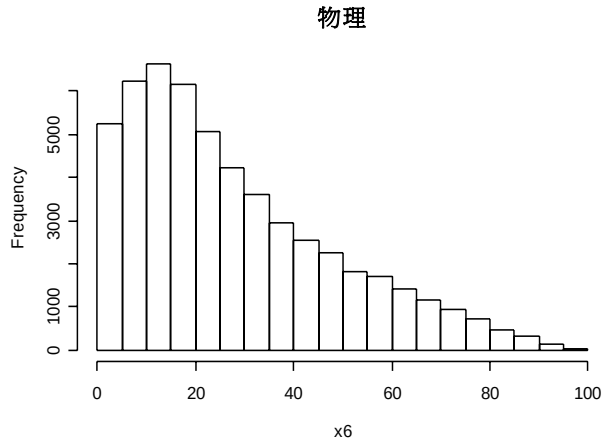

國文
Frequency
0
5000
10000
15000
0
20
40
60
80
x1

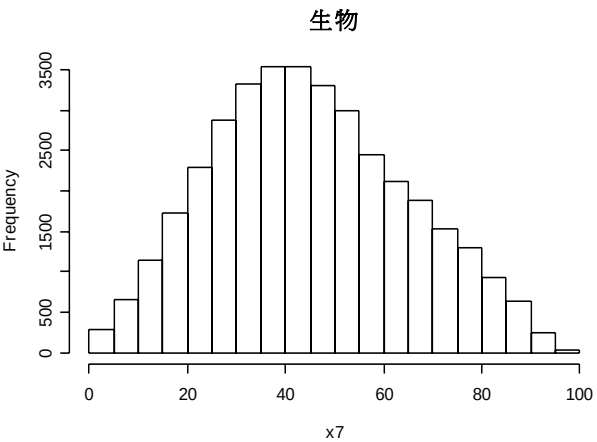

英文
Frequency
0
2000
4000
6000
8000
0
20
40
60
80
100
x2

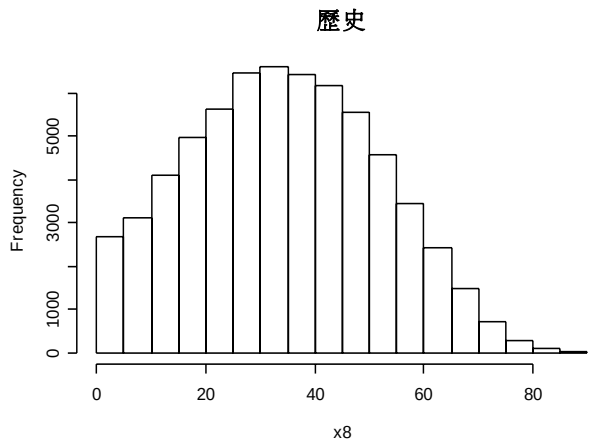

數學甲
Frequency
0
1000
2000
3000
4000
5000
0
20
40
60
80
100
x3

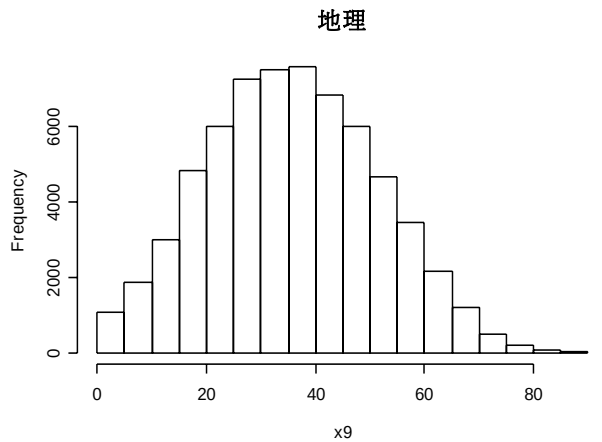

數學乙
Frequency
0
2000
6000
10000
0
20
40
60
80
100
x4

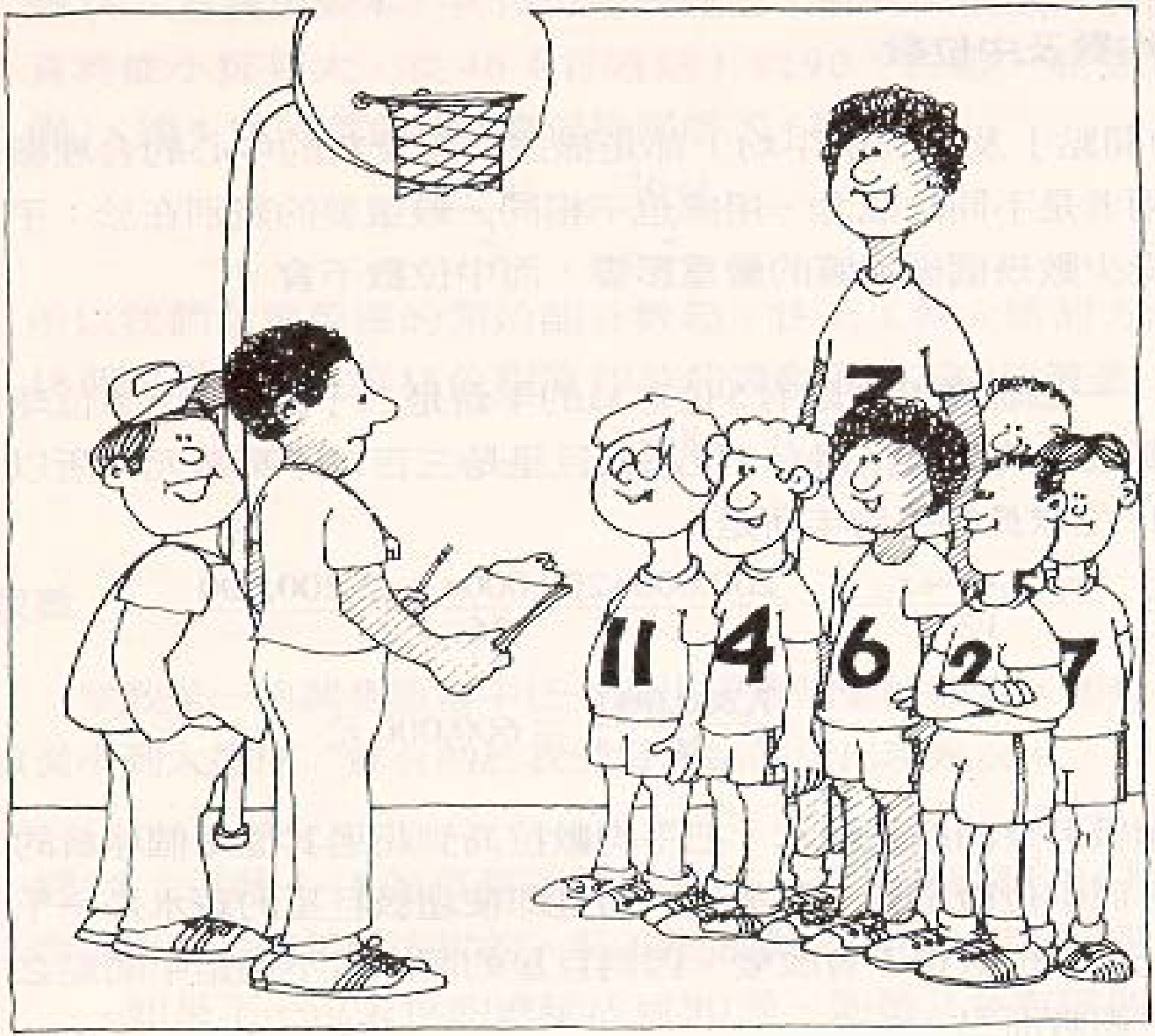

化學
Frequency
0
1000
2000
3000
4000
0
20
40
60
80
100
x5


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

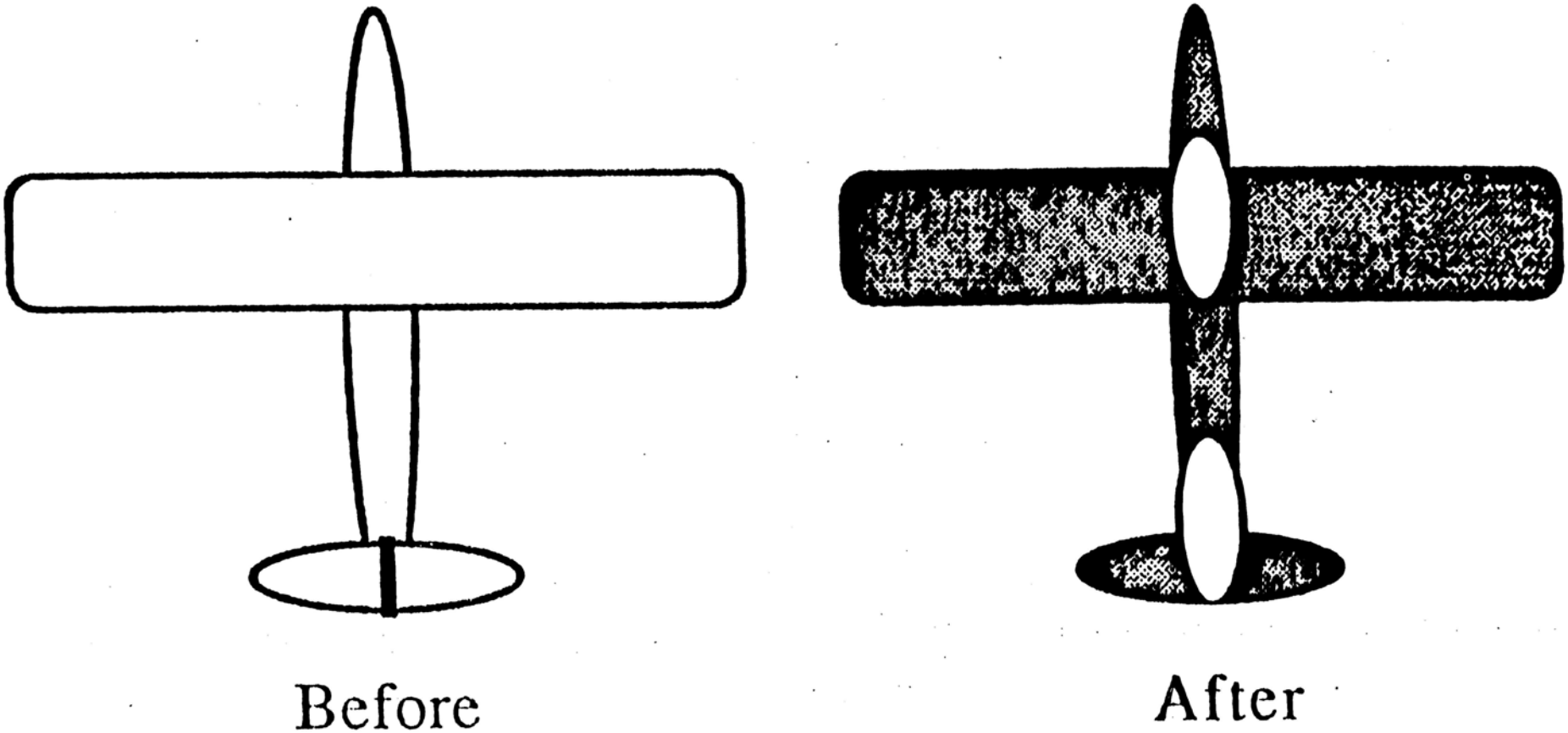

生物
Frequency
0
500
1500
2500
3500
0
20
40
60
80
100
x7


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



「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？」



A graphical depiction of ***Wald's bullethole*** *data.*

- 統計經常借重圖形、表格，提供我們重要的訊息，就如英文常見的一句話：「A picture is worth a thousand words!」
- 如果學習時遇到問題，建議你/妳先瞭解定義，再思考這些定義原先針對哪些問題設計，硬記死背頂多只能應付考試。
- 當然，平常一點一滴的累積勝過「臨時抱佛腳」，許多觀念及計算操作，需要時間才能熟練。

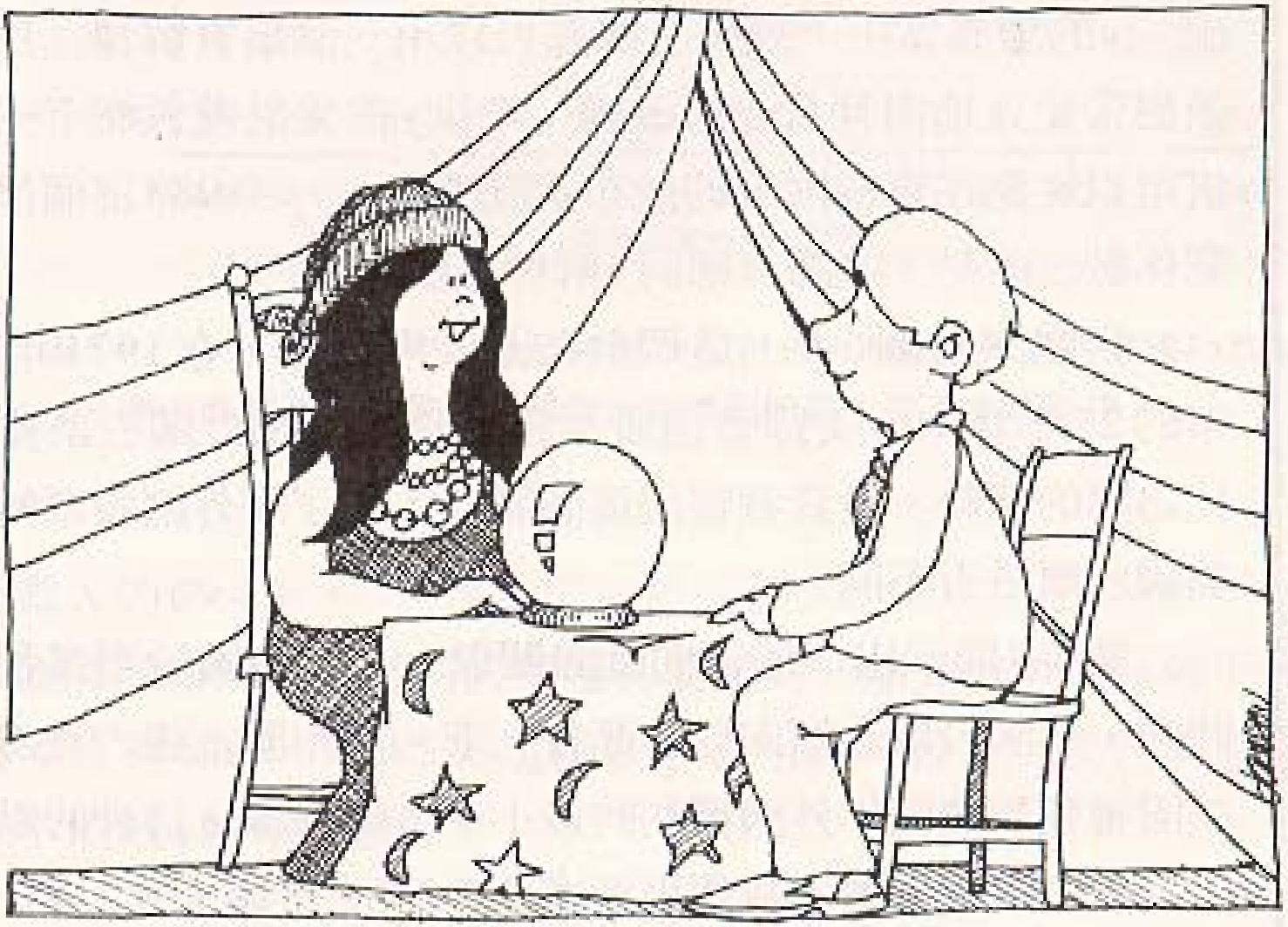
善用學習資源

- 除了每週的上課外，政治大學的統計學(及基礎科目)還有幫助同學的服務：

→每門課都有一至兩名助教，有實習課及Office Hour解決同學的問題。

→校方設置「學習資源中心」(Tutor Center)以定時、網路、分組等各種方式，協助同學基礎科目的學習。

註：同學可善用網路及搜尋引擎(Google)，解決心中的疑問。



「我怎麼會從事這一行的？是這樣的，讀大學的時候我搞不懂迴歸和相關係數，所以只好來做這種預測啦。」

問題：占星術及紫微斗數的根據？

幾本有趣的參考書籍

- 看漫畫，學統計(2003)，天下文化出版。
  →The Cartoon Guide to Statistics (1993)
- 統計，讓數字說話(1998)，鄭惟厚翻譯，天下文化出版。
- 哈佛經驗：如何讀大學－菁英學生暢談怎樣善用大學資源(2002)，立緒出版。
- 你管別人怎麼想：科學奇才費曼博士(2005)，天下文化出版。

報告完畢，

祝大家有好的開始！