統計實務

Spring 2024

授課教師：統計系 余清祥

日期：2024年2月21日

第一週：課程介紹

本課程的目標

https://leverageedu.com/blog/application-of-statistics/

- 介紹以統計思維解決實務問題。

→例如：變數間的關聯經常是研究議題，藉此提高利潤與管理績效。

- 統計不僅是一門資料分析的科學，也是一門決策科學，能協助訂定正確決策。

→未來情況視為隨機事件，統計決策提供在不確定性時的處理原理和方法，在各領域經營決策中有廣泛應用。（*MBA智庫百科*）

為什麼學統計？

https://www.eapfoundation.com/writing/essays/problemsolution/

- 當我們遭遇問題時，如何做出選擇及判斷？

→政大附近的餐廳、購物及購書的選擇、畢業後的生涯規劃？（資訊來源及可信度）

→學校篩選申請者、工作過濾求職履歷、銀行核准貸款及信用卡（推薦及評分系統）

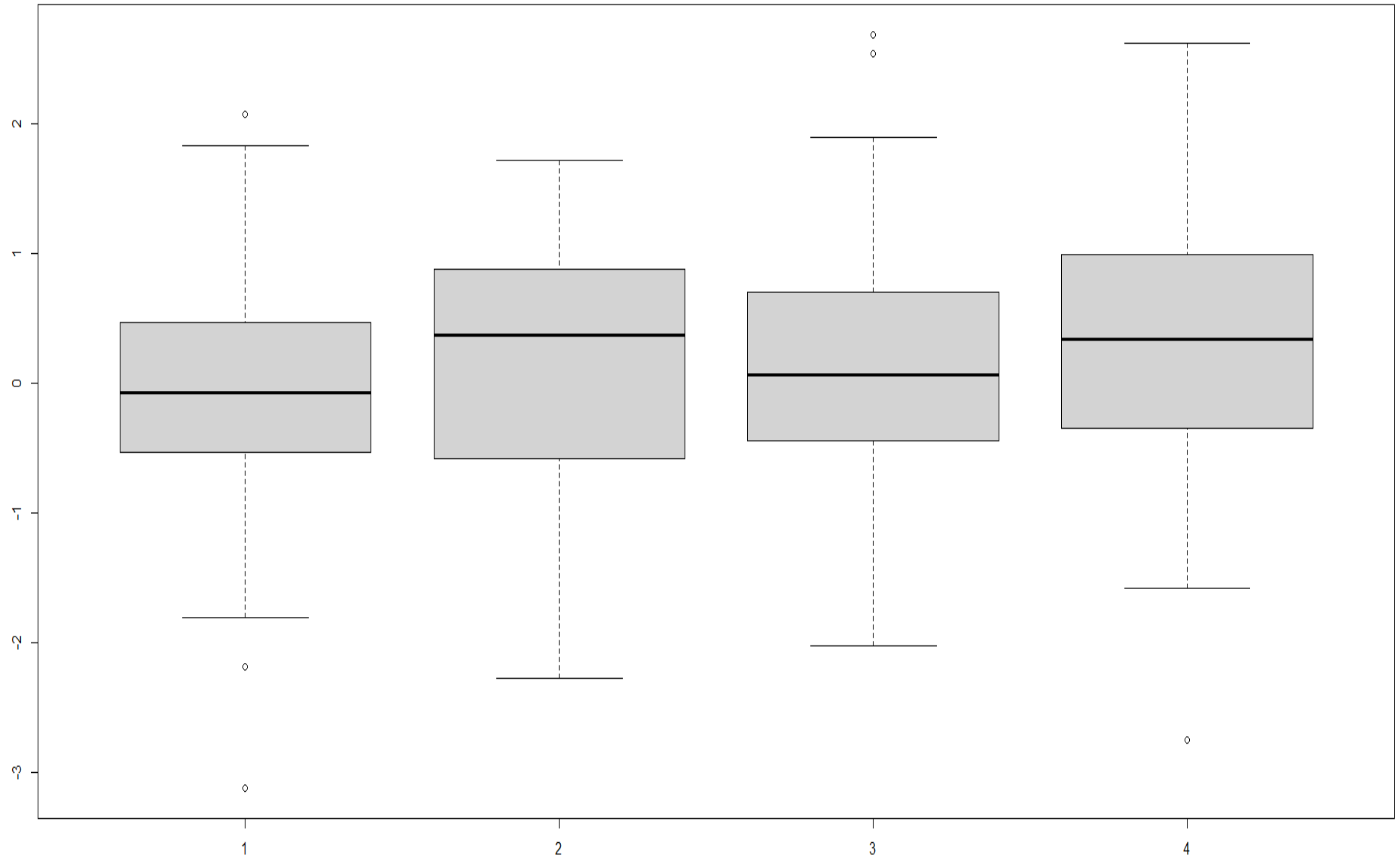
→選擇錄取的學校或公司、總統大選民調的誤差範圍是3%或6%？（解讀結果及決策）

統計思維

- **Statistical thinking** is based on **probabilistic thinking**. Learn to see probabilities everywhere and consider how you'd calculate them. Get rid of deterministic thinking altogether. Deterministic thinking is encapsulated by the phrase "*everything happens for a reason*." Probabilistic thinking's motto is "*stuff happens*." If you arise in the morning with a head cold, don't try to identify who gave it to you. It's a probabilistic event, so treat it as such. Dropping a pencil, stopping at an ATM that is out of order, finding a prime parking spot, all are probabilities. See all things in your life as events. ---*by Tim Altom*

來源：https://www.quora.com/What-are-some-examples-of-statistical-thinking

統計思維的範例



分析結果

- 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「$\mu_1 > \mu_2$」？

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

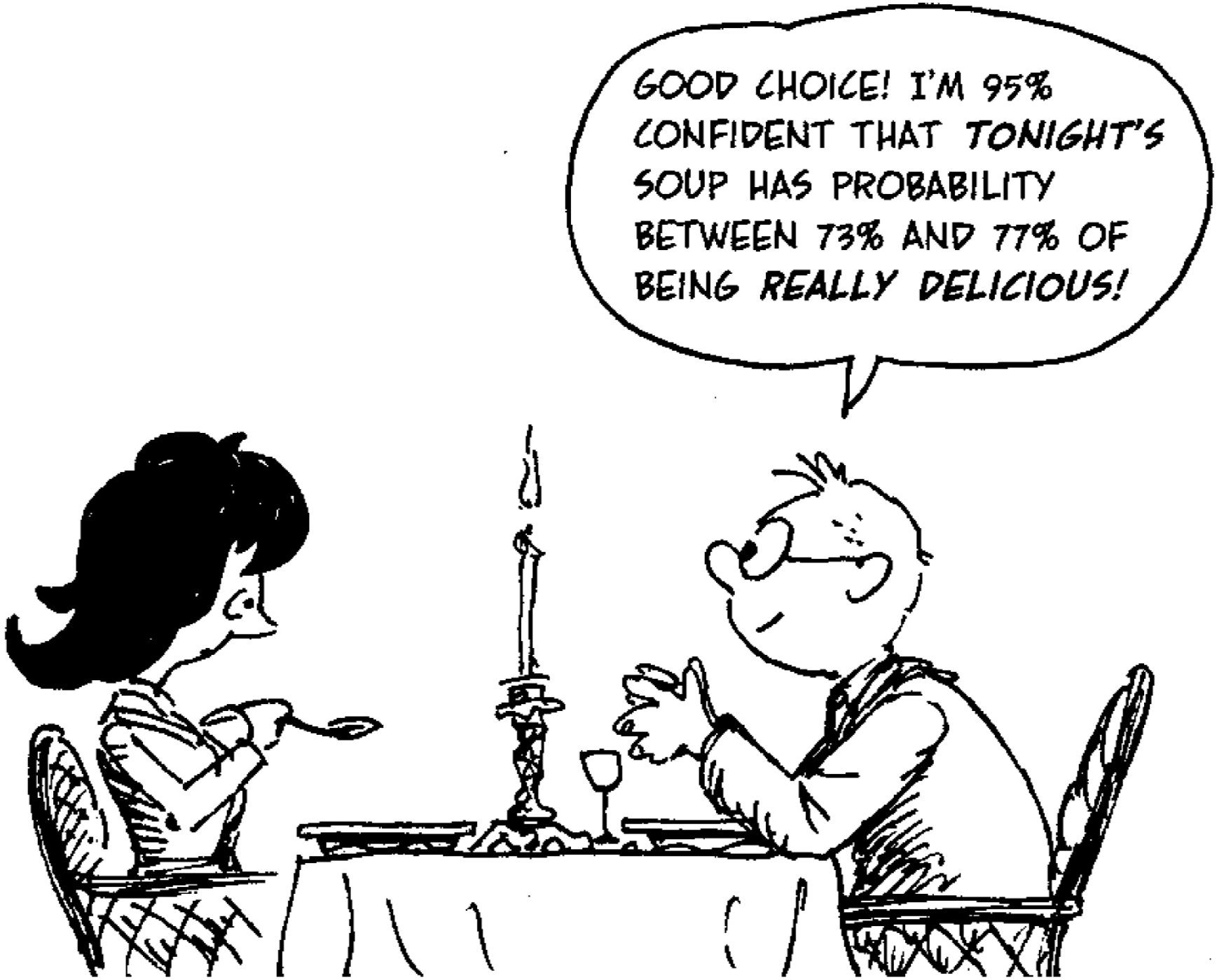
11/8 求真

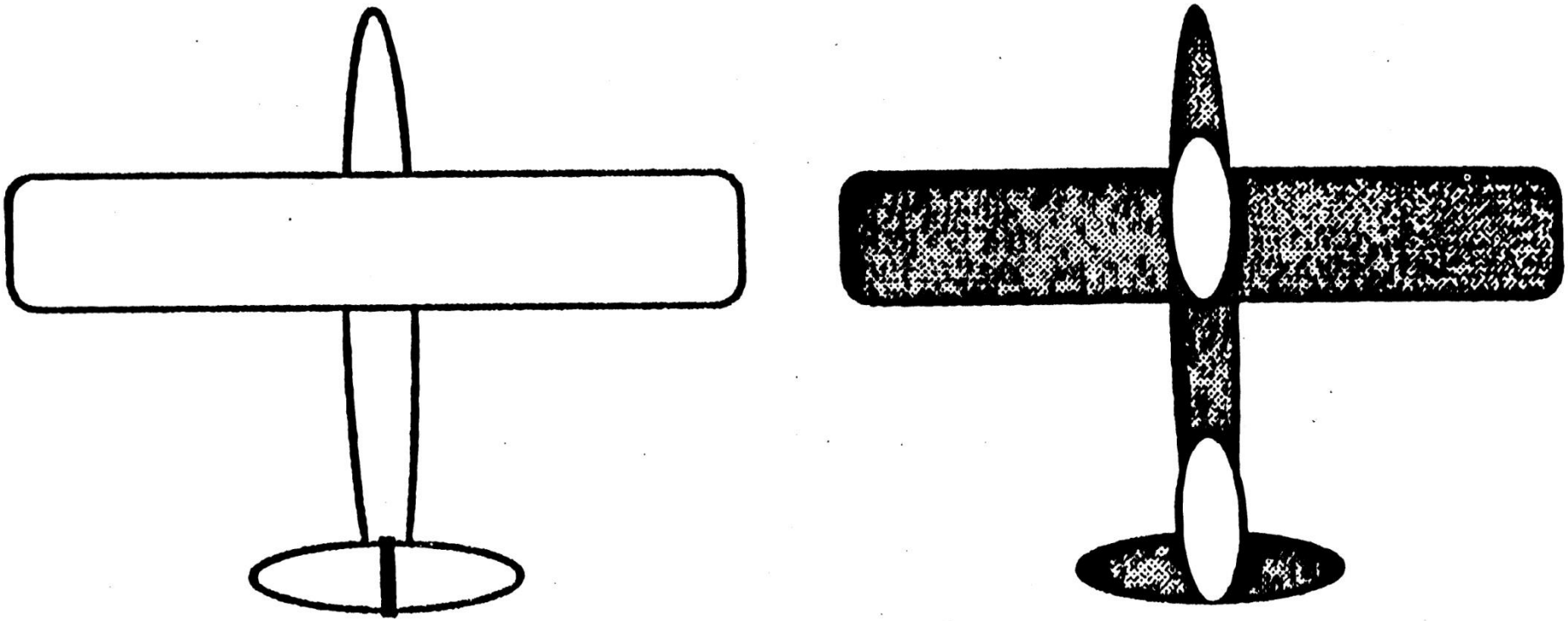
柯侯配+1

11/7 美麗島67波

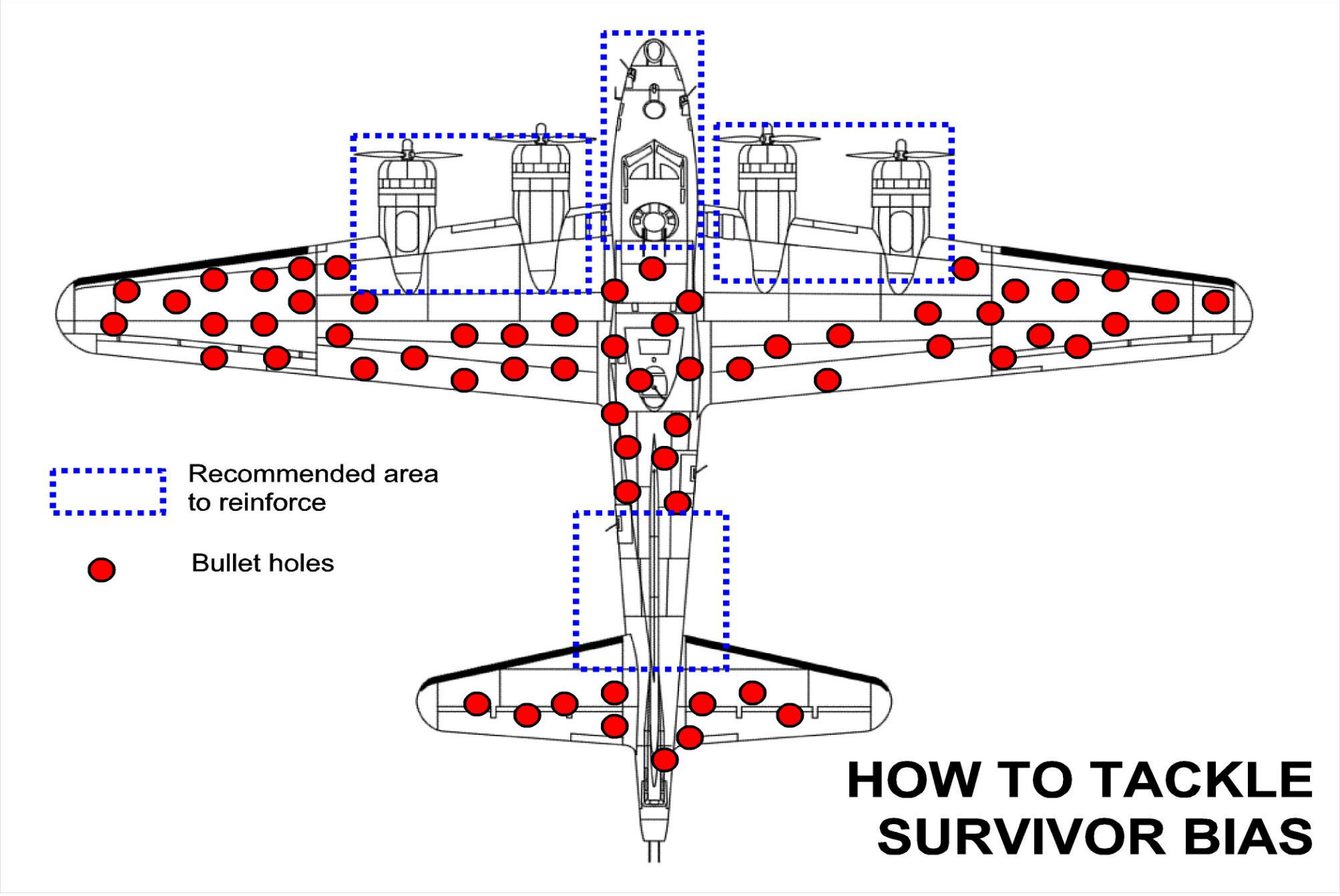
重複，採納最新版11/15

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


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



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

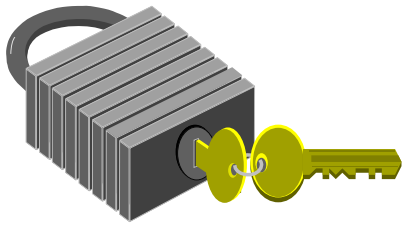


倖存者偏差（Survivorship Bias）

https://en.wikipedia.org/wiki/File:Hu_Shih_1960_color.jpg

什麼是統計？

- 統計學是在資料分析的基礎上，研究測定、收集、整理、歸納和分析反映數據資料，以便給出正確訊息的科學。（*維基百科*）
- 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。（*我的詮釋！*）



統計—21世紀的明星產業

https://www.facebook.com/Dironetv/

- 大數據的興起（IBM, 2010），激盪出新一波知識革命及產業。

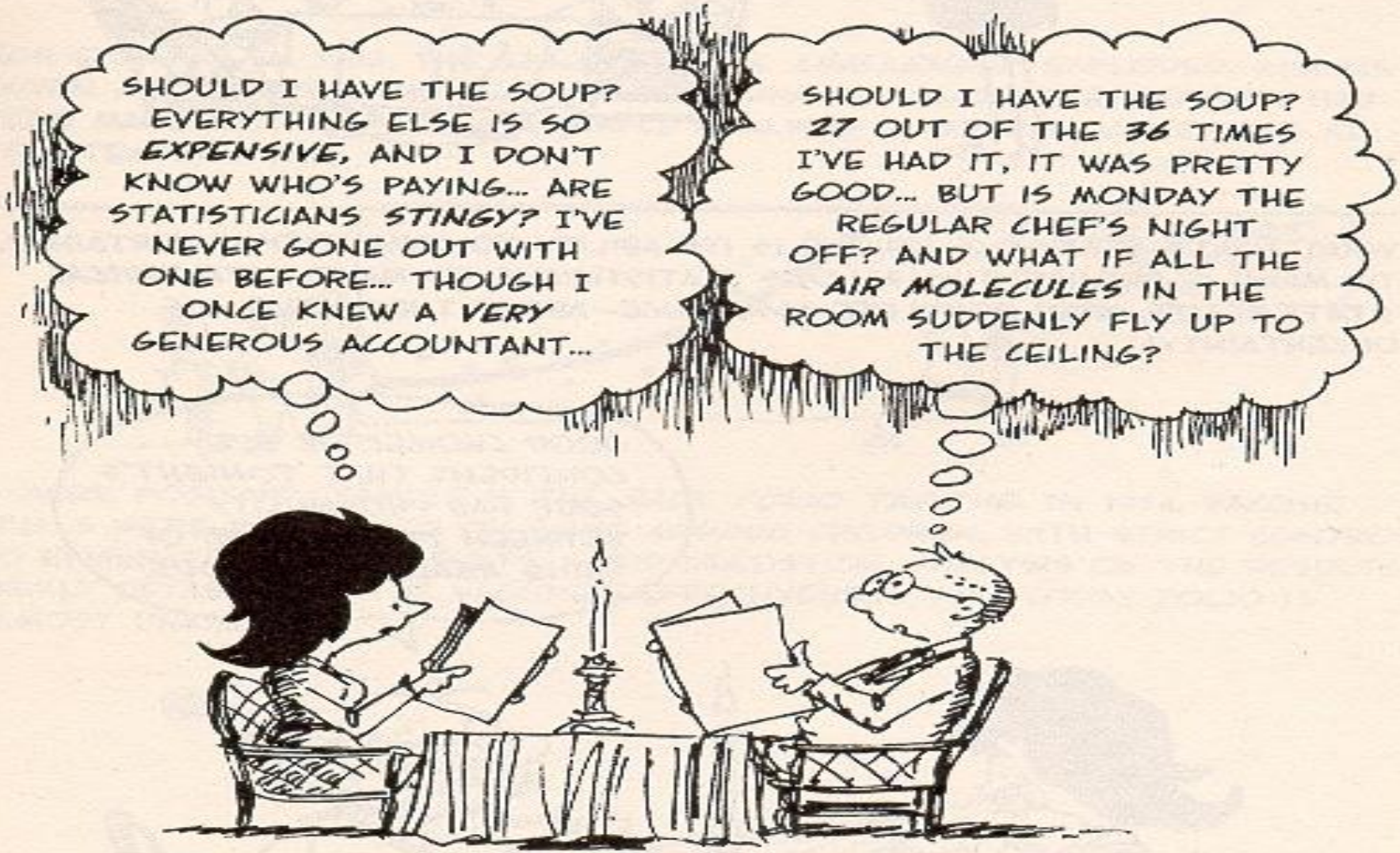
→大數據似乎無所不在（無所不能），但其中也存在許多迷思。（**請同學舉例？**）

- 問題：在大數據時代裡，統計憑藉哪些本領、可以創造哪些價值？

→統計專業人員必須具備哪些技能與知識，以因應大數據帶來的挑戰？

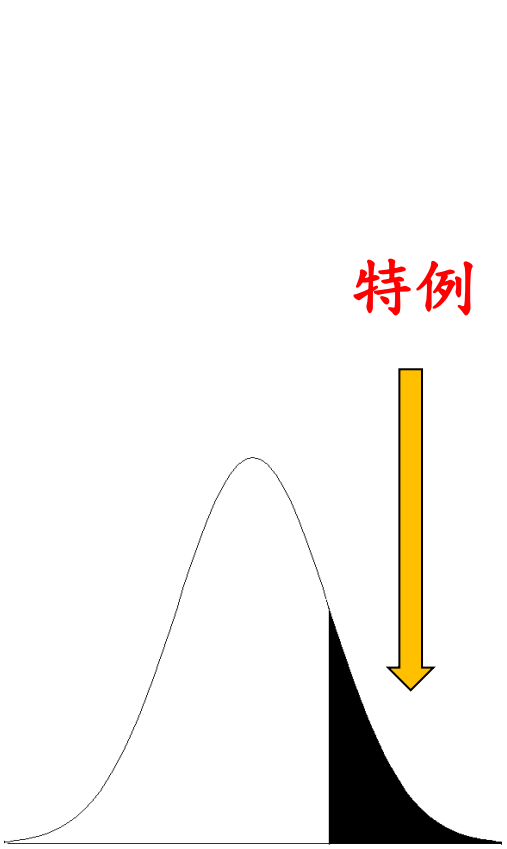
WHAT IS STATISTICS?

WE MUDDLE THROUGH LIFE MAKING CHOICES BASED ON *INCOMPLETE INFORMATION*...



統計與知識

- 統計分析屬於歸納法(Induction)，從龐雜資料找出共同趨勢，區分資料哪一種特性：



Regular — 一般、常見

Irregular — 不尋常(異常)

離群值(Outlier)

Extreme — 罕見、極端

數 據 (Data)

資 訊 (Information)

事 實 (Fact)

知 識 (Knowledge)

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？

馬克吐溫對統計的想法

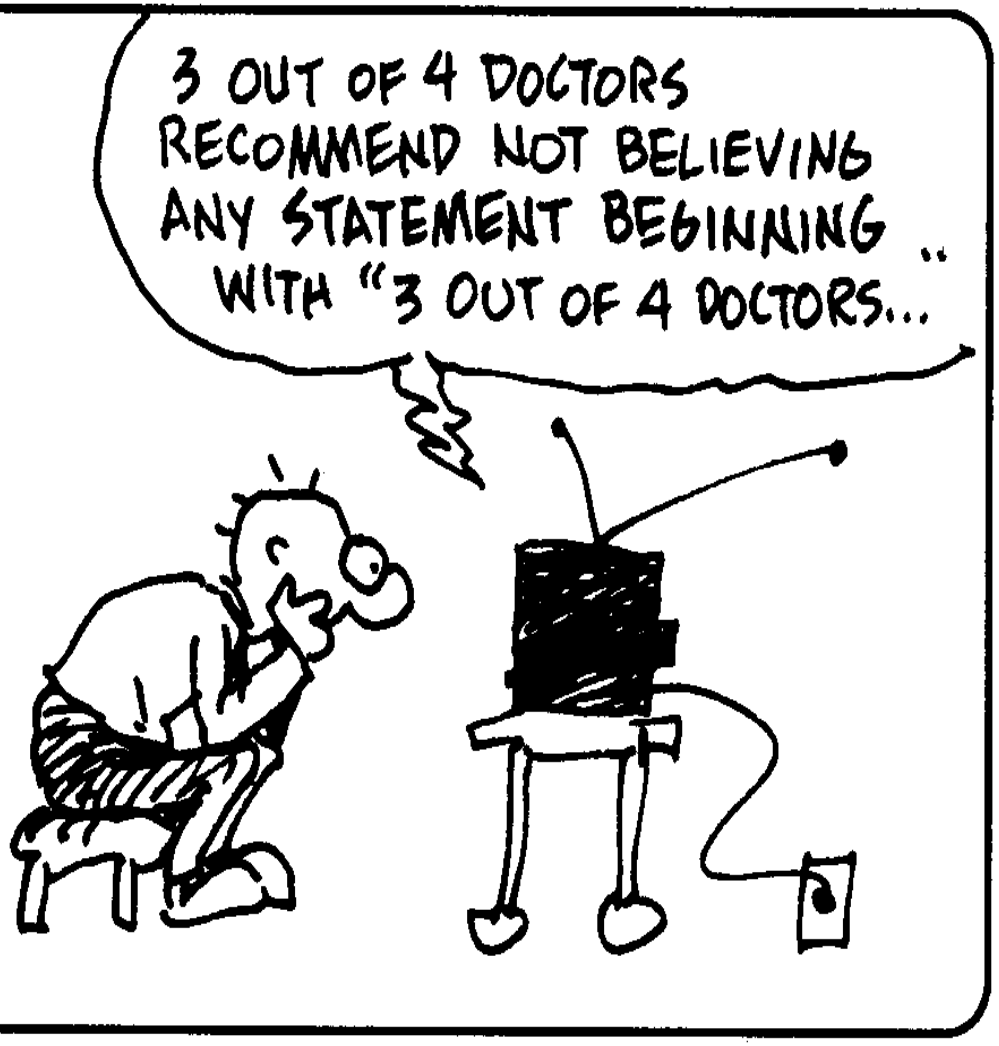
There are three kinds of lies:

Lies,

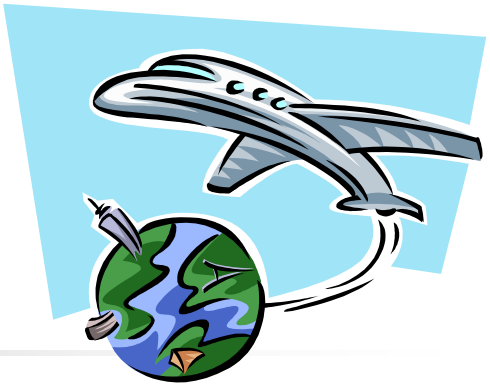
Damned lies,

and Statistics!!

FINALLY, IN DISCUSSING STATISTICS, IT'S HARD TO AVOID MENTIONING ONE OTHER THING: THE WIDESPREAD *MISTRUST* OF STATISTICS IN THE WORLD TODAY. EVERYONE KNOWS ABOUT "LYING WITH STATISTICS," WHILE GOOD STATISTICAL ANALYSIS IS NEARLY IMPOSSIBLE TO FIND IN DAILY LIFE. WHAT'S ONE TO DO?



如何解決問題？



- 解決問題(Problem Solving)的訓練，關鍵因素在於問題定義(Problem Definition)。

→《你拿什麼定義自己？：組織大師韓第的生命故事》

- 定義問題，需要背景相關知識及協助判斷的資訊，探索性資料分析可提供協助。

→《統計，改變了世界》

→《統計能為你做些什麼》

解決問題(Problem Solving)的流程

定義問題

↓

蒐集資料

↓

分析資料

↓

詮釋結果

絕大多數的
統計教學重心

資訊與知識的價值

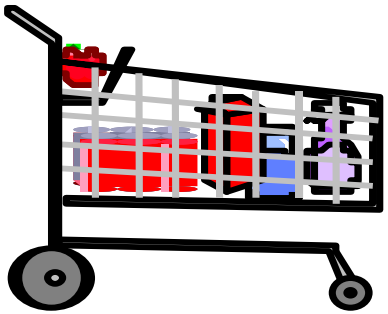
- 資料挖掘(Data Mining)的範例：＄＄＄＄

→ 協助超級市場促銷及陳設商品。

Milk, eggs, sugar, bread

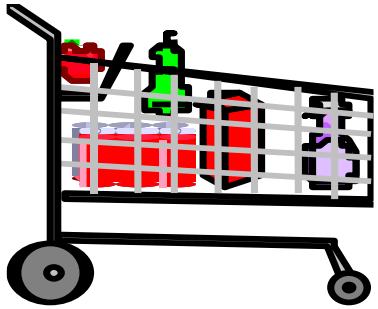
Customer1

Milk, eggs, cereal, bread



Customer2

Eggs, sugar



Customer3

https://www.pinterest.com/pin/526991593903261414/

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。

→不少美國消費者週末同時購買尿布及啤酒。

- **問題：**為什麼這兩種商品會一起購買？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，現在是全美最大百貨零售業者。

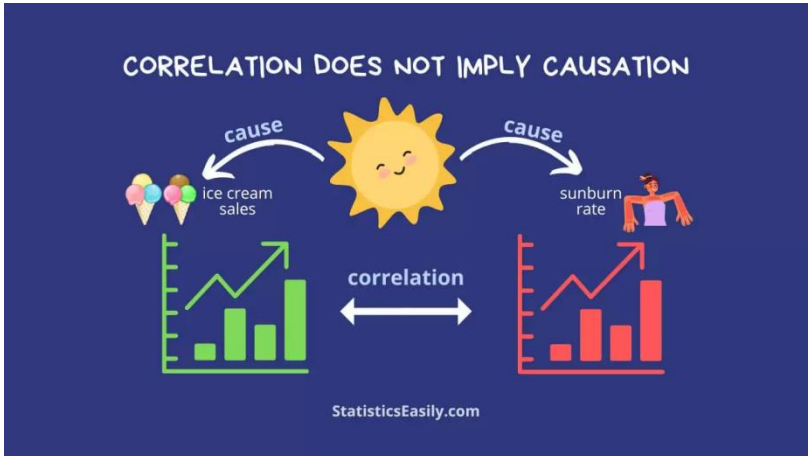
「尿布與啤酒」的延伸價值

- 問題：尿布與啤酒屬於關聯性 (Association)關係或是大家有興趣的因果關係 (Causality)？
- 關聯性的價值未必低於因果關係，像是尿布與啤酒的關連，可用於：

→商品定價與促銷；

→商品擺設（商場動線）；

→商品倉儲。



https://statisticseasily.com/correlation-vs-causality/

寶可夢旋風也能帶來商機嗎？

參考資料：https://tw.news.yahoo.com/%E5%8F%B0%E5%8D%97%E5%A1%9E%E7%88%86-%E5%B0%8F%E9%BB%83%E7%B4%99%E6%A2%9D%E6%9B%9D-%E9%BE%9C%E9%80%9F%E5%8E%9F%E5%9B%A0-%E7%B6%B2%E7%AC%91-%E4%B8%8D%E5%8F%AD%E4%BD%A0%E4%BA%86-143342818.html

「Pokémon GO Safari Zone in Tainan」今（1）日早上10點半開始將連續5天展開，而這場寶可夢盛會預計帶來的人潮預估將有20萬，各大飯店民宿住房率超過8成，甚至有業者開心表示，「比陸客來台還有用！」

三立新聞網 2018年11月1日 上午11:06

Pokémon GO Safari Zone in Tainan (寶可夢台南狩獵區)：估計主場都會公園奇美博物館有8萬人，大台南全區16萬人，連續五天活動總計主場有56萬人次，台南全區100萬人次。 六億商機！！

今日新聞NOWnews 記者陳聖璋 2018年11月5日

統計與理性判斷

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：

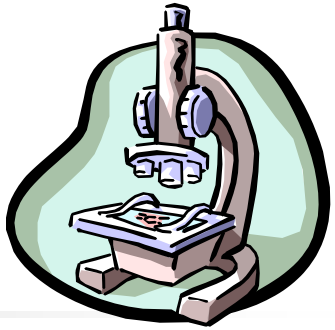
(1)健康者中有 $999{,}000 \times 2\% = 19{,}980$

(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：

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

第二意見(second opinion)



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

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

- 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。

https://www.linkedin.com/pulse/brain-stroke-detection-using-deep-learning-infogen-labs-

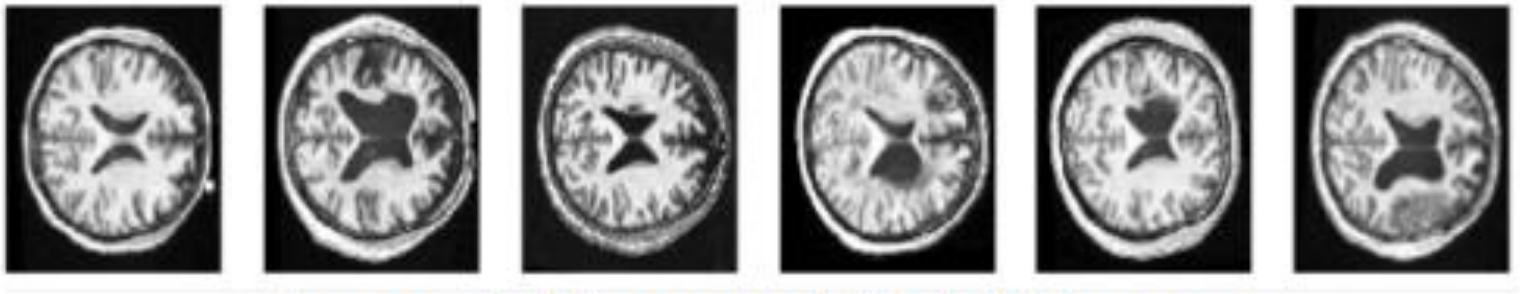


Figure 1.Abnormal Axial Images

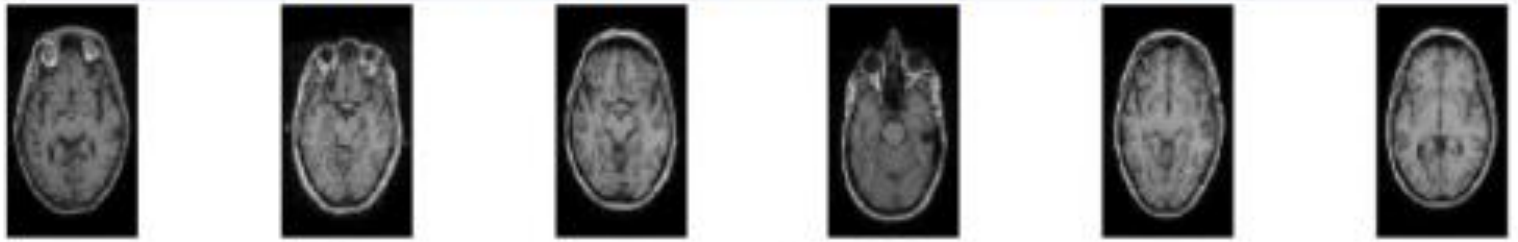
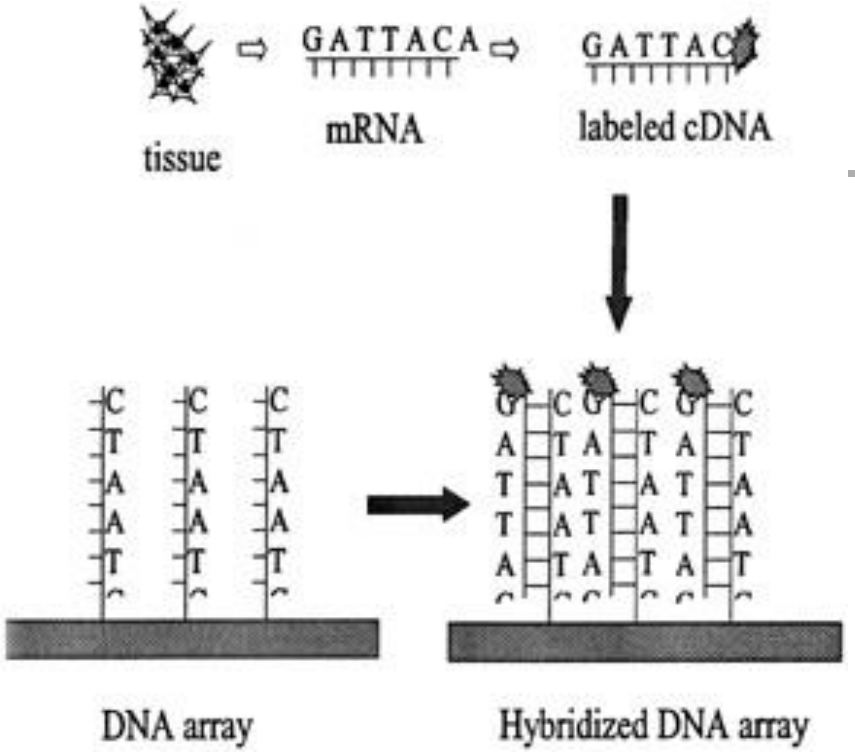


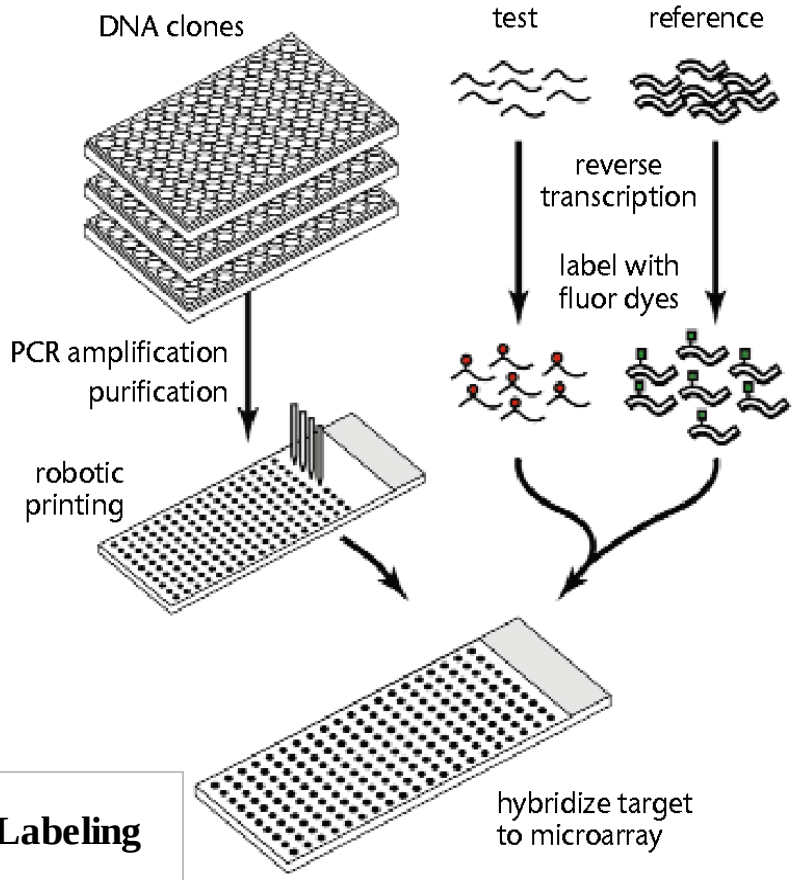
Figure 2.Normal Axial Images

生物晶片 (Micro-array Chips)

約 2 cm x 2 cm

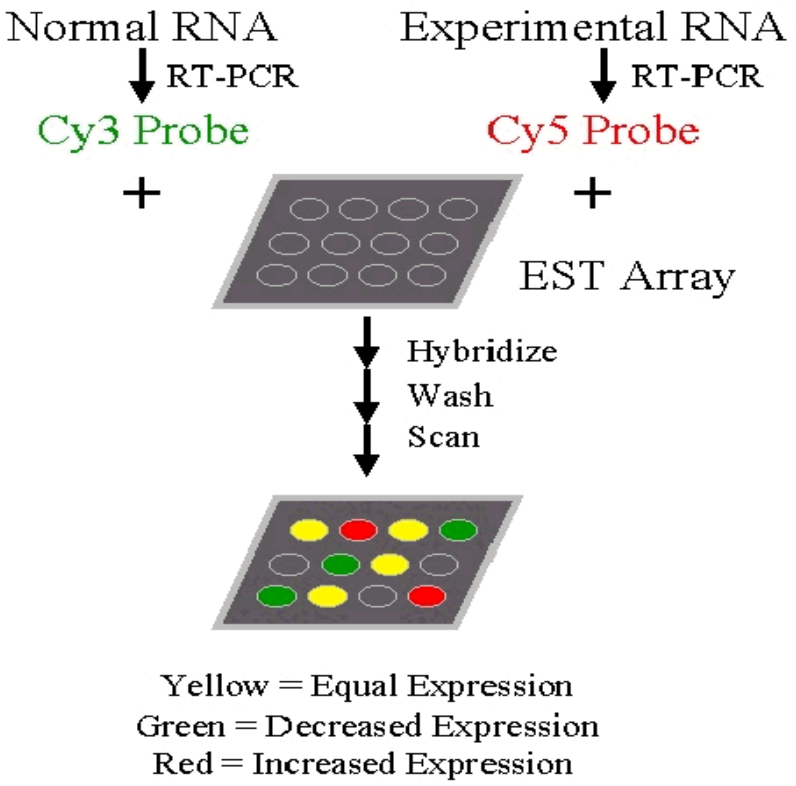
Microarray technology

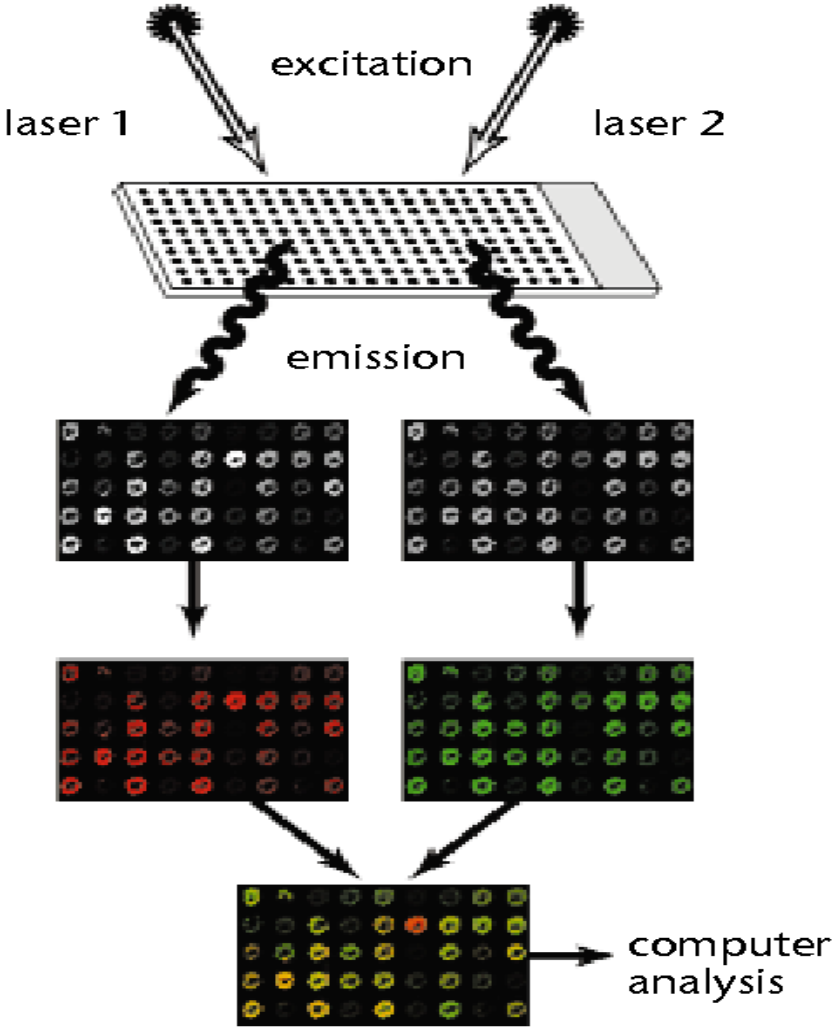




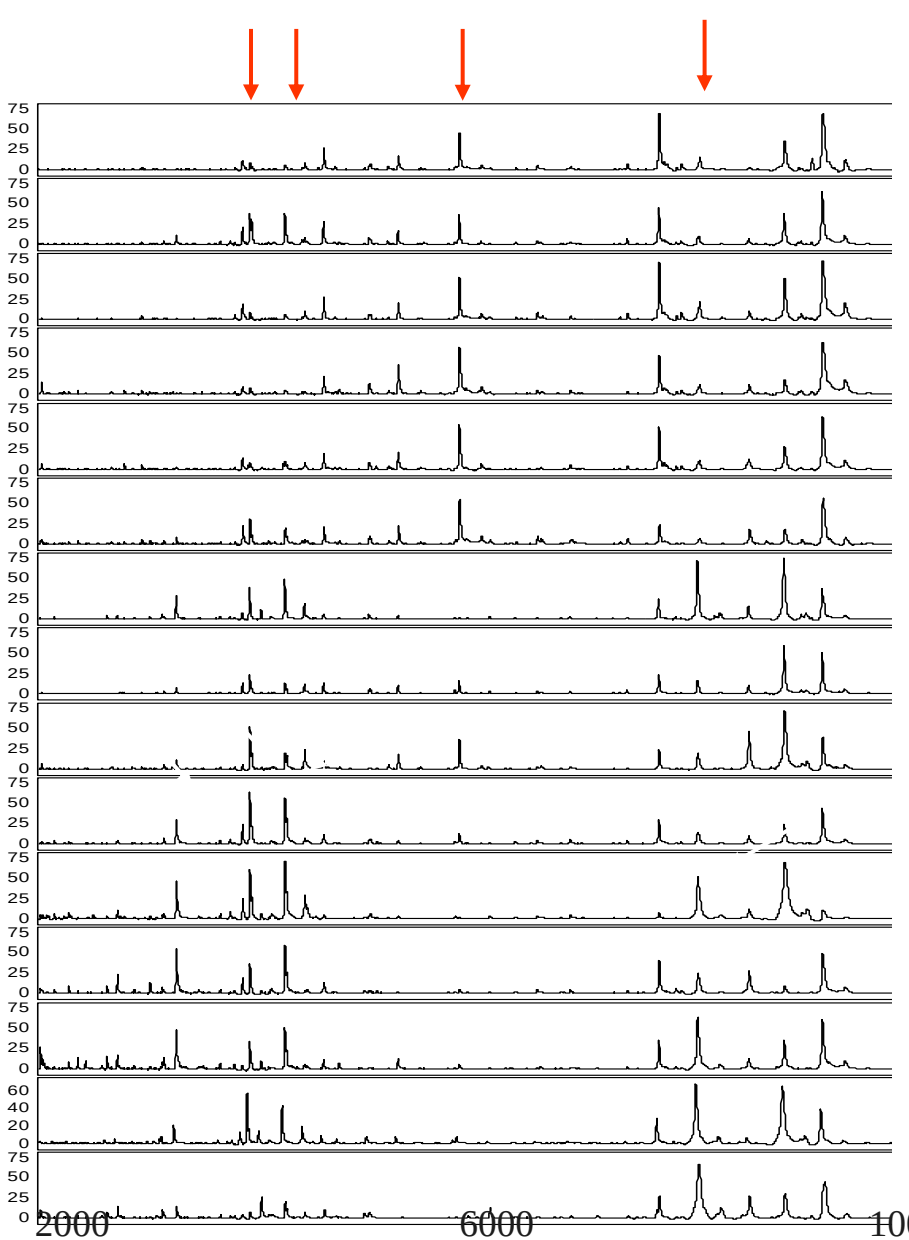
Array Type	Spot Density (per cm^2)	Probe	Target	Labeling
Nylon Macroarrays	< 100	cDNA	RNA	Radio
Nylon Microarrays	< 5000	cDNA	mRNA	Radio/Flour
Glass Microarrays	< 10,000	cDNA	mRNA	Fluor
Oligonucleotide Chips	<250,000	oligo's	mRNA	Fluor

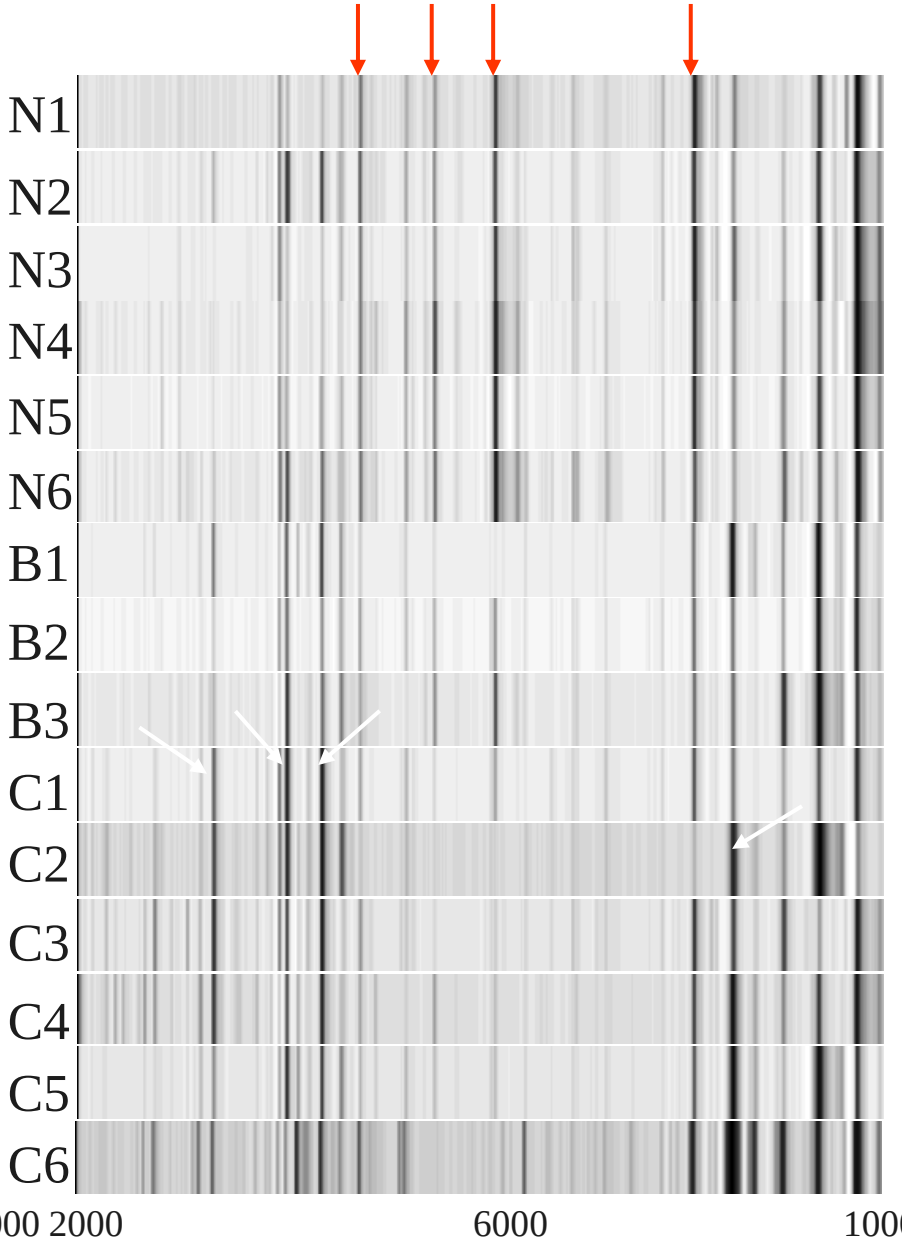
Microarray scanner





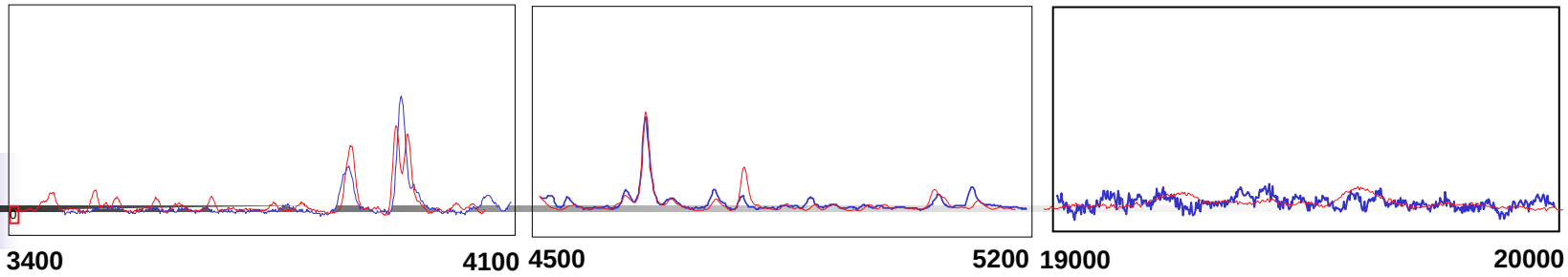
SELDI Serum Protein Profile Analysis-Prostate Cancer



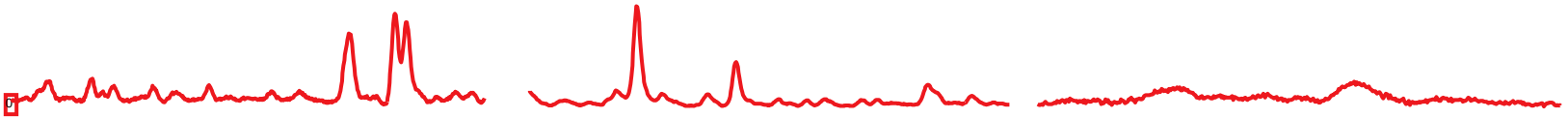


Proteomic Pattern of Sera from Patient

Stage 1 Profile

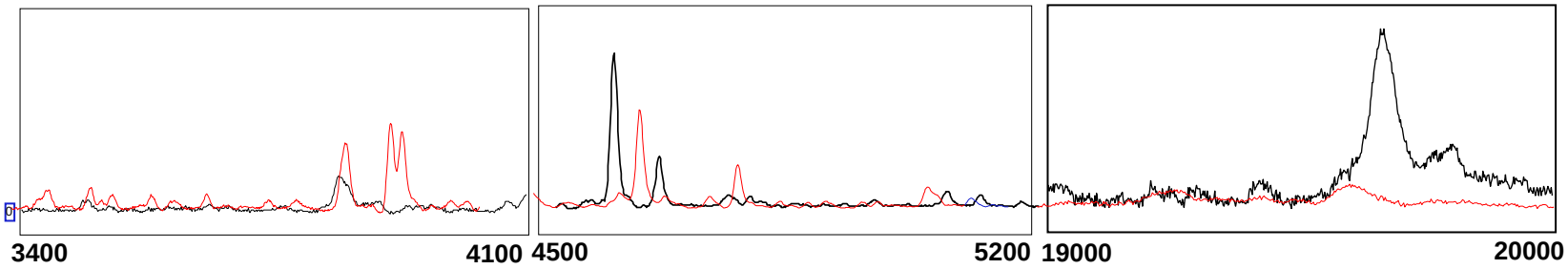


Sample



not a S2 pattern

Stage 2 Profile



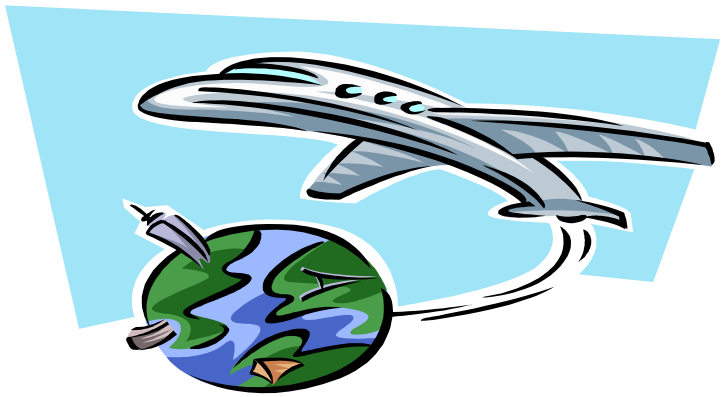
統計可用於規避風險

- 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

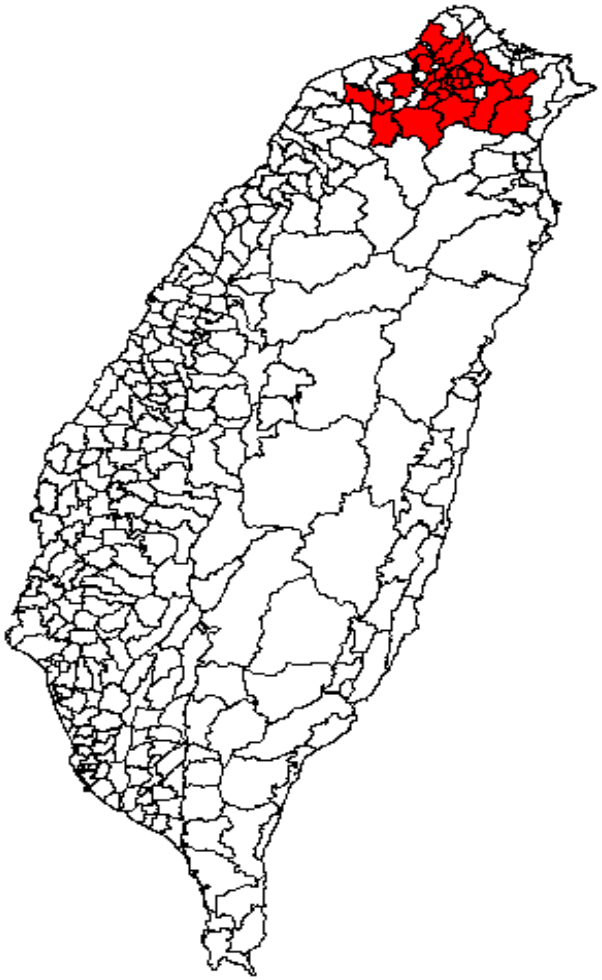
(羅吉士迴歸；logistic regression)



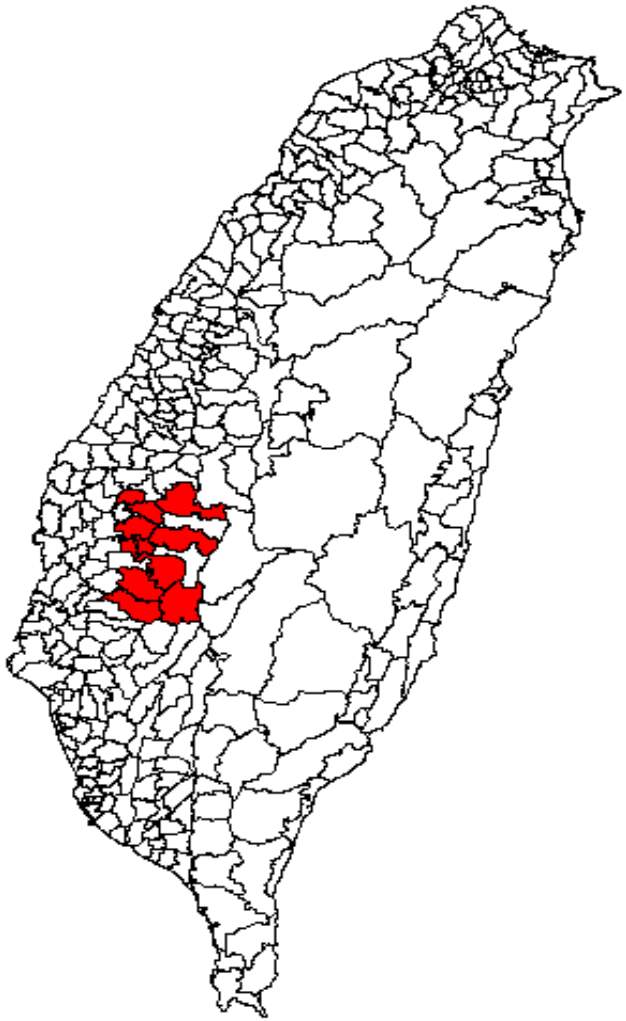
(參考書籍：天下文化「你管別人怎麼想」→費曼)

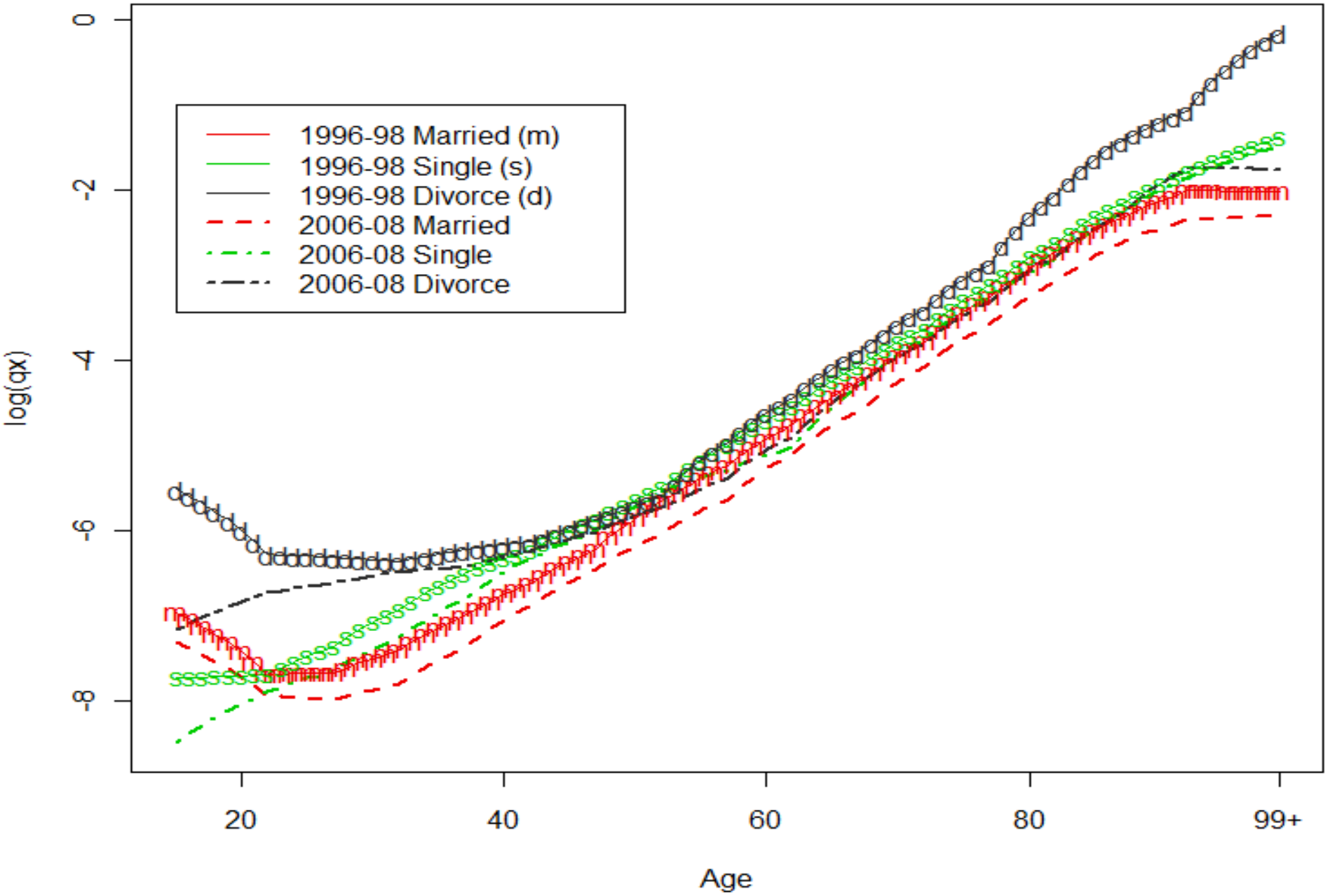
台灣地區全體人口門診次數顯著群聚圖

顯著增加



顯著減少





Taiwan's Marital Mortality Rates (Female)

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獲得賽揚獎，你們覺得哪些項目應該列入評比、比重又是多少？

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: 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。

統計在金融保險的應用

- 投資目的不外是報酬率最大、風險最小，但天下沒有白吃的午餐，這兩者很難兼顧。以符號表示，就是希望期望值最大，變異數最小， 即在固定風險值下求取最佳的報酬：

$$Max \sum_i w_i X_i, \quad given \quad Var(\sum_i w_i X_i) \le C$$

→Value at Risk (VaR)：譯作風險值，是統計應用於財務金融的熱門範例，

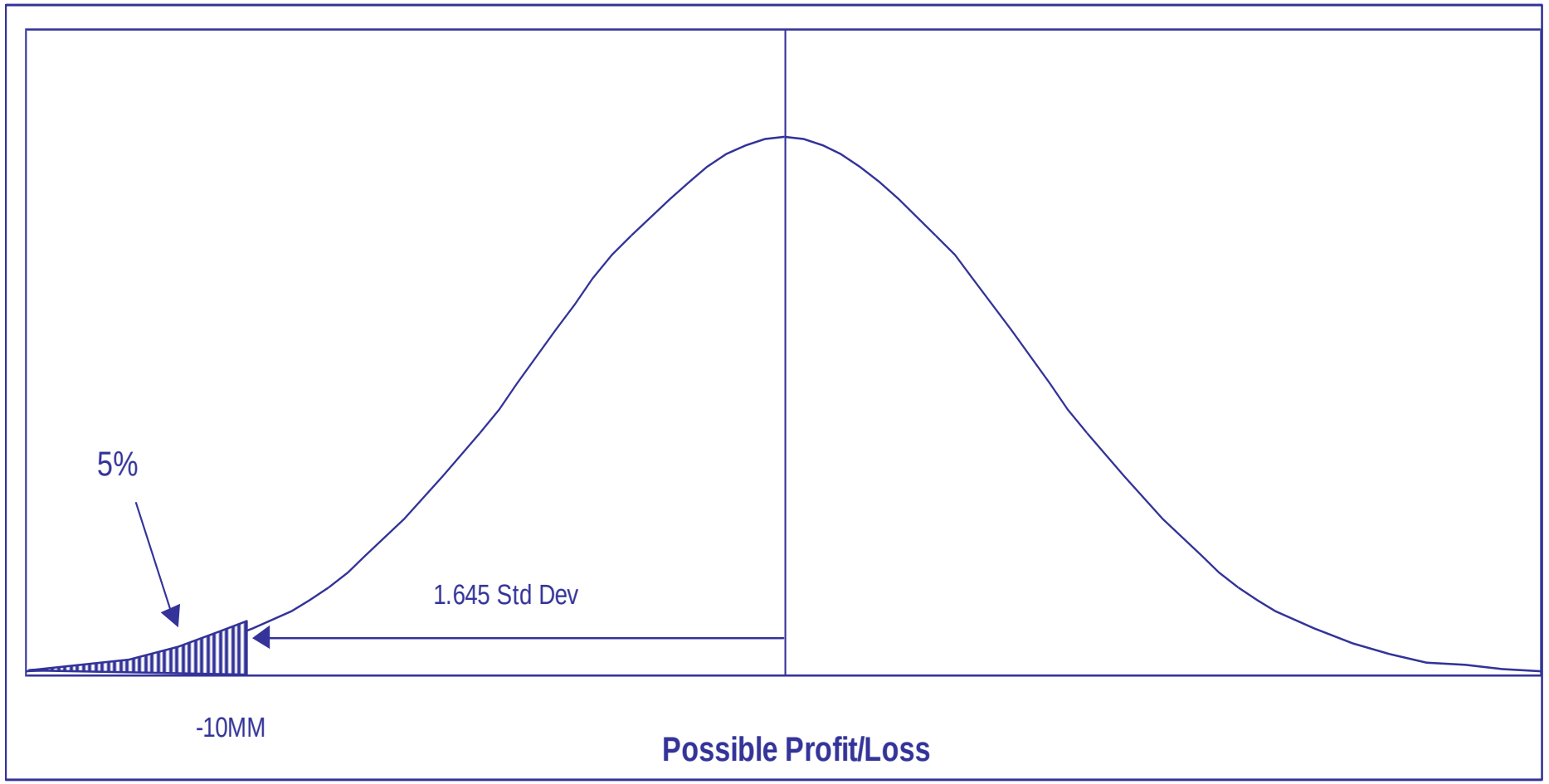
Value at Risk: Definition

The value at risk (VaR) of a portfolio is the loss in value in the portfolio that can be expected over a given period of time (e.g., 1-Day) with a probability not exceeding a given number (e.g., 5%).

$$\text{Probability (Portfolio Loss} \leq -\text{VaR}) = K$$

K = Given Probability

- A one day VAR of $10mm using a probability of 5% means that there is a 5% chance that the portfolio could lose more than $10m in the next trading day.



其他金融保險的範例

- 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。
- 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。

- 問題：假設你/妳是審查信用卡申請，或是審核消費性貸款(無抵押品)的負責人，必須考慮需要蒐集哪些個人項目？

→貸款若有抵押品，問題相同嗎？

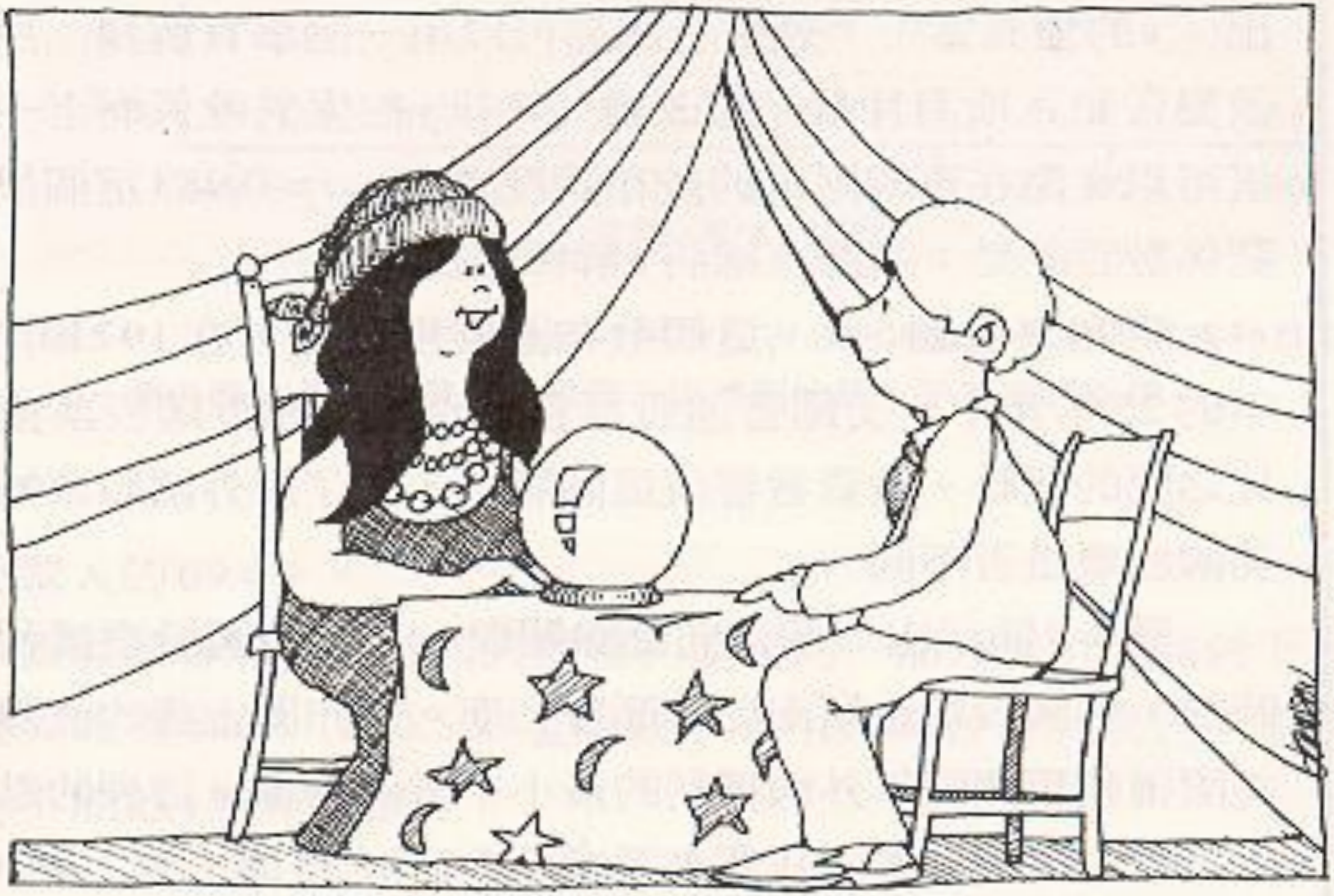
→如果考慮的不是個人，而是中小型企業貸款，需要蒐集哪些資訊？

註：消費金融 vs.企業金融

- 問題：實務上與統計分析有關的領域？

https://www.youtube.com/watch?app=desktop&v=_W_LDp735ms&ab_channel=LawShelf

https://careerdonclub.weebly.com/3709633853266684/day07



「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。」