

淺談統計與人口統計

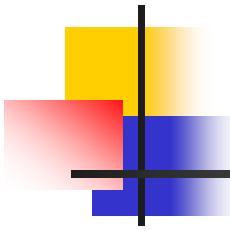
授課教師：余清祥教授

課程日期：2005年9月8日、
9月14日

資料下載：

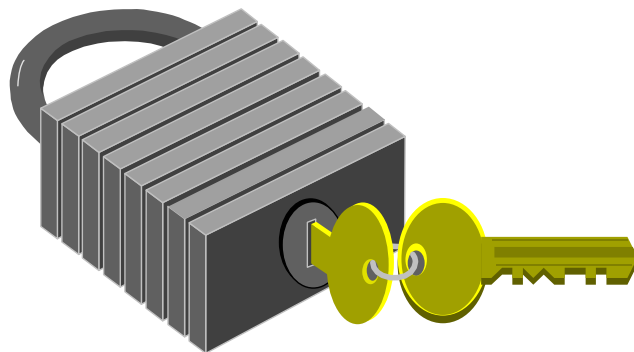
<http://csyue.nccu.edu.tw>





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。



WHAT IS STATISTICS?

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

SHOULD I HAVE THE SOUP?
EVERYTHING ELSE IS SO
EXPENSIVE, AND I DON'T
KNOW WHO'S PAYING... ARE
STATISTICIANS STINGY? I'VE
NEVER GONE OUT WITH
ONE BEFORE... THOUGH I
ONCE KNEW A VERY
GENEROUS ACCOUNTANT...

SHOULD I HAVE THE SOUP?
27 OUT OF THE 36 TIMES
I'VE HAD IT, IT WAS PRETTY
GOOD... BUT IS MONDAY THE
REGULAR CHEF'S NIGHT
OFF? AND WHAT IF ALL THE
AIR MOLECULES IN THE
ROOM SUDDENLY FLY UP TO
THE CEILING?



統計是什麼？

我們在一生當中，
常要根據不完整的資訊做決定……

我要不要叫湯？
其他東西都這麼貴，
而且還不知道是誰要付錢呢！
統計學家也不知道小不小氣？
以前沒有跟這種人約會過，
雖然曾經認識一個很大方的會計
……

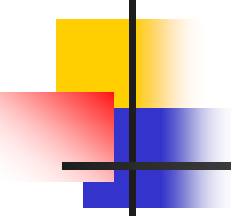
我要不要叫湯？
以前叫的36次當中，
有27次都不錯；
不過不知道平常的大廚
是不是在星期一休假？
還有萬一房間裡的空氣分子
突然全部升上天花板
怎麼辦？



以統計做決策的要件

1. 事件(Event)的可能結果含有不確定性的因素；
2. 決策所衍生的結果可被量化。
(也就是說，我們可藉由效用函數判斷哪一個決策較佳。)



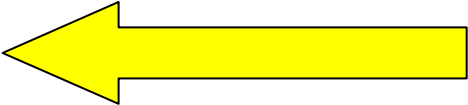


定義問題

蒐集資料

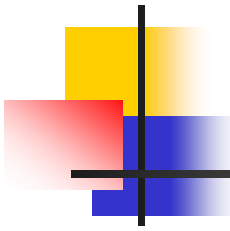
分析資料

詮釋結果



絕大多數的
統計教學重心





馬克吐溫對統計的想法

There are three kinds of lies:

Lies,

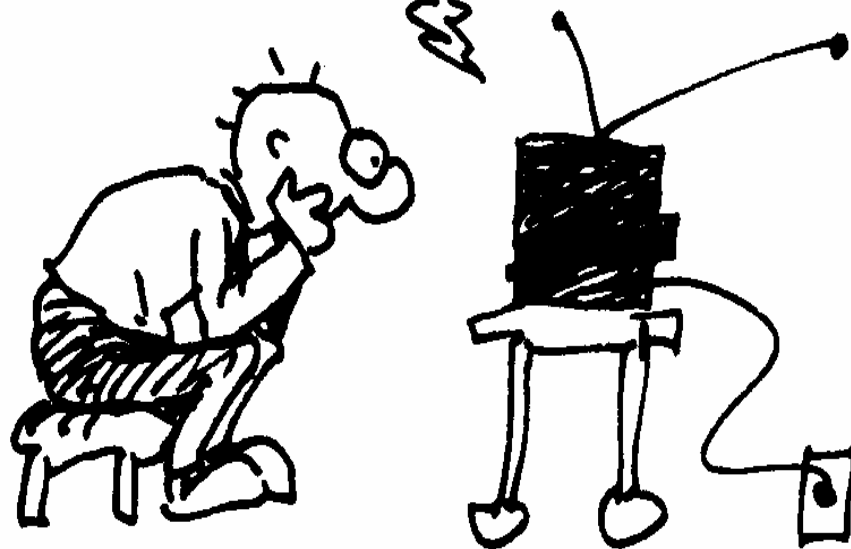
Damned lies,

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?

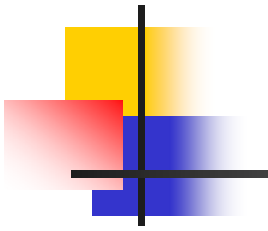
3 OUT OF 4 DOCTORS
RECOMMEND NOT BELIEVING
ANY STATEMENT BEGINNING ..
WITH "3 OUT OF 4 DOCTORS..."





統計的應用

早期的統計應用多為政府統計(包括人口統計)，直至19世紀近代統計萌芽後，應用範圍擴展至農業、生物、工業等範圍。近二十年統計在台灣的發展更多采多姿，尤其在生物資訊、資料採礦、金融財務(含保險)的表現最為耀眼，已是21世紀不可或缺的重要科學。



數據 (Data)

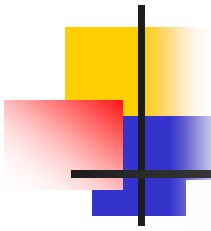
資訊 (Information)



事實 (Fact)

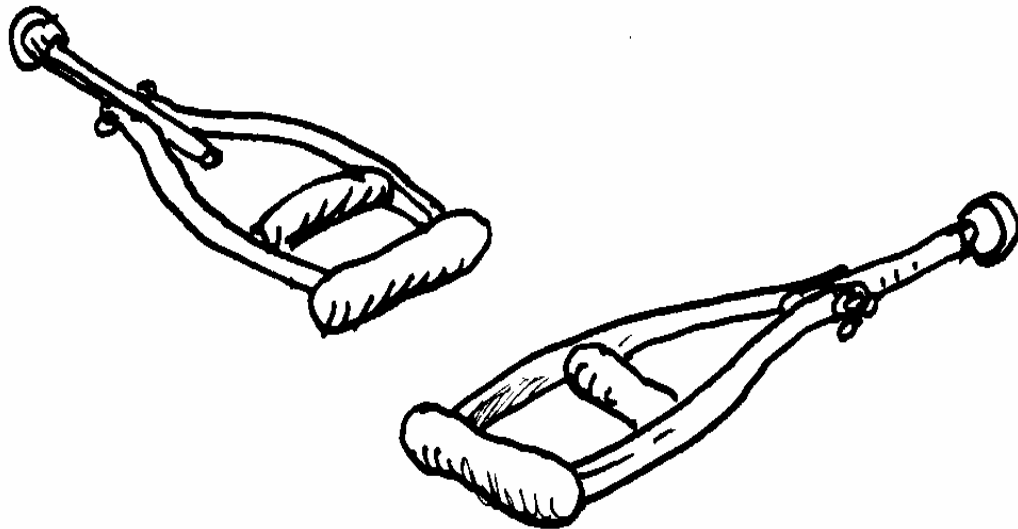
知識 (Knowledge)





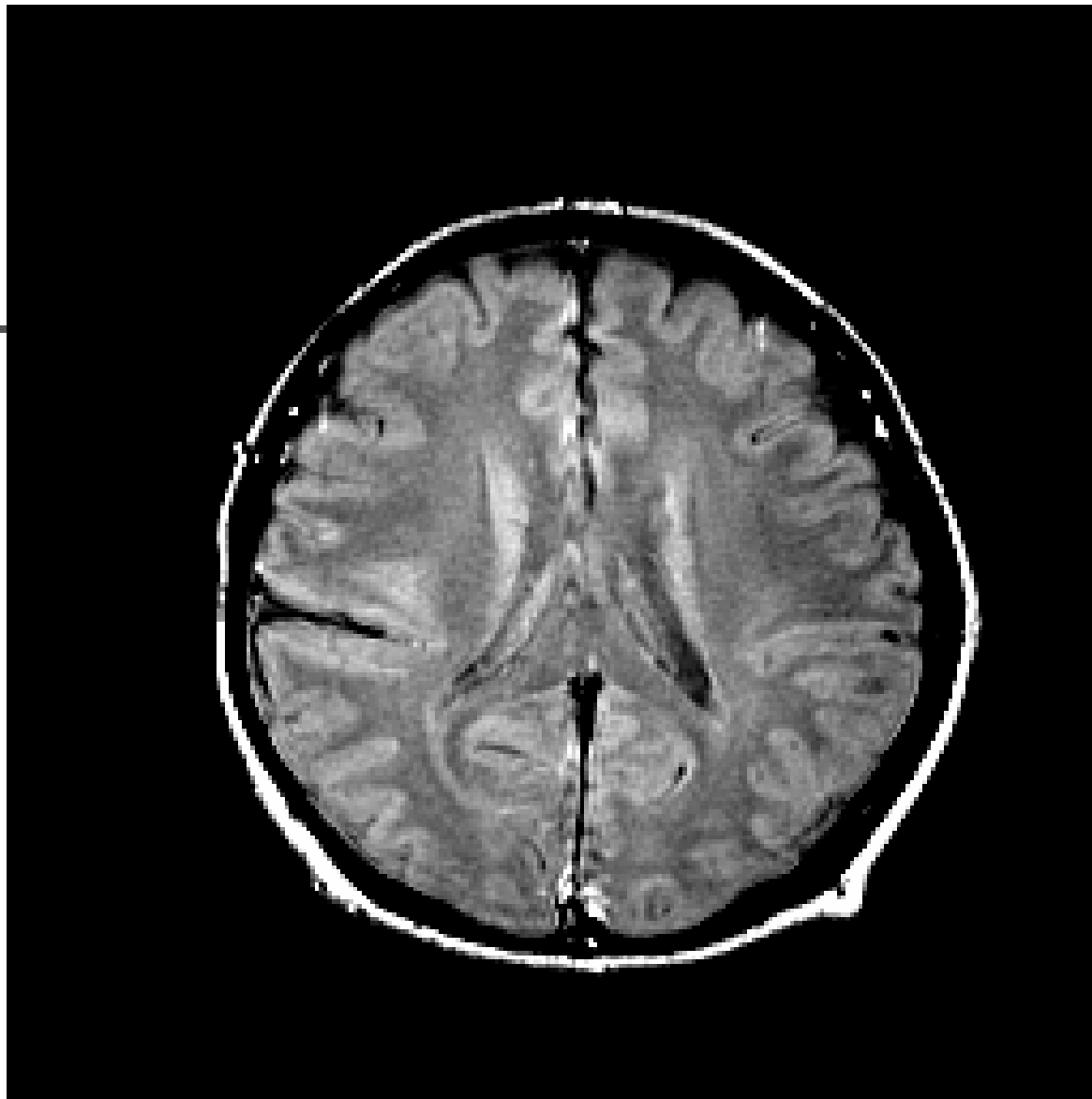
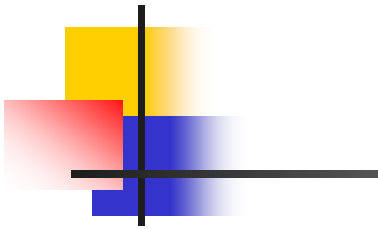
世界最大的醫學實驗(沙克疫苗)

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.

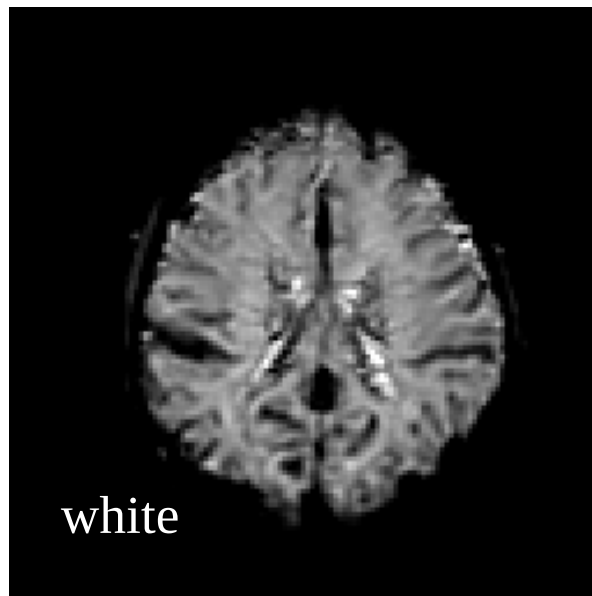
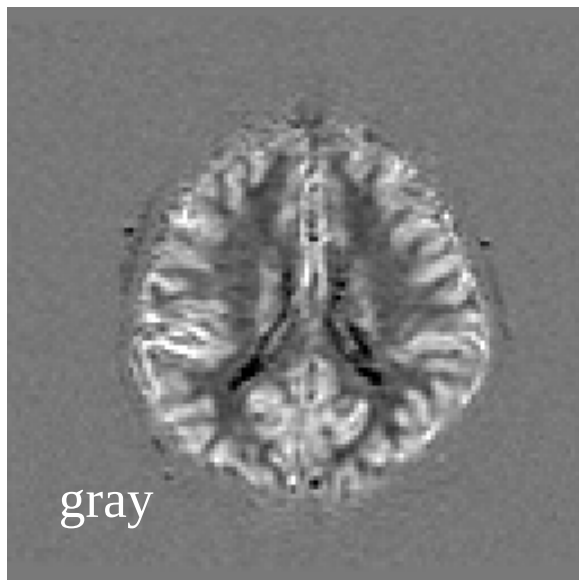
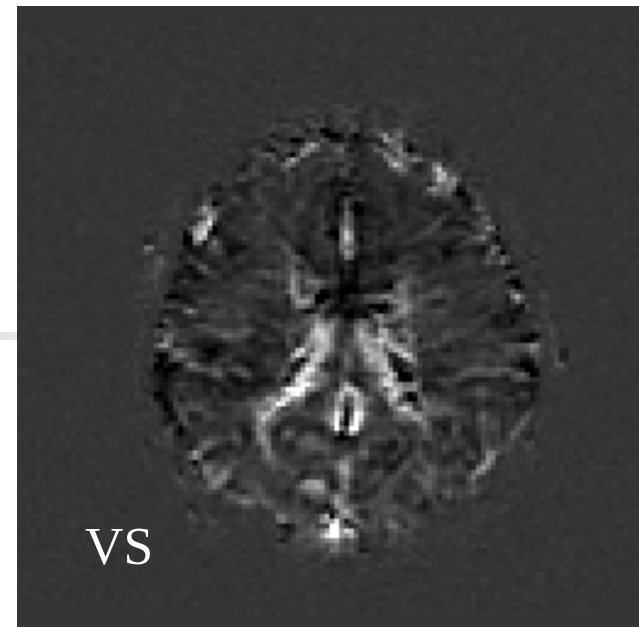
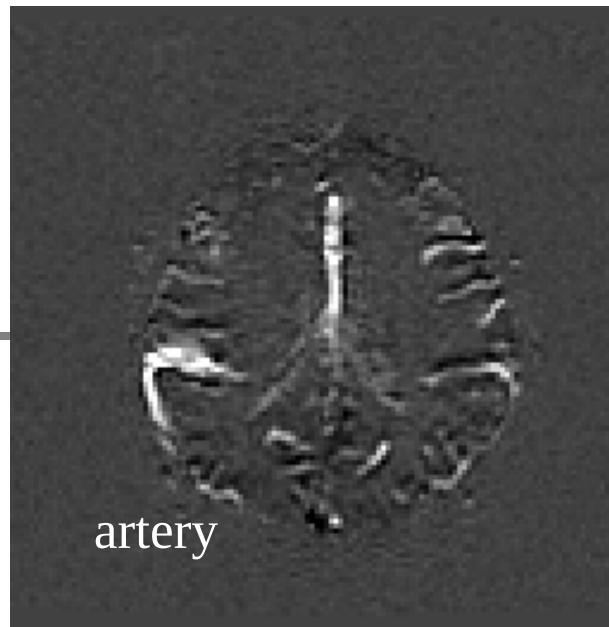
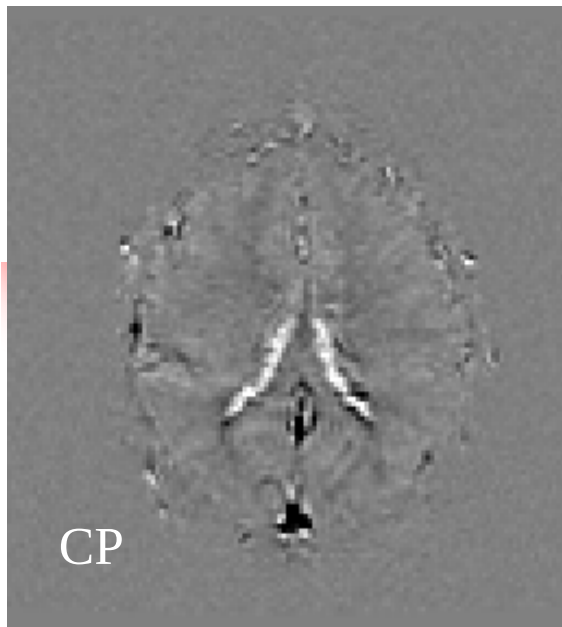


- 健康檢查時經常會抽血、驗尿、或萃取某個身體組織附近的樣本，再從檢體中判定是否罹患某種疾病。
- 檢查的結果以圖像表示較為清楚，例如：
下圖即為中風病人的檢查結果。



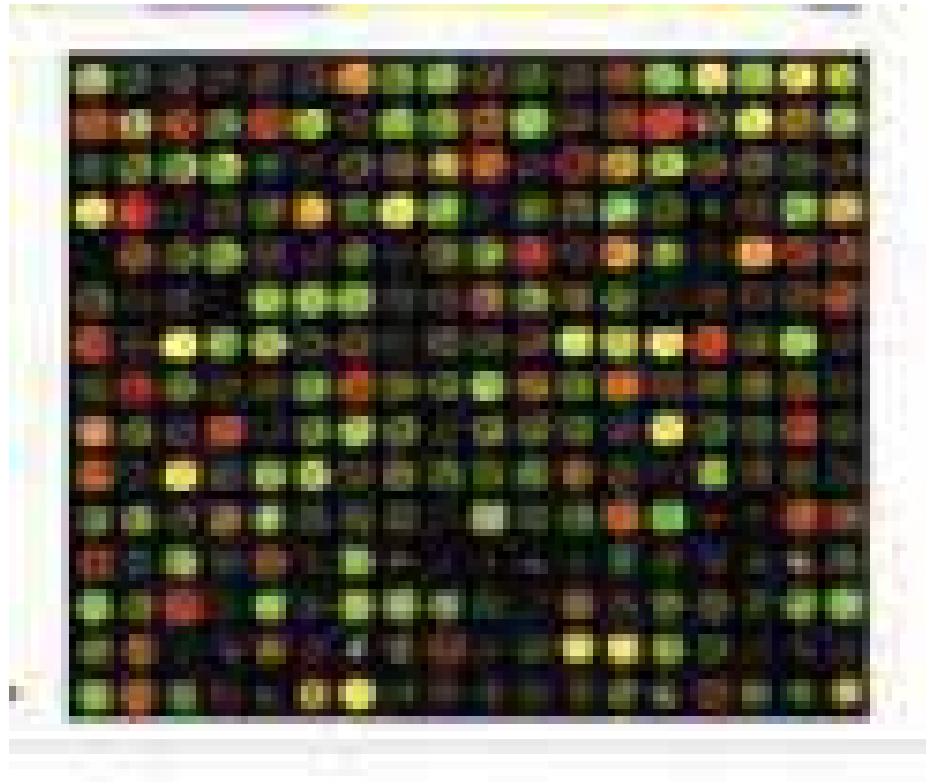


Anatomical proton-density-weighted image of human brain



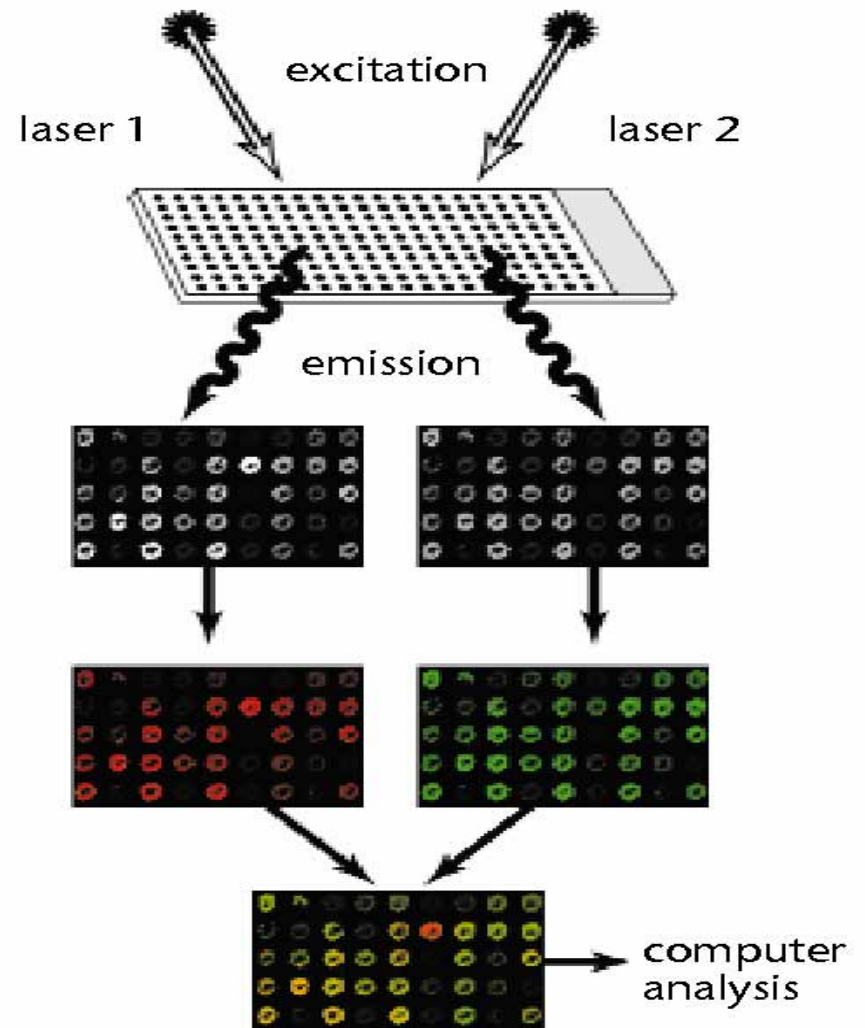
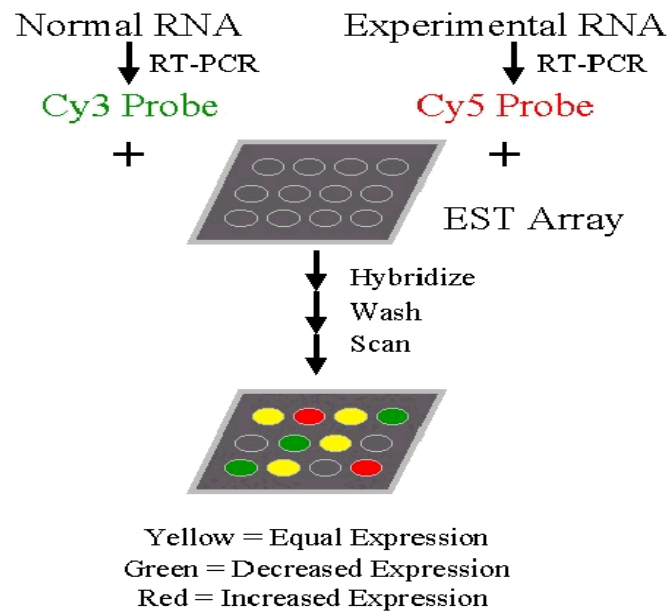
通常可分為5個
圖。

生物晶片 (Micro-array Chips)

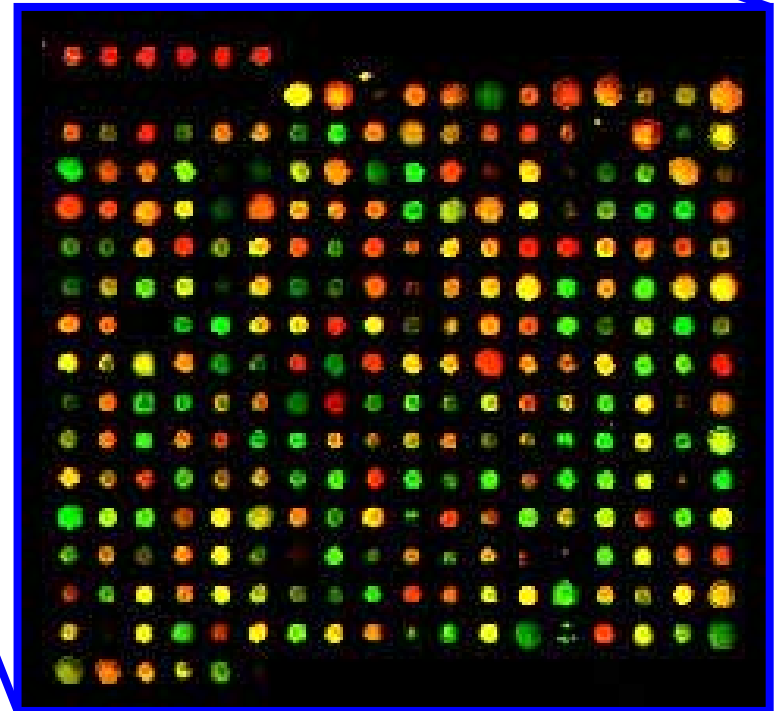
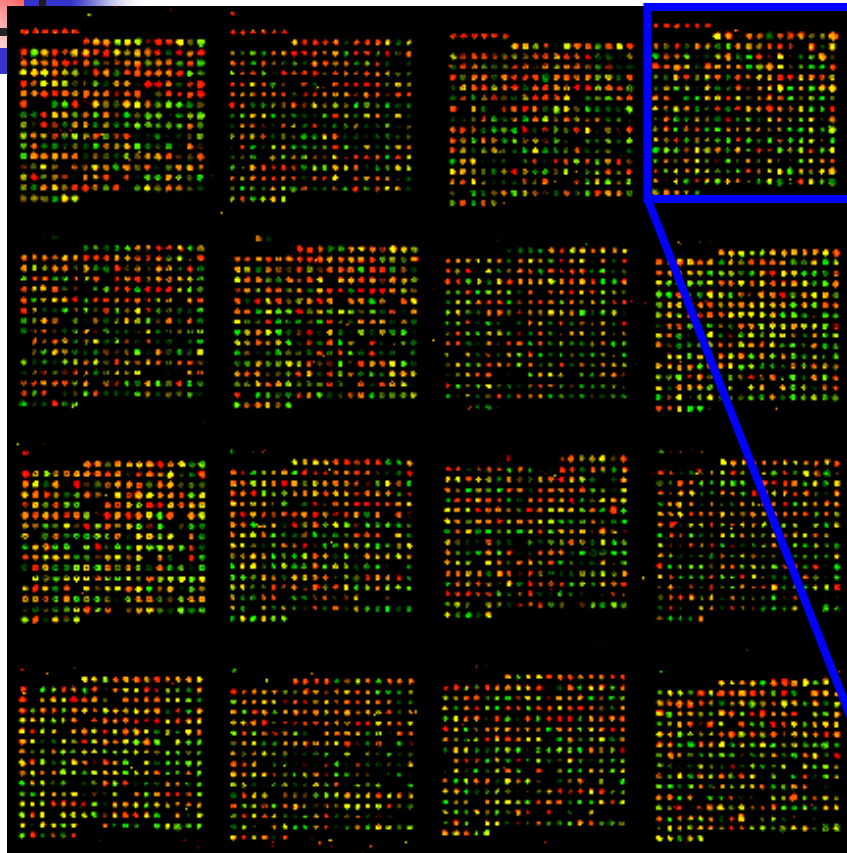


約 2 cm x 2 cm

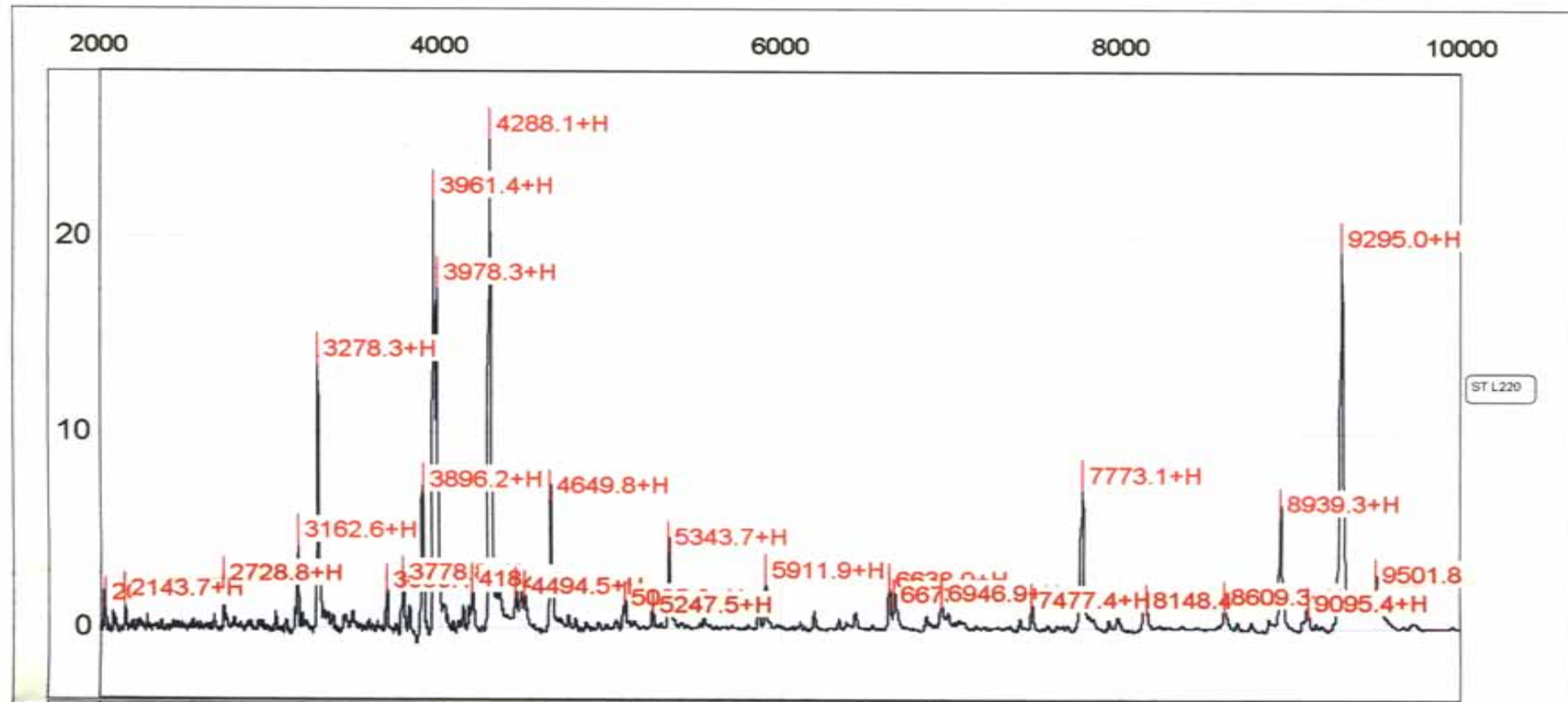
Microarray scanner

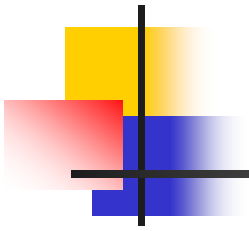


Microarray scanner

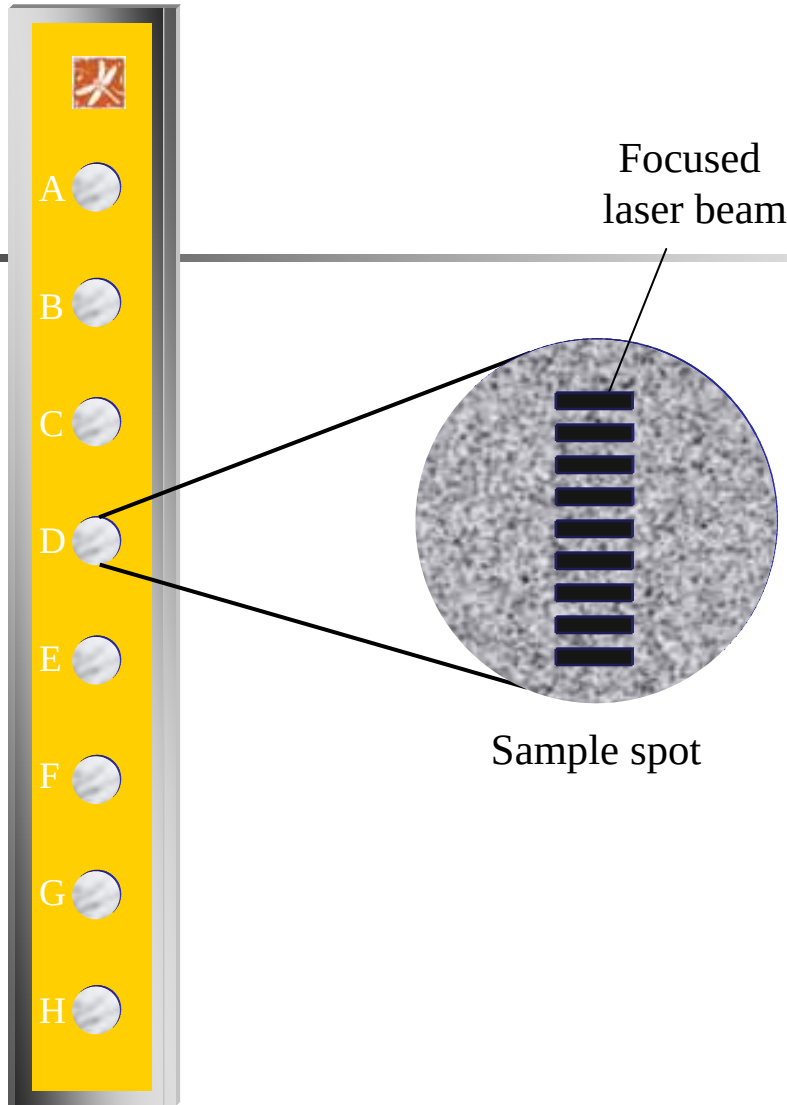


血清樣本的質譜(Spectrum)分析





ProteinChip™
Array

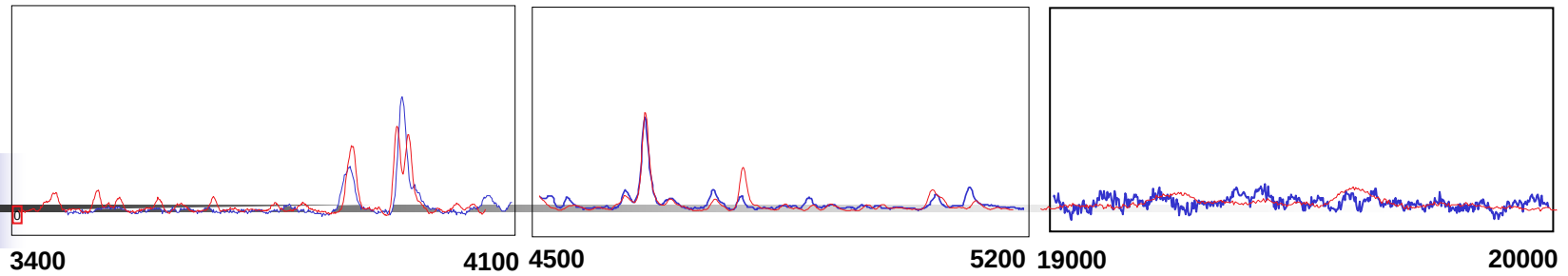


Available Array Chemistry

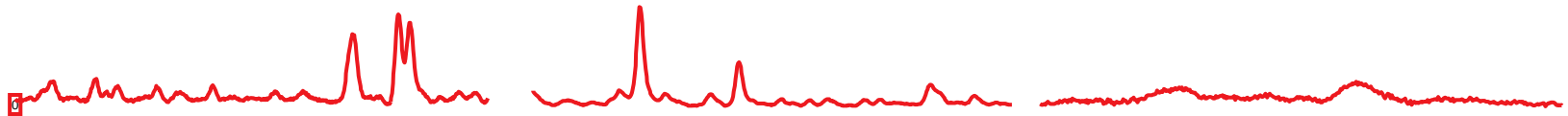
- Reverse Phase (2)
- Normal Phase (4)
- Strong Anion Exchange
- Weak Anion Exchange (2)
- IMAC
- Au Chips

Proteomic Pattern of Sera from Patient

Stage 1
Profile

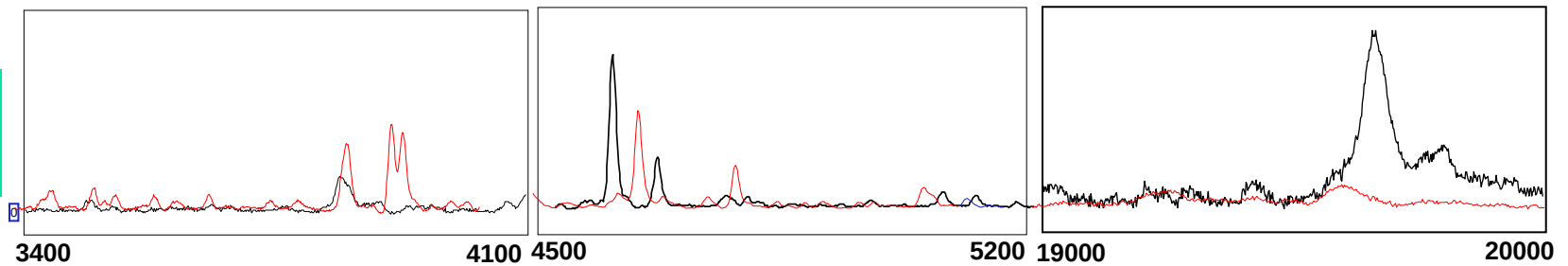


Sample



not a S2 pattern

Stage 2
Profile



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

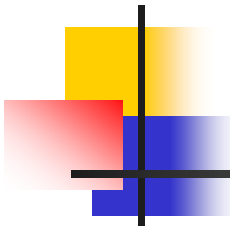


Customer2

Eggs, sugar



Customer3



其他資料採礦的範例：

- 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)
 - 舊卡戶在信用卡終止後皆發給新卡
 - 考慮發新卡的對象
- 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。

ATM提款服務R.F.M.T之顧客分群表

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，相對其風險也高，容易變成別人的顧客	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

規避不必要的風險

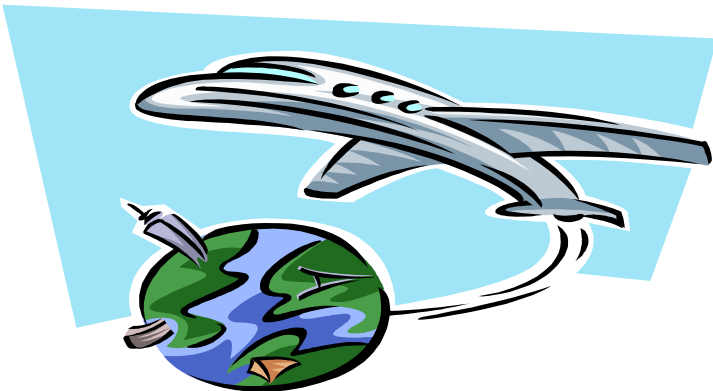
- 1986年美國挑戰者號太空梭的爆炸
→ O形環(O-Ring)在低溫下無法正常運作，造成燃料外洩而爆炸。

→ 分析過去各種溫度下的失敗比例

(羅吉士迴歸；logistic regression)

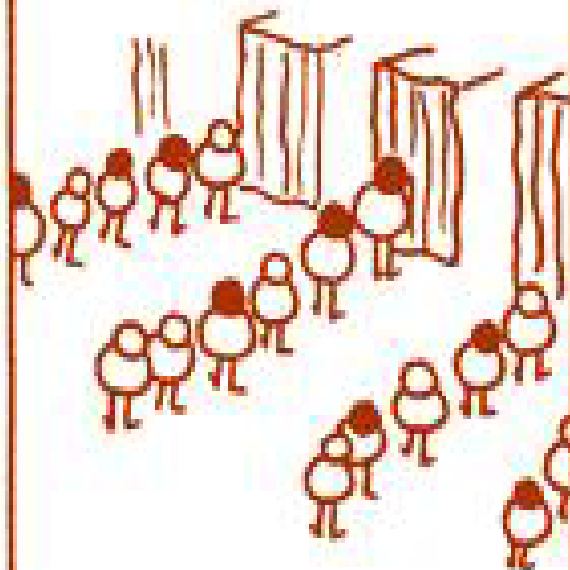
(參考書籍：天下文化

「你管別人怎麼想」→費曼)

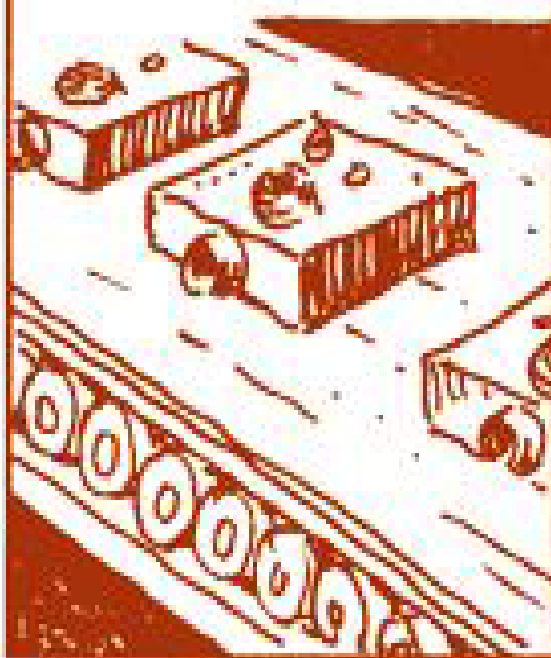


其他生活上範例

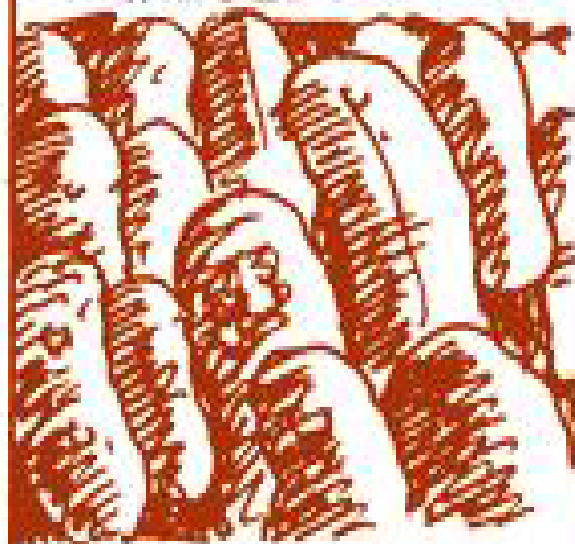
關於全體選民：
各候選人的支持率為何？



關於產品：
不良品的比例是多少？

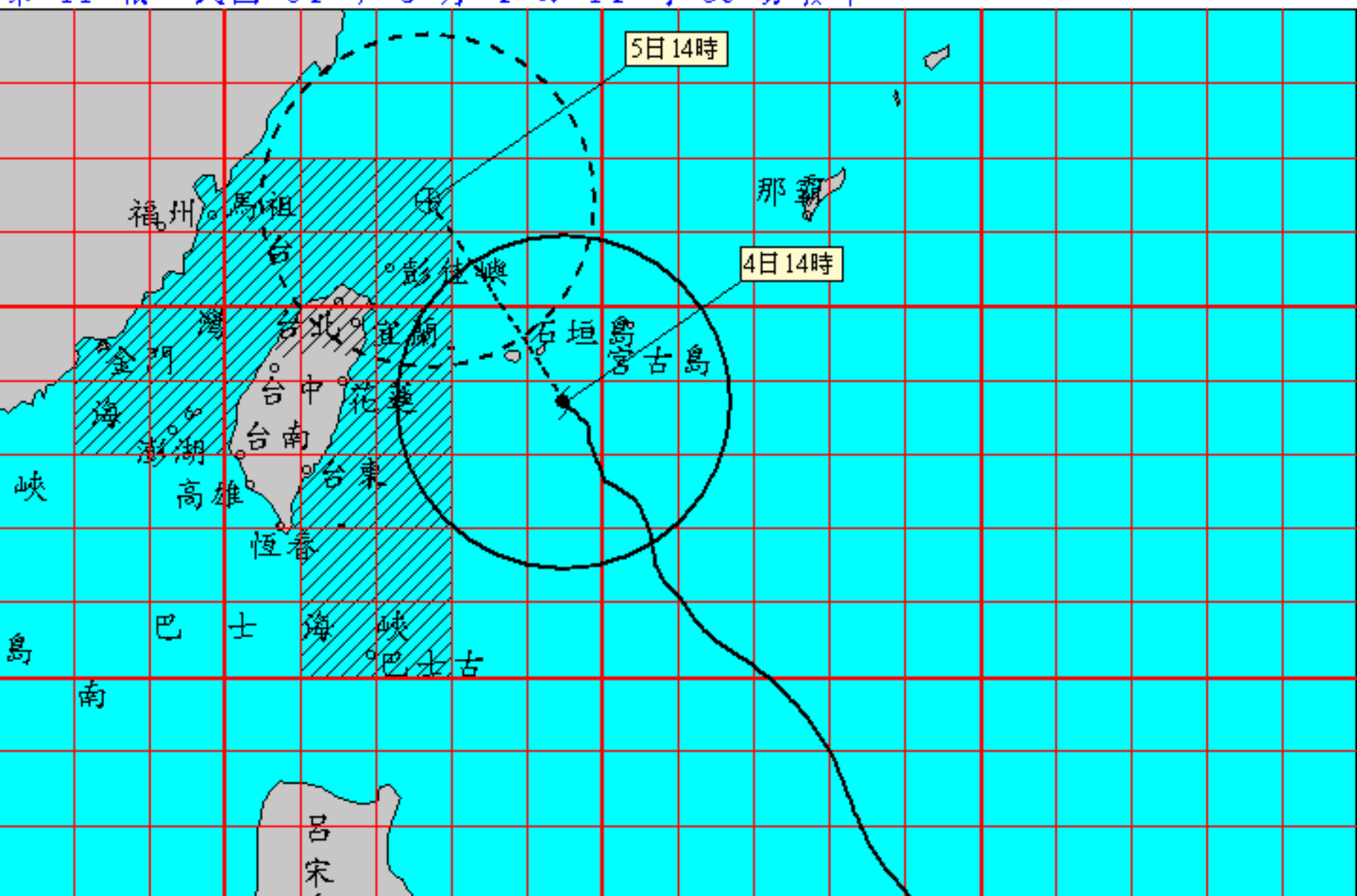


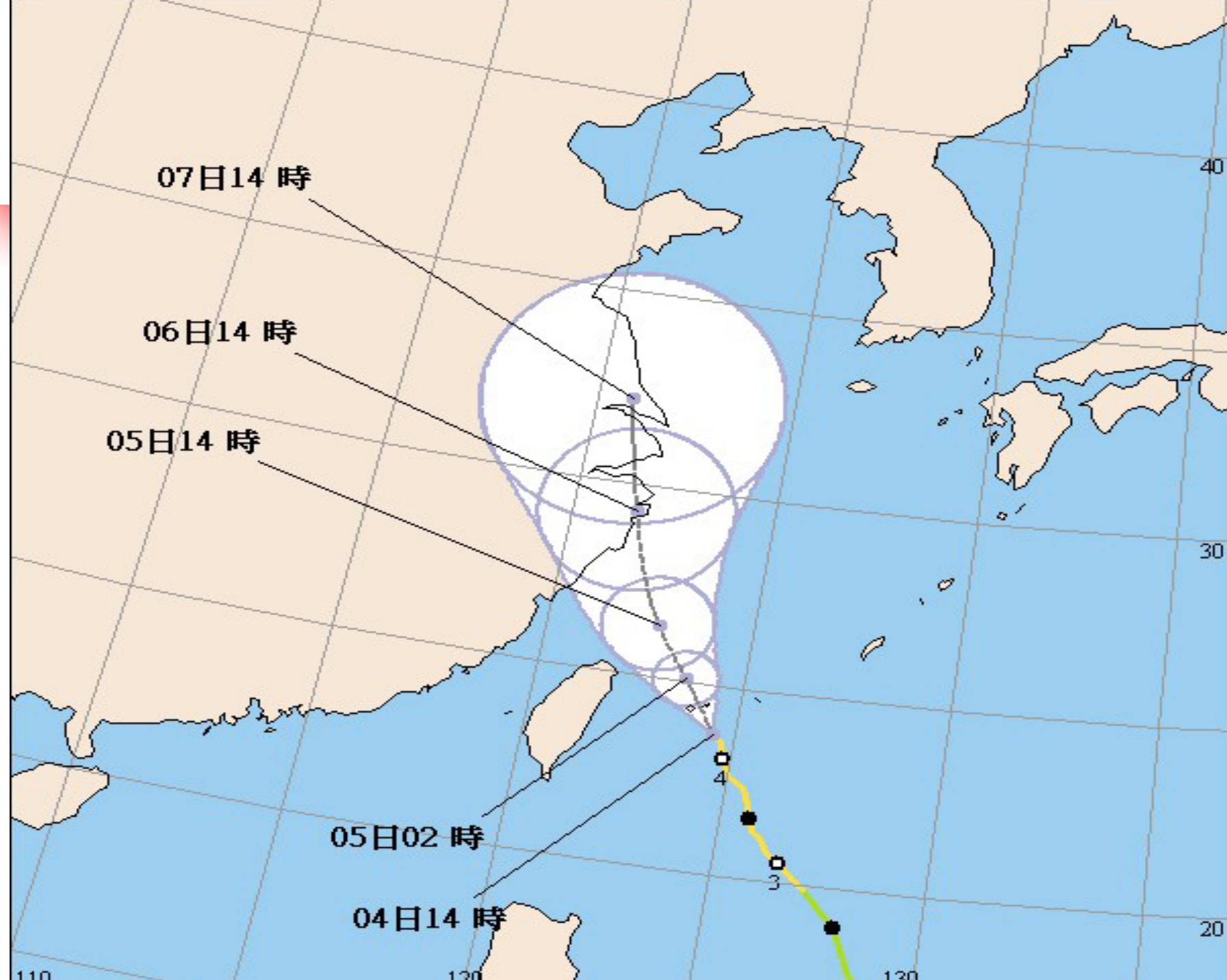
醃黃瓜：
平均長度是多少？



醃黃瓜瓶子的製造商必須知道！

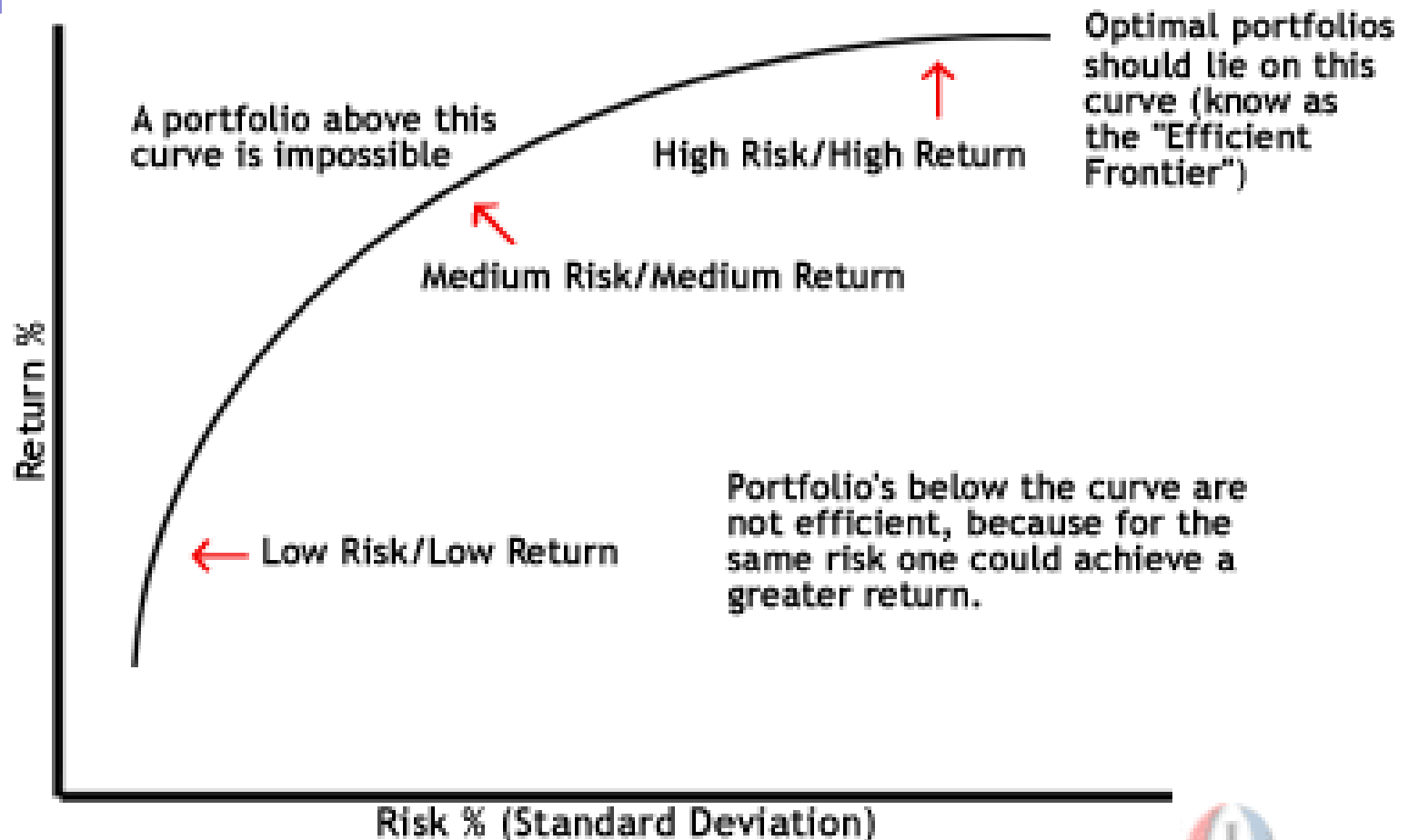
中度颱風 (編號第9號 國際命名：MATSA，中文譯名：馬莎)
第 11 報 民國 94 年 8 月 4 日 14 時 30 分發布

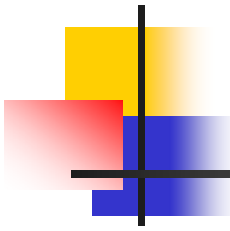




投資組合(Investment Portfolio)：

效率前緣(Efficient Frontier)

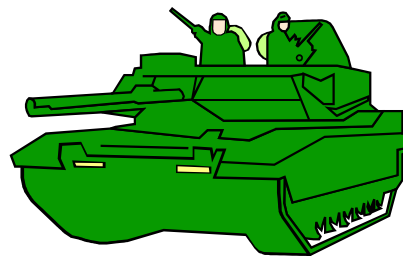
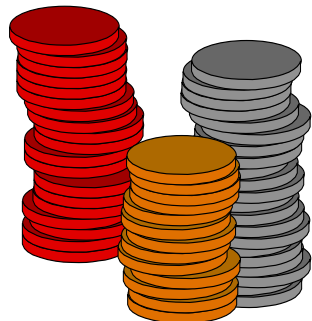




人口統計的定義



- 人口統計或人口統計學(Demography)為研究一個地區或國家人口的學門，主要涵蓋人口總數、人口結構、與人口變遷及發展等方面。其精細之意義為：「對人類人口數量及其因出生、死亡及移民所引起之變動之研究。」近年來其範圍擴大為：「對於生育、婚姻、移民及死亡等資料之蒐集及統計的分析。」



- 中國最初的統計用於清查國家的財富與人口的數字，供政府作為徵兵、徵役及課稅等項的根據。
 - 井田制度(例如：壯丁百畝)。
 - 歷朝各代多以丁男人數為課稅單位。
 - 在荒年或戰亂時代，許多人寧願放棄平民(及戶籍)的身份，成為官吏、富豪的奴隸，以逃避兵役及沈重的賦稅。

人口統計研究主要項目

- (1) 確定某區域內人口之數量及分佈情形；
- (2) 描述以往人口之成長與減退之情形；
- (3) 描繪以往及現在人口之分佈情形；
- (4) 發掘社會組織各方面影響人口趨勢之原因；
- (5) 預測人口之未來發展及可能結果。

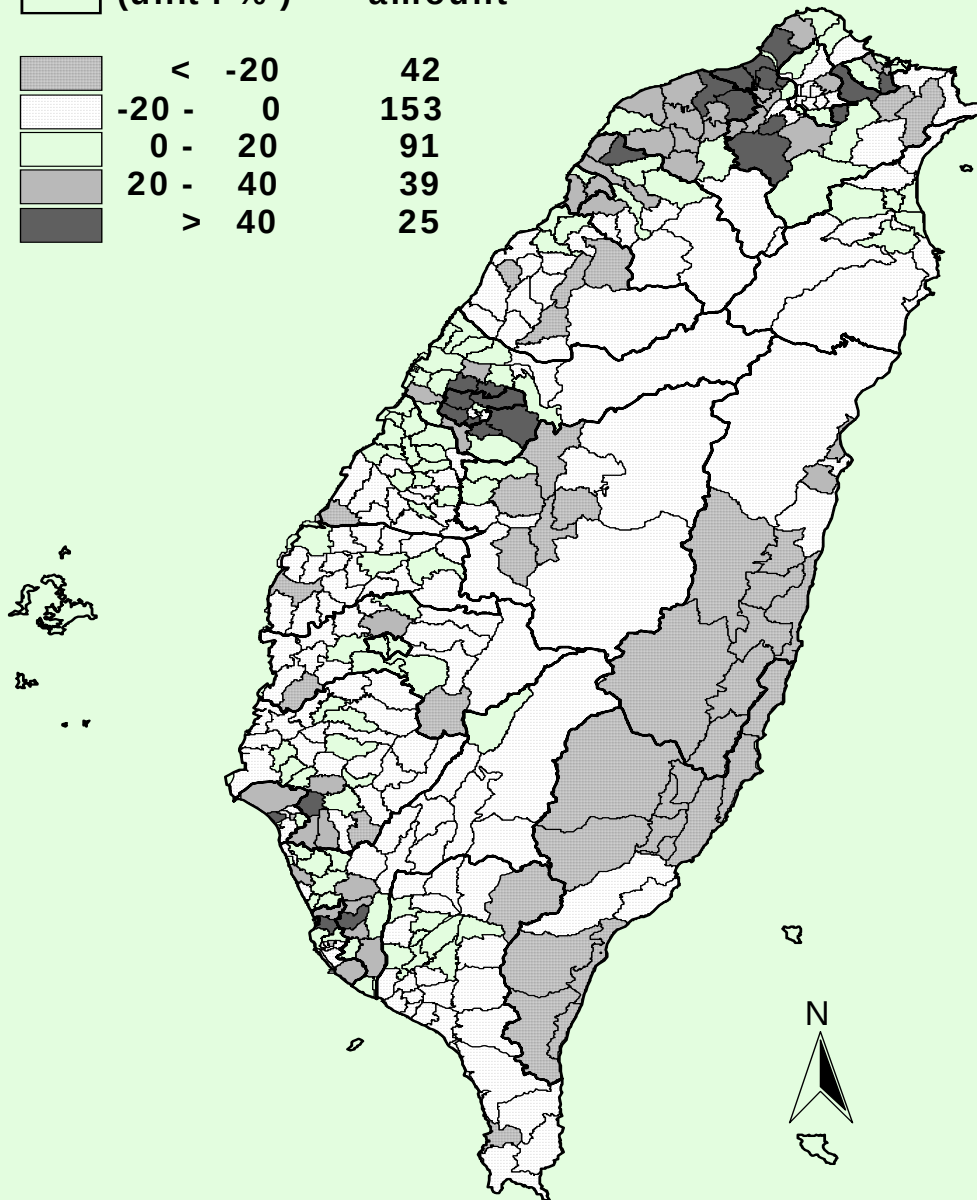


台灣人口數變化 (1990-2000)

(unit : %)

amount

	< -20	42
	-20 - 0	153
	0 - 20	91
	20 - 40	39
	> 40	25

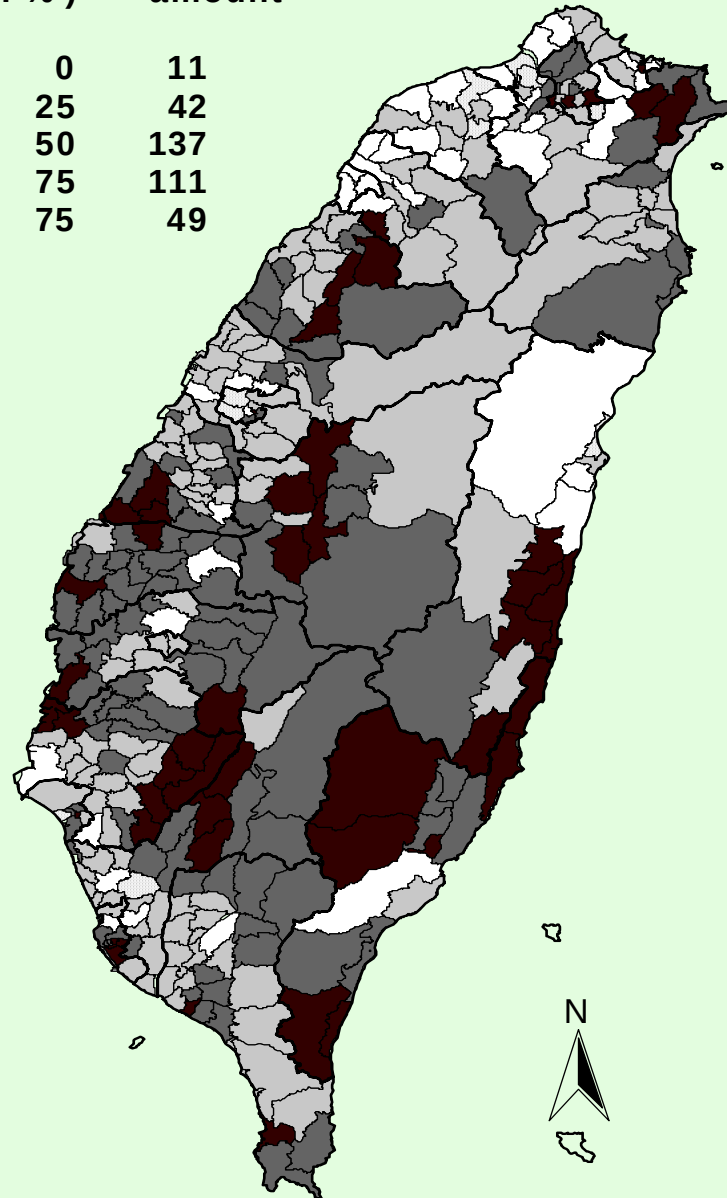


台灣高齡人口比例變化 (1990-2000)

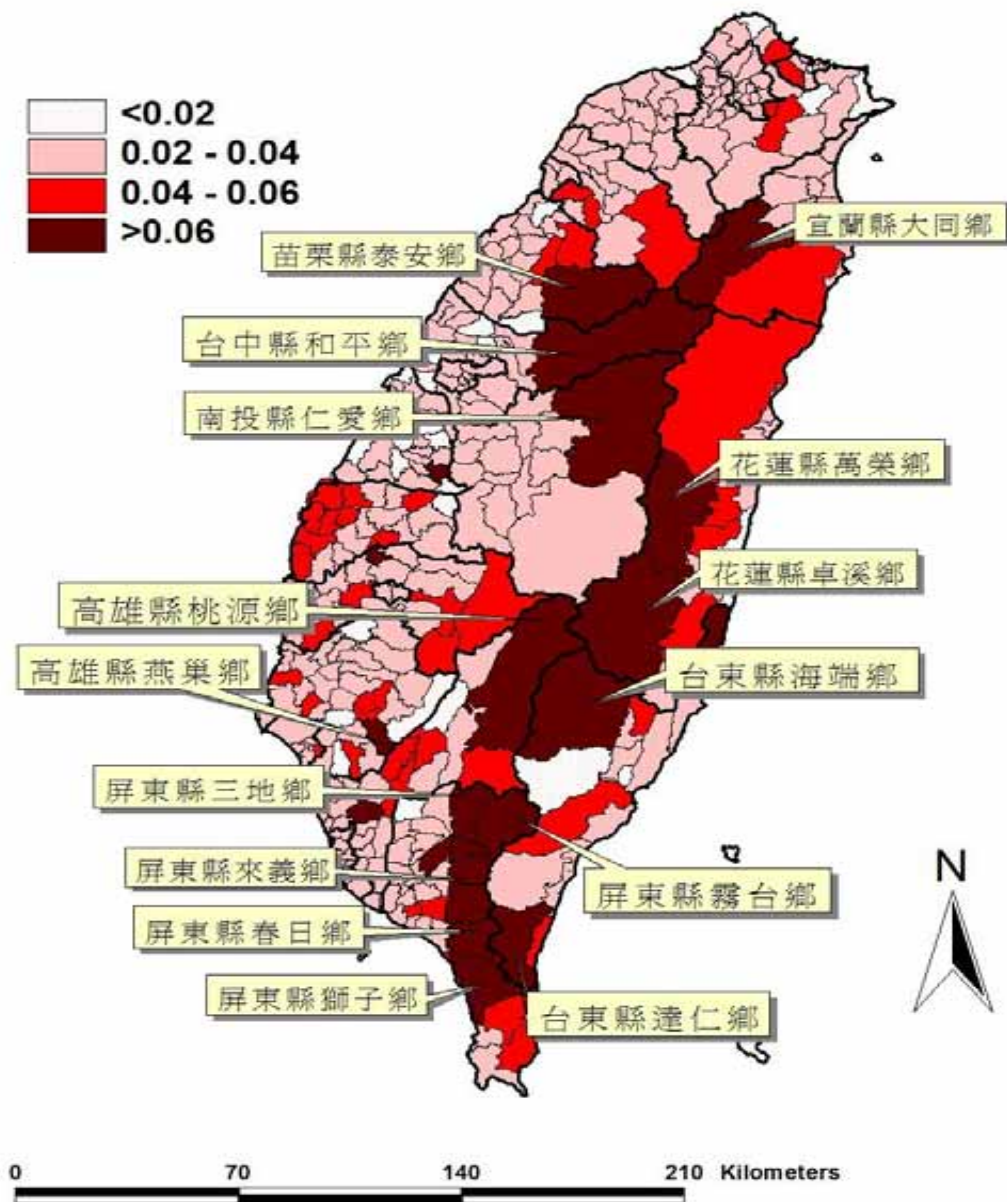
(unit : %)

amount

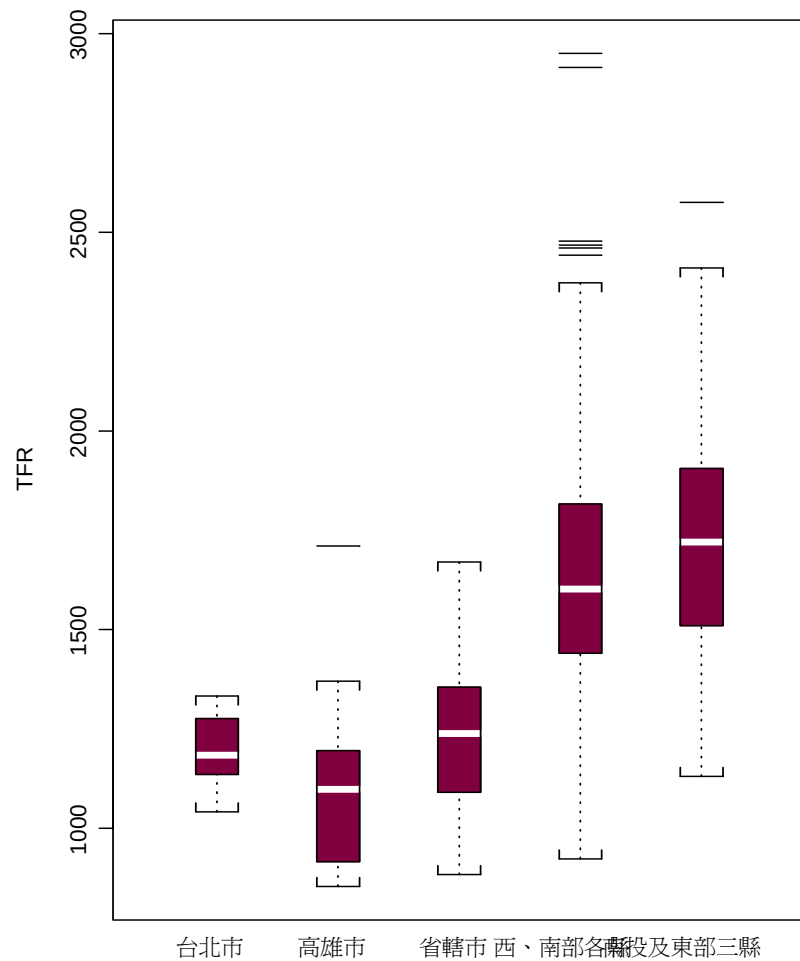
	< 0	11
	0 - 25	42
	25 - 50	137
	50 - 75	111
	> 75	49



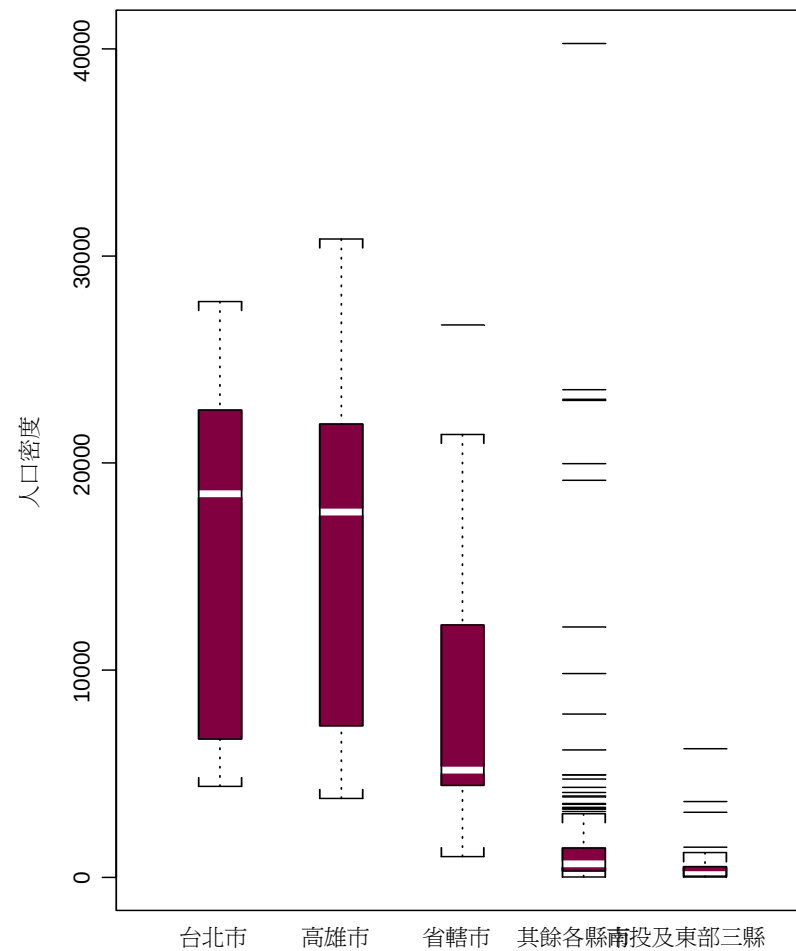
台灣2000年高齡老人輕微失能比例



2001、2002年台灣省地區婦女平均總生育率



2001、2002年台灣地區平均人口密度



台灣各鄉鎮市區婦女生育率與人口密度圖

台灣地區30~34歲育齡婦女生育率「顯著較高」地區

2001年

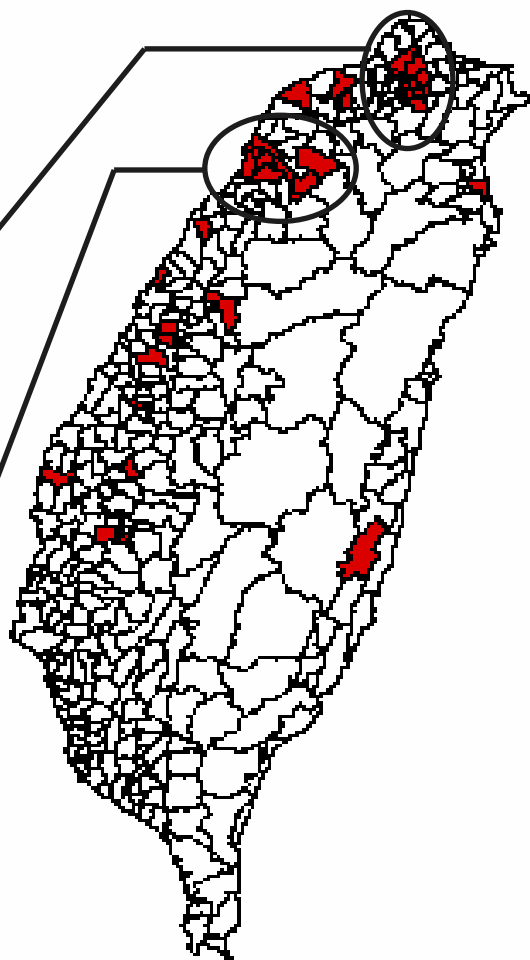
顯著個
數：36

最大聚集
區塊：12

p-value：
0

次大聚集
區塊：8

p-value：
0.001



2002年

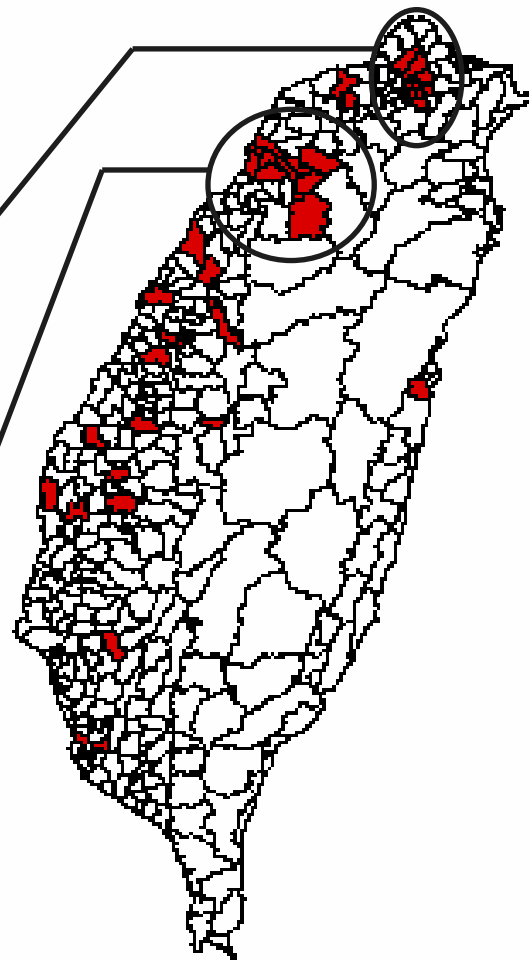
顯著個
數：43

最大聚集
區塊：12

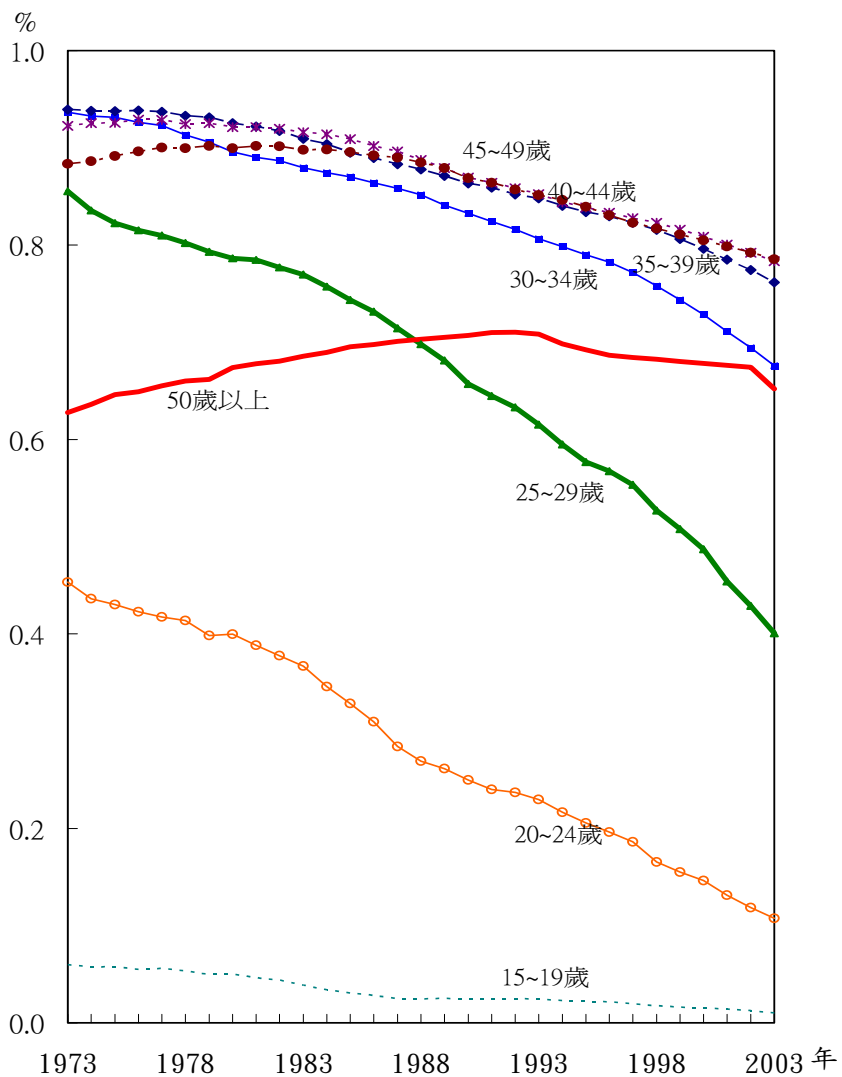
p-value：
0.0008

次大聚集
區塊：10

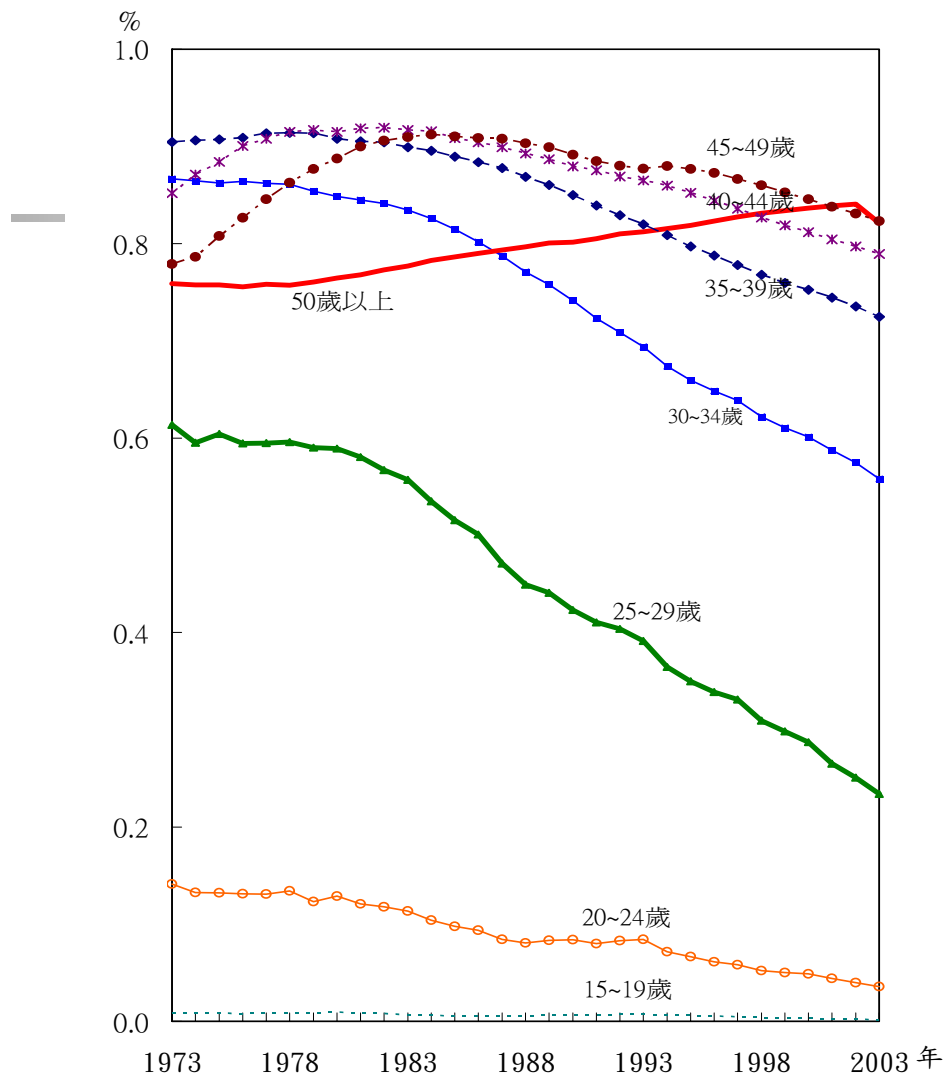
p-value：
0.001



台灣地區歷年年齡別有偶率趨勢圖



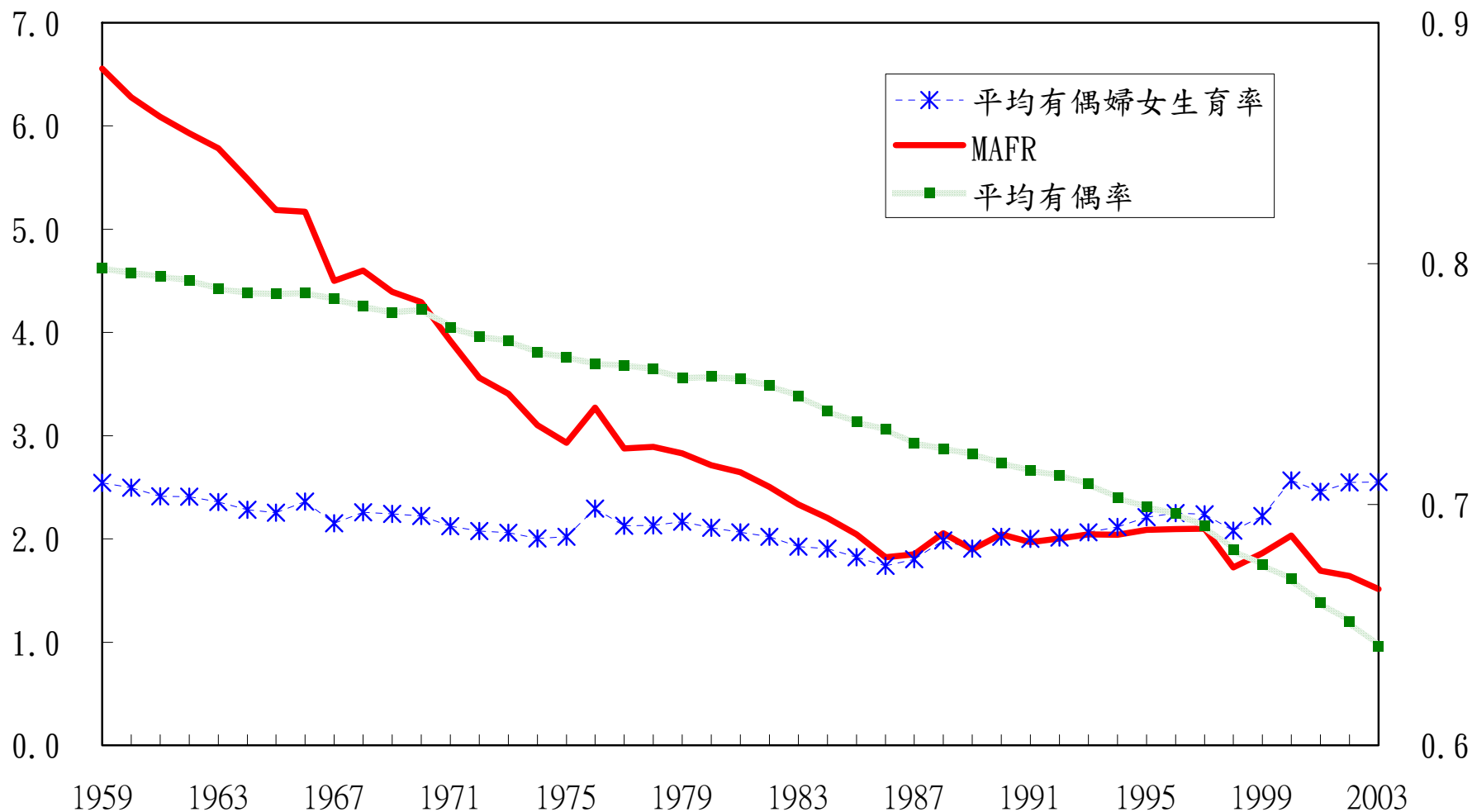
女性



男性

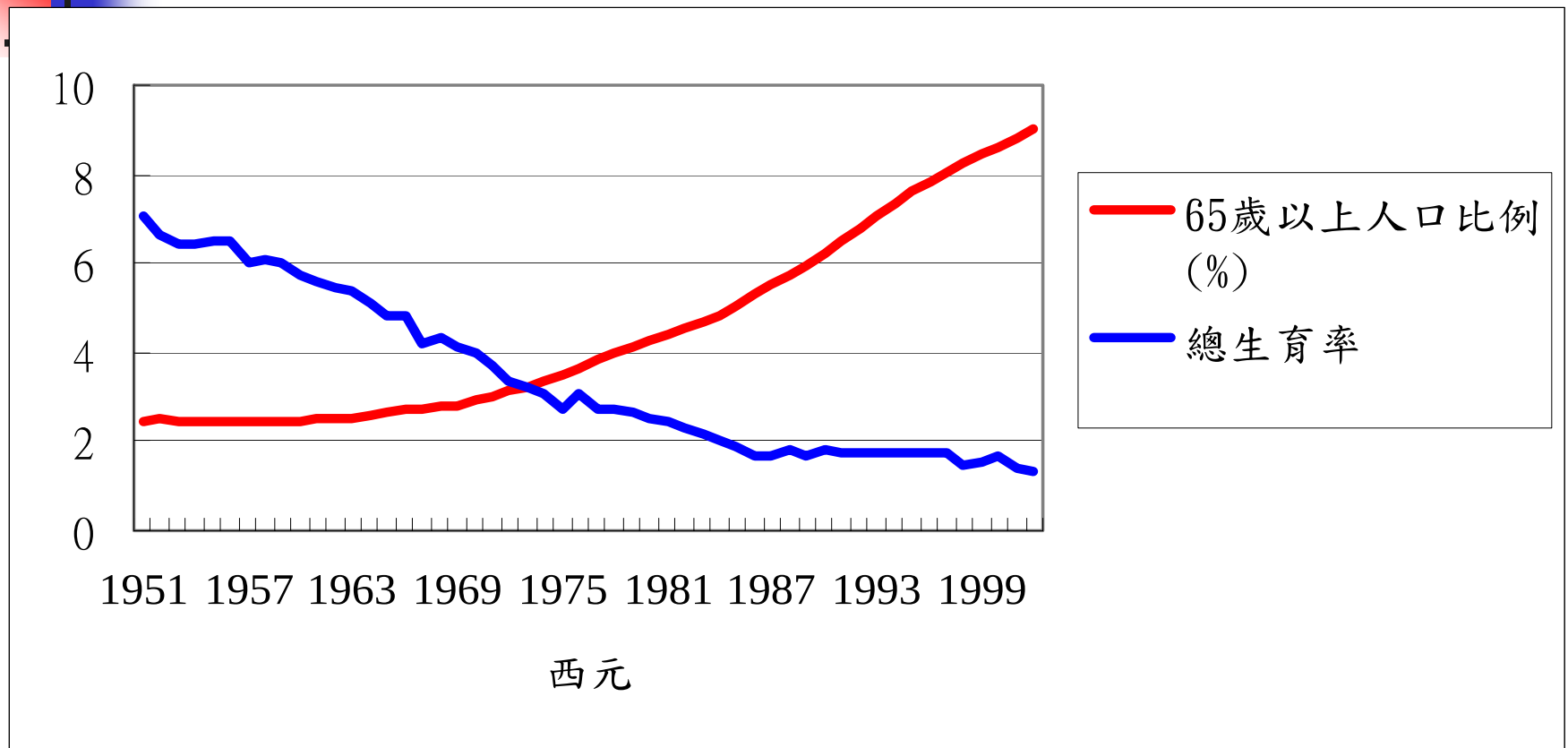
生育率:人

有偶率



台灣歷年有偶率、有偶婦女生育率與有偶婦女累積生育率圖

65歲以上老年人口比例及總生育率趨勢圖

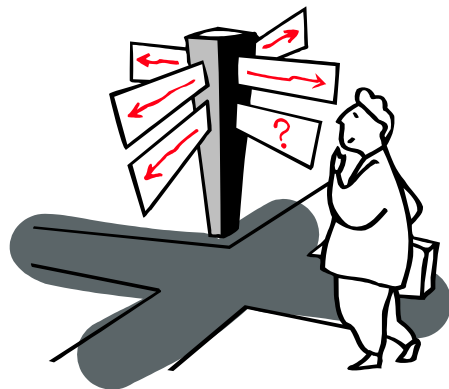


資料來源：1.人力規劃處
2.內政部統計處

- 人口統計思考問題的層面與一般不同，具有時間延滯(Time delay)的想法。

→ 「如果政府現在想辦法降低生育率，將會影響十年後的學校規模，二十年後的勞動力，三十年後下一代的人口數，以及六十年後的退休人數。」

——節錄自「複雜—走在秩序與混沌邊緣」，
天下文化出版



人口統計的重要觀念(平衡公式)

$$P(t+1) = P(t) + B(t) - D(t) + I(t) - E(t)$$

其中

$P(t)$ ：第 t 個時間的總人數

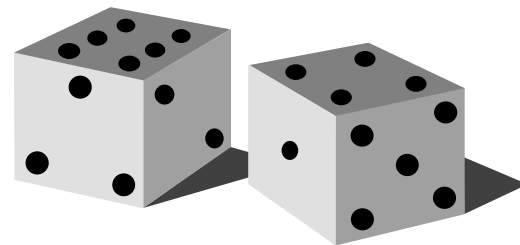
$B(t)$ ：第 t 個時間的出生人數

$D(t)$ ：第 t 個時間的死亡人數

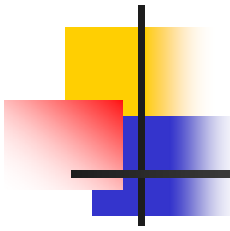
$I(t)$ ：第 t 個時間的移入人數

$E(t)$ ：第 t 個時間的移出人數





- 也就是說，一個國家或地區的人口變化由出生、死亡、移民 3個因素決定。
- 通常出生及死亡的影響較為明顯。
 - 因此，出生及死亡的變化，通常是研究人口統計最重要的課題。
 - 藉由出生率與死亡率反映變化。



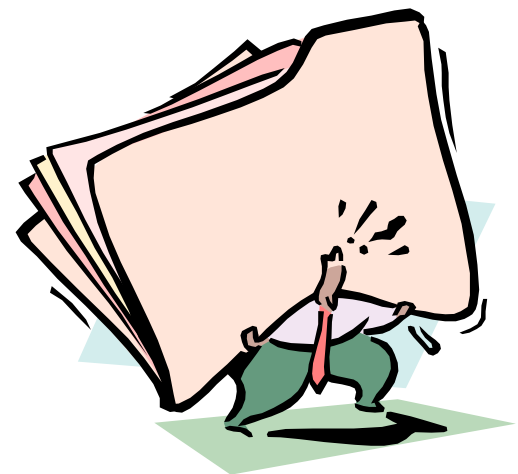
資料蒐集的方式

- 一般將資料蒐集分類成：
 1. 實驗設計(Experimental Design)
 - 包括臨床試驗 (Clinical Trials)，需要較精密計畫，一般分成實驗、對照組。
 2. 抽樣調查(Sampling Survey)
 - 設計問卷，藉由調查取得資訊。
- 目標：藉由蒐集的資料推得訊息。

人口統計資料的來源

- 人口統計的資料絕大多數來自於：

1. 戶口普查(Census)
2. 抽樣調查(Surveys)
3. 生命統計(Vital Statistics ;
或譯為人口動態統計)
4. 戶籍登記(Population Register)

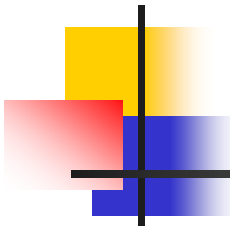




戶口普查

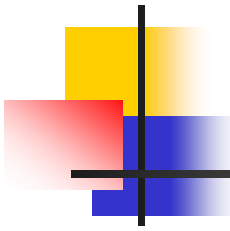


- 戶口普查指某一特定地區於某一特定時刻，對其全部人口本身的自然與法律身份事項，及其有關的經濟與社會特定資料加以調查、蒐集、整理、分析等全部作業。
- 台閩地區戶口普查每十年舉辦一次，最近一次在2000年12月進行，由主計處負責規劃。



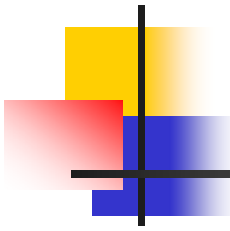
中國古代的戶籍制度

- 在什伍組織的基礎上，進行戶口的登記與統計，記載於文書之上，便成為戶籍。
- 春秋戰國時的戶籍有兩種，一為版籍、一為計簿。
 - 版籍類似出生死亡記錄，由「司民」負責管全國人口的總數。
 - 計簿統計內容較為廣泛，是徵斂兵役、賦役的依據。



中國古代的人口普查

- 《周禮》記：「小司徒之職，掌建邦之教法，以稽國中及四鄰都鄙之夫家九比之數，以辨其貴賤、老幼、廢疾
及頒比法於六鄰之大夫，使登其鄉之眾寡、六畜、車輦，辨其物，以歲時入其數，以施政教，行征令。及三年則大比，大比則受邦國之比要。」



中國古代的人口普查（續）

■ 計簿人口統計包括項目：

→人口數量

→性別

→年齡

→健康狀況

→身份

→財產

→統計時間和方法(每年校計戶口財物、每三年有一次總校，並收受邦國呈報的簿冊)

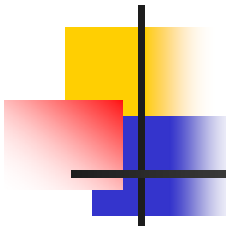


- 兩種不同普查的方法：

1. de facto (實際上)：紀錄普查當時的人口結構，聯合國推薦的方法。

2. de jure (法理上或權利上)：紀錄常住人口，我國、美國、加拿大等國使用的方法。

- 我國普查原則上採事前訪問、事後調查，並於普查後辦理抽樣複查。



- 美國的戶口普查每十年進行一次，在西元為 0 結尾的年份之四月一日進行。與台灣的情形不同，美國普查問卷有長、短問卷之分，約有 1/6 及 5/6 的比例填寫長、短問卷。
- 英國的情形卻有些許不同，戶口普查為五年或十年一次，普查標準年在以 1 或 6 結尾的年份。

調查方法

- 調查方法通常可因獲取資料方法之異，通常分為：

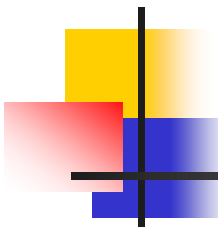
(1) 人員調查法(Personal Survey)

(2) 電話調查法(Telephone Survey)

(3) 郵件調查法(Mail Survey)

(4) 網路調查法(Internet Survey)





- 抽樣調查

- ① 偵錯

- ② 下一次普查前的資料更新

- ③ 提供普查沒有的資料

- 生命統計

- 提供出生、領養、死亡、結婚、離婚、分居等生命統計資訊。

- 戶籍登記

- 提供居住、就學、就業等資訊。

為什麼要抽樣？



■ 為什麼只看一部份的母體？

→ 普查需要較長的時間、較多的經費與人力。(兩次普查之間通常輔以問卷調查等方法彌補資料的不足。例如：1990年及2000年兩次工商業普查間，在1995年進行抽樣調查。)

→ 有時抽樣是唯一可行的方法。

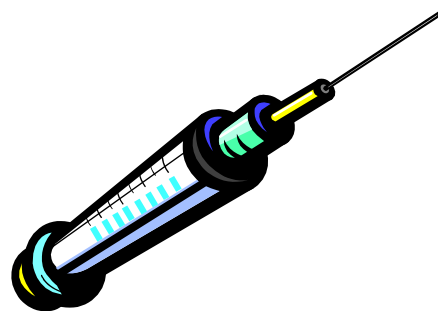
(例如：各年齡婦女墮胎率、攤販調查)

多少樣本才足夠？

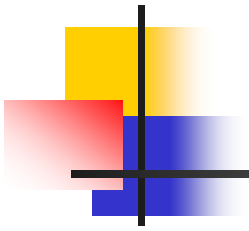
- 抽樣時常見的迷思：

- 樣本數必須至少達到母體的一定比例？

範例(1)一般抽血不多於 10 c.c.，不論大人或小孩。



範例(2)台灣與美國人口數差了10倍以上，
但民意調查多半只抽1,000份左右。



戶口普查面臨的挑戰(台灣)

1. 民眾配合度的降低

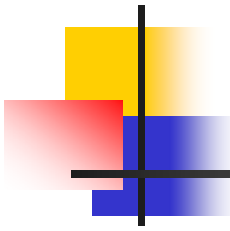
→都市化的程度愈高，完訪率愈低。

2. 調查成本的上升

→2000年的預算原為二十七億元，因九二一地震而刪為不足十億。

3. 資料的精確性

→常住人口與戶籍人口何者為主？



戶口普查面臨的挑戰(美國)

1. 民眾配合度的降低

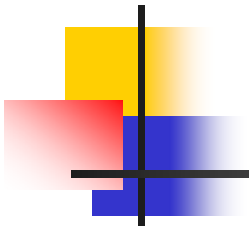
→ 美國1990年普查郵寄問卷中約有 25.9% 的住戶未寄回。

2. 調查成本的上升

→ 1990年普查總預算26億美元。

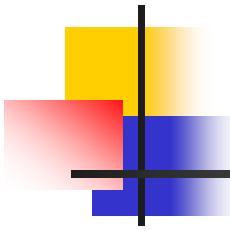
3. 資料的精確性

→ Undercount 的比率升高。



是否調整普查結果？

- 1987 — Department of Commerce 宣佈不調整涵蓋誤差(Coverage error)
- 1989 — Department of Commerce 公佈普查調整原則
- 1991 — Secretary of Commerce 宣佈不調整普查結果
- 1993 — 法院宣判商業部勝訴

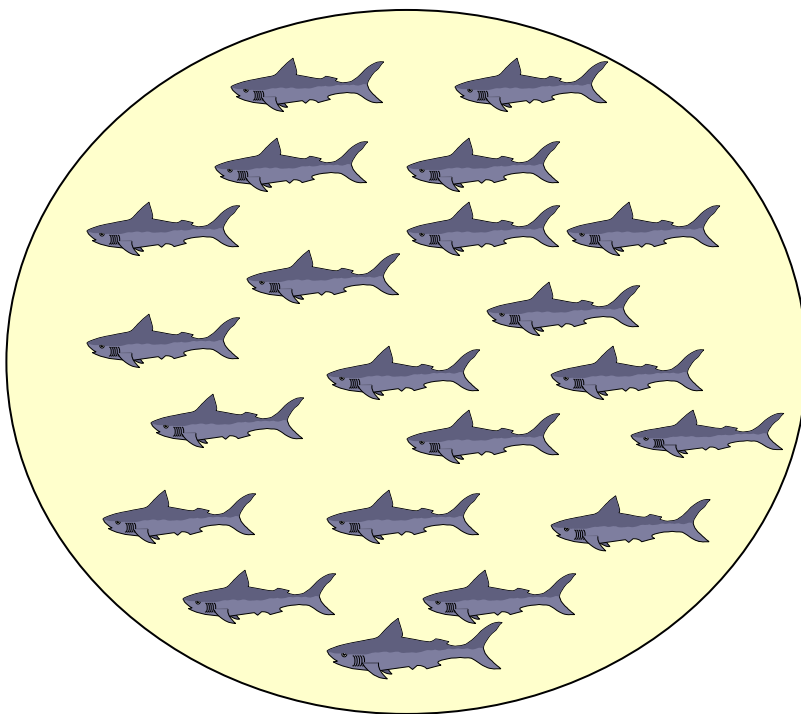


PES(事後複查法)

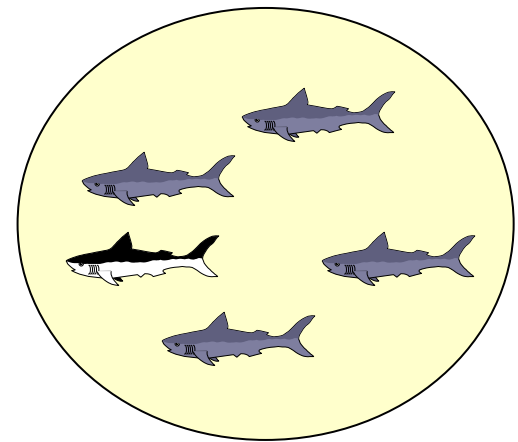
- 抽樣單位：一個或多個區集(Block)組成
- 抽出5,290個Block Cluster，分別核對其中P-sample (PES)及E-sample (普查)的人數
- 再依 Place/Size(7)、Housing Tenure(2)、Region(4)、Census Division(9)、Race(4) 分為116個Poststrata，結合Age及Sex共有1,392個Poststrata
- 全國分成13個較大區域(Evaluation Strata)調整普查結果

Capture-recapture (重複捕取)

第一次捕捉 (普查)



第二次捕捉 (複查)

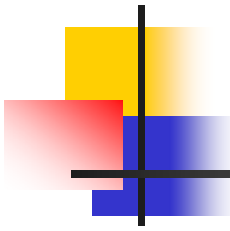


1/5 在普查中未列入！

- DSE(二重系統估計)

Census \ PES	In	Out	Total
In	N_{11}	N_{12}	N_{1+}
Out	N_{21}	N_{22}	N_{2+}
Total	N_{+1}	N_{+2}	N_{++}

$$N_{++} = (N_{+1})(N_{1+}) / N_{11}$$



有關DSE的爭議

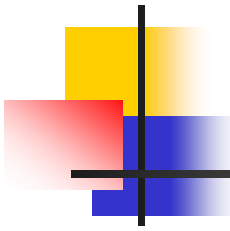
Statistical Science (1994), p. 458-537

- Breiman → Data Quality
(Errors in the PES and Matching)
- Freedman and Wachter → Heterogeneity
(Large Heterogeneity within Poststrata)
- Belin and Rolph → 建議以決策的觀點考量



•Doubts about 2000 Census

- Low response rates of the elderly in 2000 Population Census
 - 93% responded for ages 90~99 and 80% for ages 100 and over
- No Post Enumeration Survey in 2000 Population Census (Budget cut due to Earthquake 921)
 - 5% of households were surveyed in 1990 Population Census



2003年台灣老年人口普查

- 為了確定老年人口資料的正確性，台灣政府在2003年8月至2004年2月對89歲以上的老年人進行普查，人數約有5萬人。有別於2000年戶口普查，本次普查的老人必須親自回答問題。(若無法回答，則由其照護者代答。)
- 雖然耗費不少經費及人力，仍有不少老年人無法找到。

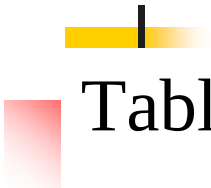


Table 1. Response and non-response rates in 2003 Census

	Age 89	Ages 90~94	Ages 95~99	Ages 100+	Total
Total Population	12,597	23,898	11,190	2,484	50,169
Response	92.17%	91.39%	88.38%	66.39%	89.67%
Non- response	7.83%	8.61%	11.62%	33.61%	10.33%

5,169 non-responses

網際網路的有用資訊

- 搜尋引擎：(中、英文各國語言皆可)

<http://www.google.com>

- 內政部統計處(人口統計名詞字典)

<http://www.moi.gov.tw/W3/stat/>

- 行政院主計處(普查局、統計局)

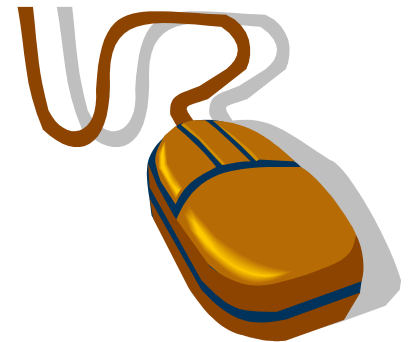
<http://www.dgbas.gov.tw/>

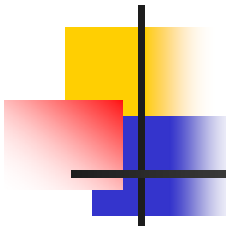
- 行政院衛生署(衛生統計)

<http://www.doh.gov.tw/>

- 台灣大學人口與性別研究中心

<http://ccms.ntu.edu.tw/~psc/index.html>





- 台灣過去因為人口資料品質良好，常常被各國人口學者研究及引用，因而台灣也有「人口實驗室」的美稱。
- 人口統計研究的基礎在於完整的資料及良好的資料品質，在座各位都是維護台灣優良資料的貢獻者，更在第一線直接面對原始資料。希望我們能繼續保持優良傳統，讓台灣在世界上持續發聲。



報告完畢，

敬請指教！

