人口統計應用與實務

政治大學統計系 余清祥

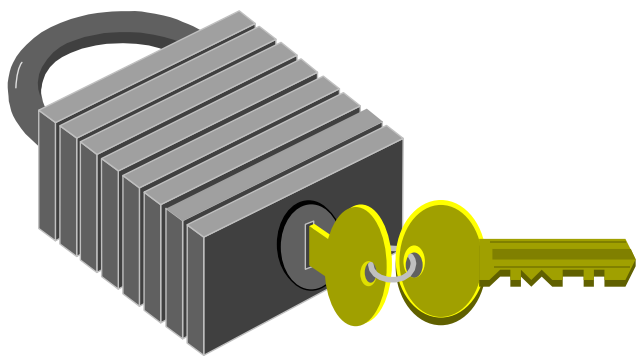
2003年12月2日

Email: csyue@nccu.edu.tw

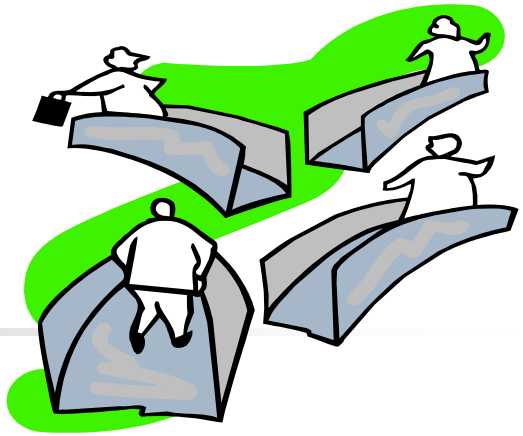
課程下載：http://csyue.nccu.edu.tw

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。



人口統計的定義



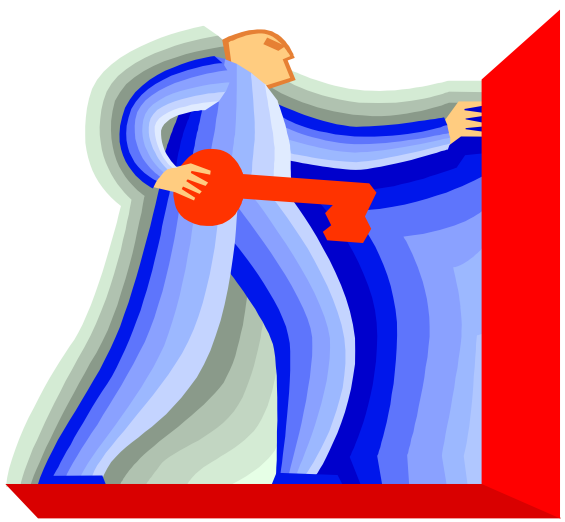
- 人口統計或人口統計學(Demography)為研究一個地區或國家人口的學門，主要涵蓋人口總數、人口結構、與人口變遷及發展等方面。其精細之意義為：「對人類人口數量及其因出生、死亡及移民所引起之變動之研究。」近年來其範圍擴大為：「對於生育、婚姻、移民及死亡等資料之蒐集及統計的分析。」

定義問題

蒐集資料

分析資料

詮釋結果



有趣(或殘酷)的範例

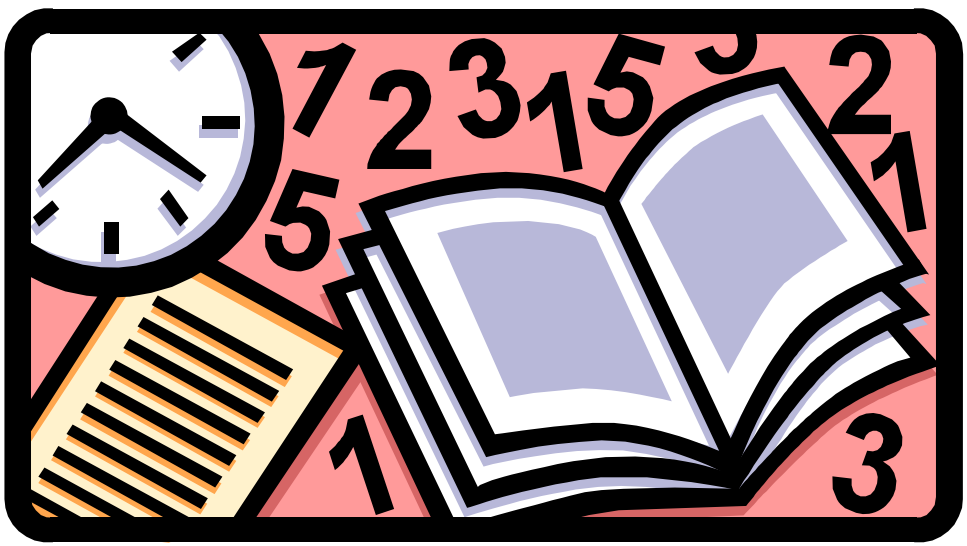
- 一位教授及其學生到非洲探勘，在一望無際的平原上被獅子追殺。眼看即將被追到，學生趕緊換上球鞋，教授說：「換上球鞋也跑不過獅子。」學生卻說：「我不必跑贏獅子，只要跑贏你就夠了。」

→ 真正的問題是甚麼？

統計的第三型誤差

- Type III error (error of the third kind):

  → Giving the “right” answer to the wrong question (Kimball, 1957)



I、統計研究的首要步驟

- 獲取研究問題的相關背景知識
- 確立問題的目標(研究目的)
- 以統計的語言定義問題

→如果與其他人合作，儘量「多發問」！

另一個定義問題的範例

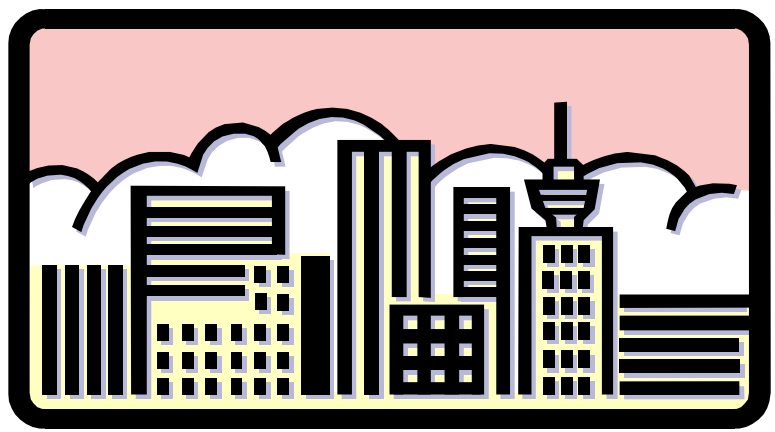
- 某家旅館重新整修內部，將客房數增為原先的1.25倍，但電梯數維持不變，房客因等待時間增長而抱怨連連。

解決方案：

→ 增加電梯數？

→ 加快電梯速度？

→ 電梯門加設鏡子？



真正的問題在哪裡？

- 有時呈現在表面的因素並非造成問題的實際原因，解決方案需從另一方向或更深層的部分去探索。
- 討論：近來關於氣象預測的話題(尤其是預測的準確性)廣為大家討論，請問你/妳覺得問題為何？

  → 如何解決？

Bargain Prices

The Situation: A local merchant on Main Street in Ann Arbor was having difficulty selling a health food mix from the rain forest called Rain Forest Crunch, which was a hot selling item in other stores. Part of the attractiveness of Rain Forest Crunch was that it was indeed from the Brazilian rain forest and part of the proceeds of the sale went to protect the rain forest. The instructions given by the store manager: *"Lower the price of the item to increase sales."* Rain Forest Crunch still did not sell. The manager lowered the price further. Still no sales. After lowering the price two more times to a level that was well below the competitors', the item still did not sell. Finally, the manager walked around the store, and studied the display of Rain Forest Crunch. Then the real problem was uncovered. The problem was not the high cost of the item; the **real problem** was that it was not in a prominent position in the store to be easily seen by the customers. Once the item was made more visible, sales began to soar.[4]

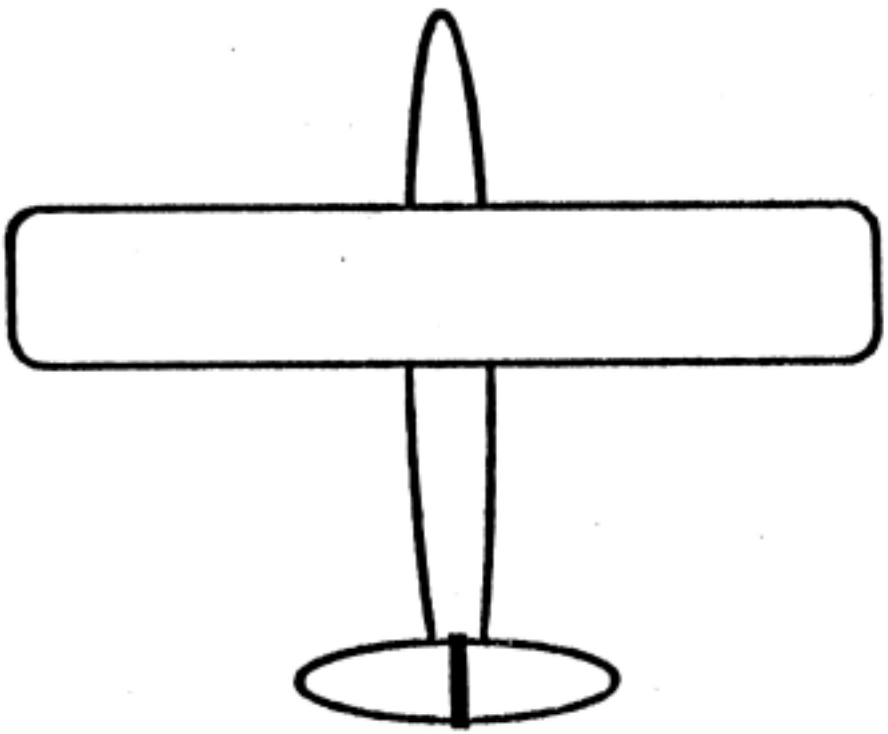
Price Reduced

~~$14.99~~

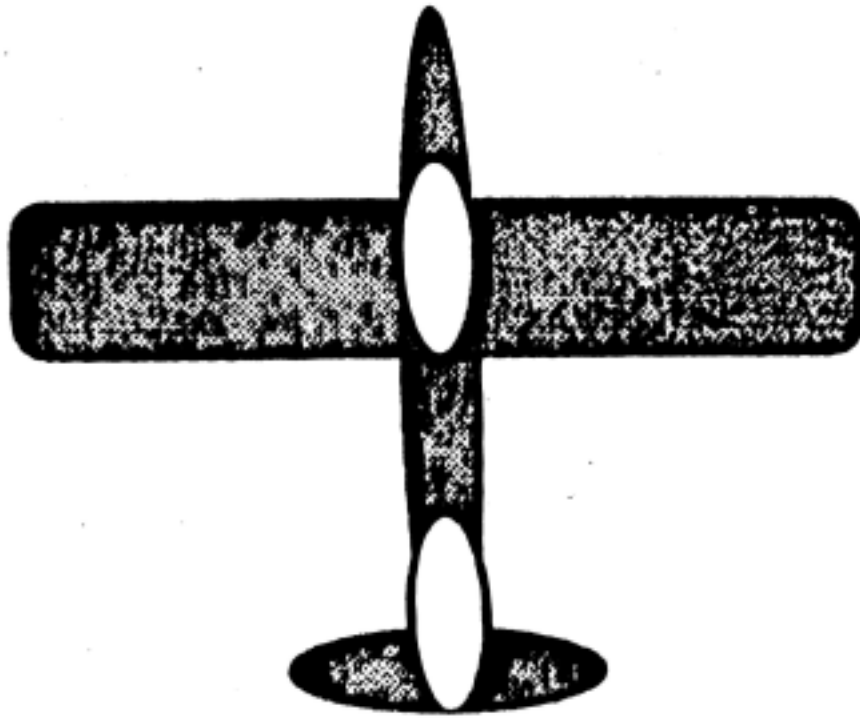
~~$12.99~~

~~$10.99~~

$9.99



Before



After

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

Face the Reality?

- 美國某家石油公司以管線的方式將阿拉斯加的天然氣輸往本土，但因天然氣中含有腐蝕性物質(二氧化硫)，連接管線間的測量表常遭腐蝕，造成天然氣外洩，該公司必須派人不定時檢修量表。

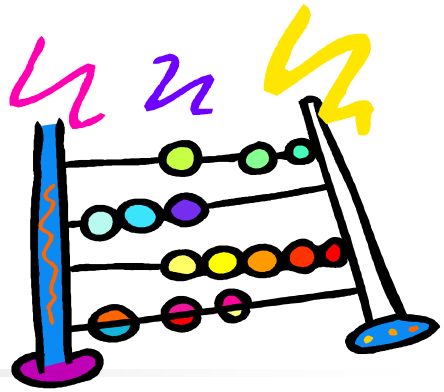
→該公司希望研發耐腐蝕的量表，但橡膠墊片會與二氧化硫作用。

"Uh, yeah, Homework Help Line? I need to have you explain the quadratic equation in roughly the amount of time it takes to get a cup of coffee."

II、如何量化與測量？

- 在確定問題的定義後，下一個關鍵的步驟是如何量化與測量。因為統計是分析數量化的資訊，不適當的測量值通常無法協助我們取得資訊，反而可能產生雜訊，干擾、誤導我們的判斷。

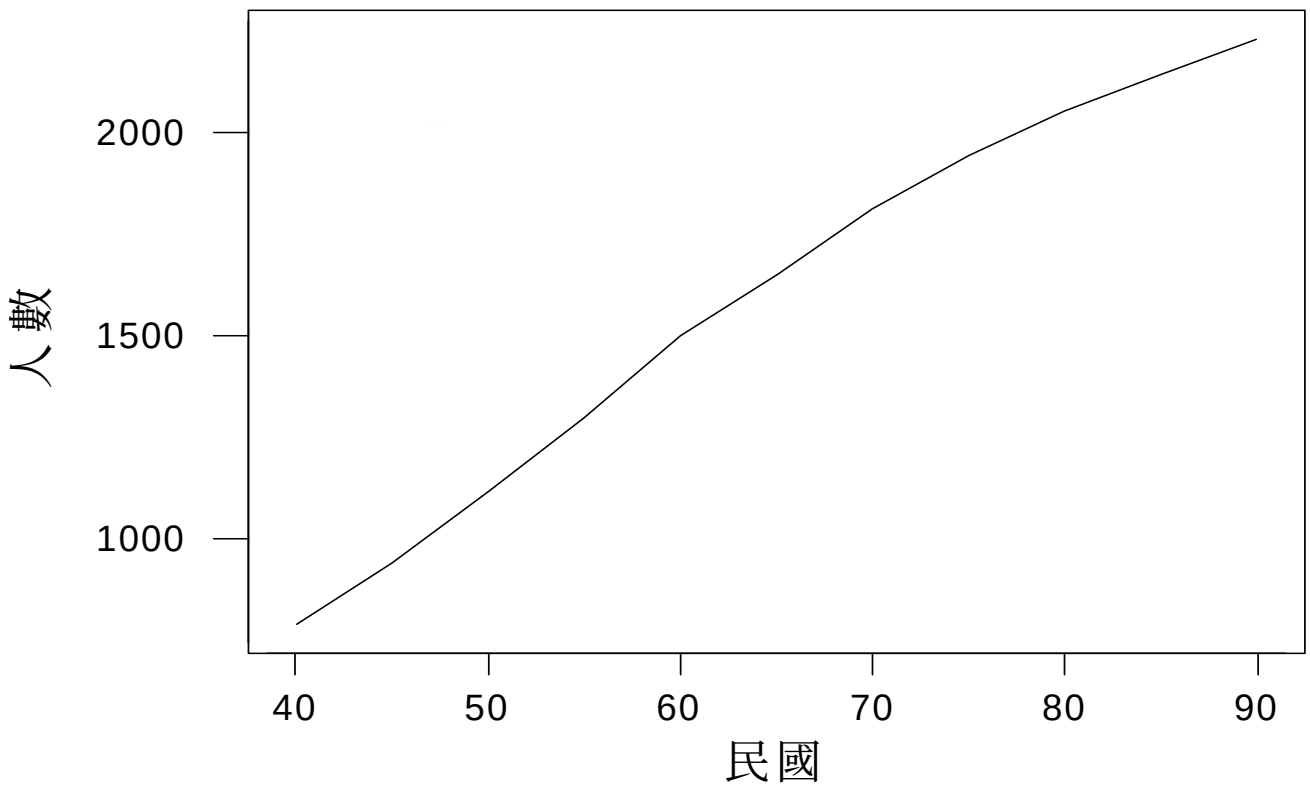
問題研究實例



- 研究自二次世界大戰後，台灣地區居民總人口數的變化。

→若以總人口數而言，台灣地區的總人口數自民國40年的787萬人，持續上升至60年的1500萬人、民國80年的2056萬人、民國90年的2234萬人，台灣地區的人口數逐年直線上升。

→似乎可使用直線迴歸預測未來人口數。



台灣地區人口數逐年直線上升！？

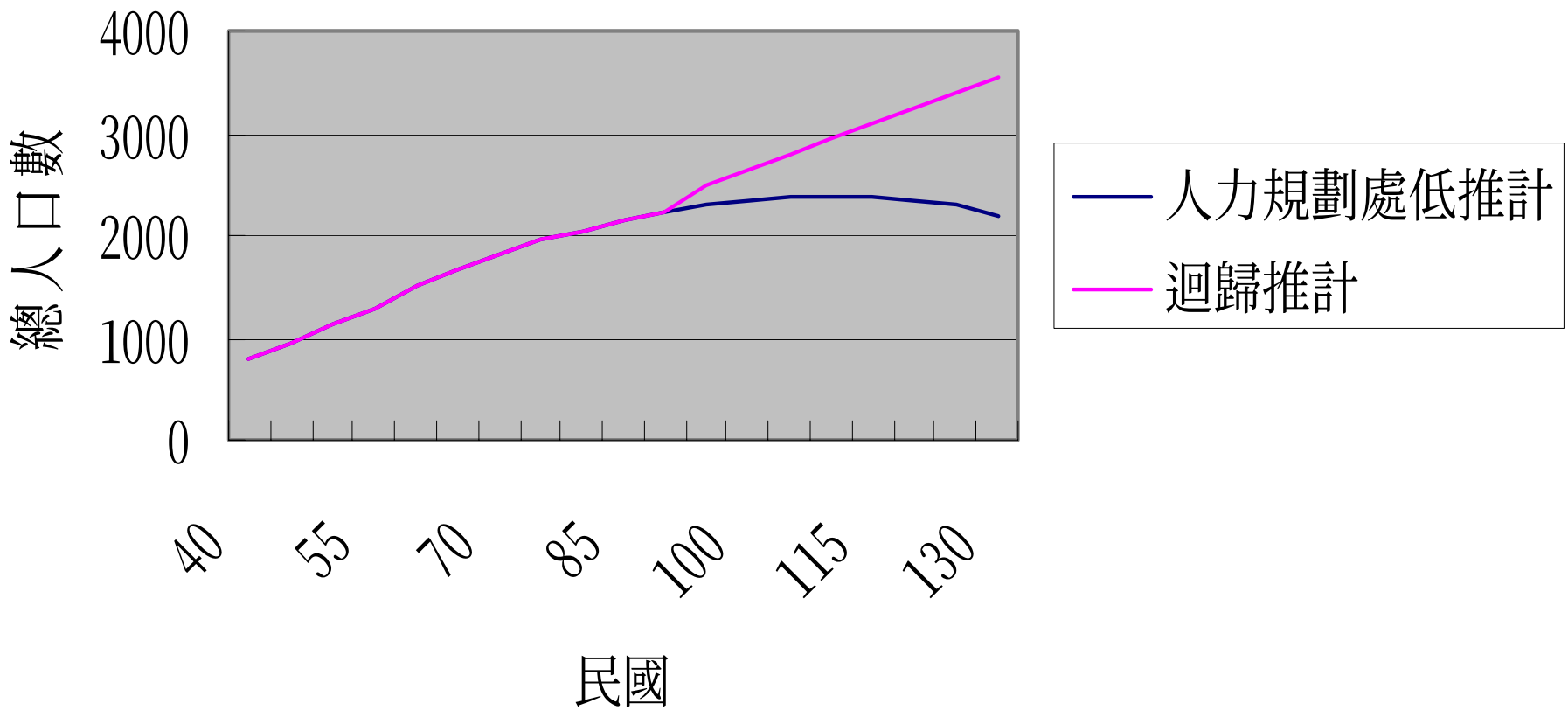
- 以時間為自變數，台灣地區人口數為應變數，得出以下結果：

The regression equation is
人數 = - 360 + 30.0 民國

Predictor	Coef	SE Coef	T	P
Constant	-359.95	79.99	-4.50	0.001
民國	29.996	1.196	25.09	0.000

S = 62.71　　R-Sq = 98.6%　　R-Sq(adj) = 98.4%

→因為R^2很大，似乎以迴歸分析來預測未來人口數頗為合適。在民國100年、120年，台灣地區人口數分別為2640萬、3240萬人。



哪一個預測方法較為正確？

問題研究實例(續)

- 台灣地區居民的人口數的變化。

→以迴歸方法推計，台灣地區似乎有人口膨脹的危機，但事實上恰好相反。單從表面上的數字來判斷，非常容易誤判。

→在人口統計的領域，如何正確地使用統計方法，通常需要使用專業知識與常識判斷，最忌諱以現象面為研究對象，任意套用統計分析。

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

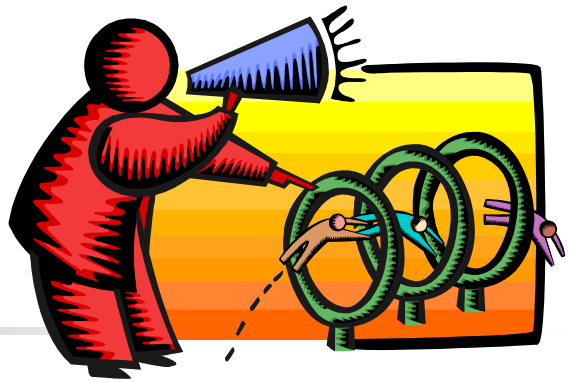
E(t)：第 t 個時間的移出人數

- 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。
- 問題：如何有效地定義出生率及死亡率？

出生率與死亡率

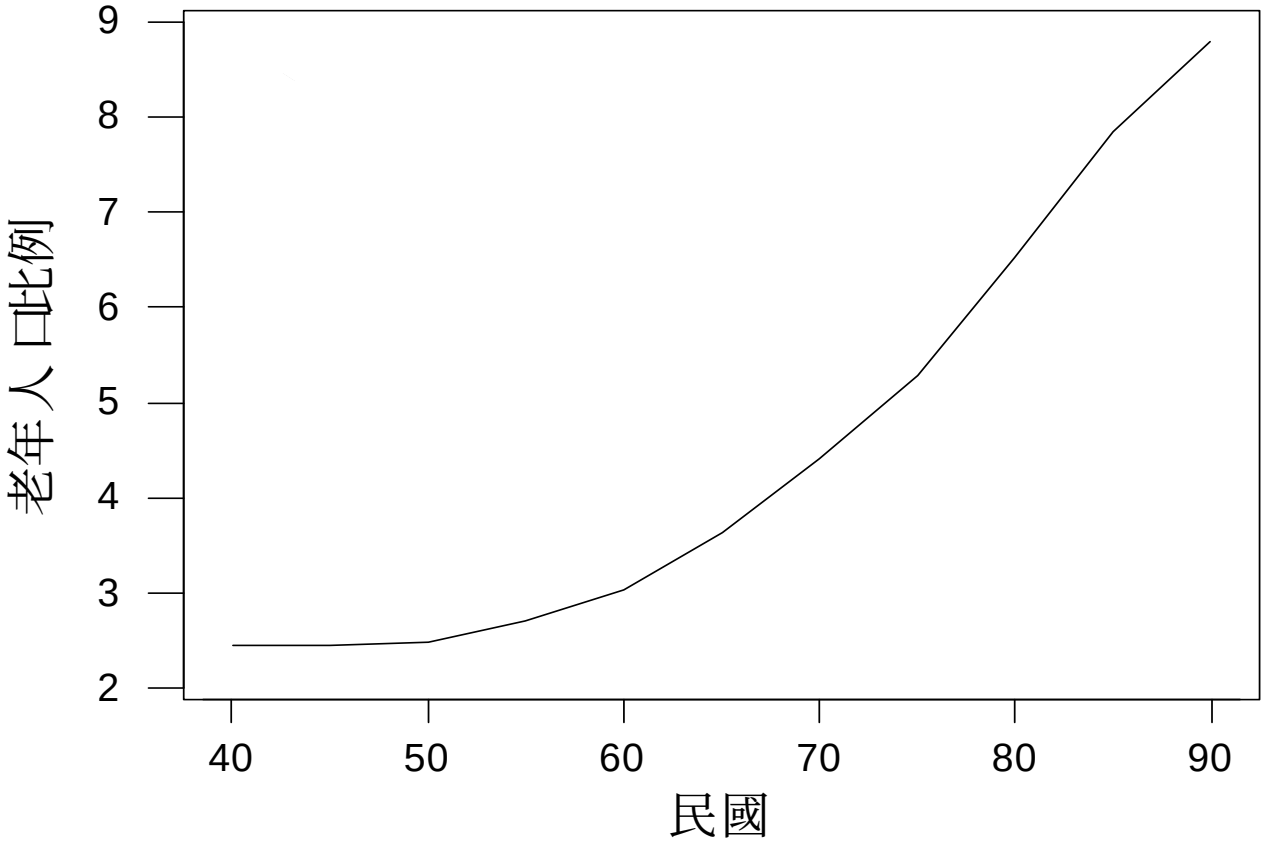


- 粗略率(Crude Rate)為直觀的定義。

  →粗出生率＝出生總人數 ÷ 總人數

  →粗死亡率＝死亡總人數 ÷ 總人數
- 粗人口增加率＝粗出生率 – 粗死亡率
- 但粗略率通常無法反映實際狀況，例如台灣現在的粗人口增加率上升，但青壯人口比例下降，人口老化現象日趨明顯。

台灣 地區二 次大戰後老年人 口比例



台灣地區老年人口比例快速上升！

修正的定義

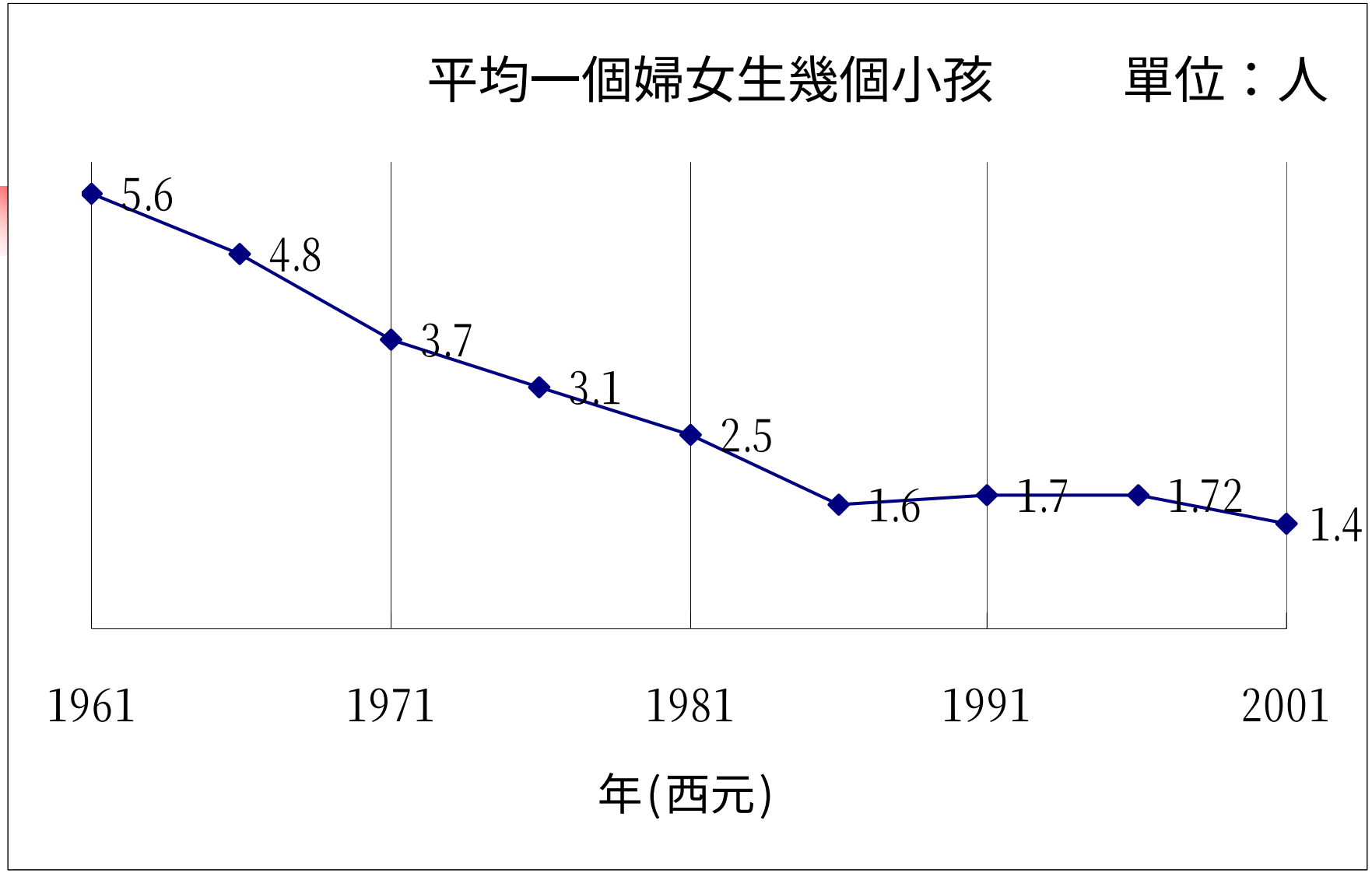
- 年齡別生育率及年齡別死亡率

→f_x=*生母x歲的出生人數÷x歲婦女人數*

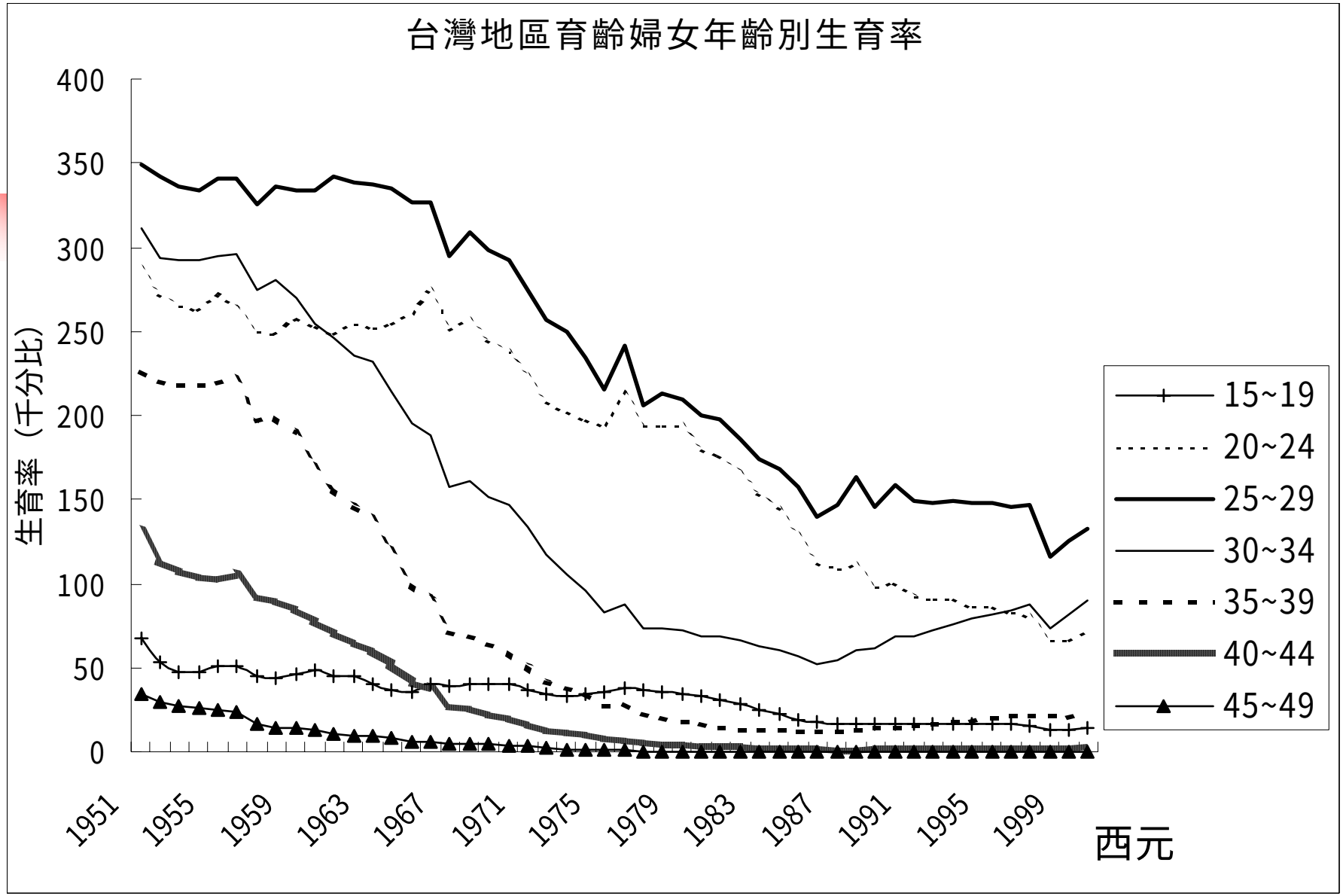
→q_x= *x歲死亡人數÷x歲總人數*

- 總生育率(Total Fertility Rates)：

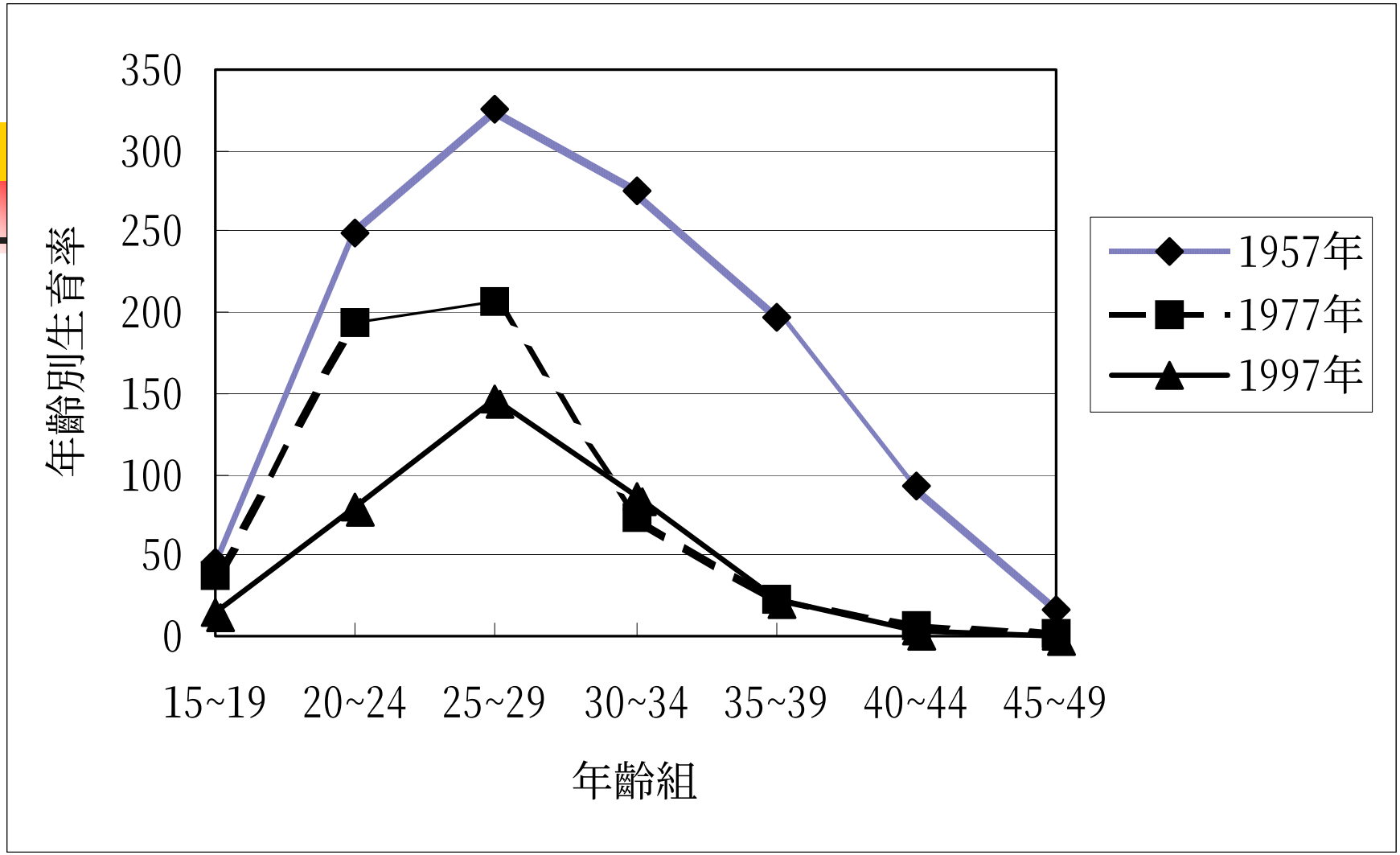
$$TFR = \sum_{x=0}^{\infty} f_x = \sum_{x=\alpha}^{\beta} f_x$$



台灣地區歷年總生育率變化圖



台灣地區育齡婦女生育率分佈圖



台灣地區育齡婦女生育率分佈圖

生育率模型的探討

- 早期的生育率分析尋求總生育率(TFR)的模型，但如同直接預測總人口數的狀況，不容易獲得正確的趨勢。

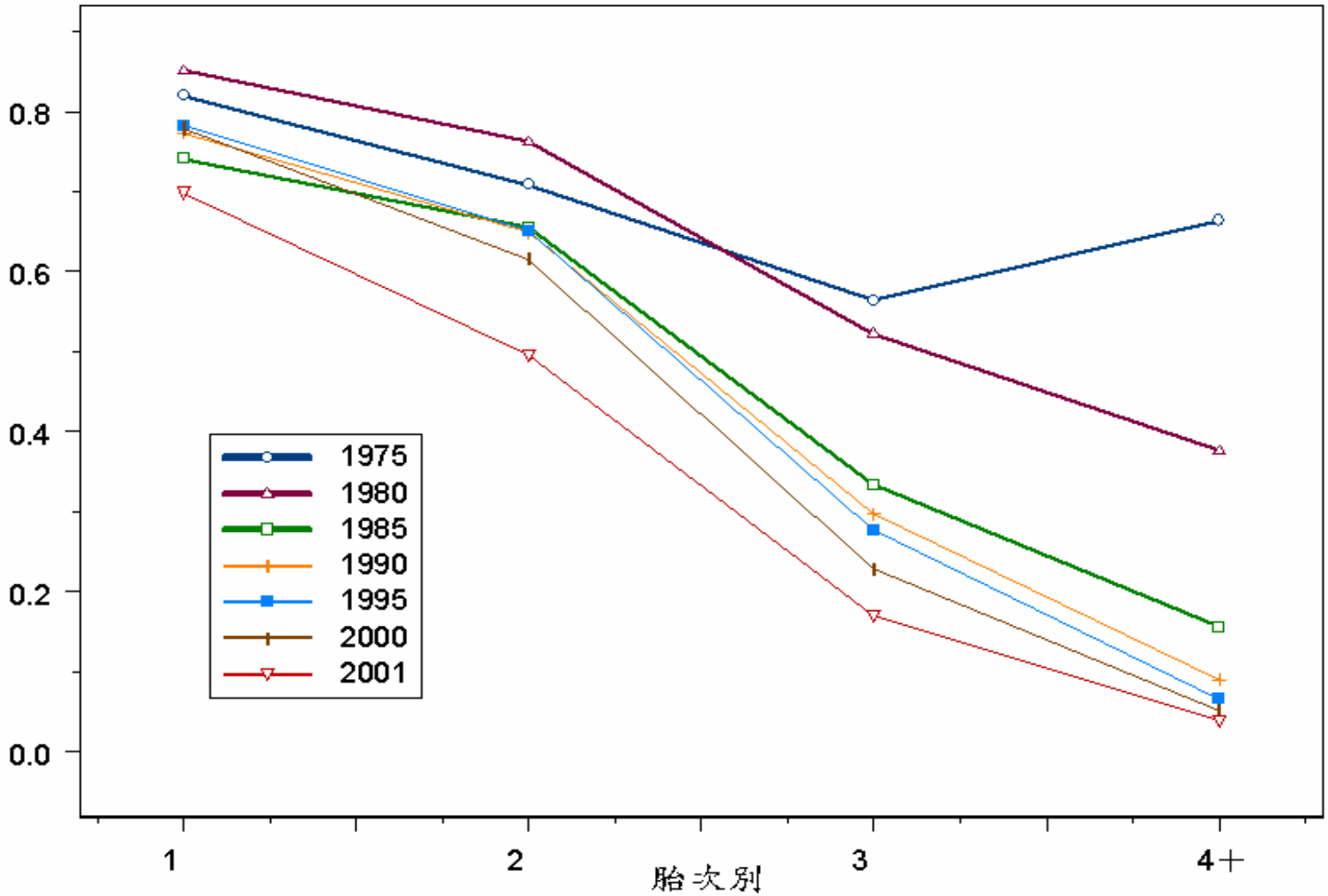
→之後，較常見的推估方式為針對每一個年齡別的婦女生育率(例如：15至49歲，每五歲一組)，分別配適統計模型。常見的模型有Gamma或Gompertz。

- 然而，近年來發現單從育齡婦女生育率著手仍然不夠精確，遂有學者提出以下幾種變通方法：

→紀錄婦女胎次別生育率。因為總生育率字面上的意義為婦女一生中的平均生育數，由胎次別著手相當合理。

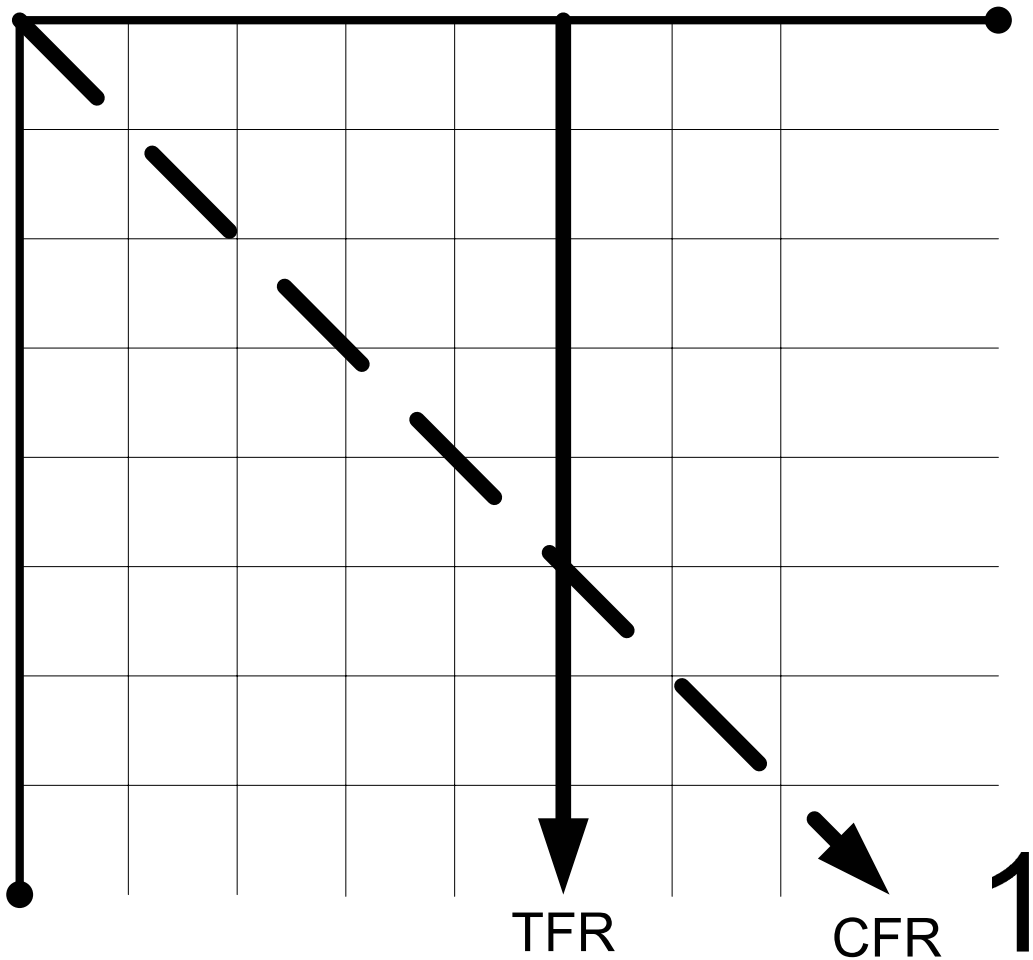
→以世代生育率(Cohort Fertility Rate；CFR)取代總生育率。

→有偶婦女生育率取代一般育齡婦女生育率。

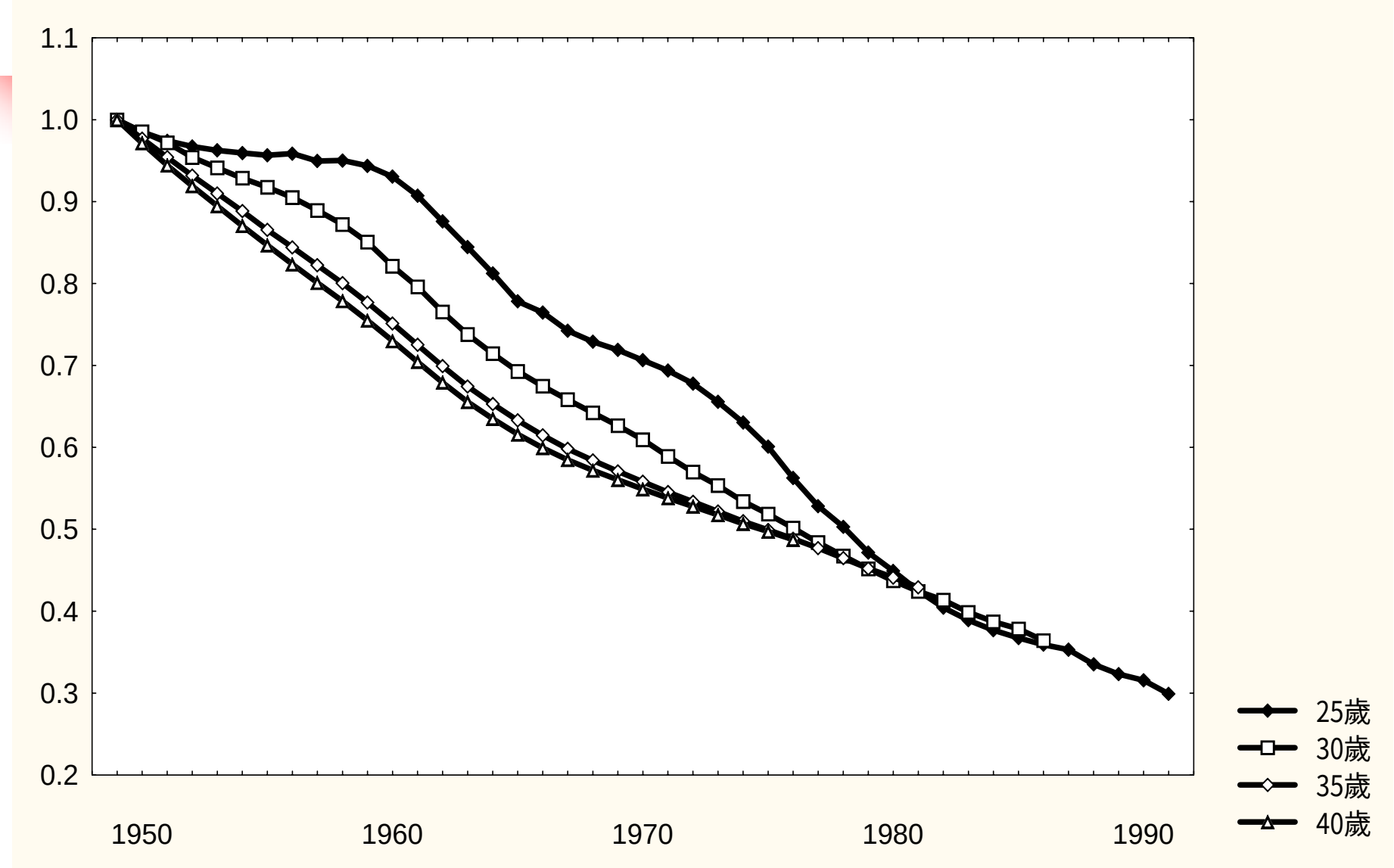


台灣地區育齡婦女胎次別生育率

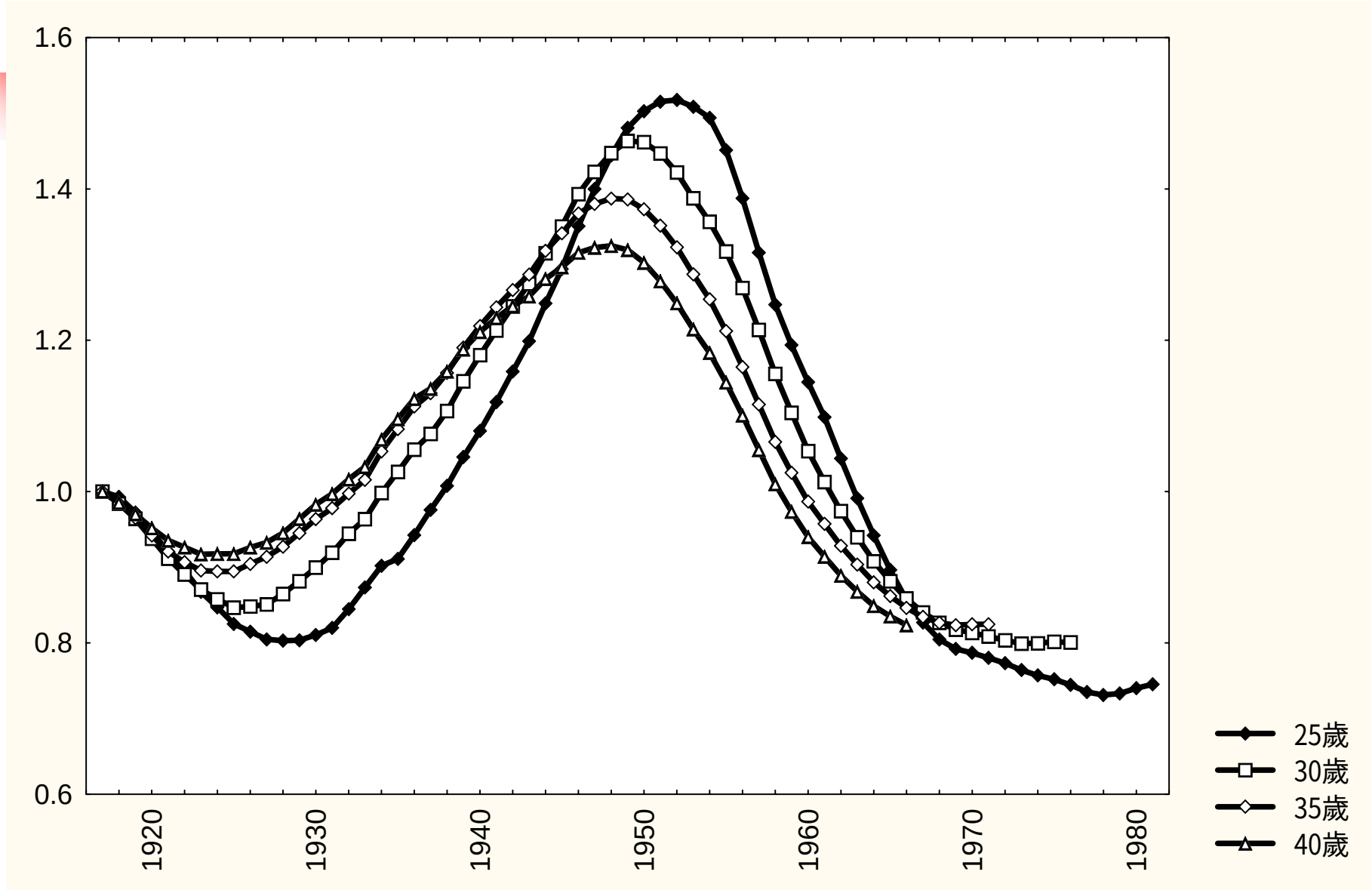
TFR與CFR比較圖

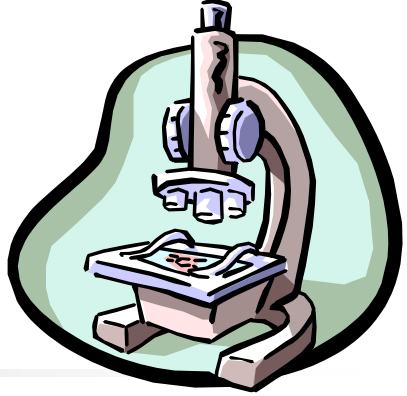


台灣CFR累積生育率


1.1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
1950
1960
1970
1980
1990
25歲
30歲
35歲
40歲

美國CFR累積生育率


1.6
1.4
1.2
1.0
0.8
0.6
1920
1930
1940
1950
1960
1970
1980
25歲
30歲
35歲
40歲

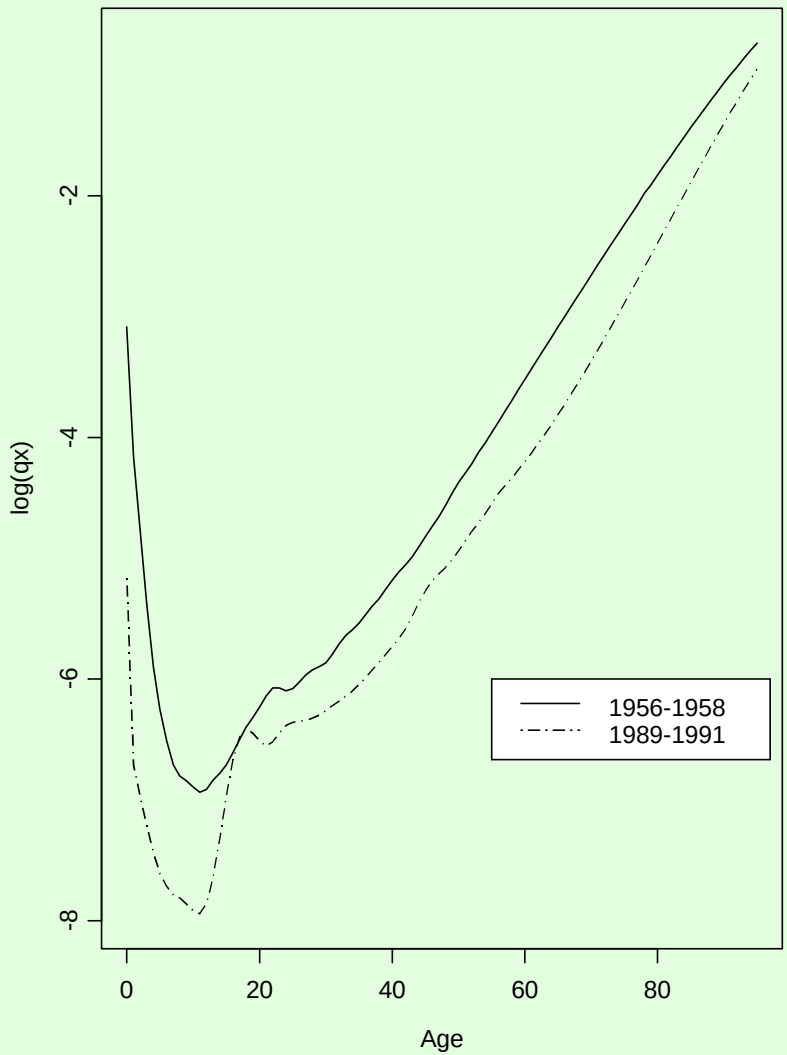


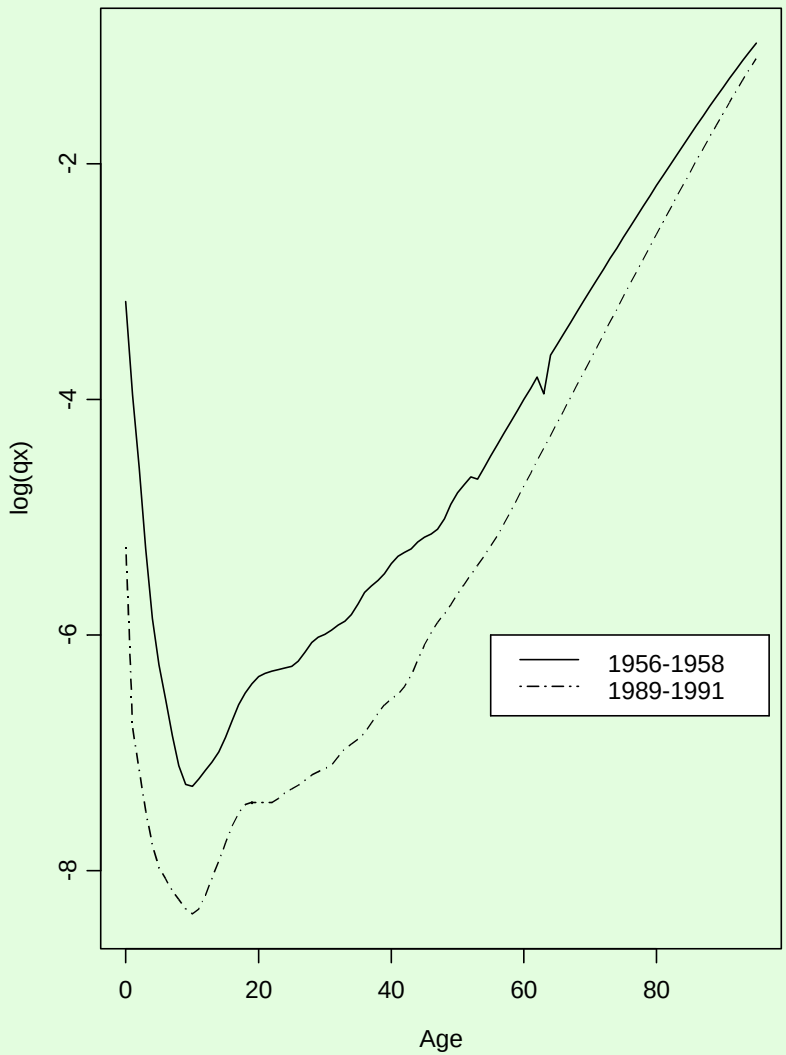
死亡率模型的探討

- 死亡率與生育率類似，因年齡而有非常大的差異，單從某個數值判斷會有偏差。
- 年齡別(Age-specific)死亡率：

→因為各年齡層的死亡率不進相同，通常在嬰幼兒時的死亡率較高，再隨年齡增加而逐步下降，在十歲前後到達最低點；之後隨年齡緩慢上升。

$$ {}_nq_x = \frac{\text{從年齡}x\text{開始}n\text{年內死亡人數}}{x\text{歲的生存人數}} = \frac{{}_nD_x}{P(x)} $$





台灣地區兩性死亡率的對數曲線

死亡率的調整值

- 因為人口結構不同，若直接以粗死亡率的數值當作死亡率高低的判斷標準，則老人比例高的地區，死亡率通常比較高。

- 根據比較地區的死亡資料是否完整，選用兩種調整方式：

1.直接調整法(Direct Method of Adjustment)

2.間接調整法(Indirect Method of Adjustment)

- 如果要比較的地區之年齡別死亡率可完整獲得，可使用直接調整法：

$$ADR_D = \frac{\sum_x {}_nP_x^s \cdot {}_nm_x^j}{\sum_x {}_nP_x^s} = \frac{\sum_x {}_nP_x^s \cdot {}_nm_x^j}{P^s}$$

其中 ${}_nP_x^s$ 為標準母體 x 到 $x+n$ 歲的人數

${}_nm_x^j$ 為第 j 地區 x 到 $x+n$ 歲的死亡率

- 如果無法獲得年齡別死亡率，只能獲得各年齡層的人數，可使用間接調整法：

$$ADR_I = \frac{D^j}{\sum_x {}_n m_x^s \cdot {}_n P_x^j} \cdot \frac{D^s}{P^s}$$

也就是說每一地區藉由與標準母體的*標準死亡比*(Standard Mortality Ratio；SMR)來比較：

$$SMR = \frac{D^j}{\sum_x {}_n m_x^s \cdot {}_n P_x^j}$$

範例(Brown)：
密西根州與佛羅里達州

MICHIGAN				
Age Group	Population on July 1, 1985 (thousands)	Percent	Deaths in 1985	${}_nm_x \cdot 10^3$
0-5	662	7.3	1,889	2.85
5-15	1,366	15.0	385	0.28
15-25	1,568	17.3	1,543	0.98
25-35	1,600	17.6	2,049	1.28
35-45	1,186	13.1	2,592	2.19
45-55	842	9.3	4,512	5.36
55-65	844	9.3	11,460	13.58
65-75	618	6.8	18,264	29.55
75-85	306	3.4	20,637	67.44
85+	95	1.0	15,381	161.91
Total	9,087		78,712	8.66

FLORIDA				
Age Group	Population on July 1, 1985 (thousands)	Percent	Deaths in 1985	${}_nm_x \cdot 10^3$
0-5	750	6.6	2,241	2.99
5-15	1,348	11.9	419	0.31
15-25	1,677	14.8	1,847	1.10
25-35	1,775	15.6	2,713	1.53
35-45	1,402	12.3	3,270	2.33
45-55	1,105	9.7	5,986	5.42
55-65	1,308	11.5	15,301	11.70
65-75	1,201	10.6	29,875	24.88
75-85	641	5.6	36,292	56.62
85+	161	1.4	23,131	143.67
Total	11,368		121,075	10.65

- 若依直接調整法計算，可得

  ADR_D(Michigan) = 9.14224

  ADR_D(Florida) = 8.12808

  →Michigan的死亡率較高！

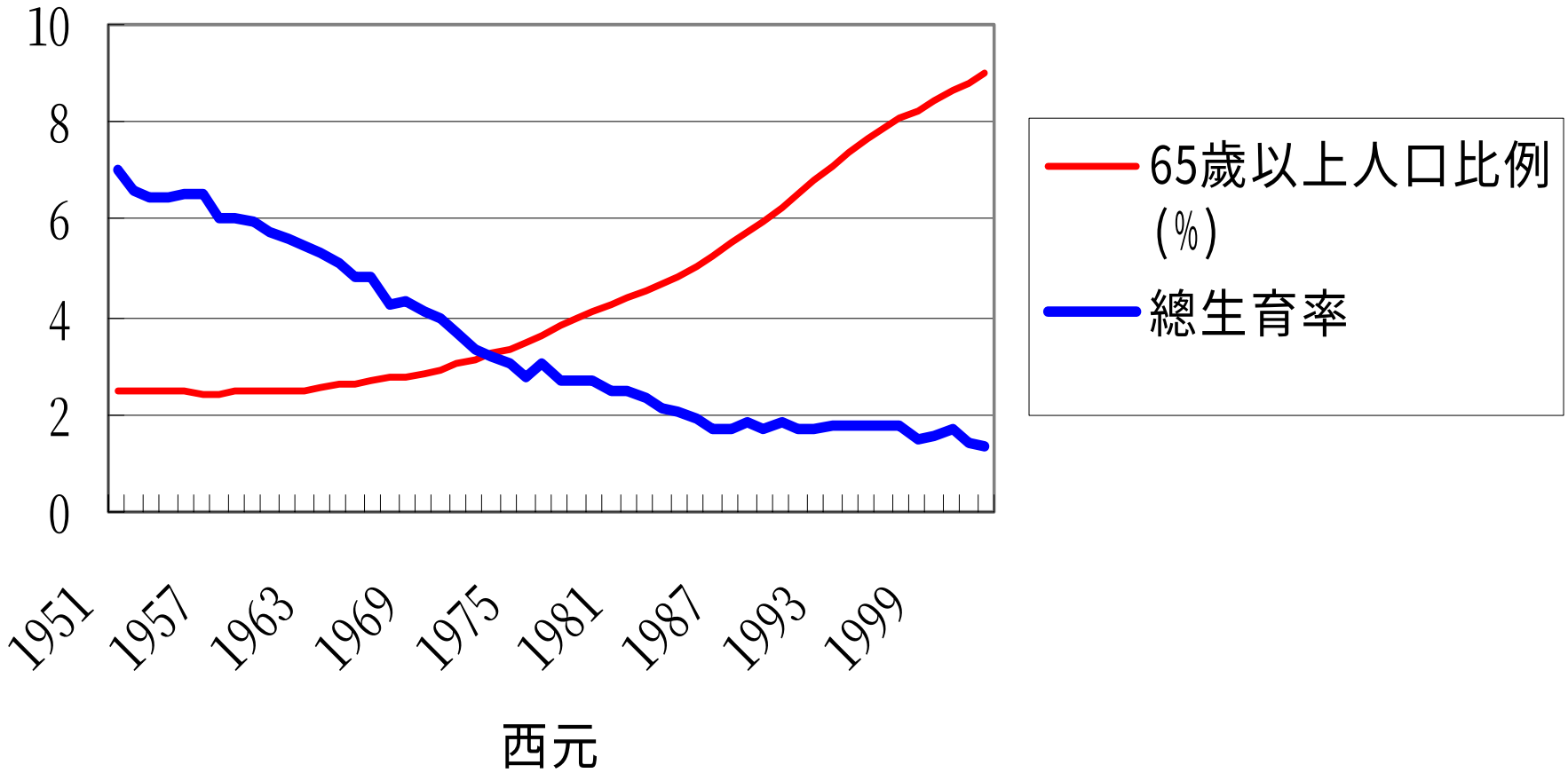
- 若依間接調整法計算，可得

  ADR_I(Michigan) = 9.14670

  ADR_I(Florida) = 8.02307

  →其中SMR(Michigan) = 1.04658

  SMR(Florida) = 0.918017



65歲以上老年人口比例及總生育率趨勢圖

III、方法優劣的評估標準

- 實證分析與理論證明不同，一般很難決定某種方法必然優於另一種方法，模型優劣的比較通常透過實際資料的驗證。然而，若純粹以模型吻合度為優劣標準，通常參數愈多的模型會愈好。一般採取以下方法修正：

→ AIC、BIC等加入參數個數的指標；

→ 交叉驗證(Cross validation)；

→資料分成估計(Training)與驗證(Testing)兩部份。

交叉驗證

- 資料依發生先後拆成兩部份，第一部份資料用於估計模型的參數，第二部份資料當作未知(或是「未來」)資料。以第一部份得出的模型「預測」第二部份的結果，再比較兩者間的差異。

→直觀上，「預測」的誤差愈小，認為模型的表現愈佳。

註：兩部份的資料量比例至少為1:2，以確保分析結果的穩定性。

模型優劣的判斷標準

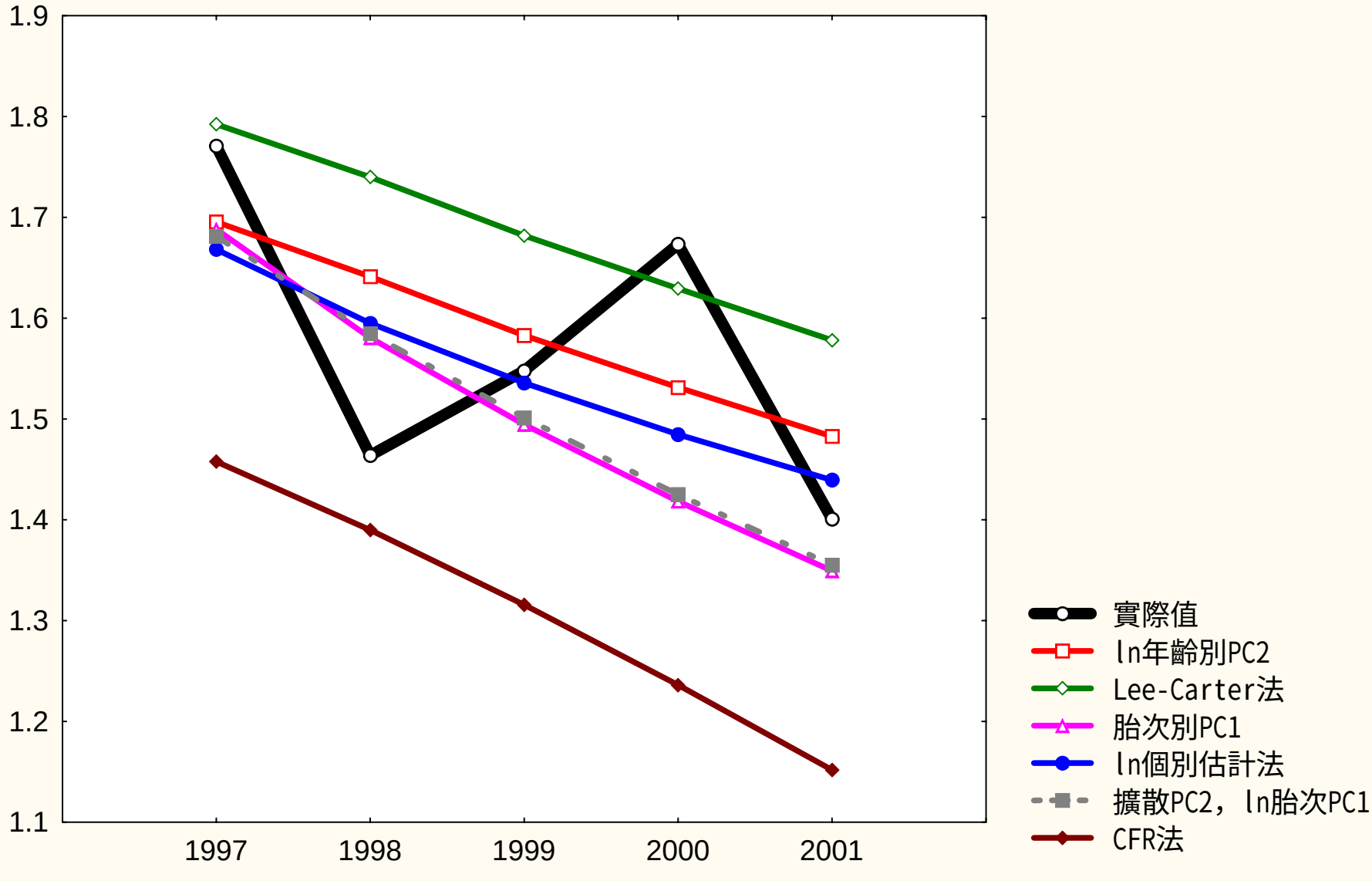
- 平均絕對誤差(Mean Absolute Percentage Error)：

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|\varepsilon_i|}{Y_i} \times 100\%$$

- 根均平方誤差(Root Mean Square Percentage Error，簡稱RMSPE)：

$$RMSPE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\frac{\varepsilon_i}{Y_i})^2} \times 100\%$$

台灣總生育率與六種方法預測之總生育率



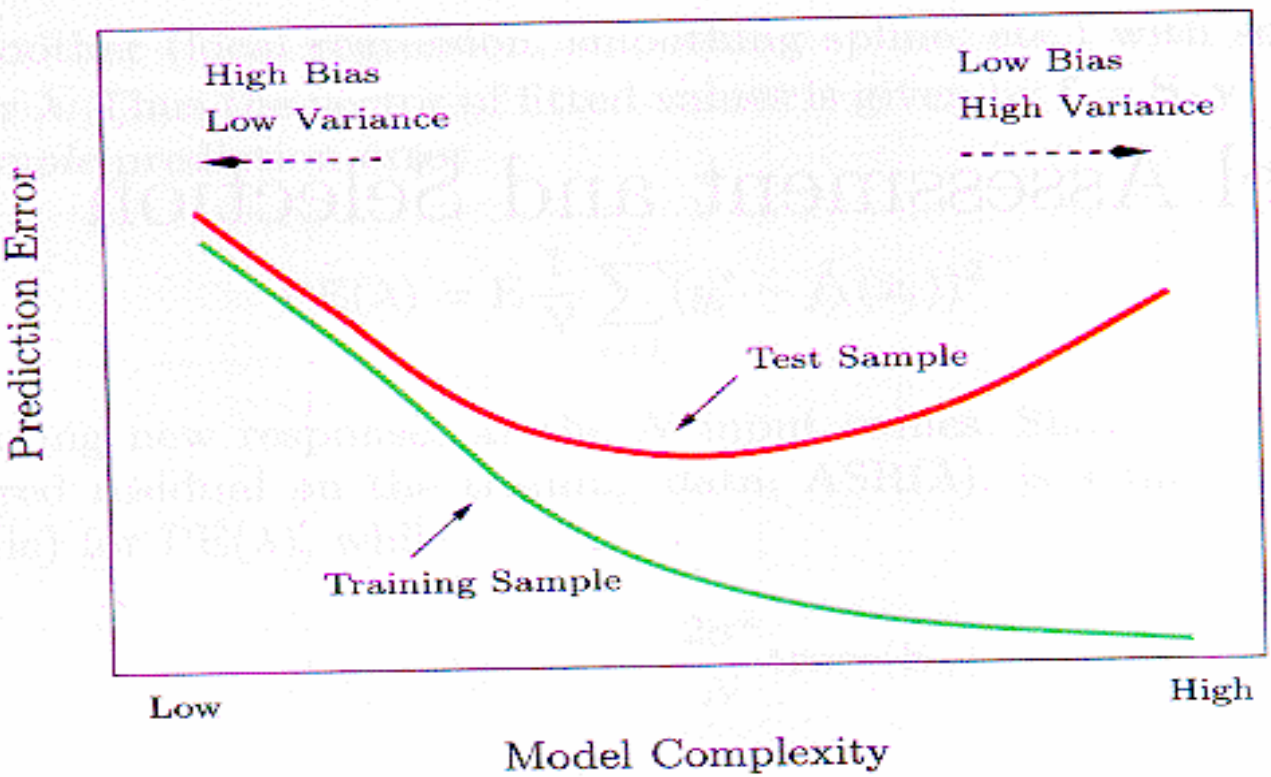
台灣總生育率預測誤差(5個年度加總)

模型	MAPE	RMSPE
ln年齡別PC2	6.60	7.44
Lee-Carter法	8.81	10.96
胎次別PC1	6.99	8.28
ln個別估計法	5.92	7.06
擴散模型PC2，ln胎次別PC1	5.94	7.7
CFR法	16.32	17.67

註：網底者為各模型中最小者

估計(Training)與驗證(Testing)

- 當資料不具先後順序時，資料隨機分成兩部份，仿照交叉驗證比較優劣。



- 當資料量較多時，也可分成三部份：:

50% 25% 25%

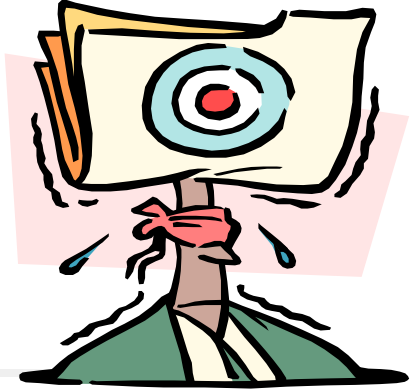
e.g. Bias-Variance Decomposition

Error(f) = Irreducible Error + Bias^2 + Variance
(e.g. Pure error)

IV、結果詮釋與推論限制

- 分析結果若能合理闡述，可達到「畫龍點睛」之效，但最忌諱忽略關鍵點，純粹就數字面來詮釋，反而變成「畫虎不成反類犬」。
- 研究結果的推論也需注意，不能僅從統計結果來看，也需瞭解推論結果代表的意義。

驟下結論(範例)



- 多數車禍發生在車速40　60公里/時，僅有少數在車速超過100公里。

→開快車比較安全？

- 美國亞歷桑那州死於肺結核的比例最高。

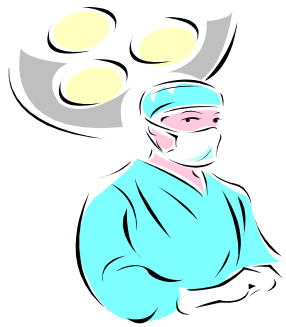
→亞歷桑那州的天氣易於感染肺結核？

- 調查小學生的拼字能力，發現腳愈大的拼字能力也較強。

→腳的大小影響拼字能力？

驟下結論(續)

- 去年大陸調查發現長壽者中，排行老大者最多。



→排行老大較長壽？

→抑或是排行老大者佔了多數？

- 英國公務統計顯示在家裡生產者，發生意外的比例較在醫院生產者高，因此孕婦都應該在醫院生產。

→為什麼有些孕婦會在醫院以外的地方生產？

不合適的迴歸分析範例

- 例題一、一般認為新生兒正常的機會(Y)與某化學藥品的成分(X)為反比關係：

The regression equation is
Y = 100 - 0.00290 X

Predictor	Coef	StDev	T	P
Constant	**100.002**	0.011	9184.91	0.000
X	-0.00290	0.00024	-12.43	0.000

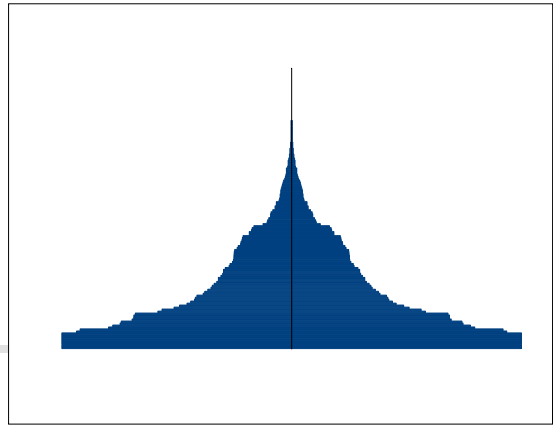
→請問上式中是否有不合理之處？

- 例題二、根據實證分析，總生育率滿足隨時間遞降的迴歸方程式，模型的殘差也無異常之處，以下為迴歸方程式：

The regression equation is
總生育率 = 2.10 – 0.290 (民國– 60)

→在使用本方程式預測未來的生育率時，需注意有何限制？

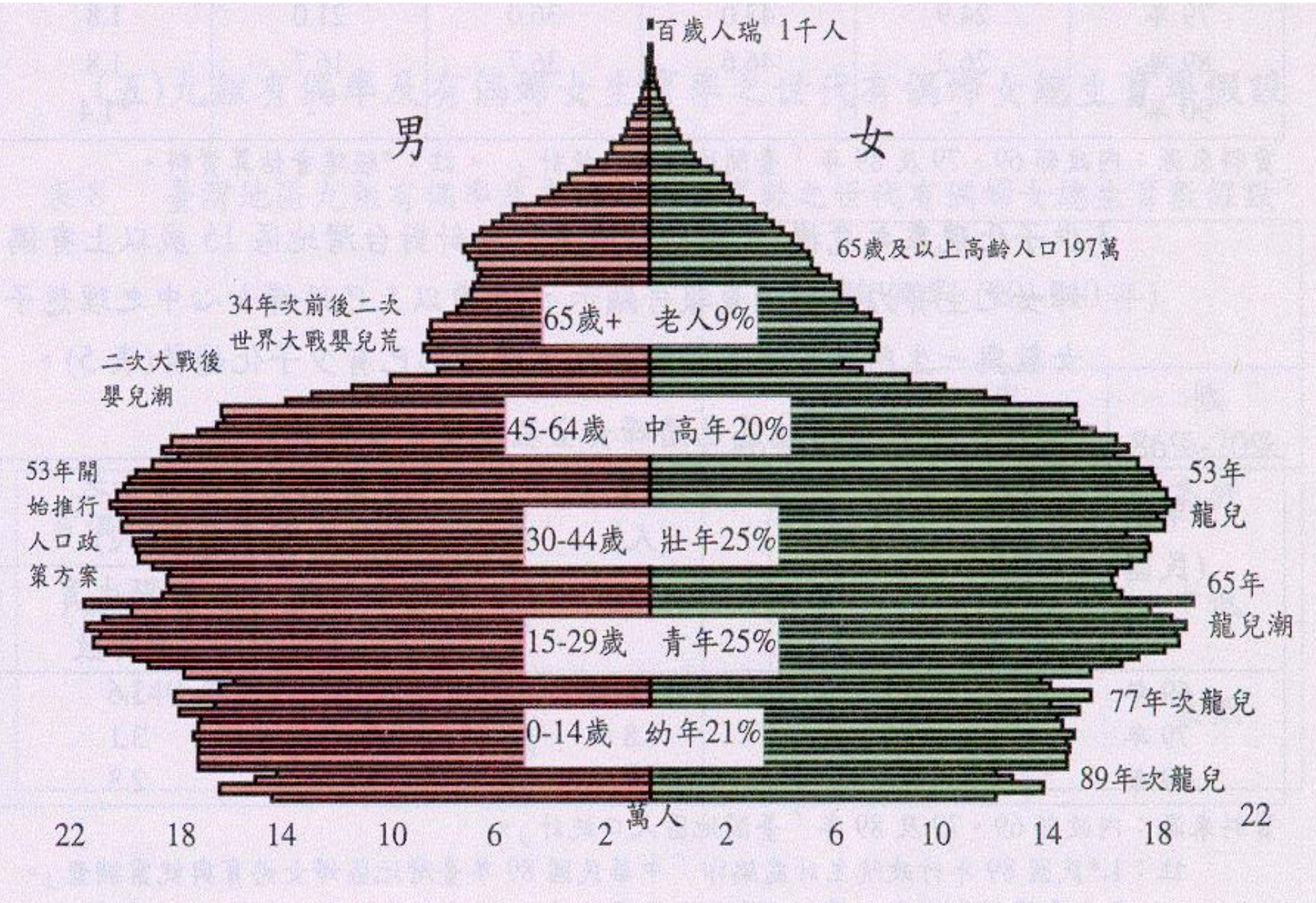
人口金字塔

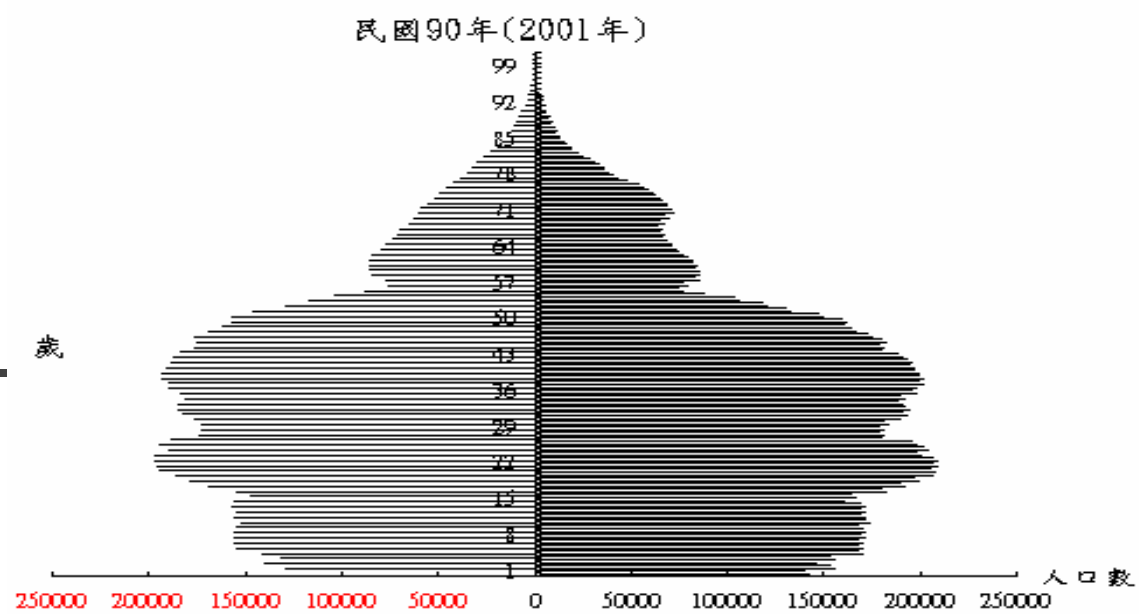


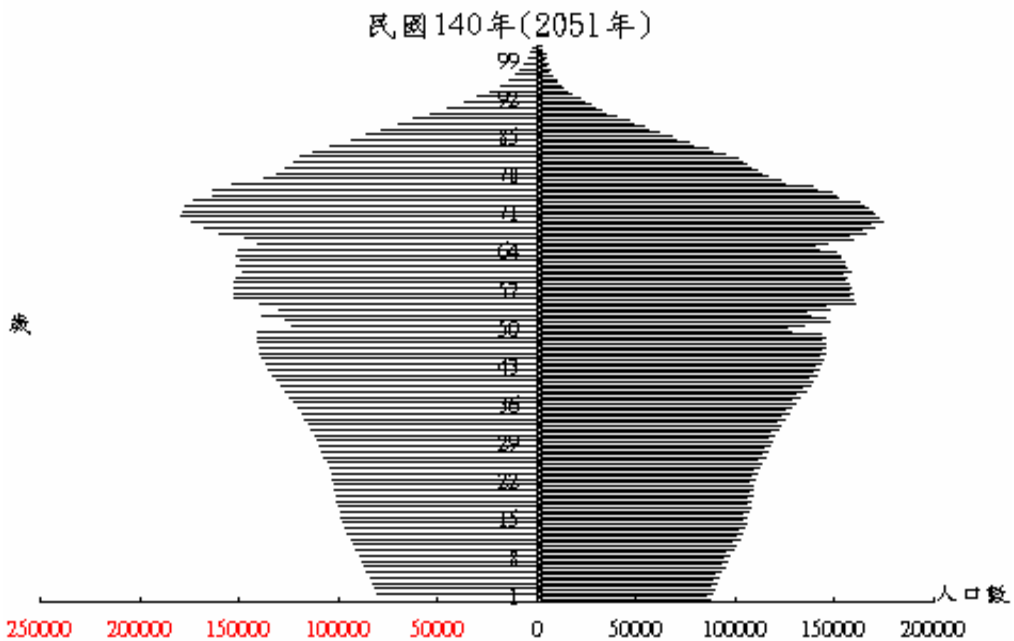
- 人口金字塔(Population Pyramid)：

→將男女兩性各年齡人口以兩側站立的直方圖(Histogram)表示，可看出各年齡的人口比例。

→因繪出的圖形多呈現類似金字塔的三角形，因此稱為人口金字塔。


百歲人瑞 1千人
男
女
65歲及以上高齡人口197萬
34年次前後二次
世界大戰嬰兒荒
二次大戰後
嬰兒潮
65歲+ 老人9%
45-64歲 中高年20%
30-44歲 壯年25%
15-29歲 青年25%
0-14歲 幼年21%
53年開
始推行
人口政
策方案
53年
龍兒
65年
龍兒潮
77年次龍兒
89年次龍兒
22 18 14 10 6 2 萬人 2 6 10 14 18 22





人口老化：
65歲以上
9%→30%

未來五十年台灣的人口結構預測

其他常用的人口統計名詞

- 平均壽命、平均餘命

→平均壽命65歲≠現年55歲的人可再活10年

- 依賴人口比例(0~14, 15~64, 65+)

→幼齡、老年人口扶養比

- 老年人口比例(65+)

→老化指數

- 有偶率

→結婚率、離婚率

- 另外，也有根據男性定義的生育率：

→男性人口一般生育率(General Fertility Rate of Men)

$$\text{男性人口一般生育率} = \frac{\text{一年內出生之活產數}}{\text{15－59歲男性年中人口數}} \times 1{,}000‰$$

→男性人口年齡別生育率(Age-specific Fertility Rate of Men)

$$\text{男性人口年齡別生育率} = \frac{\text{一年內某年齡組出生之活產數}}{\text{該年齡組男性年中人口數}} \times 1{,}000‰$$

→男性人口總生育率(Total Fertility Rate of Men) =男性人口年齡組別生育率的總和乘以五之積。

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/
- 台灣人口學會、台灣大學人口與性別研究中心
  http://ccsun57.cc.ntu.edu.tw/~psc/index.htm