

# 修勻學(Graduation) — Bootstrap(拔靴法；梯雲縱)

授課教師：余清祥教授

課程日期：2023年12月20日

資料下載：

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



# Monte Carlo vs. Bootstrap

■ 蒙地卡羅(Monte Carlo)與拔靴法(Bootstrap)兩者操作方式類似：

→ 蒙地卡羅認為母體分配已知，從這個假設產生亂數，再從這些亂數探討問題，包括期望值、變異數、百分位數等母體性質；

→ 拔靴法「假設」樣本就是母體，從這個有限母體中產生亂數，再從這些亂數計算樣本統計量的「變異數」！

註：拔靴法的功能較為侷限。

## ■ 蒙地卡羅的範例：Wilcoxon符號檢定

→ 檢查一組樣本的中位數是否等於某個定值，首先計算所有觀察值的秩(rank)，再乘以大於或小於中位數的符號，分別加總正負兩個秩( $T^+$ 、 $T^-$ )的總和，取較小者為統計檢定量。

→ 理論上， $T^+ + T^- = \frac{n(n+1)}{2}$ 。

→ 當樣本數較大時，

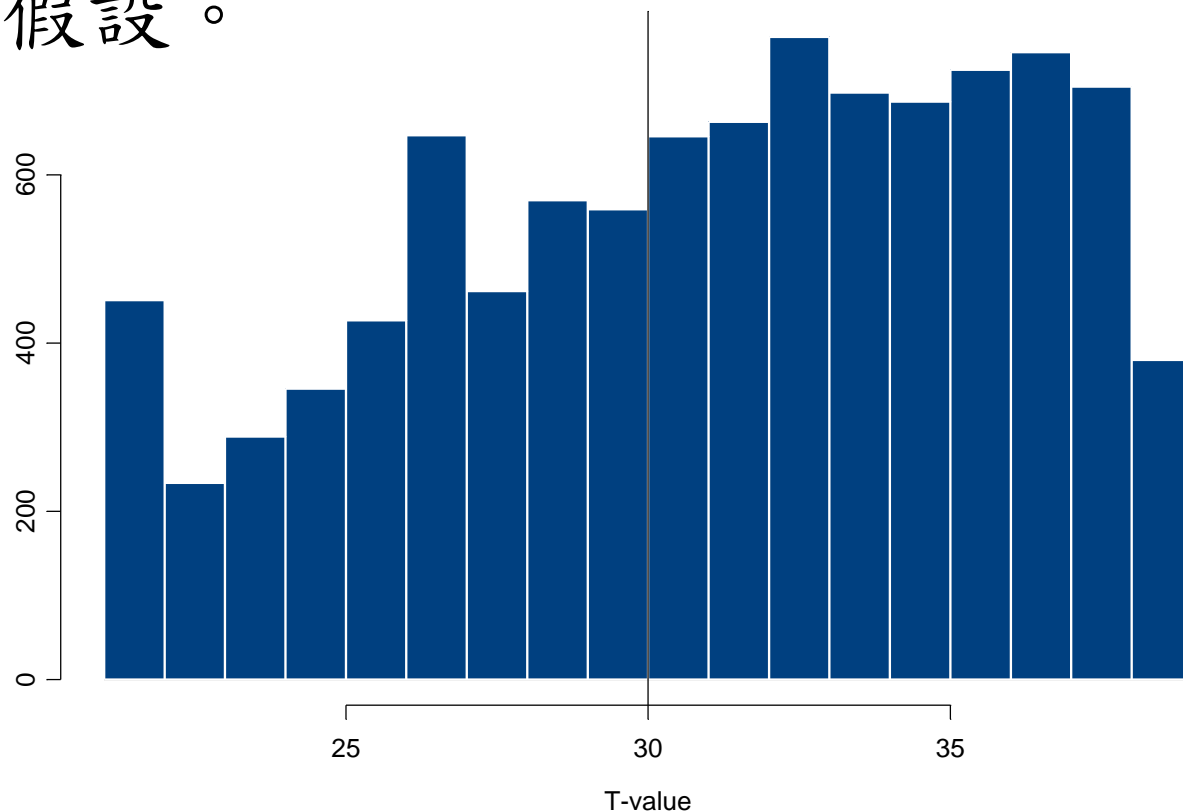
$$E(T) = \frac{n(n+1)}{4} \quad \& \quad \text{Var}(T) = \frac{n(n+1)(2n+1)}{24}.$$

註：假設  $X_i \sim i.i.d. N(m, \sigma^2), i = 1, 2, \dots, n$ .

Critical values of the Mann-Whitney-Wilcoxon test ( $\alpha=.05$ , 10,000 runs, and numbers in red are the true critical values.) → 差異不大!

	$n_1=2$	3	4	5	6	7	8	9	10
$n_2=2$	3	3	3	3	3	3	4	4	4
3	6	6	6	7	8	8	9	9	10
4	10	10	11	12	13	14	14(15)	16(15)	16
5	15	16	17	18	19	21	22	23	24
6	21	23	24	25	27	28	30	31(32)	33
7	28	30	32	34	35	37	39	41	43
8	37	39	41	43	45	47	49(50)	52	54
9	46	48	51(50)	53	56	58	61	63	66
10	56	58(59)	61	64	67	70	73	76	79

→ 以  $X_i \sim i.i.d. N(0,1), i = 1, 2, \dots, 12$  為例，假設某組觀察值的檢定量為30，重複9,999次的模擬，計算出Monte-Carlo p-value等於  $(1+3985)/(1+9999) = 0.3986$ ，不拒絕中位數為0的假設。

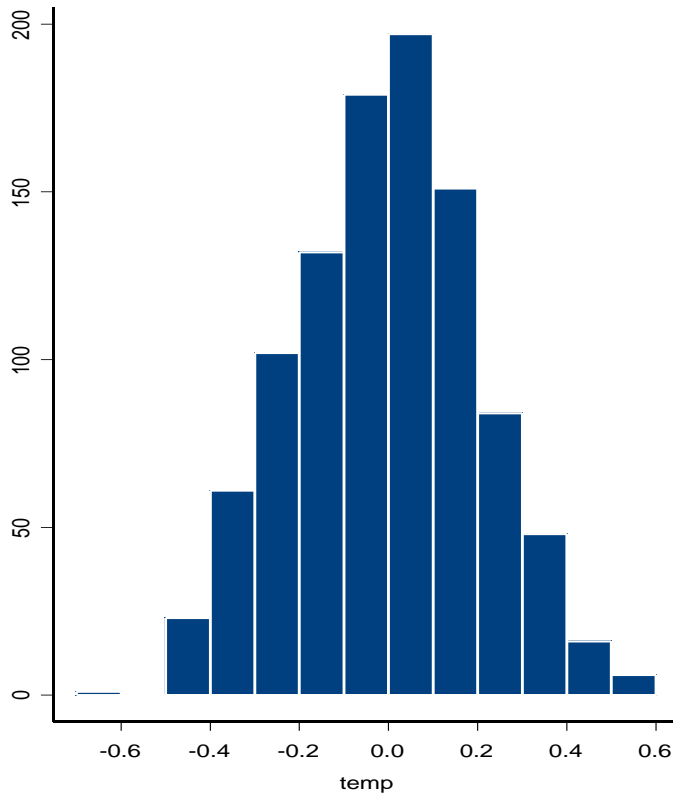


# Monte Carlo p-value & Test

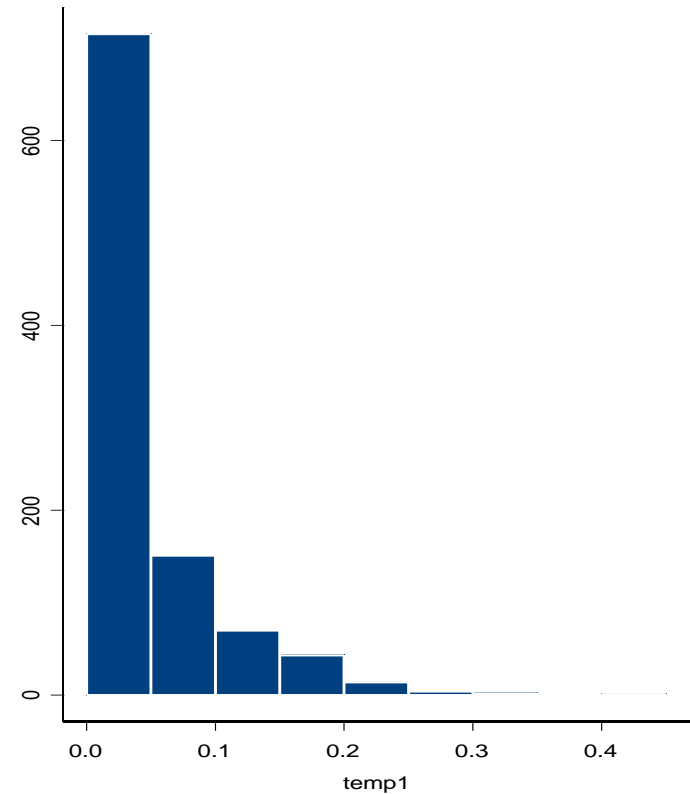
- 根據虛無假設模擬出  $n$  組樣本，如果觀察值排在第  $k$  個(也就是第  $k$  大)，則檢定的  $p\text{-value} = k/(n+1)$ 。一般顯著水準為 0.05， $n$  值多半選為 99、499、999、9,999 或 99,999 等數字。
- Q：蒙地卡羅 p-value 與排列檢定的差別？
- 蒙地卡羅檢定的作法類似，再給定模型假設下，求出統計量等相關數值的分配。
- 應用範例：迴歸分析中有人用調整過的  $R^2$ ，以消除因隨機而造成的線性相關。

- 隨機由  $N(0,1)$  產生互相獨立的25個  $X$ 、 $Y$  觀察值，根據迴歸方程式  $Y = \alpha + \beta X$  計算  $R^2$ ，重複1000次的模擬可得

Correlation between X and Y



$R^2$  of Y on X



→ 平均  $R^2 = 4.2\%$ ， $Q3 = 5.9\%$ 。

# Bootstrap(拔靴法)

- Bootstrap法可追塑至Efron於1979提出的方法，屬於重複抽樣(Resampling)方法。將已有的觀察值當作是母體重複抽樣(與Monte Carlo有真實母體不同！！)，以求取原先因資料不足而無法探討的資料特性，早期探討的特性以變異數為主。
- 舉例而言，假設  $x_1, x_2, \dots, x_N$  為來自同一分配的觀察值，而我們想瞭解這個分配的中位數與其中位數的變異數。

範例：一組由Poisson分配抽出的隨機樣本

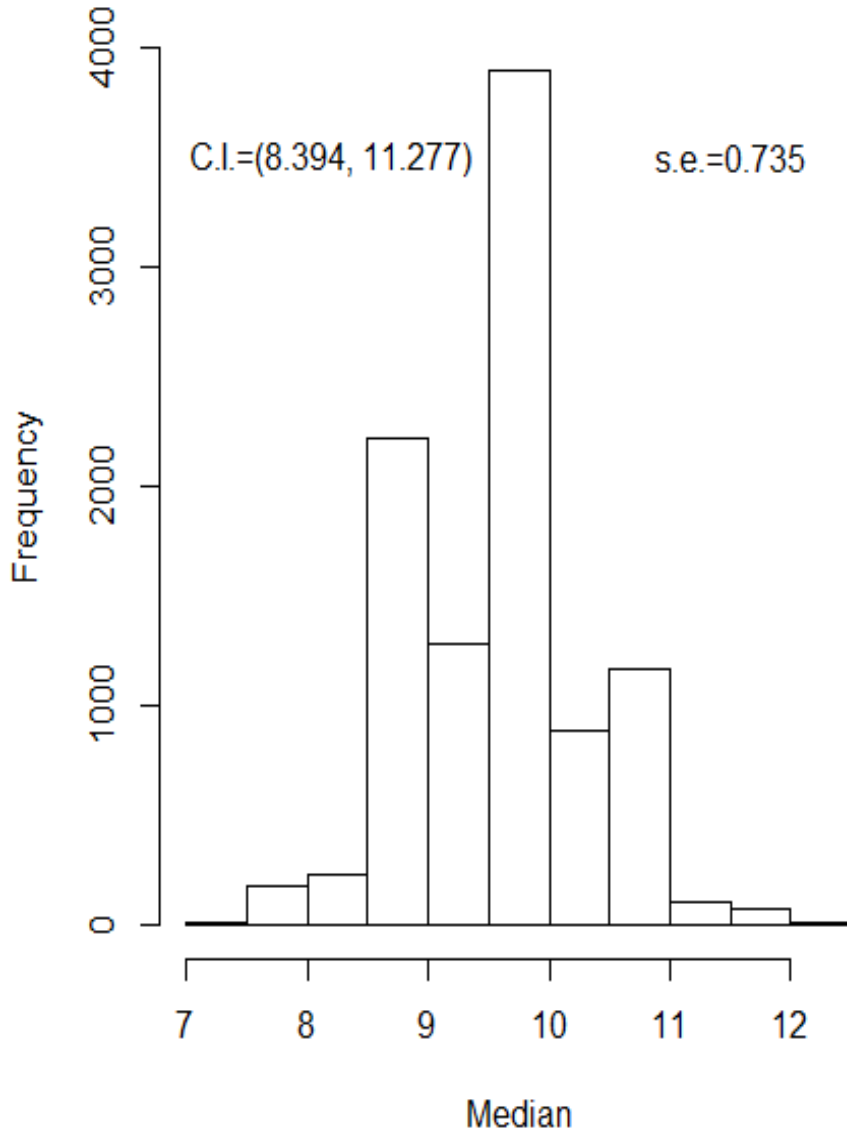
6 7 7 7 7 8 8 8 9 9 9 9 10 10 10 10  
11 11 12 13 13 13 13 14 15 15 15 15 17 20

→ 已知樣本中位數為10，欲求出母體中位數的信賴區間，可由Bootstrap模擬出標準差。10,000次模擬得出標準差估計值1.0063，中位數的95%信賴區間 (8.5586, 11.4414)

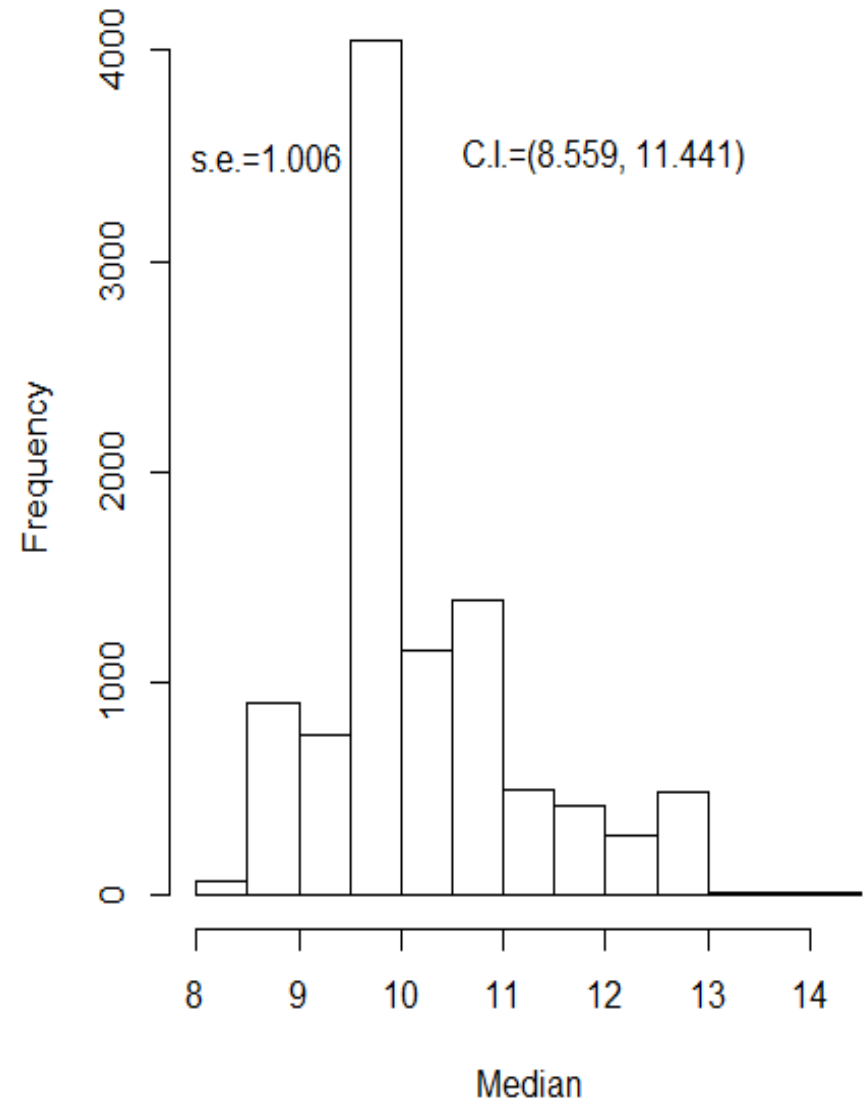
涵蓋了實際中位數10。

註：母體為Poisson(10)；蒙地卡羅95%信賴區間為(8.3939, 11.2767)。

Monte Carlo (10,000 runs)



Bootstrap (10,000 runs)

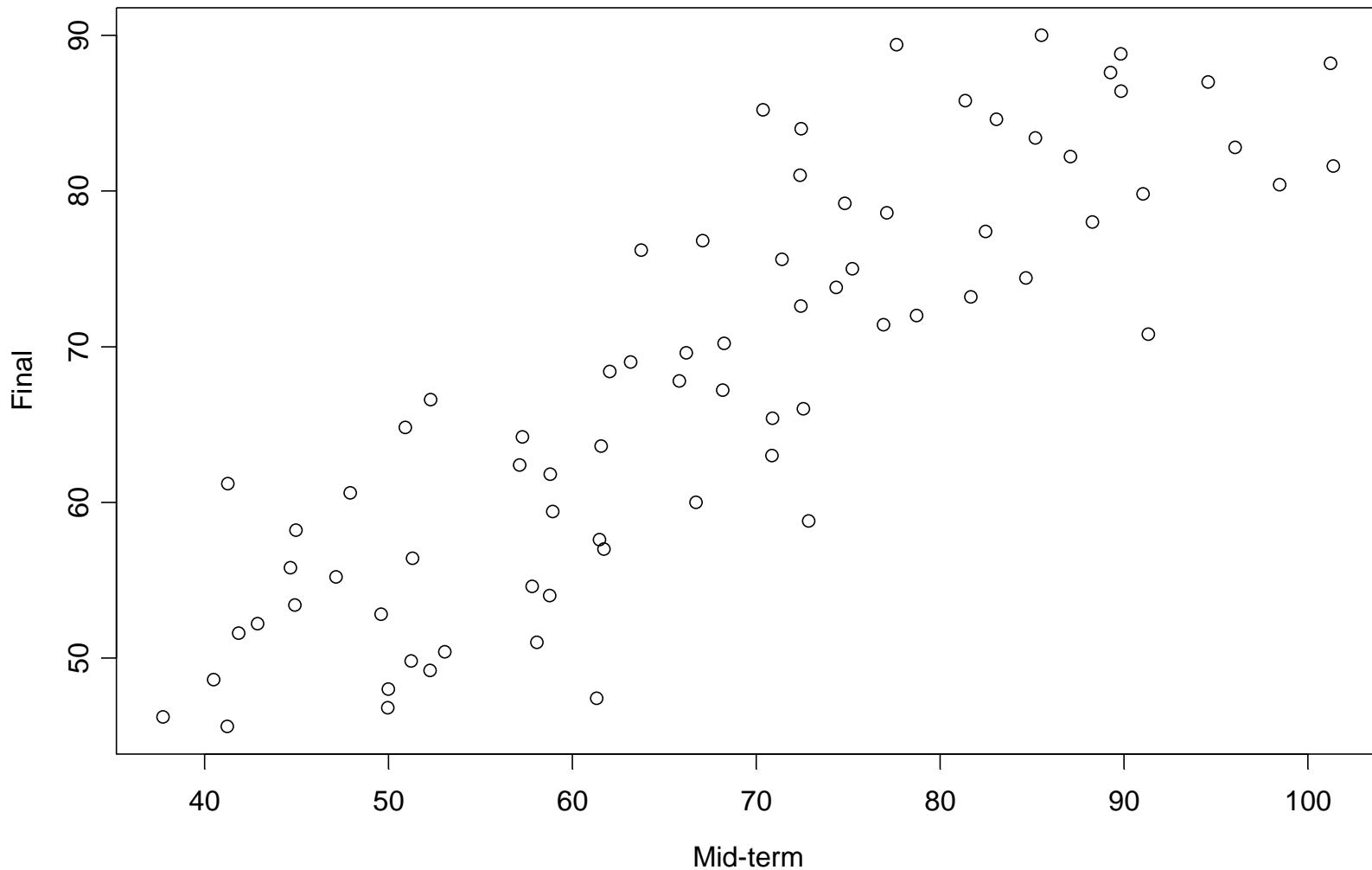


兩者的差異並不大！(標準誤0.7354 vs. 1.0063)

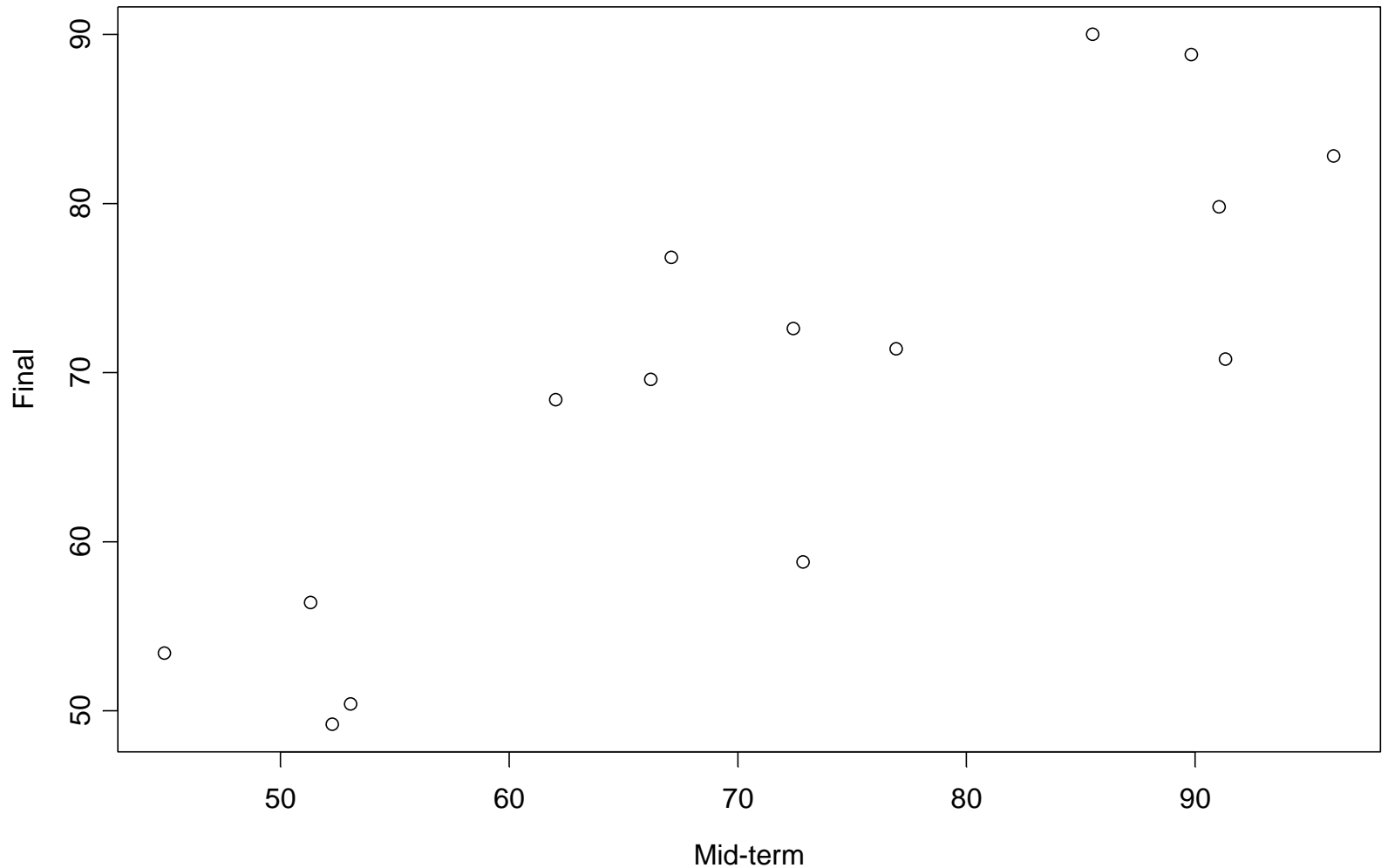
# Monte Carlo及Bootstrap的模擬程式

```
# Monte Carlo
t1=NULL
for (i in 1:10000) { x1=rpois(30,10); y1=median(x1); t1=c(t1,y1) }
# Bootstrap
x0=rpois(30,10)
t2=NULL
for (i in 1:10000) { x2=sample(x0,30,T); y2=median(x2); t2=c(t2,y2) }
#
a1=mean(t1)
b1=sqrt(var(t1))
b2=sqrt(var(t2))
par(mfrow=c(1,2))
hist(t1,xlab="Median",main="Monte Carlo (10,000 runs)")
text(8.2,3500,c("C.I.=(8.394, 11.277)"))
text(11.5,3500,c("s.e.=0.735"))
hist(t2,xlab="Median",main="Bootstrap (10,000 runs)")
text(8.7,3500,c("s.e.=1.006"))
text(12,3500,c("C.I.=(8.559, 11.441)"))
```

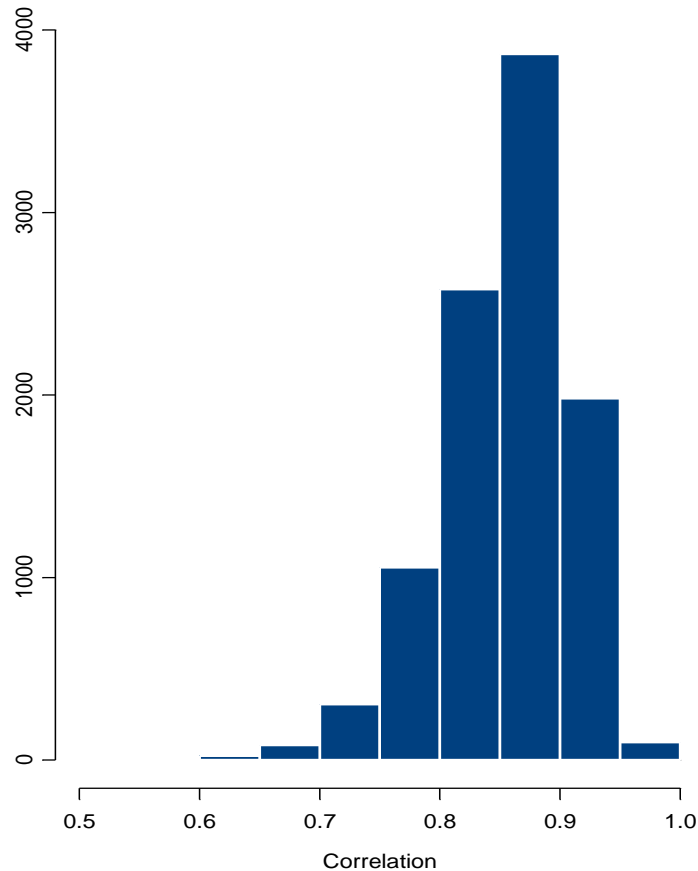
範例：75位選修統計學的學生，想瞭解期中考與總成績的相關性，只抽出15位學生。



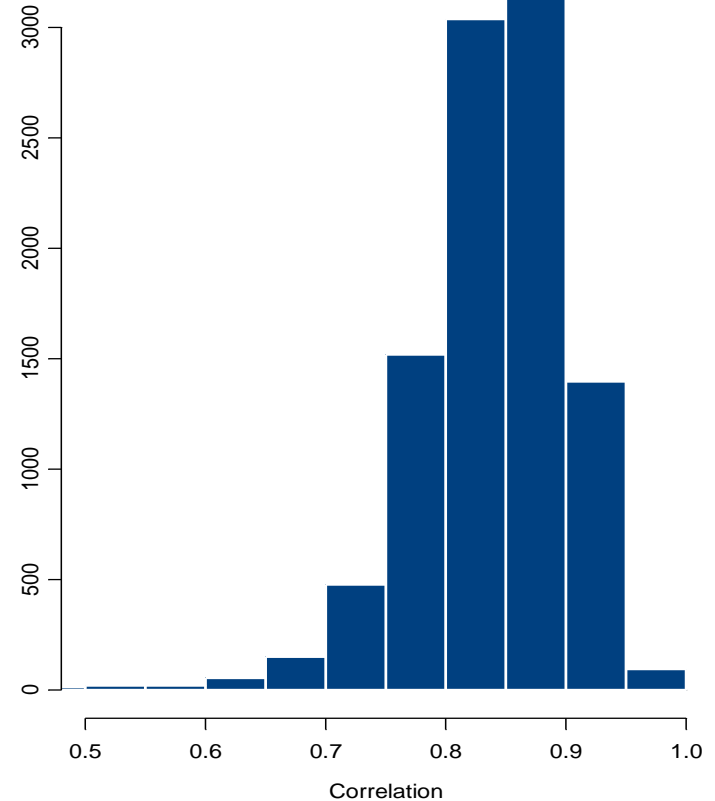
抽出的15位學生與母體特性大致接近，相關係數分別是0.8399及0.8543。



Monte Carlo



Bootstrap



比較各一萬次模擬，Monte Carlo法由母體任意抽出15個樣本得出的相關係數，與15個樣本以Bootstrap法算出的相關係數。

→ 兩者非常接近！(標準差0.0540 vs. 0.0654)



■ 註：

(1) 理論上，我們認為

Bootstrap error

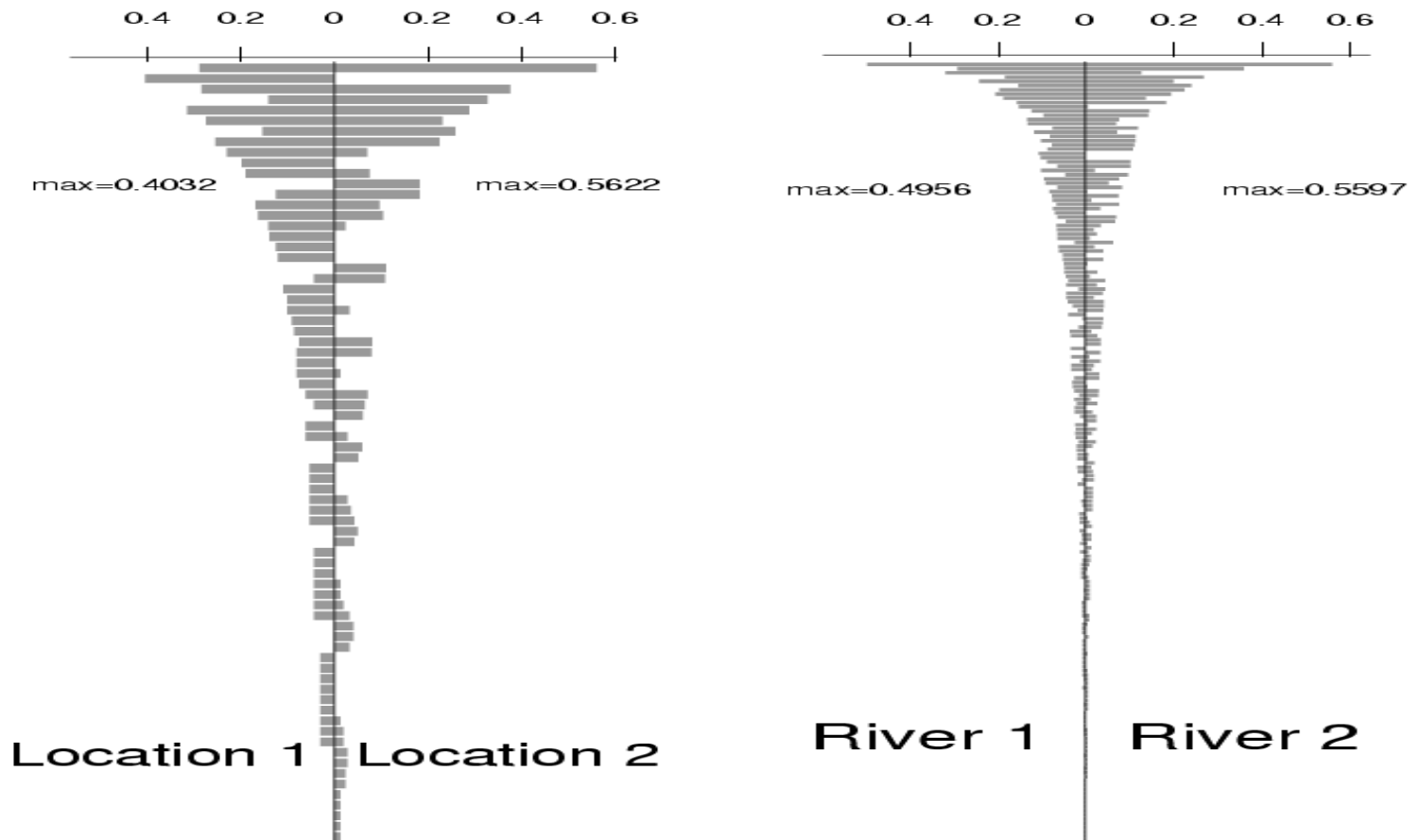
= Sampling error + bootstrap simulation error

(2) 當有充足的樣本數、且樣本具有與母體類似的特性時，Bootstrap可用來近似分配的形狀。(Shift methods!)

(3) 請參考Bootstrap講義！

# Bootstrap法估計變異數的實例

- 檢定兩個數值是否相等。



- Bootstrap法計算出的變異數(標準差)，與大樣本理論Delta 法的數值比較：

	<b>Bootstrap</b>	<b>Delta</b>
螃蟹資料	0.015	0.01426
水鳥資料	0.008	0.0083

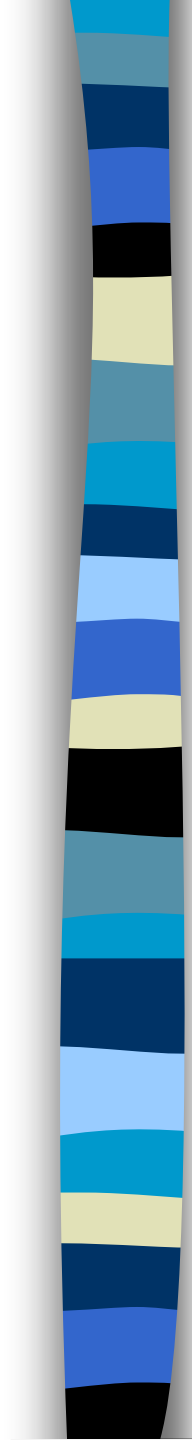


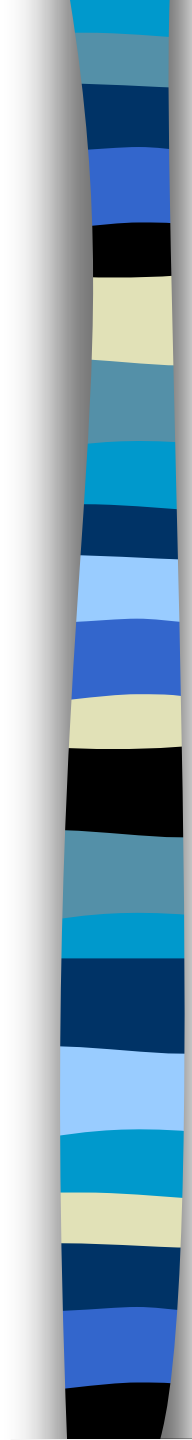
# Bootstrap用於檢定與預測

- 除了用於估計變異數外，也可用於檢定與預測。
  - 檢定可與變異數估計結合，在此將略過細節，僅以一個範例示範想法。
  - 預測多半與相關資料(Dependent data)有關，都屬於無母數方法，但至今仍無統一的方法。常見的方法有Block, Sieve, Local, Wild與Markov Bootstrap，以及Subsampling，在此只介紹Block Bootstrap。

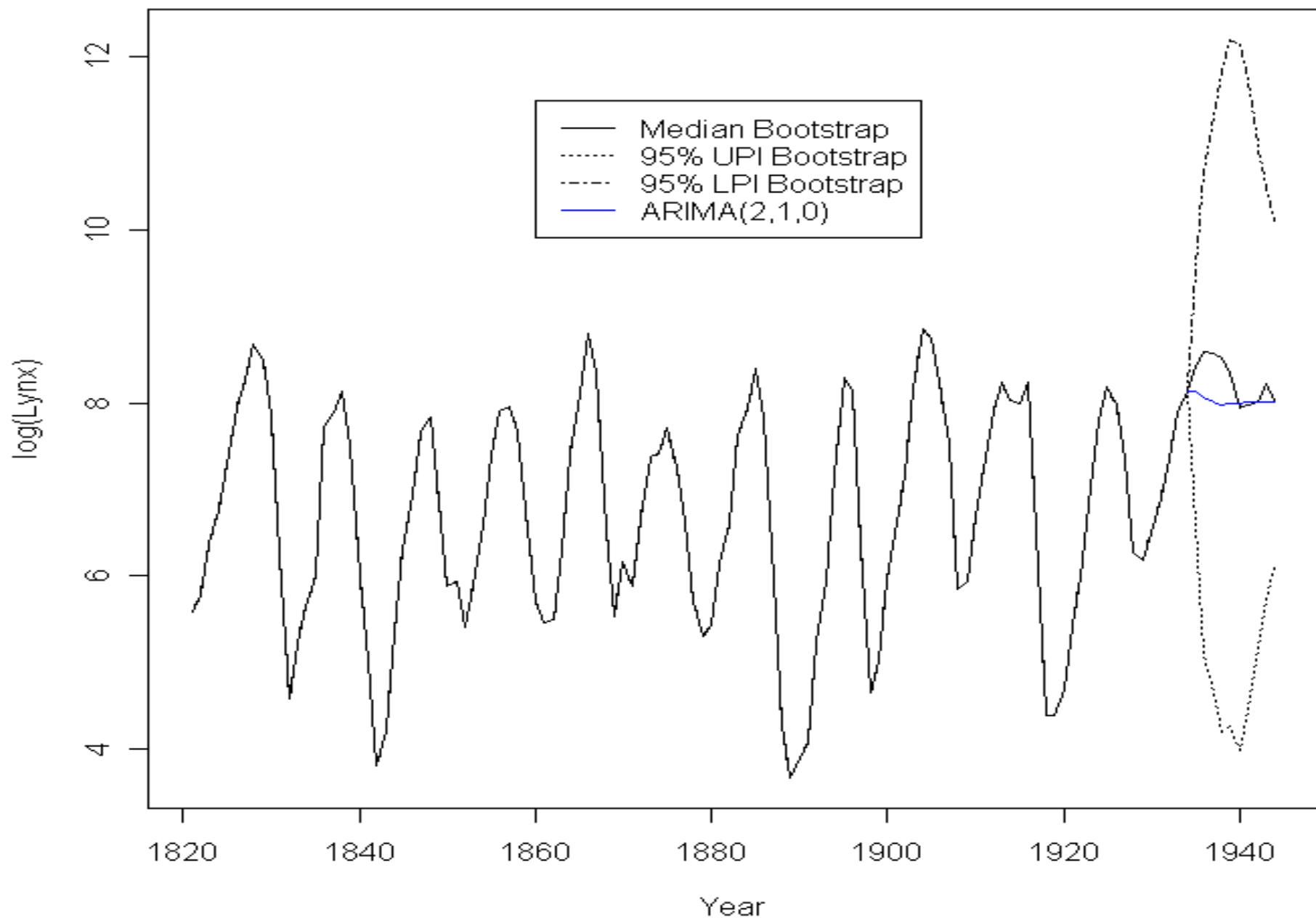
# 區塊(Block)拔靴法

- 區塊拔靴法的重複抽樣方式類似一般的拔靴法，只是每次抽取一個「區塊」的資料，當資料服從均衡(Stationary)假設時，區塊拔靴法大多都適用。
- 區塊拔靴法用於相關資料有不錯的效果，雖然不如獨立樣本時一般拔靴法的準確，但比Subsampling效果好，而且不需要資料滿足很強的條件，加上操作時不需對資料給予任何假設，實證上是很好的選擇。

- 
- 區塊拔靴法的抽樣方式類似拔靴法，但不是對觀察值直接抽樣，而是對相鄰觀察值的差異抽樣，而且抽取時將連續一串的差異值抽出。例如：若區塊長度為 $b$ 、且抽到第 $k$ 個觀察值，則第 $k$ 個至第 $k + b - 1$ 個差異值被抽出，最後一個觀察值加上這些差異值即為預測值。
  - R的模組「boot」也有處理時間數列資料(相關資料的一種)的功能，細節可查閱「tsboot」指令。(這個指令可指定區塊長度為定值、或是服從幾何分配。)

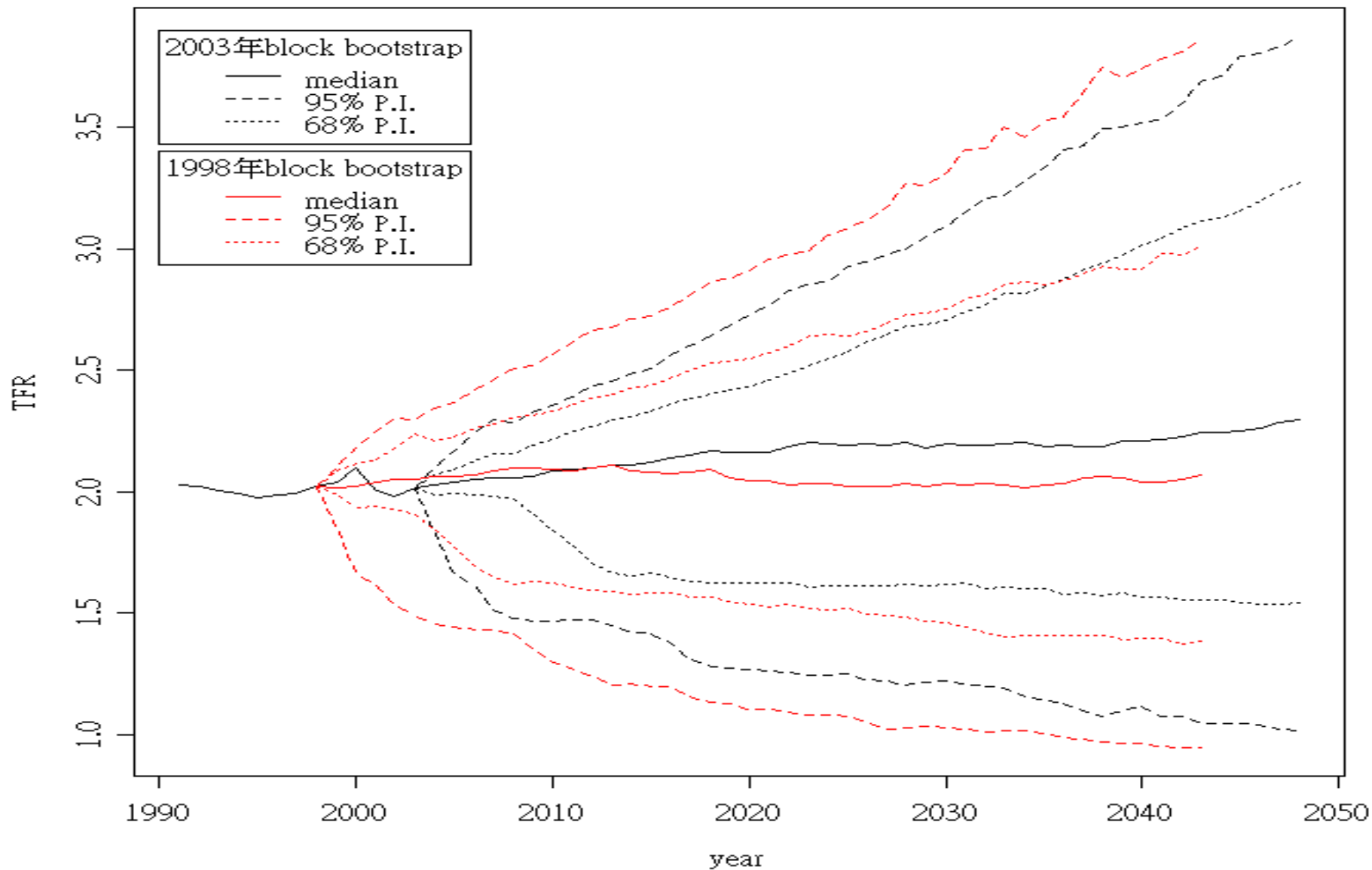
- 
- 範例：R的說明檔中使用「lynx」資料，一共有114筆資料。使用tsboot指令，可限定固定區塊長度，也可指定區塊長度服從幾何分配。（兩者得出的結果非常接近。1821~1924 vs. 1925~1934,  $b=10$ ）
  - 首先對「lynx」資料取對數，再計算相鄰時間觀察值的差值，之後抽出一整個區塊差值，加到最後一期的觀察值。下圖為依據一千次區塊拔靴法的模擬，計算而得的95%預測區間，與差分後的AR(2)模型之預測值之比較。

## Block Bootstrap in Lynx Data



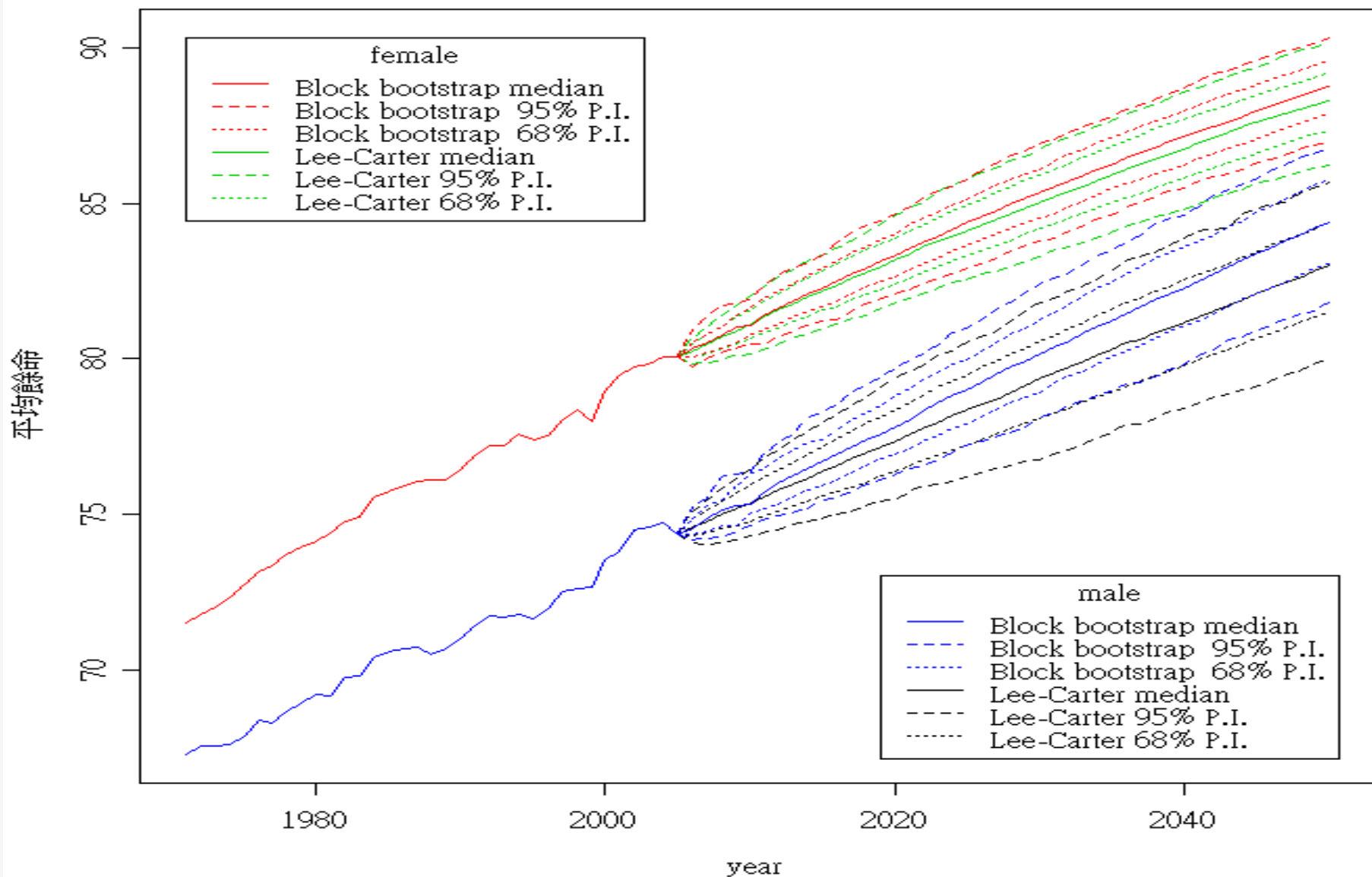
# 區塊拔靴法範例：美國總生育率預測

USA TFR

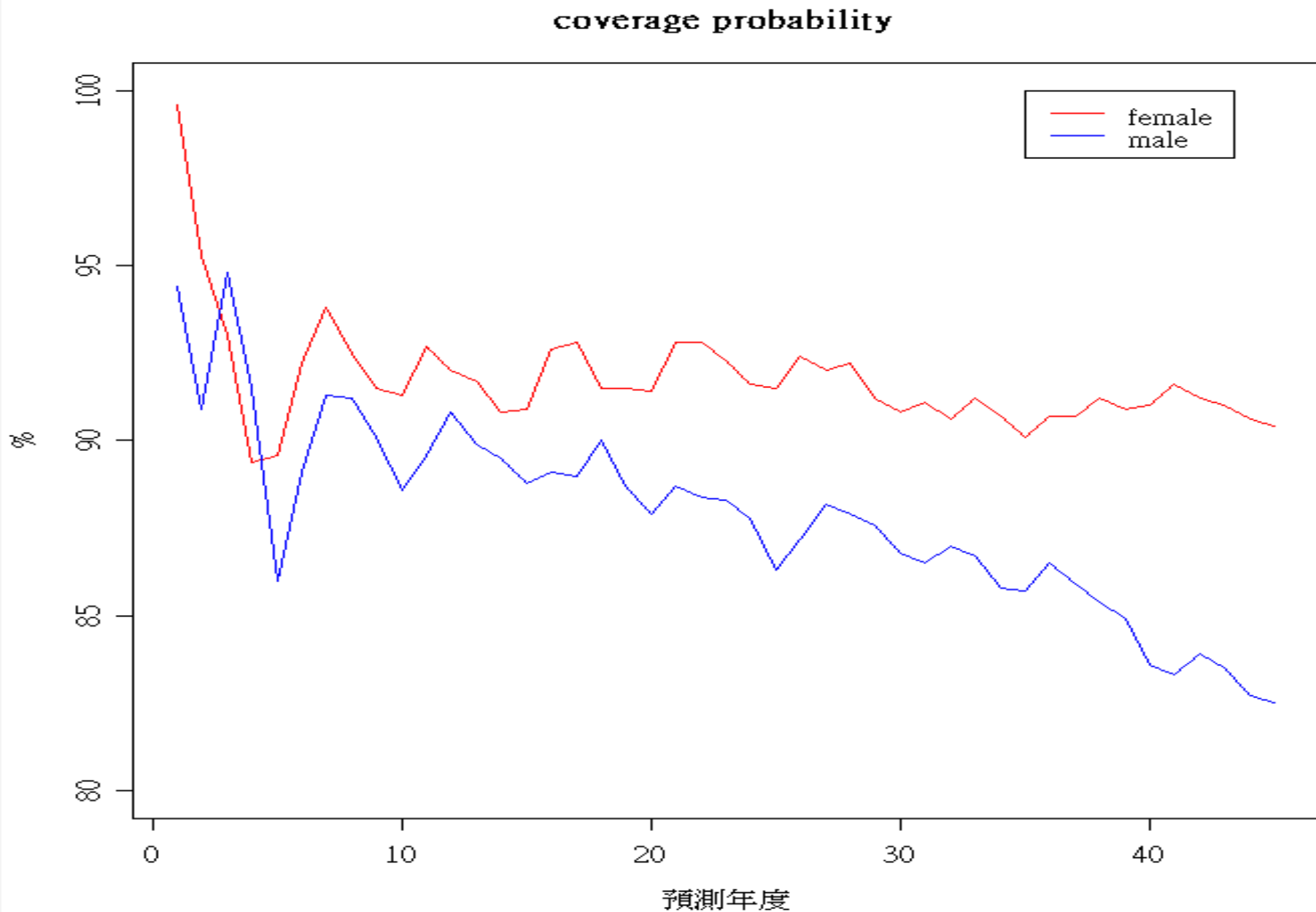


# 區塊拔靴法範例：臺灣死亡率模型預測

## TAIWAN 零歲平均餘命

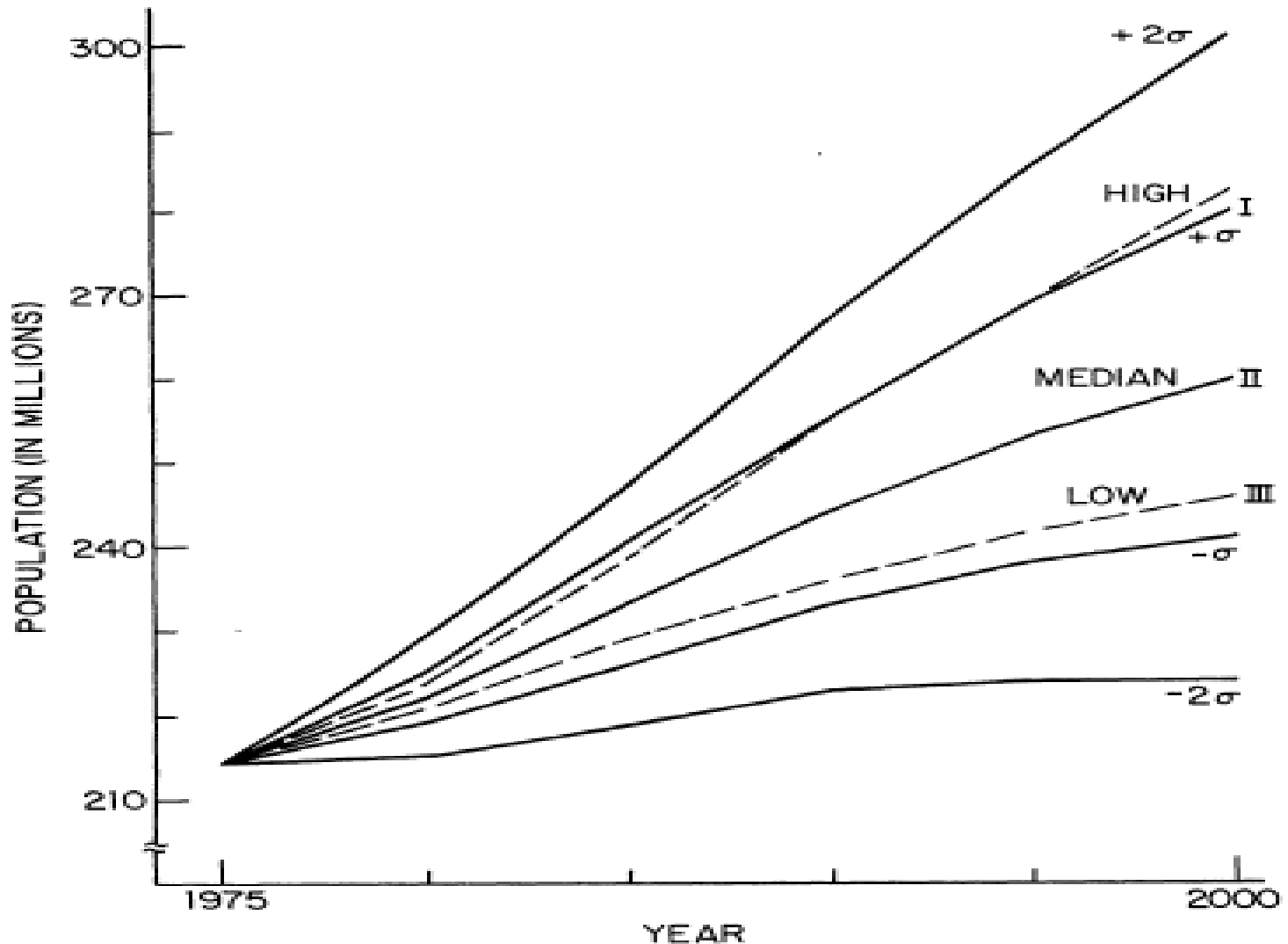


# 區塊拔靴法的涵蓋率



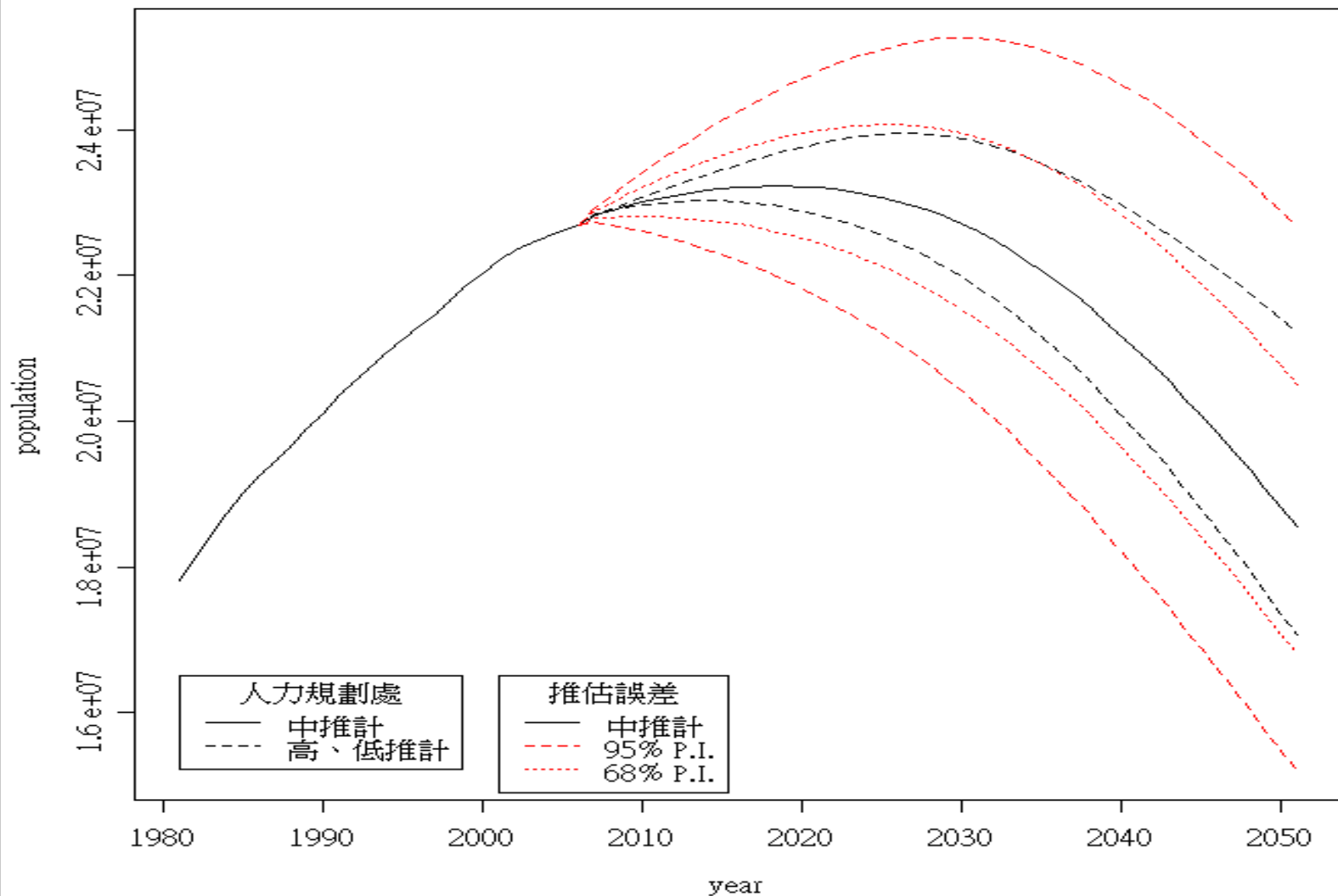
# Stoto (1983): 美國、聯合國

→ 68% 預測區間的上、下限，與美國 (Census Bureau, 1977) 的高、低推計非常接近。



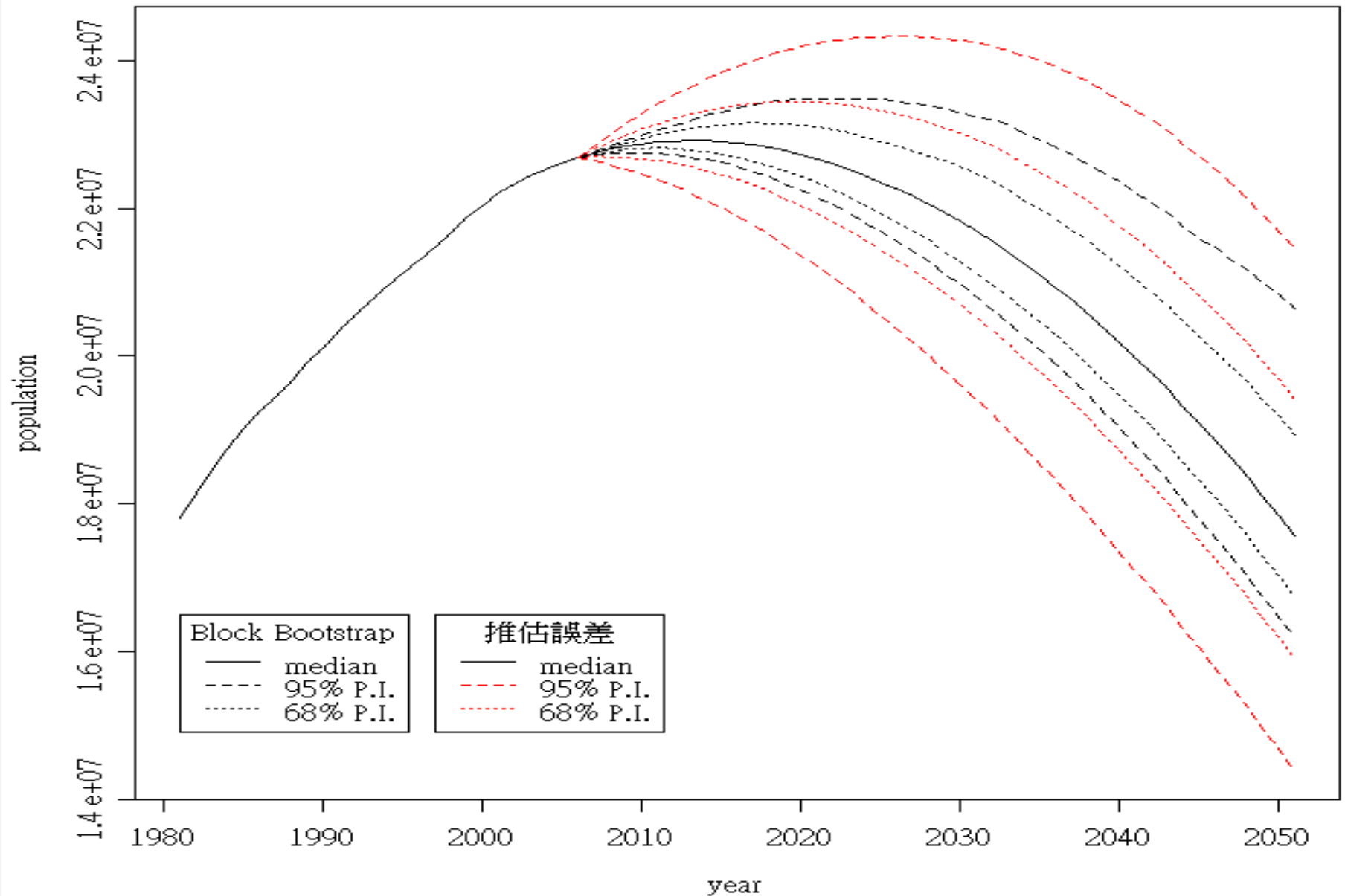
# 台灣推估誤差 (人力規劃處)

TAIWAN total population: 推估誤差 v.s. 人力規劃處

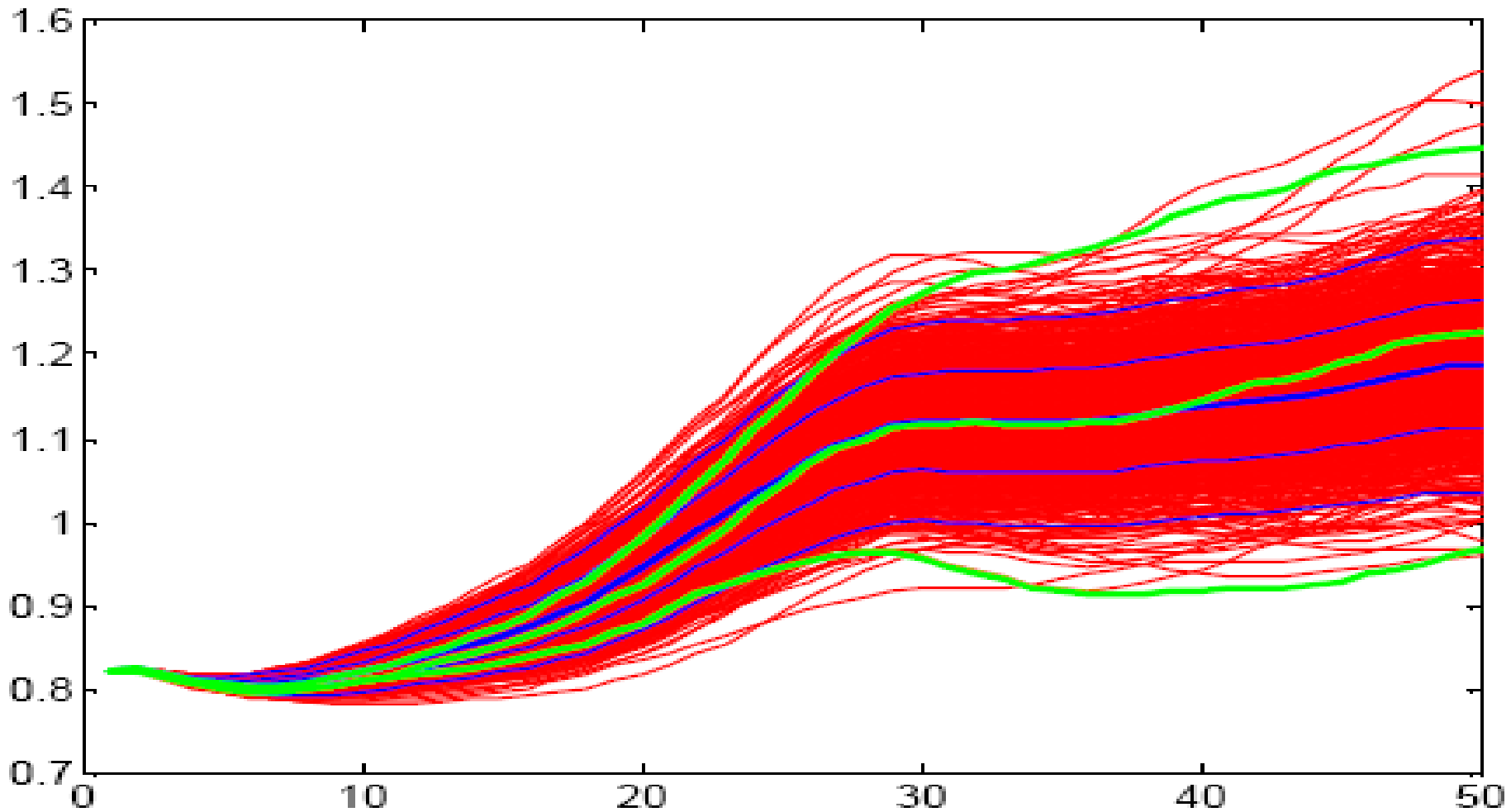


# 台灣推估誤差 (區塊拔靴法)

TAIWAN total population: 推估誤差 v.s. Block Bootstrap



# 隨機方法推估之呈現 (電腦模擬)



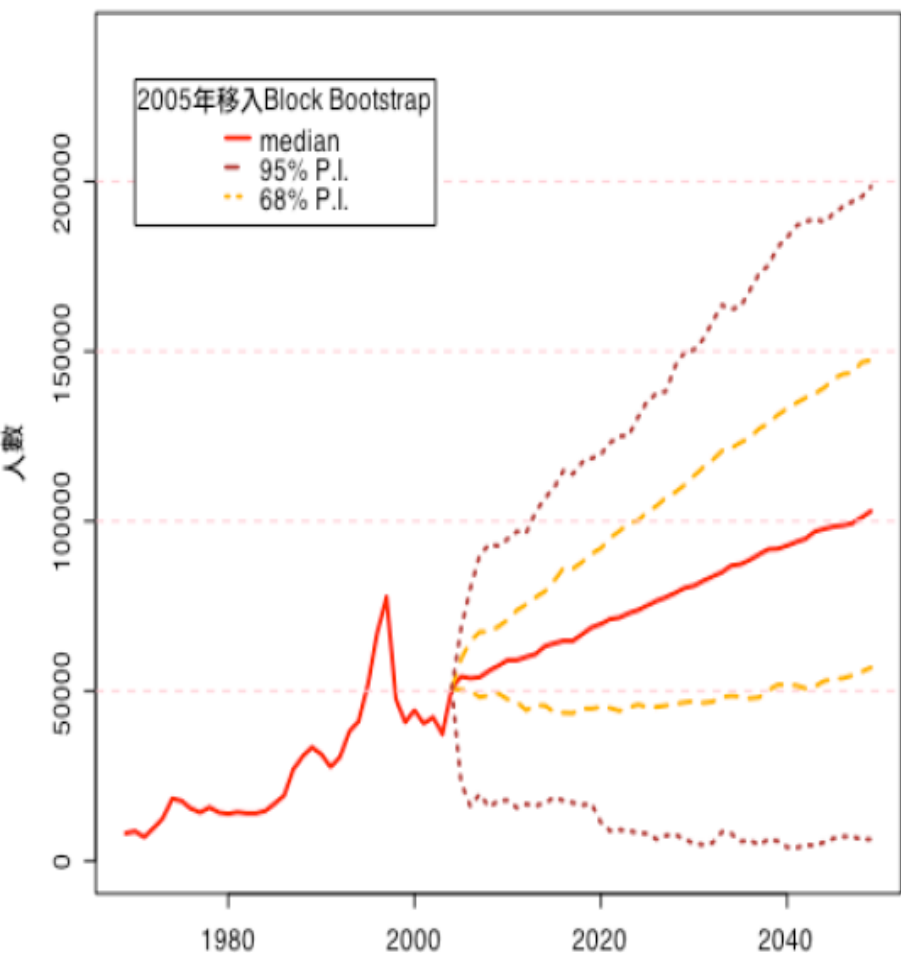
Red: 德國 2002-2050年依賴人口比之模擬路徑

Green: 加總後中位數、最小值及最大值對應之路徑

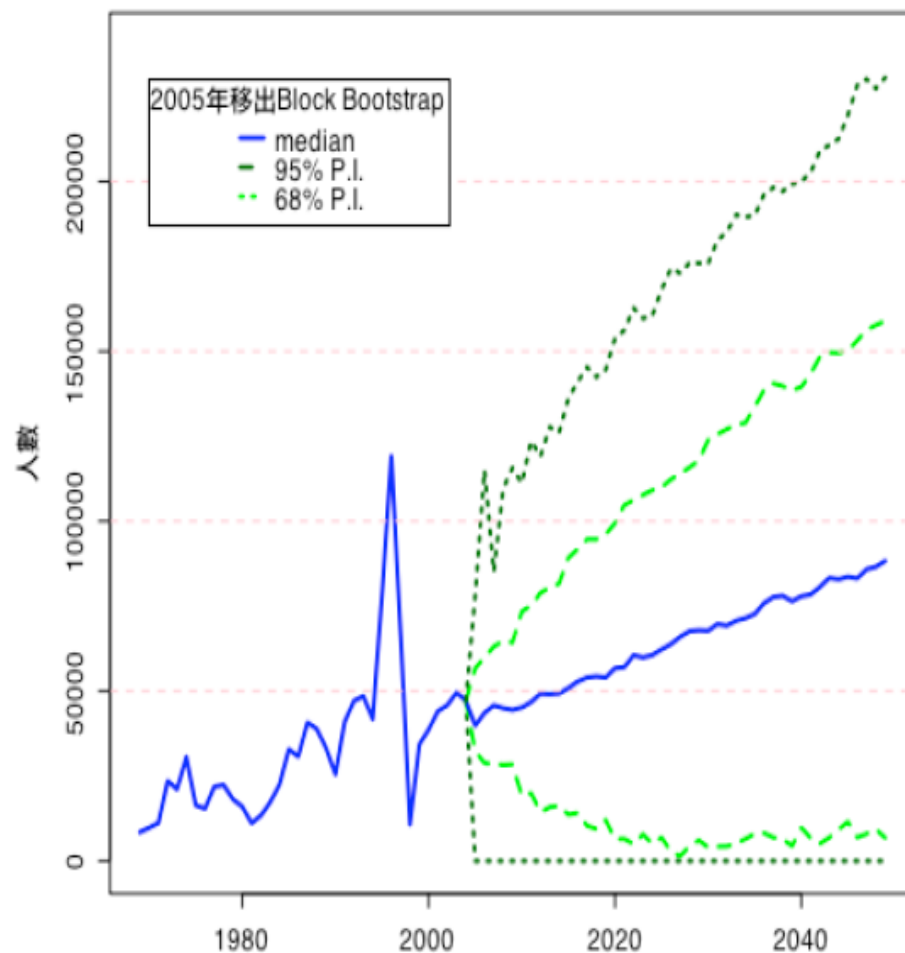
Blue: 平均值、68%預測區間及95%預測區間

# 隨機方法推估之呈現 (區塊拔靴法)

遷入台灣人口



遷出台灣人口

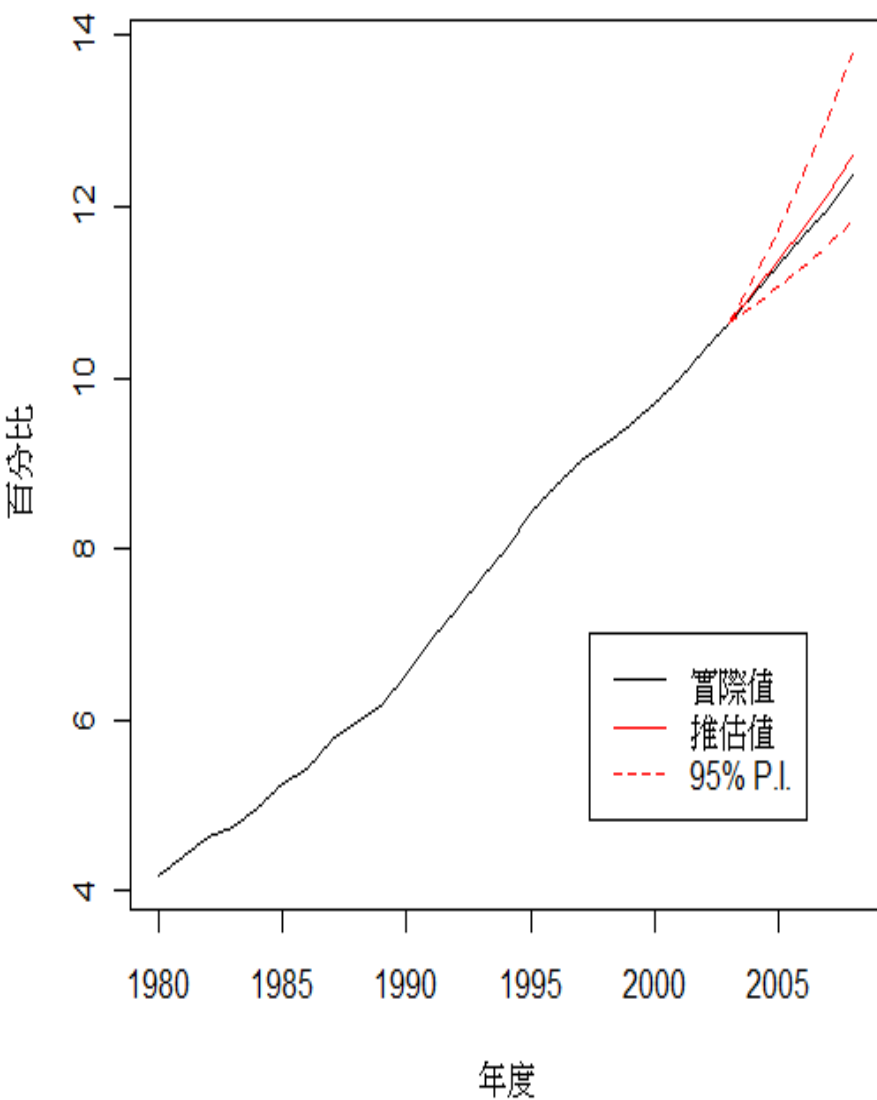


臺灣的人口遷移(遷入與遷出)推估  
(中位數、68%預測區間及95%預測區間)

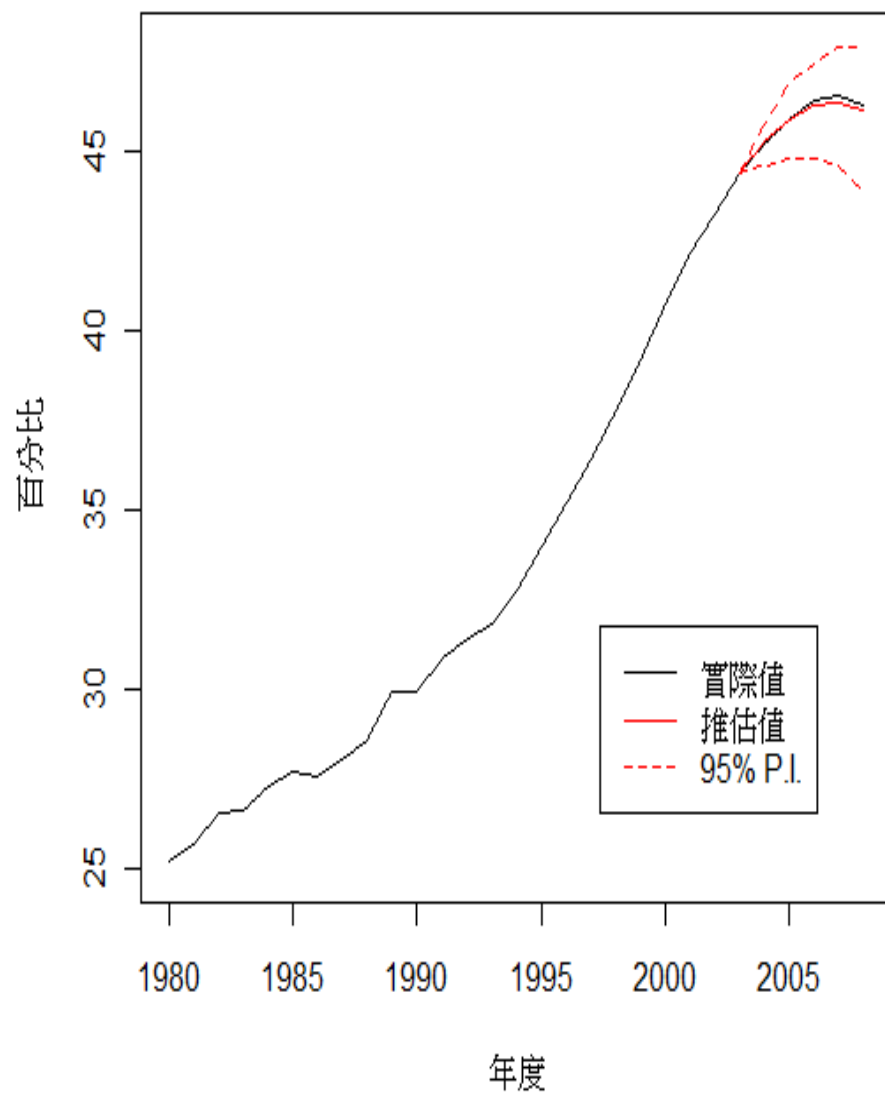
資料來源：李芯柔、余清祥(2008)

# 區塊拔靴法範例：臺北市高齡人口比例

## 65 到 99 歲人口比例



## 75 到 99 歲人口佔 65 到 99 歲人口比例

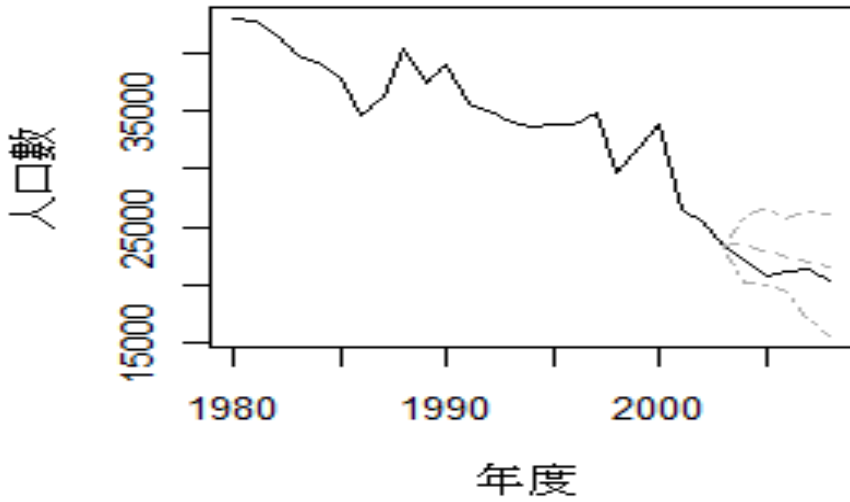


# 不同推估方法的誤差比較 (臺北市人口推估)

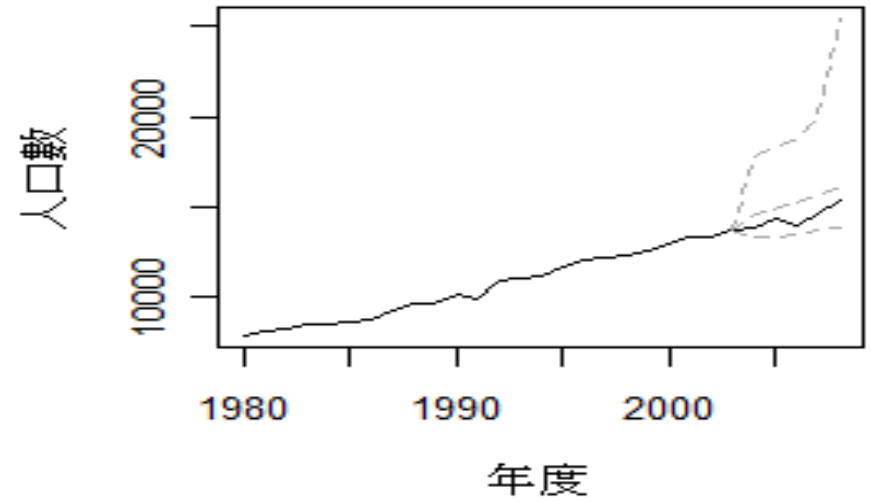
要素	本研究	臺北市主計處			
		世代生存法	迴歸分析法	ARIMA (直接)	ARIMA (間接)
總人口	0.80	1.70	3.17	0.87	1.38

# 臺北市區塊拔靴法推估值（虛線為推估值）

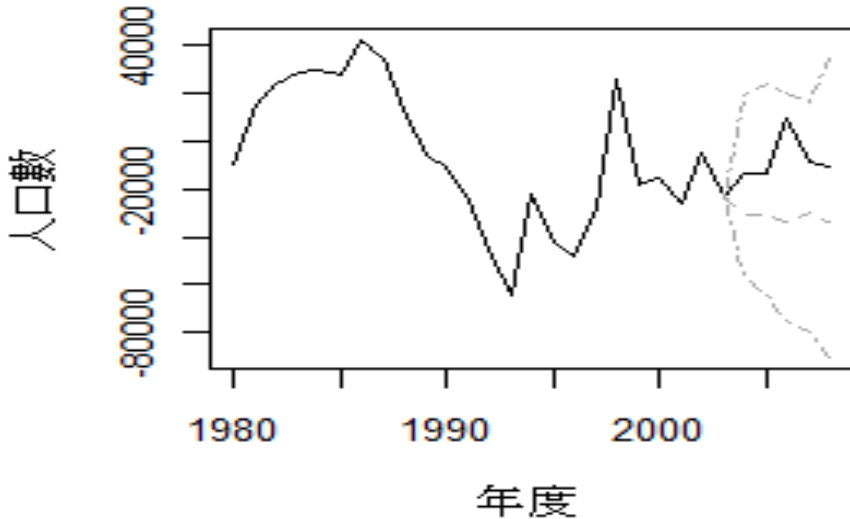
## 出生



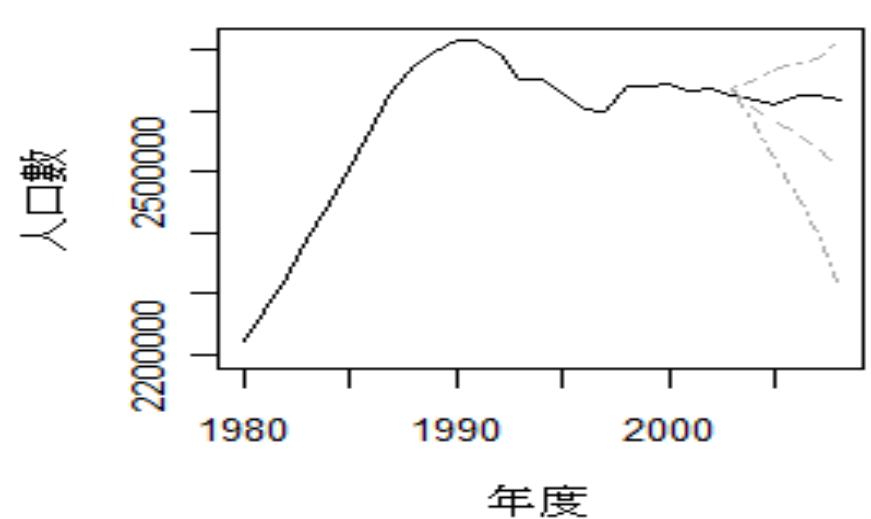
## 死亡



## 淨遷入



## 總人口數



# 幾種死亡率推估方法

方法	方法設定	參數
Block Bootstrap	Linear Weight	Block size = 5
Lee-Carter	LC	SVD
	LC	PCA
Sieve	AR	AR(p)
FPCA	BSpline	8 basis functions
	Monomial	8 basis functions
	Polygonal	Knot 數由資料決定

## 雲林及嘉義兩縣在不同方法下的平均MAPE

Estimate Forecast	Block	LC		Sieve	FPCA		
		SVD	PCA		BSpline	Polygonal	Monomial
Estimate	NA	8.03	8.03	NA	0.00	0.00	0.00
Forecast	11.73	12.03	12.03	10.26	20.21	13.42	19.28