電腦模擬與研究方法

政治大學統計系

余清祥(csyue@nccu.edu.tw)

2002年12月12日

課程下載：140.119.81.22

甚麼是模擬？

- 模擬(Simulation)：模仿虛擬

→Feign, pretend to have or feel; pretend to be, act like, resemble, …

- 模擬與我們日常生活的關係

→風洞(Wind-tunnels)：汽車與飛機

→飛行模擬器(Flight simulator)

→第一個生命的誕生(閃電高熱及電擊)

→ CPU的研發

- 模擬現在廣為學術界各領域專家使用，已是解決問題的常見工具。

→ 商業上的風險分析(Risk analysis)

→ 經濟學上的政策模擬(policy simulation)

→ 都市計畫中的交通流量(traffic flow)

→ 量子力學、積體電路設計

蒙地卡羅法(Monte Carlo Methods)

- 電腦模擬方法多以蒙地卡羅法稱之，意謂模擬的資料與統計方法的使用。

→ 資料的模擬牽涉亂數的產生(Random Number Generation)。

- 何謂亂數？
- 為什麼稱做蒙地卡羅法？

■ 為什麼要使用電腦模擬？

→ 成本較低

→ 操作較為容易

→ 無法進行真正的實驗

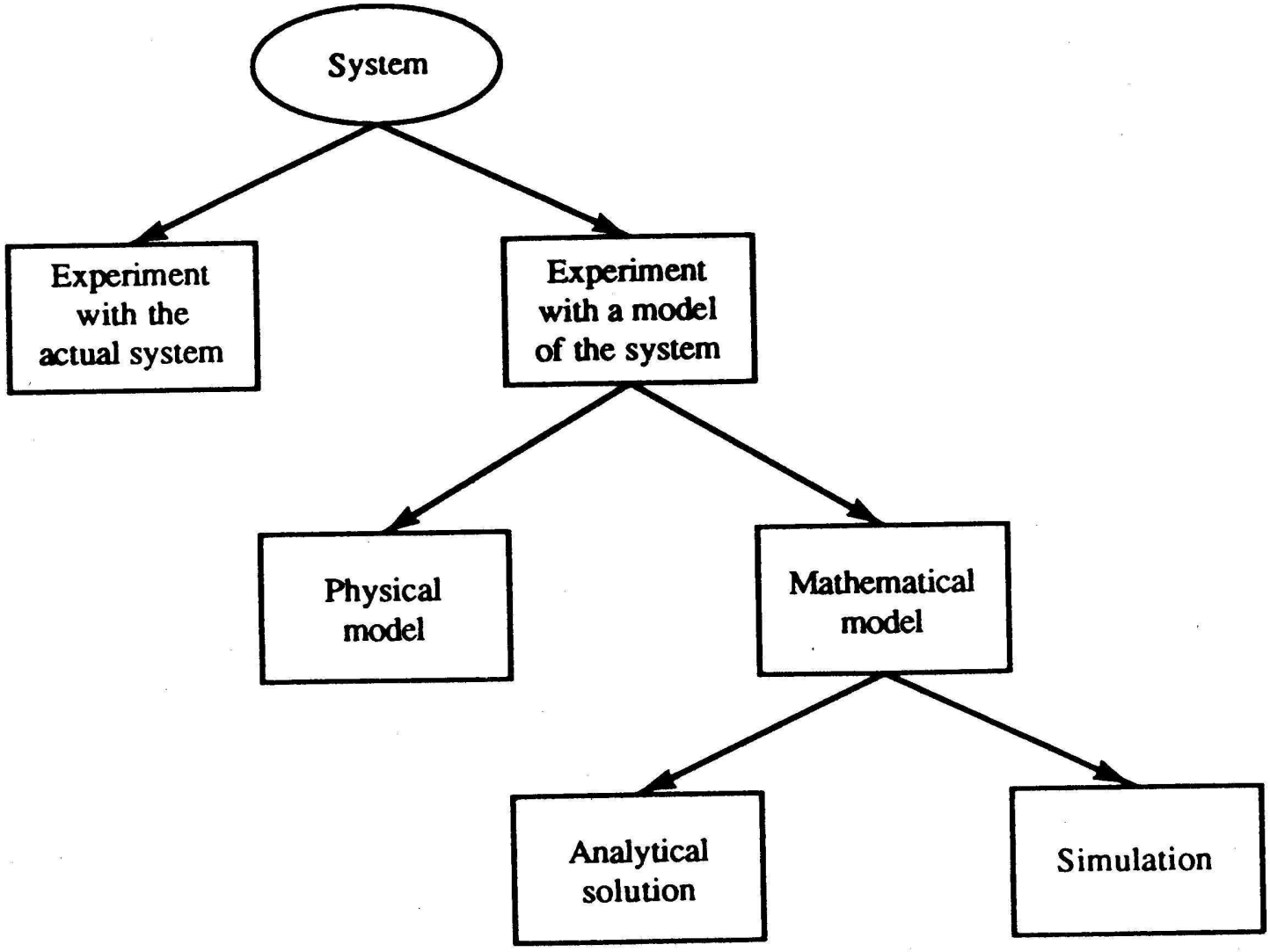
■ 統計分析的三大利器

→ 數理統計(Mathematical Statistics)

→ 實驗設計(Experimental Design)

→ 電腦模擬(Computer Simulation)

系統與模擬

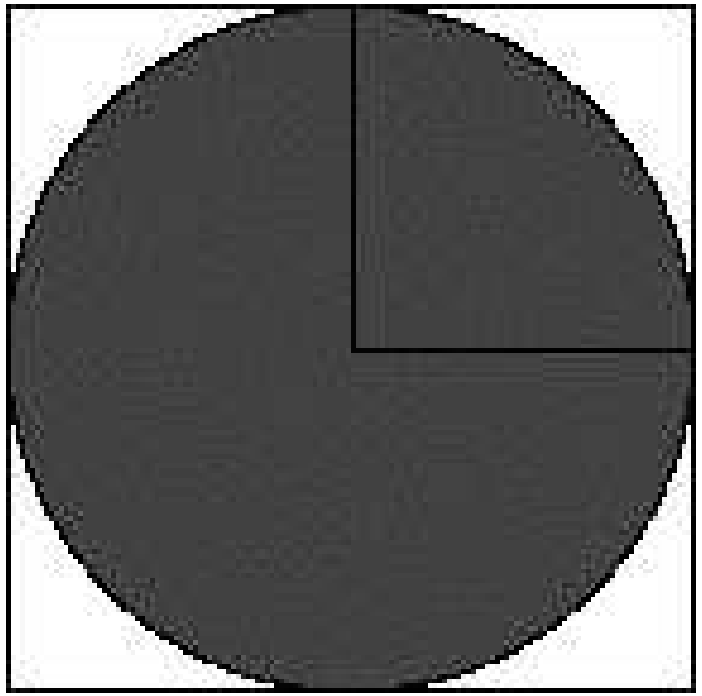

System
Experiment with the actual system
Experiment with a model of the system
Physical model
Mathematical model
Analytical solution
Simulation

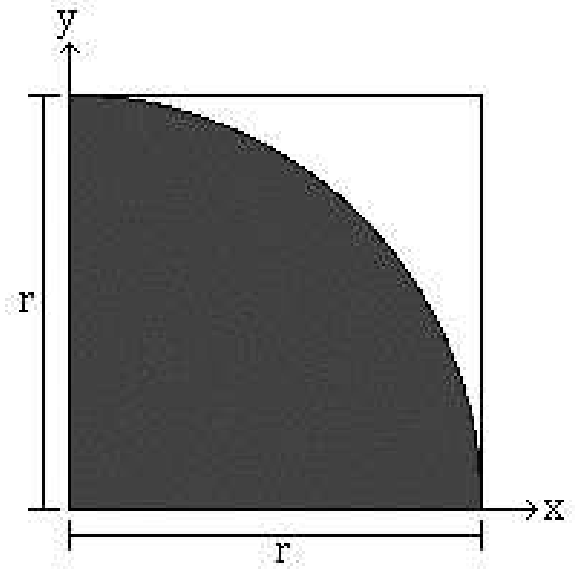
模擬分析的要件

- Probability distribution function
- Random number generator
- Sampling rule
- Scoring/Tallying
- Error estimation
- Variance Reduction techniques
- Parallelization/Vectorization

電腦模擬的實例

- 實例一、如何估計 π？

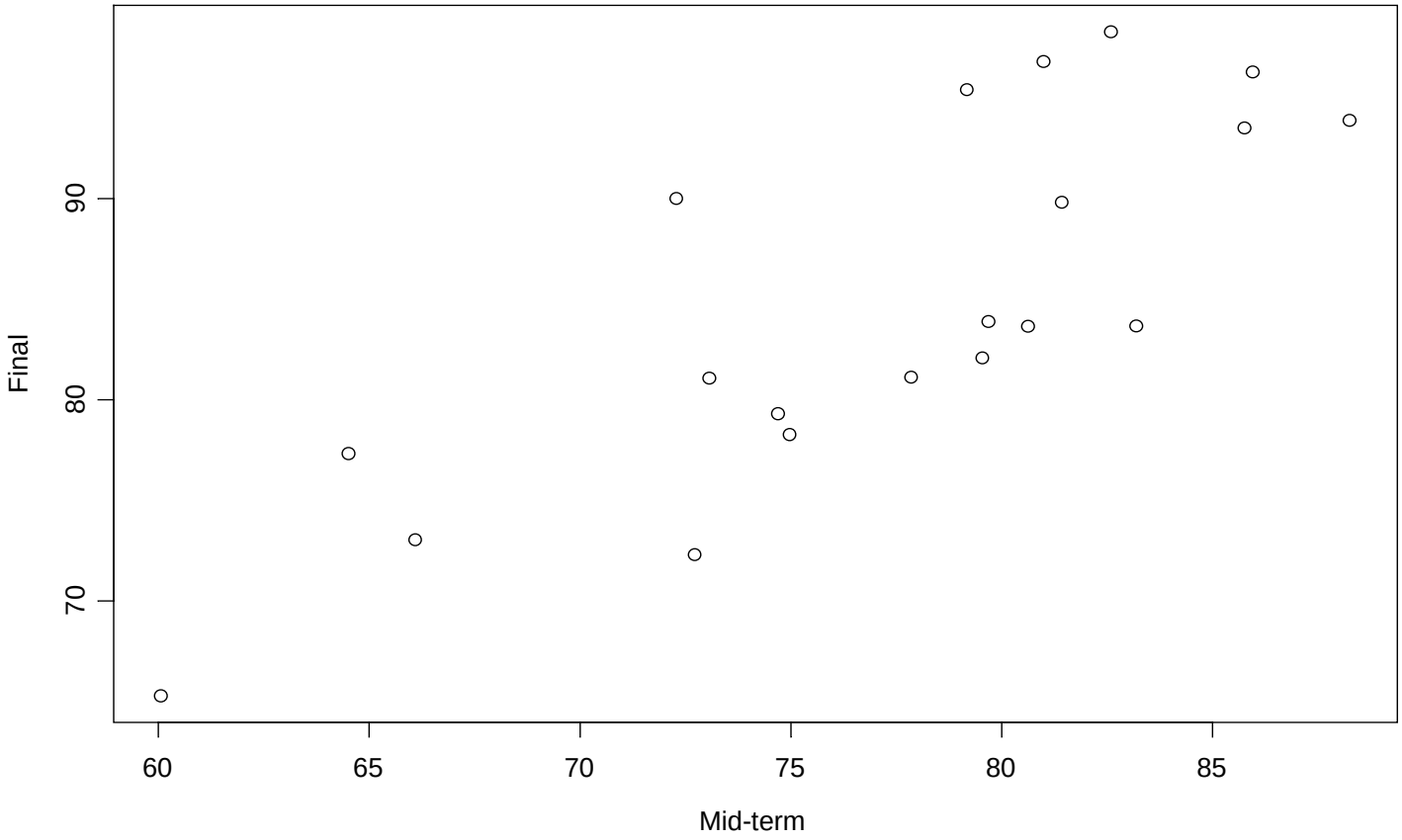




- 如果你/妳飛鏢射得很好，可以均勻地將飛鏢射在上圖的正方形中。則π的估計值可由下式獲得：

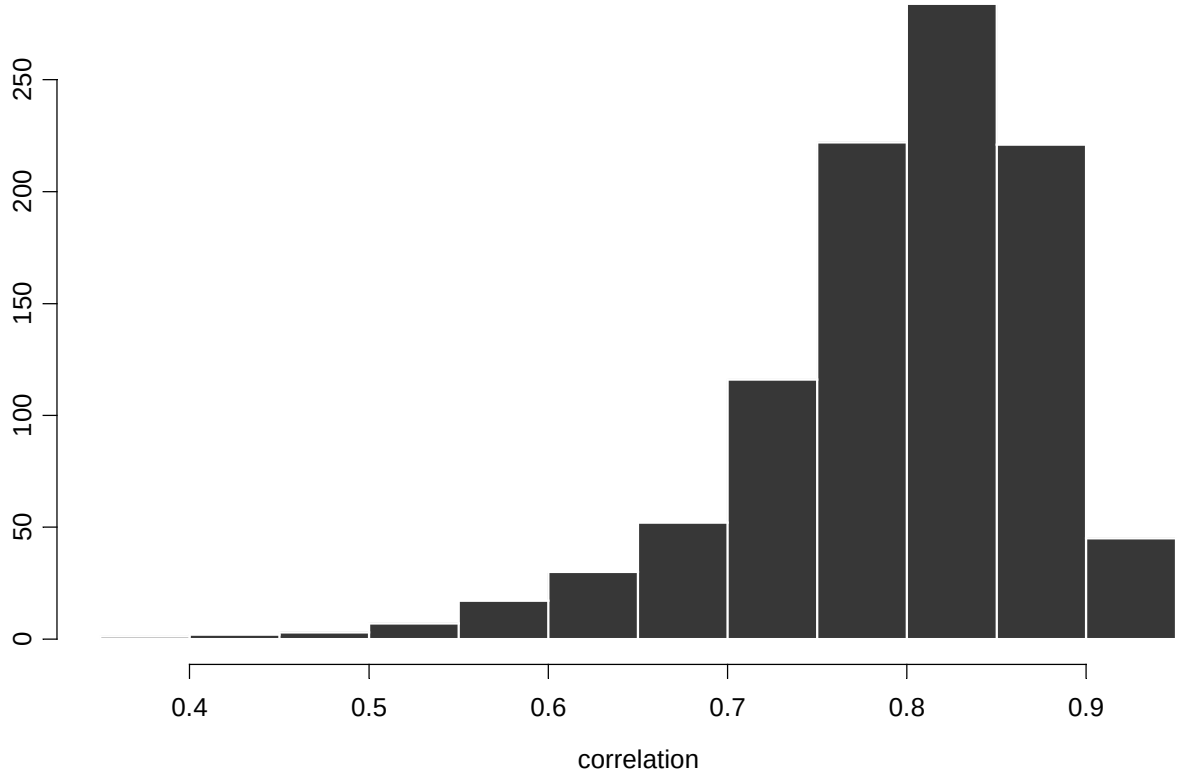
$$\frac{\text{射在藍色區域內的飛鏢數}}{\text{射在正方形內的飛鏢數}} = \frac{\text{藍色區域面積}}{\text{正方形面積}} = \frac{\pi}{4}$$

- 實例二、我想瞭解統計學期中考與期末考兩次考試的相關性，隨機在500名學生中抽出20名學生。

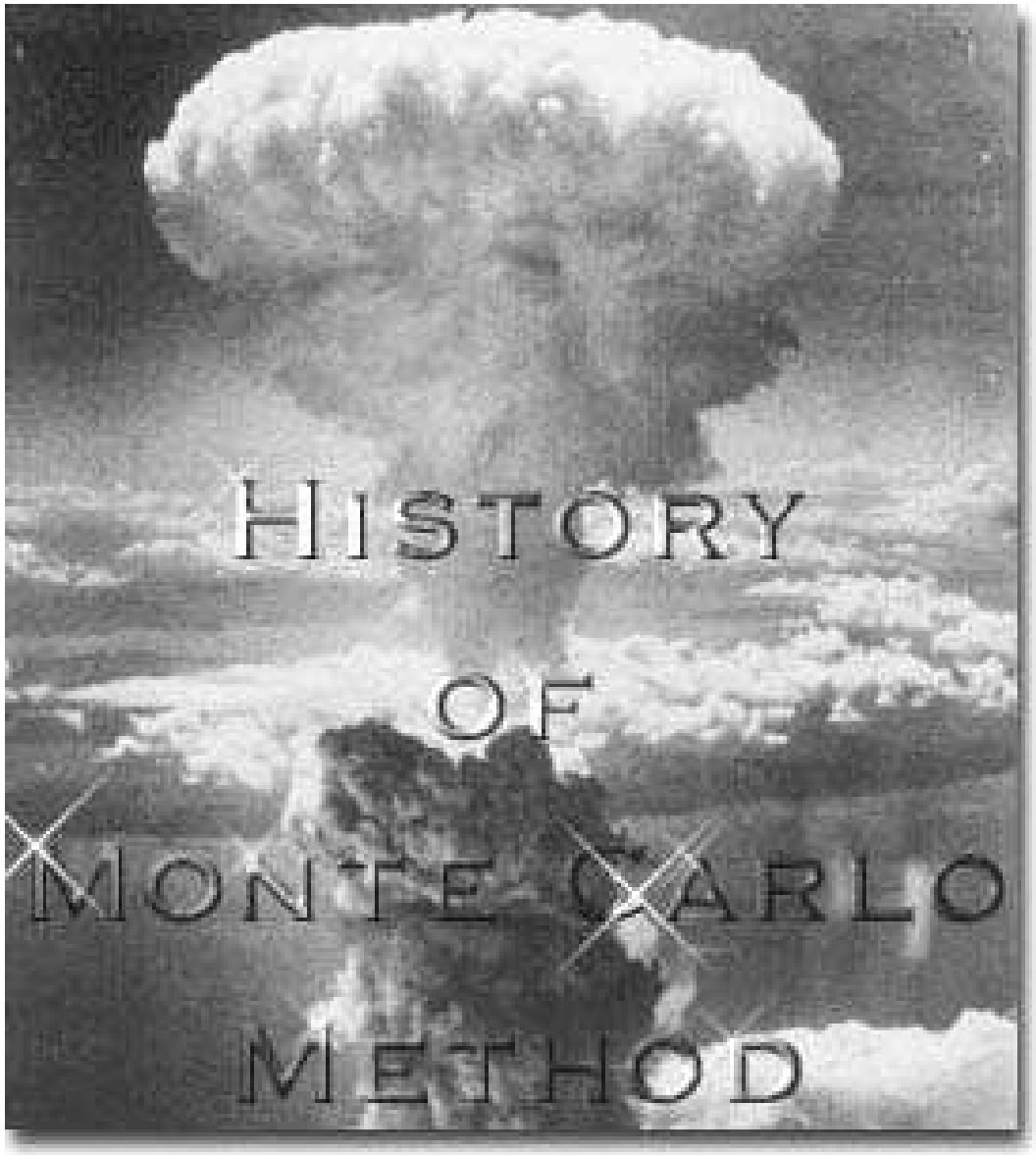


- 樣本相關係數在多數軟體中都有 (約為 0.8036)，但相關係數的標準差呢？

→ 拔靴法(Bootstrap)的1000次模擬計算可得標準差大約是0.0820。



蒙地卡羅方法的歷史



蒙地卡羅法的歷史(續)

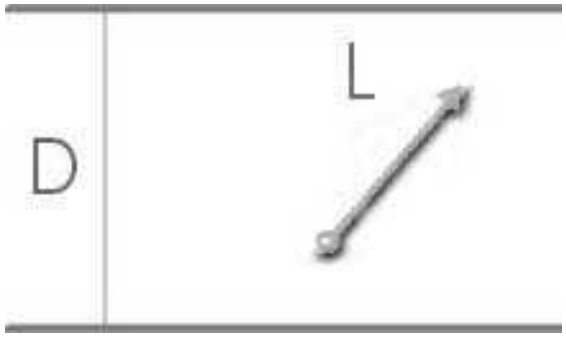
- How did Monte Carlo simulation get its name?
- The name and the systematic development of Monte Carlo methods dates from about 1940’s.
- There are however a number of isolated and undeveloped instances on much earlier occasions.

蒙地卡羅法的歷史(續)

- In the second half of the nineteenth century a number of people performed experiments, in which they threw a needle in a haphazard manner onto a board ruled with parallel straight lines and inferred the value of PI =3.14… from observations of the number of intersections between needle and lines.
- In 1899 Lord Rayleigh showed that a one-dimensional random walk without absorbing barriers could provide an approximate solution to a parabolic differential equation.

Buffon's Needle Experiment

Buffon's original form was to drop a needle of length L at random on grid of parallel lines of spacing D.



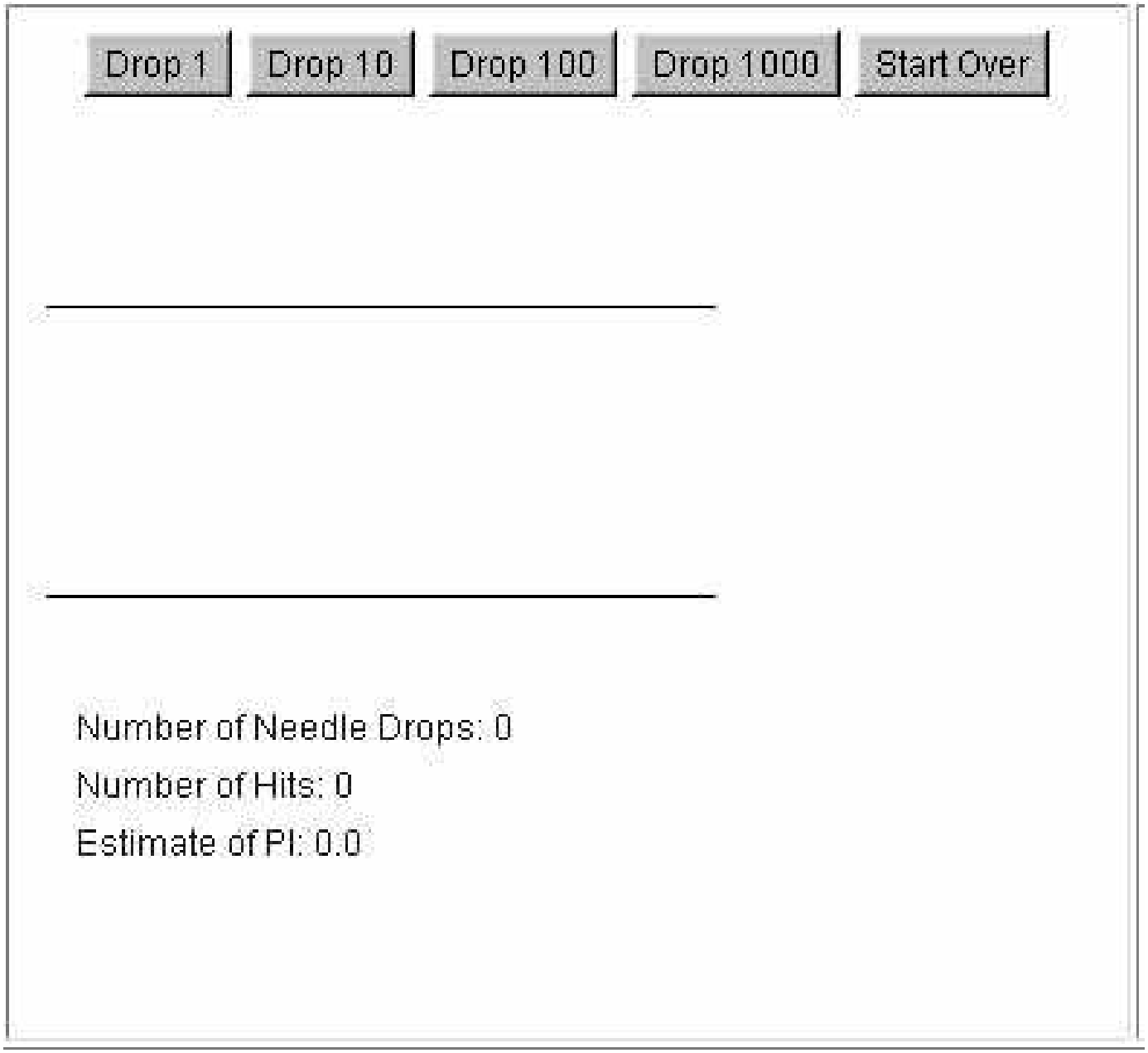
For L less than or equal D we obtain

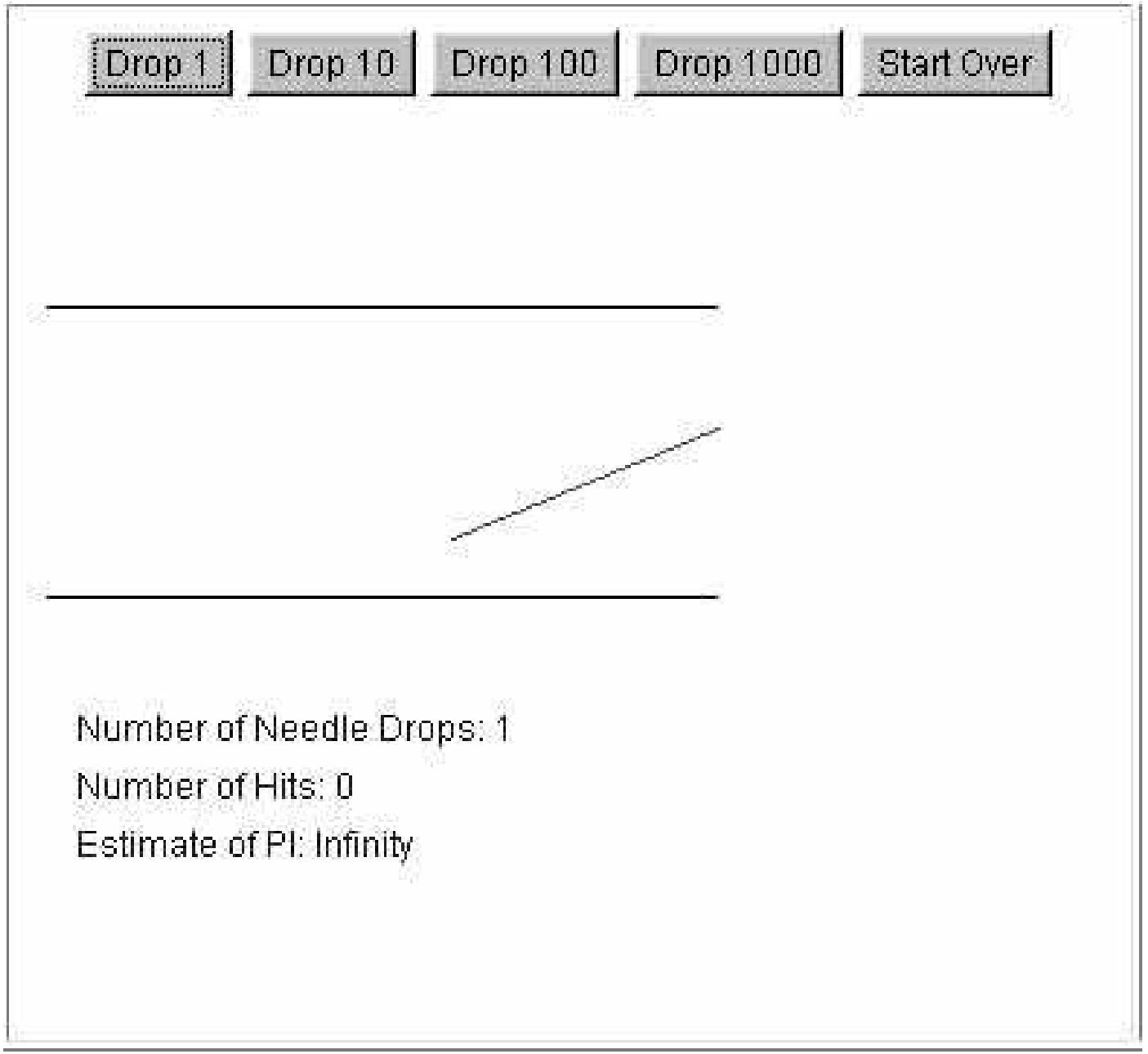
P(needle intersects the grid) = 2 • L / PI • D.

If we drop the needle N times and count R intersections we obtain

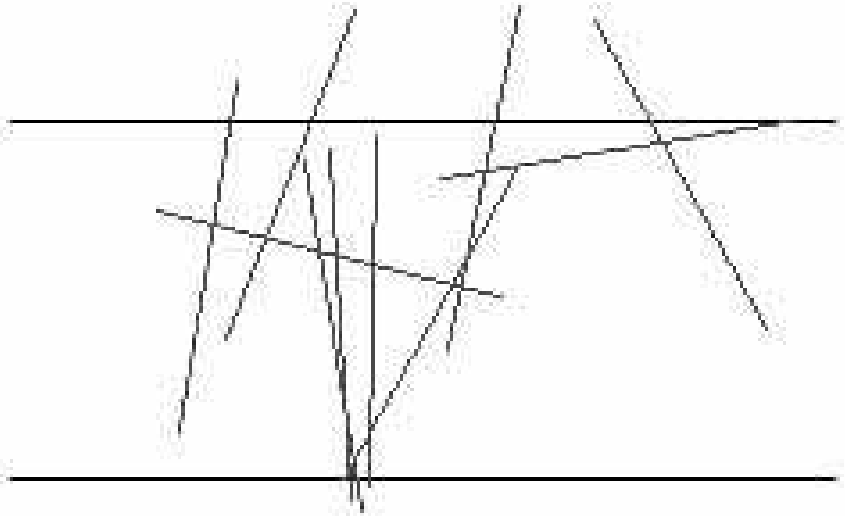
P = R / N,

PI = 2 • L • N / R • D.


Drop 1
Drop 10
Drop 100
Drop 1000
Start Over
Number of Needle Drops: 0
Number of Hits: 0
Estimate of PI: 0.0


Drop 1
Drop 10
Drop 100
Drop 1000
Start Over
Number of Needle Drops: 1
Number of Hits: 0
Estimate of PI: Infinity

Drop 1 | Drop 10 | Drop 100 | Drop 1000 | Start Over

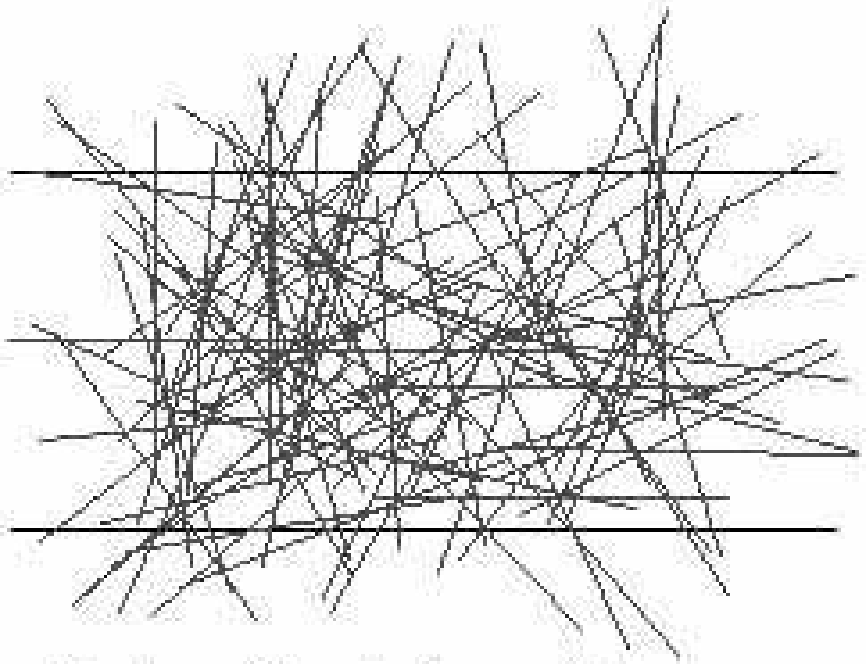


Number of Needle Drops: 11

Number of Hits: 8

Estimate of PI: 2.75

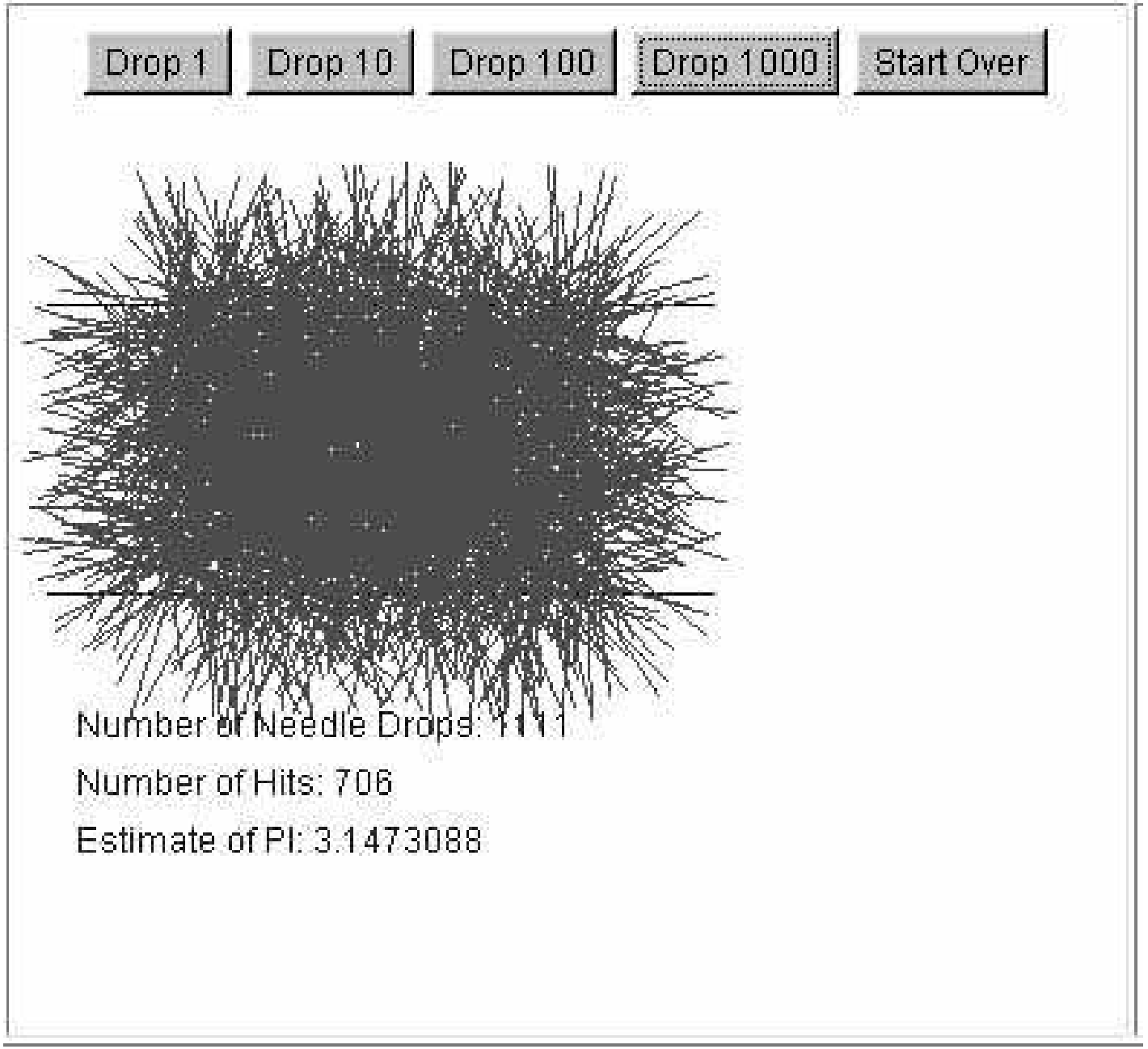
Drop 1 | Drop 10 | Drop 100 | Drop 1000 | Start Over



Number of Needle Drops: 111

Number of Hits: 69

Estimate of PI: 3.2173913


Drop 1
Drop 10
Drop 100
Drop 1000
Start Over
Number of Needle Drops: 1111
Number of Hits: 706
Estimate of PI: 3.1473088

蒙地卡羅法的歷史(續)

- In early part of the twentieth century, British statistical schools indulged in a fair amount of unsophisticated Monte Carlo work.
- In 1908 Student (W.S. Gosset) used experimental sampling to help him towards his discovery of the distribution of the correlation coefficient.
- In the same year Student also used sampling to bolster his faith in his so-called t-distribution, which he had derived by a somewhat shaky and incomplete theoretical analysis.

History of Monte Carlo Method

Student - William Sealy Gosset (1876 - 1937)

This birth-and-death process is suffering from labor pains; it will be the death of me yet. (Student Sayings)

A. N. Kolmogorov (1903-1987)

In 1931 Kolmogorov showed the relationship between Markov stochastic processes and certain integro-differential equations.

蒙地卡羅法的歷史(續)

- The real use of Monte Carlo methods as a research tool stems from work on the atomic bomb during the second world war.
- This work involved a direct simulation of the probabilistic problems concerned with random neutron diffusion in fissile material; but even at an early stage of these investigations, von Neumann and Ulam refined this particular "Russian roulette" and "splitting" methods. However, the systematic development of these ideas had to await the work of Harris and Herman Kahn in 1948.
- About 1948 Fermi, Metropolis, and Ulam obtained Monte Carlo estimates for the eigenvalues of Schrodinger equation.

John von Neumann (1903-1957)

蒙地卡羅法的歷史(續)

- In about 1970, the newly developing theory of computational complexity began to provide a more precise and persuasive rationale for employing the Mont Carlo method.
- Karp (1985) shows this property for estimating reliability in a planar multiterminal network with randomly failing edges.
- Dyer (1989) establish it for estimating the volume of a convex body in M-dimensional Euclidean space.
- Broder (1986) and Jerrum and Sinclair (1988) establish the property for estimating the permanent of a matrix or, equivalently, the number of perfect matchings in a bipartite graph.

亂數的產生

- 早期的亂數多藉由實體(Physical)方法產生，例如：擲銅板、骰子、撲克牌、輪盤。之後也有Rand公司的百萬個數字組成的亂數表，Von Neumann提議的Mid-square 法。
- 同餘(Congruential)法是現今最廣為使用的方法，基本原理為

$$X_{i+1} = aX_i + c \pmod{m}$$

其中a, c, m 為整數。

— John Von Neumann (Mid-Square Method)

e.g. 任選一個四位數的數字，計算其平方數

3307 → 10**9362**49 (取中間四位數再平方)

→ 87**6470**44

→ 41**8609**00

註：我們也可取六位數、二位數、…

e.g. 33 → 1**08**9 → **6**4 not good！

123456 →**241**383 → **265**752

(but small numbers → small numbers)

亂數的產生(續)

- 例如：$a = c = 3, m = 5, X_0 = 1$

  $\Rightarrow X_i = 3, 2, 4, 0, 3, \cdots$ (週期為4)

- 基本要求：

→週期愈長愈佳、數字出現無特殊規律

- 不錯的亂數選擇範例：

→ 十進位：$X_{i+1} = 101X_i + 1 \pmod{1000}$

→ 二進位：$X_{i+1} = (2^{17} + 3)X_i \pmod{2^{35}}$

■ 組合產生器：(多數的軟體採用的方式)

→組合產生器是利用兩個或兩個以上的亂數產生器，來組合成一組新的亂數產生器。例如：

$$x_i = 171x_{i-1} \pmod{30269}$$

$$y_i = 172y_{i-1} \pmod{30307}$$

$$z_i = 170z_{i-1} \pmod{30323}$$

$$u_i = \frac{x_i}{30269} + \frac{y_i}{30307} + \frac{z_i}{30323} \pmod{1}$$

亂數的產生(續)

- 由同餘法(或其他方法)產生介於 *0* 至*m*–1的亂數，除以 *m* 後即成為0與1間的亂數，也就是類似*U(0,1)*的隨機變數。

→其他分配的亂數可藉由*U(0,1)*產生。

(例如：*Exp(1) = – log(U(0,1))*。)

- 掌上型計算機(諸如Casio)使用的也是同餘法：

$$U_{n+1} = (\pi + U_n)^5 \pmod 1$$

- 如何確定亂數具有*U(0,1)*的特性？

  (亂數需具有哪些特質？)

→ 均勻分配(Uniformity)

多半都使用適合度(Goodness-of-fit)的檢定，確定亂數具有要求的特質，常用的方法有卡方適合度檢定(Chi-square goodness-of-fit)及Kolmogorov-Smirnov 檢定。

→ 互相獨立(Independence)

使用的多為無母數方法，例如：Gap test, Up-and-down test, run test.

好的亂數產生器的要件

- 長週期：週期長則不易發生重覆使用相同亂數數列的情形。
- 良好的統計性質：希望亂數產生器所產生的亂數數列能夠滿足統計上獨立且均勻分配的性質。
- 有效率的電腦執行：為了便於模擬，相同的亂數數列必須具有重覆執行（reproduce）的能力。演算的速度快，亂數的產生過程中佔用較少的時間和記憶體。

Inverse Transformation Method

Generates a random number from a probability density function by solving the probability density function's variable in terms of randomly generated numbers. This is achieved as follows:

- We solve the inverse of the integral of our probability density function at an arbitrary point a F(a), in terms of a random number r.
- We generate a unique random variable a, as follows:

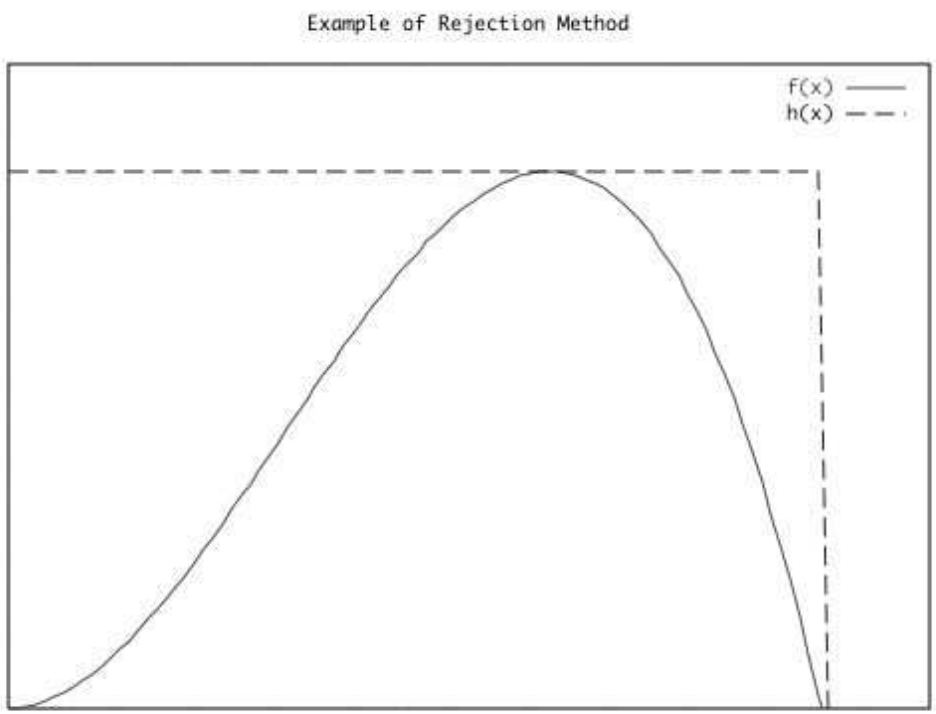
$$a = F^{-1}(r)$$

Rejection Method

Generates random numbers for a distribution function f(x). This is achieved as follows:

- Define a comparison function h(x) such that it encloses the desired function f(x).
- Choose uniformly distributed random points under h(x).
- If a point lies outside the area under f(x) reject it and choose another point.

Illustration of the Rejection Method



The following is an illustration of the rejection method using a square function for the comparison function.

常見統計軟體的亂數產生

■ 常見的統計軟體：(統計系使用)

→ SAS

→ SPSS

→ S-Plus

→ Minitab

→ EXCEL

SAS的亂數產生

- SAS 6.12 內設U(0,1)亂數產生器：

→其中乘數設為397,204,094，除數設為 $2^{31}-1$ ，先藉由下式產生 x_i：

$$x_i = 397{,}204{,}094 x_{i-1} (\bmod 2^{31}-1) \quad i = 1,2,3,...$$

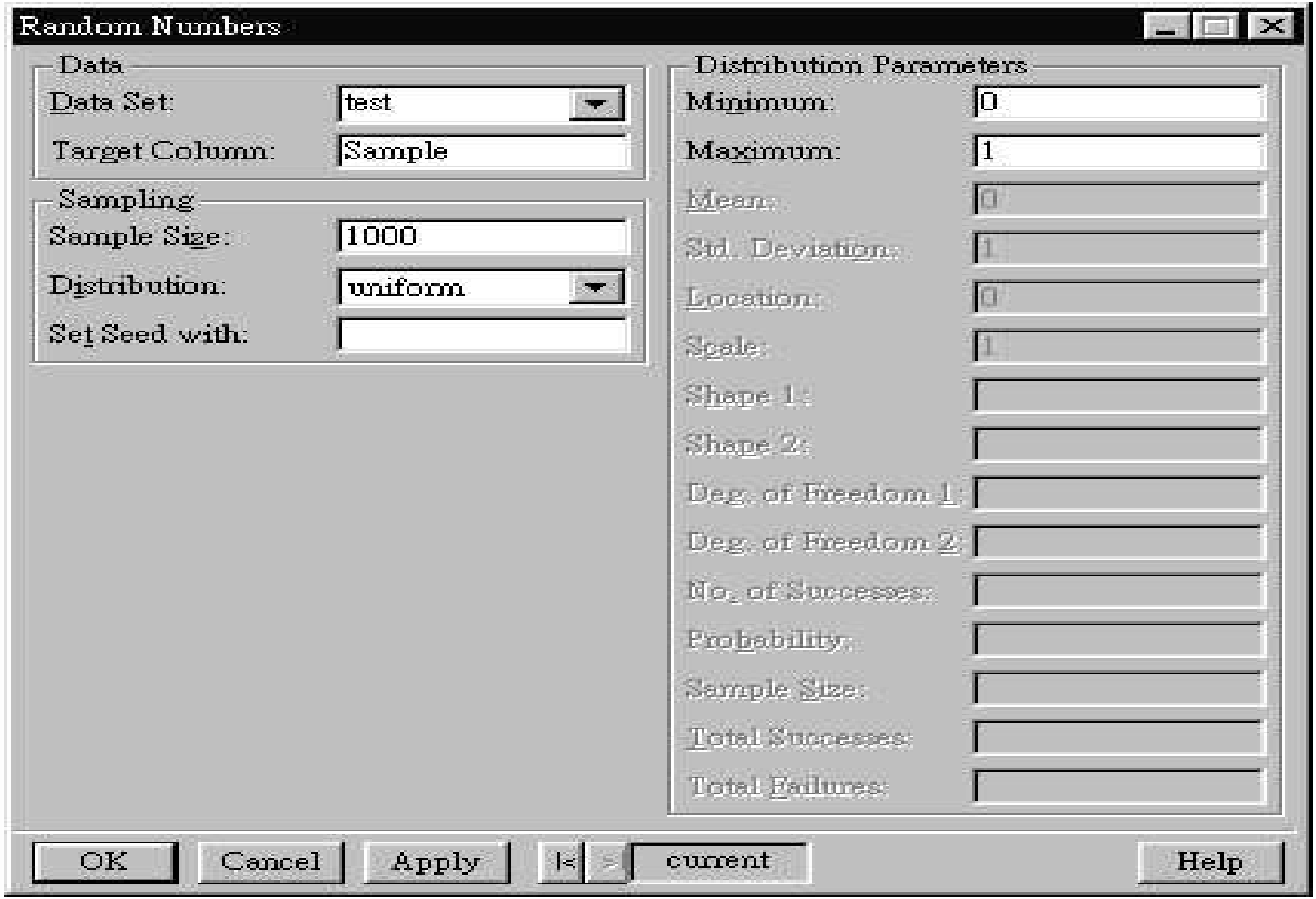
再除以 $2^{31}-1$ 而產生U(0,1)的亂數，即

$$u_i = \frac{x_i}{2^{31}-1}$$

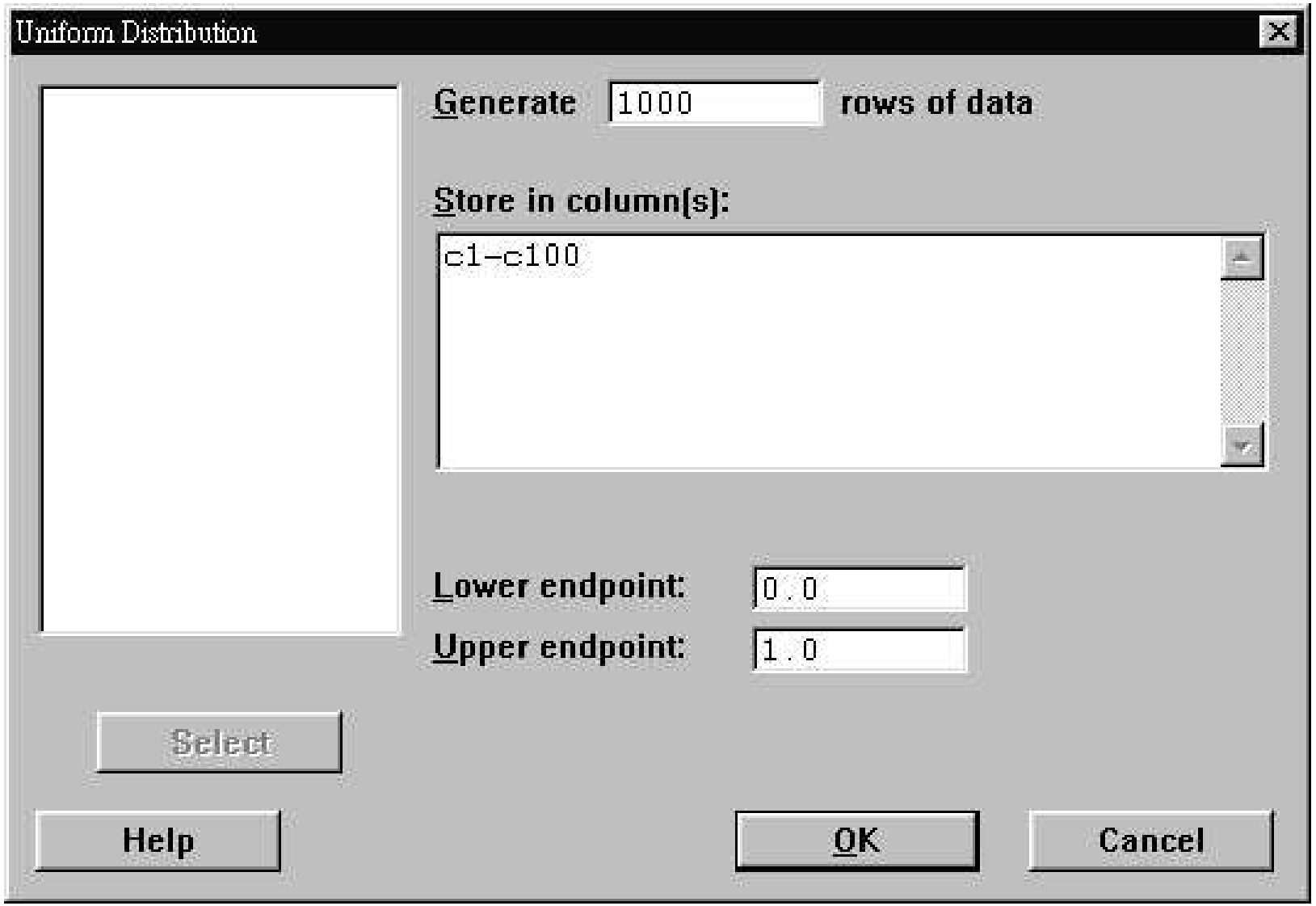
SPSS的亂數產生


計算變數
目標變數(T):
v1
=
類型註解(L)...
數值運算式(E):
UNIFORM(1)
sn
+ < > 7 8 9
- <= >= 4 5 6
* = ~= 1 2 3
/ & | 0 .
** ~ ( ) 刪除
函數(F):
TRUNC(數值表示式)
UNIFORM(最大值)
UPCASE(字串表示式)
VALUE(變數)
VARIANCE(數值表示式,數值
XDATE.DATE(日期值)
若(I)...
確定
貼上語法(P)
重設(R)
取消
輔助說明

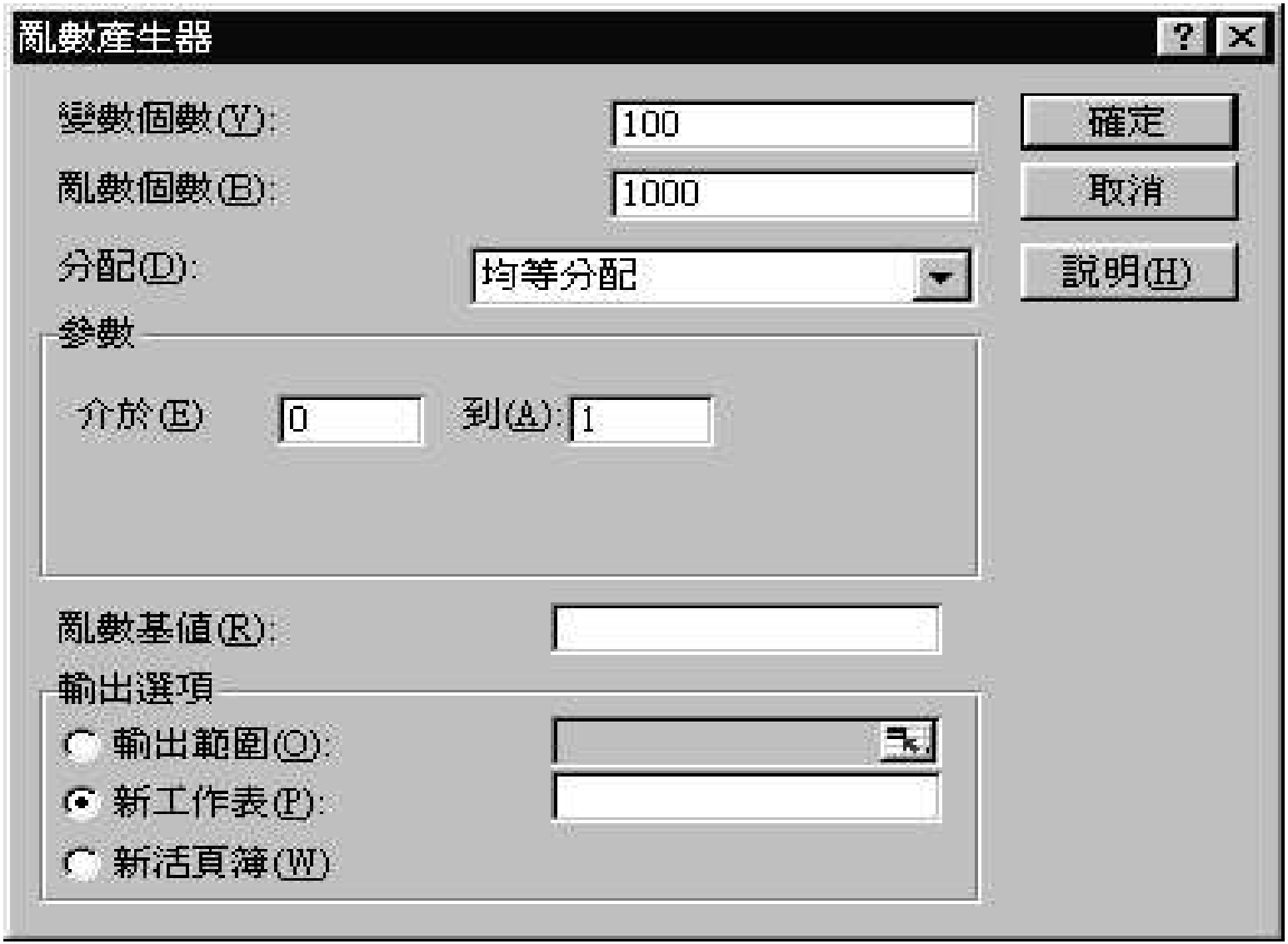
S-Plus的亂數產生


Random Numbers
Data
Data Set:
test
Target Column:
Sample
Sampling
Sample Size:
1000
Distribution:
uniform
Set Seed with:
Distribution Parameters
Minimum:
0
Maximum:
1
Mean:
0
Std. Deviation:
1
Location:
0
Scale:
1
Shape 1:
Shape 2:
Deg. of Freedom 1:
Deg. of Freedom 2:
No. of Successes:
Probability:
Sample Size:
Total Successes:
Total Failures:
OK
Cancel
Apply
current
Help

Minitab的亂數產生



EXCEL的亂數產生


亂數產生器
變數個數(V):
100
亂數個數(B):
1000
分配(D):
均等分配
確定
取消
說明(H)
參數
介於(E)
0
到(A):
1
亂數基值(R):
輸出選項
輸出範圍(O):
新工作表(P):
新活頁簿(W)

隨機變數的分配函數	SAS 6.12	S-PLUS 2000	MINITAB 12	EXCEL 97	SPSS 8.0
均勻(uniform)分配	*	*	*	*	*
常態(normal)分配	*	*	*	*	*
指數常態(lognormal)分配		*	*		*
指數(exponential)分配	*	*	*		*
貝他(beta)分配		*	*		*
咖瑪(gamma)分配	*	*	*		*
科西(cauchy)分配	*	*	*		*
羅吉斯(logistic)分配		*	*		*
韋伯(weibull)分配		*	*		*
F 分配		*	*		*
T 分配		*	*		*
卡方(chi-square)分配		*	*		*
Laplace 分配			*		*
Pareto 分配					*
伯努利(bernoulli)分配			*	*	*
二項(binomial)分配	*	*	*	*	*
幾何(geometric)分配		*	*		*
負二項(negbinomial)分配		*			*
卜瓦松(poisson)分配	*	*	*	*	*
超幾何(hypergeometric)分配		*	*		*

模擬應用範例：

■ **Monte Carlo Sampling:**

→檢查抽出的樣本是否隨機。

應用範例：

→是否應該聽從某股市名嘴的建議，投資某些「熱門股票」？下表中的平均投資報酬為7.05% [=($468.873-$438)/$438]。

→ 問題：這個報酬是否明顯高於任意選股？

Table 3.1
The stocks picked by the analyst

	Price at 12/31/85	Price at 12/31/86	Dividend for 1986
PPG INDUSTRIES INC	51.000	72.875	1.880
HILTON HOTELS CORP	64.875	67.250	1.800
GENERAL ELECTRIC CO	72.750	86.000	2.370
MCDERMOTT INTL INC	18.250	21.750	1.800
LOUISIANA LAND & EXPLORATION	30.250	27.250	1.000
DRESSER INDUSTRIES INC	18.125	19.375	0.700
MCDONALD'S CORP	80.875	60.875	0.645
WEST POINT-PEPPERELL	43.375	52.125	2.385
SERVICE CORP INTERNATIONAL	31.250	24.750	0.293
AMERADA HESS CORP	27.250	23.750	0.000
Total	438.00	456.00	12.873

隨機抽取10支股票，在1000次的模擬中，有26%(260次)的機會報酬率不小於7.05%。

→ 與隨機選股無異！！

■ **Permutation test:**

→以排列組合的方式計算事件的發生機率，常見於兩組樣本的比較，尤其當資料不是常態分配時，或是樣本數較少時。

應用範例：

→某汽油添加劑宣稱可增加每公升的里程數，以下為各測試4次添加與不添加的實驗結果。

- ♦添加：33.0, 31.0, 34.5, 34.0
- ♦不添加：29.5, 32.0, 32.9, 31.5

→因為樣本數少，除非差異較大，檢定結果通常是不能拒絕兩者的期望值相等。(t-檢定的 p-value 等於0.169)

→8個觀察值任意抽出4個，所有可能為：

$$\binom{8}{4} = \frac{8\times7\times6\times5}{4\times3\times2\times1} = 105 \text{ 種}$$

若藉由Permutation test，知道有8種可能使得兩者的差異不小於觀察差異(即差異為6.6)。因此，p-value = 8/105 = 0.076。

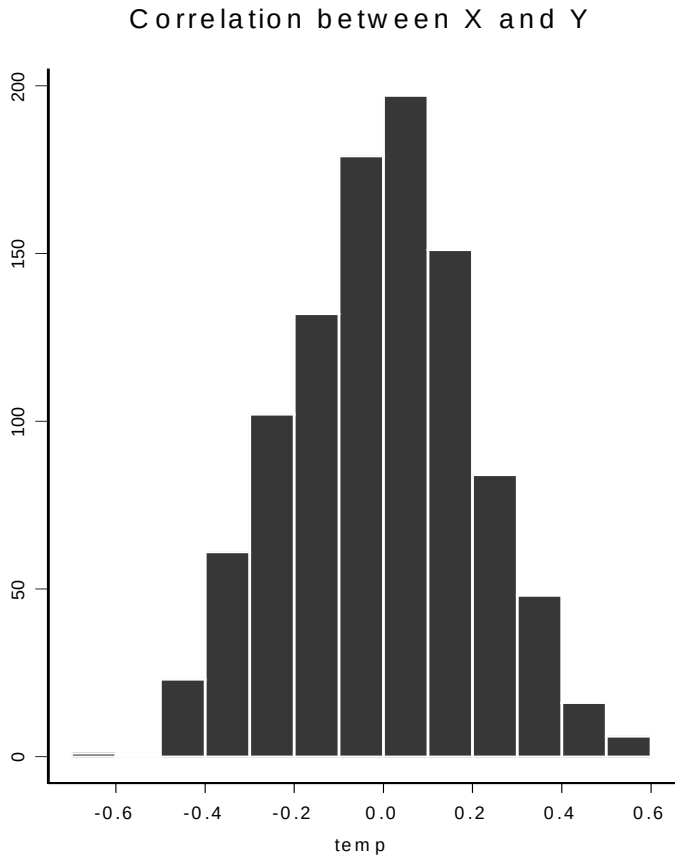
Monte Carlo p-value

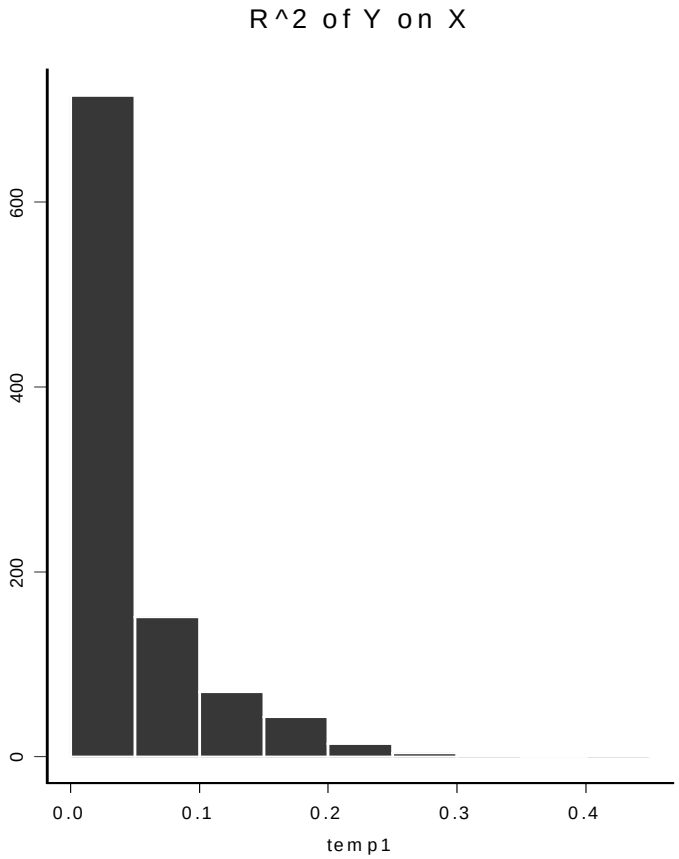
- 根據虛無假設模擬出 n 組樣本，如果我們的觀察值排在第 k 個(也就是第 k 大)，則檢定的 p-value = $k/(n+1)$。因為一般的顯著水準為0.05，n 值多半選為99、499、或是999等數字。

  應用範例：

→迴歸分析中除了 R^2 外，也有人用調整過的 R^2 來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





→平均R^2=4.2%，Q3＝5.9%。

另一個範例： $\varepsilon_i \sim i.i.d.\ N(0,\sigma^2), i=1,2,\ldots$

迴歸分析中常以殘差的符號變化，判斷殘差間是否互相獨立。例如：如果共有20個殘差，其正負號為

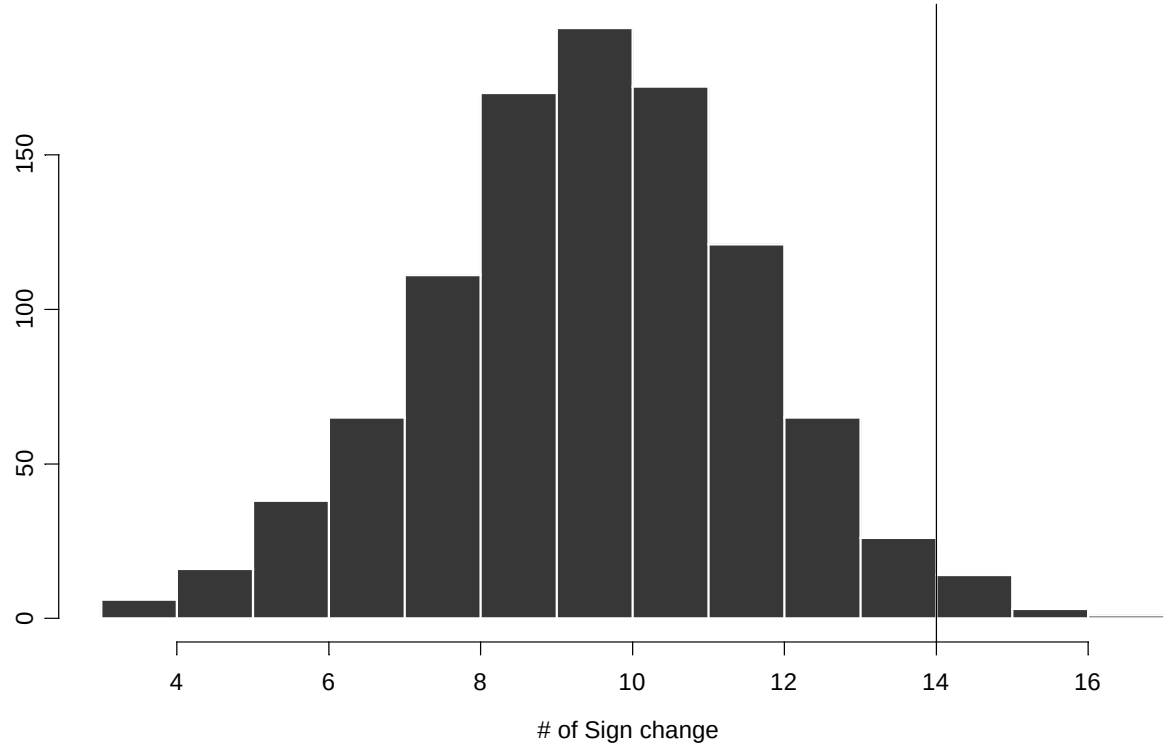
＋－－＋－＋－＋＋－＋－＋＋－－＋－－＋

共有14個符號改變，想檢查是否殘差間是否為負相關。

→ 符號改變頻繁，表示負相關；

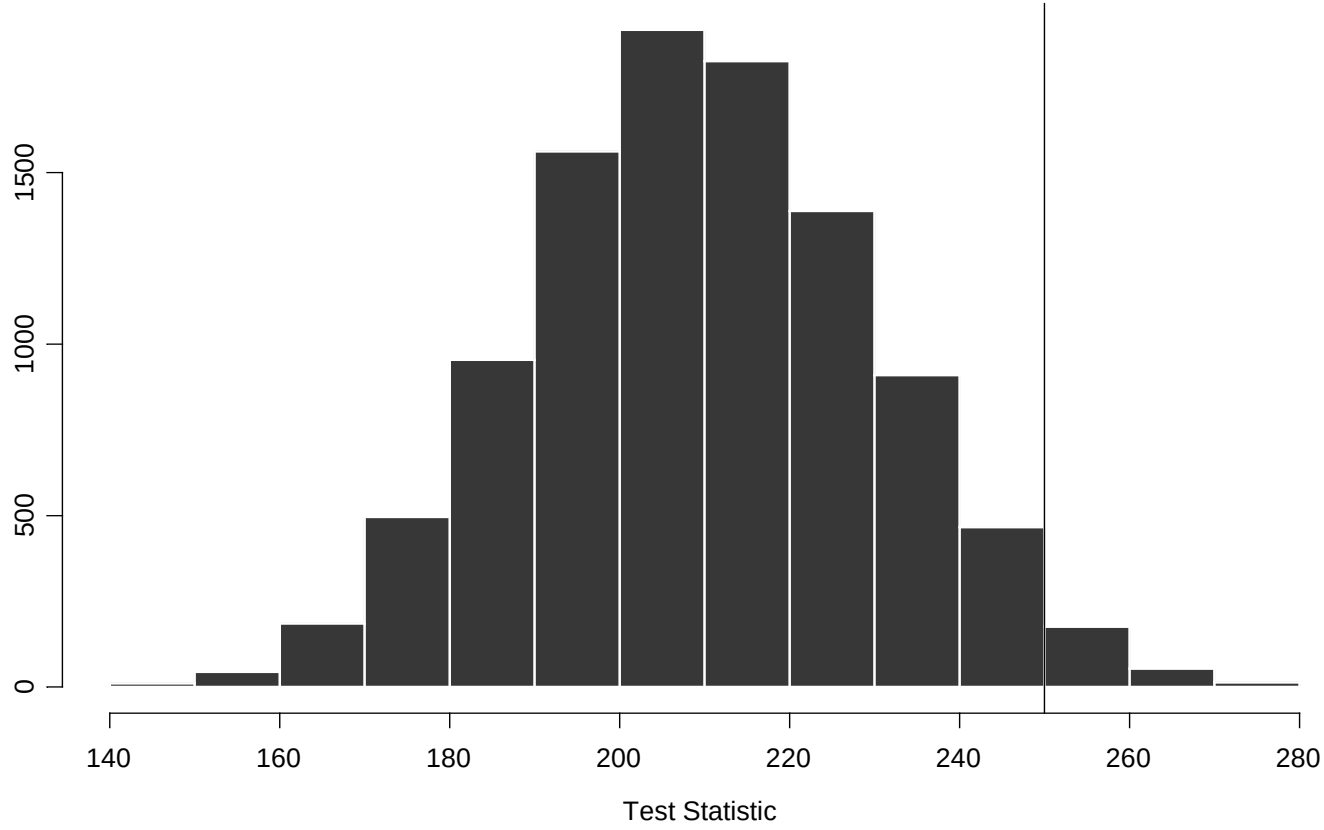
→ 符號不改變，表示正相關。

→以10個正號、10個負號為基礎，重複模擬999次，計算符號改變的次數。發現大於或等於14次符號改變的模擬次數有44次，因此 p-value = (44+1)/(999+1) = 0.045，也就是殘差間呈現負相關。



→特殊檢定的臨界值也可由模擬估算出。

例如：我們想查Wilcoxon rank sum test在比較12組及15組樣本，在α＝0.05的臨界值(一萬次模擬值＝250 vs. 理論值＝251)。

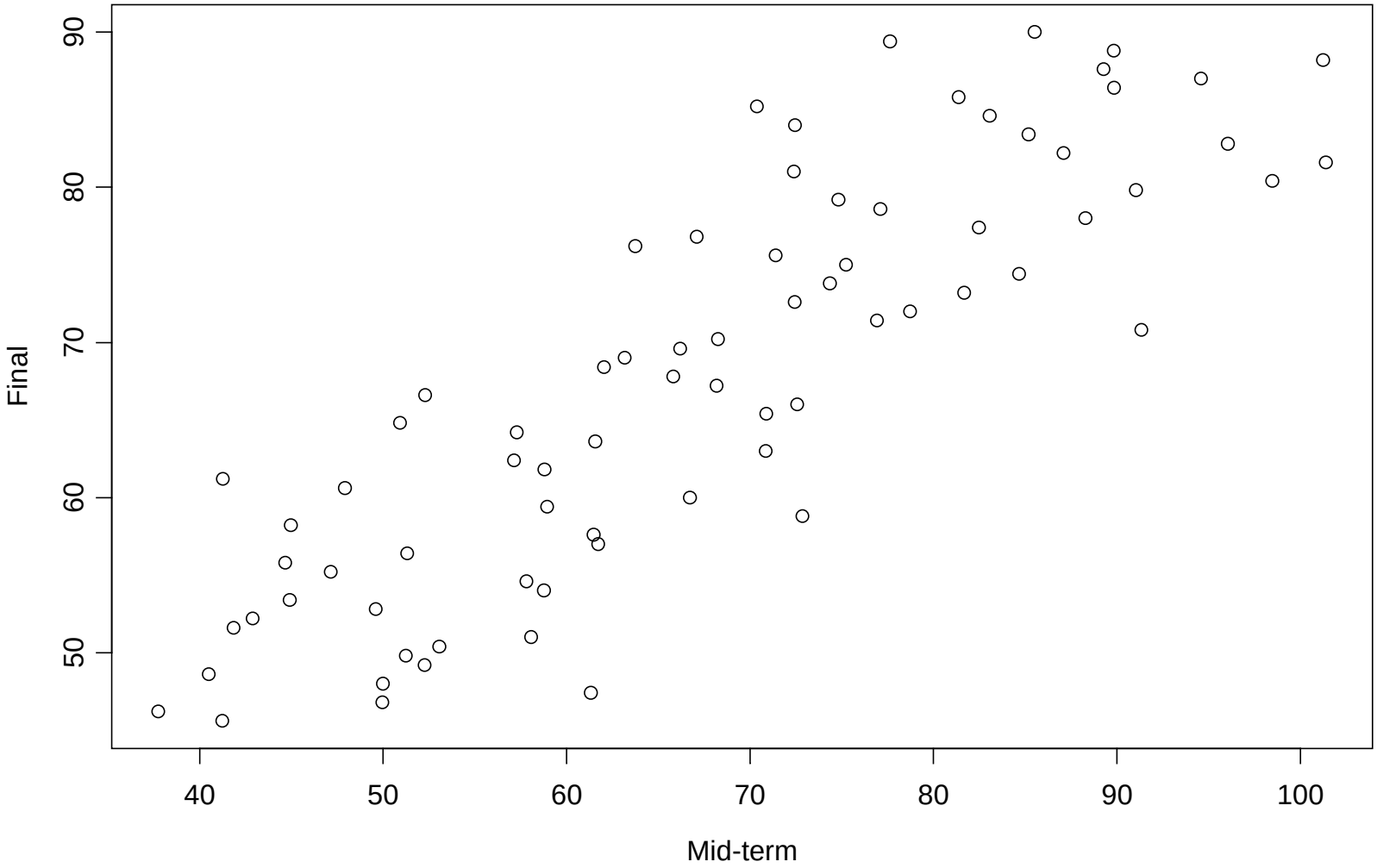


Bootstrap(拔靴法；梯雲縱)

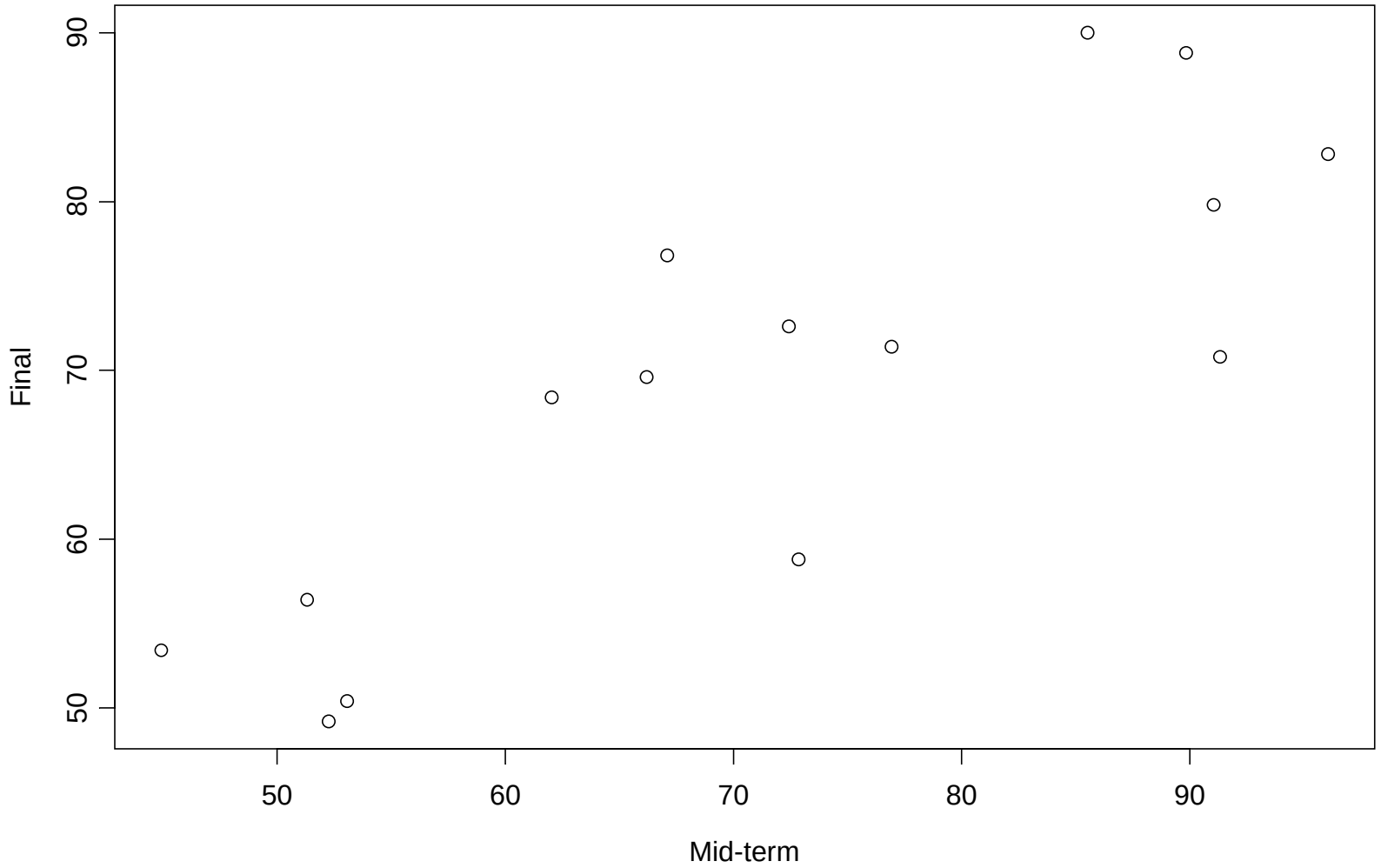
- Bootstrap法是Efron於1982所創立的方法，屬於重複抽樣(Resampling)方法。將已有的觀察值當作是母體重複抽樣，以求取原先因資料不足而無法估計的變異數。

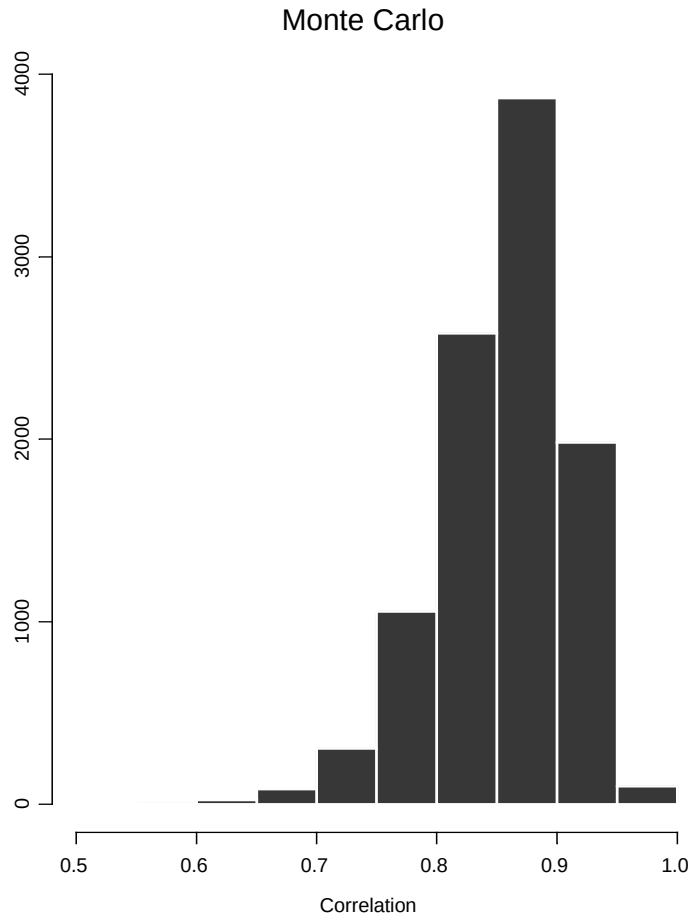
→舉例而言，假設 $x_1, x_2, \ldots, x_N$ 為來自同一分配的觀察值，而我們想瞭解這個分配的中位數與其中位數的變異數。

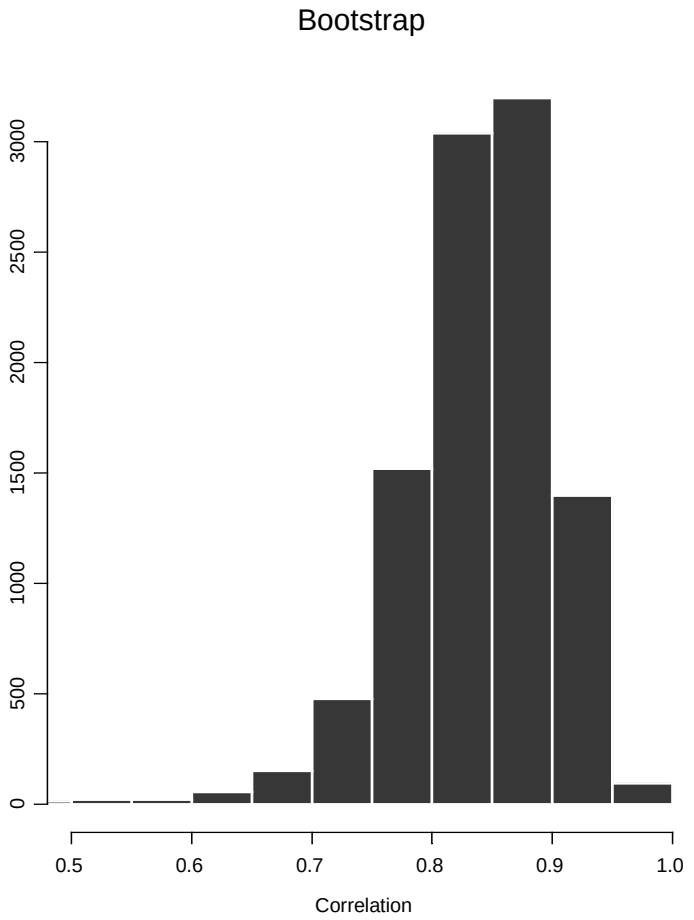
範例：75位選修統計學的學生，想瞭解期中考與總成績的相關性，只抽出15位學生。



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。





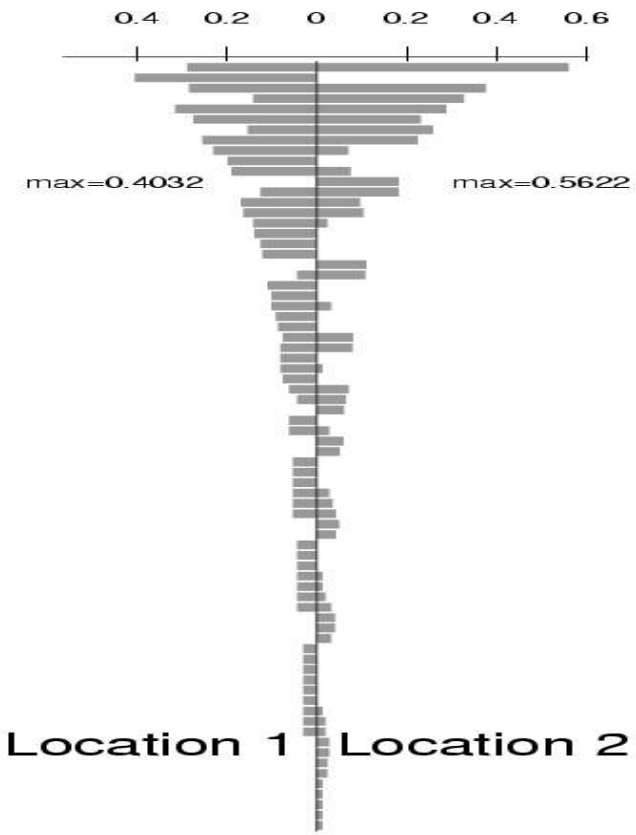


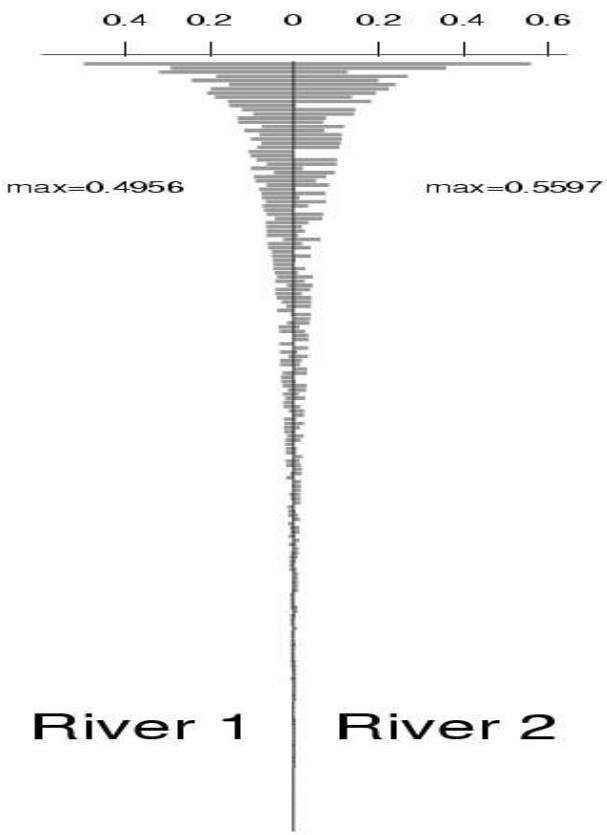
比較各一萬次模擬，Monte Carlo法由母體任意抽出15個樣本得出的相關係數，與15個樣本以Bootstrap法算出的相關係數。

→ 兩者非常接近！(標準差0.0540 vs. 0.0654)

Bootstrap法的實例

- 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：

	Bootstrap	**Delta**
螃蟹資料	0.015	0.01426
水鳥資料	0.008	0.0083

Bootstrap法的實例(續)

■ 檢定某個假設是否成立。

→ 範例：死亡率是否服從Gompertz 分配

也就是瞬間死亡率滿足

$$\mu_x = BC^x, \quad B>0, C>1$$

或是

$$log(p_{x+1})/log(p_x) = C$$

若C的信賴區間不為空集合，則不拒絕Gompertz假設。

■ 日本的資料不滿足、瑞典的資料滿足。

Fig 1-1. Bootstrap C.I. for Japanese Male

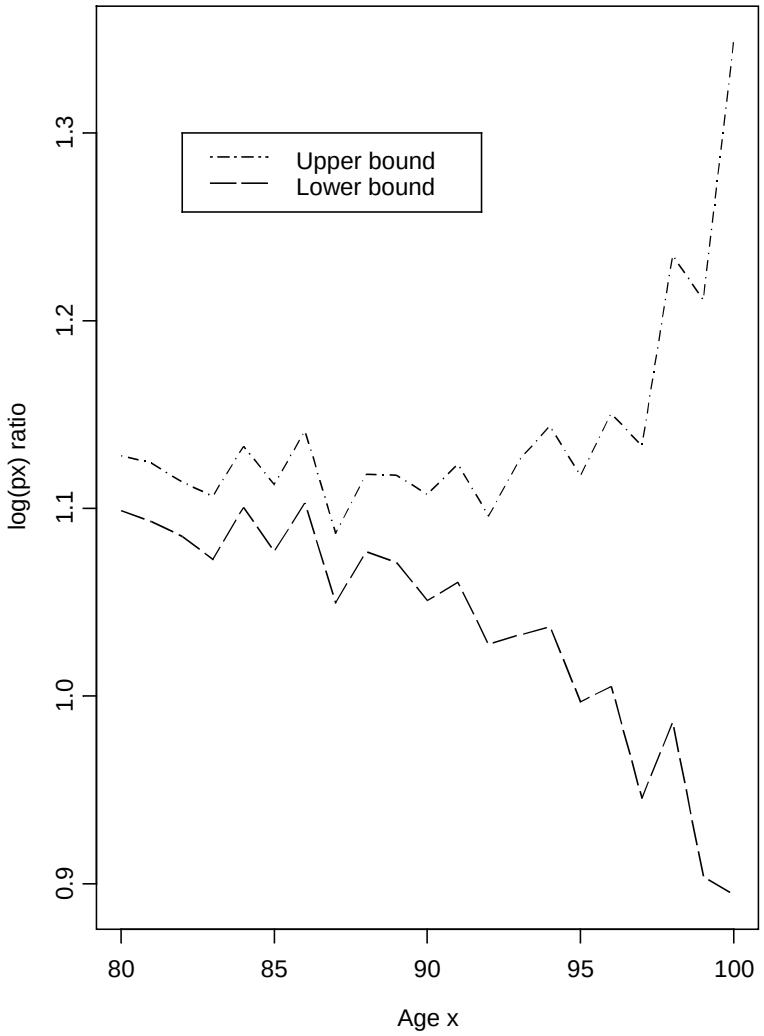
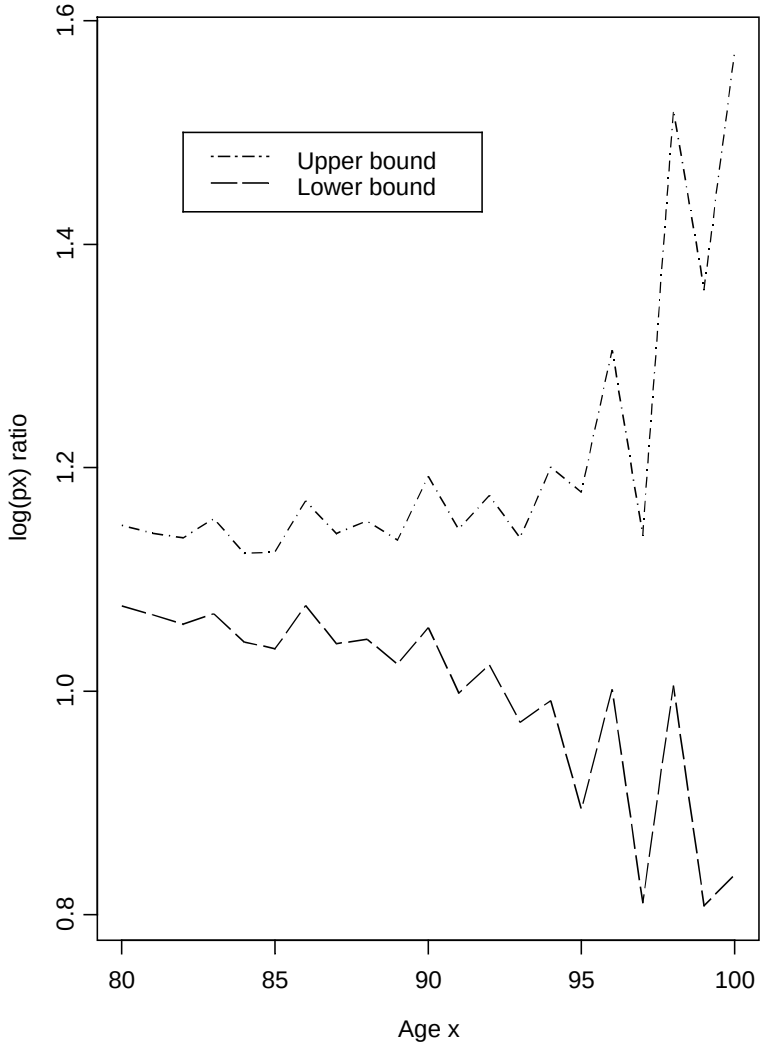
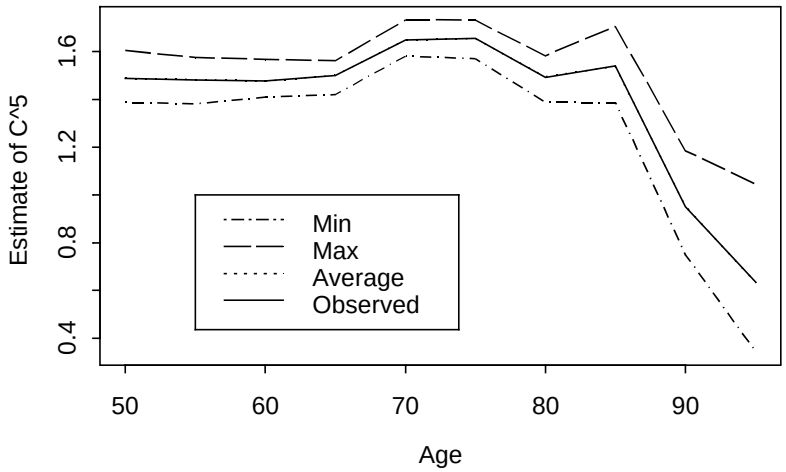


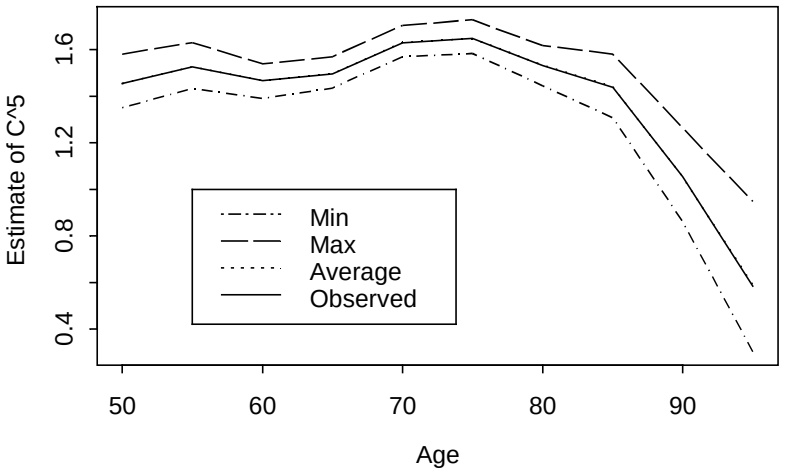
Fig 1-2. Bootstrap C.I. for Sweden Male



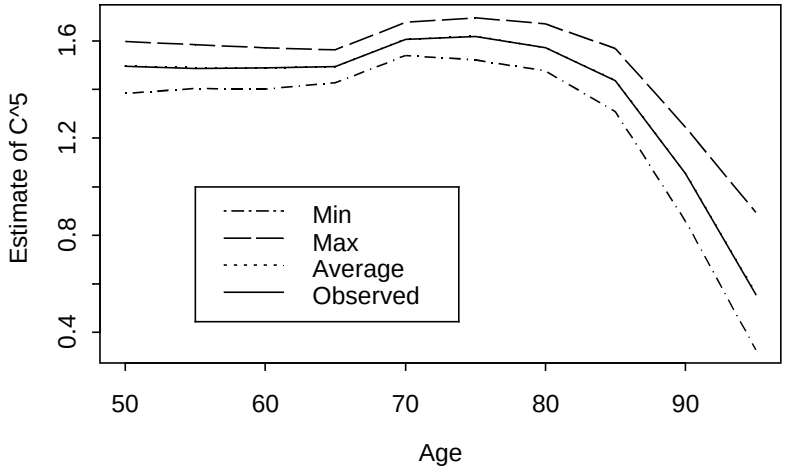
Bootstrapping C.I. for Gompertz (1998 Male)



Bootstrapping C.I. for Gompertz (1999 Male)



Bootstrapping C.I. for Gompertz (2000 Male)



1998至2000年台灣男性高齡死亡率的結果

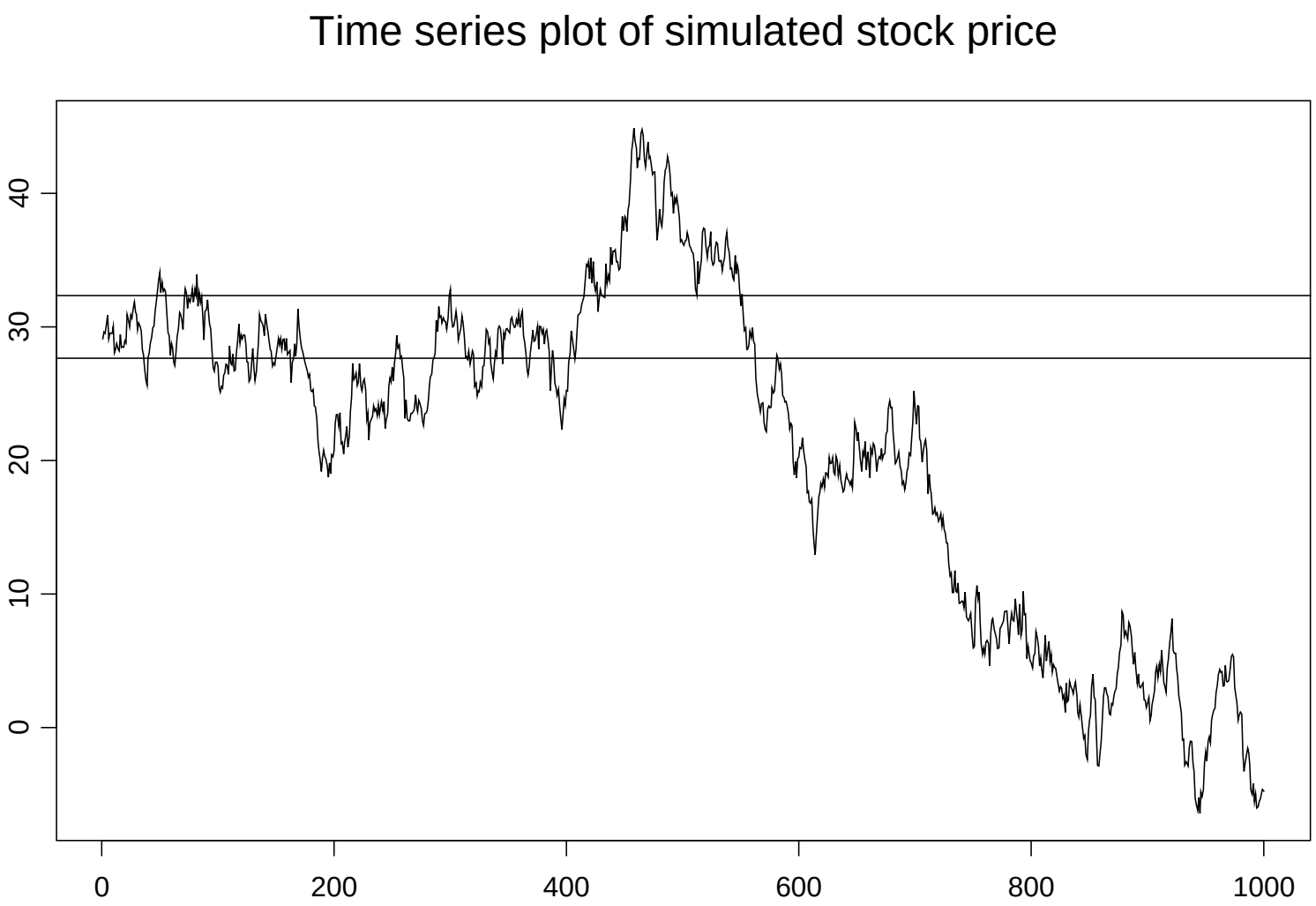
股市投資策略模擬

- 亞洲股市中據說台灣股市最接近隨機漫步(Random Walk)，如何在雜亂無章的訊息中獲取利潤？

→隨機漫步也就是假設時間 t 的股價$B(t)$：

$$B(t+s)-B(t)\,|\,B(t)\sim N(s\mu, s\sigma^2)$$

→「逢低買進、逢高賣出！」



1000次模擬的年平均報酬為16%！

排隊理論的等待時間模擬

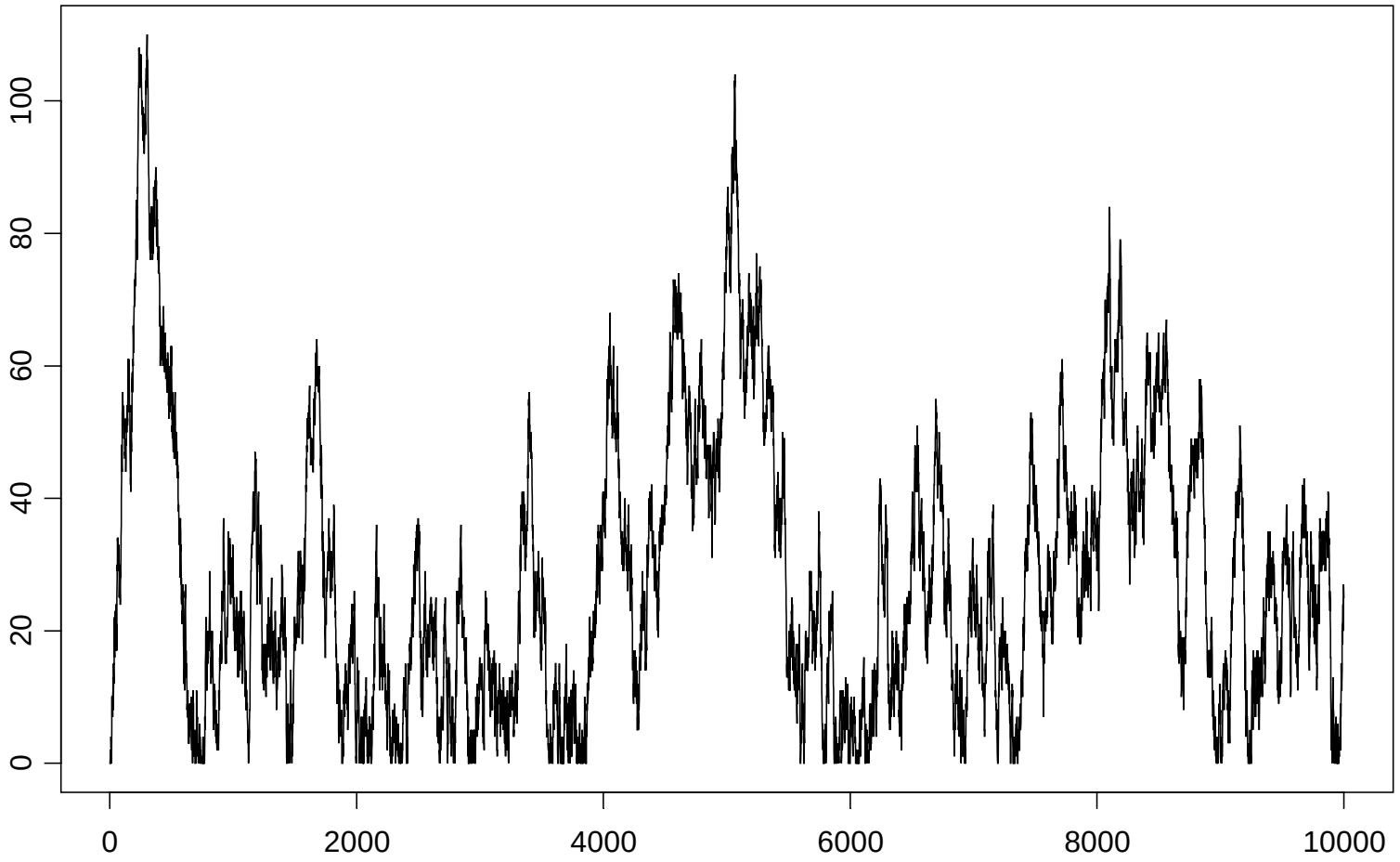
- 若服務每位顧客平均需時 μ 分鐘的指數分配，兩位顧客間的間隔時間為 λ 分鐘的指數分配；若 $\mu < \lambda$ 則等待服務的顧客隊伍才有可能消失，否則將愈排愈長。

兩組模擬比較：(利用率 = μ/λ)

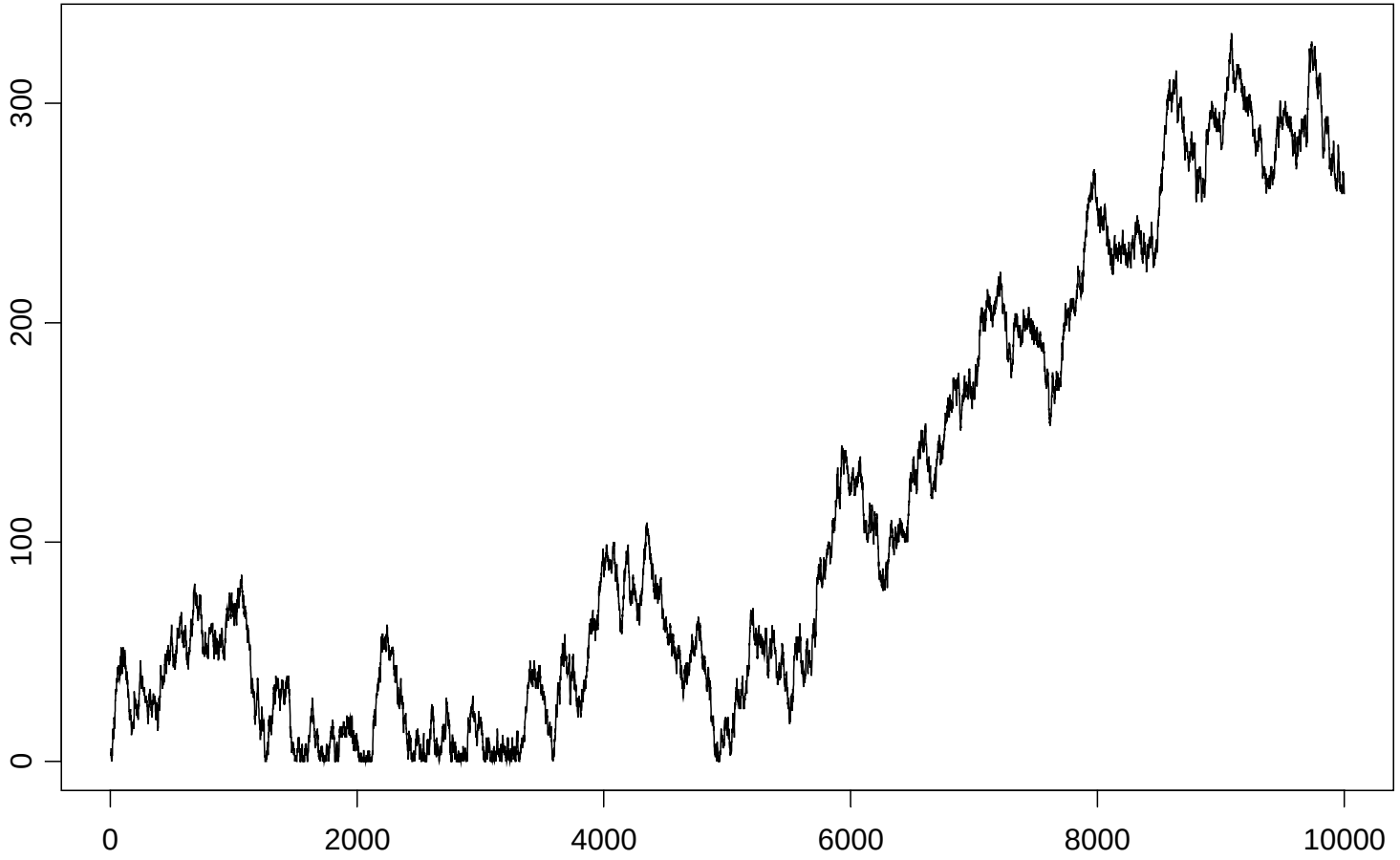
→ $\mu = 3.26 < \lambda = 3.27$；

→ $\mu = 3.26 > \lambda = 3.25$。

Time series plot when the interarrival time is bigger (3.27 v.s. 3.26)



Time series plot when the inter-arrival rate is smaller(3.25 v.s. 3.26)



介紹到此結束，你們
也不妨自己動手試試！