Ch01 機率模型

簡介

機率 (probability)

- 指隨機性與不確定性的研究
- 量測某件事情會發生的機會

- 機率的觀念是整個統計決策理論的基礎，利用機率才可以討論**不確定性**

例如

- “在年底前瓊平均指數極有可能會上漲”
- “現任者有50%的機會會尋求連任”
- “明年那門課可能至少會有一節課”
- “制定一些法令有助於盡快停止罷工”
- “預計會賣掉至少 20,000 張演唱會門票”

1.1 以數學模型做為分析和設計的工具

- **模型 (model)**
  就是實際情況的一種近似的呈現。
- 一個模型企圖要用一組簡單易懂的規則來解釋被觀察到的行為。
- 一個**數學模型**是一組有關於一個系統是如何運作的假設，這些假設是用數學關係式來描述的，關係式中包含有此系統的重要參數和變數。
- 為了要達到實用目的，模型必須要符合某特定情況的實際狀況。因此，一個模型的開發程序必須包含一連串的實驗和模型改良，如圖1.1所示。

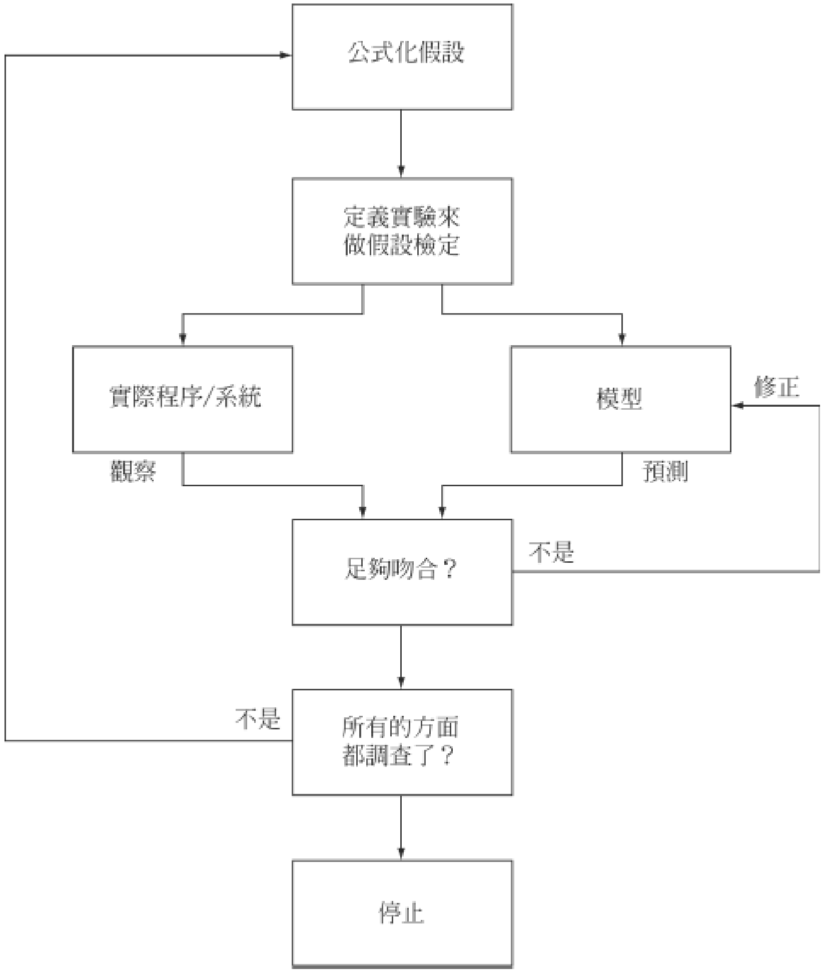


圖 1.1　模型建立過程。

1.2 確定模型

- 在**確定模型(deterministic models)**中，實驗的條件完全決定了實驗的**結果(outcome)**，一組數學方程式的解答說明了實驗的結果。
- 電路理論就是確定模型的例子。
- 例如，
  - ✓ 歐姆定律指出一個電阻的電壓電流關係為 $I = V/R$。
  - ✓ 如果測量一組電壓的實驗在相同條件下重複做了很多次，電路理論預測觀察的結果會完全一樣。

1.3 機率模型

- **實驗**(experiment) 是指任何一種行為或過程
- 隨機實驗
  - ✓ 當實驗在相同條件下重複執行，實驗結果有不可預測的變化。
- 確定模型不適用於模擬**隨機實驗**。
- 對隨機實驗需要設計**機率模型**。

1.3 機率模型

- 隨機實驗的例子：
  - ✓ 一個甕裝著三顆球，這三顆球分別標示著 0，1，2。從甕中隨機選出一顆球並記下號碼，然後把球放回甕裡。
  - ✓ 這個實驗的**結果(outcome)**是一個數，該數所有的可能值集合為 $S=\{0,1,2\}$。
  - ✓ 圖1.2為該甕實驗做100次的結果。我們可以看出這個實驗的結果是無法正確預測的。

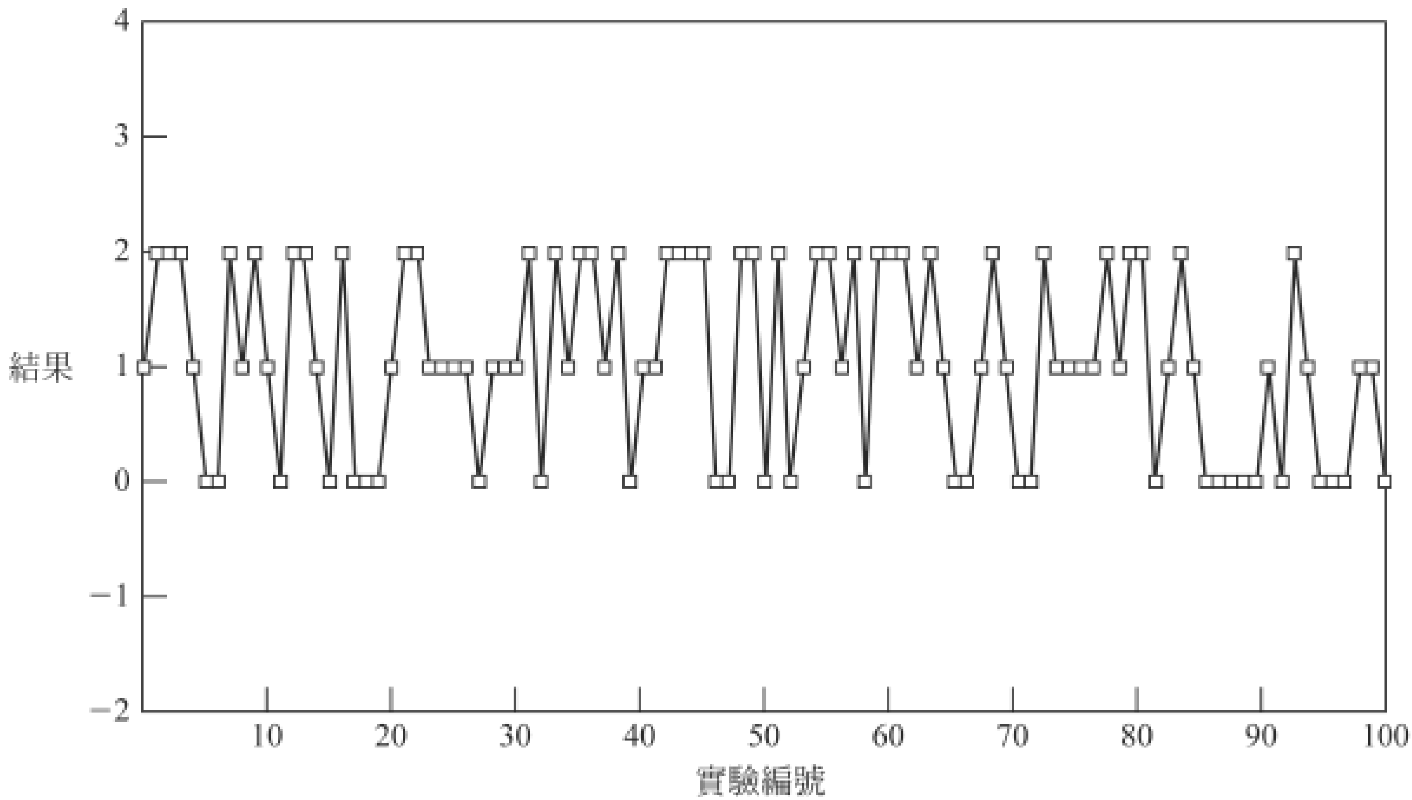


圖 1.2　甕實驗的結果。

- 集合S 稱為 **樣本空間(sample space)**。
- **樣本空間(sample space)**
  - 實驗中所有可能結果的集合

Example 1

最簡單的機率實驗是有兩種可能結果的實驗。

1.1
1.2
1.3
1.4
1.5

- 檢查一條保險絲是否有瑕疵， $\boldsymbol{S} = \{N, D\}$ ，
  - ✓ N:無缺陷，D:有缺陷
- 投擲一枚圖釘， $\boldsymbol{S} = \{U, D\}$ 。
- 觀察某間地區醫院下一個出生的小孩性別，
  $\boldsymbol{S} = \{M, F\}$。

Example 2

- 如果我們依序檢查三條保險絲並記錄每個檢查結果， 則整個實驗的任一結果是長度為3 且以 N 與 D 組合而成的序列， 因此

$$\mathscr{S} = \{NNN, NND, NDN, NDD, DNN, DND, DDN, DDD\}$$

- 如果我們投擲一枚圖釘三次， 其樣本空間只要將上述 $\mathscr{S}$ 裡的N 以U 來取代便可求得
- 以類似改變符號的作法也可以得到觀察三個新生兒性別實驗的樣本空間。

Example 3

在某一個交叉路口有兩座加油站。每一座加油站都有六部加油機器。

一實驗欲觀察在一天裡的某個特定時刻每座加油站正在使用的機器數目。

實驗結果指出在第一座加油站有幾部機器正在使用，且在第二座加油站正在使用的有幾部。可能結果為(2, 2)，另一為(4, 1)，再另一為(1, 4)。

附表列出 $\mathcal{S}$ 裡的49 種結果。投擲一顆六面骰子兩次的實驗的樣本空間會是表中刪除列0 與行0 之後的結果，共有36 種。

Example（續）

		第二座加油站						
		0	**1**	**2**	**3**	**4**	**5**	**6**
第一座加油站	**0**	(0, 0)	(0, 1)	(0, 2)	(0, 3)	(0, 4)	(0, 5)	(0, 6)
	1	(1, 0)	(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)	(1, 6)
	2	(2, 0)	(2, 1)	(2, 2)	(2, 3)	(2, 4)	(2, 5)	(2, 6)
	3	(3, 0)	(3, 1)	(3, 2)	(3, 3)	(3, 4)	(3, 5)	(3, 6)
	4	(4, 0)	(4, 1)	(4, 2)	(4, 3)	(4, 4)	(4, 5)	(4, 6)
	5	(5, 0)	(5, 1)	(5, 2)	(5, 3)	(5, 4)	(5, 5)	(5, 6)
	6	(6, 0)	(6, 1)	(6, 2)	(6, 3)	(6, 4)	(6, 5)	(6, 6)

Example 4

有一種新的D 型手電筒電池，其電壓未介於某個範圍內，則此電池被認定為失敗(F)；如果電池電壓在規定的範圍內，則認定為成功(S)。

假設實驗所測試剛從生產線產出的每一電池直到我們觀測到第一顆成功電池為止。雖然不太可能，此一實驗的可能結果為前10 (或100 或1000 或...) 顆都是 F，然後下一枚是S。

Example 4（續）

亦即，我們在遇到第一個 S 之前，可能必須要測試 n 顆電池，其中n 為任意正整數。

樣本空間 $\mathscr{S} = \{S, FS, FFS, FFFS, \ldots\}$ ，所包含的可能結果數量是無限的。

從某一時間開始，記錄每位新生兒的性別直到觀測到第一位男孩兒出生的實驗，其樣本空間也可以簡寫成上述形式。

1.3.1 統計的規律性

- 考慮實用性
  - 模型必須可以預測一個系統的未來行為
- 可預測
  - 一個現象必須要展示它的行為的規律性
- **統計的規律性(statistical regularity)**
  - 長期重複隨機實驗所得到的平均會一致性地產生出近似相同的值

1.3.1 統計的規律性

- 一個甕裝著三顆球，這三顆球分別標示著0 ， 1，2。從甕中隨機選出一顆球並記下號碼，然後把球放回甕裡。
- 在相同條件下重複 n 次。
- 令 $N_0(n)$，$N_1(n)$ 和 $N_2(n)$分別為0號球，1號球，和2號球出現的**次數**，並令 結果k 的**相對次數**為

$$f_k(n) = \frac{N_k(n)}{n} \tag{1.1}$$

- 統計的規律性
  - ✓ 指出當 n 的值變大時，以一個常數值為中心所做的變動愈來愈小，即

$$\lim_{n\to\infty} f_k(n) = p_k \tag{1.2}$$

    常數 p_k 就被稱為是結果k的**機率**。
- 式(1.2)說明
  - ✓ 一個結果的機率是多次重複的測試中次數的長期比例。
  - ✓ 式(1.2)提供了從實際量的測量到機率模型之間的關鍵連結。

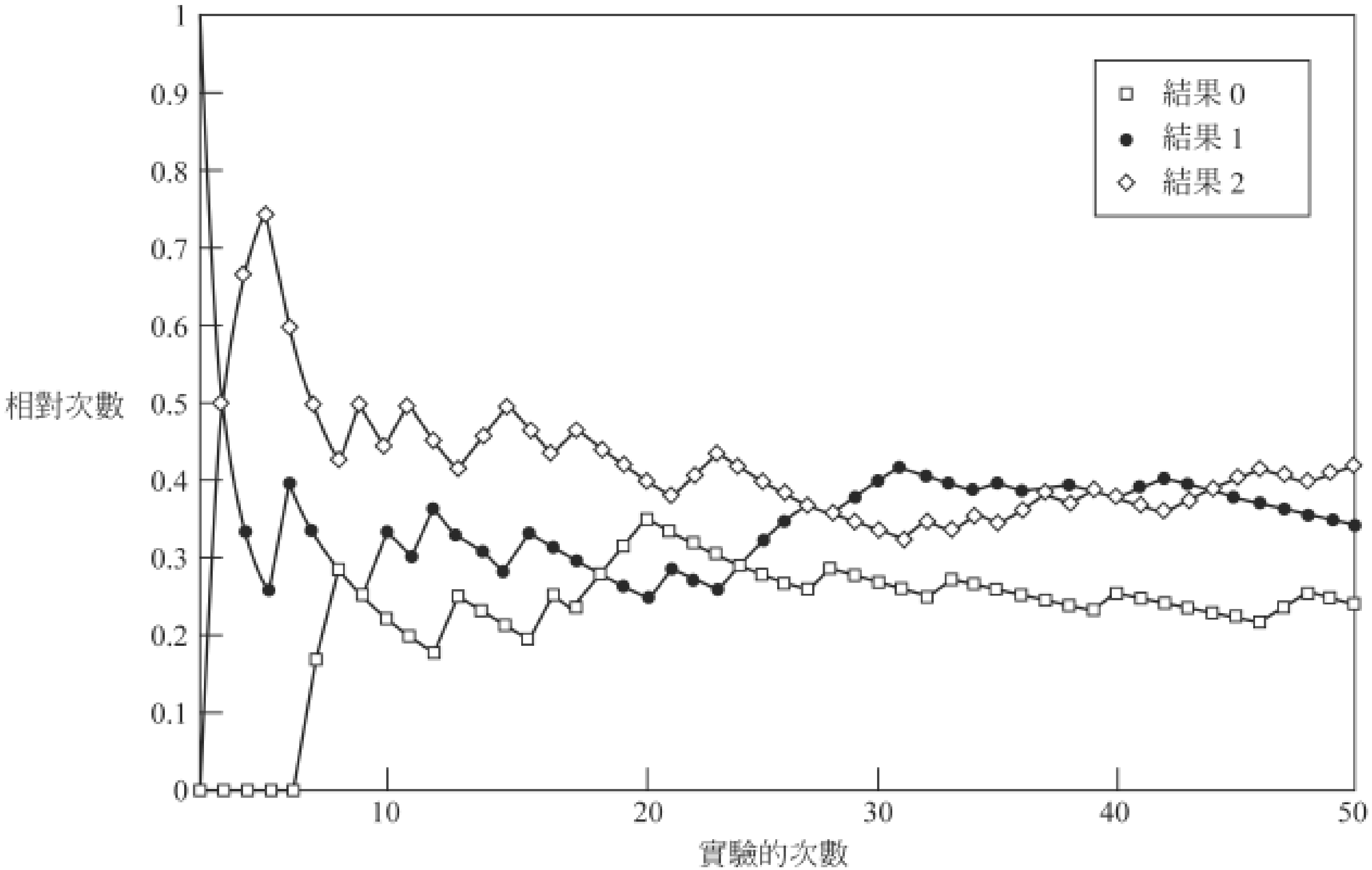


圖 1.3　在甕實驗中的相對次數。

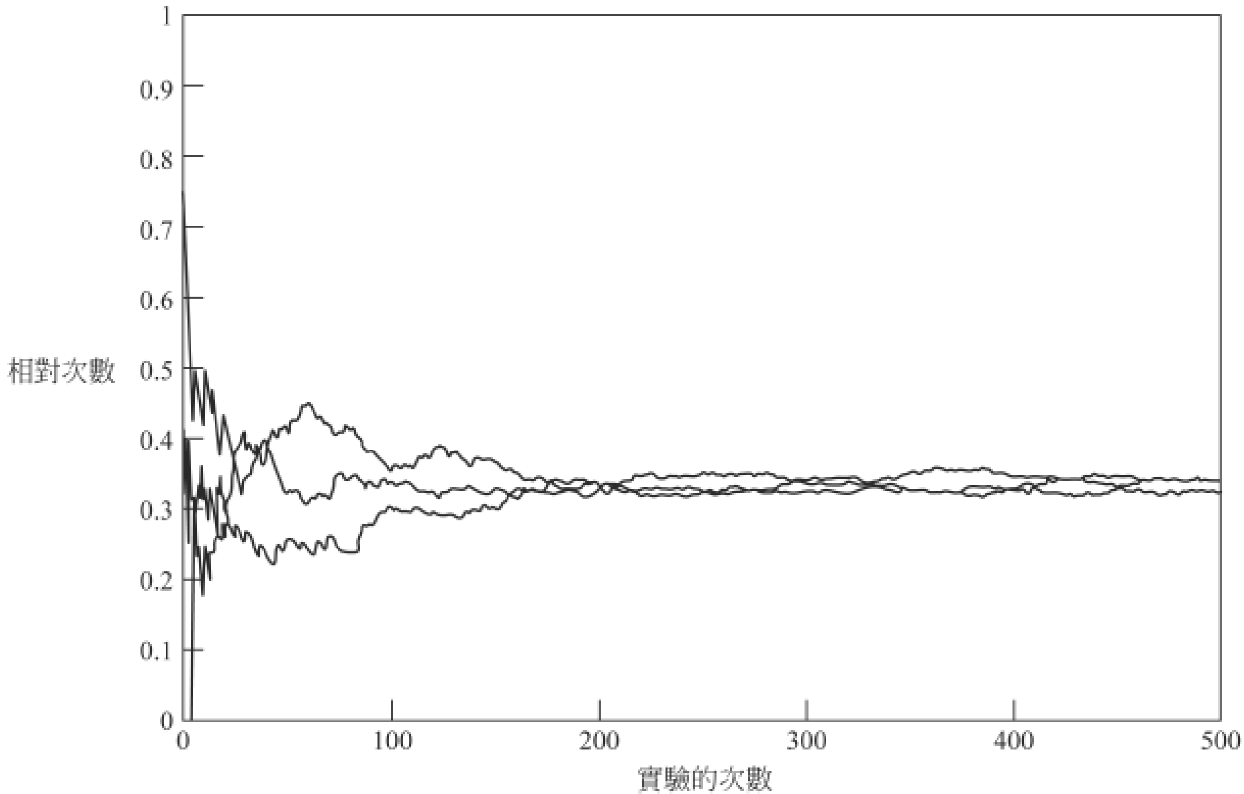


圖 1.4　在甕實驗中的相對次數。

- 若在甕中放入第四顆球，標示為0，則現在出現0號球的機率為2/4，出現1號球和2號球的機率也分別降至1/4。
- 這說明了機率模型的特性：**隨機實驗的條件決定了實驗結果的機率**。

1.3.2 相對次數的特性

- 假設一個隨機實驗有K個可能結果

$$S=\{1,2,\ldots,K\}$$

- 在n次實驗中，任何結果的出現

$$0\le N_k(n)\le n,\quad k=1,2,\ldots,K$$

$$\Rightarrow\quad 0\le \frac{N_k(n)}{n}\le 1,\quad k=1,2,\ldots,K$$

$$\Rightarrow\quad 0\le f_k(n)\le 1,\quad k=1,2,\ldots,K$$

- 所有可能結果的出現次數總和

$$\sum_{k=1}^{K} N_k(n) = n$$

$$\Rightarrow \sum_{k=1}^{K} \frac{N_k(n)}{n} = \frac{n}{n} = 1$$

$$\Rightarrow \sum_{k=1}^{K} f_k(n) = 1$$

- 事件 (event)
  - 任一包含在樣本空間中的集聚 (子集合)
  - **事件**是樣本空間的部分集
  - 簡單(simple)事件，其剛好只包含一個結果
  - 複合(compound)事件，則其包含一個以上的結果

範例 2.5

- 考慮在某個高速公路出口斜坡下記錄每三輛車向左轉(L) 或向右轉(R) 的實驗。
- 樣本空間是由八種可能的結果構成，即 LLL, RLL, LRL, LLR, LRR, RLR, RRL, 以及 RRR。
- 因此有八個簡單事件，其中有$E_1 = \{LLL\}$且$E_5 = \{LRR\}$。

範例 2.5（續）

- 一些複合事件，包括

$A = \{RLL, LRL, LLR\}$ = 三輛車中剛好有一輛車向右轉的事件

$B = \{LLL, RLL, LRL, LLR\}$ = 最多有一輛車向右轉的事件

$C = \{LLL, RRR\}$ = 三輛車全部都轉向同一個方向的事件

- 假設進行該實驗時，結果是 LLL。
  - 則發生了簡單事件 E_1，亦發生 B 和 C 事件(但 A 未發生)。

範例 6 (承範例3)

當觀察每一個有六部加油機器的加油站正在使用的機器數目時，會有49 種可能(承範例可能的結果，所以有49 個簡單事件：

$$E_1 = \{(0, 0)\}, E_2 = \{(0, 1)\}, \ldots E_{49} = \{(6, 6)\}$$ 。

範例 6（續）(承範例3)

複合事件的例子有

$A = \{(0, 0), (1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6)\}$

= 兩座加油站正在使用機器數目相同的事件

$B = \{(0, 4), (1, 3), (2, 2), (3, 1), (4, 0)\}$

= 共有四部機器正在使用的事件

$C = \{(0, 0), (0, 1), (1, 0), (1, 1)\}$

= 每座加油站至多有一部機器正在使用的事件

範例 7 (承範例4)

檢測電池實驗的樣本空間包含有無限多的結果，所以有無限個簡單事件。

複合事件包括

$A = \{S, FS, FFS\}$

= 最多檢測三顆電池的事件

$E = \{FS, FFFS, FFFFFS, \ldots\}$

= 檢查偶數顆電池的事件

1.3.3 機率理論的公設

$$\lim_{n\to\infty} f_k(n) = p_k$$

- 以長期相對次數來定義一個事件的機率會發生的問題：
  - ✓ 1.不知道在式(1.2)中的極限何時存在以及是以何種數學的意義存在
  - ✓ 2. 我們無法執行一個實驗無限多次，所以我們不可能知道精確的p_k值
  - ✓ 3. 用相對次數來定義機率會排除一些我們無法做重複實驗的情況
- 現代的機率理論始於建立一組公設，規定了機率的指派必須要滿足這組公設

1.3.3 機率理論的公設

- 對每一事件A的機率$P[A]$：

1. $0 \le P[A] \le 1$。
2. $P[S]=1$。
3. 若 A 和 B 是兩個不能同時發生的事件，則 $P[A$ 或 $B] = P[A] + P[B]$。

- 機率的理論並不關注於機率是如何獲得的，或者它的意義是什麼。
- **任何滿足上述公設的機率分配就是合法的機率。**
- 我們把權力留給理論的使用者和模型的建構者來決定機率指派應該怎麼做，以及在任何特定的應用中決定機率應該如何做解讀才合理。

機率公理

- 已知一實驗之樣本空間 $\mathcal{S}$
- 機率的目的
  - 在分配每個事件 A 一個數值 $P(A)$， 稱作 A 事件的機率

機率公理（續）

公理1　對所有事件 A，$P(A) \geq 0$。

公理2　$P(\mathcal{S}) = 1$。

公理3　如果 $A_1, A_2, A_3, \ldots$ 為無限個互不相交的事件集合，則

$$P(A_1 \cup A_2 \cup A_3 \cup \cdots) = \sum_{i=1}^{\infty} P(A_i)$$

範例 11

- 考慮往空中丟擲一枚圖釘。
- 當它掉到地上時，針頭可能會朝上(結果 U)或朝下(結果D)。
- 這個事件的樣本空間為$\mathcal{S} = \{N, D\}$。
- 公理指出$P(\mathcal{S}) = 1$，所以決定了 $P(U)$ 與 $P(D)$ 就完成機率分配工作。
- 由於U 與D 不相交，$1 = P(\mathcal{S}) = P(U) + P(D)$，故$P(D) = 1 - P(U)$

1.1
1.2
1.3
1.4
1.5

範例 11（續）

- 一個可能的機率分配方式為$P(U) = 0.5$，$P(D) = 0.5$ ，
- 另一可能的機率分配方式為$P(U) = 0.75$，$P(D) = 0.25$ 。
- 事實上，令p 表示0 與1 之間任何一個固定的數值，$P(U) = p$，$P(D) = 1 - p$ 為符合公理的一種分配方式。

範例 12

範例4 中的實驗，一個接著一個地測試剛從生產線產出的電池，直到有一個電池的電壓在指定範圍內為止。

簡單事件為$E_1 = \{S\}, E_2 = \{FS\}, E_3 = \{FFS\}, E_4 = \{FFFS\}, \ldots$ 。

範例 12（續）

若任何一顆電池合格的機率是0.99 。則 $P(E_1) = 0.99$, $P(E_2) = (0.01)(0.99)$, $P(E_3) = (0.01)^2(0.99)$, ... 是簡單事件的機率分配值，其滿足公理。

尤其，因為E_i 不相交，且 $\mathscr{S} = E_1 \cup E_2 \cup E_3 \cup \ldots$ ，則必然有

$$
\begin{aligned}
1 = P(\mathscr{S}) &= P(E_1) + P(E_2) + P(E_3) + \cdots \\
&= 0.99[1 + 0.01 + (0.01)^2 + (0.01)^3 + \cdots]
\end{aligned}
$$

範例 2.12（續）

此處我們已經使用將幾何級數加總的公式：

$$a + ar + ar^2 + ar^3 + \cdots = \frac{a}{1-r}$$

然而，另一種合法的(依據公理) “幾何” 型態的機率分配可藉由將 0.99 以另一個0 與1 之間的數值 p 來取代 (0.01 以1 - p 取代) 而求得。

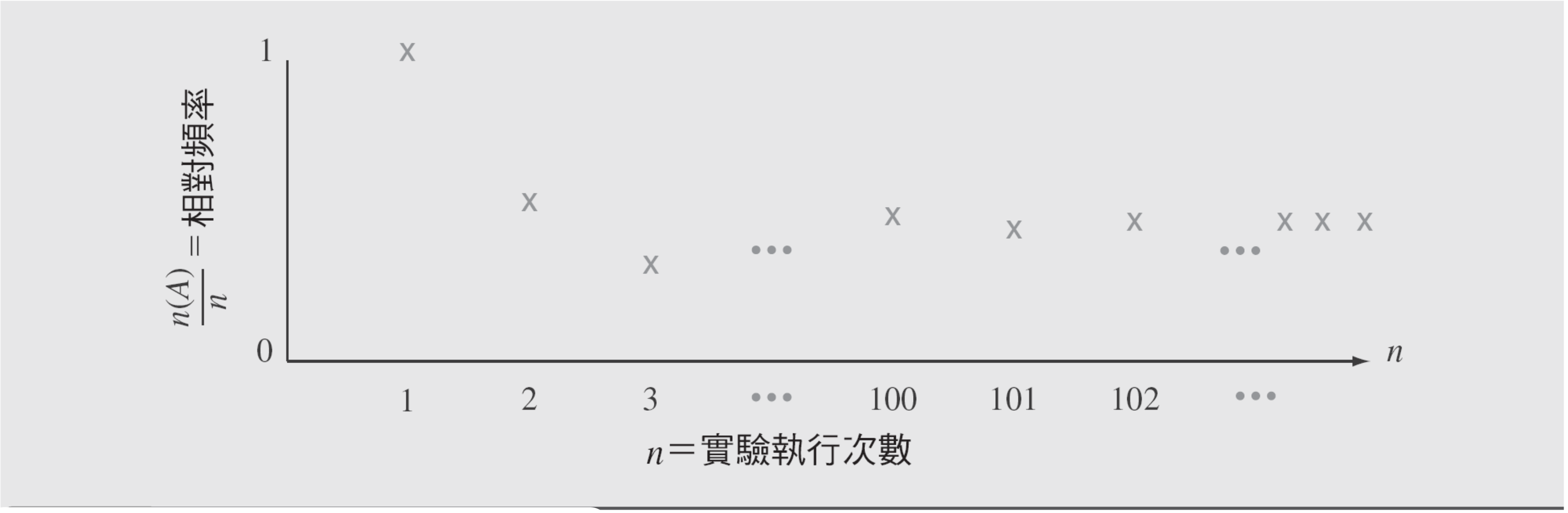
詮釋機率

最常使用且最容易懂的詮釋方法

- 以相對頻率的概念為基礎。

詮釋機率（續）

- 考慮一可以不斷地重複執行相同且獨立的實驗，並令 A 事件為一包含此實驗特定結果(出象)的集合。若這樣的實驗重複執行 n 次，令 $n(A)$ 表示重複實驗 A 發生的次數。
- 當 n 愈大時，相對頻率 $n(A)/n$ 愈穩定；當 n 增至任意大時，相對頻率會趨近一個極限值，我們稱為 A 事件的極限相對頻率，以 $P(A)$ 表示。


1
0
$\frac{n(A)}{n}$ = 相對頻率
1
2
3
100
101
102
n
n＝實驗執行次數

詮釋機率（續）

- 利用相對頻率來詮釋機率可是說客觀的，因為其係取決於實驗的特性而非與參與實驗的特定個人。
- 由於機率的客觀詮釋是基於極限頻率的概念，其應用範圍侷限於可重複的實驗。
- 語言常使用在本質上不能重複的情況。
  - 例子包括：“和平協議的機會很高”；“我們公司很有可能會取得合約”。

詮釋機率（續）

因為不同的觀察者可能會有與這種實驗情況相關的不同前景資訊與意見，機率分配可能會因人而異。在這種情況下的詮釋因此稱作是主觀的。

1.3.4 建立一個機率模型

- 如何將現實的隨機問題轉換成**機率模型 (probability model)**？
- 我們需要指出：
  (1)定義應用本身的隨機實驗
  (2)集合S以及我們所感興趣的事件
  (3)指出機率指派方式。
- 具挑戰之處在於我們所發展出來的模型必須能解釋現實問題的相關事宜。

1.4 一個詳細的例子：封包式語音傳輸系統

- 一個通訊系統必須同時將48個對話以語音資訊封包從*A*點傳到*B*點。每一個通話者所說的話會轉換成語音資訊封包。在語音封包傳輸之前，它的來源和目的地地址會附在封包上(見圖1.5)。

1.1
1.2
1.3
1.4
1.5

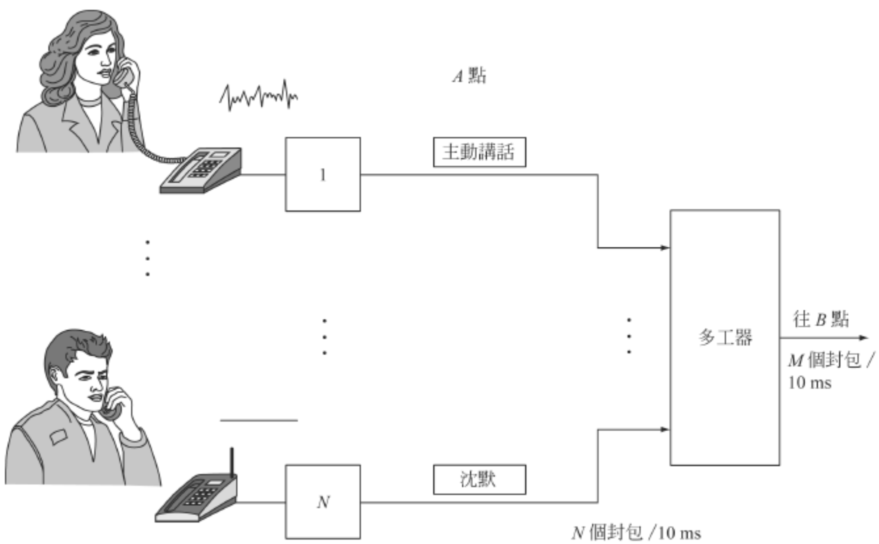


圖 1.5 一個封包式語音傳輸系統。

- 最簡單設計
  - ✓ 是每10毫秒就往每個方向傳輸48個封包。
  - ✓ 但這不有效。
  - ✓ 因為根據經驗，平均約有 2/3 的封包是沉默的。故每10毫秒48個通話者只產生約 48/ 3 = 16 個講話(非沉默)的封包。
  - ✓ 因此，考慮另一個系統，每10毫秒只傳輸$M < 48$個封包。

1.1
1.2
1.3
1.4
1.5

- 每10毫秒，這個新系統先判定那些通話者產生講話封包。
- 令產生的講話封包數為 A。
- A 可能是介於 0 (所有的通話者都沉默) 到 48 (所有的通話者都講話)。
- 如果 $A \le M$ ，則全部的講話封包都被傳輸。

但是如果$A > M$，此系統無法傳輸全部的講話封包，以致於有 $(A\text{-}M)$ 個講話封包會捨棄而導致言談資訊的消失。

所以必須控制捨棄講話封包的比例維持在一個水準，使得通話者不會感覺到怪怪的。

- 上述的實驗重複 n 次，令$A(j)$是第 j 次試驗的結果。令$N_k(n)$為主動講話封包數是k 的試驗次數。在前n 次試驗中，結果k 的相對次數 $f_k(n) = N_k(n)/n$，我們假設它收斂到機率 p_k：

$$\lim_{n\to\infty} f_k(n) = p_k, \quad 0 \le k \le 48 \tag{1.6}$$

- 圖1.6指出 p_k 和 k的關係。我們可以看到正在講話的通話者最常出現的次數是16，正在講話的通話者次數很少高於24。

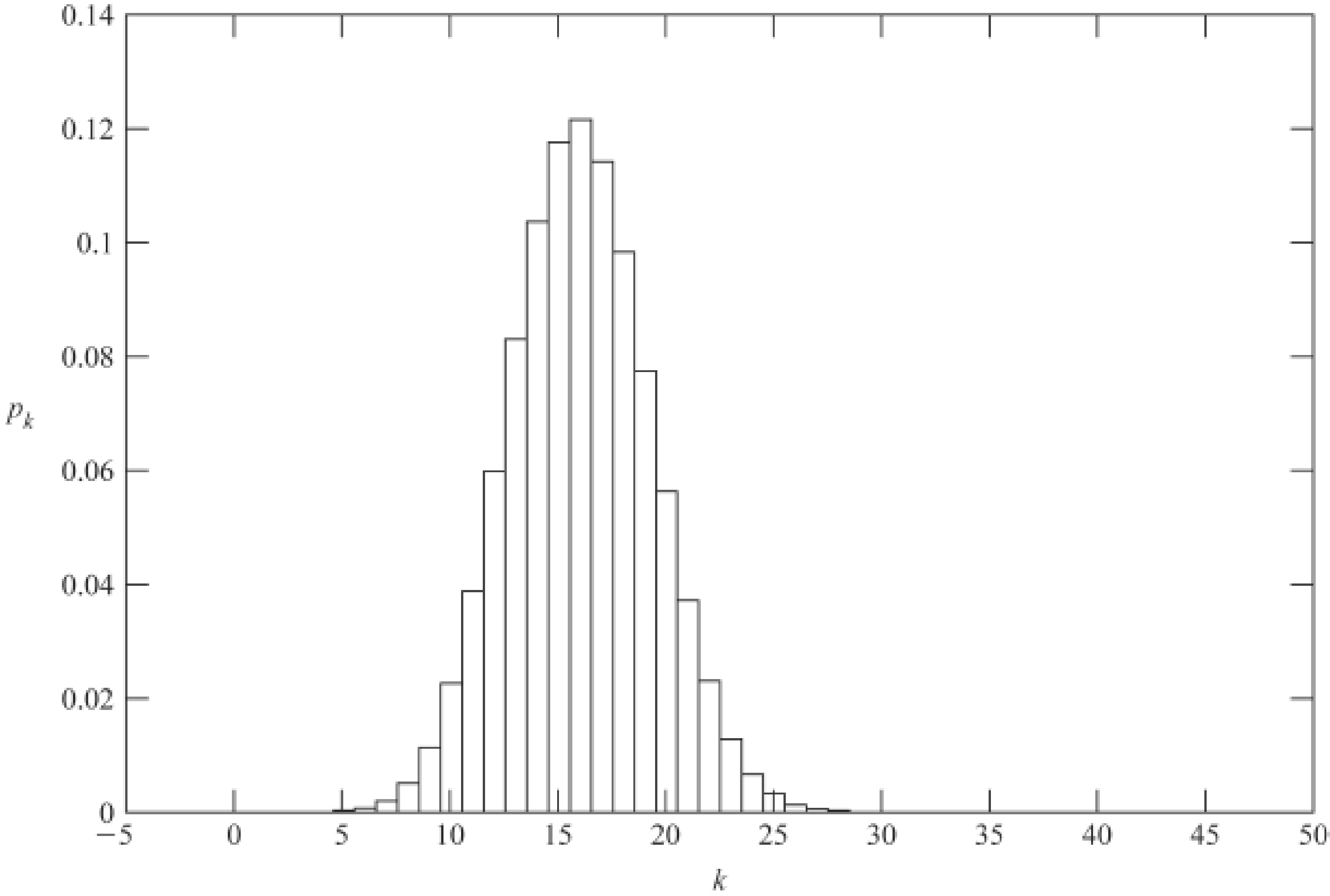


圖 1.6　一組 48 位通話者中有 *k* 位通話者正在講話的機率。

- 每10毫秒中所產生的主動講話封包平均數是由主動講話封包樣本數的**樣本平均值(sample mean)**所給定：

$$\langle A \rangle_n = \frac{1}{n}\sum_{j=1}^{n} A(j) = \frac{1}{n}\sum_{k=0}^{48} k \cdot N_k(n) \qquad (1.8)$$

- $A(j)$: 在前n次試驗中，主動講話封包的數量依據觀察順序做個加總
- $N_k(n)$: 先計數在試驗中有多少個實驗是有 k 個主動講話封包的，對每一個可能的 k 值都做計數後，做紀錄後計算總合

當n變大時，在第二個表示式中的比值$N_k(n)/n$ 會趨近於 p_k 。

因此，每10毫秒中所產生的主動講話封包平均數會趨近於

$$\langle A \rangle_n \to \sum_{k=0}^{48} k \cdot p_k(n) \triangleq E[A] \tag{1.9}$$

- 右邊的表示式定義為**A的期望值(expected value of A)**。$E[A]$完全是由機率p_k所決定，而且我們將在第三章證明
  $E[A] = 48 * 1/ 3 = 16$
- 式(1.9)指出每10毫秒所產生的主動講話封包之長期平均數為每10毫秒$E[A] = 16$位通話者。

由機率p_k所提供的資訊使我們可以設計既有效率又有好音質的系統。

例如，我們可以在每10毫秒內把傳輸量減少一半只傳24個封包，而且捨棄的主動講話封包數可以控制在無法察覺的範圍內。

1.5 其他的例子

1.5.1 在不可靠通道中的通訊

- 每T秒，發射器接受一個0或一個1，並且傳輸一個相對應的訊號。在T秒的最後，接收器依據它所接收到的訊號判定輸入是什麼。接收器的判定未必和發射器的輸入一樣。
- 在圖1.7(a)的通道機率模型中，傳輸錯誤是隨機發生且發生機率為ε 。 ε是接收端位元發生錯誤的長期比例。
- 若ε 是不可以被接受的，則必須使用錯誤控制的技術

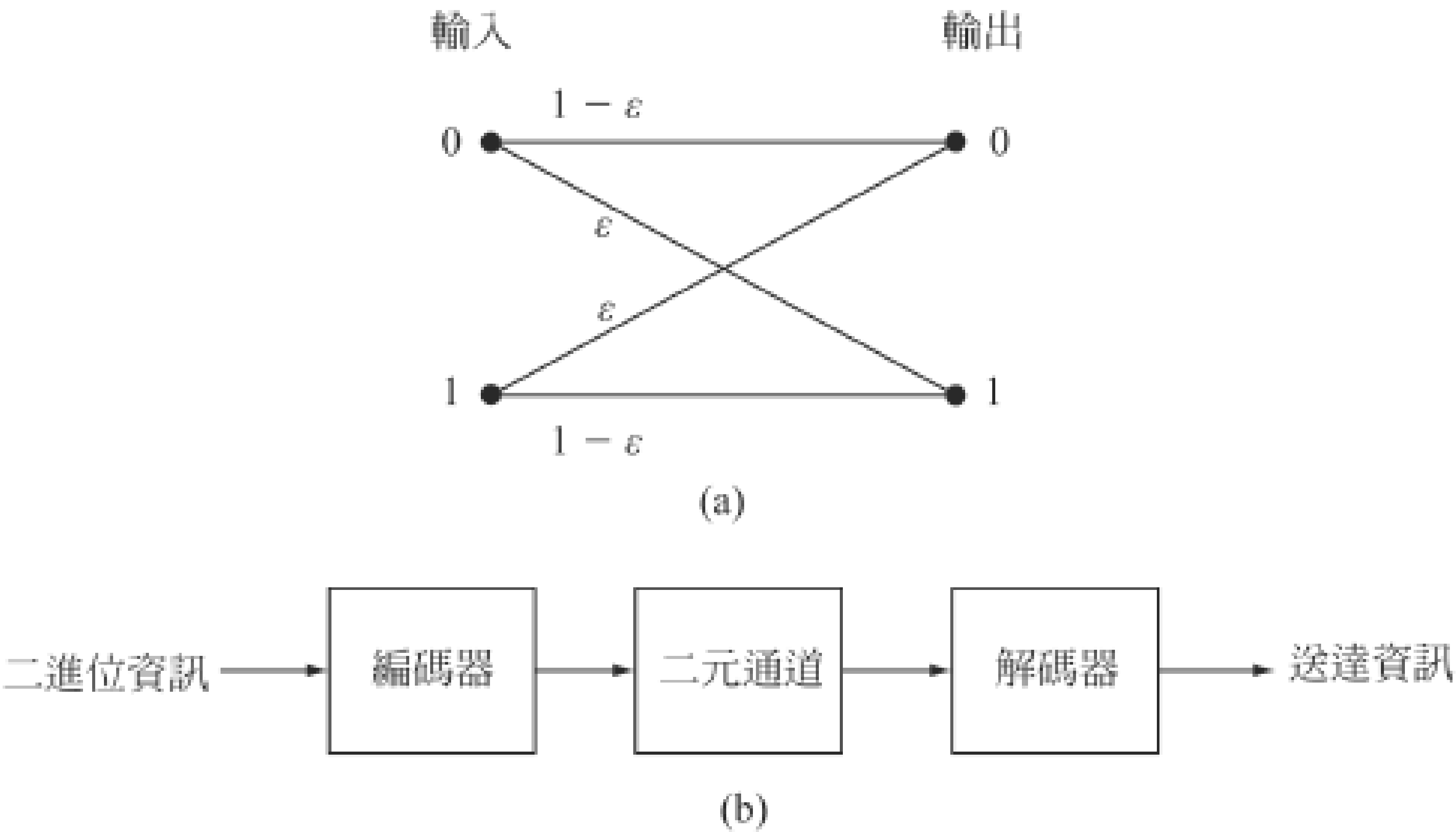


圖 1.7 (a) 二元通訊系統通道的一個模型。(b) 錯誤控制系統。

- 考慮用重複傳送碼以減少傳送錯誤發生率，其中每個資訊位元被傳輸3次：

$$0 \rightarrow 000$$
$$1 \rightarrow 111$$

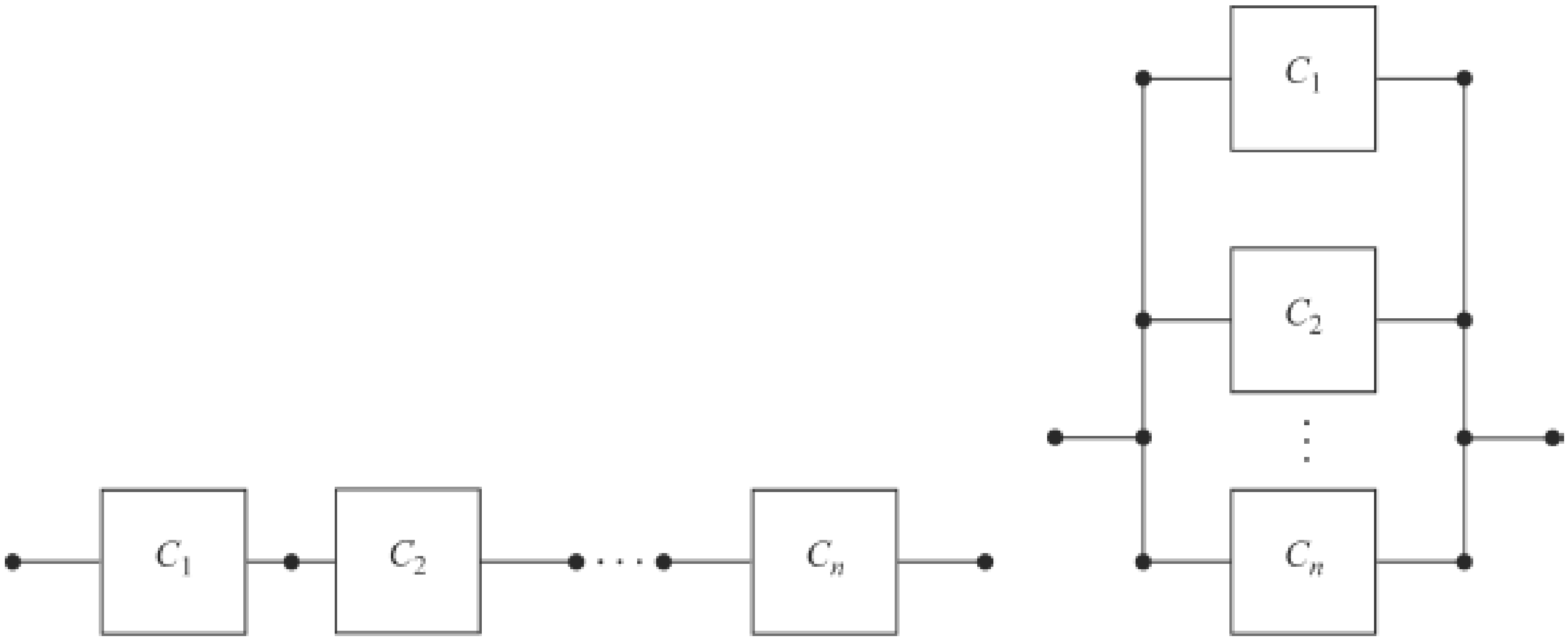
假設解碼器做判定的方式是採用接收端接收到3個位元中的多數決，那麼只有當其中出現2個或3個位元錯誤時，解碼器才會做出錯誤的判定。這個解碼器會做出錯誤判定的機率為$3(1-\varepsilon)\varepsilon^2+\varepsilon^3$。因此，若在沒有編碼的情況下這個通道的位元錯誤率為10^{-3}，那麼經過上述之重複傳送編碼解碼過程，位元錯誤率將會是$3*10^{-6}$。然而，這個好處不是沒有成本的：傳輸率下降至每$3T$秒傳送1個位元。

1.5.2 信號的壓縮

- 考慮壓縮一個音樂信號$S(t)$。壓縮技術是用預測的方式以提供有效率的表現方式，其中的下一個信號的值是使用過去的編碼值來做預測。只有在預測中的錯誤才需要被編碼，所以位元數可以降低。為了要讓這個預測系統可以運作，我們需要知道信號彼此之間是如何相關的。假如我們知道這個相關結構，我們就可以設計出最佳的預測系統。機率在解決這些問題中扮演了關鍵的角色。

1.5.3 系統的可靠度

- 如何從不可靠的元件中建立可信賴的系統？機率模型使我們可以用定量分析的方式來對付這個問題。 圖1.8(a)顯示出當全部元件都可運作時，系統才能運作，而圖1.8(b)顯示出只要有一個元件可以正常運作，系統就能運作。結合這兩種基本的配置就可以得到更複雜的系統。
- 機率理論使我們可以根據一個系統中個別元件的機率和平均來決定整個系統的機率和平均。這使得我們可以根據它們的可靠度來評估系統配置，因而選擇可靠的系統設計。



(a) 元件的串聯配置。　(b) 元件的並聯配置。

圖 1.8　有 n 個元件的系統。

1.5.4 資源分享的系統

- 許多應用程式牽涉到分享不穩定且需求是隨機的資源。需用動態資源分享來滿足客戶端的需求。許多Web伺服器系統運作方式如圖1.9所示。在任何給定時間點上，這些系統允許*c*個客戶端可以連接至伺服器。客戶端提交查詢給伺服器。查詢被放置在等待佇列中，然後由伺服器處理。在接收到伺服器的回應之後，每個客戶端會花一些時間思考，才再提出下一個查詢。在一段暫停週期過後，這個系統會終止一個正在使用的客戶端連線，另外開啟和另一個新的客戶端的連線。

- 系統需要被配置來提供快速的回應給客戶端，以避免提早終止連線，並可有效利用電腦運算的資源。這需要對以下的量做機率的描述：查詢處理時間，每次連線客戶端按下滑鼠左鍵的次數，和在按下兩次滑鼠左鍵之間的時間(思考時間)。有了這些參數之後，系統就可以根據它們決定出最佳的*c*值以及暫停週期時間。

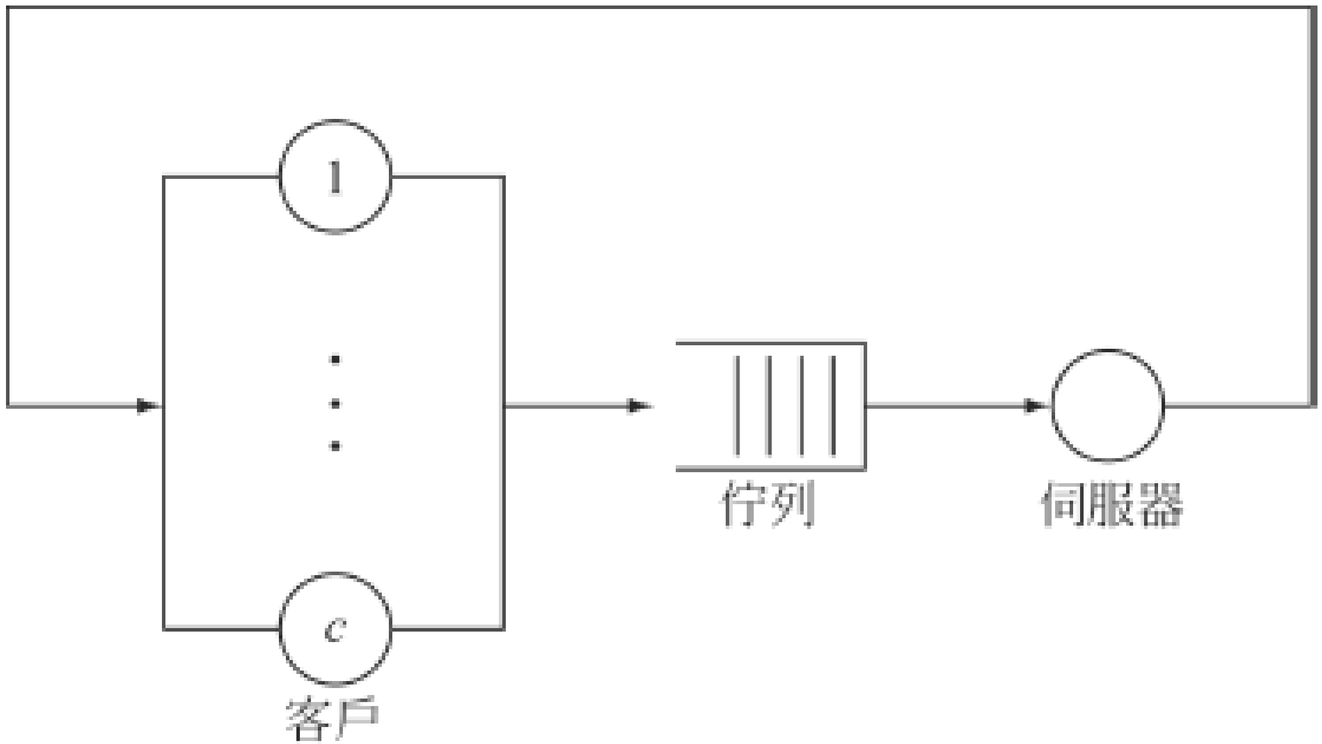


圖 1.9　Web 伺服器系統的簡易模型。

1.5.5 具Internet大小的系統

- 請參考圖1.10。機率的技術可以評估在Internet系統中節點的相對重要性，機率在搜尋引擎中扮演中核心的角色。點對點的(P2P)檔案分享和內容散佈，產生一些新的群體，與在發展可靠且可預測的管理與控制資源的方法上，機率方法扮演了重要的角色。

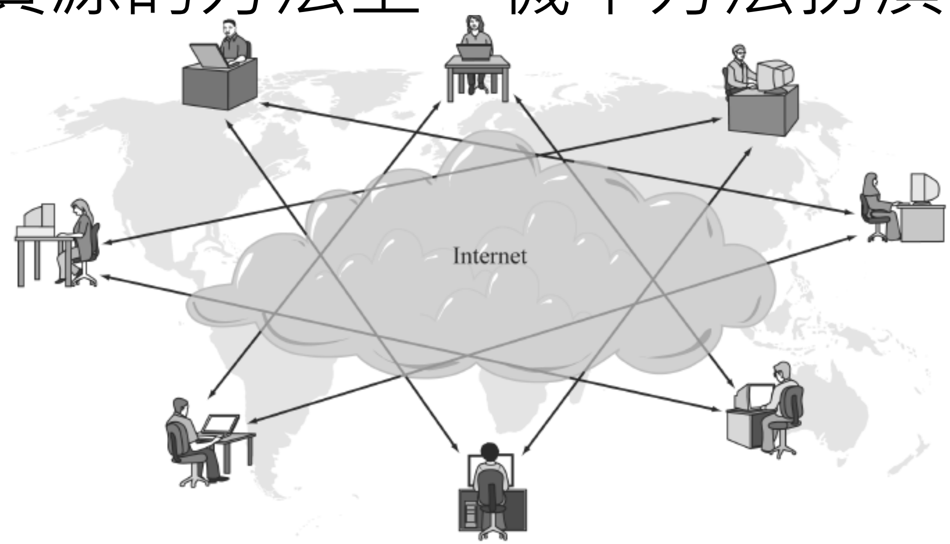


圖 1.10　一大群組的使用者橫越過 Internet 做互動。