

第十章

離散資料分析

10.1 前言

在第八章我們所討論的統計假設檢定，都是針對母體的參數。譬如說已知母體有 $\mathcal{N}(\mu, \sigma^2)$ 分佈，且 σ^2 也已知，欲檢定 $H_0: \mu = 3$ 。另外，在前幾章，我們屢屢說一隨機變數 X 有某一分佈。你可能會好奇，此假設要如何檢驗？譬如說由所收集到的資料，或由過去的經驗，或由已知母體的一些特性，猜測母體應有某一分佈，有時有必要先檢定此猜測是否正確。本章便要介紹一有用且簡易的方法。

雖然連續型的分佈，是一類重要且不可缺少的機率分佈，在本章中，我們將主要探討離散型的機率模型。對連續型的分佈，於將其離散化後，則也可適用針對離散型資料所發展出的方法。

10.2 適合度檢定

對一組觀測到的隨機樣本，有時我們會想檢定它們是否來自某一特定的母體。底下為一些可能的情況。

(i) 假設有一可產生隨機數(random numbers)的程式, $0, 1, \dots, 9$, 每個數字應有 $1/10$ 的機率出現。現欲檢驗此程式是否可靠, 我們讓它獨立地產生 n 個取值在 $0, 1, \dots, 9$ 的數, 得 $X_1, \dots, X_n$ 。根據此組觀測值, 以檢定 $H_0: 0, 1, \dots, 9$ 有相同的出現機率, vs. $H_a: 0, 1, \dots, 9$ 出現機率並非皆相同。

(ii) 在Becker and Piper(1982)一書中, 曾根據報紙及小說中的100,362個單字, 統計出英文26個字母出現之相對頻率, 而得表2.1。現對某一作家, 我們想檢定他某本書中各字母出現的頻率與表2.1所給是否相同。

表2.1 英文字母出現之相對頻率(百分比)

字母	相對頻率	字母	相對頻率
a	8.2	n	6.7
b	1.5	o	7.5
c	2.8	p	1.9
d	4.3	q	0.1
e	12.7	r	6.0
f	2.2	s	6.3
g	2.0	t	9.1
h	6.1	u	2.8
i	7.0	v	1.0
j	0.2	w	2.4
k	0.8	x	0.2
l	4.0	y	2.0
m	2.4	z	0.1

(iii) 台北銀行發行樂透彩券。每期自42個彩球中, 產生6個頭獎號碼及一特別號。自第3期至第7期, 39號球連續出現5次(有時在頭獎號碼中, 有時在特別號中)。民眾懷疑究竟是不是39號球的重量與眾不同, 使它較易出現。爲了徵信, 在第7期開獎前, 還公開在觀眾前測試, 將彩球取出秤重, 以確定彩球重量皆在標準容許範圍內。由於每期共產生7個號碼, 每

期39號球會出現之機率為 $7/42 = 1/6$ 。如何在夠多的期數後，以檢定 H_0 ：39號球出現的機率為 $1/6$ ，vs. H_a ：39號球出現的機率不為 $1/6$ ？

在上述這些情況，都可以以**多項分佈** $\mathcal{M}(n, \theta_1, \dots, \theta_k)$ 來描述： $A_1, \dots, A_k$ 為可能的結果（在(i)中為 $0, 1, \dots, 9$ ，在(ii)中為 $a, b, \dots, z$ ，在(iii)中為39與非39），發生的機率為 $\theta_1, \dots, \theta_k$ ，其中 $\theta_i = P(A_i)$ ， $i = 1, \dots, k$ 皆為未知。我們也可想成有一組樣本數為 n 之 k 維隨機向量 $\mathbf{X}_1, \dots, \mathbf{X}_n$ ，以 $\mathcal{M}(1, \theta_1, \dots, \theta_k)$ 為其共同分佈，基於所觀測到的 $\mathbf{X}_1, \dots, \mathbf{X}_n$ ，我們想檢定 H_0 ： $\theta_i = \theta_{i0}$ ， $i = 1, \dots, k$ ，vs. H_a ：在 $i = 1, \dots, k$ 中，有某個 i ， $\theta_i \neq \theta_{i0}$ 。例如，在(i)中 A_i 為數字 $i - 1$ 出現的事件， θ_i 為數字 $i - 1$ 出現的機率， $\theta_{i0} = 1/10$ ， $i = 1, \dots, 10$ 。

現今 N_i 表事件 A_i 出現的次數，又稱**觀測頻率**(observed frequency)， $i = 1, \dots, k$ 。 N_i 即 $\mathbf{X}_1 + \dots + \mathbf{X}_n$ 之第 i 分量。則 $\mathbf{N} = (N_1, \dots, N_k)$ 為 $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)$ 之一充分統計量。我們有興趣的問題即轉為基於所觀測到的有 $\mathcal{M}(n, \boldsymbol{\theta})$ 分佈之隨機向量，欲檢定

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 = (\theta_{10}, \dots, \theta_{k0}),$$

$$H_a : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0,$$

其中 $\boldsymbol{\theta} \in \Omega = \{\boldsymbol{\theta} : \theta_i \geq 0, i = 1, \dots, k, \sum_{i=1}^k \theta_i = 1\}$ 。

底下我們來導出**概似比檢定**。

首先， $\mathbf{N} = (N_1, \dots, N_k)$ 之概似函數為

$$L(\boldsymbol{\theta} | n_1, \dots, n_k) = f(n_1, \dots, n_k | \boldsymbol{\theta}) = \frac{n!}{n_1! \dots n_k!} \theta_1^{n_1} \dots \theta_k^{n_k}, \boldsymbol{\theta} \in \Omega。$$

又

$$\sup_{\boldsymbol{\theta} \in \Omega} L(\boldsymbol{\theta} | n_1, \dots, n_k) = L(\hat{\boldsymbol{\theta}} | n_1, \dots, n_k) = f(n_1, \dots, n_k | \hat{\boldsymbol{\theta}}),$$

其中 $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \dots, \hat{\theta}_k)$ 為 $\boldsymbol{\theta}$ 之MLE。由於

$$(2.1) \quad \theta_k = 1 - \sum_{i=1}^{k-1} \theta_i,$$

所以事實上 $L(\boldsymbol{\theta}|n_1, \dots, n_k)$ 可視為 $-\theta_1, \dots, \theta_{k-1}$ 之函數。參數 $(\theta_1, \dots, \theta_{k-1}) \in \{\theta_i \geq 0, i = 1, \dots, k-1, \sum_{i=1}^{k-1} \theta_i \leq 1\}$ 。

任一 MLE $\hat{\boldsymbol{\theta}}$ 滿足下述概似方程式：

$$(2.2) \quad \sum_{j=1}^k \frac{\partial}{\partial \theta_i} \log \theta_j^{n_j} = \sum_{j=1}^k \frac{n_j}{\theta_j} \frac{\partial \theta_j}{\partial \theta_i} = 0, i = 1, \dots, k-1。$$

由(2.1)式，對 $i = 1, \dots, k-1$ ， $\partial \theta_k / \partial \theta_i = -1$ ， $\partial \theta_i / \partial \theta_i = 1$ ，又 $\partial \theta_j / \partial \theta_i = 0$ ， $j = 1, \dots, k-1$ ，且 $j \neq i$ 。故(2.2)式成為

$$-\frac{n_k}{\theta_k} + \frac{n_i}{\theta_i} = 0, i = 1, \dots, k-1。$$

因此

$$(2.3) \quad \hat{\theta}_i = \frac{n_i}{n_k} \hat{\theta}_k, i = 1, \dots, k-1。$$

又 $\sum_{i=1}^k \hat{\theta}_i = 1$ ， $\sum_{i=1}^k n_i = n$ ，故(2.3)式導致

$$(2.4) \quad 1 = \sum_{i=1}^k \hat{\theta}_i = \frac{\sum_{i=1}^k n_i}{n_k} \hat{\theta}_k = \frac{n}{n_k} \hat{\theta}_k。$$

由(2.3)及(2.4)式，即得

$$(2.5) \quad \hat{\theta}_i = \frac{n_i}{n}, i = 1, \dots, k。$$

上式其實是在 $n_i > 0$ ， $i = 1, \dots, k$ ，之下導出的。不過可以證明(習題第1題)，即使某些 n_i 等於0，仍會得到如(2.5)式所給之 θ_i 的 MLE。

由上討論得概似比檢定之統計量為

$$\lambda(n_1, \dots, n_k) = \frac{L(\boldsymbol{\theta}_0|n_1, \dots, n_k)}{\sup_{\boldsymbol{\theta} \in \Omega} L(\boldsymbol{\theta}|n_1, \dots, n_k)} = \frac{f(n_1, \dots, n_k|\boldsymbol{\theta}_0)}{f(n_1, \dots, n_k|\hat{\boldsymbol{\theta}})}。$$

因此

$$(2.6) \quad -2 \log \lambda(n_1, \dots, n_k) = 2 \sum_{i=1}^n n_i \log \frac{\hat{\theta}_i}{\theta_{i0}}。$$

概似比檢定便是拒絕 H_0 ，當 $-2\log\lambda(n_1, \dots, n_k)$ 較大時。在實際應用裡，倒不常使用 $-2\log\lambda(n_1, \dots, n_k)$ ，而是利用在大樣本下， $-2\log\lambda(n_1, \dots, n_k)$ 與所謂皮爾生卡方(Pearson χ^2)統計量近似。

皮爾生卡方統計量為

$$(2.7) \quad Q = \sum_{i=1}^k \frac{(N_i - n\theta_{i0})^2}{n\theta_{i0}}。$$

此等價於(習題第2題)

$$(2.8) \quad Q = \frac{1}{n} \sum_{i=1}^k \frac{N_i^2}{\theta_{i0}} - n。$$

顯然 Q 是量測在 H_0 之下，觀測頻率 N_i 與期望頻率(expected frequency) $n\theta_{i0}$ 之差異情況。底下我們扼要地證明在 H_0 之下， n 很大時， Q 與 $-2\log\lambda(n_1, \dots, n_k)$ 很接近。

假設 n 很大。則若 $\theta = \theta_0$ ，由大數法則 $\hat{\theta}_i \doteq \theta_{i0}$ 。因此 n 很大時， $|\hat{\theta}_i - \theta_{i0}|/\theta_{i0}$ 很小。又由 $\log(1+x)$ 的泰勒展式知， x 很小時，

$$\log(1+x) \doteq x - \frac{1}{2}x^2，$$

即得

$$(2.9) \quad \begin{aligned} \log \frac{\hat{\theta}_i}{\theta_{i0}} &= \log\left(1 + \frac{\hat{\theta}_i - \theta_{i0}}{\theta_{i0}}\right) \\ &\doteq \frac{\hat{\theta}_i - \theta_{i0}}{\theta_{i0}} - \frac{(\hat{\theta}_i - \theta_{i0})^2}{2\theta_{i0}^2} \\ &= \frac{N_i - n\theta_{i0}}{n\theta_{i0}} - \frac{(N_i - n\theta_{i0})^2}{2n^2\theta_{i0}^2}。 \end{aligned}$$

因此

$$(2.10) \quad \begin{aligned} &2 \sum_{i=1}^k N_i \log \frac{\hat{\theta}_i}{\theta_{i0}} \\ &\doteq 2 \sum_{i=1}^k \frac{N_i(N_i - n\theta_{i0})}{n\theta_{i0}} - 2 \sum_{i=1}^k \frac{N_i(N_i - n\theta_{i0})^2}{2n^2\theta_{i0}^2} \end{aligned}$$

$$\begin{aligned}
&= 2 \sum_{i=1}^k \left\{ \frac{(N_i - n\theta_{i0})^2}{n\theta_{i0}} + (N_i - n\theta_{i0}) \right\} \\
&\quad - 2 \sum_{i=1}^k \left\{ \frac{(N_i - n\theta_{i0})^3}{2n^2\theta_{i0}^2} + \frac{(N_i - n\theta_{i0})^2}{2n\theta_{i0}} \right\} \\
&= \sum_{i=1}^k \frac{(N_i - n\theta_{i0})^2}{n\theta_{i0}} - \sum_{i=1}^k \frac{(N_i - n\theta_{i0})^3}{n^2\theta_{i0}^2} \circ
\end{aligned}$$

此處用到

$$(2.11) \quad \sum_{i=1}^k N_i = n = \sum_{i=1}^k n\theta_{i0} \circ$$

(2.10)式中最後一等號右側第二項滿足

$$\left| \sum_{i=1}^k \frac{(N_i - n\theta_{i0})^3}{n^2\theta_{i0}^2} \right| \leq \sum_{i=1}^k \frac{(N_i - n\theta_{i0})^2}{n\theta_{i0}} \cdot \max_{1 \leq i \leq k} \frac{|N_i - n\theta_{i0}|}{n\theta_{i0}} \circ$$

而

$$\max_{1 \leq i \leq k} \frac{|N_i - n\theta_{i0}|}{n\theta_{i0}} = \max_{1 \leq i \leq k} \frac{|n\hat{\theta}_i - n\theta_{i0}|}{n\theta_{i0}} = \max_{1 \leq i \leq k} \frac{|\hat{\theta}_i - \theta_{i0}|}{\theta_{i0}} \circ$$

又前面已提過，在 H_0 之下， n 很大時， $|\hat{\theta}_i - \theta_{i0}|/\theta_{i0}$ 很小。故(2.10)式中最後一等號右側第二項與第一項相比很小。即得

$$-2 \log \lambda(n_1, \dots, n_k) \doteq Q \circ$$

因此當 n 夠大，可以統計量 Q 當做 $-2 \log \lambda(n_1, \dots, n_k)$ 之近似。

由於在 H_0 之下， $\mathbf{N}$ 有 $\mathcal{M}(n, \theta_{10}, \dots, \theta_{k0})$ 分佈，故在 H_0 之下， Q 的分佈函數可以求出。不過由第八章定理4.2，對一般的分佈，當樣本數夠多，概似統計量便有近似的卡方分佈。那時我們並未提供證明。對此處多項分佈的情況，底下我們會證明， n 夠大時， Q 有近似的 χ_{k-1}^2 分佈。我們便是先以較容易計算的統計量 Q ，取代較複雜的統計量 $-2 \log \lambda(n_1, \dots, n_k)$ ，然後再求出 Q 的近似分佈。因此檢定可取為：

$$(2.12) \quad \text{拒絕 } H_0, \text{ 若且唯若 } Q \geq \chi_{1-\alpha, k-1}^2 \circ$$

這就是所謂的**適合度檢定**(goodness-of-fit test)。實際應用時,一般要求期望頻率 $n\theta_{i0}$ 皆至少要達到5。若有些未達到,就要將相鄰的幾項合併,而自由度也隨之減少(見稍後例2.5)。

Q 有近似的 χ^2_{k-1} 分佈之扼要證明。

設 $\mathbf{N} = (N_1, \dots, N_k)$ 有 $\mathcal{M}(n, \theta_1, \dots, \theta_k)$ 分佈。其中 $\theta_1, \dots, \theta_k$ 皆未知。則由3.7節習題第18題,

$$E(N_i) = n\theta_i, \quad \text{Var}(N_i) = n\theta_i(1 - \theta_i), \quad i = 1, \dots, k,$$

$$\text{Cov}(N_i, N_j) = -n\theta_i\theta_j, \quad i \neq j.$$

因此由**多維中央極限定理**, 當 n 很大時, $\mathbf{N} = (N_1, \dots, N_k)$ 有近似的 k 維常態分佈, 期望向量及共變異數矩陣分別為

$$\boldsymbol{\mu} = (n\theta_1, \dots, n\theta_k),$$

$$\boldsymbol{\Sigma} = n \begin{pmatrix} \theta_1(1 - \theta_1) & -\theta_1\theta_2 & -\theta_1\theta_3 & \cdots & -\theta_1\theta_k \\ -\theta_2\theta_1 & \theta_2(1 - \theta_2) & -\theta_2\theta_3 & \cdots & -\theta_2\theta_k \\ -\theta_3\theta_1 & -\theta_3\theta_2 & \theta_3(1 - \theta_3) & \cdots & -\theta_3\theta_k \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -\theta_k\theta_1 & -\theta_k\theta_2 & -\theta_k\theta_3 & \cdots & \theta_k(1 - \theta_k) \end{pmatrix}.$$

由於 $\sum_{i=1}^k \theta_i = 1$, 故 $\boldsymbol{\Sigma}$ 為奇異矩陣, $|\boldsymbol{\Sigma}| = 0$, $\boldsymbol{\Sigma}$ 之維數為 $k - 1$ 。現去掉一個 N_i , 就去掉最後一個 N_k 好了, 則 $(N_1, \dots, N_{k-1})$ 有近似的 $\mathcal{M}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ 分佈, 其中 $\boldsymbol{\mu}_1$ 為 $\boldsymbol{\mu}$ 去掉最後一分量, $\boldsymbol{\Sigma}_1$ 為 $\boldsymbol{\Sigma}$ 去掉最後一行及最後一列。

$\boldsymbol{\Sigma}_1^{-1}$ 可求出如下:

$$\boldsymbol{\Sigma}_1^{-1} = \frac{1}{n} \begin{pmatrix} \frac{1}{\theta_1 + \theta_k} & \frac{1}{\theta_k} & \frac{1}{\theta_k} & \cdots & \frac{1}{\theta_k} \\ \frac{1}{\theta_k} & \frac{1}{\theta_2 + \theta_k} & \frac{1}{\theta_k} & \cdots & \frac{1}{\theta_k} \\ \frac{1}{\theta_k} & \frac{1}{\theta_k} & \frac{1}{\theta_3 + \theta_k} & \cdots & \frac{1}{\theta_k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\theta_k} & \frac{1}{\theta_k} & \frac{1}{\theta_k} & \cdots & \frac{1}{\theta_{k-1} + \theta_k} \end{pmatrix}.$$

現今 $\mathbf{N}_1 = (N_1, \dots, N_{k-1})$, 則由第四章定理4.7, 二次式 $(\mathbf{N}_1 - \boldsymbol{\mu}_1) \boldsymbol{\Sigma}_1^{-1} \cdot (\mathbf{N}_1 - \boldsymbol{\mu}_1)'$ 有近似的 χ_{k-1}^2 分佈。又

$$\begin{aligned} & (\mathbf{N}_1 - \boldsymbol{\mu}_1) \boldsymbol{\Sigma}_1^{-1} (\mathbf{N}_1 - \boldsymbol{\mu}_1)' \\ &= \sum_{i=1}^{k-1} \frac{(N_i - n\theta_i)^2}{n\theta_i} + \frac{1}{n\theta_k} \sum_{i=1}^{k-1} \sum_{j=1}^{k-1} (N_i - n\theta_i)(N_j - n\theta_j) \\ &= \sum_{i=1}^{k-1} \frac{(N_i - n\theta_i)^2}{n\theta_i} + \frac{1}{n\theta_k} \left(\sum_{i=1}^{k-1} (N_i - n\theta_i) \right)^2 \\ &= \sum_{i=1}^k \frac{(N_i - n\theta_i)^2}{n\theta_i} = Q, \end{aligned}$$

其中最後一等式成立, 是用到 $\sum_{i=1}^{k-1} N_i = n - N_k$, 且 $\sum_{i=1}^{k-1} \theta_i = 1 - \theta_k$ 。上述結果為卡爾·皮爾生(1990)首先證出, 可說是適合度檢定的基礎。Gramér(1946)pp.417-419亦提供一藉助 $(N_1, \dots, N_k)$ 之特徵函數的證法, 也可參考。

在不少書中, 將觀測頻率以 o_i 表之, 期望頻率以 e_i 表之, 因此 Q 可寫成

$$(2.13) \quad Q = \sum_{i=1}^k \frac{(o_i - e_i)^2}{e_i}。$$

又 Q 亦可改寫成

$$(2.14) \quad Q = n \sum_{i=1}^k \frac{(N_i/n - \theta_{i0})^2}{\theta_{i0}}。$$

其中 N_i/n 即為事件 A_i 觀測到之相對頻率。有些書習於以(2.14)式來表示 Q 。本書則傾向採用(2.13)式, 二者當然是等價的。當 $k = 2$ 時, (2.7)式成為

$$Q = \frac{(N_1 - n\theta_{10})^2}{n\theta_{10}} + \frac{(n - N_1 - n(1 - \theta_{10}))^2}{n(1 - \theta_{10})} = \frac{(N_1 - n\theta_{10})^2}{n\theta_{10}(1 - \theta_{10})}。$$

此時前述卡方近似, 與基於二項分佈 $\mathcal{B}(n, \theta_1)$, 且檢定 $H_0 : \theta_1 = \theta_{10}$, vs. $H_a : \theta_1 \neq \theta_{10}$ 之常態近似吻合(參考第八章例2.2): 在 H_0 之下, 由中央極限

定理,

$$\frac{N_1 - n\theta_{10}}{\sqrt{n\theta_{10}(1 - \theta_{10})}}$$

有近似的 $\mathcal{N}(0, 1)$ 分佈, 因此

$$\frac{(N_1 - n\theta_{10})^2}{n\theta_{10}(1 - \theta_{10})}$$

有近似的 χ_1^2 分佈。在型 I 錯誤機率為 α 之下, 拒絕域可取為

$$\left\{ \frac{(N_1 - n\theta_{10})^2}{n\theta_{10}(1 - \theta_{10})} \geq \chi_{1-\alpha, 1}^2 \right\}。$$

爲了簡便, 在底下的例子中, 通常我們只敘述虛無假設 H_0 , 對立假設 H_a 如果是取成 H_0 之否定敘述, 就省略不提。

例2.1 我們想檢定一骰子是否爲公正。令 H_0 : 此骰子爲公正, 即 $H_0 : P(A_i) = 1/6, i = 1, \dots, 6$, 其中 A_i 表投擲出現點數 i 。投擲120次, 各點數出現的頻率, 及在 H_0 之下的期望頻率列於表2.2。

表2.2 投擲一骰子120次觀測頻率及期望頻率

	1	2	3	4	5	6
觀測頻率	20	22	17	18	19	24
期望頻率	20	20	20	20	20	20

由表2.2得

$$\begin{aligned}
 Q &= \sum_{i=1}^6 \frac{(o_i - e_i)^2}{e_i} \\
 &= \frac{(20 - 20)^2}{20} + \frac{(22 - 20)^2}{20} + \frac{(17 - 20)^2}{20} + \frac{(18 - 20)^2}{20} \\
 &\quad + \frac{(19 - 20)^2}{20} + \frac{(24 - 20)^2}{20} \\
 &\doteq 1.7 < 9.24 \doteq \chi_{0.90, 5}^2。
 \end{aligned}$$

故在 $\alpha = 0.10$ 之下, 不能拒絕 H_0 。換句話說, 雖然各點數出現的次數並未恰好等於期望頻率20, 有些點還偏離不少, 如點數6, 但卻不足以推翻此骰子爲公正之假設。

附帶一提, 希望讀者仍記得, 如果在 $\alpha = 0.10$ 之下不能拒絕 H_0 , 則在比 0.10 小之 α 下, 亦皆不能拒絕 H_0 。

例2.2 孟德爾(Mondel, 1822-1884)有一著名的關於豌豆(pea)生長的實驗。他將有圓黃(round yellow)種子的與有縐綠(wrinkled green)種子的豌豆雜交。依其理論, 會生長出圓黃、圓綠、縐黃及縐綠種子的後代之比率, 應分別為 $9/16$, $3/16$, $3/16$ 及 $1/16$ 。經由一組有556個樣本的實驗, 我們列出觀測頻率及期望頻率於表2.3。

表2.3 豌豆雜交後生長的觀測頻率及期望頻率

	圓黃	圓綠	縐黃	縐綠
觀測頻率	315	108	101	32
期望頻率	312.75	104.25	104.25	34.75

由表2.3得

$$\begin{aligned}
 Q &= \sum_{i=1}^4 \frac{(o_i - e_i)^2}{e_i} \\
 &= \frac{2.25^2}{312.75} + \frac{3.75^2}{104.25} + \frac{3.25^2}{104.25} + \frac{2.75^2}{34.75} \\
 &\doteq 0.470024。
 \end{aligned}$$

由附表5, 因 $\chi_{0.10,3}^2 \doteq 0.584$, $\chi_{0.05,3}^2 \doteq 0.352$, 故得 p -值介於 0.90 與 0.95 間。因此即使 α 大到 0.90 , 都無法拒絕 H_0 : 孟德爾的理論為正確。

底下我們以上述數據, 比較概似比檢定之統計量與 Q 值之差異。由(2.6)式, 因

$$\begin{aligned}
 \hat{\theta}_1 &= \frac{315}{556}, & \hat{\theta}_2 &= \frac{101}{556}, & \hat{\theta}_3 &= \frac{108}{556}, & \hat{\theta}_4 &= \frac{32}{556}, \\
 \theta_1 &= \frac{9}{16}, & \theta_2 &= \frac{3}{16}, & \theta_3 &= \frac{3}{16}, & \theta_4 &= \frac{1}{16},
 \end{aligned}$$

得

$$-2 \left(315 \log \frac{315/556}{9/16} + 101 \log \frac{101/556}{3/16} + 108 \log \frac{108/556}{3/16} \right.$$

$$+32 \log \frac{32/556}{1/16} \Big) \doteq 0.475445。$$

概似比檢定之統計量與 Q 值，二者相差的確不算大。

由於此實驗**適合的太好**(fit too well)，曾引起費雪的懷疑。也許孟德爾是重覆地做實驗，直到結果看起來很好才停止，然後只公布結果較好的那組數據。事實上，對這類統計分析，樣本數為若干必須事先確定。關於此實驗的一些討論，可參考葉偉文譯(2001)p.20，及Rice(1988)pp.285-286。

例2.3 在第二章例2.12，我們給出台灣銀行發行的公益彩券幸運獎中獎號碼。其中末一位數碼之出現頻率似乎不太均勻。由於共24期，末一位數碼0, 1, $\dots$, 9每個數出現之期望頻率 $2.4 < 5$ 。又首位不能為0，若首位不計，只看末6位數碼，則共有144個數字，0, 1, $\dots$, 9每個數字出現之期望頻率成為 $14.4 > 5$ ，我們列出各數字之觀測頻率於表2.4。

表2.4 24期公益彩券幸運獎末6位數碼0, 1, $\dots$, 9之觀測頻率

數字	0	1	2	3	4	5	6	7	8	9
觀測頻率	12	13	13	21	8	16	16	18	13	14

由表2.4得

$$Q = \sum_{i=0}^9 \frac{(o_i - e_i)^2}{e_i} = \frac{114.4}{14.4} \doteq 7.94 < 14.68 \doteq \chi_{0.90,9}^2。$$

故雖有些數字出現頻率顯得特別高(如3)，有些顯得特別低(如4)，但在 $\alpha = 0.10$ (甚至更大許多的 α)之下，不能拒絕 H_0 ：末6位數碼0, 1, $\dots$, 9隨機地出現。

現假設我們只關心3是否出現異常頻繁。由於3出現21次，非3出現123次，而非3之期望頻率為129.6，如此得

$$Q = \frac{(21 - 14.4)^2}{14.4} + \frac{(123 - 129.6)^2}{129.6} \doteq 3.36。$$

又 $\chi_{0.90,1}^2 \doteq 2.71$, $\chi_{0.95,1}^2 \doteq 3.84$ 。故在 $\alpha = 0.10$ 之下，拒絕 H_0 ：末6位數碼3出現之機率為0.1，非3出現之機率為0.9，但在 $\alpha = 0.05$ 之下，則接

受 H_0 。印證我們直觀上的想法(亦見第二章例2.11): 單獨看某一特定事件出現的頻率異常是值得可疑的, 但若事件多些, 有些事件出現頻率異常, 可能就不足為奇了。

例2.4 美國加州發行的超級樂透(SuperLotto Plus)彩券, 每週三、六各開獎一次。自1375期起, 每期自1-47的號碼中, 挑選出5個相異號碼組成頭獎, 及自1-27的號碼中, 挑選一特別號。我們列出自第1375期(2000年6月7日開獎)至1542期(2002年1月12日開獎)總共168期的特別號及頭獎號碼之出現頻率於表2.5。

表2.5 加州樂透彩券特別號及頭獎號碼之出現頻率

特別號

號碼	1	2	3	4	5	6	7	8	9	10	11	12	13	14
頻率	8	4	4	7	2	8	3	8	11	5	5	2	8	6

號碼	15	16	17	18	19	20	21	22	23	24	25	26	27
頻率	8	9	8	8	5	7	3	9	10	7	4	6	3

頭獎號碼

號碼	1	2	3	4	5	6	7	8	9	10	11	12	13
頻率	18	19	25	22	23	20	13	15	19	17	14	21	14

號碼	14	15	16	17	18	19	20	21	22	23	24	25	26
頻率	24	20	23	8	9	20	17	8	11	14	22	13	22

號碼	27	28	29	30	31	32	33	34	35	36	37	38	39
頻率	14	20	20	28	12	22	17	20	15	22	16	19	18

號碼	40	41	42	43	44	45	46	47
頻率	12	16	12	21	21	22	25	17

先看特別號, 每個號碼出現之期望頻率為

$$e_i = \frac{168}{27} = 6.2222(\text{次}) \geq 5, i = 1, \dots, 27。$$

但號碼23出現多些有10次, 號碼5則只出現2次。我們計算 Q 值:

$$Q = \sum_{i=1}^{27} \frac{(o_i - e_i)^2}{e_i} \doteq \frac{162.67}{6.2222} \doteq 26.14 < 35.56 \doteq \chi_{0.90, 26}^2。$$

故在 $\alpha = 0.10$ 之下, 不能拒絕 H_0 : 特別號為隨機地出現。

其次看頭獎號碼。如果號碼是隨機地出現, 因168期共開出 $5 \cdot 168 = 840$ 個號碼, 所以每個號碼出現之期望頻率為

$$e_i = \frac{840}{47} \doteq 17.8723(\text{次}) \geq 5, i = 1, \dots, 47。$$

但因每次挑選的5個頭獎號碼為相異, 是屬於取出後不放回, 所以47個號碼之出現頻率, 並未有多項分佈 $\mathcal{M}(840, 1/47, \dots, 1/47)$ 。因此無法以之前的卡方檢定, 來看各數字是否隨機地出現(台北銀行發行的樂透彩券, 頭獎號碼亦為此情況), 由於30出現高達28次, 底下我們來檢定是否異常。

因每次挑5個頭獎號碼, 所以30出現之機率應為 $5/47$ 。令 H_0 : 頭獎號碼中30出現之機率為 $5/47$, 非30出現之機率為 $42/47$ 。對各期之頭獎號碼, 我們將其分成兩類, 其一為有30出現, 其二為無30出現, 則前者之觀測頻率為30, 後者之觀測頻率為138。又前者之期望頻率為

$$168 \cdot \frac{5}{47} \doteq 17.8723,$$

後者之期望頻率為

$$168 \cdot \frac{42}{47} \doteq 150.1277。$$

故得

$$Q = \frac{(28 - 17.8723)^2}{17.8723} + \frac{(140 - 150.1277)^2}{150.1277} \doteq 6.719。$$

又 $\chi_{0.90, 1}^2 \doteq 2.71$, $\chi_{0.99, 1}^2 \doteq 6.63$ 。故不但在 $\alpha = 0.10$ 之下拒絕 H_0 , 即使在 $\alpha = 0.01$ 之下, 亦拒絕 H_0 。

不過如同例2.3中之說明, 單一的檢定一個數字, 較易突顯其出現頻率的異常。對樂透彩之中獎號碼, 其實並不宜只檢定某一特定數字之出現頻

率。在10.4節我們會討論，如何檢定這類取出後不放回之樂透彩中獎號碼之出現頻率。

適合度檢定並非只是檢定多項分佈的參數是否如所給定的值。一般設有一組由未知的分佈函數 $F(x)$ ，所產生之隨機樣本 $X_1, \dots, X_n$ ，欲檢定

$$(2.15) \quad H_0 : F(x) = F_0(x), x \in R,$$

其中 $F_0(x)$ 為一特別的分佈函數(連續型或離散型皆可)。這類的檢定皆通稱**適合度檢定問題**(goodness-of-fit problem)。而任一對(2.15)式的檢定，皆稱為**適合度檢定**。“fit”一詞，就是指機率模型是否合適。有如看衣服是否合身，適不適合此工作等。

更明確地說，對一隨機現象，由收集到的數據，我們想檢定某一模型對此組數據是否適合。想區別數據的偏差是由於機運所產生，或真的是因模型不夠精確，這種過程便統稱為**適合度檢定**。前面所介紹的卡方檢定，又稱**皮爾生之卡方適合度檢定**(Pearson's chi-squared goodness-of-fit test)，是最常用的適合度檢定法之一。雖然卡方檢定是針對母體有多項分佈而設計，但即使母體是源自重覆觀測一有無限多個可能離散值的結果之現象(因此不會產生多項分佈)，由於尾部之機率通常較小，適當地合併尾部，往往產生多項分佈。見下例。

例2.5 表2.6給出某打字員打一本440頁的書，各頁錯字數之觀測頻率及期望頻率。我們想檢定每頁錯字數是否遵循 $P(3)$ 分佈。

因錯字數8及9之期望頻率均小於5，所以我們將錯誤數8與9二類併入相鄰的錯字數7中。由表2.6得

$$\begin{aligned} Q &= \sum_{i=0}^7 \frac{(o_i - e_i)^2}{e_i} \\ &= \frac{(18 - 21.9)^2}{21.9} + \frac{(53 - 65.7)^2}{65.7} + \frac{(103 - 98.6)^2}{98.6} + \frac{(107 - 98.6)^2}{98.6} \\ &\quad + \frac{(82 - 73.9)^2}{73.9} + \frac{(46 - 44.0)^2}{44.0} + \frac{(18 - 22.2)^2}{22.2} + \frac{(13 - 14.3)^2}{14.3} \\ &\doteq 5.95 < 12.02 \doteq \chi_{0.90, 7}^2. \end{aligned}$$

表2.6 錯字數之觀測頻率及期望頻率

錯字數	觀測頻率	$\mathcal{P}(3)$ 之機率	期望頻率
0	18	0.0498	21.9
1	53	0.1494	65.7
2	103	0.2240	98.6
3	107	0.2240	98.6
4	82	0.1680	73.9
5	46	0.1008	44.0
6	18	0.0504	22.2
7	10	0.0216	9.5
8	2	0.0081	3.6
9	1	0.0027	1.2

故在 $\alpha = 0.10$ 之下，不能拒絕 H_0 ：每頁錯字數遵循 $\mathcal{P}(3)$ 分佈。由於5.95遠比12.02小很多，即使讓 α 變大很多，都無法拒絕 H_0 。換句話說，波松分佈很合適(good fit)當做錯字數之分佈。

在適合度的檢定中，有時只想檢定是否可以某一特定分佈當做隨機模型，至於參數則未限制。例如，在例2.5中，想檢定 H'_0 ：每頁錯字數遵循波松分佈，或等價地說檢定 H'_0 ：每頁錯字數遵循 $\mathcal{P}(\lambda)$ 分佈。此時我們要先估計參數 λ 。因波松分佈的參數等於期望值，所以用樣本平均(此亦為 λ 之MLE，見第六章例3.4)來估計 λ

$$\frac{1}{440}(0 \cdot 18 + 1 \cdot 53 + \cdots + 9 \cdot 1) = \frac{1,341}{440} \doteq 3.0477 \doteq 3。$$

然後作法如例2.5，只是 Q 的自由度要少1，成為6。現因

$$Q \doteq 5.95 < 10.64 \doteq \chi_{0.90,6}^2,$$

所以仍不能拒絕 H'_0 ：每頁錯字數遵循波松分佈。

一般若要估計的參數有 t 個, 則自由度會少掉 t 個。即自由度成為 $k - t - 1$ 。

例2.6 波松分佈雖是數學家波松所推導出, 但他並未繼續探討此分佈之性質。真正使波松分佈發揚光大的為出生於俄國的數學家波爾特凱維滋(Bortkewitz, 1868-1931, 亦寫成Bortkiewicz)。波爾特凱維滋在其所著的小數法則一書中, 描述波松分佈之一些不同的應用。例如, 書中有記錄完整的西元1875至1894年的20年間, 普魯士(Prussian)14個騎兵隊, 士兵被馬踢傷致死的統計資料(共280筆), 見表2.7。

表2.7 普魯士士兵被馬踢傷致死之觀測頻率及期望頻率

每年死亡人數	觀測頻率	$P(0.7)$ 之機率	期望頻率
0	144	0.4965	139.0
1	91	0.3476	97.3
2	32	0.1217	34.1
3	11	0.0284	8.0
4	2	0.0050	1.4
≥ 5	0	0.0007	0.2

在表2.7中仍以樣本平均

$$\frac{0 \cdot 144 + 1 \cdot 91 + 2 \cdot 32 + 3 \cdot 11 + 4 \cdot 2 + 5 \cdot 0}{280} = \frac{196}{280} = 0.7,$$

當做波松分佈之參數 λ 的估計值。可看出觀測頻率與期望頻率差異並不太大。我們可做一適合度檢定:

$$\begin{aligned} Q &= \frac{(144 - 139)^2}{139} + \frac{(91 - 97.3)^2}{97.3} + \frac{(32 - 34.1)^2}{34.1} + \frac{(13 - 9.6)^2}{9.6} \\ &= 1.91 < 4.61 \doteq \chi_{0.90, 2}^2. \end{aligned}$$

此處由於 λ 為估計值, 故自由度 $= 4 - 1 - 1 = 2$ 。即得在 $\alpha = 0.10$ 之下, 不能拒絕 H_0 : 每年被馬踢傷致死之士兵人數遵循波松分佈。

卡方檢定雖是針對離散型分佈，但亦可用來檢定連續型分佈之適合度。見下例。

例2.7 假設量測某廠牌汽車電池之壽命，得到如表2.8之40筆數據(單位為年)。

我們將表2.8之數據分成7群，列於表2.9，其中最右一行的期望頻率，底下我們會說明是如何獲得。

表2.8 電池壽命

2.2, 3.4, 2.5, 3.3, 4.7, 4.1, 1.6, 4.3, 3.1, 3.8, 3.5, 3.1, 3.4, 3.7,
3.2, 4.5, 3.3, 3.6, 4.4, 2.6, 3.2, 3.8, 2.9, 3.2, 3.9, 3.7, 3.1, 3.3,
4.1, 3.0, 3.0, 4.7, 3.9, 1.9, 4.2, 2.6, 3.7, 3.1, 3.4, 3.5。

表2.9 電池壽命之觀測頻率及期望頻率

區間	區間中點	觀測頻率	相對頻率	期望頻率
[1.45, 1.95)	1.7	2	0.050	0.73
[1.95, 2.45)	2.2	1	0.025	2.65
[2.45, 2.95)	2.7	4	0.100	6.80
[2.95, 3.45)	3.2	15	0.375	10.68
[3.45, 3.95)	3.7	10	0.250	10.29
[3.95, 4.45)	4.2	5	0.125	6.08
[4.45, 4.95)	4.7	3	0.075	2.77

由於電池壽命皆只計到小數第一位，所以區間[1.45, 1.95)若改為[1.4, 1.9]也無不可。我們這種寫法，會使得區間都接連起來。將表2.9中之相對頻率，以圖2.1表示，就是所謂相對頻率直方圖。

圖2.1顯示電池壽命相對頻率之直方圖，似呈鐘形，我們猜測可以常態分佈 $\mathcal{N}(\mu, \sigma^2)$ 當做其模式。以樣本平均及樣本變異數，分別來估計 μ 及 σ ，得

$$\hat{\mu} = 3.4125, \quad \hat{\sigma} = 0.70。$$

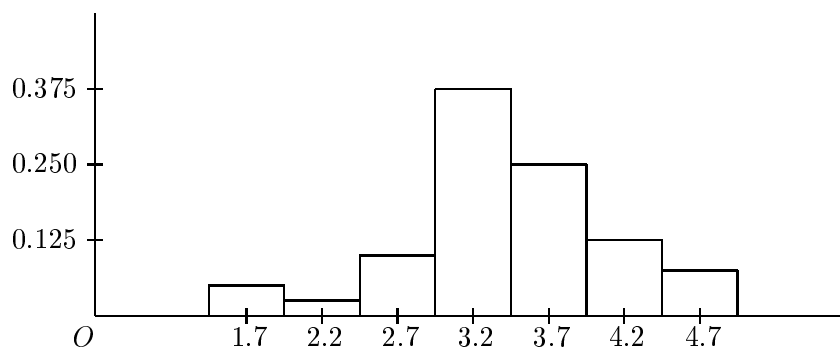


圖2.1 電池壽命之相對頻率直方圖

由此可得表2.9最右一行。例如，區間 $[1.95, 2.45)$ 之期望頻率可如下求得：

$$\begin{aligned}
 & 40 \cdot P\left(\frac{1.95 - 3.4125}{0.7} \leq Z < \frac{2.45 - 3.4125}{0.70}\right) \\
 & \doteq 40 \cdot P(-2.089 \leq Z < -1.375) \\
 & \doteq 40 \cdot (0.98165 - 0.91545) \doteq 2.65。
 \end{aligned}$$

不過第一個區間的期望頻率，是由所有在常態曲線下，1.95左側之面積而得，最後一個區間之期望頻率，是由所有在常態曲線下，4.45右側之面積而得。

現因

$$\begin{aligned}
 Q &= \frac{(7 - 10.18)^2}{10.18} + \frac{(15 - 10.68)^2}{10.68} + \frac{(10 - 10.29)^2}{10.29} + \frac{(8 - 8.85)^2}{8.85} \\
 &\doteq 2.83 < 3.84 \doteq \chi_{0.95,1}^2。
 \end{aligned}$$

其中自由度 $= 4 - 2 - 1 = 1$ 。故在 $\alpha = 0.05$ 之下，不能拒絕 H_0 ：電池壽命有常態分佈。也就是常態分佈可當做電池壽命之分佈。

例2.7提供一般對連續型分佈，以卡方檢定做適合度檢定的步驟：

將觀測值 $X_1, \dots, X_n$ 分群，即離散化。首先選取一分割

$$-\infty = a_0 < a_1 < \dots < a_{k-1} < a_k = \infty。$$

再將樣本以 $\mathbf{N} = (N_1, \dots, N_k)$ 取代，其中 N_i 表 $X_1, \dots, X_n$ 落在區間 $[a_{i-1}, a_i)$ 的個數， $i = 1, \dots, k$ 。若令

$$(2.16) \quad \theta_i = P(a_{i-1} \leq X < a_i), i = 1, \dots, k,$$

則 N 有 $\mathcal{M}(n, \theta_1, \dots, \theta_k)$ 分佈。注意 $\theta_1 = P(-\infty < X < a_1)$ ，而若 X 為取非負值之隨機變數，則取 $a_0 = 0$ 。我們便是要檢定

$$H_0 : \theta_i = \theta_{i0} = P(a_{i-1} \leq X < a_i | H_0), i = 1, \dots, k。$$

例如，某製程所生產的零件，過去有 $\mathcal{E}(1)$ 分佈，由於機器已用了一段時間，我們想檢定生產的零件壽命是否仍有 $\mathcal{E}(1)$ 分佈。取 n 個樣本做實驗，而得壽命 $X_1, \dots, X_n$ 。想檢定 $H_0: X$ 有 $\mathcal{E}(1)$ 分佈。則 $a_0 = 0$ ，且 $\theta_{i0} = e^{-a_{i-1}} - e^{-a_i}$ 。

如何選取分割呢？為了簡便常取 $a_1, \dots, a_{k-1}$ 滿足

$$(2.17) \quad a_1 = a_2 - a_1 = \dots = a_{k-1} - a_{k-2}。$$

另外，給定正整數 k ，有一些理論支持取 $a_1, \dots, a_{k-1}$ 滿足

$$(2.18) \quad \theta_{i0} = \frac{1}{k}, i = 1, \dots, k,$$

為一不錯的選擇。至於 k 又如何選取，並無明顯的規範。通常 k 當然是愈大愈好，但至少要使每一期望頻率 $e_i \geq 5$ 。Kendall and Stuart(1979)pp.455-456, 及462-463, 有一些關於分割及取 k 的討論。

例2.8 假設有下述50筆數據(見Kendall and Stuart(1979)p.457)，欲檢定 H_0 ：此組數據來自 $\mathcal{E}(1)$ 的母體：

0.01, 0.01, 0.04, 0.17, 0.18, 0.22, 0.22, 0.25, 0.25, 0.29,
0.42, 0.46, 0.47, 0.47, 0.56, 0.59, 0.67, 0.68, 0.70, 0.72,
0.76, 0.78, 0.83, 0.85, 0.87, 0.93, 1.00, 1.01, 1.01, 1.02,
1.03, 1.05, 1.32, 1.34, 1.37, 1.47, 1.50, 1.52, 1.54, 1.59,
1.71, 1.90, 2.10, 2.35, 2.46, 2.46, 2.50, 3.73, 4.07, 6.03。

(i) 採取滿足(2.17)式的分割且 $k = 4$ 。

取 $a_0 = 0, a_1 = 0.50, a_2 = 1.00, a_3 = 1.50$ 。得表2.10。

表2.10 滿足(2.17)式的分割

區間	觀測頻率	期望頻率
$[0, 0.50)$	14	19.7
$[0.50, 1.00)$	12	11.9
$[1.00, 1.50)$	10	7.2
$[1.50, \infty)$	14	11.2
	50	50

表2.10中，區間 $[0, 0.50)$ 之期望頻率乃如下求得：

$$50 \cdot (e^0 - e^{-0.50}) \doteq 50 \cdot 0.3935 \doteq 19.7,$$

餘類推。由此得

$$\begin{aligned}
 Q &= \frac{(14 - 19.7)^2}{19.7} + \frac{(12 - 11.9)^2}{11.9} + \frac{(10 - 7.2)^2}{7.2} + \frac{(14 - 11.2)^2}{11.2} \\
 &\doteq 3.44 < 6.25 \doteq \chi_{0.90,3}^2.
 \end{aligned}$$

故在 $\alpha = 0.10$ 之下，不能拒絕 H_0 。

(ii) 採取滿足(2.18)式的分割且 $k = 4$ 。

欲使4個區間之機率均為 $1/4=0.25$ ， a_1, a_2, a_3 分別約為0.288, 0.693, 1.386。得到表2.11。

表2.11 滿足(2.18)式的分割

區間	觀測頻率	期望頻率
$[0, 0.288)$	9	12.5
$[0.288, 0.693)$	9	12.5
$[0.693, 1.386)$	17	12.5
$[1.386, \infty)$	15	12.5

由表2.11得

$$\begin{aligned}
 Q &= \frac{(9 - 12.5)^2}{12.5} + \frac{(9 - 12.5)^2}{12.5} + \frac{(17 - 12.5)^2}{12.5} + \frac{(15 - 12.5)^2}{12.5} \\
 &\doteq 4.08 < 6.25 \doteq \chi_{0.90,3}^2.
 \end{aligned}$$

故在 $\alpha = 0.10$ 之下，不能拒絕 H_0 。

由於(i)及(ii)中所得之 Q 值有些差異，可看出對某些不同的 α ，有可能二者所得之推論不同。

多項分佈是二項分佈的推廣，描述一有數個可能結果的隨機現象，可說是一簡單的分佈。但基於此分佈所發展出來的卡方檢定，卻可應用到各種分佈的適合度檢定。真可說是殺牛用雞刀。

要注意一點：將一連續型的分佈離散化，會產生一些問題。例如，對於前述 $\mathcal{E}(1)$ 分佈之檢定，設有另一分佈 F ，亦滿足(2.17)式(這是很容易的)，則也會接受 H'_0 ：零件壽命以 F 為共同分佈函數。在非參數方法裡，有一**柯莫洛夫檢定**(Kolmogorov test)，可以不用對觀測值分群。另外，卡方檢定的好處是適用性廣泛。至於對一特定的分佈，有時會有更好的檢定方法。如對常態分佈，有**吉得檢定**(Geary's test)，可參考Kendaall and Stuart(1979)pp.486-487。由於常態分佈出現在很多機率模型中，關於**常態性的檢定**(Test of normality)為統計學中一重要的研究題材。

10.3 其他應用

上一節的卡方檢定，尚有不少應用，特別是在有關**類別資料**(categorical data)方面，本節我們列舉一些。

10.3.1 獨立性之檢定

假設我們想知道高雄市民對港市合一的政策，其意見與其黨性傾向是否獨立。隨機地訪問到1,000位市民，將其黨性傾向分為甲，乙，丙三類，並將其分為贊成或反對前述政策二類。觀察到的頻率列在表3.1。這就是所謂**列聯表**(contingency table)，又稱**二變量頻率表**(bivariate frequency table)，為一種表示二變量頻率分配的表。

一列聯表若有 r 列及 c 行，便稱為 $r \times c$ 列聯表。表3.1中列與行的總數，分別稱為**邊際頻率**(marginal frequency)。令 H_0 ：高雄市民對港市合

表3.1 2×3 列聯表

意見\政黨	甲	乙	丙	總數
贊成	182	213	203	598
反對	154	138	110	402
總數	336	351	313	1,000

一的意見與政黨傾向獨立。基於表3.1中的 $6(2 \times 3)$ 類觀測頻率，與在 H_0 之下的期望頻率之吻合程度，以決定接受或拒絕 H_0 。底下以 F (Favor)及 A (against)，分別表任選一市民為贊成或反對市港合一政策的事件。

利用邊際頻率可得下述估計值：

$$P(\text{甲}) = \frac{336}{1,000}, \quad P(\text{乙}) = \frac{351}{1,000}, \quad P(\text{丙}) = \frac{313}{1,000},$$

$$P(F) = \frac{598}{1,000}, \quad P(A) = \frac{402}{1,000}。$$

若 H_0 為真，即行列二變數為獨立，則應有

$$P(\text{甲} \cap F) = P(\text{甲})P(F) = \frac{336}{1,000} \frac{598}{1,000},$$

$$P(\text{甲} \cap A) = P(\text{甲})P(A) = \frac{336}{1,000} \frac{402}{1,000},$$

$$P(\text{乙} \cap F) = P(\text{乙})P(F) = \frac{351}{1,000} \frac{598}{1,000},$$

$$P(\text{乙} \cap A) = P(\text{乙})P(A) = \frac{351}{1,000} \frac{402}{1,000},$$

$$P(\text{丙} \cap F) = P(\text{丙})P(F) = \frac{313}{1,000} \frac{598}{1,000},$$

$$P(\text{丙} \cap A) = P(\text{丙})P(A) = \frac{313}{1,000} \frac{402}{1,000}。$$

每類之期望頻率之估計值，則是將上述機率分別乘上總觀測值1,000。例如， $\text{甲} \cap F$ (甲政黨傾向中贊成港市合一政策)之期望頻率估計值為

$$\frac{336}{1,000} \frac{598}{1,000} \cdot 1,000 \doteq 200.9。$$

一般期望頻率之估計值的公式為

$$(3.1) \quad \hat{e}_{ij} = \frac{\text{第}i\text{行總數} \cdot \text{第}j\text{列總數}}{\text{總觀測數}}。$$

我們將(估計的)期望頻率附於觀測頻率之右側括號中, 而得表3.2。事實上只要求第一列中任兩個期望頻率, 其餘用減法便可得到, 相信讀者不難看出。

表3.2 觀測及期望頻率

意見\政策	甲	乙	丙	總數
贊成	182(200.9)	213(209.9)	203(187.2)	598
反對	154(135.1)	138(141.1)	110(125.8)	402
總數	336	351	313	1,000

仍令

$$(3.2) \quad Q = \sum_{i=1}^r \sum_{j=1}^c \frac{(o_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}},$$

其中 o_{ij} 表觀測頻率, e_{ij} 表期望頻率。則

$$(3.3) \quad \text{拒絕} H_0, \text{若且唯若 } Q \geq \chi_{1-\alpha, (r-1)(c-1)}^2。$$

換句話說, Q 有(近似的)卡方分佈, 其中自由度 v 之公式為(見底下註3.1)

$$(3.4) \quad v = (r-1)(c-1)。$$

在本例中 $v = (2-1)(3-1) = 2$, 且

$$\begin{aligned} Q &= \frac{(182 - 200.9)^2}{200.9} + \frac{(213 - 209.9)^2}{209.9} + \frac{(203 - 187.2)^2}{187.2} \\ &\quad + \frac{(154 - 135.1)^2}{135.1} + \frac{(138 - 141.1)^2}{141.1} + \frac{(110 - 125.8)^2}{125.8} \\ &\doteq 7.85 > 5.99 \doteq \chi_{0.95, 2}^2。 \end{aligned}$$

故在 $\alpha = 0.05$ 之下, 拒絕 H_0 , 即高雄市民對港市合一政策的意見與政黨傾向並不獨立。

要留意的是，我們所依據的統計量 Q ，其分佈只是有近似的卡方分佈。 Q 的值因是由那 rc 類的頻率而得，其實是取離散的值。但只要自由度大於1，通常連續的卡方分佈，可當做離散的統計量 Q 之一很好的近似。對 2×2 的列聯表，由於自由度僅為1，為了精確些，可採葉慈連續性的更正(Yates' correction for continuity)。更正統計量為

$$Q_{\text{更正}} = \sum_{i=1}^2 \sum_{j=1}^2 \frac{(|o_{ij} - \hat{e}_{ij}| - 0.5)^2}{\hat{e}_{ij}}。$$

註3.1

$$\sum_{j=1}^c \frac{(o_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}}$$

有近似的 χ^2_{c-1} 分佈(理由與10.2節之推導相同)。因此

$$Q = \sum_{i=1}^r \left(\sum_{j=1}^c \frac{(o_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}} \right)$$

亦有近似的卡方分佈。而在對 i 求和時，同樣的理由自由度又少1，故總共之自由度為 $(r-1)(c-1)$ 。這可由在求表3.2括號中的期望頻率時，只要求出第一列中任兩個，第三個就決定了(因總數知道)，而求出第一列，第二列便也決定了，共總共只需計算 $(2-1)(3-1) = 2$ 個，其餘皆可由減法求得。

一般在卡方檢定裡，自由度亦可以下述方式決定：

$$(3.5) \quad \text{自由度} = s - t - 1,$$

其中 s = 求和中總共的項數， t = 要以估計值取代之獨立參數的個數。在(3.2)式中， $s = rc$ ，而行與列各有一限制(總數知道)，故 $t = r + c - 2$ 。因此自由度 $= s - t - 1 = rc - (r + c - 2) - 1 = (r-1)(c-1)$ 。

10.3.2 齊性之檢定

在表3.1中，是由隨機地訪問1,000位市民，依機運而得各行及各列的總數。另一種情況是行或列的總數事先決定。例如，事先決定訪問甲政

黨(傾向)200位, 乙政黨150位, 丙政黨150位, 總共500位。並調查對某項政策的意見, 分贊成、反對及未決定三類, 而得表3.3。

表3.3 3×3 列聯表

意見\政黨	甲	乙	丙	總數
贊成	82(85.6)	70(64.2)	62(64.2)	214
反對	93(88.8)	62(66.6)	67(66.6)	222
未決定	25(25.6)	18(19.2)	21(19.2)	64
總數	200	150	150	500

我們想檢定 H_0 : 各政黨支持各不同意見的比例一致。我們就是要決定各政黨的意見是否為齊性的(homogeneous)。這種檢定稱為齊性之檢定(test for homogeneity)。

假設是齊性(即 H_0 為真), 則仿10.3.1節的作法, 我們可得期望頻率之估計值, 並列於表3.3各觀測頻率之右側括號中。由於

$$Q = \sum_{i=1}^3 \sum_{j=1}^3 \frac{(o_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}} \doteq 1.53 < 9.49 \doteq \chi_{0.95, 4}^2,$$

故在 $\alpha = 0.05$ 之下, 不能拒絕 H_0 。

附帶一提, 你知道對應上述 H_0 之 H_a 為何? H_a 為 H_0 之否定, 即至少有一項意見(贊成、反對或未決定), 在各政黨間之支持比率並不相同。

再給一例。

例3.1 西元1972-1974年, 英國進行一有關甲狀腺疾病(thyroid disease)與心臟病(heart disease)的研究。二十年後做一後續的追蹤研究, 其中有些數據頗有意思。Appleton(1996)一文對其中婦女抽煙與死亡的數據做一些分析, 我們列出部分數據在表3.4。

表3.4顯示在全部參與實驗的婦女中,

$$\begin{aligned} \text{抽煙者之死亡率} &= \frac{139}{582} \doteq 23.88\%, \\ \text{不抽煙者之死亡率} &= \frac{230}{732} \doteq 31.42\%。 \end{aligned}$$

表3.4 婦女抽煙與死亡的數據

	抽煙	不抽煙	總數
死亡	139	230	369
存活	443	502	945
總數	582	732	1,314

不抽煙者之死亡率明顯地較高。

我們來檢定 H_0 : 抽煙與不抽煙之死亡率相同。因

$$\begin{aligned}
 Q &= \sum_{i=1}^2 \sum_{j=1}^2 \frac{(o_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}} \\
 &= \frac{(139 - 163.4)^2}{163.4} + \frac{(230 - 205.6)^2}{205.6} + \frac{(443 - 418.6)^2}{418.6} + \frac{(502 - 526.4)^2}{526.4} \\
 &\doteq 9.09 > 6.63 \doteq \chi_{0.99,1}^2,
 \end{aligned}$$

故在 $\alpha = 0.01$ 之下，拒絕 H_0 ，也就是接受不抽煙者之死亡率與抽煙者之死亡率不同(即較高)的推論。

這樣的結果自然是令人驚訝的，不過統計上的推論，只是證實抽煙與不抽煙兩群人的死亡率不同，且不抽煙那群人死亡率較高，並未證實抽煙是造成死亡率較低的原因，此點務必要留意，可參考Simpson(1951)對此方面的探討。在本書3.5節介紹相關係數時，我們曾有類似的說明。由於辛普生(Simpson)之首先注意此問題，後來對於兩個變數的關係，因另一變數的介入而反過來，便統稱**辛普生詭論**(Simpson's paradox)。

如果仔細分析數據，發現有一重要的變數不能忽略：年齡。初次調查時，高齡者中較少是抽煙的(這也是一值得玩味的現象)。我們對年齡分群，而得表3.5，其中以“+”表抽煙，“-”表不抽煙，第一列數字表死亡人數，第二列數字表存活人數。

經過20年，不論是否抽煙，年紀大的總是較易死亡。只要看表3.5中，當初75歲以上者，20年後無一存活即可得知。若將表3.5中，65歲以上之資料去掉，而得表3.6。則抽煙者之死亡率便高於不抽煙者了：

$$\frac{97}{533} \doteq 18.20\% > 12.06\% \doteq \frac{65}{539}。$$

表3.5 對年齡分群之抽煙與死亡的數據

18-24		25-34		35-44		45-54		55-64		65-74		≥ 75	
+	-	+	-	+	-	+	-	+	-	+	-	+	-
2	1	3	5	14	7	27	12	51	40	29	101	13	64
53	61	121	152	95	114	103	66	64	81	7	28	0	0

表3.6 不滿65歲婦女抽煙與死亡的數據

	抽煙	不抽煙	總數
死亡	97	65	162
存活	436	474	910
總數	533	539	1,072

此時因

$$\begin{aligned}
 Q &= \sum_{i=1}^2 \sum_{j=1}^2 \frac{(o_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}} \\
 &= \frac{(97 - 80.5)^2}{80.5} + \frac{(65 - 81.5)^2}{81.5} + \frac{(436 - 452.5)^2}{452.5} + \frac{(474 - 457.5)^2}{457.5} \\
 &\doteq 7.91 > 6.63 \doteq \chi_{0.99,1}^2,
 \end{aligned}$$

故在 $\alpha = 0.01$ 之下，拒絕 H_0' ：不滿65歲之抽煙者與不抽煙者之死亡率相同，即接受不滿65歲之抽煙者的死亡率較高。

10.3.3 r 組 c 維多項母體之檢定

假設有 r 組獨立的 c 維多項分佈：

$$\begin{aligned}
 &\mathcal{M}(n_1, \theta_{11}, \theta_{12}, \dots, \theta_{1c}), \\
 &\quad \vdots \\
 &\mathcal{M}(n_r, \theta_{r1}, \theta_{r2}, \dots, \theta_{rc})。
 \end{aligned}$$

其中 $\sum_{j=1}^c \theta_{ij} = 1$ 。可以想成一實驗有 c 種不同的可能結果 $A_1, \dots, A_c$ 。現有 r 組母體，第 i 組母體會出現結果 j 之機率為 θ_{ij} ，對每組母體分別觀

測 $n_1, \dots, n_r$ 次。例如, 有 r 個骰子, 每個骰子有 c 個面, 分別投擲 $n_1, \dots, n_r$ 次。則會產生 r 個多項分佈。令 o_{ij} 為第 i 組母體出現結果 j 之觀測頻率。我們有下述二種檢定。

(i) $H_0: \theta_{ij} = \theta_{ij}^{(0)}, i = 1, \dots, r, j = 1, \dots, c$, 其中 $\theta_{ij}^{(0)}$ 為給定的值。

在 H_0 之下, 期望頻率之估計值為

$$\hat{e}_{ij} = n_i \theta_{ij}^{(0)}, i = 1, \dots, r, j = 1, \dots, c。$$

如同以往,

$$(3.6) \quad Q_i = \sum_{j=1}^c \frac{(o_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}}$$

有近似的 χ_{c-1}^2 分佈, 且

$$(3.7) \quad Q = \sum_{i=1}^r Q_i = \sum_{i=1}^r \sum_{j=1}^c \frac{(o_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}}$$

有近似的 $\chi_{r(c-1)}^2$ 分佈。故得在型 I 錯誤機率為 α 之下,

$$(3.8) \quad \text{拒絕 } H_0, \text{ 若且唯若 } Q \geq \chi_{1-\alpha, r(c-1)}^2。$$

另一種較常見的情況, 是檢定 r 組多項母體, 是否有相同的類別機率(category probability)。即

(ii) $H_0: \theta_{1j} = \theta_{2j} = \dots = \theta_{rj} = \theta_j, j = 1, \dots, c$, 其中 θ_j 並未給定。

在本情況中之檢定程序, 與10.3.1及10.3.2節所討論應是相同的。令 $N = \sum_{i=1}^r n_i$ 。在 H_0 之下, θ_j 之估計值為

$$\hat{\theta}_j = \frac{\sum_{i=1}^r o_{ij}}{N} = \frac{c_j}{N},$$

其中 c_j 為列聯表中第 j 行之和。又在 H_0 之下, 期望頻率之估計值

$$\hat{e}_{ij} = n_i \hat{\theta}_j = \frac{n_i c_j}{N}。$$

令

$$Q = \sum_{i=1}^r \sum_{j=1}^c \frac{(o_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}},$$

則 Q 有近似的 $\chi^2_{(r-1)(c-1)}$ 分佈，其中自由度的計算可參考註3.1。

給一例子如下。

例3.2 假設想檢定某工廠白天、傍晚及夜間三個時段所生產的產品中瑕疵品所佔比率是否相同，而收集到如表3.7之數據。期望頻率之估計值亦附上。

表3.7 三個時段產品之數據

	瑕疵品	良品	總數
白天	45(57.0)	905(893.0)	950
傍晚	55(56.7)	890(888.3)	945
夜間	70(56.3)	870(883.7)	940
總數	170	2,665	2,835

在本例 $r = 3, c = 2$ 。由於

$$Q = \sum_{i=1}^3 \sum_{j=1}^2 \frac{(o_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}} = 6.29 > 5.99 = \chi^2_{0.95, 2},$$

故在 $\alpha = 0.05$ 之下，拒絕 H_0 ，即認為在不同時段瑕疵品之比率不同。

上述方法也可用來檢定兩筆數據所遵循的分佈是否相同，我們以下例來說明。

例3.3 某商店記錄民國89年1月及90年1月各24天每天的營業額如下(單位為萬元):

89年1月: 8.4, 6.8, 7.7, 7.7, 7.5, 10.0, 6.9, 6.8, 6.9, 6.7, 6.6, 6.7,

7.2, 7.6, 7.2, 7.4, 7.0, 7.1, 10.0, 7.4, 9.5, 7.4, 6.6, 9.8。

90年1月: 8.8, 6.8, 7.5, 7.4, 7.3, 10.2, 6.9, 6.8, 6.6, 7.1, 6.5, 6.3,

7.6, 6.3, 7.3, 7.2, 7.1, 7.3, 9.9, 7.2, 9.7, 7.3, 6.4, 9.8。

我們想檢定 $H_0 : F_1(x) = F_2(x), x \in R$ ，其中 F_1, F_2 分別表兩筆數據所遵循的分佈。即要檢定89年1月與90年1月營業額的分佈函數是否相同。

考慮區間 $[0, 7)$, $[7, 7.5)$, $[7.5, 8)$, $[8, \infty)$, 將兩筆數據各分入此四區間, 然後進行一兩組($r = 2$)四維($c = 4$)多項母體類別機率(以 θ_{ij} , $i = 1, 2$, $j = 1, 2, 3, 4$, 表之)是否相同之檢定。即檢定 $H_0' : \theta_{1j} = \theta_{2j}$, $j = 1, 2, 3, 4$ 。若判定兩組機率相同, 便接受二分佈函數相等, 否則拒絕。分類後連同在 H_0 之下, 期望頻率之估計值一併給在表3.8。

表3.8 兩年營業額之數據

	$[0, 7)$	$[7, 7.5)$	$[7.5, 8)$	$[8, \infty)$	總數
89年1月	8(8)	7(8)	4(3)	5(5)	24
90年1月	8(8)	9(8)	2(3)	5(5)	24
總數	16	16	6	10	48

由於

$$\begin{aligned}
 Q &= \sum_{i=1}^2 \sum_{j=1}^4 \frac{(o_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}} = 2 \left(\frac{0^2}{8} + \frac{1^2}{8} + \frac{1^2}{3} + \frac{0^2}{5} \right) \\
 &= \frac{11}{12} \doteq 0.92 < 6.25 \doteq \chi_{0.90, 3}^2,
 \end{aligned}$$

故在 $\alpha = 0.10$ 之下, 接受 H_0' 。即接受兩段期間營業額之分佈相同。

必須一提的是, 類如上一節最後的說明, 對本例之作法, 有些人難免會質疑, 如果是拒絕 H_0' 可能較無問題, 但接受此新的虛無假設 H_0' , 與接受原先的 H_0 , 其意義是否真的相同?

10.4 再論隨機性

先讀底下一段文字。

我翻開報紙, 新聞中這麼說:

「聯合國最新統計顯示, 全球人口正快速老化。到二〇五〇年, 慶祝百歲大壽的人, 將是現在百歲人瑞的十五倍。

而一個人如果活到一百一十四歲，一生就擁有一百萬個小時可供揮霍，稱得上是『時間的百萬富翁』。未來五十年，這種長壽的『時間的百萬富翁』將與日俱增。」

這可真是大新聞！

長壽是好事，不過可以活到一百一十四歲究竟算不算是好事，恐怕頗值得爭議。

我代替電視節目出征，站在路邊訪問經過的路人，**隨機抽樣**，盡量毫不**偏私**地呈現社會大眾的心聲。……

(91年4月14日中國時報39版，三少四壯集，黃明堅專欄，題目為活到一百一十四歲如何?)

上述這段文字有什麼問題呢(要不要算一下114年是否有一百萬小時)?

機率及統計裡，所討論的問題多半與隨機的事物有關。本來隨機的定義是很一般的，只要是事先**不可預料**(unpredictable)者都可稱為**隨機**。但在日常生活中，“隨機”常與“均勻分佈”及“i.i.d.”等涵義相同。譬如說自1至10中隨機地取一個數，其涵義即為每個數字皆有1/10的機率被取中。又如每次開獎為隨機地產生一組頭獎號碼，即表各次的頭獎號碼為i.i.d.。

前述隨機的概念，與抽籤(或拈鬮)是密不可分的。抽籤是誰發明的已不可考。但起源一定很早。在聖經民數記(英文卷名為Numbers)二十六章，耶和華(即上帝)要摩西(Moses)以拈鬮將約旦河(River Jordan)西邊的土地分給以色列各家族。耶和華說“Be sure that the land is distributed by lot”。

抽籤強調的是公平性，機會均等，正好符合這裡隨機的意義。學生要出公差，把每個人的名字寫在紙上，然後隨機地抽，抽中誰就誰去(或誰不去，當然要事先講好)。即使某些很重要的選舉，對票數相同者，抽籤也是一常見的解決方式。

如果遇到數量很大，抽籤的過程就很難執行了。譬如說要從高雄市抽出1,000位市民。高雄市人口約150萬，想將150萬張名條攪拌均勻，顯然不

是很容易。要知即使僅有52張撲克牌，或144張麻將牌，有時也不易洗均勻呢？這時候就需要**隨機數**(random numbers)，或說是**隨機數列**(random sequence)。每個人給一編號，然後依隨機數的順序，而決定那些人是被挑中的。

隨機數最簡單的定義，是指經由一隨機的程序所產生的數，此隨機程序包含重複而獨立的試驗。當我們提到0至9的隨機數，就是指每次試驗，0至9等10個數，每一數皆有0.1的機率產生。用機率的語言來說，就是有一個數列之隨機樣本(即i.i.d.) $X_1, X_2, \dots$ ，以 $P(X = i) = 0.1, i = 0, 1, \dots, 9$ ，為共同分佈。這樣產生的數列，便是隨機數列。

隨機數可用在**模擬**(simulation)、**蒙地卡羅法**(Monte Carlo methods)、抽樣調查、數值分析、計算機程式、實驗物理，以及其他許多需要隨機數之領域。時至今日已發展出不少產生隨機數(或更一般地有某一特定分佈的數列)的方法。關於隨機數發展之歷史，可參考王業鈞譯(2001)。

卡爾·皮爾生大約是第一位發現可藉助隨機數，來解決機率與統計中，一些難以求得精確解的問題之統計學者。起先隨機數是經由抽籤、搖出號碼球、輪盤(將輪盤分成10等分，每一等分對應0, 1, $\dots$, 9中之一數字)等方式產生，然後製成表，即隨機數的表。以這種方式產生的隨機數有時稱為**真正的隨機數**(truly random numbers)。

卡爾·皮爾生鼓勵英國統計學家梯培特(L.H.C. Tippett)製作一隨機數的表，以供有需要者使用。他曾說“在蒙地卡羅一個月輪盤賭的記錄，提供我們做為討論知識之基礎的資料”(The record of a month's roulette playing at Monte Carlo can afford us material for discussing the foundations of knowledge)。蒙地卡羅為蒙納哥(Monaco, 法國東南海岸之一小國，人口約只有3萬)之一城市，是一著名的賭場。蒙地卡羅法便是以此賭場命名。此法現已成為統計及許多科學中，用來解決複雜的數值問題的標準方法。利用此法，可以藉由生產出來的隨機數，進行一些簡單的計算。

西元1927年，梯培特出版一本名為“隨機抽樣數”(Random Sampling Numbers)的書，包含41,600個0, 1, $\dots$, 9的數字，每4個一組，全書共26頁

。雖只是一些無任何意義的數字之收集,但此書在當時的專業書籍中,成了暢銷書。表4.1給出該書第14頁中一部分,共有320個數字。

表4.1 梯培特隨機抽樣書中的一部分

7816	6572	0802	6314	0702	4369	9728	0198
3204	9243	4935	8200	3623	4869	6938	7481
2976	3413	2841	4241	2424	1985	9313	2322
8303	9822	5888	2410	1158	2729	6443	2943
5556	8526	6166	8231	2438	8455	4618	4445
2635	7900	3370	9160	1620	3882	7757	4950
3211	4919	7306	4916	7677	8733	9974	6732
2748	6198	7164	4148	7086	2888	8519	1620
7477	0111	1630	2404	2979	7991	9683	5125
5379	7076	2694	2927	4399	5519	8106	8501

西元1938年另兩位著名的英國統計學家費雪及耶茨(Yates, 1902-1994),以20位對數表中的第15-19位,出版了一本隨機數的書,共有15,000個數字。之後類似的書籍不少。1955年,藍德公司(Rand Corporation)出版了第一本有百萬個數字之隨機數的書,之前這類的書,最多只到25萬個數字。不過在處理大量的模擬實驗時,由於可能需要幾百萬或上千萬個隨機數,這時隨機數表就不見得那麼實用了。

給一隨機數列,如何產生各種隨機數呢? 假設我們由表4.1的第一列取出7816657208等10個數字。由於每一數字之出現皆為隨機,因此依序的5個“兩位數字”(即00至99的整數)的數: 78, 16, 65, 72, 08, 亦為隨機數。此因這是由某一隨機程序所產生的, 00至99等100個兩位數字的數,每一皆有相同的0.01之機率出現。而且這些兩位數字的數之挑選,為相互獨立。只要是隨機數列依序的每一項,為相互獨立,以 $D-U(0,9)$ 為共同分佈,都有此性質。故由右往左依序的5個兩位數字的數: 80, 27, 56, 61, 87, 亦為兩位數字之隨機數。當然由10個數字中,還可得到不少組兩位數字之隨機數,唯一的限制是同一個數字不能用兩次,如不可以取78, 81, 16, $\dots$ 等。如果一個數字至少用了兩次,隨機性便會失去。又並沒有那一組較其他組更

好。

如果需要10個兩位數字之隨機數，就要利用到20個隨機數字。例如，可自表4.1中的第三列取42412424198593132322。

隨機數表中的同一行或同一頁，不要一直重複使用。否則因隨機性而有的不可預料性會失去。通常可自隨機數表的第一頁(或自中間任何一頁)起依序取出(依列或行皆可)數字，直到夠了為止。下次若又需要隨機數，則由上次停止處往下取數字。依此類推。這是為何有些隨機數表要提供上百萬個隨機數，免得幾次就用完了。要儘量避免重複使用，以確保數字的出現為獨立及有共同分佈。

如果需要用到5個1至12間的隨機數，可先自隨機數表中產生兩位數字的數，然後去掉不是介於01至12者。如自表4.1中，先得到78, 16, 65, 72, 08, 02, 63, 14, 07, 02, 43, 69, 97, 28, 01, 98, 32, 04, ...。去掉不合者，即得02, 07, 02, 01, 04等5個所需要者。事實上可以證明(習題第1題)

$$(4.1) \quad P(i \text{ 出現} | \text{給定只有 } 0, 1, \dots, 12 \text{ 出現}) = \frac{1}{12}, i = 1, \dots, 12。$$

讀者也可想想是否有其他自表4.1產生01至12的隨機數列的方法(見習題第2及第3題)。

計算機的誕生，引進了各種產生隨機數的方法，稱為**隨機數生成器**(random number generator)，以一定的程序產生隨機數。由於是以**決定式的演算法**(deterministic algorithm)產生數列，因此被稱為**偽隨機數**(pseudo random numbers)。

隨機數的產生為何這麼麻煩？隨機並非隨便。欲自1至100中，隨機地挑出10個數字，你很可能懶得去查隨機數表，或藉助計算機以產生，通常就順口唸出10個數。問題是一般人對隨機性的判斷，並非很精準。在Hastings(1997)一書中的pp.286-287給出下例。

例4.1 人們有一奇怪的認知，就是覺得奇數較偶數隨機。曾對一班18位學生(作者承認此乃一爲了方便的樣本，而非隨機樣本)做下述實驗：每位學生從1至20的整數中，(自認)隨機地寫出5個(不重複)。

如果是簡單隨機取樣, 則5個數中有 k 個是奇數之機率為

$$P(k \text{ 個奇數}) = \frac{\binom{10}{k} \binom{10}{5-k}}{\binom{20}{5}}, k = 0, 1, \dots, 5。$$

表4.2給出觀測頻率及期望頻率。

表4.2 奇數個數之觀察頻率及期望頻率

	0	1	2	3	4	5
觀測頻率	0	3	0	9	3	3
期望頻率	0.29	2.44	6.27	6.27	2.44	0.29

本來奇數較多($k \geq 3$)與偶數較多($k \leq 2$)之機率各為 $1/2$, 但由表4.2可看出, 顯然學生有較大的傾向挑選比較多的奇數。例如, 二奇三偶與三奇二偶之機率相同, 但有9位寫出三奇二偶, 卻沒有一位寫出二奇三偶。

該書作者尚指出通常人們也認為質數較合成數隨機(有些人在捏造數字時, 會避免如20, 25等看起來“較好”的數)。稍後我們會再給出一些由人們自認隨機地寫出, 但結果卻非隨機數列之例。因此隨機數列的產生, 還是要有一科學的程序, 而非輕率為之。

如何檢驗一數列是否為隨機數列呢?

假設有一個由數字 $0, 1, \dots, 9$ 所組成的數列, 數字是以獨立的方式產生, 每次 $0, 1, \dots, 9$ 皆有 0.1 的機率出現。因此這是一隨機數列。設 x 為一介於 $0, 1$ 之間的實數, 將 x 以無限小數表之。無理數沒問題, 有理數如 0.2 則寫成 $0.1999\dots$ 。令

$$(4.2) \quad Z_k^{(n)}(x) = \frac{x\text{-之首}n\text{位數中等於}k\text{的個數}}{n}, k = 0, 1, \dots, 9, n \geq 1。$$

則若 x 為隨機數列, 顯然要有

$$(4.3) \quad \lim_{n \rightarrow \infty} Z_k^{(n)}(x) = \frac{1}{10}, k = 0, 1, \dots, 9。$$

在此一個介於 $0, 1$ 間的實數 x , 以無限小數表示後, 若滿足(4.3)式, 便稱為以10為底之簡單正常數(simple normal number to the base 10), 簡稱

為**正常數**。如果一數為 $0.12121212\cdots$ ，則顯然非正常數。在數學裡慣於將典型的事物、良好的事物，或只是令人喜愛的事物，命名為“正常”(或常態)。一正常數不一定為一隨機數列。如易見 $0.01234567890123456789\cdots$ ，即 $0, 1, \cdots, 9$, 10個數字不斷地循環出現，為一正常數，但卻不是隨機數列。對一介於 $0, 1$ 間的實數 x (且表示成無限小數，底下同)，令

$$a^{(n)}(x) = \frac{x\text{之首}n\text{位數中出現}12\text{的個數}}{n}, n \geq 1。$$

例如，若 $x = 0.312156212\cdots$ ，則 $a^{(9)}(x) = 2/9$ 。可看出若 x 為隨機數列，則

$$\lim_{n \rightarrow \infty} a^{(n)}(x) = \frac{1}{100}。$$

但前述 $0, 1, \cdots, 9$ 循環出現的小數，卻不滿足上式。因此隨機數列是比正常數還強的概念。但如果是**絕對正常數**(absolutely normal number)，即不只是以10為底，對任意底都使類似的(4.2)式成立，就與隨機數列的概念近似了，在此不擬討論。

有了一介於 $0, 1$ 間的隨機數列，就可以產生各種我們要的隨機數。如若要5位數碼，則可依序每5位一數，而得一5位數的數列。

近年來，由於計算機運算能力的強化，圓周率 π 的小數位數，不斷推進。西元1999年9月，日本東京大學的安正金田(Yasumasa Kanada)博士的研究小組，已求出 π 至小數第206,158,430,208($= 3 \cdot 2^{36}$)位。2002年9月，他們又將記錄推進了6倍多，達到1.2411兆位。此數若以A4紙印出，每頁設可印4,000位，則須 $3.10275 \cdot 10^8$ 頁。如果雙面印，每1,000頁裝訂成冊，將超過三十一萬冊。每冊以5公分厚計，全部疊起來，高度達15,513.75公尺，非常驚人。他們在一台Hitachi超級電腦上，花了四百多小時算出此 π 值。計算 π 的目的是為了測試超級電腦的效率。在數學史上很可能沒有一個數如 π 那麼吸引大家，並充滿傳奇。有關 π 的更多訊息，可見網站

<http://www.joyofpi.com/index.htm>。

我們好奇的是， π 的小數部分，是否形成一隨機數列呢？

為了本節檢定上的方便，表4.3給出幾個特別自由度之 χ^2 分佈較精確的分佈函數。當然有時在求 p -值時，難免還是要藉助內插法。

表4.3 χ^2 分佈之分佈函數 $P(\chi_n^2 \leq \chi_{p,n}^2) = p$

$n \setminus p$	0.20	0.30	0.40	0.50	0.60	0.70	0.75	0.80	0.85	0.90	0.95	0.975	0.99	0.995	0.999
3	1.01	1.42	1.87	2.37	2.95	3.66	4.11	4.64	5.32	6.25	7.81	9.35	11.34	12.84	16.27
9	5.38	6.39	7.36	8.34	9.41	10.66	11.39	12.24	13.29	14.68	16.92	19.02	21.67	23.59	27.88
41	33.25	35.81	38.11	40.34	42.65	45.22	46.69	48.36	50.36	52.95	56.94	60.56	64.95	68.05	74.74
42	34.16	36.75	39.08	41.34	43.68	46.28	47.77	49.46	51.47	54.09	58.12	61.78	66.21	69.34	76.08
43	35.07	37.70	40.05	42.34	44.71	47.34	48.84	50.55	52.59	55.23	59.30	62.99	67.46	70.62	77.42
44	35.97	38.64	41.02	43.34	45.73	48.40	49.91	51.64	53.70	56.37	60.48	64.20	68.71	71.89	78.75
45	36.88	39.58	42.00	44.34	46.76	49.45	50.98	52.73	54.81	57.51	61.66	65.41	69.96	73.17	80.08
46	37.80	40.53	42.97	45.34	47.79	50.51	52.06	53.82	55.92	58.64	62.83	66.62	71.20	74.44	81.40
47	38.71	41.47	43.94	46.34	48.81	51.56	53.13	54.91	57.03	59.77	64.00	67.82	72.44	75.70	82.72
48	39.62	42.42	44.92	47.34	49.84	52.62	54.20	55.99	58.14	60.91	65.17	69.02	73.68	76.97	84.04

例4.2 表4.4為Dodge(1996)根據安正金田博士西元1988及1995年發表有關 π 至小數60億位的資料, 所給出之 Q 值。

由於 $\chi_{0.90,9}^2 = 14.68$, 而表4.4中所列出之 Q 值皆小於14.68, 故在 $\alpha = 0.10$ 之下, 不能拒絕 H_0 : π 之小數至60億位, 形成一隨機數列。事實上表4.4中最大的 Q 值為9.24, 對應的 p -值高達0.416。故即使在 $\alpha = 0.40$ 之下, 亦不能拒絕 H_0 。

例4.2中之檢定, 其實是對正常數的檢定。尚有其他方式來檢驗 π 之小數部分的隨機性, 如見習題第5題。誠如Gardner(1996)p.99所說的“至此為止 π 通過所有隨機性之統計檢定”(So far π has passed all statistical tests for randomness)。圓周率本來非一隨機產生的數, 是一決定式的數列(deterministic sequence), 但它能通過隨機性的統計檢定, 這種數列稱為統計式的隨機(statistically random)。這是比一般普通的隨機性更有用的概念。它不管數列是如何產生的, 而只看它是否能通過統計上隨機性的檢定。我們引進兩位學者的看法來為做註腳。

Lehmer(1951): A random sequence is a vague notion embodying the idea of a sequence in which each term is unpredictable to the uninitiated and whose digits pass a certain number of tests, traditional with statisticians and depending somewhat on the uses to which the sequence is to be put.

Franklin(1963): The sequence $X_1, X_2, \dots$, is random if it has every

表 4.4 圓周率 π 至小數60億位各數字之頻率及對應之 Q 值 $K = 10^3, M = 10^6$

位數\數字	0	1	2	3	4	5	6	7	8	9	Q
1-100	8	8	12	11	10	8	9	8	12	14	4.20
1-500	45	59	54	50	53	50	48	36	53	52	6.88
1-1K	93	116	103	102	93	97	94	95	101	106	4.74
1-10K	968	1,026	1,021	974	1,012	1,046	1,021	970	948	1,014	9.32
1-50K	5,033	5,055	4,867	4,947	5,011	5,052	5,018	4,977	5,030	5,010	5.86
1-100K	9,999	10,137	9,908	10,025	9,971	10,026	10,029	10,025	9,978	9,902	4.09
1-500K	49,915	49,984	49,753	50,000	50,357	50,235	49,824	50,230	49,911	49,791	7.73
1-1M	99,959	99,758	100,026	100,229	100,230	100,359	99,548	99,800	99,985	100,106	5.51
1-10M	999,440	999,333	1,000,306	999,964	1,001,093	1,000,466	999,337	1,000,207	999,814	1,000,040	2.78
1-20M	2,001,162	1,999,832	2,001,409	1,999,343	2,001,106	2,000,125	1,999,269	1,998,404	1,999,720	1,999,630	4.17
1-30M	2,999,157	3,000,554	3,000,969	2,999,222	3,002,593	2,999,997	2,999,548	2,998,175	2,999,592	3,000,193	4.34
1-50M	4,999,632	5,002,220	5,000,573	4,998,630	5,004,009	4,999,797	4,998,017	4,998,895	4,998,494	4,999,733	6.17
1-100M	9,999,922	10,002,475	10,001,092	9,998,442	10,003,863	9,993,478	9,999,417	9,999,610	10,002,180	9,999,521	7.27
1-200M	19,997,437	20,003,774	20,002,185	20,000,410	19,999,846	19,993,031	19,999,161	20,000,287	20,002,307	20,000,562	4.13
1-3000M	299,999,143	299,995,932	299,989,126	299,992,290	300,002,257	299,979,016	300,025,447	299,975,510	300,016,550	300,024,729	9.24
1-6000M	599,963,005	600,033,260	599,999,169	600,000,243	599,957,439	600,017,176	600,016,588	600,009,044	599,987,038	600,017,038	9.00

property that is shared by all infinite sequences of independent samples of random variables from the uniform distribution.

附帶一提, Metropolis et al.(1950)根據Reitwiesner(1950)的結果, 對 e 及 π 的小數前2,000位做卡方檢定, π 當然沒問題, e 就沒有通過檢定。但Johnson and Leeming(1990)基於一些不同的**連串統計量**(run statistics), 對 $\pi, e, \sqrt{2}, \sqrt{3}, \sqrt{5}, \sqrt{7}, \sqrt{11}$ 及 $\sqrt{13}$ 等無理數的小數前100,000位, 檢定其隨機性, 皆能通過。其中 π 及 $\sqrt{7}$ 達到較高的隨機程度。

Coveyou(1969)有一篇著名的文章, 其篇名即為“隨機數的生成是如此重要, 不能僅靠機運”(Random number generation is too important to be left to chance)。Dodge(1996)一文因此建議以 π 的小數前100萬位, 或前若干億位, 當做自然的隨機數的生成器。他認為這可能比一般以任何**人工演算法**(artificial algorithm)所得到的隨機數列都隨機。

表4.5給出 π 的小數前1,000位供大家參考。

表4.5 π 之小數前1,000位

1415926535	8979323846	2643383279	5028841971	6939937510
5820974944	5923078164	0628620899	8628034825	3421170679
8214808651	3282306647	0938446095	5058223172	5359408128
4811174502	8410270193	8521105559	6446229489	5493038196
4428810975	6659334461	2847564823	3786783165	2712019091
4564856692	3460348610	4543266482	1339360726	0249141273
7245870066	0631558817	4881520920	9628292540	9171536436
7892590360	0113305305	4882046652	1384146951	9415116094
3305727036	5759591953	0921861173	8193261179	3105118548
0744623799	6274956735	1885752724	8912279381	8301194912
9833673362	4406566430	8602139494	6395224737	1907021798
6094370277	0539217176	2931767523	8467481846	7669405132
0005681271	4526356082	7785771342	7577896091	7363717872
1468440901	2249534301	4654958537	1050792279	6892589235
4201995611	2129021960	8640344181	5981362977	4771309960
5187072113	4999999837	2978049951	0597317328	1609631859
5024459455	3469083026	4252230825	3344685035	2619311881
7101000313	7838752886	5875332083	8142061717	7669147303
5982534904	2875546873	1159562863	8823537875	9375195778
1857780532	1712268066	1300192787	6611195909	2164201989

回到10.2節樂透彩券的問題。一般樂透彩是從1至 n 的號碼中, 任意挑選 r 個不同的號碼下注, 我們稱此為 n 取 r 。台北銀行目前發行的樂透彩是42取6。

經過開獎 m 期後, 各組頭獎號碼之出現頻率有多項分佈。但若欲對各組頭獎號碼之出現頻率做檢定, 卻是很不實際的。以台北銀行42取6的樂透彩為例, 共有5,245,786組頭獎號碼, 如果每期頭獎號碼都隨機地出現, 則各頭獎號碼之出現頻率有 $\mathcal{M}(m, p, \dots, p)$, 其中共有5,245,786個 p , 而 $p = (5,245,786)^{-1}$ 。若欲以卡方檢定, 且要符合各期望頻率至少為5, 則 m 至少要為 $5 \cdot 5,245,786 = 26,228,930$ 。以每週開獎二次計, 要經252,201.25年才行。這當然不可行。所以只能退而求其次, 做一些其他的檢定。譬如說對1至 n 等 n 個號碼之出現頻率做檢定。

如例2.4所述, 由於每期頭獎 r 個號碼的產生, 是屬於取出後不放回, 因此 m 期後, n 個號碼之出現頻率, 並未有多項分佈, 故10.2節對各號碼出現頻率的卡方檢定並不適用。在Joe(1993)一文中, 導出在 n 取 r 的樂透彩中, 單一號碼, 成對號碼 (i, j) , 及三碼 (i, j, k) 是否均勻地出現的檢定。特別地, 該文給出在 H_0 : 頭獎號碼各數字為均勻地出現之下, 當開獎期數 m 很大時,

$$(4.4) \quad Q_1 = \frac{n-1}{n-r} \sum_{i=1}^n \frac{(o_i - e)^2}{e},$$

有近似的 χ_{n-1}^2 分佈, 其中

$$(4.5) \quad e = m \cdot \frac{r}{n},$$

表各號碼出現之期望頻率, o_i 為各號碼出現之觀測頻率。要注意的是, 此處隱含地假設各期之頭獎號碼為獨立且有共同分佈。 Q_1 可改寫為

$$(4.6) \quad Q_1 = \frac{n(n-1)(\sum_{i=1}^n o_i^2 - m^2 r^2 / n)}{(n-r)mr}。$$

對42取6的樂透彩券, 則有

$$(4.7) \quad Q_1 = \frac{41}{36} \sum_{i=1}^{42} \frac{(o_i - m/7)^2}{m/7}。$$

例4.3 承例2.4。我們來檢定加州發行的樂透彩券 H_0 : 頭獎號碼各數字為隨機地出現。

依(4.4)及(4.5)式計算, $e = 168 \cdot 5/47 \doteq 17.87$,

$$\begin{aligned} Q_1 &\doteq \frac{47-1}{47-5} \sum_{i=1}^{47} \frac{(o_i - 17.87)^2}{17.87} \\ &= \frac{46}{42} \cdot \frac{1015.2343}{17.87} \doteq 62.22。 \end{aligned}$$

因 $Q_1 < \chi_{0.95,46}^2 \doteq 62.83$, 故在 $\alpha = 0.05$ 之下, 無法拒絕 H_0 。但 $Q_1 > \chi_{0.90,46}^2 \doteq 58.64$, 故在 $\alpha = 0.10$ 之下, 拒絕 H_0 。也就是數據顯示, 頭獎號碼之出現, 並非太隨機。附帶一提, 以附表5之 $n = 40$ 及 50 對應的 $\chi_{p,n}^2$ 值, 經由內插, 得到之 $\chi_{0.95,46}^2$ 及 $\chi_{0.90,46}^2$, 分別約為 62.83 及 58.64 , 與利用表4.3所得之差異並不太大。

例4.4 愛爾蘭國家樂透彩(Irish National Lottery)亦為42取6。Boland and Pawitan(1999)曾收集三筆數據。他們在所開設的初等統計課程(那是一個大班)中, 先向學生解釋何謂隨機數的生成, 以及愛爾蘭國家樂透彩的玩法, 然後要學生把自己當做是一隨機生成器, 每人隨機地(不論大小順序)寫出一組頭獎號碼, 如此得到234組號碼。第二筆數據是自西元1994年9月24日至1997年3月8日的樂透彩, 共產生264組頭獎號碼。第三筆數據是以S-Plus統計軟體, 模擬產生264組頭獎號碼。他們的目的是想知道以個人當做隨機生成器的能力如何?

利用(4.4)式, 表4.6給出三筆數據之 Q_1 值。因 $\chi_{0.90,41}^2 \doteq 52.95$, $\chi_{0.95,41}^2 \doteq 56.94$, $\chi_{0.99,41}^2 \doteq 64.95$, 故顯然要拒絕學生所產生的號碼符合隨機性。至於不論樂透彩之中獎號碼, 及S-Plus所產生之號碼, 均符合隨機性。

表4.6 三筆數據之 Q_1 值及 p -值

數據	學生	樂透	模擬
m	234	264	264
Q_1	261.05	39.64	37.82
p -值	$<< 0.001$	0.531	0.613

例4.5 承上例。樂透彩頭獎號碼之散佈情況，有很多量測方式。在 n 取 r 的玩法中，假設 $x_1, x_2, \dots, x_r$ 為頭獎號碼。令

$$MG = MG(x_1, x_2, \dots, x_r) = \min\{x_{(i+1)} - x_{(i)}, i = 1, \dots, r-1\},$$

其中 $x_{(1)} < x_{(2)} < \dots < x_{(r)}$ 為將 $x_1, \dots, x_r$ 按小至大排列。 MG 稱為**最小間距**(minimum gap)。Boland and Pawitan(1999)給出 MG 之分佈如下：

$$(4.8) \quad P(MG = l) = \frac{\binom{n-(l-1)(r-1)}{r} - \binom{n-l(r-1)}{r}}{\binom{n}{r}}, l = 1, \dots, \left\lceil \frac{n-1}{r-1} \right\rceil,$$

其中 $[\cdot]$ 表最大整數函數。注意 $\binom{x}{y} = 0$ ，當 x, y 為二正整數且 $y > x$ 。

以42取6的樂透彩為例。

$$P(MG = 1) = \frac{\binom{42}{6} - \binom{42-5}{6}}{\binom{42}{6}} = 1 - \frac{\binom{37}{6}}{\binom{42}{6}} \doteq 1 - 0.443 = 0.557。$$

$MG = 1$ 表有連號之事件。連號之機率大於 $1/2$ ，這可能讓很多人感到驚訝。但事實上就是如此。在42取6的樂透彩裡，因 $[(42-1)/(6-1)] = [8.2] = 8$ ，故 l 可取值 $1, \dots, 8$ 。表4.7給出 $P(MG = l)$ 之值。

表4.7 在42取6的樂透彩裡 $MG = l$ 之機率

l	1	2	3	4
$P(MG = l)$	0.557	0.270	0.116	0.0422

l	5	6	7	8
$P(MG = l)$	0.0119	0.00218	0.000175	0.00000133

對三筆數據，每一筆之每一組皆有一 MG 之觀測值，由於 MG 有多項分佈，故可以10.2節的方法，檢驗觀測到的 MG 是否遵循如表4.7所給之分佈。表4.8給出三筆數據之相關數據，括號中的為期望頻率。由於 $MG = 5, MG = 6, MG = 7, MG = 8$ 等四事件之期望頻率均小於5，故皆併入 $MG = 4$ 中。因 $\chi^2_{0.90,3} \doteq 6.25$ ， $\chi^2_{0.99,3} \doteq 11.34$ ，故即使在 $\alpha = 0.10$ 之下，樂透彩之頭獎號碼，及S-Plus所產生之號碼，其最小間距之觀測頻率，均吻合該有之分佈。但由所給之 p -值可看出，即使在 $\alpha = 0.001$ 之下，都無法接受學生所產生的號碼，其最小間距之觀測頻率，吻合該有之分佈。

表4.8 三筆數據最小間距之觀測頻率、期望頻率、 Q 值及 p -值

數據	個人	樂透	模擬
m	234	264	264
$MG = 1$	96(130.3)	145(147.0)	135(147.0)
$MG = 2$	72(63.3)	75(71.4)	70(71.4)
$MG = 3$	35(27.2)	33(30.7)	39(30.7)
$MG \geq 4$	31(13.2)	11(14.9)	20(14.9)
Q	36.45	1.40	5.00
p -值	< 0.001	0.705	0.173

上二例顯示，對一般人而言，要產生隨機數列並非易事，即使特別要求他們要“盡量隨機”，但就是很難辦到。譬如說我們會習於將數字分散開，較少寫出連號。但其實在42取6的樂透彩中，連號之出現比率應該佔一半以上。我們曾請一位具碩士學位，且對樂透內涵很清楚的助理，“隨機地”寫出234組號碼，並計算其 Q_1 值約為110.93，較例4.4中由234個不同的學生所得號碼計算出的 Q_1 值小，但仍無法通過隨機的檢定。要“盡量毫不偏私”是很困難的。讀者不妨自己練習寫出一筆頭獎號碼，並計算 Q_1 值，看你夠不夠隨機。要留意的是，由於每個號碼之期望頻率至少要達到5才適合以卡方來檢定，而

$$m \cdot \frac{6}{42} \geq 5 \Rightarrow m \geq 35,$$

故至少須寫出35組號碼才恰當。如果是要檢定成對號碼 (i, j) 是否均勻地出現，就要更多組號碼才行。

欲對樂透彩及隨機性的探討有進一步的了解，可參考黃文璋(2002)及黃文璋、洪宛頻及羅夢娜(2002)等二文。

至此，你應已經知道本節一開始所引那一段文字中的問題了。人的天性很可能缺乏隨機性，靠自己並不易達成隨機抽樣。

利用第四章例1.6，可得對 n 取 r 的樂透彩，一組號碼 $X_1, \dots, X_r$ ，樣本平均 $\overline{X}_r = (X_1 + \dots + X_r)/r$ 之期望值及變異數。即

$$E(\overline{X}_r) = \frac{n+1}{2}, \quad \text{Var}(\overline{X}_r) = \frac{(n+1)(n-r)}{12r} = \frac{n-r}{n-1} \text{Var}(\overline{Y}_r),$$

其中 $\overline{Y}_r$ 為每次取出後放回之樣本平均。 $\text{Var}(\overline{X}_r)$ 與 $\text{Var}(\overline{Y}_r)$ 之比值 $(n-r)/(n-1)$ 的倒數, 就是兩個情況下, 卡方值的比(見(4.4)式)。

在42取6的樂透彩裡,

$$E(\overline{X}_6) = 21.5, \quad \text{Var}(\overline{X}_6) = 21.5。$$

也可對此二值做檢定, 其過程在此不消細說。

最後, 我們對隨機的概念再做一些說明。

樂透彩上市後, 開出的號碼有所謂冷門及熱門, 我們引用底下的一段新聞報導。此報導登了兩次。

北銀彩券部表示, 樂透彩每一次的開獎都是「獨立事件」, 四十二個號碼出現的機率是一樣的。若以每周開獎二次來估算, 經過五萬年, 「四十二選六」所有的號碼組合(即五百二十四萬餘種)都會開出一次, 每一號碼被開出的次數就會十分接近。(91年3月31日聯合報第6版, 91年5月25日聯合報第6版, 記者黃雯雯)

如果上述報導真的是北銀彩券部的看法(但顯然該記者是相信的, 否則也不會登了兩次), 那真是令人失望的, 毫無隨機的概念。不要說不會每一組號碼都會開出一次, 且1至42, 每一號碼被開出的次數差異也會極大。只要看表4.4就可了解: π 的小數前60億位, 0, 1, $\dots$, 9出現次數的差異其實蠻大的, 最多(數字1)與最少(數字4)相差75,821次。但相對頻率(即數字出現的次數除以60億)的差異卻很小。開獎的次數愈多, 各號碼出現之相對頻率, 有**很大的機會**很接近 $1/7 (= 6/42)$ 。此即**大數法則**。但觀測頻率與期望頻率的差, 很可能愈來愈大。可參考黃文璋(1994)第五章例3.6之說明。我們藉該書pp.261-262的一段文字, 來顯示十五年來, 人們對隨機的概念並沒有更清楚。

有一陣子大家樂很風行, 賭徒藉著台銀每期開出的愛國獎券中獎號碼來對獎。經常有沒簽中的賭徒對愛國獎券之搖獎機率產生懷疑。報章雜

誌遂刊登一些教導大家“正確”了解機率學的文章。我們引用下面一段話，供大家參考，可能更有助於提醒我們須正確了解大數法則的含義(見趙慕嵩等(1987))。原文一字不易：

事實上，所謂失常的機率只是在機率學中必然性的短暫現象，其實還是正常的。譬如一顆六面的正方形骰子，上面有一到六的點數，理論上每擲一次就應該使得每個數字各有六分之一的出現機會，那麼連擲六次，是否1, 2, 3, 4, 5, 6等數字剛好各出現一次？當然不會，可是如果連擲六億次，每個點數出現的次數就非常接近一億次，而滿足於六分之一的理論機率。

如果按照理論的機率，愛券連開十五萬次獎，開出105萬組號碼，那麼00 到99這一百組號碼就有可能各出現一萬零五百次，而接近於理論機率。

希望讀者即使自己“不夠隨機”，但卻能對隨機有較清晰的概念。

習 題

第10.2節

1. 設 $\mathbf{N} = (N_1, \dots, N_k)$ 有 $\mathcal{M}(n, \boldsymbol{\theta})$ 分佈，其中 $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)$ 。又設觀測到 $N_1 = n_1, \dots, N_k = n_k$ ，其中有些 $n_i = 0$ ， $i = 1, \dots, k$ 。試證 $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \dots, \hat{\theta}_k)$ 為 $\boldsymbol{\theta}$ 之MLE，其中 $\hat{\theta}_i = n_i/n$ ， $i = 1, \dots, k$ 。
2. 試證(2.7)式等價於(2.8)式。
3. 投擲一骰子60次，點數1至6，各出現13, 19, 11, 8, 5及4次。試在 $\alpha = 0.05$ 之下，檢定此骰子是否公正。(解. $Q = 15.6 > 11.07 \doteq \chi_{0.95, 5}^2$ ，拒絕 H_0)
4. 投擲一骰子60次，點數1至6，各出現8, 11, 5, 12, 15及9次。試在 $\alpha = 0.10$ 之下，檢定此骰子是否公正。(解. $Q = 6 < 9.24 \doteq \chi_{0.90, 5}^2$ ，接受 H_0)

5. 在某次選舉前，民調公司預測4位候選人 A, B, C, D 之支持率分別為 $p_A = 0.30, p_B = 0.25, p_C = 0.25, p_D = 0.20$ 。現隨機取147位選民訪問，4位候選人分別得到46, 50, 22及29位支持。試在 $\alpha = 0.05$ 之下，對民調公司預測的支持率作一檢定。(解. $Q \doteq 10.78 > 7.81 = \chi_{0.95,3}^2$, 拒絕 H_0)
6. 試對Becker and Piper書中之英文字母出現之相對頻率，找一本英文小說的幾頁，做一檢定，看是否吻合。
7. 在例2.3中，試分別在 $\alpha = 0.10$ 及 0.05 之下，檢定 H_0 ：末6位數碼4出現之機率為 0.1 ，非4出現之機率為 0.9 。(解. $Q = 3.46, \chi_{0.90,1}^2 \doteq 2.71, \chi_{0.95,1}^2 \doteq 3.84$ ，在 $\alpha = 0.10$ 之下，拒絕 H_0 ，在 $\alpha = 0.05$ 之下，接受 H_0)
8. 在例2.4中，試分別在 $\alpha = 0.10$ 及 0.01 之下，檢定 H_0 ：頭獎號碼中17出現之機率為 $5/47$ ，非17出現之機率為 $42/47$ 。(解. $Q = 6.422, \chi_{0.90,1}^2 \doteq 2.71, \chi_{0.99,1}^2 \doteq 6.63$ ，在 $\alpha = 0.10$ 之下，拒絕 H_0 ，在 $\alpha = 0.01$ 之下，接受 H_0)
9. 某地樂透獎的頭獎號碼是自 $0, 1, \dots, 9$ 產生6個數字(可重複)。經統計在363期的頭獎號碼，第1個數字之出現頻率如下：

0	1	2	3	4	5	6	7	8	9
35	44	25	40	50	26	33	38	37	35

試在 $\alpha = 0.05$ 之下，檢定 $H_0 : 0, 1, \dots, 9$ 隨機地產生。(解. $Q \doteq 14.1 < 16.92 = \chi_{0.95,9}^2$, 拒絕 H_0)

10. 某公司之電話接線生在267天中，每日接錯次數之頻率如下：

≤ 3	4	5	6	7	8	9	10	11	12	13	14	≥ 15
6	11	14	22	43	31	40	35	20	18	12	7	8

(i) 若欲以波松分佈當做每日接錯次數之分佈，則參數 λ 應為何(求至小數第二位，且表中 ≤ 3 皆當做3， ≥ 15 皆當做15)；

(ii) 試在 $\alpha = 0.10$ 之下, 檢定 H_0 : 每日接錯次數遵循 $\mathcal{P}(\lambda)$ 分佈, 其中 λ 如(i)中所求出。

(解. (i) $\lambda = 8.74$, (ii) $Q \doteq 7.698 < 18.55 \doteq \chi_{0.90,12}^2$, 接受 H_0)

11. 欲檢定 H_0 : 某特定零件之修護時間(單位為天)有 $\mathcal{P}(3)$ 分佈。隨機地取40個零件做實驗, 而得下述結果:

修護時間	0	1	2	3	4	5	6	≥ 7
頻率	1	3	7	6	10	7	6	0

試在 $\alpha = 0.10$ 下, 做一檢定。(解. $Q \doteq 9.22 > 7.78 \doteq \chi_{0.90,4}^2$, 拒絕 H_0)

12. 在上題中, 試在 $\alpha = 0.10$ 之下, 檢定 H_0 : 修護時間有 $\mathcal{P}(\lambda)$ 分佈。(解. $\hat{\lambda} = 3.65$, $Q \doteq 1.62 < 6.25 \doteq \chi_{0.90,3}^2$, 接受 H_0)

13. 欲檢定 H_0 : 某放射性物質每秒鐘所射出之gamma射線有 $\mathcal{P}(2.4)$ 分佈。檢測300個1秒長之區間而得下述數據:

gamma射線數	0	1	2	3	4	5	6	7
頻率	19	48	66	74	44	35	10	4

試在 $\alpha = 0.05$ 之下, 做一檢定。(解. $Q \doteq 29.2 > 12.59 \doteq \chi_{0.95,6}^2$, 拒絕 H_0)

14. 觀測某地32天間每天發生的車禍次數, 而得下述數據:

0, 2, 3, 1, 3, 4, 2, 1, 2, 2, 1, 0, 3, 3, 3, 2,

4, 1, 2, 4, 2, 2, 1, 5, 1, 3, 2, 0, 3, 2, 2, 1。

試在 $\alpha = 0.05$ 之下, 檢定 H_0 : 每日車禍次數有波松分佈。(解. $\hat{\lambda} = 2.1$, $Q \doteq 1.48 < 5.99 \doteq \chi_{0.95,2}^2$, 接受 H_0)

15. 在例2.7中, 試在 $\alpha = 0.10$ 之下, 檢定 H_0 : 電池壽命有 $\mathcal{N}(3.5, 0.70)$ 分佈。(解. $Q \doteq 3.05 < 6.25 \doteq \chi_{0.90,3}^2$, 接受 H_0)

16. 自區間 $(0, 1)$ 以某一隨機的方式取點。欲檢定 H_0 : 此點之p.d.f為 $2x$, $0 < x < 1$ 。將區間分成 $A_1 = \{x|0 < x \leq 1/4\}$, $A_2 = \{x|1/4 < x \leq 1/2\}$, $A_3 = \{x|1/2 < x \leq 3/4\}$, $A_4 = \{x|3/4 < x < 1\}$ 四群。重複且獨立地產生80個點, 其中分別有6, 18, 20及36個點在 A_1, A_2, A_3 及 A_4 中。試在 $\alpha = 0.025$ 之下檢定 H_0 。(解. $Q \doteq 1.83 < 9.35 \doteq \chi_{0.975,3}^2$, 接受 H_0)

17. 有50架飛機, 各飛行1,000小時, 得飛機壞的零件個數之頻率如下:

個數	0	1	2	3	4	≥ 5
頻率	7	10	9	7	6	11

令 X 表任一架飛機壞的零件個數。

(i) 試在 $\alpha = 0.01$ 之下, 檢定 $H_0 : X \sim \mathcal{P}(2)$;

(ii) 試在 $\alpha = 0.10$ 之下, 檢定 $H_0 : X \sim \mathcal{NB}(3, 1/2)$ 。

(解. (i) $Q \doteq 16.504 > 13.28 \doteq \chi_{0.99,4}^2$, 拒絕 H_0 , (ii) $Q \doteq 0.24 < 9.24 \doteq \chi_{0.90,5}^2$, 接受 H_0)

18. 在上題中, 假設 $\overline{X} = 3.25$, 試在 $\alpha = 0.05$ 之下, 檢定 $H_0 : X \sim \mathcal{P}(\lambda)$ 。(解. $Q \doteq 18.05 > 9.49 \doteq \chi_{0.95,4}^2$, 拒絕 H_0)

19. 某項考試有75人參加, 成績如下:

16, 13, 9, 13, 19, 15, 22, 22, 14, 10, 18, 13, 6, 14, 28,
 15, 16, 16, 19, 15, 13, 11, 11, 22, 29, 15, 22, 19, 22, 17,
 15, 9, 24, 15, 20, 20, 18, 23, 13, 26, 16, 26, 15, 10, 29,
 7, 18, 12, 12, 22, 14, 10, 5, 12, 17, 16, 23, 9, 14, 20,
 16, 18, 22, 22, 19, 11, 20, 20, 19, 15, 17, 7, 19, 13, 15。

可求出樣本平均及樣本標準差分別約為16.5及5.3。試取 $k = 10$ 及滿足(2.18)式之分割, 在 $\alpha = 0.10$ 之下, 檢定 H_0 : 成績有常態分佈。(解. 分割點 $a_0 = -\infty$, $a_{10} = \infty$, $a_1, \dots, a_9$ 分別約為9.7, 12.0, 13.7, 15.2, 16.5, 17.8, 19.3, 20.9, 23.3, $Q \doteq 10.5 < 12.02 \doteq \chi_{0.90,7}^2$, 接受 H_0)

20. 試採如(2.18)式之分割, 將下述數據分成4群, 在 $\alpha = 0.05$ 之下, 檢定 H_0 : 數據遵循常態分佈:

0.8, 2.1, 3.4, 1.9, 1.3, 2.7, 2.9, 2.5, 2.4, 3.0, 3.7,
 1.1, 2.9, 1.9, 1.9, 3.0, 3.1, 0.9, 2.4, 2.2, 2.8, 0.5,
 4.2, 1.0, 4.1, 1.1, 2.2, 1.9, 2.4, 0.9, 1.0, 2.9, 1.4,
 3.9, 1.0, 0.6, 2.2, 2.0, 3.7, 2.9, 0.8。

(解. $Q \doteq 3.4 < 3.84 \doteq \chi_{0.95,1}^2$, 接受 H_0)

21. 將下述42筆數據分成 $[0, 40)$, $[40, 50)$, $[50, 60)$, $[60, \infty)$ 4群。試在 $\alpha = 0.05$ 之下, 檢定 H_0 : 數據源自 $\Gamma(20, \beta)$ 分佈, 其中 β 並未給定:

69, 70, 40, 35, 59, 40, 82, 48, 49, 35, 60, 59, 55, 46,
 78, 77, 49, 60, 37, 38, 56, 27, 46, 47, 88, 44, 52, 46,
 70, 69, 77, 41, 43, 75, 70, 37, 55, 56, 62, 28, 58, 69。

(解. $Q \doteq 4.39 < 5.99 \doteq \chi_{0.95,2}^2$, 接受 H_0)

22. Davis(1952)曾給出下述汽車在馬達依序由好至壞之間所行駛的距離(單位為1,000英哩), 其中第一次表第一次壞, 餘類推:

	第一次	第二次	第三次	第四次	第五次
0 – 20	6	19	27	34	29
20 – 40	11	13	16	20	27
40 – 60	16	13	18	15	14
60 – 80	25	15	13	15	8
80 – 100	34	15	11	8	5
100 – 120	46	18	10	3	2
120 – 140	33	5	4	1	—
140 – 160	16	2	0	—	—
160 – 180	2	2	0	—	—
180 – 200	2	2	2	—	—

(i) 試在 $\alpha = 0.05$ 之下, 檢定馬達在第一次壞之前, 所行駛之距離有指數分佈;

(ii) 試在 $\alpha = 0.10$ 之下, 檢定在第四次與第五次壞之間, 所行駛的距離有指數分佈;

(iii) 試在 $\alpha = 0.10$ 之下, 檢定在第一次壞之前, 所行駛之距離有常態分佈。

(解. (i) $\hat{\lambda}^{-1} = 96.70$, $Q \doteq 235.36 > 15.51 \doteq \chi_{0.95,8}^2$, 拒絕 H_0 ; (ii) $\hat{\lambda}^{-1} = 35.65$, $Q \doteq 4.42 < 6.25 \doteq \chi_{0.90,3}^2$, 接受 H_0 ; (iii) $\hat{\mu} \doteq 96.70$, $\hat{\sigma} \doteq 37.46$, $Q \doteq 10.31 < 12.02 \doteq \chi_{0.90,7}^2$, 接受 H_0)

第10.3節

1. Kiser and Schaefer(1949)曾根據Who's Who上所載之1,436個至少曾結婚一次的婦女的資料, 得到下述統計(見Rice(1988)pp.441-443):

學歷\結婚數	1	≥ 2
有大學學位	550	61
無大學學位	681	144

試依據上表, 在 $\alpha = 0.01$ 之下, 檢定 H_0 : 是否僅結婚一次與有沒有大學學位無關。(解. $Q \doteq 16.01 > 6.63 \doteq \chi_{0.99,1}^2$, 拒絕 H_0)

2. 欲決定立法委員的黨籍, 對政府某項預算(如社會福利)是否該增加、維持、減少的意見是否無關。隨機地訪問100位立法委員, 而得下表:

意見\黨籍	甲	乙	丙
增加	8	10	12
維持	12	17	6
減少	10	13	12

試在 $\alpha = 0.10$ 之下, 檢定 H_0 : 對該項預算之意見與黨籍無關。(解. $Q \doteq 4.54 < 7.78 \doteq \chi_{0.90,4}^2$, 接受 H_0)

3. 某校欲檢定 H_0 : 某系研究所甄試的男女錄取比率相同。經統計過去10年中, 男生有360位申請, 錄取者有97位, 女生有82位申請, 錄取者有40位。試在 $\alpha = 0.01$ 之下, 做一檢定。(解. $Q \doteq 14.93 > 6.63 \doteq \chi_{0.99,1}^2$, 拒絕 H_0)
4. 欲檢定 H_0 : 高雄、台北及台中市民對某項政策之意見無差異, 分別取400, 500及400位市民。其中高雄有232位贊成(168位反對), 台北有260位贊成(240位反對), 台中有197位贊成(203位反對)。試在 $\alpha = 0.05$ 之下, 做一檢定。(解. $Q \doteq 6.48 > 5.99 \doteq \chi_{0.95,2}^2$, 拒絕 H_0)
5. 欲檢定 H_0 : 數學程度的好壞, 對統計學課程之興趣無差異, 而得下述數據:

對統計學課程之興趣\數學程度	低	普通	高
低	63	42	15
普通	58	61	31
高	14	47	29

試在 $\alpha = 0.01$ 之下, 做一檢定。(解. $Q \doteq 32.51 > 13.28 \doteq \chi_{0.99,4}^2$, 拒絕 H_0)

6. 欲檢定 H_0 : A, B, C, D 四個不同的城市, 買菜的地方(傳統市場及超級市場)無差異, 而收集到下述數據:

地方\城市	A	B	C	D
傳統市場	219	200	181	180
超級市場	39	52	89	60
其他	42	48	30	60

試在 $\alpha = 0.05$ 之下, 做一檢定。(解. $Q \doteq 38.07 > 12.59 \doteq \chi_{0.95,6}^2$, 拒絕 H_0)

7. 民國85年及90年, 某城市分別調查506及295位居民, 問其對生活環境的滿意情況。在85年有63%表示滿意, 19%表示不滿意, 18%表示不確定。在90年有83%表示滿意, 8%表示不滿意, 9%表示不確定。試在 $\alpha = 0.01$ 之下, 檢定 H_0 : 兩次調查居民的意見一致。(解. $Q \doteq 35.1 > 9.21 \doteq \chi_{0.99,2}^2$, 拒絕 H_0)

第10.4節

- 試證(4.1)式。
- 試設計出一程序, 以自隨機數表造出一0, 1的隨機數列。
- 試設計出兩種不同的程序, 以自隨機數表造出一0, 1, 2的隨機數列。
- 對表4.4, 試驗證 π 之小數前1,000位及前10,000位之 Q 值確如所給。
- 在 π 的小數前1,000位, 依序每五位一組, 觀測其中奇數的個數(如第一組14159中, 奇數有4個), 而得下述頻率:

0	1	2	3	4	5
7	31	54	61	41	6

令 N_i 表奇數個數為 i 出現的次數, $i = 0, 1, \dots, 5$ 。則 $(N_0, N_1, \dots, N_5) \sim \mathcal{M}(200, \boldsymbol{\theta})$ 分佈, 其中 $\boldsymbol{\theta} = (\theta_0, \theta_1, \dots, \theta_5)$ 。

(i) 若0, 1, $\dots$, 9為隨機地出現, 則 $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, 試給出 $\boldsymbol{\theta}_0$;

(ii) 試在 $\alpha = 0.10$ 之下, 檢定 $H_0: \boldsymbol{\theta} = \boldsymbol{\theta}_0$ 。

(解. (i) $\boldsymbol{\theta}_0 = (1/32, 5/32, 10/32, 10/32, 5/32, 1/32)$, (ii) $Q = 4.336 < 9.24 \doteq \chi_{0.90,5}^2$, 接受 H_0)

- 試證(4.4)式之 Q_1 可改寫如(4.6)式。
- 試證(4.8)式。(提示: 先證明 $P(MG \geq l) = \binom{n-(l-1)(r-1)}{r} / \binom{n}{r}, l = 1, \dots, [(n-1)/(r-1)]$)

8. 對42取6的樂透彩券, $MG = 8$ 共有 $\binom{42-7.5}{6} = \binom{7}{6} = 7$ 種情況, 試全寫出來。(解. 1, 9, 17, 25, 33, 41; 1, 10, 18, 26, 34, 42; 1, 9, 18, 26, 34, 42; 1, 9, 17, 26, 34, 42; 1, 9, 17, 25, 34, 42; 1, 9, 17, 25, 33, 42; 2, 10, 18, 26, 34, 42)
9. 試隨機地給出35組樂透彩券之號碼, 並計算 Q_1 值以檢定隨機性。
10. 在42取6的樂透彩裡, 若欲檢定成對號碼 (i, j) 是否均勻地出現, 則至少要有多少組號碼才適合以卡方來檢定。(解. 287組)
11. 試計算114年是否有一百萬個小時。(解. 可能有999,336, 999,312, 或999,288小時)
12. 在49取6的樂透彩, 試分別求頭獎、貳獎、參獎、肆獎及普獎之機率。(解. $1/a, 6/a, 252/a, 13,545/a, 246,820/a, a = \binom{49}{6}$)
13. 在 n 取1的樂透彩, 若欲全部號碼均開出, 求所需其數之期望值。(解. $1 + n \sum_{i=1}^{n-1} i^{-1}$)

參考文獻

1. 王業鈞譯(2001). 你賭對了嗎? (Bennet, D.J.原著: Randomness)。新新聞文化事業股份有限公司, 台北。
2. 黃文章(1994). 機率論講義。全華科技圖書股份有限公司, 台北。
3. 黃文章(2002). 挑戰機率。誠品好讀月刊, 91年5月第21期, 46-47。
4. 黃文章、洪宛頻及羅夢娜(2002). 樂透彩開出號碼隨機性之檢定。中國統計學報, 第40卷第3期, 249-273。
5. 葉偉文譯(2001). 統計, 改變了世界(Salsburg, D.原著: The Lady Tasting Tea)。天下遠見出版股份有限公司, 台北。

6. 趙慕嵩、張文輝及胡不歸(1987). 羅盤、神明牌、風水、獎券科。
時報周刊民國76年10月第504期, 22-28。
7. Appleton, D.R., French, J.M. and Vanderpump, M.P.J.(1996). Ignoring a covariate: an example of Simpson's paradox. *The American Statistician*, 50, 340-341.
8. Becker, H. and Piper, F.(1982). *Cipher Systems: The Protection of Communications*. Northwood Books, London, United Kingdom.
9. Boland, P.J. and Pawitan, Y.(1999). Trying to be random in selecting numbers for lotto. *Journal of Statistical Education*, Vol.7, No.3.
(這份電子期刊之網址為<http://www.amstat.org/publications/jse/>)
10. Coveyou, R.R.(1969). Random number generation is too important to be left to chance. *Applied Mathematics*, 3, 70-111.
11. Cramér, H.(1946). *Mathematical Methods of Statistics*. Princeton University Press, Princeton, New Jersey.
12. Davis, D.J.(1952). An analysis of some failure data. *Journal of the American Statistical Association*, 47, 113-150.
13. Dodge, Y.(1996). A natural random number generator. *International Statistical Review*, 64, 329-344.
14. Franklin, J.N.(1963). Deterministic simulation of random processes. *Mathematics of Computation*, 17, 28-59.
15. Gardner, M.(1966). *New Mathematical Diversions from Scientific American*. Simon and Schuster, New York.
16. Hastings, K.J.(1997). *Probability and Statistics*. Addison-Wesley Longman, Inc., Reading, Massachusetts.

17. Joe, H.(1993). Tests of uniformity for sets of lotto numbers. *Statistics and Probability Letters*, 16, 181-188.
18. Johnson, B. and Leeming, D.(1990). A study of the digits of π , e , and certain other irrational numbers. *Sankhya Series A*, 52, 183-189.
19. Kendall, M. and Stuart, A.(1979). *The Advanced Theory of Statistics*, Vol 2: *Inference and Relationship*, 4th ed. Macmillan, New York.
20. Lehmer, D.H.(1951). Mathematical methods in large scale computing units. *Annals of Harvard Computational Laboratory*, 26, 141-146.
21. Metropolis, N., Reitwiesner, G. and van Neumann, J.(1950). Statistical treatment of values of first 2000 decimals digits of e and π , calcuted on ENIAC. *M.T.A.C.*, 4, 109-111.
22. Reitwiesner, G.(1950). An ENIAC determination of π and e to more than 2000 decimal places. *M.T.A.C.*, 4, 11-15.
23. Rice, J.A.(1988). *Mathematical Statistics and Data Analysis*. Wadsworth & Brooks/Cole, Pacific Grove, California.
24. Simpson, E.H.(1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society Series B*, 13, 238-241.