

# 第一章 水文資料的統計分析

## 1.1 水文與機率

### 1.1.1 水文資料的隨機特性

許多水文資料採用統計方法分析，主要的原因是吾人對於該水文資料的了解或觀測或計算能力不足，不能以物理或數學模型描述之。若進一步檢視水文資料的變異特性，可將變異原因歸納為三點：

- (1) 水文量的驅動變數變異性太大。例如降雨—逕流系統中的降雨強度，無論是時間與空間上的變異量均甚大。
- (2) 觀測的採樣誤差太大。例如降雨量的觀測，一般使用雨量計來量測，資料記錄方式是以一小時為單位的累積量，此種降雨強度的採樣方式，在空間上而言，採樣為空間中無限多點中的數個有限點；時間上而言，一小時之內的降雨強度隨時間變化情形便無法知道。
- (3) 對於物理現象的描述與了解不足。例如降雨現象，牽涉到大尺度的天氣系統影響(如鋒面系統等)、中尺度天氣系統的影響(如鋒面系統中的中尺度對流系統)、水汽凝結與雨滴運動等的微物理作用的影響，目前仍無法同時在不同尺度上有效模擬降雨現象。

### 1.1.2 水文模式的種類

水文模式依照模式分析所使用的物理、數學與統計描述方法，可大致分為以下四類：

- (1) 物理模式：以數學模型(如控制方程式)描述水文現象的物理、化學與生化過程，例如地下水控制方程式
- (2) 概念模式：以概念上符合物理模式的簡化參數模式描述水文現象，例如線性水庫模式
- (3) 迴歸模式：利用迴歸方法表現水文現象與驅動變數的關係，例如以矩陣法求降雨—逕流模式
- (4) 隨機模式：不考慮水文現象的物理機制，將水文變數考慮為隨機變數的

分析方法，例如降雨量與流量的頻率分析

在本章中的主要是以上的第(3)、(4)兩種模式，同時使用一些在一般的基礎機率與統計課程中所未涵蓋的統計分析項目，這些項目包括：

- (1) 非常態分佈以及偏態機率分佈函數，並且通常有個下限值為零的水文資料。
- (2) 離群值(outliers)。
- (3) 同一隨機變數時間的相關特性，或是不同時間、空間或是實現值(realizations)彼此相依(不獨立)的分析方法，例如時間序列分析與克利金法空間分析。
- (4) 具有季節性變化的水文變數資料分析。
- (5) 某隨機變數( $X$ )的機率密度函數，會受到另一定率(deterministic)或隨機變數( $Y$ )影響的分析方法，例如條件機率與貝氏定理。
- (6) 受到某種檢查方法(censored data)，通過或不通過而決定記錄資料密度的資料分析問題。例如臺灣省水利處雨量站小時雨量資料，過去均採用記錄紙記錄，並以人工將部份雨量記錄輸入電腦，目前僅輸入某特定「檢查標準」以上的暴雨記錄，檢查標準為①時雨量大於 20mm，或是②日累積雨量大於 100mm；當資料達到以上兩個檢查標準中的任何一個，便將該「降雨事件」(如颱風、暴雨等)發生日期內的所有時雨量( $24 \times n$  筆)輸入。

### 1.1.3 探索(Exploratory)統計與確認(Confirmatory)統計

傳統統計方法的分析，主要著重在分析資料，並以某種方法測試統計假設(hypothesis)的成立或不成立。其常見的主要工作項目包括求出無偏估的估計、求信心區間、和測試統計假設等。傳統統計分析方法的邏輯，是分析者心中先具有某種變數或模式的假設，再蒐集變數資料，驗證模式是否如假設一般，因此又稱為「確認統計」。

當計算器與電腦逐漸發展，分析者能利用電腦處理大量資料，並且以各種統計分析圖顯示資料的特性，因此分析者通常在心中尚無「假設」時，先

根據統計分析圖「探索」資料的分佈與統計特性。圖形顯示的目的有二：一是顯示變數的數值分佈情形(例如變數分佈梯狀圖)，與兩種變數的線性與非線性相關情形(X-Y data scatter plot)；另一則是將發現的結果向他人展示。以上的前者稱為探索資料分析(exploratory data analysis, or EDA)，此一方法是由 John Tukey 所發展。探索資料分析的結果，提供資料分佈與統計特性的“初貌”(first look)。由於 EDA 的資料分析方法是以圖形分析，逐步歸納演繹得到適當的「統計假設」，而非測試既有的「統計假設」，故有別於確認統計。資料探勘(data mining)便是探索統計的一種應用，探索的方法是先找出某個或某群變數對於目標變數的相關性或解釋能力，其次再找出兩者是否有因果關係。

## 1.2 單一變數的特性分析

### 1.2.1 樣本統計值

吾人可利用少數樣本統計值大致描述樣本資料的分佈特性。樣本統計值包括平均值(mean)、中值(median)、四分值距(interquartile range)、標準偏差(standard deviation)等。常用的樣本統計值定義如表 1-1。

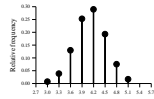
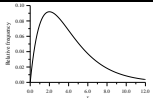
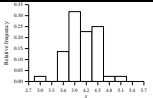
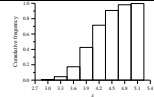
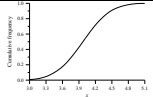
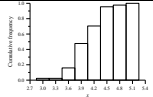
### 1.2.2 水文資料的圖形顯示

#### 一、變數分佈梯狀圖

將隨機變數樣本按照數值劃分為若干個相等區間，以變數數值為橫軸，每個數值範圍內出現的樣本個數，或是相對機率( $n_i/n$ )為縱軸，繪出的圖便是「變數分佈梯狀圖」或簡稱「梯狀圖」。

梯狀圖的數值區間選擇，Iman 和 Conover 建議，若樣本數為  $n$ ，選擇最小的整數  $k$  值，使  $2^k \geq n$  者(例如樣本數  $n=40$ ，選擇  $k=6$ )。梯狀圖雖然被廣為使用，但其缺點卻鮮少有人提到。梯狀圖的例子如圖 1-1 中的圖。圖 1-1 為淡水雨量站自西元 1960 年至 1996 年間年降雨量的兩種梯狀圖。兩個梯狀圖的資料樣本相同，但使用不同的數值區間大小，圖 1-1(a)以 150 mm 為區間，圖 1-1(b)以 300 mm 為區間。若使用大的數值區間，則分佈曲線可能會

常用的樣本統計值定義

觀念	離散分佈	連續分佈	樣本分佈
機率密度函數 (p.d.f.) 與 機率質量函數 (p.m.f.)	 機率質量函數：隨機變數等於 $k$ 的機率	 機率密度函數：累積分佈函數的一階微分 $f(x)$	 柱狀圖：隨機變數 $X$ 落於某區間內的樣本出現頻率
累積分佈函數 (c.d.f.)	 描述一隨機變數小於或等於某一特定值 $x$ 的機率	 描述一隨機變數小於或等於某一特定值 $x$ 的機率	 描述一隨機變數小於或等於某一特定值 $x$ 的樣本出現頻率
平均值 或 期望值	$\mu = \sum_{i=1}^{\infty} P(X = x_i) x_i$	$\mu = \int_{-\infty}^{\infty} x f(x) dx$	$\bar{x} = \sum_{i=1}^n \frac{x_i}{n}$
變異數	$\sigma^2 = \sum_{i=1}^{\infty} P(x_i) (x_i - \mu)^2$	$\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx$	$s^2 = \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{n-1}$
$K$ 階 中央動差值	$M_k = \sum_{i=1}^{\infty} P(x_i) (x_i - \mu)^k$	$M_k = \int_{-\infty}^{\infty} (x - \mu)^k f(x) dx$	$\tilde{M}_k = \sum_{i=1}^n \frac{(x_i - \bar{x})^k}{n-1}$
標準偏差	$\sigma \equiv \sqrt{\sigma^2}$		$s \equiv \sqrt{s^2}$
變異係數 或 相對標準偏差 (若 $\mu \neq 0$ )	$CV = \frac{\sigma}{\mu}$		$CV = \frac{s}{\bar{x}}$
偏態係數	$\gamma = \frac{M_3}{\sigma^3}$		$C_s = \frac{n \sum_{i=1}^n (x_i - \bar{x})^3}{(n-1)(n-2)s^3}$
峰度係數	$\kappa = \frac{M_4}{\sigma^4}$		$C_k = \frac{n \sum_{i=1}^n (x_i - \bar{x})^4}{(n-1)(n-2)s^4}$
$p$ 量值	$x_p$ 為 $X$ 的特定值使得 $\begin{cases} P(X < x_p) \leq p \\ P(X > x_p) \leq 1 - p \end{cases}$		$\hat{x}_p$ 為 e.d.f. 的 $p$ 分數
中值	中值 $x_{0.5}$ 為 $X$ 的任意值使得 $\begin{cases} P(X < x_p) \leq 0.5 \\ P(X > x_p) \leq 0.5 \end{cases}$		$\hat{x}_{0.5}$ ；經排序後的樣本的中間觀測值。若樣本數為偶數則為中間兩觀測值的平均值
高四分值	$x_{0.75}$		$\hat{x}_{0.75}$
低四分值	$x_{0.25}$		$\hat{x}_{0.25}$
四分值距	$x_{0.75} - x_{0.25}$		$\hat{x}_{0.75} - \hat{x}_{0.25}$

因為平滑化而失去重要資訊；若區間取得小，則會發現相鄰區間的樣本

個數，因為隨機出現的關係，出現樣本個數變異量較大的情形。無固定數值區間標準，是變數分佈梯狀圖的主要問題。

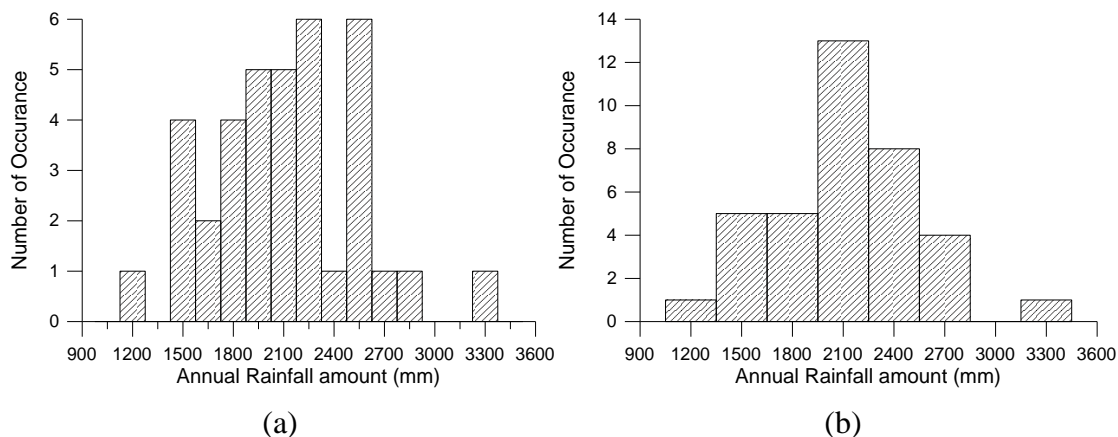


圖1-1 淡水雨量站自西元 1960 年至 1996 年間年降雨量兩種梯狀圖：  
(a)以 150 mm 為區間；(b)以 300 mm 為區間。

## 二、百分比圖(Quantile Plots)

百分比圖或累積分佈函數，顯示樣本數值分佈的累積百分比，又稱為經驗分佈函數(empirical distribution functions, or e.d.f.)。由百分比圖，樣本資料的中值(median)或高低四分值均可容易取得；由百分比圖也可以容易的看出雙峰機率等特殊變化。除此以外，因為百分比圖不需要像梯狀圖取數值區間，造成資訊的壓縮，因此較梯狀圖具優勢。

圖 1-2 將淡水雨量站自西元 1960 年至 1996 年間年降雨量以百分比圖繪出，圖中可以顯示如圖 1-1 在 3000~3150 mm 之間的空檔。若要估計一個變數值發生的機率，因為資料的資訊未遭壓縮，由百分比圖估計的結果也會優於由梯狀圖估計的結果。

繪百分比圖時，先將樣本按大小排列，最大的觀測值觀測值為  $x_{(1)}$ ，最小的為  $x_{(n)}$ ；其次，選擇一個經驗累積機率公式或點繪公式，決定各個排序樣本的累積機率或超越機率。經驗累積機率公式或點繪法公式的通式為：

$$p(X \geq x_m) = \frac{m - b}{n + 1 - 2b}$$

其中，參數  $b$  為定值，範圍介於 0 和 0.5 之間。參數  $b$  的選擇與機率密度函數有關，可參考頻率分析章節的建議。

另一種百分比圖是繪在特殊設計的機率紙上，若樣本資料點約成一直線，則該樣本屬於該種機率分佈。資料是否成一直線，可以做成一個「統計假設」，再以統計值測試假設是否通過檢定。詳細內容說明如頻率分析章節。

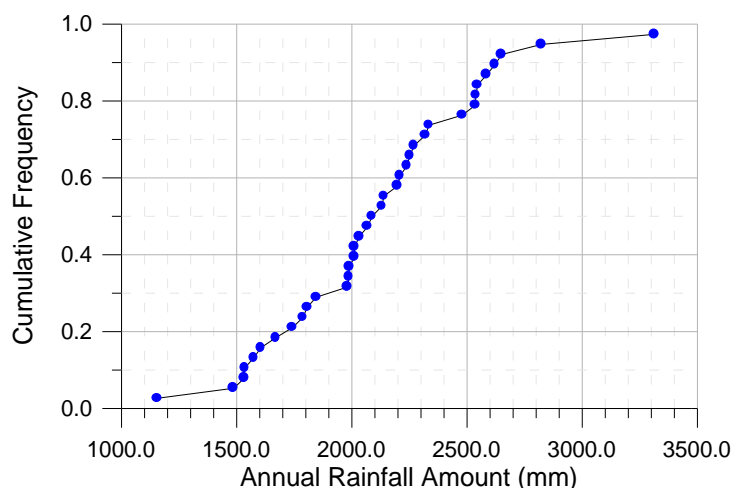


圖1-2 淡水雨量站自西元 1960 年至 1996 年間年降雨量百分比圖

### 三、方盒圖(Box Plot)

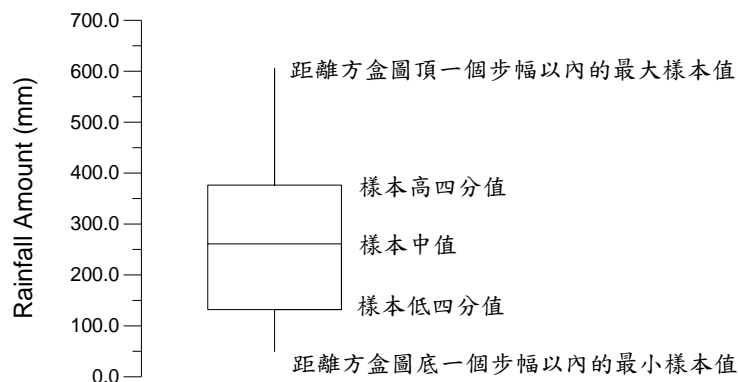
方盒圖是一個精確顯示分佈資料的繪圖方法。由方盒圖可觀察到①樣本資料的中值(median)，②樣本資料是否對稱，③是否有離群值出現。若將不同資料的方盒圖並列，可提供④不同資料分佈的初步視覺比較。如圖 1-3 所示為方盒圖解說及淡水雨量站自西元 1960 年至 1996 年間逐月降雨量方盒圖。方盒圖的繪法如下：

- (1) 方盒的中央線為樣本資料的中值，亦即資料的 50% 累積機率變數值。
- (2) 方盒的底與方盒的頂分別稱為「下關鍵點」(lower hinge)與「上關鍵點」(upper hinge)，其對應值分別為樣本資料的 25% 累積機率百分值，和樣本資料的 75% 累積機率百分值。
- (3) 方盒高差(H spread)，為 75% 累積機率百分值減 25% 累積機率百分值；「步幅」(step size)的定義為 1.5 倍的方盒高差。
- (4) 自方盒頂向上的延伸線稱為上鬚(upper whisker)，自方盒底向下的延伸

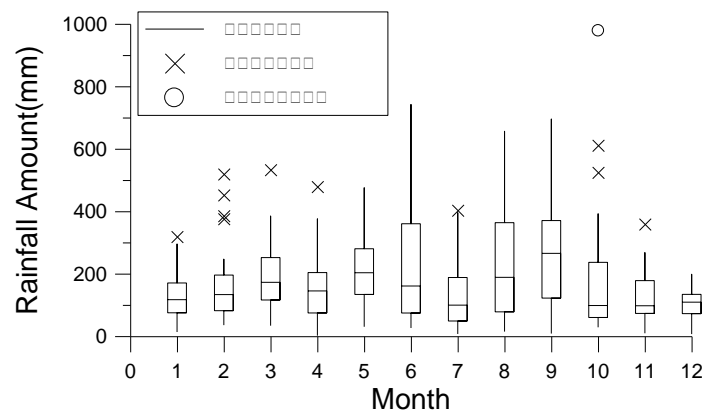
線稱為下鬚(lower whisker)；上(下)鬚延伸到不超過一個「步幅」的最大(小)樣本數值。

- (5) 超過一個「步幅」未達兩個「步幅」的外圍(outside)樣本，在圖中以“×”表示其位置。
- (6) 超過兩個「步幅」的遠外圍(far outside)樣本，在圖中以“○”表示其位置。

方盒圖可供初步視覺分析資料是否呈常態分佈。對常態分佈而言，「外圍樣本」的發生機率低於百分之一，遠外圍樣本發生的機率低於三十萬分之一。若外圍樣本或遠外圍樣本的發生機率遠大於此，表示資料的機率分佈不是常態分佈。



(a)



(b)

圖1-3 (a)方盒圖解說；(b) 淡水雨量站自西元 1960 年至 1996 年間逐月降雨量

### 1.2.3 統計假設檢定(Hypothesis Testing)

水文學中有許多可利用統計假設檢定的例子，如：

- (1) 地下水水質是否符合飲用標準？
- (2) 都市化的結果是否造成年最大洪峰流量的增加？
- (3) 汙水處理廠完成後河水中的汙染物濃度是否降低到設計標準？
- (4) 上、下兩個含水層的水力傳導係數是否相同？

在回答以上問題之前，需先將問題轉換為更嚴謹的統計假設，再利用檢定方法，由樣本是否有足夠證據否決或支持統計假設，決定檢定結果。統計假設首先確立一個測試統計值(test statistics)的「虛無假設」 $H_0$ ，以及一個對應的「替代假設」 $H_1$ 。若樣本的測試統計值偏離虛無假設，超過預設的接受統計門檻範圍，便排斥虛無假設，接受替代假設；反之，當樣本的測試統計值與虛無假設統計值的差異，不超過預設的接受統計門檻範圍，則接受虛無假設。

因為隨機變數的樣本資料具有隨機性，統計假設檢定不能證明某一個論述(假設)是正確或是錯誤，而僅是根據樣本資料的顯著傾向，選擇接受或是拒絕統計假設。根據樣本統計值所作的選擇，可能是正確的也可能是錯誤的，表 1-2 顯示統計假設檢定結果與實際情形是否符合的四種可能性，及其發生的機率。

表1-1 統計假設檢定結果及其發生機率的列聯表(contingency table)

決策	狀 態	
	$H_0$ 為真	$H_0$ 為偽
接受 $H_0$	判定正確 發生機率為 $1-\alpha$	第二型錯誤(Type II error) 發生機率為 $\beta$
拒絕 $H_0$	第一型錯誤(Type I error) 發生機率為 $\alpha$	判定正確 發生機率為 $1-\beta$

當虛無假設為正確，並且測試結果接受虛無假設的機率為  $1-\alpha$ ；發生第一型錯誤(Type I Error)的機率稱為「顯著水準」 $\alpha$ ；發生第二型錯誤(Type II



Error)的機率為 $\beta$ 。檢定效能(the power of test)的定義是當替代假設為正確，並且測試結果排斥虛無假設、接受替代假設的機率(即 $1-\beta$ )。但除非清楚定義「替代假設」，否則通常不易計算 $\beta$ ，以下利用一個例子說明。

例題1-1 一個帆布袋中有黑白兩種球，假設黑球佔 30%，即 $H_0: p=0.3$ 。試驗方法是自袋中任取樣 15 球，根據樣本中的黑球數量，決定是否接受 $p=0.3$ 的虛無假設。決策規則是：若所取的 15 個樣本中，有 2~7 顆是黑球，便接受 $p=0.3$ 的假設；若黑球數是 0~1 或 8~15 顆，便認為 $p \neq 0.3$ 。計算：

(a) 若 $p=0.3$ ，發生第一型錯誤的機率 $\alpha=?$

(b) 若實際上 $p=0.2$ 或 $p=0.4$ 時，發生第二型錯誤的機率 $\beta=?$

解：(a) 發生第一型錯誤的機會是：

$$\begin{aligned}\alpha &= P(n=0|p=0.3) + P(n=1|p=0.3) + \sum_{i=8}^{15} P(n=i|p=0.3) \\ &= \binom{15}{0} 0.3^0 0.7^{15} + \binom{15}{1} 0.3^1 0.7^{14} + \sum_{i=8}^{15} \binom{15}{i} 0.3^i 0.7^{15-i} \\ &= 0.0853\end{aligned}$$

(b) 若 $p=0.2$ ，發生第二型錯誤的機會是：

$$\beta = \sum_{i=2}^7 P(n=i|p=0.2) = \sum_{i=2}^7 \binom{15}{i} 0.2^i 0.8^{15-i} = 0.82$$

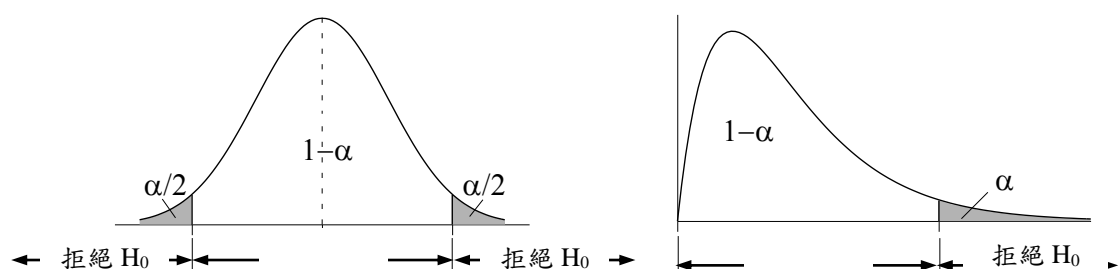
若 $p=0.4$ ，則發生第二型錯誤的機會是：

$$\beta = \sum_{i=2}^7 P(n=i|p=0.4) = \sum_{i=2}^7 \binom{15}{i} 0.4^i 0.6^{15-i} = 0.7817$$

此例題中，雖然第一型錯誤發生的機會不大( $\alpha=0.0853$ )，屬於合理範圍；但發生第二型錯誤的機會極高，因此不能算是一個好的統計假設檢定設計。要同時降低 $\alpha$ 及 $\beta$ ，必須增加樣本數量。

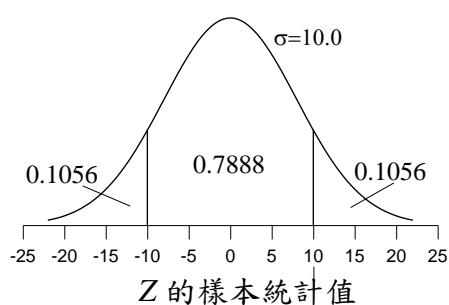
圖 1-4 介紹統計「顯著水準」(significance level)與由樣本資料「測試統計值」計算得到的「獲致顯著水準」(attained significance level)的概念。統計的顯著水準是一種客觀標準，用來決定樣本通過或是不通過檢定的一種門檻

值。舉例說明，若一變數呈常態分佈，顯著水準 $\alpha=0.05$ ，當測試統計值 $|z_s| < 1.96$ ，便通過檢定。所謂「獲致顯著水準」是由測試統計值的數值所對應的顯著水準範圍： $P(|Z| > |z_s|)$ ，「獲致顯著水準」愈高表示資料與虛無假設的情形愈接近。

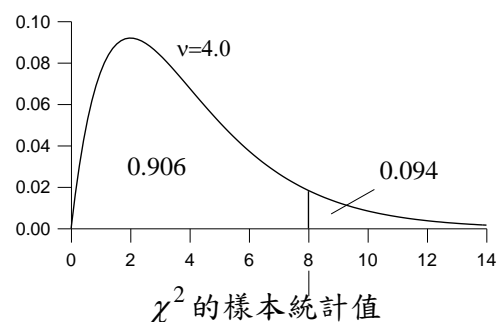


(a) 常態分佈的顯著水準 $\alpha$ 決策規則

(b) 卡方分佈的顯著水準 $\alpha$ 決策規則



(c) 獲致顯著水準 $p=0.2112$



(d) 獲致顯著水準 $p=0.094$

圖1-4 (a)(b)顯著水準 $\alpha$ 的決策規則與；(c)(d)獲致顯著水準 $p$

## 1.2.4 參數化與非參數化統計檢定

若統計假設檢定牽涉到指定資料的機率分佈，或使用變數值計算測試統計值，便屬於參數化檢定；若不指定資料的機率分佈，同時只使用排序的序號或分類個數計算測試統計值者，便是非參數化檢定。若假設的機率密度函數是正確的，則使用參數化檢定的檢定效能會略高於使用非參數化檢定的檢定效能；通常在樣本資料數增加時，兩者的表現會逐漸接近；若假設的機率密度函數與實際的機率密度函數相去甚遠，則非參數化檢定可能比參數化檢定的效能要高(more powerful)許多。不幸的是，許多機率密度函數測試的檢定水準，對於選擇錯誤的模式套配的誤差均不甚敏感，使得發生第二型錯誤的機會甚大。因此，除非有足夠大量的資訊能夠確定隨機變數的機率密度函

數，否則以參數化的統計假設檢定方法，選擇特定機率密度函數，是一個錯誤風險甚高的方法。

非參數化檢定特別適用於尾端發生比例高(heavy tail)的樣本，原因是與參數化檢定比較，非參數化檢定測試的統計值受到離群值的影響較低。參數化模式在多變數分析(multivariate)時，因為模式的多變性與分析的方便性，比非參數化檢定法較具優勢。

### 一、變數超過某特定數值機率的測試(非參數化檢定)

$$H_0 : P(A) = p \quad A = \{X > x\} \text{ or other events}$$

$$H_1 : P(A) \neq p$$

此測試不牽涉到機率密度函數的假設，也不牽涉到每個樣本數值分佈的統計，只是計算比  $x$  大的樣本個數，故屬於非參數化檢定。測試的統計值  $T$  是個數  $n$  的樣本中，大於  $x$  資料的個數。決定通過或不通過檢定的標準是：若  $T < T_L$  或  $T > T_U$ ，則排斥  $H_0$ 、接受  $H_1$ ；若  $T_L < T < T_U$  則接受  $H_0$ 、排斥  $H_1$ 。若樣本數小於 20，則如例題 1.1，利用二項(binomial)分佈計算。對於樣本數較大者，二項分佈逐漸趨近於常態分佈，可以採用常態分佈的測試值近似之。

$$T_L = np + z_{\alpha/2} \sqrt{np(1-p)}$$

$$T_U = np + z_{1-\alpha/2} \sqrt{np(1-p)}$$

其中， $z_{\alpha/2}$  與  $z_{1-\alpha/2}$  分別是標準常態分佈的  $\alpha/2$  百分比與  $1-\alpha/2$  百分比的數值。

### 二、樣本平均值測試(參數化檢定)

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu \neq \mu_0$$

測試的統計值為：

$$T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

以上變數呈 Student  $t$  分佈，自由度(degree of freedom)= $n-1$ ， $n$  為樣本數。決策的規則是：

$$\text{當 } |T| \geq t(1-\alpha/2, n-1) \Rightarrow \text{排斥 } H_0$$

$$\text{當 } |T| < t(1-\alpha/2, n-1) \Rightarrow \text{接受 } H_0$$

### 三、樣本中值測試(非參數化檢定)

$$H_0: \text{分佈群中值} = M$$

$$H_1: \text{分佈群中值} \neq M$$

其中  $M$  為假設的變數中值，必須由測試樣本以外的資訊決定（獨立資訊）。測試的統計值為：

$$T = \text{樣本中比 } M \text{ 小的樣本個數}$$

決策的規則是，在樣本數  $n \geq 20$  時：

$$\text{當 } \left| T - \frac{n}{2} \right| \geq z_{1-\alpha/2} \times \frac{\sqrt{n}}{2} \Rightarrow \text{排斥 } H_0$$

$$\text{當 } \left| T - \frac{n}{2} \right| < z_{1-\alpha/2} \times \frac{\sqrt{n}}{2} \Rightarrow \text{接受 } H_0$$

在隨機變數的分佈不是常態分佈，並且與常態分佈型態差距較大時，樣本中值測試比樣本平均值測試的第二型錯誤發生機率低，因此檢定效能較高。

## 1.3 群與群的比較

### 1.3.1 配對樣本的比較

群與群的比較是測試兩種樣本，是否有一群樣本的某種統計值有高於另一群樣本的同一統計值。所謂配對樣本，指的是兩種樣本的觀測是配對出現的，樣本為  $(x_i, y_i), i = 1, 2, \dots, n$ 。對於配對樣本，統計假設測試的虛無假設可能是平均值或中值相等，或是兩者的差值為零，意即  $D_i = x_i - y_i = 0$ 。例如：

(1) 在使用人工肥料的田地，地下水水質可能受到影響，耕種季節前與耕種

季節後，在區域淺層地下水井量測的地下水硝酸鹽濃度分別為  $x_i$  與  $y_i$ ，測試值為  $D_i = x_i - y_i$ 。

- (2) 兩種不同的流量量測方法，在同一位置、同一時間，使用兩種方法量測到的流量分別為  $x_i$  與  $y_i$ 。若假設兩種量測方法的誤差是屬於差值性的 (additive)，則測試值為  $D_i = x_i - y_i$ ；若假設兩種量測方法的誤差是屬於倍數性的 (multiplicative)，則測試值可定義為  $D_i = \log(x_i/y_i) = \log(x_i) - \log(y_i)$ 。

### 一、正負序號測試(非參數化檢定)

$H_0$ ： $x$  與  $y$  來自同一機率分佈群(population)

$H_1$ ： $x$  與  $y$  不是來自同一機率分佈群

正負序號測試屬於一種非參數化檢定，其分析步驟如下：

- (2) 去除樣本中所有的  $D_i=0$  樣本資料
- (3) 定義去除  $D_i=0$  樣本後的樣本資料個數為  $n$ ，樣本個數需大於 15，測試統計值趨近於常態分佈時，方可使用正負序號測試法。
- (4) 按  $D_i$  絕對值的大小排序，最小的  $|D_i|$  值排在第 1 個，最大的  $|D_i|$  值排在第  $n$  個
- (5) 計算所有  $D_i>0$  數字的序號的和， $R^+$ ；以及所有  $D_i<0$  數字的序號的和， $R^-$
- (6) 測試的統計值  $W^+ = \min(R^+, R^-)$
- (7) 樣本數  $n$  的樣本，計算虛無假設  $W^+$  的理論平均值與標準偏差：
$$\mu_{W^+} = \frac{n(n+1)}{4} \quad \text{與} \quad \sigma_{W^+} = \sqrt{\frac{n(n+1)(2n+1)}{24}}$$
- (8) 計算標準常態分佈的測試統計值

$$z_{sr^+} = \begin{cases} \frac{W^+ - \frac{1}{2} - \mu_{W^+}}{\sigma_{W^+}} & \text{if } W^+ > \mu_{W^+} \\ 0 & \text{if } W^+ = \mu_{W^+} \\ \frac{W^+ + \frac{1}{2} - \mu_{W^+}}{\sigma_{W^+}} & \text{if } W^+ < \mu_{W^+} \end{cases}$$

(9) 當  $|z_{sr^+}| \geq z_{1-\alpha/2} \Rightarrow$  排斥  $H_0$

當  $|z_{sr^+}| < z_{1-\alpha/2} \Rightarrow$  接受  $H_0$

## 二、配對 $t$ 測試(參數化檢定)

配對  $t$  測試是最常用的一種配對測試，因為假設測試統計值  $D_i = x_i - y_i$  呈常態分佈，因此是一種參數化的測試，測試的步驟如下：

(1) 計算樣本資料所有配對的  $D_i = x_i - y_i$  數值

(2) 計算  $D_i$  的平均值  $\bar{D}$

(3) 計算  $D_i$  的樣本變異數  $S^2 = \sum_{i=1}^n \frac{(D_i - \bar{D})^2}{n-1}$

(4) 計算標準化的  $t$  統計值  $t_p = \frac{\bar{D}}{\sqrt{S^2/n}}$

(5) 當  $|t_p| \geq t_{1-\alpha/2, n-1} \Rightarrow$  排斥  $H_0$

(6) 當  $|t_p| < t_{1-\alpha/2, n-1} \Rightarrow$  接受  $H_0$

表 1-3 為總結兩組配對樣本不同測試方法的適用情況與檢定效能。

表1-2兩組配對樣本不同測試方法的適用情況與其檢定效能

測試	$H_0$	適用情況與檢定效能
配對 $t$ 測試	$x_i - y_i$ 差值的平均值為零的常態分佈	若差值呈常態分佈時檢定效能最高 若資料有偏態或有遠離群值則不應使用
正負序號測試	$x_i - y_i$ 差值的中值為零的對稱分佈	若差值非呈常態分佈時較 $t$ 測試的檢定效能高

### 1.3.2 兩種獨立隨機變數或兩種獨立樣本的比較

此處所謂「兩種獨立隨機變數」或「兩種獨立樣本」，指的是兩種變數採樣沒有一對一對應的配對關係，測試的目的是比較兩組數據的統計值是否相當。舉例說明「兩種獨立樣本」：一個河川的集水區經過都市化的過程，欲比較開發前與開發後的洪峰流量，了解是否有增加的現象，開發前、後在同一測站取得的流量資料沒有配對關係。

#### 一、序號總和測試(Rank-Sum Test，非參數化檢定)

序號總和測試、Wilcoxon rank-sum test、Mann-Whitney test 與 Wilcoxon-Mann-Whitney rank-sum test 均為相同的測試，主要的測試目的是：兩組樣本的分佈是否相同。虛無假設是兩者相同，替代假設是兩者不相同。序號總和測試使用的是序號，而非數值，故為非參數化檢定，對樣本資料做單調轉換(monotonic transformation)，例如對數轉換，不會影響此一測試的結果。測試的步驟如下：

- (1) 將兩種樣本資料混合後排序，假設兩組資料中樣本數較多的一組有  $m$  筆資料，樣本數較少的一組有  $n$  筆資料，混合排序的總樣本數為  $N=n+m$ ，最小的數字序號為 1，最大樣本的序號為  $N$ ，若有兩個以上的數值相等，使用序號的平均。例如混合樣本由小至大排序為 77、78、78、80、...，則兩個 78 的序號皆為 2.5。
- (2) 計算樣本個數較少的一組(資料數= $n$ )所有樣本的序號總和  $W$
- (3) 計算  $W$  與理論的平均值的差距，再除以理論標準偏差，得到理論上為標準常態分佈的測試統計值  $z_{rs}$ ：

$$z_{rs} = \begin{cases} \frac{W - \frac{1}{2} - \mu}{\sigma} & \text{if } W > \mu \\ 0 & \text{if } W = \mu, \mu = \frac{n(N+1)}{2}, \sigma = \sqrt{\frac{mn(N+1)}{12}} \\ \frac{W + \frac{1}{2} - \mu}{\sigma} & \text{if } W < \mu \end{cases}$$

若樣本中，數值相等的組數比例非為極少數，則必須使用以下公式估計

$$\text{標準偏差：}\sigma = \sqrt{\frac{mn}{N(N-1)} \sum_{k=1}^N R_k^2 - \frac{nm(N+1)^2}{4(N-1)}}$$

其中  $R_k$  是排序第  $k$  個樣本在全體  $N$  個樣本中的序號，因為有部分樣本數值相同，所以會有若干個  $i \neq j$  樣本的序號相同  $R_i = R_j$ 。

- (4) 若兩組樣本的樣本個數皆大於 10，並且  $|z_{rs}| > z_{1-\alpha/2}$  則排斥虛無假設  $H_0$ 。  
若至少有一組樣本的樣本個數小於 10，使用以上常態分佈的近似法會引致較大的誤差，應該使用精確解。

## 二、兩組樣本 $t$ 測試(Two-Sample $t$ Test，參數化檢定)

「兩組樣本  $t$  測試」是最常用來比較兩種獨立隨機變數的統計測試方法。此一測試的虛無假設是：兩種樣本的平均值相等，替代假設則是：兩個變數的平均值不相等。在兩種情形下「兩組樣本  $t$  測試」不如「序號總和測試」。一是當兩個變數不是常態變數時，此一測試方法不如「序號總和測試」的檢定效能高；另一是「兩組樣本  $t$  測試」適用於分辨兩者是否增減一常數值，不適用於分辨呈一比例增減關係的兩種樣本。若兩種樣本屬於非常態分佈或是兩群組呈比例增減兩種情形中的任一種，應該使用「序號總和測試」，不應使用「兩組樣本  $t$  測試」。兩組樣本  $t$  測試的測試步驟如下：

- (1) 計算兩組樣本資料的樣本平均值與標準偏差， $\bar{x}$ 、 $s_x$ 、 $\bar{y}$  與  $s_y$ ；其中  $x$  為第一變數，樣本數為  $n$ ； $y$  為第二變數，樣本數為  $m$ 。
- (2) 計算測試的自由度。若通過  $s_x = s_y$  的統計假設檢定，則自由度  $df = n + m - 2$ ；若兩個樣本的標準偏差顯著的不同，則使用以四捨五入取整數計算自由

$$\text{度：} df = \frac{\left( \frac{s_x^2}{n} + \frac{s_y^2}{m} \right)^2}{\frac{(s_x^2/n)^2}{n-1} + \frac{(s_y^2/m)^2}{m-1}}$$

- (3) 計算測試統計值  $t$ ，
$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}}$$

- (4) 若  $|t| > t_{1-\alpha/2, df}$ ，則排斥兩組樣本平均值相等的虛無假設。



表 1-4 總結兩組獨立樣本不同測試方法的適用情況與測試的效力。

表1-3 兩組獨立樣本不同測試方法的適用情況與檢定效能

測試	$H_0$	適用情況與檢定效能
兩組樣本 $t$ 測試	兩組資料的平均值相同	假設兩分佈群為具相同變異數的常態分佈 若偏離常態分佈則會影響檢定效能
序號總和測試	兩組資料的中值相同	假設兩分佈群的機率(密度)函數相同

### 1.3.3 數種樣本的比較

一個變數可能受到不同驅動因素的影響，而呈現出不同的統計分佈特性，例如同一雨量站量測的日降雨量，可能會因為降雨型態、季節而不同。將日降雨量資料按照不同降雨型態或季節分類，欲比較各不同樣本之間的統計分佈特性問題，可用以下樣本測試方法測試之。

數種樣本的比較，可同樣分為參數化測試方法與非參數化測試方法。測試數種樣本平均值是否相等的一種方法是「變異數分析法」(Analysis of Variance, or ANOVA)，測試數種樣本的機率分佈是否相同的非參數化測試方法是 Kruskal- Wallis 測試。和以上各單元相同，當變數的分佈偏離常態分佈時，非參數化檢定方法優於參數化檢定方法，即檢定效能更高。

#### 一、單一變因樣本差異測試(Tests for Differences Due to One Factor)

資料因為某單一變因的影響，分為  $k$  個樣本群組，每組樣本群有  $n_j$  個樣本資料， $j = 1, 2, \dots, k$ 。觀測樣本  $y_{ij}$  是第  $j$  個樣本群組中的第  $i$  筆資料，資料總數為  $N$

$$N = \sum_{j=1}^k n_j$$

若每組資料的個數均相等，並且個數為  $n$ ，則  $N = k \times n$ 。

#### (一) 變異數分析法 (ANOVA)

變異數分析法藉由群組樣本平均值的權重變異數與所有樣本的變異數的比較，判斷各群組的平均值是否相等，為一種參數化檢定。測試的虛無假設為各群組的平均值相等，替代假設是不相等。測試的步驟如

下：

- (1) 計算每個樣本群各自的平均值  $\bar{y}_j$  ( $j=1,2,\dots,k$ ) 及所有樣本的總平均值  $\bar{y}$ ：

$$\bar{y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij} \quad \bar{y} = \frac{1}{N} \sum_{j=1}^k \sum_{i=1}^{n_j} y_{ij}$$

- (2) 計算群組樣本平均值總變異數  $SST$  (total sum of squares)，分析所解釋的變異數  $SSA$  (treatment sum of squares) 與殘餘誤差變異數  $SSE$  (error sum of squares)：

$$SST = \frac{1}{N} \sum_{j=1}^k \sum_{i=1}^{n_j} (y_{ij} - \bar{y})^2$$
$$SSA = \frac{1}{N} \sum_{j=1}^k n_j (\bar{y}_j - \bar{y})^2$$
$$SSE = \frac{1}{N} \sum_{j=1}^k \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2$$

若資料兩兩不相關，以上三者具有以下全等定理 (Sum-of-Squares Identity)：

$$\sum_{j=1}^k \sum_{i=1}^{n_j} (y_{ij} - \bar{y})^2 = \sum_{j=1}^k n_j (\bar{y}_j - \bar{y})^2 + \sum_{j=1}^k \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2$$

- (3)  $SSA$  的自由度為  $k-1$ ， $SSE$  的自由度為  $N-k$ ， $SST$  的自由度為  $N-1$ 。將  $SSE$  與  $SST$  分別除以其自由度，計算  $MSE$  與  $MST$ ：

$$MSA = \frac{SSA}{k-1}, \quad MSE = \frac{SSE}{N-k}, \quad MST = \frac{SST}{N-1}$$

- (4) 計算測試統計值

$$f = \frac{MSA}{MSE} \sim F(k-1, N-k)$$

- (5) 查  $F$  分佈，分子自由度為  $k-1$ ，分母自由度為  $N-k$  的表，找出顯著水準為  $\alpha$  的檢定門檻值， $F(1-\alpha, k-1, N-k)$ 。判斷測試統計值是否大於檢定門檻值，若  $f \geq F(1-\alpha, k-1, N-k)$  則排斥各群組平均值相等的虛無假設，接受替代假設。

## (二) Kruskal-Wallis 測試

Kruskal-Wallis 測試的虛無假設為各群組的機率分佈相等，替代假設是不相等。Kruskal-Wallis 測試與其他非參數化測試相同，因為用的是序號計算測試統計值，為非參數化檢定方法，若對樣本做單調函數轉換，不會影響到測試的結果。Kruskal-Wallis 測試的步驟如下：

- (1) 將所有群組樣本混合後排序，使最小的數字序號為 1，最大樣本的序號為  $N$ 。若有兩個以上的數值相等，各相等數值均使用序號的平均值（參考序號總和測試）。
- (2) 計算每一群組的序號平均值  $\bar{R}_j$ 。
- (3) 計算測試統計值  $KW$ ：

$$KW = \frac{12}{N(N+1)} \sum_{j=1}^k n_j \left( \bar{R}_j - \frac{N+1}{2} \right)^2$$

- (4) 查  $\chi^2$  分佈，自由度為樣本組數減一， $k-1$ ，找出顯著水準為  $\alpha$  的檢定門檻值， $\chi^2_{1-\alpha, k-1}$ 。若  $KW > \chi^2_{1-\alpha, k-1}$ ，則排斥虛無假設，接受替代假設。

當群組資料個數太少時(例如三個群組，每個群組內的樣本數少於等於 5 個；或是四個群組，每個群組內的樣本數少於等於 4 個)，不能使用  $\chi^2$  分佈近似的 Kruskal-Wallis 測試，必須使用直接精確解。

## 二、多變因樣本差異測試(Tests for Effects of More Than One Factor)

當隨機變數可能受到多種變因影響時，兩個不同變因可能彼此相關，因此要分析第二個變因的影響時，必須將第一個變因的影響先予去除。使用 ANOVA 兩次，分別分析兩種變因的影響，未將其他變因的影響去除，可能會得到錯誤的結論。分析多變因對某種樣本的一個可行的方案，是將所有可能變因的影響一起分析；另一種分析方法稱為「階層變異數分析」(Factorial ANOVA)。進行多變因分析時，可能不只想知道不同群樣本的分佈是否相同，往往想同時瞭解是哪一群組與其他群組不同，例如：

**Group A  $\approx$  Group B  $\ll$  Group C**

進行多變因分析之前，應該先利用 ANOVA 或 Kruskal-Wallis 法進行單變因影響的分析，若發現沒有差異，則不需繼續測試多變因分析。若檢定結果發現各樣本群組的差異顯著，沒有接受虛無假設時，再進行多變因分析。多變因分析需要決定一個「最小顯著範圍」，最小顯著範圍與兩個樣本群組的顯著水準相當不同，是隨機變數的變異數，以及樣本數的函數。另一些分析方法採用兩兩比較逐步測試法，舉例說明，若比較兩組樣本使用的顯著水準為  $\alpha_p=0.05$ ，並且有六組樣本，則兩兩比較的對數共有  $6 \times 5 / 2 = 15$  組，則全部 15 組測試，不發生第一型錯誤的機率，是每組兩兩比較差異均落在信心區間之內。一次試驗不發生第一型錯誤的的機率是  $1 - \alpha_p$ ，15 次兩兩比較的差異每次均落在信心區間之內的機會是  $(1 - \alpha_p)^{15}$ ，至少有一次落在信心區間之外的機會，或是發生第一型錯誤的機會是  $\alpha_o = 1 - (1 - \alpha_p)^{15} = 0.54$ 。在許多情形下，兩兩比較的  $\alpha_p$  與全體比較的  $\alpha_o$  有相當大的不同。若要較為精確的全體比較的  $\alpha_o$ ，可參考 Tukey 的測試。

## 1.4 常態分佈隨機變數(Gaussian Random Variables)

### 1.4.1 常態分佈隨機變數的特性

分析水文資料常是建立統計數學模型，例如迴歸模型和時間序列模型等，再透過模型遂行內插、補遺、估計和外延預報等應用目標。統計數學模型多建立在隨機變數為常態性分佈的假設上，原因是當資料呈常態分布時，數學模型只需處理一、二階動差值，即期望值與共變數，便可達到和處理整個機率密度函數的相同效果，和其他分佈比較，具有簡單、便利的優勢。

一隨機變數  $X$ ，若其發生機率與隨機變數之關係可表示為以下的關係，則此變數稱作常態分佈之隨機變數：

$$p[X = x] = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x - \mu)^2}{2\sigma^2}\right]$$

若一個向量代表  $n$  個常態隨機變數  $X$  的聯合常態分佈(jointly Gaussian distribution)，則：

$$p_{\mathbf{X}}(\mathbf{x}) = \frac{\exp\left[-\frac{1}{2} \cdot (\mathbf{x} - \mathbf{m})^T \Lambda^{-1} (\mathbf{x} - \mathbf{m})\right]}{(2\pi)^{n/2} |\Lambda|^{1/2}}$$

其中， $\Lambda$  為共變數矩陣， $\Lambda_{ij} = E[(x_i - \mu_i)(x_j - \mu_j)]$ ， $|\Lambda|$  為  $\Lambda$  矩陣的行列式值(determinant)。以上聯合常態分佈最簡單的就是二變數的聯合常態分佈：

$$P_{X,Y}(x, y) = \frac{\exp\left[-\frac{(x - \mu_X)^2 \sigma_Y^2 - 2(x - \mu_X)(y - \mu_Y)\rho_{XY}\sigma_X\sigma_Y + (y - \mu_Y)^2 \sigma_X^2}{2\sigma_X^2 \sigma_Y^2 (1 - \rho_{XY}^2)}\right]}{2\pi \cdot \sigma_X \sigma_Y (1 - \rho_{XY}^2)^{1/2}}$$

二變數的機密度函數之機率密度最高發生在  $(\mu_X, \mu_Y)$  處。等機率密度之等高線(contour)均為橢圓形，其方向和長短軸半徑由  $\sigma_X$ 、 $\sigma_Y$  和  $\rho_{XY}$  決定。圖 1-10 為以隨機變數之統計值為  $\mu_X = 60$ 、 $\mu_Y = 40$ 、 $\sigma_X = 35$ 、 $\sigma_Y = 25$  及  $\rho_{XY} = 0.8$  為例所繪製之二變數的聯合常態分佈等機率線圖，及以  $\Delta x = 5$  與  $\Delta y = 5$  之三維聯合常態分佈機率圖。若隨機變數  $X$  和  $Y$  兩者為獨立，則二獨立變數聯合常態分佈的橢圓形等機率線的長、短軸半徑會與座標軸相同，如圖 1-5。

#### 1.4.2 條件機率密度函數(Conditional p.d.f.)

若已知隨機變數實現值  $X = x$ ，則常態分佈隨機變數  $Y$  的條件機率密度函數為：

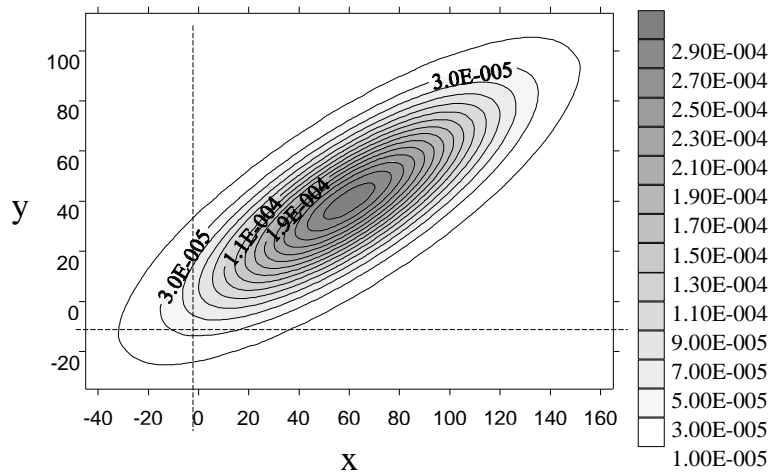
$$P[Y = y | X = x] = \frac{P(X = x, Y = y)}{P(X = x)} = \frac{\exp\left[-\frac{[(y - \mu_Y) - (\sigma_Y \cdot \rho_{xy} / \sigma_X)(x - \mu_X)]^2}{2\sigma_Y^2(1 - \rho_{xy}^2)}\right]}{\sqrt{2\pi} \cdot \sigma_Y (1 - \rho_{xy}^2)^{1/2}}$$

從以上條件機率密度函數，可以讀出和理解，常態分佈隨機變數  $Y$  的期望值和估計變異數已經分別改變成為：

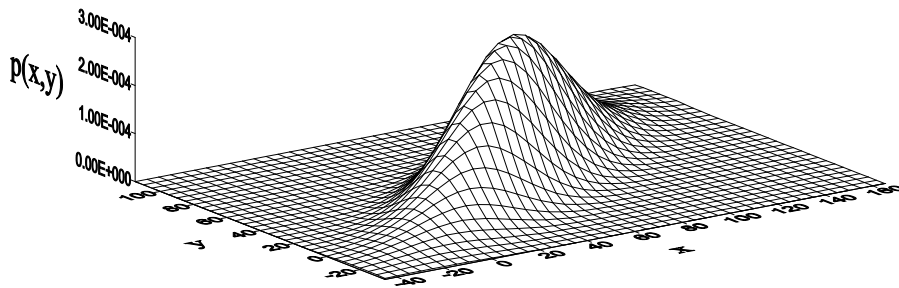
$$E[Y | X = x] = \int_{-\infty}^{\infty} y P(y | x) dy = \mu_Y + \frac{\rho_{xy} \sigma_Y}{\sigma_X} \cdot (x - \mu_X)$$

$$\text{var}[Y | X = x] = \sigma_Y^2 (1 - \rho_{xy}^2)$$

以上兩個估計也可以說是：已知  $X = x$  的情況下， $Y$  的最佳估計和估計誤差變異數。



(a)



(b)

圖1-5 (a)隨機變數之統計值為  $\mu_x = 60$ 、 $\mu_y = 40$ 、 $\sigma_x = 35$ 、 $\sigma_y = 25$  及  $\rho_{xy} = 0.8$  之二變數聯合常態分佈等機率線圖；(b)  $\Delta x = 5$  與  $\Delta y = 5$  之聯合常態分佈機率

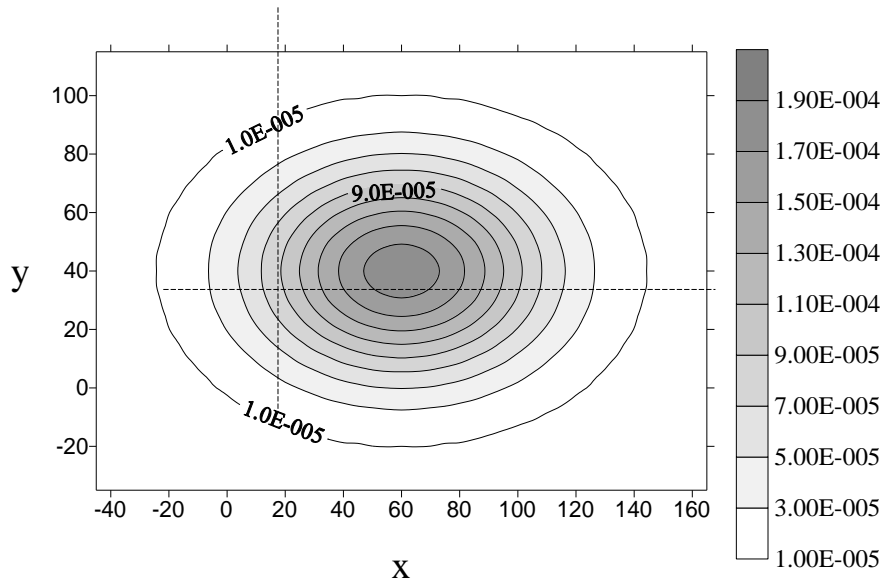


圖1-6 隨機變數之統計值為  $\mu_x = 60$ 、 $\mu_y = 40$ 、 $\sigma_x = 35$  及  $\sigma_y = 25$  之獨立二變數聯合常態分佈等機率線圖

例題1-2 假設甲河年平均流量  $X$  和乙河年平均流量  $Y$  兩個隨機變數之間的關係為： $\mu_X = 60 \text{ cms}$ 、 $\mu_Y = 40 \text{ cms}$ 、 $\sigma_X = 35$ 、 $\sigma_Y = 25$  及  $\rho_{XY} = 0.8$ ，若乙河某年資料遺失，該年甲河流量為  $65 \text{ cms}$ ，試估計該年乙河之平均流量。

$$\text{解} \quad E[y|x] = \mu_Y + \rho_{XY} \frac{\sigma_Y}{\sigma_X} \cdot (x - \mu_X) = 40 + 0.8 \times \frac{25}{35} (65 - 60) = 42.86 \text{ cms}$$

$E[]$

若  $\mathbf{Z}$  為常態分佈隨機變數向量， $\mathbf{Z} \in N(\mathbf{m}_Z, \Lambda_Z)$ ，且  $\mathbf{Z} = \begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix}$ ，

其中  $\mathbf{X} = [X_1 \ X_2 \ \cdots \ X_m]^T$ ， $\mathbf{Y} = [Y_1 \ Y_2 \ \cdots \ Y_n]^T$ ，平均值向量為  $\mathbf{m}_Z = \begin{bmatrix} \mathbf{m}_X \\ \mathbf{m}_Y \end{bmatrix}$ ，同時共變數矩陣為  $\Lambda_Z = \begin{bmatrix} \Lambda_{XX} & \Lambda_{XY} \\ \Lambda_{YX} & \Lambda_{YY} \end{bmatrix}$ ，則條件機

率的平均值向量及共變數矩陣為

$$\mathbf{m}_{Y|X} = E[\mathbf{Y} = \mathbf{y} | \mathbf{X} = \mathbf{x}] = \Lambda_{YX} \cdot \Lambda_{XX}^{-1} \cdot (\mathbf{x} - \mathbf{m}_X) + \mathbf{m}_Y$$

$$\Lambda_{Y|X} = E[(\mathbf{Y} - \mathbf{m}_{Y|X})(\mathbf{Y} - \mathbf{m}_{Y|X})^T | \mathbf{X} = \mathbf{x}] = \Lambda_{YY} - \Lambda_{YX} \cdot \Lambda_{XX}^{-1} \cdot \Lambda_{YX}^T$$

### 1.4.3 動差階層化(Moment Factoring)

若  $[X_1, X_2, \dots, X_n]^T$  為一組常態分佈隨機變數的向量，其期望值向量為  $[\bar{X}_1, \bar{X}_2, \dots, \bar{X}_n]^T$ ，共變數矩陣為  $\Lambda_{XX}$ ，則

$$E[(X_1 - \bar{X}_1)(X_2 - \bar{X}_2) \cdots (X_L - \bar{X}_L)] = \begin{cases} 0, & \text{若 } L \text{ 為奇數} \\ \sum \lambda_{j_1 j_2} \cdot \lambda_{j_3 j_4} \cdots \lambda_{j_{L-1} j_L} & \text{若 } L \text{ 為偶數} \end{cases}$$

以上  $L$  為偶數時， $\Sigma$ 指的是所有可能的不同組合的總和，不同組合的個數為  $(L-1) \times (L-3) \times \dots \times 3 \times 1$ ，以例子說明。

(1)  $L=3$  的情形

$$E[(X_1 - \bar{X}_1)(X_2 - \bar{X}_2)(X_3 - \bar{X}_3)] = 0$$

$$E[(X_1 - \bar{X}_1)^2(X_3 - \bar{X}_3)] = 0$$

(2)  $L=4$  的情形

$$E[(X_1 - \bar{X}_1)(X_2 - \bar{X}_2)(X_3 - \bar{X}_3)(X_4 - \bar{X}_4)] = \lambda_{12} \cdot \lambda_{34} + \lambda_{23} \cdot \lambda_{14} + \lambda_{13} \cdot \lambda_{24}$$

$$E[(X_1 - \bar{X}_1)^2(X_2 - \bar{X}_2)(X_3 - \bar{X}_3)] = 2\lambda_{12} \cdot \lambda_{13} + \lambda_{11} \cdot \lambda_{23}$$

$$E[(X_1 - \bar{X}_1)^2(X_2 - \bar{X}_2)^2] = \lambda_{11} \cdot \lambda_{22} + 2\lambda_{12}^2$$

$$E[(X_1 - \bar{X}_1)(X_2 - \bar{X}_2)^3] = 3\lambda_{12} \cdot \lambda_{22}$$

$$E[(X_1 - \bar{X}_1)^4] = 3\lambda_{11}^2$$

(2)  $L=6$  的情形

$$\begin{aligned} E[(X_1 - \bar{X}_1)(X_2 - \bar{X}_2)(X_3 - \bar{X}_3)(X_4 - \bar{X}_4)(X_5 - \bar{X}_5)(X_6 - \bar{X}_6)] \\ = \lambda_{12}\lambda_{34}\lambda_{56} + \lambda_{13}\lambda_{24}\lambda_{56} + \lambda_{14}\lambda_{23}\lambda_{56} + \lambda_{15}\lambda_{23}\lambda_{46} + \lambda_{16}\lambda_{23}\lambda_{45} \\ + \lambda_{12}\lambda_{35}\lambda_{46} + \lambda_{13}\lambda_{25}\lambda_{46} + \lambda_{14}\lambda_{25}\lambda_{36} + \lambda_{15}\lambda_{24}\lambda_{36} + \lambda_{16}\lambda_{24}\lambda_{35} \\ + \lambda_{12}\lambda_{36}\lambda_{45} + \lambda_{13}\lambda_{26}\lambda_{45} + \lambda_{14}\lambda_{26}\lambda_{35} + \lambda_{15}\lambda_{26}\lambda_{34} + \lambda_{16}\lambda_{25}\lambda_{34} \end{aligned}$$

$$\begin{aligned} E[(X_1 - \bar{X}_1)^2(X_2 - \bar{X}_2)^4] \\ = \lambda_{12}\lambda_{12}\lambda_{22} + \lambda_{11}\lambda_{22}\lambda_{22} + \lambda_{12}\lambda_{12}\lambda_{22} + \lambda_{12}\lambda_{12}\lambda_{22} + \lambda_{12}\lambda_{12}\lambda_{22} \\ + \lambda_{12}\lambda_{12}\lambda_{22} + \lambda_{11}\lambda_{22}\lambda_{22} + \lambda_{12}\lambda_{22}\lambda_{12} + \lambda_{12}\lambda_{22}\lambda_{12} + \lambda_{12}\lambda_{22}\lambda_{12} \\ + \lambda_{12}\lambda_{12}\lambda_{22} + \lambda_{11}\lambda_{22}\lambda_{22} + \lambda_{12}\lambda_{22}\lambda_{12} + \lambda_{12}\lambda_{22}\lambda_{12} + \lambda_{12}\lambda_{22}\lambda_{12} \\ = 3\lambda_{11}\lambda_{22}^2 + 12\lambda_{12}^2\lambda_{22} \end{aligned}$$

#### 1.4.4 常態性測試

常態分佈有其理論背景，亦即根據中央極值定理(central limit theorem)，許多來自同一個分佈的隨機變數之平均值為常態分佈(無論這隨機變數之原始分佈為何)。測試樣本資料是否呈常態分布的方法，至少包括：卡方檢定、K-S 檢定、偏態係數檢定和 Jarque-Bera 檢定等四種方法。

偏態係數檢定的邏輯是：若「樣本資料呈常態分布」(p 事件)，則「偏態係數為 0」(q 事件)，意即「若 p 則 q」命題為真。「若 p 則 q」和「非 q 則非 p」兩個命題的真偽恒等，因此若偏態係數不為 0 (非 q) 則可知「樣本資料不呈常態分布」。當資料樣本數  $N$  很大時，樣本資料偏態係數  $\hat{C}_s$  估計值的變異數為  $E[(\hat{C}_s - C_s)^2] \approx 6/N$ ，例如  $N = 40$  時， $E[(C_s - \hat{C}_s)^2] \approx 6/40$ ；



在  $\alpha$  顯著水準下，當  $|\hat{C}_s| > z_{1-\alpha/2} \sqrt{6/N}$  就拒絕水文資料為常態分佈的假設。

Jarque-Bera 檢定方法同時檢測樣本資料的偏態係數  $\hat{C}_s$  和峰度係數  $\hat{C}_k$ ，測試統計值  $JB$  的定義為：

$$JB = \frac{N}{6} \left[ \hat{C}_s^2 + \frac{(\hat{C}_k - 3)^2}{4} \right]$$

Jarque-Bera 檢定法的虛無假設是樣本資料的分佈呈常態分布，因此理論的偏態係數  $C_s=0$ ，峰度係數  $C_k=3$ 。Jarque-Bera 檢定的測試統計值  $JB$  在樣本各數  $N$  很大時，會趨近自由度為 2 的卡方分佈， $\chi^2(2)$ 。Jarque-Bera 檢定法的邏輯和偏態系數檢定的邏輯相同，只是「q 事件」同時包括  $C_s=0$  和  $C_k=3$ ，因此「非 q 事件」是  $C_s \neq 0$  或  $C_k \neq 3$ ，即測試統計值  $JB$  的數值愈大，愈傾向拒絕虛無假設。因此，在  $\alpha$  顯著水準下，當  $JB > \chi_{1-\alpha}^2(2)$ ，例如， $\alpha = 0.05$ ， $\chi_{0.95}^2(2)=5.99$ ，樣本統計值  $JB > 5.99$ ，即拒絕樣本資料呈常態分佈的假設。

#### 1.4.5 常態分佈的轉換

處理水文時間序列資料時，當資料不呈常態分佈時，處理的方法常使用 Box-Cox 轉換(Box-Cox Transformation)，轉換形式如下：

$$y_t = \begin{cases} \frac{x_t^\lambda - 1}{\lambda} & \text{for } \lambda \neq 0 \\ \ln x_t & \text{for } \lambda = 0 \end{cases}$$

在以上的 Box-Cox 轉換中  $x_t$  為原始資料， $y_t$  為轉換變數，參數  $\lambda$  可由試誤法 (trial & error) 或其他方法求得以使偏態係數  $k_s$  為零，當  $\lambda=0$  時實際上是表示將(兩參數的)對數常態(log-normally)分佈的變數，轉換成常態分佈。

#### 1.4.6 對數常態分布(Log-Normal Distribution)

若  $y = \log x$ ，且  $Y$  為常態分佈，則  $X$  呈對數常態分佈，或更清楚的定義，是二參數對數常態分佈。

$$f(X=x) = \frac{1}{x\sqrt{2\pi}\sigma_y} \exp\left(-\frac{(y-\mu_y)^2}{2\sigma_y^2}\right)$$

上式中， $\mu_y$ 和 $\sigma_y$ 分別為隨機變數 $Y$ 的期望值與標準偏差，兩者定義如下：

$$\mu_y = E[Y] = E[\log X]$$

$$\sigma_y = \sqrt{\text{var}[Y]} = \sqrt{\text{var}[\log X]}$$

二參數對數常態分佈的隨機變數 $X$ 的下邊界值為 $x=0$ ，未必適用於所有隨機變數。三參數對數常態分佈，是將隨機變數 $X$ 的固定下邊界值 $x=0$ ，改為任意值 $x=c$ ，使其更有彈性、更能適於表示隨機變數分佈，方法是將以上三式中的 $x$ 都替換為 $x-c$ ，即：

$$f(x) = \frac{1}{(x-c)\sqrt{2\pi}\sigma_y} \exp\left(-\frac{(y-\mu_y)^2}{2\sigma_y^2}\right)$$

$$\mu_y = E[y] = E[\log(x-c)]$$

$$\sigma_y = \sqrt{\text{var}[y]} = \sqrt{\text{var}[\log(x-c)]}$$

隨機變數 $X$ 和 $Y$ 前三個動差值(moments)之間的關係式如下：

$$\mu_x = c + \exp\left[\frac{1}{2}\sigma_y^2 + \mu_y\right]$$

$$\sigma_x^2 = \exp\left[2\sigma_y^2 + 2\mu_y\right] - \exp\left[\sigma_y^2 + 2\mu_y\right]$$

$$k_{s,x} = \frac{\exp(3\sigma_y^2) - 3\exp(\sigma_y^2) + 2}{\left[\exp(\sigma_y^2) - 1\right]^{3/2}}$$

若 $X$ 和 $Y$ 為時間序列隨機變數，即序率過程(stochastic process)，則兩隨機變數一階稽延時間(lag 1)相關係數之間的關係式為：

$$\rho_x = \frac{\exp(\sigma_y^2 \rho_y) - 1}{\exp(\sigma_y^2) - 1}$$

## 1.5 隨機變數轉換 (Transformation of Random Variables)

若 $X$ 為隨機變數，則經過數轉換之隨機變數 $Y = u(X)$ 亦為一個隨機變數。吾人可將 $Y$ 視為是隨機變數 $X$ 的轉換；亦可視為是隨機變數函數的機率分佈(distributions of functions of random variables)。求隨機變數 $Y$ 的機率密度函數的方法主要有兩種，第一種是「分佈函數法」(distribution function technique)，

利用隨機變數  $X$  的樣本值與其對應的機率密度，分別經過  $Y = u(X)$  函數轉換，計算得到隨機變數  $Y$  的機率密度函數： $G(y) = P[u(X) < y]$ ；第二種是「替換變數法」(change of variable technique)，方法是將  $y = u(x)$  的反函數  $x = u^{-1}(y) = w(x)$  代入  $x$  的機率密度函數  $f(x)$  中，得到隨機變數  $y$  的機率密度函數，唯在隨機變數是連續隨機變數時，還須乘以樣本值  $x$  對樣本值  $y$  導函數的「轉換賈克比」(Jacobian of transformation)，或是「賈克比矩陣」。以上兩種方法之中，「分佈函數法」往往較不容易計算，利用以下例題說明；本節中主要將介紹「替換變數法」在隨機變數為離散與連續時的轉換方式。

例題1-3 令  $X_1, X_2, X_3$  均為標準常態分佈之隨機變數，並且  $Y = X_1^2 + X_2^2 + X_3^2$ ，求隨機變數  $Y$  的累積機率密度函數  $G(y) = P[X_1^2 + X_2^2 + X_3^2 < y]$ 。

解：若  $y < 0$ ， $G(y) = 0$ 。

若  $y > 0$ ，則  $G(y) = \iiint_A \frac{1}{(2\pi)^{3/2}} \exp\left[-\frac{1}{2}(x_1^2 + x_2^2 + x_3^2)\right] dx_1 dx_2 dx_3$ ，其中積分

範圍  $A$  是以  $(x_1, x_2, x_3) = (0, 0, 0)$  為中心，半徑為  $\sqrt{y}$  的球體內所有點的集合，積分的過程相當麻煩，積分的結果得到隨機變數  $Y$  的機率密度函數是一個三個自由度的卡方分佈， $\chi^2(3)$ 。

### 1.5.1 離散隨機變數之轉換

令  $f(x_1, x_2)$  為兩個離散的隨機變數  $X_1$ 、 $X_2$  的聯合機率函數，其變數值  $(x_1, x_2)$  的集合(=樣本空間)為  $A$ ，在此集合內所有二維點  $(x_1, x_2)$  的發生機率  $f(x_1, x_2)$  均為正值，即： $f(x_1, x_2) > 0, \forall (x_1, x_2) \in A$ 。令  $y_1 = u_1(x_1, x_2)$ ， $y_2 = u_2(x_1, x_2)$  為由  $A$  到  $B$  的一對一對應轉換，則兩個隨機變數  $Y_1 = u_1(X_1, X_2)$  與  $Y_2 = u_2(X_1, X_2)$  的聯合密度函數為：

$$g(y_1, y_2) = \begin{cases} f[w_1(y_1, y_2), w_2(y_1, y_2)] & \text{for } (y_1, y_2) \in B \\ 0 & \text{elsewhere} \end{cases}$$

其中， $x_1 = w_1(y_1, y_2)$  與  $x_2 = w_2(y_1, y_2)$ ，為  $y_1 = u_1(x_1, x_2)$  與  $y_2 = u_2(x_1, x_2)$  的單點一對一反函數。

以上求  $g(y_1, y_2)$  機率密度函數的方法，是將變數值  $x_1$ 、 $x_2$  分別以  $w_1(y_1, y_2)$

與  $w_2(y_1, y_2)$  替換，代入機率密度函數  $f(x_1, x_2)$  的作法，故稱為「替換變數法」。  
使用替換變數法時，若是變數  $y$  的個數少於變數  $x$  的個數，必須增加虛擬  $y$  變數，使其變數個數與變數  $x$  的個數相同。

例題1-4 令  $X_1, X_2$  均為兩個相互獨立、機率函數為 Poisson 分佈、平均值分別為  $\mu_1$  與  $\mu_2$  的隨機變數，其聯合密度函數為：

$$f(x_1, x_2) = \begin{cases} \frac{\mu_1^{x_1} \mu_2^{x_2} e^{-\mu_1} e^{-\mu_2}}{x_1! x_2!} & \text{for } x_1 = 0, 1, 2, \dots, x_2 = 0, 1, 2, \dots \\ 0 & \text{elsewhere} \end{cases}$$

求隨機變數  $Y_1 = X_1 + X_2$  的機率函數。

解：(1) 由於轉換前的隨機變數  $X$  有兩個，轉換後的隨機變數  $Y$  只有一個，所以要增加一個虛擬的變數，令  $Y_2 = X_2$ ，則由  $(x_1, x_2)$  到  $(y_1, y_2)$  的轉換為一對一的轉換。

(2) 以  $y_1, y_2$  表示  $x_1, x_2$  的函數表示法為： $x_1 = w_1(y_1, y_2) = y_1 - y_2$  與  $x_2 = w_2(y_1, y_2) = y_2$ 。將此二函數代入  $f(x_1, x_2)$  即可求得  $y_1, y_2$  的聯合密度函數  $g(y_1, y_2)$ 。

$$g(y_1, y_2) = \frac{\mu_1^{y_1 - y_2} \mu_2^{y_2} e^{-\mu_1} e^{-\mu_2}}{(y_1 - y_2)! y_2!} \quad \text{for } (y_1, y_2) \in \mathcal{B}$$

(3) 由於所要求的是隨機變數  $Y_1 = X_1 + X_2$  的機率密度函數，而不是以上含有  $y_2$  的聯合機率密度函數，所以需要再對於  $y_2$  的範圍積分，求取  $y_1$  的邊際機率密度函數。

$$\begin{aligned} g(y_1) &= \sum_{y_2=0}^{y_1} g(y_1, y_2) \\ &= \frac{e^{-\mu_1} e^{-\mu_2}}{y_1!} \sum_{y_2=0}^{y_1} \frac{y_1!}{(y_1 - y_2)! y_2!} \mu_1^{y_1 - y_2} \mu_2^{y_2} \\ &= \frac{(\mu_1 + \mu_2)^{y_1} e^{-\mu_1} e^{-\mu_2}}{y_1!} \quad \text{for } y_1 = 0, 1, 2, \dots \end{aligned}$$

以上推導結果顯示  $Y_1 = X_1 + X_2$  是平均值為  $\mu_1 + \mu_2$ ，呈 Poisson 分佈的隨機變數。

## 1.5.2 連續隨機變數之轉換

令連續隨機變數  $X$  的機率密度函數為  $f(x)$ ， $y = u(x)$  為由  $A$  到  $B$  的一對

一對應轉換， $x = w(y)$  為  $y = u(x)$  的反函數，則隨機變數  $Y = u(X)$  的機率密度函數為：

$$g(y) = \begin{cases} f[w(y)] \cdot \left| \frac{dw(y)}{dy} \right| & \text{for } y \in \mathcal{B} \\ 0 & \text{elsewhere} \end{cases}$$

其中， $dx/dy = J$  稱為變數轉換的賈克比(Jacobian of transformation)， $J$  為其代表的數學符號， $|dw(y)/dy| = |dx/dy| = |J|$  為賈克比的絕對值。以上連續隨機變數的轉換假設  $x = w(y)$  可微分，同時  $dx/dy$  為連續函數。

例題1-5 假設  $X$  為在  $(0, 1)$  範圍內的均勻分佈連續隨機變數，其機率密度函數為：

$$f(x) = \begin{cases} 1 & \text{for } 0 < x < 1 \\ 0 & \text{elsewhere} \end{cases}$$

求隨機變數  $Y = -2 \ln X$  的機率密度函數。

解：(1) 由於  $y = u(x) = -2 \ln x$ ，所以  $x = w(y) = e^{-y/2}$ 。

$$(2) |J| = |dx/dy| = \left| -\frac{1}{2} e^{-y/2} \right| = \frac{1}{2} e^{-y/2}。$$

(3) 將  $x = w(y) = e^{-y/2}$  與  $|J| = \frac{1}{2} e^{-y/2}$  代入，即可求得隨機變數  $y$  的機率密度函數：

$$g(y) = f[w(y)] \cdot |J| = f[e^{-y/2}] \cdot \frac{1}{2} e^{-y/2} = \frac{1}{2} e^{-y/2}$$

隨機變數  $x$  的樣本空間  $\mathcal{A} = \{x, 0 < x < 1\}$ ，轉換隨機變數  $y$  的樣本空間  $\mathcal{B} = \{y, 0 < y < \infty\}$ 。

若連續的隨機變數個數為兩個時，假設  $X_1$ 、 $X_2$  的聯合機率密度函數為  $f(x_1, x_2)$ ； $y_1 = u_1(x_1, x_2)$  與  $y_2 = u_2(x_1, x_2)$  為由  $\mathcal{A}$  到  $\mathcal{B}$  的一對一對應轉換， $x_1 = w_1(y_1, y_2)$ ， $x_2 = w_2(y_1, y_2)$  為由  $\mathcal{B}$  到  $\mathcal{A}$  的一對一對應的反轉換，則兩個隨機變數  $Y_1 = u_1(X_1, X_2)$  與  $Y_2 = u_2(X_1, X_2)$  的聯合機率密度函數為：

$$g(y_1, y_2) = \begin{cases} f[w_1(y_1, y_2), w_2(y_1, y_2)] \cdot |J| & \text{for } (y_1, y_2) \in \mathcal{B} \\ 0 & \text{elsewhere} \end{cases}$$

上式中的  $|J|$  為變數轉換的賈克比(Jacobian of transformation)的行列式

(determinant)的絕對值，賈克比矩陣為：

$$J = \begin{bmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{bmatrix}$$

在樣本空間  $B$  之中， $J$  不能全等於零。

例題1-6 假設  $X_1$  與  $X_2$  分別為  $\text{Gamma}(\alpha, 1)$  與  $\text{Gamma}(\beta, 1)$  分佈、彼此獨立的兩個連續隨機變數， $Y_1 = X_1 + X_2$ ， $Y_2 = X_1 / (X_1 + X_2)$ ，求隨機變數  $Y_1$  與  $Y_2$  的聯合機率密度函數，並且證明  $Y_1$  與  $Y_2$  為兩個獨立的隨機變數。

解：(1) 由於  $X_1$  與  $X_2$  為  $\text{Gamma}$  分佈、彼此獨立的兩個連續隨機變數，其聯合機率密度函數為：

$$f(x_1, x_2) = \frac{1}{\Gamma(\alpha) \cdot \Gamma(\beta)} x_1^{\alpha-1} x_2^{\beta-1} e^{-x_1} e^{-x_2}$$

樣本空間  $A = \{(x_1, x_2), 0 < x_1 < \infty, 0 < x_2 < \infty\}$ 。

(2) 變數轉換的反函數分別為  $x_1 = y_1 y_2$  與  $x_2 = y_1(1 - y_2)$ ，轉換的賈克比矩陣為：

$$|J| = \left| \begin{bmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{bmatrix} \right| = \left| \begin{bmatrix} y_2 & y_1 \\ 1 - y_2 & -y_1 \end{bmatrix} \right| = |-y_1|, \text{ 不全等於 } 0$$

樣本空間  $B = \{(y_1, y_2), 0 < y_1 < \infty, 0 < y_2 < 1\}$ 。

(3) 隨機變數  $y$  的機率密度函數為：

$$\begin{aligned} g(y_1, y_2) &= y_1 \frac{1}{\Gamma(\alpha) \cdot \Gamma(\beta)} (y_1 y_2)^{\alpha-1} [y_1(1 - y_2)]^{\beta-1} e^{-y_1} \\ &= \frac{1}{\Gamma(\alpha) \cdot \Gamma(\beta)} [y_1^{\alpha+\beta-1} e^{-y_1}] [y_2^{\alpha-1} (1 - y_2)^{\beta-1}] \end{aligned}$$

由於  $g(y_1, y_2)$  可以寫為  $g_1(y_1) \cdot g_2(y_2)$  的形式，可以分別積分，故隨機變數  $Y_1$  與  $Y_2$  為兩個獨立的隨機變數。

(4) 對於隨機變數  $y_1$  的樣本空間積分，求隨機變數  $y_2$  的邊際機率密

度函數：

$$\begin{aligned} g_2(y_2) &= \frac{y_2^{\alpha-1}(1-y_2)^{\beta-1}}{\Gamma(\alpha) \cdot \Gamma(\beta)} \int_0^\infty y_1^{\alpha+\beta-1} e^{-y_1} dy_1 \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha) \cdot \Gamma(\beta)} y_2^{\alpha-1}(1-y_2)^{\beta-1} \quad \text{for } 0 < y_2 < 1 \end{aligned}$$

故可知  $y_2$  為呈 beta 分佈的隨機變數， $\alpha$  與  $\beta$  為其兩個參數。

- (5) 對於隨機變數  $y_2$  的樣本空間積分，求隨機變數  $y_1$  的邊際機率密度函數：

$$\begin{aligned} g_1(y_1) &= \frac{y_1^{\alpha+\beta-1} e^{-y_1}}{\Gamma(\alpha) \cdot \Gamma(\beta)} \int_0^1 y_2^{\alpha-1} (1-y_2)^{\beta-1} dy_2 \\ &= \frac{1}{\Gamma(\alpha+\beta)} y_1^{\alpha+\beta-1} e^{-y_1} \quad \text{for } 0 < y_1 < \infty \end{aligned}$$

$y_1$  呈 Gamma 分佈，兩個參數分別為  $\alpha+\beta$  與 1。