第四節

統計量的機率分布

Probability Distribution of Statistics

所謂機率分布就是將統計試驗的結果，利用數學模式來表示變數的機率，以解釋與推導試驗結果的機率值。因此，需正確判斷機率分布性質與機率分布數學模式，方不致誤用公式或查錯機率分布表。

用以表示計數值主要的機率分布包括二項分布 (binomial distribution)、超幾何分布 (hypergeometric probability distribution)、卜瓦松分布 (Poisson distribution) 等；用以表示計量值主要的機率分布則包括常態分布 (normal distribution)、學生 t - 分布 (Student's t-distribution)、F - 分布 (F-distribution) 等。

壹、計數值的機率分布表示方式

Probability Distribution for Attribute

一、二項分布

二項分布是指統計資料中只有性質不同的兩項群體的概率分布。也就是說，各個資料都可歸為兩個不同性質中的一個，這兩個性質或是觀測值是對立的，如是 / 非、成功 / 失敗、合格 / 不合格、紅球 / 白球、正面 / 反面等，因此二項分布又可說是兩個對立事件的機率分布。對無限群體數 N 進行 n 個獨立的兩項群體概率分布檢測 (如是 / 非、或成功 / 失敗) 時，其中 N 必須大於或等於 10n，而結果只會出現兩種 (如是 / 非、或成功 / 失敗) ，且抽出之樣本經檢測後，需再放回群體內混合再抽，每次試行前後互無關聯，均為獨立進行。若每次試驗的成功機率為 p (且 p 固定) ，不成功機率為 (1-p)，而 n 個獨立的檢測，成功的次數為 x，則其機率分布的結果可經由下式計算或是查二項分布累積和機率表而得：

$$p(x)=\binom{n}{x}p^{x}(1-p)^{n-x}$$

二項分布的期望值〔E(x)〕與標準差則分別為：

$$E(x)=\mu_x=np=\mu$$
$$\sigma(x)=\sigma x=\sqrt{np(1-p)}$$

如食品工廠的某一生產線，每天生產數萬瓶飲料，若產品的平均不良率為 2.0%，則檢驗人員隨機抽樣 100 瓶飲料，樣本中不合格飲料的機率分布為：

$$p(x)=\binom{100}{x}(0.02)^x(0.98)^{100-x},\ x=0,1,2,\ ...,\ 100$$

利用二項分布，我們也可以進一步計算出抽樣樣本中有一個或更少個樣本為不良品的機率為：$p(x\le 1)=p(x=0)+p(x=1)$ 。

二、超幾何分布

超幾何分布亦是統計學上常見的離散機率分布。它描述了由有限個物件中抽出 n 個物件 (N<10n)，成功抽出指定種類的物件個數（且抽出後不放回或不歸還）。如在有 N 個樣本，其中 D 個是不及格的（其機率為 $p=\frac{D}{N}$），N-D 個是及格的（其機率為 $1-p=\frac{N-D}{N}$）。那麼超幾何分布描述了在該 N 個樣本中抽出 n 個，其中 x 個是不及格的機率（或是看成 N 分割成 D 個，自 D 中抽出 x 個不及格的機率），可經由下式計算：

$$p(x)=\frac{\binom{D}{x}\binom{N-D}{n-x}}{\binom{N}{n}}$$

超幾何分布的期望值〔E(x)〕與標準差則分別為：

$$E(x)=\mu_x=\frac{nD}{N}=\mu$$

$$\sigma(x)=\sigma_x=\sqrt{\frac{nD}{N}\left(1-\frac{D}{N}\right)\left(\frac{N-n}{N-1}\right)}$$

如果 N 接近無窮大（∞）或抽樣比例很小（即 n/N 的值很小，至少小於或等於 0.1），超幾何分布則可視為二項分布（即二項分布可作為超幾何分布很好的近似值）。

三、卜瓦松分布

卜瓦松分布最早是由一位著名的法國數學家及物理學家 Simeon D. Poisson 所提出，他從二項分布的極限得到了這個日後以他命名的機率分布，適合於描述單位時間內（或單位面積等）隨機事件（或稀有事件）發生次數的機率分布。在二項隨機試驗中，當 n 很大（N≥10n，或 N 未知，則 n 取 16 以上）而隨機事件（或稀有事件）發生次數機率很小 (p<0.1) 時，我們可以用卜瓦松分布求得二項分布的近似機率值。簡單的說，我們在一個固定的時間間隔或固定範圍內，觀察某一特定事件發生的次數，會發生幾次是一個隨機變數。若某一特定事件在一個固定的時間間隔或固定範圍內發生的平均次數 λ 為已知，那麼當這些隨機行為變數具有以下性質，即符合卜瓦松分布：

1. 事件發生一次的機率與時間的長度或是區域大小成正比。
2. 在極短的時間或極小的區域內，事件發生兩次的機率幾乎為零。也就是說，任何兩事件發生，就時間而言可以區分前後，就發生地點而言可以分別彼此的範圍。
3. 任兩個不重疊的時間或區域，事件發生的次數彼此間相互獨立。

因此卜瓦松分布是在以已知特定事件在單位時間(或區域)內發生的平均次數 λ (λ=np)，去衡量一段時間事件發生次數 x 之機率，其機率密度函數可由卜瓦松機率分布表或是利用下式求得：

$$p(x)=\frac{e^{-\lambda}\cdot\lambda^{x}}{x!}$$

而卜瓦松隨機變數的期望值與變異數相同，分別為：

$$E(x)=\lambda$$
$$\sigma=\sqrt{\lambda}$$

日常生活中的許多現象也被發現是符合卜瓦松分布的，例如每小時進入某一學校大門口的人數、某一麵店每小時的來客人數，或是某個路口每次紅綠燈之間的車流量等。卜瓦松分布也是品質管制的利器，它可以幫助我們決定產品在生產過程中是否出了問題。例如假設某工廠每製作 100 個螺絲釘，平均會有 2 個不合規格，而這是合理的不合格率。但是不合格事件數量到底要有多大的改變才算是明顯的變化而反映製程的問題呢？根據卜瓦松分布，偶而出現 3 個或 4 個不合規格的螺絲釘也是正常的現象，但是如果出現的頻率太高(如出現 5 個以上)的不合規格螺絲釘，那麼生產過程就可能出現問題了。

貳、計量值的機率分布表示方式

Probability Distribution for Variables

一、常態分布

機率密度函數

當所探討之隨機變數 x 為計量值資料，其數值為量測而來，本身就存在著誤差或近似值的問題。如一個包裝食品的內容物重量為 28 公克，則數據“28”意味著無數個近似值(如 27.99999...、28.00001...)。因此計量值常以一個區間內無數個值(如 27.99999~28.00001)來看待 X 軸上之變數 x，而其所對應之機率值以「機率密度函數」f(x) 表示，也就是由區間與函數所圍成的面積做其機率之對應值。

常態分布的性質

常態分布是計量值機率分布中最重要的機率分布。若一個連續隨機變數 X，其所對應之機率密度函數 f(x) 遵循常態曲線，即其形狀為左右對稱如鐘形般之曲線（圖 3-4），此曲線只有一個眾數，且平均值與其眾數以及中位數等值，曲線的兩尾是向兩端無限延伸。而常態曲線之機率密度函數為：

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

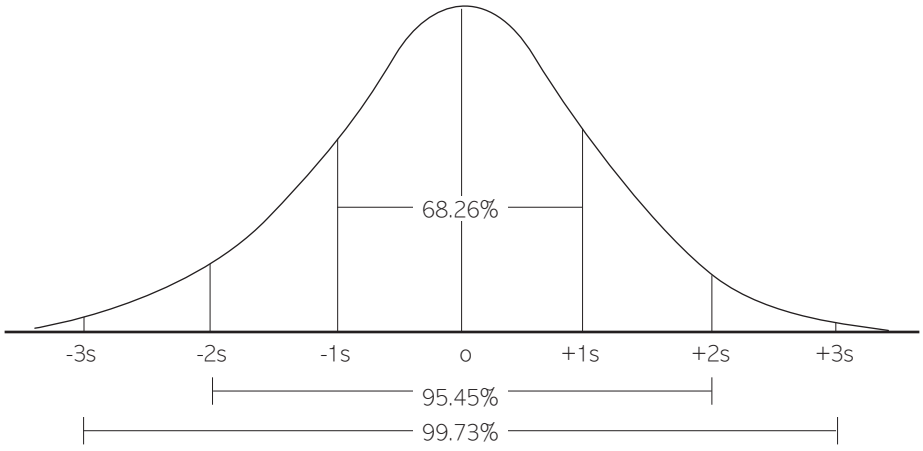

圖 3-4　常態分布機率密度函數示意圖

常態分布曲線下的機率計算方式為，將此曲線下方細切成長方形，則在某一區間內 (a ≤ x ≤ b) 其所有長方形的面積與此曲線下所有面積的比，即為相對次數，以百分率表示，即為機率 (probability) 或稱機率密度 (probability density)，亦簡稱密度 (density)。以積分公式表示為：

$$A=\int f(x)dx=\int_a^b \frac{1}{\sigma\sqrt{2\pi}}\cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}dx$$

常態分布中一些值得注意的量，包括（圖 3-5）：

- 密度函數以平均值為中心且左右對稱。
- 平均值與它的眾數以及中位數等值。
- 函數曲線下 68.26% 的面積在平均數左右的一個標準差範圍內。
- 95.45% 的面積在平均數左右兩個標準差 2σ 的範圍內。
- 99.73% 的面積在平均數左右三個標準差 3σ 的範圍內。
- 99.99% 的面積在平均數左右四個標準差 4σ 的範圍內。
- 函數曲線的反曲點 (inflection point) 發生在離平均數一個標準差距離的位置。

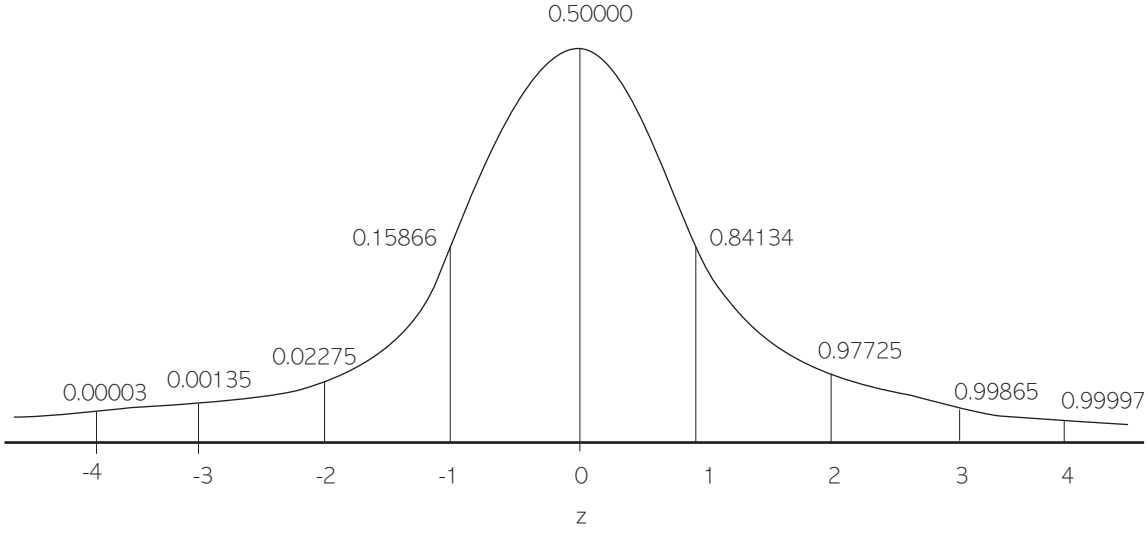

圖 3-5　Z 分布或標準常態分布示意圖

說明：數字代表 Z 分布曲線下之累積面積。

所以常態分布的偏態為 0，峰度亦為 0。除此之外，常態分布還有一個非常重要的性質，即中央極限定理 (central limit theorem)。當從母體抽取夠大的 n 個觀測值為樣品，則所有可能的平均值愈接近常態分布，且樣品的平均值會逼近母體的平均值，且標準差會愈小。中央極限定理的重要意義在於，只要在抽樣過程中大小合理，其他機率分布也可以用常態分布作為近似。換言之，有些資料本身的分布雖不呈常態分布，但若是從族群中抽取一個夠大的樣品，並計算各樣品平均值，則此樣品平均值所組成