

誤差論與最小平方法

蔡聰明

科學有實驗的一面，也有理論的一面，分別反應科學知識具有「感官經驗」與「邏輯論證」之內涵，兩者不可或缺。

實驗這一面離不開觀測度量，觀測度量就得到一些數據 (data)，數據免不了就會有誤差 (errors)。數據是我們叩問自然所得到的回音，回音伴有「雜訊」(noise)，導致誤差。雜訊包括有人為的與隨機的兩個因素，前者如儀器不精、度量技術不佳等等，後者如溫度變化、一陣風吹、重力變化等等。它們所產生的誤差分別叫做系統性誤差 (Systematic error) 與隨機性誤差 (random error)；前者可以改進，後者只能任其自然。因此，誤差恆在！誤差論就是研究隨機誤差之機率理論。物理學家 M.Born 說 [1]：

機率概念在科學活動的第一步就滲進來，因為沒有觀測是絕對精確的。我認為機率概念比因果關係更基本，理由是因果關係的成立與否，只能運用機率法則於觀測上才能判斷。

Gauss(1777-1855) 爲了利用天文觀測數據來決定行星運行的軌道，發展出最小平方法 (the least square method) 及誤差論。

自然現象的背後有規律，泛稱爲自然律。令人驚奇的是，誤差居然也有規律可循，叫做誤差律 (the law of errors)。誤差律是什麼呢？它是如何追尋出來的？這些是本文所要探討的問題。

誤差律的追尋是展示數學探索過程的一個絕佳範例，它將機率、統計學與科學方法論揉合起來，在機率論與統計學的發展史上佔有核心的地位。

一、平均律之謎

我們經常在使用平均律 (the law of averages) 的常識性法則，有時甚至「不知亦能行」。因此，它又稱爲「木匠法則」或「黃金法則」(the Golden rule)。讓我們觀察一些例子。

在記述統計學 (Descriptive statistics) 中，算術平均 (簡稱爲平均) 是最常用來表現母群體 (population) 的代表值。例如，考試過後，除了關心自己考幾分之外，一般總是還要問全班平均得幾分。換言之，若全班成績爲 x_1, x_2, \dots, x_n ，那麼就用算術平均 $\bar{x} = \frac{1}{n} \sum_{k=1}^n x_k$ 一個數字來代表全班的概況。

在機率論中，「丟一個公正銅板出現正面的機率爲 $\frac{1}{2}$ 」的意思是指：獨立地重複丟此

銅板, 得到一個隨機序列 $\xi_1, \xi_2, \xi_3, \dots$ (取值 0 或 1), 那麼平均值 $\frac{1}{n} \sum_{k=1}^n \xi_k$ 會越來越「接近」於 $\frac{1}{2}$ (當 $n \rightarrow \infty$), 這叫大數法則。

假設某物理量的真值為 θ , 這是一個未知數, 我們要用度量來估計它。由於誤差不可避免, 故只度量一次並不保險, 直覺常識告訴我們要多度量幾次, 得到 $\theta_1, \theta_2, \dots, \theta_n$, 再取其算術平均 $\bar{\theta} = \frac{1}{n} \sum_{k=1}^n \theta_k$ 來當作 θ 的估計, 這樣會比較可靠。

在統計學課堂上, 有人作過一個實驗: 老師在黑板上劃一線段, 然後要求每位同學用眼睛估計其長度, 從 53 公分到 84 公分都大有人在。老師將全班作平均得到 $61\frac{1}{8}$ 公分, 而實際度量的結果是 $61\frac{1}{4}$ 公分, 因此平均的估計相當好。這是平均的妙用, 所謂「三個臭皮匠勝過諸葛亮」(Several heads are better than one.)

遍歷論 (ergodic theory) 裡, 談物理量對空間的平均以及對時間的長期平均。機率論裡的期望值更是一種平均值。音樂的樂理也有十二平均律。

在日常生活中, 我們時常遇到「平均」。例如: 國民的平均所得; 追求一個均富的社會, 「不患貧而患不均」, 等等, 好像「平均」是一個美麗的理想遠景, 「平均」就是天堂! 但不要忘了, 熱力學的均溫狀態是一個大家都「熱死」(heat death) 的死寂世界。

我們要提防「平均」的誤用。有一個笑話說: 有兩桶水, 溫度分別是 80°C 與 10°C , 你將左、右手各放入一桶水之中, 你可以感受到平均 45°C 之溫度, 舒服極了。另外, 我們偶而也可以聽到政客誇口說: 「我要每一個人的所得皆超過國民的平均所得。」

為什麼算術平均這麼有用? 平均的魅力十足, 其道理何在?

為了揭開平均律之謎, 解讀其背後的意義與道理, 發展出了許多的數學。例如, 利用平均律追究機率意義, 我們發現了大數法則與中央極限定理, 奠下數學機率論的發展, 參見 [5]。

本文我們要追究平均律對誤差現象的意義, 這導致誤差律以及一些重要的統計概念與方法的誕生。

二、三合一又三合一

由於追尋「平均律」背後的意義, 數學家展開了誤差論的研究, 這是十九世紀初的成就。為了敘述方便, 我們先介紹較現代的機率論術語。

隨機現象的機率模型是機率空間 (Ω, \mathcal{F}, P) 。它是由樣本空間 (Sample space) Ω , 事件全體所成的 σ 代數 (σ -algebra) 以及機率測度 (probability measure) P , 三者所組成的三合一體。進一步, 隨機變數 (random variable) X 以及機率分佈 $\mu = P \circ X^{-1}$ 又是一個三合一體 (trinity), 圖解如下:

$$(\Omega, \mathcal{F}, P) \xrightarrow{X} \mu = P \circ X^{-1} \quad (\mathbb{R}, \mathcal{B})$$

整個合起來, 我們簡稱為機率論的「三合一又三合一」之理論架構。這是經過長期的試誤 (trial and error)、演化, 在 1933 年由俄國偉大數學家 Kolmogorov (1903-1987) 年提出來的。欲進入隨機現象的領域作研究, 這是必經的「方便之門」。

在上述中： \mathcal{B} 是實數軸 \mathbb{R} 的 Borel 體 (Borel field)，即由開集所生成的最小 σ 代數；隨機變數 $X : \Omega \rightarrow \mathbb{R}$ 就是一個可測映射 (measurable mapping)，即滿足 $X^{-1}(B) \in \mathcal{F}, \forall B \in \mathcal{B}$ ； μ 定義為 $\mu(B) = P(X^{-1}(B)), \forall B \in \mathcal{B}$ ，叫做 X 的機率分佈 (probability distribution)。

(Ω, \mathcal{F}, P) 是所有機率演算 (Calculus of probabilities) 之地基。就研究隨機變數 X 而言，機率分佈 μ 已含有所有的機率資訊。 μ 是集函數 (set function)。當 X 的值域為 \mathbb{R} 的情形，改用較容易掌握的點函數 (point function) 往往很方便：定義 F 為

$$F(x) = P(-\infty < X \leq x) \quad (1)$$

則 F 滿足

- (i) F 為一個遞增函數；
- (ii) $F(-\infty) = 0, F(+\infty) = 1$
- (iii) 對每一點 $x \in \mathbb{R}$ ， F 右連續且左極限存在。

我們稱 F 為 X 的累積分佈函數，簡稱為分佈函數 (distribution function)。進一步，如果 μ 相對於 Lebesgue 測度是絕對連續 (absolutely continuous)，那麼由微積分根本定理 (Radon-Nikodym 定理) 知，存在一個點函數 $p : \mathbb{R} \rightarrow \mathbb{R}$ 滿足：

- (i) $p(x) \geq 0, \forall x \in \mathbb{R}$ ；
- (ii) $\int_{-\infty}^{\infty} p(x)dx = 1$ ；
- (iii) $\mu(B) = \int_B p(x)dx, \forall B \in \mathcal{B}$ ，特別地 $F(x) = \int_{-\infty}^x p(u)du$ 。

我們稱 p 為 X 的機率密度函數 (probability density function, p.d.f.)。這些都是作機率演算時，必備的工具。

三、誤差律的追尋

假設有個物理量的真值是 θ (未知數)，我們對它度量 n 次得到 $\theta_1, \theta_2, \dots, \theta_n$ ，於是誤差 (error) 為

$$\theta_1 - \theta, \theta_2 - \theta, \dots, \theta_n - \theta \quad (2)$$

在理論上，由於 θ 是未知數，故 (2) 式中的數都是未知的，只有 $\theta_1, \theta_2, \dots, \theta_n$ 是已知的。

自古以來，按直覺常識 (良知良能) 都採用

$$\bar{\theta} = \frac{1}{n} \sum_{k=1}^n \theta_k \quad (3)$$

來估計 θ ，即所謂的平均律法則。這樣做有什麼道理呢？

面對觀測 (抽樣) 數據 $\theta_1, \theta_2, \dots, \theta_n$ ，統計學要做的工作是：先作記述統計的處理，對數據作整理、歸類並且作統計圖表；其次把數據精煉成一些代表值；最後解讀數據背後的意義，這是推理統計學的核心工作，需要透過機率模型來完成。

算術平均只是代表值的一種，我們還可以有幾何平均、調和平均、加權平均、中位數、眾數，...，等等。在諸代表值中，我們獨鍾於算術平均，初步的理由是它具有下列兩個優點：

定理 1:

- (i) 使總偏差

$$f(\theta) = \sum_{k=1}^n (\theta_k - \theta)$$

等於 0 的點是算術平均 $\bar{\theta}$ 。

(ii) 使總平方偏差

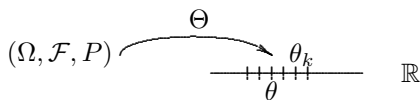
$$g(\theta) = \sum_{k=1}^n (\theta_k - \theta)^2$$

取最小值的點, 也是算術平均 $\bar{\theta}$ 。

這兩個理由只是普通常識而已。更深刻的理由是 Gauss 提出來的:

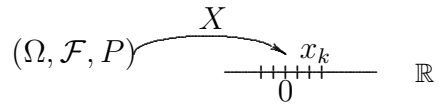
算術平均 $\bar{\theta}$ 是最佳可能值 (the most probable value)。

什麼是「最佳」呢? 在數學中, 這當然需要作明確的定義。也許我們可以先用一個類比來幫助了解: Leibniz 由我們生活所在的這個世界, 思及所有可能的世界, 然後再用哲學來論證, 這個世界是所有可能世界中最佳的世界, (the best of all possible worlds)。因此, Leibniz 被 Voltaire (伏爾泰) 諷刺為「所有可能世界中的最佳先生」(Mr. the best of all possible worlds.) 這給我們很大的啓示: 機率論的思考方式恰好是由眼見的現實 (realization) ω , 思及所有可能結果 Ω , 並且 ω 只是 Ω 中的一個元素。由於對物理量 θ 的觀測度量可得到各種不同的值, 這促使我們從「現實」提昇到「所有可能」, 把「觀測度量」想像成一個隨機變數 Θ , 定義在某個機率空間上面 (這一步的飛躍是關鍵):



或等價地, 考慮平移, 令 $X = \Theta - \theta$, 這就是誤差的隨機變數, 它的 n 次觀測值為

$$x_1 = \theta_1 - \theta, \quad x_2 = \theta_2 - \theta, \quad \dots, \quad x_n = \theta_n - \theta$$



因此, 我們得到「三合一又三合一」之機率模型。但是萬事俱備只欠機率分佈, 這是我們所要追尋的。

明確地說, 我們要找出誤差隨機變數 X 的機率密度函數 $p(x)$ 。當然 $p(x)$ 不會憑空掉下來。對於 $p(x)$ 我們已知什麼? 可以假設什麼?

由於 $p(x)$ 是機率密度函數, 天然地滿足:

$$(i) \quad p(x) \geq 0, \forall x \in \mathbb{R}; \quad (4)$$

$$(ii) \quad \int_{-\infty}^{\infty} p(x) dx = 1。$$

再加上對誤差的直觀了解, 我們可以合理地假設:

(iii) p 為一個二階連續可微分函數 (平滑性);

(iv) $p(x) = p(-x)$, 即對稱原點 (對稱性);

(v) $\lim_{x \rightarrow \pm\infty} p(x) = 0$ (漸近性)。

.....(5)

這些性質還是不能唯一決定 $p(x)$, 至少正規分佈 $y = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ 與 Cauchy 分佈 $y = \frac{1}{\pi} \frac{1}{1+x^2}$ 都滿足上述五個條件。

真正捕捉到 $p(x)$ 的是 Gauss, 他的手法如下:

對於隨機變數 Θ 作 n 次獨立的觀測, 得到 n 個「現實」 $\theta_1, \theta_2, \dots, \theta_n$ 。從而, 在理論上我們就得到 X 的 n 個觀測值

$$x_1 = \theta_1 - \theta, x_2 = \theta_2 - \theta, \dots, x_n = \theta_n - \theta \quad (6)$$

由「現實」提昇到「所有可能」, (6) 式是 n 次獨立觀測:

$$X_1, X_2, \dots, X_n$$

所實現出來的「現實」:

$$X_1 = x_1, X_2 = x_2, \dots, X_n = x_n \quad (7)$$

其中 X_1, X_2, \dots, X_n 是 X 的 n 個抄本 (copies), 它們互相獨立並且每一個都跟 X 具有相同的機率分佈 (i.i.d.)。

在此理論架構下, Gauss 提出一個深具洞察力的基本假設:

算術平均 $\bar{\theta} = \frac{1}{n} \sum_{k=1}^n \theta_k$ 是未知數 θ 的最佳可能值。這個假設的精確意思如下:

在相同條件下, 對 X 作 n 次獨立的觀測得到 (7) 式, 其聯合機率密度函數為

$$\begin{aligned} L(\theta; \theta_1, \theta_2, \dots, \theta_n) \\ &= p(x_1)p(x_2) \cdots p(x_n) \\ &= p(\theta_1 - \theta)p(\theta_2 - \theta) \cdots p(\theta_n - \theta) \end{aligned} \quad (8)$$

這叫做似然函數 (Likelihood function)。Gauss 的想法是: 既然 (7) 式之事件已發生, 理應找 θ 使得 (8) 式取最大值。那麼上述 Gauss 的基本假設就是說:

$$\theta = \bar{\theta} \text{ 爲 (8) 式的最大點} \quad (9)$$

註: 後來統計學家 R. A. Fisher (1890-1962 年, 現代統計學的創立者之一) 所提出的最大似然法 (the maximum likelihood method), 其實是起源於 Gauss 的這個想法。

Gauss 在 1809 年由 (9) 式之假設, 推導出 $p(x)$ 必為正規分佈 (由 $\theta = \bar{\theta}$ 之最佳果反推出因), 如下:

由 (9) 式知

$$\left. \frac{d}{d\theta} L(\theta; \theta_1, \theta_n) \right|_{\theta=\bar{\theta}} = 0$$

或者, 更方便地,

$$\left. \frac{d}{d\theta} \ln L(\theta; \theta_1, \dots, \theta_n) \right|_{\theta=\bar{\theta}} = 0$$

亦即

$$\sum_i \frac{p'(\theta_i - \bar{\theta})}{p(\theta_i - \bar{\theta})} = 0 \quad (10)$$

其中 $p'(x)$ 表示 $p(x)$ 的微分。換言之, $p(x)$ 滿足微分方程 (10), 如何求解呢?

這需要運用一點兒技巧。令 $u_i = \theta_i - \bar{\theta}$ 且 $\varphi(u) = \ln p(u)$, 則 (10) 式變成

$$\sum_i \varphi'(u_i) = 0 \quad (11)$$

顯然有

$$\sum_i u_i = 0 \quad (12)$$

我們要找 φ' 使其對所有滿足 (12) 式之 u_1, u_2, \dots, u_n , 都使 (11) 式成立, 這是一個函數方程 (functional equation) 的問題, 仍然不容易求解。

如果能夠求出 $\varphi'(x)$, 就可求出 $\varphi(x)$, 從而求出 $p(x)$ 。為此, 我們來推導 $\varphi(x)$ 所應滿足的微分方程, 然後求解之。

在 (12) 式中, 想像 u_3, u_4, \dots, u_n 皆固定, 只讓 u_1 與 u_2 變動, 但保持 (12) 式成立。變動的兩個數可令其為 $u_1 + \varepsilon$ 與 $u_2 - \varepsilon$, 其中 ε 為任意實數; 代入 (11) 式得到

$$\varphi'(u_1 + \varepsilon) + \varphi'(u_2 - \varepsilon) = \varphi'(u_1) + \varphi'(u_2)$$

於是

$$\frac{1}{\varepsilon} (\varphi'(u_1 + \varepsilon) - \varphi'(u_1)) = \frac{1}{\varepsilon} (\varphi'(u_2) - \varphi'(u_2 - \varepsilon))$$

令 $\varepsilon \rightarrow 0$, 得到

$$\left. \frac{d}{du} \varphi'(u) \right|_{u_1} = \left. \frac{d}{du} \varphi'(u) \right|_{u_2}$$

因爲 u_1 與 u_2 是獨立變動的, 故得

$$\varphi''(u) = c, (\text{常數}) \quad (13)$$

這就是 φ 所滿足的微分方程。

解 (13) 式得

$$\varphi'(u) = c \cdot u + c_1$$

其中 c_1 爲常數。由 (11) 式知, $c_1 = 0$ 。

於是

$$\varphi'(u) = c \cdot u \quad (14)$$

積分之, 得到

$$\varphi(u) = \frac{1}{2}c \cdot u^2 + c_2 \quad (15)$$

其中 c_2 爲常數。由 $\varphi(u) = \ln p(u)$ 得

$$p(u) = \exp\left(\frac{1}{2}c \cdot u^2 + c_2\right) = ke^{\frac{1}{2}c \cdot u^2} \quad (16)$$

今因 $|u| \rightarrow \infty$ 時, $p(u) \downarrow 0$, 故 c 必是負數, 令其爲 $-\alpha^2$, 亦即

$$p(u) = ke^{-\frac{1}{2}\alpha^2 u^2}, (\alpha > 0)$$

再由 $\int_{-\infty}^{\infty} p(u)du = 1$, 得到 $K = \frac{\alpha}{\sqrt{2\pi}}$ 。從而

$$p(x) = \frac{\alpha}{\sqrt{2\pi}} e^{-\frac{\alpha^2 x^2}{2}} \quad (17)$$

通常寫成下形:

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} \quad (18)$$

其中 $\sigma = \frac{1}{\alpha}$ 。這樣就得到世界上最重要的「正規分佈」, 又叫做「誤差曲線」。

因此, 誤差隨機變數 X 具有正規分佈 (又叫常態分佈), 期望值 $E(X) = 0$, 變異數 $E[(X - E(X))^2] = \sigma^2$, 記成 $X \sim$

$N(0, \sigma^2)$ 。從而, 度量 θ 之隨機變數 $\Theta = X + \theta$ 遵循正規分佈 $N(\theta, \sigma^2)$ 亦即 Θ 的機率密度函數爲

$$q(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\theta)^2}{2\sigma^2}} \quad (19)$$

記成 $\Theta \sim N(\theta, \sigma^2)$ 。容易驗知, 期望值 $E(\Theta) = \theta$, 變異數 $\text{Var}(\Theta) = E(\Theta - \theta)^2 = \sigma^2$ 。

定理2: (Gauss 誤差律, 1809年)

設物理量的真值爲 θ (未知參數), 度量 θ 之隨機變數爲 Θ , 誤差隨機變數 $X = \Theta - \theta$ 的機率密度函數 $p(x; \theta)$ 滿足 (5) 式。對 Θ 作 n 次獨立觀測得 $\Theta_1 = \theta_1, \Theta_2 = \theta_2, \dots, \Theta_n = \theta_n$ 。如果平均值 $\bar{\theta} = \frac{1}{n} \sum_{k=1}^n \theta_k$ 爲 θ 之最佳可能值, 即 $\bar{\theta}$ 爲似然函數 $L(\theta; \theta_1, \dots, \theta_n) = p(\theta_1; \theta) \cdots p(\theta_n; \theta)$ 之最大點, 則 $p(x; \theta)$ 依從正規分佈 $N(0, \sigma^2)$, 某 $\sigma > 0$ 。從而 $\Theta \sim N(\theta, \sigma^2)$ 。

所謂「誤差律」就是指誤差的隨機變數 X 或度量 θ 之隨機變數 Θ 遵循正規分佈的機率法則。

定義: 如果一個隨機變數 Y 的機率密度函數爲 (19) 式, 那麼我們就說 Y 依從 (或具有) 正規分佈 $N(\theta, \sigma^2)$ 。特別地, 當 $\theta = 0, \sigma^2 = 1$ 時, 叫做標準正規分佈。

「正規分佈」(normal distribution) 名稱的由來是這樣的: 因爲 Laplace 是法國人, 故在法國叫做 Laplace 分佈; 而 Gauss 是德國人, 故在德國叫做 Gauss 分佈。事實上, 應該叫做 De Moivre 分佈才對。後來法國偉大數學家 Poincaré(1854-1912) 建議改用「正規分佈」之中立名詞, 他還補充說:

正規分佈律很神秘, 物理學家相信它, 理由是因為他們相信數學家已證明了它的邏輯必然性; 而數學家相信它, 是因為他們相信物理學家在實驗室裡已驗證過它。

因為 Gauss 的名氣太大, 目前數學界通行的用語是「正規分佈」與「Gauss 分佈」, 兩者並用。

另一方面, 定理 2 之逆也成立。

定理 3: 假設隨機變數 Θ 遵循正規分佈 $N(\theta, \sigma^2)$, 其中 θ 為未知參數; 對 Θ 作 n 次獨立觀測得 $\Theta_1 = \theta_1, \dots, \Theta_n = \theta_n$ 。那麼 θ 之最佳可能值為 $\bar{\theta} = \frac{1}{n} \sum_{k=1}^n \theta_k$, 即 $\bar{\theta}$ 為似然函數 $L(\theta; \theta_1, \dots, \theta_n)$ 之最大點。

證明: 似然函數為

$$L(\theta; \theta_1, \dots, \theta_n) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (\theta_i - \theta)^2} \quad (20)$$

對 θ 求此函數的最大值: 由

$$\frac{d}{d\theta} \ln L(\theta; \theta_1, \theta_2, \dots, \theta_n) = 0$$

得到

$$\theta = \bar{\theta} = \frac{1}{n} \sum_{i=1}^n \theta_i \quad (21)$$

進一步, 利用二階微分可以驗證, $\bar{\theta}$ 確實是 L 的最大點。換言之, 算術平均 $\bar{\theta}$ 為 θ 之最佳可能值。證畢。

註: 其它機率分佈不見得具有這個性質。

定義: 使似然函數 $L(\theta; \theta_1, \dots, \theta_n)$ 取最大值的點 $\hat{\theta} = \hat{\theta}(\theta_1, \dots, \theta_n)$ 叫做 θ 的

最大似然推估值 (the maximum likelihood estimate)。提昇到「所有可能」, $\hat{\Theta} = \hat{\theta}(\Theta_1, \Theta_2, \dots, \Theta_n)$ 叫做 θ 的最大似然推估子 (the maximum likelihood estimator)。

問題: 如定理 3 之假設, 求 σ 之最大似然推估子。

上述定理 2 與定理 3 合起來構成正規分佈的一種刻劃。在文獻上, 我們還可以看到其他各式各樣的刻劃。

對於誤差律的推導, 我們也可直觀地來看: 因為誤差隨機變數 X 可以看作是無窮多個獨立的微小基本誤差之代數和; 這些基本誤差都相同, 並且取值正、負之機會均等, 有如公正銅板; 所以由 De Moivre-Laplace 的中央極限定理知道 X 遵循正規分佈。

例子: (捉、放、捉) 如何估計池塘中魚的尾數 N ?

我們先捉 N_a 條活魚上來, 作記號, 放回池中。過些時再任意捉 n 條魚, 其中有 n_a 條是有記號的。根據這些數據我們可以估計 N 的值。

(i) 按比例的想法:

$$N : N_a \doteq n : n_a$$

所以

$$N \doteq \frac{n}{n_a} \cdot N_a$$

(ii) 根據大數法則的想法:

設隨機變數 Y 表示第二度捉魚, 每捉出一條觀察的結果:

$$Y = \begin{cases} 1, & \text{若有記號, 機率為 } \frac{N_a}{N} \\ 0, & \text{若無記號, 機率為 } 1 - \frac{N_a}{N} \end{cases}$$

對 Y 作 n 之獨立觀測 (抽樣) 得到 Y_1, Y_2, \dots, Y_n , 它們獨立且同佈 (i.i.d.) 於是 $S_n = Y_1 + \dots + Y_n$ 表示第二度捉 n 條魚中, 有記號的魚數。由大數法則知 $\frac{S_n}{n}$ 收斂到 $E(Y) = \frac{N_a}{N}$, 而 $\frac{n_a}{n}$ 為 $\frac{S_n}{n}$ 的一個實現值, 故 $\frac{n_a}{n} \doteq \frac{N_a}{N}$, 從而 $N \doteq \frac{n}{n_a} \cdot N_a$ 。

(iii) 最大似然的想法

$$P(S_n = n_a) = {}_n C_{n_a} \left(\frac{N_a}{N}\right)^{n_a} \left(1 - \frac{N_a}{N}\right)^{n-n_a}$$

表示第二度捉 n 條魚中, 出現 n_a 條有記號之機率。這是未知參數 N 的函數, 令其為 $L(N)$ 。事實上, $L(N)$ 是第二度捉魚的似然函數。今求 $L(N)$ 之最大點, 令 $L'(N) = 0$, 解得

$$N = \frac{n}{n_a} \cdot N_a$$

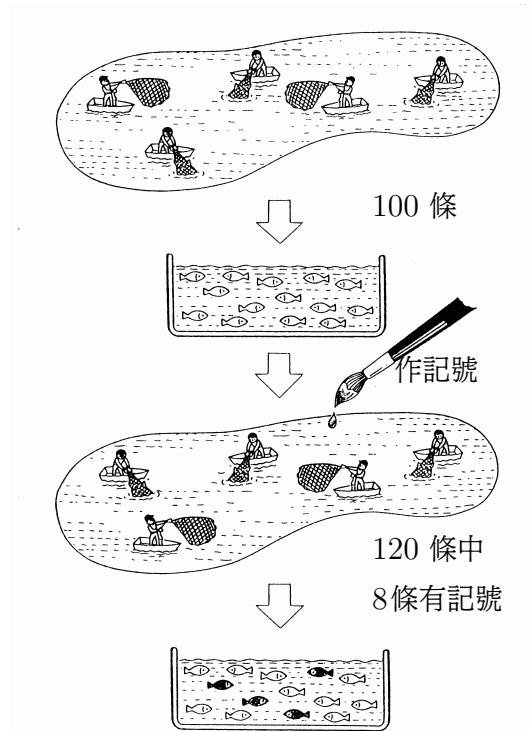
這是 N 的最大似然推估值。

例如, 在下圖中, 我們得到的數據為

$$N_a = 100, n = 120, n_a = 8,$$

所以

$$N \doteq \frac{n}{n_a} N_a = 1500.$$



四、最小平方法及其機率意義

統計學史家 Stigler 在他的「統計學史」一書 [2]中, 開宗明義就說:

最小平方法是十九世紀統計學的主題曲。從許多方面來看, 它之於統計學就相當於十八世紀的微積分之於數學。

假設對一個現象或母群體的兩個變量 x 與 y 觀測 n 次, 得到數據

$$\begin{array}{c|cccc} x & x_1 & x_2 & \cdots & x_n \\ \hline y & y_1 & y_2 & \cdots & y_n \end{array} \quad (22)$$

我們要找一個函數 $y = f(x)$ 來適配 (fits) (22) 之數據。最常見且有用的辦法是找 f 使得

$$\sum_{i=1}^n (y_i - f(x_i))^2 \quad (23)$$

取最小值。這就是所謂的最小平方方法。

但是函數太寬廣多樣，光是這樣我們無從找起。通常需要從其它方面（或理論上）作考慮，限定 f 的類型，例如限定 f 為一次函數或 m 次多項函數等等，於是只需決定係數，這就容易做了。

最小平方方法似乎也如「平均律」一樣，是直觀的「木匠法則」而已。但是，如果我們從誤差律這個不尋常的角度來觀察它，卻可得到下面的結果：

我們這樣解釋 (22) 之數據：在 x_i 之下，觀測 y 得到 y_i ，這跟實際的真值 $f(x_i)$ 就有個誤差

$$z_i = y_i - f(x_i), i = 1, 2, \dots, n \quad (24)$$

今由誤差律知道誤差的隨機變數 Z 遵循正規分佈 $N(0, \sigma^2)$ 。(24) 式就是觀測 Z 所得到的 n 個「現實」：

$$Z_1 = z_1, Z_2 = z_2, Z_n = z_n \quad (25)$$

於是似然函數為

$$L(f; z_1, z_2, \dots, z_n) = \frac{1}{(\sqrt{2\pi})^n \sigma^n} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - f(x_i))^2} \quad (26)$$

定理 4：在上述假設下，即誤差遵循正規分佈，則 f 使 L 取最大值 $\iff f$ 使 (23) 式取最小值。

換言之，最大似然推估法等價於最小平方方法。

Legendre (在 1805 年) 與 Gauss (在 1809 年) 兩人都是因為要由天文觀測數據

以決定行星運行的軌道，而發展出最小平方方法；Gauss 更進一步得到誤差的正規分佈律。Gauss 的這件工作可比美於他後來在微分幾何的偉大傑作：「曲面的一般研究」(General investigations of curved surfaces, 1827 年)。他在 1821 年到 1825 年之間測量過 Hanover 王國的土地，這個「實務」工作激發他發展出微分幾何的曲面理論，開局部方法與研究內稟幾何 (intrinsic geometry) 之先河。經驗與理論密切結合是 Gauss 研究工作的特色。

五、正規分佈律為什麼「鶴立雞群」？

正規分佈的機率法則兩度出現在數學舞台：第一次是作為銅板序列的和（再作標準化）之極限分佈 (limiting distribution) (De Moivre, 1718 年, 與 Laplace, 1801 年), 參見 [5]；第二次是作為觀測度量的誤差律 (Gauss, 1809 年)。對於兩個很不同的研究對象，居然出現相同的正規分佈律，這絕非偶然！通常數學家碰到這種情形都會很興奮，馬上意識到背後必有一個廣闊的機率天地等待探索。

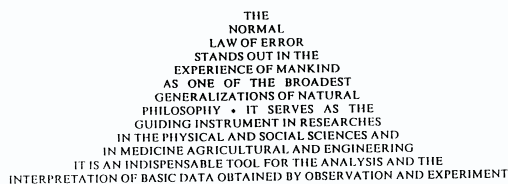
正規分佈律經過 Gauss 的大力提倡，開始暢行於天下。比利時的統計學家 Quetelet (1796 - 1874) 將它廣泛地應用到人文社會現象之研究。他把前人著重在天文現象的研究轉移到地面人間現象之研究，並且認為追求社會人文現象之機率與統計規律就好像是物理家之追求物理定律一樣。他發現人類的種種屬性，如身高、體重、能力等等，都遵循正規分佈律。理由是所有的人好像都是同一個

窯子燒出來一樣，女媧造人本欲造出理想的「平均人」(the average man, Quetelet 之創造語)，不料在創生的過程中發生誤差，導致人類社會的多樣與變異 (variations) 不過由於誤差律適用，而呈現出正規分佈。

達爾文 (Darwin, 1809-1882) 的表弟 Galton (1822-1911) 研究生物統計學、實驗心理學、人體測量學、優生學。他對正規分佈律更是推崇與讚美，他說：「我幾乎不曾見過像“誤差呈正規分佈”這麼美妙而激發人無窮想像的宇宙秩序。如果古希臘人知道這條曲線，想必會給予人格化乃至神格化。它以澄澈統治混亂，消解混亂。暴民越多，無政府狀態越顯現，它就統治得越完美。它是無理性世界中的最高法律。當我們從混沌中抽出大量的樣本時，按大小加以排列整理，那麼總是有一個不可逆料且最漂亮的規律潛伏在其中。」

當代兩位偉大的機率學家 Lévy (1886-1971) 與 Kac (1914-1985) 都曾說過，正規分佈律是他們切入機率論的「初戀情人」，具有無窮的魅力。機率學家 Loève 說：「如果我們要抽取 Lévy 的機率中心思想，那麼我們可以這樣說，自從 1919 年之後，他的研究主題曲就是正規分佈律。他一而再，再而三地以它為出點，並且堅決地又回到它……他是帶著隨機時鐘沿著隨機過程的樣本路徑作旅行的人，因此他經常在極短的時間內就會遭遇到無窮多的機運現象，從而有“Lévy demon”之美稱。」

美國國家標準局的顧問 W. J. Youden (1900-1971) 更用一段文字說明正規分佈律的重要性及其泛應用，以打字的方式排成標準正規分佈的圖形，妙趣橫生，如下圖：



THE
NORMAL
LAW OF ERROR
STANDS OUT IN THE
EXPERIENCE OF MANKIND
AS ONE OF THE BROADEST
GENERALIZATIONS OF NATURAL
PHILOSOPHY - IT SERVES AS THE
GUIDING INSTRUMENT IN RESEARCHES
IN THE PHYSICAL AND SOCIAL SCIENCES AND
IN MEDICINE AGRICULTURAL AND ENGINEERING
IT IS AN INDISPENSABLE TOOL FOR THE ANALYSIS AND THE
INTERPRETATION OF BASIC DATA OBTAINED BY OBSERVATION AND EXPERIMENT

意思是說：「誤差的正規分佈律在人類的經驗中具有“鶴立雞群”的地位，為自然哲學最廣含的推廣之一。它是各方面研究的指南，如物理學、社會科學、醫學、農業及工程等等。對於由測與度量所得的基本數據之解釋，它是不可或缺的工具。」

欲明正規分佈律更深刻的特性及重要性，需涉及高深一點的機率論，在此我們不預備介紹。

六、結語：科學方法論簡介

本文討論的兩個主題是，誤差現象以及由觀測數據找函數關係。

觀測度量含有誤差，這是人類的普遍經驗。在常識作法上，克服誤差的辦法是多作幾次度量，然後用「算術平均」來估計真值。這是我們所能做到的「最佳」結果。

其次，假設一個現象的兩個變量 x, y 它們具有函數關係 $y = f(x)$ ，但是我們不知道 f 。我們想探求 f 。這當然有許多辦法，端視我們對 f 能夠擁有什麼樣的部分資訊 (partial informations) 而定。例如，如果我們知道 f 在某一點的取值 (初期條件) 以及 f 的導函數 f' (變化機制，通常由物理定律給出)，那麼由微積分的 Newton-Leibniz 公式就可以求出 f 。這是常微分方程 (O.D.E.) 的主題。另外，如果我們直接叩問自然，得到 (22)

式之觀測數據，那麼常識性的做法是找 f ，使 (23) 式取最小值 (最小平方法)。

面對這些常識性的作法 (可能很好，也可能似是而非或錯誤)，科學要進一步追究「爲什麼」？換言之，科學要將常識經驗加以精煉，提煉出理論來說明「爲什麼」。事實上，人類因追究「爲什麼」而進步。

科學的探索步驟，歸結起來，我們不妨稱之爲「三層樓」觀點：

(一) 就一個有趣的問題或現象，收集資料或數據；在統計學上，叫做抽樣；在自然科學上，叫做實驗。這是科學的經驗基礎。

(二) 資料或數據的整理、歸類、化約、分析、表現、作圖表、找尋關係與規律等等，這是記述統計學的內容，必要的準備工作。

(三) 提出一套理論來解讀資料與數據背後的意義。用理論來適配 (fits) 既知，並且預測 (predicts) 或探索未知。理論是我們的「新眼睛」，用它來普照經驗世界的山河大地。這比美於釋迦牟尼悟道後所說的：「眼睛長出來了，光明生出來了。」

這三層樓的求知活動，不見得是由一個人單獨完成，許多時候是分工合作的結果。例如行星運動的研究，Tycho Brahe 與 Kepler 分別完成第一、二層樓的工作，接著牛頓提出萬有引力定律才完成第三層樓的建構。又如，Darwin(1809-1882) 的進化論，基本上是他一個人完成的。自從 1831 年到 1836 年之間參加 Beagle 號全球環遊的調查航行 (他說這是心智上的真正教育與歷練)，他開始蒐集動植物標本、地質資料，... 等等，長達十七年之久；接著是整理、分類、分析工作；最

後在 1859 年出版「物種源始」(the origin of species)，提出他的物種進化理論。

Gauss 對誤差論及最小平方法的研究，也都經歷這三層樓的發展。這裡尤其涉及到機率論與統計學的方法論。機率論是演繹的、綜合的、由上而下的，即由隨機現象的機率模型 (三合一又三合一)，計算出有趣的複雜事件之機率，並且推導出種種機率法則與必然的邏輯結論。統計學則反其道而行，在機率模型中含有某個未知，例如機率密度函數未知或機率密度函數的型態雖已知但其中含有未知參數，我們要透過抽樣數據來猜測未知。講究這個猜測的藝術 (the art of conjecturing) 之學問就是統計學。因此，統計學是分析的、歸納的、由下而上的。統計學是數據的物理學 (Statistics is the Physics of numbers)。機率論與統計學恰好形成了一個互補而圓足的研究迴路，統治著機運王國 (the empire of chance)。但是，作爲探索未知的科學方法，非常寬廣多樣，統計方法只是其中的一種而已。

參考文獻

1. Born, M. : Natural Philosophy of Cause and Chance, Clarendon Press, Oxford, 1948.
2. Stigler, S. M. : The History of Statistics, The measurement of Uncertainty before 1900, Harvard Univ. Press, 1986.
3. Maistrov, L. E. : Probability theory, A Historical Sketch, Academic Press, 1974.
4. Popper, K. R. : Conjectures and Refutations, the growth of scientific knowledge, Routledge and Kegan Paul, London, 1963.
5. 蔡聰明: 什麼是機率與機率法則? 數學傳播, 第十九卷第一期, 1995.

—本文作者任教於國立台灣大學數學系—