機率專題之三

機率與訊息

劉豐哲

本文作者現任職於本所

§1. 從彩燈談起

我們知道彩燈是由好幾個燈泡串起來的。要是其中有一個燈泡是壞的，則整串的燈泡就都不會亮了。如果手邊的這串彩燈不亮了，我們該怎麼樣才能儘快的找出壞的來呢？爲了簡化討論，我們假定彩燈裏一共有m個燈泡，而其中只有一個是壞的。通常的檢驗辦法是把測儀的兩根引線接到彩燈線路中的兩點，以斷定壞燈是不是在這兩點之間。在這樣的做法下，讀者不難看出，用測儀檢驗m次便足以找出壞燈。（依次的把測儀接在每一盞燈的兩邊）。很自然的，我們要問，至少要檢測幾次才能保證找到壞燈。

如果$m=2^k$，k爲一非負的整數，則我們可以證明：至少需要測k次，而且有方法在k次內找出壞燈。（只要想一下$k=0,1$的情況，然後用歸納法處理一下即可。即使一時想不出來，請不要流連其中！）

如果我們認爲這個結果不錯，而自覺滿意的話，容我們提醒自己：**在文明的進展中，重要的突破常常是歸功於那些能從簡單的事理中跨出一步的人；而這正是一般人懶得舉足的一步。**

讓我們看看，我們所得到的簡單結論，到底說明了什麼。$k=0$時，只有一個燈，這時無需測，壞的就是這一個；這是確定的情況。$k=1$時，只需測其中一個燈的兩端，就知道壞的是哪一個；在未測之前，每個燈都有壞的可能性，機率各爲 1/2。當k增大時，在未測之前，每個燈壞的可能性變小，要找出壞燈就變難了。用機率論的術語來說則是：當k增大時，壞燈所在的位置的**隨機性**也隨著增大。

各位不覺得上文中的「隨機性」使用得太含糊了嗎？戴著瓜皮帽的秀才爺爺會說，“含糊不要緊，能會意就行，就行。”但是，有點數學訓練的人就會追問：**隨機性能不能量？**

至少，在上面的例子中，隨機性是可以量的。我們知道，當$m=2^k$時，找出壞燈至少需要檢測k次才行；因此用k來表示，壞燈所在位置的隨機性是極爲自然的：把每次的檢測想成是我們所得到的一個訊息，（或「一拍」訊息，更爲傳神！），則總共需要k拍訊息才能定出壞燈的位置。

把這種論法推廣到比較一般的隨機性問題，就是訊息理論的內容。它是美國數學家兼工程師顯農氏（Claude Shannon）在 1947～1948 年間提出的。顯農氏提出訊息論的目的是解決訊號傳遞問題上的一些困難；近年來訊息論已成功地應用到許多科學的分枝；特別是它的主要概念——試驗的熵數（Entropy of Experiements）〔註〕——經蘇聯數學家 A. N. Kolmogoroff 的修訂之後在數學上有了極爲突出的貢獻。

本文的目的有二：一來淺介訊息理論，二來藉訊息理論之介紹說明機率論的方法。

首先，我們回憶一下機率論的基本概念。

〔註〕熵數的意思就是指「平均訊息量」。

§2. 機率論的基本概念

我們假定讀者具有最基本的機率論的常識。本節的目的是利用簡化的一般性討論來得到我們所需要的簡單模型——有限機率空間。機率空間與實際問題之間的關係是初學者最感迷惑的東西；初學者常常問：爲什麼需要機率空間？的確，在通常的實際問題中，事件的機率都非常顯明易見，我們只不過是採用一些計算機率的「規則」，以某些已知事件的機率來求出另外一些事件的機率。界定機率空間似乎是小題大作，多此一舉，徒然擾亂方寸（心）！關於這個問題，我們要從兩個層面來做總結的說明。第一個層面是計算機率時所採用的「規則」的合法性。這些規則所依靠的一些概念（例如隨機獨立性）常依各人的詮釋而異，因而產生了許多詭論。（請參閱本刊同期熊昭教授文：淺談幾個或然率上的詭論——編者）機率空間的提出就在建造一個適當的體系，使得相關的主要概念都能具有明確的意義，並且把大家默許的，不經意所採用的計算規則做一個清楚的說明，以確定它的角色與地位。第二個層面是我們所建造的機率空間到底跟原先所出發的實際問題有什麼關係，是不是充分的描述了原先的問題。這是一個哲學性的問題；對於科學工作者而言，這個問題的掌握便是實踐。其實，每一門數學都有這兩個層面；機率論的這兩個層面較難領會，一來機率論的年齡尚輕，我們的實踐還不充分，二來機率論的經驗基礎和其它數學比較起來，是在人類生活體驗中屬於較高層面的。這些問題的討論不是本文的目的。我們的重點是在傳達給讀者一個簡單而有用的觀點：在觀察隨機現象時，我們得到了許多資料，將這些資料去蕪存精，最後所列出來的簡明「圖表」就是機率空間。換句話說，**一個機率空間就是觀察某個隨機現象的實驗總結。**

在下面的討論中，例子很少。我們相信由讀者自己提供例子，更易達到本節的目的。

在電阻爲一歐姆的電阻器的兩端施以一伏特的電壓，然後量一量通過的電流強度，即得一安培；如果施以二伏特之電壓，量得的電流強度則爲二安培。在這個實驗中，決定電流強度的要件是電阻器兩端的電位差；由於電位差是我們能控制的，所以電流強度也是我們所能控制的。但是有的實驗卻不這樣，它的結果依著一些我們不知道的，或者我們無法控制的因素而變化；我們無法預知結果。例如隨手丟擲一枚硬幣時，我們無法預知最後出現的是正面還是反面，因爲它是由硬幣丟出瞬間的狀態、所落地面以及硬幣的各種物理性質決定；而這些因素，對隨手丟出硬幣者來講，是未知的或無法控制的因素。另外，觀察孕婦生男生女，觀察丟擲骰子出現的點數，也都是屬於這類性質的實驗。我們把這類實驗稱爲**隨機實驗**，或簡稱**試驗**。

我們可以把試驗看成是對某一隨機現象的片面觀察，例如在丟擲兩顆骰子時，觀察它們出現點數之和，或觀察點數和爲奇或爲偶數，都是對「丟擲兩顆骰子」這個隨機現象的片面觀察。科學上只考慮能夠一再地予以獨立觀察的現象，因爲只有這種現象才有可能做科學分析。我們所考慮的隨機現象也必須如此。譬如說，丟擲銅板便是一個可以重覆獨立觀察的現象，不同的人可以各自獨立地丟擲銅板，觀察銅板最後出現的面，或者同一個人也可以一再丟擲同一銅板，這些都不會影響現象本身的狀況。我們在這兒介紹的是機率論的簡單部份，因此此段假定試驗的可能結果只有有限多個。我們知道試驗的目的是在瞭解現象，如果我們無法從試驗的各個結果中找出任何規律，那麼就不可能對相關的現象提出科學性的結論，這種現象對人們而言便是迷惑的現象，須要做進一步的觀察才行。在這裡，我們只考慮具有下述規律的試驗（以後稱之爲**隨機規律性**）：

假設 $A_1, A_2, \cdots\cdots, A_n$ 爲某一試驗的所有可能結果。如果在多次獨立地觀察這個試驗的結果時，各個結果 A_i 出現的次數與試驗次數之比圍繞在某個固定數 p_i 的附近，則我們說這個試驗具有隨機規律性。$p_1, p_2, \cdots\cdots, p_n$ 這些數表現了 $A_1, A_2, \cdots\cdots, A_n$ 這些結果出現的規律性。當然，$p_i \geq 0$，$\sum_{i=1}^{n} p_i = 1$。p_i 稱爲 A_i 出現的機率，或簡稱爲 A_i 的機率。$A_1, A_2, \cdots\cdots, A_n$ 及 $p_1, p_2, \cdots\cdots, p_n$ 是多次獨立地觀察試驗所得資料的總結，因此我們用

$$<\begin{matrix}A_1, A_2, \cdots\cdots, A_n\\ p_1, p_2, \cdots\cdots, p_n\end{matrix}>$$

表示這個試驗，或簡記爲$<A_1, A_2, \cdots\cdots, A_n>$。

爲了簡單起見，我們只考慮某個隨機現象的有限多個試驗。在下面的討論中，我們先固定一個隨機現象；假設 $\alpha_1, \alpha_2, \cdots\cdots, \alpha_m$ 代表m個該隨機現象的試驗。我們用符號把 α_i 寫成 $\alpha_i=<A_1^{(i)}, A_2^{(i)}, \cdots\cdots, A_{n_i}^{(i)}>$。也就是說，在 α_i 這個試驗裏，一共有 n_i 種可能的結果，我們分別把它們用 $A_1^{(i)}, A_2^{(i)}, \cdots\cdots, A_{n_i}^{(i)}$ 表示出來。另一方面，我們可以把這m個試驗合在一起，看成是一個新的試驗。在新的試驗裏，觀察的結果將如下表示：$(A_{j_1}^{(1)}, A_{j_2}^{(2)}, \cdots\cdots, A_{j_m}^{(m)})$，其中$1 \leq j_i \leq n_i$。也就是說，在我們用這$m$個試驗來觀察同一隨機現象時，如果依次觀察到結果分別是 $A_{j_1}^{(1)}, A_{j_2}^{(2)}, \cdots\cdots, A_{j_m}^{(m)}$，則把它們集在一塊，看成是一個新的試驗的結果。我們用 $\alpha_1 \vee \alpha_2 \vee \cdots\cdots \vee \alpha_m$ 來表示以所有的 $(A_{j_1}^{(1)}, A_{j_2}^{(2)}, \cdots\cdots, A_{j_m}^{(m)})$ 爲可能結果的這個新的試驗。

舉個例吧！在隨手丟擲兩顆可分辨的骰子時，我們用 α_1 及 α_2 分別表示觀察第一顆及第二顆骰子出現的點數。即

$$\alpha_1=<A_1^{(1)}, A_2^{(1)}, \cdots\cdots, A_6^{(1)}>,\quad \alpha_2=<A_1^{(2)}, A_2^{(2)}, \cdots\cdots, A_6^{(2)}>$$

其中 $A_i^{(1)}$、$A_i^{(2)}$ 分別表示第一顆與第二顆骰子出現之點的情形。這時，

$$\alpha_1 \vee \alpha_2=<(A_1^{(1)}, A_1^{(2)}), (A_1^{(1)}, A_2^{(2)}), \cdots\cdots, (A_6^{(1)}, A_6^{(2)})>。$$

平常我們是把 $<A_i^{(1)}, A_j^{(2)}>$ 寫成 (i, j)，把 $\alpha_1 \vee \alpha_2$ 寫成 $<(1,1), (1,2), \cdots\cdots, (1,6), (2,1), \cdots\cdots, (6,6)>$。一般來說，我們把 $\alpha_1, \cdots\cdots, \alpha_m$ 想成是觀察同一隨機現象的 m 個儀器的記錄，那麼 $\alpha_1 \vee \cdots\cdots \vee \alpha_m$，就是這 m 個儀器的綜合記錄。以後我們將 $\alpha_1 \vee \cdots\cdots \vee \alpha_m$ 的每個可能結果稱爲基本事件。當然，在觀察 $\alpha_1 \vee \cdots\cdots \vee \alpha_m$ 時，有一個而且僅有一個基本事件出現。爲了方便起見，我們將每個基本事件視爲一點，而將所有這些點合攏起來記作 $\{\omega_1, \omega_2, \cdots\}$，這是個具有 $n_1 \times n_2 \times \cdots\cdots \times n_m$ 個點的集合，記之爲 Ω。Ω的任一子集A可用來表示落於A的那些基本事件的聯合事件，聯合事件A發生的意思是指A中的某個基本事件發生了。在上面隨手丟擲兩顆骰子的例子中，$\{(1,1), (1,2), (1,3), (1,4), (1,5), (1,6)\}$ 即表示第一顆骰子出現1點的事件。從現在起就稱Ω的子集爲事件，Ω的空子集ϕ爲不可能事件，Ω本身爲必然事件，而 $\{\omega_i\}$ 爲基本事件。爲了符號上的方便，我們常常不分辨 $\{\omega_i\}$ 和 ω_i。如果 A, B 爲兩個事件，則$A \cup B$稱爲 A、B 的聯合事件，$A \cap B$稱爲 A、B 的共同事件；$A \cap B = \phi$ 時，稱 A、B 爲互斥事件。

我們強調過，我們只考慮具有隨機規律性的試驗，因此，我們要求Ω的每個點 ω_i 具有機率 p_i，當然，$p_i \geq 0$，$\Sigma_i p_i = 1$。如果 $A \subset \Omega$，令 $P(A)=\Sigma_{\omega_i \in A} p_i$，這兒 $\Sigma_{\omega_i \in A} p_i$ 是指所有落於A的基本事件的機率的和；$P(A)$ 稱爲事件A的機率。根據前面所說的 p_i 的意義，就知道在多次獨立觀察試驗 $\alpha_1 \vee \alpha_2 \vee \cdots\cdots \vee \alpha_m$ 時，事件A出現的頻率將會圍繞於 $P(A)$ 的附近。

機率論的目的是在研究隨機現象裏的經驗規律的數學關係及其應用。由於上述的集合Ω與基本事件的機率綜合地表達了多次獨立觀察 $\alpha_1, \alpha_2, \cdots\cdots, \alpha_m$ 所得到的經驗規律，原先隨機現象的經驗規律變成了事件的機率，因此，機率論的目的也就是在研究機率所滿足的數學關係及其應用；就好像三角學研究的是三角函數間的關係及其應用一樣。

從數學的觀點來看，我們的起點是一個有限集合 Ω，而對 Ω 的每一個子集A，我們都附上了一個數 $P(A)$，這些 $P(A)$ 滿足下述的關係：

(1) $0 \leq P(A) \leq 1$，$P(\phi)=0$，$P(\Omega)=1$；

(2)如果 A、B 爲Ω之子集，且$A \cap B = \phi$，則 $P(A \cup B)=P(A)+P(B)$

換句話說，**有限集合Ω及滿足(1)和(2)的函數P就是數學的起點**。Ω和P合起來稱爲機率空間，簡記作

(Ω, P)。這時的機率空間 (Ω, P) 可以不具有任何實際意義，它僅僅是個數學系統；雖然如此，我們仍將 (Ω, P) 解釋爲對某個隨機現象作有限多個試驗所觀察到的經驗規律的總結，因此 Ω 的點仍然稱爲基本事件，Ω 的子集稱爲事件，$P(A)$ 稱爲 A 的機率，……。我們在機率空間裏新得到的結論，在經過這樣的解釋（翻譯）之後，就可以用來描述和預測某些隨機現象，這點非常重要，否則機率空間的研究就像是下棋、玩牌了，無法在人類文明的發展中產生積極的作用。我們抽象地研究機率空間，是希望能夠一目瞭然地看出機率與其相關概念間的關係；免得被形形色色的個別例子淹蓋了要點。我們須要適當的解釋機率空間，才能走向正確的方向（瞭解隨機現象），使得我們的心智活動不至於淪爲單純的數學遊戲。這一點，各門數學都一樣，我們不想多言。

以下我們形式地定義一些最有用的概念，而將唇齒相關的直覺意義，留給讀者自己去補上。（同時請參閱本刊同期楊維哲先生撰：機率一講。——編者）

【定　義】

(1) 如果 $P(A\cap B)=P(A)\cdot P(B)$；則我們說事件 A 與 B 互爲**隨機獨立**。（或簡稱獨立）。

(2) 如果 $P(B)\neq 0$，則 $P(A\cap B)/P(B)$ 稱爲 **A 在已知 B 時的條件機率**，記作 $P(A|B)$；

(3) 如果 $A_1, \cdots\cdots, A_n$ 爲 n 個互斥的事件，而且 $A_1\cup\cdots\cdots\cup A_n=\Omega$；則稱 $A_1, \cdots\cdots, A_n$ 組成 Ω 的一個**分割**，並記之爲 $<A_1, \cdots\cdots, A_n>$。**分割**又稱**試驗**，這時爲了明白表示每個事件 A_i 的機率，我們常把試驗記作

$$\left<\begin{matrix}A_1, & \cdots\cdots, A_n\\ P(A_1), & \cdots\cdots, P(A_n)\end{matrix}\right>。$$

(4) 設

$$\alpha=\left<\begin{matrix}A_1, & \cdots\cdots, A_n\\ P(A_1), & \cdots\cdots, P(A_n)\end{matrix}\right>,\quad \beta=\left<\begin{matrix}B_1, & \cdots\cdots, B_m\\ P(B_1), & \cdots\cdots, P(B_m)\end{matrix}\right>$$

爲兩個試驗，則

$$\left<\begin{matrix}A_1\cap B_1, & \cdots\cdots, A_n\cap B_m\\ P(A_1\cap B_1), & \cdots\cdots, P(A_n\cap B_m)\end{matrix}\right>$$

稱爲 α 和 β 的**合成試驗**，記作 $\alpha\vee\beta$，如果任何 A_i 和任何 B_j 都是隨機獨立的，即 $P(A_i\cap B_j)=P(A_i)\cdot P(B_j)$，$i=1, \cdots\cdots, n$；$j=1, \cdots\cdots, m$。則稱 α 和 β 互爲隨機獨立（或簡稱獨立）。

(5) 我們把定義在 Ω 上的任一實數值函數叫做**隨機變數**。通常用 x, y, z 等小寫字母來表示隨機變數。假設 x 是隨機變數，則 $\sum_{\omega\in\Omega}x(\omega)\ P(\omega)$ 稱爲 x 的**期望值**，記作 Ex；如果 r 是個實數，則用 $\{x=r\}$ 來表示 $\{\omega\in\Omega: x(\omega)=r\}$ 以節省筆墨，$\{x=r\}$ 是 x 取值爲 r 的事件。假設 $r_1, \cdots\cdots, r_n$ 爲 x 所可能取的值，則 $<\{x=r_1\}, \cdots, \{x=r_n\}>$ 稱爲由 x 所決定之試驗。如果 x 所決定的試驗與 y 所決定的試驗是獨立的話，則稱 x 與 y 互爲獨立的隨機變數（或簡稱 x 與 y 爲獨立變數）。

(6) 爲了方便起見，如果 $B\subset\Omega$ 是個事件，則如下定義的函數 X_B 稱爲 B 的**指示函數**：$X_B(\omega)=1$，如果 $\omega\in B$；否則 $X_B(\omega)=0$。如果 x 是個隨機變數，$<\{x=r_1\}, \cdots, \{x=r_n\}>$ 是 x 所決定的試驗，則 x 可表成

$$x=\sum_{i=1}^{n} r_i X_{\{x=r_i\}}\ ;$$

而

$$Ex=\sum_{i=1}^{n} r_i P\{x=r_i\}。$$

§3. 試驗的熵數

現在我們要簡單地討論訊息論裏最根本最重要的概念——試驗的熵數。至於它的應用，則要到下一節再討論。

假定我們現在用某個試驗來觀察一個特定的隨機現象，我們自然會問自己：能不能找到一個適當的量來度量該試驗所能提供的**訊息**（Information）？換一個角度來看，在試驗之前，我們無法預知會出現什麼結果，因此我們說試驗具有**隨機性**；試驗之後我們知道結果了，隨機性就消失了，消失了的隨機性可以看成是我們所獲得的訊息。所以，事實上，我們等於在問：能不能找到一個適當的量來度量該試驗的隨機性？

假定 $A_1, A_2, \cdots\cdots, A_n$ 是某個試驗的所有可能結果。如果 $P(A_i)>P(A_j)$，則「A_j 出現」比「A_i 出現」要使我們驚奇，就如像稀有的社會事件具有非常的新聞價值一樣。也就是說，機率不同的結果會提供不同的訊息。因此，用來量度該試驗所能提供的訊息（或試驗的隨機性）的那個量，必須是各個結果所提供的訊息的某種平均值。首先，我們來看看如何度量個別事件所能提供的訊息。我們假定有一個這樣的量，而用 $I(A)$ 來代表事件 A 所能提供的訊息，則 $I(A)$ 應當滿足：

(1) $I(A)\geq 0$；

(2) $I(A)$ 完全由 A 的機率決定，換句話說，

$$I(A)=\vartheta(P(A)),\ \vartheta \text{ 是個定義在 } 0 \text{ 到 } 1 \text{ 之間的函數；}$$

(3) 如果 $P(A)\geq P(B)$，則 $I(A)\leq I(B)$。

條件(1)僅僅表示取定一個適當的準點；條件(2)是強調機率的特點，我們說過事件的隨機規律性是由它的機率來代表的，因此和事件有關的重要數量也該是完全由事件的機率來決定。現在我們進一步考慮兩個獨立事件 A 和 B。$A\cap B$ 可以解釋爲在 A 出現的情況下，B 又出現的事件，但是 A 和 B 是獨立的，由 A 出現所得到的訊息，應該無法幫助我們預測 B 出現的可能性，因此已知 A 出現後，又知道 B 出現所提供的訊息，應當爲 A, B 各別出現所得訊息之和，亦即 $I(A\cap B)=I(A)+I(B)$。相應地，我們要求 ϑ 滿足：

(4) $\vartheta(pq)=\vartheta(p)+\vartheta(q)$，$p, q\in[0,1]$。

將(1)，(2)，(3)，和(4)綜合起來，就是一個定義〔0, 1〕在上、滿足(4)的遞減函數 ϑ。這種函數很多，譬如說，$\vartheta(t)=-\log_a t$，$t\in[0,1]$，a 爲某個正實數。這兒取不同的 a 僅僅表示選取不同的單位長度。在下面，我們取 $\vartheta(t)=-\log_2 t$，也就是令 $I(A)=-\log_2 P(A)$。

在進一步討論試驗的訊息之前，我們回頭看看 §1. 的例子。我們有 2^k 個燈泡，其中一個壞了；在檢測之前，我們認爲每個燈壞的機率都是一樣的，都是 2^{-k}。令 A_n 爲第 n 個燈泡壞了的事件，則 $I(A_n)=-\log_2 P(A_n)=-\log_2 2^{-k}=k$。這時，測出任何一個壞燈所能提供的訊息皆爲 k，因此 k 度量著測出壞燈位置所獲得之訊息，也就是壞燈位置的隨機性。一般來說，如果試驗中的每個結果具有同樣的可能性（出現的機率一樣），則 $-\log_2(1/n)=\log_2 n$ 代表著該試驗所提供的訊息，其中 n 是所有可能結果的個數。譬如說，丟擲一枚非偏倚銅板，觀察正面或反面出現所得的訊息爲 $-\log_2(1/2)=\log_2 2=1$。歷史上，第一個考慮試驗熵數的人是美國電訊工程師 Hartley（1928 年），他把試驗熵數定義爲 $\log_2 n$。他只考慮到試驗中可能出現的結果的個數，卻忽略了每個結果出現的機率。這個概念在 1947～1948 年間由顯農氏予以修正，而成了目前數學家和工程師所採用的形式。

假定 $A_1, A_2, \cdots\cdots, A_n$ 是某個試驗的所有可能結果。我們知道事件 A_i 所能提供的訊息是 $-\log_2 P(A_i)$，因此，觀察一次試驗所得到的訊息是個隨機變數。這個隨機變數的期望值就是多次獨立觀察該試驗所得的訊息的平均值。顯農氏把這平均值叫做試驗的熵數。形式上說，如果 $\alpha=\left<\begin{matrix}A_1, & \cdots\cdots, & A_n\\ P(A_1), & \cdots\cdots, & P(A_n)\end{matrix}\right>$ 是機率空間 (Ω, P) 的一個試驗，則 α 的熵數 $H(\alpha)$ 是定義爲

$$H(\alpha)=-\sum_{i=1}^{n} P(A_i)\ \log_2 P(A_i)\ 。$$

如果我們用 X_B 表示Ω中事件B的**指示函數**，則 $H(\alpha)=E\mathfrak{x}$，其中 $\mathfrak{x}=-\sum_{i=1}^{n} X_{A_i}\cdot \log_2 P(A_i)$。要是在 $H(\alpha)$ 的定義中，某個事件 A_i 的機率 $P(A_i)=0$，則令 $P(A_i)\log_2 P(A_i)=0$。這是有道理的，因爲$\lim_{t\to 0} t\log_2 t=0$，另外，如果我們令 $\eta(t)=-t\log_2 t$，則 $H(\alpha)$ 可以簡單的寫成

$$H(\alpha)=\sum_{i=1}^{n} \eta(P(A_i))\ 。$$

好了，在這些緊湊的抽象討論後，我們來回頭看看那串彩燈吧！依照上述的符號，我們要問的是試驗 $\alpha=\left<\begin{matrix}A_1,\ \cdots\cdots, A_{2^k}\\ 1/2^k, \cdots\cdots, 1/2^k\end{matrix}\right>$的熵數 $H(\alpha)$。根據剛才的定義。

$$H(\alpha)=-\sum_{i=1}^{2^k}\frac{1}{2^k}\log_2 2^{-k}=k\ ,$$

這正是我們最初的意思。

試驗的熵數具有下述性質：假設$\alpha=<A_1, A_2, \cdots\cdots, A_n>$，$\beta=<B_1, B_2, \cdots\cdots, B_m>$爲機率空間$(\Omega, P)$的兩個試驗，則

(i) $H(\alpha)\geq 0$；$H(\alpha)=0$的充要條件是某個 A_i 爲必然事件，而其餘的均爲不可能事件。

(ii) $H(\alpha)\leq\log_2 n$；$H(\alpha)=\log_2 n$ 的充要條件是 $P(A_1)=P(A_2)=\cdots\cdots=P(A_n)=1/n$。

(iii) 如果α和β是獨立的，則 $H(\alpha\vee\beta)=H(\alpha)+H(\beta)$

(i) 的證明很簡單，(ii) 的證明與機率無關，因此，我們略掉它們，而來證明 (iii)。根據定義，$\alpha\vee\beta=<A_1\cap B_1, \cdots, A_n\cap B_m>$。由於$\alpha$與$\beta$是獨立的，

$$P(A_i\cap B_j)=P(A_i)P(B_j)$$

因此

$$\begin{aligned}H(\alpha\vee\beta)&=-\sum_{i,j}P(A_i\cap B_j)\log_2 P(A_i\cap B_j)\\&=-\sum_{i,j}P(A_i)P(B_j)[\log_2 P(A_i)+\log_2 P(B_j)]\\&=-(\sum_i[P(A_i)\log_2 P(A_i)])(\sum_j P(B_j))-(\sum_j P(B_j)\log_2 P(B_j))(\sum_i P(A_i))\\&=-\sum_i P(A_i)\log_2 P(A_i)-\sum_j P(B_j)\log_2 P(B_j)\\&=H(\alpha)+H(\beta)。\end{aligned}$$

性質 (i) 可解釋爲：如果在某試驗中，會有一個必然事件產生，則觀察這個試驗是不會提供任何訊息的，也就是說，這個訊息沒有隨機性。當著試驗中的各個事件具有同樣的機率時，我們把它叫做非偏倚試驗。性質 (ii) 告訴我們，在具有n個事件的試驗中，非偏倚試驗的隨機性最大，其熵數爲 $\log_2 n$。這是合乎直覺要求的；因爲，在觀察偏倚試驗時，我們是預先就知道了某些事件比較容易發生，而另一些事件是比較不容易發生；這種含糊的預知就說明了偏倚試驗的隨機性比較小。其實，如果偏倚到了極點，就沒有隨機性了，而這正是性質 (i) 所要描述的。依照前面的說法，$\alpha\vee\beta$指的是同時觀察α和β兩個試驗，依此，性質 (iii) 可以如下敍述：如果α和β是獨立的試驗，則同時觀察α和β所得的訊息爲分別觀察α和β所得訊息的和。總結起來，我們所定義的試驗的熵數的確是描述了我們所預期的各項簡單性質，這些就注定著它會是一個重要而有用的概念。

在介紹熵數的其他性質之前，我們先談談條件熵數。假設 $\alpha=<A_1, \cdots\cdots, A_n>$爲一試驗，$B$爲一事件。則

$$\left<\begin{matrix}A_1\cap B,\ \cdots\cdots, A_n\cap B\\ \qquad\qquad n\end{matrix}\right>$$

也可以看成一個試驗。這個試驗是在 B 已經發生的情況下來觀察 α 的試驗。我們把這個試驗記爲 $\alpha\,|\,B$。試驗 $\alpha\,|\,B$ 的熵數 $H(\alpha|B)$ 表示着在事件 B 已經發生的情況下，試驗 α 所留存的隨機性。例如 $B=A_i$，則在 A_i 出現的情況下，α 已不具有任何隨機性，因此 $H(\alpha|A_i)=0$。（這點可以很容易的從性質 (i) 導出。）$H(\alpha|B)$ 稱爲試驗 α 相對於事件 B 的條件熵數。假設 $\beta=<B_1,\cdots\cdots,B_m>$ 爲另一試驗，令當

$$x=\sum_j X_{Bj}H(\alpha|B_j)。$$

B_j 發生時，x 就是 $H(\alpha|B_j)$。我們把 x 的期望值記爲 $H(\alpha|\beta)$。$H(\alpha|\beta)$ 量度的是在觀察了試驗 β 之後，α 所留存下來的隨機性。我們把 $H(\alpha|\beta)$ 叫做 α 相對於 β 的條件熵數。顯然的，$H(\alpha|\alpha)=0$，熵數的另外兩個重要性質是：假設 α,β 是兩個試驗，則

(iv) $H(\alpha\vee\beta)=H(\beta)+H(\alpha|\beta)$

(v) $0\leq H(\alpha|\beta)\leq H(\alpha)$

性質 (iii) 是性質 (iv) 的特例，而性質 (iv) 的證明又跟性質 (iii) 的完全一樣，只要把 $P(A_i\cap B_j)=P(A_i)P(B_j)$ 換成 $P(A_i\cap B_j)=P(B_j)P(A_i|B_j)$ 就行了。性質 (v) 的證明與機率概念無關，所以省略。不過，我們要提醒讀者一點：在直覺上，性質 (v) 是極爲顯然的，因爲在觀察 β 之後，我們多多少少會得到些訊息，這些訊息只可能減少 α 的隨機性。另外，從 (iii) 和 (iv) 可以看出，如果 α 和 β 是獨立的，則 $H(\alpha|\beta)=H(\alpha)$。

從上段的討論可以看出 $H(\beta)-H(\beta|\alpha)$ 量度的是試驗 β 在觀察了試驗 α 之後所減少的隨機性。因此，我們可以把 $H(\beta)-H(\beta|\alpha)$ 看成是 α 提供給 β 的訊息。我們把 $H(\beta)-H(\beta|\alpha)$ 記爲 $I(\alpha,\beta)$。有時候，$I(\alpha,\beta)$ 叫做 β 存於 α 中的訊息。由於 $H(\alpha\vee\beta)=H(\alpha)+H(\beta|\alpha)=H(\beta)+H(\alpha|\beta)$，我得到下面的關係式：

$$I(\alpha,\beta)=H(\beta)-H(\beta|\alpha)=H(\alpha)-H(\alpha|\beta)=I(\beta,\alpha)$$

在應用的時候，β 是我們要研究的對象，α 是爲了消除 β 的隨機性而考慮的輔助試驗。這些都將在下節裏詳細討論。

§4. 應　用

(一) 我們要利用 §3. 的概念來回答 §1. 的彩燈問題。如上所述，m 代表彩燈中燈泡的個數。如果我們用 γ 來表示描述壞燈位置的試驗，則 $H(\gamma)=\log_2 m$。在 §1. 中，我們是把測儀的兩根引線接到線路中的兩點以斷定兩點之間有沒有壞燈，這個做法其實也是一個試驗，它的可能結果是壞燈在兩點之間與壞燈在兩點之外。我們來算算至少要檢測幾次才能保證找到壞燈。先假設 k 次可以保證找到。把這 k 次的試驗分別用 $\alpha_1,\alpha_2,\cdots\cdots,\alpha_k$ 表示。由於 α_i 只有兩個可能結果，所以 $H(\alpha_i)\leq 1$ (§3. 之(ii))。k 次就能保證找到壞燈的意思是

$$I(\alpha_1\vee\cdots\cdots\vee\alpha_k,\gamma)=H(\gamma)=\log_2 m$$

由於

$$\begin{aligned}I(\alpha_1\vee\cdots\cdots\vee\alpha_k,\gamma)&\leq H(\alpha_1\vee\cdots\vee\alpha_k)=H(\alpha_1)+H(\alpha_2\vee\cdots\vee\alpha_k|\alpha_1)\\&\leq H(\alpha_1)+H(\alpha_2\vee\cdots\vee\alpha_k)\leq H(\alpha_1)+\cdots+H(\alpha_k)\leq k\end{aligned}$$

所以 $k\geq\log_2 m$。

要使得 k 愈小愈好，就得要求 $I(\alpha_1,\gamma)$ 愈大愈好；用俗語說，就是要使得 α_1 能夠對 γ 提供愈多的訊息愈好。假設在 α_1 中，兩根引線之中有 n 個點，則

$$\alpha_1=\begin{pmatrix}A & B\\ \frac{n}{m} & \frac{m-n}{m}\end{pmatrix},$$

其中 A 是指壞燈在兩根引線之內，B 是指在引線之外。因此，

$$I(\alpha_1,\gamma)=H(\gamma)-H(\gamma|\alpha_1)=\log_2 m-H(\gamma|\alpha_1),$$
$$H(\gamma|\alpha_1)=H(\gamma|A)P(A)+H(\gamma|B)P(B)$$
$$=(\log_2 n)P(A)+(\log_2(m-n))P(B)$$
$$=(\log_2\frac{n}{m}+\log_2 m)\frac{n}{m}+(\log_2\frac{m-n}{m}+\log_2 m)\frac{m-n}{m}$$
$$=(\log_2\frac{n}{m})\frac{n}{m}+(\log_2\frac{m-n}{m})\frac{m-n}{m}+\log_2 m$$
$$=-\eta(\frac{n}{m})-\eta(\frac{m-n}{m})+\log_2 m$$

所以我們得到

$$I(\alpha_1,\gamma)=\eta(\frac{n}{m})+\eta(\frac{m-n}{m})\quad。\tag{1}$$

(1)式的右邊剛好是一個含有兩個結果的試驗的熵數，因此，$I(\alpha_1,\gamma)$ 的極大是發生在 $n/m=1/2$ 的時候。根據這些，我們知道要使 α_1 發生最大的功用就該使 n/m 儘量的接近 1/2。譬如說，把 α_1 的兩根引線分別接在線路的一個端點與線路的中點（或近似中點）便是合乎上面的要求了。這樣做還有一個優點：可以使得 α_2 的情況與 α_1 類似。繼續這樣子的測下去，只要測〔$\log_2 m$〕次就能找到壞燈；其中〔$\log_2 m$〕是大於或等於 $\log_2 m$ 的最小的整數。當 $m=2^k$ 時，$\log_2 m=k$，和我們起初的結論一樣。

（二）某地有甲、乙兩村，甲村的人只說眞話，乙村的人只說假話。兩村村民，你來我往，十分平常。如果你走到了這個地方，卻不知是到了那一個村子，那麼你該怎麼樣的向村民請問，以便儘早知道到底是到了甲村還是乙村？（我們假定你只向第一個碰到的人請教，並假定這人只說「是」與「否」。）你有沒有辦法同時問出這人是甲村的還是乙村的。

首先，我們看看至少要問幾個問題才能確定是到了那一個村子。我們分別用 A,B 表示「到了甲村」和「到了乙村」這兩個事件。因爲事先毫無風聲，所以 $P(A)=P(B)=1/2$。令 $\beta=\begin{pmatrix}A & B\\ 1/2 & 1/2\end{pmatrix}$，則這個試驗 β 就描述了「所在何村」這個隨機現象。β 的熵數 $H(\beta)=-\frac{1}{2}\log_2\frac{1}{2}-\frac{1}{2}\log_2\frac{1}{2}=1$。

假定我們問了某個問句。我們知道村民回答時只可能說「是」與「否」，而且說「是」與「否」的可能性一樣大。令 C 表示「是」，D 表示「否」，$\alpha=\begin{pmatrix}C & D\\ 1/2 & 1/2\end{pmatrix}$。則試驗 α 就描述著這個問句所引起的隨機現象。同樣的，$H(\alpha)=1$。

依照上面的解釋，我們的目的便是要找一個好的 α，以使得 $I(\alpha,\beta)$ 的值儘量的大。要是能使得 $I(\alpha,\beta)=H(\beta)=1$，那就太好了，那就表示 α 能夠消除 β 的隨機性，α 能夠告訴我們是在那個村子。

我們先看看能不能找到這樣的 α。由於

$$I(\alpha,\beta)=H(\beta)-H(\beta|\alpha)=H(\alpha)-H(\alpha|\beta)=1-H(\alpha|\beta)$$

及

$$H(\alpha|\beta)=H(\alpha|A)P(A)+H(\alpha|B)P(B)$$
$$=\frac{1}{2}(H(\alpha|A)+H(\alpha|B))$$

$I(\alpha,\beta)=H(\beta)$ 的一個充要條件是 $H(\alpha|A)=0$ 及 $H(\alpha|B)=0$ 由於 §3. 之 (i)，$H(\alpha|A)=0$ 只發生在 $P(C\cap A)/P(A)=1$ 或 $P(D\cap A)/P(A)=1$ 的時候。同時的，$H(\alpha|B)=0$ 只發生在 $P(C\cap B)/P(B)=1$ 或 $P(D\cap B)/P(B)=1$ 的時候。因此，我們得到兩個 $I(\alpha,\beta)=H(\beta)$ 的充分條件：

$$(1)\ \begin{cases}P(C\cap A)=P(A)\\ P(D\cap B)=P(B)\end{cases}\qquad (2)\ \begin{cases}P(D\cap A)=P(A)\\ P(C\cap B)=P(B)\end{cases}$$

(1)的意思是說所問的問句必須具有下面兩個性質：

(a)如果是在甲村的話，問句的答案必須是「是」。

(b)如果是在乙村的話，問句的答案必須是「否」。

在這樣的指引下，讀者不難看出，「你住在這個村子嗎？」便是一個好問句。只要這麼問一下，便可以知道是在甲村或乙村了。此外，我們只要問一下「一加一是四嗎？」就可知道所請教的人是甲村的還是乙村的。

從上面的討論我們知道兩個問句是夠了，但是，是不是一定要兩個問句呢？

我們起初的問題是要用一些問句來判斷下面四個情況中那一個是對的：「在甲村，問甲村人」，「在甲村，問乙村人」，「在乙村，問甲村人」，「在乙村，問乙村人」。在問句提出之前，四個情況都是可能的。我們把這個試驗用 γ 來表示。因此 $H(\gamma)=\log_2 4=2$。假設 k 個問句 $\alpha_1,\alpha_2,\cdots\cdots,\alpha_k$，可以找出答案，則

$$2=I(\alpha_1\vee\alpha_2\vee\cdots\vee\alpha_k,\gamma)\leq H(\alpha_1\vee\alpha_2\vee\cdots\cdots\vee\alpha_k)\leq k。$$

所以至少需要兩個問句。這樣就完全解決了這個問題。

（三）現有九枚外表一模一樣的銅幣。其中一枚是假的。如果我們想要用天平把它找出來，並且確定它是較重或較輕，請問至少要量幾次才辦得到？

所謂用天平來找就是把相同數目的銅幣分別放在天平左右的秤盤上，觀察天平的狀態，是平衡，右傾還是左傾，然後再下判斷。這是一個試驗，它的熵數不會比 $\log_2 3$ 大。

我們把找出假幣，並確定它是較重或較輕的試驗用 γ 來代表。因爲每一枚銅幣都可能是假的，也都可能較重或較輕，而這些可能性又都一樣大，所以 $H(\gamma)=\log_2 18=\log_2 2\cdot 9=1+2\log_2 3$。

假設用天平秤 k 次便可找出假幣。我們把這 k 次的試驗分別用 $\alpha_1,\alpha_2,\cdots\cdots,\alpha_k$ 來表示。則

$$I(\alpha_1\vee\cdots\vee\alpha_k,\gamma)=H(\gamma)=1+2\log_2 3$$

但是

$$\begin{aligned}I(\alpha_1\vee\cdots\vee\alpha_k,\gamma)&\leq H(\alpha_1\vee\cdots\vee\alpha_k)\\&\leq H(\alpha_1)+\cdots+H(\alpha_k)\leq k\log_2 3,\end{aligned}$$

所以

$$k\geq 2+\frac{1}{\log_2 3}。$$

也就是說，至少要秤 3 次才行。我們來看看 3 次是不是眞的夠了。

首先，我們希望 $I(\alpha_1,\gamma)$ 能夠儘量的大。我們把 α_1 寫成

$$\alpha_1=\begin{matrix}B & L & R\\ \left\langle\dfrac{9-2i}{9}\right. & \dfrac{i}{9} & \left.\dfrac{i}{9}\right\rangle\end{matrix},$$

其中 B,L,R 分別表示平衡，左傾與右傾事件，i 表示秤盤上各有 i 個銅幣。由於 γ 可以完全確定 α_1，所以 $H(\alpha_1|\gamma)=0$，$I(\alpha_1,\gamma)=H(\alpha_1)-H(\alpha_1|\gamma)=H(\alpha_1)$。而 $H(\alpha_1)$ 的極大是發生在 $(9-2i/9)=(i/9)$ 的時候，也就是 $i=3$ 時。因此，第一次的秤法是在秤盤上各放 3 個銅幣，然後分別考慮：

(a)B 發生。這時候假幣不在秤盤上，讀者不難看出再量兩次就能找出假幣。

(b)R 發生。這時候假幣在秤盤上。請注意，我們不能丟掉天平右傾這個資料。因爲如果丟掉了它，我們從 α_1 得到的就只剩下假幣是在天平上這個事實。這件事發生的機率是 2/3，因此它所提供的訊息是 $\log_2(3/2)=\log_2 3-1$，比 α_1 所提供的少了一拍。

現在我們要在 R 爲已知的條件下，儘量的把 r 的隨機性除掉（這時 r 是六枚銅幣中的試驗）。我們以 α_2 表示第二次的量法。跟以前一樣，我們希望 $H(\alpha_2|R)$ 儘量的大；也就是說，在 R 的條件下要求 α_2 能夠提供最多的訊息。因此，在 R 的條件下，α_2 爲平衡，右傾及左傾的機率要各爲 1/3。

爲了清楚起見，我們把左邊的三枚分別叫做①，②和③，右邊的三枚分別叫做④，⑤和⑥。要使得在

R的條件下，α_2 爲平衡，右傾及左傾的機率一樣，我們可以如下安排：拿掉①和④；把②和⑤相互交換；保持③，⑥不動。在這樣的量法之下，如果天平是平衡的，則假幣便在①與④之中；如果天平仍舊右傾，則在③與⑥之中；如果天平變爲左傾，則在②與⑤之中；而且，如果假幣在①，②，③之中，則假幣比眞幣輕，否則比眞幣重。根據這些，再量一次便可完全解答原來的問題。

(c)L發生。論法與(b)相同。

因此，量三次便可完成鑑定工作。

本節所討論的例子雖然都是趣味性的問題；但是已經表現出了訊息論的一般用法。較嚴肅的應用問題（如訊息傳遞的工程問題），常常可以仿照 §3. 來處理，只是可能比較複雜，須要更精細的分析。以後有機會時，我們再介紹一些訊息論在訊息傳遞問題上的簡單應用。