# 卜松估計的漸近分析

中央研究院統計所　黃顯貴

*Among discontinuous distributions,*
*the Poisson series is of first importance.*
—Sir Ronald Aylmer Fisher
(1890–1962)

## 導論

卜松分布自卜松 (Simeon-Denise Poisson, 1781-1840) 在其「刑法及民法判決的機率分析」一書[3]首先導出後, 便以“稀有事件定律”(law of rare events) 的角色在諸多應用問題中出現。它的高度出現頻率當然與其簡單的定義有關：若 $X$ 是一卜松分布，參數為 $\lambda$，則

$$P(X=k)=\frac{l^k}{k!}e^{-l} \quad (k=0,1,\cdots)。$$

所以從聚會中生日相同的人數到每年百歲人瑞死亡人數；從印刷文件中每頁打字錯誤次數到一朵花花瓣數；從某一時間電話通話數到公路上車禍及死亡數；卜松分布都在合理的模型假定下以最自然的方式出現。

從這個稀有事件定律的角度出發，俗語所謂

無巧不成書

的「巧」字，對機率學家來說，很自然會將它與卜松分布作聯想。同理

福無雙至
禍不單行

背後也隱藏了可能的卜松分布。這些例子約略點出了為什麼卜松分布在科學研究上歷久長青。近年來分子生物學研究的蓬勃發展，又將相關卜松分布問題帶入另一境地。

我們在網路上作「Poisson distribution」這個關鍵字的搜尋，結果如下(5/29/2000)：

| | Poisson distribution | Poisson distributions | Poisson distributions* |
|---|---|---|---|
| Raging | 1314 | 6659 | 7700 |
| Google | 1178 | 6100 | ---- |

(其它 search engines 如 Yahoo 等較不適於專業搜尋；另 Google 尚未提供 OR-search 的功能。) 這其中大多數網頁屬教學性質，學術意味較低。我們再利用 AMS 的 MathSciNet 尋找 Math Reviews 所收錄自 1940 至 2000 的論文中，title 含 Poisson distribution 的文章，共找到 401 筆。若將 title 改成 anywhere，則共找到 1517 筆(搜尋日為 5/29/2000)。這些搜尋結果顯而易見地說明了卜松分布的“無所不在性”。

## 卜松估計

卜松在其書[3]上證明如下的結果：
設 $Y$ 是二項分布，參數是 $n,p$，其中 $n$ 是正整數，$0< p <1$：

$$P(Y=k)=\binom{n}{k}p^k(1-p)^{n-k} \quad (0\le k\le n).$$

若 $np\to l<\infty$ (當 $n\to\infty$ 時)，則

$$P(Y=k)\to\frac{l^k}{k!}e^{-l} \quad (k=0,1,\cdots).$$

用簡單的話來說，給定一個銅板其出現人頭的機率是 $p$，則丟 $n$ 次後人頭出現總數的極限分布是卜松($\lambda$)分布，前提是 $np\to l<\infty$；亦即 $p$ 相對於漸增的 $n$ 必須很小–從而稀有事件定律得其命名。

這個結果很自然有多方向的延伸。最常見的是去掉同分布的限制。在適當條件下，卜松分布仍以稀有事件定律的角色出現。而所謂”卜松估計”問題，便是進一步去了解逼近程度的好壞(準確性)。與本文有關，較值得一提的是 Yu. V. Prohorov (1929--) 1953 年的結果[4]。他考慮了如下的全變差距離函數：

$$d_{TV}(L(X),L(Y))=\frac{1}{2}\sum_{k\ge 0}\left|P(X=k)-P(Y=k)\right|,$$

其中 $L(X), L(Y)$ 分別表為 $X$ 及 $Y$ 的分布。他證明

$$d_{TV}(L(X), L(Y)) = \frac{p}{\sqrt{2pe}}(1+O(E))。 \qquad (1)$$

右式誤差項 $E$ 滿足

$$E = \min\left\{1, p+(np)^{-\frac{1}{2}}\right\}。$$

這是一個相當強的結果。它把卜松原始的"極限定理"作了兩個方向的延伸：

(a) 精確度：即用卜松分布來估算二項分布時誤差的大小。

(b) 稀有度：(1)式成立的條件有二：

(i) $np = O(1)$；

(ii) $p = o(1)$ 及 $np \to \infty$；

其中第(ii)個條件大大地推廣了卜松估計的適用範圍；亦即卜松分布不僅可描述稀有事件，它也可描述"中央極限定理"。這中間有無矛盾之處?答案當然是沒有。這個概念很重要，底下再進一步闡述。

我們知道，卜松分布當其參數 $l \to \infty$ 時，在適當正態化後趨近標準常態分布(即所謂中央極限定理)。所以如果適當地選取卜松分布的參數，則它既可描述稀有事件，亦可描述中央極限定理。亦即卜松分布有其漸近估計上的"均勻性"。這個均勻性的一般現象是：常態分布是一種連續分布，用它來逼近離散分布時，免不了會有跳躍點上引出的誤差。這些誤差在一般情形下都是以 1/(標準差) 的速率趨近零；而如果適當選取一較好算的離散分布時，用連續來逼近離散分布所產生的效應，不復存在；意即相對的誤差可能較易控制。

另一方面，均勻性在實用上非常重要。因為實用上所謂的無窮大是不存在的。例如，假若 $n$=10000，$p$=$10^{-3}$，則 $np$=10，這個數到底是 $n^{1/4}$ 還是 $O(1)$呢? 所以類似 (1) 式較均勻的結果，除開理論上的興趣外，也有其實用價值。

值得一提的是當 $np = O(1)$ 時，(1)式右邊變成一上界估計。

很可惜的是像 Prohorov 這種結果在卜松估計的文獻上卻是屈指可數(參閱[1])。其中一個可能的原因是其後新方法的發展使此類結果較乏人問津。這些新方法主要可分成三大類(the big three，根據 J. M. Steele 的說法[5])：

Chen-Stein method

semigroup method

coupling method

這些方法各有所長。這三類最大宗者非 Chen-Stein method 莫屬，它的條件較鬆且適於估計上界；semigroup method 基本上是複變方法(用特徵函數)的實變版本。 這些方法各有所長，也多有限制。比如說 (1) 式便不易由此三種方法得出 (主要項較容易；餘項較難)。

這些方法的優劣比較及卜松逼近近十年來的廣為重視，使我考慮採用一較直接的分析方法，來計算在一般的卜松估計問題上，卜松逼近所對應的距離函數之漸近行為。我著眼於漸近展式而不是一般的上界估計。

我在[2]中考慮了一個系列的離散隨機變數 $\{X_n\}$，其機率生成函數(pgf)漸近上很接近卜松分布的 pgf，在相當一般的條件下，我導出了底下常見距離函數的漸近展式：

$$d_{TV}(L(X_n), L(Y)) = \frac{1}{2}\sum_{j\ge 0}\left|P(X_n = j) - P(Y = j)\right|^a$$

(total variation distance)

$$d_{FM}(L(X_n), L(Y)) = \sum_{j\ge 0}\left|P(X_n \le j) - P(Y \le j)\right|^a$$

(Fortet-Mourier distance)

$$d_K(L(X_n), L(Y)) = \max_{j\ge 0}\left|P(X_n \le j) - P(Y \le j)\right|^a$$

(Kolmogorov distance)

$$d_L(L(X_n), L(Y)) = \max_{j\ge 0}\left|P(X_n = j) - P(Y = j)\right|^a$$

(point metric)

$$d_M(L(X_n), L(Y)) = \left(\frac{1}{2}\sum_{j\ge 0}\left(P(X_n = j)^a - P(Y = j)^a\right)^{1/a}\right)^a$$

(Matusita 或 Hellinger distance)

同時我也有系統地描述出一大類組合結構，使得如上的結果可以幾乎「自動」成立；另外算術數論上的卜松估計也加以探討。文中並考慮了多方向的延伸及未來發展方向。

舉個例子來說明結果的形式。令 $X_n$ 表 $\{1,\cdots,n\}$ 中任意挑選出的整數作質因數分解後

質數個數和 (不重覆計算)，其中每個整數被挑出的機率均為 $1/n$。令 $Y$ 是一卜松分布，參數是 $l = \log\log n + c$ 其中

$$c = g - 1 + \sum_{p:\,prime}\left(\log\left(1-\frac{1}{p}\right)+\frac{1}{p}\right) \approx 0.261497$$

，$\gamma$ 為 Euler 常數，

且

$$P(Y=k) = \frac{l^{k-1}}{(k-1)!}e^{-l} \quad (k = 1,2,\cdots),$$

則

$$d_{TV}(L(X_n), L(Y)) = \frac{k^{a}\,c(a)}{2^{\frac{3a+2}{2}}\,p^{\frac{a}{2}}}\,l^{-\frac{3a-1}{2}} \times\left(1+O(l^{-\frac{1}{2}})\right), \qquad (2)$$

其中

$$k = \left|1-\frac{p^2}{6}-\sum_{p:\,prime}p^{-2}\right| \approx 1.097181\cdots,$$

$$c(a) = 2\int_0^{\infty}\left|x^2-1\right|^{a} e^{-\frac{a}{2}x^2}\,dx\,。$$

其他距離函數亦有類似型式的結果。這個結果，雖然是我主要定理的一推論，在機率數論上亦是全新的式子。這個例子的參數 $l \to \infty$，再來看一個參數是有限的例子。若 $X_n$ 是二項分布，參數是 $n,p$，若 $np = l = O(1)$，則

$$d_{TV}^{(a)}(L(X_n), L(Y)) = 2^{1-a} \sum_{j\ge 0}\left(e^{-l}\frac{l^{j}}{j!}\right)^{a}\left|l^2 - 2jl + j(j-1)\right|^{a}\left(1+O(n^{-1})\right),$$

其中 $Y$ 是卜松分布參數是 $l = np$。

## 結　語

我們從底下的統計數字可約略看出卜松估計在目前文獻上的發展(或受重視)程度。「卜松估計」在 Raging(為 Altavista 之搜尋引擎)上共找到 798 個網頁(搜尋日為 5/29/2000)。而在 MathSciNet 的搜尋結果：

title: 148 篇(其中 110 篇發表於 1990 年後) ；

anywhere: 279 篇(其中 193 發表於 1990 年後) 。

雖然單從數量上無法得出太多結論，我們可以說卜松估計仍是一相當熱門的題目。

從方法學上來看，不同方法自有其優劣性。我提出的分析方法在概念上、在程序上較直接透明，比其它方法，容易了解。但在漸近估計上須較多分析工具如複變、鞍點法 ..等，這些方向進一步刺激出新的問題，比如說(2)式的餘項是否最佳？目前正從事這些及其他應用問題的進一步研究。

## 參考文獻


[1] A. D. Barbour, L. Holst and S. Janson, *Poisson approximation*, Oxford Science Publication, Clarendon Press, Oxford (1992).

[2] H.-K. Hwang, *Advances in Applied Probability*, **31**, 448 (1999).

[3] S.-D. Poisson, Recherche sur la probabilité des jugements en matière crimelle et en matière civille, Bachelier, Paris (1837).

[4] Y. V. Prohorov, *Uspekhi Matema-ticheskikh Nauk*, **8**, 135 (1953). Also in *Selected Translations in Mathematical Statistics and Probability*, volume **1**, 87.

[5] J. M. Steele, *American Mathe-matical Monthly*, **101**, 48 (1994).


*Life is good for only two things,*
*discovering mathematics and*
*teaching mathematics.*
—S.-D. Poisson