1 簡介

一公司有n台產品製造機器，此n台機器並非同一批買進，性能上有差異，在運作8小時後，我們分別紀錄此n台機器的產品不良個數，$Y_1,\ldots,Y_n$，如何用此記錄來估計n台機器生產的平均不良數$\lambda_1,\ldots,\lambda_n$呢？機器在製造時，本來就有差異，加上新機器和舊機器的運作效能不盡相同，我們考慮此n台機器生產的平均不良數服從某一分配G。如此一來，便可利用卜瓦松(Poisson)模型來詮釋我們的問題：

$$\begin{aligned} Y_i \mid \lambda_i &\sim Poisson(\lambda_i)\,, \quad i=1,\ldots,n\,, \\ \lambda_i &\sim G\,, \end{aligned} \tag{1.1}$$

其中$Y_1,\ldots,Y_n$是觀察值，$Y_i \mid \lambda_i$服從卜瓦松分配$f(Y_i \mid \lambda_i)=e^{-\lambda_i}\lambda_i^{Y_i}/Y_i!$，且給定$\lambda_i$下的$Y_i$彼此互相獨立，$\lambda_i$彼此亦獨立地服從一未知分配$G$。

在使用貝氏觀念之前，我們可以選擇用Y_i來估計λ_i，此估計量有許多重要的特性，比方說最大概似估計量、最小均方誤、不偏等等。然而當$\lambda_1,\ldots,\lambda_n$中的相異值超過三個，Stein (1955)證明在損失函數(loss function) $L(\hat{\lambda}_i,\lambda_i)=(\hat{\lambda}_i-\lambda_i)^2$之下，此估計量是不好的。

在貝氏損失函數$L(\hat{\lambda}_i,\lambda_i)=(\hat{\lambda}_i-\lambda_i)^2$之下，貝氏估計量爲後驗分配的期望值

$$\hat{\lambda}_i = E(\lambda_i \mid Y_i) = \frac{\int(\lambda_i^{Y_i+1}e^{-\lambda_i}/Y_i!)dG(\lambda_i)}{\int(\lambda_i^{Y_i}e^{-\lambda_i}/Y_i!)dG(\lambda_i)} = (Y_i+1)\frac{m(Y_i+1)}{m(Y_i)}\,, \tag{1.2}$$

其中，令$m(x)=\int(\lambda_i^x e^{-\lambda_i}/x!)dG(\lambda_i)$。

假設G爲已知的分配，便可藉由(1.2)式，估計卜瓦松均數λ_i。若不知G的分配，在估計λ_i前，我們先用$Y_1,\ldots,Y_n$對G做估計，亦即所謂的經驗貝氏（empirical Bayes，以下簡稱EB）。經驗貝氏又可分成母數經驗貝氏（parametric empirical Bayes，以下簡稱PEB）及無母數經驗貝氏（nonparametric empirical Bayes，以下簡稱NPEB），前者可假定G有卜瓦松的共軛分配型式，伽碼分配$\Gamma(a,b)$，但需用觀察值估計未知參數a、b；後者爲在不限定型式下估計G的方法，Robbins (1955)提出的估計法以及

Laird (1978) 提出的無母數最大概似（nonparametric maximum likelihood，以下簡稱NPML）便屬於NPEB 法則。

除了上述估計方法外，本文根據 Escobar (1994) 估計常態均數的方法，引進Dirichlet 過程當 G 的先驗分配，提供一個新的 NPEB 估計法 (DP)。由模擬研究得知，當均數 $\lambda_1, \ldots, \lambda_n$ 的分配 G 符合伽碼分配的型式，利用 PEB 估計的 $\hat{G}$ 當先驗分配估計 λ_i 有很好的效果。然而當真正 G 的型式爲少數幾個值的離散分配，NPML 的 λ_i 估計效果顯著優於 PEB。而採用 DP 估計 λ_i 的好處是，不論 G 爲何種型態，其效果總是介於 NPMLE 及 PEB 之間，並趨向其中較好的估計法。

本文第二節，介紹使用 Dirichlet 過程當成 G 先驗分配的觀念，討論在固定 Dirichlet 過程參數 (α_0, G_0) 之下，結合觀察值 $Y_1, \ldots, Y_n$，估計卜瓦松均數的理論和演算法；第三節進一步討論 Dirichlet 過程參數 (α_0, G_0) 所代表的意義以及對估計的影響，並給此二參數先驗分配，改進第二節的演算法；第四節介紹三個以往估計卜瓦松均數的方法：(1) Robbins，(2) NPML，(3) PEB；第五節將 DP 估計法和第四節提到的三種方法作模擬比較並對 Archibald (1948) 提供的海石花 (Armeria maritima) 樣區資料作分配估計；第六節做結論。

以下將本文所使用的特殊符號作一說明，$\lambda \sim G$ 代表 λ 服從 G 的分配，即 G 爲 λ 的分配函數 (distribution function)，$dG(\lambda) = g(\lambda)$ 表示 G 的密度函數 (density function)；粗體 $\boldsymbol{\lambda}$ 及 $\boldsymbol{Y}$ 分別表示向量 $(\lambda_1, \ldots, \lambda_n)$ 和 $(Y_1, \ldots, Y_n)$，$\boldsymbol{\lambda}^{-i}$ 則表示將 $\boldsymbol{\lambda}$ 的第 i 個元素移除。

2 Dirichlet 過程

Dirichlet 過程 (Dirichlet Process) 是一個在（機率）分配函數空間上的分配，若我們將分配函數空間上的任一種分配視爲一個「值」，針對所有「值」給定的離散分配，便稱爲「Dirichlet 過程」，因此分配由 Ferguson (1973) 首先提出，又名 Ferguson