牛奶怎麼生病的？—適合度檢定

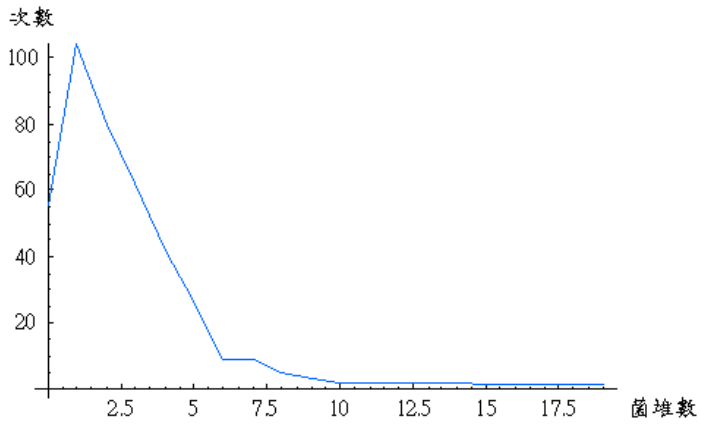
王澄祥（ 應數博 94）

為了檢測牛奶有沒有被細菌污染，必列斯與費雪（Bliss and Fisher, 1953，參考文獻一 ）做了一種實驗。實驗的作法是：將 0.01 毫升的牛奶塗在 1 平方公分的玻片上，將玻片劃分為 $n=400$ 個細格子，然後放到顯微鏡下觀察每個細方格內的菌堆數，製作菌落數的次數分配表如下：

表一：單位細格內菌落數的次數分配表

每個細格內的菌堆數（堆）	0	1	2	3	4	5	6	7	8	9	10	19
次數（次）	56	104	80	62	42	27	9	9	5	3	2	1

畫其折線圖如下：



圖一：菌堆數的折線圖

若假設細菌很均勻的分布在牛奶中，那麼每個細格出現的菌落數假設具有卜瓦松分佈（Poisson distribution）似乎是蠻直覺的想法。但是經過更深層的考慮發現，要注意兩個問題：首先，由於表面張力的緣故，在玻片上牛奶液滴的邊緣會依附較多細菌；再者，牛奶液滴薄膜並非厚度均勻，而是中間較厚邊緣較薄，這也造成細菌在玻片上並非均勻分佈。因此儘管數據上看來，菌堆數的確很像具有卜瓦松分佈，但我們還是要做進一步的確認，看菌堆數是否真的具有卜瓦松分佈？

若令單位細格內的菌堆數為 X，並假設 X 具有卜瓦松分佈，且母數（parameter）為 λ，則 X 的機率密度函數（probability density function）為

$$P(X=x)=f_X(x)=\frac{e^{-\lambda}\lambda^x}{x!},\ x=0,\ 1,\ 2,\ \cdots.$$

則可得到 λ 的最大概似估計量（Maximum Likelihood Estimator，參考文獻二 ）為

$$\hat{\lambda}=\frac{0\times56+1\times104+2\times80+\cdots+10\times2+19\times1}{400}=2.44.$$

接著將表一的末五欄合併，計算每個欄位的期望次數，以及卡方檢定統計量整理如下表：

表二：各組別的觀測次數、期望次數與卡方統計量值

菌堆數 x（堆）	0	1	2	3	4	5	6	≥7
觀測次數 O_x（次）	56	104	80	62	42	27	9	20
期望次數 E_x（次）	34.9	85.1	103.8	84.4	51.5	25.1	10.2	5.0
統計量值 $\chi_x^2=\frac{(E_x-O_x)^2}{E_x}$	12.8	4.2	5.5	6.0	1.8	0.14	0.14	45.0

其中期望次數的算法是

$$E_x = n\times P(X=x) = 400\times\frac{e^{-\hat{\lambda}}\hat{\lambda}^x}{x!} = 400\times\frac{e^{-2.44}2.44^x}{x!},\ x=0,1,2,\cdots,6$$

且

$$E_7 = n\times P(X\geq 7) = 400\times\sum_{x=7}^{\infty}\frac{e^{-2.44}2.44^x}{x!} = 400\times\left(1-\sum_{x=0}^{6}\frac{e^{-2.44}2.44^x}{x!}\right)\doteq 5.$$

由此便得卡方檢定統計量值為

$$\chi^2=\sum_{x=0}^{7}\chi_x^2\doteq 75.2.$$

而根據統計推論（參考文獻二）可知此統計量具有卡方分佈且自由度為 $d=8-1-1=6$（有 8 個類別且已先估計 $\hat{\lambda}$）。再查表知自由度 6 的卡方分佈 0.005 百分位數（Quantile）為 $\chi_6^2(0.005)\doteq 18.55$，遠低於前面算出的統計量值 75.2，所以此統計量的機率值（P-value）低於 0.005，因此無法接受「菌堆數具有卜瓦松分佈」的假設。是什麼原因造成配適這樣的分佈不當呢？觀察表二的統計量值便可看出，第一組與最後一組的統計量值偏高，這兩組的期望次數與觀測次數差距過大，造成計算出的統計量值過高，於是這樣的配適並不恰當。

筆者想順帶一提的是，統計界有句話說：「數字（或圖表）會說話。」但有時候只看表面數字或圖表就做決定，容易使人做出錯誤的判斷。最好有理論的背景作佐證，或進一步探討，出錯的可能性較低。這跟我們平時待人接物是一樣的，先了解，再處理，那麼誤會或誤判就會減少許多，與讀者共勉之。

參考文獻

一、Bliss, C.I. and Fisher, R.A., “Fitting the Negative Binomial Distribution to Biological Data.” Biometrics 9, 176-199.

二、Casella, G. and Berger R.L., “Statistical Inference, Second Edition.” Pacific Grove, CA: Duxbury, Press, 2002.