# 二项分布

## 历史

1713年，瑞士数学家雅各布·伯努利（Jakob Bernoulli）在他的著作《推测术》（Ars Conjectandi）中最早研究了伯努利试验，这是二项分布产生的源头。

1895年，二项分布作为词语最早出现在英国数学家卡尔·皮尔逊（Karl Pearson）的《对进化数学理论的贡献——II.均质材料中的倾斜变化》（Contributions to the Mathematical Theory of Evolution---II. Skew Variation in Homogeneous Material）中。原文：“这个结果似乎相当重要，我认为它还没有被注意到。它给出了任何二项式分布的均方误差。”（"This result seems of considerable importance, and I do not believe it has yet been noticed. It gives the mean square error for any binomial distribution.”）

## 二项分布的应用

EM（Expectation-Maximum）算法也称期望最大化算法，是一种迭代优化策略，由于它的计算方法中每一次迭代都分两步，其中一个为期望步（E步），另一个为极大步（M步）。EM算法受到缺失思想影响，最初是为了解决数据缺失情况下的参数估计问题，其基本思想是：首先根据己经给出的观测数据，估计出模型参数的值；然后再依据上一步估计出的参数值估计缺失数据的值，再根据估计出的缺失数据加上之前己经观测到的数据重新再对参数值进行估计，然后反复迭代，直至最后收敛，迭代结束。

如果我们只有一系列的实验结果而不知道实验的概率呢?

最大似然估计(Maximum Likelihood Estimation)~MLE

似然函数：在参数为θ时，出现这组实验结果的概率

最大似然函数：使似然函数的值最大，得到此时估计值θ，即 $\hat{\theta} = argmaxL(\theta)$

可对似然函数取对数，再求导，令导数为0，即可得到最大似然函数的θ，即极大似然估计（MLE）

### §2.3.2 EM 算法

MLE 是一种非常有效的参数估计方法, 但当分布中有讨厌参数或数据为截尾或缺失时, 其 MLE 的求取是比较困难的. 于是 Dempster 等人于 1977 年提出了 EM 算法, 其含义是把求 MLE 的过程分两步走: 第一步求期望, 以便把讨厌的部分去掉; 第二步求极大值. 本小节将简单介绍这种非常有用的方法.

#### 一、 EM 算法

下面我们以一个例子说明 EM 算法是如何进行的.

**例 2.3.7** 设一次试验可能有四个结果, 其发生的概率分别为 $\frac{1}{2}+\frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4}$, 其中 $\theta \in (0,1)$, 现进行了 197 次试验, 四种结果的发生次数分别为 125, 18, 20, 34. 试求 $\theta$ 的 MLE.

**解** 由于此时的总体分布为多项分布, 故其似然函数

$$f(\theta,\boldsymbol{y})\propto\left(\frac{1}{2}+\frac{\theta}{4}\right)^{y_1}\left(\frac{1-\theta}{4}\right)^{y_2}\left(\frac{1-\theta}{4}\right)^{y_3}\left(\frac{\theta}{4}\right)^{y_4}\propto(2+\theta)^{y_1}(1-\theta)^{y_2+y_3}\theta^{y_4},\tag{2.3.8}$$

由此可知, 其对数似然方程是关于 $\theta$ 的二次三项式.

但是我们可以通过引入一个变量 $Z$ 后, 使得求解变得比较容易. 现假设第一种结果可以分成两部分, 其发生概率分别为 $1/2$ 和 $\theta/4$, 令 $Z$ 和 $Y_1-Z$ 分别表示落入这两部分的次数. 显然, $Z$ 是我们人为引入的, 它是不可观测的 (在文献中称之为 latent 变量). 也称数据 $(\boldsymbol{Y},Z)$ 为完全数据 (complete data), 而观测到的数据 $\boldsymbol{Y}$ 称为不完全数据. 此时完全数据的似然函数为

$$f(\theta|\boldsymbol{y},z)\propto\left(\frac{1}{2}\right)^{z}\left(\frac{\theta}{4}\right)^{y_1-z}\left(\frac{1-\theta}{4}\right)^{y_2}\left(\frac{1-\theta}{4}\right)^{y_3}\left(\frac{\theta}{4}\right)^{y_4}\propto\theta^{y_1-z+y_4}(1-\theta)^{y_2+y_3},$$

其对数似然函数为

$$l(\theta|z,\boldsymbol{y})\propto(y_1-z+y_4)\ln\theta+(y_2+y_3)\ln(1-\theta).\tag{2.3.9}$$

如果 $\boldsymbol{y},z$ 均已知, 则由上式很容易求得 $\theta$ 的 MLE, 但遗憾的是, 我们仅知道 $\boldsymbol{y}$, 而不知道 $z$ 的值. 但是我们注意到, 当 $y_1$ 及 $\theta$ 已知时, $Z\sim B(y_1,2/(2+\theta))$(其原因见后面的注 2.3.4). 于是, Dempster 等人建议分两步进行:

**E 步**: 在给定 $\boldsymbol{y}$ 及 $\theta=\theta^{(i)}$ 的条件下, 求完全数据的对数似然函数关于潜在变量 $Z$ 的期望:

$$Q(\theta|\boldsymbol{y},\theta^{(i)})=E_Z l(\theta|Z,\boldsymbol{y});\tag{2.3.10}$$

**M 步**: 求 $Q(\theta|\boldsymbol{y},\theta^{(i)})$ 关于 $\theta$ 的最大值 $\theta^{(i+1)}$, 即

$$\theta^{(i+1)}=\arg\max_{\theta}Q(\theta|\boldsymbol{y},\theta^{(i)}).\tag{2.3.11}$$

重复 (2.3.10) 和 (2.3.11) 式, 直至收敛即可得到 $\theta$ 的 MLE.

对于本问题, 其 E 步为

$$\begin{aligned}Q(\theta|\boldsymbol{y},\theta^{(i)})&\propto[y_1-E_Z(Z|\boldsymbol{y},\theta=\theta^{(i)})+y_4]\ln\theta+(y_2+y_3)\ln(1-\theta)\\&=[y_1-2y_1/(2+\theta^{(i)})+y_4]\ln\theta+(y_2+y_3)\ln(1-\theta).\end{aligned}$$

执行代码如下

```cpp
#include<iostream> //输入输出流

#define Y1 125.0
#define Y2 18.0
#define Y3 20.0
#define Y4 34.0

using namespace std; //命名空间

int main()
{
    double init = -1.0;
    double next = 0.5;
    int count = 0;

    while(init != next) {
        init = next;
        next = (Y1 - ((2*Y1) / (2+init)) + Y4) / ( (Y1 - ((2*Y1) / (2+init)) + Y4) + (Y2 + Y3) );
        count++;
    }
```

```
21
22      cout << next << endl;
23      cout << count << endl;
24
25      return 0;
26  }
```

最终结果约为：0.626821