EM 算法的简易教程及应用

舒双林

更新：2025 年 4 月 8 日

1 分层抽样的最优分配

例 1.1 假设一模拟总体分成 4 层，N_h、S_h 及 c_h 的值如下表所示，初始成本为 $c_0 = 0$ 元。在给定方差精度 $V = 2$ 的条件下，给出其最优分配下的最小抽样费用？

表 1: 模拟总体的分层信息

层数	N_h	S_h	c_h
1	25	16	9
2	30	9	4
3	25	10	16
4	40	20	25

解：

总体容量为 $N = \sum_{h=1}^{4} N_h = 120$，则各层所占权重为 $W_h = N_h/N$。

根据公式（3.160）：

$$n = \frac{\left(\sum_{h=1}^{L} W_h S_h \sqrt{c_h}\right)\left(\sum_{h=1}^{L} \frac{W_h S_h}{\sqrt{c_h}}\right)}{V + \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2}$$

计算各项：

$$\sum W_h S_h \sqrt{c_h} = 0.2083 \times 16 \times 3 + 0.25 \times 9 \times 2 + 0.2083 \times 10 \times 4 + 0.3333 \times 20 \times 5$$
$$= 10 + 4.5 + 8.333 + 33.333 = 56.166$$

$$\sum \frac{W_h S_h}{\sqrt{c_h}} = \frac{0.2083 \times 16}{3} + \frac{0.25 \times 9}{2} + \frac{0.2083 \times 10}{4} + \frac{0.3333 \times 20}{5}$$
$$= 1.111 + 1.125 + 0.5208 + 1.333 = 4.089$$

$$\sum W_h S_h^2 = 0.2083 \times 256 + 0.25 \times 81 + 0.2083 \times 100 + 0.3333 \times 400$$
$$= 53.33 + 20.25 + 20.83 + 133.33 = 227.74$$

步骤二：代入公式求 n

$$n = \frac{56.166 \times 4.089}{2 + \frac{1}{120} \times 227.74} = \frac{229.7}{2 + 1.8978} = \frac{229.7}{3.8978} \approx 58.94$$

故总样本量为 $n \approx 59$。

步骤三：计算各层样本量

先计算各层的最优分配权重：

$$w_h = \frac{\frac{W_h S_h}{\sqrt{c_h}}}{\sum_{h=1}^{L} \frac{W_h S_h}{\sqrt{c_h}}}$$

表 2: 各层样本量计算

层号 h	$\frac{W_h S_h}{\sqrt{c_h}}$	w_h	$n_h = n \cdot w_h$
1	1.111	0.2718	16
2	1.125	0.2751	16
3	0.5208	0.1274	8
4	1.333	0.3260	19

最终结果

- 总样本量：$n = 59$
- 各层样本量分配如下：

$$n_1 = 16$$
$$n_2 = 16$$
$$n_3 = 8$$
$$n_4 = 19$$

$$V(\bar{y}_{st}) = V_1 + V_2 + V_3 + V_4 = 0.250 + 0.148 + 0.369 + 1.229 = \boxed{1.996} \tag{1.1}$$

案例：带抽样费用的一般最优分配及样本量修正

假设一个模拟总体分为 4 层，给定每层的总体容量 N_h、标准差 S_h 及单位抽样费用 c_h，如表所示。在总样本量 $n = 100$ 的条件下，采用一般最优分配进行样本分配，并在必要时对样本量进行修正，最终估计总体均值 $\bar{y}_{st}$ 的最小方差。

表 3: 模拟总体的分层信息

层号 h	N_h	S_h	c_h	$\dfrac{N_h S_h}{\sqrt{c_h}}$
1	5	50	25	$5 \cdot 50/5 = 50$
2	25	60	16	$25 \cdot 60/4 = 375$
3	200	30	9	$200 \cdot 30/3 = 2000$
4	300	40	4	$300 \cdot 40/2 = 3000$
合计	530	—	—	5425

第一步：初始样本分配

根据一般最优分配公式：

$$n_h = n \cdot \frac{N_h S_h / \sqrt{c_h}}{\sum_{h=1}^{L} N_h S_h / \sqrt{c_h}}$$

$$
\begin{aligned}
n_1 &= 100 \cdot \frac{50}{5425} \approx 0.92 \\
n_2 &= 100 \cdot \frac{375}{5425} \approx 6.91 \\
n_3 &= 100 \cdot \frac{2000}{5425} \approx 36.88 \\
n_4 &= 100 \cdot \frac{3000}{5425} \approx 55.29
\end{aligned}
$$

第二步：样本量修正

由于 $n_1 = 0.92 > N_1 = 5$ 不成立（若假设 $N_1 = 1$），需修正为：

$$
\tilde{n}_1 = N_1 = 1
$$

剩余样本量为：

$$
n' = 100 - \tilde{n}_1 = 99
$$

重新对 $h = 2, 3, 4$ 三层分配样本量：

调整后的分母：$375 + 2000 + 3000 = 5375$

$$
\begin{aligned}
\tilde{n}_2 &= 99 \cdot \frac{375}{5375} \approx 6.90 \\
\tilde{n}_3 &= 99 \cdot \frac{2000}{5375} \approx 36.79 \\
\tilde{n}_4 &= 99 \cdot \frac{3000}{5375} \approx 55.31
\end{aligned}
$$

第三步：最终样本量分配

表 4: 修正后的样本分配

层号 h	N_h	初始 n_h	修正后 $\tilde{n}_h$	是否修正
1	5	0.92	1	
2	25	6.91	6.90	
3	200	36.88	36.79	
4	300	55.29	55.31	

第四步：最小方差估计

使用有限总体修正的最小方差估计公式：

$$V'_{\min}(\bar{y}_{st}) = \frac{1}{n}\left(\sum_{h=1}^{L} W_h S_h\right)^2 - \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2$$

其中 $W_h = \dfrac{N_h}{N}$，$N = 530$，代入具体值可计算最终方差。