# 数据采集方法作业

**姓名：**蒋贵豪　　**学号：**B+X9bo

2021 年 10 月 25 日

**题目 1.** 邮局欲估计每个家庭的平均订报份数，该辖区共有 4000 户，划分为 400 个群，每群 10 户，现随机抽取 4 个群，取得资料如下表所示。试估计平均每户家庭的订报份数及总的订报份数，以及估计量的方差。

| 群 | 各户订报数 $y_{ij}$ | $y_i$ |
|---|---|---|
| 1 | 1，2，1，3，3，2，1，4，1，1 | 19 |
| 2 | 1，3，2，2，3，1，4，1，1，2 | 20 |
| 3 | 2，1，1，1，1，3，2，1，3，1 | 16 |
| 4 | 1，1，3，2，1，5，1，2，3，1 | 20 |

**解答.** 由题意知：$N = 400, n = 4, M_i = 10$。故**平均每户家庭的订报份数的估计为：**

$$\overline{\overline{y}} = \sum_{i=1}^{n}\sum_{j=1}^{M}\frac{y_{ij}}{nM} = \sum_{i=1}^{n}\frac{y_i}{nM_i} = \frac{19+20+16+20}{4\times 10} = 1.875$$

**于是，平均每户家庭的订报份数约为 2 份。**

**总的订报份数的估计为：**

$$\widehat{Y} = N\overline{\overline{y}} = 7500$$

**平均每户家庭的订报份数的方差为:**

$$v(\overline{\overline{y}}) = \frac{1-f}{nM}s_b^2 = \frac{1-f}{nM}\times\frac{M}{n-1}\times\sum_{i=1}^{n}(\overline{y}_i - \overline{\overline{y}})^2$$
$$= \frac{1-\frac{n}{N}}{n(n-1)}\times\sum_{i=1}^{n}(\overline{y}_i - \overline{\overline{y}})^2 = 8.86875\times 10^{-3}$$

总的订报份数的方差为:

$$v(\widehat{Y}) = N^2M^2v(\overline{y}) = 141900$$

**题目 2.** 为了便于管理，将某林区划分为 386 个小区域。现采用简单随机抽样方法，从中抽出 20 个小区域，测量树的高度，得到如下表的资料。估计整个林区树的平均高度及 95% 的置信区间。

| 区域编号 | 树木株数 $M_i$ | 平均高度 $\bar{y}_i$（尺） | 区域编号 | 树木株数 $M_i$ | 平均高度 $\bar{y}_i$（尺） |
|---|---|---|---|---|---|
| 1 | 42 | 6.2 | 11 | 60 | 6.3 |
| 2 | 51 | 5.8 | 12 | 52 | 6.7 |
| 3 | 49 | 6.7 | 13 | 61 | 5.9 |
| 4 | 55 | 4.9 | 14 | 49 | 6.1 |
| 5 | 47 | 5.2 | 15 | 57 | 6.0 |
| 6 | 58 | 6.9 | 16 | 63 | 4.9 |
| 7 | 43 | 4.3 | 17 | 45 | 5.3 |
| 8 | 59 | 5.2 | 18 | 46 | 6.7 |
| 9 | 48 | 5.7 | 19 | 62 | 6.1 |
| 10 | 41 | 6.1 | 20 | 58 | 7.0 |

**解答.** 有题意知，区域个数 $N = 386$, 抽样的区域个数 $n = 20$. 样本高度和为：

$$y = \sum_{i=1}^{n} M_i\overline{y}_i = 6180.8$$

群的高度总和均值为：

$$\overline{y} = \frac{\sum\limits_{i=1}^{n} M_i\overline{y}_i}{n} = \frac{6180.8}{20} = 309.04$$

群的平均树木株数为：

$$\overline{m} = \frac{\sum\limits_{i=1}^{n} M_i}{n} = 52.3$$

整个林区树的平均高度的估计为：

$$\overline{\overline{y}} = \sum_{i=1}^{n} \frac{M_i\overline{y}_i}{\overline{m}n} = 5.909$$

整个林区树的平均高度的方差为：

$$
\begin{aligned}
v(\overline{\overline{y}}) &= \frac{N^2(1-f)}{n(N\overline{m})^2} \times \frac{\sum\limits_{i=1}^{n}(y_i-\overline{y})^2}{n-1} \\
&= \frac{1-f}{n\overline{m}^2} \times \frac{\sum\limits_{i=1}^{n}(y_i-\overline{y})^2}{n-1} \\
&= \frac{1-\frac{20}{386}}{20\times 52.3^2} \times 3492.2 = 0.0605
\end{aligned}
$$

因此，**整个林区树的平均高度 95% 的置信区间为：**

$$
\left[\overline{\overline{y}} - t_{\frac{\alpha}{2},n-1}\sqrt{v(\overline{\overline{y}})}, \overline{\overline{y}} + t_{\frac{\alpha}{2},n-1}\sqrt{v(\overline{\overline{y}})}\right] = [5.909-2.093\times0.246, 5.909+2.093\times0.246] = [5.3941, 6.4239]
$$

**题目 3.** 某高校学生会欲对全校女生拍摄过个人艺术照的比例进行调查。全校共有女生宿舍 200 间，每间 6 人。学生会的同学运用两阶段抽样法设计了抽样方案，从 200 间宿舍中抽取了 10 间样本宿舍，在每间样本宿舍中抽取了 3 位同学进行访问，两个阶段的抽样都是简单随机抽样，调查结果如下表。试估计拍摄过个人艺术照的女生比例，并给出估计的标准差。

| 样本宿舍 | 拍照人数 | 样本宿舍 | 拍照人数 |
|---|---|---|---|
| 1 | 2 | 6 | 1 |
| 2 | 0 | 7 | 0 |
| 3 | 1 | 8 | 1 |
| 4 | 2 | 9 | 1 |
| 5 | 1 | 10 | 0 |

**解答.** 由题意知：$n=10$, $N=200$. **拍摄过个人艺术照的女生比例估计为：**

$$
\overline{\overline{p}} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij} = \frac{1}{10\times 3}\times 9 = 0.3
$$

拍摄过个人艺术照的女生比例的方差为：

$$
\begin{aligned}
v(\overline{\overline{p}}) &= \frac{1-f_1}{n}s_1^2 + \frac{f_1(1-f_2)}{nm}s_2^2 \\
&= \frac{1-f_1}{n} \times \frac{1}{n-1}\sum_{i=1}^{n}(\overline{p}_i - \overline{\overline{p}})^2 + \frac{f_1(1-f_2)}{nm} \times \frac{1}{n(m-1)}\sum_{i=1}^{n}\sum_{j=1}^{m}(p_{ij} - \overline{p}_j)^2 \\
&= \frac{1-\frac{10}{200}}{10} \times \frac{1}{10-1} \times 0.54444 + \frac{\frac{10}{200}\times(1-\frac{3}{6})}{10\times 3} \times \frac{1}{10\times(3-1)} \times 4.6667 \\
&= 5.9408\times 10^{-3}
\end{aligned}
$$

于是，**拍摄过个人艺术照的女生比例的标准差为：**

$$s(\overline{\overline{p}}) = \sqrt{v(\overline{\overline{p}})} = 0.077$$