数据采集方法作业

姓名：蒋贵豪　　**学号：**B+X9bo

2021 年 10 月 11 日

题目 1. 为了合理调配电力资源，某市欲了解 50000 户居民的日用电量，从中简单随机抽取了 300 户进行调查，现得到其日用电平均值为 $\bar{y} = 9.5$（千瓦时），$s^2 = 206$。试估计该市居民日用电量的 95% 置信区间。如果希望相对误差限不超过 10%，则样本量至少应为多少？

解答. 由题意知：$N = 50000$, $n = 300$, $\bar{y} = 9.5$, $s^2 = 206$, 则有：

$$\begin{aligned}\mathrm{Var}(\widehat{Y}) = \mathrm{Var}(N\overline{y}) &= N^2\frac{1-f}{n}s^2 \\ &= 50000^2 \times \frac{1-\frac{300}{50000}}{300} \times 206 = \frac{5119100000}{3}\end{aligned}$$

于是：

$$\sqrt{\mathrm{Var}(\widehat{Y})} = \sqrt{\frac{5119100000}{3}} \approx 41308.19$$

从而该市居民日用电量的 95% **置信区间**为:

$$\begin{aligned}&\left[N\overline{y} - z_{\frac{\alpha}{2}}\sqrt{\mathrm{Var}(\widehat{Y})}, N\overline{y} + z_{\frac{\alpha}{2}}\sqrt{\mathrm{Var}(\widehat{Y})}\right] \\ &= [50000 \times 9.5 - 1.96 \times 41308.19, 50000 \times 9.5 + 1.96 \times 41308.19] \\ &= [394035.95, 555964.05]\end{aligned}$$

又由题意知：相对误差限 $r = 0.1$，则：

$$n = \frac{{z_{\frac{\alpha}{2}}}^2 s^2}{(r\bar{y})^2 + \frac{{z_{\frac{\alpha}{2}}}^2 s^2}{N}} = \frac{1.96^2 \times 206}{(0.1 \times 9.5)^2 + \frac{1.96^2 \times 206}{50000}} \approx 861.75$$

于是，如果希望相对误差限不超过 10%，则**样本量至少应为 862**。

题目 2. 某大学有 10000 名本科生，现欲估计在暑假期间参加了各类英语培训的学生所占的比例。随机抽取了 200 名学生进行调查，得到 $p = 0.35$。试估计该大学所有本科生中暑假参加培训班的比例的 95% 的置信区间。

解答. 由题意知：

$$\begin{aligned}&\mathrm{Var}(\widehat{p}) = \tfrac{1}{n-1}(1-\tfrac{n}{N})p(1-p)\\&= \tfrac{1}{200-1} \times (1-\tfrac{200}{10000}) \times 0.35 \times 0.65 = \tfrac{4459}{3980000}\end{aligned}$$

从而，该大学所有本科生中暑假参加培训班的比例的 95% 的**置信区间**为：

$$\begin{aligned}&[p - z_{\frac{\alpha}{2}}\sqrt{\mathrm{Var}(\widehat{p})}, p + z_{\frac{\alpha}{2}}\sqrt{\mathrm{Var}(\widehat{p})}]\\&= [0.35 - 1.96 \times \sqrt{\tfrac{4459}{3980000}}, 0.35 + 1.96 \times \sqrt{\tfrac{4459}{3980000}}]\\&= [0.2844, 0.4156]\end{aligned}$$

题目 3. 研究某小区家庭用于文化方面（报刊、电视、网络、书籍等）的支出，$N = 200$，现抽取一个容量为 20 的样本，调查结果列于下表。估计该小区平均的文化支出 $\bar{Y}$，并给出置信水平 95% 的置信区间。

单位：元

编号	文化支出	编号	文化支出
1	200	11	150
2	150	12	160
3	170	13	180
4	150	14	130
5	160	15	150
6	130	16	100
7	140	17	180
8	100	18	100
9	110	19	170
10	140	20	120

解答. 由题意知：$N = 200$, $n = 20$, 则有：

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = 144.5$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2 \approx 826.0526$$

$$Var(\overline{y}) = \frac{1-\frac{n}{N}}{n}s^2 = \frac{1-\frac{20}{200}}{20} \times 826.0526 \approx 37.1234$$

$$\sqrt{Var(\overline{y})} \approx 6.097$$

于是，估计该小区**平均的文化支出** $\bar{Y} = \bar{y} = 144.5$，置信水平 95% 的**置信区间**为：

$$[\overline{y} - t_{\frac{\alpha}{2},n-1}\sqrt{Var(\overline{y})}, \overline{y} + t_{\frac{\alpha}{2},n-1}\sqrt{Var(\overline{y})}]$$
$$= [144.5 - 2.093 \times 6.907, 144.5 + 2.093 \times 6.907] = [130.0436, 158.9564]$$

题目 4. 如果在解决题目 3 的问题时可以得到这些家庭的月总支出，如下表。而全部家庭的总支出平均为 1600 元，利用比估计的方法估计平均文化支出，给出置信水平 95% 的置信区间，并比较比估计和简单估计的效率。

单位：元

编号	文化支出	总支出	编号	文化支出	总支出
1	200	2 300	11	150	1 600
2	150	1 700	12	160	1 700
3	170	2 000	13	180	2 000
4	150	1 500	14	130	1 400
5	160	1 700	15	150	1 600
6	130	1 400	16	100	1 200
7	140	1 500	17	180	1 900
8	100	1 200	18	100	1 100
9	110	1 200	19	170	1 800
10	140	1 500	20	120	1 300

解答. 由题意知：$N = 200$, $n = 20$, 且有：

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = 144.5$$

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 1580$$

$$\widehat{R} = \frac{\overline{y}}{\overline{x}} = 0.091456$$

由已知条件 $\bar{X} = 1600$, 于是，我们有：

$$\overline{y}_r = \frac{\bar{X}}{\overline{x}}\overline{y} = \frac{1600}{1580} \times 144.5 \approx 146.33$$

$$S_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2 \approx 826.05$$

$$S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \approx 99578.95$$

$$S_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) \approx 8831.58$$

于是，**平均文化支出为** 146.33 元，平均文化支出置信水平 95% 的**置信区间**为：

$$\begin{aligned}
&[\overline{y}_r - t_{\frac{\alpha}{2},n-1}\sqrt{\tfrac{1-\frac{n}{N}}{n}(S_y{}^2 + \widehat{R}^2 S_x{}^2 - 2\widehat{R}S_{xy})}, \overline{y}_r + t_{\frac{\alpha}{2},n-1}\sqrt{\tfrac{1-\frac{n}{N}}{n}(S_y{}^2 + \widehat{R}^2 S_x{}^2 - 2\widehat{R}S_{xy})}] \\
&= [146.33 \pm 2.093 \times \sqrt{\tfrac{1-\frac{20}{200}}{20}(826.05 + 0.091456^2 \times 99578.95 - 2 \times 0.091456 \times 8831.58)} \\
&= [143.4041, 149.2559]
\end{aligned}$$

对比比估计和简单估计的置信水平 95% 的置信区间，我们发现，比估计的置信区间比简单估计更窄。于是，在此情况下，**比估计比简单估计的效率更高**。