抽样调查 第四周作业

20307100013 蒋翌坤

§2.13 Exercises：12

$n_0=\frac{1.96^2}{e^2}p(1-p)=\frac{1.96^2}{0.04^2}\times 0.25=600.25$. 向上取整，得到$n_0=601$.

对于 City of Casa Grande , Phoenix, Tempe, $n\ll N$，可以忽略 fpc, $n=n_0=601$

对于 City of Gila Bend, 需要考虑 fpc，$n=\frac{n_0}{1+n_o/N}=\frac{601}{1+601/1922}=458$

对于 City of Jerome, $N<n_0$，即使抽取总体也无法满足$e\le 0.04$的要求。

补充：

由题意，$p\ge 0.15$，$e\le 0.1$，进行保守估计，取$p=0.5$，$n_0=\frac{1.96^2}{e^2}p(1-p)=\frac{1.96^2}{0.1^2}\times 0.25=96.04$，向上取整，得到$n_0=97$.

对于 City of Casa Grande , Phoenix, Tempe, $n\ll N$，可以忽略 fpc, $n=n_0=97$

对于 City of Gila Bend, 需要考虑 fpc，$n=\frac{n_0}{1+n_o/N}=\frac{97}{1+97/1922}=93$

对于 City of Jerome, 需要考虑 fpc，$n=\frac{n_0}{1+n_o/N}=\frac{97}{1+97/444}=80$

§2.13 Exercises：15

（c）采用题目中所给样本s^2来预估总体S^2. $s^2=3.705$

忽略 fpc，$n=\left(1.96\times\frac{\sqrt{3.705}}{0.5}\right)^2=57$. 因此，需要 57 的样本容量使在 0.95 置信水平下，平均年龄的最大绝对误差为 0.5。

补充：

采用题目中所给样本$\bar{y}$来预估总体$\overline{y_u}$. $\bar{y}=12.08$

忽略 fpc，$n=\frac{1.96^2}{0.1^2}\times\left(\frac{\sqrt{3.705}}{12.08}\right)^2=10$. 因此，需要 10 的样本容量使在 0.95 置信水平下，平均年龄的最大相对误差为 10%。

§2.13 Exercises：26

Let $f(n)=L(n)+C(n)=\frac{kS^2}{n}-\frac{kS^2}{N}+c_o+c_1n$

$$f'(n)=L'(n)+C'(n)=\left(\frac{kS^2}{n}-\frac{kS^2}{N}+c_o+c_1n\right)'=-\frac{kS^2}{n^2}+c_1$$

To minimize $f(n)$, $f'(n)=0\Longrightarrow -\frac{kS^2}{n^2}+c_1=0\Rightarrow n=\sqrt{\frac{kS^2}{c_1}}$

Therefore, when $n=\sqrt{\frac{kS^2}{c_1}}$, total cost $L(n)+C(n)$ is minimized.

§2.13 Exercises：42

（a） Topcoding 可能会对总体的估计造成以下影响：（1）对总体收入均值的估计会较真实值偏低；（2）对总体收入方差的估计会较真实值偏低

（b）从总体中抽取 50 个简单随机样本，得到变量 inctot 的方差$s^2=8.76\times10^7$. $n=\left(1.96\times\frac{\sqrt{8.76\times10^7}}{700}\right)^2=688$. 由于$n\ll N$，忽略 fpc，因此，需要 688 的样本容量使在 0.95 置信水平下，平均收入的最大绝对误差为 700.

（c）从总体中抽取 688 个简单随机样本，得到对总体收入的估计$\overline{y_u}=\bar{y}=8847$，$t=\hat{t}=472{,}958{,}899$，0.95 的置信区间$N\bar{y}\pm1.96N\sqrt{\frac{s^2}{n}}=[431{,}439{,}207, 514{,}478{,}591]$

补充：

利用（b）问中 50 个简单随机样本所获得的估计量（inctot 的均值$\overline{y_u}=\bar{y}=9033$、方差$s^2=8.76\times10^7$；educrec>5 的比例$\hat{p}=0.72$）来确定所需要的样本量。

要求估计 inctot 总体均值的最大相对误差不超过 10%的最小样本量：$n=\frac{1.96^2}{0.1^2}\times\left(\frac{\sqrt{8.76\times10^7}}{9033}\right)^2=413$, 要求估计 educrec>5 的人数比例最大绝对误差不超过 0.05 的最小样本量 $n=\frac{1.96^2}{0.05^2}\times0.72\times(1-0.72)=310$.

因此，要求估计 inctot 总体均值的最大相对误差不超过 10%、且同时要求估计 educrec>5 的人数比例的最大绝对误差不超过 0.05，样本量应该取 413.

§2.13 Exercises：27

（2）通过从（1）问中形成的伪总体中抽取n个 SRS 样本，可以得到$\overline{y_j}^*=302{,}179$, $s_j^*=369{,}538$, $\hat{y}_{med,j}^*=196{,}717$

（3）通过 1000 次重复 bootstrap，可以得到总体均值$\overline{y_u}$的近似置信区间$[\overline{y_L}^*,\overline{y_U}^*]=[263{,}625, 336{,}648]$。利用 Hájek 中心极限定理，300 个样本的均值为$\bar{y}=297{,}897$, $SE(\bar{y})=19{,}892.7$, $\bar{y}$的 0.95 置信区间为$\bar{y}\pm1.96SE(\bar{y})=[258{,}907, 336{,}887]$。可以发现，由 bootstrap 方法所得到的近似置信区间的范围与利用 Hájek 中心极限定理所得到的相同置信区间接近。

相同方法可以得到S的近似 0.95 置信区间：$[281{,}070, 398{,}878]$，以及y_{med}的近似 0.95 置信区间：$[168{,}051, 223{,}764]$。

附录：

解答题目所使用的 R 代码及输出请见 https://thisiskunmeng.github.io/sampling/hw4.html