

1 第一问

1.1 累计规模法介绍

记总量为 N ，各单元的规模为 $\{M_i\}_{i=1,\dots,N}$ ，累计规模 $C_k = \sum_{i=1}^k M_i, k = 1, \dots, N$ ，其中 $C_0 = 0$ 。累计规模法会生成区间列 $\{I_i\}_{i=1,\dots,N}$ ，其中 $I_i = [C_{i-1} + 1, C_i]$ 。

然后在 $[1, \max\{M_1, \dots, M_N\}]$ 生成 n 个随机整数 $\{Z_i\}_{i=1,\dots,n}$ ，其中 n 为样本容量。对每一个 $i = 1, \dots, n$ ，若 Z_i 落在了区间 I_j 中，则第 j 个总体入样。

1.2 R 语言 PPS 抽样

以数据集中的 `popn`（人口数）作为规模，确定每个州被抽中的概率。设第 i 个州的人口数为 P_i ，则其被抽中的概率应为 $\phi_i = \frac{P_i}{\sum_{j=1}^n P_j}$ 。R 语言代码如下：

```
1      df<-read.csv("statepps.csv")
2      library("sampling")
3      library("survey")
4      N=nrow(df)
5      n=10
6      pik=inclusionprobabilities(df$popn, n)
7      s=UPmultinomial(pik)
8      df.pps = df[s!=0, ]
9      Z=pik[s!=0]/n #每个单元被抽中的概率
10     Q=s[s!=0] #每个单元被抽中的次数
```

1.3 参数估计

总体总量的无偏估计应为：

$$\hat{Y}_{HH} = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{\phi_i}$$

方差的无偏估计应为：

$$Var(\hat{Y}_{HH}) = \frac{1}{n(n-1)} \sum_{i=1}^n \left(\frac{y_i}{\phi_i} - \hat{Y}_{HH} \right)^2$$

估计标准误为：

$$SE = \sqrt{Var(\hat{Y}_{HH})}$$

R 语言代码如下：

```

1      Yhh=mean(df.pps$counties/Z*Q)
2      vars=1/(n*(n-1))*sum((df.pps$counties/Z-Yhh)^2*Q)
3      std=sqrt(vars)

```

输出结果为: $\hat{Y}_{HH} = 3686.316$, $SE = 655.797$

2 第二问

2.1 参数估计原理

先用 ssu 样本估计各 psu 的 \hat{Y}_{HH_i} , 各 psu 的总量的估计为:

$$\hat{Y}_{HH_i} = \frac{M_i}{k_i} \sum_{j=1}^{k_i} y_{ij}$$

其中 k_i 是第 i 个 psu 的中抽取的 ssu 的样本个数。

然后可以求总体总量的估计:

$$\hat{Y}_{HH} = \frac{1}{n} \sum_{i=1}^n \frac{\hat{Y}_{HH_i}}{\phi_i}$$

方差的无偏估计应为:

$$Var(\hat{Y}_{HH}) = \frac{1}{n(n-1)} \sum_{i=1}^n \left(\frac{\hat{Y}_{HH_i}}{\phi_i} - \hat{Y}_{HH} \right)^2$$

估计标准误为:

$$SE = \sqrt{Var(\hat{Y}_{HH})}$$

2.2 计算

本题中, psu 样本数 $n = 10$ 。

用题中数据计算的结果如下表。

学院编号	ϕ_i	\hat{Y}_{HH_i}
14	0.08054523	113.75
23	0.03097893	31.25
9	0.05947955	12
14	0.08054523	48.75
16	0.002478315	2
6	0.07682776	139.5
14	0.08054523	65
19	0.07682776	77.5
21	0.07558860	122
11	0.05080545	225.5

算得其总量的估计为

$$\hat{Y}_{HH} = \frac{1}{n} \sum_{i=1}^n \frac{\hat{Y}_{HH_i}}{\phi_i} \approx 1371.9$$

标准误为:

$$SE = \sqrt{\frac{1}{n(n-1)} \sum_{i=1}^n \left(\frac{\hat{Y}_{HH_i}}{\phi_i} - \hat{Y}_{HH} \right)^2} \approx 372.980$$

2.3 Lahiri 方法介绍

设总体中有 N 个个体，第 i 个单元的规模为 M_i ，涉及的抽样均为有放回的抽样。

1. 抽取随机数 $i \in \{1, \dots, N\}$
2. 抽取随机数 $M \in \{1, \dots, \max\{M_1, \dots, M_N\}\}$
3. 若 $M_i \geq M$ ，则第 i 个样本入样，否则不入样。
4. 重复上述过程，直至抽满 n 个单元