数据采集方法作业

姓名：蒋贵豪　　**学号：**B+X9bo

2021 年 11 月 1 日

题目 1. 对于下面的 25 个 PSU 单元，研究人员希望抽取 10 个不等概的无放回的样本。

(1) 采用代码法抽取容量为 10 的无放回的样本。

(2) 采用拉希里法抽取容量为 10 的无放回的样本。

PSU	Z_i	PSU	Z_i
1	0.000 110	14	0.014 804
2	0.018 556	15	0.005 577
3	0.062 999	16	0.070 784
4	0.078 216	17	0.069 635
5	0.075 245	18	0.034 650
6	0.073 983	19	0.069 492
7	0.076 580	20	0.036 590
8	0.038 981	21	0.033 853
9	0.040 772	22	0.016 959
10	0.022 876	23	0.009 066
11	0.003 721	24	0.021 795
12	0.024 971	25	0.059 185
13	0.040 654		

解答. (1) 由于 Z_i 不是整数，我们将 Z_i 统一乘以 10^6，使其变为整数，即我们的代码。得到代码后，我们计算代码的累计和为 1000054，然后给出每个单元所对应的代码，该过程如下表所示。

接着，我们在 $[1, 1000054]$ 生成一个随机数为 722822，对应的单元为 18。于是，我们的第一个样本单元为 18。

然后，我们将单元 18 除去，得到剩余的 24 个单元，我们对其重新进行编码，累计代码数为 $1000054 - 34650 = 965494$。我们在 $[1, 965494]$ 中随机生成一个随机数为

353480，对应的单元为 7，于是我们的第二个样本单元为 7。

重复上述过程，直至抽到 10 个样本。使用 **Matlab** 语言对该过程进行模拟，代码见附录。**我们使用代码法抽到的样本为：**18，7，10，8，4，6，16，3，19，25。

PSU	Z_i	$1000000Zi$	累计	代码
1	0.000110	110	110	1 - 110
2	0.018556	18556	18666	111 - 18666
3	0.062999	62999	81665	18667 - 81665
4	0.078216	78216	159881	81666 - 159881
5	0.075245	75245	235126	159882 - 235126
6	0.073983	73983	309109	235127 - 309109
7	0.076580	76580	385689	309110 - 385689
8	0.038981	38981	424670	385690 - 424670
9	0.040772	40772	465442	424671 - 465442
10	0.022876	22876	488318	465443 - 488318
11	0.003721	3721	492039	488319 - 492039
12	0.024971	24971	517010	492040 - 517010
13	0.040654	40654	557664	517011 - 557664
14	0.014804	14804	572468	557665 - 572468
15	0.005577	5577	578045	572469 - 578045
16	0.070784	70784	648829	578046 - 648829
17	0.069635	69635	718464	648830 - 718464
18	0.034650	34650	753114	718465 - 753114
19	0.069492	69492	822606	753115 - 822606
20	0.036590	36590	859196	822607 - 859196
21	0.033853	33853	893049	859197 - 893049
22	0.016959	16959	910008	893050 - 910008
23	0.009066	9066	919074	910009 - 919074
24	0.021795	21795	940869	919075 - 940869
25	0.059185	59185	1000054	940870 - 1000054

(2) 由题意知，$N=25$，设 $Z^*=\max\{Z_i\}=0.078216$。我们在 $[1,25]$ 和 $[0,0.078216]$ 中随机抽取两个数为 24 和 0.041156。我们有 $Z_{24}=0.021795<0.041156$。于是该样本不入样。我们重新在 $[1,25]$ 和 $[0,0.078216]$ 中随机抽取两个数为 3 和 0.010099，我们有 $Z_3=0.062999>0.010099$。于是我们的第一个样本为 3 号单元。

去除第 3 个单元，在剩余的 24 个单元中，$Z^*=\max\{Z_i\}=0.078216$。重复上述操作，直至抽满 10 个样本。使用 **Matlab** 语言对该过程进行模拟，代码见附录。**我们使拉希里法抽到的样本为：**3，4，8，24，17，21，5，19，9，12。

题目 2. 下表给出了一个整群的总体。无放回地选择两个 PSU 单元，入样概率正比于 M_i。使用布鲁尔方法构建所有可能样本的 π_{ij} 的表格。计算霍维茨-汤普森估计量的方差。

PSU	M_i	y_{ij}	t_i
1	5	3, 5, 4, 6, 2	20
2	4	7, 4, 7, 7	25
3	8	7, 2, 9, 4, 5, 3, 2, 6	38
4	5	2, 5, 3, 6, 8	24
5	3	9, 7, 5	21

解答. 由题意知：$M_0=\sum_{i=1}^{5}M_i=25$, $Z_i=\dfrac{M_i}{M_0}$。布鲁尔方法的包含概率为：

$$\pi_{ij}=\frac{4Z_iZ_j\left(1-Z_i-Z_j\right)}{\left(1-2Z_i\right)\left(1-2Z_j\right)\left(1+\sum_{i=1}^{N}\dfrac{Z_i}{1-2Z_i}\right)}$$

霍维茨-汤普森估计量的方差为：

$$v\left(\hat{Y}_{\pi ij}\right)=\sum_{i=1}^{N}\sum_{j>i}^{N}\left(\pi_i\pi_j-\pi_{ij}\right)\left(\frac{Y_i}{\pi_i}-\frac{Y_j}{\pi_j}\right)^2=243.0652 \tag{1}$$

使用布鲁尔方法构建所有可能样本的 π_{ij} 的表格如下表所示。

样本	π_{ij}
1,2	0.068092
1,3	0.192926
1,4	0.090434
1,5	0.048549
2,3	0.147531
2,4	0.068092
2,5	0.036286
3,4	0.192926
3,5	0.106617
4,5	0.048549
求和	1

题目 3. 某市建筑行业集团共有 48 个单位，有载货汽车 186 辆。按与每个单位的车辆拥有量成比例的概率进行放回的 PPS 抽样，共抽取 10 次。对抽中单位的所有车辆调查季度运量（单位：吨）。样本数如下表所示（其中有一家单位被抽中两次，即 $i = 3, 7$）。试估计全集团的季度总运量及 95% 的置信区间。

单位编号	车辆数 M_i	单位运量总和 y_i	平均每车运量 $\bar{y}_i$
1	5	14 230	2 846
2	8	21 336	2 667
3	5	13 650	2 730
4	4	11 568	2 892
5	6	15 216	2 536
6	9	23 049	2 566
7	5	13 650	2 730
8	3	7 443	2 481
9	7	16 723	2 389
10	3	8 391	2 797

解答. 由题意知：$M_0 = 186$, **全集团季度总运量的估计为：**

$$\widehat{Y}_{HH} = \frac{1}{n}\sum_{i=1}^{10}\frac{y_i}{Z_i} = \frac{M_0}{n}\sum_{i=1}^{10}\frac{y_i}{M_i} = 495299.4$$

全集团季度总运量的方差为：

$$v(\widehat{Y}_{HH}) = \frac{1}{n(n-1)}\sum_{i=1}^{10}\left(\frac{y_i}{Z_i} - \widehat{Y}_{HH}\right)^2 = 95182398.76$$

于是，**全集团的季度总运量及 95% 的置信区间为：**

$$[\widehat{Y}_{HH} \pm t_{\frac{\alpha}{2},n-1}\sqrt{v(\widehat{Y}_{HH})}] = [495299.4 \pm 2.2622 \times 9756.15] = [473229.0375, 517369.7625]$$

附录

代码法抽取无放回的样本：

```
function [B, D] = code(A, D_1)
#输入A为单元大小M_i，D_1为对应的单元i
#输出无放回抽样后的单元大小B和对应的单元D
a = length(A);
for i = 1:a
    sumA(i) = sum(A(1:i));
end
Rand = sumA(a)*rand(1);
j = 1;
while (sumA(j) < Rand)
    j = j+1;
end
k = j;
if k == 1
    B(1:a-1) = A(2:a);
    D(1:a-1) = D_1(2:a);
else
    B(1:k-1) = A(1:k-1);
    D(1:k-1) = D_1(1:k-1);
    B(k:a-1) = A(k+1:a);
    D(k:a-1) = D_1(k+1:a);
end
```

拉希里法抽取无放回的样本：

```
function [B,D] = lahici(A,D_1)
#输入A为单元大小M_i，D_1为对应的单元i
#输出无放回抽样后的单元大小B和对应的单元D
a = length(A);
Z = max(A);
Rand = floor(a * rand(1)) + 1;
Rand_1 = floor(Z * rand(1))+1;
while A(Rand) < Rand_1
    Rand = floor(a * rand(1)) + 1;
    Rand_1 = floor(Z * rand(1)) + 1;
end
k = Rand;
if Rand == 1
    B(1:a-1) = A(2:a);
    D(1:a-1) = D_1(2:a);
else
    B(1:k-1) = A(1:k-1);
    D(1:k-1) = D_1(1:k-1);
    B(k:a-1) = A(k+1:a);
    D(k:a-1) = D_1(k+1:a);
end
```