习题6.6

a. b. 样本容量为1的不等概率抽样、等概率抽样

以县为单位对2000年人口总体进行放回的PPS抽样，选择概率 $\psi_i = M_i/M_0$ ，抽样的样本容量 $n=1$ ，则有

$$\hat{t}_\psi = \frac{1}{n}\sum_{i\in\mathcal{R}}\frac{t_i}{\psi_i} = \frac{t_i}{\psi_i}\ ,\ V(\hat{t}_\psi) = \frac{1}{n}\sum_{i=1}^{N}\psi_i\left(\frac{t_i}{\psi_i}-t\right)^2 = \sum_{i=1}^{N}\psi_i\left(\frac{t_i}{\psi_i}-t\right)^2$$

以县为单位对2000年人口总体进行放回的等概率抽样，即简单随机抽样SRS，选择概率为 $1/N = 1/13$ ，抽样的样本容量 $n=1$ ，则有

$$\hat{t}_{\text{SRS}} = \frac{t_i}{1/N} = Nt_i\ ,\ V(\hat{t}_{\text{SRS}}) = \frac{1}{n}\sum_{i=1}^{N}\frac{1}{N}\left(\frac{t_i}{1/N}-t\right)^2 = \sum_{i=1}^{N}\frac{1}{N}\left(\frac{t_i}{1/N}-t\right)^2$$

各统计量的计算代码和结果如下所示：

```r
library(knitr); library(kableExtra)
azcounties = read.csv('azcounties.csv'); N = 13
azcounties$psi_i = azcounties$population / sum(azcounties$population)
azcounties$t_psi = azcounties$housing / azcounties$psi_i
azcounties$t_SRS = N * azcounties$housing
M0 = sum(azcounties$population); t = sum(azcounties$housing)

attach(azcounties)
V_t_psi = sum(psi_i * (housing/psi_i - t)^2)
V_t_SRS = sum(1/N * (N * housing - t)^2)
detach(azcounties)

res_print = azcounties
colnames(res_print) = c('$i$','County','$M_i$','$t_i$','$\\psi_i$',
                        '$\\hat{t}_{\\psi}$','$\\hat{t}_{\\text{SRS}}$')
res_print %>% kable(format='latex', digits=4, escape=F, booktabs=T,
                    caption='不等概率抽样、等概率抽样的统计量') %>%
  kable_styling(latex_options='HOLD_position')
```

表 1: 不等概率抽样、等概率抽样的统计量

i	County	M_i	t_i	ψ_i	$\hat{t}_\psi$	$\hat{t}_{\text{SRS}}$
1	Apache	69423	31621	0.0572	553292.1	411073
2	Cochise	117755	51126	0.0969	527405.6	664638
3	Coconino	116320	53443	0.0958	558108.6	694759
4	Gila	51335	28189	0.0423	667034.6	366457
5	Graham	33489	11430	0.0276	414597.1	148590
6	Greenlee	8547	3744	0.0070	532113.6	48672
7	La Paz	19715	15133	0.0162	932417.7	196729
8	Mohave	155032	80062	0.1276	627317.4	1040806
9	Navajo	97470	47413	0.0802	590892.8	616369
10	Pinal	179727	81154	0.1480	548502.8	1055002
11	Santa Cruz	38381	13036	0.0316	412582.0	169468
12	Yavapai	167517	81730	0.1379	592659.0	1062490
13	Yuma	160026	74140	0.1317	562787.3	963820

计算得 $M_0 = 1214737$, $t = 572221$, $V(\hat{t}_\psi) = 4789282131$, $V(\hat{t}_{\text{SRS}}) = 130534375140$, 则有 $V(\hat{t}_\psi) < V(\hat{t}_{\text{SRS}})$, 即不等概率抽样比等概率抽样（这里是简单随机抽样）更有效。

c. 样本容量为3的有放回PPS抽样

采用Lahiri方法抽取样本容量为3的有放回PPS抽样，具体步骤为：记 $M^* = \max\{M_1, \ldots, M_N\}$. 先从 1~N 中等概率地抽取随机数 i , 再从 1~M^* 中等概率地产生随机数 m . 若 $m \le M_i$, 则单元 i 被抽中；否则重抽 (i, m) .

抽样的代码和结果如下：

```
M_star = max(azcounties$population); n = 3; sampleNum = c()
set.seed(12345)
while (length(sampleNum) < n){
  i = sample(1:N, 1); m = sample(1:M_star, 1)
  if (m <= azcounties$population[i]){
    sampleNum = append(sampleNum, i)
  }
}
print(paste('抽中的样本单元为', paste0(sampleNum,collapse=',')), quote=F)

azcounties_sample = azcounties[sampleNum,]
attach(azcounties_sample)
that_psi = mean(housing/psi_i)
Vhat_t_psi = 1/(n*(n-1)) * sum((housing/psi_i - that_psi)^2)
detach(azcounties_sample)
```

```
## [1] 抽中的样本单元为 3,10,8
```

根据表1计算得

$$\hat{t}_\psi = \frac{1}{n}\sum_{i\in\mathcal{R}}\frac{t_i}{\psi_i} = 577976\ ,\ \hat{V}(\hat{t}_\psi) = \frac{1}{n}\frac{1}{n-1}\sum_{i\in\mathcal{R}}\left(\frac{t_i}{\psi_i} - \hat{t}_\psi^2\right) = 616326369$$

习题6.8

1. 估计总论文发表数及标准误

根据样本数据中的 y_{ij} 计算 $\hat{t}_{ij}$ ，有

$$\hat{t}_{ij} = \frac{M_i}{m_i}\sum_{j\in\mathcal{S}_i} y_{ij}$$

然后用不等概率二阶抽样（抽取PSU时为PPS）的方法计算总论文发表数的估计量及标准误，代码和结果如下：

```
publication = read.csv('publication.csv'); n = 10
publication$t_ij = publication$sum_y / publication$m_i * publication$M_i
attach(publication)
that_psi = mean(t_ij/psi_i)
Vhat_t_psi = 1/(n*(n-1)) * sum((t_ij/psi_i - that_psi)^2)
SE_t_psi = sqrt(Vhat_t_psi)
```

计算得

$$\hat{t}_\psi = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{Q_i}\frac{\hat{t}_{ij}}{\psi_i} = 1372$$

$$\hat{V}(\hat{t}_\psi) = \frac{1}{n(n-1)}\sum_{i=1}^{N}\sum_{j=1}^{Q_i}\left(\frac{\hat{t}_{ij}}{\psi_i} - \hat{t}_\psi\right)^2 = 139113.66\ ,\ \mathrm{SE}(\hat{t}_\psi) = \sqrt{\hat{V}(\hat{t}_\psi)} = 372.98$$

2. 补充：估计人均发表论文数及95%置信区间

由下面的代码

```
Mhat0 = 1/n * sum(M_i/psi_i)
yhat_psi = that_psi / Mhat0
Vhat_y_psi = 1/(Mhat0^2*n*(n-1)) * sum((t_ij/psi_i - yhat_psi * M_i/psi_i)^2)
SE_y_psi = sqrt(Vhat_y_psi)
y_CI_lb = yhat_psi - qnorm(0.975)*SE_y_psi; y_CI_ub = yhat_psi + qnorm(0.975)*SE_y_psi
detach(publication)
```

计算得

$$\hat{M}_{0\psi} = \frac{1}{n}\sum_{i\in\mathcal{R}}\frac{M_i}{\psi_i} = 807\ ,\ \hat{\bar{y}}_\psi = \frac{\hat{t}_\psi}{\hat{M}_{0\psi}} = 1.70$$

$$\hat{V}(\hat{\bar{y}}_\psi) = \frac{1}{\hat{M}_{0\psi}^2 n(n-1)}\sum_{i=1}^{N}\sum_{j=1}^{Q_i}\left(\frac{\hat{t}_{ij}}{\psi_i} - \hat{\bar{y}}_\psi\frac{M_i}{\psi_i}\right)^2 = 0.214\ ,\ \mathrm{SE}(\hat{\bar{y}}_\psi) = \sqrt{\hat{V}(\hat{\bar{y}}_\psi)} = 0.462$$

则人均发表论文数的95%置信区间为

$$\left[\hat{\bar{y}}_\psi - z_{\alpha/2}\mathrm{SE}(\hat{\bar{y}}_\psi)\ ,\ \hat{\bar{y}}_\psi + z_{\alpha/2}\mathrm{SE}(\hat{\bar{y}}_\psi)\right] = [0.794\ ,\ 2.606]$$

习题6.12

a. 农场总数与选择概率的散点图

采用PPS抽样（依据为1992年人口数），则选择概率 $\psi_i = M_i/M_0$. 图1为农场总数与选择概率的散点图：

```
library(latex2exp)
statepop = read.csv('statepop.csv'); M0 = 255077536
statepop$psi_i = statepop$popn / M0
plot(statepop$psi_i, statepop$numfarm, pch=20,
     xlab=TeX('Probabilities of Selection $\\psi_i$'), ylab='Number of Farms')
```

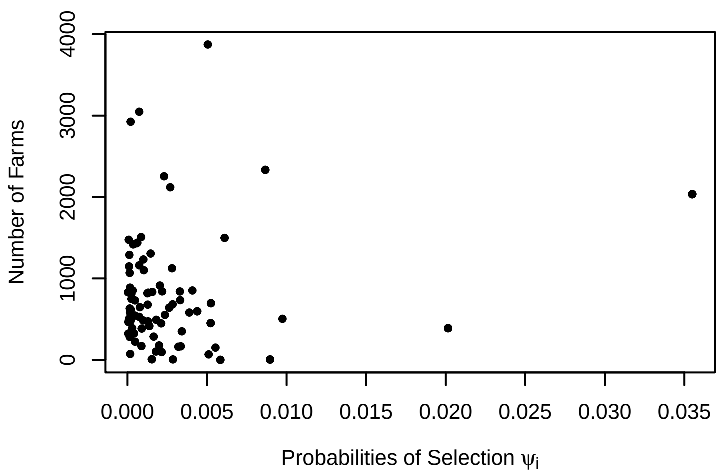

图 1: 农场总数与选择概率的散点图

由图1可知，农场总数与选择概率之间没有明显的正比例关系，所以采用不等概率抽样的估计方法不能很好地提高估计精度。

b. 估计农场总数及标准误

由下面的代码：

```
n = 100
t_i = statepop$numfarm; psi_i = statepop$psi_i
that_psi = mean(t_i/psi_i)
Vhat_t_psi = 1/(n*(n-1)) * sum((t_i/psi_i - that_psi)^2)
SE_t_psi = sqrt(Vhat_t_psi)
```

计算得

$$\hat{t}_\psi = \frac{1}{n}\sum_{i\in\mathcal{R}}\frac{t_i}{\psi_i} = 1896300$$

$$\hat{V}(\hat{t}_\psi) = \frac{1}{n}\frac{1}{n-1}\sum_{i\in\mathcal{R}}\left(\frac{t_i}{\psi_i} - \hat{t}_\psi^2\right) = 134999300087\ ,\ \mathrm{SE}(\hat{t}_\psi) = \sqrt{\hat{V}(\hat{t}_\psi)} = 367423$$

补充：用SRSWR的方法估计农场总数及标准误

根据所查阅资料，美国约有3000个县，取 $N = 3000$, 采用SRSWR的方法估计农场总数及标准误，则每个样本的选择概率 $\psi_i \equiv \psi = 1/N = 1/3000$. 由以下代码

```
psi = 1/3000
that_SRSWR = mean(t_i) / psi
Vhat_t_SRSWR = 1/(n*(n-1)) * sum((t_i/psi - that_SRSWR)^2)
SE_t_SRSWR = sqrt(Vhat_t_SRSWR)
```

计算得

$$\hat{t}_{\mathrm{SRSWR}} = \frac{1}{n\psi}\sum_{i\in\mathcal{R}} t_i = 2422800$$

$$\hat{V}(\hat{t}_{\mathrm{SRSWR}}) = \frac{1}{n}\frac{1}{n-1}\sum_{i\in\mathcal{R}}\left(\frac{t_i}{\psi} - \hat{t}_{\mathrm{SRSWR}}^2\right) = 46448265455\ ,\ \mathrm{SE}(\hat{t}_{\mathrm{SRSWR}}) = \sqrt{\hat{V}(\hat{t}_{\mathrm{SRSWR}})} = 215519$$

则SRSWR对总体总和的估计量比HH估计量偏高，而标准误偏小。这样的原因是，由图1可知，大部分县的农场总数与PPS中的选择概率都较小，而SRSWR夸大了农场总数小的县在估计总体总和时的作用，使得总体总和的估计偏高。并且SRSWR中相同的选择概率使得方差估计公式中的 t_i/ψ 更加集中，这减小了标准误。

SRSWR仅仅是不等概率抽样的一个特例（这个特例中选择概率都相同），在各层的差异较大时，将样本视为SRSWR会造成对各层差异的忽略，导致SRSWR估计量偏离真实值，所以各层差异较大时不应将SRSWR与不等概率抽样的方法混用。在各层差异较小时，为方便处理，可能可以将有放回的不等概率抽样视为SRSWR。