习题 3.20

a. 可能的非抽样误差

题中说每一层的 SRS 是来自该县的电话簿，则可能的非抽样误差有：

- 抽样框不完善：不是所有的家庭信息都在该县的电话簿上；
- 样本中有无回答单元：无法接电话或不愿意接受调查的家庭会造成样本选择偏差。

b. 计算每一层的抽样权重

每一层抽样权重的计算公式为：$w_{hi} = \dfrac{N_h}{n_h}$.

计算每一层抽样权重的代码如下所示：

```
radon = read.csv('radon.csv') # read data
radon_stratum = data.frame(countyname=unique(radon$countyname),
                           n_h=rep(0,length(unique(radon$countyname))),
                           N_h=rep(0,length(unique(radon$countyname))),
                           w_hi=rep(0,length(unique(radon$countyname))))
for (county in unique(radon$countyname)){
  n_h = radon[radon$countyname==county,3][1]
  N_h = radon[radon$countyname==county,4][1]
  radon_stratum[radon_stratum$countyname==county,c(2,3)] = c(n_h, N_h)
  radon_stratum[radon_stratum$countyname==county,4] = N_h / n_h # 每一层抽样权重
}
```

问题 1. 有些层的样本容量 $n_h = 1$

当 $n_h = 1$ 时，层内样本方差为 $s_h^2 = \dfrac{1}{n_h - 1}\sum_{i \in \mathcal{S}_h}(y_i - \bar{y}_h)^2$ ，分母 $n_h - 1 = 0$ ，故无法用此公式计算。可以采用其他方法来对 S_h^2 进行估计，比如：

- 对其他样本量大于 1 的各层的层内样本方差进行简单平均或加权平均，从而估计样本量为 1 的各层的层内样本方差；
- 从样本观测值大小、层权等角度考察层间相似性，将样本量为 1 的各层并入与其相似的层，然后用相似层的层内样本方差来估计样本量为 1 的层的层内样本方差。

对 s_h^2 进行估计之后，再计算 $\hat{V}(\bar{y}_h)$ 和 $\hat{V}(\hat{p}_h)$.

问题 2. 有些层内 0-1 变量的样本总和 $a_h = 0$ 或 n_h

层内 0-1 变量的样本总和 $a_h = 0$ 时，$\hat{p}_h = 0$ ；$a_h = n_h$ 时，$\hat{p}_h = 1$ ，则 $\hat{V}(\hat{p}_h) = \dfrac{1 - f_h}{n_h - 1}\hat{p}_h(1 - \hat{p}_h) = 0$ ，这样的估计较为极端。可以对于 $\hat{p}_h$ 进行一定的调整，比如取 $\hat{p}_h = \dfrac{a_h + 1}{n_h + 2}$ ，然后再根据公式计算 $\hat{V}(\hat{p}_h)$.

c. 估计明尼苏达州的家庭平均氡水平

各估计值的计算公式如下所示：

- 家庭平均氡水平 $\bar{y}_{str} = \sum_{h=1}^{H} W_h \bar{y}_h$ ；
- 家庭平均氡水平的标准误 $\mathrm{SE}(\bar{y}_{str}) = \sqrt{\sum_{h=1}^{H} W_h^2 \frac{1-f_h}{n_h} s_h^2}$ ；
- 家庭平均氡水平的 95% 置信区间 $[\bar{y}_{str} - \mathrm{SE}(\bar{y}_{str}), \bar{y}_{str} + \mathrm{SE}(\bar{y}_{str})]$.

对 $n_h = 1$ 的层（编号为 43、51、84 的县）取 $s_h^2 = 0$ ，则得到 $\mathrm{SE}(\bar{y}_{str})$ 的下界。对 `radon` 和 `log(radon)` 分别采用上述公式计算 $\bar{y}_U$ 的点估计值和 95% 置信区间。

代码如下所示：

```
radon$radon_log = log(radon$radon) # 在原始数据中创建变量 log(radon)
radon_stratum$W_h = radon_stratum$N_h / sum(radon_stratum$N_h) # 每一层的层权
radon_stratum$y_h = rep(0, dim(radon_stratum)[1])  # 初始化每层的样本均值
radon_stratum$s2_h = rep(0, dim(radon_stratum)[1]) # 初始化每层的样本方差
radon_stratum$y_h_log = rep(0, dim(radon_stratum)[1])  # 初始化每层的样本均值, log(radon)
radon_stratum$s2_h_log = rep(0, dim(radon_stratum)[1]) # 初始化每层的样本方差, log(radon)

for (county in unique(radon$countyname)){
  one_county = radon[radon$countyname==county,]
  radon_stratum[radon_stratum$countyname==county,6] = mean(one_county$radon)
  radon_stratum[radon_stratum$countyname==county,8] = mean(one_county$radon_log)
  if (dim(one_county)[1]==1){
    radon_stratum[radon_stratum$countyname==county,c(7,9)] = 0
  }else{
    radon_stratum[radon_stratum$countyname==county,7] = var(one_county$radon)
    radon_stratum[radon_stratum$countyname==county,9] = var(one_county$radon_log)
  }
}

# radon 均值的点估计和 95% 置信区间
radon_mean = sum(radon_stratum$W_h * radon_stratum$y_h)
radon_se = sqrt(sum(radon_stratum$W_h^2 * (1/radon_stratum$n_h-1/radon_stratum$N_h)*
                    radon_stratum$s2_h))
radon_CI_lb = radon_mean - qnorm(0.975) * radon_se
radon_CI_ub = radon_mean + qnorm(0.975) * radon_se

# radon_log 均值的点估计和 95% 置信区间
radon_log_mean = sum(radon_stratum$W_h * radon_stratum$y_h_log)
radon_log_se = sqrt(sum(radon_stratum$W_h^2 * (1/radon_stratum$n_h-1/radon_stratum$N_h)*
                    radon_stratum$s2_h_log))
radon_log_CI_lb = radon_log_mean - qnorm(0.975) * radon_log_se
radon_log_CI_ub = radon_log_mean + qnorm(0.975) * radon_log_se
```

则 radon 和 log(radon) 的均值点估计和 95% 置信区间的结果如表 1 所示：

表 1：radon 和 log(radon) 的均值点估计和 95% 置信区间

变量名	$\bar{y}_{str}$	$\mathrm{SE}(\bar{y}_{str})$	95% 置信下限	95% 置信上限
radon	4.899	0.154	4.596	5.201
log(radon)	1.301	0.029	1.245	1.358

d. 估计家庭氡水平大于等于 4 pCi/L 的总体比例

各估计值的计算公式如下所示：

- 家庭氡水平大于等于 4 的总体比例 $\hat{p}_{str} = \sum_{h=1}^{H} W_h \hat{p}_h$ ；
- 家庭氡水平大于等于 4 的总体比例的标准误 $\mathrm{SE}(\hat{p}_{str}) = \sqrt{\sum_{h=1}^{H} W_h^2 \frac{1-f_h}{n_h-1} \hat{p}_h(1-\hat{p}_h)}$ ；
- 家庭平均氡水平的 95% 置信区间 $[\hat{p}_{str} - \mathrm{SE}(\hat{p}_{str}), \hat{p}_{str} + \mathrm{SE}(\hat{p}_{str})]$.

对样本总和 $a_h = 0$ or n_h 的层的 $\hat{p}_h$ 进行调整，取 $\hat{p}_h = \frac{a_h+1}{n_h+2}$. 同时，对 $n_h = 1$ 的层（编号为 43、51、84 的县）取 $s_h^2 = 0$ ，得到 $\mathrm{SE}(\hat{p}_{str})$ 的下界。采用上述公式计算 p 的点估计值和 95% 置信区间。

代码如下所示：

```
library(dplyr)
# 求每一层的样本总和
radon_prop = radon %>%
  group_by(countyname) %>%
  summarise(prop=sum(radon>=4))

# 求每一层总体比例的估计，并进行一定调整
radon_stratum$p_h = rep(0, dim(radon_stratum)[1])
for (i in 1:dim(radon_stratum)[1]){
  if (radon_prop$prop[i] < radon_stratum$n_h[1] & radon_prop$prop[i] > 0){
    radon_stratum$p_h[i] = radon_prop$prop[i] / radon_stratum$n_h[i]
  }else{
    radon_stratum$p_h[i] = (radon_prop$prop[i]+1) / (radon_stratum$n_h[i]+2)
  }
}

radon_p_str = sum(radon_stratum$W_h * radon_stratum$p_h) # 总体比例的点估计
radon_stratum1 = radon_stratum[radon_stratum$s2_h != 0,]
radon_p_se = sqrt(sum(radon_stratum1$W_h^2 * (1-1/radon_stratum1$w_hi)/
                        (radon_stratum1$n_h-1) * radon_stratum1$p_h *
                        (1-radon_stratum1$p_h))) # 总体比例的点估计的标准误
radon_p_CI_lb = radon_p_str - 1.96 * radon_p_se # 总体比例点估计的 95% 置信下限
radon_p_CI_ub = radon_p_str + 1.96 * radon_p_se # 总体比例点估计的 95% 置信上限
```

则家庭氡水平大于等于 4 pCi/L 的总体比例的点估计和 95% 置信区间的结果如表 2 所示：

表 2：p 的点估计和 95% 置信区间

$\hat{p}_{str}$	SE($\hat{p}_{str}$)	95% 置信下限	95% 置信上限
0.4939	0.0179	0.4588	0.5290

习题 3.35

a. 按照 team 分层抽样

读入数据并对每一层的总体数进行描述性统计，代码和结果如下所示：

```
baseball_raw = read.csv('baseball.csv', header=FALSE)
baseball = baseball_raw[,c(1,4,5)]; colnames(baseball) = c('team', 'salary', 'POS')
baseball$logsal = log(baseball$salary) # 创建变量 logsal
baseball$pitcher = rep(0, dim(baseball)[1])
baseball$pitcher[baseball$POS=='P'] = 1 # 创建 0-1 变量 pitcher

baseball_strata = baseball %>%
  group_by(team) %>% summarise(N_h=n()) # 每一层的总体数
summary(baseball_strata$N_h) # 每一层总体数的描述性统计
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  24.00   26.00   26.50   26.57   27.00   29.00
```

要按比例分层抽样，理论上每一层的抽样比应相等，则每一层的抽样比为 $f_h = n/N = 150/797 \approx 0.188$ 。由每一层总体数的描述性统计可知，每一层总体数的均值为 26.57 ，且每一层总体数较为接近，所以可考虑对每一层抽取相同的样本量。故每一层的样本量为 $n_h = N_h f_h \approx 26.57 \times 0.188 \approx 5$ 。

分层抽样的代码如下所示：

```
library(sampling)
baseball_StS = strata(baseball, stratanames='team', size=rep(5,dim(baseball_strata)[1]),
                      method='srswor') # 抽取的样本保存在 baseball_StS 中
baseball_StS$logsal = baseball[baseball_StS$ID_unit, 4] # 样本的 logsal
baseball_StS$pitcher = baseball[baseball_StS$ID_unit, 5] # 样本的 pitcher
write.csv(baseball_StS, file='baseball_StS.csv', row.names=FALSE, fileEncoding='utf8')
```

b. 估计 logsal 的总体均值和 95% 置信区间

各估计值的计算公式如下所示：

- logsal 总体均值的点估计 $\bar{y}_{str} = \sum_{h=1}^{H} W_h \bar{y}_h$ ；

- `logsal` 总体均值的标准误 $\mathrm{SE}(\bar{y}_{str}) = \sqrt{\sum_{h=1}^{H} W_h^2 \frac{1-f_h}{n_h} s_h^2}$ ；
- `logsal` 总体均值的 95% 置信区间 $[\bar{y}_{str} - \mathrm{SE}(\bar{y}_{str}), \bar{y}_{str} + \mathrm{SE}(\bar{y}_{str})]$.

采用上述公式计算 $\bar{y}_U$ 的点估计值和 95% 置信区间，代码如下所示：

```
by_team = baseball_StS %>% group_by(team)
baseball_byTeam1 = by_team %>%
  summarise(logsal_mean=mean(logsal), logsal_s2=var(logsal))

baseball_strata$n_h = rep(5, dim(baseball_strata)[1]) # 每一层的样本容量
baseball_strata$f_h = baseball_strata$n_h / baseball_strata$N_h # 每一层的抽样比
baseball_strata$W_h = baseball_strata$N_h / sum(baseball_strata$N_h) # 每一层的层权
baseball_strata$logsal_mean = baseball_byTeam1$logsal_mean
baseball_strata$logsal_s2 = baseball_byTeam1$logsal_s2

# logsal 总体均值的点估计和 95% 置信区间
logsal_mean = sum(baseball_strata$W_h * baseball_strata$logsal_mean)
logsal_se = sqrt(sum(baseball_strata$W_h^2 * (1-baseball_strata$f_h)/
                       baseball_strata$n_h * baseball_strata$logsal_s2))
logsal_CI_lb = logsal_mean - qnorm(0.975) * logsal_se
logsal_CI_ub = logsal_mean + qnorm(0.975) * logsal_se
```

则 `logsal` 总体均值的点估计和 95% 置信区间的结果如表 3 所示：

表 3：logsal 总体均值的点估计和 95% 置信区间

变量名	$\bar{y}_{str}$	$\mathrm{SE}(\bar{y}_{str})$	95% 置信下限	95% 置信上限
logsal	14.023	0.086	13.853	14.192

c. 估计 `pitcher` 的总体比例和 95% 置信区间

各估计值的计算公式如下所示：

- `pitcher` 总体比例的点估计 $\hat{p}_{str} = \sum_{h=1}^{H} W_h \hat{p}_h$ ；
- `pitcher` 总体比例点估计的标准误 $\mathrm{SE}(\hat{p}_{str}) = \sqrt{\sum_{h=1}^{H} W_h^2 \frac{1-f_h}{n_h - 1} \hat{p}_h(1-\hat{p}_h)}$ ；
- `pitcher` 总体比例的 95% 置信区间 $[\hat{p}_{str} - \mathrm{SE}(\hat{p}_{str}), \hat{p}_{str} + \mathrm{SE}(\hat{p}_{str})]$.

采用上述公式计算 p 的点估计值和 95% 置信区间，代码如下所示：

```
baseball_byTeam2 = by_team %>%
  summarise(a_h=sum(pitcher))
baseball_strata$pitcher_p_h = baseball_byTeam2$a_h / baseball_strata$n_h

# pitcher 总体比例的点估计和 95% 置信区间
```

```
pitcher_mean = sum(baseball_strata$W_h * baseball_strata$pitcher_p_h)
pitcher_se = sqrt(sum(baseball_strata$W_h^2 * (1-baseball_strata$f_h)/
                        (baseball_strata$n_h-1) * baseball_strata$pitcher_p_h *
                        (1-baseball_strata$pitcher_p_h)))
pitcher_CI_lb = pitcher_mean - qnorm(0.975) * pitcher_se
pitcher_CI_ub = pitcher_mean + qnorm(0.975) * pitcher_se
```

则 `pitcher` 总体比例的点估计和 95% 置信区间的结果如表 4 所示：

表 4：pitcher 总体比例的点估计和 95% 置信区间

变量名	$\hat{p}_{str}$	SE($\hat{p}_{str}$)	95% 置信下限	95% 置信上限
pitcher	0.487	0.036	0.417	0.557

d. SRS 与 StS 估计值的比较

习题 2.32 中 SRS 的估计值与本题中 StS 的估计值如表 5 所示：

表 5：SRS 与 StS 估计值的比较

变量名	抽样方法	均值	标准误	95% 置信下限	95% 置信上限
logsal	SRS	13.950	0.090	13.774	14.126
	StS	14.023	0.086	13.853	14.192
pitcher	SRS	0.440	0.037	0.368	0.512
	StS	0.487	0.036	0.417	0.557

由表 5 可知，使用分层样本对变量 `logsal` 和 `pitcher` 进行估计的标准误均比使用简单随机样本进行估计的标准误要略小。所以，可以认为使用分层样本的估计精度更高。