习题 2.36

首先，读入数据。

```
forest = read.csv('forest.csv', header=FALSE)
colnames(forest) = c('elevation', 'Aspect', 'Slope', 'Horiz', 'Vert', 'HorizRoad',
                     'Hillshade_9am', 'Hillshade_Noon', 'Hillshade_3pm', 'HorizFire',
                     'Wilderness1', 'Wilderness2', 'Wilderness3', 'Wilderness4', 'Cover')
```

1. 抽取样本容量为 2000 的 SRS

```
set.seed(0); set.seed(12345)
n = 2000; N = dim(forest)[1]
# 定义函数 get_SRS(n, N) 来从总体中得到一个 SRS
get_SRS = function(n, N){
  S = c(); k = 0 # 初始化变量
  while (k < n){
    x = runif(1)
    R = floor(N*x) + 1
    if (! R %in% S){
      S = append(S, R)
      k = k + 1
    }
  }
  return(S)
}
sampleNum = get_SRS(n, N) # 变量 SampleNum 储存了抽取的 SRS 的样本编号
sample_forest = forest[sampleNum,]
```

2. 估计七类 Cover 的总体比例和 95% 置信区间

- 总体比例估计的计算公式为

$$\hat{p} = \frac{1}{n}\sum_{i\in\mathcal{S}} y_i$$

- 总体比例标准误的计算公式为

$$\mathrm{SE}(\hat{p}) = \sqrt{\frac{1-f}{n-1}\hat{p}(1-\hat{p})}$$

- 总体比例 95% 置信区间的计算公式为

$$[\hat{p} - z_{\alpha/2}\mathrm{SE}(\hat{p}), \hat{p} + z_{\alpha/2}\mathrm{SE}(\hat{p})]$$

```
# 初始化各变量
cover_typeNum = c(); mean_cover = c(); se_cover = c(); CI_lb_cover = c(); CI_ub_cover = c()
# 对每一类 Cover, 计算其估计量
for (i in 1:7){
```

```
  cover_typeNum_i = sum(sample_forest$Cover == i) # 第 i 类 Cover 的数量
  mean_cover_i = cover_typeNum_i / n # 第 i 类 Cover 的均值
  se_cover_i = sqrt((1-n/N)/(n-1) * mean_cover_i * (1-mean_cover_i)) # 第 i 类 Cover 的标准误
  CI_lb_cover_i = mean_cover_i - qnorm(0.975) * se_cover_i # CI lower bound
  CI_ub_cover_i = mean_cover_i + qnorm(0.975) * se_cover_i # CI upper bound

  cover_typeNum = append(cover_typeNum, cover_typeNum_i)
  mean_cover = append(mean_cover, mean_cover_i)
  se_cover = append(se_cover, se_cover_i)
  CI_lb_cover = append(CI_lb_cover, CI_lb_cover_i)
  CI_ub_cover = append(CI_ub_cover, CI_ub_cover_i)
}
# 输出结果
cover_estimate_df = data.frame(Type=1:7, TypeNum=cover_typeNum, p_hat=mean_cover,
                               SE=se_cover, CI_lb=CI_lb_cover, CI_ub=CI_ub_cover)
```

则七类 Cover 的估计量如表 1 所示：

表 1：七类 Cover 的估计量

Cover 类型	Cover 样本量	$\hat{p}$	SE($\hat{p}$)	95% 置信下限	95% 置信上限
1	769	0.3845	0.0109	0.3632	0.4058
2	920	0.4600	0.0111	0.4382	0.4818
3	143	0.0715	0.0057	0.0602	0.0828
4	11	0.0055	0.0017	0.0023	0.0087
5	27	0.0135	0.0026	0.0084	0.0186
6	58	0.0290	0.0037	0.0217	0.0363
7	72	0.0360	0.0042	0.0278	0.0442

3. 估计 elevation 的均值和 95% 置信区间

- 样本均值的计算公式为

$$\bar{y} = \frac{1}{n}\sum_{i\in\mathcal{S}} y_i$$

- 样本均值标准误的计算公式为

$$\mathrm{SE}(\bar{y}) = \sqrt{\frac{1-f}{n}s^2}$$

- 总体均值 95% 置信区间的计算公式为

$$[\bar{y} - z_{\alpha/2}\mathrm{SE}(\bar{y}), \bar{y} + z_{\alpha/2}\mathrm{SE}(\bar{y})]$$

```
mean_elevation = mean(sample_forest$elevation) # elevation 的样本平均
s2_elevation = sum((sample_forest$elevation-mean_elevation)^2)/(n-1) # elevation 的样本方差
se_elevation = sqrt((1-n/N)/n * s2_elevation) # elevation 样本平均的标准误
CI_lb_elevation = mean_elevation - qnorm(0.975) * se_elevation # CI lower bound
CI_ub_elevation = mean_elevation + qnorm(0.975) * se_elevation # CI upper bound
```

则 elevation 均值的估计及其 95% 置信区间如表 2 所示：

表 2：elevation 的均值估计及 95% 置信区间

elevation 样本量	$\bar{y}$	SE($\bar{y}$)	95% 置信下限	95% 置信上限
2000	2958.5895	6.2973	2946.2470	2970.9320

4. 估算样本量：估计变量 HorizFire 的总体均值

4.1 抽取容量为 100 的试点样本来获得估算样本量所需的总体信息

估算样本量时需要知道 HorizFire 的总体均值和标准差，故通过抽取试点样本来估计总体均值和总体标准差：

```
set.seed(0); set.seed(123)
pilot_sampleNum = get_SRS(100, N) # 试点样本的编号
pilot_sample = forest[pilot_sampleNum,] # 容量为 100 的试点样本
mean_HorizFire = mean(pilot_sample$HorizFire) # HorizFire 的样本均值
s2_HorizFire = sum((pilot_sample$HorizFire-mean_HorizFire)^2)/(100-1) # HorizFire 的样本方差
s_HorizFire = sqrt(s2_HorizFire) # HorizFire 的样本标准差
```

由上述计算可知，HorizFire 的总体均值估计值为 1960.18，总体标准差估计值为 1294.2220。

4.2 估算 HorizFire 所需的样本量

要求以 98% 的置信水平保证估计的最大相对误差不超过 5%，则

- 忽略 fpc 时，所需的样本容量为

$$n_0 = \frac{z_{\alpha/2}^2}{r^2}\left(\frac{S}{\bar{y}_u}\right)^2$$

- 考虑 fpc 时，所需的样本容量为

$$n = \frac{n_0}{1+\frac{n_0}{N}}$$

```
r = 0.05; alpha = 0.02
n0 = (qnorm(1-alpha/2) / r)^2 * (s_HorizFire / mean_HorizFire)^2 # 忽略 fpc 时的样本容量
```

由上述计算得 $n_0 = 943.7 \approx 944$，则 $n_0 \ll N$，故所需的样本容量为 944。

5. 估计 CoverType 的总体比例

要求以 96% 的置信水平保证估计的最大相对误差不超过 10%，则

- 忽略 fpc 时，所需的样本容量为

$$n_0 = \frac{z_{\alpha/2}^2}{r^2 p}(1-p)$$

- 考虑 fpc 时，所需的样本容量为

$$n = \frac{n_0}{1 + \frac{n_0}{N}}$$

可知，n_0 为 p 的函数，且关于 p 单调递减。则对于估算 CoverType=1、CoverType>3 的两类格子的总体比例，分别取 $p = 30\%$, $p = 5\%$，可达到 n_0 的上界。

```
r_cover = 0.1; alpha_cover = 0.04
p_cover1 = 0.3; p_cover3 = 0.05
# 估算 CoverType=1 的总体比例所需的样本量
n0_cover1 = (qnorm(1-alpha_cover/2)/r_cover)^2 / p_cover1 * (1-p_cover1)
# 估算 CoverType>3 的总体比例所需的样本量
n0_cover3 = (qnorm(1-alpha_cover/2)/r_cover)^2 / p_cover3 * (1-p_cover3)
```

由上述计算得，忽略 fpc 时，

- 对于估算 CoverType=1 的总体比例，所需的样本容量为 $n_0 = 984$;
- 对于估算 CoverType>3 的总体比例，所需的样本容量为 $n_0 = 8014$.

为了同时满足估计精度，取样本容量为 8014。