L08-抽样方法及 R 语言实现

郑盼盼

2024-11-20

目录

8.1 放回简单随机抽样

放回简单随机抽样方法：

1. 将 Ω 中的样本点编号为 $1, 2, \cdots, N$
2. 用取后放回的方法从编号中依次抽取数 $i_1, i_2, \ldots, i_n$ （样本量为 n）
3. 对于 $1 \leq k \leq n$，测量居民 ω_{ik} 的变量值 $X_k = X(\omega_{ik})$

一般地，称总体变量 X 的观测结果 $X_1, \ldots, X_n$ 为样本，称 n 为**样本容量**。

```
sample(1:10,10,replace=T)
```

```
##  [1]  7  1  6  8  7  9  7  3 10 10
```

8.2 简单随机抽样

简单随机抽样：依次用抽后不放回的方法从总体中抽取样本 $Y_1, \ldots, Y_n$，称这种方法为**简单随机抽样**，相应的样本为**简单随机样本**。

例 限定容量为 100，分别模拟简单放回抽样的样本均值和不放回抽样的样本均值估计总体均值，考察两种方式的估计效果

```
N = 1000 # 总体中个体的数目
n = 100  # 样本量
X = sample(1:N, n,T) # 放回随机抽样
Xbar = mean(X)  # 计算样本均值
Xbar
```

```
## [1] 510.51
```

```
Y = sample(1:N, n, F) # 简单随机抽样（不放回）
Ybar = mean(Y)
Ybar
```

```
## [1] 489.71
```

8.2.1 有放回的简单随机抽样和简单随机抽样对均值的估计

```
N <- 1000           # 总体的个数
trueMean <- mean(1:N) # 计算总体的期望
n <- 100; m <- 10000 # 样本量为 n，重复估计次数 m
# 定义向量 Xmean, Ymean, XD, YD 分别用于存储 m 次估计中，每次得到的样本均值以及与总
↪ 体均值的离差平方
Xmean <- Ymean <- XD <- YD <- 1:m
# 使用 for 循环进行 m 次估计
for (k in 1:m) {
  X <- sample(1:N, n, T)  # X 表示有放回抽取的样本
  Xmean[k] <- mean(X)     # 计算 X 的均值，作为 Xmean 的第 k 个元素
  XD[k] <- (trueMean - mean(X))^2 # 计算 X 的均值和总体均值的离差平方
  Y <- sample(1:N, n, F)  # Y 表示简单随机抽样得到的样本
  Ymean[k] <- mean(Y)     # 计算 Y 的均值，作为 Ymean 的第 k 个元素
  YD[k] <- (trueMean - mean(Y))^2 # 计算 Y 的均值和总体均值的离差平方
```

```
}

# 计算 m 次估计的平均估计值以及平均离差平方
results <- c(mean(Xmean), mean(Ymean), mean(XD), mean(YD))
results <- matrix(results, 2,2) # 将向量转换为矩阵
colnames(results) <- c(" 估计均值"," 估计均方误差") # 设置矩阵的列名
rownames(results) <- c(" 放回抽样", " 不放回抽样")  # 设置矩阵的行名
results
```

```
##              估计均值 估计均方误差
## 放回抽样    501.1016     826.3192
## 不放回抽样 500.4923     748.5452
```

8.3 等距抽样

可将总体中的所有个体排序为 $\omega_1, \dots, \omega_N$，先选定的序范围内随机抽取一个个体 $\omega_{i,1}$，然后在序列中按等间隔的原则抽取其他个体，就可以得到样本

$$X(\omega_{i,1}), \dots, X(\omega_{i,n})$$

称这种抽样方法为**等距随机抽样**，简称为**等距抽样**，相应的样本为**等距样本**。

R 语言实现

基本思路 假如我们有一个长度为 N 的总体，希望抽取 n 个样本：

1. 确定步长 $k = \lfloor N/n \rfloor$
2. 计算初始样本点的上限 $N-(n-1)K$，并在区间 $[1, N-(n-1)k]$ 之间随机选取一点作为初始位置 r
3. 依次选取位置 $r, r+k, r+2k, \dots$

例题 总体中所有个体排序为 $\omega_1, \omega_2, \dots, \omega_{401}$，总体变量 X 在个体 ω_i 处的值

$$X(\omega_i) = \sin\left(\frac{i\pi}{10}\right), \quad 1 \le i \le 401, \tag{8.1}$$

尝试使用等距抽样的方法，从中抽取出 20 个样本。

```
N <- 401 # 总体中个体总数
x <- sin((1:N) * pi / 10) # 总体数据
n <- 20 # 样本容量
```

```
pMean <- mean(x)  # 总体均值

# 等距样本抽取
groupN <- N/n  # 步长
groupN <- floor(groupN) # 向下取整，得到步长
tmpRange <- N - (n-1) * groupN # 设置第一个样本点的次序上界
tmpFirst <- sample(1:tmpRange, 1) # 抽取第一的样本点的次序
tmpIndx <- tmpFirst + (0:(n-1)) * groupN # 根据第一个样本点和步长，计算后面样本点的次
 ↪ 序
tmpS <- x[tmpIndx]
sMean <- mean(tmpS)
cat(" 样本均值和总体均值的距离", abs(pMean-sMean))
```

```
## 样本均值和总体均值的距离 0.9518271
```

8.4 分层抽样

将所关心的总体分为若干个子总体，使得各个子总体的变量值的平均值明显不同，称这样的一个子总体为**层**。在每层中按确定的样本容量 n_i 抽取简单随机样本，再把各层抽出的样本合在一起作为样本，这种抽样方法为**分层随机抽样**，简称**分层抽样**，相应的样本称为**分层样本**。

例 考虑到变量总体 (8.1)，将总体 $\Omega = \{\omega_1, \omega_2, \dots, \omega_{401}\}$ 分 “层” 为

$$\Omega_1 = \{\omega_1, \omega_2, \dots, \omega_{21}\}$$
$$\Omega_k = \{\omega_{2+20(k-1)}, \omega_{3+20(k-1)}, \cdots, \omega_{1+20k}\}$$

其中 $k = 2, 3, \dots, 20$。下列代码给出容量为 $n = 20$ 的分层抽样代码。

```
tmpIndex <- c() # 初始化一个空向量 tmpIndx 用于存储抽取出样本点的序号
# 使用 for 循环遍历 1:20
for(k in 1:20){
  # 若 k=1, 则定义" 层"subOmega, 表示 Omeaga_1
  if(k==1){
    subOmega <- 1:21
  }else{
    # 若 k 不为 1, 定义" 层" subOmega, 表示 Omega_k
    tmp <- k-1
    subOmega <- (2+20*tmp):(1+20*k)
```

```
  }
  # 从" 层" 中随机抽取一个样本放入 tmpIndex
  tmpIndex <- c(tmpIndex,
                sample(subOmega,1))
}

tmpS <- x[tmpIndex] # 从序号中得到样本 tmpS
sMean <- mean(tmpS) # 计算样本均值
cat(" 样本均值和总体均值的距离", abs(pMean - sMean))
```

```
## 样本均值和总体均值的距离 0.1493774
```

8.5 整群抽样

通常可依据辅助信息将总体分割成若干个子总体，使得各个子总体变量的均值相差无几，称这样的一个子总体为**群**。在所有群中用简单随机抽样抽取一些群，将抽出群中的所有个体变量合在一起作为样本。称这种抽样方法为**整群随机抽样**，简称**整群抽样**，相应的样本称为**整群样本**。

例 考虑到变量总体 (8.1)，将总体 $\Omega = \{\omega_1, \omega_2, \dots, \omega_{401}\}$ 分 “群” 为

$$\Omega_1 = \{\omega_1, \omega_2, \dots, \omega_{21}\}$$
$$\Omega_k = \{\omega_{2+20(k-1)}, \omega_{3+20(k-1)}, \dots, \omega_{1+20k}\}$$

其中 $k = 2, 3, \dots, 20$。如下的代码程序给出了容量为 $n = 20$ 或 $n = 21$ 的整群抽样程序。

```
# 从 20 个群中随机抽取一个
k <- sample(1:20,1)
# 若抽到第 1 个群
if(k==1){
  tmpIndex <- 1:21 # 返回 Omega_1
} else
{
  tmp <- k-1
  tmpIndex <- (2+20*tmp):(1+20*k) # 返回 Omega_k
}
tmpS <- x[tmpIndex] # 获得样本
sMean <- mean(tmpS) # 计算样本均值
cat(" 样本均值和总体均值的距离", abs(pMean - sMean)) # 计算误差
```

```
## 样本均值和总体均值的距离 0.0007706159
```

Questions

考察总体 $\Omega = \{\omega_1, \omega_2, \dots, \omega_{1000}\}$，以及变量

$$X(\omega_k) = k^2, 1 \leq k \leq 1000,$$

要用样本均值估计总体均值。

1. 在样本容量为 500 的情况下，用随机模拟方法比较放回简单随机样本和简单随机样本的估计效果。
2. 在样本容量为 500 的情况下，用随机模拟方法比较简单随机样本和等距抽样样本（按总体中个体下标排序）的估计效果，并解释引起这种结果的原因。
3. 将总体分层为

   $$\Omega_k = \{\omega_{100(k-1)+1}, \dots, \omega_{100k}\}, 1 \leq k \leq 5.$$

   在样本容量为 500 的情况下，用随机模拟方法比较简单随机样本和分层抽样样本的估计效果，并解释引起这种结果的原因。
4. 将总体分层为

   $$\Omega_k = \{\omega_{100(k-1)+1}, \dots, \omega_{100k}\}, 1 \leq k \leq 5.$$

   在样本容量为 100 的情况下，用随机模拟方法比较简单随机样本和分层抽样样本的估计效果，并解释引起这种结果的原因。