统计计算 张楠 2019秋: Monte Carlo Methods for Estimation

(返回 统计计算 张楠 2019秋)

Monte Carlo Methods for Estimation

假设$X_1,\cdots,X_n$为从总体X中抽取的随机样本, 参数θ的估计量记为$\hat\theta=\hat\theta(X_1,\cdots,X_n)$ 记$x=(x_1,\cdots,x_n)^T\in\mathcal{R}^n$, 以及$x^{(1)},x^{(2)},\cdots$为从总体$X$中抽取的一列独立的随机样本观测值. 则 有关于$\hat\theta$的性质, 可以通过估计值序列$\hat\theta(x_1^{(j)},\cdots,x_n^{(j)})$, $j=1,2,\cdots$ 来研究.

目录

Monte Carlo Estimation and Standard Error

> **例 1: 假设$X_1,X_2\ i.i.d\ \sim N(0,1)$, 估计$E|X_1-X_2|$.**

显然, $\theta=E|X_1-X_2|$ 的Monte Carlo估计可用通过从标准正态分布中 产生m个样本 $x^{(j)}=(x_1^{(j)},x_2^{(j)})$, $j=1,\cdots,m$. 然后计算$\hat\theta^{(j)}=|x_1^{(j)}-x_2^{(j)}|, j=1,\cdots,m$. 以及$\theta$的估计

$$\hat\theta=\frac{1}{m}\sum_{j=1}^m\hat\theta^{(j)}=\frac{1}{m}\sum_{j=1}^m|x_1^{(j)}-x_2^{(j)}|.$$

在R中很容易实现:

```
m <- 1000
g <- numeric(m)
for (i in 1:m) {
    x <- rnorm(2)
    g[i] <- abs(x[1] - x[2])
}
est <- mean(g)
est
```

我们也可以计算出$E|X_1-X_2|=2/\sqrt{\pi}\doteq 1.128379$ 以及方差$Var(|X_1-X_2|)=2-4/\pi$. 因此$E\hat\theta=2/\sqrt{\pi}, Var(\hat\theta)=[2-4/\pi]/m$.

对标准差的Monte Carlo 估计我们可以从一般场合出发讨论. 由于样本量为n的样本均值$\bar X$ 的标准差为$\sqrt{Var(X)/m}$, 当X的分布未知时, 可以使用“Plug-in”法估计: 由

$$\hat{Var}(X)=\frac{1}{m}\sum_{i=1}^m(x_i-\bar x)^2.$$

因此$\bar X$标准差的估计为

$$\hat{se}(\bar x)=\frac{1}{\sqrt m}\left[\frac{1}{m}(x_i-\bar x)^2\right]^{1/2}=\frac{1}{m}\left[(x_i-\bar x)^2\right]^{1/2}.$$

或者也可以使用无偏估计量

$$\hat{se}(\bar{x}) = \frac{1}{\sqrt{m}}\left[\frac{1}{m-1}(x_i - \bar{x})^2\right]^{1/2}.$$

因此, 前例中的标准差估计为

$$\hat{se}(\hat{\theta}) = \frac{1}{m}\left[\sum_{j=1}^m (\hat{\theta}^{(j)} - \hat{\theta})^2\right].$$

计算程序如下

```
sqrt(sum((g-mean(g))^2))/m
#sd(g)/sqrt(m)  #for unbiased estimator
```

Estimation of MSE

Monte Carlo 方法可以用于计算一个估计量的MSE. 一个估计量的MSE定义为$MSE(\hat{\theta}) = E[\hat{\theta} - \theta]^2$. 如果从总体$X$中产生了$m$个样本$x^{(1)}, \cdots, x^{(m)}$, 则$\hat{\theta}$的MSE的Monte Carlo估计为

$$M\hat{S}E = \frac{1}{m}\sum_{j=1}^m (\hat{\theta}^{(j)} - \theta)^2,$$

其中$\hat{\theta}^{(j)} = \hat{\theta}(x^{(j)})$.

> **例 2**: 使用Monte Carlo方法估计标准正态分布的截尾均值$\bar{X}_{[-1]}$的MSE.
>
> 当样本存在异常点时, 截尾的样本均值常常被用来估计总体的中心. 假设$X_1, \cdots, X_n$为一个随机样本, $X_{(1)}, \cdots, X_{(n)}$为相应的次序统计量, 则 一个k水平的截尾样本均值为 $\bar{X}_{[-k]} = \frac{1}{n-2k}\sum_{i=k+1}^{n-k} X_{(i)}$.

本例中, 目标参数为$\theta = E\bar{X} = E\bar{X}_{[-1]} = 0$. 记$T = \bar{X}_{[-1]}$, 则其MSE的一个 Monte Carlo 估计算法如下

1. 通过如下步骤产生m个重复$T^{(j)}, j = 1, \cdots, m$:
   - 产生总体X的样本: $x_1^{(j)}, \cdots, x_n^{(j)}$.
   - 从小到大排序 $x_{(1)}^{(j)} \leq \cdots \leq x_{(n)}^{(j)}$
   - 计算 $T^{(j)} = \frac{1}{n-2}\sum_{i=2}^{n-1} x_{(i)}^{(j)}$.
2. 计算MSE $M\hat{S}E(T) = \frac{1}{m}\sum_{j=1}^m (T^{(j)} - \theta)^2 = \frac{1}{m}\sum_{j=1}^m (T^{(j)})^2$.

实现程序如下

```
n <- 20
m <- 1000
tmean <- numeric(m)
for (i in 1:m) {
    x <- sort(rnorm(n))
    tmean[i] <- sum(x[2:(n-1)]) / (n-2)
    }
mse <- mean(tmean^2)
mse
sqrt(sum((tmean - mean(tmean))^2)) / m     #se
```

截尾均值的MSE的估计为$0.0504531(\hat{se} \doteq 0.007)$. 样本均值的 MSE为$Var(X)/n = 1/20 = 0.05$. 另一方面, 中位数本质上也是一种截尾均值, 其截掉了除中间的一个或两个点外的所有点. 样本中位数的MSE估计如下

```
n <- 20
m <- 1000
tmean <- numeric(m)
for (i in 1:m) {
    x <- sort(rnorm(n))
    tmean[i] <- median(x)
    }
mse <- mean(tmean^2)
mse
sqrt(sum((tmean - mean(tmean))^2)) / m     #se
```

从而样本中位数的MSE估计为$0.075(\hat{se} = 0.0086)$.

> **例 3:** 比较标准正态分布与如下混合(``污染")正态分布下的k水平截尾均值估计的MSE. $pN(0, \sigma^2 = 1) + (1-p)N(0, \sigma^2 = 100)$

我们写一个函数来对不同的k和p计算截尾均值$\bar{X}_{[-k]}$的MSE. 从混合正态中产生样本时, 要根据$P(\sigma = 1) = p; P(\sigma = 10) = 1 - p$来随机选择$\sigma$. 注意正态随机数产生函数 **rnorm**可以使用参数向量作为标准偏差. 考虑$p = 1.0, 0.95, 0.9$ 以及$k = 0, 1, \cdots, n/2$. 因此程序如下

```
 n <- 20; K <- n/2 - 1;  m <- 1000;
 mse <- matrix(0, n/2, 6)
 trimmed.mse <- function(n, m, k, p) {
     #MC est of mse for k-level trimmed mean of
     #contaminated normal pN(0,1) + (1-p)N(0,100)
     tmean <- numeric(m)
     for (i in 1:m) {
         sigma <- sample(c(1, 10), size = n,
             replace = TRUE, prob = c(p, 1-p))
         x <- sort(rnorm(n, 0, sigma))
         tmean[i] <- sum(x[(k+1):(n-k)]) / (n-2*k)
         }
     mse.est <- mean(tmean^2)
     se.mse <- sqrt(mean((tmean-mean(tmean))^2)) / sqrt(m)
     return(c(mse.est, se.mse))
 }
for (k in 0:K) {
    mse[k+1, 1:2] <- trimmed.mse(n=n, m=m, k=k, p=1.0)
    mse[k+1, 3:4] <- trimmed.mse(n=n, m=m, k=k, p=.95)
    mse[k+1, 5:6] <- trimmed.mse(n=n, m=m, k=k, p=.9)
}
round(n*mse,3)
```

结果表明, 均值的稳健估计(截尾均值估计)在总体分布被污染时能降低 MSE.

Estimating a confidence level

在应用统计中经常遇到的一个问题是估计总体的分布. 比如, 许多常用的统计推断 方法和工具都是基于正态性假设下的. 而在实际中, 总体分布非正态是经常的, 估计 量的分布可能不知道或者没有显示表示, 此时, Monte Carlo 方法则可以用来进行统计推断.

假设(U, V)是未知参数θ的置信区间, 则统计量U, V的分布都依赖于抽样 分布X的分布F_X. 置信水平就是区间(U, V)能够覆盖θ真值的概率. 因此 估计置信水平就是一个积分估计问题.

比如考虑方差的置信区间估计问题. 标准的方法是对正态性假设很敏感的, 在数据(样本) 偏离正态分布时, 我们来使用Monte Carlo方法估计真实的置信水平. 首先看正态假定 下方差的置信区间估计(标准方法):

例 4: 方差的置信区间 假设$X_1, \cdots, X_n,\ i.i.d \sim N(\mu, \sigma^2)$, $n \geq 2$, S^2为样本方差, 则由$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$ 知道σ^2的一个$100(1-\alpha)$置信上界为 $(0, (n-1)S^2/\chi^2_\alpha(n-1))$ 从而从正态总体$N(0, \sigma^2 = 4)$中随机抽取$n = 20$个样本, 则计算σ^2的$\%95$置信上界的 程序如下

```
n <- 20
alpha <- .05
x <- rnorm(n, mean=0, sd=2)
UCL <- (n-1) * var(x) / qchisq(alpha, df=n-1)
```

我们可以使用图来直观上判断经验的置信水平:

```
m<-100000
ucls<-numeric(m)
for(i in 1:m){
x <- rnorm(n, mean=0, sd=2)
ucls[i] <- (n-1) * var(x) / qchisq(alpha, df=n-1)
}
ind<-ucls>4
cov.rate<-cumsum(ind)/1:m
plot(2:m,cov.rate[-1],type=''l'')
abline(h=0.95)
```

经验的置信水平是通过模拟, 对理论的置信水平进行估计. 其一般做法如下

假设$X \sim F_X$, 感兴趣的参数为θ. 则对$j = 1, \cdots, m$:

1. 产生第j个随机样本$x_1^{(j)}, \cdots, x_n^{(j)}$.
2. 计算基于第j个样本的置信区间C_j
3. 计算$y_j = I(\theta \in C_j)$
4. 计算经验的置信水平$\bar{y} = \frac{1}{m}\sum_{j=1}^m y_j$.

例 5: 置信水平的Monte Carlo估计 在上一个例子中, 我们使用了**for** 循环来实现计算σ^2的m个置信上界. 我们也可以使用 **replicate**函数:

```
n <- 20
alpha <- .05
UCL <- replicate(1000, expr = {
    x <- rnorm(n, mean = 0, sd = 2)
    (n-1) * var(x) / qchisq(alpha, df = n-1)
    } )
#计算包含sigma^2=4的区间数
sum(UCL > 4)
#计算经验的覆盖率(置信水平)
mean(UCL > 4)
```

replicate函数要重复执行的代码放在**{}**中, 参数**expr**可以调用一个函数:

```
calCI <- function(n,alpha){
    x <- rnorm(n, mean = 0, sd = 2)
    return((n-1) * var(x) / qchisq(alpha, df = n-1))
    }
UCL<-replicate(1000,expr=calCI(n=20,alpha=.05))
mean(UCL>4)
```

以上所说的置信区间构造方法是建立在正态性假设之上的, 如果数据(样本)不服从正态分布, 则 真正的置信水平为

$$P(\frac{(n-1)S^2}{\chi^2_\alpha(n-1)} > \sigma^2) = P(S^2 > \frac{\sigma^2\chi^2_\alpha(n-1)}{n-1}) = 1 - G(\frac{\sigma^2\chi^2_\alpha(n-1)}{n-1}).$$

这里G为统计量S^2的分布. 因此估计置信水平等价于要估计$G(t) = P(S^2 < t) = \int_0^t g(x)dx$, 从而Monte Carlo积分估计方法可以被用来估计此积分.

例 6: 经验的置信水平 在例5中, 假设样本服从χ^2_2, 则经验的置信水平是多少? 此时 方差仍然是4.

```
n <- 20; alpha <- .05
UCL <- replicate(1000, expr = {
    x <- rchisq(n, df = 2)   # 从chi^2(2)中抽样
    (n-1) * var(x) / qchisq(alpha, df = n-1) } )
sum(UCL > 4)
mean(UCL > 4)
```

前面的例子中考虑的问题都是在总体分布已知的情形下, 对参数进行Monte Carlo估计. 因此这种 情况下的Monte Carlo方法也称为是参数**Bootstrop** 方法. 有关于**Bootstrap** 方法我们将在后面学习.
