统计计算 张楠 2019秋: Bootstrap Confidence Intervals

(返回 统计计算 张楠 2019秋)

本节中我们介绍几种在Bootstrap中构造目标参数的渐近置信区间的方法, 其中包括 标准正态Bootstrap置信区间, 基本的Bootstrap置信区间, Bootstrap百分位数(percentile)置信区间和 Bootstrap t 置信区间.

目录

The Standard Normal Bootstrap Confidence Interval
The Percentile Bootstrap Confidence Interval
The Basic Bootstrap Confidence Interval
The Bootstrap t interval

The Standard Normal Bootstrap Confidence Interval

标准正态Bootstrap置信区间是一种比较简单的方法. 假设 $\hat{\theta}$ 是参数 θ 的估计量, 以及估计量的标准差为$se(\hat{\theta})$. 若 $\hat{\theta}$ 渐近到正态分布, 即 $Z=\frac{\hat{\theta}-E\hat{\theta}}{se(\hat{\theta})}$ 渐近服从标准正态分布. 则若$\hat{\theta}$ 为θ的无偏估计, 那么θ的一个渐近的$100(1-\alpha)\%$ 标准正态 Bootstrap 置信区间为 $\hat{\theta}\pm z_{\alpha/2}\hat{se}_B(\hat{\theta})$, 其中 $z_{\alpha/2}=\Phi^{-1}(1-\alpha/2)$. 此区间容易计算, 但是有正态性假设或者CLT需成立. 以及 $\hat{\theta}$ 为θ的无偏估计.

The Percentile Bootstrap Confidence Interval

由形式 $P(L\le\hat{\theta}\le U)=1-\alpha$ 知, 可以使用Bootstrap重复的样本百分位数来估计L和U. 而$\hat{\theta}$ 为θ的估计, 因此就取θ的$1-\alpha$置信区间上下界分别为Bootstrap重复的样本$1-\alpha/2$百分位数$\hat{\theta}^*_{[(B+1)(1-\alpha/2)]}$ 和$\alpha/2$百分位数$\hat{\theta}^*_{[(B+1)\alpha/2]}$.

Efron & Tibshirani 证明了百分位数区间相比于标准正态区间有着更好的理论覆盖率. 下面我们还会介绍 *bias-corrected and accelerated*(BCa) 百分位数区间, 它是百分位数区间的一个修正版本, 有着更好的理论性质以及在应用中有着更好的覆盖率.

The Basic Bootstrap Confidence Interval

由 $P(L<\hat{\theta}-\theta<U)=1-\alpha$ 在$\hat{\theta}-\theta$ 的分布未知时, 由于Bootstrap重复$\hat{\theta}^*$ 的样本分位数$\hat{\theta}^*_{[(B+1)\alpha/2]}$和 $\hat{\theta}^*_{[(B+1)(1-\alpha/2)]}$满足 $P(\hat{\theta}^*_{[(B+1)\alpha/2]}-\hat{\theta}\le\hat{\theta}-\theta\le\hat{\theta}^*_{[(B+1)(1-\alpha/2)]}-\hat{\theta})\approx 1-\alpha$. 因此$100(1-\alpha)\%$置信区间为 $(2\hat{\theta}-\hat{\theta}^*_{[(B+1)(1-\alpha/2)]}, 2\hat{\theta}-\hat{\theta}^*_{[(B+1)\alpha/2]})$ boot包里的函数boot.ci计算五种类型的置信区间: 基本的, 正态, 百分位数, Bootstrap t和BCa.

例11 patch数据比值统计量的Bootstrap 置信区间

本例说明如何使用boot和boot.ci函数得到正态,基本的和百分位数 Bootstrap置信区间. 下面的代码计算 比值统计量的95%置信区间.

```
library(boot)       #for boot and boot.ci
 data(patch, package = "bootstrap")
 theta.boot <- function(dat, ind) {
     #function to compute the statistic
     y <- dat[ind, 1]
     z <- dat[ind, 2]
     mean(y) / mean(z)
 }
```

```
    y <- patch$y
    z <- patch$z
    dat <- cbind(y, z)
    boot.obj <- boot(dat, statistic = theta.boot, R = 2000)
    print(boot.obj)
    print(boot.ci(boot.obj,
                  type = c("basic", "norm", "perc")))
```

注意当$|\theta| < 0.2$时, 旧药和新药才被认为是等价的. 因此区间估计没有支持旧药和新药的等价性. 下面 我们根据Bootstrap置信区间的定义计算置信区间, 和上面的结果相对比.

```
   #calculations for bootstrap confidence intervals
    alpha <- c(.025, .975)
    #normal
    print(boot.obj$t0 + qnorm(alpha) * sd(boot.obj$t))
    #basic
    print(2*boot.obj$t0 -
        quantile(boot.obj$t, rev(alpha), type=1))
    #percentile
    print(quantile(boot.obj$t, alpha, type=6))
```

例12 $patch$数据中相关系数的Bootstrap置信区间 对law数据, 计算相关统计量的95%置信区间.

```
library(boot)
    data(law, package = "bootstrap")
    boot.obj <- boot(law, R = 2000,
             statistic = function(x, i){cor(x[i,1], x[i,2])})
    print(boot.ci(boot.obj, type=c("basic","norm","perc")))
```

三种置信区间都覆盖住了$\rho = .76$(此时通过完整数据集law82计算的). 百分位数置信区间和正态 置信区间的差异在于样本相关系数的分布是不是靠近正态分布. 当统计量的分布很靠近正态时, 百分位数 区间和正态区间就会一致.

The Bootstrap t interval

即使当$\hat\theta$的分布是正态分布, 且$\hat\theta$为θ的无偏估计, $Z = (\hat\theta - \theta)/se(\hat\theta)$ 的分布也不会一定是正态的, 这是因为我们估计了$se(\hat\theta)$. 我们也不能说Z的分布是t分布, 因为 Bootstrap估计$\hat{se}(\hat\theta)$的分布未知. Bootstrap t区间并没有使用t分布作为推断分布. 而是使用 再抽样方法得到一个"t类型"的统计量的分布. 假设$x = (x_1, \ldots, x_n)$为观测到得样本, 则$100(1-\alpha)\%$ Bootstrap t 置信区间为 $(\hat\theta - t^*_{1-\alpha/2}\hat{se}(\hat\theta), \hat\theta - t^*_{\alpha/2}\hat{se}(\hat\theta))$ 其中$\hat{se}(\hat\theta), t^*_{\alpha/2}$ 和$t^*_{1-\alpha/2}$由下面的方法计算:

Bootstrap t 区间

1. 计算观测到得$\hat\theta$.
2. 对每个重复, $b = 1, \cdots, B$:
    1. 从x中有放回的抽样得到第b个样本$x^{(b)} = (x_1^{(b)}, \ldots, x_n^{(b)})$.
    2. 由第b个再抽样样本计算$\hat\theta^{(b)}$.
    3. 计算标准差估计$\hat{se}(\hat\theta^{(b)})$.(对每个Bootstrap样本$x^{(b)}$, 再单独进行一个Bootstrap估计).
    4. 计算第b个重复下的 "t" 统计量: $t^{(b)} = (\hat\theta^{(b)} - \hat\theta)/\hat{se}(\hat\theta^{(b)})$.
3. 重复样本$t^{(1)}, \ldots, t^{(B)}$的分布作为推断分布. 找出样本分位数$t^*_{\alpha/2}$和$t^*_{1-\alpha/2}$.
4. 计算$\hat{se}(\hat\theta)$, 即Bootstrap重复$\{\hat\theta^{(b)}\}$的样本标准差.
5. 计算置信界 $(\hat\theta - t^*_{1-\alpha/2}\hat{se}(\hat\theta), \hat\theta - t^*_{\alpha/2}\hat{se}(\hat\theta))$.

Bootstrap t区间的一个缺点是要再次使用Bootstrap方法得到标准差的估计$\hat{se}(\hat{\theta}^{(b)})$. 这是在Bootstrap 里面嵌套Bootstrap. 若$B = 1000$, 则Bootstrap t 区间方法需要比别的方法1000倍的时间.

例13 Bootstrap t区间 本例我们写一个函数来计算一元或者多元样本下Bootstrap t 置信区间. 默认的置信水平为95%, Bootstrap重复数为500, 估计标准差的重复次数默认为100.

```
boot.t.ci <-
   function(x, B = 500, R = 100, level = .95, statistic){
       #compute the bootstrap t CI
       x <- as.matrix(x);  n <- nrow(x)
       stat <- numeric(B); se <- numeric(B)
       boot.se <- function(x, R, f) {
           #local function to compute the bootstrap
           #estimate of standard error for statistic f(x)
           x <- as.matrix(x); m <- nrow(x)
           th <- replicate(R, expr = {
               i <- sample(1:m, size = m, replace = TRUE)
               f(x[i, ])
               })
           return(sd(th))
       }
       for (b in 1:B) {
           j <- sample(1:n, size = n, replace = TRUE)
           y <- x[j, ]
           stat[b] <- statistic(y)
           se[b] <- boot.se(y, R = R, f = statistic)
       }
       stat0 <- statistic(x)
       t.stats <- (stat - stat0) / se
       se0 <- sd(stat)
       alpha <- 1 - level
       Qt <- quantile(t.stats, c(alpha/2, 1-alpha/2), type = 1)
       names(Qt) <- rev(names(Qt))
       CI <- rev(stat0 - Qt * se0)
   }
```

例14 patch数据下比值统计量的Bootstrap t 置信区间

程序如下:

```
   #boot package and patch data were loaded in Example 7.10
   #library(boot)       #for boot and boot.ci
   #data(patch, package = "bootstrap")
   dat <- cbind(patch$y, patch$z)
   stat <- function(dat) {
       mean(dat[, 1]) / mean(dat[, 2]) }
   ci <- boot.t.ci(dat, statistic = stat, B=2000, R=200)
   print(ci)
```

取自 "http://shjkx.wang/index.php?title=统计计算_张楠_2019秋:_Bootstrap_Confidence_Intervals&oldid=161447"

本页面最后编辑于2019年11月6日 (星期三) 11:48。