基础数理统计

第八章 Bootstrap 方法

问题

令 $T_n = g(X_1, X_2, \ldots, X_n)$ 为一个统计量，希望知道 T_n 的方差 $\mathrm{V}_F(T_n)$。Bootstrap 方法的思想有两个步骤：

(1) 用 $\mathrm{V}_{\widehat{F}_n}(T_n)$ 估计 $\mathrm{V}_F(T_n)$；

(2) 用随机模拟方法近似求出 $\mathrm{V}_{\widehat{F}_n}(T_n)$。

这里 $\mathrm{V}_{\widehat{F}_n}(T_n)$ 是数据服从 $\widehat{F}_n$ 分布时 T_n 的方差。

8.1 随机模拟

1 8.1 随机模拟
2 8.2 Bootstrap 方差估计
3 8.3 Bootstrap 置信区间

理论依据

大数定律：可以用随机模拟值的样本方差来近似估计方差。

8.2 Bootstrap 方差估计

1 8.1 随机模拟
2 8.2 Bootstrap 方差估计
3 8.3 Bootstrap 置信区间

Bootstrap

给定 n 个样本 $X_1, \cdots, X_n$, 我们可以计算得到经验分布函数 $\widehat{F}_n$,

1. 从分布 $\widehat{F}_n$ 中生成新样本 $\mathcal{X}_1 = \{X_1^*, \cdots, X_n^*\}$;

2. 基于样本 $\mathcal{X}_1 = \{X_1^*, \cdots, X_n^*\}$, 计算出统计量 $T^* = g(X_1^*, \cdots, X_n^*)$;

3. 重复上述步骤 m 次，得到统计量 $T_1^*, \cdots, T_m^*$.

有放回抽样

经验分布函数，

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le x),\ \forall\ x \in \mathbb{R},$$

所以从分布函数 $\widehat{F}_n(x)$ 中生成样本，等价的就是：
Bootstrap= 有放回抽样

Bootstrap

1. 从样本 $X_1, \cdots, X_n$ 中有放回抽样得到新样本 $\mathcal{X}_1 = \{X_1^*, \cdots, X_n^*\}$;
2. 基于样本 $\mathcal{X}_1 = \{X_1^*, \cdots, X_n^*\}$, 计算出统计量 $T^* = g(X_1^*, \cdots, X_n^*)$;
3. 重复上述步骤 m 次，得到统计量 $T_1^*, \cdots, T_m^*$.

注意：

- 新得到的样本中可能是有重复的因为 Bootstrap 的机制是有放回抽样。
- 新样本的容量等于原始样本的容量。

图表说明

$$\text{实际情况 } F \Rightarrow X_1, \cdots, X_n \Rightarrow T_n = g(X_1, \cdots, X_n);$$

$$\text{Bootstrap 方法} \widehat{F}_n \Rightarrow X_1^*, \cdots, X_n^* \Rightarrow T_n^\star = g(X_1^*, \cdots, X_n^*);$$

例子：均值

考虑一个抛硬币的例子，我们抛 10 次记录每次是否正面向上，得到 10 个观察值

$$X = \{x_1, x_2, \ldots, x_{10}\}.$$

我们可以用样本均值来估计正面向上的概率

$$\bar{x} = \frac{1}{10}(x_1 + x_2 + \cdots + x_{10}).$$

均值–续

要刻画上述估计的精度 (推断出 $\bar{x}$ 的分布)，我们可以考虑 Bootstrap. 首先重抽样观察值得到一个 Bootstrap 样本，可能如下：

$$X_1^* = \{x_2, x_1, x_{10}, x_{10}, x_3, x_4, x_6, x_7, x_1, x_9\}.$$

我们可以基于新样本 X_1^* 计算出 Bootstrap 均值: μ_1^*. 重复这一个过程得到第二个新样本 X_2^*，计算 μ_2^*. 以此类推，重复 100 次得到 100 个 $\mu_1^*, \mu_2^*, \ldots, \mu_{100}^*$.

- 这些通过重抽样得到的统计量 $T_1^*, \cdots, T_m^*$ 反映了原始样本统计量 T 的波动情况；
- 通过分析这些统计量，我们可以估计方差，构造出对应的置信区间。

Bootstrap 方差估计

1. 从 $\widehat{F}_n$ 分布中抽样 $X_1^\star, \ldots, X_n^\star$;
2. 计算 $T_n^\star = g(X_1^\star, \ldots, X_n^\star)$;
3. 重复第 1 步和第 2 步 B 次，得到 $T_{n,1}^\star, \ldots, T_{n,B}^\star$;
4. 令

$$v_{\text{boot}} = \frac{1}{B}\sum_{b=1}^{B}\left(T_{n,b}^\star - \frac{1}{B}\sum_{r=1}^{B}T_{n,r}^\star\right)^2.$$

8.3 Bootstrap 置信区间

1 8.1 随机模拟

2 8.2 Bootstrap 方差估计

3 8.3 Bootstrap 置信区间

一、正态区间法

基于所得到的 Bootstrap 统计量 $T_1^*, \cdots, T_m^*$, 我们可以构造置信区间

$$T_n \pm z_{\alpha/2}\widehat{\mathrm{se}}_{\mathrm{boot}},$$

其中 $\widehat{\mathrm{se}}_{\mathrm{boot}} = \sqrt{v_{\mathrm{boot}}}$ 是标准差的 Bootstrap 估计。
该区间不是很准确，除非 T_n 的分布接近正态分布。

二、枢轴量法置信区间

令 $\theta = T(F)$, $\widehat{\theta}_n = T(\widehat{F}_n)$，定义枢轴量为 $R_n = \widehat{\theta}_n - \theta$。用 $\widehat{\theta}^\star_{n,1}, \ldots, \widehat{\theta}^\star_{n,B}$ 表示 $\widehat{\theta}_n$ 的 Bootstrap 复本。$\theta^\star_\beta$ 表示 $\widehat{\theta}^\star_{n,1}, \ldots, \widehat{\theta}^\star_{n,B}$ 的 β 分位数。$1-\alpha$ 的 Bootstrap 枢轴置信区间为

$$C_n = (2\widehat{\theta}_n - \theta^\star_{1-\alpha/2}, 2\widehat{\theta}_n - \theta^\star_{\alpha/2}). \tag{1}$$

枢轴量法置信区间–续

注意到，令 $H(r) = P_F(R_n \le r)$, $C_n^\star = (a, b)$，其中

$$a = \widehat{\theta}_n - H^{-1}(1 - \frac{\alpha}{2}), \quad b = \widehat{\theta}_n - H^{-1}(\frac{\alpha}{2}),$$

则

$$P(a \le \theta \le b) = P(\widehat{\theta}_n - b \le R_n \le \widehat{\theta}_n - a) = 1 - \alpha.$$

由于 H 的 Bootstrap 估计为

$$\widehat{H}(r) = \frac{1}{B}\sum_{b=1}^{B} I(R_{n,b}^\star \le r) \text{ 其中 } R_{n,b}^\star = \widehat{\theta}_{n,b}^\star - \widehat{\theta}_n.$$

故 $\widehat{H}^{-1}(1 - \alpha/2) = \theta_{1-\alpha/2}^\star - \widehat{\theta}_n$, $\widehat{H}^{-1}(\alpha/2) = \theta_{\alpha/2}^\star - \widehat{\theta}_n$.

枢轴量法置信区间–续

定理 1

当 $T(F)$ 满足一定的条件时，

$$P_F(T(F) \in C_n) \to 1 - \alpha, \text{ as } n \to \infty.$$

三、分位区间法

Bootstrap 分位数区间定义为

$$C_n = (\theta^\star_{\alpha/2}, \theta^\star_{1-\alpha/2}).$$

例子

例 1

$X_{i1}, X_{i2}, \ldots, X_{in_i}$ 是来自分布 $F_i (i = 1, 2)$ 的简单随机样本，它们相互独立。求 F_1, F_2 的总体中位数之差的置信区间。

1. 记 $\widehat{\theta}_i$ 为 $X_{i1}, X_{i2}, \ldots, X_{in_i}$ 的样本中位数，记

$$\widehat{\theta} = \widehat{\theta}_1 - \widehat{\theta}_2.$$

2. 从 $X_{i1}, X_{i2}, \ldots, X_{in_i}$ 有放回抽样得到 $X^{\star}_{i1}, X^{\star}_{i2}, \ldots, X^{\star}_{in_i}$，计算样本中位数 $\widehat{\theta}^{\star}_1, \widehat{\theta}^{\star}_2$；该步骤重复 B 次得到的样本分位数之差记为

$$\widehat{\theta}^{\star}_r = \widehat{\theta}^{\star}_{1,r} - \widehat{\theta}^{\star}_{2,r} \ (r = 1, 2, \ldots, B)$$

$\widehat{\theta}^{\star}_r (r = 1, 2, \ldots, B)$ 的样本方差记为 v_{boot}，α 样本分位数记为 θ_α。

例子–续

1. 正态置信区间：

$$(\widehat{\theta} - z_{\alpha/2}\sqrt{\mathsf{v}_{boot}}, \widehat{\theta} + z_{\alpha/2}\sqrt{\mathsf{v}_{boot}}).$$

2. 枢轴置信区间：

$$(2\widehat{\theta} - \theta_{1-\alpha/2}, 2\widehat{\theta} - \theta_{\alpha/2}).$$

3. 分位数置信区间：

$$(\theta_{\alpha/2}, \theta_{1-\alpha/2}).$$

作业：2，3，5，7，8