自助法 和 经验似然方法
Bootstrap 和 Empirical Likelihood

Mai Zhou

Abstract

我们用不太严格 (但是易懂) 的语言从一个侧面讲述 Bootstrap 和 Empirical Likelihood 方法的一点相似和不同之处。重点是介绍并帮助大家理解 Empirical Likelihood 构造置信区间的方法，以及关于 Empirical Likelihood 的 Wilks 定理。

1 问题设定

给定随机样本 $x_1, x_2, \cdots, x_n$ 我们假定它是从总体 F_0 抽样得来。我们对于总体 F_0 没有假定它是属于某个参数分布族，所以这是一个非参数模型。另外又有一个我们感兴趣的（有限维）统计量 $T(x_1, x_2, \cdots, x_n)$ 简单记为 $T(X)$。我们不妨先假定它是一维的。(可以想像 $T(X) = 1/n \sum x_i$)。

下面我们仔细分析，比较 Bootstrap 和 Empirical Likelihood 如何做基于 $T(X)$ 的置信区间。这里的参数是 $\mathbf{E}T(X) = T(F_0)$. [1]

2 Bootstrap 的理由

对于 Bootstrap 有所了解的读者，可以略过这一节。

Bootstrap 做法的理由是:

如果我们可以得到更多从总体 F_0 抽样得来的 $x_{n+1}, \cdots, x_m$ 则不难用模拟得到 $T(X)$ 的分布。只要一遍又一遍的，使用新的 x 观察值来计算新的 $T(X)$，然后用大数定理就可以得出其分布。

但是我们只有 $x_1, \cdots, x_n$，没有更多。此路不通。

另一方面，因为 “样本是总体的一个忠实反映”，既然不能从总体抽样，那我们从它的 “忠实反映” 来抽!

这相当于从 $\hat{F}_n$ 中抽样。这里 $\hat{F}_n$ 是基于 $x_1, \cdots, x_n$ 的经验分布。

[1] 可能大家想到了用 t-test 和由它产生的置信区间。不过因为 1. 如果统计量 T 不是 $1/n \sum x_i$ 则没法用。2. 用 t-test 方法产生的区间必定是对称的。而 Bootstrap 和 Empirical likelihood 方法都有可能产生非对称的区间。

而众所周知，$\hat{F}_n$ 是 F_0 的一个很好的估计。

$$\hat{F}_n(t) = 1/n \sum_{i=1}^{n} I[x_i \le t]$$

把 $\hat{F}_n$ 看成固定，从这里面抽大小为 n 的随机样本 $y_1, y_2, \cdots, y_n$ 即所谓 Bootstrap 样本。由于 $\hat{F}_n$ 和 F_0 很接近，我们应该有

$$\sqrt{n}\,[T(y_1, y_2, \cdots, y_n) - T(\hat{F}_n)] \;\sim\; \sqrt{n}\,[T(x_1, x_2, \cdots, x_n) - T(F_0)]\,. \tag{1}$$

其中“$\sim$”理解为分布很接近。注意：左边的分布，是把 $\hat{F}_n$ 看成固定的，条件分布。在那里只有 y_j 是随机的。

如果我们有 $T(X)$ 的方差估计 $V(X)$，(例如，当 $T(X) = 1/n \sum x_i$ 时，有 $V(X) = 1/(n-1) \sum (x_i - \bar{x})^2$) 我们应该有

$$\sqrt{n}\,\frac{T(y_1, y_2, \cdots, y_n) - T(\hat{F}_n)}{\sqrt{V(Y)}} \;\sim\; \sqrt{n}\,\frac{T(x_1, x_2, \cdots, x_n) - T(F_0)}{\sqrt{V(X)}} \tag{2}$$

也有人简单的说

$$T(y_1, y_2, \cdots, y_n) \;\sim\; T(x_1, x_2, \cdots, x_n) \tag{3}$$

我们不在这儿深入讨论这些逼近说法的好坏/异同/差别。我们主要看产生 $y_1, y_2, \cdots, y_n$ 的方法。

最后，右边的分布是我们需要的。左边的分布是可以模拟出来的。当然，左右两边“很接近”即 (1)，(2) 或 (3) 这个事实需要证明，也已经被许多人证明了。

用左边的分布来逼近右边的分布就是 Bootstrap。

有了分布，置信区间就可以得出。比如，如果 (2) 成立，并且假设我们通过模拟得到对于左边的分布有

$$P\left(a < \sqrt{n}\,\frac{T(y_1, \cdots, y_n) - T(\hat{F}_n)}{\sqrt{V(Y)}} < b\right) = 0.90$$

那么，我们就近似的有

$$P\left(a < \sqrt{n}\,\frac{T(x_1, \cdots, x_n) - T(F_0)}{\sqrt{V(X)}} < b\right) \approx 0.90$$

这就可以导出 $T(F_0)$ 的一个 (Bootstrap t-) 置信区间。当然还有别的 (如 Bootstrap percentile) 方法等等。

3 Bootstrap 抽样的三种等价的视角

总结一下：Bootstrap 如何产生 $y_1, y_2, \cdots, y_n$ 的三个理解方法。

A. 使用分布函数 $\hat{F}_n$ 来产生样本。即把 $\hat{F}_n$ 当作总体。得到 Bootstrap 样本 $y_1, \cdots, y_n$.

B. 对于 $x_1, x_2, \cdots, x_n$ 做有放回的抽样。抽 n 次 得到 Bootstrap 样本 $y_1, y_2, \cdots, y_n$。

如果我们重新排列，归类一下 $y_1, y_2, \ldots, y_n$，按照这些 y 中有多少个等于 x_1，多少个等于 x_2，等等，来记录：

$$w_1 = \sum_{j=1}^{n} I[y_j = x_1],\ w_2 = \sum_{j=1}^{n} I[y_j = x_2],\ \cdots,\ w_n = \sum_{j=1}^{n} I[y_j = x_n]$$

那么，在 $x_1, x_2, \cdots, x_n$ 已经给定的情况下，不难看出 $w_1, w_2, \cdots, w_n$ 与 $y_1, y_2, \cdots\ y_n$ 等价。唯一从 w 里得不到的信息是 y 的出现先后的次序。但是在我们的“随机样本”假定下，观察值的出现次序是不带任何信息的。

不难看出，$w_1, w_2, \cdots, w_n$ 服从多项分布。这样，我们就有了第三种产生 Bootstrap 样本的方法。

C. 产生一个服从多项分布 $M(n, p = (1/n, 1/n, \cdots 1/n))$ 的随机向量 $W = (w_1, w_2, \cdots, w_n)$。将 $x_1, x_2, \cdots, x_n$ 用 W 加权就得到 $y_1^*, y_2^*\ \cdots, y_n^*$。(唯一不同的是 y^* 的排列次序未必与原 y 一样)。

如果 $w_j = 2$ 就表示，$y_1, y_2, \cdots, y_n$ 中包含了两个 x_j，等等。如果 $w_1 = 1, w_2 = 1, \cdots, w_n = 1$ 就得到原来的样本。

注：多项分布 $P(w_j = k) = \binom{n}{k}(1/n)^k(1 - 1/n)^{n-k}$ 其中 $k = 0, 1, \cdots, n$。并且 $\sum w_j = n$。

4 贝叶斯 Bootstrap

贝叶斯 (Bayesian Bootstrap) 与上面第三种观点很像。只不过把 W 服从这个多项分布改为 W/n 服从 Dirichlet 分布，参数为 $\alpha = (1, 1, \cdots, 1)$。

改完之后 W 的均值，方差，协方差都很像。只不过多项分布是离散的，可以取零值。而 Dirichlet 分布是连续的。Bayesian Bootstrap 把这个 Dirichlet 分布看作是 Posterior 分布。

从某种意义上讲，Dirichlet$(\alpha = (1, 1, \cdots, 1))$ 分布就是多项分布 $M(n, (1/n, \cdots, 1/n))$ 的连续形式。

只不过要注意正则归一化：多项分布的 W 有 $\sum w_i = n$. Dirichlet 分布的 W 有 $\sum w_i = 1$。注：对于 Dirichlet, $\mathbf{E}w_j = 1/n$。对于多项分布 $\mathbf{E}w_j = 1$。以后我们总假定 W 已经正则归一化 $\sum w_i = 1$。

5 经验似然方法

经验似然方法和上面第三种 Bootstrap 抽样 $w_1, w_2, \cdots, w_n$ 的表述有一点联系/相近。

为确定记，我们假定 w 是一个概率，即 $w_j \geq 0, \sum w_j = 1$。而且 w_j 不必是离散的，而可以取任何值，只要满足概率的要求。(这个有点像贝叶斯 Bootstrap，但是这里没有分布的假定。)

这样一个 $w_1, w_2, \cdots, w_n$ 可以看成/产生一个（离散的）分布函数:

$$F_w(t) = \sum_i w_i I[x_i \leq t] \ .$$

如果 $w_1 = 1/n, \cdots, w_n = 1/n$ 这就是 $\hat{F}_n$ (经验分布或者样本分布)。

我们将 w 看作是对于 $1/n, 1/n, \cdots, 1/n$，或者 $\hat{F}_n$ 的一个扰动，或者叫扭曲。

但是 EL 方法没有假设 w 是随机的，具有一个分布 (比如 Dirichlet) 等等。而是: EL 引进了一个距离，专门用来测量 F_w 到 $\hat{F}_n$ 的距离。或者是 $w_1, w_2, \cdots, w_n$ 到 $1/n, 1/n, \cdots, 1/n$ 的距离。

$$D(F_w, \hat{F}_n) = -2\sum_{i=1}^{n} \log \frac{w_i}{1/n} = -2\sum_{i=1}^{n} \log nw_i$$

不难证明，当且仅当 $F_w = \hat{F}_n$ 时，$D = 0$. 如若不然，$D > 0$.

注: 常数 2 只是为了后面方便。这个距离有概率意义，即 $-2\log$ 经验似然比。

引进一个距离并不特别新奇，新奇的是我们有 Wilks 定理。

5.1 Wilks 定理的应用

如果定义了上面的距离 $D(F_w, \hat{F}_n)$, 那么我们有一种构造置信区间的方法:

记

$$\mathcal{N} = \{\, F_w : s.t.\ D(F_w, \hat{F}_n) \leq 2.7 \,\}$$

这是 $\hat{F}_n$ 的一个邻域。

取 $T(\cdot)$ 在 $\mathcal{N}$ 上的最小，最大值就得到置信区间: 我们近似的有

$$P(\min_{\mathcal{N}} T(F_w) \leq T(F_0) \leq \max_{\mathcal{N}} T(F_w)) \approx 0.9 \tag{4}$$

上面的数字 2.7 和概率 0.9 的联系是来自卡方分布。因为

$$P(\chi_1^2 \leq 2.7) = 0.9$$

我们可以使用别的数字，比如把 2.7 换成 3.84，概率就变成 0.95.

我们将在下一节讲一点 Wilks 定理的证明。

严格的说，“Wilks 定理” 一般指连续随机变量，参数模型的情况。由 Wilks 在 1930 年代提出。这里用的定理是对于**非参数模型**，由 Owen 在 1980 年代首先证明。

我们不妨还沿用这个名称。不过也有人把它叫 Empirical Likelihood 定理，或者非参数 Wilks 定理。

注：虽然置信区间的定义由（4）给出，但是具体计算时有更加直接的方法。

6 Wilks 定理的证明

此节在首次阅读时可以跳过。

这里的证明是 Owen （1988，1990）给出的。我们假定前面第一节给出的问题设定。并且进一步假定 $T(X)=\int g(t)d\hat{F}_n(t)=1/n\sum g(x_i)$。其中 $g(\cdot)$ 是某个给定函数。我们记 $T(F_0)=\int g(t)dF_0(t)=\theta_0$。

首先我们叙述 Wilks 定理的最重要结论，当 $n\to\infty$

$$\min \ -2\sum_{i=1}^{n}\log(nw_i)=-2\max\sum\log(nw_i)\xrightarrow{\mathcal{D}}\chi_1^2$$

这个极大（或极小）是对于概率 w_i 而言，并且有一个等式约束：

$$\max_{w}\left\{\sum\log(nw_i)\ \ s.t.\ \sum w_i=1;w_j\geq 0;1/n\sum w_ig(x_i)=\theta_0\right\} \tag{5}$$

也就是

$$\max_{F_w}\left\{\sum\log(nw_i)\ s.t.\int g(t)dF_w(t)=\theta_0\right\}\ . \tag{6}$$

我们先来看这个极限分布它有什么用。

有了这个结论，我们就可以用卡方分布计算假设检验（列如）$H_0:\theta_0=3,\ vs\ H_A:\theta_0\neq 3$ 的（近似）$p-$ 值。只要将 θ_0 换成 3, 计算上述 max，乘 -2。最后与卡方分布比较，如果这个值太大就不像是服从卡方的（此时拒绝 H_0）：比如大于 2.7。因为卡方分布有 90% 的概率会小于 2.7。

精确的 $p-$ 值就是 $P(\chi_1^2>-2\max\sum\log(nw_i))$ 。

有了对任何 H_0 的 $p-$ 值，就可以得到置信区间：那些不能被拒绝的零假设参数值，收集在一起就构成置信区间。用数学公式写下来就是

$$\{\theta^*:s.t.-2\max_{\theta^*}\sum\log(nw_i)<2.7\}\ .$$

这里的 max 就是上面定义的 (5) 或 (6) ，只不过将 θ_0 换成 θ^*。

请大家自行证明这个置信区间等价于上面用邻域定义的置信区间。

下面我们梳理一下证明的主要步骤。首先我们使用拉格朗日乘子法，对于 $w_1,\cdots,w_n$，计算带两个等式约束的极大 (5)。约束为 $\sum w_i=1$ 和 $\sum w_ig(x_i)=\theta_0$。可以算出

$$w_i=w_i(\lambda^*)=\frac{1}{n-\lambda^*(g(x_i)-\theta_0)},$$

其中的 λ^* 选取要满足等式约束 $\sum w_i g(x_i) = \theta_0$。

从这个 $w_i(\lambda)$ 的表达式可以看到，如果 $\lambda = 0$ 那么 $w_i = 1/n$。这时 $\sum w_i g(x_i) = 1/n \sum g(x_i)$。但是这个一般来说不满足约束 $\sum w_i g(x_i) = \theta_0$。

所以 λ^* 要取另外一个非零值。不过，由于 $1/n \sum g(x_i)$ 趋向于真值 θ_0 （大数定律），所以 λ^* 离开零不远。

再多说两句：这个 $w(\lambda)$ 可以看成一个（以 λ 为参数的）单参数分布族。当然，需要将 w 归一化，成为一个概率：使得对所有 λ 有 $\sum w_i = 1$. 就是将 $w(\lambda)$ 除以这个和。

这个单参数分布族

$$F(t, \lambda) = \sum_{i=1}^{n} I[x_i \le t] \frac{w_i(\lambda)}{\sum w_j(\lambda)}$$

叫做 “最难的参数分布族”。(最不友好的参数分布族)。这个参数分布在 $\lambda = 0$ 点的 Fisher 信息数，就是在估计 θ 时的**非参信息数**。此乃后话。

Owen 的计算证明了, 如果在上面计算极大的时候，约束条件所用的 θ_0 的确是真值，那么

$$\begin{aligned} -2 \max \sum \log(n w_i) &= -2 \sum \log(n w_i(\lambda^*)) \\ &= \left[1/n \sum g(x_i) - \theta_0\right] V \left[1/n \sum g(x_i) - \theta_0\right] + o_p(1) \end{aligned} \tag{7}$$

其中

$$V = \sum (g(x_i) - \theta_0)^2$$

很显然，当 $n \to \infty$ 时，(7) 的分布趋向于卡方。计算所用到的条件就是可以 Taylor 展开以及大数定理。

7 注解，评论和推广

注 1. 参数 θ 的维数大于一时，证明类似。Owen (1990).

注 2. 注意到 EL 方法构造置信区间的一大特点：没有用到 $T(X)$ 的方差估计。只需要定义（关键的）距离 $D(F_w, \hat{F}_n)$，这个就是 $-2\log$ 经验似然比。Wilks 定理也就是关于这个距离的渐近性质。

注 3. 当样本数据是删失数据时，$-2\log$ 经验似然比, 可以由它的抽样方法的概率，比较容易写出来。例如见 Thomas and Grunkemier (1975) 或者 Owen (2001) 的书中第六章，特别是第 6.6 节。不过证明相应的 Wilks 定理就麻烦一点。这时我们可以利用 “最不友好参数分布族” 的帮助来简化证明。具体证明细节，以及计算方法和各种例子请见拙著 Empirical Likelihood Method in Survival Analysis (2016) 或者中文的 2.0 版：经验似然方法：在生存分析中的理论和实践（2023）。

注 4. 计算经验似然置信区间，相应的 `R` 计算函数包是 `emplik`，可以在 CRAN 上免费下载。具体例子在 `emplik` 包里有一些。在 [7] 里有更多。

References

[1] Efron, B. and Tibshirani, R. (1993). An introduction to the bootstrap, New York: Chapman & Hall,

[2] Rubin, D. The Bayesian Bootstrap. (1981). Ann. Statist. 130-134.

[3] Thomas, D.R. and Grunkemeier, G.L. (1975). Confidence Interval Estimation of Survival Probabilities for Censored Data. Journal of the American Statistical Association, 70, 865-871.

[4] Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. Biometrika 75, 2, 237–249.

[5] Owen, A. (1990). Empirical Likelihood Ratio Confidence Regions Ann. Statist. 18(1): 90-120. DOI: 10.1214/aos/1176347494

[6] Owen, A. *Empirical Likelihood.* Chapman & Hall (2001)

[7] Zhou, M. *Empirical Likelihood Method in Survival Analysis.* Chapman & Hall (2016).

[8] 经验似然方法：在生存分析中的理论和实践，周迈著，杨一帆周迈译