Statistical Computing
Chap. 5: Bootstrap and Permutation Test

LIU, Ran

Department of Statistics,
Beijing Normal University

May 13, 2024

北京師範大學
BEIJING NORMAL UNIVERSITY

目录

目录

Bootstrap 估计的思想

自助法，也称 Bootstrap 方法，由 Efron 于 1979 年首次提出，此后发展了大量的关于 Bootstrap 方法的研究。

首先通过一个简单例子给出 Bootstrap 估计的直观理解。

设 $X_1, X_2, \ldots, X_n$ 为来自总体分布 $F(x;\theta)$ 的样本，对于感兴趣参数 θ，通常可采用极大似然估计或者矩估计等估计方法得到 θ 的估计 $\hat{\theta}(X_1, X_2, \ldots, X_n)$。

很多时候，我们不仅关心 θ 的估计 $\hat{\theta}(X_1, X_2, \ldots, X_n)$ 本身，同时也关心估计量的分布以及分布特征，比如 $\hat{\theta}(X_1, X_2, \ldots, X_n)$ 的均值 $\mathrm{E}(\hat{\theta})$ 和方差 $\mathrm{Var}(\hat{\theta})$。

事实上，一般情况下，很难推导出 $\hat{\theta}(X_1, X_2, \ldots, X_n)$ 的分布，进而得到估计量的均值和方差。

如果 $F(x;\theta)$ 分布形式已知，可以通过蒙特卡洛方法模拟出与 $\hat{\theta}(X_1, X_2, \ldots, X_n)$ 同分布的样本，进而根据样本的信息估计 $\hat{\theta}(X_1, X_2, \ldots, X_n)$ 的分布以及分布特征。

蒙特卡洛方法近似如下：

蒙特卡洛步骤

1、从总体分布 $F(x;\theta)$ 中独立产生 n 个数据 $X_{11}, X_{12}, \cdots, X_{1n}$，得到的 θ 的估计，记为 $\hat{\theta}_1 =: \hat{\theta}_1(X_{11}, X_{12}, \cdots, X_{1n})$；
2、重复上述步骤 m 次，得到的估计分别记为 $\hat{\theta}_1, \hat{\theta}_2, \cdots, \hat{\theta}_m$。

基于样本 $\hat{\theta}_1, \hat{\theta}_2, \cdots, \hat{\theta}_m$ 推断 $\hat{\theta}(X_1, X_2, \ldots, X_n)$ 的分布，$\mathrm{E}(\hat{\theta})$ 和 $\mathrm{Var}(\hat{\theta})$ 可以分别用该样本的均值和方差估计，即 $\hat{\mathrm{E}}(\hat{\theta}) = m^{-1}\sum_{i=1}^{m}\hat{\theta}_i$，$\widehat{\mathrm{Var}}(\hat{\theta}) = m^{-1}\sum_{i=1}^{m}(\hat{\theta}_i - \hat{\mathrm{E}}(\hat{\theta}))^2$。

如果分布 $F(x;\theta)$ 未知，唯一的信息只有样本 $X_1, X_2, \ldots, X_n$。不能利用上述蒙特卡洛方法从总体 $F(x;\theta)$ 中产生数据，进而不能近似 $\hat{\theta}(X_1, X_2, \ldots, X_n)$ 的分布及其相关特征。

Bootstrap 方法就是利用样本分布 $X_1, X_2, \ldots, X_n$ 代替总体分布 $F(x;\theta)$，从分布 $X_1, X_2, \ldots, X_n$ 中有放回的产生数据，进而近似 $\hat{\theta}(X_1, X_2, \ldots, X_n)$ 的分布。

具体步骤如下：

Bootstrap 步骤

(1) 从样本 $X_1, X_2, \ldots, X_n$ 中有放回的产生数据 $X_{11}^*, X_{12}^*, \cdots, X_{1n}^*$，得到的 θ 的估计，记为 $\hat{\theta}_1^* =: \hat{\theta}_1^*(X_{11}^*, X_{12}^*, \cdots, X_{1n}^*)$；

(2) 重复上述步骤 m 次，相应的估计分别记为 $\hat{\theta}_1^*, \hat{\theta}_2^*, \cdots, \hat{\theta}_m^*$。

利用 $\hat{\theta}_1^*, \hat{\theta}_2^*, \cdots, \hat{\theta}_m^*$ 近似 $\hat{\theta}$ 的分布及其特征。

本质上，自助法相当于用样本 $X_1, X_2, \ldots, X_n$ 的经验分布 F_n 估计总体分布 F，从经验分布中再抽样来估计总体的特征。

从经验分布再抽样本质上相当于从 $X_1, X_2, \ldots, X_n$ 有放回的抽样，以相同的概率抽取到每个样本。记 X^* 表示分布函数是 F_n 的随机变量，则 $P(X^* = X_i) = n^{-1}, i = 1, \cdots, n$。

用 Bootstrap 方法抽取到样本 $X_1^*, X_2^*, \cdots, X_n^*$ 的经验分布 F_n^* 是 F_n 的逼近，F_n 是总体分布 F 的逼近。这两种逼近可以表示为 $F_n^* \to F_n \to F$。

假设我们观察到 10 个样本 $\{2,2,1,1,5,4,4,3,1,2\}$，从中抽取样本点 $1,2,3,4,5$ 的概率分别为 $0.3,0.3,0.1,0.2,0.1$。从中随机选择的一个样本 X^*，其分布函数就是经验分布函数，即

$$F_{X^*}(x) = F_n(x) = \begin{cases} 0, & x < 1; \\ 0.3, & 1 \le x < 2; \\ 0.6, & 2 \le x < 3; \\ 0.7, & 3 \le x < 4; \\ 0.9, & 4 \le x < 5; \\ 1, & x \ge 5. \end{cases}$$

需要说明的是：如果经验分布函数 $F_n(x)$ 没有靠近总体分布函数 $F(x)$，则重复抽样下的分布也不会靠近 $F(x)$。

上例中的样本实际上是从参数为 2 的泊松分布中随机产生，从样本中大量重复抽样可以很好的估计 $F_n(x)$，但是不能很好的估计 $F(x)$。因为无论重复多少次再抽样，得到的 Bootstrap 样本都不会包括总体分布的取值 0。

偏差的自助估计

基于样本数据 $X_1, X_2, \cdots, X_n$，得到感兴趣参数 θ 的估计 $\hat{\theta} =: \theta(X_1, X_2, \cdots, X_n)$。下面分别介绍估计量偏差的 Bootstrap 估计和标准差的 Bootstrap 估计。

首先讨论 $\hat{\theta}$ 的偏差 $\mathrm{Bias}(\hat{\theta}) = \mathrm{E}(\hat{\theta}) - \theta$ 的自助估计 $\widehat{\mathrm{Bias}}(\hat{\theta})$。事实上，只需要分别估计 $\mathrm{E}(\hat{\theta})$ 和 θ，θ 的估计 $\hat{\theta}$ 基于观察到的样本数据直接得到；对于 $\mathrm{E}(\hat{\theta})$ 的估计 $\widehat{\mathrm{E}}(\hat{\theta})$，利用自助法获取到的 $\hat{\theta}_1^*, \hat{\theta}_2^*, \cdots, \hat{\theta}_m^*$ 的均值来估计，也就是 $\widehat{\mathrm{E}}(\hat{\theta}) = m^{-1} \sum_{i=1}^{m} \hat{\theta}_i^*$。

因此

$$\widehat{\mathrm{Bias}}(\hat{\theta}) = m^{-1} \sum_{i=1}^{m} \hat{\theta}_i^* - \hat{\theta}$$

偏差的自助估计

具体步骤

1、从样本 $X_1, X_2, \ldots, X_n$ 中有放回的产生数据 $X_{11}^*, X_{12}^*, \cdots, X_{1n}^*$，得到 θ 的估计并记为 $\hat{\theta}_1^* =: \hat{\theta}_1^*(X_{11}^*, X_{12}^*, \cdots, X_{1n}^*)$；
2、重复上述步骤 m 次，相应的估计分别记为 $\hat{\theta}_1^*, \hat{\theta}_2^*, \cdots, \hat{\theta}_m^*$，以及基于样本 $X_1, X_2, \ldots, X_n$ 得到的估计 $\hat{\theta}$；
3、计算 $\mathrm{E}(\hat{\theta})$ 的自助估计 $\widehat{\mathrm{E}}(\hat{\theta}) = m^{-1}\sum_{i=1}^m \hat{\theta}_i^*$，以及偏差自助估计 $\widehat{\mathrm{Bias}}(\hat{\theta}) = \widehat{\mathrm{E}}(\hat{\theta}) - \hat{\theta} = m^{-1}\sum_{i=1}^m \hat{\theta}_i^* - \hat{\theta}$。

偏差的自助估计

如果偏差大于零，意味着 $\hat{\theta}$ 平均来看过高估计了 θ；而偏差小于零，意味着 $\hat{\theta}$ 平均来看过低估计了 θ。因此，经偏差修正的参数 θ 的估计量为 $\tilde{\theta} = \hat{\theta} - \widehat{\mathrm{Bias}}(\hat{\theta})$。

设数据 $X_1, X_2, \cdots, X_n$ 来自方差为 σ^2 的分布。则 σ^2 的估计为 $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$。$\hat{\sigma}^2$ 的偏差 $\mathrm{Bias}(\hat{\sigma}^2) = \mathrm{E}(\hat{\sigma}^2) - \sigma^2 = -\sigma^2/n$。

一个很自然的问题是：$\hat{\sigma}^2$ 偏差的自助估计

$$\widehat{\mathrm{Bias}}(\hat{\sigma}^2) = m^{-1}\sum_{i=1}^{m}\hat{\sigma}_i^{2*} - \hat{\sigma}^2$$

是否为 $-\sigma^2/n$ 的无偏估计？

偏差的自助估计

首先研究 $\hat{\sigma}_1^{2*} = n^{-1}\sum_{i=1}^{n}(X_{1i}^* - \bar{X}_1^*)^2$ **的性质。因为** $X_{1i}^*, i = 1, 2, \cdots, n$ **独立同分布，以相同的概率在** $X_1, X_2, \cdots, X_n$ **上取值，则**

$$\mathrm{E}\{\hat{\sigma}_1^{2*}|X_1, X_2, \cdots, X_n\} = \frac{n-1}{n}\mathrm{Var}\{X_{11}^*|X_1, X_2, \cdots, X_n\}$$

$$= \frac{n-1}{n}(\frac{1}{n}\sum_{i=1}^{n}X_i^2 - \bar{X}^2) = \frac{n-1}{n^2}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

偏差的自助估计

所以

$$
\begin{aligned}
& \mathrm{E}\{\widehat{\mathrm{Bias}}(\hat{\sigma}^2)|X_1, X_2, \cdots, X_n\} \\
= \; & \mathrm{E}\{m^{-1}\sum_{i=1}^{m}\hat{\sigma}_i^{2*} - \hat{\sigma}^2|X_1, X_2, \cdots, X_n\} \\
= \; & \mathrm{E}\{\hat{\sigma}_1^{2*}|X_1, X_2, \cdots, X_n\} - \hat{\sigma}^2 \\
= \; & \frac{n-1}{n^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 - \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \\
= \; & -\frac{1}{n^2}\sum_{i=1}^{n}(X_i - \bar{X})^2. \qquad (1)
\end{aligned}
$$

偏差的自助估计

因此

$$\mathrm{E}\{\widehat{\mathrm{Bias}}(\hat{\sigma}^2)\} = -(n-1)\sigma^2/n^2$$

根据式 (1)，条件期望 $\mathrm{E}\{\widehat{\mathrm{Bias}}(\hat{\sigma}^2)|X_1, X_2, \cdots, X_n\}$ 是 $\hat{\sigma}^2$ 理论偏差 $-\sigma^2/n$ 的估计，也就是

$$\mathrm{E}\{\widehat{\mathrm{Bias}}(\hat{\sigma}^2)|X_1, X_2, \cdots, X_n\} = -\hat{\sigma}^2/n$$

这里 σ^2 的估计 $\hat{\sigma}^2$ 为 $n^{-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$。

因为 σ^2 的估计 $\hat{\sigma}^2$ 不是无偏的，条件期望 $\mathrm{E}\{\widehat{\mathrm{Bias}}(\hat{\sigma}^2)|X_1, X_2, \cdots, X_n\}$ 也不是无偏估计。

偏差的自助估计

下面给出例子说明估计量偏差的 Bootstrap 估计。

假

设 $X_1, X_2, \cdots, X_n$ 是来自 $N(\mu, \sigma^2)$ 的样本，用 Bootstrap 法估计 σ^2 的估计量 $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ 的偏差。

```
set.seed(1)
n=100;  mu=0;  sigma=2;  m=2000;
X=rnorm(n,mu,sigma);
sigma2hat=mean((X-mean(X))^2) #基于样本信息, sigma^2的估计
sigma2Bst=NULL
for (i in (1:m)){
  O=sample(seq(1,n),n,replace=T)
  Bstrap_X=X[O] #有放回的从X中抽取新样本
  sigma2Bst[i]=mean((Bstrap_X-mean(Bstrap_X))^2)
  # 基于新样本, sigma^2的估计
}
Bias=mean(sigma2Bst)-sigma2hat #偏差的Bootstrap 估计
```

估计量标准差的 Bootstrap 估计

关于估计量 $\hat{\theta}$ 的分布，在前面讲了估计量 $\hat{\theta}$ 抽样分布的 Bootstrap 估计，则 $\hat{\theta}$ 的标准差 $\widehat{\mathrm{SE}}(\hat{\theta})$ 的 Bootstrap 估计，就是采用 $\hat{\theta}$ 分布 Bootstrap 估计的标准差。

标准差估计具体步骤

1、从分布 $X_1, X_2, \ldots, X_n$ 中有放回的产生数据 $X_{11}^*, X_{12}^*, \cdots, X_{1n}^*$，得到 θ 的估计，记为 $\hat{\theta}_1^* =: \hat{\theta}_1^*(X_{11}^*, X_{12}^*, \cdots, X_{1n}^*)$；

2、重复上述步骤 m 次，相应的估计分别记为 $\hat{\theta}_1^*, \hat{\theta}_2^*, \cdots, \hat{\theta}_m^*$；

3、计算 $\bar{\hat{\theta}}^* = m^{-1}\sum_{i=1}^m \hat{\theta}_i^*$，$\hat{\theta}$ 标准差的自助估计为

$$\widehat{\mathrm{SE}}(\hat{\theta}) = \left\{\frac{1}{m}\sum_{i=1}^m (\hat{\theta}_i^* - \bar{\hat{\theta}}^*)^2\right\}^{1/2},$$

相应的得到自助法估计的方差。

用自助法估计 $\hat{\theta}$ 的方差 $\mathrm{Var}(\hat{\theta})$，该方差无妨记为 $\widehat{\mathrm{Var}}(\hat{\theta})$，一个很自然的问题是：自助法得到的估计 $\widehat{\mathrm{Var}}(\hat{\theta})$ 和 $\mathrm{Var}(\hat{\theta})$ 之间的关系是什么？下面通过一个简单例子说明。

假设 $X_1, X_2, \cdots, X_n$ 是来自均值为 μ，方差为 σ^2 的分布，感兴趣的参数为 $\theta = \mu$。则 $\hat{\theta} = \hat{\mu} = n^{-1}\sum_{i=1}^{n} X_i$，$\mathrm{Var}(\hat{\theta}) = n^{-1}\sigma^2$。

$\mathrm{Var}(\hat{\theta})$ 的 Bootstrap 估计为

$$\widehat{\mathrm{Var}}(\hat{\theta}) = \frac{1}{m}\sum_{i=1}^{m}(\hat{\theta}_i^* - \bar{\hat{\theta}}^*)^2 =: \frac{1}{m}\sum_{i=1}^{m}(\hat{\mu}_i^* - \bar{\hat{\mu}}^*)^2.$$

估计量标准差的 Bootstrap 估计

首先计算

$$\begin{aligned} & \mathrm{E}\{\widehat{\mathrm{Var}}(\hat{\theta})|(X_1, X_2, \cdots, X_n)\} \\ = \; & \mathrm{E}\{\frac{1}{m}\sum_{i=1}^{m}\hat{\mu}_i^{*2} - \bar{\hat{\mu}}^{*2}|(X_1, X_2, \cdots, X_n)\} \\ = \; & \mathrm{E}\{\hat{\mu}_1^{*2}|(X_1, X_2, \cdots, X_n)\} - \mathrm{E}\{\bar{\hat{\mu}}^{*2}|(X_1, X_2, \cdots, X_n)\}. \end{aligned}$$

下面分别计算 $\mathrm{E}\{\hat{\mu}_1^{*2}|(X_1, X_2, \cdots, X_n)\}$ **和** $\mathrm{E}\{\bar{\hat{\mu}}^{*2}|(X_1, X_2, \cdots, X_n)\}$**。**

估计量标准差的 Bootstrap 估计

注意到

$$
\begin{aligned}
& \mathrm{E}\{\hat{\mu}_1^{*2}|(X_1, X_2, \cdots, X_n)\} \\
= \; & \mathrm{E}\left\{\frac{1}{n^2}(X_{11}^* + X_{12}^* + \cdots + X_{1n}^*)^2|(X_1, X_2, \cdots, X_n)\right\} \\
= \; & \frac{1}{n}\mathrm{E}\{X_{11}^{*2}|(X_1, X_2, \cdots, X_n)\} + \frac{n-1}{n}\mathrm{E}^2\{X_{11}^*|(X_1, X_2, \cdots, X_n)\} \\
= \; & \frac{1}{n^2}\sum_{i=1}^{n} X_i^2 + \frac{n-1}{n}(\frac{1}{n}\sum_{i=1}^{n} X_i)^2, \qquad (2)
\end{aligned}
$$

估计量标准差的 Bootstrap 估计

以及

$$
\begin{aligned}
& \mathrm{E}\{\bar{\hat{\mu}}^{*2}|(X_1, X_2, \cdots, X_n)\} \\
= \; & \mathrm{E}\{\frac{1}{m^2}(\hat{\mu}_1^* + \cdots + \hat{\mu}_m^*)^2|(X_1, X_2, \cdots, X_n)\} \\
= \; & \frac{1}{m}\mathrm{E}\{\hat{\mu}_1^{*2}|(X_1, X_2, \cdots, X_n)\} + \frac{m-1}{m}\mathrm{E}^2\{\hat{\mu}_1^*|(X_1, X_2, \cdots, X_n)\} \\
= \; & \frac{1}{m}\left\{\frac{1}{n^2}\sum_{i=1}^n X_i^2 + \frac{n-1}{n}\left(\frac{1}{n}\sum_{i=1}^n X_i\right)^2\right\} \\
& + \frac{m-1}{m}\left(\frac{X_1 + \cdots + X_n}{n}\right)^2,
\end{aligned} \tag{3}
$$

估计量标准差的 Bootstrap 估计

结合式 (2)，(3)，可得

$$\begin{aligned}
& \mathrm{E}\{\widehat{\mathrm{Var}}(\hat{\theta})|(X_1, X_2, \cdots, X_n)\} \\
= & \frac{m-1}{m}\left\{\frac{1}{n^2}\sum_{i=1}^{n} X_i^2 + \frac{n-1}{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2\right\} \\
& -\frac{m-1}{m}\left(\frac{X_1+\cdots+X_n}{n}\right)^2 \\
= & \frac{m-1}{m}\left\{\frac{1}{n^2}\sum_{i=1}^{n} X_i^2 - \frac{1}{n}\bar{X}^2\right\} = \frac{m-1}{mn^2}\sum_{i=1}^{n}(X_i-\bar{X})^2 = \frac{m-1}{mn}\hat{\sigma}^2,
\end{aligned}$$

这里 $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$。因此自助法得到的估计 $\widehat{\mathrm{Var}}(\hat{\theta})$ 的条件期望为 $(mn)^{-1}(m-1)\hat{\sigma}^2$，是方差 $\mathrm{Var}(\hat{\theta}) = n^{-1}\sigma^2$ 的估计。

估计量标准差的 Bootstrap 估计

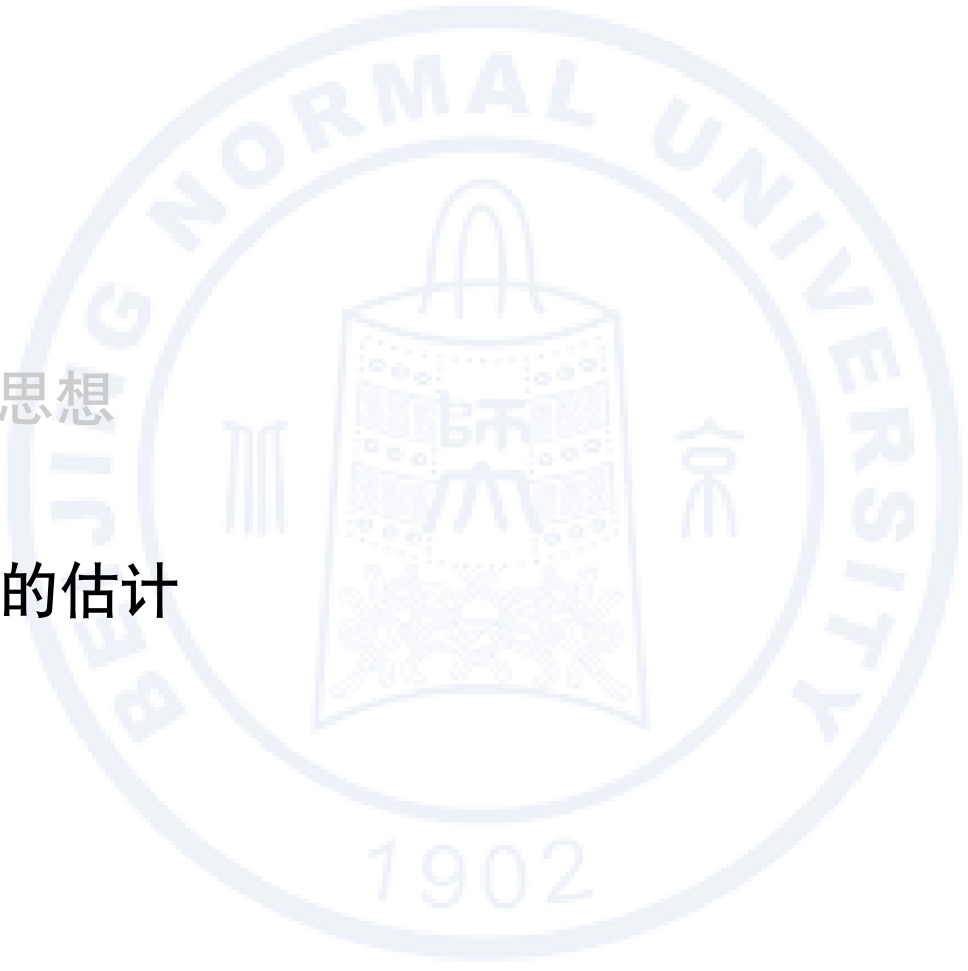
假

设 $X_1, X_2, \cdots, X_n$ 是来自 $N(\mu, \sigma^2)$ 的样本，用 $\hat{\mu} = n^{-1} \sum_{i=1}^{n} X_i$ 估计 μ，并求该估计量方差的 Bootstrap 估计。

```
set.seed(1)
n=100; mu=0; sigma=2; m=2000
X=rnorm(n,mu,sigma);
muhat=mean(X) #基于样本信息，mu的估计
muBst=NULL
for (i in (1:m)){
  O=sample(seq(1,n),n,replace=T)
  Bstrap_X=X[O] #有放回的从X中抽取新样本
  muBst[i]=mean(Bstrap_X) #基于新样本,mu的估计
}
var_muBst=mean((muBst-mean(muBst))^2) #方差的Bootstrap估计
```

其实 Bootstrap 可以想成是近似采样估计量的分布，所以当知道分布的样本时，分布的各阶矩自然能够估计出来。

目录



基于 Jackknife 法的估计

在 Bootstrap 方法提出之前，另外一种有影响力的重抽样方法是由 Quenouille(1949，1956) 提出的 Jackknife 法。

Jackknife 估计的基本思想是，对于从样本 $X_1, X_2, \cdots, X_n$ 中获得的感兴趣参数 θ 的估计量 $\hat{\theta}(X_1, X_2, \cdots, X_n)$，其分布特征（如估计量的偏差和方差）可以用 Jackknife 法得到。

对于给定样本 $X_1, X_2, \cdots, X_n$，每次删除其中一个 (或者几个) 样本点，基于剩下的样本采用相同的估计量公式得到 θ 的估计，经过逐个删除并分别计算估计之后，便可以得到一系列估计值，基于这些估计值进而估计 $\hat{\theta}$ 的分布特征。

基于 Jackknife 法的估计

去掉一个数据的 Jackknife 计算 θ 的具体步骤如下：

Jackknife 计算步骤

1、从观测样本 $X_1, X_2, \ldots, X_n$ 中去掉第 i 个数据 X_i 后的剩余样本，定义为第 i 个 Jackknife 样本，记为 $X_{(-i)} = (X_1, \cdots, X_{i-1}, X_{i+1}, \cdots, X_n)$；

2、基于第 i 个 Jackknife 样本 $X_{(-i)}, i = 1, 2, \cdots, n$，得到相应的估计 $\hat{\theta}_{(-i)} = \hat{\theta}(X_{(-i)}), i = 1, \cdots, n$。

Jackknife 法就是利用 $\hat{\theta}_{(-i)} = \hat{\theta}(X_{(-i)}), i = 1, \cdots, n$ 得到估计量 $\hat{\theta}(X_1, X_2, \cdots, X_n)$ 的分布特征。

估计量偏差的 Jackknife 估计

基于 $\hat{\theta}_{(-i)}$，$\hat{\theta}$ 的偏差 $\mathrm{Bias}(\hat{\theta}) = \mathrm{E}(\hat{\theta}) - \theta$ 的 Jackknife 估计为

$$\widehat{\mathrm{Bias}}(\hat{\theta}) = (n-1)\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{(-i)} - \hat{\theta}\right) = \frac{n-1}{n}\sum_{i=1}^{n}(\hat{\theta}_{(-i)} - \hat{\theta}). \tag{4}$$

以总体方差 $\theta = \sigma^2$ 为例，对于方差的估计 $\hat{\theta} = n^{-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$。估计量 $\hat{\theta}$ 是 σ^2 的有偏估计，偏差为 $\mathrm{Bias}(\hat{\theta}) = \mathrm{E}(\hat{\theta}) - \theta = -n^{-1}\sigma^2$。

对于每一个 Jackknife 估计 $\hat{\theta}_{(-i)}$，基于样本量 $n-1$ 的样本 $X_{(-i)} = (X_1, \cdots, X_{i-1}, X_{i+1}, \cdots, X_n)$ 构造，因此

$$\begin{aligned} & \mathrm{E}(\hat{\theta}_{(-i)} - \hat{\theta}) = \mathrm{E}(\hat{\theta}_{(-i)} - \theta) - \mathrm{E}(\hat{\theta} - \theta) = \mathrm{Bias}(\hat{\theta}_{(-i)}) - \mathrm{Bias}(\hat{\theta}) \\ = & -\frac{\sigma^2}{n-1} - (-\frac{\sigma^2}{n}) = -\frac{\sigma^2}{n(n-1)} = \frac{\mathrm{Bias}(\hat{\theta})}{n-1}. \end{aligned}$$

估计量偏差的 Jackknife 估计

根据上式可得

$$\mathrm{Bias}(\hat{\theta}) = (n-1)\mathrm{E}(\hat{\theta}_{(-i)} - \hat{\theta}). \tag{5}$$

根据式 (4)，(5), 可得

$$\begin{aligned} \mathrm{E}\{\widehat{\mathrm{Bias}}(\hat{\theta})\} &= \frac{n-1}{n}\mathrm{E}\left\{\sum_{i=1}^{n}(\hat{\theta}_{(-i)} - \hat{\theta})\right\} \\ = \quad & (n-1)\mathrm{E}(\hat{\theta}_{(-1)} - \hat{\theta}) = \mathrm{Bias}(\hat{\theta}). \end{aligned}$$

因此偏差 Jackknife 估计 $\widehat{\mathrm{Bias}}(\hat{\theta})$ 是偏差 $\mathrm{Bias}(\hat{\theta})$ 的无偏估计。这也是为什么 jacknife 估计偏差有个系数 (n-1).

估计量标准差的 Jackknife 估计

对于估计量 $\hat{\theta}$，其标准差的 Jackknife 估计定义为

$$\widehat{\mathrm{SE}}_{Jack}(\hat{\theta}) = \left\{ \frac{n-1}{n} \sum_{i=1}^{n} (\hat{\theta}_{(-i)} - \bar{\hat{\theta}}_{(\cdot)})^2 \right\}^{1/2}, \tag{6}$$

这里 $\bar{\hat{\theta}}_{(\cdot)} = n^{-1} \sum_{i=1}^{n} \hat{\theta}_{(-i)}$。相应地，根据式 (6) 可得到方差的 Jackknife 估计。

估计量标准差的 Jackknife 估计

以 θ 是总体均值为例，$\hat{\theta} = \bar{X} = n^{-1}\sum_{i=1}^{n} X_i$，其方差为 $\mathrm{Var}(\hat{\theta}) = n^{-1}\sigma^2$。由于

$$\hat{\theta}_{(-i)} = \frac{n\bar{X} - X_i}{n-1},$$

可得

$$\bar{\hat{\theta}}_{(\cdot)} = n^{-1}\sum_{i=1}^{n}\hat{\theta}_{(-i)} = n^{-1}\sum_{i=1}^{n}\frac{n\bar{X} - X_i}{n-1} = \frac{(n^2-n)\bar{X}}{n(n-1)} = \bar{X}.$$

基于上述结果，不难得到

$$\widehat{\mathrm{Var}}_{Jack}(\hat{\theta}) = \frac{n-1}{n}\sum_{i=1}^{n}(\hat{\theta}_{(-i)} - \bar{\hat{\theta}}_{(\cdot)})^2$$

$$= \frac{n-1}{n}\sum_{i=1}^{n}(\frac{n\bar{X} - X_i}{n-1} - \bar{X})^2 = \frac{1}{n(n-1)}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

因此 $\mathrm{E}(\widehat{\mathrm{Var}}_{Jack}(\hat{\theta})) = n^{-1}\sigma^2$，也就是估计量 θ 的方差的 Jackknife 估计是无偏估计。

目录

Permutation Test 置换检验

置换检验（permutation test）是统计学上一种基于反证法、重抽样原则的非参数性检验.

置换检验的零假设 (虚无假设) 为 $H_0 : F = G$, 即所有样本都服从同一分布。

置换检验通过对比样本置换后的检验统计量与置换前的检验统计量来决定是否拒绝零假设 $H_0 : F = G$、接受备择假设 H_1.

p 值为假设检验中假设零假设为真时观测到的至少与实际观测样本相同的样本的概率。很小的 p 值说明在零假设下观测到的概率很小。

进行置换检验前，首先计算两样本（样本容量设为 n_A 和 n_B）之间原本的检验统计量。检验统计量可以是两样本间平均数之差 $(\bar{X}_A - \bar{X}_B)$、方差之差 $(S_A^2 - S_B^2)$，或 t 值 (t)、卡方检验中的卡方值 (χ^2) 等。

随后，将两个样本打乱后再重新选出两组容量等于之前两样本的新样本（即两个样本容量同样为 n_A 和 n_B 的样本），并计算新的检验统计量。

如接受零假设 $H_0 : F = G$，即样本源于同一分布，则随机抽样计算出的新检验统计量应不难大于最初置换前算出的两样本间检验统计量（如为双侧检验，则是其绝对值应不难大于置换前算出的两样本间检验统计量），即这个概率应大于设定的 I 型错误（假阳性）概率 α.

此处置换检验的检验统计量为两样本间平均数之差 $\hat{\mu}_1 - \hat{\mu}_2$。置换检验中，首先将两个样本混合打乱后，再分别抽出 4 个数与 5 个数，重新计算平均数之差，之后再计算出有多少次置换中得到的新样本间平均数之差大于置换前两样本间平均数之差.