



# 《Python统计计算》

( 2021年秋季学期 )

翟祥  
北京林业大学

E-mail: zhaixbh@126.com

## 第13章 自助法 (Bootstrap)

# 统计计算



- Bootstrap是一个很通用的工具，用来估计标准误差、置信区间和偏差。由Bradley Efron于1979年提出，用于计算任意估计的标准误差
- 术语“Bootstrap”来自短语“to pull oneself up by one's **bootstraps**”（源自西方神话故事“The Adventures of Baron Munchausen”，男爵掉到了深湖底，没有工具，所以他想到了拎着鞋带将自己提起来）
  - ◆ 计算机的引导程序boot也来源于此
  - ◆ 意义：不靠外界力量，而靠自身提升自己的性能，翻译为自助/自举
- 1980年代很流行，因为计算科学被引入统计实践中来

# 统计计算



在统计建模中，伴随着参数的估计值，应该同时给出估计的“标准误差”。设总体  $X \sim F(x, \theta)$ ,  $\theta \in \Theta$ ,  $\hat{\phi}$  是总体的一个参数  $\phi$  的估计量，称  $SE = \sqrt{\text{Var}(\hat{\phi})}$  为  $\hat{\phi}$  的标准误差。实际工作中 SE 一般是未知的，SE 的估计也称为  $\hat{\phi}$  的标准误差。对有偏估计，除了标准误差外我们还希望能够估计偏差。进一步地，我们还可能希望得到统计量  $\hat{\phi}$  的分布，称为抽样分布。

例 3.6.1. 设  $X_i, i = 1, \dots, n$  是总体  $X \sim F(x)$  的样本，样本平均值  $\hat{\phi} = \bar{X} = \frac{1}{n} \sum_i X_i$  为  $\phi = EX$  的点估计， $SE(\bar{X}) = \sqrt{\text{Var}(X)/n}$ ，可以用  $S/\sqrt{n}$  估计  $SE(S^2$  为样本方差)。根据中心极限定理和强大数律，当样本量  $n$  较大时可以取  $EX$  的近似 95% 置信区间为  $\bar{X} \pm 2SE(\bar{X})$ 。□

$$\left( \bar{x} - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right) \quad \left( \bar{x} - t_{\alpha/2} \frac{s_{n-1}}{\sqrt{n}}, \bar{x} + t_{\alpha/2} \frac{s_{n-1}}{\sqrt{n}} \right)$$

# 统计计算



- $T_n = g(X_1, \dots, X_n)$  是一个统计量，或者是数据的某个函数，数据来自某个未知的分布  $F$ ，我们想知道  $T_n$  的某些性质（如偏差、方差和置信区间）
- 假设我们想知道  $T_n$  的方差  $\mathbb{V}_F(T_n)$ 
  - 如果  $\mathbb{V}_F(T_n)$  的形式比较简单，可以直接用学习的嵌入式估计量  $\mathbb{V}_{\hat{F}_n}(T_n)$  作为  $\mathbb{V}_F(T_n)$  的估计
  - 例： $T_n = n^{-1} \sum_{i=1}^n X_i$ ，则
    - $\mathbb{V}_F(T_n) = \sigma^2/n$ ，其中  $\sigma^2 = \int (x - \mu)^2 dF(x)$ ,  $\mu = \int x dF(x)$
    - $\mathbb{V}_{\hat{F}_n}(T_n) = \hat{\sigma}^2/n$ ，其中  $\hat{\sigma}^2 = \sum_{i=1}^n (X_i - \bar{X}_n)^2 / n$
  - 问题：若  $\mathbb{V}_F(T_n)$  的形式很复杂（任意统计量），如何计算/估计？

1. 所有样本指标（如均值、比例、方差等）所形成的分布称为抽样分布
2. 是一种理论概率分布
3. 随机变量是 **样本统计量**
  - ◆ 样本均值, 样本比例等
4. 结果来自容量相同的所有可能样本

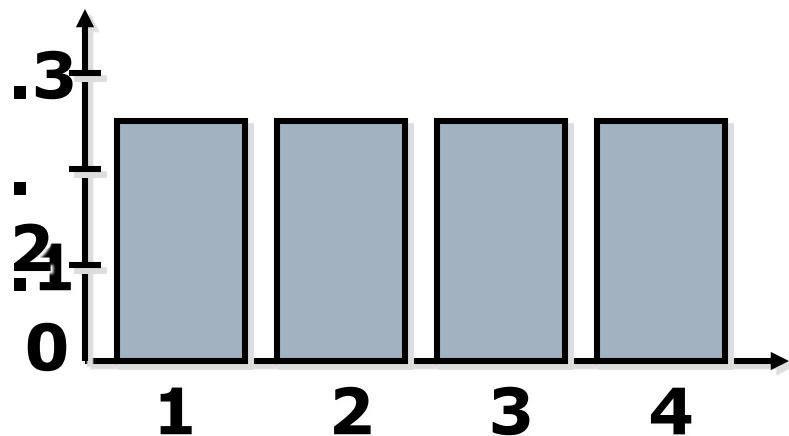
# 样本均值的抽样分布 python™

【例】 设一个总体，含有4个元素（个体），即总体单位数 $N=4$ 。4 个个体分别为 $X_1=1$ 、 $X_2=2$ 、 $X_3=3$  、 $X_4=4$ 。总体的均值、方差及分布如下

均值和方差

$$\mu = \frac{\sum_{i=1}^N X_i}{N} = 2.5$$
$$\sigma^2 = \frac{\sum_{i=1}^N (X_i - \mu)^2}{N} = 1.25$$

总体分布



# 样本均值的抽样分布 python™

➡ 现从总体中抽取 $n=2$ 的简单随机样本，在重复抽样条件下，共有 $4^2=16$ 个样本。所有样本的结果如下表

■所有可能的 $n = 2$ 的样本（共16个）				
■第一个 ■观察值	■第二个观察值			
	■1	■2	■3	■4
■1	■1,1	■1,2	■1,3	■1,4
■2	■2,1	■2,2	■2,3	■2,4
■3	■3,1	■3,2	■3,3	■3,4
■4	■4,1	■4,2	■4,3	■4,4

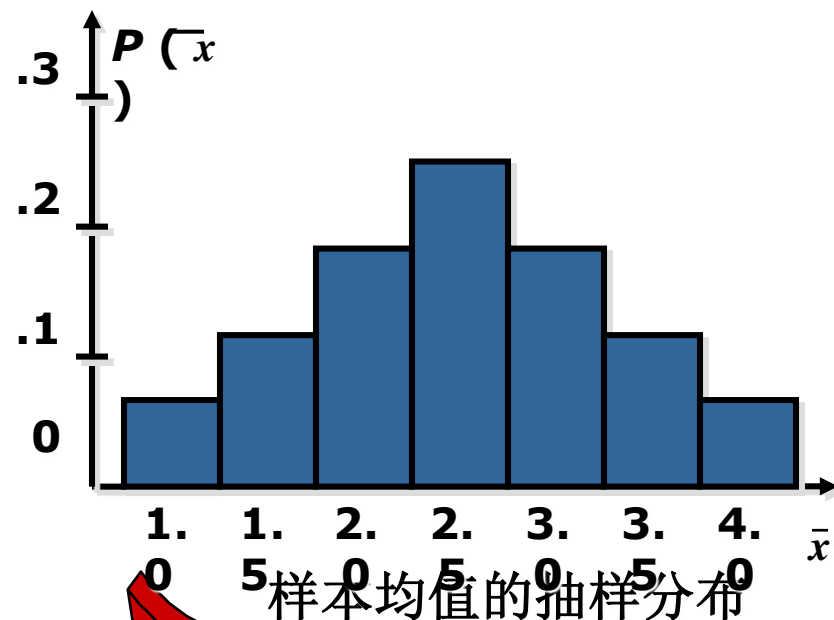


# 样本均值的抽样分布 (一个例子)



➡ 计算出各样本的均值，如下表。并给出样本均值的抽样分布

■16个样本的均值 ( $\bar{x}$ )				
■第一个观察值	■第二个观察值			
	■1	■2	■3	■4
■1	■1.0	■1.5	■2.0	■2.5
■2	■1.5	■2.0	■2.5	■3.0
■3	■2.0	■2.5	■3.0	■3.5
■4	■2.5	■3.0	■3.5	■4.0



样本均值的抽样分布

# 所有样本均值的均值和方差 python™

---

$$\mu_{\bar{x}} = \frac{\sum_{i=1}^n \bar{x}_i}{M} = \frac{1.0 + 1.5 + \dots + 4.0}{16} = 2.5 = \mu$$
$$\sigma_{\bar{x}}^2 = \frac{\sum_{i=1}^n (\bar{x}_i - \mu_{\bar{x}})^2}{M}$$
$$= \frac{(1.0 - 2.5)^2 + \dots + (4.0 - 2.5)^2}{16} = 0.625 = \frac{\sigma^2}{n}$$

式中： $M$ 为样本数目

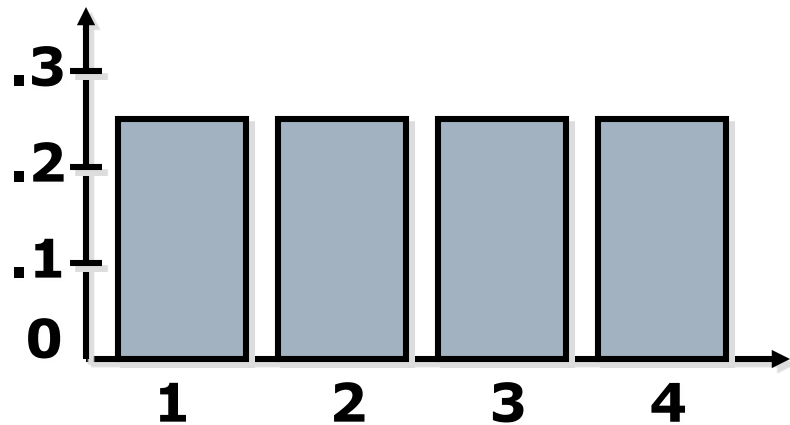
比较及结论：1. 样本均值的均值（数学期望）等于总体均值

---

2. 样本均值的方差等于总体方差的 $1/n$

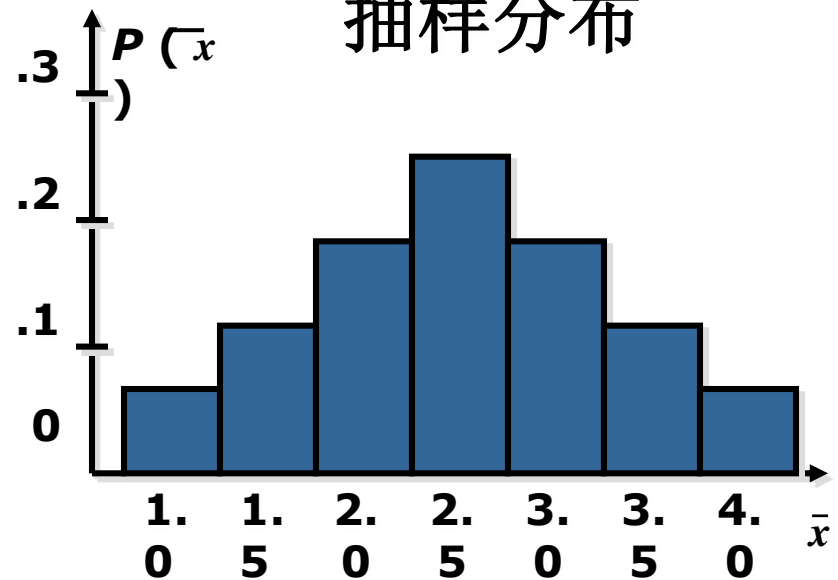
# 样本均值的分布与总体分布的比较

总体分布



$$\mu = 2.5$$
$$\sigma^2 = 1.25$$

抽样分布

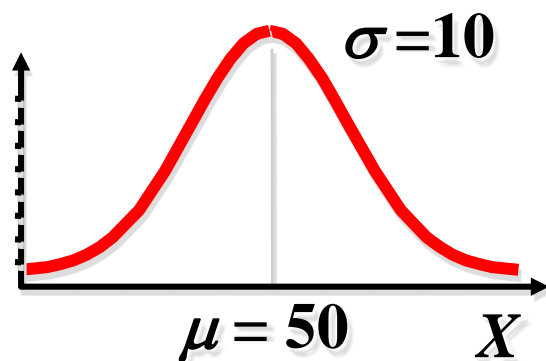


$$\mu_{\bar{x}} = 2.5$$
$$\sigma_{\bar{x}}^2 = 0.625$$

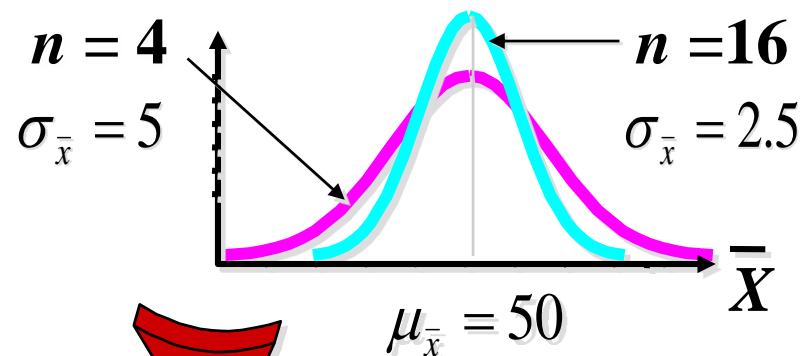
# 样本均值的抽样分布 与中心极限定理



当总体服从正态分布 $N \sim (\mu, \sigma^2)$ 时，来自该总体的所有容量为 $n$ 的样本的均值  $\bar{X}$ 也服从正态分布， $\bar{X}$  的数学期望为 $\mu$ ，方差为 $\sigma^2/n$ 。即  $\bar{X} \sim N(\mu, \sigma^2/n)$



总体分布

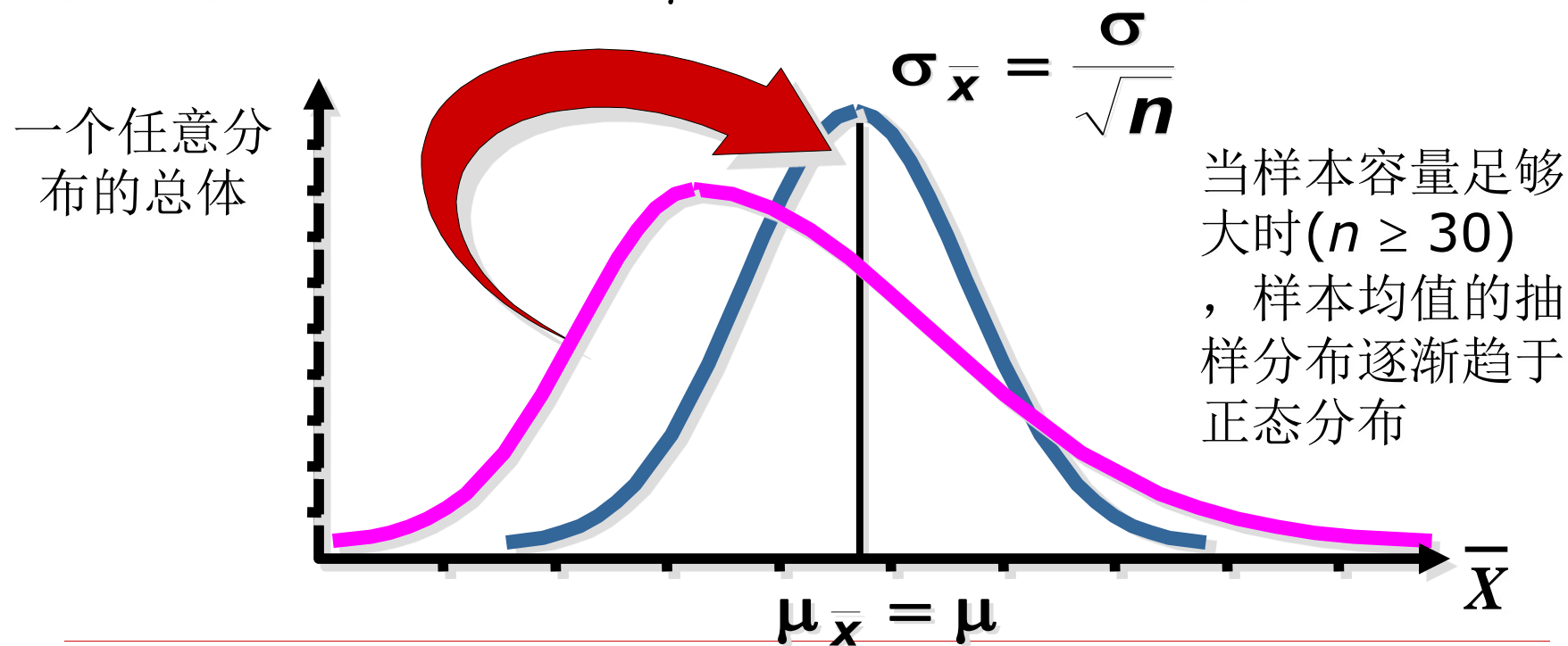


抽样分布

# 中心极限定理



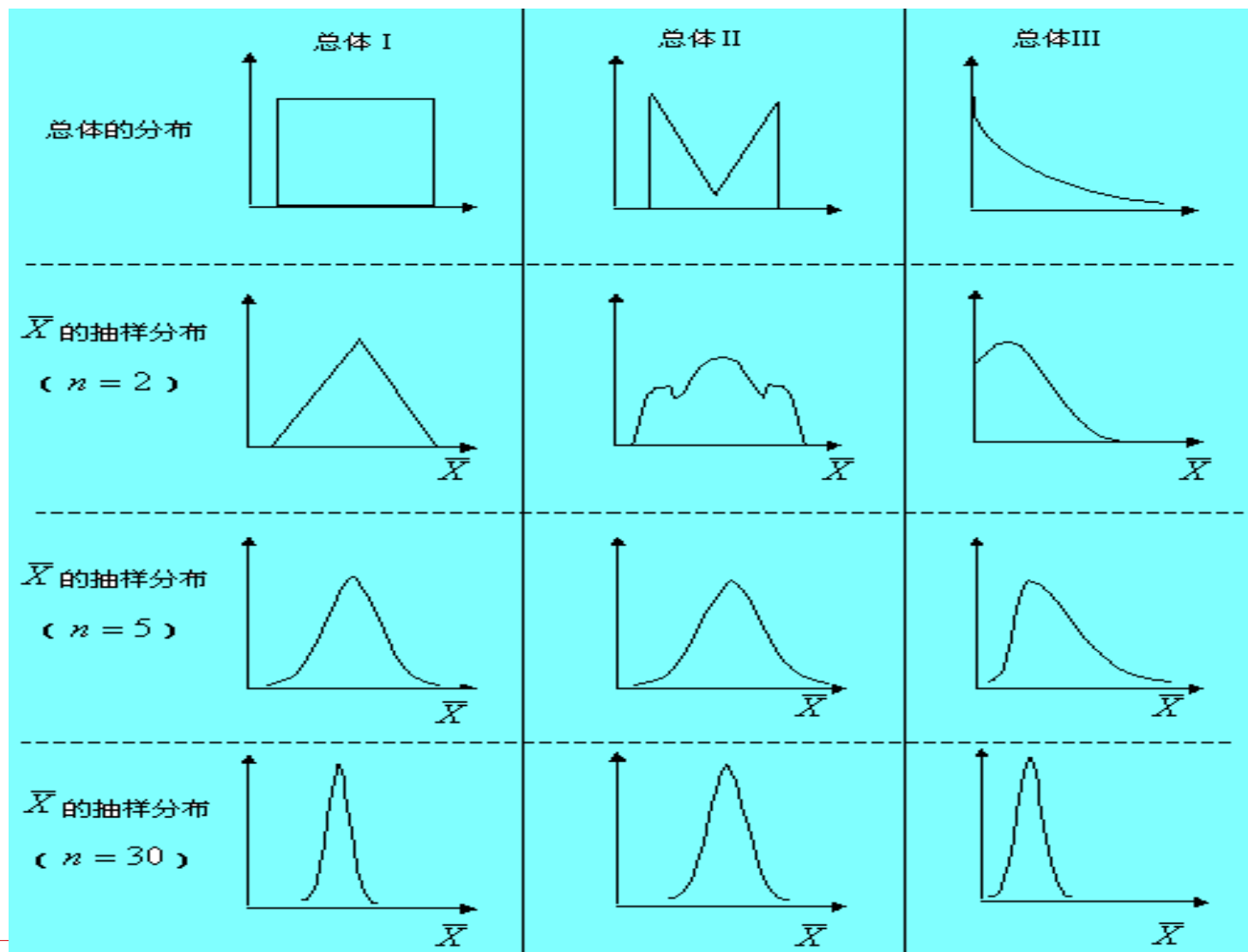
**中心极限定理：** 设从均值为 $\mu$ ，方差为 $\sigma^2$ 的一个任意总体中抽取容量为 $n$ 的样本，当 $n$ 充分大时，样本均值的抽样分布近似服从均值为 $\mu$ 、方差为 $\sigma^2/n$ 的正态分布



# 中心极限定理



$\bar{X}$  的分布趋于正态分布的过程



# Bootstrap 方法的引入

---

## □ 估计标准误差的困难

1、计算参数估计的标准误差不一定总有简单的公式。

例如，需要估计的参数不一定是均值这样的简单特征，像中位数、相关系数这样的参数估计的标准误差就比均值的估计的标准误差要困难得多。

2、在线性模型估计的例子中，如果独立性、线性或者正态分布的假定不满足则求参数估计方差阵变得很困难，比如稳健回归系数的标准误差就很难得到理论公式。

3、在最大似然估计问题中，最大似然估计不一定总是渐近正态的，信息量有时不存在或难以计算，从而无法用上面的方法给出标准误差。

# Bootstrap

---

- Bootstrap: 利用计算手段进行重/再采样 (Resample)
  - 一种基于数据的模拟 (simulation) 方法，用于统计推断。基本思想是：利用样本数据计算统计量和估计样本分布，而不对模型做任何假设 (非参数bootstrap)
  - 无需标准误差的理论计算，因此不关心估计的数学形式有多复杂
  - Bootstrap有两种形式：非参数bootstrap和参数化的bootstrap，但基本思想都是模拟
-



Efron 在1979,1981和1982年的工作中引入和进一步发展了Bootstrap方法, 此后发表了大量的关于 此方法的研究.

Bootstrap方法是一类非参数Monte Carlo方法, 其通过再抽样对总体分布进行估计. 再抽样方法将 观测到的样本视为一个有限总体, 从中进行随机(再)抽样来估计总体的特征以及对抽样总体作出统计推断. 当目标总体分布没有指定时, Bootstrap方法经常被使用, 此时, 样本是唯一已有的信息.

Bootstrap 一词可以指非参数Bootstrap, 也可以指参数Bootstrap(上一讲中). 参数Bootstrap是指总体分布 完全已知, 利用Monte Carlo方法从此总体中抽样进行统计推断; 而非参数Bootstrap是指总体分布完全未知, 利用再抽样

# 重采样

□ 通过从原始数据  $X = X_1, \dots, X_n$  进行n次有放回采样n个数据，得到bootstrap样本

◆ 对原始数据进行有放回的随机采样，抽取的样本数目同原始样本数目一样（也可以不一样）

□ 如：若原始样本为  $X = X_1, X_2, X_3, X_4, X_5$

□ 则bootstrap样本可能为

$$X_2^* = X_1, X_3, X_1, X_4, X_5$$

$$X_1^* = X_2, X_3, X_5, X_4, X_5$$

# Bootstrap

设总体  $X$  服从某个未知分布  $F(x)$ ,  $\mathbf{X} = (x_1, x_2, \dots, x_n)$  是  $X$  的一个样本,  $\phi$  是  $F$  的一个参数, 可以把  $\phi$  看成  $F$  的一个泛函  $\phi(F)$ , 用统计量  $\hat{\phi} = g(\mathbf{X})$  估计  $\phi$ , 设  $\psi = \psi(g, F, n)$  是统计量  $\hat{\phi}$  的某种分布特征 ( $\hat{\phi}$  的抽样分布的数字特征)。例如  $\psi = \sqrt{\text{Var}(\bar{X})}$  为统计量  $\bar{X}$  的标准误差, 又如取  $\psi = E\hat{\phi} - \phi$  为统计量  $\hat{\phi}$  的偏差。可以用随机模拟的方法估计  $\psi$ 。

- (1) 从样本  $\mathbf{X}$  估计总体分布  $F$  为  $\hat{F}$ ;
- (2) 从  $\hat{F}$  抽取  $B$  个独立样本  $\mathbf{Y}^{(b)}$ ,  $b = 1, \dots, B$ , 每一个  $\mathbf{Y}^{(b)}$  样本量为  $n$ , 称  $\mathbf{Y}^{(b)}$  为 bootstrap 样本。
- (3) 从每个 bootstrap 样本  $\mathbf{Y}^{(b)}$  可以估计得到  $\hat{\phi}^{(b)} = g(\mathbf{Y}^{(b)})$ ,  $b = 1, \dots, B$ 。
- (4)  $\hat{\phi}^{(b)}$ ,  $b = 1, \dots, B$  是  $g(\mathbf{Y})$  在  $\hat{F}$  下的独立同分布样本, 可以用标准的估计方法估计关于  $g(\mathbf{Y})$  在  $\hat{F}$  下的分布特征  $\hat{\psi} = \psi(g, \hat{F}, n)$ , 估计结果记作  $\tilde{\psi}$ , 并以  $\tilde{\psi}$  作为统计量  $\hat{\phi}$  的抽样分布的数字特征  $\psi(g, F, n)$  的估计值。

可以基于参数的bootstrap, 也可以基于非参的bootstrap, 但基本思想都是模拟

- 假设我们从  $T_n$  的分布  $G_n$  中抽取 IID 样本  $T_{n,1}, \dots, T_{n,B}$ ，当  $B \rightarrow \infty$  时，根据大数定律，

$$\bar{T}_n = \frac{1}{B} \sum_{b=1}^B T_{n,b} \xrightarrow{P} \int t dG_n \quad t = \mathbb{E} T_n$$

- 也就是说，如果我们从  $G_n$  中抽取大量样本，我们可以用样本均值  $\bar{T}_n$  来近似

- ◆ 当样本数目  $B$  足够大时，样本均值  $\bar{T}_n$  与期望  $\mathbb{E} T_n$  之间的差别可以忽略不计

□ 更一般地, 对任意均值有限的函数 $h$ , 当有

$$\frac{1}{B} \sum_{b=1}^B h(T_{n,b}) \xrightarrow{P} \int h(t) dG_n(t) = \mathbb{E} h(T_n)$$

□ 则当  $h(T_{n,b}) = T_{n,b} - \bar{T}_n$  时, 有

$$\frac{1}{B} \sum_{b=1}^B (T_{n,b} - \bar{T}_n)^2 \xrightarrow{P} \mathbb{E} (T_n - \bar{T}_n)^2 = \mathbb{V} T_n$$

◆ 用模拟样本的方差来近似方差

$$\mathbb{V} T_n$$

## □ 怎样得到 $T_n$ 的分布？

- ◆ 已知的只有  $X$ ，但是我们可以讨论  $X$  的分布  $F$
- ◆ 如果我们可以从分布  $F$  中得到样本  $X_1^*, \dots, X_n^*$ ，我们可以计算
$$T_n^* = g(X_1^*, \dots, X_n^*)$$

## □ 怎样得到 $F$ ？用 $\hat{F}_n$ 代替（嵌入式估计量）

## □ 怎样从 $\hat{F}_n$ 中采样？

- ◆ 因为  $\hat{F}_n$  对每个数据点  $X_1, \dots, X_n$  的质量都为  $1/n$
- ◆ 所以从  $\hat{F}_n$  中抽取一个样本等价于从原始数据随机抽取一个样本
- ◆ 也就是说：为了模拟  $X_1^*, \dots, X_n^* \sim \hat{F}_n$  可以通过有放回地随机抽取  $n$  个样本（bootstrap 样本）来实现

# Bootstrap: 一个重采样过程



## □ 重采样:

- ◆ 通过从原始数据  $X = X_1, \dots, X_n$  进行有放回采样  $n$  个数据, 得到bootstrap样本

$$X_b^* = X_{1,b}^*, \dots, X_{n,b}^*$$

## □ 模拟:

- ◆ 为了估计我们感兴趣的统计量  $T_n = g(\mathbf{X}) = g(X_1, \dots, X_n)$  的方差/中值/均值, 我们用 bootstrap样本对应的统计量 (bootstrap复制)  $T_{n,b}^* = g(\mathbf{X}_b^*) = g(X_{1,b}^*, \dots, X_{n,b}^*)$  近似, 其中  $b = 1, \dots, B$

$$T_{n \text{ boot}} = \frac{1}{B} \sum_{b=1}^B T_{n,b}^* = \frac{1}{B} \sum_{b=1}^B g(\mathbf{X}_b^*)$$

# 例：中值

$X = (3.12, 0, 1.57, 19.67, 0.22, 2.20)$

Mean=4.46

$X1 = (1.57, 0.22, 19.67, 0, 0, 2.2, 3.12)$

Mean=4.13

$X2 = (0, 2.20, 2.20, 2.20, 19.67, 1.57)$

Mean=4.64

$X3 = (0.22, 3.12, 1.57, 3.12, 2.20, 0.22)$

Mean=1.74

$$Mean_{boot} = \frac{1}{3} 4.13 + 4.64 + 1.74 = 3.50$$



# Bootstrap方差估计

□ 方差:  $\mathbb{V}_F T_n = \sigma_T^2 / n$

□ 其中  $\sigma_T^2 = \int (t - \mu_T)^2 dG_n(t)$ ,

◆ 注意:  $F$ 为数据 $X$ 的分布,  $G$ 为统计量 $T$ 的分布

$$\mu_T = \int t dG_n(t)$$

□ 通过两步实现:

◆ 第一步: 用  $\mathbb{V}_{\hat{F}_n} T_n$  估计  $\mathbb{V}_F T_n$

➤ 插入估计, 积分符号变成求和

◆ 第二步: 通过从 $\hat{F}_n$ 中采样来近似计算  $\mathbb{V}_{\hat{F}_n} T_n$

➤ Bootstrap采样+大数定律近似

$$\mathbb{V}_{\hat{F}_n} T_n = \frac{1}{B} \sum_{b=1}^B T_{n,b}^* - \bar{T}_n^*{}^2, \quad \bar{T}_n^* = \frac{1}{B} \sum_{b=1}^B T_{n,b}^*$$

# Bootstrap: 方差估计

## □ Bootstrap的步骤:

□ 1. 画出  $X_1^*, \dots, X_n^* \sim F_n$  (计算bootstrap样本)

□ 2. 计算  $T_n^* = g(X_1^*, \dots, X_n^*)$  (计算bootstrap复制)

□ 3. 重复步骤1和2共 $B$ 次, 得到  $T_{n,1}^*, \dots, T_{n,B}^*$

□ 4. 
$$v_{boot} = \frac{1}{B} \sum_{b=1}^B \left( T_{n,b}^* - \frac{1}{B} \sum_{r=1}^B T_{n,r}^* \right)^2$$

(大数定律)

# Bootstrap: 标准误差估计



- 在标准误差估计中,  $T_n$  可为任意统计量
  - ◆ 如均值（混合高斯模型的例子）
  - ◆ 中值
  - ◆ 偏度
  - ◆ 极大值
  - ◆ ...
- 除了用来计算方差外, 还可以用作其他应用
  - ◆ CDF近似、偏差估计、置信区间估计

# 相关系数的标准误差估计

例 3.6.5. 设  $(H, W)$  为某地小学五年级学生的身高和体重的总体,  $(H, W) \sim F(\cdot, \cdot)$ , 考虑  $H$  和  $W$  的相关系数  $\phi$  的估计。设调查了  $n = 10$  个学生的身高和体重的数据  $(h_i, w_i), i = 1, 2, \dots, n$ :

$h_i$	144	166	163	143	152	169	130	159	160	175
$w_i$	38	44	41	35	38	51	23	51	46	51

计算得  $\hat{\phi} = g(h_1, w_1, \dots, h_n, w_n) = 0.904$ 。令  $SE(\hat{\phi}) = [\text{Var}(\hat{\phi})]^{1/2} = \psi(g, F, n)$ 。设  $\hat{F}$  为  $F$  的估计, 取为经验分布  $F_n$ , 则 bootstrap 方法用随机模拟方法估计  $\psi(g, F_n, n)$ , 然后当作  $SE(\hat{\phi})$  的估计。计算步骤如下:

- (1) 从  $F_n$  中作  $n = 10$  次独立抽样, 即从  $\{(h_1, w_1), \dots, (h_n, w_n)\}$  中有放回独立抽取  $n$  次, 得到  $\hat{F} = F_n$  的一组样本  $\mathbf{Y}^{(1)} = ((h_1^{(1)}, w_1^{(1)}), \dots, (h_n^{(1)}, w_n^{(1)}))$ ;
- (2) 重复第 (1) 步, 直到获取了  $B$  组 bootstrap 样本  $\mathbf{Y}^{(b)}, b = 1, \dots, B$ ;
- (3) 对每一样本  $\mathbf{Y}^{(b)}$  计算样本相关系数  $\hat{\phi}^{(b)} = g(\mathbf{Y}^{(b)}), b = 1, \dots, B$ ;
- (4) 把  $\hat{\phi}^{(b)}, b = 1, \dots, B$  作为  $\hat{F}$  下  $n = 10$  的样本相关系数的简单随机样本, 估计其样本标准差  $S$ , 以  $S$  作为  $\psi(g, \hat{F}, n)$  的估计, 进而用  $S$  估计  $\hat{\phi}$  在真实的总体分布  $F$  下的标准误差  $SE(\hat{\phi})$ 。

取  $B = 10000$  的一次 bootstrap 计算得到的标准误差估计为  $S = 0.101$ 。当  $B \rightarrow \infty$  时  $S \rightarrow \psi(g, F_n, n)$ , 但是要注意, 由于抽样误差影响,  $\psi(g, F_n, n)$  和  $\psi(g, F, n)$  之间的误差无法避免。

# CDF近似

□ 令  $G_n(t) = \mathbb{P}(T_n \leq t)$  为  $T_n$  的CDF

□ 则  $G_n$  的bootstrap估计为

$$\hat{G}_n^*(t) = \frac{1}{B} \sum_{b=1}^B I(T_{n,b}^* \leq t)$$

# 偏差估计（校正）

- 设  $\mathbf{X} = (x_1, x_2, \dots, x_n)$  为总体  $F(\cdot)$  的样本，总体参数  $\phi = \phi(F)$  的估计为  $\hat{\phi} = g(\mathbf{X})$ ,  $b = E\hat{\phi} - \phi$  为估计偏差,  $\text{Var}(\hat{\phi})$  为估计方差。
- 估计的均方误差可以分解为

$$E[\hat{\phi} - \phi]^2 = \text{Var}(\hat{\phi}) + b^2. \quad (3.80)$$

- 如果  $\hat{b}$  是  $b$  的估计, 则参数  $\phi$  的一个改善的估计为  $\tilde{\phi} = \hat{\phi} - \hat{b}$ , 新的估计在减小了偏差的同时一般也减小了均方误差。
- 设  $b = \psi(g, F, n)$ ,  $\hat{F}$  是总体分布  $F$  的一个估计,  $\hat{F}$  取为经验分布  $F_n$ 。
- 可以用  $\hat{b} = \psi(g, \hat{F}, n) = Eg(\mathbf{Y}) - \hat{\phi}$  来估计偏差, 其中  $\mathbf{Y}$  是总体  $F_n$  的样本量为  $n$  的样本,  $\hat{\phi}$  恰好是总体分布为  $F_n$  时的参数  $\phi$ , 即  $\hat{\phi} = \phi(F_n)$ 。

# 偏差估计（校正）

- 如果  $\hat{b}$  不能通过理论公式计算，可以用 bootstrap 方法估计  $\hat{b}$ ，步骤如下：

- (1) 从  $\{x_1, x_2, \dots, x_n\}$  独立有放回地抽取  $n$  个，记为  $\mathbf{Y}^{(1)} = (Y_1^{(1)}, Y_2^{(1)}, \dots, Y_n^{(1)})$ 。
- (2) 重复第 (1) 步，直到获取了  $B$  组 bootstrap 样本  $\mathbf{Y}^{(b)}, b = 1, 2, \dots, B$ ；
- (3) 从每个 bootstrap 样本  $\mathbf{Y}^{(b)}$  可以估计得到  $\hat{\phi}^{(b)} = g(\mathbf{Y}^{(b)})$ ,  $b = 1, \dots, B$ 。
- (4) 用  $\hat{\phi}^{(b)}, b = 1, \dots, B$  作为  $g(\mathbf{Y})$  在  $F_n$  下的独立同分布样本，估计  $\hat{b} = \psi(g, F_n, n)$  为  $\tilde{b} = \frac{1}{B} \sum_{b=1}^B \hat{\phi}^{(b)} - \hat{\phi}$ 。

最后，取  $\tilde{\phi} = \hat{\phi} - \tilde{b} = 2\hat{\phi} - \frac{1}{B} \sum_{b=1}^B \hat{\phi}^{(b)}$  为改善的估计。

# 偏差估计

- 偏差的bootstrap估计定义为:

$$Bias_{boot} T_n = \mathbb{E}_F T_n^* - T_n$$

- Bootstrap偏差估计的步骤为:

- ◆ 得到 $B$ 个独立bootstrap样本  $X_1^*, \dots, X_B^*$
- ◆ 计算每个bootstrap样本  $X_b^*$  对应的统计量的值

$$T_{n,b}^* = g(X_b^*) = g(X_{1,b}^*, \dots, X_{n,b}^*)$$

- ◆ 计算bootstrap期望:  $\bar{T}_n^* = \frac{1}{B} \sum_{r=1}^B T_{n,r}^*$

- ◆ 计算bootstrap偏差:  $Bias_{boot} T_n = \bar{T}_n^* - T_n$



# 例：混合高斯模型：

## □ 标准误差估计

◆ 在标准误差估计中， $B$ 为50到200之间结果比较稳定

$B$	10	20	50	100	500	1000	10000
$se_{boot}$	0.1386	0.2188	0.2245	0.2142	0.2248	0.2212	0.2187

## □ 偏差估计

$$\bar{X}_n = 4.997$$

$B$	10	20	50	100	500	1000	10000
$\mathbb{E}_F \bar{X}^*$	5.0587	4.9551	5.0244	4.9883	4.9945	5.0035	4.9996
$Bias_{boot}$	0.0617	-0.0417	0.0274	-0.0087	-0.0025	0.0064	0.0025

- 如果模型更为复杂，比如，总体分布类型未知， $\hat{b}_1$  这样的简单偏差估计很难得到，这种情况下可以用 bootstrap 方法进行偏差校正。
- 步骤如下：
  - (1) 对  $b = 1, 2, \dots, B$  重复：从  $x_1, x_2, \dots, x_n$  独立有放回地抽取  $n$  个，组成 bootstrap 样本  $\mathbf{Y}^{(b)} = (y_1^{(b)}, \dots, y_n^{(b)})$ ;
  - (2) 对每个 bootstrap 样本计算  $\hat{\phi}^{(b)} = \left( \frac{1}{n} \sum_{i=1}^n y_i^{(b)} \right)^2$ ;
  - (3) 用  $\tilde{\phi} = 2\bar{X}^2 - \frac{1}{B} \sum_{b=1}^B \hat{\phi}^{(b)}$  作为  $\mu^2$  的改善的估计。

# Bootstrap置信区间

枢轴量法是构造置信区间的最基本的方法。设  $\phi$  是总体  $F(\cdot)$  的一个参数, 看成  $F$  的一个泛函  $\phi = \phi(F)$ 。  $\mathbf{X} = (x_1, x_2, \dots, x_n)$  为总体的样本,  $g(\mathbf{X})$  为与  $\phi$  有关系的一个统计量, 经常是  $\phi$  的估计量。如果有变换  $W = h(g(\mathbf{X}), \phi)$  使得  $W$  的分布不依赖于任何未知参数, 则设  $W$  的左右两侧分位数分别为  $w_{\frac{\alpha}{2}}$  和  $w_{1-\frac{\alpha}{2}}$ , 有

$$P(w_{\frac{\alpha}{2}} < h(T, \phi) < w_{1-\frac{\alpha}{2}}) = 1 - \alpha, \quad (3.84)$$

反解上面的不等式可以得到  $\phi$  的置信区间。

如果对枢轴量  $W$  很难求分位数时, 可以用 bootstrap 方法获得置信区间。设  $\hat{F}$  为总体分布  $F$  的估计, 设  $\mathbf{Y} = (y_1, \dots, y_n)$  为总体  $\hat{F}$  的样本,  $\hat{\phi} = \phi(\hat{F})$  为与总体  $\hat{F}$  对应的参数  $\phi$  的值, 实际是  $\phi$  的估计值, 则  $V = h(g(\mathbf{Y}), \hat{\phi})$  与  $W$  的分布相近, 可以用  $V$  的分位数近似  $W$  的分位数。

例 3.6.8. 设总体  $X \sim F(x, \theta)$ ,  $\theta$  为总体的未知参数,  $\phi = EX$ ,  $\mathbf{X} = (x_1, x_2, \dots, x_n)$  为总体的样本, 则  $g(\mathbf{X}) = \bar{X}$  是  $\phi$  的估计, 若  $W = h(g(\mathbf{X}), \phi) = \bar{X} - EX$  的分布与  $\theta$  无关, 求  $W$  的分位数  $w_{\frac{\alpha}{2}}$  和  $w_{1-\frac{\alpha}{2}}$  就可以构造  $\phi = EX$  的置信区间  $(\bar{X} - w_{1-\frac{\alpha}{2}}, \bar{X} - w_{\frac{\alpha}{2}})$ 。

若  $W$  的分位数无法求得, 用经验分布  $F_n$  作为总体分布  $F$  的估计, 这时  $\phi(F_n) = \bar{X}$ , 设  $\mathbf{Y} = (Y_1, \dots, Y_n)$  为  $F_n$  的样本,  $V = h(g(\mathbf{Y}), \bar{X}) = \bar{Y} - \bar{X}$ , 这里  $\bar{X}$  作为已知值, 可以用  $V$  的分位数近似代替  $W$  的分位数。求  $V$  的分位数, 只要有放回独立抽样方法从  $x_1, x_2, \dots, x_n$  抽取  $F_n$  的  $B$  组样本  $\mathbf{Y}^{(b)} = (y_1^{(b)}, \dots, y_n^{(b)}), b = 1, 2, \dots, B$ , 对每组样本计算平均值  $\bar{Y}^{(b)}$ , 定义  $V^{(b)} = \bar{Y}^{(b)} - \bar{X}$ , 用  $V^{(b)}, b = 1, 2, \dots, B$  的样本分位数估计  $V$  的分位数, 作为  $W$  的分位数  $w_{\frac{\alpha}{2}}$  和  $w_{1-\frac{\alpha}{2}}$  的近似。 □

# Bootstrap置信区间

□ 正态区间:  $T_n \pm z_{\alpha/2} se$

◆ 简单, 但该估计不是很准确, 除非  $T_n$  接近正态分布

□ 百分位区间:  $C_n = T_{\alpha/2}^*, T_{1-\alpha/2}^*$  , 对应  $T_{n,1}^*, \dots, T_{n,B}^*$  的样本分位数

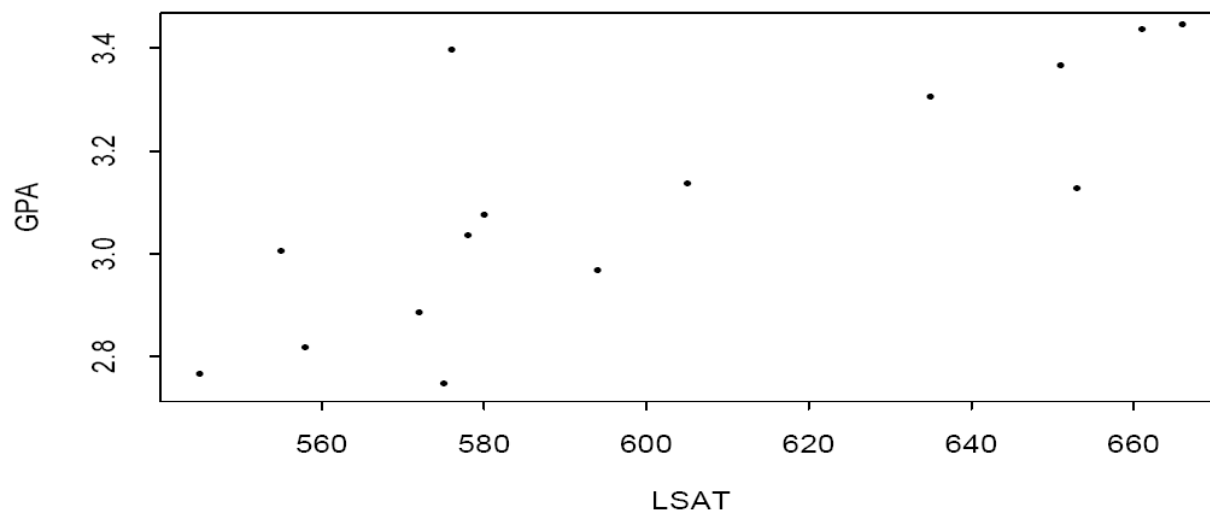
□ 还有其他一些计算置信区间的方法

◆ 如枢轴置信区间:  $C_n = 2\hat{T} - T_{1-\alpha/2}^*, 2\hat{T} - T_{\alpha/2}^*$ ,

# 例：Bootstrap置信区间

- 例：Bootstrap方法的发明者Bradley Efron给出了下列用语解释Bootstrap方法的例子。这些数据是LAST分数（法学院的入学分数）和GPA。计算相关系数及其标准误差。

LSAT (Y)	576	635	558	578	666	580	555	661
	651	605	653	575	545	572	594	
GPA (Z)	3.39	3.30	2.81	3.03	3.44	3.07	3.00	3.43
	3.36	3.13	3.12	2.74	2.76	2.88	2.96	



## 例8.6（续）

□ 相关系数的定义为：

$$\theta = \frac{\iint (y - \mu_Y)(z - \mu_Z) dF_{y,z}}{\sqrt{\int (y - \mu_Y)^2 dF_y \int (z - \mu_Z)^2 dF_z}}$$

□ 相关系数的嵌入式估计量为：

$$\hat{\theta} = \frac{\sum_i (Y_i - \bar{Y})(Z_i - \bar{Z})}{\sqrt{\sum_i (Y_i - \bar{Y})^2 \sum_i (Z_i - \bar{Z})^2}} = 0.776$$

□ Bootstrap得到的相关系数插入估计的标准误差为：

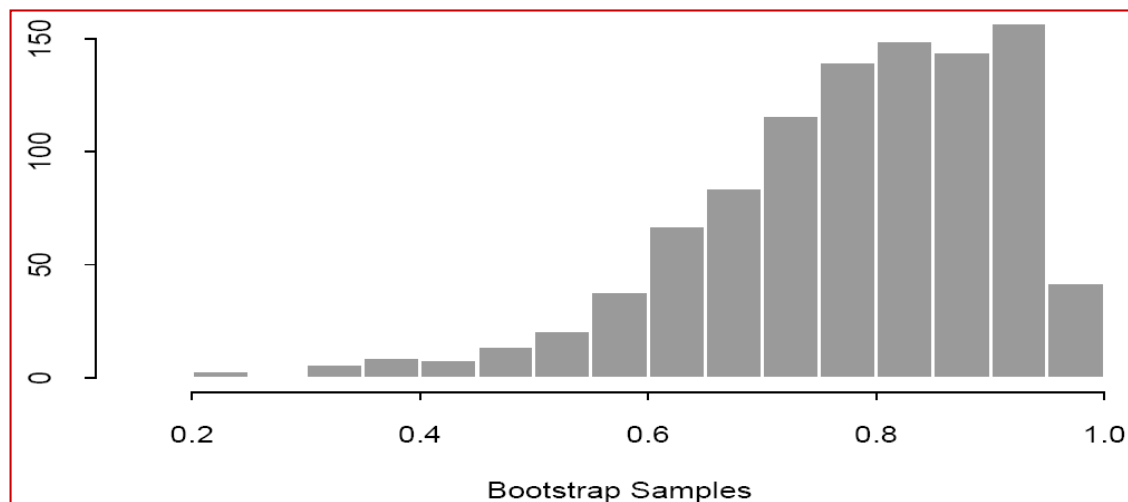
$B$	25	50	100	200	400	800	1600	3200
$se_{boot}$	0.140	0.142	0.151	0.143	0.141	0.137	0.133	0.132

标准误差趋向稳定于

$$se_{boot} \approx 0.132$$

## 例8.6 （续）

- 当 $B=1000$ 时,  $se_{boot} \approx 0.137$
- $\hat{\theta}_1^*, \dots, \hat{\theta}_B^*$  的直方图为下图, 可近似为从 $\hat{\theta}$  的分布采样



- 95%的正态区间为:  $0.78 \pm 2se = 0.51, 1.00$
- 95%的百分点区间为:  $0.46, 0.96$
- 当大样本情况下, 这两个区间趋近于相同

# 非参数bootstrap过程总结

□ 对原始样本数据  $X = X_1, \dots, X_n$  进行重采样, 得到  $B$  个 bootstrap 样本  $X_b^* = X_1^*, \dots, X_n^*$ , 其中  $b=1, \dots, B$

□ 对每个 bootstrap 样本  $X_b^*, b=1, \dots, B$ , 计算其对应的统计量的值 (bootstrap 复制)

$$T_{n,b}^* = g(X_b^*) = g(X_{1,b}^*, \dots, X_{n,b}^*)$$

□ 根据 bootstrap 复制  $T_{n,b}^*, b=1, \dots, B$ , 计算其方差、偏差和置信区间等

□ 称为非参数 bootstrap 方法, 因为没有对  $F$  的先验 (即  $F$  的知识仅从样本数据中获得)



- 统计量/统计函数:  $T = T(F)$
- ◆ 没有对 $F$ 的先验,  $F$ 的知识仅从样本数据中获得 (CDF估计), 统计函数的估计变为嵌入式估计

$$T = T(F), \quad T_n = T(\hat{F}_n) = g(X_1, \dots, X_n)$$

- ◆ 真实世界:  $F_n \Rightarrow X_1, \dots, X_n \Rightarrow T_n = g(X_1, \dots, X_n)$
- ◆ Bootstrap世界:  $F_n \Rightarrow X_1^*, \dots, X_n^* \Rightarrow T_n^* = g(X_1^*, \dots, X_n^*)$

- ◆ 如方差计算中, 发生了两个近似

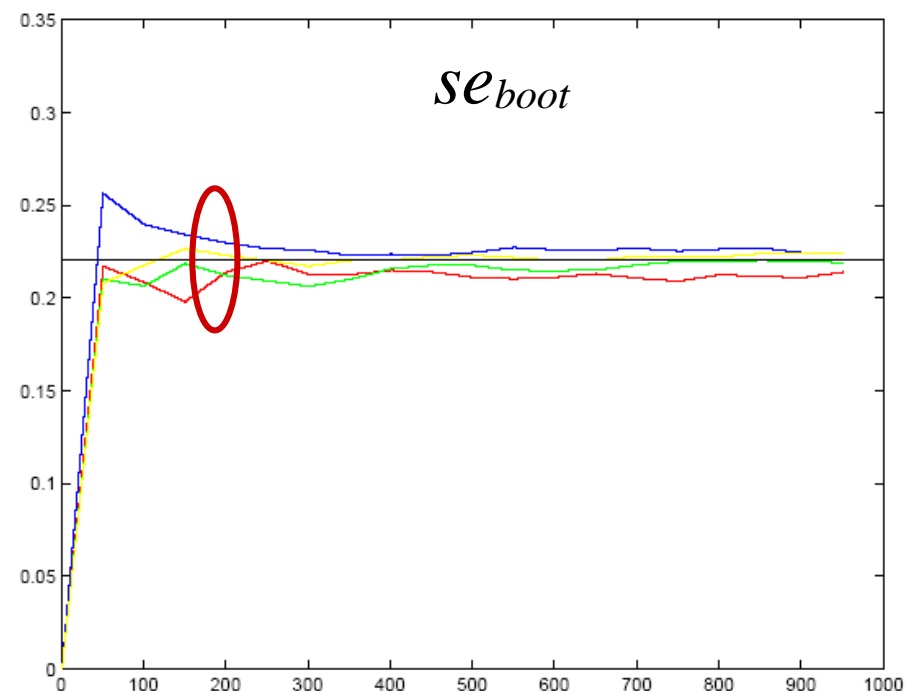
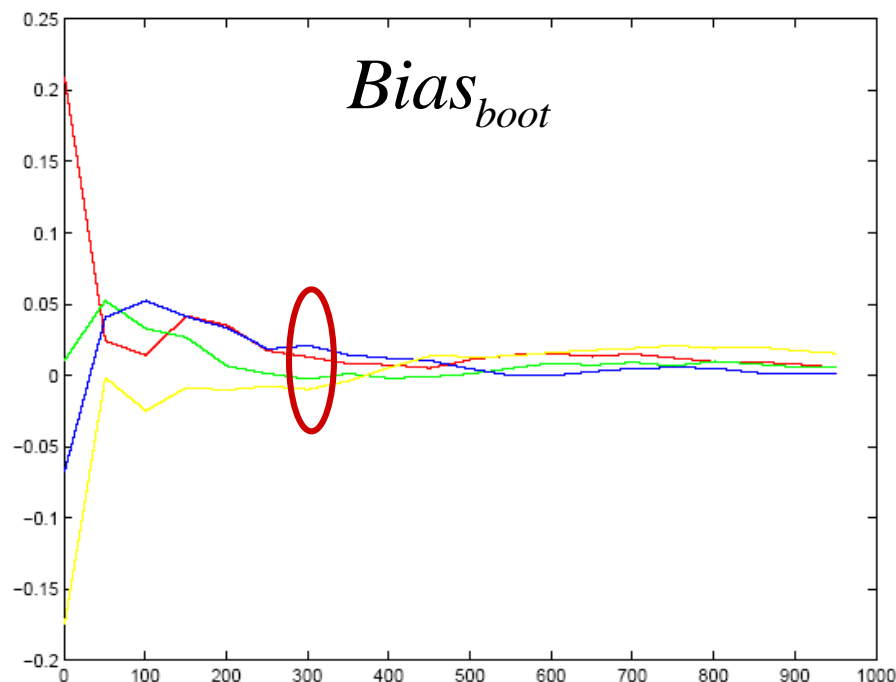
$$O(1/\sqrt{n}) \qquad O(1/\sqrt{B})$$

$$\mathbb{V}_F(T_n) \approx \mathbb{V}_{F_n}(T_n) \approx v_{boot}$$

- ◆ 近似的程度与样本数目 $n$ 及bootstrap样本的数目 $B$ 有关

# Bootstrap的收敛性

- 例：混合高斯模型：  $T_n = \bar{X}_n$
- ◆  $F: F(X) = 0.2N(1, 2) + 0.8N(6, 1)$
- ◆  $n=100$ 个观测样本  $\mathbf{X} = X_1, \dots, X_{100}$
- ◆ 4次试验得到不同  $B$  的偏差和方差的结果



# Bootstrap的收敛性

---

□  $B$ 的选择取决于

◆ 计算机的可用性

$$T_{n \text{ boot}} = \frac{1}{B} \sum_{b=1}^B T_{n,b}^*$$

◆ 问题的类型：标准误差/偏差/置信区间/...

◆ 问题的复杂程度

- 在一次bootstrap采样中，某些原始样本可能没被采到，另外一些样本可能被采样多次
- 在一个bootstrap样本集中不包含某个原始样本  $X_i$  的概率为

$$\mathbb{P} X_j \neq X_i, j=1, \dots, n = \left(1 - \frac{1}{n}\right)^n \approx e^{-1} \approx 0.368$$

- ◆ 一个bootstrap样本集包含了大约原始样本集的  $1 - 0.368 = 0.632$ ，另外0.368的样本可能永远不会被抽到。
- ◆ 这个在集成学习中成为out-of-bag (oob)。

# Bootstrap（参数/非参数）不适合的情况 python™

---

- 小样本（ $n$ 太小）
  - ◆ 原始样本不能很好地代表总体分布
  - ◆ Bootstrap只能覆盖原始样本的一部分，带来更大的偏差
- 结构间有关联
  - ◆ 如时间/空间序列信号
  - ◆ 因为bootstrap假设个样本间独立
- 脏数据
  - ◆ 奇异点(outliers)给估计带来了变化

- Bootstrap方法并不总是最佳的。其中一个主要原因是bootstrap样本是从 $\hat{F}_n$ 产生而不是从 $F$ 产生。
- 问题：能完全从 $F$ 采样或重采样吗？
  - ◆ 如果样本数目为 $n$ ，答案是否定的！
  - ◆ 若样本数目为 $m$  ( $m < n$ )，则可以从 $F$ 中找到数目为 $m$ 的采样/重采样，通过从原始样本 $X$ 得到不同的子集就可以！
- 寻找原始样本的不同子集相当于从观测  $X_1, \dots, X_n$  进行无放回采样，得到数目为 $m$ 的重采样样本（在此称为子样本）这就是jackknife的基本思想。

□ Jackknife样本定义为：一次从原始样本中留出一个样本  $X_i, i=1, \dots, n$  :  $X = X_1, \dots, X_n$

$$\begin{aligned} X &= X_1, \dots, X_{i-1}, \boxed{X_i}, X_{i+1}, \dots, X_n \\ X_{(-i)} &= X_1, \dots, X_{i-1}, \quad X_{i+1}, \dots, X_n \end{aligned}$$

- ◆ Jackknife样本中的样本数目为  $m=n-1$
- ◆ 共有  $n$  个不同的jackknife样本
- ◆ 无需通过采样手段得到 jackknife样本

谢谢Q/A