

ブートストラップ法の概要

秋田 智之

平成 18 年 9 月 26 日

1 導入

学部 of 確率統計で習った推測統計は「母集団から無作為抽出された標本 X_1, X_2, \dots, X_n をもとに母集団の母平均などの母数を推測」するものであった。

これに対し、ブートストラップ法は「標本の分布は母集団の分布を再現している」という考えのもとで、この“標本の分布”から再度(復元)抽出して得られたデータ $X_1^*, X_2^*, \dots, X_n^*$ をもとに母集団の母数を推測する方法である。

ここでは次節以降用いられる記号の意味を挙げておく：

- F_0 : 母集団の分布 (函数)
- X_1, \dots, X_n : 母集団 (F_0) から抽出された標本
- $\mathcal{X} = \{X_1, \dots, X_n\}$
- F_1 : 標本の分布 (函数)
- X_1^*, \dots, X_n^* : 標本 (F_1) から (復元) 抽出された再標本 (リサンプル)
- $\mathcal{X}^* = \{X_1^*, \dots, X_n^*\}$
- θ : 母集団の母数 (パラメータ)
- $\hat{\theta}$: θ の推定量 . $\theta[\mathcal{X}], \theta(F_1)$ と書くこともある
- F_2 : 再標本の分布 (函数)

2 原理

幾分抽象的であるがブートストラップ法の原理を紹介する。

t をパラメータとして汎函数 $(F_0, F_1) \mapsto f_t(F_0, F_1)|F_0$ を考える。これについて

$$E[f_t(F_0, F_1)|F_0] = 0 \quad (1)$$

となるような $t = t_0$ を求めるという問題を考えてみる。しかし母集団 F_0 が未知なのでこれを解くことは

できない。ここでブートストラップ法の理論が登場する。これは上の (F_0, F_1) を (F_1, F_2) で置き換えて

$$E[f_t(F_1, F_2)|F_1] = 0 \quad (2)$$

となるような $t = \hat{t}_0$ でもって t_0 を推定しようとするものである。 F_1 は既知なのでこれは(厳密にもしくは数値的に)解くことができる。

次に f_t の具体的な形を例の中で見ていく。

例 1 (バイアスを修正する問題). f_t を次で定義する：

$$f_t(F_0, F_1) = \theta(F_1) - \theta(F_0) + t \quad (3)$$

このとき (1) は次のようになる。

$$\theta(F_0) = E[\theta(F_1)|F_0] + t_0$$

即ち t は母数 θ と推定量 $\hat{\theta}$ の偏差を表している。

例 2 (信頼区間を求める問題). f_t を次で定義する：

$$f_t(F_0, F_1) = I\{\theta(F_1) - t \leq \theta(F_0) \leq \theta(F_1) + t\} - 0.95 \quad (4)$$

このとき (1) は次のようになる。

$$P(\theta(F_1) - t_0 \leq \theta(F_0) \leq \theta(F_1) + t_0) = 0.95$$

ここでは t は母数 θ の 95% 信頼区間を構成するときの誤差限界である。

次に標本の分布 (函数) F_1 の構成法について述べる。まず母集団から無作為に n 個の標本 X_1, \dots, X_n を抽出する。このとき母集団の分布の形が正規分布のようなパラメトリックな分布と仮定できるときと、そうでないときで構成法は若干異なる。

ノンパラメトリックの場合。このとき F_1 は経験分布函数である。即ち

$$F_1(x) = \frac{1}{n} \#\{i | X_i \leq x\}$$

パラメトリックの場合．この場合，数個のパラメータが分かれば分布が再現できる．そこで母集団の母数 θ を λ から求めた推定量 $\hat{\theta}$ で置き換えた分布を F_1 にする．

例 3 (母集団が正規分布と仮定できるときの F_1 の構成法)．母数 (μ, σ^2) の代わりにそれぞれの最尤推定量

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

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

を用いた正規分布 $N(\bar{X}, S^2)$ を F_1 にする．

この例のように母数の推定量には最尤推定量がよく用いられる．ちなみに S^2 の代わりに不偏分散

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

を用いても問題はない．

ところでバイアス修正や信頼区間の問題でデータ数 n によっては (2) を (丁度) 満たす t が存在しないことがある．このときは近い値で代用する．

例 4 (信頼区間の問題)．このとき \hat{t}_0 は次で代用する．

$$\inf\{t | P(\theta(F_2) - t \leq \theta(F_1) \leq \theta(F_2) + t | F_1) - 0.95 \geq 0\}$$

3 例

3.1 バイアス修正

f_t を次で定義する：

$$f_t(F_0, F_1) = \theta(F_1) - \theta(F_0) + t$$

このとき標本方程式 (2) は

$$E[f_t(F_1, F_2) | F_1] = E\{\theta(F_2) - \theta(F_1) + t | F_1\} = 0$$

となる．ここで $\theta(F_1)$ は F_1 が既知であり，また t はパラメータなので $E\{\cdot\}$ の外に出すことができる．

$$E\{\theta(F_2) | F_1\} - \theta(F_1) + t = 0$$

従って

$$t = \hat{t}_0 = \theta(F_1) - E\{\theta(F_2) | F_1\}$$

であるので，バイアスを修正した θ の推定量は次のようになる：

$$\hat{\theta}_1 = \hat{\theta} + \hat{t}_0 = \theta(F_1) + \hat{t}_0 = 2\theta(F_1) - E\{\theta(F_2) | F_1\} \quad (5)$$

近似計算． $E\{\theta(F_2) | F_1\}$ を厳密に求めることができないときでもモンテカルロ法で近似値を得ることができる．まず F_1 から独立に再標本群 $\mathcal{X}_1^*, \mathcal{X}_2^*, \dots, \mathcal{X}_B^*$ をとる．ここで F_{2b} ($b = 1, 2, \dots, B$) を次のように作る．

ノンパラメトリックの場合． \mathcal{X}_b^* の経験分布関数を F_{2b} にする．

パラメトリックの場合．再標本 \mathcal{X}_b^* から計算された母数の推定値 $\hat{\lambda} = \lambda[\mathcal{X}_b^*]$ をパラメータとする分布を F_{2b} とする．

このとき大数の法則により $B \rightarrow \infty$ のとき

$$\frac{1}{B} \sum_{b=1}^B \theta(F_{2b}) \rightarrow E\{\theta(F_2) | F_1\}$$

となる．よって十分大きい B を用いて

$$\frac{1}{B} \sum_{b=1}^B \theta(F_{2b})$$

を $E\{\theta(F_2) | F_1\}$ にする．

$E[\theta(F_2) | F_1]$ が計算できる例．ここでは次のパラメータ

$$\theta(F_0) = \mu^3 = \left\{ \int x dF_0 \right\}^3$$

の (ブートストラップ) 推定値

$$\theta(F_1) = \bar{X}^3 = \left\{ \int x dF_1 \right\}^3$$

のバイアスを修正する問題を考える．

$E\{\theta(F_1) | F_0\}$ を計算すると

$$E\{\theta(F_1) | F_0\} = E \left\{ \frac{1}{n} \sum_{i=1}^n X_i \right\}^3$$

$$= E \left\{ \mu + \frac{1}{n} \sum_{i=1}^n (X_i - \mu) \right\}^3$$

$$\begin{aligned}
&= E \left\{ \mu^3 + 3\mu^2 \frac{1}{n} \sum_{i=1}^n (X_i - \mu) \right. \\
&\quad \left. + 3\mu \left\{ \frac{1}{n} \sum_{i=1}^n (X_i - \mu) \right\}^2 + \frac{1}{n^3} \left\{ \sum_{i=1}^n (X_i - \mu) \right\}^3 \right\} \\
&= E \left\{ \mu^3 + 3\mu^2 \frac{1}{n} \sum_{i=1}^n (X_i - \mu) \right. \\
&\quad \left. + 3\mu \frac{1}{n^2} \left\{ \sum_{i=1}^n (X_i - \mu) \right\}^2 \right. \\
&\quad \left. - \sum_{i \neq j} (X_i - \mu)(X_j - \mu) \right\} \\
&\quad \left. + \frac{1}{n^3} \left\{ \sum_{i=1}^n (X_i - \mu) \right\}^3 \right. \\
&\quad \left. + \sum_{i=1}^n (X_i - \mu) \sum_{j \neq i} \sum_{k \neq i} (X_j - \mu)(X_k - \mu) \right\} \\
&= \mu^3 + \frac{3}{n} \mu \sigma^2 + \frac{1}{n^2} \gamma
\end{aligned}$$

ここで $E(X_i - \mu) = 0$ を使った．また $\sigma^2 = E\{(X_1 - \mu)^2\}$, $\gamma = E\{(X_i - \mu)^3\}$ である．

例 5 (母集団の分布が未知のとき)．ブートストラップ法の原理を使って (6) を書き直すと次のようになる．

$$E\{\theta(F_2)|F_1\} = \bar{X}^3 + \frac{3}{n} \bar{X} \hat{\sigma}^2 + \frac{1}{n^2} \hat{\gamma}$$

ここで

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i, \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2, \hat{\gamma} = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^3 \quad \hat{\theta}_1 = 2\theta(F_1) - E\{\theta(F_2)|F_1\} = \bar{X}^3 \left(1 - \frac{3}{n} - \frac{2}{n^2}\right) \quad (10)$$

である．従ってバイアスを修正したブートストラップ推定量は (5) より

$$\begin{aligned}
\hat{\theta}_1 &= 2\theta(F_1) - E\{\theta(F_2)|F_1\} \\
&= 2\bar{X}^3 - \left(\bar{X}^3 + \frac{3}{n} \bar{X} \hat{\sigma}^2 + \frac{1}{n^2} \hat{\gamma}\right) \\
&= \bar{X}^3 - \frac{3}{n} \bar{X} \hat{\sigma}^2 - \frac{1}{n^2} \hat{\gamma}
\end{aligned} \quad (7)$$

となる．

母集団の分布の型が既知のときはもっと精密な推定ができる．

例 6 (母集団が正規分布のとき)．母集団が $N(\mu, \sigma^2)$ のとき $\gamma = E\{(\bar{X} - \mu)^3\} = 0$ となるので (6) は

$$E\{\theta(F_1)|F_0\} = \mu^3 + \frac{3}{n} \mu \sigma^2$$

となる．次に母数 (μ, σ^2) を最尤推定量 $(\bar{X}, \hat{\sigma}^2)$ で推定し, $N(\bar{X}, \hat{\sigma}^2)$ から標本を抽出することを考える．このときも先程と同様にブートストラップ法の原理から

$$E\{\theta(F_2)|F_1\} = \bar{X}^3 + \frac{3}{n} \bar{X} \hat{\sigma}^2$$

であり, バイアスを修正した推定量は

$$\hat{\theta}_1 = 2\theta(F_1) - E\{\theta(F_2)|F_1\} = \bar{X}^3 - \frac{3}{n} \bar{X} \hat{\sigma}^2 \quad (8)$$

となる．

なお, $\hat{\sigma}^2$ の代わりに不偏分散 $\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ を使って, $N(\bar{X}, \hat{\sigma}^2)$ からの標本抽出を考えると

$$\hat{\theta}_1 = \bar{X}^3 - \frac{3}{n} \bar{X} \hat{\sigma}^2 \quad (9)$$

となる．これを用いてもよい．

例 7 (母集団が指数分布のとき)．今度は母集団が平均 μ の指数分布 $Ex(\mu)$ をときを考える．このとき $\sigma^2 = \mu^2$, $\gamma = 2\mu^3$ なので (6) は

$$E\{\theta(F_1)|F_0\} = \mu^3 \left(1 + \frac{3}{n} + \frac{2}{n^2}\right)$$

となる．ブートストラップ法の原理に従って μ を最尤推定量 \bar{X} で推定して (2) は

$$E\{\theta(F_2)|F_1\} = \bar{X}^3 \left(1 + \frac{3}{n} + \frac{2}{n^2}\right)$$

となるので, バイアスを修正した推定量は

$$\hat{\theta}_1 = 2\theta(F_1) - E\{\theta(F_2)|F_1\} = \bar{X}^3 \left(1 - \frac{3}{n} - \frac{2}{n^2}\right) \quad (10)$$

である．

3.2 信頼区間

次に信頼区間を求める問題を考える． f_t を次で定義する：

$$f_t(F_0, F_1) = I\{\theta(F_1) - t \leq \theta(F_0) \leq \theta(F_1) + t\} - 0.95$$

このとき (1) は

$$\begin{aligned}
&E\{f_t(F_0, F_1)|F_0\} \\
&= 1 \cdot P(\theta(F_1) - t \leq \theta(F_0) \leq \theta(F_1) + t) - 0.95 = 0
\end{aligned}$$

即ち

$$P(\theta(F_1) - t \leq \theta(F_0) \leq \theta(F_1) + t) = 0.95$$

となる．ブートストラップ法の原理に基づいてこれを書き換えると

$$P(\theta(F_2) - t \leq \theta(F_1) \leq \theta(F_2) + t) = 0.95 \quad (11)$$

となり，この解を $t = t_0$ とすると $(\theta(F_1) - t_0, \theta(F_1) + t_0)$ は 95 % 信頼区間となる．これを両側パーセンタイル区間という．

なお，(11) の厳密な解が存在しないときは

$$\inf\{t : P(\theta(F_2) - t \leq \theta(F_1) \leq \theta(F_2) + t | F_1) - 0.95 \geq 0\}$$

を t_0 とする．

両側パーセンタイル区間以外の信頼区間も同様に作ることができる．例えば

$$P(\theta(F_1) \leq \theta(F_2) - t | F_1) - 0.025 = 0$$

の解 \hat{t}_{01} および

$$P(\theta(F_1) \leq \theta(F_2) + t | F_1) - 0.975 = 0$$

の解 \hat{t}_{02} を用いた区間 $(\theta(F_1) - \hat{t}_{01}, \theta(F_1) + \hat{t}_{02})$ は両側等裾信頼区間と呼ばれ，また

$$P(\theta(F_1) \leq \theta(F_2) + t | F_1) - 0.95 = 0$$

の解 \hat{t}_{03} を用いた区間 $(-\infty, \theta(F_1) + \hat{t}_{03})$ は片側信頼区間と呼ばれる．

次に母集団がパラメトリックな場合を詳しく見ていく．

例 8 (母集団が正規分布のとき)．このときは正規分布のパラメータ (μ, σ^2) を最尤推定量 $(\bar{X}, \hat{\sigma}^2)$ で推定することになる．このとき (11) は $\theta(F_1) = \bar{X}$, $\theta(F_2) = \bar{X}^*$ を用いて

$$P(|\bar{X}^* - \bar{X}| \leq t | F_1) = 0.95 \quad (12)$$

と書ける．ここで \bar{X}^* は正規分布 $N(\bar{X}, \hat{\sigma}^2/n)$ に従うので学生化統計量

$$\frac{\bar{X}^* - \bar{X}}{\sqrt{\hat{\sigma}^2/n}}$$

は標準正規分布 $N(0, 1)$ に従う．そこで N を $N(0, 1)$ に従う確率変数とし

$$P(|N| \leq x_\alpha) = \alpha$$

で x_α を定義すると (12) は

$$P\left(\left|\frac{\hat{\sigma}}{\sqrt{n}}N\right| \leq t \mid F_1\right) = 0.95$$

となり，従ってブートストラップ信頼区間は

$$\left(\bar{X} - \frac{\hat{\sigma}}{\sqrt{n}}x_{0.95}, \bar{X} + \frac{\hat{\sigma}}{\sqrt{n}}x_{0.95}\right)$$

である．

ここで被覆誤差を定義する．

定義 1 (被覆誤差)．パラメータ μ の信頼区間 I の被覆誤差を次で定義する：

$$P(\mu \in I) - (1 - \alpha)$$

この例の場合，被覆誤差は

$$P\left(\bar{X} - \frac{\hat{\sigma}}{\sqrt{n}}x_{0.95} \leq \mu \leq \bar{X} + \frac{\hat{\sigma}}{\sqrt{n}}x_{0.95}\right) - 0.95$$

であり，これが $O(n^{-1})$ であることが知られている．

例 9 (母集団が指数分布のとき)．このときも指数分布のパラメータ λ を最尤推定量 \bar{X} で推定すると (11) は $\theta(F_1) = \bar{X}$, $\theta(F_2) = \bar{X}^*$ を用いて

$$P(|\bar{X}^* - \bar{X}| \leq t | F_1) = 0.95 \quad (13)$$

となる．ここで \bar{X}^* について考えてみる． X_1^*, \dots, X_n^* は独立に平均 \bar{X} の指数分布に従うので，これらの和 $X_1^* + \dots + X_n^*$ はガンマ分布 $Ga(n, \bar{X})$ に従う．従って

$$\frac{X_1^* + \dots + X_n^*}{\bar{X}} = \frac{n\bar{X}^*}{\bar{X}}$$

はガンマ分布 $Ga(n, 1)$ に従う．そこで Y を $Ga(n, 1)$ に従うような確率変数とすると (13) は

$$P(\bar{X}|n^{-1}Y - 1| \leq t | F_1) = 0.95$$

となる．そこで

$$P(|n^{-1}Y - 1| \leq y_\alpha) = \alpha$$

で y_α を定義するとブートストラップ信頼区間は

$$(\bar{X} - y_{0.95}\bar{X}, \bar{X} + y_{0.95}\bar{X})$$

である．このときも被覆誤差は $O(n^{-1})$ であることが知られている．

これまでの区間推定の方法はパーセンタイル法と呼ばれる。これは簡便な方法でよいのだが、例えば母集団が正規分布のときの母集団方程式

$$P(|\theta(F_1) - \theta(F_0)| \leq t) = 0.95$$

について $|\theta(F_1) - \theta(F_0)|$ の分布はこれから推定する μ 以外に未知のパラメータ σ^2 にも依存している。このように分布が推定したいパラメータ以外の未知のパラメータに依存しているときはどうしても誤差が大きくなる。そこで $\theta(F_1) - \theta(F_0)$ の代わりに F_1 から求められる量 $\hat{\tau} = \tau(F_1)$ を用いて $T = (\theta(F_1) - \theta(F_0))/\hat{\tau}$ とスケール変換して未知のパラメータに依存しないようにすることを考える。

定義 2 (pivotal). $T(\theta, \mathcal{X}) = T(\theta, X_1, \dots, X_n)$ を未知母数とデータを共に含む函数とする。

(1) もし $\forall \theta, \theta' \in \Theta$ に対し $T(\theta, \mathcal{X})$ と $T(\theta', \mathcal{X})$ の分布が等しいとき (つまり T の分布が未知母数に依存していないとき), T は *pivotal* であるという。

(2) もし $a_n T + b_n$ が θ に依存しない退化しない分布を持つとき T は *asymptotically pivotal* であるという。

例 10 (母集団が正規分布のとき). f_t を次で定義する:

$$f_t(F_0, F_1) = I\{|\theta(F_1) - \theta(F_0)| \leq t\tau(F_1)\} - 0.95$$

ここでスケール推定量 $\hat{\tau} = \tau(F_1)$ は

$$\hat{\tau} = \hat{\sigma} = \sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2}$$

である。このとき母集団方程式 (1) は

$$P\left(\left|\frac{\theta(F_1) - \theta(F_0)}{\sigma(F_1)}\right| \leq t|F_0\right) = 0.95$$

であり、標本方程式 (2) は

$$P\left(\left|\frac{\theta(F_2) - \theta(F_1)}{\sigma(F_2)}\right| \leq t|F_1\right) = 0.95 \quad (14)$$

となる。ここで

$$\theta(F_2) = \bar{X}^*, \theta(F_1) = \bar{X}, \sigma(F_2)^2 = \frac{1}{n} \sum_{i=1}^n (X_i^* - \bar{X}^*)^2$$

である。また $\hat{\sigma}^2$ と不偏分散

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

について

$$\hat{\sigma}^2 = \frac{n-1}{n} \hat{\sigma}^2$$

の関係があることに注意すると

$$\frac{\bar{X}^* - \bar{X}}{\sqrt{\hat{\sigma}^2}} = \frac{\bar{X}^* - \bar{X}}{\sqrt{(n-1)\frac{\hat{\sigma}^2}{n}}} = \frac{1}{\sqrt{n-1}} T_{n-1}$$

となる。ここで T_{n-1} は自由度 $n-1$ の t 分布に従う確率変数である。これより (14) は

$$P\left(\frac{1}{\sqrt{n-1}} |T_{n-1}| \leq t|F_1\right) = 0.95$$

そこで $w_\alpha = w_\alpha(n)$ を

$$P(|T_{n-1}| \leq w_\alpha) = \alpha$$

で定義すると解は

$$t = \hat{t}_0 = (n-1)^{-1/2} w_{0.95}$$

であり、ブートストラップ信頼区間は

$$\left(\bar{X} - \frac{1}{\sqrt{n-1}} w_{0.95} \hat{\sigma}, \bar{X} + \frac{1}{\sqrt{n-1}} w_{0.95} \hat{\sigma}\right)$$

となる。これをパーセンタイル t 区間という。

例 11 (母集団が指数分布のとき). f_t , 母集団方程式および標本方程式は正規分布の場合と同じ。スケール推定量を

$$\hat{\tau} = \tau(F_1) = \theta(F_1) = \bar{X}$$

とおく。ここで

$$Y = \frac{X_1^* + \dots + X_n^*}{\bar{X}} = \frac{n\bar{X}^*}{\bar{X}}$$

とおくと Y はガンマ分布 $Ga(n, 1)$ に従う。また

$$\frac{\bar{X}^* - \bar{X}}{\bar{X}^*} = \frac{(Y/n)\bar{X} - \bar{X}}{(Y/n)\bar{X}} = \frac{(Y/n) - 1}{Y/n}$$

であるので

$$P\left(\left|\frac{(Y/n) - 1}{Y/n}\right| \leq t|F_1\right) = 0.95$$

となる。これを満たす t を $t_0 = w_{0.95}$ とおくと、信頼区間は

$$(\bar{X} - w_{0.95} \bar{X}, \bar{X} + w_{0.95} \bar{X})$$

となる。

両側パーセンタイル t 区間以外の信頼区間はパーセンタイル法のとおり同様に行うことができる。標本方程式

$$P\left(\frac{\theta(F_2) - \theta(F_1)}{\tau(F_2)} \leq t | F_1\right) = \alpha$$

の解を $t = \hat{v}_\alpha$ とする。ここで $\hat{\tau}^* = \tau(F_2)$ はブートストラップリサンプリングから計算されたスケール推定量である。このとき

$$(-\infty, \hat{\theta} - \hat{\tau}\hat{v}_{1-\alpha})$$

は片側パーセンタイル t 区間と呼ばれ、また

$$(\hat{\theta} - \hat{\tau}\hat{v}_{\alpha/2}, \hat{\theta} - \hat{\tau}\hat{v}_{(1-\alpha)/2})$$

は両側等裾パーセンタイル t 区間と呼ばれる。

モンテカルロ法による標本方程式の解法。ここでは標本方程式

$$P\left(\frac{\theta(F_2) - \theta(F_1)}{\tau(F_2)} \leq t | F_1\right) = \alpha$$

の解 $t = \hat{v}_\alpha$ をモンテカルロ・シミュレーションで求める方法を述べる。再標本群の個数 B を固定し $\nu = \alpha(B+1)$ となるような整数 ν をとる。ここで標本の分布 F_1 から B 個の再標本群 $\mathcal{X}_1^*, \dots, \mathcal{X}_B^*$ を独立に選ぶ。母集団分布がパラメトリックなときとそうでないときのそれぞれの場合で $F_{2,b}$ ($1 \leq b \leq B$) を次のように定義する。

ノンパラメトリックの場合。このとき $F_{2,b}$ は \mathcal{X}_b^* の経験分布関数である。

パラメトリックの場合。母集団の母数 θ を \mathcal{X}_b^* から求めた推定量 $\hat{\theta}$ で置き換えた分布を $F_{2,b}$ にする。

ここで

$$T_b^* = \frac{\theta(F_{2,b}) - \theta(F_1)}{\tau(F_{2,b})}$$

とおき、 $T_1^*, T_2^*, \dots, T_B^*$ を求める。ここで T_1^*, \dots, T_B^* を小さいほうから並べたものを

$$T_{(1)}^*, T_{(2)}^*, \dots, T_{(B)}^*$$

とし、 $\hat{v}_{\alpha,B} = T_{(\nu)}^*$ とおく。このとき \mathcal{X} の条件下で

$$\hat{v}_{\alpha,B} \rightarrow \hat{v}_\alpha \quad (B \rightarrow \infty)$$

が成り立つ。そこで \hat{v}_α を十分大きい B を用いて $\hat{v}_{\alpha,B}$ で代用する。

3.3 短い信頼区間

ここでは信頼区間を短くすることを考える。 $\hat{\tau} = \tau(F_1)$ をスケール推定量とし

$$P\left(\frac{\hat{\theta} - \theta}{\hat{\tau}} \leq t\right) = \alpha$$

の解を $t = v_\alpha$ とする。 $0 < \alpha < 1 - \alpha$ に対し

$$J(a) = (\hat{\theta} - \hat{\tau}v_{\alpha+a}, \hat{\theta} - \hat{\tau}v_a)$$

は θ の水準 α の信頼区間を表している。即ち

$$P\{\theta \in J(a)\} = \alpha$$

である。ここで $J(a)$ の長さは $v_{\alpha+a} - v_a$ に比例するので $J(a)$ の長さを最小にするには $v_{\alpha+a} - v_a$ を最小にするような $a = a_1$ を用いて

$$J_{SH} = (\hat{\theta} - \hat{\tau}v_{\alpha+a_1}, \hat{\theta} - \hat{\tau}v_{a_1})$$

とすればよい。そこで f_t を

$$f_t(F_0, F_1) = I\{\theta(F_1) - \theta(F_0) \leq \tau(F_1)t\} - \alpha$$

としたときの母集団方程式 (1) の解は $v_\alpha = v_\alpha(F_0)$ である。そこでこの f_t の標本方程式 (2) を解くことによって v_α のブートストラップ推定量 $\hat{v}_\alpha = v_\alpha(F_1)$ を得ることができる。そこで $\hat{v}_{\alpha+a} - \hat{v}_a$ を最小にするような a を \hat{a}_1 とおくと J_{SH} のブートストラップ信頼区間は

$$\hat{J}_{SH} = (\hat{\theta} - \hat{\tau}\hat{v}_{\alpha+\hat{a}_1}, \hat{\theta} - \hat{\tau}\hat{v}_{\hat{a}_1})$$

となる。

信頼区間を短くすることにより母数の推定がより精密にできる。しかし $J(a)$ の長さは θ を被覆することに失敗したとき短くなり、 θ を被覆したときに長くなる特徴がある。そこで

$$l(a) = E\{\hat{\tau}(v_{\alpha+a} - v_a) | \theta \in J(a)\} \\ = \alpha^{-1}(v_{\alpha+a} - v_a) E\{\hat{\tau}I(\hat{\theta} - \hat{\tau}v_{\alpha+a} \leq \theta \leq \hat{\theta} - \hat{\tau}v_a)\}$$

を最小にするような $a = a_2$ を用いた

$$J_{LS} = (\hat{\theta} - \hat{\tau}v_{\alpha+a_2}, \hat{\theta} - \hat{\tau}v_{a_2})$$

がよりいい信頼区間といえる。 a_2 のブートストラップ推定量 \hat{a}_2 は

$$\hat{\lambda}(a) = \alpha^{-1}(\hat{v}_{\alpha+a} - \hat{v}_a) \\ \cdot E[\tau(F_2)I\{\tau(F_2)\hat{v}_a \leq \theta(F_2) - \theta(F_1) \leq \tau(F_2)\hat{v}_{\alpha+a}\} | F_1]$$

を最小化することによって得ることができる。