（Ｊ言語研究会 2005/3.26）

ブートストラップ法(Bootstrap method)について

帝京平成大学　鈴木義一郎

一般に、連続な分布関数$F(x)$に従う確率変数$X_1, X_2, \cdots\cdots X_n$に対して、大きさの順番に並べ直した$\{X_1^*, X_2^*, \cdots\cdots X_n^*\}$を「順序統計量」という。さらに

$$F_n(x)=\{no.of\ i; X_i \le x\}/n = i/n\,(X_i^* \le x < X_{i+1}^*)\ i=1,2,\cdots\cdots,n\ (X_0^*=0, X_{n+1}^*=\infty)$$

で定義されるものを「経験分布関数」と呼ばれている。

そこで、未知の分布関数$F(x)$の母集団からの大きさｎの標本を$\{X_1, X_2, \cdots\cdots X_n\}$として、$F(x)$のあるパラメータ$\theta$を統計量$\hat{\theta}_n=\hat{\theta}(X_1, X_2, \cdots\cdots X_n)$を用いて推定する問題を考える。このような推定量の評価尺度は、

①　$$B\{F)=E_F\{\hat{\theta}_n-\theta\},\quad V(F)=E_F\{[(\hat{\theta}_n-E_F\{\hat{\theta}_n\}^2]\},\quad SE\{F\}=\sqrt{V(F)}$$

のような「バイアス」と「分散」そして「標準誤差」である。さらに、推定量の標本分布が分かれば、パラメータθに対する信頼区間を構成することができる。

そこで、統計量$T_n=\sqrt{n}(\hat{\theta}_n-\theta)$の標本分布と$100\alpha$％点

$$H_n(x,F)=\Pr_F\{\sqrt{n}(\hat{\theta}_n-\theta) < x\},\quad x_n(\alpha)=\min\{x : H_n(x,F)\ge\alpha\}$$

を求めることができる。

さて、「ブートストラップ法」とは、これらの量の推定を個々の推定量に関して解析的に行う代わりに、コンピュータを利用して数値的に実行する統計手法で、次のような手順で実行することができる。

（１）未知の分布関数$F(x)$の母集団からの大きさｎの標本$\{X_1, X_2, \cdots\cdots X_n\}$を用いて経験分布関数$F_n(x)=F_n(x\,|\,X_1, X_2, \cdots\cdots X_n)$を求める。そして、①の評価尺度の推定値としては、$F(x)$を$F_n(x)$で置き換えた

②　$$B\{F_n)=E_{F_n}\{\hat{\theta}_n-\theta\},\quad V(F_n)=E_{F_n}\{[(\hat{\theta}_n-E_{F_n}\{\hat{\theta}_n\}^2]\},\quad SE\{F_n\}=\sqrt{V(F_n)}$$

で推定する。ここで、$\hat{\theta}_n=\hat{\theta}(X_1, X_2, \cdots\cdots X_n)$はパラメータ$\theta$の推定量である。

（２）次に、この経験分布関数$F_n(x)$を母集団のように考えて、ここから大きさｎの標本$X_n^*=\{X_1^*, X_2^*, \cdots\cdots X_n^*\}$を無作為抽出する。これを「ブートストラップ標本」という。このブートストラップ標本にもとづく推定量を$\hat{\theta}_n^*=\hat{\theta}(X_n^*)=\hat{\theta}(X_1^*, X_2^*, \cdots\cdots X_n^*)$とする。

（３）（２）の手順をＮ回繰り返すことにより、$\{X_n^*(i): i=1,2,\cdots\cdots,N\}$というＮ組のブートストラップ標本と対応する Ｎ 個の推定量$\{\hat{\theta}_n^*(i)=\hat{\theta}(X_n^*(i)): i=1,2,\cdots\cdots,N\}$を求めることができるから、②に対する値を次式で近似することができる。

③　$$B\{F_n^*)=\bar{\theta}_n^*-\hat{\theta}_n\ ,\quad V(F_n^*)=\frac{1}{N-1}\sum_{i=1}^{N}\{\hat{\theta}_n^*(i)-\bar{\theta}_n^*\}^2\quad(\bar{\theta}_n^*=\frac{1}{N}\sum_{i=1}^{N}\hat{\theta}_n^*(i)\})$$

ブートストラップ回帰分析

目的変数 y と p 個の説明変数 $\mathbf{x}=\left(x_1,x_2,\cdots\cdots,x_p\right)$に関して、n 個の観測値

$$\{(x_1,y_1),(x_2,y_2),\cdots\cdots,(x_n,y_n)\}$$

が与えられているとする。そして、y と $\mathbf{x}$ の間には

$$y_i=f(x_i;\beta)+\varepsilon_i \quad i=1,2,\cdots\cdots,n$$

のような関係式を仮定する。ここで、f の関数形は既知、β は未知のパラメータベクトルとする。さらに誤差項 $\varepsilon_1,\varepsilon_2,\cdots\cdots,\varepsilon_n$ は、互いに独立で同一の未知の確率分布に従うものとし、$E\{\varepsilon_i\}=0$ と仮定する。

①未知のパラメータベクトル β に対して、最小二乗法による推定値 $\hat{\beta}$ を用いて

$$e_i=y_i-f(x_i;\hat{\beta}) \quad i=1,2,\cdots\cdots,n$$

といった残差から経験分布関数 $F_n(x)=F_n(x\,|\,e_1,e_2,\cdots\cdots e_n)$ を求める。

②経験分布関数 $F_n(x)$ から大き n のブートストラップ標本 $\{e_1^{\,*},e_2^{\,*},\cdots\cdots,e_n^{\,*}\}$ から

$$y_i^{\,*}=f(x_i;\hat{\beta})+e_i^{\,*} \quad i=1,2,\cdots\cdots,n$$

という値を算出する。

③$\sum_{i=1}^{n}\{y_i^{\,*}-f(x_i;\beta)\}^2$ を最小にするような β を $\hat{\beta}^*_{(1)}$ とする。

④以上の②と③の手順を N 回繰り返して得られる $\{\hat{\beta}^*_{(1)},\hat{\beta}^*_{(2)},\cdots\cdots,\hat{\beta}^*_{(N)}\}$ から推定値 $\hat{\beta}$ に対する誤差評価を行うことができる。

```
nrnd0=:3 : '_6+(+/?(12,y.)$1000)%1000'
nrndh=:3 : '<.49.5+10*nrnd0 y.'     NB.平均 50、標準偏差 10 の整数乱数
```

平均 50、標準偏差 10 の 100 個の整数乱数を 5 組以下のように生成する：

```
D1=: nrndh 100
D2=: nrndh 100
D3=: nrndh 100
D4=: nrndh 100
D5=: nrndh 100

5 20 $ D1
```

61 49 42 45 45 35 22 52 65 53 32 48 45 41 41 40 41 40 43 47
67 39 52 51 51 67 35 40 55 39 47 49 44 35 58 58 37 60 57 50
56 45 55 29 54 43 57 63 49 52 37 51 56 34 39 52 41 43 49 30
48 46 59 62 54 75 44 51 42 43 54 45 33 67 43 30 47 52 48 20
51 43 37 55 38 42 39 52 34 56 41 58 63 48 63 36 38 43 64 35

5 20 $ D2

44 36 59 53 49 53 58 51 41 51 48 58 58 35 56 48 40 57 55 54
63 45 53 56 63 45 56 39 38 47 49 32 40 56 58 54 51 44 32 52
37 42 59 44 43 35 57 59 42 57 52 63 40 44 46 47 52 49 53 35
26 58 50 54 30 53 62 55 43 54 45 52 48 49 49 62 45 56 60 50
53 52 37 52 47 57 56 54 45 48 50 48 59 65 52 60 31 50 50 37

5 20 $ D3

9 58 58 47 40 61 31 56 54 53 60 54 36 57 42 58 50 61 58 43
45 46 47 38 48 47 23 53 74 52 42 49 61 50 47 58 41 48 53 74
45 70 61 35 71 41 49 41 41 43 54 33 63 46 42 57 42 41 39 41
42 41 61 63 51 64 33 51 60 58 63 30 54 39 42 55 48 38 47 49
51 42 68 53 49 36 54 48 38 48 60 67 55 45 65 43 59 52 42 43

5 20 $ D4

51 58 57 49 31 36 55 46 39 53 29 63 47 59 43 78 32 45 54 51
54 61 44 48 42 55 40 47 44 54 46 37 45 47 59 58 57 54 52 35
62 48 33 41 43 35 53 63 56 43 65 53 57 51 52 57 55 45 54 40
44 45 51 43 32 49 66 50 46 30 45 53 49 49 30 43 51 41 46 61
47 50 58 53 55 54 30 51 36 35 43 45 41 34 39 45 50 52 44 41

5 20 $ D5

26 42 70 50 28 53 36 60 54 53 57 61 39 46 53 38 37 54 58 59
72 49 47 45 39 53 45 55 62 52 55 53 55 37 45 52 41 42 28 38
61 53 48 38 51 29 39 58 63 50 49 54 57 43 38 59 52 55 58 51
44 54 54 52 40 46 59 45 43 45 63 45 63 68 50 70 57 56 51 52
60 35 71 45 60 44 55 41 43 53 41 47 48 47 32 57 40 51 52 56

(mean;var)D1

47.12	105.206

(mean;var)D2

49.37	70.2331

(mean;var)D3

50.04	104.958

(mean;var)D4

47.88	83.7056

(mean;var)D5

49.8	92.78

```
bstp_s=:3 :'(?r$r=.#y.){y.'

     bstp=:4 :0
(sample=.([:?#$#){])r=.,."
mean=.+/%#
while.x.>#r
   do. r=.r,m,v=.mean*:s-m=.mean s=.sample y.
end.
mean r
)
     1000 bstp D1
47.0906 103.573
     1000 bstp D2
49.3773 69.6324
     1000 bstp D3
50.0633 103.972
     1000 bstp D4
47.8105 82.6012
     1000 bstp D5
49.8546 91.8698
     10000 bstp D1
47.1098 104.16
     10000 bstp D2
49.371 69.4345
     10000 bstp D3
50.057 103.674
     10000 bstp D4
47.884 82.7596
     10000 bstp D5
49.7947 91.9343
```