2011/04/20

福中公輔

ブートストラップ法

1　ブートストラップ法とは

例えば，母集団から得られた標本数が 7 個のような少数の値であったとします。通常，このような小標本下では，実用に耐えられる精度で誤差を見積もることはほぼ不可能であると考えられてきました。しかし，このブートストラップ法を利用すると，小さな標本から得られた母数の推定値の誤差を推定することが可能になり，その精度を調べることができるのです。

あるいは，観測変数が正規分布していなかったとします。このような場合には，最尤法を利用するのは難しいです。しかし，このブートストラップ法を利用すると，観測変数が正規分布しないような場合にも，標準誤差や信頼区間が評価できるようになるのです。

ブートストラップ法（Bootstrap Method）とは，k 個の小標本 $x_1, x_2, \ldots, x_k$ から，繰り返しを許してランダムに k 個の標本 $X_1^{*b}, X_2^{*b}, \ldots, X_k^{*b}$ を選び，平均や分散など母数の推定値を繰り返し求め，その分布から母数の確率分布や誤差を推定する方法です（吉岡，2007)。このブートストラップ法は，1979 年に Efron によって提案された比較的新しいリサンプリング手法です。このとき提案されたブートストラップ法は，確率分布を仮定せずにデータの経験分布（empirical distribution）に基づいて推論を行うので，ノンパラメトリック・ブートストラップ（Nonparametiric Bootstrap）法と呼ばれています。一方，データの確率モデルを仮定し，それを利用する手法は，パラメトリック・ブートストラップ（Parametiric Bootstrap）法と呼ばれています。一般的にブートストラップ法といった場合，伝統的なノンパラメトリック・ブートストラップ法を指していることが多いようです。

次にブートストラップ法とは具体的にどのような手法なのかを見てみましょう。例えば，ブートストラップ平均の推定では，得られた k 個の標本から繰り返しを許して k 個を抽出し，平均を求めます。繰り返しを許す，すなわち復元抽出なので，k 個からなる一組のデータがすべて同一標本が選ばれる可能性もありますが，すべて異なる場合もありえます。こうして選ばれた k 個の標本の平均を n 回計算し，その分布を調べます。

また，ブートストラップ分散の推定も同様です。アルゴリズムを文章で書くと以下のようになります。

1. データ $x_1, x_2, \ldots, x_k$ から無作為復元抽出を k 回行い，大きさ k のブートストラップ標本 $X_1^{*b}, X_2^{*b}, \ldots, X_k^{*b}$ を構成し，ブートストラップ統計量 $\hat{\theta}^{*b} = \hat{\theta}(X_1^{*b}, X_2^{*b}, \ldots, X_k^{*b})$ を計算する。
2. ステップ（1）を n 回繰り返すことにより $\hat{\theta}^{*1}, \ldots, \hat{\theta}^{*n}$ を計算する。
3. ブートストラップ分散推定量を次式により近似計算する。

$$\hat{\sigma}_b^2 = \sum_{b=1}^{n} \frac{(\hat{\theta}^{*b} - \hat{\theta}^*_{\cdot})^2}{n-1} \tag{1}$$

ただし，$\hat{\theta}^*_{\cdot} = \frac{1}{n}\sum_{b=1}^{n} \hat{\theta}^{*b}$ はブートストラップ統計量のモンテカルロ平均です。ブートストラップ反復回数 n は，標本数 k の大きさと推定量の複雑さに応じて適切に定める必要があり，一般に k が大きければ n も大きくとる必要があります。

最後にブートストラップ信頼区間について簡単に述べておきます。ブートストラップ信頼区間とは，推定値の信頼区間を構成するためにブートストラップ法を利用して計算された信頼区間のことです。このブートストラップ信頼区間の中で，特に有名なものが 3 つあります。1 つ目はパーセンタイル法と呼ばれるのもであり，これが計算的に最も簡単な方法です。2 番目の方法はブートストラップ t 法と呼ばれるものですが，この方法では推定量の分散の推定が必要となります。分散推定値があまり信頼できない場合には，この方法の適用には注意が必要です。第 3 の方法は BCa 法と呼ばれるもので，パーセンタイル法やブートストラップ t 法を改良した方法です。BCa 法は推定量の偏りとその分布のゆがみを同時に補正した形で信頼区間を計算することができます。ただし，この方法を適用するためには，偏り修正量と，加速定数と呼ばれる量を推定しなければならなりません。ブートストラップ信頼区間を詳しく知りたい方は，専門書等を参考にして下さい。

2　ブートストラップ法の実施方法

ブートストラップ法を実際に使いたい場合，フリーの統計解析ソフト『R』を利用すると簡単です。（ただし，アルゴリズムが簡単なので，エクセルのマクロを

自分で書いても可能ではあります。）R を用いた場合，’simpleboot’ というライブラリの中にある’one.boot’ という関数を利用するのが最も簡単です。ブートストラップ法の標準的な分析手順は以下の通りです。

1. データを準備する。
2. 通常の方法で分析し，結果を見る。
3. 関心のある値あるいはベクトルを出力する関数を作成する。
4. ブートスラップ法を適用する。
5. ブートストラップ法適用後の信頼区間を計算する。

2.1 データの準備

ここでの分析には，以下のような花粉飛散量の測定データを用います。このデータは宮城県多賀城市鶴ヶ谷 2 丁目で 2007 年 2 月に測定されたものであり，花粉の種類はすべてスギ花粉でした。また，飛散量は $1cm^2$ あたりの発見された花粉の個数で記録されています。

2007/2/06	スギ花粉	1.9 個/cm^2
2007/2/07	スギ花粉	0.6 個/cm^2
2007/2/10	スギ花粉	6.2 個/cm^2
2007/2/13	スギ花粉	1.5 個/cm^2
2007/2/14	スギ花粉	1.5 個/cm^2
2007/2/15	スギ花粉	6.5 個/cm^2
2007/2/17	スギ花粉	29.0 個/cm^2
2007/2/18	スギ花粉	32.1 個/cm^2
2007/2/19	スギ花粉	20.7 個/cm^2
2007/2/20	スギ花粉	76.5 個/cm^2
2007/2/21	スギ花粉	53.7 個/cm^2
2007/2/22	スギ花粉	179.9 個/cm^2
2007/2/23	スギ花粉	330.6 個/cm^2
2007/2/24	スギ花粉	38.9 個/cm^2
2007/2/25	スギ花粉	42.3 個/cm^2
2007/2/26	スギ花粉	13.6 個/cm^2
2007/2/27	スギ花粉	45.4 個/cm^2

このデータを折れ線グラフとして描いてみると以下のようになります。

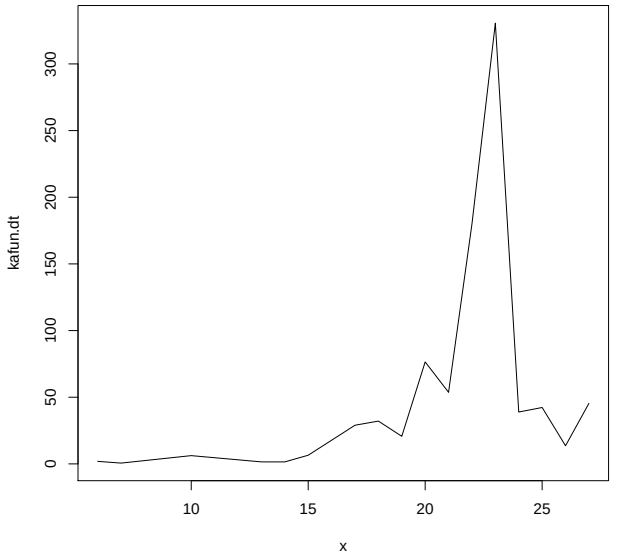

図 1　花粉飛散量の時系列推移

このように，2007 年 2 月のスギ花粉は 23 日をピークとして飛散していたことがわかります。

2.2 平均値に関するブートストラップ推定

まず始めに，花粉飛散量の単純な平均値を求めてみましょう。すると結果は『51.81765』のようになりました。値を確認するまでもなく，2 月 22 日や 23 日の外れ値に引っ張られてしまい，平均値が大きくなってしまっています。これでは 2007 年 2 月の花粉飛散量の代表値として適切だとは言えませんし，ここから算出された分散の値も正しいとは言えません。したがって，このままでは平均値が存在する範囲（95% 信頼区間など）を見積もることができません。

そこで，次にブートストラップ法を適用してみます。これより以下はブートストラップ法に関する R の関数の使い方なので，R の使い方がよく分からない人は，まずはそちらを勉強して下さい。

```
library(simpleboot)
library(boot)
kafun.dt <- c(1.9,0.6,6.2,1.5,1.5,6.5,29.0,
              32.1,20.7,76.5,53.7,179.9,
              330.6,38.9,42.3,13.6,45.4)
b.mean <- one.boot(kafun.dt, mean, 10000)
boot.ci(b.mean)
```

なお，'one.boot' に関して，第 1 引数にはオリジナルデータ（花粉飛散量）を，第 2 引数にはブートストラップを行いたい関数を，第 3 引数には反復回数を指定します。今回の反復回数は 10000 回としました。

結果は以下の通りです。なお，ブートストラップ法の性質上，分析を繰り返しても以下の結果と全く同じになることはありません。このようにして推定された範囲は，ただ点推定値として平均値と標準誤差を求めて信頼区間を計算するよりも『統計的に正しい』結果となっています。

```
Intervals :
Level      Normal              Basic
95%   (13.12, 90.63 )   ( 7.79, 83.32 )

Level     Percentile            BCa
95%   ( 20.32,  95.84 )   ( 25.11, 117.31 )
```

3 参考文献

Efron, B. and Tibshirani, R. (1993) *An Introduction to the Bootstrap*. Chapman and Hall, New York, London.

汪金芳・手塚集・上田修功・田栗正章（2003）計算統計 I–確率計算の新しい手法– 岩波書店。

永田靖・棟近雅彦 (2001) 多変量解析法入門 サイエンス社。

吉岡茂（2007）C プログラミングの基礎と統計プログラミング 現代図書。