R とブートストラップ

1. ブートストラップとは

統計学の主な目的の 1 つは、標本データを用いて母集団の性質を推測することである。同じ母集団から抽出した標本であっても、無作為であるため標本を構成する要素、標本のサイズが異なると、それらの統計量(比率、平均、分散など)は異なる。従って、標本データを用いて母集団の性質を推測する際には常に誤差が伴う。

正規分布 $N(\mu,\sigma^2)$の母集団から抽出した大きさ n の無作為標本の平均は $N(\mu,\sigma^2/n)$ に従うことが知られている。σ は一定の条件のもとでは標本の不偏標準偏差を用いることも可能である。このように正規分布、t分布、x^2分布などの確率分布を用いて母数やモデルの推定およびその推定の誤差を計算することができる。しかし、問題によっては確率分布を仮定できないケースも少なくない。そこで、1970 年代にエフロン(Efron)は確率分布の性質に頼らないブートストラップ(bootstrap)という方法を提唱した。ブートストラップの語源に関しては、インターネットでも検索できる。

データサイエンスの分野では、1 つの標本から復元抽出を繰り返して大量の標本を生成し、それらの標本から推定値 を計算し、母集団の性質やモデルの推測の誤差などを分析する方法をブートストラップ法と呼ぶ。

ブートストラップ法では母数 の推定量は標本から生成したブートストラップ標本の推定量 を用いて推定する。1 つの標本からリサンプリングを繰り返して生成する標本をブートストラップ標本と呼ぶ。図1にブートストラップ法のイメージを示す。

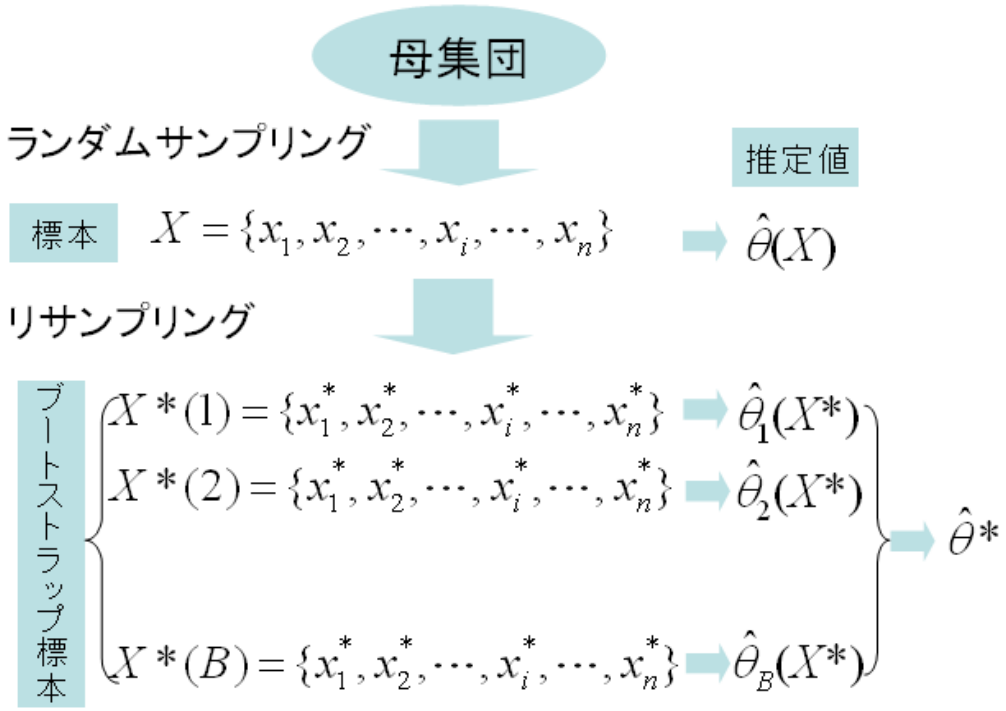

図 1　ブートストラップ法のイメージ

2. ブートストラップ標本の生成

ブートストラップ標本の生成には幾つかの方法が提案されているが、確率分布型を仮定するパラメトリック・ブートストラップ法と確率分布型を仮定しないノンパラメトリック・ブートストラップ法に大別される。そのアルゴリズムの例を次に示す。

パラメトリック・ブートストラップ法

(1)標本サイズが n である標本データ$\{x_1,x_2,\cdots,x_i,\cdots,x_n\}$の平均$\bar{x}$、標準偏差 sを計算する。

(2)n 個の正規乱数 $z_1,z_2,\cdots,z_i,\cdots,z_n$ を生成し、$x_i^* = \bar{x} + z_i s$ で新しい標本

$\{x_1^*,x_2^*,\cdots,x_i^*,\cdots,x_n^*\}$を生成する。この標本による推定値を$\hat{\theta}_1^*$ （例えば、平均$\bar{x}_1$）とする。

ノンパラメトリック・ブートストラップ法

(1)区間(0,1)を n 等分した各区間の値を標本データ$\{x_1,x_2,\cdots,x_i,\cdots,x_n\}$に1対1で対応させる。
(2)n 個の一様乱数 $u_1,u_2,\cdots,u_i,\cdots,u_n$ を生成し、u_i の値が含まれる区間に対応する x_k を x_i^*とし、新しい標本データ$\{x_1^*,x_2^*,\cdots,x_i^*,\cdots,x_n^*\}$を生成する。この標本から得られた推定値を$\hat{\theta}_1^*$ とする。

両方法ともステップ(2)を B 回繰り返し、B 個の標本の推定値$\hat{\theta}_1^*,\hat{\theta}_2^*,\cdots,\hat{\theta}_3^*,\cdots,\hat{\theta}_B^*$ を求める。その推定値、標準偏差、バイアスはそれぞれ次の式で求める。

$$\bar{\hat{\theta}} = \frac{1}{B}\sum_{i=1}^{B}\hat{\theta}_i^*$$

$$s(\bar{X}) = \sqrt{\frac{1}{B-1}\sum_{i=1}^{B}(\hat{\theta}_i^* - \bar{\hat{\theta}})^2}$$

$$bias(\bar{X}) = \frac{1}{B}\sum_{i=1}^{B}(\hat{\theta}_i^* - \hat{\theta})$$

また、確率分布関数は

$$Pr(\bar{X} \le x) = \frac{1}{B}\{x \ge \bar{x}_i \text{の個数}\}$$

により推定できる。B 個の推測値を大小順に並べた $B\times\alpha$ 番目の値を 100α%点とする。繰り返しの回数 B については、推定値の標準誤差を求める場合は約 100～200 回、確率分布関数の値や 100α%点を求める場合は 1000～2000 回が必要であるとされている(小西(1988, 2004))

3. 区間推定

ブートストラップ標本を生成し、平均の信頼区間を推定する例を次に示す。用いるデータは平均 170、標準偏差 6 である正規分布の母集団から抽出したサイズが 20 の標本であるとする。標本データ生成のコマンドを示す。

```
> set.seed(20)
> sam<-rnorm(20,170,6)
> round(sam[1:5],2)
[1] 165.75 166.71 170.16 169.54 165.03
```

パラメトリック・ブートストラップ法による 2000 個のブートストラップ標本を生成し、その平均 を求めるコマンドを次に示す。ただし、標準偏差は標本の標準偏差を用いる。

```
> tt<-numeric(0)
> ME<-mean(sam); SD<-sd(sam)
> for(i in 1:2000){z<-rnorm(20,0,1);bx<-ME+z*SD; tt<-cbind(tt,mean(bx))}
```

求めた 2000 個の標本平均の平均と 95%の信頼区間を 100 %点で求める例を示す。100 %点は、関数 quantile を用いて求めることができる。

```
> mean(tt)
[1] 170.2390
> quantile(tt,p=c(0.025,0.975))
    2.5%    97.5%
168.1689  172.3980
```

ノンパラメトリックのブートストラップ標本は、関数 sample を用いて生成することができる。関数 sample を用いて生成したブートストラップ標本を用いて区間推定を行うコマンドを次に示す。

```
> tt<-numeric(0)
> for(i in 1:2000){bs<- sample(sam,20,replace =
TRUE)me<-mean(bs)tt<-cbind(tt,me) }
> mean(tt)
[1] 170.1977
> quantile(tt,p=c(0.025,0.975))
    2.5%    97.5%
168.1203  172.3522
```

R にはブートストラップに関連するパッケージ boot, simpleboot, bootstrap などがある。いずれも CRAN ミラーサイトからダウンロードできる。

パッケージ boot の中にはブートストラップの推定量を求める関数 boot があるが引数

が若干多いので、ここではパッケージ simpleboot の中の関数 one.boot を用いることにする。関数 one.boot の書き式を次に示す。ただし、引数に関しては最も基本的な項目のみを示す。

one.boot(data, FUN, R, ...)

引数 data には用いる標本データ、引数 FUN には推定する統計量の関数(mean, median,あるいは自作関数)を指定する。引数 R には生成するブートストラップ標本の数を指定する。デフォルトの値のままブートストラップの推定値のベクトル $\hat{\theta}_1^*, \hat{\theta}_2^*, \cdots, \hat{\theta}_3^*, \cdots, \hat{\theta}_B^*$ を生成するためには、これらの3つの引数のみを指定すればよい。

標本データ sam を用いた10個のブートストラップ標本の平均値を求める例を次に示す。ブートストラップの平均ベクトルは$t に格納されている。生成されたブートストラップの統計量はノンパラメトリック法によるものである。

```
>library(simpleboot); set.seed(20)
> b.mean <- one.boot(sam, mean, 10)
> b.mean$t
         [,1]
[1,] 170.4848
＜中略＞
[10,] 169.1167
```

パッケージ simpleboot の関数が返す結果のオブジェクト形式は、パッケージ boot を用いた結果と一致する。パッケージ boot には信頼区間を求める関数 boot.ci がある。関数 boot.ci の書き式を次に示す。ただし、引数に関しては最も基本的な項目のみを示す。

boot.ci(boot.out, conf = 0.95, type = "all", ...)

引数 boot.out はパッケージ boot、あるいは simpleboot で生成したブートストラップの結果オブジェクトである。引数 conf は信頼区間であり、デフォルトには 95%の信頼区間が指定されている。引数 type は信頼区間を求める方法であり、正規分布の近似(normal) 法、basic 法、ブートストラップ *t*(studentized)法、パーセンタイル(percentile)法、BCa 法の5種類の方法を指定することができる。ただし、ブートストラップ *t* 法に関しては、ブートストラップ標本の推定値を生成する際にブートストラップ *t* 法を用いなければならない。関数 boot、one.boot には関連の引数がある。これらのアルゴリズムに

関しては汪(2006)が詳しい。

関数の boot.ci の引数 type をデフォルト値"all"のままで実行するとブートストラップ t 法を除いた4種類の信頼区間を返す。

関数boot.ciに実装された5種類の区間推定法の中の正規分布の近似法を除いた4種類の境界値の計算法を表1に示す。

表 1　4 種類の区間推定の計算式

方法	有意水準 α の点
Basic	$2\hat{\theta} - \hat{F}_b^{-1}(\alpha)$
Studentized	$\hat{\theta} - s(\hat{\theta})\hat{F}_s^{-1}(\alpha)$
Percentile	$\hat{F}_b^{-1}(\alpha)$
BCa	$\hat{F}_b^{-1}\{\Phi[c + \frac{c - z_\alpha}{1 - a(c + z_\alpha)}]\}$

表の中の $\hat{\theta}$ は標本データの推定値、$\hat{F}_b$ はブートストラップの累積分布関数、z_α はブートストラップ t の累積分布関数、$\hat{F}_s$ は有意水準 α の標準正規分布の値、a,c は定数である。定数 a、c の決め方に関しては汪・田栗(2003)，丹後(2000)に紹介されている。ブートストラップの標本の数を2000とした標本平均の信頼区間を、関数boot.ciを用いて求める例を次に示す。

```
> library(boot)
> b.mean <- one.boot(sam, mean, 2000)
```

```
> boot.ci(b.mean)
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 2000 bootstrap replicates

CALL :
boot.ci(boot.out = b.mean)

Intervals :
Level      Normal              Basic
95%   (168.0, 172.4 )   (167.9, 172.3 )

Level     Percentile            BCa
95%   (168.1, 172.5 )   (168.2, 172.7 )
Calculations and Intervals on Original Scale
```

返された結果 Intervals の下部の Normal は通常よく用いられている正規分布を近似した区間推定の結果である。Percentile の結果は推定値を小さいものから大きい順に並べた数列の 100α%点である。

異なる計算方法であるのにもかかわらず4種類の推定値が非常によく近似している。乱数を用いる計算結果は、同じ方法を繰り返してもその結果が同じになるとは限らない。勿論、同一の乱数シードを用いるなどの工夫を行うことで再現することは可能である。

より安定した区間推定値を得るためにはブートストラップ標本の数を増やすことが1つの方法である。上記の例のような計算量であればブートストラップ標本を1万にしてもRでは直ちに計算結果が返される。

4. 回帰分析とブートストラップ

ブートストラップ法は多くの統計データ処理に用いられている。本項では目的変数を $Y=\{y1,y2,\cdots,yn\}$、説明変数を $X=\{x1,x2,\cdots,xm\}$としたブートストラップ法による回帰分析の例を紹介する。そのアルゴリズムを次に示す。

(1)観測の標本データを用いて回帰モデル $\hat{Y}=R(X)$ を作成する。

(2)残差 $E=Y-\hat{Y}$ を用いてノンパラメトリック・ブートストラップ標本 $E^*=\{\varepsilon 1^*,\varepsilon 2^*,\cdots,\varepsilon n^*\}$ を生成する。

(3)回帰モデルの予測値に残差のブートストラップ標本を加え、新しい目的変数のブートストラップ標本 $Y^*=\hat{Y}+E^*$ を作成する。

(4) X^*,Y^*を用いて回帰モデルを作成する。
(5)ステップ(2)～(4)を B 回繰り返す。

回帰係数は B 回のブートストラップ標本の回帰係数の平均、回帰係数の推定誤差はブートストラップ標本の回帰係数の標準偏差を用いて求める。

R の中のデータ cars を用いた例を次に示す。

```
> data(cars)
> car.lm<-lm(dist~speed,data=cars)
> car.boot<-lm.boot(car.lm, R = 2000)
> summary(car.boot)
BOOTSTRAP OF LINEAR MODEL  (method = rows)

Original Model Fit
------------------
Call:
lm(formula = dist ~ speed, data = cars)

Coefficients:
(Intercept)        speed
    -17.579        3.932

Bootstrap SD's:
(Intercept)        speed
  5.9178481    0.4170457
```

パッケージ simpleboot には関数 lm.boot のブートストラップ単回帰結果の作図関数 plot.lm.simpleboot があり、plot に略して用いる。この関数は散布図の回帰直線にブートストラップの 2 倍の標準誤差の区間を表示する。その例を次に示す。

```
> plot(car.boot,xlab="speed", ylab="dist")
```

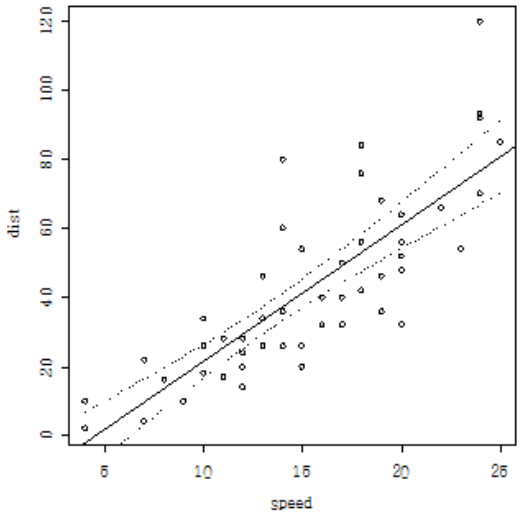

図 2　lm.boot の回帰図

パッケージ simpleboot には、局所多項式回帰関数 loess による回帰結果についてブートストラップ回帰を行う関数 loess.boot がある。データ cars を用いた 2000 回のブートストラップの回帰結果のグラフを次に示す。

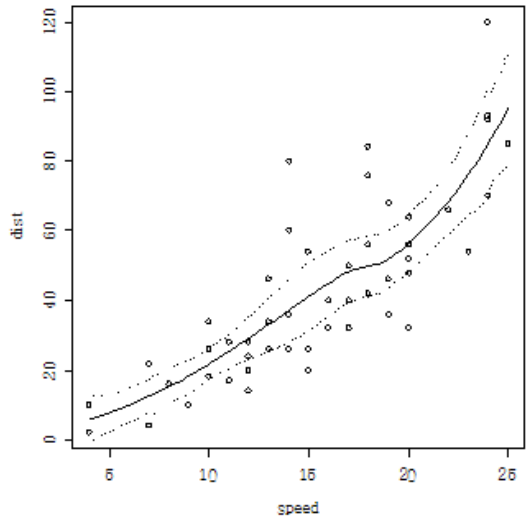

図 3　関数 loess.boot の回帰図

パッケージ boot, bootstrap にもブートストラップ回帰関数が用意されている

5. クラスター分析とブートストラップ

階層的クラスター分析で得られた樹形図がデータの構造を表す真の樹形図であるどうかは判断しがたい。従って、樹形図を用いたクラスターの結果を解釈する際には主観的になりがちである。

そこで、得られた樹形図が真の樹形図であるかどうかをブートストラップ法によって評価する研究が行われている。

クラスター分析におけるブートストラップ法の適応は、ブートストラップ標本が仮説を支持する相対頻度(ブートストラップ確率)を仮説検定の確率値として用いる。下平(2002)は、標本データの変数の次元と異なるブートストラップ標本を生成するマルチスケールブートストラップ法によるブートストラップ確率を用いて樹形図を評価する方法を提案した。そのアルゴリズムを含む解説論文の PDF ファイルがインターネット上で入手できるので、その説明は省略する。その理論に基づいたパッケージ pvclust は CRAN ミラーサイトからダウンロードできる。クラスターの生成とブートストラップ確率を計算する関数は pvclust である。その書き式を次に示す。

```
pvclust(data,
method.hclust="average",method.dist="correlation",use.cor="pairwise.complete.obs
",nboot=1000,…)
```

引数 method.hclust は、関数 hclust の引数 method に対応する。引数 method.dist では距離関数 dist の引数および相関係数のような類似度関数を指定する。引数 use.cor は、相関係数行列を求める関数 cor の引数 use に相当する。引数 nboot はブートストラップ標本の数であり、1000 以上が推奨されているが、データサイズが大きい場合は計算結果が返されるのを待つのに辛抱を要する。

データ iris の中の異なる2種類のデータの一部分を用いた例を次に示す。ここではキャンベラ距離を用いる。

```
> y<-t(iris[c(51:60,141:150),1:4])
> library(pvclust)
> irisy.pvwar <- pvclust(y, method.dist="can",nboot=5000,
method.hclust="ward",r=seq(.5,1,by=.1))
> plot(irisy.pvwar,hang=-1,cex.pv=1,  col.pv=c(4,2,8),float=0.02 )
```

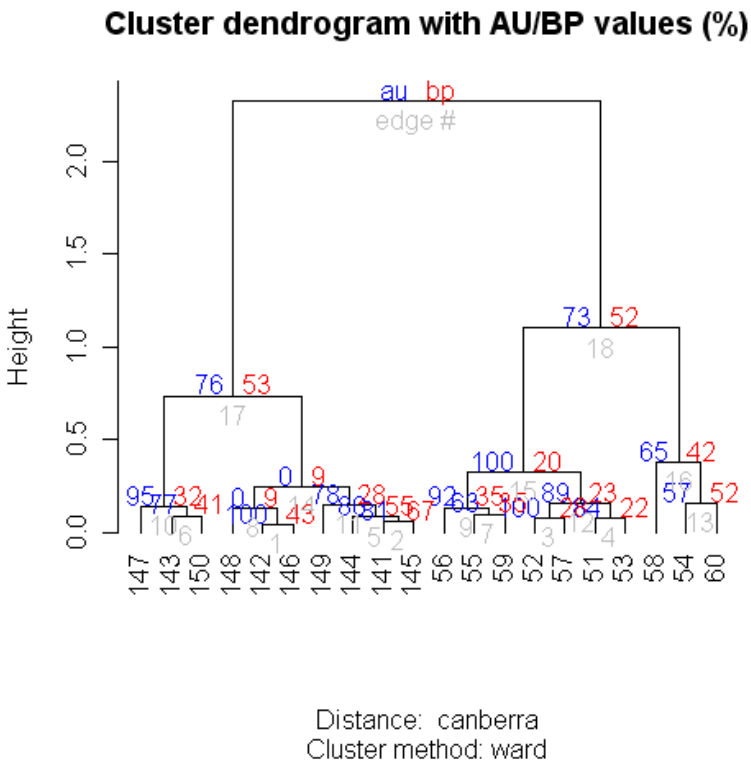

図 4　ブートストラップ確率と樹形図1

この結果の場合はブートストラップ確率(ノードの右側の数値 bp)が 5%以下のものはない。2つのクラスターに分ける際のそれぞれのクラスターの最上部のノードのブートストラップ確率は 50%を超えている。都合よくこの2つのクラスターは真のクラスターと一致する。パッケージの中には確率値を用いてクラスターを見つける関数 pvpick がある。

```
> pvpick(irisy.pvwar, alpha=0.50, pv="bp")
```
＜結果は省略＞

参考のため、同じキャンベラ距離に最遠隣法を用いたブートストラップ標本数を 5000 とした結果を図 5 に示す。このように、同じデータと距離に対して、異なるクラスター方法を用いると、そのブートストラップ確率も大きく異なる。

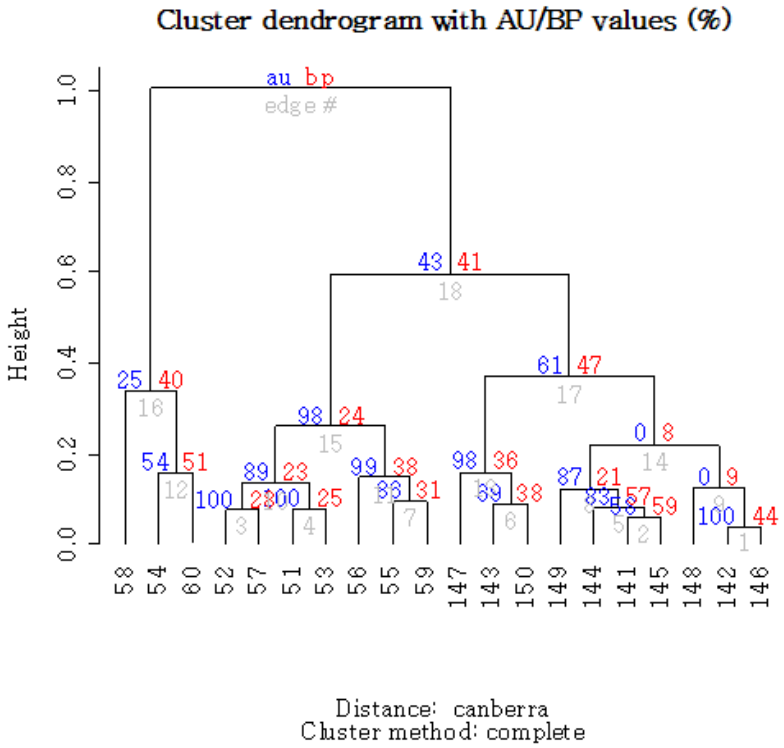

図 5　ブートストラップ確率と樹形図2

参考文献

小西　貞則(1988). ブートストラップ法による推定量の誤差評価、パソコンによるデータ解析(村上・田村編)、p.123-142, 朝倉書店

小西　貞則, 北川　源四郎(2004). 情報量基準、朝倉書店

丹後　俊郎(2000). 統計モデル入門, 朝倉書店

下平　英寿(2002). ブートストラップ法によるクラスター分析のバラツキ評価、統計数理、第 50 巻第 1 号、p.33-44(http:// www.ism.ac.jp/editsec/toukei/pdf/50-1-033.pdf)

汪　金芳・田栗　正章(2003). ブートストラップ法入門,「計算統計Ⅰ:統計科学のフロンティア 11」, 岩波書店

「ブートストラップ法を教えれ」

無茶を言うな。名前を聞いたことしかないわ。

と思いつつ、知らないでいるのも何なので少しだけ調べてみた。

とりあえず一日調べた感じだと、「標本集団からのリサンプリングを繰り返すことにより母集団の性質を推定する」方法らしい。 これだけだと何のことやらわからないかもしれない。

というわけで、僕の勉強を兼ねて少し丁寧に見ていこう。

1.母集団の性質を推定する

たとえば、何らかのデータについて「平均値」という性質を知りたいとしよう。もしも存在する全てのデータ（これを母集団と呼ぶ） を得ることが可能であれば、単純に全てのデータをデータの個数で割れば「平均値」は得られる。

しかし、そのようなケースはほとんどない。例えば、地球上の全てのカラスの体長を知ることができるだろうか？また、 実験で得られたデータを考えても、過去に行われた実験、さらに未来に行われるであろう実験についてまでデータを得られるだろうか？ 母集団は未知なのが普通なのだ。

そこで通常は可能な範囲でデータを(ふつう無作為に)集めることになる。こうして集められたデータ、つまり母集団の一部であるが、 これを標本集団、または単に標本と呼ぶ。

じゃあ、標本集団のデータの平均を取れば母集団の「平均値」が分かる…という訳にはいかない。 ランダムに採取されたデータは当然ばらつくのだから、 **標本集団から母集団の真の性質を知ることは不可能である**。

そこで、通常は標本集団を用いて母集団の性質を「推定」する。

例えば「平均値」については、母集団が平均値 μ、分散 σ^2 のなんらかの分布をしている場合、そこから抽出された大きさ n の標本集団は「平均値」が平均値 μ、 分散 σ^2/n の正規分布に従うということが知られている。分散が小さければ良い推定値となるので、 大きさ n を増やせば、つまりたくさんのサンプルを採れば母集団の平均値の良い推定が得られる。まあ当たり前といえば当たり前。

（2007.11.08 追記：ごめん、ウソ。俺が馬鹿だった。正しくは、平均値は正規分布し、分散比はＦ分布し、 期待値からの外れ具合はカイ二乗分布する…といったように、求めたい統計量に対し、何か確率分布を過程できるならそいつの推定は簡単ですよ、 というような感じです。そのほかいろいろ修正しました）

ちなみに、今の例だと母集団の分散が分からなければ標本集団の平均値の分散も分からないではないか、 というツッコミが入るかもしれない。しかし、 母集団の分散(母分散という)の代わりに標本より計算される分散(不偏分散という)を用いた場合、 t 分布と呼ばれる正規分布に良く似た分布(実際、n を増やせば正規分布に近づく)をするということも知られているので、 やはり母集団の平均(母平均)を推定することは可能なのである。詳しくは専門の解説書を参考に。

削除: 要するに、母集団の分布の仕方(正規分布に限らない)が明らかであれば、 標本集団の平均値の分布の仕方も分かるということである。 そして、そのほかの性質(分散、標準偏差、 中央値 etc...)についても標本集団より推定ができる。

2.ブートストラップ法

求めたい統計量の分布の仕方が明らかであれば標本集団より母集団の性質を知ることができる。

しかし、分布を仮定できない統計量というのも少なからず存在する。2 種のデータの中央値の差だとか、平均値の比だとか （僕が不勉強なだけですでにこれらの確率分布も知られているのかもしれないけど)、そんなデータがどんな分布をするのかを考えるのは難しい。

1979 年、スタンフォード大学のブラッドリー・エフロンは、確率分布にたよらず母集団の性質の推定量を得る「ブートストラップ法」 を発表した。ブートストラップという語には、ブーツの紐を引っ張り、自分で自分を持ち上げる、つまり「自力で」というような意味合いがある。 コンピュータを起動するときの「ブート」も「ブートストラップ」より来ている。

ブートストラップ法では、標本集団より重複を許したリサンプリングを多数回繰り返し、 そうして得られた新たな標本集団(これをブートストラップ集団と呼ぶ)より母集団の性質を推定する。

3.R を使ってやってみる

実際にやってみるのが一番わかりやすいと思うので、R を使って実際に簡単なブートストラップ法をやってみよう。ちなみに、 ブートストラップ法には確率分布を仮定するパラメトリック・ブ

ートストラップ法と、確率分布を仮定しないノンパラメトリック・ ブートストラップ法があるが、ここではノンパラメトリック・ブートストラップ法を行っている。

まず、標本集団を用意する。変な分布をしているもののほうが面白いと思うので、次のようなデータを使ってみる。

```
x <- c(1,1,1,1,1,1,1,1,10,10,10,1,1,1,10,10,10,1,1,1)
```

データ数は 20。このデータの母集団の平均値を推定しよう(なお、平均値の誤差は正規分布することが分かっているので、 本来ブートストラップ法を用いる必要は無い)。

最初に、ブートストラップ標本平均のデータを入れる入れ物を作っておく。

```
my.boot <- numeric(0) #my.boot という名前の入れ物
```

そしてリサンプリング。2000 回ほどやってみよう。

```
for(i in 1:2000){              #i が 2000 になるまで i を 1 ずつ増やしながら繰り返し
 y <- sample(x,20,replace=T)#重複を許して x より 20 個リサンプリング
 my.boot[i] <- mean(y)      #リサンプリングの平均を my.boot の i 番目へ代入
 }
```

これで、2000 個のブートストラップ標本平均データが bootstrap という入れ物に入った。あとは区間推定をするだけ。 ブートストラップ法による信頼区間にはいくつか種類があるが、ここでは一番簡単なパーセンタイル法による信頼区間を求める。

パーセンタイル法は、ブートストラップ標本の推定量(今回は平均)の集合を大きさ順に並べたとき、 100×α%の位置を信頼区間とするというシンプルな方法。R では、 クォンタイル点を求めるための関数 quantile に引数を与えることで、何%の順位のデータでも得ることができる。さきほど得られたブートストラップ標本平均のデータより、2.5%点と 97.5%点のデータを求める(要するに、 95%信頼区間を求める)には、次のようにする。

```
quantile(my.boot,p=c(0.025,0.975)) #2.5%点、97.5%点の計算→95%信頼区間の推定
```

そうすると、次のように結果が得られる。

```
  2.5% 97.5%
1.9    5.5
```

これがブートストラップ法により推定された平均値の信頼区間である。

以上で説明はおしまい。R にはブートストラップ法のためのパッケージがいくつかあるので、興味のある方、もっと難しいことをしたい方、 もっと楽をしたい方は是非参照のこと。

参考文献

- 金明哲、 R とブートストラップ、ESTRELA 2007 年 3 月
- 汪金芳、 ブートストラップ法入門
- デイヴィッド・サルツブルグ 「統計学を拓いた異才たち―経験則から科学へ進展した一世紀」

「ブートストラップ法を教えれ」

無茶を言うな。名前を聞いたことしかないわ。

と思いつつ、知らないでいるのも何なので少しだけ調べてみた。

とりあえず一日調べた感じだと、「標本集団からのリサンプリングを繰り返すことにより母集団の性質を推定する」方法らしい。 これだけだと何のことやらわからないかもしれない。

というわけで、僕の勉強を兼ねて少し丁寧に見ていこう。

1.母集団の性質を推定する

たとえば、何らかのデータについて「平均値」という性質を知りたいとしよう。もしも存在する全てのデータ（これを母集団と呼ぶ） を得ることが可能であれば、単純に全てのデ

ータをデータの個数で割れば「平均値」は得られる。

しかし、そのようなケースはほとんどない。例えば、地球上の全てのカラスの体長を知ることができるだろうか？また、 実験で得られたデータを考えても、過去に行われた実験、さらに未来に行われるであろう実験についてまでデータを得られるだろうか？ 母集団は未知なのが普通なのだ。

そこで通常は可能な範囲でデータを(ふつう無作為に)集めることになる。こうして集められたデータ、つまり母集団の一部であるが、 これを標本集団、または単に標本と呼ぶ。

じゃあ、標本集団のデータの平均を取れば母集団の「平均値」が分かる…という訳にはいかない。 ランダムに採取されたデータは当然ばらつくのだから、 標本集団から母集団の真の性質を知ることは不可能である。

そこで、通常は標本集団を用いて母集団の性質を「推定」する。

例えば「平均値」については、母集団が平均値μ、分散$\sigma 2$のなんらかの分布をしている場合、 そこから抽出された大きさ n の標本集団は「平均値」が平均値μ、 分散$\sigma 2/n$の正規分布に従うということが知られている。分散が小さければ良い推定値となるので、 大きさ n を増やせば、つまりたくさんのサンプルを採れば母集団の平均値の良い推定が得られる。まあ当たり前といえば当たり前。

要するに、母集団の分布の仕方(正規分布に限らない)が明らかであれば、 標本集団の平均値の分布の仕方も分かるということである。 そして、そのほかの性質(分散、標準偏差、 中央値 etc...)についても標本集団より推定ができる。

（2007.11.08 追記：ごめん、ウソ。俺が馬鹿だった。正しくは、平均値は正規分布し、分散比は F 分布し、 期待値からの外れ具合はカイ二乗分布する…といったように、求めたい統計量に対し、何か確率分布を過程できるならそいつの推定は簡単ですよ、 というような感じです。そのほかいろいろ修正しました）

ちなみに、今の例だと母集団の分散が分からなければ標本集団の平均値の分散も分からないではないか、 というツッコミが入るかもしれない。しかし、 母集団の分散(母分散という)の代わりに標本より計算される分散(不偏分散という)を用いた場合、 t 分布と呼ばれる正規分布に良く似た分布(実際、n を増やせば正規分布に近づく)をするということも知られているので、 やはり母集団の平均(母平均)を推定することは可能なのである。詳しくは専

門の解説書を参考に。

2.ブートストラップ法

求めたい統計量の分布の仕方が明らかであれば標本集団より母集団の性質を知ることができる。

しかし、分布を仮定できない統計量というのも少なからず存在する。2 種のデータの中央値の差だとか、平均値の比だとか （僕が不勉強なだけですでにこれらの確率分布も知られているのかもしれないけど)、そんなデータがどんな分布をするのかを考えるのは難しい。

1979 年、スタンフォード大学のブラッドリー・エフロンは、確率分布にたよらず母集団の性質の推定量を得る「ブートストラップ法」 を発表した。ブートストラップという語には、ブーツの紐を引っ張り、自分で自分を持ち上げる、つまり「自力で」というような意味合いがある。 コンピュータを起動するときの「ブート」も「ブートストラップ」より来ている。

ブートストラップ法では、標本集団より重複を許したリサンプリングを多数回繰り返し、そうして得られた新たな標本集団(これをブートストラップ集団と呼ぶ)より母集団の性質を推定する。

3.R を使ってやってみる

実際にやってみるのが一番わかりやすいと思うので､R を使って実際に簡単なブートストラップ法をやってみよう。ちなみに、 ブートストラップ法には確率分布を仮定するパラメトリック・ブートストラップ法と、確率分布を仮定しないノンパラメトリック・ ブートストラップ法があるが、ここではノンパラメトリック・ブートストラップ法を行っている。

まず、標本集団を用意する。変な分布をしているもののほうが面白いと思うので、次のようなデータを使ってみる。

```
x <- c(1,1,1,1,1,1,1,1,10,10,10,1,1,1,10,10,10,1,1,1)
```

データ数は 20。このデータの母集団の平均値を推定しよう（なお、平均値の誤差は正規分布することが分かっているので、 本来ブートストラップ法を用いる必要は無い)。

最初に、ブートストラップ標本平均のデータを入れる入れ物を作っておく。
my.boot <- numeric(0) #my.boot という名前の入れ物

そしてリサンプリング。2000 回ほどやってみよう。
for(i in 1:2000){ #i が 2000 になるまで i を 1 ずつ増やしながら繰り返し
y <- sample(x,20,replace=T)#重複を許して x より 20 個リサンプリング
my.boot[i] <- mean(y) #リサンプリングの平均を my.boot の i 番目へ代入
}

これで、2000 個のブートストラップ標本平均データが bootstrap という入れ物に入った。
あとは区間推定をするだけ。 ブートストラップ法による信頼区間にはいくつか種類があるが、ここでは一番簡単なパーセンタイル法による信頼区間を求める。

パーセンタイル法は、ブートストラップ標本の推定量(今回は平均)の集合を大きさ順に並べたとき、 100×α%の位置を信頼区間とするというシンプルな方法。R では、 クォンタイル点を求めるための関数 quantile に引数を与えることで、何%の順位のデータでも得ることができる。 さきほど得られたブートストラップ標本平均のデータより、2.5%点と 97.5%点のデータを求める(要するに、 95%信頼区間を求める)には、次のようにする。
quantile(my.boot,p=c(0.025,0.975)) #2.5%点、97.5%点の計算→95%信頼区間の推定

そうすると、次のように結果が得られる。
2.5% 97.5%
1.9 5.5

これがブートストラップ法により推定された平均値の信頼区間である。

以上で説明はおしまい。R にはブートストラップ法のためのパッケージがいくつかあるので、興味のある方、もっと難しいことをしたい方、 もっと楽をしたい方は是非参照のこと。

参考文献
·金明哲、 R とブートストラップ、ESTRELA 2007 年 3 月
·汪金芳、 ブートストラップ法入門

·デイヴィッド・サルツブルグ　「統計学を拓いた異才たち―経験則から科学へ進展した一世紀」