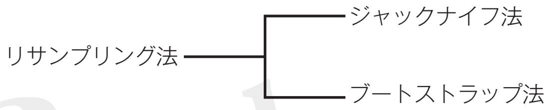
D.1 リサンプリング法による標準誤差の推定

リサンプリング法

母平均、母分散、母相関係数といった母数の値は、1つのデータセットから推定されるので、データセットが異なれば、母数の推定値も異なるものとなる。そこで、推定値のばらつきの大きさを把握する必要がある。このための数値を標準誤差という。標準誤差は理論的に導かれた数学的公式から求めることができるが、実際のデータは理論どおりの状況にあるとは限らない。そこで、手元にあるデータを使って標準誤差を把握する方法が考えられた。この方法として、リサンプリング（Resampling）法と呼ばれる方法が提案されている。リサンプリング法の代表的な手法として、ジャックナイフ（Jackknife）法とブートストラップ（Bootstrap）法がある。



リサンプリング法の代表的な手法

ジャックナイフ法

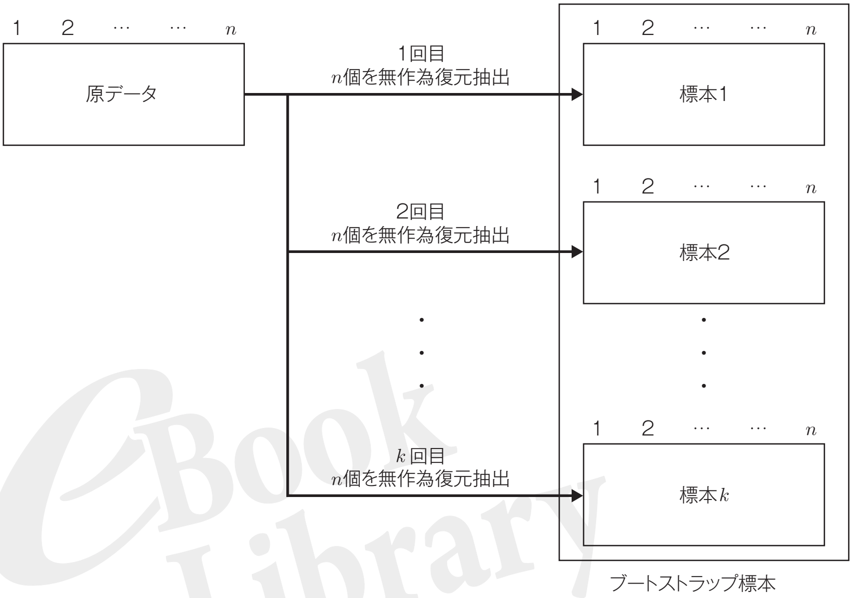
n 個のデータを使って、ある母数の推定値 θ_n を計算したとする。第 i 番目のデータを除いた $(n-1)$ 個のデータを使って計算される推定量を $\theta_{n-1}(i)$ とし、次に示すような θ_i を定義する。

$$\theta_i = n \times \theta_n - (n-1) \times \theta_{n-1}(i) \qquad (i = 1 \sim n)$$

この θ_i を擬似値と呼んでいる。ジャックナイフ法の狙いは、推定値 θ_n の標準誤差を擬似値の標準偏差で把握することにある。なお、擬似値 θ_i の平均値を推定値 θ_n のジャックナイフ推定値と呼んでいる。

ブートストラップ法

n個のデータがあるときに、このデータからn個を無作為に復元抽出する（同じデータが選ばれてもよい）。次に、その標本を使って、注目している母数の推定値θを計算する。それをk回繰り返すと、推定値θがk個得られる。そこで、このk個のθ_i（$i = 1 \sim k$）から、推定値θのばらつきの大きさを調べようとするものである。この方法のイメージを次の図に示す。



ブートストラップ法の概念

繰り返し回数kは1000から10000が使われることが多い。

D.2 ブートストラップ法の実際

例題1：ブートストラップ法による区間推定

次に示すような20個のデータがあるとしよう。このデータを使って、ブートストラップ法による母平均の95%信頼区間を求める方法を解説する。

```
98 78 79 77
72 83 80 67
74 82 81 81
41 64 79 74
68 83 51 98
```

原データ

今、この原データから無作為に20個のデータを復元抽出することを考える。それを標本1として、標本1の平均値を求める。再び、原データから20個のデータを無作為に復元抽出して、それを標本2として、標本2の平均値を求める。このような作業を1000回繰り返すと、1000個の標本が作成され、平均値も1000個求められる。この1000個の平均値の標準偏差を計算して、その値を平均値のばらつき（標準誤差）と見なそうというのがブートストラップ法の考え方である。

では、以下に1000個のブートストラップ標本を生成する実施例を示す。

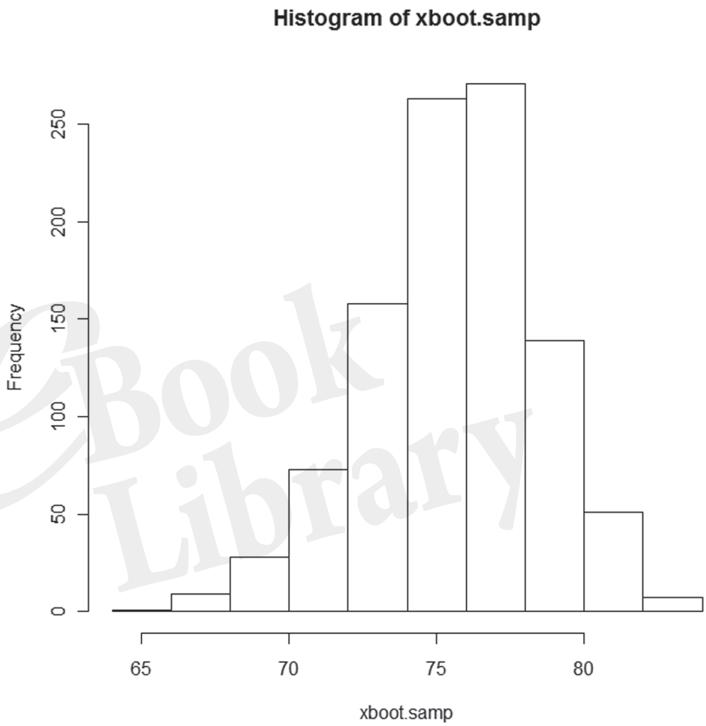
Rによるブートストラップ法

```
> x <- c(98,72,74,41,68,78,83,82,64,83,79,80,81,79,51,77,67,81,74,98)
> xboot.samp <- numeric(1000)
> for(i in 1:1000){
+  xboot <- sample(x, 20, replace=T)
+  xboot.samp[i] <- mean(xboot)
+  }
> quantile(xboot.samp, c(0.025,0.975))◀——パーセンタイルを求める関数
 2.5%   97.5%
69.35   81.05
```

```
> mean(xboot.samp)
[1] 75.6264
> sd(xboot.samp)
[1] 2.905586
```

「平均値の標準偏差」（標準誤差）は2.905586、95％信頼区間は69.35〜81.05と求められている。

ブートストラップ標本の1000個の平均値をヒストグラムで表現すると、次のようになる。



ヒストグラム

パラメトリック・ブートストラップ法

ブートストラップ法の別な方法として、正規乱数を使う方法がある。この原データの平均値と標準偏差は次のような値となる。

```
> x <- c(98,72,74,41,68,78,83,82,64,83,79,80,81,79,51,77,67,81,74,98)
> m <- mean(x)
> s <- sd(x)
> m
[1] 75.5
> s
[1] 13.27641
```

平均値：　75.5
標準偏差：13.27641

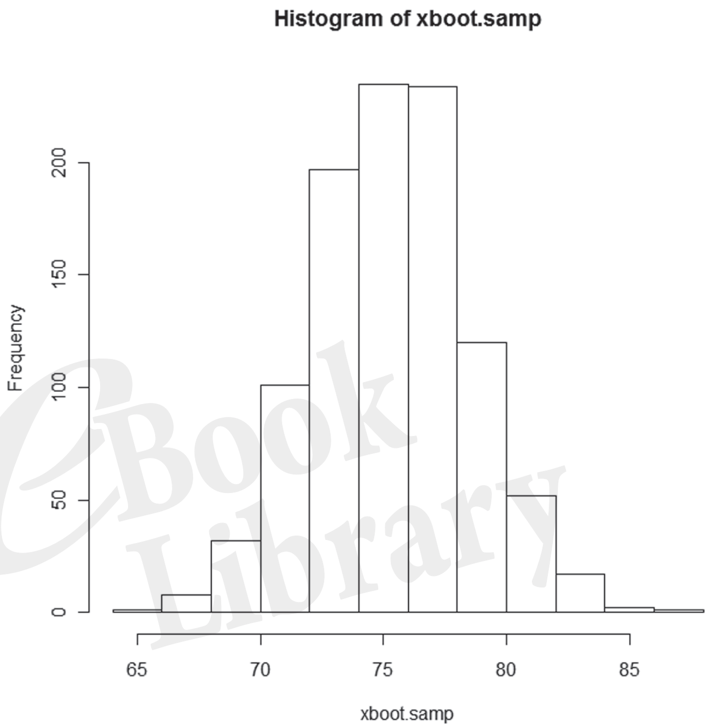
そこで、平均値 ＝ 75.5、標準偏差 ＝ 13.27641の正規乱数を20個発生させて、1つの標本を生成する。この作業を1000回繰り返すことで、1000個のブートストラップ標本が生成され、平均値も1000個求められ、母平均の95%信頼区間を求めることができる。正規乱数を使うこのようなブートストラップ法はパラメトリック・ブートストラップ法と呼ばれている。次にRによる実施例を示す。

```
> x <- c(98,72,74,41,68,78,83,82,64,83,79,80,81,79,51,77,67,81,74,98)
> m <- mean(x)
> s <- sd(x)
> n <- 20
> xboot.samp <- numeric(1000)
> for(i in 1:1000){
+  xboot <- rnorm(n)*s+m
+  xboot.samp[i] <- mean(xboot)
+  }
> quantile(xboot.samp, c(0.025,0.975))
    2.5%      97.5%
69.34167   81.65887
> mean(xboot.samp)
[1] 75.35837
```

```
> sd(xboot.samp)
[1] 3.120452
```

「平均値の標準偏差」(標準誤差)は3.120452、95%信頼区間は69.34167〜81.65887と求められている。

ブートストラップ標本の1000個の平均値をヒストグラムで表現すると、次のようになる。



ヒストグラム

Rには、ブートストラップ法を実施するためのパッケージとして、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。