ブートストラップ法の数値実験

2014ss036 木下彰大
指導教員：小藤俊幸

1　はじめに

ブートストラップ法とは, ある標本集団から母集団の性質を測定するための方法で, 標本集団と標本集団と同じ数だけランダムに値を再抽出し, 新しいデータセットを取得し統計値を計算する方法だ. 例えば,1000 回繰り返したら,1000 個分のデータの平均値と標準偏差, 信頼区間を算出することができる. また, 統計学におけるブートストラップ法は様々な目的に用いられる統計的推論の手法であり再標本化法に分類される一つであり, モンテカルロ法の一つである.

今回は男女における自殺率のデータを用いてブートストラップ法を使い実験をしたいと思う. 日本における自殺率は非常に高い水準で移行し続け, 大きな社会的関心を読んでいる. こういった現象の背後には様々な原因ガ考えられる.

2　データについて

ここで使うデータは平成 9 年から平成 18 年までの 10 年間の女性の自殺者と失業者のデータ（出典　警察庁生活安全地域課（2007）) を使いブートストラップ法で解き勧めてく.

表 1　男性自殺率

	男性自殺者数	男性自殺率	完全失業者数
平成 9 年	16416	26.6	2304167
平成 10 年	23013	37.2	2789167
平成 11 年	23512	37.9	3170000
平成 12 年	22727	36.6	3195833
平成 13 年	22144	35.6	3399167
平成 14 年	23080	37.1	3585333
平成 15 年	24963	40.1	3500000
平成 16 年	23272	37.4	3133333
平成 17 年	23540	37.8	2943333
平成 18 年	22813	36.6	2748333

ここでは,自殺率 = (自殺者数 $*10^5$)/人口 とおく.

3　定義

これからこの平成 9 年から平成 18 年までの 10 年間の男性による自殺率データを取り上げ,R を用いて様々な計算を行う. これらのデータを x_1,...,x_{10} と表すことにする.

3.1　自殺率のデータの平均, 分散

まずは, このデータの平均 $\overline{x}$ と分散 s_n^2 を求める.

>men.h< −c(26.6,37.2,37.9,...,37.4,37.8,36.6)
>men.h
平成 9 年から平成 18 年までの男性の自殺率を men.h に格納
[1]26.6 37.2 37.9 36.6 35.6 37.1 40.1 37.4 37.8 36.6
>mean(men.h)
標本平均
[1]36.29
>var(men.h)
標本不偏分散
[1]12.96322
以上の結果から,$\overline{x} = \sum_{i=1}^{10} x_i/10 = 36.29$
$s_n^2 = \sum_{i=1}^{10}(x_i - \overline{x})^2/(10-1) = 12.96322$ だということが分かる。

4　ブートストラップ標本の抽出

ここでは無作為標本を $x_1, ..., x_{10}$ とする.

4.1　手順 1

先ず,1 から 10 までの整数から重複を許して 10 個の整数を無造作に抽出する. それらを $i_1, i_2, ..., i_{10}$ とします. ここで R を使う.

>set.seed(314) 1 から 10 までの数字を無造作に抽出
>(b< −sample(1:10,replace=TRUE))
[1]1 3 8 3 3 4 3 4 6 8

4.2　手順 2

次にこの $i_1, i_2, ..., i_{10}$ を用いてブートストラップ標本 $X_1^* = x_{i1}, ..., X_{10}^* = x_{i10}, ...$ を構成する.

>(men.b< −men.h[b]) 1 回目のブートストラップ標本
[1] 26.6 37.9 37.4 37.9 37.9 36.6 37.9 36.6 37.1 37.4
この時のブートストラップ標本平均は
>mean(men.b) ブートストラップ標本平均
[1]36.33

4.3　以上から分かること

この場合, ブートストラップ標本平均 $\overline{X}^* = \sum_{i=1}^{10} X_i^*/10 = 14.11$ は初期標本に基づく標本平均 $\overline{x} = 36.29$ に近似する. ブートストラップ標本は毎回異なる可能性が極めて高くこのようにブートストラップ標本に基づく平均値も毎回異なる事が想像できる. 手順 1 により,$X_1^*, ..., X_{10}^*$ は元のデータ $x_1^*, ..., x_{10}^*$ から無作為標本になっているため, ブートストラップ標本平均の期待値は容易に確かめる事ができる. すなわち, 以下で求めたように乱数を絡めて作成した自殺率データ $X_1^*, ..., X_{10}^*$ は, 元の自殺率データ $x_1, ..., x_{10}$ に含まれているかなりの情報を持っているはずだ.

5　2000 回の標本抽出

>(b< −sample(1:10,replace=TRUE))

1 から 10 までの整数から重複を許して 10 回無造作に抽出

[1]4 2 3 3 6 4 8 6 1 4

>(men.b< −men.h[b])

2 回目のブートストラップ標本

[1]36.6 37.2 37.9 37.9 37.1 36.6 37.4 37.1 26.6 36.6

>mean(women.b)

2 回目のブートストラップ標本平均

[1]36.1

この場合のブートストラップ標本平均も元の標本平均 $\bar{x}$ と非常に近い値になっている. 大量のブートストラップ標本抽出によって, 本質的に重要な情報が得られる可能性が期待できる. これから R プログラムを用いてブートストラップ標本抽出を 2000 回行い,2000 個のブーストラップ平均値 $\overline{X}^{*1}$,...,$\overline{X}^{*2000}$ を計算する. R プログラミング

mean.boot< −numeric(2000)

set.seed(314)

for(b in 1:2000)

i< −sample(1:10,replace=TRUE)

1 から 10 まで無造作に抽出

women.boot< −women.h[i]

b 回目のブートストラップ標本

mean.boot[b]< −mean(women.boot)

b 回目のブートストラップ標本平均

hist(mean.boot,freq=F,xlab=”bootstrap mean”,main=””)

平均のヒストグラム ここでブートストラップ分布のパーセント点を用いて,「真の自殺率」に関する信頼区間を求める.

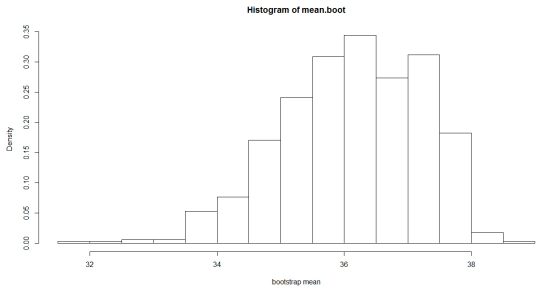

図 1　男性自殺率データを用いた 2000 回の標本抽出によるブートストラップ平均のヒストグラム

>sort(mean.boot)[c(0.025*2000,0.975*2000)]

両側 95 パーセントの信頼区間

[1]13.62 14.48

よって, 両側 95 パーセントの信頼区間は [13.62,14.48] となる.

これらより,2000 回ブートストラップ標本を出すとほとんどが 13.62 から 14.48 の間の数字になることが分かる.

6　女性自殺率のデータを使ったブートストラップ法

女性自殺率のデータから上のブートストラップ法を用いて

標本平均、14.06　標本不偏、0.5848889

1 回目のブートストラップ標本

12.4 14.7 13.8 14.7 14.7 14.2 14.7 14.2 13.9 13.8

1 回目のブートストラップ標本平均、14.11

両端 95 パーセントの信頼区間、[13.62,14.48]

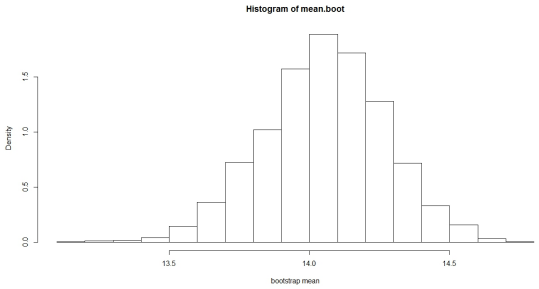

図 2　女性の自殺率データを用いて 2000 回の標本抽出によるブートストラップ平均のヒストグラム

7　分布表を用いた男女の信頼区間

男性の両側 95 パーセントの信頼区間 [26.5157321,46.0642679]

女性の両側 95 パーセントの信頼区間 [13.4199377,14.000623]

8　まとめ

「真の自殺率」に対する両側 95 パーセントの信頼区間は男性の場合,[33.98,37.89] 女性の場合,[13.62,14.48] だった. このブートストラップ標本平均は約 95 パーセントは, 男性で 33.62 以上 37.89 以下の間となり, 女性の場合 13.62 以上 14.48 以下の間の数値になる. この両側 95 パーセントの信頼区間, すなわち「真の自殺率」を求めるのにブートストラップ法を使うメリットは, 普通に信頼区間を求めるのに必要とされているさまざまな分布表を使うことが一切無いので簡潔に計算することができる, またグラフで見ることも出来るので理論的に説明し辛いところも簡潔に説明することが出来る. デメリットとしては, 正規分布を用いる手法と違い物事を大雑把にとらえているので正確な数値が出ない.

9　参考文献

汪金芳・桜井裕二,「R で学ぶデータサイエンス 4　ブートストラップ入門」, 共立出版, 東京,2011.

汪金芳・田栗正章,「計算統計　統計科学のフロンティア 11　ブートストラップ入門」, 岩波書店, 東京,2003

白旗慎吾,「統計解析入門」, 共立出版, 東京,1992.