回帰分析におけるブートストラップ法の特性

2015SS067 住田匡拡
指導教員：小藤俊幸

1　はじめに

ブートストラップ法とは，Efron(1979) によって提唱された，データ解析の確からしさを評価するための統計的手法のひとつである．誤差推定，信頼区間の構成，仮説検定などに用いられ，従来の複雑な数式に基づく理論を莫大な数値計算による単純なシミュレーションで置き換えることができる．[1]

日本における自殺者数は高い水準で推移し続け，大きな社会的関心を呼んでいる．本研究では，日本における離婚件数，自殺者数，失業率のデータを用いてブートストラップ法による分析を行い，自殺者数と離婚件数・失業率の相関に関する考察を行う．

2　分析するデータ

「平成 29 年人口動態統計の年間推計」[2],「平成 29 年中における自殺の状況」[3],「平成 29 年労働力調査」[4] から，平成 8 年から平成 29 年までの 22 年間の離婚件数，自殺者数，失業率のデータを用いて分析を行う．

3　ブートストラップ法を用いた分析

3.1　信頼区間を求めるために必要な反復回数

教科書では，両側 95 %信頼区間を求めるために必要な反復回数を 2000 回としている．ここで，反復回数を 100 回，40000 回にした場合，両側 95 %信頼区間にどの程度差が出るのかを離婚件数と男性の自殺者数のデータを用いて検証する．

3.1.1　データの準備

> divorce <- c(206955, 222635, ⋯, 216798, 212000)
＃ 22 年間の離婚件数を divorce に格納
> m.suicide <- c(15393, 16416, ⋯, 15121, 14826)
＃ 22 年間の男性の自殺者数を m.suicide に格納

3.1.2　ブートストラップ標本の抽出

まず，1 から 22 までの整数 $1, 2, \cdots, 22$ から重複を許して 22 個の整数を無作為に抽出する．

> b <- sample(1:22, 22, replace=T)
＃ 1 から 22 までの整数から重複を許して 22 回無作為に抽出
[1] 14 20 17 22 20 ⋯ 19 2 13 14 11

次に、抽出した整数を用いて，ブートストラップ標本を構成する．

> divorce.b <- divorce[b]
＃ 1 回目の離婚件数のブートストラップ標本を divorce.b に格納
[1] 253353 270804 231383 ⋯ 222635 251378 222107
> m.suicide.b <- m.suicide[b]
＃ 1 回目の男性の自殺者数のブートストラップ標本を m.suicide.b に格納
[1] 23472 23272 18787 ⋯ 16416 22283 17386
> cor(divorce.b, m.suicide.b)
＃ 1 回目の離婚件数と男性の自殺者数のブートストラップ標本に基づく相関係数
[1] 0.8517168

この場合，ブートストラップ標本に基づく相関係数 0.8517168 は初期標本に基づく相関係数 0.8639531 とかなり近い値になっていることが分かる．ブートストラップ標本は毎回異なる可能性がきわめて高く，このように得られるブートストラップ標本に基づく相関係数も毎回変動することが予想できる．しかし，ブートストラップ標本は元のデータからの無作為標本となっているため，ブートストラップ標本に基づく相関係数の期待値は初期標本に基づく相関係数と一致する．

3.1.3　反復回数の選択

反復回数が 2000 回の場合

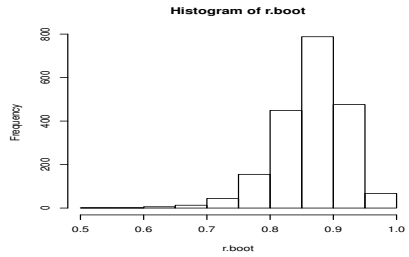

図 1　2000 組のブートストラップ標本に基づく離婚件数と男性の自殺者数の相関係数のヒストグラム

> sort(r.boot)[c(0.025*2000, 0.975*2000)]
＃両側 95 %信頼区間
[1] 0.7350273 0.9530029

よって，両側 95 %信頼区間は (0.7350273, 0.9530029) である．

反復回数が 100 回の場合

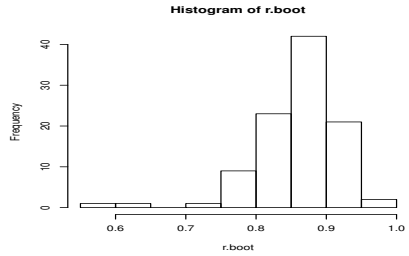

図 2　100 組のブートストラップ標本に基づく離婚件数と男性の自殺者数の相関係数のヒストグラム

両側 95 %信頼区間は (0.6103702, 0.9442927) である．

反復回数が 40000 回の場合

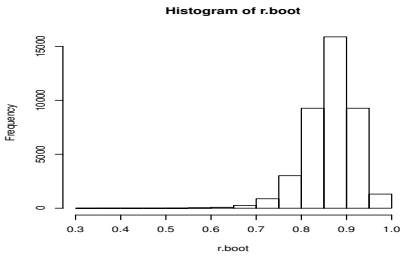

図 3　40000 組のブートストラップ標本に基づく離婚件数と男性の自殺者数の相関係数のヒストグラム

両側 95 %信頼区間は (0.7398235, 0.9531642) である．

反復回数を 2000 回から 100 回に減らした場合，ブートストラップ信頼区間が広くなり精度が悪くなっている．また，反復回数を 40000 回に増やした場合，ブートストラップ信頼区間にほとんど差が出ていない．よって，反復回数は 2000 回で十分であると考え，以降の反復回数を 2000 回とする．

3.2　離婚件数と女性の自殺者数の相関

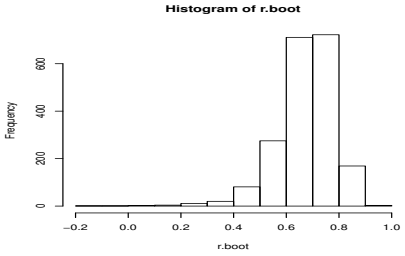

図 4　2000 組のブートストラップ標本に基づく離婚件数と女性の自殺者数の相関係数のヒストグラム

両側 95 %信頼区間は (0.4151210 0.8365684) である．

3.3　男性の自殺者数と失業率の相関

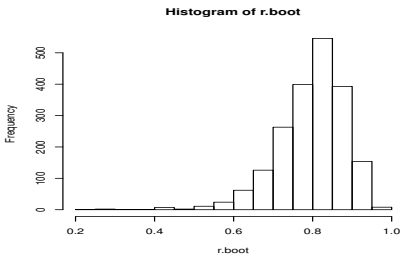

図 5　2000 組のブートストラップ標本に基づく男性の自殺者数と失業率の相関係数のヒストグラム

両側 95 %信頼区間は (0.6005244, 0.9273907) である．

3.4　女性の自殺者数と失業率の相関

両側 95 %信頼区間は (0.4808620, 0.9204715) である．

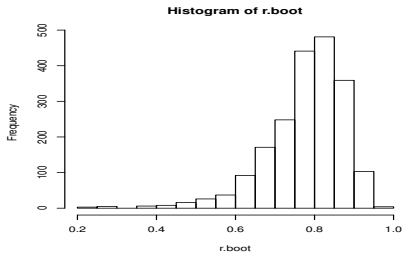

図 6　2000 組のブートストラップ標本に基づく女性の自殺者数と失業率の相関係数のヒストグラム

4　考察

「真の離婚件数と男性の自殺者数の相関係数」に対する両側 95 %信頼区間は (0.7350273, 0.9530029) であり，離婚件数と男性の自殺者数に強い相関がみられることが分かる．ただし，離婚をした男性が自殺をしていると，ただちに判断することはできない．離婚件数は年間に数万件程度であるため，男性の自殺者数に占めるその年に離婚した人の割合があまり大きいとは考えられない．

「真の男性の自殺者数と失業率の相関係数」に対する両側 95 %信頼区間は (0.6005244, 0.9273907) であり，男性の自殺者数と失業率にも強い相関がみられることが分かる．

一方で女性は，どちらにも相関はみられるものの，男性よりも弱いことが分かる．このことから，男性は離婚や失業によって自殺のリスクが高まるが，女性はそのような理由に対して，ある程度の耐性があると考えられる．

5　おわりに

ブートストラップ反復回数は，ごく小さい場合を除き精度には関係しないことを確認することができた．

ブートストラップ法では，他の統計的手法で必要となる様々な分布表を一切使わないため，信頼区間を簡潔に求めることができた．また，ヒストグラムを描くことで「相関係数のばらつき方の可能性」を見ることができた．

しかし，データの分布がでたらめになっている場合，信頼区間が広くなってしまい正確な数値が出ないため，適用できる問題は限られていると考えられる．

参考文献

[1] 汪　金芳・桜井裕仁：「ブートストラップ入門」．共立出版，東京，2011.

[2] 厚生労働省：「平成 29 年人口動態統計の年間推計」．https://www.mhlw.go.jp/toukei/saikin/hw/jinkou/suikei17/index.html

[3] 警察庁：「平成 29 年中における自殺の状況」．https://www.npa.go.jp/publications/statistics/safetylife/jisatsu.html

[4] 総務省統計局：「平成 29 年労働力調査」．https://www.stat.go.jp/data/roudou/sokuhou/nen/ft/index.html