

被験者内反復測定による信頼性係数の ブートストラップ推定

教育心理学コース 奥 村 太 一

Bootstrap estimation of reliability coefficient by repeated measurement

Taichi OKUMURA

In this article, the parametric bootstrap estimation of the reliability coefficient by repeated measurement is proposed. At first, a hierarchical model for calculating reliability from repeated measurement is presented. Next, the parametric bootstrap method for calculating the precision of estimation of the reliability coefficient is presented. The validity of applying the bootstrap method to this problem is examined by a simulation study. The result of the simulation study tells that the calculated bootstrap confidence interval tends to underestimate the statistical power, while the type I error is controlled.

目 次

1. 反復測定による信頼性係数の推定

A はじめに

B モデル

2. ブートストラップ法による信頼性係数の信頼区間

A 推定精度を求めるための数値的方法

B パラメトリック・ブートストラップ法による信頼区間の算出

3. シミュレーション

A 概要

B 結果

4. 結論

5. プログラム

1. 反復測定による信頼性係数の推定

A はじめに

テストの信頼性を知ることは、そのテストを運用する上で重要な問題である。テストの信頼性が低ければそもそもテスト得点の値が十分な精度を持つと保証できないわけであるし、その得点を用いて統計解析を行う場合にも相関の希薄化など好ましくない現象が生じることが知られている。

テストの信頼性を推定する方法には、再検査法、折半法、内的一貫性による方法などいくつかある。再検

査法とは被験者に 2 度同じ検査を実施し、得られた測定値間の相関係数をもって信頼性係数の推定値とするものである。折半法はテストを何らかの方法で 2 つの部分テストに折半し、その点数間の相関係数とスピアマン・ブラウンの公式を用いて信頼性係数を推定する。内的一貫性による方法では、 α 係数と呼ばれる指標を用いて信頼性係数を推定する。ただし、これらの方法で推定された信頼性係数が妥当なものであるためには、測定が平行測定であるなどの制約が満たされている必要がある。

ここで、信頼性という概念はいわゆる「テスト」といわれる形式のみに限定されるものではないことに注意する必要がある。むしろ、自然科学一般の実際的場面から考えれば、「測定」という行為に付随するものと考えるのが自然であろう。すなわち、ある測定についてどの程度の精度が保証されているのかということである。こうした概念に立ち返れば、測定の信頼性を確かめるのに最も基礎的な方法は、「繰り返し測定する」ことである。なぜならば、折半法や内的一貫性による方法は、測定が何らかの下位尺度から成り立っていることを前提としており、従って結局のところ「テスト」という概念から抜け出せないでいるからである。加えて、異なる項目が平行測定の仮定を満たしているというのは実際問題非現実的である。従って、上記の 3 つの方法のうち、信頼性を確認する上で測定の基

本的概念からして最も忠実であると思われるは再検査法であろう。しかしながら、再検査法では測定回数が2回という限定がついている。これは実際に再検査法によって信頼性を推定する上で大きな制約となる。というものも、第一に2回を前提として測定をしたとしても、すべての被験者について2回の測定値が完全に得られるという保証はない。すなわち、欠測が生じる可能性がある。このような場合、1回しか測定値を得ることのできなかった被験者についてどのように扱えばよいのか、明確な答えがあるわけがない。そうした被験者を除いて相関係数を算出しているというのが現状ではなかろうか。これは明らかに情報の損失である。第二に、2回でなくともそれ以上の回数測定を行った方が信頼性を高い精度で推定できるのではないかという期待もある。実際問題、生理学的指標を用いて測定を行うような領域においては測定誤差の影響を除くために多数回同じ刺激について反復測定を実施し、その平均値を統計解析に用いるというような行為が日常的に行われているのである。こうした測定がなされたとき、その測定の信頼性を知ることは重要ではないだろうか。

以上を踏まえて、本研究では、新たに各被験者に任意回数の測定を行った場合の信頼性係数の推定方法について考察する。また、ブートストラップ法によって算出される信頼性係数の信頼区間の妥当性をシミュレーションによって確認する。

B モデル

j 番目の被験者($j = 1, \dots, J$)における i 回目の測定値($i = 1, \dots, n_j$)を x_{ij} とする。このとき、 x_{ij} を j 番目の被験者の真値 t_j および全体平均 μ 、 μ からのランダムな個人間変動 u_j および個人内変動 r_{ij} を用いて、

$$x_{ij} = t_j + r_{ij}$$

および

$$t_j = \mu + u_j$$

と表す。ただし、 $r_{ij} \sim N(0, \sigma^2)$ および $u_j \sim N(0, \tau^2)$ とする。このモデルは、階層的線形モデル(hierarchical linear models; HLM)における一元配置分散分析モデル(one-way ANOVA model)に相当する(Raudenbush & Bryk, 2002)。

このモデルにおいて、テストの信頼性係数 ρ はいわゆる級内相関係数に相当するものであり、

$$\rho = \frac{\tau^2}{\sigma^2 + \tau^2}$$

で表される。従って、信頼性係数の点推定値を求める

には上記のモデルにおいて σ^2 および τ^2 を推定すればよい。これらの推定値は、EMアルゴリズム(Dempster, Rubin, & Tsutakawa, 1981)や反復一般化最小2乗推定法(Goldstein, 1986)など一連の数値的解法を用いて求めることができる。これらの数値的解法は多くの統計ソフトウェアに標準で実装されている。

2. ブートストラップ法による信頼性係数の信頼区間

A 推定精度を求めるための数値的方法

以上で反復測定モデルおよびそれにもとづくテスト信頼性係数の点推定値の算出は行えるわけであるが、これではその推定精度がわからない。当然のごとく、信頼性係数の推定値は標本変動するわけであるから、点推定値のみならずその推定精度を考慮することは重要である。ただし、上記のモデルにおいて信頼性係数 ρ の信頼区間を解析的に導出するのは数学的に極めて煩雑となる。そこで、ここでは数値的な方法にもとづいてそれを算出する方法について考察をしたい。

統計量の推定精度を求めるための数値的方法としては、ジャックナイフ法、ブートストラップ法が一般的である。ジャックナイフ法は、データの一部分を取り除いて統計量を計算することを繰り返し、その変動をもって推定精度を求めようというものであり、PISAなどの国際学力調査の分析にも用いられている(OECD, 2005)。ブートストラップ法には大きく分けてノンパラメトリック・ブートストラップ法とパラメトリック・ブートストラップ法がある(汪・大内・景・田栗, 1992)。前者は、標本から復元抽出されたブートストラップ標本から統計量を計算することを繰り返し、その変動をもって推定精度を求めようというものである。これらはいずれもノンパラメトリックな手法にもとづくものであり、簡便ではあるが複雑なモデルに対応することが難しいともいえる。これに対して、パラメトリック・ブートストラップ法は、一度標本から算出された統計量の推定値をもとに新たなブートストラップ標本を発生させ、それぞれについて統計量を求める変動をもって推定精度とするものである。手間はかかるが、複雑なモデルにも柔軟に対応することができる。本研究で採用したモデルは、各測定値が各被験者にネストしているという階層的構造を持っている。こうした複雑なモデルに対しては前二者のノンパラメトリックな方法では対応が難しい。そこで、本研究ではパラメトリック・ブートストラップ法を用いて信頼性係数の推定精度を求めるることを考える。

B パラメトリック・ブートストラップ法による信頼区間の算出

まず、パラメトリック・ブートストラップ法を用いて統計量の信頼区間を求める方法についてまとめておく。ここでは、信頼区間を求める方法として最も一般的なパーセンタイル法を用いる。母集団分布を F_0 とする。

- ① F_0 から大きさ n の標本 $\{x_1, x_2, \dots, x_n\}$ を無作為抽出する。
- ② ①で得られた標本 $\{x_1, x_2, \dots, x_n\}$ にもとづき、 F_0 に含まれるすべての未知パラメタを推定する。
- ③ F_0 における未知パラメタを、②で求めた推定値で置きかえた分布を F_1 とする。
- ④ ③で求めた F_1 から、大きさ n の標本 $\{x_1^*, x_2^*, \dots, x_n^*\}$ を無作為復元抽出する。ここから、統計量 θ の点推定値を求める。
- ⑤ ④を K 回繰り返し、 θ に関する K 個のブートストラップ推定値を得る。
- ⑥ K 個のブートストラップ推定値から区間 $\{\omega_{\alpha/2}, \omega_{1-\alpha/2}\}$ を求める。

以上の手順を経て求められた区間 $\{\omega_{\alpha/2}, \omega_{1-\alpha/2}\}$ が、パーセンタイル法による統計量 θ の $100(1-\alpha)\%$ ブートストラップ信頼区間である。

ここまでで明らかなように、ブートストラップ法によって信頼区間を算出することは計算機と基礎的なプログラミングの技術さえあれば非常に容易である。しかし、こうした数値的方法には一つ重要な欠点がある。それは、この方法がある統計量について妥当かどうかは実際にシミュレーションによって検証してみないとわからないということである。ブートストラップ法は統計学の漸近論にその理論的根拠を求めることができるが、それが適用可能であるかどうかをこうした複雑なモデルについて理論的に判定するのは難しい。すなわち、本研究で提示したモデルによって推定された信頼性係数のブートストラップ信頼区間が妥当なものであるかどうかは、シミュレーションによって検証する必要がある。

3. シミュレーション

A 概要

本章では、被験者内反復測定によって得られた信頼性係数のブートストラップ信頼区間が妥当なものであ

るか、特に信頼区間から計算される危険率および検定力に焦点を当て、シミュレーションによって検証する。シミュレーションにはフリーのデータ解析環境である R (R Development Core Team, 2005) を用いる。また、信頼性係数の点推定値の算出には nlme パッケージにおける lme 関数を用いる。

シミュレーションでは、信頼性係数のブートストラップ信頼区間から計算される危険率および検定力が真のそれをきちんと再現できているかどうかで判断することにする。真の検定力を解析的に導出するのは困難なため、 σ^2 および τ^2 を適当に設定してデータを発生させ、そこから信頼性係数 ρ の帰無分布および非心分布を数値的に発生させる。また、ブートストラップ信頼区間から計算される検定力は、各データセットについてブートストラップ信頼区間が真の ρ を含んでいれば 0、含んでいなければ 1 とすることを繰り返し 1 となる割合をもって計算することにする。シミュレーションにおいて帰無仮説は $\rho = 0.5$ に設定する。シミュレーションにおいて操作した σ^2 、 τ^2 および ρ の値を表 1 に示す。いずれのシミュレーションにおいても、 $\mu = 10$ とし、20人の被験者それぞれに 2 回ずつ反復測定を実施した状況を想定した。データセットは 1000 個発生させ、またそれぞれのデータセットについてブートストラップ標本は 1000 個発生させた。また、検定の際の第 1 種の誤りは $\alpha = 0.05$ とした。

表 1

	条件1	条件2	条件3	条件4
σ^2	4.00	3.00	2.00	1.00
τ^2	4.00	4.00	4.00	4.00
ρ	0.50	0.57	0.67	0.80

B 結果

シミュレーションの結果を表 2 にまとめる。「見かけ上の検定力」とは、 σ^2 および τ^2 を表 1 のように設定してデータを発生させ、信頼性係数 ρ の帰無分布および非心分布を数値的に得ることによって得られたものである。これに対して、「実際の検定力」とは、前項で述べたようにブートストラップ信頼区間を用いて計算された検定力である。

表 2

	条件1	条件2	条件3	条件4
見かけ上の検定力	0.05	0.096	0.238	0.702
実際の検定力	0.058	0.078	0.185	0.629

この結果を見ると、2つのことがわかる。まず、条件1における結果からブートストラップ信頼区間は検定における第1種の誤りをある程度統制していると考えられる。これに対して、条件2から条件4の結果からブートストラップ信頼区間から計算された検定力は本来あるべき値よりも少し低くなってしまっていることがわかる。すなわち、ブートストラップ信頼区間を用いて統計的検定を行った場合、その結果はある程度保守的になることが予想される。

4. 結論

本研究では、被験者に任意回の反復測定を実施した状況で測定の信頼性係数の推定精度をブートストラップ推定することの妥当性について統計的検定における第1種の誤りおよび検定力の観点から調べた。シミュレーションの結果、算出されたブートストラップ信頼区間は、第1種の誤りは統制できているものの、検定力は本来あるべき値よりもある程度低めになってしまったことがわかった。このことは、解析的に結果を求められない問題に対して容易に従来ある数値的方法で対処しようとするこの危うさを示したものであるといえるだろう。今後は、被験者内反復測定によって得られた信頼性係数の推定精度を正しく算出するための方法についてさらに検討を進める必要がある。

5. プログラム

本研究で用いたRプログラムのソースコードを以下に提示する。

```
setwd("C:/Documents and Settings/Taichi OKUMURA/
My Documents/研究/KIYO/programs/")
#-----
# Data Generation
#----- Generate Data -----
gendata <- function(n,mu,sigma,tau,ndata,filename) {
J <- length(n)
datalist <- list()
for(i in 1:ndata) {
testdata <- c()
for(j in 1:J) {
tj <- mu + rnorm(1,mean=0,sd=sqrt(tau))
x <- c(tj+rnorm(n[j],mean=0,sd=sqrt(sigma)))
person <- rep(j,n[j])
datanew <- cbind(x,person)
testdata <- rbind(testdata,datanew)
}
testdata <- as.data.frame(testdata)
datalist[[i]] <- testdata
}
return(datalist)
}
#-----
# Sampling distribution of rho
#----- sampledist -----
sampledist <- function(datalist,filename) {
ndata <- length(datalist)
J <- max(datalist[[1]][,2])
N <- nrow(datalist[[1]])
library(nlme)
rho <- c()
for(i in 1:ndata) {
# Data-no-henkan
lstx <- list()
attach(datalist[[i]])
nj <- tapply(person,person,sum) /
tapply(person,person,mean)
p <- 1
for(j in 1:J) {
lstx[[j]] <- x[p:sum(nj[1:j])]
p <- p + nj[j]
}
}
```

```

}

detach(datalist[[i]])
# REML estimation
fit0 <- lme(x~1, random=~1 | person, data=datalist[[i]])
s0 <- as.numeric(VarCorr(fit0)[2,1])
t0 <- as.numeric(VarCorr(fit0)[1,1])
rho[i] <- t0 / (s0 + t0)
}
save(rho, file=filename)
}
#----- h1dist -----
h1dist <- function(datalist,filename) {

ndata <- length(datalist)
J <- max(datalist[[1]][,2])
N <- nrow(datalist[[1]])
library(nlme)
rho1 <- c()
for(i in 1:ndata) {
# Data-no-henkan
lstx <- list()
attach(datalist[[i]])
nj <- tapply(person,person,sum) / tapply(person,person,
mean)
p <- 1
for(j in 1:J){
lstx[[j]] <- x[p:sum(nj[1:j])]
p <- p + nj[j]
}
detach(datalist[[i]])
# REML estimation
fit0 <- lme(x~1, random=~1 | person, data=datalist[[i]])
s0 <- as.numeric(VarCorr(fit0)[2,1])
t0 <- as.numeric(VarCorr(fit0)[1,1])
rho1[i] <- t0 / (s0 + t0)
}
save(rho1, file=filename)
}
#----- sampledist 2 -----
sampledist2 <- function(datalist) {
ndata <- length(datalist)
J <- max(datalist[[1]][,2])
N <- nrow(datalist[[1]])
library(nlme)
rho <- c()
for(i in 1:ndata) {
# Data-no-henkan
lstx <- list()
attach(datalist[[i]])
nj <- tapply(person,person,sum) / tapply(person,person,
mean)
p <- 1
for(j in 1:J){
lstx[[j]] <- x[p:sum(nj[1:j])]
p <- p + nj[j]
}
detach(datalist[[i]])
nj <- as.vector(nj)
# REML estimation
fit0 <- lme(x~1, random=~1 | person, data=datalist[[i]])
g0 <- fixef(fit0)[1]
}
}

```

```

s0 <- as.numeric(VarCorr(fit0)[2,1])
t0 <- as.numeric(VarCorr(fit0)[1,1])
bootstrapdata <-
gendata2(n=nj,mu=g0,sigma=s0,tau=t0,ndata=nitr)
rho.boot[i,] <- sampledist2(datalist=bootstrapdata)
}
save(rho.boot,file=filename)
}
#-----
# Calculate True Power
#-----
calc.true.power <- function(rho0, rho1, alpha){
critical.value.l <- quantile(rho0, alpha/2)
critical.value.u <- quantile(rho0, 1-alpha/2)
(mean((rho1 < critical.value.l) | (critical.value.u < rho1)))
}
#-----
# Calculate Bootstrap Power
#-----
calc.power <- function(truerho, rho.boot, alpha){
ndata <- dim(rho.boot)[1]
reject <- c()
for(i in 1:ndata){
reject[i] <- ((truerho > quantile(rho.boot[i,], 1-alpha/2)) ||
(truerho < quantile(rho.boot[i,], alpha/2)))
}
(mean(reject))
}
#-----
# RUN SIMULATION
#-----
gendata(n=c(2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2),
mu=10, sigma=4, tau=4, ndata=1000, filename=
"dataset.data")
load("dataset.data")
sampledist(datalist=datalist, filename="nulldist.data")
load("nulldist.data")
bootstrap(datalist=datalist, nitr=1000, filename="rho.data")
load("rho.data")
calc.true.power(rho0=rho, rho1=rho, alpha=0.05)
calc.power(truerho=0.5, rho.boot=rho.boot, alpha=0.05)
#
# (指導 南風原朝和教授)

```

引用文献

- Dempster, A. P., Rubin, D. B., & Tsutakawa, R. K. 1981 Estimation in covariance components models. *Journal of the American Statistical Association*, 76, 341-353.
- Goldstein, H. 1986 Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika*, 73, 43-56.
- OECD 2005 PISA 2003 Data analysis manual.
- 汪金芳・大内俊二・景平・田栗正章 1992 ブートストラップ法 行動計量学, 19, 50-81.
- Raudenbush, S. W. & Bryk, A. S. 2002 Hierarchical linear models: Application and data analysis methods (2nd ed.). Sage.
- R Development Core Team 2005 R: A language and environment for statistical computing.