ブートストラップ信頼区間・ノンパラメトリック検定

樋口さぶろお https://hig3.net

龍谷大学大学院 理工学研究科 数理情報学専攻

理論物理学特論　L14 (2023-01-11 Wed)

最終更新: Time-stamp: "2023-01-14 Sat 08:48 JST hig"

今日の目標

- パラメトリックブートストラップ法を説明できる
- ブートストラップ信頼区間を求められる
- (ブートストラップでない) ノンパラメトリック統計を説明できる

L13-Q3
Quiz 解答: ブートストラップ標本抽出

1. $\frac{1}{20}[8+8+\cdots+12]=11.$ バイアス $b=11-10=1.$
2. $10-(11-10)=2\times 10-11=9.$
3. $\frac{1}{20-1}[(8-11)^2+\cdots+(12-11)^2]=\frac{24}{19}.$ 標準誤差の推定値は $(24/19)^{1/2}.$
4. $0.05, 0.95$ quantile は $8, 12.$ よって, $10-(12-10)<\theta<10+(8-10)].$

ここまで来たよ

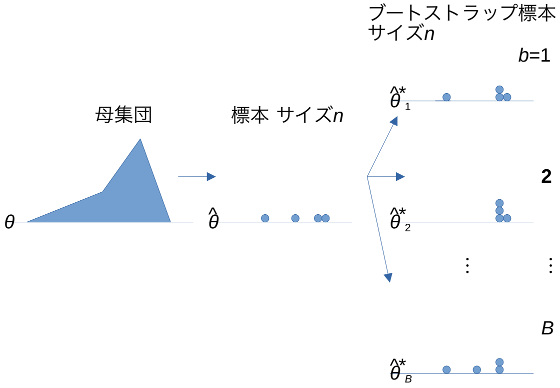
ブートストラップ標本 I

ブートストラップ標本とは, 経験分布から抽出したサイズ $k = n$ の標本
(n は経験分布の元になった標本のサイズ).
別の言い方: 標本を母集団と思い直して復元抽出して孫標本.
ふつう, 多数個 (B 個) のブートストラップ標本を考える.

(母集団) : (標本) ≈(標本) : (ブートストラップ標本)
親 : 子 ≈子 : 孫

親 $\xrightarrow{\text{標本抽出}}$ 子

$$\text{真のパラメタ}\theta - \text{標本からの推定値}\hat{\theta}$$
$$\approx \text{標本からの推定値}\hat{\theta} - \text{ブートストラップ推定値}\hat{\theta}^*$$


ブートストラップ標本
サイズn
b=1
母集団
標本 サイズn
θ
$\hat{\theta}$
$\hat{\theta}^*_1$
2
$\hat{\theta}^*_2$
B
$\hat{\theta}^*_B$

ブートストラップ推定値

ブートストラップ標本を抽出する試行を考えたとき, その母期待値 $\mathrm{E}[\hat{\theta}^*]$ をブートストラップ推定値という.

Monte Carlo 近似

ブートストラップ推定値は, Monte Carlo 近似できる. B 個のブートストラップ標本 b の統計量 $\hat{\theta}_b^*$ を求めて, 標本期待値をとる.

$$\hat{\theta}^* = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^*$$

.

B: ブートストラップ反復回数

バイアス, バイアス補正値

バイアス $b = \mathrm{E}[\text{標本推定値}\hat{\theta}] - \text{真の値}\theta$.

$$
\begin{aligned}
\text{真の値}\theta =& \mathrm{E}[\text{標本推定値}\hat{\theta}] - \text{バイアス}b \\
\approx& \text{標本推定値}\hat{\theta} - (\text{ブートストラップ推定値}\hat{\theta}^* - \text{標本推定値}\hat{\theta}) \\
=& 2 \times \hat{\theta} - \hat{\theta}^* \\
=& \text{バイアス補正 (された) 値}
\end{aligned}
$$

ブートストラップ推定値は, ブートストラップ反復回数 $B = 50$–300 の Monte Carlo 近似で求める.
ブートストラップの近似値が標本推定値より左に出たら, 補正値は標本推定値より右に置け. 標本推定値がもともと左にでちゃってただろうから, その分戻す, という考え方.

ここまで来たよ

パラメトリックブートストラップ法

- X の母分布の関数形 f が既知, パラメタ θ が未知とする: $f(x|\theta)$.
- r を, X の標本から計算される量とする
- r の分布を知りたい

アイデア: ブートストラップ法なら, ブートストラップ標本から r を計算することもできるが, 母分布を推定してそこから乱数で取り出した方が正確 or 速いかも.

例

- $\boldsymbol{X} = (X, Y) \sim \mathrm{N}(\boldsymbol{\mu}, \Sigma)$. $\theta = (\boldsymbol{\mu}, \Sigma)$ 母平均値ベクトル, 母分散母共分散行列 は未知.
- r を, $\boldsymbol{X}$ のサイズ n の標本から計算される標本相関係数とする
- 標本相関係数 r の分布を知りたい

パラメトリックブートストラップ法の手続き

1. 標本から θ の推定値 $\hat{\theta}$ を得る (ブートストラップと無関係な方法で)
2. $f(X|\hat{\theta})$ にしたがう擬似乱数を生成し, サイズ n の標本を B 個発生する
3. B 個の標本からそれぞれ r を計算し, r の経験分布を得て, 母分布の近似と思う.

ここまで来たよ

パーセンタイルブートストラップ信頼区間

パーセンタイルブートストラップ信頼区間を求める手続き

パラメタ θ の推定値 $\hat{\theta}$ の信頼係数 $1-\alpha=0.99$ の信頼区間を求める.

1. 標本からブートストラップ標本 $b=1,\ldots,B$ を作り, ブートストラップ推定値 $\hat{\theta}_b^*$ を得る.
2. $\hat{\theta}_b^*$ を大きさの順に並べかえたものを, $\hat{\theta}_{(1)}^* \leq \cdots \leq \theta_{(B)}^*$ のように添字に括弧をつけて書く.
3. 両側の極端な $B \cdot \frac{\alpha}{2}$ 個を取り除いた

$$\hat{\theta}^*_{(B\cdot\frac{\alpha}{2})} \leq \theta \leq \hat{\theta}^*_{(B\cdot(1-\frac{\alpha}{2}))}$$

がパーセンタイルブートストラップ信頼区間 (percentile bootstrap confidence interval) である.

パーセンタイルブートストラップ信頼区間の根拠

- 母分布=経験分布 と思うと, 得られる $\hat{\theta}$ のうち $1-\alpha$ が収まる区間を与えている.
- バイアスの推定のときに使った,「子から親の性質を知るには, 子と孫の関係を逆に使う」ことをしていない.
  - その結果, 真の θ のまわりの $\hat{\theta}$ の分布が左右対称であるときのみ, 正確と期待できる.
  - パラメタを $\xi = h(\theta)$ のように変換するときの性質がよいので, 上の欠点にもかかわらず使われることがある.
- 歴史的には, 当初, 深く考えずに提案された (?)

ここまで来たよ

13 ブートストラップ法

14 ブートストラップ信頼区間・ノンパラメトリック検定
- 復習: ブートストラップ法, 標準誤差, バイアス補正
- パラメトリックブートストラップ法
- パーセンタイルブートストラップ信頼区間
- 基本的ブートストラップ信頼区間
- ノンパラメトリック検定

ブートストラップ信頼区間 I

母分布で, 区間内が $1-\alpha$ になるように調節された, 両端の位置を $t(\frac{\alpha}{2}), t(1-\frac{\alpha}{2})$ とおく.

$$1-\alpha = P(t(\tfrac{\alpha}{2}) \le \hat{\theta} - \theta \le t(1-\tfrac{\alpha}{2})).$$

不等式を母分布の量 θ について解いて,

$$1-\alpha = P(\hat{\theta} - t(1-\tfrac{\alpha}{2}) \le \theta \le \hat{\theta} - t(\tfrac{\alpha}{2})).$$

を, $\hat{\theta}^*_b - \hat{\theta}$ の下側 $\frac{\alpha}{2}$ の位置の値で置き換える.

$$t(\tfrac{\alpha}{2}) = (\hat{\theta}-\theta)\text{の下側}\tfrac{\alpha}{2} = \hat{\theta}\text{の下側}\tfrac{\alpha}{2} - \theta \simeq \hat{\theta}^*_{(B\cdot\frac{\alpha}{2})} - \hat{\theta}$$

定義 (基本的ブートストラップ信頼区間)

$$\hat{\theta} - (\hat{\theta}^*_{(B\cdot(1-\frac{\alpha}{2}))} - \hat{\theta}) \leq \theta \leq \hat{\theta} - (\hat{\theta}^*_{(B\cdot\frac{\alpha}{2})} - \hat{\theta})$$

ブートストラップ標本で左にでがちだったら，もともとの標本推定値が左にでてるだろうから，そのずれくらい，右に信頼区間を広げておかなければいけない．

よりハイテクなブートストラップ信頼区間

ブートストラップ-t $\hat{\theta} - \theta$ のかわりに $(\hat{\theta} - \theta)/\hat{\sigma}$ の分布で考える．この分布が偶関数であるなら正確.

BC$_a$ 上の量の分布が, 近似的に「正規分布から非対称にずれた分布」(歪度が 0 でない分布) にしたがうとして, 歪度を見積もる.

ここまで来たよ

13 ブートストラップ法

14 ブートストラップ信頼区間・ノンパラメトリック検定

- 復習: ブートストラップ法, 標準誤差, バイアス補正
- パラメトリックブートストラップ法
- パーセンタイルブートストラップ信頼区間
- 基本的ブートストラップ信頼区間
- ノンパラメトリック検定

ノンパラメトリック検定

t 検定, カイ二乗検定, 二項検定などは, 母分布が正規分布, 正規分布, 二項分布にしたがうことを仮定して, 母分布のパラメタに関する帰無仮説を立てて始めた.
一方, 次のような状況を考えたいこともある.

1 標本問題・1 群の検定, t 検定相当

未知の分布にしたがう X の標本を得た. 帰無仮説「X の中央値は m に等しい」を棄却できるか?

2 標本問題・2 群の検定, 2 群の t 検定相当

それぞれ未知の分布にしたがう X と Y の標本を得た. 帰無仮説「X の分布は Y の分布より中央値は Y の中央値と等しい」は棄却できるか?

定義 (ノンパラメトリック検定 (広義))

母分布の関数形を具体的には仮定しない検定手法群

定義 (ノンパラメトリック検定 (狭義))

母分布の関数形を具体的には仮定しない検定手法群のうち，標本に含まれるデータの大小の順位から作った検定統計量を使う検定.

- 1 標本問題: ウィルコクソンの符号付き順位和検定 (signed rank sum test)=ウィルコクソン検定 (より単純化したものとして符号検定)
- 2 標本問題: ウィルコクソンの順位和検定 (rank sum test)=マン-ホイットニーの U 検定
- 対応ありの 2 標本問題: 1 標本問題に帰着

条件が多いと，適用範囲は狭くなるが検出力 (power) は大きくなる．つまり，正規分布にしたがう場合，2 群の t 検定では棄却できるけど，ウィルコクソンの符号和検定では棄却できない，となったりする.

ウィルコクソンの順位和検定

ウィルコクソンの順位和検定

- データ: 確率密度関数 $f(x), f(x-\theta)$ にしたがう 2 つの標本 $\{X_1, \ldots, X_m\} \sim f(x), \{Y_1, \ldots, Y_n\} \sim f(x-\theta)$.
- 帰無仮説 X, Y は同じ確率分布にしたがう．すなわち $\theta = 0$
- 対立仮説 $\theta \neq 0$
- 検定統計量 $U_2 = mn + \frac{1}{2}n(n+1) - R, R = \sum_{i=1}^{n} R_i$.
  - $\{X_1, \ldots, X_m\}, \{Y_1, \ldots, Y_n\}$ を混ぜて小さい方から並べたときの, Y_i の順位を R_i とする．$1 \leq R_i \leq m+n$.
  - U_2 は, Y_i が X_j より小さいようなペア (i,j) の個数.
- 棄却域は, 手でも計算できるが, 専用の表が本や Web に載ってる

- 帰無仮説のもとで, $U_2 = mn + \frac{1}{2}n(n+1) - n\frac{1}{2}(m+n+1) = \frac{1}{2}mn$ くらいになりそう．$\theta < 0$ のとき, U_2 はこれより小さい.
- n, m 大のとき, 実は, $U_2 \sim \mathrm{N}(\frac{1}{2}mn, \frac{1}{12}mn(m+n+1))$ となる．正規分布表で棄却域を作れる.
- 標本の中に同点 (タイ (tie)) がある場合も小修正で.
- $U_1 = mn + \frac{1}{2}m(m+1) - R', R' = \sum_{j=1}^{m} R'_j$.

L14-Q1

Quiz(ウィルコクソンの順位和検定)

Y の標本は $\{1.1, 3.3\}$, X の標本は $\{2.3, 4.2, 5.5, 6.3\}$ で与えられる.

1. $Y_i < X_j$ となるペア (i, j) の個数を求めよう.
2. U_2 を求めよう.
3. $Y_i > X_j$ となるペア (i, j) の個数を求めよう.
4. U_1 を求めよう.
5. X,Y は同じ分布にしたがう, という帰無仮説のもとで, $U_2 \geq 7$ となる確率を求めよう.
6. 帰無仮説は有意水準 $\alpha = 0.05$ で棄却されるか.

L14-Q2

Quiz(ウィルコクソンの順位和検定)

Y の標本は $\{1.1, 2.3\}$, X の標本は $\{3.3, 4.2, 5.5\}$ で与えられる.

1. $Y_i < X_j$ となるペア (i, j) の個数を求めよう.
2. U_2 を求めよう.
3. $Y_i > X_j$ となるペア (i, j) の個数を求めよう.
4. U_1 を求めよう.
5. X,Y は同じ分布にしたがう, という帰無仮説は有意水準 $\alpha = 0.05$ で棄却されるか.