# L12 ブートストラップ法

樋口さぶろお　https://hig3.net

龍谷大学 先端理工学部 数理・情報科学課程

多変量解析☆演習　L12(2021-12-23 Thu)

最終更新: Time-stamp: "2021-12-24 Fri 09:35 JST hig"

今日の目標

- ブートストラップ法の考え方を説明できる
- ブートストラップ法で推定量の標準誤差を推定できる
- ブートストラップ法で信頼区間を求められる
- (ブートストラップ法で推定量のバイアスを求められる)

# ここまで来たよ

11 主成分分析 (2)

12 ブートストラップ法
- 母集団・標本抽出・推定ふたたび
- 標準誤差のブートストラップ推定
- ブートストラップ信頼区間
- バイアスのブートストラップ推定

# 母集団と標本 (1) 有限母集団

岩薙林 確率・統計 §§5.1,5.2 | 確率統計 I(2021)L10

## 某アイドルグループの身長ふたたび

- 某アイドルグループ全員 ($\to$ 有限母集団) の身長 $x_i$ の平均値 $\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ を求めたい!
  - メンバー 1 名を等確率で選んでくる, という試行を考えると, 確率変数 $X$ の母平均値 $\mu = \mathrm{E}[X]$.
- メンバー全員分のデータがあれば定義の式使うだけ
- 握手会でメンバー 1 人ずつに質問しなければいけないとしたら?
- 握手会参加券 40 枚集めないで何とかすませたい.

$\rightsquigarrow$ 質問できたメンバー 5 人の身長 (=標本)(独立同分布にしたがう確率変数 $X_1, X_2, \cdots, X_5$) から推定したい.
5 人を '無作為に' 選ぶ (=標本抽出する)
母集団サイズ=46, 標本サイズ=5, 標本の個数=1.

# 母集団と標本 (2) 離散 or 連続型確率変数

岩薙林 確率・統計 §5.1,5.2

賞金額, 個数が謎のスピードくじ (引いて賞金額を見た後で箱に戻す).
賞金額 $X$ は離散型確率変数 $\to$ 無限母集団 (何回でもひけるから).

- 賞金の母平均値 $\mu = \mathrm{E}[X] = \sum_x x \cdot p(x)$ を求めたい.
- くじの中を見れば ($p(x)$ の式を知れば) 定義の式使うだけ.
- しかし, 中を見ることはできない.
- $+\infty$ 回くじを買わず, 何とかすませたい.

$\rightsquigarrow$ 引いた 5 枚のくじの賞金額=標本)(独立同分布にしたがう確率変数 $X_1, X_2, \cdots, X_5$) から推定したい.
5 枚を '無作為に' 選ぶ (=標本抽出する).
母集団サイズ=$+\infty$, 標本サイズ=5, 標本の個数=1.

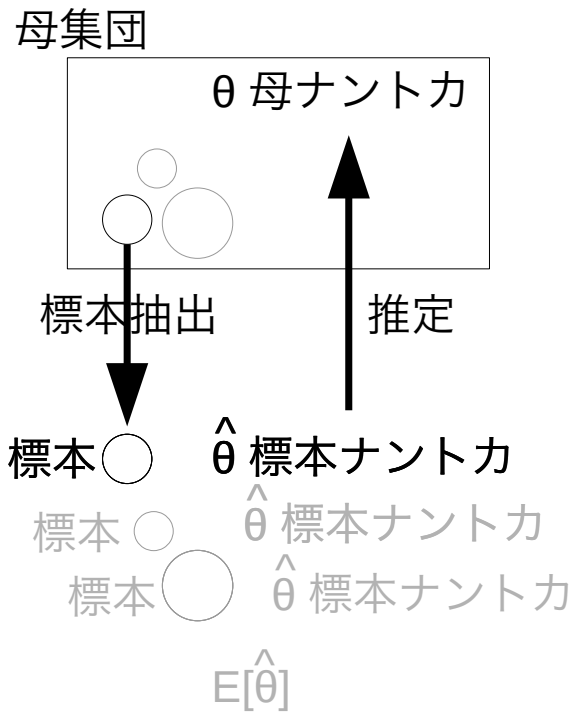
## 母集団・標本抽出・推定

岩薩林 確率・統計 例 11(p.115)

- 母集団 population =考えたい集団. どんな分布, 母平均値, 母分散, などわかっていないことがあるが, 全体を調べるわけにはいかない集団.
- 標本=sample (名詞) =母集団から '無作為に' とってきた一部分
- 標本抽出する sample(動詞)=母集団から '無作為に' とってくる $\rightsquigarrow$ sampling (動名詞)
- 推定 する estimate(動詞) =標本を調べて母集団について正しそうな事実を見つける $\rightsquigarrow$ estimation (名詞)
- 確率変数 $X$, $\overline{X}$ 分布をもつ変数
- 実現値, 観測値 $x$, $\overline{x}$ 標本を 1 つとって確定した値

推定には誤差あるかも. 標本の選び方ごとに答は違うし.

岩薩林 確率・統計 図 p.109,115,137,167



# 標本ナントカによる推定ってどのくらい正確なの?

サイズ $n$ の標本 (sample) $X_1, \ldots, X_n$.

母平均値

母平均値 $\theta = \mu = \mathrm{E}[X]$ の推定量である標本平均値 (sample mean)
$\hat{\theta} = \overline{X} = \frac{1}{n}[X_1 + \cdots + X_n]$.
$X$ が正規分布にしたがうなら t 分布を使った, 信頼係数 $1-\alpha$ の信頼区間 確率統計 I(2021)L1 1

母期待値

母期待値 $\theta = \mathrm{E}[f(X)]$ の推定量である標本期待値 $\hat{\theta} = \overline{X} = \frac{1}{n}[f(X_1) + \cdots + f(X_n)]$.
$f(X)$ が正規分布にしたがうなら t 分布を使った, 信頼係数 $1-\alpha$ の信頼区間 確率統計 I(2021)L11

母分散

母分散 $\theta = \sigma^2 = \mathrm{V}[X]$ の推定量である不偏標本分散 (unbiased sample variance)
$\hat{\theta} = S^2 = \frac{1}{n-1}[(X_1 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2]$.
$X$ が正規分布にしたがうなら カイ二乗分布を使った, 信頼係数 $1-\alpha$ の信頼区間
確率統計 I(2021)L11

**Example**

心配

- 正規分布にしたがってないときは? ふつうしたがってないでしょ. 何分布かわからないこともある.
- 母平均値, 母分散以外の量は? 例:中央値, 第 1 四分位点, パーセント位点, 四分位範囲, …

ブートストラップ (bootstrap) 法は, これらに計算機で力づくで答える方法

母比率 の推定での, ベルヌイ分布の仮定は成立してない心配は少ない.

# ここまで来たよ

11 主成分分析 (2)

12 ブートストラップ法
- 母集団・標本抽出・推定ふたたび
- 標準誤差のブートストラップ推定
- ブートストラップ信頼区間
- バイアスのブートストラップ推定

サイズ $n$ (固定) の標本を繰り返し抽出して, 推定量 (例:標本平均値)$\hat{\theta}$ を繰り返し計算するとする.
この $\hat{\theta}$ は確率変数と思える.

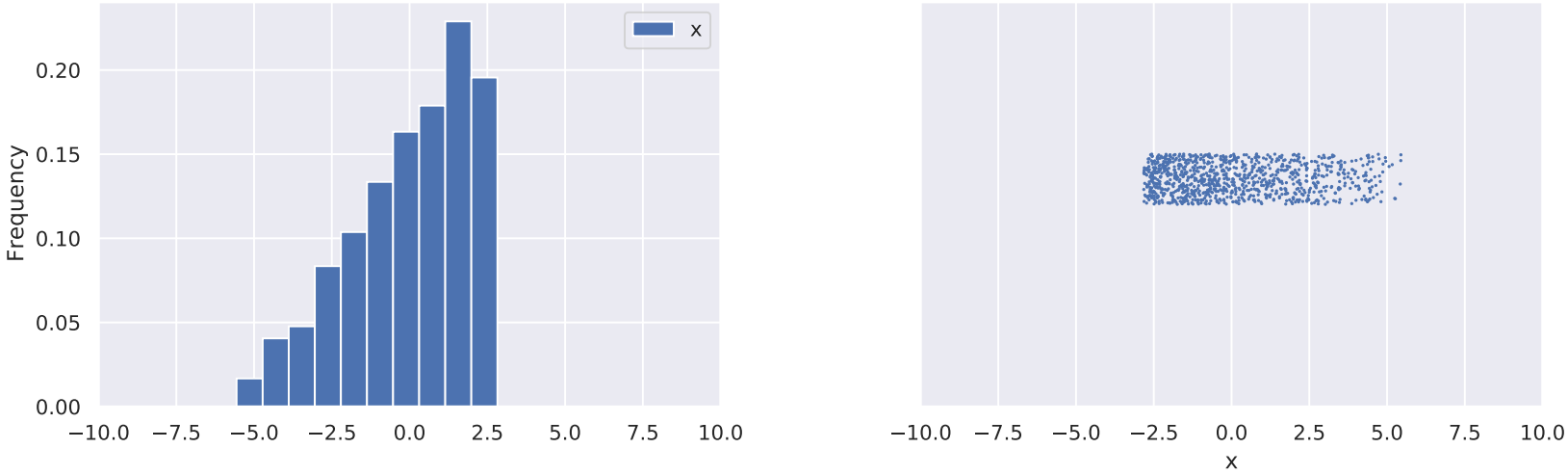
## 標準誤差 (standard error)

$\hat{\theta}$ の誤差は, 次で定まる標準誤差くらいと思える.

$$(\text{標準誤差})^2 = \mathrm{V}[\hat{\theta}]$$

標準誤差は母標準偏差とは異なる. 標準誤差は, 標本サイズ $n$ を大きくすると小さくなる.
母分布がわからないときに, 標準誤差を知りたい.
アイデア 1 標本がたくさん作れれば (「標本」の標本があれば) 推定できる! でも, 母分布がわからないから作れない. そもそも, 標本 1 個からやるのがルール.

アイデア 2 標本分布 (経験分布) は母分布に似てるはず．母分布=標本分布と思っちゃえ．
標本分布 $(x_1, \ldots, x_n,\ n = 1000\ )$ の一例．母分布は正規分布じゃなさそう…

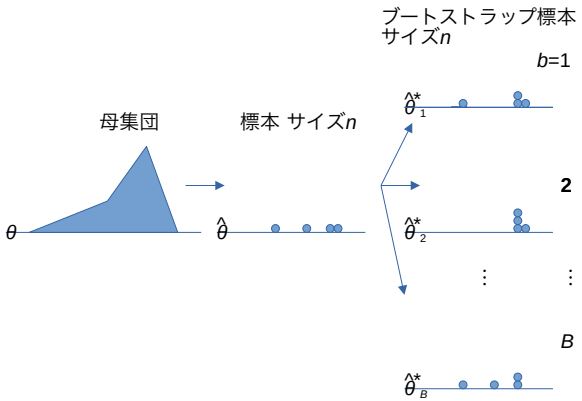


ここから (母分布の確率密度関数をあてずっぽうで 1 次関数で書くのでなく), 次の離散型分布を考え, これを母分布の近似として使う.

$$\text{確率関数 } p(x) = \begin{cases} 1/n & (x = x_1) \\ 1/n & (x = x_2) \\ \vdots & \vdots \\ 1/n & (x = x_n) \\ 0 & (\text{他}) \end{cases}$$

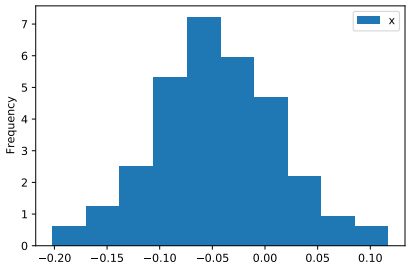
つまり, 標本のデータ点を袋に入れ, 1 点ずつ引く (戻す).
ここから計算機を使って $n$ 個のデータ点を取り出す (取り出した後戻す=重複を許す). こうして得たものをブートストラップ標本 $(x_1^*, \ldots, x_n^*)$ という.

> boot=くつ (ブーツ), strap=ひも (ストラップ).




ここから, ブートストラップ推定量 $\hat{\theta}^* = \hat{\theta}(x_1^*, \ldots, x_n^*)$ が計算できる.
例えば $\frac{1}{n}[x_1^* + \cdots + x_n^*]$.

ブートストラップ標本を, $B$ 個作る ($B = 200$ とか?). ブートストラップ推定量も $\hat{\theta}^*(b)$ $(b = 1, \ldots, B)$ の $B$ 個できる.
平均値のブートストラップ推定量のヒストグラム.




ここから, 次で, 標準誤差の 2 乗を推定する.

## 標準誤差のブートストラップ推定

$$\overline{\hat{\theta}^*} = \frac{1}{B}[\hat{\theta}^*(1) + \cdots + \hat{\theta}^*(B)],$$

$$(\text{標準誤差})^2 \overset{\text{推定}}{\simeq} \frac{1}{B-1}[(\hat{\theta}^*(1) - \overline{\hat{\theta}^*})^2 + \cdots + (\hat{\theta}^*(B) - \overline{\hat{\theta}^*})^2].$$

L12-Q1

## Quiz(ブートストラップ法による標準誤差)

(理解チェックのための不自然なブートストラップ標本です)
標本で, 推定量 $\hat{\theta} = 10$ となった. $B = 20$ ブートストラップ標本で, 推定量 $\hat{\theta}^*$ は, 大きさの順に,
$8, 8, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 12, 12, 12, 12, 12, 12$
となった.

1. 標準誤差を推定しよう.
2. $\alpha = 0.1$ の基本的ブートストラップ区間推定を行おう.

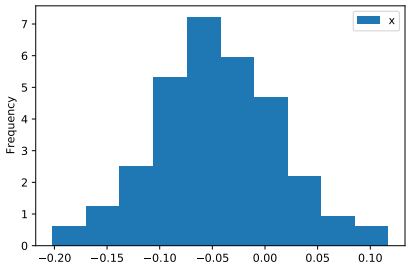
# ここまで来たよ

11 主成分分析 (2)

12 ブートストラップ法
- 母集団・標本抽出・推定ふたたび
- 標準誤差のブートストラップ推定
- ブートストラップ信頼区間
- バイアスのブートストラップ推定

# ブートストラップ信頼区間

推定値のブートストラップ分布のヒストグラム




信頼係数 $\alpha$ の信頼区間を求めるには (=区間推定するには), この上下 $\alpha/2$ を捨てればいいような気がするけど, 実はそうじゃない.

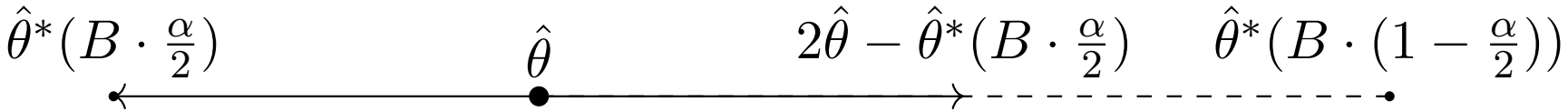
> この間違ったやり方には, パーセンタイル信頼区間という名前までついている. 特殊な場合には正解と同じになることがある.

# 基本的 (basic) ブートストラップ信頼区間

## 母分布の量 $\theta$ の基本的ブートストラップ信頼区間

$\hat{\theta}$: 標本からの推定量
$\hat{\theta}^*(b)$: (小さい方から数えて)$b = 1, \ldots, B$ 個目のブートストラップ推定量.
$\hat{\theta}^*(B \cdot \frac{\alpha}{2})$: 小さい方から数えて, $b = B \cdot \frac{\alpha}{2}$ 個目の $\hat{\theta}^*(b)$.
このとき, $\hat{\theta}$ の信頼係数 $\alpha$ の信頼区間は

$$2\hat{\theta} - \hat{\theta}^*(B(1 - \tfrac{\alpha}{2})) < \theta < 2\hat{\theta} - \hat{\theta}^*(B \cdot \tfrac{\alpha}{2})$$




直感的理解: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.

母分散の区間推定でも, カイ二乗分布の値が不等式の反対側の分母に出てきたでしょ?

$\hat{\theta} - \theta$ の分布を見たとき, 区間内の確率が $1 - \alpha$ になるように調節された, 両端の位置 $\hat{\theta} - \theta = w(\frac{\alpha}{2}), w(1 - \frac{\alpha}{2})$ を考える. ($\alpha$ が小さければふつうは負, 正).

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

不等式を母分布の量 $\theta$ について解いて,

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

$\hat{\theta}$ をブートストラップ推定値 $\hat{\theta}^*$ で近似,
$w(\frac{\alpha}{2})$ を $\hat{\theta}^*(B \cdot \frac{\alpha}{2}) - \hat{\theta}$ で近似,

$$1 - \alpha = P(\hat{\theta} - (\hat{\theta}^*(B \cdot (1 - \tfrac{\alpha}{2})) - \hat{\theta}) \leq \theta \leq \hat{\theta} - (\hat{\theta}^*(B \cdot \tfrac{\alpha}{2}) - \hat{\theta})).$$

## Quiz(ブートストラップ法による標準誤差)

(理解チェックのための不自然なブートストラップ標本です)
標本で, 推定量 $\hat{\theta} = 10$ となった. $B = 20$ ブートストラップ標本で, 推定量 $\hat{\theta}^*$ は, 大きさの順に,
$8, 8, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 12, 12, 12, 12, 12, 12$
となった.

1. 標準誤差を推定しよう.
2. $\alpha = 0.1$ の基本的ブートストラップ区間推定を行おう.

# ここまで来たよ

11 主成分分析 (2)

12 ブートストラップ法
- 母集団・標本抽出・推定ふたたび
- 標準誤差のブートストラップ推定
- ブートストラップ信頼区間
- バイアスのブートストラップ推定

# バイアスのブートストラップ推定

標準誤差のブートストラップ推定は, 「推定値の母分散の推定」だった. 「推定値の母平均値の推定」ってあるの?

$$\frac{1}{B}[\hat{\theta}^*(1) + \cdots + \hat{\theta}^*(B)$$

ふつうは, ここから母平均値 $\theta$ を引いた量 バイアス (bias) で語る.

$$\text{バイアス } b = \mathrm{E}[\hat{\theta}] - \theta$$

$$\text{バイアスのブートストラップ推定値 } b^* = \frac{1}{B}[\hat{\theta}^*(1) + \cdots + \hat{\theta}^*(B)] - \hat{\theta}$$

バイアスが $0$ になることが保証されている推定量も多い.

- 例: 標本平均値, 不偏標本分散, ...

| 不偏=unbiased=バイアスがない |
|---|

そうでない量 (例: 中央値, ...) では役立つ. さらに, バイアスのブートストラップ推定値を使って, 推定値を改善する (バイアス補正) するという方法がある. (区間推定で $\alpha \to 1$ としたようなもの.