

ブートストラップ法

東京大学大学院数理科学研究科

小池 祐太

2023年11月25日



東京大学大学院数理科学研究科 **フェスタ**
Graduate School of Mathematical Sciences, THE UNIVERSITY OF TOKYO 国立大学2023

1

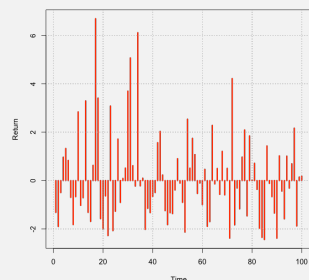
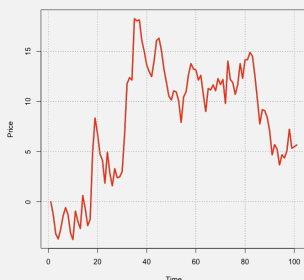
本日の話の流れ

- 導入：株式に正の収益率を期待できるか？
- 数理モデル
- ブートストラップ法
- ブートストラップ法はなぜうまくいくか？
- 正規近似 vs ブートストラップ法
- まとめ

2

導入：株式に正の収益率を期待できるか？

- ある株式に投資したときに長期的に収益を期待できるか調べたい
➡ 株価のリターン（価格変化）が「平均的には」正となるかどうかを調べればよい



3

導入：株式に正の収益率を期待できるか？

- この株式銘柄の過去 n 期分のリターンのデータ

$$X_1, X_2, \dots, X_n$$

が手元にあるとする

- 平均リターンを計測するには、データの平均値

$$\bar{X}_n = \frac{X_1 + X_2 + \dots + X_n}{n}$$

を計算するのが自然

4

導入：株式に正の収益率を期待できるか？

- この平均値 \bar{x}_n が正であれば、「この株式銘柄は長期的には正の収益を生み出す」と主張するに十分か？
- 統計学の観点では、これでは十分とは言えない
 - このデータにおいて、 \bar{x}_n の値が「たまたま」正に振れただけかもしれないため
- では、どうすればよいか？

5

導入：株式に正の収益率を期待できるか？

- \bar{x}_n の値が「十分」大きければ、「たまたま」正の値に振れた可能性はほぼ排除できるはず
- 実際にどれだけ大きければ「十分」大きいと判断できるか？
- 「十分大きい」の意味を明確にするために、数理モデルを導入する

6

数理モデル

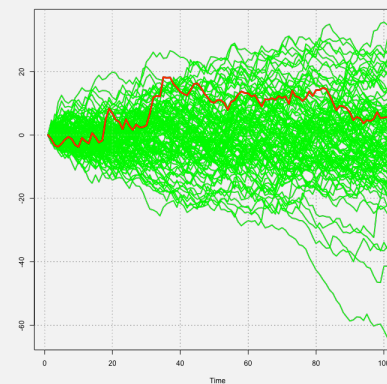
- リターンデータ X_1, X_2, \dots, X_n は、「ベースラインのリターン」に「ランダムな変動」がのった形で表現されているとモデル化する

$$X_i = \mu + \epsilon_i, \quad i = 1, 2, \dots, n$$

ベースラインの
リターン

ランダムな
変動

7



- 前頁の形の数理モデルに基づいて株価系列を100パターン生成した一例
- このように無数に起こりうるシナリオのうちの1パターンが、実際の観測データとして得られていると想定する

8

数理モデル

- ランダムな変動たちは、各期で独立同分布かつ偏りなく発生すると仮定（「偏りなく」とは、数学的には期待値が0であるという意味）
- このとき、ランダムな変動たちの平均は、どのシナリオでも長期的には（確率1で）0に近づく（**大数の法則**）：

$$\frac{\epsilon_1 + \epsilon_2 + \dots + \epsilon_n}{n} \rightarrow 0 \quad (n \rightarrow \infty)$$

9

数理モデル

- 従って、リターンの平均は長期的には（確率1で） μ に近づく：

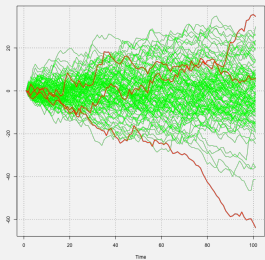
$$\bar{X}_n \rightarrow \mu \quad (n \rightarrow \infty)$$

➡「 $\mu > 0$ 」ならば、この株式の平均リターンは長期的には正となる

10

数理モデル

- 「 $\mu > 0$ である」かどうかをデータから検証する方法を考える
- ここでの問題は、 \bar{X}_n は μ に近いけれどもズレがあり、そのズレはどのシナリオを観測したかで変わってくる



観測されるシナリオによっては大きく上下にぶれたりするケースがありえる

11

ブートストラップ法

- 実際にどのシナリオが観測されたかは「神のみぞ知る」ところなので、観測されたデータに対して、誤差

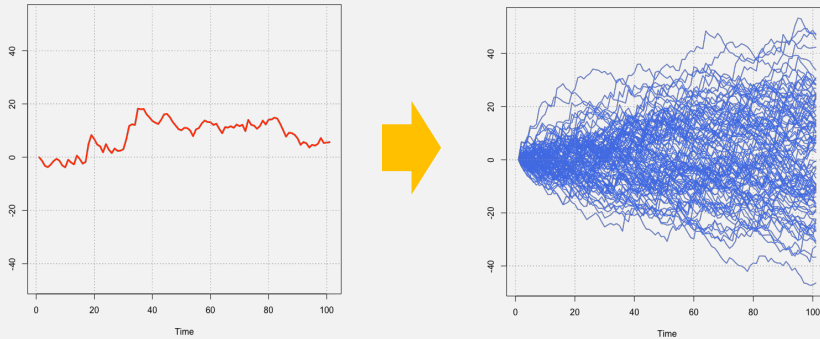
$$\bar{X}_n - \mu$$

がどの程度か知るのほとんど不可能なように思える

ブートストラップ法

観測された**ただ一つのシナリオ**を用いて、他に起こりえた無数のシナリオを擬似的にシミュレーションして生成することで、推定誤差の大きさを評価する方法

12



ブートストラップ法を用いると、左図のデータだけが観測されたときに、そのデータだけから右側のグラフをシミュレートできる

13

アイデア：観測データ

$$X_1, X_2, \dots, X_n$$

から重複を許して n 個ランダムにサンプリングしたもの

$$X_1^*, X_2^*, \dots, X_n^*$$

で「実現しなかったシナリオ」を代用する

- $X_1^*, X_2^*, \dots, X_n^*$ はブートストラップ標本と呼ばれる
- ブートストラップ標本はいくらでもシミュレートできるので、いくらでも異なるシナリオを生成できる！

14

ブートストラップ法による誤差の評価

知りたいこと

起こりうるシナリオの中で、誤差 $\bar{X}_n - \mu$ の値がどのように分布するか？（ヒストグラムはどのような形か？）

- 上で見たように、ブートストラップ標本から計算された平均

$$\bar{X}_n^* = \frac{X_1^* + X_2^* + \dots + X_n^*}{n}$$

の値は \bar{X}_n と同じように分布しているように見える

15

ブートストラップ法による誤差の評価

- μ は \bar{X}_n の「期待値」であり、これはわからない
- 一方で、 \bar{X}_n^* の（データが与えられた下での）期待値は

$$\bar{X}_n$$

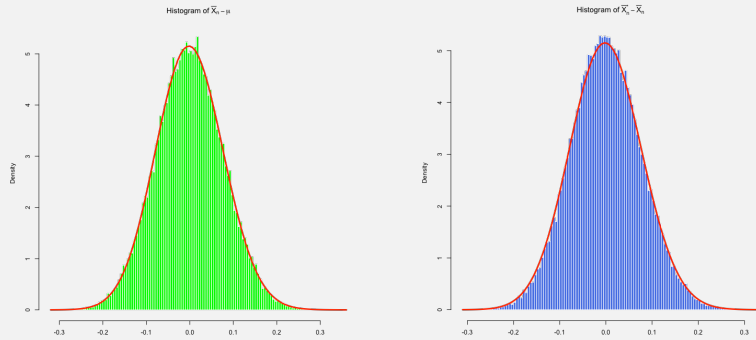
であることが計算するとわかる (X_1, X_2, \dots, X_n が入った箱から同様に確からしく n 回復元抽出したときの期待値だから)

ブートストラップ標本から計算された平均の誤差

$$\bar{X}_n^* - \bar{X}_n$$

はシミュレーションできる！

16



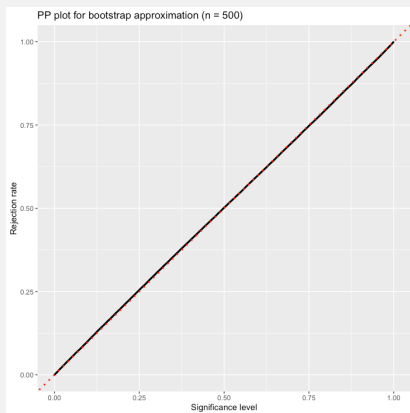
- $n = 500$ の場合の $\bar{X}_n - \mu$ のヒストグラム(左)と $\bar{X}_n^* - \bar{X}_n$ のヒストグラム(右)
- 左図は実際は観測できない多数のシナリオのシミュレーション結果に基づき作成。赤線は理論値(実際はもちろん未知)
- 右図は1つのシナリオの観測データのみを用いてブートストラップ標本を多数シミュレートした結果に基づき作成

17

実際に「 $\mu > 0$ である」ことをデータから検証するには、典型的には次のようにする(統計的仮説検定)

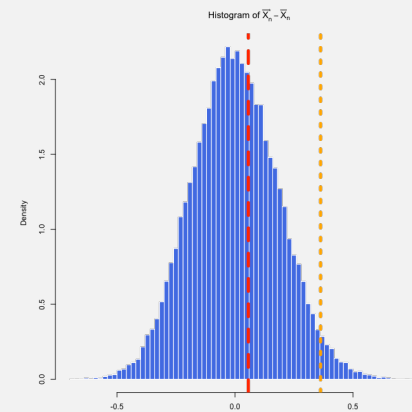
- 「非常に小さい」と思える確率を定める(ここでは5%とする)
- $\bar{X}_n^* - \bar{X}_n$ のヒストグラムから、「 $\bar{X}_n^* - \bar{X}_n > c$ というシナリオが起きる割合」が5%となるような c を求める(たくさんシミュレートした $\bar{X}_n^* - \bar{X}_n$ の値の上位5%のものを採用すればよい)
- 「 $\bar{X}_n^* - \bar{X}_n > c$ というシナリオが起きる割合」 \approx 「 $\bar{X}_n - \mu > c$ というシナリオが起きる確率」なので、仮に $\mu \leq 0$ であったならば「 $\bar{X}_n > c$ 」というシナリオが起きる確率は(およそ)5%以下
- $\bar{X}_n > c$ となっているか否かはデータから検証可能 → もしそうになっていた場合、 $\mu \leq 0$ であると信じるのは無理筋(5%の奇跡が起こったことを認めないといけな) → $\mu > 0$ であると考えるのが妥当だろう

18



- p を0から1まで動かしながら、「 $\bar{X}_n - \mu$ が $\bar{X}_n^* - \bar{X}_n$ の値の上位 $100p\%$ を超えるシナリオが起きる確率」を計算し、前者を横軸、後者を縦軸にとったグラフ(P-Pプロットと呼ばれる)
- 赤点線は45度線
- ブートストラップ法が誤差分布をよく近似しているのであれば、グラフは45度線上に来るはず

19



冒頭の例で示したデータで検証した場合
 橙線: $\bar{X}_n^* - \bar{X}_n$ の上位5%のライン
 赤線: \bar{X}_n の値
 赤線は橙線よりもずっと左側にあるので、この場合は $\mu > 0$ であるとは言い切れない

20

ブートストラップ法はなぜうまくいくか？

- ブートストラップ法は常にうまくいくわけではない

- 極端な例では、 $n = 1$ の場合は明らかにうまくいかない
- n が十分大きくてもうまくいかない場合もある

➡ 理論的な正当化が必要！

正当化のための典型的な方法

中心極限定理を用いて、元の統計量とブートストラップ統計量に対して共通の正規近似が成立することを示す

21

中心極限定理

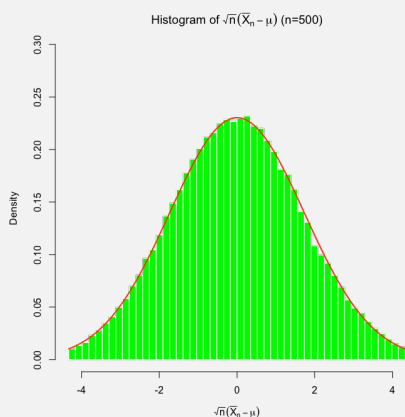
- ランダムな変動 ϵ_i の (理論上の) 標準偏差を σ とする (0 より大きい有限値であると仮定する)
- このとき、 n が十分大きければ、任意の定数 $a < b$ に対して

$$P(a \leq \sqrt{n}(\bar{X}_n - \mu) \leq b) \approx \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} dx$$

という近似が成り立つ

かつこの事象が起きる確率を表す

22



- $\sqrt{n}(\bar{X}_n - \mu)$ のヒストグラムの様子 ($n = 500$)
 - 10万パターン of シナリオのシミュレーションに基づく
 - 棒グラフの面積 = 底辺の区間に値が入る割合
 - 赤線は正規密度関数のグラフ:

$$y = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}$$

23

ブートストラップ法の正当化

- データが観測済みという条件の下では、ブートストラップ標本 $X_1^*, X_2^*, \dots, X_n^*$ は、その構成から、
 - 独立同分布な確率変数の列で、
 - 期待値はデータの平均値 \bar{X}_n 、
 - 標準偏差はデータの標準偏差

$$s_n = \sqrt{\frac{(X_1 - \bar{X}_n)^2 + (X_2 - \bar{X}_n)^2 + \dots + (X_n - \bar{X}_n)^2}{n}}$$

で与えられる

24

ブートストラップ法の正当化

- 従って、ブートストラップ統計量に対しても中心極限定理が適用できて、以下のような正規近似が得られる (厳密には少し議論が必要)

$$P^*(a \leq \sqrt{n}(\bar{X}_n^* - \bar{X}_n) \leq b) \approx \int_a^b \frac{1}{\sqrt{2\pi s_n^2}} e^{-\frac{x^2}{2s_n^2}} dx$$

- P^* は観測データが与えられた下での条件付き確率

25

ブートストラップ法の正当化

- $\sqrt{n}(\bar{X}_n - \mu)$ の分布を近似する正規分布と $\sqrt{n}(\bar{X}_n^* - \bar{X}_n)$ の(条件付き)分布を近似する正規分布の間には、標準偏差が異なる分だけ齟齬がある (前者は σ 、後者は s_n)
- しかし、データ数 n が十分大きいとき、 $s_n \approx \sigma$ という近似が成立する (大数の法則: データの標準偏差は理論上の標準偏差に近づく) → 上の齟齬は近似的には問題ない

$$\int_a^b \frac{1}{\sqrt{2\pi s_n^2}} e^{-\frac{x^2}{2s_n^2}} dx \approx \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} dx$$

26

ブートストラップ法の正当化

- 以上の議論より、

$$P(a \leq \sqrt{n}(\bar{X}_n - \mu) \leq b) \approx P^*(a \leq \sqrt{n}(\bar{X}_n^* - \bar{X}_n) \leq b)$$

という近似が成立する

→ ブートストラップ法による推定誤差の近似が正当化される!

27

直接正規近似する

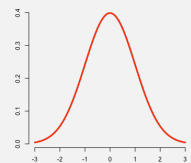
- 上の議論では、正規近似を介してブートストラップ近似を正当化した
- はじめから直接正規近似を使ってはダメなのか? → **実はOK**
- つまり、次のように近似してしまえばよい:

$$P(a \leq \sqrt{n}(\bar{X}_n - \mu) \leq b) \approx \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} dx \approx \int_a^b \frac{1}{\sqrt{2\pi s_n^2}} e^{-\frac{x^2}{2s_n^2}} dx$$

28

直接正規近似する

- つまり、 $\sqrt{n}(\bar{X}_n - \mu)$ の分布は平均0、標準偏差 s_n の正規分布で近似できる
- s_n はデータから計算可能なので、この正規分布のグラフは描画できる → $\sqrt{n}(\bar{X}_n - \mu)$ の分布のヒストグラムを近似可能！

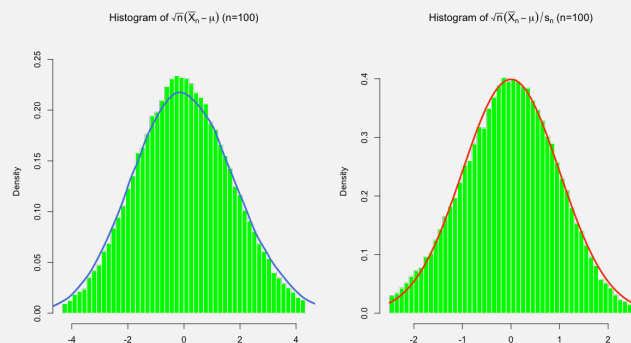


29

正規近似 vs ブートストラップ法

- 正規近似ではなくブートストラップ法を使う意味はあるのか？
 - 感覚的には、ブートストラップ法の方が精度がよさそう？
 - 一方で、上の議論は暗に「ブートストラップ \approx 正規近似」を示唆？
- ➡ 実は、後者の主張が一般には正しいことが知られている
- つまり、(いまのやり方では) 正規近似に対してブートストラップ法を優先して使う意味はない

30



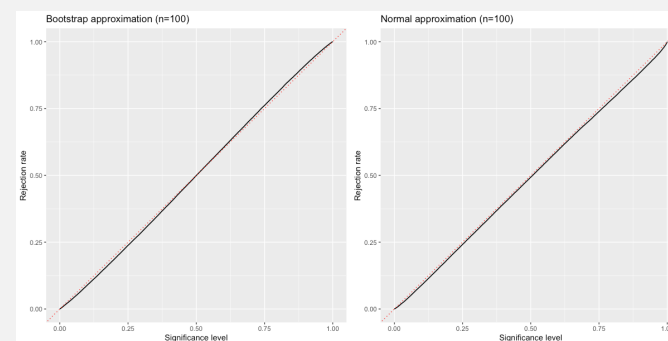
誤差分布のヒストグラムの近似 ($n = 100$)

左図青線：ブートストラップ統計量の誤差分布による近似

右図赤線：正規曲線による近似

このケースでは、むしろ正規近似の方が精度がよいように見える

31



P-Pプロット

左図：ブートストラップ統計量の誤差分布による近似

右図：正規曲線による近似

45度線に近いほど近似がうまくいっている

32

正規近似 vs ブートストラップ法

- なぜブートストラップ法は正規近似を改善しないのか？
- その理由も上の議論の中で暗に示されている
 - 元の推定誤差とブートストラップ法の推定誤差の正規近似では、標準偏差の分だけ齟齬があった（前者は σ 、後者は s_n ）
 - この誤差のせいでブートストラップ法は正規近似を改善しない
- 標準偏差に齟齬がでないようにすることでブートストラップ近似を改善できないか？

33

正規近似 vs ブートストラップ法

- $\sqrt{n}(\bar{X}_n - \mu)$ の標準偏差 σ は未知であり、これをデータから計算した標準偏差 s_n で置き換える必要があったことが問題
- $\sqrt{n}(\bar{X}_n - \mu)$ をあらかじめ s_n で割っておくと、その標準偏差は近似的に 1 となるため置き換えの操作は不要



「標準化した」誤差

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{s_n}$$

をブートストラップ法で近似してはどうか？

34

正規近似 vs ブートストラップ法

- つまり、 \bar{X}_n だけでなく、 s_n もブートストラップ版

$$s_n^* = \sqrt{\frac{(X_1^* - \bar{X}_n^*)^2 + (X_2^* - \bar{X}_n^*)^2 + \dots + (X_n^* - \bar{X}_n^*)^2}{n}}$$

で置き換える。ここに、

$$\bar{X}_n^* = \frac{1}{n} \sum_{i=1}^n X_i^*$$

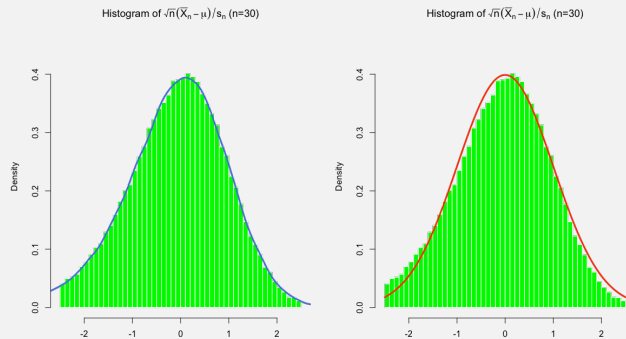
35

正規近似 vs ブートストラップ法

- 上と同様の正規近似を介した議論により次のように近似できることがわかる

$$P\left(a \leq \frac{\sqrt{n}(\bar{X}_n - \mu)}{s_n} \leq b\right) \approx P^*\left(a \leq \frac{\sqrt{n}(\bar{X}_n^* - \bar{X}_n)}{s_n^*} \leq b\right)$$

36



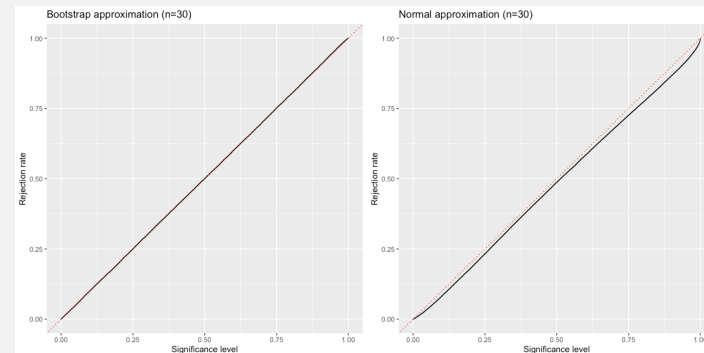
誤差分布のヒストグラムの近似 ($n = 30$)

左図青線：標準化したブートストラップ統計量の誤差分布による近似

右図赤線：正規曲線による近似

特に n が小さい場合にブートストラップ法が正規近似を改善している

37



P-Pプロット

左図：標準化したブートストラップ統計量の誤差分布による近似

右図：正規曲線による近似

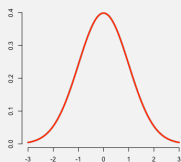
45度線に近いほど近似がうまくいっている

38

エッジワース展開

- 上のアイデアは、標準偏差の齟齬が解消されることは説明しているが、近似自体が改善する理由にはなっていない
- 近似が改善される理由は、中心極限定理の精密化である**エッジワース展開**（漸近展開の一種）によって説明できる
- 標準正規分布の確率密度関数を $\phi(x)$ と書くことにする：

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$



39

エッジワース展開

- 元の統計量のエッジワース展開：

$$P\left(a \leq \frac{\sqrt{n}(\bar{X}_n - \mu)}{s_n} \leq b\right) \approx \int_a^b \phi(x) dx - \frac{\gamma}{6\sqrt{n}} \int_a^b (2x^3 - 3x)\phi(x) dx$$

ここに、 γ は**歪度**と呼ばれるデータの分布の歪みを表す量

- ブートストラップ統計量のエッジワース展開：

$$P^*\left(a \leq \frac{\sqrt{n}(\bar{X}_n^* - \bar{X}_n)}{s_n^*} \leq b\right) \approx \int_a^b \phi(x) dx - \frac{\gamma_n}{6\sqrt{n}} \int_a^b (2x^3 - 3x)\phi(x) dx$$

γ_n はデータの観測値から計算した歪度

40

エッジワース展開

- エッジワース展開から、**正規近似**は一般には $1/\sqrt{n}$ 程度の大きさの誤差がある
- 一方で、 γ_n と γ の誤差も $1/\sqrt{n}$ 程度の大きさであることが知られているので、エッジワース展開より、**ブートストラップ近似**の誤差の大きさは $1/n$ 程度だとわかる

➡ **ブートストラップ近似**が**正規近似**を改善する！

41

まとめ

- **ブートストラップ法**：観測データから重複を許してリサンプリングすることで、「実現しなかった」シナリオを擬似的にシミュレートする方法
- ブートストラップ法の正当性は、典型的には**正規近似（中心極限定理）**を介して示される
- 一般に、ナイーブなブートストラップ法は正規近似を改善しないが、多くの場合、標準化した統計量をブートストラップすることで正規近似を改善できる（その正当化には**漸近展開**を使うのが標準的¹⁾）

¹⁾Hall (1992) *The bootstrap and Edgeworth expansion*, Springer 参照

42

ブートストラップ法の他のメリット

- 本日は触れなかったが、ブートストラップ法が有用な別の場面として、誤差分布が容易に近似できない統計量に対しても適用可能な点がある
 - 正規分布で近似できることはわかっているが、その標準偏差が容易に推定できない場合（例：メディアン¹⁾）
 - そもそも正規分布で近似できない場合（例：説明変数に外れ値が含まれる線形回帰モデル²⁾）
 - もちろん常に適用可能とは限らないので、理論的な正当化が必要。そのために様々な極限定理が利用される

¹⁾Bickel & Freedman (1981, *Ann. Statist.* **9**, 1196-1217), Proposition 5.1 参照

²⁾Mammen (1992) *When does bootstrap work?*, Springer, Chapter 6 参照

43