高知工科大学 経済・マネジメント学群

統計学 2

8. 標本平均と母平均

矢内 勇生（やない ゆうき）

https://yukiyanai.github.io

yanai.yuki@kochi-tech.ac.jp

このトピックの目標

- 大数(たいすう)の法則を理解する
- 標本分布と標準誤差を理解する
- 標本平均を利用して母平均を推定する方法を身につける
  - 点推定
  - 区間推定

©2020 Yuki Yanai

大数の法則

大数の法則
(Law of Large Numbers; LLN)

- ある試行（例：コイン投げ）においてある事象が起こる（例：表が出る）確率が θ（例：0.5）だとする（さらに、各試行は他の試行に影響を及ぼさない[各試行は独立である]とする）

★ **大数の法則**：試行回数を増やすにつれて、ある事象が起こる割合（比率）は θ に近づく

★　厳密には、「大数の弱法則（Weak Law of Large Numbers）」と「大数の強法則（Strong Law of Large Numbers）」を区別する必要がある

©2020 Yuki Yanai

LLNの例1：コイン投げ（1）

- 正しいコインを投げたとき、表が出る確率 $\theta = 0.5$
- 実際にコインを N 回投げたとき、表が出る割合 p は $p = 0.5$ になる？

©2020 Yuki Yanai

LLNの例1：コイン投げ（2）

表が出る割合 p は0.5になるとは限らない！

- 実際にコインを投げてみると

  - 1回目に表が出た → $p = 1/1 = 1$

  - 2回目も表が出た → $p = 2/2 = 1$

  - 3回目も表が出た → $p = 3/3 = 1$

  - 4回目は裏が出た → $p = 3/4 = 0.75$

©2020 Yuki Yanai

LLNの例1：コイン投げ（3）

- コインを投げる回数が少ないとき、表が出る割合 p は0.5から離れた値になりやすい
  - LLN：投げる回数 Nを大きくすれば、p が0.5に近づく！（シミュレーションで示す）

★ コイン投げの回数 = 標本サイズ

©2020 Yuki Yanai

実際にやってみよう

- 実習課題
  - https://yukiyanai.github.io/jp/classes/stat2/contents/R/LLN.html

©2020 Yuki Yanai

標本分布

標本分布（sampling distribution）

- 母集団からサイズ N の標本を単純無作為抽出するとき、選ばれる個体の組み合わせは何通りもある
- 抽出可能な組み合わせをすべて考え、それぞれの標本で統計量（例：女性の割合）を求めると、その統計量は分布する

- こうして求められる分布（統計量の標本ごとの分布）を**標本分布**と呼ぶ

©2020 Yuki Yanai

標本分布の例（1）

個人	A	B	C	D	E
身長　(cm)	158	159	160	161	162

5人の母集団から標本サイズ2の標本を1つ抽出する：

全部で ${}_5C_2 = 10$ 通りの選び方

標本	{A, B}	{A, C}	{A, D}	{A, E}	{B, C}	{B, D}	{B, E}	{C, D}	{C, E}	{D, E}
標本平均	158.5	159	159.5	160	159.5	160	160.5	160.5	161	161.5

©2020 Yuki Yanai

標本分布の例（2-1）

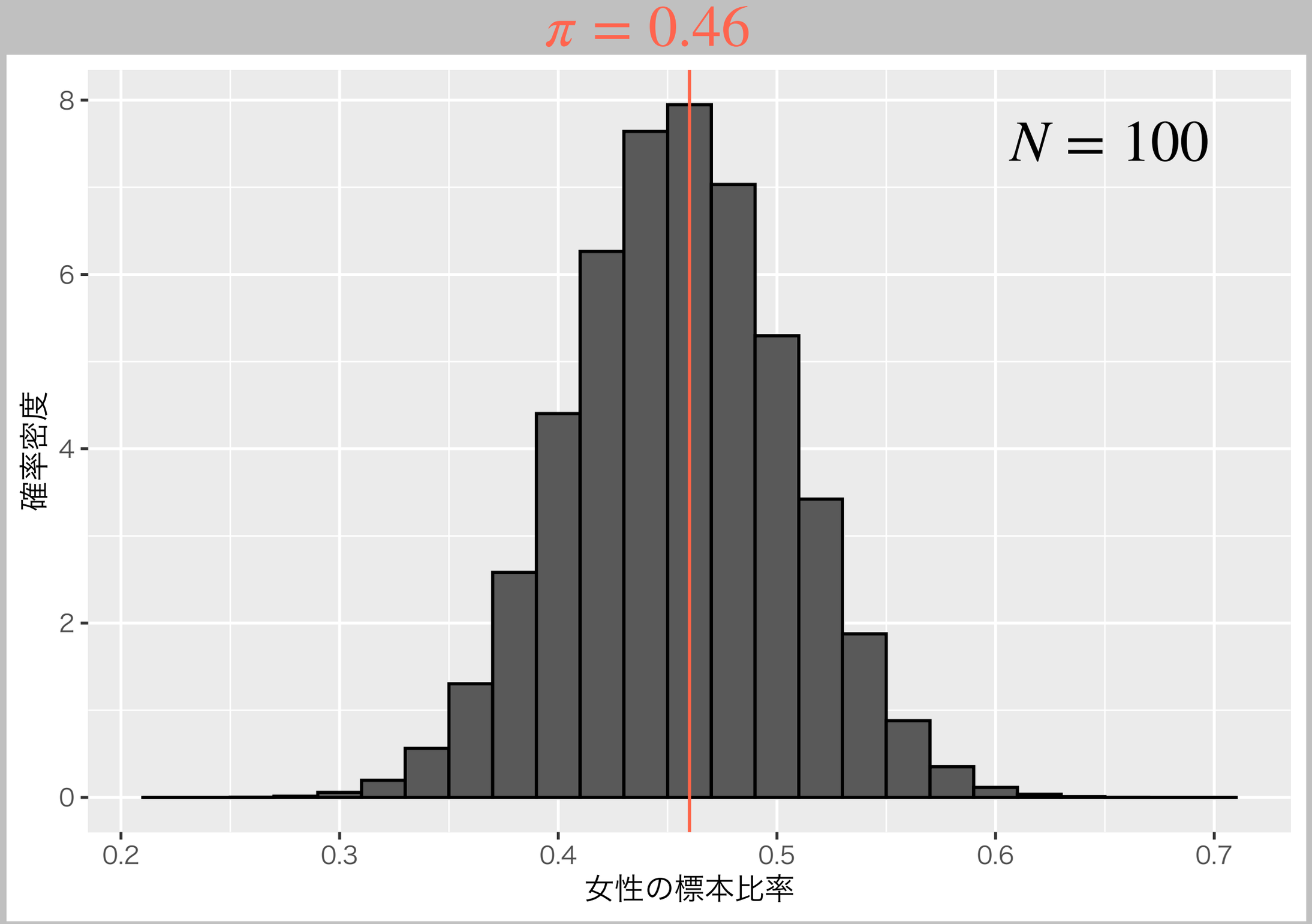
- 男5400人、女4600人の計1万人から100人選ぶ
- 1万人から100人を選ぶ方法は全部で約6.5×10^{241}通り
  → 全部の組み合わせ：標本分布

➡ 数が多いので全部の組み合わせを考えるのは困難

➡ コンピュータ・シミュレーションで $N = 100$ のサンプルを100万個抽出し、擬似的な標本分布を得る

©2020 Yuki Yanai

標本分布の例（2-2）



©2020 Yuki Yanai

標本分布と標本サイズ

- 標本サイズ N を変えれば、得られる標本分布も変わる
  - 10人の母集団から2人を選ぶ：45通り
  - 10人の母集団から3人を選ぶ：120通り
  - 10人の母集団から9人を選ぶ：10通り

➡ 可能な組み合わせが異なれば、統計量の分布も変化する

©2020 Yuki Yanai

標準誤差 (standard error; SE)

- **標準誤差** = 標本分布に現れるばらつき

  = **統計量の標準偏差**

- 母集団が十分大きい（目安：母集団が標本サイズ N の100倍以上）と

き、標準誤差 SE は

$$\mathrm{SE} = \frac{\textbf{母標準偏差}}{\sqrt{\textbf{標本サイズ}}} = \frac{\sigma}{\sqrt{N}}$$

または

$$\mathrm{SE} = \frac{\textbf{母標準偏差の推定値}}{\sqrt{\textbf{標本サイズ}}} = \frac{s}{\sqrt{N}}$$

©2020 Yuki Yanai

標準誤差の特徴

- 標準誤差の大きさは、標本サイズ N の平方根に反比例する
  - N を4倍にすれば、SE は半分になる！

➡ N が大きいほど、統計的推測が正確になる

©2020 Yuki Yanai

標準誤差の例（1）

- 女性比率が0.46 である100万人の母集団から N人の標本を抽出し、女性比率を求めるときの標準誤差

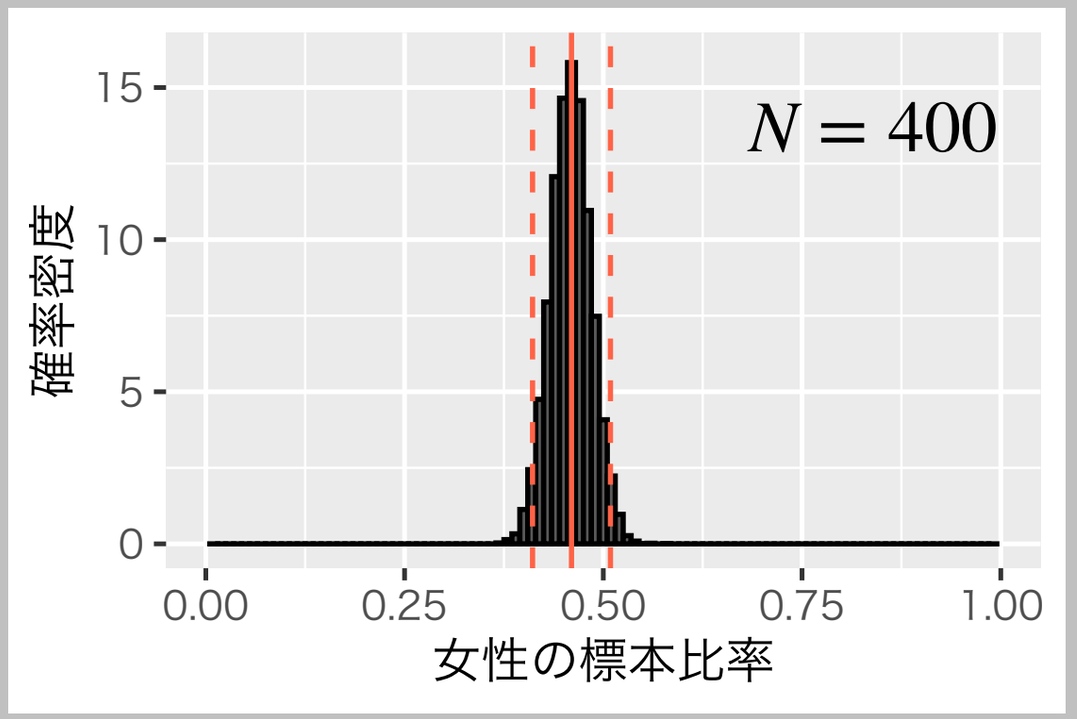
  - $N = 25$ のとき：$\frac{\sigma}{\sqrt{N}} = \frac{\sqrt{0.46(1-0.46)}}{\sqrt{25}} \approx \frac{0.5}{5} = 0.1$

  - $N = 100$ のとき：$\frac{\sigma}{\sqrt{N}} = \frac{\sqrt{0.46(1-0.46)}}{\sqrt{100}} \approx \frac{0.5}{10} = 0.05$

  - $N = 400$ のとき：$\frac{\sigma}{\sqrt{N}} = \frac{\sqrt{0.46(1-0.46)}}{\sqrt{400}} \approx \frac{0.5}{20} = 0.025$

©2020 Yuki Yanai

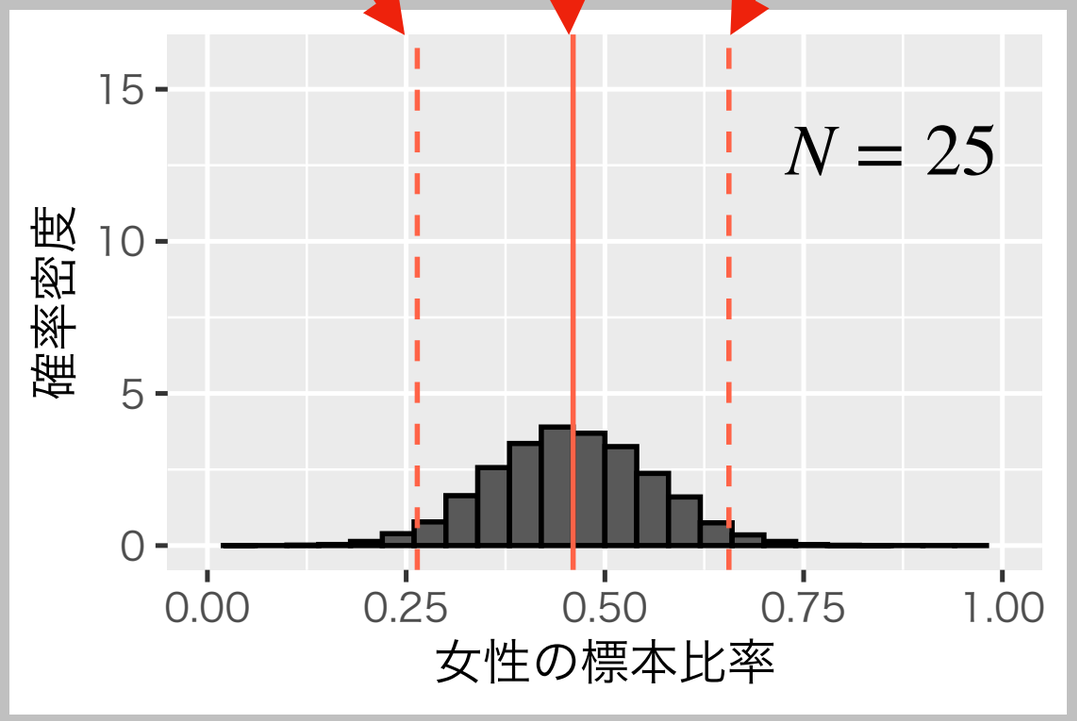
標準誤差の例（2）

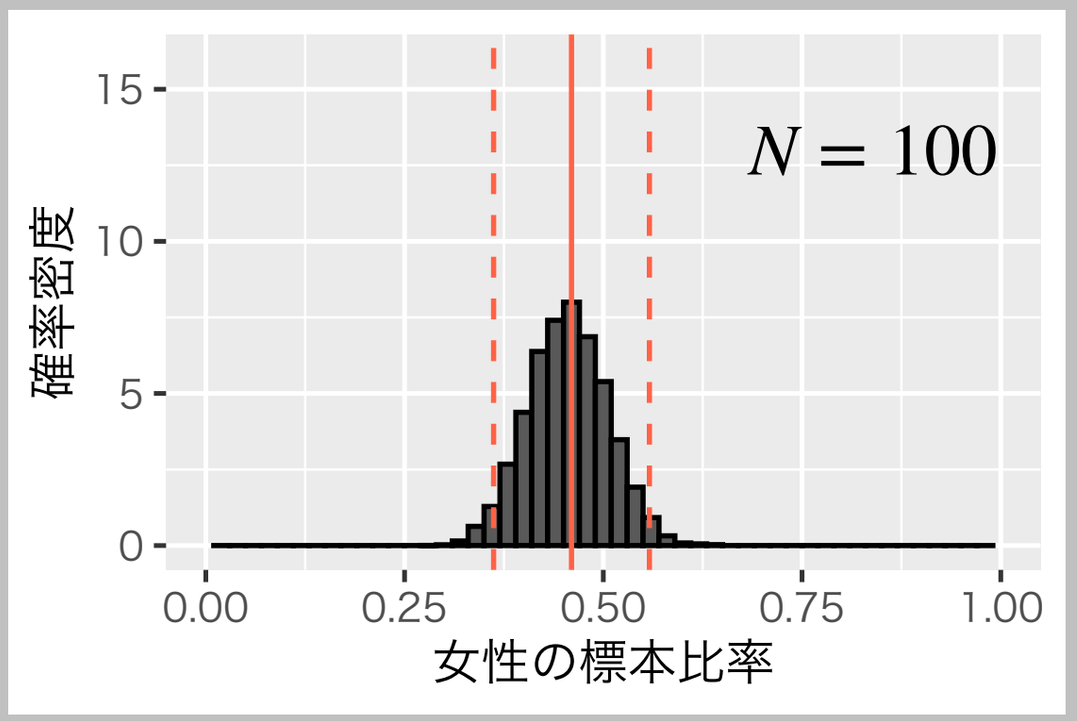


$\pi = 0.46$

$\pi - 1.96 \cdot \mathrm{SE}$

$\pi + 1.96 \cdot \mathrm{SE}$





©2020 Yuki Yanai

母集団の大きさと標本サイズ（1）

- 都道府県ごとに増税に賛成か反対か調べたい
  - 東京都の人口：約1400万人
  - 滋賀県の人口：約140万人

▸ 東京の標本サイズは滋賀の標本サイズの10倍にすべきか？

©2020 Yuki Yanai

母集団の大きさと標本サイズ（2）

- 東京でも滋賀でもちょうど半分の人が増税に賛成（反対）だとする（仮定）
- 標本サイズを100にすると
  - 東京の標準誤差 = 0.5/10 = 0.05
  - 滋賀の標準誤差 = 0.5/10 = 0.05
- 母集団の人口が10倍でも、同じ標本サイズで同じ精度の調査ができる

★（母集団が十分大きいとき）**標準誤差は母集団の大きさに依存しない**！

©2020 Yuki Yanai

母集団の大きさと標本サイズ（3）

- Good news：東京の調査を滋賀と同じ精度で行うためには、滋賀と同じサイズの標本を抽出すればよい
- Bad news：滋賀の調査を東京の調査と同じ精度にするためには、東京と同じサイズの標本を抽出しなければならない

©2020 Yuki Yanai

標本平均と母平均

標本平均と母平均

- 標本平均：標本から求められる平均値（統計量）
- 母平均：母集団の平均値（母数）
- 例：日本の成人男性（母集団）の身長を知るために、単純無作為抽出で1000人抽出した

- 標本平均：1000人の身長の平均値

↓ 推測する！

- 母平均：日本の成人男性身長の平均値（未知、興味の対象）

©2020 Yuki Yanai

標本平均と母平均の差

- 標本平均の誤差 = 標本平均 − 母平均
- 「標本平均 = 母平均」になるとは限らない

➡ 標本平均には誤差がある

➡ 標本平均は分布する！

©2020 Yuki Yanai

標準誤差 ≤ 母標準偏差

- 標準誤差 ≤ 母標準偏差

➡標本平均のばらつき ≤ 母集団のばらつき

- なぜ？

  - シミュレーションで示す

©2020 Yuki Yanai

標本分布のシミュレーション

★母集団の身長が平均170、 標準偏差5.5の正規分布に従うことを知っているとする

★このとき、標本平均の分布をシミュレーションで手に入れる（標本の数 [標本サイズではない！] =10万）

- 目的：標本平均のばらつきが母集団のばらつき以下になる理由を理解する

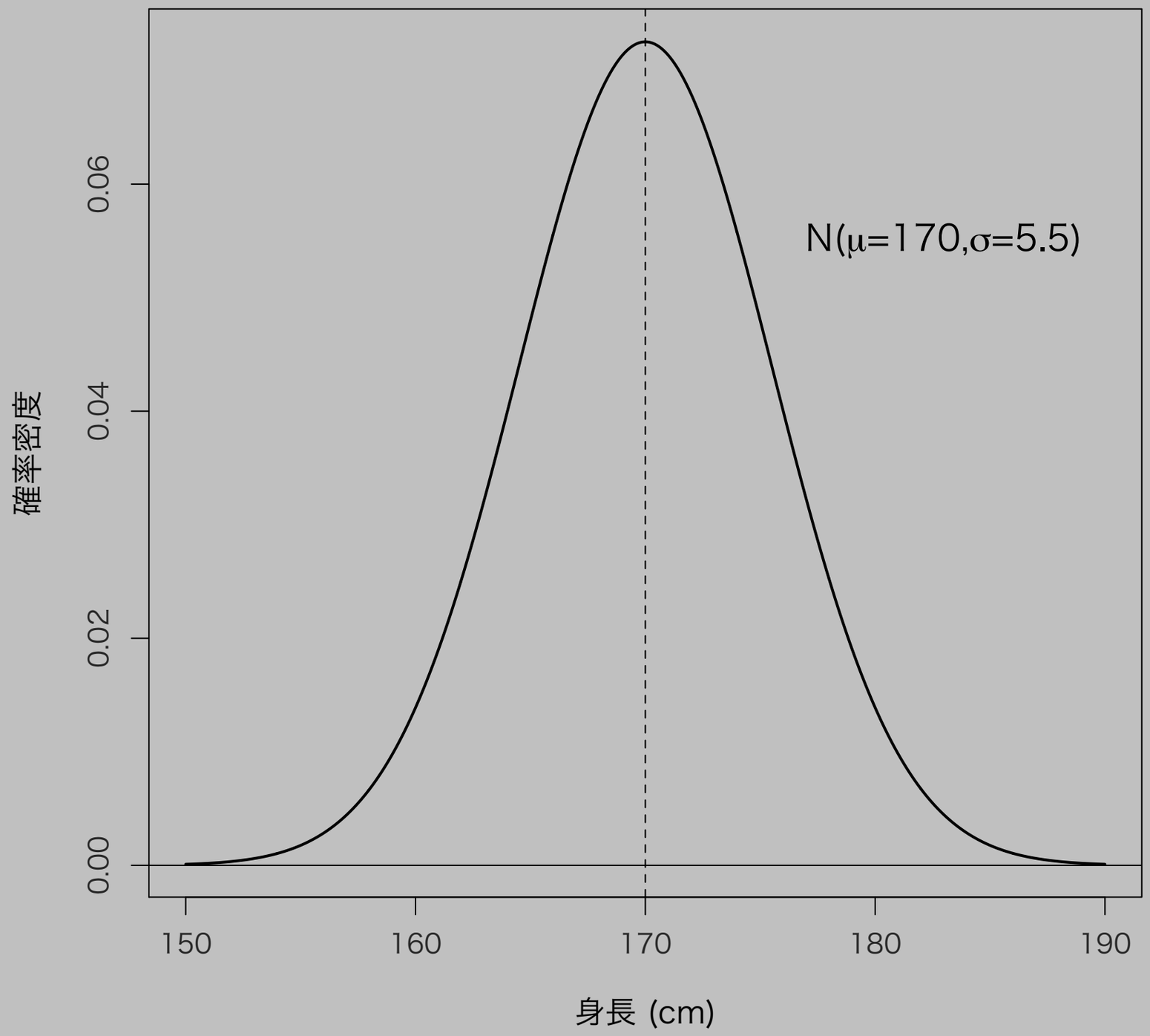
©2020 Yuki Yanai

標本サイズ（N）が1のとき

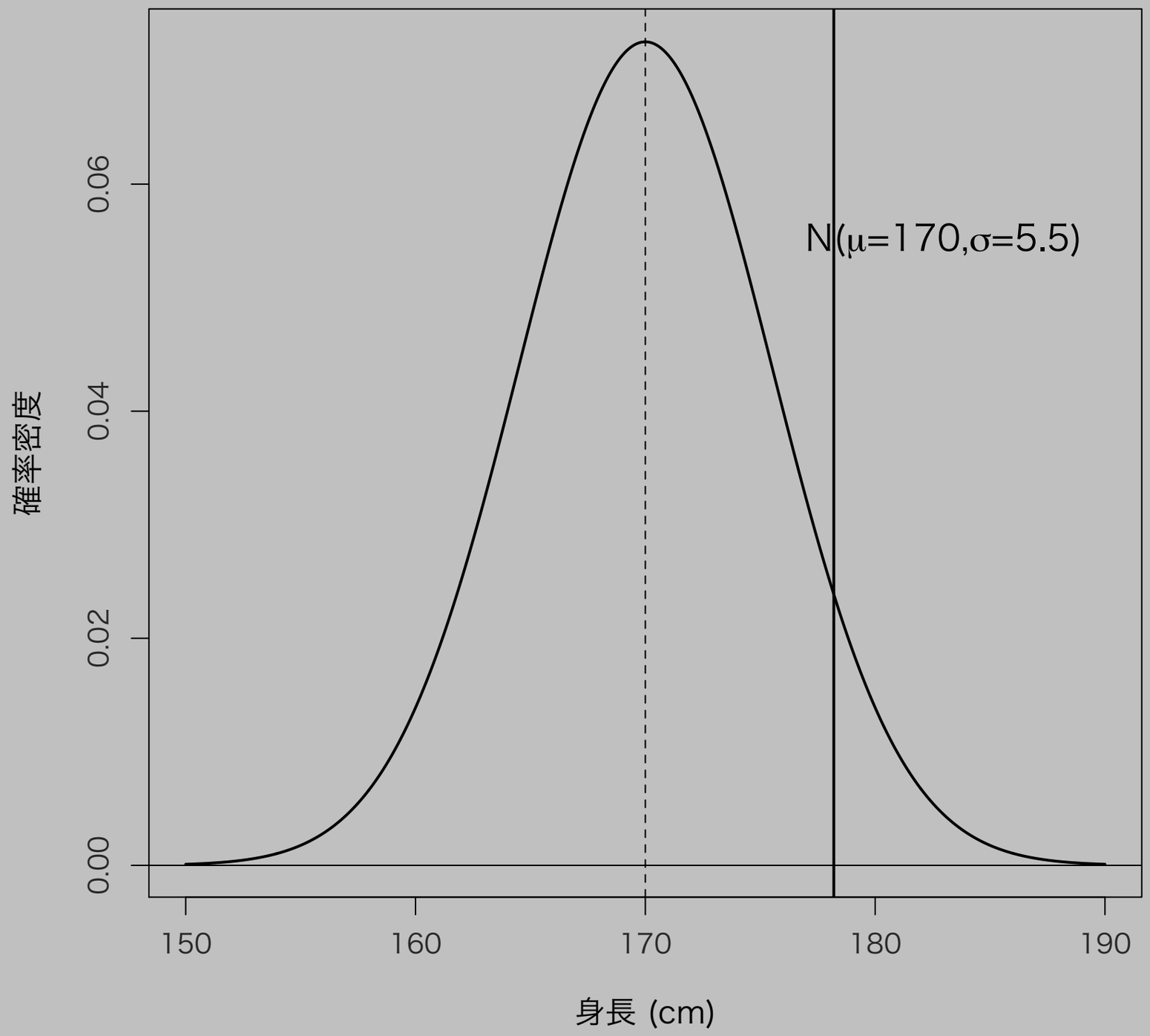
- 標本平均は標本として抽出された値そのもの
- 標準誤差　$\mathrm{SE} = \dfrac{\sigma}{\sqrt{N}} = \dfrac{\sigma}{\sqrt{1}} = \sigma$

- 「標本平均のばらつき = 母集団のばらつき」になるはず

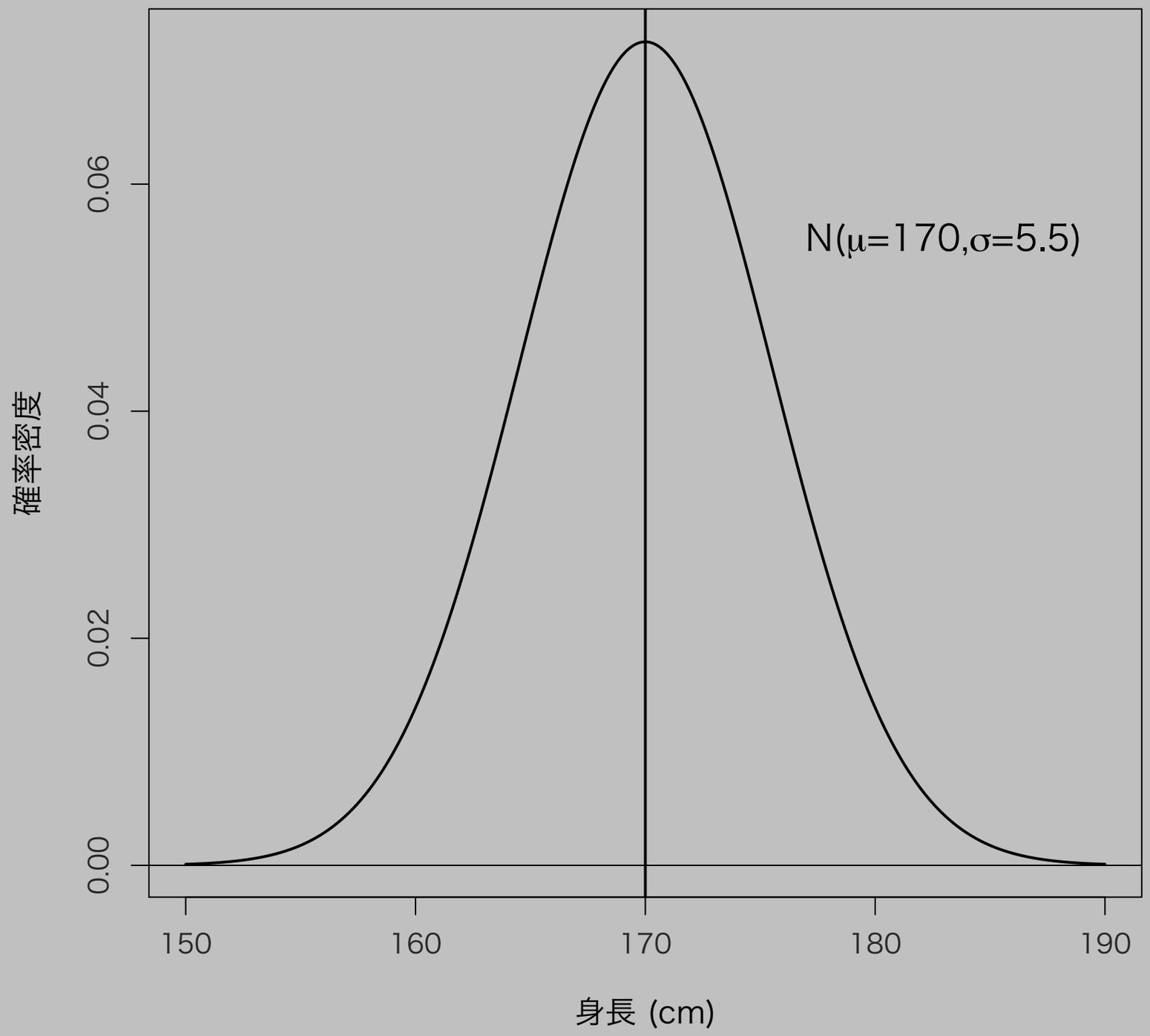
©2020 Yuki Yanai



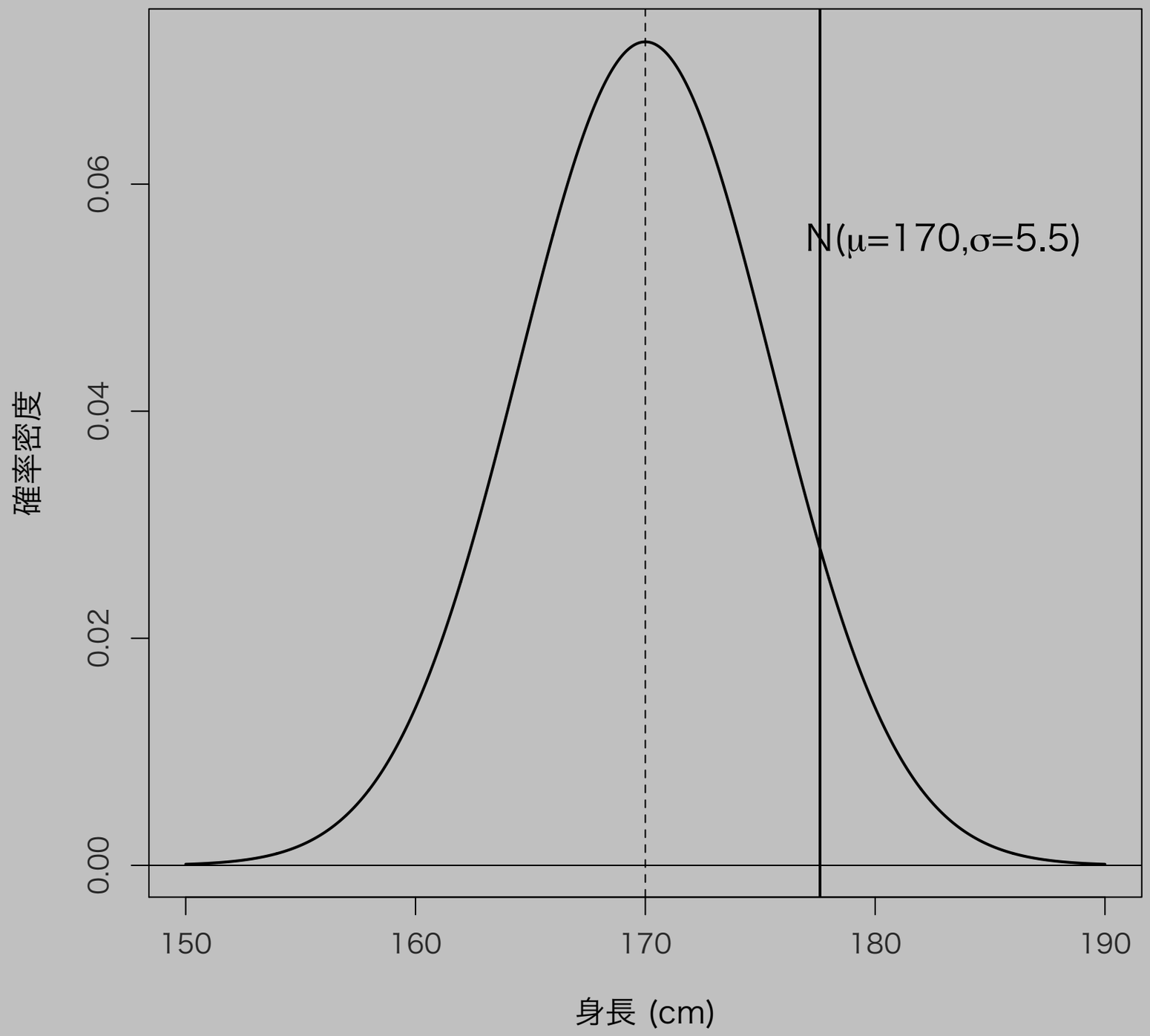
©2020 Yuki Yanai



©2020 Yuki Yanai



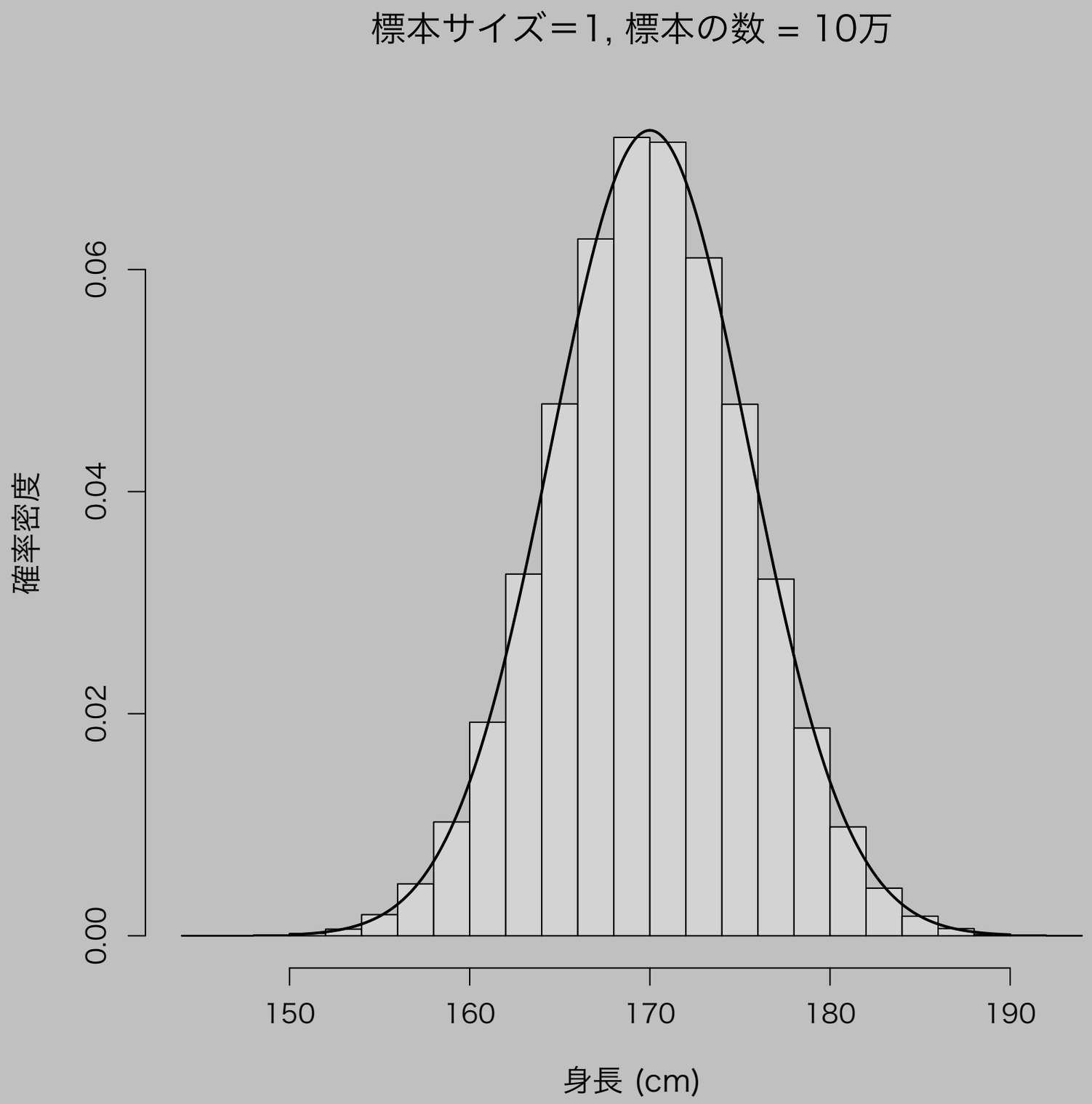
©2020 Yuki Yanai



©2020 Yuki Yanai

https://yukiyanai.github.io/jp/classes/stat2/contents/R/sampling_distribution_n1.mp4

©2020 Yuki Yanai



©2020 Yuki Yanai

標本サイズ（N）が2のとき

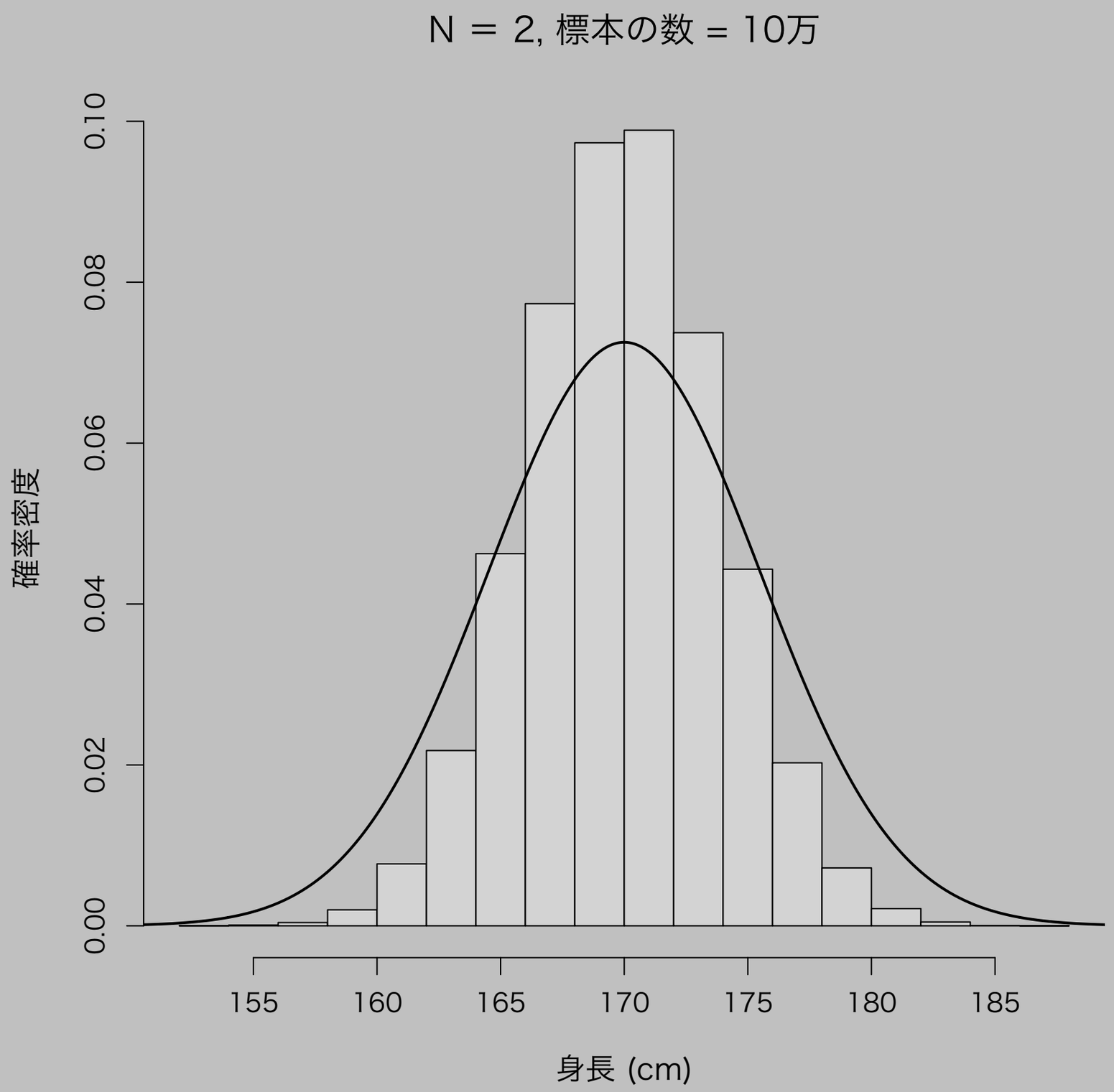
- 標本平均は標本として抽出された2つの値の平均値
- 標準誤差　$\mathrm{SE} = \dfrac{\sigma}{\sqrt{N}} = \dfrac{\sigma}{\sqrt{2}} \approx 0.7\sigma$

▸ 「標本平均のばらつき = 母集団のばらつきの約0.7倍」になるはず

©2020 Yuki Yanai

https://yukiyanai.github.io/jp/classes/stat2/contents/R/sampling_distribution_n2.mp4

©2020 Yuki Yanai



©2020 Yuki Yanai

標本サイズ（N）が10のとき

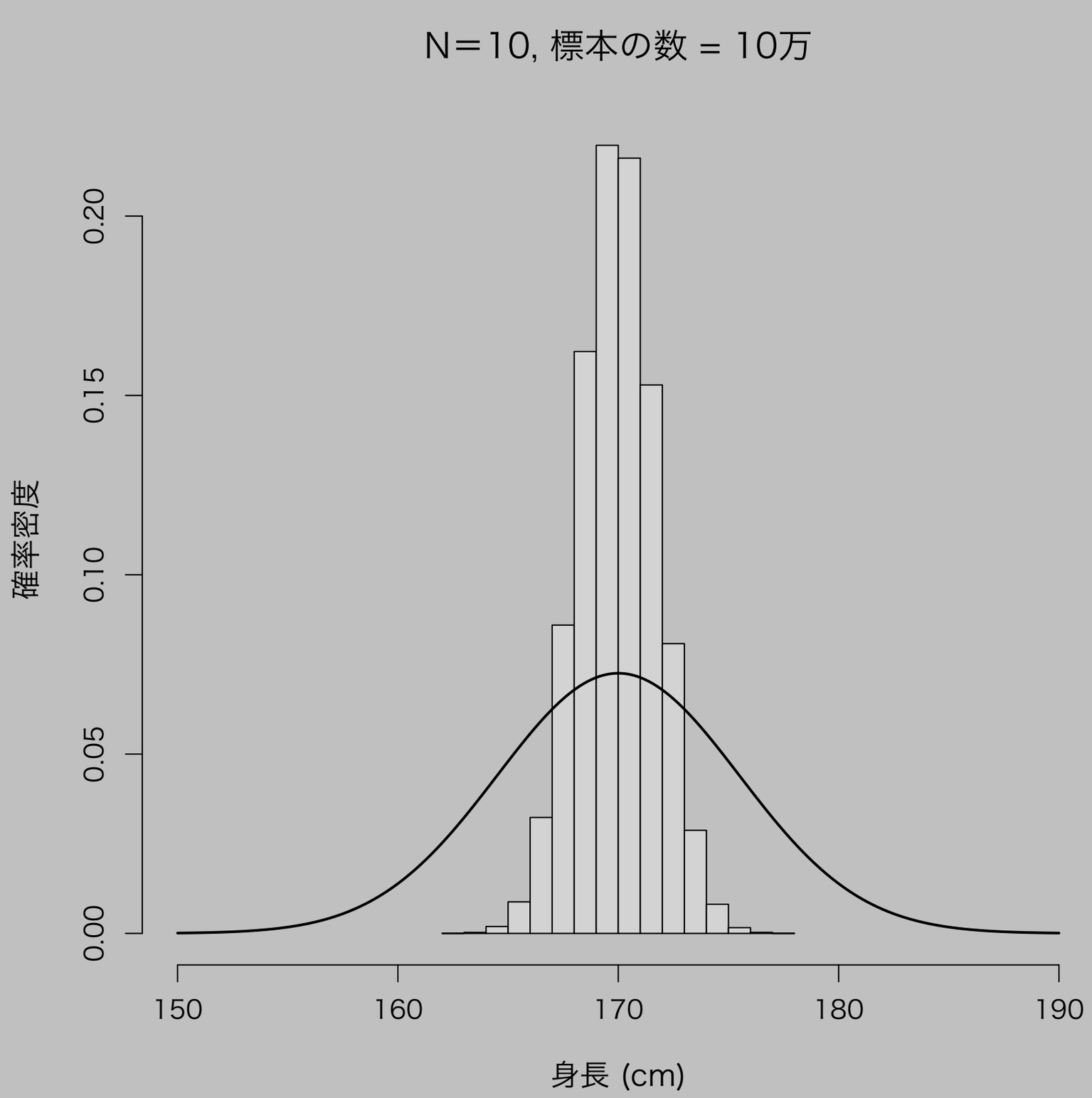
- 標本平均は標本として抽出された10個の値の平均値
- 標準誤差 $\mathrm{SE} = \dfrac{\sigma}{\sqrt{N}} = \dfrac{\sigma}{\sqrt{10}} \approx 0.3\sigma$

- 「標本平均のばらつき = 母集団のばらつきの約0.3倍」になるはず

©2020 Yuki Yanai

https://yukiyanai.github.io/jp/classes/stat2/contents/R/sampling_distribution_n10.mp4

©2020 Yuki Yanai



©2020 Yuki Yanai

標準誤差の特徴

- 標準誤差の大きさは、標本サイズ N の平方根に反比例する
  - N を4倍にすれば、SE は半分になる！
  - N が大きいほど、統計的推測が正確になる（大数の法則による）

©2020 Yuki Yanai

ここまでのまとめ

- 大数の法則：標本サイズが大きくなるほど、誤差が小さくなる
- 標本には誤差がつきもの
  - 標本分布と標準誤差
  - 「平均すれば」うまく推定できる
  - しかし、1つひとつの標本平均は信用できない
  - では、どうする？（次のテーマ）
- 実習：
  - https://yukiyanai.github.io/jp/classes/stat2/contents/R/sampling-distribution.html

©2020 Yuki Yanai

母平均の推定

標本平均から母平均を推測する：点推定

- 点推定：母数（パラメタ）を1つの値で推定する
- 母平均の点推定値 (point estimate) ＝ 標本平均
  - **標本平均の平均（標本分布の中心）は母平均に一致する**
  - **標本平均は母平均の不偏推定量 (unbiased estimator)**

©2020 Yuki Yanai

不偏性 (unbiasedness)

. 標本平均 $\bar{x}$ には誤差がある：$\bar{x} \neq \mu_x$

. しかし、標本平均の平均は母平均 μ_xに一致する

$$\mathbb{E}[\bar{x}] = \mu_x$$

- この性質を**不偏性**と呼ぶ：推定値に望まれる性質の1つ

©2020 Yuki Yanai

標本平均から母平均を推測する：区間推定

- 区間推定：母数を区間で推定する
  - 推定に幅をもたせる
- 区間推定に用いる区間：信頼区間（confidence interval; CI）

©2020 Yuki Yanai

標準正規分布の特徴を利用して推測する

- 標準正規分布の特徴：[-1.96, 1.96] の区間にデータの95%が収まる
- 正規分布に従う変数を標準化することで、標準正規分布を使える

★ 標本サイズ（N）が大きくなれば、誤差の分布は正規分布に近づく（中心極限定理）

©2020 Yuki Yanai

標本平均を標準化する

- 標本平均の平均 = 母平均 μ
- 標本平均の標準偏差 = 標準誤差 (SE)

▸ 標本平均の z 値は、

$$z = \frac{\bar{x} - \mu_x}{\mathrm{SE}} = \frac{\bar{x} - \mu_x}{\frac{\sigma}{\sqrt{N}}}$$

©2020 Yuki Yanai

z 値の95%が [-1.96, 1.96] にある

- 標本平均の z 値のうち、95%は 区間 [-1.96, 1.96] に収まるはず
- つまり、たくさんある標本の95%について、次の式が成り立つ：

$$-1.96 \leq z \leq 1.96$$

$$-1.96 \leq \frac{\bar{x} - \mu_x}{\frac{\sigma}{\sqrt{N}}} \leq 1.96$$

©2020 Yuki Yanai

95%信頼区間を求める（1）

$$-1.96 \leq \frac{\bar{x} - \mu_x}{\frac{\sigma}{\sqrt{N}}} \leq 1.96$$

- 既知のもの：$N, \bar{x}$（σ も知っているとする）
- 推定の対象：μ_x
- 上の不等式を μ_xについて解けば、μ_x（x の母平均）の95%信頼区間が得られる

©2020 Yuki Yanai

95%信頼区間を求める（2）

$$-1.96 \leq \frac{\bar{x} - \mu_x}{\frac{\sigma}{\sqrt{N}}} \leq 1.96$$

$$\Rightarrow \bar{x} - 1.96\frac{\sigma}{\sqrt{N}} \leq \mu_x \leq \bar{x} + 1.96\frac{\sigma}{\sqrt{N}}$$

★ μ_xの95%信頼区間は

$$\left[\bar{x} - 1.96\frac{\sigma}{\sqrt{N}},\ \ \bar{x} + 1.96\frac{\sigma}{\sqrt{N}}\right] = \left[\bar{x} - 1.96 \cdot \mathrm{SE},\ \ \bar{x} + 1.96 \cdot \mathrm{SE}\right]$$

©2020 Yuki Yanai

95%以外の信頼区間を求める

- 求めるパーセントに応じて、1.96 の代わりに適切な数字を選ぶ
  - 標準正規分布表を使う
  - R では qnorm() で求める

(例)

- 50%信頼区間：標本平均 ± 0.67SE
- 99%信頼区間：標本平均 ± 2.58SE
- 99.9%信頼区間：標本平均 ± 3.29SE

©2020 Yuki Yanai

qnorm() の使い方

- 標準正規分布で分布の中央部分95%の区間がどこにあるか求める方法
    - 正規分布は左右対称：中央95%を得るには、両側2.5% (0.025) ずつ除外すればよい
    - つまり、標準正規分布で、左端から2.5%の点と97.5%の点を見つければよい
- Rでは

```
qnorm(c(0.025, 0.975))
```

とすると、-1.96 と 1.96 という結果が得られる

©2020 Yuki Yanai

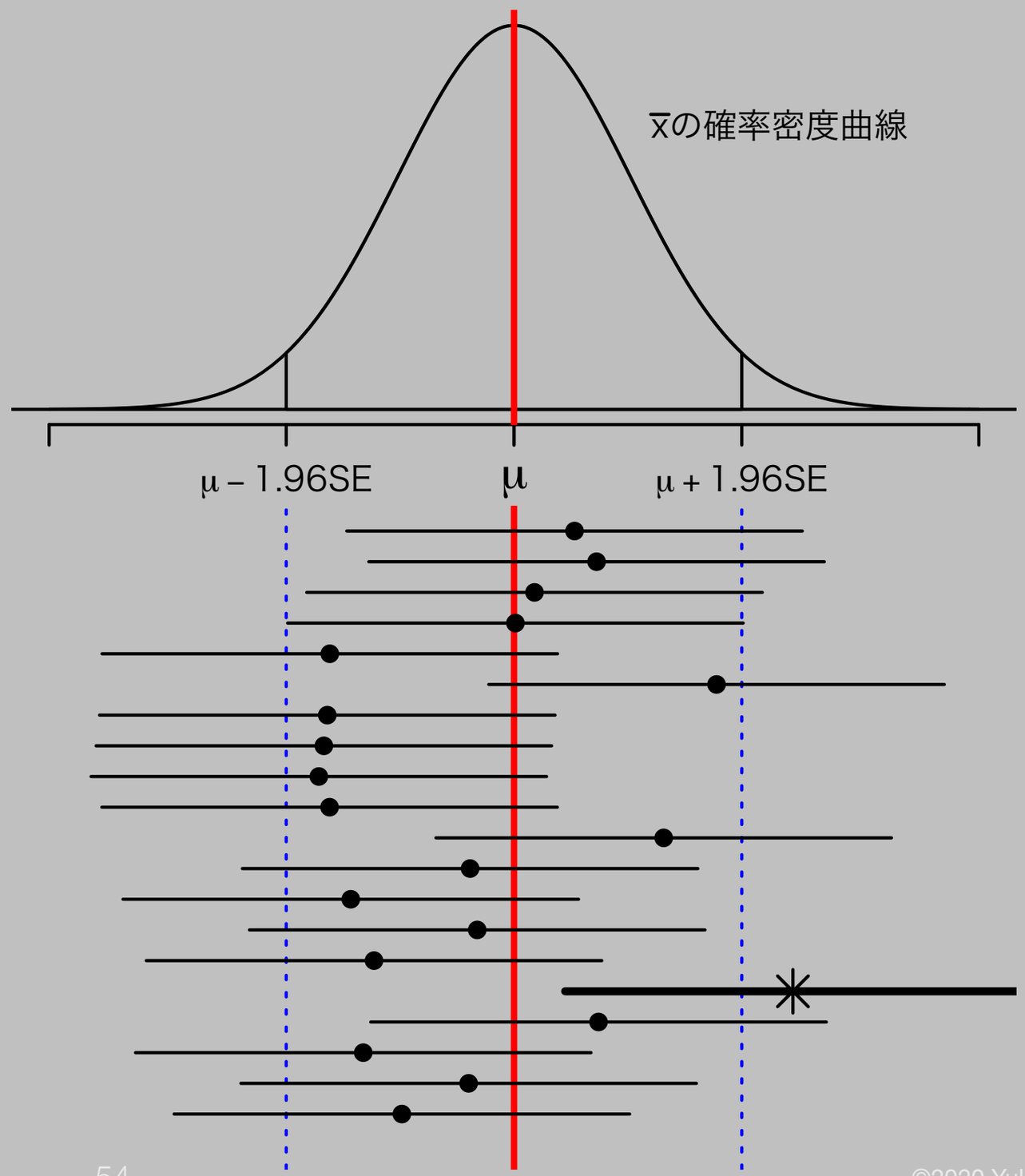
信頼区間についての注意

- 95%信頼区間 ≠ 母平均がその区間にある確率が95%
- 95%信頼区間 = 標本をたくさん抽出し、同じ手続でそれぞれの標本から信頼区間を求めたとき、母平均を含んでいる区間を得る確率が95%
- よって、１つの標本から得た95%信頼区間が母平均（真の値）を含んでいる確率は1か0（つまり、含んでいるか含んでいないかのどちらか）

©2020 Yuki Yanai

様々な標本の95%信頼区間

- 標本によって、信頼区間は変わる
- 通常、手元には1つの標本しかない
- 手元にある標本の信頼区間は、「母数を含む」か「母数を含まない」のどちらか一方



©2020 Yuki Yanai

練習問題

- ある人の血圧の計測値を母集団とすると、現在の実際の血圧 μ を母平均として、母標準偏差が10の正規分布をしている。

(1) 1回だけ血圧を測ったら、計測値は130だった。95%信頼区間は？

(2) 4回血圧を測ったら、計測値は{131, 135, 140, 138}だった。95%信頼区間は？

©2020 Yuki Yanai

母平均の推定方法のまとめ

- 標本平均から母平均を推測する
  - 母平均の点推定値 = 標本平均 (不偏推定量)
  - 母平均の95%信頼区間 = [標本平均 ± 1.96SE]
- 1つの標本から計算した95%信頼区間に母数（パラメタ、真の値）が含まれる確率は95%**ではない**
- 実習：
  - https://yukiyanai.github.io/jp/classes/stat2/contents/R/estimating-pop-mean.html

©2020 Yuki Yanai

次回予告

9. t 分布と母平均の推定