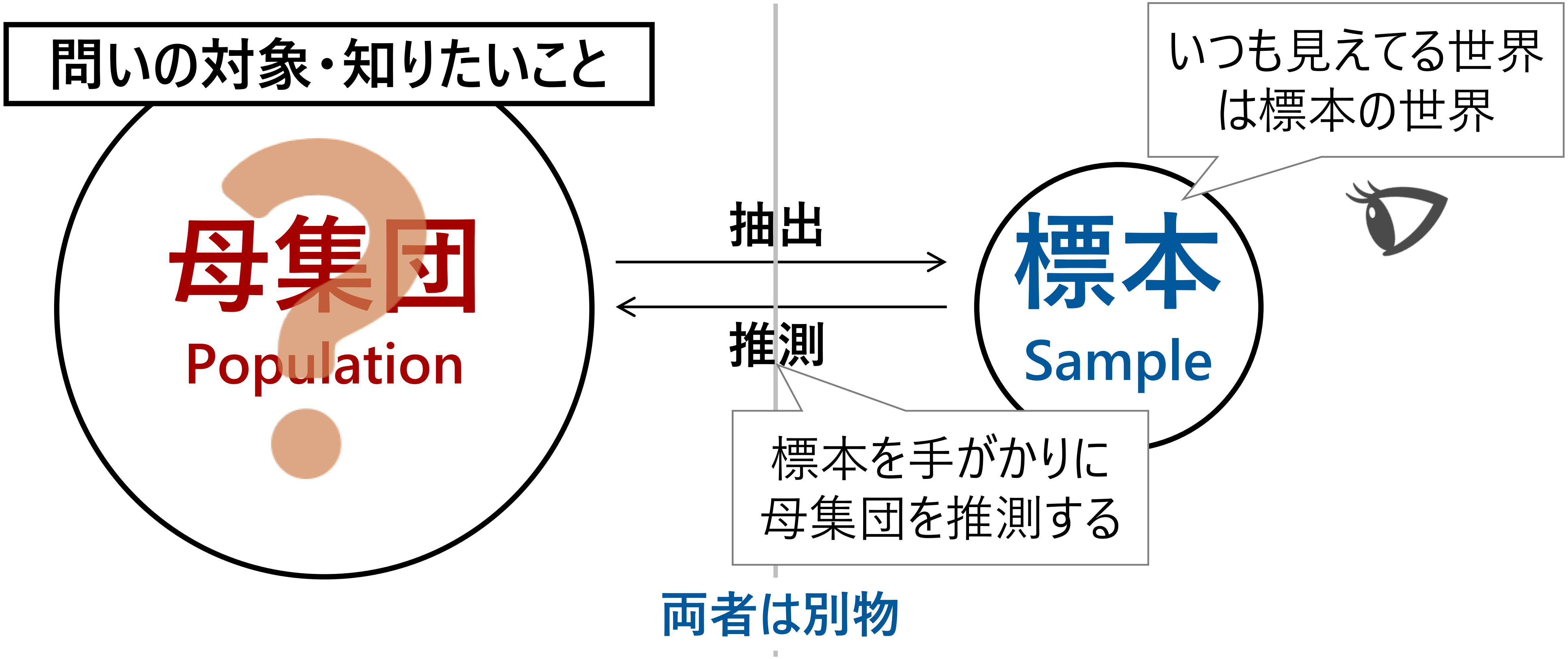
いちばん理解しやすい統計学ベーシック講座

講義スライド

セクション１：記述統計

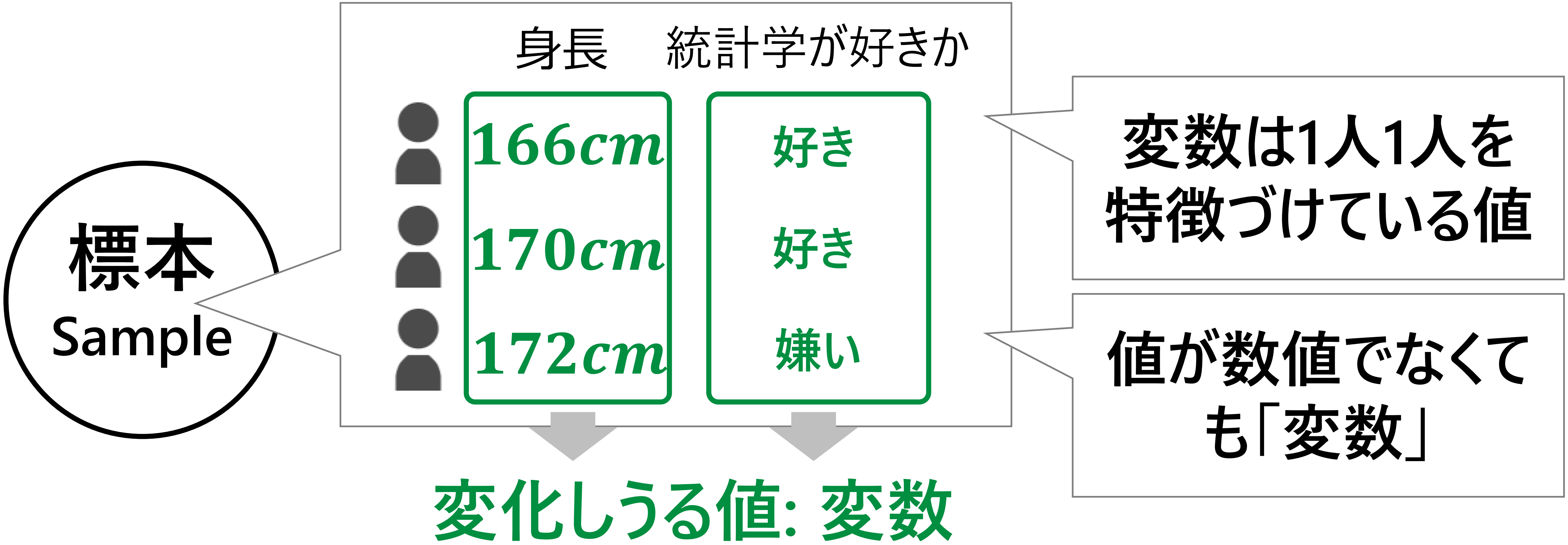
母集団と標本

問いの対象である母集団を標本を手がかりに推測する



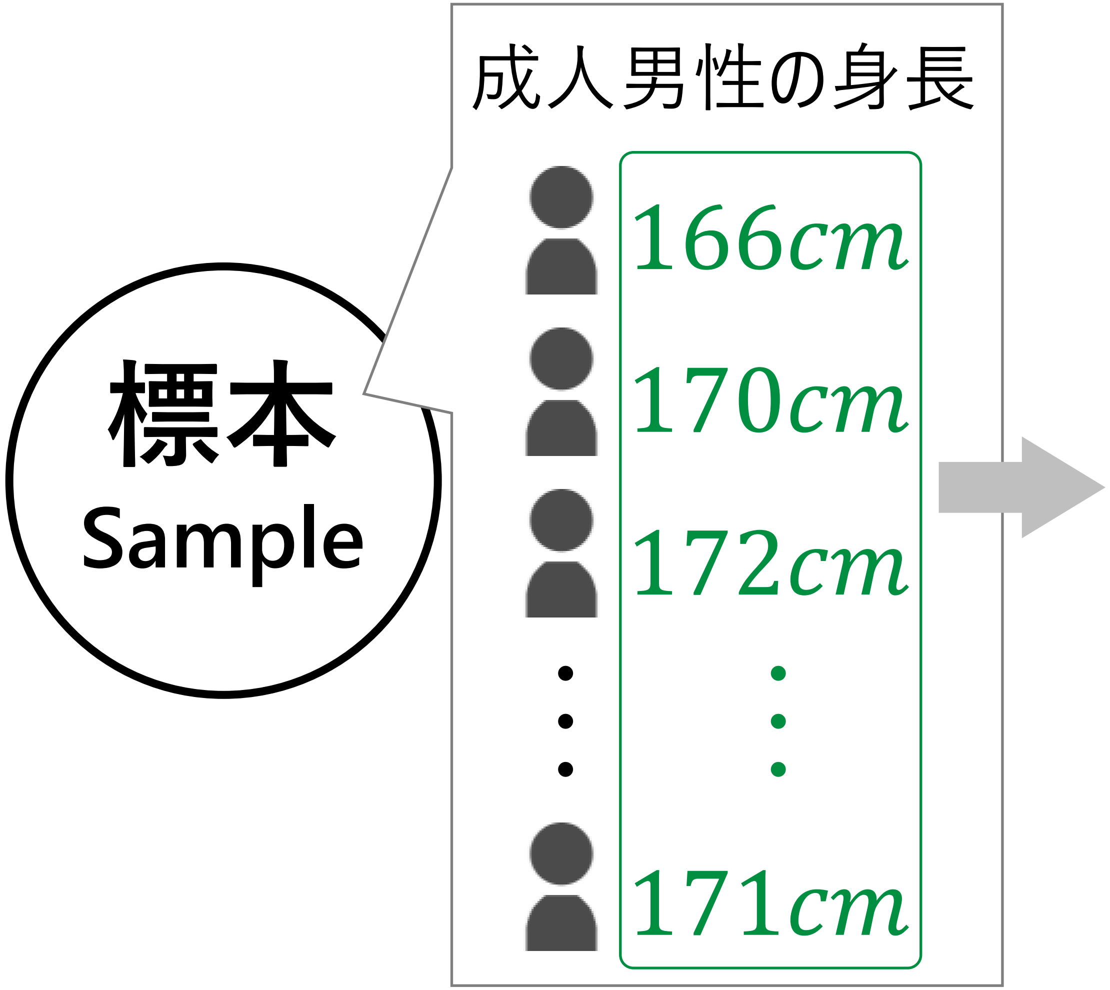
変数

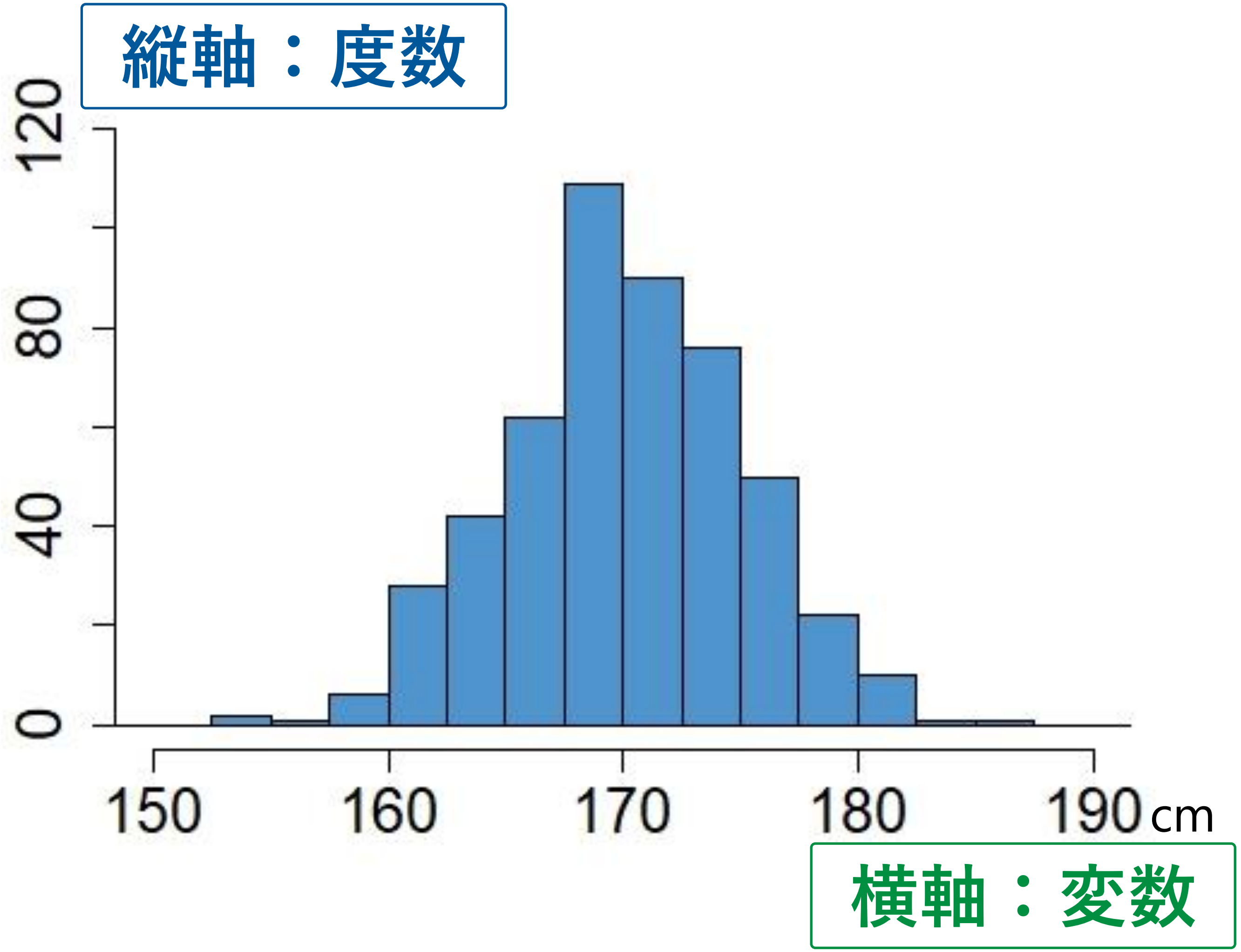
変数は「変化する値」のこと



ヒストグラム（度数分布図）

ヒストグラムは変数を可視化する基本の図





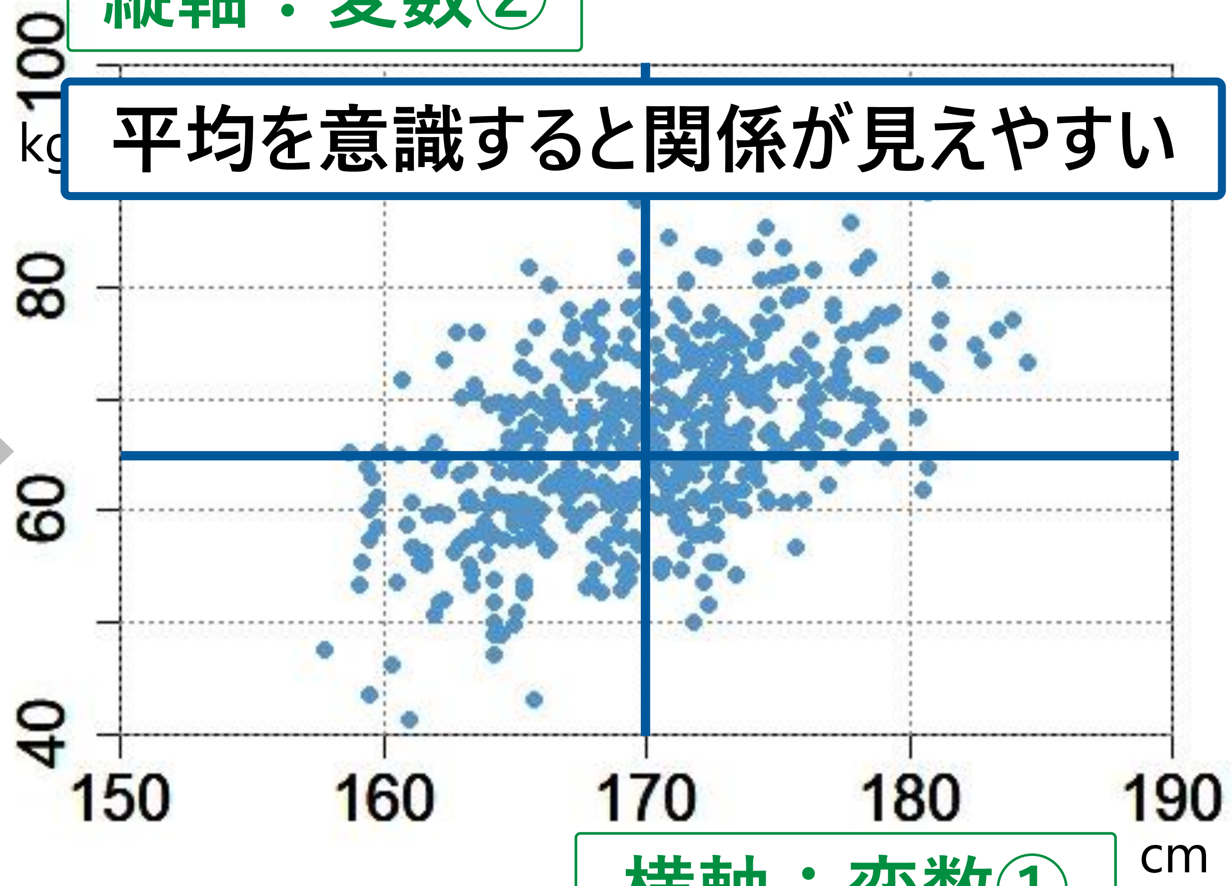
散布図

2つの変数の関係を可視化する基本の図

標本
Sample

	身長	体重
	$166cm$	$62kg$
	$170cm$	$68kg$
	$172cm$	$65kg$
⋮	⋮	⋮
	$171cm$	$70kg$

縦軸：変数②



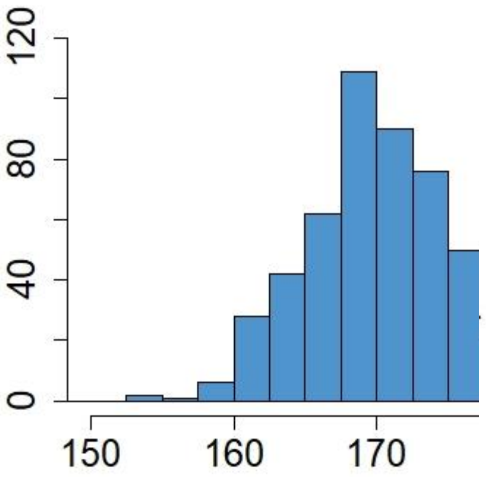
横軸：変数①

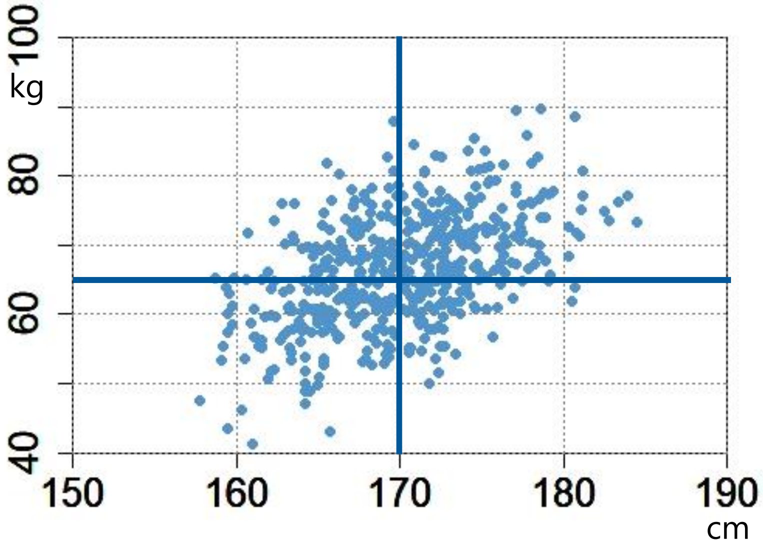
記述統計

変数を具体的な値で記述する

標本
Sample

身長	体重
$166cm$	$62kg$
$170cm$	$68kg$
$172cm$	$65kg$
⋮	⋮
$171cm$	$70kg$



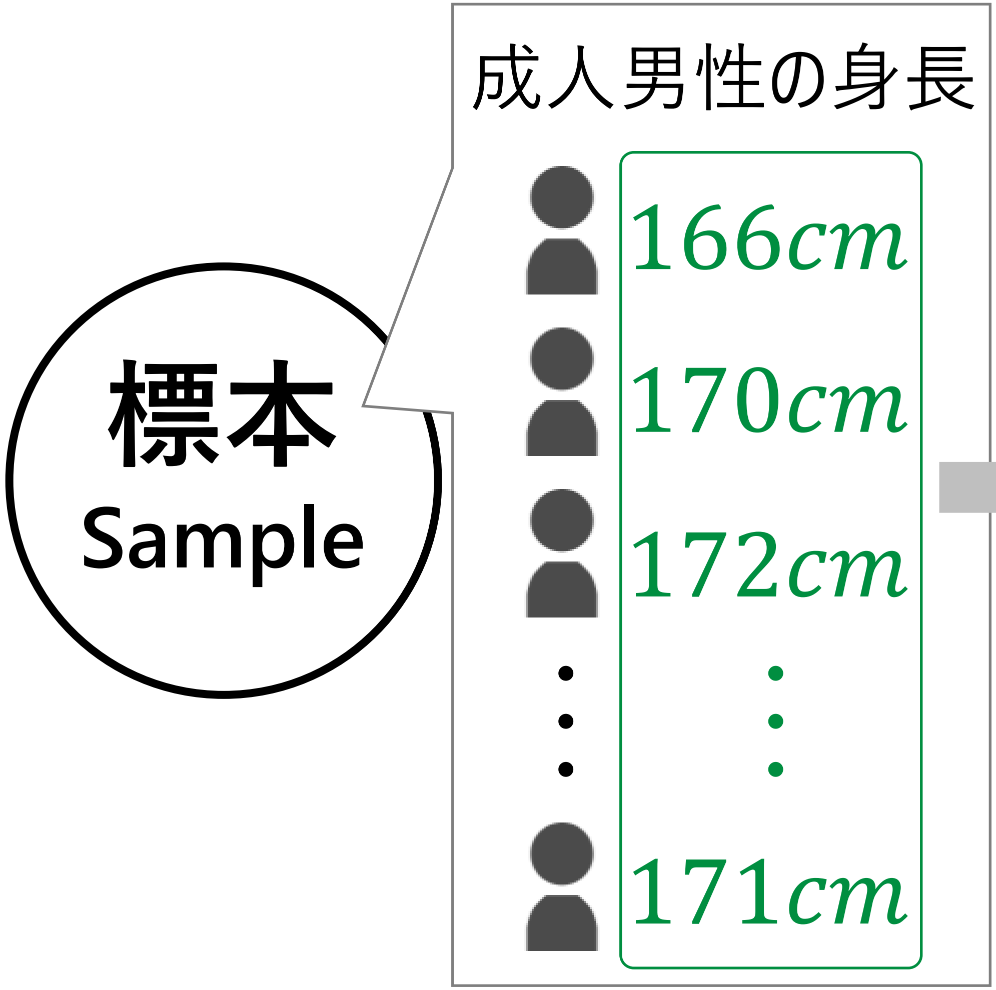


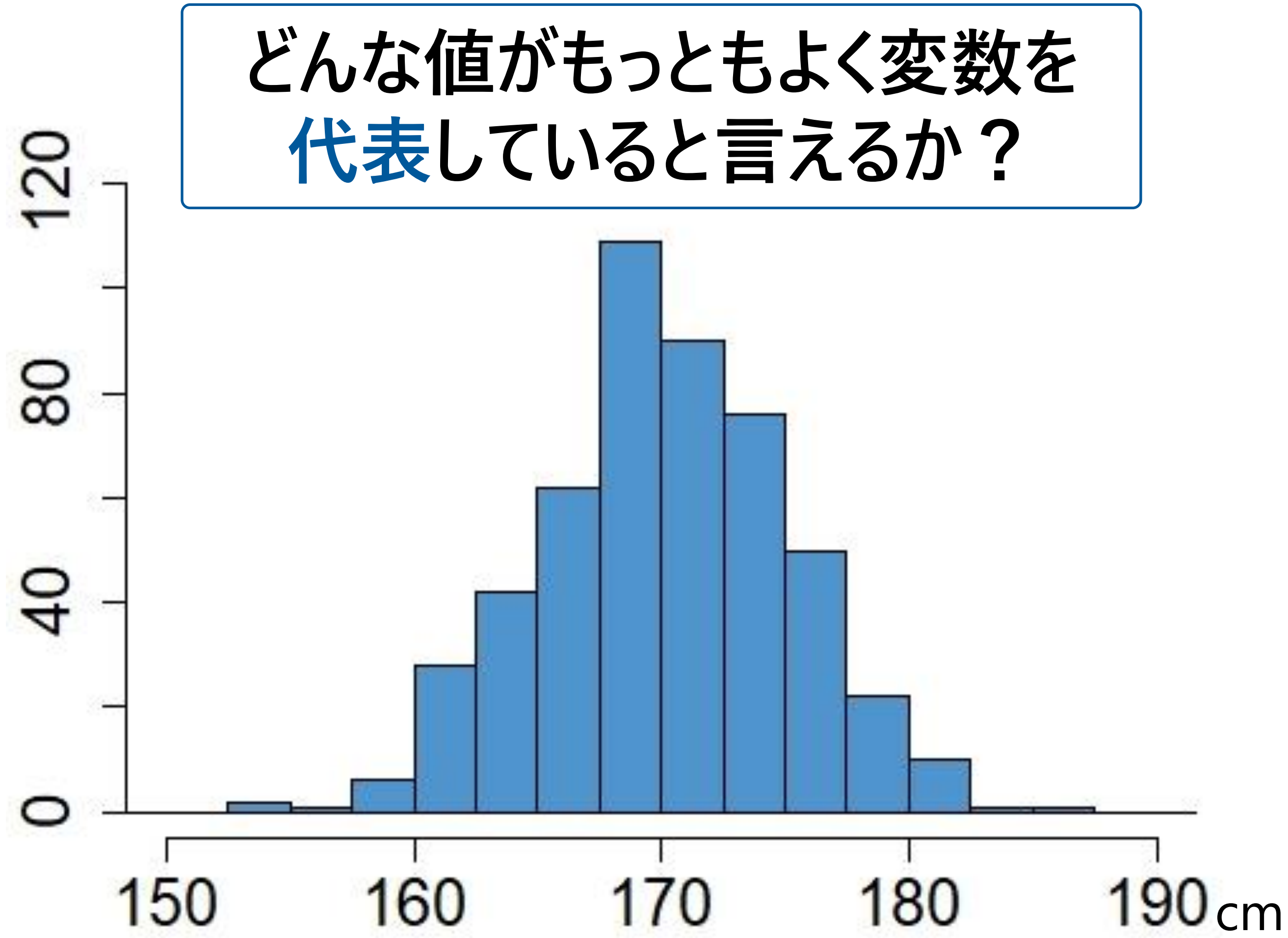
具体的な1つの値で
記述できないか？

記述統計

代表値

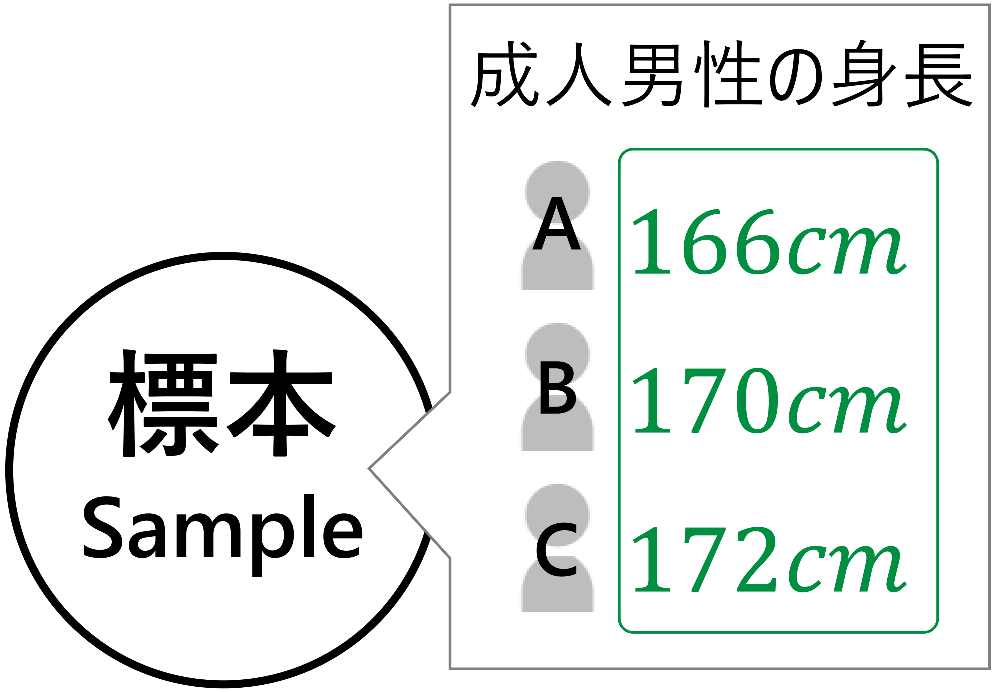
その変数を代表する1つの値





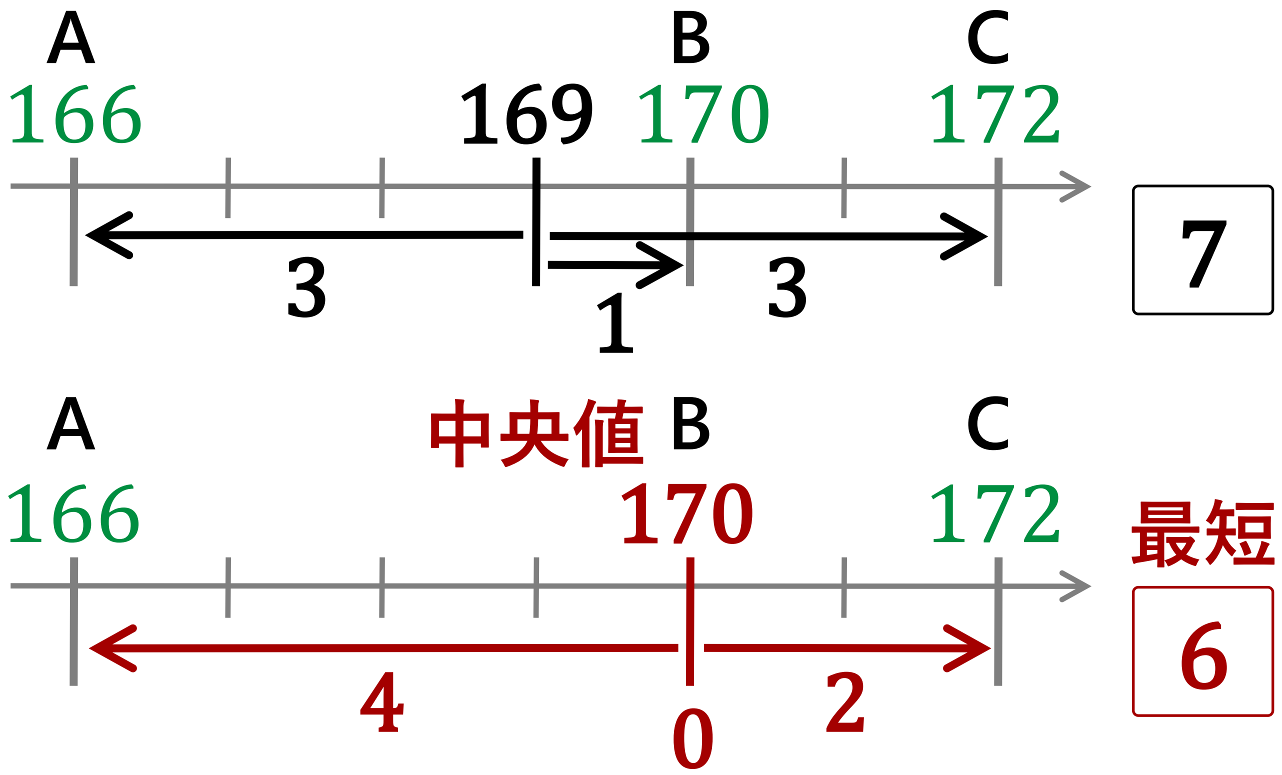
中央値

中央値はちょうど真ん中の順位の値



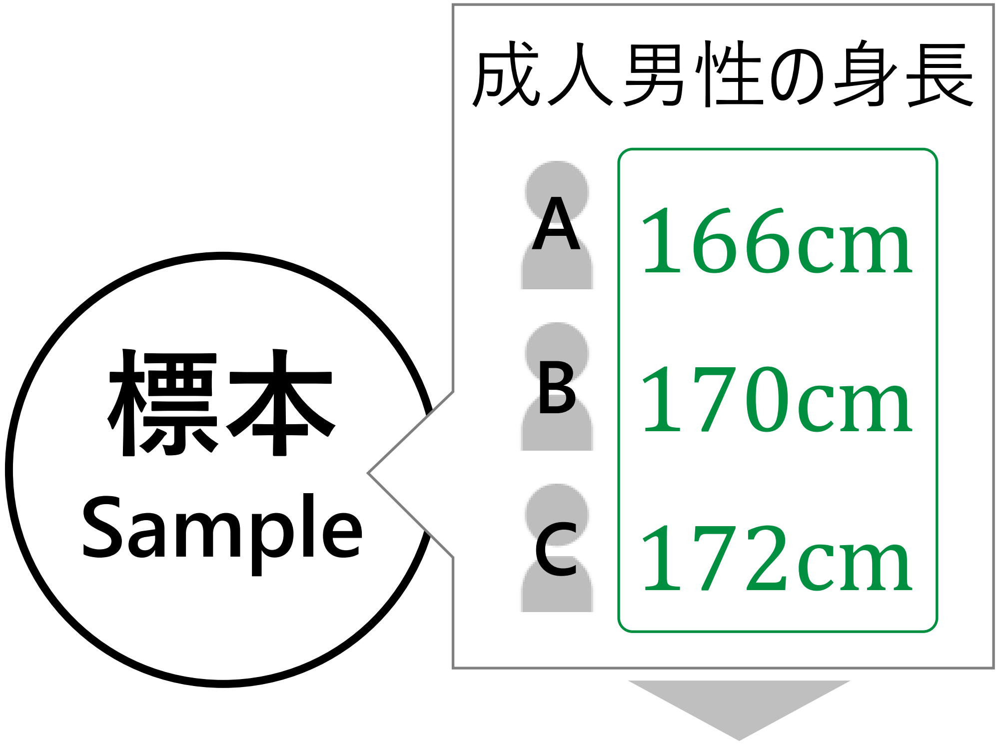
中央値 ＝ 170cm

統計的には…
「各値との距離の合計を最短にする値」



平均

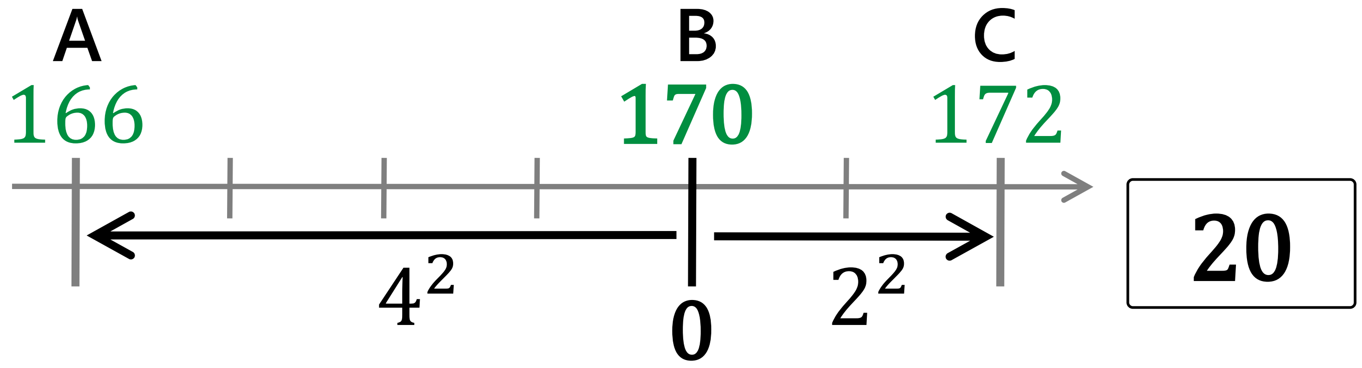
平均はすべて足して標本の大きさで割った値

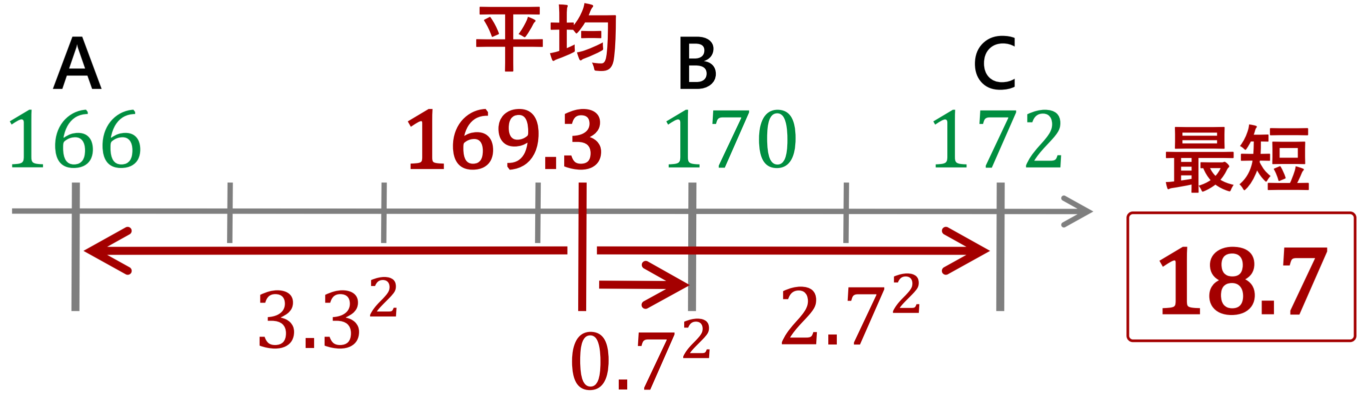


$$\text{平均} = \frac{166+170+172}{3}$$
$$\fallingdotseq 169.3cm$$

統計的には...

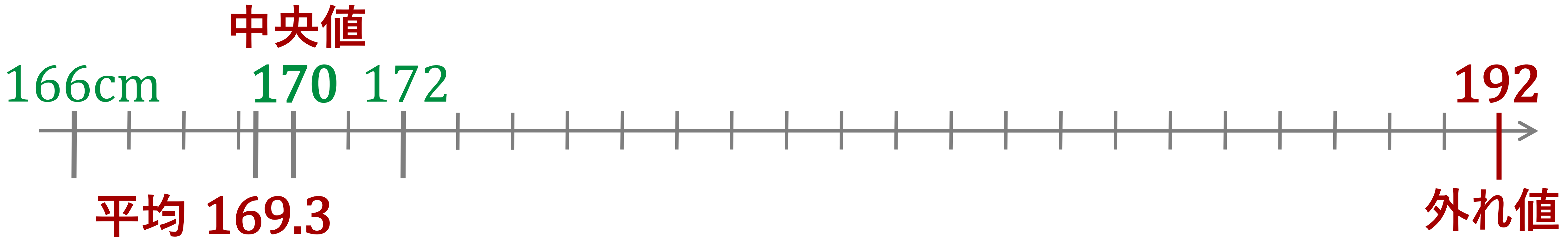
「各値との距離の2乗の合計を最短にする値」





外れ値

外れ値の影響は中央値よりも平均の方が受けやすい



中央値

170cm ⇒ 171cm
順位が１つずれるだけなので
外れ値の影響は小さい

平均

169.3cm ⇒ 175cm
値の離れ度合いも考慮されるので
外れ値の影響が大きい

- ✓ 外れ値にも「情報」はある
- ✓ 中央値と平均の両方を確認することが大切

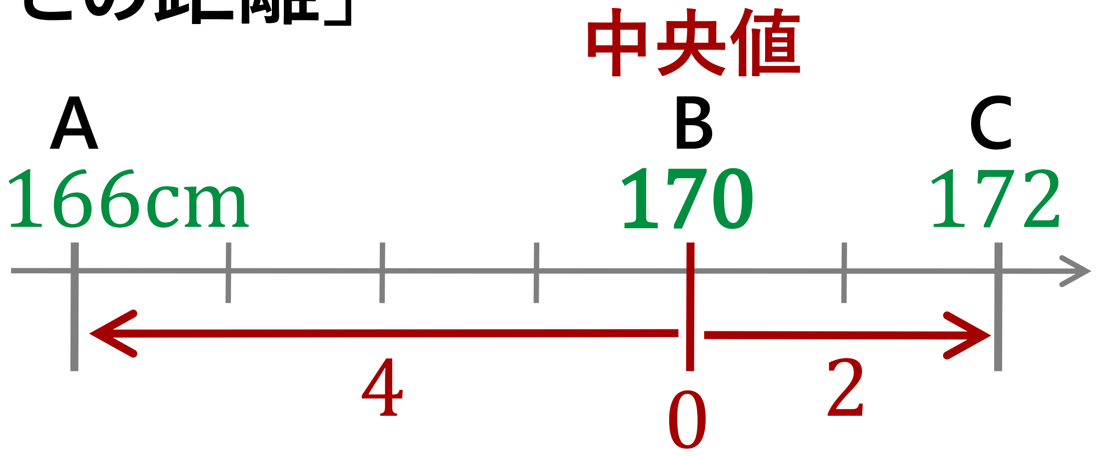
平均偏差

変数における平均的な「中央値※との距離」

標本
Sample

成人男性の身長

A	166cm
B	170cm
C	172cm



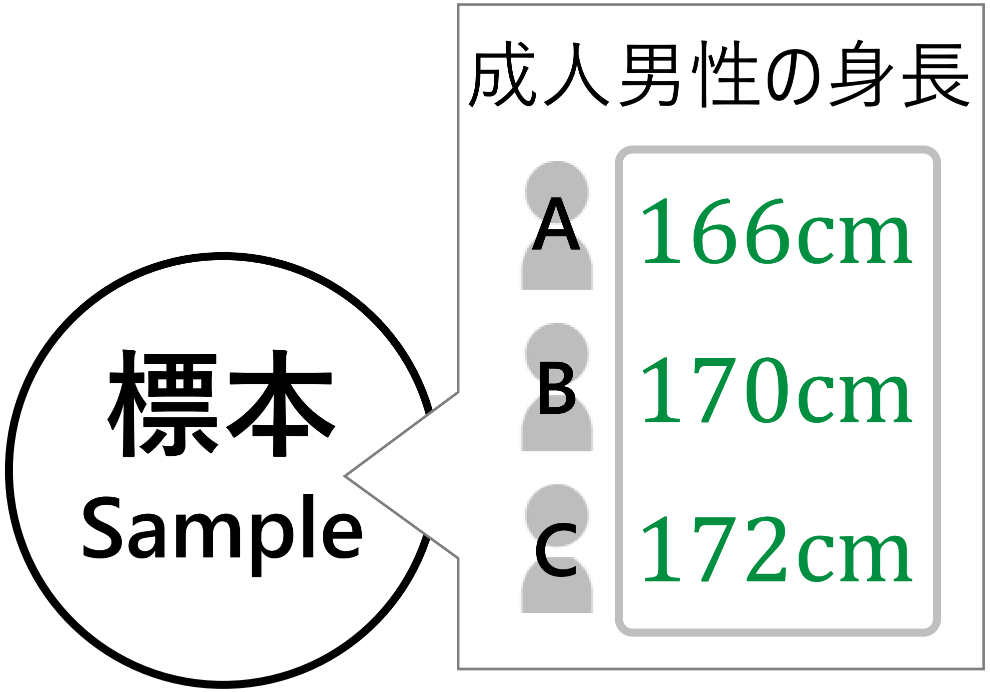
$$\text{平均偏差} = \frac{\text{各値と中央値}^{※}\text{の距離の合計}}{\text{標本の大きさ}}$$

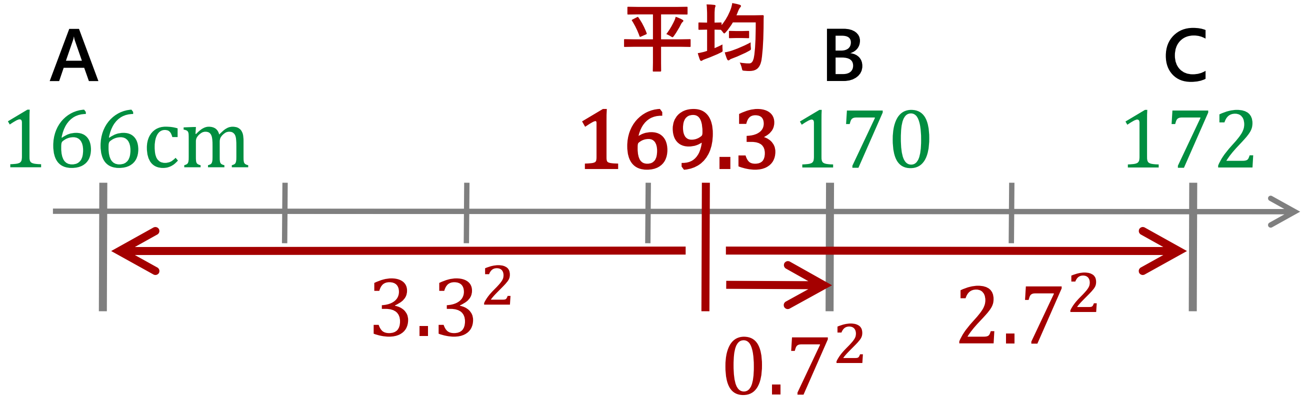
$$= \frac{4+0+2}{3} = 2$$

※中央値でなく平均を用いることもあります

分散

「平均との距離の2乗」を標本の大きさ※で割ったもの





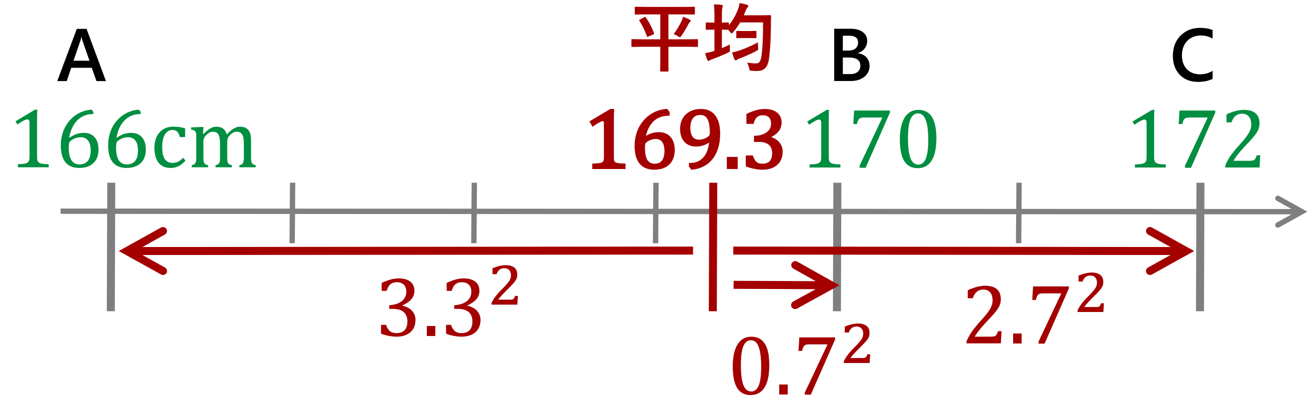
$$分散^{※} = \frac{各値と平均の距離の2乗の合計}{標本の大きさ}$$

$$= \frac{3.3^2 + 0.7^2 + 2.7^2}{3} \fallingdotseq 6.2$$

※標本分散。不偏分散についてはセクション2にてご説明します。

標準偏差

分散の尺度を元に戻して解釈しやすくした値



$$\text{分散}^{※} = \frac{\text{各値と平均の距離の2乗の合計}}{\text{標本の大きさ}}$$

$$= \frac{3.3^2 + 0.7^2 + 2.7^2}{3} \fallingdotseq 6.2$$

$$\text{標準偏差} = \sqrt{\text{分散}^{※}}$$

$$\fallingdotseq \sqrt{6.2}$$

$$\fallingdotseq 2.5$$

※標本分散。不偏分散についてはセクション2にてご説明します。

標準化

特定の平均と標準偏差となるような変数の変換

平　　均：国92点 算80点
標準偏差：国2.9　算17.2

	国語	算数
A	90点	85点
B	92点	65点
C	91点	95点
D	95点	75点

国語と算数の基準をそろえる

標準化

①平均を引く
②標準偏差で割る

平　　均：国0点 算0点
標準偏差：国1.0 算1.0

	国語	算数
A	-0.69点	0.29点
B	0点	-0.87点
C	-0.35点	0.87点
D	1.04点	-0.29点

標準化 修正版

特定の平均と標準偏差となるような変数の変換

平　　均：国92点 算80点
標準偏差：国1.9　算11.2

修正箇所

	国語	算数
A	90点	85点
B	92点	65点
C	91点	95点
D	95点	75点

国語と算数の

平　　均：国0点 算0点
標準偏差：国1.0 算1.0

動画内スライドならびに講師のコメントに誤りがございました。正しくは赤字の通り国語の標準偏差が「1.9」、算数の標準偏差が「11.2」となります。申し訳ございません。
なお、ここでは標本分散から標準偏差を計算しております。標本分散とは別に不偏分散という値がございますが、これについてはセクション2（点推定）にてご説明いたします。
引き続きどうぞよろしくお願いいたします。

D	1.04点	-0.29点

セクション2：点推定

点推定とは

標本統計量（推定量）で母数をピンポイントに推定する

両者は別物

母集団

抽出 →

標本

真値

母数
（母集団を特徴づける
- ✓母平均
- ✓母分散
- ✓母標準偏差 etc.

ある1つの値で
ピンポイントに推定

←

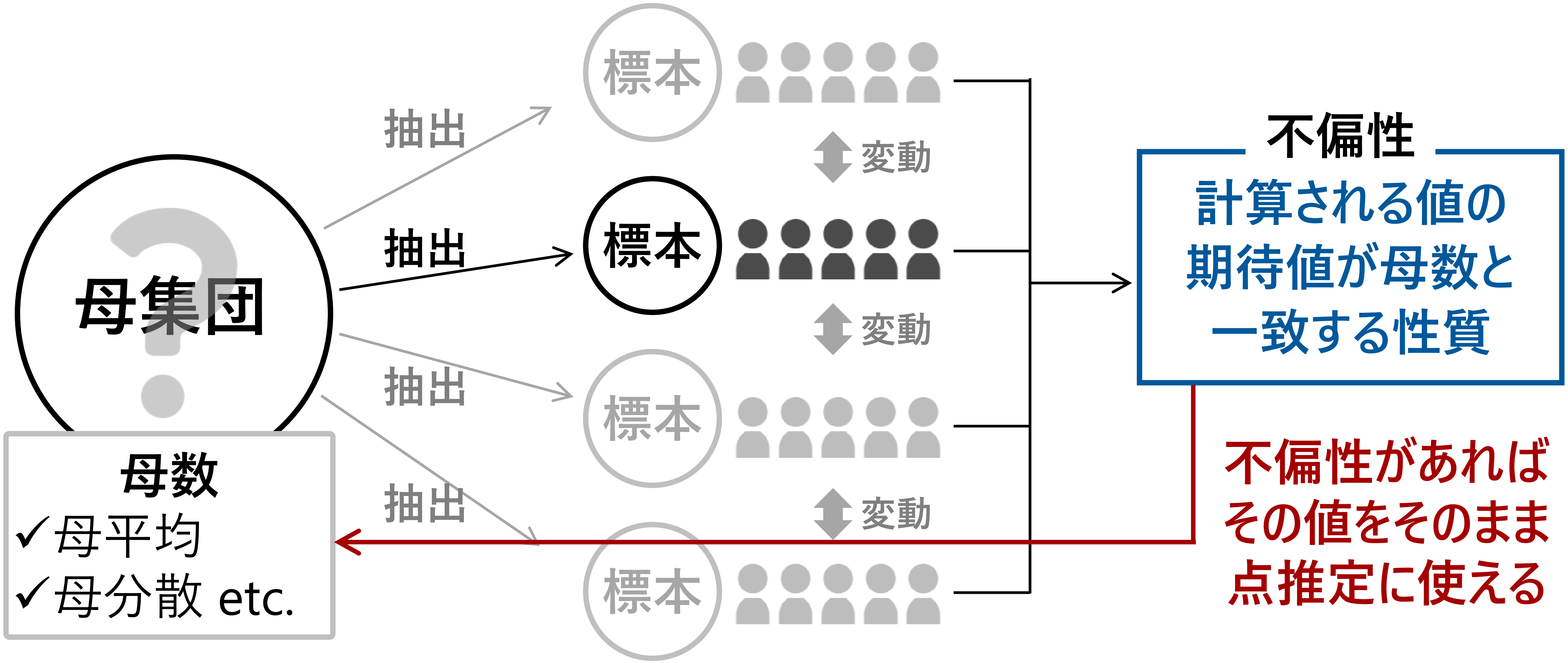
変動

推定量

標本統計量
本から計算できる値）
- ✓標本平均
- ✓標本分散・不偏分散
- ✓標本標準偏差 etc.

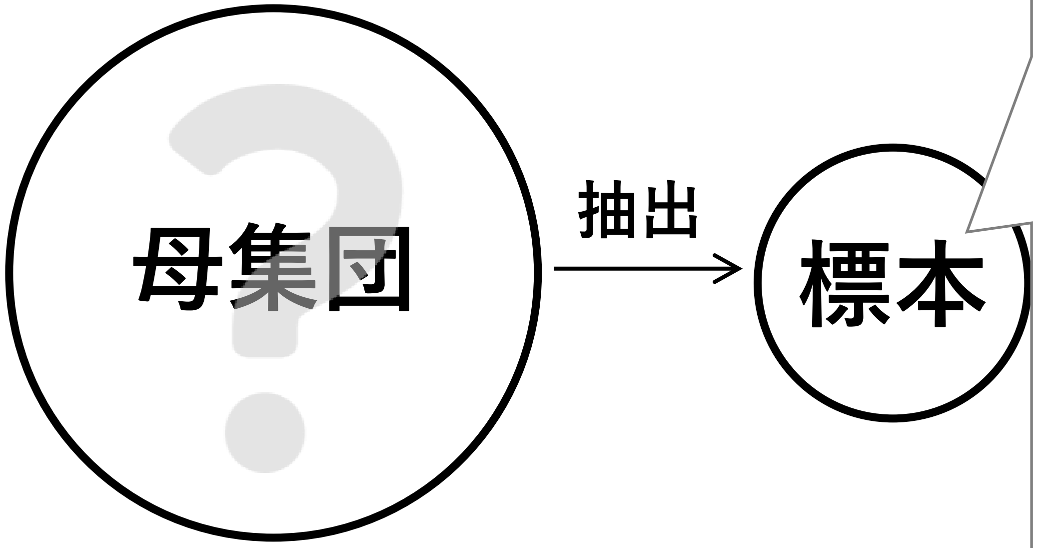
不偏性

平均的には母数と一致する（偏らない）性質



平均の点推定

標本平均で母平均を点推定



	身長
A	164cm
B	170cm
C	173cm

標本平均 $\boxed{\text{不偏性○}}$

$$= \frac{164 + 170 + 173}{3}$$

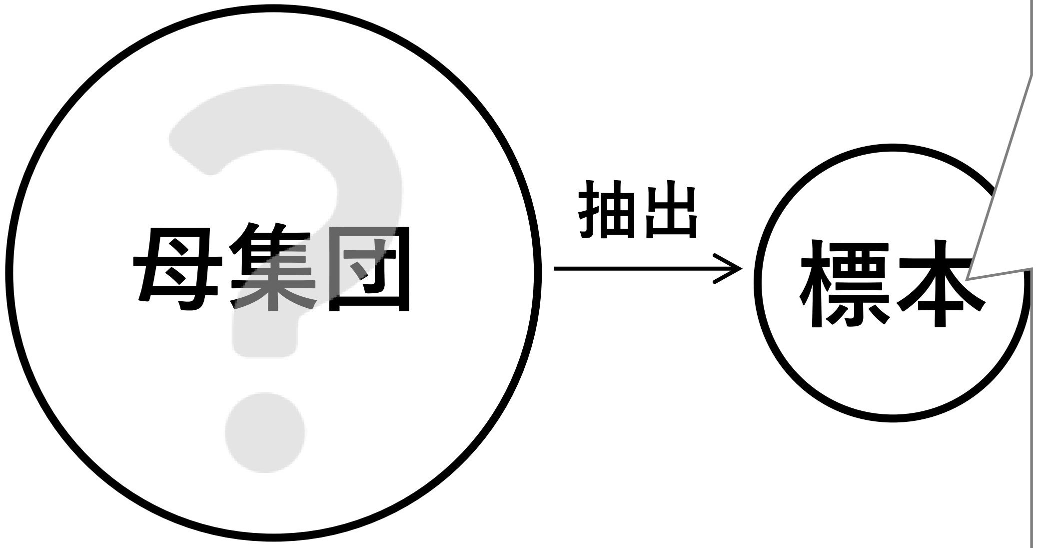
$$= 169$$

母平均＝？

母平均の点推定に
標本平均を使える

分散の点推定

分散の推定には不偏分散を用いる



	身長
A	164cm
B	170cm
C	173cm
	平均169cm

標本分散 不偏性×

$$= \frac{5^2 + 1^2 + 4^2}{3} = 14$$

不偏分散 不偏性○

$$= \frac{5^2 + 1^2 + 4^2}{3-1} = 21$$

母分散＝？

母分散の点推定には不偏分散を使う

標本分散と不偏分散

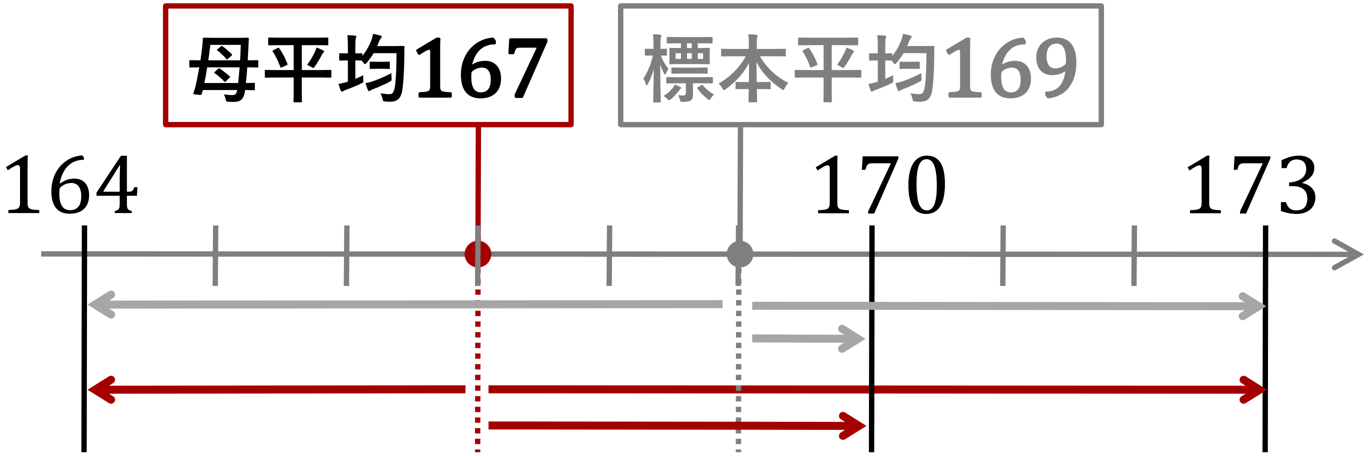
標本分散は過小評価されてしまっている

$$\text{標本分散} = \frac{\text{標本平均との距離の2乗の合計}}{\text{標本の大きさ}}$$

母分散を過小評価

$$\text{不偏分散} = \frac{\text{標本平均との距離の2乗の合計}}{\text{標本の大きさ}-1}$$

過小評価分を補正



イメージ

標本分散は分子の「距離の２乗の合計」を過小に見積もっている

セクション3：確率分布

なぜ「確率」か

母集団を標本を手がかりに確率論で推測する

両者は別物

母集団　—抽出→　標本

真値

母数
（母集団を特徴づける値）
- ✓母平均
- ✓母分散
- ✓母標準偏差 etc.

←推定

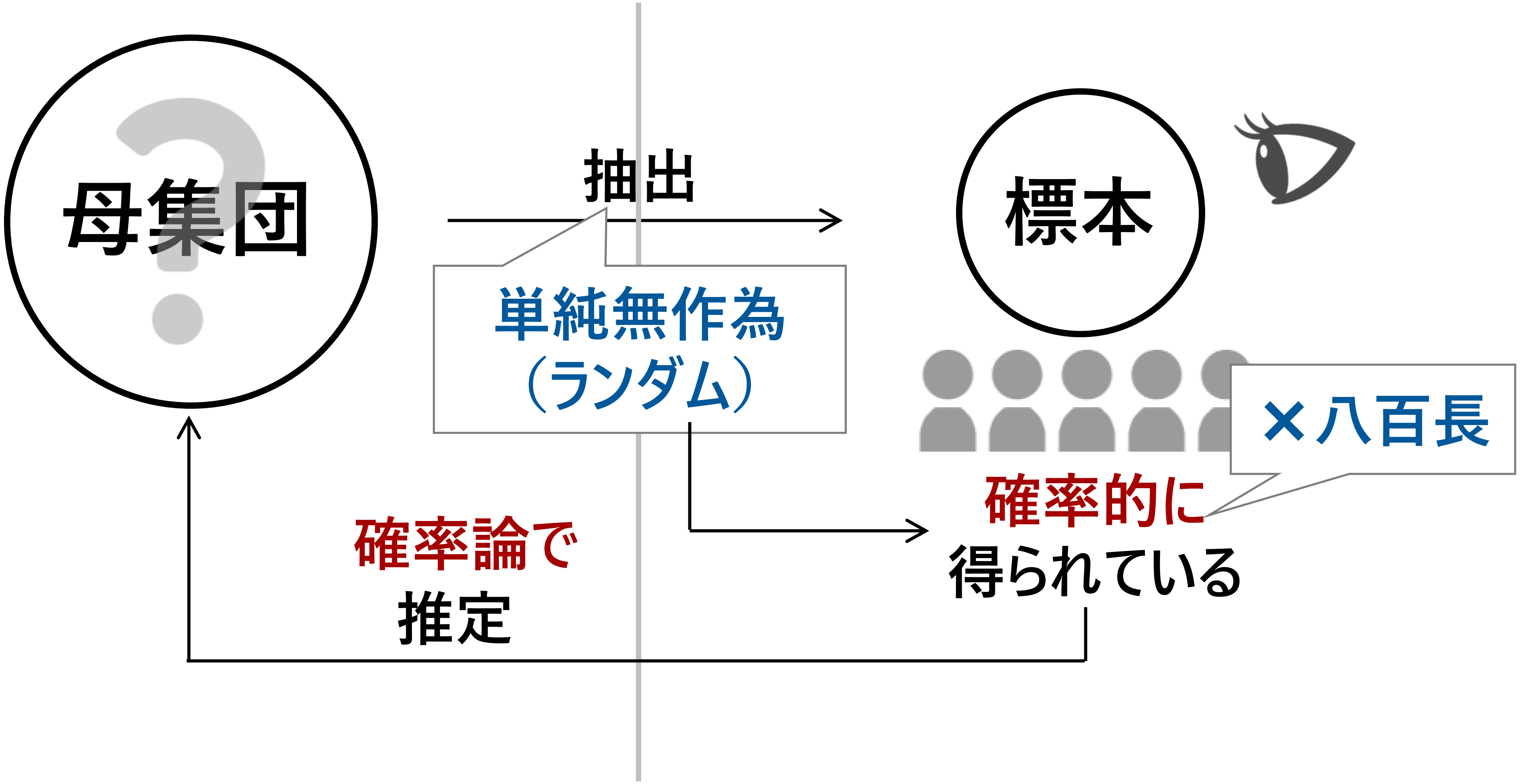
確率論を用いる

変動

標本統計量
（標本から計算できる値）
- ✓標本平均
- ✓標本分散・不偏分散
- ✓標本標準偏差 etc.

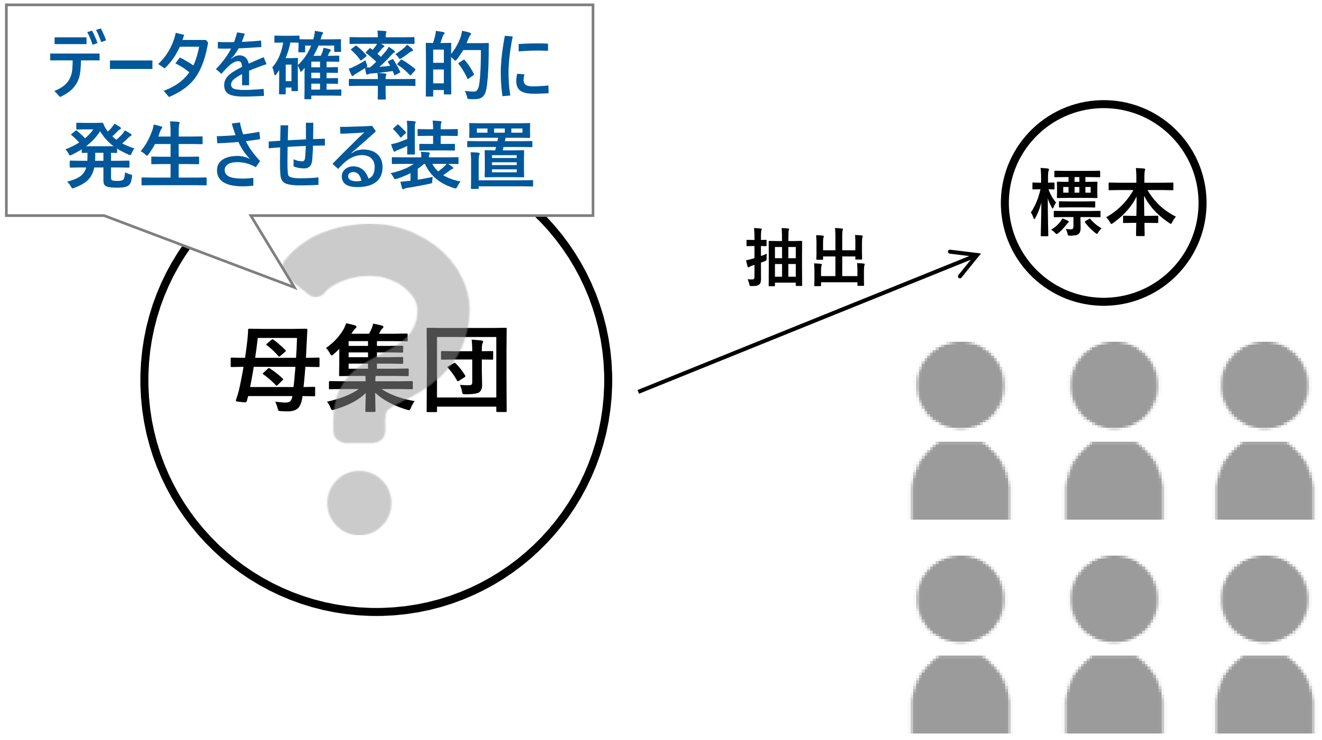
単純無作為抽出

母集団からランダムに標本を抽出する



確率モデル

標本は母集団から確率的に発生していると考える



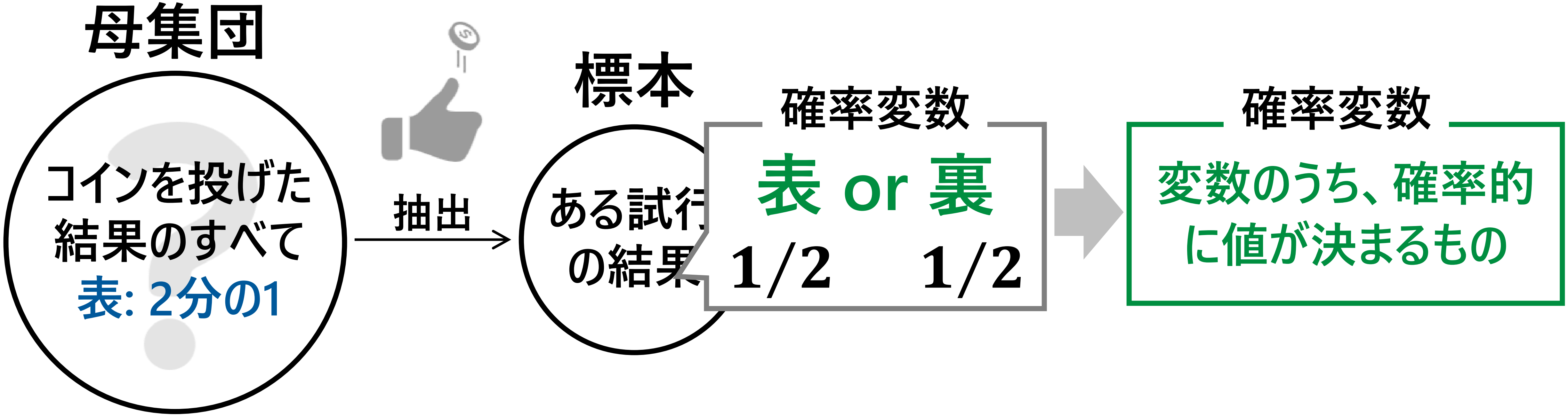
確率モデル

1つ1つのデータの
発生は、母集団の
確率的なデータ発生の
しくみにしたがっている

確率変数

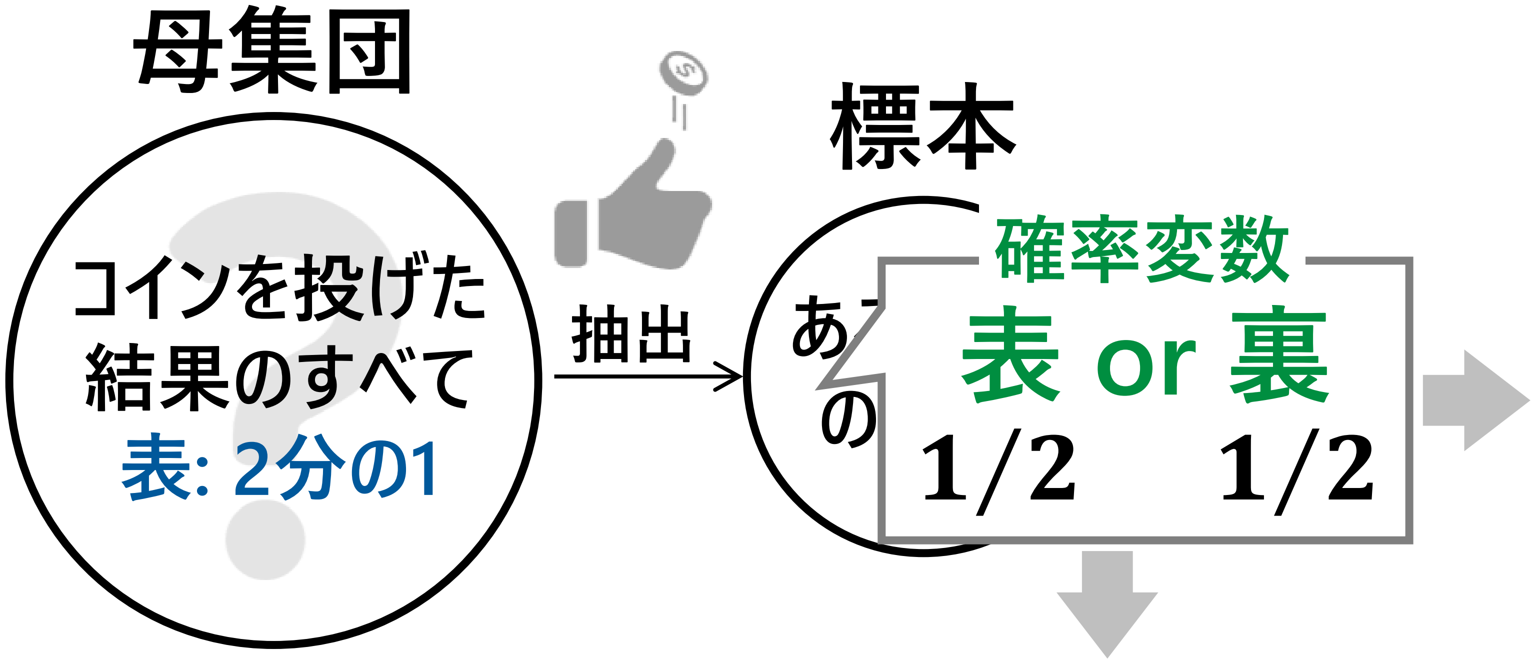
確率的に値が決まる変数

例：2分の1の確率で表が出るコインを投げる

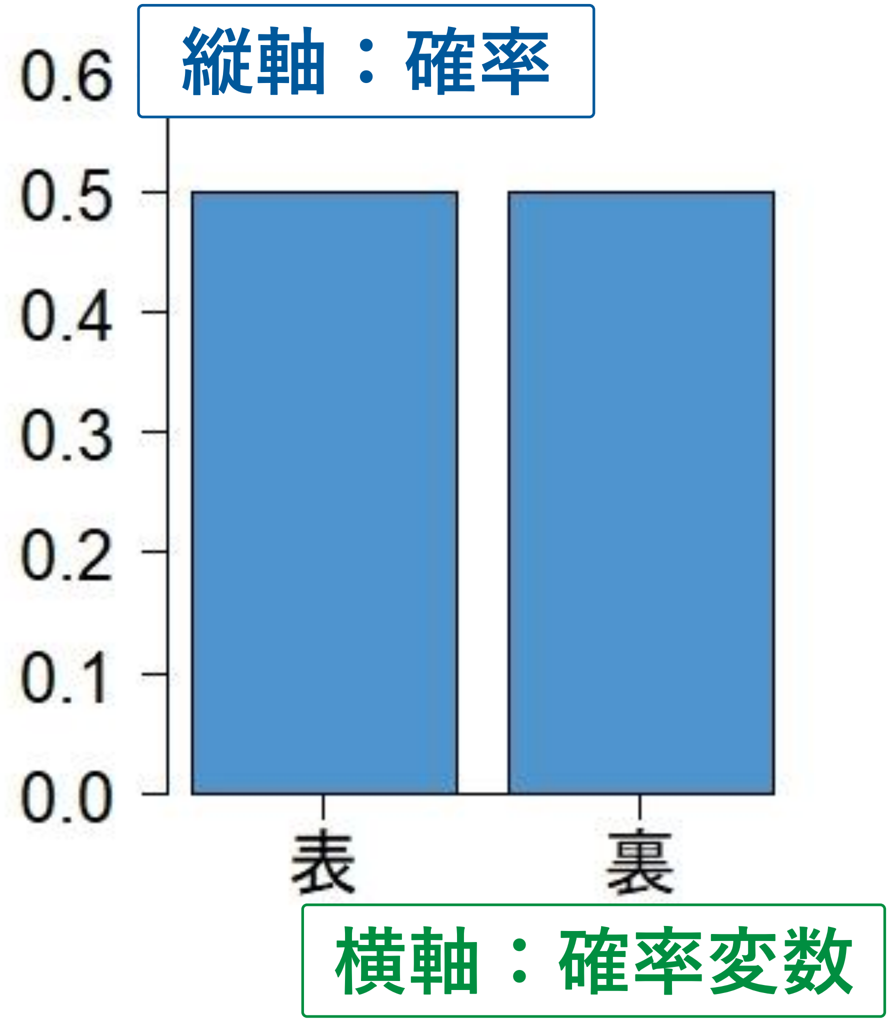


確率分布

確率変数のそれぞれの値が発生する確率の分布

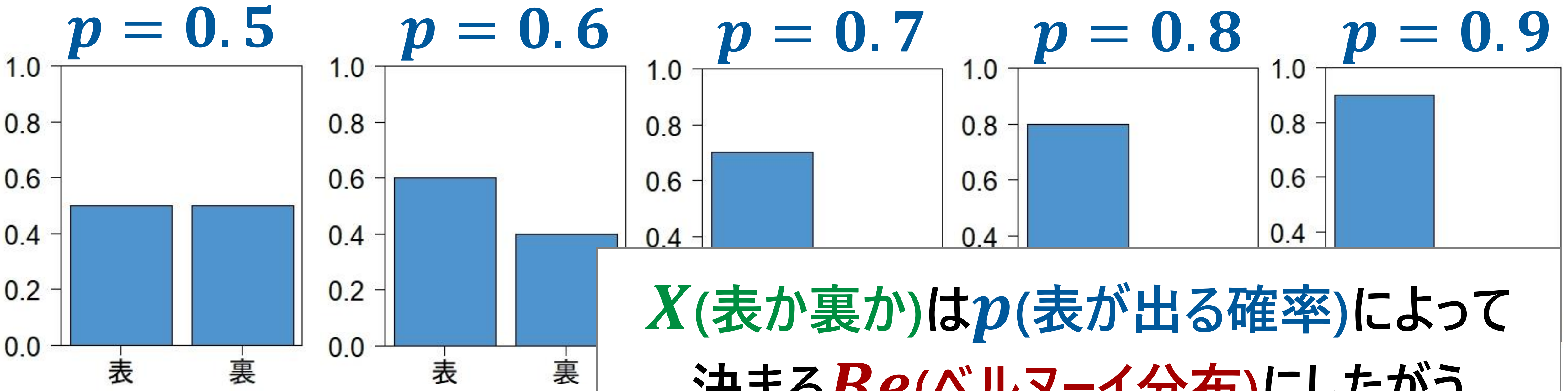


確率変数	表	裏	合計
確率	1 / 2	1 / 2	1



確率関数とパラメータ

確率関数は確率変数と発生確率のルールを式にしたもの



X(表か裏か)はp(表が出る確率)によって決まるBe(ベルヌーイ分布)にしたがう

ルールを記号で表記すると…

表が出る確率$= p$

裏が出る確率$= 1 - p$

$$X \sim Be(p)$$

確率分布の表記

確率分布の表記は「●●が"●●"にしたがう」

ルールを記号で表記すると…

表が出る確率$= p$

裏が出る確率$= 1 - p$

$$\underset{①}{X} \underset{②}{\sim} \underset{③}{Be}(\underset{④}{p})$$

①確率変数

例)表か裏か

②「したがう」

③分布の型

④パラメータ

二項分布

何回かコインを投げて表が出る回数がしたがう分布

例: 表の確率2分の1のコインを4回投げる

1回目　2回目　3回目　4回目

表 0 回の確率 = $(1/2)^4$ ×1通り = 0.0625

表 1 回の確率 = $(1/2)^4$ ×4通り = 0.25

表 2 回の確率 = $(1/2)^4$ ×6通り = 0.375

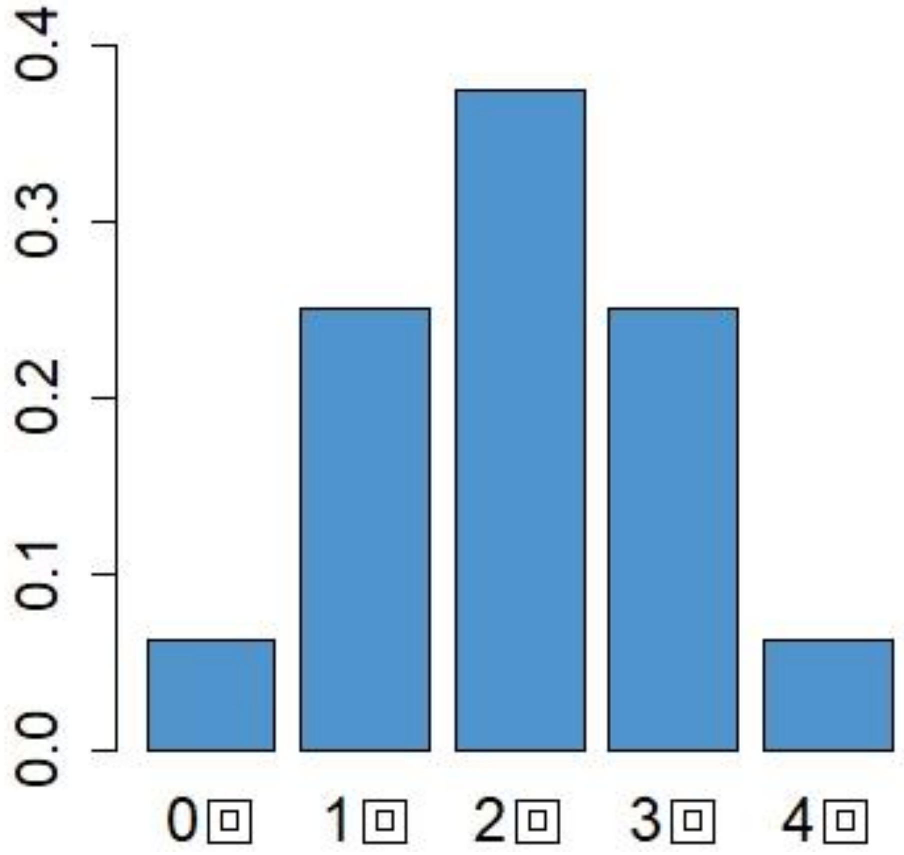
表 3 回の確率 = $(1/2)^4$ ×4通り = 0.25

表 4 回の確率 = $(1/2)^4$ ×1通り = 0.0625

式にすると…

$$P(X = k\text{回}) = \boxed{{}_nC_k\, p^k (1-p)^{n-k}}$$

$$X \sim B(n = 4, p = 0.5)$$

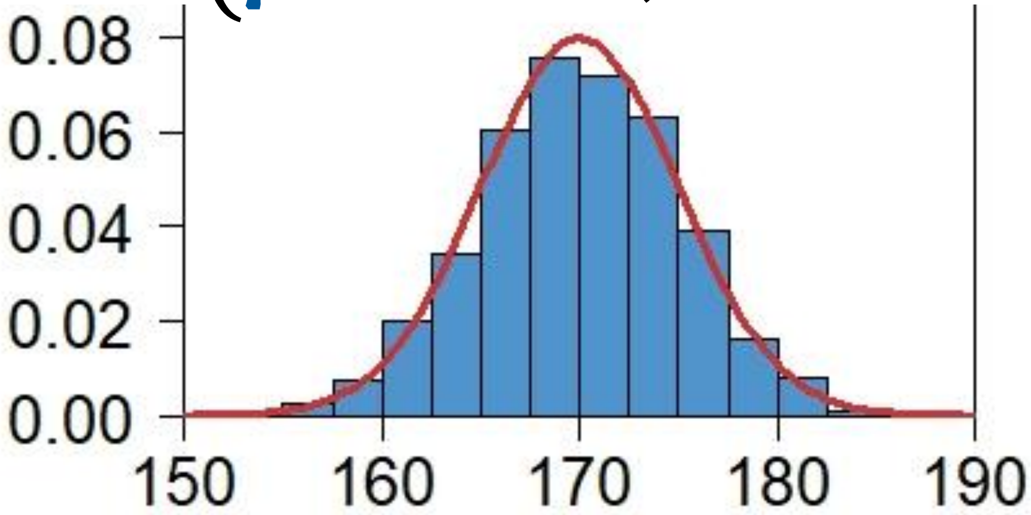


正規分布

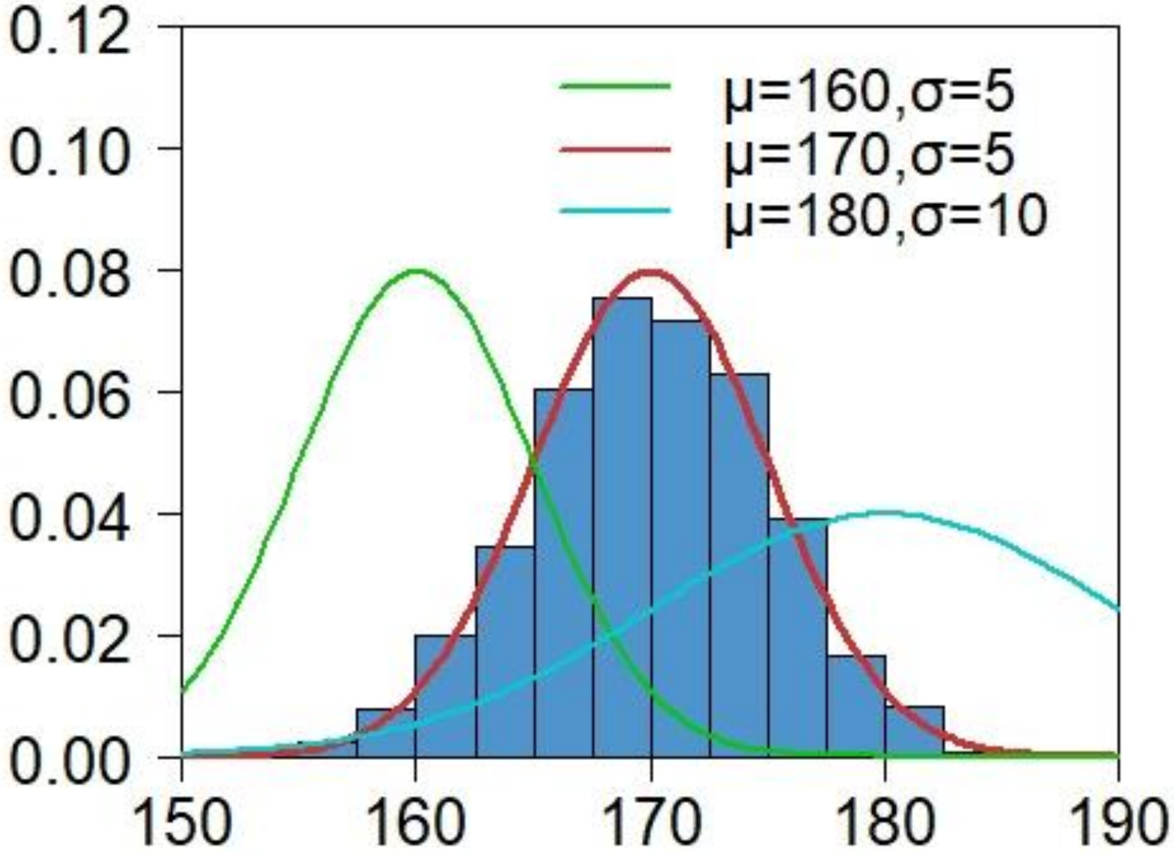
最も重要なふつうで自然な確率分布

例: 成人男性の身長の確率分布

$$X \sim N(\mu = 170, \sigma^2 = 5^2)$$



$$X \sim N(\mu, \sigma^2)$$



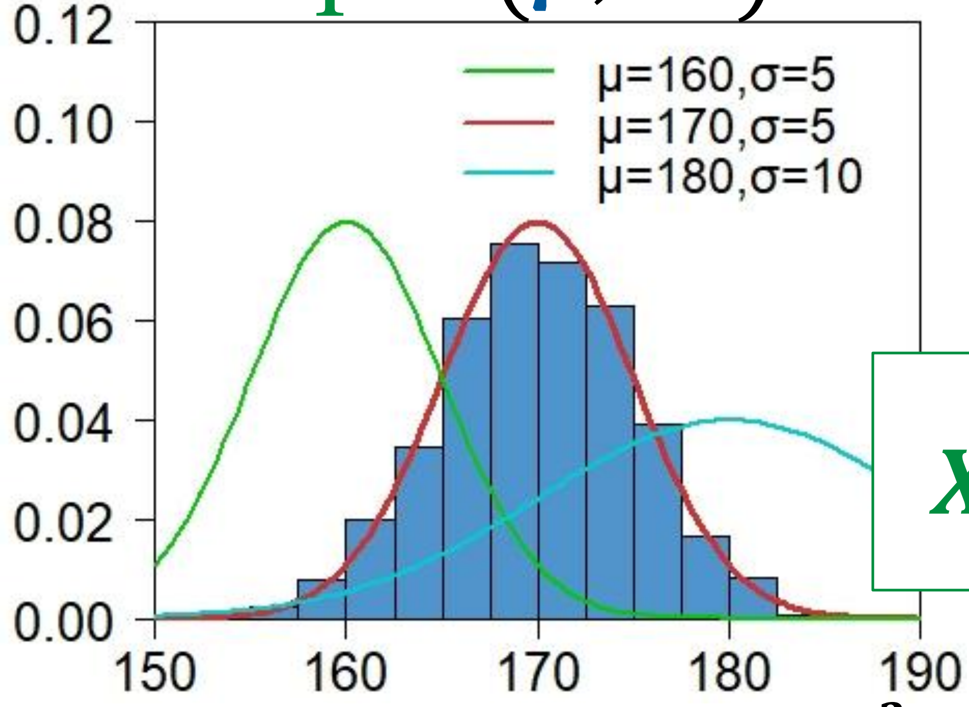
式にすると…

$$f(x) = \boxed{\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}}$$

標準正規分布

正規分布を平均0、分散(標準偏差)1に標準化したもの

$$X_1 \sim N\left(\mu, \sigma^2\right)$$



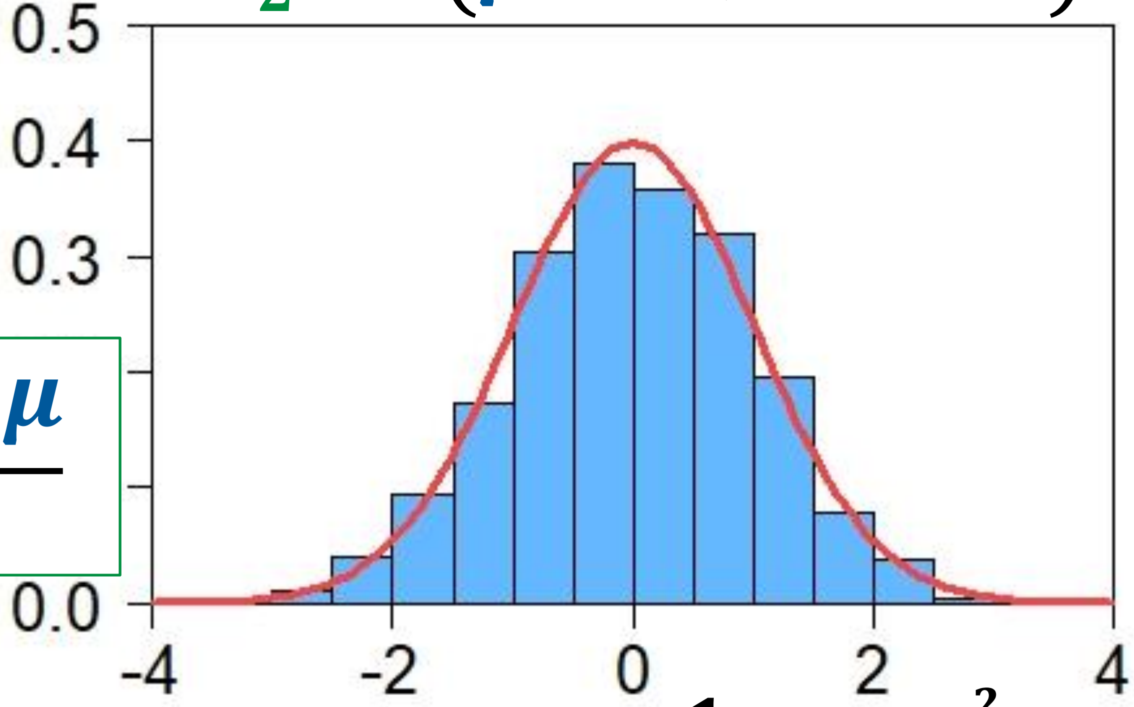
再生性

標準化

$$X_2 = \frac{X_1 - \mu}{\sigma}$$

割る

$$X_2 \sim N\left(\mu = 0, \sigma^2 = 1\right)$$

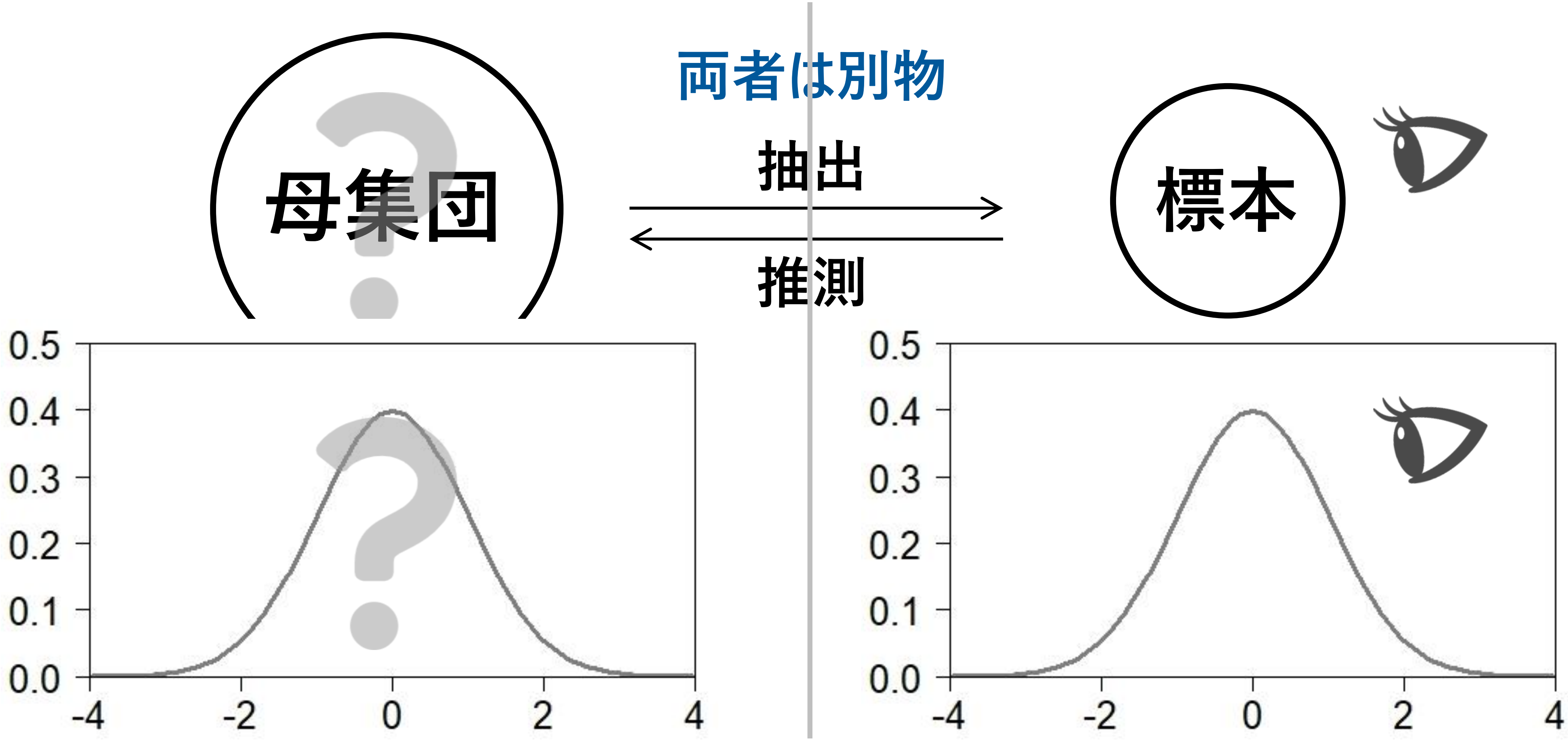


$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

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

母集団分布と標本分布

母集団分布と標本分布は必ず分けて考える



セクション4：区間推定①

区間推定とは

母数を確率的に区間で推定する

両者は別物

母集団 抽出→ 標本

母集団 ←推測 標本

真値

母数
（母集団を特徴づける値）

✓母平均
✓母分
✓母標
✓母比率 etc.

推定 ←

変動

標本統計量
（標本から計算できる値）

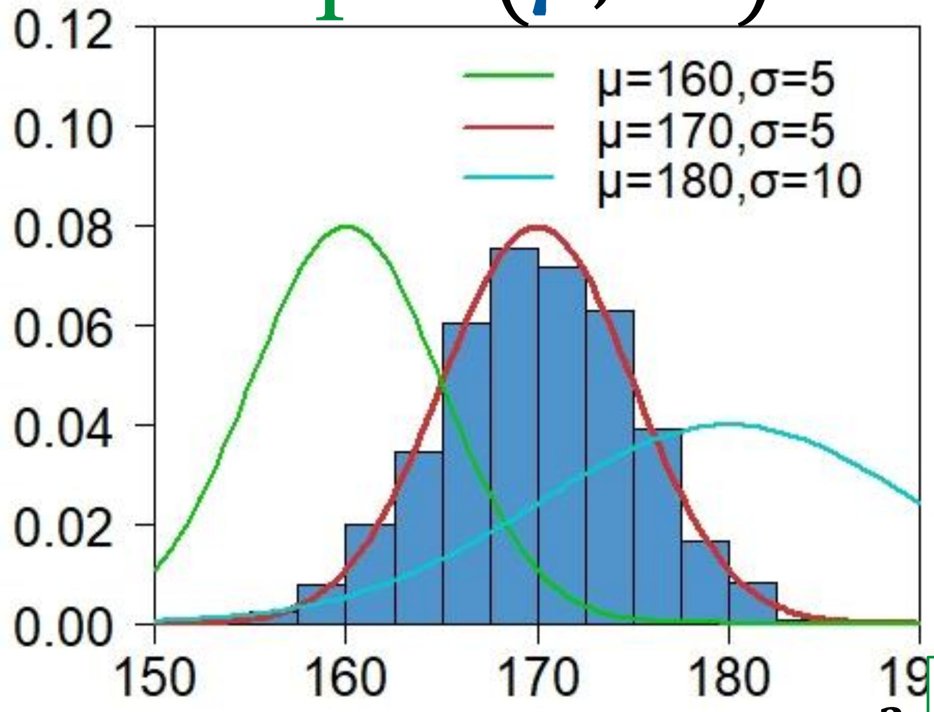
✓標本平均
不偏分散
偏差
✓標本比率 etc.

✓ 値がどんな範囲をとるか区間で推定
✓ 確率的に（100回中95回は…）

正規分布・標準正規分布（復習）

確率的に考えるために（標準）正規分布を前提とする

$X_1 \sim N(\mu, \sigma^2)$

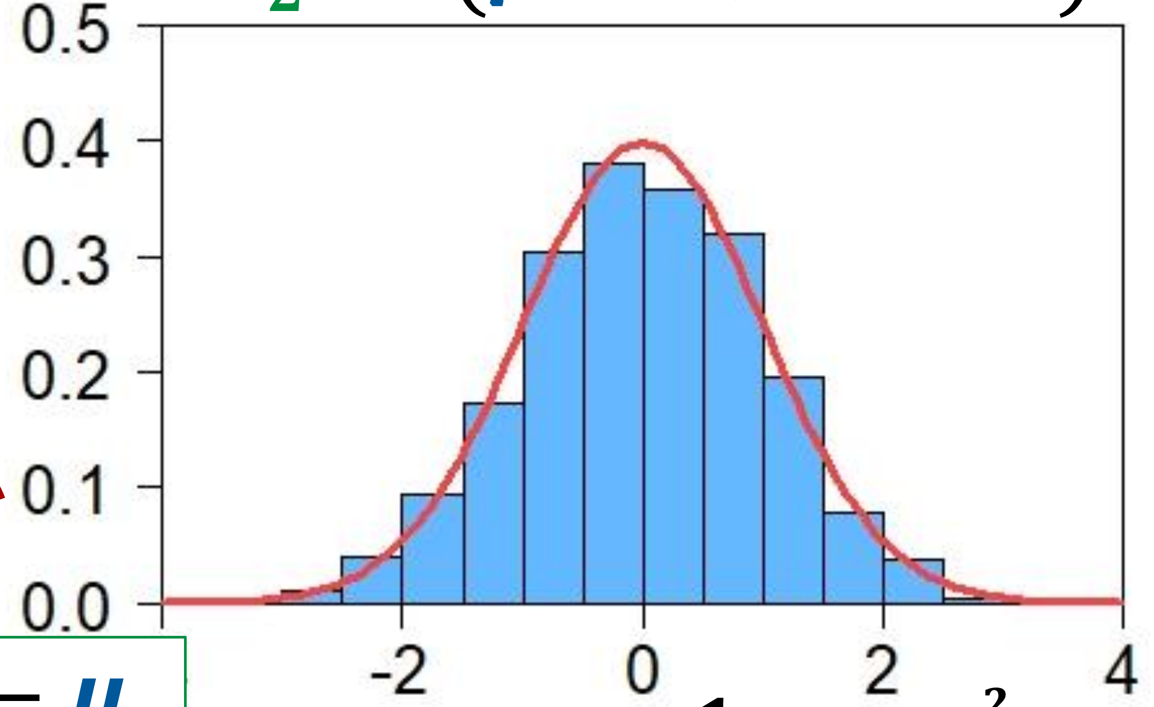


再生性

標準化

①平均を引く
②標準偏差で割る

$X_2 \sim N(\mu = 0, \sigma^2 = 1)$



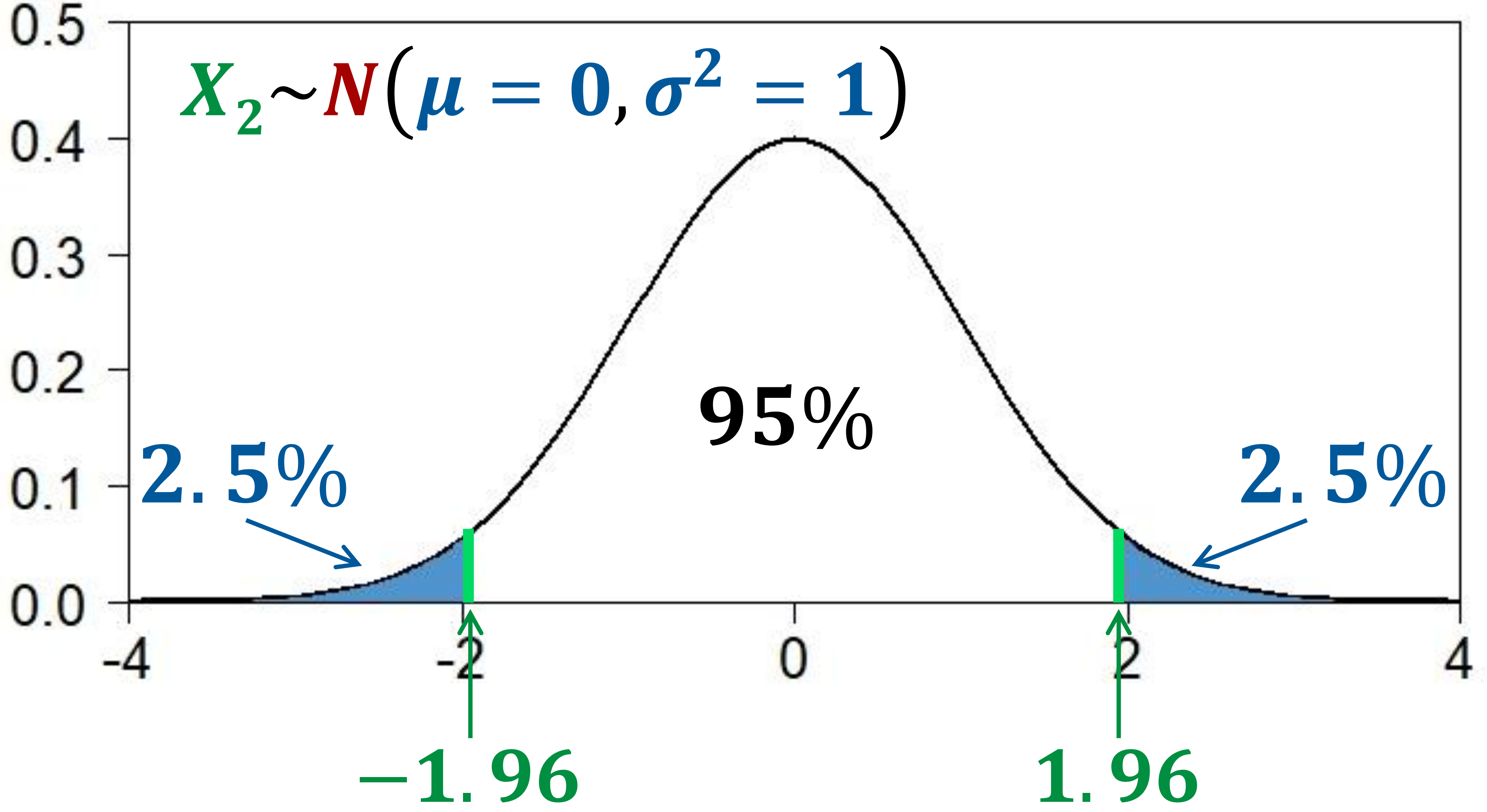
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$X_2 = \frac{X_1 - \mu}{\sigma}$$

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

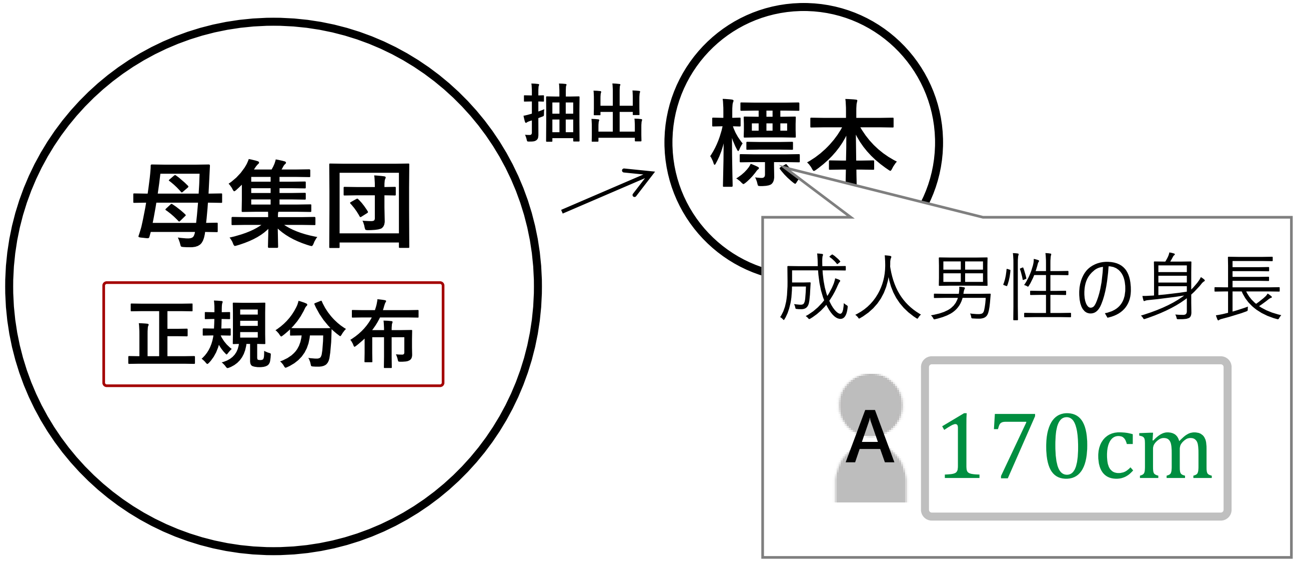
標準正規分布の両側5%点

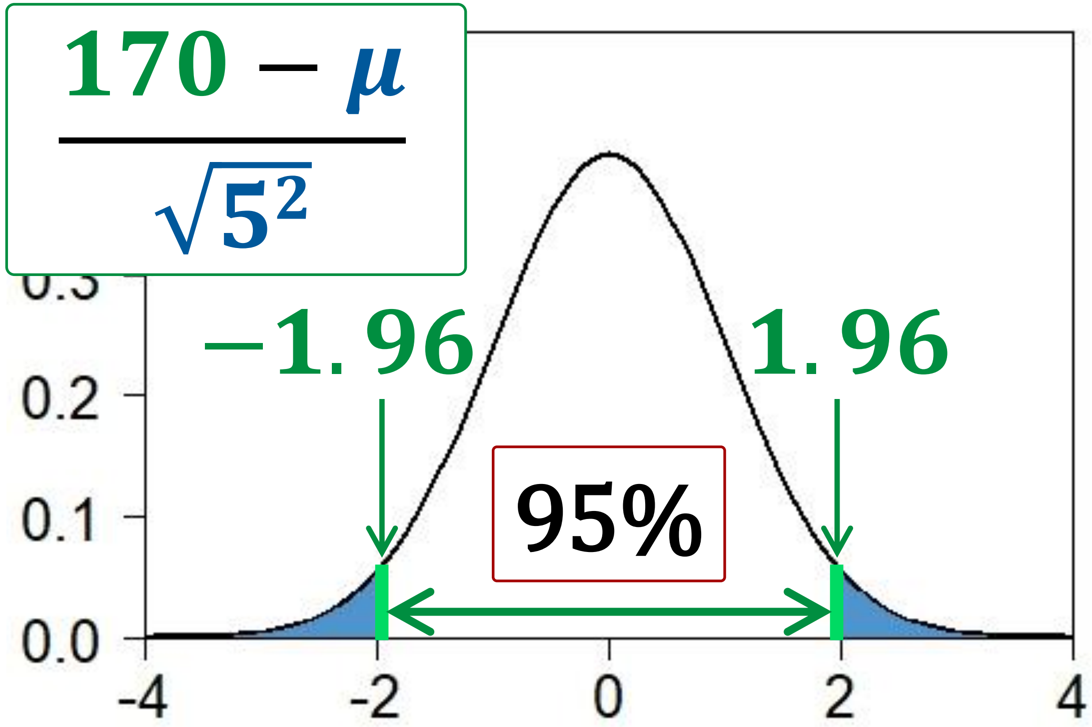
100%を95%と5%に分ける点（境界ライン）：±1.96



母平均の区間推定

標本から母平均を区間推定





$$-1.96 \le \frac{170-\mu}{\sqrt{5^2}} \le 1.96$$

母平均μ = ?

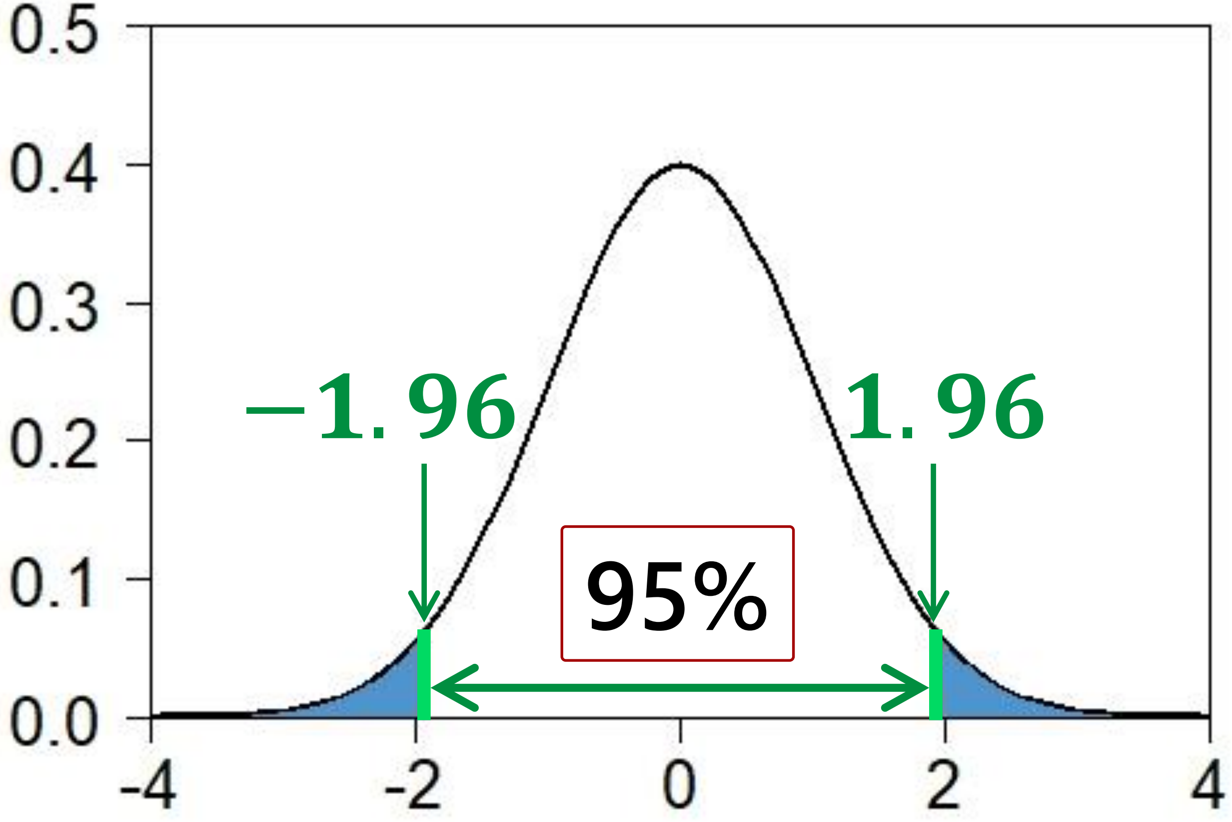
母分散$\sigma^2 = 5^2$

信頼度95%で成り立つ
母平均の信頼区間

$$160.2cm \le \mu \le 179.8cm$$

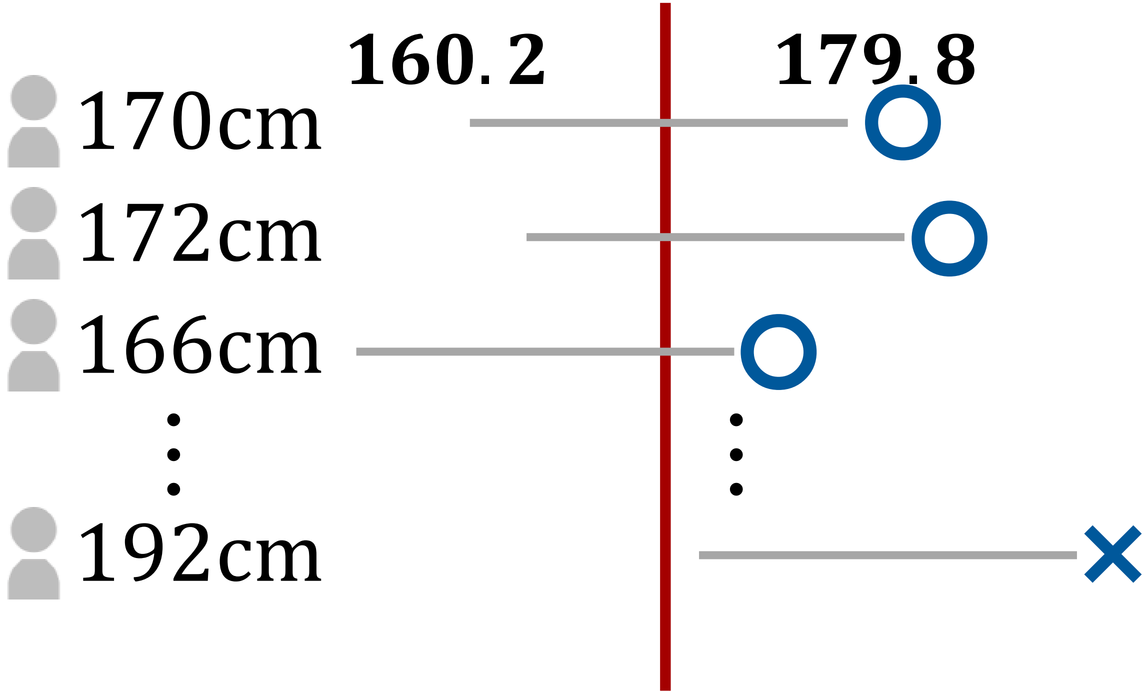
信頼度と信頼区間

信頼度：標本を100回抽出したとき確率的に何回成り立つか



$$160.2cm \leq \mu \leq 179.8cm$$

信頼度95%の信頼区間

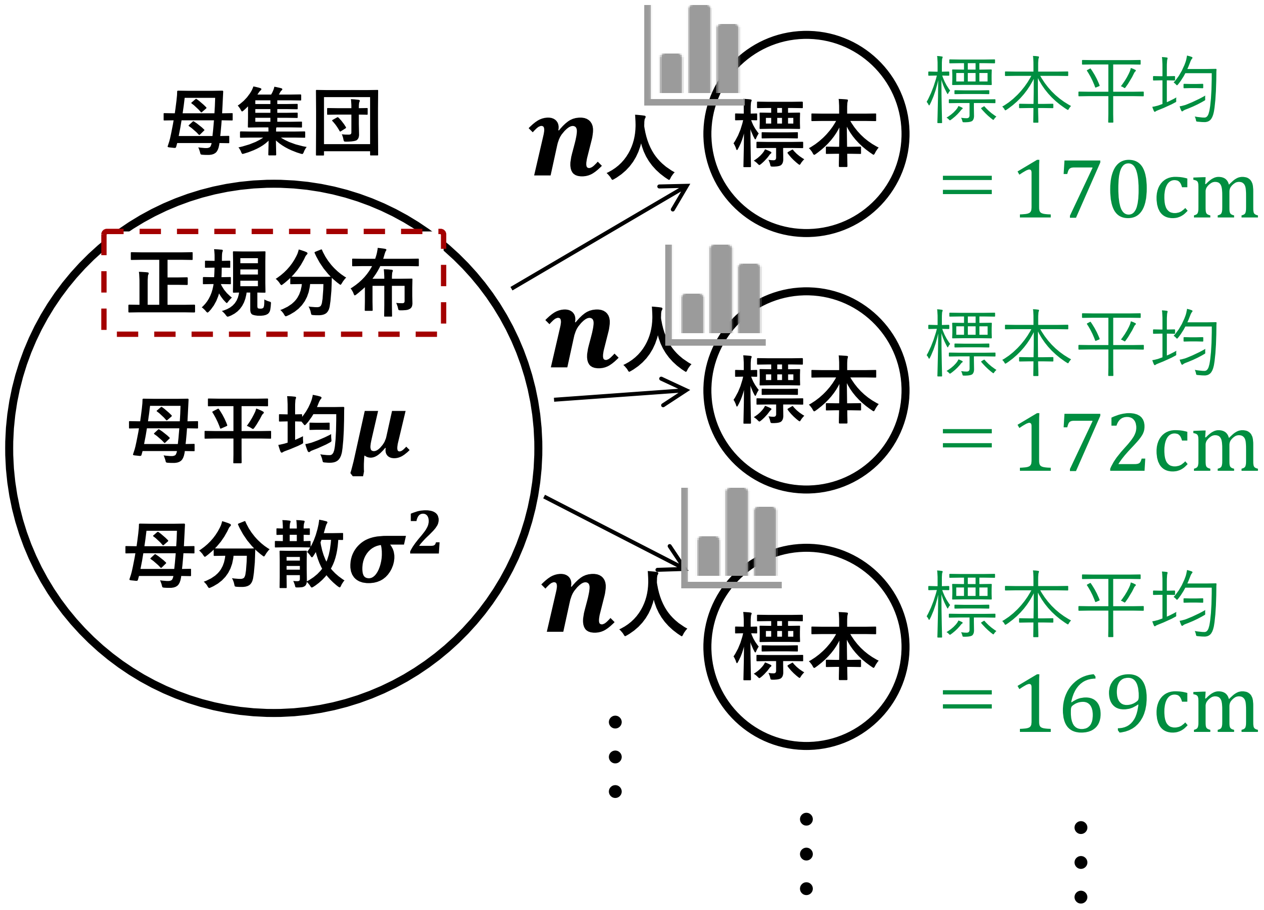


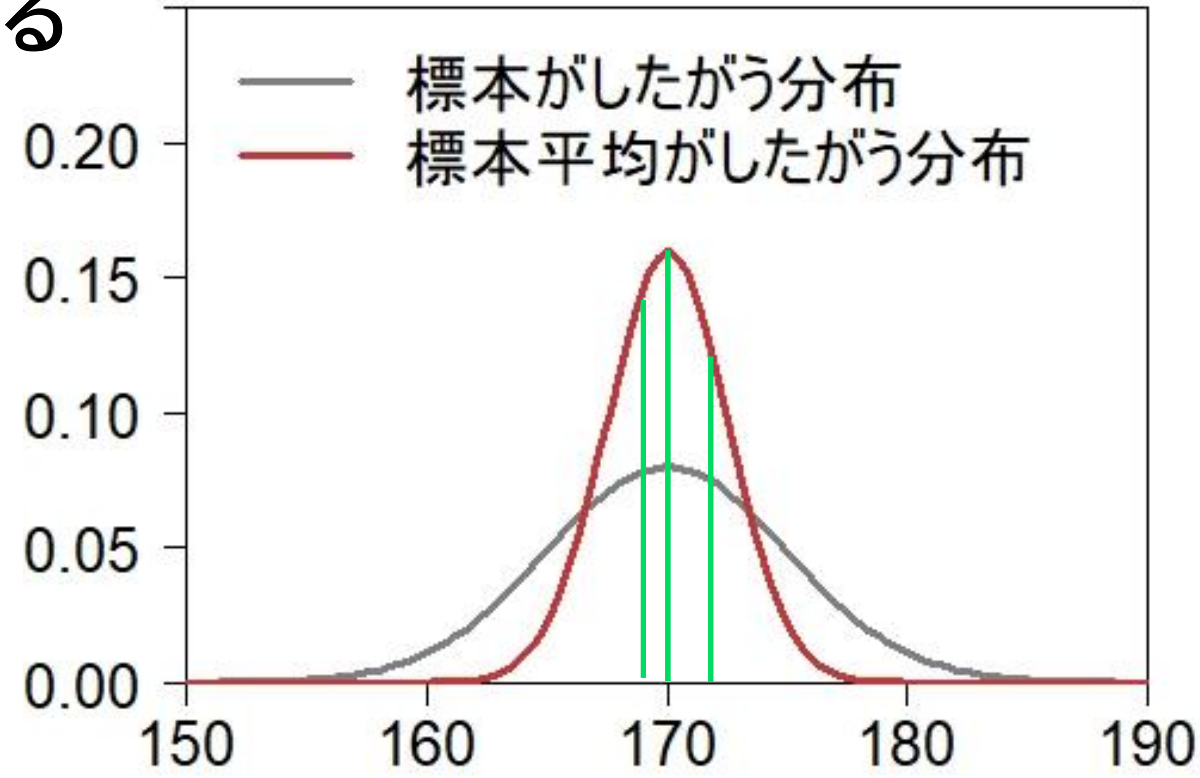
$$\mu = 170cm$$

100回標本を抽出したら
95回は成り立つ区間

標本平均の性質

平均をとることでばらつきを抑制できる





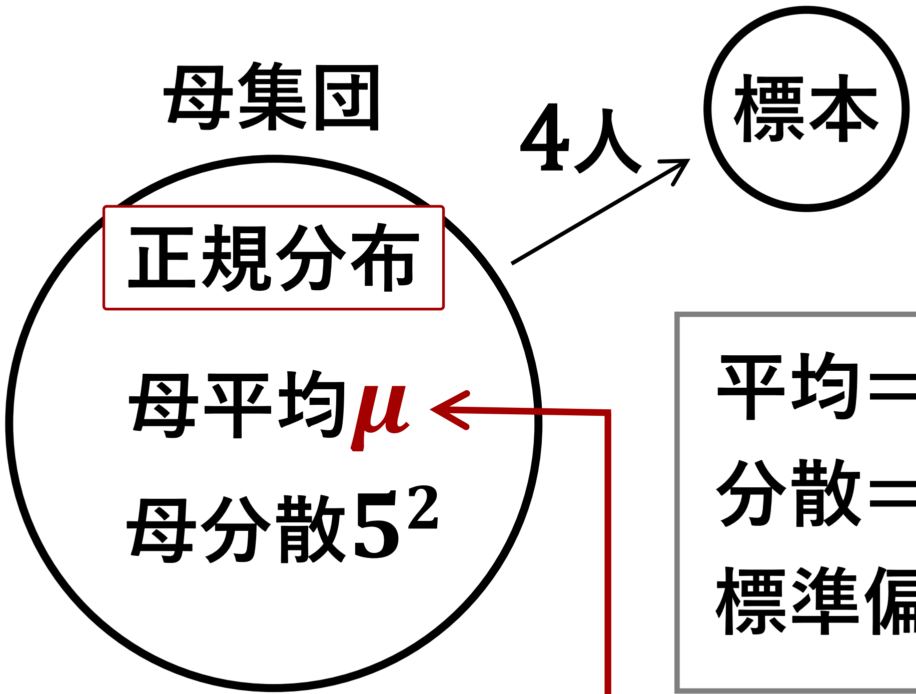
標本平均の平均＝ μ

標本平均の分散＝ σ^2/n

標本平均の標準偏差＝ $\sqrt{\sigma^2/n}$

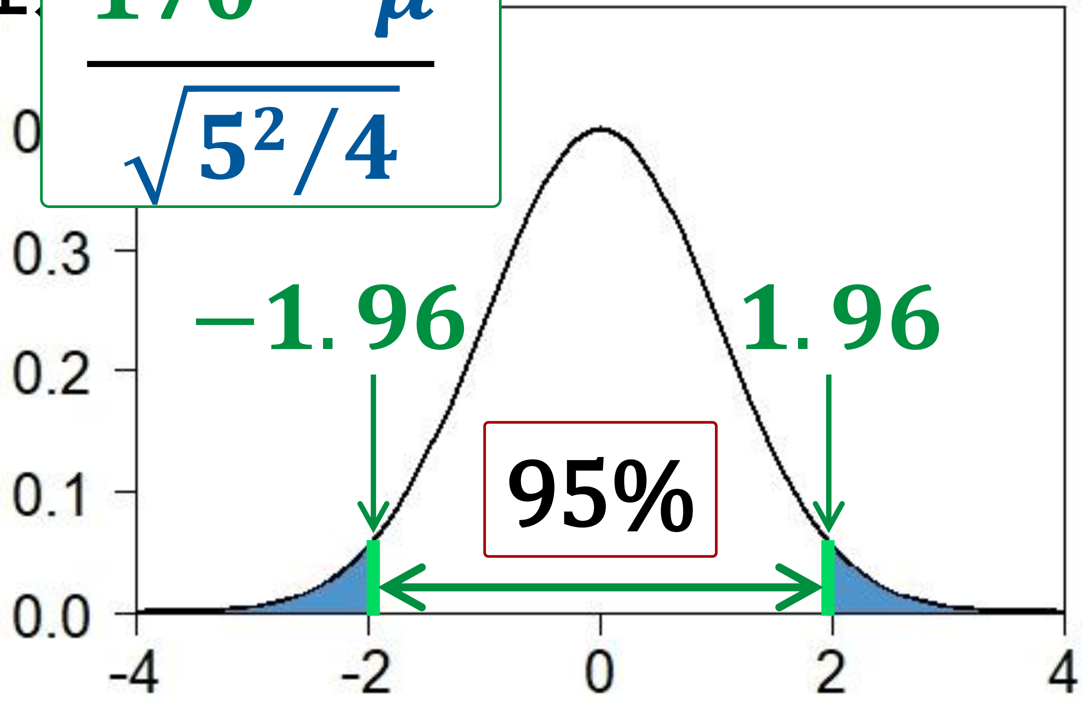
標本平均を用いた区間推定

標本平均の性質を利用して母平均を推定



平均＝ μ
分散＝ $5^2/4$
標準偏差＝ $\sqrt{5^2/4}$

$$\frac{170-\mu}{\sqrt{5^2/4}}$$



$$-1.96 \leq \frac{170-\mu}{\sqrt{5^2/4}} \leq 1.96$$

$$165.1cm \leq \mu \leq 174.9cm$$

分散と信頼区間

分散（標準偏差）が小さい方が信頼区間を狭められる

1人のデータ

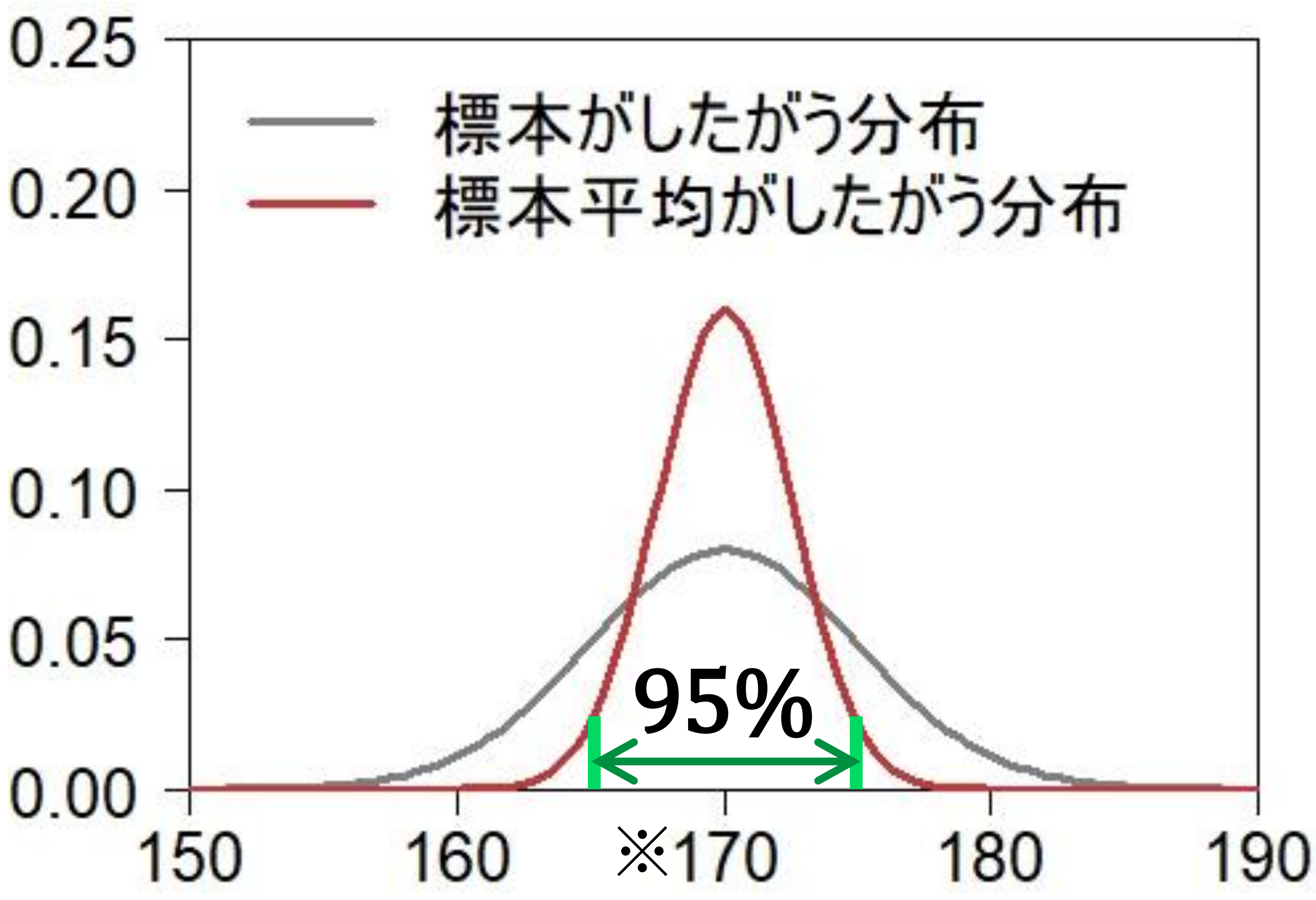
分散＝ 5^2

$$160.2cm \leq \mu \leq 179.8cm$$

4人の「平均」

「平均」の分散＝ $5^2/4$

$$165.1cm \leq \mu \leq 174.9cm$$



※仮にμ=170としたとき

セクション5：区間推定②

t 分布

母分散を不偏分散で代用すると標本平均は「 t 分布 」にしたがう

母集団

正規分布

母平均μ

母分散σ^2 ?

n=4人

標本

標本平均＝170cm

不偏分散＝5^2

標本平均の性質

標本平均の平均＝ μ

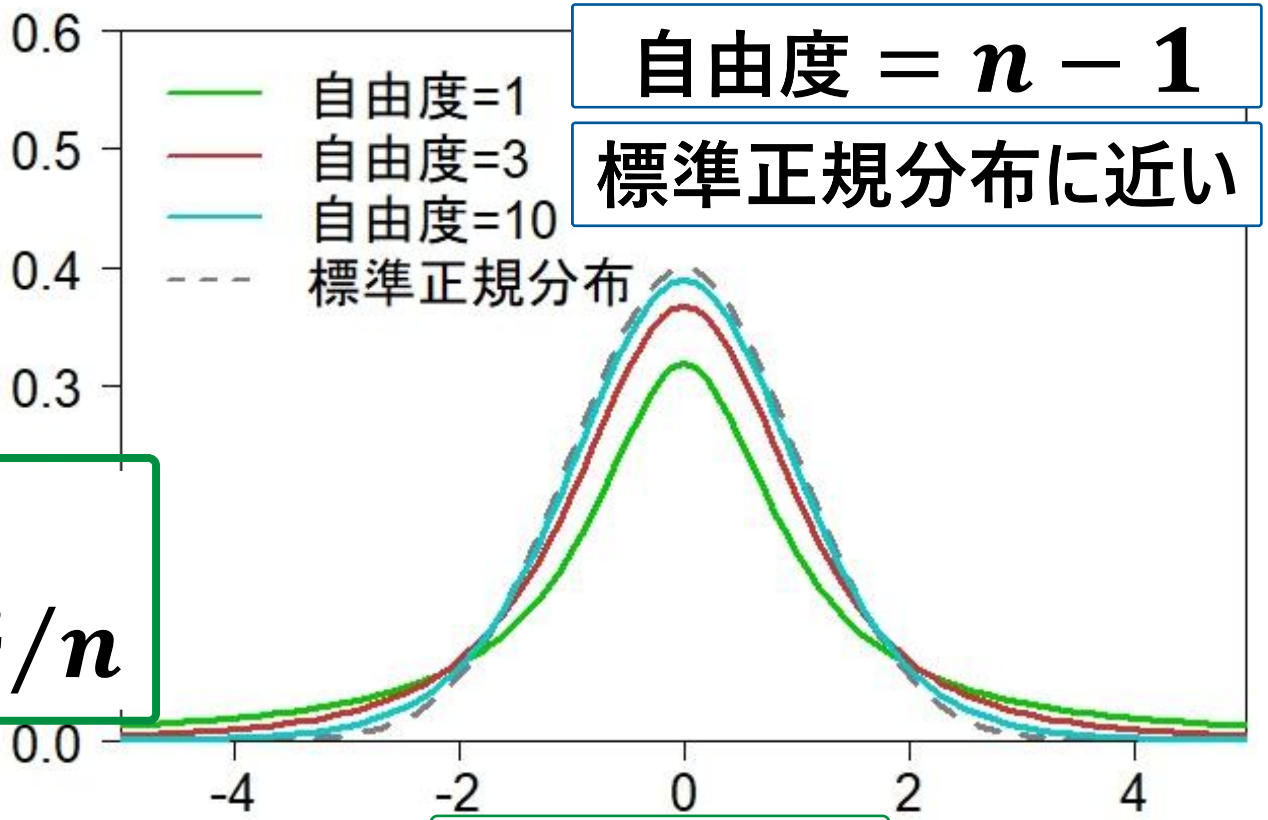
標本平均の分散＝ σ^2/n

$$\frac{170-\mu}{\sqrt{\sigma^2/4}}$$

不偏分散で代用

$$\frac{170-\mu}{\sqrt{5^2/4}}$$

t 分布にしたがう



自由度 ＝ $n-1$

標準正規分布に近い

t 分布による区間推定

母平均を t 分布から区間推定

標本平均＝170cm

不偏分散＝5^2

標本

n=4人

母集団

正規分布

母平均μ

母分散σ^2 ?

標本平均の性質

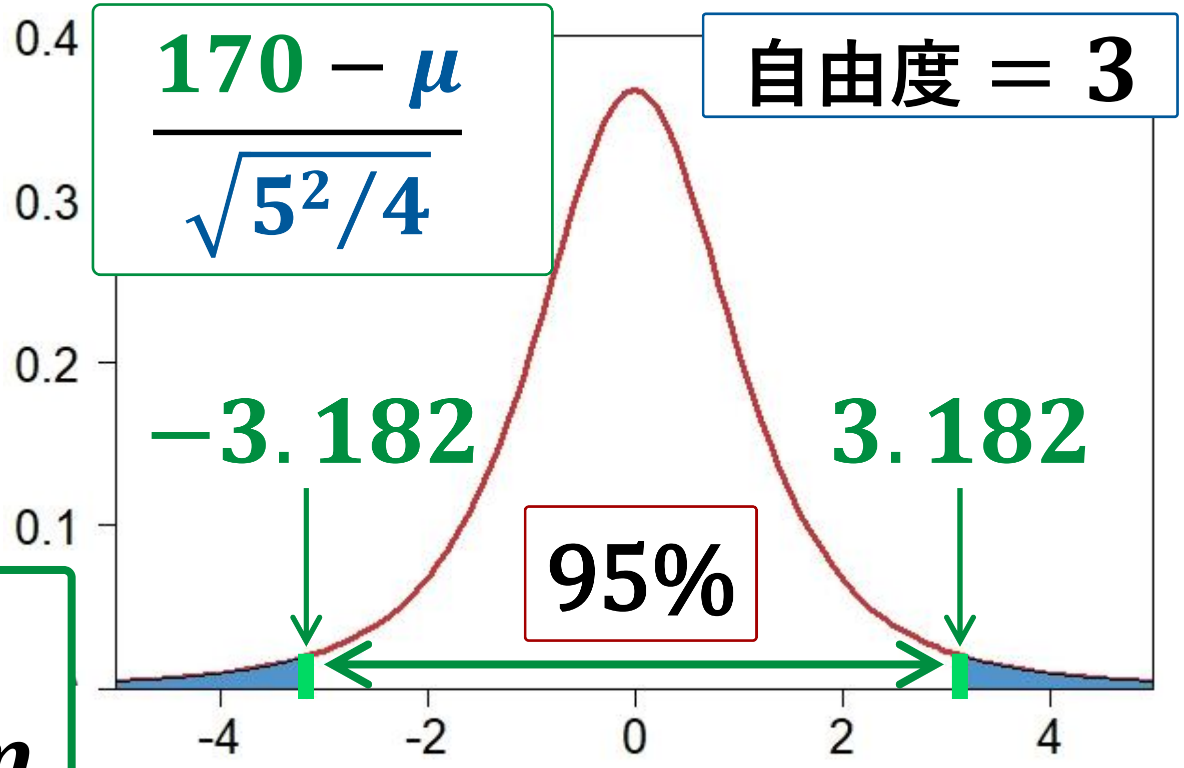
標本平均の平均＝ μ

標本平均の分散＝ σ^2/n

不偏分散で代用 → $\dfrac{170-\mu}{\sqrt{5^2/4}}$

自由度 ＝ 3

$\dfrac{170-\mu}{\sqrt{5^2/4}}$

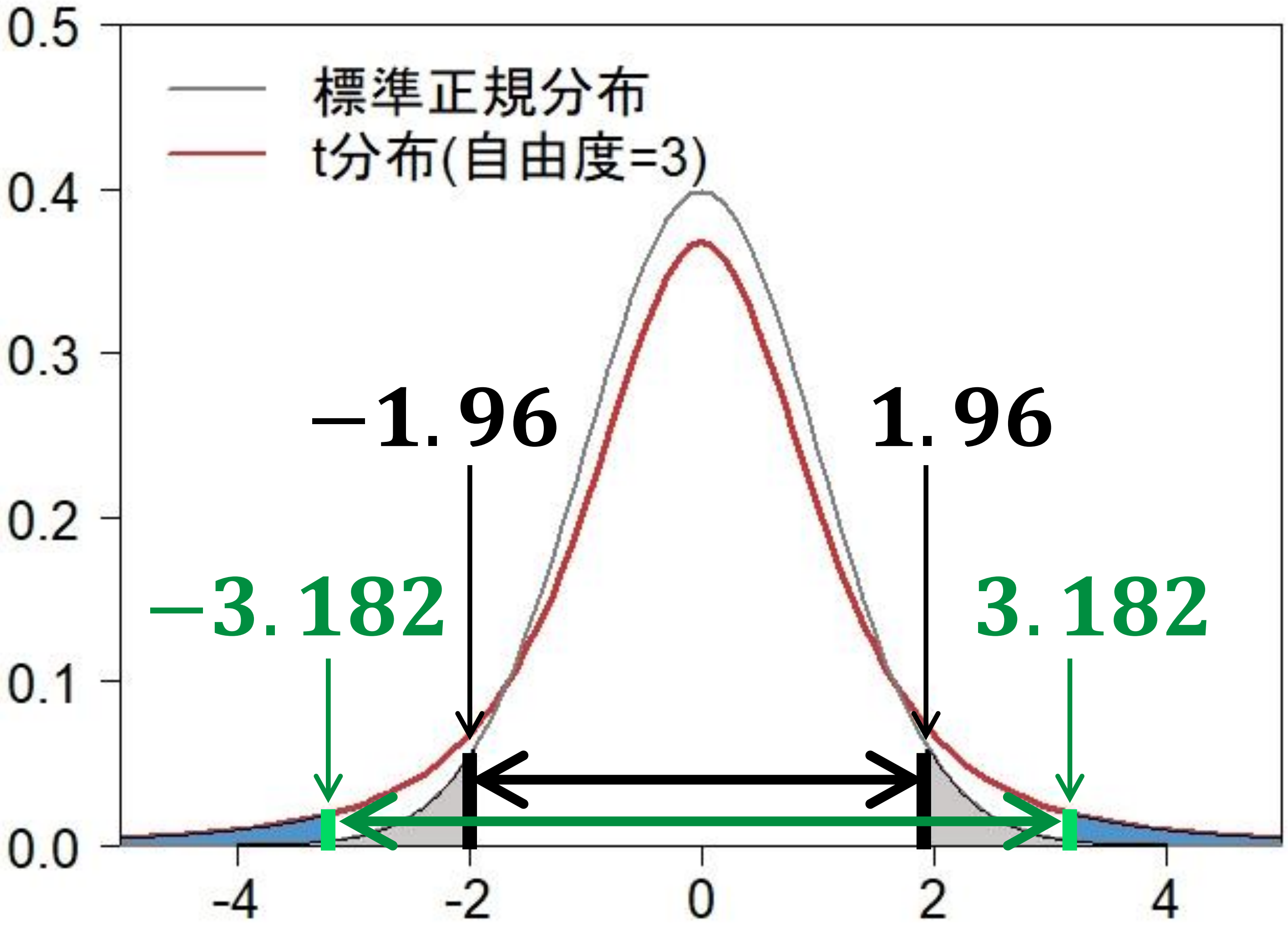


$$-3.182 \le \frac{170-\mu}{\sqrt{5^2/4}} \le 3.182$$

$$162.0cm \le \mu \le 178.0cm$$

t 分布による区間推定の特徴

t 分布は正規分布より裾が広がるため推定区間も広がる



母分散が既知

母分散$= 5^2$

$$165.1cm \leq \mu \leq 174.9cm$$

母分散が未知

不偏分散$= 5^2$で代用
（t 分布にしたがう）

$$162.0cm \leq \mu \leq 178.0cm$$

母集団分布が未知の場合

母集団分布が正規分布でない場合にどうするか

母集団

正規分布 → ?

母平均μ

母分散σ^2 → ?

4人

標本

標本平均＝170cm

不偏分散＝5^2

不偏分散で代用
（ t 分布で推定）

標本平均の性質

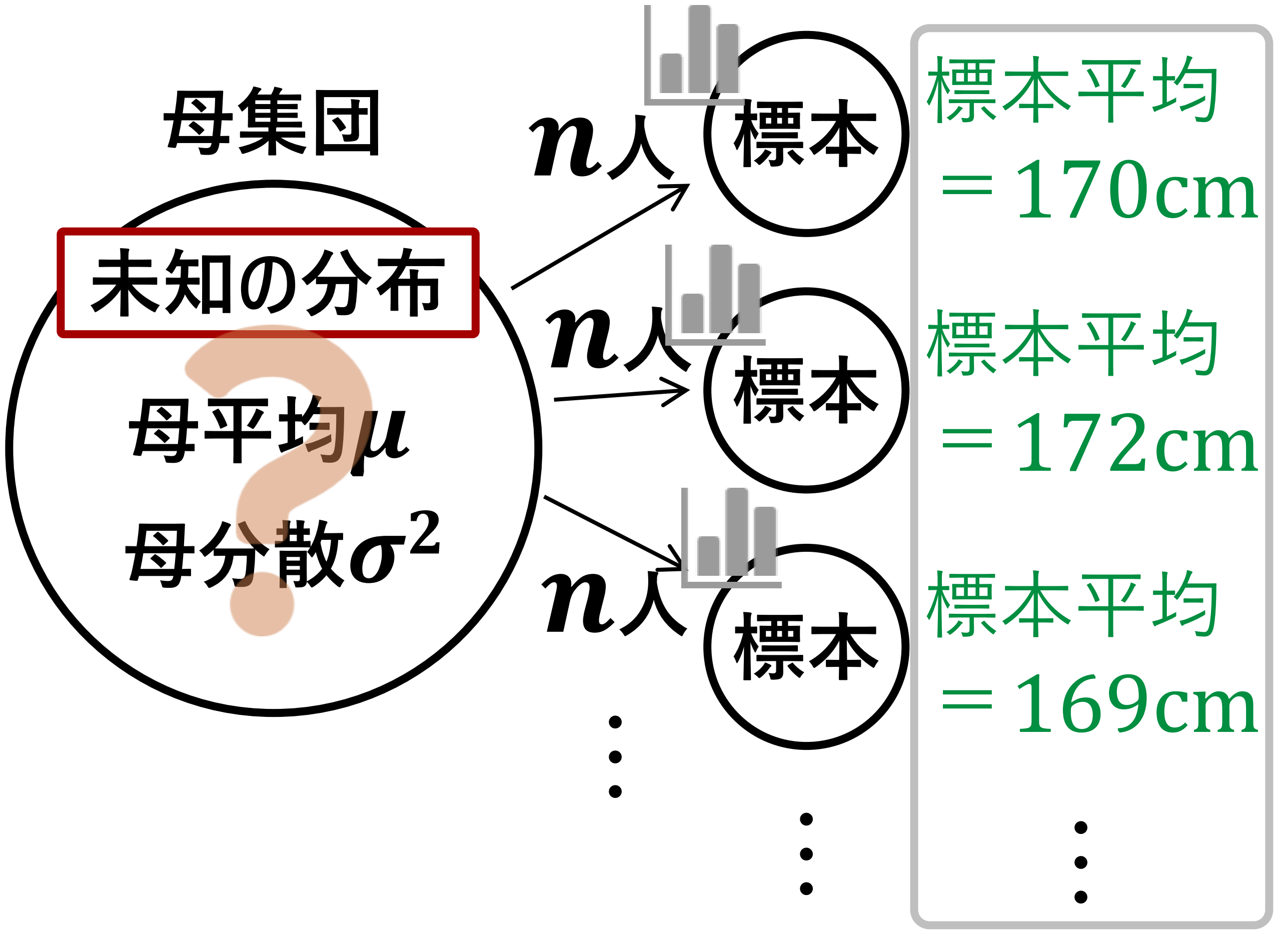
平均 $= \mu$

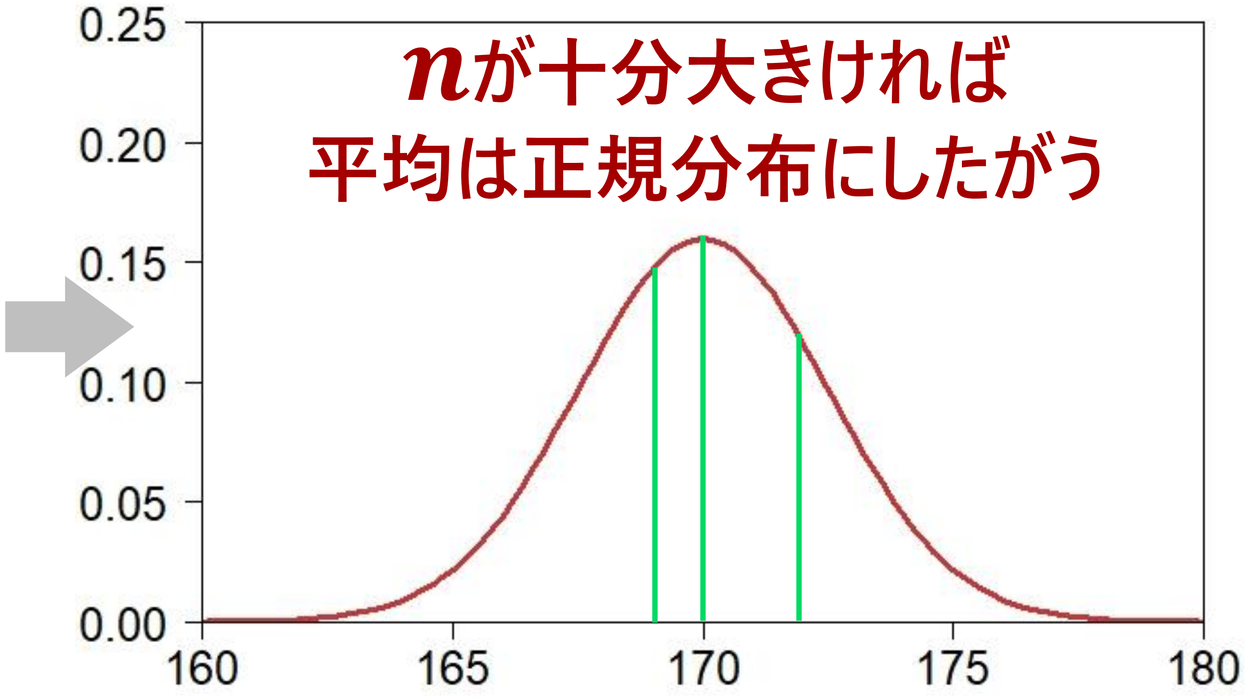
分散 $= \sigma^2/4$

$5^2/4$

中心極限定理

標本が十分大きければ「平均」は正規分布にしたがう





※概ね$n = 30$以上あれば安心

中心極限定理を利用した区間推定

標本が十分大きければ正規分布を使える

n=30人

標本

標本平均＝170cm

不偏分散＝$\boxed{5^2}$

母集団

未知の分布

母平均μ

母分散σ^2

？

標本平均の性質

平均＝ μ

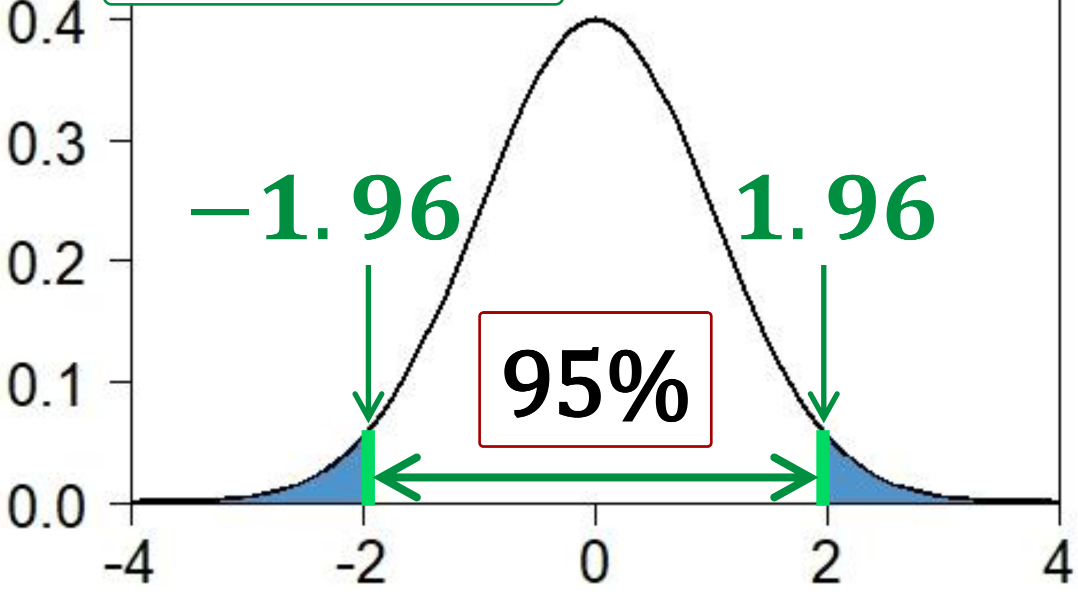
分散＝ $\sigma^2/30$

正規分布
（中心極限定理）

不偏分散で代用
（nが十分大きいから）

$$\frac{170-\mu}{\sqrt{5^2/30}}$$

$$\frac{170-\mu}{\sqrt{5^2/30}}$$



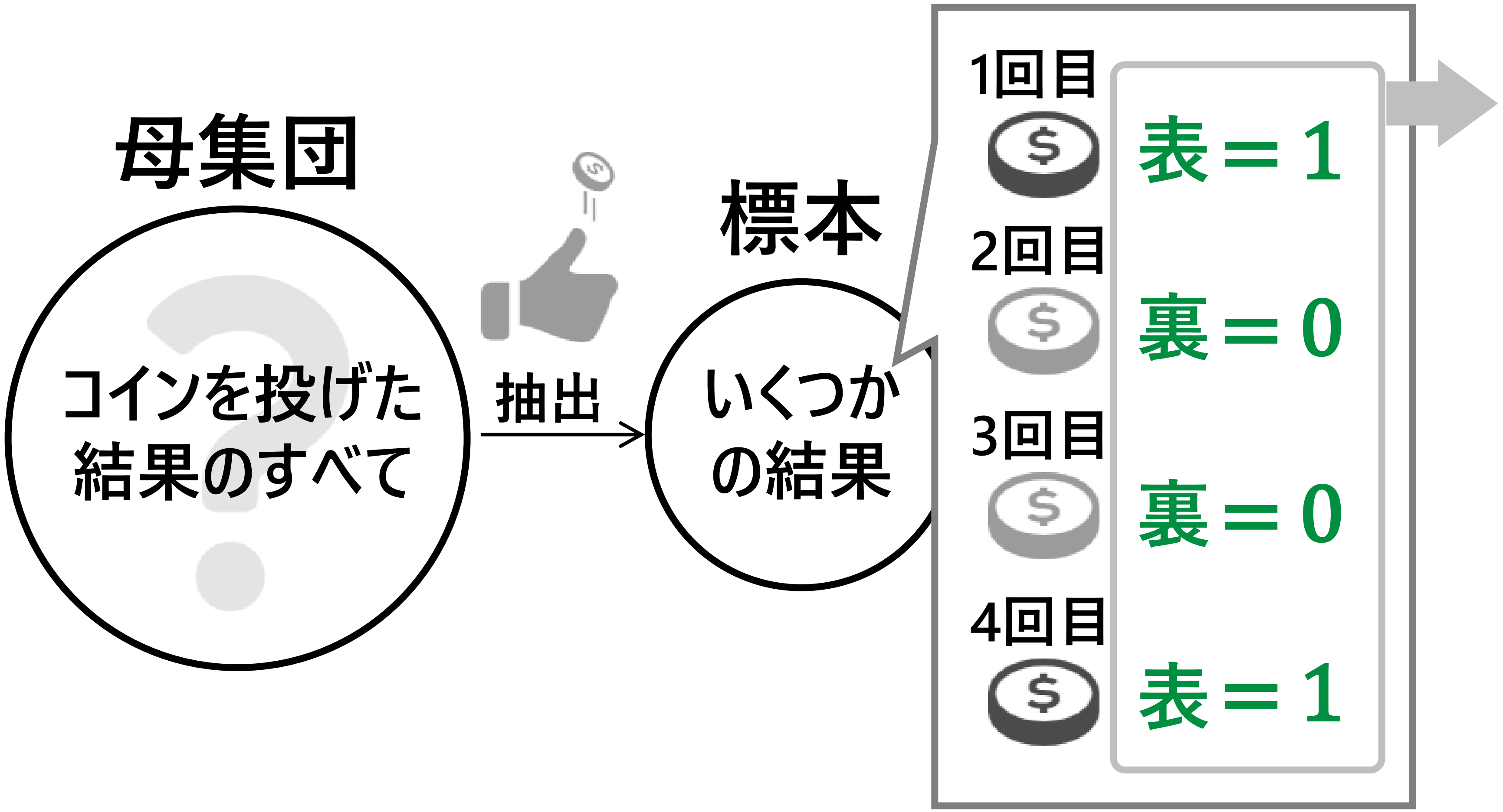
$$-1.96 \le \frac{170-\mu}{\sqrt{5^2/30}} \le 1.96$$

$$168.2cm \le \mu \le 171.8cm$$

セクション6：区間推定③

二値変数と比率

比率は二値変数の平均と言い換えることができる



二値変数

結果が 2 通りのみ

比率

二値変数の**平均**

例: 表が出る比率

$$= \frac{1 + 0 + 0 + 1}{4}$$

$$= 50\%$$

ベルヌーイ分布

二値変数はベルヌーイ分布にしたがう

標本

二値変数

1回目　表 = 1

2回目　裏 = 0

3回目　裏 = 0

⋮

母集団

コインを投げた結果のすべて

抽出

いくつか
の結果

母比率 = $\boldsymbol{p}$
（表が出る比率）

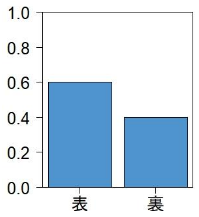
ベルヌーイ分布

二値変数は

平均 = p

分散 = $p(1-p)$

のベルヌーイ分布にしたがう



平均の性質と**中心極限定理**

二値変数の比率(平均)は

平均 = p

分散 = $p(1-p)/n$

の**正規分布**にしたがう

母比率の区間推定

平均の性質と中心極限定理を使う

n=400人

標本

抽出

支持者160人

標本比率=0.4

母集団

全有権者の支持/不支持

母比率$= p$

母分散$= p(1-p)$

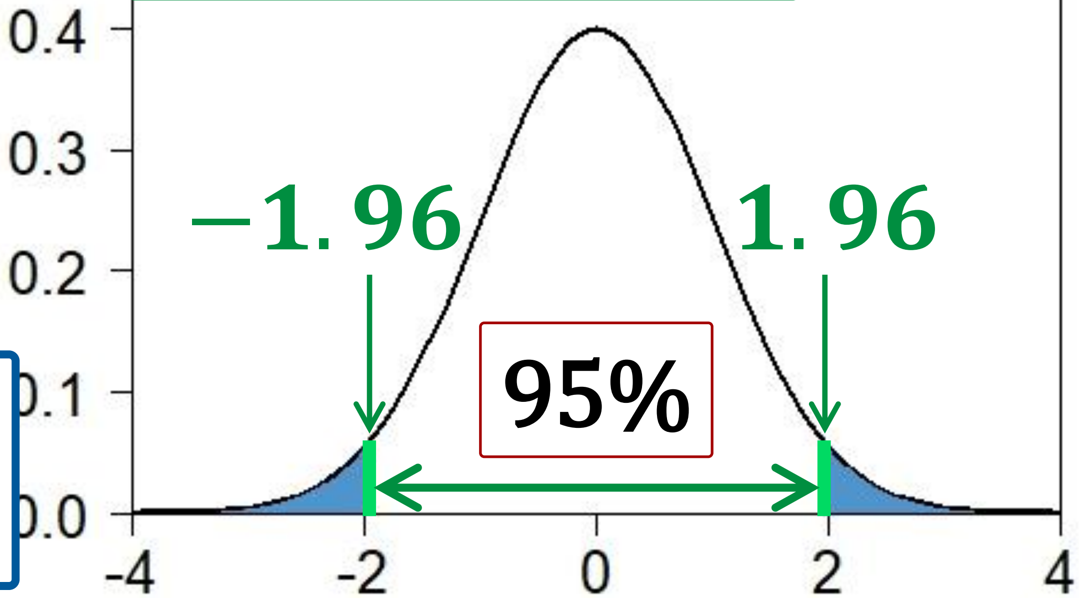
標本比率(平均)の性質

平均$= p$

分散$= p(1-p)/400$

正規分布
（中心極限定理）

$$\frac{0.4-p}{\sqrt{p(1-p)/400}}$$



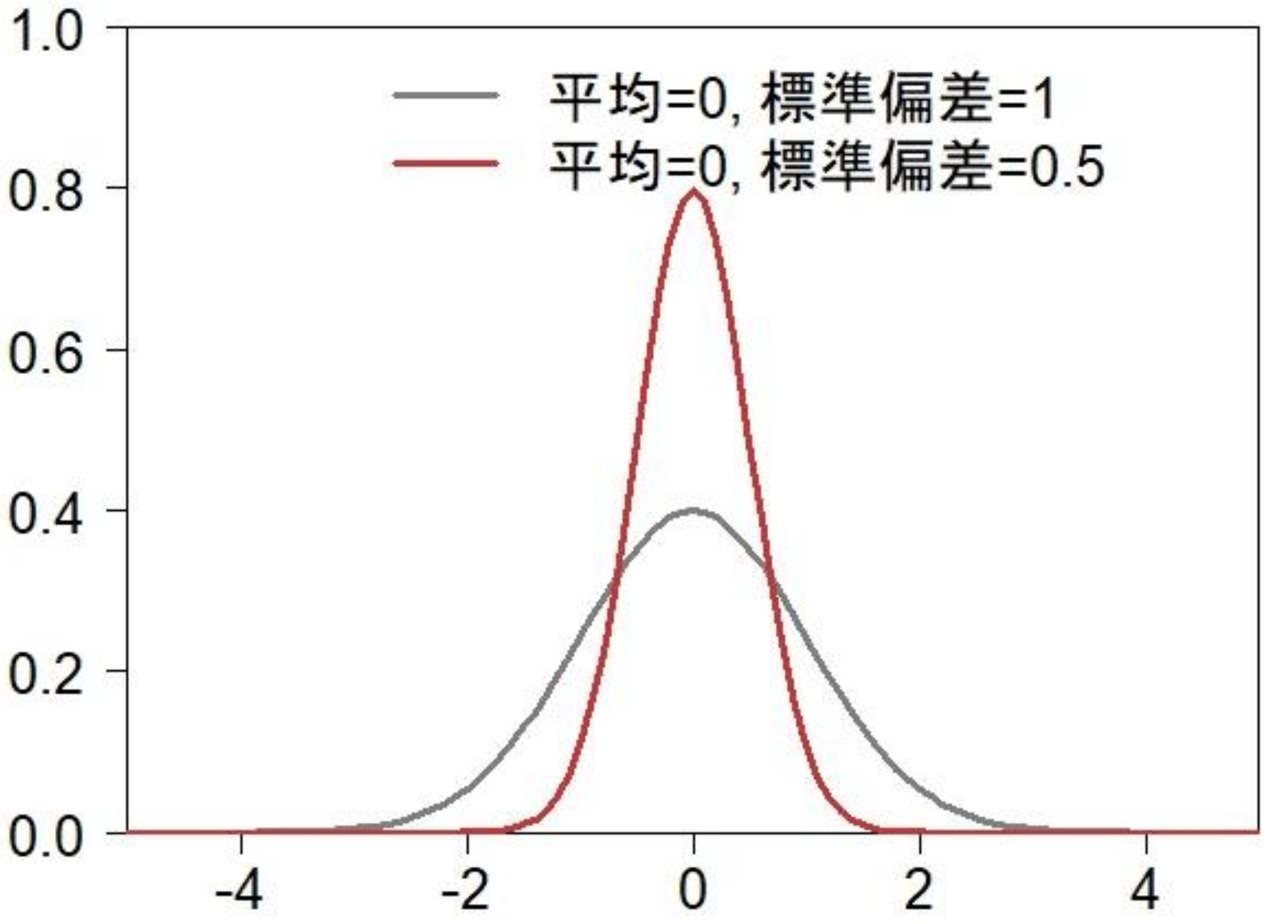
$$-1.96 \le \frac{0.4-p}{\sqrt{p(1-p)/400}} \le 1.96$$

標本比率0.4で代用

$$0.352 \le p \le 0.448$$

母分散とは

母分散も問いの対象となりうる



母分散を知りたいケース

- ✓ 製造物の寸法や品質
- ✓ 従業員の残業時間
- ✓ 店舗の混雑状況
- ✓ 株価　etc.

カイ二乗分布

正規分布にしたがう確率変数の「ばらつき」がしたがう分布

母集団
（ある製品の寸法誤差）

正規分布

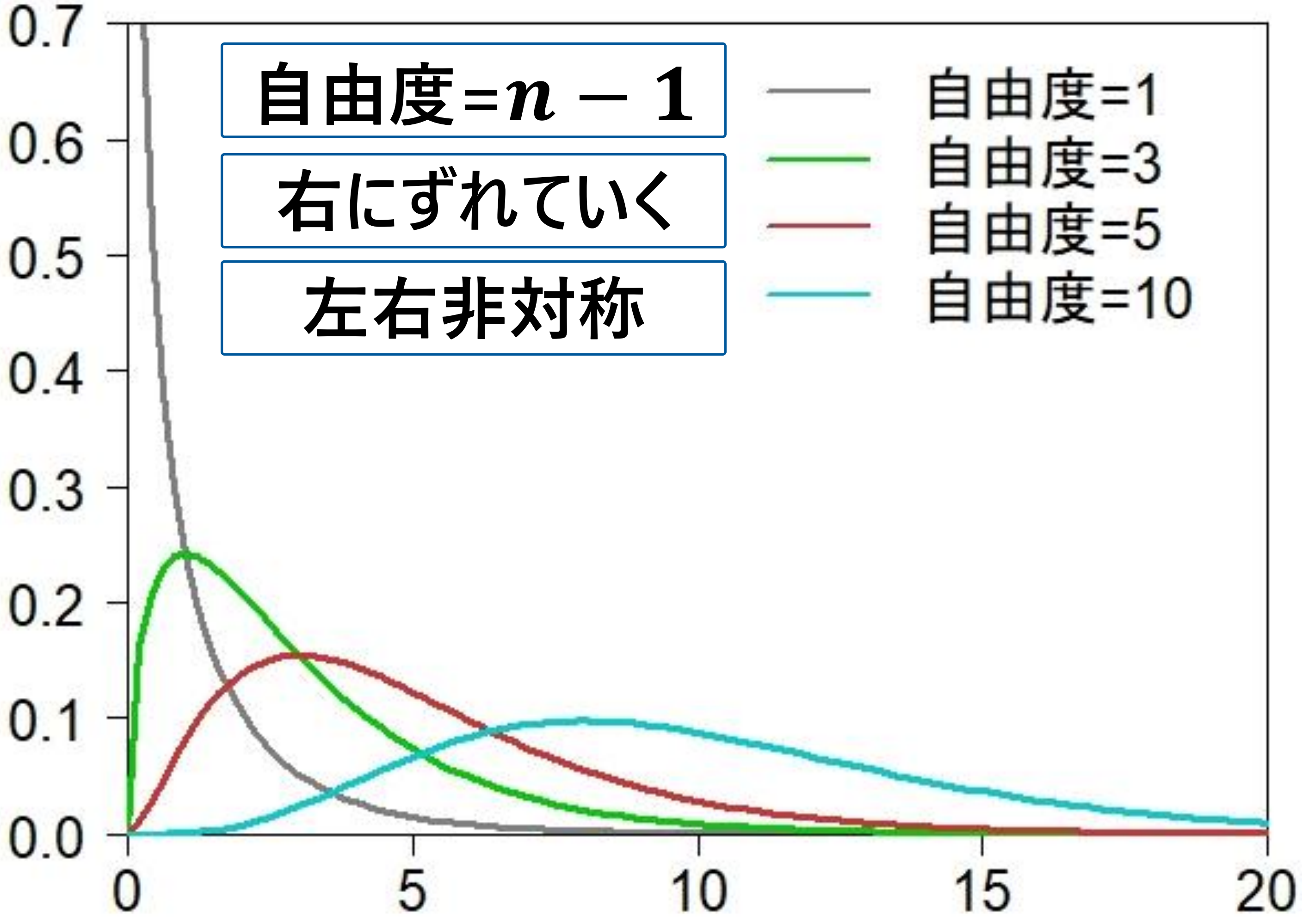
母平均μ
母分散σ^2

n個 → 標本

標本平均
不偏分散

カイ二乗

$$\chi^2 = \frac{(n-1) \times \text{不偏分散}}{\text{母分散}\sigma^2}$$



母分散の区間推定

カイ二乗分布を使って母分散を区間推定する

$$\frac{(10-1)\times 0.2^2}{\sigma^2}$$

母集団
（ある製品の寸法誤差）

正規分布

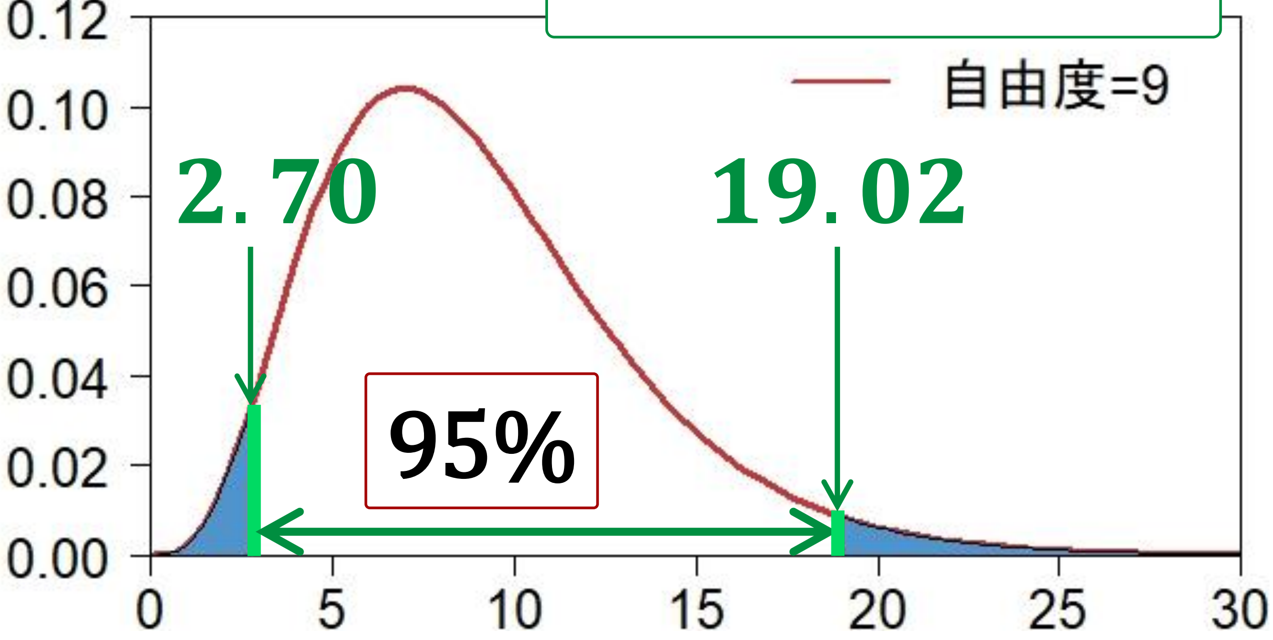
母平均μ
母分散σ^2

10個 → 標本

標本平均=0cm
不偏分散=0.2^2

カイ二乗

$$\chi^2 = \frac{(n-1)\times \text{不偏分散}}{\text{母分散}\sigma^2}$$



$$2.7 \le \frac{(10-1)\times 0.2^2}{\sigma^2} \le 19.02$$

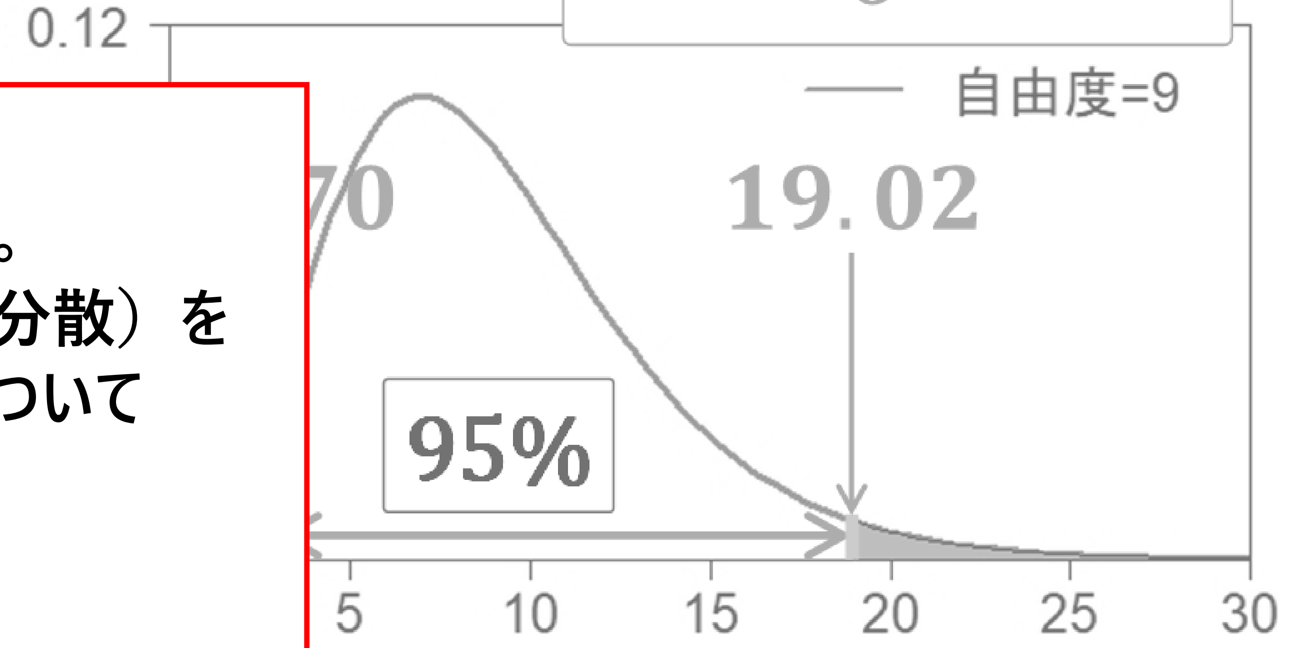
$$0.1332 \le \sigma^2 \le 0.3652$$

母分散の区間推定 修正版

カイ二乗分布を使って母分散を区間推定する

$$\frac{(10-1)\times 0.2^2}{\sigma^2}$$

母集団



動画内スライドに誤記がございました。
正しくは「$0.0189 \leq \sigma^2 \leq 0.1333$」となります。
また、動画内（4:00～4:30あたり）でσ^2（母分散）を平方根をとって標準偏差に計算しなおした値について講師が「0.36から0.6」とコメントしておりますが、
正しくは「0.1376から0.3651」となります。
申し訳ありません。よろしくお願いいたします。

カイ二乗

$$\chi^2 = \frac{(n-1)\times \text{不偏分散}}{\text{母分散}\sigma^2}$$

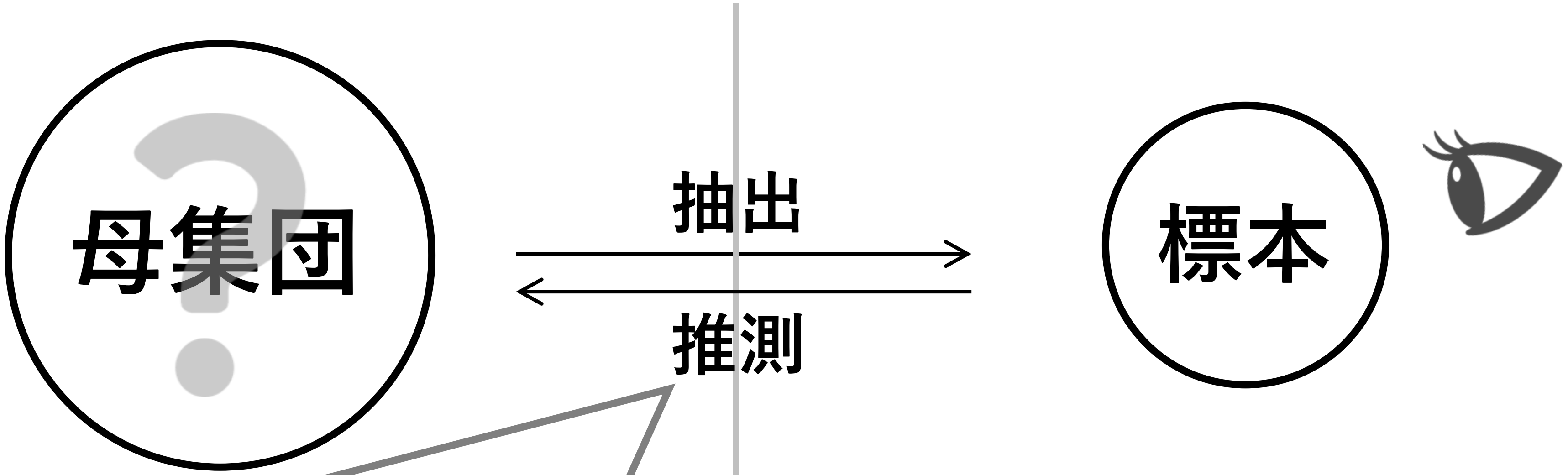
$$2.7 \leq \frac{(10-1)\times 0.2^2}{\sigma^2} \leq 19.02$$

修正箇所 → $0.0189 \leq \sigma^2 \leq 0.1333$

セクション7：検定①

検定とは

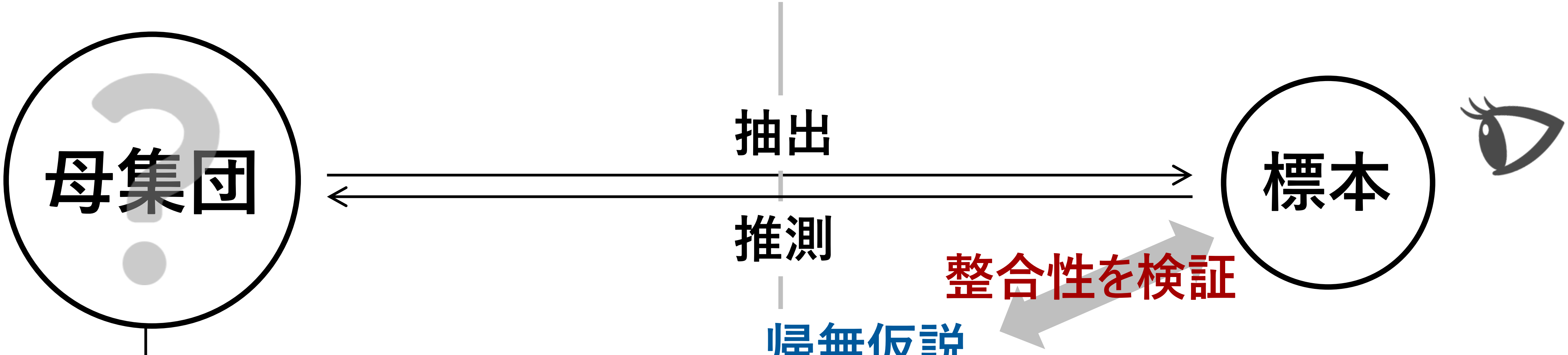
統計的な「お作法」で確率的に論じる



- ✓ 母集団についての仮説を設定する
- ✓ 仮説と標本を見比べて確率論で検証する
- ✓ 仮説についての結論を導く
- ✓ 計算の考え方は区間推定と同様

帰無仮説と対立仮説

否定したい仮説を棄却することで主張したい仮説を支持する



帰無仮説

否定したい仮説（例: 母集団の平均身長は$165cm$だ）

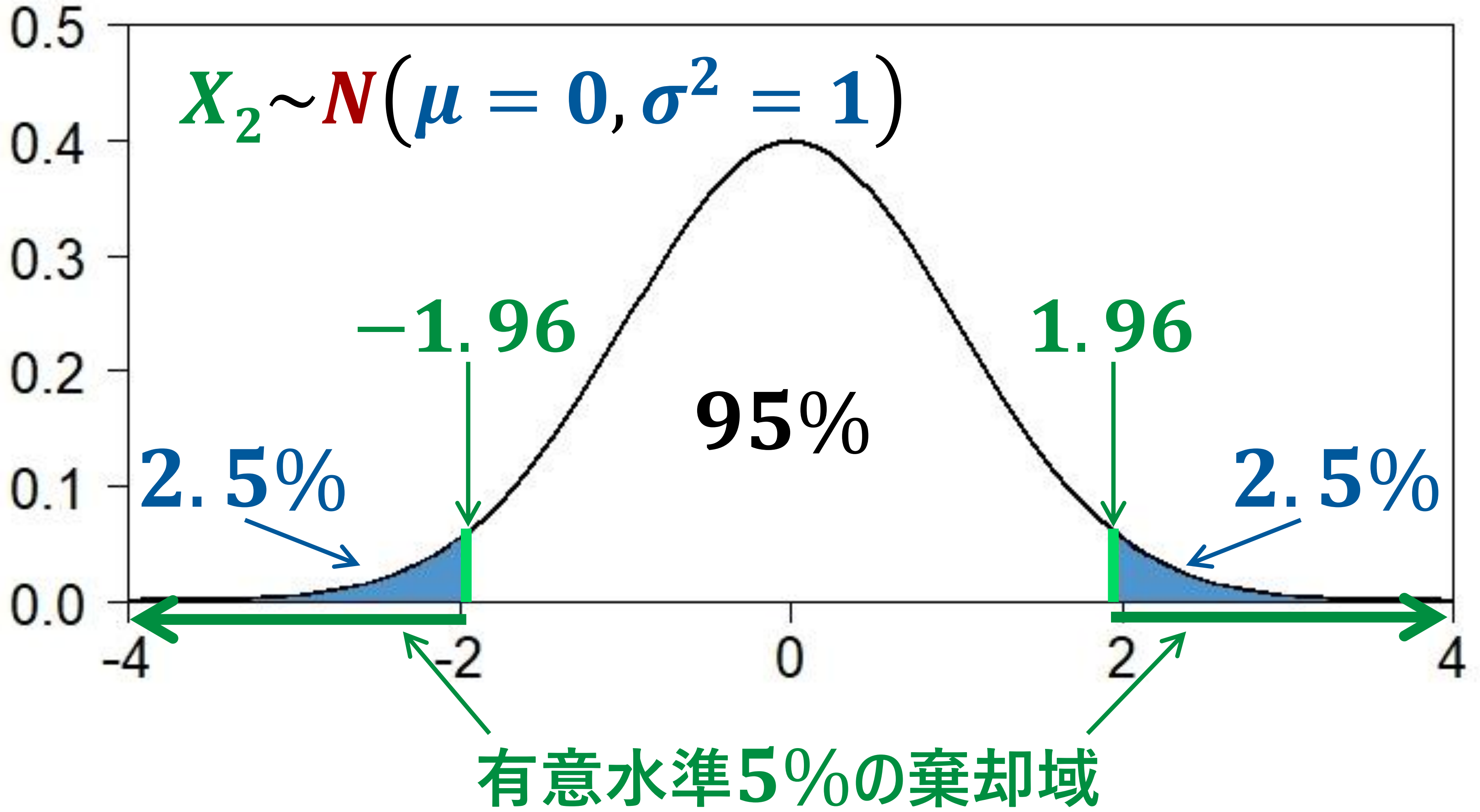
vs.

対立仮説

主張したい仮説（例: 母集団の平均身長は$165cm$ではない）

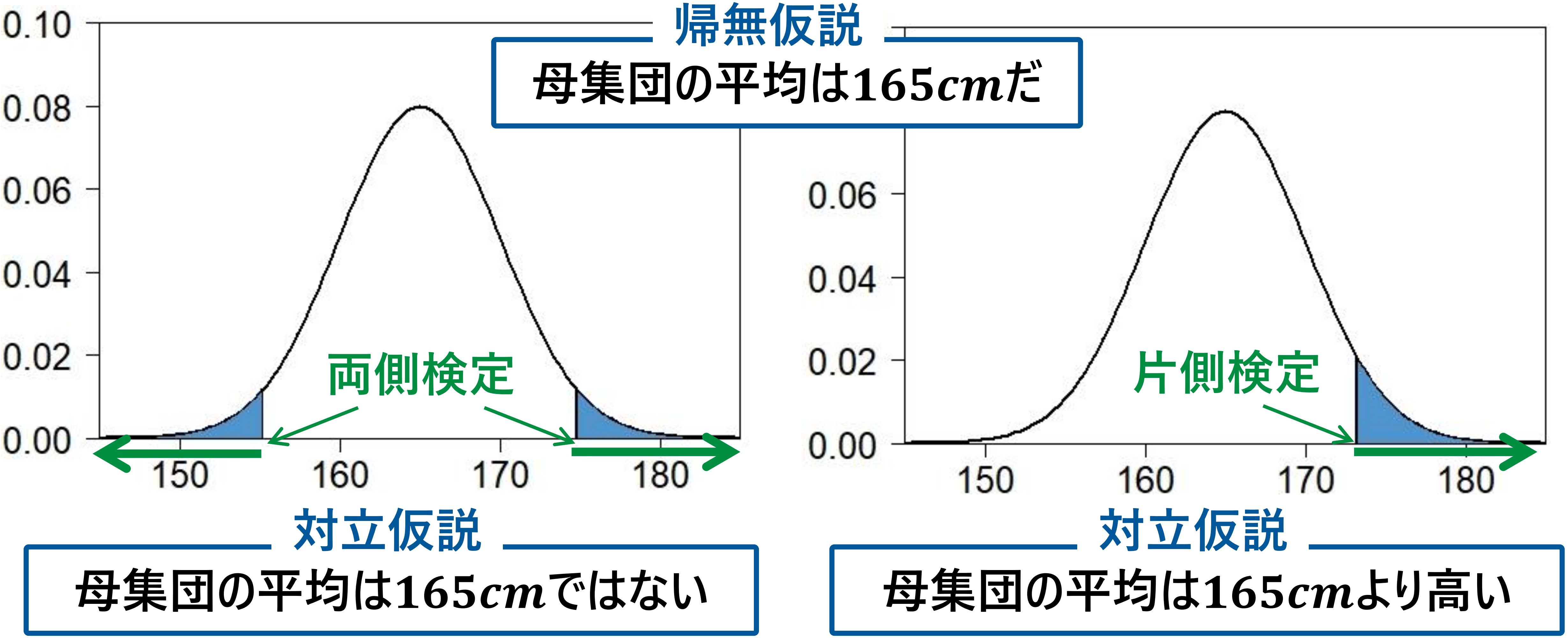
有意水準と棄却域

帰無仮説が正しい場合において「起こったら不自然な領域」を考える



両側検定と片側検定

棄却域は対立仮説によって両側か片側かが決まる



検定の流れ

仮説検定には決められた手順がある

①帰無仮説を設定する

②対立仮説を設定する

③帰無仮説が正しいと仮定する

④有意水準・棄却域の設定

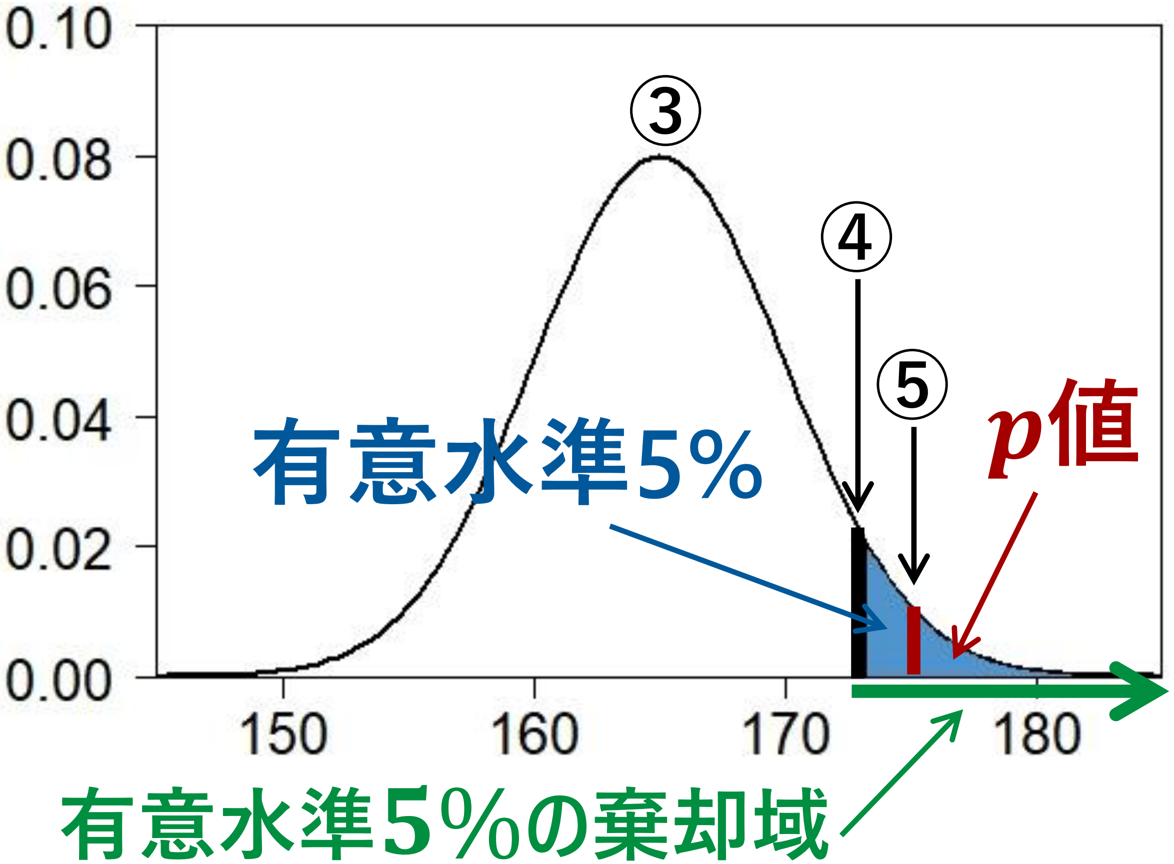
⑤標本から検定統計量を計算

⑥検定統計量と棄却域を見比べる

⑦結論(帰無仮説棄却の判断)

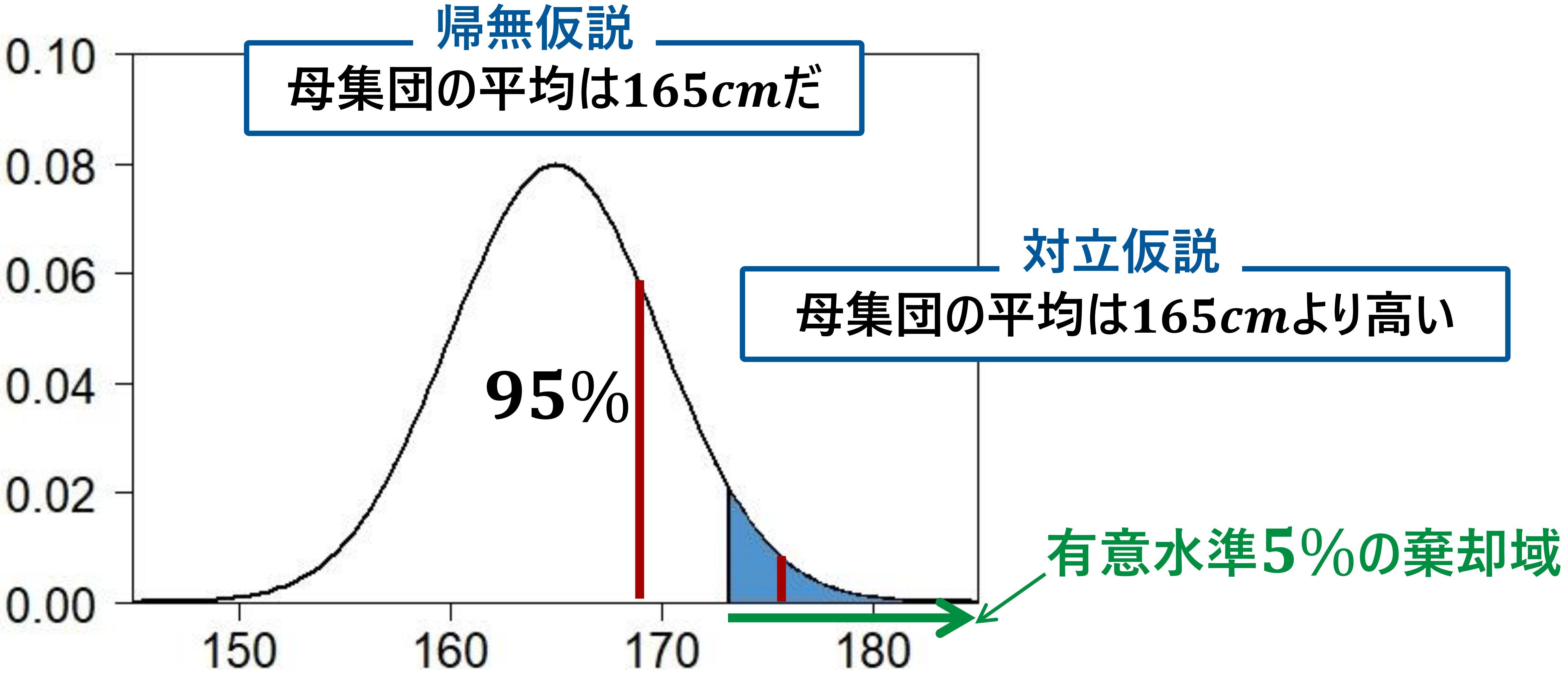
①母集団の平均は165cmだ

②母集団の平均は165cmより高い



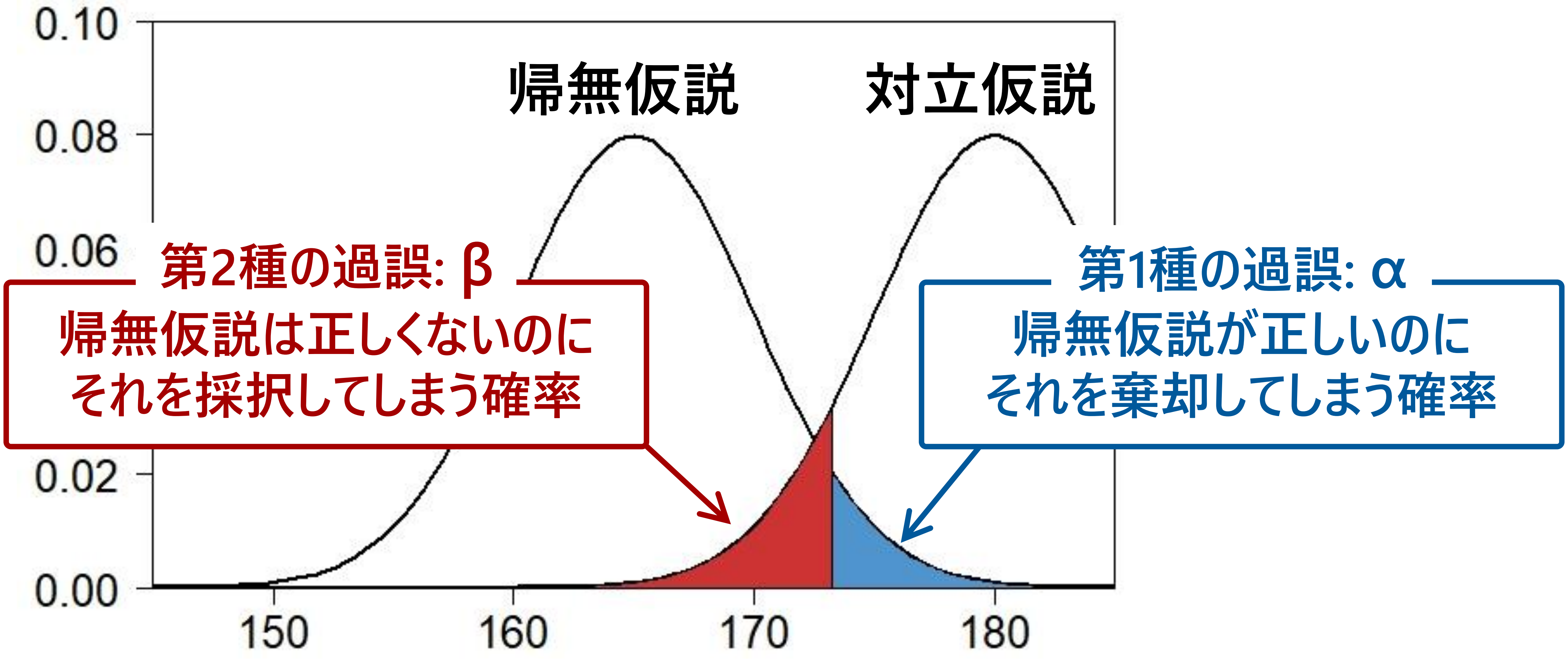
検定結果の解釈

帰無仮説を棄却できるか否か



第１種の過誤、第２種の過誤

あわてて棄却する誤りとぼんやり見過ごす誤り



セクション8：検定②

母平均の検定（t 検定）

t 分布を用いて母平均を検定する

母集団

正規分布

母平均μ

母分散σ^2

n=9人

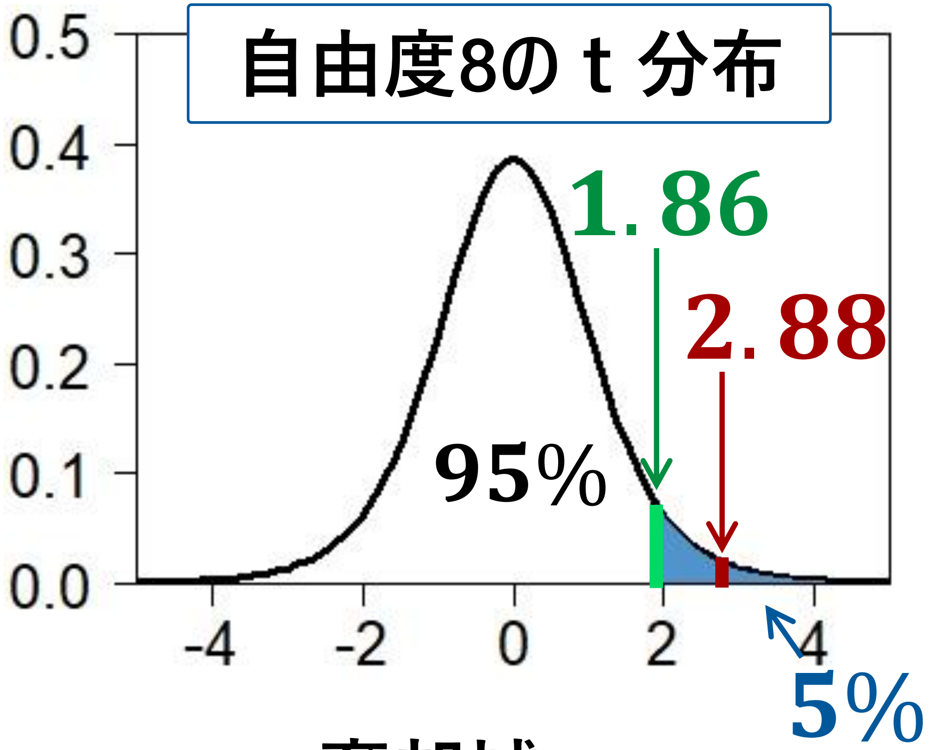
標本

標本平均＝$170cm$

不偏分散＝5.2^2

検定統計量

$$\frac{170-165}{\sqrt{5.2^2/9}} = 2.88$$



帰無仮説: 母平均μは165cmである

有意水準: 片側5%

対立仮説: 母平均μは165cmより高い

棄却域

1.86より大きい領域

結論: 帰無仮説を棄却できる

母平均の差の検定（ウェルチの t 検定）

2つの異なる母集団の平均に差があるかを検定する

A高校男子
正規分布
→ 9人 → 標本 a
標本平均 = $170cm$
不偏分散 = 5.2^2

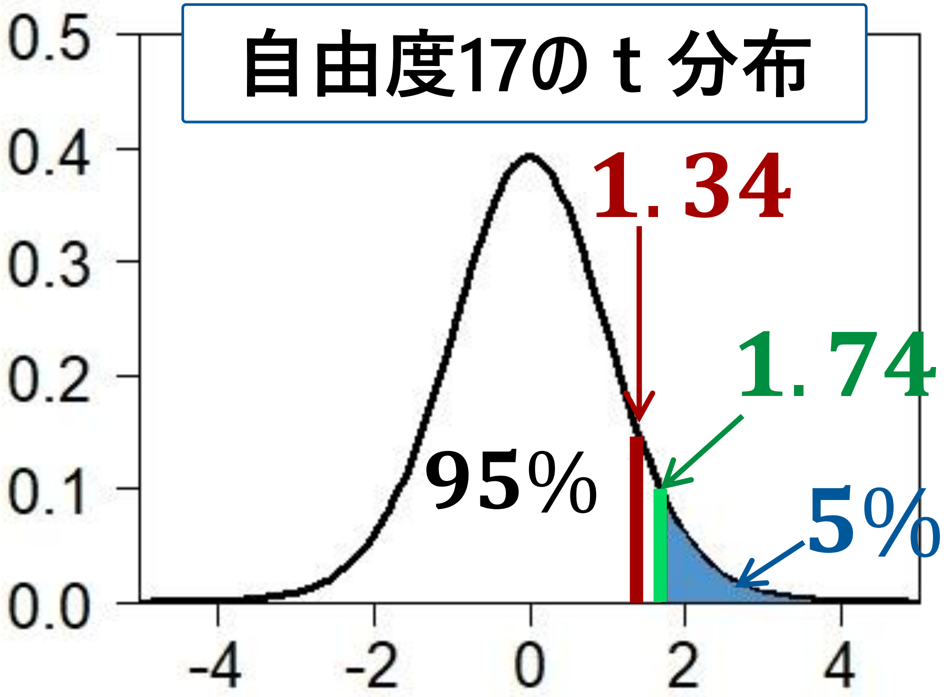
検定統計量 = 1.34

B高校男子
正規分布
→ 12人 → 標本 b
標本平均 = $167cm$
不偏分散 = 4.9^2

帰無仮説: AとBの平均に差はない

有意水準: 片側5%

対立仮説: Aの平均はBの平均より高い

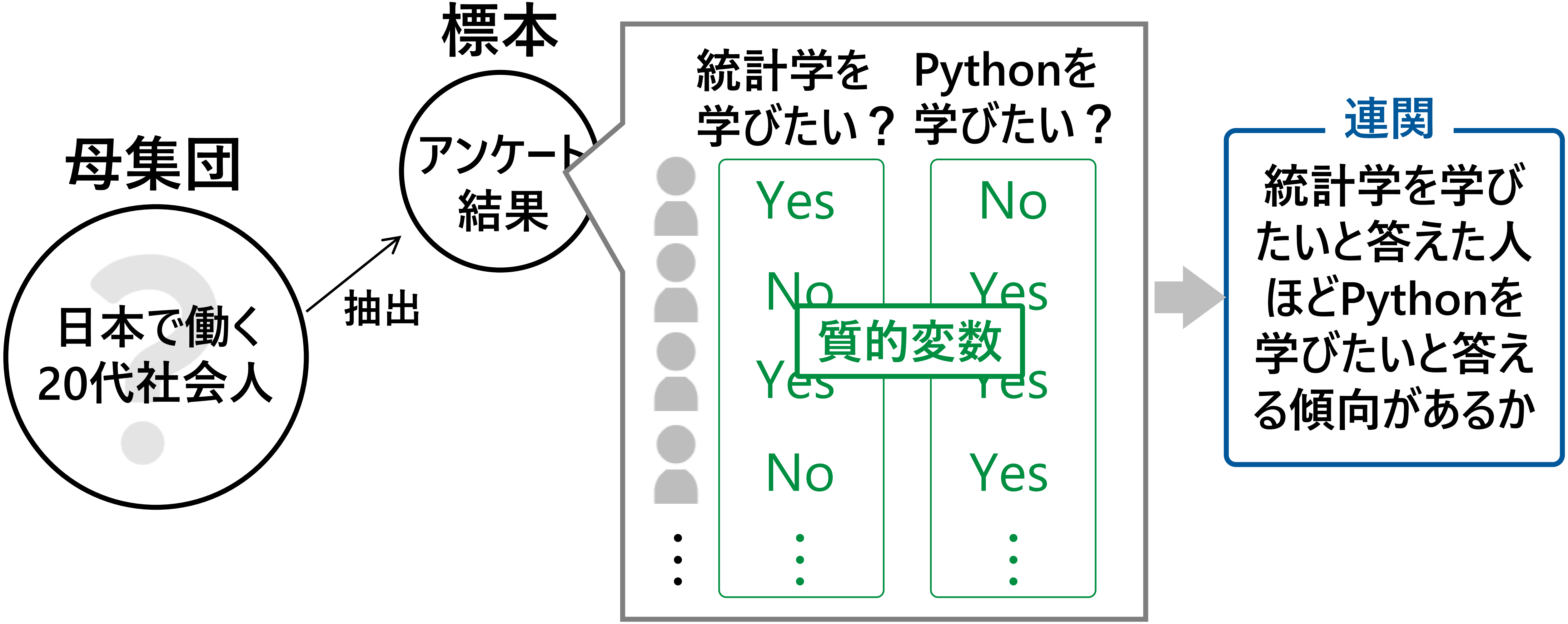


棄却域
1.74より大きい領域

結論: 帰無仮説を棄却できない

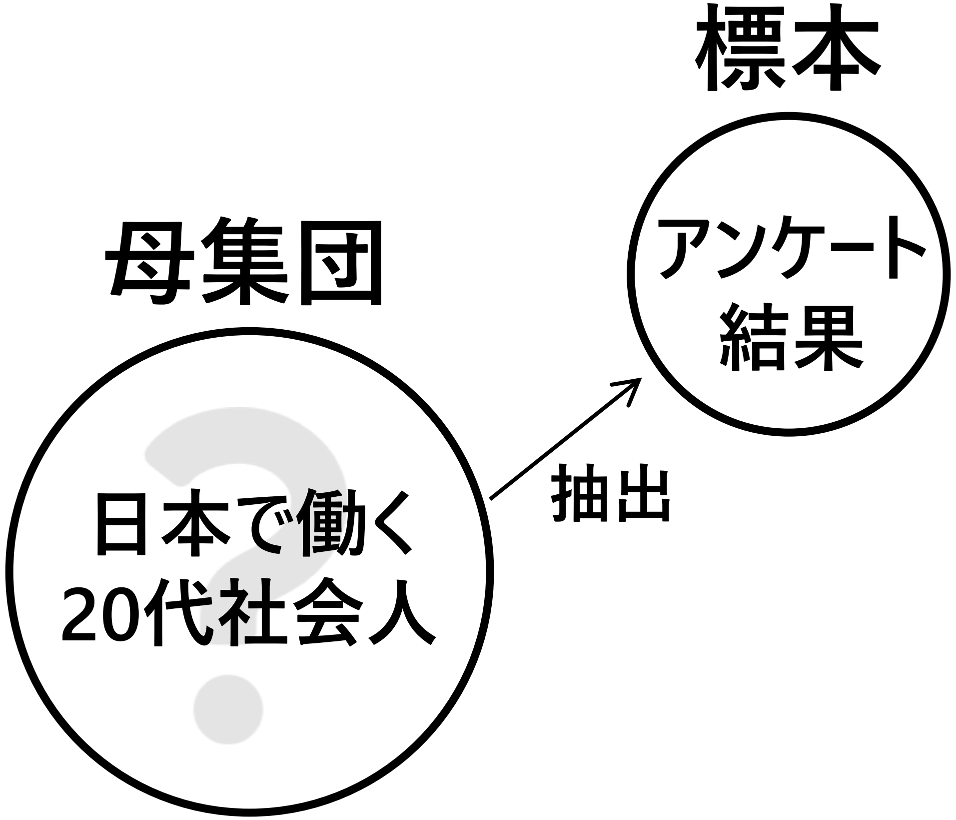
連関

2つの質的変数の間の関連性の有無



クロス集計表

質的変数を集計して表にしたもの



統計学	Python		合計
	はい	いいえ	
はい	7	3	10
いいえ	13	27	40
合計	20	30	50

周辺度数

周辺度数

独立性

２つの質的変数間に何の連関もないと仮定する

標本データ(集計結果)

統計学	Python はい	Python いいえ	合計
はい	7	3	10
いいえ	13	27	40
合計	20	30	50

クロスするセルを一旦忘れる

統計学	Python はい	Python いいえ	合計
はい	[redacted]		10
いいえ			40
合計	20	30	50

合計の比率から数字を1行当て込む

統計学	Python はい	Python いいえ	合計
はい	4	6	10
いいえ	[redacted]		40
合計	20	30	50

もう1行も同様に

独立なときの自然な結果

	Python		
はい	4	6	10
いいえ	16	24	40
合計	20	30	50

カイ二乗検定（独立性の検定）

カイ二乗分布を用いて連関の有無を検定する

標本データ

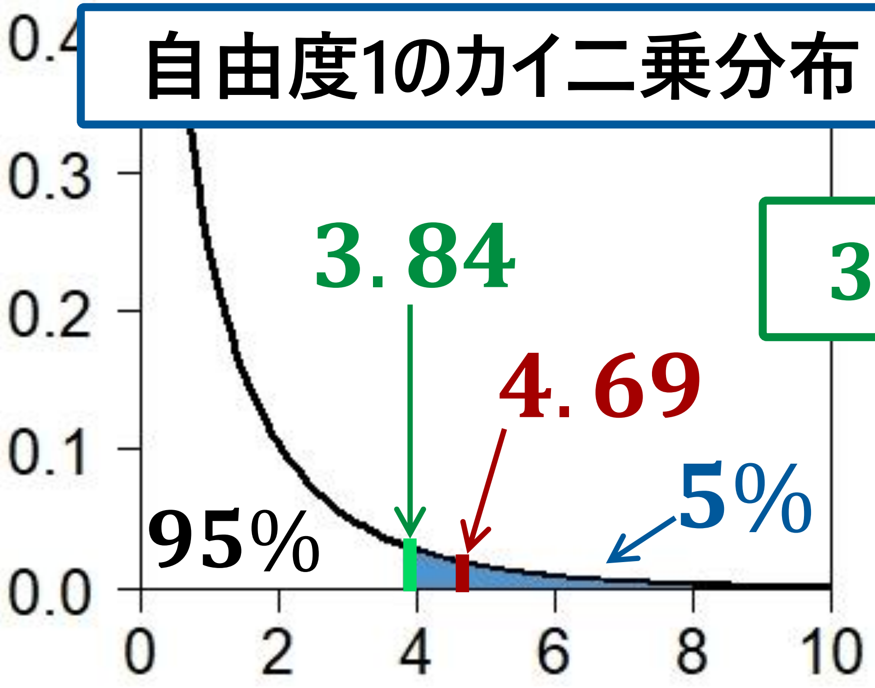
統計学	Python		合計
はい	7	3	10
いいえ	13	27	40
合計	20	30	50

帰無仮説

統計学	Python		合計
はい	4	6	10
いいえ	16	24	40
合計	20	30	50

検定統計量

$$\frac{(7-4)^2}{4}+\cdots+\frac{(27-24)^2}{24}=4.69$$



棄却域

3.84より大きい領域

結論: 帰無仮説を棄却できる

以　上