情報統計
第２回

2022年8月2日　　神奈川工科大学

櫻井 望

国立遺伝学研究所
生命情報・DDBJセンター

統計の基本と用語

学習目標

以下の統計用語をマスターします

- 平均値、中央値
- 分散
- 標準偏差
- 母集団
- ランダムサンプリング
- 標本
- 統計的推定
- 母平均、母分散
- 標本平均、標本分散、不偏標本分散
- 分布
- 正規分布（ガウス分布）
- 標準誤差

統計って？

集団の状況を
数値で表したもの

目的：集団の〇〇を知りたい

統計学

- データを集める
- 解析する
- 解釈する

ための方法論

結果：集団の〇〇がわかった！

第1回の
身長データを使って
解析してみる

目的：このクラスの人の身長はどのくらい？

データ

集団の状況を表す代表的な値を計算

平均値
中央値

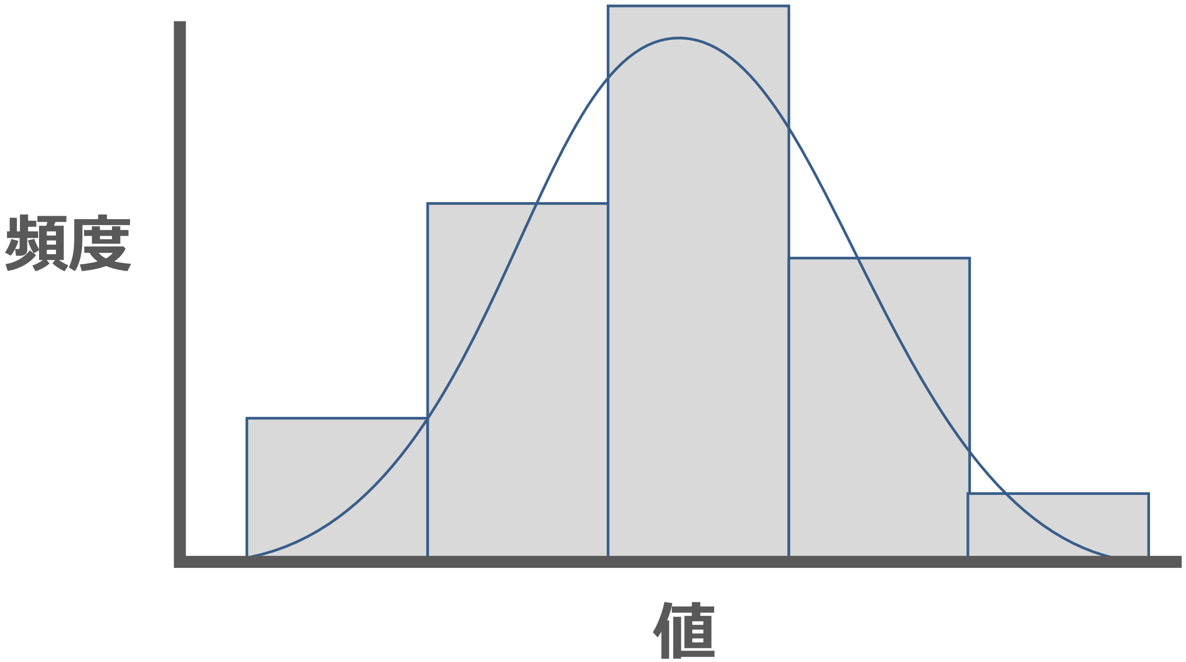
中心を表す値

分散
標準偏差

ばらつきを表す値

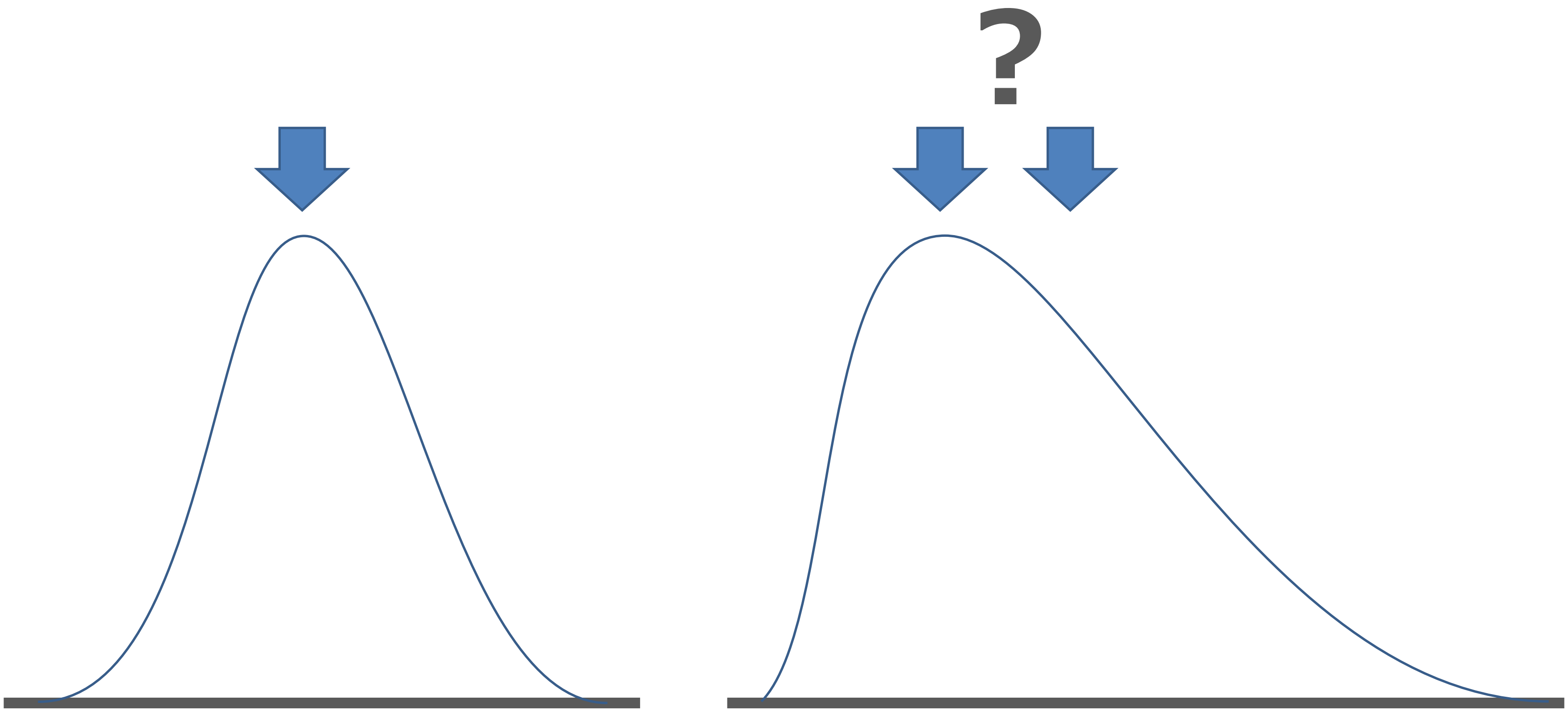
（基本・基礎）統計量

分布　データの散らばり具合



ヒストグラム（頻度分布図）

イメージ　データの中心



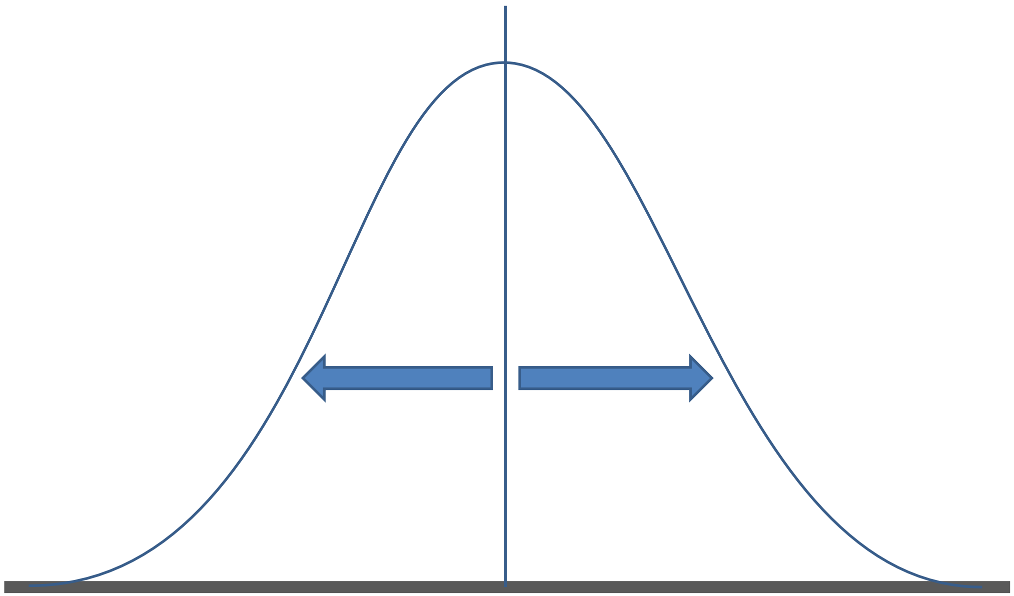
偏りのないデータ

身長の分布など

偏っているデータ

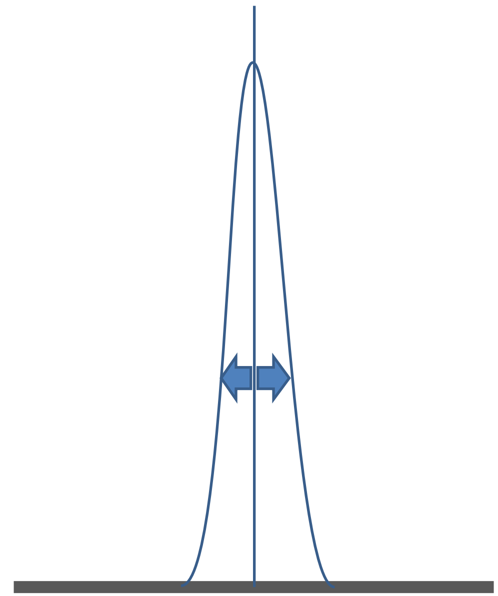
体重の分布など

イメージ　ばらつき



ばらつき大きい

中心からの差が
全体的に大きい



ばらつき少ない

中心からの差が
全体的に小さい

平均値

- 合計を計算
- 要素数で割る

中央値

- 小さい順（大きい順）にならべる
- 要素が奇数の場合、真ん中の値を採用
- 要素が偶数の場合、中央の２要素の平均値を計算

ばらつきとは？

分散、標準偏差

平均値からのずれの大きさ

分散

- 平均値を計算
- 各要素-平均値を計算
- その値を2乗
- その平均値を計算

分散

②要素iと平均値の差

①平均値

⑤要素数nで
割って平均
にする

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

③その2乗

④その全要素(iが1からnまで)の合計

分散…2乗された値

計測値と単位を
そろえるため

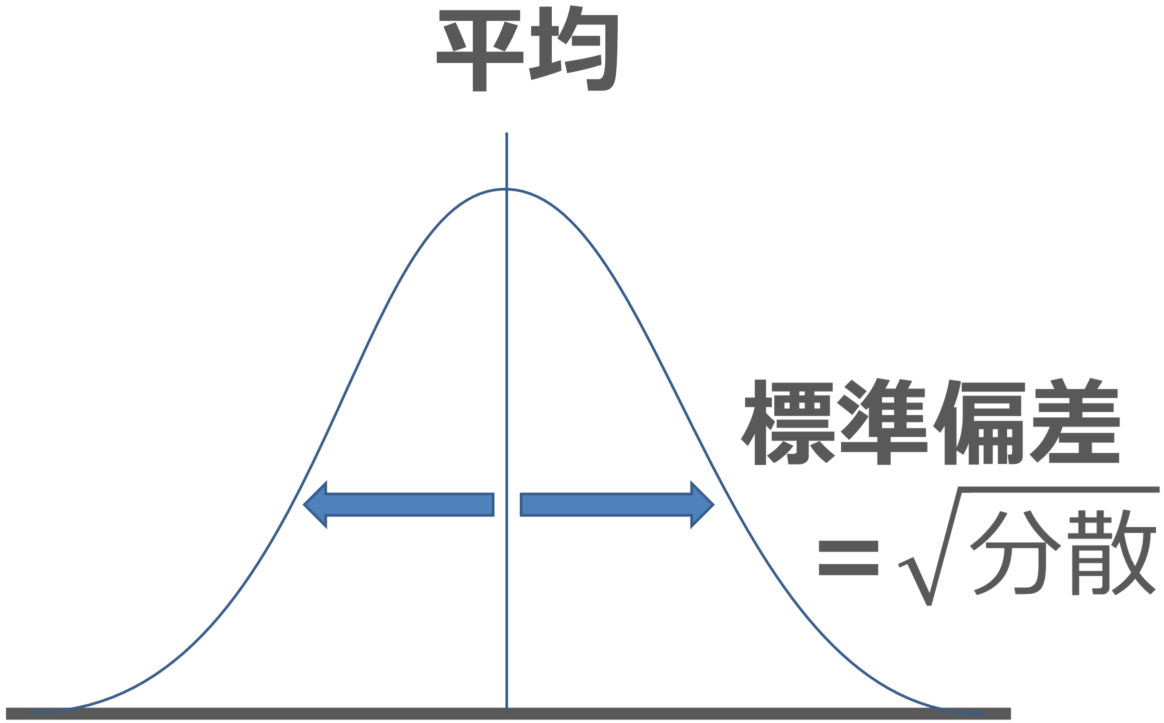
平方根を計算

標準偏差

目的：このクラスの人の身長はどのくらい？

	平均		標準偏差
男性		±	
女性		±	

イメージ



もっと広い
世界が知りたい

目的：このクラスの人の
身長はどのくらい？

目的：日本人の身長はど
のくらい？

全員の身長を測定して計算する

現実的ではない。
コストもかかる

何名かを抜き取り調査する

サンプリング（抽出）

サンプリング

偏りなくランダムに選ぶことが原則

ランダムサンプリング
（無作為抽出）

サンプリングされた要素

標本
（サンプル）

今回の目的の場合、
サンプリングされた人のこと

サンプリング前の要素全体

母集団　＝解析の対象

今回の目的の場合、日本人全員のこと

標本の数が多いほど、正確になる！

目的：日本人の身長はどのくらい？

↓

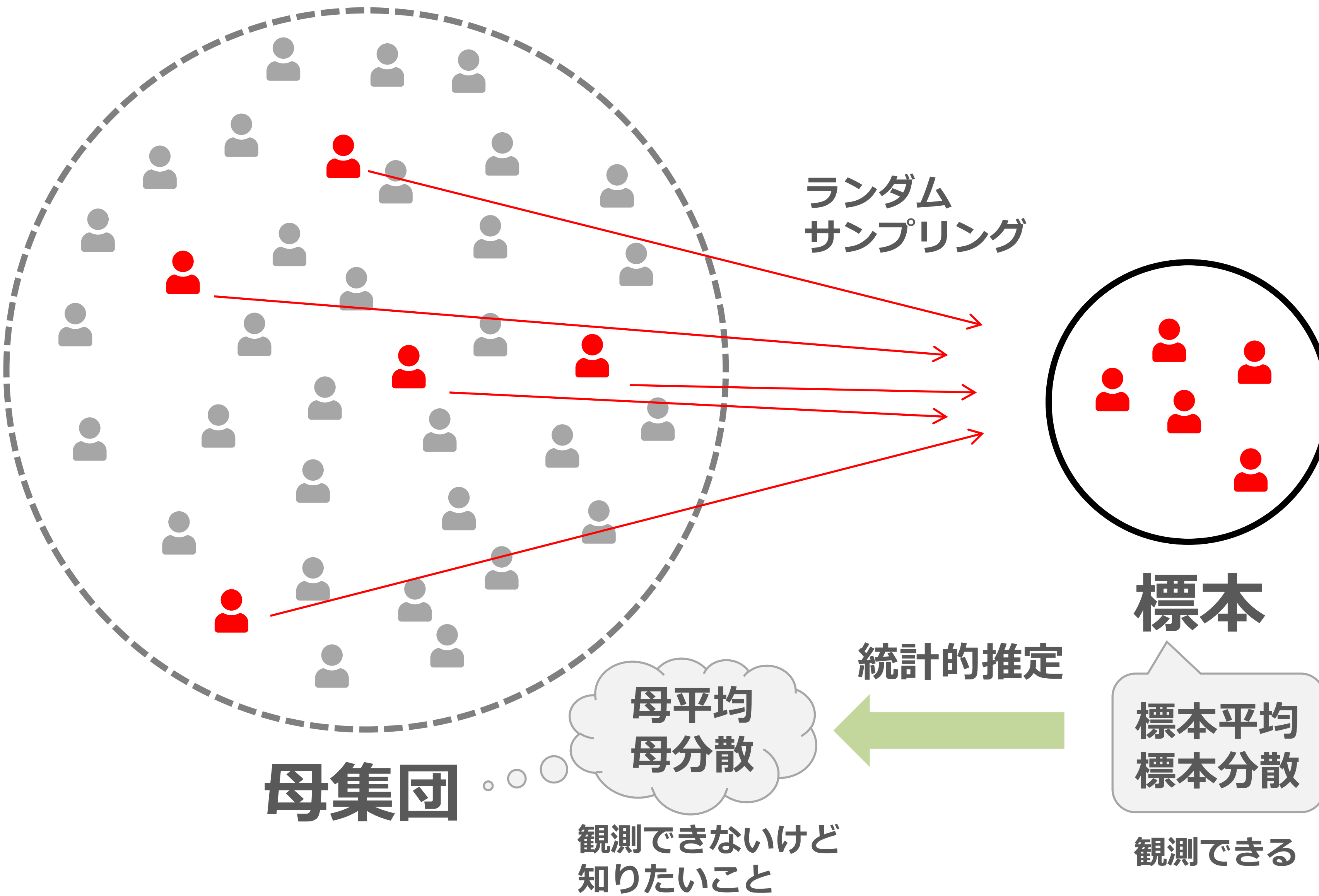
限られた標本から母集団（日本人全体）の

- 推定の平均値や
- 推定のばらつき

を計算する、という問題

統計的推定

母集団が大きい、あるいは無限で、直接観測できないとき、標本を観測することで、母集団の性質を調べる。



母平均μ ← 標本平均$\bar{x}$

一致が期待できる

母分散σ^2 ~~←~~ 標本分散s^2

母集団の全標本を観測できる場合は一致するが、
そうでない場合は、**実は一致が期待できない**

一致が期待できる

不偏(標本)分散v^2

真の値から外れていないことを、
不偏性があると言うので。

標本分散

②要素iと平均値の差

①標本平均

⑤要素数nで割って平均にする

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

③その2乗

④その全要素(iが1からnまで)の合計

不偏(標本)分散

⑤n-1で割る

$$\frac{1}{\boldsymbol{n-1}}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

n-1で割る？

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- 標本の数nが母集団の数N（大きな数）に近づくと、母分散に近くなる

➡ 母分散の推定に使える

- 自由度を表している

自由度＝互いに影響を与えない（独立した）値の数

上の式で、一つの観測値x(i=a)は他と完全に独立ではなく、それ以外の(n-1)個の独立した観測値と平均値$\bar{x}$によって求められる。

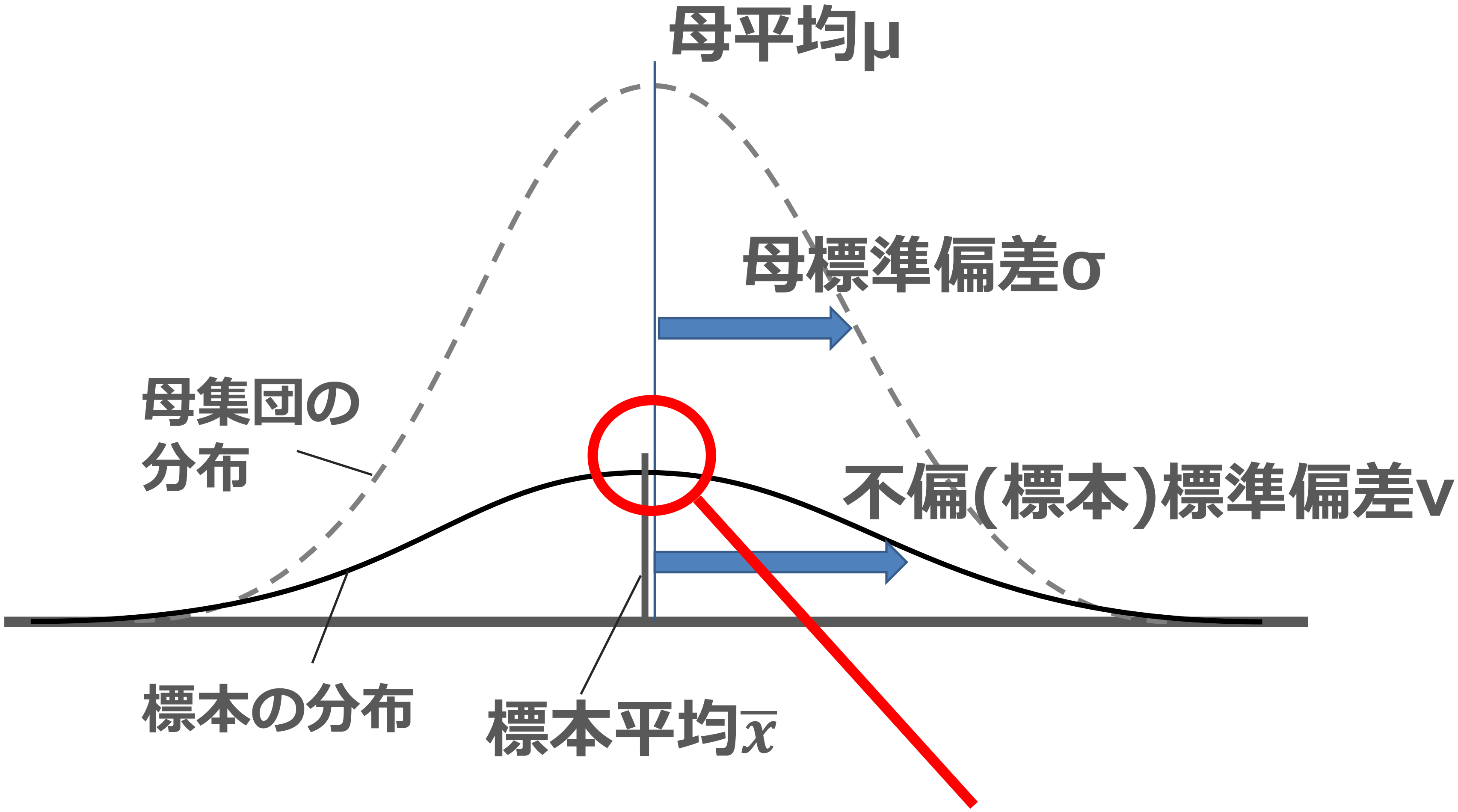
用語より、
n-1で割っているか
どうかに注目

書籍によって、標本分散s^2を不偏標本分散（不偏分散）のこととして記述しているものもあります。「（不偏）標本分散」と記述されることもあります。標本を考える時点で、そもそも母集団の推定を前提としていることが多いためです。

nで割っていたら、**観測値**の話
n-1で割っていたら、**推定値**の話

です

イメージ

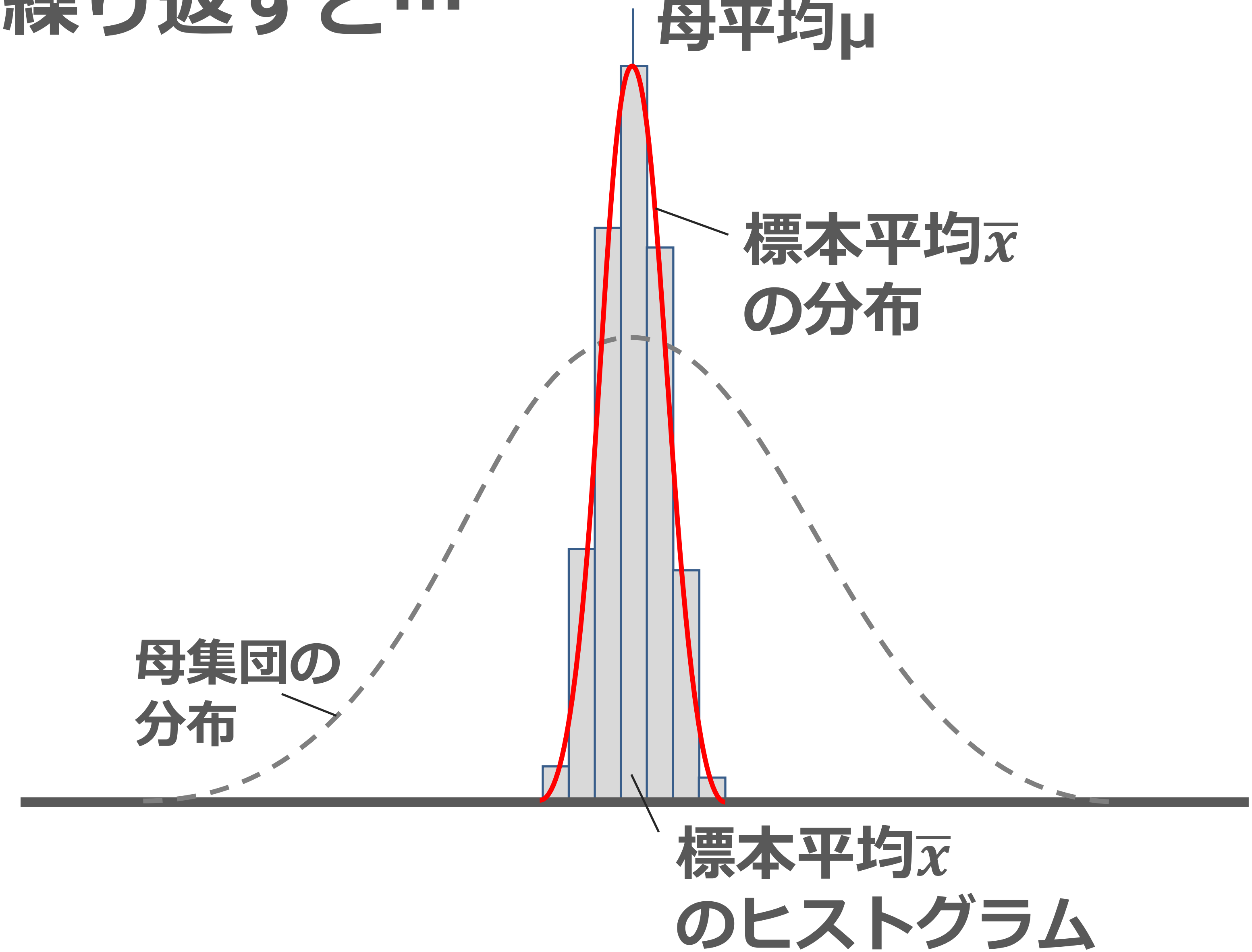


きっとズレが生じている

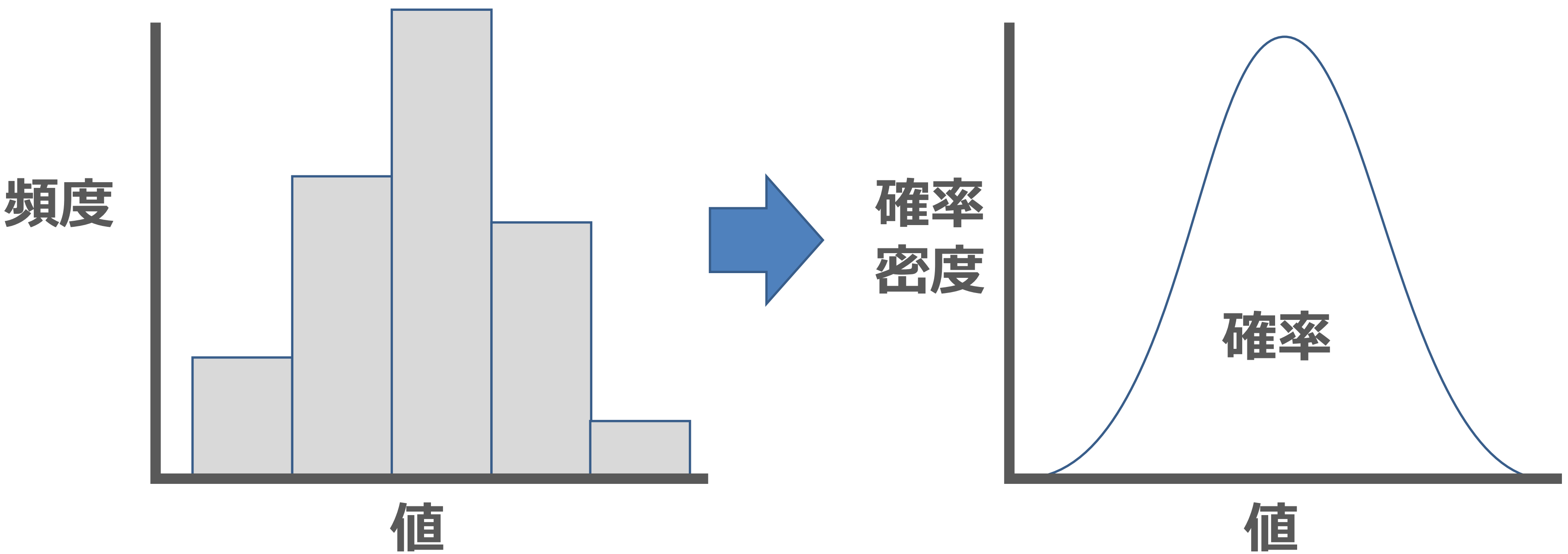
誤差

- サンプリング誤差
- 測定誤差

サンプリングして標本平均$\overline{x}$を算出して、を繰り返すと…



分布　データの散らばり具合



ヒストグラム

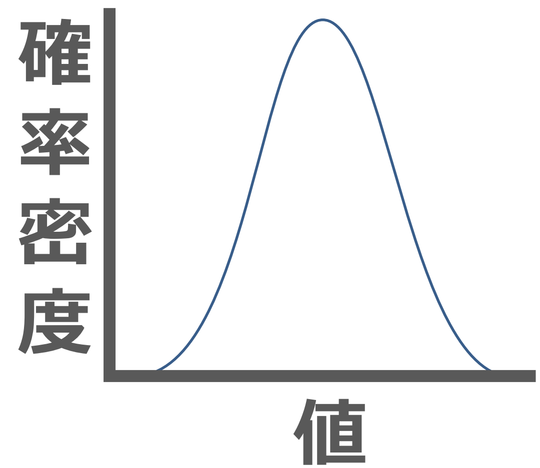
観測結果

確率密度関数

事象の起こる確率を表すモデル

正規分布（ガウス分布）

● 平均値が中心で、
● 平均値に近いものが多く、
● 左右に均等な釣り鐘状の分布



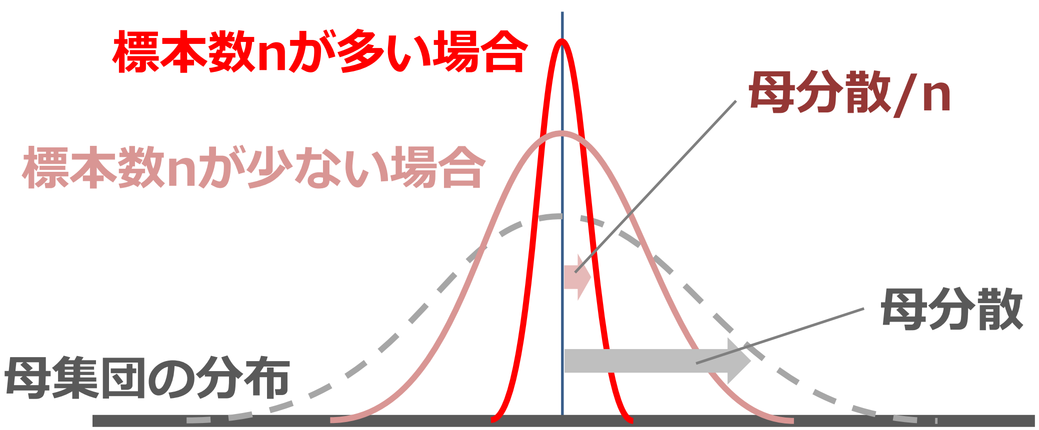
均等な確率で生じたばらつき
の場合にとる分布

✓ 身長の分布
✓ 測定誤差の分布
✓ 自然界で起こるゆらぎ　　　など

標本平均$\overline{x}$の分布

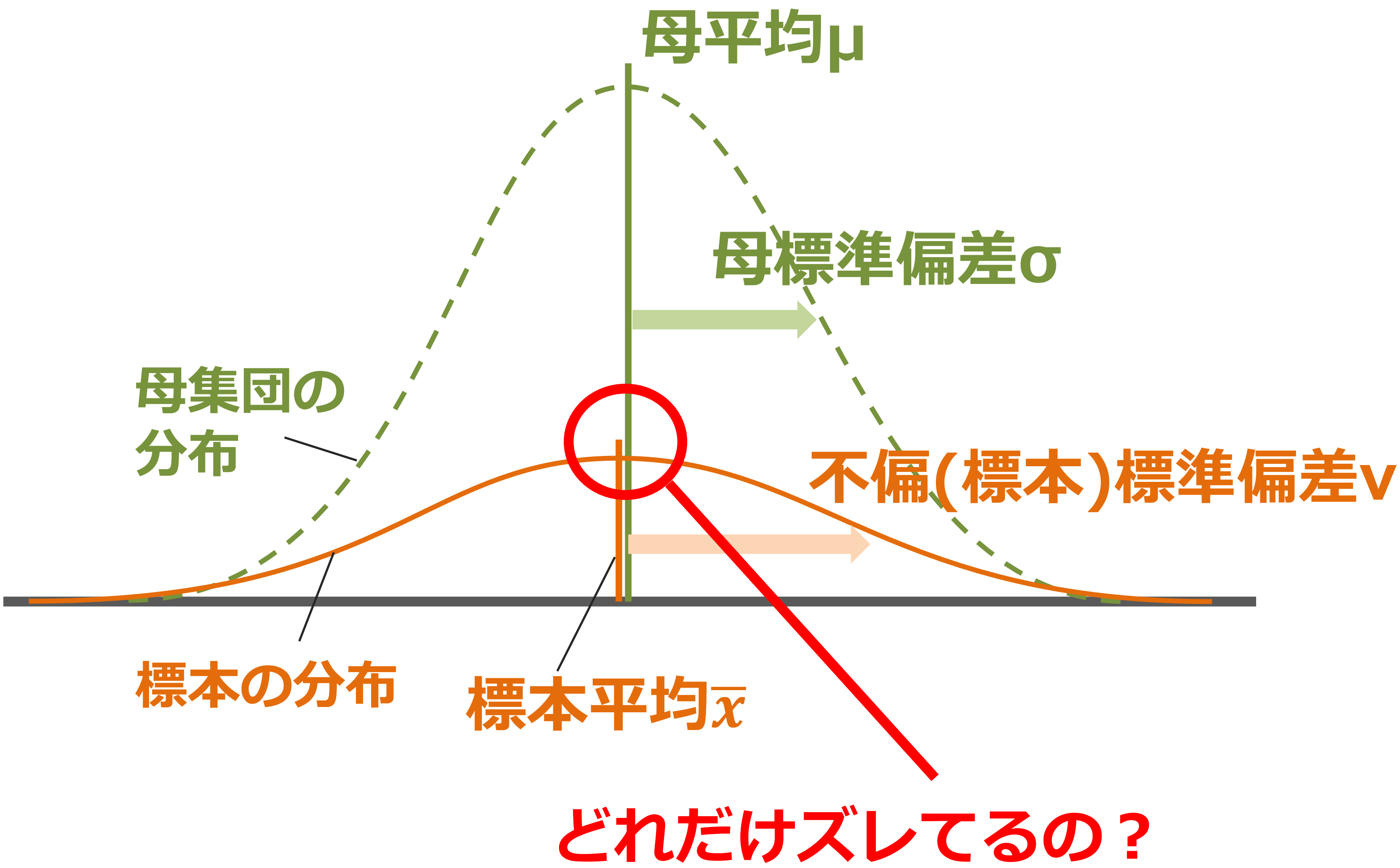
- 正規分布に従う
- 標本の数nが大きいほど、母平均μの推定確度は高まり、分散が小さくなる
- 分散は**母分散σ^2の1/n**になることが知られている

n=母集団数Nなら、全数検査なので、母平均μとのずれはゼロになる。
n=1なら、母集団のうち一つずつを測定するのと同じなので、分散も同じ。



ばらつき度合いを、標本数nで表せる

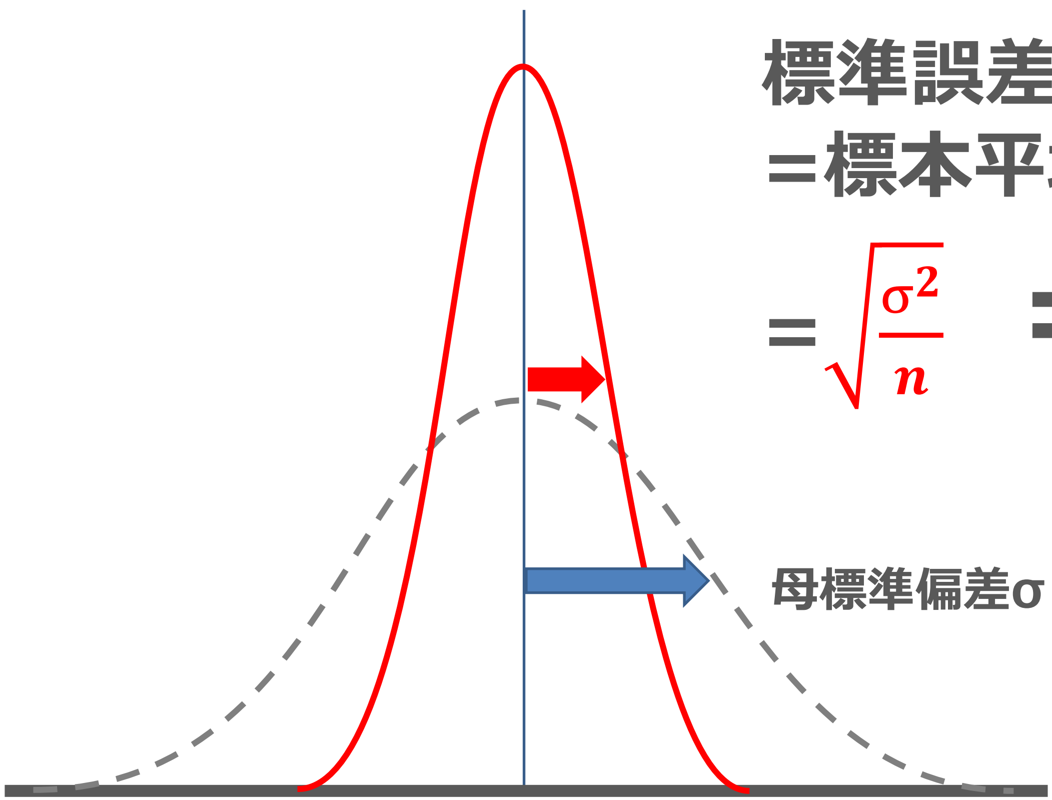
中心極限定理


母平均μ
母標準偏差σ
母集団の
分布
不偏(標本)標準偏差v
標本の分布
標本平均$\overline{x}$
どれだけズレてるの？

➡ 標本数nで示せる!!

標準誤差

- 標本平均$\bar{x}$の分布の標準偏差のこと。
  つまり、母平均μの推定値のばらつきを表す
- **母分散σ^2の1/nの平方根**



標準誤差
＝標本平均$\bar{x}$の標準偏差

$$= \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

標準偏差と標準誤差

論文などでよく見る図

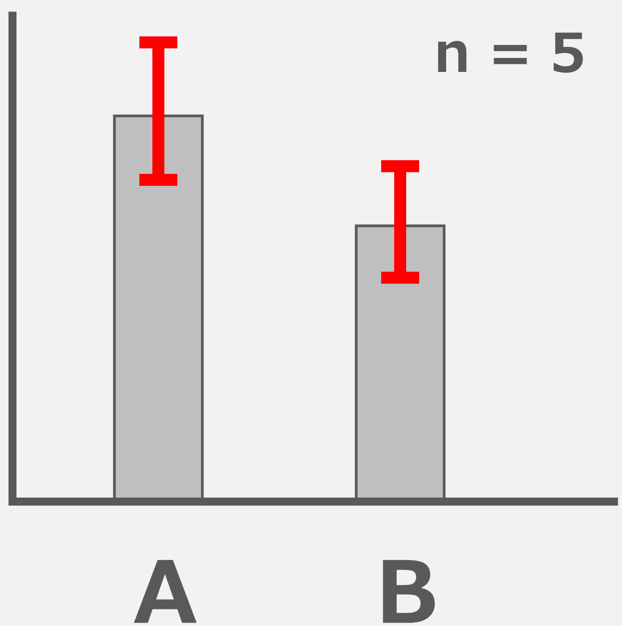


図１ A群とB群の＊＊の違い
それぞれ5個体を測定した。
エラーバーは標準偏差を表す

エラーバーが標準偏差

測定した標本自体の平均値を論じている

エラーバーが標準誤差

測定した標本から推定される母集団の平均値について論じている

標準誤差は標準偏差の$1/\sqrt{n}$なので、エラーバーは短くなり、より明確な差がありそうな見栄えになります。標準誤差を示すことが適当なのかどうかを、正しく判断しながらデータを解釈しましょう。

計算してみよう

このクラスの身長データからいくつかのデータを抜き出し、クラスの身長の平均値を推定してみる