データサイエンスへの誘い

第7回: データからの推論

瓜生真也（デザイン型AI教育研究センター・助教）

講義内容（予定）

講義に関する資料（スライド、補足資料等）を GitHubに置いておきます

https://github.com/uribo/INNV1250

ダウンロード可能

1. ガイダンス、データサイエンスとは何か
2. 現代社会におけるデータサイエンスの活用事例
3. データ処理の手法
4. データの要約
5. データの可視化
6. データと確率
7. データからの推論
8. 複数のデータを比較する
9. 統計のウソ
10. 統計的モデリング
11. 統計的学習
12. さまざまなデータサイエンスの手法
13. 機械学習とAI
14. コンピューターを用いた分析
15. ビッグデータの扱い
16. 期末試験（8月1日）

今日の目標

統計的推測のための

推定と仮説検定を知る

統計的推測

観測されたデータを用いて、未知の母集団の特性を得るための一連の手続き

分布は？平均値は？分散は？

統計的推定と**仮説検定**の2つの主要な手法を利用する

統計的推定

標本の特性から母集団の特性を推測

点推定と区間推定が含まれる

（よく使われる例え）味噌汁の味を確認するための味見

仮説検定

データを用いて特定の仮説が真であるかを評価する

確認された事象が偶然生じたものなのか

→いずれも「統計学」「確率」を利用する

~~まずは確率とその背景について理解を深めよう~~

概念と基礎の理解を目指そう

week07/statistical_inference.ipynb

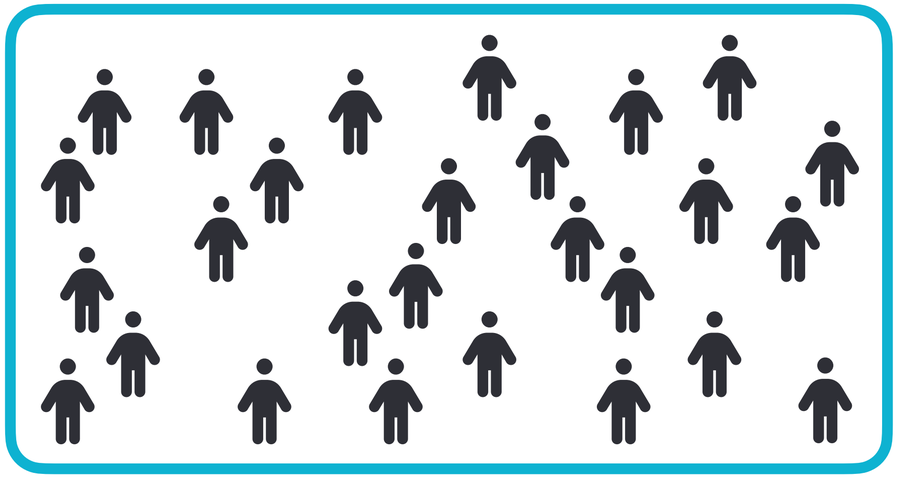
母数と統計量

母集団と標本、母数と統計量

平均値や分散は
母集団の性質を反映する
→**母数**

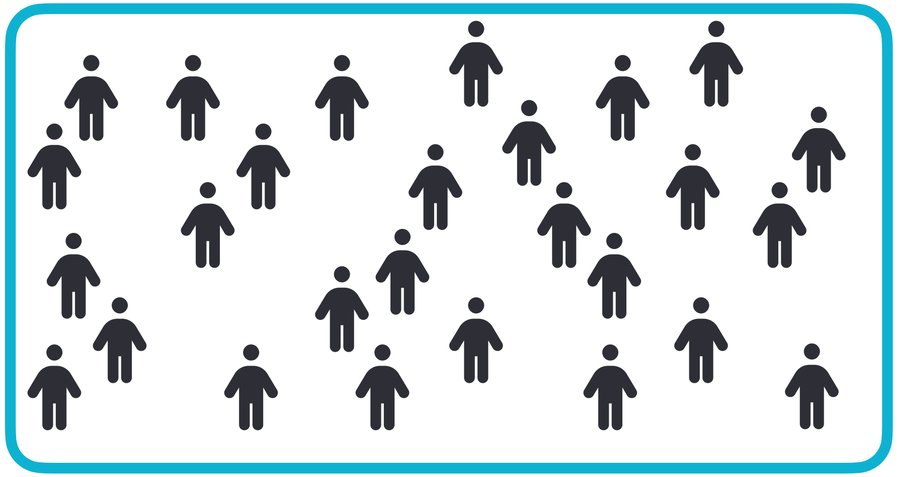
全数調査をしない場合は、一般に未知の値

母集団



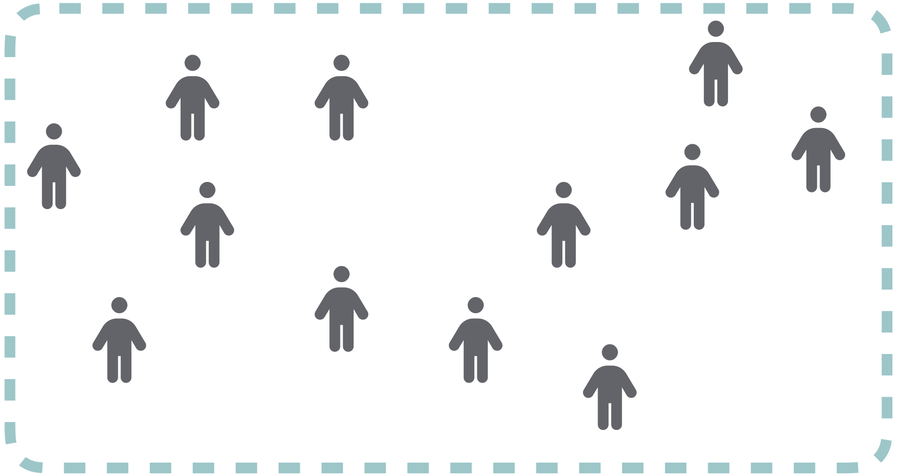
標本から母集団の性質を
推測するために抽出される情報
→**統計量**

全数調査



母集団全体を調査。国勢調査など

標本調査



母数と統計量

区別せず使われることもあるので、文脈による判断が必要

対象	母数	統計量
データの値の合計をデータの数で割った値	**母**平均	**標本**平均
データの値が平均値からどれだけ離れているかを示す尺度	**母**分散	**標本**分散
データの値が平均値からどれだけ離れているかを示す尺度	**母**標準偏差	**標本**標準偏差
…	…	…

母数と統計量で対応関係をもつ

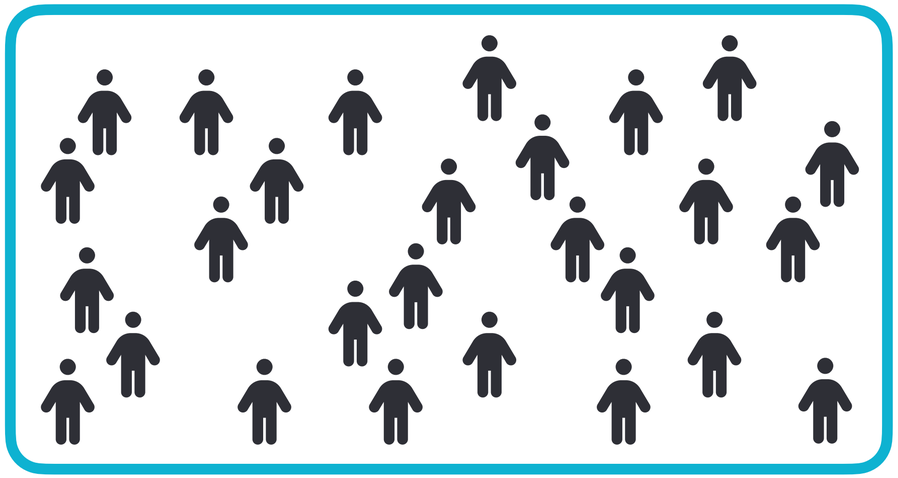
week07/statistical_inference.ipynb

統計的推定

点推定

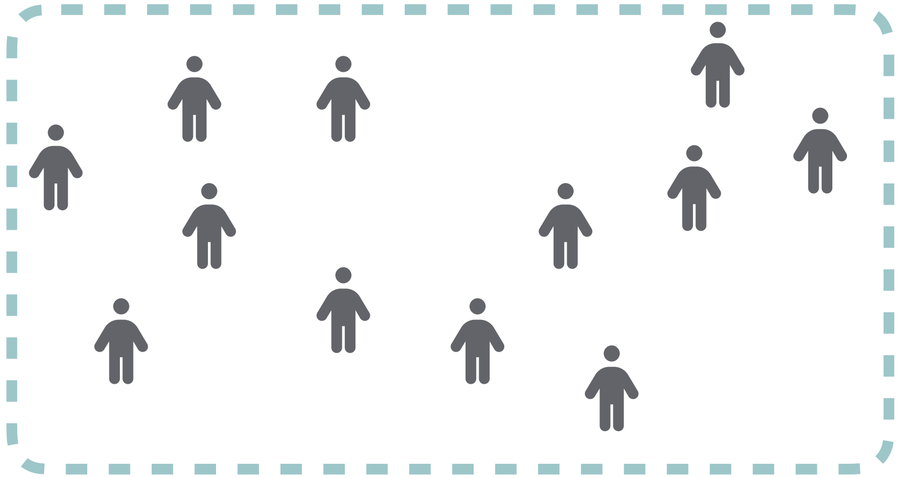
母数の値を1つの数値で推定する

あるクラスの生徒（母集団）



母平均は未知

無作為抽出（標本）

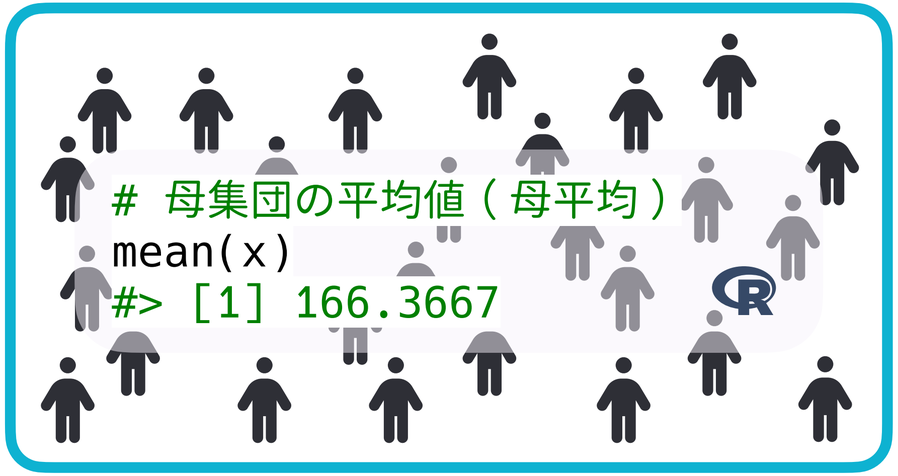


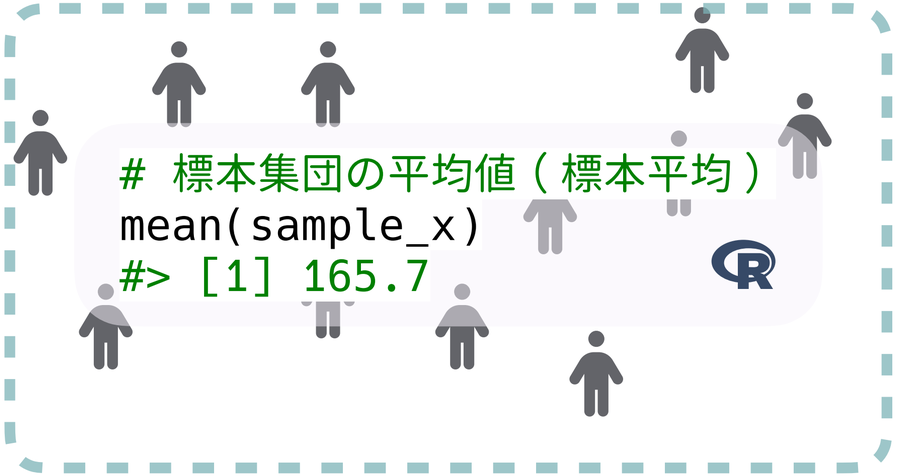
平均が計算できる（標本平均）
→この値を母集団の平均値としたい

標本が十分に大きい、偏りのない標本であれば
母集団の性質を反映できていると考えられる

偏りのない標本抽出の重要性

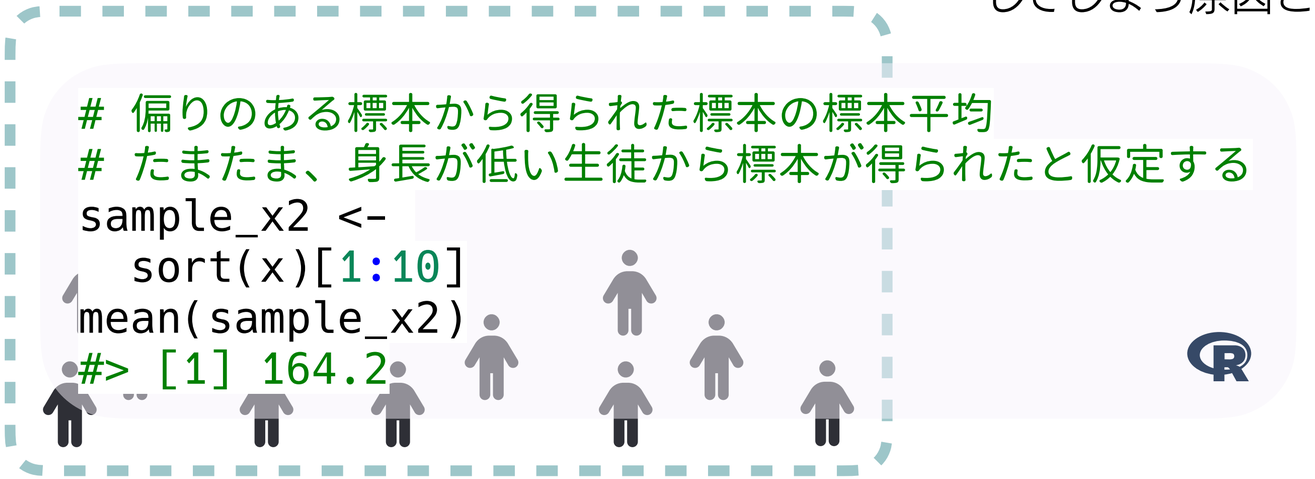
母集団の値を知っている「神の視点」をもつとして…





母集団について誤った理解をしてしまう原因となる

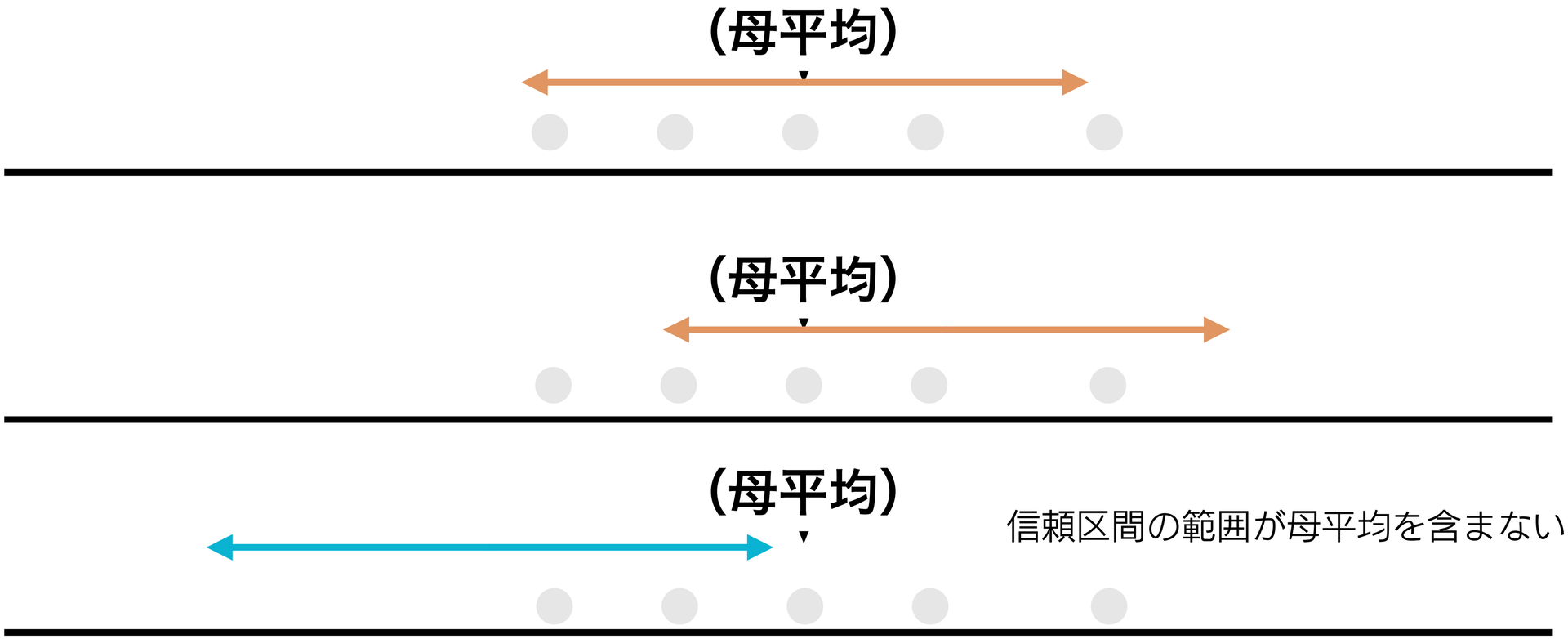
標本に偏りがあると
真の値（母数）との
差が大きくなる



区間推定

点推定の結果に対する不確実性を確率的に評価する

95%信頼区間　複数の標本による信頼区間の範囲に母平均を含む確率が95%



信頼区間の範囲が母平均を含まない標本集団もある

week07/statistical_inference.ipynb

仮説検定

母集団の特性についての仮説の検証

得られたデータから仮説が正しいか、稀なことか（偶然ではないこと）を評価する

帰無仮説 →保持したい主張や状況のない場合を表すもの。無に帰したい

対立仮説 →本当に証明したい仮説

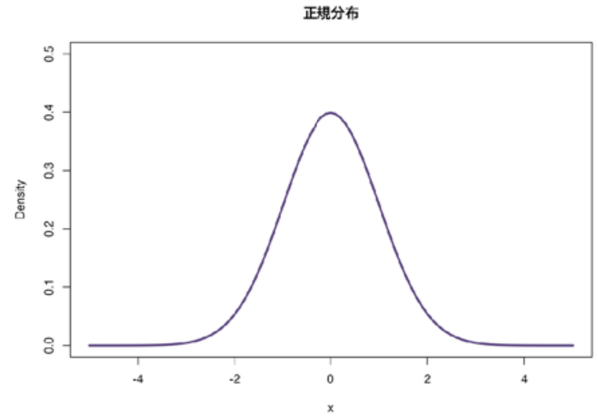
データと確率を用いて判断

棄却 あるいは 採択

帰無仮説を棄却することで対立仮説を採択する

仮説検定の手順

1. 帰無仮説と対立仮説を設定する
2. 有意水準αを設定する(5%、1%など)
3. 検定統計量を計算する
4. 検定統計量の分布を決定する
5. 検定統計量の実現値を計算する
6. p値を計算する



仮説検定の例

あるクラスの生徒の身長について、男女で平均値に差があるか

帰無仮説 →男女間で身長の平均値の差はない

対立仮説 →男女間で身長の平均値の差がある

```r
# 男女別でのクラスの生徒の身長
height_male <-
  c(160, 162, 165, 167, 168, 170, 171, 172, 173, 173)
height_female <-
  c(156, 158, 161, 163, 164, 162, 167, 154, 169, 171)
```