統計学入門
～データ活用社会と統計学～

2025年度1学期:　　月曜2限

担当教員:　　石垣　司

超スマート社会の実現へ

Cyber-Physical System (CPS) ではデータが"資源"

（CPS：現実世界のデータから仮想空間で情報・知識を創出し，現実世界での価値を提供するシステム）

（Society 5.0：日本の政策(2016年閣議決定)。第5期科学技術基本計画で提唱された目指すべき未来社会の姿）

- AIやDXを手段として，データを活用した社会・ビジネス変革があらゆる産業での現代的なテーマに

政府広報オンラインHPより　https://www.gov-online.go.jp/cam/s5/pdf/Society5.0_20d.pdf (最終閲覧2019年10月1日)

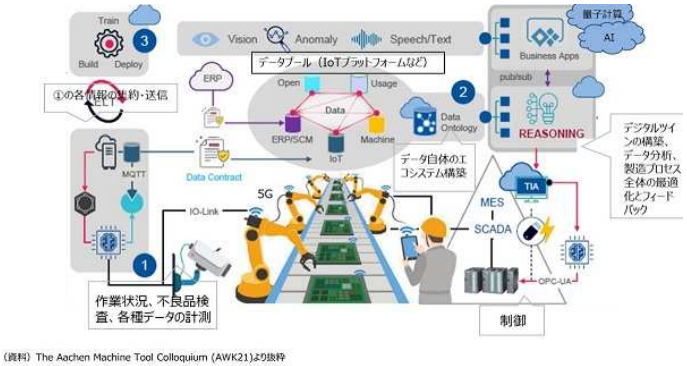
DX（デジタルトランスフォーメーション）

経済産業省によるDXの定義：「企業がビジネス環境の激しい変化に対応し，データとデジタル技術を活用して，顧客や社会のニーズを基に，製品やサービス，ビジネスモデルを変革するとともに，業務そのものや，組織，プロセス，企業文化・風土を変革し，競争上の優位性を確立すること」

（経済産業省 DX推進ガイドライン，2018.12 ⇒ **デジタルガバナンスコード** 2.0，2022.9）

DXの意義は変革と価値創造（デジタル化，IT化との差異）

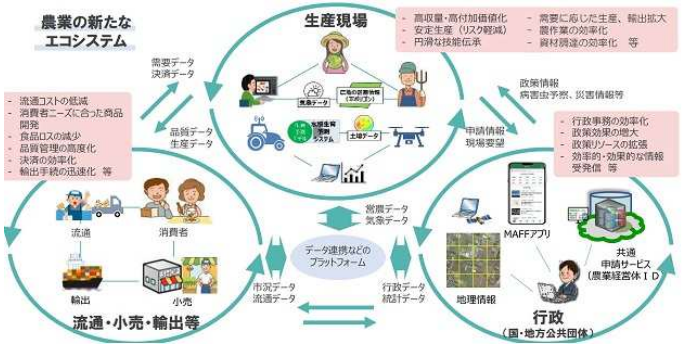
- **「DX」という単語自体はバズワードだが，その意義と必要性の本質は現代日本のあらゆる産業において普遍的**



（資料）The Aachen Machine Tool Colloquium（AWK21）より抜粋

製造業：スマートファクトリー（経済産業省）

https://www.meti.go.jp/shingikai/mono_info_service/sangyo_cyber/wg_seido/wg_kojo/pdf/002_03_00.pdf



農業DX（農林水産省）

https://www.maff.go.jp/j/kanbo/dx/attach/pdf/index-61.pdf

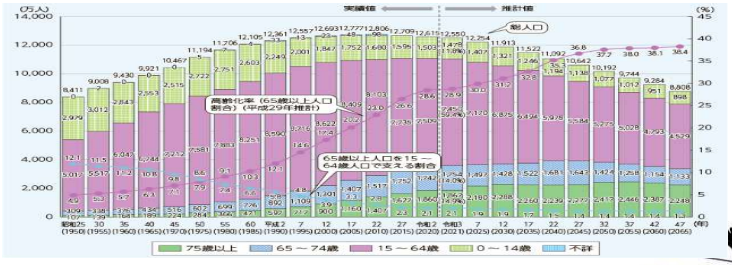
大学・研究（東北大学）

https://www.dx.tohoku.ac.jp/

日本社会とDX・データ活用のそもそも論

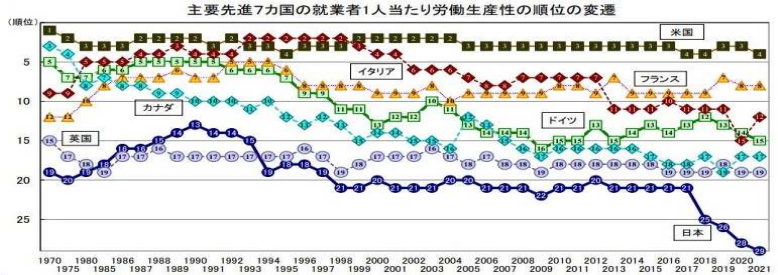
DXやデータ活用は，日本が抱える諸問題と向き合いながら安全で安心な生活を維持していくための一つの手段

急激な生産年齢人口の減少
(2020年:7450万人→2065年:4529万人)



総務省 情報通信白書 2022

国際的に低い生産性
(1970年:38か国中19位 → 2021年:29位)



日本生産性本部 労働生産性の国際比較2022

DX, AI, データの活用で
我々の生活の破綻を回避！

難しく思われがちだけど，

「足りない労働力はコンピューターやロボットで代替」
「定型作業はコンピューターで自動化。人間は創造的作業に集中」
「コンピューター・AIで人間の意思決定を高度にサポート」

という，ごく自然なモチベーション

データ ≠ 情報

社会・学術・経済・経営の問題解決に必要なのは「情報」

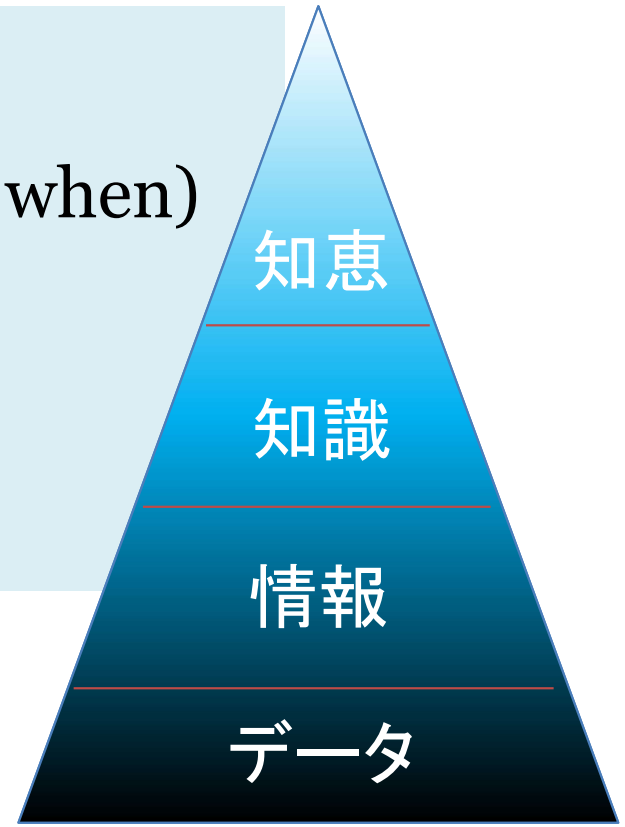
– **データ活用のためにはデータから情報を抽出・創出する**

データ: know-nothing, **符号や信号**

情報: know-what (who, where, how many, when)

知識: know-how, **パターンや情報の集合**

知恵: know-why, what to do,
why do, what is best.



DIKWピラミッド

「情報」を「知識/知恵」に昇華させることがデータ活用の理想形

本授業と統計学

統計学とは？

- 集団現象を観察し分析する方法を研究する学問（大辞林）
- 確率論を基盤にして，集団全体の性質を一部の標本を調べることによって推定するための処理・分析方法について研究する学問（デジタル大辞泉）

本授業で目指す統計学：

「社会や学術上の問題を解決するために数学とコンピュータを用いてデータを活用する方法の学問」

統計の起源① 国の実態をとらえるための統計

古代：国家統治のための統計（例：ピラミッド建設）

- Statistics の語源はラテン語 Status（国家・状態）
- センサス：アウグストゥス（古代ローマ帝国初代皇帝）の治世の頃に行われた人口・土地の調査

Augustus
(27 BC - AD 14)

17世紀以降の欧州：国家間の勢力比較

- ドイツの国勢学

  国家の基礎事項（人口、土地面積、貿易規模など）を数量化

  18世紀以降, デンマーク, 米, 蘭, 英などで初の近代的センサス

現在：国勢調査

- 公的統計は民主的な社会の情報システムの不可欠な要素
  （「公的統計の基本原則」国連統計委員会採択1994）
- 日本では, 日本に住んでいるすべての人と世帯を対象として5年に1度実施

※総務省統計局補助教材より　https://warp.da.ndl.go.jp/info:ndljp/pid/11489777/www.stat.go.jp/teacher3/episode01.html （2025年3月）

統計の起源① 現在の日本 #1

例：経済統計

- 社会全体で利用されるべき情報基盤として官庁により体系的に整備されている
- 個別主体：人口, 労働, 家計など
- 経済全体：GDP, 景気動向など
- 国民経済計算（新SNA：System of National Account）
  国連が定める国際基準に準拠
  日本では1979年より内閣府が公表（統計法に基づき作成）

官庁の調査は統計法に従う

- 調査目的の明示, 調査の実施方法, 結果の公表の方法など

統計の起源① 現在の日本 #2

統計法：公的統計に関する基本法

- 公的統計の体系的・計画的整備，統計データの利用促進，統計の公表，統計調査の対象者の秘密の保護，「かたり調査」の禁止，統計委員会の設置

統計法 第1条

- 「公的統計が国民にとって合理的な意思決定を行うための基盤となる重要な情報であることにかんがみ，公的統計の作成及び提供に関し基本となる事項を定めることにより，公的統計の体系的かつ効率的な整備及びその有用性の確保を図り，もって国民経済の健全な発展及び国民生活の向上に寄与することを目的とする」

統計の起源② 大量の事象をとらえるための統計

政治算術学派（17世紀イギリス）

- 政治的解剖（社会構造を解剖し国政へ活用）
- 大量観測により法則や因果関係の発見を重視
- 数と量と尺度によって社会を対象とした議論を展開

- **ペストの予測**（John Grant）
  - 死亡統計表（生命表）(1662)の分析
  - ロンドンの人口の見積もり（200万人⇒約38万人）

- **保険料算出**（Edmond Halley）
  - 当時の保険料はギャンブル要素強
  - 死亡統計表から合理的な保険料を算出
  - 生命保険事業の基礎

William Petty (1632-87)

※総務省統計局補助教材より　https://warp.da.ndl.go.jp/info:ndljp/pid/11489777/www.stat.go.jp/teacher3/episode01.html （2025年3月）

統計の起源② 日本の物価指数 #1

消費者物価指数(CPI)

- 経済の体温計。総務省
- 582品目の調査(2020年基準, 5年毎更新)

国内企業物価指数(PPI)

- 企業間取引の価格動向。日本銀行
- 515品目の調査(2020年基準)

GDPデフレーター

- 名目GDPと実質GDPの比。内閣府
- すべての経済活動に伴う新たな生産物の価格変動の指標

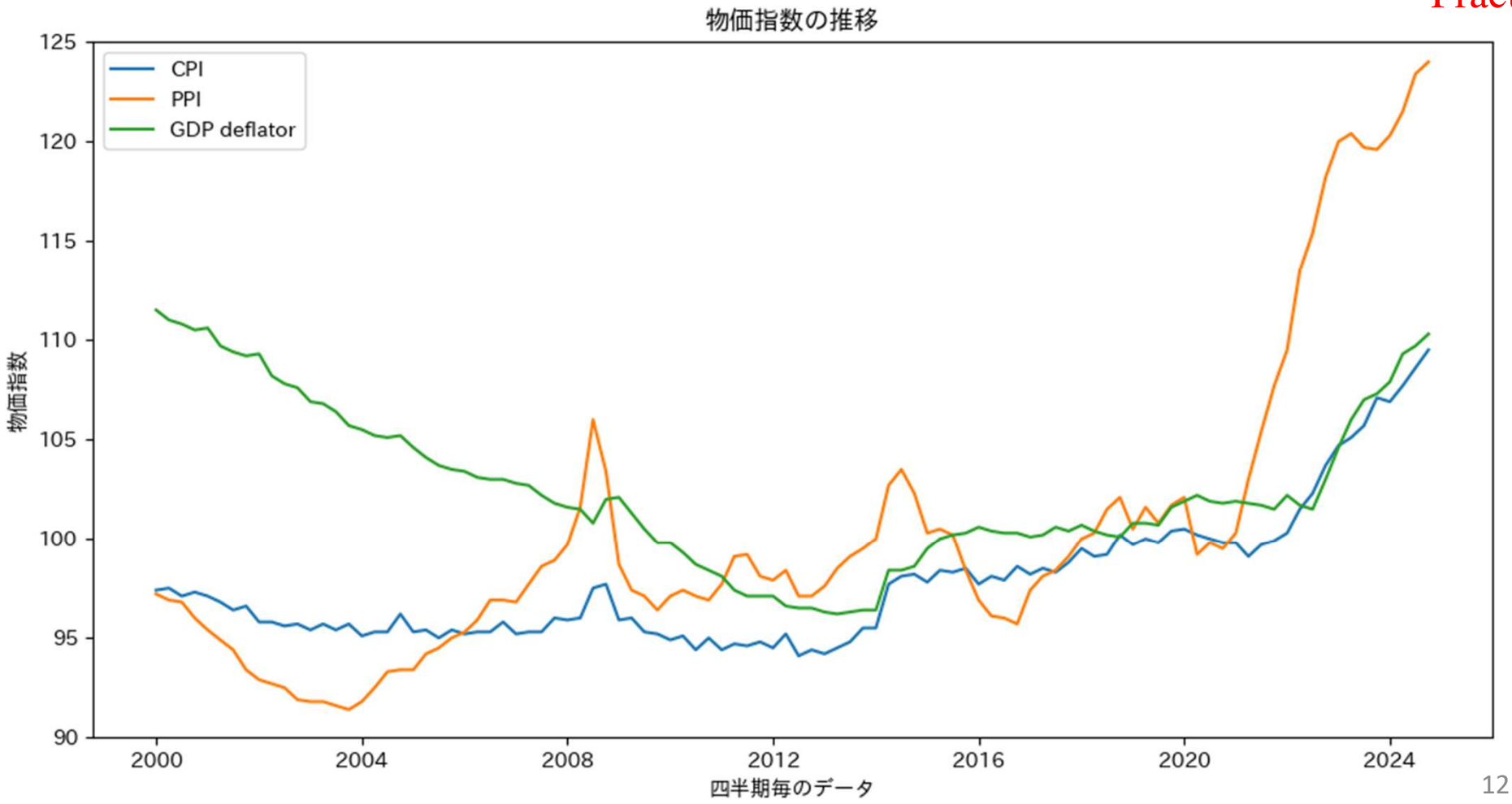
CPIとPPIは調査対象品目が限定的

統計の起源② 日本の物価指数 #2

それぞれ異なる意味を持つ指標なので注意

- **例：**2004 ~ 08**年頃**, CPI **横ばい**, PPI **上昇**, GDP **下降**

Practice!



統計の起源③ 確率的事象をとらえるための統計

ギャンブル（サイコロ賭博やトランプ）の研究から確率論が誕生

Gerolamo Cardano
(1501-76)
標本空間の概念
「さいころ遊びについて」

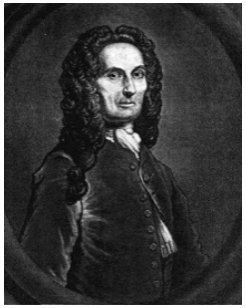
Galileo Galilei
(1564-1642)
小論「サイコロゲーム
についての考察」

Blaise Pascal
(1623-62)

Pierre de Fermat
(1607-65)

標本空間の基本的概念
（パスカルとフェルマーの往復書簡）

Jakob Bernoulli
(1654-1705)
天然痘の罹病率,
死亡率の計算



Abraham de Moivre
(1667-1754)
年金論への応用

Thomas Bayes
(1701-61)
ベイズの定理

Leonhard Euler
(1707-83)
標本調査に基づく
全体の推計方法

Joseph-Louis Lagrange
(1736-1813)

Pierre-Simon Laplace
(1749-1827)

確率論の大成

※総務省統計局補助教材より　https://warp.da.ndl.go.jp/info:ndljp/pid/11489777/www.stat.go.jp/teacher3/episode01.html （2025年3月）

近代統計学(19世紀以降)

3つの流れを「統計」として Quetelet が統一

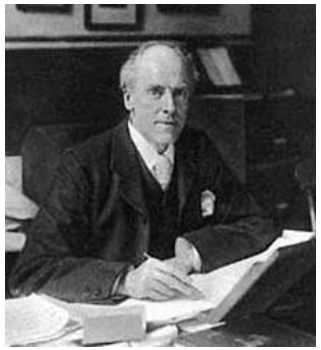
- 社会現象の解明に確率論を導入
  - 出生, 犯罪, 結婚, 自殺などの発生率へ正規分布
  - 個人の行動の集合も科学として追及

Adolphe Quetelet (1796-1874)
「近代統計学の父」

この授業で扱う統計学

記述統計学(Pearsonが大成)

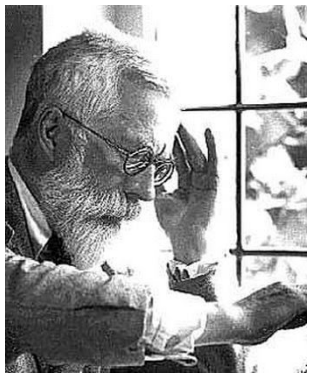
- データの平均・分散, 可視化などから分布を議論し, 対象の傾向や性質を把握
  - ヒストグラム, 標準偏差, 相関係数など



Karl Pearson (1857-1936)

推測統計学(Fisherが体系化)

- 部分の標本から全体の構造を推定
  - 仮説検定, 最尤推定, 実験計画など



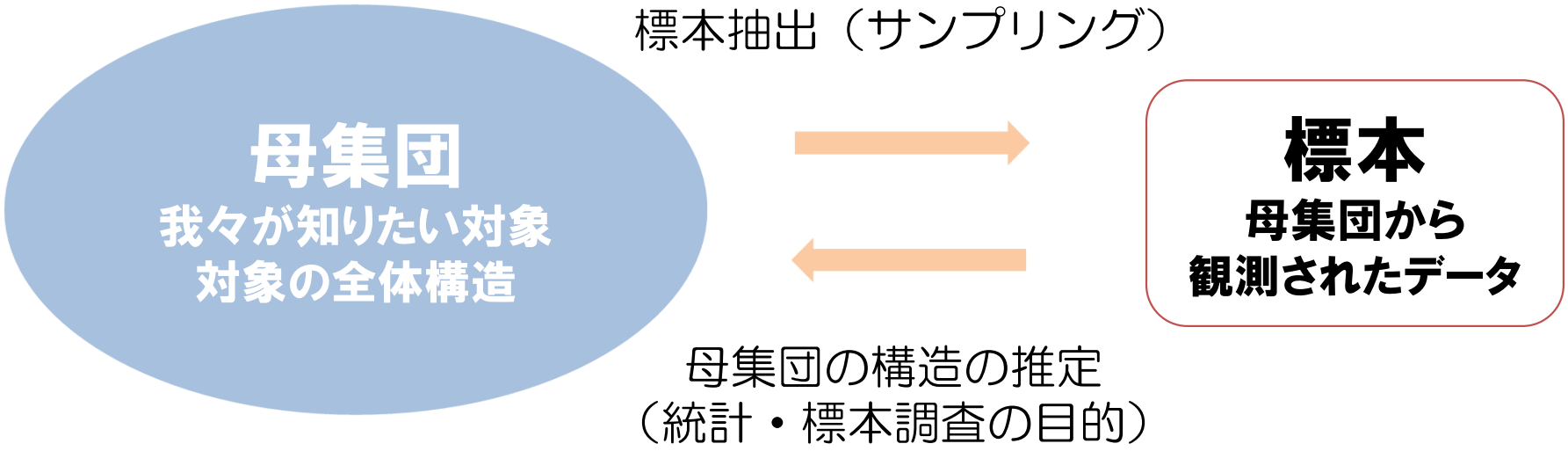
Ronald Fisher (1890-1962)

近代統計学における母集団と標本

統計的推測：標本から母集団の性質を推定

母集団：我々が考える対象の全体

標本：母集団から抽出された部分



例1： 仙台市立 A 高校3年生の身長　170cm, 148cm, 165cm, 181cm ...
例2： 日本の納税者の所得　400万, 1500万, 30万, 420万, ...

調査と標本抽出 #1

全数調査

- 母集団に含まれる全対象を観測して母集団について調べる
- 長所：母集団の知りたい対象を網羅できる
- 短所：母集団が大きい場合はコストや労力が膨大になる

標本調査

- 母集団から観測される標本を元に母集団の特徴を推定
- 長所：全数調査と比べて少ないコスト・労力で実施可能
- 短所：本質的に誤差を含む。選択バイアスが生じる可能性

調査と標本抽出 #2

有意抽出

- 調査実施者が調査対象を決めて標本を抽出する

  実施者の主観に調査結果が依存し，選択バイアスが大きくなる傾向

選択バイアス

- 抽出された標本が母集団の特性を反映していない偏り
- 標本調査では選択バイアスに常に注意が必要

  例：日本の納税者の平均所得の調査で東京都港区のみから標本抽出。港区民の平均所得1112万円（全自治体の中で1位）。（仙台市337万円）

  ※平均所得＝課税対象所得÷納税義務者数　総務省「平成28年度 市町村税課税状況等の調（しらべ）」より
  http://www.soumu.go.jp/main_sosiki/jichi_zeisei/czaisei/czaisei_seido/ichiran09_16.html

無作為抽出

- くじや乱数によりランダムに調査対象を決めて標本を抽出する方法。調査実施者による恣意性を排除

選択バイアスの例

1936年の米国大統領選挙の結果予測

- 共和党 Landon 氏 vs 民主党 Roosevelt 氏

- The Literary Digest の予測
  - 老舗総合雑誌で過去5回の大統領選的中
  - 雑誌の読者200万人に調査
  - 予測：得票率57%で Landon 氏の勝利

- アメリカ世論研究所（George Gallup）の予測
  - 前年に世論調査に初参戦
  - 適切に割当てた3000人への調査
  - 予測：得票率54%で Roosevelt 氏の勝利

George Gallup (1901-81)

結果：得票率60%でRoosevelt 氏の勝利

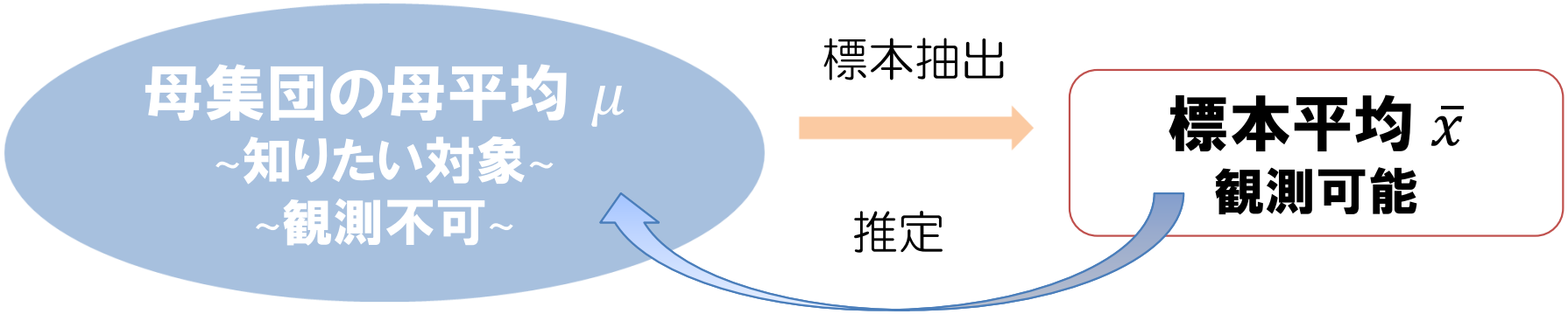
標本の代表値

代表値：標本の特性を代表する値

- 与えられたデータから計算される標本平均，中央値，最頻値，尖度，歪度，標準偏差，分散，パーセント点など
  - 記述統計量，基本統計量，要約統計量と同義

統計学での推定対象は母集団の特性

- 例：母平均 μ　　母集団の平均（標本調査では観測できない）
- 例：標本平均 $\bar{x}$　データの平均（データから計算可能）



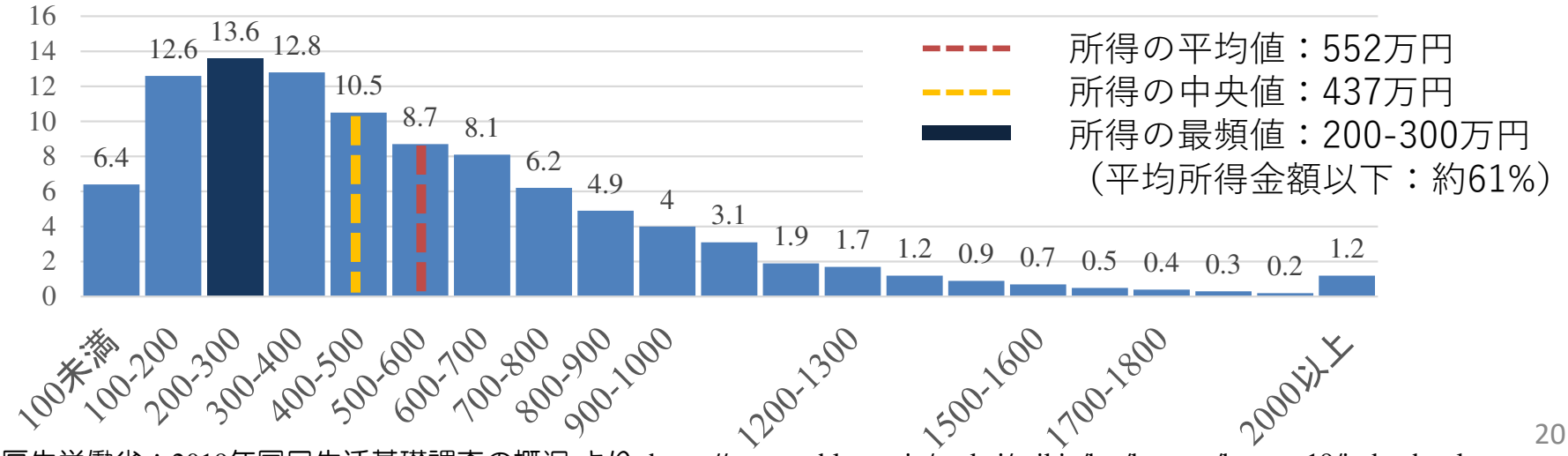
標本の平均, 中央値, 最頻値

標本 $\{x_1, x_2, \cdots, x_n\}$ の“真ん中”を表す代表値の例

- **標本平均: $\bar{x} = {}^1\!/_n \sum_{i=1}^{n} x_i$**
- **中央値: データを大小の順に並べた真ん中の値**
- **最頻値: データの中に最も多くあらわれた値**

問題: 適切な代表値は？

- **2019年調査の日本の世帯別所得の平均値は552万円である。多くの世帯で552万円の所得があるとして政策等を決定することは妥当であるか？**



厚生労働省：2019年国民生活基礎調査の概況 より https://www.mhlw.go.jp/toukei/saikin/hw/k-tyosa/k-tyosa19/index.html

演習問題

1. **全国から無作為抽出された固定電話の電話番号に電話をかけてアンケート調査を行った。また, その調査は2022年4月18日午後2時に行われた。この標本調査で生じ得る選択バイアスについて議論しなさい**

2. **標本** $\{x_1, x_2, \cdots, x_n\}, (n > 100)$ **が与えられたとき, その標本の平均値と中央値が一致するのはどのような場合か議論しなさい**