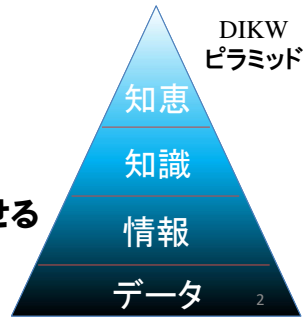
統計学入門 補助資料
～調査と標本～

2022年度1学期:　月曜2限
担当教員:　石垣　司

“データ” ≠ “情報”

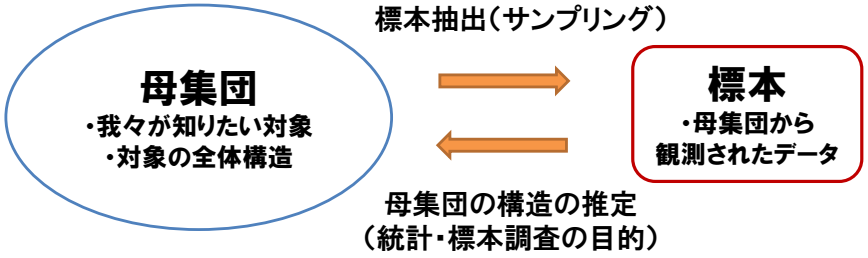
- データと情報は異なる意味をもつ
  - データ: 単なる符号, 信号
  - 情報: 5W1H(What, Who, Where, When, How many)等
  - 知識: ノウハウ, 情報の集合 等
  - 知恵: なぜ？, 何をすべき？, 何がベスト？等
- データ活用のために必要なこと
  - データから情報を創出する
  - 情報を知識や知恵へ深化する
  - 情報, 知識, 知恵を意思決定に反映させる

#メモ: DIKWピラミッド: 初出はT.S. Eliot(1934)の詩と言われている。
各段階の内容は様々な人が様々に定義



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, ...

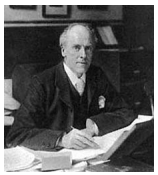
再掲:近代統計学(19世紀以降)

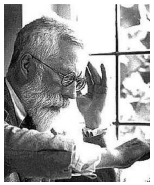
- 3つの流れを「統計」として整理
  - 社会統計に確率論を導入(A. Quetelet「近代統計学の父」)
    - 社会現象・自然現象を数量的にとらえる


Adolphe Quetelet 1796-1874

- 記述統計学(K. 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
- 推測統計学(R. 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

この講義で扱う統計学


Karl Pearson 1857-1936


Ronald Fisher 1890-1962

全数調査と標本調査

- **全数調査**
  - 母集団に含まれる全対象をデータとして観測
    - 例：国勢調査、東北大学経済学部生の出身高校
  - 長所：母集団の知りたい対象を網羅
  - 短所：母集団が大きい場合、費用や手間が膨大
- **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
    - 例：世論調査、選挙での当確予想
  - 長所：全数調査と比べて、少ない費用・手間で実施可能
  - 短所：<u>選択バイアス</u>（標本の偏り）が生じる可能性

標本調査と選択バイアス

- **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
- **大きな選択バイアスが生じる例：**
  - 仙台市立A高校3年生の平均身長の調査
    標本を女子のみから採取 ⇒ 平均身長を小さく推定
  - 日本の納税者の平均所得の調査
    標本を東京都港区のみから採取 ⇒ 平均所得を大きく推定
    - 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

選択バイアスの実例

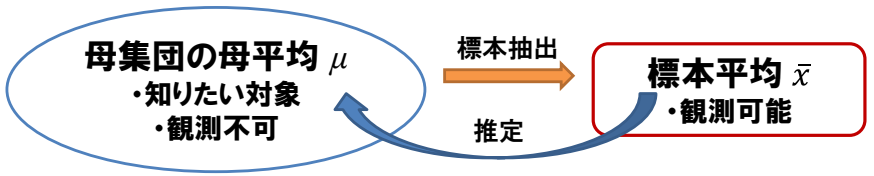
- 1936年　米国大統領選挙の結果予測
  - 共和党：Landon 氏　vs　民主党：Roosevelt 氏
- The Literary Digest
  - 老舗総合雑誌（過去5回の大統領選的中）
  - 雑誌の読者200万人に調査
  - 予測：得票率57％で Landon 氏の勝利

- **アメリカ世論研究所**（George Gallup）
  - 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

- **結果：得票率60％でRoosevelt 氏の勝利**

標本の代表値

- **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<u>母集団の特性</u>**
  - 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）

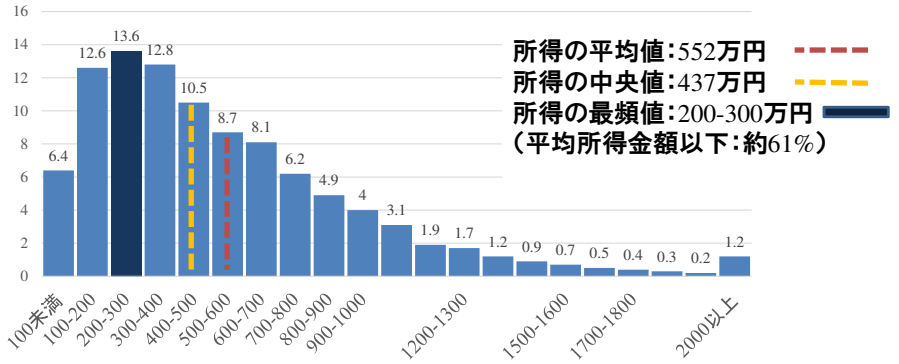


標本の平均, 中央値, 最頻値

- 標本 $\{x_1, x_2, \cdots, x_n\}$ の“真ん中”を表す代表値
  - 標本平均　$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
  - 中央値(メジアン)　データを大小の順に並べた真ん中の値
  - 最頻値(モード)　データの中に最も多くあらわれた値
    - 最頻値はただ一つに定まるとは限らない
- 問題
  - 標本{1, 1, 2, 2, 2, 4, 4, 5, 6}の平均、中間値、最頻値を求めなさい
  - 標本{1, 1, 2, 3, 4, 4}の平均、中間値、最頻値を求めなさい

適切な代表値は？

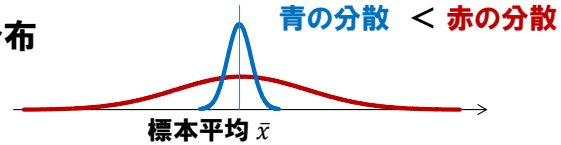
- 問題
  - 2019年調査(2020年はCovid-19の影響で調査中止)の日本の世帯別所得の平均値は552万円である。多くの世帯で552万円の所得があるとして政策等を決定することは妥当であるか？



厚生労働省：2019年国民生活基礎調査の概況　より

分散と標準偏差

- データのばらつきの大きさの代表値
  - 標本分散　$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
    - 推定や仮説検定で重要な役割
  - 例:データの分布



  - 例:推測結果の当たり方

経済統計

- 官庁が行う統計調査の結果
  - 社会全体で利用されるべき情報基盤
  - 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: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
- 「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、国民経済の健全な発展及び国民生活の向上に寄与する」(総務省統計局HPより)

物価指数

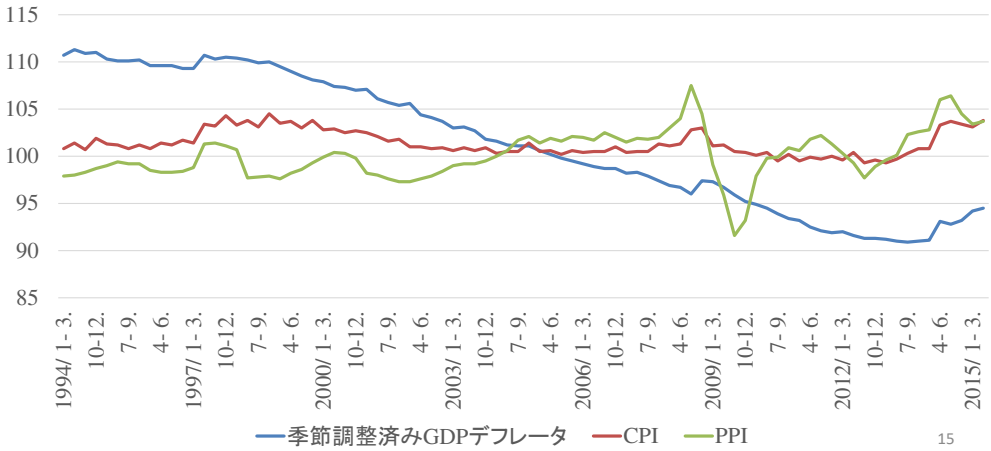
- 消費者物価指数(CPI)
  - 経済の体温計。総務省
  - 582品目の調査(2020年基準、5年毎更新)

- 国内企業物価指数(PPI)
  - 企業間取引の価格動向。日本銀行
  - 746品目の調査(2015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

3つの物価指数

- 各"物価指数"でそれぞれの傾向が異なる。
  - 例:2003~05年頃:CPI(横ばい)、PPI(上昇)、GDP(下降)



景気動向指数

- 先行系列
  - 景気を先取り傾向(11種)
  - 鉱工業生産財在庫率指数、新規求人数、東証株価指数など
- 一致系列
  - 景気の現状を示す傾向(10種)
  - 鉱工業生産指数、営業利益、有効求人倍率(除学卒)など
- 遅行系列
  - ある程度時間がたった傾向(9種)
  - 法人税収入、完全失業率、第3次産業活動指数など

(内閣府　第11次改訂:平成27年度)
内閣府HP　http://www.esri.cao.go.jp/jp/stat/di/di3.html#link000

景気動向指数(DIとCI)

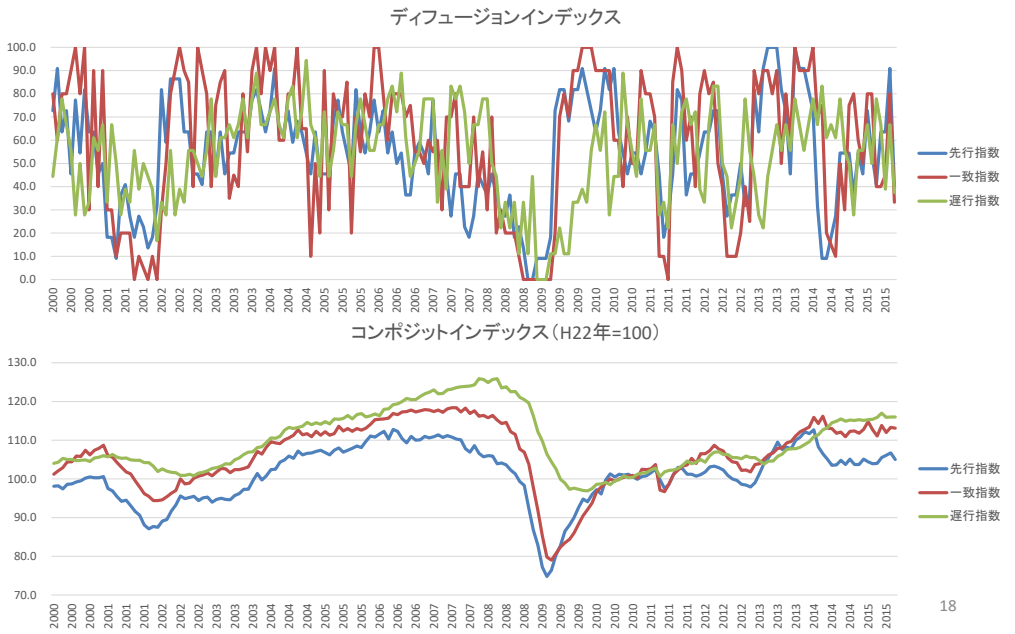
- **ディフュージョン・インデックス(DI)**
  - **定性的指標　景気拡張の動きの各経済部門への波及度**
  - DI = **拡張系列数 / 採用系列数** × 100%
    ※採用系列の3ヶ月前の値と比較して
    増加は"1"、変化なし"0.5"、減少 "0"をつけ、重みを算出
  - 50%**を目安に景気の方向性を判断**

- **コンポジット・インデックス(CI)**
  - **定量的指標　景気変動の相対的大きさ、テンポ**
  - **各系列の前月との変化率を過去5年分の平均、分散、標準化変化率と先行・一致・遅行の各指標の平均より算出**

- **指数が上昇⇒景気拡張局面。下降⇒後退局面**

内閣府HP参照:http://www.esri.cao.go.jp/jp/stat/di/di3.html#link000

景気動向指数(DIとCI)



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**