

第1章

データの整理と簡約化

入門 統計学 (P.1~37)

橋本 智雄 著

発表者 吉富 楓雅

目次

- 統計学を始める
- データの特性値
- 相関の理論

統計学を始める



統計学の歴史

- 歴史自体は浅い
 - * 250年程と、他の諸科学に比べると比較的新しい
 - ▶ 源流は経済学・社会学・生物学・数学など自然科学系
- 統計学は生のデータを通じてほぼ全ての科学と関わり発展してきた

統計学とは何か？

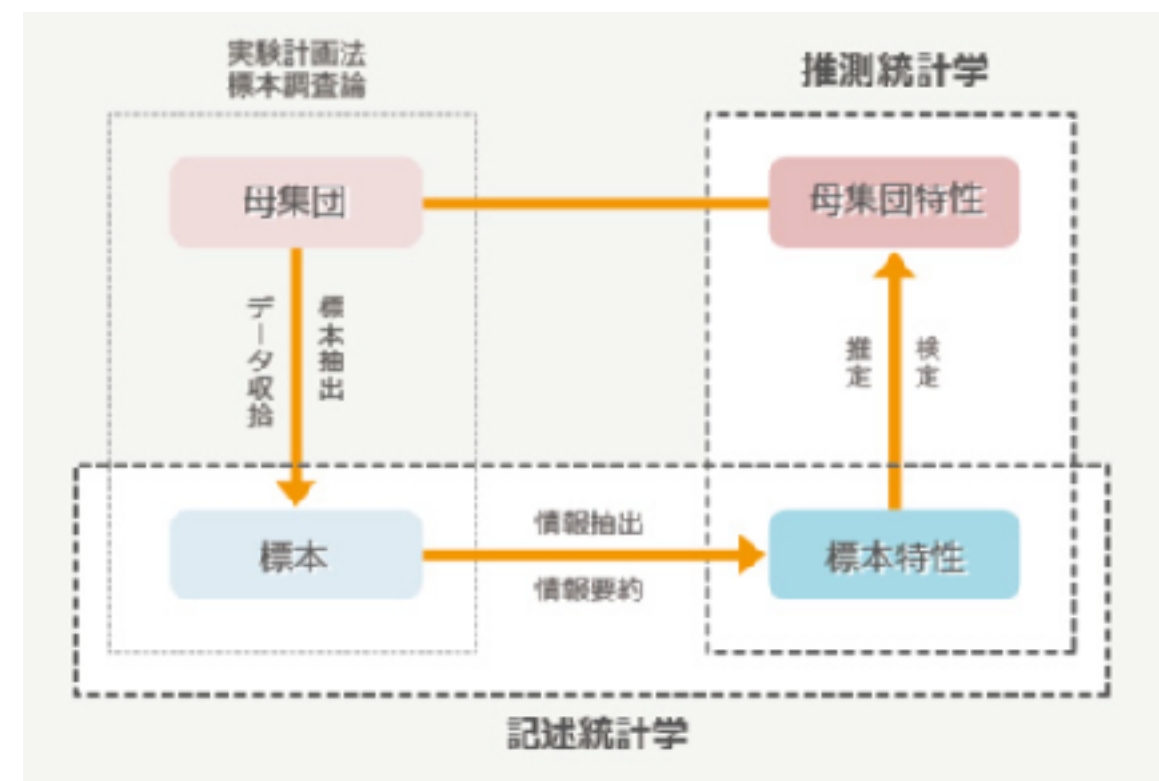
- 元の大きなデータの性質を推測したりするための方法論を体系化したもの
 - * ある程度以上の数のバラツキ（不確実性）のあるデータの性質を調べる
 - * 大きなデータ（母集団）から一部を抜き取って、その抜き取ったデータ（標本）の性質を調べる

統計学の種類

- 統計学の種類

1. 記述統計学（古典統計学）

2. 推測統計学（近代統計学）



統計学の分類

データの特性値



代表値

- 平均値
- メディアン（中央値）

$$Me = \begin{cases} x_{(\frac{n+1}{2})} & (n = \text{奇数}) \\ \frac{1}{2} \{ x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)} \} & (n = \text{偶数}) \end{cases}$$

- モード（最頻値）
- その他

* 最大値・最小値など

分散と標準偏差

- 分散 (Variance)

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

- 標準偏差 (Standard deviation)

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

平均値, 分散の計算法

- データの総数 n が大きくなったり, 測定値そのものの数値が大きかったりすると平均値・分散を定義から直接手計算するのは面倒

変量 x_i を変量 $u_i = \frac{x_i - x_0}{a}$ (a, x_0 は定数で $a > 0$)に

変換する時, 次の式が成立する↵

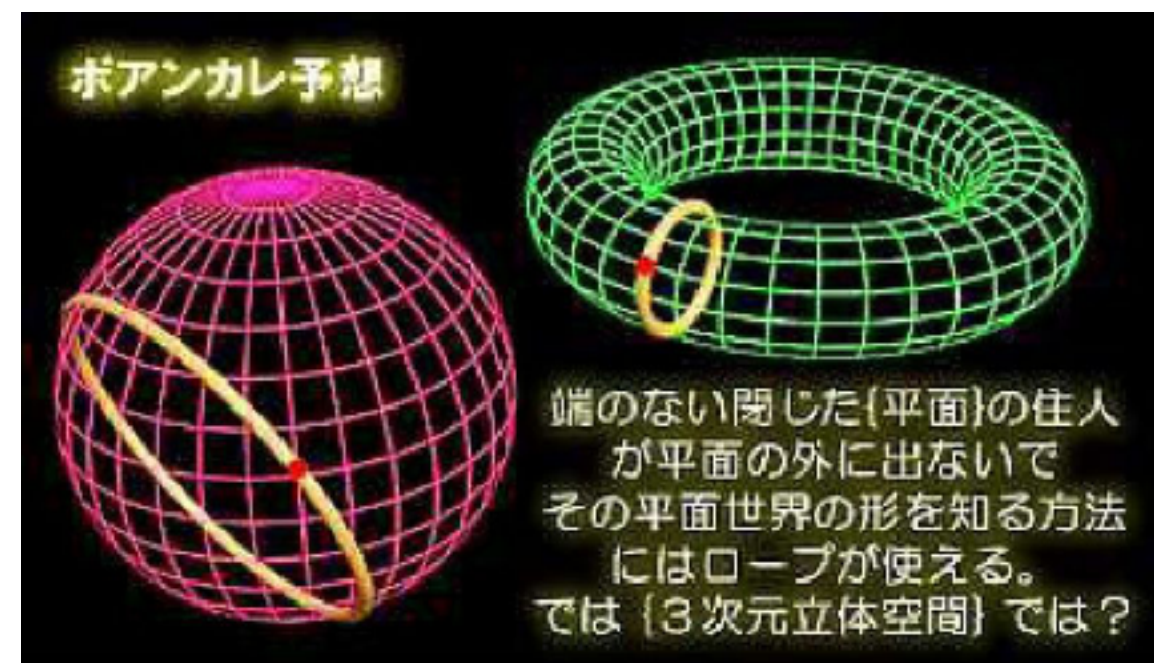
↵

$$(1) \bar{x} = a\bar{u} + x_0 \quad (\bar{u} \text{は} u_1, u_2, \dots, u_n \text{の} \text{平均値}) \quad \text{↵}$$

$$(2) S^2 = a^2 S_u^2 \quad (S_u \text{は} u_1, u_2, \dots, u_n \text{の} \text{分散}) \quad \text{↵}$$

- a は度数分布の階級の幅, x_0 は度数の多い階級値をとる

相関の理論



相関とは

- 世の中のデータは2変数などの多元的な情報の方がはるかに多い

* 例として

▶ 身長と体重・資本金と売上高

- 散布図

* 平面上にxy座標をとり,点をプロットすることで得られる図

散布図の種類

散布図

正の相関がある



正の相関がありそう



相関がない



負の相関がありそう



負の相関がある



回帰直線

- データが1つの直線 $y=ax + b$ の周りに集まっていると考える
 - * 一方の変量 x が与えられた時他方の変量 y を予測することができる
 - ▶ a と b を求めることができれば良い
- 最小2乗法を用いて a と b を求める
 - * データの各点から直線 $y=ax + b$ までの鉛直線の距離の2乗の和

$$\sum_{i=1}^n d_i^2 = \sum_{i=1}^n \{y_i - (ax_i + b)\}^2 = f(a, b)$$

を最小にする a と b を定める

傾きと切片の求め方

$f(a, b)$ は a, b についての2次式であることに注意し, $\sum x_i = n\bar{x}$, $\sum y_i = n\bar{y}$ を用いると次のように変形される

$$\begin{aligned} f(a, b) &= \sum \{y_i - (ax_i + b)\}^2 \\ &= nb^2 - 2n(\bar{y} - a\bar{x})b + (\sum x_i^2)a^2 \\ &\quad - 2\sum(x_i y_i)a + \sum y_i^2 \end{aligned}$$

$$\begin{aligned} &= n\{b - (\bar{y} - a\bar{x})\}^2 + (\sum x_i^2 - n\bar{x}^2)a^2 \\ &\quad - 2(\sum(x_i y_i - n\bar{x}\bar{y}))a + \sum y_i^2 - n\bar{y}^2 \\ &= n\{b - (\bar{y} - a\bar{x})\}^2 + (\sum x_i^2 - n\bar{x}^2) \left\{ a - \frac{\sum(x_i y_i - n\bar{x}\bar{y})}{\sum x_i^2 - n\bar{x}^2} \right\}^2 + \sum y_i^2 - n\bar{y}^2 - \frac{\sum(x_i y_i - n\bar{x}\bar{y})^2}{\sum x_i^2 - n\bar{x}^2} \end{aligned}$$

これより

$$b = \bar{y} - a\bar{x}, \quad a = \frac{\sum(x_i y_i - n\bar{x}\bar{y})}{\sum x_i^2 - n\bar{x}^2}$$

と定めると平方和 $\sum d_i^2$ は最小になる

相関係数

- 直線的傾向の強さを示す値 (R)
- R=0の場合は無相関であり,データはいかなる直線的傾向も示さない

$$R = \frac{S_{XY}}{S_X S_Y} = \frac{1}{n} \sum_{i=1}^n \left(\frac{X_i - \bar{x}}{S_X} \right) \left(\frac{Y_i - \bar{y}}{S_Y} \right) \leftarrow$$
$$(-1 \leq r \leq 1) \leftarrow$$