

第2章

1つの変数の記述統計

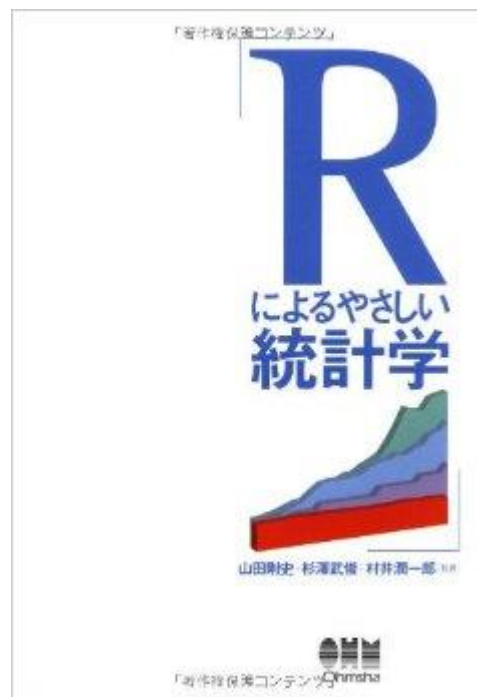
B3 松尾 大典

資料

「Rによるやさしい統計学」

著者

- 山田剛史
- 杉澤武俊
- 村井潤一郎



この章の概要

1つの変数をどう記述するか、の視点から
データ解析の出発点の作業について説明する

数値要約

- データの特徴を1つにまとめたもの
例) テストの点数をただ並べただけでは良し悪しが分かりにくいので、平均をとって分かりやすくする

変数の種類

質的変数

- データを構成する物を分類する変数

量的変数

- 大きさの大小を表す変数

EX

- 二値変数

- 選択肢が 2 つの変数

- 質的変数と量的変数は性質の違いから区別される

データの視覚的表現

度数

- カテゴリのデータの個数

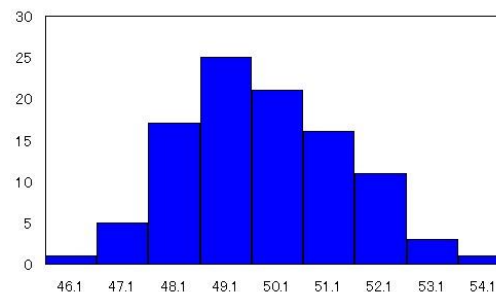
得点 (点)	人数(人)
40~50	4
50~60	8
60~70	13
70~80	21
80~90	11
90~100	3
計	60

度数分布

- 全てのカテゴリについて度数を纏めたもの

ヒストグラム

- 棒と棒の隙間のない図



代表値

- 分布の中心的位置を示すデータ
- 平均を代表値とするのが一般的
- 平均
 - データを合計して個数で割ったもの
 - データ中に他のデータから大きく離れた値（外れ値）がある場合は大きな影響を受ける

平均以外の代表値

中央値

- データを大きさ順に並べた場合に中央にくる値
- データの数が偶数だと一般に2つの数の平均を中央値とする

最頻値

- データ中で最も頻繁に現れる値

EX

- データ中に外れ値がある場合は中央値を代表値とする
- 量的変数の代表値は平均値、次に中央値をよく使う
- 質的変数の代表値は数値化できないので最頻値を使う

散布度

散布度はデータの特徴を示す数値で
データがどの程度バラバラなのかの指標
→分散と標準偏差

分散

- 分散 =
$$\frac{(\text{データ 1} - \text{平均})^2 + (\text{データ 2} - \text{平均})^2 + (\text{データ 3} - \text{平均})^2 + \dots}{n}$$
- 標本分散
 - 偏差の二乗の和をデータの個数nで割ったもの
 - 全てのデータを持っている状態で、データの散らばりを知りたいときに用いる
- 不偏分散
 - 偏差の二乗の和をデータの個数n-1で割ったもの
 - 全てのデータを持っていない状態で、母集団を想定して値を推測するときに用いる

標準偏差

- 標準偏差 = $\sqrt{\text{分散}}$
- 一般に散布度の指標
- 値が大きいほどデータの散らばりが大きい

分散、標本偏差以外の散布度

- 平均偏差
 - 「平均からの偏差」の絶対値の平均
- 範囲（レンジ）
 - データ中の最大のものと最小のものの差

以上の散布度の表し方があるが、普通は分散と標準偏差を用いる

標準化

- データを式で変換して、平均値と標準偏差を目的の値に変えること
- 異なるデータ同士を比較するときに標準化しておくことで差異が分かりやすくなる

標準得点

- 標準化で変換された値

z 得点

- 平均を 0, 標準偏差を 1 になるよう変換した標準得点
- コンピュータの端数処理の関係(丸め誤差)で平均がきっちり 0 にならないことがある

$$z\text{得点} = \frac{\text{データの値} - \text{平均}}{\text{標準偏差}}$$

偏差値

- データを平均 50、標準偏差を 10 になるように変換した標準得点

$$\text{偏差値} = z\text{得点} \times 10 + 50$$

まとめ

- 今回は 1 つの変数の扱い方を説明
- 代表値は変数の性質と外れ値に依存する
- 散布度は分散と標準偏差で示す
- データを標準化して捉えやすくする