

# 第2回 データの整理 (2)

村澤 康友

2020年4月28日

## 今日のポイント

1. データの記述統計量の復習 (平均・分散・標準偏差・共分散・相関係数).
2. 考察の対象全体を母集団, 母集団のうち実際に観察される部分を標本という. 標本から母集団について推測することを統計的推測という.
3. どの個体も等確率で取り出される標本抽出を無作為抽出という. 本格的な標本調査では, 単純無作為抽出より効率的な無作為抽出 (層化2段抽出など) を用いる.

## 目次

1	データの種類	1
1.1	個票データと集計データ (p. 16) . . .	1
1.2	横断面データと時系列データ (p. 16)	1
1.3	1変量データと多変量データ . . . . .	2
2	1変量データの整理	2
2.1	度数分布 (p. 17) . . . . .	2
2.2	記述統計量 (p. 18) . . . . .	2
3	2変量データの整理	2
3.1	共分散 (p. 27) . . . . .	2
3.2	相関係数 (p. 27) . . . . .	3
3.3	相関と因果 (p. 28) . . . . .	3
4	母集団と標本	3
4.1	記述統計学と推測統計学 (p. 23) . . .	3
4.2	母集団と標本 (p. 22) . . . . .	3

4.3	全数調査と標本調査 (p. 22) . . . . .	3
5	標本抽出法	3
5.1	標本抽出 . . . . .	3
5.2	無作為抽出 (p. 24) . . . . .	4
5.3	層化抽出 (p. 25) . . . . .	4
5.4	集落抽出 . . . . .	4
5.5	2段抽出 (p. 25) . . . . .	4
6	今日のキーワード	4
7	次回までの準備	4

## 1 データの種類

### 1.1 個票データと集計データ (p. 16)

定義 1. 調査における個別の調査票を個票という.

定義 2. 調査対象の個別のデータを個票データという.

例 1. 個別の学生のテストの点数.

定義 3. 個票データを集計したデータを集計データという.

例 2. 学生全体のテストの平均点.

### 1.2 横断面データと時系列データ (p. 16)

定義 4. 複数の個体についてある時点で記録したデータを横断面データという.

例 3. あるクラスの学生全員のテストの点数.

定義 5. 1つの個体について時間を通じて記録したデータを時系列データという.

例 4. ある学生のテストの点数の推移.

定義 6. 複数の個体について時間を通じて記録したデータをパネル・データという.

例 5. あるクラスの学生全員のテストの点数の推移.

### 1.3 1変量データと多変量データ

定義 7. 1つの変量を各個体について観測したデータを1変量データという.

例 6. テストの点数 (のみ).

定義 8. 複数の変量を各個体について観測したデータを多変量データという.

例 7. テストの点数と勉強時間.

注 1. 因果関係の分析には多変量データが必要.

## 2 1変量データの整理

### 2.1 度数分布 (p. 17)

まず最初に観測値の範囲をいくつかの階級に分割する.

定義 9. ある階級に含まれる観測値の数を, その階級の度数という.

定義 10. (度数) / (観測値の総数) を相対度数という.

定義 11. 横軸に値をとり, 各階級の(相対)度数を柱の面積で表したグラフをヒストグラム(柱状グラフ)という.

注 2. 棒の高さで表す棒グラフとは異なる.

注 3. ヒストグラムの印象は階級の取り方により異なる. 粗すぎても細かすぎてもダメ.

定義 12. ある階級以下の度数の和を, その階級までの累積度数という.

定義 13. (累積度数) / (観測値の総数) を累積相対度数という.

定義 14. 累積(相対)度数の折れ線グラフを累積(相対)度数グラフという.

注 4. 階級が細かいほど滑らかなグラフとなる.

### 2.2 記述統計量 (p. 18)

1変量データを  $(x_1, \dots, x_n)$  とする.

定義 15. (観測値の総和) / (観測値の総数) を(算術)平均という.

注 5. 式で表すと

$$\mu := \frac{1}{n} \sum_{i=1}^n x_i$$

定義 16. 平均からの偏差の2乗の平均を分散という.

注 6. 式で表すと

$$\sigma^2 := \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$$

定理 1.

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \mu^2$$

証明. 復習テスト. □

定義 17. 分散の平方根を標準偏差という.

定理 2. データを  $y_i := ax_i + b$  と一次変換すると,

$$\begin{aligned} \mu_y &= a\mu_x + b \\ \sigma_y^2 &= a^2\sigma_x^2 \end{aligned}$$

ただし  $\mu_x, \mu_y$  は平均,  $\sigma_x^2, \sigma_y^2$  は分散を表す.

証明. 復習テスト. □

## 3 2変量データの整理

### 3.1 共分散 (p. 27)

2変量データを  $((x_1, y_1), \dots, (x_n, y_n))$  とする.

定義 18. 各変量の平均からの偏差の積の平均を共分散という.

注 7. 式で表すと

$$\sigma_{xy} := \frac{1}{n} \sum_{i=1}^n (x_i - \mu_x)(y_i - \mu_y)$$

注 8.  $x_i$  が大きいと  $y_i$  も大きいなら共分散は正,  $x_i$  が大きいと  $y_i$  は小さいなら共分散は負, 「無関係」なら 0 となる.

定理 3.

$$\sigma_{xy} = \frac{1}{n} \sum_{i=1}^n x_i y_i - \mu_x \mu_y$$

証明. 復習テスト. □

### 3.2 相関係数 (p. 27)

定義 19. 変量の値から平均を引き, 標準偏差で割る変換を標準化という.

注 9. 式で表すと

$$z_i := \frac{x_i - \mu_x}{\sigma_x}$$

注 10. 標準化した変量の平均は 0, 分散は 1 となる.

定義 20. 標準化した 2 変量の共分散を相関係数という.

注 11. 式で表すと

$$\begin{aligned} \rho_{xy} &:= \frac{1}{n} \sum_{i=1}^n \left( \frac{x_i - \mu_x}{\sigma_x} - \frac{1}{n} \sum_{i=1}^n \frac{x_i - \mu_x}{\sigma_x} \right) \\ &\quad \left( \frac{y_i - \mu_y}{\sigma_y} - \frac{1}{n} \sum_{i=1}^n \frac{y_i - \mu_y}{\sigma_y} \right) \\ &= \frac{1}{n} \sum_{i=1}^n \frac{x_i - \mu_x}{\sigma_x} \frac{y_i - \mu_y}{\sigma_y} \end{aligned}$$

注 12. 「関係」が強いほど 1 か -1 に近くなる.

定理 4.

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

証明. 復習テスト. □

### 3.3 相関と因果 (p. 28)

因果関係があれば相関が生じる. 逆に相関があっても因果関係があるとは限らない.

定義 21. 因果関係のない相関を見かけ上の相関という.

注 13. 2 変量の原因となる第 3 の変量が存在する場合に生じる.

例 8. 親の所得と子どもの学力.

## 4 母集団と標本

### 4.1 記述統計学と推測統計学 (p. 23)

定義 22. データ整理の手法の体系を記述統計学という.

注 14. 大量観察による法則の発見を目的とする.

定義 23. 一部の観察から全体について推測することを統計的推測という.

定義 24. 統計的推測の理論体系を推測統計学という.

### 4.2 母集団と標本 (p. 22)

定義 25. 考察の対象全体を母集団という.

例 9. 日本国民の有権者全体.

定義 26. 母集団のうち実際に観察される部分を標本という.

注 15. 標本から母集団について推測するのが統計的推測.

### 4.3 全数調査と標本調査 (p. 22)

定義 27. 母集団全体を調査することを全数調査という.

例 10. 国勢調査.

定義 28. 標本を調査することを標本調査という.

例 11. 世論調査.

## 5 標本抽出法

### 5.1 標本抽出

定義 29. 母集団から標本を取り出すことを標本抽出という.

定義 30. 標本に含まれる個体の数を標本の大きさという.

注 16.  $n$  個の個体を含む標本は大きさ  $n$  の 1 つの

標本であり、 $n$  個の標本ではない。

## 5.2 無作為抽出 (p. 24)

**定義 31.** どの個体も等確率で取り出される抽出を無作為抽出という。

**定義 32.** どの個体の組合せも等確率で取り出される抽出を単純無作為抽出という。

## 5.3 層化抽出 (p. 25)

**定義 33.** 母集団に関する事前情報を補助情報という。

**例 12.** 国勢調査による居住地・性別・生年月・婚姻状態・学歴・就業状態・職業。

**定義 34.** 補助情報で分類した部分母集団を層という。

**定義 35.** 母集団を層に分けることを層化という。

**定義 36.** 母集団を層化し、層ごとに抽出する方法を層化抽出という。

**定義 37.** 各層から単純無作為抽出する層化抽出を層化無作為抽出という。

**例 13.** 男女の各層から同人数を単純無作為抽出。

## 5.4 集落抽出

**定義 38.** 複数の個体から成る抽出単位を集落という。

**例 14.** 市町村。

**定義 39.** 集落を抽出する方法を集落抽出という。

**定義 40.** 集落を単純無作為抽出する方法を単純集落抽出という。

**定義 41.** 各抽出単位の抽出確率を何かに比例させる方法を確率比例抽出という。

**定義 42.** 各集落の抽出確率を集落の大きさに比例させる方法を確率比例集落抽出（規模比例確率抽出）という。

## 5.5 2 段抽出 (p. 25)

**定義 43.** まず集落を抽出し、次に各集落から抽出する方法を2段抽出という。

**定義 44.** 第1段を単純集落抽出する2段抽出を単純2段抽出という。

注 17. 第2段は比例配分で単純無作為抽出。

**定義 45.** 第1段を確率比例集落抽出する2段抽出を確率比例2段抽出という。

注 18. 第2段は同数配分で単純無作為抽出。

**定義 46.** 母集団を層化し、層ごとに2段抽出する方法を層化2段抽出という。

**例 15.** 市町村を規模別に層化して、まず市町村を抽出し、次に各市町村から抽出。

## 6 今日のキーワード

個票, 個票データ, 集計データ, 横断面データ, 時系列データ, パネル・データ, 1 変量データ, 多変量データ, 度数, 相対度数, ヒストグラム (柱状グラフ), 累積度数, 累積相対度数, 累積 (相対) 度数グラフ, (算術) 平均, 分散, 標準偏差, 共分散, 標準化, 相関係数, 見かけ上の相関, 記述統計学, 統計的推測, 推測統計学, 母集団, 標本, 全数調査, 標本調査, 標本抽出, (標本の) 大きさ, 無作為抽出, 単純無作為抽出, 補助情報, 層, 層化, 層化抽出, 層化無作為抽出, 集落, 集落抽出, 単純集落抽出, 確率比例抽出, 確率比例集落抽出 (規模比例確率抽出), 2 段抽出, 単純 2 段抽出, 確率比例 2 段抽出, 層化 2 段抽出

## 7 次回までの準備

提出 宿題 1

復習 教科書第 2 章, 復習テスト 2

予習 教科書第 3 章 1-2 節