第 2 回　データの整理（2）

村澤 康友

2023 年 10 月 3 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．どの個体の組合せも等確率で取り出される標本抽出を単純無作為抽出という．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　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. 勉強時間とテストの点数.

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 $(x_1, \ldots, x_n)$ 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,

$$\mu_y = a\mu_x + b$$
$$\sigma_y^2 = a^2\sigma_x^2$$

ただし μ_x, μ_y は平均，σ_x^2, σ_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 $((x_1, y_1), \ldots, (x_n, y_n))$ 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

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

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（因果関係がなくても相関は生じうる）.

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標本抽出を**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標本抽出を**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・職業の分布の情報.

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，各層から個体を抽出する方法を**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　系統抽出

定義 43. 抽出枠から一定の間隔で個体を抽出する方法を**系統抽出**という.

注 17. 一定の間隔で並ぶ個体の集まりを 1 つの集落とした集落抽出.

注 18. 抽出枠を無作為に並べれば（非復元）単純無作為抽出と同等.

5.6　2 段抽出（p. 25）

定義 44. まず集落を抽出し，次に各集落から個体を抽出する方法を *2* **段抽出**という.

定義 45. 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 19. 第 2 段は比例配分で単純無作為抽出.

定義 46. 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 20. 第 2 段は同数配分で単純無作為抽出.

定義 47. 母集団を層化し，各層から 2 段抽出する方法を**層化** *2* **段抽出**という.

例 15. 全国を市町村に層化し，各市町村から調査区・個体の順に 2 段抽出.

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），系統抽出，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, 4 節