

第5章 データの分析

2.5 データの整理, データの代表値

① データ

運動の記録や所属クラスなどのように, ある特性を表す数量を **変量** といい, ある変量の測定値や観測値, 調査結果などの集まりを **データ** という。
また, データにおける測定値や観測値の個数を, そのデータの **大きさ** という。

② 度数分布

度数分布表 データのとり値をいくつかの区間に区切って, 各区間に含まれるデータの値の個数をまとめた表。
度数分布表において, 区切られた区間を **階級**, 区間の幅を **階級の幅**, 各階級に含まれる値の個数を **度数** という。
また, 各階級の真ん中の値を **階級値** という。

ヒストグラム 度数分布表を柱状のグラフで表したもの。

③ 代表値

平均値 変量 x のデータの値が x_1, x_2, \dots, x_n であるとき, このデータの平均値 \bar{x} は
$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \dots + x_n) \quad \leftarrow (\text{平均値}) = \frac{(\text{データの値の総和})}{(\text{データの大きさ})}$$
最頻値 (モード) データにおいて, 最も個数の多い値。
データが度数分布表に整理されているときは, 度数が最も大きい階級の階級値。
中央値 (メジアン) データを値の大きさの順に並べたとき, 中央の位置にくる値。
データの大きさが偶数のときは, 中央の2つの値の平均値。

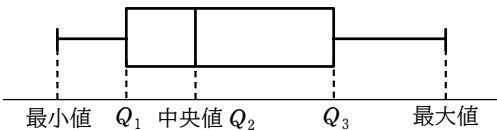
2.6 データの散らばりと四分位数

① 四分位数

範囲 データの最大値から最小値を引いた差。
データの範囲が大きいほど, 散らばりの度合いが大きいと考えられる。
四分位数 データの値を大きさの順に並べたとき, 4等分する位置の値。
小さい方から順に, **第1四分位数**, **第2四分位数**, **第3四分位数** といい, 順に Q_1, Q_2, Q_3 で表す。
第2四分位数 Q_2 はデータの中央値である。
四分位範囲 データの第3四分位数 Q_3 から第1四分位数 Q_1 を引いた差 $Q_3 - Q_1$ 。
データの四分位範囲が大きいほど, 散らばりの度合いが大きいと考えられる。
補足 四分位範囲を2で割った値を **四分位偏差** という。

② 箱ひげ図

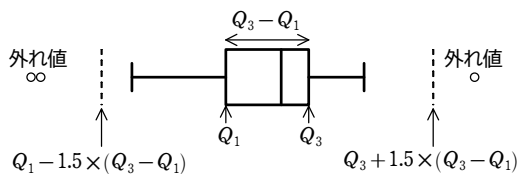
箱ひげ図 データの分布を, 次の5つの値で表した図。
平均値を記入することもある。
最小値, 第1四分位数 (Q_1), 中央値 (Q_2), 第3四分位数 (Q_3), 最大値



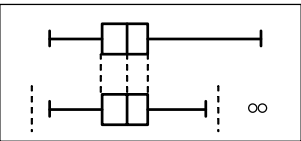
外れ値 データの中に, 他の値から極端に離れた値が含まれるとき, そのような値を **外れ値** という。
外れ値の基準は複数あるが, 四分位範囲を利用した外れ値の基準として, 次のものがある。

$\{(\text{第1四分位数}) - 1.5 \times (\text{四分位範囲})\}$ 以下の値
 $\{(\text{第3四分位数}) + 1.5 \times (\text{四分位範囲})\}$ 以上の値

外れ値がある場合, 箱ひげ図において, 次の図のように外れ値を。などで表すことがある。また, 箱ひげ図の左右のひげは, データから外れ値を除いたときの最小値または最大値まで引く。



注 外れ値を。で表す箱ひげ図をかく場合でも, 四分位数は外れ値を除かないすべてのデータの四分位数であり, その値にもとづいて箱をかく。



2.7 分散と標準偏差

① 分散, 標準偏差

変量 x のデータの値を x_1, x_2, \dots, x_n , その平均値を \bar{x} とする。
偏差 変量 x のデータの各値から平均値 \bar{x} を引いた差 $x_1 - \bar{x}, x_2 - \bar{x}, \dots, x_n - \bar{x}$ 。
これを $x - \bar{x}$ で表す。
分散 偏差の2乗の平均値。 s^2 で表す。

$$s^2 = \frac{1}{n} \{ (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2 \} \quad \leftarrow (\text{分散}) = \frac{(\text{偏差の2乗の総和})}{(\text{データの大きさ})}$$

分散と平均値の関係式

$$(\text{\textit{x}のデータの分散}) = (\text{\textit{x}のデータの平均値}) - (\text{\textit{x}のデータの平均値})^2$$

標準偏差 分散の正の平方根。 s で表す。

$$s = \sqrt{\text{分散}} = \sqrt{\frac{1}{n} \{ (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2 \}}$$

データの値が平均値の周りに集中しているほど, それぞれの偏差の絶対値は小さくなり, 分散, 標準偏差ともに小さくなる傾向がある。

研究 変量の変換

a, b は定数とする。変量 x のデータから $y = ax + b$ によって新しい変量 y のデータが得られるとき, x, y のデータの平均値を \bar{x}, \bar{y} , 分散を s_x^2, s_y^2 , 標準偏差を s_x, s_y とすると $\bar{y} = a\bar{x} + b, s_y^2 = a^2s_x^2, s_y = |a|s_x$

補足 とくに, $u = \frac{x - x_0}{c}$ すなわち $x = cu + x_0$ (c, x_0 は定数) として新しい変量 u を作るとき, x_0 を **仮平均** という。

参考 変量 x について, その平均値 \bar{x} と標準偏差 s_x を用いて, $z = \frac{x - \bar{x}}{s_x}$ として

新しい変量 z を作るとき, 変量 $10z + 50$ を **偏差値** という。

変量 x のデータについて, あるデータの値 x_k の偏差値は

$$10 \times \frac{x_k - \bar{x}}{s_x} + 50$$

28 2つの変量の間の関係

1 散布図

散布図 2つの変量からなるデータを点として平面上に図示したもの。

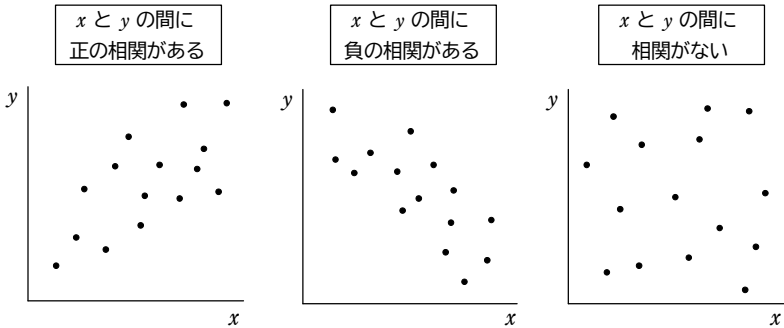
2 正の相関，負の相関

正の相関 2つの変量からなるデータにおいて，一方が増えると他方も増える傾向がみられるとき，2つの変量の間には 正の相関 があるという。

負の相関 2つの変量からなるデータにおいて，一方が増えると他方が減る傾向がみられるとき，2つの変量の間には 負の相関 があるという。

注 正の相関と負の相関のいずれの傾向もみられないときは，2つの変量の間には相関がないという。

補足 正の相関関係がある，負の相関関係がある，相関関係がない，ということもある。



2つの変量の間に正の相関あるいは負の相関があるとき，散布図における点の分布が1つの直線に接近しているほど 相関が強い といい，散らばっているほど 相関が弱い という。

3 相関係数

共分散 x の偏差と y の偏差の積の平均値。 s_{xy} で表す。

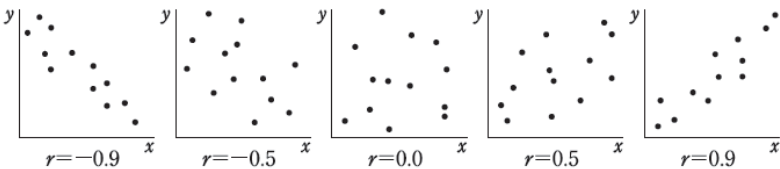
$$s_{xy} = \frac{1}{n} \{ (x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + \cdots + (x_n - \bar{x})(y_n - \bar{y}) \}$$

x と y の間に，正の相関があるとき共分散は正となり，
負の相関があるとき共分散は負となる。

相関係数 相関の正負と強弱を表す値。 r で表す。

$$\begin{aligned} r &= \frac{s_{xy}}{s_x s_y} \quad \leftarrow (\text{相関係数}) = \frac{(\text{x と y の共分散})}{(\text{x の標準偏差}) \times (\text{y の標準偏差})} \\ &= \frac{\frac{1}{n} \{ (x_1 - \bar{x})(y_1 - \bar{y}) + \cdots + (x_n - \bar{x})(y_n - \bar{y}) \}}{\sqrt{\frac{1}{n} \{ (x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2 \}} \sqrt{\frac{1}{n} \{ (y_1 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2 \}}} \\ &= \frac{(x_1 - \bar{x})(y_1 - \bar{y}) + \cdots + (x_n - \bar{x})(y_n - \bar{y})}{\sqrt{\{ (x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2 \} \{ (y_1 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2 \}}} \end{aligned}$$

相関係数 r については， $-1 \leq r \leq 1$ を満たす。
 r の値は，正の相関が強いほど 1 に近づき，負の相関が強いほど -1 に近づく。
また，相関がないとき， r の値は 0 に近い値をとる。



注 相関係数は，外れ値の影響を受けやすい値である。

4 相関関係と因果関係

2つの変量の間に相関があるからといって，一方が原因で他方が起こる因果関係があるとは断定できない。

5 質的データをとる2つの変量の間の関係

量的データ 身長や体重など，数値として得られるデータ。

質的データ 所属クラスや都道府県など，数値ではないものとして得られるデータ。

分割表 たとえば，合格が判定されるある試験において，受験者 100 人全員を対象に，教材 A を使用して学習したか調べるとする。2つの変量 (合格，使用したかどうか) の間の関係を調べて，右の図のようにまとめたとき，この表を分割表 または クロス集計表 という。

| | 合 | 否 | 計 |
|---------|----|----|-----|
| A の使用：有 | 9 | 5 | 14 |
| A の使用：無 | 42 | 44 | 86 |
| 計 | 51 | 49 | 100 |

29 仮説検定の考え方

1 仮説検定の考え方

得られたデータをもとに，ある主張が妥当かどうかを判断する，次のような手法を仮説検定 という。

- ① 妥当かどうか判断したい主張 [1] と，それに反する仮説 [2] を立てる。
また，基準となる確率を定める。
- ② 仮説 [2] のもとで，調査や実験の結果が起こる確率を調べる。
- ③ ② で求めた確率が，基準となる確率より小さければ，仮説 [2] が正しい可能性は低い，すなわち主張 [1] が妥当であると判断してよい。
- ④ ② で求めた確率が，基準となる確率より小さくなければ，仮説 [2] は否定できず，主張 [1] が妥当であるとは判断できない。

注 このとき，仮説 [2] が妥当であると判断できるわけではない。
補足 仮説検定において，妥当かどうかを判断したい主張 [1] に反する仮説として立てた主張 [2] を 帰無仮説 といい，主張 [1] を 対立仮説 という。

発展 仮説検定と反復試行の確率

1 仮説検定と反復試行の確率

1 回の試行で事象 A の起こる確率を p とする。この試行を n 回繰り返し行うとき，事象 A がちょうど r 回起こる確率は ${}_n C_r p^r (1 - p)^{n-r}$

補足 ${}_n C_r$ は，異なる n 個のものから r 個を取り出して作る組合せの総数を表す。
また，正の数 a に対して， $a^0 = 1$ と定める。

この「反復試行の確率」を用いて，上の「仮説検定の考え方」の ② の確率を計算することができる。