第５章　データの分析

２５　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。

2　度数分布

度数分布表　データのとる値をいくつかの区間に区切って階級を定め，各階級に度数を対応させた表。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，……，x_n であるとき，このデータの平均値 $\overline{x}$ は　$\overline{x}=\dfrac{1}{n}(x_1+x_2+\cdots\cdots+x_n)$

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。

２６　データの散らばりと四分位数

1　四分位数

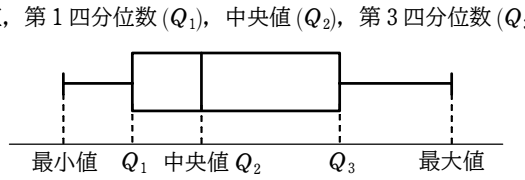
範囲　データの最大値から最小値を引いた差。

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

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。外れ値の基準として，たとえば次のようなものがある。
(第 1 四分位数 − 1.5 × 四分位範囲) 以下の値
(第 3 四分位数 + 1.5 × 四分位範囲) 以上の値

２７　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，……，x_n，その平均値を $\overline{x}$ とする。

偏差　変量 x のデータの各値から平均値 $\overline{x}$ を引いた差 $x_1-\overline{x}$，$x_2-\overline{x}$，……，$x_n-\overline{x}$
これを $x-\overline{x}$ で表す。

分散　偏差の 2 乗の平均値。s^2 で表す。

$$s^2=\frac{1}{n}\{(x_1-\overline{x})^2+(x_2-\overline{x})^2+\cdots\cdots+(x_n-\overline{x})^2\}$$

(x のデータの分散) = (x^2 のデータの平均値) − (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。

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

(x のデータの標準偏差) $=\sqrt{(x^2\text{ のデータの平均値})-(x\text{ のデータの平均値})^2}$

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 $\overline{x}$，$\overline{y}$，分散を ${s_x}^2$，${s_y}^2$，標準偏差を s_x，s_y とすると　$\overline{y}=a\overline{x}+b$，　${s_y}^2=a^2{s_x}^2$，　$s_y=|a|s_x$

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。

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。

2　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。

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

相関係数　相関の正負と強弱を表す値。r で表す。

$$r=\frac{s_{xy}}{s_x s_y}$$

$$=\frac{\frac{1}{n}\{(x_1-\overline{x})(y_1-\overline{y})+\cdots\cdots+(x_n-\overline{x})(y_n-\overline{y})\}}{\sqrt{\frac{1}{n}\{(x_1-\overline{x})^2+\cdots\cdots+(x_n-\overline{x})^2\}}\sqrt{\frac{1}{n}\{(y_1-\overline{y})^2+\cdots\cdots+(y_n-\overline{y})^2\}}}$$

$$=\frac{(x_1-\overline{x})(y_1-\overline{y})+\cdots\cdots+(x_n-\overline{x})(y_n-\overline{y})}{\sqrt{(x_1-\overline{x})^2+\cdots\cdots+(x_n-\overline{x})^2}\sqrt{(y_1-\overline{y})^2+\cdots\cdots+(y_n-\overline{y})^2}}$$

相関係数 r は $-1\leqq r\leqq 1$ であり，r が 1 に近いほど正の相関が強く，-1 に近いほど負の相関が強い。相関がないとき，r 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。

3　相関関係と因果関係

一般に，2 つの変量の間に相関関係があるからといって，必ずしも因果関係があるとはいえない。

4　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，合否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 (合否，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 [1] が正しい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 主張 [1] と反する仮定を立てる。(主張 [2] とする。)
② 主張 [2] のもとで，実際に起こった出来事が起こりにくい出来事かどうかを調べる。
③ ② で調べた結果，実際に起こった出来事は十分起こりにくいと判断するとき，主張 [2] の仮定は正しくないと判断できる。
④ 主張 [1] は正しいと判断してもよいと考えられる。

補足　③ で，実際に起こった出来事が十分起こりにくいと判断しないときは，主張 [2] の仮定は否定できず，主張 [1] は正しいと判断できない。このとき，主張 [2] が正しいと判断できるわけではない。

発展　**仮説検定の考え方と反復試行の確率**

1　組合せ

n 個から r 個取る組合せ (異なる n 個のものから異なる r 個を取り出して作る組合せ) の総数は　${}_nC_r=\dfrac{n(n-1)\cdots\cdots(n-r+1)}{r(r-1)\cdots\cdots\cdot 3\cdot 2\cdot 1}$　ただし，${}_nC_0=1$ とする。

${}_nC_r$ の性質　${}_nC_r={}_nC_{n-r}$

2　試行と事象

同じ条件のもとで繰り返すことができ，その結果が偶然によって決まる実験や観測を **試行** という。また，試行の結果として起こる事柄を **事象** という。

3　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　${}_nC_r p^r(1-p)^{n-r}$
ただし，正の数 a に対して，$a^0=1$ と定める。