## 高校生のための統計用語

変量（variable）　：ある集合に含まれる要素の性質を数量的に表現するもの

データ（data）　：観測，調査，実験などの結果として得られた変量の集まり

分布（distribution）　：ある変量について，その測定値の広がりの状態，散らばり方

代表値　：データの分布の状況を特徴的な１つの数値で表すときの値

平均値（mean value）　：変量 $x$ におけるデータが， $n$ 個の値 $x_1, x_2, \cdots, x_n$ であるとき，

それらの総数 $n$ で割ったもの

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_3}{n}$$

中央値（median）　：データをその値の大きさの順に並べたときに中央の位置にくる値

データの個数が偶数の場合は，最も中央に近い２個のデータの平均

最頻値（mode）　：データの中で最も個数の多い値

最大値（maximum value）：データの中で最も大きな値

最小値（minimum value）：データの中で最も小さな値

範囲（range）　：最大値と最小値の差

度数分布（frequency distribution）：量の大小の順で並べ，範囲を一定幅の小区間（＝階級）で区切り，

その階級に存在する度数で表現したもの

ヒストグラム　：階級幅を一定にとり，階級の度数がグラフの高さで表される棒グラフ

階級値　：各階級の真ん中の値（階級の最大値と最小値の平均）

相対度数　：$\dfrac{\text{階級の度数}}{\text{度数の合計}}$

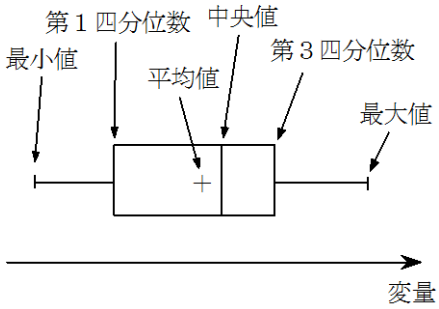
四分位数（quartie）　：データを昇順に並べ，それを４等分した個数で区切ったときに，境界に現れる値

第１四分位数（$Q_1$）　：中央値よりも小さいデータの中央値

第２四分位数（$Q_2$）　：中央値のこと

第３四分位数（$Q_3$）　：中央値よりも大きいデータの中央値

四分位範囲（IQR）　：第１四分位数 $Q_1$ と第３四分位数 $Q_3$ との差




四分位偏差（quartie deviation）：四分位範囲を２で割った値

箱ひげ図　：データの分布を可視化したもの

最小値，第１四分位数，中央値，第３四分位数，最大値を箱とひげで表したもの

平均値を「＋」で入れる場合と入れない場合がある

偏差　　：各データと平均値との差　$x_n - \overline{x}$

分散（variance）　　：データの散らばり具合を表す数値，偏差を２乗した値の平均値

$$\frac{\left(x_1-\overline{x}\right)^2+\left(x_2-\overline{x}\right)^2+\cdots+\left(x_n-\overline{x}\right)^2}{n}$$　　$s^2$ や $v$ で表すことが多い

$s^2$ は次のように求めることもできる

$$s^2=\overline{x^2}-\left(\overline{x}\right)^2$$　　（２乗の平均)ー(平均)$^2$

標準偏差（standard deviation)：分散の平方根　$s=\sqrt{\dfrac{\left(x_1-\overline{x}\right)^2+\left(x_2-\overline{x}\right)^2+\cdots+\left(x_n-\overline{x}\right)^2}{n}}$

相関（correlation）　：２つの変量の関係に着目して母集団の性質を定量化する手法

散布図（scattergram)：$x$ 軸，$y$ 軸に２つの変量を対応させ，各データをプロットしたグラフ

共分散（covarience）：$x$ の偏差と $y$ の偏差の積の平均　　$s_{xy}$ で表すことが多い

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

相関係数（correlation coefficient)：共分散 $s_{xy}$ を標準偏差 $s_x$, $s_y$ の積で割った値

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

（注１）和の記号Σを使えば，簡潔に表現できる。

$$\overline{x}=\frac{1}{n}\sum_{k=1}^{n}x_k\,,\quad s^2=\frac{1}{n}\sum_{k=1}^{n}\left(x_k-\overline{x}\right)^2,\quad s=\sqrt{\frac{1}{n}\sum_{k=1}^{n}\left(x_k-\overline{x}\right)^2}\,,\quad s_{xy}=\frac{1}{n}\sum_{k=1}^{n}\left(x_k-\overline{x}\right)\left(y_k-\overline{y}\right)$$

（注２）

偏差を単純に足し合わせると必ず０になることから，分散の計算には偏差の２乗の平均が使用されるが，

各データの値と平均値の差を絶対値を用いて表せば，偏差の平均は

$$\frac{\left|x_1-\overline{x}\right|+\left|x_2-\overline{x}\right|+\cdots+\left|x_n-\overline{x}\right|}{n}$$

となる。これを平均絶対偏差（MAD：mean absolute deviation）や平均偏差という。

絶対値を含む計算は，計算機では用意なので，こちらが使われることもある。