医学統計学講義：データの代表値

芳賀昭弘 *

1　データの代表値の計算

1.1　平均 mean

平均と言えば、普通は算術平均をいう。しかし、平均にはこの他にも幾何平均、加重平均などがある。

算術平均 Arithmetic mean；

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

幾何平均 (geometric mean)；（外れ値[*1]に若干強い。データが常に 0 より大きい場合に使うことができる）

$$m_g = (\prod_{i=1}^{n} x_i)^{1/n} \tag{2}$$

両辺の対数を取ると、その意味がわかるでしょう。

加重平均 weighted mean；

$$E_w[x] = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \tag{3}$$

w_i は x_i の重みを表す。$\boldsymbol{w}^T = (w_1, w_2, \cdots, w_n)$ と $\boldsymbol{x}^T = (x_1, x_2, \cdots, x_n)$ というベクトルを新たに定義すると、加重平均は

$$E_w[x] = \boldsymbol{w}^T \boldsymbol{x} \tag{4}$$

とも書ける（$\sum_{i=1}^{n} w_i = 1$ となるように規格化されているとした)。

更にもう少し深く考えてみよう。今 $\boldsymbol{w}$ の成分を全て足すと 1 であり、その成分が非負であるとすると、w_i は x_i の生起確率であるとみなすことができる。w_i は x_i に依存するので、x_i の生起確率を w_i の代わりに $P(x_i)$ とおいて、

$$E[x] = \sum_{i=1}^{n} P(x_i) x_i \tag{5}$$

という表現が導かれる。これは確率分布 $P(x)$ に対する x の期待値と一致する。

---

* Electronic address: haga@tokushima-u.ac.jp

*1 他のデータに比べ極端に大きい or 小さい値を持つデータ

確率分布 $P(x)$ は離散的で n 個の階級値を持つ時、

$$\sum_{i=1}^{n} P(x_i) = 1 \tag{6}$$

および

$$P(x_i) \geq 0, \ \ (n = 1, 2, \cdots n) \tag{7}$$

また、確率分布 $P(x)$ が区間 $[a, b]$ で定義された連続関数であるとき、

$$\int_a^b P(x)dx = 1 \tag{8}$$

および

$$P(x) \geq 0, \ \ (a \leq x \leq b) \tag{9}$$

という性質を持つ。

1.2 最大値 max と最小値 min

データの最大値、最小値もデータの代表値の１つである。

1.3 中央値 median 及び最頻値 mode

中央値はデータを小さい順（もしくは大きい順）に並べた時に中央に位置する値（データ数が偶数の場合は、中央２つの平均値）のこと。
ヒストグラムのデータで最大度数（最大頻度）を持つ階級値のことを、最頻値という。

1.4 四分位数 Quartile と四分位範囲

中央値より小さいデータ（A）と中央値より大きいデータ（B）の 2 群に分け、A のデータ内での中央値を第 1 四分位数、B のデータ内での中央値を第 3 四分位数と呼ぶ。第 3 四分位数から第 1 四分位数を引いた値を四分位範囲という。

1.5 分散 variance と標準偏差 standard deviation

$$s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{10}$$

は標本分散を与える。s を標本標準偏差という。
標本データから不偏分散 σ^2 を推定する場合は以下を用いる。

$$\sigma^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{11}$$

σ を不偏標準偏差という。

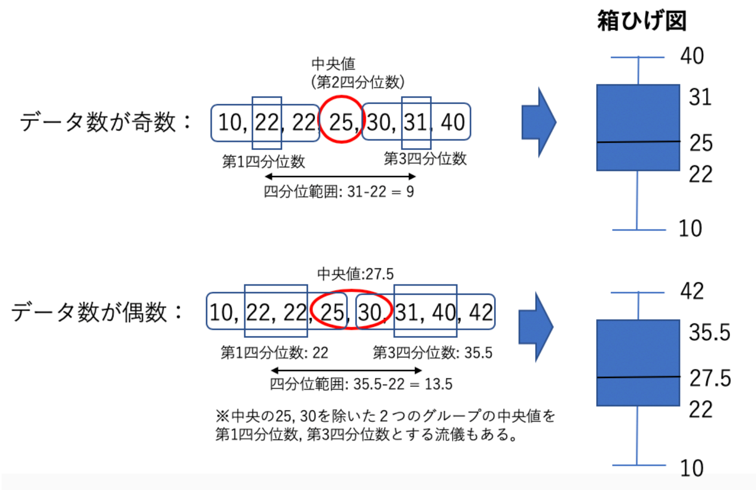

図１　中央値・四分位数と四分位範囲

データ x_i の生起確率分布が $P(x_i)$ である時を考えてみよう。$(x-\bar{x})^2$ の期待値は加重平均の時と同じように、

$$E[(x-E[x])^2] = \sum_{i=1}^{n} P(x_i)(x_i - E[x])^2 \tag{12}$$

となる。つまり、式（10）で与えられる分散は $P(x_i) = 1/n$ というようにn それぞれのデータの生起確率が一様である、という暗黙の仮定のもとで与えられている。では、期待値の計算で行ったのと同じように、データが発生する頻度が x によって異なる場合、分散はどのように表現できるであろうか。x に応じたデータの生起確率が確率分布 $P(x)$ で与えられている場合、x の分散を $V[x]$ と書くことにすると、

$$V[x] = E[(x-E[x])^2] = \sum_{i=1}^{n} P(x_i)x_i^2 - \sum_{i=1}^{n} P(x_i)2x_iE[x] + \sum_{i=1}^{n} P(x_i)E[x]^2 \tag{13}$$

$$= E[x^2] - 2E[x]\sum_{i=1}^{n} P(x_i)x_i + E[x]^2 \sum_{i=1}^{n} P(x_i) \tag{14}$$

$$= E[x^2] - E[x]^2 \tag{15}$$

より、

$$V[x] = E[x^2] - E[x]^2 \tag{16}$$

が成り立つことがわかる。$E[(x-E[x])^m]$ を m 次のモーメントといい、分散は２次のモーメントのことである。３次のモーメントは歪度 skewness、４次のモーメントは尖度 kurtosis という統計量の計算に使われる。

1.6　共分散 covariance

共分散 (covariance) は２つのデータ間の相関（関係性）を見る時に有効な指標となる。標本共分散 (covariance)；

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) \tag{17}$$

不偏共分散：

$$\sigma_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \tag{18}$$

で与えられる。

共分散が正のとき正の相関（一方が大きくなると一方も大きくなる)、負の時は負の相関（一方が大きくなると一方は小さくなる）があると言える。しかし、その値の大きさは、データで扱っている値の大きさに依存してしまうため、相関を見る場合には次の相関係数を用いる。

1.7 相関係数 cross correlation

相関係数（correlation coefficient）は2つの変量の関係を表す統計量で、

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{19}$$

と定義される。

$(x_i - \bar{x})$ が i 番目の成分となるベクトル $\boldsymbol{x}$、$(y_i - \bar{y})$ が i 番目の成分となるベクトル $\boldsymbol{y}$、を定義しよう。式(20）は次のように書けることがわかる；

$$r = \frac{\boldsymbol{x}^T\boldsymbol{y}}{|\boldsymbol{x}||\boldsymbol{y}|} \tag{20}$$

つまり、相関係数 r とは、同じ次元を持つ2つのベクトル $\boldsymbol{x}, \boldsymbol{y}$ のなす角 θ の余弦 $(\cos\theta)$ である。ベクトルが同じ方向を向くとき 1、直交するとき 0、反対を向くとき-1 となることが直感的にわかるであろう。