分布の特性値

確率変数の分布の特性値

確率変数Xの確率密度関数を$f(x)$、分布関数を$F_{X(x)}$とする.

このとき中央値と最頻値は以下のように定義される.

- 中央値 : $F_{X(m)} = 0.5$ となる m （または $P(X \leq m) = 0.5$）
- 最頻値: $f(x)$ が最大となる x

また、標準偏差と四分位範囲は以下のように定義される.

- 標準偏差: $\sigma = \sqrt{V[X]}$
- 四分位範囲: 第 3 四分位数と第 1 四分位数の差

  $P(X \leq Q_3) = 0.75, \quad P(X \leq Q_1) = 0.25$ となる $Q_3 - Q_1$

ここで、分布関数 $F_{X(x)}$ の逆関数 $F^{-1}(\alpha) = \inf\left\{x \mid F_{X(x)} \geq \alpha\right\}$ をXの**分位点関数**という. これを用いると、四分位範囲は $F^{-1}(0.75) - F^{-1}(0.25)$ と表せる.

$$\frac{\sqrt{V[X]}}{E[X]}$$

を**変動係数**と定義する. 分散は値の大きさに依存するため、平均値が大きく異なるデータ間では単純比較が難しい. 平均値で割ることでスケールの影響を除き、相対的な散らばりを比較することができる.

位置の指標や散らばりの指標以外の特性値として、分布の歪みの指標である**歪度**と、分布の裾の重さの指標である**尖度**がある.

$$\text{歪度} = \frac{E[(X-\mu)^3]}{V[X]^{\frac{3}{2}}}, \quad \text{尖度} = \frac{E[(X-\mu)^4]}{V[X]^2}$$

で定義される.

歪度は平均周りの 3 次モーメントを用いるため、分布の非対称性を表す. 分布の裾が右側に長く伸びているほど正の大きな値を取り、左側に伸びているほど負の値を取る.

尖度は 4 次モーメントを用いるため値は必ず正になり、平均から大きく離れた値の影響を強く受ける. 正規分布では尖度が 3 になるため、尖度から 3 を引いた値を**超過尖度**として定義し、正規分布との乖離を表すことがある.

同時分布の特性値

2 つの確率変数X, Yの相関を表す指標として**共分散**や**相関係数**がある. 共分散$\mathrm{Cov}(X, Y)$は

$$\mathrm{Cov}(X,Y) = E[(X-E[X])(Y-E[Y])] = E[XY] - E[X]E[Y]$$

で定義される. 正の相関があるときは共分散は正になり, 負の相関があるときは共分散は負になる. しかし, 共分散の大きさは元の確率変数の分散に依存しているため, これを基準化した

$$\rho[X,Y] = E\left[\left(\frac{X-E[X]}{\sqrt{V[X]}}\right)\left(\frac{Y-E[Y]}{\sqrt{V[Y]}}\right)\right] = \frac{\mathrm{Cov}(X,Y)}{\sqrt{V[X]V[Y]}}$$

を**相関係数**という.

相関係数は$[-1,1]$の範囲を取り, 相関係数の絶対値が 1 であればX,Yには一次式の関係$Y = aX+b$が成り立つ. 一方X,Yが独立であれば共分散, 相関係数は 0 になる.

---

2 つの確率変数X,Yに別の確率変数Zが影響を与えているときに, X,Yの相関は強くなりやすい. これを**疑似相関**という. このような場合, Zの影響を取り除いた相関を考えたい. ある変数の影響を除いた相関係数として**偏相関係数**がある. Zの影響を取り除いたX,Yの偏相関係数は

$$\rho[X,Y|Z] = \frac{\rho[X,Y]-\rho[X,Z]\rho[Y,Z]}{\sqrt{(1-\rho[X,Z]^2)(1-\rho[Y,Z]^2)}}$$

である.

---

2 つの確率変数X,Yについて一方の変数が与えられたもとでの期待値や分散をそれぞれ**条件付き期待値**, **条件付き分散**という. Xが与えられたもとでのYの条件付き期待値は

$$E[Y|X] = \int_{-\infty}^{\infty} y f_{Y|X}(y)dy$$

であり, Xが与えられたもとでのYの条件付き分散は

$$V[Y|X] = E[Y^2|X] - (E[Y|X])^2$$

である.

特性値の性質

期待値の性質

$$E[aX+bY+c] = aE[X]+bE[Y]+c$$

またX,Yが独立のとき

$$E[XY] = E[X]E[Y]$$

分散の性質

$$V[aX+b] = a^2V[X], V[X\pm Y] = V[X]+V[Y]\pm 2\,\mathrm{Cov}[X,Y]$$

条件付き期待値の性質

$$E[E[X|Y]] = E[X]$$

$$V[X] = E[V[X|Y]] + V[E[X|Y]]$$

データの特性値

これまで平均では$|(x) = \frac{1}{n}\sum_{i=1}^{n} x_i$という算術平均を用いてきたが平均にはこれ以外にもある.

加重平均は重み$w_1...w_n(w_i > 0, w_1 + ...w_n = 1)$に対する$x_1...x_n$の加重平均は$\sum_{i=1}^{n} w_i x_i$として定義される.これは観測値 x_i が割合 w_i で得られる場合の全平均を計算したものである.

幾何平均は$x_1...x_n(x_i > 0)$に対して$(x_1 \times ... \times x_n)^{\frac{1}{n}}$として定義される.

調和平均は$x_1...x_n(x_i > 0)$に対して$\frac{1}{x_1} + ... + \frac{1}{x_n}$の平均$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{x_i}\right)$の逆数として定義される

平均ベクトルと分散共分散行列

$\boldsymbol{X} = (X_1...X_k)^\top$を k 次元確率ベクトルとする.$\mu_i = E[X_i]$を要素とする k 次元ベクトル$\boldsymbol{\mu} = (\mu_1...\mu_n)^\top$を**期待値ベクトル**あるいは**平均ベクトル**と呼ぶ.またX_iとX_jの共分散$\sigma_{ij} = E\big[(X_i - E[X_i])\big(X_j - E\big[X_j\big]\big)\big]$を$(i,j)$要素とする行列

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1k} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{k1} & \sigma_{k2} & \cdots & \sigma_{kk} \end{pmatrix}$$

を**分散共分散行列**とよぶ.分散共分散行列の対角成分はX_iの共分散である.同様に対角要素を 1 としてX_i, X_jの相関係数ρ_{ij}を(i,j)要素とする行列を**相関係数行列**あるいは**相関行列**と呼ぶ.