確率統計学

多次元の確率分布

会津大学 コンピュータ理工学研究科 コンピュータ情報システム学専攻 髙橋輝

同時確率分布と周辺確率分布

同時確率分布

n 個の確率変数 $X_1, X_2, \cdots, Xn$ の**同時確率分布**とは、確率変数の組 $(X_1, X_2, \cdots, X_n) \in R_n$に確率を対応させる関数のこと.

2次元の場合,

$$P(X = x, Y = y) = f(x, y)$$

ただし,

$$f(x, y) \geq 0 \text{かつ} \sum_{x \in X} \sum_{y \in Y} f(x, y) = 1$$

事象

一般に, 確率測度に対して可測である集合Aを**事象**と呼ぶ.

……なんて？？？

(僕らが考える上では)そんな難しいことわかんないので、
(測度論の話は)省略

$$\Pi_{i=1}^{n} X_i \quad \text{の部分集合}$$

くらいのイメージでいいはず。

ちゃんと説明してほしい人がいたら、僕ができる範囲でする。

余談

ここら辺の確率変数の定義とかは、この本では(多分)意図的にぼかされて書いてある。
本来は、確率空間から可測空間への写像で定義されている。
前の章で扱った分布や、以下で扱う積分等も、本来は(確率空間と)可測写像の逆像によって定義されている。

事象の確率

(二次元の)事象Aの確率は、同時確率分布fを用いて、以下のように定義される。

$$P((x,y)=A)=\sum_{x\in A} f(x,y)$$

同時確率密度関数

X_iが連続型の確率変数の時には, $f(x_i, \cdots, x_n)$を**同時確率密度関数**と呼ぶ.
$f(x_i, \cdots, x_n)$は,

$$f(x_i, \cdots, x_n) \geq 0 \text{かつ} \int \cdots \int_{\mathbb{R}^n} f(x_1, \cdots, x_n) dx_1 \cdots dx_n = 1$$

を満たす。
ここで, $\mathbb{R}^n$はn次元のユークリッド空間全体を指し、標本空間Ωとも言われる。

この$f(x_i, \cdots, x_n)$によって, 事象Aの確率は以下のように定義される。

$$P((x_1, \cdots, x_n) = A) = \int \cdots \int_A f(x_1, \cdots, x_n) dx_1 \cdots dx_n$$

一般に、Aは領域(Region)となる。

周辺確率分布

同時確率分布から求められる, X, Y 単独の確率分布のこと。

$$g(x) = \sum_{y \in Y} f(x, y) \qquad h(y) = \sum_{x \in X} f(x, y)$$

それぞれ、X, Y の周辺確率分布と呼ぶ。
連続型の場合、周辺確率密度関数として、同時確率密度関数から以下のように定義される。

$$g(x) = \int_{-\infty}^{\infty} f(x, y) dy \qquad h(y) = \int_{-\infty}^{\infty} f(x, y) dx$$

周辺確率分布(密度関数)から同時確率分布(密度関数)を求めることはできない。
(Exception. 変数が独立の場合)

共分散と相関係数

2変数X, Yの間に関連があれば、一方の変化は他方に及ぶと考えられるから、一般に分散の単純な加法は成り立たない。

$$V(X+Y) \neq V(X) + V(Y)$$

定義に基づいて計算すると,

$$V(X+Y) = V(X) + V(Y) + 2Cov(X, Y)$$

となる。

実際の定義に基づいた共分散の計算

$$
\begin{aligned}
V(X+Y) = \mathbb{E}[(X+Y-\mathbb{E}[X+Y])^2] \\
= \mathbb{E}[(X^2+Y^2+2XY) - 2(X\mathbb{E}[X]+X\mathbb{E}[Y]+Y\mathbb{E}[X]+Y\mathbb{E}[Y]) \\
+(\mathbb{E}[Y]^2+\mathbb{E}[X]^2+2\mathbb{E}[X]\mathbb{E}[Y])] \\
= \mathbb{E}[(X^2+Y^2-2X\mathbb{E}[X]-2Y\mathbb{E}[Y]+\mathbb{E}[Y]^2+\mathbb{E}[X]^2) \\
-2(XY-X\mathbb{E}[Y]-Y\mathbb{E}[X]+\mathbb{E}[X]\mathbb{E}[Y]) \\
= \mathbb{E}[X^2]-\mathbb{E}[X]^2+\mathbb{E}[Y^2]-\mathbb{E}[Y]^2-2[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] \\
= V(X)+V(Y)-2\mathrm{Cov}(X,Y)
\end{aligned}
$$

……つかれた。

共分散の意味

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

式を見ればわかると思うが、共分散はXとYの平均値からのばらつきが同じ方向にあるか、逆方向にあるかを示す指標として考えることができる。

相関係数

しかし、このままでは、異なる単位の変数の共分散を比較することができない。そこで、共分散を標準化したものが相関係数である。
確率変数X, Yの相関係数は以下のように定義される。

$$\rho(X,Y) = \rho_{XY} = \frac{Cov(X,Y)}{\sqrt{V(X)V(Y)}}$$

$\rho(X,Y)$は必ず$-1 \leq \rho \leq 1$の範囲に収まる。(証明はP.138中段)

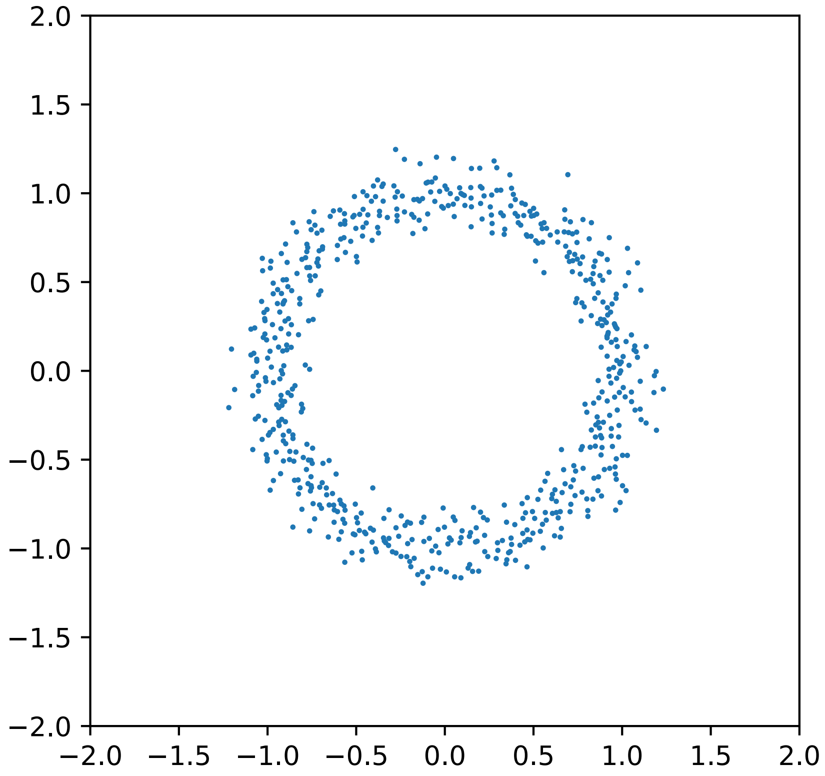
相関係数の意味

導出から明らかなように
$\rho > 0$ならば、Xが大きくなるとYも大きくなり、
$\rho < 0$ならば、Xが大きくなるとYは小さくなる傾向がある。

$\rho = \pm 1$ならば、XとYは完全に線形に関連している。 (i.e. $X = aY + b$が成り立つ)

逆に、$\rho = 0$ならば、XとYはどちらの関係を持つとも言えない。
（ただし、XとYが独立であるとは限らない。）

無相関でも独立ではない例



共分散の簡単な計算方法

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

導出は、

$$\begin{aligned}\mathrm{Cov}(X,Y) &= \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] \\ &= \mathbb{E}[XY - X\mathbb{E}[Y] - Y\mathbb{E}[X] + \mathbb{E}[X]\mathbb{E}[Y]] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[Y] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]\end{aligned}$$

離散型確率変数の共分散

$$\mathrm{Cov}(X, Y) = \sum_{x \in X} \sum_{y \in Y} (x - \mu_X)(y - \mu_Y) f(x, y)$$

で求められるが、

$$\mathbb{E}[XY] = \sum_{x \in X} \sum_{y \in Y} xy \cdot f(x, y)$$

を用いて求めることが多い。

連続型確率変数の共分散

$$\mathrm{Cov}(X, Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x - \mu_X)(y - \mu_Y)f(x, y)\mathrm{d}x\mathrm{d}y$$

$$\mathbb{E}[XY] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy \cdot f(x, y)\mathrm{d}x\mathrm{d}y$$

例はぶっ飛ばします！

よめばわかる。

条件付確率と独立な確率変数

条件付確率の定義