確率密度と確率の関係について

確率変数 X, Y があるとする。X の実現値が a より大きく $a+\Delta a$ 以下になる確率は

$$\begin{aligned} P(a < X \leq a+\Delta a) &= F_X(a+\Delta a) - F_X(a) \\ &= \int_a^{a+\Delta a} f_X(x)dx \end{aligned}$$

である。もし f_X が a の前後でおおむね一定であるならば

$$P(a < X \leq a+\Delta a) \simeq f_X(a)\Delta a$$

が成立する。
一方、X の実現値が a より大きく $a+\Delta a$ 以下になりかつ Y の実現値が b より大きく $b+\Delta b$ 以下になる確率は

$$\begin{aligned} & P(a < X \leq a+\Delta a, b < Y \leq b+\Delta b) \\ = \ & F_{XY}(a+\Delta a, b+\Delta b) - F_{XY}(a,b) \\ = \ & \int_b^{b+\Delta b}\int_a^{a+\Delta a} f_{XY}(x,y)dxdy \end{aligned}$$

である。もし f_{XY} が点 (a,b) 周囲でおおむね一定であるならば、

$$P(a < X \leq a+\Delta a, b < Y \leq b+\Delta b) \simeq f_{XY}(a,b)\Delta a\Delta b$$

が成立する。

2009-4-1

確率変数の独立性のより一般的な定義

確率変数 X と Y の独立性を、連続型であるか離散型であるかに関わらず、以下のように定義できる。

すべての x, y について

$$F_{XY}(x, y) = F_X(x) \cdot F_Y(y)$$

が成立するときに、X と Y が統計的に独立であると言う。確率変数が両方とも連続型であるときまたは両方とも離散型であるとき、上記の定義は既に紹介した定義と同じになる。

2009-4-2

今日の内容：多次元の確率分布パート２
(教科書の７章)
３変数以上の統計的独立性

n 個の確率変数について同時累積分布関数、同時確率分布、同時確率密度関数は２個の確率変数の場合と同じように定義される。
n 個の確率変数 $X_1,\ X_2,\ \ldots,\ X_n$ があるとする。それらがすべて離散型のときすべての $x_1,\ \ldots,\ x_n$ について

$$P_{X_1X_2\ldots X_n}(x_1,x_2,\ldots,x_n)=P_{X_1}(x_1)P_{X_2}(x_2)\cdots P_{X_n}(x_n)$$

を満たせば、$X_1,\ \ldots,\ X_n$ が統計的に独立であると言う。
$X_1,\ X_2,\ \ldots,\ X_n$ がすべて連続型であるときは、すべての $x_1,\ \ldots,\ x_n$ について

$$f_{X_1X_2\ldots X_n}(x_1,x_2,\ldots,x_n)=f_{X_1}(x_1)f_{X_2}(x_2)\cdots f_{X_n}(x_n)$$

を満たせば、$X_1,\ \ldots,\ X_n$ が統計的に独立であると言う。

2009-4-3

$X_1, \ldots, X_n$ が統計的に独立ならば、すべての $i \neq j$ について X_i と X_j は統計的に独立になるが、逆は成立しない。
例えば X_1, X_2 を 0 または 1 を等確率で取る離散型確率変数で X_1 と X_2 は統計的に独立であるとする。X_3 を $X_1 + X_2$ を 2 で割った余りの確率変数とする。このとき X_1 と X_3 および X_2 と X_3 は統計的に独立になるが、X_1, X_2, X_3 は統計的に独立ではない。

共分散

確率変数 X, Y があるときにその和 $X+Y$ の分散を考えたい。定義に従い

$$
\begin{aligned}
& V(X+Y) \\
= \; & E[(X+Y)^2] - E[X+Y]^2 \\
= \; & E[X^2] - E[X]^2 + E[Y^2] - E[Y]^2 \\
& +2E[XY] - 2E[X]E[Y] \\
= \; & V[X] + V[Y] + 2(E[XY] - E[X]E[Y])
\end{aligned}
$$

となる。第３項の $E[XY] - E[X]E[Y]$ を共分散と呼び $\mathrm{Cov}[X, Y]$ と表記する。
$m_X = E[X]$, $m_Y = E[Y]$ とする。

$$
\begin{aligned}
& E[XY] - m_X m_Y \\
= \; & E[XY] - 2m_X m_Y + m_X m_Y \\
= \; & E[XY] - m_X E[Y] - m_Y E[X] + m_X m_Y \\
= \; & E[XY - m_X Y - m_Y X + m_X m_Y] \\
= \; & E[(X - m_X)(Y - m_Y)]
\end{aligned}
$$

直近の式 $E[(X-m_X)(Y-m_Y)]$ が、共分散の意味を明らかにしている。$(X-m_X)(Y-m_Y)$ は X と Y が期待値から同じ方向にずれるときに正になり、反対方向にずれるときに負になる。従って $X+Y$ の分散は、共分散が大きいほど大きくなるが、共分散は X と Y が期待値から同じ方向にずれる確率が高いときに大きくなり、素朴な印象と一致している。
逆に X と Y が期待値から反対方向にずれる確率が大きければ共分散は負の値になり、X と Y の影響が打ち消し合うから $X+Y$ の分散は X の分散と Y の分散の和よりも小さくなる。
また X と Y が統計的に独立ならば $E[XY]=E[X]E[Y]$ なので共分散は 0 になる。

相関係数

複数の確率変数が統計的にどの程度依存しているか定量的に捉える概念を紹介する。共分散がそのために使えそうに見えるが X または Y の分散が大きくなると共分散も大きくなるため、例えば X と Y の相互依存関係の強さと X と Z の相互依存関係の強さを比べるために共分散はそのままでは使えない。

X と Y の相関係数を

$$\rho_{XY} = \frac{\mathrm{Cov}[X,Y]}{\sqrt{V[X]V[Y]}}$$

で定義する。相関係数はどのような確率変数についても -1 以上 1 以下の値を取る。理由は以下の通り：

$$V(tX+Y) = t^2V[X] + 2t\mathrm{Cov}[X,Y] + V[Y]$$

が成立するが、分散は非負なので t の値によらず右辺は非負になる必要がある。従って右辺を t の二次式と見るとその判別式 $\mathrm{Cov}[X,Y]^2 - V[X]V[Y]$ は 0 以下になる必要がある。従って $-1 \leq \rho_{XY} \leq 1$ である。また X と Y が統計的に独立ならば共分散が 0 になるので相関係数も 0 である。$\rho_{XY} = 0$ となるとき X と Y は無相関であると言う。

2009-4-7

相関係数の例

X, Y が 1 または -1 の値を取る確率変数で同時確率分布が $P_{XY}(1,-1) = P_{XY}(-1,1) = 0.5$ となっているとする。このとき $E[X] = E[Y] = 0$, $V[X] = E[X^2] - E[X]^2 = 1$ 同様に $V[Y] = 1$ である。分散が 1 なので

$$\rho_{XY} = \mathrm{Cov}[XY] = E[XY] - E[X]E[Y] = -1$$

である。

一方確率変数 Z を $Z = 2Y + 2$ で定義すると X と Z の同時確率分布は $P_{XZ}(1,0) = P_{XZ}(-1,4) = 0.5$ になる。このとき X と Z の関係は X と Y に本質的に同じであること注意して欲しい。$V[Z] = 4$, $E[Z] = 2$ であり

$$\mathrm{Cov}(X,Z) = E[XZ] - E[X]E[Z] = -2$$

である。従って $\rho_{XZ} = \mathrm{Cov}[X,Z]/\sqrt{V[X]V[Z]} = -1$ である。共分散の値は異なるが相関係数は同じになっていることに注目して欲しい。

無相関と統計的独立性の関係

X と Y が統計的に独立ならば無相関である。しかし X と Y が無相関なのに X と Y が統計的に独立ではないことがある。
例えば $P_{XY}(1,1) = P_{XY}(0,-2) = P_{XY}(-1,1) = 1/3$ という同時確率分布を X と Y が持つとする。$E[X] = E[Y] = 0$ である。$\mathrm{Cov}[X,Y] = E[XY] - E[X]E[Y] = 0$ であるから共分散と相関係数は両方とも 0 である。しかし X と Y は統計的に独立ではない。

2009-4-9

演習問題

原則として授業終了までに答案を提出して下さい。終らない場合は授業が行われた日の夜までに南３号館３１１号室前の箱に演習答案を提出して下さい。

31. 2009-4-4 ページの X_1, X_2, X_3 について以下のことを確認せよ。

- すべての x_1, x_3 について $P_{X_1X_3}(x_1, x_3) = P_{X_1}(x_1) P_{X_3}(x_3)$
- すべての x_2, x_3 について $P_{X_2X_3}(x_2, x_3) = P_{X_2}(x_2) P_{X_3}(x_3)$

さらに $P_{X_1X_2X_3}(x_1, x_2, x_3) \neq P_{X_1}(x_1)P_{X_2}(x_2)P_{X_3}(x_3)$ となる (x_1, x_2, x_3) の組を１つ明示せよ。

32. 正四面体に 0,1,2,3 の数字が書いてある 4 面サイコロを考える。このサイコロを振ると各数字が等しい確率で出るとする (床に面している数字が「出た」と見なす)。このサイコロの出た目を確率変数 X で表わし、X を 3 で割った余りを Y $(= 0,1,2)$ とし、X を 2 で割った余りを Z $(= 0,1)$ とする。Y と Z の共分散と相関係数を求めよ。

33. 2009-4-9 ページの X と Y が統計的に独立ではない理由を説明せよ。

34. U を 0 以上 2 以下のすべての値に一様に実現値を取る一様連続確率変数とする。累積分布関数 F_U と確率密度関数 f_U を図示せよ (式が無くてもよい)。
35. U を前問と同じ確率変数とし、V を U と同じ分布を持ち U と統計的に独立な確率変数とする。同時累積分布関数 F_{UV} を求め、同時確率密度関数 f_{UV} が存在するか否か答えよ。(デルタ関数を知っている学生へ：この問では密度関数にデルタ関数を用いることはできないとする)
36. U を前問と同じ確率変数とし、W を確率変数 U の実現値が u であるときに W の実現値も必ず u になる確率変数とする。同時累積分布関数 F_{UW} を求め、同時確率密度関数 f_{UW} が存在するか否か答えよ。(デルタ関数を知っている学生へ：この問では密度関数にデルタ関数を用いることはできないとする)
37. 授業でわかりにくかった所や要望を書いて下さい。また前回の答案で×になったところで、理由がわからないところは申し出て下さい。