

# ゲノムリテラシー講座 講義資料 1

加藤 毅

<http://www.net-machine.net/~kato/>

平成 21 年 7 月 25 日

## 確率論

確率分布について簡単に紹介する．

### 確率関数

整数のみの値をとりうる離散確率変数  $x$  に対して，確率分布  $P(x)$  は次の性質を満たさなくてはならない．

$$\forall x \in \mathbb{Z} : \quad P(x) \geq 0, \\ \sum_{x \in \mathbb{Z}} P(x) = 1.$$

ただし， $\mathbb{Z}$  は整数の集合を表す． $P(x)$  はしばしば，確率関数と呼ばれる．

例 1 (サイコロ) 典型的なサイコロの目  $x$  は， $1, \dots, 6$  の値を等しい確率でとる．よって，その確率関数は

$$P(x) = \begin{cases} \frac{1}{6} & (x = 1, \dots, 6) \\ 0 & (\text{それ以外}) \end{cases}$$

とあらわされる．

例 2 (コイン) コインをなげる試行を考え，表が出る事象に  $x = 1$  を割り当て，裏が出る事象に  $x = 0$  を割り当てるとする．表と裏が出る確率は等しいとすると，確率変数  $x$  の確率関数は

$$P(x) = \begin{cases} \frac{1}{2} & (x = 0, 1) \\ 0 & (\text{それ以外}) \end{cases}$$

とあらわされる．

複数の離散確率変数に対しても確率分布を定義できる．2 つの離散確率変数  $x$  および  $y$  の確率関数は

$$\forall x, \forall y \in \mathbb{Z} : \quad P(x, y) \geq 0 \\ \sum_{x \in \mathbb{Z}} \sum_{y \in \mathbb{Z}} P(x, y) = 1$$

を満たさなくてはならない．

$P(x, y)$  はこの 2 つの整数の組  $(x, y)$  が同時に生起する確率を表しているので，同時確率と呼ばれる．

例 3 天候によって A 君が遅刻する確率が変化するとし，離散確率変数  $x$  と  $y$  に，事象を

$$\begin{array}{ll} x = 1 & \text{A 君が遅刻する,} \\ y = 1 & \text{雨が降る,} \end{array} \quad \begin{array}{ll} x = 0 & \text{遅刻しない} \\ y = 0 & \text{降らない} \end{array}$$

のように割り当てるとする．

同時確率  $P_{x,y}(x, y)$  は次の表に従うとする：

|          | $y = 0$               | $y = 1$               | $P_x(x)$       |
|----------|-----------------------|-----------------------|----------------|
| $x = 0$  | $P_{x,y}(0, 0) = 0.4$ | $P_{x,y}(0, 1) = 0.2$ | $P_x(0) = 0.6$ |
| $x = 1$  | $P_{x,y}(1, 0) = 0.1$ | $P_{x,y}(1, 1) = 0.3$ | $P_x(1) = 0.4$ |
| $P_y(y)$ | $P_y(0) = 0.5$        | $P_y(1) = 0.5$        |                |

A 君が遅刻する確率  $P_x(x)$  は

$$P_x(x = 1) = P_{x,y}(x = 1, y = 0) + P_{x,y}(x = 1, y = 1) = 0.4$$

と与えられ，雨が降る確率は

$$P_y(y = 1) = P_{x,y}(x = 0, y = 1) + P_{x,y}(x = 1, y = 1) = 0.5$$

で与えられる．

$P_x(x)$ ,  $P_y(y)$  は，それぞれ， $x$  の周辺分布， $y$  の周辺分布 と呼ばれる．

このとき，雨が降る条件の下で A 君が遅刻する条件付き確率は

$$\begin{aligned} P_{x|y}(x = 1|y = 1) &= \frac{P_{x,y}(x = 1, y = 1)}{P_{x,y}(x = 0, y = 1) + P_{x,y}(x = 1, y = 1)} \\ &= \frac{P_{x,y}(x = 1, y = 1)}{P_y(y = 1)} = \frac{0.3}{0.5} = 0.6 \end{aligned}$$

で与えられる．

一般に，2つの確率変数  $x$  および  $y$  に対して，

$$P_{x|y}(x|y) = \frac{P_{x,y}(x, y)}{P_y(y)}$$

が成り立ち，ベイズの定理と呼ばれる．

## 確率密度関数

実数値をとる連続確率変数  $x$  に対して，確率分布は確率密度関数  $p(x)$  で決定される：

$$\begin{aligned} \forall x \in \mathbb{R} : \quad & p(x) \geq 0 \\ & \int_{-\infty}^{+\infty} dx p(x) = 1. \end{aligned}$$

ただし， $\mathbb{R}$  は実数の集合を表す．

例 4 (一様分布) 連続確率変数  $x$  が 0 以上 1 以下では等しい確率で生起し，それ以外の値の確率が 0 であるとき，その確率密度関数は

$$p(x) = \begin{cases} 1 & (0 \leq x \leq 1) \\ 0 & (\text{それ以外}) \end{cases}$$

とあらわされる．

複数の連続確率変数に対しても確率分布を定義できる．2つの連続確率変数  $x$  および  $y$  の確率密度関数は

$$\begin{aligned} \forall x, \forall y \in \mathbb{R} : \quad & p(x, y) \geq 0 \\ & \int_{-\infty}^{+\infty} dx \int_{-\infty}^{+\infty} dy p(x, y) = 1 \end{aligned}$$

を満たさなくてはならない．

## 期待値と分散

離散確率変数  $x$  に対して, ある確率分布  $P(x)$  の下での期待値は

$$E(x) = \sum_{x \in \mathbb{Z}} P(x)x$$

で定義される. 期待値のことを平均ともいう.

より一般には, ある関数  $f(x)$  に対して期待値を定義でき,

$$E(f(x)) = \sum_{x \in \mathbb{Z}} P(x)f(x)$$

とあらわされる.  $E(x)$  は  $f(x) = x$  とおいた場合に相当する.

連続確率変数  $x$  に対しては, 関数  $f(x)$  の期待値は

$$E(f(x)) = \int_{-\infty}^{+\infty} dx p(x)f(x)$$

とあらわされる.

関数  $f(x)$  の分散は,

$$\text{var}(f(x)) = E((f(x))^2) - (E(f(x)))^2$$

で定義される. 標準偏差は

$$\sigma(f(x)) = \sqrt{\text{var}(f(x))}$$

で定義される.

例 5 (サイコロ) サイコロの目  $x$  の平均は,

$$\begin{aligned} E(x) &= \sum_{x \in \mathbb{Z}} P(x)x \\ &= \frac{1}{6} (1 + 2 + 3 + 4 + 5 + 6) \\ &= \frac{7}{2}, \end{aligned}$$

となる.

$x^2$  の平均は,

$$\begin{aligned} E(x^2) &= \sum_{x \in \mathbb{Z}} P(x)x^2 \\ &= \frac{1}{6} (1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2) = \frac{91}{6} \end{aligned}$$

となる.

これより, 分散と標準偏差は

$$\begin{aligned} \text{var}(x) &= E(x^2) - (E(x))^2 \\ &= \frac{91}{6} - \left(\frac{7}{2}\right)^2 \\ &= \frac{35}{12}, \\ \sigma(x) &= \sqrt{\text{var}(x)} \\ &= \sqrt{\frac{35}{12}} \end{aligned}$$

となる.

例 6 (一様分布変数) 一様分布に従う確率変数  $x$  の平均は ,

$$\begin{aligned} E(x) &= \int dx p(x) x \\ &= \int_0^1 dx x \\ &= \frac{1}{2}, \end{aligned}$$

となる .  
 $x^2$  の平均は ,

$$\begin{aligned} E(x^2) &= \int_0^1 dx x^2 \\ &= \frac{1}{3} \end{aligned}$$

となる .  
これより , 分散と標準偏差は

$$\begin{aligned} \text{var}(x) &= E(x^2) - (E(x))^2 \\ &= \frac{1}{3} - \left(\frac{1}{2}\right)^2 \\ &= \frac{1}{12}, \\ \sigma(x) &= \sqrt{\text{var}(x)} \\ &= \frac{1}{\sqrt{12}} \end{aligned}$$

となる .

確率密度関数  $p(x, y)$  にしたがう 2 つの連続確率変数  $x$  および  $y$  に対して , 2 変数関数  $f(x, y)$  の期待値は

$$E(f(x, y)) = \int_x \int_y dx dy p(x, y) f(x, y)$$

と定義される .

この定義を使って , 2 つの確率変数  $x$  および  $y$  に対して , それらの共分散は

$$\text{cov}(x, y) = E(xy) - E(x)E(y)$$

とあらわされる .  
相関係数は

$$\rho(x, y) = \frac{\text{cov}(x, y)}{\sqrt{\text{var}(x)\text{var}(y)}}$$

と定義される .

## \*平成 19 年度 問 21

以下に示した 6 個の数値のデータが与えられたとする．このデータにおける 3 個の基本統計量（平均値，中央値，最頻値）の値として正しい組み合わせを選択肢の中から一つ選べ．

ここで中央値はメジアン，最頻値はモードとも呼ばれる．

データ：2, 3, 8, 2, 6, 9

|   | 平均値 | 中央値 | 最頻値 |
|---|-----|-----|-----|
| 1 | 5   | 4.5 | 2   |
| 2 | 5   | 5   | 5   |
| 3 | 4.5 | 4.5 | 2   |
| 4 | 4.5 | 2   | 5   |

## 解説

平均値，中央値，最頻値は，確率変数に対する統計量であるが，ここでは，標本集合に対する標本平均値，標本中央値，標本最頻値をさしている．

サイズ  $n$  の標本集合  $\mathcal{X} = (x_1, \dots, x_n)$  が所与のとき，その標本平均値は

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

で定義される．

標本中央値は，標本を

$$x'_1 \leq \dots \leq x'_n$$

のように並べ替えたものを使って，

$$\begin{cases} x'_{\frac{n+1}{2}} & (n \text{ が奇数のとき}) \\ \frac{1}{2} \left( x'_{\frac{n}{2}} + x'_{\frac{n+1}{2}} \right) & (n \text{ が偶数のとき}) \end{cases}$$

で定義される．

標本最頻値は，標本集合の中で最も頻度の多かった値で定義される．

では，標本集合  $\mathcal{X} = \{2, 3, 8, 2, 6, 9\}$  に対しては，標本平均値は

$$\frac{1}{6}(2 + 3 + 8 + 2 + 6 + 9) = 5$$

となる．

この標本集合を昇順に整列すると 2, 2, 3, 6, 8, 9 となるので，標本中央値は

$$\frac{1}{2}(3 + 6) = 4.5$$

となる．

値 2 の頻度が 2 で最も多いので，標本最頻値は

$$2$$

となる．

よって，選択肢 1 が正しい．

## \*平成 19 年度 問 22

10,000 人が受験した試験があるとする．その採点結果を集計したところ，得点分布が平均 60 点，標準偏差 10 点の正規分布にしたがっていることが分かった．ここで，受験生の A さんが 80 点を獲得したとき，A さんの成績に関する記述としてもっとも適切な説明を選択肢の中から一つ選べ．

ここで必要であれば，下記の標準正規分布表を利用してよい．なお，標準正規分布表では， $z$  で示された位置よりも上側の確率を示している．

| $z$ | 0.0     | 0.2     | 0.4     | 0.6     | 0.8     |
|-----|---------|---------|---------|---------|---------|
| 0.0 | 0.50000 | 0.42074 | 0.34458 | 0.24196 | 0.21186 |
| 1.0 | 0.15866 | 0.11507 | 0.08076 | 0.05480 | 0.03593 |
| 2.0 | 0.02275 | 0.01390 | 0.00820 | 0.00466 | 0.00256 |
| 3.0 | 0.00135 | 0.00069 | 0.00034 | 0.00016 | 0.00007 |

- 1 A さんの成績は，ほぼ平均であり，5,000 位前後であろう．
- 2 A さんの成績は，平均より高く，およそ 1,600 位前後であろう．
- 3 A さんの成績は，平均より高く，およそ 230 位前後であろう．
- 4 A さんの成績は，平均より高く，およそ 15 位前後であろう．

## 解説

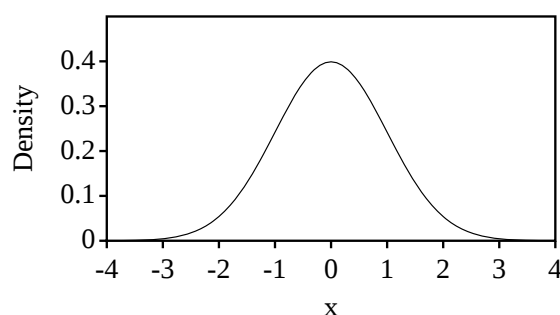
正規分布は確率分布の一つで，その確率密度関数は，2 つのパラメータ  $\mu, \sigma^2$  を用いて

$$\mathcal{N}(x | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$

で表される．

この 2 つのパラメータ  $\mu, \sigma^2$  は特に平均，分散と呼ばれている．

平均  $\mu$ ，分散  $\sigma^2$  を持つ正規分布は  $\mathcal{N}(\mu, \sigma^2)$  と表される．



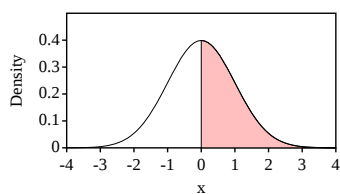
$\mu = 0, \sigma^2 = 1$  とおいた正規分布  $\mathcal{N}(0, 1)$  は標準正規分布と呼ばれる．

標準正規分布表は  $\mathcal{N}(0, 1)$  において， $z$  で示された値より大きな数値が生起される確率

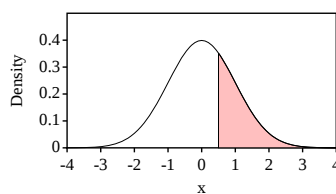
$$P(Z \geq z) = \int_z^{\infty} dt \mathcal{N}(t | 0, 1)$$

を表される．

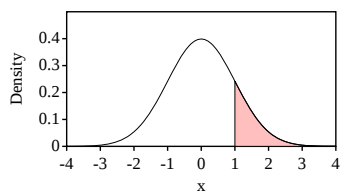
$z$  の値を 0.0, 0.5, 1.0, 1.5, 2.0, 2.5 の場合の確率を以下に図示する：



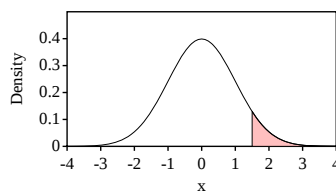
$$z = 0.0, P(Z \geq z) = 0.500$$



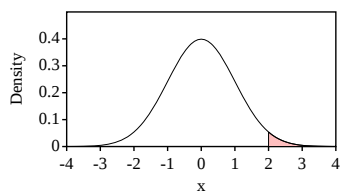
$$z = 0.5, P(Z \geq z) = 0.309$$



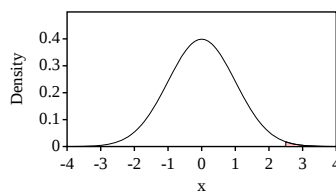
$$z = 1.0, P(Z \geq z) = 0.159$$



$$z = 1.5, P(Z \geq z) = 0.067$$



$$z = 2.0, P(Z \geq z) = 0.023$$



$$z = 2.5, P(Z \geq z) = 0.006$$

$\mathcal{N}(\mu, \sigma^2)$  に従う確率変数  $x$  に対して,

$$z = \frac{x - \mu}{\sigma}$$

という変換 ( $z$  変換と呼ばれる) を行くと,  $z$  は  $\mathcal{N}(0, 1)$  に従う.

設問の仮定では, 得点分布が  $\mathcal{N}(60, 10^2)$  に従うとなっている.

A さんは  $x = 80$  点獲得したので, この  $z$  値は

$$z = \frac{80 - 60}{10} = 2$$

で与えられるため,

$$P(X \geq 80) = P(Z \geq 2) = 0.02275$$

となることが分かる.

よって, A さんは 10,000 人中だいたい 227.5 位になるはずである.

したがって, 選択肢 3 が正しい.

## \*平成 19 年度 問 25

2つの独立な変数  $X$  と  $Y$  があり、それぞれ平均 0、分散 1 の標準正規分布に従うことがわかっているとする。次に示したこの 2 つの確率変数についての説明文の中で、間違っているものはどれか。一つ選べ。

- 1  $X$  と  $Y$  の積  $XY$  の平均値は 0 である。
- 2  $X + Y$  は、正規分布に従う。
- 3  $X + Y$  の分散は、1 である。
- 4  $X - Y$  の分散は、2 である。

## 解説

確率密度関数  $p(X, Y)$  をもつ 2 つの確率変数  $X, Y$  に対して、2 変数関数  $f(X, Y)$  の期待値は

$$E(f(X, Y)) = \int_X \int_Y dX dY p(X, Y) f(X, Y)$$

と定義されている。

この期待値を使って以下の統計量が定義されている：

$f(X)$  の分散

$$V(f(X), g(X)) = E(f(X)^2) - (E(f(X)))^2$$

$f(X)$  と  $g(X)$  の共分散

$$\text{cov}(f(X), g(X)) = E(f(X)g(X)) - E(f(X))E(g(X)),$$

$f(X)$  と  $g(X)$  の相関係数

$$\rho(f(X), g(X)) = \frac{\text{cov}(f(X), g(X))}{\sqrt{V(f(X))V(g(X))}}$$

正規分布は確率分布の一つで、その確率密度関数は、2 つのパラメータ  $\mu, \sigma^2$  を用いて

$$p(X) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(X-\mu)^2}{2\sigma^2}\right)$$

で表される。

この 2 つのパラメータ  $\mu, \sigma^2$  は特に平均、分散と呼ばれている。

平均  $\mu$ 、分散  $\sigma^2$  を持つ正規分布は  $\mathcal{N}(\mu, \sigma^2)$  と表される。

$\mu = 0, \sigma^2 = 1$  とおいた正規分布  $\mathcal{N}(0, 1)$  は標準正規分布と呼ばれる。

$\mathcal{N}(\mu_U, \sigma_U^2)$  に従う確率変数  $U$  と  $\mathcal{N}(\mu_V, \sigma_V^2)$  に従う確率変数  $V$  を考え、この 2 つの確率変数は互いに独立であるとする。

任意のスカラー  $a, b$  に対して、線形結合  $aU + bV$  は正規分布  $\mathcal{N}(a\mu_U + b\mu_V, a^2\sigma_U^2 + b^2\sigma_V^2)$  に従うことが知られている。

選択肢 1 2 つの確率変数  $X$  と  $Y$  に対して、

$$p(X, Y) = p(X)p(Y)$$

が成り立つとき、 $X$  と  $Y$  は統計的独立である、という。

これより、 $XY$  の期待値は、

$$\begin{aligned} E(XY) &= \int_X \int_Y dX dY p(X, Y) XY \\ &= \int_X \int_Y dX dY p(X)p(Y) XY \\ &= \left( \int_X dX p(X) X \right) \left( \int_Y dY p(Y) Y \right) \\ &= E(X)E(Y) = 0 \cdot 0 = 0 \end{aligned}$$

を得る。

選択肢 2  $X + Y$  は  $\mathcal{N}(0, 2)$  に従う . よって , この選択肢は正しい .

選択肢 3  $X + Y$  は  $\mathcal{N}(0, 2)$  に従うので , 分散は 2 である . よって , この選択肢には間違いがある .

選択肢 4  $X - Y$  は  $\mathcal{N}(0, 2)$  に従うので , 分散は 2 である . よって , この選択肢は正しい .

## \*平成 19 年度 問 23

1 つのサンプルから 2 つの測定値 ( 例 , ある人の身長と体重 , ある人の国語の成績と算数の成績 など ) が得られる場合 , この 2 つの測定値に相関関係があるかどうかを判定する方法に相関係数  $r$  がある . 1 つのサンプルが ,  $x$  と  $y$  の 2 つの測定値をもつような集団に対して測定を行い , 測定結果を散布図を用いて表現した結果を以下に示した . このとき , 相関係数  $r$  の値は , どのような範囲の値を持つと考えられるか . 適当なものを選択肢の中から一つ選べ .



- 1  $1 < r$
- 2  $0 < r \leq 1$
- 3  $r = 0$
- 4  $-1 \leq r < 0$

## 解説

2 変量の標本集合

$$\mathcal{X} = \{(x_1, y_1), \dots, (x_\ell, y_\ell)\}$$

から , 標本相関係数は

$$\rho = \frac{\sum_{i=1}^{\ell} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{\ell} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{\ell} (y_i - \bar{y})^2}}$$

で表される .

ただし ,  $\bar{x}$  および  $\bar{y}$  は , 標本平均

$$\bar{x} = \frac{1}{\ell} \sum_{i=1}^{\ell} x_i, \quad \bar{y} = \frac{1}{\ell} \sum_{i=1}^{\ell} y_i$$

を表す .

この定義より標本相関係数は

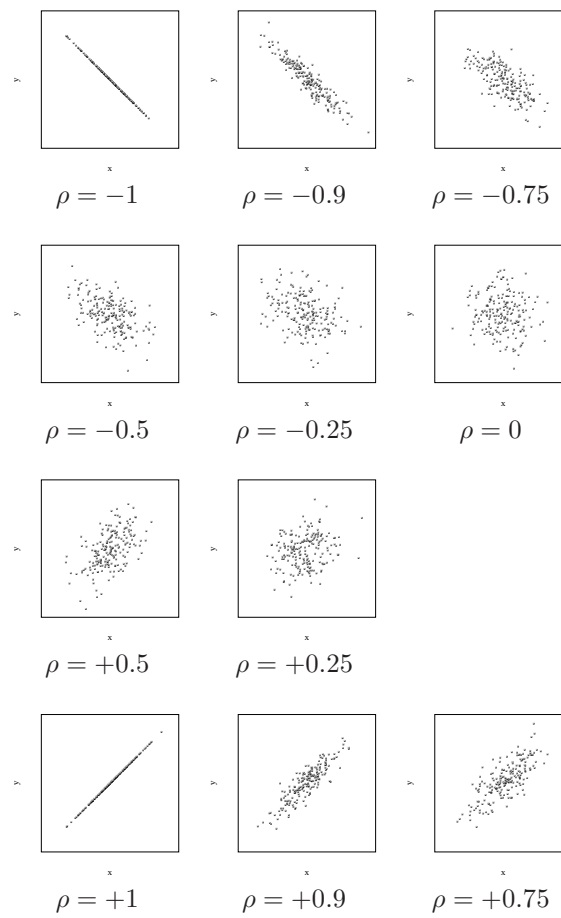
$$-1 \leq \rho \leq +1$$

を満たす .

この定義に基づいて標本相関係数を計算すると , 次のような傾向がある :

- $x$  が大きくなるほど  $y$  も大きくなる時 ,  $\rho > 0$  になる傾向にある .
- $x$  が大きくなるほど  $y$  も小さくなる時 ,  $\rho < 0$  になる傾向にある .
- $x$  と  $y$  が無関係な分布をしているとき ,  $\rho = 0$  に近い値になる傾向にある .

以下に例を示す :



よって、選択肢 4 が正しい。

## \*平成 20 年度 問 34

2つの確率変数  $X, Y$  を考える．不適切な記述を選択肢の中から一つ選べ．ここで，それぞれの記号は次の意味で用いられる．

$P(A, B)$   $A$  と  $B$  の同時確率  
 $E(A)$   $A$  の平均  
 $V(A)$   $A$  の分散

- 1  $X, Y$  が独立のとき， $X, Y$  の同時確率  $P(A, B)$  は，それぞれの確率の積  $P(A)P(B)$  に等しい．
- 2  $Y$  が生じたという条件のもとで  $X$  が生じる条件付確率  $P(X|Y)$  は，次の式で表される：

$$\frac{P(X, Y)}{P(Y)}.$$

- 3  $X$  の分散  $V(X)$  は， $X^2$  の平均と  $X$  の平均の二乗を用いて次のように表される：

$$E(X^2) - (E(X))^2.$$

- 4  $X, Y$  が独立ではないとき，積  $XY$  の平均  $E(XY)$  は，それぞれの平均の積  $E(X)E(Y)$  に等しい．

## 解説

確率分布  $P(X)$  に従う確率変数  $X$  に対して，関数  $f(X)$  の値の期待値は

$$E(f(X)) = \sum_X P(X)f(X)$$

と定義される．

この期待値を使って，様々な統計量が定義されている．  
 代表的なものとして，

$$X \text{ の平均: } M(X) = E(X)$$

$$X \text{ の分散: } V(X) = E(X^2) - (E(X))^2$$

$$X \text{ の標準偏差: } \sigma(X) = \sqrt{V(X)}$$

がある．

同時確率 事象  $X$  と  $Y$  がどちらも生起する確率を同時確率といい， $P(X, Y)$  と書く．

条件付き確率 事象  $Y$  が生起した条件の下で  $X$  が生起する確率を条件付き確率といい， $P(X, Y)$  と書く．

周辺確率 2つの事象  $X$  と  $Y$  を考えたとき，事象  $Y$  にかかわらず，事象  $X$  の確率を，周辺確率と呼ぶ．

選択肢 1 2つの確率変数  $X$  と  $Y$  に対して，

$$P(X, Y) = P(X)P(Y)$$

が成り立つとき，また，その時に限り， $X$  と  $Y$  は統計的独立である，という．よって，この選択肢は正しい．

選択肢 2 ベイズの定理より，

$$P(X|Y) = \frac{P(X, Y)}{P(Y)}$$

が成り立つ．よって，この選択肢は正しい．

選択肢 3 分散の定義より, この選択肢は正しい.

選択肢 4 この選択肢は正しくない. 反例をあげる. ベルヌーイ変数  $X$  と  $Y$  が

$$P_X(X) = \begin{cases} 1/2 & \text{if } X = 1, \\ 1/2 & \text{if } X = 0, \end{cases}$$

$$P_{Y|X}(Y|X) = \begin{cases} 1 & \text{if } Y = X, \\ 0 & \text{if } Y \neq X, \end{cases}$$

に従うとする.

このとき, それぞれの変数の平均値は

$$\begin{aligned} E(X) &= \sum_{X \in \{0,1\}} P_X(X)X \\ &= \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot 0 = \frac{1}{2}, \\ E(Y) &= \sum_{Y \in \{0,1\}} P_Y(Y)Y \\ &= \sum_{Y \in \{0,1\}} \left( \sum_{X \in \{0,1\}} P_{X,Y}(X,Y) \right) Y \\ &= \sum_{Y \in \{0,1\}} \sum_{X \in \{0,1\}} P_{Y|X}(Y|X)P(X)Y \\ &= P_{Y|X}(Y=1|X=1)P(X=1)1 \\ &\quad + P_{Y|X}(Y=1|X=0)P(X=0)1 \\ &\quad + P_{Y|X}(Y=0|X=1)P(X=1)0 \\ &\quad + P_{Y|X}(Y=0|X=0)P(X=0)0 \\ &= 1 \cdot \frac{1}{2} \cdot 1 + 0 \cdot \frac{1}{2} \cdot 1 + 0 \cdot \frac{1}{2} \cdot 0 + 0 \cdot \frac{1}{2} \cdot 0 \\ &= \frac{1}{2}. \end{aligned}$$

となる.

また,  $XY$  の期待値は

$$\begin{aligned} E(XY) &= \sum_{Y \in \{0,1\}} \sum_{X \in \{0,1\}} P_{X,Y}(X,Y)XY \\ &= \sum_{Y \in \{0,1\}} \sum_{X \in \{0,1\}} P_{Y|X}(Y|X)P(X)XY \\ &= P_{Y|X}(Y=1|X=1)P(X=1)1 \cdot 1 \\ &\quad + P_{Y|X}(Y=1|X=0)P(X=0)0 \cdot 1 \\ &\quad + P_{Y|X}(Y=0|X=1)P(X=1)0 \cdot 1 \\ &\quad + P_{Y|X}(Y=0|X=0)P(X=0)0 \cdot 0 \\ &= 1 \cdot \frac{1}{2} \cdot 1 + 0 \cdot \frac{1}{2} \cdot 0 + 0 \cdot \frac{1}{2} \cdot 0 + 0 \cdot \frac{1}{2} \cdot 0 \\ &= \frac{1}{2} \end{aligned}$$

となる.

これは  $E(X)E(Y) = \frac{1}{4}$  と等しくない.

## \*平成 20 年度 問 33

コインを投げ、表が出たか、裏が出たかを調べるという試行を繰り返す。このコインの表、裏が出る確率はそれぞれ  $1/2$  であり、各試行は独立であるとする。この試行に関連する記述として不適切なものを選択肢の中から一つ選べ。

- 1 この試行を繰り返したとき、10 回続けて表が出た。このとき、その次の試行では裏が出る確率が高い。
- 2 このコイン投げを 100 回行ったとき、表が出る回数の期待値は 50 回である。
- 3 5 回の試行、表が 3 回、裏が 2 回出る確率は、

$$\frac{5!}{3!2!} \left(\frac{1}{2}\right)^5$$

と計算される。

- 4 この問題では、コインの表あるいは裏が出ること以外の事象（コインが立つ等）の確率は 0 と仮定されている。

## 解説

コイン投げはベルヌーイ試行の一つであり、その確率分布はベルヌーイ分布と呼ばれる。

ベルヌーイ分布に従う確率変数  $x$  は、0 か 1 の値をとり、パラメータ  $\pi$  を使って

$$\begin{aligned} p(x) &= \begin{cases} \pi & \text{if } x = 1, \\ 1 - \pi & \text{if } x = 0, \end{cases} \\ &= \pi^x (1 - \pi)^{1-x} \end{aligned}$$

で表される。

この設問の場合、表が出る事象に  $x = 1$  を割り当て、裏が出る事象に  $x = 0$  を割り当てるとすると、 $\pi = 1/2$  とおくことができる。

1 回の試行で表が出る回数の期待値は

$$\begin{aligned} E(x) &= \sum_x p(x)x \\ &= \pi \cdot 1 + (1 - \pi) \cdot 0 = \pi \end{aligned}$$

と表される。

選択肢 1 各試行は独立と仮定されているので、確率は前回までの結果に依存しない。

よって、この選択肢に誤りがある。

選択肢 2 直感的にも平均 50 回表が出ると分かるが、数学的には次のように証明できる。

確率  $p(x)$  に従う確率変数  $x$  に対して、 $x$  の期待値は、

$$E(x) = \sum_x p(x)x$$

と定義される。

今回の場合、100 回試行を行ったので、100 個のベルヌーイ分布に従う確率変数  $x_1, \dots, x_{100}$  を考えると、表が出る回数は

$$\sum_{i=1}^{100} x_i$$

で表すことができる。

この期待値をとると，

$$\begin{aligned} E\left(\sum_{i=1}^{100} x_i\right) &= \sum_{i=1}^{100} E(x_i) \\ &= \sum_{i=1}^{100} \pi = 100\pi = 100 \frac{1}{2} = 50 \end{aligned}$$

を得る．

選択肢 3 ベルヌーイ試行を  $n$  回繰り返して，ちょうど  $k$  回成功する (表が出る) 確率は，

$$\frac{n!}{k!(n-k)!} \pi^k (1-\pi)^{n-k}$$

で表されることが知られており，特に 2 項分布と呼ばれる．

この設問では， $n=5$ ， $k=3$  の場合なので，代入すると，確かに，5 回の試行で，表が 3 回，裏が 2 回出る確率は，

$$\frac{5!}{3!2!} \left(\frac{1}{2}\right)^5$$

と表される．

選択肢 4 コインの表，裏が出る確率はそれぞれ  $1/2$  なので，その和は 1 である．

よってそれ以外の事象が生起する確率は 0 である．

## \*平成 19 年度 問 34

サンプルに対する測定結果を用いて、サンプルについての性質（例、罹患、非罹患など）の陽性もしくは陰性を予測することができる計算手法があるとする。以下の表は、独立な 100 件のサンプルについて、この手法を用いて事前に得た予測結果と、厳密な実験で確認した陽性もしくは陰性の確定結果を比較し、それぞれの場合ごとの件数を示したものである。この手法に関する記述として、適切ではないものを選択肢から一つ選べ。

|      |    | 確定結果 |      |
|------|----|------|------|
|      |    | 陽性   | 陰性   |
| 予測結果 | 陽性 | 48 人 | 12 人 |
|      | 陰性 | 2 人  | 38 人 |

- 1 確定結果が陽性であるものについて、およそ 96% の確率で正しく陽性と予測できる。
- 2 陽性と予測されるサンプルのうちおよそ 20% は、実際には陰性である。
- 3 陰性と予測されるサンプルのうちおよそ 5% は、実際には陽性である。
- 4 確定結果が陰性であるものについて、およそ 48% の確率で正しく陰性と予測できる。

## 解説

選択肢 1 確定結果が陽性である人に対して、正しく陽性と予測できた割合は

$$\frac{48}{48 + 2} = 0.96$$

である。よって、この選択肢は正しい。

選択肢 2 陽性と予測された人のうち、実際は陰性であった割合は

$$\frac{12}{48 + 12} = 0.2$$

である。よって、この選択肢は正しい。

選択肢 3 陰性と予測された人のうち、実際は陽性であった割合は

$$\frac{2}{2 + 38} = 0.05$$

である。よって、この選択肢は正しい。

選択肢 4 確定結果が陰性である人に対して、正しく陰性と予測できた割合は

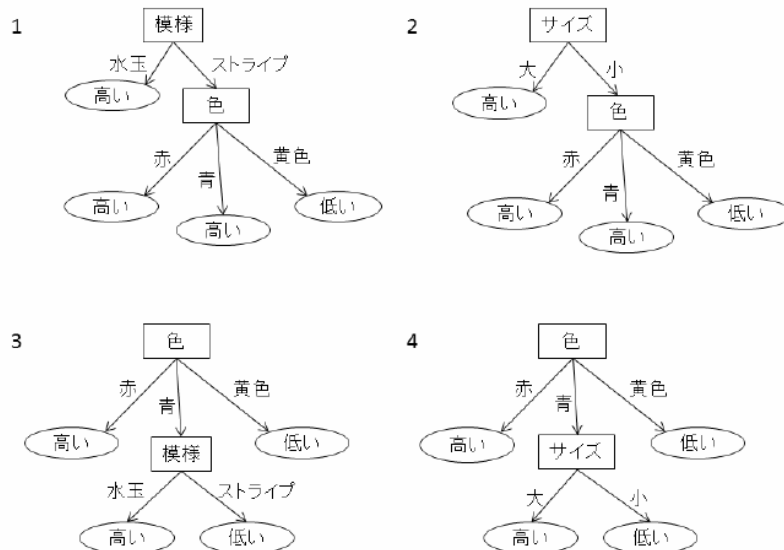
$$\frac{38}{38 + 12} = 0.76$$

である。よって、この選択肢に誤りがある。

## \*平成 19 年度 問 35

意志の決定プロセスなどをグラフを用いて表現する方法に決定木がある。ある商店において、売られている 6 種類のマグカップの売れ行きを調べた結果、次のような結果が得られた。この表を元にして売れ行きを予測するための決定木を作成したとき、もっとも適切なものを選択肢の中から一つ選べ。

|     | マグカップの特徴<br>サイズ 色 模様 |   |       | マグカップの<br>売れ行き |
|-----|----------------------|---|-------|----------------|
| (a) | 大                    | 赤 | ストライプ | 高い             |
| (b) | 小                    | 赤 | 水玉    | 高い             |
| (c) | 大                    | 青 | ストライプ | 低い             |
| (d) | 大                    | 青 | 水玉    | 高い             |
| (e) | 小                    | 黄 | ストライプ | 低い             |
| (f) | 大                    | 黄 | 水玉    | 低い             |



## 解説

H20 問 39 の解説を参照。

### \*平成 19 年度 問 36

次に示した説明文の中で、予測手法の性能評価の際に行われるクロスバリデーション法 (交差検定法) の説明として、不適切なものはどれか、一つ選べ。

- 1 クロスバリデーションは、未知データにも対応できるかを検査する目的で行われる。
- 2 一部のデータを学習に使わずに残しておき、テスト用に用いて予測性能を測定する。
- 3 leave-one-out 法は、データのうち 2 個のみをテスト用に残しておく方法である。
- 4  $n$ -fold 法は、データのうち  $1/n$  をテスト用に残しておく方法である。

#### 解説

データのモデルを学習する際に、しばしば複数のモデルを考えることができる。

交差検証法は、汎化性能が最も良いと予測されるモデルを選択するために使われる。

汎化性能とは、未知データにどれほどフィットするかに関する能力のことをさす。

交差検証法は基本的には次のように行う。

- (1) すでに得られているデータ集合を訓練用集合と検証用集合に分ける。
  - (2) まず、あるモデルに対して、パラメータを訓練用集合にフィットさせる。
  - (3) 得られたパラメータに対して、検証用集合にどれほどフィットしているか測定する。
- 交差検証法として、 $n$ -fold 交差検証法と一つ抜き (leave-one-out) 交差検証法がよく用いられている。

**$n$ -fold 交差検証法** データ集合を  $n$  個のグループに分割し、 $n - 1$  個のグループを訓練用集合として使い、残りの 1 個のグループを検証用集合を用いるものである。

すべてのグループが正確に 1 回だけ検証用集合に使われるよう、これを  $n$  回繰り返し、検証用集合に対する結果を平均して、一つの推定値を得る。

**一つ抜き (leave-one-out) 交差検証法**  $n$ -fold 交差検証法におけるグループ数をサンプル数にした場合、一つ抜き交差検証法となる。

すなわち、データ集合のうち、1 サンプルだけ検証用に使い、残り全てを訓練用に用いる。

すべてのサンプルが正確に 1 回だけ検証用に使われるよう繰り返し、それらの結果を平均して、一つの推定値を得る。

では、選択枝を一つずつ検証する。

選択枝 1 この記述は正しい。

選択枝 2 この記述は正しい。

選択枝 3 leave-one-out 法は、データのうち 1 個のみをテスト用に残しておく方法である。よって、この選択枝は誤りを含む。

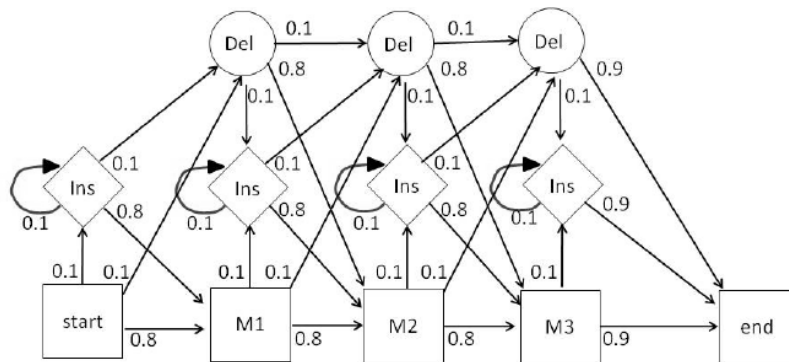
選択枝 4 この記述は正しい。

## \*平成 20 年度 問 40

下図のようなプロファイル HMM がある．状態間の遷移確率は図中に示されている．各状態での文字の出力確率は， $M_1, M_2, M_3$  の各状態では以下になっている．

$$\begin{array}{llll} M_1 : & P(A) = 0.0, & P(T) = 0.5, & P(G) = 0.5, & P(C) = 0.0, \\ M_2 : & P(A) = 1.0, & P(T) = 0.0, & P(G) = 0.0, & P(C) = 0.0, \\ M_3 : & P(A) = 0.0, & P(T) = 0.0, & P(G) = 0.9, & P(C) = 0.1, \end{array}$$

また Ins 状態では，位置によらず， $P(A) = P(T) = P(G) = P(C) = 0.25$  である．この HMM から「start  $\rightarrow$  T  $\rightarrow$  A  $\rightarrow$  G  $\rightarrow$  A  $\rightarrow$  end」という出力文字列が観測されたとする．以下の選択肢に示された状態遷移パスのうち，この出力文字列を発生させる確率がもっとも大きいものを一つ選べ．



- |   |                     | T                 | A                 | G                 | A                     |
|---|---------------------|-------------------|-------------------|-------------------|-----------------------|
| 1 | start $\rightarrow$ | Ins $\rightarrow$ | $M_1 \rightarrow$ | $M_2 \rightarrow$ | $M_3 \rightarrow$ end |
| 2 | start $\rightarrow$ | $M_1 \rightarrow$ | Ins $\rightarrow$ | $M_2 \rightarrow$ | $M_3 \rightarrow$ end |
| 3 | start $\rightarrow$ | $M_1 \rightarrow$ | $M_2 \rightarrow$ | Ins $\rightarrow$ | $M_3 \rightarrow$ end |
| 4 | start $\rightarrow$ | $M_1 \rightarrow$ | $M_2 \rightarrow$ | $M_3 \rightarrow$ | Ins $\rightarrow$ end |

## 解説

プロファイル HMM は，マルコフ過程に基づいている．マルコフ過程において，状態系列  $s_{\text{start}}, s_1, \dots, s_T, s_{\text{end}}$  が与えられたとき，その状態系列から観測系列  $o_1, \dots, o_T$  が生成される確率は

$$P(o_1, \dots, o_T | s_1, \dots, s_T) = P(o_1 | s_1) \dots P(o_T | s_T)$$

とあらされる．

選択肢 1 状態遷移パス start  $\rightarrow$  Ins  $\rightarrow$   $M_1 \rightarrow$   $M_2 \rightarrow$   $M_3 \rightarrow$  end から観測系列 T  $\rightarrow$  A  $\rightarrow$  G  $\rightarrow$  A を生起させる確率は

$$\begin{aligned} P(o_1 | s_1) \dots P(o_T | s_T) &= P(T | \text{Ins})P(A | M_1)P(G | M_2)P(A | M_3) \\ &= 0.25 \cdot 0.0 \cdot 0.0 \cdot 0.0 = 0.0 \end{aligned}$$

選択肢 2 状態遷移パス start  $\rightarrow$   $M_1 \rightarrow$  Ins  $\rightarrow$   $M_2 \rightarrow$   $M_3 \rightarrow$  end から観測系列 T  $\rightarrow$  A  $\rightarrow$  G  $\rightarrow$  A を生起させる確率は

$$\begin{aligned} P(o_1 | s_1) \dots P(o_T | s_T) &= P(T | M_1)P(A | \text{Ins})P(G | M_2)P(A | M_3) \\ &= 0.5 \cdot 0.25 \cdot 0.0 \cdot 0.0 = 0.0 \end{aligned}$$

選択肢 3 状態遷移パス  $\text{start} \rightarrow M_1 \rightarrow M_2 \rightarrow \text{Ins} \rightarrow M_3 \rightarrow \text{end}$  から観測系列  $T \rightarrow A \rightarrow G \rightarrow A$  を生起させる確率は

$$\begin{aligned} P(o_1 | s_1) \dots P(o_T | s_T) &= P(T | M_1)P(A | M_2)P(G | \text{Ins})P(A | M_3) \\ &= 0.5 \cdot 1.0 \cdot 0.25 \cdot 0.0 = 0.0 \end{aligned}$$

選択肢 4 状態遷移パス  $\text{start} \rightarrow M_1 \rightarrow M_2 \rightarrow M_3 \rightarrow \text{Ins} \rightarrow \text{end}$  から観測系列  $T \rightarrow A \rightarrow G \rightarrow A$  を生起させる確率は

$$\begin{aligned} P(o_1 | s_1) \dots P(o_T | s_T) &= P(T | M_1)P(A | M_2)P(G | M_3)P(A | \text{Ins}) \\ &= 0.5 \cdot 1.0 \cdot 0.1 \cdot 0.25 = 1/80 \end{aligned}$$

となる。

よって、選択肢 4 が  $T \rightarrow A \rightarrow G \rightarrow A$  を生起させる確率が最も大きい。

### \*平成 20 年度 問 36

次のクラスタ解析の手法の中で非階層的クラスタリングではないものを選択肢の中から一つ選べ．

- 1  $K$ -平均法
- 2 自己組織化マップ
- 3 Fuzzy  $c$ -means
- 4 UPGMA

#### 解説

UPGMA は階層的クラスタリングの一つで，系統樹を作る．よって，UPGMA が非階層的クラスタリングではない．

\*問題文はバイオインフォマティクス技術者認定試験（日本バイオインフォマティクス学会主催）問題（平成19年度または平成20年度）から引用

平成19 年度日本バイオインフォマティクス学会（JSBi）バイオインフォマティクス技術者認定試験試験問題  
Copyright© 2007 Japanese Society for Bioinformatics. All Rights Reserved. :

[http://www.jsbi.org/modules/jsbi/index.php/nintei/H19/H19mondai\\_kaitou.pdf](http://www.jsbi.org/modules/jsbi/index.php/nintei/H19/H19mondai_kaitou.pdf)

平成20 年度日本バイオインフォマティクス学会（JSBi）バイオインフォマティクス技術者認定試験試験問題  
Copyright©2008 Japanese Society for Bioinformatics. All Rights Reserved. :

[http://www.jsbi.org/modules/jsbi/index.php/nintei/H20/H20mondai\\_kaitou.pdf](http://www.jsbi.org/modules/jsbi/index.php/nintei/H20/H20mondai_kaitou.pdf)

日本バイオインフォマティクス学会： <http://www.jsbi.org/>