期待値、確率分布

1　確率試行と確率変数

確率試行コイン投げ**確率変数**とは実験の結果 Ω から数値（実数）を対応させるもの一般に $\{X \leq a\} = \{\omega \in \Omega | X(\omega) \leq a\}$ と表し，これが事象となるもの。すなわち $\{X \leq a\} \in \mathcal{F}$ とします。

５枚のコインを投げて，表のでた枚数を X とおけば，各結果を H（表，Head），T（裏，Tail）を使うと，$\{X = 0\} = \{TTTTT\}$, $\{X = 2\} = \{HHTTT, HTHTT, \cdots, TTTHH\}$, $\{X = 3\} = \{HHHTT, HHTHT, \cdots, TTTHH\}$ となります。また $\{X \leq 2\} = \{X = 0\} \cup \{X = 1\} \cup \{X = 2\}$, $\{X \leq \sqrt{2}\} = \{X = 0\} \cup \{X = 1\}$ などが成り立ちます。これは集合の包含関係で要素が同じであることを表しています。

2　確率分布

コイン投げやさいころ振りを繰り返して、この実験結果を整理まとめていくと、ある傾向が求められます。たとえば，５枚のコインを投げて，何枚表が出たかを記録してします。大体このような実験では、確率的に変動しますから、起こりやすいところ、起こらないところの大小で、高さを描いたグラフが記述統計で調べた度数分布表です。これに対して，理論的にあるいは仮想的にそれぞれそのくらいの可能性があるかを計算します。これが，コイン投げの確率分布です。コイン投げの分布では，可能性をできるだけ調べ上げて，当てはまるの事象の大きさを測ります。つまり確率を計算します。このときによく用いる概念が２項係数などで，場合の数を効率よく数え上げようとするものです。上の例でいえば，2^5 通りの結果から，表が１枚出る事象は５通りあるので，$P(X = 1) = \frac{5}{2^5}$ となります。

和の分布；数枚のコインを投げる場合では，各コインの結果（表が１，裏がゼロ）をまとめて，和を取ると，表の出た枚数となります。一般にこれは和の分布を計算することです。独立な確率変数について，

$$P(X_1 + X_2 = a) = \sum_k P(X_1 = k) P(X_2 = a - k)$$

同時分布（結合分布）；２変量 (X, Y) の結果について，積事象 $\{X = a\} \cap \{Y = b\}$ を $(X, Y) = (a, b)$ と表し，$f_{X,Y}(a, b) = P(X = a, Y = b)$ を同時密度とよびます。周辺分布とは $f_X(a) = \sum_b f_{X,Y}(a, b), f_Y(b) = \sum_a f_{X,Y}(a, b)$ をいいます。行列の表形式で与えられた２次元の度数分布表から周辺度数（縦計や横計）を求めたことに相当します。また和の分布は

$$P(X_1 + X_2 = a) = \sum_{(x,y)|x+y=a} f_{X,Y}(x, y)$$

という意味で，ここで和 $\sum$ は $x + y = a$ となるすべての (x, y) にわたる和をとります。最後の項が独立であれば，確率の積になります。

以上は離散型確率変数ですが，連続型の場合もほぼ同様です。

離散型確率変数；

$$P(a \leq X \leq b) = \sum_{\{k|a \leq k \leq b\}} P(X = k)$$

連続型確率変数；

$$P(a \leq X \leq b) = \int_a^b f_X(x) dx$$

ここで $F_X(x) = P(X \leq x)$ とおき，微分ができるばあい，その微分係数（導関数）を $f_X(c) = \frac{dF_X(x)}{dx}|_{x=c}$ とします。

3　期待値

平均の定義式；

$$E(X) = \begin{cases} \sum_i x_i f_X(x_i) & \text{離散型} \\ \int_{-\infty}^{\infty} x f_X(x) dx & \text{連続型} \end{cases}$$

離散型 では $E[(X-a)^2] = \sum_i (x_i - a)^2 f_X(x_i) = \sum_i x_i^2 f_X(x_i) - 2a \sum_i x_i f_X(x_i) + a^2$ 連続型では $E[(X-a)^2] = \int_{-\infty}^{\infty} (x-a)^2 f_X(x) dx = \int_{-\infty}^{\infty} x^2 f_X(x) dx - 2a \int_{-\infty}^{\infty} x f_X(x) dx + a^2$ となります。
分散 $V(X)$ の定義は，上の式で，$a = E(X)$ とおいたもので，

$$V(X) = \begin{cases} \sum_i x_i^2 f_X(x_i) - \{E(X)\}^2 & \text{離散型} \\ \int_{-\infty}^{\infty} x^2 f_X(x) dx - \{E(X)\}^2 & \text{連続型} \end{cases}$$

一般に２変量確率変数 (X, Y) から定めた実数値関数の確率変数 $h(X, Y)$ の期待値は

$$E(h(X,Y)) = \begin{cases} \sum_i \sum_j h(x_i, y_j) f_{X,Y}(x_i, y_j) & \text{離散型} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x,y) f_{X,Y}(x,y) dx dy & \text{連続型} \end{cases}$$

２重積分やら，２重和をとることが入って，すこし難しいかも知れませんが，１変量が面積の計算，２変量が体積の計算をしているだけです。平均は重心の計算と同じです。和の分散には，$h(x,y) = \{(x+y) - a\}^2$ として計算します。ここで $a = E(X+Y)$ は既に定数となっていることに注意。

問 3.1
X が平均 μ, 分散 σ^2 をもつとき，つぎの値を μ, σ^2 で表しなさい。

(1) $E(3X+2)$　　(2) $E(X^2)$
(3) $E(2X^2+2X-1)$　　(4) $E[(X-2)^2]$

問 3.2
分散に関するつぎの式を導きなさい。

(1) $V(aX+b) = a^2 V(X)$ (b にはよらない)

(2) X, Y が独立ならば，$V(X-Y) = V(X) + V(Y)$ (符号に注意，差の分散でも分散の和になる)

問 3.3
つぎの表は離散型確率変数の結合密度を表すとします。

y の値		1	2	3
x	0	0.15	0.10	0.15
の	1	0.10	0.0	0.10
値	2	0.15	0.10	0.15

このとき，

(1) $E(XY) = E(X)E(Y)$ が成り立つが, 独立ではないことを示せ。

(2) $V(X+Y)$ を計算しなさい。

4　コインとサイコロ

コインを投げると，表か裏のいずれか，2 通りのうちのひとつが結果としておこる。結果を表す変数を確率変数といい，この変数がとり得る値の集まりを考え，その起こりえる可能性を尺度化することで，確率分布が定められる。コイン投げの 2 通りの結果がある場合をベルヌーイ分布といい，サイコロ投げのように 6 通りの目の出方が同じ確からしさをもつ場合，一様分布とよぶ。

5　２項分布

表の出る確率が p のコインを一枚投げる実験では，$X = \begin{cases} 1, & p \\ 0, & 1-p \end{cases}$ で，これを n 枚投げる（繰り返す）ことは $Y_n = X_1 + X_2 + \cdots + X_n$ とおき，$k(k = 0, 1, 2, \cdots, n)$ 枚表が出る事象の確率 $P(Y_n = k)$ は $X_i(i = 1, 2, \cdots, n)$ のなかで 1 となっているものの数え上げであるから 2 項係数で表現できる。

$$P(Y_n = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, 2, \cdots, n$$

これをパラメータ n, p の 2 項分布という。 2 項分布の平均 μ と分散 σ^2 は

$$\mu = np, \quad \sigma^2 = np(1-p)$$

とくに $n = 1$ はパラメータ p のベルヌイ分布とよばれる。ベルヌイ分布や 2 項分布から生じるさまざまな現象は，ランダムウオーク（酔歩），正規分布やポアソン分布などいろいろな分布の解析に結びつく。

6　超幾何分布

つぼの中からボールを取り出すとき，取り出されたボールを元のつぼに戻すかどうかで次に取り出すボールの結果に影響を与える。復元抽出と非復元抽出である。ボールの総個数が有限個しかないならば，復元抽出は同じ状況の繰り返しであるが，非復元の場合には取り出しの度毎に変わっていく。条件付きの確率を考えることになる。ボールの総個数を N とし，2 種類のボール，たとえば赤 r と黒 $N - r$ のボールがあるとき n 個のボールを非復元抽出するとき，この中に赤ボールが X 個含まれる，つまり黒ボールが $n - X$ 個となる確率，

$$P(X = k) = \frac{\binom{r}{k}\binom{N-r}{n-k}}{\binom{N}{n}}, \quad k = 0, 1, 2, \cdots, r$$

をパラメータ N, n, r の超幾何分布という。ボールの総数 N がかなり大きいときには，ボールをもとに戻してもつぎのボールの取出しにはほとんど影響ない。すなわち，同じ状況の繰り返しであるから，これはパラメータ n, p の 2 項分布になる。復元抽出（繰り返しのばあい）には 2 項分布であるが，非復元抽出（もとに戻さない取り出し）では，超幾何分布である。ここで $\lim_N r/N = p, \lim_N (N-r)/N = 1-p$ それぞれ赤ボール，黒ボールの比率を表す。上の超幾何分布の確率は $N \to \infty$ とすれば，2 項分布の確率に近づく。