考える習慣を身に着ける

数理統計学入門*

慶應義塾大学先端生命科学研究所・環境情報学部
専任講師 斎藤輪太郎

平成 22 年 3 月 15 日

1 離散確率分布

変数 X が様々な値を取り、またどのような値を取るかが確率的に決まる時、X を確率変数と呼ぶ。例えばサイコロを振ったときに出た目を X とすると、X は確率的に決まるため、確率変数である。X がある値 x を取る確率を $P(X = x)$ と表す。例えば X をサイコロを振ったときに出る目とすれば、サイコロの目が 2 になる確率は $P(X = 2) = 1/6$ と表される。誤解を招く恐れがないときは、$P(x)$ と省略する。

X が取り得る値 $x_1, x_2, x_3, \cdots$ を横軸に、その値を取る確率 $p_1, p_2, p_3 \cdots$ を縦軸にしたヒストグラムを描くと、変数 X の特性が分かる。例えば横軸にサイコロの目、縦軸にその目が出る確率をヒストグラムで表すと図 1(a) のようになるし、またサイコロを 3 回振ったときの目の合計を横軸、その確率を縦軸にすると、図 1(b) のようになる。このように変数 X の確率的な分布を表したものを確率分布と呼ぶ。

図 1 の場合、ヒストグラムの各棒の幅を 1 とすれば、各棒の面積が確率変数が取り得る各値の確率となり、さらに以下の 2 つの重要な性質が得られる。

1. 棒の面積の合計が 1 となる
2. X が a 以上、b 未満となる確率 $P(a \leq X < b)$ は、以下の式で与えられる。

$$P(a \leq X < b) = \sum_{a \leq x_i < b} P(X = x_i) \tag{1}$$

逆に必要条件として、確率分布は上記 2 つの性質を満たすもの、と約束する。こうすることによって様々な確率分布に対してその特性の計算方法を統一できる。サイコロを 3 回振る例で、目の合計が 9 以上 11 以下になる確率は $P(9) + P(10) + P(11)$、すなわち図 1 の色の濃い棒の面積の和で表されることが分かる。

*To my wife and Dr. Nozomu Yachie.

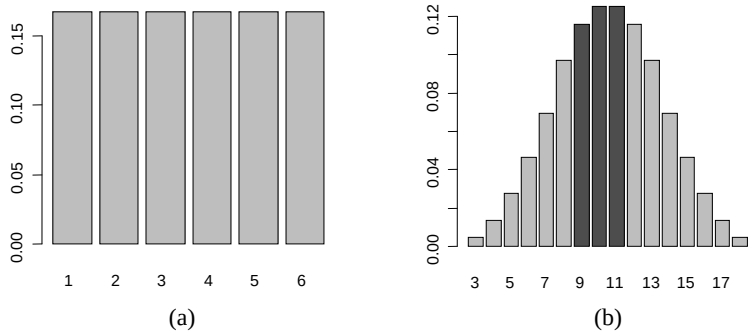

図 1: サイコロを例にした確率分布の例

(a) サイコロの各目が出る確率 (b) サイコロを 3 回振ったときの目の合計とその確率。合計が 9 以上 11 以下の確率に該当する棒の色を濃く示した。

上記の例では確率分布をヒストグラムで表したが、確率変数 X が値 x を取る確率 $P(x)$ を数式で表現できれば最も客観的である。これを考えると、図 1(a) の分布は $P(x) = \frac{1}{6}$ と表されるし、図 1(b) の分布は $P(x) = \sum_{i=0}^{\lfloor \frac{1}{6}x-2 \rfloor} \frac{3(x-6i-1)(x-6i-2)}{i!(3-i)!}(-1)^i$ である[1]。但し、$\lfloor \alpha \rfloor$ は α 以下の最大の整数を表す (小数以下を切捨てた整数)。このように確率変数 X が各値 x を取る確率を関数で表したものを確率関数と呼ぶ。

確率分布と度数分布の間には密接な関係がある。N 回の試行で $X = x_i$ となる現象が n_i 回起こったとする。このとき、N が大きくなるにつれて、n_i/N は $P(X = x_i)$ に近づいてゆく。

問題 1-1:各年に特定の都市 A で大地震が起きる確率は $\frac{1}{60}$ であるとする。今年の始めから数えて x 年の間に 1 回も地震が起ない確率を $P(x)$ としたとき、$P(x)$ の確率関数を求めよ。

2 連続確率分布

サイコロの例では確率変数 X は離散値、すなわち 1,2,3,4,5,6 と整数値を取り、2.5 など間の数を取ることはなかったし、1 から 6 という範囲の中で取り得る値は 6 個と有限であった。確かに現実で起こる確率的と思われる現象の中には、気象を例に取ればある地域における年間の台風発生数、落雷数など離散的な数で表されるものもある。しかし、降水量や平均気温など取り得る値が実数で無限にある確率変数も存在する。これを連続量と呼ぶが、連続量に対しては確率分布や確率関数をどのように考えればよいだろうか。図 2 の”サイコロルーレット”を例にとって説明しよう。この装

[1] より一般的にはサイコロの目の数を q(一般的なサイコロは $q = 6$)、振る回数を n として、

$$P(x) = \sum_{i=0}^{\lfloor \frac{x-n}{q} \rfloor} \frac{(x-qi-1)!}{(n-1)!(x-qi-n)!} \cdot \frac{n!}{i!(n-i)!}(-1)^i \tag{2}$$

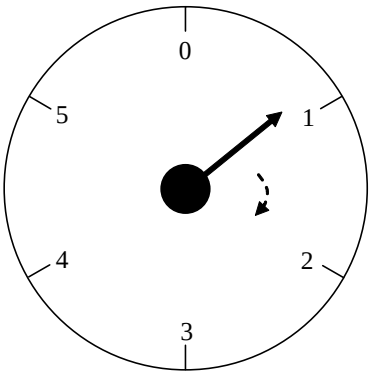

図 2: サイコロルーレット

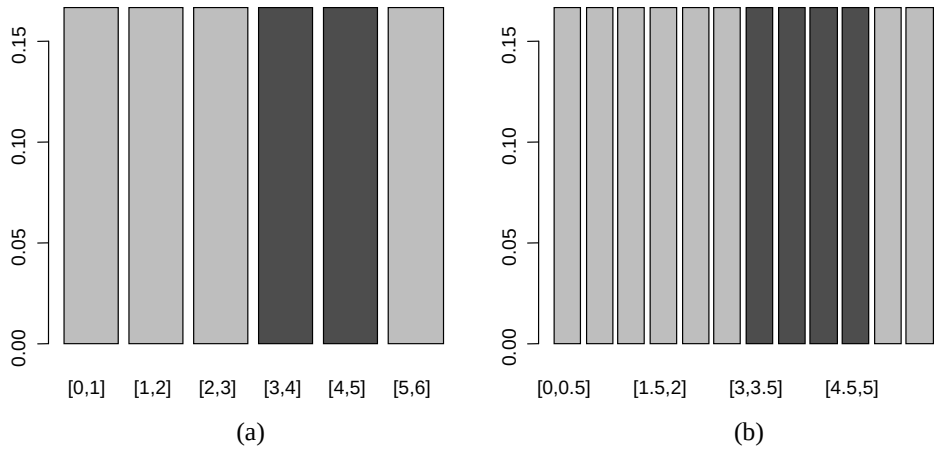

図 3: 連続分布の離散分布への変換

置は円盤と自由に動く針で構成されている。円盤には 0 から 5 までの数が刻んである。針を力強く弾くと、針は何回転もした後、摩擦で停止する。停止したときに針が指し示す実数を X とすれば、針は 0 と 1 の間などどこにでも停止しうるため、X は連続値となる。確率変数が離散値を取るときと同じような論法で X が例えば 3 を取る確率を論じることはできない。針が寸分の狂いもなくちょうど 3 のところで停止することはあり得ないからである (確率は 0)。

そこで連続値を離散値に直すことを考えてみよう。例えば、サイコロルーレットの針が差す数値が 0 以上 1 未満になる確率、1 以上 2 未満になる確率、... というように連続値を各範囲に小分けにして離散値として扱ってはどうだろうか。すると図 3(a) に示すように、X の取り得る範囲は 6 つの領域に分かれ、各範囲内に X が収まる確率は全て $\frac{1}{6}$ となる。またこの確率関数は、p.1 で述べた確率分布の 2 つの性質を満たす。すなわち、

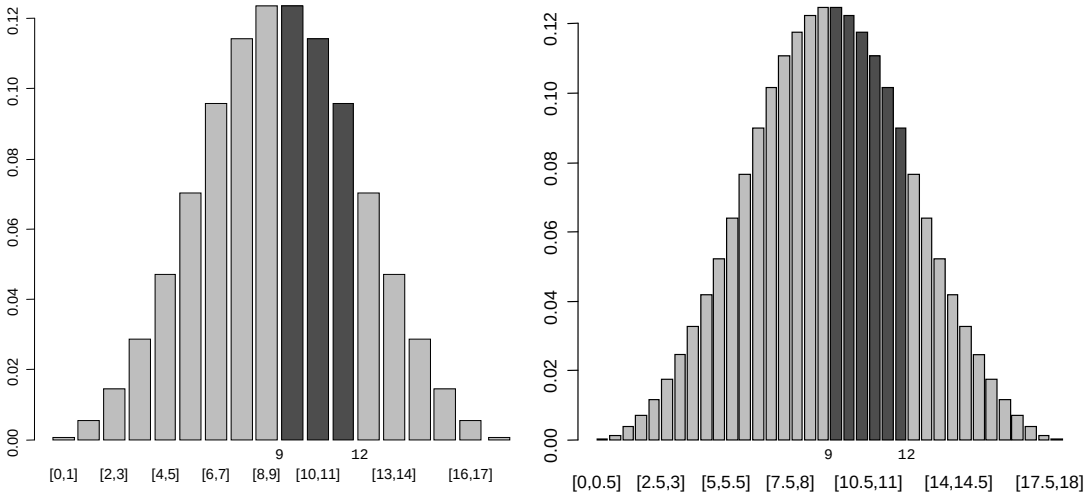

図 4: サイコロルーレットを 3 回まわしたときの合計値の離散確率分布による近似

$$P(a \leq X < b) = \sum_{a \leq i < b} P(i \leq X < i+1) \tag{3}$$

なので 2 番目の性質は満たされている。但し、a, b, i は全て整数で、$0 \leq a < 6, 0 < b \leq 6$ である。また、

$$P(0 \leq X < 6) = \sum_{0 \leq i < 6} P(i \leq X < i+1) = 1 \tag{4}$$

より、1 の性質も満たす。X が 3 以上 5 未満になる確率は濃い色の棒の面積の合計で表されることも確認されたい。

区画をさらに細かくして、0 以上 0.5 未満、0.5 以上 1 未満、$\cdots$ とすれば、図 3(b) のような分布が出来上がる。区間が小さくなれば、X がその区画に入る確率は当然小さくなってゆく。このとき注意しなければならないのは、区間の幅 $\Delta x = 0.5$ となっているため、$X = x$ の地点における棒の高さ=その地点の確率としてしまうと、棒の面積は棒の高さ $\times 0.5$ により、実際の確率の半分になってしまう。そこで (X が入る区画の確率/Δx) という補正をかける必要がある。

これをサイコロルーレットを 3 回回したときの合計値の分布で考えてみよう。区画幅を 1 にしたときは図 4(a) のような分布になり、区間幅を 0.5 とより細かくしたときは、図 4(b) のようによりスムーズな曲線に近づく。区間幅 Δx を極限まで小さくしていったときに、幅が限りなく 0 に近い棒の高さが一定値に収束するならば、それこそが連続確率分布となる。

例えば図 3 の分布は図 5(a) の分布に収束するし、図 4 の分布は、図 5(b) の分布に収束する[2]。

[2]この連続分布は $f(x) = \frac{1}{2(n-1)!} \sum_{k=0}^{n} (-1)^k \frac{n!}{k!(n-k)!} (x-k)^{n-1} \mathrm{sgn}(x-k)$ で表される [2, 3]。

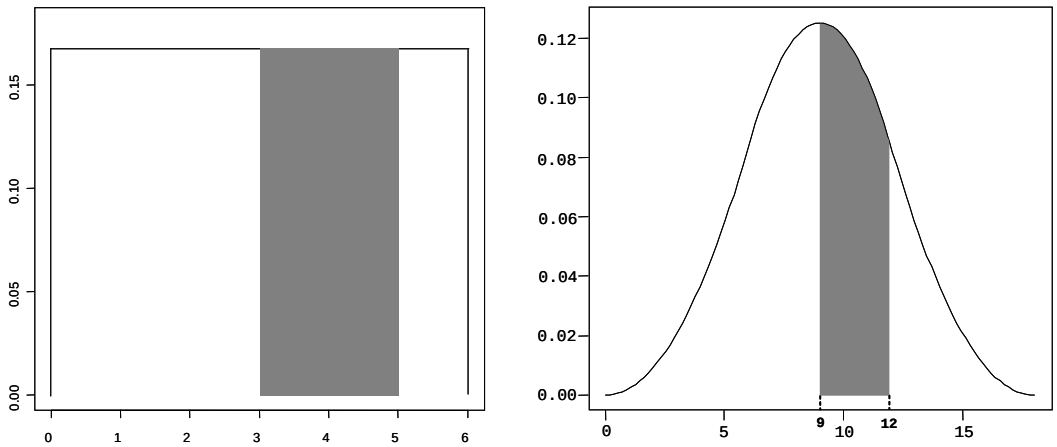

図 5: サイコロルーレットの合計値の連続分布

さてこれらの分布を幅が限りなく 0 に近い棒の集合と考えるならば、離散確率分布のところで述べた 2 番目の性質から、

$$P(a \leq X < b) = \sum_{a \leq x_i < b} P(X = x_i) = \sum_{a \leq x_i < b} f(x_i)\Delta x \tag{5}$$

式 5 は $\Delta x \to 0$ のとき、積分の定義から、

$$P(a \leq X < b) = \int_a^b f(x)dx \tag{6}$$

となる。つまり、確率変数がある範囲の値を取る確率は、確率密度関数の該当面積、すなわち積分で表されるということである。節 1 で述べた離散確率分布の 2 つの重要な性質を連続確率分布に当てはめると、確率密度関数を $f(x)$ として、以下のようになる。

1. 確率変数が取りうる全範囲の確率の合計は 1 となる。すなわち、

$$\int_{-\infty}^{+\infty} f(x)dx = 1 \tag{7}$$

2. X が a 以上、b 未満となる確率 $P(a \leq X < b)$ は、以下の式で与えられる。

$$P(a \leq X < b) = \int_a^b f(x)dx \tag{8}$$

問題 2-1: 確率変数 X の従う確率密度関数 $f(x)$ が

$$f(x) = \begin{cases} x & \text{for } \ 0 \leq x \leq \sqrt{2} \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

で与えられるとき、X が 0 以上 1 以下になる確率を求めよ。

3　平均と分散

また X が取り得る値 $x_1, x_2, x_3, \cdots$ とその確率 $P(X = x_1), P(X = x_2), P(X = x_3), \cdots$ の積の総和を X の期待値と呼び、$E(X)$ で表す。

$$E(X) = \sum_i x_i P(X = x_i) \tag{10}$$

例えば先に出たサイコロの例では、期待値は 3.5 となる。以上は X が離散値の場合だが、連続値の場合は、X の取りうる範囲を Δx の大きさに分割すると、

$$E(X) = \sum_i x_i f(x_i) \Delta x \tag{11}$$

となる。$\Delta x \to 0$ を取れば式 11 は、

$$E(X) = \int_{-\infty}^{+\infty} x f(x) dx \tag{12}$$

に収束する。$E(X)$ を X の平均とも呼ぶ。特に X が密度関数 $f(x)$ に従うとき、$E(X)$ を $f(x)$ の平均という。

分散 σ^2 は $E((X - E(X))^2)$ で定義され、離散確率の場合は平均を $E(X) = \mu$ としたとき、、

$$\sigma^2 = E((X - E(X))^2) = \sum_i (x_i - \mu)^2 P(X = x_i) \tag{13}$$

で表される。

問題 3-1: 連続確率について、分散を表す式を導け。
問題 3-2: 確率変数 $X_1, X_2, \cdots, X_n$ について、以下の数式が成立することを示せ。

$$E(X_1 + X_2 + \cdots + X_n) = E(X_1) + E(X_2) + \cdots + E(X_n) \tag{14}$$

問題 3-3: 以下の数式が成立することを示せ。

$$\sigma^2 = E((X - E(X))^2) = E(X^2) - \{E(X)\}^2 = E(X^2) - \mu^2 \tag{15}$$

4　標本分布

ある確率変数 X の性質を調べたいときに、X がどのような値を取るかを次々に見ていく方法がある。例えば X を日本人 20 歳男子の身長だとしよう。確実に分布を得るためには、日本人 20 歳男子全ての身長を測る方法があるだろう。しかしながらこれは時間・費用・労力の面から効率の良い方法とは言えないし、多くの場合現実的ではないだろう。そこで多くの場合、無作為に標本を取り、

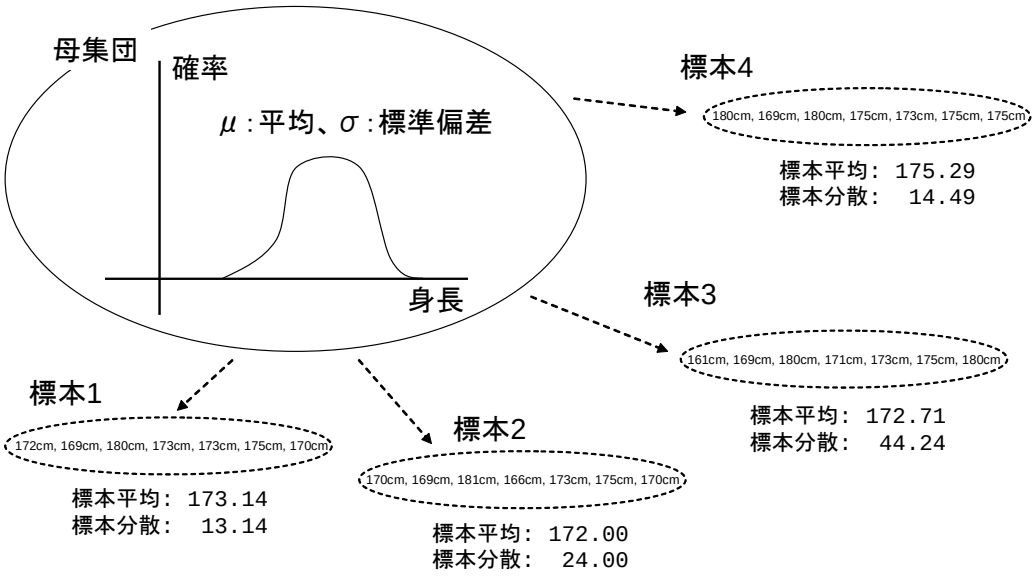

図 6: 身長の平均の例

その標本の性質から全体の性質を推測することが行われる。標本のことから母数のことがどれほどのことがどれほどの精度で分かるだろうか?

まずは母数の身長の平均値 μ とその標本 i の標本中の平均 $\bar{x_i}$ との関係を調べよう。例えば 7 人分の標本を採れば、その標本に対して標本平均 $\bar{x}$ と標本分散 s^2 を計算できる。この $\bar{x}$ や s^2 はどのような性質を持っているだろうか。まずは $\bar{x}$ の方から見てみよう。$\bar{x}$ の性質を調べるためには、$\bar{x}$ 自体の分布、すなわち 7 人分の標本を採るという操作を繰り返し行ったとき、$\bar{x}$ がどのような分布に従うかを調べるのがよい。図 6 では 7 人分の標本を何回も採っている。

5 正規分布

$N(\mu, \sigma^2)$ に従う正規分布 $f(x)$ は

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{16}$$

で表される。

数式の中に何故円周率の π が出てくるのか、不思議に思う読者のために、直感的な説明を以下に与えよう。式 16 を簡略化した

$$f(x) = e^{-x^2} \tag{17}$$

で与えられる関数を考え、その積分

$$I = \int_0^{+\infty} f(x)dx = \int_0^{+\infty} e^{-x^2} dx \tag{18}$$

を定義し、また $y = xu$ を考える。式 18 の x を y に変えても式全体の値は変わらないからこれを自乗して[3]、

$$I^2 = \left(\int_0^{+\infty} e^{-x^2} dx \right)^2 = \left(\int_0^{+\infty} e^{-x^2} dx \right) \left(\int_0^{+\infty} e^{-y^2} dy \right) = \int_0^{+\infty} \int_0^{+\infty} e^{-(x^2+y^2)} dy dx \quad (19)$$

ここで $\frac{dy}{du} = x$ すなわち $dy = xdu$ より、

$$\begin{aligned} I^2 &= \int_0^{+\infty} \int_0^{+\infty} e^{-(x^2+y^2)} dy dx = \int_0^{+\infty} \left(\int_0^{+\infty} e^{-(1+u^2)x^2} x du \right) dx \\ &= \int_0^{+\infty} \int_0^{+\infty} e^{-(1+u^2)x^2} x dx du = \int_0^{+\infty} \left[-\frac{1}{2(1+u^2)} \cdot e^{-(1+u^2)x^2} \right]_0^{+\infty} du \\ &= \int_0^{+\infty} \frac{du}{2(1+u^2)} = \frac{1}{2} \left[\tan^{-1} u \right]_0^{+\infty} = \frac{\pi}{4} \end{aligned} \quad (20)$$

6 多変量の確率関数

確率関数は多変量に対しても拡張できる。まずは 2 変量で考えてみよう。確率密度関数を $f(x, y)$ とすれば、節 2 で述べた確率密度関数の性質を以下のように拡張できる。

1. 確率変数が取りうる全範囲の確率の合計は 1 となる。すなわち、

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y) dx dy = 1 \quad (21)$$

2. X, Y が範囲 D の値に収まる確率は以下の式で与えられる。

$$P((X, Y) \in D) = \int \int_D f(x, y) dx dy \quad (22)$$

図形的に表せば図 7 のように、X, Y が領域 D に入る確率は領域 D と $f(x, y)$ の部分で囲まれた空間の体積で表される。すなわち、

$$P(x_i \leq X < x_i + \Delta x \wedge y_i \leq Y < y_i + \Delta y) \approx f(x_i, y_i) \Delta x \Delta y \quad (23)$$

領域 D に $\Delta x, \Delta y$ の長方形を敷き詰め、i 番目の長方形の隅の座標を $x_i, y_i \in D$ として合計すると、$\sum_{x_i, y_i \in D} f(x_i, y_i) \Delta x \Delta y$ であり、$\Delta x, \Delta y \to 0$ のとき、式 22 に収束する。

[3] $y = xu$ の制約の中でも、u の値を調整すれば、x, y は正のあらゆる組み合わせの値を取り得ることに注意。式の中に 3 つの変数全て (x, y, u) が登場すると、これらの 3 つの変数の取り得る値の組み合わせは限定される。

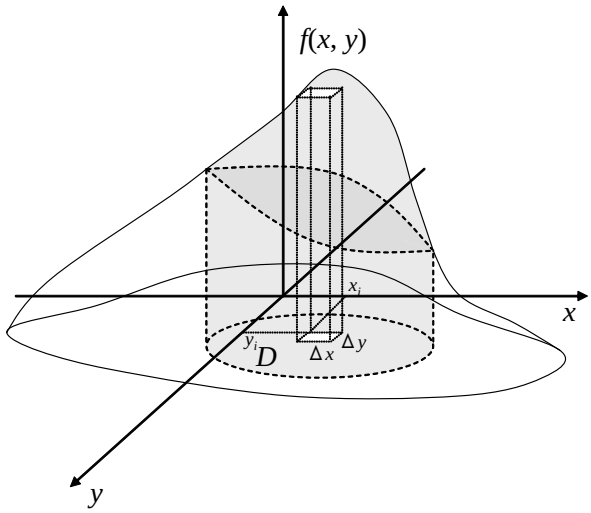

図 7: 3 次元空間上の連続分布

7　テイラー展開

$f(x), f'(x), \cdots, f^{(n-1)}(x)$ は $[a,b]$ で連続で (a,b) で $f^{(n)}(x)$ が存在すれば、

$$f(b) = f(a) + \frac{f'(a)}{1!}(b-a) + \frac{f''(a)}{2!}(b-a)^2 + \cdots + \frac{f^{(n-1)}(a)}{(n-1)!}(b-a)^{n-1} + \frac{f^{(n)}(c)}{n!}(b-a)^n \quad (a < c < b) \tag{24}$$

となるような c が存在する。

証明：

$$f(b) = f(a) + \frac{f'(a)}{1!}(b-a) + \frac{f''(a)}{2!}(b-a)^2 + \cdots + \frac{f^{(n-1)}(a)}{(n-1)!}(b-a)^{n-1} + \frac{k}{n!}(b-a)^n \tag{25}$$

とおく。k 以外の変数が固定されていても、k の値を調整すれば、式 25 を成立させることができる。右辺の a を x に置き換え、$f(b)$ から引くと、

$$F(x) = f(b) - f(x) - \frac{f'(x)}{1!}(b-x) - \frac{f''(x)}{2!}(b-x)^2 - \cdots - \frac{f^{(n-1)}(x)}{(n-1)!}(b-x)^{n-1} + \frac{k}{n!}(b-x)^n \tag{26}$$

$F(a) = 0, F(b) = 0$ であり、Rolle の定理から $F'(c), a < c < b$ となる c が存在する。

$$F'(c) = -\frac{f^{(n)}(c)}{(n-1)!}(b-c)^{n-1} + \frac{k}{(n-1)!}(b-x)^{n-1} = 0 \tag{27}$$

より $k = f^{(n)}(c)$ を得る。

□

8 積率母関数

8.1 積率母関数の定義

x を確率変数とし、t を実数とするとき、e^{tx} の期待値を x の積率母関数 (moment generating function) といい、$M_x(t)$ で表す。すなわち x が離散型であれば、$f(x)$ を確率関数として、

$$M_x(t) = E(e^{tx}) = \sum_x e^{tx} f(x) \tag{28}$$

となる。但し、σ_x は x の可能な値全てについての合計を表す。また x が連続型で $-\infty < x < \infty$ とすれば、$f(x)$ を密度関数として、

$$M_x(t) = E(e^{tx}) = \int_{-\infty}^{\infty} e^{tx} f(x) dx \tag{29}$$

である。

e^{tx} をマクローリン展開すれば、

$$e^{tx} = 1 + \frac{tx}{1!} + \frac{(tx)^2}{2!} + \cdots \tag{30}$$

であるから、これを式 29 に代入して、

$$\begin{aligned} E(e^{tx}) &= \int_{-\infty}^{\infty} f(x)dx + \int_{-\infty}^{\infty} txf(x)dx + \frac{1}{2}\int_{-\infty}^{\infty} (tx)^2 f(x)dx + \cdots \\ &= 1 + tE(x) + \frac{t^2}{2}E(x^2) + \cdots = \sum_{k=0}^{\infty} \frac{t^k}{k!} E(x^k) \end{aligned} \tag{31}$$

となる。すなわち各項は原点回りの k 次の積率 $E(x^k)$ を含んでいる。

式 31 に関する k 階の導関数で $t = 0$ とおくとその値は $E(x^k)$ すなわち原点回りの k の積率に等しくなる。すなわち、

$$\left.\frac{\partial^k M_x(t)}{\partial t^k}\right|_{t=0} = E(x^k) \qquad k = 1, 2, \cdots \tag{32}$$

これが $M_x(t)$ が積率母関数と呼ばれる所以である。以下、$\partial^k M_x(t)/\partial t^k|_{t=0}$ を単に $M_x^{(k)}(0)$ と表すことにする。例えば $k = 1$ であれば、

$$M_x'(0) = E(x)$$

で平均が得られる。また $k = 2$ ならば、

$$M_x''(0) = E(x^2)$$

すなわち、原点周りの 2 次の積率が得られるから、

$$E(x^2) - E(x)^2 = \sigma^2$$

によって分散が得られる。

例題 8-1: 正規分布 $N(\mu, \sigma^2)$ の積率母関数を求めてみよう。

$$\begin{aligned} M_x(t) &= \int_{\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx \\ &= \int_{\infty}^{\infty} \exp\left[-\frac{1}{2\sigma^2}\{(x-\mu)^2 - 2\sigma^2 tx\}\right] dx \end{aligned}$$

ところで $\{\cdots\}$ の中は $(x-\mu)^2 - 2\sigma^2 tx = x^2 - 2\mu x + \mu^2 - 2\sigma^2 tx = x^2 - 2(\mu - \sigma^2 t)x + \mu^2 = \{x - (\mu + \sigma^2 t)\}^2 - 2\sigma^2\mu t - \sigma^4 t^2$ であるから、

$$\begin{aligned} M_x(t) &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2\sigma^2}\left\{x - (\mu + \sigma^2 t)\right\}^2 + \mu t + \frac{\sigma^2}{2} t^2\right] dx \\ &= \exp\left(\mu t + \frac{\sigma^2}{2} t^2\right) \cdot \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2\sigma^2}\left\{x - (\mu + \sigma^2 t)\right\}^2\right] dx \end{aligned}$$

ここで、$\frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2\sigma^2}\left\{x - (\mu + \sigma^2 t)\right\}^2\right] dx$ は平均 $\mu + \sigma^2 t$、分散 σ^2 の正規分布密度関数の積分に他ならないから、1 に等しい。それゆえ、

$$M_x(t) = \exp\left(\mu t + \frac{\sigma^2}{2} t^2\right) \tag{33}$$

である。

$M_x'(t) = (\mu + \sigma^2 t)\exp(\mu t + \frac{\sigma^2}{2} t^2)$ であるから、

$$M_x'(0) = \mu \tag{34}$$

となり、また $M_x''(t) = \sigma^2 \exp(\mu t + \frac{\sigma^2}{2} t^2) + (\mu + \sigma^2 t)^2 \exp(\mu t + \frac{\sigma^2}{2} t^2)$ から、$M''(0) = \sigma^2 + \mu^2$、すなわち、

$$M''(0) - M'(0)^2 = \sigma^2 \tag{35}$$

となり、正規分布の平均は μ、分散は σ^2 であることが示された。

8.2 確率変数の関数の積率母関数

x を確率変数、t を実数とするとき、e^{tx} の期待値を x の積率母関数ということは既に述べた。ここで、$\varphi(x)$ を x の連続な関数としたとき、$\varphi(x)$ の積率母関数は $e^{t\varphi(x)}$ の期待値と定義され、$M_{\varphi(x)}(t)$ で表される。

x が連続型で $-\infty < x < \infty$ とすれば、$f(x)$ を密度関数として、

$$M_{\varphi(x)}(t) = E(e^{t\varphi(x)}) = \int_{-\infty}^{\infty} e^{t\varphi(x)} f(x)dx \tag{36}$$

である。

$e^{t\varphi(x)}$ をマクローリン展開すれば、

$$e^{t\varphi(x)} = 1 + \frac{t\varphi(x)}{1!} + \frac{(t\varphi(x))^2}{2!} + \cdots \tag{37}$$

であるから、これを式 36 に代入して、

$$\begin{aligned} E(e^{t\varphi(x)}) &= \int_{-\infty}^{\infty} f(x)dx + \int_{-\infty}^{\infty} t\varphi(x)f(x)dx + \frac{1}{2}\int_{-\infty}^{\infty} \{t\varphi(x)\}^2 f(x)dx + \cdots \\ &= 1 + tE[\varphi(x)] + \frac{t^2}{2}E[\{\varphi(x)\}^2] + \cdots \end{aligned} \tag{38}$$

であるから、

$$M^k(0) = E[\{\varphi(x)\}^k] \qquad k = 1, 2, \cdots \tag{39}$$

により、$\varphi(x)$ の原点回りの k 次の積率が計算できる。

9 中心極限定理

$x_1, \cdots, x_n$ を密度関数 $f(x)$ からの大きさ n の無作為標本とし、x の平均を μ、分散を σ^2 とする。またのこの標本の標本平均を $\bar{x}$、標本平均を標準化した変数を z_* とする。すなわち、

$$z_* = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = z\sqrt{n} \tag{40}$$

このとき、z_* の積率母関数は、

$$\begin{aligned} M_z(t) &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{tz_*} f(x_1) \cdots f(x_n) dx_1 \cdots dx_n \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\left(\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} t\right) f(x_1) \cdots f(x_n) dx_1 \cdots dx_n \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\left(t \sum_{i=1}^{n} \frac{x_i - \mu}{\sigma/\sqrt{n}}\right) f(x_1) \cdots f(x_n) dx_1 \cdots dx_n \end{aligned}$$

$$= \left\{ \int_{-\infty}^{\infty} \exp\left(\frac{x-\mu}{\sigma/\sqrt{n}}t\right) f(x)dx \right\}^n = \left\{ M_z\left(\frac{t}{\sqrt{n}}\right) \right\}^n = \left\{ \int_{\infty}^{\infty} \exp\left(\frac{t}{\sqrt{n}}z\right) f(z)dz \right\}^n$$

ここで、

$$\exp\left(\frac{t}{\sqrt{n}}z\right) = 1 + \frac{tz}{\sqrt{n}} + \frac{1}{2}\left(\frac{tz}{\sqrt{n}}\right)^2 + \frac{1}{6}\left(\frac{tz}{\sqrt{n}}\right)^3 + \cdots \tag{41}$$

となるから、

$$\left\{ M_z\left(\frac{t}{\sqrt{n}}\right) \right\}^n = \left[\int_{-\infty}^{\infty} \left\{ 1 + \frac{tz}{\sqrt{n}} + \frac{1}{2}\left(\frac{tz}{\sqrt{n}}\right)^2 + \frac{1}{6}\left(\frac{tz}{\sqrt{n}}\right)^3 + \cdots \right\} f(z)dz \right]^n$$
$$= \left[\int_{-\infty}^{\infty} f(z)dz + \int_{\infty}^{\infty} \frac{t}{\sqrt{n}} z f(z)dz + \int_{-\infty}^{\infty} \frac{t^2}{2n} z^2 f(z)dz + \int_{-\infty}^{\infty} \frac{t^3}{6n^{3/2}} z^3 f(z)dz + \cdots \right]^n \tag{42}$$

となる。ここで、$\int_{-\infty}^{\infty} f(z)dz = 1$(確率密度関数の定義、全確率の合計は 1)、$\int_{-\infty}^{\infty} z f(z)dz = 0$($z$ の平均は 0)、$\int_{-\infty}^{\infty} z^2 f(z)dz = 1$($z$ の分散) より、式 42 は、

$$M_{z*}(t) = \left[1 + 0 + \frac{t^2}{2n} + (n^{-3/2}\text{の項}) + (n^{-2}\text{の項}) + \cdots \right]^n \tag{43}$$

かくして、両辺の $n \to \infty$ における極限を取れば[4]

$$\lim_{n\to\infty} M_{z_*}(t) = e^{\frac{t^2}{2}} \tag{44}$$

となり、標準正規分布の積率母関数と一致する。

10　二項分布と正規分布

a, b および p を一定とするとき、確率

$$P(a \le x \le b) = \sum_{a\le x\le b} \frac{n!}{x!(n-x)!} p^x q^{n-x}, \quad q = 1-p \tag{45}$$

は $n \to \infty$ のとき、

$$\int_{(a-np)/\sqrt{npq}}^{(b-np)/\sqrt{npq}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz \tag{46}$$

に収束する。

証明：

標準化された $z = (x-np)/\sqrt{npq}$ の積率母関数は、

[4]十分大きな n に対し、$\frac{1}{n} < \frac{1}{n} + h(n) < \frac{p}{n}$ なら、$(1+\frac{1}{n})^n < (1+\frac{1}{n}+h(n))^n < (1+\frac{p}{n})^n$ より、$n \to \infty$ をとって、$e < \lim_{n\to\infty}(1+\frac{1}{n}+h(n))^n < e^p$。さらに $n \to \infty$ のとき、p がいくらでも 1 に近づけるなら、$\lim_{n\to\infty}(1+\frac{1}{n}+h(n))^n = e$。

$$M_z(t) = \sum_{x=0}^{n} \frac{n!}{x!(n-x)!} p^x q^{n-x} e^{t(x-np)/\sqrt{npq}} = e^{-\frac{np}{\sqrt{npq}}t} \sum_{x=0}^{n} \frac{n!}{x!(n-x)!} (pe^{\frac{t}{\sqrt{npq}}})^x q^{n-x}$$
$$= e^{-\frac{np}{\sqrt{npq}}t} (q + pe^{\frac{t}{\sqrt{npq}}})^n \tag{47}$$

従って、

$$\log M_z(t) = -\frac{np}{\sqrt{npq}}t + n\log(q + pe^{\frac{t}{\sqrt{npq}}}) = -\frac{np}{\sqrt{npq}}t + n\log(1+u) \tag{48}$$

但し $u = p(-1 + e^{t/\sqrt{npq}})$ となる。$\log(1+u)$ を展開すれば、

$$\log(1+u) = u - \frac{u^2}{2} + \frac{u^3}{3} - \cdots \tag{49}$$

となる。また u は $e^{t/\sqrt{npq}}$ の展開により、

$$u = p(-1 + e^{\frac{t}{\sqrt{npq}}}) = p\left\{ \left(\frac{t}{\sqrt{npq}}\right) + \frac{1}{2!}\left(\frac{t}{\sqrt{npq}}\right)^2 + \frac{1}{3!}\left(\frac{t}{\sqrt{npq}}\right)^3 + \cdots \right\} \tag{50}$$

となる。式 50 を式 49 に代入し、これをさらに式 48 に代入して整理すれば、

$$\log M_z(t) = \frac{t^2}{2} + \sum_{k=3}^{\infty} c_k \left(\frac{t}{\sqrt{npq}}\right)^k \tag{51}$$

となる。但し右辺の第 2 項は $\frac{t}{\sqrt{npq}}$ の 3 次以上の項よりなる多項式で、$c_3, c_4, \cdots$ は n を含まない係数である。そこで両辺の極限を取れば $\lim_{n\to\infty} \log M_x(t) = \frac{t^2}{2}$、すなわち

$$\lim_{n\to\infty} M_z(t) = e^{\frac{t^2}{2}} \tag{52}$$

となる。これは正規分布 $N(0,1)$ の積率母関数に他ならない。よって、$z = (x-np)/\sqrt{npq}$ は $N(0,1)$ に分布収束する。このことは、$P(a \le x \le b) = P((a-np)/\sqrt{npq} \le z \le (b-np)/\sqrt{npq}$ が式 46 に収束することを意味する。

□

参考文献

[1] 冨田勝 (監修) 斎藤輪太郎 (著) 「バイオインフォマティクスの基礎〜ゲノム解析プログラミングを中心に〜」サイエンス社 (2005/07)

[2] Hall, P(1927) "The Distribution of Means for Samples of Size N Drawn from a Population in which the Variate Takes Values Between 0 and 1, All Such Values Being Equally Probable". Biometrika, Vol. 19, No. 3/4., pp. 240-245.

[3] Irwin JO (1927) "On the Frequency Distribution of the Means of Samples from a Population Having any Law of Frequency with Finite Moments, with Special Reference to Pearson's Type II". Biometrika, Vol. 19, No. 3/4., pp. 225-239.