正規分布

確率変数が連続的な値を取るものを、連続分布と言う。その中で最も重要なのが、**正規分布** (normal distribution) である。正規分布とは、ずいぶん堅苦しい言い方だが、normal とは、標準的なとか平均的などの意味である。正規分布とは、いろんな所にあまねく存在する分布である。ある一つの物を何度も測定すると、その測定値は、ほとんど同じであるが、皆少しずつ異なっている。この測定誤差の分布が正規分布である。そんなわけで、この分布は、別名、**誤差分布**とも呼ばれる。

正規分布は、最初、ド・モアブルにより二項分布の極限として導入された。そして、この分布を詳しく調べたのが、ドイツの数学者カール・フリードリッヒ・ガウス (1777〜1855) であったので、ガウス分布とも言う。確率変数 x の取り得る範囲は、マイナス無限大からプラス無限大までの全領域である。平均が μ、標準偏差が σ で与えられる正規分布の確率密度関数 $f(x)$ は、

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

で与えられる。これを $N(\mu, \sigma^2)$ と表す。

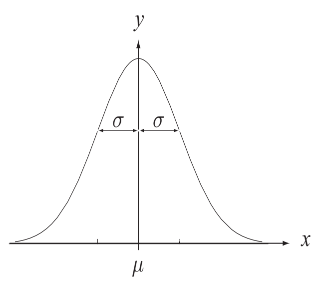

図 1: 正規分布

正規分布の特徴は、確率密度関数が、平均値 $x = \mu$ に対して対象なことである。

正規分布のような連続分布において、確率変数 x が、ある特定の値を取る確率というのは意味を持たない。なぜならばその確率は常にゼロだからである。連続分布において意味があるのは、確率変数 x がある値の範囲を取る場合である。すべての確率についての和は、

$$\frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{+\infty} e^{(x-\mu)^2/2\sigma}\ dx = 1$$

と表される。確率変数 x が、a から b までの範囲を取る確率 $P(a < x < b)$ は、確率密度関数 $f(x)$ を a から b まで積分することにより得られる。

$$\mathrm{Prob}(a < x < b) = \int_a^b f(x)\ dx \tag{1}$$

この値を解析的に求めることはできないので、数値積分を行うしかない。しかし、すべての正規分布は、変数変換により、平均が 0、分散が 1 の正規分布に変換することができる。これを**標準正規分布**と言う。標準正規分布についての積分表が与えられていれば、すべての正規分布についての確率が分ることになる。

標準正規分布

平均が μ、標準偏差が σ^2 の正規分布 $N(\mu,\sigma^2)$ を、標準正規分布 $N(0,1)$ に変換するには、正規分布 $N(\mu,\sigma^2)$ の確率変数 x に線形変換 $z=(x-\mu)/\sigma$ を行う。そのためには、正規分布関数 $f(x)=\dfrac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ に、$\delta(z-\frac{x-\mu}{\sigma})$ を掛けて、変数 x について積分すればよい。ここで、デルタ関数の性質、$\delta(z-\frac{x-\mu}{\sigma})=\sigma\delta(\sigma z-x+\mu)$ を用いると、標準正規分布関数 $f_N(z)$ は、

$$\begin{aligned}
f_N(z) &= \int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\delta\Big(z-\frac{x-\mu}{\sigma}\Big)\,dx \\
&= \int_{-\infty}^{+\infty}\frac{\sigma}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\delta(\sigma z-x+\mu)\,dx \\
&= \frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}
\end{aligned}$$

すなわち、

$$f_N(z)=\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}$$

を得る。

たたみこみ積分

確率変数 x_1、x_2 の確率密度関数は、各々、 $f_1(x_1)$、$f_2(x_2)$ で与えられるとしよう。確率変数が、$x=x_1+x_2$ の確率密度関数を $f(x)$ を表すと、これは、デルタ関数をもちいて、

$$f(x)=\int\delta(x_1+x_2-x)f_1(x_1)f_2(x_2)\;dx_1dx_2 \tag{2}$$

と表される。これを x_1 について積分したものは、単に、$f_1(x_1)f_2(x_2)$ において $x_1=x-x_2$ を代入すればよいので、

$$f(x)=\int f_1(x-x_2)f_2(x_2)\;dx_2$$

となる。これが、関数 $f_1(x_1)$、$f_2(x_2)$ の**たたみこみ積分**である。

正規分布の線形結合

x が正規分布 $N(\mu,\sigma^2)$ に従っているとき、その線形変換 $y=ax+b$ は正規分布 $N(a\mu+b,a^2\sigma^2)$ に従う。

証明

確率変数 x は、平均 μ、分散 σ^2 の正規分布 $f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ に従うとする。確率変数が $y=ax+b$ である分布を $f_L(y)$ と書くことにすると、

$$\begin{aligned}
f_L(y) &= \int_{-\infty}^{+\infty}\delta(ax+b-y)f(x)\,dx \\
&= \frac{1}{a}f\Big(\frac{y-b}{a}\Big) \\
&= \frac{1}{\sqrt{2\pi a^2\sigma^2}}\exp\left[-\frac{\big(y-(a\mu+b)\big)^2}{2a^2\sigma^2}\right]
\end{aligned}$$

これは、確率変数 y が平均 $a\mu+b$、分散 $a^2\sigma^2$ の正規分布 $N(a\mu+b, a^2\sigma^2)$ に従うことを表している。

正規分布の一次結合

確率変数 x_1、x_2 が互いに独立で、それぞれ正規分布 $N(\mu_1,\sigma_1^2)$ $N(\mu_2,\sigma_2^2)$ に従っているとき、確率変数 $x=x_1+x_2$ は正規分布 $N(\mu_1+\mu_2,\sigma_1^2+\sigma_2^2)$ に従う。

それでは、２つの正規分布関数をたたみこんだ関数 $f(x)$ を計算しよう。

$$f(x) \quad = \quad \frac{1}{\sqrt{2\pi\sigma_1^2}}\frac{1}{\sqrt{2\pi\sigma_2^2}}\int \exp\left[-\frac{(x-x_2-\mu_1)^2}{2\sigma_1^2}-\frac{(x_2-\mu_2)^2}{2\sigma_2^2}\right]dx_2$$

指数関数の肩の部分を整理すると、

$$\begin{aligned}&-\frac{(x-x_2-\mu_1)^2}{2\sigma_1^2}-\frac{(x_2-\mu_2)^2}{2\sigma_2^2}\\&=-\frac{1}{2\sigma_1^2\sigma_2^2}\left\{(\sigma_1^2+\sigma_2^2)\left[x_2-\frac{\sigma_2^2(x-\mu_1)+\sigma_1^2\mu_2}{\sigma_1^2+\sigma_2^2}\right]^2+\frac{\sigma_1^2\sigma_2^2}{(\sigma_1^2+\sigma_2^2)}(x-\mu_1-\mu_2)^2\right\}\end{aligned}$$

となる。ここで、x_2 についての積分は、

$$\int_{-\infty}^{+\infty}\exp\left\{-\frac{(\sigma_1^2+\sigma_2^2)}{2\sigma_1^2\sigma_2^2}\left[x_2-\frac{\sigma_2^2(x-\mu_1)+\sigma_1^2\mu_2}{\sigma_1^2+\sigma_2^2}\right]^2\right\}dx_2=\sqrt{\frac{2\pi\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}}$$

となるので、

$$f(x)=\frac{1}{\sqrt{2\pi(\sigma_1^2+\sigma_2^2)}}\exp\left[-\frac{(x-\mu_1-\mu_2)^2}{2(\sigma_1^2+\sigma_2^2)}\right]$$

となる。これは、平均が $\mu_1+\mu_2$、分散が $\sigma_1^2+\sigma_2^2$ の正規分布を表す。

上記の結果と、先に示した正規分布の線形変換の結果を合わせると、一般的に次のことが言える。

確率変数 $x_1,\ x_2,\ x_3,\ \cdots,\ x_n$ が互いに独立で、それらの分布が正規分布 $N(\mu_1,\sigma_1^2)$, $N(\mu_2,\sigma_2^2)$, $N(\mu_3,\sigma_3^2),\cdots,\ N(\mu_n,\sigma_n^2)$ であるとき、

$$y=a_0+a_1x_1+a_2x_2+a_3x_3+\cdots+a_nx_n$$

は、

$$\text{平均}\quad \mu_y=a_0+a_1\mu_1+a_2\mu_2+a_3\mu_3+\cdots+a_n\mu_n, \tag{3}$$

$$\text{分散}\quad \sigma_y^2=a_1^2\sigma_1^2+a_2^2\sigma_2^2+a_3^2\sigma_3^2+\cdots+a_n^2\sigma_n^2 \tag{4}$$

の正規分布 $N(\mu_y,\sigma_y^2)$ に従う。