

**補 足 7**

**チェビシェフの不等式**

本文中ではチェビシェフの不等式をどんな分布であれ平均値±2SDの範囲に少なくとも3/4のデータが含まれると説明したが、もう少し一般化して書くと、どんな分布であれ、平均値±k×SDの範囲を超える領域の確率が1/k<sup>2</sup>以下になる、ということである。k=2のときこの確率の上限は1/4であるから残った3/4が「平均値±2SDの範囲の確率」の下限になるのだ。

このことを平均μ、分散σ<sup>2</sup>(>0)を持つ連続変数xの確率密度関数f(x)について式で表すと、

$$\int_{\mu-k\sigma}^{\mu+k\sigma} f(x) dx \geq 1 - \frac{1}{k^2}$$

ということである。なぜそう言えるのかというと、まずこの分散σ<sup>2</sup>は、

$$\begin{aligned} \sigma^2 &= E((x - \mu)^2) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx \\ &= \int_{-\infty}^{\mu-k\sigma} (x - \mu)^2 f(x) dx + \int_{\mu-k\sigma}^{\mu+k\sigma} (x - \mu)^2 f(x) dx + \int_{\mu+k\sigma}^{\infty} (x - \mu)^2 f(x) dx \\ &\geq \int_{-\infty}^{\mu-k\sigma} (x - \mu)^2 f(x) dx + \int_{\mu+k\sigma}^{\infty} (x - \mu)^2 f(x) dx \end{aligned}$$

(3つの領域に積分を分けて、真ん中を取り除いただけのことである。)

ここで、第一項についても第二項についてもμより±kσ以上離れた領域のことを考えているため|x - μ| ≥ kσ > 0であり、ゆえに(x - μ)<sup>2</sup> ≥ k<sup>2</sup>σ<sup>2</sup> > 0である。

よって、

$$\begin{aligned} \sigma^2 &\geq \int_{-\infty}^{\mu-k\sigma} (x - \mu)^2 f(x) dx + \int_{\mu+k\sigma}^{\infty} (x - \mu)^2 f(x) dx \\ &\geq \int_{-\infty}^{\mu-k\sigma} k^2 \sigma^2 f(x) dx + \int_{\mu+k\sigma}^{\infty} k^2 \sigma^2 f(x) dx \\ &= k^2 \sigma^2 \left( \int_{-\infty}^{\mu-k\sigma} f(x) dx + \int_{\mu+k\sigma}^{\infty} f(x) dx \right) \\ &= k^2 \sigma^2 \left( 1 - \int_{\mu-k\sigma}^{\mu+k\sigma} f(x) dx \right) \\ &\Leftrightarrow \frac{1}{k^2} \geq 1 - \int_{\mu-k\sigma}^{\mu+k\sigma} f(x) dx \end{aligned}$$

$$\Leftrightarrow \int_{\mu-k\sigma}^{\mu+k\sigma} f(x) dx \geq 1 - \frac{1}{k^2}$$

となり、このチェビシェフの不等式が示された。