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 $1/k^2$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$\sigma^2(>0)$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、

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

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^2は、

$$\sigma^2 = E((x-\mu)^2) = \int_{-\infty}^{\infty}(x-\mu)^2 f(x)dx$$

$$= \int_{-\infty}^{\mu-k\sigma}(x-\mu)^2 f(x)dx + \int_{\mu-k\sigma}^{\mu+k\sigma}(x-\mu)^2 f(x)dx + \int_{\mu+k\sigma}^{\infty}(x-\mu)^2 f(x)dx$$

$$\geqq \int_{-\infty}^{\mu-k\sigma}(x-\mu)^2 f(x)dx + \int_{\mu+k\sigma}^{\infty}(x-\mu)^2 f(x)dx$$

（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$|x-\mu| \geqq k\sigma > 0$であり、ゆえに$(x-\mu)^2 \geqq k^2\sigma^2 > 0$である。

よって、

$$\sigma^2 \geqq \int_{-\infty}^{\mu-k\sigma}(x-\mu)^2 f(x)dx + \int_{\mu+k\sigma}^{\infty}(x-\mu)^2 f(x)dx$$

$$\geqq \int_{-\infty}^{\mu-k\sigma} k^2\sigma^2 f(x)dx + \int_{\mu+k\sigma}^{\infty} k^2\sigma^2 f(x)dx$$

$$= k^2\sigma^2\left(\int_{-\infty}^{\mu-k\sigma} f(x)dx + \int_{\mu+k\sigma}^{\infty} f(x)dx\right)$$

$$= k^2\sigma^2\left(1 - \int_{\mu-k\sigma}^{\mu+k\sigma} f(x)dx\right)$$

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

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

となり、このチェビシェフの不等式が示された。