プログラムで理解する視覚でわかる統計確率論　第2回

関連データ

中心極限定理と正規分布

浪平 博人

品質管理や金融工学，数理統計などで扱う分布は，そのほとんどが正規分布を仮定しています．しかし，研究開発の現場できれいな分布を扱うことはあまり経験をしません．それでも，得られた分布をあたかも正規分布のように扱って処理を進めることが多くあります．これでよいのでしょうか．結論としては，これでよいのです．その理由が確率論・統計学における極限定理のうち一つ，中心極限定理です．（筆者）

正規分布と中心極限定理

正規分布は，実験や研究，検証などでの判断材料としてよく使われています．

中心極限定理とは，任意の一つの分布からデータをn個とり，それらを平均した値zを考えると，zの分布はデータの数nが大きくなると正規分布に近づくという定理ですです(**図1**)．今回は，この内容を詳しく調べてみましょう．

畳み込み分布とは

中心極限定理を理解するには，まず，n個の変数を足すことの意味から考えます．二つの分布F_1，F_2があり，それぞれから確率変数x_1，x_2が発生するとき，それらの和の分布$z = x_1 + x_2$を**図2**に示します．

分布F_1の変数x_1は1から3の，分布F_2のx_2は1から4のどれかなので，zの範囲は2から7までです．そして，$z = 2$は$[x_1 = 1, x_2 = 1]$のとき以外はありません．その確率は，$0.3 \times 0.2 = 0.06$です．$z = 3$が生じるのは$[x_1 = 1, x_2 = 2]$，$[x_1 = 2, x_2 = 1]$の二つの組み合わせで，その確率は合計して0.22となります．

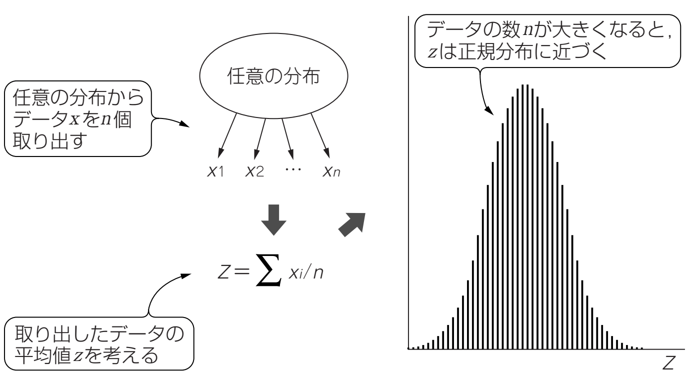

図1　中心極限定理の概要

指定したzのx_1とx_2の組み合わせと，その発生確率を**図3(a)**の表に示します．これを視覚化したのが**図3(b)**の分布Zです．分布ZをF_1とF_2の畳み込み(Convolution)分布といいます．$Z = x_1 + x_2 + \cdots + x_n$の分布を考えるには，起こり得るすべての$z$を生じさせる$[x_1, x_2, \cdots, x_n]$の組み合わせを数え上げ，その確率を計算して求めます．

変数の数nが3以上になれば，組み合わせを数え上げるのは大変面倒です．nが3以上のときは，数え方を工夫します．x_1，x_2，x_3の三つの組み合わせを考える場合，まず$x_1 + x_2$の分布を求めます．その結果と残ったx_3との組み合わせを計算するという手順でzを求めます．$n = 4$の場合は，$n = 3$として求めた3次の分布を使って，4次の畳み込み分布を求めます．下の次数の畳み込み分布を順次求めながら，目的のn次まで進んでいくのです．

中心極限定理の意味

さて，中心極限定理に話を戻します．

それでは，n個の確率変数の和がどのように分布するかを実証的に検証してみましょう．最初に任意の分布Fを設定し，これから取り出したn個の変数の和の分布を計算します．すなわち，n次の畳み込み分布を計算します．これを視覚的に表示したのが**図4**です．

原分布は任意の分布ですが，畳み込みが4次(四つの変数の和の分布)になると正規分布に近くなります．6次ではほぼ正規分布です．わずか6個の変数の和の分布がこのように正規に近づくことが，**図4**のように視覚化すれば直感的にわかります．従来の数式の理論展開から推測するのは難しいことです．