正規分布

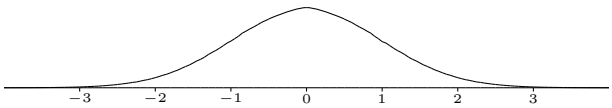
年収や身長の分布には上限や下限が存在している。しかし、理論的には確率変数 X の範囲は実数全体で考えるとよい場合が多い。年収や身長にしても、平均値を 0 にしておけば、平均より大きい値が正の実数、平均より小さい値が負の実数に対応させられる。もっとも、実際の値が 0 から際限なく離れることはないのだが、0 から極端に離れた値 X の確率 $P(X)$ は 0 と考えれば済む話なので、連続的な確率変数に対する上限・下限を決めない方が都合がよい。

そこで実数全体を範囲とする確率変数 X の確率密度関数を

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

によって定義する。π はおなじみの円周率である。e は**自然対数の底**と呼ばれる定数で、およそ 2.71828 程度の無理数である。σ は標準偏差を表し、これまでは s を用いてきたが、ここからは習慣にしたがい σ を使うことにする。

突然、恐ろしげな式を与えたが、e は定数であることから、$f(x)$ は基本的に指数関数 $g(x) = ka^{-x^2}$ と同等である。指数関数ならば右上がりのグラフを想像すると思うし、知識があれば、指数が負の値であることから右下がりのグラフであると思えるだろう。さらに指数は $^{-x^2}$ なので、$g(x)$ のグラフは 0 を中心に左右対称であるはずだ。つまり、グラフの形状としては、左右対称の右上がりかつ右下がりの曲線であろうと思える。実際、$\sigma = 1$, $m = 0$ のときの $f(x)$ のグラフは



のような形状である。この場合、確率密度関数は $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ であり、この分布は**標準正規分布**と呼ばれる。

* * *

$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}$ は、平均 m、標準偏差 σ の**正規分布**といい、$N(m,\ \sigma^2)$ で表して、確率変数 X は正規分布 $N(m,\ \sigma^2)$ にしたがうという。一般に

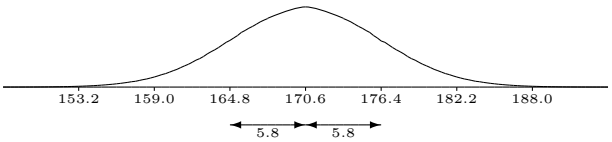
$$\begin{aligned} P(m-\sigma \leqq X \leqq m+\sigma) &\approx 0.6827 \\ P(m-2\sigma \leqq X \leqq m+2\sigma) &\approx 0.9545 \\ P(m-3\sigma \leqq X \leqq m+3\sigma) &\approx 0.9973 \end{aligned}$$

であることが分かっているが、正規分布曲線 $f(x)$ が区間 $[a,\ b]$ で囲む面積が $a \leqq X \leqq b$ の確率を表すので、平均から標準偏差 σ 分だけ離れた範囲の確率は 68% ほどであることが分かる。これまでに何かと標準偏差を求めてきたが、平均値と標準偏差を組み合わせることで、対象とする確率変数の値が平均的なグループに対

して、どの程度の位置にあるかが分かるのである。たとえば、平均から標準偏差 σ の 2 倍以内の範囲は全体の 95% 以上を占めるので、そこから外れた確率変数の値は 5% 未満である。そして、半々が上位と下位に外れている。よって、平均より 2σ 以上離れた値というのは、だいぶ稀（まれ）な存在と言えるだろう。▌

正規分布を仮定する

前節で 17 歳男子の身長の分布を示し、平均値はおよそ 170.6cm であった。標準偏差はだいたい 5.8 ぐらいのようである。この分布を正規分布と仮定すると



のようなグラフが、17 歳男子の身長の分布を表している。この場合、（描いてないが）縦軸は割合を表すので、当然、前節のグラフとは縦・横方向の縮尺が異なっている。標準正規分布はグラフの囲む面積が 1 であるから、実数を積み上げたヒストグラムと違って当然である。しかし、ヒストグラムの縦・横方向の縮尺を標準正規分布に合うようにすれば、グラフは大体重なるだろう。

縮尺を変えることでグラフが重なるならば、正規分布と思（おぼ）しき分布は $N(0,\ 1^2)$ に変換できるはずだ。

正規分布の標準化

身長の分布は正規分布に近いようだが、ここでは正しく正規分布 $N(m,\ \sigma^2)$ を標準正規分布 $N(0,\ 1^2)$ に変換することを考えよう。その際、確率変数の平均と標準偏差の性質に目を向ける必要がある。たとえば

X	x_1	x_2	$\cdots$	$\cdots$	x_n	計
P	p_1	p_2	$\cdots$	$\cdots$	p_n	1
(i)	(i_1)	(i_2)	$\cdots$	$\cdots$	(i_n)	(N)

のような分布の平均はどのように求めるべきだろうか。( ) 内は実際の個数を表している。一般に、平均 $E(X)$ は（総和）÷（総数）で求めるので、

$$E(X) = \frac{x_1 i_1 + x_2 i_2 + \cdots + x_n i_n}{N}$$

であるが、$\dfrac{i_1}{N}$ は X が値 x_1 をとる確率 p_1 であるから、

$$\begin{aligned} E(X) &= x_1\frac{i_1}{N} + x_2\frac{i_2}{N} + \cdots + x_n\frac{i_n}{N} \\ &= x_1p_1 + x_2p_2 + \cdots + x_np_n \end{aligned} \tag{1}$$

となる。これを確率変数 X の平均というが、実際は**期待値**を求めているのである。

また、確率変数 X の分散 $V(X)$ は「偏差平方和、の平均」なので、確率変数 X の平均を $E(X) = m$ とすると

$$\begin{aligned} V(X) &= \frac{(x_1-m)^2i_1 + (x_2-m)^2i_2 + \cdots + (x_n-m)^2i_n}{N} \\ &= (x_1-m)^2p_1 + (x_2-m)^2p_2 + \cdots + (x_n-m)^2p_n \end{aligned} \tag{2}$$

となる。この正の平方根が X の標準偏差になるので $\sigma(X) = \sqrt{V(X)}$ と書いておこう。

さて、これら平均と標準偏差の性質であるが、$Y = aX + b$ の変換を考える。このときの分布は

Y	ax_1+b	ax_2+b	$\cdots$	$\cdots$	ax_n+b	計
P	p_1	p_2	$\cdots$	$\cdots$	p_n	1

であるから、(1) を参考に

$$\begin{aligned} E(Y) &= (ax_1+b)p_1 + (ax_2+b)p_2 + \cdots + (ax_n+b)p_n \\ &= a(x_1p_1 + x_2p_2 + \cdots + x_np_n) + b(p_1 + p_2 + \cdots + p_n) \\ &= aE(X) + b \end{aligned}$$

である。$p_1 + p_2 + \cdots + p_n = 1$ であることに注意しよう。つまり確率変数 Y の平均は $am + b$ に変わる。

また、$V(Y)$ は (2) を参考に

$$\begin{aligned} &V(Y) \\ &= \{(ax_1+b) - (am+b)\}^2p_1 + \{(ax_2+b) - (am+b)\}^2p_2 + \cdots + \{(ax_n+b) - (am+b)\}^2p_n \\ &= a^2(x_1-m)^2p_1 + a^2(x_2-m)^2p_2 + \cdots + a^2(x_n-m)^2p_n \\ &= a^2V(X) \end{aligned}$$

である。

確率変数 X の平均 $E(X)$、分散 $V(X)$ に対して、$Y = aX + b$ で変換すると

$$E(Y) = E(aX + b) = aE(X) + b$$
$$V(Y) = V(aX + b) = a^2V(X)$$

これでようやく、正規分布 $N(m,\ \sigma^2)$ を標準正規分布 $N(0,\ 1^2)$ に変換する準備が整った。確率変数 X が正規分布 $N(m,\ \sigma^2)$ にしたがうとき、$Z = \dfrac{X - m}{\sigma}$ とおく。このとき、上記性質により

$$E(Z) = E\left(\frac{X - m}{\sigma}\right) = \frac{1}{\sigma}\{E(X) - m\} = \frac{1}{\sigma}(m - m) = 0$$
$$V(X) = V\left(\frac{X - m}{\sigma}\right) = \frac{1}{\sigma^2}V(X) = \frac{1}{\sigma^2} \cdot \sigma^2 = 1$$

である。すなわち、確率変数 Z は平均 0、分散 1（標準偏差は $\sqrt{1} = 1$）である。よって、Z の確率分布は標準正規分布 $N(0,\ 1^2)$ にしたがうことが分かった。

この場合、確率変数 X が中心 0 から一定の標準偏差だけ離れた確率 P は

$$P(-1 \leqq X \leqq 1) \approx 0.6827$$
$$P(-2 \leqq X \leqq 2) \approx 0.9545$$
$$P(-3 \leqq X \leqq 3) \approx 0.9973$$

ということである。

二項分布で正規分布を近似する

正規分布について調べると、二項分布とは少し違う面が見えてくる。実際、データ J で見た $B\left(10,\ \dfrac{1}{6}\right)$ のヒストグラムは、明らかに左右非対称のグラフである。ところが、n が大きくなると少しずつではあるが、山のピークが右へ移り左右の裾（すそ）がなだらかに対称性を帯びてくる。この様子は Microsoft Excel で見ることができる。

◇	A	B	C	D	E	F
1	n\r	0	5	10	15	**5** ずつ増→
2	10	（※ **B2**）	右へコピーする→	→	→	→
3	20	↓下へコピーする	↓→			
4	30	↓				
5	40	↓				
6	50	↓				
7	↓以下 **10** 増	↓				
8	↓	↓				

※ セルの式

（**B2**）　**=COMBIN($A2,B$1)*(1/6)^B$1*(5/6)^($A2-B$1)**

B2 セルには ${}_n\mathrm{C}_r(\frac{1}{6})^r(\frac{5}{6})^{n-r}$ に相当する式を入力する。この式はそのまま下方向と右方向にコピーするのだが、n を大きするということは、表のサイズも大きくすることなので、必要なだけコピーをしてもらいたい。その際、余分なセルにはエラー "#NUM!" が表示されるが無視してよい。

このとき、数値が示された表を眺めていても分布の様子は分からないだろうから、2 行目から順に B 列から右のセルを 1 行分選択してグラフにしてみよう。3 行目、4 行目、... と 1 行ずつグラフにするのは手間だが、見比べるとグラフが変化するのが分かるだろう。

もし、グラフをたくさん並べるのが煩雑に思えたら、2 行目だけをずっと右方向へコピーして、その範囲を一旦グラフにしておこう。この場合、3 行目以下は不要になる。その上で、A2 セルの数値を 10, 20, 30, 40, ... と変化させればその都度グラフも変化するので、アニメーションのように変化を見ることになる。1 行目の r のとり方を変えるなど、いろいろ試すとよいだろう。