III-2-4. 正規分布

III-2-4-1. 二項分布から正規分布を導く

二項分布は、比率データーに関する確率分布ですから不連続ですが、nが大きくなったら、次第になめらかな曲線に近くなってくるでしょう。nを無限個にすれば、完全に滑らかになるはずです。nを無限大にするときに、2 つの方向が考えられます。一つは、pを一定にして、nを無限大に大きくする方向です。これが正規分布です。すでに、二項分布のところで、pを一定にしながらnを大きくするということを試してみました。これら例からわかるように、次第に、左右相称になり、分散が一定の値に近づきます。二項分布の正規分布への拡張の目的の一つは、身長や体重のような連続した値をとるデーターを統計的に扱うための拡張です。いくつかのグループのデータを比較して、その差の有意性を判断することを可能にするためです。拡張のもう一つの方向は、平均値npを一定にして、p を小さくしてnを大きくしていく方向です。その結果、分布は大きく偏っていきます。これがポアソン分布です。pを小さくすることからわかるように、ポアソン分布は極めてまれに起こる現象についての分析に使います。ポアソン分布は水産の世界では、たとえば、プランクトンの計数などのときに、きわめて稀な種類が、計数版の方形枠の中に現れたり現れなかったりする場合などに使います。ポアソン分布 III-2-3 で説明したので、ここでは正規分布について考えます。

二項分布の正規分布への拡張

二項分布と正規分布の一番大きな違いは、二項分布の確率が不連続関数であるのに対して、正規分布は連続関数だということです。二項分布のnを無限大にしたものだという説明もあります。確かにそうなのですが、二項分布はnの増加に伴って右側に広がっていきますが、正規分布は平均値を中心としたシンメトリックな分布という違いもあります。無限大に拡大したという説明だけでは、二項分布と正規分布の関係が具体的にわからないし、どのようにして二項分布から正規分布を導くのかわかりません。筆者は、二項分布から正規分布を導く作業の通り、正規分布は二項分布のテイラー展開だと説明しています。テイラー展開については、III-3-1 で詳しく説明していますから、そちらを読んでいただければよいのですが、最も簡単に説明すると、複雑な関数の全体を、ある一点での高次導関数の和で表す方法ということです。これを公式に表すと以下の公式になります。

$$f(x) = \lim_{n\to\infty} \sum_{i=1}^{n} \frac{f^{(i)}(a)}{i!}(x-a)^i$$

この式の右辺は無限級数なので、左辺と右辺を等号で結べますが、実際には有限個の項数に近似することになるので、以下の式になります。

$$f(x) \fallingdotseq \sum_{i=1}^{n} \frac{f^{(i)}(a)}{i!}(x-a)^i$$

これらの式で、$f^{(i)}$は i 次導関数数を表します。.

二項係数の確率は以下の式になることは III-2-2 で説明しました。

$$W(k) =_n C_k p^k q^{(1-k)}$$

コンビネーション記号を書き換えて分数で表すと

$$W(k) = \frac{n!}{k!(n-k)!} p^k q^{(n-k)} \qquad \text{p+q=1}$$

III-2-2 では、入れ子型の合成関数として微分する方法と、対数変換したものを微分する方法の２つを紹介しました。入れ子型の合成関数の微分法を知っていれば、１階微分は簡単なのですが、合成関数にする過程で１回対数化しますし、2 階微分の時に関数の積の形になって、どちらにしても面倒だから、対数にしたものを微分して、後から元に戻すことにします。

両辺の対数をとります。

$$\log W(x) = \log(n!) - \log(x!) - \log(n-x)! + k\log(p) + (n-x)\log(q)$$

ここで、　$\int_1^x log\, t\, dt$ ≒logx!をつかって

$$\log W(x) \fallingdotseq \log(n!) - \int_1^x \log t dt - \int_1^{n-x} \log t dt + x\log p + (n-x)\log q$$

これを微分すると

$$\{\log W(x\}' \fallingdotseq -[\log t]_1^x + [\log t]_1^{n-x} + \log p - \log q$$

p+q=1、log1=0 ですから

$$\{\log W(x\}' \fallingdotseq -[\log t]_1^x + [\log t]_1^{n-x} + \log p - \log 1 - p$$

$$\fallingdotseq -\log x + \log(n-x) + \log p - \log(1-p)$$

$$\fallingdotseq \log\frac{(n-x)p}{x(1-p)}$$

$$\{\log W(x\}'' \fallingdotseq \{-\log x + \log(n-x) + \log p - \log(1-p)\}'$$

$$\fallingdotseq \{-\log x\}' + \{\log(n-x)\}'$$

$$\fallingdotseq -\frac{1}{x} - \frac{1}{n-x}$$

これで、1 階微分と 2 階微分ができて、1 次導関数と 2 次導関数が計算できます。どこか一点の微分値を求めて、その近傍で Taylor 展開すれば、3 階微分以降は値が十分小さく無視できます。できれば全体を代表する点で Taylor 展開したい。考えられるのは分布のピークでの展開です。

$$\{\log W(x\}' = \log\frac{(n-x)p}{x(1-p)} = 0$$

$$\frac{(n-x)p}{x(1-p)} = 1$$

$$(n-x)p = x(1-p)$$

$$np - xp = x - xp$$

$$x = np$$

２次導関数を計算する前に、この点がどんな点なのかを考えておきます。
まず確率の総和は１だから、$p+q=1$です。また、起こった回数をx起こらない回数を z とすると、総試行回数nは$n=x+z$です。

$$\frac{(n-x)p}{x(1-p)}=1$$

$$\frac{(n-(n-z))(1-q)}{(n-z)(1-(1-q))}=1$$

$$\frac{z(1-q)}{(n-z)q}=1$$

右辺が１だから、左辺の分母・分子を入れ替えて

$$\frac{(n-z)q}{z(1-q)}=1$$

これは$\frac{(n-x)p}{x(1-p)}=1$と同じだから

$$z=nq$$

$$\mu=nq$$

となります。右から見ても左から見ても、式の形は同じということですね。また、もともと２項分布なのですから、p を一定にして n を大きくしていけば、左右対称に近づきます。nを無限大にすれば、その分布の形も左右平等です。つまり、期待値＝最頻値＝中央値ということです。（これは２項分布の性質でもありますね。）
２階微分は

$$\{\log W(x)\}'' \fallingdotseq -\frac{1}{x}-\frac{1}{n-x}$$

$$\{\log W(np)\}'' \fallingdotseq -\frac{1}{np}-\frac{1}{n-np}$$

$$\fallingdotseq -\frac{1}{n}\left(\frac{1}{p}+\frac{1}{1-p}\right)$$

$$\fallingdotseq -\frac{1}{np(1-p)}$$

二項分布では$np(1-p)=\sigma^2$ですから

$$\{\log W(x)\}'' \fallingdotseq -\frac{1}{\sigma^2}$$

従って２次導関数までの Taylor 展開は次のようになります

$$\log \mathrm{W}(x)=\log(\mathrm{n}!)-\log(x!)-\log(\mathrm{n}-x)!+\mathrm{k}\log(\mathrm{p})+(\mathrm{n}-x)\log(\mathrm{q})$$

$$\fallingdotseq \log \mathrm{W}(\mu)+\frac{(\log(\mu))'}{1!}(x-\mu)+\frac{(\log(\mu))''}{2!}(x-\mu)^2$$

$(\log(\mu))'=0$ だから

$$\log \mathrm{W}(x) = \log W(\mu) + \frac{(\log(x))''}{2}(x-\mu)^2$$

$$=\log W(\mu) + \frac{-\frac{1}{\sigma^2}}{2}(x-\mu)^2$$

右辺を一つの対数にまとめます。

$$\log W(x) \fallingdotseq \log W(\mu) + \frac{-1}{2\sigma^2}(x-\mu)^2$$

$$= \log_e W(\mu) + \frac{-1}{2\sigma^2}(x-\mu)^2 \log_e e$$

$$\because \log_e e = 1$$

$$= \log_e W(\mu) + \log_e e^{\frac{-1}{2\sigma^2}(x-\mu)^2}$$

$$= \log_e W(\mu) + \log_e e^{\frac{-1}{2\sigma^2}(x-\mu)^2}$$

$$= \log_e W(\mu) e^{\frac{-1}{2\sigma^2}(x-\mu)^2}$$

となるので、対数の中だけを考えれば、

$$\mathrm{W(x)} \fallingdotseq \mathrm{W(\mu)} e^{\frac{-1}{2\sigma^2}(x-\mu)^2}$$

著者注

ちなみに、ここで e は自然対数の底として知られるもので、数学的にはネイピア数と言います。高校の数学でネイピア数とは何かしっかりとした説明を受けていない人が多いということを最近知りました。そこで、ネイピア数についての解説を（3-4-3.ネイピア数）に書いておきました。参考にしてください。しっかり理解すると、以下の説明がわかりやすくなります。なお記号の約束事として、特に断らない限り対数$\log A$と書いたときの$\log$　は$\log_e A$のことで、対数はネイピア数を底とする自然対数だと理解してください。なお、対数の微分$\frac{d \log x}{dx} = \frac{1}{x}$、指数の微分$\frac{de^x}{dx} = e^x$は知っているものとして話を進めます。これがわからない人は（III-3-2.ネイピア数）を読んでください。

元に戻って

$$\mathrm{W(x)} \fallingdotseq \mathrm{W(\mu)} e^{\frac{-1}{2\sigma^2}(x-\mu)^2}$$

数学の答えとしてはこれで良いのかもしれませんが、これではあまりよく意味がわからな

いし、正規分布として私たちが知っている式とも表現の仕方が違います。式に含まれているW(μ)はμが与えられば一定の値として定数になるはずですが、これがどのような値なのかは少なくとも知りたいところです。そこで、何らかの条件を与えて、W(μ)の値を求めることを考えます。すぐに気が付く条件は、この式は確率分布の式なのだからその面積の総和は1ということです。つまり-∞から∞まで積分すれば、その値は1になるということです。

ですから、W(μ) = Aとして A について以下の式を解けばよいことになります。

$$\int_{-\infty}^{\infty} W(\mu) e^{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

$$= \int_{-\infty}^{\infty} A\, e^{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

$$= \mathrm{A} \int_{-\infty}^{\infty} e^{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

指数のカッコの中の$\left(\frac{x-\mu}{\sigma}\right)$について考えます。この値は、期待値（母集団の平均値・中央値）と実際に得られたデーターを、標準偏差で割ったものですね。μを起点(0)としたときに、μからデーターxまでの距離を標準偏差σを1単位として表したものです。つまり、正規分布するデーターをそのばらつきの大きさにかかわらず標準化して表した距離ということになります。そういうことも意識しながら、

$$\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)^2 = X^2$$

と置いて、式を単純化します。

$$\mathrm{X} = \frac{x-\mu}{\sqrt{2}\sigma}$$

両辺を x で微分すると

$$\frac{dX}{dx} = \frac{1}{\sqrt{2}\sigma}$$

計算の便宜上、$\mathrm{dx} = \sqrt{2}\sigma \mathrm{dX}$と分離できるものとして、

$\mathrm{A} \int_{-\infty}^{\infty} e^{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$を

$$\mathrm{A} \int_{-\infty}^{\infty} e^{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = \mathrm{A} \int_{-\infty}^{\infty} e^{-X^2} \sqrt{2}\sigma dX$$

$$= \sqrt{2}\sigma A \int_{-\infty}^{\infty} e^{-X^2} dX$$

と変形します。これは分布の中心を0としてσを単位にした距離に変換する標準化のための作業です。

つまり、この問題は、

$$\int_{-\infty}^{\infty} e^{-X^2}\,dX$$

の答えを出す問題という問題に還元されます。

答えを先に言うと

$$\int_{-\infty}^{\infty} e^{-X^2}\,dX = \sqrt{\pi}$$

です。

一般の証明で、変数を X と書くのはあまり一般的でないので、変数をxと表して説明します。

$$I = \int_{-\infty}^{\infty} e^{-x^2}\,dx$$

原点を中心に左右対称なので、

$$\frac{I}{2} = \int_{0}^{\infty} e^{-x^2}\,dx$$

重積分と座標変換

突然ですが、ここで、両辺を二乗します。

$$\frac{I^2}{4} = \int_{0}^{\infty} e^{-x^2}\,dx \times \int_{0}^{\infty} e^{-x^2}\,dx$$

右辺の１番目の定積分と２番目の定積分を区別して別々に計算するものとして、２番目の定積分の変数を y として書き換えます

$$\frac{I^2}{4} = \int_{0}^{\infty} e^{-x^2}\,dx \times \int_{0}^{\infty} e^{-y^2}\,dy$$

これは積分したものの掛け算なのですが、x と y が互いに独立で直行しているとすれば、積分したものを掛け合わせることと、重積分することは同じ結果になります。

$$\int_{0}^{\infty} e^{-x^2}\,dx \times \int_{0}^{\infty} e^{-y^2}\,dy = \int_{0}^{\infty}\int_{0}^{\infty} e^{-x^2} \times e^{-y^2}\,dxdy$$

ということです。

3-4）

$$\int_{0}^{\infty} e^{-x^2}\,dx \times \int_{0}^{\infty} e^{-y^2}\,dy = \int_{0}^{\infty}\int_{0}^{\infty} e^{-x^2} \times e^{-y^2}\,dxdy$$

$$= \int_{0}^{\infty}\int_{0}^{\infty} e^{-(x^2+y^2)}\,dxdy$$

何をやっているのかというと、2 つの変数の積分の積を、２つの変数の積の積分として表し、それを $x=r\cos\theta$、$y=r\sin\theta$という極座標に変換して、θで積分するため、無理やり、x^2+y^2 を作っているのです。座標変換については III-3-3.ヤコビアン、III-3-4.座標変換を差

参考にしてください。準備が出来たので次のように座標変換します。

$$\mathrm{x} = \mathrm{r}\cos\theta$$

$$\mathrm{y} = \mathrm{r}\sin\theta$$

で極座標変換すると

$$\frac{I^2}{4} = \int_0^\infty \int_0^\infty e^{-(x^2+y^2)}\,dxdy = \int_0^{\frac{\pi}{2}} \int_0^\infty e^{-r^2}\,rdrd\theta$$

まず、内側の積分$\int_0^\infty e^{r^2}\,rdr$について

$$r^2 = s$$

とおいて

$$2\mathrm{r} = \frac{ds}{dr}$$

$$rdr = \frac{1}{2}ds$$

$$\int_0^\infty e^{-r^2}\,rdr = \int_0^\infty e^{-s}\frac{1}{2}ds$$

$$= \frac{1}{2}\int_0^\infty e^{-s}\,ds$$

$$= \frac{1}{2}[-e^{-s}]_0^\infty$$

$$= \frac{1}{2}\left\{-\frac{1}{e^\infty} - \left(-\frac{1}{e^0}\right)\right\}$$

$$= \frac{1}{2}\{0-(-1)\}$$

$$= \frac{1}{2}$$

$\frac{I^2}{4}$の式に戻って

$$\frac{I^2}{4} = \int_0^{\frac{\pi}{2}}\int_0^\infty e^{-r^2}\,rdrd\theta$$

$$= \int_0^{\frac{\pi}{2}}\int_0^\infty e^{-s}\frac{1}{2}ds\,d\theta$$

$$= \int_0^{\frac{\pi}{2}}\frac{1}{2}d\theta$$

$$= \frac{1}{2}\int_0^{\frac{\pi}{2}}d\theta$$

$$= \frac{1}{2}[\theta]_0^{\frac{\pi}{2}}$$

$$= \frac{1}{2}\left(\frac{\pi}{2} - 0\right)$$

$$= \frac{\pi}{4}$$

したがって、

$$\frac{I^2}{4} = \frac{\pi}{4}$$

$$I^2 = \pi$$

$$\mathrm{I} = \sqrt{\pi}$$

ところで、制約条件は

$$\int_{-\infty}^{\infty} A\, e^{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = 1$$

です。

$$\mathrm{A}\int_{-\infty}^{\infty} e^{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = \sqrt{2}\sigma A \int_{-\infty}^{\infty} e^{-X^2} dX$$

$$= \sqrt{2}\sigma A I$$

$$= \sqrt{2}\sigma\sqrt{\pi}A$$

$$= \sqrt{2\pi}\sigma\mathrm{A}$$

したがって

$$\sqrt{2\pi}\sigma A = 1$$

$$\mathrm{A} = \frac{1}{\sqrt{2\pi}\sigma}$$

となって、A の値が決まります。
もう忘れてしまったかもしれませんが

$$\mathrm{W}(\mu) = \mathrm{A}$$

としたのでしたね。

$$\mathrm{W}(x) \fallingdotseq \mathrm{W}(\mu) e^{\frac{-1}{2\sigma^2}(x-\mu)^2}$$

$$\mathrm{W}(x) \fallingdotseq \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-1}{2\sigma^2}(x-\mu)^2}$$

$$\mathrm{W}(x) \fallingdotseq \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

これは教科書などに出てくる正規分布の式です。

平均がμ、分散がσ^2の正規分布を$\mathrm{N}(\mu,\sigma)$と書きあらわします。$\mathrm{N}(0,1)$の正規分布を標準正規分布と言います。

xが$\mathrm{N}(\mu,\sigma)$に従うとき

$$\mathrm{P}(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

式 17

確かにnが十分に大きければ、２項分布は正規分布に近づくことが示されました。正規分布は二項分布の極限だと説明されることが多いと思います。確かにそうなのですが、それ以上に、２個の同じ二項分布を重ね合わせて畳み込んで４つ折りすることによって、確率の式の中に、分散という中心からの距離の尺度を持ち込んだことが大きいと思います。このプロセスを理解すると、正規分布の式がなぜそうなるのか理解できます。