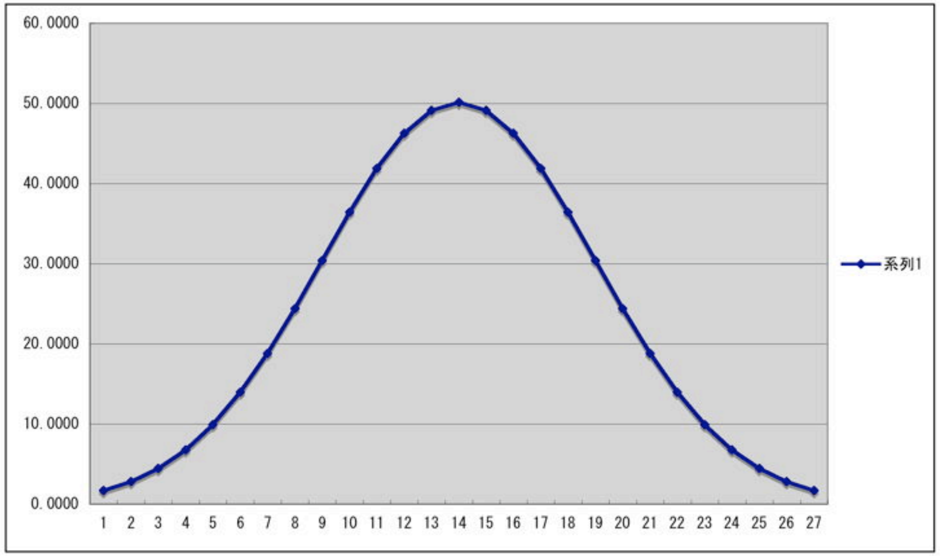
正規分布（step2）とガンマ関数

やっとこれからが正規分布へのアプローチである。
アプローチへの考え方だが、

・ 二項分布の場合は、値が離散的（1,2,3・・・）のような整数である。
・ ここで x を x+1 として f(x+1)/f(x) を求める。これも離散的なものである。
・ これを離散的なもの(x)から、連続的な数値(z)に拡大して行く、という手法で解を求めている。近似式など他のものを知らないで済む方法である。
・ まず、理解しなければならないことがある。式の工夫で $z=\frac{x-\mu}{\sigma}=\frac{x-np}{\sqrt{np(1-p)}}\cdots(1)$ とすることである。ここが正規化のはじまりである。
・ x は離散的数値、μ は平均値、σ は偏差、n はサンプル数、p はある事象が発生する確率、(1-p)は前事象とならない確率、である。「平均値を基準としてサンプルがどれだけ偏差があるかという比」であり、(1)は x から平均値を引き、それを偏差で割っているのでここで正規化がなされてくると言えるだろう。それをさらに大元の変数 n,p で表現している「ほぼ実際と合う一つの定義」と考える他はないだろう。ということである。最小二乗法などもあって、中間では二乗が使われているが、ここでは σ のように二乗ではなくなっているので次元も合っている。
・ このようにすると、(x)は離散的な値なのだが、n を∞にして行くと(z)は連続的な値であると考えて良くなる。



まず、勘を養うためにマイクロソフト･エクセルでの計算結果を示しておこう。

式は $h(x)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\cdots(2)$ であり、グラフはその結果である。

計算は上式のように「何でもよいのだが
・(x)の平均値(μ=5)を 5 と定め、
・ある標準偏差(σ=0.05)を定め、
・平均値前後の値(y)をある一定間隔(0.01)で値を定め、
計算しグラフ化したものである。

　標準偏差が小さければ、当然だが平均値に近いところに値は集中し、山も高くなることが推察できるだろう。

離散から連続へ　いよいよやっと計算式の導出への取り組みである。

$z=\frac{x-\mu}{\sigma}=\frac{x-np}{\sqrt{np(1-p)}}\cdots(1)$　の値は離散的な x に対して $\Delta z=\frac{x-\mu}{\sigma}=\frac{x}{\sigma}-\frac{\mu}{\sigma}$　のように等間隔に並んでいることがわかる。

即ち、$x\to x+\Delta x$　$z\to z+\Delta z$　ただし、Δx=1（離散）である。

$\frac{x+\Delta x-\mu}{\sigma}=\frac{x+\Delta x-np}{\sqrt{np(1-p)}}=\frac{x-np}{\sqrt{np(1-p)}}(=z)+\frac{\Delta x}{\sqrt{np(1-p)}}(=\Delta z)=z+\frac{\Delta x}{\sqrt{np(1-p)}}$　　$\Delta z=\frac{\Delta x}{\sqrt{np(1-p)}}=\frac{1}{\sqrt{np(1-p)}}\cdots(2a)$　　$\sqrt{np(1-p)}=\frac{1}{\Delta z}\cdots(2b)$

$z=\frac{x-\mu}{\sigma}=\frac{x-np}{\sqrt{np(1-p)}}=\frac{x-np}{\Delta z}$　　$x=np+\frac{z}{\Delta z}\cdots(2c)$　などの変形が出来、(2a)からは n→∞により離散的な数値を連続的な数値として見なして良くなって行くことがわかる。

ここからは導関数的なものの導出を行うことになる。それを積分すれば関数が求まる。そのために、ある離散値と次の離散値の比（導関数相当）を考える。比というのは極限では微分になる。このように次の x との比を求めているのは、良くやる方法に思われる。

$$f(x)=\frac{n\ !}{x\ !(n-x)\ !}\cdot p^{x}(1-p)^{(n-x)}$$

$$\frac{f(x+1)}{f(x)}=\frac{\frac{n\ !}{x\ !(x+1)(n-x-1)\ !}\cdot p^{x+1}(1-p)^{(n-x-1)}}{\frac{n\ !}{x\ !(n-x)\ !}\cdot p^{x}(1-p)^{(n-x)}}=\frac{\frac{1}{(x+1)(n-x-1)\ !}\cdot p(1-p)^{(-1)}}{\frac{1}{(n-x)\ !}}=\frac{(n-x)p}{(x+1)(1-p)}\cdots(3a)$$

Z の分布を g(z)とすると、

$$g(z)=\frac{\Delta x}{\Delta z}f(x)=\frac{1}{\Delta z}f(x)\qquad \because \Delta x=1\qquad g(z+\Delta z)=\frac{1}{\Delta z}f(x+1)\qquad \therefore\ \frac{g(z+\Delta z)}{g(z)}=\frac{f(x+1)}{f(x)}=\frac{(n-x)p}{(x+1)(1-p)}\cdots(3b)$$

(3b)に(2c)を代入して x を消去する。

$$\frac{g(z+\Delta z)}{g(z)}=\frac{\frac{1}{(\Delta z)^2}-p\frac{z}{\Delta}}{\left(np+\frac{z}{\Delta z}+1\right)(1-p)}=\frac{np(1-p)-\frac{pz}{\Delta z}}{np(1-p)+(1-p)\left(\frac{z}{\Delta z}+1\right)}\qquad np(1-p)=\frac{1}{(\Delta z)^2}\qquad \frac{g(z+\Delta z)}{g(z)}=\frac{\frac{1}{(\Delta z)^2}-p\frac{z}{\Delta}}{\frac{1}{(\Delta z)^2}+(1-p)\left(\frac{z}{\Delta z}+1\right)}=\frac{1-pz\Delta z}{1+(1-p)\left\{z\Delta z+(\Delta z)^2\right\}}$$

$$\frac{g(z+\Delta z)-g(z)}{\Delta z}=\left(\frac{g(z+\Delta z)}{g(z)}-1\right)\frac{g(z)}{\Delta z}=\frac{1-pz\Delta z-1-(1-p)\left\{z\Delta z+(\Delta z)^2\right\}}{1+(1-p)\left\{z\Delta z+(\Delta z)^2\right\}}\frac{g(z)}{\Delta z}=\frac{-z\Delta z-(1-p)(\Delta z)^2}{1+(1-p)\left\{z\Delta z+(\Delta z)^2\right\}}\frac{g(z)}{\Delta z}$$

結局　$\dfrac{g(z+\Delta z)-g(z)}{\Delta z}=\dfrac{-z-(1-p)\Delta z}{1+(1-p)\left\{z\Delta z+(\Delta z)^2\right\}}g(z)$　　$\Delta z\to 0$　　$\dfrac{dg(z)}{dz}=-zg(z)\cdots(4)$　の微分方程式となる。

導関数的なものを求めようとした結果、微分方程式が出てきたと言うことであるのだが、このように前後の関連から導関数を求めようとする場合には、自然に微分方程式が出てくるものかも知れない。
この部分は突破するのにスターリングの近似式などというものを使っているものもあるのだが、それを覚えるのも嫌と言うことで探した結果、こんな計算を発見した。とは言え、間違いなくコツコツと進むしかない。

更に、$\dfrac{1}{g(z)}dg'(z)=\dfrac{1}{g(z)}\dfrac{dg(z)}{dz}=-zdz$　　$\int\dfrac{1}{g(z)}\dfrac{dg(z)}{dz}dz=\int -zdz$　　$\int\dfrac{1}{g(z)}dg(z)=\int -zdz$　　$\log\ g(z)=-\dfrac{z^2}{2}+C'$

$$g(z)=e^{\left(-z^2/2+C'\right)}=e^{c'}e^{-z^2/2}=Ce^{-z^2/2}\cdots(5)$$

結果を先に示せば、積分の結果は１であるから、$1=\int_{-\infty}^{\infty}g(z)dz=C\int_{-\infty}^{\infty}e^{-z^2/2}dz=C\sqrt{2\pi}$　　$\therefore C=\dfrac{1}{\sqrt{2\pi}}\cdots(6)$

となるが、(6)のＣがわかるためには、ガンマ関数がわからなければならない。
こんなことで正規分布とは、統計・確率・平均値・偏差・微分方程式・ガンマ関数などがわからなければ導けない結構な代物なのであることがわかる。そこでガンマ関数をおさらいすることになる。

ガンマ関数

・ガンマ関数は　$\Gamma(s)=\int_0^{\infty}x^{s-1}e^{-x}dx\cdots(g1)$　で表される関数である。学生時代は、何の説明もなくいきなりこれが出てきたような記憶がある。何に使うのかもわからないし、授業だけでその後の私の経過では使った記憶もない。同時にベータ関数も示された。
・それから 50 年経過したが、正規分布というものを数式まで厳密に理解しようとすると最終段階で、$\sqrt{\pi}$　などという要素が出てくることがわかった。通常の積分ではなさそうである、わからない。$\sqrt{\pi}$?不思議である。そう言えば　$\Gamma(1/2)=\sqrt{\pi}$　と言う記憶がある。
・考えてみると、正規分布は離散的な数値の二項分布を正規化して連続関数化するものであり、ここには　n！要素が含まれている。ガンマ関数は離散的な　n！と言うものを連続関数化するものであるから、当然重要な関係を持つものと思って差し支えないだろう。

・そんなことで、正規分布の追求を最終段階で一旦中断し、ガンマ関数のおさらいをすることになった。最初に言っておきたいのだが n！の連続関数化が何故(1)なのかはわからない。これは数学者の追求の結果なのだろう。
(2)は基本的な部分積分の確認である。(3)も部分積分である。公式など覚えるのは少ない方がよい。滅多にないことだし忘れるので、私は $\left(f(x)g(x)\right)' = f'(x)g(x) + f(x)g'(x)$ と大元からはじめる。x は自然数として、

$$\left(e^{-x}\right)' = \left(\frac{1}{e^{x}}\right)' = \frac{-e^{x}}{\left(e^{x}\right)^{2}} = \frac{-1}{e^{x}} = -e^{-x} \cdots (g2) \qquad \left(x^{s-1}e^{-x}\right)' = (s-1)x^{s-2}e^{-x} - x^{s-1}e^{-x} \to x^{s-1}e^{-x} = (s-1)x^{s-2}e^{-x} - \left(x^{s-1}e^{-x}\right)' \cdots (g3)$$

$$\int_0^\infty x^{s-1}e^{-x}dx = \int_0^\infty (s-1)x^{s-2}e^{-x} - \int_0^\infty \left(x^{s-1}e^{-x}\right)' dx = \int_0^\infty (s-1)x^{s-2}e^{-x} - \left[x^{s-1}e^{-x}\right]_0^\infty = \int_0^\infty (s-1)x^{s-2}e^{-x} - (0-0) = \int_0^\infty (s-1)x^{s-2}e^{-x} \cdots (g4)$$

$$\therefore \quad \Gamma(s) = (s-1)\Gamma(s-1) \cdots (g5)$$

ただし、$\left[x^{s-1}e^{-x}\right]$の無限大での収斂は何かを見ていただこう。

$$\begin{cases} \Gamma(s) = (s-1)\Gamma(s-1) \\ \Gamma(s-1) = (s-2)\Gamma(s-2) \\ \cdots \\ \Gamma(2) = (1)\Gamma(1) \end{cases}$$

辺々積をとれば、$\left(\Gamma(s) = (s-1)(s-2)\cdots 2\cdot 1\cdot \Gamma(1)\right.$ となる。

これでガンマ関数がわかったかというとそうではないが、何かの本でも見ていただこう。

さて正規分布の式の追求をやって行くと、(5)(6)から、
$N = \int_{-\infty}^{\infty} e^{-x^2} \cdots (g6)$ のような積分に突き当たった。(g1)式で s=1 の場合に似ている。一見簡単な積分そうだが、実はそうではなくて、ちょっとしたテクニックを要する積分である。この非積分関数は、数学の王と呼ばれたガウス（1777-1855）が天文台勤務のとき天体の観測誤差を集計していて発見した関数で、観測結果が正確な値から x だけずれることの起こりやすさが e^{-x^2} に比例することを見出したと言うことである。したがって、多くの確率現象は、この関数により記述され、統計学や物理学や経済学では頻繁に利用されると言うことらしい。計算は下記のように重積分に拡大して何の計算もなくあっさりと行える。

$I = \iint_D e^{-(x^2+y^2)}dxdy = \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} e^{-x^2}e^{-y^2}dx\right]dy = \int_{-\infty}^{\infty} e^{-y^2}\left[\int_{-\infty}^{\infty} e^{-x^2}dx\right]dy \cdots (g7)$ もし〔〕内が N という値になるならば、x,y は対照的であり、x と y での積分は同じであるから、$I = \int_{-\infty}^{\infty} e^{-y^2}Ndy = N\int_{-\infty}^{\infty} e^{-y^2}dy = N^2 \cdots (g8)$ のようになり、I が求まればルートをとる

ことによってＮは求まると言うものである。このまま直交座標で求めるのは難しいらしい、そこで、極座標を使う。

$x = r\ \cos\theta \quad y = r\ \cos\theta \qquad dxdy = rdrd\theta$

$$I = \iint_D e^{-(x^2+y^2)}dxdy = \iint_E e^{-r^2(\cos\theta^2+sin\theta^2)} \cdot r \cdot drd\theta = \iint_E re^{-r^2}drd\theta = \int_0^\infty re^{-r^2}\left[\int_0^{2\pi}1d\theta\right]dr = 2\pi\int_0^\infty re^{-r^2}dr = 2\pi\left[\frac{-e^{-r^2}}{2}\right]_0^\infty = \pi = N^2$$

$\therefore\ N = \sqrt{\pi}$

$dxdy = rdrd\theta$ は直感的に考えていただければよいだろう（厳密に考えるとまた難しそう）。

さて、(g1)はｙ軸対称であるから、半分は $N = \int_0^\infty e^{-x^2} = \dfrac{\sqrt{\pi}}{2}$ である。これがわかった後はこんな風になる。

ここで、$t = \dfrac{x^2}{2} \qquad dt = xdx \qquad x = \sqrt{2t} \qquad dx = \dfrac{1}{x}dt = \dfrac{1}{\sqrt{2}\sqrt{t}}dt = \dfrac{1}{\sqrt{2}}t^{-\frac{1}{2}}dt$

$$\int_0^\infty e^{-\frac{x^2}{2}}dx = \int_0^\infty e^{-t}\frac{1}{\sqrt{2}}t^{-\frac{1}{2}}dt = \frac{1}{\sqrt{2}}\int_0^\infty t^{-\frac{1}{2}}e^{-t}dt = \frac{1}{\sqrt{2}}\int_0^\infty t^{\left(\frac{1}{2}-1\right)}e^{-t}dt = \frac{1}{\sqrt{2}}\Gamma\left(\frac{1}{2}\right) = \frac{\sqrt{2\pi}}{2}$$ となる。

−∞から+∞の積分なら倍になり $\sqrt{2\pi}$ である。こんなことでやっと、正規分布の関数の原形 $\dfrac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}$ が求まることになる。

$z = \dfrac{x-\mu}{\sigma} = \dfrac{x-np}{\sqrt{np(1-p)}}\cdots(1) \qquad h(x) = \dfrac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ であるが､ｘ軸にμだけ平行移動して $z' = \dfrac{x}{\sigma}$ のように考えてもよいから（無限大までの積分なので、一旦簡略化して、結果でμ戻せばよい）、これで考えてみると、

$t = \dfrac{1}{2}\left(\dfrac{x}{\sigma}\right)^2 \qquad dt = \left(\dfrac{x}{\sigma}\right)dx \qquad x = \sqrt{2t\sigma^2} \qquad dx = \dfrac{\sigma}{x}dt = \dfrac{1}{\sqrt{2\sigma^2}\sqrt{t}}dt = \dfrac{1}{\sqrt{2\sigma^2}}t^{-\frac{1}{2}}dt$ となり $h(t) = \dfrac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(t-\mu)^2}{2\sigma^2}}$ が得られる。