正規分布（step1）

・二項分布はちょっと根気を入れれば理解できるものである。そこをベースとしたポアソン分布は会社に入ってしばらくして、式を導出計算して確認したことがある。仕事上で数表を見る必要があったからである。その後、ポアソン分布はイギリスの将校が「兵隊が馬にけられて死亡する確率」から定義されたものと見たことがある。
・ところが、正規分布というものは頻繁に言葉を聞きながら、何だかわからないもので、数表も使ったこともない。そのくせ「３σだから」などと話をすることがある。
・「わかっても居ないで、数表しか使えず、したり顔」も出来ない。そこで理解へのチャレンジをはじめることにした。結果だが「完全に理解するには、結構過程が多く、その過程の中に面倒なところ（近似）部分があって時間がかかりそう」と言うことがわかったが、やってみることにした。しかし、この年齢では使うこともないだろう。
・この分布はガウス分布とも言われており、ガウスが式を導出したものらしいが、彼ほどの頭の人間はあっという間にやってしまうのだろう。

・何でも良いのだが、例えば「ある寸法の棒をつくろうとした場合」沢山つくって行けば、狙いの寸法の前後に集中して行くことは素直に理解してよいだろう（大数の法則）。狙いを定めてつくっているのだから、当然である。
・二項分布と言うものは、ｎを 1,2,3・・・のように離散変数（整数値しかとり得ない数）だが、これでは不連続な函数なので、何とか連続関数にしたいということは当然の欲求である（こんなことは「（ガンマ）函数に似ている）。
・式を先に示してしまおう。式は $h(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ のようになり、一見して二項分布の連続関数化により何で$\sqrt{2\pi}$のようなものが出てくるのかは不思議に思うのだが、これはいろいろなところで出てくる係数で、数学に長けている人にはわかるのではないだろうか。

・そこで「二項分布の連続関数化を試みる」と言うのが狙いで、連続関数化でき、計算容易ならば、それは貴重なものとなるのは明かである。
・もし、そうならばここでの追求も考えやすくわかりやすそうである。何でそんな関数が必要なのかもわからず追求するほど無為なものはない。
・ところが、授業などは「正規分布の式はこうである、表を見れば数値がわかる」である。こんな事では、核心がわかっていないのだから、使う気にはならない。

(1) やろうとすると、まず出てくるのは「偏差・分散」である。学校などでは先生は「二乗すると」で済ませてしまう。何で二乗するのかと言うことは過去から疑問に思っていたのだが、聞いたことも調べたこともない。ここからがはじまりである。
(2) いずれ、詳細に調べてみようと思うのだが、ちょっとかじって推測すればこんなことなのではないだろうか。
・ 平均値は説明の必要はないだろう。
・ 偏りなどを数値化する場合、平均の両側に偏ることになる。＋と−でΣは零となり具合が悪い。二乗すれば＋と−のどちらの偏りでも＋となり具合がよい。それをわかって利用するなら、何ら問題はない。
・ こんな推測を書いた後、目にした記事があった。その記事を読んだところ、多少は当たっているような感じになっている。

では「正規分布」へのチャレンジと開始しようとしたのだが、なかなか適切なものがない。どんな点かというと、ある説明では途中に「スターリングの近似式」と言うものが出てくる。私から観ると本末転倒な気がする。どうしてもそれを理解しないと通過できないのなら仕方がないのだが、そんなものは知らないのだから、使わなくても済む説明が欲しいのである。

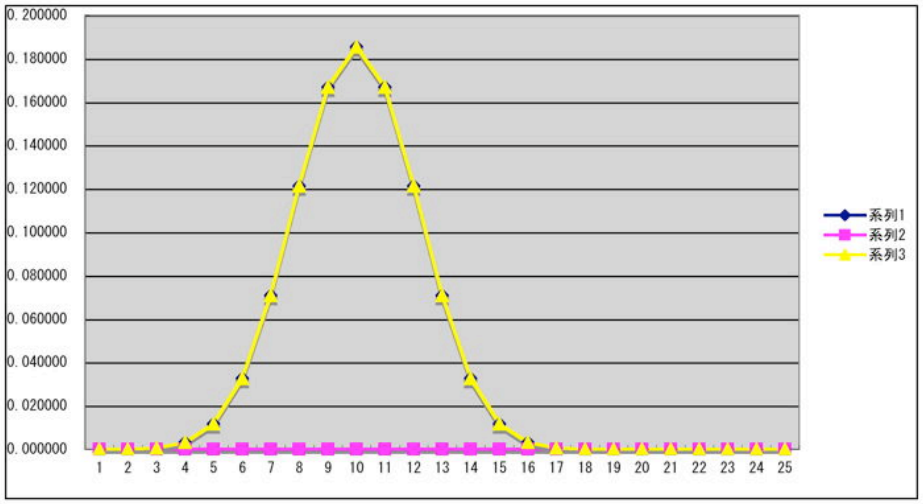
二項分布からの発展で「わかる」と思われる説明があったので、それをベースにチャレンジすることにした。まずは、二項分布からのおさらいである。
・ 白と黒が同じ数均質に混じり合った碁石が数多く(n)個ある。
・ 目隠しでもして、この中の(x)個を取り出すと、例えば全部が白となる確率は $\left(\frac{1}{2}\right)^x$ である。
・ 後の確率は、白と黒が混じったものまたは全部黒で、その確率は $\left\{1-\left(\frac{1}{2}\right)^x\right\}$ である。ただし「大数の法則」といって数多くこれを繰り返した結果が、この確率に近づくということである。
・ もし、ここで碁石が白(p)、黒(q)の割合で混じっているものとする。この場合は、碁石を一つとったとき白に当たる確率は(p)であるから、(x)個をとった場合に、全部が白となる確率は p^x である。
・ 同様にして、一つだけ黒が混じる確率は $p^{(x-1)}\cdot q^1 = p^{(x-1)}\cdot(1-p)^1 = p^{(x-1)} - p^x$ である。
・ 従って、白が(x)個である確率は、$p^x\cdot q^{(n-x)}\cdots(1)$ となる。この計算結果だが、P=q=0.5の場合には $0.5^x\cdot 0.5^{(n-x)} = 0.5^n$ という明らかに一定値となる。P=0.1 q=0.9のような場合には、一見すると値は徐々に大きくなるように感じるのだが、計算してみればわかるように、当初のある部分で最大に達して、急激に減少して行くことになる。
・ 今度は組合せである。白黒の碁石に番号があるものと考え、(x)個の中の白でも黒でも番号が一つでも違うものがある場合は、別の

組合せであるから、この組合せの数がどれだけあるかを考えなければならない。白が(m)である場合の組合せ数は

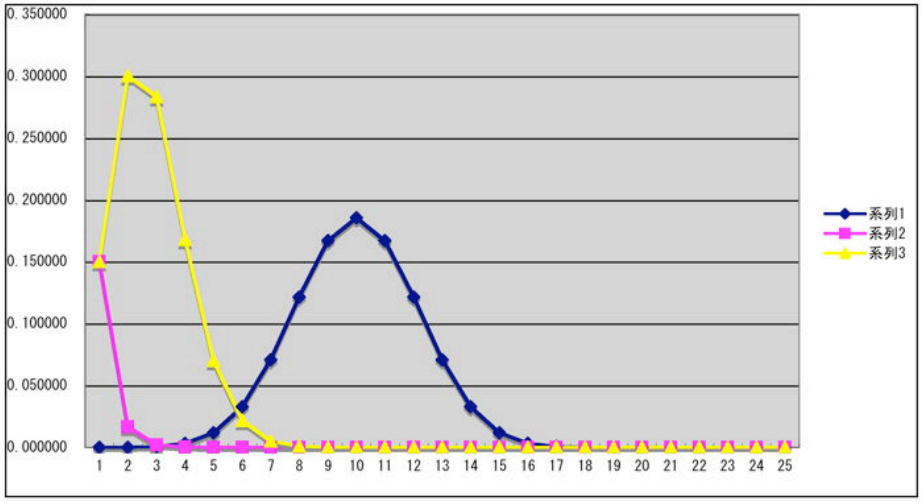
$${}_nC_x = \frac{n\ !}{x\ !(n-x)\ !} \cdots (2)$$ である。この分布は左右対称の山形になる。

・ 全体の確率は　f(x)=（その混合の組合せ数）X（混合確率）で、$$f(x) = \frac{n\ !}{x\ !(n-x)\ !} \cdot p^x q^{(n-x)} \cdots (3)$$ である。(1)(2)の積が結果としてどうなるかだが、(1)の影響を大きく受け形としては山を左右に移動させたようなものになる（山の高さは勿論変わる）。

計算は実際にn=12や18としてエクセルなどで実施したらどうだろう。私が学生時代の50年も前は、今のように高速なパソコンなどはないから、頭と式だけで理解し実際に計算してみることなどなかった。頭の良い人はそれでも感覚を得てしまうのだろうが、私などは駄目だったし、その後実験や仕事上で使うことのなく還暦を過ぎ、古希も過ぎてしまった。しかし、計算してみると少々感覚をつかめたような気分になっている。エクセルでの計算結果を示してみよう（n=18として：赤=(1)、青=(2)、黄=(3)



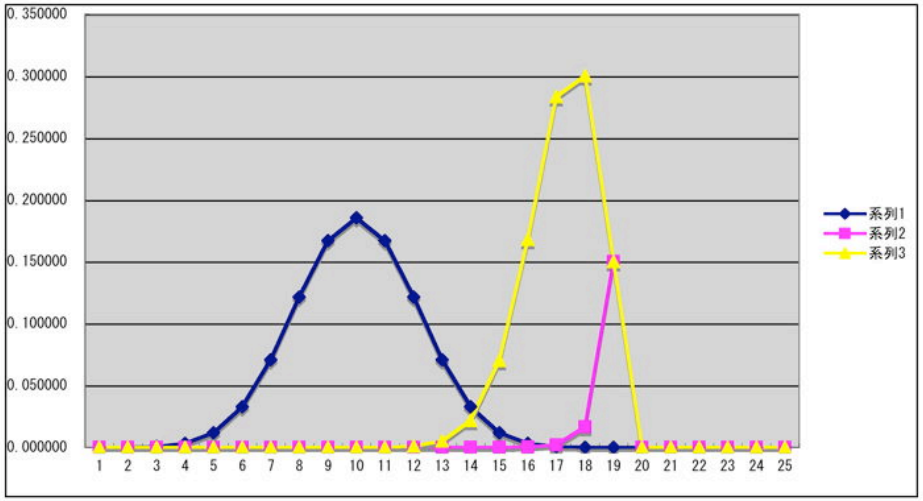
p=q=0.5の場合　二つのグラフが重なり合っている。



P=0.1 q=0.9 の場合

P の小さな、これだけは少々使ったことがある。ポアソン分布（Poisson）と言うことなのだが、通信分野に居て、電話の加入者数に対してどれだけ交換局内部や局間の中継の線などを用意すれば話中率をある限度内に収めることが出来るかという計算をして線を用意しなければならないからである。特に、一般の家庭の電話の場合などは、電話の利用はある時間帯に集中（最繁時などという）するのだが、その時間帯ですら平均すれば利用度は低いからである。共有される電話線というものは例えば 1/10 程しか設置されていない。だから、異常事態では電話がつながらなくなって当たり前である。しかし、細かいことは忘れてしまった。伝送路とコンピュータチップというインターネット時代になって、今こんな計算が使われているかどうかは知らない。

そもそも、これはトラフィック（trafiic）計算というものであって、高速道路なども同じ考えで計算出来るものなのである。ただ、電話線の場合は束で敷設でき場所もあまりとらない。高速道路はそうは行かない。インフラの重みが何桁違いである。



P=0.9 q=0.1 の場合
結果は前グラフと対象になっているのがわかるだろう（数式上もそうだが）。

先の正規分布に進むために、二項分布の性質を知っておかなければならない。
まず「平均値」である。

$(p+q)^x$ の展開そのものが、(3)の $f(x)=\dfrac{n\ !}{x\ !(n-x)\ !}\cdot p^x q^{(n-x)}=f(0)+f(1)+f(2)\cdots+f(n)$ であり、かつ全体は１であることはわかっている。

更に $\sum x\cdot f(x)=\dfrac{f(0)+f(1)+f(2)\cdots+f(n)}{n}=\mu$（平均値$\mu$）であることもわかるだろうから、これを頭に置いておく。

(3)の $\sum f(x)=\sum {}_nC_x\cdot p^x q^{(n-x)}=(p+q)^n\cdots(a1)$

両辺を p で微分すると $\sum x\cdot {}_nC_x\cdot p^{(x-1)}q^{(n-x)}=n(p+q)^{n-1}\cdots(a2)$ （ x や q はこの場合常数として扱って）

更に両辺に p を掛けて $\sum x\cdot {}_nC_x\cdot p^{(x)}q^{(n-x)}=np(p+q)^{n-1}\cdots(a3)$ 左辺は平均値（μ）であり、左辺は p+q=1 だから np と表せる。

よって、$\mu=np\cdots(4)$ である。

平均値は単に n と言う全体数と p と言う混じり割合率の積となっていると言うことである。
微分などで騙されたようだが、こんな式操作ができるのは、辿り着くところが推測できる人に出来ることなのだろう。p+q=1 であるが、最初にこれを常数としてしまっては、先に進まなくなる。変数として扱ってきて、最後に常数化するのが味噌である。

今度は「分散=σ^2」である。

(a2)を更に p で微分する。 $\sum x(x-1)\cdot {}_nC_x\cdot p^{(x-2)}q^{(n-x)}=n(n-1)(p+q)^{n-2}\cdots(b1)$

両辺に p^2 を掛けると、 $\sum x(x-1)\cdot {}_nC_x\cdot p^{(x)}q^{(n-x)}=\sum x^2\cdot {}_nC_x\cdot p^{(x)}q^{(n-x)}-\sum x\cdot {}_nC_x\cdot p^{(x)}q^{(n-x)}=n(n-1)p^2(p+q)^{n-2}\cdots(b2)$

更に、 $\sum x(x-1)\cdot {}_nC_x\cdot p^{(x)}q^{(n-x)}=\sum x^2\cdot {}_nC_x\cdot p^{(x)}q^{(n-x)}-np=n(n-1)p^2\cdots(b3)$

更に、 $\sum x^2\cdot {}_nC_x\cdot p^{(x)}q^{(n-x)}=n(n-1)p^2+np\cdots(b4)$

「二項分布の分散σ^2」であるが、定義は$\sigma^2=\sum(x-\mu)^2 f(x)$、であり、これを求めなければならない。
そこでまた、戻らなければならない。今度はモーメントである。モーメントを次のように定義するのだが、これは力学などで出てく

る初歩的なものと同じなので問題ないだろう。

$$E\left[\varphi(X)\right]=\begin{cases}\sum\varphi(x_i)f(x_i)=\varphi(x_1)f(x_1)+\varphi(x_2)f(x_2)+\cdots+\varphi(x_n)f(x_n)\cdots(c1)\\ \int_{-\infty}^{\infty}\varphi(x)f(x)dx\cdots(c2)\end{cases}$$

(c1)は離散的な場合、(c2)は連続的な場合である。

$E\left[\varphi(X)\right]$は$\varphi(X)$の期待値(expectation value)と言う。

二項分布のモーメントは

$\sigma^2=E\left[(X-\mu)^2\right]=\int_{-\infty}^{\infty}\left(x^2-2\mu x+\mu^2\right)f(x)dx=E\left[X^2\right]-2\mu E[X]+\mu^2E[1]=E\left[X^2\right]-2\mu^2+\mu^2=E\left[X^2\right]-\mu^2$　のように書けるのだが、この辺のことは、先へ進みたいので置いて行き、機会あったら理解したい

と言うことで、$\sigma^2=\sum(x-\mu)^2f(x)=\sum_{x=0}^{n}x^2f(x)-\mu^2=\sum_{x=0}^{n}x^2{}_nC_xp^xq^{n-x}-\mu^2$　に(4)と(b4)を代入すると、

$$\sigma^2=\sum(x-\mu)^2f(x)=\sum_{x=0}^{n}x^2f(x)-\mu^2=\sum_{x=0}^{n}x^2{}_nC_xp^xq^{n-x}-\mu^2=n(n-1)p^2+np-(np)^2=np(1-p)\cdots(5)$$

と言うことでやっと二項分布の分散が求まった。