JAPLA 研究会資料　2016/10/22

χ^2をめぐって、私なりの統計学への再考
－専門用語のネーミングまで－

西川　利男

統計学では、○○分布という用語、たとえば
　２項分布、正規分布、から χ^2 分布、F 分布など
たくさん出てくる。しかし、これらを一緒くたに、まとめて説明してしまうことは、統計学の理解を誤らせることにもなる。

先月の JAPLA 例会で、志村正人氏と山本洋一氏のお二人により、χ^2 をテーマに２つの論文が提出された。私としては、今まで分かったつもりでいたのが、そうでなく大いに反省させられた。あらためて統計学を勉強し直している。[1]

［1］蓑谷千凰彦「統計学入門」東京図書、新装合本(2004).

１．χ^2値からはじめる

まずは、鈴木義一郎先生の「J 言語による統計分析」第３章§3.7 理論と実際(p. 58 以下)にそって、式の展開などをていねいに追っていくことにする。

χ^2 値とは、理論的な値と現実の統計の実測値とのずれの大きさを示すものとして定義される。

そして同書では、次のように具体的に説明されている。ある病院で生まれた子どもを調べたら、900 人のうち男児は 480 人、したがって女児は 900-480=420 人という統計データが得られた。

一方、男女の生まれる割合は 1/2 という理論があったとする。上の統計データに基づいてこの理論は正しいかどうかを検証したい、という問題である。

理論値と実測値とのずれの平方の和の理論値に対する割合を c^2 として示す。

$$c^2 = \frac{(480-450)^2}{450} + \frac{(420-450)^2}{450}$$
$$= 4$$

上の記述を一般的に式を用いておこなう。

統計として収集された人数を n 人として、そのうち x 人が理論的に属性を持つ割合を p とする。その人数は np 人で、そうでない人数は $n(1-p)$ 人となる。上の式はつぎのようになる。

$$c^2 = \frac{(x-np)^2}{np} + \frac{((n-x)-n(1-p))^2}{n(1-p)}$$
$$= \frac{1}{np(1-p)}\{(x-np)^2(1-p) + p(n-x-n+np)^2\}$$
$$= \frac{1}{np(1-p)}(x-np)^2$$
$$= \left(\frac{x-np}{\sqrt{np(1-p)}}\right)^2$$

ここでの最後の式が、ふつう χ^2 値と呼ばれる。

この後は同書では、、次のように説明されている。

xを２項分布に従う変量と考えると

$$Z = \frac{x - np}{\sqrt{np(1-p)}}$$

は平均が0、標準偏差が1の標準正規分布に従うものと考えてよい。

したがって

$$P\{C^2 \geq 4\} = P\{|Z| \geq 2\} = 0.05$$

という関係が成立している。だから、理論と実際とのずれの値C^2が4を越えるようなら実際の値は理論からずれている。（途中略）男女半々ずつ生まれるという仮定は正しくないと判断される。

ここで、Jで上のχ^2値の計算を追ってみよう。

```
   x =: 480
   n =: 900
   p =: 0.5
   *: (x - n*p) % %: (n*p) * (1-p)
4
```

普通は経験的に、 p=0.514 と言われている。[1] p.187

```
   p =: 0.514
   *: (x - n*p) % %: (n*p) * (1-p)
1.34666
```

するとχ^2値は上のようになる。

このように「さらっと」記述されているが、統計学での多くの基礎知識を前提にしているので、その内容を理解するのは、初学者にとっては仲々難しいと、私は思う。

例えば、xを２項分布に従う変量とすると、nが大きくなると正規分布として扱うことが出来る、とされているが、これをキチンと証明するには、かなりな数学が必要となる。ポイントは組み合わせの数に対して、kが大きいときのスターリングの公式

$$k! = \sqrt{2\pi k}\, k^k e^{-k}$$

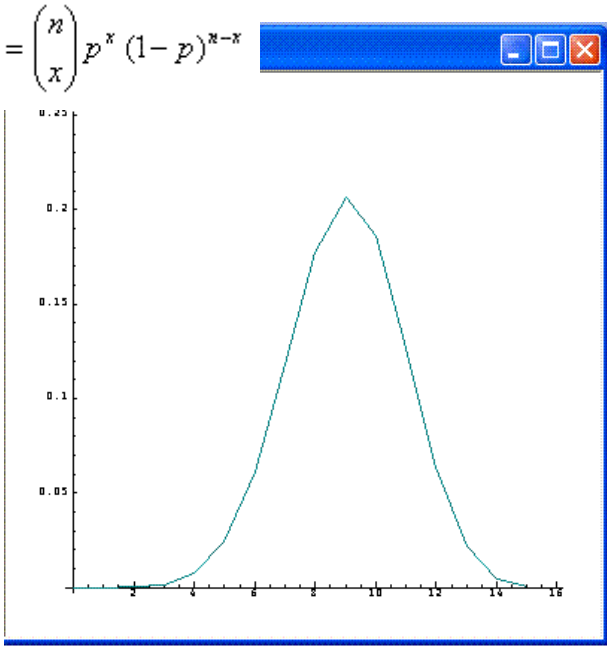
を適用して計算する。先の蓑谷先生の書[1]では、ド・モアブルーラプラスの定理としてp.196-198の３ページにわたって数式による証明がなされている。

このように、統計学の書は、いろいろなレベルの読者に向けてと言えば聞こえがよいが、数式を出したもののその説明が天下り式で、ブラックボックス化している。

ここでは、Jのグラフィックスの図示で、納得することにしよう。同書p.185から

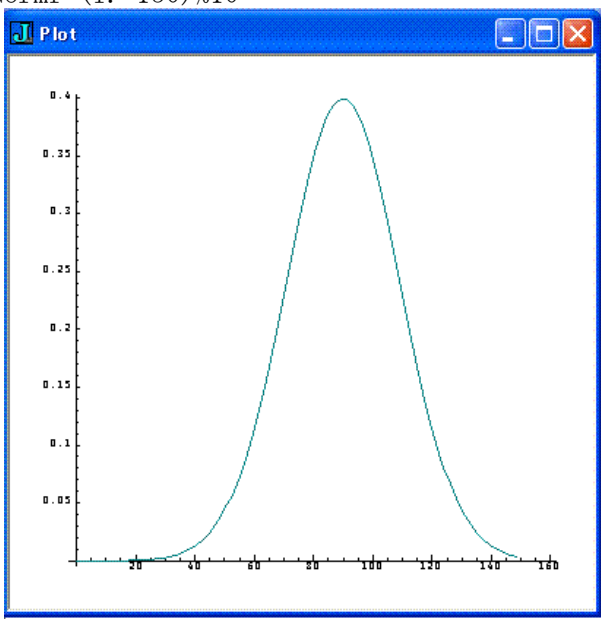
２項分布 $f(x) = \binom{n}{x} p^x (1-p)^{n-x}$

```
   n =: 15
   x =: i. >: n
   p =: 0.6
   require 'plot'
   plot (x!n) * (p^x) * (1-p)^(n-x)
```



正規分布 $$f(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{1}{2\sigma^2}(x-\mu)^2)$$

```
Normf =: 3 :                                        0
(0, 1) Normf y.
:
'm s' =. x. NB. m: mean, s: st. dev.
(% %: o. 2) * ^ - -: *: (y. - m) % s
)
plot (9, 1.9) Normf (i. 150)%10
```



私としては、統計学そのものの解説などするつもりはない。ただ、以下に、統計学に接した私なりの感想をいささか述べさせていただく。

２．統計学になぜ確率が必要なのか？

・統計学はまず収集したサンプル値を整理して、見易くする道具として役立つ。
この段階では確率は必要ない。……>記述統計学
・次に、多くの場合は収集したサンプル値から、収集の元となった母集団の性質を推測したい。……>推測統計学
・しかし、知りたいのは母集団の性質なので、当然これは未知数である。
・そこで、母集団のモデルとなるための人工の値の集合がほしい。このモデルの作成のために、いろいろな確率の問題を利用する。
・コインを投げたり、サイコロを振ったり、というのはそれにより確率のモデルと人工の母集団を作るためである。
・あらためて、従来の数学では、元となる変数はあるときはとびとびの整数、またあるときは連続した実数が変数として任意に、何か操作をしたり、現象が起ったりしたときは、それに対応して、関数として値が返さる。
ところが、コイン投げやサイコロなどの場合には、1/2や1/6のような決まった値しか変数として取れない。そしてこれが分布を持っている。これが確率変数であるが、普通の変数ー>関数という数学と勝手がちがう。
・確率変数のモデルとしては、２項分布から幾何分布、ポアソン分布など数学の式で表されるものが母集団として使われる。
・自然現象は連続であり、これに対してド・モアブル、ラプラス、ガウスらにより正規分布が出された。そして、実際の多くの統計の母集団は、正規分布をなすとして計算処理される。先述したが、ここにいたる真の理解は仲々難解である。
・用語の使用として、私としては、種々の確率分布関数とは違うので、推定、検定のパラメータである χ^2 値、F値は分布と呼ばない方が良いと思う。

３．おわりに、大隈良典先生のノーベル賞テーマーオートファジーについて一言

今年もひきつづき、ノーベル賞生理医学賞受賞者として、東京工業大学の大隈良典教授が選ばれた。日本人として誇らしいビッグニュースである。

日本科学未来館でもこれについていろいろ紹介している。その中で気になった一言。

受賞されたテーマである「オートファジー(Autophagy)」なる専門用語は、岩波生物学辞典によると、自食作用(Self Eating)として、その分野では以前から使われている用語だそうである。

私は最初、一時はやりのファジー洗濯機、ファジー扇風機の「ファジー(Fuzzy)」と混同してしまった、多くの人がそうであろうが。また、テレビなどマスコミがあえて、この言葉で話題にしているように見えてならない。

テレビに出ていらした大隈先生自身が地味な研究者で、純粋に自然の不思議さを強調していらっしゃるにもかかわらず、マスコミは、すぐ新薬に向けての新しいビジネスなどと煽りたてている。

言葉の使い方は難しい、いたずらにありがたそうにして、ブラックボックスとしたり、ネーミングによりビジネスとしたり、毎日聞かされていると洗脳されてしまう。専門の用語を覚えるのではなく、その内容をかみしめて自分のものにしたいと思う。