Ｊ研究会資料(2001.2.24)

ベータ２項分布の当てはめ

統計数理研究所　鈴木 義一郎

§１ 出生児性比と２項分布・ベータ２項分布の当てはめ

男児の出生数が女児の出生数に比べて、わずかに多いことはよく知られた統計的事実である。Fisher[１]では、ちょうど８人の子供をもつ家庭 53680組について調べた Geissler のデータについて、男児の生れる確率をｐとして、２項分布Ｂ(８,ｐ）に従うものかどうかを議論している。

まず、表１から直接、平均と分散の値を求めてみると

$$m = 221023/53680 = 4.117418$$
$$v = 1021023/53680 - 4.1174182 = 2.074167$$

のように算出される。一方、２項分布を想定したときには

$$m = np = 8\cdot p \ , \ v = npq = 8\cdot p(1-p)$$

であるから、ｐについての推定値

$$P = 4.117418/8 = 0.514677$$

を用いてｖ(分散)の値を求めてみると

$$v = 8\cdot P(1-P) = 8\times 0.514677\times 0.485323 = 1.998277$$

のように算出される。データから直接計算した分散の値より

$$2.074167 - 1.998277 = 0.075890(3.798\%)$$

だけ大きい。

ところで、この分散に対する標準誤差の平方は、次のように算出される。

$$\frac{2npq}{N-1} + \frac{npq(1-6pq)}{N} = \frac{2\cdot 8\cdot 0.514677\cdot 0.485323}{53679} + \frac{8\cdot 0.514677\cdot 0.485323(1-6\cdot 0.514677\cdot 0.485323)}{53680}$$
$$= \frac{2\cdot 1.998277}{53679} + \frac{1.998277\cdot(-0.498708)}{53680}$$
$$= 0.000055888 = (0.007476)2$$

つまり、データから直接計算した分散の値は、２項分布を仮定した

ときの分散より、その標準誤差の１０倍も大きいことが分かる。実際、表２の結果をみても、男だけとか女だけといった家庭の数値が２項分布を仮定した理論度数をかなり上まわっている。さらに、男児が５または３という家庭の数が、逆に少なすぎる傾向もみられる。変動が大きすぎる原因の１つとして、多生児の影響が考えられる。いま、８人の子供の中に、“１卵性”双生児が丁度１組だけいるという家庭(A)での男児の数をXとすると、XはB(６,p）に従う変数Yと、B(１,p）に従う変数Zの２倍の和で表すことができる。したがってXの分散は

$$V\{X\} = V\{Y\}+V\{2Z\} = 6pq+22\times pq = 10pq$$

となる。

そこで、Aタイプの家庭の比率をαとして、それ以外の家庭(１－α)での男児の数は２項分布B(８,p）に従うと考えてみよう。ここで２卵性双生児については、同性と異性になる可能性はほぼ等しいものと考えられるし、２組以上の１卵性双生児がいたり、双生児以外の多生児がいる家庭については、無視して考える。すると、全体としての男児の数の分散は

$$\alpha\cdot 10pq+(1-\alpha)\cdot 8pq = (1+\alpha/4)8pq$$

のように表すことができる。ところで、このデータの例のように 3.8 %もの分散の増加があるということから

$$\alpha/4 = 0.0382$$

つまり、α＝ 0.152 もの比率でAタイプの家庭が存在しなければならないことになる。

例えば、1000 回の出生のうち 12 回は双生児であり、さらに双生児のうち 1/4 は“１卵性”であるといったデータがある。従って 1000 回の出生のうち 3 回が１卵性　であるから、結局、125(＝1000/8)組のうち3組(＝ 2.4 %)の家庭が“１卵性双生児”をもつという勘定になる。結局、このデータでの分散の大きさは、１卵性双生児の影響だけで説明するには無理があるということになる。

ところで、分散が２項分布のものより大きい場合には、ベータ２項分布を当てはめるほうがよさそうである。この分布の確率密度関数や平均、分散などが

表２　２項分布(ｇ)とベータ２項分布(ｈ)の適合度

男児数 (x)	観測度数 (f)	期待度数 (g)	偏差 (d=f-g)	d2/g	期待度数 (h)	(f-h)2/h
0	215	165.22	+49.78	14.998	193.03	2.501

1	1485	1401.69	+83.31	4.952	1518.18	0.725
2	5331	5202.65	+128.35	3.166	5334.89	0.003
3	10649	11034.65	-385.65	13.478	10940.00	7.741
4	14959	14627.60	+331.40	7.508	14319.30	28.581
5	11929	12409.87	-480.87	18.633	12250.10	8.416
6	6678	6580.24	+97.76	1.452	6689.27	0.019
7	2092	1993.78	+98.22	4.839	2131.72	0.740
8	342	264.30	+77.70	22.843	303.54	4.873
SUM	53680	53680.00		91.869	53680.00	53.599

β(k;n,a,b) = nCkB(k+a,n-k+b)/B(a,b) , (k=0,1,2,……,n)

M = na/(a+b) , V = nab(a+b+n)/(a+b)2(a+b+1)

のように与えられる。与えられた平均と分散の値から、ａ，ｂを推定すると

a = 94.35 , b = 88.97

といった値が算出される。表２には、ｈ(ｋ)＝β(ｋ;８,94.35,88.97)を当てはめたときの期
待度数も示してある。当てはまりがかなり改良されていることが分かる。

観測値から算出される平均と分散を与えて、ベータ２項分布のａ，ｂを推定する関数を次のように定義する：

```
   est_ab=:4 :0
NB. estimate a,b of β(k;n,a,b) input n(left);M,V(right)
a=.(m*m*q=.1-p)-(p=.(m=.{.y.)%x.)*v=.{:y.
a,(q%p)*a=.a%v-m*q
)
   8 est_ab 4.117418 2.074167
94.3494 88.9682
```

ベータ２項分布β(ｋ；ｎ,ａ,ｂ)の確率関数は

β(k;n,a,b) = nCkB(k+a,n-k+b)/B(a,b) , (k=0,1,2,……,n)

$$= nCk \cdot \frac{\Gamma(a+k)}{\Gamma(a)} \cdot \frac{\Gamma(b+n-k)}{\Gamma(b)} \cdot \frac{\Gamma(a+b)}{\Gamma(a+b+n)}$$

のように表現できる。そこで、Γ(ａ＋ｋ)/Γ(ａ)というガンマ関数の比を(ガンマ関数の原始関数「！」を使わないで)算出する関数を次のように定義する。

```
   ratio=:4 :0
NB. calculate Γ(x.+y.)/Γ(y.) not using gammma function(!@<:)
if. x.=0 do. r=.1 erase. r=.a=.<:x.+y.
while. a>y. do. r=.x:r*a=.a-1 end.
end.
)
   bbfn=:4 :0
NB. beta-binomial probability function
r=.(( |.k)ratio"0{:y.)*(k=.i.x.+1)ratio"0{.y.
(k!x.)*r%x.ratio+/y.
)
   0.4": 8 bbfn 94.35 88.97
0.0036 0.0283 0.0994 0.2038 0.2668 0.2282 0.1246 0.0397 0.0057
   0.2": 53580*8 bbfn 94.35 88.97
```

192.67 1515.35 5324.95 10919.63 14292.60 12227.26 6676.81 2127.75 302.98

§２ ベータ２項分布の変形

大相撲の幕内力士は、毎場所１５人の相手と対戦する。表３のＮ(ｘ)の行に示したのは、平成６年１月から平成７年７月までの１０場所について、途中休場などしなかった力士の勝星数の分布である。どんな相手にも勝率ｐで戦える力士の勝星数の確率は、２項分布Ｂ(15,ｐ）で与えられる。Ａ(ｘ)の行には、平均勝率Ｐの値が０．５という場合の数値を示してみた(観測データと合わせるために、２項分布の確率に３８６という数値を掛けて丸めたもの)。

実際のデータと比べてみて、かなりずれている。そもそも、幕内の全ての力士の勝率Ｐが同じ０．５であるとは考えられない。優勝を争うような力士のｐは０．７以上かもしれないし、反対に不調な下位力士のｐは０．３程度かもしれない。

そこで、ｐの値を“混ぜこぜ”にした「複合２項分布」というモデルを考えてやればよい。特に、２項分布のパラメータｐをベータ分布β(ｔ;ａ,ｂ)で加重平均したものがベータ２項分布Ｂ(ｋ;15,ａ,ｂ）である。与えられたデータの平均と分散の値から、ａ，ｂを推定すると

$$a = 9.59 \ , \ b = 9.47$$

と算出される。そこで、

$$h(k) = h(k;15,9.59,9.47)$$

というモデルを当てはめた結果が、表３のＢ(ｘ)の行に示してある。このモデルにしたら、大分、実際のデータに近くなったが、まだ７勝、８勝の辺りではかなりの“ずれ”が見られる。

ところで８勝と７勝とでは、次の場所の番付の上下に雲泥の差があり、勝越しをかけた一番は是が非でも勝ちたいと奮起せざるを得ない。したがって、八百長の有無はともかくも、８勝力士の多くなるのは当然すぎる結果かもしれない。

表４に、１４日で７勝７敗だった力士の数と、千秋楽に勝って８勝を達成した力士の数を示してみた。１０場所全体でみると、７割５分以上の高率を示している。

そこで、１４日目までの勝星数(Ｙ)にはベータ２項分布ｇ(ｙ)＝Ｂ(ｙ;14,ａ,ｂ）を当てはめる。そして、千秋楽での勝星数(Ｚ)の分布としては、１４日目でＹの値が７である力士の勝率には、“奮起度”がプラスされるものと想定して、

$$\Pr\{Z=1\} = p$$
$$\Pr\{Z=1\ Y=7\} = p+\alpha$$
$$\Pr\{Z=1\ Y\neq 7\} = p-\alpha\delta\ ,\ \delta = g(7)/\{1-g(7)\}$$

といったように与える。このαというパラメータは、ＺのＹに対する従属性の度合いを示すもので、ある種の“dependent outlier”と解釈できる。

次に、１５日間での勝星数(Ｘ＝Ｙ＋Ｚ)の分布をｈ(ｘ)とすると

$$h(7) = \mu(7)+\alpha\delta\{d(7)-1\}\ ,\ h(8) = \mu(8)+\alpha\delta\{d(8)+1\}$$
$$h(x) = \mu(x)+\alpha\delta d(x)\ \ (x\neq 7,8)$$

のように表現できる。ここで

$$\mu(x) = \{g(x)+g(x-1)\}/2\ ,\ d(x) = \{g(x)-g(x-1)\}\ ,[g(-1)=g(15)=0]$$

である。さらにαというパラメータも

$$\alpha = \{h(8)-h(7)\}/\delta\{2[1-g(7)]+g(6)+g(8)\}$$
$$= \{h(8)-h(7)\}/g(7)\{2+[g(6)+g(8)]/[1-g(7)]$$

のように表される。Ｙの期待値を７と想定すると

$$E\{YZ\} = \Sigma_y \Pr\{Z=1,Y=y\} = \Sigma_y g(y)\Pr\{Z=1\ Y=y\}$$
$$= (p-\alpha\delta)\Sigma_y g(y)+(p+\alpha)7g(7) = p\Sigma_y g(y)+\alpha\delta[7-\Sigma_y g(y)]$$
$$= E\{Z\}E\{Y\}+\alpha\delta[7-E\{Y\}] = E\{Z\}E\{Y\}$$

となることから、ＹとＺは無相関になる。したがって

$$V\{Y\} = V\{X\}-V\{Z\} = V\{X\}-p(1-p)$$

という関係より、ｐ＝１/２と想定してＹの分散も推定できるから、ｇ(ｙ）の分布が特定できることになり、結局、αの値も推定できる([２]参照)。

このように、“dependent outlier”を加えた変形モデルを用いて導出した結果が表３に示したＣ(ｘ)の行の数値である。Ｂ(ｘ)と比べると、７勝、８勝の辺で改良されている。この数値を、実際のデータに重ね合わせてみたのが図１で、ほどほどには適合していることが分かる。

上述のような変形ベータ２項分布のモデルを当てはめる関数を次のように定義する。ここで出力結果の最初の値は、従属性のパラメータであるαの推定値で、以下が当てはめた確率関数の値である。

```
   sumou=:3 :0
NB.modified beta-binomial model for OZUMOU
v=.+/f**:x -+/(x=.i.16)*f=.(%+/)y.
b=.m%1-m=.7{g=.14 bbfn 14 est_ab 7,v-0.25
c=.b*a=.( -/8 7{f)%m*2+(+/6 8{g)%1-m
h=.( -:g1+g2)+c*d=.(g2=.g,0)-g1=.0,g
a,h+t,(( -c),c),t=.7$0
)
   F1=.sumou f1=.0 0 6 13 24 38 53 37 102 54 10 17 17 7 5 3
   ({.F1);0.1":386*}.F1
```

0.4715775	0 2 6 15 40 51 25 90 49 37 24 13 5 2 0

```
   +/ 0 2 6 15 40 51 25 90 49 37 24 13 5 2 0
385
```

【整数に丸めてしまったので、合計が 386 にならない】

```
   F2=.sumou f2=.0 2 8 25 37 80 71 48 150 93 46 23 13 19 9 6
   ({.F2);0.1":626*}.F2
```

0.474372	1 4 12 25 44 64 81 39 141 78 61 41 23 10 3 0

また図２には、昭和５１年から５３年までの１８場所についての勝星数の分布と、それに当てはめたモデルの結果も示してみた([２])。この時代は、輪島・北の湖そして三重海・二代目若の花が横綱で、それに先代の貴ノ花や旭国といった個性派大関が活躍していた時代である。やはり７勝の“落ちこみ”は、現在のパターンと同じような傾向を見せている。しかし、１０勝近辺での“落ちこみ”は今ほど顕著でない。つまり、優勝にからむ力士が大勢いる限り、１０勝力士が極端に少なくなることはなく、したがってここで示したモデルでほぼ説明がつくということにもなる。

【参考文献】

[１] Fisher,R.A.(1970)：研究者のための統計的方法(遠藤・鍋谷訳 森北出版)
(Statistical Merhods for Research Workers)

[２] 鈴木義一郎(1979)：ある種の Dependent Outlier を含む確率模型
統計数理研究所彙報 26-1,23-32.