

正の二項分布に関する統計的推測

岩崎 学^{*1}, 吉田 清隆^{*2}

Statistical Inference for the Positive Binomial Distribution

Manabu IWASAKI^{*1} and Kiyotaka YOSHIDA^{*2}

ABSTRACT : The binomial distribution formed by omission of zero class is called the positive binomial distribution. We discuss estimation problems for this distribution. The ordinary maximum likelihood (ML) procedure requires some numerical algorithm to calculate the estimate. A simple iterative procedure for the ML estimate is given in this paper. When observed number is one out of n independent trials, the maximum likelihood estimate tends to zero, which causes negative bias of the ML estimator. A modified new estimator that requires no iteration is proposed here. It is pointed out that the new estimator has less bias than the ML estimator for a wide range of parameter values. It is also pointed out that the standard error of our new estimator is less than that of the ML estimator for almost all parameter values, and hence the new estimator is more efficient and useful than the ML estimator.

Keywords : zero-truncated distribution, maximum likelihood, bias, functional iteration

(Received September 1, 2007)

1. はじめに

ある事象の生起確率が θ である試行を独立に n 回繰り返すときの事象の生起回数を表わす確率変数を X としたとき, X はパラメータ n, θ の二項分布 (binomial distribution) に従い, このことを $X \sim B(n, \theta)$ と書く。二項分布は最も基本的な離散型分布であり, 初等的な教科書を始め, 多くの書物でその性質が論じられている (詳細は Johnson, Kemp and Kotz (2005), 竹内・藤野 (1981) などを参照されたい)。ところが実際のデータ解析では, X のすべての値が必ずしも得られるとは限らない場合がある。特に, X が 0 のときは報告がなく, X が 1 以上のときのみ事象の生起回数のデータが得られることがある。このときの確率分布を正の二項分布 (positive binomial distribution) もしくはゼロトランケートされた二項分布 (zero-truncated binomial distribution) といい (Johnson, *et al.* (2005), Section 3.11), ここではそれを $PB(n, \theta)$ と書く。

本論では, 第 2 節で $PB(n, \theta)$ の基本的な性質を概観した後, パラメータ θ の最尤推定法を与える。最尤推定値の計算には反復計算が必要となり, さらに生起回数が 1 のとき推定値が 0 になってしまうという不都合が生じることを注意する。このことから最尤推定量は負の偏りを持つことが示される。その不都合をある程度回避するため, 第 3 節では反復計算を用いない簡便な推定法を提案し, その推定量はパラメータの多くの範囲で最尤推定量よりも偏りが小さく, さらに標準誤差も小さいという統計的によい性質を持つことを示す。そして最後の第 4 節で簡単な議論を行なう。

2. 正の二項分布の性質と最尤推定

パラメータ n, θ の正の二項分布 $PB(n, \theta)$ に従う確率変数を Y とする (以降, X を通常二項分布 $B(n, \theta)$ に従う確率変数とする)。 $B(n, \theta)$ は取り得る値が 1 以上の条件付き二項分布でもあるので, その確率関数 $p(y)$ は, $\Pr(X=0) = (1-\theta)^n$ より

^{*1}: 情報科学科教授 (iwasaki@st.seikei.ac.jp)

^{*2}: 情報科学科助手

$$p(y) = \Pr(Y = y) = \frac{{}_n C_y \theta^y (1-\theta)^{n-y}}{1-(1-\theta)^n}, \quad y=1, 2, \dots, n \quad (1)$$

となる。このとき、相続く y に関する確率の比は

$$\frac{p(y+1)}{p(y)} = \frac{n-y}{y+1} \cdot \frac{\theta}{1-\theta} \quad (2)$$

となる。この形は二項分布のときの確率の比と同じである（(1)における分母が比によりキャンセルアウトするためである）。これより、 $PB(n, \theta)$ の最頻値（モード）は、二項分布のときと同様、 $(n+1)\theta$ を超えない最大の整数で与えられ、 $(n+1)\theta$ が整数となる場合には、 $y=(n+1)\theta-1$ と $y=(n+1)\theta$ で確率は同じ最大値を取ることが示される（竹内・藤野（1981）参照）。特に (2) で $y=1$ とした不等式

$$\frac{p(2)}{p(1)} = \frac{n-1}{2} \cdot \frac{\theta}{1-\theta} \leq 1$$

より関係式 $\theta \leq 2/(n+1)$ を得る。このときすべての y で $p(y) \geq p(y+1)$ となることから、 $\theta \leq 2/(n+1)$ が、確率 $p(y)$ が y に関し単調減少となるための条件となる（二項分布 $B(n, \theta)$ のときの確率が単調減少であるための条件は $\theta \leq 1/(n+1)$ である）。

$PB(n, \theta)$ に従う確率変数 Y の期待値と分散はそれぞれ

$$E[Y] = \frac{n\theta}{1-(1-\theta)^n} \quad (3)$$

および

$$V[Y] = \frac{n\theta(1-\theta)}{1-(1-\theta)^n} - \frac{n^2\theta^2(1-\theta)^n}{\{1-(1-\theta)^n\}^2}$$

で与えられる（Johnson, *et al.* (2005), Section 3.1 参照）。 Y は値 0 を取らないので、 $E[X] \leq E[Y]$ である（等号は $\theta=1$ のときのみ）。これは、(1) で分母が $1-(1-\theta)^n \leq 1$ であることから示される。また、 $E[Y]$ は θ の単調増加関数であり、 $\theta \rightarrow 0$ の極限は、ロピタルの定理より

$$\lim_{\theta \rightarrow 0} \frac{n\theta}{1-(1-\theta)^n} = \lim_{\theta \rightarrow 0} \frac{(n\theta)'}{\{1-(1-\theta)^n\}'} = \lim_{\theta \rightarrow 0} \frac{n}{n(1-\theta)^{n-1}} = 1$$

となる。よって、 $0 < \theta \leq 1$ に対し $1 < E[Y]$ であることが分かる。分散については、 Y は値 0 を取らないことから $V[Y] \leq V[X]$ となる。 θ が 1 に近いときには $B(n, \theta)$ で $X=0$ となる確率は小さく、 $B(n, \theta)$ と $PB(n, \theta)$ との違いがほとんどなくなるので、 θ が 1 に近づくにつれ $PB(n, \theta)$ の期待値も分散も $B(n, \theta)$ との差は小さくなる。

次に、 n 回の独立な試行で事象の生起が y 回得られたときの θ の最尤推定値 $\hat{\theta}$ を求める。 $Y=y$ が与えられたときの θ の対数尤度関数は (1) より

$$l(\theta) = \log L(\theta) = \log \left\{ \frac{{}_n C_y \theta^y (1-\theta)^{n-y}}{1-(1-\theta)^n} \right\} \\ = -\log \{1-(1-\theta)^n\} + \log {}_n C_y + y \log \theta + (n-y) \log(1-\theta)$$

であるので、これを θ で微分して 0 と置き、尤度方程式

$$l'(\theta) = -\frac{n(1-\theta)^{n-1}}{1-(1-\theta)^n} + \frac{y}{\theta} - \frac{n-y}{1-\theta} \\ = \frac{y\{1-(1-\theta)^n\} - n\theta}{\theta(1-\theta)\{1-(1-\theta)^n\}} = 0$$

を得る。尤度方程式の分母は $0 < \theta < 1$ で正であるので、分子を 0 と置くことにより、関係式

$$n\theta = y\{1-(1-\theta)^n\} \quad (4)$$

が導かれ、これを満足する θ が求める最尤推定値となる。(4) を変形すると

$$y = \frac{n\theta}{1-(1-\theta)^n} \quad (5)$$

となる。(5) は Y の期待値の式 (3) の左辺の期待値を観測値 y で置き換えたものになっている。すなわち、この場合の最尤推定値はモーメント法による推定値でもある。関係式 (4) より

$$\theta = \frac{y}{n} \{1-(1-\theta)^n\} \quad (6)$$

となるので、最尤推定値の具体的な値は、 $\theta^{(0)}$ を初期値とした反復計算

$$\theta^{(i+1)} = \frac{y}{n} \{1-(1-\theta^{(i)})^n\} \quad (7)$$

の収束により求められる。初期値 $\theta^{(0)}$ としては y/n を取ればよいであろう。計算は EXCEL で簡単に実行できる（例 2.1 参照）。経験的に (7) の反復は $y \geq 2$ で素早く収束する。(7) の反復法は単純関数反復法と呼ばれ（Lange (1999), 岩崎 (2004) などを参照）、その収束条件は岩崎 (2001) で与えられている。 $y=1$ では (4) を満足する θ は 0 であることが示されるが、 $\theta=0$ では事象が生起しないのであるから推定値にはなり得ず、この場合、最尤推定値は存在しないことになる（限りなく 0 に近い値ともいえる）。このため、最尤推定量（観測値 y の代わりに確率変数 Y を用いたもの）は負の偏りを持つ（具体的な数値は次節参照）。

例 2.1

$n=10$ とし、 $y=3$ が観測されたとすると、初期値 $\theta^{(0)} = 3/10 = 0.3$ から出発し、表 1 のような推移により、最尤推定値 $\hat{\theta} = 0.290272$ に収束する。その他の数値設定でも概ね 10 回以内の反復で十分な精度での収束が得られる。

表 1 $n = 10, y = 3$ のときの最尤推定法の反復の推移

ITE	Theta
0	0.300000
1	0.291526
2	0.290442
3	0.290295
4	0.290275
5	0.290272
6	0.290272

より一般に, $PB(n, \theta)$ に従う独立な観測値が $y_1, \dots, y_N$ と N 個得られた場合には, それらの標本平均を $\bar{y} = \frac{1}{N} \sum_{i=1}^N y_i$ とし, (5) 同様

$$\bar{y} = \frac{n\theta}{1 - (1 - \theta)^n}$$

を満たす θ として最尤推定値が得られる。実際の計算は (7) 同様,

$$\theta^{(r+1)} = \frac{\bar{y}}{n} \{1 - (1 - \theta^{(r)})^n\}$$

の反復を実行すればよい。Mantel (1951) は反復法を用いない方法として f_1 を $y=1$ となった度数として

$$\theta^* = \frac{\bar{y} - f_1 / N}{n - f_1 / N}$$

なる推定値を提案した。Thomas and Gart (1971) は, 最尤推定値 $\hat{\theta}$ と Mantel の推定値 θ^* を比較し, θ^* のほうが比較的正確で偏りが少ないことを示している。Mantel の時代 (1950 年代) には計算環境が貧弱で, 現在では簡単に実行できる反復計算による推定値も容易には得られなかったことから簡便な推定法が提案されたのであろうが, 簡便法のほうがよりよい推定値を与えるという例でもある。

3. 新しい推定法

前節で述べた最尤推定法は, EXCEL による反復法で容易に推定値は得られるが $y = 1$ のときには推定値が 0 に収束してしまうという欠点があり, その結果推定量は負の偏りを持つ。ここでは, 反復法によらない簡便な推定法を提案し, その特性を吟味する。

尤度方程式から導かれる関係式 (4) における右辺の θ は, 特に θ があまり小さくない場合には二項分布における推定値 y/n により推定される。これより

$$n\theta = y \left\{ 1 - \left(1 - \frac{y}{n} \right)^n \right\}$$

となるが, n が十分大きいときは $(1 - y/n)^n \approx e^{-y}$ であるので, 結局, 推定値

$$\tilde{\theta} = \frac{y}{n} (1 - e^{-y}) \quad (8)$$

が得られる。(8) がここで提案する推定値であり, 以下その性質を調べる。 $y = 1$ のときは $1 - e^{-1} \approx 0.632$ であるので, $\tilde{\theta} = 0.632/n$ と推定される。なお, (8) で観測値 y の代わりに確率変数 Y を代入した推定量も同じ記号 $\tilde{\theta}$ で表わす。

例 3.1

$n = 10$ のときの各観測値 $y = 1, \dots, 10$ に対する最尤推定値 $\hat{\theta}$ と (8) の推定値 $\tilde{\theta}$ の実際の値は表 2 のようである (便宜上 $y=1$ のときの最尤推定値を 0 とした)。 θ がある程度大きいときにはほぼ同じ推定値となることが分る。

表 2 $n = 10$ のときの最尤推定値と新推定値の値

y	MLE	NEW
1	0.000000	0.063212
2	0.168343	0.172933
3	0.290272	0.285064
4	0.397478	0.392674
5	0.499507	0.496631
6	0.599937	0.598513
7	0.699996	0.699362
8	0.800000	0.799732
9	0.900000	0.899889
10	1.000000	0.999955

最尤推定量と (8) の推定量の期待値と標準誤差は, 数学的に簡潔には表現できないが, 数値的には容易に求めることができる。それらを $n = 10, 20, 30, 40, 50$ および $\theta = 0.05 (0.05) 0.5$ の範囲で計算した結果が表 3 および表 4 である。表 3 より, 最尤推定量 $\hat{\theta}$ は常に負の偏りを持つことが分かる。それに対し, (8) の推定量 $\tilde{\theta}$ は, θ が小さいときを除いて偏りは概ね $\hat{\theta}$ よりも小さい。また, 表 4 より, ほぼすべてのパラメータ値に対し $\tilde{\theta}$ のほうが $\hat{\theta}$ よりも標準誤差が小さい。すなわち, ここで提案した推定量 $\tilde{\theta}$ は, 反復計算が不要なこと, $y = 1$ であっても推定値が得られることの利点の他, その統計的性質も最尤推定量よりも優れているといえよう。

4. おわりに

事象の生起回数が 1 以上のみデータが観測される正の二項分布 (ゼロトランケートされた二項分布) におけるパラメータの推定法を議論した。正の二項分布は, 一般に生起回数がある範囲でのみ観測されるトランケートされた二項分布 (truncated binomial distribution) の特別な場合であるが, 一般のトランケートされた二項分布は, Johnson, et al. (2005) Section 3.11 でも触れられているように, ゼロトランケートの場合を除き, あまり実際の場面では遭遇しない。

表3 最尤推定量と提案する推定量の期待値

Theta	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
n = 10										
E[MLE]	0.0399	0.0837	0.1309	0.1808	0.2326	0.2854	0.3386	0.3917	0.4443	0.4963
E[NEW]	0.0903	0.1222	0.1587	0.1996	0.2442	0.2916	0.3410	0.3917	0.4430	0.4945
n = 20										
E[MLE]	0.0425	0.0908	0.1423	0.1948	0.2469	0.2984	0.3492	0.3997	0.4499	0.4999
E[NEW]	0.0624	0.1017	0.1474	0.1966	0.2473	0.2982	0.3490	0.3994	0.4497	0.4999
n = 30										
E[MLE]	0.0442	0.0948	0.1471	0.1987	0.2495	0.2998	0.3499	0.4000	0.4500	0.5000
E[NEW]	0.0544	0.0986	0.1481	0.1988	0.2494	0.2998	0.3499	0.4000	0.4500	0.5000
n = 40										
E[MLE]	0.0455	0.0972	0.1489	0.1997	0.2499	0.3000	0.3500	0.4000	0.4500	0.5000
E[NEW]	0.0513	0.0986	0.1491	0.1996	0.2499	0.3000	0.3500	0.4000	0.4500	0.5000
n = 50										
E[MLE]	0.0466	0.0985	0.1496	0.1999	0.2500	0.3000	0.3500	0.4000	0.4500	0.5000
E[NEW]	0.0500	0.0990	0.1496	0.1999	0.2500	0.3000	0.3500	0.4000	0.4500	0.5000

表4 最尤推定量と提案する推定量の標準誤差

Theta	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
n = 10										
SE[MLE]	0.0794	0.1105	0.1318	0.1469	0.1571	0.1633	0.1665	0.1674	0.1666	0.1646
SE[NEW]	0.0555	0.0825	0.1043	0.1225	0.1371	0.1483	0.1562	0.1611	0.1634	0.1635
n = 20										
SE[MLE]	0.0561	0.0766	0.0889	0.0964	0.1015	0.1052	0.1081	0.1102	0.1115	0.1119
SE[NEW]	0.0430	0.0650	0.0810	0.0921	0.0995	0.1046	0.1081	0.1104	0.1117	0.1120
n = 30										
SE[MLE]	0.0455	0.0607	0.0692	0.0751	0.0800	0.0840	0.0872	0.0895	0.0908	0.0913
SE[NEW]	0.0371	0.0554	0.0670	0.0745	0.0799	0.0841	0.0872	0.0895	0.0909	0.0913
n = 40										
SE[MLE]	0.0391	0.0510	0.0581	0.0638	0.0686	0.0725	0.0754	0.0775	0.0787	0.0791
SE[NEW]	0.0334	0.0486	0.0575	0.0638	0.0687	0.0725	0.0754	0.0775	0.0787	0.0791
n = 50										
SE[MLE]	0.0345	0.0445	0.0512	0.0567	0.0613	0.0648	0.0675	0.0693	0.0704	0.0707
SE[NEW]	0.0306	0.0434	0.0510	0.0567	0.0613	0.0648	0.0675	0.0693	0.0704	0.0707

本論では、正の二項分布において、一般に用いられる最尤法による推定は、反復計算を必要とし、かつ $y=1$ の場合に推定値が得られず、しかも負の偏りを持つという欠点があることから、最尤推定法に代わる新たな推定法を提案し、その統計的な性質を議論した。新しい推定法は、簡便であるが、上述の最尤推定法の持つ欠点を回避し、偏りおよび標準誤差の観点から最尤法よりも優れていることが示された。

ゼロトランケーションは、二項分布以外にも多くの実際問題で重要な役割を果たし、特にポアソン分布に関しては多くの理論的な研究成果がある (Johnson, et al. (2005) Section 4.10 を参照)。また、岩崎・廉 (2007) ではゼロトランケートされたポアソン分布およびガンマポアソン分布に関するいくつかの興味深い実際例が示されている。

謝 辞

本論で扱った問題は、著者の一人 (岩崎) が成蹊大学大学院工学研究科および東京大学大学院教育学研究科における授業で取り上げたものでもある。有益なコメントを頂戴した受講生諸君、特に阿部大輔、楠田貴至の両氏に感謝する。本論の内容は両氏のコメントに触発されたものである。

参考文献

- 岩崎 学 (2001) 非線型方程式の単純反復解法とその統計的応用. 応用統計学, **30**, 107-118.
- 岩崎 学 (2004) 統計的データ解析のための数値計算法入門. 朝倉書店。
- 岩崎 学・廉 民善 (2007) ゼロトランケーションのあるカウントデータの解析. 行動計量学, **34**, 91-100.
- Johnson, N. L., Kemp, A. W. and Kotz, S. (2005) *Discrete Univariate Distributions, Third Edition*. John Wiley & Sons, New York.
- Lange, K. (1999) *Numerical Analysis for Statisticians*. Springer, New York.
- Mantel, N. (1951) Evaluation of a class of diagnostic tests. *Biometrics*, **3**, 240-246.
- 竹内 啓・藤野和健 (1981) 2項分布とポアソン分布. 東京大学出版会。
- Thomas, H. A. and Gart, J. J. (1971) Small sample performance of some estimators of the truncated binomial distribution. *Journal of the American Statistical Association*, **66**, 169-177.