日水研報告, (39): 9—17, 1989.
Bull. Japan Sea Reg. Fish. Res. Lab., (39): 9—17, 1989.

An Interval Estimation for Extraction using Bayesian Statistics

Tatsuro AKAMINE[1)]

Abstract

The statistical model for extraction is a binomial distribution. The conventional method for employing this binomial model is based on approximation to a normal distribution. The Bayesian statistical method, which assumes that the prior distribution of parameter is uniform, is preferable to the conventional method, and two theorems demonstrate that this model corresponds well with the conventional method. Furthermore, this model is simpler to understand and easier to calculate by micro-computer than the conventional method.

Key words Bayesian statistics, extraction, binomial distribution, normal distribution, uniform distribution

Introduction

Extraction is the simplest method to estimate total number. Let n: total number of individuals to estimate, r: number of extracted individuals and p: extract ratio (Fig. 1). Although the binomial distribution gives r when n and p are fixed, in this problem, we require n when r and p are fixed. AKAMINE (1981) approached this problem graphically, but he depended on the conventional method.

The prior distribution of n, which is a uniform distribution, provides the simplest method to estimate a confidence interval for n. First, two important theorems for binomial distributions will be proven. Next, the Bayesian statistical method according to the above theorems will be demonstrated. This method is more logical and simpler than, and corresponds with, the conventional method.

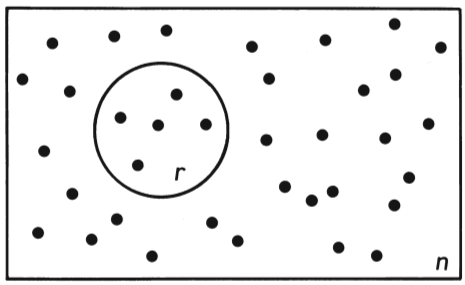

Fig. 1. The image of extraction ($p=r/n$).

Accepted: October 20, 1988. Contribution A *No.* 453 from the Japan Sea Regional Fisheries Research Laboratory.

1) Japan Sea Regional Fisheries Research Laboratory, Suido-cho, Niigata 951, Japan.
(〒951 新潟市水道町1丁目5939-22 日本海区水産研究所)

Theorems for a binomial distribution

1. The binomial theorem

The binomial distribution is expressed as follows:

$$P(n, r) = \binom{n}{r} p^r q^{n-r}, \quad p+q=1. \tag{2.1}$$

Where the binomial coefficient (number of combinations):

$$\binom{n}{r} = \frac{n!}{r!(n-r)!} = \frac{n(n-1)\cdots\cdots(n-r+1)}{r(r-1)\cdots\cdots 1}. \tag{2.2}$$

The binomial theorem is as follows:

$$\sum_{r=0}^{n} P(n, r) = 1. \tag{2.3}$$

The proof of this theorem is quite simple. The binomial expansion:

$$\begin{aligned}(p+q)^n &= \binom{n}{0} p^0 q^n + \binom{n}{1} p^1 q^{n-1} + \cdots\cdots + \binom{n}{n} p^n q^0 \\ &= 1\end{aligned} \tag{2.4}$$

is the proof as well as the definition of the binomial distribution.

2. Theorem 1

The arrangement of $P(n, r)$ is shown in Table 1. The binomial theorem gives the total sum of the row. The following theorem gives the total sum of the column.

[Theorem 1]

For any r, the following expression holds.

$$S_r = \sum_{n=r}^{\infty} P(n, r) = \frac{1}{p}. \tag{2.5}$$

[Proof]

First, the convergence of S_r is demonstrated.

$$\frac{P(n+1, r)}{P(n, r)} = q\,\frac{n+1}{n-r+1} \quad (n=\infty) \longrightarrow q. \tag{2.6}$$

Then, the ratio test certifies the convergence of this series.

Next, the convergence value is required. From (2.6)

$$P(n+1, r) = q\,\frac{n+1}{n-r+1} P(n, r). \tag{2.6'}$$

From recursion of this equation,

$$\begin{aligned} S_r &= P(r, r)\left\{1 + \frac{r+1}{1} q\left(1 + \frac{r+2}{2} q\left(1 + \cdots\cdots\right)\right)\right\} \\ &= p^r\left\{1 + \frac{r+1}{1!} q + \frac{(r+1)(r+2)}{2!} q^2 + \cdots\cdots\right\} \\ &= p^r \sum_{i=0}^{\infty} \binom{r+i}{i} q^i. \end{aligned} \tag{2.7}$$

On the other hand, the binomial coefficient can be expanded to include negative

numbers as follows:

$$\binom{-n}{r}=\frac{(-n)(-n-1)\cdots\cdots(-n-r+1)}{r(r-1)\cdots\cdots 1}$$

$$=(-1)^r\binom{n+r-1}{r},\ n>0. \tag{2.8}$$

Therefore, the binomial theorem can be expanded to include negative numbers as follows:

$$(1-q)^{-r-1}=\binom{-r-1}{0}(-q)^0+\binom{-r-1}{1}(-q)^1+\cdots\cdots$$

$$=\binom{r}{0}+\binom{r+1}{1}q+\binom{r+2}{2}q^2+\cdots\cdots$$

$$=\sum_{i=0}^{\infty}\binom{r+i}{i}q^i. \tag{2.9}$$

Then, from (2.7) and (2.9) the result is:

$$S_r=p^r(1-q)^{-r-1}=1/p. \qquad \text{[Q. E. D.]}$$

[Another proof]

S_0 is a geometrical series by ratio q. Then, S_0 is simply given as follows:

$$S_0=1/(1-q)=1/p. \tag{2.10}$$

Binomial coefficients have the following formula:

$$\binom{n}{r}=\binom{n-1}{r-1}+\binom{n-1}{r}. \tag{2.11}$$

The following formula for $P(n, r)$ is obtained from this formula.

$$P(n, r)=pP(n-1, r-1)+qP(n-1, r). \tag{2.12}$$

From this equation,

$$P(n-1,\ r-1)=\frac{1}{p}P(n, r)-\frac{q}{p}P(n-1, r). \tag{2.12'}$$

Where,

$$\frac{1}{p}-\frac{q}{p}=1. \tag{2.13}$$

Therefore, Fig. 2 illustrates the following equation:

$$S_{r-1}=S_r \tag{2.14}$$

(2.5) is proved by (2.10) and (2.14).

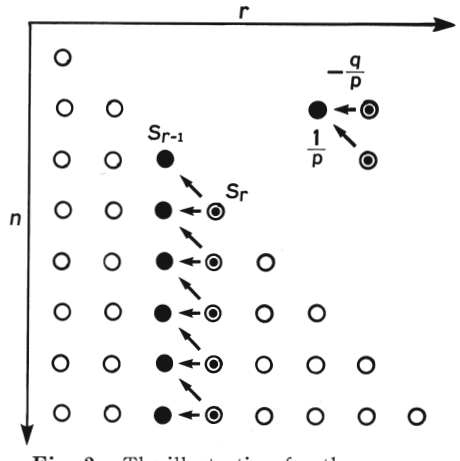

Fig. 2. The illustration for the arrangement of binomial distributions to prove the Theorem 1.

3. A BASIC program

For (2.6') let $f(n)$ be the following:

$$f(n)=q\frac{n+1}{n-r+1}$$

$$=q\left(1+\frac{r}{n-r+1}\right). \tag{2.15}$$

Then, $f(n)$ is monotone decreasing. In addition,

$$f\binom{r}{p}-1)=q\frac{r/p}{r/p-r}=\frac{q}{1-p}=1. \tag{2.16}$$

Let $n_0=\max\{n\,|\,n\leqq r/p\}$, where n is a natural number. The result from (2.16) is:

$$\max P(n)=P(n_0). \tag{2.17}$$

In addition to, when $n_0=r/p$,

$$\max P(n)=P(n_0)=P(n_0-1). \tag{2.17'}$$

Therefore, point estimation of n is n_0.

From (2.6')

$$P(n-1,\ r)=\frac{1}{q}\frac{n-r}{n}P(n,\ r). \tag{2.6''}$$

An example of BASIC programs is shown in 'Program 1.' This program calculates $P(n,\ r)$ by (2.6') and (2.6'').

4. Theorem 2

In Table 1, the percent points for rows correspond with that of columns. Theorem 2 gives this relation.

[Theorem 2]

For any n, r, the following equation holds.

$$p\sum_{i=r}^{n-1}P(i,\ r)=\sum_{j=r+1}^{n}P(n,\ j) \tag{2.18}$$

[Proof]

Fig. 3–a gives this proof. Let

$$C_k=\sum_{i=r}^{r+k-1}P(i,\ r)=\sum_{i=1}^{k}P(r+i-1,\ r)=\sum_{i=1}^{k}c_i, \tag{2.19}$$

$$R_k=\sum_{j=r+1}^{r+k}P(r+k,\ j)=\sum_{j=1}^{k}P(r+k,\ r+j)=\sum_{j=1}^{k}r_j. \tag{2.20}$$

The following equation is obvious.

Table 1. The arrangement of binomial coefficients :

$$P(n,\ r)=\binom{n}{r}p^rq^{n-r}.$$

		r							
		0	1	2	3	4	5	6	7
n	0	1							
	1	q	p						
	2	q^2	$2pq$	p^2					
	3	q^3	$3pq^2$	$3p^2q$	p^3				
	4	q^4	$4pq^3$	$6p^2q^2$	$4p^3q$	p^4			
	5	q^5	$5pq^4$	$10p^2q^3$	$10p^3q^2$	$5p^4q$	p^5		
	6	q^6	$6pq^5$	$15p^2q^4$	$20p^3q^3$	$15p^4q^2$	$6p^5q$	p^6	
	7	q^7	$7pq^6$	$21p^2q^5$	$35p^3q^4$	$35p^4q^3$	$21p^5q^2$	$7p^6q$	p^7

$$pC_1=R_1\,. \tag{2.21}$$

If $pC_k=R_k$,

From (2.19) $C_{k+1}=C_k+c_{k+1}$,

From Fig. 3–a $R_{k+1}=R_k+pc_{k+1}$ $(\because\ p+q=1)$.

Then, $pC_{k+1}=R_{k+1}$.

Therefore, Theorem 2 is proved. [Q.E.D.]

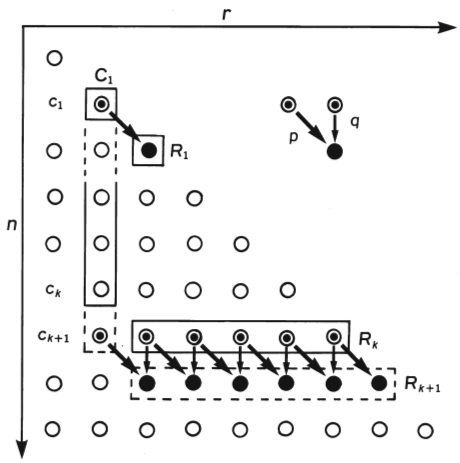

Fig. 3–a. The illustration for the arrangement of binomial distributions to prove the Theorem 2.

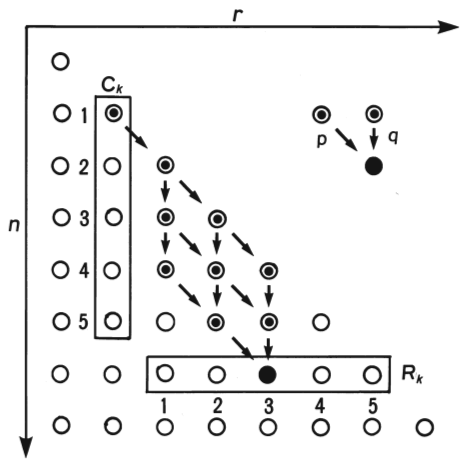

Fig. 3–b. The illustration for the arrangement of binomal distributions to prove the Theorem 2 by the other way.

[Another proof]

Fig. 3–b gives this proof. The contribution of c_1 to r_3 is

$$p\binom{4}{2}p^2q^2\,.$$

Then, the total contribution of c_1 to R is

$$p\left\{\binom{4}{0}q^4+\binom{4}{1}pq^3+\cdots\cdots+\binom{4}{4}p^4\right\}=p(p+q)^4=p\,.$$

Similarly, the contribution of the other c_i to R is p, then

$$pC_k=R_k\,. \qquad \text{[Q.E.D.]}$$

Estimation for a confidence interval

1. Bayesian statistical method

Bayesian statistical method is as follows: Let θ: parameter to estimate, t: data, $P(\theta, t)$: probability of data for each θ. $P^\circ(\theta)$: prior distribution of θ. Where,

$$\sum P^\circ(\theta)=1\,. \tag{3.1}$$

Let $P^*(\theta)$: posterior distribution of θ. Then

$$P^*(\theta)=\frac{P^\circ(\theta)P(\theta,t)}{\sum P^\circ(\theta)P(\theta,t)}. \tag{3.2}$$

Where,

$$\sum P^*(\theta)=1. \tag{3.3}$$

In particular, the prior distribution of θ is a uniform distribution:

$$P^\circ(\theta)\equiv\varepsilon=\text{const}. \tag{3.4}$$

The posterior distribution becomes as follows:

$$P^*(\theta)=\frac{P(\theta,t)}{\sum P(\theta,t)}. \tag{3.5}$$

In this case, the prior distribution of n is the uniform distribution from r to ∞. It is better to regard this as follows: The prior distribution of n is the uniform distribution from r to n_1. Then,

$$P^\circ(n)\equiv\varepsilon=1/N,\quad N=n_1-r+1. \tag{3.6}$$

Therefore, the posterior distribution is (3.5). Next, let $n_1\to\infty$. Then the posterior distribution is

$$P^*(n)=pP(n,r). \tag{3.7}$$

[Example 1] Estimate n when $p=0.2$ and $r=20$.

Point estimation is $n=r/p=100$.

Interval estimation is as follows: The result of Program 1 is shown in Table 2

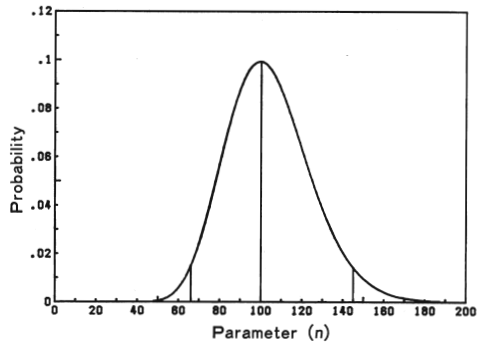

Fig. 4. The graph of $P(n, r)$ when $p=0.2$ and $r=20$.

Table 2. Values of binomial distributions: $P(n, r)$ when $p=0.2$ and $r=20$.

n	P	$\sum P$
20	1.048576×10^{-14}	1.048576×10^{-14}
65	.01294324	.07890761
66	.01485659	.09376420
67	.01694284	.11070703
68	.01920188	.12990892
98	.09904946	2.10339186
99	.09930021	2.20269208
100	.09930021	2.30199229
101	.09905503	2.40104732
144	.01527861	4.82707068
145	.01417855	4.84124923
146	.01314329	4.86656300
147	.01217048	4.87782069
149	.01040246	4.88822315
200	.00006090	4.99953125
∞	.00000000	5.00000000

and Fig. 4. The 95% confidence interval is $n=66\sim145$. Length for the interval of n is the minimum.

2. The conventional method

The non-Bayesian statistical method is as follows: When $n\to\infty$, a binomial distribution approaches to a normal distribution with $\mu=np$ and $\sigma=\sqrt{npq}$. Let z be as follows:

$$z=\frac{np-r}{\sqrt{npq}},\qquad p+q=1. \tag{3.8}$$

Then z distributes according to the standardized normal distribution $N(0, 1)$. The confidence interval is easily obtained (ex. 95% confidence interval is $-1.96\leqq z\leqq1.96$). From (3.8)

$$n=\frac{r}{p}+z\sqrt{\frac{nq}{p}}. \tag{3.9}$$

Let p of the last term be fixed. Then we obtain the rough estimator as variance of n:

$$V(n)=\sigma^2(n)=\frac{nq}{p}=\frac{r(1-p)}{p^2}. \tag{3.10}$$

The rough 95% confidence interval is given by $r/p\pm2\sigma$.

[Example 1′] Estimate n when $p=0.2$ and $r=20$.

Point estimation is $n=r/p=100$.

Interval estimation is as follows:

Let $z=\pm1.96\fallingdotseq\pm2$. Then (3.8) squared becomes

$$4=(0.2n-20)^2/0.16n$$

$$16n=(n-100)^2$$

$$n^2-216n+10000=0$$

$$n=108\pm40.8=67.2,\ 148.8$$

Then $n=67\sim149$. If $z=\pm1.96$, then $n=68\sim148$. These results are larger than for the interval and are shifted to large side of n from those obtaines with the Bayesian statistical method.

3. Relation between the Bayesian statistical method and the conventional method.

Fig. 5, which is an inverse of Table 1 for the ordinate, shows an image of (r, n) coordinates. The Bayesian statistical model is along the ordinate in Fig. 5 and the conventional model is along the abscissa. Theorem 2 proves that the per-

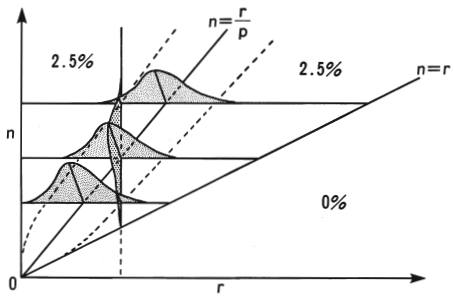

Fig. 5. The image of (r, n) coordinates.

cent points for the ordinate correspond with those for the abscissa. Although the distribution for the abscissa is almost symmetric, that for the ordinate is not. Therefore, the confidence interval of the conventional method is larger than and shifted to the large side of n from that of the Bayesian statistical method (Fig. 6). The difference of results from both method is not so large. However, the conventional method is difficult to obtain the strict solution using the binomial distribution itself. The method presented in this paper based on Bayesian statistics is simpler, easier and more logical than the conventional method.

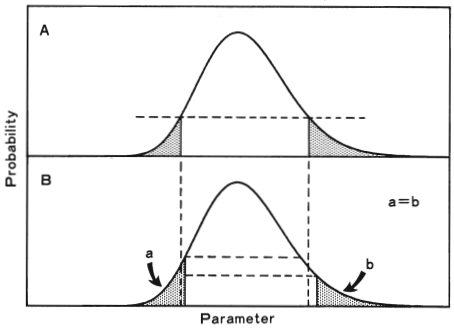

Fig. 6. The comparison of the confidence interval for Bayesian statistical method (A) and that for the conventional method (B).

Acknowledgements

The author is indebted to Dr. Y. MATSUMIYA of the Ocean Research Institute of the University of Tokyo and Mr. K. HIRAMATSU of the Far Seas Fisheries Research Laboratory for their kind advice. The author is also grateful to Mr. K. NOGAMI and Mr. S. UMEZAWA of Japan Sea Regional Fisheries Research Laboratory for their critical reading of the manuscript.

References

AKAMINE, T. (1981) A handy method for selecting and counting bivalve larvae at planktonic stage. *Bull. Japan Sea Reg. Fish. Res. Lab.*, (32), 77–81. (In Japanese with English abstract)

ベイズ統計による抽出法の区間推定

赤 嶺 達 郎

抽出法のモデルは二項分布と一致する．従来の手法は二項分布を正規分布に近似して行うものであった．母数の事前分布を一様分布と仮定するベイズ統計の手法は従来の手法より優れている．二項分布に関する二つの定理によってこのモデルは従来の手法とよく一致することが示される．このモデルは従来の手法より単純で理解しやすく小型計算機で容易に計算できる．

Appendix

Program 1. An example of BASIC programs to calculate a confidence interval for extraction using Bayesian statistics.

```
100  '-----------------------------------------------------------
110  '          Interval estimation for Extraction
120  '               (Binomial distribution)
130  '                                    by Tatsuro Akamine
140  '                                                 1988-08-31
150  '-----------------------------------------------------------
1000 DEFINT I-N
1010 DEFDBL A-H,O-Z
1020 P1=.2# : IR1=20
1030 N1=INT(IR1/P1) : Q1=1#-P1
1040 C1=1#
1050 FOR I=0 TO IR1-1
1060    C1=C1*(N1-I)/(IR1-I)*P1
1070 NEXT I
1080 FOR J=1 TO N1-IR1 : C1=C1*Q1 : NEXT J
1090 PROB=C1 : AREA=C1*P1
1100 PRINT N1,PROB,AREA
2000 BPRO=PROB*Q1*(N1+1)/(N1-IR1+1)
2010 SPRO=PROB*(N1-IR1)/Q1/N1
2020 N1S=N1 : N1B=N1
2030 *REPEAT
2040  IF BPRO>SPRO GOTO *RIGHT
3000 *LEFT
3010  N1S=N1S-1
3020  N1S9=N1S : SPRO9=SPRO
3030  AREAS=SPRO*P1
3040  AREA=AREA+AREAS
3050  SPRO=SPRO*(N1S-IR1)/Q1/N1S
3060  GOTO *CHECK
4000 *RIGHT
4010  N1B=N1B+1
4020  N1B9=N1B : BPRO9=BPRO
4030  AREAB=BPRO*P1
4040  AREA=AREA+AREAB
4050  BPRO=BPRO*Q1*(N1B+1)/(N1B-IR1+1)
5000 *CHECK
5010  IF AREA<.95# GOTO *REPEAT
5020  PRINT N1S9,SPRO9,AREAS
5030  PRINT N1B9,BPRO9,AREAB
5040  PRINT AREA
5050 END
```