PART II

CHAPTER 6

ESTIMATION IN THE TRUNCATED BINOMIAL DISTRIBUTION

6.1 Introduction

The uses of zero-truncated binomial distribution were discussed by Fisher [18], Haldane ([20], [21]) and Finney [17]. For example, in problems of human genetics, in estimating the proportion of albino children produced by couples capable of producing albinos, sampling has necessarily to be restricted to families having atleast one albino child. The problem of estimating the proportion p of the zero-truncated binomial distribution has been discussed by Fisher [18], Haldane [21], Finney [17], Patil ([29], [30]) and Rider [36]. The different types of the estimators of the proportion p in this case can be classified as (i) the maximum likelihood estimator, (Fisher [18], Haldane [20], [21], Finney [17] and Patil [29]), (ii) the ratio-estimator (Patil [30]) and (iii) the method of moments estimator (Rider [36]).

In Section 6.2, we consider the binomial distribution in which k classes from the lower end of the distribution are truncated. We derive here the asymptotic variance of the method of moments estimator of the proportion p. An important application of interest in genetics arises when $k = 1$. In this case, the asymptotic efficiency of the method of moments estimator has been compared with that of maximum likelihood estimator and also with that of the ratio-estimator.

In Section 6.3, we consider doubly truncated binomial distribution when k classes from the lower end and h classes from the upper end are truncated. We derive here the method of moments estimator of p and its asymptotic variance. Comparison of the asymptotic efficiency of the method of moments estimator of p has been made with that of the maximum likelihood estimator in the case $h = k = 1$ and $n = 10$.

6.2 Singly truncated binomial distribution

In this section, we consider the truncated binomial distribution in which k classes from the lower end of the distribution are truncated and the estimation of its parameter by the method of moments and the comparison of the method of moments estimator

with the maximum likelihood estimator and ratio estimator.

6.2.1 The method of moments estimator of p

The probability law of the binomial distribution in which k classes from the lower end of the distribution are truncated is

$$p_j = S_j/G, \qquad (j = k, k+1, \ldots, n) \tag{6.2.1.1}$$

where $S_j = \binom{n}{j} p^j q^{n-j}$ and $G = \sum_{j=k}^{n} S_j$; $0 < p < 1$ and $q = 1-p$. Let $\mu'_r = E(j^r)$ be the rth raw moment of the probability distribution (6.2.1.1). By considering the recurrence relation between the probabilities p_j , one easily derives the following relation between the moments μ'_r .

$$\mu'_1 = np + qkp_k , \tag{6.2.1.2}$$

$$\begin{aligned} \mu'_{r+1} = &- np(k-1) + [k + p(n-1)]\,\mu'_r \\ &+ p\sum_{j=1}^{r-1}\Big[n\Big\{\binom{r}{j} + \binom{r-1}{j}\Big\} \\ &- \Big\{\binom{r}{j-1} - k\binom{r-1}{j-1}\Big\}\Big]\mu'_j , \end{aligned} \tag{6.2.1.3}$$

$$r = 1, 2, 3, \ldots .$$

Putting $r = 1$ in (6.2.1.3), we get

(6.2.1.4) $$\mu_2' - k\mu_1' = pA,$$

where $A = (n-1)\mu_1' - n(k-1)$.

Now, since

$$A = \sum_{j=k}^{n} [(n-1)j - n(k-1)]p_j$$

$$= \sum_{j=k}^{n} [n(j-k) + (n-j)]p_j ,$$

and the bracketed quantity in A is always positive as $k \leq j \leq n$, it follows that $A > 0$. Hence, we have from (6.2.1.4)

(6.2.1.5) $$p = (\mu_2' - k\mu_1')/A.$$

Now consider a random sample of size N from the truncated binomial distribution (6.2.1.1). Let n_x be the observed frequency corresponding to x in the sample. Let $m_r = \sum_{x=k}^{n} x^r n_x / N$ denote the rth order raw sample moment. Then, equation (6.2.1.5) suggests that we can estimate p by substituting m_r for μ_r' in the r.h.s. expression of (6.2.1.5), i.e., by

(6.2.1.6) $$\hat{p} = (m_2 - km_1)/[(n-1)m_1 - n(k-1)].$$

The denominator in (6.2.1.6) is always positive by the same argument as for A with n_j in place of p_j. The estimator $\hat{p}$ was first given by Rider [36].

By using the δ-method (Kendall and Stuart [23], 10.6), the asymptotic variance of $\hat{p}$ has been derived. Let δx denote the differential of x. Taking differentials, we can derive from (6.2.1.6)

$$A\delta\hat{p} = \delta m_2 - D\delta m_1 , \tag{6.2.1.7}$$

where $D = k + p(n-1)$ and A has the same meaning as in (6.2.1.4). Squaring both sides of (6.2.1.7) and taking expectations, we get the asymptotic variance of $\hat{p}$ as

$$V(\hat{p}) = \left[\alpha_{22} - 2D\alpha_{12} + D^2\alpha_{11}\right]/NA^2 , \tag{6.2.1.8}$$

where $\alpha_{ij} = \mu'_{i+j} - \mu'_i \mu'_j$. Using the recurrence relation (6.2.1.3) between the moments, the asymptotic variance of $\hat{p}$ can be seen to be

$$V(\hat{p}) = \left[pq\{p(n-3)-(k-1)\}A + p\mu'_1\{(n-1)2q - (k-1)(k+2q-nk)\}\right]/NA^2 . \tag{6.2.1.9}$$

6.2.2 Comparison with the maximum likelihood estimator of p

The maximum likelihood estimator p^* of p of the distribution (6.2.1.1) is given by

$$m_1 = np^* + q^*kp_k^* , \tag{6.2.2.1}$$

where q^* and p_k^* denote the values of q and p_k when $p = p^*$. The asymptotic variance of p^* is given by

(6.2.2.2) $$V(p^*) = p^2q^2/N\mu_2 ,$$

where μ_2 = variance of the distribution (6.2.1.1) = $\mu_1'(np + q - \mu_1') + \mu_1'(k-1) - np(k-1)$.

A special case of interest in genetics arises when $k = 1$. In this case (6.2.1.6) and (6.2.2.1) become

(6.2.2.3) $$\hat{p} = (m_2 - m_1)/m_1(n - 1),$$

(6.2.2.4) $$m_1 = np^* + q^*p_1^* .$$

Also, (6.2.1.9) and (6.2.2.2) become

(6.2.2.5) $$V(\hat{p}) = pq(np-3p+2)/N(n-1)\mu_1'^2 ,$$

(6.2.2.6) $$V(p^*) = p^2q^2/N\mu_1'(np+q-\mu_1') .$$

Thus, when $k = 1$, the asymptotic efficiency E of the method of moments estimator relative to the maximum likelihood estimator is given by

(6.2.2.7) $$E = (n-1)pq/(np+q-\mu_1')(np-3p+2),$$

wherein $\mu_1' = np/(1-q^n)$. The values of E are tabulated in Table 6.2.2.1 for $p = 0.25, 0.50$, and 0.75 and $n = 3, 4, 5, 6, 7, 8, 9, 10$.

TABLE 6.2.2.1

ASYMPTOTIC EFFICIENCY OF $\hat{p}$ FOR $k = 1$

n	Values of p		
	0.25	0.50	0.75
3	0.925	0.875	0.875
4	0.871	0.818	0.859
5	0.831	0.795	0.870
6	0.802	0.789	0.886
7	0.781	0.794	0.901
8	0.766	0.803	0.913
9	0.755	0.815	0.923
10	0.749	0.826	0.931

6.2.3 Comparison with the ratio-estimator of p

The ratio-estimator p' of p due to Patil [30] of the distribution (6.2.1.1) is given by

$$p' = t_1/(t_1 + t_2), \tag{6.2.3.1}$$

where $t_1 = \sum_{x=k+1}^{n} xn_x/(n - x + 1)$ and $t_2 = \sum_{x=k}^{n-1} n_x$. The asymptotic variance of p' (Patil [30]) is given by

$$V(p') = q^2\left[q^2R - Pp^2 + 2p^2p_{n-1}\right]/NP^2 , \tag{6.2.3.2}$$

where $P = \sum_{x=k}^{n-1} p_x$, $R = \sum_{x=k+1}^{n} x p_x / (n-x+1)$.

In the special case when $k = 1$, the asymptotic efficiency of the ratio-estimator in comparison with the maximum likelihood estimator has been tabulated by Patil [30] and is reproduced here in Table 6.2.3.1 for the sake of comparison with the method of moments estimator.

TABLE 6.2.3.1

ASYMPTOTIC EFFICIENCY OF RATIO-ESTIMATOR p' FOR $k = 1$

n	Values of p		
	0.25	0.50	0.75
3	0.924	0.875	0.875
4	0.909	0.769	0.772
5	0.919	0.715	0.664
6	0.933	0.694	0.563
7	0.947	0.693	0.523
8	0.952	0.705	0.481
9	0.956	0.723	0.435
10	0.959	0.776	0.388

Comparison of Tables 6.2.2.1 and 6.2.3.1 shows that the method of moments estimator is more efficient than ratio-estimator when $p = 0.50$ and 0.75, while

ratio-estimator is more efficient than the method of moments estimator when $p = 0.25$.

6.2.4 An illustrative example

As an illustration, we take K. Pearson's [33] data on albinism in man. Table 6.2.4.1 gives the number of families (n_x), each of five children, having exactly x albino children in the family ($x = 1, 2, 3, 4, 5$).

TABLE 6.2.4.1

NUMBER OF ALBINOS IN FAMILIES HAVING FIVE CHILDREN

No. of albinos in a family x	No. of families n_x
1	25
2	23
3	10
4	1
5	1
Total	60

For this example, we obtain $m_1 = 1.8333$, $m_2 = 4.1333$ and using (6.2.2.3), we find that the method of moments estimate of p is $\hat{p} = 0.3136$.

Using (6.2.2.5), we find that the estimate of the asymptotic variance of $\hat{p}$ is $V(\hat{p}) = 0.001285$.

Using (6.2.2.4), we find that the maximum likelihood estimate of p is $p^* = 0.3088$ and using (6.2.2.6), we find that the estimate of the asymptotic variance of p^* is $V(p^*) = 0.001030$.

Using (6.2.3.1) for $k = 1$, we find that the ratio-estimate of p is $p' = 0.3257$ and using (6.2.3.2) for $k = 1$, we find that the estimate of the asymptotic variance of p' is $V(p') = 0.001341$.

Thus, we obtain that (i) the estimated asymptotic efficiency of the ratio-estimate relative to the maximum likelihood estimate is 0.768, and (ii) the estimated asymptotic efficiency of the method of moments estimate relative to the maximum likelihood estimate is 0.802.

6.3 Doubly truncated binomial distribution

In this section, we consider the estimation of the parameter of doubly truncated binomial distribution by the method of moments and the comparison of the method of moments estimator with the maximum likelihood estimator.

6.3.1 The method of moments estimator of the parameter p

The probability law of the binomial distribution in which k classes from the lower end and h classes from the upper end of the distribution are truncated is

$$(6.3.1.1) \qquad p_j = S_j/G', \quad (j = k,\ k+1,\ \ldots,\ n-h)$$

where $G' = \sum_{j=k}^{n-h} S_j$ and S_j has the same meaning as in (6.2.1.1). Let $\mu_r' = E(j^r)$ be the rth order raw moment of the distribution (6.3.1.1). By considering the recurrence relation between the probabilities p_j given by (6.3.1.1), the following recurrence relation between moments μ_r' is obtained.

$$(6.3.1.2) \qquad \mu_1' = X - Y + np,$$

$$(6.3.1.3) \qquad \mu_{r+1}' = k^r X - a^r Y + np + p \sum_{j=1}^{r} \left\{ n\binom{r}{j} - \binom{r}{j-1} \right\} \mu_j', \qquad r = 1, 2, \ldots .$$

where $X = qkp_k$, $Y = php_{n-h}$, $a = (n-h+1)$, and $q = 1-p$. Taking $r = 1$ and 2, we get from (6.3.1.3)

$$(6.3.1.4) \qquad \mu_2' = kX - aY + p(n-1)\mu_1' + np,$$

$$(6.3.1.5) \qquad \mu_3' = k^2 X - a^2 Y + p(n-2)\mu_2' + p(2n-1)\mu_1' + np.$$

Eliminating X and Y from (6.3.1.2), (6.3.1.4) and (6.3.1.5), we obtain the solution for p as

$$(6.3.1.6) \qquad p = \left[\mu_3' - (k+a)\mu_2' + ka\mu_1'\right]/Q,$$

where $d = (n-1)(h-k) - n(n-2)$ and $Q = [(n-2)\mu_2' + d\mu_1' + n(n-h)(k-1)]$. Now Q can be written as

$$(6.3.1.7) \qquad Q = \sum_{j=k}^{n-h} [(n-2)j^2 + dj + n(n-h)(k-1)]p_j .$$

When $h = k = 1$, $Q = \sum_{j=1}^{n-1} (j^2 - nj)p_j < 0$. Further, the bracketed quantity in Q in (6.3.1.7) is a convex function of j and therefore has its maximum at either $j = k$ or $j = n-h$. At $j = k$, it is $-(n-k)(n-h-k)$ which is negative as $n > k$ and $n > h+k$. At $j = n-h$, it is $-(n-h)(n-h-k)$ which is also negative. Hence $Q \neq 0$.

Now consider a random sample of size N from the truncated distribution (6.3.1.1). Let n_x be the observed frequency corresponding to x in the sample. Let $m_r = \sum_{x=k}^{n-h} x^r n_x / N$ denote the rth order raw sample moment. The equation (6.3.1.6) suggests that an estimator of p is obtained by substituting m_r for μ_r' in the r.h.s. expression in (6.3.1.6), i.e., by

$$(6.3.1.8) \qquad p = \frac{[m_3 - (k+a)m_2 + kam_1]}{[(n-2)m_2 + dm_1 + n(n-h)(k-1)]} .$$

The denominator in (6.3.1.8) can be seen to be non-zero by following the same argument as employed in showing that $Q \neq 0$ in (6.3.1.7) with n_x in place of p_x.

By the δ-method used previously to find the asymptotic variance of $\hat{p}$ in Section 6.2.1, we obtain the asymptotic variance of $\hat{p}$ as

$$V(\hat{p}) = B' \Lambda B / NQ^2, \tag{6.3.1.9}$$

where

$$B = \begin{bmatrix} b_1 \\ b_2 \\ 1 \end{bmatrix}, \qquad \Lambda = \begin{bmatrix} \alpha_{11}, & \alpha_{12}, & \alpha_{13} \\ \alpha_{21}, & \alpha_{22}, & \alpha_{23} \\ \alpha_{31}, & \alpha_{32}, & \alpha_{33} \end{bmatrix},$$

$b_1 = ka - pd$, $b_2 = -(k+a) - p(n-2)$, $\alpha_{ij} = \mu'_{i+j} - \mu'_i \mu'_j$,

while a, d and Q have respectively the same meanings as in (6.3.1.3), (6.3.1.6) and (6.3.1.6).

6.3.2 Comparison with the maximum likelihood estimator of p

By applying the well-known principle of maximum likelihood estimation, we see that the maximum likelihood estimator p^* of the parameter p of a doubly truncated binomial distribution (6.3.1.1) is given by

$$m_1 = X^* - Y^* + np^*, \tag{6.3.2.1}$$

where X^*, Y^* and p^* denote the values of X, Y and p when $p = p^*$, X and Y having the same meanings as in (6.3.1.2). The asymptotic variance of p^* is given by

(6.3.2.2) $$V(p^*) = (pq)^2/Nu_2 ,$$

where μ_2 is the variance of the distribution (6.3.1.1), and is given by $\mu_2 = kX - aY - (X - Y + np)(X - Y + p)$. Then, the asymptotic efficiency of the method of moments estimator $\hat{p}$ relative to the maximum likelihood estimator p^* is given by

(6.3.2.3) $$(pq)^2Q^2/\mu_2 B' \wedge B.$$

6.3.3 A special case : $h = k = 1$

Let us consider a particular case when $h = k = 1$, i.e., when one class from the lower end and one class from the upper end of binomial distribution are truncated. From (6.3.1.8), by putting $h = k = 1$, we obtain the method of moments estimator $\hat{p}$ as

(6.3.3.1) $$\hat{p} = [m_3 - (n+1)m_2 + nm_1]/(n-2)(m_2 - nm_1).$$

The asymptotic variance of p as obtained from (6.3.1.9) by taking $h = k = 1$ is given by

(6.3.3.2) $$V(\hat{p}) = B' \wedge B/NQ^2,$$

where

$$B = \begin{bmatrix} n(1-2p+np) \\ -n-1-np+2p \\ 1 \end{bmatrix}, \qquad Q = (n-2)(m_2 - nm_1) ,$$

and has the same meaning as in (6.3.1.9).

From (6.3.2.1), by taking $h = k = 1$, we obtain the maximum likelihood estimator p^* as given by

$$(6.3.3.3) \qquad m_1 = n(p^* - S_n^*)/(1 - S_0^* - S_n^*),$$

where $S_n^* = p^{*n}$, $S_0^* = (1 - p^*)^n$ and m_1 = the sample mean. To facilitate the solution of the equation (6.3.3.3), the function $F(p, n) = (p - S_n)/(1 - S_0 - S_n)$ is tabulated in Table 6.3.3.1 for $p = 0.10, 0.15. 0.20, 0.25, 0.30. 035, 0.40, 0.45, 0.50$ and $n = 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20$. Equation (6.3.3.3) can be written as

$$(6.3.3.4) \qquad (m_1/n) = F(p^*, n).$$

Thus, the value of p^* can be obtained by inverse interpolation from Table 6.3.3.1. It can be easily verified that

$$(6.3.3.5) \qquad F(1-p, n) = 1 - F(p, n).$$

Hence, noting this symmetry, the values of $F(p, n)$ for $p = 0.55, 0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.90$ can be obtained from the values of $F(p, n)$ for $p = 0.45, 0.40, 0.35, 0.30, 0.25, 0.20, 0.15, 0.10$.

TABLE 6.3.3.1

VALUES OF F(p, n)

n \ p	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
4	0.29058	0.31308	0.34458	0.36207	0.38827	0.41537	0.44318	0.47147	0.50000
5	0.24418	0.26954	0.29674	0.32692	0.35873	0.39233	0.42737	0.46342	0.50000
6	0.21342	0.24082	0.27099	0.30392	0.33946	0.37731	0.41707	0.45818	0.50000
7	0.19168	0.22078	0.25306	0.28846	0.32676	0.36762	0.41053	0.45432	0.50000
8	0.17558	0.20618	0.24032	0.27779	0.31830	0.36137	0.40644	0.45287	0.50000
9	0.16324	0.19522	0.23100	0.27029	0.31260	0.35735	0.40391	0.45166	0.50000
10	0.15353	0.18677	0.22406	0.26492	0.30872	0.35476	0.40237	0.45095	0.50000
11	0.14573	0.18014	0.21879	0.26102	0.30605	0.35308	0.40143	0.45054	0.50000
12	0.13936	0.17487	0.21476	0.25818	0.30421	0.35200	0.40086	0.45031	0.50000
13	0.13408	0.17063	0.21163	0.25608	0.30293	0.35130	0.40052	0.45017	0.50000
14	0.12966	0.16718	0.20920	0.25454	0.30205	0.35084	0.40031	0.45010	0.50000
15	0.12593	0.16436	0.20729	0.25338	0.30143	0.35055	0.40019	0.45005	0.50000
16	0.12275	0.16203	0.20579	0.25253	0.30100	0.35036	0.40011	0.45003	0.50000
17	0.12001	0.16010	0.20461	0.25189	0.30070	0.35023	0.40007	0.45002	0.50000
18	0.11766	0.15850	0.20367	0.25142	0.30049	0.35015	0.40004	0.45001	0.50000
19	0.11562	0.15717	0.20292	0.25106	0.30034	0.35010	0.40002	0.45000	0.50000
20	0.11384	0.15605	0.20233	0.25080	0.30024	0.35006	0.40001	0.45000	0.50000

The asymptotic variance of p^* can be obtained from (6.3.2.2) by taking $h = k = 1$ and is given by

$$V(p^*) = (pq)^2/N\mu_2 , \tag{6.3.3.6}$$

where $\mu_2 = X - nY - (X - Y + np)(X - Y + p)$, and $X = qS_1/G'$, $Y = pS_{n-1}/G'$, $G' = \sum_{j=1}^{n-1} S_j$, $S_j = \binom{n}{j}p^j q^{n-j}$, $j = 1, 2, \ldots, (n-1)$. The asymptotic efficiency of the method of moments estimator $\hat{p}$ relative to the maximum likelihood estimator p^* can be obtained by dividing (6.3.3.6) by (6.3.3.2) and is tabulated in Table 6.3.3.2 for $n = 10$ and $p = 0.1$, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9.

TABLE 6.3.3.2

ASYMTOTIC EFFICIENCY OF $\hat{p}$

FOR $h=k=1$ AND $n=10$

p			efficiency
0.1	or	0.9	0.9206
0.2	or	0.8	0.9032
0.3	or	0.7	0.9121
0.4	or	0.6	0.9292
0.5			0.9346

6.3.4 An illustrative example

For illustrating the results of Section 6.3, we

consider the data of Table 6.3.4.1 (taken from Rider's paper [36]), about the number of boys in families having eight children. Let one class from the lower end and one class from the upper end of the distribution be truncated.

TABLE 6.3.4.1

NUMBER OF BOYS IN FAMILIES

HAVING EIGHT CHILDREN

No. of boys	No. of families
0	215
1	1,485
2	5,331
3	10,649
4	14,959
5	11,929
6	6,678
7	2,092
8	342
Total	53,680

Then, we get

$$N = 53123 \quad , \quad m_1 = 4.10909,$$
$$m_2 = 18.80795 \quad , \quad m_3 = 92.9948,$$
$$m_4 = 487.775 \quad , \quad m_5 = 2681.405,$$
$$m_6 = 15312.745 \, .$$

Using (6.3.3.1) and substituting the values of m_1, m_2, m_3 and $n = 8$, we find the method of moments estimate of p to be $p = 0.51434$. Using (6.3.3.2) and substituting the values of m_1, m_2, m_3, m_4, m_5, m_6 for μ_1', μ_2', μ_3', μ_4', μ_5', μ_6', we find that the estimate of the asymptotic variance of $\hat{p}$ is given by $V(\hat{p}) = 0.035751/N$.

Further, using (6.3.3.4) and Table 6.3.3.1, we find the maximum likelihood estimate of p to be $p^* = 0.51378$. Using (6.3.3.6) we find that the estimate of the asymptotic variance of p^* is given by $V(p^*) = 0.03306/N$. Thus the estimate of the asymptotic efficiency of the method of moments estimator relative to the maximum likelihood estimator is 0.9294. We note that the method of moments estimator is easy to compute while the loss of efficiency is not much.