ON MINIMUM VARIANCE UNBIASED ESTIMATION FOR TRUNCATED BINOMIAL AND NEGATIVE BINOMIAL DISTRIBUTIONS

T. CACOULLOS AND CH. CHARALAMBIDES

(Received Oct. 14, 1972)

1. Introduction

The estimation of the parameter p for binomial and negative binomial distributions truncated away from zero has been considered in a few papers. Thus, Rider [9] gave an intuitive estimator of p without any minimum variance or unbiasedness considerations. Fisher [5] and Patil [7] examined maximum likelihood (ML) estimation in the binomial case. Sampford [10] gave moment and ML estimators of p for the negative binomial distribution under truncation at zero.

This paper is concerned with uniformly minimum variance unbiased (UMVU) estimation under truncation away from zero. In this respect the results of Patil [8] are quite pertinent; treating the binomial as a generalized power series distribution with parameter $\theta=p/q$, it follows from Corollary 1, p. 1052 (op. cit.) that there is no UMVU estimator of θ since the range of values of the binomial variable is finite. Moreover, it can easily be verified that no UMVU estimator exists for p either. On the other hand, there is a UMVU estimator for p in the negative binomial case, which we are actually going to construct. As regards the binomial case, it is interesting to ignore momentarily the non-existence of a UMVU estimator and proceed, heuristically, to construct an unbiased estimator of the odds ratio θ based on a sufficient statistic, namely, the sum T of the observations. This serves a double purpose. First it illustrates the dependence of an unbiased estimator on the range of the distribution and, second, it leads to a sufficient estimator $\tilde{\theta}_0(T)$ whose relative bias is practically negligible (see Section 2.2). Furthermore, the expression of $\tilde{\theta}_0(T)$ (see (2.7)) in terms of certain analogues of the Stirling numbers of the second kind can be used to show the intrinsic relations in the estimation problems of the parameters of the Poisson (cf. [11]), binomial and negative binomial distributions, each truncated away from zero.

A main feature of this paper is the derivation of the distribution of the sufficient statistic T in each case by applying the Poincaré formula and certain combinatorial results. A similar method was used in [2] for the Poisson case. Ahuja [1] recently derived the distribution of the sum of truncated negative binomial variables by applying some results of Patil [8]. The distribution of T in the binomial case, apparently, has not been given so far in any explicit form.

The truncated negative binomial distribution arises in several applications in biology (see Sampford [10]). The truncated binomial occurs when one observes the presence or absence of a given attribute in certain finite groups where at least a specified number of their members have the attribute. For example, the group may be a family and the attribute chicken pox; here the specified number would be one and the situation calls for a truncated binomial away from zero.

2.1. The distribution of the sum of truncated binomial variables

PROPOSITION 2.1. Let $X_1, \cdots, X_n$ be a random sample from the truncated binomial distribution away from zero, that is, with probability distribution

$$\text{(2.1)} \qquad \mathrm{P}\,[X_i=x]=\frac{1}{1-q^r}\binom{r}{x}p^x q^{r-x}\,, \qquad x=1,\cdots,r,\ i=1,\cdots,n$$

($r>1$, $q=1-p$). Then the probability distribution $b_0(t)$, say, of the sufficient statistic $T=X_1+\cdots+X_n$ for the parameter p is given by

$$\text{(2.2)} \qquad b_0(t)=\frac{n!r^n}{(1-q^r)^n}\frac{C^t_{r,n}}{t!}p^t q^{rn-t} \qquad t=n,\, n+1,\cdots,rn$$

where the numbers $C^t_{r,n}$ are defined for all integers r, n, t ($n>0$, $t>0$, $r\neq 0$) by

$$\text{(2.3)} \qquad C^t_{r,n}=\frac{(-1)^n}{n!r^n}\sum_{k=1}^{n}(-1)^k\binom{n}{k}(rk)_t\,.$$

PROOF. We have, by (2.1) and the independence of X_i,

$$\text{(2.4)} \qquad b_0(t)=\mathrm{P}\,[T=t]=\mathrm{P}\left[\sum_{i=1}^{n}X_i=t\right]=\sum_{(x_1,\cdots,x_n)}\prod_{i=1}^{n}\frac{1}{1-q^r}\binom{r}{x_i}p^{x_i}q^{r-x_i}$$

$$=\frac{\binom{rn}{t}p^t q^{rn-t}}{(1-q^r)^n}\sum_{(x_1,\cdots,x_n)}\frac{\binom{r}{x_1}\cdots\binom{r}{x_n}}{\binom{rn}{t}}$$

where the summation extends over all ordered n-tuples $(x_1,\cdots,x_n)$ of integers x_i, $1\leqq x_i\leqq r$, such that $x_1+\cdots+x_n=t$.

In order to evaluate the last sum, consider the following coupon collector's problem (see, e.g. Feller [3], p. 111). An urn contains r identical groups of balls each group consisting of n balls numbered 1, $2, \cdots, n$; t $(t \geqq n)$ balls are drawn without replacement one after another. Then the probability that each of the numbers $1, 2, \cdots, n$ is included in the sample is given by the last sum in (2.4). This probability p_0, say, can be obtained by Poincaré formula as follows: Let A_k denote the event: the kth number does not show up $(k=1, \cdots, n)$. Then

$$p_0 = 1 - \mathrm{P}\left[\bigcup_{k=1}^{n} A_k\right] = 1 - \sum_{k=1}^{n} (-1)^{k-1} S_k$$

where

$$S_k = \sum \mathrm{P}\,[A_{i_1}, \cdots, A_{i_k}] = \binom{n}{k} \frac{(rn-rk)_t}{(rn)_t} \qquad (k=1, 2, \cdots, n)$$

so that p_0, by (2.3) also, becomes

$$p_0 = \sum_{k=0}^{n-1} (-1)^k \binom{n}{k} \frac{(r[n-k])_t}{(rn)_t} = \frac{n!r^n}{(rn)_t} C_{r,n}^t \,.$$

Therefore, in virtue of (2.4), we obtain (2.2).

The distribution of T can also be obtained by inverting its characteristic function. This is given by

$$\varphi_T(u) = [\varphi_X(u)]^n = \frac{1}{(1-q^r)^n} [(pe^{iu}+q)^r - q^r]^n$$

where φ_X denotes the characteristic function of (2.1). By the Fourier inversion formula, the probability distribution $b_0(t)$ of T is given by

$$\begin{aligned} b_0(t) &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \varphi_T(u) e^{-itu} du \\ &= \frac{1}{(1-q^r)^n} \frac{1}{2\pi} \int_{-\pi}^{\pi} \left[\sum_{k=0}^{n} (-1)^{n-k} \binom{n}{k} (pe^{iu}+q)^{rk} q^{r(n-k)}\right] e^{-itu} du \end{aligned}$$

which, after expanding $(pe^{iu}+q)^{rk}$ and taking into account

$$\int_{-\pi}^{\pi} e^{iu(s-t)} du = 0 \qquad \text{for every } s \neq t\,,$$

gives (2.2).

2.2. An asymptotically UMVU estimator of the odds ratio θ

As stated in the introduction, there is no UMVU estimator for $\theta = p/q$ as shown in [7]; actually, there is no unbiased estimator based on the sufficient statistic T in which case, T being also complete, would imply the existence of a UMVU estimator of θ by the well known theorems of Rao-Blackwell and Lehmann-Scheffé. Nonetheless, ignoring this fact, let us denote by $\tilde{\theta}_0(t)$ an estimate of θ based on T and let us proceed to see where unbiasedness fails. The condition of unbiasedness is

$$\sum_{t=n}^{rn} \tilde{\theta}_0(t) b_0(t) \equiv \theta \tag{2.5}$$

and since $b_0(t)$ in (2.2) can be written as

$$b_0(t) = \frac{n! r^n}{[(1+\theta)^r - 1]^n} \frac{C_{r,n}^t}{t!} \theta^t ,$$

condition (2.5) gives

$$\sum_{t=n}^{rn} \tilde{\theta}_0(t) \frac{C_{r,n}^t}{t!} \theta^t \equiv \sum_{t=n}^{rn} \frac{C_{r,n}^t}{t!} \theta^{t+1} \qquad (0<\theta<\infty) . \tag{2.6}$$

Equating the coefficients of θ^t on both sides shows that (2.6) can be satisfied if, and only if,

$$\tilde{\theta}_0(t) = t \frac{C_{r,n}^{t-1}}{C_{r,n}^t} \qquad t=n, n+1, \cdots, rn-1 \tag{2.7}$$

$$\frac{C_{r,n}^{rn}}{(rn)!} = \frac{1}{n! r^n} = 0 \tag{2.8}$$

where we have used the fact that, by definition,

$$C_{r,n}^{rn} = \frac{(rn)!}{n! r^n} \qquad \text{for } r=1, 2, \cdots . \tag{2.9}$$

Obviously, however, (2.8) cannot be satisfied and $\tilde{\theta}_0(t)$, as defined in (2.7), has relative bias, caused by the last term θ^{rn+1} in (2.6), equal to

$$\beta(\tilde{\theta}_0) = \mathrm{E}\left(\frac{\tilde{\theta}_0}{\theta}\right) - 1 = -b_0(rn) = -\frac{p^{rn}}{(1-q^r)^n} . \tag{2.10}$$

Since $p^r/(1-q^r) \leqq 1$ with strict inequality for $0<p<1$ and moreover this is an increasing function of p, it follows that $\beta(\tilde{\theta}_0)$ approaches zero very fast even for moderate values of the sample size n, and the relative bias is practically negligible for all values of p not close to 1 (not too large values of θ). It can be easily verified that $|\beta(\tilde{\theta}_0)|$ as a function of r is strictly decreasing and $\tilde{\theta}_0(t)$ is asymptotically $(r \to \infty)$ unbiased.

Let us now look at the behavior of $\tilde{\theta}_0(t)$ as $r \to \infty$ and $p \to \infty$. From the preceding discussion we see that $\beta(\tilde{\theta}_0) \to 0$. This asymptotic unbiasedness of $\tilde{\theta}_0(t)$ follows also from (2.6) where now we have an infinite power series in θ and therefore $\tilde{\theta}_0(t)$ as defined in (2.7) becomes unbiased as $r\to\infty$. In fact, we will show immediately that

$$r\tilde{\theta}_0(t)\to\tilde{\lambda}_0(t)=t\frac{\mathfrak{S}^n_{t-1}}{\mathfrak{S}^n_t} \tag{2.11}$$

where $\tilde{\lambda}_0(t)$ denotes the UMVU estimator of the parameter λ of the truncated Poisson at zero obtained in [11] and $\mathfrak{S}^n_t$ denotes the Stirling number of the second kind defined by

$$\mathfrak{S}^n_t=\begin{cases}\dfrac{1}{n!}\sum_{k=1}^{n}(-1)^{n-k}\binom{n}{k}k^t\,, & t=n,\, n+1,\cdots\\ 0\,, & t<n\,.\end{cases}$$

To show (2.11) we use the following property of the numbers $C^t_{r,n}$:

$$\lim_{r\to\infty}\frac{C^t_{r,n}}{r^{t-n}}=\mathfrak{S}^n_t\,. \tag{2.12}$$

It should be noted that (2.11) reflects the Poisson approximation to the binomial as $r\to\infty$ and $p\to 0$ so that $rp\to\lambda$. Thus $r\theta\to\lambda$ and $r\tilde{\theta}_0(t)\to\tilde{\lambda}_0(t)$ as $r\to\infty$ and $p\to 0$.

An intrinsic relation between the Poisson and binomial estimation problems under truncation is further pointed out by the following. The UMVU estimator $\tilde{\lambda}_0(t)$ can be written (see [11]) in the form

$$\tilde{\lambda}_0(t)=\frac{t}{n}\left(1-\frac{\mathfrak{S}^{n-1}_{t-1}}{\mathfrak{S}^n_t}\right)\,.$$

A similar relation holds for $\tilde{\theta}_0(t)$, namely,

$$\tilde{\theta}_0(t)=\frac{t}{rn-t+1}\left(1-\frac{C^{t-1}_{r,n-1}}{C^t_{r,n}}\right)\,. \tag{2.13}$$

This is a consequence of the recurrence relation (easily verifiable):

$$C^t_{r,n}=(rn-t+1)C^{t-1}_{r,n}+C^{t-1}_{r,n-1} \tag{2.14}$$

which can be used also to show, by induction, that the $C^t_{r,n}$ are integers.

The quantity

$$\bar{C}(r,\,n,\,t)=1-\frac{C^{t-1}_{r,n-1}}{C^t_{r,n}} \tag{2.15}$$

may be interpreted as the correction factor by which the modified ML estimator $t/(rn-t+1)$ of θ in the usual case of no truncation has to be multiplied to give the asymptotically ($r\to\infty$) UMVU estimator $\tilde{\theta}_0(t)$ (the usual ML $\hat{\theta}$ of θ is $\hat{\theta}=t/(rn-t)$. Note that $\bar{C}(r,n,n)=0$ and $\bar{C}(r,n,t)$ increases with t reaching the value 1 for $t=rn-1$ and $t=rn$.

3.1. The distribution of $T=\sum X_i$ in the negative binomial case

PROPOSITION 3.1. Let $X_1, \cdots, X_n$ be a random sample from the negative binomial distribution truncated away from zero, that is, with probability distribution

$$\text{(3.1)}\qquad P[X_i=x]=\frac{1}{1-q^r}\binom{r+x-1}{x}p^x q^r \qquad x=1, 2, \cdots,\ i=1, 2, \cdots.$$

Then the probability distribution of the sufficient statistic

$$T=X_1+\cdots+X_n$$

is given by

$$\text{(3.2)}\qquad g_0(t)=\frac{r^n n!}{(1-q^r)^n}\frac{S^t_{r,n}}{t!}p^t q^{rn}, \qquad t=n, n+1, \cdots$$

where, for all positive integers r, n, t, $S^t_{r,n}$ is defined by

$$\text{(3.3)}\qquad S^t_{r,n}=(-1)^{t-n}C^t_{-r,n}=\frac{1}{n!r^n}\sum_{k=1}^{n}(-1)^{n-k}\binom{n}{k}(rk+t-1)_t$$

and $C^t_{r,n}$ was defined in (2.3) (we prefer the $S^t_{r,n}$ to the $C^t_{-r,n}$ since the latter are alternately positive and negative integers, whereas $S^t_{r,n}$ are non-negative integers).

PROOF. We have

$$\text{(3.4)}\qquad g_0(t)=P\left[\sum_{i=1}^{n}X_i=t\right]=\sum_{(x_1,\cdots,x_n)}\prod_{i=1}^{n}\left[\frac{1}{1-q^r}\binom{r+x_i-1}{x_i}p^{x_i}q^r\right]$$

$$=\frac{\binom{rn+t-1}{t}p^t q^{rn}}{(1-q^r)^n}\sum_{(x_1,\cdots,x_n)}\frac{\binom{r+x_1-1}{x_1}\cdots\binom{r+x_n-1}{x_n}}{\binom{rn+t-1}{t}}$$

where the summation extends over all ordered n-tuples $(x_1, \cdots, x_n)$ of integers $x_i \geqq 1$ with $x_1+\cdots+x_n=t$.

In order to evaluate the last sum in (3.4) consider the following problem. Suppose that $t \geqq n$ indistinguishable balls are distributed at random in rn cells constituting n groups of r cells each. Then the probability that each group of cells contains at least one ball is given by the last sum in (3.4). This probability $P_{r,n}(t)$ say, can be found by applying the Poincaré formula as follows. Let A_k denote the event: the kth group of cells contains no ball. Then

$$P_{r,n}(t)=1-P\left[\bigcup_{k=1}^{n}A_k\right]=1-\sum_{k=1}^{n}(-1)^{k-1}\sum_{1\leq i_1<\cdots<i_k\leq n}P[A_{i_1}\cdots A_{i_k}]$$

$$=1-\sum_{k=1}^{n}(-1)^{k-1}\binom{n}{k}\frac{\binom{rn-rk+t-1}{t}}{\binom{rn+t-1}{t}}=\frac{n!r^n}{(rn+t-1)_t}S_{r,n}^t\,.$$

Hence (3.4) yields (3.2).

Remark. The probability distribution (3.2) of $T=\sum_{i=1}^{n}X_i$ can also be obtained easily by inverting the characteristic function of T_0. The characteristic function of (3.1) is found to be

$$\varphi_X(u)=\frac{q^r}{1-q^r}[(1-pe^{iu})^{-r}-1]$$

therefore that of $T=\sum_{i=1}^{n}X_i$ is

$$\varphi_T(u)=[\varphi_X(u)]^n=\frac{q^{rn}}{(1-q^r)^n}[(1-pe^{iu})^{-r}-1]^n\,.$$

Thus the probability distribution of T is given by

$$\begin{aligned}g_0(t)&=\frac{1}{2\pi}\int_{-\pi}^{\pi}\varphi_T(u)e^{-itu}du\\&=\frac{q^{rn}}{(1-q^r)^n}\frac{1}{2\pi}\int_{-\pi}^{\pi}\sum_{k=0}^{n}(-1)^{n-k}\binom{n}{k}(1-pe^{iu})^{-rk}e^{-itu}du\\&=\frac{q^{rn}}{(1-q^r)^n}\sum_{k=0}^{n}(-1)^{n-k}\binom{n}{k}\sum_{s=0}^{\infty}\binom{rk+s-1}{s}p^s\frac{1}{2\pi}\int_{-\pi}^{\pi}e^{-iu(t-s)}du\end{aligned}$$

and since

$$\int_{-\pi}^{\pi}e^{iu(s-t)}du=0 \qquad \text{for every } s\neq t$$

we get (3.2).

As noted in Section 1, the probability distribution (3.2) is obtained by Ahuja [1] by applying a results of Patil [8] for the generalized power series distribution.

3.2. UMVU estimator of p

For the construction of a UMVU estimator of p in (3.1) on the basis of a random sample from this distribution, it suffices to find an unbiased estimator of p which is a function of the complete sufficient statistic T.

Let $\tilde{p}_0(t)$ be such an unbiased estimator of p. Then the condition of unbiasedness gives the identity

$$\sum_{t=n}^{\infty} \tilde{p}_0(t) g_0(t) = p \qquad (0<p<1)$$

which by (3.2) after some simplification gives

$$\sum_{t=n}^{\infty} \tilde{p}_0(t) \frac{S^t_{r,n}}{t!} p^t = \sum_{t=n}^{\infty} \frac{S^t_{r,n}}{t!} p^{t+1} ;$$

this holds if, and only if

$$p_0(t) = t \frac{S^{t-1}_{r,n}}{S^t_{r,n}} . \tag{3.5}$$

Thus we have obtained the UMVU estimator $\tilde{p}_0(t)$ of p.

By the analogue of the recurrence relation (2.14) in terms of the numbers $S^t_{r,n}$, namely,

$$S^t_{r,n} = (rn - t - 1) S^{t-1}_{r,n} + S^{t-1}_{r,n-1} , \tag{3.6}$$

we can write $\tilde{p}_0(t)$ in the form

$$\tilde{p}_0(t) = \frac{t}{rn+t-1} \left(1 - \frac{S^{t-1}_{r,n-1}}{S^t_{r,n}}\right) . \tag{3.7}$$

It should be observed that $t/(rn+t-1)$, $\hat{p}$ say, is the UMVU estimator of p in the usual (non-truncated) negative binomial distribution. To obtain the UMVU estimator $\tilde{p}_0(t)$ we must multiply $\hat{p}$ by the correction factor (cf. (2.15))

$$\bar{S}(r, n, t) = 1 - \frac{S^{t-1}_{r,n-1}}{S^t_{r,n}} . \tag{3.8}$$

It is interesting to note here also, as in the binomial case, that the estimate of rp as $r \to \infty$ and $p \to 0$ so that $rp \to \lambda$ converges to the estimate $\tilde{\lambda}_0(t)$. This is expected in view of the corresponding convergence of the negative binomial (3.1) to the truncated Poisson at zero with parameter λ. Indeed, by the limiting property of $S^t_{r,n}$ which is the analogue of (2.12), that is,

$$\lim_{r\to\infty} \frac{S^t_{r,n}}{r^{t-n}} = \mathfrak{S}^n_t ,$$

we obtain:

$$r\tilde{p}_0(t) = t \frac{S^{t-1}_{r,n}/r^{t-n+1}}{S^t_{r,n}/r^{t-n}} \xrightarrow[r\to\infty]{} t \frac{\mathfrak{S}^{t-1}_n}{\mathfrak{S}^t_n} = \lambda_0(t) .$$

The ML estimator $\hat{p}$ of p, which can be found as the solution of the equation (see Patil [7])

$$\frac{\hat{p}}{(1-\hat{p})[1-(1-\hat{p})^r]}=\frac{t}{rn}$$

is obviously different than $\tilde{p}$. Therefore, by the uniqueness of the unbiased estimators based on t, $\hat{p}$ is a biased estimator of p.

The asymptotic variance of $\hat{p}$ is given by:

$$\operatorname{Var}(\hat{p})=\frac{p(1-p)^2[1-(1-p)^r]^2}{rn[1-(1+rp)(1-p)^r]} .$$

This is also the Cramér-Rao lower bound for the variances of unbiased estimators. We will show that there is no unbiased estimator whose variance attains this lower bound. Indeed, if $f(x_1,\cdots,x_n;p)$ denotes the joint probability function of n independent truncated negative binomial variables, then a necessary and sufficient condition for a Cramér-Rao estimator to exist is that there exists a function $g(p)$ such that the expression

$$p+g(p)\frac{\partial}{\partial p}\log f(x_1,\cdots,x_n;p)$$

is independent of p for all values of $(x_1,\cdots,x_n)$. Since

$$\frac{\partial}{\partial p}\log f(x_1,\cdots,x_n)=\frac{\sum x_i}{p}-\frac{rn(1-p)^{-r-1}}{(1-p)^{-r}-1} ,$$

no such function $g(p)$ exists. Therefore we may write:

$$\operatorname{Var}(\tilde{p})>\frac{p(1-p)^2[1-(1-p)^r]^2}{rn[1-(1+rp)(1-p)^r]} .$$

4. Asymptotic expressions for the estimates

For values of t which are large compared to n, approximate expressions for the estimates may be obtained by using the following asymptotic expressions:

$$C^t_{r,n}\sim\frac{(rn)_t}{r^n n!} , \qquad S^t_{r,n}\sim\frac{(rn+t-1)_t}{r^n n!} ;$$

these can easily be verified by considering the limits of $C^t_{r,n}/(rn)_t$ and $S^t_{r,n}/(rn+t-1)_t$ as $t\to\infty$. The estimates (2.13) and (3.7) can now be written in the form:

$$\tilde{\theta}_0(t)=\frac{t}{rn-t+1}\left(1-\frac{(rn-r)_{t-1}}{(rn-1)_{t-1}}\right)$$

and

$$\tilde{p}_0(t)=\frac{t}{rn+t-1}\left(1-\frac{(r(n-1)+t-2)_{t-1}}{(rn+t-1)_{t-1}}\right),$$

respectively.

UNIVERSITY OF ATHENS, GREECE

REFERENCES

[1] Ahuja, J. C. (1971). Distribution of the sum of independent decapitated negative binomial variables, *Ann. Math. Statist.*, **42**, 383-384.

[2] Cacoullos, T. (1961). A combinatorial derivation of the distribution of the truncated Poisson sufficient statistic, *Ann. Math. Statist.*, **32**, 904-905.

[3] Feller, W. (1968). *An Introduction to Probability Theory and Its Applications* 1, Third Ed., Wiley, New York.

[4] Haldane, J. B. S. (1945). On a method of estimating frequencies, *Biometrika*, **33**, 222-225.

[5] Fisher, R. A. (1936). The effects of method of ascertainment upon estimation of frequencies, *Ann. Eugen. London*, **6**, 13-25.

[6] Jordan, C. (1950). *Calculus of Finite Differences*, Chelsea, New York.

[7] Patil, G. P. (1962). Maximum likelihood estimation for generalized power series distributions and its applications to a truncated binomial distribution, *Biometrika*, **49**, 227-232.

[8] Patil, G. P. (1963). Minimum variance unbiased estimation and certain problems of additive number theory, *Ann. Math. Statist.*, **34**, 1050-1056.

[9] Rider, P. R. (1955). Truncated binomial and negative binomial distributions, *J. Amer. Statist. Ass.*, **50**, 877-883.

[10] Sampford, M. R. (1955). The truncated negative binomial distribution, *Biometrika*, **42**, 58-69.

[11] Tate, R. F. and Goen, R. L. (1958). Minimum variance unbiased estimation for the truncated Poisson distribution, *Ann. Math. Statist.*, **29**, 755-765.