Sankhyā : The Indian Journal of Statistics
1984, Volume 46, Series B, Pt. 3, pp. 301-305.

A NEW VARIANCE FORMULA FOR UNBIASED ESTIMATION IN INVERSE SAMPLING

By P. K. PATHAK

University of New Mexico and *Indian Statistical Institute*

and

Y. S. SATHE

University of Bombay

SUMMARY. Let T_r denote the unbiased minimum variance estimator of a proportion obtained by sampling until r successes turn up. Then

$$V(T_r) = pq\, E[(1+X_{r-1})^{-1}]$$

where X_{r-1} denotes the number of trials necessary for $(r-1)$ successes in a sequence of Bernoulli trials with probability of success p, $q = 1-p$. An important advantage of this formula is that it permits the power series expansion :

$$V(T_r) = pq \left[\sum_{j=1}^{r-k-1} (-1)^{j+1} \frac{p^j}{\binom{r-2}{j}} + (-1)^{r-k-1} \frac{(r-k)p^{r-k-1}}{\binom{r-2}{r-k-1}} E(k, r-k) \right]$$

where $E(k, r-k) = p^k q^{-r} \int_0^q y^{r-1} (1-y)^{-k}\, dy$. This expansion allows us to derive sharper bounds for $V(T_r)$ and can be used to evaluate $V(T_r)$ to any desired degree of accuracy.

1. Statement of the problem

Let $U = (u_1, u_2, \ldots)$ denote a sequence of Bernoulli trials with an unknown probability of success p, i.e., $P(u = 1) = p$ and consider the problem of estimating the parameter p by inverse sampling, i.e., the outcomes u_i's are observed sequentially until r successes turn up for the first time. Let X_r denote the number of trials required. Then $T_r = (r-1)/(X_r-1)$ is the customary unbiased minimum variance estimator of p. There is, however, no simple tractable expression for $V(T_r)$ which can be used to evaluate $V(T_r)$ accurately (cf. Best, 1974 and Mikulski and Smith, 1976 in this connection). The object of this note is to develop an exact expression for $V(T_r)$ which admits a simple power series expansion in p. It is shown that this power series can also be used to derive simple and sharper bounds for $V(T_r)$.

2. Derivation of the formula

Let V_r denote the variance of $T_r = (r-1)/(X_r-1)$ and $E(s, m) = E[(X_s+m)^{-1}]$, where X_s denotes the number of trials necessary for s successes in a sequence of Bernoulli trials with probability of success p. The following lemma plays a central role in the derivation of a simple formula for V_r.

AMS (1970) *subject classification* : Primary 62D05 ; Secondary 60C05.

Key words and phrases : Inverse sampling ; Unbiased estimation ; Variance expression ; Variance bound.

Lemma 2.1 : *For each positive integer $s \geqslant 1$,*

$$(m+s)E(s, m) = p+(m+1)qE(s, m+1), \quad \dots \quad (2.1)$$

and for each positive integer $s > 1$,

$$(s-1)E(s, m) = p-(m+1)pE(s-1, m+1). \quad \dots \quad (2.2)$$

Proof : It is clear that

$$\begin{aligned} 1-(m+s)E(s, m) &= E[(X_s-s)/(X_s+m)] \\ &= \sum_{j=s+1}^{\infty} \frac{(j-s)}{(j+m)} \begin{bmatrix} j-1 \\ s-1 \end{bmatrix} p^s q^{j-s} \\ &= q \sum_{j-1=s}^{\infty} \frac{(j-1)}{(j-1+m+1)} \begin{bmatrix} j-2 \\ s-1 \end{bmatrix} p^s q^{j-1-s} \\ &= qE[X_s/(X_s+m+1)] \\ &= q[1-(m+1)E(s, m+1)]. \quad \dots \quad (2.3) \end{aligned}$$

Formula (2.1) easily follows from (2.3). Formula (2.2) can be established on similar lines.

This lemma allows us to establish the following theorem :

Theorem 2.1 : *Let $r > 1$. Then*

$$V_r = pqE(r-1, 1). \quad \dots \quad (2.4)$$

Proof : It is easily seen that

$$V_r = (r-1)pE(r-1, 0)-p^2. \quad \dots \quad (2.5)$$

Letting $s = (r-1)$ and $m = 0$ in (2.1) and then replacing $E(r-1, 0)$ in terms of $E(r-1, 1)$ from (2.1) establishes (2.4).

Corollary 1 : *For $r > 2$ and for each $k \geqslant 1$, $1 \leqslant k \leqslant r-1$,*

$$V_r = pq \left[\sum_{j=1}^{r-k-1} (-1)^{j+1} \frac{p^j}{\begin{bmatrix} r-2 \\ j \end{bmatrix}} + (-1)^{r-k-1} \frac{(r-k)}{\begin{bmatrix} r-2 \\ r-k-1 \end{bmatrix}} p^{r-k-1} E(k, r-k) \right] \quad \dots \quad (2.6)$$

where

$$E(k, r-k) = p^k q^{-r} \int_0^q y^{r-1}(1-y)^{-k}\, dy. \quad \dots \quad (2.7)$$

The corollary is based on the repeated use of the recurrence relation given by (2.2) and the easily verifiable identity (2.7).

The corollary furnishes a simple power series expansion for V_r in increasing powers of p and can be easily used to compute V_r to any preassigned degree of accuracy. To get a rough idea of the remainder term in (2.6), note that

$$\begin{aligned} E(k, r-k) &= E[(X_k+r-k)^{-1}] \geqslant [E(X_k+r-k)]^{-1} \\ &= p/[r-(r-k)q]. \quad \dots \quad (2.8) \end{aligned}$$

3. SHARPER BOUNDS FOR V_r

Mikulski and Smith (1976), Sahai (1980) and Sathe (1977) have furnished upper and lower bounds for V_r. Their investigations show that asymptotically $V_r \approx p^2q/r$. We now proceed to demonstrate how our techniques can be used in a unified way to derive sharper bounds for V_r.

For example, letting $k = (r-1)$ and $k = (r-2)$ in (2.6) and (2.8) yields the following bounds

$$\frac{p^2q}{(r-q)} < V_r < \frac{p^2q}{(r-2q)}. \quad \ldots \quad (3.1)$$

These bounds are better than the Mikulski-Smith bounds, but worse than Sathe's bounds.

To obtain bounds which are sharper than those of (3.1), let us note that by the Cauchy-Schwarz inequality

$$E(k, r-k-1)E(k, r-k+1) > E^2(r, k). \quad \ldots \quad (3.2)$$

So, on evaluating $E(k, r-k \pm 1)$ in terms of $E(k, r-k)$ from (2.1) and simplifying, we get

$$q(k-1)E^2(k, r-k) - p(rp+kq)E(k, r-k) + p^2 \leqslant 0. \quad \ldots \quad (3.3)$$

Hence

$$E(k, r-k) > 2p/[r-(r-k)q + \sqrt{\{(r-(r-k+2)q)^2 + 4pq(r-k+1)\}}]. \quad \ldots \quad (3.4)$$

Now letting $k = (r-1)$ and $(r-2)$ in (3.4) and (2.6) leads to the following bounds for $r > 2$:

$$\frac{2p^2q}{[(r-q)+\sqrt{\{(r-3q)^2+8pq\}}]} < V_r < \frac{p^2q}{(r-2)}\left[1 - \frac{4p}{(r-2q)+\sqrt{\{(r-4q)^2+12pq\}}}\right]$$
$$\ldots \quad (3.5)$$

It is interesting to note that the lower bound to V_r in (3.5) is the same as that of Sathe (1977). His upper bound is

$$V_r < \frac{2p^2q}{(r-2q)+\sqrt{\{(r-2q)^2+4pq\}}}. \quad \ldots \quad (3.6)$$

A direct comparison of the upper bound in (3.5) with that of (3.6) shows that for $r = 3$, Sathe's bound is sharper, while for $r \geqslant 4$, the upper bound in (3.5) is sharper.

Next letting $k = (r-3)$ and $k = (r-4)$ in (3.4) and (2.6), we obtain the following bounds:

For $r \geqslant 4$,

$$V_r > \frac{p^2q}{(r-2)}\left[1 - \frac{2p}{(r-3)} + \frac{12p^2}{(r-3)(r-3q)+\sqrt{\{(r-5q)^2+16pq\}}}\right], \quad \ldots \quad (3.7)$$

and for $r \geqslant 5$,

$$V_r \leqslant \frac{p^2q}{(r-2)}\left[1-\frac{2p}{(r-3)}+\frac{6p^2}{(r-3)(r-4)}-\frac{48p^3}{(r-3)(r-4)[r-4q+\sqrt{\{(r-6q)^2+20pq\}}]}\right] \quad \ldots \quad (3.8)$$

A numerical comparison of the lower bound in (3.5) with that of (3.7) shows that for $r = 4$, (3.5) is sharper; for $r = 5$, (3.5) is sharper for $p \geqslant \cdot 292$; while for $r \geqslant 6$, (3.7) is sharper than (3.5). A similar numerical comparison of the upper bound in (3.8) with that of (3.5) shows that for $r = 5$; (3.5) is sharper; for $r = 6$, (3.5) is sharper for $p \geqslant \cdot 163$; for $r = 7$, (3.5) is sharper for $p \geqslant \cdot 565$; while for $r \geqslant 8$, (3.8) is sharper than (3.5).

It is interesting to note that a simple integration by parts of the formula in (2.7) yields the following recurrence relation

$$rV_r = pq-(r-1)(q/p)V_{r+1}. \quad \ldots \quad (3.9)$$

This equation can be used to obtain yet another system of bounds from the existing bounds. For example, using the lower bound for V_{r+1} from (3.5), we find that

$$V_r \leqslant \frac{pq}{r}-\left(1-\frac{1}{r}\right)\frac{2pq^2}{[(r+1-q)+\sqrt{\{(r+1-3q)^2+8pq\}}]}. \quad \ldots \quad (3.10)$$

Numerical comparisons of (3.10) with other upper bounds of this paper shows that (3.10) is sharper than (3.6), for $r = 3$ and 4, (3.10) is sharper than (3.5); while for each $r > 4$, there is a $p(r) > 0$ such that (3.5) is sharper than (3.10) for $p \leqslant p(r)$ with $p(r)$ approaching one in the limit.

From (3.7) and (3.9), one can also derive the following upper bound

$$V_r \leqslant \frac{p^2q}{r}\left[1+\frac{2q}{(r-2)}-\frac{12pq}{(r-2)(r+1-3q)+\sqrt{\{(r+1-5q)^2+16pq\}.)}}\right]. \quad \ldots \quad (3.11)$$

In our numerical investigation this bound has performed remarkably well. It is sharper than (3.5) for all $r \geqslant 3$. It is sharper than (3.6) for $r \geqslant 4$; while for $r = 3$, it is sharper than (3.6) for $p \geqslant \cdot 255$. For large values of r, the bound given by (3.8), is, however, sharper than (3.11). We have not included Sahai's upper bounds (Sahai, 1980) into this investigation since the derivation of his upper bounds is erroneous.

Likewise from (3.9) and (3.11), one gets the following lower bound

$$V_r \geqslant \frac{p^2q}{r}\left[1+\frac{2q}{(r+1)}+\frac{12q^2}{(r+1)(r+2-3q+\sqrt{\{(r+2-5q)^2+16pq\}})}\right]. \quad \ldots \quad (3.12)$$

This lower bound has also performed quite well in our numerical investigation. It is sharper than (3.5). It is also sharper than Sahai's (1980) lower bound. It is sharper than (3.7) for $p \geqslant p(r)$, where $p(r)$ depends on r and increases to $\cdot 43$ for $r \geqslant 27$.

Acknowledgements. The authors are grateful to Professor S. K. Mitra and the referee for their valuable remarks which have significantly improve the quality of this paper.

This research has been partially supported by the NSF Grant INT-8020450, and completed during the first author's sabbatical leave at the Indian Statistical Institute, New Delhi.

REFERENCES

BEST, D. J. (1974): The variance of the inverse binomial estimator. *Biometrika*, **61**, 385-386.

KENDALL, M. G. and STUART, A. (1978): *The Advanced Theory of Statistics*, Vol. 2, 2nd edition, London: Griffin.

MIKULSKI, P. W. and SMITH, P. J. (1976): A variance bound for unbiased estimation in inverse sampling. *Biometrika*, **63**, 216-127.

SAHAI, A. (1980): Improved variance bounds for unbiased estimation in inverse sampling. *J. Statist. Planning and Inference*, **4**, 213-216.

SATHE, Y. S. (1977): Sharper variance bounds for unbiased estimation in inverse sampling. *Biometrika*, **64**, 425-426.

Paper received: December, 1981.
Revised: April, 1983.