The Annals of Probability
1989, Vol. 17, No. 1, 396–400

THE TOTAL VARIATION DISTANCE BETWEEN THE BINOMIAL AND POISSON DISTRIBUTIONS

BY J. E. KENNEDY AND M. P. QUINE

University of Sydney

The exact total variation distances are obtained between a binomial distribution with parameters n and p and Poisson distributions with means np and $-n\log(1-p)$, for small values of p. It is shown that the latter distance is smaller for $0 < p < c_n$ and larger for $c_n < p < a'_{n0}$, where as $n \to \infty$, $nc_n \to 1.596\ldots$ and $na'_{n0} \to 3.414\ldots$.

1. Introduction. Let X be binomial with parameters n and p and Y be Poisson with parameter $n\nu$. We consider the total variation distance $d(X, Y) = \Sigma\{P(X = j) - P(Y = j)\}$, the sum being over those j for which $P(X = j) > P(Y = j)$, in two cases: $\nu = p$ and $\nu = -\log(1-p)$ $(= \lambda$, say), and denote the two distances $d(X, Y)$ by d_p and d_λ, respectively. We derive the exact distance for small values of p in each case and prove analogues to some of the asymptotic results of Deheuvels and Pfeifer (1986). The problem of computation of $d(X, Y)$ is also discussed.

2. The Poisson approximation for the binomial distribution. We will use the notation

$$f_j(\nu) = \binom{n}{j} p^j (1-p)^{n-j} - (n\nu)^j e^{-n\nu}/j!, \qquad j = 0, \ldots, n.$$

THEOREM 1. *For $n \geq 1$ and $0 < np \leq 2 - \sqrt{2}$,*

$$d_p = f_1(p), \tag{2.1}$$

and for $n \geq 1$ and $np \leq 1/(1 + p/12 + p^2/2)$,

$$d_\lambda = f_1(\lambda). \tag{2.2}$$

The equality in (2.1) *can be replaced by* $\geq$ *for $n \geq 1$ and $np < 2$, and that in* (2.2) *by* $\geq$ *for $n \geq 1$ and $0 < p < 1$.*

PROOF. Clearly $f_0(p) \leq 0$ and $f_1(p) > 0$ if and only if $-(n-1)\log(1-p) < np$. Now $-(n-1)\log(1-p) \leq (n-1)p(1-p/2)/(1-p) < np$ if $np < 1$. For $n \geq 2$, we use the fact that $f_2(p) \leq 0$ if and only if $A = \log(1 - 1/n) + (n-2)\log(1-p) \leq -np$. Now $A \leq -np + a + b$, with $na = 2np - 1 - (np)^2/2 \leq 0$ for $np \leq 2 - \sqrt{2}$ and $n^2 b = (np)^2 - \frac{1}{2} \leq 0$ for $np < 1/\sqrt{2}$, so $f_2(p) \leq 0$ if $np \leq 2 - \sqrt{2}$. For $n \geq 3$, we also need the following argument: We have

$$f_j(p) \leq n^{j-1} p^j \{(n-j+1)(1-p)^{n-j+1} - n(1-p)e^{-np}\}/\{j!(1-p)\}$$

Received July 1987; revised February 1988.
AMS 1980 *subject classifications*. Primary 60F05; secondary 62E20.
Key words and phrases. Total variation distance, binomial, Poisson.

with equality for $j = 1, 2$. Now $g(x) = x(1-p)^x - n(1-p)e^{-np}$ has a unique maximum and the preceding results imply $g(1) > 0$, $g(2) \le 0$ and it is easily seen that $g(n) < 0$ for $n \ge 3$ and $np < 1$. Thus $f_j(p) \le 0$ for $j = 2, \ldots, n$, and (2.1) follows.

Now consider (2.2). We have $f_0(\lambda) = 0$, and $f_1(\lambda) \ge 0$ from the standard inequality $(1+x)\log(1+x) \ge x$. For $n \ge 2$, we must see when $f_2(\lambda) \le 0$, or equivalently when $1 - 1/n \le \{(1-p)\log(1-p)\}^2/p^2$, which, using the inequality

$$(2.3) \qquad (1+x)\log(1+x) \le x + x^2/2 - x^3/6 + x^4/3,$$

is implied by $np \le 1/(1 + p/12 + p^2/2)$. For $n \ge 3$ we must also check when $f_j(\lambda) \le 0$ for $j = 3, \ldots, n$. But this is equivalent to

$$\frac{1}{j}\sum_{i=1}^{j-1} \log\left(1 - \frac{i}{n}\right) \le \log\{-(1-p)\log(1-p)\} - \log p$$

and the left-hand side is easily seen to be decreasing in j so that $f_2(\lambda) \le 0$ implies $f_j(\lambda) \le 0$, $j = 3, \ldots, n$. Thus we have established (2.2). The rest of the proof is straightforward. □

3. Comparisons and computational aspects. In this section we give some inequalities on the roots of those equations used for computing the total variation distance for any values of n and p, and in addition use these to give a precise comparison between d_p and d_λ. For convenience of exposition we set $g_j(p) = f_j(\lambda)$, $j = 0, \ldots, n$. The following lemma is crucial.

LEMMA 3.1. *Let* $0 < p < 1$ *and* $n \ge 7$. *Then*

$$(3.1) \qquad f_0(p) < 0;$$

$$(3.2) \qquad f_1(p) = 0 \text{ has exactly one root } a_{n,2} \text{ and } f_1'(a_{n,2}) < 0;$$

$$f_j(p) = 0 \text{ has exactly two roots denoted}$$

$$(3.3) \qquad a_{n,j-2} < a'_{n,j-2} \quad \text{for } j = 2,3,$$
$$a_{n,j-1} < a'_{n,j-1} \quad \text{for } j = 4, \ldots, n-1, \text{ and}$$

$$f_j'(a_{n,l}) > 0, \qquad f_j'(a'_{n,l}) < 0, \qquad \begin{cases} l = j-2, & j = 2,3, \\ l = j-1, & j = 4, \ldots, n-1; \end{cases}$$

$$(3.4) \qquad f_n(p) = 0 \text{ has exactly one root } a_{n,n-1} \text{ with } f_n'(a_{n,n-1}) > 0;$$

$$(3.5) \qquad a_{n,0} < a_{n,1} < \cdots < a_{n,n-1}, \qquad a'_{n,0} < a'_{n,1} < a'_{n,3} < \cdots < a'_{n,n-2};$$

$$(3.6) \qquad a_{n,4} < a'_{n,0} < a_{n,5} < a'_1;$$

$$(3.7) \qquad g_0(p) = 0, \qquad g_1(p) > 0;$$

$$(3.8) \qquad g_j(p) = 0 \text{ has exactly one root } b_{n,j-2} \text{ and } g_j'(b_{n,j-2}) > 0 \text{ for } j = 2, \ldots, n;$$

$$(3.9) \qquad b_{n,0} < b_{n,1} < \cdots < b_{n,n-2};$$

$$(3.10) \qquad 0 < a_{n,0} < b_{n,0} < a_{n,1} < a_{n,2} < b_{n,1}.$$

Note that similar results are easy to compute for $2 \leq n \leq 6$. We defer a sketch of the proof of this lemma to the Appendix and concentrate here on its implications:

1. In calculating d_p, (3.1)–(3.6) imply that $d_p = f_1(p)$ for $0 < p \leq a_{n,0}$,

$$d_p = \begin{cases} \sum_{i=1}^{j+2} f_i(p), & a_{n,j} < p \leq a_{n,j+1},\ j = 0, 1, \\ \sum_{i=2}^{j+1} f_i(p), & a_{n,j} < p \leq a_{n,j+1},\ j = 2, 3, \end{cases}$$

and $d_p = \sum_{i=2}^{5} f_i(p)$ for $a_{n,4} < p \leq a'_{n,0}$. For computation purposes, the $a_{n,j}$'s are unnecessary: One simply adds up consecutive positive $f_j(p)$'s.

2. Similarly, (3.7)–(3.9) imply that $d_\lambda = f_1(\lambda)$ for $0 < p \leq b_{n,0}$ and for $k = 0, \ldots, n-2$, $d_\lambda = \sum_{j=1}^{k+2} f_j(\lambda)$ for $b_{n,k} < p \leq b_{n,k+1}$. This leads to an elementary proof of the result [see Deheuvels and Pfeifer (1986), Theorem 1.3] that as $n \to \infty$, $np \to a \in (0, \infty)$,

$$d_\lambda \sim \frac{1}{2} np^2 \frac{a^{[a]}}{[a]!} e^{-a}.$$

3. If $n \geq 100$ and $np \leq 0.999$, then $np \leq 1/(1 + p/12 + p^2/2)$ so (2.2) is satisfied. Because of (3.8), the fact (established directly) that $g_2(0.999/n) < 0$ for $n = 84, \ldots, 99$ shows that (2.2) holds for $np \leq 0.999$ and $n \geq 84$. This method will work with 0.999 replaced by any $\theta < 1$.

4. By looking at the explicit forms of d_λ and d_p for p in each of the intervals implicit in (3.10), we have the following complement (in the iid case) to the asymptotic result of Deheuvels and Pfeifer [(1986), Corollary 2.1]:

$$d_\lambda < d_p, \qquad 0 < p < c_n,$$
$$d_\lambda > d_p, \qquad c_n < p < a'_{n0},$$

where $c_n \in (a_{n,1}, a_{n,2})$ and as $n \to \infty$,

$$nc_n \to 1 + (\sqrt{2} + 1)^{1/3} - (\sqrt{2} - 1)^{1/3}.$$

5. It is straightforward to check that as $n \to \infty$,

$$na_{n,0} \to 2 - \sqrt{2}, \qquad na'_{n,0} \to 2 + \sqrt{2},$$
$$na_{n,i} \to i + 2 - \sqrt{(i+2)}, \qquad na'_{n,i} \to i + 2 + \sqrt{(i+2)}, \qquad i = 1, 2,$$
$$na_{n,i} \to i + 1 - \sqrt{(i+1)}, \qquad na'_{n,i} \to i + 1 + \sqrt{(i+1)}, \qquad i \geq 3,$$

and

$$nb_{n,i} \to i + 1, \qquad i \geq 0.$$

6. Two of the best-known bounds available for d_p are $d_p \leq \frac{1}{2}p/\sqrt{(1-p)} = R_p$ and $d_p \leq (1 - e^{-np})p = B_p$, due to Romanowska (1977) and Barbour and Hall (1984), respectively. The bounds are exhibited for the case $n = 10$ in Figure 1.

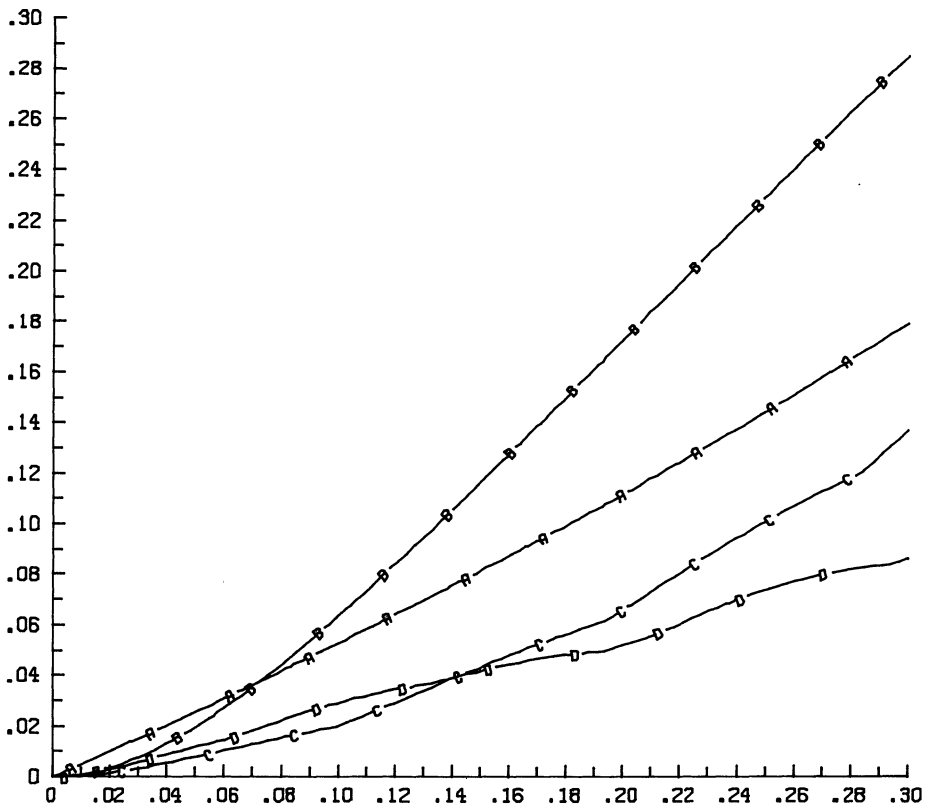

FIG. 1. $A = R_p$, $B = B_p$, $C = d_\lambda$ *and* $D = d_p$.

APPENDIX

PROOF OF LEMMA 3.1. Consider for $j = 1, \ldots, n$ the function $f_j^*(p) = j! f_j(p)/(np)^j$. Clearly, $f_j(p)$ and $f_j^*(p)$ have the same sign, and for $j = 1, \ldots, n-1$,

$$f_j^{*\prime}(p) = -n f_{j+1}^*(p), \tag{A1}$$

$$f_j^*(0) < 0, \quad j = 2, \ldots, n, \qquad f_j^*(1) < 0, \quad j = 1, \ldots, n-1, \qquad f_n^*(1) > 0, \tag{A2}$$

and $f_n^*(p) = 0$ has only one root, namely $a_{n,n-1} = \log(n^n/n!)/n$, which lies in $(0,1)$. Now $f_{n-1}^*(a_{n,n-1}) > n! n^{-n}(\log(2\pi n) - 2)/2 > 0$ for $n \geq 2$. Thus using (A1) and (A2) it follows that $f_{n-1}^*(p)$ has exactly two roots $a_{n,n-2}, a'_{n,n-2}$ in $(0,1)$ such that $a_{n,n-2} < a_{n,n-1} < a'_{n,n-2}$. Now consider $f_{n-2}^*(p)$. By (A1) we know that $f_{n-2}^*(p)$ is increasing on $(0, a_{n,n-2})$ and $(a'_{n,n-2}, 1)$ and decreasing on $(a_{n,n-2}, a'_{n,n-2})$. We claim $f_{n-2}^*(a_{n,n-2}) > 0$ and $f_{n-2}^*(a'_{n,n-2}) < 0$. For otherwise we must have $f_{n-2}^*(p) < 0$ for $p \in (0,1)$. Then by (A1), $f_{n-3}^*(p)$ is always

increasing. However, $f^*_{n-3}(p)$ is negative at $p = 1$. So this would mean $f^*_{n-3}(p) < 0$ for $p \in (0,1)$ as well. In view of (A1) and (A2) this argument can be repeated for $j = n - 4$ down to $j = 2$. But that $f^*_2(1/n) > 0$ is easily verified. Thus we have a contradiction. Hence $f^*_{n-2}(p)$ has two roots $a_{n,n-3}, a'_{n,n-3}$ in $(0,1)$ such that $a_{n,n-3} < a_{n,n-2} < a'_{n,n-3} < a'_{n,n-2}$. Clearly similar reasoning yields (3.3). Since $f^*_1(0) = 0$, $f^*_1(p)$ has only one root in $(0,1)$. Note that, in addition, we have the relationship

$$a_{n,j-1} < a_{n,j} < a'_{n,j-1} < a'_{n,j} \quad \text{for } j \geq 3.$$

To complete the proof of (3.6) it suffices to show $a_{n4} < 3/n < a'_{n0} < 3.5/n < a_{n5}$, which can be done by checking that $f_5(3/n) > 0$, $f_2(3/n) > 0$, $f_2(3.5/n) < 0$, $f_6(3.5/n) < 0$, $f_6(4/n) > 0$ and $f_3(4/n) > 0$. Now $f_5(3/n) = (1 - 1/n)\cdots(1 - 4/n)(1 - 3/n)^{n-5} - e^{-3}$. Taking logs and using $x(1 + x/2)/(1 + x) < \log(1 + x) < x(1 - x/2)$, $-1 < x \leq 0$, gives $f_5(3/n) > 0$. Similar arguments prove $f_2(3/n) > 0$ and similar but messier arguments using (2.3) prove the remaining four inequalities.

Now consider (3.8)–(3.9). We show each $g_j(p)$ $(2 \leq j \leq n)$ has at least one root in $(0,1)$ by checking that $g_j(p)$ changes sign. We use (2.3) to show $g_2(1/2n) < 0$, and $g_j(1/2n) < 0$ for $2 < j \leq n$ follows by induction. Also $g_n(1 - 1/n) = (1 - 1/n)^n - (\log n)^n/n! > 0$ and $g_j(1 - 1/n) > 0$ for $2 \leq j < n$ by induction. Now write $g_j(p) = (1 - p)^{n-j}n^j h_j(p)/j$. To establish that for $2 \leq j \leq n$, $g_j(p)$ has a unique root in $(0,1)$, it suffices to show that for any $b_{n,j-2}$ such that $g_j(b_{n,j-2}) = 0$, we have $h'_j(b_{n,j-2}) > 0$, which can be done by writing $h'_j(b_{n,j-2})$ as a function of j, n and $b_{n,j-2}$. Finally to prove $b_{nj} < b_{n,j+1}$ for $j = 0, \ldots, n - 3$ we show $g_{j+2}(b_{n,j+1}) > 0$. Now $g_{j+3}(b_{n,j+1}) = 0$, and it follows that

$$g_{j+2}(b_{n,j+1}) = b_{n,j+1}^{j+2}\Big\{(n-1)\cdots(n-j-1) - n^{j+1}\big[(n-1)\cdots(n-j-2)/n^{j+2}\big]^{(j+2)/(j+3)}\Big\},$$

which is easily shown to be positive. The proof of (3.10) follows the same lines. □

Acknowledgment. The authors are grateful to the referee for helpful comments which led in particular to improvements in the statement and proof of Theorem 1.

REFERENCES

BARBOUR, A. D. and HALL, P. (1984). On the rate of Poisson convergence. *Math. Proc. Cambridge Philos. Soc.* **95** 473–480.

DEHEUVELS, P. and PFEIFER, D. (1986). A semi-group approach to Poisson approximation. *Ann. Probab.* **14** 663–676.

ROMANOWSKA, M. (1977). A note on the upper bound for the distance in total variation between the binomial and the Poisson distribution. *Statist. Neerlandica* **31** 127–130.

STATISTICAL LABORATORY
UNIVERSITY OF CAMBRIDGE
16 MILL LANE
CAMBRIDGE CB2 1SB
ENGLAND

DEPARTMENT OF MATHEMATICAL STATISTICS
UNIVERSITY OF SYDNEY
NEW SOUTH WALES 2006
AUSTRALIA