PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Prediction for Small Subgroups

D. R. Cox and A. C. Davison

Phil. Trans. R. Soc. Lond. B 1989 **325**, 185-187
doi: 10.1098/rstb.1989.0084

Phil. Trans. R. Soc. Lond. B **325**, 185–187 (1989) [185]
Printed in Great Britain

PREDICTION FOR SMALL SUBGROUPS

By D. R. COX, F.R.S.[1], and A. C. DAVISON[2]

[1] *Nuffield College, Oxford OX1 1NF, U.K.*
[2] *Department of Mathematics, Imperial College, London SW7 2BZ, U.K.*

Prediction limits are calculated for the number of events likely to occur in a specified time period in an exponentially growing epidemic. The basis for the prediction is the total number of events observed in the past.

1. Introduction

In predicting the number of new cases of AIDS to be diagnosed in some future period, there are essentially three sources of error. These arise respectively from:

(*a*) use of an incorrect empirical formula or model on which to base the prediction;

(*b*) errors in estimating unknown parameters in the model, e.g. errors in estimating the doubling time were exponential growth to be assumed;

(*c*) Poisson-distributed variations in observed numbers to be anticipated even if the systematic part of the variation were to be correctly specified.

For predicting some way ahead with fairly large numbers, the first type of error is likely to predominate. This is the case, for example, in predicting the number of AIDS cases in England and Wales for three or four years ahead. This is essentially because many different shapes of curve are consistent with the data on which the prediction is based. For rather short-term prediction (*a*) becomes less important and ultimately, when predicting rather small numbers a short time ahead, Poisson type errors, (*c*), will be the major source of uncertainty. This is particularly relevant if we wish to predict events within a small geographical area for a rather short time ahead.

One approach to such a prediction would be based on a careful analysis of the spatial distribution of the epidemic leading to a small area prediction formula based on local characteristics and experience. Although this could be of considerable interest, here we develop a much simpler approach in which the total number of cases diagnosed up to the current point is used as a basis for prediction.

The central idea is to assume that the broad pattern of growth is determined via a large body of data and is essentially the same throughout but that the local rate varies from place to place and can be determined only via the total number of cases observed locally.

2. Formulation

Suppose then that in the area under study, diagnoses occur in a Poisson process of rate $b\lambda\, \mathrm{e}^{bt}$, where b is a known constant but λ is unknown. Thus the doubling time is $0.693/b$ years. Suppose further that in the area or subgroup in question n_0 cases have been diagnosed up to time t_0 and that it is required to predict the number n' of new cases to be diagnosed in a future time period (t', t''). The arguments apply with minor modification if the rate is $\lambda h(t)$, where $h(t)$ is any known function of time.

Define a new time scale, which we call operational time, by

$$s(t) = \mathrm{e}^{bt}.$$

On this time scale cases occur in a Poisson process of constant rate λ and the key time points $-\infty, t_0, t', t''$ are transformed to $0, s_0 = \mathrm{e}^{bt_0}, s' = \mathrm{e}^{bt'}, s'' = \mathrm{e}^{bt''}$.

The value n' to be predicted is such that the split (n_0, n') is consistent with a binomial distribution with probability of 'success'

$$(s''-s')/(s_0+s''-s') = p,$$

say; this is known.

Thus the upper and lower $1-2\epsilon$ limits for n' are respectively the largest n^* and the smallest n_* such that

$$\epsilon \leqslant \sum_{r=n^*}^{n_0+n^*} (1-p)^{n_0+n^*-r} p^r \binom{n_0+n^*}{r},$$

$$\epsilon \leqslant \sum_{r=0}^{n_*} (1-p)^{n_0+n^*-r} p^r \binom{n_0+n_*}{r}.$$

The values of n^* and n_* can be found numerically as set out below. A simple point estimate of n' is found by writing

$$n'/(n_0+n') = p$$

and solving for n'.

The values of n^* and n_* are most straightforwardly found for given n_0 and p by enumeration of the probabilities:

$$p^* = \sum_{r=n^*}^{n_0+n^*} p^r (1-p)^{n_0+n^*-r} \binom{n_0+n^*}{r},$$

and

$$p_* = \sum_{r=0}^{n^*} p^r (1-p)^{n_0+n''-r} \binom{n_0+n_*}{r}.$$

Direct calculation is tedious, and a large saving in computer time is obtained by noting that

$$p^* = I_p(n^*, n_0+1) \quad \text{and} \quad p_* = I_{1-p}(n_0, n_*+1),$$

where

$$I_x(a,b) = \frac{1}{B(a,b)} \int_0^x t^{a-1}(1-t)^{b-1}\,\mathrm{d}t$$

is the incomplete beta function. Press *et al.* (1986) give FORTRAN and PASCAL algorithms for numerical evaluation of $I_x(a,b)$.

For large values of n_0 a normal approximation is available. Suppose that X is a binomial random variable with index n_0+n and probability p. Then approximately,

$$\mathrm{pr}(X \leqslant n) = \Phi\left\{\frac{n-(n+n_0)\,p}{\surd[(n+n_0)\,p(1-p)]}\right\} = g(n),$$

say, where $\Phi(.)$ is the standard normal cumulative distribution function. The values n_* and n^* roughly satisfy the equations $g(n) = \epsilon$ and $g(n) = 1-\epsilon$. Thus they are the solutions of

$$An^2+Bn+C = 0,$$

where $A = (1-p)^2$, $B = p(p-1)\,(2n+z_\epsilon^2)$, and $C = np[np-(1-p)\,z_\epsilon^2]$.

Here $\Phi(z_\epsilon) = \epsilon$, and z_ϵ is found from tables of the normal distribution.

PASCAL programs to calculate exact and approximate values of n^* and n_* are available from the second author.

3. Numerical results

Table 1 shows the 90% prediction limits obtained from the formulae of §2; three values of doubling time have been used. The main qualitative conclusion to be drawn from the results is the surprising insensitivity to the doubling time, confirming the remarks in §1.

Table 1. Local predictions of numbers of events likely to arise in next year and the year after next

(Doubling time A: 1.25 years, B: 1.75 years, C: 2.5 years.)

number of events observed so far				90% prediction limits					
	next year			year after next			next 2 years combined		
	A	*B*	*C*	*A*	*B*	*C*	*A*	*B*	*C*
0	(0, 3)	(0, 2)	(0, 2)	(0, 5)	(0, 3)	(0, 2)	(0, 7)	(0, 4)	(0, 3)
2	(0, 6)	(0, 4)	(0, 3)	(0, 10)	(0, 6)	(0, 4)	(0, 14)	(0, 9)	(0, 6)
4	(0, 8)	(0, 6)	(0, 4)	(1, 14)	(0, 8)	(0, 5)	(2, 20)	(1, 13)	(0, 8)
6	(1, 13)	(0, 8)	(0, 5)	(2, 17)	(1, 10)	(0, 7)	(4, 26)	(2, 16)	(1, 11)
8	(1, 13)	(1, 9)	(0, 6)	(3, 21)	(1, 13)	(0, 8)	(6, 32)	(3, 20)	(1, 13)
10	(2, 15)	(1, 10)	(0, 7)	(5, 24)	(2, 15)	(1, 9)	(9, 37)	(5, 23)	(2, 15)
15	(5, 20)	(3, 14)	(1, 10)	(10, 33)	(5, 20)	(2, 12)	(16, 50)	(9, 31)	(5, 20)
20	(7, 25)	(4, 17)	(2, 12)	(14, 41)	(7, 24)	(3, 15)	(24, 63)	(13, 39)	(7, 25)
25	(10, 29)	(6, 20)	(3, 14)	(19, 49)	(10, 29)	(5, 18)	(32, 75)	(18, 46)	(10, 29)
30	(13, 34)	(8, 23)	(4, 16)	(24, 57)	(12, 33)	(6, 21)	(40, 87)	(23, 53)	(13, 34)
40	(19, 43)	(9, 26)	(6, 20)	(35, 72)	(18, 42)	(9, 26)	(58, 110)	(33, 67)	(19, 43)
50	(25, 52)	(11, 29)	(9, 24)	(46, 67)	(24, 51)	(13, 31)	(75, 133)	(43, 81)	(25, 51)
60	(31, 60)	(18, 41)	(11, 28)	(57, 102)	(30, 59)	(16, 36)	(93, 156)	(53, 95)	(31, 60)

The simple quadratic approximation at the end of §2 gives adequate accuracy provided that the lower limit is at least 10.

As an example, in Wales up to 30 June 1988, 27 cases had been reported so that, taking a local doubling time of 1.25 years, anything between 12 and 30 new cases could be expected in the following year.

Reference

Press, W. H., Flannery, B. P., Teukolsky, S. A. & Vetterling, W. T. 1986 *Numerical recipes: the art of scientific computing*. Cambridge University Press.