STATISTICS

The Meaning of Binomial Distribution

Two generalizations of the simple binomial distribution are common in statistical text-books, one due to W. Lexis and the other to S. D. Poisson. Lexis considered the case in which the probability of an event occurring, p , is constant in the N trials of one experiment, but varies among several such experiments. He showed that the mean and variance of the resulting binomial distribution are $N\bar{p}$ and $N\bar{p}\bar{q} + N(N-1)V(p)$, where $\bar{p} = 1 - \bar{q}$ is the mean, and $V(p)$ the variance, of p between experiments. The variance thus exceeds that of the simple binomial distribution with the same mean. Poisson considered the case in which p takes the value p_i at the i th trial in each experiment, and showed that the mean and variance of the resulting distribution are $N\bar{p}$ and $N\bar{p}\bar{q} - NV'(p)$, where $V'(p)$ is the variance of p within experiments.

In order to specify the exact form of the modified binomial distribution, it is necessary to postulate a distribution for p , in the first case, and the values p_i in the second. A convenient distribution for p is the β -distribution, and the resulting Lexian binomial has been described and used by Edwards¹ in connexion with work on the human sex ratio. The Poisson binomial is not commonly used, although to do so would be a simple, if tedious, matter. Parkes² gives some data on the sex ratio in pigs which exhibit unexpectedly small variances, and if he had known of this distribution he need not have observed that his was "a finding very hard to interpret".

David³ describes Lexian and Poissonian variation adequately, including the case in which both types occur simultaneously. But she perpetrates a common fallacy when she concludes that if the observed variance of a binomial distribution does not differ significantly from $N\bar{p}\bar{q}$ then "the probability may be assumed to be constant throughout the observations". For the two types of variation may be masking each other, and the only valid inference is that there is no evidence for variations in the probability.

I have recently found a third generalization of the simple binomial distribution useful. Suppose that the probability is p at each trial, but that the result of a trial is causally correlated to the result of the previous one, with correlation coefficient r . Then, formulating the problem like a Markov process, the matrix of transition probabilities from one trial to the next is:

$$\begin{pmatrix} p + rq & q - rq \\ p - rp & q + rp \end{pmatrix}$$

and the generating function of the 'Markov binomial' distribution is:

$$G(s) = (ps \ q) \begin{pmatrix} p + rq & q - rq \\ p - rp & q + rp \end{pmatrix}^{N-1} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

where s is a dummy variable.

The matrix of transition probabilities has the interesting property:

$$\begin{pmatrix} p + rq & q - rq \\ p - rp & q + rp \end{pmatrix}^N = \begin{pmatrix} p + r^Nq & q - r^Nq \\ p - r^Np & q + r^Np \end{pmatrix}$$

Thus the correlation coefficient between trials separated by $N-1$ other trials is r^N . The coefficient is, in fact, the determinant of the matrix.

The mean of the distribution is clearly Np , and the variance can be found from the formula for the variance of the sum of N correlated variables:

$$\begin{aligned} \sigma^2 &= Npq + 2(N-1)rpq + 2(N-2)r^2pq + \\ &\quad 2(N-3)r^3pq \dots + 2r^{N-1}pq \\ &= Npq + 2pq \cdot \frac{r}{(1-r)^2} (N(1-r) - 1 + r^N) \end{aligned}$$

This variance exceeds Npq if the correlation is positive, but is less than Npq if the correlation is negative.

In practice it is impossible to separate out the three effects here described when the data are solely in the form of binomial distributions; but if the data are available broken down into series, an analysis may be made:

It can be shown that the correlation coefficient between successive trials is equal to:

$$r + (1-r) \frac{V(p)}{pq}$$

where r is the correlation coefficient between successive trials within experiments and $V(p)$ is the variance of p between experiments. Small variations in p between trials are second-order effects in the expression and may be neglected. Thus the correlation coefficient between trials separated by $(N-1)$ other trials is:

$$r^N + (1-r^N) \frac{V(p)}{pq}$$

and this formula enables V and r to be estimated. Variation in p between trials may, of course, be estimated directly.

We are thus equipped to analyse binomial data on the fairly comprehensive hypothesis that the probability of an event occurring at the i th trial in the j th experiment is $p_i + p_j$, where the p_i 's are small and $\sum_i p_i = 0$, there being an additional real correlation, with coefficient r , between the results of successive trials in an experiment. Further generalization may, of course, be made, at the risk of some complexity, but it is doubtful whether any data exist that warrant it.

In conclusion, it must again be emphasized that no useful inferences can be made from the fact that data fit the simple binomial distribution satisfactorily; for, on one hand, variability of p between experiments, and a positive correlation between successive trials, increase the variance, and on the other, variability of p between trials, and a negative correlation, reduce it.

A. W. F. EDWARDS

Department of Genetics,
University of Cambridge.
Feb. 12.

¹ Edwards, A. W. F., *Ann. Hum. Genet.*, **23**, 6 (1958).

² Parkes, A. S., *Biometrika*, **15**, 373 (1923).

³ David, F. N., "Probability Theory for Statistical Methods" (Cambridge University Press, 1951).