APPLICATIONES MATHEMATICAE
22,3 (1994), pp. 331–337

**W. KÜHNE (Dresden), P. NEUMANN (Dresden),
D. STOYAN (Freiberg) and H. STOYAN (Freiberg)**

PAIRS OF SUCCESSES IN BERNOULLI TRIALS AND A NEW n-ESTIMATOR FOR THE BINOMIAL DISTRIBUTION

Abstract. The problem of estimating the number, n, of trials, given a sequence of k independent success counts obtained by replicating the n-trial experiment is reconsidered in this paper. In contrast to existing methods it is assumed here that more information than usual is available: not only the numbers of successes are given but also the number of pairs of consecutive successes. This assumption is realistic in a class of problems of spatial statistics. There typically $k = 1$, in which case the classical estimators cannot be used. The quality of the new estimator is analysed and, for $k > 1$, compared with that of a classical n-estimator. The theoretical basis for this is the distribution of the number of success pairs in Bernoulli trials, which can be determined by an elementary Markov chain argument.

1. Introduction. The standard statistical problem associated with the binomial distribution is that of estimating its probability, p, of success.

A much less well studied and considerably harder problem is that of estimating the number, n, of trials. The papers by Olkin, Petkau and Zidek [6] and Carroll and Lombard [3] study this problem for the case where k independent success counts $s_1, \ldots, s_k$ are given. Their methods cannot be applied if only one count is considered, $k = 1$. But just this case appears in some problems of spatial statistics.

An important application consists in estimating the fraction of chips on a silicon wafer which are faulty because of technological reasons. In a general setting the spatial problem is as follows. A rectangle is divided into

1991 *Mathematics Subject Classification*: 62E25, 62F10, 62M30.
Key words and phrases: binomial distribution, Markov chain, n-estimator, silicon wafer, simulation.

$M \times N$ cells. A fraction, f, of these cells has a property F. For example, these cells are chips which are technologically faulty, or these cells represent areas in a forest where certain mushrooms cannot live. The other cells have independently from one another a property R or not. The probability that a cell which does not have the F-property has property R is p. Cells which have this property are considered to be "successes". Cells which do not belong to the R-class cannot be discriminated from the cells with property F.

In our examples, "successes" represent chips free of failures or areas in which mushrooms are detected. Our problem is estimating the number $n = M \times N \times (1 - f)$. If n is estimated, then we determine p by dividing the total number of successes by the estimated n.

We assume that the union of all cells with property F forms an unknown connected subarea of perhaps elliptical shape in the rectangle.

The estimation method bases on a success count procedure in the whole rectangle. We count all successes and all *pairs* of consecutive successes which appear in the same horizontal line of cells in the rectangle; see Fig. 1.

F F F F F F F F F F F F F F F F
F F F F F F F F ∗ F F F F F F F
F F F F F F ∗ F F F F F F F
F F F F ∗ F F F F F F F F
F F F ∗ ∗ ∗ ∗ F F F F F F F
F F F ∗ ∗ F F F F F F F F
F F ∗ ∗ F F F F F F F F
F ∗ ∗ F F F F F F F F F
F F ∗ F F F F F F F F F
F F F ∗ F F F F F F F F
F F F ∗ ∗ F F F F F F F
F F F ∗ ∗ F F F F F F F
F F F ∗ ∗ F F F F F F F
F F F F ∗ F F F F F F F
F F F F F ∗ F F F F F F
F F F F F F F F F F F F

Fig. 1. Cells with property F (F), R (empty), and not-R ($*$). The aim of the statistical procedure is the estimation of the number of cells without "F", where F-cells and $*$-cells cannot be discriminated

In the following we do not continue the discussion of the spatial statistical problem. Of course, for it our estimation procedure is an approximation only because of edge effects at the boundary of the subregions of cells with and without property F. Probably, methods for restoring dirty images could yield an adequate solution, in particular Bayesian inference methods; see Besag and Green [1] and Besag, York and Mollié [2]. For further discussion of the chip problem we refer to Kühne [5].

Here we discuss the problem for the binomial distribution assuming that information about pairs of successes is available.

2. The number of pairs of successes. It is well known that the number of successes in n trials has a binomial distribution. But what about the number of consecutive pairs in n trials? (We repeat that in a series of three successes we count two pairs and in a series of four successes three pairs. In contrast, in the probabilistic literature usually "runs" are considered, i.e. series of consecutive successes. But see also Janson [4].) It seems to be difficult to give a simple formula for the probabilities $p_{n,l}$ of having l pairs of successes in n trials, $l = 0, 1, \ldots, n-1$. But, nevertheless, these probabilities can be calculated analytically by means of a simple iteration procedure.

For this purpose, let us consider the following Markov chain. It has the states E_l and M_l, $l = 0, 1, \ldots$, where

$$E_l = l \text{ success pairs and the last trial was a success,}$$
$$M_l = l \text{ success pairs and the last trial was not a success.}$$

Let $e_{n,l}$ be the probability that after the nth trial the chain is in E_l and $m_{n,l}$ the corresponding one for M_l. For these probabilities the following recurrence relation is true:

$$e_{n+1,l} = pe_{n,l-1} + pm_{n,l}, \qquad e_{n,-1} \equiv 0, \tag{2.1}$$

$$m_{n+1,l} = (1-p)e_{n,l} + (1-p)m_{n,l}, \qquad l = 1, 2, \ldots \tag{2.2}$$

For $n = 3$ we have

$$e_{3,0} = p(1-p), \qquad m_{3,0} = (1-p)^3 + 2p(1-p)^2,$$
$$e_{3,1} = p^2(1-p), \qquad m_{3,1} = p^2(1-p).$$
$$e_{3,2} = p^3,$$

The other probabilities are zero.

Clearly,

$$p_{n,l} = e_{n,l} + m_{n,l}.$$

Analogously, the joint distribution of the numbers of successes and pairs of successes can be determined. The states of the corresponding more complicated Markov chain describe both numbers. This distribution is useful for the investigation of the estimator $\widehat{n}$ in the next section.

Of particular importance for the investigation of this estimator is the probability, $p_{n,0}$, that in n trials there is no pair of successes. It can be separately determined, without using the formulae (2.1) and (2.2).

Consider for this purpose a Markov chain which describes the behaviour before first appearance of a pair of successes. It has the states E, M and D. E means that the last Bernoulli trial was a success, M that it was not

a success, and D that it was a success following a success. If D is entered, then the first pair of successes is obtained. This state is an absorbing state. The one-step transition probabilities p_{ij} are

i \ j	E	M	D
E	0	$1-p$	p
M	p	$1-p$	0
D	0	0	1

Let the state probabilities of this Markov chain be e_n, m_n and d_n. Then

$$p_{n,0} = 1 - d_n.$$

The recursive relations

$$\begin{aligned} e_{n+1} &= p \cdot m_n, \quad e_1 = p, \\ m_{n+1} &= (1-p)(e_n + m_n), \quad m_1 = 1-p, \end{aligned} \tag{2.3}$$

lead to

$$m_{n+1} = (1-p)(m_n + pm_{n-1}), \quad n = 2, 3, \ldots \tag{2.4}$$

By means of (2.4), (2.3) and

$$e_n + m_n + d_n = 1,$$

$p_{n,0}$ can be easily calculated. The problem considered here is equivalent to the problem of finding the distribution of the waiting time to the first success-run of length two. For its solution also generating functions are used.

3. The estimation procedures. Consider a series of n independent Bernoulli trials with success probability p. Let n_1 be the number of successes and n_2 the number of pairs of consecutive successes. Then the mean of n_1 is

$$En_1 = np. \tag{3.1}$$

The mean of n_2 is

$$En_2 = (n-1)p^2. \tag{3.2}$$

The proof of (3.2) is easy. Let $X_1, \ldots, X_n$ be i.i.d. random variables which take only the values 0 and 1, with $P(X_1 = 1) = p$. Furthermore, let Z be the number of pairs with $X_i = 1$ and $X_{i+1} = 1$. Then

$$Z = \sum_{i=1}^{n-1} X_i X_{i+1}.$$

Hence, the mean En_2 of Z is

$$EZ = \sum_{i=1}^{n-1} EX_i X_{i+1} = (n-1)p^2.$$

The formulas (3.1) and (3.2) suggest the n-estimator

$$\widehat{n} = \frac{n_1^2}{n_2}. \tag{3.3}$$

Since n is an integer, in practice instead of $\widehat{n}$ the nearest integer to $\widehat{n}$ is taken as the estimator.

By simulation and numerical experiments we learned that it was better than an estimator originally used by the first author. This estimator had used the number of success pairs without overlappings. That means, in an isolated sequence of three consecutive successes only one pair is counted, while in a series of four successes two pairs are counted. Clearly, the number of pairs in our counting procedure is greater than in the original counting. This may explain the better quality. "Better" means mainly "smaller variance of estimation"; the biases are similar for both methods.

Table 1 shows parameters which characterize the quality of $\widehat{n}$.

T A B L E 1. Means and *standard deviations* of $\widehat{n}$

p \ n	10	20	50	100	200
0.3					211.3
					39.9
0.5				103.3	203.0
				12.2	*14.9*
0.7		22.0	51.7	101.6	200.9
		3.6	*3.6*	*4.8*	*6.1*
0.9	11.3	21.2	51.2	101.2	200.6
	1.1	*0.9*	*1.0*	*1.2*	*1.6*

The values of n and p are such that $p_{n,0} = P(n_2 = 0)$ is very small. The values of mean and standard deviation for $n < 50$ result from an exact iteration procedure such as mentioned in Section 2. The other values were obtained by Monte Carlo simulation. (In the simulations the case $n_2 = 0$ did not appear; the calculated means and standard deviations are under the condition that $n_2 > 0$.)

It is not surprising that the biases (mean of $\widehat{n}$ − true value) and the standard deviation decrease with increasing p for fixed n. In our opinion, the estimator $\widehat{n}$ seems to be a good estimator for p not too small.

If we have to consider k success counts (for example, k silicon wafers or k forest areas), then the estimator (3.5) in [6] could be an alternative. This estimator is a stabilized method of moments estimator:

$$\widetilde{n} = \max\{s^2\phi^2/(\phi - 1), s_{\max}\} \tag{3.4}$$

where s is the sample variance to the success counts $s_1, \ldots, s_k$ and $s_{\max}$ is the maximum of the k counts. Furthermore,

$$\phi = \begin{cases} \overline{x}/s^2 & \text{if } \overline{x}/s^2 \geq 1 + 1/\sqrt{2}, \\ \max\{(s_{\max} - \overline{x})/s^2, 1 + \sqrt{2}\} & \text{if } \overline{x}/s^2 < 1 + 1/\sqrt{2}. \end{cases}$$

If information about success pairs is available, then (3.3) can be used for constructing two further estimators in the case $k > 1$:

$$\widehat{\overline{n}} = (\widehat{n}^{(1)} + \ldots + \widehat{n}^{(k)})/k \tag{3.5}$$

where $\widehat{n}_i$ is the result of (3.3) for the ith success count and

$$\widehat{n}^{(k)} = \frac{(\text{number of all successes in all } k \text{ counts})^2}{\text{number of all success pairs in all } k \text{ counts}}. \tag{3.6}$$

We have compared the estimators (3.4)–(3.6) by a Monte Carlo experiment.

It was carried out as in [6]. The whole procedure consisted in 1000 steps. At each step, values of n, p, and k were generated at random and then k sequences of n trials. All three estimators were used and the winner (smallest absolute difference between n and the n-estimator) was determined.

Clearly, such a comparison only makes sense for great values of p. We restricted the p-values to

$$p \geq 1.033 - 0.0133n \quad \text{for } n \leq 40$$

and

$$p \geq 0.633 - 0.0033n \quad \text{otherwise.}$$

For these values $p_{n,0}$ is smaller than 0.001. (If n_2 became zero, $\widehat{n}$ was of course the looser estimator.) The values for n and k were taken uniformly between 10 and 100 and 3 and 25, as in [6].

The result of this comparison was clear: In 652 cases $\widehat{\overline{n}}$ was the winner, in 348 the winner was $\widetilde{n}$; $\widehat{n}^{(k)}$ was never the winner.

References

[1] J. Besag and P. J. Green, *Spatial statistics and Bayesian computation*, J. Roy. Statist. Soc. B 55 (1993), 25–37.

[2] J. Besag, J. C. York and A. Mollié, *Bayesian image restauration, with two applications in spatial statistics* (*with discussion*), Ann. Inst. Statist. Math. 43 (1991), 1–59.

[3] R. J. Carroll and F. Lombard, *A note on N estimators for the binomial distribution*, J. Amer. Statist. Assoc. 80 (1985), 423–426.

[4] S. Janson, *Runs in m-dependent sequences*, Ann. Probab. 12 (1984), 805–818.

[5] W. Kühne, *Some results in subdividing the yield in microelectronic production by measurable parameters* (in preparation) (1994).

[6] I. Olkin, A. J. Petkau and J. V. Zidek, *A comparison of n estimators for the binomial distribution*, J. Amer. Statist. Assoc. 76 (1981), 637–642.

WOLFGANG KÜHNE
BEILSTR. 11
01277 DRESDEN, GERMANY

DIETRICH STOYAN
HELMUT STOYAN
TU BERGAKADEMIE FREIBERG
FACHBEREICH MATHEMATIK
09596 FREIBERG, GERMANY

PETER NEUMANN
TECHNISCHE UNIVERSITÄT DRESDEN
FACHBEREICH MATHEMATIK
01062 DRESDEN, GERMANY

Received on 8.6.1993