The Annals of Statistics
1973, Vol. 1, No. 2, 373-379

STATISTICAL INFERENCE IN BERNOULLI TRIALS WITH DEPENDENCE[1]

BY JEROME KLOTZ

University of Wisconsin at Madison

A model for Bernoulli trials with Markov dependence is developed which possesses the usual frequency parameter $p = P[X_i = 1]$ and an additional dependence parameter $\lambda = P[X_i = 1 \mid X_{i-1} = 1]$. Sufficient statistics for the model with p and λ unknown are found and an exact closed form expression for their small sample joint distribution is given. Large sample distribution theory is also given and small sample variances compared with large sample approximations. Easily computed estimators of p and λ are recommended and shown to be asymptotically efficient. With p unknown the u.m.p. unbiased test of independence is noted to be the run test. An application to a rainfall example is given.

1. Summary. A model for Bernoulli trials with Markov dependence is developed which possesses the usual frequency parameter p and an additional dependence parameter λ. Small and large sample distribution theory for the sufficient statistics of the model is presented. Easily computed estimators of p and λ are recommended and shown to be asymptotically efficient. Lastly, with p unknown, the u.m.p. unbiased test of independence is noted to be the run test.

2. Introduction and model. Consider a sequence of random variables $X_1, X_2, \cdots, X_n$ which each take on the values 1 and 0 as in the Bernoulli model. We consider the following generalization with Markov dependence between successive observations and

$$P[X_i = 1] = 1 - P[X_i = 0] = p\,, \qquad i = 1, 2, \cdots, n\,, \tag{2.1}$$

$$P[X_i = 1 \mid X_{i-1} = 1] = \lambda\,, \qquad i = 2, 3, \cdots, n\,. \tag{2.2}$$

From (2.1) and (2.2) it follows that

$$P[X_i = 0 \mid X_{i-1} = 1] = 1 - \lambda \tag{2.3}$$

$$P[X_i = 1 \mid X_{i-1} = 0] = (1 - \lambda)p/q \tag{2.4}$$

$$P[X_i = 0 \mid X_{i-1} = 0] = (1 - 2p + \lambda p)/q \tag{2.5}$$

using $p = P[X_i = 1, X_{i-1} = 1] + P[X_i = 1, X_{i-1} = 0]$. We have $0 \leqq p \leqq 1$ and $\max(0, (2p - 1)/p) \leqq \lambda \leqq 1$ so that the transition probabilities are between zero and one. When $\lambda = p$ the model reduces to independent Bernoulli trials, when $\lambda > p$ clustering will occur among the ones and among the zeros, and when $\lambda < p$ a lack of clustering is present.

Received November 17, 1971; revised June 22, 1972.

[1] This investigation was supported at the University of California Berkeley by PHS Research Grant No. GM-10525-08, National Institutes of Health, Public Health Service.

Using (2.1) thru (2.5) and the Markov assumption, the joint distribution can be written

$$
\begin{aligned}
&P[X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n] \\
&\quad = P[X_1 = x_1]P[X_2 = x_2 \mid X_1 = x_1] \cdots P[X_n = x_n \mid X_{n-1} = x_{n-1}] \\
&\quad = p^{x_1}q^{1-x_1} \prod_{i=2}^{n} \lambda^{x_i - 1^{x_i}}(1-\lambda)^{x_{i-1}(1-x_i)}[(1-\lambda)p/q]^{(1-x_{i-1})x_i} \\
&\qquad \times [(1-2p+\lambda p)/q]^{(1-x_{i-1})(1-x_i)} \\
&\quad = \lambda^r(1-\lambda)^{2(s-r)-t}(1-2p+\lambda p)^{n-1-2s+t+r}p^{s-r}q^{-(n-2-s+t)} \\
&\quad = C_{1n}\eta_1^{\,r}\eta_2^{\,s}\eta_3^{\,t}
\end{aligned} \tag{2.6}
$$

where

$$
\begin{aligned}
C_{1n} &= (1-2p+\lambda p)^{n-1}/q^{n-2} \\
\eta_1 &= \lambda(1-2p+\lambda p)/(p(1-\lambda)^2) \\
\eta_2 &= (1-\lambda)^2pq/(1-2p+\lambda p)^2 \\
\eta_3 &= (1-2p+\lambda p)/(q(1-\lambda))
\end{aligned} \tag{2.7}
$$

and $r = \sum_{i=2}^{n} x_{i-1}x_i$, $s = \sum_{i=1}^{n} x_i$, $t = x_1 + x_n$. As a check when $\lambda = p$, (2.6) reduces to $p^s q^{n-s}$ for the Bernoulli model.

3. Distribution of the sufficient statistics. From (2.6) using the factorization theorem, $R = \sum_{i=2}^{n} X_{i-1}X_i$, $S = \sum_{i=1}^{n} X_i$, and $T = X_1 + X_n$ are sufficient for $\theta = (\lambda, p)$. In (2.6) we see that the joint probability is constant for fixed values (r, s, t) of (R, S, T) so the joint distribution can be obtained by counting the number of sequences $(x_1, x_2, \cdots, x_n)$ of zeros and ones that give (r, s, t). In earlier applications of the model in which p was assumed known [6], [8] this number was found to be $\binom{2}{t}\binom{s-1}{r}\binom{n-s-1}{s-r-t}$ so that

$$
P[R = r, S = s, T = t] = \binom{2}{t}\binom{s-1}{r}\binom{n-s-1}{s-r-t}C_{1n}\eta_1^{\,r}\eta_2^{\,s}\eta_3^{\,t} \tag{3.1}
$$

with the convention $\binom{-1}{-1} = 1$ adopted for the case $s = n$ and C_{1n} and η_i given by (2.7). A combinatorial argument relating the number of zero runs (W) between the first and last ones in the sequence $(x, x_2, \cdots, x_n)$ was used to prove (3.1) in [8] ($R = S - 1 - W$).

If we write $X = (R - (n-1)\lambda p)/n^{\frac{1}{2}}$, $\gamma = (s - np)/n^{\frac{1}{2}}$ then the limiting joint distribution of (X, Y) and T can be shown to be a bivariate normal $N((0, 0), \Sigma)$ and an independent binomial $B(2, p)$ with the asymptotic variance covariance matrix

$$
\Sigma = \begin{pmatrix} \lambda p(1-\lambda p) + (2\lambda^2pq^2/(1-\lambda)) & 2\lambda^2pq^2/(1-\lambda) \\ 2\lambda^2pq^2/(1-\lambda) & pq(1-2p+\lambda)/(1-\lambda) \end{pmatrix}. \tag{3.2}
$$

This result follows from (3.1) using Stirling's approximation and a log expansion when taking the limit as n becomes large.

4. Asymptotic estimation theory. The model (2.6) is a particular case of the general Markov models discussed by Billingsley [2] and the general theory de-

veloped there can be applied. In the notation of [2], we have

$$f(x_{i-1}, x_i; \boldsymbol{\theta}) = \lambda^{x_{i-1}x_i}(1-\lambda)^{x_{i-1}(1-x_i)}[(1-\lambda)p/q]^{(1-x_{i-1})x_i} \times [(1-2p+\lambda p)/q]^{(1-x_{i-1})(1-x_i)}.$$

If we call $\hat{\boldsymbol{\theta}} = (\hat{\lambda}, \hat{p})$ a root in the region $\Theta = \{(\lambda, p): 0 \leqq p \leqq 1, \max(0, (2p-1)/p) \leqq \lambda \leqq 1\}$ that simultaneously satisfies the two partial derivative likelihood equations, it follows from Theorem (2.2) of [2] when (λ, p) is in the interior of Θ that $n^{\frac{1}{2}}(\hat{\lambda}-\lambda, \hat{p}-p)$ has a limiting bivariate normal distribution $N((0,0), \Lambda)$ with variance covariance matrix

$$\Lambda = \begin{pmatrix} \lambda(1-\lambda)/p & q\lambda \\ q\lambda & pq(1-2p+\lambda)/(1-\lambda) \end{pmatrix} \tag{4.1}$$

obtained by inverting the information matrix. The derivative likelihood equation derived from the partial derivatives of (2.6) are

$$\frac{r}{\lambda} - \frac{(2(s-r)-t)}{1-\lambda} + \frac{(n-1-2s+t+r)p}{(1-2p+\lambda p)} = 0 \tag{4.2a}$$

$$\frac{s-r}{p} + \frac{(n-2-s+2)}{q} + \frac{(n-1-2s+t+r)(\lambda-2)}{(1-2p+\lambda p)} = 0. \tag{4.2b}$$

It is apparent from (4.2) that closed form expressions are difficult to obtain and unrewarding although numerical solutions can be computed. However, from (3.2) we note that the asymptotic variance of $\bar{X} = S/n$ is the same as that of $\hat{p}$ given in (4.1). When p is known, the maximum likelihood estimator $\hat{\lambda}(p)$ can be shown [8] to be the positive radical root of the quadratic equation derived from (4.2a),

$$\hat{\lambda}(p) = \frac{[r-q(2s-t)+(m)p]+([r-q(2s-t)+(m)p]^2+4r(1-2p)(m)p)^{\frac{1}{2}}}{2(m)p} \tag{4.3}$$

where $m = n-1$. Because of the simplicity of computation, the estimators $\tilde{\boldsymbol{\theta}} = (\hat{\lambda}(\bar{X}), \bar{X})$ are recommended in view of the following.

THEOREM. *The estimators $\hat{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}$ are asymptotically equivalent and asymptotically efficient for $\boldsymbol{\theta}$ in an open subset of Θ.*

PROOF. For our simple Markov model, regularity conditions can be verified and, in particular, Theorem 1 of Roussas [11], page 254 gives local asymptotic normality as required in Assumption (3.1) of Hájek [5]. As a consequence, Theorem 4.2 of [5] is valid and Conditions (4.15) and (4.16) of this theorem hold for both estimates $\hat{p}$ and $\bar{X}$. Thus $p - \lim n^{\frac{1}{2}}(\hat{p}-\bar{X}) = 0$ and we have asymptotic equivalence. To show large sample equivalence of $\hat{\lambda}$ and $\tilde{\lambda} = \hat{\lambda}(\bar{X})$ where $\hat{\lambda}(p)$ is given by (4.3) we note from (4.1) that $R/(n-1) \to_p \lambda p$, $\bar{X} \to_p p$ and $T/n \to_p 0$. Substituting $\bar{X}$ for p in (4.3) and using Slutsky's theorem we have $\tilde{\lambda} = \hat{\lambda}(\bar{X}) \to_p \lambda$ as n becomes large. Next, denote the log of (2.6) by $L(\boldsymbol{\theta}) = L(\lambda, p)$ and the partial derivatives by $g_\lambda(\lambda, p) = \partial L/\partial\lambda$, $g_{\lambda\lambda} = \partial^2 L/(\partial\lambda)^2$, $g_{\lambda p} = \partial^2 L/\partial\lambda\,\partial p$. Expand-

ing $g_\lambda(\tilde{\lambda}, \bar{X})/n^{\frac{1}{2}}$ about $(\hat{\lambda}, \hat{p})$, we have

$$g_\lambda(\tilde{\lambda}, \bar{X})/n^{\frac{1}{2}} = (g_\lambda(\hat{\lambda}, \hat{p})/n^{\frac{1}{2}}) + n^{\frac{1}{2}}(\bar{X} - \hat{p})g_{\lambda p}(\lambda^*, p^*)/n + n^{\frac{1}{2}}(\tilde{\lambda} - \hat{\lambda})g_{\lambda\lambda}(\lambda^*, p^*)/n \tag{4.4}$$

where (λ^*, p^*) is between $(\tilde{\lambda}, \bar{X})$ and $(\hat{\lambda}, \hat{p})$. The left side of (4.4) and the first term on the right are zero since $\tilde{\lambda} = \hat{\lambda}(\bar{X})$ and $\hat{\lambda}$ are both solutions to the derivative equation (4.2a). Since $(\hat{\lambda}, \hat{p})$, $(\tilde{\lambda}, \bar{X})$ are consistent so is (λ^*, p^*). It follows that $-g_{\lambda p}(\lambda^*, p^*)/n \to_p \gamma_{12}$, $-g_{\lambda\lambda}(\lambda^*, p^*)/n \to_p \gamma_{11}$ the corresponding terms of the information matrix

$$\Gamma = \begin{pmatrix} \dfrac{p(1-2p+\lambda)}{\lambda(1-\lambda)(1-2p+\lambda p)} & \dfrac{-1}{1-2p+\lambda p} \\ \dfrac{-1}{1-2p+\lambda p} & \dfrac{1-\lambda}{pq(1-2p+\lambda p)} \end{pmatrix} = \Lambda^{-1}. \tag{4.5}$$

Using $n^{\frac{1}{2}}(\hat{p} - \bar{X}) \to_p 0$ we conclude from (4.4) that $n^{\frac{1}{2}}(\tilde{\lambda} - \hat{\lambda}) \to_p 0$. Asymptotic efficiency of $(\tilde{\lambda}, \bar{X})$ follows from that of $(\hat{\lambda}, \hat{p})$ and the equivalence.

5. Small sample variance comparisons. If we write out the transition probability matrix using (2.2) thru (2.5) we have

$$\mathbf{P} = \begin{pmatrix} (1-2p+\lambda p)/q & (1-\lambda)p/q \\ 1-\lambda & \lambda \end{pmatrix}. \tag{5.1}$$

We can diagonalize P using its right and left eigen vectors and its eigen values

$$\mathbf{P} = \begin{pmatrix} 1 & p \\ 1 & -q \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & (\lambda - p)/q \end{pmatrix}\begin{pmatrix} q & p \\ 1 & -1 \end{pmatrix} \tag{5.2}$$

and so

$$\mathbf{P}^k = \begin{pmatrix} 1 & p \\ 1 & -q \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & [(\lambda - p)/q]^k \end{pmatrix}\begin{pmatrix} q & p \\ 1 & -1 \end{pmatrix}. \tag{5.3}$$

Using this kth step transition probability matrix we compute

$$\begin{aligned} \operatorname{Cov}(X_i, X_{i+k}) &= EX_iX_{i+k} - p^2 = P[X_i = 1, X_{i+k} = 1] - p^2 \\ &= p(p + q[(\lambda - p)/q]^k) - p^2 = pq[(\lambda - p)q]^k \end{aligned} \tag{5.4}$$

and, after some algebra,

$$\operatorname{Var} S = npq + \frac{2pq(\lambda - p)}{1-\lambda}\left[(n-1) - \frac{(\lambda - p)}{(1-\lambda)}(1 - [(\lambda - p)/q]^n)\right]. \tag{5.5}$$

Similar computations give $\operatorname{Cov}(X_iX_{i+1}, X_jX_{j+1}) = \lambda^2 pq[(\lambda - p)/q]^{j-i-1}$ and

$$\begin{aligned} \operatorname{Var} R = {} & (n-1)\lambda p(1-\lambda p) \\ & + \frac{2\lambda^2 pq^2}{(1-\lambda)}\left[(n-2) - \frac{9}{(1-\lambda)}(1 - [(\lambda - p)/q]^{n-1})\right]. \end{aligned} \tag{5.6}$$

Both (5.5) and (5.6) give the corresponding entries in (3.2) when n becomes large.

The distribution of T is given by

$$
\begin{aligned}
P[T = 0] &= q^2 + pq[(\lambda - p)/q]^n \\
P[T = 1] &= 2pq - 2pq[(\lambda - p)/q]^n \\
P[T = 2] &= p^2 + pq[(\lambda - p)/q]^n
\end{aligned} \tag{5.7}
$$

which approaches the binomial $B(2, p)$ when n is large.

6. Testing the hypothesis of independence. For the hypothesis $H: \lambda = p$, corresponding to independent trials, we have a complete sufficient statistic for the unknown parameter p given by $S = X_1 + X_2 + \cdots + X_n$. Thus unbiased tests are similar with Neyman structure and are conditional tests with conditional level α given S ([9] page 130). For a one-sided test of $H: \lambda \leqq p$ against $A: \lambda > p$ the conditional likelihood ratio given S is equivalent to $\eta_1{}^R\eta_3{}^T = \eta_3\eta_1{}^S\Delta(u, v)$ where η_1, η_3 are given by (2.7), $U = S - R - T + 1$, $V = S - R$, and Δ is defined by Lehmann ([9] page 156). By showing that Δ is essentially a decreasing function of $Z = U + V = 2(S - R) - T + 1$, the total number of runs, Lehmann shows that the most powerful similar test coincides with the run test at its natural significance levels. This test rejects for small values of Z where

$$
\begin{aligned}
P_H[Z = 2k \mid S = s] &= 2\binom{s}{k-1}\binom{n-s-1}{k-1}/\binom{n}{s} \\
P_H[Z = 2k + 1 \mid S = s] &= \left[\binom{s-1}{k}\binom{n-s-1}{k-1} + \binom{s-1}{k-1}\binom{n-s-1}{k}\right]/\binom{n}{s} .
\end{aligned} \tag{6.1}
$$

This distribution is well known [9], [14] and tabled in [12]. A minor modification obtained by neglecting T, rejects for large values of R given S. This modification is asymptotically equivalent and has the hypergeometric distribution

$$
P[R = r \mid S = s] = \binom{s-1}{r}\binom{n-s+1}{s-r}/\binom{n}{s} \tag{6.2}
$$

[7] with more comprehensive tables [10].

7. An example with rainfall data. We consider an application of the model to rainfall data as do Gabriel and Neuman [3], [4] who use different methods of Markov chain inference. June days with measurable amounts of precipitation (at least .01 inches) at Madison, Wisconsin are given in the table for the years 1961 thru 1971 [13]. Assuming that the model holds between consecutive days in June and that independence holds between different years, model (3.1) is modified [8] for K independent samples, each of length n_k, to

$$
P[R = r, S = s, T = t] = \binom{2K}{t}\binom{n-s-K}{s-r-t}\binom{s-K}{r}C_{Kn}\eta_1{}^r\eta_2{}^s\eta_3{}^t , \tag{7.1}
$$

where $R = \sum_{k=1}^K R_k$, $S = \sum_{k=1}^K S_k$, $T = \sum_{k=1}^K T_k$, $n = \sum n_k$, $C_{Kn} = \prod_{k=1}^K C_{1n_k}$, and C_{1n} and η_i are given by (2.7). Using the same motivation as before, the recommended estimator $\hat{\lambda}$ is modified for K samples by taking $m = \sum_{k=1}^K (n_k - 1) = n - K$ and $\hat{p} = S/n$ in(4.3). Thus for the rainy month of June, the estimated conditional probability of rain given rain the previous day is increased to nearly one-half from the unconditional probability estimate of approximately one-third. This estimate is in agreement with others [1] for nearby locations.

TABLE 1

June days with measurable precipitation at Madison, Wisconsin

Year	June Day Number	R_k,	S_k,	T_k
1961	1, 7, 8, 10, 12, 13, 19, 22, 24	2,	9,	1
1962	3, 4, 8, 10, 11, 17, 18, 22	3,	8,	0
1963	5, 7, 8, 9, 13, 19, 26, 27	3,	8,	0
1964	2, 11, 12, 14, 15, 17, 21, 22	3,	8,	0
1965	1, 5, 6, 20, 22, 23, 27	2,	7,	1
1966	2, 3, 6, 7, 9, 11, 12, 15, 20, 26, 27, 28	5,	12,	0
1967	6, 7, 9, 10, 11, 12, 13, 15, 16, 17, 19, 24, 28, 29	8,	14,	0
1968	1, 9, 10, 11, 14, 15, 18, 21, 23, 24, 25, 26, 27, 29, 30	8,	14,	2
1969	1, 2, 4, 6, 7, 11, 12, 17, 18, 25, 26, 27, 29	6,	13,	1
1970	1, 2, 12, 16, 17, 20, 26	2,	7,	1
1971	1, 7, 11, 18, 19, 20, 22, 24	2,	8,	1
	$\hat{\lambda} \doteq .462$, $\hat{p} \doteq .327$, $n = 330$, $m = 319$. Totals	44,	108,	7
		R,	S,	T

Acknowledgment. The model arose out of a consulting problem with the late Charles Johnson of Argonne Labs. Deep appreciation goes to Lucien LeCam for extremely helpful suggestions. Conversations with Bernie Harris, C. J. Park, E. Wahl, and George Roussas were also helpful as were the referee's suggestions.

REFERENCES

[1] BARK, L. D., BURROWS, W. C. and FEYERHERM, A. M. (1965). Probabilities of sequences of wet and dry days in Wisconsin. Kansas Agricultural Experimental Station Technical Bulletin 139 g.

[2] BILLINGSLEY, P. (1961). *Statistical Inference for Markov Processes.* Univ. of Chicago Press.

[3] GABRIEL, K. R. (1959). The distribution of the number of successes in a sequence of dependent trials. *Biometrika* **46** 454–460.

[4] GABRIEL, K. R. and NEUMAN, J. (1962). A Markov chain model for daily rain-fall occurrence at Tel-Aviv. *Quart. J. Roy. Meteorological Soc.* **88** 90–95.

[5] HÁJEK, J. (1972). Local asymptotic minimax and admissibility in estimation. *Proc. Sixth Berkeley Symp. Math. Statist. Prob.* To appear.

[6] JOHNSON, C. and KLOTZ, J. (1971). The atom probe and Markov chain statistics of clustering. Univ. of Wisconsin Technical Report No. 267, Department of Statistics.

[7] JOHNSON, N. L. and KOTZ, S. (1969). *Discrete Distributions.* Houghton Mifflin, Boston.

[8] KLOTZ, J. (1972). Markov chain clustering of births by sex. *Proc. Sixth Berkeley Symp. Math. Statist. Prob.* To appear.

[9] LEHMANN, E. (1959). *Testing Statistical Hypotheses.* Wiley, New York.

[10] LIEBERMAN, G. and OWEN, D. (1961). *Tables of the Hypergeometric Distribution.* Stanford Univ. Press.

[11] ROUSSAS, G. (1968). Some applications of the asymptotic distribution of likelihood functions to the asymptotic efficiency of estimates. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **10** 252–260.

[12] SWED, F. and EISENHART, C. (1943). Tables for testing randomness of grouping in a sequence of alternatives. *Ann. Math. Statist.* **14** 66–87.

[13] U. S. DEPARTMENT OF COMMERCE (1961–1971). Local Climatological Data, Madison, Wisconsin, Truax Field, June 1961-June 1971. Superintendent of Documents, U. S. Govt. Printing Office, Washington, D.C. 20402.

[14] WOLFOWITZ, J. (1943). On the theory of runs with some applications to quality control. *Ann. Math. Statist.* **14** 280–288.

DEPARTMENT OF STATISTICS
UNIVERSITY OF WISCONSIN
1210 WEST DAYTON STREET
MADISON, WISCONSIN 53706