THE GAP TEST FOR RANDOM SEQUENCES

By Eve Bofinger[1] and V. J. Bofinger[2]

North Carolina State College

Summary. This paper is concerned with the gap test for random sequences, first proposed by Kendall and Babington-Smith [7], and with various extensions to this test. One of these extensions is the test proposed by Meyer, Gephart and Rasmussen [8], another is, asymptotically, a partitioning of the χ^2 statistic of Kendall and Babington-Smith [7], and others are likelihood ratio tests based on Markov chain models.

Notation. Consider a long chain of observations, $a_1, a_2, \cdots, a_N$, arising from a Markov process of order $\nu - 1$, with two states denoted by 0 and 1, and with positive transition probabilities $p_{r_1 \cdots r_\nu}$. That is, $p_{r_1 \cdots r_\nu}$ is the conditional probability of the state r_ν given that the preceding $\nu - 1$ states are $r_1, r_2, \cdots, r_{\nu-1}$. Assume the process starts in a stationary state (although this can easily be seen not to affect the asymptotic results given), so that the occupation probabilities $P(r_1 \cdots r_{\nu-1})$ may be derived from the transition probabilities by the relation $\sum_{r_1} P(r_1 \cdots r_{\nu-1}) p_{r_1 \cdots r_\nu} = P(r_2 \cdots r_\nu)$.

Let $n_{r_1 \cdots r_t}$ be the number of times that $r_1, \cdots, r_t$ appears as a connected sequence within the observation chain $a_1, \cdots, a_N$. In the case where $\nu = 1$ (the case of independent observations or random binary numbers) it seems reasonable (so that the algebra will be tidier) to follow Kendall and Babington-Smith [7] and consider a "cyclic" sequence in which $a_{i+N} \equiv a_i$. In this case we define $n_{r_1 \cdots r_t}$ as the number of times $r_1, \cdots, r_t$ appears as a connected sequence in one cycle of the observation chain. The difference in $n_{r_1 \cdots r_t}$ under the cyclic or non-cyclic definitions is at most $t - 1$.

The gap test is concerned with the number of non-zero digits (all denoted here by 1) between zero digits, but we could easily apply the results to gaps between any particular class of digits, for example, between even digits. For random decimal digits we have $\nu = 1$ and $p_0 = 0.1$ while $p_1 = 0.9$.

Let $N_x = n_{r_1 r_2 \cdots r_{x+1} r_{x+2}}$ and $M_x = n_{r_1 r_2 \cdots r_{x+1}}$, where $r_1 = r_{x+2} = 0$ and $r_2 = \cdots = r_{x+1} = 1$. That is, if there are at least two zeros in the sequence $a_1, \cdots, a_N$, then N_x is the number of gaps of length x and, in the cyclic case, M_x is the number of gaps of length x or greater.

Since we may wish to pool some classes to give "reasonably large expectations", let

$$N_{x \to x+s} = \sum_{i=0}^{s} N_{x+i}$$

Received January 19, 1959; revised May 26, 1960.

[1] At present at the University of Sydney, Sydney, N.S.W., Australia.

[2] At present at the University of New South Wales, Kensington, N.S.W., Australia.

and

$$M_{x\to x+s} = \sum_{i=0}^{s} M_{x+i}\,.$$

For small values of x, s will probably be taken as zero and may always be taken as zero if we are concerned only with gaps of small size rather than all possible gap sizes. We shall, of course, choose values of x and s so that we have non-overlapping classes.

Since we shall mainly be concerned with an independent sequence, we take $p_0 = p$ and $p_1 = q$. Let $g_x = pq^x$ and $g_{x\to x+s} = \sum_{i=0}^{s} g_{x+i} = q^x(1 - q^{s+1})$.

In what follows, $\sum$, with no superscripts or subscripts, is taken to mean summation over all appropriate pairs of values of x and s. That is, the summation is over all classes considered, which may be numbered $0, 1, \cdots, L$.

Formulation of the Problem. We shall consider the following possible tests of the null hypothesis $p_{r_1\cdots r_\nu} = p_{r_\nu} = p$ when $r_\nu = 0$, where p has a specified value:

Test A: Kendall and Babington-Smith [7] suggested using

$$X_A^2 = \sum \frac{(N_{x\to x+s} - n_0\, g_{x\to x+s})^2}{n_0\, g_{x\to x+s}}$$

which they considered to be asymptotically (with respect to N) distributed as χ^2 with L degrees of freedom, where the $L + 1$ classes include all possible gap sizes, the last class including all those from a certain convenient finite size up to size $N - 2$.

Test B: Meyer, Gephart and Rasmussen [8] have suggested their "strong" gap test using

$$X_B^2 = \sum \frac{(N_{x\to x+s} - Npg_{x\to x+s})^2}{Npg_{x\to x+s}}$$

which they have taken to be asymptotically distributed as χ^2 with $L + 1$ degrees of freedom. We shall show that this is not quite the case.

Test C: We suggest that it may be more useful to base a test on

$$X_C^2 = \sum \frac{(N_{x\to x+s} - pM_{x\to x+s})^2}{Npqg_{x\to x+s}}\,.$$

We shall show that this is asymptotically distributed as χ^2 with $L + 1$ degrees of freedom, and, in fact, that for each of the $L + 1$ classes

$$\frac{(N_{x\to x+s} - pM_{x\to x+s})^2}{Npqg_{x\to x+s}}$$

is asymptotically distributed as χ^2 with 1 degree of freedom.

An alternative procedure is to use

$$\frac{(N_{x\to x+s} - pM_{x\to x+s})^2}{pqM_{x\to x+s}}$$

which, by Cramér [5], 20.6, is asymptotically equivalent (that is, has the same asymptotic distribution) under the null hypothesis.

The advantage of test C is that we may examine each of these separate contributions, since these are asymptotically independent. It is difficult, however, to relate these separate contributions to likelihood ratio tests in the way that (as will be seen later) we can relate the total X_C^2.

Asymptotic Distributions of X_A^2, X_B^2 and X_C^2. Now Bartlett [2] has shown that the $N^{-\frac{1}{2}}[n_{r_1\cdots r_t} - E(n_{r_1\cdots r_t})]$ for various $r_1, \cdots, r_t$ have asymptotically a joint normal distribution. Hence each of the sets of variables

a. $N^{-\frac{1}{2}}(N_{x\to x+s} - n_0 g_{x\to x+s})$

b. $N^{-\frac{1}{2}}(N_{x\to x+s} - Np\, g_{x\to x+s})$

c. $N^{-\frac{1}{2}}(N_{x\to x+s} - p\, M_{x\to x+s})$

for various finite values of x and s, being linear combinations (or very nearly so in the non-cyclic case) of the $N^{-\frac{1}{2}}[n_{r_1\cdots r_t} - E(n_{r_1\cdots r_t})]$, for t at least as large as the largest value of $x + s + 2$ considered, have asymptotically a joint normal distribution. We may easily see that the expected value of any one of the three variables (labelled a, b, or c) is zero and so we only need to find the variance-covariance matrix for the above sets of variables to find the asymptotic distributions of X_A^2, X_B^2 and X_C^2.

Following Billingsley [3] we let

$$\alpha_i = \begin{cases} 1 & \text{if the sequence } a_i, a_{i+1}, \cdots, a_{i+t-1} \\ & \text{is the sequence } r_1, r_2, \cdots, r_t \\ 0 & \text{otherwise} \end{cases}$$

and let

$$\beta_j = \begin{cases} 1 & \text{if the sequence } a_j, a_{j+1}, \cdots, a_{j+t+K-1} \\ & \text{is the sequence } s_1, s_2, \cdots, s_{t+K} \text{ where } K \text{ is a finite non-} \\ & \quad \text{negative integer} \\ 0 & \text{otherwise} \end{cases}$$

Now $n_{r_1\cdots r_t} = \sum_{i=1}^N \alpha_i$ (or $\sum_{i=1}^{N-t+1} \alpha_i$ for the non-cyclic case) and $n_{s_1\cdots s_{t+K}} = \sum_{j=1}^N \beta_j$. Hence

$$\operatorname{Cov}(n_{r_1\cdots r_t}, n_{s_1\cdots s_{t+k}}) = \sum_{i,j=1}^{N} \operatorname{Cov}(\alpha_i, \beta_j).$$

The evaluation of this expression is greatly simplified in the case we are considering of an independent sequence and is further simplified here, since we consider such sequences as $r_1, \cdots, r_t$ with $r_1 = r_t = 0$ and $r_2 = \cdots = r_{t-1} = 1$.

We find that

$$\operatorname{Var}(N_x) = Np^2q^x(1 + 2pq^x - (2x + 3)p^2q^x)$$

$$\operatorname{Cov}(N_x, N_y) = Np^3q^{x+y}(2 - (x + y + 3)p)$$

and

$$\mathrm{Cov}(N_x, n_0) = Np^2q^x(2 - (x + 2)p)$$

Hence

$$\mathrm{Var}(N_x - n_0g_x) = Npg_x(1 - g_x)$$

and

$$\mathrm{Cov}(N_x - n_0g_x, N_y - n_0g_y) = -Npg_xg_y .$$

Hence

$$\mathrm{Var}(N_{x\to x+s} - n_0g_{x\to x+s}) = Npg_{x\to x+s}(1 - g_{x\to x+s})$$

and

$$\mathrm{Cov}(N_{x\to x+s} - n_0g_{x\to x+s}, N_{y\to y+t} - n_0g_{y\to y+t}) = -Npg_{x\to x+s}g_{y\to y+t} .$$

The variance covariance matrix of the $N_x - pM_x$ may be obtained by noticing that

$$\begin{aligned}\mathrm{Cov}(n_{r_1\cdots r_t} - p_{r_t}n_{r_1\cdots r_{t-1}}, n_{s_1\cdots s_t} - p_{s_t}n_{s_1\cdots s_{t-1}})\\ = N\delta_{s_1\cdots s_{t-1}}^{r_1\cdots r_{t-1}}\, p_{r_1} \cdots p_{r_t}(\delta_{s_t}^{r_t} - p_{s_t})\end{aligned}$$

where

$$\delta_{s_1\cdots s_{t-1}}^{r_1\cdots r_{t-1}} = \begin{cases}1 & \text{if } r_1 = s_1, \cdots, r_{t-1} = s_{t-1}\\ 0 & \text{otherwise.}\end{cases}$$

This may be seen by a slight modification of the work of Anderson and Good man [1] or by expressing the above covariance as $\sum_{i,j=1}^{N} f(i, j)$, where $f(i, j)$ contains four terms of the type $\mathrm{Cov}(\alpha_i, \beta_j)$ corresponding to the four terms of the product

$$(n_{r_1\cdots r_t} - p_{r_t}n_{r_1\cdots r_{t-1}})(n_{s_1\cdots s_t} - p_{s_t}n_{s_1\cdots s_{t-1}}).$$

We find that $f(i, j)$ is zero unless $i = j$ and that

$$f(i, i) = \delta_{s_1\cdots s_{t-1}}^{r_1\cdots r_{t-1}}\, p_{r_1} \cdots p_{r_t}(\delta_{s_t}^{r_t} - p_{s_t}).$$

Hence

$$\begin{aligned}\mathrm{Cov}(n_{r_1\cdots r_t} - p_{r_t}n_{r_1\cdots r_{t-1}}, n_{s_1\cdots s_{t+K}} - p_{s_{t+K}}n_{s_1\cdots s_{t+K-1}})\\ = N\delta_{s_{K+1}\cdots s_{t+K-1}}^{r_1\cdots r_{t-1}}\, p_{s_1} \cdots p_{s_{t+K}}(\delta_{s_{t+K}}^{r_t} - p_{r_t})\end{aligned}$$

and so

$$\mathrm{Cov}(N_x - pM_x, N_y - pM_y) = 0 \quad \text{for} \quad x \neq y.$$

Hence

$$\mathrm{Cov}(N_{x\to x+s} - pM_{x\to x+s}, N_{y\to y+t} - pM_{y\to y+t}) = 0 \quad \text{for} \quad x + s < y.$$

Also

$$\mathrm{Var}(N_x - pM_x) = Np^2q^{x+1}$$

and hence

$$\mathrm{Var}(N_{x\to x+s} - pM_{x\to x+s}) = Npqg_{x\to x+s}\,.$$

Now we may find the asymptotic distributions of X_A^2, X_B^2 and X_C^2.

Test A: Consider classes numbered 0, 1, 2, $\cdots$, L with associated variates $f_0, f_1, f_2, \cdots, f_L$ such that for $i, j = 0, 1, 2, \cdots L$ we have $E(f_i) = 0$, $\mathrm{Var}(f_i) = Kp_i(1 - p_i)$ and $\mathrm{Cov}(f_i, f_j) = -Kp_ip_j$ for $i \neq j$ and p_i is a positive number with $\sum_{i=0}^{L} p_i = P \leqq 1$.

We may easily show that the variance covariance matrix of $z_i = f_i(Kp_i)^{-\frac{1}{2}}$ has L latent roots equal to 1 and one equal to $1 - P$.

To apply this to the variates $N^{-\frac{1}{2}}(N_{x\to x+s} - n_0g_{x\to x+s})$ we notice that $P = 1 - q^{N-1}$ if all gap sizes are included and hence, by Cochran [4],

$$\sum \frac{(N_{x\to x+s} - n_0\, g_{x\to x+s})^2}{Npg_{x\to x+s}},$$

(and also, by Cramér [5], 20.6,

$$\sum \frac{(N_{x\to x+s} - n_0\, g_{x\to x+s})^2}{n_0\, g_{x\to x+s}})$$

is asymptotically distributed as χ^2 with L degrees of freedom.

This may also be seen by usual multinomial theory.

If, however, not all gap sizes are considered, but only, say, those of sizes 0 to k in the $L + 1$ classes ($L \leqq k$), then L latent roots still equal 1 but the $(L + 1)$th equals q^{k+1}. Provided k is large enough this would have a small effect, but in dealing with decimal numbers $q = 0.9$ and with $L = k = 4$, say, the fifth latent root is 0.59, which constitutes an appreciable effect.

Test B: From the variances and covariances determined above we find that

$$\mathrm{Var}\{N_x/(Npg_x)^{\frac{1}{2}}\} = 1 + 2pq^x - (2x + 3)p^2q^x$$

and

$$\mathrm{Cov}\left\{\frac{N_x}{(Npg_x)^{\frac{1}{2}}},\quad \frac{N_y}{(Npg_y)^{\frac{1}{2}}}\right\} = pq^{\frac{1}{2}(x+y)}\,[2 - (x + y + 3)p].$$

First let us consider the case where no pooling takes place. That is, we consider the statistic

$$\sum_0^L \frac{(N_x - Np^2\, q^x)^2}{Np^2\, q^x}\,.$$

We can easily find that the corresponding variance-covariance matrix has $L - 1$ of its latent roots equal to 1 and, after some algebra, we find that the remaining 2 latent roots are given by the expressions $\frac{1}{2}(\alpha \pm \beta)$, where

$$\alpha = 1 + q + (2L + 3)pq^{L+1}$$

and

$$\beta = \{(1 - q^{L+1})[(1 + q)^2(1 - q^{L+1}) - 4(L + 1)(L + 2)p^2q^{L+1}]\}^{\frac{1}{2}}.$$

If we pool so that the Lth class consists of all those gaps of length L or greater the variance-covariance matrix is modified in the following way. Let

$$N_* = N_L + N_{L+1} + \cdots + N_{N-2}, \; M_* = M_L + M_{L+1} + \cdots + M_{N-2}$$

and $g_* = g_L + g_{L+1} + \cdots + g_{N-2}$.

In practice, many of the components of N_* will be small or zero and N_* may be calculated as $n_0 - \sum_{x=0}^{L-1} N_x$.

We have considered the above method for pooling classes since it seems to be a reasonable one and any general pooling scheme as in Test A is too awkward algebraically.

The asymptotic expected value of N_* is Npq^L and so the test statistic is now

$$\sum_0^{L-1} \frac{(N_x - Np^2q^x)^2}{Np^2q^x} + \frac{(N_* - Npq^L)^2}{Npq^L}.$$

We can easily show that $\text{Var}\{N_*(Npq^L)^{-\frac{1}{2}}\}$ is asymptotically equal to $1 - (2L + 1)pq^L$ and that

$$\text{Cov}\{N_x(Np^2q^x)^{-\frac{1}{2}}, N_*(Npq^L)^{-\frac{1}{2}}\}$$

is asymptotically equal to

$$p^{\frac{1}{2}}q^{\frac{1}{2}(L+x)}[2 - (L + x + 3)p - q].$$

It follows that the corresponding variance-covariance matrix has $L - 1$ of its latent roots equal to 1 as before, and, after some algebra, has the remaining 2 latent roots equal to

$$\tfrac{1}{2}(1 + q)\{1 \pm (1 - 4q^{L+1}(1 + q)^{-2})^{\frac{1}{2}}\}.$$

This means that, for large L, one of the latent roots is approximately $1 + q$ and the other is approximately zero.

TEST C: For each class denoted by $x \to x + s$,

$$\frac{(N_{x\to x+s} - pM_{x\to x+s})^2}{Npqg_{x\to x+s}}$$

is asymptotically independent of similar contributions from the other L classes and is asymptotically distributed as χ^2 with 1 degree of freedom. Hence X_C^2 is asymptotically distributed as χ^2 with $L + 1$ degrees of freedom.

Discussion of Tests A, B and C. In the following we restrict ourselves to the case where we consider gaps of sizes 0, 1, 2, $\cdots$, $L - 1$ and pool gaps of size L or greater.

Now

$$\sum_{x=0}^{L-1} \frac{(N_x - n_0\,g_x)^2}{Npg_x} + \frac{(N_* - n_0\,g_*)^2}{Npg_*} = \sum_0^{L-1} \frac{(N_x - pM_x)^2}{Npqg_x} = X_{C*}^2.$$

This can easily be seen to be true for $L = 1$ and may be proved by induction, remembering that $M_L = N_*$. Hence X^2_{C*} is asymptotically equivalent to X^2_A. Asymptotically then, X^2_{C*} is a partitioning of X^2_A into independent χ^2 variates, each with one degree of freedom. Notice that X^2_{C*} is a modified form of X^2_C.

Also we may show (using the asymptotic likelihood found by Bartlett [2]) that the likelihood ratio test of the null hypothesis

$$p_{r_1\cdots r_{L+1}} = p \text{ (a specified value) if } r_{L+1} = 0$$

against the alternative

$p_{r_1\cdots r_{L+1}} = p_{r_x\cdots r_{L+1}} \neq p$ if $r_x = r_{L+1} = 0$ and $r_{X+1} = \cdots = r_L = 1$ for $x = 1, 2, \cdots, L$ (where these probabilities are unspecified)

and

$p_{r_1\cdots r_{L+1}} = p$ (specified) if $r_1 = \cdots = r_L = 1$ and $r_{L+1} = 0$,

is given by

$$-2\log\lambda = 2\sum_{x=0}^{L-1}\left\{N_x \log\left(\frac{N_x}{pM_x}\right) + (M_x - N_x)\log\left(\frac{M_x - N_x}{qM_x}\right)\right\}.$$

This may be shown (using methods similar to those used by Anderson and Goodman [1]) to be asymptotically equivalent, under the null hypothesis, to

$$\sum_{x=0}^{L-1}\frac{(N_x - pM_x)^2}{pqM_x},$$

which is asymptotically equivalent to X^2_{C*} and hence to X^2_A.

In the case where the null hypothesis does not specify the value p the likelihood ratio test for the null and alternative hypotheses above is given by

$$-2\log\lambda = 2\sum_{x=0}^{L-1}\left\{N_x\log\left(\frac{N_x N}{n_0 M_x}\right) + M_{x+1}\log\left(\frac{M_{x+1}N}{(N-n_0)M_x}\right)\right\}$$
$$+ 2N_*\log\left(\frac{N_* N}{n_0 M_*}\right) + 2(M_* - N_*)\log\left(\frac{(M_* - N_*)N}{(N-n_0)M_*}\right),$$

which, under the null hypothesis, is asymptotically distributed as χ^2 on L degrees of freedom and is asymptotically equivalent to

$$X^2_D = \sum_{x=0}^{L-1}\frac{(N_x - \hat{p}M_x)^2}{\hat{p}\hat{q}M_x} + \frac{(N_* - \hat{p}M_*)^2}{\hat{p}\hat{q}M_*}$$

where $\hat{p} = 1 - \hat{q} = n_0/N$.

It is interesting to note that

$$X^2_D = \sum_{x=0}^{L-1}\frac{(N_x - pM_x)^2}{\hat{p}\hat{q}M_x} + \frac{(N_* - pM_*)^2}{\hat{p}\hat{q}M_*} - \frac{(n_0 - Np)^2}{N\hat{p}\hat{q}}.$$

Under the null hypothesis

$$p_{r_1 \cdots r_{L+1}} = p \text{ (specified)} \quad \text{if} \quad r_{L+1} = 0$$

(which is the null hypothesis for the first likelihood ratio test considered) this is asymptotically equivalent to

$$X_{C*}^2 + \frac{(N_* - pM_*)^2}{Npqg_*} - \frac{(n_0 - Np)^2}{Npq} = X_C^2 - \frac{(n_0 - Np)^2}{Npq}.$$

It may be seen that X_C^2 is asymptotically equivalent, under the null hypothesis

$$p_{r_1 \cdots r_{L+1}} = p \text{ (specified)} \quad \text{if} \quad r_{L+1} = 0,$$

to the likelihood ratio test of this null hypothesis against the alternative hypothesis

$$p_{r_1 \cdots r_{L+1}} = p_{r_x \cdots r_{L+1}} \neq p \quad \text{if } r_x = r_{L+1} = 0 \text{ and } r_{x+1} = \cdots = r_L = 1 \text{ for}$$

$$x = 1, 2, \cdots, L.$$

(Notice that this is a slight modification of the first likelihood ratio rest considered, the only difference being that the alternative hypothesis here does not specify $p_{r_1 \cdots r_{L+1}}$ where $r_1 = r_2 = \cdots = r_L = 1$.)

The result for the asymptotic distribution of X_B^2 may be illustrated by noting that $X_B^2 = X_{C*}^2 + q(n_0 - Np)^2/Npq$.

Now $(n_0 - Np)^2/Npq$ and X_{C*}^2 are asymptotically distributed as χ^2 variates with 1 and L degrees of freedom respectively. However, these two χ^2 variates are not asymptotically independent. In fact we may see that X_{C*}^2 may be partitioned as follows:

$$X_{C*}^2 + \frac{(N_* - pM_*)^2}{Npqg_*} = \sum_{x=0}^{L-1} \frac{(N_x - pM_x - g_x(n_0 - Np))^2}{Npqg_x}$$

$$+ \frac{(n_0 - Np)^2}{Npq} + \frac{(N_* - pM_* - g_*(n_0 - Np))^2}{Npqg_*}$$

and for large values of L, and hence small values of g_*, the last terms on either side of this equation are approximately asymptotically equivalent and the first term on the right hand side is asymptotically distributed as an approximate χ^2 variate on $L - 1$ degrees of freedom, independently of $(n_0 - Np)^2/Npq$.

Hence for large values of L, the asymptotic distribution of X_B^2 is approximately that of a χ^2 variate on $L - 1$ degrees of freedom plus $(1 + q)$ multiplied by an independent χ^2 variate on 1 degree of freedom, this last χ^2 variate arising from the term $(n_0 - Np)^2 / Npq$.

This explains the result for the latent roots of the variance-covariance matrix associated with X_B^2. $L - 1$ of these latent roots are equal to 1 and, for large L, one is approximately equal to zero and the last is approximately equal to $1 + q$.

The approximation in the above is asymptotically $O_p(q^L)$ and q^L may not be particularly small in cases of interest. We are unlikely to be interested in extremely large values of L.

Extension of Test C to the case of Dependent Sequences. We consider now a non-cyclic dependent sequence. It will be easier to find first the likelihood ratio test and then the related χ^2 test, which may be regarded as an extension of Test C.

Consider the null hypothesis

$p_{r_1 \cdots r_{L+1}} = p_{r_{L-\mu+1} \cdots r_{L+1}} = p$ (specified) if $r_{L-\mu+1} = \cdots = r_L = 1$ and $r_{L+1} = 0$ (where μ is a fixed integer and $0 \leqq \mu \leqq L$)

against the alternative

$p_{r_1 \cdots r_{L+1}} = p_{r_{L-x} \cdots r_{L+1}} \neq p$ if $r_{L-x} = 0, r_{L-x+1} = \cdots = r_L = 1$ and $r_{L+1} = 0$ and $x = \mu, \mu + 1, \cdots, L - 1$

and

$p_{r_1 \cdots r_{L+1}} = p$ (specified) if $r_1 = \cdots = r_L = 1$ and $r_{L+1} = 0$.

Then if λ is the appropriate likelihood ratio

$$-2 \log \lambda = 2 \sum_{x=\mu}^{L-1} \left\{ n_{\cdots r_{L-x} \cdots r_L 0} \log \left(\frac{n_{\cdots r_{L-x} \cdots r_L 0}}{p n_{\cdots r_{L-x} \cdots r_L \cdot}} \right) + n_{\cdots r_{L-x} \cdots r_L 1} \log \left(\frac{n_{\cdots r_{L-x} \cdots r_L 1}}{q n_{\cdots r_{L-x} \cdots r_L \cdot}} \right) \right\}$$

and under the null hypothesis this is asymptotically distributed as χ^2 on $L - \mu$ degrees of freedom and is asymptotically equivalent to

$$\sum_{x=\mu}^{L-1} \frac{Y_{x,L}^2}{pqM_{x,L}}$$

where $Y_{x,L} = n_{\cdots r_{L-x} \cdots r_{L+1}} - p\, n_{\cdots r_{L-x} \cdots r_L \cdot}$ and

$$M_{x,L} = n_{\cdots r_{L-x} \cdots r_L} = \sum\nolimits_{r_1, \cdots, r_{L-x-1}} n_{r_1 \cdots r_{L-x-1} r_{L-x} \cdots r_L}$$

where $r_{L-x} = r_{L+1} = 0$ and $r_{L-x+1} = \cdots = r_L = 1$.

Consider the hypothesis $p_{r_1 \cdots r_{L+1}} = p_{r_{L-\mu+1} \cdots r_{L+1}}$, where the probabilities are specified, and in the particular case $r_{L-\mu+1} = \cdots = r_L = 1$ and $r_{L+1} = 0$ let us denote $p_{r_{L-\mu+1} \cdots r_{L+1}}$ by p. That is, we are considering a Markov chain of order at most μ with specified transition probabilities.

Under this hypothesis the likelihood ratio test statistic above is asymptotically equivalent to

$$\sum_{x=\mu}^{L-1} \frac{Y_{x,L}^2}{Npq\, P(r_{L-x} \cdots r_L)},$$

which is asymptotically distributed as a χ^2 variate with $L - \mu$ degrees of freedom. This is a possible extension of test C for μ-dependent sequences. Also it may be shown that each term $Y_{x,L} / (Npq\, P(r_{L-x} \cdots r_L))$ is asymptotically distributed as an independent χ^2 variate on 1 degree of freedom.

However, the likelihood ratio test is perhaps preferable since it is not necessary for the likelihood ratio test to specify, under the null hypothesis, the value of $p_{r_1 \cdots r_{L+1}}$ where any one of $r_{L-\mu+1}, \cdots, r_L$ is zero.

A test that is perhaps more useful is the likelihood ratio test where the null hypothesis does not specify values of probabilities. That is, we test the null hypothesis:

$p_{r_1 \cdots r_{L+1}} = p_{r_{L-\mu+1} \cdots r_{L+1}} = p$ (unspecified) where $r_{L-\mu+1} = \cdots = r_L = 1$ and $r_{L+1} = 0$

against the alternative

$p_{r_1 \cdots r_{L+1}} = p_{r_{L-x} \cdots r_{L+1}} \neq p$ when $r_{L-x} = r_{L+1} = 0$ and $r_{L-x+1} = \cdots = r_L = 1$ for $x = \mu, \mu + 1, \cdots, L - 1$.

In this case

$$\begin{aligned} -2 \log \lambda = 2 \sum_{x=\mu}^{L-1} \Bigg\{ & n_{\cdots r_{L-x} \cdots r_L 0} \log \left(\frac{n_{\cdots r_{L-x} \cdots r_L 0}}{\hat{p} n_{\cdots r_{L-x} \cdots r_L \cdot}} \right) \\ & + n_{\cdots r_{L-x} \cdots r_L 1} \log \left(\frac{n_{\cdots r_{L-x} \cdots r_L 1}}{\hat{q} n_{\cdots r_{L-x} \cdots r_L \cdot}} \right) \Bigg\} \\ & + 2 n_{s_1 \cdots s_L 0} \log \left(\frac{n_{s_1 \cdots s_L 0}}{\hat{p} n_{s_1 \cdots s_L \cdot}} \right) \\ & + 2 n_{s_1 \cdots s_L 1} \log \left(\frac{n_{s_1 \cdots s_L 1}}{\hat{q} n_{s_1 \cdots s_L}} \right) \end{aligned}$$

where $s_1 = s_2 = \cdots = s_L = 1$, $r_{L-x} = 0$, $r_{L-x+1} = \cdots = r_L = 1$. Also $\hat{p} = 1 - \hat{q} = n_{\cdots r_{L-\mu+1} \cdots r_L 0} / (n_{\cdots r_{L-\mu+1} \cdots r_L \cdot})$ where $r_{L-\mu+1} = \cdots = r_L = 1$.

This is an obvious extension of the likelihood ratio test associated with X_D^2 and under the null hypothesis $-2 \log \lambda$ given here is asymptotically distributed as χ^2 on $L - \mu$ degrees of freedom.

Some of the tests given above are anticipated in a statement by Goodman [6] on possible tests on his $_ik_j$ which is the same as our N_x with $x = i$ in the particular case $j = 1$, and evaluation of some of the variances and covariances is related to some work by Good [5a].

REFERENCES

[1] T. W. Anderson and Leo A. Goodman, "Statistical inference about Markov chains," *Ann. Math. Stat.*, Vol. 28 (1957), pp. 89–110.

[2] M. S. Bartlett, "The frequency goodness of fit test for probability chains," *Proc. Camb. Phil. Soc.*, Vol. 47 (1951), pp. 86–95.

[3] Patrick Billingsley, "Asymptotic distributions of two goodness of fit criteria," *Ann. Math. Stat.*, Vol. 27 (1956), pp. 1123–1129.

[4] W. G. Cochran, "The distribution of quadratic forms in a normal system with application to the analysis of covariance," *Proc. Camb. Phil. Soc.*, Vol. 30 (1934), pp. 178–191.

[5] Harold Cramér, *Mathematical Methods of Statistics*, Princeton University Press, Princeton, 1946.

[5a] I. J. GOOD, "The serial test for sampling number and other tests for randomness," *Proc. Camb. Philos. Soc.*, Vol. 49 (1953), pp. 276–284.

[6] LEO A. GOODMAN, "Simplified run tests and likelihood ratio tests for Markov chains," *Biometrika*, Vol. 45 (1958), pp. 181–197.

[7] M. G. KENDALL AND B. BABINGTON-SMITH, "Randomness and random sampling numbers," *J. Roy. Stat. Soc.*, Vol. 101 (1938), pp. 147–166.

[8] H. A. MEYER, L. S. GEPHART AND N. L. RASMUSSEN, "On the generation and testing of random digits," WADC Technical Report 54-55, Wright-Patterson Air Force Base, Ohio, 1954.