DISTRIBUTION OF MANN AND WHITNEY AND ALLIED STATISTICS FOR TWO STATE MARKOV CHAIN

P. V. KRISHNA IYER & D. RAY

Defence Science Laboratory, Delhi

In this paper the distribution of ab transitions separated by 0, 1, 2, ...r observations and also of T_3, T_4, and T_n for a Markov Chain having two states a and b has been discussed. The power of these statistics for comparing the randomness of a binomial sequence against a Markov Chain alternative wherein the asymptotic probabilities of the two states are the same for the null hypothesis H_0 (which is possible only when the alternative hypothesis is a Markov Chain) has been considered. The relative efficiencies of the various statistics show that the tests based on ab transitions separated by nk observations (when k is a fraction) will in general be better than Mann and Whitney statistics for testing two samples where the distributions are not known.

In an earlier paper [1] we obtained the first two moments and product moments for the number of transitions like aa, ab etc., between adjoining observations of a two-state Markov Chain. Iyer and Singh[2] discussed the distribution of the statistics T_r which is the sum of transitions separated by 0, 1, 2, 3.....r—2 observations for binomial and multinomial sequences without any restriction on the observations separating them. In this paper we propose to discuss the distributions of T_3, T_4, and T_n for a two-state Markov Chain. We also discuss the distribution of transitions like ab separated by 0, 1, 2, 3, ...r observations which may be defined as transitions of 1, 2, ...$(r+1)th$ order respectively.

Goodman[3] has discussed the distribution of t-tuple which represents a set of given consecutive observations defined by $n = (u_1, u_2 \ldots\ldots\ldots u_t)$ where $u_1, u_2 \ldots\ldots$ are the various states of the mth order Markov Chain. Bhatt[4] has obtained the first two moments of t-tuple in discussing the distribution of psi-square defined by $\Psi^{t2} = \Sigma_u [(n_u - m_u)^2/mn]$ where n_u is the frequency of the t-tuple in a sequence of length $n+t-1$ and m_u its expected value in a new sequence of the same length. Results obtained so far will not enable us to consider the distribution of T_r 's. The distribution of T_3 & X_2 and the first two moments of the Statistics $X_1 (= T_2)$, X_2, X_{r+1}, T_3, T_4 and the relative powers of Mann and Whitney (Wilcoxon) and allied Statistics as compared to t-test are discussed in this paper.

DISTRIBUTION OF TRANSITIONS OF DIFFERENT ORDER

Discussion of the distribution of the number of transitions of order $(r+1)$ and of T_r 's are extremely complicated and can be done only in matrix form. For the sake of illustration we shall obtain the difference equations satisfied by the probability generating function for the distribution of transitions of order 2 and of T_3. The procedure followed here can be extended readily for the general case.

DISTRIBUTION OF TRANSITIONS OF ORDER TWO (X_2)

Let $\phi\,[n, (t_{jk})]$ where (t_{jk}) is the 2×2 matrix for the states a and b, be the probability generating function for the joint distribution of the number of transitions like (jk) of order 2, for a sequence of n observation. For the sake of simplicity $\phi\,[n, (t_{jk})]$ is denoted by $\phi\,(n)$ in subsequent discussions. Assume $\phi_{jk}\,(n)$ to be the conditional probability generating function when the last two observations are jk. It can be easily seen that

$$\phi\,(n) = \phi_{aa}\,(n) + \phi_{ab}(n) + \phi_{ba}(n) + \phi_{bb}\,(n) \qquad (1)$$

Further $\phi_{jk}\,(n)$ satisfies the difference equation $\phi_{jk}\,(n) = \sum_{i\,=\,a\,or\,b} t_{ik}\,\phi_{ij}\,(n-1)$ (2)

Following the arguments given in an earlier paper by Iyer[5] and using (1) and (2) it can be seen that $\phi\,(n)$ satisfies the difference equation:

$$\begin{bmatrix} E - P_{aa}\,t_{aa} & O & -p_{aa}\,t_{ba} & O \\ -p_{ab}\,t_{ab} & E & -p_{ab}\,t_{ba} & O \\ O & p_{ba}\,t_{aa} & E & -p_{ba}\,t_{ba} \\ O & p_{bb}\,t_{ab} & O & E - p_{bb}\,t_{bb} \end{bmatrix} \phi\,(n) = O$$

or $\{E^4 - E^3(p_{aa}\,t_{aa} + p_{bb}\,t_{bb}) + E^2\,t_{aa}\,t_{bb}\,(p_{aa}\,p_{bb} - p_{ab}\,p_{ba}) + E\,p_{ab}\,p_{ba}\,(t_{aa}\,t_{bb} - t_{ab}\,t_{ba})\,(p_{aa}\,t_{aa} + p_{bb}\,t_{bb}) - p_{aa}\,p_{bb}\,p_{ab}\,p_{ba}\,(t_{aa}\,t_{bb} - t_{ab}\,t_{ba})^2\}\,\phi\,(n) = O \qquad (3)$

The general solution of this equation is given by

$$\phi\,(n) = \sum_{r\,=\,1}^{4} A_r\,\lambda_r^{\,n}$$

Where λ's are the roots of the biquadratic equation (3). A's are determined by using the initial conditions. As it is difficult to determine the roots of the above equation it would be possible to obtain the generating functions by using the difference equation in succession. It may be mentioned that for large values of n, $\phi\,(n)$ may be approximated to $A\,\hat{\lambda}^r$ where $\hat{\lambda}$ is the maximum root of (3). The moments can be obtained by the procedure developed by Iyer and Kapur[6]. We have however obtained the moments from simpler considerations in subsequent sections.

Before considering the moments we shall show that the distribution can also be obtained on the following lines which is similar to that given by Goodman[7]. Assuming that $n_{a.a}$, $n_{a.b}$, $n_{b.a}$ and $n_{b.b}$ be the number of transitions of second order in a given sequence of n observations where the first two initial observations are say, aa and the last two are, say ab, the transitions for $a.a$ can be considered to be the sum of the triplets aaa and aba. Similarly $a.b$, $b.a$ and $b.b$ are the sum of the frequencies for the triplets noted below:

$$n_{a.a} = n_{aaa} + n_{aba.}$$
$$n_{a\cdot b} = n_{aab} + n_{aba.}$$
$$n_{b\cdot a} = n_{baa} + n_{bba.}$$
$$n_{b.b} = n_{bab} + n_{bbb.}$$

In view of these relations, the second order transitions $n_{a.a}$, $n_{a.b}$, $n_{b.a}$ and $n_{b\cdot b}$ can arise from all bi-partitions of the numbers $n_{i.j}$, each partition representing the number of triplets in the sequence. Thus the probability for second order transitions can be written by modifying Goodman's procedure by forming contingency tables.

TABLE 1

CONTINGENCY TABLE SHOWING THE NUMBER OF TRIPLETS IN THE SEQUENCE COMMENCING IN *aa* AND ENDING IN *ab*

Ending observation of the triplet / Ist two observations of the triplet	*a*	*b*	Total
aa	n_{aaa}	n_{aab}	$n_{aa.}$
ab	n_{aba}	n_{abb}	$n_{ab.}$
ba	n_{baa}	n_{bab}	$n_{ba.}$
bb	n_{bba}	n_{bbb}	$n_{bb.}$
Total	$n_a - 2$	n_b	$n-2$

The probability for sequences having the triplets indicated in the above contingency table is equal to

$$\frac{n_{aa.}!\, n_{ab.}!\, n_{ba.}!\, n_{bb.}!\, (n_a - 2)!\, n_b\,!}{n_{aaa}\,!\, n_{aba}\,!\, n_{baa}\,!\, n_{bba}\,!\, n_{aab}\,!\, n_{abb}\,!\, n_{bab}\,!\, n_{bbb}\,!\, (n-2)\,!} \times \frac{n_a\,(n_a - 1)}{n\,(n-1)}\, G$$

when G is the co-factor of the matrix

$$\begin{array}{c} \\ aa \\ ab \\ ba \\ bb \end{array} \begin{array}{c} \begin{array}{cccc} aa & & ab & \end{array} \\ \begin{bmatrix} 1 - \dfrac{n_{aaa}}{n_{aa.}} & 0 & - \dfrac{n_{baa}}{n_{ba.}} & 0 \\ - \dfrac{n_{aab}}{n_{aa.}} & 1 & - \dfrac{n_{bab}}{n_{ba.}} & 0 \\ 0 & - \dfrac{n_{aba}}{n_{ab.}} & 1 & - \dfrac{n_{bba}}{n_{bbb.}} \\ 0 & - \dfrac{n_{abb}}{n_{ab.}} & 0 & 1 - \dfrac{n_{bbb}}{n_{bb.}} \end{bmatrix} \end{array} \begin{array}{c} \\ ba \\ bb \end{array}$$

obtained by removing the rows for *aa* and columns for *ab*. Similarly if it ends in *ba*, we remove the rows for *aa* and columns for *ba*; for *aa* it will be *aa* and *aa*, for *bb* it will be *bb* and *bb*. Let these be denoted by G_1, G_2, G_3 and G_4. The number of possible arrangements is equal to $\dfrac{n!}{n_a\,!\, n_b\,!}$

The probability for each of these tables is obtained by multiplying the number of arrangements for the respective table by

$$P_a\, p_{aa}\, (p_{aa})^{n_{aaa} + n_{bba}} \cdot (p_{ab})^{n_{aab} + n_{bab}} \cdot (p_{ba})^{n_{aba} + n_{bba}} \cdot (p_{bb})^{n_{abb} + n_{bbb}}$$

The sum of the four probabilities gives the probability for the sequence to begin in (*aa*). Similarly the probabilities for sequences starting in *ab*, *ba* and *bb* are determined. The sum of all these sixteen values give the probability for given number of *a*'s and *b*'s say n_a and n_b. Similar expressions can be obtained for all possible values of n_a and n_b which are partitions of n (the number of observations in the sequence).

DISTRIBUTION OF T_3

Following the same lines as in (*a*) we can obtain the difference equation satisfying the probability generating function for T_3 and it reduces to

$$\left\{ E^4 - E^3 (p_{aa} t_{aa}^2 + p_{bb} t_{bb}^2) + E^2 t_{aa} t_{bb} (p_{aa} p_{bb} t_{aa} t_{bb} - p p_{ba} t_{ab} t_{ba}) \right.$$

$$- E p_{ab} p_{ba} t_{ab} t_{ba} (p_{aa} t_{aa}^2 + p_{bb} t_{bb}^2) (t_{aa} t_{bb} - t_{ab} t_{ba})$$

$$\left. - p_{aa} p_{bb} p_{ab} p_{ba} t_{aa} t_{bb} t_{ab} t_{ba} (t_{aa} t_{bb} - t_{ab} t_{ba}^2) \right\} \phi(n) = O$$

The probability generating function for different values of n can be obtained in succession from the lower ones. As already mentioned above it is not easy to obtain the cumulants for the above distribution. We shall, therefore, obtain the moments of this distribution and for others like T_4 and T_n from other considerations. Using the procedure followed by Iyer and Kapur, it can be shown that cumulants of T_r are linear in n when r is finite and therefore the distributions of T_2, T_3, T_4.....approaches the normal form asymptotically for finite values of r. For large values of r, it has been found in connection with a number of other investigations that the cumulants take the form

$$\kappa_t = \phi_{1t}(n, r)\ \phi_t\ (r)$$

where $\phi_{1t}(n, r)$ is a function in n and r of degree 1 of the form $(fn + gr + h)$ and $\phi_t\ (n, r)$ is a function in r of degree t,

$$\text{Now } \gamma_1 = \frac{\kappa_3}{\kappa_2^{3/2}} = \frac{\phi_{13}(n, r)\,\phi_3(r)}{\{\phi_{12}(n, r)\}^{3/2}\ \{\phi_2(r)\}^{3/2}} \sim O\left(\frac{1}{\sqrt{n}}\right)$$

$$\gamma_2 = \frac{\kappa_4}{\kappa_2^2} = \frac{\phi_{14}(n, r)\phi_4(r)}{\{\phi_{12}(n, r)\}^2\ \{\phi_2(r)\}^2} \sim O\left(\frac{1}{n}\right)$$

Hence the distribution of statistics tend to normal form asymptotically for all values of r.

CUMULANTS FOR *ab* TRANSITIONS OF $(r+1)th$ ORDER

Let $(r+1)_{th}$ order statistic be represented by

$$X_{r+1} = x(1,r+2) + x(2,r+3) + \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots x(n-r-1,n)$$

where $x\ (s,\ r+s+1)$ stands for transitions between s_{th} and $(r+s+1)_{th}$ observation. Assume $x(s,r+s+1)$ takes the value 1 if transition is ab, and 0 otherwise. It has been shown[1] that for a simple two-state chain, the conditional probability for the s_{th} observation to be in state a or b when the t_{th} one $(t<s)$ is in a given state are as follows:

$$P_a\ (s/t_a) = (p_2 + q_1\delta^{s-t})/(1 - \delta)$$
$$P_a\ (s/t_b) = p_2(1 - \delta^{s-t})/(1 - \delta)$$
$$P_b\ (s/t_a) = q_1(1 - \delta^{s-t})/(1 - \delta)$$
$$P_b\ (s/t_b) = (q_1 + p_2\delta^{s-t})/(1 - \delta)$$

where $\begin{pmatrix} p_1 & p_2 \\ q_1 & q_2 \end{pmatrix}$ is the transition matrix and $\delta = p_1 - p_2$. Using these results the first two moments of this statistics are given by

$$E(X_{r+1}) = (n-r-1)\frac{p_2q_1}{1-\delta}(1-\delta^{r+1}) \tag{4}$$

and

$$V(X_{r+1}) = \frac{p_2q_1}{(1-\delta)^2}(1-\delta^{r+1})(n-r-1)$$
$$+\frac{p_2^2q_1^2}{(1-\delta)^4}(1-\delta^{r+1})^2\left\{(r+1)(3r+5)-n(2r+3)-\frac{2\delta}{1-\delta}D_{n-2r-3}\right\}$$
$$+\frac{2p_2^2q_1^2}{(1-\delta)^4}\left\{\frac{r(n-3r-3)}{2}+\frac{(n-r-2)\delta^{r+1}}{1-\delta}-\delta^2\frac{1-\delta^{r-1}}{(1-\delta)^2}-\frac{(n-2r-1)\delta}{1-\delta}\right.$$
$$+\frac{(n-r-2)\delta^2}{1-\delta^2}-\frac{(n-r-2)\delta^{r+2}}{1-\delta}-\delta^4\cdot\frac{1-\delta^{2r-2}}{(1-\delta^2)^2}+\delta^{r+1}\cdot\frac{1-\delta^{r-1}}{(1-\delta)^2}$$
$$\left.-\frac{(n-2r-1)\delta^{2r+2}}{1-\delta^2}+\frac{(n-2r-1)\delta^{2r+2}}{1-\delta}\right\}$$
$$+\frac{2p_2q_1(p_2^2+q_1^2)}{(1-\delta)^4}\left\{\frac{n-r-2}{1-\delta}\delta-\delta^2\frac{1-\delta^{r-1}}{(1-\delta)^2}-\frac{(n-2r-1)\delta^{r+1}}{1-\delta}\right.$$
$$\left.-\frac{r}{2}(2n-3r-3)\delta^{r+1}\right\} \tag{5}$$

where

$$D_{n-2r-3} = \left\{(n-2r-3)-\frac{\delta}{1-\delta}(1-\delta^{n-2r-3})\right\}$$

Putting $r=0$ and 1 we get the moments for X_1 and X_2 the first and second order statistics. They are given below :

$$E(X_1) = (n-1)\frac{p_2q_1}{1-\delta} \tag{6}$$

$$V(X_1) = (n-1)\frac{p_2q_1}{1-\delta}+\frac{p_2^2q_1^2}{(1-\delta)^2}\left\{(-3n+5)-\frac{2\delta}{1-\delta}D_{n-3}\right\} \tag{7}$$

$$E(X_2) = (n-2)\frac{p_2q_1}{1-\delta}(1+\delta) \tag{8}$$

$$V(X_2) = (n-2)\frac{p_2q_1(1+\delta)}{1-\delta}+\frac{p_2^2q_1^2}{(1-\delta)^2}(1+\delta)^2\left\{(-5n+16)-\frac{2\delta}{1-\delta}D_{n-5}\right\}$$
$$+\frac{2(n-3)p_2q_1}{(1-\delta)^3}(p_2q_1+p_2^2\delta+q_1^2\delta+p_2q_1\delta^2) \tag{9}$$

CUMULANTS OF T's

By definition $T_r = \sum_{k=1}^{r-1} X_k$, the first two moments for T_3, T_4 and T_n are given below. It is obvious that $T_2 = X_1$

(i) T_3, the sum of transitions of ab of 1st and 2nd order

$$E(T_3) = (2n-3)\frac{p_2q_1}{1-\delta}+(n-2)\frac{p_2q_1\delta}{1-\delta} \tag{10}$$

and

$$V(T_3) = (4n-7)\frac{p_2q_1}{1-\delta} + (5n-12)\frac{p_2q_1\delta}{1-\delta} + \frac{p_2^2q_1^2}{(1-\delta)^2}\left\{(-14n+35) - \frac{2\delta}{1-\delta}(D_{n-3} + 2D_{n-4} + D_{n-5})\right\}$$
$$+ \frac{2p_2^2q_1^2\delta}{(1-\delta)^2}\left\{(-10n+29) - \frac{2\delta}{1-\delta}(D_{n-4} + D_{n-5})\right\} + \frac{p_2^2q_1^2\delta^2}{(1-\delta)^2}\left\{(-5n+16) - \frac{2\delta}{1-\delta}D_{n-5}\right\} \quad (11)$$

(ii) T_4, the sum of transitions of ab of 1st, 2nd and 3rd orders

$$E(T_4) = (3n-6)\frac{p_2q_1}{1-\delta} + (2n-5)\frac{p_2q_1\delta}{1-\delta} + \frac{p_2q_1\delta^2}{1-\delta}(n-3) \quad (12)$$

and

$$V(T_4) = (9n-22)\frac{p_2q_1}{1-\delta} + (16n-49)\frac{p_2q_1\delta}{1-\delta} + (9n-33)\frac{p_2q_1\delta^2}{1-\delta}$$
$$+ \frac{p_2^2q_1^2}{(1-\delta)^2}\left\{(-33n+110) - \frac{2\delta}{1-\delta}(D_{n-3} + 2D_{n-4} + 3D_{n-5} + 2D_{n-6} + D_{n-7})\right\}$$
$$+ \frac{2p_2^2q_1^2\delta}{(1-\delta)^2}\left\{(-34n+125) - \frac{2\delta}{1-\delta}(D_{n-4} + 2D_{n-5} + 2D_{n-6} + D_{n-7})\right\}$$
$$+ \frac{p_2^2q_1^2\delta^2}{(1-\delta)^2}\left\{(-60n+245) - \frac{2\delta}{1-\delta}(3D_{n-5} + 4D_{n-6} + 3D_{n-7})\right\}$$
$$+ \frac{2p_2^2q_1^2\delta^3}{(1-\delta)^2}\left\{(-14n+62) - \frac{2\delta}{1-\delta}(D_{n+6} + D_{n-7})\right\} + \frac{p_2^2q_1^2\delta^4}{(1-\delta)^2}\left\{(-7n+33) - \frac{2\delta}{1-\delta}D_{n-7}\right\} \quad (13)$$

(iii) T_n, sum of ab up to order $(n-1)$

$$E(T_n) = \frac{n(n-1)p_2q_1}{2(1-\delta)^2} - \frac{(n-1)\delta p_2q_1}{(1-\delta)^3} \quad (14)$$

$$V(T_n) = \frac{n(n-1)(2n-1)}{6(1-\delta)^2}p_2q_1 - \frac{(n-1)(1+\delta)-2\delta}{(1-\delta)^4}p_2q_1\delta$$
$$- \frac{p_2^2q_1^2\, n(n-1)(2n-3)}{2(1-\delta)^4} + \frac{p_2q_1\delta(n-1)}{3(1-\delta)^5}\left\{3p_2q_1n(n-1) + (n-2)(n-3)(2p_2^2 + 2q_1^2 - 5p_2q_1)\right\}$$
$$- \frac{p_2q_1\delta^2}{(1-\delta)^6}\left\{p_2q_1(n-1)^2 + 2(n-2)(n-3)(2p_2^2 + 2q_1^2 - 5p_2q_1)\right\}$$
$$+ \frac{12p_2q_1\delta^3(n-3)(p_2^2 + q_1^2 - 3p_2q_1)}{(1-\delta)^7} \quad (15)$$

These moments for T_n have been calculated on the basis of results given in Appendix. It may be noted that T_2 corresponds to ab transition between successive observations considered by Goodman[7], Bartlett[8] and others and T_n to the Mann and Whitney statistics U for two samples with the difference that U refers to finite sampling whereas T_n refers to sampling with fixed probabilities.

RELATIVE EFFICIENCY OF DIFFERENT STATISTICS

The statistics Tr's discussed above can be used for testing the randomness of a binomial sequence. Taking for example, a given sequence of observations, each assuming the states a and b with probabilities p and q respectively, the randomness of the sequence can be tested by noting the statistics T_3, T_4..... T_n and applying the usual standardised deviate test for the hypothesis $p=p_o$, $q=(1-p_o)$. We may compare the relative efficiencies of T_r's when the alternative H_1 is a Markov Chain sequence defined by the transition matrix (p_{ij}), $i, j,=1, 2$, such that asymptotic probabilities for the two states are the same as for H_o. It will be observed that such an alternative can exist only for a Markov Chain and does not seem to have been considered before.

Pitman [9] has shown that the asymptotic relative efficiency of two statistics T_n and T_n * which are the functions of a parameter say $\theta_\circ$, are given by

$$\underset{n \to \infty}{\text{Lt}} \quad \frac{1}{\sigma_n{}^2(\theta_\circ)} \left(\frac{d\mu_n(\theta_\circ)}{d\theta_\circ} \right)^2 \Big/ \frac{1}{\sigma_n{}^{*2}(\theta_\circ)} \left(\frac{d\mu_n{}^*(\theta_\circ)}{d\theta_\circ} \right)^2$$

where $\mu_n(\theta_\circ)$ and $\mu_n{}^*(\theta_\circ)$ are the expected values of T_n and T_n * and $\sigma^2{}_n(\theta_\circ)$ and $\sigma_n{}^{*2}(\theta_\circ)$ are the corresponding variances of the statistics. Mood [10] has shown that this result can also be established by considering the relative changes in the power of the tests for the alternative hypothesis on the assumption that there is no change in the variances of the respective statistics under the alternative hypothesis. This assumption is not fully justified. This can be seen from Table 2 which gives the expected values and variances of the statistics T_2, T_3, T_4 and T_n for different hypothesis.

TABLE 2

EXPECTED VALUES AND VARIANCES OF DIFFERENT STATISTICS FOR TWO STATE MARKOV CHAIN OF SIZE 3

Particulars of chain / Statistics	p = ·5, q = ·5, δ = ·01		p = ·5, q = ·5, δ = ·001		p = ·6, q = ·4, δ = ·005	
	Expected value	Variance	Expected value	Variance	Expected value	Variance
T_2 (= X_1)	7·18	1·937	7·24	1·938	6·92	2·062
T_3	14·11	4·262	14·24	4·195	13·61	4·980
T_4	20·79	7·160	20·98	7·016	20·06	9.914
T_n	108·68	600·307	108·74	590·177	104·36	629·866

Table 2 shows that both the mean and the variances change under varying hypothesis. In view of the fact that the variances for the two hypothesis also differ, the conclusions regarding the relative efficiencies of different tests on the basis of Pitman's results need further examination. We shall examine the relative efficiencies of the tests based on T_r 's by evaluating the power of the different statistics taking into account the changes in both

TABLE 3

INCREASE IN POWER $(\xi + \eta^2/\sigma_0^2)$ FOR DIFFERENT STATISTICS FOR TESTING RANDOMNESS OF A BINOMIAL SEQUENCE AGAINST MARKOV CHAIN ALTERNATIVE SEQUENCE

Statistics / Hypothesis				T_2 (= X_1)	T_3	T_4	T_h
	n	p	q				
1	2	3	4	5	6	7	8
0·01	30	·5	·5	·002623	·019226	·023506	·019097
		·6	·4	·000917	·015466	·019138	·018877
		·75	·25	—·004385	·007365	·011400	·018045
	50	·5	·5	·004629	·021020	·025659	·019566
		·6	·4	·002472	·016672	·020326	·019404
		·75	·25	—·003683	·007733	·011681	·018928
	100	·5	·5	·009606	·024197	·028661	·019894
		·6	·4	·006853	·018932	·022076	·019562
		·75	·25	—·001501	·007555	·012133	·019551
0·005	30	·5	·5	·000655	·009281	·011498	·009497
		·6	·4	—·001070	·007481	·009390	·009389
		·75	·25	—·002484	·003601	·005650	·008980
	50	·5	·5	·001157	·010534	·012396	·009731
		·6	·4	·000305	·008238	·009862	·009652
		·75	·25	—·002303	·003714	·005755	·009407
	100	·5	·5	·002401	·010951	·013454	·009895
		·6	·4	·001363	·008602	·010441	·009731
		·75	·25	—·001792	·003945	·005892	·009727
0·001	30	·5	·5	·000026	·001804	·002260	·001891
		·6	·4	—·000112	·001456	·001849	·001870
		·75	·25	—·000543	·000707	·001122	·001789
	50	·5	·5	·000046	·001892	·002412	·001938
		·6	·4	—·000098	·001508	·001926	·001885
		·75	·25	—·000541	·000720	·001105	·001875
	100	·5	·5	·000096	·001980	·002555	·001971
		·6	·4	—·000058	·001562	·001996	·001938
		·75	·25	—·000525	·000733	·001152	·001938

the expected values and their variances. Recently Bhattacharyya[10] has obtained a general expression for the power of two tailed test based on any statistics T for an alternative H_1 against H_0 as follows :

$$P = \alpha + k\ \phi(k)\left[\xi + \frac{\eta^2}{\sigma_0^2} + \frac{\xi^2(k^2-3)}{4} + \frac{k^4-10k^2+15}{24}\xi^3 + \frac{(k^2-3)}{2}\xi\frac{\eta^2}{\sigma_0^2} + \ldots\right]$$

for the significance level α defined by

$$\alpha = 1 - \int_{-k}^{k}\phi(t)\,dt \text{ where } \phi(t) = \frac{1}{2\pi}\,e^{-t^2/2} \text{ and } \phi(k) = [\phi(t)]_{t=k}$$

It is assumed here that the asymptotic distribution of T is normal with mean and variance m_0 and σ_0^2 for H_0 and $m_1 = m_0 + \eta$ and $\sigma_1^2 = \sigma_0^2(1+\xi)$ for H_1.

Using the above expression we find that the increase in power is proportional to $\xi + \eta^2/\sigma_0^2$ neglecting the higher powers of ξ and η^2. The value $\xi + \eta^2/\sigma_0^2$ has been tabulated in Table 3 for the various statistics T_2, T_3, T_4 and T_n against different two state Markov Chain alternatives for testing the randomness of a binomial sequence. As pointed out earlier it has been assumed that for the Markov Chain alternative the asymptotic probabilities for a and b i.e. $p_2/1-\delta$ and $q_2/1-\delta$ are equal to p and q respectively where p and q are the probabilities of occurrence of a and b in the binomial sequence.

These powers have also been shown in the graphs 1, 2 and 3. It will be seen from the graph that when p lies between 5 and ·75, the powers of T_4 is definitely more than that of T_n for all positive values of δ.

For negative values of δ the increase in the power is negative for all the statistics except T_2 as can be seen from Table 4. The maximum decrease in power is for T_n.

TABLE 4

INCREASE IN POWER FOR DIFFERENT STATISTICS FOR $n = 100$

Statistics / Alternative hypothesis	T_2 (= X_1)	T_3	T_4	T_n
p = ·5, q = ·5, δ = —·01	·009609	—·000052	—·022026	—·019489
p = ·6, q = ·4, δ = —·01	·009661	—·000708	—·017562	—·019169

The increase in power has also been calculated assuming the null hypothesis to be a Markov Chain sequence. The results for such cases have been tabulated in Table 5 for a few hypothesis.

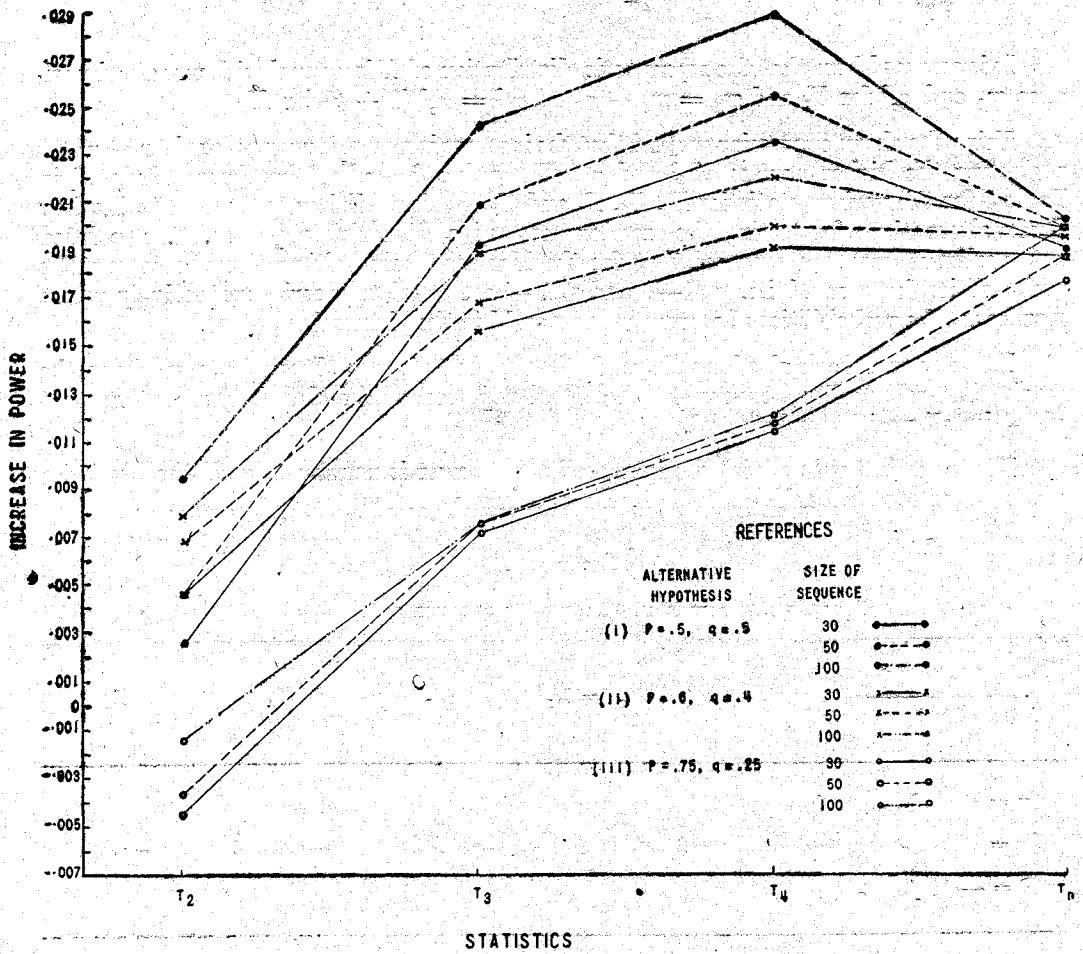

Fig. 1—Graph showing the increase in power $(\xi + \eta^2/\sigma_0^2)$ for different statistics and $\delta = \cdot 01$

TABLE 5

INCREASE IN POWER WHEN BOTH THE NULL AND ALTERNATIVE HYPOTHESIS ARE MARKOV CHAIN SEQUENCES

Hypothesis		Increase in power for Statistics											
		T_2 (=X_1)			T_3			T_4			T_n		
H_0	H_1	$n=30$	$n=50$	$n=100$	$n=30$	$n=50$	$n=100$	$n=30$	$n=50$	$n=100$	$n=30$	$n=50$	$n=100$
p=·5 q=·5 δ=·01	p=·5 q=·5 δ=·005	·000748	·001249	·002500	–·008506	–·008640	–·008252	–·010972	–·011519	–·011713	–·009412	—·009640	—·009800
p=·5 q=·5 δ=·005	p=·5 q=·5 δ=·01	·000608	·001105	·002352	·009222	·009864	·010707	·011489	·012388	·013454	·009506	·009737	·009930
p=·5 q=·5 δ=·01	p=·6 q=·4 δ=·001	·081918	·114261	·016514	·221056	·286442	·422565	·319435	·435347	·665902	·077970	·101839	·16544

The Table shows that when the asymptotic probabilities for states a and b remain same for both H_0 and H_1, δ being positive and different such that δ of H_1 is less than that of H_0, the increase in power is negative for all statistics except T_2. But when the δ of H_1 is greater than δ of H_0; the increase in power is positive for all and T_4 is more powerful than T_n . When the null and alternative hypothesis have different transitions matrix such that asymptotic probabilities are also different the increase in power is again positive and T_3 and T_4 are more powerful than T_n.

The approximate efficiency of statistics T_4 compared to T_n for different positive values of δ and for different hypothesis is given in Table 6.

Our previous investigations on random sequences have shown that for $n=100$, the power for T_{15} is maximum and is 148% for H_0: $p=\cdot 5$, $q=\cdot 5$ and H_1: $p=\cdot 55$, $q=\cdot 45$ and 160% for H_0: $p=\cdot 2$, $q=\cdot 8$ and H_1: $p=\cdot 15$, $q=\cdot 85$. These findings are in line with our present investigation wherein the alternative hypothesis include various types of Markov Chains. Therefore it is clear that tests based on T_3, T_4, etc. are in general definitely better than T_n.

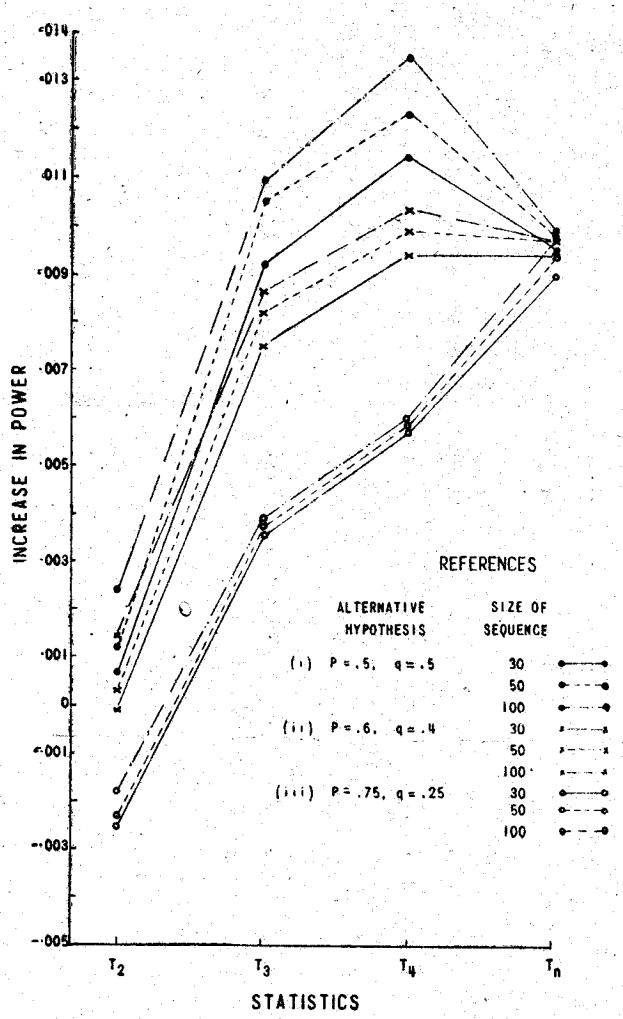

Fig. 2—Graph showing the increase in power $(\xi + \eta^2/\sigma_0^2)$ for different statistics and $\delta = \cdot 005$

TABLE 6

EFFICIENCY OF T_4 AS COMPARED TO T_n

Hypothesis	Percentage relative efficiency of T_4 with respect to T_n		
	$n = 30$	$n = 50$	$n = 100$
(1) p = ·5 q = ·5 δ = ·01	123	131	144
(2) p = ·6 q = ·4 δ = ·005	100	102	107
(3) p = ·75 q = ·25 δ = ·001	63	59	59

We may now examine how far these conclusions are applicable for comparing two samples on the basis of Mann and Whitney's test. In this test, we arrange the two samples A and B by pooling them together in ascending or descending order and identifying the observations as a and b according as the observation belongs to A and B. We may consider such a sequence a binomial one with probabilities $p=n_1/(n_1+n_2)$ and $q=n_2/(n_1+n_2)$ where n_1 and n_2 are the sizes of the samples. If n_1 and n_2 are fairly large, the distribution of T_n approximates to that of Mann and Whitney Statistics.

Table 7 shows expectation and variance of T_n and U for different sample sizes.

TABLE 7

EXPECTED VALUES AND VARIANCES FOR T_n & U

$n_1 = n_2$	$n = 2n_1$	T_n		U		Difference of T_n and U as % of U	
		Expected value	Variance	Expected value	Variance	Expected value	Variance
50	100	1237·5	21140·6	1250·0	21041·7	1·00	0·47
100	200	4975·0	167906·0	5000·0	187500·0	0·50	0·24
200	400	1995·0	1338312·5	20000·0	1336666·7	0·25	0·12

The percentage differences in the moments of these two statistics are very small and are of the order of $1/n$ and therefore the conclusions arrived at from considerations of binomial sequences or from those of two samples sequences will not differ much from each other for all practical purposes.

It would be noted that when the two samples belong to two distribution functions $F(x)$ and $G(x)$ the probability that any of the observations say x_r of the ordered sequence belongs to F is equal to $\frac{n_1 f(x_r)}{n_1 f(x_r) + n_2 g(x_r)}$. This consideration shows that the sequence can be approximated to a binomial one with probabilities $p = \frac{n_1}{n_1 + n_2}$ and $q = \frac{n_2}{n_1 + n_2}$

This is true when $f(x) = g(x)$. If $f(x) \neq g(x)$, the change in p can be approximated to $p+t$ which is the alternative H_1 discussed in an earlier paper[2]. Thus for discussions of relative power it appears to be more rational and simpler to treat the sequence as a binomial one

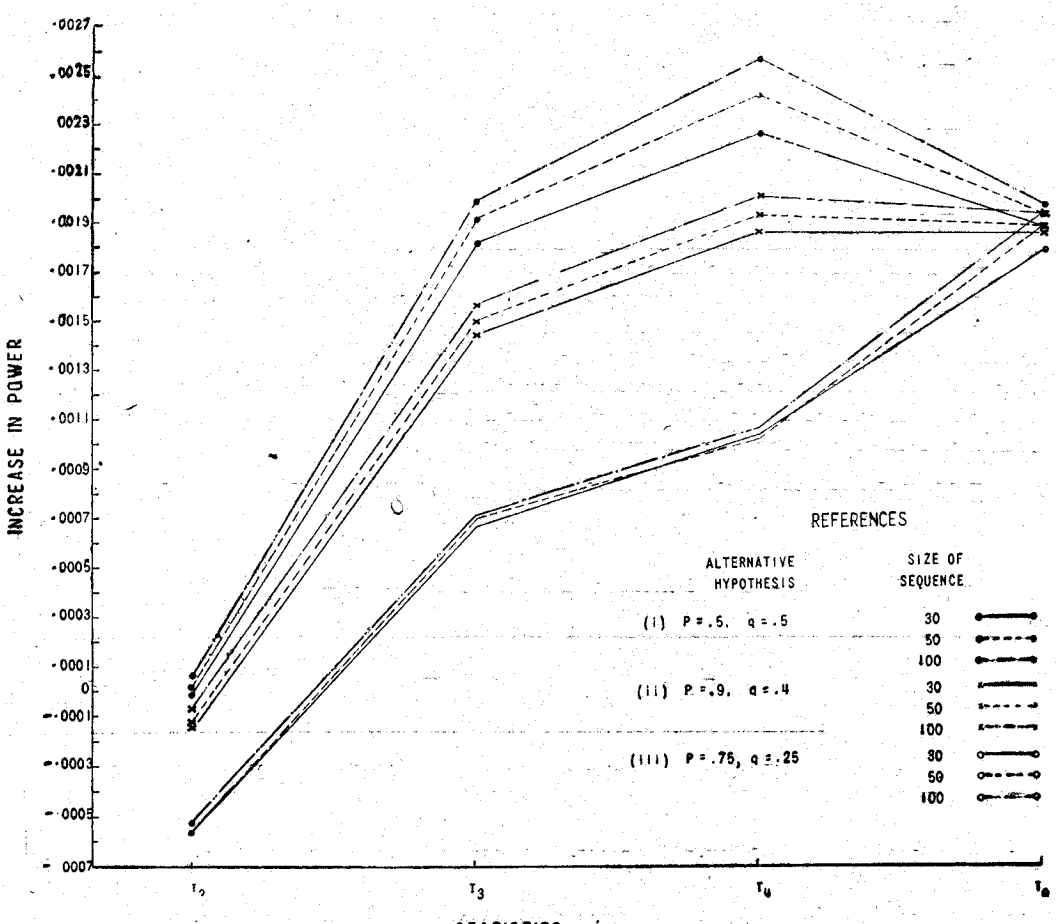

Fig. 3—Graph showing increase in power $(\xi + \eta^2/\sigma_0^2)$ for different statistics and $\delta = \cdot 001$

than as an ordered one arising from the pooled n_1 observations. Therefore the conclusions based on T_n is true to a large extent for U also when the forms of F is not definite.

Earlier investigations show that $T_{n/6}$ is more efficient than T_n for binomial sequences. It therefore, follows that these tests will be better than U *i.e.* Mann and Whitney's test. Investigations of Wetherill[11] show that two sample Wilcoxon's test is a little more robust than the t-test to differences in population variances. For location, when the populations are identical but non-normal, Wilcoxon's test is to be preferred. Comparing these conclusions with Table 3 it suggests that T_3 and T_4 and other extended forms of these tests will be better than the t-test when the variances are different and the populations are non-normal, for samples of nearly equal sizes.

Some simple results allied to T_2 have been discussed by Goodman[12] and Granger[13]. Goodman shows that *ab* transitions between successive observations will give a sufficient statistics on estimates of transition probabilities and therefore it may be argued that no other statistics will provide more information than the simple transitions between successive observations. This argument does not appear to be valid because in dealing with a binomial sequence there is only one parameter involved in H_0 while in the Markov sequence H_1, there are two parameters, one for the proportions of a and b and the other δ which defines the extent of randomness in the sequence. While in H_o we confine only to variation in p, in H_1 we vary both p and δ. Thus H_0 is a case of simple hypothesis whereas H_1 is a case of composite hypothesis. Therefore the conclusions arising from maximum likelihood and likelihood ratio tests do not hold in these cases.

In view of these considerations there appears to be no inconsistency in our findings that T_k when $k << n$, is better than T_2 and T_n. More rigorous studies on the relative powers of these tests including the case of two samples belong to populations F and G are in progress and the results will be published in due course.

ACKNOWLEDGEMENT

Our sincere thanks are due to Shri S. P. Varma for the complicated numerical calculations.

REFERENCES

1. Iyer, P. V. K. & Ray; D., *J. Indian Soc. Agric. Statist.*, **10** (1958), 23.
2. Iyer, P. V. K. & Singh, B.N., *J. Indian Soc. Agric. Statist.* **7** (1955)., 127.
3. Goodman, L.A., *Ann. Math. Statist.*, **7** (1958), 127.
4. Bhatt, B.R., *Ann. Math. Statist*, **32** (1961), 49.
5. Iyer, P.V.K., 'The use of the difference equation in solving distribution problem, *Internat. Statist. Inst.* **XXXIII** Part II (1951).
6. Iyer, P.V.K. & Kapur, M.N., "*Biometrika*" **41** (1954), 553.
7. Bartlett, M.S., "*An introduction to Stochastic process*" (Cambridge University Press), 1955.
8. Pitman, E.J.C., "*Lecture notes on non-parametric statistical inference*", Univ. of North Carolina, Institute of Statistics, 1948.
9. Mood, A.M., *Ann. Math. Statist.*, **25** (1954), 514.
10. Bhattacharyya M.N., 'Distribution free tests of hypothesis with reference to two sequences of alternatives' Approved D. Phil Thesis of the Calcutta University, 1962.
11. Wetherill, G.B., *J. Roy, Statist. Soc.*, series **B**. **22** (1960), 402.
12. Goodman, L.A., "*Biometrika*", **45** (1958), 181.
13. Granger, C.W.J., *J. Amer. Statist. Ass.*, **58** (1963), 728,

APPENDIX

In this section we give some auxiliary results which have been used for calculating the expected value and variance of T_n. It is hoped that the results will be useful in other similar investigations.

It can be seen that

$$E(T_n) = \sum_{k=2}^{n} E_k(a) \quad \text{for sequences ending in } b. \tag{1}$$

where

$$E_k(a) = \sum_{i=1}^{k-1} P(a_i\ b_k)$$

$$= \sum_{i=1}^{k-1} P(a_i)\ P(b_k/a_i)$$

$P(a_i\ b_k)$ being the probability for ith and kth observation to be a and b respectively which is equal to the probability for ith observation to be multiplied by the conditional probability of kth observation to be b when ith one is a. Obviously $E_k(a)$ represents the number of a's for a binomial chain of k observation where-in the last observation is b.

For obtaining the variance of T_n we have to evaluate the sum of the expectations for the different ways of obtaining two as's from the sequence[1]. Defining this by $E_n(2, ab)$ we can write

$$E_n(2,ab) = \sum_{3}^{n} E_k(aa) + \sum_{3}^{n} E_k(ab) + \sum_{4}^{n} E_k(aba) + \sum_{4}^{n} E_k(aab) \tag{2}$$

for sequences ending in b

where

$$E_k(aa) = \sum_{\substack{i=1 \\ i<j<k}}^{k-2} \sum_{j=2}^{k-1} P(a_i\ a_j\ b_k) = \sum_{i=1}^{k-2} \sum_{j=2}^{k-1} P(a_i)\ P(a_j/a_i)\ P(b_k/a_j)$$

$$E_k(ab) = \sum_{\substack{i=1 \\ i<j<k}}^{k-2} \sum_{j=2}^{k-1} P(a_i\ b_j\ b_k) = \sum_{\substack{i=1 \\ i<j<k}}^{k-2} \sum_{j=2}^{k-1} P(a_i)\ P(b_j/a_i)\ P(b_k/b_j)$$

$$E_k(aba) = \sum_{\substack{i=1 \\ i<j<l<k}}^{k-3} \sum_{j=2}^{k-2} \sum_{l=3}^{k-1} P(a_i\ b_j\ a_l\ b_k) = \sum_{\substack{i=1 \\ i<j<l<k}}^{k-3} \sum_{j=2}^{k-2} \sum_{l=3}^{k-1} P(a_i)\ P(b_j/a_i)\ (a_l/b_j)\ P(b_k/a_l)$$

and

$$E_k\,(aab) = \sum_{\substack{i=1\\ i<j<l<k}}^{k-3}\ \sum_{j=2}^{k-2}\ \sum_{l=3}^{k-1} P(a_i\ a_j\ b_l\ b_k) = \sum_{\substack{i=1\\ i<j<l<k}}^{k-3}\ \sum_{j=2}^{k-2}\ \sum_{l=3}^{k-1} P(a_i)\,P(a_j/a_i)\,P(b_l/a_j)\,P(b_k/b_l)$$

$P(a_i\ a_j\ b_l\ b_k)$ is the probability for ith, jth, lth, and kth observations to be in states a, a, b and b respectively. Other P's are defined in a similar manner.

We now give below the basic results such as $E_k\,(a)$ and $E_k\,(b)$ for sequences ending in b and $E_k\,(a)$ and $E_k\,(b)$ for sequences ending in a. These results will be useful for evaluating any type of configuration *e.g.*, $E_k\,(aa)$, $E_k\,(aba)$,.....etc.

Sequences ending in b

$$\acute{E}_k\,(a) = \frac{p_2q_1}{(1-\delta)^2}\,D_{k-1} + \frac{(Pq_1-Qp_2)}{(1-\delta)^2}\,q_1\left\{\frac{1-\delta^{k-1}}{1-\delta} - (k-1)\,\delta^{k-1}\right\} \tag{3}$$

Where P and Q are the initial probabilities for the Markov Chain to start with states a and b respectively.

$$\sim \frac{p_2q_1}{(1-\delta)^2}\,D_{k-1} \text{ as } (Pq_1 - Qp_2)\to 0 \text{ asymptotically} \tag{4}$$

and

$$\acute{E}_k\,(b) = \frac{q_1}{(1-\delta)^2}\left\{(k-1)q_1+p_2\delta\,\frac{1-\delta^{k-1}}{1-\delta}\right\} - \frac{(Pq-Qp_2)}{(1-\delta)^2}\left\{q_1\,\frac{1-\delta^{k-1}}{1-\delta}+p_2(k-1)\delta^{k-1}\right\} \tag{5}$$

$$\sim \frac{q_1}{(1-\delta)^2}\left\{(k-1)q_1 + p_2\delta\,\frac{1-\delta^{k-1}}{1-\sigma}\right\} \tag{6}$$

Sequences ending in a

$$\acute{E}_k\,(a) = \frac{p_2}{(1-\delta)^2}\left\{p_2(k-1)+q_1\delta\,\frac{1-\delta^{k-1}}{1-\delta}\right\} + \frac{(Pq_1-Qp_2)}{(1-\delta)^2}\left\{p_2\,\frac{1-\delta^{k-1}}{1-\delta}+q_1\,(k-1)\delta^{k-1}\right\} \tag{7}$$

$$\sim \frac{p_2}{(1-\delta)^2}\left\{p_2\,(k-1) + q_1\delta\,\frac{1-\delta^{k-1}}{1-\delta}\right\} \tag{8}$$

$$\acute{E}_k\,(b) = \frac{p_2q_1}{(1-\delta)^2}\,D_{k-1} - \frac{(Pq_1-Qp_2)}{(1-\delta)^2p_2}\,p_2\left\{\frac{1-\delta^{k-1}}{1-\delta} - (k-1)\delta^{k-1}\right\} \tag{9}$$

$$\approx \frac{p_2q_1}{(1-\delta)^2}\,D_{k-1} \tag{10}$$

where

$$D_{k-s} = \left\{(k-s) - \frac{\delta}{1-\delta}\,(1-\delta^{k-s})\right\}$$

$E(T_n)$ can be calculated from (1) by summing it up for all values of k from 2 to n. By using these we can get the values for $E_k\,(aa)$, $E_k\,(ab)$,..... For example, suppose we are interested in evaluating $E_k\,(aa)$. Since by definition

$$E_k\,(aa) = \sum_{i=1}^{k-2}\ \sum_{j=2}^{k-1} P\,(ai)\ \ P\,(a_j/a_i)\ \ P\,(b_k/a_j) \tag{11}$$

we can easily identify the expression $P(a_j/a_i)\,P(b_k/a_j)$ in the right hand side of (11) with $E_k(a)$ which is equal to $P(a_i)\,P(b_k/a_j)$, the only difference being that $P(a_i)$ in $E_k(a)$ has been replaced by the conditional probability $P(a_j/a_i)$ here. Substituting the value of $E_k(a)$ with this necessary alteration in (11), $E_k(aa)$ is obtained easily. By similar argument values for $E_k(aba)$, $E_k(aab)$.....will be calculated from $E_k(aa)$, $E_k(ab)$....etc. Thus by such successive operations we are able to get values for any type of configuration.

For the sake of illustration we also give here the values for those expressions which are necessary for the calculation of $V(T_n)$ namely $E_k(aa)$, $E_k(ab)$, $E_k(aba)$ and $E_k(aab)$.

$$E_k(aa) = \frac{p_2 q_1}{(1-\delta)^3}\left[p_2 \frac{(k-2)(k-3)}{2} - \frac{(p_2 - q_1)\delta}{1-\delta} D_{k-2} + q_1 \left\{ \frac{k-2}{1-\delta}\delta^k - \delta^2 \frac{1-\delta^{k-2}}{(1-\delta)^2} \right\} \right] \tag{12}$$

$$E_k(ab) = \frac{p_2 q_1}{(1-\delta)^3}\left[q_1 \frac{(k-2)(k-3)}{2} + \frac{(p_2 - q_1)\delta}{1-\delta} D_{k-2} + p_2 \left\{ \frac{k-2}{1-\delta}\delta^k - \delta^2 \frac{1-\delta^{k-2}}{(1-\delta)^2} \right\} \right] \tag{13}$$

$$E_k(aba) = \frac{p_2^2 q_1^2}{(1-\delta)^4}\left[\frac{(k-1)(k-2)(k-3)}{6} + \frac{6\delta^2}{(1-\delta)^2} D_{k-3} - \frac{3\delta}{1-\delta}\left\{ \frac{(k-2)(k-3)}{2} - \frac{(k-3)\delta^{k-1}}{1-\delta} + \delta^2 \frac{1-\delta^{k-3}}{(1-\delta)^2} \right\} + \frac{(k-2)(k-3)}{2(1-\delta)}\delta^k + \delta^k \frac{k-3}{(1-\delta)^2} - \delta^3 \frac{1-\delta^{k-3}}{(1-\delta)^3} \right] \tag{14}$$

$$E_k(aab) = \frac{p_2 q_1}{(1-\delta)^4}\left[(p_2^2 + q_1^2 - p_2 q_1)\left\{ \frac{(k-2)(k-3)}{2(1-\delta)} - \frac{2\delta^2}{(1-\delta)^2} D_{k-3} - \frac{k-3}{(1-\delta)^2}\delta^k + \delta^3 \frac{1-\delta^{k-3}}{(1-\delta)^3} \right\} + p_2 q_1 \left\{ \frac{(k-1)(k-2)(k-3)}{6} + \frac{(k-2)(k-3)}{2(1-\delta)}\delta^k + \frac{k-3}{(1-\delta)^2}\delta^k - \delta^3 \frac{1-\delta^{k-3}}{(1-\delta)^3} \right\} \right] \tag{15}$$

From the above expressions $E_n(2, ab)$ can be obtained by virtue of (2). Now since $E_n(2,ab)$ is equal to $\frac{1}{2}\times$second factorial moment[1] for ab* and also

$$\mu_2 = \mu_{[2]} + \mu'_1 - (\mu'_1)^2$$

we get

$$V(T_n) = 2E_n(2, ab) + E(T_n) - \{E(T_n)\}^2$$

* Iyer, P.V.K., *Ann. Math. Statist.* **29** (1958), 254.