SUB-SEQUENCE INCIDENCE ANALYSIS WITHIN SERIES OF BERNOULLI TRIALS: APPLICATION IN CHARACTERIZATION OF TIME SERIES DYNAMICS

Paper number 04/07

RICHARD H. G. JACKSON[*]

University of Exeter, UK

Abstract

This paper presents a widely applicable nonparametric approach to the characterisation of time series dynamics. The approach involves analysis of the incidence of occurrence of patterns in the direction of movement of the series, and may readily be applied to time series data measured on any scale. The paper includes derivations of analytic forms for two (infinite) families of distributions under the associated null hypothesis, and of a useful analytic form for the generation of the moments of these distributions. Areas of application in accounting and finance are suggested.

*Contact Author: Richard Jackson Senior Lecturer in Accounting and Finance, School of Business and Economics, Streatham Court, Rennes Drive, University of Exeter, EX4 4PU, UK; Tel: +44 (0)1392 263216; E-mail richard.jackson@exeter.ac.uk

ISSN 1473 2904

INTRODUCTION

The time series properties of a wide range of variables are of long-standing and continuing interest in the accounting and finance literature. For example, O'Hanlon (1995) reports that cited motivations for the study of earnings dynamics include, *inter alia*, the desire to understand the true earnings process in order to identify earnings smoothing practices; the desire to observe the impact of accounting policy changes on the earnings generating process; and the role of earnings forecasts in equity valuation. In this last respect, Peasnell (1982) and Ohlson (1991) made explicit the import of the time series properties of residual income, and paved the way for a wealth of theoretical development and empirical testing. In capital markets research, investigation of the dynamics of asset prices and / or returns has been central to investigation of market efficiency. After the early, empirical work of Kendall (1953), which suggested, in essence, a random walk generating process for asset prices, the formulation of questions concerning the efficiency of capital markets was refined, and a large body of theoretical and empirical literature, of increasing sophistication, has developed (Fama, 1970, 1991).

Often central within in the literature have been questions concerning the suitability of application of various time series modelling techniques, and a recurring issue has been whether or not time series may be best described by random walk, submartingale or martingale models. In the case of earnings dynamics, considerable effort has been expended upon identification, for example, of the form of autoregressive integrated moving average process best suited to modelling earnings series. The forecasting performance, however, of such models has generally been found not to dominate random walk models of behaviour (see, for example, Watts & Leftwich, 1977; Callen, Cheung, Kwan & Yip, 1993). With a regard to a range of financial ratios, Konings and Roodhooft (1997) considered dynamic evolution of the cross-sectional distribution using a non-parametric Markovian approach, questioning the pertinence of estimating partial adjustment models in earlier work and the associated maintained assumption of convergence. They demonstrated that there is no transition of financial ratios towards some industry average, and that the ratios show rich dynamics. In capital markets research, an established paradigm of widespread support for the efficient markets hypothesis, as acknowledged by Jensen (1978), has been subject to increasing challenge in recent years, with the development of so-called 'behavioural finance' (Schleifer, 2000).

As is well known, there are technical difficulties in distinguishing whether or not series are truly random, and a constant vigilance against apparently parsimonious but mis-specified models must be maintained. Therefore, in accounting and finance, as elsewhere in the social and natural sciences, there is strong interest in developing the battery of relevant tests.

This paper presents a new framework for testing hypotheses concerning the dynamics of time series by analysis of the incidence of patterns in the direction of movement of the series as against a null hypothesis of symmetrical random behaviour. It may be used as a primary methodology for the characterisation of time series dynamics, and also as a complement or response to findings from other tests which provide inferences as to the dynamics, stationarity and / or random walk nature of series. It may be applied widely - to time series which are measured on the nominal, ordinal, interval or ratio scales - and may be combined with a variety of specific statistical tests.

Thus, the framework provides a new approach to the investigation of hypothesis concerning the dynamics of interval or ratio scale time series, and also may be of particular interest to researchers who seek to analyse and draw inferences from ordinal time series data, such as business confidence survey results, brokers' buy and sell recommendations, *etc.* over time. A further application may be in the testing of pseudo random number generators, in addition or as a complement to the extant theoretical and empirical tests[1].

The paper proceeds as follows: section 2 sets out the framework; section 3 deals with generation of distributions under the null hypothesis; section 4 discusses calculation of the moments of probability distributions under the null; section 5 gives an illustrative application; and section 6 concludes. There are two appendices.

DATA CHARACTERIZATION, HYPOTHESES AND TESTING

Data characterization

Given a time series x_t, for t = 0 to n, measured on at least an interval scale and stripped of drift and time trend effects, the first difference time series may be generated as follows:

$$\Delta x_t = x_t - x_{t-1} \qquad\qquad t = 1 \text{ to } n \qquad (1)$$

This may then be transformed into the binary variable $\Delta' x_t$:

$$\Delta' x_t = \begin{Bmatrix} \uparrow & \text{if } \Delta x_t > 0 \\ \downarrow & \text{otherwise} \end{Bmatrix} \qquad\qquad t = 1 \text{ to } n \qquad (2)$$

Alternatively, x_t might be measured on an ordinal scale. In this case, let the scale's equivalence classes be defined by true attribute *A(x)* and denoted by labels *L(x)* – both of which may be ranked with a meaningful comparator relation ">".[2] Further, let the lowest and highest ranked equivalence classes (where either or both exist) be denoted L_{lowest} and $L_{highest}$ respectively. Then $\Delta' x_t$ may be generated as follows:

$$\Delta' x_t = \left\{ \begin{array}{ll} \uparrow & \text{if } L(x_t) > L(x_{t-1}) \\ \uparrow & \text{if } L(x_t) = L(x_{t-1}) = L_{highest} \\ \downarrow & \text{otherwise} \end{array} \right\} \quad t = 1 \text{ to } n \tag{3}$$

In the particular case of an ordinal scale upon which $A(x_t)$ represents a comparison between some matter at time *t* and that matter at time *t-1* (e.g. "more confident"), then $\Delta' x_t$ may be generated as follows given *m*, $L_{highest} > m \geq L_{lowest}$:

$$\Delta' x_t = \left\{ \begin{array}{ll} \uparrow & \text{if } L(x_t) > m \\ \downarrow & \text{otherwise} \end{array} \right\} \quad t = 0 \text{ to } n \tag{4}$$

Finally, x_t might be measured on a nominal scale. Let the scale's equivalence classes be defined by true attribute *A(x)* and denoted by labels *L(x)*, and let one of these equivalence classes be denoted *B*. Then $\Delta' x_t$ may be generated as follows:

$$\Delta' x_t = \left\{ \begin{array}{ll} \uparrow & \text{if } L(x_t) = B \\ \downarrow & \text{otherwise} \end{array} \right\} \quad t = 0 \text{ to } n \tag{5}$$

Null hypothesis

The null hypothesis is one of symmetrical random behaviour, i.e. H_0: $\Delta' x_t$ is a random binary sequence, with probability [↑] = probability [↓]. For x_t measured on at least the interval scale, the null hypothesis is equivalent to the hypothesis that x_t follows a pure random walk, i.e. $x_t = x_{t-1} + u_t$, the u_t being independent stochastic error terms with zero mean. The homoscedasticity (or otherwise) of the u_t has no impact on the analysis which follows.

Alternative hypotheses and hypothesis testing

By application of some alternative hypothesis (*H_1*) as to the dynamics of the series x_t, sub-sequences of $\Delta' x_t$ may be identified whose incidence of occurrence will be of particular interest in comparison to expectations under the null. For example, consider analysis of a time series of annual data, where the alternative hypothesis under investigation is that the series is cycling with period between four and six years (as against the null set out above). It is inferred that, *inter alia*, the number of incidences of periods of short term (say two to three year) sustained increase in x_t immediately followed by short term decrease, or vice versa, should be greater than that expected under H_0. Therefore, incidences of occurrence of the following sequences of $\Delta' x_t$ are of special interest: ↑↑↓↓, ↓↓↑↑, ↑↑↓↓↓, ↓↓↑↑↑, ↑↑↑↓↓, ↓↓↓↑↑, ↑↑↑↓↓↓, ↓↓↓↑↑↑. (Note that the terminology 'sub-sequence' is dropped at this point in favour of the less cumbersome 'sequence').

Having decided upon those sequences whose incidence of occurrence is of interest, the number of occurrences of any such sequence, *S*, may then be counted to yield the count I_S. This may then be compared to the distribution of the number of occurrences of that sequence generated under the null hypothesis and statistical inferences drawn. A variety of specific tests might be employed, including Kolmogorov-Smirnov or other 'goodness of fit' tests. Further, writing the expected number of occurrences of the sequence of interest under the null as μ_S and its standard deviation as σ_S, and given *M* time series in the data set which are subject to the same hypotheses, then application of the central limit theorem yields, for sufficiently large *M*, the standard normal z-statistic:

$$Z_S = \frac{\sum_{j=1}^{M} I_S - M\mu_S}{\sigma_S \sqrt{M}} \tag{6}$$

subject to mutual independence and common distribution of the random variables, and existence of the mean and variance for each (Feller, 1968; Lindeberg, 1922).[3]

DISTRIBUTIONS OF OCCURRENCE OF SUB-SEQUENCES UNDER THE NULL

Definition: Let B_n denote a series of outcomes of *n* independent Bernoulli trials, with $n \in N^+$, Prob ["success"] ≡ Prob [↑] = Prob ["failure"] ≡ Prob [↓] = 0.5.

The series $\Delta' x_t$ under the null hypothesis may then be represented as one of the 2^n possible series B_n. The task in hand, therefore, is to calculate the distribution of the number of occurrences of a sequence of interest over all 2^n possible series B_n. This may be approached by computational exhaustion, but, approached in this way, the task grows exponentially as n increases. Therefore, an analytic expression for the distribution is desirable.

Definitions: Let S_l be a sequence of outcomes of l Bernoulli trials, $l \in N^+$. Let $S_l(a,b)$, with $a \in N^+$, $b \in N^+$ and $1 \le a \le b \le l$, be the sub-sequence from the a^{th} to the b^{th} terms (inclusive) of S. Let the *overlap order* of S be denoted $p(S_l)$ and defined as follows: $p(S_l)$ = max(i) such that $S_l(1,i) \equiv S_l(l-i+1,l)$, $i \in N$. Let O_p denote the set of sequences of outcomes of Bernoulli trials with overlap order p. It is noted that $0 \le p \le l$. It is further noted that $S_l(1,0)$ and $S_l(l+1,l)$ are not defined, so in the case of no overlap $p = 0$ is correct.

The concept of overlap order is demonstrated in the following examples, with parentheses used to highlight the maximum potential overlap as each of the example sequences is repeated:

↑↑↑↓↓↓ is in O_0 (↑ ↑ ↑ ↓ ↓ ↓)(↑ ↑ ↑ ↓ ↓ ↓)(↑ ↑ ↑ ↓ ↓ ↓) (• • •

↑↓↓↓↑ is in O_1 (↑ ↓ ↓ ↓{↑)↓ ↓ ↓(↑}↓ ↓ ↓{↑)↓ • • •

↑↑↓↓↓↑↑ is in O_2 (↑ ↑ ↓ ↓ ↓{↑ ↑)↓ ↓ ↓(↑ ↑}↓ ↓ ↓{↑ ↑)↓ • • •

Definition: Let $X(n,i,l,p)$ be the number of series B_n in which a sequence S_l from set O_p occurs i times, $i \in N$.

The following are evident:

if $n = l$ then $X(n,1,l,p) = 1$ (7)

if $n = 2l - p$ then $X(n,2,l,p) = 1$ (8)

and, generally, for $i > 0$:

if $n = il - (i-1)p = i(l-p) + p$ then $X(n,i,l,p) = 1$ (9)

$$\text{if} \qquad n < i(l-p)+p \qquad \text{then} \qquad X(n,i,l,p)=0 \tag{10}$$

The distribution of $X(n,i,l,p)$ represents the distribution under the null hypothesis which is sought in respect of a sequence of interest of length l from set O_p. An analytic expression for this distribution is derived in Appendix 1 for the cases $p = 0$ and $p = 1$. This expression is in the form of a backwards recursive formula involving an intermediate variable, $Y(n,i,l,p)$, which is also derived in Appendix 1. Appendix 2 gives a numerical illustration of the reasoning in these derivations.

The analytic expressions are as follows:

$$X(n,i,l,p) = \begin{cases} Y(n,i,l,p) - \sum_{j>i} {}^{j}C_{i} X(n,j,l,p) & \text{for } n \geq i(l-p)+p \\ 0 & \text{for } n < i(l-p)+p \end{cases} \tag{11}$$

where, writing $n - [i(l-p)+p] = k$:

case p = 0

$$Y(n,i,l,0) = \begin{cases} {}^{i+k}C_{k} \cdot 2^{k} & \text{for } n \geq il \\ 0 & \text{for } n < il \end{cases} \tag{12}$$

case p = 1

$$Y(n,i,l,1) = \begin{cases} 2^{n} & \text{for } i = 0 \\ (-1)^{k} \sum_{j=\max(0,k-(i-1))}^{k} {}^{i-1}C_{k-j} \cdot {}^{i+j}C_{j} \cdot 2^{j} \cdot (-1)^{j} & \text{for } i > 0 \text{ and } n \geq i(l-1)+1 \\ 0 & \text{for } n < i(l-1)+1 \end{cases} \tag{13}$$

These expressions have been verified computationally for n up to 31 for various l.

PROBABILTY DISTRIBUTION MOMENTS UNDER THE NULL HYPOTHESIS

For given *n*, *l*, and *p*, we write:

$$Y(n,i,l,p) = Y_i \tag{14}$$

$$X(n,i,l,p) = X_i \tag{15}$$

The probability distribution, X'_i, for the number of series B_n in which sequence of interest of length *l* from set O_p occurs *i* times is given by:

$$X'_i = \frac{X_i}{2^n} \tag{16}$$

We also calculate Y'_i as follows: $Y'_i = \frac{Y_i}{2^n}$ (17)

From expression (11), we deduce that:

$$Y'_i = \sum_{j \geq i} {}^jC_i X'_i \tag{18}$$

Expression (18) encapsulates the useful result that Y'_i is a factorial moment generating function[4] for X'_i. Using $E(\cdot)$ to denote expectation, and with readily calculable $\alpha_1, \alpha_2, \cdots, \alpha_{i-1} \in Z$:

$$Y'_i = \frac{E\left((X')^i\right) + \alpha_{i-1} E\left((X')^{i-1}\right) + \cdots + \alpha_1 E(X')}{i!} \tag{19}$$

Expression (19) allows the central moments of the distribution of X' to be deduced readily. In particular, the fist and second order central moments are given by:

$$E(X') = Y'_1 \tag{20}$$

$$SD(X') = \sqrt{2 \cdot Y'_2 + Y'_1 - (Y'_1)^2} \tag{21}$$

ILLUSTRATIVE APPLICATION

Consider a time series x_t of annual data measured on an interval scale over, say, 32 years. The binary time series, $\Delta' x_t$, generated by application of expressions (1) and (2) then has 31 terms. It is decided to analyse, *inter alia*, incidences of occurrence of ↓↓↑↑ as a sub-sequence of $\Delta' x_t$, with the specific alternative hypothesis that the number of occurrences of this sequences will be greater than that expected under the null. In this case, n = 31, l = 4 and p = 0. Table 1 shows the pertinent distributions of *Y*, *Y'*, *X* and *X'* as calculated from expressions (12), (17), (11) and (16) respectively, and includes the mean and standard deviation of *X*' as calculated from expressions (20) and (21) respectively. It also includes the cumulative probability distribution of *X'*.

***** Table 1 about here *****

If the observed number of occurrences of ↓↓↑↑ in the series is four, say, then we may deduce from the cumulative probability distribution that the that the null hypothesis may be rejected in favour of the alternative with 95.7% confidence; if the observed number is five, the confidence level is 99.6%; etc.

If we have, say, a sample of say, 53 such series (all subject to the same hypotheses, and subject to usual caveats concerning mutual independence, *etc.*), and the number of observed occurrences of ↓↓↑↑ across the whole sample is 112, then expression (6) yields the standard normal z-statistic 2.60; and the null hypothesis may be rejected in favour of the alternative with confidence of over 99.5% (one tail test).

CONCLUSIONS AND FURTHER WORK

Analytic expressions for distributions of incidence of occurrence of sequences with overlap order equal to 0 or 1 within series of Bernoulli trials, and for the moments of those distributions, have been produced under the null hypothesis that the series of Bernoulli trials are symmetrically random. These expressions may readily be used for speedy calculation of statistics which allow the testing of a range of hypotheses concerning dynamics of time series measured on any scale.

The restriction of the analytic expressions to the cases of overlap order $p = 0$ and $p = 1$ is not onerous. For example, if the incidence of monotone increase (or decrease) is of interest, sequences of interest for test purposes might be chosen as ↓↑↑, ↓↑↑↑, ↓↑↑↑↑, *etc.*, which are all in set O_0. Similarly for incidence of monotone decrease. If the incidence of monotone increase followed by monotone decrease (or vice-versa) is of interest, sequences of type ↓↓↑↑, ↑↑↑↓↓↓, ↓↓↓↓↓↓↓↑↑↑↑, *etc.* are also all in the set O_0. If the incidence of monotone increase or decrease of some exact duration in time periods is of interest, sequences of the type ↓↑↑↑↓, ↓↑↑↑↑↓, *etc.* are all in the set O_1. Nevertheless, theoretical work to further generalise the analytic expressions is desirable.

APPENDIX 1: DERIVATION OF ANALYTIC EXPRESSIONS

Introduction and overview

Definitions of B_n, S_l, $S_l(a,b)$, *overlap order* $p(S_l)$, O_p and $X(n,i,l,p)$ (including the permissible arguments for these) are as per the third section of the paper.

The derivations are based upon combinatorial mathematics. The basic approach is to consider a base series of Bernoulli trials which contains a number of occurrences of a sequence of interest; to count the ways in which this may be augmented by the addition of terms whilst preserving the occurrences of the sequence of interest; and to thereby generate general analytic expressions for $X(n,i,l,p)$ for the cases p = 0 and p = 1. Therefore, the derivations start with some definitions designed to unambiguously define a framework in which we may discuss the building of series of Bernoulli trials and the 'legality' of those builds.

Definitions

Given a series B_n containing i occurrences of a sequence S_l:

> Let 'allowable building positions' (ABPs) be defined as follows, in order to define exact positions at which terms may be added / inserted to augment the series B_n. The idea is to allow addition of terms before, between or after occurrences of S_l. (Note that arbitrary choices between possible candidate positions have been made):
>
> > If i > 1, let 'allowable building positions' (ABPs) denote: (i) the positions at the end of the series B_n - giving two ABPs, to be termed 'exterior ABPs'; and (ii) the positions immediately to the right of the first (i-1) occurrences of S_l - giving a further (i-1) ABPs, $i \geq 1$, to be termed 'interior ABPs'.
> >
> > If i = 1, let 'allowable building positions' (ABPs) denote the positions at the end of the series B_n - giving two ABPs, to be termed 'exterior ABPs'. Note that in this case there are no interior ABPs.

If $i = 0$, let ‘allowable building position’ (ABP) denote the position at the end of the series B_n - giving a single ABP only.

Let ‘build’ denote the generation of the series B_{n+j} from the series B_n by the addition of j terms one by one to ABPs, $j \in N^+$.

Let ‘legal build’ denote a build from B_n to B_{n+j} which, with each term added, maintains the original i occurrences of the sequence S_l.

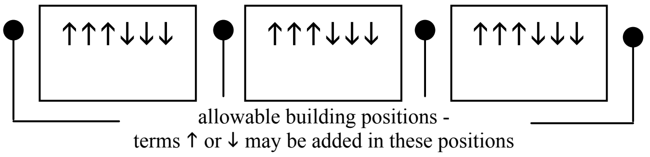
Let ‘illegal build’ denote any build which is not a legal build.

Let ‘illegal addition’ denote the addition of a term in an allowable building position but which results in an illegal build.

Derivation: case p = 0

In this case, statement (9) becomes: if $n = il$ then $X(n,i,l,0) = 1$. We now consider the specific case in which l divides n where $il = N_l$, say. There are then i contiguous and non-overlapping occurrences of the sequence of interest of length l. There are i+1 allowable building positions where terms of either type (i.e. ↑ or ↓) may be added to generate legal builds as n is increased beyond N_l. See, for example, Figure 1.

Figure 1: example: $i = 3$, $l = 6$, $p = 0$,
Situation when $n = 18$, $X(18,3,6,0) = 1$


↑↑↑↓↓↓
↑↑↑↓↓↓
↑↑↑↓↓↓
allowable building positions -
terms ↑ or ↓ may be added in these positions

Therefore, $Y(n,i,l,0)$ defined as follows represents the number of series B_n in which the sequence of interest occurs at least i times[5]:

$$Y(n,i,l,0) = \begin{cases} \left({}^{(i+n-il)}C_{(n-il)}\right)2^{(n-il)} & \text{for } n \geq il \\ 0 & \text{for } n < il \end{cases} \tag{22}$$

Writing $n - [i(l-p)+p] = k$, and given that we are dealing with case $p = 0$, expression (22) can be seen to be equivalent to expression (12) (*QED*). For ease of reading, notice that $k = (n-il)$ is the number of Bernoulli trials in the series in excess of the number N_1 at which $X(n,i,l,0)$ equalled 1, i.e. k represents the number of terms added to the series B_{N_1} in which i occurrences of the sequence of interest was first achieved.

Now, the $Y(n,i,l,0)$ as defined take no account of the fact that as n increases beyond N_1 it will reach $(i+1)l$, $(i+2)l$, and so on; therefore, it ignores the possible advent of occurrence of $(i+1)$, $(i+2)$, etc. incidences of the sequence of interest. In order to derive $X(n,i,l,0)$, the $Y(n,i,l,0)$ must be reduced to remove the number of series which need be counted in $X(n,j,l,0)$ rather than in $X(n,i,l,0)$, $j \in N^+$, $j > i$. Consider the case l divides n where, say, $jl = N_2$ and $X(n,j,l,0) = 1$. There are then j contiguous non-overlapping occurrences of the sub-sequence interest of length l, and the point of interest here is to deduce the count within $Y(N_2,i,l,0)$ which is (properly) accounted for by $X(N_2,j,l,0)$. Imagine the series containing j contiguous occurrences of the sequence of interest as being built (by the addition of terms to the series) from one which contained exactly i occurrences: these original i occurrences of the sequence of interest may be seen to coincide with any of the j occurrences of the sequence of interest in the series which is built, i.e. $X(N_2,j,l,0)$ properly accounts for ${}^{j}C_i$ of the count within $Y(N_2,i,l,0)$. Therefore, the $X(n,i,l,0)$ may be calculated by adjustment of the $Y(n,i,l,0)$ by application of the following backwards recursive formula:

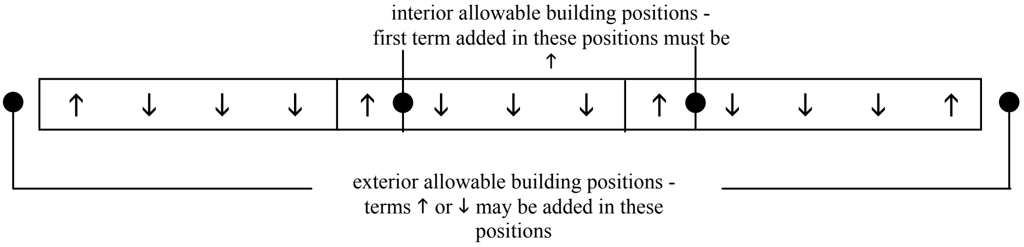
$$X(n,i,l,0) = \begin{cases} Y(n,i,l,0) - \sum_{j>i} {}^{j}C_i X(n,j,l,0) & \text{for } n \geq il \\ 0 & \text{for } n < il \end{cases} \tag{23}$$

It is noted that when $i = 0$ this formula may be re-arranged to give, as expected, $\sum_{j\geq 0} X(n,j,l,0) = 2^n$. Expression (23) is equivalent to expression (11) for $p = 0$ *(QED)*.

Derivation: case p = 1

In this case, statement (9) becomes: if $n = i(l-1)+1$ then $X(n,i,l,p)=1$. We now consider the specific case in which $i(l-1)+1 = N_3$, say. There are then i occurrences of the sequence of interest of length l, each of which overlaps its right and left hand immediate neighbours (where such exist) by one term. There are i+1 allowable building positions where terms of either type (i.e. ↑ or ↓) may be added to generate builds as n is increased beyond N_3. In order, however, that such builds are legal builds, the first term added to each of the interior allowable building positions must be of the same type as that of the overlap term in the sequence of interest. There is no such restriction on any terms added to the exterior allowable building positions, or on the second subsequent terms added to interior allowable building positions. See, for example, Figure 2.

Figure 2:
Example: sequence of interest ↑↓↓↓↑, i = 3, l = 5, p = 1
Situation when n = 13, X(13,3,5,1) = 1



interior allowable building positions - first term added in these positions must be ↑
↑ ↓ ↓ ↓ ↑ ↓ ↓ ↓ ↑ ↓ ↓ ↓ ↑
exterior allowable building positions - terms ↑ or ↓ may be added in these positions

Therefore, to calculate the number of distinct series B_n in which the sequence of interest occurs at least i times, being $Y(n,i,l,1)$, we adopt the following approach. This approach is illustrated numerically in Appendix 2.

Definition: Let L_j to be the number of possible distinct series B_n for $n > N_3$ which can be built from B_{N_3} by addition of terms one by one to allowable building positions, the first j of which additions are illegal additions to distinct interior allowable building positions, $j \in N$, $1 \le j \le i-1$.

We may then deduce the number of distinct B_n for $n > N_3$ built from B_{N_3} via illegal builds to be:

$$L_1 - (L_2 - (L_3 - \cdots - (L_{i-2} - L_{i-1})\cdots)) = (-1)\sum_{j=1}^{i-1}(-1)^j L_j \tag{24}$$

Writing $k = n - [i(l-1)+1]$, we calculate L_j as:

$$L_j = \left({}^{i-1}C_j \cdot 1^j\right)\cdot\left({}^{i+k-j}C_{k-j} \cdot 2^{k-j}\right) \tag{25}$$

The total number of distinct B_n for $n > N_3$ built from B_{N_3} via all builds (legal and illegal) is given by $Y(n,i,l,0)$, calculated using expression (22) adapted to the following and again writing $k = n - [i(l-1)+1]$:

$$Y(n,i,l,0) = \begin{cases} {}^{i+k}C_k\, 2^k & \text{for } n \geq i(l-1)+1 \\ 0 & \text{for } n < i(l-1)+1 \end{cases} \tag{26}$$

Therefore, the number of distinct B_n built from B_{N_3} via legal builds, $Y(n,i,l,1)$, is given by the following expression, deduced by combination of expressions (24), (25) and (26):

$$Y(n,i,l,1) = \begin{cases} 2^n & \text{for } i = 0 \\ (-1)^k \displaystyle\sum_{j=\max(0,k-(i-1))}^{k} {}^{i-1}C_{k-j} \cdot {}^{i+j}C_j \cdot 2^j \cdot (-1)^j & \text{for } i > 0 \text{ and } n \geq i(l-1)+1 \\ 0 & \text{for } n < i(l-1)+1 \end{cases} \tag{27}$$

This expression is the same as expression (13) *(QED)*.

The logic of calculation of $X(n,i,l,1)$ from $Y(n,i,l,1)$, and generally of $X(n,i,l,p)$ from $Y(n,i,l,p)$, follows similarly to that used in the case $p = 0$. Therefore, we have:

$$X(n,i,l,1) = \begin{Bmatrix} Y(n,i,l,1) - \sum_{j>i} {}^{j}C_{i} X(n,j,l,1) & \text{for } n \geq i(l-1)+1 \\ & \\ 0 & \text{for } n < i(l-1)+1 \end{Bmatrix} \quad (28)$$

which is equivalent to expression (11) for $p = 1$ *(QED)*.

More generally, given an analytic expression for $Y(n,i,l,p)$:

$$X(n,i,l,p) = \begin{Bmatrix} Y(n,i,l,p) - \sum_{j>i} {}^{j}C_{i} X(n,j,l,p) & \text{for } n \geq i(l-p)+p \\ & \\ 0 & \text{for } n < i(l-p)+p \end{Bmatrix} \quad (29)$$

APPENDIX 2: NUMERICAL ILLUSTRATION OF REASONING IN THE DERIVATION FOR CASE $p = 1$

Consider the sequence of interest ↑↓↓↓↓↑, for which $l = 6$ and $p = 1$, and suppose that $X(26,4,6,1)$ is sought, *i.e.* the number of series of 26 Bernoulli trials in which the sequence of interest occurs four (and only four) times.

When n equals $i(l-1)+1 = 21$, there is just one series of 21 Bernoulli trials in which the sequence of interest is repeated four times, i.e. $Y(21,4,6,1) = 1$. We now seek to add 26 – 21 = 5 = k terms to that series, maintaining at each addition the four 'original' occurrences of the sequence of interest. There are two exterior allowable building positions where either ↑ or ↓ may be added; and there are $i - 1 = 3$ interior allowable building positions where, in each case, the first term added must be ↑.

We are concerned, therefore, with: (a) counting the number of ways in which B_{26} may be built from the B_{21}; and (b) deducting the number of such builds which are illegal.

Calculation (a) Is given by expression (26), yielding: ${}^{9}C_{5} \cdot 2^{5}$ = 4,032

Calculation (b) Requires calculation of L_1, L_2 and L_3 as given by expression (25)

> L_1 = number of series of 26 Bernoulli trials built from the original series of 21 Bernoulli trials by first making the illegal addition of a single ↑ to one of the three interior allowable building positions, then addition of four more terms of either type amongst the five allowable building positions = $\left({}^{3}C_{1} \cdot 1^{1}\right)\left({}^{8}C_{4} \cdot 2^{4}\right)$ = 3,360
>
> L_2 = number of series of 26 Bernoulli trials built from the original series of 21 Bernoulli trials by first making the illegal addition of a single ↑ to two of the three interior allowable building positions, then addition of three more terms of either type amongst the five allowable building positions = $\left({}^{3}C_{2} \cdot 1^{2}\right)\left({}^{7}C_{3} \cdot 2^{3}\right)$ = 840
>
> L_3 = number of series of 26 Bernoulli trials built from the original series of 21 Bernoulli trials by first making the illegal addition of a single ↑ to each of the three interior building

positions, then addition of two more terms of either type amongst the five allowable building positions = $\left({}^{3}C_{3}\cdot 1^{3}\right)\cdot\left({}^{6}C_{2}\cdot 2^{2}\right)$= 60

In illustration of expression (24), note that L_3 is counted in L_2, so $L_2 - L_3$ = 780 series of 26 Bernoulli trials are built from the original series of 21 Bernoulli trials by first making exactly two non-allowable additions into two separate interior allowable building positions, and then proceeding with “legal” additions. But these $L_2 - L_3$ are counted in L_1, so $L_1 - (L_2 - L_3)$ = 2,580 series of 26 Bernoulli trials are built from the original series of 21 Bernoulli trials by first making exactly one non-allowable addition into an interior allowable position, and then proceeding with “legal” additions. It is this number which must be eradicated from the count made under Calculation (a). This is equivalent to imposing the condition that in building the series of 26 Bernoulli trials by the addition of terms, we must start and continue using only legal additions.

Therefore, $Y(26,4,6,1)$ = 4,032 – 2,580 = 1,452.

This calculation is encapsulated and generalised in expression (27).

Since $Y(26,5,6,1)$ = 1, because 26 = $5(l-1)+1$, and $Y(26,i,6,1)$ = 0 for all i > 5, $X(26,4,6,1)$ may be calculated from expression (28) as: $1{,}452-{}^{5}C_{4}\cdot 1$ = 1,447.

NOTES

1. See, for example, Knuth (1998) section 3.

2. Nomenclature regarding attributes, labels and ranking of ordinal scale classes follows Siegel and Castellan (1988) section 3.3.

3. Existence of mean and variance being satisfied (see sections on distributions and their moments), conduct of a z-test is against the null hypothesis as expanded to include the mutual independence of the time series under investigation.

4. Kendall *et al* (1987) sections 3.7-3.11 give a general treatment of factorial moments and associated generating functions.

5. Noting that the number of possible distinguishable arrangements of *a* indistinguishable objects into *b* distinguishable compartments is ${}^{a+b-1}C_a$, where ${}^{x}C_y$ represents combination and equals $\frac{x!}{y!(x-y)}$ with $y \in N$, $(x+y) \in N^+$ (see, for example, Gray, 1967 pp. 97-98).

REFERENCES

Callen, J., Cheung, C., Kwan, C. and Yip, R. (1993). 'An Empirical Investigation of the Random Character of Annual Earnings, *Journal of Accounting, Auditing and Finance*, pp. 151-162

Fama, E. F. (1970). 'Efficient Capital Markets: A Review of Theory and Empirical Work', *The Journal of Finance*, Vol. 25, Issue 2, May, pp. 383-417.

Fama, E. F. (1991). 'Efficient Capital Markets: II', The Journal of Finance, Vol. 46, Issue 5, December, pp. 1575-1617.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications, Volume 1*, 3rd Edition, John Wiley & Sons Inc.

Gray, J.R. (1967). *Probability*, Oliver and Boyd Limited.

Jensen, M. C. (1978). 'Some Anomalous Evidence Regarding Market Efficiency', *Journal of Financial Economics*, 6, pp. 95-101.

Kendall, M. G. (1953). 'The Analysis of Economic Time-Series – Part I: Prices', *Journal of the Royal Statistical Society, Series A (General)*, Vol. 116, Issue 1, pp. 11-25.

Kendall, M., Stuart, A., Ord, J. K. (1987). *Kendall's Advanced Theory of Statistics, Volume 1: Distribution Theory*, 5th Edition, Charles Griffin & Co.

Knuth, D. E. (1998). *The Art of Computer Programming, Volume 2: Seminumerical Algorithms*, 3rd Edition., Addison-Wesley.

Konings, J and Roodhooft, F. (1997), 'Financial Ratio Cross-Section Dynamics: A Non-Parametric Approach', *Journal of Business Finance & Accounting*, 24 (9) & (10), October & December, pp. 1331-1342.

Lindeberg, J. W. (1922). 'Eine neue Herleitung des Exponentialgesetzes in der Wahrsheinlichkeitsrechnung', *Mathematische Zeitschrift*, Vol. 15, pp. 211-225, as cited by Feller (1968).

O'Hanlon, J. (1995). 'The Univariate Time Series Modelling of Earnings: A Review', *British Accounting Review*, 27, pp. 187-210.

Ohlson, J. (1991). 'Earnings, Book Value and Dividends in Security Valuation', Working Paper, University of British Columbia and Columbia University.

Peasnell, K. V. (1982). 'Some Formal Connections Between Economic Values and Yields and Accounting Numbers', *Journal of Business Finance & Accounting*, October, pp. 361-381.

Schleifer, A. (2000). *Inefficient Markets*, Oxford.

Siegel, S., Castellan, N. J. (1988). *Nonparametric Statistics for the Behavioural Sciences*, 2nd Edition, McGraw-Hill.

Watts, R. L. and Leftwich, R. W. (1977). 'The Time Series of Annual Accounting Earnings', *Journal of Accounting Research*, Vol. 15, No. 2, Autumn, pp. 253-271.

TABLE 1

Distributions in respect of the incidence of occurrence of sequence ↓↓↑↑ within series of 31 independent Bernoulli trials

	Number of occurrences of sequence of interest (*i*)									Mean	SD
	0	1	2	3	4	5	6	7	8 or more		
Y(31,i,4,0)	2,147,483,648	3,758,096,384	2,516,582,400	807,403,520	127,008,768	8,945,664	219,648	960	0	n/a	n/a
Y'(31,i,4,0)	1.0000	1.7500	1.1719	0.3760	0.0591	0.0042	0.0001	0.0000	0.0000	n/a	n/a
X(31,i,4,0)	216,847,936	682,524,224	770,242,368	384,465,728	85,541,568	7,647,936	212,928	960	0	n/a	n/a
X'(31,i,4,0) = probability	0.1010	0.3178	0.3587	0.1790	0.0398	0.0036	0.0001	0.0000	0.0000	1.7500	1.0155
cum *X'(31,i,4,0)* = cum probability	0.1010	0.4188	0.7775	0.9565	0.9963	0.9999	1.0000	1.0000	n/a	n/a	n/a