Sub-sequence incidence analysis within series of Bernoulli trials: application in characterization of time series dynamics

Richard H. G. Jackson*

School of Management, University of Bath, Claverton Down, Bath, BA2 7AY, UK

This paper presents a new and widely applicable nonparametric approach to the characterization 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 null hypothesis of random behaviour, and of a useful analytic form for the generation of the moments of these distributions. The distributions are asymptotically normal, so allowing for straightforward application of the approach presented in the paper to long series of high frequency and/or extended time period data. Areas of application in finance and accounting are suggested.

Keywords: Bernoulli trials; time series dynamics; nonparametric test

JEL Classification: C14; C22; C55

1. Introduction

As the twenty-first century unfolds there is a renewed enthusiasm for investigation and interpretation of series dynamics – in finance and accounting, certainly, and much more widely across the social and biological sciences. This is in some large part driven by the availability of the ever larger datasets, often with previously-unseen levels of granularity and richness, which can now be collected, stored, shared and processed in the era of 'big data'; and the desire best to exploit those datasets from a variety of academic, social, political and commercial perspectives.[1]

As is well known, there are technical difficulties in distinguishing whether or not series are truly random; and whether or not patterns which we might (wish to) perceive by inspection of data are actually present with any statistical significance. A constant vigilance against apparently parsimonious but misspecified models must be maintained. In finance and accounting, as elsewhere in the social and natural sciences, there is strong interest in developing the battery of relevant tests.

* Email: R.H.G.Jackson@bath.ac.uk

This paper presents a new framework for testing hypotheses concerning the dynamics of 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. Relevant distributions under the null hypothesis are derived step-by-step using combinatorial mathematics. The mathematical approach employed in this paper is not the only way to proceed: alternative combinatorial approaches are available; and also applicable would be the flexible technique of finite Markov chain imbedding (FMCI) (Fu and Koutras 1994; Fu and Lou 2003).[2] As detailed in Section 3, the principal mathematical burden of this paper is to deduce the distribution of the number of occurrences of a given sequence in a series of independent Bernoulli trials. FMCI allows numerical evaluation of such distributions for arbitrary sequences – without the restrictions on symmetry of Bernoulli trials and 'overlap order' as stipulated in Section 3 of this paper.[3] FMCI does not, however, lead to analytic expressions for the distributions or their moments; whereas such expressions are provided via the combinatorial approach of this paper and by alternative combinatorial approaches.

The framework developed in this paper 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. Further, the approach 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 paper 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 or brokers' buy/sell recommendations, 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.[4]

The time series properties of a wide range of variables are of long-standing and continuing interest in the finance and accounting 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, Ohlson (1995) makes 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 suggests, 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 (e.g. Watts and Leftwich 1977; Callen, Cheung, Kwan, and Yip 1993). Konings and Roodhooft (1997) question the pertinence of the partial adjustment models in earlier work; and demonstrate that a range of financial ratios show rich dynamics. Tippett and Warnock (1997), implementing a continuous time formulation of the Garman-Ohlson framework, find that the theory allows for the possibility of complex, firm-specific dynamics in the evolution of, inter alia, earnings – including harmonic behaviour. As regards valuation in particular, Nichols et al. (2017) considers the link between accounting fundamentals and market prices, and identifies incidences of mispricing in both the cross section of firms and the time series. In capital markets research, an established paradigm of widespread support for the efficient markets hypothesis has been subject to increasing challenge in recent years, with the development of so-called 'behavioural finance' (Schleifer, 2000), and a re-found interest in technical analysis (Lo and Hasanhodzic, 2010). At the same time, significant attention is attracted by the availability of high-frequency data and trading opportunities, and the application of big data analytics (e.g. Seddon and Currie, 2017).

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 those distributions; and Section 5 discusses their asymptotic properties. Section 6 then gives simple illustrative applications; and Section 7 concludes. There are three appendices.

2. Data characterization, hypotheses and testing

2.1 *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:

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

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

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

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 ">".[5] 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:

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

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*, $L_{highest} > m \geq L_{lowest}$.

$$\Delta' x_t = \begin{cases} \uparrow & \text{if } L(x_t) > m \\ \downarrow & \text{otherwise} \end{cases} \qquad t = 0 \text{ to } n \tag{4}$$

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:

$$\Delta' x_t = \begin{cases} \uparrow & \text{if } L(x_t) = B \\ \downarrow & \text{otherwise} \end{cases} \qquad t = 0 \text{ to } n \tag{5}$$

As will be clear from the foregoing, the framework presented in this paper is for application to data summarised into binary form. One advantage is that the framework is widely applicable, and may be used upon data measured on any scale – including qualitative or quantitative data measured on ordinal or nominal scales (e.g., survey data captured on Likert-type scales). In application of the framework to ratio or interval scale data, however, concern may pertain that the transformation of the data to binary form results in loss of (or disregard for) some of the information in the data – by ignoring the size of series increments, and preserving only a series

record of increase versus decrease. This cannot be gainsaid: loss of information is usual when non-parametric tests are applied to ratio or interval scale data.[6] Nevertheless, there are a number of areas in finance and accounting in which analysis of binary representations of ratio or ordinal scale data is both common and useful. Examples include testing for the weak form of the efficient markets hypothesis, and identification and analysis of patterns pertinent in technical analysis of stock price movements.[7] Further, non-parametric approaches can outperform parametric approaches in the analysis of ratio/interval scale data – in the presence, for example, of significant violation of the assumption, as made by many parametric tests, of normality in the data. [8,9]

2.2 *Null hypothesis*

The null hypothesis is one of symmetrical random behaviour, so 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.

2.3 *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).[10] It is inferred that, inter alia, the number of incidences of periods of short term 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 sequences ↑↑↓↓ and ↓↓↑↑ would be of special interest.[11] More generally, where cyclical behaviour is theorised or suspected, then incidence of occurrence of sequences of the form ↑…↑↓…↓ and ↓...↓↑...↑, detecting, respectively, local maxima and local minima ('hilltops' and 'valleys') in the series, would be of interest; as would those of the form ↑↓…↓↑ and ↓↑…↑↓, detecting runs from hilltop to valley or vice-versa. Alternatively, if a process of repeated innovation/shock and subsequent erosion is hypothesised (e.g., at firm-level, a profitability-boosting investment or other intervention, followed by reduction of profitability as a consequence of competitive forces), then sequences of the form ↓↑↓…↓ and ↓↑↑↓…↓ might be of particular interest. Or if close oscillation around some level (e.g. some competitive mean or theoretical equilibrium) is hypothesised, then sequences of the form ↓↑↓, ↑↓↑, ↓↑↑↓↑↓, ↑↓↓↑↓↑, etc, could be the focus.

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).[12]

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 $n \in \mathbb{N}^+$, and 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, $l \in \mathbb{N}^+$. Let $S_l(a, b)$, with $a, b \in \mathbb{N}^+$ and $1 \leq a \leq b \leq 1$, be the sub-sequence from the a^{th} to the b^{th} terms (inclusive) of S_l. Let the *overlap order* of S_l be denoted $p(S_l)$ and defined as follows: $p(S_l)$ is the largest i_0 such that $S_l(1, i) \equiv S_l(l - i + 1, l)$ for all $i \leq i_0 \in \mathbb{N}$. Let V_p denote the set of sequences of outcomes of Bernoulli trials with overlap order p. It is noted that $0 \leq p \leq l$. It is further noted that $S_l(1,0)$ and $S_l(l + 1, l)$ are not defined, and 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 V_0 $(\uparrow\ \uparrow\ \uparrow\ \downarrow\ \downarrow\ \downarrow)(\uparrow\ \uparrow\ \uparrow\ \downarrow\ \downarrow\ \downarrow)(\uparrow\ \uparrow\ \uparrow\ \downarrow\ \downarrow\ \downarrow)\ (\cdots$

↑↓↓↓↑ is in V_1 $(\uparrow\ \downarrow\ \downarrow\ \downarrow\{\uparrow)\downarrow\ \downarrow\ \downarrow(\uparrow\}\downarrow\ \downarrow\ \downarrow\{\uparrow)\downarrow\ \cdots$

↑↑↓↓↓↑↑ is in V_2 $(\uparrow\ \uparrow\ \downarrow\ \downarrow\ \downarrow\{\uparrow\ \uparrow)\downarrow\ \downarrow\ \downarrow(\uparrow\ \uparrow\}\downarrow\ \downarrow\ \downarrow\{\uparrow\ \uparrow)\downarrow\ \cdots$

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

The following are evident:

$$\text{if } n = l \qquad \text{then } X(n,1,l,p) = 1 \tag{7}$$

$$\text{if } n = 2l - p \qquad \text{then } X(n,2,l,p) = 1 \tag{8}$$

and, generally, for $i > 0$:

$$\text{if } n = il - (i-1)p = i(l-p) + p \qquad \text{then } X(n,i,l,p) = 1 \tag{9}$$

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

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 V_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

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

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

4. Moments of the probability distributions under the null hypothesis

For given *n*, *l*, and *p*, we simplify the nomenclature by writing:

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

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

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

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

Similarly, we calculate Y_i' as follows:

$$Y_i' = \frac{Y_i}{2^n} \tag{17}$$

From expression (11), we deduce that:

$$Y_i' = \sum_{j \geq i} {}^{j}C_i X_j' \tag{18}$$

Expression (18) encapsulates the useful result that Y_i' is a factorial moment generating function.[13] Writing W for the count of the number of occurrences of the sequence of interest (of length *l* and overlap order *p*) within a series of *n* Bernoulli trials), using $E(\cdot)$ to denote expectation, and with readily ascertainable $\alpha_1, \alpha_2, \cdots, \alpha_{i-1} \in \mathbb{Z}$:

$$Y_i' = \frac{E(W^i) + \alpha_{i-1}E(W^{i-1}) + \cdots + \alpha_1 E(W)}{i!} \tag{19}$$

Expressions (18) and (19) allow the moments of the distribution of W to be deduced readily. In particular, for *i* = 1 to 4:

$$\left.\begin{aligned} Y_1' &= \sum_{j\geq 1} jX_j' = E(W) \\ Y_2' &= \sum_{j\geq 2} j(j-1)X_j'/2! = [E(W^2) - E(W)]/2 \\ Y_3' &= \sum_{j\geq 3} j(j-1)(j-2)X_j'/3! = [E(W^3) - 3E(W^2) + 2E(W)]/6 \\ Y_4' &= \sum_{j\geq 4} j(j-1)(j-2)(j-3)X_j'/4! = [E(W^4) - 6E(W^3) + 11E(W^2) - 6E(W)]/24 \end{aligned}\right\} \tag{20}$$

from which we may retrieve the mean and higher order moments about the origin as follows:

$$E(W) = Y_1' \tag{21a}$$

$$E(W^2) = 2Y_2' + Y_1' \tag{21b}$$

$$E(W^3) = 6Y_3' + 6Y_2' + Y_1' \tag{21c}$$

$$E(W^4) = 24Y_4' + 36Y_3' + 14Y_2' + Y_1' \tag{21d}$$

and note that

$$E(W^i) = i!\,Y_i' + f(Y_1', Y_2', \cdots, Y_{i-1}') \tag{22}$$

where *f* represents a linear combination.

Therefore, we deduce the variance and standard deviation to be:

$$VAR(W) = 2Y_2' + Y_1' - (Y_1')^2 \tag{23a}$$

$$SD(W) = \sqrt{2Y_2' + Y_1' - (Y_1')^2} \tag{23b}$$

5. More on moments and asymptotic properties of the probability distributions under the null hypothesis

A further analytic expression for $Y_i^{'}$ is derived in Appendix 3 for the cases $p = 0$ and $p = 1$, and is as follows:

$$Y_i' = \frac{n^i}{2^{il} i!} + O\left(n^{i-1}\right) \tag{24}$$

where $O(n^x)$ represents orders of magnitude in n of x or less.

Adopting the notation that μ_r^* represents the r'th moment about zero (so $E(W^r) = \mu_r^*$), from expressions (22) and (24) we deduce that for the cases $p = 0$ and $p = 1$:

$$\mu_r^* = \frac{n^r}{2^{rl}} + O(n^{r-1}) \qquad\qquad r \epsilon \mathbb{N}^+ \tag{25}$$

We now call upon the general result for the conversion from moments about an arbitrary point to central moments (i.e. moments about the mean), in the case that the arbitrary point concerned is zero, and adopting the notation that μ_r represents the r'th central moment:

$$\mu_r = \sum_{j=0}^{r} {}^{r}C_j \, \mu_{r-j}^* (-\mu_1^*)^j \qquad\qquad r \epsilon \mathbb{N}^+ \text{ and } r \geq 2 \tag{26}$$

Each term in the sum indicated by this expression is of the form $(-1)^j \, {}^{r}C_j \mu_{r-j}^* (\mu_1^*)^j$. Referring to expression (25), this is equivalent to $(-1)^j \, {}^{r}C_j \left(\frac{n^r}{2^{rl}}\right) + O(n^{r-1})$; and the coefficient of $\left(\frac{n^r}{2^{rl}}\right)$ in the sum of expression (26) is, therefore, $\sum_{j=0}^{r} (-1)^j \, {}^{r}C_j$. This is a sum of standard binomial coefficients of alternating sign, and so is equal to zero. We deduce that the μ_r (for $r \geq 2$) have order of magnitude in n of at most n^{r-1}. Consistent with this, the second central moment, from expressions (12), (13), (17) and (23a), is as follows:

$$\mu_2 = \begin{Bmatrix} \frac{2^l + 1 - 2l}{2^{2l}} n + O(n^0) & case\ p = 0 \\ \frac{2^l + 5 - 2l}{2^{2l}} n + O(n^0) & case\ p = 1 \end{Bmatrix} \tag{27}$$

and the standard deviation is of order of magnitude $\sqrt{n}$.

We turn to consideration of the standardised forms of the distributions introduced in this paper, i.e. the distributions translated to be mean-centred, and scaled by standard deviation. In the standardised distributions, the *r*'th central moment is scaled by $\mu_2^{r/2}$; which here means scaling by $O\left(n^{r/2}\right)$ and, in particular, scaling the third central moments by $O\left(n^{3/2}\right)$. The foregoing establishes that that the pre-scaled third central moments are of order no more than $O(n^2)$; and we now address the size of their n^2 coefficient. Expressions (21a), (21b), (21c) and (26) may be combined to give the following in the case *r* = 3:

$$\mu_3 = 6Y_3' + 6Y_2' + Y_1' - 3(2Y_2' + Y_1')Y_1' + 2(Y_1')^3 \tag{28}$$

Algebraic expansions in terms of *n* and *l* for Y_1', Y_2' and Y_3' (cases $p = 0$ and $p = 1$), inter alia, are found in Appendix 3. Table 1 sets out a breakdown of the contributions of the terms in expression (28) to the coefficient of n^2 in μ_3.

*** insert Table 1 about here ***

Table 1 demonstrates that pre-scaled third central moments have no terms in n^2, and so are of order no more than $O(n^1)$. Therefore, the third moment of the standardised distributions tends to zero as *n* tends to infinity. This is suggestive that the distributions presented in this paper are asymptotically standard normal. Indeed, asymptotic normality of the distributions follows formally from the Central Limit Theorem for m-dependent sequences.[14] This allows for straightforward application of the approach presented in the paper to long series of high frequency and/or extended time period data. As regards such application, the following expressions in *n* and *l*, derived from expressions (12), (13), (17), (21a) and (23b) will be useful:

$$Mean(W) = (n + 1 - l).2^{-l} \qquad cases\ p = 0,\ 1 \tag{29}$$

$$SD(W) = \begin{cases} 2^{-l}\sqrt{(2^l + 1 - 2l)n + 3l^2 - 4l + 1 - 2^l(l-1)} & case\ p = 0 \\ 2^{-l}\sqrt{(2^l + 5 - 2l)n + 3l^2 - 12l + 9 - 2^l(l-1)} & case\ p = 1 \end{cases} \tag{30}$$

6. Simple illustrative applications

6.1 *Example distribution under the null hypothesis*

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 2 shows the pertinent distributions of Y, Y', X and X' as calculated from expressions (12), (17), (11) and (16) respectively. It also includes the cumulative probability distribution X'. The mean and standard deviation of W are 1.7500 and 1.0155, as calculated from expressions (29) and (30) respectively.

*** insert Table 2 about here ***

If the observed number of occurrences of ↓↓↑↑ in the series is, say, four, then we may deduce from the cumulative probability distribution that the that the null hypothesis may be rejected in favour of the alternative with greater than 5% significance; and if the observed number is five or more, the significance level is greater than 1%.

The distribution presented in Table 2 is next applied to a synthetic example, and then to an example involving real historic data – with both examples designed such that the parameters n = 31, l = 4, p = 0 (and so also Table 2) remain pertinent.

6.2 *Synthetic example application and comparison with the Wald-Wolfowitz runs test*

This subsection considers a time series determined by the following expression:

$$x_t = c_1 \sin\left(c_2 \frac{t}{2\pi m}\right) \tag{31}$$

where $m \in \mathbb{N}^+$ represents the number of periods (elements) of the time series; $t \in \mathbb{N}^+$ is the series ordinal, $1 \leq t \leq m$; and c_1, $c_2 \in \mathbb{R}^+$ are constants.

Expression (31) defines a sine wave which has c_2 cycles of amplitude c_1 over time period m. We focus on the particular time series where m = 32 (in line with sub-section 6.1 above), $c_1 = 1$ (unit amplitude of the sine wave) and $c_2 = 9$. Figure 1 presents a graphical representation of the series, along with its underlying sine wave.

*** insert Figure 1 about here ***

Applying expressions (1) and (2), we then deduce the following series of directions of movement, length 31:

↓ ↓ ↑ ↓ ↓ ↑ ↑ ↓ ↓ ↑ ↑ ↓ ↑ ↑ ↓ ↓
↑ ↑ ↓ ↑ ↑ ↓ ↓ ↑ ↑ ↓ ↓ ↑ ↓ ↓ ↑

The standard and well-known Wald-Wolfowitz runs test yields a test statistic of 0.5545; so we are unable to reject the runs test null hypothesis of randomness in the original underlying series at any generally acceptable level of significance.[15] Using the approach developed in this paper, however, and focusing on the number of occurrences of observed number of occurrences of the sequence ↓↓↑↑ (which is four), we may reject the null of randomness in the original underlying series with significance at more than 5%.[16] We might, of course, focus on alternative sequences using the approach developed in this paper to obtain further, similar results.

The comparison between the Wald-Wolfowitz test and that developed in this paper merits further consideration: concern might pertain as to apparent difference in alternative hypothesis. The Wald-Wolfowitz runs test and the test (or family of tests) proposed in the paper take as their null hypothesis random behaviour of the binomial series under investigation; more specifically, that the terms of the series are identically and independently distributed. Both take as their alternative hypothesis non-random behaviour in the series, i.e. non-independence of the series terms. The approach proposed in this paper, however, allows the adoption of a wide range of operational forms for the alternative hypothesis which are more specific than the operational form of the alternative hypothesis of the standard runs test (whose focus, as its name suggests, is fixed on the number of runs in the series). Rather than a cause for concern, this is a distinct strength of the proposed approach, for two reasons. First, the form of alternative hypothesis may be chosen to test with respect to some particular theory, type of data generating process, or pattern tentatively identified from visual inspection of the data. Second, in allowing for the adoption of a more specific forms of alternative hypothesis, the approach in this paper allows for a more specific form of test which, in turn, has greater power (ceteris paribus) than the standard test.

6.3 *Example based on real, historic data: the dynamics of corporate earnings*

This sub-section considers the mean-adjusted earnings series of listed UK manufacturing and services companies over the period 1968 to 1999. The earnings series are defined by $\boldsymbol{\rho_{i,t}} = \left(\boldsymbol{\pi_{i,t}} - \overline{\boldsymbol{\pi}}_{\boldsymbol{t}}\right)/\overline{\boldsymbol{\pi}}_{\boldsymbol{t}}$, the proportional deviation of the *i*'th firm's profitability from the norm at time *t*, where $\boldsymbol{\pi_{i,t}}$ is the de-geared pre-taxation return on capital of the of the *i*'th firm at time *t*, and the profitability norm, $\overline{\boldsymbol{\pi}}_{\boldsymbol{t}}$, is proxied by the population mean of such returns at time *t*.[17] Unbroken time series of necessary data over the whole period 1968-1999 are available for only 53 companies; but the population mean, value weighted by total assets, subsumes all FTSE All Share manufacturing and services companies for which data is available each year. The ρ variable is a ratio measure; it is deflated by total assets; it is immune to the effects of differing gearing and

taxation regime (across sample and over time); and it is mean-adjusted, so general business cycle effects are removed, or, at least, mitigated. Therefore, it might reasonably be expected to be one of the 'better behaved' measures of earnings, in comparison, for example, with non-deflated earnings series, clean surplus earnings, and so forth.

Applying expressions (1) and (2) to each of the 53 firm-level time series, we deduce 53 series of direction of movement in mean-adjusted earnings, each having 31 terms. Using standard runs tests upon these series, the null of randomness may not be rejected at the 1% significance level for any firm; and may be rejected at the 5% significance level in only two cases. Further, in the two cases where non-randomness is indicated by the runs test, the test does not speak to what pattern or type of non-random dynamics are being detected.

On plotting and inspecting the time series of ρ, however, one feature is that some of the series seem to contain cycles with period of around four to six years, albeit with clear evidence also of other forms of dynamic.[18] This observation as regards possible cyclic behaviour in some cases leads to the following inferences, which are susceptible to direct test testing via the approach developed in this paper: (i) the number of incidences of periods of short term (say, two-year) sustained growth or decrease in ρ should be greater than would be expected were ρ to be random; and (ii) the number of reversals in ρ (that is, monotone increase followed by monotone decrease, or vice-versa) should also be greater than that expected were ρ to be random. Therefore, the frequency of occurrence of the following sequences within the series of direction of movement in ρ are of special interest: sequences capturing two-year monotone increasing/decreasing behaviour (bounded for exact length), ↓↑↑↓ and ↑↓↓↑; and sequences capturing 'turning' within cycles, ↑↑↓↓ and ↓↓↑↑.

Table 3 shows number of firm-level time series for which the null of randomness ρ may be rejected at generally acceptable levels of significance based on analysis of incidence of occurrence of these identified sequences of special interest. This provides clear evidence of non-randomness in the mean-adjusted earnings for more firms and with greater significance than is possible via the standard runs test. The incidence of occurrence of the sequence ↑↑↓↓ is particularly telling. Furthermore, as and where non-randomness is indicated, the alternative, non-random dynamic is explicit in the sequences investigated.

*** insert Table 3 about here ***

In practice, series will often be longer, sometimes very much longer indeed, than the example series of this sub-section; and the asymptotic normality discussed in Section 5, with expressions (29) and (30), will be pertinent to allow straightforward application of the approach presented in

this paper. Commonly periods of sustained monotone increase or decrease in the series will be of interest – as regards transitions from local maxima to local minima (and vice-versa) in cycles, and periods of sustained growth or decay. From inspection of the data, perhaps tabulation of runs lengths, or according to expectations from theory, or by some other means, it might be decided, for example, that the incidence of occurrence of monotone increasing sequences of length seven, i.e. ↓↑↑↑↑↑↑↓, is of particular interest. If the series has, say, 10,000 terms, then relevant statistics under the null hypothesis, readily calculated from expressions (29) and (30), are mean 19.5156 occurrences with standard deviation 4.3612. Alternatively, there might be interest in the number of monotone increasing sequences of length seven *or more*. In this case, focus is upon the sequences of the form ↓↑↑↑↑↑↑; and the relevant statistics under the null are then mean 39.0352, standard deviation 6.0621.[19]

7. 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. Also provided is a demonstration the distributions are asymptotically normal, and, therefore, application to very large datasets is straightforward.

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., all have overlap order $p = 0$. 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. also all have overlap order $p = 0$. 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 have overlap order $p = 1$. Nevertheless, theoretical work to further generalise the analytic expressions is desirable.

Notes

[1] See Sivarajah et al. (2017) for an interesting review of the challenges and methods associated with big data.
[2] In addition to presenting an approach for the distribution theory of runs based on FMCI, Fu and Koutras (1994) also provide a number of references to various combinatorial approaches in this area.
[3] 'Symmetry' in Bernoulli trials meaning that the probability of one outcome equals the probability of the other (both being 0.5); and 'overlap order' being as defined in the third paragraph of Section 3.
[4] See, for example, Knuth (1998) chapter 3.
[5] Nomenclature regarding attributes, labels and ranking of ordinal scale classes follows Siegel and Castellan (1988) section 3.3.
[6] This is true for the standard Wald-Wolfowitz runs test – against which the framework presented in this paper is compared in Section 6, and against which it is shown to compare favourably.
[7] See Edwards et al. (2007) for an excellent overview of technical analysis.
[8] See, for example, Hallin and Mélard (1988).
[9] It is not uncommon in literature for empirical analysis to be based on the assumption of normally distributed data, whilst at the same time reporting test statistics (e.g. Jarque-Bera statistics) which imply the assumption is violated – and without recognition or discussion of the impact of non-normality.
[10] This is the case in the illustrative application set out in Section 6.3.
[11] Note that the terminology 'sub-sequence' is dropped at this point in favour of the less cumbersome 'sequence'.
[12] 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.
[13] Kendall et al. (1987) sections 3.7-3.11 gives a general treatment of factorial moments and associated generating functions.
[14] See, for example, Billingsley (1995, 364), Ferguson (1996, 70) and Bradley (2007).
[15] It is, of course, straightforward to 'fool' the standard runs test using a variety of synthetic series. In the set-up of Section 6.2, this may be achieved simply by adjusting parameter c_2 in expression (31). Here, any value of c_2 chosen from the integer range [6, 10] results in inability to reject the null of randomness under the standard runs test at a generally acceptable level of significance.
[16] Referring to the cumulative probability distribution under the null hypothesis, as shown in Table 2.
[17] All necessary variables obtained from Datastream.
[18] In some cases the variable appears to cycle, then undergo an innovation (either cycling upwards/downwards, or via a jump) to a new level around which cycles recommence. This is highly suggestive of a concatenation of differing generating processes, rather than processes which are constant in terms of structure and parameters. Inspection also suggests that the variables' dynamics appear to be largely firm-idiosyncratic: common patterns are hard to detect, even amongst firms in broadly similar industry categories.
[19] Notice that in the sequence ↓↑↑↑↑↑↑↑↓ (for which $l = 9$ and $p = 1$), terms '↓' at the beginning and the end demarcate a run of exactly seven '↑' terms. In the sequence ↓↑↑↑↑↑↑↑ (for which $l = 8$ and $p = 0$), the end of the run of seven increases is not demarcated with a '↓' term.

Acknowledgments

I am very grateful to two anonymous referees for the care which they took in review of this paper and for the expertise which they brought to bear. Also to Mark Tippett for his constructive criticism and his support in the development of this work. Any remaining errors remain my sole responsibility.

References

Billingsley, P. 1995. *Probability and Measure*, 3rd Edition. New York: Wiley.

Bradley, R. 2007. *Introduction to strong mixing conditions*. Heber City, Utah: Kendrick Press.

Callen, J., C. Cheung, C. Kwan, and R. Yip. 1993. "An Empirical Investigation of the Random Character of Annual Earnings." *Journal of Accounting, Auditing and Finance* 8(2): 151-162.

Edwards, R. D., J. Magee, and W. H. C. Bassetti. 2007. *Technical Analysis of Stock Trends*, 9th Edition. Boca Raton, Florida: CRC Press.

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

Fama, E. F. 1991. "Efficient Capital Markets: II." *The Journal of Finance* 46(5): 1575-1617.

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

Ferguson, T.S. 1996. *A Course in Large Sample Theory*. Boca Raton, Florida: CRC Press.

Fu, J. C. and M. V. Koutras. 1994. "Distribution Theory of Runs: A Markov Chain Approach." *Journal of the American Statistical Association* 89(427): 1050-1058.

Fu, J. C. and W. Y. Lou. 2003. *Distribution Theory of Runs and Patterns and its Applications: A Finite Markov Chain Imbedding Approach*. River Edge, New Jersey: World Scientific Publishing.

Gray, J. R. 1967. *Probability*. London: Oliver and Boyd Limited.

Hallin, M. and G. Mélard. 1988. “Rank-Based Tests for Randomness Against First-Order Serial Dependence.” *Journal of the American Statistical Association* 83(404): 1117-1128.

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

Kendall, M. G., A. Stuart, and J. K. Ord. 1987. *Kendall’s Advanced Theory of Statistics, Volume 1: Distribution Theory*, 5th Edition. London: Charles Griffin & Co.

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

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

Lindeberg, J. W. 1922. “Eine neue Herleitung des Exponentialgesetzes in der Wahrsheinlichkeits-rechnung.” *Mathematische Zeitschrift* 15: 211-225 (as cited by Feller, 1968).

Lo, A. W. and J. Hasanhodzic. 2010. *The Evolution of Technical Analysis: Financial Prediction from Babylonian Tablets to Bloomberg Terminals*. Wiley.

Nichols, D. C., J. M Wahlen, and M. M. Wieland. 2017. “Pricing and Mispricing of Accounting Fundamentals in the Time-Series and in the Cross Section.” *Contemporary Accounting Research* 34(3): 1378-1417.

O’Hanlon, J. 1995. “The Univariate Time Series Modelling of Earnings: A Review.” *British Accounting Review* 27(3): 187-210.

Ohlson, J. A. 1995. “Earnings, Book Values, and Dividends in Equity Valuation.” *Contemporary Accounting Research* 11(2): 661-687.

Schleifer, A. 2000. *Inefficient Markets*. Oxford: Oxford University Press.

Seddon, J. J. J. M. and W. L. Currie. 2017. “A model for unpacking big data analytics in high-frequency trading.” *Journal of Business Research* 70: 300-307.

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

Sivarajah, U., M. M. Kamal, Z. Irani, and V. Weerakkody. 2017. "Critical analysis of Big Data challenges and analytical methods." *Journal of Business Research* 70: 263–286.

Tippett, M. and T. Warnock. (1997). "The Garman-Ohlson structural system." *Journal of Business Finance and Accounting*, 24(7-8): 1075-1099.

Watts, R. L. and R. W. Leftwich. 1977. "The Time Series of Annual Accounting Earnings." *Journal of Accounting Research* 15(2): 253-271.

Appendix 1. Derivation of analytic expressions for distributions under the null

A1.1 *Introduction and overview*

Definitions of B_n, S_l, overlap order $p(S_l)$, 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 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.

A1.2 *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:

If $i > 1$, let the ABPs be: (i) the positions at the end of the series B_n – giving two 'exterior' ABPs; and (ii) the positions immediately to the right of the first $(i - 1)$ occurrences of S_l – giving a further $(i - 1)$ 'interior' ABPs, $i \geq 1$.

If $i = 1$, let the ABPs be 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 the single ABP be the position at the end of the series B_n.

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, $j \in \mathbb{N}^+$.

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; and let 'legal addition' denote the addition of a term in an allowable building position which results in a legal build.

Let 'illegal build' denote any build which is not a legal build; and let 'illegal addition' denote the addition of a term in an allowable building position which results in an illegal build.

A1.3 *Derivation: case p = 0*

In this case, expression (9) becomes: if $n = il$ then $X(n, i, l, 0) = 1$. We now consider the specific case in which l divides n where $il = N_1$, 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_1. See, for example, Figure A1.

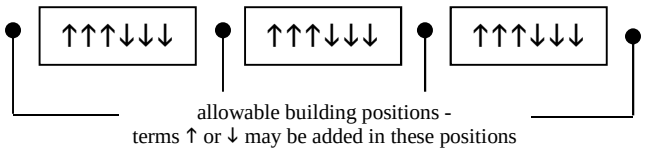

Figure A1. Example: $i = 3$, $l = 6$, $p = 0$; situation when $n = 18$, $X(18,3,6,0) = 1$

It is a standard result in combinatorics that the number of possible distinguishable arrangements of a indistinguishable objects into b distinguishable compartments is ${}^{a+b-1}C_a$, where C represents combination (e.g. Gray 1967, 97-98). 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:

$$Y(n, i, l, 0) = \begin{cases} \left({}^{(i+n-il)}C_{(n-il)}\right) 2^{(n-il)} & \text{for } n \geq il \\ 0 & \text{for } n < il \end{cases} \qquad \text{(A1)}$$

Writing $n - [i(l - p) + p] = k$, and given that we are dealing with case $p = 0$, expression (A1) 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 of 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 i 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 $X(n,j,l,0)$ may be calculated by adjustment of the $Y(n,i,l,0)$ by application of the following backwards recursive formula:

$$X(n,i,l,0) = \begin{cases} Y(n,i,l,0) - \sum_{j>i} {}^{j}C_{i} X(n,j,l,0) & \text{for } n \geq il \\ 0 & \text{for } n < il \end{cases} \qquad \text{(A2)}$$

It is noted that from this expression together with expression (A1), substituting $i = 0$ and rearranging, we obtain $\sum_{j\geq 0} X(n,j,l,0) = 2^n$, which is as expected. Expression (A2) is equivalent to expression (11) for $p = 0$ (QED).

A1.4 ***Derivation: case p = 1***

In this case, statement (9) becomes: if $n = i(l-1)+1$ then $X(n,i,l,1) = 1$. We now consider the specific case in which $i(l-1)+1 = N_3$, say; and there are 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 and subsequent terms added to interior allowable building positions. See, for example, Figure A2.

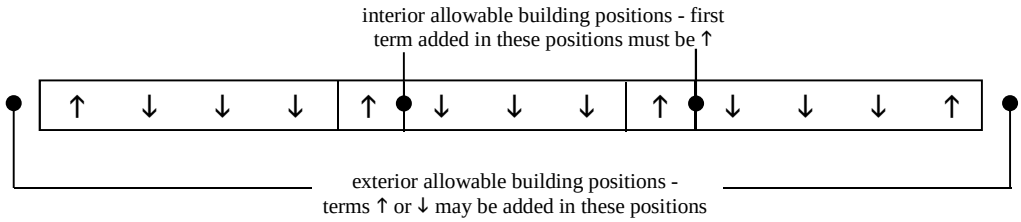

Figure A2. Sequence of interest ↑↓↓↓↑, $i = 3$, $l = 5$, $p = 1$; situation when $n = 13$, $X(13,3,5,1) = 1$

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 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, j of which additions are illegal additions to distinct interior allowable building positions, $j \in \mathbb{N}$, $1 \leq j \leq i-1$.

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

$$L_1 - \Big(L_2 - \big(L_3 - \cdots - (L_{i-2} - L_{i-1})\big)\Big) = (-1)\sum_{j=1}^{i-1}(-1)^j L_j \qquad \text{(A3)}$$

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

$$L_j = \left({}^{i-1}C_j \cdot 1^j\right)\left({}^{i+k-j}C_{k-j} \cdot 2^{k-j}\right) \tag{A4}$$

The total number of distinct B_n for $n > N_3$ built from B_{N_3} via all builds (legal and illegal), D, is given by adaption of the calculation of expression (A1) to the following, again writing $k = n - [i(l-1)+1]$:

$$D = \begin{cases} {}^{i+k}C_k \cdot 2^k & \text{for } n \geq i(l-1)+1 \\ 0 & \text{for } n < i(l-1)+1 \end{cases} \tag{A5}$$

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 (A3), (A4) and (A5):

$$Y(n,i,l,1) = \begin{cases} 2^n & \text{for } i = 0 \\ (-1)^k \sum\limits_{\substack{j=max \\ (0,k-(i-1))}}^{k} {}^{i-1}C_{k-j} \cdot {}^{i+j}C_j \cdot 2^j \cdot (-1)^j & \text{for } i > 0 \text{ and } n \geq i(l-1)+1 \\ 0 & \text{for } n < i(l-1)+1 \end{cases} \tag{A6}$$

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:

$$X(n,i,l,1) = \begin{cases} Y(n,i,l,1) - \sum\limits_{j>i} {}^{j}C_i X(n,j,l,1) & \text{for } n \geq i(l-1)+1 \\ 0 & \text{for } n < i(l-1)+1 \end{cases} \tag{A7}$$

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

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

$$X(n,i,l,p) = \begin{cases} Y(n,i,l,p) - \sum\limits_{j>i} {}^{j}C_i X(n,j,l,p) & \text{for } n \geq i(l-p)+p \\ 0 & \text{for } n < i(l-p)+p \end{cases} \tag{A8}$$

Appendix 2. Numerical illustration of reasoning the Appendix 1 derivation in 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 (A5), 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 (A4)

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 $=\left({}^{3}C_{3}\cdot 1^{3}\right)\left({}^{6}C_{2}\cdot 2^{2}\right)=60$

In illustration of expression (A4), 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 making exactly two non-allowable additions into two separate interior allowable building positions, and otherwise

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 making exactly one non-allowable addition into an interior allowable position, and otherwise 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 (A6).

Since $Y(26{,}5{,}6{,}1) = 1$, because $5(l-1) + 1 = 26$, and $Y(26, i, 6{,}1) = 0$ for all $i > 5$, $X(26{,}4{,}6{,}1)$ may be calculated from expression (28) as: $1{,}452 - {}^{5}C_4 \cdot 1 = 1{,}447$.

Appendix 3. Derivation of further analytic expressions for Y'_i

A3.1 *Case p* = 0

From expressions (12) and (17), and noting that for $p = 0$, $k = n - il$:

For $n \geq il$, $Y'_i = {}^{1+n-il}C_{n-il} \cdot 2^{-il}$ and, writing $O(n^x)$ to represent orders of magnitude in n of x or less:

$$Y'_1 = (1 + n - l)2^{-l} = \frac{n}{2^l} + O(n^0)$$

$$Y'_2 = (2 + n - 2l)(1 + n - 2l)2^{-2l}/2! = \frac{n^2}{2^{2l}2!} + O(n^1)$$

$$Y'_3 = (3 + n - 3l)(2 + n - 3l)(1 + n - 3l)2^{-3l}/3! = \frac{n^3}{2^{3l}3!} + O(n^2)$$

$$Y'_4 = (4 + n - 4l)(3 + n - 4l)(2 + n - 4l)(1 + n - 4l)2^{-4l}/4! = \frac{n^4}{2^{4l}4!} + O(n^3)$$

and, generally,

$$Y'_i = \frac{n^i}{2^{il}i!} + O\left(n^{i-1}\right) \tag{A9}$$

A3.2 *Case p = 1*

From expressions (13) and (17), and noting that for $p = 1$, $k = n - [i(l-1) + 1]$:

For $n \geq i(l-1) + 1$

$$\begin{aligned} Y'_1 &= {}^0C_0 \cdot {}^{1+n-l}C_{n-l} \cdot 2^{-l} \\ &= (1 + n - l)2^{-l} \\ &= \frac{n}{2^l} + O(n^0) \end{aligned}$$

$$\begin{aligned} Y'_2 &= {}^1C_0 \cdot {}^{n-2l+3}C_{n-2l+1} \cdot 2^{-2l+1} - {}^1C_1 \cdot {}^{n-2l+2}C_{n-2l} \cdot 2^{-2l} \\ &= (n - 2l + 3)(n - 2l + 2)2^{-2l+1}/2! - (n - 2l + 2)(n - 2l + 1)2^{-2l}/2! \\ &= \frac{2n^2}{2^{2l}2!}\left(1 - \frac{1}{2}\right) + O(n^1) = \frac{n^2}{2^{2l}2!} + O(n^1) \end{aligned}$$

$$
\begin{aligned}
Y_3' = {} & {}^{2}C_0 \cdot {}^{n-3l+5}C_{n-3l+2} \cdot 2^{-3l+2} - {}^{2}C_1 \cdot {}^{n-3l+4}C_{n-3l+1} \cdot 2^{-3l+1} \\
& + {}^{2}C_2 \cdot {}^{n-3l+3}C_{n-3l} \cdot 2^{-3l} \\
= {} & (n-3l+5)(n-3l+4)(n-3l+3)2^{-3l+2}/3! \\
& - 2(n-3l+4)(n-3l+3)(n-3l+2)2^{-3l+1}/3! \\
& + (n-3l+3)(n-3l+2)(n-3l+1)2^{-3l}/3! \\
= {} & \frac{2^2 n^3}{2^{3l}3!}\left(1 - 2 \cdot \frac{1}{2} + \frac{1}{4}\right) + O(n^2) = \frac{n^3}{2^{3l}3!} + O(n^2)
\end{aligned}
$$

$$
\begin{aligned}
Y_4' = {} & {}^{3}C_0 \cdot {}^{n-4l+7}C_{n-4l+3} \cdot 2^{-4l+3} - {}^{3}C_1 \cdot {}^{n-4l+6}C_{n-4l+2} \cdot 2^{-4l+2} \\
& + {}^{3}C_2 \cdot {}^{n-4l+5}C_{n-4l+1} \cdot 2^{-4l+1} - {}^{3}C_3 \cdot {}^{n-4l+4}C_{n-4l} \cdot 2^{-4l} \\
= {} & (n-4l+7)(n-4l+6)(n-4l+5)(n-4l+4)2^{-4l+3}/4! \\
& -3(n-4l+6)(n-4l+5)(n-4l+4)(n-4l+3)2^{-4l+2}/4! \\
& +3(n-4l+5)(n-4l+4)(n-4l+3)(n-4l+2)2^{-4l+1}/4! \\
& -(n-4l+4)(n-4l+3)(n-4l+2)(n-4l+1)2^{-4l}/4! \\
= {} & \frac{2^3 n^4}{2^{4l}4!}\left(1 - 3 \cdot \frac{1}{2} + 3 \cdot \frac{1}{4} - \frac{1}{8}\right) + O(n^3) = \frac{n^4}{2^{4l}4!} + O(n^3)
\end{aligned}
$$

We recognise, in the last line of each case, that the bracket multiplier to the expression in n^i is equivalent to the binomial expansion of $\left(1 - \frac{1}{2}\right)^{i-1}$, and again we have generally:

$$
Y_i' = \frac{n^i}{2^{il}i!} + O\left(n^{i-1}\right) \tag{A10}
$$

It is noted that expressions (A9) and (A10) are equivalent; and are the same as expression (24) (QED).

Table 1. Coefficient of n^2 in μ_3

Term in expression (28)	Contribution to n^2coefficient in μ_3 ($\times 2^{-3l}$)	
	Case $p = 0$	Case $p = 1$
$+6Y_3'$	$-9l + 6$	$-9l + 18$
$+6Y_2'$	3.2^l	3.2^l
$+Y_1'$	0	0
$-3(2Y_2' + Y_1')Y_1'$	$15l - 12 - 3.2^l$	$15l - 24 - 3.2^l$
$+2(Y_1')^3$	$-6l + 6$	$-6l + 6$
Total	0	0

Table 2. Distribution of the incidence of occurrence of sequence ↓↓↑↑ within series of 31 independent Bernoulli trials ($n = 31$, $l = 4$, $p = 0$)

	Number of occurrences of sequence of interest (i)							
	0	1	2	3	4	5	6	7
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
Y'(31,i,4,0)	1.0000	1.7500	1.1719	0.3760	0.0591	0.0042	0.0001	0.0000
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
X'(31,i,4,0) = probability	0.1010	0.3178	0.3587	0.1790	0.0398	0.0036	0.0001	0.0000
cum *X'(31,i,4,0)* = cum probability	0.1010	0.4188	0.7775	0.9565	0.9963	0.9999	1.0000	1.0000

Table 3. Number of firms for which the null of randomness in mean-adjusted earnings may be rejected at generally acceptable levels of significance

Sequence	Number of firms for which we may reject the null	
	at 1% significance	at 5% significance
↓↑↑↓	1	3
↑↓↓↑	1	1
↑↑↓↓	0	7
↓↓↑↑	2	2

Notes: Tests performed using the approach developed in this paper, by considering the incidence of occurrence of different sub-sequences of interest. Earnings series of a sample of 53 UK listed manufacturing and services companies, 1968-1999.

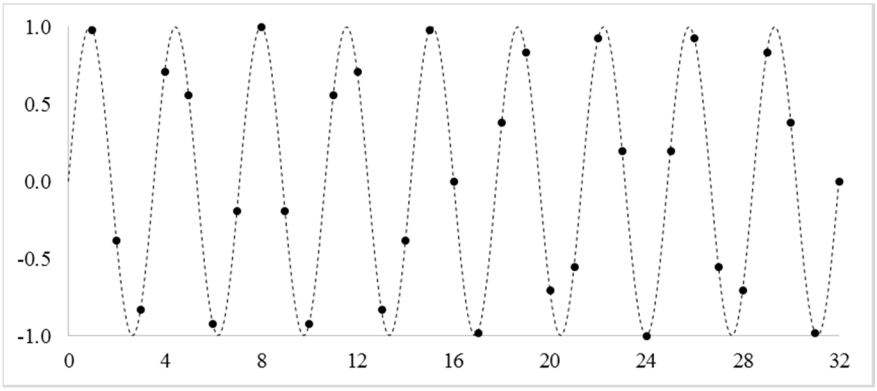

Figure 1. Synthetic example time series data with underlying sine wave.