Hassler, Uwe; Hosseinkouchack, Mehdi

Understanding nonsense correlation between (independent) random walks in finite samples

Statistical Papers

Provided in Cooperation with:
Springer Nature

Suggested Citation: Hassler, Uwe; Hosseinkouchack, Mehdi (2021) : Understanding nonsense correlation between (independent) random walks in finite samples, Statistical Papers, ISSN 1613-9798, Springer, Berlin, Heidelberg, Vol. 63, Iss. 1, pp. 181-195, https://doi.org/10.1007/s00362-021-01237-0

Statistical Papers (2022) 63:181–195
https://doi.org/10.1007/s00362-021-01237-0

REGULAR ARTICLE

Understanding nonsense correlation between (independent) random walks in finite samples

Uwe Hassler[1] · **Mehdi Hosseinkouchack**[2]

Received: 15 June 2020 / Accepted: 21 April 2021 / Published online: 6 May 2021

Abstract
Consider two independent random walks. By chance, there will be spells of association between them where the two processes move in the same direction, or in opposite direction. We compute the probabilities of the length of the longest spell of such random association for a given sample size, and discuss measures like mean and mode of the exact distributions. We observe that long spells (relative to small sample sizes) of random association occur frequently, which explains why nonsense correlation between short independent random walks is the rule rather than the exception. The exact figures are compared with approximations. Our finite sample analysis as well as the approximations rely on two older results popularized by Révész (Stat Pap 31:95–101, 1990, *Statistical Papers*). Moreover, we consider spells of association between correlated random walks. Approximate probabilities are compared with finite sample Monte Carlo results.

Keywords Coin tossing · Concordance · Discordance · Maximum length of association

Mathematics Subject Classification 60G50 · 62H20

1 Introduction

The puzzle why "we sometimes get nonsense-correlation between time-series" has first been addressed in the seminal paper by Yule (1926). One model that he sug-

The authors thank Matei Demetrescu, Tobias Hartl, Marc-Oliver Pohle and Jan Reitz for many helpful comments. Moreover, suggestions by two anonymous reviewers are gratefully acknowledged.

✉ Uwe Hassler
hassler@wiwi.uni-frankfurt.de

[1] Statistics and Econometric Methods, Goethe University Frankfurt, Theodor-W.-Adorno-Platz 4, 60323 Frankfurt, Germany

[2] University of Mannheim, Mannheim, Germany

gested to explain correlation between independent series was the random walk, called "conjunct series the differences of which are random" by Yule (1926, p. 26). For independent random walks Yule (1926, p. 33) provided experimental evidence, obtained by drawing playing cards from shuffled packs, that "The frequency-distribution of the correlations of samples of 10 observations [...] are much more widely dispersed than the correlations from samples of random series". His findings were accomplished by the computer experimental evidence on spurious regressions by Granger and Newbold (1974) for independent random walks of length 50, see also Palm and Sneek (1984) for further Monte Carlo results. Phillips (1986) showed that nonsense correlation between independent random walks is not a finite sample problem only. From Phillips (1986, Thm. 1) the limiting distribution of the sample correlation is available: it converges to a nondegenerate random variable. More recently, Ernst et al. (2017) determined the variance of this limit, and numerical evaluation showed that it equals 0.240522 (Ernst, Shepp and Wyner (2017, p. 1807)). Of course, such findings cannot fully explain why nonsense correlation occurs between random walks in small samples.[1]

In this note, we return to the finite sample puzzle. Yule (1926, Fig. 14) observed that random walks may trend in the same direction (concordance) or in the opposite direction (discordance) for certain periods of time. This is an intuitive explanation for nonsense correlation: there will be cluster of association between independent random walks. To add some rigour to this intuition, we would like to know: what is the maximum length to be expected for such spells of concordance or discordance given a fixed sample size? How large is the mode of this maximum length? And how large is the probability to observe values equal to or even larger than the mode? These questions will be answered by means of the corresponding probability distribution given in Corollary 1, building on the little-known Hungarian paper by Székely and Tusnády (1976-1979), see Révész (1990, Thm. 7) and Révész (2013, Thm. 2.7) for a reference. For independent random walks of length $n = 25$ we learn for instance: The probability that the maximum length of spells with consecutive concordance, or consecutive discordance, is at least equal to 4 amounts to 84.76%. Hence, long spells of random association (relative to the small sample size) are rather likely. The merits of exact results will be demonstrated by comparison with approximations. Asymptotic results in Proposition 2 can be traced back to Földes (1975), which is again a Hungarian paper referenced by Révész (1990, Thm. 6). Further, Gordon et al. (1986) provided approximations that allows for correlated random walks, too, which will be evaluated at the end of our note. Since no results for exact probabilities are available, we confront the asymptotic results with finite sample Monte Carlo figures.

The rest of this paper is organized as follows. The next section motivates this study with some Monte Carlo results. Section 3 becomes precise on random association and provides the exact distributional result under independence. The latter is evaluated numerically in Sect. 4 to shed light on why nonsense correlation is likely between independent random walks in finite samples. Section 5 compares the exact results with approximations. Section 6 is devoted to the extension of correlated random walks. A short summary is contained in the final section.

[1] Note that nonsense correlation is not limited to independent random walks but arises similarly due to (neglected) time-variation in mean from series with constant autocovariance structure, see Hassler (2003, Prop. 1).

A word on notation before we begin. Let $\lfloor x \rfloor$ denote the integer part of $x \in \mathbb{R}$, with fractional part $\{x\} := x - \lfloor x \rfloor$. Let $\log_b$ stand for the logarithm to the base b, while ln denotes the natural logarithm.

2 Some experimental evidence

Consider a bivariate random walk $(X_i, Y_i)_{i=0,1,\ldots,n}$ defined by

$$X_i = X_{i-1} + \varepsilon_i \text{ and } Y_i = Y_{i-1} + \eta_i\,, \quad i = 1, \ldots, n\,, \tag{1}$$

where (X_0, Y_0) is an arbitrary starting value. Before we begin with the theory, let us collect some experimental evidence. For computer simulation, the differences $(\Delta X_i, \Delta Y_i) = (\varepsilon_i, \eta_i)$ are drawn from a bivariate normal distribution:

$$\begin{pmatrix} \varepsilon_i \\ \eta_i \end{pmatrix} \sim \mathcal{N}_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right). \tag{2}$$

We simulated random walks with $(X_0, Y_0) = (0, 0)$ and computed the sample correlation:

$$\widehat{\rho} = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^n (X_i - \overline{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \overline{Y})^2}}.$$

Then we took the absolute value $|\widehat{\rho}|$ (since it is known that $\widehat{\rho}$ varies symmetrically around zero for $\rho = 0$). We report the average over 10^5 replications for growing sample size. Clearly, there is massive evidence in favour of nonsense correlation for $\rho = 0$, and the absolute correlation coefficients are of the same size for small n as for large n, see Table 1. For moderate correlation $\rho = 0.2, 0.4$, the sample correlation still exaggerates the true values, while $\rho = 0.6$ results in averages $|\widehat{\rho}| \approx 0.6$, and $\rho = 0.8$ yields on average $|\widehat{\rho}| \approx 0.77$, and these figures are rather robust over the sample size n, too.

In this paper we offer the length of random association between independent random walks or between moderately correlated random walks of small and medium sample sizes as an explanation for nonsense correlation or exaggerated correlation as documented in Table 1.

3 Spells of concordance and discordance

We maintain a bivariate random walk $(X_i, Y_i)_{i=0,1,\ldots,n}$ defined by equation (1), where (X_0, Y_0) is an arbitrary starting value. We now focus on independence (to be relaxed in Assumption 2). More precisely, the differences $(\Delta X_i, \Delta Y_i) = (\varepsilon_i, \eta_i)$ meet the following set of assumptions.

Table 1 Absolute value of sample correlation, $|\widehat{\rho}|$

n	25	50	100	200	400
$\rho = 0$	0.4246	0.4231	0.4215	0.4225	0.4228
$\rho = 0.2$	0.4451	0.4415	0.4404	0.4404	0.4400
$\rho = 0.4$	0.5015	0.5008	0.4997	0.4995	0.5018
$\rho = 0.6$	0.6052	0.6052	0.6031	0.6036	0.6058
$\rho = 0.8$	0.7649	0.7651	0.7658	0.7646	0.7647

Average over 10^5 replications of absolute values of sample correlations of Gaussian random walks with starting values equal to zero

Assumption 1 *Let* $(\varepsilon_i, \eta_i)_{i=1,\ldots,n}$ *be a sequence of independent, identically distributed and continuous random variables with*

$$p_\varepsilon := \mathrm{P}(\varepsilon_i < 0),\ \mathrm{P}(\varepsilon_i > 0) = 1 - p_\varepsilon, \quad p_\eta := \mathrm{P}(\eta_i < 0),\ \mathrm{P}(\eta_i > 0) = 1 - p_\eta,$$

$p_\varepsilon, p_\eta \in (0, 1)$. *Further,* ε_i *and* η_i *are independent, and at least one of the probabilities equals* $1/2$: $p_\varepsilon = 1/2$ *or* $p_\eta = 1/2$.

Remark 1 Note that the asymptotic theory by Phillips (1986) or Ernst et al. (2017) requires $\mathrm{E}(\varepsilon_i) = \mathrm{E}(\eta_i) = 0$, which we do not need. For Proposition 1 and 2 we just need that the median of ε_i or η_i equals zero.

We say that the variables from (1) are concordant on the ith interval if X_i and Y_i move in the same direction; if they move in the opposite direction, they are called discordant. In terms of the usual sign function this provides the following definition.

Definition 1 *Concordance on the* i*th interval means that* $\mathrm{sign}(\Delta X_i \Delta Y_i) = 1$. *Discordance on the* i*th interval means that* $\mathrm{sign}(\Delta X_i \Delta Y_i) = -1$.

Note that we rule out $\Delta X_i = 0$ or $\Delta Y_i = 0$ with probability 1 by assumption. For convenience, we define C_i as concordance indicator, taking on the value 0 if ΔX_i and ΔY_i have the same sign:

$$C_i = \begin{cases} 0 \text{ if } \mathrm{sign}(\Delta X_i \Delta Y_i) = 1 \\ 1 \text{ if } \mathrm{sign}(\Delta X_i \Delta Y_i) = -1 \end{cases}, \quad i = 1, \ldots, n\,. \tag{3}$$

By Assumption 1, it holds that

$$p := \mathrm{P}(C_i = 0) = 1 - p_\varepsilon - p_\eta + 2p_\varepsilon p_\eta = \frac{1}{2} = \mathrm{P}(C_i = 1)\,.$$

Consider a subsequence of consecutive zeros in $(C_i)_{i=1,\ldots,n}$, called a zero run. Let Z_n stand for the length of the longest zero run, which corresponds to the length of the longest spell without interruption where X_i and Y_i move in the same direction. The probabilities $\mathrm{P}(Z_n = k)$ for given n can be expressed in terms of generalized Fibonacci numbers. We adopt the most convenient definition for our purposes by Spickerman and Joyner (1984, p. 327).

Definition 2 *A Fibonacci sequence of order* ℓ, $(f_m^{(\ell)})_{m=1,2,\ldots}$ *for* $\ell \in \{1, 2, \ldots\}$, *is defined by the linear difference equation*

$$f_m^{(\ell)} = \sum_{i=1}^{\ell} f_{m-i}^{(\ell)} \quad \text{for } m > \ell\,,$$

with $f_m^{(\ell)} = 2^{m-1}$ *for* $m = 1, \ldots, \ell$.

The trivial case $\ell = 1$ covers a sequence of ones. For $\ell = 2$, the usual Fibonacci numbers are obtained. The case $\ell = 3$ has been called 'tribonacci' sequence, see e.g. Spickerman (1982). The following table corresponds to Székely and Tusnády (1976-1979, p. 149).

m	1	2	3	4	5	6	7	8	9
$f_m^{(1)}$	1	1	1	1	1	1	1	1	1
$f_m^{(2)}$	1	2	3	5	8	13	21	34	55
$f_m^{(3)}$	1	2	4	7	13	24	44	81	149
$f_m^{(4)}$	1	2	4	8	15	29	56	108	208

Trivially, $\mathrm{P}(Z_n < k) = 1$ for $k > n$. The general probability distribution is given next. Révész (1990, Thm. 7) and Révész (2013, Thm. 2.7) stated it without proof referring to Székely and Tusnády (1976-1979).

Proposition 1 *Let* $(X_i, Y_i)_{i=0,1,\ldots,n}$ *from equation (1) satisfy Assumption 1. It then holds that*

$$P(Z_n < k) = \frac{f_{n+1}^{(k)}}{2^n}\,, \quad 1 \le k \le n\,.$$

Proof See Székely and Tusnády (1976-1979). For completeness and easier accessibility, a separate proof is provided in the Appendix. □

By Proposition 1, it immediately follows that

$$\mathrm{P}(Z_n = k) = \frac{f_{n+1}^{(k+1)} - f_{n+1}^{(k)}}{2^n}\,, \quad 1 \le k \le n\,. \tag{4}$$

Further, $Z_n = 0$ corresponds to a sequence of n ones with probability $\mathrm{P}(Z_n = 0) = 1/2^n$.

More generally, we are interested in the length of the longest spell of consecutive intervals where X_i and Y_i are concordant or discordant without interruption. This corresponds to the maximum length of zero runs or runs of ones in (C_i). Let S_n stand for this length of the longest spell of consecutive ones or zeros. With Proposition 1, it is straightforward to establish the following distribution.

Table 2 Measures of S_n

	$n = 25$	$n = 50$	$n = 100$	$n = 200$	$n = 400$
expect	4.9799	5.9783	6.9774	7.9770	8.9768
var	2.6419	2.9983	3.2134	3.3401	3.4134
skew	1.2601	1.2173	1.1759	1.1465	1.1274
kurt	5.7909	5.6780	5.5361	5.4366	5.3742

Expected value, variance, skewness and kurtosis coefficients computed with probabilities from Corollary 1

Corollary 1 *Under the assumptions of Proposition 1 holds that*

$$P(S_n < k) = P(Z_{n-1} < k-1) = \frac{f_n^{(k-1)}}{2^{n-1}}, \quad 2 \le k \le n,$$

and $P(S_n < 1) = 0$.

Proof See Appendix. □

By Corollary 1, it immediately follows that

$$\mathrm{P}(S_n = k) = \frac{f_n^{(k)} - f_n^{(k-1)}}{2^{n-1}}, \quad 2 \le k \le n. \tag{5}$$

From (4) and (5) one obtains with $\mathrm{P}(S_n = 1) = 2^{1-n} = \mathrm{P}(Z_{n-1} = 0)$ that

$$\mathrm{P}(S_n = k) = \mathrm{P}(Z_{n-1} = k-1), \quad k = 1, \ldots, n, \tag{6}$$

which will be used below.

4 Numerical work

Given the relation in (6), our numerical evaluation will be restricted to the length of the longest spell of consecutive zeros or ones, S_n. The computation requires to determine (generalized) Fibonacci numbers. We employ the recursion from Definition 2 and do not bother about explicit solutions.

Statistical measures of S_n are given in Table 2, and they are illustrated by the plots in Fig. 1. For the expected values from Table 2 one observes a logarithmic rate: Doubling n adds roughly 1 to $\mathrm{E}(S_n)$; an asymptotic explanation for this feature will be given in the next Section. While the variance mildly grows with n, the skewness and the kurtosis decrease with the sample size. All in all, we find the spread in S_n rather small.

Looking more closely into the figures behind Fig. 1 reveals that the five outcomes with highest probabilities including the most probably value (mode) cover roughly 90% of the probability mass:

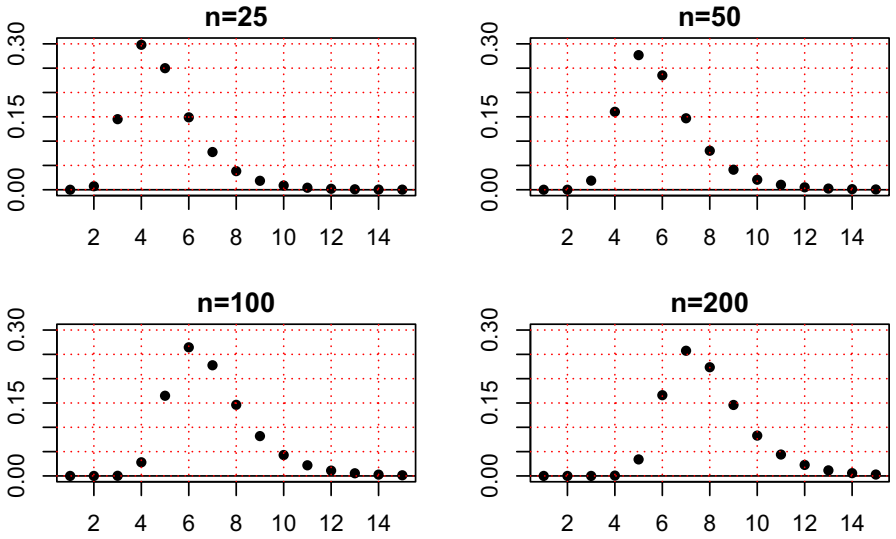

Fig. 1 $\mathrm{P}(S_n = k), k = 1, \ldots, 15$

$$\mathrm{P}(3 \le S_{25} \le 7) = 0.9195\,, \quad \mathrm{P}(4 \le S_{50} \le 8) = 0.8995\,,$$
$$\mathrm{P}(5 \le S_{100} \le 9) = 0.8850\,, \quad \mathrm{P}(6 \le S_{200} \le 10) = 0.8758\,.$$

Table 3 looks more closely at the mode, mod_n. While Fig. 1 and Table 2 are restricted to $n = 2^s \cdot 25$ for $s = 0, 1, 2, \ldots$, we consider now more generally $n = 2^s \cdot B$ and vary $B \in \{25, 30, 35\}$. From Table 3 we observe a logarithmic rate, $mod_n = \lfloor \log_2 n \rfloor = s + \lfloor \log_2 B \rfloor$;[2] as with the expectation this feature calls for an explanation provided in the next section. As we know from Fig. 1, the maximum probability decreases with n. For large n this probability seems to settle around 0.25 or slightly below, and an approximate explanation will be provided again in the next section. At the same time, it is interesting to look at the probabilities for larger values, say larger than the mode, $\mathrm{P}(S_n > mod_n)$: Throughout, the probability for the maximum length of a spell to exceed the mode is varies only very little with s given $n = 2^s \cdot B$, but the probability depends on B, which will be again clarified in the subsequent section. In any case we observe large probabilities $\mathrm{P}(S_n > mod_n)$: Long spells relative to the sample size of concordance or discordance will be the rule rather than the exception. This is in line with the experimental evidence documented in Table 1 for no correlation.

5 Approximate results

In this section we compare our exact figures from Table 2 and 3 and Fig. 1 with approximate figures. The following approximation can be traced back to Földes (1975), see Révész (1990, Thm. 6). Easier to access is the proof by Földes (1979), while

[2] This holds for our choices of n. In fact, we verified that it holds for the majority of values of n, but we found counterexamples, too, where $mod_n = \lfloor \log_2 n \rfloor \pm 1$.

Table 3 Mode of S_n, $n = 2^s \cdot B$

	$n = B$	$n = 2 \cdot B$	$n = 2^2 \cdot B$	$n = 2^3 \cdot B$	$n = 2^4 \cdot B$
$B = 25$ with $\lfloor \log_2 25 \rfloor = 4$					
$mod_n = \lfloor \log_2 n \rfloor$	4	5	6	7	8
$\mathrm{P}(S_n = \lfloor \log_2 n \rfloor)$	0.2980	0.2768	0.2645	0.2574	0.2534
$\mathrm{P}(S_n > \lfloor \log_2 n \rfloor)$	0.5496	0.5441	0.5423	0.5419	0.5418
$B = 30$ with $\lfloor \log_2 30 \rfloor = 4$					
$mod_n = \lfloor \log_2 n \rfloor$	4	5	6	7	8
$\mathrm{P}(S_n = \lfloor \log_2 n \rfloor)$	0.2743	0.2601	0.2511	0.2456	0.2425
$\mathrm{P}(S_n > \lfloor \log_2 n \rfloor)$	0.6255	0.6160	0.6119	0.6100	0.6091
$B = 35$ with $\lfloor \log_2 35 \rfloor = 5$					
$mod_n = \lfloor \log_2 n \rfloor$	5	6	7	8	9
$\mathrm{P}(S_n = \lfloor \log_2 n \rfloor)$	0.2784	0.2627	0.2542	0.2495	0.2469
$\mathrm{P}(S_n > \lfloor \log_2 n \rfloor)$	0.4102	0.4139	0.4167	0.4185	0.4196

mod_n denotes the finite sample mode of S_n; the probabilities build on Corollary 1

extensions have been provided by Gordon et al. (1986, Thm. 1), see also Proposition 3 below. For this and the next section, remember the definition of the fractional part of a real number x, $\{x\} := x - \lfloor x \rfloor$, with $\lfloor \cdot \rfloor$ being the usual floor function.

Proposition 2 *Under the assumptions of Proposition 1 it holds uniformly for any integer z that*

$$P(Z_n - \lfloor \log_2 n \rfloor < z) = F_n(z) + o(1)\,,$$

where $F_n(z) := \exp\left(-2^{-(z+1-\{\log_2 n\})}\right)$.

Proof Földes (1979, Thm. 4). □

Now we are equipped to turn to an approximation of S_n with $S_n \approx Z_n + 1$ building on $\mathrm{P}(S_n = k) \approx \mathrm{P}(Z_n = k - 1)$ for large n according to (6). The distribution of Z_n can be approximated by truncating a Gumbel distribution with distribution function F_n. Let V_n be Gumbel distributed with parameters $\{\log_2 n\} - 1$ and $1/\ln 2$ such that

$$\mathrm{E}(V_n) = \{\log_2 n\} - 1 + \frac{\gamma}{\ln 2} \quad \text{and} \quad \mathrm{Var}(V_n) = \frac{\pi^2}{6}\frac{1}{\ln^2 2}\,,$$

where $\gamma \approx 0.5772$ is Euler's constant. It is known that $F_n(v) = \mathrm{P}(V_n \le v)$, $v \in \mathbb{R}$, with F_n given in Proposition 2, the mode is $\mathrm{mod}(V_n) = \{\log_2 n\} - 1$, i.e. the density $f_n(v)$ is maximized at $\mathrm{mod}(V_n)$, and the median is $\mathrm{med}(V_n) = \mathrm{mod}(V_n) - \ln(\ln 2)/\ln 2$. We then have by Proposition 2 that $Z_n - \lfloor \log_2 n \rfloor \approx \lfloor V_n \rfloor$ in the sense that

$$\mathrm{P}(Z_n - \lfloor \log_2 n \rfloor \le z - 1) \approx \mathrm{P}(V_n \le z) = \mathrm{P}(\lfloor V_n \rfloor \le z - 1).$$

Consequently,

$$S_n \approx \lfloor V_n \rfloor + \lfloor \log_2 n \rfloor + 1 \,. \tag{7}$$

Because of

$$\mathrm{P}(\lfloor V_n \rfloor = z - 1) = \mathrm{P}(z - 1 \leq V_n < z) = \int_{z-1}^{z} f_n(v) \mathrm{d}v \,,$$

the mode $\mathrm{mod}(V_n)$ with $-1 < \mathrm{mod}(V_n) < 0$ suggests that $\mathrm{mod}(\lfloor V_n \rfloor) = -1$. Hence, (7) suggests that $\mathrm{mod}(S_n) = \lfloor \log_2 n \rfloor$, which was observed in Table 3.

Remark 2 Note that the approximation in (7) builds on the convergence result in Proposition 2, which, however, may not be interpreted as a limiting distribution: The approximating random variable V_n with the distribution function F_n does not converge with n, simply because the fractional part $0 \leq \{\log_2 n\} < 1$ does not.

More loosely speaking, it follows from Proposition 2 that $(k = 1, 2, \ldots)$[3]

$$\mathrm{P}(S_n \leq k) \approx \mathrm{P}(Z_n < k) \approx \exp\left(-2^{-(k+1-\log_2 n)}\right) = F_n(k - \lfloor \log_2 n \rfloor) \,. \tag{8}$$

Hence, $\mathrm{P}(S_n = k)$ can be approximated by $P_n(k)$ defined as follows:

$$\mathrm{P}(S_n = k) \approx P_n(k) := F_n(k - \lfloor \log_2 n \rfloor) - F_n(k - 1 - \lfloor \log_2 n \rfloor) \,. \tag{9}$$

As in Table 3, consider $n = 2^s \cdot B$ such that $\lfloor \log_2 n \rfloor = s + \lfloor \log_2 B \rfloor$ with $\{\log_2(2^s \cdot B)\} = \{\log_2 B\}$. Obviously, $P_n(\lfloor \log_2(2^s \cdot B) \rfloor)$ is constant for all s,

$$P_n(\lfloor \log_2(2^s \cdot B) \rfloor) = \exp\left(-2^{\{\log_2 B\}-1}\right) - \exp\left(-2^{\{\log_2 B\}}\right) .$$

Since $\{\log_2 B\} \in [0, 1)$, it is straightforward to verify that $P_n(\lfloor \log_2(2^s \cdot B) \rfloor)$ varies only between 0.233 and 0.250, which explains $\mathrm{P}(S_n = \lfloor \log_2 n \rfloor)$ in Table 3, in particular $P_n(\lfloor \log_2(2^s \cdot 25) \rfloor) = 0.2482$, $P_n(\lfloor \log_2(2^s \cdot 30) \rfloor) = 0.2383$, and $P_n(\lfloor \log_2(2^s \cdot 35) \rfloor) = 0.2438$. Similarly,

$$\mathrm{P}(S_n > \lfloor \log_2 \left(2^s \cdot B\right) \rfloor) \approx 1 - F_n(0) = 1 - \exp\left(-2^{\{\log_2 B\}-1}\right) .$$

Again, for $B \in \{25, 30, 35\}$ this very well explains the figures from Table 3 since

$$\mathrm{P}(S_n > \lfloor \log_2 \left(2^s \cdot B\right) \rfloor) \approx \begin{cases} 0.5422 \text{ for } B = 25 \\ 0.6084 \text{ for } B = 30 \\ 0.4212 \text{ for } B = 35 \end{cases} .$$

[3] Note that this is only a pragmatic approximation since Prop. 2 does not guarantee that $\mathrm{P}(Z_n < k) = \mathrm{P}(Z_n - \lfloor \log_2 n \rfloor < k - \lfloor \log_2 n \rfloor)$ equals $\exp\left(-2^{-(k+1-\log_2 n)}\right) + o(1)$ since $k - \lfloor \log_2 n \rfloor$ is not a finite integer z for growing n.

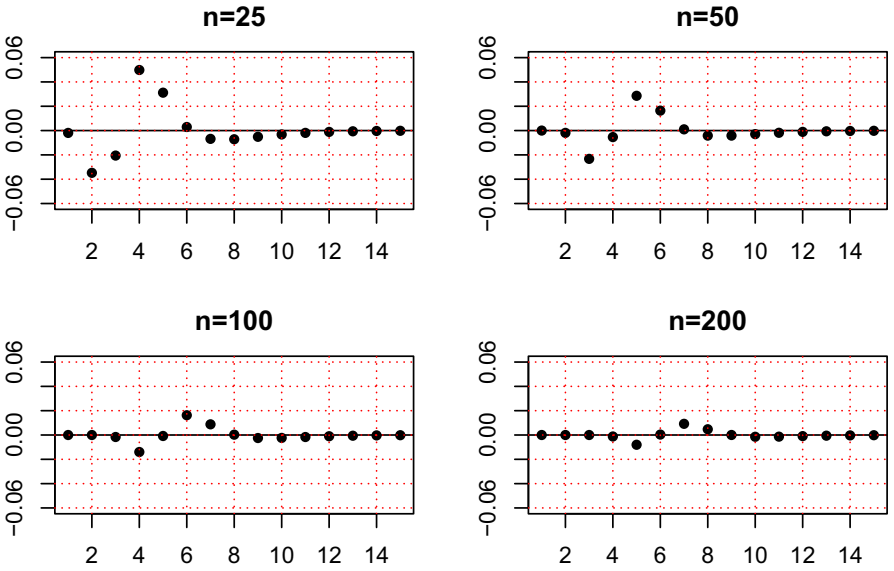

Fig. 2 $\mathrm{P}(S_n = k) - P_n(k)$, see (9), $k = 1, \ldots, 15$

Further, Fig. 2 displays selected differences of the exact and the approximate probabilities, $\mathrm{P}(S_n = k) - P_n(k)$: (9) does a fairly good job in approximating the single exact probabilities from Corollary 1 for $n \geq 100$, while for $n = 25$ or $n = 50$ the deviations may be considerable.

Using $\mathrm{E}(V_n)$ and $\mathrm{Var}(V_n)$, we could roughly approximate $\mathrm{E}(S_n)$ and $\mathrm{Var}(S_n)$, but more elaborate results are available from the literature. Because of (6) we have

$$\mathrm{E}(S_n) = \sum_{k=1}^{n} k\mathrm{P}(S_n = k) = \sum_{k=0}^{n-1}(k+1)\mathrm{P}(Z_{n-1} = k) = \mathrm{E}(Z_{n-1}) + 1\,.$$

Gordon et al. (1986, Thm. 2) provided $\mathrm{E}(Z_n) \approx \log_2 n + \gamma/\ln 2 - 3/2$. It follows that

$$\mathrm{E}(S_n) \approx \mu_n := \log_2 n + \frac{\gamma}{\ln 2} - \frac{1}{2}\,. \tag{10}$$

More precisely, one has, see Guibas and Odlyzko (1980, Thm. 4.1), that

$$\mathrm{E}(S_n) = \mu_n + r(n) + o(1)\,,$$

where $r(n)$ does not vanish but is small: $|r(x)| \leq 1.6 \cdot 10^{-6}$ for all x according to Guibas and Odlyzko (1980, p. 245). Due to $r(n)$, S_n does not converge with n even if demeaned by μ_n, see Remark 2. Still, the mean can be very well approximated as the evaluation of (10) demonstrates:

A look at Table 2 demonstrates the close correspondance with the exact expectation even for small n. Finally, Gordon et al. (1986, Thm. 2) provided an approximation of

	$n = 25$	$n = 50$	$n = 100$	$n = 200$	$n = 400$
μ_n	4.9766	5.9766	6.9766	7.9766	8.9766

the variance, too. Since $\mathrm{Var}(S_n) = \mathrm{Var}(Z_n)$ we have from their paper that

$$\mathrm{Var}(S_n) \approx \frac{\pi^2}{6 \ln^2 2} + \frac{1}{12} \approx 3.5070\,,$$

see also Guibas and Odlyzko (1980, Thm. 4.1). This value independent of n does not explain well the exact variances for small n given in Table 2.

6 Correlated random walks

Drawing from the paper by Gordon et al. (1986) we briefly consider an extensions of Proposition 2. We now relax Assumption 1 and allow for correlation between the random walks. In terms of the concordance from Definition 1, correlation allows for $\mathrm{P}(C_i = 0) \neq \mathrm{P}(C_i = 1)$. Technically, this means we have a Bernoulli process without symmetry, which is the model for tossing a coin that is not fair. The stronger the positive correlation between the two random walks is, the larger is the probability of concordance p,

$$p := \mathrm{P}(C_i = 0) \quad \text{and} \quad q := 1 - p = \mathrm{P}(C_i = 1).$$

Negative correlation implies $p < 1/2$.

Assumption 2 *Let $(\Delta X_i, \Delta Y_i)_{i=1,\ldots,n}$ be a sequence of independent, identically distributed and continuous random variables with $0 < p < 1$.*

From Gordon et al. (1986, Thm. 1) we have the following result, see also Arratia, Gordon and Waterman (1990, Coro. 3).

Proposition 3 *Let $(X_i, Y_i)_{i=0,1,\ldots,n}$ from equation (1) satisfy Assumption 2. It then holds uniformly for any integer z that*

$$P(Z_n - \lfloor m_{n,p} \rfloor < z) = \exp\left(-p^{z-\{m_{n,p}\}}\right) + o(1)\,,$$

where $m_{n,p} := \log_{1/p}(nq)$.

Proof The result follows from Gordon et al. (1986, Thm. 1); details are provided in the Appendix. □

Note that $m_{n,1/2} = \log_2(n) - 1$ and $\{\log_2(n) - 1\} = \{\log_2 n\}$, such that Proposition 2 arises as a special case. Further, Proposition 3 allows to approximate in the sense of (8) that

$$\mathrm{P}(Z_n \leq k) \approx \exp\left(-p^{k+1-m_{n,p}}\right)\,. \tag{11}$$

Table 4 Probabilities $P(Z_n > \lfloor \log_2 n \rfloor)$ for varying p

n	$n = 25$	$n = 50$	$n = 100$	$n = 200$	$n = 400$
$\lfloor \log_2 n \rfloor$	4	5	6	7	8
Approximate					
$p = 0.23$	0.0123	0.0057	0.0026	0.0012	0.0006
$p = 0.42$	0.1726	0.1472	0.1252	0.1062	0.0900
$p = 0.50$	0.3234	0.3234	0.3234	0.3234	0.3234
$p = 0.58$	0.4980	0.5504	0.6044	0.6590	0.7129
$p = 0.77$	0.7891	0.9090	0.9751	0.9966	0.9998
Monte Carlo estimates					
$p = 0.23$	0.0111	0.0054	0.0027	0.0010	0.0004
$p = 0.42$	0.1574	0.1375	0.1207	0.1022	0.0893
$p = 0.50$	0.3139	0.3161	0.3174	0.3185	0.3224
$p = 0.58$	0.5166	0.5633	0.6135	0.6667	0.7183
$p = 0.77$	0.9297	0.9766	0.9955	0.9997	1.0000

Approximate probabilities from (11); Monte Carlo estimation builds on computer experiments with 10^5 replications

This formula underlies Table 4 dedicated to the effect of p on $P(Z_n > \lfloor \log_2 n \rfloor)$ building on the approximation from (11). Our choices of p equal the probabilities if $(\Delta X_i, \Delta Y_i)$ are jointly normal with a correlation of ρ: $p = 0.23, 0.42, 0.5, 0.58, 0.77$ arise from $\rho = -0.75, -0.25, 0, 0.25, 0.75$. It is intuitively clear that $p > 0.5$ increases the probabilites of long zero runs, and it does so dramatically e. g. for $p = 0.77$. For $p < 0.5$ on the other hand, zero runs become less likely because the random walks tend to drift in a discordant manner. This does of course not reduce the correlation between the random walks. Let O_n stand for the length of the longest sequence of ones in $(C_i)_{i=1,\ldots,n}$. It is clear from (11) that

$$P(O_n \leq k) \approx \exp\left(-q^{k+1-\mu_{n,q}}\right),$$

where $\mu_{n,q}$ is defined analogously to $m_{n,p}$ from Proposition 3: $\mu_{n,q} := \log_{1/q}(np)$. Table 4 formalizes the following intuition: The stronger the correlation between the random walks is, i. e. the larger $|p - 0.5|$ is, the more likely are long runs of zeros or ones in (C_i), depending on the sign of $p - 0.5$. The approximate results from the first panel of Table 4 are well supported by finite sample Monte Carlo estimates for p being not too large; for $p = 0.77$, however, the approximate figures are too conservative in that the Monte Carlo estimates are sizeable larger.

7 Summary

There exists a well understood asymptotic theory why one gets nonsensical correlation between independent long random walks, see Phillips (1986, Thm. 1). In this note we

focus on finite samples with a special interest on small sizes. What is, for instance, the maximum length of random association (consecutive concordance or consecutive discordance) between two independent random walks of sample size $n = 50$? Evaluating Corollary 1, one can verify that the probability of the maximum length of random association being equal to 5 amounts to 27.68% (see also Fig. 1). The exact probability that this maximum length is at least equal to 5 amounts to 82.09% (see Table 3), and the expected value is 5.9783 (Table 2). Hence, long episodes (relative to the small sample size) of random association occur frequently, which explains why nonsense correlation arises between independent short random walks. We also included the case of correlated random walks where long episodes of association are of course more likely, see Table 4 for a quantification.

Acknowledgements Open Access funding enabled and organized by Projekt DEAL.

Appendix

Let $(B_i)_{i=1,\ldots,n}$ denote a sequence of independent Bernoulli trials with equal probability: $\mathrm{P}(B_i = 1) = \mathrm{P}(B_i = 0) = 1/2$. We call (B_i) a Bernoulli-Laplace [BL] process, which is the mathematical model for tossing a fair coin. Let $\mathcal{B}_n$ be the set of all possible BL sequences of length n, such that $\#(\mathcal{B}_n) = 2^n$, where $\#(\mathcal{S})$ denotes the number of elements of some set $\mathcal{S}$. By Assumption 1, the concordance indicators $(C_i)_{i=1,\ldots,n}$ from (3) form a BL process.

Proof of Proposition 1

By definition, Z_n stands for the length of the longest zero run of a BL sequence. To determine its probability distribution, we define the set $\mathcal{Z}_n(k) \subseteq \mathcal{B}_n$ containing all sequences subject to $Z_n < k$. By definition, $\mathrm{P}(Z_n < k) = \#(\mathcal{Z}_n(k))/2^n$. For brevity, let $N_n^{(k)} := \#(\mathcal{Z}_n(k))$:

$$\mathrm{P}(Z_n < k) = \frac{N_n^{(k)}}{2^n}.$$

Hence, we are left with determining $N_n^{(k)}$. Obviously, $N_n^{(1)} = 1$, since $\mathcal{Z}_n(1) = \{(1, 1, \ldots, 1)\}$. All sequences contained in $\mathcal{Z}_n(2)$ necessarily begin with '1' or with '0,1' glued to sequences from $\mathcal{Z}_{n-1}(2)$ and $\mathcal{Z}_{n-2}(2)$, respectively. Consequently, $N_n^{(2)} = N_{n-1}^{(2)} + N_{n-2}^{(2)}$. Analogously, all elements in $\mathcal{Z}_n(3)$ are made up by '1', '0,1'

or '0,0,1' followed by sequences from $\mathcal{Z}_{n-1}(3)$, $\mathcal{Z}_{n-2}(3)$ and $\mathcal{Z}_{n-3}(3)$, respectively, and so on. The general recursion becomes:

$$N_n^{(k)} = \sum_{i=1}^{k} N_{n-i}^{(k)}\,, \quad 1 \le k \le n\,. \tag{12}$$

To initialize this recursion, one requires starting values $N_m^{(k)}$ for $m < k$. In this case, all zero runs are necessarily shorter than k, i.e. all elements in $\mathcal{B}_m$ satisfy $Z_m < k$, such that

$$N_m^{(k)} = 2^m\,, \quad 0 \le m < k\,, \tag{13}$$

which formally covers the case $N_0^{(k)} = 1$, too. By Definition 2 it holds that $N_m^{(\ell)} = f_{m+1}^{(\ell)}$, $m = 0, 1, \ldots$, which completes the proof.

Proof of Corollary 1

By definition, S_n is the maximum length of a spell (of zeros or ones). Denote by $\mathcal{S}_n(k)$ the subset of $\mathcal{B}_n$ meeting $S_n < k$. Obviously, $\mathcal{S}_n(1)$ equals the empty set. Generally, a new spell begins exactly when C_{i+1} differs from C_i. Define the corresponding difference indicator

$$D_i = \begin{cases} 1 \text{ if } C_{i+1} \neq C_i \\ 0 \text{ if } C_{i+1} = C_i \end{cases}\,, \quad i = 1, \ldots, n-1\,.$$

By construction, $\mathrm{P}(D_i = 1) = \mathrm{P}(D_i = 0) = 1/2$, and $(D_i)_{i=1,\ldots,n-1}$ is a new BL process. Further, a zero run of length $k-1$ in (D_i) is equivalent to a run of zeros or a run of ones of length k in (C_i). Therefore, $\#(\mathcal{S}_n(k)) = 2\,\#(\mathcal{Z}_{n-1}(k-1))$, and with the previous notation this mean that

$$\#(\mathcal{S}_n(k)) = 2\,N_{n-1}^{(k-1)} = 2\,f_n^{(k-1)}\,, \quad 2 \le k \le n\,. \tag{14}$$

With $\mathrm{P}(S_n < k) = \#(\mathcal{S}_n(k))/2^n$, the proof is complete.

Proof of Proposition 3

Define as in Gordon et al. (1986) $V_{n,p} := W/\ln(1/p) + \{m_{n,p}\}$, where W follows a standard Gumbel distribution; note that $V_{n,1/2} = V_n$ from Sect. 5. The corresponding distribution function is known to become

$$\begin{aligned} F_{n,p}(v) &:= \mathrm{P}(V_{n,p} \le v) = \exp\left(-\mathrm{e}^{-(v-\{m_{n,p}\})\ln(1/p)}\right) \\ &= \exp\left(-p^{v-\{m_{n,p}\}}\right)\,. \end{aligned}$$

From Gordon et al. (1986, Thm. 1) we have that uniformly in z

$$\mathrm{P}\left(Z_n - m_{n,p} \le z - 1\right) = \mathrm{P}\left(\lfloor V_{n,p} \rfloor - \{m_{n,p}\} \le z - 1\right) + o(1)\,, \tag{15}$$

see also Arratia, Gordon and Waterman (1990, Coro. 3). Using

$$\mathrm{P}\left(Z_n - m_{n,p} \le z - 1 - \{m_{n,p}\}\right) = \mathrm{P}\left(Z_n - \lfloor m_{n,p} \rfloor \le z - 1\right)$$

together with

$$\mathrm{P}\left(\lfloor V_{n,p} \rfloor \le z - 1\right) = \mathrm{P}\left(V_{n,p} < z\right) = F_{n,p}(z)\,,$$

the claim follows.

References

Arratia R, Gordon L, Waterman MS (1990) The Erdös-Rényi law in distribution, for coin tossing and sequence matching. Ann Stat 18:539–570

Ernst PA, Shepp LA, Wyner AJ (2017) Yule's "nonsense correlation" solved! Ann Stat 45:1789–1809

Földes A (1975) On the limit distribution of the longest heads (in Hungarian). Matematikai Lapok 26:105–116

Földes A (1979) The limit distribution of the length of the longest head-run. Period Math Hung 10:301–310

Gordon L, Schilling MF, Waterman MS (1986) An extreme value theory of long head runs. Probab Theory Relat Fields 72:279–287

Granger CWJ, Newbold P (1974) Spurious regressions in econometrics. J Econ 2:111–120

Guibas LJ, Odlyzko AM (1980) Long repetitive patterns in random sequences. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 53:241–262

Hassler U (2003) Nonsense regressions due to neglected time-varying means. Stat Pap 44:169–182

Palm FC, Sneek JM (1984) Significance tests and spurious correlation in regression models with autocorrelated errors. Stat Pap 25:87–105

Phillips PCB (1986) Understanding spurious regressions in econometrics. J Econ 33:311–340

Révész P (1990) Regularities and irregularities in a random 0, 1 sequence. Stat Pap 31:95–101

Révész P (2013) Random walk in random and non-random environments, 3rd edn. World Scientific Publishing, Singapore

Spickerman WR (1982) Binet's formula for the tribonacci sequence. Fibonacci Q 20:118–120

Spickerman WR, Joyner RN (1984) Binet's formula for the recursive sequence of order K. Fibonacci Q 22:327–331

Székely G, Tusnády G (1976-1979) Generalized fibonacci numbers and the number of "pure heads" (in Hungarian). Matematikai Lapok 27:147–151

Yule GU (1926) Why do we sometimes get nonsense-correlations between time-series? A study in sampling and the nature of time-series. J R Stat Soc 89:1–63