CORRELATION ANALYSIS OF EXCEL'S RAND() FUNCTION USED FOR SHUFFLING DATA

B. Hughes, 2017 August 18

1. INTRODUCTION

In the course of determining the statistics associated with the results of Klondike Solitaire[1], the question of the suitability of online random number generators arose regarding their use in online Solitaire games, particularly for producing the shuffling process.

One online random number generator with common and easy access is Excel's RAND() function and its subsidiary RANDBETWEEN. Even though it is highly unlikely that RAND() would be used for Klondike Solitaire, its ready access prompted a desire to examine the correlational properties that existed between the "cards" in a given shuffle when the shuffle was produced using a sort process that started with Excel's RAND(). As well, the correlational properties between different shuffles (but for the same "card" position), and the correlations between the full card distributions from shuffle-to-shuffle, have been examined to determine whether the "shuffles" and the "card" locations within a "shuffle" were indeed statistically independent in a measurable way. The results of this examination not only provide conclusions that pertain to RAND() quite separately from Klondike Solitaire, but it is likely that they are indicative of possibilities for other random number generators as well.

All computations were carried out using Excel and three different randomization processes: (i) a direct use of RANDBETWEEN for the numbers 1 to 52, simulating a deck of cards in a straightforward way; (ii) a use of the MOD function (i.e. modulus) on RAND() multiplied by 10^9 to simulate a closer result to the Growly Solitaire software application[2] (although that software uses ISAAC rather than RAND() for the basic random number generation process); and (iii) a double modulus process that used the randomization of the MOD process multiplied by 10^9 and subjected once more to Excel's MOD function, as in (ii). In using Excel, the "calculation" preference was set to Manual with only 1 iteration for the part of the process that creates the random sets of data. A semi-automatic arrangement was used to produce the number of sets needed for the sample averages, and it was tested to ensure that Excel wasn't proceeding too quickly for the calculations to be performed appropriately.

2. GENERATING A SHUFFLE

(a) RANDBETWEEN: Because successive calls to RANDBETWEEN will only produce random values between specified integers, a series of calls can result in any given integer being produced several times. To produce a *set* of random values in which each value occurs once

and only once, a special process is necessary. In order to facilitate this the Fisher-Yates[3] process was followed.

Firstly, 52 values were produced ranging between 1 and 52 using RANDBETWEEN for each but with each successive call using an interval reduced by one:

$$X_1 = RANDBETWEEN(1, 52) \quad (1a)$$
$$X_2 = RANDBETWEEN(1, 51) \quad (1b)$$
$$\ldots\ldots\ldots\ldots$$
$$X_n = RANDBETWEEN(1, 52 - [n-1]) \quad (1c)$$
$$\ldots\ldots\ldots\ldots$$
$$X_{51} = RANDBETWEEN(1, 2) \quad (1d)$$
$$X_{52} = 1 \quad (1e)$$

Secondly, from these values, the first integer (representing a "card") was selected by choosing, *and removing*, the number in the X_1^{th} location in the set (1,2,3,…,52) and closing up the set of the remaining 51 values. The second integer was selected by choosing, *and removing*, the number in the X_2^{th} location in the remaining set of 51 values and closing up the set of the remaining 50 values, and this process was repeated until all 52 integers were selected. The result of this is a set of 52 ostensibly randomly-ordered integer values between 1 and 52 inclusive with each integer occurring once and only once.

(b) MOD: For this process, the only difference from RANDBETWEEN is in the very first portion when the initial 52 random values are created. The function RANDBETWEEN is replaced by (i) RAND() which when called produces a real value ≥ 0 and < 1 from a "uniform" distribution over that interval, and (ii) the product of RAND() and 10^9 (Excel's precision is basically 15 figures), and (iii) the modulus of that over an interval of 52 (for the first one), and then (iv) that converted to an integer with 1 added. As happened in the RANDBETWEEN formulation, the interval used for the modulus is decremented by one for each successive call until all 52 are created. The same sort process is then used to obtain a set of 52 "randomly-ordered" integers with each integer occurring once and only once. The initial calls are as follows:

$$X_1 = INT\{ MOD[RAND() \times 10^9, 52]\} + 1 \quad (2a)$$
$$X_2 = INT\{ MOD[RAND() \times 10^9, 51]\} + 1 \quad (2b)$$
$$\ldots\ldots\ldots\ldots$$
$$X_n = INT\{ MOD[RAND() \times 10^9, 52 - (n-1)]\} + 1 \quad (2c)$$
$$\ldots\ldots\ldots\ldots$$
$$X_{51} = INT\{ MOD[RAND() \times 10^9, 2]\} + 1 \quad (2d)$$
$$X_{52} = 1. \quad (2e)$$

Here, $MOD(x, N)$ is defined as

$$MOD(x, N) = x - N\{ \text{Integer Part of}(x/N)\} \quad (3)$$

therefore the integer part of $MOD(x, N)$ itself is an integer between 0 and $N - 1$, and the set $\{X_n\}$ comprises integers between 1 and 52.

The main reason for using the MOD approach was to produce a separate randomizing process that was less dependent on Excel's random number generator. A secondary reason was to have a more direct comparison to Growly Solitaire's process[3] for generating a "shuffled" deck of cards.

(c) DOUBLE MODULUS: Again the sort process is the same as in the previous two methods, but the initial 52 random values are produced similarly to the MOD method but with RAND() replaced as follows:

$$X_1 = INT\{ MOD[MOD(RAND() \times 10^9, 997) \times 10^9, 52]\} + 1 \quad (4a)$$
$$X_2 = INT\{ MOD[MOD(RAND() \times 10^9, 997) \times 10^9, 51]\} + 1 \quad (4b)$$
$$\ldots\ldots\ldots\ldots$$
$$X_n = INT\{ MOD[MOD(RAND() \times 10^9, 997) \times 10^9, 52 - (n-1)]\} + 1 \quad (4c)$$
$$\ldots\ldots\ldots\ldots$$
$$X_{51} = INT\{ MOD[MOD(RAND() \times 10^9, 997) \times 10^9, 2]\} + 1 \quad (4d)$$
$$X_{52} = 1 \quad (4e)$$

In these formulae, the term 997 was chosen because it is the largest prime number less than 1000 and it was considered that this would produce a suitable random remainder.

All three randomizing processes were tested – without including the sorting process – to determine whether they produced uniformly distributed random variables. For each, 2×10^5 integer samples were generated spanning the range 1 to 52, and histograms of these were produced. This was repeated several times to obtain different sets of the random variables. The Kolmogorov-Smirnov non-parametric test was used[4] to compare the histograms to a theoretical one that was uniform, and with better than 90% confidence the generated histograms compared sufficiently well that none could be failed. The standard deviation of the variation in uniformity across the 52 values in the histograms was typically 1.65% of the average of the histogram's values. This is very close to the theoretical expectation (see Appendix A for derivation) which is 1.6% if the 2×10^5 random variables are statistically independent and uniformly distributed.

3. TESTING THE RANDOMNESS BY COVARIANCES.

There are 3 different covariance protocols that have been used: "within a shuffle" in which autocovariances are formed using the set of 52 values; "inline" in which the autocovariances are formed from a set of shuffles but for each location in the 52 separately; and "cross"-covariances that are formed using the entire set of 52 but cross-correlated with subsequent shuffles.

The basic definition of the covariance $C(n)$ for the set of 52 random variables is

$$C(n) = \frac{1}{52}\sum_{i=1}^{52} r(i)\, r(i+n) \tag{5}$$

but with suitable variations for each of the 3 covariance protocols, particularly the "inline" one. Here i refers to the location in the set and

$$r(i) = R(i) - 26.5 \tag{6}$$

with

$$R(i) = \text{set of integers } \{1,2 \ldots 52\} \text{ in randomized order} \tag{7}$$

Thus, the term $r(i)$ refers to the i^{th} location in the result of the shuffle and the value of $r(i)$ is the value in that location – the value of the "card" – and it is a random variable as defined above. It can be seen from Eq. (6) that the average of $r(i)$ over the set of 52 as used in the covariance $C(n)$ is zero so there is no need to subtract it off in the covariance definition in Eq. (5) and to reduce the divisor to 51 to obtain a non-biased covariance.

This can be generalized to a set of integers from 1 to M so that Eq. (5) becomes

$$C(n) = \frac{1}{M}\sum_{i=1}^{M} r(i)\, r(i+n) \tag{5a}$$

with the factor 26.5 in Eq. (6) replaced by $(M+1)/2$. It can be further generalized in notation to refer to each shuffle specifically by

$$C_j(n) = \frac{1}{M}\sum_{i=1}^{M} r_j(i)\, r_j(i+n) \tag{5b}$$

where the subscript j refers to the results of the j^{th} shuffle. The following sections will specialize the above equations and definitions to each of the 3 covariance protocols.

In all of the analysis that follows, it will be assumed that each $r(i)$ comprises a set of M random variables for which each realization is "i.i.d", that is, statistically *independent* from all other sets and each set being *identically distributed*. It will also be the case that within each set of M random values, each value will occur once and only once in accordance with the sort process described above.

(a) *Within a Shuffle*: With any of the above randomizing techniques, a set of 52 randomly-ordered integers is produced in which each integer from 1 to 52 occurs once and only once. The average of this set, 26.5, is subtracted from each value so the average of the resulting set of 52 is zero. Autocovariances have been calculated on each zero-average set with a repeating

extension beyond the 52 values, and because of the repeating extension the autocovariances for lags 0 to 25 are identical to those with lags from 27 to 52 but in reverse order. Here the autocovariance is defined as follows:

$$C_j(n) = \frac{1}{52}\sum_{i=1}^{52}{}' r_j(i)\, r_j(i+n) \tag{6}$$

where the ' on the summation sign indicates the repeating extension, so that $i+n$ is treated as $MOD(i+n, 52)$ where MOD is defined in Eq. (3).

On repeating this with different initial sets of the 52 integers and averaging the resulting autocovariances, an approximation to a statistical expectation is obtained. If arrangements of the 52 integers render them uncorrelated the average of the autocovariances should tend towards a δ-function located at lag zero. However, the sorting process does have some residual order inherent in it because when 51 of the values are put in their randomly-selected order, there is no choice regarding the 52nd-value's location. The result of this is an expectation of the autocovariance for lag ≠ 0 or 52 that is non-zero. The sum of the autocovariances of the zero-average data-sets over all lags from 1 to 52 must be zero because the sum of r over all locations is zero by construct:

$$\sum_{n=1}^{52}{}' C_j(n) = \frac{1}{52}\sum_{i=1}^{52}{}' r_j(i) \sum_{n=1}^{52}{}' r_j(i+n) = 0 \tag{7}$$

The *expectation* of $C_j(n)$, however, is non-zero for all n and all j, and this can be shown as follows. In Eq. (7) the last sum on the right is zero by construct, so that sum taken from 1 to 51 is equal to $-r_j(i+52)$, but by definition of the ' sign, this is the same as $-r_j(i)$. By reducing the upper limit in the sum on n in Eq. (7) to 51 and using this result,

$$\sum_{n=1}^{51} C_j(n) = -\frac{1}{52}\sum_{i=1}^{52} r_j^2(i) \tag{8}$$

and because the sum on the right side of this spans all locations (and thus all integers in the set) once and only once

$$\sum_{n=1}^{51} C_j(n) = -\frac{1}{52}\sum_{i=1}^{52} r_j^2(i) = -\frac{1}{52}\sum_{i=1}^{52}(i-26.5)^2 \tag{9}$$

In more general terms,

$$\sum_{n=1}^{M-1} C_j(n) = -\frac{1}{M}\sum_{i=1}^{M}\left(i - \frac{M+1}{2}\right)^2 = -\frac{M^2-1}{12} \tag{10}$$

(For the present case in which $M = 52$, the numerical value of this is -225.25.) From the general definition of $C_j(n)$ in Eq. (6), but generalized by replacing '52' by M

$$\langle C_j(n)\rangle = \frac{1}{M}\sum_{i=1}^{M}{}' \langle r_j(i) r_j(i+n)\rangle \tag{11}$$

In Appendix B it is shown that $\langle C_j(n)\rangle$ is independent of n for $n \neq 0$ or M and that it is given by

$$\langle C_j(n)\rangle = -\frac{M^2-1}{12(M-1)} \tag{12}$$

For $n = 0$ or M, $\langle C_j(n)\rangle$ can be obtained directly from Eq. (11) recognizing that $\langle r_j^2(i)\rangle$ is the same as the right side of Eqs.(8), (9) and (10) but with the opposite sign.

$$\langle C_j(0)\rangle = \langle C_j(M)\rangle = \frac{M^2-1}{12} \tag{13}$$

for all j. Finally, Eqs. (12) and (13) can be combined, and for all n

$$\langle C_j(n)\rangle = \left(\frac{M^2-1}{12}\right)\begin{cases} 1 & n = 0, M \\ -\dfrac{1}{M-1} & 1 \le n < M \end{cases} \tag{14}$$

independent of j.

To complete the normalization and convert the covariances into correlations, $C(n)$ can be divided by $C(0)$ and this removes the $(M^2-1)/12$ factor on the right side of Eq. (14):

$$\langle \text{Autocorrelation}\rangle = \begin{cases} 1 & n = 0, M \\ -1/(M-1) & 1 \le n < M \end{cases} \tag{15}$$

The numerical results do show these patterns (for all randomizing models), i.e. 1 or -1/51, and examples are given in Figures 1a and 1b.

The average of the autocovariances over different realizations should tend to zero inversely as $\sqrt{N}$ when the constant is subtracted off, and indeed this is the case for all randomization models as shown in Figures 2a (RANDBETWEEN), 2b (MOD) and 2c (Double Modulus). The black line in each of these is the expectation of the root-mean-square (rms) value of the averaged autocovariance (minus the constant) where the rms is taken over all lag-values minus one, i.e.,

51 values. The constant, λ_W, that defines the black line in each can be obtained by assuming each set of 52 randomly-ordered integer-based values is independent of each other set.

It is shown in Appendix C that as M increases

$$\lambda_W \to \frac{(M+1)\sqrt{M-2}}{12} \tag{16}$$

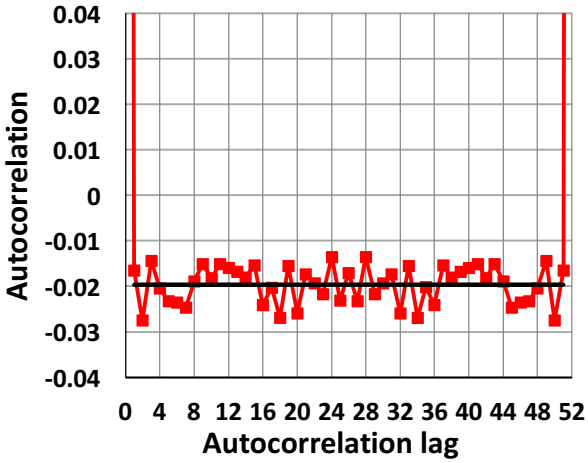

Figure 1a. RANDBETWEEN

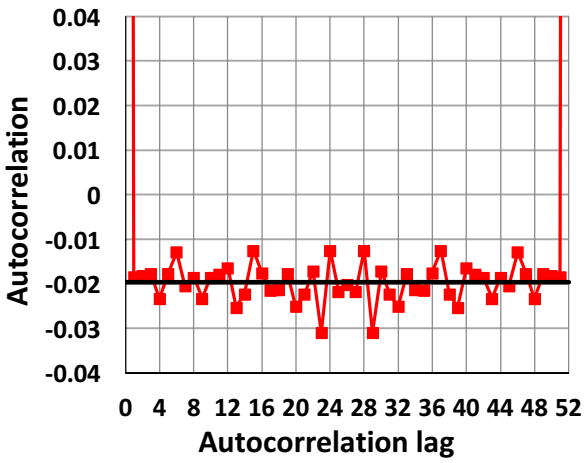

Figure 1b. Modulus

The black lines show the expectation (-0.0196) and the red lines show the measured values using an average over 1000 samples. The expectation and measured values at lag = 0 and 52 are 1. The autocorrelation values show a strong tendency towards the expectation, and the mirrored values for lags 27-51 compared to 1-25 are quite apparent (due to the repeating extension in taking the covariances). The Double Modulus autocorrelations show very similar results.

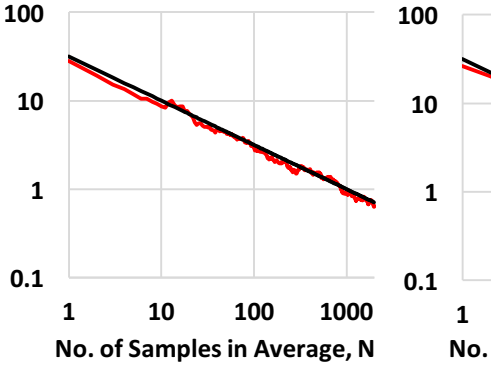

Figure 2a. RANDBETWEEN

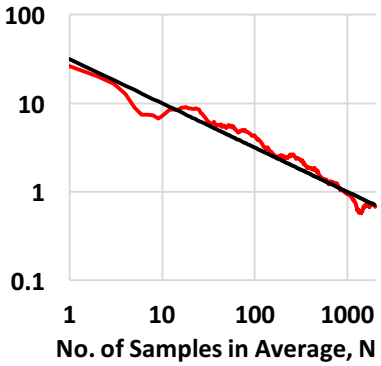

Figure 2b. Modulus

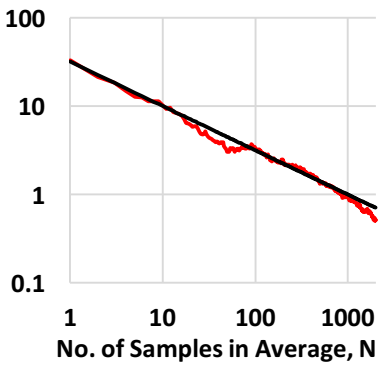

Figure 2c. Double Modulus

The black lines show the expectation $31.588/\sqrt{N}$ and the red lines show the root-mean-squares over lags 1 to 51 of the covariances averaged over a range of 1 to 2000 samples.

where the subscript W signifies "Within a shuffle". In the present case M is 52, and the more exact formula in Appendix C gives

$$\lambda_W = 31.588\ldots \tag{17}$$

(b) *Shuffle-to-Shuffle Covariances at a Particular Location in the Set of 52*: In this case the autocovariances are calculated for each location in the 52 integers rather than across the integers themselves, and the dimension along which the autocovariances are calculated is generated by repeatedly producing samples of the sets of 52 randomly-arranged integers using the processes described in section 3. For the present purpose, 2000 sets were produced. For each location, autocovariances with the following lags, n, were calculated for all averaging lengths, N, from 4 to $2000 - n$:

$$n = 1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1000. \tag{21}$$

The rms of these over all locations was also calculated and it is shown for the different randomizing models in Figures 3a, 3b and 3c.

$$rms \text{ autocovariance} \equiv C_{rms}(N,n) = \sqrt{\frac{\left(\sum_{i=1}^{M} C_i^2\,(N,n)\right)}{M}} \tag{22}$$

The autocovariance $C_i(N,j)$ is given in the standard non-biased form with *i* as the location in the set of 52 and $r_j(i)$ as the value in the i^{th} location of the j^{th} sample set (as in the previous section)

$$C_i(N,n) = \frac{1}{N-1}\sum_{k=1}^{N}\{r_k\,(i) - \rho_o(N,i)\}\{r_{k+n}(i) - \rho_n(N,i)\}) \tag{23}$$

where ρ is the average over the N samples

$$\rho_q(N,i) = \frac{1}{N}\sum_{l=1}^{N} r_{l+q}(i) \tag{24}$$

and q = 0 for the first average and q = n for the second.

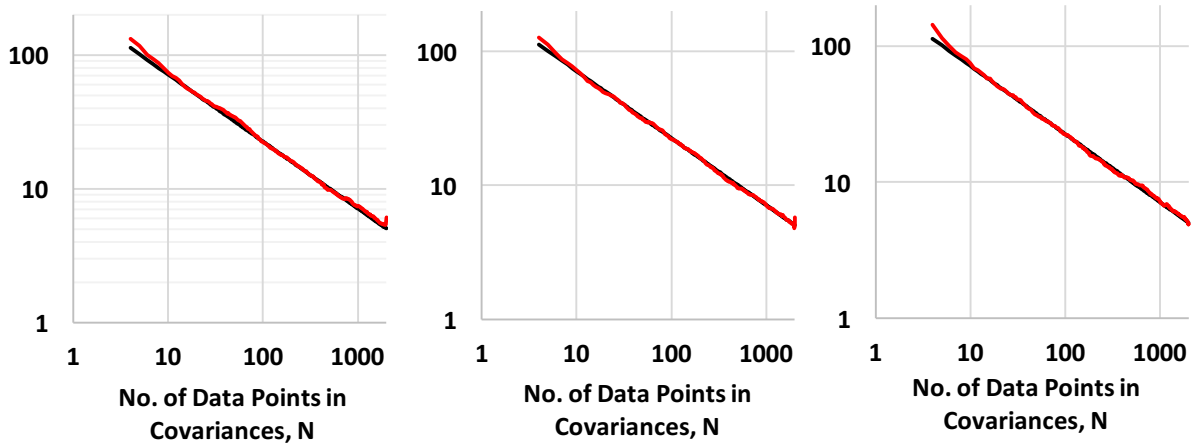

Figure 3a. RANDBETWEEN Figure 3b. Modulus Figure 3c. Double Modulus

The black lines show the asymptotic expectation $225.25/\sqrt{N}$ and the red lines show the inline root-mean-squares over locations 1 to 52 of the autocovariances with their numbers of data points ranging from 4 to 2000 samples. The resulting comparisons between each measured line and the expectation line are very close for all randomizing processes.

As in the case of averages within a shuffle these also tend to zero as $\lambda_I/\sqrt{N}$. (The subscript I signifies an "Inline" covariance.) In Appendix D the general form for λ_I is determined to be $(M^2-1)/12$ for large N, which results in

$$C_{rms}(N,n) \to \frac{(M^2-1)}{12\sqrt{N}} \tag{25}$$

for all n as $N \to \infty$.

For the present case In which M is 52

$$C_{rms}(N,n) \to \frac{225.25}{\sqrt{N}} \tag{26}$$

(c) *Determining Cross-Covariances of the Full Set of 52 with Subsequent Sets:* In this case the covariance between a set of 52 and a subsequent set of 52 is determined for various lags, where the lag is the number of sets from one set to the other. For a lag of 1, the covariance is determined between adjacent sets. The procedure was to produce 2000 sets of shuffled integers (with averages of 26.5 subtracted from each value in the each data set) and to form the covariance between set_1 and set_{n+1}, the covariance between set_2 and set_{n+2}, etc. until all $2000-n$ covariances were computed. This was carried out for the set of 11 lags, n, which is given in Eq. (21).

For this method covariances were calculated as in Eq. (23) but with no need to allow for the subtraction of the averages because they are all zero by construct. A simple unbiased estimator is appropriate for the same reason:

$$C_k(n) = \frac{1}{M}\sum_{i=1}^{M} r_k\,(i) r_{k+n}(i) \tag{27}$$

One of the chosen statistics, defined as $S(N)$, used to display some of the characteristics of $C_k(n)$ is the rms over all 11 lags each of which comprises running averages of $C_k(n)$, so that

$$S^2(N) = \frac{1}{11}\sum_{n}\left\{\frac{1}{N}\sum_{k=1}^{N} C_k(n)\right\}^2 \tag{28}$$

It is shown in Appendix E that if the samples are statistically independent $S(N)$ will decrease as $\lambda_X/\sqrt{N}$, which is the square-root of the expectation of $S^2(N)$. (The subscript X on λ signifies a "Cross" covariance.) Examples of $S(N)$ are shown in Figures (4a)-(4c).

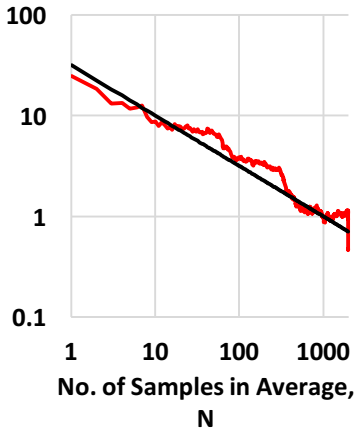

Figure 4a. RANDBETWEEN

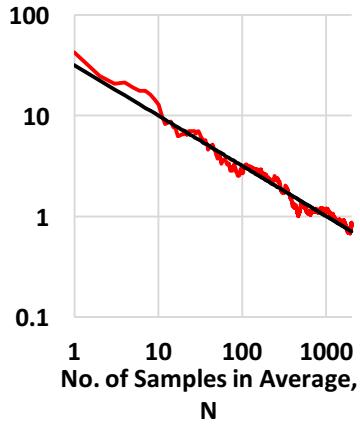

Figure 4b. Modulus

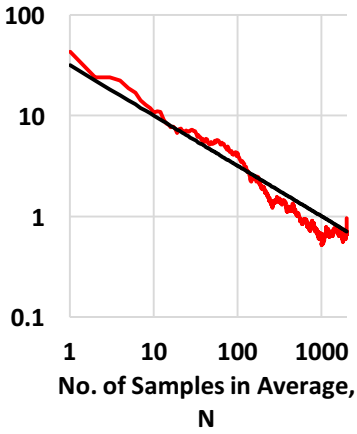

Figure 4c. Double Modulus

The black lines show the expectation $31.54/\sqrt{N}$ and the red lines show the cross-covariances with their averages ranging from 1 to 2000 samples, and each averaged over all of the chosen lags (11 values). Much more variation between each measured line and the expectation line is apparent here compared to the other protocols (Figures 2a-2c and 3a-3c) although the trends are very similar to the expectation.

In Appendix E the general form for λ_X is determined to be $(M+1)\sqrt{M-1}/12$ which results in

$$\frac{\lambda_X}{\sqrt{N}} = \frac{(M+1)\sqrt{M-1}}{12\sqrt{N}} \tag{29}$$

For the present case In which M is 52

$$\lambda_X = 31.54\ldots \tag{30}$$

4. THE SIGNIFICANCE OF THE COVARIANCES.

(a) *Within a Shuffle*: Significance tests were carried out on the 2000 covariance values using the Kolmogorov-Smirnov (KS) testing method[4]. The 2000 covariances for each location were first tested for normality using the Lilliefors test[5], and all three randomization methods had 3 or less fails out of 25 (the number of lags in the test). Even though 3 out of 25 is 12% (i.e. slightly more than 10% for 2 of the methods), they were all treated as if they were Gaussian. The theoretical model in the KS comparison was a Gaussian that used the theoretically determined mean (μ) and standard deviation (λ_W). Because the sets were extended in a circular manner, the number of lags with different covariances is 25: lags 27-51 duplicate lags 1-25, and the 26th doubles up on the cross-products and so its statistics are different from the others and were not used. Of the 25 the number in the KS test that fell outside the 90% level were counted for each of the 3 randomization process, and are shown in Table 1. If the proportion of the KS probabilities that are outside the 90% level is greater than 10%, the covariance cannot be considered to be zero with 90% confidence. It can be seen that the RANDBETWEEN process shows the most results outside the 90% probability, 2 out of 25.

CIRCULAR "WITHIN" (LAGS 1-25)	Number Probably NOT μ-Mean Gaussian	Out of	Should be < ~10%	Number of fails in Lags 1-4
RANDBETWEEN	2	25	OK	0
Modulus	1	25	OK	0
Double Modulus	0	25	OK	0

Table 1. The second column shows the number of cases of the 2000-sample autocovariance averages for the circularly extended samples within a "shuffle" for which the mean falls outside the 90% probability level compared to the expected mean μ using the Kolmogorov-Smirnov test and the assumption that the averages are normally distributed. The third column shows the number of cases tested (lags 1-25). Because of the 90% probability level, approximately 10% could possibly fail the test and that comparison is shown in column 4. Column 5 shows which lag fails in the low-numbered lags.

(b) *Inline — Shuffle to Shuffle at a Particular Location*: Significance tests on the covariances calculated with $2000 - n$ values in the covariance sum were again determined using the KS test. This was performed assuming the 52 locations were independent and that the covariances presented a Gaussian distribution over these 52 locations. The covariances were converted by the Fisher z-transform using the theoretical standard deviation of λ_I from section 3(b) and Appendix D, with the large-N value of 225.25, and the transformed variables were then tested against a Gaussian with a mean of zero and standard deviation of $1/\sqrt{2000 - lag - 3}$. The results are shown in Table 2 which gives the number of lags in which the KS probability is greater than 90% that the tested data is different from the zero-mean Gaussian, which is taken to indicate that the covariances do not show appropriate independence from location-to-location. Only the RANDBETWEEN randomization process presents a proportion of probabilities that are larger than 10% namely 3 out of 11 cases.

"INLINE" (11 LAGS)	Number Probably NOT zero-Mean Gaussian	Out of	Should be < ~10%	Number of fails in Lags 1-4
RANDBETWEEN	3	11	Fail	lag 4
Modulus	0	11	OK	0
Double Modulus	0	11	OK	0

Table 2. The second column shows the number of cases of the 2000-sample autocovariance averages for the samples at a particular location in the set (i.e. inline) for which the mean falls outside the 90% probability level compared to the expected value of zero using the Kolmogorov-Smirnov test. The third column shows the number of cases tested (lags 1-26). Because of the 90% probability level, approximately 10% could fail the test and that comparison is shown in column 4. Only RANDBETWEEN fails the test. Column 5 shows which lag fails in the low-numbered lags.

(c) *Cross — Shuffle to Shuffle*: The covariances were converted to correlations by dividing each by the variance of the r's which is the same for all covariances and is given by $(M^2-1)/12$, i.e. 225.25. The correlations were transformed into Fisher z-scores which were also multiplied by $\sqrt{M-3}$ to normalize them to a unit standard deviation, and these were tested against a Gaussian distribution with zero mean and unit standard deviation for each of the 11 chosen lags. The number with a KS probability greater than 90% for the differences from the Gaussian distribution were counted and are shown in Table 3 for each of the 3 randomizing methods. It is expected that about 1 out of the 11 will fail the test but it can be seen that RANDBETWEEN shows about 3 times that much, calling it into question.

"CROSS" (11 LAGS)	Number Probably NOT zero-Mean Gaussian	Out of	Should be < ~10%	Number of fails in Lags 1-4
RANDBETWEEN	3	11	Fail	lag 4
Modulus	0	11	OK	0
Double Modulus	0	11	OK	0

Table 3. The second column shows the number of cases of the 2000-sample averages for the shuffle-to-shuffle cross-covariances for which the mean falls outside the 90% probability level compared to the expected value of zero using the Kolmogorov-Smirnov test. The third column shows the number of cases tested. Because of the 90% probability level, approximately 10% should fail the test and that comparison is shown in column 4. Only RANDBETWEEN fails the test. Column 5 shows which lag fails in the low-numbered lags.

5. SUMMARY AND CONCLUSIONS

a) All randomizing processes produce random variables that have distributions that are in close agreement with statistical expectation for uniformity, and they show no "fails" to 90% confidence when compared to a theoretical uniform distribution.

b) All randomizing processes and covariance protocols display the expected $\lambda/\sqrt{N}$ behaviour as the number N of samples in the covariance estimate or the average of the

covariance estimates increases, and the constant λ for each protocol is in approximate agreement with the expected value.

c) The covariances display the approximately δ-function behaviour expected for independent samples for the autocovariances within a "shuffle".

d) The KS tests of significance show that the overall number of "fails" at 90% confidence is different for RANDBETWEEN than for the other methods, particularly as is seen in Table 4.

e) From this analysis, the sorting process and the *modulus-based* randomizing processes are shown to be acceptable, with 90% confidence, in achieving the desired result of statistically independent "shuffling", not only from shuffle-to-shuffle but also from location-to-location within a shuffle. The *RANDBETWEEN* process is suspect in this regard.

OVERALL NUMBER OF FAILS		Out of	Percentage failed
RANDBETWEEN	8	47	17%
Modulus	1	47	2%
Double Modulus	0	47	0%

Table 4. The first column shows the number of cases from all covariance protocols (within, inline and cross) which fail at the 90% probability level compared to expectations, using the Kolmogorov-Smirnov test. The second column shows the number of cases tested. Because of the 90% probability level, approximately 10% could fail the test and that comparison is shown in column 3. Only RANDBETWEEN falls outside the 10% level.

Given that the results here imply that Excel's RAND() is an adequate starting point to create random sets of data for solitaire games at least, any better random number generator being used as a starting point is more likely to produce statistically separate "shuffles" or statistically separate cards within a shuffle. For the most part RANDBETWEEN also displays some adequacy for producing the random numbers before sorting,

6. REFERENCES

1. Hughes, B. A., *Probabilities For Klondike Solitaire* , 2017 May 11
2. Mason, Chris, *email communication*, Growlybird Software, 2017 April 27
3. For a good description of this method see *https://en.wikipedia.org/wiki/Fisher–Yates_shuffle*
4. Zaiontz, Charles, *Real-Statistics, http://www.real-statistics.com/tests-normality-and-symmetry/statistical-tests-normality-symmetry/kolmogorov-smirnov-test/kolmogorov-distribution/*

5. Zaiontz, Charles, *Real-Statistics,* http://www.real-statistics.com/tests-normality-and-symmetry/statistical-tests-normality-symmetry/lilliefors-test-normality/
6. Wikipedia, *https://en.wikipedia.org/wiki/Binomial_distribution - Variance*

Appendix A

Determination of Histogram Standard Deviation

In the main body of the report random variables are produced as integers in the range 1 to M where M itself is in the range from 2 to 52. As a test of the uniformity of the resulting integer values, for each of the randomizing processes 200,000 values were produced with $M = 52$, and the histograms of the resulting 200,000 values were produced with bins equal to the integers 1 to M. An example of such a histogram is shown in Figure (A1). As one of the measures of uniformity, standard deviations of the histogram values $H(m)$, $1 \leq m \leq M$ were computed over the M values and then divided by the average over the same range.

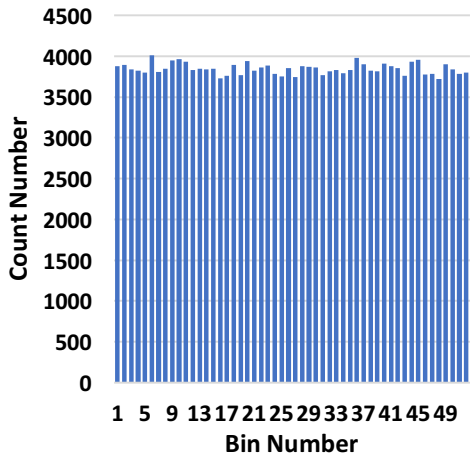

Figure A1. Histogram of 200,000 random samples generated by the Modulus process with 52 as the modulus parameter. The standard deviation of the counts across all Bins is 1.72% of the average count number 3846.1.. ($2x10^5/52$)

The statistical expectation of the result can be determined assuming that the random variables are integer values in the range 1 to M and all are statistically independent and drawn from a uniform distribution.

In this determination

$$H(m) = \sum_{j=1}^{N} \delta_{mR_j} \qquad \text{(A1)}$$

where δ_{mR_j} is the Kronecker delta – unity if the two subscripts are the same and zero if they are not – and R_j is the j^{th} random variable R, and R is a random integer in the range 1 to M. It can be noted that the sum of $H(m)$ over all m is simply the number of data points N, and the average of $H(m)$ over all m is N/M.

This is a sum of N independent random binomial variables and the variance $\sigma^2(m)$ of such a sum is given by[6]

$$\sigma^2(m) = NQ(m)\big(1 - Q(m)\big) \qquad \text{(A2)}$$

where $Q(m)$ is the probability that a random variable will fall in the m^{th} bin. Because there are M bins and all are equally likely,

$$Q(m) = \frac{1}{M} \qquad \text{(A3)}$$

and

$$\sigma^2(m) = \frac{N(M-1)}{M^2} \tag{A4}$$

As already noted the average of $H(m)$, $\overline{H(m)}$, is N/M and so the standard deviation $\sigma(m)$ divided by the average is

$$\frac{\sigma(m)}{\overline{H(m)}} = \sqrt{\frac{M-1}{N}} \tag{A5}$$

For the values $M = 52$ and $N = 200{,}000$

$$\frac{\sigma(m)}{\overline{H(m)}} = .0159687\ldots \tag{A6}$$

Appendix B

Determination of $\langle r(i)r(i+n)\rangle$

The subscript j on $r(i)$ is suppressed because the probabilities are assumed to be the same for all sample sets. The expectation is defined to be

$$\langle r(i)r(i+n)\rangle \;=\; \sum_{r(i)=1}^{M}\sum_{r(i+n)=1}^{M} r(i)r(i+n)jpdf\{r(i),r(i+n)\} \qquad \text{(B1)}$$

where $jpdf\,(x,y)$ is uniform over the range of x and uniform over the range of y except that one point in y's range is missing. This happens because of the sorting condition that values are to appear once and only once. Thus for whatever value $r(i)$ takes, that value is forbidden to $r(i+n)$, and even if $n>1$ all other values are available to $r(i+n)$ over all the realizations. In the description of the sort process in section 2 of the report, if a random value for $r(i)$ is obtained that value is removed from the remaining values to be sorted, i.e. $r(i+n)$ is chosen from random values that do not contain that value but can contain all others. With this in mind the $jpdf$ can be constructed using a δ-function:

$$jpdf\{k,l\} = \frac{1-\delta_{kl}}{M(M-1)} \qquad \text{(B2)}$$

where M is the range of k and because l is reduced In range by one point, $M-1$ is the range of l.

$$\langle r(i)r(i+n)\rangle \;=\; \frac{1}{M(M-1)}\sum_{k=1}^{M}\sum_{l=1,l\neq k}^{M}\left(k-\frac{(M+1)}{2}\right)\left(l-\frac{(M+1)}{2}\right) \qquad \text{(B3)}$$

Break the second summation into 2 parts, $1\le l\le k-1$, and $k+1\le l\le M$. Then Eq. (B3) becomes

$$\begin{aligned}\langle r(i)r(i+n)\rangle &\\ &= \frac{1}{M(M-1)}\sum_{k=1}^{M}\left(k-\frac{(M+1)}{2}\right)\left\{\sum_{l=1}^{k-1}\left(l-\frac{(M+1)}{2}\right)\right.\\ &\left.+\sum_{l=k+1}^{M}\left(l-\frac{(M+1)}{2}\right)\right\} \qquad \text{(B4)}\end{aligned}$$

The last summation can be written in two parts again, $1\le l\le M$ and a subtracted sum over $1\le l\le k$:

$$\langle r(i)r(i+n)\rangle = \frac{1}{M(M-1)}\sum_{k=1}^{M}\left(k-\frac{(M+1)}{2}\right)\left\{\sum_{l=1}^{k-1}\left(l-\frac{(M+1)}{2}\right)+\sum_{l=1}^{M}\left(l-\frac{(M+1)}{2}\right)-\sum_{l=1}^{k}\left(l-\frac{(M+1)}{2}\right)\right\}$$

The second-to-last sum returns zero, and the third-to-last sum cancels all but the $l = k$ value in the last sum. This simplifies everything to

$$\langle r(i)r(i+n)\rangle = -\frac{1}{M(M-1)}\sum_{k=1}^{M}\left(k-\frac{(M+1)}{2}\right)^2 \quad \text{(B5)}$$

which by Eq. (12) in the main body of the report is

$$\langle r(i)r(i+n)\rangle = -\frac{M^2-1}{12(M-1)} = -\frac{M+1}{12} \quad \text{(B6)}$$

This is independent of n, so the sum in Eq. (11) simply cancels the $1/M$ portion and returns Eq. (12).

Appendix C

Adding All the Terms to $\lambda_W = \sqrt{M-2}\,\dfrac{(M+1)}{12}$

The analysis in this Appendix pertains to the covariance protocol "Within a Shuffle". From the definition of the covariance $C_j(n)$ as given in Eq. (5b), the constant λ_W in Eq. (18) is defined as the expectation of $\sqrt{N}$ times the root-mean-square (*rms*) value of the averaged autocovariance (minus its average) where the *rms* is taken over all lag-values minus one, i.e., $M-1$ values:

$$\frac{\lambda_W^2}{N} = \left\langle \frac{1}{M-1}\sum_{n=1}^{M-1}\left\{\frac{1}{N}\sum_{j=1}^{N}(C_j(n)-\mu)\right\}^2 \right\rangle \tag{C1}$$

where

$$\mu = \frac{1}{M-1}\sum_{n=1}^{M-1} C_j(n) = -\frac{M+1}{12} \tag{C2}$$

The right side of Eq. (C1) can be expanded to produce

$$\frac{1}{M-1}\sum_{n=1}^{M-1}\left\{\frac{1}{N}\sum_{j=1}^{N}(C_j(n)-\mu)\right\}^2 = \frac{1}{M-1}\sum_{n=1}^{M-1}\left\{\left\{\frac{1}{N}\sum_{j=1}^{N}C_j(n)\right\}^2 - \frac{2\mu}{N}\sum_{j=1}^{N}C_j(n)+\mu^2\right\} \tag{C3}$$

On taking the sum on n inside the large braces on the right and using Eq. (C2),

$$\frac{1}{M-1}\sum_{n=1}^{M-1}\left\{\frac{1}{N}\sum_{j=1}^{N}(C_j(n)-\mu)\right\}^2 = \frac{1}{M-1}\sum_{n=1}^{M-1}\left\{\frac{1}{N}\sum_{j=1}^{N}C_j(n)\right\}^2 - \mu^2 \tag{C4}$$

$$\frac{1}{M-1}\sum_{n=1}^{M-1}\left\{\frac{1}{N}\sum_{j=1}^{N}(C_j(n)-\mu)\right\}^2 = \frac{1}{(M-1)N^2}\sum_{n=1}^{M-1}\left\{\sum_{j=1}^{N}\sum_{k=1}^{N}C_j(n)C_k(n)\right\} - \mu^2 \tag{C5}$$

By replacing the product $C_j(n)C_k(n)$ by its definition in terms of the r's and taking the expectation,

$$\langle C_j(n)C_k(n)\rangle = \frac{1}{M^2}\sum_{i=1}^{M}{}'\sum_{l=1}^{M}{}'\langle r_j(i)r_j(i+n)\rangle\langle r_k(l)r_k(l+n)\rangle \qquad \text{(C6)}$$

which from Eq. (C1) provides

$$\frac{\lambda_W^2}{N} = \frac{1}{(M-1)N^2}\sum_{n=1}^{M-1}\left\{\sum_{j=1}^{N}\sum_{k=1}^{N}\frac{1}{M^2}\sum_{i=1}^{M}{}'\sum_{l=1}^{M}{}'\langle r_j(i)r_j(i+n)\rangle\langle r_k(l)r_k(l+n)\rangle\right\} - \mu^2 \qquad \text{(C7a)}$$

and where the dashes (') on the sums over i and l signify that the r's are to be extended in a repeating fashion (see section 3a).

The sums on j and k can be carried out by interchanging the order of the sums on k, j with the sums on l, i and removing $k = j$ from the sum on k and adding it back in as a separate term. The expectation in the sum on k, j now comprises two independent terms – the product of the j-subscripted terms and the k-subscripted terms, and this separates the 4-fold expectation into the product of two independent ones :

$$\frac{\lambda_W^2}{N} = \frac{1}{(M-1)M^2}\sum_{n=1}^{M-1}\left\{\sum_{l=1}^{M}{}'\sum_{i=1}^{M}{}'\frac{1}{N^2}\sum_{j=1}^{N}\sum_{k=1}^{N}\langle r_j(i)r_j(i+n)\rangle\langle r_k(l)r_k(l+n)\rangle\right\} - \mu^2 \qquad \text{(C7b)}$$

$$\frac{1}{N^2}\sum_{j=1}^{N}\sum_{k=1}^{N}\langle r_j(i)r_j(i+n)r_k(l)r_k(l+n)\rangle$$

$$= \frac{1}{N^2}\sum_{j=1}^{N}\langle r_j(i)r_j(i+n)r_j(l)r_j(l+n)\rangle$$

$$+ \frac{1}{N^2}\sum_{j=1}^{N}\sum_{k=1}^{N-1}\langle r_j(i)r_j(i+n)\rangle\langle r_k(l)r_k(l+n)\rangle \qquad \text{(C7c)}$$

Note that the upper limit on the remaining sum on k has been reduced by 1.

In the last term on the right each expectation is given by Eqs. (B5) and (B6) in Appendix B, and the definition of μ in Eq. (C2), and each is independent of i, n, j and k. The expectation in the first term on the right is independent of the subscript j and so the sum simply returns N. The same applies to the second term and the sums return $N(N-1)$. Because the subscript j is no

longer relevant, all r's will not have a subscript but will be recognized as being from the same statistical sample. With these changes

$$\frac{1}{N^2}\sum_{j=1}^{N}\sum_{k=1}^{N}\langle r_j(i)r_j(i+n)r_k(l)r_k(l+n)\rangle = \frac{\langle r(i)r(i+n)r(l)r(l+n)\rangle}{N} + \frac{(N-1)\mu^2}{N} \qquad \text{(C7d)}$$

and Eq. (C7b) becomes

$$\frac{\lambda_W^2}{N} = \frac{1}{(M-1)M^2}\sum_{n=1}^{M-1}\left\{\sum_{l=1}^{M\ '}\sum_{i=1}^{M\ '}\frac{\langle r(i)r(i+n)r(l)r(l+n)\rangle}{N} + \frac{(N-1)\mu^2}{N}\right\} - \mu^2 \qquad \text{(C7e)}$$

$$\frac{\lambda_W^2}{N} = \frac{1}{N(M-1)M^2}\sum_{n=1}^{M-1}\left\{\sum_{l=1}^{M\ '}\sum_{i=1}^{M\ '}\langle r(i)r(i+n)r(l)r(l+n)\rangle\right\} - \frac{\mu^2}{N} \qquad \text{(C7f)}$$

(In passing it can be noted that λ_W^2 is the variance of $C_j(n)$.) The expectation of the r's will devolve into a linear combination of the two terms $\langle r^2(i)\rangle^2$ and $\langle r^4(i)\rangle$ where the expectation $\langle r^2(i)\rangle^2$ is given by

$$\langle r^2(i)\rangle^2 = \left\{\frac{1}{M}\sum_{\tau=1}^{M}\left(\tau - \frac{M+1}{2}\right)^2\right\}^2 = (M-1)^2\mu^2 = \frac{(M^2-1)^2}{12^2} \qquad (C8a)$$

and, with $\varphi \equiv \langle r^4(i)\rangle/[M(M-1)^2]$,

$$\langle r^4(i)\rangle = \frac{1}{M}\sum_{\tau=1}^{M}\left(\tau - \frac{M+1}{2}\right)^4 = M(M-1)^2\varphi = \left(\frac{(M^2-1)^2}{12^2}\right)\left(\frac{9M^2-21}{5M^2-5}\right) \qquad \text{(C8b)}$$

The $jpdf$ of the r's is needed to carry out the expectation in Eq. (Cf7) and to determine λ_W^2.

Underlying all the $jpdf$'s is the basic pdf of each of the r's, which is uniform over the range $-M/2$ to $+M/2$. In practice this will be shifted so the range is from 1 to M with the statistical variable shifted by its mean to compensate:

$$pdf\{r(i)\} = pdf\left\{x - \frac{M+1}{2}\right\} = \frac{H\{x\}H\{M+1-x\}}{M} \qquad \text{(C9)}$$

where H is the Heavyside operator and is 1 if its argument is positive and 0 if it is negative or 0, and x is an integer such that $1 \le x \le M$. From this the $jpdf$'s and the expectations can be obtained.

There are 6 cases that arise, and even though $r(i)$ is never the same random variable as $r(i+n)$, they are correlated (Appendix B) so they must be treated specially in the $jpdf$. The same applies to $r(l)$ and $r(l+n)$. The $jpdf$ is written as $jpdf(x,y,z,t)$ where x, y, z and t are more convenient ways of writing $r(i), r(i+n), r(l)$, and $r(l+n)$ respectively. It will be even more convenient to write these as $\bar{x}, \bar{y}, \bar{z}, \bar{t}$ at times where the overbar indicates that the mean $(M+1)/2$ has been subtracted off but the variable's sums are from 1 to M.

With respect to the sums on i and l, the special cases are all included in the following:

(i) $l \neq i$, and $l \neq i-n$, and $l \neq i+n$, i.e. none of x, y, z, t are equal |
(ii) $l = i$, i.e. $x = z$, $y = t$ and $x \neq y$ |
(iii) $l = i-n, n \neq M/2$, i.e. $x = t$ and $y \neq x, z$ and $x \neq z$ (C10)
(iv) $l = i+n, n \neq M/2$, i.e. $x = y$ and $z \neq x, t$ and $t \neq x$ |
(v) $l = i-n, n = M/2$, i.e. $x = t$, $y = z$ and $x \neq z$ |
(vi) $l = i+n, n = M/2$, i.e. $x = y$, $z = t$ and $x \neq z$ |

The conditions in (v) and (vi) only pertain if M is even. If M is odd, they don't occur.

The overall expectation of the r's can be separated into a sum over these 6 special cases with each special case having suitable δ–functions associated with it to select the relevant variables. The δ-functions will simplify the summations appropriately. Commensurate with the 6 cases, the expectations, E, are as follows:

(i) $E_1 = \sum_{x,y,z,t=1}^{M} \bar{x}\bar{y}\bar{z}\bar{t}(1-\delta_{xy})(1-\delta_{zt})(1-\delta_{xz})(1-\delta_{yt})(1-\delta_{xt})(1-\delta_{yz})/Norm1$ |
(ii) $E_2 = \sum_{x,y=1}^{M} \bar{x}^2\bar{y}^2(1-\delta_{xy})/Norm2$ |
(iii) $E_3 = \sum_{x,y,z=1}^{M} \bar{x}^2\bar{y}\bar{z}(1-\delta_{xy})(1-\delta_{xz})(1-\delta_{yz})/Norm3$ |
(iv) $E_4 = \sum_{x,z,t=1}^{M} \bar{x}^2\bar{z}\bar{t}(1-\delta_{xz})(1-\delta_{xt})(1-\delta_{zt})/Norm3$ (C11)
(v) $E_5 = \sum_{x,z=1}^{M} \bar{x}^2\bar{z}^2(1-\delta_{xz})/Norm2$ |
(vi) $E_6 = \sum_{x,z=1}^{M} \bar{x}^2\bar{z}^2(1-\delta_{xz})/Norm2$ |

It can be seen that $E_3 = E_4$ and $E_5 = E_6$. It can also be seen that $E_5$and $E_6$are special cases of E_2 (for $n = M/2$, and if M is odd they both are zero). With this notation, the total expectation of the r's is

$$\langle r(i)r(i+n)r(l)r(l+n)\rangle = \left\{ (1-\delta_{li})(1-\delta_{l(i-n)})(1-\delta_{l(i+n)})E_1 + \delta_{li}E_2 + 2\delta_{l(i-n)}\left(1-\delta_{n\frac{M}{2}}\right)E_3 + 2\left(\delta_{l(i-n)}\delta_{n\frac{M}{2}}\right)E_2 \right\} \quad \text{(C12)}$$

which requires the determination of the 3 expectations. These will be obtained in order of the number of variables in the $jpdf$ argument.

(A) 2-variable case – $\overline{x}^2\overline{y}^2$ for E_2, E_5 and E_6

$$jpdf(x,y) = \frac{(1-\delta_{xy})}{Norm2} \tag{C13}$$

which gives

$$Norm2 = \sum_{x=1}^{M}\sum_{y=1}^{M}(1-\delta_{xy}) = M(M-1) \tag{C14}$$

The expectations are given by

$$\frac{1}{M(M-1)}\sum_{x=1}^{M}\sum_{y=1}^{M}\left(x-\frac{(M+1)}{2}\right)^2\left(y-\frac{(M+1)}{2}\right)^2(1-\delta_{xy})$$

$$= \frac{1}{M(M-1)}\left[\left\{\sum_{x=1}^{M}\left(x-\frac{M+1}{2}\right)^2\right\}^2 - \sum_{x=1}^{M}\left(x-\frac{M+1}{2}\right)^4\right]$$

which by Eqs. (C8a) and (C8b) are

$$E_2 = E_5 = E_6 = M(M-1)\,(\mu^2-\varphi) \tag{C15}$$

(C) 3-variable case — $\overline{x}^2\overline{y}\overline{z}$ for E_3 and E_4

$$jpdf(x,y,z) = \frac{(1-\delta_{xy})(1-\delta_{xz})(1-\delta_{yz})}{Norm3} \tag{C16}$$

which gives

$$Norm3 = \sum_{z=1}^{M}\sum_{y=1}^{M}\sum_{x=1}^{M}(1-\delta_{xy})(1-\delta_{xz})(1-\delta_{yz}) = M(M-1)(M-2) \tag{C17}$$

From this the expectations are given by

$$E_3 = E_4 = -\frac{1}{M(M-1)(M-2)}\left[\left\{\sum_{x=1}^{M}\left(x-\frac{M+1}{2}\right)^2\right\}^2 - 2\sum_{x=1}^{M}\left(x-\frac{M+1}{2}\right)^4\right] \quad \text{(C18)}$$

and so

$$E_4 = E_5 = -\frac{M(M-1)(\mu^2-2\varphi)}{(M-2)} \quad \text{(C19)}$$

(C) 4-variable case — $\overline{xyzt}$ for E_1

$$jpdf(x,y,z,t) = \frac{(1-\delta_{xy})(1-\delta_{zt})(1-\delta_{xz})(1-\delta_{yt})(1-\delta_{xt})(1-\delta_{yz})}{Norm1}$$

$$= \frac{(1-\delta_{xy}-\delta_{xz}+\delta_{xy}\delta_{xz})(1-\delta_{xt})(1-\delta_{yz}-\delta_{yt}+\delta_{yz}\delta_{yt})(1-\delta_{zt})}{Norm1}$$

$$= \frac{1}{Norm1}(1-\delta_{xy}-\delta_{xz}-\delta_{xt}+\delta_{xy}\delta_{xz}+\delta_{xy}\delta_{xt}+\delta_{xt}\delta_{xz}-\delta_{xy}\delta_{xz}\delta_{xt})\times$$

$$(1-\delta_{yz}-\delta_{yt}-\delta_{zt}+\delta_{yz}\delta_{yt}+\delta_{zt}\delta_{yz}+\delta_{zt}\delta_{yt}-\delta_{zt}\delta_{yz}\delta_{yt}) \quad \text{(C20)}$$

The term on the right when expanded produces 64 separate terms and will not be written out in full here. Its sum over all x, y, z and t provides the expression for $Norm1$

$$Norm1 = \sum_{x,y,z,t=1}^{M} jpdf(x,y,z,t) = M^4 - 6M^3 + 11M^2 - 6M$$

which simplifies to

$$Norm1 = M(M-1)(M-2)(M-3) \quad \text{(C21)}$$

The expectation E_1 becomes

$$E_1 = \frac{1}{M(M-1)(M-2)(M-3)}\sum_{x,y,z,t=1}^{M}\left(x-\frac{M+1}{2}\right)\left(y-\frac{M+1}{2}\right)\left(z-\frac{M+1}{2}\right)\left(t-\frac{M+1}{2}\right) jpdf(x,y,z,t)$$

where the sum is

$$\sum_{x,y,z,t=1}^{M}\left(x-\frac{M+1}{2}\right)\left(y-\frac{M+1}{2}\right)\left(z-\frac{M+1}{2}\right)\left(t-\frac{M+1}{2}\right)(1-\delta_{xy})(1-\delta_{xz})(1-\delta_{xt})(1-\delta_{yz})(1-\delta_{yt})(1-\delta_{zt})$$

The final result of all 4 summations is

$$=\frac{M^2(M-1)^2}{M(M-1)(M-2)(M-3)}(3\mu^2-6\varphi) \tag{C22}$$

which gives

$$E_1=\frac{M(M-1)(3\mu^2-6\varphi)}{(M-2)(M-3)} \tag{C23}$$

It can be seen that all expectations E_1 through E_6 are functions only of M and specifically not functions of i,j,k,l or n. With this in mind the expectation of the r's can be substituted into Eq. (C12) and the sums on i,j,k,l and n can be carried out. Before doing this it is helpful to expand $(1-\delta_{li})(1-\delta_{l(i-n)})(1-\delta_{l(i+n)})$, the multiplier for E_1. The result is $1-\delta_{li}-\delta_{l(i-n)}-\delta_{l(i+n)}$ + products of the δ-functions all of which are mutually exclusive because of their subscripts, and so their sum over l is zero.

$$\begin{aligned}\langle r(i)r(i+n)r(l)r(l+n)\rangle &\\ =\Big\{&(1-\delta_{li}-\delta_{l(i-n)}-\delta_{l(i+n)})E_1+\delta_{li}E_2+2\delta_{l(i-n)}\left(1-\delta_{n\frac{M}{2}}\right)E_3\\ &+2\left(\delta_{l(i-n)}\delta_{n\frac{M}{2}}\right)E_2\Big\}\end{aligned} \tag{C24}$$

With this

$$\frac{\lambda_W^2}{N}=-\frac{\mu^2}{N}+\frac{1}{N}\left\{\left(1-\frac{3}{M}\right)E_1+\left(\frac{1}{M}\right)E_2+\frac{2}{M}\left(1-\frac{\delta_{Me}}{M-1}\right)E_3+\left(\frac{2\delta_{Me}}{M(M-1)}\right)E_2\right\} \tag{C25}$$

or,

$$\frac{\lambda_W^2}{N}=-\frac{\mu^2}{N}+\frac{1}{MN}\left\{(M-3)E_1+E_2+\left(\frac{2M-2(1+\delta_{Me})}{M-1}\right)E_3+\frac{2\delta_{Me}E_2}{(M-1)}\right\} \tag{C26}$$

In terms of $\mu^2$and φ, and after multiplying through by N, this simplifies to

$$\lambda_W^2 = -\mu^2 + \frac{1}{M}\left\{\frac{3M(M-1)(\mu^2-2\varphi)}{(M-2)} + M(M-1)(\mu^2-\varphi) - \left(\frac{2M-2(1+\delta_{Me})}{M-1}\right)\frac{M(M-1)(\mu^2-2\varphi)}{(M-2)} + 2\delta_{Me}M(\mu^2-\varphi)\right\} \quad \text{(C27)}$$

which on further simplification provides

$$\lambda_W = \sqrt{(M-2)}\left(\frac{M+1}{12}\right)\sqrt{1+\frac{6(M-3)(M+1)-2(1-\delta_{Me})(5M^2-4M-13)}{5(M-2)(M-1)(M+1)}} \quad \text{(C28)}$$

It can be noted that the fraction in the large radical on the right is of order $1/M$ and so for the present case of $M = 52$ it is expected to be relatively unimportant numerically.

If M is even

$$\lambda_W = \sqrt{(M-2)}\left(\frac{M+1}{12}\right)\sqrt{1+\frac{6(M-3)}{5(M-2)(M-1)}} \quad \text{(C29)}$$

and if M is odd

$$\lambda_W = \sqrt{(M-2)}\left(\frac{M+1}{12}\right)\sqrt{1-\frac{4(M+2)}{5(M-2)(M+1)}} \quad \text{(C30)}$$

For $M = 52$, from Eq. (C29),

$$\lambda_W = 31.588\ldots \quad \text{(C31)}$$

Appendix D

Derivation of $\lambda_I = (M^2 - 1)/12$

For this Appendix the covariance is defined as

$$C_i(N,n) = \frac{1}{N-1}\sum_{k=1}^{N}\{r_k(i) - \rho_0(N,i)\}\{r_{k+n}(i) - \rho_{k+n}(N,i)\} \quad \text{(D1)}$$

where k refers to the sample, i to the location within the sample and N as the number of samples in the average. The nomenclature here is somewhat different from that in the main body of the report,

$$r_k(i) = R(i) - (M+1)/2 \quad \text{(D2)}$$

$$R(i) = \text{set of integers } \{1,2 \ldots M\} \text{ in randomized order} \quad \text{(D3)}$$

$$\rho_n(N,i) = \frac{1}{N}\sum_{l=1}^{N} r_{l+n}(i) \quad \text{(D4)}$$

and

$$\frac{\lambda_I^2}{N} = \frac{1}{M}\sum_{i=1}^{M}\langle C_i^2(N,n)\rangle \quad \text{(D5)}$$

In Eq, (D1) incorporate the r's in the sums defining the ρ's

$$r_k(i) - \rho_0(N,i) = -\frac{1}{N}\sum_{l=1}^{N} r_l(i)(1 - N\delta_{lk}) \quad \text{(D6a)}$$

$$r_{k+n}(i) - \rho_n(N,i) = -\frac{1}{N}\sum_{m=1}^{N} r_{m+n}(i)(1 - N\delta_{mk}) \quad \text{(D6b)}$$

Then

$$C_i(N,n) = \frac{1}{N^2(N-1)}\sum_{k=1}^{N}\sum_{l=1}^{N}\sum_{m=1}^{N} r_l(i)r_{m+n}(i)(1 - N\delta_{lk})(1 - N\delta_{mk}) \quad \text{(D7)}$$

$$C_i^2(N,n) = \frac{1}{N^4(N-1)^2}\sum_{k=1}^{N}\sum_{l=1}^{N}\sum_{m=1}^{N}\sum_{j=1}^{N}\sum_{p=1}^{N}\sum_{q=1}^{N} r_l(i)r_{m+n}(i)r_p(i)r_{q+n}(i)(1-N\delta_{lk})(1 - N\delta_{mk})\big(1-N\delta_{pj}\big)\big(1-N\delta_{qj}\big) \quad \text{(D8)}$$

By carrying out the sums on k and j, the right side becomes

$$= \frac{1}{N^2(N-1)^2}\sum_{l=1}^{N}\sum_{m=1}^{N}\sum_{p=1}^{N}\sum_{q=1}^{N} r_l(i)r_{m+n}(i)r_p(i)r_{q+n}(i)(1-N\delta_{lm})\big(1-N\delta_{pq}\big) \quad \text{(D9a)}$$

$$= \frac{1}{N^2(N-1)^2}\sum_{l=1}^{N}\sum_{m=1}^{N}\sum_{p=1}^{N}\sum_{q=1}^{N} r_l(i)r_{m+n}(i)r_p(i)r_{q+n}(i)\big(1-N\delta_{lm}-N\delta_{pq}+N^2\delta_{pq}\delta_{lm}\big)$$

and so

$$\begin{aligned} C_i^2(N,n) \;=& \frac{1}{N^2(N-1)^2}\sum_{l=1}^{N}\sum_{m=1}^{N}\sum_{p=1}^{N}\sum_{q=1}^{N} r_l(i)r_{m+n}(i)r_p(i)r_{q+n}(i) \\ &- \frac{2}{N(N-1)^2}\sum_{m=1}^{N}\sum_{p=1}^{N}\sum_{q=1}^{N} r_m(i)r_{m+n}(i)r_p(i)r_{q+n}(i) \\ &+ \frac{1}{(N-1)^2}\sum_{m=1}^{N}\sum_{q=1}^{N} r_m(i)r_{m+n}(i)r_q(i)r_{q+n}(i) \end{aligned} \quad \text{(D9b)}$$

In the last term, rewrite the sum on q leaving out $q = m$, and $q = m + n$ then in the remainder of the sum, $r_q(i)$ is never the same as any of the other 3 r's (it can't be the same as $r_{q+n}(i)$ because n is never zero) and it is isolated and becomes an i.i.d. zero-mean variable which makes the expectation of the r's zero. These two terms must be added back in and that transforms the last term as follows:

$$\begin{aligned} &\frac{1}{(N-1)^2}\sum_{m=1}^{N}\sum_{q=1}^{N} r_m(i)r_{m+n}(i)r_q(i)r_{q+n}(i) \\ &\qquad = \frac{1}{(N-1)^2}\sum_{m=1}^{N}\left\{\sum_{q=1}^{m-1} + \sum_{q=m+1}^{N} r_m(i)r_{m+n}(i)r_q(i)r_{q+n}(i)\right\} \\ &\qquad + \frac{1}{(N-1)^2}\sum_{m=1}^{N} r_m^2(i)r_{m+n}^2(i) + \frac{1}{(N-1)^2}\sum_{m=1}^{N} r_m(i)r_{m+n}^2(i)r_{m+2n}(i) \end{aligned} \quad \text{(D9c)}$$

The expectation of the first two terms on the right is zero because one of the r's in each is isolated from the others and zero-mean. The last term's expectation is also zero because each of the r-terms is isolated through n never being zero. The expectation of the sum of the terms is non-zero because of the third term only. Its expectation is the square of the expectation of each of the r^2 terms and the result is

$$\frac{1}{(N-1)^2}\sum_{m=1}^{N}\sum_{q=1}^{N}\langle r_m(i)r_{m+n}(i)r_q(i)r_{q+n}(i)\rangle = \frac{N\mu^2}{(N-1)^2} \qquad \text{(D9d)}$$

The second term on the right in Eq. (D9b) can be reduced using the same process by removing $m = p$, and $m = q + n$ from the sum on m (again, $n \neq 0$ so m is never equal to $m + n$ and that condition does not need to be removed explicitly).

$$\begin{aligned}
&-\frac{2}{N(N-1)^2}\sum_{m=1}^{N}\sum_{q=1}^{N}\sum_{p=1}^{N} r_m(i)r_{m+n}(i)r_p(i)r_{q+n}(i) \\
&\qquad = -\frac{2}{N(N-1)^2}\sum_{p=1}^{N}\sum_{q=1}^{N}\left\{\sum_{m=1}^{p-1} + \sum_{m=p+1}^{q+n-1} + \sum_{m=q+n+1}^{N} r_m(i)r_{m+n}(i)r_p(i)r_{q+n}(i)\right\} \\
&\qquad -\frac{2}{N(N-1)^2}\sum_{p=1}^{N}\sum_{q=1}^{N} r_p^2(i)r_{p+n}(i)r_{q+n}(i) \\
&\qquad -\frac{2}{N(N-1)^2}\sum_{p=1}^{N}\sum_{q=1}^{N} r_{q+n}^2(i)r_{q+2n}(i)r_p(i)
\end{aligned}$$

For $p < q + n$ the 3 terms on the right in the large braces have an expectation that is zero because $r_m(i)$ is never the same as either of the other r-terms because of the limits of the sum on m and because $n \neq 0$. The same holds true if $p > q + n$ as can be seen by breaking the m-sum at $q + n$ first and then at p. If $p = q + n$ the expectation is still zero: the r's become $r_m(i)r_{m+n}(i)r_{p+n}^2(i)$ which if $p = m$ leaves $r_p(i)$ isolated, and if $m = p + n$ it leaves $r_{p+2n}(i)$ isolated.

The full equation reduces to the last 2 terms which are dealt with in the same manner.

$$\begin{aligned}
&-\frac{2}{N(N-1)^2}\sum_{m=1}^{N}\sum_{q=1}^{N}\sum_{p=1}^{N} r_m(i)r_{m+n}(i)r_p(i)r_{q+n}(i) \\
&\qquad = -\frac{2}{N(N-1)^2}\sum_{q=1}^{N}\sum_{p=1}^{N} r_p^2(i)r_{p+n}(i)r_{q+n}(i) \\
&\qquad -\frac{2}{N(N-1)^2}\sum_{q=1}^{N}\sum_{p=1}^{N} r_{q+n}^2(i)r_{q+2n}(i)r_p(i)
\end{aligned}$$

where the order of the sums on the right have been interchanged. Again with the p-sum separated into three parts by removing $p = q$ and $p = q + n$ in the first term and $p = q + n$ and $p = q + 2n$ in the second, the equation becomes

$$
\begin{aligned}
&-\frac{2}{N(N-1)^2}\sum_{m=1}^{N}\sum_{q=1}^{N}\sum_{p=1}^{N} r_m(i)r_{m+n}(i)r_p(i)r_{q+n}(i) \\
&\qquad = -\frac{2}{N(N-1)^2}\sum_{q=1}^{N}\left\{\sum_{p=1}^{q-1}+\sum_{p=q+1}^{q+n-1}+\sum_{p=q+n+1}^{N} r_p^2(i)r_{p+n}(i)r_{q+n}(i)\right\} \\
&\qquad -\frac{2}{N(N-1)^2}\sum_{q=1}^{N-n}\left\{\sum_{p=1}^{q+n-1}+\sum_{p=q+n+1}^{q+2n-1}+\sum_{p=q+2n+1}^{N} r_{q+n}^2(i)r_{q+2n}(i)r_p(i)\right\} \\
&\qquad -\frac{2}{N(N-1)^2}\sum_{q=1}^{N} r_q^2(i)r_{q+n}^2(i) - \frac{2}{N(N-1)^2}\sum_{q=1}^{N-n} r_{q+n}^3(i)r_{q+2n}(i) \\
&\qquad -\frac{2}{N(N-1)^2}\sum_{q=1}^{N-2n} r_{q+n}^2(i)r_{q+2n}^2(i)
\end{aligned}
$$

The first 3 terms in the large braces on the right have an expectation that is zero because $r_{p+n}(i)$ is never the same as either of the other r-terms because of the limits of the sum on p and because $n \neq 0$. For the second 3 terms in large braces $r_p(i)$ is also isolated because of the limits and $n \neq 0$ and so their expectation is also zero. The expectation of the second to last term is also zero because $n \neq 0$ and both terms are isolated and of zero mean. If $N \leq 2n$ the last term is zero. The full equation reduces to the 2 remaining terms which provide the overall result

$$
\begin{aligned}
\langle C_i^2(N,n)\rangle \;=\; &\frac{1}{N^2(N-1)^2}\sum_{l=1}^{N}\sum_{m=1}^{N}\sum_{p=1}^{N}\sum_{q=1}^{N}\langle r_l(i)r_{m+n}(i)r_p(i)r_{q+n}(i)\rangle \\
&-\frac{\{2N+2(N-2n)\epsilon_{N,2n}\}(M-1)^2\mu^2}{N(N-1)^2}+\frac{N(M-1)^2\mu^2}{(N-1)^2}
\end{aligned}
\qquad \text{(D9e)}
$$

where $\epsilon_{a,b}$ is zero if $a \leq b$ and 1 if $a > b$.

When the 4-fold sum is simplified in the same way this becomes

$$
\begin{aligned}
\langle C_i^2(N,n)\rangle \;=\; &\frac{(M-1)^2\mu^2}{N^2(N-1)^2}\,[N^3 - N^2 - N + \epsilon_{N,n}(N-n)(2N-n-2+6H) \\
&-2N(N-2n)\epsilon_{N,2n}]
\end{aligned}
\qquad \text{(D9f)}
$$

$$\langle C_i^2(N,n)\rangle \;=\; \frac{(M-1)^2\mu^2}{N^2(N-1)^2}\,[N^3-N^2-N+\epsilon_{N,n}(2N^2-2N-4nN+2n^2+2n+6(N-n)H)\;-2N(N-2n)\epsilon_{N,2n}]$$

where $H=\langle r^4(i)\rangle/\langle r^2(i)\rangle^2\;\;(=M\varphi/\mu^2)$.

Noting that the right side of $\langle C_i^2(N,n)\rangle$ is independent of i, λ_I can be determined from Eq. (D5) and because of the ϵ's it splits into 3 parts depending on the relative sizes of N and n, the averaging length and the lag:

$$\frac{\lambda_I^2}{N}=\;(M-1)^2\mu^2\begin{cases}\dfrac{\{N^3-N^2-3N+2n^2+2n+6H(N-n)\}}{N^2(N-1)^2} & N>2n \quad \text{(D10a)}\\[2ex] \dfrac{\{N^3-N^2-N+(2N-2n-2+6H)(\mathrm{N}-n)\}}{N^2(N-1)^2} & n<N\le 2n \quad \text{(D10b)}\\[2ex] \dfrac{(N^3-N^2-N)}{N^2(N-1)^2} & N\le n \quad \text{(D10c)}\end{cases}$$

The result for λ_I is

$$\lambda_I=\;\frac{(M^2-1)}{12}\begin{cases}\sqrt{1+\dfrac{1}{(N-1)}-\dfrac{1}{(N-1)^2}+\dfrac{2[n^2-(1-3H)(N-n)]}{N(N-1)^2}} & N>2n \text{ (D11a)}\\[2ex] \sqrt{1+\dfrac{1}{(N-1)}-\dfrac{1}{(N-1)^2}+\dfrac{2[N-n-1+3H](N-n)}{N(N-1)^2}} & \text{(D11b)}\\[2ex] \sqrt{1+\dfrac{1}{(N-1)}-\dfrac{1}{(N-1)^2}} & N\le n \text{ (D11c)}\end{cases}$$

The terms in the radicals are $\{1\;+O(1/N)\}$ as N increases, so as $N\to\infty$ for all n

$$\lambda_I\to\;\frac{(M^2-1)}{12} \qquad \text{(D12)}$$

It can also be noted that with $H=M\varphi/\mu^2$

$$H=1.8\frac{M^2-\frac{7}{3}}{M^2-1}\approx 1.8 \qquad \text{(D13)}$$

and with $M=52$,

$$\lambda_I\to\;225.25 \qquad \text{(D14)}$$

Appendix E

Derivation of

$$\lambda_X = \frac{(M+1)\sqrt{M-1}}{12}$$

In section 3(c) λ_X is defined by

$$\frac{\lambda_X^2}{N} \equiv \langle\, S^2(N)\rangle \;= \frac{1}{11N^2}\sum_n \langle\left(\sum_{k=1}^{N} C_k(n)\right)^2\rangle \qquad \text{(E1)}$$

and with the definition of $C_k(n)$ from Eq. (29),

$$\frac{\lambda_X^2}{N} = \frac{1}{11N^2M^2}\sum_n \sum_{k=1}^{N}\sum_{j=1}^{N}\sum_{i=1}^{M}\sum_{m=1}^{M}\langle r_k(i)r_{k+n}(i)r_j(m)r_{j+n}(m)\rangle \qquad \text{(E2)}$$

The expectation of the r's presents 2 cases: (i) $j = k$ excluding $m = i$, and (ii) $j = k$ and $m = i$ simultaneously. If the subscripts are all different the expectation is zero because the r's for different subscripts are assumed to be statistically independent. If 2 of the subscripts are equal but the other two are separately different, again the expectation is zero because of the independence of the unmatched r's. This case arises if $k = j + n$ because then the remaining subscripts would be j and $k + n$ which would become j and $j + 2n$ which are never equal. The same situation occurs for $j = k + n$ in which k would have to also equal $j + n$ which would become $k + 2n$. The only case in which there is any matching occurring amongst all r's is $j = k$. There is also no case in which all 4 subscripts are simultaneously equal.

If $j = k$ there are two circumstances that will arise and will be treated separately, namely excluding $m = i$ and including it. They present different statistics.

(i) $j = k$, excluding $m = i$

The expectation for this case becomes

$$\langle r_k(i)r_{k+n}(i)r_j(m)r_{j+n}(m)\rangle_1 \;=\; \langle r_k(i)r_k(m)r_{k+n}(i)r_{k+n}(m)\rangle\delta_{jk}(1-\delta_{mi}) \qquad \text{(E3)}$$

The expectation separates into two parts, one with k-subscripts and one with $(k + n)$-subscripts and each is simply the covariance given in Appendix B Eq. (B6):

$$\langle r_k(i)r_{k+n}(i)r_j(m)r_{j+n}(m)\rangle_1 \;=\; \delta_{jk}\frac{(M+1)^2}{12^2}(1-\;\delta_{mi}) \tag{E4}$$

(ii) $j = k$; $m = i$ simultaneously

For the second case the expectation separates into 2 parts each of which is a square

$$\langle r_k(i)r_{k+n}(i)r_j(m)r_{j+n}(m)\rangle_2 \;=\; \delta_{jk}\delta_{mi}\langle r_k^2(m)r_{k+n}^2(m)\rangle \tag{E5}$$

The separation happens because of the assumed statistical independence of the k-subscript r's and the $(k+n)$-subscript r's, and as a result

$$\langle r_k(i)r_{k+n}(i)r_j(m)r_{j+n}(m)\rangle_2 \;=\; \delta_{jk}\delta_{mi}\langle r_k^2(m)\rangle\langle r_{k+n}^2(m)\rangle \tag{E6}$$

or, because

$$\langle r_k^2(m)\rangle \;= \frac{M^2-1}{12} \tag{E7}$$

$$\langle r_k(i)r_{k+n}(i)r_j(m)r_{j+n}(m)\rangle_2 \;=\; \delta_{jk}\delta_{mi}\frac{(M^2-1)^2}{12^2} \tag{E8}$$

The full expectation is the sum of these two cases:

$$\langle r_k(i)r_{k+n}(i)r_j(m)r_{j+n}(m)\rangle \;=\; \delta_{jk}\frac{(M+1)^2}{12^2}(1-\;\delta_{mi}) + \;\delta_{jk}\delta_{mi}\frac{(M^2-1)^2}{12^2} \tag{E9}$$

and from this

$$\frac{\lambda_X^2}{N} = \frac{1}{11N^2M^2}\sum_n\sum_{k=1}^{N}\sum_{j=1}^{N}\sum_{i=1}^{M}\sum_{m=1}^{M}\left\{\delta_{jk}\frac{(M+1)^2}{12^2}(1-\;\delta_{mi}) + \;\delta_{jk}\delta_{mi}\frac{(M^2-1)^2}{12^2}\right\} \tag{E10}$$

The sums can be carried out and the result is

$$\frac{\lambda_X^2}{N} = \frac{(M+1)^2(M-1)}{12^2N} \tag{E11}$$

which gives

$$\lambda_X = \frac{(M+1)\sqrt{M-1}}{12} \tag{E12}$$

For the present case of $M = 52$,

$$\lambda_X = 31.54 \ldots . \tag{E13}$$

If the covariance in Eq. (E1) is normalized so that it becomes the correlation,

$$\lambda_{Xnorm} = \frac{1}{\sqrt{M-1}} \tag{E14}$$

and for $M = 52$

$$\lambda_{Xnorm} = 0.140 \ldots \tag{E15}$$

The normalization is accomplished by dividing the covariance by the product of the standard deviations of each random variable. In this case the standard deviations are the same and the normalizing factor is simply $C_k(0)$ which is given by Eq. (E7).