2013

Using Excel to Develop Random Number Sense

Thomas C. Juster
University of South Florida, juster@usf.edu

Spreadsheets in Education (eJSiE)

Volume 6 | Issue 2 Article 4

6-20-2013

Using Excel to Develop Random Number Sense

Thomas C. Juster
University of South Florida, Tampa, juster@usf.edu

Using Excel to Develop Random Number Sense

Abstract
People have difficulty creating and recognizing random sequences, but can become better at these tasks through practice and feedback. This paper describes how simulations produced with Excel can be used to visually display a sequence of random events, allowing users to develop their randomness cognition. Instructions are provided for simulating simple binary events, dice rolls, and events generated by a Poisson process. The results are then visualized using Excel's conditional formatting option or by displaying a bar graph. A survey of students learning the concept of recurrence interval showed that most enjoyed interacting with the Excel simulations and gained insights not apparent from reading about the concept or viewing static diagrams.

Keywords
Random sequences, Probability, Conditional Formatting, Simulation, Randomness Cognition

Using Excel to Develop Random Number Sense

Abstract

People have difficulty creating and recognizing random sequences, but can become better at these tasks through practice and feedback. This paper describes how simulations produced with Excel can be used to visually display a sequence of random events, allowing users to develop their randomness cognition. Instructions are provided for simulating simple binary events, dice rolls, and events generated by a Poisson process. The results are then visualized using Excel's conditional formatting option or by displaying a bar graph. A survey of students learning the concept of recurrence interval showed that most enjoyed interacting with the Excel simulations and gained insights not apparent from reading about the concept or viewing static diagrams.

Keywords: Random sequences, probability, conditional formatting, simulation, randomness cognition

1. Introduction

Random sequences[1] arise in many mathematical and scientific contexts. They occur in disciplines that are inherently probabilistic, like statistical mechanics or epidemiology, but also in any field that studies random or quasi-random events. In the physical sciences, for example, phenomena such as earthquakes, volcanic eruptions, tornadoes, hurricanes, and floods are largely stochastic, producing temporal random sequences that scientists must analyse in order to evaluate risk. In the business world, stock prices represent a sequence of values that can be modelled as a random walk [3]. Random sequences even play an important role in our everyday lives—from the weather, to games, to determining the prices we pay to insure our homes against disaster.

Despite their importance, most people have difficulty both generating and recognizing random sequences [2, 4-7]. This difficulty has been attributed to belief in the "Law of Small Numbers" (LOSN) [8], the erroneous assumption that if a process is random, then its statistical properties will be found in every sequence it generates, no matter how small. Where the process is unquestionably random, as in games of chance, the LOSN leads to mistaken predictions about future events [9]. For example, following a sequence of three coin flips of heads, many people assume a tail is imminent. This prediction is based on probabilities of the coin inferred from the sample of four: if the last flip is a tail, then the probability of a tail becomes 1/4 (HHHT), which is more representative of the true probability than what would be inferred if the last flip were a head (HHHH = 0/4). This fallacy has been called the "gambler's fallacy" (GF).

People also fall for the logical contrapositive to the LOSN: that if the statistical properties measured in a small sequence are unlike those of a random variable, then the generating process isn't random. In this case the (contrapositive of the) LOSN leads to a mistaken conclusion about the process that generated the sequence [9]. This fallacy has been called the "hot hand fallacy" (HHF) [10] because it arises in sports, especially basketball, where a series of successes is assumed to be the result of a shooting streak, or "hot hand", not chance[2]. In fact, the HHF may be the more common fallacy because people instinctively assume that sequences are generated by deterministic processes, only accepting a random explanation when all other explanations fail [13]. Thus except where the process is unambiguously random—as it is in games of chance—people are more likely to fall prey to the HHF than the GF.

---

[1] There is a rich literature on the definition of random sequences, a concept which was introduced by Von Mises in 1919 [1]. For the purposes of this paper I will simply define a random sequence as a set of numbers each of which is the result of an identical sequentially independent stochastic process. By this definition a sequence is random if it is generated by a random process, regardless of whether or not structure appears in the sequence through serendipity. For a discussion of the difference between sequences generated from random processes vs. intrinsically random sequences, see Nickerson [2].

[2] Whether or not streak shooting is actually a statistical artefact has been the subject of vigorous debate among psychologists and statisticians. Recent studies [11, 12] have concluded that it probably is.

Fortunately, there is some evidence that people can improve their random number sense. Students trained in either mathematics [14] or statistics [15] perform better on tests of their randomness cognition, as do college students in general [16], suggesting that the acquisition of random number sense is a component of quantitative literacy. More directly, experiments demonstrate that subjects can be coached to improve their ability to generate random sequences by providing detailed instructions about the task, or feedback about the statistical properties of the sequences they create [15, 17]. Since the subjects clearly must learn how to recognize flaws in their purportedly random sequences in order to improve them, these studies also implicitly suggest that people can be trained to improve their perception of random sequences as well.

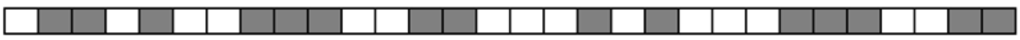
My goal in this paper is to describe a teaching method that persuades students to reject the LOSN and its contrapositive, thus discrediting both the HHF and GF. The method uses visualizations of the variability of sequences produced by stochastic processes. Because this variability is difficult to explain using static diagrams, a dynamic simulation method is used instead.

2. Visualizing Random Sequences with Conditional Formatting

Excel is an ideal tool to create and display random sequences simply, without requiring knowledge of computer programming. For example, suppose we want to generate a sequence of outcomes representing the results of flipping a coin. For this binomial case we define 1 = heads, 2 = tails, and simulate the outcome of a single flip with the RANBETWEEN() function in a cell; e.g.,

=RANDBETWEEN(1,2)

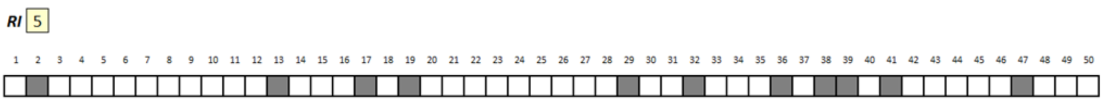
To highlight the result, conditional formatting is used to shade the cell background and font the same dark color when the outcome is 1 ("heads"), and white when the outcome is 2 ("tails). A dark outside border is used to make cell stand out against the background. If we wish to simulate 25 successive coin flips, we then copy the formula and formatting in this cell to an array of 24 neighbouring cells:



The same process can be used to simulate binomial events with a probability that can be expressed as the reciprocal of an integer. In these cases it is often easier to use the recurrence interval ($RI = 1/P$) to describe the frequency of the event. In the next example the recurrence interval is entered in a separate cell to serve as an adjustable constant. The steps to create the simulated sequence are the same, except that the formula in the conditionally formatted cell now references the constant cell; e.g.,

=RANDBETWEEN(1, *Con*)

where *Con* is the absolute address of the cell that contains the *RI*. We now define a result of 1 to indicate an event occurred, and any other value to mean it didn't. Here's an example with a *RI* of 5 (probability = 0.2), simulated over 50 time periods:

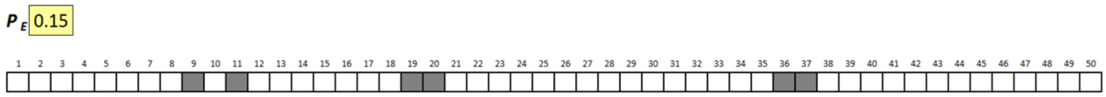


3

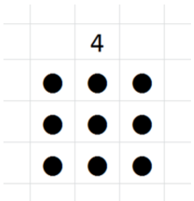
For probabilities that cannot be expressed as the reciprocal of an integer, the RAND() function, which generates a real number between 0 and 1, can be used instead. If the probability of an event is given as P_E ($0 < P_E < 1$) then its random occurrence can be simulated with the statement

=IF(RAND()<=P_E ,1, 0)

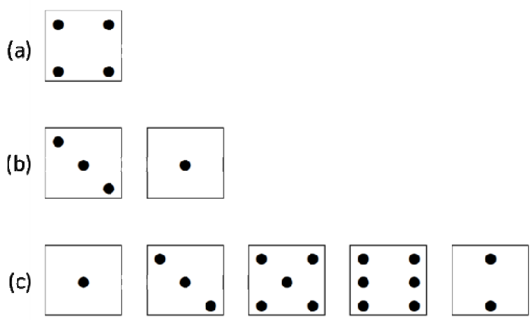
In this case a result of 1 indicates the event occurred, while a result of 0 indicates it didn't. The example below shows a sequence of 50 random events each with a probability $P_E = 0.15$. Once again, the value of P_E has been placed in a separate cell so it can be easily varied.



Dice can be simulated with another trick of conditional formatting. First, the outcome of a roll is simulated with the function RANDBETWEEN(1,6). Then the circle symbol (●) is input into each of nine cells in a block, and the cell width and height of each adjusted so it looks like a 'dot':



Finally, conditional formatting is used to turn on or off each of the nine dots depending on the value of the random number (which can then be hidden by changing its font color to white). To complete the illusion the internal cell borders are set to white and a bold black border is drawn around the entire 'die':



The first example (a) simulates a single roll of a die. Games like craps or Monopoly can be simulated with two dice, as in (b), and Yahtzee with five dice (c). To "roll" the dice, users recalculate the entire spreadsheet, generating new random numbers.

Some events have probabilities that are best described with a Poisson distribution. For example, natural phenomena like floods, earthquakes, and hurricanes can occur continuously in time even though outcomes are typically recorded over a time period, such as one year. Simulating these events with Excel is a little more complicated.

Poisson probabilities depend on the rate , called the "mean" in Excel with symbol x. This is the long-term average probability of the event, with units of number of events per time period (e.g., number of floods/year). The probability of n or fewer events in a time period is then computed by the function POISSON.DIST(n - 1, x, TRUE). For example, if $x = 0.3$ then the probability of two or fewer events is given by:

=POISSON.DIST(1, 0.3, TRUE)

Using POISSON.DIST we first build up a reference table of the probabilities of $n = 0$, 1, 2 . . . or fewer events, as shown below for $x = 0.3$ and :

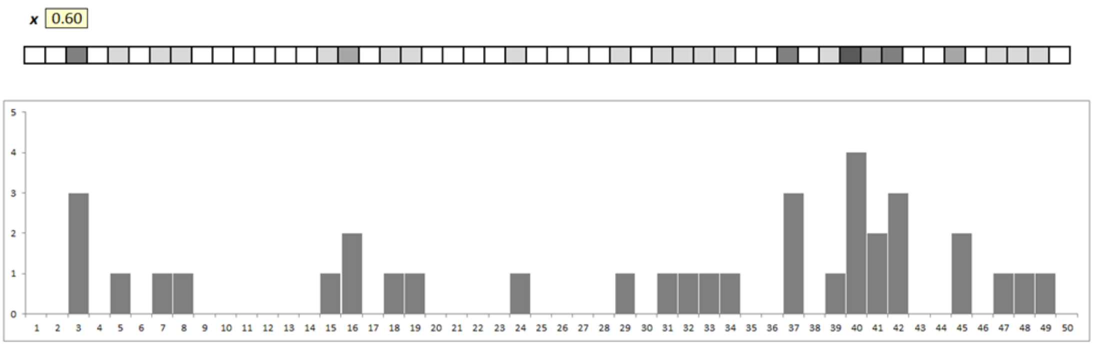
	A	B
1	*x*	0.3
2		
3	***P(n)***	***n***
4	0.00000	0
5	0.74082	1
6	0.96306	2
7	0.99640	3
8	0.99973	4
9	0.99998	5

	A	B
1	*x*	0.3
2		
3	***P(n)***	***n***
4	0	0
5	=POISSON.DIST(B5-1,$B$1,TRUE)	1
6	=POISSON.DIST(B6-1,$B$1,TRUE)	2
7	=POISSON.DIST(B7-1,$B$1,TRUE)	3
8	=POISSON.DIST(B8-1,$B$1,TRUE)	4
9	=POISSON.DIST(B9-1,$B$1,TRUE)	5

The left panel shows the results of the formulas (normal view); the right panel shows the equations in each cell. Then it's a simple matter of using a lookup table [VLOOKUP()] to determine the number of simulated events during a given time period, given a random number between 0 and 1 generated with RAND(). For example, using the Poisson table of probabilities above, the following formula is entered into each cell:

=VLOOKUP(RAND(),$A$4:$B$9,2,TRUE)

The outcome represents the number of occurrences of the event during one time period, produced by a random draw from the Poisson distribution. Once again the results can be shaded using conditional formatting to enhance visualization, but now each cell is not a binomial result so the method must be modified. One option would be to choose a colour (or shade) that corresponds to the number of events in each cell. Another option is simply to use a bar graph. An example of 50 periods of data with x = 0.6 using both visualization approaches is shown below.



5

3. Discussion

With all of these simulations, a key advantage of Excel is that the user can continue to explore additional sequences produced by the same process by simply pressing a single key (<F9> on a PC) or key sequence (<CMD> + '=' on a Mac) which recalculates the spreadsheet and generates new random numbers. These different realizations allow users to quickly develop number sense about the variety of sequences produced from a single random processes. As an example, notable gaps and clusters—which most people do not expect to find in a random sequence [6]—appear in all the examples above. An appreciation for the frequency with which these gaps and clusters appear can readily be achieved in a few minutes.

An additional advantage of the Excel simulations is their simplicity, allowing students with even limited skills not only to use them, but to create them. Research has shown that students who construct their own spreadsheets are more invested in the results and learn more than students who passively work with a program someone else wrote [18]. They are also more likely to explore 'what-if' scenarios; i.e., what happens if I double the recurrence interval? All of the simulations described above can be created with nothing more than a basic understanding of Excel functions and a quick lesson in conditional formatting. The conditional formatting has the additional benefit of requiring students to apply their comprehension of logical statements, which is needed to define the formatting rules.

The following four activities illustrate how instructors might use the tools described above to teach about random sequences in the classroom. The first two use simulations created by the instructor, while the last two take advantage of the insights that occur when the students build the spreadsheet themselves.

Activity 1: *Learning how to generate random coin flips*. The key to training people to generate a random binary sequence (like coin flips) is recognizing what a truly random sequence looks like. The most common mistake people make is to assume that a random sequence produces too many flips and too few runs[3] [4, 17]. Students first create what they consider to be a random sequence of 100 coin flips. The instructor then displays an Excel simulation of 100 coin flips and allows the students to repeatedly explore additional sequences by recalculating the spreadsheet. The teacher then leads a discussion of how the simulated sequences differ from the ones generated by the students, and the students are given an opportunity to try again. This can be repeated until the teacher is satisfied the students have a good understanding of the nature of a true binary random sequence.

Activity 2: Learning the meaning of the "10-year flood". In earth science, probabilities of infrequent natural events are commonly expressed in terms of *RI*. This terminology leads to the common misconception that the events occur at regular intervals [19, 20]. Even when students are taught that the *RI* is the average interval between events they do not appreciate how much these intervals can vary. Students are first

---

[3] By *flip* I mean a change from one outcome to the other, while for *run* I mean a sequence of identical outcomes. For example, if a sequence is generated by flipping a coin then a flip would be TH or HT and a run might be HHH or TTTT.

instructed to create a 100-year timeline on which they mark the occurrences of a "10-year flood" event. The teacher then displays an Excel simulation of these events over 100 years, and encourages the students to explore additional realizations by recalculating the spreadsheet. This is followed by a discussion where the students re-examine their assumptions and the teacher emphasizes the implications of their conclusions. For example, the teacher might ask: is it possible that 20 years go by without a single 10-year flood? How does the probability of a 10-year flood this year change if one occurred last year?

Activity #3: Confronting the gambler's fallacy. When gamblers fall victim to the GF they implicitly believe that events are not independent, even if they accept that they are random [21, cited in 2]. In other words, the die or coin somehow knows what has happened before, and attempts to correct the distribution towards its long-term average [22]. If students create the simulation program, however, they cannot escape the reality that the events are independent. In this activity groups of students are instructed to create a spreadsheet that simulates the roll of a die. They then use this program to play a game of "Pig", a simple dice game in which players score points by rolling a single die, and must decide whether or not to jeopardize their previous gains by continuing to play and risk rolling a one, which voids all points won during that turn [23]. Students who play the game without being misled by the GF will be more successful.

Activity #4: Is there such a thing as a "hot hand" in sports? The HHF endures in part because people don't believe that a random process can produce long runs of a single outcome [11]. In this activity students are charged with simulating outcomes of a binary process ("success" vs. "fail") using Excel. "Success" could be defined as making a shot in basketball, getting a hit in baseball, or converting a penalty kick in football (soccer).The key educational value in having the students create the spreadsheet is that it ensures they believe the generating process is truly random. After the students have created their spreadsheets, they can then assess the validity of the HHF by using simulated sequences to answer the following types of questions: "assuming a player makes 45% of his shots in basketball, how often will a run of 5 straight made shots occur in a typical game in which he shoots 20 times?"; or "assuming a baseball player has a 0.300 batting average, how likely is it that she will go hitless in six straight at-bats"?

Activities like these offer several advantages. They are easily incorporated into inquiry- or problem-based learning activities in which students must construct their own understanding by solving problems, a strategy that has been shown to promote a deeper understanding of the scientific process [24, 25]. Simulations can be run to explore data that are relevant to students, a key to teaching statistics [26]. The activities can be written to appeal to the modern student's love of technology without requiring them to learn a computer language. And, finally, they are likely to be viewed as more "fun" than watching a lecture or reading a book.

To assess the effectiveness of Excel simulations of random events, a survey was administered to students in an introductory geology course for engineering students. These students were mathematically sophisticated, but most had not taken a course in statistics. Students were quizzed on their knowledge of the concept of RI three times: once before the survey, once after reading a standard description of the topic compiled from two websites of the United States Geological Survey [27, 28] (which included several figures), and once again after interacting with a simple bimodal Excel simulation program written by the instructor. The survey asked the students a series of open-ended essay questions to determine whether they enjoyed using the Excel simulations, whether they thought the simulations were more or less useful than reading the text, and whether they were surprised at any of the results of the simulations (Table 1). Students overwhelmingly liked the simulations and considered them more useful than the text, while 20% considered the combination of text + simulations to be the most effective strategy. Fully three-quarters of the respondents admitted to being surprised at the results of the simulation program, with the most common reason being the large variation in the return intervals. Again, this is consistent with the prediction that untrained students are seduced by the LOSN, and underestimate the degree of variation in randomly generated events.

Table 1: Survey Results

	USGS Text	Simulation	Combination
Which method was more useful? (*n* = 25)	20%	60%	20%

	Yes	No	Explanations cited more than three times
Did the Excel simulations provide any additional insights into the distribution of random events that you didn't get from reading about them? If so, what were they? (*n* = 25)	96%	4%	Interval between events very variable It is normal to have large gaps

	Yes	No	Explanations cited more than three times
Did anything surprise you about the random distributions simulated with Excel? If so, what? (*n* = 24)	75%	25%	Interval between events very variable (cited by nearly 50% of respondents)

References

[1] Von Mises, R. (1919). Grundlagen der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift* 5: 52-99.

[2] Nickerson, R.S. (2002). The production and perception of randomness. *Psychological Review* 109(2): 330-357.

[3] Cootner, P. (1964). *The random character of stock market prices*. Cambridge: MIT Press.

[4] Reichenbach, H. (1934/1949). *The theory of probability*. Berkeley: University of California Press.

[5] Bar-Hillel, M. and Wagenaar, W.A. (1991). The perception of randomness. *Advances in Applied Mathematics* 12(4): 428-454.

[6] Gilovich, T.R. (1991). How we know what isn't so: The fallibility of human reason in everyday life. New York: The Free Press. Kindle edition.

[7] Nickerson, R.S. (2004). *Cognition and chance: The psychology of probabilistic reasoning*. Mahway, New Jersey: Lawrence Erlbaum Associates.

[8] Tversky, A. and Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin* 76(2): 105-110.

[9] Ayton, P. and Fischer, I. (2004). The hot hand fallacy and the gambler's fallacy: Two faces of subjective randomness? *Memory and Cognition* 32 (8): 1369-1378.

[10] Gilovich, T.R., Vallone, R., and Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology* 17: 295-314.

[11] Koehler, J.J. and Conley, C.A. (2003). The "Hot Hand" myth in professional basketball. *Journal of Sport and Exercise Psychology* 25: 253-259.

[12] Bar-Eli, M., Avugos, S., and Raab, M. (2006). Twenty years of "hot hand" research: Review and critique. *Psychology of Sport and Exercise* 7: 525-553.

[13] Diener, D. and Thompson, W.B. (1985). Recognizing randomness. *American Journal of Psychology* 98 (3): 433-447.

[14] Chapanis, A. (1953). Random number guessing. *American Psychologist* 8: 332.

[15] Lopes, L.L. and Oden, G.C. (1987). Distinguishing between random and non-random events. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 13: 392-400.

[16] Ross, B.M. and Levn, N. (1958). Patterned prediction of chance events by children and adults. *Psychological Reports* 4: 87-124.

[17] Neuringer, A. (1986). Can people behave "randomly?": The role of feedback. *Journal of Experimental Psychology: General* 115: 62-75.

[18] Baker, J. and Sugden, S. J. (2007). Spreadsheets in education—The first 25 years. *Spreadsheets in Education (eJSiE)*: 1(1): 18-43.

[19] Costa, J.E. (1978). The dilemma of flood control in the United States. *Environmental Management* 2(4): 313-322.

[20] Lutz, T. (2011). Toward a new conceptual framework for teaching about flood risk in introductory geoscience courses. *Journal of Geoscience Education* 59: 5-12.

[21] Gold, E. and Hester, G. 1989. *The gambler's fallacy and the coin's memory*. Unpublished manuscript. Carnegie-Mellon University, Pittsburg, PA.

[22] Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science* 185 (4157): 1124-1131.

[23] Tredd, W. (2013). *Dice Games New and Old*. Oldeander Press, Amazon Kindle edition.

[24] Byers, A. and Fitzgerald, M.A. (2002). Networking for leadership, inquiry, and systemic thinking: A new approach to inquiry-based learning. *Journal of Science Education and Technology* 11(1): 81-91.

[25] Savery, J.R. (2006). Overview of problem-based learning: Definitions and distinctions. *Interdisciplinary Journal of Problem-based Learning* 1(1): 5-22.

[26] Gould, R. (2010). Statistics and the modern student. *International Statistical Review* 78(2): 297-315.

[27] Holmes, R.R. Jr. and Dinicola, K. (2010). *100-year flood—it's all about chance*. United States Geological Survey General Information Product 106, 1p. Available at: http://pubs.usgs.gov/gip/106/.

[28] Dinicola, K. (1996). The 100-year flood. United States Geological Survey Fact Sheet 229-96, 1p. Available at: http://pubs.usgs.gov/fs/FS-229-96/pdf/FS 229-96.pdf.