Testing the negative recency effect among teacher students trying to generate random sequences

Olav Gravir Imenes

Oslo Metropolitan University, Norway; ogim@oslomet.no

Students of a teacher training program (N=124) were asked to generate, unaided, a binary sequence of 40 symbols appearing to be as random as possible. It was found that the average probability of alternation, P(A), was 0.61, which agrees with previous experiments described in the literature, and that the students tend to underestimate the occurrence of runs with four or more equal symbols. When comparing to random sequences I argue that only asking for sequences of length 40 will overestimate P(A).

Keywords: Negative recency effect, randomness, probabilistic reasoning.

Introduction

One of the goals of education in the subject of probability is to give students a better understanding of what constitutes random events and how to recognise random events. A democracy needs citizens that can understand and evaluate quantitative information and statistical analyses (Utdannings-direktoratet, 2012). Media sometimes reports clusters, for example the occurrence of many cancer cases in small communities, without acknowledging that such clusters may arise as a result of randomness. Knowledge of what to expect of random events is necessary in order to know how to process such information. However, it is difficult to get a good grasp of probability, and especially the occurrence of clusters. One type of cluster occurs in binary sequences. A binary sequence is a sequence that consists of two symbols, for example the sequence generated by tossing a coin multiple times and noting whether the outcome was head or tail, or throwing a die and noting whether it displays an even or odd number. The former could look like THHHHTTHTT, while the latter could look like 0111100100. A cluster in a binary sequence is a long run of the same symbol, for example four heads or four odd numbers in a row. Research (Bryant & Nunes, 2012; Chiesi & Primi, 2009; Falk & Konold, 1997; Williams & Griffiths, 2013) shows that people overestimate the number of changes in order for a binary sequence to be random and underestimate the average length of runs, i.e., the number of times a symbol is repeated consecutively.

In Norwegian schools in the last 20 years there has been a stronger emphasis on probability than previously, especially with the introduction of curriculum reform in 1994 for high school, R94 (Kirke, utdannings- og forskningsdepartementet, 1999), and in the curriculum reform in 2006 for both primary and secondary school, LK06 (Kunnskapsdepartementet, 2013). For example, the curriculum for the first year of high school demands that the student should be able to produce examples and simulations of random experiments (Kunnskapsdepartementet, 2013, p. 11).

One might ask what the stronger emphasis on probability means to students' understanding of clusters. If, successively, an even number has been thrown four times with a fair die, and students are asked if there is still a 50% chance to throw an even number in the next throw, most will immediately answer yes. However, when asked to construct sequences, they may tend to

underestimate such probabilities, i.e., seemingly assigning a less than 50% chance of getting the same result as in the previous throw. This paper will focus on Norwegian teacher students' understanding of what constitutes a random binary sequence. In order to measure the randomness of binary sequences, Falk and Konold (1997) presents a framework for analysing binary sequences mathematically, using among other indicators a probability of alternation, *P*(A). I will be using parts of this framework.

This paper sets out to explore the following question: To what extent are Norwegian teacher students able to understand the probability of forming of clusters in a binary sequence? The guesses will be measured by the calculated *P*(A), and by testing whether or not the students' guesses for the number of runs of given lengths differ significantly from that of a fair die thrown.

Review of the literature

Children's and adults' understanding of what constitutes a random sequence tends to show some common misconceptions. One such misconception is the negative recency bias (Bryant & Nunes, 2012, p. 10). This is the assumption that after a long sequence of the same result, e.g., six heads in row when tossing a coin, a tail is more likely in the next toss. The counterpart of this misconception is the positive recency bias (Bryant & Nunes, 2012, p. 10), in which people estimate that a certain result is more likely to be the next outcome because it has occurred frequently in the past. Positive recency bias occurs in for example estimating scoring in baseball (Gilovich, Vallone, & Tversky, 1985). Many studies have been carried out investigating the negative recency effect. For surveys of such investigations, see for example Batanero and Sanchez (2005), Bryant and Nunes (2012), Chiesi and Primi (2009) and Falk and Konold (1997).

Some of the studies carried out ask the participant to tell which of two or more sequences or patterns are random and which are not (Batanero & Serrano, 1999; Falk & Konold, 1997; Kahneman & Tversky, 1972) or which result is more likely to come next (Chiesi & Primi, 2009; Fischbein & Schnarch, 1997), and such studies are examples of judgement tasks. Others ask the participants to generate random sequences (Bakan, 1960; Towse & Mclachlan, 1999), and are production tasks. A variation is studied by Rapoport and Budescu (1992), where the participants generated random numbers as part of a game. Bar-Hillel and Wagenaar (1991) argue that judgement tasks are a purer way of studying the perception of randomness. However, the basic biases were discovered by production tasks, since these tasks were the ones in the early research. Another problem with judgement tasks is that when a researcher asks whether or not a given sequence is random, he may not himself know the answer. For example, Green (1982, p. 157) provides an example of two binary sequences of length 150 and 153 respectively and asks the subject to determine which sequence is made up. The one he would classify as not made up has a *P*(A) as low as 0.44 in addition to having one run of length 9, which is not expected for such a short sequence. At least when assigning production tasks, the researcher does not need to say anything wrong. In addition, when assigned production tasks, the students cannot guess among the alternatives, thus taking a more active role.

Falk and Konold (1997) present a framework for analysing binary sequences with respect to randomness. For every binary sequence a number called probability of alternation, *P*(A), may be

calculated. It is given by $P(A) = \frac{r-1}{n-1}$, where r is the number of runs, and n is the length of the sequence (Falk & Konold, 1997). In an infinitely long, truly random binary sequence, the expected P(A) is 0.5. The P(A) measure first order dependencies. Another measure is the second-order entropy (EN). This is based on the relative frequency of all ordered pairs, and is a measure of the amount of new information provided by the second symbol of the pair (Falk & Konold, 1997). The second order entropy is maximal (EN = 1) when all the four pairs, 00, 01, 10, 11, are equally probable. It is minimal (EN = 0) when P(A) = 1. Yet another measure is the complexity of the sequence, and can be defined as the "bit length of the shortest computer program that can reproduce the sequence" (Falk & Konold, 1997, p. 306). Another method Falk and Konold describes, due to Garner (1970), is to sort all sequences of a given length n into different disjoint sets based on their P(A). The most random sequences are the ones contained in the set which consist of the maximal number of sequences. In this paper we will focus on the P(A), and also of the number of subsets of length m with P(A) = 0, that is, the number of runs with a given length.

The typical value of P(A) for a binary sequence constructed by a person seems to be around 0.6. The nine studies referenced by Falk and Konold (1997) have P(A) ranging from 0.56 to 0.63. For nine experiments, referenced in the same paper, where participants are judging whether or not a certain sequence is random the P(A) range from 0.57 to 0.65 in eight of the experiments while one has a P(A) of between 0.7 and 0.8 (Gilovich et al., 1985). Thus, the literature seems fairly consistent regarding the extent to which people overestimate the number of runs. However, there seems to be a lack in reporting standard deviation. It is also interesting to examine to what extent people tend to underestimate the possibility of long runs, which we may call clusters. The most consistent result on binary sequences is that both in generating and perception people tend to underestimate the number of long runs, i.e., consecutively equal results (Falk & Konold, 1997, p. 302).

Method

I had 127 students enrolled in the first-year teacher education study program providing me with data. For entry into the program the students need more than a passing grade in mathematics from high school. In Norway, the lowest grade in high school is a zero, while 6 is the best obtainable grade. The requirement for entry to the teacher education program is a 4, while 2 is the lowest passing grade. Thus, the enrolled students had a somewhat better understanding of mathematics than that required for other study programs. As they were about to start learning about probability and combinatorics, the students were asked to imagine the following task: Throw a die 40 times and mark for each throw whether you get an even or an odd number. Mark an even number with 0 and an odd with 1. The students were asked to write down as realistic a sequence as possible. Afterwards they were asked to actually throw a die 40 times, and compare the results. Thus, I had about 5000 actual die throws to compare. The students were informed that the data they provided could be used for research. All data were collected anonymously. Among the participants three students did not produce a binary sequence of at least 40 symbols, and thus they were discarded. Therefore 124 student generated guessed sequences were analysed.

In order to determine if the student made a good guess or not, the $P(A)$ was the first indicator, and was compared with the theoretical $P(A)$ of 0.5 and the $P(A)$ of the actual die throws. In the latter comparison, a *t*-test was used since it was possible to calculate standard deviations. The second criterion was that the number of runs of each length in a guessed sequence should be approximately equal to the number of runs of that length in the thrown results. A list of the number of runs each student guessed for any given length from one to seven was made, and for each run length a *t*-test was carried out to compare to the corresponding list for the sequences obtained from the actual die thrown. The data was tested for normality using the Shapiro-Wilk test. A *t*-test ordinarily requires a normal distribution. However, the central limit theorem ensures that when the number of participants is large, it can be used even though the data does not have a normal distribution. Typically, this occurs when the number of participants is more than 30.

The tests described above were carried out using R: A language and environment for statistical computing (R Core Team, 2016).

Results

Each student produced one sequence of 40 binary digits, representing even and odd numbers. The sequence guessed by participant number 85 is provided in Figure 1. An even number is represented by 0 while an odd number is represented by 1. Another example is given in Figure 2, the sequence of participant number 93.

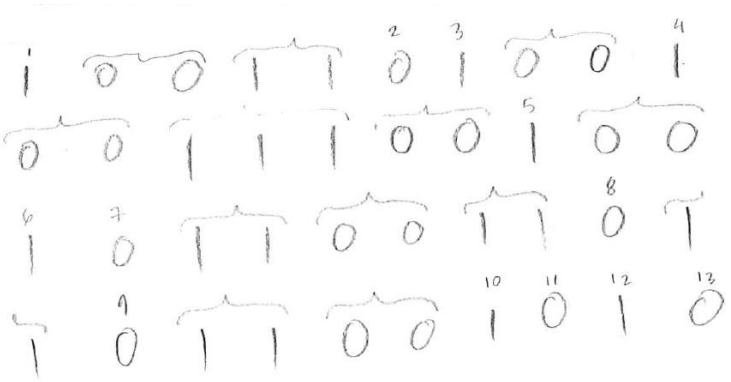

Figure 1: The guessed sequence of participant number 85

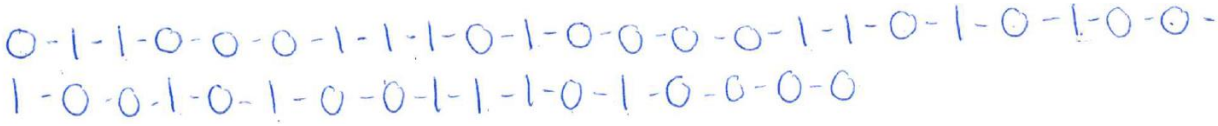

Figure 2: The guessed sequence of participant number 93

In addition to guessing how such a sequence would look like, the students each threw a die 40 times, and recorded the outcome. An example of such a sequence, form participant number 88, is given in Figure 3

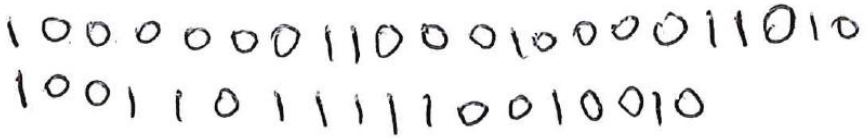

Figure 3: The actually thrown sequence of participant number 88

The probability of alternation was calculated for each participant. The mean $P(\mathrm{A})$ among the participants was 0.61 (SD = 0.12). For the random numbers the students threw, the mean $P(\mathrm{A})$ were 0.51 (SD = 0.08), quite near the theoretical value of 0.5 for infinitely long sequences. Ten students made guesses with $P(\mathrm{A}) < 0.51$. The $P(\mathrm{A})$ for the estimated results were not normally distributed, the Shapiro-Wilk normality test giving $p < .001$, where for the random generated sequences, that is when the die was actually thrown, the $P(\mathrm{A})$ was normally distributed, with the Shapiro-Wilk normality test giving $p = .08$. There was a significant difference between the made-up sequences compared to the actual throws ($t(213.72) = -8.23,\ p < .001$).

The number of runs of a given length in the estimated sequences varied. In Figure 4 the median and the quartiles of the number of runs of a given length for each student's guesses are shown in a boxplot. For each run of a given length, the corresponding result for the actual throws is shown to the right for comparison. The box labeled RLG1 shows the first and third quartile of the number of runs of length one in the guessed data as the lower and upper sides of the box. The solid line in the middle of the box is the median, and the mean is given by an asterisk. Similarly, RLG2 shows the number of runs of length two. RLT1 shows the number of runs of length one for actually thrown dice, and so on.

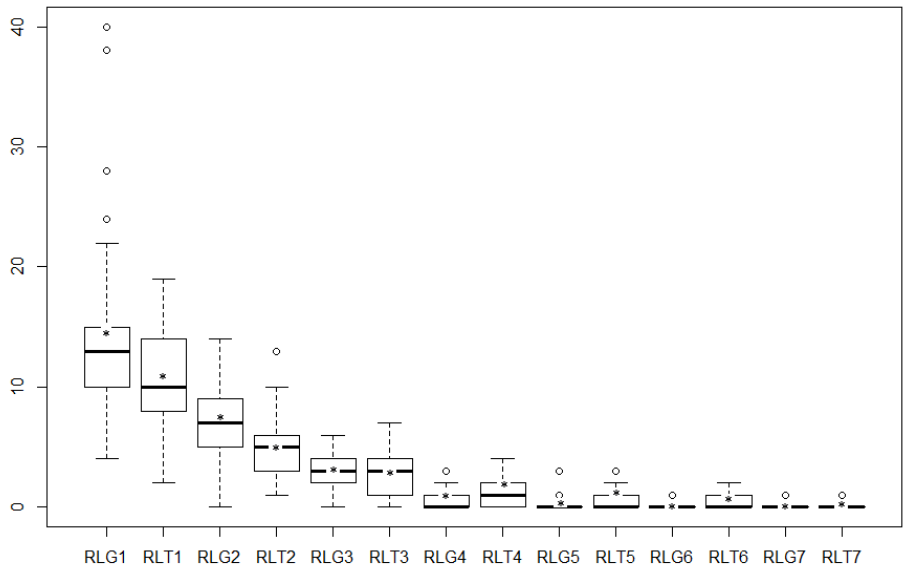

Figure 4: A boxplot of the number of runs of a given length

We note that the students tend to underestimate the number of runs of length four and five, and overestimate the number of runs of length one and two. For each run length, a t-test were carried out, and in particular it showed that for every run length except three, there was a significant difference ($p < .002$) between the guessed results and the actual throws. For runs of length three, the

difference was not significant ($t(243.31) = 1.00$, $p = .32$). With the sequence length being only 40, we cannot from this data conclude that students underestimate the numbers of run of length six or more. The reason is that with 20 expected runs, and only a $\frac{1}{2^{(6-1)}} = \frac{1}{32}$ chance of getting a run of length six for every run started, a student trying for the most random sequence should not include runs of length six or more. Also, in Figure 4, especially note that the median of the number of runs of length five for the actual thrown dice is zero, even though the average is 0.58. However, remember that the probability to get a run of length five or more is double that of obtaining a run of exactly length five. Therefore, we should expect student guesses to include one run of length five or more, but not be surprised if none were included. We have not included runs of length longer than seven in the boxplot as no students included runs of that length in their guesses.

Among the students there are some that made better guesses than others. Among the 124 students which completed the task, 49 included at least one run of length four or more in their guessing. Of these, 14 students included a run of length five or more.

Discussion

The average results for the probability of alternation, $P(A)$, when the students are guessing, is in line with the literature, being 0.61. This is significantly different from a random binary sequence. I have not found the standard deviations from the previous experiments, so no hypothesis test could be done comparing this $P(A)$ to the $P(A)$'s listed by Falk and Konold (1997). However, I conclude that this number seems fairly constant across countries and time. The list compiled by Falk and Konold (1997) seems to indicate that the $P(A)$ becomes smaller as the length of the guessed sequence increase. The result from the current experiment seems to agree with experiment where sequences of comparable length has been used. One reason for a lower $P(A)$ when longer sequences are used may be that participants then will expect some very long runs. For sequences as short as 40, a student who knew about the correct distribution could argue, that since the median of runs of length five or higher is zero, such sequences should not be included in the guess, even though the mean is closer to one than zero, and that the expectation value for the number of runs of five or more in length is more than one. Excluding runs of length five or higher should lead to a somewhat higher $P(A)$ than 0.5. Thus, it is not surprising that students overestimate the $P(A)$. However, the observed average $P(A)$ of 0.61 is higher than should be expected even when excluding runs of length five or more.

From the results (Figure 4) we observe that the students' guesses are relatively accurate concerning runs of length five, and this is confirmed by the *t*-test. The problems with underestimating the number of long runs begin with runs of length four or more. Each student should have expected at least one run of length five or more, as this will happen in one of 16 cases when starting on a run, and each student would be expected to start 20 runs. The approximate number of runs is $P(A)$ multiplied by the sequence length, in this case $0.5 \cdot 40$. Out of 124 students, 49 included at least one run of length four or more in their guessing. Of these, 14 students included a run of length five or more. Thus, few students think that clusters are as common as they actually are.

The probability of clusters arising in a binary sequence is comparatively easy to understand compared to the probability of clusters of for example diseases arising in a general population.

Therefore it may be a start to educate students of clusters in such a setting, when known biases can be used to give the students a better understanding of how they tend to underestimate the occurrence of clusters. It is also easier to show them their own bias when using production tasks.

Conclusion

Teacher students do significantly underestimate the number of long runs, i.e., runs of four or more equal symbols when trying to construct binary sequences that shall appear random. They also overestimate the number of runs of length one and two significantly. They do not seem to underestimate the number of runs of length three significantly; the median guess is actually the same as for a random sequence. Also, a *t*-test could not determine a significant difference when it comes to sequences of length 3.

For further research it would be interesting to ask participants to generate longer sequences, for example of length 150, where 75 runs would be expected, thus at least one run of at least seven, since the probability of getting a run of seven is $\frac{1}{2^{(7-1)}} = \frac{1}{64}$. The current results suggest that students would grossly underestimate the occurrence of the really long runs, but a sequence length of 40 does not allow us to conclude in this case. Another interesting experiment would be to tell some of the students about what constitutes a random binary sequence and tell some of them beforehand that you are to test their sequences on the *P*(A) and the number of runs of each given length. Then the results of this group of students could be compared to the students not having been given such instructions. Thus one could observe what effects such instructions would have.

It would also be very interesting to ask the students to explain their thinking during the production of these sequences. This would add a qualitative dimension to the results and be useful for improving the education in order to help students understand clusters better.

The author wishes to thank Trude Sundtjønn, Siri Krogh Nordby and Grethe Kjensli for help with data collection and George Hitching for comments.

References

Bakan, P. (1960). Response-tendencies in attempts to generate random binary series. *The American Journal of Psychology, 73*(1), 127–131. doi:10.2307/1419124

Bar-Hillel, M., & Wagenaar, W. (1991). The perception of randomness. *Advances in Applied Mathematics, 12*, 428–454. doi:10.1016/0196-8858(91)90029-I

Batanero, C., & Serrano, L. (1999). The meaning of randomness for secondary school students. *Journal for Research in Mathematics Education, 30*(5), 558–567. doi: 10.2307/749774

Batanero, C., & Sanchez, E. (2005). What is the nature of high school students' conceptions and misconceptions about probability? In G. Jones (Ed.), *Exploring probability in school: Challenges for teaching and learning* (pp. 241–266). New York, NY: Springer. doi:10.1007/0-387-24530-8_11

Bryant, P., & Nunes, T. (2012). *Children's understanding of probability: A literature review (full report)*. London, England: Nuffield Foundation.

Chiesi, F., & Primi, C. (2009). Recency effects in primary-age children and college students. *International Electronic Journal of Mathematics Education, 4*(3), 259–274.

Falk, R., & Konold, C. (1997). Making sense of randomness: Implicit encoding as a basis for judgement. *Psychological Review, 104*(2), 301–318. doi:10.1037/0033-295X.104.2.301

Fischbein, E., & Schnarch, D. (1997). The evolution of age of probabilistic, intuitively based misconceptions. *Journal for Research in Mathematics Education, 28*(1), 96–105. doi: 10.2307/749665

Garner, W. R. (1970). Good patterns have few alternatives. *American Scientist, 58,* 34–42.

Gilovich, T., Vallone, R., & Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology, 17*(3), 295–314. doi:10.1016/0010-0285(85)90010-6

Green, D. (1982). *Probability concepts in school pupils aged 11 – 16 years* (Doctoral dissertation, Loughborough University of Technology, Leicestershire, UK). Retrieved from https://dspace.lboro.ac.uk/2134/7409

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgement of representativeness. *Cognitive Psychology, 5*, 207–232. doi:10.1016/0010-0285(72)90016-3

Kirke-, utdannings- og forskningsdepartementet. (1999). *Matematikk: Felles allment fag i alle studieretninger.* Oslo, Norway: Kirke-, utdannings- og forskningsdepartementet. Retrieved from https://www.udir.no/globalassets/upload/larerplaner/felles-allmenne-fag/5/lareplan_matematikk.rtf

Kunnskapsdepartementet. (2013). *Læreplan i matematikk fellesfag: Fastsett som forskrift av Kunnskapsdepartementet 21. juni 2013.* Oslo, Norway: Kunnskapsdepartementet. Retrieved from http://data.udir.no/kl06/MAT1-04.pdf?lang=nno

Rapoport, A., & Budescu, D. (1992). Generation of random series in two-person strictly competitive games. *Journal of experimental Psychology, 121*(3), 352–363. doi: 10.1037/0096-3445.121.3.352

R Core Team. (2016). *R: A language and environment for statistical computing (version 3.3.2).* Vienna, Austria: The R Foundation for Statistical Computing.

Towse, J., & Mclachlan, A. (1999). An exploration of random generation among children. *British Journal of Developmental Psychology, 17*(3), 363–380. doi:10.1348/026151099165348

Utdanningsdirektoratet. (2012). *Læringsstøttende prøver: Matematikk 5. –10.årstrinn. Ressurshefte. Statistikk. Sannsynlighet. Kombinatorikk.* Oslo, Norway: Utdanningsdirektoratet.

Williams, J., & Griffiths, T. (2013). Why are people bad at detecting randomness? A statistical argument. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(5), 1473–1490. doi:10.1037/a0032397