The Entropy Method, or How to Detect the Hot Hand and Other Dependencies in Data

Ben Johnson-Laird and Phil Johnson-Laird

Everyone knows that athletes can get the ‘hot hand’ and have a streak in which they perform better than normal, or an awful slump. But, a trio of psychologists – Gilovich, Vallone, and Tversky – published an analysis in 1985 showing that two NBA teams showed no such effects in either field goals or free throws. Later in 1989, Tversky and Gilovich debunked the hot hand in a further article in this journal. Basketball players were not the only ones to be dumbfounded. Yet, belief in the hot hand has been blamed for all sorts of anomalous bets on outcomes in sports, games in casinos, and trades in the stock market.

In the probability calculus, if the conditional probability of a hit, such as field goal in basketball, given a preceding hit, differs from the overall probability of a hit, i.e., $P(\text{Hit}_n | \text{Hit}_{n-1}) \neq P(\text{Hit})$, then hits are *dependent* on one another. But, a robust effect calls for the difference to be statistically significant, i.e., to have too low a probability by chance *not* to be attributable to a dependency. So, analysts have used a variety of statistical techniques to search for hot or cold hands, including analyses of correlations. A few years ago, *Putting the ‘Hot Hand’ on Ice* in this journal used linear regression to take into account various aspects of ice-hockey games, such as the quality of the goalie, in order to show that players in the NHL were less likely to score after having just scored. And a recent study, *Surprised by the hot hand fallacy? A truth in the law of small numbers*, showed that for any finite sequence of hits and misses generated independently, the probability of a hit following a streak is less than the underlying

conditional probability of a hit. This proof undermines many studies of the hot hand, including the original one.

We needed a general procedure applicable to the analyses of dependencies in anything from a season of games to the choices people make in psychological experiments. The analysis of correlations was dubious – a significant correlation can occur in data that can be split into two subsets in which neither yields a reliable correlation. And the use of regression models is impossible in the absence of data about relevant factors. So, we developed a new method, which depends on an analysis of entropy.

The Entropy Method

Suppose you believe that the ‘hits’ (goals, runs, whatever) that a player makes in a sequence of games are independent of one another. But, you notice that a run of games occurs in which the player more hits than you expect, and in other runs of games no hits at all. And so you wonder about their independence, and you are right to do so. Is there a way to transform your intuition into a proper test?

There is: the method of entropy. The first step measures Shannon’s H , the entropy of a sequence of events, i.e., the degree of their disorder or unpredictability. And entropy is measured in bits: when it is high, the greater the amount of memory needed to store the sequence. The distribution of hits is unpredictable. When entropy is low, the sequence is more predictable: the hits come in streaks and slumps. But, we need to assess the statistical significance of the reduction in entropy. To do so, the

method's second step is to make a large number of Monte Carlo simulations of the sequence based on the overall probability of a hit. The final step is to determine the frequency with which sequences with entropies equal to or less than that of the original data occur in the simulations. A real sequence with significantly less entropy than its simulations has something constraining how it unfolds: the data are dependent. And the dependency in sport can be a streak or a slump or both.

To explain how to compute the method, let's start with a player's number of hits in each of a season of 38 games, and let's keep the number of hits low. In the first game, the player has 1 hit, and in the second game 0 hits, and so on, in this sequence:

101001001101100030111012001000000011021

In the season as a whole the player makes exactly one hit in 15 out of 38 games, so the probability of one hit is $P = .395$. The entropy, H , of a single hit in the i th game in the season equals the product of this probability with its logarithm to base 2:

$$P_i \log_2 P_i$$

The $\log_2$ of .395 equals: -1.341 , and the product of these two numbers is $-.53$, which is the entropy of the hit in the first game. Entropy, H , is an additive measure, and so the entropy of a season as a whole is the sum of the entropies of each of its games:

$$H = - \sum P_i \log_2 P_i$$

where the minus sign eliminates the negative values of logs of numbers less than 1.

The season above has the following entropy:

Score of 0 hits in 20/38 games: $P = .526$. $\log_2 P = -.927$. $P_i \log_2 P_i = -.488$

Score of 1 hit in 15/38 games: $P = .395$ $\log_2 P = -1.340$. $P_i \log_2 P_i = -.529$

Score of 2 hits in 2/38 games: $P = .053$ $\log_2 P = -4.238$. $P_i \log_2 P_i = -.225$

Score of 3 hits in 1/38 game: $P = .026$ $\log_2 P = -5.265$. $P_i \log_2 P_i = -.137$

Total H: $= -\sum P_i \log_2 P_i = 1.379$

So, the entropy for the season is: 1.379 bits. Suppose that the player's 22 hits had been more evenly distributed over the 38 games, so the four categories of 0 through 3 hits had these frequencies in the season as a whole: 6, 6, 5, and 5. Now, entropy equals 1.99 bits, just shy of the maximum of 2 bits for 4 categories. But, does the total entropy of the actual season show that hits over the season are dependent – that a hot hand or cold hand is affecting their distribution? Or, could the actual distribution over the season have occurred by chance? The next step in the method tells us.

Monte Carlo Simulations as Resampling Tests

The way to find out whether the entropy in the player's season departs from chance, and therefore contains streaks or slumps, is to run a Monte-Carlo re-sampling test. It uses the player's actual probabilities of making 0, 1, 2, or 3 hits in a game, and no other constraints, to simulate a season of games, and it makes a large number of such simulations, say, 10,000. The procedure is straightforward. Assign the four preceding probabilities to cumulative intervals in the unit line $[0, 1]$, i.e., the boundaries are:

For each game, simulate a player's performance by choosing a number at random from 0 to 1, and its value yields the number of hits from 0 to 3 according to the unit line above. Repeat this procedure 38 times, and it simulates a single season according to the assumption that the player's performance in any one game is independent from

performance in any other game. The entropy for a single simulated season yields a representative value on the assumption of independence. So, the mean entropy for 10,000 simulated seasons yields a very good estimate for a season of scores in games that are independent of one another. The proportion of simulations in which entropy is equal to, or less than, that of the actual season yields the probability of the actual season given the assumption of independence.

The Hot and Cold Boots of Soccer Players in the English Premier League

Now it can be revealed that the scores in the season above are the number of goals that best striker in the English Premier league of soccer, Mohamed Salah of Liverpool, scored in the 2018 season. Figure 1 is a histogram of his goals, and you can see that he started the season steadily, his boot warmed up for games 16 through 26, but then cooled off until the final streak starting in game 33.

Figure 1. The number of Mohamed Salah's goals in each of his 38 games in the 2018 English Premier League soccer season

Readers might wonder whether it takes an exceptional scorer to show significant streaks and slumps. So, we analyzed the records of the top 32 strikers in the Premier league for the same season. But, to put them on a level playing field, we normalized their entropies. If a player only ever scored 0 or 1 goal in any game, then his maximum entropy is 1 bit per game, whereas if a player, such as Salah, scored 0, 1, 2, or 3 goals, then his maximum entropy is 2 bits per game. So, we normalized the players' entropies by dividing each of them by their respective maximum entropies. Table 1 presents these normalized entropies of the 32 best strikers in 2018, the mean of the normalized entropies of 10,000 simulations, and the difference between them, i.e., the reduction in entropy, if any, of their actual seasons. Figure 2 is a histogram of the reduction of the normalized entropy of the 32 strikers' actual seasons in comparison with that of their 10,000 simulations. To answer our earlier question: no reliable correlation occurred between the number of goals a player scored and their tendency to have a hot boot (Kendall's rank-order correlation coefficient, $\tau = -.18$, $P > .9$).

Like Salah, some strikers did have streaks and slumps, and others did not. As Table 2 and Figure 2 illustrate, 23 out of the 32 players' seasons had a reduction in entropy consistent with the occurrence of streaks and slumps. The tendency to have such a reduction was highly significant: a Monte Carlo resampling procedure yielded $P < .001$. So, despite the variability, strikers in professional soccer at the highest level tend to have dependent games of scoring or not scoring – hot boots or cold.

Player's name	Number of goals	Normalized entropy	Mean entropy in 10K simulations	Reduction in entropy of actual season
1. Alexandre Lacazette	13	0.674	0.766	0.091
2. Raul Jimenez	13	0.650	0.741	0.091
3. Ashley Barnes	12	0.636	0.721	0.085
4. Luka Milivojevic	12	0.628	0.713	0.085
5. Salomon Rondon	11	0.658	0.742	0.084
6. Anthony Martial	10	0.685	0.763	0.078
7. Leroy Sane	10	0.641	0.718	0.077
8. Wilfried Zaha	10	0.610	0.686	0.076
9. Roberto Firmino	12	0.547	0.618	0.071
10. Ayoze Perez	12	0.520	0.589	0.069
11. Gerard Deulofeu	10	0.526	0.583	0.057
12. Lucas Moura	10	0.505	0.559	0.054
13. Mohamed Salah	22	0.689	0.744	0.055
14. Raheem Sterling	17	0.678	0.706	0.028
15. Eden Hazard	16	0.623	0.649	0.026
16. Glenn Murray	13	0.673	0.686	0.013
17. Gylfi Sigurosson	13	0.673	0.686	0.013
18. Harry Kane	17	0.820	0.832	0.012
19. Richarlison	13	0.702	0.713	0.011
20. Callum Wilson	14	0.777	0.786	0.009
21. Son Heung-min	12	0.719	0.727	0.008
22. Chris Wood	10	0.576	0.580	0.004
23. Marko Arnautovic	10	0.691	0.692	0.001
24. Jamie Vardy	18	0.826	0.813	-0.013
25. Sergio Aguero	21	0.834	0.820	-0.014
26. Aleksandar Mitrovic	11	0.605	0.589	-0.016

27. Joshua King	12	0.668	0.651	-0.017
28. Romelu Lukaku	12	0.704	0.685	-0.019
29. Marcus Rashford	10	0.885	0.864	-0.021
30. Paul Pogba	13	0.680	0.658	-0.022
31. Sadio Mane	22	0.893	0.868	-0.025
32. Pierre Aubameyang	22	0.890	0.865	-0.025

Table 1. The 32 best strikers in the 2018 English Premier soccer league, the numbers of goals they scored in the season, the normalized entropies of their season, its mean values in 10,000 simulations of the season based on the probabilities of the numbers of goals per game, and the difference between them, i.e., the reduction in entropy, if any, between their actual season and these simulations.

Figure 2. The reduction, if any, in the entropy of the actual seasons of the 32 best strikers in the 2018 Premier league season in comparison with 10,000 simulations for each player.

Do People Seek Independent Tests of Hypotheses?

The entropy method has widespread applications. We illustrate them with an example remote from sports. In a well-known psychological task, which the late Peter Wason devised, participants have to select potential evidence to test a given hypothesis. His idea was to model scientific thinking and to examine whether individuals grasp the relevance of falsification. In the original test, the hypothesis was an abstract one about letters and numbers:

If a card has an E on one side then it has a 4 on the other side.

The participants knew that every card has a letter on one side and a number on the other side; and their task was to make a parsimonious selection from four cards of those they needed to turn over to find out whether the hypothesis was true or false about these cards:

Most people selected the E and 4 cards, some selected E alone, and a few selected E, 4, and 7. What was remarkable was the neglect of 7. Yet, if it has an E on its other side, then the hypothesis is false. This finding of a neglect of evidence that could falsify a hypothesis sparked a surge in experiments and theories. At least 16 different theories of the task existed at one time, and the equivalent of the population of a small town – over 14 thousand people – was tested in over 200 experiments. These studies

also examined sensible everyday hypotheses, and principles governing permissions, and both elicited many more potentially falsifying selections than the abstract task.

One theoretical issue was whether people select each piece of evidence independently from the others, i.e., they decide whether or not to select each of the four cards, E, K, 4, and 7, independently of their other selections. Over half the theories of people's performance implied that they do. To test independence, some investigators examined correlations between the selections of different pairs of cards, but their results were mixed.

The entropy method answered the question. With four pieces of potential evidence, e.g., the cards above, there are 16 possible selections, including selecting all four cards or none at all. In fact, the entropy for the 99 experiments that reported the frequencies of all 16 selections was often much smaller than the maximum of 4 bits. Suppose that in an actual experiment, 244 participants made the following numbers of selections:

E & 4: 116

E: 62

E & 7: 41

E, 4 & 7: 21

K: 4

The individual probabilities of selecting each of the four cards are: $P(E) = .984$, $P(K) = .016$, $P(4) = .561$, $P(7) = .254$. To simulate independent selections, a program makes each individual selection by deciding whether or not to select each card according to

these probabilities, i.e., it make independent selections. It repeats the process 244 times to simulate the entire experiment. And it simulates the experiment 10,000 times in this way. The entropy of the original experiment is: 1.846. Its mean value for the simulations was 2.00, and in 9,765 of them it was larger than that of the actual experiment, so the probability it yielded dependent selections is $P < .025$. In the 99 experiments using the task, the overall differences between their mean entropies and those of their simulations were highly significant: 2.33 vs. 2.60 for abstract hypothesis, 2.43 vs. 2.75 for sensible everyday hypotheses, and 1.57 vs. 1.80 for rules about permissions ($P < .0001$ in all three comparisons). The smaller entropy in actual performance shows that the participants' selections were dependent. For example, they tended not to select the 7 card in the abstract version of the task unless they selected the E card. The entropy method eliminated nine theories of the task.

One unexpected finding was a difference between those experiments carried out face-to-face and those carried out on-line. The face-to-face experiments had a smaller entropy than that of the on-line experiments. And this difference was highly reliable. So, the selection of evidence in the on-line studies was more unpredictable than in the face-to-face studies. Participants tested on line are likely to care less and to pay less attention. Indeed, unlike those in the face-to-face studies, some of them selected no evidence whatsoever. Comparisons between the two ways of carrying out experiments have seldom reported any systematic differences between them. The absence of such reports may be because investigators haven't examined the entropies of the two sets of data.

Conclusion

If you need to test whether or not there's a dependency between events, then the method of entropy, which combines Shannon's measure, H , of entropy with Monte Carlo re-sampling is helpful. It shows that the hot boot occurs in soccer, that people think about potential evidence in dependent ways, and that it can matter whether an experiment is carried out on line or face to face.

Further Reading

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

Miller, J. B., and Sanjurjo, A. 2018. Surprised by the hot hand fallacy? A truth in the law of small numbers. *Econometrica*, 86, 2019-2047.

Ragni, M., Kola, I., and Johnson-Laird, P. N. 2018. On selecting evidence to test hypotheses. *Psychological Bulletin*, 144, 779-796.

Shannon, C. E. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27, 379-423.

Tversky, A., & Gilovich, T. (1989). The 'hot hand': Statistical reality or cognitive illusion?. *CHANCE*, 2, 31-34.

Vesper, A. (2015). Putting the 'Hot Hand' on ice. *CHANCE*, 28, 13-18.

Premier League data are from <https://fbref.com>. Its use is allowed from its parent website: https://www.sports-reference.com/data_use.html It states: "As an aside, copyright law is clear that facts cannot be copyrighted, so you are free to reuse facts found on this site in accordance with copyright laws."

Appendix

About the authors

Ben Johnson-Laird is a software engineer and data scientist. He received his BSc (Hons) in Computer Science from the University of Kent, and his MPhil in Computer Speech and Language Processing from the University of Cambridge. His chief extra-curricular interest is in listening to vinyl records of classical music and modern jazz.

Phil Johnson-Laird is Stuart Professor of Psychology, Emeritus, Princeton University, and a Visiting Scholar in the Psychology Department of New York University. He is a member of the National Academy of Sciences and the American Philosophical Society, and a fellow of the Royal Society (of London). His main research is into the role of mental models in reasoning. He too listens to classical music and modern jazz, but not on vinyl.