# Probability in Decline

Dean M. Brooks
Ekaros Analytical Inc., Box 408, 125A-1030 Denman Street, Vancouver, B.C. V6G 2M6
Blog: www.declineeffect.com
Declineeffect@gmail.com
JSM 2010 abstract #308013

## Abstract

Spencer-Brown advanced a provocative thesis in *Probability and Scientific Inference* (1957). From experiments with early "chance machines," he argued that a fundamental flaw exists in our view of randomness. Long sequences of random digits generated by a variety of methods show long-term declines in repetition of rare items or sequences.

I revisit this neglected topic and show that a variety of modern random number generators (including pseudo-random algorithms) exhibit this same property. For simple schemes the decline is short and the system soon lapses into classic equipartition. For sufficiently complex schemes the decline continues indefinitely. Standard tests like DIEHARD do not detect this pattern.

I suggest the principle of maximum entropy as the underlying cause. Similar decline patterns show up empirically in epidemiology, Web traffic, and other probabilistic settings.

**Key Words:** maximum entropy, decline, epidemiology, Web traffic, Spencer-Brown, randomization

## 1. Spencer-Brown's Critique of Probability

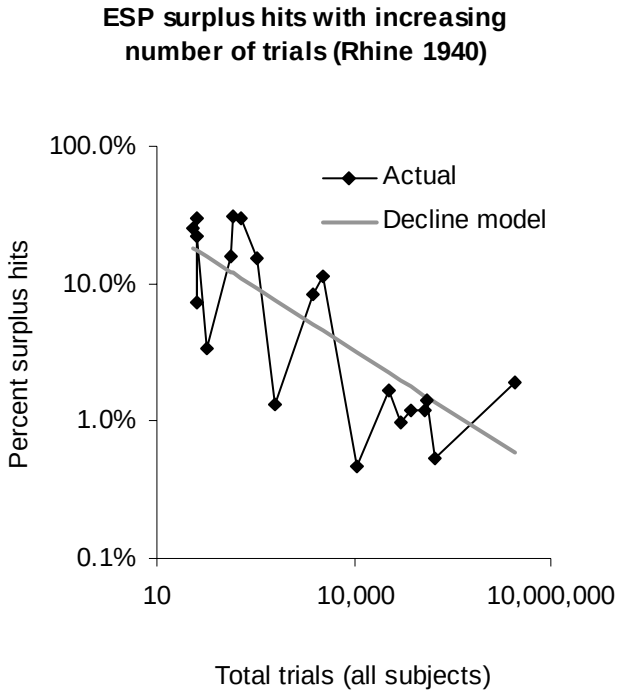
In the 1950's, experiments in extra-sensory perception (ESP) were still novel enough that they could be written up in *Nature* and other reputable journals. However, several decades of work on the subject had so far failed to produce a convincing demonstration of a stable talent for clairvoyance or telepathy or any other variant of ESP, and hopes were fading that any would be found.

George Spencer-Brown, who did his post-graduate work under both Bertrand Russell and Ludwig Wittgenstein, considered the possible reasons for the failure of ESP experiments and turned them into a novel critique of the classical foundations of probability. His argument first appeared in *Nature* in 1953, then as a monograph, *Probability and Scientific Inference* (1957).

The puzzle, as Spencer-Brown observed, wasn't that ESP experiments were unrepeatable. It was that they failed to be repeatable in much the same way every time. Their failure was itself a predictable pattern. Initially the subject would score well above chance, but after a few dozen trials, or a few hundred, the margin of superiority and the significance of the subject's score would both sink. Eventually many subjects lapsed into 'psi-missing,' that is, their rate of successful guessing would be not at, but rather significantly below the rate predicted by classical chance. According to ESP researcher Robert Thouless, such declines had been observed since sometime in the 19th century:

> It is not easy to give a date for the first discovery, although they were first singled out as a significant feature of the ESP response by Rhine in his 1934 book Extra-Sensory Perception. They had, however, been noticed earlier. Of the Creery sisters, for example, it was reported that 'the average of successes gradually declined' (Gurney et al., 1886). A similar decline was also pointed out by Estabrooks in an early study of ESP (Estabrooks, 1927). Since then, decline effects (both episodic and long-period) have been found by so many workers that one must regard decline as one of the best attested and most often repeated observations in ESP research.

The accepted explanation for this pattern (among scientists who were not ESP enthusiasts) was the 'file drawer' effect, in which successful experiments with high significance were reported, but unsuccessful ones with low significance were simply tossed aside or put in a file drawer. Extending or repeating the occasional high-significance result would prove impossible and average scores would appear to decline, but overall, there was no such thing as a decline in guessing ability.




In 1940, Joseph Banks Rhine constructed a table summarizing dozens of different ESP experiments. Rhine was and remains today the best-known writer on the scientific study of ESP. The purpose of Rhine's meta-analysis was to compare various test schemes to see which succeeded best at delaying the eventual lapse of the subject's talent. What is interesting is that if we simply take the entire set of results, and plot the decline of significance on a log-log scale, we get something close to a straight line, as shown above. This doesn't follow in any obvious way from the 'file drawer' effect.

Spencer-Brown hypothesized that the failure to find convincing evidence of ESP ironically exposed a real scientific puzzle worth solving. If there was no ESP, then there must be something lacking in probability theory.

> This is quite plausible in light of the fact that psychical research is perhaps the only present-day science which has looked for something (not already known to exist) for sixty years and failed to find it; and if it happened that what it was looking for did not exist, we should have in effect sixty years of pure probability experiments which there is no reason to suppose should have fared, in terms of significance, better than the best (and the worst) of all the pure probability experiments down the ages. It would thus be its remarkable additions to our experimental picture of pure probability for which we owe the most thanks to modern psychical research.

The book and *Nature* essay did not have much impact compared with his later classic *Laws of Form*, but the questions Spencer-Brown raised still remain relevant today.

### 1.1 Philosophical Concerns About Classical Probability

Spencer-Brown attacked conventional probability from two angles. First, he attempted to demonstrate its philosophical contradictions and inadequate foundation. This approach owed something to the influence of Wittgenstein. One especially useful line of inquiry, according to Spencer-Brown, is the proper way to distinguish between 'atomic' and 'molecular' events. Classical probability focuses on the 'atomic' level. For example, if we throw a six-sided die 100 times, we treat this as 100 independent events. To work out the likelihood of two successive results of '6', we combine these 'atomic' events. But we do not, as a rule, make observations of 'molecular' events to see if their frequencies actually conform to theory. We are confident that for a fair die, the six sides will over a long span of time come up equally often. This is thought to guarantee that over an even longer span, all possible permutations on the 'molecular' level will do so as well.

But this is begging the question, according to Spencer-Brown. It assumes independence instead of proving it empirically. A non-classical theoretical framework may give a different answer, one that fits the actual facts much better. By choosing to experiment and reason within the classical framework, we set up the expectation of a certain kind of result.

It was widely understood in the 1950's that the state of the art in 'chance machines' at that time was quite primitive. Mechanical methods like applying a strobe light to a spinning wheel necessarily involved many different risks of bias. In practice, any project of generating random numbers involved re-processing whatever sequences were produced, by some kind of algorithm, and throwing out some of the data. For example the first attempt by the RAND Corporation in 1955 to publish a table of one million random digits involved extensive re-processing, as the original data set showed unacceptably large bias in variables as simple as the balance between even and odd values. It was specifically noted by the RAND researchers that the various biases evolved over time: 'Apparently the machine had been running down despite the fact that periodic electronic checks indicated it had remained in good order.'

This kind of problem would show up in any long series of machine-generated 'atomic' events; the derived, 'molecular' outcomes would not show up in equal numbers. There would be an observable bias, typically one that changed between the beginning and the end of the run. The more complex the 'molecular' event, the greater these biases tended to be.

The philosophical perspective that probability is genuinely ‘atomic’ influenced what data researchers would do with these results. Because early ‘chance machines’ were not able to produce truly unbiased data, the standard for genuine randomness had to be set fairly low. Researchers gradually abandoned the effort to purge randomly generated sequences of all discernable bias, seeing this as practically impossible and theoretically unnecessary. They settled instead for monitoring and controlling bias in the ‘atomic’ events, or at most for the very simplest ‘molecular’ events. But what they could have done, according to Spencer-Brown, was to reconsider whether their concepts regarding randomness were correct, and to look more carefully at the raw output from the machines for clues about how randomness really works.

There is no reason in principle why a series cannot gradually become less and less biased at the ‘atomic’ level, but remain biased on the various higher ‘molecular’ levels for arbitrarily long spans.

For example, in throwing a six-sided die it is theoretically possible to arrange the results such that there are nearly equal numbers of results ‘1’ through ‘6’, and at the same time a growing shortage of ‘11’ through ‘66’ relative to results like ‘25’ or ‘34’. This particular effect has actually been observed by gamblers, and reproduced as part of the work described in this paper. Many similar kinds of decline in the rarer ‘molecular’ outcomes have also been observed as described below.

### 1.2 The Empirical Case in 1957

There was a substantial body of evidence available to Spencer-Brown. After noting a variety of cases like those mentioned above in which ‘chance machines’ produced a certain predictable pattern of bias similar to the ESP decline effect, Spencer-Brown proceeded to run his own series of experiments. These were of the general ESP type, but without a human test subject, and showed the same result. For example, instead of having a subject guess the values of a deck of cards, Spencer-Brown proposed that we use a second deck of cards to simulate the guesses. The question of the subject growing fatigued or having an erratic ESP talent is then moot, and the whole topic of ESP is irrelevant.

These experiments pointed to a common anomaly: rare items would cluster near the start, then gradually grow rarer. The simplest way to demonstrate this trend, given the diverse range of experiments to be considered, is to compare quartiles. Even in a short test with a few hundred trials, the first quartile of the test will tend to have significantly more rare items ($p < 0.05$) than the last quartile. Critics attacked the experimental aspect of Spencer-Brown’s work starting with the *Nature* paper, but then a controlled test performed by a critic showed the same result.

## 2. More Recent Empirical Cases in Nonlinear Probability

Spencer-Brown's theorizing did not produce much of a response. His non-classical, non-linear framework of probability was, so to speak, childless. But explaining failed ESP experiments is far from the only application of his idea. There are dozens of what we might call 'orphan' laws in need of explanation, *ad hoc* empirical patterns that have been known for decades (or even centuries) and that remain outside the classical framework. The childless theory and the orphan empirical laws might yet form a family.

Nonlinear probability is a broad term, really a matter of definition by negatives. It is *not* the familiar linear probability that is taught and used in every university. Nonlinear probability distributions will *not* satisfy the usual criteria applied in probability work, such as conformity to the central limit theorem; nor will it exhibit the usual properties of linear systems. Nonlinear distributions are diverse, each case potentially *sui generis*.

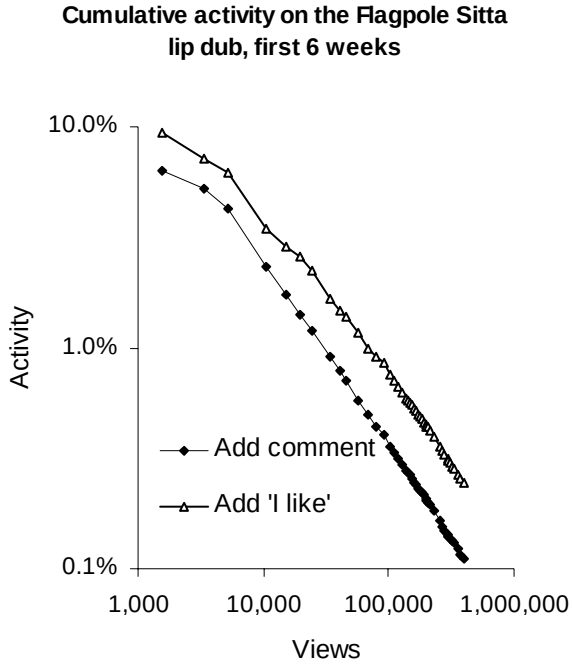
In nature, though, we find one outstandingly common pattern, across a broad range of fields including species abundance, cellular metabolism, economics, epidemiology, Web traffic, military history, voting, participation in religion, and much more. The common pattern in this field is logarithmic decline in rare events with increasing set size, in the manner of the plot from Rhine (1940). The form of the decline when plotted on log-log axes is a straight line.

For example, take Smeed's Law. This is a pattern governing traffic accident fatalities, first observed in 1949. Smeed observed that traffic fatalities per capita in different countries decrease in a very orderly way as the absolute number of drivers increases. The relation is a power law of the form $p(\text{fatality/year}) = kN^a$, where N is the number of drivers, a is roughly 0.7, and k is a constant.

The literature on Smeed's Law, not very abundant, offers no convincing explanation for this relationship. It is an 'orphan' law, an *ad hoc* observation that awaits integration into the broad system of scientific knowledge.

Another more recent example is the anomalous decline in participation by members of online communities. For example, take video sharing websites like YouTube. Most such sites keep track of how many viewers have seen a particular item. They also allow viewers to submit a comment, or to press a button saying they like or dislike it. The near-universal pattern on such sites is for the rate of commenting to decline relative to the rate of viewing. As audiences grow from 10 viewers to 10,000 to 10 million, the rate of commenting plunges, typically from 10 percent to 1 percent to 0.1 percent or less. These dramatic swings in interest are made even more mysterious because they are so orderly.

The example below is from the Vimeo video sharing website in 2007. The video is known as the 'Flagpole Sitta Lip Dub' and featured an office full of people singing along with a pop song. It earned more than 400,000 views in its first six weeks. The participation rate as a percentage of cumulative viewers to date dropped in very orderly fashion. Two rates are shown, for people who wrote a comment, and those who pressed the 'I like' button.




The comments and ‘I like’ responses grow steadily rarer over the six-week debut of the video, falling so rapidly that the rate is noticeably different in the morning and the afternoon on any given day. This law of mass participation is of tremendous interest and practical relevance.

Another highly interesting application is in epidemiology. One longstanding problem in that field is how to know, at the start of an epidemic, what the mortality rate is likely to be over its course. The standard approach is to assume a stationary mean, or at most a series of stationary means for different risk groups. That is, for adults 18-45 there will be an average mortality rate that is observable at the beginning, middle, and end. The mortality rate may vary for children or the elderly when considered separately, but these too will be stationary averages. Likewise for the transmission risk, or reproduction number R commonly used by epidemiologists.

The historical record does not bear out these assumptions. For many diseases there is evidence of mortality falling between the start and end of an epidemic, as well as evidence of declining transmission rates. The cases go back centuries.

The standard epidemiological model is not of much help in such situations. A nonlinear probability scheme can perhaps do better. The global H1N1 pandemic in 2009 offered a chance to validate a log-log probability model.

In May 2009, the first reports emerged from Mexico of an outbreak of the H1N1 virus that had frighteningly high mortality rates. The medical personnel attending the first few dozen cases were coming down with the disease, and Internet rumors claimed one or more had already died.

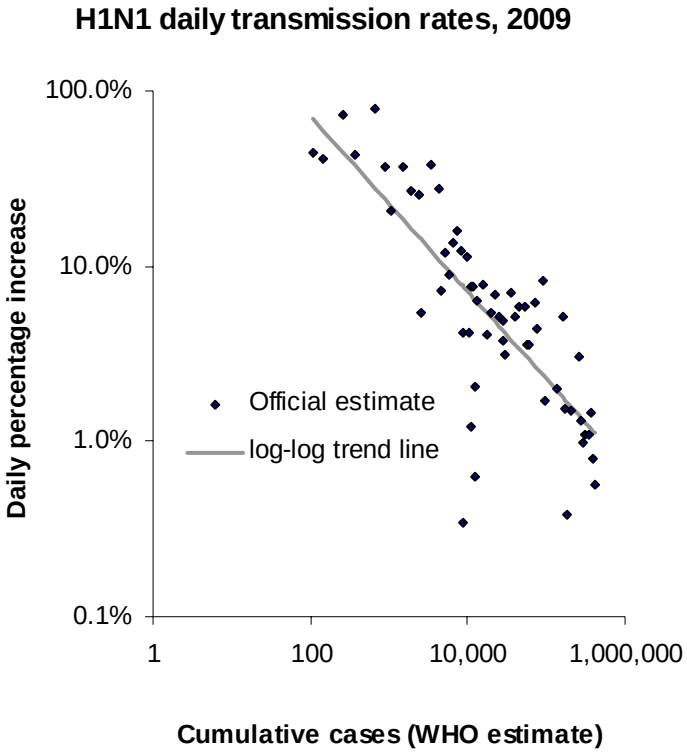
The Mexican government and the World Health Organization issued initial statements about the outbreak that were only a little less alarming than the rumors. But week by week, the picture became more optimistic.

**Table 1:** Selected falling mortality estimates for H1N1 pandemic, 2009

| *Mortality* | *Case count* | *Date* | *Issuing authority* |
|---:|---:|---:|---:|
| 9.6 % | 73 | 30 Apr 2009 | WHO lab confirmations Mexico |
| 5.3 % | 3,000 | 30 Apr 2009 | Mexican government |
| 1.7 % | 4,900 | 31 May 2009 | WHO lab confirmations Mexico |
| 0.8 % | 17,400 | 31 May 2009 | WHO lab confirmations global |
| 1.4 % | 8,300 | 30 Jun 2009 | WHO lab confirmations Mexico |
| 0.7 % | 52,000 | 30 Jun 2009 | WHO lab confirmations global |
| 0.15 % | 30 million | 15 Aug 2009 | U.S. Presidential Panel |
| 0.9 % | Unknown | Sep 2009 | Media reports |
| 0.5 % | Unknown | Oct 2009 | Malaysia |
| 0.02 % | 22 million | Nov 2009 | Centers for Disease Control |

The spread of the epidemic also failed to live up to the first official forecasts.
The original expectation was that H1N1 would spread until a billion people, or perhaps several billion, had caught it. This followed from the standard epidemiological assumption of constant, stationary transmission rates. The only force that is recognized as restraining spread of an epidemic (once it has escaped quarantine) is acquired immunity.

Here, just as in the Vimeo video example, the behavior of the virus was noticeably different from day to day. The points on the graph cover the period from April 27 through July 6. After that date, the WHO ceased to provide country-by-country coverage of specific cases and shifted to regional estimates. Even for the period covered by WHO bulletins, we can be sure that the data increasingly understate the real transmission rate as more and more cases were going unrecorded. However, the trend is obviously not stationary. The final size of the epidemic (in early 2010) is unknown, but most likely between 100 and 200 million cases.




Precisely why these declines occur, that is, the medical mechanisms involved, is beyond the scope of this paper, but a promising lead is a measurable reduction in the 'viral load' of later patients. The average number of viral particles per millilitre of blood is smaller for successive cohorts. This reduces the severity of symptoms (lower mortality) as well as the efficiency of transmission.

There are similar trends to be found in criminology, voter turnout, adoption of new products, religious participation, and dozens of other fields. Thus reviving Spencer-Brown's 1957 project in 2010, and developing a general methodology of nonlinear probability, is far from being an idle or purely mathematical diversion. If we can understand this decline process better, there are enormous practical benefits waiting.

## 2. The Bayesian, Maximum-Entropy Framework of Jaynes

Coincidentally, in the same year that Spencer-Brown was suggesting that classical probability lacked a proper foundation and had unsolved empirical problems lying on all sides, the physicist Edwin Jaynes put forward a new framework in which a solution

could be found, the ‘principle of maximum entropy’. Here is how the principle is described byWikipedia:

> Let some testable information about a probability distribution function be given. Consider the set of all trial probability distributions that encode this information. Then, the probability distribution that maximizes the information entropy is the true probability distribution with respect to the testable information prescribed.

This is essentially a Bayesian approach to empirical problems in nonlinear probability. In his initial 1957 paper, Jaynes first proposed that lacking a detailed model, or when faced with multiple possible dimensions of measurement for a complex system, one could use as a ‘prior’ for a given system whatever distribution maximizes uncertainty about the outcome.

For a simple, classically linear case like an unbiased six-sided die, Jaynes noted that the long-run maximum entropy distribution is the same as the classical distribution. Our uncertainty about the outcome is maximized by setting the likelihoods for the six sides equal to one another. There is thus no conflict between the two frameworks. At least, Jaynes believed there was none, in the long run. As it turns out, experimentally there is a difference. But this is good news for Jaynes and not so good news for the classical approach, as Jaynes’ approach is adaptable to the observed facts of decline.

Moreover, for any more complex nonlinear case, lacking a detailed account of the mechanism or a classically linear model, we are often at a loss how to proceed. We do not have even a plausible approximation. This is where Jaynes’ method offers support and insight.

One aspect of Jaynes’ work that is vital for our purposes is that there is no preferred perspective, no single way of looking at the data that qualifies as being causative in a way that other perspectives are not. Thus we would not focus necessarily on the ‘atomic’ sequence of individual throws of a die. It would be just as appropriate to characterize the system in terms of two-throw groups, or four-throw groups, or some other ‘molecular’ configuration. The theory is not a theory of what is physically happening, so much as a theory of what observations we can hope to make of the system. One method of making observations is in principle much like another.

This perspective helps to overcome the difficulties raised by Spencer-Brown in 1957, about experimenters throwing out significant patterns of behavior by their ‘chance machines’.

## 4. Basic Procedure for Testing for Decline in 'Chance Machines'

To properly assess whether a given 'chance machine' exhibits decline requires some novel test procedures and a different perspective on what data are to be kept or thrown out. Here are some key points:

1. Careful study must be made of the apparatus and its possibilities for compound or 'molecular' events. These exist along two dimensions. A single die thrown multiple times yields sequential compounds like '11' and '34'. Two dice thrown together yield simultaneous compounds as well as sequential ones. Dice or coins or cards should always be distinguishable, so that for example if a red die and a blue die are thrown together, the sequential compounds for the red die can be identified as distinct from those for the blue die.

   Some experiments will have to be done with a person in the loop. For example, a slot machine in which the precise moment that the lever is pulled influences the random number selection process. In such cases we must consider each subject-machine combination to be a distinct apparatus. Decline will occur for each combination. If multiple experimenters play the same machine, their results must be distinguished in the analysis.

   The apparatus should also be studied for its axes of internal symmetry. These turn out to be quite numerous and relevant. For example, an American roulette wheel has 38 spaces. Because of symmetry, we can also think of the wheel as evenly divided into 19 'double' spaces, taking two adjacent spaces as one. European roulette wheels only have 37 spaces and so lack this internal symmetry. When determining the odds of compound events and their relative rarity, this kind of internal symmetry can have a large impact, particularly because of rule #2 below.

2. The rarer the item in nominal, classical terms, the greater the surplus will generally be at the start, and the longer the series must be before occurrences of the item slip below the classical expectation. The length of the sequence needed to exhaust the surplus is roughly proportional to the odds of the event. Thus if the apparatus is a pair of brand-new dice, and the test is for occurrences of 'doubles,' then the surplus of doubles will usually be exhausted after several hundred throws. If the apparatus is three dice, and the test is for occurrence of 'triples,' the surplus will take longer to exhaust, perhaps thousands of throws instead of hundreds.

3. Every trial, especially early trials, must be recorded. The more usual practice with a newly constructed or acquired random number generator is to 'run it in,' and not to start sampling data until many hundreds or

thousands of results have been generated. This will undermine the purpose of the experiment, as it will not only limit the study to data from the later, much slower period of decline, when rare items have already become much rarer, but itwill also render moot the full size of the set of results.

Failure to do this will not make the tests useless, merely limited. For example, in one early test, some exploratory experiments were done with a 12-sided die but not fully recorded. Then a long series of 1,500 throws was made looking for repeats. As expected, the series of 1,500 throws under-repeated. The rare event of a repeat had already declined below the classical level, so that only 97 repeats ($p < 0.005$) were obtained. It would clearly have been better, however, if the result in this case could have been compared with the early part of the run in which repeats were more plentiful.

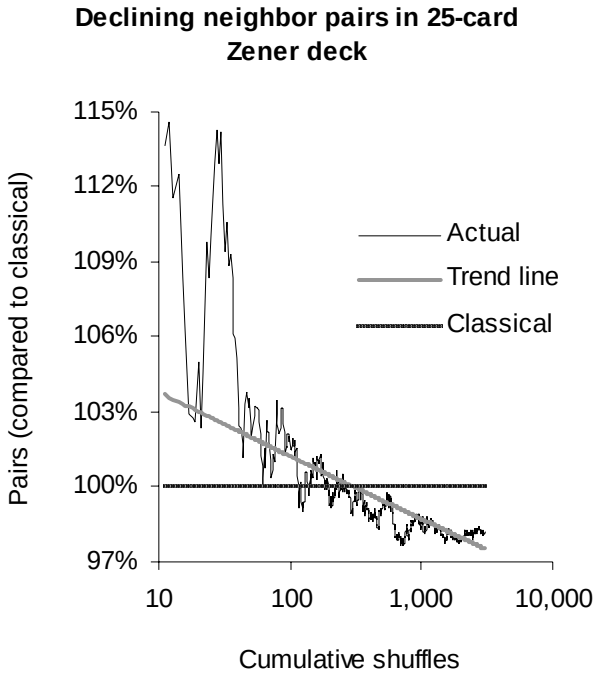
4. Trends should be evaluated cumulatively. Because our hypothesis is that the system is accumulating entropy, the variable of greatest interest is the rate of rare occurrences over the full set of results. As they become rarer, these events effectively increase the overall entropy of the system.

   This rule of evaluation in cumulative context contrasts strongly with the state of the art in random number generation at present. The suite of DIEHARD tests endorsed by the National Institute for Standards and Testing (NIST) works on arbitrarily sampled blocks of data, typically 1,000 bits in length, to determine if there are, for example, too many long strings of 111111 or 000000.Unfortunately, a proper test can only be done contextually, not in relation to the string's immediate neighborhood but in relation to the whole output history of the RNG to date. Thus standard testing regimes are not suited to finding the phenomenon under study here.

   This is important because declines do show up in purely digital random-number generation schemes. In one case, several years of output from a brand-new Keno game based on a digital RNG was found to exhibit significant long-run bias—not of a kind that players could exploit to win the game, but of a kind that conformed to the decline hypothesis and that ought to raise doubts about the stability of the algorithm. If a test procedure never considers long-run trends of this kind, then long-run stability issues can scarcely be discussed.

   It is important as well because simulations of physical games often exhibit the same kind of decline as the physical game itself. A deck of 25 cards with five symbols, when shuffled, should yield an average of four 'repeats' per shuffle, in which successive cards have the same symbol. Digitally simulated decks using a spreadsheet show a small but significant decline,

as do actual decks. Here is the cumulative plot of decline in a simulated deck shuffled 3,000 times:




The net shortage after 3,000 shuffles was 213 missed repeats, or 1.8 percent. This was hardly something that would be noticed casually, but it was significant.

5. Apart from graphing the data cumulatively, another handy standard format for comparison is by quartiles. This was common practice in ESP studies in the early 20th century and it remains very useful today. Whether the data consist of a few hundred dice throws, or a few hundred thousand 'I like' responses from viewers of a video, a comparison of quartiles should yield a common pattern. The difference between the first and fourth quartiles will generally be the largest. The exact difference will depend on the particular apparatus and the rare items being tracked.

## 5. Summary of Recent Testing

The purpose of testing conducted between 2003 and 2009 was to extend the range of different ‘chance machines’ considered, beyond the fairly limited number that Spencer-Brown tried. Virtually all of these have generated significant ($p < 0.05$) declines of the expected kind. In every category there has been third-party data generated prior to the experiment, sometimes in great quantity, that also showed the same type of decline.

— 4-sided, 6-sided, 8-sided, 10-sided, 12-sided, and 20-sided dice
— decks of cards (52-card, 25-card Zener, multi-deck Baccarat, custom decks)
— simulated decks of cards (all the same types above)
— coins (spun on a table or flipped in the air, singly or in groups)
— digital Keno games (online and in casinos)
— roulette wheels (in casinos and using smaller replicas)
— slot machines (casino, online, PC software simulator)

The significance of the results will vary. Occasionally the difference will only yield $p < 0.10$, or even $p < 0.20$. Differences of $p < 0.05$ to $p < 0.01$ are quite common. Extremely large differences of $p < 0.00001$ or smaller have occurred numerous times.

Typically the decline is not large enough to overcome the house edge in any game of chance. Sometimes the decline is striking but not really relevant to the game. For example, a 2 percent shortage of pairs in an eight-deck blackjack shoe will not give either the house or the player any sort of meaningful advantage, despite the overall house edge being smaller than 2 percent.

In preparing this paper I am conscious of the extreme skepticism that is likely to be brought to bear by readers on any challenge to classical probability.

It should be kept firmly in mind that the strength of the argument for decline does not lie in any one test having high significance, but rather in the difficulty of finding ‘chance machines’ that do *not* exhibit decline. A dozen experiments using different apparatus, half of them based on data collected by third parties, each with an expectation of zero long-run bias, each with decline of significance $p < 0.1$, constitute a more compelling argument than a single wildly improbable result. A hundred such experiments are, for me at least, enormously provocative.

Also, as I have tried to show, ultimately the motive for investigation is not simply to revisit half-century-old concerns about the classical model being wrong, but to acknowledge the present challenges that we face in interpreting and plotting a wealth of highly relevant phenomena, like Web traffic and the spread of H1N1 virus.

These experiments, if done correctly, can only end by strengthening our understanding of probability. I hope that readers will be encouraged to pursue the topic on their own. Ultimately, science needs to understand this, wherever it might lead, and whatever 'sacred cows' it might gore along the way.

## Acknowledgements

I would like to thank Milo Schield of Augsburg College, who chaired our session at JSM 2010. Milo has offered great encouragement for my work over the years and I am grateful for the invitation to speak.

The 90 or so mathematicians in attendance at the talk (at 8:30 on a Monday morning) came up with a number of good questions afterward and I thank them for their lively interest in the subject.

## References

Jaynes, E. T. (1957). "Information Theory and Statistical Mechanics". *Physical Review* Series II 106 (4): 620–630. doi:10.1103/PhysRev.106.620. MR87305

Rhine, J. B. *Extra-Sensory Perception After Sixty Years* (1940)

Spencer-Brown, G. *Probability and Scientific Inference* (1957)

Thouless, R. *From anecdote to experiment in psychical research* (1972)