

Remarks on “Random Sequences”

Branden Fitelson* and Daniel Osherson**

*Department of Philosophy, Rutgers University

**Department of Psychology, Princeton University

Abstract

We show that standard statistical tests for randomness of finite sequences are language-dependent in an inductively pernicious way.

1 SETUP

We consider evidence relevant to whether a (possibly idealized) physical process is producing its output randomly. For definiteness, we’ll consider a coin-flipper \mathcal{C} which reports “H” for heads and “T” for tails. By \mathcal{C} producing its output “randomly,” we mean H and T have equal probability and trials are independent. If \mathcal{C} produces its output randomly (in the above sense), then we’ll say that \mathcal{C} is a *random device*.

There are many potential reasons to believe that \mathcal{C} is a random device (or that it’s not). We might know something about its manufacture, or be told that \mathcal{C} is random on good authority, *etc.* Our question is whether there is any information in \mathcal{C} ’s *output* that bears on whether \mathcal{C} is random. We’ll consider two statistics (concerning output sequences generated by \mathcal{C}) that are often taken to provide information [6] about \mathcal{C} ’s randomness. The first statistic is the number of runs in an output sequence [6]. The second statistic is the number of heads *versus* tails.

In order to evaluate the utility of these statistical tests for randomness, we’ll focus on the following two potential output sequences:

(A) HTTHTHHHT

(B) HHHHHTTTT

A *run* in a sequence is a maximal non-empty segment consisting of adjacent equal elements. For example (A) has six runs whereas (B) has just two. If Hs and Ts alternate randomly then the number of runs after N trials is a random variable whose cumulative distribution is given by counting the number of sequences of length N with r or fewer runs (or conversely, r or greater runs). Doing the relevant calculations for (A), we deduce: If \mathcal{C}

is a random device then the probability is 0.363 of producing these many (*viz.*, six) runs or more in a sequence of length nine. Because of this, advocates of the runs test say that producing (*A*) does *not* strongly disconfirm that \mathcal{C} is a random device. For (*B*) the same calculations imply: If \mathcal{C} is a random device then the probability is 0.035 of producing this many runs (*viz.*, two) or fewer in a sequence of length nine. In this case, advocates of the runs test say that \mathcal{C} 's generating (*B*) *does* strongly disconfirm \mathcal{C} 's randomness.¹

The binomial test gives the probability of throwing at least x heads in n tosses of the coin (or the probability of throwing at most x heads if they are fewer than $\frac{n}{2}$). In both (*A*) and (*B*), we see 5 heads in 9 tosses. We compute that if \mathcal{C} is a random device, then producing a sequence with five or more heads has probability 0.5. Because of this, advocates of the binomial test say that the fact that \mathcal{C} generates either sequence does *not* strongly disconfirm the claim that \mathcal{C} is a random device.

We've exploited two statistical tests to evaluate evidence regarding whether \mathcal{C} is a random device. If (*B*) is the output, the first test ("runs") classifies this as strongly disconfirmatory of \mathcal{C} 's randomness. If the output is (*A*), the first test does not deem this to be strongly disconfirmatory. The second test ("binomial") views neither case (*A*) nor (*B*) as constituting strong evidence against \mathcal{C} 's randomness. While these tests may disagree with each other, they each seem to be perfectly self-consistent. But, there is a problem . . .

2 THE PROBLEM

At a given position of the sequence produced by \mathcal{C} , there are more potential events than just "heads" and "tails." For example, let $X = \{1, 4, 9\}$, and define:

- Position i of \mathcal{C} 's output holds a *hail* (**h**) iff either $i \in X$ and position i holds a head (**H**), or $i \notin X$ and position i holds a tail (**T**).
- Position i of \mathcal{C} 's output holds a *tead* (**t**) iff either $i \in X$ and position i holds a tail (**T**), or $i \notin X$ and position i holds a head (**H**).

Given these definitions of teads and hails, we see that \mathcal{C} generates (*A*) iff \mathcal{C} generates (*a*), and \mathcal{C} generates (*B*) iff \mathcal{C} generates (*b*).

(<i>A</i>)	HTTHTHHHT	(<i>a</i>)	hhhhhtttt
(<i>B</i>)	HHHHHTTTT	(<i>b</i>)	htththhht

Let us respond at once to the concern that teads and hails are "unnatural," "position dependent," or otherwise "gerrymandered." Such characterizations seem no more applicable to teads/hails than to heads/tails. For, we have the following symmetry:

¹Standard objections to evidential interpretations of classical statistical tests have been recently surveyed in [5] and [2]. Our objection will be somewhat different from earlier concerns.

- Position i of \mathcal{C} 's output holds a tail (T) iff either $i \in X$ and position i holds a tead (t), or $i \notin X$ and position i holds a hail (h).
- Position i of \mathcal{C} 's output holds a head (H) iff either $i \in X$ and position i holds a hail (h), or $i \notin X$ and position i holds a tead (t).

Someone who thinks in terms of heads/tails may well find teads/hails to be “derivative.” But someone who thinks in terms of teads/hails will make the parallel claim about heads/tails. It's not obvious how to break the symmetry. Moreover, \mathcal{C} produces an unbiased, independent sequence of heads/tails iff \mathcal{C} produces an unbiased, independent sequence of teads/hails. (This is easy to verify.) Therefore, the runs test applied to teads/hails is as relevant to the randomness of \mathcal{C} as the runs test applied to heads/tails. Unfortunately, applying the runs test to teads/hails leads to a *reversal* of our initial assessment (in terms of heads/tails).

To see this, just count the number of runs of teads/hails in (a) and (b), above. We see that (a) has two runs and (b) has six. Doing the relevant calculations for (a), we deduce: If \mathcal{C} is a random device then the probability is 0.035 of producing a sequence (of length nine) with so few t/h runs. The advocate of the runs test should say that this constitutes strong evidence against the claim that \mathcal{C} is a random device. And, for (b), we deduce: If \mathcal{C} is a random device then the probability is 0.363 of producing a sequence (of length nine) with so many t/h runs. The advocate of the runs test should say that that this does *not* constitute strong evidence against the claim that \mathcal{C} is a random device. Thus, the use of teads/hails instead of heads/tails reverses the evidential verdict implied by the runs test!

Underlying this phenomenon is alteration of the “rejection set” in the passage from heads/tails to teads/hails. The rejection set is composed of the sequences whose number of runs is “too extreme” to be easily compatible with \mathcal{C} 's randomness. A given, potential output from \mathcal{C} might be considered extreme when the rejection set is reckoned in terms of runs of heads/tails but not teads/hails, and conversely. So the runs test is ambiguous unless some reason can be given to favor one way of counting runs over all the competing ways (and finding such a reason seems problematic).

The same sort of reversal can be achieved for the binomial test as well. To wit, consider the following pair of potential outcome sequences:

(A) HTTHTHHHT

(D) TTTTTTTTT

Then, let $Y = \{2, 3, 5, 9\}$, and define:

- Position i of \mathcal{C} 's output holds a *schmail* (t) iff either $i \in Y$ and position i holds a tail (T), or $i \notin Y$ and position i holds a head (H).
- Position i of \mathcal{C} 's output holds a *schmead* (h) iff either $i \in Y$ and position i holds a head (H), or $i \notin Y$ and position i holds a tail (T).

Similarly to before, \mathcal{C} is a random device for generating heads/tails iff \mathcal{C} is a random device for generating schmails/schmeads. But, \mathcal{C} produces (A) or (D) iff \mathcal{C} produces (c) or (d) , respectively. So we can apply the binomial test to both (pairs of) sequences of events:

(A)	HTTHTHHHT	(c)	ttttttttt
(D)	TTTTTTTTT	(d)	htththhht

Applying the binomial test to the schmeads and schmails in (c) yields: If \mathcal{C} is a random device then the probability is 0.004 of producing an event with so few hs. The advocate of the binomial test should therefore view \mathcal{C} 's generating (c) as strong evidence against the claim that \mathcal{C} is a random device. We saw earlier that the binomial test does *not* imply that (A) is an improbable sequence if generated randomly. As such, advocates of the binomial test should not view \mathcal{C} 's generation of (A) as strong evidence against \mathcal{C} 's randomness. The same reversal affects (D) and (d) . Once again, the test's implications about evidential relevance depend on which concepts we employ.²

3 WHAT THE PROBLEM IS *Not*

The teads/hails terminology resonates with Goodman's [1, Ch. 3] use of grue/bleen to question the basis of projections to the future. But this is not the point of the present discussion. Indeed, whether one reckons an output sequence as HTTHTHHHT versus hhhhtttt has no bearing on predictions about the next coin toss. After the 9th output, heads are invariably teads and tails hails. So if you expect a head [tail] there is no harm in announcing a tead [hail]. The situation is thus different from Goodman's since projecting the greenness of emeralds ultimately conflicts with projecting grueness (after time t the two kinds of emeralds look different). The same remarks apply to schmeads and schmails.

In contrast, the choice between heads/tails versus teads/hails appears to alter the verdict of standard statistical tests about the here and now, namely, whether \mathcal{C} is producing its output randomly. Driving the ambiguity is the fact that \mathcal{C} *issues heads and tails in a uniform, independent way just in case the same is true for teads and hails*, hence, the tests apply equally in the two cases. Preserving the "null hypothesis" of uniformity and independence across shifts in vocabulary is not a feature of the grue/bleen puzzle.³

Of course, at a more abstract level, both teads/hails and grue/bleen point to the language dependence of inductive inference. If we denoted both heads and tails by *theds* without specialized vocabulary for each then we might be struck by the fact that \mathcal{C} produces nothing but theds. But our point is more specific. Standard statistical tests for the randomness of \mathcal{C} yield conflicting results even though \mathcal{C} is random with respect to one vocabulary if and only if it is random with respect to the other. Unless a principled choice can be made among candidate vocabularies, the tests are bound to offer equivocal verdicts.

²Such reversals will plague *any* statistical test for randomness that we have encountered (see [3, Ch. 2] for a recent survey).

³In this sense, the present phenomenon is perhaps more similar to Miller's [4, Ch. 11] language-dependencies than Goodman's.

Embracing the language dependence of the tests, moreover, does not seem to be a viable response to the ambiguity. It makes no sense to declare different levels of confidence for \mathcal{C} 's being random “in the sense of heads/tails” compared to \mathcal{C} 's being random “in the sense of teads/hails.” For (to repeat), *\mathcal{C} is random in one sense if and only if it is random in the other.* A better response, it seems to us, is to abandon the tests altogether, along with any other attempt to harness \mathcal{C} 's output to compute its likelihood assuming randomness within a null hypothesis framework.

4 LESSONS LEARNED

What is the value of a statistical test whose outcome is so sensitive to the concepts used to describe the data? It would appear that this kind of null hypothesis testing in the service of evaluating the randomness of \mathcal{C} is of little epistemic value. Indeed, it is often noted that all sequences of a given length have the same probability of being generated by an unbiased independent source. So there's no such thing as an “atypical” sequence that is “unlikely” to be generated if \mathcal{C} is random. *All sequences are atypical, surprising, coincidental, etc.*⁴

Yet, intuitively, it seems reasonable (in some sense) to be sceptical about the randomness of a source that relentlessly produces heads. What explanation can we offer for such doubt? Prior to seeing any output, there are many alternatives to the hypothesis that \mathcal{C} is random. One alternative is that *a human mind controls the output*. The human-control hypothesis enjoys a relatively elevated prior probability because there are so many human minds in the neighborhood. (If we lived far away, we might be surrounded by teads/hails speakers, leading to different priors about the character of \mathcal{C} .) The likelihood of a long initial stretch of heads — *given human control* — is relatively high (simply because that's the kind of thing a human would do), so the posterior probability of human-control comes to swamp the priors.

On our view, belief that \mathcal{C} is random should not be based solely on \mathcal{C} 's output. Ideally, it is inspection of \mathcal{C} 's mechanism that grounds convictions about randomness (perhaps because of symmetries discovered, or for deeper reasons involving quantum theory, *etc.*). On this view, a sequence of events is “random” iff it has been generated by a random device.

REFERENCES

- [1] Goodman, N. (1955), *Fact, Fiction and Forecast*, Harvard.
- [2] Greco, D. (2011), “Significance Testing in Theory and Practice,” *British Journal for the Philosophy of Science*, 62: 607–637.
- [3] Li, M. and Vitanyi, P. (2008), *An Introduction to Kolmogorov Complexity and Its Applications*, third edition, Springer.

⁴Ironically, some advocates of statistical tests for randomness [3, §1.8.1] seem to think that this reveals a *shortcoming* of purely probabilistic assessments of output sequences. On the contrary, we think the present considerations show that there is something wrongheaded about any approach to randomness that appeals solely to properties of output sequences.

- [4] Miller, D. (2006), *Out of Error: Essays in Critical Rationalism*, Ashgate.
- [5] Royall, R. (1997), *Statistical Evidence: A Likelihood Paradigm*, Chapman & Hall.
- [6] Wald, A. and Wolfowitz, J. (1940), “On a test whether two samples are from the same population,” *Annals of Mathematical Statistics* 11: 147–162.

B. Fitelson: branden@fitelson.org

D. Osherson: osherson@princeton.edu