How Long Until a Random Sequence Decreases?

JACOB A. SIEHLER
Washington & Lee University
Lexington, VA 24450
siehlerj@wlu.edu

Waiting for the fall Imagine observing a stream of random real numbers: If you saw the sequence

$$0.0478, 0.1429, 0.1667, 0.2204, 0.8124, 0.8226, 0.3101 \ldots,$$

with the first decrease occurring in the seventh position, you might feel that this was an unusually long time to wait for that first decrease—even if you're not exactly sure how long such a run "usually" lasts in a random sequence. The average position of the first decrease in a stream of random numbers depends on precisely what you mean by random; that is, it depends on the distribution of random numbers that you're sampling. Surprisingly, though, for *continuous* distributions, the question has a very specific (and delightful) answer that is independent of how the random numbers are distributed.

We will uncover the answer in due time—in Proposition 2, to be precise. In the meantime, exercise your intuition by making a guess in advance about how long, on average, a monotone run like the one above will last in a sequence of random numbers. If you've been around math much, you can probably make a shrewd guess based solely on the fact that I described the answer as "delightful." But we will start by discussing a special case of the problem, when the random numbers are generated simply by rolling dice.

The answers to these questions have been known to specialists for some time. In fact, one can find the answer to the main question as an exercise in Knuth [**5**], and most of the results here can be found, in a more general setting, in Guy Louchard's thorough analysis of monotone runs [**6**]. However, they do not seem to be well known generally, despite their accessibility and interest for students with a basic undergraduate calculus background.

The die-rolling game One of the fringe benefits of teaching a course on probability and statistics is that it affords an excellent excuse to keep an assortment of toys on my desk, especially all sorts of dice. This article had its beginning when I was rolling an ordinary 6-sided die and got what I felt was an unusually long run before the first decrease in the numbers occurred. It might have been something like this:

$$1, 3, 3, 3, 4, 6, 5$$

and I decided to make a game of it: I would award myself 7 points for that run, since I got to roll seven times (including the final, decreasing roll that ended the game). Naturally, I was wondering what a typical score in my new game might be. But I also had my eye on the 4- and 20-sided dice lying nearby, and wondered if one of those might give me a better chance to get a large score. (Exercise your intuition again: can I expect to get a better score by choosing one of the other dice?)

Math. Mag. **83** (2010) 374–379. doi:10.4169/002557010X529798. © Mathematical Association of America

This content downloaded from
138.236.178.90 on Mon, 24 Dec 2018 16:34:30 UTC
All use subject to https://about.jstor.org/terms

Mathematically, an n-sided die is modeled by a discrete random variable that is uniformly distributed over the set $\{1, 2, \dots, n\}$; by this we mean that all of the outcomes in the set occur with the same probability, $1/n$. But we're not interested in single rolls of the die; rather, we want to study nondecreasing runs from repeated rolls of the die, and that motivates the following. Given any random variable X we define an associated random variable $R(X)$, the *run-length variable* for X, as follows: We sample X until the first decrease occurs, then let $R(X)$ be the total number of samples we took, including the final decrease that ends the experiment. Call X the *underlying variable* of $R(X)$.

The $R(X)$ notation emphasizes the fact that the experiment depends on the underlying variable X, but we'll suppress the argument and simply refer to R when the underlying variable is clear from context. And we'll call the run-length variable R_n when X is an n-sided die roll.

Our goal now is to study the expected value of R, particularly when the underlying variable is an n-sided die roll. Informally, the expected value of a random variable is the long-term average of its outcomes; by definition, the expected value of R is

$$E[R] = \sum_r rf(r),$$

where the summation is over all possible outcomes r that might occur, and $f(r)$ denotes the probability of getting outcome r.

No matter what the underlying variable is, we always get an outcome of at least 2 for R. On the other hand, there's no upper bound on the potential length of a nondecreasing run, so the possible outcomes of R are $\{2, 3, 4, \dots\}$, and we can rewrite the expected value calculation with more explicit limits of summation:

$$E[R] = \sum_{r=2}^{\infty} rf(r).$$

Next we find an explicit formula for $f(r)$ in the case of an n-sided die roll. Take $n \geq 2$, and let $a_n(r)$ denote the number of nondecreasing sequences of length r that can be formed from the set $\{1, 2, \dots, n\}$. This is a problem of selection with repetition allowed, and any combinatorics text will tell you that $a_n(r)$ is given by a binomial coefficient: $a_n(r) = \binom{n+r-1}{n-1}$. In fact, we do not need to defer to a text for this: Imagine making your nondecreasing selection by distributing r stones among boxes numbered 1 through n, all in a row. Place a stick between each adjacent pair of boxes; now, let the boxes vanish (leaving their contents behind). What remains is a sequence of r stones and $(n-1)$ sticks which uniquely codes for your selection, and the number of such sequences is counted by the binomial coefficient we have given.

With this notation the probability that a sequence of length r is nondecreasing is $a_n(r)/n^r$. The probability that a random sequence of length r decreases for the first time in the last position, then, is the probability that it increases for $r-1$ steps, minus the probability that it increases for r steps:

$$f_n(r) = \frac{a_n(r-1)}{n^{r-1}} - \frac{a_n(r)}{n^r}, \tag{1}$$

which can simplified (just combine fractions and cancel factorials) to

$$f_n(r) = \binom{n+r-2}{r} \cdot \frac{(r-1)}{n^r}. \tag{2}$$

With a formula for the mass function established, we can proceed to the expected value problem.

This content downloaded from
138.236.178.90 on Mon, 24 Dec 2018 16:34:30 UTC
All use subject to https://about.jstor.org/terms

PROPOSITION 1. *The expected value of R_n is given exactly by*

$$E[R_n] = \left(\frac{n}{n-1}\right)^n$$

Proof. With formula (2) in hand, we can write the summation for the expected value as

$$\sum_{r=2}^{\infty} r \cdot \binom{n+r-2}{r} \cdot \frac{(r-1)}{n^r} \qquad (*)$$

and massage the form until the sum can be evaluated, as follows:

$$\begin{aligned}(*) &= n(n-1)\sum_{r=2}^{\infty}\binom{n+r-2}{r-2}\left(\frac{1}{n}\right)^r \\ &= \frac{n-1}{n}\sum_{r=0}^{\infty}\binom{n+r}{r}\left(\frac{1}{n}\right)^r \\ &= \frac{n-1}{n}\left(1-\frac{1}{n}\right)^{-(n+1)},\end{aligned}$$

the last line following from the binomial series expansion of $\left(1-\frac{1}{n}\right)^{-(n+1)}$ which appears in the previous step. The series does converge, since we are assuming $n \geq 2$, and of course the resulting expression reduces to the form in the statement of the proposition. ■

Let's look back at a few questions we can now answer about the die-rolling game:

1. By formula (2), the probability that I would get a score of 7, using an ordinary 6-sided die as in the example, is just

$$f(7) = \frac{\binom{11}{7} \cdot 6}{6^7},$$

which is about 0.007—small enough that you might suspect my example is fictitious. (With a little more work you can check that the probability that I would get a score of 7 *or more* is just a tiny bit less than 1%, which is probably a more relevant fact.)

2. By Proposition 1, the average score for a game with a 6-sided die would be $(6/5)^6$, or just barely under 3.
3. Since the formula $(n/(n-1))^n$ is strictly decreasing in n (a popular exercise), I'd have a higher score, on average, if I switched to a 4-sided die, and a lower score if I used the 20-sided die. In fact, to maximize your score, your best bet for this game would be to toss a coin with sides labelled 1 and 2. You'd expect an average score of 4 in that case.

If you're still thinking about the question posed in the introduction, you might stop to consider this: What can you say about the expected value of R if X has a very large number of equally likely outcomes?

A variation: strictly increasing die rolls If we change the rules of the die-rolling game slightly to insist on strictly increasing numbers, we get slightly different results

This content downloaded from
138.236.178.90 on Mon, 24 Dec 2018 16:34:30 UTC
All use subject to https://about.jstor.org/terms

for the mass function and expected score. In the strictly increasing game, for example, rolling

$$1, 3, 5, 5$$

would cause the game to end with a score of 4. We can briefly establish results analogous to those of the previous section for the strictly increasing game. The expected value calculation may seem even simpler, as it uses the more familiar version of the binomial theorem, where the exponent is a positive integer.

Given any random variable X, define another random variable $R_s(X)$ as follows: we sample X until we get a result which is *not* strictly greater than the previous result, then let $R_s(X)$ be the total number of samples we took. If X represents an n-sided die roll, then the probability mass function for $R_s(X)$ is given by

$$f_s(r) = \binom{n+1}{r} \cdot \frac{(r-1)}{n^r}. \tag{3}$$

The details are left as an exercise; the derivation is very similar to the nondecreasing case. And if X represents an n-sided die roll then the expected value of $R_s(X)$ is given by

$$E[R_s(X)] = \left(\frac{n+1}{n}\right)^n \tag{4}$$

To verify this, notice that in this variant we have an upper bound on the possible scores: If we use an n-sided die, then our score must come from the set $\{2, \ldots, n+1\}$. That means that the expected value calculation involves only a finite sum:

$$\begin{aligned} E[R_s] &= \sum_{r=2}^{n+1} r \cdot \binom{n+1}{r} \frac{(r-1)}{n^r} \\ &= \left(\frac{n+1}{n}\right) \sum_{r=0}^{n-1} \binom{n-1}{r} \left(\frac{1}{n}\right)^r \\ &= \left(\frac{n+1}{n}\right)\left(1+\frac{1}{n}\right)^{n-1}, \end{aligned}$$

with the last line following by the binomial theorem. And this simplifies to $\left(\frac{n+1}{n}\right)^n$ as claimed.

QUESTION. *In the strictly increasing game, what sort of die (how many sides) should you choose to maximize your expected score?*

The continuous game Now, instead of rolling dice to generate our random numbers, suppose we have a continuous (real) random variable X as our source of randomness.

The only assumption we will make is that X is described by a probability density function—that is, there is a nonnegative function $p(x)$ on $\mathbb{R}$ with the property that the probability that X is between a and b is given by

$$P(a \le X \le b) = \int_a^b p(x)\,dx.$$

With a continuous random variable, there is zero probability of any sample duplicating an earlier number in the sequence. Therefore, the probability that a randomly

This content downloaded from
138.236.178.90 on Mon, 24 Dec 2018 16:34:30 UTC
All use subject to https://about.jstor.org/terms

generated sequence of length r is nondecreasing is the same as the probability that it is strictly increasing, $1/r!$ in both cases. As in equation (1), we can compute the probability of a monotone run of length as a difference of probabilities:

$$f(r) = \frac{1}{(r-1)!} - \frac{1}{r!} = \frac{(r-1)}{r!} \tag{5}$$

PROPOSITION 2. *Let R be the run-length variable for any continuous random variable X. Then the expected value of R is exactly*

$$E[R] = e,$$

the base of the natural logarithm.

Proof. This is just an easy corollary of the formula for the mass function; working directly from the definition of expected value we have

$$E[R] = \sum_{r=2}^{\infty} r f(r) = \sum_{r=2}^{\infty} \frac{r(r-1)}{r!} = \sum_{r=2}^{\infty} \frac{1}{(r-2)!}$$

and this last expression is exactly the beloved series expansion

$$\frac{1}{0!} + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \cdots = e. \qquad \blacksquare$$

Based on the previous sections, we might have arrived at this result heuristically as follows: a continuous distribution is, loosely speaking, like a distribution with infinitely many equally likely outcomes. Since the expected run length when there are n equally likely outcomes is $E[R_n] = (n/(n-1))^n$, we could have guessed that

$$E[R] = \lim_{n\to\infty} \left(\frac{n}{n-1}\right)^n$$

and this is another famous limit expression for e. (We could just as well have used a limit of formula (4) from the strictly increasing game.) There is a pleasant symmetry in the way the strictly increasing and nondecreasing versions of the discrete game approach the continuous game, matched by the use of the binomial theorem with positive and negative exponents. This can be seen in the mass functions as well as the expected value result: taking the limit as n goes to infinity in equations (2) or (3) gives the mass function for the continuous case.

The appearance of e in this problem is reminiscent of its appearance in the "Hat-Check Problem" [**2, 3**], where $1/e$ occurs as the approximate probability that a random permutation of an n-element set has no fixed points; the probabilities converge to e as n gets large. In the hat-check problem, $1/e$ is an excellent approximation to the true probability even for relatively small n. In our problem, the expected value for R_n converges much more slowly to e; roughly, you have to use an $n = 10^k$-sided die for $E[R_n]$ to match k decimal digits of e. The slow convergence of this sequence is discussed in an interesting article by Knox and Brothers [**4**].

A combinatorial connection The discrete distribution described by equation (5) can be viewed as a limiting case of either of the two families of distributions described by equations (2) and (3). Neither the families nor the limiting distribution appear to be familiar enough to have a widely-known name attached to them. For the statistically-inclined and curious, further investigation of these distributions might begin with their

This content downloaded from
138.236.178.90 on Mon, 24 Dec 2018 16:34:30 UTC
All use subject to https://about.jstor.org/terms

variance and higher moments. In the case of the run-length distribution for continuous variables, that would entail considering sums of the form

$$\sum_{r=2}^{\infty} \frac{r^k (r-1)}{r!} \tag{6}$$

for different exponents k. Starting with $k = 1$, this will yield a sequence beginning

$$e, 3e, 10e, 37e, 151e, 674e, \ldots.$$

The sequence of coefficients (sequence A005493 in Sloane's index [**8**]) has a combinatorial interpretation in its own right, but may be better recognized as first differences in the sequence of Bell numbers,

$$\{B_k\} = 1, 2, 5, 15, 52, 203, 877, \ldots,$$

which count the number of ways to partition a k-element set. The connection can be seen by splitting (6) as

$$\sum_{r=2}^{\infty} \frac{r^k}{(r-1)!} - \sum_{r=2}^{\infty} \frac{r^{k-1}}{(r-1)!}$$

and recognizing these as Dobinski's summations [**1, 7**] for B_{k+1} and B_k (albeit with their first terms deleted). This fact probably doesn't afford us any winning insight into dice games, but the emergence of the Bell numbers here, hand in hand with Euler's e, seems to mark this problem as perfectly poised on the boundary between continuous and discrete mathematics, and a satisfying demonstration of the interplay between them.

REFERENCES

1. Louis Comtet, *Advanced combinatorics*, D. Reidel, Dordrecht, 1974.
2. Karl David, Rencontres reencountered, *Coll. Math. J.* **19** (1988) 139–148. doi:10.2307/2686175
3. Steve Fisk, The secretary's packet problem, *Math. Mag.* **61** (1988) 103–105. doi:10.2307/2690041
4. John A. Knox and Harlan J. Brothers, Novel series-based approximations to *e*, *Coll. Math. J.* **30** (1999) 269–275. doi:10.2307/2687664
5. Donald E. Knuth, *The Art of Computer Programming, Vol. 2: Seminumerical Algorithms*, Addison-Wesley, Reading, MA, 1981.
6. Guy Louchard, Monotone runs of uniformly distributed integer random variables: A probabilistic analysis, *Theor. Comp. Sci.* **346** (2005) 358–387. doi:10.1016/j.tcs.2005.08.015
7. Gian-Carlo Rota, The number of partitions of a set, *Amer. Math. Monthly* **71** (1964) 498–504. doi:10.2307/2312585
8. N. J. A. Sloane, *The on-line encyclopedia of integer sequences*, http://www.research.att.com/~njas/sequences.

Summary Increasing runs of numbers are a naturally attractive feature in any randomly-generated sequence. Surprisingly, the average length of such runs is easy to compute and does not depend on the distribution of the random numbers, at least in the case of continuous random variables. We prove this, along with similar results for runs in sequences generated by rolling dice.

This content downloaded from
138.236.178.90 on Mon, 24 Dec 2018 16:34:30 UTC
All use subject to https://about.jstor.org/terms