AN ELEMENTARY "EXCELLENT APPROXIMATION" TO SUCCESS RUN PROBABILITIES

Stephen M. Samuels

by

Purdue University

Technical Report #99-10

Department of Statistics
Purdue University
West Lafayette, IN USA

June 1999

*

AN ELEMENTARY "EXCELLENT APPROXIMATION" TO SUCCESS RUN PROBABILITIES

by

Stephen M. Samuels
Purdue University

Abstract

Here is a method of approximating success run probabilities which is elementary enough to fit nicely into almost any first course in probability, yet every bit as accurate as Feller's celebrated partial fractions expansion approach. The title is somewhat of a play on words: "excellent approximation" were Feller's words, while the method in this paper can be implemented with a fairly routine use of Excel software. In addition, you can use your spreadsheet software to quite easily compute the exact probabilities.

An Elementary "Excellent Approximation" to Success Run Probabilities

Stephen M. Samuels*

Abstract

Here is a method for approximating success run probabilities which is elementary enough to fit nicely into almost any first course in probability, yet every bit as accurate as Feller's celebrated partial fractions expansion approach. The title is somewhat of a play on words: "excellent approximation" were Feller's words, while the method in this paper can be implemented with a fairly routine use of Excel software. In addition, you can use your spreadsheet software to quite easily compute the exact probabilities.

KEY WORDS: Coin tossing; Bernoulli Trials; Excel.

1 Introduction

This paper is inspired by something I read in the **Science Times** section of *The New York Times* on August 4, 1998. An article by Malcolm W. Browne begins—

> Dr. Theodore P. Hill asks his mathematics students at the Georgia Institute of Technology to go home and either flip a coin 200 times and record the results. or merely pretend to flip a coin and fake 200 results. The following day he runs his eye over the homework data, and to the students' amazement, he easily fingers nearly all those who faked their tosses.

*Stephen M. Samuels is Professor, Department of Statistics, Purdue University, 1399 Math Sciences Bldg, West Lafayette, IN 47907-1399 (Email: ssamuels@stat.purdue.edu)

> 'The truth is,' he said in an interview, 'most people don't know the real odds of such an exercise. so they can't fake data convincingly.'

—and later quotes Hill's 1998 *American Scientist* article where he said

> A sequence of 200 truly random coin tosses of a fair coin contains a run of six heads or six tails with very high probability—the exact calculation is quite involved—yet the average person trying to fake a random sequence very rarely writes such long runs.

I resolved to try Ted Hill's gambit on opening day of my undergraduate probability course (which I did, with similar results). But I also wanted to be able to teach my class—later in the semester—how to obtain the relevant probability. Naturally I went first to the celebrated text by Feller (1968) What I found there (and will describe below) was fascinating, but as Hill says, "quite involved." I looked at some other textbooks but was unable to find a suitable method of approximating the probabilities, so I put together my own argument which I offer here. As you will see, its agreement with Feller's method is breathtakingly close. And, as you will also see, the approximation is truly excellent.

2 Feller's Analytic Approach

Chapter XIII of Feller (1968) is entitled "Recurrent Events. Renewal Theory." And Section 7 of that Chapter is "Application to the Theory of Success Runs." For a sequence of independent Bernoull trials, with success probability $p \equiv 1 - q$, let the random variable T_r be the number of trials until the first occurrence of r consecutive successes. Feller derives the probability generating function,

$$(1) \qquad F_r(s) \equiv Es^{T_r} = \frac{p^r s^r (1 - ps)}{1 - s + qp^r s^{r+1}} = \frac{p^r s^r}{1 - qs(1 + ps + \cdots + p^{r-1}s^{r-1})}.$$

¿From this, he gets the mean and variance of T_r

$$(2) \qquad \mu_r = \frac{1 - p^r}{qp^r}, \qquad \sigma_r^2 = \frac{1}{(qp^r)^2} - \frac{2r + 1}{qp^r} - \frac{p}{q^2}.$$

One could now make the normal approximation to the distribution of T_r. For example, if $p = 1/2 = q$ and $r = 5$, then $\mu = 62$, $\sigma = 58.22$ and 200 is about 2.37 standard deviations above the mean, so $P(T_5 > 200)$ is estimated to be 0.009. But, as we shall see shortly, this is a very poor estimate because the distribution of T_r is quite skewed. What Feller does is to use a method of partial fraction expansions (described in his Section XI.4) to derive the following "excellent approximation" to the tail of the distribution of T_r:

$$P(T_r > n) \sim \frac{1 - px}{(r + 1 - rx)q} \cdot \frac{1}{x^{n+1}} \tag{3}$$

where x is the unique positive root of the denominator in the second formula of (1). For $r = 2$ and $p = 1/2$, Feller leaves it to us to show that $x = (\sqrt{5} - 1) = 1.23607$. For this case, Feller does give a brief table of both exact and approximate values to demonstrate how good the approximation is, even for very small n. For $r = 5$ and $p = 1/2$ (which, as we'll see, is relevant to what Hill did in class), a little help from Mathematica yields $x = 1.01732$ and the values in Column 3 of Table 1. For reasons which I will explain shortly, I used $n - 1$ rather than n in (3) for my table values. You can see from the table that $P(T_5 \geq 200) = 0.0346$ which is almost four times the 0.009 from the crude normal approximation above.

Of course, Hill's problem is different. He is looking for runs of length 6 of *either heads or tails.* Feller covers this problem, too, in Section 8, "More General Patterns," the first example of which is *Runs of either kind.* He again gives us the probability generating function of the waiting time,

$$\tilde{F}_r(s) = \frac{p^r s^r (1 - ps)(1 - q^r s^r) + q^r s^r (1 - qs)(1 - p^r s^r)}{1 - s + qp^r s^{r+1} + pq^r s^{r+1} - p^r q^r s^{2r}}. \tag{4}$$

He also gives us the mean (with an obvious misprint in my copy of the book),

$$\tilde{\mu}_r = \frac{(1 - p^r)(1 - q^r)}{qp^r + pq^r - p^r q^r}, \tag{5}$$

but not the variance.

When $p = 1/2 = q$, $\tilde{\mu}_r = 2^r - 1 = 1 + \mu_{r-1}$. This is more than just a

coincidence. Indeed, letting $u = s/2$,

$$\begin{aligned} \tilde{F}_r(u) &= \frac{2u^r(1-u)(1-u^r)}{1-2u+2u^{r+1}-u^{2r}} \\ &= \frac{2u^r(1-u)(1-u^r)}{(1-2u+u^r)(1-u^r)} \\ &= \frac{2u^r(1-u)}{(1-2u+u^r)} \\ &= 2uF_{r-1}(u). \end{aligned} \tag{6}$$

Thus, for a fair coin, the waiting time, for a run of length r of either heads or tails, has the same distribution as one plus the waiting time for a run of $r-1$ heads—which with hindsight is obvious, isn't it?—and Feller's "excellent approximation" (for 199 tosses) can be applied to Hill's case.

3 Quick Approximations

Let us now specialize to runs *of either kind* of length r in n tosses of a fair coin. We'll start with two crude approximations which are **upper bounds on the probability of no such run**.

Perhaps the simplest such approximation looks only at say k consecutive non-overlapping blocks of r tosses and the probability of no run in any block, namely

$$\begin{aligned} &P(\text{no } r\text{-run in tosses } jr+1, jr+2, \ldots, jr+r;\ j=0,1,\ldots,k-1) \\ &\qquad = (1-2^{-(r-1)})^k < e^{-k/2^{r-1}}. \end{aligned} \tag{7}$$

If $n = kr + d$, with $0 \le d < r$, then (7) gives an upper bound for the probability of no run of length r. In Hill's case, where $r = 6$ and $k = 33$, this probability is $(31/32)^{33} = .3507$ (and $e^{-33/32} = .3566$). So already we know that the odds are (at least) almost two to one in favor of a run of length six in 200 tosses.

We can do a little bit better (replacing the 33 by 38 in Hill's case) with very little extra work, as follows: The number of tosses, T_r, until a run of length r, is r plus the sum of geometrically many independent "short blocklengths;" i.e., runs of length less than r. Thus

$$T_r = r + \sum_{i=1}^{\tau} L_i \tag{8}$$

where $L_1, L_2, \ldots$ are IID with a truncated augmented geometric($p = 1/2$) distribution

$$P(L = \ell) = 2^{-\ell}/(1 - 2^{-(r-1)}) \quad \ell = 1, 2, \ldots, r-1, \tag{9}$$

and τ is independent of the L_i's with a geometric($p = 2^{-(r-1)}$) distribution, so

$$P(\tau > k') = (1 - 2^{-(r-1)})^{k'} < e^{-k'/2^{r-1}}. \tag{10}$$

Now, if $n = r + k'(r-1) + d'$ with $0 \le d' < r$, then $\tau \le k'$ implies that T_r must be less that or equal to n. So (10) gives an upper bound on the probability of no run. In Hill's case, $200 = 6 + 38(5) + 4$ so $k' = 38$. Thus an improved simple upper bound on the probability of no run is $(31/32)^{38} = .2993$ (and $e^{-38/32} = .3050$). So now we know that the odds are more than two to one in favor of a run of length six in 200 tosses.

4 Our Own "Excellent Approximation"

The latter of the two approximations described above is just a piece of a package. Here is the whole package:

Let us now use the Law of Total Probability to write $P(T_r > n)$, conditioning on τ, the number of short runs before the first run of length r. Thus, for $n - r = k'(r-1) + d'$ (i.e., $k' = [(n-r)/(r-1)]$),

$$\begin{aligned} P(T_r > n) &= \sum_{j=1}^{n-r} P(\tau = j) P(T_r > n | \tau = j) \\ &= 0 + \sum_{j=k'+1}^{n-r} P(\tau = j) P\left(\sum_{i=1}^{j} L_i > n - r \right) \\ &\approx \sum_{j=k'+1}^{n-r} 2^{-(r-1)} \left(1 - 2^{-(r-1)}\right) \left[1 - \Phi\left(\frac{n - r + \frac{1}{2} - j\mu_L}{\sqrt{j}\sigma_L} \right) \right]. \end{aligned} \tag{11}$$

Here we are using the normal approximation—with continuity correction—to the distribution of $L_1 + \cdots + L_j$, where the L_i's are IID with distribution given by (9), with mean μ_L and standard deviation σ_L. For $r = 6$, $\mu_L = 57/31 \approx 1.84$ and $\sigma_L = \sqrt{1122/961} \approx 1.08$. When k' is substantially larger

than r, as it is when $n = 200$ and $r = 6$, we can be quite confident that the normal approximation would work very well. But as we shall see, it works quite well even for the smallest values of n.

Formula (11) is a "piece of cake" to implement with Excel, using its `NORMDIST` function. To get column 2 of Table 1, I first put the integers 1 to 194 into the first 194 rows of Column A. Then I put "`=(1/32)*(31/32)^A1`" into cell B1 and copied the formula $(1/32) * (31/32)^j$ into the next 193 rows of Column B. Next I entered

```
=1-NORMDIST($E$1+0.5,$E$2*A1,$E$3*SQRT(A1),TRUE).
```

into `C1` and copied the formula into the next 193 rows of Column C. (Cells `E1`, `E2` and `E3` hold $n - r$, μ_L and σ_L, respectively, and the word "`TRUE`" tells Excel to give us the CDF rather than the density of the standard normal.) Finally I put products of `B`j and `C`j into `D`j and the sum of the `D`j's—my final answer—into `E4`. So each time I entered a new $n - 5$ into `E1` I got a new $P(T_5 > n)$ in `E4`, which I then copied into a little table elsewhere on my spreadsheet.

It's true that for n's smaller than 200, my summation goes beyond the limits, k' and $n - r$. But no harm is done because the contribution from those terms is negligible.

As you can see, in a wide range of n's there is agreement to about four significant digits between Feller's approximation and mine. So it seems safe to recommend mine—which uses only elementary ideas from probability—as a proxy for Feller's.

In addition, you can use Excel or some other spreadsheet software to quite easily get the exact probabilities (Column 4 of the Table) via an elementary six-state Markov Chain approach. For example, start by entering the numbers $1, 0, 0, 0, 0, 0$ into cells A1, B1, C1, D1, E1 and F1, respectively. Then enter "`=.5*SUM(A1:E1)`" into A2, "`=.5*A1`" into B2, "`=.5*B1`" into C2, "`=.5*C1`" into D2, "`=.5*D1`" into E2, "`=F1+.5*E1`" into F2, and "`=1-F2`" into G2. Now copy these formulas, row by row, into as many rows as you like. The contents of cell Gn will become $P(T_5 > n)$, the probability of no run of length 6 or more in the first n tosses. For example, for $n = 200$, the exact probability is 0.0346872, while Feller's approximation gives 0.0345924 and mine gives 0.0346857.

Table 1.
Two Approximations to the Probability of No Run of Either Kind, of Length 6 or More, in n Tosses of a Fair Coin.

n	Samuels'	Feller's	Exact
6	0.964808	0.970523	0.968750
7	0.950507	0.953999	0.953125
8	0.938236	0.937757	0.937500
10	0.906182	0.906098	0.906250
20	0.763187	0.763132	0.763123
30	0.642760	0.642723	0.642710
45	0.496793	0.496775	0.496760
50	0.455915	0.455903	0.455887
70	0.323385	0.323385	0.323369
100	0.193185	0.193194	0.193179
150	0.0818581	0.0818679	0.0818587
180	0.0489007	0.0489087	0.0489021
200	0.0346857	0.0345924	0.0346872

References

[1] Browne, M. W. (1998), "Following Benford's Law, or Looking Out for No. 1," *The New York Times,* (August 4, 1998, page B10).

[2] Feller, W. (1968), *An Introduction to Probability Theory and Its Application, Volume I,* (3rd ed.), New York: Wiley.

[3] Hill, T. P. (1998), "The first digit phenomenon," *American Scientist,* 86, 358–363.