Surprised by the Gambler's and Hot Hand Fallacies? A Truth in the Law of Small Numbers

Joshua B. Miller and Adam Sanjurjo *†‡

November, 2016 (with some updates)

Abstract

We prove that a subtle but substantial bias exists in a standard measure of the conditional dependence of present outcomes on streaks of past outcomes in sequential data. The magnitude of this novel form of selection bias generally decreases as the sequence gets longer, but increases in streak length, and remains substantial for a range of sequence lengths often used in empirical work. The bias has important implications for the literature that investigates incorrect beliefs in sequential decision making—most notably the Hot Hand Fallacy and the Gambler's Fallacy. Upon correcting for the bias, the conclusions of prominent studies in the hot hand fallacy literature are reversed. The bias also provides a novel structural explanation for how belief in the law of small numbers can persist in the face of experience.

JEL Classification Numbers: C12; C14; C18;C19; C91; D03; G02.

Keywords: Law of Small Numbers; Alternation Bias; Negative Recency Bias; Gambler's Fallacy; Hot Hand Fallacy; Hot Hand Effect; Sequential Decision Making; Sequential Data; Selection Bias; Finite Sample Bias; Small Sample Bias.

*Miller: Department of Decision Sciences and IGIER, Bocconi University, Sanjurjo: Fundamentos del Análisis Económico, Universidad de Alicante. Financial support from the Department of Decision Sciences at Bocconi University, and the Spanish Ministry of Economics and Competitiveness under project ECO2012-34928 is gratefully acknowledged.

†Both authors contributed equally, with names listed in alphabetical order.

‡This draft has benefitted from helpful comments and suggestions from the editor and anonymous reviewers, as well as Jason Abaluck, Jose Apesteguia, David Arathorn, Jeremy Arkes, Maya Bar-Hillel, Phil Birnbaum, Daniel Benjamin, Marco Bonetti, Colin Camerer, Juan Carrillo, Gary Charness, Ben Cohen, Vincent Crawford, Martin Dufwenberg, Jordan Ellenberg, Florian Ederer, Jonah Gabry, Andrew Gelman, Ben Gillen, Tom Gilovich, Maria Glymour, Uri Gneezy, Daniel Goldstein, Daniel Houser, Richard Jagacinski, Daniel Kahan, Daniel Kahneman, Erik Kimbrough, Dan Levin, Elliot Ludvig, Mark Machina, Daniel Martin, Filippo Massari, Guy Molyneux, Gidi Nave, Muriel Niederle, Christopher Olivola, Andreas Ortmann, Ryan Oprea, Carlos Oyarzun, Judea Pearl, David Rahman, Justin Rao, Alan Reifman, Pedro Rey-Biel, Yosef Rinott, Aldo Rustichini, Ricardo Serrano-Padial, Bill Sandholm, Vernon Smith, Lones Smith, Connan Snider, Joel Sobel, Charlie Sprenger, Daniel Stone, Sigrid Suetens, Dmitry Taubinsky, Richard Thaler, Nat Wilcox, and Bart Wilson. We would also like to thank seminar participants at Caltech, City U. London, Chapman U., Claremont Graduate School, Columbia U., Drexel U., George Mason U., New York University, NHH Norway, Max Planck Institute for Human Development (ABC), Microsoft Research, U. of Minnesota, Naval Postgraduate School, the Ohio State U., Santa Clara U., Stanford U., Tilburg U., U. de Alicante, U. del País Vasco, U. of Amsterdam, UC Berkeley, UC Irvine, U. of Pittsburg, UC Santa Cruz, UC San Diego, U. New South Wales, U. Southern California, U. of Queensland, U. of Wellington, U. of Wisconsin, U. of Zurich, Washington State U., WZB Social Science Center, as well as conference participants at Gary's Conference, IMEBE Rome 2016, M-BEES Maastricht 2015, SITE Stanford U. 2016, 11th World Congress of The Econometric Society, The 30th Annual Congress of the European Economic Association, and the 14th TIBER Symposium on Psychology and Economics. All mistakes and omissions remain our own.

1 Introduction

Jack takes a coin from his pocket and decides to flip it, say, one hundred times. As he is curious about what outcome typically follows a heads, whenever he flips a heads he commits to writing the outcome of the next flip on the scrap of paper next to him. Upon completing the one hundred flips, Jack of course expects the proportion of heads written on the scrap of paper to be one-half. Shockingly, Jack is wrong. For a fair coin, the expected proportion of heads is smaller than one-half.

We prove that for any finite sequence of binary data, in which each outcome of "success" or "failure" is determined by an i.i.d. random variable, the proportion of successes among the outcomes that immediately follow a streak of consecutive successes is expected to be strictly less than the underlying (conditional) probability of success.[1] While the magnitude of this novel form of selection bias generally decreases as the sequence gets longer, it increases in streak length, and remains substantial for a range of sequence lengths often used in empirical work.

We show that the bias has considerable implications for beliefs and decision making in environments that involve sequential data. First, we find that prominent studies in the influential *hot hand fallacy* literature (see Gilovich, Vallone, and Tversky [1985]; for a brief review see Section 3.4) have employed a biased estimation procedure analogous to Jack's.[2] Crucially, upon correcting for the bias we find that the long-standing conclusion of the seminal hot hand fallacy study reverses. Second, we use the bias to develop a novel structural explanation for how the well-known *gambler's fallacy* can persist, even for individuals who have extensive experience.[3] These and other implications are further discussed below.

To see why Jack's procedure in the opening example leads to a bias, consider the simplest case in which he flips the coin just three times. While Jack will generate only a single sequence of heads and tails, there are eight possible sequences. Each of these is listed in column one of Table 1. In column two are the number of flips that Jack would record (write down) on his scrap of paper for each sequence (these flips are underlined in column one), and in column three the corresponding proportion of heads among the flip outcomes recorded on his scrap of paper. Observe that the set of values the proportion can take is $\{0, 1/2, 1\}$, and that the number of sequences that yield each proportion leads to a skewed probability distribution of $(3/6, 1/6, 2/6)$ over these respective values. As a result, the expected proportion is $5/12$ rather than $1/2$. To provide a rough intuition for this result, we begin with the observation that the number of recorded flips (column 2) varies across sequences. In general, to have an opportunity to record more flips, more heads must be packed into the first $n-1$ flips of a length n sequence. This forces the heads to run together, which in turn

[1] This assumes only that the length of the sequence n satisfies $n \geq 3$, and that the streak length k satisfies $1 \leq k < n-1$.
[2] See Miller and Sanjurjo (2014) for a complete review of the literature.
[3] The gambler's and hot hand fallacies reflect opposite beliefs about the sign of sequential dependence in a random process. See Ayton and Fischer (2004) and Rabin (2002) for alternative approaches to reconciling the two.

Table 1: *The proportion of heads on those flips that immediately follow one or more heads, and the number of flips recorded, for the 8 equally likely sequences that can be generated when Jack flips a coin three times. In the bottom row the expected value of the proportion is reported under the assumption that the coin is fair.*

3-flip sequence	# of recorded flips	proportion of Hs on recorded flips
TTT	0	-
TTH	0	-
THT	1	0
HTT	1	0
THH	1	1
HTH	1	0
HHT	2	$\frac{1}{2}$
HHH	2	1
Expected Proportion (fair coin):		$\frac{5}{12}$

increases the proportion of heads on the flips that immediately follow heads in these sequences, as can be seen with HHT and HHH. This implies that sequences that have more recorded flips will tend to have a higher proportion of heads among these flips. Because sequences that have more recorded flips are given the same weight as sequences that have fewer, any recorded flip in such a sequence will be weighted less, which means that the heads are weighted less, resulting in the bias.[4]

In Section 2 we prove the bias for the general case. The proof uses straightforward Bayesian reasoning, which is made possible by operationalizing the expected proportion as a conditional probability. The proof highlights how a researcher who uses this proportion is implicitly following an estimation procedure that is *contingent* on the sequence of data that she observes. To derive an explicit formula for the bias we extend the intuition provided above in the simple three flip example, i.e. that the proportion is related to the way in which trial outcomes of one kind run together in finite sequences. While the formula does not appear, in general, to admit a simple representation, for the special case of streaks of length $k = 1$ (as in the examples discussed above) we provide one. For the more general case of $k > 1$, we use an analogous combinatorial argument to reduce the

[4]If Jack were instead to control the number of flips he records by flipping the coin until he records the outcomes of exactly m flips that immediately follow a heads, rather than flipping the coin exactly n times, the proportion would be unbiased. This method of controlling the effective sample size is known as *inverse sampling*, which provides an alternative intuition for the bias in which Jack's sampling procedure—the criterion he uses for recording flips—can be viewed as a form of repeated negative binomial sampling (e.g. see Haldane (1945)). Another unbiased method for estimating the conditional probability, which does not control the effective sample size, involves eliminating the overlapping nature of the measure. In particular, for a sequence of n flips, take each run of ones, and if it is of even length 2ℓ, divide it into blocks of two flips; if it is of odd length $2\ell - 1$ include the right adjacent tail and divide it into blocks of two flips. In each case, the run of ones contributes ℓ observations.

dimensionality of the problem, which yields a formula for the bias that is numerically tractable for sequence lengths commonly used in empirical work.

In Section 2.1 we show that the bias can be decomposed into a form of sampling-without-replacement and an additional bias that relates to the *overlapping words paradox* (Guibas and Odlyzko 1981). In particular, the additional bias results from the overlapping nature of the selection procedure that selects the trial outcomes used to calculate the proportion. For the simple case of $k = 1$, we show that the bias can be understood entirely in terms of sampling-without-replacement, which we use to reveal its near equivalence to the following known biases and paradoxes: (1) the Monty-Hall problem (Friedman 1998; Nalebuff 1987; Selvin 1975; Vos Savant 1990), and other classic probability puzzles, (2) a form of selection bias known in the statistics literature as Berkson's bias, or Berkson's paradox (Berkson 1946; Roberts, Spitzer, Delmore, and Sackett 1978), for which our approach provides new insights, and (3) a form of finite sample bias that shows up in autoregressive coefficient estimators (Shaman and Stine 1988; Yule 1926). For the more general case of $k > 1$, the bias is typically far stronger than sampling-without-replacement, and has no direct analog.

One implication of the bias is for the analysis of streak effects in binary (or binarized) sequential data. In Section 3 we revisit the well-known "hot hand fallacy," which refers to the conclusion of the seminal work of Gilovich et al. (1985; henceforth GVT), in which the authors found that despite the near ubiquitous belief among basketball fans and experts in the hot hand, i.e. "streak" shooting, statistical analyses of shooting data did not support this belief. The result has long been considered a surprising and stark exhibit of irrational behavior, as professional players and coaches have consistently rejected the conclusion, and its implications for their decision making. Indeed, in the years since the seminal paper was published a consensus has emerged that the hot hand is a "myth," and the associated belief a "massive and widespread cognitive illusion" (Kahneman 2011; Thaler and Sunstein 2008).

We find that GVT's critical test of hot hand shooting is vulnerable to the bias. As a result, we re-examine the raw data from GVT, using two different approaches to provide de-biased tests. We find that both approaches yield strong evidence of streak shooting, with considerable effect sizes. Further, we find similar results when correcting for the bias in other controlled tests of streak shooting that replicated GVT's original result using similar statistical tests (Koehler and Conley 2003; Miller and Sanjurjo 2015b). Lastly, we discuss studies in which each player takes sufficiently many shots to test for streak shooting on the individual level. We find significant and substantial evidence of the hot hand in each study (Jagacinski, Newell, and Isaac 1979; Miller and Sanjurjo 2014, 2015b).

On the basis of our evidence, we must conclude that the hot hand is not a myth, and that the associated belief is not a cognitive illusion. In addition, because researchers have: (1) accepted the null hypothesis that players have a fixed probability of success, and (2) treated the *mere* belief

in the hot hand as a cognitive illusion, the hot hand fallacy itself can be viewed as a fallacy. Nevertheless, evidence that the belief in the hot hand is justified does not imply that peoples' beliefs are accurate in practice. In fact, GVT provided evidence that players' beliefs in the hot hand are not accurate. In particular, GVT conducted a betting task, which was paired with their shooting task, and found that players' bets on shot outcomes are no better than what chance betting would predict. In Section 3.5 we show how GVT have misinterpreted their estimates, and observe that their tests are underpowered. Then, upon re-analyzing GVT's betting data, we find that their players successfully predict shot outcomes at rates significantly (and substantially) better than what chance would predict. This suggests that players can profitably exploit their beliefs in the hot hand. Further, we discuss findings from a separate study that show how players are able to identify which of their teammates have a tendency to get the hot hand (Miller and Sanjurjo 2014). Nevertheless, we observe that while these results are not inconsistent with the possibility that decision-makers detect the hot hand in real time, and have reasonably well-calibrated beliefs, they do not guarantee either. We suggest avenues for future research.

In Section 4 we present implications of the bias for the *Gambler's fallacy*, i.e. the tendency to believe that streaks are more likely to end than the underlying probability dictates. While the existence of the gambler's fallacy is commonly attributed to a mistaken belief in the *law of small numbers* (Rabin 2002; Tversky and Kahneman 1971), there exist no formal accounts for how it could persist in the face of experience (Nickerson 2002). Given this gap in the literature, we introduce a simple model in which a decision maker updates her beliefs as she observes finite sequences of outcomes over time. The model allows for the possibility that sequences are given equal weights, or variable weights according to sample size (e.g. the number of recorded flips in the Jack example). If sample size is correctly accounted for, then gambler's fallacy beliefs disappear with sufficient experience. However, with sufficient insensitivity to sample size the bias implies that a believer in the gambler's fallacy will never abandon his or her incorrect beliefs. The model has testable implications, as the degree of decision-maker bias will depend on the: (1) length of finite sequences observed, (2) length of streaks attended to, and (3) sensitivity to sample size.

Finally, because the bias is subtle and (initially) surprising, even for the sophisticated, those unaware of it may be susceptible to being misled, or exploited.[5] On the most basic level, in line with the discussion of the gambler's fallacy above, a naïve observer can be convinced that negative sequential dependence exists in an i.i.d. random process if sample size information is obscured. More subtly, the bias can also be leveraged to manipulate people into believing that the outcomes

[5]In informal conversations with researchers, and surveys of students, we have found a near-universal belief that the sample proportion should be equal to the underlying probability, in expectation. The conviction with which these beliefs are often held is notable, and reminiscent of the arguments which surrounded the classic Monty Hall Puzzle. Indeed, as mentioned above, in Section 2.1 (and Appendix F.1) we explain that the Monty Hall problem is essentially equivalent to the simplest version of the bias, with $n = 3$ and $k = 1$.

of an unpredictable process can be predicted at rates better than chance.[6] Aside from manipulation of beliefs, the bias can be applied in a straightforward way to construct gambling games that appear actuarially fair, but are not.[7]

Our identification of the bias in this sample proportion has revealed an underlying truth in the law of small numbers that intimately links the gambler's and hot hand fallacies. In particular, the bias implies that streaks within finite sequences are expected to end more often than continue (relative to the underlying probability), which can lead both the gambler to think that an i.i.d process has a tendency towards reversal, and the hot hand researcher to think that a process is i.i.d. when it actually has a tendency towards momentum. Absent a formal correction for the bias, the intuitive metric for probability of success on the trials of interest, the sample proportion, is expected to confirm the respective priors of both the gambler and the researcher.

Section 2 contains our main theoretical results, and Sections 3 and 4 the applications to the hot hand and gambler's fallacies, respectively.

2 The Bias

Let $\boldsymbol{X} = \{X_i\}_{i=1}^n$ be a sequence of binary random variables, with $X_i = 1$ a "success" and $X_i = 0$ a "failure." A natural procedure for estimating the probability of success on trial t, conditional on trial t immediately following k consecutive successes, is to first select the subset of trials that immediately follow k consecutive successes $I_k(\boldsymbol{X}) := \{i : \prod_{j=i-k}^{i-1} X_j = 1\} \subseteq \{k+1, \ldots, n\}$, then calculate the proportion of successes on these trials.[8] The following theorem establishes that when $\{X_i\}_{i=1}^n$ is a sequence of i.i.d random variables, with probability of success p and fixed length n, this procedure yields a biased estimator of the conditional probability, $\mathbb{P}(X_t = 1 \mid \prod_{j=t-k}^{t-1} X_j = 1) \equiv p$.

Theorem 1 *Let $\boldsymbol{X} = \{X_i\}_{i=1}^n$, $n \geq 3$, be a sequence of independent Bernoulli trials, each with probability of success $0 < p < 1$. Let $\hat{P}_k(\boldsymbol{X})$ be the proportion of successes on the subset of trials $I_k(\boldsymbol{X})$ that immediately follow k consecutive successes, i.e. $\hat{P}_k(\boldsymbol{X}) := \sum_{i \in I_k(\boldsymbol{X})} x_i / |I_k(\boldsymbol{X})|$. $\hat{P}_k$ is a biased estimator of $\mathbb{P}(X_t = 1 \mid \prod_{j=t-k}^{t-1} X_j = 1) \equiv p$ for all k such that $1 \leq k \leq n-2$. In particular,*

[6] For example, suppose that a predictor observes successive realizations from a binary (or binarized) i.i.d. random process (e.g. daily stock price movements), and is evaluated according to the *success rate* of her predictions over, say, three months. If the predictor is given the freedom of *when* to predict, then she can exceed chance in her expected success rate simply by predicting a reversal whenever there is a streak of consecutive outcomes of the same kind.

[7] A simple example is to sell the following lottery ticket for \$5. A fair coin will be flipped 4 times. For each flip the outcome will be recorded if and only if the previous flip is a heads. If the proportion of recorded heads is strictly greater than one-half then the ticket pays \$10; if the proportion is strictly less than one-half then the ticket pays \$0; if the proportion is exactly equal to one-half, or if no flip is immediately preceded by a heads, then a new sequence of 4 flips is generated. While, intuitively, it seems that the expected value of the lottery must be \$5, it is actually \$4. Curiously, the willingness-to-pay for the lottery ticket may be higher for someone who believes in the independence of coin flips, as compared to someone with Gambler's fallacy beliefs.

[8] In fact, this procedure yields the maximum likelihood estimate for $\mathbb{P}(X_t = 1 \mid \prod_{j=t-k}^{t-1} X_j = 1)$.

$$E\left[\ \hat{P}_k(\boldsymbol{X})\ \middle|\ I_k(\boldsymbol{X}) \neq \emptyset\ \right] < p \tag{1}$$

Proof: See Appendix A.[9]

Though biased, it is straightforward to show that $\hat{P}_k(\boldsymbol{X})$ is a consistent estimator of $\mathbb{P}(X_t = 1 \mid \prod_{j=t-k}^{t-1} X_j = 1)$.[10]

2.1 Mechanism behind the bias: sampling-without-replacement and streak overlap

Any sequence $\boldsymbol{x} \in \{0,1\}^n$ with $I_k(\boldsymbol{x}) \neq \emptyset$ that a researcher encounters will contain a certain number of successes $N_1(\boldsymbol{x}) = n_1$ and failures $n_0 := n - n_1$. To estimate the conditional probability of interest, the researcher will first select the subset of trials $I_k(\boldsymbol{x})$ from the sequence, and then compute the proportion of successes on those trials, i.e. $\hat{P}_k(\boldsymbol{x})$. For each case $N_1(\boldsymbol{x}) = n_1 \in \{k, \ldots, n\}$ the prior odds in favor of a success on any given trial in the sequence are $n_1/n_0 : 1$. By contrast, Theorem 1 shows that the odds are strictly less than this for any given trial in $I_k(\boldsymbol{x})$. An intuition for why can be obtained by considering the following equation, which follows from Bayes' rule (see Web Appendix D; Equation 19):

$$\frac{\mathbb{P}(x_t = 1 | \tau = t)}{\mathbb{P}(x_t = 0 | \tau = t)} = \frac{E\left[\ \frac{1}{M}\ \middle|\ \prod_{t-k}^{t-1} x_i = 1, x_t = 1\right]}{E\left[\ \frac{1}{M}\ \middle|\ \prod_{t-k}^{t-1} x_i = 1, x_t = 0\right]} \frac{n_1 - k}{n_1} \frac{n_1}{n_0} \tag{2}$$

Equation 2 gives the posterior odds $\frac{\mathbb{P}(x_t=1|\tau=t)}{\mathbb{P}(x_t=0|\tau=t)}$ in favor of observing $x_t = 1$ (relative to $x_t = 0$), for a representative trial $\tau = t$ drawn at random from $I_k(\boldsymbol{x})$.[11] Observe that the prior odds ratio n_1/n_0 is multiplied by two separate updating factors. Below we show that each factor is strictly less than one when $t < n$, resulting in posterior odds that are smaller than the prior.

The first updating factor $(n_1 - k)/n_1 < 1$ reflects the restriction that the finite number of available successes places on the procedure for selecting trials into $I_k(\boldsymbol{x})$. In particular, it can

[9] If the researcher were instead to control the number of selected trials by repeating the experiment until he generates exactly m trials that immediately follow k consecutive successes, then the proportion would be unbiased. Alternatively, if the researcher were to eliminate the overlapping nature of the measure, there would be no bias, even though the number of selected trials would still be random. In particular, for a sequence of n trials, one can take each run of successes, and if it is of even length 2ℓ, divide it into blocks of two trials; if it is of odd length $2\ell - 1$ include the right adjacent tails and divide it into blocks of two trials. In each case, the run of successes contributes ℓ observations.

[10] To demonstrate the consistency of $\hat{P}_k(\boldsymbol{X})$ first define $Y_{k,i} := \prod_{j=i-k+1}^{i} X_j$ for $i \geq k$. With this, $\hat{P}_k(\boldsymbol{X}) = \sum_{i=k+1}^{n} Y_{k+1,i} / \sum_{i=k}^{n-1} Y_{k,i}$. Note that each of the respective sequences $\{Y_{k,i}\}$, $\{Y_{k+1,i}\}$ are asymptotically uncorrelated (k fixed). Therefore, their time averages converge to their respective means almost surely, i.e. $1/(n-k)\sum_{i=k}^{n-1} Y_{k,i} \xrightarrow{a.s.} E[Y_{k,i}] = p^k$, and $1/(n-k)\sum_{i=k+1}^{n} Y_{k+1,i} \xrightarrow{a.s.} E[Y_{k+1,i}] = p^{k+1}$. This implies that $\hat{P}_k(\boldsymbol{X}) \xrightarrow{a.s.} p = \mathbb{P}(X_t = 1 \mid \prod_{j=t-k}^{t-1} X_j = 1)$, which in turn implies consistency.

[11] Note that the expected proportion is simply the probability of success for a trial that is randomly drawn from $I_k(\boldsymbol{x})$.

be thought of as the information provided upon learning that k of the n_1 successes are no longer available, which leads to a sampling-without-replacement effect on the prior odds n_1/n_0. This effect is made most transparent by re-expressing the (intermediate) posterior odds, $\frac{n_1-k}{n_1}\frac{n_1}{n_0}$ (before the second updating factor is applied), as $\frac{n_1-k}{n-k}/\frac{n_0}{n-k}$.[12,13] Clearly, the attenuation in the odds due to this factor increases in the streak length k.

The second updating factor $\frac{E\left[\ \frac{1}{M}\ \middle|\ \prod_{t-k}^{t-1} x_i=1, x_t=1\right]}{E\left[\ \frac{1}{M}\ \middle|\ \prod_{t-k}^{t-1} x_i=1, x_t=0\right]} < 1$, for $t < n$ (see Appendix D), reflects an additional restriction that the arrangement of successes and failures in the sequence places on the procedure for selecting trials into $I_k(\boldsymbol{x})$. It can be thought of as the additional information provided by learning that the k successes, which are no longer available, are consecutive and immediately precede t. To see why the odds are further attenuated in this case, we begin with the random variable M, which is defined as the number of trials in $I_k(\boldsymbol{x})$. The probability of any particular trial $t \in I_k(\boldsymbol{x})$ being selected at random is $1/M$. Now, because the expectation in the numerator conditions on $x_t = 1$, this means intuitively that $1/M$ is expected to be smaller in the numerator than in the denominator, where the expectation instead conditions on $x_t = 0$. The reason why is that for a sequence in which $x_t = 1$ the streak of 1's continues on, meaning that trial $t+1$ must also be in $I_k(\boldsymbol{x})$, and trials $t+2$ through $t+k$ each may also be in $I_k(\boldsymbol{x})$. By contrast, for a sequence in which $x_t = 0$ the streak of 1's ends, meaning that trials $t+1$ through $t+k$ cannot possibly be in $I_k(\boldsymbol{x})$, which leads the corresponding $1/M$ to be smaller in expectation.[14] This last argument provides intuition for why the attenuation of the odds due to this factor increases in k.[15]

Interestingly, for the special case of $k = 1$, $\frac{E\left[\ \frac{1}{M}\ \middle|\ x_{t-1}=1, x_t=1\right]}{E\left[\ \frac{1}{M}\ \middle|\ x_{t-1}=1, x_t=0\right]} = 1 - \frac{1}{(n-1)(n_1-1)} < 1$ when $t < n$, and $\frac{E\left[\ \frac{1}{M}\ \middle|\ x_{n-1}=1, x_n=1\right]}{E\left[\ \frac{1}{M}\ \middle|\ x_{n-1}=1, x_n=0\right]} = \frac{n_1}{n_1-1} > 1$ when $t = n$.[16] These contrasting effects combine to yield the familiar sampling-without-replacement formula:

$$E\left[\ \hat{P}_1(\boldsymbol{X})\ \middle|\ I_1(\boldsymbol{X}) \neq \emptyset,\ N_1(\boldsymbol{X}) = n_1\ \right] = \frac{n_1 - 1}{n - 1} \tag{3}$$

as demonstrated in Lemma 1, in Appendix A.2. On the other hand, when $k > 1$ the bias is

[12] The numerator of the latter expression is the probability of drawing a 1 at random from an urn containing n_1 1's and n_0 0's, once k 1's (and no 0's) have been removed from the urn. The denominator is the probability of drawing a 0 from the same urn.

[13] This effect calls to mind the key behavioral assumption made in Rabin (2002), that believers in the law of small numbers view outcomes from an i.i.d. process as if they were instead generated by random draws without replacement.

[14] This is under the assumption that $t \leq n - k$. In general, the event $x_t = 0$ excludes the next $\min\{k, n-t\}$ trials from $t+1$ to $\min\{t+k, n\}$ from being selected, while the event $x_t = 1$ leads trial $t+1$ to be selected, and does not exclude the next $\min\{k, n-t\} - 1$ trials from being selected.

[15] The likelihood ratio does not admit a simple representation; see footnote 66 of Web Appendix D.

[16] The likelihood ratios can be derived following the proof of Lemma 1 in Appendix A.2. In particular, for the equivalent likelihood ratio, $\frac{\mathbb{P}(\tau=t|x_{t-1}=1,x_t=1)}{\mathbb{P}(\tau=t|x_{t-1}=1,x_t=0)}$, the approach used to derive the numerator can also be used to show that the denominator is equal to $\frac{1}{n-2}\left(\frac{n_0-1}{n_1} + \frac{n_1-1}{n_1-1}\right)$. Further, in the case of $t = n$, it is clear that $\mathbb{P}(\tau = n|x_{n-1} = 1, x_n = 0) = \frac{1}{n_1}$. Each likelihood ratio then follows from dividing and collecting terms.

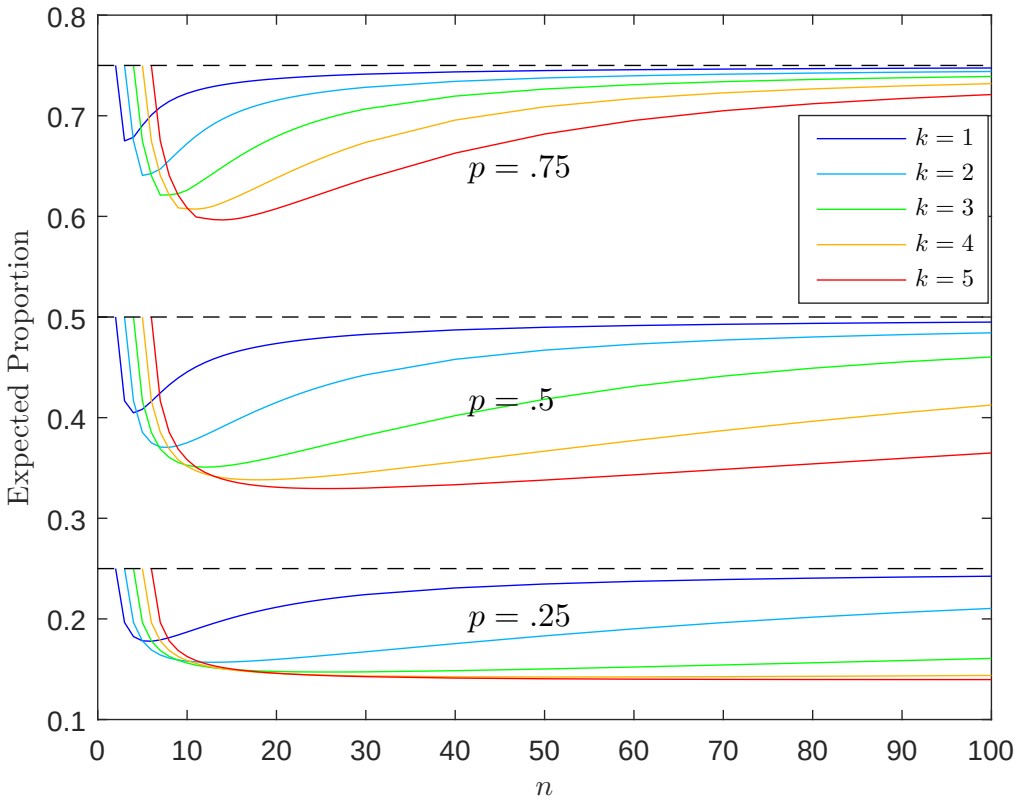

Figure 1: *The expected value of the proportion of successes on trials that immediately follow k consecutive successes, $\hat{P}_k(\boldsymbol{X})$, as a function of the total number of trials n, for different values of k and probabilities of success p, using the formula provided in Web Appendix E (Theorem 5, combined with Equation 20).*

substantially stronger than sampling-without-replacement (see Figure 6 in Appendix D). For further discussion on the relationship between the bias, sampling-without-replacement, and the *overlapping words paradox* (Guibas and Odlyzko 1981) see Web Appendix F.

2.2 Quantifying the bias.

In order to derive a formula for $E[\ \hat{P}_k(\boldsymbol{X}) \mid I_k(\boldsymbol{X}) \neq \emptyset\ ]$, and quantify the magnitude of the corresponding bias, we first derive a formula for the expected proportion that is conditional on the number of success in the sequence, $N_1(\boldsymbol{x}) := n_1$. Then we compute the unconditional expectation using the distribution $\mathbb{P}(N_1(\boldsymbol{x}) = n_1 | I_k(\boldsymbol{X}) \neq \emptyset)$. The conditional expectation can, in principle, be obtained directly by computing $\hat{P}_k(\boldsymbol{x})$ for each sequence that contains n_1 successes, and then taking the average across sequences, as performed in Table 1. However, the number of sequences required

for the complete enumeration is typically too large.[17] Consequently, we instead derive a formula that is numerically tractable by identifying, and enumerating, the set of sequences for which $\hat{P}_k(\boldsymbol{x})$ is constant, which greatly reduces the dimensionality of the problem. The set of such sequences is determined both by the number of successes n_1 and how many runs of successes of each length there are. This observation can be used to derive an explicit formula by way of combinatorial argument. While the formula does not admit a simple representation for $k > 1$, it is numerically tractable for the sequence and streak lengths that are empirically relevant. For the special case of $k = 1$ a simple representation exists, which we provide in Appendix A.2.

Figure 1 contains a plot of $E[\, \hat{P}_k(\boldsymbol{X}) \mid I_k(\boldsymbol{X}) \neq \emptyset \,]$, as a function of the number of trials in the sequence n, and for different values of k and p.[18] The dotted lines in the figure represent the true probability of success for $p = 0.25, 0.50$, and 0.75, respectively. The five solid lines immediately below each dotted line represent the respective expected proportions for each value of $k = 1, 2, \ldots, 5$. Observe that while the bias does generally decrease as n increases, it can remain substantial even for long sequences. For example, in the case of $n = 100$, $p = 0.5$, and $k = 5$, the magnitude of the bias is $.35 - .50 = -0.15$, and in the case of $n = 100$, $p = 0.25$, and $k = 3$, the magnitude of the bias is $.16 - .25 = -0.09$.[19]

3 Application to the Hot Hand Fallacy

> *This account explains both the formation and maintenance of the erroneous belief in the hot hand: if random sequences are perceived as streak shooting, then no amount of exposure to such sequences will convince the player, the coach, or the fan that the sequences are in fact random.* (Gilovich, Vallone, and Tversky [GVT] 1985)

In their seminal paper GVT find no evidence of hot hand shooting in their analysis of basketball shooting data, despite the near-unanimous belief in the hot hand among players, coaches, and fans. As a result, they conclude that belief in the hot hand is a "powerful and widely shared cognitive illusion." (p. 313).

[17] For example, the GVT basketball data that we analyze in Section 3 has shot sequences of length $n = 100$ and a design target of $n_1 = 50$ made shots, resulting in a computationally unwieldy $\binom{100}{50} > 10^{29}$ distinguishable sequences.

[18] All values are exact. For $k > 1$ the figure was produced using the formula for the expectation found in Theorem 5, which conditions on $N_1(\boldsymbol{X})$, in conjunction with the distribution $\mathbb{P}(N_1(\boldsymbol{x}) = n_1 | I_k(\boldsymbol{X}) \neq \emptyset)$. See Web Appendix E.

[19] The non-monotonicity in n of the curves presented in Figure 1 arises because for any streak length k there is no bias when $n = k + 1$ (because there are only two feasible sequences, which are equally likely), or in the limit (see Footnote 10).

3.1 GVT's analysis

Empirical approach

GVT's "Analysis of Conditional Probabilities" is their main test of hot hand shooting, and provides their only measure of the magnitude of the hot hand effect. The goal of their analysis is to determine whether a player's hit probability is higher following a streak of hits than it is following a streak of misses.[20] To this end, GVT reported each player i's shooting percentage conditional on having: (1) hit the last k shots, $\hat{P}^i(hit|k\ hits)$, and (2) missed the last k shots, $\hat{P}^i(hit|k\ misses)$, for streak lengths $k = 1, 2, 3$ (Table 4, p. 307).[21] After informally comparing these shooting percentages for individual players, GVT performed a paired t-test of whether $E[\hat{P}^i(hit|k\ hits) - \hat{P}^i(hit|k\ misses)] = 0$, for $k = 1, 2, 3$.[22,23]

In the remainder of this section, we focus our discussion on streaks of length three (or more), as in, e.g. Koehler and Conley (2003); Rao (2009b), given that: (1) shorter streak lengths exacerbate attenuation bias due to measurement error (see Footnote 23 and Appendix B), and (2) people typically perceive streaks as beginning with the third successive event (Carlson and Shu 2007). In any case, robustness checks using different streak lengths yield similar results (see Footnotes 31 and 35 in Section 3.3).

Data

GVT analyze shot sequences from basketball players in three contexts: NBA field goal data, NBA free-throw data, and a controlled shooting experiment with NCAA collegiate players. The shooting experiment was GVT's controlled test of hot hand shooting, designed for the purpose of "eliminating the effects of shot selection and defensive pressure" (p. 34), which makes it central to their main conclusions. Thus, we focus on this data below when discussing the relevance of the bias to GVT's

[20] GVT explicitly treat hot hand and streak shooting as synonymous (Gilovich et al. 1985, pp. 296-297). Miller and Sanjurjo (2014) provide an analysis that distinguishes between hot hand and cold hand shooting, and find hot hand shooting across all extant controlled shooting datasets, but little in the way of cold hand shooting. Thus, in the present analysis we use the terms streakiness and hot hand shooting interchangeably.

[21] We abuse our notation from Section 2 here in order to facilitate comparison with GVT's analysis: we use $\hat{P}^i(hit|k\ hits)$ for both the random variable $\hat{P}_k(\boldsymbol{X})$ and its realization $\hat{P}_k(\boldsymbol{x})$. Similarly, we use $\hat{P}^i(hit|k\ misses)$ for the proportion of successes on trials that immediately follow k consecutive failures.

[22] Under the null hypothesis the difference between each i's pair of shooting percentages is drawn from a normal distribution with mean zero.

[23] While GVT's analysis of conditional probabilities provides their only measure of the magnitude of the hot hand, they also analyze the number of runs, serial correlation, and variation of shooting percentage in 4-shot windows. Miller and Sanjurjo (2014) show that the runs and serial correlation tests, along with the conditional probability test for $k = 1$, all amount to roughly the same test, and moreover, that they are not sufficiently powered to identify hot hand shooting. The reason why is due to measurement error: the act of hitting a single shot is only a weak signal of a change in a player's underlying probability of success, which leads to an attenuation bias in the estimate of the increase in the probability of success associated with entering the hot state (see Appendix B and Stone (2012)'s work on measurement error when estimating autocorrelation in ability). The test of variation in 4-shot windows is even less powered than the aforementioned tests (Miller and Sanjurjo 2014; Wardrop 1999).

results.[24,25]

In GVT's controlled shooting experiment 26 players from the Cornell University Mens' (14) and Womens' (12) basketball teams participated in an incentivized shooting task. Each player shot 100 times at a distance from which the experimenters determined he/she would make around 50 percent of the shots. Following each shot the player had to change positions along two symmetric arcs—one facing the basket from the left, and the other from the right.

Results

In Columns 4 and 5 of Table 2 we use the raw data from GVT to reproduce the shooting percentages, $\hat{P}^i(hit|3\ hits)$ and $\hat{P}^i(hit|3\ misses)$, for each of the 26 players (these are identical to Columns 2 and 8 of Table 4 in GVT). As indicated in GVT, players on average hit .49 when on a hit streak, versus .45 when on a miss streak. GVT's paired t-test finds the difference to be statistically indistinguishable from zero, and we replicate this result ($p = .49$).

3.2 The bias in GVT's analysis

While GVT's null hypothesis that $E[\hat{P}^i(hit|k\ hits) - \hat{P}^i(hit|k\ misses)] = 0$ seems intuitively correct for a consistent shooter with a fixed probability of success p^i (i.i.d. Bernoulli), Theorem 1 reveals a flaw in this reasoning. In particular, we have established that $\hat{P}^i(hit|k\ hits)$ is expected to be less than p^i, and $\hat{P}^i(hit|k\ misses)$ greater than p^i (by symmetry). In fact, in Appendix A.3 we show that the difference $\hat{P}^i(hit|k\ hits) - \hat{P}^i(hit|k\ misses)$ is not only expected to be negative, but that its magnitude is more than double the bias in either of the respective proportions.[26]

[24] From the statistical point of view, the in-game field goal data that GVT analyze (Study 2: 76ers, 1980-81 season: 9 players, 48 home games) is not ideal for the study of hot hand shooting for reasons that are unrelated to the bias. The most notable concern with in-game field goal data is that the opposing team has incentive to make *costly* strategic adjustments to mitigate the impact of the "hot" player (Dixit and Nalebuff 1991, p. 17). This concern has been emphasized by researchers in the hot hand literature (Aharoni and Sarig 2011; Green and Zwiebel 2017), and is not merely theoretical, as it has a strong empirical basis. While GVT observed that a shooter's field goal percentage is lower after consecutive successes, subsequent studies have shown that with even partial controls for defensive pressure (and shot location), this effect is eliminated (Bocskocsky, Ezekowitz, and Stein 2014; Rao 2009a). Further, evidence of specific forms of strategic adjustment has been documented (Aharoni and Sarig 2011; Bocskocsky et al. 2014). See Miller and Sanjurjo (2014) for further details.

[25] The in-game free throw data that GVT analyze (Study 3: Celtics, 1980-81, 1981-82 seasons: 9 players), while arguably controlled, is not ideal for the study of hot hand shooting for a number of reasons: (1) hitting the first shot in a pair of isolated shots is not typically regarded by fans and players as hot hand shooting (Koehler and Conley 2003), presumably due to the high prior probability of success ($\approx .75$), (2) hitting a single shot is a weak signal of a player's underlying state, which can lead to severe measurement error (Arkes 2013; Stone 2012), (3) it is vulnerable to an omitted variable bias, as free throw pairs are relatively rare, and shots must be aggregated across games and seasons in order to have sufficient sample size (Miller and Sanjurjo 2014). In any event, subsequent studies of free throw data have found evidence that is inconsistent with the conclusions that GVT drew from the Celtics' data (Aharoni and Sarig 2011; Arkes 2010; Goldman and Rao 2012; Miller and Sanjurjo 2014; Wardrop 1995; Yaari and Eisenmann 2011).

[26] That the difference is expected to be negative does not follow immediately from Theorem 1, as the set of sequences for which the difference is well-defined is a strict subset of the set corresponding to either of the respective proportions. Nevertheless, the reasoning of the proof is similar. See Theorem 3 of Appendix A.3.

Table 2: *Columns 4 and 5 reproduce Columns 2 and 8 of Table 4 from Gilovich et al. (1985) (note: 3 hits (misses) includes streaks of 3, 4, 5, etc.). Column 6 reports the difference between the reported proportions, and column 7 adjusts for the bias (mean correction), based on each player's shooting percentage (probability in this case) and number of shots.*

					$\hat{D}_3 := \hat{P}(hit\|3\ hits) - \hat{P}(hit\|3\ misses)$	
Player	# shots	$\hat{P}(hit)$	$\hat{P}(hit\|3\ hits)$	$\hat{P}(hit\|3\ misses)$	GVT est.	bias adj.
Males						
1	100	.54	.50	.44	.06	.14
2	100	.35	.00	.43	−.43	−.33
3	100	.60	.60	.67	−.07	.02
4	90	.40	.33	.47	−.13	−.03
5	100	.42	.33	.75	−.42	−.33
6	100	.57	.65	.25	.40	.48
7	75	.56	.65	.29	.36	.47
8	50	.50	.57	.50	.07	.24
9	100	.54	.83	.35	.48	.56
10	100	.60	.57	.57	.00	.09
11	100	.58	.62	.57	.05	.14
12	100	.44	.43	.41	.02	.10
13	100	.61	.50	.40	.10	.19
14	100	.59	.60	.50	.10	.19
Females						
1	100	.48	.33	.67	−.33	−.25
2	100	.34	.40	.43	−.03	.07
3	100	.39	.50	.36	.14	.23
4	100	.32	.33	.27	.07	.17
5	100	.36	.20	.22	−.02	.08
6	100	.46	.29	.55	−.26	−.18
7	100	.41	.62	.32	.30	.39
8	100	.53	.73	.67	.07	.15
9	100	.45	.50	.46	.04	.12
10	100	.46	.71	.32	.40	.48
11	100	.53	.38	.50	−.12	−.04
12	100	.25	.	.32	.	.
Average		.47	.49	.45	.03	.13

Under GVT's design target of each player taking $n = 100$ shots and making half ($p = .5$) of them, we use the results from Section 2 and Appendix A.3 to find that the expected difference (and the strength of the bias) is -8 percentage points.[27] Therefore, the difference between the average proportion of +4 percentage points observed by GVT is actually +12 percentage points higher than the difference that would be expected from a Bernoulli i.i.d. shooter. Thus, the bias has long disguised evidence in GVT's data that may well indicate hot hand shooting.

3.3 A bias-corrected statistical analysis of GVT

A straightforward way to adjust for the bias in GVT's analysis is simply to shift the difference for each shooter by the amount of the corresponding bias, then repeat their paired t-test. While this test yields a statistically significant result ($p < .05$), the paired t-test limits statistical power because it reduces each player's performance to a single number, ignoring the number of shots that the player attempted in each category, i.e. "3 hits" and "3 misses." In addition, adjusting for the bias based on the assumption that $p = .5$ assumes that GVT's design target was met precisely.

As a result, for each player we again compute the bias under the null hypothesis that trials are i.i.d. Bernoulli (i.e. "consistent" shooting) but now with a probability of success equal to the player's observed shooting percentage (Column 3 of Table 2), and using the number of shots taken in each category to inform our standard errors. With this approach the average difference goes from +3 to a considerable +13 percentage points ($p < .01$, $S.E. = 4.7$pp).[28,29] To put the magnitude of +13 percentage points into perspective, the difference between the median three point shooter and the top three point shooter in the 2015-2016 NBA season was 12 percentage points.[30] Further, this is a *conservative* estimate because in practice the data generating processes (i.e. shooters) clearly differ from i.i.d. Bernoulli trials, and the bias becomes much larger under various models of hot hand shooting because of measurement error (see Appendix B).

GVT also informally discussed the heterogeneity across players, and asserted that most players shot relatively better when on a streak of misses than when on a streak of hits. By contrast, Figure 2 shows that once the bias correction is made to the differences 19 of the 25 players directionally exhibit hot hand shooting, which is itself significant ($p < .01$, binomial test).[31] Further, as indi-

[27] See Figure 4 in Appendix A.3 for the bias in the difference as n, p and k vary.

[28] The standard error is computed based on the assumption of independence across the 2600 trials, and normality. In particular, defining player i's difference $\hat{D}^i_k := \hat{P}^i(hit|k\ hits) - \hat{P}^i(hit|k\ misses)$, the variance satisfies $\widehat{Var}(\hat{D}^i_k) = \widehat{Var}(\hat{P}^i(hit|k\ hits)) + \widehat{Var}(\hat{P}^i(hit|k\ misses))$ for each player i. Simulations reveal that the associated $(1-\alpha)\times 100\%$ confidence intervals with radius $z_{\alpha/2} \times \widehat{Var}(\bar{D}_k)^{1/2}$ (where the mean difference is given by $\bar{D}_k := (1/n)\sum_{i=1}^{n} \hat{D}^i_k$) have the appropriate coverage—i.e. $(1-\alpha/2)\times 100\%$ of the time the true difference is greater than $\bar{D}_k - z_{\alpha/2} \times \widehat{Var}(\bar{D}_k)^{1/2}$, for both Bernoulli trials and the positive feedback model discussed in Section B.

[29] For an alternative approach that involves pooling shots across players, and yields similar results, see Appendix C.

[30] ESPN, "NBA Player 3-Point Shooting Statistics - 2015-16." http://www.espn.com/nba/statistics/player/_/stat/3-points [accessed September 24, 2016].

[31] Repeating the tests for longer ($k = 4$) or shorter ($k = 2$) streak lengths yields similar results that are consistent with

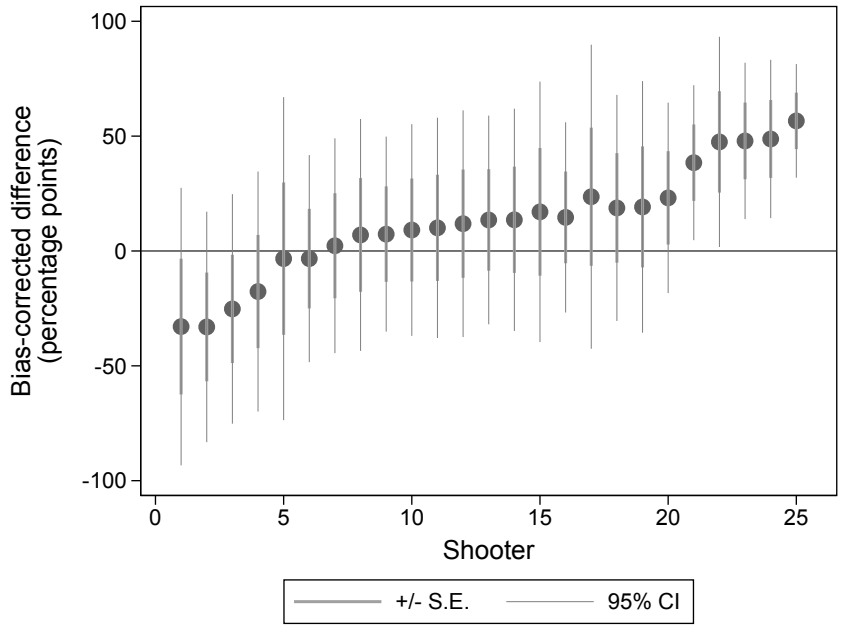

Figure 2: *The bias-corrected difference* $\hat{D}_3^i = \hat{P}^i(hit|3\ hits) - \hat{P}^i(hit|3\ misses)$ *for each player, under the assumption that his/her probability of success is equal to his/her overall shooting percentage.*

cated by the confidence intervals, t-tests reveal that 5 of the players exhibit statistically significant evidence of hot hand shooting ($p < .05$, t-test), which, for a set of 25 independent tests, is itself significant ($p < .01$, binomial test).

Non-parametric robustness test

As a robustness check we perform permutation tests, which are (by construction) invulnerable to the bias. The null hypothesis for a permutation test is that a player is a consistent shooter, i.e. has an i.i.d. fixed (unknown) probability of success. The first step to test for streak shooting in player i is to observe his/her shot sequence and compute the difference in proportions, $\hat{P}^i(hit|k\ hits) - \hat{P}^i(hit|k\ misses)$. The second step is to compute this difference for each unique rearrangement of the observed sequence; each of these *permutations* is equally likely because player i's probability of

the attenuation bias in estimated effect sizes discussed in Footnote 23. In particular, If we instead define a streak as beginning with 4 consecutive hits, which is a stronger signal of hot hand shooting, then the average bias-adjusted difference in proportions is 10 percentage points ($p = .07$, $S.E. = 6.9$, one-sided test), and four players exhibit statistically significant hot hand shooting ($p < .05$), which is itself significant ($p < .01$, binomial test). On the other hand, if we define a streak as beginning with 2 consecutive hits, which is a weaker signal of hot hand shooting, then the average bias-adjusted difference in proportions is 5.4 percentage points ($p < .05$, $S.E. = 3$, one-sided test), and four players exhibit statistically significant hot hand shooting ($p < .05$), which is itself significant ($p < .01$, binomial test).

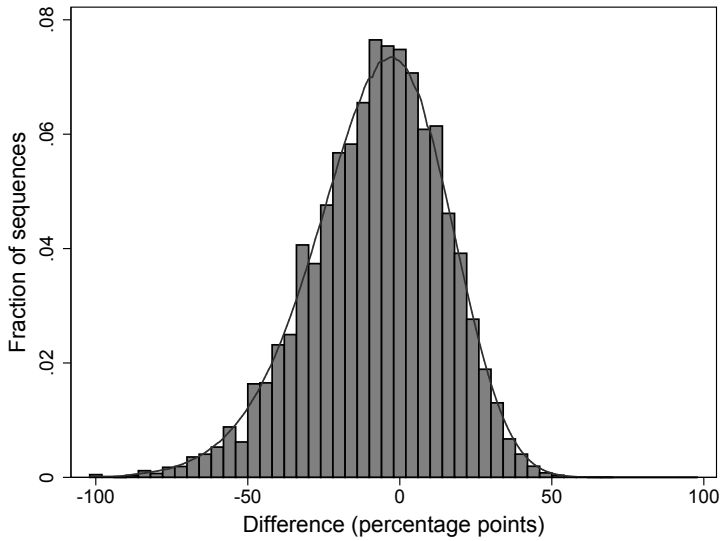

Figure 3: *The histogram and kernel density plot of the (exact) discrete probability distribution of* $\hat{P}^i(hit|k\ hits) - \hat{P}^i(hit|k\ misses)$*, a single player* i *with* $n = 100$ *and* $n_1 = 50$ *(using the formula for the distribution provided in the proof of Theorem 6, with a bin width of 4 percentage points).*[33]

success is fixed under the null hypothesis.[32] The set of unique differences computed in the second step, along with their associated relative frequencies, constitutes the exact sampling distribution of the difference under the null hypothesis (conditional on the observed number of hits). This distribution can then be used for statistical testing (See Appendix C.2 for details). The distribution is negative-skewed, and can be represented by histograms such as the one shown in Figure 3, which uses Theorem 6 of Web Appendix E to provide the *exact* distribution for a player who has hit 50 out of 100 shots.

Results of the permutation tests agree with those of the bias-corrected tests reported above. In particular, the average difference across shooters indicates hot hand shooting with a similar level of significance ($p < .01$).[34] Also as before, 5 individual players exhibit significant hot hand shooting ($p < .01$, binomial test).[35]

[32] Thus, the permutation procedure directly implements GVT's idea of comparing a "player's performance...to a sequence of hits and misses generated by tossing a coin" (Gilovich et al. 1985, p. 296)

[33] The values for the difference are grouped based on the first 6 decimal digits of precision. For this precision, the more than 10^{29} distinguishable sequences take on 19,048 distinct values. In the computation of the expected value in Figures 1 and 4, each difference is instead represented with the highest floating point precision available.

[34] The procedure in this pooled test involves stratifying the permutations by player. In particular, we conduct a test of the average of the standardized difference, where for each player the difference is standardized by shifting its mean and scaling its variance under H_0. In this case H_0: $\mathbb{P}(\text{success on trial } t \text{ for player } i) = p^i$ for all t, i.

[35] As in Footnote 31, the results of the permutation test are robust to varying streak length k.

3.4 The hot hand (and bias) in other controlled and semi-controlled studies

A close replication of GVT's controlled shooting experiment is found in Avugos, Bar-Eli, Ritov, and Sher (2013a), a study that mimics GVT's design and analysis, but with olympian rather than collegiate players, and fewer shots ($n = 40$) per player. From the authors' Table 1 (p. 6), one can derive the average $\hat{p}(hit|3\ hits)$ and $\hat{p}(hit|3\ misses)$ across players, which are roughly .52 and .54, respectively, yielding an average difference in shooting percentages of -2 percentage points.[36] However, Figure 4 in Appendix A.3 shows that the strength of the bias for $n = 40$ shots and $p = .5$ (the design target) is -.20. Thus, once the bias is corrected for in this small sample the average difference across shooters becomes roughly +18 percentage points.[37]

Koehler and Conley (2003) test for the hot hand in the NBA three point shooting contest, which has been described as an ideal setting in which to study the hot hand (Thaler and Sunstein 2008). The authors find no evidence of hot hand shooting in their analysis of four years of data. However, as in GVT and Avugos et al. (2013a), the conditional probability tests that the authors conduct are vulnerable to the bias. By contrast, Miller and Sanjurjo (2015b) collect 28 years of data, which yields 33 players that have taken at least 100 shots; using this dataset, we find that the average bias-corrected difference across players is +8 percentage points ($p < .01$).[38] Further, 8 of the 33 players exhibit significant hot hand shooting ($p < .05$), which itself is statistically significant ($p < .001$, binomial test).

The only other controlled shooting studies that we are aware of are Jagacinski et al. (1979) and Miller and Sanjurjo (2014).[39,40] Both studies have few shooters (6 and 8, respectively) but many shots across multiple shooting sessions for each player (540 and 900+ shots, respectively). The bias-adjusted average difference in the studies are +7 and +4 percentage points, respectively. In addition, Miller and Sanjurjo (2014) find substantial and persistent evidence of hot hand shooting in individual players.[41]

[36] We could not analyze the raw data because the authors declined to provide it to us. The data that represents a close replication of GVT is from the betting game phase. Using Table 1, we have $\hat{p}(hit|3\ hits) = (.56 + .52)/2$ and $\hat{p}(hit|3\ misses) = (.54 + .49)/2$, which is the average of the shooting percentage of Group A in Phase 1 with that of Group B from Phase 2.

[37] The authors also had another treatment, in which they had shooters rate, before each shot, from 0-100% on a certainty scale whether they would hit the next shot. If we repeat the analysis on the data from this treatment then the average $\hat{p}(hit|3\ hits)$ and $\hat{p}(hit|3\ misses)$ across players are roughly .56 and .65, respectively, yielding an average difference of -9 percentage points, and a bias-adjusted difference of +11 percentage points.

[38] Miller and Sanjurjo (2015b) also implement the unbiased permutation test procedure of Section 3.3.

[39] The one exception is a controlled shooting study that involved a single shooter Wardrop (1999). After personal communication with the shooter, who conducted the study herself (recording her own shots), we viewed it as not having sufficient control to warrant analysis.

[40] We thank Tom Gilovich for bringing the study of Jagacinski et al. to our attention. It had gone uncited in the hot hand literature until Miller and Sanjurjo (2014).

[41] See Avugos, Köppen, Czienskowski, Raab, and Bar-Eli (2013b) for a meta-analysis of the hot hand, which includes sports besides basketball. Tversky and Gilovich (1989) argue that evidence for the hot hand in other sports is not relevant to their main conclusion because so long as the hot hand does not exist in basketball, then the perception of

Thus, once the bias is accounted for, *conservative* estimates of hot hand effect sizes across all extant controlled and semi-controlled shooting studies are consistently moderate to large.[42]

3.5 Belief in the Hot Hand

The results of our reanalysis of GVT's data lead us to a conclusion that is the opposite of theirs: belief in the hot hand is not a cognitive illusion. Nevertheless, it remains possible, perhaps even likely, that professional players and coaches sometimes infer the presence of a hot hand when it does not exist. Similarly, even when in the presence of the hot hand, players may overestimate its influence and respond too strongly to it. By contrast, a hot hand might also go undetected, or be underestimated (Stone and Arkes 2017). These questions are important because understanding the extent to which decision makers' beliefs and behavior do not correspond to the actual degree of hot hand shooting could have considerable implications for decision-making more generally.

While GVT's main conclusion was of a binary nature, i.e. based on the question of whether belief in the hot hand is either fallacious or not, they explored hot hand beliefs via a survey of player and coach beliefs, and an incentivized betting task with the Cornell players. In the survey they find that the near universal beliefs in the hot hand do not accord with the lack of hot hand shooting evidence that resulted from their analysis of the shooting data, and in the betting task they found that players were incapable of predicting upcoming shot outcomes successfully, which suggests that even if there were a hot hand, it could not be detected successfully.

However, in light of the results presented in the present paper subjects' responses in GVT's unincentivized survey are actually qualitatively consistent with the evidence presented above.[43] More substantively, GVT's statistical analysis of betting data has recently been shown to be considerably underpowered, as the authors conduct many separate individual bettor level tests rather than pooling the data across bettors (Miller and Sanjurjo 2017). In addition, GVT misinterpret their measures of bettors' ability to predict. In light of these limitations, Miller and Sanjurjo (2017) reanalyze GVT's betting data, and find that players on average shoot around +7 percentage points higher when bettors have predicted that the shot will be a hit, rather than a miss ($p < .001$). This increase is comparable in magnitude to an NBA shooter going from slightly above average to elite in three point percentage.[44]

---

the hot hand by fans, players and coaches must necessarily be a cognitive illusion (see also Alter and Oppenheimer [2006]).

[42] The magnitudes of all estimated effect sizes are conservative for two reasons: (1) if a player's probability of success is not driven merely by feedback from previous shots, but also by other time-varying player (and environment) specific factors, then the act of hitting consecutive shots will serve as only a noisy proxy of the hot state, resulting in measurement error, and an attenuation bias in the estimate (see Appendix B), and (2) if the effect of consecutive successes on subsequent success is heterogenous in magnitude (and sign) across players, then an average measure will underestimate how strong the effect can be in certain players.

[43] See Appendix B of Miller and Sanjurjo (2017) for details.

[44] ESPN, "NBA Player 3-Point Shooting Statistics - 2015-16." http://www.espn.com/nba/statistics/player/_/stat/3-

Miller and Sanjurjo (2014) present complementary evidence on beliefs, in which semi-professional players rank their teammates' respective increases in shooting percentage when on a streak of three hits (relative to their base rates) in a shooting experiment that the rankers do not observe. Players' rankings are found to be highly correlated with their teammates' actual increases in shooting percentage in this out-of-sample test, yielding an average correlation of -0.60 ($p < .0001$; where 1 is the rank of the shooter with the perceived largest percentage point increase).

In sum, while it remains possible that professional players' and coaches' hot hand beliefs are poorly calibrated, this claim is not clearly supported by the existing body of evidence.

4 Application to the Gambler's Fallacy

> *Why, if the gambler's fallacy is truly fallacious, does it persist? Why is it not corrected as a consequence of experience with random events?* (Nickerson 2002)

A classic result on the human perception of randomness in sequential data is that people believe the outcomes of randomly generated sequences to alternate more than they actually do. For example, if a (fair) coin flip lands heads, then a tails is thought to be more likely on the next flip (Bar-Hillel and Wagenaar 1991; Nickerson 2002; Oskarsson, Boven, McClelland, and Hastie 2009; Rabin 2002).[45] Further, as a streak of identical outcomes (e.g. heads) increases in length, it is believed that the alternation rate on the outcome that follows becomes even larger, which is known as the *Gambler's Fallacy* (Bar-Hillel and Wagenaar 1991).[46] Gambler's fallacy beliefs are widespread among novice gamblers, with adherents that have included at least one historically eminent mathematician (D'Alembert 1761, pp. 13-14).[47] The fallacy has been attributed to the mistaken belief in the "Law of Small Numbers," by which large sample properties are incorrectly thought to also hold within small samples (Tversky and Kahneman 1971), so if, for example, several heads flips have occurred in a row, then tails is deemed more likely on the next flip to help "balance things out."

points [accessed September 24, 2016].

[45] This *alternation bias* is also sometimes referred to as *negative recency bias*.

[46] For simplicity, in the following discussion we assume that a decision maker keeps track of the alternation rate of a single outcome (e.g. for heads, $1 - \hat{p}(H|H)$), which seems especially reasonable for applications in which outcomes appear qualitatively different (e.g. rainy/sunny days). On the other hand, in the case of flipping a fair coin there may be no need to discriminate between an alternation that follows heads, or tails, respectively. In this special case, the overall alternation rate, (# alternations for streaks of length 1)/(number of flips $-$ 1), is expected to be 0.5. Nevertheless, it is easy to demonstrate that the overall alternation rate computed for any other streak length ($k > 1$) is expected to be strictly greater than 0.5 (the explicit formula can be derived using an argument identical to that used in Theorem 6).

[47] In particular, D'Alembert famously argued in favor of his gambler's fallacy beliefs. In response to the problem: "When a fair coin is tossed, given that heads have occurred three times in a row, what is the probability that the next toss is a tail?" D'Alembert argued that the probability of a tail is greater than 1/2 because it is unlikely that a probable event will never occur in a finite sequence of trials (D'Alembert 1761, pp. 13-14); see Gorroochurn (2012, p. 124) for a discussion.

The opening quote by Nickerson (2002) poses an important question: given that the gambler's fallacy is an error, why does experience fail to correct it? One explanation is that there may be insufficient incentive, or opportunity to learn, given that people are often mere passive observers of random sequential data, or have little at stake.[48] However, this explanation is unsatisfying as it presupposes no advantage to holding correct beliefs per se, and ignores their option value. Therefore a potentially more satisfying explanation for the persistence of the gambler's fallacy is one that is capable of addressing how it could be robust to experience.

Based on the results from Section 2, we propose a simple model of how a mistaken belief in the gambler's fallacy can persist. Consider a decision maker (DM) who repeatedly encounters finite length sequences of "successes" and "failures." DM begins with prior beliefs regarding the conditional probability of "success," given that an outcome immediately follows k consecutive successes. Naturally, for each encounter with a finite sequence, DM attends to the outcomes that immediately follow k consecutive successes, and updates accordingly.

Importantly, when updating his prior, we allow for the possibility that DM focuses on the *strength evidence*, i.e. the proportion of successes on the outcomes that follow a streak of successes, rather than the *weight of evidence*, i.e. the effective sample size used in the calculation of the proportion. This feature of the model is consistent with results on how people weight evidence when updating their beliefs (Griffin and Tversky 1992). In particular, sample size neglect has been documented extensively (Benjamin, Rabin, and Raymond 2014; Kahneman and Tversky 1972), and is sometimes attributed to working memory capacity limitations (Kareev 2000).

More formally, DM has beliefs regarding the conditional probability $\theta = \mathbb{P}(X_i = 1 | \prod_{j=i-k}^{i-1} X_j = 1)$, with a prior $\mu(\theta)$ over the support $[0, 1]$. When DM encounters a sequence $\{X_i\}_{i=1}^{\ell}$, he attends to those trials that immediately follow k (or more) successes, defined as $I' := \{i \in \{k+1, \ldots, \ell\} : \prod_{j=i-k}^{i-1} X_j = 1\}$. Thus, he effectively observes $\boldsymbol{Y} := (Y_i)_{i=1}^{M} = (X_i)_{i \in I'}$, where $M := |I'|$. Whenever a sequence contains trials worthy of attending to (i.e. $I' \neq \emptyset$), DM calculates the proportion of successes $\hat{p}$ on those trials, weighting it according to his perception of the sample size $w = w(M)$. Given w, DM's posterior distribution for θ follows:

$$p(\theta | \boldsymbol{Y}) = \frac{\theta^{w\hat{p}}(1-\theta)^{w(1-\hat{p})}\mu(\theta)}{\int \theta'^{w\hat{p}}(1-\theta')^{w(1-\hat{p})}\mu(\theta')}$$

Using this simple setup, we now briefly explore under what conditions gambler's fallacy beliefs can persist. Suppose that DM encounters an i.i.d. sequence of Bernoulli random variables $\{X_i\}_{i=1}^{\ell}$

[48] In casino games such as roulette, people make active decisions based on events that are sequentially independent. While there is typically no additional cost to placing one's bets on an event that hasn't occurred for some time, rather than another event, the fallacy can be costly if it leads one to bet larger amounts (given that expected returns are negative). See Rabin (2002), Ayton and Fischer (2004), Croson and Sundali (2005), and Chen, Moskowitz, and Shue (2014) for further discussion.

in which each trial has probability of success p. Further, DM is a believer in the law of small numbers, and holds a strong prior towards gambler's fallacy beliefs. In the case that he observes few sequences, experience will have little effect on DM's beliefs, regardless of whether or not he accounts for sample size. In the case that DM observes many sequences, the degree to which his gambler's fallacy beliefs persist will depend on (1) the extent to which he neglects sample size $w(\cdot)$, (2) the length of the sequences he is exposed to (ℓ), and (3) the threshold streak length (k) that leads him to attend to outcomes. To illustrate the role of sample size sensitivity, let $w(M) := M^\alpha$ for some $\alpha \geq 0$. On one extreme, DM does not discriminate between different sample sizes, weighting all proportions the same with $\alpha = 0$. In this case, as the number of sequences increases, DM's beliefs, μ, approach point mass on the fully biased (unweighted) expected proportion given in Section 2.[49] As in the gambler's fallacy, these beliefs are strictly less than p, and become more biased as k increases. On the other extreme DM may fully discriminate between sample sizes, weighting proportions according to their sample size with $\alpha = 1$. In this case, there is no asymptotic bias in the proportion, so his beliefs will be correct in the limit.[50,51] Perhaps more plausibly, if DM has *some* degree of sensitivity to sample size then the asymptotic beliefs will be biased, and will lie somewhere between the two extremes just given, depending on the sensitivity $0 < \alpha < 1$.

We are not the first to propose that beliefs may be influenced by the statistical properties of finite samples. For example, in the psychology literature, it has been proposed that associations learned via experience may be influenced by the smallness of samples that people are typically exposed to (Kareev 1995a,b, 2000; Kareev, Lieberman, and Lev 1997).[52] More recently, and closely

[49] To see this, first note that DM will observe a sequence of i.i.d. proportions $\hat{p}_i$, with $E[\hat{p}_i] := \theta^* < p$ (by Theorem 1). The strong law of large numbers applies in this case, and $\bar{p}_n := \sum_{i=1}^n \hat{p}_i/n$ will converge to θ^* almost surely (a.s.). After the n^{th} sequence, DM's posterior odds in favor of θ (relative θ^*) become $\left[\left(\frac{\theta}{\theta^*}\right)^{\bar{p}_n}\left(\frac{1-\theta}{1-\theta^*}\right)^{1-\bar{p}_n}\right]^n \frac{\mu(\theta)}{\mu(\theta^*)}$. The posterior probability will converge to point mass on θ^* (a.s.) because the posterior odds in favor of θ converge to zero (a.s.) for all $\theta \neq \theta^*$, which follows because $\theta \neq \theta^*$ implies $\left(\frac{\theta}{\theta^*}\right)^{\theta^*}\left(\frac{1-\theta}{1-\theta^*}\right)^{1-\theta^*} < 1$.

[50] The weighted average satisfies $\sum_{i=1}^n M_i\hat{p}_i/\sum_{i=1}^n M_i = \sum_{i=1}^n \sum_{j=1}^{M_i} x_{ij}/\sum_{i=1}^n M_i$, where x_{ij} is the j^{th} outcome from the i^{th} sequence. This weighted average is the maximum likelihood estimator for the transition probability p from the state "a trial is immediately preceded by k successes" to itself (with $\sum_{i=1}^n M_i$ total observations), in the associated irreducible and ergodic 2^k-state Markov chain, and converges to the transition probability p almost surely (see e.g. Grimmett and Stirzaker (2001, p. 358)). Following the argument in footnote 49, we conclude the DM's step n posterior odds in favor of θ relative to p converge to 0 (a.s.), which implies the asymptotic posterior probability will have point mass on p (a.s.).

[51] There are two alternative statistical approaches that do not require an infinite sample of sequences for the decision maker to obtain an unbiased estimate of the conditional probability, see footnote 9 for details.

[52] In a review article, Kareev (2000) observes that the sampling distribution of the correlation coefficient between any two variables is strongly skewed for small samples, which implies that measures of central tendency in the sampling distribution of the correlation can be substantially different than the true correlation, which can influence belief formation. Interestingly, in earlier work Kareev (1992) observes a finite sample property for the alternation rate in a sequence. In particular, while the expected overall alternation rate for streaks of length $k = 1$ is equal to 0.5 (when not distinguishing between a preceding heads or tails), people's experience can be made to be consistent with an alternation rate that is greater than 0.5 if the set of observable sequences that they are exposed to is restricted to those that are subjectively "typical" (e.g. those with an overall success rate close to 0.5). In fact, for streaks of length

related, Hahn and Warren (2009) conjecture that the gambler's fallacy may arise from the small sample properties of the distribution of finite length strings, which relates to the *overlapping words paradox* (Guibas and Odlyzko [1981]; also see Web Appendix F.2). In particular, the authors note that in a sequence of length $n > 4$, the pattern HHHT is more likely to occur than the pattern HHHH, which may explain why people believe that the probability of tails is greater than 1/2 after three heads in a row. While this conjecture has sparked some debate, it does not yet appear to have been empirically tested (Hahn and Warren 2010a,b; Sun, Tweney, and Wang 2010a,b; Sun and Wang 2010).[53] In a formal comment based on an earlier version of this paper, Sun and Wang (2015) relate the bias that we find to this debate, but argue that its implications for human judgement and decision-making are limited. Instead, the authors emphasize the primacy of the waiting time distribution of finite length patterns in infinite sequences, rather than the distribution of sample statistics in finite length sequences.

In our view, this model offers a plausible account for the persistence of the gambler's fallacy, which also has testable implications. First, in terms of plausibility, there is ample evidence that people tend to adapt to the natural statistics in their environment (Atick 1992; Simoncelli and Olshausen 2001), with the sample proportion being an example of a statistic that humans find intuitive and tend to assess relatively accurately (Garthwaite, Kadane, and O'Hagan 2005). Second, in terms of testability, our model predicts that the magnitude of bias in peoples' beliefs should depend on the following measurable and experimentally manipulable factors: (1) the length of sequences (ℓ), (2) the streak lengths (k) that immediately precede the outcomes attended to, and (3) sensitivity to sample size $w(\cdot)$.

The explanation provided here can be thought of as complementary to Rabin (2002) and Rabin and Vayanos (2010). In particular, it provides a structural account for why the central behavioral primitive of their model—the belief in the law of small numbers—should persist in the face of experience. Further, our approach relates to Benjamin et al. (2014) in that it illustrates how a limited sensitivity to sample size can affect inference.

5 Conclusion

We prove that in a finite sequence of data that is generated by repeated realizations of a binary i.i.d. random variable, the expected proportion of successes, on those realizations that immediately follow a streak of successes, is *strictly less than* the underlying probability of success. The mechanism

$k > 1$, this restriction is not necessary, as the expected overall alternation rate across all sequences is greater than 0.5 (the explicit formula that demonstrates this can be derived using an argument identical to that used in Theorem 6).

[53] The focus on fixed length string patterns has a few limitations with regard to testability: (1) some patterns with lower associated proportions e.g. HTHT, have much lower probabilities than patterns with high associated proportions, such as TTHH, (2) for most patterns the difference in the probability is small, even for patterns in which the proportion associated with the pattern varies considerably.

is a form of selection bias that arises from the sequential structure of the finite data. A direct implication of the bias is that empirical approaches of the most prominent studies in the hot hand fallacy literature are incorrect. Upon correcting for the bias we find that the data that had previously been interpreted as providing substantial evidence that belief in the hot hand is a fallacy, reverses, instead providing substantial evidence that it is not a fallacy to believe in the hot hand. Another implication of the bias is a novel structural explanation for the persistence of gambler's fallacy beliefs in the face of experience. Finally, we find that the respective errors of the gambler and hot hand fallacy researcher are analogous: the gambler sees reversal in an i.i.d. process, while the researcher sees an i.i.d. process when there is momentum.

References

Aharoni, G. and O. H. Sarig (2011): "Hot hands and equilibrium," *Applied Economics*, 44, 2309–2320.

Alter, A. L. and D. M. Oppenheimer (2006): "From a fixation on sports to an exploration of mechanism: The past, present, and future of hot hand research," *Thinking & Reasoning*, 12, 431–444.

Arkes, J. (2010): "Revisiting the Hot Hand Theory with Free Throw Data in a Multivariate Framework," *Journal of Quantitative Analysis in Sports*, 6.

——— (2013): "Misses in 'Hot Hand' Research," *Journal of Sports Economics*, 14, 401–410.

Atick, J. J. (1992): "Could information theory provide an ecological theory of sensory processing?" *Network: Computation in neural systems*, 3, 213–251.

Avugos, S., M. Bar-Eli, I. Ritov, and E. Sher (2013a): "The elusive reality of efficacy performance cycles in basketball shooting: analysis of players' performance under invariant conditions," *International Journal of Sport and Exercise Psychology*, 11, 184–202.

Avugos, S., J. Köppen, U. Czienskowski, M. Raab, and M. Bar-Eli (2013b): "The "hot hand" reconsidered: A meta-analytic approach," *Psychology of Sport and Exercise*, 14, 21–27.

Ayton, P. and I. Fischer (2004): "The hot hand fallacy and the gamblers fallacy: Two faces of subjective randomness?" *Memory & Cognition*, 21, 1369–1378.

Bai, D. S. (1975): "Efficient Estimation of Transition Probabilities in a Markov Chain," *The Annals of Statistics*, 3, 1305–1317.

Balakrishnan, N. and M. V. Koutras (2011): *Runs and scans with applications*, vol. 764, John Wiley & Sons.

Bar-Hillel, M. and W. A. Wagenaar (1991): "The perception of randomness," *Advances in Applied Mathematics*, 12, 428–454.

Benjamin, D. J., M. Rabin, and C. Raymond (2014): "A Model of Non-Belief in the Law of Large Numbers," Working Paper.

Berkson, J. (1946): "Limitations of the application of fourfold table analysis to hospital data," *Biometrics Bulletin*, 47–53.

Bocskocsky, A., J. Ezekowitz, and C. Stein (2014): "The Hot Hand: A New Approach to an Old 'Fallacy'," 8th Annual Mit Sloan Sports Analytics Conference.

Carlson, K. A. and S. B. Shu (2007): "The rule of three: how the third event signals the emergence of a streak," *Organizational Behavior and Human Decision Processes*, 104, 113–121.

Chen, D., T. J. Moskowitz, and K. Shue (2014): "Decision-Making under the Gamblers Fallacy: Evidence from Asylum Judges, Loan Officers, and Baseball Umpires," Working Paper.

Croson, R. and J. Sundali (2005): "The Gamblers Fallacy and the Hot Hand: Empirical Data from Casinos," *Journal of Risk and Uncertainty*, 30, 195–209.

D'Alembert, J. (1761): *Opuscules mathmatiques*, David, Paris.

Dixit, A. K. and B. J. Nalebuff (1991): *Thinking Strategically: The Competitive Edge in Business, Politics, and Everyday Life*, W.W. Norton & Company.

Friedman, D. (1998): "Monty Hall's Three Doors: Construction and Deconstruction of a Choice Anomaly," *American Economic Review*, 88, 933–946.

Garthwaite, P. H., J. B. Kadane, and A. O'Hagan (2005): "Statistical Methods for Eliciting Probability Distributions," *Journal of the American Statistical Association*, 100, 680–700.

Gibbons, J. D. and S. Chakraborti (2010): *Nonparametric Statistical Inference*, New York: CRC Press, Boca Raton, Florida.

Gilovich, T., R. Vallone, and A. Tversky (1985): "The Hot Hand in Basketball: On the Misperception of Random Sequences," *Cognitive Psychology*, 17, 295–314.

Goldman, M. and J. M. Rao (2012): "Effort vs. Concentration: The Asymmetric Impact of Pressure on NBA Performance," 6th Annual Mit Sloan Sports Analytics Conference.

Gorroochurn, P. (2012): *Classic problems of probability*, New Jersey: John Wiley & Sons.

Green, B. S. and J. Zwiebel (2017): "The Hot Hand Fallacy: Cognitive Mistakes or Equilibrium Adjustments?" *Management Science*, working Paper.

Griffin, D. and A. Tversky (1992): "The weighing of evidence and the determinants of confidence," *Cognitive Psychology*, 24, 411–435.

Grimmett, G. R. and D. R. Stirzaker (2001): *Probability and Random Processes*, Oxford University Press.

Guibas, L. J. and A. M. Odlyzko (1981): "String overlaps, pattern matching, and nontransitive games," *Journal of Combinatorial Theory, Series A*, 30, 183–208.

Hahn, U. and P. A. Warren (2009): "Perceptions of randomness: why three heads are better than four," *Psychological Review*, 116, 454–461.

——— (2010a): "Postscript: All together now: 'Three heads are better than four'." *Psychological Review*, 117, 711–711.

——— (2010b): "Why three heads are a better bet than four: A reply to Sun, Tweney, and Wang (2010)." *Psychological Review*, 117, 706–711.

Haldane, J. B. S. (1945): "On a Method of Estimating Frequencies," *Biometrika*, 33, 222–225.

Imbens, G. W. and M. Kolesar (2016): "Robust Standard Errors in Small Samples: Some Practical Advice," Working Paper, March.

Jagacinski, R. J., K. M. Newell, and P. D. Isaac (1979): "Predicting the Success of a Basketball Shot at Various Stages of Execution," *Journal of Sport Psychology*, 1, 301–310.

Kahneman, D. (2011): *Thinking, Fast and Slow*, Farrar, Straus and Giroux.

Kahneman, D. and A. Tversky (1972): "Subjective Probability: A Judgement of Representativeness," *Cognitive Psychology*, 3, 430–454.

Kareev, Y. (1992): "Not that bad after all: Generation of random sequences." *Journal of Experimental Psychology: Human Perception and Performance*, 18, 1189–1194.

——— (1995a): "Positive bias in the perception of covariation." *Psychological Review*, 102, 490–502.

——— (1995b): "Through a narrow window: working memory capacity and the detection of covariation," *Cognition*, 56, 263–269.

——— (2000): "Seven (indeed, plus or minus two) and the detection of correlations." *Psychological Review*, 107, 397–402.

Kareev, Y., I. Lieberman, and M. Lev (1997): "Through a narrow window: Sample size and the perception of correlation." *Journal of Experimental Psychology: General*, 126, 278–287.

Koehler, J. J. and C. A. Conley (2003): "The "hot hand" myth in professional basketball," *Journal of Sport and Exercise Psychology*, 25, 253–259.

Konold, C. (1995): "Confessions of a coin flipper and would-be instructor," *The American Statistician*, 49, 203–209.

Miller, J. B. and A. Sanjurjo (2014): "A Cold Shower for the Hot Hand Fallacy," Working Paper.

——— (2015a): "A Bridge from Monty Hall to the (Anti-)Hot Hand: Restricted Choice, Selection Bias, and Empirical Practice," Working Paper, December 31.

——— (2015b): "Is the Belief in the Hot Hand a *Fallacy* in the NBA Three Point Shootout?" Working Paper.

——— (2017): "A Visible Hand? Betting on the hot hand in Gilovich, Vallone, and Tversky (1985)," *Working Paper*.

Nalebuff, B. (1987): "Puzzles: Choose a curtain, duel-ity, two point conversions, and more," *Journal of Economic Perspectives*, 157–163.

Nerlove, M. (1971): "Further Evidence on the Estimation of Dynamic Economic Relations from a Time Series of Cross Sections," *Econometrica*, 39, 359–382.

Neyman, J. and E. L. Scott (1948): "Consistent Estimates Based on Partially Consistent Observations," *Econometrica*, 16, 1–32.

Nickell, S. (1981): "Biases in Dynamic Models with Fixed Effects," *Econometrica*, 49, 1417–1426.

Nickerson, R. S. (2002): "The production and perception of randomness," *Psychological Review*, 109, 350–357.

——— (2007): "Penney Ante: Counterintuitive Probabilities in Coin Tossing," *The UMAPJournal*.

Oskarsson, A. T., L. V. Boven, G. H. McClelland, and R. Hastie (2009): "What's next? Judging sequences of binary events," *Psychological Bulletin*, 135, 262–385.

Rabin, M. (2002): "Inference by Believers in the Law of Small Numbers," *Quarterly Journal of Economics*, 117, 775–816.

Rabin, M. and D. Vayanos (2010): "The Gamblers and Hot-Hand Fallacies: Theory and Applications," *Review of Economic Studies*, 77, 730–778.

Rao, J. M. (2009a): "Experts' Perceptions of Autocorrelation: The Hot Hand Fallacy Among Professional Basketball Players," Working Paper.

——— (2009b): "When the Gambler's Fallacy becomes the Hot Hand Fallacy: An Experiment with Experts and Novices," Working Paper.

Rinott, Y. and M. Bar-Hillel (2015): "Comments on a 'Hot Hand' Paper by Miller and Sanjurjo," Federmann Center For The Study Of Rationality, The Hebrew University Of Jerusalem. Discussion Paper # 688 (August 11).

Riordan, J. (1958): *An Introduction to Combinatorial Analysis*, New York: John Wiley & Sons.

Roberts, R. S., W. O. Spitzer, T. Delmore, and D. L. Sackett (1978): "An empirical demonstration of Berkson's bias," *Journal of Chronic Diseases*, 31, 119–128.

Selvin, S. (1975): "A Problem in Probability (letter to the editor)," *The American Statistician*, 29, 67.

Shaman, P. and R. A. Stine (1988): "The bias of autoregressive coefficient estimators," *Journal of the American Statistical Association*, 83, 842–848.

Simoncelli, E. P. and B. A. Olshausen (2001): "Natural Image Statistics And Neural Representation," *Annual Review of Neuroscience*, 24, 1193–1216.

Stone, D. F. (2012): "Measurement error and the hot hand," *The American Statistician*, 66, 61–66, working paper.

Stone, D. F. and J. Arkes (2017): "March Madness? Underreaction to Hot and Cold Hands in NCAA Basketball," *Working Paper*, –.

Sun, Y., R. D. Tweney, and H. Wang (2010a): "Occurrence and nonoccurrence of random sequences: Comment on Hahn and Warren (2009)." *Psychological Review*, 117, 697–703.

——— (2010b): "Postscript: Untangling the gambler's fallacy." *Psychological Review*, 117, 704–705.

Sun, Y. and H. Wang (2010): "Gamblers fallacy, hot hand belief, and the time of patterns," *Judgement and Decision Making*, 5, 124–132.

——— (2015): "Alternation Bias as a Consequence of Pattern Overlap: Comments on Miller and Sanjurjo (2015)," Working Paper, November 8.

Thaler, R. H. and C. R. Sunstein (2008): *Nudge: Improving Decisions About Health, Wealth, and Happiness*, Yale University Press.

Tversky, A. and T. Gilovich (1989): "The cold facts about the "hot hand" in basketball," *Chance*, 2, 16–21.

Tversky, A. and D. Kahneman (1971): "Belief in the Law of Small Numbers," *Psychological Bulletin*, 2, 105–110.

Vos Savant, M. (1990): "Ask Marilyn," *Parade Magazine*, 15.

Wardrop, R. L. (1995): "Simpson's Paradox and the Hot Hand in Basketball," *The American Statistician*, 49, 24–28.

——— (1999): "Statistical Tests for the Hot-Hand in Basketball in a Controlled Setting," *Unpublished manuscript*, 1, 1–20.

Yaari, G. and S. Eisenmann (2011): "The Hot (Invisible?) Hand: Can Time Sequence Patterns of Success/Failure in Sports Be Modeled as Repeated Random Independent Trials?" *PLoS One*, 6, 1–10.

Yule, G. U. (1926): "Why do we Sometimes get Nonsense-Correlations between Time-Series?–A Study in Sampling and the Nature of Time-Series," *Journal of the Royal Statistical Society*, 89, 1–63.

A Appendix: Section 2 Proofs

A.1 Proof of Theorem 1 (Section 2)

Define $F := \{\boldsymbol{x} \in \{0,1\}^n : I_k(\boldsymbol{x}) \neq \emptyset\}$ to be the sample space of sequences for which $\hat{P}_k(\boldsymbol{X})$ is well defined. The probability distribution over F is given by $\mathbb{P}(A|F) := \mathbb{P}(A \cap F)/\mathbb{P}(F)$ for $A \subseteq \{0,1\}^n$, where $\mathbb{P}(\boldsymbol{X} = \boldsymbol{x}) = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}$.

Let the random variable X_τ represent the outcome of the randomly "drawn" trial τ, which is selected as a result of the two-stage procedure that: (1) draws a sequence $\boldsymbol{x}$ at random from F, according to the distribution $\mathbb{P}(\boldsymbol{X} = \boldsymbol{x}|F)$, and (2) draws a trial τ at random from $\{k+1, \ldots, n\}$, according to the distribution $\mathbb{P}(\tau = t|\boldsymbol{X} = \boldsymbol{x})$. Let τ be a uniform draw from the trials in sequence $\boldsymbol{X}$ that immediately follow k consecutive successes, i.e. for $\boldsymbol{x} \in F$, $\mathbb{P}(\tau = t|\boldsymbol{X} = \boldsymbol{x}) = 1/|I_k(\boldsymbol{x})|$ for $t \in I_k(\boldsymbol{x})$, and $\mathbb{P}(\tau = t|\boldsymbol{X} = \boldsymbol{x}) = 0$ for $t \in I_k(\boldsymbol{x})^C \cap \{k+1, \ldots, n\}$.[54] It follows that the unconditional probability distribution of τ over all trials that can possibly follow k consecutive successes is given by $\mathbb{P}(\tau = t|F) = \sum_{\boldsymbol{x} \in F} \mathbb{P}(\tau = t|\boldsymbol{X} = \boldsymbol{x}, F)\mathbb{P}(\boldsymbol{X} = \boldsymbol{x}|F)$, for $t \in \{k+1, \ldots, n\}$. The probability that this randomly drawn trial is a success, $\mathbb{P}(X_\tau = 1|F)$, must be equal to the expected proportion, $E[\hat{P}_k(\boldsymbol{X})|F]$.[55]

Note that $\mathbb{P}(X_\tau = 1|F) = \sum_{t=k+1}^n \mathbb{P}(X_t = 1|\tau = t, F)\mathbb{P}(\tau = t|F)$, and $\mathbb{P}(\tau = t|F) > 0$ for $t \in \{k+1, \ldots, n\}$. Below, we demonstrate that $\mathbb{P}(X_t = 1|\tau = t, F) < p$ when $t < n$, and that $\mathbb{P}(X_t = 1|\tau = n, F) = p$, which, taken together, guarantee that $\mathbb{P}(X_\tau = 1|F) < p$.

First we observe that $\mathbb{P}(X_t = 1|\tau = t, F) = \mathbb{P}(X_t = 1|\tau = t, F_t)$, where $F_t := \{\boldsymbol{x} \in \{0,1\}^n : \prod_{i=t-k}^{t-1} x_i = 1\}$. Bayes Rule then yields:

$$\begin{aligned}\frac{\mathbb{P}(X_t = 1|\tau = t, F_t)}{\mathbb{P}(X_t = 0|\tau = t, F_t)} &= \frac{\mathbb{P}\left(\tau = t\,|X_t = 1, F_t\right)\mathbb{P}(X_t = 1|F_t)}{\mathbb{P}\left(\tau = t\,|X_t = 0, F_t\right)\mathbb{P}(X_t = 0|F_t)} \\ &= \frac{\mathbb{P}\left(\tau = t\,|X_t = 1, F_t\right)}{\mathbb{P}\left(\tau = t\,|X_t = 0, F_t\right)}\frac{p}{1-p}.\end{aligned}$$

Therefore, for the case of $t \in \{k+1, \ldots, n-1\}$, in order to show that $\mathbb{P}(X_t = 1|\tau = t, F) = \mathbb{P}(X_t = 1|\tau = t, F_t) < p$ it suffices to show that $\mathbb{P}(\tau = t|X_t = 1, F_t) < \mathbb{P}(\tau = t|X_t = 0, F_t)$, which follows below:

[54] For $\boldsymbol{x} \in F^C$ no trial is drawn, which we can represent as $\mathbb{P}(\tau = 1|\boldsymbol{X} = \boldsymbol{x}) = 1$ (for example).

[55] The identity follows by the law of total probability, with the key observation that $\hat{P}_k(\boldsymbol{x}) = \sum_{t \in I_k(\boldsymbol{x})} x_t \cdot \frac{1}{|I_k(\boldsymbol{x})|} = \sum_{t=k+1}^n \mathbb{P}(X_t = 1|\tau = t, \boldsymbol{X} = \boldsymbol{x}, F)\mathbb{P}(\tau = t|\boldsymbol{X} = \boldsymbol{x}, F)$.

$$
\begin{aligned}
\mathbb{P}(\tau = t | X_t = 0, F_t) &= \sum_{\boldsymbol{x} \in F_t : x_t = 0} \mathbb{P}(\tau = t | X_t = 0, \boldsymbol{X} = \boldsymbol{x}, F_t) \mathbb{P}(\boldsymbol{X} = \boldsymbol{x} | X_t = 0, F_t) \\
&= \sum_{\boldsymbol{x} \in F_t : x_t = 0} \mathbb{P}(\tau = t | X_t = 0, \boldsymbol{X}_{-t} = \boldsymbol{x}_{-t}, F_t) \mathbb{P}(\boldsymbol{X}_{-t} = \boldsymbol{x}_{-t} | X_t = 0, F_t) \quad (4) \\
&> \sum_{\boldsymbol{x} \in F_t : x_t = 0} \mathbb{P}(\tau = t | X_t = 1, \boldsymbol{X}_{-t} = \boldsymbol{x}_{-t}, F_t) \mathbb{P}(\boldsymbol{X}_{-t} = \boldsymbol{x}_{-t} | X_t = 1, F_t) \quad (5) \\
&= \sum_{\boldsymbol{x} \in F_t : x_t = 1} \mathbb{P}(\tau = t | X_t = 1, \boldsymbol{X} = \boldsymbol{x}, F_t) \mathbb{P}(\boldsymbol{X} = \boldsymbol{x} | X_t = 1, F_t) \\
&= \mathbb{P}(\tau = t | X_t = 1, F_t)
\end{aligned}
$$

where in (4), given $\boldsymbol{x}$, we define $\boldsymbol{x}_{-t} := (x_1, \ldots, x_{t-1}, x_{t+1}, \ldots, x_n)$. To obtain the inequality in (5) we observe that: (i) $\mathbb{P}(\boldsymbol{X}_{-t} = \boldsymbol{x}_{-t} | X_t = 0, F_t) = \mathbb{P}(\boldsymbol{X}_{-t} = \boldsymbol{x}_{-t} | X_t = 1, F_t)$ because $\boldsymbol{X}$ is a sequence of i.i.d. Bernoulli trials, and (ii) $\mathbb{P}(\tau = t | X_t = 1, \boldsymbol{X}_{-t} = \boldsymbol{x}_{-t}, F_t) < \mathbb{P}(\tau = t | X_t = 0, \boldsymbol{X}_{-t} = \boldsymbol{x}_{-t}, F_t)$ because τ is drawn at random (uniformly) from the set $I_k(\boldsymbol{x})$, which contains at least one more element (trial $t+1$) if $x_t = 1$ rather than $x_t = 0$.

For the case of $t = n$ we follow the above steps until (5), at which point an equality now emerges as $X_n = 1$ no longer yields an additional trial from which to draw, because trial n is terminal. This implies that $\mathbb{P}(\tau = n | X_n = 1, F_n) = \mathbb{P}(\tau = n | X_n = 0, F_n)$.

Taking these two facts together: (i) $\mathbb{P}(X_t = 1 | \tau = t, F) < p$, for $k+1 \leq t < n$, and (ii) $\mathbb{P}(X_n = 1 | \tau = n, F) = p$, it immediately follows that $\mathbb{P}(X_\tau = 1 | F) < p$.[56]

■

A.2 Formula for the expected proportion (special case of $k = 1$)

The following lemma shows that the expected proportion $\hat{P}_1(\boldsymbol{X})$, conditional on a known number of successes $N_1(\boldsymbol{X}) = n_1$, satisfies the sampling-without-replacement formula, which for any given trial is less than the probability of success $\mathbb{P}(X_i | N_1(\boldsymbol{X}) = n_1) = \frac{n_1}{n}$.

Lemma 1 *Let $n > 1$. Then*

$$
E\left[\hat{P}_1(\boldsymbol{X}) \;\middle|\; I_1(\boldsymbol{X}) \neq \emptyset,\ N_1(\boldsymbol{X}) = n_1 \right] = \frac{n_1 - 1}{n - 1} \quad (6)
$$

for $0 \leq n_1 \leq n$.

[56] Note that the proof does not require that the Bernoulli trials be identically distributed. Instead, we could allow the probability distribution to vary, with $\mathbb{P}(X_i = 1) = p^i$ for $i = 1, \ldots, n$, in which case our result would be that $\mathbb{P}(X_\tau = 1 | F) < E[p_\tau | F]$.

Proof: As in the proof of Theorem 1, let τ be drawn at random from $I_1(\boldsymbol{X})$, which is non-empty when $N_1(\boldsymbol{X}) = n_1 \geq 2$ (the result is trivial when $n_1 = 1$). In order to ease notation we let probability $\mathbb{P}(\cdot)$ represent the conditional probability $\mathbb{P}(\cdot|N_1(\boldsymbol{X}) = n_1)$, which is defined over the subsets of $\{\boldsymbol{x} \in \{0,1\}^n : N_1(\boldsymbol{x}) = n_1\}$.

$$
\begin{aligned}
E[\hat{P}_1(\boldsymbol{X})|N_1(\boldsymbol{X}) = n_1, I_1(\boldsymbol{X}) \neq \emptyset] &= \mathbb{P}(X_\tau = 1) && (7)\\
&= \mathbb{P}(X_\tau = 1|\tau < n)\mathbb{P}(\tau < n) + \mathbb{P}(X_n = 1|\tau = n)\mathbb{P}(\tau = n)\\
&= \sum_{t=2}^{n-1} \mathbb{P}(X_t = 1|\tau = t)\frac{1}{n-1} + \mathbb{P}(X_n = 1|\tau = n)\frac{1}{n-1} && (8)\\
&= \frac{n-1}{n-2}\left(\frac{n_1}{n} - \frac{1}{n-1}\right)\frac{n-2}{n-1} + \frac{n_1}{n}\frac{1}{n-1} && (9)\\
&= \frac{n_1 - 1}{n-1}
\end{aligned}
$$

In (7), equality follows by an argument analogous to that provided in the proof of Theorem 1. In (8), equality follows from the fact that $\mathbb{P}(\tau = t) = 1/(n-1)$ for all $t \in \{2, 3, \ldots, n\}$.[57] In (9), equality follows from using an application of Bayes rule to derive $\mathbb{P}(X_t = 1|\tau = t)$, which satifies:

$$
\mathbb{P}(X_t = 1|\tau = t) = \begin{cases} \frac{n-1}{n-2}\left(\frac{n_1}{n} - \frac{1}{n-1}\right) & \text{for } t = 2, \ldots, n-1 \\ \frac{n_1}{n} & \text{for } t = n \end{cases} \tag{10}
$$

In particular,

$$
\begin{aligned}
\mathbb{P}(X_t = 1|\tau = t) &= \frac{\mathbb{P}(\tau = t|X_{t-1} = 1, X_t = 1)\mathbb{P}(X_{t-1} = 1|X_t = 1)\mathbb{P}(X_t = 1)}{\mathbb{P}(\tau = t)}\\
&= \mathbb{P}(\tau = t|X_{t-1} = 1, X_t = 1)\frac{n_1(n_1-1)}{n} && (11)
\end{aligned}
$$

where for all t, $\mathbb{P}(X_{t-1} = 1|X_t = 1) = (n_1 - 1)/(n-1)$, which is the likelihood that relates to sampling-without-replacement. For $t < n$, $\mathbb{P}(\tau = t|X_{t-1} = 1, X_t = 1)$, which is the likelihood that

[57] Note that $\mathbb{P}(\tau = t) = \sum_{\boldsymbol{x}:N_1(\boldsymbol{x})=n_1} \mathbb{P}(\tau = t|\boldsymbol{X} = \boldsymbol{x})\mathbb{P}(\boldsymbol{X} = \boldsymbol{x}) = \sum_{\boldsymbol{x}:N_1(\boldsymbol{x})=n_1, x_{t-1}=1} \frac{1}{n_1 - x_n}\frac{1}{\binom{n}{n_1}} = \frac{1}{\binom{n}{n_1}}\left[\binom{n-2}{n_1-1}\frac{1}{n_1} + \binom{n-2}{n_1-2}\frac{1}{n_1-1}\right] = \frac{1}{n-1}$.

relates to the arrangement of successes and failures, satisfies:

$$\begin{aligned}\mathbb{P}(\tau = t | X_{t-1} = 1, X_t = 1) &= E\left[\frac{1}{M} \;\middle|\; X_{t-1} = 1, X_t = 1 \right] \\ &= \sum_{x \in \{0,1\}} E\left[\frac{1}{M} \;\middle|\; X_{t-1} = 1, X_t = 1, X_n = x \right] \mathbb{P}(X_n = x | X_{t-1} = 1, X_t = 1) \\ &= \frac{1}{n_1} \frac{n_0}{n-2} + \frac{1}{n_1 - 1} \frac{n_1 - 2}{n-2} \\ &= \frac{1}{n-2} \left(\frac{n_0}{n_1} + \frac{n_1 - 2}{n_1 - 1} \right)\end{aligned}$$

where $M := |I_1(\boldsymbol{X})|$, i.e. $M = n_1 - X_n$. In the case that $t = n$, clearly $\mathbb{P}(\tau = n | X_{n-1} = 1, X_n = 1) = \frac{1}{n_1 - 1}$.

■

Formulae for expected value of the proportion

The conditional expectation in Lemma 1 can be combined with $\mathbb{P}(N_1(\boldsymbol{X}) = n_1 | I_1(\boldsymbol{X}) \neq \emptyset)$ to express the expected proportion in terms of just n and p.[58]

Theorem 2 *Let $n > 2$ and $0 < p < 1$. Then*

$$E\left[\hat{P}_1(\boldsymbol{X}) \;\middle|\; I_1(\boldsymbol{X}) \neq \emptyset \right] = \frac{\left[p - \frac{1-(1-p)^n}{n} \right] \frac{n}{n-1}}{1 - (1-p)^{n-1}} < p \tag{12}$$

Proof: We first observe that in light of Lemma 1, Equation 12 can be written as follows:

$$\begin{aligned}E\left[\hat{P}_1(\boldsymbol{X}) \;\middle|\; I_1(\boldsymbol{X}) \neq \emptyset \right] &= E\left[E\left[\hat{P}_1(\boldsymbol{X}) \;\middle|\; I_1(\boldsymbol{X}) \neq \emptyset,\ N_1(\boldsymbol{X}) = n_1 \right] \right] \\ &= E\left[\frac{N_1(\boldsymbol{x}) - 1}{n-1} \middle| I_1(\boldsymbol{X}) \neq \emptyset \right]\end{aligned}$$

[58] In a comment written about this paper, Rinott and Bar-Hillel (2015) provide an alternative proof for this theorem.

The expected value can then be computed using the binomial distribution, which yields:

$$
\begin{aligned}
E\left[\frac{N_1(\boldsymbol{x})-1}{n-1}\middle| I_1(\boldsymbol{X})\neq\emptyset\right] &= C\sum_{n_1=1}^{n} p^{n_1}(1-p)^{n-n_1}\left[\binom{n}{n_1}-U(n,n_1)\right]\cdot\frac{n_1-1}{n-1} \\
&= \frac{\sum_{n_1=2}^{n}\binom{n}{n_1}p^{n_1}(1-p)^{n-n_1}\frac{n_1-1}{n-1}}{1-(1-p)^n-p(1-p)^{n-1}} \\
&= \frac{\frac{1}{n-1}\left[\left(np-np(1-p)^{n-1}\right)-\left(1-(1-p)^n-np(1-p)^{n-1}\right)\right]}{1-(1-p)^n-p(1-p)^{n-1}} \\
&= \frac{\left[p-\frac{1-(1-p)^n}{n}\right]\frac{n}{n-1}}{1-(1-p)^{n-1}}
\end{aligned}
$$

where, as in the discussion below Equation 20 in Web Appendix E.2, $U(n,n_1)$ is the number of sequences with n_1 successes for which the proportion is undefined, and C is the constant that normalizes the total probability to 1. The second line follows because $U_1(n,n_1)=0$ for $n_1>1$, $U_1(n,0)=U_1(n,1)=1$, and $C=1/[1-(1-p)^n-p(1-p)^{n-1}]$.

Finally, by letting $q:=1-p$ it is straightforward to show that the bias in $\hat{P}_1(\boldsymbol{X})$ is negative:

$$
\begin{aligned}
E\left[\,\hat{P}_1(\boldsymbol{X})-p\;\middle|\;I_1(\boldsymbol{X})\neq\emptyset\,\right] &= \frac{\left[p-\frac{1-q^n}{n}\right]\frac{n}{n-1}}{1-q^{n-1}}-p \\
&= \frac{(n-1)(q^{n-1}-q^n)-(q-q^n)}{(n-1)(1-q^{n-1})} \\
&< 0
\end{aligned}
$$

The inequality follows from $f(x)=q^x$ being strictly decreasing and convex, which implies that $q-q^n>(n-1)(q^{n-1}-q^n)$.

■

A.3 The expected difference in proportions

Let D_k be the difference in the probability of success when comparing trials that immediately follow k consecutive successes with trials that immediately follow k consecutive failures. That is, $D_k := \mathbb{P}(X_t=1 \mid \prod_{j=t-k}^{t-1} X_j=1)-\mathbb{P}(X_t=1 \mid \prod_{j=t-k}^{t-1}(1-X_j)=1)$. An estimator of D_k that is used in the hot hand fallacy literature (see Section 3) is $\hat{D}_k(\boldsymbol{x}) := \hat{P}_k(\boldsymbol{x})-[1-\hat{Q}_k(\boldsymbol{X})]$, where $\hat{Q}_k(\boldsymbol{X})$ is the proportion of failures on the subset of trials that immediately follow k consecutive failures, $J_k(\boldsymbol{X}) := \{j : \prod_{i=j-k}^{j-1}(1-X_i)=1\} \subseteq \{k+1,\ldots,n\}$.

A.3.1 Proof of the bias in the difference

We extend the proof of Theorem 1 to show that $\hat{D}_k(\boldsymbol{X})$ is a biased estimator of D_k. Recall that $I_k(\boldsymbol{X})$ is the subset of trials that immediately follow k consecutive successes, i.e. $I_k(\boldsymbol{X}) := \{i : \prod_{j=i-k}^{i-1} X_j = 1\} \subseteq \{k+1, \ldots, n\}$. Analogously, let $J_k(\boldsymbol{X})$ be the subset of trials that immediately follow k consecutive failures, i.e. $J_k(\boldsymbol{X}) := \{j : \prod_{i=j-k}^{j-1}(1 - X_i) = 1\} \subseteq \{k+1, \ldots, n\}$.

Theorem 3 *Let $\boldsymbol{X} = \{X_i\}_{i=1}^n$, $n \geq 3$, be a sequence of independent Bernoulli trials, each with probability of success $0 < p < 1$. Let $\hat{P}_k(\boldsymbol{X})$ be the proportion of successes on the subset of trials $I_k(\boldsymbol{X})$ that immediately follow k consecutive successes, and $\hat{Q}_k(\boldsymbol{X})$ be the proportion of failures on the subset of trials $J_k(\boldsymbol{X})$ that immediately follow k consecutive failures. $\hat{D}_k(\boldsymbol{x}) := \hat{P}_k(\boldsymbol{x}) - [1 - \hat{Q}_k(\boldsymbol{X})]$ is a biased estimator of $D_k := \mathbb{P}(X_t = 1 \mid \prod_{j=t-k}^{t-1} X_j = 1) - \mathbb{P}(X_t = 1 \mid \prod_{j=t-k}^{t-1}(1 - X_j) = 1) \equiv 0$ for all k such that $1 \leq k < n/2$. In particular,*

$$E[\ \hat{D}_k(\boldsymbol{X}) \mid I_k(\boldsymbol{X}) \neq \emptyset, J_k(\boldsymbol{X}) \neq \emptyset\ ] < 0 \tag{13}$$

Proof: Following the notation used in the proof of Theorem 1, let $F := \{\boldsymbol{x} \in \{0,1\}^n : I_k(\boldsymbol{x}) \neq \emptyset\}$ and $G := \{\boldsymbol{x} \in \{0,1\}^n : J_k(\boldsymbol{x}) \neq \emptyset\}$. We will show the following:

$$\begin{aligned}
E[\hat{D}_k(\boldsymbol{x})|F,G] &= E[\hat{P}_k(\boldsymbol{X})|F,G] + E[\hat{Q}_k(\boldsymbol{X})|F,G] - 1 \\
&= \mathbb{P}(X_\tau = 1|F,G) + \mathbb{P}(X_\sigma = 0|F,G) - 1 &(14)\\
&< p + (1-p) - 1 &(15)\\
&= 0 &(16)
\end{aligned}$$

where in (14), as in the proof of Theorem 1, τ is a random draw from $I_k(\boldsymbol{x})$ and σ is an analogous random draw from $J_k(\boldsymbol{x})$. In particular, we will demonstrate that the inequality in (15) holds by showing that $\mathbb{P}(X_\tau = 1|F,G) < p$, which, by symmetry, implies that $\mathbb{P}(X_\sigma = 0|F,G) < 1 - p$.

To show that $\mathbb{P}(X_\tau = 1|F,G) < p$ we use an approach similar to that presented in the proof of Theorem 1. In particular, note that $\mathbb{P}(X_\tau = 1|F,G) = \sum_{t=k+1}^n \mathbb{P}(X_t = 1|\tau = t, F, G)\mathbb{P}(\tau = t|F,G)$, and $\mathbb{P}(\tau = t|F,G) > 0$ for $t \in \{k+1, \ldots, n\}$. In what follows, we demonstrate that $\mathbb{P}(X_t = 1|\tau = t, F, G) < p$ when $t < n$, and that $\mathbb{P}(X_t = 1|\tau = n, F, G) = p$, which, taken together, guarantee that $\mathbb{P}(X_\tau = 1|F,G) < p$.

First we observe that $\mathbb{P}(X_t = 1|\tau = t, F, G) = \mathbb{P}(X_t = 1|\tau = t, F_t, G)$, where $F_t := \{\boldsymbol{x} \in \{0,1\}^n :$

$\prod_{i=t-k}^{t-1} x_i = 1\}$. Bayes Rule then yields:

$$\begin{aligned}\frac{\mathbb{P}(X_t = 1|\tau = t, F_t, G)}{\mathbb{P}(X_t = 0|\tau = t, F_t, G)} &= \frac{\mathbb{P}\left(\tau = t, G\,|X_t = 1, F_t\right)}{\mathbb{P}\left(\tau = t, G\,|X_t = 0, F_t\right)} \frac{\mathbb{P}(X_t = 1|F_t)}{\mathbb{P}(X_t = 0|F_t)} \\ &= \frac{\mathbb{P}\left(\tau = t, G\,|X_t = 1, F_t\right)}{\mathbb{P}\left(\tau = t, G\,|X_t = 0, F_t\right)} \frac{p}{1-p}.\end{aligned}$$

Therefore, for the case of $t \in \{k+1, \ldots, n-1\}$, in order to show that $\mathbb{P}(X_t = 1|\tau = t, F, G) = \mathbb{P}(X_t = 1|\tau = t, F_t, G) < p$ it suffices to show that $\mathbb{P}(\tau = t, G|X_t = 1, F_t) < \mathbb{P}(\tau = t, G|X_t = 0, F_t)$, which follows below:

$$\begin{aligned}\mathbb{P}(\tau = t, G|X_t = 0, F_t) &= \sum_{\substack{\boldsymbol{x} \in F_t \cap G: \\ x_t = 0}} \mathbb{P}(\tau = t, \boldsymbol{X} = \boldsymbol{x}|X_t = 0, F_t) \\ &= \sum_{\substack{\boldsymbol{x} \in F_t \cap G: \\ x_t = 0 \\ (1, \boldsymbol{x}_{-t}) \in F_t \cap G}} \mathbb{P}(\tau = t, \boldsymbol{X} = \boldsymbol{x}|X_t = 0, F_t) \qquad (17) \\ &\quad + \sum_{\substack{\boldsymbol{x} \in F_t \cap G: \\ x_t = 0 \\ (1, \boldsymbol{x}_{-t}) \notin F_t \cap G}} \mathbb{P}(\tau = t, \boldsymbol{X} = \boldsymbol{x}|X_t = 0, F_t) \\ &\geq \sum_{\substack{\boldsymbol{x} \in F_t \cap G: \\ x_t = 0 \\ (1, \boldsymbol{x}_{-t}) \in F_t \cap G}} \mathbb{P}(\tau = t, \boldsymbol{X} = \boldsymbol{x}|X_t = 0, F_t) \\ &= \sum_{\substack{\boldsymbol{x} \in F_t \cap G: \\ x_t = 0 \\ (1, \boldsymbol{x}_{-t}) \in F_t \cap G}} \mathbb{P}(\tau = t|\boldsymbol{X} = \boldsymbol{x}, X_t = 0, F_t) \mathbb{P}(\boldsymbol{X} = \boldsymbol{x}|X_t = 0, F_t) \\ &= \sum_{\substack{\boldsymbol{x} \in F_t \cap G: \\ x_t = 0 \\ (1, \boldsymbol{x}_{-t}) \in F_t \cap G}} \mathbb{P}(\tau = t|X_t = 0, \boldsymbol{X}_{-t} = \boldsymbol{x}_{-t}, F_t) \mathbb{P}(\boldsymbol{X}_{-t} = \boldsymbol{x}_{-t}|X_t = 0, F_t) \\ &> \sum_{\substack{\boldsymbol{x} \in F_t \cap G: \\ x_t = 0 \\ (1, \boldsymbol{x}_{-t}) \in F_t \cap G}} \mathbb{P}(\tau = t|X_t = 1, \boldsymbol{X}_{-t} = \boldsymbol{x}_{-t}, F_t) \mathbb{P}(\boldsymbol{X}_{-t} = \boldsymbol{x}_{-t}|X_t = 1, F_t) \\ &\qquad (18) \\ &= \sum_{\substack{\boldsymbol{x} \in F_t \cap G: \\ x_t = 1 \\ (1, \boldsymbol{x}_{-t}) \in F_t \cap G}} \mathbb{P}(\tau = t|\boldsymbol{X} = \boldsymbol{x}, X_t = 1, F_t) \mathbb{P}(\boldsymbol{X} = \boldsymbol{x}|X_t = 1, F_t) \\ &= \sum_{\substack{\boldsymbol{x} \in F_t \cap G: \\ x_t = 1}} \mathbb{P}(\tau = t, \boldsymbol{X} = \boldsymbol{x}|X_t = 1, F_t) \\ &= \mathbb{P}(\tau = t, G|X_t = 1, F_t)\end{aligned}$$

where in (17), given $\boldsymbol{x}$, we define $\boldsymbol{x}_{-t} := (x_1, \ldots, x_{t-1}, x_{t+1}, \ldots, x_n)$, and $(b, \boldsymbol{x}_{-t}) := (x_1, \ldots, x_{t-1}, b, x_{t+1}, \ldots, x_n)$.[59] The inequality in (18) follows for the same reason as the inequality in line (5) of Theorem 1. In particular, $\mathbb{P}(\boldsymbol{X}_{-t} = \boldsymbol{x}_{-t} | X_t = 0, F_t) = \mathbb{P}(\boldsymbol{X}_{-t} = \boldsymbol{x}_{-t} | X_t = 1, F_t)$ because $\boldsymbol{X}$ is a sequence of i.i.d. Bernoulli trials, and $\mathbb{P}(\tau = t | X_t = 1, \boldsymbol{X}_{-t} = \boldsymbol{x}_{-t}, F_t) < \mathbb{P}(\tau = t | X_t = 0, \boldsymbol{X}_{-t} = \boldsymbol{x}_{-t}, F_t)$ because τ is drawn at random (uniformly) from the set $I_k(\boldsymbol{x})$, which contains at least one more element (trial $t+1$) if $x_t = 1$ rather than $x_t = 0$.

For the case of $t = n$ we follow the above steps until (18), at which point an equality now emerges, as $X_n = 1$ no longer yields an additional trial from which to draw, because trial n is terminal. This implies that $\mathbb{P}(\tau = n | X_n = 1, F_n, G) = \mathbb{P}(\tau = n | X_n = 0, F_n, G)$.

Taking these two facts together: (i) $\mathbb{P}(X_t = 1 | \tau = t, F, G) < p$, for $k+1 \leq t < n$, and (ii) $\mathbb{P}(X_n = 1 | \tau = n, F, G) = p$, it immediately follows that $\mathbb{P}(X_\tau = 1 | F, G) < p$.

■

A.3.2 Formula for the expected difference in proportions (special case of $k = 1$)

In the case of $k = 1$ the expected difference in proportions admits a simple representation that is independent of p.

Theorem 4 *Let* $\hat{D}_1(\boldsymbol{X}) := \hat{P}_1(\boldsymbol{X}) - (1 - \hat{Q}_1(\boldsymbol{X}))$. *If* $n > 2$ *and* $0 < p < 1$ *then*

$$E\left[\hat{D}_1(\boldsymbol{X}) \;\middle|\; I_1(\boldsymbol{X}) \neq \emptyset, J_1(\boldsymbol{X}) \neq \emptyset \right] = -\frac{1}{n-1}$$

Proof: The method of proof is to first show that if $n > 2$ and $1 \leq n_1 \leq n-1$ then:

$$E\left[\hat{D}_1(\boldsymbol{X}) \;\middle|\; N_1(\boldsymbol{X}) = n_1, I_1(\boldsymbol{X}) \neq \emptyset, J_1(\boldsymbol{X}) \neq \emptyset \right] = -\frac{1}{n-1}$$

which leaves us just one step from the desired result.

First, consider the case that $1 < n_1 < n-1$. In this case $\hat{D}_1(\boldsymbol{x}) := \hat{P}_1(\boldsymbol{x}) - (1 - \hat{Q}_1(\boldsymbol{x}))$ is defined for all sequences. Therefore, by the linearity of the expectation, and applying Lemma 1 to $\hat{Q}_1(\boldsymbol{X})$ (by symmetry), we have:

$$\begin{aligned} E[\hat{D}_1(\boldsymbol{X}) | N_1(\boldsymbol{X}) = n_1] =& E[\hat{P}_1(\boldsymbol{X}) | N_1(\boldsymbol{X}) = n_1] - E(1 - \hat{Q}_1(\boldsymbol{X}) | N_1(\boldsymbol{X}) = n_1] \\ =& \frac{n_1 - 1}{n-1} - \left(1 - \frac{n_0 - 1}{n-1}\right) \\ =& -\frac{1}{n-1} \end{aligned}$$

[59] Note that the second sum will have no terms for $t \geq n - k$.

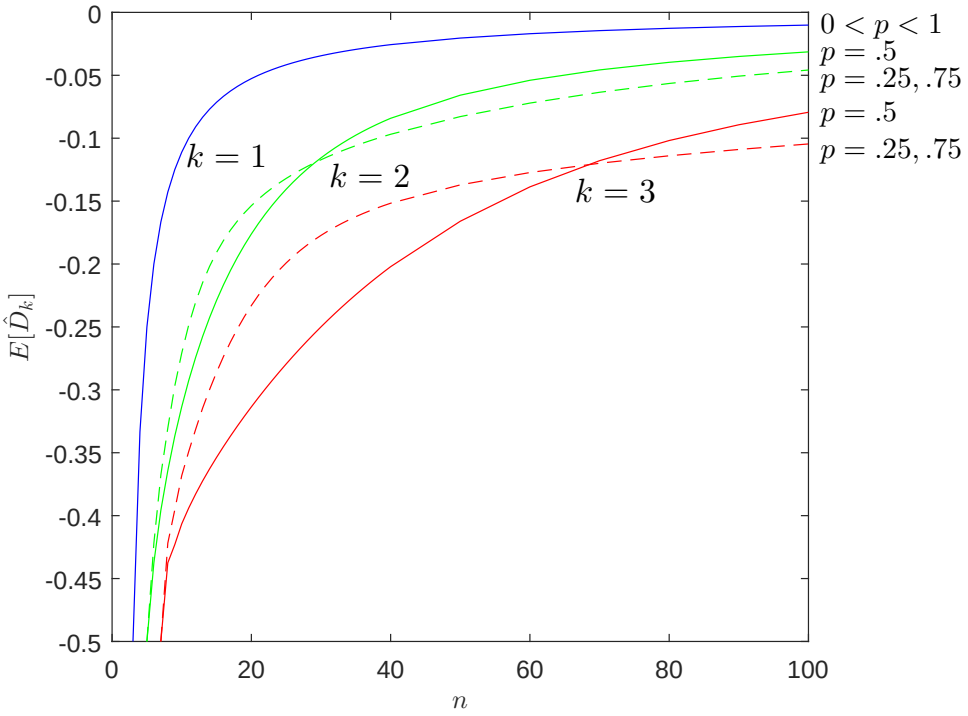

Figure 4: *For the expected difference in the proportion of successes, as a function of n, three values of k, and various probabilities of success p, using the formulas of Web Appendix E.3,(Theorem 6 combined Equation 22.*

If $n_1 = 1$ then $\hat{D}_1$ is defined for all sequences that do not have a 1 in the final position; there are $n-1$ such sequences. The sequence with the 1 in the first position yields $\hat{D}_1 = 0$, while the other $n-2$ sequences yield $\hat{D}_1 = -1/(n-2)$. Therefore, $E\left[\, \hat{D}_1(\boldsymbol{X}) \;\middle|\; N_1(\boldsymbol{X}) = 1, I_1(\boldsymbol{X}) \neq \emptyset, J_1(\boldsymbol{X}) \neq \emptyset \,\right] = -1/(n-1)$.

Now consider the case of $n_1 = n-1$. The argument for this case is analogous, with $\hat{D}_1$ undefined for the sequence with the zero in the last position, equal to 0 for the sequence with the zero in the first position, and equal to $-1/(n-2)$ for all other sequences.

Finally, that the conditional expectation is independent of $N_1(\boldsymbol{x})$ implies that $E[D_1(\boldsymbol{X}) \mid I_1(\boldsymbol{X}) \neq \emptyset, J_1(\boldsymbol{X}) \neq \emptyset\,]$ is independent of p, yielding the result.

■

A.3.3 Quantifying the bias for the difference

Figure 4 contains a plot of $E[\, \hat{D}_k(\boldsymbol{X}) \mid I_k(\boldsymbol{X}) \neq \emptyset, J_k(\boldsymbol{X}) \neq \emptyset\,]$ as a function of the number of trials n, and for $k = 1, 2$, and 3. Because the bias is dependent on p when $k > 1$, the difference is plotted for various values of p. These expected differences are obtained by combining Theorem 4

with the results in Web Appendix E. The magnitude of the bias is simply the absolute value of the expected difference. As with the bias in the proportion (see Figure 1), the bias in the difference is substantial even when n is relatively large.

B Appendix: Size of the bias when the DGP is hot hand/streak shooting

In Section 3.3 the correction to GVT's estimate of the hot hand effect (and test statistic) is based on the magnitude of the bias under the assumption that the shooter has a fixed probability of success (Bernoulli process). However, if the underlying data generating process (DGP) instead represents hot hand or streak shooting, then the size of the bias changes. While many DGPs can produce hot hand shooting, arguably the most natural are those discussed in Gilovich et al. (1985), as they reflect lay conceptions of the hot hand and streak shooting. While GVT take no particular stance on which lay definition is most appropriate, they do identify hot hand and streak shooting with: (1) "non-stationarity" (the zone, flow, in the groove, in rhythm), and (2) "positive association" (success breeds success). We label (1) as a *regime shift* model, and interpret it as capturing the idea that a player's probability of success may increase due to some factor that is unrelated to previous outcomes, so unobservable to the econometrician. This can be modeled naturally as a hidden markov chain over the player's (hidden) ability state. We label (2) as a *positive feedback* model, because it can be interpreted as capturing the idea that positive feedback from a player's previous shot outcomes can affect his/her subsequent probability of success. This can be modeled naturally as an autoregressive process, or equivalently as a markov chain over shot outcomes.[60]

In Figure 5 we plot the bias in the estimator of the difference in probability of success when on a hit streak rather than miss streak, $\hat{D}_3$, for three alternative DGPs, each of which admits the Bernoulli process as a special case.[61] The first panel corresponds to the "regime shift" DGP in which the difference in the probability of success between the "hot" state and the "normal" state is given by d (where $d = 0$ represents Bernoulli shooting),[62] the second panel corresponds to the "positive feedback" DGP in which hitting (missing) 3 shots in a row increases (decreases) the probability of success by $d/2$, and the third panel corresponds to the "positive feedback (for hits)" DGP in which positive feedback operates for hits only, making the probability of success increase by d whenever

[60] A positive feedback model need not be stationary.

[61] Each point is the output of a simulation with 10,000 repetitions of 100 trials from the DGP.

[62] In particular, let Q be the hidden markov chain over the "normal" state (n) and the "hot" state (h), where the probability of success in the normal state is given by p_n, and the probability of success in the hot state is given by p_h, with the shift in probability of success given by $d := p_h - p_n$

$$Q := \begin{pmatrix} q_{nn} & q_{nh} \\ q_{hn} & q_{hh} \end{pmatrix}$$

Where q_{nn} represents the probability of staying in the "normal" state, and q_{nh} represents the probability of transitioning from the "normal" to the "hot" state, etc. Letting $\pi = (\pi_n, \pi_h)$ be the stationary distribution, we find that the magnitude of the bias is not very sensitive to variation in the stationary distribution and transition probabilities within a plausible range (i.e. $\pi_h \in [.05, .2]$ and $q_{hh} \in (.8, .98)$), while it varies greatly with the difference in probabilities d and the overall expected shooting percentage $p = p_n + \pi_h d$. In the graph, for each d and p (FG%), we average across values for the stationary distribution (π_h) and transition probability (q_{hh}).

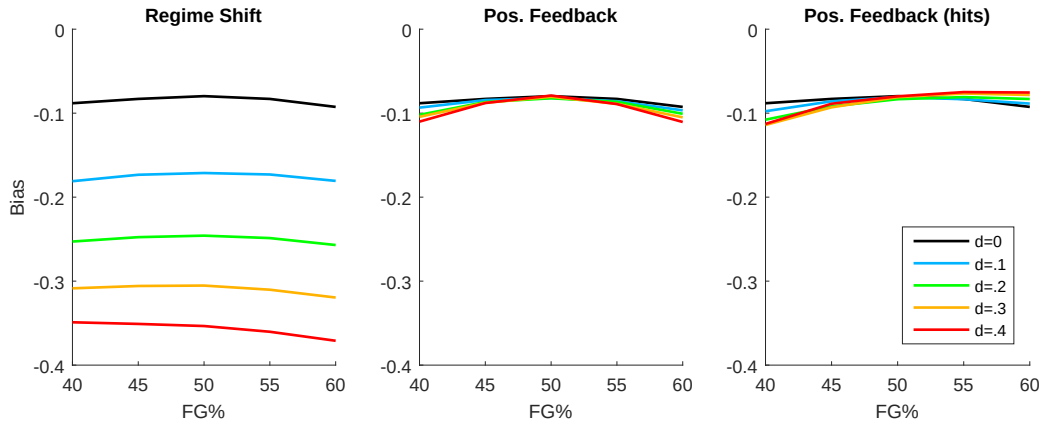

Figure 5: *The bias for three types of hot hand and streak shooting data generating processes (DGPs), where FG% is the expected overall field goal percentage from the DGP, and d represents the change in the player's underlying probability of success. When $d = 0$ each model reduces to a Bernoulli process. Therefore the black line represents the bias in a Bernoulli proccess ($n = 100$ trials, $k = 3$).*

3 hits in a row occurs. Within each panel of the figure, the bias, which is the expected difference between $\hat{D}_3$, the estimator of the shift in the probability of success, and d, the true shift in the probability of success, is depicted as a function of the expected overall shooting percentages (from 40 percent to 60 percent), for four true shifts in the underlying probability ($d \in \{.1, .2, .3, .4\}$).[63]

Observe that when the true DGP is a player with a hot hand, the bias is typically more severe, or far more severe, than the bias associated with a Bernoulli DGP. In particular, the bias in the "regime shift" model is particularly severe, which arises from two sources: (1) the bias discussed in Section 2, and (2) an attenuation bias, due to measurement error, as hitting 3 shots in a row is an imperfect proxy for the "hot state."[64] The bias in the positive feedback DGP is uniformly below the bias for a Bernoulli shooter. For the DGP in which positive feedback operates only for hits, the bias is stronger than that of Bernoulli shooters for expected shooting percentages below 50 percent (as in GVTs data), and slightly less strong for shooting percentage above 50 percent. As the true DGP is likely some combination of a regime shift and positive feedback, it is reasonable to conclude that the empirical approach in Section 3.3 should be expected to (greatly) understate the true magnitude of any underlying hot hand.

The intuition for why the introduction of regime shift elements increases the strength of the bias so considerably is that if a player's probability of success is not driven merely by feedback from previous shots, but also by other time-varying player (and environment) specific factors, then the act of hitting consecutive shots will serve as only a noisy proxy of the hot state, resulting in measurement error, and an attenuation bias in the estimate. This type of measurement error is

[63] Results are similar if the DGP instead has negative feedback, i.e. $d \in \{-.1, -.2, -.3, -.4\}$.

[64] In practice observers may have more information than the econometrician (e.g. shooting mechanics, perceived confidence, or lack thereof, etc.), so may be subject to less measurement error.

similar to what Stone (2012) uncovered in the relationship between autocorrelation in outcomes and autocorrelation in ability when considering a DGP that contains autocorrelation in ability.

C Appendix: Additional analyses, and details for Section 3

C.1 An alternative (pooled) analysis of shooting data

An alternative approach to testing for streak shooting across players is to pool all shots from the "3 hits" and "3 misses" categories (discarding the rest), then use a linear probability model to estimate the effect of a shot falling in the "3-hits" category. If the implementation of GVT's design met the goal of placing each player in a position in which his or her probability of success is .5, then this approach would be analogous to re-weighting the under-weighted coin flips in Table 1 of Section 1. With 2515 shots, the bias is minimal and the estimate in this case is +17 percentage points ($p < .01$, $S.E. = 3.7$). Because GVT's design goal is difficult to implement in practice, this approach will introduce an upward bias, due to aggregation, if the probability of success varies across players. Adding fixed effects in a regression will control for this aggregation bias, but strengthens the selection bias related to streaks.[65] As a result, a bias correction is necessary. In this case, the estimated effect is +13.9 percentage points ($p < .01$, $S.E. = 5.8$), which has larger standard errors because the heteroscedasticity under the assumption of different player probabilities necessitates the use of robust variants (in this case, Bell and McCaffrey standard errors, see Imbens and Kolesar [2016]). The magnitude of the estimated effect has a different interpretation than the one given for the estimate of the average difference across players; it should be thought of as the hot hand effect for the average shot rather than the average player. This interpretation arises because pooling shots across players generates an unbalanced panel, which results in the estimate placing greater weight on the players that have taken more shots. As such, in the extreme it is even possible that the majority of players exhibit a tendency to have fewer streaks than expected by chance, yet, because they have generated relatively few observations, their data becomes diluted by many observations from a single streak shooter.

C.2 Details on the hypothesis testing with the permutation test procedure

Let $\boldsymbol{X} \in \{0,1\}^n$ be a sequence of shot outcomes from some player, i. The null hypothesis is that the shots are i.i.d. with $\mathbb{P}(X_t = 1) = p^i$. This implies that conditional on the number of hits, $N_1(\boldsymbol{X}) = n_1$, each rearrangement is equally likely. Considering only sequences for which both $\hat{P}^i(hit|k\ hits)$ and $\hat{P}^i(hit|k\ misses)$ are defined, the hot hand hypothesis predicts that the difference $\hat{P}^i(hit|k\ hits) - \hat{P}^i(hit|k\ misses)$ will be significantly larger than what one would expect by chance. Let $\hat{D}_k(\boldsymbol{X})$ be this difference for sequence $\boldsymbol{X}$. For an observed sequence $\boldsymbol{x}$, with $N_1(\boldsymbol{x}) = n_1$ hits, to test the null hypothesis at the α level, one simply checks if $\hat{D}_k(\boldsymbol{x}) \geq c_{\alpha,n_1}$, where the critical

[65] In this panel regression framework, the bias from introducing fixed-effects is an example of an incidental parameter problem of Neyman and Scott (1948), and is essentially equivalent to that discussed in Nerlove (1971) and Nickell (1981), and itself is closely related to the bias in estimates of autocorrelation in time series mentioned in the Introduction.

value c_{α,n_1} is defined as the smallest c such that $\mathbb{P}(D_k(\boldsymbol{X}) \geq c \mid H_0,\ N_1(\boldsymbol{X}) = n_1) \leq \alpha$, and the distribution $\mathbb{P}(D_k(\boldsymbol{X}) \geq c \mid H_0,\ N_1(\boldsymbol{X}) = n_1)$ is generated using the enumeration provided in Theorem 6 of Web Appendix E.3. For the quantity $\mathbb{P}(D_k(\boldsymbol{X}) \geq c \mid H_0,\ N_1(\boldsymbol{X}) = n_1)$ it may be the case that for some c^*, it is strictly greater than α for $c \leq c^*$, and equal to 0 for $c > c^*$. In this case, for any sequence with $N_1(\boldsymbol{X}) = n_1$ one cannot reject H_0 at an α level of significance. From the ex ante perspective, a test of the hot hand at the α level of significance consists of a family of such critical values $\{c_{\alpha,n_1}\}$. It follows immediately that $\mathbb{P}(\text{reject}|H_0) \leq \alpha$ because $\mathbb{P}(\text{reject}|H_0) = \sum_{n_1=1}^{n} \mathbb{P}(D_k(\boldsymbol{X}) \geq c_{\alpha,n_1}|H_0,\ N_1(\boldsymbol{X}) = n_1)\mathbb{P}(N_1(\boldsymbol{X}) = n_1|H_0) \leq \alpha$. Lastly, for any arbitrary test statistic $T(\boldsymbol{X})$, the fact that the distribution of $\boldsymbol{X}$ is *exchangeable* conditional on $N_1(\boldsymbol{X}) = n_1$ means that $\mathbb{P}(T(\boldsymbol{X}) \geq c \mid H_0,\ N_1(\boldsymbol{X}) = n_1)$ can be approximated to appropriate precision with Monte-Carlo permutations of the sequence $\boldsymbol{x}$.

D Web Appendix: Bias Mechanism and Quantitative Comparison to Sampling without Replacement, for Section 2.1

Suppose that the researcher were to know the overall proportion of successes in the sequance, $\hat{p} = n_1/n$. Consider the following two ways of learning that trial t immediately follows k consecutive successes: (1) a trial τ_N, drawn uniformly at random from $\{k+1,\ldots,n\}$, ends up being equal to trial t, and preceded by k consecutive successes, or (2) a randomly drawn trial τ_I from $I_k(\boldsymbol{x}) \subseteq \{k+1,\ldots,n\}$ is trial t. In each case, the prior probability of success is $\mathbb{P}(x_t = 1) = n_1/n$, which can be represented as an odds ratio in favor of $x_t = 1$ (relative to $x_t = 0$) equal to $\mathbb{P}(x_t = 1)/\mathbb{P}(x_t = 0) = n_1/n_0$.

In the first case the probability is given by $\mathbb{P}(\tau_N = t) = 1/(n-k)$ for all $t \in \{k+1,\ldots,n\}$, and is independent of $\boldsymbol{x}$. Upon finding out that $\tau_N = t$ one then learns that $\prod_{t-k}^{t-1} x_i = 1$. As a result, the posterior odds yield a sampling-without-replacement formula, via Bayes rule:

$$
\begin{aligned}
\frac{\mathbb{P}(x_t = 1|\tau_N = t)}{\mathbb{P}(x_t = 0|\tau_N = t)} &= \frac{\mathbb{P}(x_t = 1, \prod_{t-k}^{t-1} x_i = 1|\tau_N = t)}{\mathbb{P}(x_t = 0, \prod_{t-k}^{t-1} x_i = 1|\tau_N = t)} \\
&= \frac{\mathbb{P}(\tau_N = t|x_t = 1, \prod_{t-k}^{t-1} x_i = 1)}{\mathbb{P}(\tau_N = t|x_t = 0, \prod_{t-k}^{t-1} x_i = 1)} \frac{\mathbb{P}(\prod_{t-k}^{t-1} x_i = 1|x_t = 1)}{\mathbb{P}(\prod_{t-k}^{t-1} x_i = 1|x_t = 0)} \frac{\mathbb{P}(x_t = 1)}{\mathbb{P}(x_t = 0)} \\
&= \frac{\mathbb{P}(\prod_{t-k}^{t-1} x_i = 1|x_t = 1)}{\mathbb{P}(\prod_{t-k}^{t-1} x_i = 1|x_t = 0)} \frac{\mathbb{P}(x_t = 1)}{\mathbb{P}(x_t = 0)} \\
&= \frac{\frac{n_1-1}{n-1} \times \cdots \times \frac{n_1-k}{n-k}}{\frac{n_1}{n-1} \times \cdots \times \frac{n_1-k+1}{n-k}} \frac{n_1}{n_0} \\
&= \frac{n_1 - k}{n_1} \frac{n_1}{n_0} \\
&= \frac{n_1 - k}{n_0}
\end{aligned}
$$

Observe that the prior odds in favor of success are attenuated by the likelihood ratio $\frac{n_1-k}{n_1}$ of producing k consecutive successes given either hypothetical state of the world, $x_t = 1$ or $x_t = 0$, respectively.

In the second case, the probability that $\tau_I = t$ is drawn from $I_k(\boldsymbol{x})$ is completely determined by $M := |I_k(\boldsymbol{x})|$, and equal to $1/M$. Upon learning that $\tau_I = t$ one can infer the following three things: (1) $I_k(\boldsymbol{x}) \neq \emptyset$, i.e. $M \geq 1$, which is informative if $n_1 \leq (k-1)(n-n_1)+k$, (2) t is a member of $I_k(\boldsymbol{x})$, and (3) $\prod_{t-k}^{t-1} x_i = 1$, as in sampling-without-replacement. As a result, the posterior odds

can be determined via Bayes Rule in the following way:

$$
\begin{aligned}
&\frac{\mathbb{P}(x_t = 1|\tau_I = t)}{\mathbb{P}(x_t = 0|\tau_I = t)} \\
&\qquad = \frac{\mathbb{P}(x_t = 1, \prod_{t-k}^{t-1} x_i = 1, M \geq 1|\tau_I = t)}{\mathbb{P}(x_t = 0, \prod_{t-k}^{t-1} x_i = 1, M \geq 1|\tau_I = t)} \\
&\qquad = \frac{\mathbb{P}(\tau_I = t|x_t = 1, \prod_{t-k}^{t-1} x_i = 1)}{\mathbb{P}(\tau_I = t|x_t = 0, \prod_{t-k}^{t-1} x_i = 1)} \frac{\mathbb{P}(M \geq 1|x_t = 1, \prod_{t-k}^{t-1} x_i = 1)}{\mathbb{P}(M \geq 1|x_t = 0, \prod_{t-k}^{t-1} x_i = 1)} \frac{\mathbb{P}(\prod_{t-k}^{t-1} x_i = 1|x_t = 1)}{\mathbb{P}(\prod_{t-k}^{t-1} x_i = 1|x_t = 0)} \frac{\mathbb{P}(x_t = 1)}{\mathbb{P}(x_t = 0)} \\
&\qquad = \frac{E\left[\ \frac{1}{M} \ \middle|\ \prod_{t-k}^{t-1} x_i = 1, x_t = 1\right]}{E\left[\ \frac{1}{M} \ \middle|\ \prod_{t-k}^{t-1} x_i = 1, x_t = 0\right]} \frac{\mathbb{P}(\prod_{t-k}^{t-1} x_i = 1|x_t = 1)}{\mathbb{P}(\prod_{t-k}^{t-1} x_i = 1|x_t = 0)} \frac{\mathbb{P}(x_t = 1)}{\mathbb{P}(x_t = 0)} \\
&\qquad = \frac{E\left[\ \frac{1}{M} \ \middle|\ \prod_{t-k}^{t-1} x_i = 1, x_t = 1\right]}{E\left[\ \frac{1}{M} \ \middle|\ \prod_{t-k}^{t-1} x_i = 1, x_t = 0\right]} \frac{n_1 - k}{n_1} \frac{n_1}{n_0}
\end{aligned}
\tag{19}
$$

The conditional argument is omitted in the first term of the first equality because the event $M \geq 1$ is implied by the event $\prod_{t-k}^{t-1} x_i = 1$. The formula in the final line indicates that the posterior odds in favor of $x_t = 1$ can be thought of as arising from the following two-stage Bayesian updating procedure: (1) updating with the sampling-without-replacement factor $\frac{n_1-k}{n_1}$, and (2) updating with the factor that relates to how the successes and failures are arranged in the sequence: $\frac{E[\ \frac{1}{M} \ \mid \ \prod_{t-k}^{t-1} x_i=1, x_t=1]}{E[\ \frac{1}{M} \ \mid \ \prod_{t-k}^{t-1} x_i=1, x_t=0]}$. This second factor reveals that if the expected probability of choosing any given trial (including t) is larger in the state of the world in which $x_t = 0$, rather than $x_t = 1$, then the posterior odds will decrease beyond what sampling-without-replacement alone would suggest. This is natural to expect in the case that $t < n$, as $x_t = 0$ makes it impossible for the $\min\{n-t, k\}$ trials that follow trial t to be members of $I_k(\boldsymbol{x})$, whereas $x_t = 1$ assures that trial $t+1$ is a member of $I_k(\boldsymbol{x})$, and does not exclude the $\min\{n-t, k\} - 1$ trials that follow it from also being in $I_k(\boldsymbol{x})$. In the proof of Theorem 1 it was shown that this factor is strictly less than 1 in the general case, when $t < n$ and $\hat{p} = n_1/n$ is unknown. For the case in which $\hat{p} = n_1/n$ is known, and $k = 1$, see the discussion that precedes Equation 3 in Section 2.1.[66]

A quantitative comparison with sampling-without-replacement

For the general case, in which $\hat{p} = n_1/n$ is unknown, juxtaposing the bias with sampling-without-replacement puts the magnitude of the bias into context. Let the probability of success be given by $p = \mathbb{P}(X_t = 1)$. In Figure 6, the expected empirical probability that a randomly drawn trial

[66] When $k > 1$, the computation is combinatorial in nature, and utilizes the dimension reduction arguments that are provided in Web Appendix E to produce the formula presented in Lemma 2. This formula is in turn employed in Theorem 5, with $1/M = 1/f_{1k}$ or $1/M = 1/(f_{1k}-1)$ depending on the final k trials in the sequence. The calculation of $E\left[\ \frac{1}{M} \ \middle|\ N_1 = n_1, M \geq 1\right]$ requires the distribution of M. It appears that all known formulations of this distribution do not admit a simple representation of the expectation (Balakrishnan and Koutras 2011, p.188).

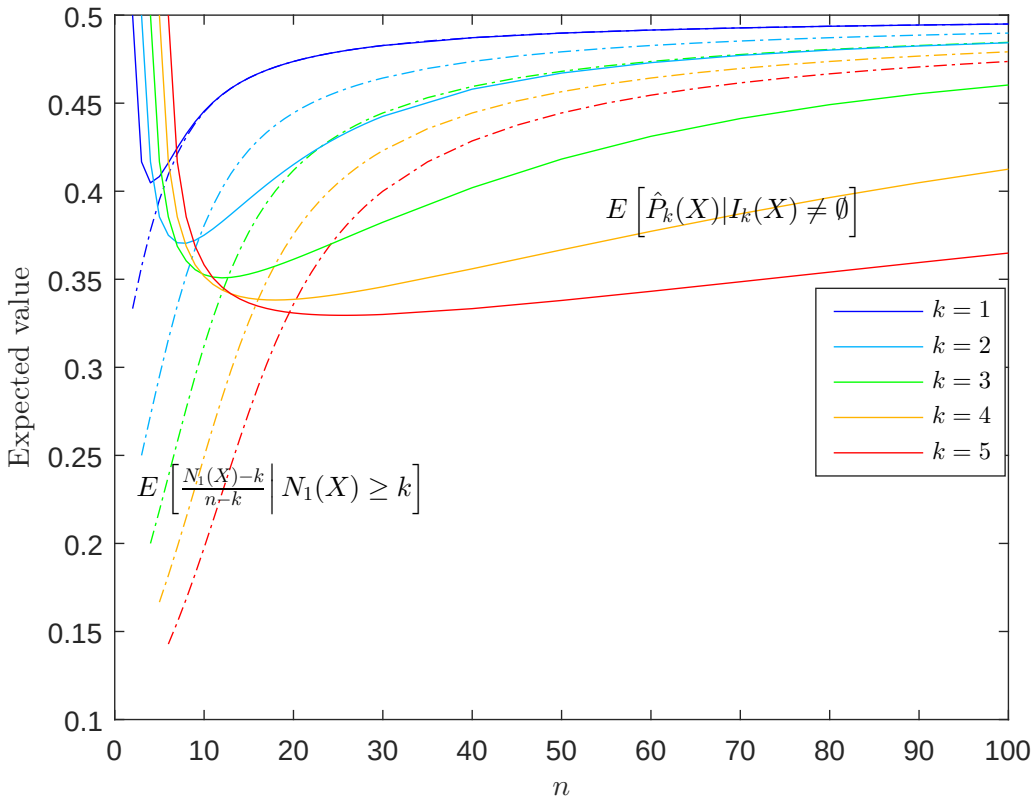

Figure 6: *The dotted lines correspond to the bias from sampling-without-replacement. It is the expected probability of a success, given that k successes are first removed from the sequence (assuming $p = .5$). The solid lines correspond to the expected proportion from Figure 1.*

in $I_k(\boldsymbol{X})$ is a success, which is the expected proportion, $E[\hat{P}_k(\boldsymbol{X}) \mid I_k(\boldsymbol{X}) \neq \emptyset]$, is plotted along with the expected value of the probability that a randomly drawn trial $t \in \{1, \ldots, n\} \backslash T_k$ is a success, given that the k success trials $T_k \subseteq \{1, \ldots, n\}$ have already been drawn from the sequence (sampling-without-replacement), $E\left[\frac{N_1(\boldsymbol{X})-k}{n-k}\middle| N_1(\boldsymbol{X}) \geq k\right]$. The plot is generated using the combinatorial results discussed in Section 2.2. Note that in the case of $k = 1$, the bias is identical to sampling-without-replacement, as shown in Equation 3.[67] Observe that for $k > 1$, and n not too small, the bias in the expected proportion is considerably larger than the corresponding bias from

[67] This appears to contradict what was shown in Section 2.1, i.e. that the bias in the procedure used to select the subset of trials $I_{1k}(\boldsymbol{x})$, is *stronger* than sampling-without-replacement for $t < n$, whereas it is non-existent (thus weaker) for $t = n$. This disparity is due to the second updating factor, which relates to the arrangement. It turns out that for $k = 1$, the determining aspect of the arrangement that influences this updating factor is whether or not the final trial is a success, as this determines the number of successes in the first $n-1$ trials, where $M = n_1 - x_n$. If one were to instead fix M rather than n_1, then sampling-without-replacement relative to the number of successes in the first $n-1$ trials would be an accurate description of the mechanism behind the bias, and it induces a negative dependence between any two trials within the first $n-1$ trials of the sequence. Therefore, it is sampling-without-replacement with respect to M that determines the bias when $k = 1$.

sampling-without-replacement. In the case of sampling-without-replacement, the selection criteria for sequences, $N_1(\boldsymbol{X}) \geq k$, can be thought of as a generalization of Berkson's paradox for binary data. In the case of the bias in the expected proportion, the sequence weight updating factor, analogous to the likelihood ratio in equation 19, is determined by the number of successes in the sequence, but not by their arrangement.[68]

[68] In particular, the sequence weight that corresponds to $1/M$, is $1/\binom{N_1}{k}$, i.e. the reciprocal of the number of ways to choose k successes from N_1 successes.

E Web Appendix: Explicit formulas for the expected proportion, and difference in proportions

We begin by sketching our approach to deriving a formula for the expected value of each of our statistical measures of interest. Let $T(\boldsymbol{x})$ be a statistic that is well-defined for sequences in $\boldsymbol{x} \in A \subset \{0,1\}^n$. Then, the law of total probability implies that:

$$E[T(\boldsymbol{X})|A] = \sum_{n_1=0}^{n} E[T(\boldsymbol{X}) \mid N_1(\boldsymbol{X}) = n_1, A)]\mathbb{P}(N_1(\boldsymbol{X}) = n_1|A)$$

with $\mathbb{P}(N_1(\boldsymbol{X}) = n_1|A)$ satisfying:

$$\mathbb{P}(N_1(\boldsymbol{X}) = n_1|A) = \left[\binom{n}{n_1} - U_T(n, n_1)\right] \times p^{n_1}(1-p)^{n-n_1} \times C$$

where $U_T(n_1, n)$ is the number of sequences with n_1 successes for which T is undefined, and C is the constant that normalizes the total probability to 1. More precisely, $U_T(n, n_1) := |\{\boldsymbol{x} \in \{0,1\}^n : N_1(\boldsymbol{x}) = n_1\} \cap A^C|$ and $C := 1/\left(1 - \sum_{n_1=0}^{n} U_T(n, n_1)p^{n_1}(1-p)^{n-n_1}\right)$. In the theorems of this appendix we derive $E[T(\boldsymbol{X}) \mid N_1(\boldsymbol{X}) = n_1, A]$ and $U_T(n, n_1)$ for the estimators of interest.

For the statistics considered below, this approach leads to tractable formulae, because the expected value is constant over a large set of sequences when conditioning on $N_1(\boldsymbol{X}) = n_1$. For the proportion $\hat{P}_k(\boldsymbol{x})$, in particular, the set of sequences over which it is constant is in turn determined by the joint distribution of the number of runs of different lengths. Therefore the distribution of the proportion, conditional on $N_1(\boldsymbol{X}) = n_1$, is determined by the the joint distribution of runs.

To illustrate this using a simple case, consider the proportion of successes on the trials that immediately follow k consecutive successes, when $k = 2$.

First, we observe that the proportion of successes on the trials that immediately follow two consecutive successes $\hat{P}_2(\boldsymbol{x})$ can be represented simply as the ratio of the number of streaks of at least three successes in n trials to the number of streaks of at least two success in the first $n-1$ trials. In particular, for a sequence $\boldsymbol{x} \in \{0,1\}^n$ of successes and failures, we have:

$$\begin{aligned}\hat{P}_2(\boldsymbol{x}) &:= \frac{\sum_{i \in I_2(\boldsymbol{x})} x_i}{|I_2(\boldsymbol{x})|} \\ &= \frac{\#_{111}(\boldsymbol{x})}{\#_{111}(\boldsymbol{x}) + \#_{110}(\boldsymbol{x})}\end{aligned}$$

where $I_2(\boldsymbol{x}) := \{i \in \{3, \ldots, n\} : x_{i-1}x_{i-2} = 1\}$, as in Section 2, and $\#_{111}(\boldsymbol{x})$ is defined as the

number of overlapping runs of three consecutive successes, with $\#_{110}(\boldsymbol{x})$ defined similarly.[69] We show that, given $N_1(\boldsymbol{x})$ successes in n trials, this ratio can be written in terms of the number of non-overlapping runs of ones of exactly length one, $R_{11}(\boldsymbol{x})$, and the number of non-overlapping runs of ones of length two or more, $S_{12}(\boldsymbol{x})$. This means that $\hat{P}_k(\boldsymbol{x})$ is constant for the set of sequences $\{\boldsymbol{x} \in \{0,1\}^n : N_1(\boldsymbol{x}) = n_1, R_{11}(\boldsymbol{x}) = r_{11}, S_{12}(\boldsymbol{x}) = s_{12}\}$. This allows us to considerably reduce the number of sequences under consideration. The expected value of $\hat{P}_2(\boldsymbol{x})$, conditional on $N_1(\boldsymbol{x}) = n_1$, is now fully determined by the joint distribution of $(R_{11}(\boldsymbol{x}), S_{12}(\boldsymbol{x}))$, conditional on $N_1(\boldsymbol{x}) = n_1$. This distribution is derived by way of combinatorial argument in Lemma 2 in Section E.2, where we determine $C_{1k} := |\{\boldsymbol{x} \in \{0,1\}^n : N_1(\boldsymbol{x}) = n_1, R_{11}(\boldsymbol{x}) = r_{11}, S_{12}(\boldsymbol{x}) = s_{12}\}|$. In Theorem 5 we derive the formula for the expected proportion, conditional on $N_1(\boldsymbol{x}) = n_1$. A similar approach is used to derive a formula for the expected difference in proportions.

E.1 Definitions

Given the sequence $\boldsymbol{x} = (x_1, \ldots, x_n) \in \{0,1\}^n$, we provide precise definitions of runs and streaks, and the variables corresponding to their frequency of occurrence within a sequence. We illustrate the definitions with an the example sequence $\boldsymbol{z} = (1,1,0,1,1,0,0,1,1,1)$, which has $n = 10$ trials, $N_1(\boldsymbol{z}) = n_1 = 7$ successes, and $N_0(\boldsymbol{z}) = n_0 = 3$ failures.

A run of 1s is a subsequence of consecutive 1s in $\boldsymbol{x}$ that is flanked on each side by a 0 or an endpoint.[70] A run of 0s is defined analogously. Let $R_{1j}(\boldsymbol{x}) = r_{1j}$ be the number of runs of 1s of exactly length j for $j = 1, \ldots, n_1$. Let $R_{0j}(\boldsymbol{x}) = r_{0j}$ be defined similarly. For example, in sequence $\boldsymbol{z}$, for runs of 1s we have $r_{11} = 0$, $r_{12} = 2$, $r_{13} = 1$ and $r_{1j} = 0$ for $j \geq 4$, and for runs of 0s we have $r_{01} = 1$, $r_{02} = 1$, and $r_{0j} = 0$ for $j \geq 3$. The number of runs of 1s of length j or more, $S_{1j}(\boldsymbol{x}) = s_{1k}$, is given by $S_{1j}(\boldsymbol{x}) := \sum_{i=j}^{n_1} R_{1j}(\boldsymbol{x})$ for $j = 1, \ldots, n_1$, with $S_{0j}(\boldsymbol{x}) = s_{0j}$ defined similarly. For example, in sequence $\boldsymbol{z}$, for runs of 1s we have $s_{11} = 3$, $s_{12} = 3$, $s_{13} = 1$ and $s_{1j} = 0$ for $j \geq 4$, and for runs of 0s we have $s_{01} = 2$, $s_{02} = 1$, and $s_{0j} = 0$ for $j \geq 3$. Note that Let $R_1(\boldsymbol{x}) = r_1$, be the number of runs of 1s, i.e. $R_1(\boldsymbol{x}) = S_{11}(\boldsymbol{x})$, and $R_0(\boldsymbol{x}) = r_0$ be the number of runs of 0s. Let $R(\boldsymbol{x}) = r$ be the total number of runs, i.e. $R(\boldsymbol{x}) := R_1(\boldsymbol{x}) + R_0(\boldsymbol{x})$. For example, in sequence $\boldsymbol{z}$, the total number of runs is $r = 5$. Further, let a streak of success of length k be an instance of k consecutive successes, and the number of such (overlapping) instances, $F_{1k}(\boldsymbol{x}) = f_{1k}$, be defined as $F_{1k}(\boldsymbol{x}) := \sum_{j=k}^{n_1}(j-k+1)R_{1j}(\boldsymbol{x})$, with $F_{0k}(\boldsymbol{x}) = f_{0k}$ defined analogously. For example, in sequence $\boldsymbol{z}$, for streaks of 1s we have $f_{11} = 7$, $f_{12} = 4$, $f_{13} = 1$, and $f_{1j} = 0$ for $j \geq 4$, and for streaks of 0s we have $f_{01} = 3$, $f_{02} = 1$, and $f_{0j} = 0$ for $j \geq 3$. Notice that $f_{1k} = |I_k(\boldsymbol{x})|$ if $\exists i > n-k$ with $x_i = 0$, and $f_{1k} = |I_k(\boldsymbol{x})| + 1$ otherwise. Also note that $n_1 = f_{11} = \sum_{j=1}^{n_1} j r_{1j}$ and $n_0 = f_{01} = \sum_{j=1}^{n_0} j r_{0j}$.

[69] Note that the pattern 110 cannot overlap with itself.

[70] More precisely, it is a subsequence with successive indices $j = i_1 + 1, i_1 + 2, \ldots, i_1 + k$, with $i_1 \geq 0$ and $i_1 + k \leq n$, in which $x_j = 1$ for all j, and: (1) either $i_1 = 0$, or if $i_1 > 0$ then $x_{i_1} = 0$, and (2) either $i_1 + k = n$, or if $i_1 + k < n$ then $i_1 + k + 1 = 0$

E.2 Expected Proportion ($k > 1$)

Let F be the set of sequences over which the proportion is well-defined, i.e. $F := \{\boldsymbol{x} \in \{0,1\}^n : I_k(\boldsymbol{x}) \neq \emptyset\}$. By the law of total probability it follows that:

$$E[\hat{P}_k(\boldsymbol{X})|F] = \sum_{n_1=k}^{n} E[\hat{P}_k(\boldsymbol{X}) \mid N_1(\boldsymbol{X}) = n_1, F]\ \mathbb{P}(N_1(\boldsymbol{X}) = n_1|F) \tag{20}$$

Therefore, it suffices to find formulas for $E[\hat{P}_k(\boldsymbol{X}) \mid N_1(\boldsymbol{X}) = n_1, F]$ and $\mathbb{P}(N_1(\boldsymbol{X}) = n_1|F)$.

Note that $\mathbb{P}(N_1(\boldsymbol{X}) = n_1|F)$ is determined by the number of sequences $U_{1k}(n, n_1)$, for which the difference is undefined:

$$\mathbb{P}(N_1(\boldsymbol{X}) = n_1|F) = \left[\binom{n}{n_1} - U_{1k}(n, n_1)\right] \times p^{n_1}(1-p)^{n-n_1} \times C$$

where C is the constant that normalizes the total probability to 1. More precisely, $U_{1k}(n, n_1) := |\{\boldsymbol{x} \in \{0,1\}^n : N_1(\boldsymbol{x}) = n_1\} \cap F^C|$ and $C := 1/\left(1 - \sum_{n_1=0}^{n} U_{1k}(n, n_1)p^{n_1}(1-p)^{n-n_1}\right)$. In Theorem 6 below we derive a formula for $E[\hat{P}_k(\boldsymbol{X}) \mid N_1(\boldsymbol{X}) = n_1, F]$, which uses Lemma 2 and $U_k(n, n_1)$. From these formulae the expected difference can also be computed.

To shorten some of the expressions in this section we will assume that our sample space of sequences are those in which the proportion of successes immediately after a streak of k successes, $\hat{P}_k(\boldsymbol{x})$, is well defined. We define $P_{1k} = \hat{P}_1(\boldsymbol{x})$ as the induced random variable, with support $\{p_{1k} \in [0,1] : p_{1k} = \hat{P}_k(\boldsymbol{x}) \text{ for } \boldsymbol{x} \in \{0,1\}^n,\ I_k(\boldsymbol{x}) \neq \emptyset\}$.

Our first step is to obtain an explicit formula for $E[P_{1k}|N_1 = n_1]$. That $E[P_{1k}|N_1 = n_1]$ was shown in Lemma 1 to be equal to $(n_1 - 1)/(n - 1)$ when $k = 1$ suggests the possibility that, in the spirit of sampling-without-replacement, the expression $(n_1 - k)/(n - k)$ represents the expected proportions for $k > 1$. That this formula does not hold in the case of $k > 1$ is easy to confirm by setting $k = 2$, $n_1 = 4$, and $n = 5$.[71] As in Section 2, it is not possible to determine $\hat{P}_k(\boldsymbol{x})$ directly by computing its value for each sequence, as the number of admissible sequences is typically too large.

We observe that the number of trials in the proportion satisfies $|I_k(\boldsymbol{x})| = F_{1k}(\boldsymbol{x}_{-n})$, i.e. it is equal to the frequency of length k 1-streaks in the sub-sequence that does not include the final term. Further we note that $F_{1k+1}(\boldsymbol{x})$ is the number of trials in the proportion that are themselves equal to 1. Therefore the proportion $\hat{P}_k(\boldsymbol{x})$ can be represented as

$$\hat{P}_k(\boldsymbol{x}) = \frac{F_{1k+1}(\boldsymbol{x})}{F_{1k}(\boldsymbol{x}_{-n})} \quad \text{if } F_{1k}(\boldsymbol{x}_{-n}) > 0 \tag{21}$$

[71] If $k = 2$ then for $n_1 = 4$ and $n = 5$, $E[P_{1k}|N_1 = n_1] = (0/1+1/1+1/2+2/2+2/3)/5 = 19/30 < 2/3 = (n_1-k)/(n-k)$ (see Section D for intuition).

where $\hat{P}_k(\boldsymbol{x})$ is otherwise undefined. Further, because $F_{1k}(\boldsymbol{x}_{-n}) = F_{1k}(\boldsymbol{x}) - \prod_{i=n-k+1}^{n} x_i$, it follows that

$$\hat{P}_k(\boldsymbol{x}) = \frac{F_{1k+1}(\boldsymbol{x})}{F_{1k}(\boldsymbol{x}) - \prod_{i=n-k+1}^{n} x_i} \quad \text{if } F_{1k}(\boldsymbol{x}) > \prod_{i=n-k+1}^{n} x_i$$

Because $F_{1k}(\boldsymbol{x}) := \sum_{j=k}^{n_1}(j-k+1)R_{1j}(\boldsymbol{x})$, we can, in principle, calculate $E[P_k(\boldsymbol{x},1)|N_1 = n_1]$ using the joint distribution $(R_{11},\ldots,R_{1n_1})$, which appears in a classic reference for non-parametric statistical theory (Gibbons and Chakraborti 2010, Theorem 3.3.2, p.87).[72]

Nevertheless, the calculation unfortunately does not appear to be computationally feasible for the sequence lengths of interest here. As a result, we make the key observation that for all sequences with $R_{1j}(\boldsymbol{x}) = r_{1j}$ for $j = 1,\ldots,k-1$ and $S_{1k}(\boldsymbol{x}) = s_{1k}$, the proportion $\hat{P}_k(\boldsymbol{x})$ is: (i) constant and equal to $(f_{1k} - s_{1k})/f_{1k}$ for those sequences that have a 0 in one of the final k positions, and (ii) constant and equal to $(f_{1k} - s_{1k})/(f_{1k} - 1)$ for those sequences that have a 1 in each of the final k positions. This is true because $f_{1k+1} = f_{1k} - s_{1k}$, and $f_{1k} = n_1 - \sum_{j=1}^{k-1} j r_{1j} - (k-1)s_{1k}$. Notice that for each case $\hat{P}_k(\boldsymbol{x}) = G(R_{11}(\boldsymbol{x}),\ldots,R_{1k-1}(\boldsymbol{x}),S_{1k}(\boldsymbol{x}))$ for some G. Therefore, by finding the joint distribution of $(R_{11},\ldots,R_{1k-1},S_{1k})$, conditional on N_1, it is possible to obtain $E[P_{1k}|N_1 = n_1]$. With $\binom{n}{n_1}$ sequences $\boldsymbol{x} \in \{0,1\}^n$ that satisfy $N_1(\boldsymbol{x}) = n_1$, the joint distribution of $(R_{11}(\boldsymbol{x}),\ldots R_{1k-1}(\boldsymbol{x}),S_{1k}(\boldsymbol{x}))$ is fully characterized by the number of distinguishable sequences $\boldsymbol{x}$ that satisfy $R_{11}(\boldsymbol{x}) = r_{11},\ldots R_{1k-1}(\boldsymbol{x}) = r_{1k-1}$, and $S_{1k}(\boldsymbol{x}) = s_{1k}$, which we obtain in the following lemma. In the proof of the lemma we provide a combinatorial argument that we later use repeatedly in the proof of Theorem 5.

Lemma 2 *Let $n \geq 1$, $n_1 \leq n$, $s_{1k} \geq 0$, and $r_{1j} \geq 0$. If*

$$C_{1k} := \#\{\boldsymbol{x} \in \{0,1\}^n : N_1(\boldsymbol{x}) = n_1, S_{1k}(\boldsymbol{x})) = s_{1k}, R_{11}(\boldsymbol{x}) = r_{11},\ldots,R_{1k-1}(\boldsymbol{x}) = r_{1k-1}\}$$

then

$$C_{1k} = \frac{r_1!}{s_{1k}!\prod_{j=1}^{k-1} r_{1j}!}\binom{n_0+1}{r_1}\binom{f_{1k}-1}{s_{1k}-1}$$

for $j = 1,\ldots k-1$, where $r_1 = \sum_{j=1}^{k-1} r_{1j} + s_{1k}$, $f_{1k} = n_1 - \sum_{j=1}^{k-1} j r_{1j} - (k-1)s_{1k}$, and $\binom{n}{k} = n!/k!(n-k)!$ if $n \geq k \geq 0$ and $\binom{n}{k} = 0$ otherwise (except for the special case $\binom{-1}{-1} = 1$).[73]

[72] The theorem in Gibbons and Chakraborti (2010) is not quite correct; the distribution presented in the theorem is for $(R_{11},\ldots,R_{1n_1})$ conditional only on N_1 (unconditional on R_1). For the distribution conditional on R_1 and N_1 it is straightforward to show that

$$\mathbb{P}(R_{11} = r_{11},\ldots,R_{1n_1} = r_{1n_1}|N_1 = n_1, R_1 = r_1) = \frac{r_1!}{\binom{n_1-1}{r_1-1}\prod_{j=1}^{n_1} r_{1j}!}$$

[73] Note that with this definition of $\binom{n}{k}$, we have $C_{1k} = 0$ if $r_1 > n_0 + 1$, or $\sum_{j=1}^{k-1} j r_{1j} + k s_{1k} > n_1$ (the latter occurs if

Proof:

Any sequence with $r_{11}, \ldots, r_{1k-1}$ runs of 1s of fixed length, and s_{1k} runs of 1s of length k or more can be constructed in three steps by: (1) selecting a distinguishable permutation of the $r_1 = \sum_{j=1}^{k-1} r_{1j} + s_{1k}$ cells that correspond to the r_1 runs, which can be done in $r_1!/(s_{1k}! \prod_{j=1}^{k-1} r_{1j})$ unique ways, as for each j, the $r_{1j}!$ permutations of the r_{1j} identical cells across their fixed positions do not generate distinguishable sequences (nor for the s_{1k} identical cells), (2) placing the r_1 1s into the available positions to the left or the right of a 0 among the n_0 0s; with n_0+1 available positions, there are $\binom{n_0+1}{r_1}$ ways to do this, (3) filling the "empty" run cells, by first filling the r_{1j} run cells of length j with exactly jr_{1j} 1s for $j < k$, and then by filling the s_{1k} indistinguishable (ordered) run cells of length k or more by: (a) adding exactly $k-1$ 1s to each cell, (b) with the remaining f_{1k} 1s (the number of 1s that succeed some streak of $k-1$ or more 1s), taking an ordered partition of these 1s into a separate set of s_{1k} cells, which can be performed in $\binom{f_{1k}-1}{s_{1k}-1}$ ways by inserting $s_{1k}-1$ dividers into the $f_{1k}-1$ available positions between 1s, and finally (c) adjoining each cell of the separate set of (nonempty and ordered) cells with its corresponding run cell (with exactly $k-1$ 1s), which guarantees that each s_{1k} cell has at least k 1s.

■

Below we state the main theorem, which provides the formula for the expected value of $\hat{P}_{1k}(\boldsymbol{x})$, conditional on the number of 1s:

Theorem 5 *Let n, n_1 and k satisfy $1 < k \leq n_1 \leq n$. Then*

$$E[P_{1k}|N_1 = n_1] = \frac{1}{\binom{n}{n_1} - U_{1k}} \sum_{\substack{r_{11},\ldots,r_{1k-1},s_{1k} \\ \sum_{j=1}^{k-1} jr_{1j} < n_1 - k \\ s_{1k} \geq 1}} C_{1k} \left[\frac{s_{1k}}{n_0+1} \binom{f_{1k}-s_{1k}}{f_{1k}-1} + \frac{n_0+1-s_{1k}}{n_0+1} \binom{f_{1k}-s_{1k}}{f_{1k}} \right]$$

where f_{1k} and C_{1k} are defined as in Lemma 2 and depend on $n_0, n_1, r_{11}, \ldots, r_{1k-1}$, and s_{1k},[74] *and U_{1k} is defined as the number of sequences in which $\hat{P}_{1k}(\boldsymbol{x})$ is undefined and satisfies*

$$U_{1k} = \sum_{r_1=1}^{\min\{n_1, n_0+1\}} \binom{n_0+1}{r_1} \sum_{\ell=0}^{\lfloor \frac{n_1-r_1}{k-1} \rfloor} (-1)^\ell \binom{r_1}{\ell} \binom{n_1-1-\ell(k-1)}{r_1-1}$$
$$+ \delta_{n_1 k} + \sum_{r_1=2}^{\min\{n_1-k+1, n_0+1\}} \binom{n_0}{r_1-1} \sum_{\ell=0}^{\lfloor \frac{n_1-k-r_1+1}{k-1} \rfloor} (-1)^\ell \binom{r_1-1}{\ell} \binom{n_1-k-1-\ell(k-1)}{r_1-2}$$

Proof:

$s_{1k} > \lfloor \frac{n_1 - \sum_j jr_{1j}}{k} \rfloor$, or $r_{1\ell} > \lfloor \frac{n_1 - \sum_{j \neq \ell} jr_{1j} - ks_{1k}}{\ell} \rfloor$ for some $\ell = 1, \ldots, k-1$, where $\lfloor \cdot \rfloor$ is the floor function). Further, because $r_1 > n_1$ implies the latter condition, it also implies $C_{1k} = 0$.

[74] Note that $\sum_{j=1}^{k-1} jr_{1j} < n_1 - k$ implies $f_{1k} > s_{1k} \geq 1$, which guarantees that $f_{1k} \geq 2$.

When considering all sequences $\boldsymbol{x} \in \{0,1\}^n$ that have n_1 1s, there are three possible cases for how $\hat{P}_{1k}(\boldsymbol{x})$ is determined by r_{1j} $j < k$ and s_{1k}: (1) $\hat{P}_{1k}(\boldsymbol{x})$ is not defined, which arises if: (i) $f_{1k} = 0$ or (ii) $f_{1k} = 1$ and $\sum_{i=n-k+1}^{n} x_i = k$, (2) $\hat{P}_{1k}(\boldsymbol{x})$ is equal to $(f_{1k} - s_{1k})/(f_{1k} - 1)$, which arises if $f_{1k} \geq 2$ and $\sum_{i=n-k+1}^{n} x_i = k$, or (3) $\hat{P}_{1k}(\boldsymbol{x})$ is equal to $(f_{1k} - s_{1k})/f_{1k}$, which arises if $f_{1k} \geq 1$ and $\sum_{i=n-k+1}^{n} x_i < k$.

In Case 1i, if $f_{1k} = 0$ then the number of terms, which we denote U_{1k}^1, satisfies:

$$
\begin{aligned}
U_{1k}^1 &:= \sum_{\substack{r_{11},\ldots,r_{1k-1},s_{1k} \\ \sum_{j=1}^{k-1} j r_{1j} = n_1 \\ s_{1k}=0}} C_{1k} \\
&= \sum_{\substack{r_{11},\ldots,r_{1k-1} \\ \sum_{j=1}^{k-1} j r_{1j} = n_1}} \frac{r_1!}{\prod_{j=1}^{k-1} r_{1j}!} \binom{n_0+1}{r_1} \\
&= \sum_{r_1=1}^{\min\{n_1,n_0+1\}} \binom{n_0+1}{r_1} \sum_{\substack{r_{11},\ldots,r_{1k-1} \\ \sum_{j=1}^{k-1} j r_{1j} = n_1 \\ \sum_{j=1}^{k-1} r_{1j} = r_1}} \frac{r_1!}{\prod_{j=1}^{k-1} r_{1j}!} \\
&= \sum_{r_1=1}^{\min\{n_1,n_0+1\}} \binom{n_0+1}{r_1} \sum_{\ell=0}^{\lfloor \frac{n_1-r_1}{k-1} \rfloor} (-1)^{\ell} \binom{r_1}{\ell} \binom{n_1-1-\ell(k-1)}{r_1-1}
\end{aligned}
$$

where the last line follows by first noting that the inner sum of the third line is the number of compositions (ordered partitions) of $n_1 - k$ into $r_1 - 1$ parts, which has generating function $(x + x^2 + \cdots x^{k-1})^{r_1}$ (Riordan 1958, p. 124). Therefore, the inner sum can be generated as the coefficient on x^{n_1} in the multinomial expansion of $(x + x^2 + \cdots x^{k-1})^{r_1}$. The inner sum of binomial coefficients in the fourth line corresponds to the coefficient on x^{n_1} in the binomial expansion of an equivalent representation of the generating function $x^{r_1}(1 - x^{k-1})^{r_1}/(1 - x)^{r_1} = (x + x^2 + \cdots x^{k-1})^{r_1}$. The coefficient in the binomial expansion must agree with the coefficient in the multinomial expansion.[75]

In Case 1ii, if $f_{1k} = 1$ and $\sum_{i=n-k+1}^{n} x_i = k$ (in which case $\hat{P}_{1k}(\boldsymbol{x})$ is also undefined) all sequences that satisfy this criteria can be constructed by first forming a distinguishable permutation of the $r_1 - 1$ runs of 1s that are not the final run of k 1s, which can be done in $r_1!/(\prod_{j=1}^{k-1} r_{1j}!)$ ways, then placing these runs to the left or the right of the available n_0 0s, not including the right end

[75] The binomial expansion is given by:

$$
x^{r_1} \frac{(1-x^{k-1})^{r_1}}{(1-x)^{r_1}} = x^{r_1} \left[\sum_{t_1=0}^{r_1} \binom{r_1}{t_1} (-1)^{t_1} x^{t_1(k-1)} \right] \cdot \left[\sum_{t_2=0}^{+\infty} \binom{r_1-1+t_2}{r_1-1} x^{t_2} \right]
$$

therefore the coefficient on x^{n_1} is $\sum (-1)^{t_1} \binom{r_1}{t_1} \binom{r_1-1+t_2}{r_1-1}$ where the sum is taken over all t_1, t_2 such that $r_1 + t_1(k-1) + t_2 = n_1$.

point, which can be done in $\binom{n_0}{r_1-1}$ ways given the n_0 positions. Summing over all possible runs, the number of terms U_{1k}^2 satisfies:

$$\begin{aligned}
U_{1k}^2 &:= \sum_{\substack{r_{11},\ldots,r_{1k-1},s_{1k} \\ \sum_{j=1}^{k-1} j r_{1j} = n_1 - k \\ s_{1k}=1}} \frac{s_{1k}}{n_0+1} C_{1k} \\
&= \sum_{\substack{r_{11},\ldots,r_{1k-1},s_{1k} \\ \sum_{j=1}^{k-1} j r_{1j} = n_1 - k \\ s_{1k}=1}} \frac{(r_1-1)!}{\prod_{j=1}^{k-1} r_{1j}!} \binom{n_0}{r_1-1} \\
&= \delta_{n_1 k} + \sum_{r_1=2}^{\min\{n_1-k+1,n_0+1\}} \binom{n_0}{r_1-1} \sum_{\substack{r_{11},\ldots,r_{1k-1} \\ \sum_{j=1}^{k-1} j r_{1j} = n_1 - k \\ \sum_{j=1}^{k-1} r_{1j} = r_1 - 1}} \frac{(r_1-1)!}{\prod_{j=1}^{k-1} r_{1j}!} \\
&= \delta_{n_1 k} + \sum_{r_1=2}^{\min\{n_1-k+1,n_0+1\}} \binom{n_0}{r_1-1} \sum_{\ell=0}^{\lfloor \frac{n_1-k-r_1+1}{k-1} \rfloor} (-1)^{\ell} \binom{r_1-1}{\ell} \binom{n_1-k-1-\ell(k-1)}{r_1-2}
\end{aligned}$$

and we assume that $\sum_{j=m}^{n} a_j = 0$ if $m > n$. The Kronecker delta in the third line appears because when $s_{1k} = 1$ and $\sum_{j=1}^{k-1} j r_{1j} = n_1 - k$ there is only one sequence for which $\hat{P}_{1k}(\boldsymbol{x})$ is undefined. The last line follows because the inner sum of the third line can be generated as the coefficient on x^{n_1-k} in the multinomial expansion of $(x + x^2 + \cdots x^{k-1})^{r_1-1}$, which, as when determining U_{1k}^1, corresponds to the coefficient on the binomial expansion. Taking case 1i and 2ii together, the total number of sequences in which $\hat{P}_{1k}(\boldsymbol{x})$ is undefined is equal to $U_{1k} = U_{1k}^1 + U_{1k}^2$.

In Case 2, in which $\hat{P}_{1k}(\boldsymbol{x})$ is defined with $\sum_{i=n-k+1}^{n} x_i = k$ and $f_{1k} \geq 2$, it must be the case that $\sum_{j=1}^{k-1} j r_{1j} < n_1 - k$, and that all sequences that satisfy this criteria can be constructed in three steps that are analogous to those used in Lemma 2: (1) selecting a distinguishable permutation of the $r_1 - 1$ remaining runs, (2) placing the $r_1 - 1$ 1s into the n_0 available positions to the left or the right of a 0, and (3) filling the "empty" run cells. For a given $(r_{11}, \ldots, r_{1k-1}, s_{1k})$ the total number of sequences that satisfy this criteria is:

$$\frac{(r_1-1)!}{(s_{1k}-1)! \prod_{j=1}^{k-1} r_{1j}!} \binom{n_0}{r_1-1} \binom{f_{1k}-1}{s_{1k}-1} = \frac{s_{1k}}{n_0+1} C_{1k}$$

In Case 3, in which $\hat{P}_{1k}(\boldsymbol{x})$ is defined with $\sum_{i=n-k+1}^{n} x_i < k$ and $f_{1k} \geq 1$, it must be the case that $\sum_{j=1}^{k-1} j r_{1j} \leq n_1 - k$, as all sequences that satisfy this criteria can again be constructed in the same three steps as before. We consider each of two subcases separately: sequences that terminate in a 1 (i.e. a run of 1s of length less than k) and sequences that terminate in a 0 (i.e. a run of 0s). When considering those sequence that terminate in a 1, for a given $(r_{11}, \ldots, r_{1k-1}, s_{1k})$ the total

number of sequences satisfying this criteria is:

$$\left(\frac{r_1!}{s_{1k}!\prod_{j=1}^{k-1} r_{1j}!} - \frac{(r_1-1)!}{(s_{1k}-1)!\prod_{j=1}^{k-1} r_{1j}!}\right)\binom{n_0}{r_1-1}\binom{f_{1k}-1}{s_{1k}-1} = \frac{r_1-s_{1k}}{n_0+1}C_{1k}$$

where $(r_1-1)!/((s_{1k}-1)!\prod_{j=1}^{k-1} r_{1j}!)$ is the number of sequences that terminate in a run of 1s of length k or more. When considering those sequences that terminate in a 0, for a given $(r_{11},\ldots,r_{1k-1},s_{1k})$ the total number of sequences satisfying this criteria is:

$$\frac{r_1!}{s_{1k}!\prod_{j=1}^{k-1} r_{1j}!}\binom{n_0}{r_1}\binom{f_{1k}-1}{s_{1k}-1} = \frac{n_0+1-r_1}{n_0+1}C_{1k}$$

Therefore, the expectation of $\hat{P}_{1k}(\boldsymbol{x})$ across all sequences for which it is defined satisfies:

$$\begin{aligned}E[\hat{P}_{1k}|N_1=n_1]\left[\binom{n}{n_1}-U_{1k}\right] = &\sum_{\substack{r_{11},\ldots,r_{1k-1},s_{1k}\\ \sum_{j=1}^{k-1} jr_{1j}<n_1-k\\ s_{1k}\geq 1}} C_{1k}\frac{s_{1k}}{n_0+1}\frac{f_{1k}-s_{1k}}{f_{1k}-1}\\ &+\sum_{\substack{r_{11},\ldots,r_{1k-1},s_{1k}\\ \sum_{j=1}^{k-1} jr_{1j}\leq n_1-k\\ s_{1k}\geq 1}} C_{1k}\frac{r_1-s_{1k}}{n_0+1}\frac{f_{1k}-s_{1k}}{f_{1k}}\\ &+\sum_{\substack{r_{11},\ldots,r_{1k-1},s_{1k}\\ \sum_{j=1}^{k-1} jr_{1j}\leq n_1-k\\ s_{1k}\geq 1}} C_{1k}\frac{n_0+1-r_1}{n_0+1}\frac{f_{1k}-s_{1k}}{f_{1k}}\end{aligned}$$

This expression can then be reduced to the formula in the theorem by first combining the final two terms, then summing over all runs that satisfy $\sum_{j=1}^{k-1} jr_{1j} < n_1-k$, and finally by combining with the first term (because $f_{1k}-s_{1k}=0$ if $\sum_{j=1}^{k-1} jr_{1j}=n_1-k$).[76]

■

[76] While the first term has a closed form representation: $\sum C_{1k}\frac{s_{1k}}{n_0+1}\frac{f_{1k}-s_{1k}}{f_{1k}-1} = \binom{n_1-1}{k}/\binom{n-1}{k}$, this does not appear to be the case for the other terms. Further, even if the other terms were to have a closed form, $E[P_{1k}|N_1=n_1]$ cannot possibly have one, as the term U_{1k} does not allow it.

E.3 Expected Difference in Proportions

The difference in the rate of success between trials that immediately follow a streak of successes and trials that immediately follow a streak of failures is given by:

$$\hat{D}_k(\boldsymbol{x}) := \hat{P}_k(\boldsymbol{x}) - (1 - \hat{Q}_k(\boldsymbol{x}))$$

where $\hat{P}_k(\boldsymbol{x}) := \frac{\sum_{i \in I_k(\boldsymbol{x})} x_i}{|I_k(\boldsymbol{x})|}$, with $I_k(\boldsymbol{x}) := \{i \in \{k+1, \ldots, n\} : \prod_{j-k}^{i-1} x_j = 1\}$ and $\hat{Q}_k(\boldsymbol{x}) := \frac{\sum_{i \in J_k(\boldsymbol{x})} 1 - x_i}{|J_k(\boldsymbol{x})|}$, with $J_k(\boldsymbol{x}) := \{j \in \{k+1, \ldots, n\} : \prod_{i-k}^{j-1}(1 - x_i) = 1\}$.

The exact formula for the expected difference can be obtained with an approach similar to the one used in the previous section. Let H be the set of sequences over which the difference is well-defined, i.e. $H := \{\boldsymbol{x} \in \{0,1\}^n : I_k(\boldsymbol{x}) \neq \emptyset, J_k(\boldsymbol{x}) \neq \emptyset\}$. By the law of total probability it follows that:

$$E[\hat{D}_k(\boldsymbol{X})|H] = \sum_{n_1=k}^{n} E[\hat{D}_k(\boldsymbol{X}) \mid N_1(\boldsymbol{X}) = n_1, H]\mathbb{P}(N_1(\boldsymbol{X})|H) \tag{22}$$

Therefore, it suffices to find formulas for $E[\hat{D}_k(\boldsymbol{X}) \mid N_1(\boldsymbol{X}) = n_1, H]$ and $\mathbb{P}(N_1(\boldsymbol{X}) = n_1|H)$.

Note that $\mathbb{P}(N_1(\boldsymbol{X}) = n_1|H)$ is determined by the number of sequences $U_k(n, n_1)$ for which the difference is undefined:

$$\mathbb{P}(N_1(\boldsymbol{X}) = n_1|H) = \left[\binom{n}{n_1} - U_k(n, n_1)\right] \times p^{n_1}(1-p)^{n-n_1} \times C$$

where C is the constant that normalizes the total probability to 1. More precisely, $U_k(n, n_1) := |\{\boldsymbol{x} \in \{0,1\}^n : N_1(\boldsymbol{x}) = n_1\} \cap H^C|$ and $C := 1/\left(1 - \sum_{n_1=0}^{n} U_k(n, n_1)p^{n_1}(1-p)^{n-n_1}\right)$. In Theorem 6 below we derive a formula for both $E[\hat{D}_k(\boldsymbol{X}) \mid N_1(\boldsymbol{X}) = n_1, H]$ and $U_k(n, n_1)$, from which the expected difference can be computed.

First, let D_k be the random variable defined by the distribution of $\boldsymbol{X}$, conditional on $\hat{D}(\boldsymbol{X})$ being well-defined. There are three categories of sequences for which $\hat{D}(\boldsymbol{X})$ is well-defined: (1) a sequence that ends in a run of 0s of length k or more, with $f_{0k} \geq 2$ and $f_{1k} \geq 1$, and the difference equal to $D_k^1 = (f_{1k} - s_{1k})/f_{1k} - (s_{0k} - 1)/(f_{0k} - 1)$, (2) a sequence that ends in a run of 1s of length k or more, with $f_{0k} \geq 1$ and $f_{1k} \geq 2$, and the difference equal to $D_k^2 := (f_{1k} - s_{1k})/(f_{1k} - 1) - s_{0k}/f_{0k}$, and (3) a sequence that ends in a run of 0s of length $k-1$, or less, or a run of 1s of length $k-1$, or less, with $f_{0k} \geq 1$ and $f_{1k} \geq 1$, and the difference equal to $D_k^3 := (f_{1k} - s_{1k})/f_{1k} - s_{0k}/f_{0k}$. For all other sequences the difference is undefined.

Theorem 6 *Let n, n_1, n_0 and k satisfy $n_0 + n_1 = n$ and $1 < k \leq n_0, n_1 \leq n$. Then the expectation of the difference D_k in the rate of success between the respective collections of trials $I_k(\boldsymbol{x})$ and $J_k(\boldsymbol{x})$*

satisfies

$$E[D_k \mid N_1 = n_1] = \frac{1}{\binom{n}{n_1} - U_k} \left[\sum_{\substack{r_{01},\ldots,r_{0k-1},s_{0k} \\ r_{11},\ldots,r_{1k-1},s_{1k} \\ \sum_{j=1}^{k-1} j r_{0j} < n_0 - k, s_{0k} \geq 1 \\ \sum_{j=1}^{k-1} j r_{1j} \leq n_1 - k, s_{1k} \geq 1 \\ r_0 \geq r_1}} C_k \left[\frac{s_{0k}}{r_0} D_k^1 + \frac{r_0 - s_{0k}}{r_0} D_k^3 \right] \right.$$

$$\left. + \sum_{\substack{r_{01},\ldots,r_{0k-1},s_{0k} \\ r_{11},\ldots,r_{1k-1},s_{1k} \\ \sum_{j=1}^{k-1} j r_{0j} \leq n_0 - k, s_{0k} \geq 1 \\ \sum_{j=1}^{k-1} j r_{1j} < n_1 - k, s_{1k} \geq 1 \\ r_1 \geq r_0}} C_k \left[\frac{s_{1k}}{r_1} D_k^2 + \frac{r_1 - s_{1k}}{r_1} D_k^3 \right] \right]$$

where $D_k^1 = (f_{1k} - s_{1k})/f_{1k} - (s_{0k} - 1)/(f_{0k} - 1), D_k^2 := (f_{1k} - s_{1k})/(f_{1k} - 1) - s_{0k}/f_{0k}, D_k^3 := (f_{1k} - s_{1k})/f_{1k} - s_{0k}/f_{0k}$,

$$C_k := \frac{r_0!}{s_{0k}! \prod_{i=1}^{k-1} r_{0i}!} \frac{r_1!}{s_{1k}! \prod_{i=1}^{k-1} r_{1i}!} \binom{f_{0k} - 1}{s_{0k} - 1} \binom{f_{1k} - 1}{s_{1k} - 1}$$

and U_k *(see Expression* $*$ *on page 60) is the number of sequences for which* $I_k(\boldsymbol{x}) = \emptyset$ *or* $J_k(\boldsymbol{x}) = \emptyset$.

Proof:

Note that for the case in which $|r_1 - r_0| = 1$, C_k is the number of sequences with $N_1 = n_1$ in which the number of runs of 0s, and runs of 1s satisfy run profile $(r_{01}, \ldots, r_{0k-1}, s_{0k}; r_{11}, \ldots, r_{1k-1}, s_{1k})$; for the cases in which $r_1 = r_0$, C_k is equal to half the number of these sequences (because each sequence can end with a run of 1s, or a run of 0s). The combinatorial proof of this formula, which we omit, is similar to the one used in the proof of Lemma 2.

The sum total of the difference, across all sequences for which the difference is defined and

$N_1 = n_1$ is:

$$E[D_k \mid N_1 = n_1] \cdot \left[\binom{n}{n_1} - U_k\right]$$
$$= \sum_{\substack{r_{01},\ldots,r_{0k-1},s_{0k} \\ r_{11},\ldots,r_{1k-1},s_{1k} \\ \sum_{j=1}^{k-1} jr_{0j} < n_0-k, s_{0k} \geq 1 \\ \sum_{j=1}^{k-1} jr_{1j} \leq n_1-k, s_{1k} \geq 1 \\ r_0 \geq r_1}} \frac{s_{0k}}{r_0} C_k D_k^1 \quad + \sum_{\substack{r_{01},\ldots,r_{0k-1},s_{0k} \\ r_{11},\ldots,r_{1k-1},s_{1k} \\ \sum_{j=1}^{k-1} jr_{0j} \leq n_0-k, s_{0k} \geq 1 \\ \sum_{j=1}^{k-1} jr_{1j} < n_1-k, s_{1k} \geq 1 \\ r_1 \geq r_0}} \frac{s_{1k}}{r_1} C_k D_k^2$$
$$+ \sum_{\substack{r_{01},\ldots,r_{0k-1},s_{0k} \\ r_{11},\ldots,r_{1k-1},s_{1k} \\ \sum_{j=1}^{k-1} jr_{0j} \leq n_0-k, s_{0k} \geq 1 \\ \sum_{j=1}^{k-1} jr_{1j} \leq n_1-k, s_{1k} \geq 1 \\ r_0 \geq r_1}} \frac{r_0 - s_{0k}}{r_0} C_k D_k^3 \quad + \sum_{\substack{r_{01},\ldots,r_{0k-1},s_{0k} \\ r_{11},\ldots,r_{1k-1},s_{1k} \\ \sum_{j=1}^{k-1} jr_{0j} \leq n_0-k, s_{0k} \geq 1 \\ \sum_{j=1}^{k-1} jr_{1j} \leq n_1-k, s_{1k} \geq 1 \\ r_1 \geq r_0}} \frac{r_1 - s_{1k}}{r_1} C_k D_k^3$$

where the first sum relates to those sequences that end in a run of 0s of length k or more (whence $r_0 \geq r_1$, the multiplier s_{0k}/r_0, and $\sum_{j=1}^{k-1} jr_{0j} < n_0 - k$);[77] the second sum relates to those sequences that end in a run of 1s of length k or more (whence $r_1 \geq r_0$, the multiplier s_{1k}/r_1, and $\sum_{j=1}^{k-1} jr_{1j} < n_1 - k$); the third sum relates to those sequences that end on a run of 0s of length k-1 or less (whence $r_0 \geq r_1$, the multiplier $(r_0 - s_{0k})/r_0$, and $\sum_{j=1}^{k-1} jr_{0j} < n_0 - k$);[78] and the fourth sum relates to those sequences that end on a run of 1s of length $k-1$ or less (whence $r_1 \geq r_0$, the multiplier $(r_1 - s_{1k})/r_1$, and $\sum_{j=1}^{k-1} jr_{1j} < n_1 - k$). These four terms can be combined into the following two terms:

$$E[D_k \mid N_1 = n_1] \cdot \left[\binom{n}{n_1} - U_k\right] = \sum_{\substack{r_{01},\ldots,r_{0k-1},s_{0k} \\ r_{11},\ldots,r_{1k-1},s_{1k} \\ \sum_{j=1}^{k-1} jr_{0j} < n_0-k, s_{0k} \geq 1 \\ \sum_{j=1}^{k-1} jr_{1j} \leq n_1-k, s_{1k} \geq 1 \\ r_0 \geq r_1}} C_k \left[\frac{s_{0k}}{r_0} D_k^1 + \frac{r_0 - s_{0k}}{r_0} D_k^3\right]$$
$$+ \sum_{\substack{r_{01},\ldots,r_{0k-1},s_{0k} \\ r_{11},\ldots,r_{1k-1},s_{1k} \\ \sum_{j=1}^{k-1} jr_{0j} \leq n_0-k, s_{0k} \geq 1 \\ \sum_{j=1}^{k-1} jr_{1j} < n_1-k, s_{1k} \geq 1 \\ r_1 \geq r_0}} C_k \left[\frac{s_{1k}}{r_1} D_k^2 + \frac{r_1 - s_{1k}}{r_1} D_k^3\right]$$

[77] Note that $\sum_{j=1}^{k-1} jr_{0j} < n_0 - k \iff f_{0k} \geq 2$.

[78] The multiplier $(r_0 - s_{0k})/r_0$ arises because the number of distinguishable permutations of the runs of 0s that end with a run of length $k-1$ or less is equal to the total number of distinguishable permutations of the runs of 0s minus the number of distinguishable permutations of the runs of 0s that end in a run of length k or more, i.e.

$$\frac{r_0!}{s_{0k}! \prod_{i=1}^{k-1} r_{0i}!} - \frac{(r_0-1)!}{(s_{0k}-1)! \prod_{i=1}^{k-1} r_{0i}!} = \frac{r_0 - s_{0k}}{r_0} \frac{r_0!}{s_{0k}! \prod_{i=1}^{k-1} r_{0i}!}$$

which can readily be implemented numerically for the finite sequences considered here.[79] The total number of sequences for which the difference is undefined, U_k, can be counted in a way that is analogous to what was done in the proof of Theorem 5, by using an application of the inclusion-exclusion principle:

$$
\begin{aligned}
U_k := & \sum_{\substack{r_{11},\ldots,r_{1k-1},s_{1k} \\ \sum_{j=1}^{k-1} j r_{1j}=n_1 \\ s_{1k}=0}} C_{1k} \quad + \sum_{\substack{r_{11},\ldots,r_{1k-1},s_{1k} \\ \sum_{j=1}^{k-1} j r_{1j}=n_1-k \\ s_{1k}=1}} \frac{s_{1k}}{n_0+1} C_{1k} \quad + \sum_{\substack{r_{01},\ldots,r_{0k-1},s_{0k} \\ \sum_{j=1}^{k-1} j r_{0j}=n_1 \\ s_{0k}=0}} C_{0k} \quad + \sum_{\substack{r_{01},\ldots,r_{0k-1},s_{0k} \\ \sum_{j=1}^{k-1} j r_{0j}=n_1-k \\ s_{0k}=1}} \frac{s_{0k}}{n_1+1} C_{0k} \\
& - \sum_{\substack{r_{01},\ldots,r_{0k-1},s_{0k} \\ r_{11},\ldots,r_{1k-1},s_{1k} \\ \sum_{j=1}^{k-1} j r_{0j}=n_0, s_{0k}=0 \\ \sum_{j=1}^{k-1} j r_{1j}=n_1, s_{1k}=0 \\ |r_0-r_1|\leq 1}} \left(2 \cdot \mathbb{1}_{\{r_1=r_0\}} + \mathbb{1}_{\{|r_1-r_0|=1\}}\right) C_k \\
& - \sum_{\substack{r_{01},\ldots,r_{0k-1},s_{0k} \\ r_{11},\ldots,r_{1k-1},s_{1k} \\ \sum_{j=1}^{k-1} j r_{0j}=n_0-k, s_{0k}=1 \\ \sum_{j=1}^{k-1} j r_{1j}=n_1, s_{1k}=0 \\ r_0 \geq r_1}} \frac{s_{0k}}{r_0} C_k \quad - \sum_{\substack{r_{01},\ldots,r_{0k-1},s_{0k} \\ r_{11},\ldots,r_{1k-1},s_{1k} \\ \sum_{j=1}^{k-1} j r_{0j}=n_0, s_{0k}=0 \\ \sum_{j=1}^{k-1} j r_{1j}=n_1-k, s_{1k}=1 \\ r_1 \geq r_0}} \frac{s_{1k}}{r_1} C_k
\end{aligned}
$$

where C_{0k} is a function of $r_{01}, \ldots, r_{0k-1}, s_{0k}; n_0, n_1$ and defined analogously to C_{1k}. We can simplify the above expression by first noting that the third term, which corresponds to those sequences in

[79] In the numerical implementation one can consider three sums: $r_0 = r_1 + 1$, $r_1 = r_0 + 1$, and for the case of $r_1 = r_0$ the sums can be combined.

which $I_k(\boldsymbol{x}) = \emptyset$ and $J_k(\boldsymbol{x}) = \emptyset$, can be reduced to a sum of binomial coefficients:

$$
\begin{aligned}
&\sum_{\substack{r_{01},\ldots,r_{0k-1},s_{0k}\\ r_{11},\ldots,r_{1k-1},s_{1k}\\ \sum_{j=1}^{k-1} jr_{0j}=n_0,s_{0k}=0\\ \sum_{j=1}^{k-1} jr_{1j}=n_1,s_{1k}=0\\ |r_0-r_1|\leq 1}} \left(2\cdot\mathbb{1}_{\{r_1=r_0\}}+\mathbb{1}_{\{|r_1-r_0|=1\}}\right) C_k \\
&\quad= \sum_{\substack{r_{01},\ldots,r_{0k-1},s_{0k}\\ r_{11},\ldots,r_{1k-1},s_{1k}\\ \sum_{j=1}^{k-1} jr_{0j}=n_0,s_{0k}=0\\ \sum_{j=1}^{k-1} jr_{1j}=n_1,s_{1k}=0\\ |r_0-r_1|\leq 1}} \left(2\cdot\mathbb{1}_{\{r_1=r_0\}}+\mathbb{1}_{\{|r_1-r_0|=1\}}\right) \frac{r_0!}{s_{0k}!\prod_{i=1}^{k-1} r_{0i}!}\frac{r_1!}{s_{1k}!\prod_{i=1}^{k-1} r_{1i}!} \\
&\quad= \sum_{\substack{r_{01},\ldots,r_{0k-1}\\ r_{11},\ldots,r_{1k-1}\\ \sum_{j=1}^{k-1} jr_{0j}=n_0\\ \sum_{j=1}^{k-1} jr_{1j}=n_1\\ |r_0-r_1|\leq 1}} \left(2\cdot\mathbb{1}_{\{r_1=r_0\}}+\mathbb{1}_{\{|r_1-r_0|=1\}}\right) \frac{r_0!}{\prod_{i=1}^{k-1} r_{0i}!}\frac{r_1!}{\prod_{i=1}^{k-1} r_{1i}!} \\
&\quad= \sum_{r_1=1}^{\min\{n_1,n_0+1\}} \sum_{r_0=\max\{r_1-1,1\}}^{\min\{r_1+1,n_0\}} \left(2\cdot\mathbb{1}_{\{r_1=r_0\}}+\mathbb{1}_{\{|r_1-r_0|=1\}}\right) \sum_{\substack{r_{01},\ldots,r_{0k-1}\\ r_{11},\ldots,r_{1k-1}\\ \sum_{j=1}^{k-1} jr_{0j}=n_0\\ \sum_{j=1}^{k-1} jr_{1j}=n_1\\ \sum_{j=1}^{k-1} r_{0j}=r_0\\ \sum_{j=1}^{k-1} r_{1j}=r_1}} \frac{r_0!}{\prod_{i=1}^{k-1} r_{0i}!}\frac{r_1!}{\prod_{i=1}^{k-1} r_{1i}!} \\
&\quad= \sum_{r_1=1}^{\min\{n_1,n_0+1\}} \sum_{r_0=\max\{r_1-1,1\}}^{\min\{r_1+1,n_0\}} \left(2\cdot\mathbb{1}_{\{r_1=r_0\}}+\mathbb{1}_{\{|r_1-r_0|=1\}}\right) \sum_{\substack{r_{01},\ldots,r_{0k-1}\\ \sum_{j=1}^{k-1} jr_{0j}=n_0\\ \sum_{j=1}^{k-1} r_{0j}=r_0}} \frac{r_0!}{\prod_{i=1}^{k-1} r_{0i}!} \sum_{\substack{r_{11},\ldots,r_{1k-1}\\ \sum_{j=1}^{k-1} jr_{1j}=n_1\\ \sum_{j=1}^{k-1} r_{1j}=r_1}} \frac{r_1!}{\prod_{i=1}^{k-1} r_{1i}!} \\
&\quad= \sum_{r_1=1}^{\min\{n_1,n_0+1\}} \sum_{r_0=\max\{r_1-1,1\}}^{\min\{r_1+1,n_0\}} \left(2\cdot\mathbb{1}_{\{r_1=r_0\}}+\mathbb{1}_{\{|r_1-r_0|=1\}}\right) \sum_{\ell_0=0}^{\lfloor\frac{n_0-r_0}{k-1}\rfloor} (-1)^{\ell_0}\binom{r_0}{\ell_0}\binom{n_0-1-\ell_0(k-1)}{r_0-1} \\
&\qquad\times \sum_{\ell_1=0}^{\lfloor\frac{n_1-r_1}{k-1}\rfloor} (-1)^{\ell_1}\binom{r_1}{\ell_1}\binom{n_1-1-\ell_1(k-1)}{r_1-1}
\end{aligned}
$$

For the final two negative terms in the formula for U_k, we can apply a similar argument in order to represent them as a sum of binomial coefficients. For the first four positive terms we can use the argument provided in Theorem 5 to represent them as sums of binomial coefficients. Therefore, U_k

reduces to the following sum of binomial coefficients:

$$
\begin{aligned}
U_k = & \sum_{r_1=1}^{\min\{n_1,n_0+1\}} \binom{n_0+1}{r_1} \sum_{\ell=0}^{\lfloor \frac{n_1-r_1}{k-1} \rfloor} (-1)^\ell \binom{r_1}{\ell} \binom{n_1-1-\ell(k-1)}{r_1-1} \qquad (*) \\
& + \delta_{n_1 k} + \sum_{r_1=2}^{\min\{n_1-k+1,n_0+1\}} \binom{n_0}{r_1-1} \sum_{\ell=0}^{\lfloor \frac{n_1-k-r_1+1}{k-1} \rfloor} (-1)^\ell \binom{r_1-1}{\ell} \binom{n_1-k-1-\ell(k-1)}{r_1-2} \\
& + \sum_{r_0=1}^{\min\{n_0,n_1+1\}} \binom{n_1+1}{r_0} \sum_{\ell=0}^{\lfloor \frac{n_0-r_0}{k-1} \rfloor} (-1)^\ell \binom{r_0}{\ell} \binom{n_0-1-\ell(k-1)}{r_0-1} \\
& + \delta_{n_0 k} + \sum_{r_0=2}^{\min\{n_0-k+1,n_1+1\}} \binom{n_1}{r_0-1} \sum_{\ell=0}^{\lfloor \frac{n_0-k-r_0+1}{k-1} \rfloor} (-1)^\ell \binom{r_0-1}{\ell} \binom{n_0-k-1-\ell(k-1)}{r_0-2} \\
& - \Bigg[\sum_{r_1=1}^{\min\{n_1,n_0+1\}} \sum_{r_0=\max\{r_1-1,1\}}^{\min\{r_1+1,n_0\}} \left(2 \cdot \mathbb{1}_{\{r_1=r_0\}} + \mathbb{1}_{\{|r_1-r_0|=1\}}\right) \times \\
& \qquad \times \sum_{\ell_0=0}^{\lfloor \frac{n_0-r_0}{k-1} \rfloor} (-1)^{\ell_0} \binom{r_0}{\ell_0} \binom{n_0-1-\ell_0(k-1)}{r_0-1} \sum_{\ell_1=0}^{\lfloor \frac{n_1-r_1}{k-1} \rfloor} (-1)^{\ell_1} \binom{r_1}{\ell_1} \binom{n_1-1-\ell_1(k-1)}{r_1-1} \Bigg] \\
& - \Bigg[\delta_{n_0 k} + \sum_{r_0=2}^{\min\{n_0-k+1,n_1+1\}} \sum_{r_1=\max\{r_0-1,1\}}^{\min\{r_0,n_1\}} \sum_{\ell_0=0}^{\lfloor \frac{n_0-k-r_0+1}{k-1} \rfloor} (-1)^{\ell_0} \binom{r_0-1}{\ell_0} \binom{n_0-k-1-\ell_0(k-1)}{r_0-2} \times \\
& \qquad \times \sum_{\ell_1=0}^{\lfloor \frac{n_1-r_1}{k-1} \rfloor} (-1)^{\ell_1} \binom{r_1}{\ell_1} \binom{n_1-1-\ell_1(k-1)}{r_1-1} \Bigg] \\
& - \Bigg[\delta_{n_1 k} + \sum_{r_1=2}^{\min\{n_1-k+1,n_0+1\}} \sum_{r_0=\max\{r_1-1,1\}}^{\min\{r_1,n_0\}} \sum_{\ell_1=0}^{\lfloor \frac{n_1-k-r_1+1}{k-1} \rfloor} (-1)^{\ell_1} \binom{r_1-1}{\ell_1} \binom{n_1-k-1-\ell_1(k-1)}{r_1-2} \times \\
& \qquad \times \sum_{\ell_0=0}^{\lfloor \frac{n_0-r_0}{k-1} \rfloor} (-1)^{\ell_0} \binom{r_0}{\ell_0} \binom{n_0-1-\ell_0(k-1)}{r_0-1} \Bigg]
\end{aligned}
$$

■

F Web Appendix: The relationship with known biases and paradoxes

F.1 Sampling-without-replacement and the bias for streaks of length $k = 1$.

A brief inspection of Table 1 in Section 1 reveals how the dependence between the first $n-1$ flips in the sequence arises. In particular, when the coin is flipped three times, the number of Hs in the first 2 flips determines the number of observations of flips that immediately follow an H. Because TT must be excluded, the first two flips will consist of one of three equally likely sequences: HT, TH or HH. For the two sequences with a single H—HT and TH—if a researcher were to find an H within the first two flips of the sequence and then select the adjacent flip for inspection, the probability of heads on the adjacent flip would be 0, which is strictly less than the overall proportion of heads in the sequence. This can be thought of as a sampling-without-replacement effect. More generally, across the three sequences, HT, TH, and HH, the expected probability of the adjacent flip being a heads is $(0 + 0 + 1)/3 = 1/3$. This probability reveals the (negative) sequential dependence that exists between the first two flips of the sequence. Further, the same negative dependence holds for *any two flips* in the first $n - 1$ flips of a sequence of length n, *regardless of their positions*. Thus, when $k = 1$ it is neither time's arrow nor the arrangement of flips within the sequence that determines the bias.

This same sampling-without-replacement feature also underlies a classic form of selection bias known as Berkson's bias (aka Berkson's paradox). Berkson (1946) presented a hypothetical study of the relationship between two diseases that, while not associated in the general population, become negatively associated in the population of hospitalized patients. The cause of the bias is subtle: patients are hospitalized only if they have *at least one* of the two particular diseases. To illustrate, assume that someone from the general population has a given disease (Y="Yes") or does not (N="No"), with equal chances. Just as in the coin flip example, anyone with neither disease (NN) is excluded, while a patient within the hospital population must have one of the three equally likely profiles: YN, NY, or YY. Thus, just as with the coin flips, the probability of a patient having another disease, given that he already has one disease, is 1/3.

The same sampling-without replacement feature again arises in several classic conditional probability paradoxes. For example, in the Monty Hall problem the game show host inspects two doors, which can together be represented as one of three equally likely sequences GC, CG, or GG (G="Goat", C="Car"), then opens one of the G doors from the realized sequence. Thus, the host effectively samples G without replacement (Nalebuff 1987; Selvin 1975; Vos Savant 1990).[80]

[80] The same structure also appears in what is known as the boy-or-girl paradox (Miller and Sanjurjo 2015a). A slight modification of the Monty-Hall problem makes it identical to the coin flip bias presented in Table 1 (see Miller and Sanjurjo [2015a]).

Sampling-without-replacement also underlies a well-known finite sample bias that arises in standard estimates of autocorrelation in time series data (Shaman and Stine 1988; Yule 1926). This interpretation of finite sample bias, which does not appear to have been previously noted, allows one to see how this bias is closely related to those above. To illustrate, let $\boldsymbol{x}$ be a randomly generated sequence consisting of n trials, each of which is an i.i.d. draw from some continuous distribution with finite mean and variance. For a researcher to compute the autocorrelation she must first determine its sample mean $\bar{x}$ and variance $\hat{\sigma}^2(\boldsymbol{x})$, then calculate the autocorrelation $\hat{\rho}_{t,t+1}(\boldsymbol{x}) = \hat{cov}_{t,t+1}(\boldsymbol{x})/\hat{\sigma}^2(\boldsymbol{x})$, where $\hat{cov}_{t,t+1}(\boldsymbol{x})$ is the autocovariance.[81] The total sum of values $n\bar{x}$ in a sequence serves as the analogue to the number of Hs (or Gs/Ys) in a sequence in the examples given above. Given $n\bar{x}$, the autocovariance can be represented as the expected outcome from a procedure in which one draws (at random) one of the n trial outcomes x_i, and then takes the product of its difference from the mean $(x_i - \bar{x})$, and another trial outcome j's difference from the mean. Because the outcome's value x_i is essentially drawn from $n\bar{x}$, without replacement, the available sum total $(n\bar{x} - x_i)$ is averaged across the remaining $n-1$ outcomes, which implies that the expected value of another outcome j's $(j \neq i)$ difference from the mean is given by $E[x_j|x_i, \bar{x}] - \bar{x} = (n\bar{x} - x_i)/(n-1) - \bar{x} = (\bar{x} - x_i)/(n-1)$. Therefore, given $x_i - \bar{x}$, the expected value of the product $(x_i - \bar{x})(x_j - \bar{x})$ must equal $(x_i - \bar{x})(\bar{x} - x_i)/(n-1) = -(x_i - \bar{x})^2/(n-1)$, which is independent of j. Because x_i and j were selected at random, this implies that the expected autocorrelation, given $\bar{x}$ and $\hat{\sigma}^2(\boldsymbol{x})$, is equal to $-1/(n-1)$ for all $\bar{x}$ and $\hat{\sigma}^2(\boldsymbol{x})$. This result accords with known results on the $O(1/n)$ bias in discrete-time autoregressive processes (Shaman and Stine 1988), and happens to be identical to the result in Theorem 4 for the expected difference in proportions (see Appendix A.2).[82]

F.2 Pattern overlap and the bias for streaks of length $k > 1$.

In Figure 6 of Web Appendix D we compare the magnitude of the bias in the (conditional) expected proportion to the pure sampling-without-replacement bias, in a sequence of length n. As can be seen, the magnitude of the bias in the expected proportion is nearly identical to that of sampling-without-replacement for $k = 1$. However, for the bias in the expected proportion, the relatively

[81] The autocovariance is given by $\hat{cov}_{t,t+1}(\boldsymbol{x}) := \frac{1}{n-1}\sum_{i=1}^{n-1}(x_i - \bar{x})(x_{i+1} - \bar{x})$.

[82] In a comment on this paper, Rinott and Bar-Hillel (2015) assert that the work of Bai (1975) (and references therein) demonstrate that the bias in the proportion of successes on the trials that immediately follow a streak of k or more successes follows directly from known results on the finite sample bias of Maximum Likelihood estimators of transition probabilities in Markov chains, as independent Bernoulli trials can be represented by a Markov chain with each state defined by the sequence of outcomes in the previous k trials. While it is true that the MLE of the corresponding transition matrix is biased, and correct to note the relationship in this sense, the cited theorems do not indicate the direction of the bias, and in any event do not directly apply in the present case because they require that transition probabilities in different rows of the transition matrix not be functions of each other, and not be equal to zero, a requirement which does not hold in the corresponding transition matrix. Instead, an unbiased estimator of each transition probability will exist, and will be a function of the overall proportion.

stronger sampling-without-replacement effect that operates within the first $n-1$ terms of the sequence is balanced by the absence of bias for the final term.[83] On the other hand, for $k > 1$ the bias in the expected proportion is considerably stronger than the pure sampling-without-replacement bias. One intuition for this is provided in the discussion of the updating factor in Section 2.1. Here we discuss another intuition, which has to do with the overlapping nature of the selection criterion when $k > 1$, which is related to what is known as the *overlapping words paradox* (Guibas and Odlyzko 1981).

For simplicity, assume that a sequence is generated by $n = 5$ flips of a fair coin. For the simple case in which streaks have length $k = 1$, the number of flips that immediately follow a heads is equal to the number of instances of H in the first $n - 1 = 4$ flips. For any given number of Hs in the first four flips, say three, if one were to sample an H from the sequence and then examine an adjacent flip (within the first four flips), then because any H could have been sampled, across all sequences with three Hs in the first four flips, any H appearing within the first four flips is given equal weight regardless of the sequence in which it appears. The exchangeability of outcomes across equally weighted sequences with an H in the sampled position (and three Hs overall) therefore implies that for any other flip in the first four flips of the sequence, the probability of an H is equal to $\frac{3-1}{4-1} = \frac{2}{3}$, regardless of whether or not it is an adjacent flip. On the other hand, for the case of streaks of length $k = 2$, the number of opportunities to observe a flip that immediately follows two consecutive heads is equal to the number of instances of HH in the first 4 flips. Because the pattern HH can overlap with itself, whereas the pattern H cannot, then for a sequence with three Hs, if one were to sample an HH from the sequence and examine an adjacent flip within the first 4 flips, it is not the case that any two of the Hs from the sequence can be sampled. For example, in the sequence HHTH only the first two Hs can be sampled. Because the sequences HHTH and HTHH each generate just one opportunity to sample, this implies that the single instance of HH within each of these sequences is weighted twice as much as any of the two (overlapping) instances of HH within the two sequences HHHT and THHH that each allow two opportunities to sample, despite the fact that each sequence has three heads in the first four flips. This implies that, unlike in the case of $k = 1$, when sampling an instance of HH from a sequence with three heads in the first four flips, the remaining outcomes H and T are no longer exchangeable, as the arrangements HHTH and HTHH, in which every adjacent flip within the first four flips is a tails, must be given greater weight than the arrangements HHHT and THHH, in which half of the adjacent flips are heads.

This consequence of pattern overlap is closely related to the *overlapping words paradox*, which states that for a sequence (string) of finite length n, the probability that a pattern (word) appears, e.g. _HTTHH_, depends not only on the length of the pattern relative to the length of the sequence,

[83] The reason for this is provided in the alternative proof of Lemma 1 in Appendix A.2

but also on how the pattern *overlaps* with itself (Guibas and Odlyzko 1981).[84] For example, while the expected number of (potentially overlapping) occurrences of a particular two flip pattern—TT, HT,TH or HH—in a sequence of four flips of a fair coin does not depend on the pattern, it's probability of occurrence does.[85] The pattern HH can overlap with itself, so can have up to three occurrences in a single sequence (HHHH), whereas the pattern HT cannot overlap with itself, so can have at most two occurrences (HTHT). Because the expected number of occurrences of each pattern must be equal, this implies that the pattern HT is distributed across more sequences, meaning that any given sequence is more likely to contain this pattern.[86]

[84] For a simpler treatment which studies a manifestation of the paradox in the non-transitive game known as "Penney's" game, see Konold (1995) and Nickerson (2007).

[85] That all fixed length patterns are equally likely ex-ante is straightforward to demonstrate. For a given pattern of heads and tails of length ℓ, $(y_1, \ldots, y_\ell)$, the expected number of occurrences of this pattern satisfies $E[\sum_{i=\ell}^{n} 1_{[(X_{i-\ell+1},\ldots,X_i)=(y_1,\ldots,y_\ell)]}] = \sum_{i=\ell}^{n} E[1_{[(X_{i-\ell+1},\ldots,X_i)=(y_1,\ldots,y_\ell)]}] = \sum_{i=\ell}^{n} 1/2^\ell = (n-\ell+1)/2^\ell$.

[86] Note that the proportion of heads on flips that immediately follow two consecutive heads can be written as the number of (overlapping) HHH instances in n flips, divided by the number of (overlapping) HH instances in the first $n-1$ flips (see equation 21 in Web Appendix E).