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

Joshua B. Miller and Adam Sanjurjo *†‡

November 8, 2016

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 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 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, Daniel Martin, Mark Machina, 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, Vernon 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., NHH Norway, Microsoft Research, U. of Minnesota, Naval Postgraduate School, the Ohio State U., Santa Clara U., Stanford U., Tilburg U., U de Alicante, U. of Amsterdam, UC Berkeley, UC Irvine, UC Santa Cruz, UC San Diego, U. New South Wales, U. Southern California, U. of Queensland, U. of Wellington, U. of Zurich, 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 typically happens after 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.[1] 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.

Jack's procedure for writing down flips illustrates a novel form of selection bias which we identify, and prove more generally. We find that this 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 3.3) 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, the bias can be used 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 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 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]

[1] Assume that whenever Jack flips a tails, he will not write down the outcome of the next flip.

[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.

[4] If Jack were instead to control the number of flips he records by flipping the coin until he records the outcomes of

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):		$\frac{5}{12}$

In Section 2 we generalize this result and find 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.[5] 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 dimensionality of the problem, which yields a formula for the bias that is numerically tractable for sequence lengths commonly used in empirical work. We find

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.

[5] This assumes only that the length of the sequence n satisfies $n \geq 3$, and that the streak length k satisfies $1 \leq k < n-1$.

that while the 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.

In Section 2.2 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 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 3.4 we show how GVT have misinterpreted their estimates, and further observe that their tests are underpowered. We re-analyze GVT's betting data and, in contrast with their findings, we show that players can successfully predict shot outcomes at rates significantly (and substantially) greater than chance would predict. This suggests that players can profitably exploit their belief in the hot hand. Further, we discuss findings from a separate study which shows that players can identify which teammates have a tendency to get the hot hand (Miller and Sanjurjo 2014). While these results are important and show that decision makers have some of the necessary ingredients to make well-informed decisions based on the hot hand, we conclude that the existing evidence is ultimately unclear about whether decision-makers can detect the hot hand in real-time, and whether their responses are well-calibrated. We suggest avenues for future research.

In Section 4 we show how the bias has implications for the study of 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.[6] 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.

[6]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 2.2 (and Appendix D.1) we explain that the Monty Hall problem is essentially equivalent to the simplest version of the bias, with $n = 3$ and $k = 1$.

More subtly, the bias can also be leveraged to manipulate people into believing that the outcomes of an unpredictable process can be predicted at rates better than chance.[7] 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.[8]

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

In Section 2.1 we provide a proof of the bias. In Section 2.2 we discuss the two mechanisms behind the bias, and relate the bias to other known biases and paradoxes. In Section 2.3 we quantify the bias for empirically relevant parameter values, and graphically depict the bias as a function of various sequence lengths, streak lengths, and underlying probability of success.

2.1 A proof of the bias in the estimator

Let $\{X_i\}_{i=1}^n$ 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 conditional probability of success on trial t, given that trial t immediately follows k consecutive successes, is to first select all of the trials t that immediately follow k consecutive successes ($\prod_{j=t-k}^{t-1} X_j = 1$), then calculate the proportion of successes on the selected trials.[9] The following theorem establishes that when $\{X_i\}_{i=1}^n$ is a sequence

[7] 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.

[8] 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.

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

of i.i.d random variables, with probability of success $p := \mathbb{P}(X_t = 1) \equiv \mathbb{P}(X_t = 1 \mid \prod_{j=t-k}^{t-1} X_j = 1)$, this procedure yields a biased estimator of the conditional probability.

Theorem 1 *Let $\{X_i\}_{i=1}^n$, $n \geq 3$, be a sequence of independent Bernoulli trials, each with probability of success $0 < p < 1$. Let $I_{1k}(\boldsymbol{X}) := \{i : \prod_{j=i-k}^{i-1} X_j = 1\} \subseteq \{k+1, \ldots, n\}$ be the subset of trials that immediately follow k consecutive successes, and $\hat{P}_{1k}(\boldsymbol{X})$ the proportion of successes in $I_{1k}(\boldsymbol{X})$. For $1 \leq k \leq n-2$, $\hat{P}_{1k}(\boldsymbol{x})$ is a biased estimator of p. In particular,*

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

Proof: See Appendix A

The main intuition behind the proof can be illustrated by modifying the opening example from Section 1 in order to operationalize the expected proportion as a conditional probability. In particular, suppose that a researcher will generate a predetermined number of i.i.d. Bernoulli trials $\{x_i\}_{i=1}^n$, with $n > 2$. For each trial t, the researcher will record the outcome on his scrap of paper if and only if the previous k outcomes are successes, i.e. for $t \in I_{1k}(\boldsymbol{x})$. Next, if he has recorded at least one outcome, i.e. $I_{1k}(\boldsymbol{x}) \neq \emptyset$, then he will circle *one* of the outcomes on his scrap of paper (uniformly at random). If the researcher were to know all of the outcomes from the sequence $\boldsymbol{x}$, then the probability of circling a success would be $\hat{P}_{1k}(\boldsymbol{x})$. However, before generating the sequence, the probability of circling a success, conditional on recording at least one outcome, is instead $E[\ \hat{P}_{1k}(\boldsymbol{X}) \mid I_{1k}(\boldsymbol{X}) \neq \emptyset\ ]$. Now, if the researcher were to circle the outcome corresponding to trial $t < n$, then by not having circled trial $t+1$'s outcome, the posterior odds in favor of a sequence in which he could *not* have done so (because $t+1$'s outcome had not been written down),[10] as opposed to an otherwise identical sequence in which he *could* have circled trial $t+1$'s outcome (because the outcome had been written down), will be strictly greater than the prior odds. From this it follows immediately that the posterior odds in favor of the circled outcome being a failure will also be greater than the prior odds. On the other hand, if the researcher were to instead circle the outcome corresponding to trial $t = n$, then the posterior odds would remain unchanged, as there would have been no trial $t+1$, thus no associated outcome that could have been circled. Finally, because he does not know which trial's outcome he will circle, and each of the $n-k$ feasible trials from $k+1$ to n has a nonzero probability of being circled before the sequence is generated, then if he were to circle the outcome corresponding to trial t, the posterior odds that it is a failure would be strictly greater than the prior odds. This implies that the probability of the researcher circling a success is less than the prior probability of success (p), which in turn implies that, in expectation, the proportion of successes among the flip outcomes recorded on his scrap of paper will be less than the probability of success, i.e. $E[\ \hat{P}_{1k}(\boldsymbol{X}) \mid I_{1k}(\boldsymbol{X}) \neq \emptyset\ ] < p$.

[10] In fact, the number of trials ruled out are typically more than this. If trial $t+1$ is not written down, this implies trial t is a failure, and therefore no trial $i \in \{t+1, \ldots, \min\{t+k, n\}\}$ can be written down.

2.2 The mechanism behind the bias, sampling-without-replacement, and relation to known results

Any sequence $\boldsymbol{x} \in \{0,1\}^n$ with $I_{1k}(\boldsymbol{x}) \neq \emptyset$ that a researcher encounters will contain a certain number of successes $N_1(\boldsymbol{x}) = n_1$ and failures $n_0 := n - n_1$, where $n_1 \in \{k, \ldots, n\}$. To estimate the conditional probability of interest, the researcher will select only the trials in the sequence that satisfy $t \in I_{1k}(\boldsymbol{x})$, and then compute the proportion of successes on those trials, i.e. $\hat{P}_{1k}(\boldsymbol{X})$. Assuming that n_1 is known, the prior odds in favor of a success on any given trial in the sequence are $n_1/n_0 : 1$, whereas the odds are strictly less than this for any given trial in $I_{1k}(\boldsymbol{x})$. An intuition for why, which also reveals the mechanism behind the bias, can be obtained by considering the following equation, which we derive in Appendix A.3 (equation 11):

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

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 $I_{1k}(\boldsymbol{x})$.[11,12] Observe that the prior odds ratio n_1/n_0 is multiplied by two separate updating factors. Each of these factors is strictly less than one when $t < n$, as we will now discuss. Thus, each acts to attenuate the prior odds, resulting in posterior odds that are smaller than the prior.

The first updating factor $(n_1 - k)/n_1 < 1$ reflects the constraint that the finite number of available successes places on the procedure for selecting the trials $I_{1k}(\boldsymbol{x})$. In particular, it can 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 of n_1/n_0. This effect is perhaps easier to see 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 $\frac{n_1-k}{n-k}/\frac{n_0}{n-k}$. 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 have been removed from the urn. The denominator is the probability of drawing a 0 from the same urn, given that no 0's have previously been removed. Clearly, the strength of the bias, via 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 A.3), reflects an additional constraint that the arrangement of successes and failures in the sequence places on the procedure for selecting trials into $I_{1k}(\boldsymbol{x})$. It can be thought of as the additional information gained by learning that the k successes, which are no longer available, are consecutive and immediately

[11] This is the same selection procedure that is described with the intuition for the proof of Theorem 1. It operationalizes the expected proportion of interest as a conditional probability.

[12] See Appendix A.2 for a derivation of the posterior odds in the case that $\hat{p} = n_1/n$ is unknown.

precede t. To see why the odds are further attenuated in this case, we begin with the random variable M, which is defined to be the number of trials in $I_{1k}(\boldsymbol{x})$. The probability of any particular trial $t \in I_{1k}(\boldsymbol{x})$ being selected at random is $1/M$.[13] 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 $x_t = 1$, trial $t+1$ must also be in $I_{1k}(\boldsymbol{x})$, and trials $t+2$ through $t+k$ each may also be in $I_{1k}(\boldsymbol{x})$. By contrast, for a sequence in which $x_t = 0$, trials $t+1$ through $t+k$ cannot possibly be in $I_{1k}(\boldsymbol{x})$, which leads one to expect the corresponding $1/M$ to be smaller.[14] This last argument provides intuition for why the strength of the bias, via this factor, also increases in k.[15]

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

$$E\left[\, \hat{P}_{11}(\boldsymbol{X}) \;\middle|\; I_{11}(\boldsymbol{X}) \neq \emptyset,\ N_1(\boldsymbol{X}) = n_1 \,\right] = \frac{n_1 - 1}{n - 1} \tag{3}$$

as demonstrated in Lemma 2, in Appendix B.[16] In Appendix D.1 we show that sampling-without-replacement reasoning alone can be used to demonstrate that when $k = 1$ the bias is essentially identical to:[17] (1) a classic form of selection bias known as Berkson's bias, or Berkson's paradox (Berkson 1946; Roberts et al. 1978), (2) classic conditional probability puzzles such as the Monty Hall problem (Nalebuff 1987; Selvin 1975; Vos Savant 1990), and (3) finite sample bias in autocorrelation estimators (Shaman and Stine 1988; Yule 1926).[18] In addition to identifying the connections between these biases—which have not been noted before—we provide a novel insight into Berkson's bias by establishing conditions under which it can be expected to be empirically relevant.

[13] Following the intuition from the Introduction, $1/M$ represents the implicit weight placed on each trial $t \in I_{1k}(\boldsymbol{x})$ in the sequence $\boldsymbol{x}$.

[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 81.

[16] This follows from Equation 14 in the discussion of the alternative proof of Lemma 2 in Appendix B.

[17] While the overall bias is equal in magnitude to that of sampling-without-replacement in Equation 3, we have seen that the bias in the procedure used to select 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 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. It is sampling-without-replacement with respect to M which determines the bias when $k = 1$.

[18] We thank Fillipo Massari for insisting that the connection between the bias and the Monty Hall problem may extend beyond their shared paradoxical qualities.

On the other hand, when $k > 1$ the bias is substantially stronger than sampling-without-replacement (see Figure 4 in Appendix A.3). An intuition that complements the explanation of the mechanism given above is that the bias is determined not by the number of successes n_1 in a sequence of length n, but by the number of (overlapping) instances of k consecutive successes within the first $n-1$ trials, which depends on both the number of successes and their *arrangement*. Specifically, in Appendix C we show that the essential feature of the arrangement is how the successes and failures are grouped into *runs*. In addition, in Appendix D.2, we explain how this feature of the bias relates to what is known as the *overlapping words paradox* (Guibas and Odlyzko 1981).[19]

Our demonstration of the relationship between the bias and sampling-without-replacement calls to mind the key behavioral assumption made in Rabin (2002), that believers in the law of small numbers view signals from an i.i.d. process as if they were instead generated by random draws without replacement. Indeed, in Section 4 we use the bias to provide a novel structural explanation for how such a belief can persist in the face of experience.

2.3 Quantifying the bias.

In order to derive an explicit formula for $E[\ \hat{P}_{1k}(\boldsymbol{X}) \mid I_{1k}(\boldsymbol{X}) \neq \emptyset\ ]$, and quantify the magnitude of the corresponding bias, we first derive a formula for the conditional expectation, given the number of successes in the sequence, $N_1(\boldsymbol{x}) := \sum_{i=1}^{n} x_i$. It then follows from the law of total expectations that,

$$E\left[\ \hat{P}_{1k}(\boldsymbol{X})\ \middle|\ I_{1k}(\boldsymbol{X}) \neq \emptyset\ \right] = E\left[E\left[\ \hat{P}_{1k}(\boldsymbol{X})\ \middle|\ I_{1k}(\boldsymbol{X}) \neq \emptyset,\ N_1(\boldsymbol{X}) = n_1\ \right]\right] \tag{4}$$

The value of $E[\ \hat{P}_{1k}(\boldsymbol{X}) \mid I_{1k}(\boldsymbol{X}) \neq \emptyset,\ N_1(\boldsymbol{X}) = n_1]$ can, in principle, be obtained directly by first computing $\hat{P}_{1k}(\boldsymbol{x})$ for each sequence that contains n_1 successes, 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. 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. Our solution to this problem is to derive a numerically tractable formula by identifying, and enumerating, the set of sequences for which $\hat{P}_{1k}(\boldsymbol{x})$ 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 for Equation 4, by way of combinatorial argument (see Appendix C). While the formula does not admit a simple representation for $k > 1$, it is numerically tractable for the sequence and streak lengths that are

[19] We thank Kurt Smith for suggesting that the work of Guibas and Odlyzko (1981) could be related.

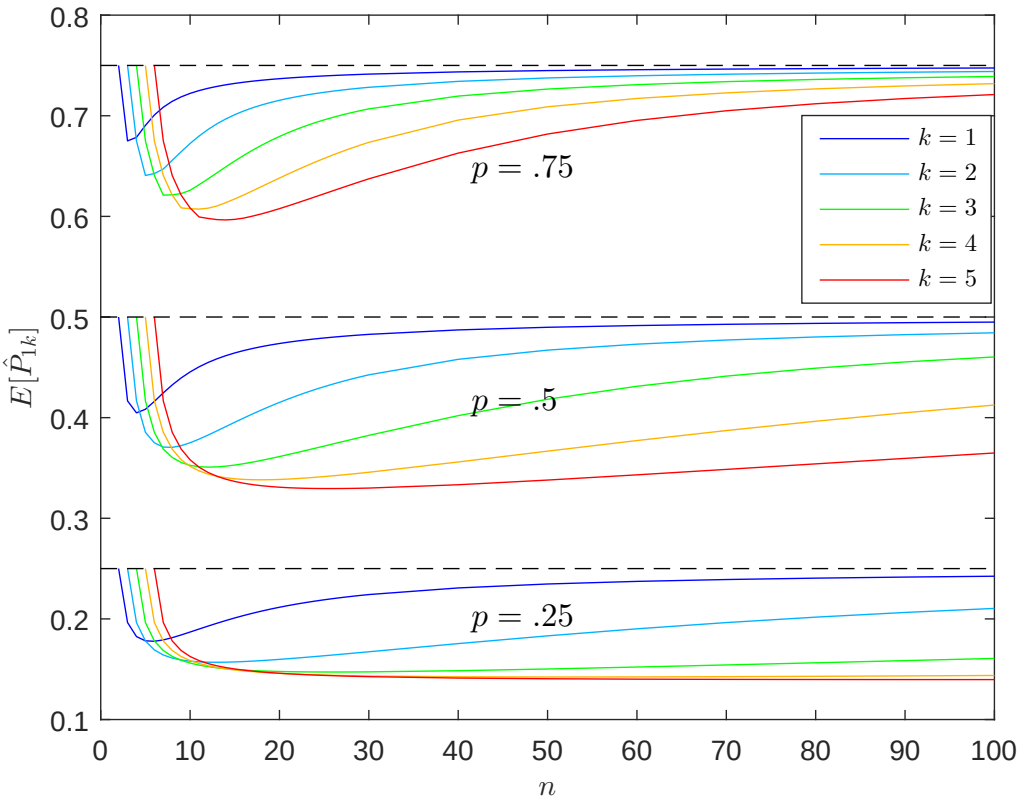

Figure 1: *The expected value of the proportion of successes on trials that immediately follow k consecutive successes, $\hat{P}_{1k}(\boldsymbol{X})$, as a function of the total number of trials n, for different values of k and probabilities of success p (using the formula provided in Theorem 6, combined with Equation 4).*

empirically relevant. For the special case of $k = 1$ a simple representation exists, and is presented in Appendix B.

2.3.1 The magnitude of the bias in the expected proportion

Figure 1 contains a plot of $E[\ \hat{P}_{1k}(\boldsymbol{X}) \mid I_{1k}(\boldsymbol{X}) \neq \emptyset\ ]$, as a function of the number of trials in the sequence n, and for different values of k and p.[20] 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, $p - E[\hat{P}_{1k}]$, 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 $.16 - .25 = -0.09$.

[20] For $k > 1$ the figure was produced by combining Equation 4 with the formula provided in Theorem 6 (Appendix C).

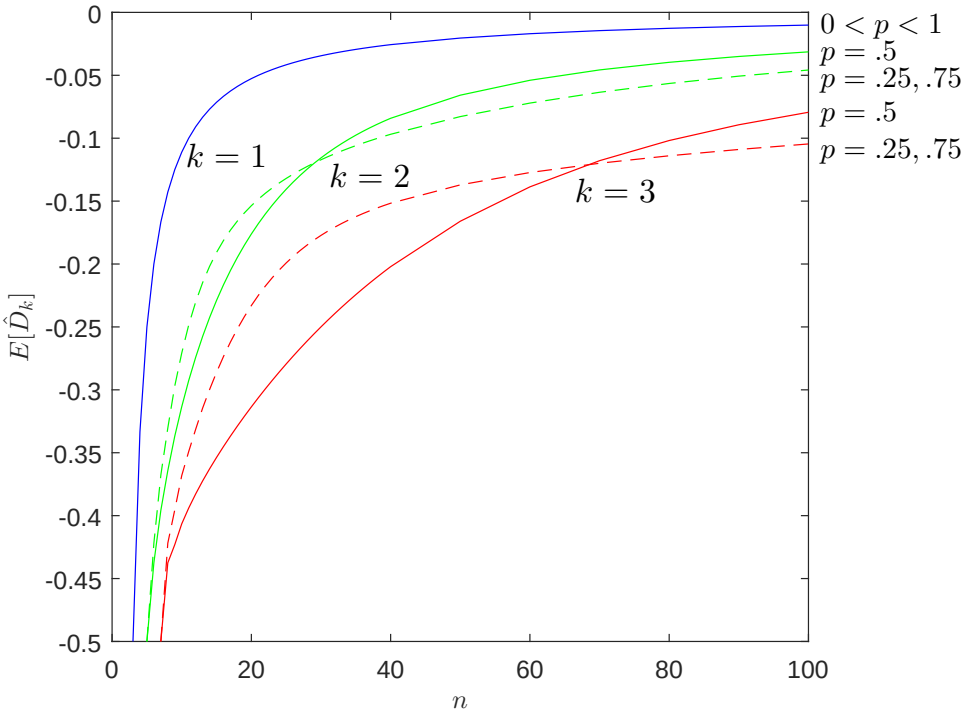

Figure 2: *The expected difference in proportions, $\hat{D}_k := \hat{P}_{1k} - \hat{P}_{0k}$, where the proportion of successes $\hat{P}_{1k}$ is computed for the trials that immediately follow a streak of k or more successes, and the proportion of successes $\hat{P}_{0k}$ is computed for the trials that immediately follow a streak of k or more failures, as a function of n, three values of k, and various probabilities of success p (using the formula provided in Theorem 7, combined with Equation 4).*

2.3.2 The magnitude of the bias for the difference in proportions

Let $\hat{D}_k(\boldsymbol{x}) := \hat{P}_{1k}(\boldsymbol{x}) - \hat{P}_{0k}(\boldsymbol{x})$, where $\hat{P}_{0k}(\boldsymbol{x})$ is the proportion of successes on the trials that immediately follow k consecutive failures, i.e. $I_{0k}(\boldsymbol{x}) := \{i : \prod_{j=i-k}^{i-i}(1 - x_j) = 1\} \subseteq \{k+1, \ldots, n\}$. This difference is a biased estimator of the true difference in (conditional) probabilities, and is relevant for statistical tests used in the hot hand fallacy literature, as discussed in Section 3. The magnitude of the bias in the difference is slightly greater than double the bias in the proportion. For the simple case in which $k = 1$ the bias is independent of p, and the expected difference in proportions admits the simple representation $E[\, \hat{D}_k(\boldsymbol{X}) \mid I_{1k}(\boldsymbol{X}) \neq \emptyset, I_{0k}(\boldsymbol{X}) \neq \emptyset \,] = -1/(n-1)$. We prove this in Appendix B. For the case of $k > 1$, refer to Appendix C.

Figure 2 contains a plot of $E[\, \hat{D}_k(\boldsymbol{X}) \mid I_{1k}(\boldsymbol{X}) \neq \emptyset, I_{0k}(\boldsymbol{X}) \neq \emptyset \,]$ as a function of the number of trials n, and for $k = 1, 2$, and 3.[21] Because the bias is dependent on p when $k > 1$, the difference is

[21] While Figure 1 also includes the cases $k = 4, 5$ in the plot of the expected value of $\hat{P}_{1k}(\boldsymbol{x})$, these cases are not plotted in Figure 2 because of the computational requirements arising from the number of terms in the sum.

plotted for various values of p. These expected differences are obtained by combining Theorem 4 with the results in Appendix C. The magnitude of the bias is obtained by comparing the expected difference to zero, and as in the case of the proportion, can remain substantial even as n gets large.

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 1985)

The hot hand fallacy refers to the mistaken belief that success tends to follow success (hot hand), when in fact observed patterns of successes and failures are consistent with the typical fluctuations of an i.i.d. random process. The seminal paper of Gilovich, Vallone, and Tversky (1985; henceforth GVT) introduced the hot hand fallacy, finding that while basketball players believe that a shooter has "a better chance of making a shot after having just made his last two or three shots than he does after having just missed his last two or three shots," the evidence from their analysis of shooting data shows that players' beliefs are wrong.

Because the incorrect beliefs are held by experts who, despite the evidence, continue to make high-stakes decisions based on these beliefs, the hot hand fallacy has come to be known as a "massive and widespread cognitive illusion"(Kahneman 2011).[22,23] Further, it has had a pronounced influence on empirical and theoretical work in economics, finance, and psychology.[24] This is due, presumably, to the surprising nature of the original result, the striking irrationality of basketball professionals' refusal to accept it, and the fact that the perception of patterns in sequential data is relevant in many domains of decision making.

In the following subsections we explain the bias in GVT's analysis of data, conduct a de-biased analysis using two separate approaches, and, in light of our results, re-assess evidence for both the

[22] GVT's result has had a notable impact on popular culture (see Gould (1989) [link]). The existence of the fallacy itself has been highlighted in the popular discourse as a salient example of how statistical analysis can reveal the flaws of expert intuition (e.g. see Davidson (2013, May 2) [link]).

[23] For evidence that players continue to make consequential decisions based on their hot hand beliefs see Aharoni and Sarig (2011); Attali (2013); Avugos, Köppen, Czienskowski, Raab, and Bar-Eli (2013b); Bocskocsky, Ezekowitz, and Stein (2014); Rao (2009a).

[24] The hot hand fallacy has been given considerable weight as a candidate explanation for various puzzles and behavioral anomalies identified in the domains of financial markets, sports wagering, casino gambling, and lotteries (Arkes 2011; Avery and Chevalier 1999; Barberis and Thaler 2003; Brown and Sauer 1993; Camerer 1989; Croson and Sundali 2005; De Bondt 1993; De Long, Shleifer, Summers, and Waldmann 1991; Durham, Hertzel, and Martin 2005; Galbo-Jørgensen, Suetens, and Tyran 2015; Guryan and Kearney 2008; Kahneman and Riepe 1998; Lee and Smith 2002; Loh and Warachka 2012; Malkiel 2011; Narayanan and Manchanda 2012; Paul and Weinbach 2005; Rabin and Vayanos 2010; Sinkey and Logan 2013; Smith, Levere, and Kurtzman 2009; Sundali and Croson 2006; Xu and Harvey 2014; Yuan, Sun, and Siu 2014).

hot hand and the hot hand fallacy.

3.1 The bias in GVT

GVT proposed that if the hot hand (or "streak" shooting) exists, then regardless of how it is defined, player performance records—patterns of hits (successes) and misses (failures)—should "differ from sequences of heads and tails produced by [weighted] coin tosses"(Gilovich et al. 1985).[25] While this proposal allows one to test for the *existence* of hot hand shooting, in order to evaluate the *relevance* one must estimate the magnitude of the effect, i.e. the associated change in a shooter's probability of hitting a shot. GVT (and subsequent studies) operationalize this measure as the percentage point difference in a player's field goal percentage (proportion of hits) between shots taken on a hit streak and shots taken on a miss streak. In particular, a player is on a hit (miss) streak if the previous k consecutive shot outcomes are identical (Avugos, Bar-Eli, Ritov, and Sher 2013a; Gilovich et al. 1985; Koehler and Conley 2003).[26,27]

The idea of testing whether a player's field goal percentage depends on the outcome of the immediately preceding shots appears to be a sound one: if each shot has the same probability of success (the null hypothesis), then whether a given shot follows a streak of hits or a streak of misses is determined by an independent chance event. Therefore, it is natural to treat these two sets of shot attempts as statistically independent treatments. In particular, for each shot i, if the preceding k shots $i-1$ through $i-k$ are hits, let i be assigned to the "k-hits" treatment, whereas if the preceding k shots are misses, let i be assigned to the "k-misses" treatment.[28] Given the independence assumption, the null hypothesis for GVT's associated statistical test is that the (mean) field goal percentage is equal across treatments. However, the independence assumption, while intuitive, is incorrect. What it overlooks is that, given a sequence of finite length, the act of assigning each shot to a treatment based on the outcome of the previous shot(s) happens to

[25] In particular GVT highlight patterns relating to two types of mechanisms for hot hand shooting: (1) feedback from preceding shot outcomes into a player's probability of success ("autocorrelation"), (2) shifts in a player's probability of success unrelated to previous outcomes ("non-stationarity").

[26] The commonly used cutoff for the definition of a streak is three, which happens to agree with the "rule of three" for people to perceive consecutive outcomes of the same kind to be a streak (Carlson and Shu 2007). While this definition is arbitrary, in practice the choice of the cutoff involves a trade-off. The larger the cutoff k, the higher the probability that the player is actually hot on those shots that are immediately preceded by a streak of hits, which reduces measurement error. On the other hand, as k gets larger, the bias from Section 2 increases, and fewer shots are available, which leads to a smaller sample size and reduced statistical power. See Miller and Sanjurjo (2014) for a more thorough discussion, which investigates statistical power and measurement error for a number of plausible hot hand models.

[27] GVT also conduct a runs test, a test of serial correlation, and test if the proportion of hits is influenced by whether a shot is preceded by a hit or by a miss (conditional probability test). We find that all three of these tests amount to the same test, and moreover, that they are not powered to identify hot hand shooting. The reason why is that the act of hitting a single shot is only a weak signal of a change in a player's underlying probability of success.

[28] If shot $i \leq k$, or if it does not immediately follow a streak of at least k outcomes of the same kind, then it is not assigned to either treatment.

increase the likelihood that any given "assigned" shot is different than the previous shot. For example, as shown in Section 2, if a player has a constant probability p of hitting a shot then the probability that a randomly chosen ("representative") shot is a hit, among those "assigned" to the k-hits treatment, is strictly less than p. This bias has important implications for GVT's study of basketball shooting data.

GVT's study of basketball shooting

GVT analyzed shooting data from three sources: live ball field goal data from the NBA's Philadelphia 76ers (1980-81 season: 9 players, 48 home games), dead ball free throw data from the NBA's Boston Celtics (1980-81, 1981-82 seasons: 9 players), and a shooting study that they conducted with Cornell's intercollegiate (NCAA) basketball teams (26 players, 100 shots each from a fixed distance, with varying locations). The Cornell study was designed for the purpose of "eliminating the effects of shot selection and defensive pressure" and is GVT's most controlled test of hot hand shooting, which makes it central to their main conclusions. Therefore, we focus on this data below when discussing the relevance of the bias on GVT's results.[29,30]

Upon calculating the size of the bias that affects GVT's statistical tests we observe that it is large enough to make their conclusions hinge on whether it is accounted for, or not. To see why, for each of the 26 players, the authors first assign each of the 100 shot attempts to either the k-hits or k-misses treatment, separately for $k = 1$ then $k = 2$, then $k = 3$ (as defined above), discarding any shots that are not immediately preceded by a streak of k. Next, for each of the players but one they find the differences in field goal percentages across treatments to be statistically indistinguishable.[31] Of course, given the results that we report in Section 2.3, these differences are biased, which makes the tests based on the differences biased as well. In particular, given the parameters of the study, and the most commonly considered streak length ($k = 3$), the expected difference for a *consistent*

[29] From the statistical point of view, the 76ers' in-game field goal data is not ideal for the study of hot hand shooting for reasons unrelated to the bias (see, e.g. Miller and Sanjurjo (2014)). 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 2013), 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 et al. 2014; Rao 2009a). Further, evidence of specific forms of strategic adjustment has been documented (Aharoni and Sarig 2011; Bocskocsky et al. 2014).

[30] The Celtics' in-game free throw data is not ideal, 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) there is a potential for 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 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).

[31] The significant effect size GVT found in a single player does not control for multiple comparisons.

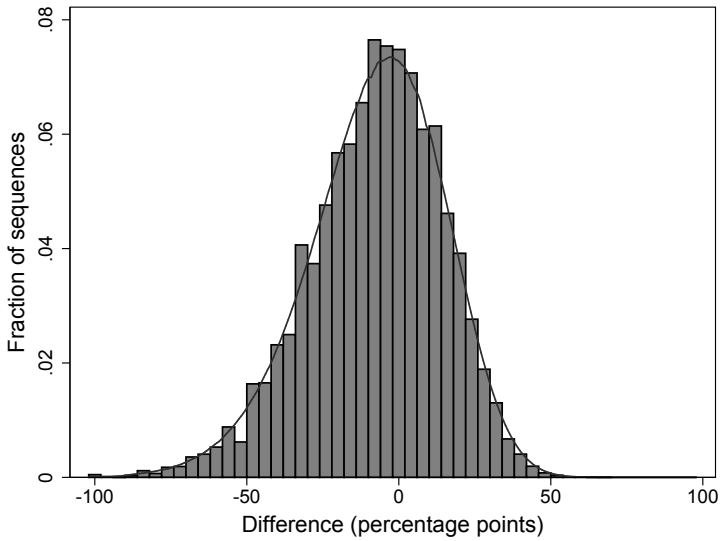

Figure 3: *The histogram and kernel density plot of the (exact) discrete probability distribution of* $\hat{D}_3|N_1 = n_1$*, where* $\hat{D}_3 := \hat{P}_{13} - \hat{P}_{03}$ *is the difference between the proportion of successes on those trials that immediately follow a streak of* 3 *or more successes, and the proportion of successes on those trials that immediately follow a streak of* 3 *or more failures, for a single player with* $n = 100$ *and* $n_1 = 50$ *(using the formula for the distribution provided in the proof of Theorem 7, with a bin width of 4 percentage points).*[33]

shooter, i.e. a shooter who hits with constant probability of success p, is not 0, but instead -8 percentage points. Moreover, the distribution of the differences has a pronounced negative skew. To illustrate, Figure 3 gives the *exact* distribution of the difference, based on the enumeration used in Theorem 7 of Appendix C. The distribution is generated using the target parameters of the original study: sequences of length $n = 100$, $n_1 = 50$ hits, and streaks of length $k = 3$, or more. The skewness in the distribution is pronounced, with 63 percent of observations below 0, and a median of -.06. To get a sense of how GVT's results compare with the bias, in their summary statistics the authors report average field goal percentages of 49 percent in the 3-hits treatment, and 45 percent in the 3-misses treatment (unweighted, across players).[32] However, this difference of 4 percentage points happens to be 12 percentage points higher than what would be expected from a consistent shooter with a probability of success equal to .5. This observation reveals that the bias has long disguised evidence that may well indicate hot hand shooting.

[32] Gilovich et al. (1985), Table 4, p. 307.

[33] The values for $\hat{D}_3$ 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 when calculating $\hat{D}_3$. In the computation of the expected value in Figures 1 and 2, each difference is instead represented with the highest floating point precision available.

3.2 An unbiased statistical analysis of GVT

We present two approaches to de-biasing GVT's estimates and conducting unbiased statistical tests. The first approach modifies GVT's individual player t-tests, and then extends this analysis to test for an average effect. The second approach provides an exact test of GVT's null hypothesis that each player shoots with a constant probability of success.

Bias-adjusted t-test

A straightforward way to adjust for the bias in GVT's test statistic is to shift the estimated difference used in each of GVT's 26 t-tests by the corresponding bias. In particular, for each player we compute the bias under the null hypothesis that trials are Bernoulli (i.e. consistent shooting) with a probability of success equal to the player's observed field goal percentage. This bias adjustment is *conservative*, as the bias becomes much larger if we instead assume that the underlying data generating process involves hot hand shooting (see Appendix E). Table 2 reproduces two columns from Table 4 (p. 307) of Gilovich et al. (1985), providing shooting performance records for each of the 14 male and 12 female Cornell University basketball players who participated in the controlled shooting experiment. From left to right, the table includes the number of shots taken ("# shots"), the overall proportion of hits ("$\hat{p}(hit)$"), the proportion of hits in the 3-hits treatment ("$\hat{p}(hit|3\ hits)$"), the proportion of hits in the 3-misses treatment ("$\hat{p}(hit|3\ misses)$"), the observed difference in proportions across the 3-hits and 3-misses treatments ("GVT est."), and the bias-adjusted difference ("bias adj."). The bias adjustment is made by subtracting the expected difference from each player's observed difference, which results in 19 of the 25 players directionally exhibiting hot hand shooting ($p < .01$, binomial test).

The bias-adjusted version of GVT's individual t-tests reveals 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).[34] GVT's study did not attempt to estimate an average hot hand effect, or conduct tests regarding whether or not the average is positive, presumably because beliefs about the hot hand typically pertain to individuals, not groups. In any case, we conduct such a test despite the possibility that sufficient heterogeneity across individuals could in principle mask any evidence of hot hand shooting within certain individuals. We find the across player average (bias adjusted) difference to be 13 percentage points ($p < .01$,

[34] This test is robust to how a streak is defined. If we instead define a streak as beginning with 4 consecutive hits, which is a stronger signal of hot hand shooting 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, four players exhibit statistically significant hot hand shooting ($p < .05$), which is itself significant ($p < .01$, binomial test).

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 field goal 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

$SE = 4.7$pp).[35,36] To put this number 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.[37,38]

Permutation test

As noted in Section 3.1, the idea behind GVT's analysis is to compare players' performance over a sequence of shots to what one would expect from a sequence of Bernoulli trials.[39] A procedure that directly implements this idea, and that is invulnerable to the bias, is a permutation test. The rationale for using a permutation test is the fact that each arrangement of a player's observed sequence of hits and misses is equally likely under the null hypothesis that the player shoots with a constant probability of success. This yields an exact statistical test, with a null distribution that is constructed by computing the value of the statistic of interest for each equally likely arrangement. This approach has the additional advantage of generating a test for any performance pattern that can be expressed as a function of the observed sequence of hits and misses, including several that GVT discussed as indicative of hot hand shooting, but could not test for.[40]

[35] The standard error is computed based on the assumption of independence across players and trials, i.e. $\widehat{Var}(\hat{D}_k^i) = \widehat{Var}(\hat{P}_{1k}^i) + \widehat{Var}(\hat{P}_{0k}^i)$ for each player i. Simulations reveal that the associated $(1-\alpha)\times 100\%$ confidence intervals with radius $z_{\alpha/2} \times \widehat{Var}(\hat{D}_k)^{1/2}$, have the appropriate coverage—i.e. $(1-\alpha/2)\times 100\%$ of the time the true difference is greater than $\hat{D}_k^i - z_{\alpha/2} \times \widehat{Var}(\hat{D}_k)^{1/2}$, for both Bernoulli trials and the positive feedback model discussed in Section E.

[36] An alternative approach involves pooling shots from both treatments into a regression framework, with a coefficient indicating the treatment "3-hits". If the implementation of GVTs 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 ($p < .01$, $SE = 3.7$). 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 framework will control for this aggregation bias, but strengthens the selection bias related to streaks. As a result, a bias adjustment is necessary. In this case, the estimated effect is 13.9 percentage points ($p < .01$, $SE = 5.8$), 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 should be thought of as the hot hand effect for the average shot rather than the average player, which is a different interpretation than the one given for the estimate of the average difference across players. This different interpretation arises because pooling shots across players generates an unbalanced panel and therefore the estimate will place greater weight on players with more shots. In the extreme, it is possible that the majority of players exhibit a tendency to have fewer streaks than expected by chance, yet, because they have relatively few observations, their data becomes diluted by many observations from a single streak shooter.

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

[38] Average estimates are also robust to how a streak is defined. 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 ($p = .07$, $SE = 6.9$, one-sided 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 ($p < .05$, $SE = 3$, one-sided test).

[39] "The player's performance, then, can be compared to a sequence of hits and misses generated by tossing a coin" (Gilovich et al. 1985, p. 296).

[40] When introducing streak shooting, GVT define it in terms of the length of extended runs of hits, and their frequency, relative to what one would expect from a coin: "Consider a professional basketball player who makes 50% of his

To conduct the permutation test, upon observing a sequence of n_1 hits and n_0 misses, first the difference in proportions, $\hat{D}_k := \hat{p}(hit|k\ hits) - \hat{p}(hit|k\ misses)$ is computed. Next, for each equally likely *rearrangement* of the original sequence, the difference D_k is computed. This yields the distribution of values D_k across the unique rearrangements of the original sequence, which is equal to the exact sampling distribution of the difference under the null hypothesis of consistent shooting. The distribution is negative-skewed, and can be represented by histograms such as the one presented in Figure 3. Finally, this null-distribution can be used to formally test whether the observed difference of the shooter constitutes significant evidence of streak shooting.[41,42]

Using the permutation test to analyze GVT's shooting data yields results that agree with those of the bias-adjusted t-tests reported above. In particular, 5 players exhibit significant hot hand shooting ($p < .01$, binomial test).[43] Further, a Monte-Carlo re-sampling procedure permits one to stratify the permutation by player and allows for a direct statistical test of the average difference in proportions across shooters. As with the bias adjusted t-test, the result of this test indicates hot hand shooting with a similar level of significance ($p < .01$).[44]

---

shots. This player will occasionally hit four or more shots in a row. Such runs can be properly called streak shooting, however, only if their length or frequency exceeds what is expected on the basis of chance alone" (Gilovich et al. 1985, p. 296). GVT do not conduct statistical tests for these patterns of hit streaks, presumably because the distributions for the associated measures are not well-approximated parametrically.

[41] More precisely, let $\boldsymbol{x} \in [0,1]^n$ be a sequence of shot outcomes for which $D_k(\boldsymbol{x})$ is defined. The hot hand hypothesis predicts that $D_k(\boldsymbol{x})$ will be significantly larger than what one would expect by chance. To test the null hypothesis at the α level, with n_1 hits in n trials, one simply checks if $D_k(\boldsymbol{x}) \geq c_{\alpha,n_1}$, where the critical value c_{α,n_1} is defined as the smallest c such that $\mathbb{P}(D_k(\boldsymbol{X}) \geq c \mid H_0,\ \sum_{i=1}^n X_i = n_1) \leq \alpha$, and the distribution $\mathbb{P}(D_k(\boldsymbol{X}) \geq c \mid H_0,\ \sum_{i=1}^n X_i = n_1)$ is generated using the enumeration provided in Theorem 7 of Appendix C.2. For the quantity $\mathbb{P}(D_k(\boldsymbol{X}) \geq c \mid H_0,\ \sum_{i=1}^n X_i = n_1)$ 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 $\sum_{i=1}^n x_i = n_1$ 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 $\mathbb{P}(\text{reject}|H_0) = \sum_{n_1=1}^n \mathbb{P}(D_k(\boldsymbol{X}) \geq c_{\alpha,n_1}|H_0,\ \sum_{i=1}^n X_i = n_1)\mathbb{P}(\sum_{i=1}^n X_i = n_1|H_0) \leq \alpha$. Lastly, for any arbitrary test statistic $T(\boldsymbol{x})$, the fact that the distribution of $(\boldsymbol{X}|H_0, \sum_{i=1}^n X_i = n_1)$ is *exchangeable* means that $\mathbb{P}(T(\boldsymbol{X}) \geq c \mid H_0,\ \sum_{i=1}^n X_i = n_1)$ can be approximated to appropriate precision with Monte-Carlo permutations of the sequence $\boldsymbol{x}$.

[42] Miller and Sanjurjo (2014) use this hypothesis test procedure, and propose three test statistics, along with a composite statistic, that they show to have greater statistical power than previous measures. These statistics include measures of streak shooting which GVT discuss, but do not test for: the length of the longest run of hits and the frequency of extended runs of hits (see footnote 40). Using these statistics, Miller and Sanjurjo (2014) find significant and substantial evidence of the hot hand across all extant controlled shooting datasets.

[43] As in footnote 34, the results of the permutation test are robust to varying streak length k.

[44] Specifically, 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, we have H_0: $\mathbb{P}(\text{success on trial } t \text{ for player } i) = p_i$ for all t, i. Two observations about this test are in order: (1) While the hot hand hypothesis predicts that the average difference will be larger than expected, a failure to reject would not amount to evidence against the hot hand. That is, it is entirely possible for streak shooting to exist for some players, but for it to be canceled out in the average due to the existence of players exhibiting anti-streaky shooting, (2) While a rejection is evidence of hot hand shooting, the test cannot distinguish between a few players exhibiting hot hand shooting (with most of the remaining players having no effect) and most players exhibiting hot hand shooting. GVT's dataset is not sufficiently powered to test for heterogeneity in effect size.

3.3 Evidence of the hot hand in other controlled and semi-controlled studies

While we have shown that there is strong evidence of hot hand shooting in GVT's seminal study, it is not the only controlled shooting study to test for the hot hand. In particular, there are four other controlled (or semi-controlled) basketball shooting studies: Jagacinski et al. (1979), Koehler and Conley (2003), Avugos et al. (2013a) and Miller and Sanjurjo (2014). We obtain the raw data for all of these studies but Avugos et al. (2013a).[45,46] It is important that we analyze the raw data, as many of the tests in the original studies exhibit the same methodological problems that are found in GVT. This in turn means that any meta-analysis that involves the previous results is invalid.[47]

Among these studies, the most prominent is Koehler and Conley (2003)'s analysis of four years of data collected from the NBA's Three Point shooting contest, which has been described as "an ideal situation in which to study the hot hand" (Thaler and Sunstein 2008). We observe that Koehler and Conley (2003)'s study is severely underpowered at just 49 shots per-player (median), and applies a testing procedure that has the same bias found in GVT. Miller and Sanjurjo (2015b) revisit this setting, and increase power by collecting an additional 24 years of data. Upon applying the de-biased test procedure from Section 3.2 we find that players' exhibit an estimated 8 percentage point average increase in field goal percentage following three consecutive hits (relative to three consecutive misses), which is substantial and highly significant ($p < .01$).

The studies of Jagacinski et al. (1979) and Miller and Sanjurjo (2014), on the other hand, allow for testing on the individual shooter level. This is important because it is commonly believed that some players are "streak shooters" while others are not, and that the heterogeneity in this characteristic is decision relevant (Gilovich et al. 1985). While early analyses, including GVT, do conduct individual-level tests, they do not have sufficient statistical power to identify individual streak shooters (Gilovich et al. 1985; Larkey, Smith, and Kadane 1989; Tversky and Gilovich 1989b).[48] Similarly, the NBA's Three Point Contest does not provide sufficiently many shots

[45] For Avugos et al. (2013a), the authors declined to make their data available to us. Avugos et al. (2013a) is a close replication of GVT, with olympian players. Because they used the same conditional probability test as GVT, but had fewer shots per session (40), the bias is particularly severe (-20 percentage points).

[46] There exists a controlled shooting study involving 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 be included in our analysis.

[47] See Avugos et al. (2013b) for a meta-analysis of the hot hand, which includes sports besides basketball. Tversky and Gilovich (1989a) 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 the hot hand by fans, players and coaches must necessarily be a cognitive illusion (also see Alter and Oppenheimer (2006)). While we do not study other sports, we do not find this argument convincing. In our view other sports should shed light on the existence of the hot hand phenomenon in basketball as the known relationships between confidence and performance (Bandura 1982), and the influence of increased focus, attention, and motor control on performance (Churchland, Afshar, and Shenoy 2006; Csikszentmihalyi 1988; Kahneman 1973) should apply to all sports.

[48] Gilovich et al. (1985), noted that with just nine players on the 76ers, they may not have included a "real streak shooter" in their sample. They conducted an informal poll having fans list "streak shooters." Andrew Toney, who was included in their 76ers sample, was generally regarded as a streak shooter, but GVT found no evidence to support

per player (median 148) for an analysis based on the data from a single shooter.[49] By contrast, Jagacinski et al. (1979), which has previously gone uncited in the literature, and Miller and Sanjurjo (2014), both contain a small sample of players, but have sufficiently many shots per player to detect the hot hand on the individual player level.[50] Upon applying the de-biased testing procedure from Section 3.2 to this data we find substantial and persistent evidence of hot hand shooting in individual players. Further, the average effect sizes across shooters are 7 percentage points (6 shooters), and 4 percentage points (8 shooters), respectively.[51]

Finally, it is important to observe that the magnitudes of these 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 E), 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.

3.4 The Belief in the Hot Hand

The combination of GVT's evidence that: (1) players' believe in the hot hand, and (2) the hot hand does not exist in basketball shooting, led them to the stark conclusion that belief in the hot hand is a cognitive illusion (Gilovich et al. 1985; Tversky and Gilovich 1989a). By contrast, the result of the present analysis, which uses the same data, leads to the opposite conclusion: 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

this belief. In an early response to GVT, the statisticians Larkey, Smith, and Kadane (1989) analyzed shooting data from Vinnie "The Microwave" Johnson, a player widely believed to have the tendency to get the hot hand, and found evidence to support of this belief. However, Tversky and Gilovich (1989b) subsequently found a coding error in Larkey et. al.'s data, which ended the debate. Neither of these analyses could identify a streak shooter as they do not control for the strategic adjustments of the opposing defense.

[49] On the other hand, the NBA's Three Point Data is suitably powered to detect a subset of shooters with the hot hand. In particular, Miller and Sanjurjo (2015b) find significant evidence of substantial hot hand shooting ($p < .05$) for 8 out of the 33 participants in the history of the NBA's Three Point contest (that have taken at least 100 shots), which is itself statistically significant ($p < .001$, binomial test).

[50] We thank Tom Gilovich for bringing the study of Jagacinski et al. to our attention

[51] In Jagacinski et al. (1979), each of six players participated in 9 sessions of 60 shots each. Even before adjusting for the bias, the difference $\hat{D}_3$ is significant ($p < .05$) for two out of eight players, which is itself significant ($p < .05$, Binomial probability). The average $\hat{D}_3$ is 7 percentage points across the six players. In Miller and Sanjurjo (2014) eight players participated in multiple sessions spaced across 6 months. One of the eight players exhibited consistent evidence of hot hand shooting across sessions, with an 11 percentage point difference ($\hat{D}_3$). Moreover, his tendency to get "hot" was predicted out-of-sample in a survey of teammate opinion (and based on a prior shooting study).

respond too strongly to it.[52,53] An understanding of the extent to which decision makers' beliefs and behavior do not correspond to the actual degree of hot hand shooting may have important 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, their study included a survey of player and coach beliefs. The survey itself is predominantly qualitative, and focuses on questions that relate to whether players believe that they tend to perform better after recent success, and whether they make decisions based on these beliefs.[54] In contrast with the original conclusion, and in light of the present findings, we observe that the *qualitative* beliefs of players (and fans) reported in GVT are actually consistent with the evidence of hot hand shooting that results from an un-biased analysis of the data.

GVT also administered quantitative survey questions to fans, which indicate that fans' beliefs are not well-calibrated, e.g. fans estimate that a "hypothetical" 50 percent field goal shooter will hit 61 percent of his field goals after hitting one and 42 percent after missing one (Gilovich et al. 1985).[55] Taken at face value, fans' beliefs are almost certainly too strong. On the other hand, these unincentivized survey measures of fan beliefs about hypothetical players have limitations. Most importantly, the questions elicit stated beliefs about a shooter's numerical field goal percentage when shooting immediately after a streak of hits, which aside from being potentially unnatural from the perspective of the responder, may be different from the beliefs that are operational in decision making.[56] These concerns are not merely theoretical, as a relatively recent study finds survey measures and decision-based measures to differ. In particular, Rao (2009b)'s decision-based measure consisted of an incentivized prediction task that was designed to study the influence of prior outcomes on subsequent predictions. He finds that the same people who, when surveyed, express a belief in the hot hand for a player that hits a single shot, do not reveal this belief when

[52] Of course, it is also possible that a hot hand goes undetected, or that a detected hot hand is underestimated.

[53] For instance, there is anecdotal evidence that NBA players believe that a shooter with the hot hand can sometimes disrupt his team's offensive flow (Blakely 2016, April 26) [link]

[54] In Gilovich et al. (1985) there is a five question survey of 76ers players: (1) six out of eight players reported that they have on occasion felt that after having made a few shots in a row they "know" they are going to make their next shot, i.e they "almost can't miss." (2) Five out of seven players believed that a player "has a better chance of making a shot after having just made his last two or three shots than he does after having just missed his last two or three shots." (3) Seven of the eight players reported that after having made a series of shots in a row, they "tend to take more shots than they normally would." (4) All of the players believed that it is important "for the players on a team to pass the ball to someone who has just made several (two, three, or four) shots in a row." (5) Five players and the coach also made numerical estimates, which are further discussed just below in the text. Five of these six respondents estimated their field goal percentage for shots taken after a hit (mean: 62.5%) to be higher than their percentage for shots taken after a miss (mean: 49.5%).

[55] Fans also estimate that a hypothetical 70 percent free throw shooter will hit 74 percent of his free throws after hitting one and 66 percent after missing one.

[56] Other limitations include (1) because the shooter in the survey is hypothetical, this does not control for fan background beliefs regarding the player's tendency to get the hot hand, and (2) the fact that the questions are unincentivized may lead to systematic response bias due to demand effects, or anchoring.

asked to make a prediction; instead, they increase their propensity to predict miss after a player has hit a single shot, with hot hand beliefs emerging only for longer streaks of hits.[57]

The above discussion suggests that peoples' operational beliefs may be better elicited by (incentivized) prediction tasks than by survey responses. Accordingly, when conducting their controlled shooting task GVT simultaneously conducted a paired betting task. In the betting task both the shooter and an observer bet on the outcome of each upcoming shot by either placing a "high" or "low" stakes bet on hit. Under reasonable assumptions, a "high" bet can be viewed as a hit prediction and a "low" bet can be viewed as a 'miss" prediction.[58] GVT found that while observer predictions are positively correlated with the outcome of the previous shot (.42 on average), their predictions are relatively uncorrelated with the outcome of the shot that is bet on (.04 on average). Further, GVT do not find significant evidence of any exemplary individual predictors. This seems to suggest that players are generally unable to predict the outcome of shots, which would in turn imply that they cannot detect the hot hand.[59]

However, what the above reasoning fails to account for is that the hot hand is, by definition, an infrequent phenomenon. This means that if a predictor does have the ability to detect the instances in which a shooter's probability of success shifts, this ability will be obscured by the predictions made in the more numerous situations in which the shooter's probability of success is relatively stable. To illustrate, imagine a hypothetical shooter who ordinarily hits with probability .45, but becomes hot on 15 percent of his shots, in which case he hits with probability .55. If a predictor were to *perfectly* detect the shooter's hot hand whenever it occurs, then the expected correlation between predictions and shots would be .07. Observe that this is not vastly different than the average correlation of .04 that GVT estimated. This suggests that GVT's betting data may contain evidence of players successfully predicting outcomes at rates better than chance would predict, but that it has gone undetected because GVT conducted underpowered individual predictor level tests, rather than pooling the data together.

As such, we perform a re-analysis of GVT's betting task using the available data, which corresponds to 22 of the 26 shooters (44 sequences of predictions in total from observers and shooters).[60]

[57] There may be a difference between how fans predict and how players predict, as this contrast between stated beliefs (survey) and operational beliefs (decisions) has not been found with players in other data sets. Rao (2009a) finds that players are more likely to take more difficult shots, less likely to pass, and more likely to attempt the next shot if they made the previous shot. Similar results have been found in subsequent studies (Aharoni and Sarig 2011; Attali 2013; Bocskocsky et al. 2014; Cao 2011; Neiman and Loewenstein 2011). Given the strategic confounds present in live-ball data, it is not obvious that these responses are systematically wrong.

[58] The bettors vary their stakes between the "high" condition, in which they earn +\$0.05 for a hit and −\$0.04 for a miss, and the "low" condition, in which they earn +\$0.02 for a hit and −\$0.01 for a miss. A risk neutral bettor should bet "high" if she believes that the probability of a hit is greater than .5, and "low" otherwise.

[59] This positive correlation with previous shots contrasts with Rao (2009b)'s finding that subjects predict hit less often after a single hit, than after a single hit. On the other hand, inline with Rao's findings, the bettors are significantly more likely bet on the continuation of the same outcome as the streak length increases.

[60] We thank Tom Gilovich for providing us with this data. We were informed that the data for the remaining four

We find the average correlation between predictions and shot outcomes to be $\hat{\rho} = .07$ across predictors, which is highly significant ($p < .001$, permutation test with predictor/shooter stratification).[61] This correlation is large, which can be illustrated by noting the correlation is approximately equal to the (average) percentage point increase in a player's field goal percentage whenever a predictor predicts a hit (relative to predicting a miss).[62] In fact, the actual (average) increase is 7.7 percentage points ($p < .001$, $SE = 1.8$), which is comparable to an NBA shooter going from a slightly above average three point percentage to a super-star one.[63,64]

Our finding that shooters and their teammates have the ability to successfully predict shot outcomes is novel, and contrasts sharply with GVT's original findings. This ability is *consistent* with the possibility that observers (and the shooters themselves) are detecting the hot hand as it occurs, and exploiting it. On the other hand it is also possible that players merely have a default belief in the hot hand that just happens to successfully "predict" shot outcomes in GVT's betting task because the hot hand exists in GVT's data.[65]

Assessing the ability of players, coaches, and fans to detect the hot hand, and respond to it appropriately, is a challenging problem. Nevertheless, while the evidence that we have presented is not definitive, it is does indicate that decision makers can exploit the hot hand, either by detecting it in individual shooters as it occurs or by applying generally correct hot-hand heuristics. Further, Miller and Sanjurjo (2014) present complementary evidence that decision makers are also able to identify which shooters have more (less) of a tendency to get the hot hand. In particular, semi-professional players' rankings of their teammates' respective increases in field goal percentage when on a streak of three hits are found to be highly correlated with the actual increase in performance, yielding an average correlation of -0.60 ($p < .0001$; where 1 is the rank of the shooter with the perceived largest percentage increase).[66] Given this result, a natural question is how players come

shooters could not be located. For the 44 sequences of predictions and outcomes the data was entered by two independent coders, and cross-checked.

[61] In the case of shooters predicting their own shots the average correlation is $\hat{\rho} = .07$ ($p < .01$). In the case of the predictions of an observer the average correlation is $\hat{\rho} = .066$ ($p < .01$).

[62] The average correlation is close to the average OLS regression coefficient $\hat{\beta} \equiv \hat{p}(\text{hit}|\text{predict hit}) - \hat{p}(\text{hit}|\text{predict miss})$, because the variance in bets is close to the variance in hits for most bettors.

[63] The standard errors are computed exactly as in footnote 35. In the case that shooters predict their own shots, they are 8.7pp more likely to hit a shot after a prediction of hit ($p < .001$, $SE = 2.7$). In the case that predictions are made by an observer, shooters are 6.6pp more likely to hit a shot after a prediction of hit ($p < .01$, $SE = 2.4$).

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

[65] To illustrate how this could occur, suppose that one were to bet "hit" each time a shooter hit three shots in a row, "miss" each time a shooter missed three shots in a row, and otherwise bet hit (miss) if the shooter's overall (unknown) field goal percentage is greater (less) than .5. With this heuristic, as one might anticipate given the results of Section 3.2, on average, shooters would perform significantly better (5.5 percentage points) after a"hit" bet than after a "miss" bet.

[66] We also elicited ratings and numerical estimates. The teammates' ratings (on a scale of -3 to 3) of how much each of their teammates' shooting percentage tends to increase has a 0.49 correlation with the actual increase in performance ($p < .0001$). The teammates' numerical estimates of each shooter's percentage-point change do not exhibit significant

to learn each others' tendencies. Players could, for example, notice over time how successes and failures tend to cluster in other players' aggregate shooting performance. This would be enough to predict such patterns out of sample. However, players may additionally be able to detect the hot hand in real-time, when it emerges as a consequence of a shift in a player's underlying probability of success. Doing so would require information beyond the outcomes of recent shots, as a few observations of binary data is simply too weak of a signal to clearly diagnose a shift in a player's probability of success.[67]

Accordingly, if players learn to associate other cues, such as shooting technique, body language, and eagerness for the ball with subsequent performance, then it may be possible for them to detect the hot state. This suggests the possibility of conducting experiments in which experienced players (or coaches) are incentivized to predict the shot outcomes of players that they are familiar with, but only predict when they feel sufficiently confident about their ability to do so accurately. Such expert predictions could then be compared to similar predictions made by amateurs who are not familiar with the shooters.

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).[68] 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).[69] Gambler's fallacy beliefs

correlation with the ratings, the rankings, or the actual increase in performance. This may be because providing a numerical estimate is less natural and more difficult for players than rankings or ratings. This conjecture is consistent with the corresponding responder attrition rates to these questions that have been observed in both GVT and Miller and Sanjurjo (2014)

[67] To illustrate, suppose that a player's hit rate is .6 in the "hot" state and .4 in the "normal" state, and that the player is in the hot state on 20 percent of his shots. The likelihood of him hitting three in a row is $(.6/.4)^3 \approx 3.38$ times higher when a player is in the hot state. Thus, upon observing three hits in a row, the odds in favor of the player being hot must increase by this factor. Nevertheless, because the prior odds are just 1:4 in favor, the posterior odds become 3.38:4, indicating slightly less than fair odds of detecting a true hot hand.

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

[69] 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

are widespread among novice gamblers, with adherents that have included at least one historically eminent mathematician (D'Alembert 1761, pp. 13-14).[70] 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."

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.[71] 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).

---

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 7).

[70] 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.

[71] 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.

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 $\{X_i\}_{i=1}^{\ell}$ 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.[72] 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.[73,74] Perhaps more plausibly, if DM has

[72] 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 $\left(\frac{\theta}{\theta^*}\right)^{\theta^*}\left(\frac{1-\theta}{1-\theta^*}\right)^{1-\theta^*} < 1$.

[73] 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 72, 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.).

[74] There are two alternative statistical approaches that do not require an infinite sample of sequences for the decision

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).[75] More recently, and closely 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 Appendix D.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).[76] 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

maker to obtain an unbiased estimate of the conditional probability, see footnote 4 for details.

[75] 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 $k > 1$, 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 7).

[76] 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.

(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 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.

——— (2011): "Do Gamblers Correctly Price Momentum in NBA Betting Markets?" *Journal of Prediction Markets*, 5, 31–50.

——— (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.

Attali, Y. (2013): "Perceived Hotness Affects Behavior of Basketball Players and Coaches," *Psychological Science*, forthcoming.

Avery, C. and J. Chevalier (1999): "Identifying Investor Sentiment from Price Paths: The Case of Football Betting," *Journal of Business*, 72, 493–521.

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.

Bandura, A. (1982): "Self-Efficacy Mechanism in Human Agency," *American Psychologist*, 37, 122–147.

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

Barberis, N. and R. Thaler (2003): "A survey of behavioral finance," *Handbook of the Economics of Finance*, 1, 1053–1128.

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.

Blakely, A. S. (2016, April 26): "Millsap's Hot Hand Took Hawks Out Of Their Game Sunday," *CNS New England*, (link).

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.

Brown, W. A. and R. D. Sauer (1993): "Does the Basketball Market Believe in the Hot Hand? Comment," *American Economic Review*, 83, 1377–1386.

Camerer, C. F. (1989): "Does the Basketball Market Believe in the 'Hot Hand,'?" *American Economic Review*, 79, 1257–1261.

Cao, Z. (2011): "Essays on Behavioral Economics," Ph.D. thesis, Oregon State University.

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.

Churchland, M. M., A. Afshar, and K. V. Shenoy (2006): "A Central Source of Movement Variability," *Neuron*, 52, 1085–1096.

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.

Csikszentmihalyi, M. (1988): "The flow experience and its significance for human psychology," in *Optimal experience: Psychological studies of flow in consciousness*, ed. by M. Csikszentmihalyi and I. S. Csikszentmihalyi, New York, NY, US: Cambridge University, chap. 2.

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

Davidson, A. (2013, May 2): "Boom, Bust or What?" *New York Times*, MM28, (link).

De Bondt, W. P. (1993): "Betting on trends: Intuitive forecasts of financial risk and return," *International Journal of Forecasting*, 9, 355–371.

De Long, J. B., A. Shleifer, L. H. Summers, and R. J. Waldmann (1991): "The Survival of Noise Traders In Financial-markets," *Journal of Business*, 64, 1–19.

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

Durham, G. R., M. G. Hertzel, and J. S. Martin (2005): "The Market Impact of Trends and Sequences in Performance: New Evidence," *Journal of Finance*, 60, 2551–2569.

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

Galbo-Jørgensen, C. B., S. Suetens, and J.-R. Tyran (2015): "Predicting Lotto Numbers A natural experiment on the gamblers fallacy and the hot hand fallacy," *Journal of the European Economic Association*, forthcoming, working Paper.

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.

Gould, S. J. (1989): "The streak of streaks," *Chance*, 2, 10–16.

Green, B. S. and J. Zwiebel (2013): "The Hot Hand Fallacy: Cognitive Mistakes or Equilibrium Adjustments?" 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.

Guryan, J. and M. S. Kearney (2008): "Gambling at Lucky Stores: Empirical Evidence from State Lottery Sales," *American Economic Review*, 98, 458–473.

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. (1973): *Attention and Effort.*, Prentice Hall.

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

Kahneman, D. and M. W. Riepe (1998): "Aspects of Investor Psychology: Beliefs, preferences, and biases investment advisors should know about," *Journal of Portfolio Management*, 24, 1–21.

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.

Larkey, P. D., R. A. Smith, and J. B. Kadane (1989): "Its Okay to Believe in the Hot Hand," *Chance*, 2, 22–30.

Lee, M. and G. Smith (2002): "Regression to the mean and football wagers," *Journal of Behavioral Decision Making*, 15, 329–342.

Loh, R. K. and M. Warachka (2012): "Streaks in Earnings Surprises and the Cross-Section of Stock Returns," *Management Science*, 58, 1305–1321.

Malkiel, B. G. (2011): *A random walk down Wall Street: the time-tested strategy for sucessful investing*, New York: W. W. Norton & Company.

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.

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

Narayanan, S. and P. Manchanda (2012): "An empirical analysis of individual level casino gambling behavior," *Quantitative Marketing and Economics*, 10, 27–62.

Neiman, T. and Y. Loewenstein (2011): "Reinforcement learning in professional basketball players," *Nature Communications*, 2:569.

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.

Paul, R. J. and A. P. Weinbach (2005): "Bettor Misperceptions in the NBA: The Overbetting of Large Favorites and the 'Hot Hand'," *Journal of Sports Economics*, 6, 390–400.

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.

Sackett, D. L. (1979): "Bias in analytic research," *Journal of Chronic Diseases*, 32, 51–63.

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.

Sinkey, M. and T. Logan (2013): "Does the Hot Hand Drive the Market?" *Eastern Economic Journal*, Advance online publication, doi:10.1057/eej.2013.33.

Smith, G., M. Levere, and R. Kurtzman (2009): "Poker Player Behavior After Big Wins and Big Losses," *Management Science*, 55, 1547–1555.

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

Sundali, J. and R. Croson (2006): "Biases in casino betting: The hot and the gamblers fallacy," *Judgement and Decision Making*, 1, 1–12.

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

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

——— (1989b): "The"Hot Hand": Statistical Reality or Cognitive Illusion?" *Chance*, 2, 31–34.

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," Working paper, University of Wisconsin - Madison.

Xu, J. and N. Harvey (2014): "Carry on winning: The gambler's fallacy creates hot hand effects in online gambling," *Cognition*, 131, 173 – 180.

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.

Yuan, J., G.-Z. Sun, and R. Siu (2014): "The Lure of Illusory Luck: How Much Are People Willing to Pay for Random Shocks," *Journal of Economic Behavior & Organization*, forthcoming.

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 2.1)

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

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 $\tau|\boldsymbol{X}$ be a uniform draw from the trials in sequence $\boldsymbol{X}$ that immediately follow k consecutive successes, i.e. $\mathbb{P}(\tau = t|\boldsymbol{X} = \boldsymbol{x}) = 1/|I_{1k}(\boldsymbol{x})|$ for $t \in I_{1k}(\boldsymbol{x})$, and $\mathbb{P}(\tau = t|\boldsymbol{X} = \boldsymbol{x}) = 0$ for $t \in I_{1k}(\boldsymbol{x})^C \cap \{k+1,\ldots,n\}$. It follows that the unconditional probability distribution of τ over all trials that can possibly follow k consecutive successes is given by $\mathbb{P}(\tau = t|F) = \sum_{\boldsymbol{x}\in F} \mathbb{P}(\tau = t|\boldsymbol{X} = \boldsymbol{x})\mathbb{P}(\boldsymbol{X} = \boldsymbol{x}|F)$, 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 of successes in the set of trials $I_{1k}(\boldsymbol{x})$. The intuition why is explained in Section 2, just below Theorem 1. Here, for completeness, we provide the following formal derivation.

$$
\begin{aligned}
&E\left[\, \hat{P}_{1k}(\boldsymbol{X}) \;\middle|\; I_{1k}(\boldsymbol{x}) \neq \emptyset \,\right] \\
&\qquad = \sum_{\boldsymbol{x}\in F} \hat{P}_{1k}(\boldsymbol{X})\mathbb{P}(\boldsymbol{X} = \boldsymbol{x} \mid I_{1k}(\boldsymbol{x}) \neq \emptyset) \\
&\qquad = \sum_{\boldsymbol{x}\in F} \sum_{t\in I_{1k}(\boldsymbol{x})} x_t \cdot \frac{1}{|I_{1k}(\boldsymbol{x})|}\mathbb{P}(\boldsymbol{X} = \boldsymbol{x}|F) \\
&\qquad = \sum_{\boldsymbol{x}\in F} \sum_{t=k+1}^{n} \mathbb{P}(X_t = 1|\tau = t, \boldsymbol{X} = \boldsymbol{x})\mathbb{P}(\tau = t|\boldsymbol{X} = \boldsymbol{x})\mathbb{P}(\boldsymbol{X} = \boldsymbol{x}|F) \\
&\qquad = \sum_{\boldsymbol{x}\in F} \sum_{t=k+1}^{n} \mathbb{P}(X_t = 1|\tau = t, \boldsymbol{X} = \boldsymbol{x}, F)\mathbb{P}(\tau = t|\boldsymbol{X} = \boldsymbol{x}, F)\mathbb{P}(\boldsymbol{X} = \boldsymbol{x}|F) \\
&\qquad = \sum_{\boldsymbol{x}\in F} \sum_{t=k+1}^{n} \mathbb{P}(X_t = 1, \tau = t, \boldsymbol{X} = \boldsymbol{x}|F) \\
&\qquad = \mathbb{P}(X_\tau = 1|F)
\end{aligned}
$$

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 $\mathbb{P}(X_\tau = 1|F) < p$.

Suppose that $t < n$. To demonstrate that $\mathbb{P}(X_t = 1|\tau = t, F) < p$ it suffices to show that $\mathbb{P}(\tau = t|X_t = 0, F) > \mathbb{P}(\tau = t|X_t = 1, F)$. This inequality is established next by ap-

plying the law of total probability, conditioning (within F) on values of the random variable $\boldsymbol{X}_{-t}|\prod_{i=t-k}^{t-1} X_i = 1$, whose realizations, $\boldsymbol{x}_{-t} \in \{0,1\}^{n-1}$ (with $\prod_{i=t-k}^{t-1} x_i = 1$), must satisfy $(\boldsymbol{x}_{-t}, x_t) := (x_1, \ldots, x_{t-1}, x_t, x_{t+1}, \ldots, x_n) \in F$ for $x_t \in \{0,1\}$.

$$
\begin{aligned}
&\mathbb{P}(\tau = t | X_t = 0, F) \\
&= \sum_{\substack{\boldsymbol{x}_{-t} \in \{0,1\}^{n-1}: \\ \prod_{i=t-k}^{t-1} x_i = 1}} \mathbb{P}\left(\tau = t \,|X_t = 0, \boldsymbol{X}_{-t} = \boldsymbol{x}_{-t}, F\right) \mathbb{P}\left(\boldsymbol{X}_{-t} = \boldsymbol{x}_{-t} \,\middle|\, X_t = 0, \textstyle\prod_{j=i-k}^{t-1} X_j = 1, F\right) \\
&= \sum_{\substack{\boldsymbol{x}_{-t} \in \{0,1\}^{n-1}: \\ \prod_{i=t-k}^{t-1} x_i = 1}} \mathbb{P}\left(\tau = t \,|X_t = 0, \boldsymbol{X}_{-t} = \boldsymbol{x}_{-t}\right) \mathbb{P}\left(\boldsymbol{X}_{-t} = \boldsymbol{x}_{-t} \,\middle|\, X_t = 0, \textstyle\prod_{j=i-k}^{t-1} X_j = 1\right) \qquad (5) \\
&= \sum_{\substack{\boldsymbol{x}_{-t} \in \{0,1\}^{n-1}: \\ \prod_{i=t-k}^{t-1} x_i = 1}} \mathbb{P}\left(\tau = t \,|X_t = 0, \boldsymbol{X}_{-t} = \boldsymbol{x}_{-t}\right) \mathbb{P}\left(\boldsymbol{X}_{-t} = \boldsymbol{x}_{-t} \,\middle|\, \textstyle\prod_{j=i-k}^{t-1} X_j = 1\right) \qquad (6) \\
&> \sum_{\substack{\boldsymbol{x}_{-t} \in \{0,1\}^{n-1}: \\ \prod_{i=t-k}^{t-1} x_i = 1}} \mathbb{P}\left(\tau = t \,|X_t = 1, \boldsymbol{X}_{-t} = \boldsymbol{x}_{-t}\right) \mathbb{P}\left(\boldsymbol{X}_{-t} = \boldsymbol{x}_{-t} \,\middle|\, \textstyle\prod_{j=i-k}^{t-1} X_j = 1\right) \qquad (7) \\
&= \sum_{\substack{\boldsymbol{x}_{-t} \in \{0,1\}^{n-1}: \\ \prod_{i=t-k}^{t-1} x_i = 1}} \mathbb{P}\left(\tau = t \,|X_t = 1, \boldsymbol{X}_{-t} = \boldsymbol{x}_{-t}, F\right) \mathbb{P}\left(\boldsymbol{X}_{-t} = \boldsymbol{x}_{-t} \,\middle|\, X_t = 1, \textstyle\prod_{j=i-k}^{t-1} X_j = 1, F\right) \\
&= \mathbb{P}(\tau = t | X_t = 1, F)
\end{aligned}
$$

In (5), equality follows for the first term in the sum because for all $\boldsymbol{x}_{-t} \in \{0,1\}^{n-1}$ with $\prod_{i=t-k}^{t-1} x_i = 1$, we have that $\{\boldsymbol{x'} \in \{0,1\}^n : \boldsymbol{x'}_{-t} = \boldsymbol{x}_{-t}\} \subset F$, and for the second term in the sum because for $t \geq k+1$, we have that $\{\boldsymbol{x} \in \{0,1\}^n : \prod_{j=i-k}^{t-1} x_i = 1\} \subset F$. In (6), equality follows for the second term in the sum because Bernoulli trials are independent.[77] In (7), the inequality follows because for each common set of trial outcomes $\boldsymbol{x}_{-t}$ with $\prod_{i=t-k}^{t-1} x_i = 1$, the likelihood that the randomly drawn trial $\tau|\boldsymbol{X} = \boldsymbol{x}$ is equal to t will be lower when $x_t = 1$ than when $x_t = 0$. This is because when $x_t = 1$ there is at least one more trial to choose from—trial $t+1$—compared to when $x_t = 0$. As a result, the respective set of trials that are (uniformly) drawn from satisfy $|I_{1k}(\boldsymbol{x}_{-t}, 1)| \geq |I_{1k}(\boldsymbol{x}_{-t}, 0)| + 1$.

For the case of $t = n$ we can follow the above steps until (6), at which point an equality now emerges between (6) and (7), as $x_t = x_n = 1$ no longer translates into one more trial to choose from, because nothing can follow trial n. This implies that $\mathbb{P}(\tau = n | X_n = 1, F) = \mathbb{P}(\tau = n | X_n = 0, F)$.

Taking these two facts together: (1) $\mathbb{P}(X_t = 1 | \tau = t) < p$, for $k+1 \leq t < n$, and (2)

[77] If we were to condition on $\boldsymbol{X}_{-t}$ rather than $\boldsymbol{X}_{-t}|\prod_{i=t-k}^{t-1} X_i = 1$, then in the second term on the first line the conditioning factors would be $X_t = 0$ and F, and we would not be able to drop $X_t = 0$ to reach a version of (6) in which the second term had just F as a conditioning factor, because $\{\boldsymbol{x} \in \{0,1\}^n : x_t = 0\} \nsubseteq F$.

$\mathbb{P}(X_n = 1|\tau = n) = p$, it immediately follows that $\mathbb{P}(X_\tau = 1|F) < p$.[78]

■

A.2 The mechanism behind the bias in the selection procedure (Section 2.2)

The proof of Theorem 1 in Section A.1 begins with a representative trial τ being drawn at random from the set of trials selected by the researcher, $I_{1k}(\boldsymbol{X})$. Learning that $\tau = t$ provides three pieces of information about the sequence from which to update: (1) $I_{1k}(\boldsymbol{X}) \neq \emptyset$, (2) $t \in I_{1k}(\boldsymbol{X})$, and (3) $\prod_{t-k}^{t-1} X_i = 1$. Letting $M := |I_{1k}(\boldsymbol{X})|$, the posterior odds in favor of $X_t = 1$ (relative to $X_t = 0$) are given by:

$$\frac{\mathbb{P}(X_t = 1|\tau = t)}{\mathbb{P}(X_t = 0|\tau = t)} \tag{8}$$

$$= \frac{\mathbb{P}(M \geq 1, \prod_{t-k}^{t-1} X_i = 1, X_t = 1|\tau = t)}{\mathbb{P}(M \geq 1, \prod_{t-k}^{t-1} X_i = 1, X_t = 0|\tau = t)}$$

$$= \frac{\mathbb{P}(\tau = t|\prod_{t-k}^{t-1} X_i = 1, X_t = 1)}{\mathbb{P}(\tau = t|\prod_{t-k}^{t-1} X_i = 1, X_t = 0)} \frac{\mathbb{P}(M \geq 1|X_t = 1, \prod_{t-k}^{t-1} X_i = 1)}{\mathbb{P}(M \geq 1|X_t = 0, \prod_{t-k}^{t-1} X_i = 1)} \frac{\mathbb{P}(\prod_{t-k}^{t-1} X_i = 1|X_t = 1)}{\mathbb{P}(\prod_{t-k}^{t-1} X_i = 1|X_t = 0)} \frac{\mathbb{P}(X_t = 1)}{\mathbb{P}(X_t = 0)}$$

$$= \frac{\mathbb{P}(\tau = t|\prod_{t-k}^{t-1} X_i = 1, X_t = 1)}{\mathbb{P}(\tau = t|\prod_{t-k}^{t-1} X_i = 1, X_t = 0)} \frac{\mathbb{P}(X_t = 1)}{\mathbb{P}(X_t = 0)}$$

$$= \frac{E\left[\frac{1}{M} \;\middle|\; \prod_{t-k}^{t-1} X_i = 1, X_t = 1\right]}{E\left[\frac{1}{M} \;\middle|\; \prod_{t-k}^{t-1} X_i = 1, X_t = 0\right]} \frac{\mathbb{P}(X_t = 1)}{\mathbb{P}(X_t = 0)} \tag{9}$$

The event $M \geq 1$ in the conditional argument is omitted in the first term of the second equality because it is implied by the event $\prod_{t-k}^{t-1} X_i = 1$. For the Bayes updating factor in the second equality, $\frac{\mathbb{P}(\prod_{t-k}^{t-1} X_i = 1|X_t = 1)}{\mathbb{P}(\prod_{t-k}^{t-1} X_i = 1|X_t = 0)}$, if $\hat{p} = n_1/n$ were known, it would operate like sampling-without-replacement and be equal to $(n_1 - k)/n_1$. However, in the present case, with no prior knowledge of the sequence, the event $\prod_{t-k}^{t-1} X_i = 1$ cannot provide information regarding the remaining $n - k$ outcomes, which implies that the likelihood of this event is constant, regardless of the value of x_t. The updating factor that drives the bias is revealed in the final line, which follows from applying the law of total probability, conditioning on M, and using the fact that $\mathbb{P}(\tau_I = t|M = m) = 1/m$.

To identify whether the updating factor in (9) increases or decreases the odds in favor of $X_t = 1$, we apply the results established within the proof of Theorem 1. In particular, it was found that $\mathbb{P}(X_t = 1|\tau = t) < p$ for $t < n$ and $\mathbb{P}(X_t = 1|\tau = t) = p$ for $t = n$. This implies that the posterior odds in favor of success are strictly less than the prior odds for $t < n$, and equal for $t = n$. Therefore

[78] Note that the proof did 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 $i = 1, \ldots, n$, and our result would be that $\mathbb{P}(X_\tau = 1|F) < E[p_\tau|F]$.

the updating factor satisfies,

$$\frac{E\left[\ \frac{1}{M}\ \middle|\ \prod_{t-k}^{t-1} X_i = 1, X_t = 1\right]}{E\left[\ \frac{1}{M}\ \middle|\ \prod_{t-k}^{t-1} X_i = 1, X_t = 0\right]} < 1 \quad \text{for } t = k+1, \ldots, n-1 \tag{10}$$

with equality for $t = n$. This factor has a straightforward interpretation for $t < n$: for a sequence in which $X_t = 1$, and $\prod_{t-k}^{t-1} X_i = 1$ the likelihood of drawing any given trial from $I_{1k}(\boldsymbol{X})$, including trial t—given by the reciprocal of the number of selected trials $1/M$—is expected to be lower for a sequence in which $X_t = 0$ and $\prod_{t-k}^{t-1} X_i = 1$. In particular, a sequence with $X_t = 1$ and $\prod_{t-k}^{t-1} X_i = 1$ will typically have more trials selected, i.e. larger M, than a sequence with $X_t = 0$ and $\prod_{t-k}^{t-1} X_i = 1$, because the event $X_t = 0$ excludes the next k trials from $t+1$ to $t+k$ from being selected, whereas the event $X_t = 1$ leads trial $t+1$ to be selected, and does not exclude the next $k-1$ trials from being selected.[79] This implies that, in expectation, for the sequences in which there are k consecutive successes from trial $t-k$ to trial $t-1$, those in which the next trial t is a success are given a lower (relative) weight, than the sequences in which the next trial t is a failure, because the sequences in which trial t is a success are (essentially) expected to have more effective observations—i.e. trials that immediately follow k consecutive successes.[80]

A.3 A comparison with sampling-without-replacement (Section 2.2)

Suppose that the researcher knows the overall proportion of successes in the sample, $\hat{p} = n_1/n$. 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 $I_{1k}(\boldsymbol{x}) \subseteq \{k+1, \ldots, n\}$ 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:

[79] 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.

[80] More accurately, they are expected to have a lower value for the reciprocal of the number of effective observations.

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

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 $I_{1k}(\boldsymbol{x})$ is completely determined by $M:=|I_{1k}(\boldsymbol{x})|$, and equal to $1/M$. Upon learning that $\tau_I=t$ one can infer the following three things: (1) $I_{1k}(\boldsymbol{x})\neq\emptyset$, i.e. $M\geq 1$, which is informative if $n_1\leq(k-1)(n-n_1)+k$, (2) t is a member of $I_{1k}(\boldsymbol{x})$, 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:

$$
\begin{aligned}
&\frac{\mathbb{P}(x_t=1|\tau_I=t)}{\mathbb{P}(x_t=0|\tau_I=t)} \\
&\quad= \frac{\mathbb{P}(x_t=1,\prod_{t-k}^{t-1}x_i=1,M\geq 1|\tau_I=t)}{\mathbb{P}(x_t=0,\prod_{t-k}^{t-1}x_i=1,M\geq 1|\tau_I=t)} \\
&\quad= \frac{\mathbb{P}(\tau_I=t|x_t=1,\prod_{t-k}^{t-1}x_i=1)}{\mathbb{P}(\tau_I=t|x_t=0,\prod_{t-k}^{t-1}x_i=1)}\frac{\mathbb{P}(M\geq 1|x_t=1,\prod_{t-k}^{t-1}x_i=1)}{\mathbb{P}(M\geq 1|x_t=0,\prod_{t-k}^{t-1}x_i=1)}\frac{\mathbb{P}(\prod_{t-k}^{t-1}x_i=1|x_t=1)}{\mathbb{P}(\prod_{t-k}^{t-1}x_i=1|x_t=0)}\frac{\mathbb{P}(x_t=1)}{\mathbb{P}(x_t=0)} \\
&\quad= \frac{E\left[\ \frac{1}{M}\ \middle|\ \prod_{t-k}^{t-1}x_i=1,x_t=1\right]}{E\left[\ \frac{1}{M}\ \middle|\ \prod_{t-k}^{t-1}x_i=1,x_t=0\right]}\frac{\mathbb{P}(\prod_{t-k}^{t-1}x_i=1|x_t=1)}{\mathbb{P}(\prod_{t-k}^{t-1}x_i=1|x_t=0)}\frac{\mathbb{P}(x_t=1)}{\mathbb{P}(x_t=0)} \\
&\quad= \frac{E\left[\ \frac{1}{M}\ \middle|\ \prod_{t-k}^{t-1}x_i=1,x_t=1\right]}{E\left[\ \frac{1}{M}\ \middle|\ \prod_{t-k}^{t-1}x_i=1,x_t=0\right]}\frac{n_1-k}{n_1}\frac{n_1}{n_0}
\end{aligned}
\tag{11}
$$

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 poste-

rior 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 $I_{1k}(\boldsymbol{x})$, whereas $x_t = 1$ assures that trial $t+1$ is a member of $I_{1k}(\boldsymbol{x})$, and does not exclude the $\min\{n-t, k\}-1$ trials that follow it from also being in $I_{1k}(\boldsymbol{x})$. In Section A.2 we proved 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 $k = 1$, we prove this result in the discussion of the alternate proof to Lemma 2 in Appendix B.[81]

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 4, the expected empirical probability that a randomly drawn trial in $I_{1k}(\boldsymbol{X})$ is a success, which is the expected proportion, $E[\hat{P}_{1k}(\boldsymbol{X}) \mid I_{1k}(\boldsymbol{X}) \neq \emptyset]$, 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 2.3. Observe that for $k > 1$, and n not too small, the bias in the expected proportion is considerably larger than the corresponding bias from sampling-without-replacement. In the case of the 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 11, is determined by the number of successes in the sequence, but not by their arrangement.[82]

[81] When $k > 1$, the computation is combinatorial in nature, and utilizes the dimension reduction arguments in Appendix C to reach the formula in Lemma 5, which is employed in Theorem 6—note: $1/M = 1/f_{1k}$ 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. All known formulations of this distribution appear to be combinatorial in nature. This is evidenced in a standard reference on runs and scans, Balakrishnan and Koutras (2011, p.188). The authors consider the Type III binomial distribution of order k, represented by the variable $M_{n,k}$, which is the number of (overlapping) instances of k consecutive successes in n trials conditional on a fixed number of successes. The variable M in our case is the number of (overlapping) instances of k consecutive success in the first $n-1$ trials.

[82] 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.

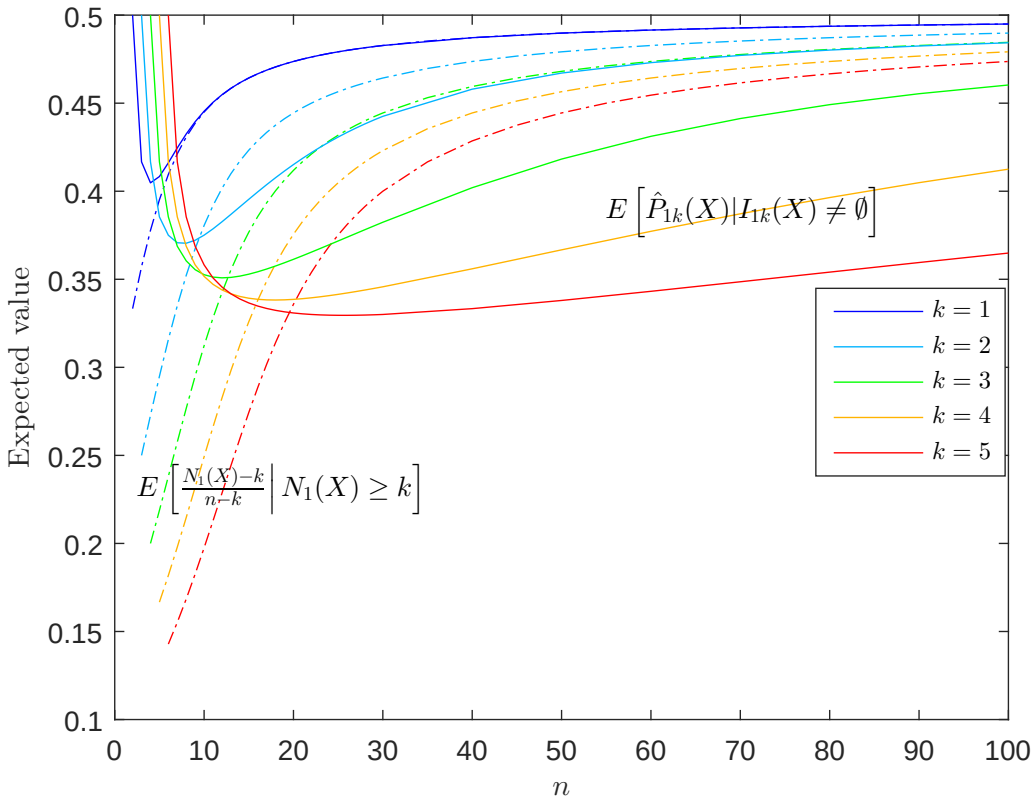

Figure 4: *The dotted lines correspond to the bias from sampling-without-replacement. It is the expected probability of a success, given that k successes are first removed from the sequence (assuming $p = .5$). The solid lines correspond to the expected proportion from Figure 1.*

B Appendix: Proofs for the special case of $k = 1$

The following lemma employs a similar approach to that provided in Appendix C for the general case of $E[\hat{P}_{1k}(\boldsymbol{X}) \mid I_{1k}(\boldsymbol{X}) \neq \emptyset,\ N_1(\boldsymbol{X}) = n_1]$, an alternative, simpler proof, follows immediately below.

Lemma 2 *For $n > 1$ and $n_1 = 1, 2, \ldots, n$*

$$E\left[\ \hat{P}_{11}(\boldsymbol{X})\ \middle|\ I_{11}(\boldsymbol{X}) \neq \emptyset,\ N_1(\boldsymbol{X}) = n_1\ \right] = \frac{n_1 - 1}{n - 1} \tag{12}$$

Proof: For $n_1 = 1$, clearly $\hat{P}_{11}(\boldsymbol{x}) = 0$ for all $\boldsymbol{x}$, and the identity is satisfied. For $n_1 > 1$ this quantity cannot be computed directly by calculating its value for each sequence because the number of admissible sequences is typically too large.[83] In order to handle the case of $n_1 > 1$, we

[83] For example, with $n = 100$, $n_1 = 50$ and $k = 1$ there are $\binom{100}{50} > 10^{29}$ such sequences.

first define $R_1(\boldsymbol{x})$ as the number of runs of ones, i.e. the number of subsequences of consecutive ones in sequence $\boldsymbol{x}$ that are flanked by zeros or an end point.[84] The key observation is that for all sequences with $R_1(\boldsymbol{x}) = r_1$, $\hat{P}_{11}(\boldsymbol{x})$ is (i) constant and equal to $(n_1 - r_1)/n_1$ across all of the sequences that terminate with a zero, and (ii) constant and equal to $(n_1 - r_1)/(n_1 - 1)$ across all of the sequences that terminate with a one. The number of sequences in each of these cases can be counted using a combinatorial argument.

Any sequence with n_1 ones can be constructed, first, by building the runs of ones of fixed length with an ordered partition of the n_1 ones into r_1 cells (runs), which can be performed in $\binom{n_1-1}{r_1-1}$ ways by inserting $r_1 - 1$ dividers into the $n_1 - 1$ available positions between ones, and second, by placing the r_1 runs into the available positions to the left or the right of a zero among the n_0 zeros to form the final sequence. For the case in which $x_n = 0$ there are n_0 available positions to place the runs, thus $\binom{n_0}{r_1}$ possible placements, while in the case in which $x_n = 1$ (which must end in a run of ones) there are n_0 available positions to place the $r_1 - 1$ remaining runs, thus $\binom{n_0}{r_1-1}$ possible placements. Note that for $n_1 > 1$, we have that the proportion is defined for all sequences, and $r_1 \leq n_1$, thus:

$$E[P_{11}|N_1 = n_1] = \frac{1}{\binom{n}{n_1}} \sum_{x_n \in \{0,1\}} \sum_{r_1=1}^{\min\{n_1, n_0+x_n\}} \binom{n_1-1}{r_1-1} \binom{n_0}{r_1 - x_n} \frac{n_1 - r_1}{n_1 - x_n}$$

For the case in which $x_n = 0$, the inner sum satisfies:

$$\begin{aligned}
\sum_{r_1=1}^{\min\{n_1,n_0\}} \binom{n_1-1}{r_1-1}\binom{n_0}{r_1}\frac{n_1-r_1}{n_1} &= \sum_{r_1=1}^{\min\{n_1,n_0\}} \binom{n_1-1}{r_1-1}\binom{n_0}{r_1}\left(1 - \frac{r_1}{n_1}\right) \\
&= \binom{n-1}{n_0-1} - \frac{1}{n_1}\sum_{r_1=1}^{\min\{n_1,n_0\}} \binom{n_1-1}{r_1-1}\binom{n_0}{r_1} r_1 \\
&= \binom{n-1}{n_0-1} - \frac{n_0}{n_1}\sum_{r_1=1}^{\min\{n_1,n_0\}} \binom{n_1-1}{r_1-1}\binom{n_0-1}{r_1-1} \\
&= \binom{n-1}{n_0-1} - \frac{n_0}{n_1}\sum_{x=0}^{\min\{n_1-1,n_0-1\}} \binom{n_1-1}{x}\binom{n_0-1}{x} \\
&= \binom{n-1}{n_0-1} - \frac{n_0}{n_1}\binom{n-2}{n_1-1}
\end{aligned}$$

The left term of the second line follows because it is the total number of sequences that can be formed in the first $n - 1$ positions with $n_0 - 1$ zeros and $n_1 = n - n_0$ ones. The final line follows

[84] The number of runs of ones can be defined explicitly to be the number of trials in which a one occurs and is immediately followed by a zero on the next trial or has no following trial, i.e. $R_1(\boldsymbol{x}) := |\{i \in \{1, \ldots, n\} : x_i = 1 \text{ and, if } i < n \text{ then } x_{i+1} = 0\}|$

from an application of Vandermonde's convolution.[85]

For the case in which $x_n = 1$, the inner sum can be reduced using similar arguments:

$$\begin{aligned}
\sum_{r_1=1}^{\min\{n_1,n_0+1\}} \binom{n_1-1}{r_1-1}\binom{n_0}{r_1-1}\frac{n_1-r_1}{n_1-1} &= \frac{n_1}{n_1-1}\binom{n-1}{n_0} - \frac{1}{n_1-1}\sum_{r_1=1}^{\min\{n_1,n_0+1\}} \binom{n_1-1}{r_1-1}\binom{n_0}{r_1-1} r_1 \\
&= \binom{n-1}{n_0} - \frac{1}{n_1-1}\sum_{r_1=1}^{\min\{n_1,n_0+1\}} \binom{n_1-1}{r_1-1}\binom{n_0}{r_1-1}(r_1-1) \\
&= \binom{n-1}{n_0} - \frac{n_0}{n_1-1}\sum_{r_1=2}^{\min\{n_1,n_0+1\}} \binom{n_1-1}{r_1-1}\binom{n_0-1}{r_1-2} \\
&= \binom{n-1}{n_0} - \frac{n_0}{n_1-1}\sum_{x=0}^{\min\{n_1-2,n_0-1\}} \binom{n_1-1}{x+1}\binom{n_0-1}{x} \\
&= \binom{n-1}{n_0} - \frac{n_0}{n_1-1}\binom{n-2}{n_1-2}
\end{aligned}$$

Combining both cases we have:

$$\begin{aligned}
E[P_{11}|N_1 = n_1] &= \frac{1}{\binom{n}{n_1}}\left[\binom{n-1}{n_0-1} - \frac{n_0}{n_1}\binom{n-2}{n_1-1} + \binom{n-1}{n_0} - \frac{n_0}{n_1-1}\binom{n-2}{n_1-2}\right] \\
&= \frac{1}{\binom{n}{n_1}}\left[\binom{n}{n_0} - \frac{n_0}{n-1}\binom{n}{n_1}\right] \\
&= \frac{n_1-1}{n-1}
\end{aligned}$$

■

Alterative Proof of Lemma 2:

As discussed in Section 2.2, in the case of known $N_1(\boldsymbol{X}) = n_1$, the expected proportion $E[\hat{P}_{11}(\boldsymbol{X})|N_1(\boldsymbol{X}) = n_1, I_{11}(\boldsymbol{X}) \neq \emptyset]$ is equal to the probability of success, $\mathbb{P}(X_\tau = 1|N_1(\boldsymbol{X}) = n_1)$, for a randomly drawn trial $\tau \in I_{11}(\boldsymbol{X})$. We assume hereafter that all probabilities $\mathbb{P}(\cdot)$ are conditional on $N_1(\boldsymbol{X}) = n_1$,

[85] Vandermonde's convolution is given as

$$\sum_{k=\max\{-m,n-s\}}^{\min\{r-m,n\}} \binom{r}{m+k}\binom{s}{n-k} = \binom{r+s}{m+n}$$

from which one can derive the following identity, which we apply

$$\sum_{k=\max\{-m,-n\}}^{\min\{\ell-m,s-n\}} \binom{\ell}{m+k}\binom{s}{n+k} = \sum_{k=\max\{-m,-n\}}^{\min\{\ell-m,s-n\}} \binom{s}{n+k}\binom{\ell}{(\ell-m)-k} = \binom{\ell+s}{\ell-m+n}$$

and because the result is trivial for $n_1 = 1$, we start by considering the case in which $n_1 \geq 2$. Note that $\mathbb{P}(X_\tau = 1) = \sum_{t=2}^{n} \mathbb{P}(X_t = 1|\tau = t)\mathbb{P}(\tau = t)$, and that each posterior is equally likely to be reached, with $\mathbb{P}(\tau = t) = 1/(n-1)$ for all $t \in \{2, 3, \ldots, n\}$.[86] To determine each posterior $\mathbb{P}(X_t = 1|\tau = t)$ for $t = 2, \ldots, n$, we apply Bayes rule

$$\begin{aligned}\mathbb{P}(X_t = 1|\tau = t) &= \frac{\mathbb{P}(\tau = t|X_{t-1} = 1, X_t = 1)\mathbb{P}(X_{t-1} = 1|X_t = 1)\mathbb{P}(X_t = 1)}{\mathbb{P}(\tau = t)} \\ &= \mathbb{P}(\tau = t|X_{t-1} = 1, X_t = 1)\frac{n_1(n_1-1)}{n} \qquad (13)\end{aligned}$$

where $\mathbb{P}(X_{t-1} = 1|X_t = 1) = (n_1 - 1)/(n-1)$ is the sampling-without-replacement likelihood. For the likelihood $\mathbb{P}(\tau = t|X_{t-1} = 1, X_t = 1)$, in the case that $t < n$, it satisfies:

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

where $M := |I_{11}(\boldsymbol{X})|$, with $M = n_1 - X_n$. This shows that the value of the likelihood is independent of t, for $t < n$. In the case that $t = n$, clearly $\mathbb{P}(\tau = n|X_{n-1} = 1, X_n = 1) = \frac{1}{n_1 - 1}$.

The posterior probability for each t follows from substituting the likelihood into (13), which yields (after collecting terms),

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

Summing across ex-ante equally likely trial draws, we have

$$\begin{aligned}\mathbb{P}(X_\tau = 1) &= \mathbb{P}(X_\tau = 1|\tau < n)\mathbb{P}(\tau < n) + \mathbb{P}(X_n = 1|\tau = n)\mathbb{P}(\tau = n) \\ &= \frac{n-1}{n-2}\left(\frac{n_1}{n} - \frac{1}{n-1}\right)\frac{n-2}{n-1} + \frac{n_1}{n}\frac{1}{n-1} \\ &= \frac{n_1 - 1}{n-1}\end{aligned}$$

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

■

Discussion of alternative proof of Lemma 2

As discussed in Section 2.2 the updating factor relating to the arrangement of successes and failures, $\mathbb{P}(\tau = t|X_{t-1} = 1, X_t = 1)/\mathbb{P}(\tau = t|X_{t-1} = 1, X_t = 0)$, is determined by the (expected) reciprocal of the number of effective observations in the sequence. The equation for the likelihood $\mathbb{P}(\tau = t|X_{t-1} = 1, X_t = 1) = \frac{1}{n-2}\left(\frac{n_0}{n_1} + \frac{n_1-2}{n_1-1}\right)$ derived within in the alternative proof to Lemma 2 can be used to demonstrate that this updating factor shrinks the odds beyond sampling-without-replacement for $t < n$ and inflates the odds beyond sampling-without-replacement for $t = n$. First we note that an analogous argument to the one presented in the alternative proof yields the likelihood $\mathbb{P}(\tau = t|X_{t-1} = 1, X_t = 0) = \frac{1}{n-2}\left(\frac{n_0-1}{n_1} + \frac{n_1-1}{n_1-1}\right)$. Further, in the case of $t = n$, it is clear that $\mathbb{P}(\tau = n|X_{n-1} = 1, X_n = 0) = \frac{1}{n_1}$. The likelihood ratio thus becomes (after collecting terms),

$$\frac{\mathbb{P}(\tau = t|X_{t-1} = 1, X_t = 1)}{\mathbb{P}(\tau = t|X_{t-1} = 1, X_t = 0)} = \begin{cases} 1 - \frac{1}{(n-1)(n_1-1)} & \text{for } t = 2, \ldots, n-1 \\ \frac{n_1}{n_1-1} & \text{for } t = n \end{cases} \tag{14}$$

which is clearly strictly less than 1 for $t < n$, and strictly greater than 1 for $t = n$. Because the above likelihood ratio is independent of t for $2 \leq t \leq n-1$, the randomly drawn trial that determines the probability of interest $\mathbb{P}(X_\tau = 1)$, has a $(n-2)/(n-1)$ chance of leading to an updating of the odds that is stronger than sampling-without-replacement would suggest, and a $1/(n-1)$ chance of leading to an updating of the odds that is weaker than sampling-without-replacement would suggest. As shown above, in this special case with $k = 1$, these countervailing forces precisely balance.

Formulae for expected value of the proportion (and the difference in proportions)

Lemma 2 can be combined with Equation 4 to express the expected proportion in terms of just n and p.[87]

Theorem 3 *For $p > 0$*

$$E\left[\, \hat{P}_{11}(\boldsymbol{X}) \;\middle|\; I_{1k}(\boldsymbol{X}) \neq \emptyset \,\right] = \frac{\left[p - \frac{1-(1-p)^n}{n}\right]\frac{n}{n-1}}{1-(1-p)^{n-1}} < p \tag{15}$$

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

Proof: First we observe that in light of Lemma 2, Equation 15 can be written as follows:

$$\begin{aligned} E\left[\ \hat{P}_{11}(\boldsymbol{X})\ \middle|\ I_{11}(\boldsymbol{X})\neq\emptyset\ \right] &= E\left[E\left[\ \hat{P}_{1k}(\boldsymbol{X})\ \middle|\ I_{11}(\boldsymbol{X})\neq\emptyset,\ N_1(\boldsymbol{X})=n_1\ \right]\right] \\ &= E\left[\frac{N_1(\boldsymbol{x})-1}{n-1}\middle| I_{11}(\boldsymbol{X})\neq\emptyset\ \right] \end{aligned}$$

Let $U_{11}(n,n_1) := |\{\boldsymbol{x}\in\{0,1\}^n : N_1(\boldsymbol{x})=n_1\ \&\ I_{11}(\boldsymbol{X})=\emptyset\}|$, i.e. the number of sequences for which $\hat{P}_{11}(\boldsymbol{x})$ is undefined, and C be the constant that normalizes the total probability to 1.[88] The expected value can then be computed using the binomial distribution, which yields:

$$\begin{aligned} E\left[\frac{N_1(\boldsymbol{x})-1}{n-1}\middle| I_{11}(\boldsymbol{X})\neq\emptyset\ \right] &= C\sum_{n_1=1}^{n} p^{n_1}(1-p)^{n-n_1}\left[\binom{n}{n_1}-U_{11}(n,n_1)\right]\cdot\frac{n_1-1}{n-1} \\ &= \frac{\sum_{n_1=2}^{n}\binom{n}{n_1}p^{n_1}(1-p)^{n-n_1}\frac{n_1-1}{n-1}}{1-(1-p)^n-p(1-p)^{n-1}} \\ &= \frac{\frac{1}{n-1}\left[\left(np-np(1-p)^{n-1}\right)-\left(1-(1-p)^n-np(1-p)^{n-1}\right)\right]}{1-(1-p)^n-p(1-p)^{n-1}} \\ &= \frac{\left[p-\frac{1-(1-p)^n}{n}\right]\frac{n}{n-1}}{1-(1-p)^{n-1}} \end{aligned}$$

where the second line follows because $U_{11}(n,n_1)=0$ for $n_1>1$, $U_{11}(n,0)=U_{11}(n,1)=0$, and $C=1/[1-(1-p)^n-p(1-p)^{n-1}]$.[89]

By letting $q:=1-p$ it is straightforward to show that the bias in $\hat{P}_{11}(\boldsymbol{X})$ is negative:

$$\begin{aligned} E\left[\ \hat{P}_{11}(\boldsymbol{X})-p\ \middle|\ I_{11}(\boldsymbol{X})\neq\emptyset\ \right] &= \frac{\left[p-\frac{1-q^n}{n}\right]\frac{n}{n-1}}{1-q^{n-1}}-p \\ &= \frac{(n-1)(q^{n-1}-q^n)-(q-q^n)}{(n-1)(1-q^{n-1})} \\ &< 0 \end{aligned}$$

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

[88] More precisely, $C := 1/\left(1-\sum_{n_1=0}^{n}U_{11}(n,n_1)p^{n_1}(1-p)^{n-n_1}\right)$.

[89] Recall $U_{11}(n,n_1) := |\{\boldsymbol{x}\in\{0,1\}^n : N_1(\boldsymbol{x})=n_1\ \&\ I_{11}(\boldsymbol{X})=\emptyset\}|$, i.e. the number of sequences for which $\hat{P}_{11}(\boldsymbol{x})$ is undefined

The expected difference in proportions

For $k = 1$, we find that the expected difference in proportions is independent of p, and depends only on n.

Theorem 4 *Letting* $\hat{D}_1(\boldsymbol{x}) := \hat{P}_{11}(\boldsymbol{x}) - \hat{P}_{01}(\boldsymbol{x})$*, then for any* $0 < p < 1$ *and* $n > 2$*:*

$$E\left[\ \hat{D}_1(\boldsymbol{X})\ \middle|\ I_{11}(\boldsymbol{X}) \neq \emptyset, I_{01}(\boldsymbol{X}) \neq \emptyset\ \right] = -\frac{1}{n-1}$$

Proof: We show that for $n > 2$ and $n_1 = 1, \ldots, n-1$:

$$E\left[\ \hat{D}_1(\boldsymbol{X})\ \middle|\ N_1(\boldsymbol{X}) = n_1, I_{11}(\boldsymbol{X}) \neq \emptyset, I_{01}(\boldsymbol{X}) \neq \emptyset\ \right] = -\frac{1}{n-1}$$

If $1 < n_1 < n-1$ then $\hat{D}_1(\boldsymbol{x}) := \hat{P}_{11}(\boldsymbol{x}) - \hat{P}_{01}(\boldsymbol{x})$ is defined for all sequences. Therefore, by linearity of the expectation, and a symmetric application of Lemma 2 to $\hat{P}_{01}(\boldsymbol{X})$, we have:

$$\begin{aligned} E[\hat{D}_1(\boldsymbol{X}) | N_1(\boldsymbol{X}) = n_1] =& E[\hat{P}_{11}(\boldsymbol{X}) | N_1(\boldsymbol{X}) = n_1] - E(\hat{P}_{01}(\boldsymbol{X}) | N_1(\boldsymbol{X}) = n_1] \\ =& \frac{n_1 - 1}{n-1} - \left(1 - \frac{n_0 - 1}{n-1}\right) \\ =& -\frac{1}{n-1} \end{aligned}$$

If $n_1 = 1$ then there are $n-1$ possible sequences in which $\hat{D}_1$ is defined (i.e. with 1 not in the final position). For the sequence in which 1 is in the first position, $\hat{D}_1 = 0$. For the other $n-2$ sequences, $\hat{D}_1 = -1/(n-2)$. Therefore, $E\left[\ \hat{D}_1(\boldsymbol{X})\ \middle|\ N_1(\boldsymbol{X}) = 1, I_{11}(\boldsymbol{X}) \neq \emptyset, I_{01}(\boldsymbol{X}) \neq \emptyset\ \right] = -1/(n-1)$. The argument for $n_1 = n-1$ is analogous, with $\hat{D}_1$ undefined for the sequence in which there is a 0 in the last position, equal to 0 for the sequence in which there is 0 in the first position, and equal to $-1/(n-2)$ for all other sequences.

That the conditional expectation is independent of $N_1(\boldsymbol{x})$ implies that $E[D_1(\boldsymbol{X}) \mid I_{11}(\boldsymbol{X}) \neq \emptyset, I_{01}(\boldsymbol{X}) \neq \emptyset\ ]$ is independent of p, yielding the result.

■

C Appendix: A derivation of the formula for the expected proportion, and difference in proportions, for $k > 1$

In this section, for $k > 1$, we obtain the expected proportion of 1s for k/1-streak successor trials, and the expected difference in the proportion of 1s, between k/1-streak successor trials and k/0-streak successor trials.

Similar to what was done in the proof of the $k = 1$ case, representing the proportion in terms of runs allows us to reduce the dimensionality of the problem by identifying the set of sequences over which $\hat{P}_{1k}(\boldsymbol{x})$ is constant. We begin with some basic definitions.

Given the sequence $\boldsymbol{x} = (x_1, \ldots, x_n)$, recall that 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.[90] Define runs of 0s analogously to runs of 1s. 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. Let $S_{1j}(\boldsymbol{x}) = s_{1k}$ be the number of runs of 1s of length j or more, i.e. $S_{1j}(\boldsymbol{x}) := \sum_{i=j}^{n_1} R_{1j}(\boldsymbol{x})$ for $j = 1, \ldots, n_1$, with $S_{0j}(\boldsymbol{x}) = s_{0j}$ defined similarly. 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. $R(\boldsymbol{x}) := R_1(\boldsymbol{x}) + R_0(\boldsymbol{x})$. Further, let the k/1-streak *frequency* statistic $F_{1k}(\boldsymbol{x}) = f_{1k}$ be defined as the number of (overlapping) 1-streaks of length k, i.e. $F_{1k}(\boldsymbol{x}) := \sum_{j=k}^{n_1}(j-k+1)R_{1j}(\boldsymbol{x})$, with $F_{0k}(\boldsymbol{x}) = f_{0k}$ defined analogously. Notice that $f_{1k} = |I_{1k}(\boldsymbol{x})|$ if $\exists i > n-k$ with $x_i = 0$, and $f_{1k} = |I_{1k}| + 1$ otherwise. Also note that $n_1 = f_{11} = \sum_{j=1}^{n_1} j r_{1j}$ and $n_0 = f_{01} = \sum_{j=1}^{n_0} j r_{0j}$.

To illustrate the definitions, consider the sequence of 10 trials 1101100111. The number of 1s is given by $n_1 = 7$. For $j = 1, \ldots, n_1$, the number of runs of 1s of exactly length j is given by $r_{11} = 0$, $r_{12} = 2$, $r_{13} = 1$ and $r_{1j} = 0$ for $j \geq 4$; the number of runs of 1s of length j or more is given by $s_{11} = 3$, $s_{12} = 3$, $s_{13} = 1$ and $s_{1j} = 0$ for $j \geq 4$. The total number of runs is $r = 5$. The k/1-streak *frequency* statistic satisfies $f_{11} = 7$, $f_{12} = 4$, $f_{13} = 1$, and $f_{1j} = 0$ for $j \geq 4$. Finally, the proportion satisfies $p_{11} = 4/6$, $p_{12} = 1/3$, with p_{1j} undefined for $j \geq 3$.

C.1 Expected Proportion ($k > 1$)

In this section we obtain the expected value of the proportion of 1s on k/1-streak successor trials $k > 1$. To shorten the expressions in this section we assume that our sample space of sequences are those in which $\hat{P}_{1k}(\boldsymbol{x})$ is well defined, and we define $P_{1k} = \hat{P}_{1k}(\boldsymbol{x})$ as the induced random variable, with support $\{p_{1k} \in [0,1] : p_{1k} = \hat{P}_{1k}(\boldsymbol{x}) \text{ for } \boldsymbol{x} \in \{0,1\}^n,\ I_{1k}(\boldsymbol{x}) \neq \emptyset\}$.

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 to be equal to $(n_1 - 1)/(n-1)$ for $k = 1$ in Lemma 2 suggests the possibility that, in the spirit of sampling-without-replacement, the expression $(n_1 - k)/(n-k)$ determines the expected proportions for $k > 1$. That this formula does not hold in the case of $k > 1$ can easily be confirmed by setting $k = 2$, $n_1 = 4$, and $n = 5$.[91] As in Section 2, it is not possible to determine $\hat{P}_{1k}(\boldsymbol{x})$ directly by computing its value for each sequence, as the number of admissible sequences is typically

[90] More precisely, it is a subsequence with successive indices $j = i_1 + 1, i_1 + 2, \ldots, i_1 + k$, with $i_1 >= 0$ and $i_1 + k <= n$, 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 $i_1 + k + 1 = 0$

[91] 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 A.3 for intuition).

too large.

We observe that the number of k/1-streak successors satisfies $|I_{1k}(\boldsymbol{x})| = F_{1k}(\boldsymbol{x}_{-n})$, 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 length k/1-streak successors that are themselves equal to 1. Therefore the proportion $\hat{P}_{1k}(\boldsymbol{x})$ can be represented as

$$\hat{P}_{1k}(\boldsymbol{x}) = \frac{F_{1k+1}(\boldsymbol{x})}{F_{1k}(\boldsymbol{x}_{-n})} \quad \text{if } F_{1k}(\boldsymbol{x}_{-n}) > 0 \tag{16}$$

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

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

A classic reference for non-parametric statistical theory Gibbons and Chakraborti (2010) contains a theorem (Theorem 3.3.2, p.87) for the joint distribution $(R_{11}, \ldots, R_{1n_1})$, conditional on N_1 and R_1, from which, in principle, $E[P_{1k}(\boldsymbol{x})|N_1 = n_1]$ can be calculated directly.[92] Unfortunately, the calculation does not appear to be computationally feasible for the sequence lengths of interest here. As a result, we instead follow an approach similar to that in the proof of Lemma 2, making 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 $\hat{P}_{1k}(\boldsymbol{x})$ 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 $\hat{P}_{1k}(\boldsymbol{x}) = G(R_{11}(\boldsymbol{x}), \ldots, R_{1k-1}(\boldsymbol{x}), S_{1k}(\boldsymbol{x}))$ for some G, and 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 lemma's proof we provide a combinatorial argument that we apply repeatedly in the proof of Theorem 6.

Lemma 5 *The number of distinguishable sequences $\boldsymbol{x} \in \{0,1\}^n$, $n \geq 1$, with $n_1 \leq n$ 1s, $r_{1j} \geq 0$*

[92] 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

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

runs of 1s of exactly length j for $j = 1, \ldots k-1$, and $s_{1k} \geq 0$ runs of 1s of length k or more satisfies:

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

where $r_1 = \sum_{j=1}^{k-1} r_{1j} + s_{1k}$ and $f_{1k} = n_1 - \sum_{j=1}^{k-1} j r_{1j} - (k-1)s_{1k}$. Further, let $\binom{n}{k} = n!/k!(n-k)!$ if $n \geq k \geq 0$ and $\binom{n}{k} = 0$ otherwise, except for the special case $\binom{-1}{-1} = 1$.[93]

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 $j r_{1j}$ 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 6 *For n, n_1 and k such that $1 < k \leq n_1 \leq n$*

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

where f_{1k} and C_{1k} depend on $n_0, n_1, r_{11}, \ldots, r_{1k-1}$, and s_{1k}, and are defined as in Lemma 5.[94] U_{1k}

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

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:

For all sequences $\boldsymbol{x} \in \{0,1\}^n$ with n_1 1s, we have 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, with $f_{1k} = 0$, 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

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

coefficient in the binomial expansion must agree with the coefficient in the multinomial expansion.[95]

In case 1ii, with $f_{1k} = 1$ 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 not including the final run of k 1s, which can be done in $r_1!/(\prod_{j=1}^{k-1} r_{1j}!)$ ways, and second placing these runs to the left or the right of the available n_0 0s, not including the right end point, which can be done in $\binom{n_0}{r_1-1}$ ways with 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 in 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 all sequences that satisfy this criteria can be constructed in three steps analogous to those used in Lemma 5 by (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

[95] 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 $r_1 + t_1(k-1) + t_2 = n_1$.

right of a 0, (3) filling the "empty" run cells. For a given $(r_{11}, \ldots, r_{1k-1}, s_{1k})$ the total number of sequences satisfying 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 before all sequences that satisfy this criteria can be constructed in three steps as before, and we consider two subcases, 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). For 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:

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

with $(r_1-1)!/((s_{1k}-1)!\prod_{j=1}^{k-1} r_{1j}!)$ being the number of sequences that terminate in a run of 1s of length k or more. For 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 sum of $\hat{P}_{1k}(\boldsymbol{x})$ across all sequences for which it is defined satisfies:

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

and this reduces to the formula in the theorem because the final two terms can be combined, and then can be summed over only runs that satisfy $\sum_{j=1}^{k-1} j r_{1j} < n_1 - k$, and finally combined with the first term (because $f_{1k} - s_{1k} = 0$ if $\sum_{j=1}^{k-1} j r_{1j} = n_1 - k$).[96]

[96] While the first term has the 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

■

C.2 Expected Difference in Proportions

The exact formula for the expected difference between the proportion of 1s for k/1-streak successor trials and the proportion of 1s for k/0-streak successor trials can be obtained with an approach similar to that of the previous section. The difference satisfies $\hat{D}_k(\boldsymbol{x}) := \hat{P}_{1k} - \hat{P}_{0k}(\boldsymbol{x})$, and there are three categories of sequences for which D_k is 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 $D_k^2 := (f_{1k} - s_{1k})/(f_{1k} - 1) - s_{0k}/f_{0k}$, (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 7 *For n, n_1, n_0 and k such that $n_0 + n_1 = n$, and $1 < k \leq n_0, n_1 \leq n$, the expected difference in the proportion of 1s on k/1-streak successor trials and the proportion of 1s on k/0-streak successor trials, $D_k := P_{1k} - (1 - P_{0k})$, satisfies*

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

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

be the case for the other terms. Even if the other terms have a closed form, $E[P_{1k}|N_1 = n_1]$ cannot, as the term U_{1k} does not allow one.

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

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

and U_k *(see expression* $*$ *on page 24) is the number of sequences in which there are either no k/1-streak successors, or no k/0-streak successors.*

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 5.

The sum total of the differences, across all sequences for which the difference is defined and $N_1 = n_1$ is

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

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 $\sum_{j=1}^{k-1} jr_{0j} < n_0 - k$);[97] 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 $\sum_{j=1}^{k-1} jr_{0j} < n_0 - k$);[98] 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

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

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

terms:

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

which can readily be implemented numerically for the finite sequences considered here.[99] 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 6, by using an application of the inclusion-exclusion principle:

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

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

number of distinguishable permutations of the 0 runs that end in a run of length k or more, i.e.

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

[99] 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.

the above expression by first noting that the third term, which corresponds to those sequences in which there are no k/1-streak successors and no k/0-streak successors, 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 \\
&= \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}!} \\
&= \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}!} \\
&= \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}!} \\
&= \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}!} \\
&= \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 6 to represent them as sums of binomial coefficients, and therefore

U_k reduces to a 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}
$$

■

D Appendix: The relationship with known biases and paradoxes

D.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).[100]

Sampling-without-replacement also underlies a well-known finite sample bias that arises in stan-

[100] 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]).

dard 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 $\hat{\sigma}(\boldsymbol{x})$, then calculate the autocorrelation $\hat{\rho}_{t,t+1}(\boldsymbol{x}) = \hat{cov}_{t,t+1}(\boldsymbol{x})/\hat{\sigma}(\boldsymbol{x})$, where $\hat{cov}_{t,t+1}(\boldsymbol{x})$ is the autocovariance.[101] 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 $\hat{\sigma}(\boldsymbol{x})$, is equal to $-1/(n-1)$ for all $\bar{x}$ and $\hat{\sigma}(\boldsymbol{x})$. 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 B).[102,103]

While drawing connections between these biases is useful for understanding their common underlying source, it also yields further insights. In particular we find that the comparison with the biases here leads to a natural generalization of Berkson's bias that provides conditions under which one should expect the bias to be empirically relevant. Suppose that for disease $i \in \{1, 2, \ldots, n\}$, an individual either has it $(x_i = 1)$, or does not $(x_i = 0)$, where x_i are i.i.d with probability of

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

[102] In Appendix D.3, we also find that the least squares estimators for the linear probability model for $\mathbb{P}(X_i = 1|\prod_{j=i-k}^{i-1} X_j = 1)$, $x_i = \beta_0 + \beta_1 \prod_{j=i-k}^{i-1} x_j$, satisfy $\hat{P}_{1k}(\boldsymbol{x}) = \hat{\beta}_0(\boldsymbol{x}) + \hat{\beta}_1(\boldsymbol{x})$, and in the special case of $k = 1$, $\hat{\beta}_1(\boldsymbol{x}) = \hat{D}_1(\boldsymbol{x})$.

[103] 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 success on 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.

success p. The patient is admitted to the hospital if he or she has at least one disease, i.e. if $N_1(\boldsymbol{x}) \geq 1$. This selection criterion for sequences is nearly identical to that used when calculating the proportion of successes after success, where it is required that $N_1(x_1, \ldots, x_{n-1}) \geq 1$. For both criteria, the outcomes x_i and x_j become negatively associated for $i \neq j$; in Berkson's bias the negative association is for all i, while in the case of the bias presented in Section 2.1 it was shown that the negative association is for $i, j \leq n-1$. Because the bias in the proportion becomes negligible as n gets arbitrarily large, Berkson's bias must also become negligible as the number of potential diseases increases, just as with the finite sample bias in autocorrelation. This asymptotic unbiasedness can explain the elusiveness of evidence to date in support of the empirical relevance of Berkson's bias, and why when evidence has been discovered, the bias has been found to be small in magnitude (Roberts et al. 1978; Sackett 1979).

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

In Figure 4 of Appendix A.3 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 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.[104] 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 2.2. 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

[104] The reason for this is provided in the alternative proof of Lemma 2 in Appendix B

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, but also on how the pattern *overlaps* with itself (Guibas and Odlyzko 1981).[105] 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.[106] 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.[107]

D.3 More on the relationship to finite sample bias in a least-squares linear probability model of time series data

For $\boldsymbol{x} \in \{0,1\}^n$, $x_i = \beta_0 + \beta_1 \prod_{j=i-k}^{i-1} x_j$ is the linear probability model for $\mathbb{P}(X_i = 1 | \prod_{j=i-k}^{i-1} X_j = 1)$, which is the conditional probability of success on trial i, given that it immediately follows k consecutive successes. The theorem below establishes that the least squares estimators satisfy

[105] 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).

[106] 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 $E[\sum_{i=\ell}^{n} 1_{[(X_{i-\ell+1},\ldots,X_i)=(y_1,\ldots,y_\ell)]}] = \sum_{i=\ell}^{n} E[1_{[(X_{i-\ell+1},\ldots,X_i)=(y_1,\ldots,y_\ell)]}] = \sum_{i=\ell}^{n} 1/2^\ell = (n-\ell+1)/2^\ell$.

[107] 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 16 in Appendix C).

$\hat{\beta}_0(\boldsymbol{x}) + \hat{\beta}_1(\boldsymbol{x}) = \hat{P}_{1k}(\boldsymbol{x})$. When trials are independent, with $\mathbb{P}(X_t = 1) = p$ for all t, the formula for the bias from Section 2.3 can be applied directly to these estimators. In the special case of $k = 1$, $\hat{D}_1(\boldsymbol{x}) = \hat{\beta}_1(\boldsymbol{x})$ and it follows from Theorem 4 that the expected value of $\hat{\beta}_1(\boldsymbol{x})$ is equal to $-1/(n-1)$, as found in Section D.1, which accords with known results on the $O(1/n)$ bias in discrete-time autoregressive processes (Shaman and Stine 1988).[108]

Theorem 8 *Let $\boldsymbol{x} \in \{0,1\}^n$ with $I_{1k}(\boldsymbol{x}) \neq \emptyset$. If $\boldsymbol{\beta}_k(\boldsymbol{x}) = (\beta_{0k}(\boldsymbol{x}), \beta_{1k}(\boldsymbol{x}))$ is defined to be the solution to the least squares problem, $\boldsymbol{\beta}_k(\boldsymbol{x}) \in \operatorname{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^2} \|\boldsymbol{x} - [\mathbf{1}\ \boldsymbol{d}]^\intercal \boldsymbol{\beta}\|^2$ where $\boldsymbol{d} \in \{0,1\}^n$ is defined so that $d_i := \prod_{j=i-k}^{i-1} x_i$ for $i = 1, \ldots, n$, then*[109]

$$\hat{P}_{1k}(\boldsymbol{x}) = \beta_{0k}(\boldsymbol{x}) + \beta_{1k}(\boldsymbol{x})$$

Proof:
If $\boldsymbol{\beta}_k(\boldsymbol{x})$ minimizes that sum of squares then $\beta_{1k}(\boldsymbol{x}) = \sum_{i=1}^n (x_i - \bar{x})(d_i - \bar{d}) / \sum_{i=1}^n (d_i - \bar{d})^2$. First, working with the numerator, letting $I_{1k} \equiv I_{1k}(\boldsymbol{x})$ we have

$$\begin{aligned}
\sum_{i=1}^n (x_i - \bar{x})(d_i - \bar{d}) &= \sum_{i \in I_{1k}} (x_i - \bar{x})(d_i - \bar{d}) + \sum_{i \in I_{1k}^C} (x_i - \bar{x})(d_i - \bar{d}) \\
&= \left(1 - \frac{|I_{1k}|}{n}\right) \sum_{i \in I_{1k}} (x_i - \bar{x}) - \frac{|I_{1k}|}{n} \sum_{i \in I_{1k}^C} (x_i - \bar{x}) \\
&= \left(1 - \frac{|I_{1k}|}{n}\right) \sum_{i \in I_{1k}} x_i - \frac{|I_{1k}|}{n} \sum_{i \in I_{1k}^C} x_i - \left(1 - \frac{|I_{1k}|}{n}\right) |I_{1k}| \bar{x} + \frac{|I_{1k}|}{n} (n - |I_{1k}|) \bar{x} \\
&= |I_{1k}| \left(1 - \frac{|I_{1k}|}{n}\right) \left(\frac{\sum_{i \in I_{1k}} x_i}{|I_{1k}|} - \frac{\sum_{i \in I_{1k}^C} x_i}{n - |I_{1k}|} \right)
\end{aligned}$$

[108] We are not aware of existing results that quantify the bias for higher-order autoregressive processes with interactions.

[109] When $I_{1k}(\boldsymbol{x}) = \emptyset$ the solution set of the least squares problem is infinite, i.e. $\operatorname{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^2} \|\boldsymbol{x} - \beta_0\|^2 = \{(\beta_0, \beta_1) \in \mathbb{R}^2 : \beta_0 = (1/n) \sum_{i=1}^n x_i\}$. If we treat $\boldsymbol{\beta}_k(\boldsymbol{x})$ as undefined in this case, then the bias from using the proportion of success restricted to trials that immediately follow a success is equal to the finite sample bias in the coefficients of the associated linear probability model. If instead we define $\beta_{1k}(\boldsymbol{x}) = 0$, then the bias in the coefficients of the associated linear probability model will be less than the bias in the in this proportion.

second, with the denominator of $\beta_{1k}(\boldsymbol{x})$ we have

$$\begin{aligned}\sum_{i=1}^{n}(d_i-\bar{d})^2 &= \sum_{i\in I_{1k}}\left(1-\frac{|I_{1k}|}{n}\right)^2+\sum_{i\in I_{1k}^C}\left(\frac{|I_{1k}|}{n}\right)^2\\ &= |I_{1k}|\left(1-\frac{|I_{1k}|}{n}\right)^2+(n-|I_{1k}|)\left(\frac{|I_{1k}|}{n}\right)^2\\ &= |I_{1k}|\left(1-\frac{|I_{1k}|}{n}\right)\end{aligned}$$

therefore we have

$$\beta_{1k}(\boldsymbol{x})=\frac{\sum_{i\in I_{1k}}x_i}{|I_{1k}|}-\frac{\sum_{i\in I_{1k}^C}x_i}{n-|I_{1k}|}$$

now

$$\begin{aligned}n\beta_{0k}(\boldsymbol{x}) &= n(\bar{x}-\beta_1(\boldsymbol{x})\bar{d})\\ &= \sum_{i=1}^{n}x_i-\left(\frac{\sum_{i\in I_{1k}}x_i}{|I_{1k}|}-\frac{\sum_{i\in I_{1k}^C}x_i}{n-|I_{1k}|}\right)|I_{1k}|\\ &= \sum_{i\in I_{1k}^C}x_i+\frac{|I_{1k}|\sum_{i\in I_{1k}^C}x_i}{n-|I_{1k}|}\\ &= \frac{n\sum_{i\in I_{1k}^C}x_i}{n-|I_{1k}|}\end{aligned}$$

and the result follows from summing both coefficients.

■

Note that the bias in the coefficients follows from the bias in the estimate of the conditional probability, i.e. $E[\hat{P}_{1k}(\boldsymbol{x})]<p$ implies that $E[\beta_{1k}(\boldsymbol{x})]<0$ and $E[\beta_{0k}(\boldsymbol{x})]>p$.[110]

E Appendix: Quantifying the bias for hot hand/steak shooting DGPs

In Section 3.2 the adjustment 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). The bias when the underlying data generating process (DGP) reflects hot hand or streak shooting differs. While there are many DGPs which may produce hot hand shooting, the most natural ones to consider 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 stand

[110] Note that $p=E\left[(1/n)\sum_{i=1}^{n}x_i\right]$, and the sum can be broken up and re-arranged as in the theorem.

on which lay definition is most appropriate, they identify hot hand and streak shooting with (1) "non-stationarity" (the zone, flow, in the groove, in rhythm), (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 unrelated to previous outcomes, and is therefore not observable to the econometrician. This can be most simply modeled as a hidden markov chain over the player's (hidden) ability state. We label (2) as *positive feedback* model, and interpret it as capturing the idea that there may be positive feedback from previous shot outcomes into a player's subsequent probability of success. This can be modeled as an autoregressive process, which is equivalent to a markov chain over shot outcomes.[111]

In Figure 5 we plot the bias in the estimate of the change in field goal percentage due to the hot hand, $\hat{D}_3$, for three alternative DGPs, each of which admits the Bernoulli process as a special case.[112] 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),[113] 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, and hitting 3 shots in a row increases the probability of success by d. Within each panel of the figure, the bias, which is the expected difference between $\hat{D}_3$, the estimate 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 field goal percentages (from 40 percent to 60 percent), for four true shifts in the underlying probability ($d \in \{.1, .2, .3, .4\}$).

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 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."[114] The bias in the positive feedback DGP is uniformly below the bias

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

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

[113] 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 based on the difference in probabilities d and the overall expected field goal 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}).

[114] In practice observers may have more information than the econometrician (e.g. shooting mechanics, perceived confi-

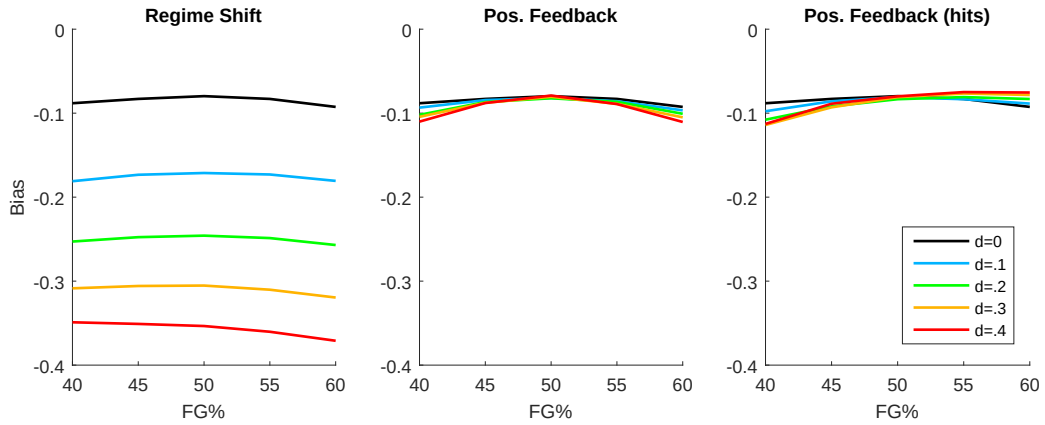

Figure 5: *The bias for three types of hot hand and streak shooting data generating processes (DGPs), where FG% is the expected overall field goal percentage from the DGP, and d represents the change in the player's underlying probability of success. When $d = 0$ each model reduces to a Bernoulli process,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

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 field goal percentages below 50 percent (as in GVTs data), and slightly less strong for field goal 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 3.2 should be expected to (greatly) understate the true magnitude of any underlying hot hand.

dence,or lack thereof, etc.), and may be subject to less measurement error.