SUPPLEMENT TO "SURPRISED BY THE HOT HAND FALLACY? A TRUTH IN THE LAW OF SMALL NUMBERS"

(*Econometrica*, Vol. 86, No. 6, November 2018, 2019–2047)

JOSHUA B. MILLER
Fundamentos del Análisis Económico (FAE), Universidad de Alicante

ADAM SANJURJO
Fundamentos del Análisis Económico (FAE), Universidad de Alicante

APPENDIX D: ONLINE

D.1. *Streak Selection Bias and a Quantitative Comparison to Sampling Without Replacement*

WE SHOW HOW THE DOWNWARD BIAS in the estimator $\hat{P}_k(X)$ is driven by two sources of selection bias. One is related to sampling-without-replacement, and the other to the overlapping nature of streaks.

Recall from the proof of Theorem 1 that $E[\hat{P}_k(\mathbf{X})|I_k(\mathbf{X}) \neq \emptyset] = \mathbb{P}(X_\tau = 1|I_k(\mathbf{X}) \neq \emptyset)$, where τ is drawn (uniformly) at random from $I_k(\mathbf{X})$. Because any sequence $\mathbf{X} \in \{0,1\}^n$, such that $I_k(\mathbf{X}) \neq \emptyset$, that a researcher encounters will contain a certain number of successes $N_1(\mathbf{X}) = n_1$ and failures $n_0 := n - n_1$, for $n_1 = k, \ldots, n$ we can write $\mathbb{P}(X_\tau = 1|I_k(\mathbf{X}) \neq \emptyset) = \sum_{n_1=k}^{n} \mathbb{P}(X_\tau = 1|N_1(\mathbf{X}) = n_1, I_k(\mathbf{X}) \neq \emptyset)\mathbb{P}(N_1(\mathbf{X}) = n_1|I_k(\mathbf{X}) \neq \emptyset)$. To explore the nature of the downward bias, we discuss why $\mathbb{P}(X_\tau = 1|N_1(\mathbf{X}) = n_1, I_k(\mathbf{X})) < \mathbb{P}(X_t = 1|N_1(\mathbf{X}) = n_1) = n_1/n$, that is, why the probability that a randomly drawn trial from $I_k(\mathbf{X})$ is less than the overall proportion of successes in the sequence $\hat{p} = n_1/n$, that is, the prior probability that a trial is a success when it is drawn (uniformly) at random from $1, \ldots, n$ under the knowledge that $N_1(\mathbf{X}) = n_1$.[58]

Suppose that the researcher were to know the overall proportion of successes $\hat{p} = n_1/n$ in the sequence. Now, consider the following two ways of learning that trial t immediately follows k consecutive successes: (i) a trial τ_N, drawn uniformly at random from $\{k+1, \ldots, n\}$, is revealed to be trial $\tau_N = t$, and preceded by k consecutive successes, or (ii) a trial τ_I, drawn (uniformly) at random from $I_k(\mathbf{X}) = \{i : \prod_{i=t-k}^{t-1} X_i = 1\} \subseteq \{k+1, \ldots, n\}$, is revealed to be trial $\tau_I = t$. In each case, the prior probability of success is $\mathbb{P}(X_t = 1) = n_1/n$, which can be equivalently represented with the odds ratio $\mathbb{P}(X_t = 1)/\mathbb{P}(X_t = 0) = n_1/n_0$, indicates the $n_1/n_0 : 1$ prior odds in favor of $X_t = 1$ (relative to $X_t = 0$).

In the first case, the probability distribution for τ_N is given by $\mathbb{P}(\tau_N = t) = 1/(n-k)$ for all $t \in \{k+1, \ldots, n\}$, and is independent of $\mathbf{X}$. 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 can be represented by a sampling-

Joshua B. Miller: joshua.benjamin.miller@gmail.com
Adam Sanjurjo: sanjurjo@ua.es

[58]Note that $\mathbb{P}(N_1(\mathbf{X}) = n_1|I_k(\mathbf{X}) \neq \emptyset) > \mathbb{P}(N_1(\mathbf{X}) = n_1)$ because the exclusion of sequences without a streak of k successes in the first $n-1$ trials biases upwards the number of successes. We do not consider this upward bias here as Theorem 1 shows that the downward biases predominate.

© 2018 The Econometric Society https://doi.org/10.3982/ECTA14943

without-replacement formula, via Bayes's rule:

$$\begin{aligned}
\frac{\mathbb{P}(X_t=1|\tau_N=t)}{\mathbb{P}(X_t=0|\tau_N=t)} &= \frac{\mathbb{P}\left(X_t=1,\prod_{t-k}^{t-1}X_i=1\,\middle|\,\tau_N=t\right)}{\mathbb{P}\left(X_t=0,\prod_{t-k}^{t-1}X_i=1\,\middle|\,\tau_N=t\right)}\\
&= \frac{\mathbb{P}\left(\tau_N=t\,\middle|\,X_t=1,\prod_{t-k}^{t-1}X_i=1\right)}{\mathbb{P}\left(\tau_N=t\,\middle|\,X_t=0,\prod_{t-k}^{t-1}X_i=1\right)}\frac{\mathbb{P}\left(\prod_{t-k}^{t-1}X_i=1\,\middle|\,X_t=1\right)}{\mathbb{P}\left(\prod_{t-k}^{t-1}X_i=1\,\middle|\,X_t=0\right)}\frac{\mathbb{P}(X_t=1)}{\mathbb{P}(X_t=0)}\\
&= \frac{\mathbb{P}\left(\prod_{t-k}^{t-1}X_i=1\,\middle|\,X_t=1\right)}{\mathbb{P}\left(\prod_{t-k}^{t-1}X_i=1\,\middle|\,X_t=0\right)}\frac{\mathbb{P}(X_t=1)}{\mathbb{P}(X_t=0)}\\
&= \frac{\frac{n_1-1}{n-1}\times\cdots\times\frac{n_1-k}{n-k}}{\frac{n_1}{n-1}\times\cdots\times\frac{n_1-k+1}{n-k}}\frac{n_1}{n_0}\\
&= \frac{n_1-k}{n_1}\frac{n_1}{n_0}\\
&= \frac{n_1-k}{n_0}.
\end{aligned}$$

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. That this is a sampling-without-replacement effect can be made most transparent by re-expressing the posterior odds as $\frac{n_1-k}{n-k}/\frac{n_0}{n-k}$.[59,60]

In the second case, the probability that $\tau_I=t$ is drawn from $I_k(\mathbf{X})$ is completely determined by $M:=|I_k(\mathbf{X})|$, and equal to $1/M$. Upon learning that $\tau_I=t$, one can infer the following three things: (i) $I_k(\mathbf{X})\neq\emptyset$, that is, $M\geq 1$, which is informative if $n_1\leq(k-1)(n-n_1)+k$, (ii) t is a member of $I_k(\mathbf{X})$, and (iii) $\prod_{t-k}^{t-1}X_i=1$, as in sampling-without-replacement. As a result, the posterior odds can be determined via Bayes's rule

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

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

in the following way:

$$
\begin{aligned}
&\frac{\mathbb{P}(X_t=1|\tau_I=t)}{\mathbb{P}(X_t=0|\tau_I=t)}\\
&\quad=\frac{\mathbb{P}\left(X_t=1,\prod_{t-k}^{t-1}X_i=1,M\geq 1\,\middle|\,\tau_I=t\right)}{\mathbb{P}\left(X_t=0,\prod_{t-k}^{t-1}X_i=1,M\geq 1\,\middle|\,\tau_I=t\right)}\\
&\quad=\frac{\mathbb{P}\left(\tau_I=t\,\middle|\,X_t=1,\prod_{t-k}^{t-1}X_i=1,M\geq 1\right)}{\mathbb{P}\left(\tau_I=t\,\middle|\,X_t=0,\prod_{t-k}^{t-1}X_i=1,M\geq 1\right)}\frac{\mathbb{P}\left(X_t=1,\prod_{t-k}^{t-1}X_i=1,M\geq 1\right)}{\mathbb{P}\left(X_t=0,\prod_{t-k}^{t-1}X_i=1,M\geq 1\right)}\\
&\quad=\frac{\mathbb{P}\left(\tau_I=t\,\middle|\,X_t=1,\prod_{t-k}^{t-1}X_i=1\right)}{\mathbb{P}\left(\tau_I=t\,\middle|\,X_t=0,\prod_{t-k}^{t-1}X_i=1\right)}\frac{\mathbb{P}\left(\prod_{t-k}^{t-1}X_i=1\,\middle|\,X_t=1\right)}{\mathbb{P}\left(\prod_{t-k}^{t-1}X_i=1\,\middle|\,X_t=0\right)}\frac{\mathbb{P}(X_t=1)}{\mathbb{P}(X_t=0)} \qquad (16)\\
&\quad=\frac{E\left[\frac{1}{M}\,\middle|\,\prod_{t-k}^{t-1}X_i=1,X_t=1\right]}{E\left[\frac{1}{M}\,\middle|\,\prod_{t-k}^{t-1}X_i=1,X_t=0\right]}\frac{\mathbb{P}\left(\prod_{t-k}^{t-1}X_i=1\,\middle|\,X_t=1\right)}{\mathbb{P}\left(\prod_{t-k}^{t-1}X_i=1\,\middle|\,X_t=0\right)}\frac{\mathbb{P}(X_t=1)}{\mathbb{P}(X_t=0)}\\
&\quad=\frac{E\left[\frac{1}{M}\,\middle|\,\prod_{t-k}^{t-1}X_i=1,X_t=1\right]}{E\left[\frac{1}{M}\,\middle|\,\prod_{t-k}^{t-1}X_i=1,X_t=0\right]}\frac{n_1-k}{n_1}\frac{n_1}{n_0}. \qquad (17)
\end{aligned}
$$

For the first term in (16), the event $M\geq 1$ is dropped from the conditional argument because it is implied by the event $\prod_{t-k}^{t-1}X_i=1$, and the term $\frac{\mathbb{P}(M\geq 1|X_t=1,\prod_{t-k}^{t-1}X_i=1)}{\mathbb{P}(M\geq 1|X_t=0,\prod_{t-k}^{t-1}X_i=1)}$ does not appear because it is equal to 1.

Equation (17) gives the posterior odds $\frac{\mathbb{P}(X_t=1|\tau_I=t)}{\mathbb{P}(X_t=0|\tau_I=t)}$ in favor of observing $X_t=1$ (relative to $X_t=0$), for a representative trial $\tau=t$ drawn at random from $I_k(\mathbf{X})$. Observe that the prior odds ratio n_1/n_0 is multiplied by two separate updating factors, which we now discuss.

The first updating factor $\frac{n_1-k}{n_1}$ is clearly strictly less than 1 and reflects the restriction that the finite number of available successes places on the procedure for selecting trials into $I_k(\mathbf{X})$. 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 n_1/n_0. Clearly, the attenuation in the odds due to this factor increases in the streak length k.

The second updating factor $\frac{E[\frac{1}{M}|\prod_{t-k}^{t-1} X_i=1, X_t=1]}{E[\frac{1}{M}|\prod_{t-k}^{t-1} X_i=1, X_t=0]} < 1$, for $t < n$, reflects an additional restriction that the arrangement of successes and failures in the sequence places on the procedure for selecting trials into $I_k(\mathbf{X})$. It can be thought of as the additional information provided by learning that the k successes, which are no longer available, are consecutive and immediately precede t. To see why the odds are further attenuated in this case, we begin with the random variable M, which is defined as the number of trials in $I_k(\mathbf{X})$. The probability of any particular trial $t \in I_k(\mathbf{X})$ being selected at random is $1/M$. Now, because the expectation in the numerator conditions on $X_t = 1$, this means 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 the same as that given in the proof of Theorem 1. For a sequence in which $X_t = 1$, the streak of 1's continues on, meaning that trial $t+1$ must also be in $I_k(\mathbf{X})$, and trials $t+2$ through $t+k$ each may also be in $I_k(\mathbf{X})$. By contrast, for a sequence in which $X_t = 0$, the streak of 1's ends, meaning that trials $t+1$ through $t+k$ cannot possibly be in $I_k(\mathbf{X})$, which leads the corresponding $1/M$ to be smaller in expectation.[61] This last argument provides intuition for why the attenuation of the odds due to this factor increases in k.

Interestingly, for the special case of $k = 1$, $\frac{E[\frac{1}{M}|x_{t-1}=1, x_t=1]}{E[\frac{1}{M}|x_{t-1}=1, x_t=0]} = 1 - \frac{1}{(n-1)(n_1-1)} < 1$ when $t < n$, and $\frac{E[\frac{1}{M}|x_{n-1}=1, x_n=1]}{E[\frac{1}{M}|x_{n-1}=1, x_n=0]} = \frac{n_1}{n_1-1} > 1$ when $t = n$.[62] These contrasting effects combine to yield the familiar sampling-without-replacement formula:

$$E\big[\hat{P}_1(\mathbf{X})|I_1(\mathbf{X}) \neq \emptyset, N_1(\mathbf{X}) = n_1\big] = \frac{n_1 - 1}{n - 1} \tag{18}$$

as demonstrated in Lemma 1, in Appendix A.3. On the other hand, when $k > 1$, the bias is substantially stronger than sampling-without-replacement (see Figure 6), though the formula does not admit a simple representation.[63] For further discussion on the relationship between the bias, sampling-without-replacement, and the *overlapping words paradox* (Guibas and Odlyzko (1981)), see Supplemental Material Appendix F.

A Quantitative Comparison With Sampling-Without-Replacement

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

[61] 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.

[62] The likelihood ratios can be derived following the proof of Lemma 1 in Appendix A.3. In particular, for the equivalent likelihood ratio, $\frac{\mathbb{P}(\tau=t|x_{t-1}=1, x_t=1)}{\mathbb{P}(\tau=t|x_{t-1}=1, x_t=0)}$, the approach used to derive the numerator can also be used to show that the denominator is equal to $\frac{1}{n-2}(\frac{n_0-1}{n_1} + \frac{n_1-1}{n_1-1})$. Further, in the case of $t = n$, it is clear that $\mathbb{P}(\tau = n|x_{n-1} = 1, x_n = 0) = \frac{1}{n_1}$. Each likelihood ratio then follows from dividing and collecting terms.

[63] See Supplemental Zip file for the formula used to produce Figure 6.

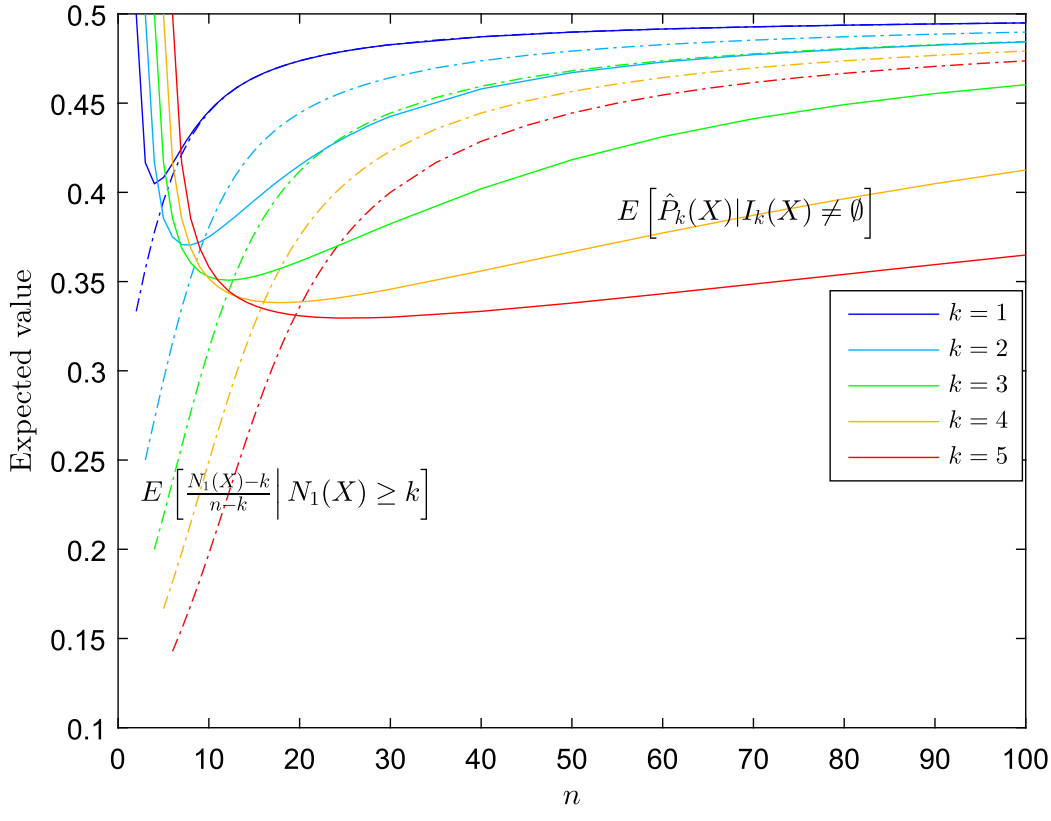

FIGURE 6.—The dotted lines correspond to the bias from sampling-without-replacement. It is the expected probability of a success, given that k successes are first removed from the sequence (assuming $p = .5$). The solid lines correspond to the expected proportion from Figure 1.

$\{1, \ldots, n\}$ have already been drawn from the sequence (sampling-without-replacement), $E\left[\frac{N_1(\mathbf{X})-k}{n-k} \mid N_1(\mathbf{X}) \geq k\right]$. The plot is generated using the combinatorial results discussed in Section 2.1. Note that in the case of $k = 1$, the bias is identical to sampling-without-replacement, as shown in Equation (18).[64] 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.

[64]This appears to contradict Equation (17), that is, that the bias in the procedure used to select the subset of trials $I_{1k}(\mathbf{X})$ is *stronger* than sampling-without-replacement for $t < n$, whereas it is non-existent (thus weaker) for $t = n$. This disparity is due to the second updating factor, which relates to the arrangement. It turns out that for $k = 1$, the determining aspect of the arrangement that influences this updating factor is whether or not the final trial is a success, as this determines the number of successes in the first $n - 1$ trials, where $M = n_1 - X_n$. If one were to instead fix M rather than n_1, then sampling-without-replacement relative to the number of successes in the first $n - 1$ trials would be an accurate description of the mechanism behind the bias, and it induces a negative dependence between any two trials within the first $n - 1$ trials of the sequence. Therefore, it is sampling-without-replacement with respect to M that determines the bias when $k = 1$.

APPENDIX E: THE FORMULA USED TO GENERATE THE SAMPLING DISTRIBUTION AND CALCULATE EXPECTATIONS

We describe the formula used to build the exact sampling distribution of the proportion, and difference in proportions, from which we calculate expectations and plot histograms.

E.1. *Proportion*

Given n trials and streaks of length k, we observe that the proportion of successes on the trials that immediately follow k consecutive success $\hat{P}_k(\mathbf{x})$ can be represented simply as the the number of successes on trials that immediately follow a streak of k consecutive successes divided by the total number of trials—i.e. failures and success—that immediately follow a streak of k consecutive successes. In particular, for a sequence $\mathbf{x} \in \{0,1\}^n$ of successes and failures, we have:

$$\hat{P}_k(\mathbf{x}) = \frac{M^1(\mathbf{x})}{M^0(\mathbf{x}) + M^1(\mathbf{x})},$$

where $M^0(\mathbf{x}) := |\{i \in \{k+1,\ldots,n\} : (1-x_i)\prod_{j=i-k}^{i-1} x_j = 1\}|$ is the number of failures that immediately follow k consecutive successes (suppressing the k to ease notation). Similarly, the number of successes that immediately follow k consecutive successes is defined as $M^1(\mathbf{x}) := \{i \in \{k+1,\ldots,n\} : x_i \prod_{j=i-k}^{i-1} x_j = 1\}$. Finally, the expected value of $\hat{P}_k(\mathbf{x})$ is uniquely determined by the joint distribution of counts $\mathbb{P}((M^0(\mathbf{X}), M^1(\mathbf{X})) = (m^0, m^1))$.

The algorithm described below (recursively) constructs the exact joint distribution of counts, by associating each unique count realization, which we call a *key*, with its corresponding probability.[65] In general, for a sequence of length n and a streak of length k this joint distribution can be represented as a *dictionary* of (key:probability) pairs $D := (\mathbf{m} : p_D(\mathbf{m}))_{\mathbf{m}\in D_c}$, where $\mathbf{m} := (m^0, m^1)$ is a unique pair, D_c corresponds to the set of count realizations with non-zero probability, i.e.

$$D_c := \left\{\mathbf{m} \in \mathbb{N}^2 | \, p_D(\mathbf{m}) > 0\right\}$$

and $p_D(\mathbf{m}) := \mathbb{P}((M^0(\mathbf{X}), M^1(\mathbf{X})) = (m^0, m^1))$.

Table E.I reports the distribution over the sample space of sequences, and the corresponding dictionary, for the simple case of $n = 3$ and $k = 1$. From the dictionary one can derive the sampling distribution of the proportion and directly compute the expected proportion:

$$E\left[\hat{P}_k(\mathbf{x}) | I_k(\mathbf{x}) \neq \emptyset\right] = \sum_{\mathbf{m}\in D_c^*} \frac{m^1}{m^0 + m^1} p_D^*(\mathbf{m}),$$

where $D_c^* = D_c \setminus \{(0,0)\}$ and $p_D^*(\mathbf{m}) := p_D(\mathbf{m}) / \sum_{\mathbf{m}'\in D_c^*} p_D(\mathbf{m}')$.

Let $D(\ell, r)$ be the dictionary that represents the count–probability pairs for the remaining r trials of a sequence that has $\ell \leq k$ consecutive successes immediately preceding the

[65] This algorithm, which builds upon an algorithm suggested by Michael J. Wiener, replaces an exact formula based on the joint distribution of runs of various lengths that we derived in a previous working paper version of this manuscript. The previous formula, while numerically tractable, was less efficient.

TABLE E.I

DICTIONARY REPRESENTATION OF COUNT-PROBABILITY PAIRS[a]

Sample space of sequences		
Sequence	Probability	Count
000	q^3	$(0,0)$
001	q^2p	$(0,0)$
010	q^2p	$(1,0)$
100	q^2p	$(1,0)$
011	qp^2	$(0,1)$
101	qp^2	$(1,0)$
110	qp^2	$(1,1)$
111	p^3	$(0,2)$

Dictionary		
Count $\mathbf{m}$	:	Probability $p_D(\mathbf{m})$
$(0,0)$	:	q^2
$(1,0)$	:	$(q+q^2)p$
$(0,1)$	:	qp^2
$(1,1)$	:	qp^2
$(0,2)$	:	p^3

[a]In the table to the left column one lists the sample space of eight possible sequence realizations from three trials. Column two lists the probability with which the sequence occurs, where p is the probability of success and q is the probability of failure. The third column lists the number of (failures, successes) that immediately follow a success. In the table to the right the joint distribution is represented as a dictionary of count–probability pairs. Each unique count $\mathbf{m} = (m^0, m^1)$ has a unique associated probability equal to the sum of the probabilities of all sequences with the same associated count (see the table on the left).

current trial. For example, if $k = 2$ then $D(0,0) = D(1,0) = D(2,0) = ((0,0) : 1)$, as when zero trials remain in the sequence the only count possible is $(0,0)$, which occurs with probability 1. Also note that $D(1,1) = ((0,0) : 1)$, $D(2,1) = ((1,0) : q, (0,1) : p)$, and $D(2,2) = ((1,0) : q, (1,1) : pq, (0,2) : p^2)$, as a trial can only be counted as a fail or success if it is immediately preceded by $\ell = k = 2$ consecutive successes. The key observation is that given the initial condition $D(\ell, 0) = ((0,0) : 1)$ for $0 \le \ell \le k$, the dictionaries $D(\ell, r)$ can be defined recursively for $r > 0$ and $0 \le \ell \le k$, and take the following form:

$$D(\ell, r) = \begin{cases} D(0, r-1)^{(0,0):q} \uplus D(\ell+1, r-1)^{(0,0):p}, & \text{if } \ell < k, \\ D(0, r-1)^{(1,0):q} \uplus D(k, r-1)^{(0,1):p}, & \text{if } \ell = k, \end{cases}$$

where: (i) the operation $D^{\mathbf{m}':p'} := (\mathbf{m} + \mathbf{m}' : p_D(\mathbf{m}) \times p')_{\mathbf{m} \in D_c}$ increments each count $\mathbf{m}$ with the addition of $\mathbf{m}'$, and scales its corresponding probability $p_D(\mathbf{m})$ by the probability p' of the increment, and (ii) given the dictionaries A and B, the operation $A \uplus B := (\mathbf{m} : (p_A + p_B)(\mathbf{m}))_{\mathbf{m} \in A_c \cup B_c}$ defines the union of two dictionaries as the union of their counts, where the corresponding probabilities for a key that appears in both dictionaries are summed together (we assume that $p_A(\mathbf{m}) = 0$ for $\mathbf{m} \notin A_c$; also for B). If a trial is immediately preceded by $\ell < k$ consecutive successes, then with probability q (p) the next trial to its right will be immediately preceded by 0 $(\ell + 1)$ consecutive successes; regardless of the outcome of the trial, $\mathbf{m}' = (0,0)$ additional failures and successes will be counted as immediately preceded by k successes and $r - 1$ trials will remain. If, on the other hand, a trial is immediately preceded by $\ell = k$ consecutive successes (and there is at least one trial remaining, i.e. $r > 0$), then with probability q (p) the next trial to its right will be immediately preceded by 0 (k) consecutive successes and we will count $\mathbf{m}' = (1,0)$ $((0,1))$ additional failures and successes; regardless of the outcome of the trial, $r-1$ trials will remain.

Algorithm 1 describes the complete recursive procedure.

Algorithm 1 Recursive formula that builds the collection of dictionaries D. Of interest are the dictionaries $D(0,n)$ for $n=k+1,\ldots,N$ which correspond to the joint distribution of the total number of (successes, failures) that immediately follow k consecutive successes in n trials.

Function `Count_Distribution`(N, k, p)**:**
  /* For the definition of $D(\ell,r)$, $A^{m':p'}$ and $A \uplus B$ below, see text. */
  $q \leftarrow 1-p$
  for $n=0,\ldots,N$ **do**
    $L \leftarrow \min\{k,n\}$
    for $\ell = L,\ldots,0$ **do**
      $r \leftarrow n-\ell$
      if $r=0$ **then**
        $D(\ell,r) \leftarrow ((0,0):1)$
      else if $r>0$ **then**
        if $\ell<k$ **then**
          $D(\ell,r) \leftarrow D(0,r-1)^{(0,0):q} \uplus D(\ell+1,r-1)^{(0,0):p}$
        else if $\ell=k$ **then**
          $D(\ell,r) \leftarrow D(0,r-1)^{(1,0):q} \uplus D(k,r-1)^{(0,1):p}$
    end
  end
  return D

E.2. *Difference in Proportions*

The difference in proportions can be computed from a dictionary $D := (\mathbf{m} : p_D(\mathbf{m}))_{\mathbf{m}\in D_c}$, where D_c corresponds to the set of count realizations with non-zero probability i.e.

$$D_c := \left\{\mathbf{m} \in \mathbb{N}^4 | p_D(\mathbf{m}) > 0\right\}$$

and $p_D(\mathbf{m}) := \mathbb{P}((M_0^0(\mathbf{X}), M_0^1(\mathbf{X}), M_1^0(\mathbf{X}), M_1^1(\mathbf{X})) = (m_0^0, m_0^1, m_1^0, m_1^1))$. The variables $M_1^0(\mathbf{X})$ and $M_1^1(\mathbf{X})$ yield the total number of failures and successes (respectively) on those trials that immediately follow a streak of k successes, whereas $M_0^0(\mathbf{X})$ and $M_0^1(\mathbf{X})$ yield the total number of failures and successes (respectively) on those trials that immediately follow a streak of k failures.

Let $D(\ell_0,\ell_1,r)$ be the dictionary that represents the count–probability pairs for the remaining r trials of a sequence in which there are $\ell_0 \le k$ consecutive failures and $\ell_1 \le k$ consecutive successes on the immediately preceding trials (so that $\ell_0\ell_1 = 0$). These dictionaries can be constructed recursively in a way similar to that shown in Supplemental Material Appendix E.1:

$$D(\ell_0,\ell_1,r) = \begin{cases} D(\ell_0+1,0,r-1)^{(0,0,0,0):q} \uplus D(0,\ell_1+1,r-1)^{(0,0,0,0):p}, & \\ & \text{if } \max\{\ell_0,\ell_1\} < k, \\ D(k,0,r-1)^{(1,0,0,0):q} \uplus D(0,1,r-1)^{(0,1,0,0):p}, & \text{if } \ell_0 = k, \\ D(1,0,r-1)^{(0,0,1,0):q} \uplus D(0,k,r-1)^{(0,0,0,1):p}, & \text{if } \ell_1 = k. \end{cases}$$

See Supplemental Zip File for the corresponding code.

APPENDIX F: THE RELATIONSHIP BETWEEN THE STREAK SELECTION BIAS AND KNOWN BIASES AND PARADOXES

F.1. *Sampling-Without-Replacement and the Bias for Streaks of Length* $k = 1$

A brief inspection of Table I 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 H's in the first two 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 (Selvin (1975), Nalebuff (1987), Vos Savant (1990)).[66]

Sampling-without-replacement also underlies a well-known finite sample bias that arises in standard estimates of autocorrelation in time series data (Yule (1926), Shaman and Stine (1988)). 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 $\mathbf{x}$ be a randomly generated sequence consisting of n trials, each of which is an i.i.d. draw from some continuous distribution with finite mean and variance. For a researcher to compute the autocorrelation, she must first determine its sample mean $\bar{x}$ and variance $\hat{\sigma}^2(\mathbf{x})$, then calculate the autocorrelation $\hat{\rho}_{t,t+1}(\mathbf{x}) = \hat{\text{cov}}_{t,t+1}(\mathbf{x})/\hat{\sigma}^2(\mathbf{x})$, where $\hat{\text{cov}}_{t,t+1}(\mathbf{x})$ is the autocovariance.[67] The total sum of values $n\bar{x}$ in a sequence serves as the

[66]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 I (see Miller and Sanjurjo (2015a)).

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

analogue to the number of H's (or G's/Y's) in a sequence in the examples given above. Given $n\bar{x}$, the autocovariance can be represented as the expected outcome from a procedure in which one draws (at random) one of the n trial outcomes x_i, and then takes the product of its difference from the mean $(x_i - \bar{x})$, and another trial outcome j's difference from the mean. Because the outcome's value x_i is essentially drawn from $n\bar{x}$, without replacement, the available sum total $(n\bar{x} - x_i)$ is averaged across the remaining $n - 1$ outcomes, which implies that the expected value of another outcome j's $(j \neq i)$ difference from the mean is given by $E[x_j|x_i, \bar{x}] - \bar{x} = (n\bar{x} - x_i)/(n-1) - \bar{x} = (\bar{x} - x_i)/(n-1)$. Therefore, given $x_i - \bar{x}$, the expected value of the product $(x_i - \bar{x})(x_j - \bar{x})$ must equal $(x_i - \bar{x})(\bar{x} - x_i)/(n-1) = -(x_i - \bar{x})^2/(n-1)$, which is independent of j. Because x_i and j were selected at random, this implies that the expected autocorrelation, given $\bar{x}$ and $\hat{\sigma}^2(\mathbf{x})$, is equal to $-1/(n-1)$ for all $\bar{x}$ and $\hat{\sigma}^2(\mathbf{x})$. This result accords with known results on the $O(1/n)$ bias in discrete-time autoregressive processes (Yule (1926), Shaman and Stine (1988)), and happens to be identical to the result in Theorem 4 for the expected difference in proportions (see Appendix A.3). In the context of time series regression, this bias is known as the *Hurwicz bias* (Hurwicz (1950)), which is exacerbated when one introduces fixed effects into a time series model with a small number of time periods (Nerlove (1967, 1971), Nickell (1981)).[68,69]

F.2. *Pattern Overlap and the Bias for Streaks of Length $k > 1$*

In Figure 6 of Supplemental Material Appendix D, we compare the magnitude of the bias in the (conditional) expected proportion to the pure sampling-without-replacement bias, in a sequence of length n. As can be seen, the magnitude of the bias in the expected proportion is nearly identical to that of sampling-without-replacement for $k = 1$. However, for the bias in the expected proportion, the relatively 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.[70] 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 Supplemental Material Appendix D. 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

[68]The bias that is exacerbated by the introduction of of exogenous variables is commonly known as the "Nickell bias," which was first explored by simulation by Nerlove (1967, 1971). It is an example of what is known as the *incidental parameter problem* (Neyman and Scott (1948), Lancaster (2000)).

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

[70]The reason for this is provided in the alternative proof of Lemma 1 in Appendix A.3.

follow a heads is equal to the number of instances of H in the first $n-1=4$ flips. For any given number of H's 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 H's 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 H's overall) therefore implies that for any other flip in the first four flips of the sequence, the probability of an H is equal to $\frac{3-1}{4-1}=\frac{2}{3}$, regardless of whether or not it is an adjacent flip. On the other hand, for the case of streaks of length $k=2$, the number of opportunities to observe a flip that immediately follows two consecutive heads is equal to the number of instances of HH in the first four flips. Because the pattern HH can overlap with itself, whereas the pattern H cannot, then for a sequence with three H's, if one were to sample an HH from the sequence and examine an adjacent flip within the first four flips, it is not the case that any two of the H's from the sequence can be sampled. For example, in the sequence HHTH, only the first two H's 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, for example, _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)).[71] 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, its probability of occurrence does.[72] 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.[73]

REFERENCES

BAI, D. S. (1975): "Efficient Estimation of Transition Probabilities in a Markov Chain," *The Annals of Statistics*, 3, 1305–1317.[10]

[71]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).

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

[73]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.

BALAKRISHNAN, N., AND M. V. KOUTRAS (2011): *Runs and Scans With Applications*, Vol. 764. John Wiley & Sons.
BERKSON, J. (1946): "Limitations of the Application of Fourfold Table Analysis to Hospital Data," *Biometrics Bulletin*, 2 (3), 47–53.[9]
GIBBONS, J. D., AND S. CHAKRABORTI (2010): *Nonparametric Statistical Inference*. New York: CRC Press, Boca Raton, Florida.
GUIBAS, L. J., AND A. M. ODLYZKO (1981): "String Overlaps, Pattern Matching, and Nontransitive Games," *Journal of Combinatorial Theory, Series A*, 30, 183–208.[4,10,11]
HURWICZ, L. (1950): "Least Square Bias in Time Series," in *Statistical Inference in Dynamic Economic Models*, ed. by T. Koopmans. New York: Wiley.[10]
KONOLD, C. (1995): "Confessions of a Coin Flipper and Would-be Instructor," *The American Statistician*, 49, 203–209.[11]
LANCASTER, T. (2000): "The Incidental Parameter Problem Since 1948," *Journal of Econometrics*, 95, 391–413.[10]
MILLER, J. B., AND A. SANJURJO (2015a): "A Bridge From Monty Hall to the Hot Hand: Restricted Choice, Selection Bias, and Empirical Practice," Working Paper, Available at OSF, https://doi.org/10.31219/osf.io/dmgtp. [9]
NALEBUFF, B. (1987): "Puzzles: Choose a Curtain, Duel-ity, Two Point Conversions, and More," *Journal of Economic Perspectives*, 1 (1), 157–163.[9]
NERLOVE, M. (1967): "Experimental Evidence on the Estimation of Dynamic Economic Relations From a Time Series of Cross-Section," *The Economic Studies Quarterly (Tokyo 1950)*, 18, 42–74.[10]
——— (1971): "Further Evidence on the Estimation of Dynamic Economic Relations From a Time Series of Cross Sections," *Econometrica*, 39, 359–382.[10]
NEYMAN, J., AND E. L. SCOTT (1948): "Consistent Estimates Based on Partially Consistent Observations," *Econometrica*, 16, 1–32.[10]
NICKELL, S. (1981): "Biases in Dynamic Models With Fixed Effects," *Econometrica*, 49, 1417–1426.[10]
NICKERSON, R. S. (2007): "Penney Ante: Counterintuitive Probabilities in Coin Tossing," *The UMAP Journal*, 28, 503–532.[11]
RABIN, M. (2002): "Inference by Believers in the Law of Small Numbers," *Quarterly Journal of Economics*, 117 (3), 775–816.[2]
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," Vol. 11. Discussion Paper # 688. [10]
RIORDAN, J. (1958): *An Introduction to Combinatorial Analysis*. New York: John Wiley & Sons.
SELVIN, S. (1975): "A Problem in Probability (Letter to the Editor)," *The American Statistician*, 29 (1), 67.[9]
SHAMAN, P., AND R. A. STINE (1988): "The Bias of Autoregressive Coefficient Estimators," *Journal of the American Statistical Association*, 83, 842–848.[9,10]
VOS SAVANT, M. (1990): "Ask Marilyn," *Parade Magazine, 15*. [9]
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), 1–63.[9,10]

Co-editor Itzhak Gilboa handled this manuscript.

Manuscript received 17 December, 2016; final version accepted 13 February, 2018; available online 14 March, 2018.