UNCERTAINTY IN THE HOT HAND FALLACY: DETECTING STREAKY ALTERNATIVES TO RANDOM BERNOULLI SEQUENCES

By

David M. Ritzwoller
Joseph P. Romano

Technical Report No. 2020-02
February 2020

Department of Statistics
STANFORD UNIVERSITY
Stanford, California 94305-4065

UNCERTAINTY IN THE HOT HAND FALLACY: DETECTING STREAKY ALTERNATIVES TO RANDOM BERNOULLI SEQUENCES

By

David M. Ritzwoller
Joseph P. Romano
Stanford University

Technical Report No. 2020-02
February 2020

Department of Statistics
STANFORD UNIVERSITY
Stanford, California 94305-4065

http://statistics.stanford.edu

Uncertainty in the Hot Hand Fallacy: Detecting Streaky Alternatives to Random Bernoulli Sequences

David M. Ritzwoller, Stanford University*
Joseph P. Romano, Stanford University

February 13, 2020

Abstract

We study a class of tests of the randomness of Bernoulli sequences and their application to analyses of the human tendency to perceive streaks of consecutive successes as overly representative of positive dependence—the hot hand fallacy. In particular, we study tests of randomness (i.e., that trials are i.i.d.) based on test statistics that compare the proportion of successes that directly follow k consecutive successes with either the overall proportion of successes or the proportion of successes that directly follow k consecutive failures. We derive the asymptotic distributions of these test statistics and their permutation distributions under randomness, under a set of general stationary processes, and under a Markov model of "streakiness", which allow us to evaluate their local asymptotic power. The results are applied to evaluate tests of randomness implemented on data from a basketball shooting experiment, whose conclusions are disputed by Gilovich, Vallone, and Tversky (1985) and Miller and Sanjurjo (2018a). We establish that substantially larger data sets are required to derive an informative measurement of the deviation from randomness in basketball shooting. Although multiple testing procedures reveal that one shooter in the experiment exhibits a shooting pattern significantly inconsistent with randomness – supplying strong evidence that basketball shooting is not random for all shooters all of the time – we find that the evidence against randomness is limited to this shooter. Our results provide a mathematical and statistical foundation for the design and validation of experiments that directly compare deviations from randomness with human beliefs about deviations from randomness in Bernoulli sequences.

Keywords: Bernoulli Sequences, Hot Hand Fallacy, Hypothesis Testing, Permutation Tests

JEL Codes: C12, D9, Z20

*E-mail: ritzwoll@stanford.edu, romano@stanford.edu. DR acknowledges funding from the Stanford Institute for Economic Policy Research (SIEPR). We thank Tom DiCiccio, Maya Durvasula, Matthew Gentzkow, Tom Gilovich, Zong Huang, Victoria de Quadros, Joshua Miller, Linda Ouyang, Adam Sanjurjo, Azeem Shaikh, Jesse Shapiro, Hal Stern, Marius Tirlea, Shun Yang, Molly Wharton, Michael Wolf, and seminar audiences at Stanford University and the California Econometrics Conference for helpful comments and conversations.

1 Introduction

Suppose that, for each i in $1,\ldots,N$, we observe n consecutive Bernoulli trials $\mathbf{X}_i = \{X_{ij}\}_{j=1}^{n}$, with $X_{ij} = 1$ denoting a success and $X_{ij} = 0$ denoting a failure. We are interested in testing either the individual hypotheses

$$H_0^i : \ \mathbf{X}_i \text{ is i.i.d.},$$

the multiple hypothesis problem that tests the hypotheses H_0^i simultaneously, or the joint hypothesis

$$H_0 : \ \mathbf{X}_i \text{ is i.i.d. for each } i \text{ in } 1,\ldots,N$$

against alternatives in which the probabilities of success and failure immediately following streaks of consecutive successes or consecutive failures are greater than their unconditional probabilities.

The interpretation of results of tests of this form have been pivotal in the development of behavioral economics, and in particular, theories of misperception of randomness. In a formative paper, Tversky and Kahneman (1971) hypothesize that people erroneously believe that small samples are highly representative of the "essential characteristics" of the population from which they are drawn. For example, investors who observe a period of increasing returns to an asset will perceive the increase to be representative of the dynamics of the asset and expect increases in returns to persist (Greenwood and Shleifer 2014, Barberis et. al 2015). Similarly, people perceive streaks of ones in Bernoulli sequences to be overly representative of deviations from randomness, and thereby underestimate the probability of streaks when randomness is true (Bar-Hillel and Wagenaar 1991, Rabin 2002).

Gilovich, Vallone, and Tversky (1985), henceforth GVT, test this hypothesis – that people significantly underestimate the probability of streaks in random processes – by analyzing basketball shooting data collected from the Cornell University men and women's varsity and junior varsity basketball teams. They are unable to reject the hypothesis that the sequences of shots they observe are i.i.d. and conclude that the belief in the "hot hand" – that basketball players are more likely to make a shot after one or more successful shots than after one or more misses – is a pervasive cognitive illusion or fallacy. This conclusion became the academic consensus for the following three decades (Kahneman 2011) and provided central empirical support for many economic models in which agents are overconfident in conclusions drawn from small samples (Rabin and Vayanos

2009).

The GVT results were challenged by Miller and Sanjurjo (2018a), henceforth MS, who note that there is a significant small-sample bias in estimates of the probability of success following streaks of successes or failures. They argue that when the GVT analysis is corrected to account for this small-sample bias, they are able to reject the null hypothesis that shots are i.i.d., in favor of positive dependence consistent with expectations of streakiness in basketball.[1]

Miller and Sanjurjo (2018b) argue that their work "uncovered critical flaws ... sufficient to not only invalidate the most compelling evidence against the hot hand, but even to vindicate the belief in streakiness." A more conservative interpretation of their conclusions resulted in persisting uncertainty about the empirical support for textbook theories of misperception of randomness. Benjamin (2018) indicates that MS "re-opens–but does not answer–the key question of whether there is a hot hand *bias* ... a belief in a stronger hot hand than there really is."

The objective of this paper is to clarify and quantify this uncertainty by developing the asymptotic properties of the tests considered by GVT and MS, measuring the tests' finite-sample power with a set of local asymptotic approximations and simulations, and providing a comprehensive presentation and interpretation of the results of these tests implemented on data from the GVT shooting experiment. We focus our empirical analysis on the GVT data because the conclusions reached in GVT and MS based on the same data are starkly different and have resulted in both the former consensus and current uncertainty.[2]

Following MS and GVT, we study the test statistics $\hat{P}_{n,k}(\mathbf{X}_i)$, $\hat{Q}_{n,k}(\mathbf{X}_i)$, and $\hat{D}_{n,k}(\mathbf{X}_i)$, defined as follows. Each X_{ij} has probability of success p_i, which may depend on i. Let each individual's observed probability of success be given by $\hat{p}_{n,i} = \frac{1}{n}\sum_{j=1}^{n} X_{ij}$, and let $\hat{P}_{n,k}(\mathbf{X}_i)$ denote the proportion of successes following k consecutive successes. That is, letting $Y_{ijk} = \prod_{m=j}^{j+k} X_{im}$ and $V_{ik} = \sum_{j=1}^{n-k} Y_{ijk}$, then $\hat{P}_{n,k}(\mathbf{X}_i)$ is given by

$$\hat{P}_{n,k}(\mathbf{X}_i) = V_{ik} \big/ V_{i(k-1)}. \tag{1.1}$$

[1]The MS results earned extensive coverage in the popular press, garnering expository articles in the New York Times (Johnson 2015 and Appelbaum 2015), the New Yorker (Remnick 2015), the Wall Street Journal (Cohen 2015), and on ESPN (Haberstroh 2017), among many other media outlets. MS was the 10th most downloaded paper on SSRN in 2015. Statistics sourced from http://ssrnblog.com/2015/12/29/ssrn-top-papers-of-2015/, accessed on July 21st, 2019.

[2]Miller and Sanjurjo (2019) administer their own controlled shooting experiment and reach conclusions similar to those in their analysis of the GVT shooting experiment. While we do not study the Miller and Sanjurjo (2019) experiment and results in detail, we highlight where our asymptotic results are informative for the power of some of the tests that they perform.

Likewise, let $\hat{Q}_{n,k}(\mathbf{X}_i)$ denote the proportion of failures following k consecutive failures. Letting $Z_{ijk} = \prod_{m=j}^{j+k}(1 - X_{im})$ and $W_{ik} = \sum_{j=1}^{n-k} Z_{ijk}$, then $\hat{Q}_{n,k}(\mathbf{X}_i)$ is given by

$$\hat{Q}_{n,k}(\mathbf{X}_i) = W_{ik}/W_{i(k-1)}. \tag{1.2}$$

Let $\hat{D}_{n,k}(\mathbf{X}_i)$ denote the difference between the proportion of successes following k consecutive successes and k consecutive failures, given by

$$\hat{D}_{n,k}(\mathbf{X}_i) = \hat{P}_{n,k}(\mathbf{X}_i) - \left(1 - \hat{Q}_{n,k}(\mathbf{X}_i)\right). \tag{1.3}$$

Section 2 derives the asymptotic distributions of $\hat{P}_{n,k}(\mathbf{X}_i)$, $\hat{Q}_{n,k}(\mathbf{X}_i)$, and $\hat{D}_{n,k}(\mathbf{X}_i)$ and their permutation distributions under H_0. We give analytical expressions for the normal asymptotic distributions of these test statistics, showing that tests relying on a normal approximation, applied by both GVT and MS, control type 1 error asymptotically. Additionally, we show that the permutation distributions of these statistics converge to the statistics' normal asymptotic distributions, implying that the permutation tests applied by MS behave similarly to tests relying on normal approximations.

Section 3 analyzes the asymptotics of the test statistics and their permutation distributions under stationary processes. First, we characterize the normal asymptotic distributions of the test statistics under general α-mixing processes (Bradley 2005). We use these results to study the asymptotic distributions of the test statistics under a Markov model of "streakiness", in which the probability of a success or a failure is increased directly following m consecutive successes or failures, respectively. We give expressions for the local asymptotic power of the hypothesis tests under consideration against these streaky alternatives.

We show that the tests rejecting for large values of $\hat{D}_{n,1}(\mathbf{X}_i)$ are asymptotically equivalent to the Wald-Wolfowitz runs test (Wald and Wolfowitz 1940), which is known to be the uniformly most powerful unbiased statistic against first-order Markov chains and is the standard test statistic used to test randomness in Bernoulli sequences (Lehmann 1998).[3] As a byproduct of our analysis, we derive the limiting local power function for the Wald-Wolfowitz runs test, which appears to be new. In turn, we show that the test rejecting for large values of $\hat{D}_{n,k}(\mathbf{X}_i)$ has the maximum power within the hypothesis tests that we consider against alternatives in which streakiness be-

[3]Miller and Sanjurjo (2019) note this approximate equivalence. Their results are not asymptotic but are supported by simulation of the correlation of between various test statistics.

gins after k consecutive successes or failures. Simulation evidence indicates that our asymptotic approximations to the power against streaky alternatives perform remarkably well in the sample sizes considered in GVT and MS.

Despite the long history of the tests that we study, our asymptotic results are new. Though some of our initial arguments are fairly standard, deriving the limiting behavior of the permutation distributions proved challenging, even under the null (Theorem 2.2). The standard approach is to verify Hoeffding's condition (see Theorem 15.2.3 in Lehmann and Romano 2005). To do so, we develop a novel application of the Rinott (1994) central limit theorem, which is based on Stein's method. Our derivation of the limiting behavior and local power of the permutation tests under dependent processes is more complex. We obtain the limiting behavior of the permutation distribution under deterministic sequences (i.e., when the number of successes is fixed) with a novel equicontinuity argument (Lemma C.1 in the Online Appendix). This result (Theorem 3.5) holds without probabilistic qualification, unlike results obtained from verifying Hoeffding's condition, and allows us to derive the limiting behavior of the test statistics under dependent sequences (Theorem 3.6).

Section 4 presents several standard methods for testing the hypotheses H_0^i simultaneously and for testing the joint null hypothesis H_0 that all sequences $\mathbf{X}_i$ are i.i.d. We extend our approximation of the power of tests of the individual hypotheses H_0^i to tests of the joint hypothesis H_0 against alternatives in which each $\mathbf{X}_i$ follows the Markov model of streakiness developed in Section 3 with probability θ and are i.i.d. with probability $1-\theta$. Again, we implement a set of simulations to verify the quality of these approximations in the sample sizes considered in GVT and MS. These results significantly reduce the computational expense of power analyses in the design of future experiments.

Having established the asymptotic properties of the tests considered by GVT and MS under the null and under alternative models of streakiness, in Section 5 we evaluate the implications of the GVT shooting experiment outcomes for the question posed in Benjamin (2018): "whether there is ... a belief in a stronger hot hand than there really is." A conclusive answer to this question requires informative estimates of the deviation from randomness and expectations of the deviation from randomness in basketball shooting.

First, we consider evidence for streakiness in the GVT data. When testing the hypotheses H_0^i simultaneously, we find that we are able to reject i.i.d. shooting consistently for only one shooter out of twenty-six, identified as "Shooter 109" in the GVT data. This shooter's shot sequence is remarkably streaky: he makes 16 shots in a row directly following a period in which he misses

15 out of 18 shots. This is strong evidence that basketball shooting is not perfectly random for all basketball players all of the time.[4] However, we find that the tests considered by GVT and MS do not have adequate power to detect parameterizations of our alternative model of streakiness consistent with the variation in NBA field goal and free throw shooting percentages. Moreover, tests of the joint null hypothesis H_0 at the 5% level are not robust to the exclusion of Shooter 109 from the sample. These results indicate that the GVT data are insufficient to provide informative estimates of the deviation from randomness in basketball shooting.

We are not the first to observe that the GVT data are underpowered for the Markov alternatives that we consider. Miller and Sanjurjo (2019), Miyoshi (2000), and Wardrop (1999) measure the power of individual tests of H_0^i against specific parameterizations of similar models with simulation.[5] We contribute to these analyses by deriving analytical approximations of the power, studying a significantly richer set of parameterizations of these models, informing our choices of alternatives by comparison to NBA shooting percentages, and explicitly considering simultaneous tests of H_0^i and tests of the joint null H_0.[6]

Second, we assess the available evidence on expectations of streakiness. We highlight methodological limitations of the surveys of basketball fans presented in GVT. We note a variety of observational estimates (e.g., Rao 2009, Bocskocsky et. al 2014, Lantis and Nesson 2019) consistent with large expected deviations from randomness, but find that the available estimates of beliefs are not directly comparable to measurements of streakiness.

We conclude that larger data and more structured elicitation of beliefs are required to resolve the uncertainty in the hot hand fallacy. We provide a mathematical and statistical foundation for future work with this objective.

The hypothesis tests studied in this paper are applicable to a wider class of questions. Tests of the randomness of stochastic processes against nonrandom, persistent, or streaky alternatives have been studied extensively within finance, economics, and psychology, including in large literatures that develop tests of the efficient market hypothesis (see Fama 1965, Malkiel and Fama 1970, and Malkiel 2003) and tests designed to detect whether mutual funds consistently outperform their benchmarks (see Jensen 1968, Hendricks et. al 1993, Carhart 1997, and Romano and Wolf 2005).

[4]GVT observe the rejection of H_0^i for Shooter 109, but concede "we might expect one significant result out of 26 by chance." We show that the rejection of H_0^i for Shooter 109 is robust to standard multiple testing corrections. Waldrop (1999) notes that the p-value for the Wald-Wolfowitz run test of H_0^i for Shooter 109 is extremely small.

[5]Korb and Stillwell (2003) and Stone (2012) measure power against particular non-stationary alternatives.

[6]Our results align with the conclusions of Stern and Morris (1993), who show that tests of the randomness of hitting streaks in baseball applied in Albright (1993) have limited power.

More broadly, our paper contributes to the literature on inference in Markov Chains (see Billingsley 1961, Chapter 5 of Bhat and Miller 2000, and references therein).

Section 6 concludes. The appendix presents a detailed overview and replication of the analyses of GVT and MS. Online Appendix A and Online Appendix B include supplemental tables and figures relevant to our analysis, respectively. Proofs of all mathematical results presented in the main body of this paper are given in Online Appendix C.

2 Asymptotics Under I.I.D. Processes

In this section, we derive the asymptotic unconditional sampling distributions of $\hat{P}_{n,k}(\mathbf{X}_i)$, $\hat{P}_{n,k}(\mathbf{X}_i) - \hat{p}_{n,i}$, and $\hat{D}_{n,k}(\mathbf{X}_i)$ under H_0^i. We also derive the limiting behavior of the corresponding permutation distributions of these test statistics.

For ease of notation, we drop the dependence on the individual i. The asymptotic distribution of $\hat{Q}_{n,k}(\mathbf{X})$ can be obtained by replacing p with $1-p$ in the expressions for the asymptotic distributions of $\hat{P}_{n,k}(\mathbf{X})$. Note that $\hat{P}_{n,k}(\mathbf{X})$, $\hat{Q}_{n,k}(\mathbf{X})$, and $\hat{D}_{n,k}(\mathbf{X})$ are not defined for every sequence $\mathbf{X}$, that is they are not defined for sequences without instances of k consecutive successes or failures. However, the statistics are defined with probability approaching one exponentially quickly as n grows to infinity.

2.1 Asymptotic Behavior of the Test Statistics

First, we evaluate the asymptotic distributions of $\hat{P}_{n,k}(\mathbf{X})$, $\hat{P}_{n,k}(\mathbf{X}) - \hat{p}_n$, and $\hat{D}_{n,k}(\mathbf{X})$ under H_0. Despite the long history of this problem, such distributions have not been provided to date. Miller and Sanjurjo (2014) claim that $\hat{P}_{n,k}(\mathbf{X})$ is asymptotically normal, referencing Mood (1940), but are unable to provide explicit formulae for the asymptotic variances. Note that, even in the null i.i.d. case, the test statistics are functions of overlapping subsequences of observations, thus central limit theorems for dependent data are required. In order to analyze the asymptotic behavior of the permutation distributions, we are aided by an appropriate central limit theorem using Stein's method (see Rinott 1994 and Stein 1986).

Theorem 2.1. *Under the assumption that* $\mathbf{X} = \left\{X_j\right\}_{j=1}^n$ *is a sequence of i.i.d. Bernoulli(p) random variables,*

(i) $\hat{P}_{n,k}(\mathbf{X})$*, given by (1.1), is asymptotically normal with limiting distribution given by*

$$\sqrt{n}\left(\hat{P}_{n,k}(\mathbf{X})-p\right) \xrightarrow{d} N\left(0,\sigma_P^2(p,k)\right), \tag{2.1}$$

where $\sigma_P^2(p,k) = p^{1-k}(1-p)$ and $\xrightarrow{d}$ denotes convergence in distribution,

(ii) *$\hat{P}_{n,k}(\mathbf{X})-\hat{p}$, where $\hat{p} = n^{-1}\sum_{i=1}^n X_i$, is asymptotically normal with limiting distribution given by*

$$\sqrt{n}\left(\hat{P}_{n,k}(\mathbf{X})-\hat{p}_n\right) \xrightarrow{d} N\left(0,\sigma_{\tilde{P}}^2(p,k)\right), \tag{2.2}$$

where $\sigma_{\tilde{P}}^2(p,k) = p^{1-k}(1-p)\left(1-p^k\right)$,

(iii) *and $\hat{D}_{n,k}(\mathbf{X})$, given by (1.3), is asymptotically normal with limiting distribution given by*

$$\sqrt{n}\hat{D}_{n,k}(\mathbf{X}) \xrightarrow{d} N\left(0,\sigma_D^2(p,k)\right), \tag{2.3}$$

where $\sigma_D^2(p,k) = (p(1-p))^{1-k}\left((1-p)^k+p^k\right)$.

Remark 2.1. Note that $\sigma_D^2\left(\frac{1}{2},k\right) = 2^{k-1}$ increases quite rapidly with k, stemming from an effectively reduced sample size when considering successes, or failures, following only streaks of length k. ■

Remark 2.2. Theorem 2.1 can be generalized to a triangular array $\mathbf{X}_n = \left\{X_{n,j}\right\}_{j=1}^n$ of i.i.d. Bernoulli trials with probability of success p_n converging to p. Specifically, we have that,

$$n^{1/2}\left(\hat{P}_k(\mathbf{X}_n)-p_n\right) \xrightarrow{d} N\left(0,\sigma_P^2(p,k)\right),$$
$$n^{1/2}\left(\hat{P}_k(\mathbf{X}_n)-\hat{p}_n\right) \xrightarrow{d} N\left(0,\sigma_{\tilde{P}}^2(p,k)\right), \text{ and}$$
$$n^{1/2}\hat{D}_k(\mathbf{X}_n) \xrightarrow{d} N\left(0,\sigma_D^2(p,k)\right).$$

This result implies that we can consistently approximate the quantiles of the distributions of $\hat{P}_{n,k}(\mathbf{X}_n)$ and $\hat{D}_{n,k}(\mathbf{X}_n)$ with the parametric bootstrap, which approximates the distribution of $\sqrt{n}\hat{D}_{n,k}(\mathbf{X})$ under p by that of $\sqrt{n}\hat{D}_{n,k}(\mathbf{X})$ under $\hat{p}_n$. ■

Remark 2.3. MS show that, under H_0^i, the expectations of $\hat{P}_{n,k}(\mathbf{X})-\hat{p}_n$ and $\hat{D}_{n,k}(\mathbf{X})$ under the null are significantly less than 0 in small samples. Although exact expressions for the expectations of these statistics appear to be unknown for $k>1$, in Online Appendix D we obtain the second order approximations

$$\mathbb{E}\left[\hat{P}_{n,k}(\mathbf{X})-\hat{p}_n\right]=n^{-1}p\left(1-p^{-k}\right)+O\left(n^{-2}\right) \text{ and}$$
$$\mathbb{E}\left[\hat{D}_{n,k}(\mathbf{X})\right]=n^{-1}\left(1-(1-p)^{1-k}-p^{1-k}\right)+O\left(n^{-2}\right).$$

■

Remark 2.4. Note that the asymptotic variance of $\hat{D}_{n,k}(\mathbf{X})$ is equal to the sum of the asymptotic variances of $\hat{P}_{n,k}(\mathbf{X})$ and $\hat{Q}_{n,k}(\mathbf{X})$, suggesting that

$$n\operatorname{Cov}\left(\hat{P}_{n,k}(\mathbf{X}),\hat{Q}_{n,k}(\mathbf{X})\right)\to 0. \tag{2.4}$$

In fact, in Online Appendix D, we show that $\operatorname{Cov}\left(\hat{P}_{n,k}(\mathbf{X}),\hat{Q}_{n,k}(\mathbf{X})\right)=O\left(n^{-2}\right)$. GVT and MS approximate the variance of $\hat{D}_{n,k}(\mathbf{X})$ with estimators that implicitly assume (2.4). MS cite a simulation exercise supporting their assumption. Our results justify this assumption mathematically. Additionally, the asymptotic variance of $\hat{P}_{n,k}(\mathbf{X})-p$ is equal to the sum of the asymptotic variance of $\hat{P}_{n,k}(\mathbf{X})-\hat{p}_n$ and $p(1-p)$, which implies $\hat{P}_{n,k}(\mathbf{X})-\hat{p}_n$ and $\hat{p}_n$ are asymptotically independent. ■

2.2 Asymptotic Behavior of the Permutation Distribution

Next, we will consider the permutation distribution for various test statistic sequences $T=\{T_n\}$. As a robustness check to their results relying on a normal approximation, MS perform a permutation test, rejecting for large values of $\hat{D}_{n,k}(\mathbf{X})$. In general, the permutation, or randomization, distribution for $\sqrt{n}T_n$ is given by

$$\hat{R}_n(t)=\frac{1}{n!}\sum_{\pi}I\left\{\sqrt{n}T_n\left(X_{\pi(1),\ldots,}X_{\pi(n)}\right)\le t\right\}, \tag{2.5}$$

where $\pi=(\pi(1),\ldots,\pi(n))$ is a permutation of $(1,\ldots,n)$. Of course, the permutation distribution is just the distribution of $\sqrt{n}T_n$ conditional on the number of successes. By sufficiency, $\hat{R}_n$ does not depend on p and, by completeness of the number of successes, permutation tests are the only tests that are exactly level α. Therefore, in practice, we will use permutation tests. Deriving these tests' asymptotic distributions allows us to analyze their power.

Theorem 2.2. *Let $\Phi(\cdot)$ denote the standard normal cumulative distribution function. Assuming*

$X_1, X_2, \ldots$ are i.i.d Bernoulli (p) variables, then

(i) *the permutation distribution of $\sqrt{n}T_n$ based on the test statistic $T_n = \hat{D}_{n,k}(X_1, \ldots, X_n)$ satisfies*

$$\sup_t \left|\hat{R}_n(t) - \Phi(t/\sigma_D(p,k))\right| \xrightarrow{P} 0, \tag{2.6}$$

where $\xrightarrow{P}$ denotes convergence in probability, and

(ii) *the permutation distribution of $\sqrt{n}T_n$ based on the test statistic $T_n = \hat{P}_{n,k}(X_1, \ldots, X_n) - \hat{p}_n$ satisfies*

$$\sup_t \left|\hat{R}_n(t) - \Phi(t/\sigma_{\tilde{P}}(p,k))\right| \xrightarrow{P} 0, \tag{2.7}$$

where $\sigma_D(p,k)$ and $\sigma_{\tilde{P}}(p,k)$ are given in Theorem 2.1.

In particular, part (i) shows that the (random) permutation distribution of $\sqrt{n}\hat{D}_{n,k}(\mathbf{X})$ behaves asymptotically like the true unconditional sampling distribution of $\sqrt{n}\hat{D}_{n,k}(\mathbf{X})$. Note, however, that due to the need to center $\hat{P}_{n,k}(\mathbf{X})$ by $\hat{p}_n$, the same is not true for the sampling distribution of $\sqrt{n}\left(\hat{P}_{n,k}(\mathbf{X}) - p\right)$.

3 Asymptotics Under General Stationary Processes

In this section, we describe the asymptotic distributions and permutation distributions of $\hat{D}_{n,k}(\mathbf{X})$ and $\hat{P}_{n,k}(\mathbf{X}) - \hat{p}_n$ under a general stationary process $\mathbb{P}$. When considering the asymptotic distributions of the statistics, we confine the class of processes considered to those satisfying a particular notion of asymptotic independence, or mixing.

3.1 A General Convergence Theorem Under α-Mixing

Define the measure of dependence

$$\alpha(\mathscr{A}, \mathscr{B}) = \sup\left\{\left|\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)\right| A \in \mathscr{A}, B \in \mathscr{B}\right\}, \tag{3.1}$$

where $\mathscr{A}$ and $\mathscr{B}$ are two sub σ-fields of the σ-field $\mathscr{F}$. For $\mathbf{X} = (X_i, i \in \mathbb{Z}_+)$, a sequence of random variables, let us define the mixing coefficient

$$\alpha(\mathbf{X}, n) = \sup_{j \in \mathbb{Z}} \alpha\left(\mathscr{F}_{-\infty}^{j}(\mathbf{X}), \mathscr{F}_{j+n}^{\infty}(\mathbf{X})\right), \tag{3.2}$$

where the σ-field $\mathscr{F}_J^K(\mathbf{X})$ is given by $\sigma(X_i, J \leq i \leq K)$, with $\sigma(\ldots)$ denoting the σ-field generated by $(\ldots)$. We say $\mathbf{X}$ is α-mixing if $\alpha(\mathbf{X}, n) \to 0$ as $n \to \infty$. Additionally, for $\mathbf{G} = (G_i, i \in \mathbb{Z}_+)$, a stationary sequence of random vectors, let

$$\Sigma(\mathbf{G}) = \operatorname{Var}(G_1) + 2\sum_{i=2}^{\infty} \operatorname{Cov}(G_1, G_i). \tag{3.3}$$

By appealing to Theorem 1.7 of Ibragimov (1962), we can give a general form for the asymptotic distributions of the test statistics under α-mixing processes.

Theorem 3.1. *Assuming $\mathbf{X} = \left(X_j, j \in \mathbb{Z}_+\right)$ is a stationary, α-mixing, Bernoulli sequence such that $\sum_{j=1}^{\infty} \alpha(\mathbf{X}, j) < \infty$, with $\alpha(\mathbf{X}, j)$ given by (3.2), then*
(i) *$\hat{P}_{n,k}(\mathbf{X})$, where $\hat{P}_{n,k}(\mathbf{X})$ is given by (1.1), is asymptotically normal with limiting distribution given by*

$$\sqrt{n}\left(\hat{P}_{n,k}(\mathbf{X}) - \frac{\mathbb{E}\left[Y_{jk}\right]}{\mathbb{E}\left[Y_{j(k-1)}\right]}\right) \xrightarrow{d} N\left(0, \mathbb{E}\left[\Psi_j\right]^\top \Sigma\left(\Psi_j\right) \mathbb{E}\left[\Psi_j\right]\right), \tag{3.4}$$

where, $\Psi_j = \left[Y_{jk}, Y_{j(k-1)}\right]^\top$ and $\Sigma\left(\Psi_j\right)$ is given by (3.3), and
(ii) *$\hat{P}_{n,k}(\mathbf{X}) - \hat{p}_n$ is asymptotically normal with limiting distribution given by*

$$\sqrt{n}\left(\left(\hat{P}_{n,k}(\mathbf{X}) - \hat{p}_n\right) - \left(\frac{\mathbb{E}\left[Y_{jk}\right]}{\mathbb{E}\left[Y_{j(k-1)}\right]} - p\right)\right) \xrightarrow{d} N\left(0, \mathbb{E}\left[\Gamma_j\right]^\top \Sigma\left(\Gamma_j\right) \mathbb{E}\left[\Gamma_j\right]\right), \tag{3.5}$$

where $\Gamma_j = \left[Y_{jk}, Y_{j(k-1)}, X_i\right]^\top$ and $\Sigma\left(\Gamma_j\right)$ is given by (3.3), and
(iii) *$\hat{D}_{n,k}(\mathbf{X})$, given by (1.3), is asymptotically normal with limiting distribution given by*

$$\sqrt{n}\left(\hat{D}_{n,k}(\mathbf{X}) - \left(\frac{\mathbb{E}\left[Y_{jk}\right]}{\mathbb{E}\left[Y_{j(k-1)}\right]} - \left(1 - \frac{\mathbb{E}\left[Z_{jk}\right]}{\mathbb{E}\left[Z_{j(k-1)}\right]}\right)\right)\right) \xrightarrow{d} N\left(0, \mathbb{E}\left[\Lambda_j\right]^\top \Sigma\left(\Lambda_j\right) \mathbb{E}\left[\Lambda_j\right]\right), \tag{3.6}$$

where $\Lambda_j = \left[Y_{jk}, Y_{jk-1}, Z_{jk}, Z_{jk-1}\right]^\top$ and $\Sigma\left(\Lambda_j\right)$ is given by (3.3).

Remark 3.1. Note that $\mathbb{E}\left[Y_{jk}\right]/\mathbb{E}\left[Y_{j(k-1)}\right]$ is equal to the probability of a success following k consecutive successes, given by $\gamma_P(\mathbb{P},k) = \mathbb{P}\left(X_{j+k} = 1 | X_{j+k-1} = 1, \ldots, X_j = 1\right)$. Likewise, the asymptotic mean of $\hat{D}_{n,k}(\mathbf{X})$ is equal to the difference in the probability of successes following k consecutive successes and failures, given by

$$\gamma_D(\mathbb{P},k) = \mathbb{P}\left(X_{j+k} = 1 | X_{j+k-1} = 1, \ldots, X_i = 1\right) - \mathbb{P}\left(X_{j+k} = 1 | X_{j+k-1} = 0, \ldots, X_j = 0\right).$$

The parameters $\gamma_P(\mathbb{P},k)$ and $\gamma_D(\mathbb{P},k)$ are functionals of the underlying stationary process $\mathbb{P}$ and the value of k.

Theorem 3.1 implies that

$$\sqrt{n}\left(\hat{P}_{n,k}(\mathbf{X}) - \gamma_P(\mathbb{P},k)\right) \xrightarrow{d} N\left(0, \tau_P^2(\mathbb{P},k)\right) \text{ and}$$
$$\sqrt{n}\left(\hat{D}_{n,k}(\mathbf{X}) - \gamma_D(\mathbb{P},k)\right) \xrightarrow{d} N\left(0, \tau_D^2(\mathbb{P},k)\right)$$

where the limiting variances $\tau_P^2(\mathbb{P},k)$ and $\tau_D^2(\mathbb{P},k)$ are also parameters or functionals of the underlying process $\mathbb{P}$ and k. In particular, $\tau_D^2(\mathbb{P},k) = \mathbb{E}\left[\Lambda_j\right]^\top \Sigma\left(\Lambda_j\right)\mathbb{E}\left[\Lambda_j\right]$, as in part (iii) of Theorem 3.1. If $\hat{\tau}_P^2(k)$ and $\hat{\tau}_D^2(k)$ are consistent estimators of $\tau_P^2(\mathbb{P},k)$ and $\tau_D^2(\mathbb{P},k)$, then $\hat{P}_{n,k}(\mathbf{X}) \pm \hat{\tau}_P(k)\frac{z_{1-\alpha/2}}{\sqrt{n}}$ and $\hat{D}_{n,k}(\mathbf{X}) \pm \hat{\tau}_D(k)\frac{z_{1-\alpha/2}}{\sqrt{n}}$ are asymptotically valid confidence intervals for $\gamma_P(\mathbb{P},k)$ and $\gamma_D(\mathbb{P},k)$ respectively. Of course, when H_0^i is true, $\tau_P^2(\mathbb{P},k) = \sigma^2(p,k)$, where p is the marginal probability of success at any time point for the process $\mathbb{P}$. ■

Remark 3.2. For a fixed stationary model, the limiting variances of $\hat{P}_{n,k}(\mathbf{X})$, $\hat{P}_{n,k}(\mathbf{X}) - \hat{p}_n$, and $\hat{D}_{n,k}(\mathbf{X})$ can be quite complicated. However, they, as well as their entire sampling distributions, can be estimated with general bootstrap methods for stationary time series (see Lahiri 2003), such as the moving blocks bootstrap (Liu and Singh 1992 and Künsch 1989), the stationary bootstrap (Politis and Romano 1994), or subsampling (Politis et. al 1999). Such methods provide asymptotically valid confidence intervals for general parameters, such as $\gamma_P(\mathbb{P},k)$ and $\gamma_D(\mathbb{P},k)$. ■

Remark 3.3. If we consider a stationary sequence of alternatives that is contiguous to H_0^i for some p, then by LeCam's 3rd lemma, we expect that $\hat{P}_{n,k}(\mathbf{X})$, $\hat{P}_{n,k}(\mathbf{X}) - \hat{p}_n$, and $\hat{D}_{n,k}(\mathbf{X})$ have limiting distributions with shifted means and that their limiting variances are the same as under H_0^i. In this case, $\hat{P}_{n,k}(\mathbf{X}) \pm \sigma_P(\hat{p}_n,k)\frac{z_{1-\alpha/2}}{\sqrt{n}}$ and $\hat{D}_{n,k}(\mathbf{X}) \pm \sigma_D(\hat{p}_n,k)\frac{z_{1-\alpha/2}}{\sqrt{n}}$ are asymptotically valid confidence intervals for $\gamma_P(\mathbb{P},k)$ and $\gamma_D(\mathbb{P},k)$ under stationary alternatives contiguous to H_0^i. As we have identified the expression for the mean of the limiting distributions of $\hat{P}_{n,k}(\mathbf{X})$, $\hat{P}_{n,k}(\mathbf{X}) -$

$\hat{p}_n$, and $\hat{D}_{n,k}(\mathbf{X})$, and have previously calculated their limiting variances under H_0, we can now anticipate their limiting distributions under contiguous alternatives. This will allow us to calculate the limiting power for $\hat{P}_{n,k}(\mathbf{X})$, $\hat{P}_{n,k}(\mathbf{X}) - \hat{p}_n$, and $\hat{D}_{n,k}(\mathbf{X})$ under various alternatives. Note that we have not verified the conditions in LeCam's 3rd lemma. However, we will formally verify the limiting behavior of the test statistics under consideration in some Markov Chain models in the next subsection. ■

3.2 Power Against a Markov Chain Model of Streakiness

In this section, we study the asymptotic power of tests of randomness using the test statistics $\hat{P}_{n,k}(\mathbf{X}) - \hat{p}_n$ and $\hat{D}_{n,k}(\mathbf{X})$ against a class of Markov Chain alternative models of streakiness, wherein persistence begins after streaks of m successive successes or failures. First, we evaluate the exact asymptotic distribution of $\hat{P}_{n,k}(\mathbf{X}) - \hat{p}_n$ and $\hat{D}_{n,k}(\mathbf{X})$ when m is equal to 1.

Theorem 3.2. *Assuming $X_1, X_2, \ldots$ is a two-state stationary Markov Chain on $\{0,1\}$ with transition matrix given by*

$$\mathscr{P} = \begin{bmatrix} \frac{1}{2}+\varepsilon & \frac{1}{2}-\varepsilon \\ \frac{1}{2}-\varepsilon & \frac{1}{2}+\varepsilon \end{bmatrix}, \tag{3.7}$$

where $0 \le \varepsilon < \frac{1}{2}$, then

(i) *$\hat{D}_{n,1}(\mathbf{X})$, given by (1.3) with k equal to 1, is asymptotically normal with limiting distribution given by*

$$\sqrt{n}\left(\hat{D}_{n,1}(\mathbf{X}) - 2\varepsilon\right) \xrightarrow{d} N\left(0, 1-4\varepsilon^2\right). \tag{3.8}$$

(ii) *$\hat{P}_{n,1}(\mathbf{X})$, given by (1.1) with k equal to 1, is asymptotically normal with limiting distribution given by*

$$\sqrt{n}\left(\hat{P}_{n,1}(\mathbf{X}) - \frac{1}{2} - \varepsilon\right) \xrightarrow{d} N\left(0, \frac{1}{2} - 2\varepsilon^2\right). \tag{3.9}$$

(ii) *$\hat{P}_{n,1}(\mathbf{X}) - \hat{p}_n$, given by (1.1) with k equal to 1, is asymptotically normal with limiting distribution given by*

$$\sqrt{n}\left(\hat{P}_{n,1}(\mathbf{X}) - \hat{p}_n - \varepsilon\right) \xrightarrow{d} N\left(0, \frac{1-2\varepsilon+16\varepsilon^2}{4-8\varepsilon}\right). \tag{3.10}$$

Remark 3.4. The argument for Theorem 3.2 holds if we let ε vary with n such that $\varepsilon_n = \varepsilon + O\left(n^{-1/2}\right)$. If we take $\varepsilon_n = \frac{h}{\sqrt{n}}$, then $\sqrt{n}\left(\hat{D}_{n,1}(\mathbf{X}) - \frac{2h}{\sqrt{n}}\right) \xrightarrow{d} N(0,1)$ and therefore the power of the test that rejects when $\sqrt{n}\hat{D}_{n,1}(\mathbf{X}) > z_{1-\alpha}$ is given by

$$\begin{aligned}\mathbb{P}\left(\sqrt{n}\hat{D}_{n,1}(\mathbf{X}) > z_{1-\alpha}\right) &= \mathbb{P}\left(\sqrt{n}\left(\hat{D}_{n,1}(\mathbf{X}) - \frac{2h}{\sqrt{n}}\right) > z_{1-\alpha} - 2h\right) \\ &\to 1 - \Phi(z_{1-\alpha} - 2h).\end{aligned} \tag{3.11}$$

The same limiting power results if $z_{1-\alpha}$ is replaced by the permutation quantile. ■

Next, we verify the asymptotic normality of $\hat{P}_{n,k}(\mathbf{X}) - \hat{p}_n$ and $\hat{D}_{n,k}(\mathbf{X})$ for deviations from independence occurring at general m.

Theorem 3.3. *Let $0 \leq \varepsilon < \frac{1}{2}$. Assume $X_1, X_2, \ldots$ is a two-state stationary Markov chain of order m on $\{0,1\}$ such that the probability of transitioning from 1 to 1 (0 to 0) is $\frac{1}{2} + \varepsilon$ after m successive 1's ($0's$) and $\frac{1}{2}$ after any other sequence of m states with at least one 1 and one 0, then*

$$\begin{gathered}\sqrt{n}\left(\hat{P}_{n,k}(\mathbf{X}) - \mu_P(k,m,\varepsilon)\right) \xrightarrow{d} N\left(0, \sigma_P^2(k,m,\varepsilon)\right), \\ \sqrt{n}\left(\hat{P}_{n,k}(\mathbf{X}) - \hat{p}_n - \mu_{\tilde{P}}(k,m,\varepsilon)\right) \xrightarrow{d} N\left(0, \sigma_{\tilde{P}}^2(k,m,\varepsilon)\right), \text{ and} \\ \sqrt{n}\left(\hat{D}_{n,k}(\mathbf{X}) - \mu_D(k,m,\varepsilon)\right) \xrightarrow{d} N\left(0, \sigma_D^2(k,m,\varepsilon)\right)\end{gathered} \tag{3.12}$$

where $\mu_P(k,m,\varepsilon)$, $\mu_{\tilde{P}}(k,m,\varepsilon)$, and $\mu_D(k,m,\varepsilon)$ are given explicitly in the proof and $\sigma_P^2(k,m,\varepsilon)$, $\sigma_{\tilde{P}}^2(k,m,\varepsilon)$, and $\sigma_D^2(k,m,\varepsilon)$ are functions of k, m, and ε.

Remark 3.5. The functions $\sigma^2(k,m,\varepsilon)$ are continuous in ε, so if we take $\varepsilon_n = \frac{h}{\sqrt{n}}$ then we expect that $\sigma_P^2(k,m,\varepsilon)$, $\sigma_{\tilde{P}}^2(k,m,\varepsilon)$, and $\sigma_D^2(k,m,\varepsilon)$ would converge to the asymptotic variances of $\hat{P}_{n,k}(\mathbf{X})$, $\hat{P}_{n,k}(\mathbf{X}) - \hat{p}_n$, and $\hat{D}_{n,k}(\mathbf{X})$ under H_0, respectively. This is verified formally for the case of $m = 1$ in Remark 3.4 and can be shown more generally by tracing the proof of Theorem 3.2, though the details are omitted. Therefore, if $\varepsilon_n = \frac{h}{\sqrt{n}}$, then

$$\sqrt{n}\left(\hat{D}_{n,k}(\mathbf{X}) - \mu_D(k,m,\varepsilon_n)\right) \xrightarrow{d} N\left(0, 2^{k-1}\right), \tag{3.13}$$

where 2^{k-1} is the asymptotic variance of $\hat{D}_{n,k}(\mathbf{X})$ under H_0, given by Theorem 2.1. Let $\phi_D(k,m,h)$ denote the limit of $\frac{\sqrt{n}\mu_D(k,m,\varepsilon_n)}{\sqrt{2^{k-1}}}$ and $z_{1-\alpha}$ be the $1-\alpha$ quantile of the standard normal distribution. The power of the test that rejects when $\frac{\sqrt{n}\hat{D}_{n,k}(\mathbf{X})}{\sqrt{2^{k-1}}} > z_{1-\alpha}$ where under the Markov Chain model

	m			
k	1	2	3	4
1	$2h$	h	$\frac{h}{2}$	$\frac{h}{4}$
2	$\sqrt{2}h$	$\sqrt{2}h$	$\frac{h}{\sqrt{2}}$	$\frac{h}{2\sqrt{2}}$
3	h	h	h	$\frac{h}{2}$
4	$\frac{h}{\sqrt{2}}$	$\frac{h}{\sqrt{2}}$	$\frac{h}{\sqrt{2}}$	$\frac{h}{\sqrt{2}}$

Table 1: Value of $\phi_D(k,m,h)$ For Small Values of k and m

Notes: Table displays the limit as n grows to infinity of the $\sqrt{n}$ scaled ratio of the mean $\mu_D(k,m,\varepsilon_n)$ and the standard deviation $\sigma_D(k,m,\varepsilon_n)$ of the asymptotic distribution of $\hat{D}_{n,k}$ under the Markov Chain alternatives considered in Section 3.2 for m and k between 1 and 4. An explicit expression for $\mu_D(k,m,\varepsilon_n)$ is given in the proof of Theorem 3.3. We consider local perturbations $\varepsilon_n = \frac{h}{\sqrt{n}}$ which imply that $\sigma_D(k,m,\varepsilon_n)$ converges to the asymptotic variance of $\hat{D}_{n,k}(\mathbf{X})$ under H_0.

considered in Theorem 3.3 is given by

$$\mathbb{P}\left(\frac{\sqrt{n}\hat{D}_{n,k}(\mathbf{X})}{\sqrt{2^{k-1}}} > z_{1-\alpha}\right) = \mathbb{P}\left(\sqrt{n}\left(\frac{\hat{D}_{n,k}(\mathbf{X})}{\sqrt{2^{k-1}}} - \frac{\mu_D(k,m,\varepsilon_n)}{\sqrt{2^{k-1}}}\right) > z_{1-\alpha} - \frac{\sqrt{n}\mu_D(k,m,\varepsilon_n)}{\sqrt{2^{k-1}}}\right)$$
$$\to 1 - \Phi\left(z_{1-\alpha} - \phi_D(k,m,h)\right). \tag{3.14}$$

■

Table 1 displays the values of $\phi_D(k,m,h)$ for m and k between 1 and 4. The tests that reject for large values of $\hat{D}_{n,k}(\mathbf{X})$ for $k = m$ have the largest power against the alternative where streakiness begins after m consecutive successes or failures. The test that rejects for large values $\hat{D}_{n,k}(\mathbf{X})$ for $k = 1$ against the alternative with $m = 1$ has the largest power over any combination of the test statistics and alternatives that we consider. Thus, when we present results measuring finite-sample power, the power of the test using $\hat{D}_{n,k}(\mathbf{X}_i)$ for $k = 1$ against the alternative with $m = 1$ gives an upper bound to the power of any of the hypothesis tests against any model of streakiness that we consider.

Figure 1 displays the power for the permutation test rejecting at level 0.05 for large values of $\hat{D}_{n,k}(\mathbf{X})$ for k between 1 and 4 and $n = 100$ against the model considered in Theorem 3.3 with $m = 1$ over a grid of ε.[7] The solid lines display the power of each test measured with a simulation, drawing and implementing the tests on 2,000 replicates of sequences for each value of ε. The

[7]Most shooters take 100 shots in the experiment considered in GVT and MS. Three shooters take 90, 75, and 50 shots, respectively.

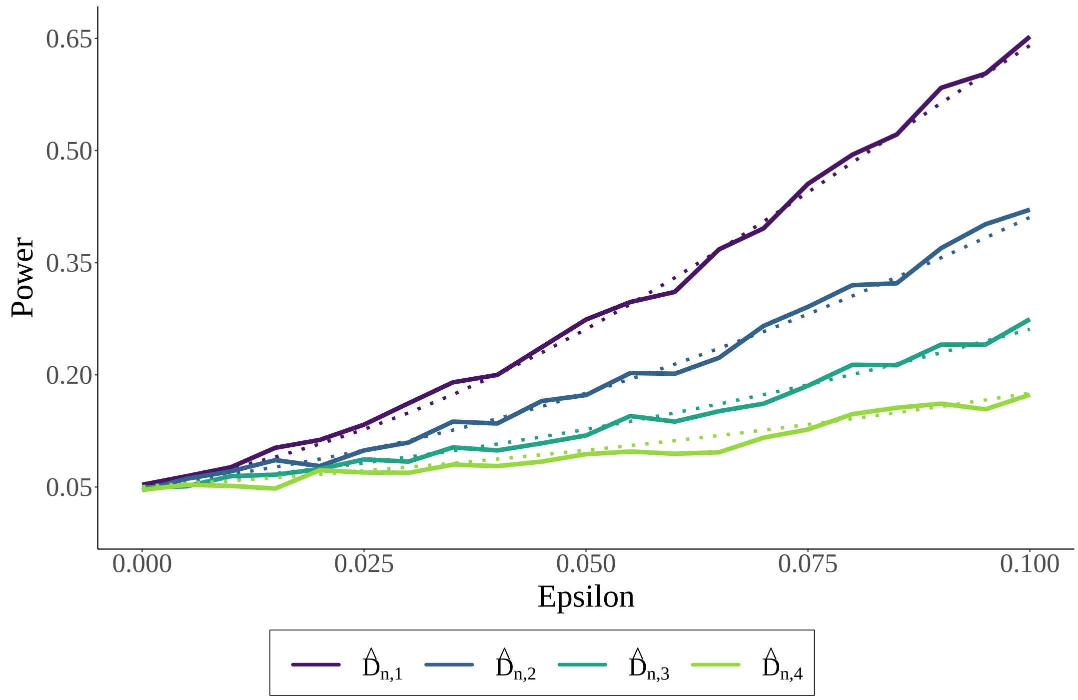

Figure 1: Power Curve for Permutation Test Rejecting for Large $\hat{D}_{n,k}(\mathbf{X})$

Notes: Figure displays the power for the permutation test rejecting at level α for large values of $\hat{D}_{n,k}(\mathbf{X})$ for a range of ε in the alternative given by (3.7), $n = 100$, and each k in $1, \ldots, 4$. The solid lines display the power measured by a simulation, taking the proportion of $2,000$ replications which reject at the 5% level for each value of ε. The dashed lines display the power calculated by the analytic approximation given by (3.14).

dashed lines display the power approximated with the asymptotic expression derived in equation (3.14).

The finite-sample simulation and asymptotic approximation results are remarkably close. The permutation test rejecting for large values of $\hat{D}_{n,1}(\mathbf{X})$ has the largest power, and in fact, in the following section we show that it is asymptotically equivalent to the uniformly most powerful unbiased test.

3.3 Asymptotic Equivalence to the Wald-Wolfowitz Runs Test

The Wald-Wolfowitz Runs Test (Wald and Wolfowitz 1940) rejects for small values of the number of runs R, or equivalently, for large values of

$$Z_n = \left(\frac{\frac{-R}{2n} + \hat{p}_n (1 - \hat{p}_n)}{\hat{p}_n (1 - \hat{p}_n)} \right). \tag{3.15}$$

As shown in Wald and Wolfowitz (1940), under i.i.d. Bernoulli trials, $\sqrt{n}Z_n \xrightarrow{d} N(0,1)$, so the runs test may use either $z_{1-\alpha}$ or a critical value determined exactly from the permutation distribution. Note that the runs test is known to be the uniformly most powerful unbiased test against the Markov Chain alternatives considered in Section 3.2; see Lehmann and Romano (2005), Problems 4.29–4.31. The following Theorem shows the runs test is asymptotically equivalent to the test based on $\hat{D}_{n,1}(\mathbf{X})$.

Theorem 3.4. *The Wald Wolfowitz Runs Test and the test based on $\hat{D}_{n,1}(\mathbf{X})$ are asymptotically equivalent in the sense that they reach the same conclusion with probability tending to one, both under the null hypothesis and under contiguous alternatives. In particular, we show the following:*
(i) *Under i.i.d Bernoulli trials,*

$$\sqrt{n}\left(\hat{D}_{n,1}(\mathbf{X}) - Z_n\right) \xrightarrow{P} 0. \tag{3.16}$$

Therefore, if both statistics are applied using $z_{1-\alpha}$ as a critical value, they both lead to the same decision with probability tending to one.
(ii) *Since (3.16) implies the same is true under contiguous alternatives to Bernoulli sampling (for some p), the same conclusion holds.*
(iii) *The same conclusion holds if $z_{1-\alpha}$ is replaced by critical values obtained by the permutation distribution.*
(iv) *Both tests have the same local limiting power functions under some sequence of contiguous alternatives, and in particular, under the Markov Chain model considered in Section 3.2, where the limiting local power function is given in Remark 3.4.*

Remark 3.6. The permutation test based on the standardized first sample autocorrelation divided by the sample variance, which is not known to have any optimality properties for binary data, is equivalent to the permutation test based on $\sum_{j=1}^{n} X_i X_{i+1}$ by the invariance of the sample mean and variance under permutations. In turn, the permutation test based on $\sum_{j=1}^{n} X_i X_{i+1}$ is asymptotically equivalent to the permutation test based on $\hat{P}_{n,1}(\mathbf{X})$; See Wald and Wolfowitz (1943). It also follows from (C.5) in the Online Appendix that the test based on $\hat{P}_{n,1}(\mathbf{X}) - \hat{p}_n$ and $\hat{D}_{n,1}(\mathbf{X})$ are asymptotically equivalent. Therefore, the permutation tests based on Z_n, $\hat{D}_{n,1}(\mathbf{X})$, $\hat{P}_{n,1}(\mathbf{X}) - \hat{p}_n$, and the first sample autocorrelation are asymptotically equivalent and Theorem 3.4 can be applied to any of the four tests. Miller and Sanjurjo (2019) note this approximate equivalence. Their results are not asymptotic and are based on an approximate algebraic equivalence supported by simulation of correlations between the various test statistics. ■

3.4 Asymptotic Behavior of the Permutation Distribution in Non I.I.D. Settings

Previously, we considered the permutation distribution for various test statistic sequences $T = \{T_n\}$. The permutation distribution itself is random, but depends only on the number of successes in the data set. Under i.i.d. Bernoulli trials, the number of successes is binomial. We now wish to study the behavior of the permutation distribution in possibly non i.i.d. settings (such as the Markov Chain models considered in Section 3.2). But first, we will study the behavior of the permutation distribution for fixed (nonrandom) sequences of number of successes, in which case the permutation distribution is not random, but its limiting distribution is nontrivial.

In order to do this, the following notation is useful. Let $L_n(h)$ be the permutation distribution based on a data set of length n with

$$S_n = S_n(h) = \lfloor \frac{n}{2} + h\sqrt{n} \rfloor \tag{3.17}$$

successes and $n - S_n$ failures, where $\lfloor x \rfloor$ denotes the largest integer less than or equal to x . So, for a given h, $S_n(h)$ is the greatest integer less than or equal to $n/2 + h\sqrt{n}$. Note that if S_n is an integer between 0 and n, then $h = n^{-1/2}\left(S_n - \frac{n}{2}\right)$. Note $L_n(0)$ is then the permutation distribution when you observe $n/2$ successes in n trials if n is even (and $(n-1)/2$ successes if n is odd). We wish to derive the limiting distribution of $L_n(0)$.

The claim is that when $T_n = \hat{D}_{n,1}(\mathbf{X})$, given by (1.3), $L_n(0)$ converges in distribution to $N(0,1)$. In fact, $L_n(h_n)$ has the same limit whenever $h_n \to h$ for some finite h. Note that the permutation distribution $\hat{R}_n(\cdot)$ previously considered for i.i.d. sampling can be expressed as $L_n\left(\hat{h}_n\right)$, where

$$\hat{h}_n = n^{-1/2}\left(S_n - \frac{n}{2}\right), \tag{3.18}$$

and S_n is the number of successes in n Bernoulli trials.

We can now prove a theorem for the behavior of the permutation distribution for the statistic $\hat{D}_{n,1}(\mathbf{X})$ under nonrandom sequences. Note that, if h_n is nonrandom, so is $L_n(h_n)$ and the limit result then does not require any probabilistic qualification (such as convergence in probability or almost surely).

Theorem 3.5. *Assume $h_n \to h$. Let $L_n(h_n)$ be the permutation distribution based on $\lfloor \frac{n}{2} + \sqrt{n}h_n \rfloor$ number of successes (and the remaining failures). Equivalently, if S_n is the number of successes at* *time n, then assume $n^{-1/2}\left(S_n - \frac{n}{2}\right) \to h$. Then,*

$$L_n(h_n) \xrightarrow{d} N(0,1). \tag{3.19}$$

Remark 3.7. The argument generalizes if h_n is defined to be the permutation distribution based on $\lfloor np + \sqrt{n}h_n \rfloor$ number of successes, so that the fixed number of successes at time n, S_n, satisfies $n^{-1/2}(S_n - np) \to h$. ■

Corollary 3.1. *The same argument generalizes to $\hat{D}_{n,k}(\mathbf{X})$ or $\hat{P}_{n,k}(\mathbf{X}) - \hat{p}_n$ for general k. Rather than $N(0,1)$ as the limit, one gets the same unconditional limiting distribution for these statistics as would be obtained under i.i.d. sampling.*

It also follows that we can derive the behavior of the permutation distribution for non i.i.d. processes, such as the Markov Chains considered in Section 3.2.

Theorem 3.6. *Suppose that $X_1, X_2, \dots$ is a possibly dependent stationary Bernoulli sequence. Let $\hat{S}_n$ denote the number of successes in n trials. Assume, for some $p \in (0,1)$, $\sqrt{n}\left(\hat{S}_n - np\right)$ converges in distribution to some limiting distribution. Then, the permutation distribution for $\hat{D}_{n,1}(\mathbf{X})$ converges to $N(0,1)$ in probability; that is*

$$\sup_t |\hat{R}_n(t) - \Phi(t)| \xrightarrow{P} 0. \tag{3.20}$$

Remark 3.8. In the Markov Chain model considered in Section 3.2. we know from the proof of Theorem 3.2 that

$$n^{-1/2}\left(\hat{S}_n - n/2\right) \xrightarrow{d} N\left(0, \frac{1}{4} + \frac{\varepsilon}{1-2\varepsilon}\right), \tag{3.21}$$

and so Theorem 3.6 applies. More generally, the assumption that $n^{-1/2}\left(\hat{S}_n - np\right)$ converges in distribution can be weakened to the assumption that $\mathbf{X}$ is an α-mixing process, as the former condition follows from the latter assumption by Theorem 1.7 of Ibragimov (1962). ■

4 Multiple and Joint Hypothesis Testing

The previous sections dealt with the statistical properties of the statistics $\hat{G}_{n,k}(\mathbf{X}_i)$, applied to an individual i, given a choice of $G \in \{D, P, Q\}$. The GVT controlled shooting experiment measures shooting sequences $\mathbf{X}_i$ for several individuals. In this section, we outline standard methods for testing the hypotheses H_0^i simultaneously and for testing the joint hypothesis H_0 that all sequences $\mathbf{X}_i$ are i.i.d.

We then extend our asymptotic approximation of the power of tests of the individual hypotheses H_0^i to tests of the joint hypothesis H_0 against alternatives in which each of N individuals has probability θ of being drawn from the Markov Chain alternative considered in Section 3.2 under a specified ε. We present simulations of the finite-sample power of the joint hypothesis testing procedures to verify the quality of our approximations. These results significantly reduce the computational burden of power analyses in future experiments.

4.1 Multiple Hypothesis Testing Procedures

In this section, we outline two standard multiple hypothesis testing procedures that can be applied to test the individual hypotheses H_0^i simultaneously. The application of these procedures allow for valid inference in the sense that the familywise error rate (FWER) is controlled at the nominal level α.

Let ρ_i denote a p-value for the a test of H_0^i. In our application, we compute p-values by comparing the observed statistics to their permutation distributions. However, the multiple testing procedures presented in this section are valid for any method of computing p-values that exhibits rigorous type 1 error control. Let the p-values ordered from lowest to highest be $\rho_{(1)},\dots,\rho_{(N)}$ with associated hypotheses $H_0^{(1)},\dots,H_0^{(N)}$, j be the minimal index such that

$$\rho_{(j)} > \frac{\alpha}{m+1-j},$$

and l be the maximal index such that

$$\rho_{(l)} \leq \frac{\alpha}{m+1-l}.$$

First, the Bonferroni-Šidák procedure rejects H_0^i for each i such that $\rho_i \leq \left(1-(1-\alpha)^{1/N}\right)$. The Bonferroni-Šidák procedure is marginally more powerful than the canonical Bonferroni procedure, but can fail to control the FWER if there is negative dependence between tests. In the setting considered in this paper, the GVT basketball shooting experiment, the tests are independent, so the Bonferroni-Šidák procedure is justified.

Second, the Hochberg (1988) Step-Up procedure rejects all $H_0^{(i)}$ with $(i) < (l)$. The Hochberg

Step-Up procedure is more powerful than the related Holm (1979) step-down procedure, but can fail to control the FWER if there is negative dependence between tests. Again, as the tests in the setting considered in this paper are independent, the Hochberg Step-Up procedure is justified.

4.2 Joint Hypothesis Testing Procedures

In this section, we consider four procedures for testing the joint hypothesis H_0 using a single test statistic $\hat{G}_{n,k}(\mathbf{X}_i)$, given a choice of $G \in \{D, P, Q\}$ and a value of k. We then present two methods for testing H_0 which combine several tests of H_0 using different test statistics into one overall test statistic.

4.2.1 Combining Results for Several Individuals with One Statistic

The four procedures that test the joint hypothesis H_0 using a single statistic $\hat{G}_{n,k}(\mathbf{X}_i)$ combined across individuals are as follows:

Average Value of $\hat{G}_{n,k}(\mathbf{X}_i)$: The first procedure rejects for large values of the average of the appropriately centered mean of the test statistic over individuals $\bar{G}_k$. Specifically,
$\bar{D}_k = N^{-1}\sum_{i=1}^{N} D_{n,k}(\mathbf{X}_i)$, $\bar{P}_k = N^{-1}\sum_{i=1}^{N} \hat{P}_{n,k}(\mathbf{X}_i) - \hat{p}_{n,i}$, and $\bar{Q}_k = N^{-1}\sum_{i=1}^{N} \hat{Q}_{n,k}(\mathbf{X}_i) - (1-\hat{p}_{n,i})$. MS implement this procedure and approximate the critical values of the test rejecting for large $\bar{G}_k$ with a normal approximation and with a stratified permutation procedure wherein each individual's observed sequence of trials is permuted separately. We will refer to the distribution of a statistic computed on each of the permuted replicates of each individual's sequence of trials as the stratified permutation distribution. In each stratified permutation, $\bar{G}_k$ is computed over all individuals with $\hat{G}_{n,k}(\mathbf{X}_i)$ defined.

Minimum p-value: Let $\rho_G(k,i)$ denote the p-value for individual i for a test of the hypothesis H_0^i which rejects for extreme values of $\hat{G}_{n,k}(\mathbf{X}_i)$. The minimum p-value joint hypothesis testing procedure rejects for small values of $\hat{\psi}_{G,k} = \min_{1\le i\le N}(\rho_G(k,i))$. The critical values of the test rejecting for small values of $\hat{\psi}_{G,k}$ can be approximated by the stratified permutation distribution of $\hat{\psi}_{G,k}$.[8]

[8]One can also use the Bonferroni-Šidák critical value outlined in Section 4.1, so permutation is not strictly necessary.

Fisher's Method: The Fisher joint hypothesis test statistic (Fisher 1925) is given by $\hat{f}_{G,k} = -2\sum_i \log(\rho_G(k,i))$. If $\rho_G(k,i)$ are p-values for independent tests, then $\hat{f}_{G,k}$ has a chi-squared distribution with $2 \cdot N$ degrees of freedom under H_0. However, we need to account for the fact that $\hat{D}_{n,k}(\mathbf{X}_i)$, $\hat{P}_{n,k}(\mathbf{X}_i)$, and $\hat{Q}_{n,k}(\mathbf{X}_i)$ can be undefined for some sequences. By assigning a p-value of 1 to these sequences, the critical values of the test rejecting for large values of $\hat{f}_{G,k}$ can be approximated with the stratified permutation distribution of $\hat{f}_{G,k}$.

Tukey's Higher Criticism: The Tukey Higher Criticism test statistic is given by

$$\hat{T}_{G,k} = \max_{0<\delta<\delta_0} [T_\delta] = \max_{0<\delta<\delta_0} \left[\frac{\sqrt{\tilde{N}}\,(\xi_\delta - \delta)}{\sqrt{\delta\,(1-\delta)}} \right], \tag{4.1}$$

where

$$\xi_\delta = \tilde{N}^{-1} \mathbb{I}\{\rho_G(k,i) \leq \delta\} \tag{4.2}$$

is the fraction of individuals that are significant at level δ for a given test of H_0^i rejecting for large values of $\hat{G}_{n,k}(\mathbf{X}_i)$, $\tilde{N}$ is the number of individuals for which $\hat{G}_{n,k}(\mathbf{X}_i)$ is defined, δ_0 is a tuning parameter, and $\mathbb{I}\{\cdot\}$ is the indicator function. Again, critical values of the test rejecting for large values of $\hat{T}_{G,k}$ can be approximated with the stratified permutation distribution of $\hat{T}_{G,k}$. See Donoho and Jin (2004) for further discussion. MS implement binomial tests (Clopper and Pearson 1934) that reject for large proportions of significant individuals. A binomial test chooses a specified threshold of significance δ, and rejects H_0 at level α if the number of individuals significant at level δ exceeds the $1-\alpha$ quantile of the distribution of a binomial variable with parameters N and δ. Tukey's Higher Criticism is a refinement of this testing procedure that allows for a data-driven choice of the significance threshold δ.

4.2.2 Combining the Results of Several Joint Test Statistics

The results of any of the procedures that test the joint hypothesis for a single test statistic can be combined with the results from tests using different test statistics with Fisher's method or by computing the minimum p-value. Specifically, let $\rho_G(k)$ be the p-value of a test of the joint null using the test statistic $\hat{G}_{n,k}(\mathbf{X}_i)$ for $G \in \{D,P,Q\}$ and k in $1,\ldots,K$. The Fisher test statistic is given by

$$\hat{\mathbf{F}} = -2\log \sum_{G\in\{D,P,Q\}} \sum_{k=1}^{K} \rho_G(k) \tag{4.3}$$

and the minimum p-value test statistic is given by

$$\hat{\Psi} = \min\left\{\rho_G(k) \,|\, G \in \{D,P,Q\}, 1 \leq k \leq K\right\}. \tag{4.4}$$

The critical values for the tests rejecting for large values of $\hat{\mathbf{F}}$ and small values of $\hat{\Psi}$ can be approximated with the stratified permutation distribution of $\hat{\mathbf{F}}$ and $\hat{\Psi}$, respectively.

4.3 Power of Joint Hypothesis Testing Procedures

In this section, we obtain an approximation to the the power of the joint hypothesis testing methods presented in Section 4.2 against alternatives in which each of the N individuals independently follow the Markov model with $m = 1$, studied in Section 3.2, with probability θ and H_0^i with probability $1-\theta$. We then implement a set of comprehensive simulations to verify the quality of our approximation.

Recall that the power of the test that rejects for large values of $\hat{D}_{n,k}(\mathbf{X}_i)$ with $k = 1$ against the Markov model with $m = 1$ gives an upper bound to the power of any of the tests against any of the alternatives that we consider. Therefore, in order to attain an upper bound on the power of the joint hypothesis tests, we confine our theoretical analysis to statistics that combine values of $\hat{D}_{n,k}(\mathbf{X}_i)$ for $k = 1$ across individuals under the Markov model with $m = 1$.

We present analytic results for the power of the test using $\bar{D}_1$. However, we note that the asymptotic results presented in Sections 2 and 3 allow for computationally efficient measurement of the finite sample power of the minimum p-value, Fisher, and Tukey's Higher Criticism joint hypothesis tests by drawing from the normal asymptotic distributions of the individual test statistics. This procedure has a significantly lower computational burden than directly simulating Bernoulli sequences $\mathbf{X}_i$ and computing permutation distributions for each draw of a joint test statistic. In the simulations that follow, we implement the latter more expensive simulation approach to highlight the finite sample quality of our asymptotic approximation.

Theorem 4.1. *For each i in $1,\ldots,N$, let B_i be a Bernoulli(θ) random variable and let $\mathbf{X}_i^0 = \left\{X_{ij}^0\right\}_{j=1}^n$ and $\mathbf{X}_i^1 = \left\{X_{ij}^1\right\}_{j=1}^n$ be two-state Markov Chains on $\{0,1\}$ following transition matrices*

$$\begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} \quad \textit{and} \quad \begin{bmatrix} \frac{1}{2}+\varepsilon_{n,N} & \frac{1}{2}-\varepsilon_{n,N} \\ \frac{1}{2}-\varepsilon_{n,N} & \frac{1}{2}+\varepsilon_{n,N} \end{bmatrix}$$

where $\varepsilon_{n,N} = \frac{h}{\sqrt{nN}}$, respectively. Assume that B_i, $\mathbf{X}_i^0$, and $\mathbf{X}_i^1$ are mutually independent for all i in $1, \ldots, N$. The random variable

$$\bar{D}_1 = N^{-1} \sum_{i=1}^{N} \left(B_i \hat{D}_{n,1} \left(\mathbf{X}_i^1 \right) + (1 - B_i) \hat{D}_{n,1} \left(\mathbf{X}_i^0 \right) \right) \tag{4.5}$$

is asymptotically normal with limiting distribution given by

$$\sqrt{nN} \bar{D}_1 \xrightarrow{d} N(2\theta h, 1) \tag{4.6}$$

as $n \to \infty$ and $N \to \infty$.

Remark 4.1. It is straightforward to generalize Theorem 4.1 to give the limiting distribution of

$$\bar{D}_k = N^{-1} \sum_{i=1}^{N} \left(B_i \hat{D}_{n,k} \left(\mathbf{X}_i^1 \right) + (1 - B_i) \hat{D}_{n,k} \left(\mathbf{X}_i^0 \right) \right) \tag{4.7}$$

where $\mathbf{X}_i^0$ is i.i.d. Bernoulli(1/2), $\mathbf{X}_i^1$ follows the Markov model of Theorem 3.3 for general m, and B_i is Bernoulli(θ). In this case, the limiting distribution is given by

$$\sqrt{nN} \bar{D}_k \xrightarrow{d} N\left(\sqrt{2^{k-1}} \theta \phi_D(k, m, h), 2^{k-1} \right) \tag{4.8}$$

where $\phi_D(k,m,h)$ is defined in Remark 3.5. ■

Remark 4.2. Under the conditions of Theorem 4.1, the power of the test that rejects when $\sqrt{nN}\bar{D}_1 > z_{1-\alpha}$, where $z_{1-\alpha}$ is the $1-\alpha$ quantile of the standard normal distribution, is given by

$$\begin{aligned} \mathbb{P}\left(\sqrt{nN} \bar{D}_1 > z_{1-\alpha} \right) &= \mathbb{P}\left(\sqrt{nN} \left(\bar{D}_1 - \frac{2\theta h}{\sqrt{nN}} \right) > z_{1-\alpha} - 2\theta h \right) \\ &\to 1 - \Phi(z_{1-\alpha} - 2\theta h). \end{aligned} \tag{4.9}$$

Therefore, the product of the number of individuals N and observations per individuals n should be approximately

$$\left(\frac{z_{1-\alpha} - \Phi^{-1}(1-\beta)}{2\theta\varepsilon} \right)^2. \tag{4.10}$$

in order for a test rejecting for large values of $\bar{D}_1$ to achieve power β against an alternative model specified in Theorem 4.1 for fixed values of ε and θ. ■

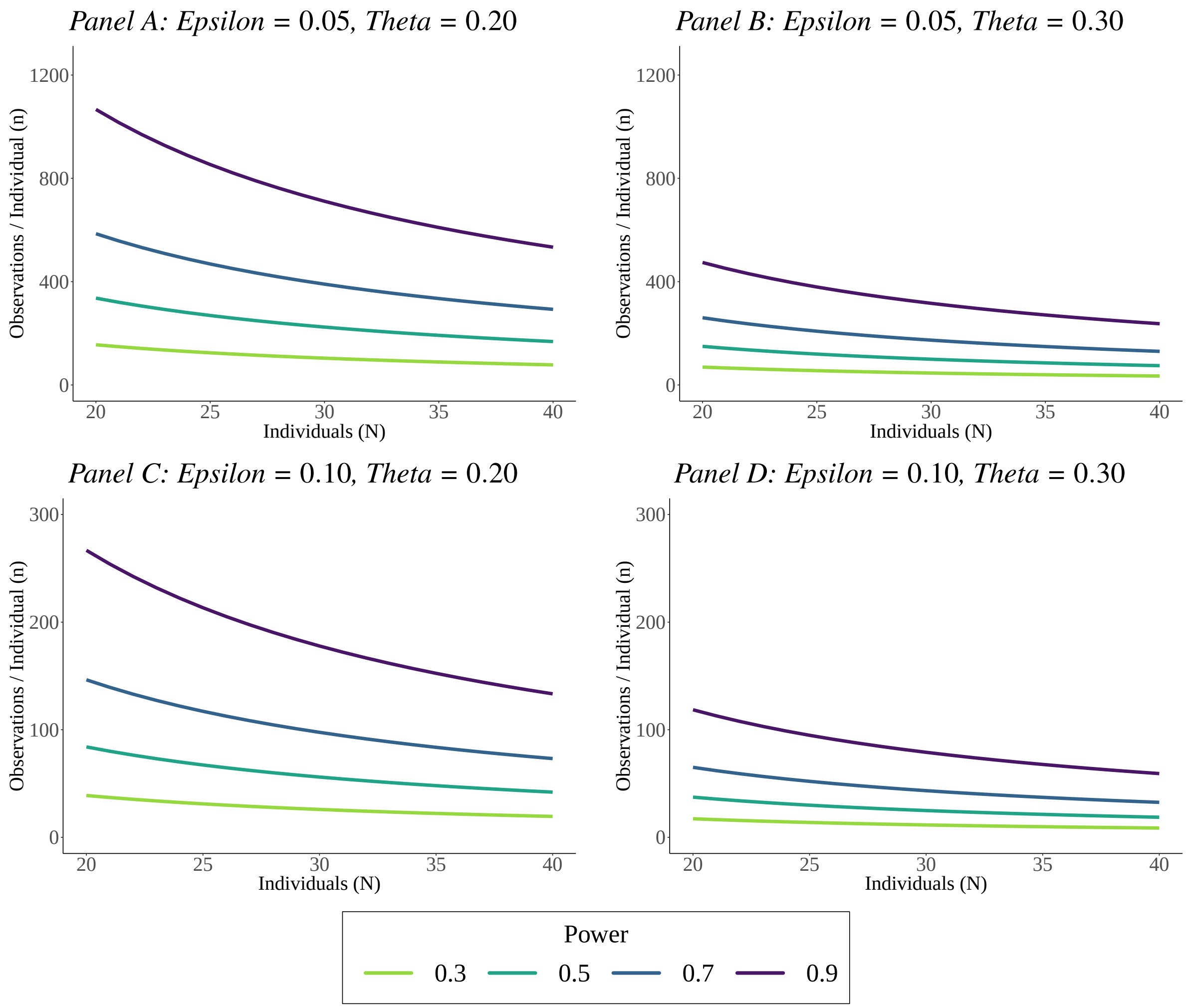

Figure 2: Requisite Sample Size for Power of Tests of the Joint Null Against Specified Alternatives

Notes: Figure displays the power of the test of the joint null H_0 using the test statistic $\bar{D}_1$ against the alternative specified in Theorem 4.1 calculated by the analytic approximation in (4.10). Each panel gives the power for the test in data for different sample sizes n and N under a specified ε and θ.

Figure 2 displays the power of the test rejecting for large values of $\bar{D}_1$ at level $\alpha = 0.05$ against four different parameterizations of the alternative model specified by the conditions of Theorem 4.1 for a grid of values of n and N. Of the four alternatives that we display, an experiment with $n = 100$ and $N = 26$ only has reasonable power against the parameterization $\varepsilon = 0.1$ and $\theta = 0.3$.

Next, we implement a simulation measuring the finite-sample power of the joint hypothesis testing methods presented in Section 4.2 against the alternative specified by the conditions of Theorem 4.1 for a grid of ε and θ. For all simulations, we take $1{,}000$ draws of $N = 26$ individuals, each with $n = 100$ observed trials, under specified values of ε and θ.[9] For each simulated individual, we compute the p-value for the permutation test rejecting for large values of $\hat{D}_{n,k}(\mathbf{X})$ for $k = 1$. For each set of N simulated individuals, we compute each of $\bar{G}_1$, $\hat{\psi}_{D,1}$, $\hat{f}_{D,1}$, and $\hat{T}_{D,1}$ as well as their permutation distributions. Recall that our simulation obtains an estimate of the upper bound on the power of the joint hypothesis testing procedures that we consider.

Figure 3 displays the proportion of replicates in which joint tests using $\hat{D}_{n,k}(\mathbf{X})$ with $k = 1$ reject H_0 at the 5% level over a grid of ε and θ between 0 and 0.1 and 0 and 1, respectively. Note that the tests have insignificant power against alternatives with ε on the order of 0.03, even when the proportion of nonrandom individuals is large.

Panel A of Figure 3 overlays the asymptotic approximation to the power of the test using $\bar{D}_1$ given by (4.9) on the simulation measurement of the finite sample power. The approximation is accurate for most parameterizations of the model, but overestimates the power for parameterizations where the finite sample power is close to 0.9.

Our asymptotic results significantly reduce the computational expense of power analysis in future experiments. The simulations presented in this section require extensive parallelization, as the permutation distributions of the joint statistics of each draw of N individuals need to be computed.[10] In contrast, measuring the minimum n and N necessary to achieve a desired power against a wide range of ε and θ is almost instantaneous with the analytic approximation in (4.10).

[9] There are 26 individuals who participate in the GVT controlled shooting experiment. For all but three individuals, we observe 100 shots. We simulate 100 draws for each individual and so compute a slight upper bound to the power of the tests that we consider.

[10] The simulation utilizes 2,600 nodes, each equipped with 15 cores. If the script were run in serial, it would take approximately five years and six months to run to completion.

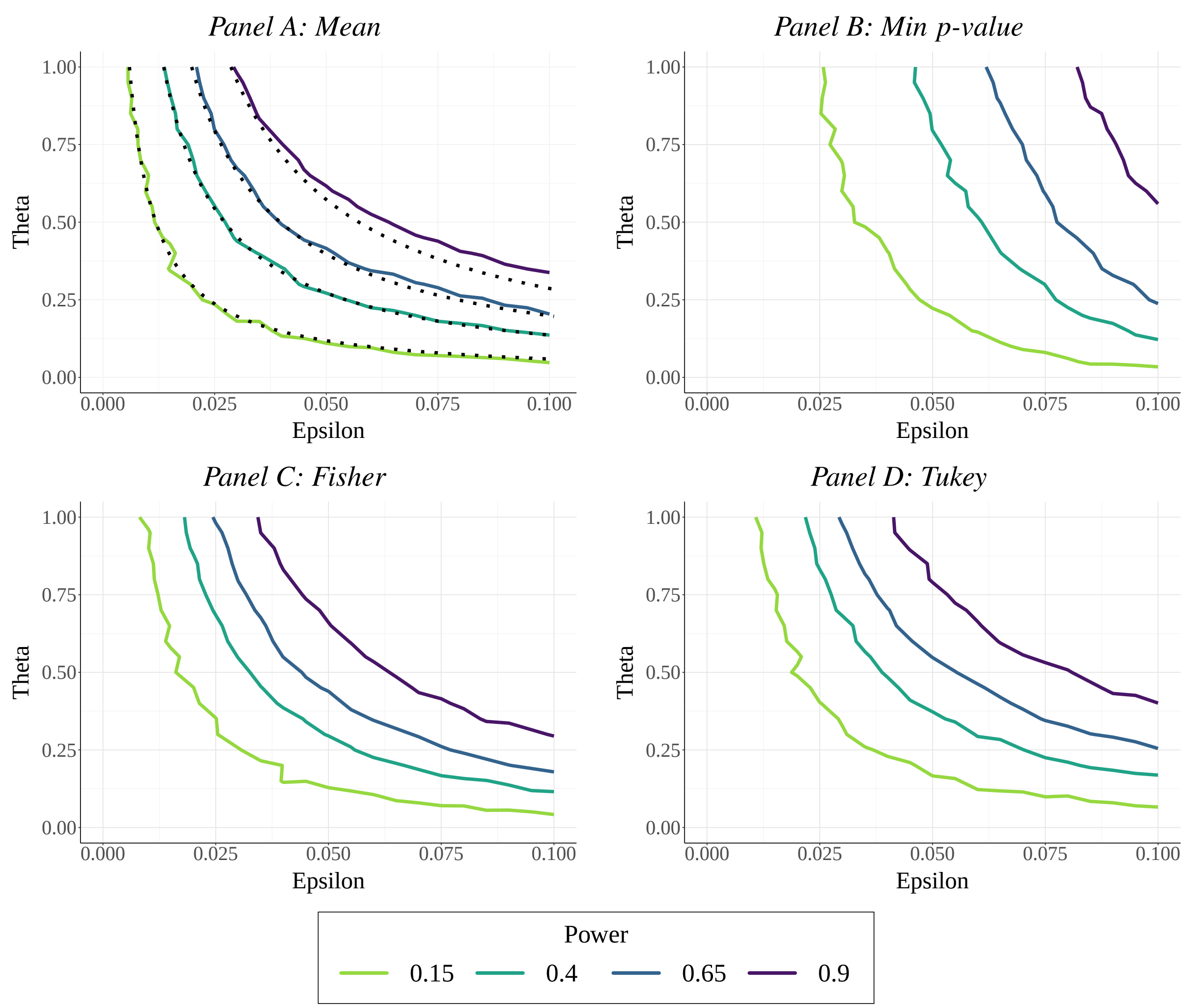

Figure 3: Contours of the Power Surfaces of Tests of the Joint Null

Notes: Figure displays contours of the power surface for tests of the joint null H_0, which use the test statistic $\hat{D}_{n,k}(\mathbf{X})$ for $k = 1$ against the alternative where each individual's sequence $\mathbf{X}_i$ follows the Markov Chain given by (3.7) for different values ε with probability θ and follows H_0^i with probability $1-\theta$. At each grid point, we simulate 1000 replicates of $N = 26$ Bernoulli sequences $\mathbf{X}_i$ of length $n = 100$. The dashed line in Panel A display the power calculated by the analytical approximation given by (4.9).

5 Uncertainty in the Hot Hand Fallacy

Are beliefs about the hot hand stronger than what is observed in reality? In the preceding three sections, we established the validity and measured the power of the tests considered by GVT and MS and are now equipped to evaluate the implications of the GVT controlled shooting experiment for this question.

It is important to distinguish between three distinct components of this question. First, does basketball shooting deviate from an i.i.d. process? Second, how streaky is basketball shooting on average? Finally, and substantively, do people systematically overestimate the positive dependence in basketball shooting?

We begin by outlining the details of the GVT controlled shooting experiment, before analyzing these three questions in succession. In the Appendix, we give a detailed overview and replication of the results of GVT and MS, studying normal approximation and permutation tests of the individual hypotheses H_0^i and permutation tests of the joint hypothesis H_0.

5.1 The GVT Controlled Shooting Experiment

We observe shooting sequences for 26 members of the Cornell University men and women's varsity and junior varsity basketball teams.[11] Fourteen of the players are men, and twelve of the players are women. For all but three players, we observe 100 shots. We observe 90, 75, and 50 shots for three of the men. The experimenters determined distances from which each player's shooting percentage was approximately 50% and placed two arcs 60 degrees from the baseline on the left and right hand sides of the basket. Each individual took 50% of their shots from each side of the basket. The experiment was incentivized.[12]

5.2 Does Basketball Shooting Deviate from an I.I.D. Process?

The first question, whether all basketball shooting is random, can be assessed by simultaneously testing the hypotheses H_0^i that each shooter's sequence is random. Table 2 displays the number of rejections of H_0^i at level $\alpha = 0.05$ when the p-values from the one-sided individual shooter

[11] We obtained the data from https://www.econometricsociety.org/sites/default/files/14943_Data_and_Programs.zip on April 19, 2019.

[12] Prior to each shot, each individual chose whether to bet "high" or "low." If the individual bets high (low) and makes the shot they win 5 (2) cents. If the individual bets high (low) and misses the shot they lose 4 (1) cents. To the best of our knowledge, the betting data are not publicly available.

	k			
	1	2	3	4
$\hat{D}_{n,k}(\mathbf{X}_i)$	1	2	1	0
$\hat{P}_{n,k}(\mathbf{X}_i)$	1	1	1	0
$\hat{Q}_{n,k}(\mathbf{X}_i)$	1	0	0	0

Table 2: Number of Rejections of H_0^i Under Two Multiple Hypothesis Testing Procedures

Notes: Table displays the number of rejections of H_0^i at level $\alpha = 0.1$. for each of $\hat{D}_{n,k}(\mathbf{X}_i)$, $\hat{P}_{n,k}(\mathbf{X}_i)$, or $\hat{Q}_{n,k}(\mathbf{X}_i)$ and each k in $1,\ldots,4$ corrected with both the Bonferroni-Šidák and Hochberg Step-Up multiple hypothesis testing procedures implemented on the p-values from the one-sided permutation test.

permutation tests are corrected with both the Bonferroni-Šidák and Hochberg Step-Up multiple hypothesis testing procedures. Both give the same results.[13] Both procedures consistently reject H_0^i for only one shooter, identified as "Shooter 109," over the set of test statistics considered.

The rejection of H_0^i for Shooter 109 for most test statistics, robust to standard multiple hypothesis testing corrections, is strong evidence that some basketball players exhibit streaky shooting some of the time. The large extent to which Shooter 109 deviates from randomness is emphasized by Panel A of Figure 4, which plots his sequence of makes and misses. Shooter 109 begins by missing 9 shots in a row. Shortly thereafter, he makes 16 out of 17 shots, followed by a sequence where he misses 15 out of 18 shots and a sequence where he makes 16 shots in a row.

It is unlikely that a random Bernoulli sequence would generate this pattern, even among $N = 26$ sequences.[14] Panel B of Figure 4 plots the permutation distribution of $\hat{D}_{n,1}(\mathbf{X}_i)$ for Shooter 109's shooting sequence, superimposing the observed value of $\hat{D}_{n,1}(\mathbf{X}_i)$ with a vertical black line and our asymptotic approximation to this distribution with a black curve. The p-value of the individual permutation test using $\hat{D}_{n,1}(\mathbf{X}_i)$ for shooter 109 is given by the proportion of permutations that are to the right of the observed value.

It is also unlikely, however, that the streakiness exhibited by Shooter 109 is indicative of what should be expected from a representative basketball player. Figure 5 displays histograms and empirical distribution functions of the field goal and free throw shooting percentages of NBA

[13] The results are identical if we take $\alpha = 0.1$.

[14] GVT observe that the rejection of the individual hypothesis H_0^i of Shooter 109 is significant, but neither GVT nor MS consider the multiple testing problem.

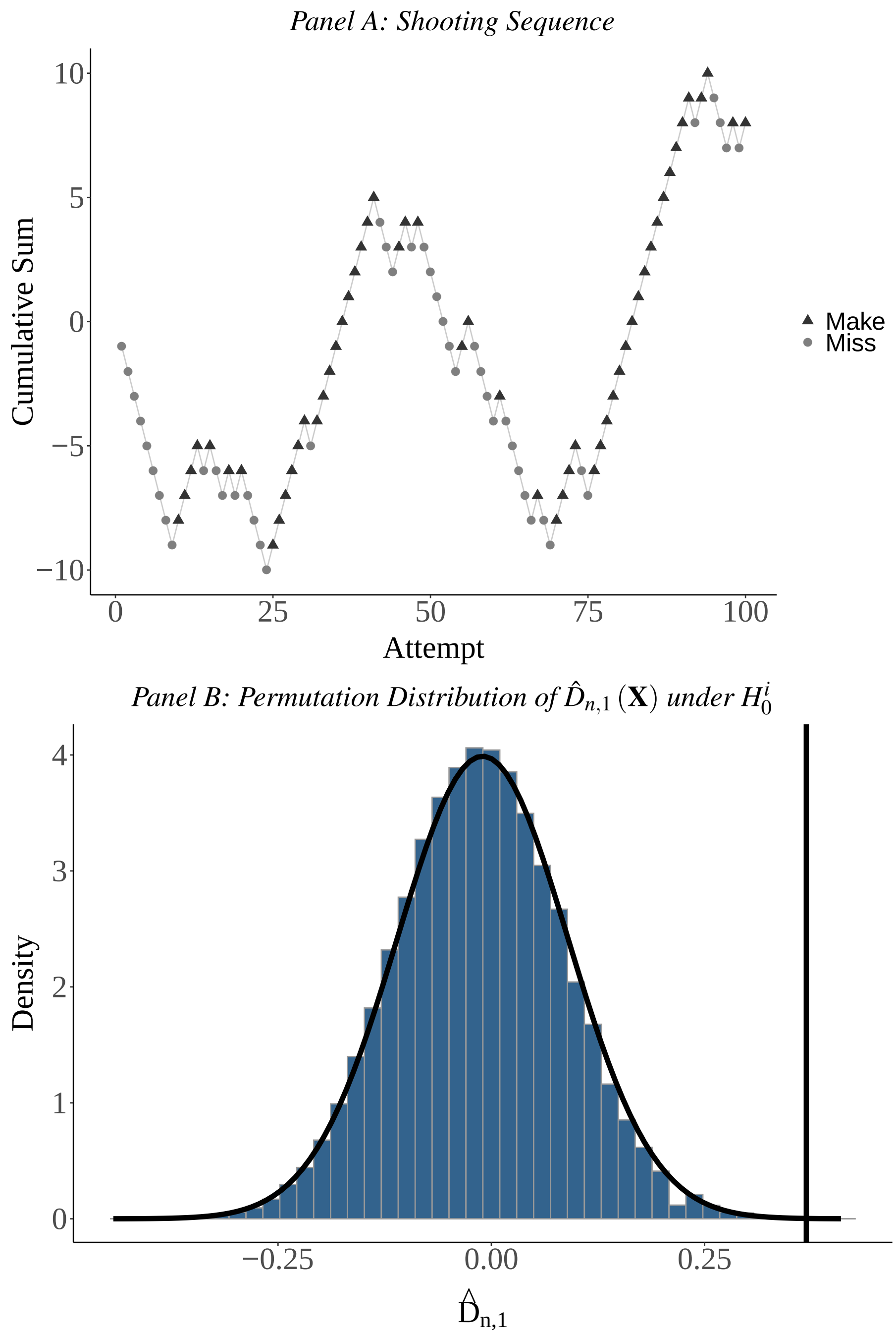

Figure 4: Shooter 109 Shooting Sequence and Permutation Distribution

Notes: Panel A displays the cumulative sum of the sequence of makes and misses for Shooter 109. Made baskets are coded as a 1 and displayed with a black triangle and missed baskets are coded as a -1 and displayed as a grey circle. Panel B displays a density histogram of $\hat{D}_{n,1}(\mathbf{X}_i)$ computed for $100,000$ permutations of shooter 109's observed shooting sequence. The observed value of $\hat{D}_{n,1}(\mathbf{X}_i)$ is displayed with a vertical black line. The density histogram is superimposed with $N\left(\beta_D(n,k,\hat{p}_i), n^{-1}\sigma_D^2(\hat{p}_i,k)\right)$, which is the asymptotic approximation for the permutation distribution of $\hat{D}_{n,1}(\mathbf{X}_i)$ derived in Theorem 2.2, where $\sigma_D^2(\hat{p}_i,k)$ is given in the statement of Theorem 2.1, shifted by a Monte Carlo approximation for the small-sample bias $\beta_D(n,k,\hat{p}_i)$ discussed in the Appendix.

players in the 2018–2019 regular season.[15] The x-axes of the empirical distribution function plots have been relabelled such that the medians of the distributions are displayed as 0 and ε corresponds to the difference, in terms of shooting percentage, between the x-axis positions and the medians. The value of the bias-corrected-difference in the proportion of successful shots after making and missing the previous shot, $\hat{D}_{n,k}(\mathbf{X}_i) - \beta_D(n,k,\hat{p}_{n,i})$ for $k = 1$, for Shooter 109 is 0.38, analogous to an ε of 0.19 in the Markov model of streakiness developed in Section 3.2 with $m = 1$.[16] An ε of this size is equivalent to varying between shooting at a rate similar to the best or worst shooter in the NBA, depending on whether a shooter made or missed their previous shot.

5.3 How Streaky is Basketball Shooting on Average?

The conclusion that basketball shooting is not always random for all shooters is insufficient in evaluating whether people believe that there is a stronger hot hand than is observed in reality We would prefer to have an estimate of the streakiness of an average or representative shooter, and optimally some sense of the variation in streakiness over shooters.

However, measurement of the magnitude of the average streakiness of basketball shooting can only be addressed if reasonable deviations from randomness can be detected. The tests studied in this paper implemented on the GVT shooting data do not have sufficient power to detect parameterizations of the Markov model of streaky shooting consistent with the variation in NBA shooting percentages.

Suppose that 50% of players shoot at a rate equivalent to the 75$^{\text{th}}$or 25$^{\text{th}}$ percentile of the distribution of field goal percentage of NBA players after making a shot or missing a shot, respectively. This is parameterized as $\varepsilon = 0.038$ and $\theta = 0.5$. We think of this parameterization as an upper bound on the set of deviations from randomness consistent with the variation in NBA shooting percentages. Under this parameterization, the test rejecting for large values of $\bar{D}_1$ has a power of

[15]The data were downloaded from https://www.basketball-reference.com/leagues/NBA_2019_totals.html#totals_stats::fg_pct and https://www.basketball-reference.com/leagues/NBA_2019_totals.html#totals_stats::ft_pct on July 16, 2019. Following the minimum requirements established by www.basketball-reference.com, the free throw sample includes players who have attempted more than 125 free throws and the field goal sample includes players who have attempted more than 300 field goals.

[16]The bias $\beta_D(n,k,\hat{p}_{n,i})$ is the parametric bootstrap estimate of expected value of $\hat{D}_{n,k}(\mathbf{X}_i)$ under randomness $\mathbb{E}_{H_0^i}\left[\hat{D}_{n,k}(\mathbf{X}_i)\right]$. MS show that this term is negative and substantial in the sample sizes we consider. See Appendix A.1 for further discussion.

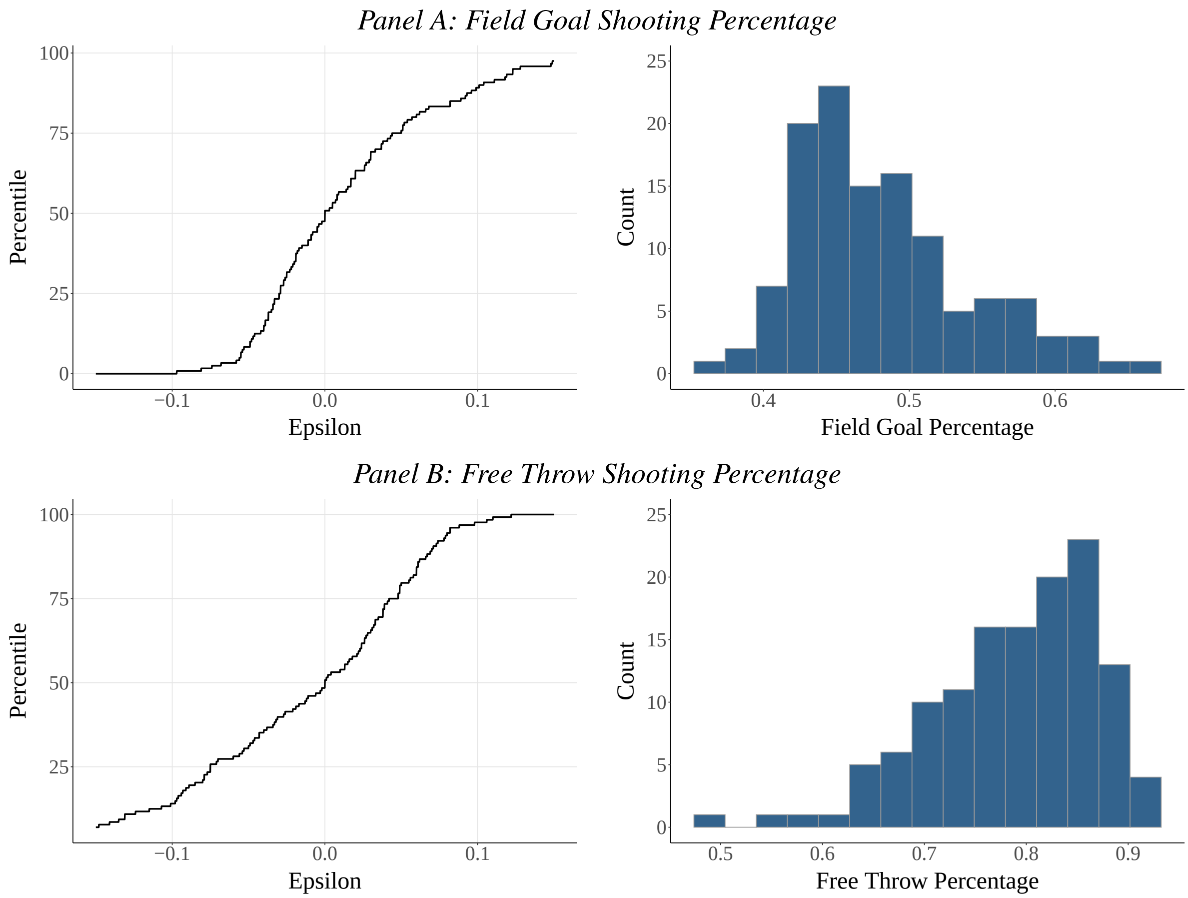

Figure 5: Distribution of Field Goal and Free Throw Shooting Percentages in the 2018-2019 NBA Season

Notes: Figure displays the distributions of the field goal and free throw shooting percentages of NBA players in the 2018–2019 regular season. Players shooting fewer than 300 field goals or 125 free throws are omitted when displaying the distributions of field goal and free throw shooting percentage, respectively. Panels A and B display a truncated empirical cumulative distribution and a histogram of the field goal and free throw shooting percentages, respectively. To parallel the model developed in Section 3.2, in both panels the x-axis of the truncated cumulative distribution is transformed such that the median is displayed as 0, and ε corresponds to the difference, in terms of shooting percentage, between the x-axis position and the median. The median free throw shooting percentage is 80.6%, and the median field goal shooting percentage is 46.7%.

only 0.62.[17,18]

Moreover, any evidence of positive dependence appears to be confined to Shooter 109. Table 3 presents the p-values for the four tests of H_0 outlined in Section 4 implemented with each test statistic $\hat{D}_{n,k}(\mathbf{X}_i)$, $\hat{P}_{n,k}(\mathbf{X}_i)$, and $\hat{Q}_{n,k}(\mathbf{X}_i)$ for each k between 1 and 4. The majority of tests using individual test statistics reject H_0 at the 5% level. The Fisher test statistic $\hat{\mathbf{F}}$, specified in (4.3), is highly significant for the test using the means of the test statistics, for Tukey's Higher Criticism, and for the test using the minimum p-value. $\hat{\mathbf{F}}$ is significant at the 10% level for the test using the Fisher test statistic. The minimum p-value test statistic $\hat{\mathbf{\Psi}}$ is highly significant for all four tests.

The rejection of H_0 at the 5% level is not robust to the exclusion of Shooter 109 from the sample. Table 3 also displays the p-values for the tests of H_0 implemented without the inclusion of Shooter 109 in the sample. Now, at most three of the p-values for tests of H_0 using a single test statistic for each method of testing the joint null are significant at the 5% level. $\hat{\mathbf{F}}$ and $\hat{\mathbf{\Psi}}$ are no longer significant at the 5% level for tests using the means of the test statistics over shooters and Tukey's Higher Criticism and are no longer significant at the 10% level for tests using the minimum p-value and Fisher's test statistic.

To conclude, the GVT shooting experiment is insufficiently powered to detect deviations from randomness that we argue would be consistent with a realistic parameterization of positive dependence in basketball shooting. We are therefore unable to provide an informative estimate of the mean or dispersion of streakiness in basketball shooting. This conclusion could be challenged by a strong and robust rejection of H_0, but the rejection of H_0 is sensitive to inclusion of an outlier. This result cuts both ways. The data are insufficient to make strong statements about the magnitude of positive dependence in basketball shooting, either small or substantial.

[17]The expression for the limiting power function (3.14) indicates that tests of the individual hypotheses are even more under-powered. With a sample size of 100 shots, $\varepsilon = 0.038$, and $m = 1$, the power for the test using $\hat{D}_{n,1}(\mathbf{X}_i)$ is equal to 0.19. Even if the sample size were increased to 300 shots the power is only 0.37. A sample size of approximately 1050 shots is required for a power of 0.8.

[18]The controlled shooting experiment studied in Miller and Sanjurjo (2019) is powered to detect this alternative. In their "panel" they observe three sessions of eight players who shoot 300 shots, giving the test using $\bar{D}_1$ a power of 0.95 against the alternative we consider. Similarly, they observe five sessions of 300 shots for one player that they identified as particularly streaky in the first session. We caution that this parameterization is an upper bound to the set of models consistent with NBA shooting percentages and that, likewise, the power of the test using $k = 1$ against the alternative model with $m = 1$ was chosen as an upper bound to the power of any of the tests we consider. As indicated in Figures 1, 2, and 3 and in Remark 4.1 power is significantly attenuated for less extreme parameterizations, for tests using larger values of k, and for alternatives using larger values of m.

	k	Mean $\bar{G}_k$		Min. p-value $\hat{\psi}_{G,k}$		Fisher $\hat{f}_{G,k}$		Tukey HC $\hat{T}_{G,k}$	
		w/ 109	w/o 109	w/ 109	w/o 109	w/109	w/o 109	w/109	w/o 109
$\hat{D}_k(\mathbf{X}_i)$	1	0.1464	0.3678	0.0030	0.4940	0.0428	0.3654	0.1079	0.2165
	2	0.0402	0.1261	0.0010	0.0354	0.0021	0.0648	0.2834	0.4096
	3	0.0036	0.0125	0.0204	0.1279	0.0021	0.0213	0.0483	0.0678
	4	0.0716	0.1294	0.1404	0.1346	0.0054	0.0165	0.0079	0.0213
$\hat{P}_k(\mathbf{X}_i)$	1	0.1548	0.3520	0.0008	0.4868	0.0150	0.2939	0.0096	0.0417
	2	0.0323	0.0879	0.0021	0.1273	0.0047	0.0911	0.0001	0.5405
	3	0.0418	0.0882	0.0131	0.3555	0.0385	0.2470	0.2445	0.3299
	4	0.3035	0.4095	0.2721	0.2642	0.1690	0.3337	0.5154	0.6729
$\hat{Q}_k(\mathbf{X}_i)$	1	0.1492	0.3917	0.0035	0.4916	0.0631	0.4340	0.0571	0.2777
	2	0.1891	0.3446	0.1356	0.2796	0.1398	0.3679	0.0113	0.2764
	3	0.0126	0.0259	0.1539	0.1458	0.0279	0.0459	0.0877	0.1502
	4	0.0361	0.0555	0.3360	0.3238	0.0543	0.0566	0.0126	0.0097
$\hat{\mathbf{F}}$		68.6796	49.6754	96.2418	36.5842	93.0266	51.2112	82.2148	46.7255
p-value		0.0191	0.0746	0.0002	0.1342	0.0849	0.1339	0.0019	0.0672
$\hat{\Psi}$		0.0036	0.0125	0.0008	0.0354	0.0021	0.0165	0.0001	0.0097
p-value		0.0271	0.0828	0.0041	0.21097	0.0907	0.1509	0.0016	0.0880

Table 3: Tests of the Joint Null Hypothesis H_0
with and without Shooter 109

Notes: Table displays the p-values for four tests of the joint null hypothesis H_0 for $\hat{D}_k(\mathbf{X}_i)$, $\hat{P}_k(\mathbf{X}_i)$, or $\hat{P}_k(\mathbf{X}_i)$ and each k in $1,\ldots,4$ with and without the inclusion of shooter 109. The minimum p-value procedure, Fisher joint hypothesis testing procedure, and Tukey's Higher Criticism procedure use the p-values from the one-sided individual shooter permutation test. We choose $\delta_0 = 0.5$ for computing $\hat{T}_{G,k}$. The p-values for all four procedures are estimated by permuting each shooter's observed shooting sequence $100,000$ times, computing the test statistics for each set of permuted shooting sequences, and computing the proportion of test statistics greater than or equal to the observed test statistics. We compute Fisher's statistic $\hat{\mathbf{F}}$ for all four procedures by taking -2 times the log of the sum of the p-values for each $\hat{D}_k(\mathbf{X}_i)$, $\hat{P}_k(\mathbf{X}_i)$, or $\hat{Q}_k(\mathbf{X}_i)$ and each k in $1,\ldots,4$. We compute the minimum p-value statistic $\hat{\Psi}$ for all four procedures by taking the minimum of the p-values for each $\hat{D}_k(\mathbf{X}_i)$, $\hat{P}_k(\mathbf{X}_i)$, or $\hat{Q}_k(\mathbf{X}_i)$ and each k in $1,\ldots,4$. The p-values for $\hat{\mathbf{F}}$ and $\hat{\Psi}$ are computed by estimating the stratified permutation distributions of $\hat{\mathbf{F}}$ and $\hat{\Psi}$.

5.4 Do People Overestimate Positive Dependence in Basketball Shooting?

If we had an informative estimate of the streakiness of an average shooter, a comparison with evidence on expectations of streakiness in basketball shooting would provide a direct test of the hot hand fallacy. We find that the available evidence on expectations of streakiness in basketball shooting suffers either from methodological flaws or is not directly comparable to estimates of streakiness.

GVT measure beliefs by implementing a survey of one hundred basketball fans from Cornell and Stanford. The fans were asked to consider a hypothetical basketball player who makes 50% of their shots. The average expected field goal percentages for this player after having just made and missed a shot were 61% and 42%, respectively. Similarly, when asked to consider a hypothetical player who makes 70% of shots from the free throw line, fans expected that the average free throw percentages for second free throws after having made and missed the first were 74% or 66%, respectively.

Taken at face value, the surveys can be interpreted as eliciting expectations of ε when $m = 1$ and indicating that these expectations are approximately 0.1 and 0.04, respectively. An ε of 0.1 can be detected with probability close to one for reasonably large proportions of streaky shooters, but as we argue in Section 5.3, there is insufficient power to detect an ε of 0.04. If the surveys were credible, they would provide mixed evidence that people systematically overestimate the streakiness in basketball shooting.

However, there are severe methodological limitations to the GVT survey. First, there is considerable evidence that surveys eliciting beliefs about hypothetical events can be prone to substantial bias (Harrison and Rutström 2008). Second, the results may be biased by framing (Tversky and Kahneman 1981) – that is, the language of the survey questions may be suggestive of positive dependence.

Optimally, we would be able to infer beliefs from observations of incentivized decisions. In our review of the literature, we are unable to find estimates of beliefs in streakiness that directly translate to estimates of people's expectations of ε for any m. GVT allow players to bet on whether they will make or miss their next shot. GVT and Miller and Sanjurjo (2017) analyze these wagers. Both conclude that players believe that the probability of success is greater after a make than a miss, although they disagree in their assessment of the quality of the player's predictions. Nevertheless, these wagers do not provide an estimate of ε. Similarly, Rao (2009), Bocskocsky et. al (2014), and Lantis and Nesson (2019) explore shot selection and defensive pressure in NBA games, and find that players behave as if they believe that the probability of a make is higher after a streak of makes than after a streak of misses. Again, these studies do not provide an estimate of beliefs directly comparable to estimates of the streakiness in basketball shooting.

Future studies should estimate expectations of streakiness that are directly comparable to measurements of streakiness. Manski (2004) advocates for surveys of probabilistic expectations in non-hypothetical settings. Data from surveys of this form have been valuable in informing behavioral models of expectation formation in financial markets (Greenwood and Shleifer 2014, Barberis et. al 2015). We support a design in which an observer of a shooter in a controlled shooting experiment is asked to record their expectation of the probability that the shooter makes their next shot prior to each shot. If the shot is made, the observed is rewarded for submitting large probabilities and punished for submitting small probabilities. If the shot is missed, the converse is true. This design would not suffer from the framing bias of the GVT survey and would directly provide an estimate of ε.

6 Conclusion

The purpose of this paper is to clarify and quantify the uncertainty in the empirical support for the tendency to perceive streaks as overly representative of positive dependence – the hot hand fallacy. Following Gilovich, Vallone, and Tversky (1985), the results of a class of tests of randomness implemented on data from a basketball shooting experiment have provided central empirical support for textbook models of misperception of randomness. The results and conclusions of these tests were called into question by Miller and Sanjurjo (2018), who observe that there is a substantial small sample bias in the test statistics that had been applied. We evaluate the implications, limitations, and interpretation of these tests by establishing their validity, approximating their power, and re-evaluating their application to the Gilovich, Vallone, and Tversky (1985) shooting experiment.

Our theoretical and simulation analyses show that the tests considered are insufficiently powered to detect effect sizes consistent with the observed variation in NBA shooting percentages with high probability. Substantially larger data sets are required for informative estimates of the streakiness in basketball shooting. We are able to reject i.i.d. shooting consistently for only one participant in the experiment. This rejection is robust to standard multiple testing corrections, providing strong evidence that basketball shooting is not perfectly random. However, evidence against randomness is limited to this player.

Future research should directly test the accuracy of people's predictions of streakiness in stochastic processes and should be implemented in settings with reasonable power against sensible alternatives. We provide a mathematical and statistical theory to serve as a foundation for future analyses with this objective. Our analytic power approximations significantly reduce the computational burden of power analyses in the design of these studies. Additionally, we contribute an emphasis on the differentiation of individual, simultaneous, and joint hypothesis testing that can more clearly delineate the conclusions and limitations of inferences on deviations from randomness.

References

Albright, S.C., 1993. A statistical analysis of hitting streaks in baseball. *Journal of the American Statistical Association*, 88(424), pp.1175-1183.

Appelbaum, Binyamin. Streaks Like Daniel Murphy's Aren't Necessarily Random. *The New York Times*, The New York Times, 27 Oct. 2015, www.nytimes.com/2015/10/27/upshot/trust-your-eyes-a-hot-streak-is-not-a-myth.html.

Bar-Hillel, M. and Wagenaar, W.A., 1991. The perception of randomness. *Advances in Applied Mathematics*, 12(4), pp.428-454.

Barberis, N., Greenwood, R., Jin, L. and Shleifer, A., 2015. X-CAPM: An extrapolative capital asset pricing model. *Journal of Financial Economics*, 115(1), pp.1-24.

Benjamin, D.J., 2018. Errors in probabilistic reasoning and judgment biases (No. w25200). *National Bureau of Economic Research*.

Bhat, U.N. and Miller, G.K., 2002. Elements of applied stochastic processes (Vol. 3). Hoboken, NJ: Wiley-Interscience.

Bocskocsky, A., Ezekowitz, J. and Stein, C., 2014. The hot hand: A new approach to an old 'fallacy'. In 8th Annual MIT Sloan Sports Analytics Conference (pp. 1-10).

Bradley, R.C., 2005. Basic properties of strong mixing conditions. A survey and some open questions. *Probability Surveys*, 2, pp.107-144.

Carhart, M.M., 1997. On persistence in mutual fund performance. *The Journal of Finance*, 52(1), pp.57-82.

Clopper, C.J. and Pearson, E.S., 1934. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4), pp.404-413.

Cohen, Ben. The 'Hot Hand' Debate Gets Flipped on Its Head. *The Wall Street Journal*, Dow Jones & Company, 1 Oct. 2015, www.wsj.com/articles/the-hot-hand-debate-gets-flipped-on-its-head-1443465711.

Fama, E.F., 1965. The behavior of stock-market prices. *The Journal of Business*, 38(1), pp.34-105.

Gilovich, T., Vallone, R. and Tversky, A., 1985. The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology,* 17(3), pp.295-314.

Greenwood, R. and Shleifer, A., 2014. Expectations of returns and expected returns. *The Review of Financial Studies*, 27(3), pp.714-746.

Haberstroh, Tom. He's Heating up, He's on Fire! Klay Thompson and the Truth about the Hot Hand. *ESPN*, ESPN Internet Ventures, 12 June 2017, www.espn.com/nba/story/_/page/presents-19573519/heating-fire-klay-thompson-truth-hot-hand-nba.

Harrison, G.W. and Rutström, E.E., 2008. Experimental evidence on the existence of hypothetical bias in value elicitation methods. Handbook of experimental economics results, 1, pp.752-

767.

Hendricks, D., Patel, J. and Zeckhauser, R., 1993. Hot hands in mutual funds: Short-run persistence of relative performance, 1974–1988. *The Journal of Finance*, 48(1), pp.93-130.

Hochberg, Y., 1988. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4), pp.800-802.

Holm, S., 1979. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pp.65-70.

Ibragimov, I.A., 1962. Some limit Theorems for stationary processes. *Theory of Probability & Its Applications*, 7(4), pp.349-382.

Jensen, M.C., 1968. The performance of mutual funds in the period 1945–1964. *The Journal of Finance*, 23(2), pp.389-416.

Johnson, George. Gamblers, Scientists and the Mysterious Hot Hand. *The New York Times*, The New York Times, 17 Oct. 2015, www.nytimes.com/2015/10/18/sunday-review/gamblers-scientists-and-the-mysterious-hot-hand.html.

Kahneman, Daniel. *Thinking, Fast and Slow*. Macmillan, 2011.

Korb, K.B. and Stillwell, M., 2003. The story of the hot hand: Powerful myth or powerless critique. I*nternational Conference on Cognitive Science*.

Künsch, H.R., 1989. The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, pp.1217-1241.

Lantis, R.M. and Nesson, E.T., 2019. Hot Shots: An Analysis of the Hot Hand in NBA Field Goal and Free Throw Shooting (No. w26510). National Bureau of Economic Research.

Lahiri, S.N., 2013. Resampling methods for dependent data. Springer, NY.

Lehmann, Erich L., *Nonparametrics; Statistical Methods Based on Ranks*. 1998, Prentice Hall.

Lehmann, Erich L., and Joseph P. Romano. *Testing Statistical Hypotheses*. 2005, Springer, NY.

Liu, R.Y. and Singh, K., 1992. Moving blocks jackknife and bootstrap capture weak dependence. *Exploring the Limits of Bootstrap*, 225, p.248.

Malkiel, B.G., 2003. The efficient market hypothesis and its critics. *Journal of Economic Perspectives*, 17(1), pp.59-82.

Malkiel, B.G. and Fama, E.F., 1970. Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2), pp.383-417.

Manski, C.F., 2004. Measuring expectations. *Econometrica*, 72(5), pp. 1329-1376

Marcus, R., Eric, P. and Gabriel, K.R., 1976. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3), pp.655-660.

Miller, J.B. and Sanjurjo, A., 2017. A Visible (Hot) Hand? Expert Players Bet on the Hot Hand and Win. Expert Players Bet on the Hot Hand and Win. University of Alicante mimeo.

Miller, J.B. and Sanjurjo, A., 2018. Surprised by the hot hand fallacy? A truth in the law of small numbers. *Econometrica*, 86(6), pp.2019-2047.

Miller, Joshua, and Adam Sanjurjo. Momentum Isn't Magic--Vindicating the Hot Hand with the Mathematics of Streaks. *Scientific American*, 28 Mar. 2018b, www.scientificamerican.com/article/momentum-isnt-magic-vindicating-the-hot-hand-with-the-mathematics-of-streaks/.

Miller, J.B. and Sanjurjo, A., 2019. A cold shower for the hot hand fallacy. University of Alicante mimeo.

Miyoshi, H., 2000. Is the "hot-hands" phenomenon a misperception of random events?. *Japanese Psychological Research*, 42(2), pp.128-133.

Mood, A.M., 1940. The distribution theory of runs. *The Annals of Mathematical Statistics*, 11(4), pp.367-392.

Politis, D.N. and Romano, J.P., 1994. The stationary bootstrap. *Journal of the American Statistical Association*, 89(428), pp.1303-1313.

Politis, D.N., Romano, J.P. and Wolf, M., 1999. *Subsampling*. Springer, NY.

Rabin, M., 2002. Inference by believers in the law of small numbers. *The Quarterly Journal of Economics*, 117(3), pp.775-816.

Rabin, M. and Vayanos, D., 2010. The gambler's and hot-hand fallacies: Theory and applications. *The Review of Economic Studies*, 77(2), pp.730-778.

Rao, J.M., 2009. Experts' perceptions of autocorrelation: The hot hand fallacy among professional basketball players. Unpublished technical manuscript. San Diego, CA.

Remnick, David. Bob Dylan and the 'Hot Hand.' *The New Yorker*, The New Yorker, 19 June 2017, www.newyorker.com/culture/cultural-comment/bob-dylan-and-the-hot-hand.

Rinott, Y., 1994. On normal approximation rates for certain sums of dependent random variables. *Journal of Computational and Applied Mathematics*, 55(2), pp.135-143.

Stein, C., 1986. Approximate computation of expectations. IMS. Hayward, CA.

Stern, H.S. and Morris, C.N., 1993. A statistical analysis of hitting streaks in baseball: Comment. *Journal of the American Statistical Association*, 88(424), pp.1189-1194.

Stone, D.F., 2012. Measurement error and the hot hand. *The American Statistician*, 66(1), pp.61-66.

Romano, J.P., Shaikh, A. and Wolf, M., 2011. Consonance and the closure method in multiple testing. *The International Journal of Biostatistics*, 7(1), pp.1-25.

Romano, J.P. and Wolf, M., 2005. Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4), pp.1237-1282.

Tversky, A. and Kahneman, D., 1971. Belief in the law of small numbers. *Psychological Bulletin*, 76(2), p.105.

Tversky, A. and Kahneman, D., 1981. The framing of decisions and the psychology of choice.

Science, 211(4481), pp.453-458.
Wald, A. and Wolfowitz, J., 1940. On a test whether two samples are from the same population. *The Annals of Mathematical Statistics*, 11(2), pp.147-162.
Wald, A. and Wolfowitz, J., 1943. An exact test for randomness in the non-parametric case based on serial correlation. *The Annals of Mathematical Statistics*, 14(4), pp.378-388.
Wardrop, R.L., 1999. Statistical tests for the hot-hand in basketball in a controlled setting. *American Statistician*, 1, pp.1-20.

A Appendix: Overview and Replication of GVT and MS

To test the individual shooter hypotheses H_0^i, GVT and MS choose the test statistic $\hat{D}_{n,k}(\mathbf{X}_i)$. MS show formally that, while $\hat{D}_{n,k}(\mathbf{X}_i)$ converges to 0 in probability as n increases, the expectation of $\hat{D}_{n,k}(\mathbf{X}_i)$ for finite n is strictly less than 0, and argue numerically that this difference can be substantial for the sample sizes considered in the GVT shooting data. MS argue that if the GVT analysis is corrected to account for the small-sample bias, the results are reversed, and there is evidence for significant deviations from randomness.

A.1 Normal Approximation Confidence Intervals

For their main analyses, GVT and MS use tests that rely on normal approximations to the distributions of $\hat{D}_{n,k}(\mathbf{X}_i)$, $\hat{P}_{n,k}(\mathbf{X}_i)$, and $\hat{Q}_{n,k}(\mathbf{X}_i)$. We have shown in Section 2 that these tests control the probability of a type 1 error asymptotically.

We are interested in giving confidence intervals for $\gamma_{\tilde{P}}(\mathbb{P},k)$, $\gamma_{\tilde{Q}}(\mathbb{P},k)$, and $\gamma_D(\mathbb{P},k)$, where

$$\begin{aligned}
\gamma_{\tilde{P}}(\mathbb{P},k) &= \mathbb{P}\left(X_{j+k}=1|X_{j+k-1}=1,\ldots,X_j=1\right)-\mathbb{P}\left(X_j=1\right),\\
\gamma_{\tilde{Q}}(\mathbb{P},k) &= \mathbb{P}\left(X_{j+k}=0|X_{j+k-1}=0,\ldots,X_j=0\right)-\mathbb{P}\left(X_j=0\right), \text{ and}\\
\gamma_D(\mathbb{P},k) &= \mathbb{P}\left(X_{j+k}=1|X_{j+k-1}=1,\ldots,X_j=1\right)-\mathbb{P}\left(X_{j+k}=1|X_{j+k-1}=0,\ldots,X_i=0\right).
\end{aligned}$$

Recall from Remark 3.3 that $\hat{P}_{n,k}(\mathbf{X}_i)-\hat{p}_{n,i}\pm\hat{\sigma}_{\tilde{P}}(\hat{p}_{n,i},k)\frac{z_{1-\alpha/2}}{\sqrt{n}}$, $\hat{Q}_{n,k}(\mathbf{X}_i)-(1-\hat{p}_{n,i})\pm\hat{\sigma}_{\tilde{Q}}(\hat{p}_{n,i},k)\frac{z_{1-\alpha/2}}{\sqrt{n}}$, and $\hat{D}_{n,k}(\mathbf{X}_i)\pm\hat{\sigma}_{\tilde{D}}(\hat{p}_{n,i},k)\frac{z_{1-\alpha/2}}{\sqrt{n}}$ are asymptotically valid confidence intervals for $\gamma_{\tilde{P}}(\mathbb{P},k)$, $\gamma_{\tilde{Q}}(\mathbb{P},k)$, and $\gamma_D(\mathbb{P},k)$ under stationary alternatives contiguous to H_0^i, respectively.

MS depart from GVT by correcting for finite-sample bias. Specifically, let $\beta_P(n,k,p_i)$ denote $\mathbb{E}_{H_0^i}\left[\hat{P}_{n,k}(\mathbf{X}_i)-p_i\right]$. Then, $\hat{P}_{n,k}(\mathbf{X}_i)-\hat{p}_{n,i}-\beta_P(n,k,p_i)$ is an unbiased estimator for $\gamma_{\tilde{P}}(\mathbb{P},k)$ and has asymptotic variance equal to $\sigma^2_{\tilde{P}}(p,k)$ under H_0^i. Therefore, $\hat{P}_{n,k}(\mathbf{X}_i)-\hat{p}_{n,i}-\beta_P(n,k,p_i)\pm\sigma^2_{\tilde{P}}(p,k)\frac{z_{1-\alpha/2}}{\sqrt{n}}$ is an asymptotically valid confidence interval for $\gamma_{\tilde{P}}(\mathbb{P},k)$ that likely has improved coverage in finite samples under H_0^i. Likewise, let $\beta_Q(n,k,p_i)$ and $\beta_D(n,k,p_i)$ denote $\mathbb{E}_{H_0^i}\left[\hat{Q}_{n,k}(\mathbf{X}_i)-(1-p_i)\right]$, and $\mathbb{E}_{H_0^i}\left[\hat{D}_{n,k}(\mathbf{X}_i)\right]$. Bias corrected confidence intervals for $\gamma_{\tilde{Q}}(\mathbb{P},k)$, and $\gamma_{\tilde{Q}}(\mathbb{P},k)$ are formed similarly. Note that these parameters are all 0 under the null hypothesis H_0^i.

MS approximate $\beta_P(n,k,p_i)$, $\beta_Q(n,k,p_i)$, and $\beta_D(n,k,p_i)$ using the parametric bootstrap, computing the means of $\hat{D}_{n,k}(\mathbf{X}_i)$, $\hat{P}_{n,k}(\mathbf{X}_i)-\hat{p}_i$, and $\hat{Q}_{n,k}(\mathbf{X}_i)-(1-\hat{p}_i)$ over the many replicates of $\mathbf{X}_i$ drawn as i.i.d. Bernoulli sequences with probability of success $\hat{p}_i$. We denote these estimates

$\beta_D(n,k,\hat{p}_{n,i})$, $\beta_P(n,k,\hat{p}_{n,i})$, and $\beta_Q(n,k,\hat{p}_{n,i})$ for each choice of streak length k and shooter i.[19]

Online Appendix Figures 1, 2, and 3 replicate the GVT and MS results, displaying 95% confidence intervals for $\gamma_{\tilde{P}}(\mathbb{P},k)$, $\gamma_{\tilde{Q}}(\mathbb{P},k)$, and $\gamma_D(\mathbb{P},k)$ for each shooter and streak length $k = 1,\ldots,4$. We estimate the variance of $\hat{P}_{n,k}(\mathbf{X}_i)-\hat{p}_i$, $\hat{Q}_{n,k}(\mathbf{X}_i)-(1-\hat{p}_i)$, and $\hat{D}_{n,k}(\mathbf{X}_i)$ by plugging in each shooter's observed shooting percentage $\hat{p}_{n,i}$ into the respective formulae of the asymptotic variances.[20] The $100\cdot(1-\alpha)\,\%$ confidence intervals for $\gamma_{\tilde{P}}(\mathbb{P},k)$, $\gamma_{\tilde{Q}}(\mathbb{P},k)$, and $\gamma_D(\mathbb{P},k)$ are given by

$$\begin{aligned}
\hat{P}_{n,k}(\mathbf{X}_i)-\hat{p}_{n,i}-\beta_P(n,k,\hat{p}_i) \;&\pm\; t_{n,1-\alpha/2}\left(n^{-1/2}\sigma_{\tilde{P}}(\hat{p}_i,k)\right),\\
\hat{Q}_{n,k}(\mathbf{X}_i)-(1-\hat{p}_{n,i})-\beta_Q(n,k,\hat{p}_i) \;&\pm\; t_{n,1-\alpha/2}\left(n^{-1/2}\sigma_{\tilde{Q}}(1-\hat{p}_i,k)\right),\text{ and}\\
\hat{D}_{n,k}(\mathbf{X}_i)-\beta_D(n,k,\hat{p}_i) \;&\pm\; t_{n,1-\alpha/2}\left(n^{-1/2}\sigma_D(\hat{p}_i,k)\right),
\end{aligned}$$

respectively, where $t_{n,1-\alpha/2}$ denotes the $1-\alpha/2$ quantile of the t distribution with n degrees of freedom.[21]

These confidence interval constructions are valid for parameters where the underlying process $\mathbb{P}$ is i.i.d. or nearly so (i.e., a sequence contiguous to i.i.d.). However, our main application is to determine whether the confidence intervals include 0, corresponding to the true parameters for an i.i.d. process. As mentioned before, for general and potentially non-contiguous stationary sequences, one could apply block resampling methods as a means of confidence interval construction. Given the relatively small sample sizes of the data in the study under consideration, we do not pursue the more general problem of confidence interval construction using the bootstrap. Under the null, the asymptotic variances of the test statistics only depend on the underlying success rate,

[19] The expectations of the permutation distributions of $\hat{P}_k(\mathbf{X}_i)-\hat{p}_i$, $\hat{Q}_k(\mathbf{X}_i)-(1-\hat{p}_i)$, and $\hat{D}_k(\mathbf{X}_i)$ are also consistent estimates of $\gamma_{\tilde{P}}(\mathbb{P},k)$, $\gamma_{\tilde{Q}}(\mathbb{P},k)$, and $\gamma_D(\mathbb{P},k)$. In Online Appendix D, we provide second order approximations to $\gamma_{\tilde{P}}(\mathbb{P},k)$, $\gamma_{\tilde{Q}}(\mathbb{P},k)$, and $\gamma_D(\mathbb{P},k)$ and demonstrate that they perform accurately in small samples. The second order approximations may be computationally advantageous for problems with a large sample size.

[20] Any consistent estimate of p_i can be plugged into the asymptotic variances of $\hat{P}_{n,k}(\mathbf{X}_i)$, $\hat{Q}_{n,k}(\mathbf{X}_i)$, and $\hat{D}_{n,k}(\mathbf{X}_i)$ to produce a set of consistent estimators. This includes $\hat{P}_{n,k}(\mathbf{X}_i)$ for all k. Additionally, the variances can be estimated consistently with the permutation distribution or with the bootstrap. In Online Appendix E, we show that $\hat{P}_{n,k}(\mathbf{X}_i)\left(1-\hat{P}_{n,k}(\mathbf{X}_i)\right)/V_{ik}$ is also a consistent estimator for the asymptotic variance of $\hat{P}_{n,k}(\mathbf{X}_i)$. MS estimate the variance of $\hat{D}_{n,k}(\mathbf{X}_i)$ with

$$\left(\frac{(V_{ik}-1)\,s_{p,i}^2+(W_{ik}-1)\,s_{q,i}^2}{V_{ik}+W_{ik}-2}\right)\left(\frac{1}{V_{ik}}+\frac{1}{W_{ik}}\right) \tag{A.1}$$

where $s_{p,i}^2=\left(\frac{V_{ik}}{V_{ik}-1}\right)\hat{P}_{n,k}(\mathbf{X}_i)\left(1-\hat{P}_{n,k}(\mathbf{X}_i)\right)$ and $s_{q,i}^2=\left(\frac{W_{ik}}{W_{ik}-1}\right)\hat{Q}_{n,k}(\mathbf{X}_i)\left(1-\hat{Q}_{n,k}(\mathbf{X}_i)\right)$. This estimator is typically employed when $\hat{P}_{n,k}(\mathbf{X}_i)$ and $\hat{Q}_{n,k}(\mathbf{X}_i)$ are the sample means of i.i.d. populations assumed to have equal variances. This is not the case in our setting, where the variances of $\hat{P}_{n,k}(\mathbf{X}_i)$ and $\hat{Q}_{n,k}(\mathbf{X}_i)$ are not equal and the covariance of $\hat{P}_{n,k}(\mathbf{X}_i)$ and $\hat{Q}_{n,k}(\mathbf{X}_i)$ is not equal to 0. However, in Online Appendix E, we show that the ratio of (A.1) and the asymptotic variance of $\hat{D}_{n,k}(\mathbf{X}_i)$ converges to 1 in probability.

[21] We use the t quantiles to be consistent with MS, though they are no more justified asymptotically than the normal quantiles.

which is much easier to estimate consistently than the limiting variances in Theorem 3.1.

The 95% confidence intervals for $\gamma_{\tilde{P}}(\mathbb{P},k)$, $\gamma_{\tilde{Q}}(\mathbb{P},k)$, and $\gamma_D(\mathbb{P},k)$ are above 0 for at most 1 shooter for $k=1$, 3 shooters for $k=2$, 4 shooters for $k=3$, and 2 shooters for $k=4$. For each statistic and for k equal to 1 and 2, the bias-corrected estimates of $\gamma_{\tilde{P}}(\mathbb{P},k)$, $\gamma_{\tilde{Q}}(\mathbb{P},k)$, and $\gamma_D(\mathbb{P},k)$ are approximately evenly split above and below 0. For k equals 3 and 4, approximately 60% of the shooters have bias-corrected estimates of $\gamma_{\tilde{P}}(\mathbb{P},k)$, $\gamma_{\tilde{Q}}(\mathbb{P},k)$, and $\gamma_D(\mathbb{P},k)$ greater than 0. MS approximate the variance of $\hat{D}_{n,k}(\mathbf{X}_i)$ differently. For their approximation, the 95% confidence intervals for $\gamma_D(\mathbb{P},k)$ with $k=3$ are above 0 for 5 shooters.

A.2 Tests of Individual Shooter Hypotheses H_0^i with Permutation Tests

We now test the individual hypotheses H_0^i with permutation tests. MS give the results of permutation tests as a robustness check. Permutation tests have the advantage of accounting for finite-sample bias automatically. We remarked in Section 2.2 that permutation tests are the only tests that are exactly level α.

GVT present results for tests using $\hat{D}_{n,k}(\mathbf{X}_i)$ for k in $1,\ldots,3$, and MS present results using $\hat{D}_{n,k}(\mathbf{X}_i)$ for $k=3$ and note that the results for $k=2$ and 4 are consistent. We display results for all tests using $\hat{D}_{n,k}(\mathbf{X}_i)$, $\hat{P}_{n,k}(\mathbf{X}_i)$, and $\hat{Q}_{n,k}(\mathbf{X}_i)$ for k between 1 and 4, as these tests all have maximal power within the class of statistics we consider against different plausible models of hot hand shooting.

Online Appendix Figure 4 overlays the estimates for $\hat{D}_{n,k}(\mathbf{X}_i)$ onto the estimated permutation distributions for each shooter and streak length $k=1,\ldots,4$. Each panel displays the density of the statistics of interest for each shooter over the permutation replications in a white-to-black gradient. The 97.5$^{\text{th}}$ and 2.5$^{\text{th}}$ quantiles of the estimated permutation distributions are denoted by black horizontal line segments. The observed estimates for $\hat{D}_{n,k}(\mathbf{X}_i)$ are denoted by grey horizontal line segments. Online Appendix Figures 5 and 6 show equivalent plots for the tests statistics $\hat{P}_{n,k}(\mathbf{X}_i)-\hat{p}_{n,i}$, and $\hat{Q}_{n,k}(\mathbf{X}_i)-(1-\hat{p}_{n,i})$, respectively. The observed values of $\hat{D}_{n,k}(\mathbf{X}_i)$ are above the 97.5$^{\text{th}}$ quantile of the permutation distribution for 1 shooter for k equal to 1, 3 shooters for k equal to 2 and 4, and 4 shooters for k equal to 3.

Online Appendix Figures 7, 8, and 9 display the p-values of the one-sided permutation tests using $\hat{D}_{n,k}(\mathbf{X}_i)$, $\hat{P}_{n,k}(\mathbf{X}_i)$, and $\hat{Q}_{n,k}(\mathbf{X}_i)$ for each k in $1,\ldots,4$. Under H_0 we would expect the p-values to vary about the black line drawn on the diagonal. Almost all p-values are below the diagonal line. However, relatively few p-values are below the canonical thresholds of 0.05 and 0.1. Roughly, the separation between the p-values and the diagonal line increases with k.

A.3 Tests of the Joint Hypothesis H_0

In this section, we implement the procedures outlined in Section 4 that test the joint null H_0 and enable us to infer whether any shooters deviate from randomness.

The primary evidence MS provide in support of significant hot hand shooting effects are rejections of two tests of the joint null H_0. First, they reject H_0 for the test using $\bar{D}_k$ with $k = 3$, and note that the results for $k = 2$ and 4 are consistent. Second, they perform a set of binomial tests, rejecting for large proportions of individuals significant at the 5% and 50% levels. The binomial tests are sensitive to the choice of the significance thresholds 5% and 50% and Online Appendix Figure 7 Panel C indicates that these choices were fortuitous, in the sense that H_0 is rejected for these choices and not for others. Additionally, when the individual hypotheses H_0^i are tested simultaneously by applying the 5% binomial test, the 50% binomial test, or Tukey's Higher Criticism with the closed testing procedure of Markus et. al. (1976), no individual hypotheses are rejected at the 5% level, including Shooter 109.[22]

Figure 10 overlays the estimates for $\bar{D}_k$, $\bar{P}_k$, and $\bar{Q}_k$ onto the estimated permutation distributions for each streak length $k = 1, \ldots, 4$. Each panel displays the density of the statistics of interest over the permutation replications in a white-to-black gradient. The 97.5^{th} and 2.5^{th} quantiles of the computed permutation distributions are denoted by dark black horizontal line segments. The observed estimates for $\bar{D}_k$, $\bar{P}_k$, and $\bar{Q}_k$ are indicated by grey horizontal line segments. Although each of the observed values of $\bar{D}_k$, $\bar{P}_k$, and $\bar{Q}_k$ are above the means of the respective permutation distributions, only the observed values of $\bar{D}_k$ and $\bar{Q}_k$ for k equals\ to 3 are above the 97.5^{th} quantile of their permutation distributions.

[22]The closed testing procedure rejects an individual hypotheses H_0^i at level α if all possible intersection hypotheses containing H_0^i are rejected by a joint testing procedure at level α. Note that any coherent multiple testing procedure that controls the familywise error rate must arise from a closed testing procedure; see Theorem 2.1 of Romano et. al (2011).