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

David M. Ritzwoller
Stanford University

Joseph P. Romano*
Stanford University

April 1, 2021

Abstract

We study a class of permutation tests of the randomness of a collection 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 permutation tests of the null hypothesis 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 characterize the asymptotic distributions of these test statistics and their permutation distributions under randomness, under a set of general stationary processes, and under a class of Markov chain alternatives, which allow us to derive their local asymptotic power. The results are applied to evaluate the empirical support for the hot hand fallacy provided by four controlled basketball shooting experiments. We establish that substantially larger data sets are required to derive an informative measurement of the deviation from randomness in basketball shooting. In one experiment, for which we were able to obtain data, multiple testing procedures reveal that one shooter exhibits a shooting pattern significantly inconsistent with randomness – supplying strong evidence that basketball shooting is not random for all shooters all of the time. However, we find that the evidence against randomness in this experiment 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, and thereby constitute a direct test of the hot hand fallacy.

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 and the National Science Foundation under the Graduate Research Fellowship Program. JR acknowledges funding from the National Science Foundation (MMS-1949845). 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 we observe s Bernoulli sequences of length n. We are interested in testing the null hypothesis that these sequences are independent and identically distributed (i.i.d.) against alternatives in which the probability of success following a streak of consecutive successes is greater than it is either unconditionally or following a streak of consecutive failures. The interpretation of results of tests of this form have been pivotal in the development of behavioral economics.

In an influential paper, Tversky and Kahneman (1971) hypothesize that people tend to believe that small samples are overly representative of the "essential characteristics" of the population from which they are drawn. They describe this phenomenon as "belief in the law of small numbers," and present several evocative examples in support of this claim. For instance, they show that academic researchers tend to substantially underestimate sample sizes necessary to achieve adequate statistical power against reasonable alternatives in the design of experiments – evidently finding samples overly representative of the populations from which they are drawn. Similarly, in what has been termed the "gambler's fallacy," when asked to successively predict outcomes of an i.i.d. Bernoulli sequence, experimental subjects tend to underestimate the probability of streaks of consecutive successes or failures – apparently perceiving streaks to be overly representative of non-randomness. Subsequently, this insight into misperception of randomness has been formalized (Rabin, 2002) and integrated into standard models in behavioral economics and finance (Barberis and Thaler, 2003; Barberis, 2018).[1]

The "hot hand fallacy," proposed and studied in Gilovich et al. (1985), henceforth GVT, is a behavioral bias attributable to belief in the law of small numbers. As subsequently formalized in Rabin and Vayanos (2010), the hot hand fallacy refers to a positive bias in beliefs about the dependence in a Bernoulli process following the observation of a streak of consecutive successes. In particular, when faced with a streak of consecutive successes, believers in the law of small numbers overestimate the positive dependence in the sequence – perceiving streaks to be overly representative of dependence.[2] GVT aim to document the hot fallacy using data from a controlled basketball shooting experiment and results from a survey on beliefs in the serial dependence of

[1] Bar-Hillel and Wagenaar (1991) review the psychological literature on models of misperception of randomness, highlighting their implications for judgment of dependence in random Bernoulli sequences. Benjamin (2019) reviews the psychological and behavioral economics literatures on errors in probabilistic reasoning, surveying the available empirical support for proposed biases and highlighting areas of economics where these biases are relevant.

[2] We refer the reader to Rabin and Vayanos (2010) for a precise analysis of conditions under which belief in the law of small numbers implies the hot hand and gambler's fallacies.

basketball shooting. They fail to reject the hypothesis that the sequences of shots they observe are i.i.d., but document a widespread 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. Thus, they conclude that the belief in the hot hand is a pervasive cognitive illusion or fallacy, giving provocative evidence in favor of models that incorporate belief in the law of small numbers. This result became the academic consensus for the following three decades (Kahneman, 2011; Thaler and Sunstein, 2009), and provided a central empirical support for economic models in which agents are overconfident in conclusions drawn from small samples (Barberis and Thaler, 2003).

The GVT results were challenged by Miller and Sanjurjo (2018d), henceforth MS, who discovered a significant small-sample bias in plug-in estimates of the probability of success following streaks of successes or failures. They argue that when they correct the GVT analysis for this small-sample bias, they are able to reject the null hypothesis that shots are i.i.d., in favor of positive dependence that is consistent with expectations of streakiness in basketball shooting.[3] 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 suggests that their work creates persisting uncertainty about the empirical support for textbook theories of misperception of randomness. Benjamin (2019) writes 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 the uncertainty in the evidence that controlled basketball shooting experiments have contributed to our understanding of the hot hand fallacy and, by implication, economic models incorporating belief in the law of small numbers. Towards this goal, we develop a formal statistical framework for testing the randomness of a set of Bernoulli sequences and measure the finite-sample power of these tests with local asymptotic approximations. Equipped with these theoretical results, we then provide a comprehensive analysis of the design and interpretation of the outcomes of four controlled basketball shooting experiments and give recommendations for the design and methodology of future empirical work.

Section 2 develops a formal statistical framework for assessing the positive serial dependence in basketball shooting using data from controlled shooting experiments. We emphasize the dis-

[3]The MS results received extensive coverage in the popular press, including expository articles in the New York Times (Johnson 2015 and Appelbaum 2015), the New Yorker (Remnick, 2017), the Wall Street Journal (Cohen, 2015), and on ESPN (Haberstroh, 2017), among 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.

tinctions between individual, simultaneous, and joint hypothesis testing. We specify our null hypothesis – that observed shooting outcomes are i.i.d. – and a set of alternative hypotheses in which the probability of a make following a streak of consecutive makes is greater than it is either unconditionally or following a streak of consecutive misses. We denote this class of alternatives as "streaky", as the probability of a streak of makes is larger than it would be under randomness. These alternatives motivate a set of natural plug-in test statistics, studied previously in GVT and MS: the observed differences between the proportions of makes following a streak of consecutive makes and either the overall proportion of makes or the proportion of makes following a streak of consecutive misses.

Section 3 develops methods for testing the randomness of a collection of Bernoulli sequences against streaky alternatives using these test statistics. We derive the asymptotic distributions of the test statistics specified in Section 2 under the null hypothesis of randomness and under general stationary alternatives. We highlight the substantial small-sample biases of these approximations under the null hypothesis, which were discovered and studied in MS. This bias motivates the application of permutation tests, which we show are the only tests with exact type 1 error control. We conclude the Section by characterizing the asymptotics of the test statistics' permutation distributions under the null hypothesis and under general stationary alternatives.

In Section 4, these results allow us to derive asymptotic approximations to the power of the permutation tests developed in Section 3 against a specific class of Markov chain streaky alternatives with a local asymptotic approximation. These results significantly reduce the computational expense of power analyses in the design of future experiments. Simulation evidence indicates that our asymptotic power approximations perform remarkably well in the sample sizes considered in available controlled basketball shooting experiments.

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. The standard approach is to verify Hoeffding's condition; see Theorem 3.4. 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 K.1 in the Online Appendix). This result (Lemma K.2 in the Online Appendix) 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 3.4).

Having developed a formal statistical framework for testing the randomness of a collection of Bernoulli sequences, in Section 5 we evaluate the implications of the outcomes of four controlled basketball shooting experiments for the question posed in Benjamin (2019): "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 actual deviation from randomness and expectations of the deviation from randomness in basketball shooting.

First, we analyze the design and results of four controlled basketball shooting experiments. We find that there is strong evidence that basketball shooting is not perfectly random for all basketball players all of the time. In data from the GVT experiment, we find that we are able to reject i.i.d. shooting consistently after accounting for multiplicity for only one shooter out of twenty-six, identified in the dataset as "Shooter 109". 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.[4] However, we argue that the four controlled shooting experiments do not have adequate power to detect parameterizations of the Markov chain streaky alternative, studied in Section 4, consistent with the variation in NBA field goal shooting percentages.[5] Moreover, evidence against randomness in the GVT experiment appears to be confined to Shooter 109.[6]

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

[4]GVT observe the rejection of the null hypothesis for Shooter 109, but concede "we might expect one significant result out of 26 by chance." We show that the rejection of the null hypothesis for Shooter 109 is robust to standard multiple testing corrections. Wardrop (1999) notes that the p-value for standard tests of the randomness of the shooting sequence for Shooter 109 is extremely small.

[5]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.

[6]We are not the first to observe that the GVT data are underpowered for the Markov chain alternatives. Miller and Sanjurjo (2019), Miyoshi (2000), and Wardrop (1999) measure the power of individual tests against specific parameterizations of similar models with simulation. Korb and Stillwell (2003) and Stone (2012) measure power against particular non-stationary alternatives. 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 and joint null hypothesis tests.

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

Tests of the randomness of stochastic processes against nonrandom, persistent, or serially dependent alternatives have been studied extensively within finance and economics; the framework and methods that we develop are applicable to these settings. In Online Appendix A, we outline the application of our methods to two problems in empirical finance: tests of the weak form efficient market hypothesis (Fama, 1970; Malkiel, 2003) and tests of persistence in the performance of mutual funds relative to benchmarks (Jensen, 1968; Hendricks et al., 1993; Carhart, 1997). In particular, in the spirit of Fama (1965), in Online Appendix A.1 we implement the individual permutation tests that we develop in Section 3.2 on two datasets of stock price sequences. Moreover, the problem of testing for and estimating state dependence – the causal effect of an outcome in the previous period on the current period's outcome – is widely studied in microeconomics (see e.g., Heckman 1981; Chay et al. 1999; Keane 1997). In Online Appendix B, we show that our methods provide a test for state dependence under appropriate unconfoundedness type assumptions.[7]

Section 6 concludes. Online Appendices A-J give supplementary results and discussion that will be introduced at appropriate points throughout the paper. Proofs of all Theorems presented in the main body of this paper are given in Online Appendix K.

2 Posing the Problem

Do people overestimate positive serial dependence in basketball shooting? Three components of this question are often conflated:

- Is there any positive serial dependence in basketball shooting?
- If so, how widespread and substantial is it?
- And finally, do people systematically overestimate this dependence?

In this section, we provide a formal framework that will enable us to develop inferential methods for answering the first two questions. In Section 5, we provide a discussion of methods for

[7] As they do not play a role in our empirical application, we do not consider covariates or instruments. The extension of our methods to account for observed and unobserved heterogeneity across time and individuals is important for their application to these contexts and may be fruitful. Torgovitsky (2019) develops a partial identification approach to bounding state dependence in these settings.

addressing the third question and a review of the evidence on beliefs.

The Null Hypothesis: Suppose that we observe s shooters; each shoots n consecutive shots under identical conditions. Let $\mathbf{X}_i = \{X_{ij}\}_{j=1}^{n}$ denote the vector of shot outcomes for shooter i and $\mathbf{X} = \{\mathbf{X}_i\}_{i=1}^{s}$ the matrix of these outcomes for all shooters, with $X_{ij} = 1$ denoting a made shot and $X_{ij} = 0$ denoting a missed shot.

We would like to test the hypothesis that there is no positive serial dependence in the outcomes of the observed shots. A test of the joint null hypothesis

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

assesses whether basketball shooting is a random process for all shooters in the sample. In contrast, tests of the individual hypotheses

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

or the multiple hypothesis problem that tests the hypotheses H_0^i simultaneously assess whether basketball shooting is a random process for shooter i or for each of the shooters in the sample simultaneously, respectively.[8]

Rejection of the joint null hypothesis H_0 indicates that there is non-zero serial dependence for at least one shooter in the sample, but does not indicate *which* shooters deviate from randomness. In order to identify any such shooters, we apply multiple testing methods that control the familywise error rate (FWER), i.e., the probability of at least one false rejection of an individual hypothesis H_0^i. Note that a test of the joint null hypothesis H_0 is more liberal than simultaneous tests of H_0^i, in the sense that tests of H_0 can be rejected even if there is insufficient evidence to reject any of the individual hypotheses H_0^i at the same level.[9]

[8]There is an ambiguity in the literature on the hot hand fallacy whether belief in the hot hand refers to a biased belief in a serial correlation or a causal effect of an outcome of a shot on subsequent shots. In Online Appendix B.1, we characterize the relationship between the null hypothesis that $\mathbf{X}_i$ is i.i.d. and the null hypothesis that, for each shot, the outcomes of the preceding m shots have no causal effect on the probability of a make. We provide an unconfoundedness type assumption under which these conditions are equivalent.

[9]Indeed, the closure method for constructing multiple tests that control the FWER is based on tests of joint hypotheses; in order for H_0^i to be rejected, tests of all joint null hypotheses for any subset of shooters containing shooter i must be rejected, not just the subset consisting of all shooters. In fact, any multiple hypothesis testing method that controls the FWER must be constructed with the closure method (Romano et al., 2011).

Streaky Alternatives: In general, stationary processes are a broad class of alternatives to i.i.d. processes, allowing for quite arbitrary dependence. We maintain the assumption that the shot outcomes $\mathbf{X}_i$ follow stationary Bernoulli (p_i) processes $\mathbb{P}_i$ and are independent across shooters, with $\mathbb{P} = \{\mathbb{P}_i\}_{i=1}^{s}$ collecting these processes in a 2^s-vector valued Bernoulli process.[10]

However, some stationary alternatives to H_0 are inconsistent with notions of "the hot hand" or "streak shooting". GVT argue that in most conceptions of the hot hand, the probability of making a shot following a series of made shots is higher than both the marginal probability of making a shot and the probability of making a shot following a series of missed shots. Thus, following GVT and MS, in order to assess whether there is positive serial dependence in basketball shooting, we study tests of H_0 against alternatives in which the parameters

$$\bar{\theta}_P^k(\mathbb{P}) = \frac{1}{s}\sum_{i=1}^{s}\theta_P^k(\mathbb{P}_i) \quad \text{and} \quad \bar{\theta}_D^k(\mathbb{P}) = \frac{1}{s}\sum_{i=1}^{s}\theta_D^k(\mathbb{P}), \tag{2.1}$$

where

$$\theta_P^k(\mathbb{P}_i) = \mathbb{P}_i\left\{X_{i,j+k} = 1 | \prod_{l=0}^{k-1} X_{i,j+l} = 1\right\} - \mathbb{P}_i\{X_{ij} = 1\} \tag{2.2}$$

$$\theta_D^k(\mathbb{P}_i) = \mathbb{P}_i\left\{X_{i,j+k} = 1 | \prod_{l=0}^{k-1} X_{i,j+l} = 1\right\} - \mathbb{P}_i\left\{X_{i,j+k} = 1 | \prod_{l=0}^{k-1}(1 - X_{i,j+l}) = 1\right\}, \tag{2.3}$$

are greater than zero for some integer k. Throughout, we refer to alternatives of this form as "streaky", as the probability of a streak of made shots of length $k+1$ is higher than it would be under an i.i.d. process. A rejection of the null hypothesis H_0 against streaky alternatives provides an affirmative answer to the first question posed at the beginning of this section – that there is non-zero and positive serial dependence in basketball shooting.

Test Statistics: Following GVT and MS, we study tests of the individual and joint null hypotheses of randomness against streaky alternatives that use natural plug-in estimators for the parameters $\theta_P^k(\mathbb{P}_i)$ and $\theta_D^k(\mathbb{P}_i)$, as well as $\bar{\theta}_P^k(\mathbb{P})$ and $\bar{\theta}_D^k(\mathbb{P}_i)$, as test statistics, respectively. These statistics are defined as follows. Let each individual's observed shooting percentage 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 made shots following k consecutive made shots. That is, letting $Y_{ijk} = \prod_{l=j}^{j+k} X_{il}$ and $V_{ik} = \sum_{j=1}^{n-k} Y_{ijk}$, then $\hat{P}_{n,k}(\mathbf{X}_i)$ is given by

[10] Under H_0, $p_i = \mathbb{P}_i\{X_{ij} = 1\}$ may vary across shooters.

$$\hat{P}_{n,k}(\mathbf{X}_i) = \frac{V_{ik}}{V_{i(k-1)}}. \tag{2.4}$$

Likewise, let $\hat{D}_{n,k}(\mathbf{X}_i)$ denote the difference between the proportion of made shots following k consecutive made shots and k consecutive missed shots. That is, letting $Z_{ijk} = \prod_{l=j}^{j+k}(1 - X_{il})$ and $W_{ik} = \sum_{j=1}^{n-k} Z_{ijk}$, then $\hat{D}_{n,k}(\mathbf{X}_i)$ is given by

$$\hat{D}_{n,k}(\mathbf{X}_i) = \frac{V_{ik}}{V_{i(k-1)}} - \frac{W_{ik}}{W_{i(k-1)}} \tag{2.5}$$

Intuitively, $\hat{P}_{n,k}(\mathbf{X}_i) - \hat{p}_{n,i}$ and $\hat{D}_{n,k}(\mathbf{X}_i)$ are natural plug-in estimators for $\theta_P^k(\mathbb{P}_i)$ and $\theta_D^k(\mathbb{P}_i)$, respectively. The averages of these estimators over the shooters in the sample are denoted by

$$\bar{P}_k(\mathbf{X}) = \frac{1}{s}\sum_{i=1}^{s} \hat{P}_{n,k}(\mathbf{X}_i) - \hat{p}_{n,i} \quad \text{and} \quad \bar{D}_k(\mathbf{X}) = \frac{1}{s}\sum_{i=1}^{s} \hat{D}_{n,k}(\mathbf{X}_i) \tag{2.6}$$

and are natural plug-in estimators for $\bar{\theta}_P^k(\mathbb{P})$ and $\bar{\theta}_D^k(\mathbb{P})$, respectively. Note that $\hat{P}_{n,k}(\mathbf{X}_i)$ and $\hat{D}_{n,k}(\mathbf{X}_i)$ are not defined for every sequence $\mathbf{X}_i$. Specifically, they are not defined for sequences without instances of k consecutive ones or zeros. However, under our null hypothesis of randomness and the alternatives that we consider, the statistics are defined with probability approaching one exponentially quickly as n grows to infinity.

Estimation: Rejection of either the joint null hypothesis H_0 or an individual hypothesis H_0^i, after accounting for simultaneity, indicates that there is non-zero serial dependence for at least one shooter in the sample. It does not, however, provide a quantification of this dependence. Estimates and confidence intervals for $\bar{\theta}_P^k(\mathbb{P})$ and $\bar{\theta}_D^k(\mathbb{P})$ provide metrics for quantifying the observed serial dependence. These metrics are adopted by GVT and MS. As suggested above, $\bar{P}_k(\mathbf{X})$ and $\bar{D}_k(\mathbf{X})$ are natural estimators for $\bar{\theta}_P^k(\mathbb{P})$ and $\bar{\theta}_D^k(\mathbb{P})$, respectively. While not the emphasis of our analysis, we discuss methods for constructing confidence intervals for these parameters in Section 3.1.

3 Testing Randomness Against Streaky Alternatives

In this Section, we develop methods for testing the randomness of a collection of Bernoulli sequences against streaky alternatives using the plug-in statistics presented in Section 2.

3.1 Asymptotic Behavior of the Test Statistics

We begin by characterizing the asymptotic distributions of the plug-in statistics $\hat{P}_{n,k}(\mathbf{X}_i) - \hat{p}_n$ and $\hat{D}_{n,k}(\mathbf{X}_i)$ under the null hypothesis of randomness H_0^i. To date, such distributions have not been derived. Miller and Sanjurjo (2018a) claim that $\hat{P}_{n,k}(\mathbf{X}_i)$ is asymptotically normal under the null hypothesis, referencing Mood (1940), but do not provide explicit formulae for its asymptotic variance. Even in the i.i.d. case, the test statistics are functions of overlapping subsequences of observations, so central limit theorems for dependent data are required.

Theorem 3.1. *Under the assumption that $\mathbf{X}_i = \{X_{ij}\}_{j=1}^n$ is a sequence of i.i.d. Bernoulli(p_i) random variables,*
(i) *$\hat{P}_{n,k}(\mathbf{X}_i) - \hat{p}_{n,i}$, with $\hat{P}_{n,k}(\mathbf{X}_i)$ given by (2.4) and $\hat{p}_{n,i} = n^{-1}\sum_{j=1}^n X_{ij}$, is asymptotically normal with limiting distribution given by*

$$\sqrt{n}\left(\hat{P}_{n,k}(\mathbf{X}_i) - \hat{p}_{n,i}\right) \xrightarrow{d} N\left(0, \sigma_P^2(p_i,k)\right), \tag{3.1}$$

as $n \to \infty$, where $\sigma_P^2(p_i,k) = p_i^{1-k}(1-p_i)\left(1-p_i^k\right)$, and
(ii) *$\hat{D}_{n,k}(\mathbf{X}_i)$, given by (2.5), is asymptotically normal with limiting distribution given by*

$$\sqrt{n}\hat{D}_{n,k}(\mathbf{X}_i) \xrightarrow{d} N\left(0, \sigma_D^2(p_i,k)\right), \tag{3.2}$$

as $n \to \infty$, where $\sigma_D^2(p_i,k) = (p_i(1-p_i))^{1-k}\left((1-p_i)^k + p_i^k\right)$.

Remark 3.1. Note that $\sigma_D^2\left(\frac{1}{2},k\right) = 2^{k-1}$ increases exponentially with k, stemming from an effectively reduced sample size – limiting to those outcomes that follow sequences of ones or zeros of length k. ■

Remark 3.2. Theorem 3.1 can be generalized to a triangular array $\mathbf{X}_{n,i} = \{X_{n,i,j}\}_{j=1}^n$ of i.i.d. Bernoulli trials with probability of success $p_{n,i}$ converging to p_i. Specifically, we have that, under $p_{n,i}$ (3.1) and (3.2) continue to hold. This result implies that we can consistently approximate the quantiles of the distributions of $\hat{P}_{n,k}(\mathbf{X}_{n,i}) - \hat{p}_{n,i}$ and $\hat{D}_{n,k}(\mathbf{X}_{n,i})$ under the null hypothesis with the parametric bootstrap, which approximates the distribution of $\sqrt{n}\hat{D}_{n,k}(\mathbf{X}_i)$ under p_i using that of $\sqrt{n}\hat{D}_{n,k}(\mathbf{X}_i)$ under $\hat{p}_{n,i}$. ■

Remark 3.3. For a set $\mathbf{X} = \{\mathbf{X}_i\}_{i=1}^s$ of Bernoulli sequences of length n each having probability of success p_i, Theorem 3.1 implies that the statistics $\sqrt{ns}\bar{P}_k(\mathbf{X})$ and $\sqrt{ns}\bar{D}_k(\mathbf{X})$ have normal

limiting distributions with means equal to zero and variances equal to the averages of $\sigma_P^2(p_i,k)$ and $\sigma_D^2(p_i,k)$ over the individuals $1,\ldots,s$, respectively.[11] ■

Remark 3.4. In Online Appendix F, we show that under a stationary α-mixing process $\mathbb{P}_i$, $\hat{P}_{n,k}(\mathbf{X}_i)-\hat{p}_{n,i}$ and $\hat{D}_{n,k}(\mathbf{X}_i)$ are asymptotically normal with means $\theta_P^k(\mathbb{P}_i)$ and $\theta_D^k(\mathbb{P}_i)$, respectively. Stationary α-mixing processes provide a general class of alternatives to the null hypothesis of randomness, and allow for quite general forms of dependence between shots that are close to each other.[12] ■

Remark 3.5. For a stationary process $\mathbb{P}_i$, the limiting variances of $\hat{P}_{n,k}(\mathbf{X}_i)-\hat{p}_{n,i}$ and $\hat{D}_{n,k}(\mathbf{X}_i)$ 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 (2013)), such as the moving blocks bootstrap (Liu and Singh, 1992; 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 $\theta_P^k(\mathbb{P}_i)$ or $\theta_D^k(\mathbb{P}_i)$. ■

There is a severe second-order bias in the finite sample performance of these asymptotic approximations. Specifically, let $\beta_P^{n,k}(\mathbb{P}_i)$ and $\beta_D^{n,k}(\mathbb{P}_i)$ denote the expectations of $\hat{P}_{n,k}(\mathbf{X}_i)-\hat{p}_{n,i}$ and $\hat{D}_{n,k}(\mathbf{X}_i)$ under the stationary process $\mathbb{P}_i$. With a minor abuse of notation, we let $\beta_P^{n,k}(p_i)$ and $\beta_D^{n,k}(p_i)$ denote these parameters when $\mathbb{P}_i$ is an i.i.d. Bernoulli process with marginal success rate p_i. MS show that $\beta_P^{n,k}(p_i)$ and $\beta_D^{n,k}(p_i)$ are less than $\theta_P^k(\mathbb{P}_i)$ and $\theta_D^k(\mathbb{P}_i)$ under H_0^i with marginal success rate p_i. These differences converge to zero as n increases.[13]

Thus, the statistics $\hat{P}_{n,k}(\mathbf{X}_i)-\hat{p}_{n,i}$ or $\hat{D}_{n,k}(\mathbf{X}_i)$ have a negative bias when considered as estimators for $\theta_P^k(\mathbb{P}_i)$ and $\theta_D^k(\mathbb{P}_i)$ under H_0^i. Equivalently, procedures that test H_0^i against streaky alternatives by comparing $\hat{P}_{n,k}(\mathbf{X}_i)-\hat{p}_{n,i}$ or $\hat{D}_{n,k}(\mathbf{X}_i)$ to quantiles of their limiting distributions – without correcting for these finite-sample biases – have a type 1 error rate below the desired level in finite-samples. To illustrate, suppose that $n=100$ and that $\mathbf{X}_i=\{X_{ij}\}_{j=1}^n$ is an i.i.d. Bernoulli(p_i) sequence with $p_i=1/2$. Columns (1) and (3) in Table 1 give the expectations of $\hat{P}_{n,k}(\mathbf{X}_i)-\hat{p}_{n,i}$ and

[11] In Online Appendix C, we discuss the asymptotic distributions of these statistics when the parameters p_i are realizations of a sequence of i.i.d. random variables. We use this result in Section 4 to approximate the power of the tests that we develop against a set of alternatives in which individuals independently deviate from the null hypothesis with a pre-specified probability.

[12] In an α-mixing process, the dependence between two shots X_{ij} and $X_{i(j+t)}$ approaches zero as t grows to infinity. For example, any Markov Chain with finite state space that is irreducible and aperiodic is α-mixing (Bradley, 1986). In Online Appendix B.2, we show that stationaryα-mixing alternatives are a natural class of alternatives to consider in a dynamic potential outcomes framework.

[13] Exact expressions for the expectations of these statistics in finite-samples appear to be unknown for $k>1$. In Online Appendix D, we obtain the second order approximations $\beta_P^{n,k}(p_i)=n^{-1}p_i\left(1-p_i^{-k}\right)+O\left(n^{-2}\right)$ and $\beta_D^{n,k}(p_i)=n^{-1}\left(1-(1-p_i)^{1-k}-p_i^{1-k}\right)+O\left(n^{-2}\right)$.

k	$\hat{P}_{n,k}(\mathbf{X}_i) - \hat{p}_{n,i}$		$\hat{D}_{n,k}(\mathbf{X}_i)$	
	Expectation	Type 1 Error Rate	Expectation	Type 1 Error Rate
	(1)	(2)	(3)	(4)
1	-0.005	0.044	-0.010	0.039
2	-0.016	0.032	-0.032	0.029
3	-0.041	0.023	-0.080	0.020
4	-0.090	0.013	-0.177	0.010

Table 1: Finite-Sample Behavior of Plug-in Statistics

Notes: Table displays simulated estimates of the finite sample expectations of $\hat{P}_{n,k}(\mathbf{X}_i) - \hat{p}_{n,i}$ and $\hat{D}_{n,k}(\mathbf{X}_i)$ as well as the type 1 error rates of the hypothesis tests that reject H_0^i if $\hat{P}_{n,k}(\mathbf{X}_i) - \hat{p}_{n,i}$ and $\hat{D}_{n,k}(\mathbf{X}_i)$ exceed the 0.95 quantile of their asymptotic distributions. We take 100,000 draws of Bernoulli$(1/2)$ random variables of length 100. We compute expectations by taking the mean of $\hat{P}_{n,k}(\mathbf{X}_i) - \hat{p}_{n,i}$ and $\hat{D}_{n,k}(\mathbf{X}_i)$ computed on each draw. We compute type 1 error rates by taking the proportion of draws in which $\hat{P}_{n,k}(\mathbf{X}_i) - \hat{p}_{n,i}$ and $\hat{D}_{n,k}(\mathbf{X}_i)$ exceed the 0.95 quantiles of their asymptotic distributions.

$\hat{D}_{n,k}(\mathbf{X}_i)$ for k in $1, \ldots, 4$, respectively. In contrast, $\theta_P^k(\mathbb{P}_i)$ and $\theta_D^k(\mathbb{P}_i)$, defined in (2.2) and (2.3), are both equal to zero. Columns (2) and (4) of Table 1 give the probabilities that $\hat{P}_{n,k}(\mathbf{X}_i) - \hat{p}_{n,i}$ and $\hat{D}_{n,k}(\mathbf{X}_i)$ are greater than the 0.95 quantiles of the normal distributions with means zero and variances $\sigma_D^2(p_i, k)$ and $\sigma_P^2(p_i, k)$ for k in $1, \ldots, 4$, respectively. The probabilities are significantly below 0.05 and decrease with k. Hence, to conduct more powerful tests of randomness, it is necessary to account for this bias – at least implicitly.

In their analysis of controlled basketball shooting experiments, GVT test the individual hypotheses H_0^i by comparing $\hat{P}_{n,k}(\mathbf{X}_i) - \hat{p}_{n,i}$ and $\hat{D}_{n,k}(\mathbf{X}_i)$ to quantiles of approximations to their limiting distributions without correcting for finite-sample bias. MS argue that GVT's conclusion that the null hypotheses H_0^i cannot be rejected is sensitive to correction for this bias, i.e., the implementation of tests with more accurate control of the type 1 error rate. In the subsequent subsection, we discuss permutation tests, show that they automatically account for finite-sample biases, and prove that they are in fact the only tests that control the type 1 error rate exactly in finite samples. We advocate for their choice as the default test in our setting.

3.2 Permutation Tests, Bias-Corrected Estimation, and Simultaneous Inference

In this subsection, we outline permutation tests of the individual hypothesis H_0^i and the joint null hypothesis H_0 that control the type 1 error rate exactly in finite-samples. We then propose a set

of estimators of the individual parameters $\theta_P^k(\mathbb{P}_i)$ and $\theta_D^k(\mathbb{P}_i)$ and the joint parameters $\bar{\theta}_P^k(\mathbb{P})$ and $\theta_D^k(\mathbb{P}_i)$ that are exactly unbiased under the null hypothesis. Finally, we lay out a standard multiple hypothesis testing procedure that can be applied to test the individual hypotheses H_0^i simultaneously.

Individual and Joint Tests: Based on the data $\mathbf{X}_i = \{X_{ij}\}_{j=1}^n$, it is desired to test the null hypothesis H_0^i that the underlying observations are i.i.d. Bernoulli with some unknown success probability p_i. Under H_0^i, the distribution of $\mathbf{X}_i$ is invariant under permutations; that is $(X_{i,1}, \ldots, X_{i,n})$ and $\left(X_{i,\pi(1)}, \ldots, X_{i,\pi(n)}\right)$, where π is a permutation of $(1, \ldots, n)$, have the same joint distribution. This property is a special case of the *randomization hypothesis* specified in Section 15.2 of Lehmann and Romano (2005) and allows for the construction of permutation tests.

In a permutation test, a test statistic is recomputed on every permutation of a data set. The distribution of these recomputed statistics is used as a null or reference distribution for comparison with the observed value of the test statistic. The proportion of recomputed statistics exceeding the observed test statistic is the p-value of the permutation test. Permutation tests are exact level α for any choice of test statistic. In particular, let $T_n(\mathbf{X}_i)$ be any real-valued test statistic for testing H_0^i. Let $\mathbf{X}_{i,\pi} = \left(X_{i,\pi(1),\ldots,}X_{i,\pi(n)}\right)$, where π is an element of $\Pi(n)$, be the set of permutations of $\{1, \ldots, n\}$. The permutation, or randomization, distribution for $\sqrt{n}T_n(\mathbf{X}_i)$ is given by

$$\hat{R}_n^T(t) = \frac{1}{n!} \sum_{\pi \in \Pi(n)} I\left\{\sqrt{n}T_n(\mathbf{X}_{i,\pi}) \leq t\right\}.$$

For a nominal level α, $0 < \alpha < 1$, the permutation test rejects at level α if $\sqrt{n}T_n(\mathbf{X}_i)$ is greater than the $1-\alpha$ quantile of $\hat{R}_n^T$.[14] Define the permutation test function $\varphi(\mathbf{X}_i)$ to be equal to one if the permutation test rejects and zero otherwise. By Theorem 15.2.1 in Lehmann and Romano (2005), $\mathbb{E}[\varphi(\mathbf{X}_i)] = \alpha$ if H_0^i is true. What may be less obvious is that any test φ that is exactly level α for testing H_0^i *must* be a permutation test.

Theorem 3.2. *Suppose $\varphi = \varphi(\mathbf{X}_i)$ is any test function such that $\mathbb{E}[\varphi(\mathbf{X}_i)] = \alpha$ whenever $\mathbf{X}_i$ is*

[14] As the permutation distribution is discrete, the exact permutation test may require randomization when $\sqrt{n}T_n(\mathbf{X}_i)$ is equal to the $1-\alpha$ quantile of $\hat{R}_n^T$. In practice, we use a slightly conservative approach by not randomizing; that is, we reject H_0^i only if $\sqrt{n}T_n(\mathbf{X}_i)$ exceeds the $1-\alpha$ quantile of $\hat{R}_n^T$.

i.i.d. Bernoulli with some unknown success rate p_i. Then, φ must be a permutation test; that is

$$\frac{1}{n!}\sum_{\pi\in\Pi(n)}\varphi\left(\mathbf{X}_{i,\pi}\right)=\alpha.$$

In practice, one does not need to compute all $n!$ permutations. Instead, if permutations are sampled at random, then one can still attain valid finite-sample p-values. Both $\hat{P}_{n,k}\left(\mathbf{X}_i\right)-\hat{p}_{n,i}$ and $\hat{D}_{n,k}\left(\mathbf{X}_i\right)$, given in (2.4) and (2.5), are appropriate choices for $T_n\left(\mathbf{X}_i\right)$. In the following subsection, we characterize the asymptotic distribution of $\hat{R}_n^T$ for these choices.

Similarly, the joint null hypothesis H_0 can be tested with a stratified permutation test wherein each Bernoulli sequence $\mathbf{X}_i$ is permuted separately. Specifically, let $K_{n,s}\left(\mathbf{T}\right)$ be a general function of the individual test statistics $\mathbf{T}=\left\{T_n\left(\mathbf{X}_i\right)\right\}_{i=1}^s$. The stratified permutation distribution for $\sqrt{ns}K_{n,s}$ is given by

$$\hat{R}_{n,s}^{K,T}\left(t\right)=\frac{1}{\left(n!\right)^s}\sum_{\left(\pi_1,\ldots,\pi_s\right)\in\Pi(n)^s}I\left\{\sqrt{ns}K_{n,s}\left(T_n\left(\mathbf{X}_{i,\pi_i}\right),\ldots,T_n\left(\mathbf{X}_{i,\pi_s}\right)\right)\leq t\right\},$$

where $\Pi\left(n\right)^s$ is the set of all s-vectors of permutations of $\left(1,\ldots,n\right)$.[15] A stratified permutation test rejects H_0 at level α if $\sqrt{ns}K_{n,s}$ exceeds the $1-\alpha$ quantile of $\hat{R}_{n,s}^{K,T}$. Both $\bar{P}_k\left(\mathbf{X}\right)$ and $\bar{D}_k\left(\mathbf{X}\right)$, given in (2.6), are appropriate choices for the joint test statistic $K_{n,s}$.[16]

The use of permutation tests bypasses the need for explicit bias-correction. Specifically, the expected value of the mean of the permutation distribution $\hat{R}_n^T\left(t\right)$ is exactly that of $\sqrt{n}T_n\left(\mathbf{X}_i\right)$ under the null hypothesis. Thus, one can avoid approximating finite sample biases explicitly, because the permutation distributions account for these biases automatically.

Bias-Corrected Estimation: The equality between expectations of the means of permutation distributions and expectations of their associated statistics under the null hypothesis can be leveraged to construct bias-corrected estimators. In particular, let

$$\hat{\eta}\left(\mathbf{X}_i,T_n\right)=\frac{1}{n!}\sum_{\pi\in\Pi(n)}T_n\left(\mathbf{X}_{i,\pi}\right)$$

[15] We note that $K_{n,s}$ must be computed over all individuals i where the statistic $T_n\left(\mathbf{X}_i\right)$ is defined.

[16] In Online Appendix G, we outline three additional choices for joint test statistics that combine p-values of individual permutation tests across individuals. Each of these choices of statistics will have power against different alternatives. Additionally, we outline two methods for combining p-values of different joint tests to compute a singular composite p-value.

denote the mean of the permutation distribution of the statistic $T_n(\mathbf{X}_i)$. Under the null hypothesis, the expectation of $\hat{\eta}(\mathbf{X}_i, T_n)$ is exactly equal to the expectation of $T_n(\mathbf{X}_i)$. This observation suggests the bias-corrected estimators

$$\tilde{P}_{n,k}(\mathbf{X}_i) = \hat{P}_{n,k}(\mathbf{X}_i) - \hat{p}_{n,i} - \hat{\eta}\left(\mathbf{X}_i, \hat{P}_k - \hat{p}_i\right) \text{ and } \tilde{D}_{n,k}(\mathbf{X}_i) = \hat{D}_{n,k}(\mathbf{X}_i) - \hat{\eta}\left(\mathbf{X}_i, \hat{D}_{n,k}\right) \quad (3.3)$$

and their averages

$$\bar{\tilde{P}}_k(\mathbf{X}) = \frac{1}{s}\sum_{i=1}^{s} \tilde{P}_{n,k}(\mathbf{X}_i) \text{ and } \bar{\tilde{D}}_k(\mathbf{X}) = \frac{1}{s}\sum_{i=1}^{s} \tilde{D}_{n,k}(\mathbf{X}_i). \quad (3.4)$$

These estimators are exactly unbiased under the null hypothesis, and are consistent under streaky alternatives.[17]

Simultaneous Tests: Suppose that the joint null hypothesis H_0 is rejected. In this case, in order to characterize which of the Bernoulli sequences $\mathbf{X}_i$ are non-random, we would like to know which of the individual hypothesis H_0^i can be rejected. A problem of this form – testing a finite number of individual hypotheses simultaneously – is a "multiple testing" or "simultaneous inference" problem; see Chapter 9 of Lehmann and Romano (2005) for a textbook treatment.

If the hypotheses H_0^i are each tested at level α, then the probability of a false rejection of at least one individual hypothesis H_0^i increases rapidly with s. In fact, when s is equal to 10, then the probability of at least one false rejection when all individual hypotheses are true is equal to approximately 0.4. Thus, we apply methods that control the familywise error rate (FWER), i.e., the probability of at least one false rejection of an individual null hypothesis H_0^i.

In particular, we apply a stepdown procedure with Šidák critical values. Let ρ_i denote the p-value for a permutation test of H_0^i and let the p-values ordered from lowest to highest be $\rho_{(1)}, \ldots, \rho_{(s)}$, with associated hypotheses $H_0^{(1)}, \ldots, H_0^{(s)}$. Fix a nominal level α, $0 < \alpha < 1$, and let r be the maximal index such that $\rho_{(1)} < \alpha_1, \cdots, \rho_{(r)} < \alpha_r$ and $\rho_{(r+1)} > \alpha_{r+1}$, where

$$\alpha_i = 1 - (1-\alpha)^{(1/(s-i+1))},$$

[17] Alternatively, the bias can be approximated by the parametric bootstrap, i.e., $\hat{P}_{n,k}(\mathbf{X}_i) - \hat{p}_{n,i} - \beta_P^{n,k}(\hat{p}_{n,i})$ and $\hat{D}_{n,k}(\mathbf{X}_i) - \beta_D^{n,k}(\hat{p}_{n,i})$ where $\beta_P^{n,k}(\hat{p}_{n,i})$ and $\beta_D^{n,k}(\hat{p}_{n,i})$ are computed with simulation. MS take this approach. These estimators are only approximately unbiased under the null hypothesis. It is straightforward to show that the expectations of these statistics are $O\left(n^{-2}\right)$ by replacing p_i with $\hat{p}_{n,i}$ in the second order approximations given in Online Appendix D.

Then, if the tests of H_0^i are independent, the stepdown procedure with Šidák critical values rejects the hypotheses $H_0^{(1)}, \ldots, H_0^{(r)}$ and has FWER less than or equal to α. If the tests of H_0^i are independent, as they are in our application to controlled basketball shooting experiments, then the stepdown procedure with Šidák critical values is optimal in a maximin sense; see Section 9.2 of Lehmann and Romano (2005) for a detailed discussion.[18]

3.3 Asymptotic Behavior of the Permutation Distributions

In this section, we describe the limiting behavior of the permutation distributions of $\sqrt{n}\left(\hat{P}_{n,k}\left(\mathbf{X}_i\right) - \hat{p}_{n,i}\right)$ and $\sqrt{n}\hat{D}_{n,k}\left(\mathbf{X}_i\right)$ under the null hypothesis that $\mathbf{X}_i$ is i.i.d. and under general stationary alternatives. In Section 4, these results allow us to study the power of the permutation tests outlined in Section 3.2 against particular stationary alternatives. We are aided by an appropriate central limit theorem using Stein's method (see Rinott 1994 and Stein 1986). The permutation distribution itself is random, but depends only on the number of ones in $\mathbf{X}_i$, which, under i.i.d. sampling, is binomial.

Theorem 3.3. *Under the assumption that $\mathbf{X}_i = \{X_{ij}\}_{j=1}^{\infty}$ are $i.i.d$ Bernoulli (p_i) variables, then* ***(i)*** *the permutation distribution of $\sqrt{n}T_n$ based on the test statistic $T_n = \hat{D}_{n,k}\left(X_{i1}, \ldots, X_{in}\right)$ satisfies*

$$\sup_t \left|\hat{R}_n^T\left(t\right) - \Phi\left(t/\sigma_D\left(p_i, k\right)\right)\right| \xrightarrow{P} 0$$

as $n \to \infty$, where $\xrightarrow{P}$ denotes convergence in probability and $\Phi\left(\cdot\right)$ denotes the standard normal cumulative distribution function, and

(ii) *the permutation distribution of $\sqrt{n}T_n$ based on the test statistic $T_n = \hat{P}_{n,k}\left(X_{i1}, \ldots, X_{in}\right) - \hat{p}_{n,i}$ satisfies*

$$\sup_t \left|\hat{R}_n^T\left(t\right) - \Phi\left(t/\sigma_P\left(p_i, k\right)\right)\right| \xrightarrow{P} 0$$

as $n \to \infty$, where $\sigma_P\left(p_i, k\right)$ and $\sigma_D\left(p_i, k\right)$ are given in Theorem 3.1.

Next, we study the limiting behavior of the permutation distributions of $\sqrt{n}\left(\hat{P}_{n,k}\left(\mathbf{X}_i\right) - \hat{p}_{n,i}\right)$ and $\sqrt{n}\hat{D}_{n,k}\left(\mathbf{X}_i\right)$ in possibly non-i.i.d. settings. We provide the details of this argument for $\sqrt{n}\hat{D}_{n,1}\left(\mathbf{X}_i\right)$ and note that the argument generalizes to $\sqrt{n}\hat{D}_{n,k}\left(\mathbf{X}_i\right)$ or $\sqrt{n}\left(\hat{P}_{n,k}\left(\mathbf{X}_i\right) - \hat{p}_{n,i}\right)$ for general k. We begin by considering the behavior of the permutation distribution of $\sqrt{n}\hat{D}_{n,1}\left(\mathbf{X}_i\right)$

[18] For cases where tests of H_0^i are not independent, the stepdown method of Romano and Wolf (2005) can be applied. The first step of this procedure can be used as a test of the joint null hypothesis H_0.

for fixed (nonrandom) sequences of the number of ones in n Bernoulli trials. In this case, the permutation distribution is not random, but deriving its limiting behavior is nontrivial and requires the application of a novel equicontinuity argument. In particular, let $L_n(h)$ be the permutation distribution for $\sqrt{n}T_n$ based on a data set of length n with

$$a_n = a_n(h) = \lfloor \frac{n}{2} + h\sqrt{n} \rfloor$$

ones and $n - a_n$ zeros, where $\lfloor x \rfloor$ denotes the largest integer less than or equal to x . Observe that if a_n is an integer between 0 and n, then $h = n^{-1/2}\left(a_n - \frac{n}{2}\right)$. In Online Appendix Lemma K.2, we show that under nonrandom sequences $h_n \to h$ and for $T_n = \hat{D}_{n,1}(\mathbf{X}_i)$, we have that $L_n(h_n) \xrightarrow{d} N(0,1)$. The argument generalizes if $L_n(h_n)$ is defined to be the permutation distribution for $T_n = \hat{D}_{n,1}(\mathbf{X}_i)$ based on $\lfloor np + \sqrt{n}h_n \rfloor$ number of ones, so that the fixed number of ones at time n, a_n, satisfies $n^{-1/2}(a_n - np) \to h$.

We then generalize this result to derive the limiting permutation distribution for $T_n = \hat{D}_{n,1}(\mathbf{X}_i)$ under stationary alternatives in which the number of ones in n Bernoulli trials converges in distribution under an appropriate normalization. Note that the permutation distribution $\hat{R}_n^T$ can be expressed as $L_n\left(\hat{h}_n\right)$, where

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

and $\hat{a}_n$ is the number of ones in n Bernoulli trials.

Theorem 3.4. *Suppose that $\mathbf{X}_i = \{X_{ij}\}_{j=1}^{\infty}$ is a possibly dependent, stationary Bernoulli sequence. Let $\hat{a}_n$ denote the number of ones in the first n elements of $\mathbf{X}_i$ and p_i denote the marginal probability of a success. Assume that $n^{-1/2}(\hat{a}_n - np_i)$ converges in distribution to some limiting distribution as $n \to \infty$. Then, the permutation distribution for $\sqrt{n}T_n$ based on the test statistic $T_n = \hat{D}_{n,1}(X_{i1}, \ldots, X_{in})$ converges to $N(0,1)$ in probability; that is*

$$\sup_t |\hat{R}_n^T(t) - \Phi(t)| \xrightarrow{P} 0$$

as $n \to \infty$, where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function.

Remark 3.6. The same argument can be applied to generalize Theorem 3.4 for statistics T_n equal to $\hat{D}_{n,k}(\mathbf{X}_i)$ or $\hat{P}_{n,k}(\mathbf{X}_i) - \hat{p}_{n,i}$ for general k. ■

Corollary 3.1. *Suppose that* $n^{-1/2}\left(\hat{a}_n - np_i\right)$ *converges in distribution to some limiting distribution as* $n \to \infty$*. Then, if the test statistic* T_n *is equal to* $\hat{P}_{n,k}\left(\mathbf{X}_i\right) - \hat{p}_{n,i}$ *or* $\hat{D}_{n,k}\left(\mathbf{X}_i\right)$*, the permutation distribution for* $\sqrt{n}T_n$ *satisfies*

$$\sup_t \left| \hat{R}_n^T\left(t\right) - \Phi\left(t/\sqrt{\sigma_T^2\left(p_i,k\right)}\right) \right| \xrightarrow{P} 0.$$

That is, rather than $N\left(0,1\right)$ *as the limit, one gets the same unconditional limiting distribution for these statistics as would be obtained under i.i.d. sampling with success probability* p_i*, where* p_i *denotes the marginal probability of success.*

Remark 3.7. The assumption that $n^{-1/2}\left(\hat{a}_n - np_i\right)$ converges in distribution can be weakened to the assumption that $\mathbf{X}_i$ is an α-mixing process, as the former condition follows from the latter assumption under stationarity by Theorem 1.7 of Ibragimov (1962). ■

Remark 3.8. In Online Appendix C, we discuss the asymptotic distribution of the stratified permutation distributions for the test statistics $\sqrt{ns}\bar{P}_k\left(\mathbf{X}\right)$ and $\sqrt{ns}\bar{D}_k\left(\mathbf{X}\right)$ under the condition that $n^{-1/2}\left(\hat{a}_n - np_i\right)$ converges in distribution for each individual i in $1,\ldots,s$. In particular, we find that the limiting stratified permutation distribution is the same unconditional limiting distribution for these statistics as would be obtained under i.i.d. sampling with success probability p_i for i in $1,\ldots,s$. ■

4 Power Against a Class of Markov Chain Streaky Alternatives

In this section, we study the power of the permutation tests developed in Section 3 against specific models of streaky alternatives.

4.1 A Class of Markov Chain Streaky Alternatives

We are interested in measuring the power of the permutation tests developed in Section 3 against stationary alternatives in which the parameters $\bar{\theta}_P^k\left(\mathbb{P}\right)$ and $\bar{\theta}_D^k\left(\mathbb{P}\right)$ are greater than zero. In this section, we specify a parsimonious class of Markov chain alternatives of this form. Each instance of these alternatives parameterizes $\bar{\theta}_P^k\left(\mathbb{P}\right)$ and $\bar{\theta}_D^k\left(\mathbb{P}\right)$ with two terms: ϵ and ζ. The parameter ϵ determines the "magnitude" of the deviation from randomness. The parameter ζ determines the "prevalence" of the deviation from randomness across individuals.

There are s total individuals who fall into one of two types – random and streaky. For each individual, the probability that they are streaky is ζ. That is, the number of streaky individuals in the sample is Binomial(s, ζ). For each individual i in $1, \ldots, s$, there is a Bernoulli sequence $\mathbf{X}_i = \{X_{ij}\}_{j=1}^{n}$ of length n. Each $\mathbf{X}_i$ follows a Markov chain of order 2^m. The states of the Markov chain are given by the 2^m binary tuples $\{0,1\}^m$. The event that $\mathbf{X}_i$ is in state $(x_1, \ldots, x_m) \in \{0,1\}^m$ at time j corresponds to the event

$$X_{ij} = x_1, X_{i(j-1)} = x_2, \ldots, X_{i(j-(m-1))} = x_m.$$

The sequence $\mathbf{X}_i$ is i.i.d. Bernoulli(p_i) for each random individual i. That is, for each $(x_1, \ldots, x_m)$ in $\{0,1\}^m$, the probabilities of transitioning to $(1, x_1, \ldots, x_{m-1})$ and $(0, x_1, \ldots, x_{m-1})$ are equal to p_i and $(1 - p_i)$, respectively. Streaky individuals deviate from randomness after streaks of m ones or m zeros. For these individuals the probability of a one or a zero increases by ϵ after a streak of m ones or m zeros, respectively. Formally, for each streaky individual i, the probabilities of transitioning from $\mathbf{1}^m$ to itself and $\mathbf{0}^m$ to itself are $p_i + \epsilon$ and $(1 - p_i) + \epsilon$, where $\mathbf{1}^m$ is an m-vector of ones, $\mathbf{0}^m$ is an m-vector of zeros, and ϵ is a positive real number less than $\min(1 - p_i, p_i)$. That is, the probabilities of a one or a zero after a sequence of m ones or m zeros are equal to $p_i + \epsilon$ and $(1 - p_i) + \epsilon$, respectively. For all other states, the transition probabilities are the same as for a random individual. Observe that, for a streaky individual i, $\theta_P^m(\mathbb{P}_i) = \epsilon$ and $\theta_D^m(\mathbb{P}_i) = 2\epsilon$. For a random individual, these parameters are equal to zero. Likewise, we have that under this model, $\bar{\theta}_P^m(\mathbb{P}) = \zeta\epsilon$ and $\bar{\theta}_D^m(\mathbb{P}) = 2\zeta\epsilon$.

Throughout this section, we specialize to the case that $p_i = 0.5$ for all individuals. The results are easily generalized to arbitrary p_i with more involved notation. In our empirical setting of controlled basketball shooting experiments, shot locations are chosen such that shooting percentages should be close to 0.5. The average shooting percentages for the GVT and Miller and Sanjurjo (2018a) shooting experiments are 52% and 50%, respectively.

4.2 Analytic Power Approximation

In this subsection, we derive analytic asymptotic approximations to the power of the permutation tests presented in Section 3.1 against the class of streaky alternatives specified in Section 4.1. For the sake of parsimony, we present the details of our argument when $m = k = 1$ and conclude with a discussion of the generalization of these results to cases with m and k greater than one. The

details of this generalization appear in Online Appendix H.

First, we characterize the exact asymptotic distribution of $\hat{P}_{n,1}(\mathbf{X}_i) - \hat{p}_{n,i}$ and $\hat{D}_{n,1}(\mathbf{X}_i)$ computed on the Bernoulli sequence $\mathbf{X}_i$ of a streaky individual. By invoking the characterizations of the limiting permutation distributions developed in Section 3.3, we can compute the limiting power of the permutation tests that we study against streaky alternatives local to the null hypothesis.

Theorem 4.1. *Assume that $\mathbf{X}_i = \{X_{ij}\}_{j=1}^n$ is a two-state stationary Markov Chain on $\{0,1\}$ with transition matrix given by*

$$\mathcal{P} = \begin{bmatrix} \frac{1}{2}+\epsilon & \frac{1}{2}-\epsilon \\ \frac{1}{2}-\epsilon & \frac{1}{2}+\epsilon \end{bmatrix}, \tag{4.1}$$

where $0 \leq \epsilon < \frac{1}{2}$. Then:

(i) *$\hat{P}_{n,1}(\mathbf{X}_i) - \hat{p}_{n,i}$, with $\hat{P}_{n,1}(\mathbf{X}_i)$ given by (2.4) with k equal to 1 and $\hat{p}_{n,i} = n^{-1}\sum_{j=1}^n X_{ij}$, is asymptotically normal with limiting distribution given by*

$$\sqrt{n}\left(\hat{P}_{n,1}(\mathbf{X}_i) - \hat{p}_{n,i} - \epsilon\right) \xrightarrow{d} N\left(0, \frac{1-2\epsilon+16\epsilon^2}{4-8\epsilon}\right),$$

and

(ii) *$\hat{D}_{n,1}(\mathbf{X}_i)$, given by (2.5) with k equal to 1, is asymptotically normal with limiting distribution given by*

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

as $n \to 0$.

Remark 4.1. The argument for Theorem 4.1 holds if we let ϵ vary with n such that $\epsilon_n = \epsilon + O\left(n^{-1/2}\right)$. In particular, if we take $\epsilon_n = \frac{h}{\sqrt{n}}$, then

$$\sqrt{n}\left(\hat{P}_{n,1}(\mathbf{X}_i) - \hat{p}_{n,i}\right) \xrightarrow{d} N(h, 1/4) \quad \text{and} \quad \sqrt{n}\hat{D}_{n,1}(\mathbf{X}) \xrightarrow{d} N(2h, 1).$$

Additionally, under the conditions of Theorem 4.1, Remark 3.7 indicates that as $\mathbf{X}_i = \{X_{ij}\}_{j=1}^n$ is α-mixing, $n^{-1/2}(\hat{a}_n - np_i)$ converges in distribution to some limiting distribution as $n \to \infty$, where $\hat{a}_n$ denotes the number of ones in the first n elements of $\mathbf{X}_i$. Thus, by Corollary 3.1 and Lemma 11.2.1 of Lehmann and Romano (2005), the $1-\alpha$ quantile of the permutation distribution for $\hat{D}_{n,1}(\mathbf{X})$ converges in probability to $z_{1-\alpha}$ – the $1-\alpha$ quantile of the standard normal distribution. Hence, by Slutsky's Theorem, the power of the permutation test with test statistic T_n equal to

$\hat{D}_{n,1}(\mathbf{X}_i)$ is given by

$$\mathbb{P}_i\left\{\sqrt{n}\hat{D}_{n,1}(\mathbf{X}_i) > \hat{r}_n^{\hat{D}_1}(1-\alpha)\right\} = \mathbb{P}_i\left\{\sqrt{n}\left(\hat{D}_{n,1}(\mathbf{X}_i) - \frac{2h}{\sqrt{n}}\right) > \hat{r}_n^{\hat{D}_1}(1-\alpha) - 2h\right\}$$
$$\to 1 - \Phi(z_{1-\alpha} - 2h)$$

as $n \to \infty$, where $\hat{r}_n^{\hat{D}_1}(1-\alpha)$ denotes the $1-\alpha$ quantile of the permutation distribution of $\hat{D}_{n,1}(\mathbf{X}_i)$. An analogous result holds for the permutation tests with test statistic T_n equal to $\sqrt{n}\left(\hat{P}_{n,1}(\mathbf{X}_i) - \hat{p}_{n,i}\right)$. This argument implies the following Corollary. ■

Corollary 4.1. *Consider the permutation test of the null hypothesis H_0^i that the Bernoulli sequence $\mathbf{X}_i = \{X_{ij}\}_{j=1}^n$ is i.i.d. rejecting for large values of the test statistic T_n. If the test statistic T_n is equal to $\hat{P}_{n,1}(\mathbf{X}_i) - \hat{p}_{n,i}$ or $\hat{D}_{n,1}(\mathbf{X}_i)$, then the power of this test against the alternative that $\mathbf{X}_i$ is a two-state Markov Chain on $\{0,1\}$ with transition matrix given by (4.1) and $\epsilon = h/\sqrt{n}$ converges to $1 - \Phi(z_{1-\alpha} - 2h)$ as $n \to \infty$.*

Moreover, Theorem 4.1 allows us to characterize the limiting distributions of $\bar{P}_k(\mathbf{X})$ and $\bar{D}_k(\mathbf{X})$ under the Markov chain streaky alternatives. In turn, this result allows us to derive an expression for the limiting power of stratified permutation tests of the joint null hypothesis H_0 against streaky alternatives local to the joint null hypothesis that use $\bar{P}_k(\mathbf{X})$ and $\bar{D}_k(\mathbf{X})$ as test statistics.

Corollary 4.2. *Assume that a population of s individuals are associated with the two-state stationary Markov chains $\mathbf{X}_i = \{X_{ij}\}_{j=1}^\infty$ on $\{0,1\}$ for each i in $1,\ldots,s$, such that each sequence $\mathbf{X}_i$ has probability ζ of having transition matrix given by (4.1) with $\epsilon = h/\sqrt{ns}$ and is otherwise i.i.d. Bernoulli$(1/2)$. Then:*
(i) *$\bar{P}_1(\mathbf{X})$, given by (2.6) with $k=1$, is asymptotically normal with limiting distribution given by*

$$\sqrt{ns}\bar{P}_1(\mathbf{X}) \xrightarrow{d} N(\zeta h, 1/4),$$

and
(ii) *$\bar{D}_1(\mathbf{X})$, given by (2.6) with $k=1$, is asymptotically normal with limiting distribution given by*

$$\sqrt{ns}\bar{D}_1(\mathbf{X}) \xrightarrow{d} N(2\zeta h, 1)$$

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

(iii) *The power of the stratified permutation test of the joint null hypothesis H_0 rejecting for large values of the test statistic $K_{n,s}$, for $K_{n,s}$ equal to $\bar{P}_1(\mathbf{X})$ or $\bar{D}_1(\mathbf{X})$, against the alternative specified in the conditions of this corollary converges to $1-\Phi(z_{1-\alpha}-2\zeta h)$ as $n\to\infty$ and $s\to\infty$.*

Now, we discuss the extension of these results to cases with general m and k. Details of these extensions are given in Online Appendix H. Consider the Bernoulli sequence of a single individual $\mathbf{X}_i$. The power of the permutation test of the null hypothesis H_0^i that individual i is random against the alternative that individual i is streaky with $\epsilon = h/\sqrt{n}$, rejecting for large values of the test statistic, T_n converges to

$$1-\Phi\left(z_{1-\alpha}-\phi_T(k,m,h)\right)$$

for T equal to P or D when T_n is equal to $\hat{P}_{n,k}(\mathbf{X}_i)-\hat{p}_{n,i}$ or $\hat{D}_{n,k}(\mathbf{X}_i)$, respectively. The constant $\phi_T(k,m,h)$ is a function of k, m, and h and is given by

$$\phi_T(k,m,h)=\lim_{n\to\infty}\frac{\sqrt{n}\mu_T(k,m,\epsilon_n)}{\sqrt{\sigma_T^2(1/2,k)}}$$

where $\mu_P(k,m,\epsilon_n)$ and $\mu_D(k,m,\epsilon_n)$ are the asymptotic means of $\hat{P}_{n,k}(\mathbf{X}_i)-\hat{p}_{n,i}$ or $\hat{D}_{n,k}(\mathbf{X}_i)$ when $\mathbf{X}_i$ corresponds to a streaky individual with $\epsilon_n=h/\sqrt{n}$, and $\sigma_P^2(1/2,k)$ and $\sigma_D^2(1/2,k)$ are the asymptotic variances of $\hat{D}_{n,k}(\mathbf{X}_i)$ under H_0, given by Theorem 3.1. We give expressions for $\mu_P(k,m,\epsilon)$ and $\mu_D(k,m,\epsilon)$ in terms of k, m, and ϵ in Online Appendix H. Corollary 4.1 shows that in the case that m and k are equal to one, $\phi_T(k,m,h)=2h$ for both T equal to D and P.

Table 2 displays the values of $\phi_D(k,m,h)$ for m and k between one and four. The permutation tests that reject for large values of $\hat{D}_{n,k}(\mathbf{X}_i)$ with k equal to m have the largest power against the alternative where deviations from randomness begin after m consecutive ones or zeros. The permutation test that rejects for large values $\hat{D}_{n,k}(\mathbf{X})$ with k equal to one against the alternative with m equal to one has the largest power over any combination of the test statistics and alternatives that we consider. Thus, the power of the test using $\hat{D}_{n,k}(\mathbf{X}_i)$ with k equal to one against the alternative with m equal to one gives an upper bound to the power of any of the permutation tests that we consider against any of the Markov chain streaky alternatives for a given value of ϵ. In fact, in Online Appendix I, we show that the permutation tests rejecting for large values of $\hat{D}_{n,k}(\mathbf{X}_i)$ and $\hat{P}_{n,k}(\mathbf{X}_i)-\hat{p}_{n,i}$ for k equal to one are asymptotically equivalent to the uniformly most powerful unbiased test against first order Markov chains.

Similarly, consider a collection of Bernoulli sequences for s individuals. The power of the

	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 2: Value of $\phi_D\left(k,m,h\right)$ 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 and standard deviation of the asymptotic distribution of $\hat{D}_{n,k}\left(\mathbf{X}_i\right)$ under the Markov chain alternatives considered in Section 4.1 for m and k between one and four with local perturbations $\epsilon_n = \frac{h}{\sqrt{n}}$.

stratified permutation test of the joint null hypothesis H_0 – that all of the individuals are random – against the alternative that each individual is streaky with probability ζ and $\epsilon = h/\sqrt{ns}$ rejecting for large values of the test statistic K_n^T converges to

$$1 - \Phi\left(z_{1-\alpha} - \phi_T\left(k,m,h\right)\zeta\right)$$

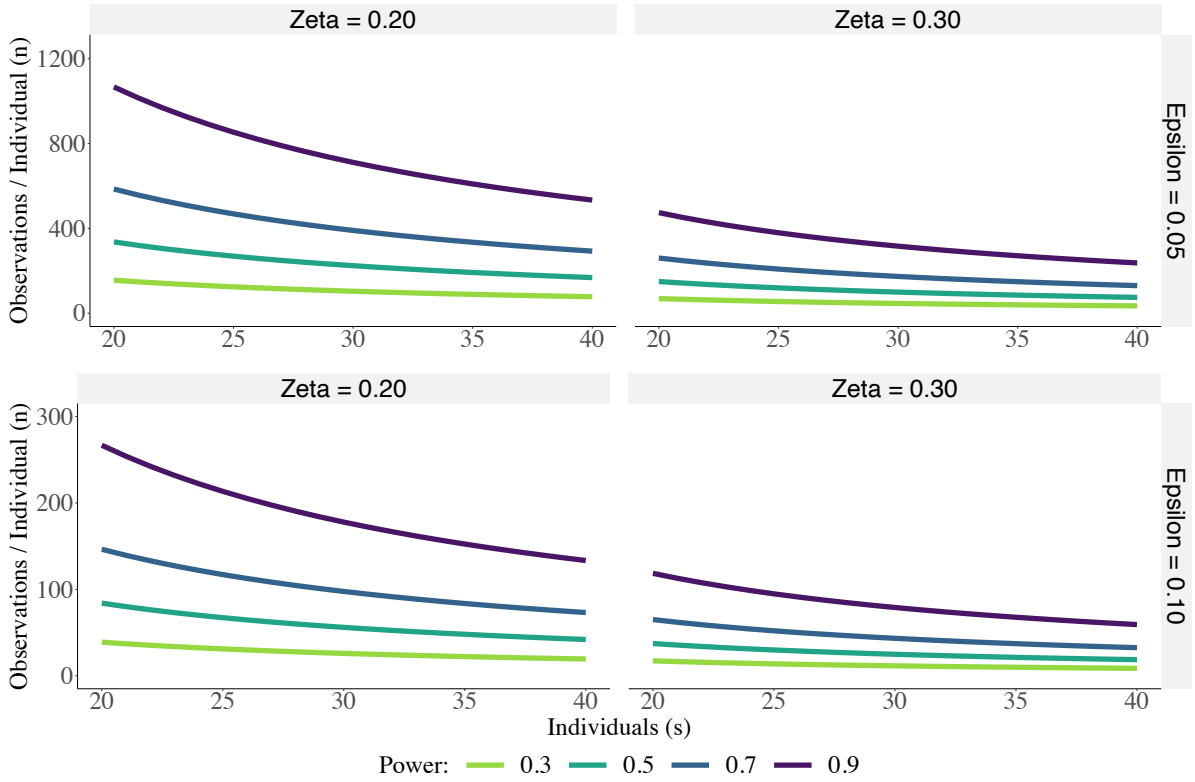
for T equal to P or D when the test statistic $K_{n,s}$ is equal to $\bar{P}_k\left(\mathbf{X}\right)$ or $\bar{D}_k\left(\mathbf{X}\right)$, respectively.

These results are very useful for power calculations when planning or assessing experiments. Suppose that we were planning on implementing an experiment where we would collect Bernoulli sequences $\mathbf{X}_i$ of length n from s individuals, and would like to test the joint null hypothesis H_0 that all sequences are i.i.d. against an alternative where each sequence is streaky with probability ζ for a given ϵ and m equal to one with a desired power of β. If we use the test statistic T_n equal to $\bar{D}_1$, our results demonstrate that the product of the number of individuals s and observations per individual n should be approximately

$$\left(\frac{z_{1-\alpha} - z_{1-\beta}}{2\zeta\epsilon}\right)^2. \tag{4.2}$$

This calculation is straightforward for any choice of parameter values, test statistic, and m by plugging in $h = \epsilon\sqrt{ns}$ and solving for ns in the limiting power expression from Corollary 4.1. Figure 1 displays the power of the test rejecting for large values of $\bar{D}_1\left(\mathbf{X}\right)$ at level $\alpha = 0.05$ against four parameterizations of ϵ and ζ for the Markov chain streaky alternative with $m = 1$ for

Figure 1: Requisite Sample Size for Power of Tests of the Joint Null Against Specified Alternatives



Notes: Figure displays the power of the permutation test of the joint null hypothesis H_0 using the test statistic $\bar{D}_1(\mathbf{X})$ against the Markov chain streaky alternative with $m = 1$ calculated by the analytic approximation in Corollary 4.2. Each panel gives the power for the test for different sample sizes n and s under a specified ϵ and ζ.

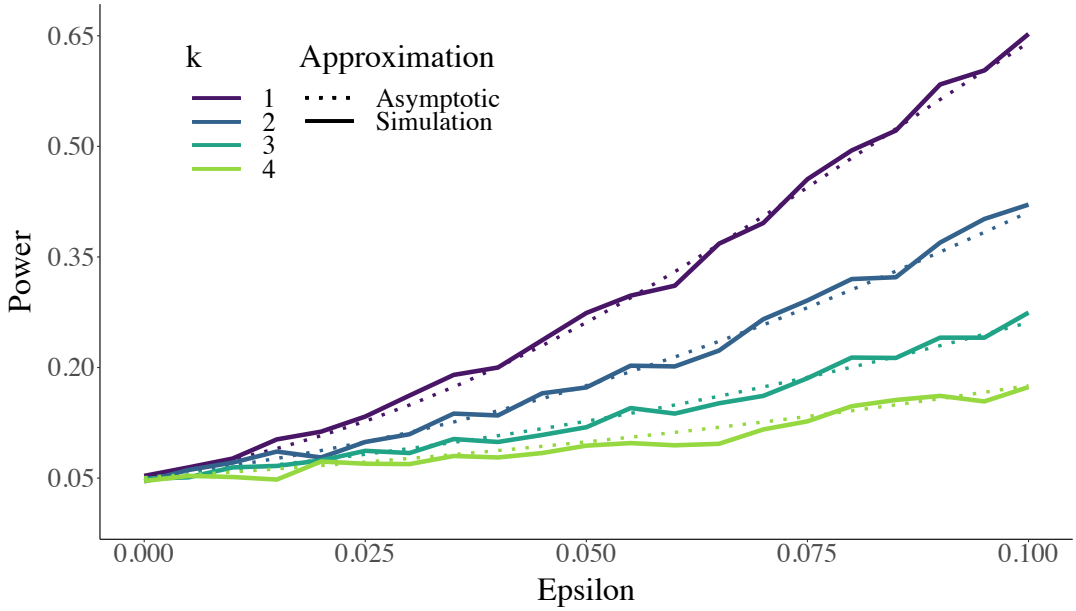
a grid of values of n and s. As we outline in the subsequent section, measuring these power curves with simulation is computationally very costly.

4.3 Simulation Analysis

In this section, we study the finite-sample quality of the asymptotic approximations to the power of the permutation tests that we consider against the Markov chain streaky alternative specified in Section 4.1. We focus on permutation tests of the individual hypotheses H_0^i that use the test statistic $\hat{D}_{n,k}(\mathbf{X}_i)$ and of the joint hypothesis H_0 that use the test statistic $\bar{D}_k(\mathbf{X})$. The results for permutation tests using $\hat{P}_{n,k}(\mathbf{X}_i) - \hat{p}_{n,i}$ and $\bar{P}_k(\mathbf{X})$ are very similar.

The simulations presented in this section require extensive parallelization. The computation is particularly expensive for measurement of the power of tests of the joint hypothesis H_0, as the

Figure 2: Power Curve for Permutation Test Rejecting for Large $\hat{D}_{n,k}(\mathbf{X}_i)$



Notes: Figure displays the power for the permutation test rejecting at level $\alpha = 0.05$ for large values of $\hat{D}_{n,k}(\mathbf{X}_i)$ for a range of ϵ in the alternative given by (4.1), $n = 100$, and each k in $1, \ldots, 4$. The solid lines display the power measured by a simulation, which takes the proportion of $2,000$ replications of Bernoulli sequences $\mathbf{X}_i$ following the transition matrix (4.1) on which the permutation test using $\hat{D}_{n,k}(\mathbf{X}_i)$ rejects H_0^i at 5% level for each value of ϵ. The dashed lines display the power calculated by the analytic approximation given by Corollary 4.1.

permutation distributions of joint test statistics of each draw of s individuals need to be computed.[19] In contrast, measuring the minimum n and s necessary to achieve a desired power against a wide range of ϵ and ζ is instantaneous with the analytic approximation given in Corollary 4.2.

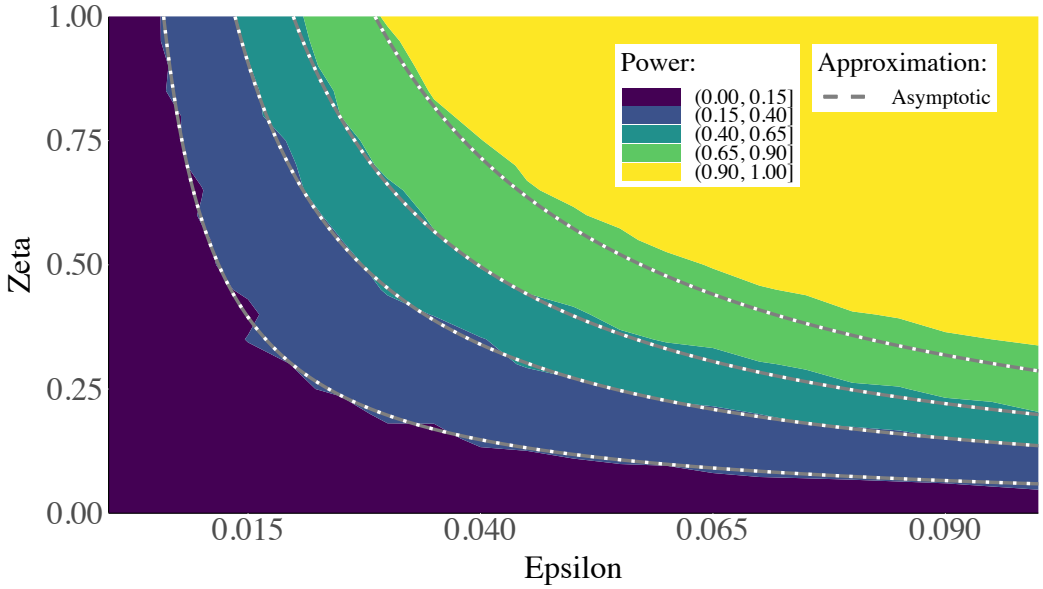
Figure 2 displays the power for the permutation test that rejects at level 0.05 for large values of $\hat{D}_{n,k}(\mathbf{X}_i)$ for k between 1 and 4 and n equal to 100 against the alternative that $\mathbf{X}_i$ is a Bernoulli sequence associated with a streaky individual with m equal to one over a grid of ϵ.[20] 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 dashed lines display the power approximated with the asymptotic expression given in Corollary 4.1. The finite-sample simulation and asymptotic approximations are remarkably close.

Figure 3 displays contours of the power surface on ϵ and ζ for the stratified permutation test rejecting at level 0.05 for large values of $\bar{D}_1(\mathbf{X})$ against the streaky alternative specified in Section

[19]The simulation underlying Figure 3 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.

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

Figure 3: Power Contours for Permutation Test Rejecting for Large $\bar{D}_1(\mathbf{X})$



Notes: Figure displays contours of the power surface on ϵ and ζ for the stratified permutation test rejecting at level 0.05 for large values of $\bar{D}_1(\mathbf{X})$ against the streaky alternative specified in Section 4.1 for n equal to 100, s equal to 26, and $m = 1$. We draw 1,000 replicates of s Bernoulli sequences $\mathbf{X}_i$ according to the streaky alternative specified in Section 4.1 with $m = 1$ for each ϵ and ζ. The estimate of the power at each ϵ and ζ is given by the proportion of replicates in which the stratified permutation test using $\bar{D}_1(\mathbf{X})$ rejects H_0 at level 0.05. The estimates of power are grouped into five regions, corresponding to sets of ϵ and ζ values with estimated power in five mutually exclusive intervals on $(0, 1]$. The white dotted curves give the asymptotic approximations to the ζ values at which the permutation test rejecting for large values of $\bar{D}_1(\mathbf{X})$ at level 0.05 has power equal to 0.15, 0.40, 0.65, 0.90, and 1.00 as a function of ϵ. The expressions for these curves are obtained by solving expression (4.2) for ζ for a given value of β in terms of ϵ.

4.1 for n equal to 100, s equal to 26, and $m = 1$.[21] For each ϵ and ζ on a two dimensional grid, we measure the power of the permutation test rejecting for large $\bar{D}_1(\mathbf{X})$ with simulation by drawing and implementing the test on 1,000 replicates of s sequences. We find that our asymptotic approximation is very accurate for most parameterizations of the model at the sample size that we study. However, our approximation appears to overestimate the power for parameterizations where the finite-sample power is close to 0.9.

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

5 Uncertainty in the Hot Hand Fallacy

In the preceding two sections, we developed inferential methods for testing whether a collection of Bernoulli sequences deviates from randomness. Equipped with these methods, we now examine two empirical questions posed formally in Section 2. First, is there evidence of positive serial dependence in basketball shooting? Second, if so, how widespread and substantial is this dependence?

We begin by providing an overview of the available data from controlled basketball shooting experiments, before addressing these two questions in succession. We conclude by discussing evidence on beliefs in serial dependence in basketball shooting, outlining a formal framework for addressing the final, behaviorally substantive, question from Section 2 – whether people systematically overestimate positive serial dependence in basketball shooting.

5.1 Controlled Shooting Experiments

We examine the evidence for serial dependence in basketball shooting provided by controlled shooting experiments. In a controlled shooting experiment, each individual is observed taking a sequence of shots under identical conditions. Although live game data from professional and collegiate basketball is abundant, these data are subject to large and ambiguous selection biases. In a live game setting, making a shot may subsequently affect defensive pressure, shot selection, and offensive strategy. Controlling for these effects is a complicated computational and statistical problem (Bocskocsky et al., 2014; Lantis and Nesson, 2019). Controlled shooting experiments provide a significantly cleaner statistical setting.

We consider the design and results of four controlled shooting experiments. The GVT shooting experiment is the only experiment designed for tests for serially dependent shooting whose data are publicly available and whose results have been peer-reviewed. Moreover, the conclusions reached in GVT and MS based on the data from the GVT shooting experiment are starkly different and have resulted in both the former consensus and current uncertainty concerning the empirical support for the hot hand fallacy in economics. Thus, we focus on the results of this experiment.

In the GVT shooting experiment, we observe shooting sequences for 26 members of the Cornell University men and women's varsity and junior varsity basketball teams.[22] Fourteen of the players are men and twelve of the players are women. For all but three players, we observe 100 shots. We

[22] We obtained the data from Miller and Sanjurjo (2018c), available at https://www.econometricsociety.org/sites/default/files/14943_Data_and_Programs.zip on April 19, 2019.

	Stratified Permutation Test of H_0 p-Value		Number of Simultaneous Rejections of H_0^i	
k	$\bar{P}_k(\mathbf{X})$	$\bar{D}_k(\mathbf{X})$	$\hat{P}_{n,k}(\mathbf{X}_i) - \hat{p}_{n,i}$	$\hat{D}_{n,k}(\mathbf{X}_i)$
	(1)	(2)	(3)	(4)
1	0.155	0.146	1	1
2	0.032	0.040	1	2
3	0.042	0.004	1	1
4	0.303	0.072	0	0

Table 3: Results of Simultaneous and Joint Hypothesis Tests for the GVT Experiment

Notes: Table displays the results of simultaneous tests of the individual null hypotheses H_0^i and tests of the joint null hypothesis H_0 for the GVT controlled shooting experiment. Columns (1) and (2) display the p-values of the stratified permutation test of H_0 using the statistics $\bar{P}_k(\mathbf{X})$ and $\bar{D}_k(\mathbf{X})$, respectively. We estimate the stratified permutation distribution of each statistic with 100,000 stratified permutations. Columns (3) and (4) display the number of rejections of H_0^i at level $\alpha = 0.05$ using the test statistics $\hat{D}_{n,k}(\mathbf{X}_i)$ and $\hat{P}_{n,k}(\mathbf{X}_i) - \hat{p}_{n,i}$ for each k in $1, \ldots, 4$, respectively. We use the stepdown procedure with Šidák critical values implemented on the p-values from the one-sided permutation test.

observe 90, 75, and 50 shots for three of the men. The experimenters determined distances from the basket at 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.

Miller and Sanjurjo (2018a) and Miller and Sanjurjo (2019) study the results of three additional controlled shooting experiments. Miller and Sanjurjo (2018a) implement an experiment with ten semi-professional Spanish basketball players. Two shooters took 300 consecutive shots in one session, seven shooters took 300 consecutive shots in each of three sessions, and one shooter took 300 shots in each of five sessions. Miller and Sanjurjo (2018a) also study data from the controlled shooting experiment presented originally in Jagacinski et al. (1979), in which six former collegiate players took 60 shots in each of nine sessions. The implementations of these experiments are otherwise very similar to the GVT experiment. Miller and Sanjurjo (2019) study the results from the annual NBA Three Point Shooting contest, in which players compete by taking rounds of 25 consecutive three point shots. They consider all 34 players who have taken more than 100 shots in this contest over the course of their careers. The average number of shots taken in this sample is 166.

5.2 Is There Positive Serial Dependence in Basketball Shooting?

The first question posed in Section 2– whether all basketball shooting is random – can be assessed with a test of the joint null hypothesis H_0. Columns (1) and (2) of Table 3 display the p-values for the stratified permutation tests of H_0 using $\bar{P}_k(\mathbf{X})$ and $\bar{D}_k(\mathbf{X})$ in the GVT shooting experiment, respectively. Both tests reject at the 5% level for k equal to 2 and 3. The test using $\bar{D}_k(\mathbf{X})$ for k equal to 3 rejects at the 1% level. These results provide reasonably strong evidence that basketball shooting is not random. However, one may be concerned that the rejection of H_0 is not overwhelmingly strong.

Assuaging this concern, columns (3) and (4) of Table 3 display the number of rejections of H_0^i at level $\alpha = 0.05$ when the p-values from the individual shooter permutation tests using $\hat{P}_{n,k}(\mathbf{X}_i) - \hat{p}_{n,i}$ and $\hat{D}_{n,k}(\mathbf{X}_i)$ are corrected with the stepdown procedure with Šidák critical values. The results are identical at level $\alpha = 0.1$. The procedure consistently rejects H_0^i for only one shooter, identified as "Shooter 109," over the set of test statistics considered.[23]

The rejection of H_0^i for Shooter 109 in the GVT experiment 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 substantial 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.

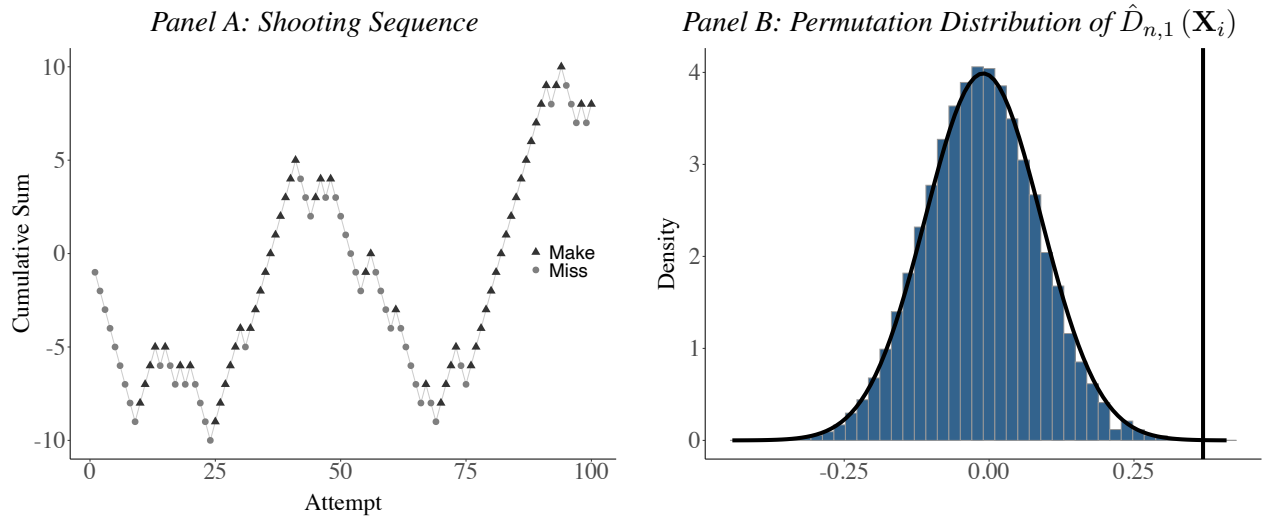
It is unlikely that a random Bernoulli sequence would generate this pattern, even among $s = 26$ sequences.[24] Panel B of Figure 4 plots the permutation distribution of $\hat{D}_{n,1}(\mathbf{X}_i)$ for Shooter 109's shooting sequence, denoting the observed value of $\hat{D}_{n,1}(\mathbf{X}_i)$ with 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 with recomputed statistics that are to the right of the observed value; this p-value is equal to 0.0001.

In fact, any evidence of positive dependence in the GVT data appears to be confined to Shooter 109. Figure 5 overlays the realized values of $\bar{D}_k(\mathbf{X})$ and $\bar{P}_k(\mathbf{X})$ from the GVT experiment on their stratified permutation distributions, displayed with horizontal black to white gradients, with

[23] Tables giving the p-values of the individual permutation tests using $\hat{P}_{n,k}(\mathbf{X}_i) - \hat{p}_{n,i}$ and $\hat{D}_{n,k}(\mathbf{X}_i)$ for k in $1, \ldots, 4$ for each shooter in the GVT shooting experiment are given in Online Appendix J.

[24] 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



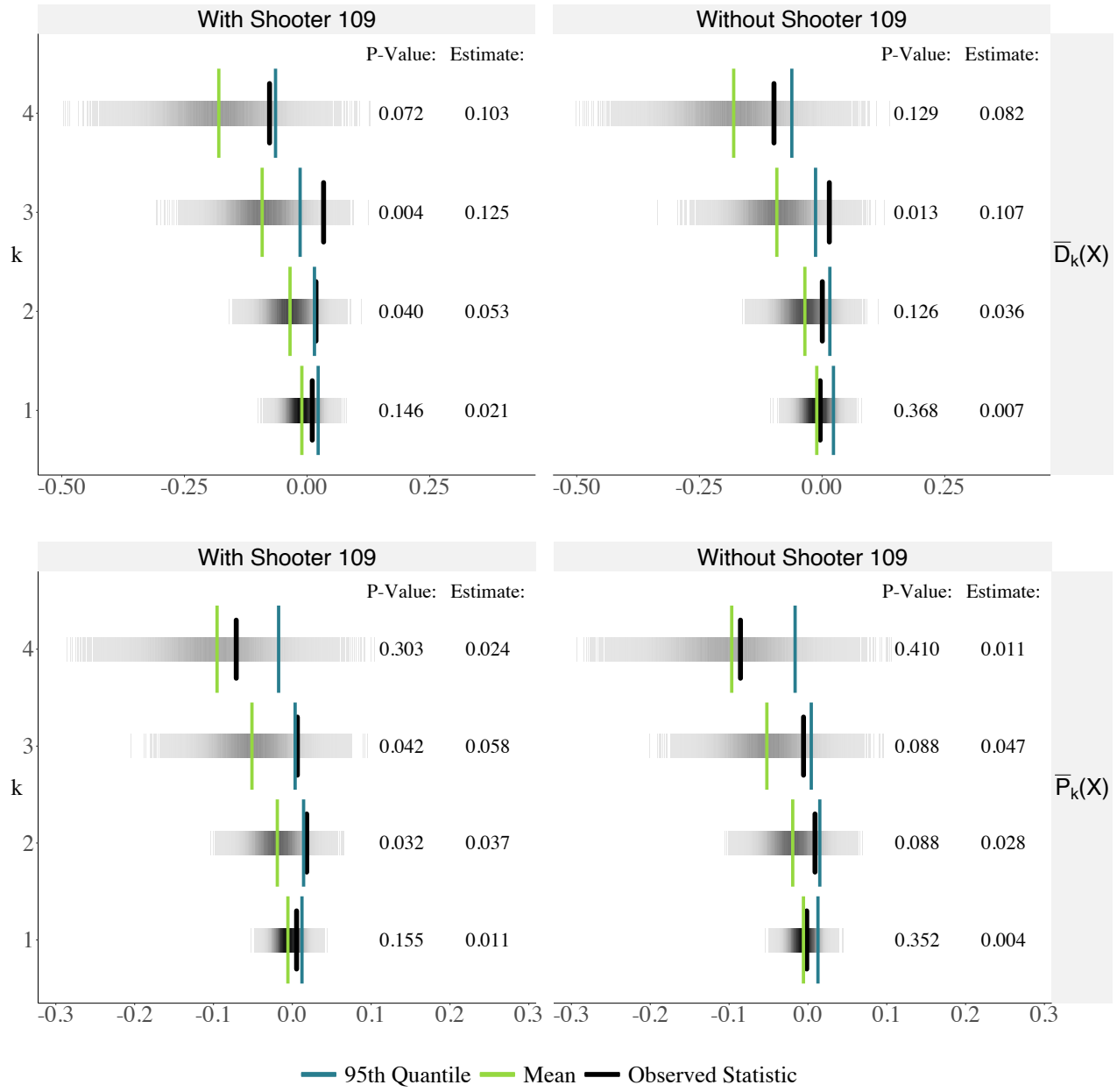
Notes: Panel A displays the cumulative sum of the sequence of makes and misses for Shooter 109. Made baskets are coded as a 1 and displayed with a black triangle and missed baskets are coded as a -1 and displayed as a grey circle. Panel B displays a density histogram of $\hat{D}_{n,1}(\mathbf{X}_i)$ 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 $N\left(\beta_D^{n,1}(p_i), n^{-1}\sigma_D^2(\hat{p}_{n,i}, 1)\right)$ in black, which is the asymptotic approximation for the permutation distribution of $\hat{D}_{n,1}(\mathbf{X}_i)$ derived in Theorem 3.1, where $\sigma_D^2(\hat{p}_{n,i}, 1)$ is given in the statement of Theorem 3.1, shifted by the small-sample bias $\beta_D^{n,1}(p_i)$. By Theorem 4 of MS, for $k = 1$, we have that $\beta_D^{n,1}(p_i) = -1/(n-1)$.

and without the inclusion of Shooter 109. The 95^{th} quantiles of these distributions are denoted by lines with squared ends. The observed statistics are denoted with vertical lines with rounded ends. The p-values of the stratified permutation tests are displayed to the right of the corresponding permutation distributions. When Shooter 109 is removed from the sample, only the joint test using $\bar{D}_k(\mathbf{X})$ for k equal to 3 is significant at the 5% level.[25]

These results are broadly consistent with the evidence from Miller and Sanjurjo (2018a) and Miller and Sanjurjo (2019). Miller and Sanjurjo (2018a) report that p-values from the stratified permutation tests of H_0 using $\bar{P}_k(\mathbf{X})$ for k equal to 3 are equal to 0.008 and 0.341 using the data from their experiment and from the Jagacinski et al. (1979) experiment, respectively, but do not report these results for other values of k or for tests using $\bar{D}_k(\mathbf{X})$. Likewise, they highlight one player from their experiment and one player from the Jagacinski et al. (1979) experiment that are uniquely streaky. The p-values of the permutation tests of H_0^i using $\hat{P}_{n,k}(\mathbf{X}_i) - \hat{p}_{n,i}$ with k equal

[25] In Online Appendix G.4, we implement a similar exercise for joint tests that use three different statistics as well as two joint testing methods that combine the results of several tests. The finding that joint tests are no longer significant after the removal of Shooter 109 is robust to these different choices of test statistics.

Figure 5: Stratified Permutation Tests of H_0 using $\bar{D}_k(\mathbf{X})$ and $\bar{P}_k(\mathbf{X})$



Notes: Figure displays the observed values of $\bar{D}_k(\mathbf{X})$ and $\bar{P}_k(\mathbf{X})$ overlaid onto the stratified permutation distribution of $\bar{D}_k(\mathbf{X})$ and $\bar{P}_k(\mathbf{X})$ under H_0 for each k in $1, \ldots, 4$. The observed values of $\bar{D}_k(\mathbf{X})$ and $\bar{P}_k(\mathbf{X})$ are indicated by vertical line segments with squared ends. The estimated 95^{th} quantile and mean of the permutation distributions under H_0 are denoted by vertical line segments with squared ends, respectively. We estimate the permutation distribution of $\bar{D}_k(\mathbf{X})$ and $\bar{P}_k(\mathbf{X})$ under H_0^i by permuting each of the $\mathbf{X}_i$'s 100,000 times separately and recomputing $\bar{D}_k(\mathbf{X})$ and $\bar{P}_k(\mathbf{X})$ for each permuted collection of sequences. The estimates of the permutation distribution are displayed in horizontal white to black gradients, shaded by the proportion of permutations whose recomputed values of $\bar{D}_k(\mathbf{X})$ or $\bar{P}_k(\mathbf{X})$ that lie in a fine partition of the x-axis. The p-values of the stratified permutation test using $\bar{D}_k(\mathbf{X})$ or $\bar{P}_k(\mathbf{X})$ are reported to the right of each distribution for each k. The difference between the observed values of $\bar{D}_k(\mathbf{X})$ and $\bar{P}_k(\mathbf{X})$ and the means of the permutation distributions, denoted by $\bar{\bar{P}}_k(\mathbf{X})$ and $\bar{\bar{D}}_k(\mathbf{X})$ and defined in (3.4), are displayed to the right of the p-values.

k	$\tilde{P}_{n,k}(\mathbf{X}_i)$	$\tilde{D}_{n,k}(\mathbf{X}_i)$
1	0.182	0.379
2	0.263	0.487
3	0.324	0.561
4	0.330	0.593

Table 4: Estimates of $\theta_P^k(\mathbb{P}_i)$ and $\theta_D^k(\mathbb{P}_i)$ for Shooter 109

Notes: Table displays the statistics $\tilde{P}_{n,k}(\mathbf{X}_i)$ and $\tilde{D}_{n,k}(\mathbf{X}_i)$, defined in (3.3), computed using data from Shooter 109 from the GVT controlled shooting experiment.

to 3 are equal to 0.003 and 0.0001, respectively. Likewise, in the analysis of the NBA Three Point Shooting contest in Miller and Sanjurjo (2019), the stratified permutation test of H_0 using $\bar{P}_k(\mathbf{X})$ for k equal to 3 has a p-value less than 0.001. While they do not report p-values of individual tests, correct for multiplicity, or report the results of the individual tests using $\hat{P}_{n,k}(\mathbf{X}_i) - \hat{p}_{n,i}$ or $\hat{D}_{n,k}(\mathbf{X}_i)$, there is one player with an abnormally significant rejection of H_0^i using a more complicated test statistic.

5.3 Which Streaky Alternatives Can Be Detected?

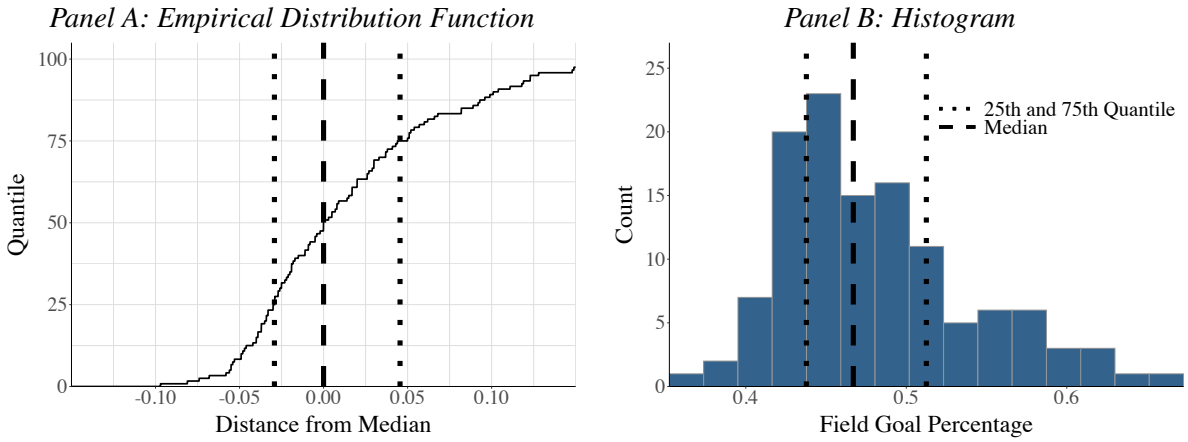
A tempting conclusion from the analysis presented in the previous subsection might be that streakiness in basketball shooting is confined to a small number of shooters – that is, there are a small number of shooters with very hot hands. However, we argue that (i) the deviation from randomness exhibited by Shooter 109 is unlikely to be indicative of what could realistically be expected from even a small proportion of basketball players and that (ii) the existing controlled shooting experiments do not have sufficient power to detect what would be realistic alternatives.

In support of the former point, Table 4 displays the statistics $\tilde{P}_{n,k}(\mathbf{X}_i)$ and $\tilde{D}_{n,k}(\mathbf{X}_i)$, defined in (3.3), computed on Shooter 109's shooting sequence. We caution that these statistics are only bias-corrected under the null hypothesis. Nevertheless, taken as estimates of $\theta_P^k(\mathbb{P}_i)$ and $\theta_D^k(\mathbb{P}_i)$, they correspond to massive deviations from randomness.

For perspective, Figure 6 displays a histogram and an empirical distribution function of the field goal percentages of NBA players in the 2018–2019 regular season.[26] The x-axis of the empirical

[26]The data were downloaded from Basketball Reference (2019). The field goal sample includes players who have attempted more than 300 field goals. Field goals are shots taken in any context of a live basketball game, other than free throws.

Figure 6: Distribution of Field Goal Shooting Percentage in the 2018-2019 NBA Season



Notes: Figure displays the distribution of the field goal shooting percentages of NBA players in the 2018–2019 regular season. Players shooting fewer than 300 field goals are omitted when displaying the distributions. The left panel displays a truncated empirical cumulative distribution and the right panel displays a histogram of the shooting percentages. To parallel the Markov streaky alternatives specified in Section 4.1, 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 vertical dashed lines give the medians of the distributions. The vertical dotted lines give the 25^{th} and 75^{th} quantiles of the distributions.

distribution function plot has been relabelled such that the median of the distribution is displayed as 0. The statistic $\tilde{D}_{n,k}(\mathbf{X}_i)$ with k equals one for Shooter 109 is equal to 0.379. Taken as an estimate of $\theta_D^k(\mathbb{P}_i)$, this corresponds 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.

Now, perhaps more realistically, suppose that the marginal shooting percentage for all shooters is 50%. For half of these shooters, shooting percentage increases and decreases by half of the interquartile range of the distribution of field goal percentages of NBA players after making their previous m shots or missing their previous m shots, respectively. The other half of shooters remain 50% shooters at all times. This is an instance of Markov chain streaky alternative studied in Section 4.1, parameterized as $\epsilon = 0.038$ and $\zeta = 0.5$ for a given value of m.[27]

We argue that this parameterization is a conservative upper bound on the set of deviations from randomness consistent with the variation in NBA shooting percentages.[28] In our choice of ϵ, in effect, we assume that the variation in shooting percentages within players is less than the variation

[27] A similar parameterization is obtained if we consider half the distance between the top and bottom terciles of the NBA free throw distribution. In fact, in this case, $\epsilon = 0.0384$. However, the distribution of free throw percentages has a larger median (0.806) and a long left tail. The data were downloaded from Basketball Reference (2019). The free throw sample includes players who have attempted more than 125 free throws.

[28] Note that, additionally, the power of the joint permutation tests using $\tilde{D}_{n,k}(\mathbf{X}_i)$ decreases if marginal shooting percentages are different than $1/2$ or are sampled from a distribution.

in shooting percentages across players. As the proportion of players in each experiment with large values of $\tilde{P}_{n,k}(\mathbf{X}_i)$ and $\tilde{D}_{n,k}(\mathbf{X}_i)$ is small, imposing $\zeta = 0.5$ is likely to be very conservative.[29] We consider two relaxations of this upper bound, which in our judgment do not seem less reasonable: reducing the proportion of streaky individuals to 25% ($\zeta = 0.25$) and assuming that streaky individuals increase and decrease their shooting percentages by half the distance between the 66$^{\text{th}}$ and 33$^{\text{rd}}$ quantiles of the NBA field goal percentage distribution after making or missing m shots ($\epsilon = 0.024$).

We consider $m = 3$ as the benchmark parameterization based on the emphasis in GVT and MS, although parameterizations with $m = 1$ give a more conservative upper bound on power. Specifically, GVT describe streaks of three makes (misses) as "hot" ("cold") periods and emphasize statistics with $k = 3$. Likewise, Miller and Sanjurjo (2018a) and Miller and Sanjurjo (2019) denote streaks of three makes (misses) as "hot" ("cold") streaks. Miller and Sanjurjo (2018a) argue that they are interested in detecting alternatives with $m = 3$, citing literature in psychology indicating that people perceive streaks to begin at three (Carlson and Shu, 2007) and only implementing statistics with k equal to 3.

Figure 7 displays our asymptotic approximation to the power of the stratified permutation test of H_0 using the test statistic $\bar{D}_k(\mathbf{X})$ against the Markov chain streaky alternative studied in Section 4.1 over a grid of ϵ, for ζ equal to 0.25 and 0.5, $m = k$ for k in $1, \ldots, 4$, and for n and s equal to their values in the GVT, Miller and Sanjurjo (2018a), Jagacinski et al. (1979), and Miller and Sanjurjo (2019) controlled basketball shooting experiments.[30] The vertical black lines denote $\epsilon = 0.024$ and $\epsilon = 0.038$.

All four experiments lack adequate power against conservative parameterizations of the Markov chain streaky alternative specified above.[31] No experiment has adequate power for $\zeta = 0.25$ or for $\epsilon = 0.024$ for any value of m. No experiment has adequate power for $\epsilon = 0.038$ and $\zeta = 0.5$ for $m = 3$. For $\epsilon = 0.038$ and for $\zeta = 0.5$, Miller and Sanjurjo (2018a) and Miller and Sanjurjo (2019) have reasonable power for $m = 1$ and Miller and Sanjurjo (2018a) has reasonable power

[29] See Online Appendix Figures 7 and 8.

[30] Note that in the setting of Miller and Sanjurjo (2019) – the NBA Three Point Shooting contest – participants take different numbers of shots. For this example we replace ns with s times the average number of shots taken, obtained from Table 1 of Miller and Sanjurjo (2019) , which gives an approximation to the power of the stratified permutation test using the sample-size weighted average of $\hat{D}_{n,k}(\mathbf{X}_i)$ across individuals.

[31] It follows that multiple tests of the individual hypotheses are even less powerful. Indeed, as any multiple hypothesis testing method that controls the FWER must be constructed with the closure method (Romano et al., 2011), and under the closure method the individual hypothesis H_0^i is rejected if all joint tests of subsets of $\left\{H_0^i : i \in 1, \ldots, s\right\}$ containing H_0^i are rejected, then even the rejection of the joint null hypothesis H_0 is not sufficient to obtain any rejections of the individual hypotheses H_0^i.

for $m = 2$.

5.4 How Much Positive Serial Dependence is There in Basketball Shooting?

An answer to the second question – how widespread and substantial is dependence in basketball shooting — can be assessed with estimates of the individual parameters $\theta_P^k(\mathbb{P}_i)$ and $\theta_D^k(\mathbb{P}_i)$ and the average parameters $\bar{\theta}_P^k(\mathbb{P})$ and $\bar{\theta}_D^k(\mathbb{P})$. However, measurement of the magnitude of the average streakiness of basketball shooting can only be distinguished from zero if reasonable deviations from randomness can be detected, and as we have argued in the previous subsection, this is not the case for the existing controlled shooting experiments.

Returning to Figure 5 and the GVT shooting experiment, the statistics $\bar{\bar{P}}_k(\mathbf{X})$ and $\bar{\bar{D}}_k(\mathbf{X})$, defined in (3.3), are given by the difference between the observed statistics and the means of their permutation distributions and are displayed under "Estimate" on the far right hand side of each panel. These statistics are unbiased for $\bar{\theta}_P^k(\mathbb{P})$ and $\bar{\theta}_D^k(\mathbb{P})$ under H_0. These estimates are large for k equal to 2 and 3, and particularly for $\bar{\theta}_D^k(\mathbb{P})$ with k equal to 3. However, with the exception of $\bar{\theta}_D^k(\mathbb{P})$ for $k = 3$, the permutation tests no longer reject at the 5% level after the removal of Shooter 109.

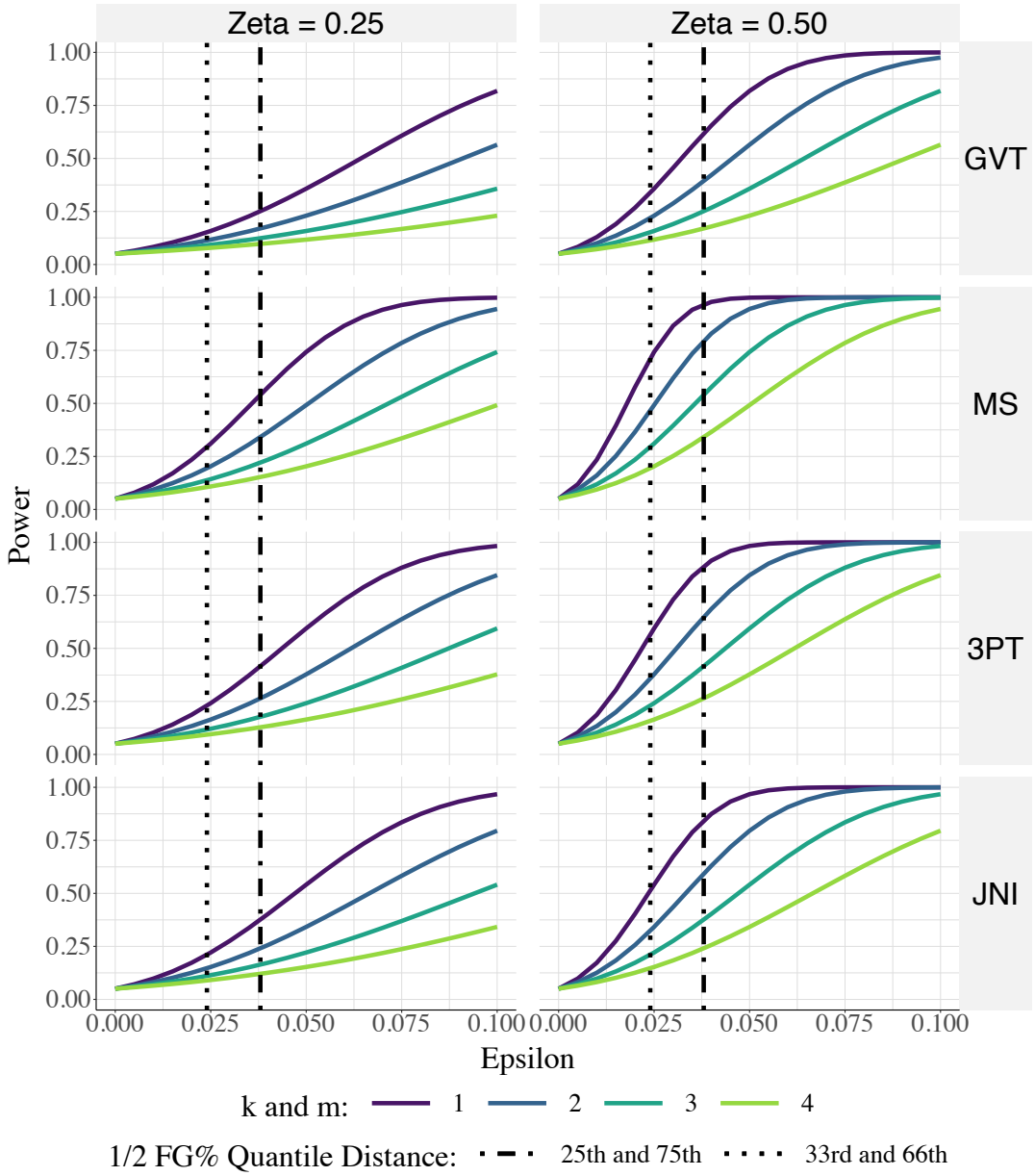
Recall that in Theorem 3.2, we show that permutation tests are the only tests with exact type 1 error control. This implies that if a stratified permutation test of H_0 using the test statistic $K_{n,s}$ does not reject the null hypothesis, then any lower confidence bound constructed with test inversion using the test statistic $K_{n,s}$ that obtains exact 95% coverage will be below zero. Thus, with the exception of $\bar{\theta}_D^k(\mathbb{P})$ for $k = 3$, any lower confidence bound for the parameters $\bar{\theta}_P^k(\mathbb{P})$ and $\bar{\theta}_D^k(\mathbb{P})$ constructed with statistics $\bar{\bar{P}}_k(\mathbf{X})$ and $\bar{\bar{D}}_k(\mathbf{X})$ using test inversion that obtains exact 95% coverage will be below zero. The analogous lower confidence bound for $\bar{\theta}_D^k(\mathbb{P})$ for $k = 3$ will be close to zero.

In practice, for experiments with significantly larger samples, we advocate for the application of general bootstrap methods for stationary time series (see Lahiri (2013)), such as the moving blocks bootstrap (Liu and Singh, 1992; Künsch, 1989), the stationary bootstrap (Politis and Romano, 1994), or subsampling (Politis et al., 1999).[32]

To conclude, existing randomized shooting experiments are insufficiently powered to detect deviations from randomness that we argue would be consistent with a realistic parameterization

[32]Additionally, simultaneous confidence regions for the individual parameters $\theta_P^k(\mathbb{P}_i)$ and $\theta_D^k(\mathbb{P}_i)$ can be constructed using a Šidák correction for multiplicity.

Figure 7: Power of Joint Tests of H_0 for Four Controlled Basketball Shooting Experiments



Notes: Figure displays asymptotic approximations to the power of the stratified permutation tests of the joint null hypothesis H_0 using the test statistic $\bar{D}_k(\mathbf{X})$ against the Markov chain streaky alternative specified in Section 4.1 over a grid of ϵ, for ζ equal to 0.25 or 0.5, for $m = k$ for k in $1, \ldots, 4$, and for n and s at the values of the GVT, Miller and Sanjurjo (2018a) (MS), Jagacinski et al. (1979) (JNI), and Miller and Sanjurjo (2019) (3PT) controlled basketball shooting experiments. An expression for this asymptotic approximation is given in Online Appendix Corollary H.3, with $h = \epsilon\sqrt{ns}$. The vertical dotted and dot-dashed black lines denote the $\epsilon = 0.024$ and $\epsilon = 0.038$, which are consistent with half the distance between the 33rd and the 66th quantiles and the 25th and the 75th quantiles of the distribution of NBA field goal percentages, respectively. Note that in the setting of Miller and Sanjurjo (2019), the NBA Three Point Shooting contest, participants take different numbers of shots. For this example we replace ns with s times the average number of shots taken, which gives an approximation to the power of the stratified permutation test using the sample-size weighted average of $\hat{D}_{n,k}(\mathbf{X}_i)$ across individuals.

of positive dependence in basketball shooting. These experiments are therefore unable to provide an informative estimate of the mean or dispersion of the serial dependence in basketball shooting. This conclusion could be challenged by a strong and robust rejection of H_0, but the rejection of H_0, at least in the case of the GVT experiment, 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.

5.5 Do People Overestimate Positive Serial Dependence?

If we had an informative estimate of the positive serial dependence of an average shooter, a comparison with evidence on expectations of serial dependence in basketball shooting would provide a direct test of the hot hand fallacy. Specifically, we advocate for a test of the null hypothesis that $\bar{\theta}_P^k(\mathbb{P})$ and $\bar{\theta}_D^k(\mathbb{P})$ are equal to an audience's expectations of $\bar{\theta}_P^k(\mathbb{P})$ and $\bar{\theta}_D^k(\mathbb{P})$ against the alternative that the audience's expectations are larger. We find that the available evidence on expectations of streakiness in basketball shooting suffers either from prohibitive methodological flaws or is not directly comparable to estimates of $\bar{\theta}_P^k(\mathbb{P})$ and $\bar{\theta}_D^k(\mathbb{P})$.

GVT measure beliefs with two methods. First, they implement 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 $\theta_D^k(\mathbb{P}_i)$ when $k = 1$ and indicating that these expectations are approximately 0.1 and 0.04, respectively. 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 serial dependence.

Second, GVT attempt to infer beliefs from observations of incentivized decisions. In their controlled shooting experiment, prior to each shot, each shooter and an observer choose whether to bet "high" or "low." If an 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.[33] Miller and Sanjurjo (2017) find that the average correlation between the bets and the shot outcomes is 0.07, that the increase in the probability of predicting a make after a make is 0.077, and that these estimates are significantly different from zero.

Unfortunately, these estimates do not pin down an estimate of $\bar{\theta}_P^k(\mathbb{P})$ or $\bar{\theta}_D^k(\mathbb{P})$. Assume that individuals only bet on a make if they believe that there is greater than a 50% chance of a make. In this case, if we observe infinite shots, the proportion of shots in which an individual predicts a make is equal to the proportion of shots in which the individual expects that the probability of a make is greater than 50%. This proportion is not in general equal to the individual's average expectation of the probability of a make.

Miller and Sanjurjo (2018a) implement a survey of the participants in their experiment, asking the basketball players to rate how likely their teammates are to make a shot following a sequence of three makes on an arbitrary scale from -3 to 3. Again, this does not identify an estimate of the player's expectations of the serial dependence in basketball shooting.

In our review of the literature, we are unable to find estimates of beliefs in serial dependence in basketball shooting that directly translate to estimates of people's expectations of $\bar{\theta}_P^k(\mathbb{P})$ or $\bar{\theta}_D^k(\mathbb{P})$. 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 $\bar{\theta}_P^k(\mathbb{P})$ or $\bar{\theta}_D^k(\mathbb{P})$.

Future studies should estimate expectations of $\bar{\theta}_P^k(\mathbb{P})$ or $\bar{\theta}_D^k(\mathbb{P})$ that are directly comparable to measurements of $\bar{\theta}_P^k(\mathbb{P})$ or $\bar{\theta}_D^k(\mathbb{P})$. 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 observer 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 provide a direct estimate of beliefs in $\bar{\theta}_P^k(\mathbb{P}_i)$ and $\bar{\theta}_D^k(\mathbb{P}_i)$. Nevertheless, it is important to ensure that these measurements of beliefs are made on a sample large enough to ensure adequate precision for an

[33] These data are not publicly available.

informative comparison to measurements of $\bar{\theta}_P^k(\mathbb{P}_i)$ and $\bar{\theta}_D^k(\mathbb{P}_i)$.

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 et al. (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 (2018d), 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 four controlled basketball shooting experiments.

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 Gilovich et al. (1985) shooting experiment. This rejection is robust to standard multiple testing corrections, providing strong evidence that basketball shooting is not perfectly random. However, evidence against randomness in that experiment 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.

Data Availability Statement: The data and code underlying this article are available on Zenondo as "Replication package for: Uncertainty in the Hot Hand Fallacy: Detecting Streaky Alternatives to Random Bernoulli Sequences" at http://doi.org/10.5281/zenodo.4563661.

References

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

Appelbaum, B. (2015). Streaks like daniel murphy's aren't necessarily random. *The New York Times*.

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

Barberis, N. (2018). Psychology-based models of asset prices and trading volume. In *Handbook of Behavioral Economics: Applications and Foundations 1*, volume 1, pages 79–175. Elsevier.

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

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

Basketball Reference (2019). *2018-19 NBA Player Stats: Totals*. https://www.basketball-reference.com/leagues/NBA_2019_totals.html [Accessed: July 16, 2019].

Benjamin, D. J. (2019). Errors in probabilistic reasoning and judgment biases. In *Handbook of Behavioral Economics: Applications and Foundations 1*, volume 2, pages 69–186. Elsevier.

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*. Citeseer.

Bradley, R. C. (1986). Basic properties of strong mixing conditions. In *Dependence in Probability and Statistics*, pages 165–192. Springer.

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

Carlson, K. A. and Shu, S. B. (2007). The rule of three: How the third event signals the emergence of a streak. *Organizational Behavior and Human Decision Processes*, 104(1):113–121.

Chay, K. Y., Hoynes, H. W., and Hyslop, D. R. (1999). A non-experimental analysis of true state dependence in monthly welfare participation sequences. *Proceedings of the American Statistical Association*, pages 9–17.

Cohen, B. (2015). The 'hot hand' debate gets flipped on its head. *The Wall Street Journal*.

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

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

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

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

Haberstroh, T. (2017). He's heating up, he's on fire! klay thompson and the truth about the hot hand. *ESPN*.

Harrison, G. W. and Rutström, E. E. (2008). Experimental evidence on the existence of hypothetical bias in value elicitation methods. *Handbook of Experimental Economics*, 1:752–767.

Heckman, J. J. (1981). Heterogeneity and state dependence. *Studies in Labor Markets*, pages 91–140.

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):93–130.

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

Jagacinski, R. J., Newel, K. M., and Isaac, P. D. (1979). Predicting the success of a basketball shot at various stages of execution. *Journal of Sport Psychology*, 1(4):301 – 310.

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

Johnson, G. (2015). Gamblers, scientists and the mysterious hot hand. *The New York Times*.

Kahneman, D. (2011). *Thinking, Fast and Slow*. Macmillan.

Keane, M. P. (1997). Modeling heterogeneity and state dependence in consumer choice behavior. *Journal of Business & Economic Statistics*, 15(3):310–327.

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

Künsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *Annals of Statistics*, 17(3):1217–1241.

Lahiri, S. N. (2013). *Resampling Methods for Dependent Data*. Springer, NY.

Lantis, R. M. and Nesson, E. T. (2019). Hot shots: An analysis of the 'hot hand' in nba field goal and free throw shooting. Technical report, National Bureau of Economic Research.

Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses*. Springer, NY, 3^{rd} edition.

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

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

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

Miller, J. B. and Sanjurjo, A. (2017). A visible (hot) hand? expert players bet on the hot hand and win. *University of Alicante mimeo*.

Miller, J. B. and Sanjurjo, A. (2018a). A cold shower for the hot hand fallacy: Robust evidence that belief in the hot hand is justified. *University of Alicante mimeo*.

Miller, J. B. and Sanjurjo, A. (2018b). Momentum isn't magic–vindicating the hot hand with the mathematics of streaks. *Scientific American*.

Miller, J. B. and Sanjurjo, A. (2018c). Supplement to "surprised by the hot hand fallacy? a truth in the law of small numbers.". *Econometrica*, 86(6):2019–2047.

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

Miller, J. B. and Sanjurjo, A. (2019). Is it a fallacy to believe in the hot hand in the nba three-point contest? *University of Alicante mimeo*.

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

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

Politis, D. N. and Romano, J. P. (1994). The stationary bootstrap. *Journal of the American Statistical Association*, 89(428):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):775–816.

Rabin, M. and Vayanos, D. (2010). The gamblers and hot-hand fallacies: theory and applications. *Review of Economic Studies*, 77(2):730–778.

Rao, J. M. (2009). Experts' perceptions of autocorrelation: The hot hand fallacy among professional basketball players.

Remnick, D. (2017). Bob dylan and the 'hot hand.'. *The New Yorker*.

Rinott, Y. (1994). On normal approximation rates for certain sums of dependent random variables. *Computational and Applied Mathematics*, 55(2):134–143.

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

Romano, J. P. and Wolf, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association*, 100(469):94–108.

Stern, H. S. and Morris, C. N. (1993). A statistical analysis of hitting streaks in baseball: Comment. *Journal of the American Statistical Association*, 88(242):1189–1194.

Stone, D. F. (2012). Measurement error and the hot hand. *Scientific American*, 66(1):61–66.

Thaler, R. H. and Sunstein, C. R. (2009). *Nudge: Improving Decisions about Health, Wealth, and Happiness*. Penguin.

Torgovitsky, A. (2019). Nonparametric inference on state dependence in unemployment. *Econometrica*, 87(5):1475–1505.

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

Tversky, A. and Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458.

Wardrop, R. L. (1999). Statistical tests for the hot-hand in basketball in a controlled setting.