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

David M. Ritzwoller
(joint with J.P. Romano)

Stanford University

UC Berkeley CDAR , February 2021

The Problem

- Observe s Bernoulli sequences of length n
- Interested in testing null hypothesis that sequences are i.i.d.
- Interpretation pivotal in development of behavioral economics

Context

- Tversky and Kahneman (1971, *Psychological Bulletin*)
  - Belief in the Law of Small Numbers
  - Small samples overly "representative" of "essential characteristics"
- **Example:** Choice of experimental sample size
- **Impact:** Influential in Behavioral Economics and Finance
  - Barberis and Thaler (2003, *Handbook of the Economics of Finance*)
  - Barberis (2018, *Handbook of Behavioral Economics*)

Context

- **Hot Hand Fallacy:** Behavioral bias attributable to belief in the law of small numbers
  - Tendency to perceive streaks of consecutive successes as overly representative of positive dependence

- Gilovich, Vallone, and Tversky (1985, *Cognitive Psychology)*
  - Document widespread belief in "the hot hand"
  - Controlled basketball shooting experiment
  - Fail to reject i.i.d. shooting

- **Impact:** Academic consensus for following three decades
  - Kahneman (2011, *Thinking Fast and Slow*)
  - Thaler and Sunstein (2009, *Nudge*)

Context

Miller and Sanjurjo (2018, *Econometrica*) challenge the GVT results

- Identify a subtle, significant, second-order bias
- Revisit the GVT shooting data
- Argue that the GVT results are reversed

Context

"Uncovered critical flaws ... sufficient to not only invalidate the most compelling evidence against the hot hand, but even to vindicate the belief in streakiness." - Miller and Sanjurjo *(2018, Scientific American)*

"These new analyses re-opens–but does not answer–the key question of whether there is a hot hand bias, i.e., a belief in a stronger hot hand than there really is." - Benjamin *(*2018, *Handbook of Behavioral Economics)*

Objective

Objective: Clarify and quantify uncertainty in the evidence that basketball shooting experiments have contributed to our understanding of the hot hand fallacy

Objective

Objective: Clarify and quantify uncertainty in the evidence that basketball shooting experiments have contributed to our understanding of the hot hand fallacy

- **Develop** formal statistical framework for testing the randomness of a set of Bernoulli sequences
- **Emphasize** distinction between individual, joint, and multiple testing
- **Measure** finite-sample power with a local asymptotic approximation
- **Provide** comprehensive re-analysis of evidence from controlled basketball shooting experiments

Objective

Objective: Clarify and quantify uncertainty in the evidence that basketball shooting experiments have contributed to our understanding of the hot hand fallacy

- **Develop** formal statistical framework for testing the randomness of a set of Bernoulli sequences
- **Emphasize** distinction between individual, joint, and multiple testing
- **Measure** finite-sample power with a local asymptotic approximation
- **Provide** comprehensive re-analysis of evidence from controlled basketball shooting experiments

- **This Talk:** Interweave problem formalization, theoretical development, and empirics

Outline

Controlled Basketball Shooting Experiments

- Basketball players are observed taking a sequence of incentivized shots under identical conditions
- We focus on the GVT shooting experiment
  - Data are publicly available
  - Results from GVT and MS based on these data are starkly different
  - Drives former consensus and current uncertainty
- Shooting sequences of 100 shots for 26 members of Cornell's men and women's varsity and junior varsity basketball teams
- Miller and Sanjurjo analyze three additional experiments
  - We discuss the design and results of these experiments

Outline

Data

Observe s shooters; each shoots n consecutive shots:

- $\mathbf{X}_i = \left\{X_{ij}\right\}_{j=1}^{n} :=$ Vector of outcomes for shooter i
- $\mathbf{X} = \{\mathbf{X}_i\}_{i=1}^{s} :=$ Matrix of outcomes for all shooters

Null Hypotheses

Three Testing Problems:

- Joint null hypothesis:

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

- Individual hypotheses:

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

- Multiple hypothesis problem: Testing H_0^i simultaneously

Streaky Alternatives

Stationarity: Assume shot outcomes follow stationary Bernoulli (p_i) processes $\mathbb{P}_i$, independent across shooters, with $\mathbb{P} = \{\mathbb{P}_i\}_{i=1}^{s}$

Streakiness: We consider alternatives in which

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

where

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

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

are greater than zero for some integer k

Test Statistics

Test Statistics:

$\hat{p}_{n,i} := \frac{1}{n}\sum_{j=1}^{n} X_{ij}$

$\hat{P}_{n,k}(\mathbf{X}_i) := \frac{V_{ik}}{V_{i(k-1)}}$, with $V_{ik} = \sum_{j=1}^{n-k} Y_{ijk}$ and $Y_{ijk} = \prod_{l=j}^{j+k} X_{il}$,
i.e., proportion of makes after k consecutive makes

$\hat{D}_{n,k}(\mathbf{X}_i) := \frac{V_{ik}}{V_{i(k-1)}} - \frac{W_{ik}}{W_{i(k-1)}}$ with $W_{ik} = \sum_{j=1}^{n-k} Z_{ijk}$ and $Z_{ijk} = \prod_{l=j}^{j+k}(1 - X_{il})$,
i.e., proportion of makes after k consecutive makes minus proportion of makes after k consecutive misses

Averages Over Shooters:

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

Outline

Multiplicity

- Suppose that we would like to identify *which* shooters deviate from randomness
- The probability of a false rejection of at least one H_0^i increases rapidly with s
- **Example:** If s is equal to 10, all H_0^i true and testing at level $\alpha = 0.05$, then probability of at least one false rejection ≈ 0.4.
- **Goal:** Control the familywise error rate (FWER), i.e., the probability of at least one false rejection
- **Solution:** Simultaneous Inference / Multiple Testing
  - Stepdown procedure with Šidák critical values

Outline

Asymptotic Behavior of the Test Statistics

Theorem (Asymptotics Under Null Hypothesis)

Under H_0^i,

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

as $n \to \infty$, *where*

$$\sigma_P^2\left(p_i,k\right)=p_i^{1-k}\left(1-p_i\right)\left(1-p_i^k\right) \text{ and } \sigma_D^2\left(p_i,k\right)=\left(p_i\left(1-p_i\right)\right)^{1-k}\left(\left(1-p_i\right)^k+p_i^k\right)$$

Miller and Sanjurjo (2018): Finite-Sample Bias

Table: Finite-Sample Behavior of Plug-in Statistics

	$\hat{P}_{n,k}(\mathbf{X}_i) - \hat{p}_{n,i}$		$\hat{D}_{n,k}(\mathbf{X}_i)$	
k	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

Note: We take 100,000 draws of Bernoulli$(1/2)$ random variables of length 100. Type 1 error rate refers to the proportion of realizations of $\hat{P}_{n,k}(\mathbf{X}_i) - \hat{p}_{n,i}$ and $\hat{D}_{n,k}(\mathbf{X}_i)$ that exceed the 0.95 quantile of their asymptotic distributions.

Permutation Tests

- Permutation tests automatically account for finite-sample biases
- Permutation distribution for $\sqrt{n}T_n(\mathbf{X}_i)$, any real-valued test statistic, is given by

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

  where $\Pi(n)$ is the set of permutations of $\{1, \ldots, n\}$.

- Reject at level α if $\sqrt{n}T_n(\mathbf{X}_i)$ is $>$ the $1-\alpha$ quantile of $\hat{R}_n^T$
- Control the type 1 error rate exactly under H_0^i
  - In fact, *only* tests that control type 1 error rate exactly
- Similar procedure holds for joint statistics $\bar{P}_k(\mathbf{X})$ and $\bar{D}_k(\mathbf{X})$, permuting each sequence separately

Bias-Corrected Estimation

- Consider the mean of the permutation distribution of the statistic $T_n(\mathbf{X}_i)$

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

- Under H_0^i, the expectation of $\hat{\eta}(\mathbf{X}_i, T_n)$ is exactly equal to the expectation of $T_n(\mathbf{X}_i)$

- Suggests the bias-corrected estimators

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

  and their averages

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

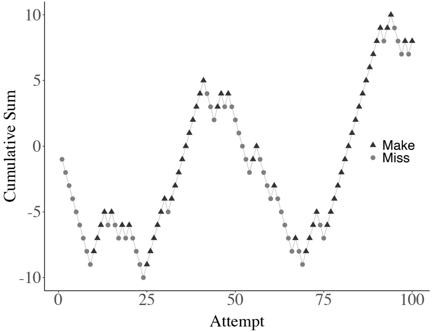
Outline

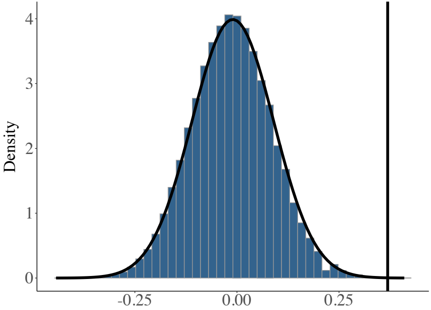
Implementation on GVT Data

k	Stratified Test of H_0 p-Value		Simultaneous Rejections of H_0^i	
	$\bar{P}_k(\mathbf{X})$	$\bar{D}_k(\mathbf{X})$	$\hat{P}_{n,k}(\mathbf{X}_i) - \hat{p}_{n,i}$	$\hat{D}_{n,k}(\mathbf{X}_i)$
	(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

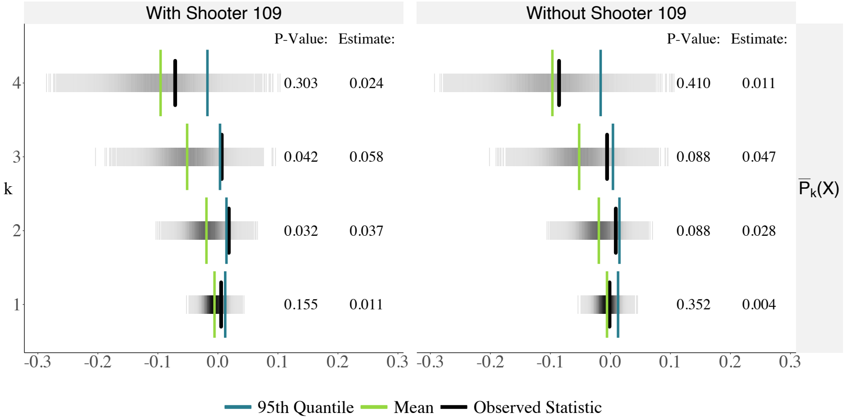
Table: Results of Simultaneous and Joint Hypothesis Tests for the GVT Experiment

Shooter 109: Shooting Sequence and Permutation Distribution

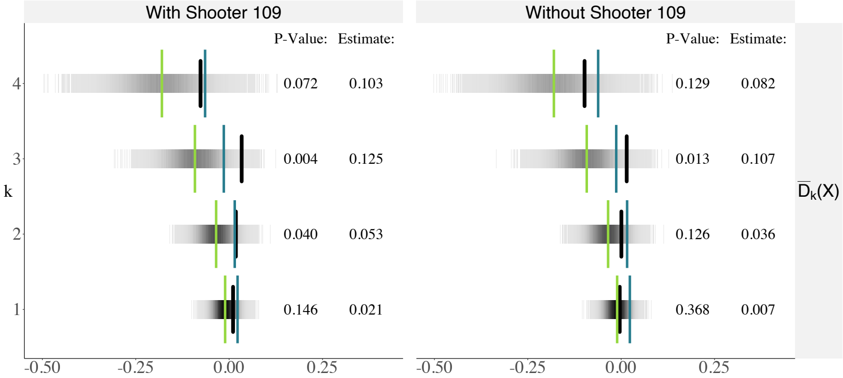




Stratified Permutation Tests



Stratified Permutation Tests



Taking Stock

A Tempting Conclusion:

- Streakiness is confined to a small number of players
- Small number of shooters with very hot hands

We Argue:

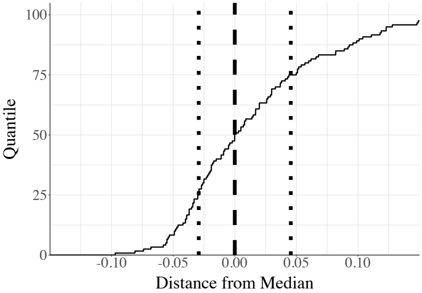
1. Deviation from randomness by Shooter 109 is unlikely to be indicative of what could realistically be expected from even a small proportion of basketball players

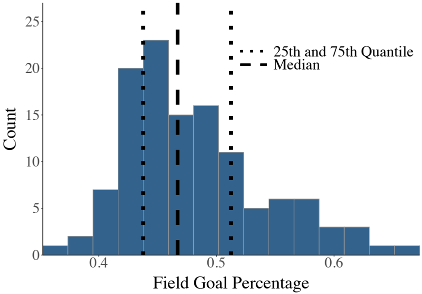
Bias-Corrected Estimates for Shooter 109

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

Comparison to NBA Shooting





Taking Stock

A Tempting Conclusion:

- Streakiness is confined to a small number of players
- Small number of shooters with very hot hands

We Argue:

1. Deviation from randomness by Shooter 109 is unlikely to be indicative of what could realistically be expected from even a small proportion of basketball players
2. Existing controlled shooting experiments do not have sufficient power to detect what would be realistic alternatives

Outline

Permutation Asymptotics under the Null Hypothesis

Goal: Approximate power of the permutation tests that we apply

Approach: Characterize the asymptotic behavior of the permutation distributions

Theorem (Asymptotics under the Null Hypothesis)

Under H_0^i, we have that

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

as $n \to \infty$, where T is equal to D or P if $T_n = \hat{D}_{n,k}(X_{i1},\ldots,X_{in})$ or $T_n = \hat{P}_{n,k}(X_{i1},\ldots,X_{in}) - \hat{p}_{n,i}$, respectively

Proof Outline

- **To Show:** Hoeffding's condition, i.e., $\sqrt{n}T_n(\mathbf{X}_i)$ and $\sqrt{n}T_n(\mathbf{X}_{i,\Pi})$ are asymptotically independent, with Π a random permutation
- **Idea:** Condition on Π
  - Apply a central limit theorem of Stein (1986) for dependent random variables
- Hoeffding's condition will not hold for every permutation π
  - e.g., the identity permutation
- We show that it holds with probability $\to 1$ as $n \to \infty$

Permutation Asymptotics under Alternatives

- Consider $\hat{D}_{n,1}(\mathbf{X}_i)$ for simplicity
- Let $\hat{a}_n$ denote the number of ones in the first n elements of $\mathbf{X}_i$

Theorem (Permutation Asymptotics under Alternatives)

If $n^{-1/2}(\hat{a}_n - np_i)$ converges in distribution to some limiting distribution as $n \to \infty$, then

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

with $T_n = \hat{D}_{n,1}(X_{i1}, \ldots, X_{in})$, as $n \to \infty$, where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function

Proof Outline

- **Idea:** Consider fixed (nonrandom) sequences
  - Let $L_n(h)$ be the permutation distribution based on a_n ones with
    $$h = n^{-1/2}(a_n - np)$$
  - We show that if $h_n \to h$, i.e., $n^{-1/2}(a_n - np) \to h$, then $L_n(h_n) \xrightarrow{d} N(0,1)$
- Complete argument with appeal to almost sure representation theorem

Outline

A Class of Markov Chain Streaky Alternatives

Goal: Measure the power against streaky alternatives

- There are s individuals; each associated with a Bernoulli sequence $\mathbf{X}_i = \{X_{ij}\}_{j=1}^{n}$ of length n
- Two types:
  - **Random:** $\mathbf{X}_i$ is i.i.d. Bernoulli(p_i)
  - **Streaky:** Probability of a one increases after streak of m ones by ε; probability of a one decreases after streak of m zeros by ε
- Each individual is Streaky with probability ζ, i.e., number of streaky individuals is Binomial(s, ζ)
- Specialize to the case $p_i = 0.5$

Analytic Power Approximation

Theorem

Power of permutation test of H_0^i against the alternative that individual i is streaky with $\varepsilon = h/\sqrt{n}$ converges to

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

for T equal to P or D when T_n is equal to $\hat{P}_{n,k}\left(\mathbf{X}_i\right) - \hat{p}_{n,i}$ or $\hat{D}_{n,k}\left(\mathbf{X}_i\right)$, respectively

Corollary

Power of stratified permutation test of H_0 against the alternative that each individual is streaky with probability ζ and $\varepsilon = h/\sqrt{ns}$ converges to

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

for T equal to P or D

Analytic Power Approximation

Example: If we use $\bar{D}_1$ then set

$$ns = \left(\frac{z_{1-\alpha} - z_{1-\beta}}{2\zeta\varepsilon} \right)^2$$

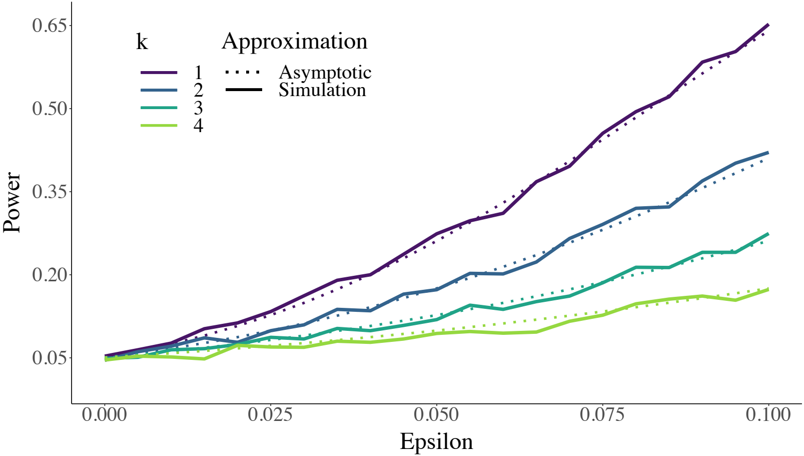
for power β

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

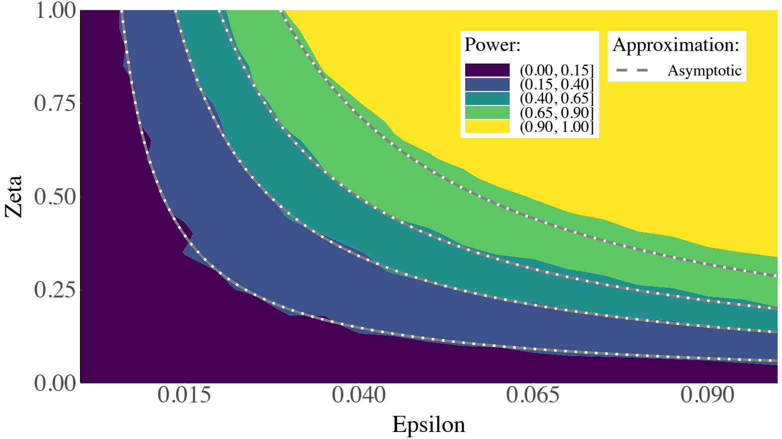
Simulation Analysis

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



Simulation Analysis

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

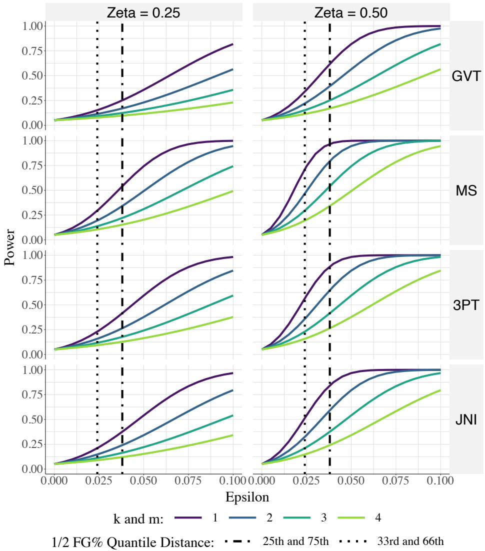


Outline

Returning to the Hot Hand

- **A Reasonable Alternative**:
  - The marginal shooting percentage for all shooters is 50%
  - For half of shooters, shooting percentage increases and decreases by half of interquartile range of distribution of field goal percentages of NBA players after making their previous m shots or missing their previous m shots
  - Other half is always i.i.d.
- This is equivalent to $\varepsilon = 0.038$ and $\zeta = 0.5$ in our model
- We argue this is a conservative upper bound and consider relaxations
- For GVT and MS, $m = 3$ is of particular interest

Returning to the Hot Hand



Taking Stock Redux

- Existing randomized shooting experiments are insufficiently powered to detect deviations from randomness consistent with realistic parameterizations of positive dependence in basketball shooting
- Unable to provide an informative estimate of the mean or dispersion of the serial dependence in basketball shooting
- Could be challenged by a strong and robust rejection of H_0
  - At least in the case of the GVT experiment, rejection of H_0 is sensitive to inclusion of an outlier.
- This result cuts both ways

Outline

Do People Overestimate Positive Serial Dependence?

- **Testing the Hot Hand Fallacy:**
  - Compare $\bar{\theta}_P^k(\mathbb{P})$ and $\bar{\theta}_D^k(\mathbb{P})$ with audience's expectations of $\bar{\theta}_P^k(\mathbb{P})$ and $\bar{\theta}_D^k(\mathbb{P})$
- **Available Evidence on Beliefs:**
  - Prohibitive methodological flaws (e.g., the GVT survey)
  - Not directly comparable to estimates of $\bar{\theta}_P^k(\mathbb{P})$ and $\bar{\theta}_D^k(\mathbb{P})$ (e.g., the MS survey)
- **Alternative Design**: Eliciting probabilistic expectations
  - Observer of a shooter asked to record expectation of probability shooter makes their next shot
  - If shot made, rewarded for submitting large probabilities, etc.,

Outline

Conclusion

- **Objective:** 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
- Substantially larger data sets are required for informative estimates of the streakiness in basketball shooting
- We reject i.i.d. shooting consistently for only one participant in the GVT shooting experiment
  - Strong evidence that basketball shooting is not perfectly random
  - Evidence against randomness limited to that player
- Future research should directly test the accuracy predictions of streakiness in stochastic processes

Thank You!