Hot or Not? A Nonparametric Formulation of the Hot Hand in Baseball

Amanda K. Glazer
Department of Statistics
University of California Berkeley
amandaglazer@berkeley.edu

Lisa R. Goldberg
Department of Statistics
University of California Berkeley
lrg@berkeley.edu

March 18, 2020

Abstract

A nonparametric analysis of player plate appearances (PA) in the 2018 Major League Baseball (MLB) season provides no evidence of a batter hot hand. Players with more than 100 PAs in the 2018 season are analyzed using one-sided permutation tests stratified by player. Based on recent literature, we use the correlation between lagged on-base percentage (OBP)and a binary indicator of on-base performance. We discuss the strengths and weaknesses of this test statistic as well as others in the literature. A common criticism of no-hot-hand findings for individual players is low power, and a frequently proposed remedy is pooling data across players. Through simulation, we show that pooling data conflates long-term ability and recent performance. Another common criticism of no-hot-hand findings is emphasis on recent performance. We show that long lags, which de-emphasize recent performance, can lead to counter intuitive results. In contrast to much of the recent literature, which uses parametric methods, we argue that our nonparametric method is the most appropriate way to analyze the existence of the hot hand in baseball as well as numerous other inference questions.

> I never blame myself when I'm not hitting. I just blame the bat and if it keeps up, I change bats. After all, if I know it isn't my fault that I'm not hitting, how can I get mad at myself?
>
> - Yogi Berra

1 Introduction

Streaks of hits and misses are familiar to everyone who plays or watches sports. Performance seems to be dictated by magical streaky periods in which players appear to be "hot" or "cold." Are we to believe in these streaky periods, or should we believe Yogi Berra that there is no one to blame for strings of hits and misses, and they are especially not the fault of the players themselves?

The hot hand in sports is the idea that a player who has recently been successful has an elevated probability of success in the near future. The hot hand has been a hot topic since Gilovitch, Vallone, and Tversky's 1985 study [7] found no statistical evidence of a hot hand in basketball. They argued that, perceptions notwithstanding, a basketball player's chance of making a shot showed no dependence on whether it followed a hit or a miss [7]. This classical study examined the difference in player i's shooting percentage conditional on having made the last k shots and missed the last k shots (with $k = 1, 2, 3$):

$$\hat{P}^i(\text{hit}|k \text{ hits}) - \hat{P}^i(\text{hit}|k \text{ misses})$$

Paired t-tests indicate that there is no significant difference in the chance that a player makes the next shot based on whether they missed or hit the last k shots.

The findings in the original study were widely debated, with many players and sports enthusiasts refusing to believe that the hot hand did not exist. Thirty years after its publication, this study was found to contain a small sample bias that some argue led to the incorrect conclusion [4]. There is no consensus on whether the hot hand exists or not in basketball.

The original study and much of the literature since, including recent studies on baseball, have largely focused on parametric models to test for the hot hand [3, 2]. These parametric models rely on a significant p-value of the coefficient of interest (generally in a linear model) to conclude that a hot hand exists. They usually control for a number of factors, such as ballpark characteristics and ability of the pitcher.

In order to delve more deeply into the question of whether the hot hand exists in baseball, we must clearly formulate our question. In analyzing whether the hot hand exists, researchers tend to focus their attention on one of the following two questions:

1. Does a player perform better if they have performed well recently (e.g., in the last L plate appearances or shots), outside of the effects of all other factors?

2. Does fan's *perception* of the hot hand, players that have performed well recently will continue to do so, exist?

The key difference in these questions is whether we think it is important to control for all other factors when searching for the existence of the hot hand. We argue that most fans and players perceiving a hot hand are not mentally adjusting for factors. Rather, they are reacting to a streak. Generally, someone does not think, "LeBron James has been making a higher number of shots then you would expect given the opposing team, level of defense, arena, day of the week, etc." A more realistic formulation centers around a fan's heart rate increasing because LeBron James just made his tenth basket in a row. While question 1 is interesting and has its own merits and challenges (e.g., it is difficult to take into account all factors that could affect performance), for this paper we will consider question 2 because we believe it is a more realistic formulation of the hot hand phenomenon.

In trying to sort out the hot hand issue outlined in question 2, some previous research in basketball has considered a nonparametric approach stratified by player. For example, in Daks, Desai and Goldberg (2018), strings of hits (1) and misses (0) were considered (e.g., "100101110") for Steph Curry, Klay Thompson, and Kevin Durant. The proportion of "11"

followed by "1" minus the proportion of "00" followed by "1" was used as a test statistic in a permutation test. This study found no evidence that the Golden State Warriors players exhibited hot hands [1].

Relative to a parametric analysis, a nonparametric approach is based on fewer assumptions and is conceptually clearer. For example, the p-value of a permutation test is based on the proportion of random shufflings of a variable that result in a test statistic larger than or equal to the test statistic calculated on the actual data. The core assumption is that if we believe that the shuffled variable does not make a difference, then we should be able to scramble its realizations without significantly changing the test statistic.

While most researchers have not found evidence of the hot hand in baseball, those that have found evidence have relied on data pooled across players rather than considering them individually; see, for example, [2, 6]. While the motivation for pooling data in order to increase statistical power is understandable, pooling data leads to a large type I error rate and erroneous conclusions. We demonstrate this with a simple simulation.

In this paper, we adopt permutation tests, a nonparametric method, stratified by player, to examine whether or not the hot hand exists in baseball. This analysis is inspired by Daks, Desai and Goldberg's 2018 study (although using a different test statistic) [1]. We illustrate the benefits of a nonparametric approach over a parametric approach, demonstrate the perils of pooling data, and compare various test statistics. Finally, we illustrate the potential use of nonparametric analysis for a variety of inference questions in baseball.

2 Data and Methodology

2.1 Data

We used play by play Major League Baseball (MLB) data from all 30 teams in the 2018 season. The data is publicly available from Retrosheet.org. We analyzed all 447 players with more than 100 plate appearances in the 2018 season. Table 1 gives summary statistics about the OBP and number of PAs for the 2018 players included in the analysis.

	Max	Mean	Minimum
OBP	0.460	0.315	0.162
PA	745	387	101

Table 1: Summary statistics about the 447 players from 2018 included in the analysis.

2.2 Defining the batter hot hand in baseball

As discussed in the introduction, we consider the following question:

> Is a batter with a higher on base percentage (OBP) over the last L plate appearances (PAs) more likely to get on base at their next PA?

OBP is approximately equal to the number of times on base divided by PAs. If the batter hot hand exists, the answer to this question would be yes, higher OBPs over the last L PAs should be positively correlated with whether the batter gets on base in the next PA. In order to investigate this, we used permutation tests stratified by player and season.

The binary on-base vector OB_k has jth entry equal to 1 if player k got on base at the jth PA and 0 if not. Following Green and Zwiebel (2018)[2], we define player state to be OBP for the last L PAs, where the lag L can take on different values. Let state_{kj} be the state of player k just prior to plate appearance j,

$$\text{state}_{kj} = \frac{1}{L} \sum_{i=j-L}^{j-1} OB_{ki},$$

and let state_k denote the player k's vector of states.

2.3 Test Statistic

Our test statistic T_k is the correlation between OB_k and state_k:

$$T_k = \text{corr}(OB_k, \text{state}_k).$$

If there is a batter hot hand, we would expect higher state values to correlate with a higher likelihood of reaching base. Thus, we would expect that a batter hot hand would lead to larger test statistic values relative to random shufflings of player PAs. This test statistic is similar to Green and Zwiebel (2018)'s use of the state coefficient, in the logistic regression regressing on base indicator on state as well as other controls, as a measure of the batter hot hand.

In Section 3 we consider other possible test statistics, autocorrelation and the test statistic in the Gillovich et al. study [7], but find that T_k has the highest power among the test statistics we consider.

2.4 Permutation Tests

For each player in our data set, we ran a one-sided permutation test. The null hypothesis is that state has no effect on outcome of the next PA.

If the null hypothesis is true, then shuffling the PAs, recalculating state, and calculating the correlation between state and OB should lead to a correlation that is not too different from our original value for player k. The p-value is the proportion of shufflings that result in a correlation as large or larger than the correlation we observed.

Advantages of permutation tests over parametric methods are that permutation tests make fewer assumptions, and those assumptions they do make are conceptually very clear. For example, permutation tests do not make assumptions about the distributions from which data are drawn. Instead, the p-value in a permutation test is derived from the assumption that under the null hypothesis, you should be able to shuffle, for example, player PAs and not see a large difference in the test statistic if the hot hand does not exist.

We must acknowledge and correct for the fact that we are running multiple hypothesis tests, and some will show significance purely by chance. The same is true for running

numerous regression models. Under the null hypothesis, we expect, for a significance level α, that a fraction α of the tests will lead to a false rejection (i.e., an indication of a hot hand when it is not there). In other words, if there is no hot hand, we expect nevertheless to see significance at level α in a fraction α of the tests.

2.5 Choice of Lag

There is no clear choice for the lag L. Green and Zwiebel (2018) used lags of length 10, 25 and 40. In streaky data, counter-intuitive results can occur when the lag length is longer than the streak length. We ran several simple simulations to illustrate potential issues with certain lag lengths.

Consider a string of plate appearances composed of 24 1s followed by 24 0s followed by 24 1s and so on for the entire season. We would hope that our test would pick up on the extreme streakiness in this sequence. However, a lag 25 results in a *negative* correlation of -0.136 between OB and state. On the other hand, lag 10 and lag 5 result in positive correlations. A rule of thumb is that our test statistic will miss a streak that is shorter than the lag used to compute state.

For this reason, lags of length 10 or 5 seem more appropriate as they seem more likely to pick up on a variety of streak lengths. In our hot hand analysis, we used values of 5, 10 and 25 for L to evaluate if results are sensitive to the length of history used to calculate state.

3 Type I Error and Power

Previous studies have pooled data across players to increase power in parametric tests for the hot hand in baseball [2, 6]. But pooling data can lead to erroneous significant results. If we do not pool data, however, we might be concerned about the ability of our permutation tests to correctly identify a hot hand if it does exist. Simulations can give us a general sense of the power of permutation tests in this setting.

3.1 Pooling Data

Pooling data can inflate the type I error rate. Consider, for example the simple simulation where we generate 400 baseball players, each of which has a different OBP p evenly spaced from .250 to .450. For each player, we generated on base indicators for 500 PAs from the binomial distribution: $Binom(500, p)$. There is no hot hand because data is generated at random from this binomial distribution.

We follow the same set up as in Green and Zwiebel (2018) [2]. Ability is defined as the OBP of a player outside of the 50 PAs before and after the current PA. If we regress the current plate appearance on state (as defined in the previous section with $L = 25$) and ability, we would hope that state would not be significant because there is no hot hand in this simulation. However, over 1000 simulations, we find that 99% of the time, state *is* significant (at the 0.05 level). In other words, the type I error rate (or false positive rate) is 99%.

The source of the significance is differences in player OBP and not a hot hand. However, the linear regression conflates differences in OBP with differences in state, even when

we account for an ability variable. We must be cautious about drawing conclusions from regressions that pool data across players with varied abilities.

3.2 Power

Since we are concerned with the hot hand at the player level, and since pooling data can lead to high type I error, it makes sense to consider tests for streakiness at the player level. However, previous research has expressed concern about the potential loss of power due to the smaller sample size when considering players individually rather than in aggregate [2, 6].

Before we run permutation tests that are stratified by player, we investigate and address power concerns. There is no closed form formula for power in permutation tests. However we can use simulation to assess power is in various situations. There are lots of ways that binary sequences can deviate from random. We consider two different ways of creating correlated binary strings and evaluate the power of our stratified permutation tests in each.

3.2.1 Autocorrelated Binary Strings

There is no precise level of autocorrelation in a binary string that corresponds to a hot hand, the two concepts are loosely connected. By generating binary strings with different levels of autocorrelation, we can get a better sense of what level of autocorrelation our method is able to detect.

We simulate correlated binary variables with specified marginal means and correlations. Then, we calculate the probability that we can detect a hot hand (with lag equal to 5, 10 and 25), as characterized by correlation between state and on base performance, using a permutation test. We simulate binary variables $Y_1, ..., Y_n$) with correlation ρ using the conditional linear family method as outlined in [5]:

1. Generate $Y_1 \sim \text{Bern}(p)$
2. Generate $Y_2, ..., Y_n$ each with mean $E(Y_j|Y_{j-1}) = p + \rho(y_{j-1} - p)$

We set $p = 0.318$, the average OBP for the 2018 MLB season (for all players), and $n = 500$. We consider several different values of the correlation, ρ, which gives us an idea of what the power (for significance level 0.05) would be for our permutation tests with lag 5, 10 and 25. Results are reported in Table 2. Power is smallest for $L = 25$ and largest for $L = 5$. Both $L = 5, 10$ have fairly reasonable power for autocorrelation greater than or equal to 0.4.

3.2.2 Markov Model

Next, we consider a two-state Markov model with transition probability 0.05. The two states are hot and cold. We consider hot OBP/cold OBP of 0.6/0.2, 0.55/0.25 and 0.5/0.3. When a player is in a hot (cold) state whether he makes it on base will be a Bernoulli draw with probability equal to the hot (cold) OBP. According to this model we generate binary strings of length 500 and over 1000 simulations calculate the proportion of time that our permutation test methodology with various test statistics will detect a hot hand.

Lag	ρ		
	0.2	0.4	0.6
5	0.635	0.991	1
10	0.343	0.841	0.994
25	0.139	0.478	0.761

Table 2: Power (significance level is 0.05) over 1000 simulations for permutation tests defined in section 2 for various lag lengths and correlated binary variables with correlation ρ. Marginal mean is set to 0.318 and n is set to 500.

Test Statistic	Hot/Cold OBP		
	0.6/0.2	0.55/0.25	0.5/0.3
Lag 5 Correlation	1	0.88	0.32
Lag 10 Correlation	0.99	0.85	0.34
Autocorrelation	0.9	0.51	0.18
Gilovich et al. Statistic	0.97	0.65	0.23

Table 3: Power (significance level is 0.05) over 1000 simulations for four different test statistics for a two-state markov chain with transition probability 0.05 and various hot/cold OBPs. Binary strings of length 500 were generated.

The results in Table 3 show that in our two-state Markov model simulations, power dropped off steeply as the difference between hot and cold states gets smaller. The Lag 5 and 10 correlation test statistics perform the best.

Results from both of our power simulations indicate that smaller lags have higher power. Therefore, we believe that lag 5 and 10 are more reasonable lags to consider in our tests. Not only can larger lags such as 25 and 40 produce counter-intuitive results, but larger lags also appear to have lower power.

4 Hot Hand Results

We found no evidence of a hot hand in the 2018 MLB season. Table 4 shows the proportion of player p-values that were significant at level α for α equal to 0.05 and 0.1, with state calculated using lags 5, 10 and 25. Under the null hypothesis that there is no streakiness: we would expect the proportion of p-values that are significant at level α to be close to α.

Figure 1 shows CDFs for player p-values from the permutation tests with state calculated using lags 5, 10 and 25. Under the null hypothesis that there is no streakiness, we expect the CDF to look similar to the uniform distribution.

Our results are in line with what we would expect for the type I error rate, if there were no hot hand. Results from the 2012 season (see Appendix A) show similar results. Our analysis yields no evidence of the batter hot hand in baseball. Other permutation test

α	Lag		
	5	10	25
0.05	0.087	0.067	0.076
0.01	0.018	0.020	0.025

Table 4: Proportion of p-values for player permutation tests significant at level α for state calculated with various lag lengths.

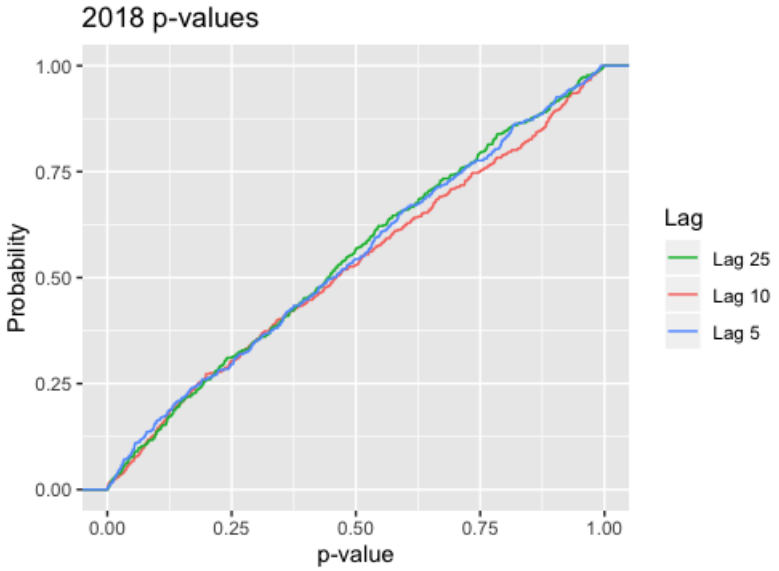

Figure 1: CDFs for player p-values for permutation tests with L = 5, 10 and 25

statistics, such as autocorrelation and the state coefficient in a linear regression, also show similar results.

5 Other Nonparametric Formulations of the Hot Hand in Baseball

In the previous section, we looked for streakiness in strings of player plate appearances with no attention to clustering by game. In this section, we consider other formulations of the hot hand in baseball, and we analyze them with non-parametric methods.

To analyze streakiness at the game level, we formulate the hot hand as whether a player exhibits cross-game streakiness. In other words, do games in which a batter has above average performance cluster? This clustering is a necessary (but perhaps not sufficient) condition for cross-game hot hand. There are challenges to this formulation. For example, the game OBPs will be very noisy due to the relatively small number of plate appearances in a game. However, it will give us a sense of whether a player exhibits cross-game streakiness or not.

Consider, for example, Mike Trout's 2019 season. His OBP for the 2019 was 0.438, and

we calculate his OBP for each game. For his plate appearances in a game, we calculate the proportions of hits (H), walks (BB) and hits by pitch (HBP, Trout's game OBP). We then create an indicator variable, called above average indicator, that is 1 when the game OBP is above his season average and 0 when it is below his season average. If Trout were to exhibit streaky playing, we would expect that the autocorrelation of his above average indicator to be high relative to random shufflings of his game OBPs.

We run a permutation test in order to evaluate this claim. We randomly shuffle the game level OBP and calculate the autocorrelation of the above average indicator. The p-value is the proportion of random shuffles or permutations that have autocorrelation greater than or equal to the observed autocorrelation. The p-value is 0.743, providing no evidence that above and below average OBP games are clustered together.

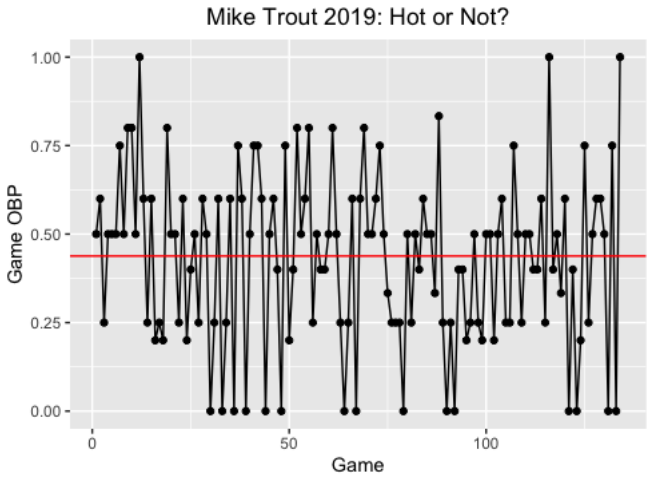

Figure 2: Game OBP for Mike Trout's 2019 season

From Figure 2, it is clear that there are many game OBPs that are close to his season average. It is not necessarily true that a streak of 0.5 game OBPs would feel like a hot hand to a fan. However, since we observe an insignificant p-value, there is no reason to believe that even that level of streakiness exists.

6 Other Applications of Nonparametric Analysis in Baseball

Permutation tests are the ideal methodology for many inference questions in baseball. Since permutation tests are nonparametric, they make minimal assumptions about the data, as opposed to parametric methods such as linear or logistic regression, which make numerous assumptions. In this section, we sketch examples of other uses of permutation tests to answer inference questions in baseball.

Suppose we want to evaluate the dependence of a batter's OBP on the number of outs prior to plate appearance. Consider Mike Trout's 2019 season as an example.

If the number of outs do not make a difference in the likelihood that Trout will get on base or not, then we should be able to shuffle the columns that indicates the number of

Outs	0	1	2
OBP	.430	.449	.431
PA	200	247	153

Table 5: Mike Trout's OBP and PA based on number of outs for the 2019 season.

outs for a PA and not see much of a difference. Using a chi-squared test statistic in the permutation test, we get a p-value of 0.90. There is no evidence that Trout's OBP depends on the number of outs at the time of his plate appearance.

We use similar analysis to evaluate whether Trout's OBP is significantly different for home versus away games. In 2019 Trout had an OBP of .450 for home games (280 PA) and .428 for away games (320 PA). Running a permutation test with difference in OBP between home and away games as the test statistic yields a p-value of 0.326. There is no evidence of a significant difference between Trout's 2019 OBP at home versus away games.

A similar method could be used to evaluate whether OBP (or any other statistic) varies depending on whether or not it is a clutch situation.

7 Conclusion

Nonparametric tests of player on base performance showed no evidence of the hot hand phenomenon, despite its perception by sports enthusiasts. We ran permutation tests on all MLB players with more than 100 plate appearances in the 2018 season. We used lags of 5, 10 and 25 plate appearance to determine hot and cold states, and we asked whether these states were correlated with the subsequent plate appearance. The proportion of tests that were significant aligned with the type I error rate, providing no evidence that the hot hand as formulated exists.

Crucially, the permutation tests are stratified by player, so the results are not corrupted by the conflation of state and ability, which plagued the results in Green and Zwiebel 2018 [2]. Through simulation, we show that if there is no hot hand but players with varying OBP are pooled together then linear regression will reveal a hot hand effect when there is not one. Since permutation tests stratified by player adequately control for the type I error rate at the cost of statistical power. It is clear that this is the proper tradeoff.

In general, nonparametric tests are the ideal methodology for answering many inference questions in baseball because they rely on relatively few assumptions, and the assumptions they do make are conceptually very clear. If we believe that a factor should not make a difference in outcome, then we should be able to shuffle the realizations of that factor and not see much difference in our chosen test statistic. There are no assumptions about the distribution of our data which could lead to erroneous results. We advocate for the use of permutation tests for inference questions in baseball.

A Appendix

A.1 2012 Results

We also ran our permutation test analysis on players with more than 100 plate appearances in the 2012 season (459 players). Our results are in line with our 2018 results: we find no evidence of a batter hot hand.

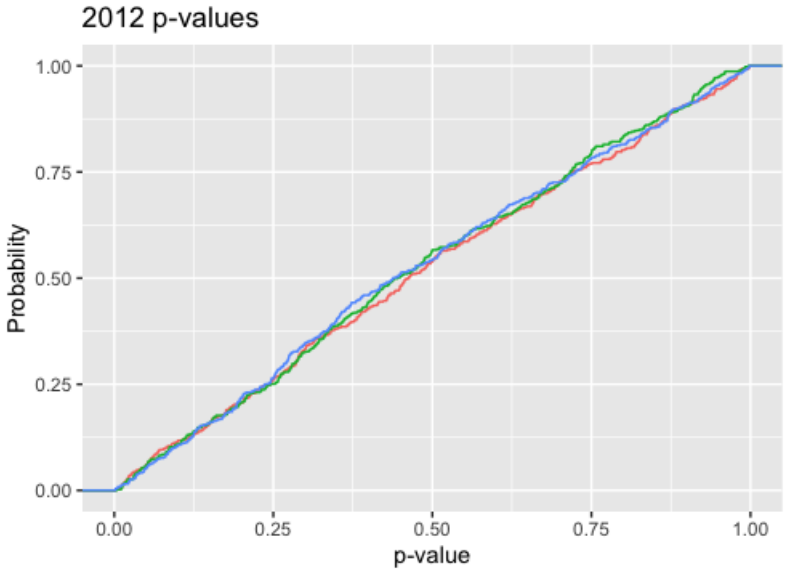

Figure 3: CDFs for player p-values for permutation tests for lags $L = 5, 10, 25$ for the 2012 season

α	Lag		
	5	10	25
0.05	0.052	0.059	0.059
0.01	0.009	0.004	0.002

Table 6: Proportion of p-values for player permutation tests significant at level α for state calculated with various lag lengths for all players with more than 100 PAs in the 2012 MLB season.

References

[1] N. D. Alon Daks and L. Goldberg. Do the golden state warriors have hot hands? *The Mathematical Intelligencer*, 2018.

[2] B. Green and J. Zwiebel. The hot-hand fallacy: Cognitive mistakes or equilibrium adjustments? evidence from major league baseball. *Management Science*, 2018.

[3] S. A. Michael Bar-Eli and M. Raab. Twenty years of "hot hand" research: Review and critique. *Psychology of Sport and Exercise*, 2006.

[4] J. Miller and A. Sanjurjo. Surprised by the hot hand fallacy? a truth in the law of small numbers. *Econometrica*, 2018.

[5] J. S. Preisser and B. F. Qaqish. A comparison of methods for simulating correlated binary variables with specified marginal means and correlations. *Journal of Statistical Computation and Simulation*, 2014.

[6] H. S. Stern and C. N. Morris. A statistical analysis of hitting streaks in baseball: Comment. *Journal of the American Statistical Association*, 1993.

[7] R. V. Thomas Gilovich and A. Tversky. The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology*, 1985.